Universidade De Vigo
Departmento de Enxeññería Telemática
Grupo de Tecnologías de la Información

Doctoral Thesis

International Mention

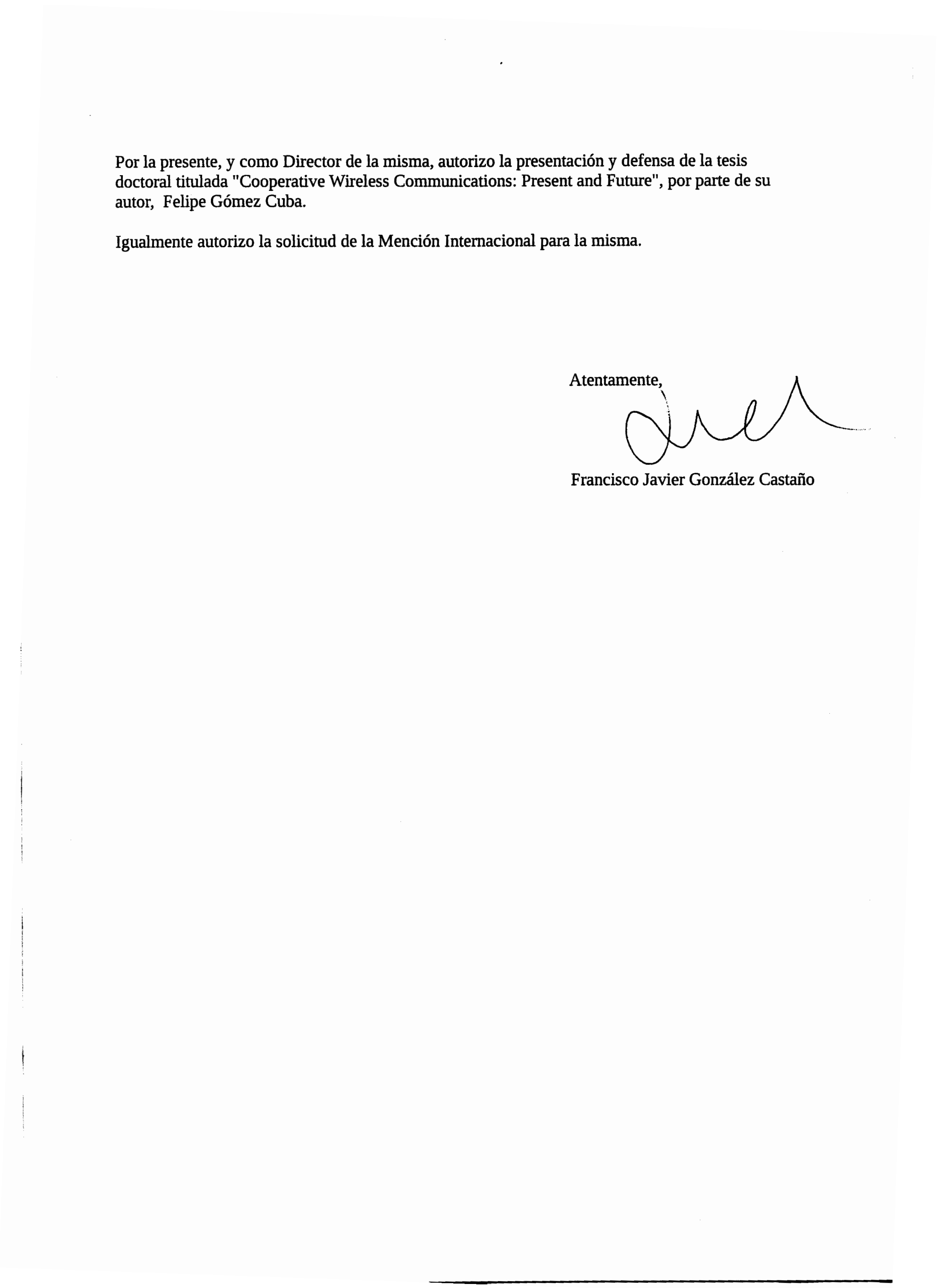
Cooperative Wireless Communications: Present and Future

Author:
Felipe Gómez Cuba

Supervisor:
Prof. Francisco J. González Castaño

Submitted in fulfilment of the requirements
for the degree of Doctor por la Universidad de Vigo

March 10, 2015



Por la presente, y como Director de la misma, autorizo la presentación y defensa de la tesis doctoral titulada "Cooperative Wireless Communications: Present and Future", por parte de su autor, Felipe Gómez Cuba.

Igualmente autorizo la solicitud de la Mención Internacional para la misma.

Atentamente,

Francisco Javier González Castaño

"Everyman for himself is not going to work. It's time to start organizing. We need to figure out how we're going to survive here. Now, I found water. Fresh water, up in the valley. I'll take a group in at first light. If you don't want to go come then find another way to contribute. Last week most of us were strangers, but we're all here now. And god knows how long we're going to be here. But if we can't live together, we're going to die alone."

Jack Shepard

Lost

Agradecementos

Esta tese non vale nada se non é capaz de repagar, polo menos nunha pequena parte, o traballo de tantos que me apoiaron e sufriron conmigo ao longo de todos estes anos.

En primeiro lugar, quero agradecer á miña familia todo o seu apoio e fe. A meus pais, pois de non ser por eles que me deron a vida, o esforzo e o sacrificio invertidos neste longo recorrido non terían razón de ser. E a meus tíos e padriños que me apoiaron aquí en Vigo.

En segundo lugar, a todos aqueles que se embarcaron comigo nesta singradura. Moi poucas persoas teñen a fortuna de poder dicir que teñen demasiados amigos para ser capaces de contalos a todos. Será mellor que non o intente. Aos amigos de toda a vida, quero agradecer a paciencia que tivestes cando deixei de estar ao lado todos os días e o calor co que me recibides sempre que podo estar. Aos amigos que fixen en Vigo, agradecer todos os momentos compartidos estes anos e todo o que aprendín xunto con vos. Quero facer unha mención especial para os que estiveron comigo cando cheguei a esta cidade, a os compañeiros de piso e de coche de volta á casa, aos que empezaron o doutorado comigo, aos que se apuntaban aos bolos o Xoves, e unha grandísima aperta final para os que nos atopamos de novo nestes últimos anos para compartir as últimas etapas desta viaxe.

En terceiro lugar, debo moitísimo do acadado nesta tese aos compañeiros do GTI, de Gradiant, e do resto da EET. Quero agradecer especialmente a o meu titor Javier Castaño, por terme guiado e apoiado en cada tolería que se me pasou pola cabeza durante todos estes anos. Aos Compañeiros do grupo que traballaron ao meu lado, e especialmente a Rafa Asorey, Jorge Muñoz, Carlos Perez e Juan García que traballaron en publicacións desta tese. Aos antigos compañeiros de Gradiant, que me acolleron cando me unin a eles e me ensinaron tanto do mundo. Tamén quero agradecer aos outros profesores e persoal da EET que me axudaron, e en especial a Nuria González que estivo dispoñible para resolver as miñas dúbidas.

I would like to thank all the people at NYU Polytechnic School of Engineering that allowed me to join them for a brief time. Thanks to Profs. Elza Erkip and Sundeep Rangan for working with me, challenging my skills, and putting up with my endless stream of words. And thanks also to the students in their groups, that worked with me and showed me so many new things.

Finally, I would like to thank the colleages that met, advised and questioned me, at conferences and other visits. All the people that works to keep the research world ticking, and allow us to keep dreaming. And all the people who shared the wind with me in each leg along this long journey.

Abstract

Wireless data traffic grows exponentially while, at the same time, communications in known point-to-point channels seem to have almost achieved the Shannon Limit. In order to satisfy demand the fundamental point-to-point limitation needs to be circumvented by considering communications in wireless networks jointly in all nodes. Cooperative wireless communications achieve greater communication performance through networked collaboration not unlike the networking of computers has empowered computation.

This topic covers a broad range of theoretical and practical aspects, as well as the evolution of wireless networks in an extended span of time from the state of the art and current standards, to the backwards-compatible modification of current technologies in the near future, while they are still in the middle of their life cycle, and the preparation of theoretical ground works for the writing of new standards in the distant future.

The first part of the thesis focuses on the present. In the theoretic aspect, a survey is performed on the state of the art in research literature about the analysis of cooperative diversity gains at multiple levels of the protocol stack. In the practical aspect, the implementation of a detailed system level simulator puts to test the rules for relaying in the most recent cellular standard, which is an early instance of cooperation.

The second part of the thesis focuses on the near future. The scarcity of spectrum in the microwave bands employed by current technologies is identified as a major problem that can be tackled with cooperation. In the theoretic aspect, an analytical study is performed to compute the social spectrum gains in cognitive radio cooperative spectrum leasing mechanisms. In the practical aspect, the results illustrate that the limited gains of cooperation in current standards can be improved in some emergent application niches such as machine-to-machine communications, much more suitable to be used with cooperative spectrum leasing than personal data services.

The last part of the thesis focuses on the distant future. The spectrum crunch will be palliated through the introduction of massive degrees of freedom made possible by moving communication up to millimeter-wave bands that allow more bandwidth, antennas and node density. In the theoretical aspect, a scaling law analysis proves that there is a limit to the increment of capacity through massive resources, and that said limit must be pushed further back with cooperative multi-hop communications in new standards. In the practical aspect, a summary covers the future research topics in information theory, physical and medium access control layers that must pave the way for the design of the hardware in future wireless networks.

Resumo (Galego)

Segundo informes publicados recentemente pola industria [1], as redes en fíos modernas están a experimentar un intensivo incremento na cantidade de trafico de datos que debe ser cursada polo sistema, tanto na cantidade de datos xerados por cada usuario como na expansión do número de usuarios. Nembargantes, á vista dalgunhas declaracións efectuadas por representantes da industria parecera que non é posible aumentar máis as taxas de transmisión nos enlaces sen fíos, que estarían xa preto do que estes individuos chaman o Límite de Shannon.

> “A industria sen fíos chegou ao limite de como de rápido poden ir as redes.”
>
> K. Fitcher, revista Connected Planet

> “Estamos ao 99%” do camiño “cara a barreira coñecida como Límite de Shannon.”
>
> D. Warren, Director de Tecnoloxía da Asociación GSM

A clave para entender como se pode harmonizar perspectivas tan dispares está na distinta definición formal de ambas medidas do tráfico. O feito é que as palabras Límite de Shannon tal e como se entende por parte dos enunciados antes mencionados non son formalmente correctas, en tanto que fan referencia soamente a unha pequenísima parte da teoría desenvolvida polo coñecido matemático. Deste xeito, mentres que a primeira clase de perspectivas optimistas se centraría no volume total de datos intercambiado dentro dunha rede enteira, a segunda clase de perspectivas pesimistas estaría guiada por unha confusión do ”todo pola parte”, constatando que efectivamente estamos próximos ao Límite de Shannon, pero soamente na taxa acadable por un único enlace punto a punto entre dous dispositivos situados no baleiro. En realidade, o Límite de Shannon para a capacidade das comunicacións dunha rede con tres ou máis nodos (coa definición correcta) é un problema aínda por resolver, pero é facilmente demostrable que dito límite está moi por riba da taxa dunha simple configuración cun único enlace entre dous nodos que podemos acadar hoxe en día.

O significado desta distinción entre a taxa acadable por unha rede e a taxa acadable entre dous nodos illados é que, se queremos incrementar a capacidade das nosas redes para satisfacer a demanda, debemos deixar de pensar nas redes coma se foran simples coleccións de conexións un a un que simplemente teñen lugar unhas ao lado de outras.

No canto, debemos ter en conta a capacidade de terceiros para influír positivamente nas comunicacións entre un emisor e un receptor.

Nesta tese explóranse as ”comunicacións sen fíos cooperativas” nun sentido amplo. Esta é unha disciplina transversal que cobre calquera tipo de rede de comunicacións sen fíos en que se produce a intervención de terceiros na comunicación entre un transmisor e un receptor. Existen multitude de exemplos na literatura de distintos problemas deste tipo, os cales adoitan desenvolver as súas linguaxes e culturas propias, dificultando a fertilización entre distintos campos do saber. Varios exemplos de comunicacións cooperativas son.

- A canle relay en teoría da información
- O encamiñamento multi-salto en redes sen fíos
- A codificación de rede
- Os conxuntos de antenas virtuais
- A coordinación de interferencias
- A diversidade en planificadores multi-usuario.

Esta tese estuda o rol das comunicacións cooperativas nas redes sen fíos, para o cal se identificaron tres etapas distintas para a incorporación da cooperación nas redes sen fíos en xeral, e nas redes celulares de telefonía e datos en particular. A tese está dividida en tres partes que se corresponden con cada unha destas etapas. A primeira parte da tese céntrase no presente, en que o modelado teórico de canles cooperativas está relativamente desenvolvido, pero a súa implementación en dispositivos reais é practicamente inexistente coa excepción dalgúns leves xestos de aproximación como os relays estáticos na última versión do estándar de redes celulares. A segunda parte da tese céntrase no futuro próximo (aproximadamente dende o presente ata a década de 2020), que se corresponde coa extensión do ciclo de vida das tecnoloxías da actual xeración. Dado que neste período a compatibilidade entre dispositivos da mesma familia será unha prioridade, as posibles melloras han de chegar na forma de refinamentos na eficiencia de uso das mesmas frecuencias (microondas) e arquitecturas (dúplex estático e topoloxías en árbore). A última parte correspondese co futuro máis avanzado (a partir de 2020), cando se extinga a actual xeración de comunicacións e novos estándares tomen o lugar dos actuais. Durante este período abrirase unha ventá de oportunidade para a introdución de cambios drásticos na filosofía de deseño, e será posible expandir o espectro das redes cara novas bandas (ondas milimétricas) ou novas topoloxías (dúplex dinámico e topoloxía completamente conectada).

Cada unha destas tres partes consta de dous capítulos, de modo que os contidos desta tese se concentran en seis capítulos principais (mais unha introdución e as conclusións). Estes seis capítulos organízanse segundo dous eixos, un temporal e outro teórico-práctico. Como xa se dixo, o eixo do tempo correspondese coas tres partes nas que a tese está dividida, e que se corresponden coas tres etapas futuras da incorporación de mecanismos de cooperación nas redes sen fíos. Por outra banda, no eixo teórico-práctico cada parte da tese consta de dous capítulos. O primeiro capítulo de cada parte contén as contribucións nesta tese ao corpo teórico da cooperación nas comunicacións sen fíos. Así, os capítulos 2, 4 e 6 conteñen na súa maior parte discusión para cada etapa sobre o modelado teórico das transmisións cooperativas, e a análise da súa capacidade a través da teoría da información, demostrando que a cooperación introduce ganancias sobre cada modelo non-cooperativo correspondente. O segundo capítulo de cada parte, pola contra, contén as contribucións nesta tese de cara á realización práctica das redes sen fíos cooperativas. Así, os capítulos 3, 5 e 7 conteñen na súa maior parte discusión sobre a posible realización práctica de cada etapa. Sen perda de xeneralidade, tódolos exemplos prácticos nesta tese céntranse na aplicación da cooperación ás redes celulares de datos modernas. O estudo deste tipo de rede é de gran importancia dado o incremento exponencial na demanda de tráfico móbil de datos. Ademais, o aprendido sobre estas redes é de aplicación directa para moitas outras redes sen fíos (coma redes de emerxencia ou de sensores) dado que as redes celulares se contan entre os sistemas sen fíos máis grandes e complexos existentes.

Na primeira parte da tese, centrada no presente, o aspecto teórico do segundo capítulo presente unha escolma do estado do arte da análise das ganancias de diversidade cooperativa en distintos niveis da pila de protocolos. Os niveis estudados son tres:

- No nivel da teoría da información, a capacidade dunha transmisión cooperativa é estudada a través de modelos derivados da canle relay teórica. Nesta tese determinase que existen varios modelos con resultados diferentes, dependendo dunha serie de propiedades: A función de relay, que determina o sinal transmitido polo relay en función do recibido $P(X_r = f(Y_r))$ (como por exemplo AF, DF...), a división en tempo entre as transmisións da fonte e do relay (estática ou dinámica), a ortogonalidade de ditas transmisións (é dicir, se a fonte debe gardar silencio cando o relay transmite) e a posibilidade de engadir múltiples capas de codificación no mesmo sinal (como no protocolo Enhanced-DDF, que engade un terceiro sinal da fonte ao protocolo DDF).

- No nivel da capa física, observamos que o procesado necesario para combinar distintas sinais de distintas canles cooperativas pode ter lugar en distintos puntos da etapa de radiofrecuencia: pode realizarse no nivel de control, mediante a escolla do mellor relay para transmitir un único sinal; ao nivel da codificación de canle, mediante códigos distribuídos concatenados cooperativos; ao nivel de codificación de fonte, mediante codificación de rede; ou a nivel de procesado multi-antena, a través de códigos multi-antena distribuídos.

- No nivel da capa MAC, esta tese detectou que tódolos protocolos realizan cinco operacións fundamentais: mapeado dos veciños, deseño de conxuntos de relays óptimos, selección entre transmisión cooperativa ou directa, notificación aos relay, e deseño da transmisión cooperativa.

Pola outra banda, o aspecto práctico da primeira parte da tese, cuberto no terceiro capítulo, trata sobre a implementación de comunicacións cooperativas con relays no estándar LTE-A. Neste capítulo, implementouse un simulador de nivel de sistema detallado para verificar as regras especificadas no estándar. A simulación revelou diversos problemas en capas superiores da rede que investigacións previas de capas máis baixas tiñan ignorado.

- O uso dun parámetro global en común a tódalas celas e relays para determinar a división en tempo entre conexións dos usuarios e dos relays impide que se poda balancear a distribución dos recursos para cada camiño de dous saltos, correspondente a un usuario, de maneira independente. Como tódolos usuarios teñen un balance óptimo propio, pero se debe fixar un valor único para todos, estes experimentan un colo de botella importante.
- Ao introducir relays a interferencia incrementase de maneira esperada. Sen embargo, dado que os recursos non están correctamente balanceados, en algúns casos este incremento ocorre en van, sen ningún tipo de contrapartida positiva.
- A obrigatoriedade de que o relay se conecte en certos instantes concretos dana a flexibilidade requirida polos planificadores que explotan a diversidade multiusuario, que requiren a capacidade de pospoñer ou adiantar a activación dun enlace para atopar mellores realizacións da súa canle. Este fenómeno produce unha caida da ganancia na capa MAC que anula a ganancia de potencia introducida polos relays na capa física.
- Os algoritmos baseados en incentivos propostos na literatura para re-balancear estes planificadores só poden paliar as perdas, pero non eliminalas completamente. No peor dos casos con incentivos o planificador perde toda a diversidade multiusuario.
- A cobertura dun relay é moi pequena, polo que é necesario desenvolver regras de control de admisión que rexeiten a incorporación de relays que non producen beneficios.

A segunda parte da tese, centrada no futuro próximo, fai fincapé na escaseza de espectro nas bandas de microondas empregadas polas tecnoloxías actuais. Esta escaseza é un dos problemas máis transcendentes para este marco de tempo que pode ser paliado con cooperación e con radio cognitiva. Na parte teórica, o cuarto capítulo consiste nun estudo analítico para computar a ganancia social de espectro feita posible a través da combinación de técnicas cooperativas e de préstamo de espectro propias da radio cognitiva.

- Modelase a ganancia de espectro social como a cantidade de espectro que un primario pode deixar gracias á presenza de cooperación mentres que se siga acadando as mesmas taxas que o primario observaría nunha transmisión directa en todo o espectro dispoñible.
- Observase que a probabilidade de que exista algunha ganancia converxe a 1 cando o primario experimenta unha canle cunha distribución mala comparada coas canles do secundario, confirmando que o escenario onde estas técnicas son beneficiosas é aquel onde o propietario dun espectro o emprega de maneira ineficiente.
- As ganancias de espectro ergódicas acadables mediante unha toma de decisións estática a longo prazo experimenta un punto de inflexión: se o primeiro é bo, a ganancia é cero, pero se é malo, a ganancia rapidamente sobe a o 90% do espectro orixinal.
- A distribución das ganancias de espectro instantáneas acadables mediante unha toma de decisións dinámica a curto prazo evoluciona de maneira suave: para primarios malos a probabilidade probabilidade dunha grande ganancia de espectro é elevada, pero ademais para primarios que en media son bos, o sistema dinámico pode explotar os poucos momentos en que o primario sofre canles excepcionalmente malas para cooperar temporalmente e obter unha pequena ganancia adicional no espectro.

No aspecto práctico dedicado ao futuro próximo, o quinto capitulo da tese discute a limitada ganancia que o préstamo de espectro cooperativo pode ofrecer nos estándares actuais, debido ás limitacións antes mencionadas. Sen embargo, está previsto que emerxan novas aplicacións de datos con características distintas neste período, e algunhas destas aplicacións poden ser máis axeitadas para explotar as ganancias de espectro obtidas.

- Un só relay produce unha ganancia moi pequena, da orde dunha única parte por mil do espectro. Ademais a área de cobertura deste relay é moi pequena.

- Dado que na filosofía cognitiva os relays cooperativos son esencialmente gratis para a cela, e a área da cela se pode dividir entre un grande número destes, é posible acumular un grande número destes relays de modo que as súas ganancias se agregan. Aínda así, un número masivo de relays distribuídos intelixentemente pode acadar soamente ganancias da orde do 20% do espectro. E distribuídos aleatoriamente, a ganancia é aínda menor.

- Para realizar a cooperación, os dispositivos cognitivos requiren dispoñer de tecnoloxía LTE. Cabe supoñer que utilizarán parte desta mesma tecnoloxía para realizar as súas propias transmisións sobre o espectro gañado. Deste xeito, podemos estimar informalmente as taxas acadadas multiplicando a eficiencia espectral de LTE pola ganancia de espectro.

- O resultado de tal estimación é da orde de decenas de Mbps. Esta taxa non é suficiente para as comunicacións persoais do futuro. Sen embargo, é competitiva dabondo con outras tecnoloxías de comunicacións entre máquinas como por exemplo ZigBee (250Kbps). Este resultado confirma que os relays de LTE-A, aínda co apoio masivo obtido a través do préstamo masivo de espectro, estarán seguramente limitados a aplicacións de nichos menores.

A terceira e derradeira parte desta tese centrase no futuro distante. A carestía de espectro será paliada a través da introdución dun número masivo de graos de liberdade, feito posible polo traslado das comunicacións cara bandas máis altas de ondas milimétricas que permiten máis ancho de banda, máis antenas por área de silicio, e máis densidade espacial de nodos. Esta transición vai requirir un drástico re-deseño da enxeñería das redes celulares de comunicacións.

No aspecto teórico, esta tese presenta unha análise das leis de escalado que demostra que hai un límite para o incremento da capacidade dunha rede a través dun número masivo de recursos. Este límite depende da arquitectura das comunicacións empregada e pode ser aumentado substituíndo as comunicacións directas con comunicacións cooperativas en múltiples saltos.

O resultado de escalado obtense comparando a cantidade de recursos dispoñibles (area, ancho de banda, número de estacións base, número de antenas por estación base...) nunha rede sen fíos celular co número de nodos nesta. Expresando a capacidade como unha función crecente con número de nodos, o resultado obtén unha cota superior á capacidade e unha taxa acadable co protocolo multi-salto que escalan co mesmo expoñente. Ademáis obtiveronse outras taxas acadables derivadas doutros protocolos distintos, que non son quen de igualar a capacidade.

- A capacidade dunha rede celular que escala no número de nodos experimenta un "escalado crítico do ancho de banda" da mesma maneira que unha canle punto a punto experimenta unha transición dun primeiro réxime limitado en ancho de banda a un segundo réxime limitado en potencia cando o ancho de banda é grande. Cando a rede entra nun réxime de escalado da capacidade limitado en ancho de banda, acelerar o escalado deste aumenta a capacidade. Pero se o ancho de banda escala moi rápido, a rede entra nun réxime de escalado limitado en potencia e acelerar o escalado do ancho de banda non aumenta a capacidade.

- Protocolos diferences experimentan estas limitacións en limiares diferentes. O protocolo cooperativo baseado en múltiples saltos é o mellor, posto que garante que se acade o escalado óptimo da cota superior para calquera valor dos parámetros. Os límites de cada protocolo dependen da distancia de transmisión típica, e por tanto as transmisións multi-salto realizadas con relays dedicados que están máis separados que os usuarios teñen unha limitación teórica máis cativa que as comunicacións entre usuarios en termos de explotación do ancho de banda, pero aínda así máis grande que a limitación dun protocolo que simplemente envía transmisións directas de longa distancia cara cada un dos destinatarios.

- O modelo do tráfico ten unha extensa influencia nos resultados de escalado da capacidade. En redes ad-hoc con distribucións densas de usuarios (área constante co escalado do número de usuarios) hai técnicas de cooperación xerárquica moi elaboradas que son optimas por riba do simple protocolo multi-salto. Sen embargo, segundo os resultados obtidos para as redes celulares, un protocolo equivalente que combina cooperación xerárquica e uso de infraestrutura non é óptimo. Isto ocorre porque nas redes ad-hoc as transmisións directas poden dirixirse a calquera destinatario na rede, incluso a través da área da rede enteira, mentres que en redes celulares as transmisións directas se dirixen ao punto de acceso máis próximo, a unha moito menor distancia de transmisión. Deste xeito, a transmisión directa é ineficiente en redes ad-hoc densas con pouco ancho de banda, pero non en redes celulares cos mesmos recursos. Isto significa que moitas publicacións que analizan as redes con infraestrutura utilizando modelos de tráfico de ad-hoc apoiados pola infraestrutura non son suficientemente axeitados para modelar redes celulares.

O sétimo capítulo aborda a parte práctica da realización das comunicacións cooperativas con múltiples saltos nas futuras redes celulares de quinta xeración. Os temas tratados neste capítulo son os resultados máis recentes e algúns aínda non foron publicados en revista. Este capítulo escolma resultados obtidos en varios niveis de cara a deseñar e facer posible a construción de tecnoloxías de quinta xeración cooperativos.

Nestas redes o espectro non será tan escaso como ata o de agora, permitindo maior ancho de banda; os radios das celas serán máis pequenos, favorecendo a diversidade espacial; e o número de antenas por área de silicio incrementará ao reducirse as lonxitudes de onda.

- No nivel da teoría da información, a capacidade dunha canle cun grande ancho de banda precisa ser entendida mellor. Aínda que existe abundante literatura no tocante ás canles punto a punto no límite cando o ancho de banda tende a infinito, as ramificacións do problema cando se teñen en conta tamén outros factores non están tan claras. Nesta tese unificase a análise de dous tipos de sinais que ata o de agora se consideraban independentes: os sinais con picos, que en teoría acadan a capacidade cun ancho de banda infinito, e os sinais sen picos, que en teoría teñen un ancho de banda crítico finito e a súa taxa binaria descende se o ancho de banda é excesivo. Sorprendentemente, o resultado unificado é que a cantidade de picos do sinal non é relevante se se mide a cantidade física axeitada: a ocupación de ancho de banda do sinal, que coincide co ancho de banda nos sinais sen picos pero se reduce ao ancho de banda só nos picos cando os sinais os teñen, sempre experimenta un límite crítico finito. Ademáis introdúcese outra xeneralización do concepto de ancho de banda crítico para canles con múltiples usuarios, demostrando que ademais do coñecido fenómeno polo cal as transmisións superpostas son mellores en anchos de banda pequenos e as transmisións ortogonais son mellores para anchos de banda grandes, existe un réxime intermedio en que o ancho de banda é demasiado grande para a superposición pero demasiado pequeno para a ortogonalidade. Resulta sorprendente que a literatura de ondas milimétricas non teña prestado apenas atención a estes problemas no pasado.
- No nivel da capa física existen múltiples opcións para xerar sinalización en ondas milimétricas. Pódese diferenciar entre as estratexias que teñen que afrontar a limitación en ancho de banda mencionada anteriormente, e as estratexias que introducen recursos adicionais (tipicamente, un número masivo de antenas receptoras) para rodear a limitación e evitala saíndose por unha tanxente. Entre as primeiras, está a sinalización tradicional utilizada en LTE-A, que non ten picos e se debe limitar ao ancho de banda crítico; e mais as modulacións con picos tradicionais como FSK e a súa mellora moderna m-FSK, que poden utilizar todo o ancho de banda pero sempre deben satisfacer a restrición da ocupación de ancho de banda crítica. Na segunda clase de sinalización, esta tese analiza o caso en que un número suficientemente grande de antenas no receptor completamente a limitación do ancho de banda crítico e permite a capacidade na canle non-coherente crecer co ancho de banda sen un límite superior. Sen embargo, o número de antenas debe

crecer co cadrado do ancho de banda -$\Theta(B^2)$- e o coste dos dispositivos pode non merecer a pena a medio prazo.

- No nivel da capa MAC, a alta directividade das antenas en ondas milimétricas introduce un profundo cambio de paradigma. Dado que ocorre un illamento natural das interferencias, as estacións base e relays en celas distintas non teñen que estar a transmitir á vez de maneira sincronizada, e o dúplex pode ser dinamicamente escollido para cada nodo de maneira distinta, axustado para optimizar o rendemento. Ademáis, a posibilidade de comunicarse con múltiples puntos de acceso sen causar interferencia a outros permite a asociación simultánea do usuario con múltiples estacións base. A resolución dun problema de optimización para o planificador tendo en conta o caso xeral é relativamente complexa, polo que nesta tese se estudaron as ganancias de dous sub-problemas máis sinxelos. O primeiro combina o dúplex dinámico co procesado multi-usuario multi-antena nunha rede de topoloxía clásica en árbore. O segundo combina o dúplex dinámico coa multi-asociación que da lugar a unha nova topoloxía máis densamente conectada, pero que ao deixar de ser unha árbore estruturada non permite estudar facilmente o procesado multi-usuario multi-antena á vez. Demostramos que, en ambas estratexias, as taxas poden incrementar significativamente gracias ao dúplex dinámico en comparación co mecanismo estático empregado nas redes celulares tradicionais.

Publications

The following is a list of journal and conference publications that have been produced as a result of the work on this thesis.

Journal publications

1. [2] **Gomez-Cuba, F.**; Rangan, S.; Erkip, E. and Gonzalez-Castano, F.J., ”Capacity Scaling of Cellular Networks with Arbitrary Infrastructure Density and Bandwidth,” in preparation

2. [3] García-Rois, J.; **Gomez-Cuba, F.**; Gonzalez-Castano, F.J.; Burguillo-Rial, J.C:., Rangan, S. and Lorenzo, B. ”On the Analysis of Scheduling in Dynamic Duplex Multi-Hop mmWave Cellular Systems” IEEE Transactions on Wireless Communications (submitted)

3. [4] **Gomez-Cuba, F.**; Du, J.; Médard, M. and Erkip, E., ” “Unified Capacity Limit of Non-coherent Wideband Fading Channels” IEEE Transactions on Information Theory (submitted)

4. [5] **Gomez-Cuba, F.**; Asorey-Cacheda, R; and Gonzalez-Castano, F.J., ”Smart Grid Last-Mile Communications Model and Its Application to the Study of Leased Broadband Wired-Access,” , IEEE Transactions on Smart Grid, vol.4, no.1, pp.5,12, March 2013

5. [6] **Gomez-Cuba, F.**; Asorey-Cacheda, R.; Gonzalez-Castano, F.J. and Huang, H., ”Application of Cooperative Diversity to Cognitive Radio Leasing: Model and Analytical Characterization of Resource Gains,” IEEE Transactions on Wireless Communications, vol.12, no.1, pp.40,49, January 2013

6. [7] Sendín-Raña, P.; González-Castaño, F. J.; **Gómez-Cuba, F.**; Asorey-Cacheda, R. and Pousada-Carballo, J. M., ”Improving Management Performance of P2PSIP for Mobile Sensing in Wireless Overlays”. Sensors, 13(11), 15364-15384 2013.

7. [8] **Gomez-Cuba, F.**; Asorey-Cacheda, R. and Gonzalez-Castano, F.J., "A Survey on Cooperative Diversity for Wireless Networks," IEEE Communications Surveys & Tutorials , vol.14, no.3, pp.822,835, Third Quarter 2012

Conference publications

1. [9] Ford, R.; **Gomez-Cuba , F.**; Mezzavilla, M.; Rangan, S., "Dynamic Time-domain Duplexing for Self-backhauled Millimeter Wave Cellular Networks" to appear in IEEE International Converence on Communications (ICC), June 2015
2. [10] Chowdhury, M.; Manolakos, A.; **Gomez-Cuba, F.**; Erkip, E. and Goldsmith, A.J. "Capacity Scaling in Noncoherent Wideband Massive SIMO Systems" to appear in IEEE Information Theory Workshop (ITW), April 2015
3. [11] **Gomez-Cuba, F.**; Rangan, S. and Erkip, E., "Scaling laws for Infrastructure Single and multihop wireless networks in wideband regimes," IEEE International Symposium on Information Theory (ISIT), July 2014
4. [12] **Gomez-Cuba, F.** and Gonzalez-Castano, F.J., "Improving third-party relaying for LTE-A: A realistic simulation approach," IEEE International Conference on Communications (ICC), June 2014
5. [13] **Gomez-Cuba, F.**; Gonzalez-Castano, F. J. and Munoz-Castaner, Jorge, "Is Cooperative Spectrum Leasing by Third-Party Relays Advantageous in Next-Generation Cellular Networks?," 20th European Wireless Conference, May 2014
6. [14] **Gomez-Cuba, F.**; Gonzalez-Castano, F.J. and Perez-Garrido, C.P., "Practical Smart Grid traffic management in leased Internet access networks," IEEE International Energy Conference (ENERGYCON), May 2014
7. [15] **Gomez-Cuba, F.**; Asorey-Cacheda, R. and Gonzalez-Castano, F.J., "WiMAX for smart grid last-mile communications: TOS traffic mapping and performance assessment," 3rd IEEE PES Conference on Innovative Smart Grid Technologies Europe (ISGT Europe), Oct. 2012

Contents

List of Figures

List of Tables

Abbreviations

3GPP	**3rd G**eneration **P**artnership **P**roject
4G	**Fourth** Generation (mobile standard)
5G	**Fifth** Generation (mobile standard)
ABSF	**A**lmost **B**lank **S**ub**F**rames
ACK	**Ack**nowledgement
ADFM	**A**d-hoc-like **D**ata **F**low **M**odel
AF	**A**mplify and **F**orward
AMC	**A**daptive **M**odulation and **C**oding
AP	**A**ccess **P**oint
ARQ	**A**utomatic **R**epeat re**Q**uest
AWGN	**A**dditive **W**hite **G**aussian **N**oise
BER	**B**it **E**rror **R**ate
BC	**B**roadcast **C**hannel
BPSK	**B**inary **P**hase **S**hift **K**eying
BS	**B**ase **S**tation
CA	**C**arrier **A**ggregation
CBR	**C**onstant **B**it **R**ate
CC	**C**onvolutional **C**ode/ing
C.D.F. or *CDF*	**C**ummulative probability **D**ensity **F**unction
CDFM	**C**ellular-like **D**ata **F**low **M**odel
CDMA	**C**ode **D**ivision **M**ultiple **A**ccess
CFC	**C**oherent **F**ading **C**hannel
CFNC	**C**omplex **F**ield **N**etwork **C**oding
CQI	**C**hannel **Q**uality **I**ndicator
CR	**C**ognitive **R**adio

CSI	**C**hannel **S**tate **I**nformation
CSI-R	**C**hannel **S**tate **I**nformation available at the Receiver
CSI-T	**C**hannel **S**tate **I**nformation available at the Transmitter
CSL	**C**ooparetive **S**pectrum **L**easing
CTS	**C**lear / **C**onsent **T**o **S**end
D2D	**D**evice **to** **D**evice
DCF	**D**istributed **C**oordination **F**unction
DD	**D**ynamic **D**uplex
DDF	**D**ynamic **D**ecode and **F**orward
DeNB	**D**onnor **E**volved **N**ode **B**
DF	**D**ecode and **F**orward
DL	**D**own**L**ink
DSTBC	**D**istributed **S**pace **T**ime **B**lock **C**ode
EDDF	**E**xtended **D**ynamic **D**ecode and **F**orward
eNB	**E**volved **N**ode **B** (the name of base stations in LTE)
ESDF	**E**xtended **S**tatic **D**ecode and **F**orward
E-UTRAN	**e**volved **U**MTS **T**errestrial **R**adio **A**ccess **N**etwork
FCFS	**F**irst **C**ome **F**irst **S**erved
FDD	**F**requency **D**ivision **D**uplex
FDMA	**F**requency **D**ivision **M**edium **A**ccess
FEC	**F**orward **E**rror **C**orrection code
FSK	**F**requency **S**hift **K**eying
GFNC	**G**aulois **F**ield **N**etwork **Coding**
HAN	**H**ome **A**rea **N**etwork
HC	**H**ierarchical **C**ooperation
IHC	**I**nfrastructure **H**ierarchical **C**ooperation
i.i.d.	**I**ndependent and **I**dentically **D**istributed
IMH	**I**nfrastructure **M**ulti-**H**op
IMT	**I**nternational **M**obile **T**elecommunications
IMT-A	IMT-**A**dvanced
IoT	**I**nternet **of** **T**hings
IP	**I**nternet **P**rotocol
IRH	**I**nfrastructure **R**elay-multi-**H**op

ISD	**I**nter-**S**tation **D**istance
ISH	**I**nfrastructure **S**ingle-**H**op
ITU	**I**nternational **T**elecommunication **U**nion
LAN	**L**ocal **A**rea **N**etwork
LTE	**L**ong **T**erm **E**volution (3GPP standard release 8-9)
LTE-A	**L**ong **T**erm **E**volution **A**dvanced (3GPP standard release 10-11)
LTE-B	**L**ong **T**erm **E**volution Advanced **2nd** Phase (3GPP standard release 12-13)
LTE-Direct	**L**ong **T**erm **E**volution Advanced **D**irect Device to Device Communications
LTW-AF	**L**aneman, **T**se and **W**ornell's **A**mplify and **F**orward
M2M	**M**achine **To** **M**achine communciations
MAC	**M**edium **A**ccess **C**ontrol protocol or **M**ultiple **A**ccess **C**hannel
MBMS	**M**obile **B**roadcast **M**ulticast **S**ervice
MI	**M**utual **I**nformation
MIMO	**M**ultiple **I**nput **M**ultiple **O**utput
MISO	**M**ultiple **I**nput **S**ingle **O**utput
MRC	**M**aximal **R**atio **C**ombining
mmWave	**M**illi**m**etric **Wave**lenth band
MU-MIMO	**M**ulti-**U**ser **M**ultiple **I**nput **M**ultiple **O**utput
NACK	**N**egative **Ack**nowledgement
NAF	**N**on-orthogonal **A**mplify and **F**orward
NAN	**N**eighborhood **A**rea **N**etwork
NBK-AF	**N**abar, **B**ölcskei and **K**neubühler's policy-I **A**mplify and **F**orward
NC	**N**etwork **C**oding
NFC	**N**on-coherent **F**ading **C**hannel
ODE	**O**rdinary **D**ifferential **E**quation
OFDM	**O**rthogonal **F**requency **D**ivision **M**ultiplexing
OFDMA	**O**rthogonal **F**requency **D**ivision **M**ultiple **A**ccess
OR	**O**pportunistic **R**elaying
OSI	**O**pen **S**ystems **I**nterconnect
P2P	**P**eer to **P**eer
PaC	**P**ick **a**nd **C**ompare
PAM	**P**ulse **A**mplitude **M**odulation
p.d.f. or *pdf*	**p**robability **d**ensity **f**unction

PF	**P**roportional **F**air
PHY	**Phy**sical layer
PLC	**P**ower **L**ine **C**ommunicatons
PPM	**P**ulse **P**osition **M**odulation
PSD	**P**ower **S**pectral **D**ensity
RB	**R**esource **B**lock
RC	**R**elay **C**hannel
RN	**R**elay **N**ode
RR	**R**ound **R**obing
RRC	**R**adio **R**esource **C**ontrol
RRH	**R**emote **R**adio **H**ead
RDSTBC	**R**andomized **D**istributed **S**pace **T**ime **B**lock **C**ode
RTS	**R**equest **T**o **S**end
SDR	**S**oftware **D**efined **R**adio
SG	**S**mart **G**rid
SGLM	**S**mart **G**rid **L**ast **M**ile
SIMO	**S**ingle **I**nput **M**ultiple **O**utput
SIR	**S**ignal to **I**nterference **R**atio
SINR	**S**ignal to **I**nterference-plus-**N**oise **R**atio
SL	**S**pectrum **L**easing
SNR	**S**ignal to **N**oise **R**atio
STBC	**S**pace **T**ime **B**lock **C**ode
STC	**S**pace **T**ime **C**oding
TB	**T**ransport **B**lock
TDD	**T**ime **D**ivision **D**uplex
TDMA	**T**ime **D**ivision **M**edium **A**ccess
TTI	**T**ransmission **T**ime **I**nterval
UE	**U**ser **E**quipment (the name of user nodes in LTE-A)
UL	**U**p**L**ink
UMTS	**U**niversal **M**obile **T**elecommunications **S**ervice
VBR	**V**ariable **B**it **R**ate
VoIP	**V**oice **o**ver **IP**
μWave	**micro-Wave** band

Notation

Standard mathematical notation is followed throughout this thesis. The following non-exhaustive list defines the more uncommon notation symbols employed.

$\circledast$	Circular convolution operator
$\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$, $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$	Complex circular and Real Gaussian random distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$
$\mathrm{Exp}(\lambda)$	is the exponential distribution with mean $\frac{1}{\lambda}$
$\mathrm{E}_X\left[f(X)\right]$	Expectation of expression $f(X)$ over the random variable X
$P(x)$	Probability of event x
$\det(\mathbf{A})$, $\mathrm{tr}(\mathbf{A})$	Determinant and trace of $\mathbf{A}$
$\mathbf{A} \bullet \mathbf{B}$	Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbb{X}^{a \times b}$	Set of all matrices with a rows and b columns with entries in $\mathbb{X}$
$(\cdot)^T$, $(\cdot)^H$	Transpose and conjugate transpose
$\lvert \cdot \rvert_\ell$ (resp. $\lvert \cdot \rvert$)	norm ℓ (resp. norm 2)

Para mamá y papá

Chapter 1

Introduction

Contents

1.1 Motivation

According to the industry predictions [1], global mobile data traffic will increase nearly 11-fold between 2013 and 2018. By 2018 there will be nearly 1.4 mobile devices per capita, half of which will be “smart” devices, each generating almost 3 GB of data per month. Nonetheless, some industry representatives have suggested that the Shannon Limit to capacity has been reached, or that we are so close that any further improvement will be negligible.

> "The wireless industry has reached the theoretical limit of how fast networks can go."
>
> K. Fitcher, Connected Planet

> "We are 99%" of the way "to the barrier known as Shannon's limit."
>
> D. Warren, GSM Association Sr. Dir. of Tech.

But, as pointed out in [18], these statements come from the misunderstanding of taking the whole (the capacity of wireless networks) for the part (the Shannon limit of a point to point channel). The fact is that, when several directions of information exist, such as in a channel with feedback or an interference channel, not to speak of a whole network; we do not even know the limit of capacity, let alone how to achieve it.

This tells us that the increment in demand may be satisfied by wireless networks, but the problem must be attacked from other flanks instead of by merely improving point-to-point transmission technology. Increasing the number of transmission and reception antennas, the density of devices, or the bandwidth are reasonable approaches. In fact, they have been, by several orders of magnitude, greater enablers of rate increase, compared to point-to-point transmission/reception, over the last decade [19].

Another important side mechanism that can be exploited to circumvent the limitations of point to point channels is given by the existence of multiple communication streams occurring in the same network. This is, sometimes when one transmitter desires to deliver information to a receiver, and there are more nodes in the same radio environment, it is possible to make use of them to achieve a higher capacity. The simplest example of this is the Relay Channel (RC), in which an auxiliary repeater helps the transmitter reach the receiver.

This thesis explores "cooperative wireless communications" in a broad sense, as any form of wireless communication in a network where devices other than the source and the destination participate in their communication. Cooperation is not often implemented in today's networks, but this thesis shows that its use is imperative to make future networks capable of satisfying the growing demand for data.

1.2 Structure of this Thesis

There are three different stages in the incorporation of cooperation to wireless networks: the present, when few instances of cooperation are employed but the theory of cooperative communications is solid; the near future, when cooperation will have to be patched into networks already in existence rising compatibility issues; and the distant future, for which yet-to-be-standardized wireless technologies will be designed with cooperative support from the onset.

The contents of this thesis are separated in six chapters that can be viewed as two dimensional picture represented in Fig. 1.1. In the "time" axis, this thesis is divided in three parts that roughly correspond with the state of the art of analysis and implementation of cooperative communications, in the present; the prospective gains that can be expected through the a-posteriori introduction of cooperation into wireless technologies currently in use, in the near future; and the design of new wireless communication systems with a-priory support for cooperation, in the long-term future.

In the "theory-practice" axis, each part of this thesis consists in two chapters. The first chapter of each part contains the contributions of the thesis to the corpus of theoretical analysis of cooperative wireless communications. The second chapter of each part contains the corresponding contributions to the practical implementation of cooperative communications.

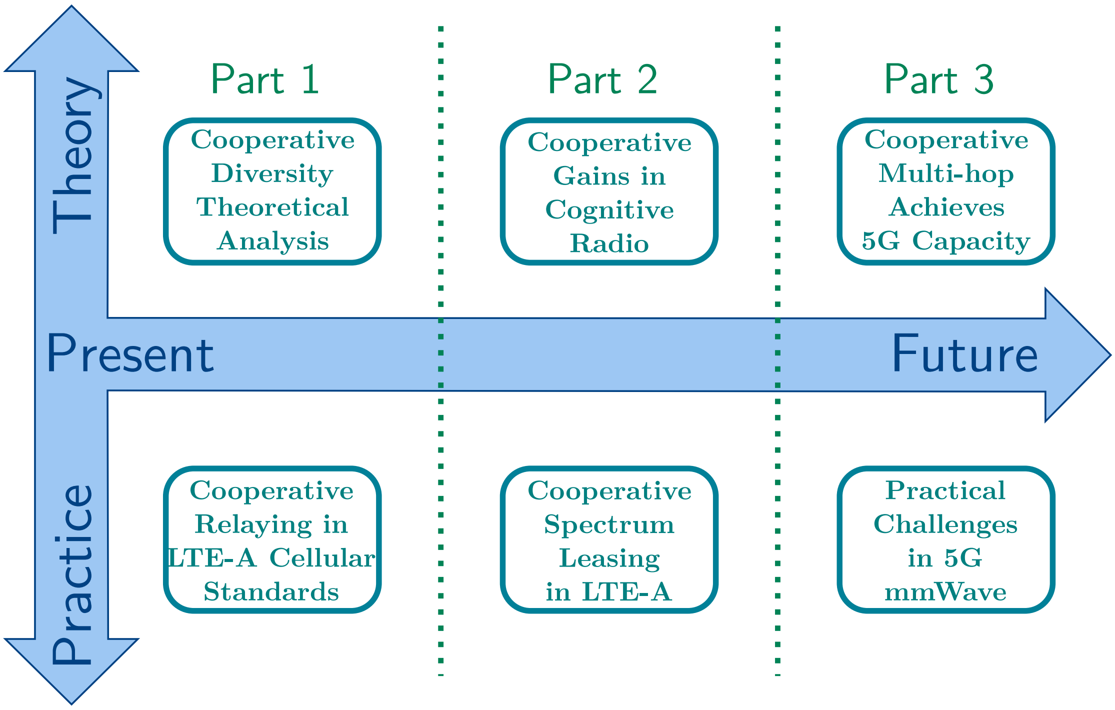

Theory
Practice
Present
Future
Part 1
Part 2
Part 3
Cooperative Diversity Theoretical Analysis
Cooperative Gains in Cognitive Radio
Cooperative Multi-hop Achieves 5G Capacity
Cooperative Relaying in LTE-A Cellular Standards
Cooperative Spectrum Leasing in LTE-A
Practical Challenges in 5G mmWave

FIGURE 1.1: Conceptual Map of the Thesis.

Chapters 2, 4 and 6 contain the theoretical contribution in each stage. These chapters discuss the theoretical modeling of cooperative transmissions in wireless networks and the information theoretical analysis of their capacity, showing that cooperation can introduce gains over the corresponding non-cooperative equivalents.

Chapters 3, 5 and 7 describe the practical contributions in each stage. Without loss of generality, all practical examples in this thesis focus on the application of cooperation to wireless cellular data networks. The study in this type of network is of great importance because of the exponential increase in demand for mobile data services. Moreover, the outcome of the discussion of practical issues in wireless cellular data networks can be easily translated to other types of wireless networks –such as sensors networks, ad-hoc networks, emergency networks...– since the former are amongst the largest and most complex wireless systems in existence.

1.3 Cooperative Wireless Communications

1.3.1 Description

The point-to-point channel is the simplest communication topology. Its model consists in only one transmitter and one receiver existing in an otherwise empty space. Even with such a simple topology, communications depend on the properties of the environment and the problem is not always trivial. The capacity of a communications system composed of more than those two essential devices is even harder to determine, due to their interactions, and it is impossible to calculate the capacity region as the mere union of independent point-to-point scalar problems.

A naive first approach when two communications occur in the same space is to consider the presence of the first as an obstacle for the second. A clever engineer would realize that two communications can have needs in common, opening the possibility to share resources. An example of this change in mentality is the two-way relay channel [20], where designers realized that in two symmetric transmissions between nodes mutually connected through a relay, one should not require twice the time to perform all connections compared to a single-sided transmission, but that in fact it is possible to implement a reciprocal link without any additional transmission slots.

The broad term “cooperative wireless communications” does not refer to a specific, isolated field of knowledge. It is more like a transversal concept that covers any type of communication problem in a wireless network where one transmitter desires to deliver information to a receiver, and where one or several cooperative nodes nearby help in the communication, so that the transmitter can achieve a different (possibly higher or more stable) rate than with a point-to-point transmission. A few examples of a cooperative wireless communications in different fields are:

- The Relay Channel (RC) in information theory.

- Multi-hop routing.
- Cooperative virtual antenna arrays.
- Network Coding (NC).
- Interference Coordination.
- Multi-user diversity in scheduling.

The different models of cooperative communications are not independent. On the contrary, they tend to be highly overlapped. For example, a multi-hop transmission can only achieve the capacity of one or multiple concatenated RC at best. Or, in a virtual antenna array (a cluster of single-antenna devices cooperating to collectively implement multiple-antenna techniques), the Multiple-Input-Multiple-Output (MIMO) technique of "selection combining" (of all pairs of antennas, choose the best one at each time) boils down to a multi-user diversity scheduling model. More generally, the study of cooperation in wireless communications is a foggy area that often admits the representation of the same reality with wildly different mental pictures, an issue that makes it difficult to navigate through the literature and condense results from different lines of research. Chapter 2 surveys and discusses results on the theoretical framework for the measurement and achievement of *cooperative diversity gains* in wireless networks.

In practice, all past cellular standards specified only one-hop direct communications from the Evolved Node B (eNB) to the User Equipment (UE) and vice-versa. That is, the UEs in a cell attached solely to a single eNB and exchanging transmissions only with it. Moreover, in current systems the downlink (DL) and uplink (UL) transmissions are synchronized across all cells using either Frequency Division Duplex (FDD) or Time Division Duplex (TDD), so that at a given time and sub-band there are only two possible states: either all eNBs in the system are transmitting and all UEs receiving, or vice versa. The Third Generation Partnership Project (3GPP) fourth generation (4G) cellular standard, Long Term Evolution (LTE), introduced Relay Nodes (RNs) in its Advanced version (LTE-A) approved in 2012 (Release 10), allowing in-band two-hop communications for the first time in a cellular system. Chapter 3 studies the current practical feasibility of these concepts through the implementation of a system level simulation.

1.3.2 Prior Work

1.3.2.1 Cooperative Diversity

Chapter 2 surveys theoretical works on cooperative wireless communications that are relevant to this thesis. For further documentation, the reader may refer to the recent books on the subject [21–23].

The term diversity gain comes from the analysis of wireless transmissions with a random (slow) fading channel, in which capacity is expressed in terms of outage probability for a given power and transmission rate [24]. Contrary to the rapid vanishing of Bit Error Rate (BER) with power in Additive White Gaussian Noise (AWGN) channels, in a point-to-point fading channel outage probability decreases only linearly with increasing Signal to Noise Ratio (SNR). We say that this represents a diversity index of $d = 1$ because this is the exponent of the inverse of a linear function SNR^{-1}. Through the introduction of redundant transmission devices (redundant carrier frequencies, repetition coding longer than the fading duration, or multiple antennas), it is possible to increase the exponent of outage probability as a function of SNR, SNR^{-d}, $d > 1$. Thus, for example, for a 2-antenna transmission with appropriate encoding that makes outage probability vanish quadratically with power (SNR^{-2}), we say that the diversity index is $d = 2$. The extension of the concept to a cooperative network is straightforward, in cases when there is cooperation in a network and data can be delivered by N relays, N multi-hop routes, a virtual array of N transmitters, an N-flow network coding scheme, etc. Then, if we can find an encoding scheme such that end-to-end outage probability vanishes as a function of power with exponent SNR^{-N}, we say the cooperative scheme achieves the *full cooperative diversity* $d = N$.

Even though the measurement of diversity in fading channels gained relevance with the recent popularization of MIMO, the use of multiple paths to obtain capacity gains can be traced much further back, to the first analysis of the RC by Cover and el Gamal [25] in 1979. These first approaches to cooperation, however, lack any explicit mention to cooperative communications, which is understandable: on the one hand, the term cooperative diversity was not popularized until Nicholas Laneman's thesis in 2002 [26]. On the other hand, people intuitively used relaying to extend range since the early days of wireless communications, and the early work had enough ground on the fact that these systems were already in use. Asking these early authors to foresee that their result would be extrapolated to a communications paradigm decades latter might have been a little excessive.

1.3.2.2 LTE-A Relaying

The performance of LTE relays has been assessed by several authors [27–31]. A major drawback of that work is that relay channels are modeled following theoretical characterizations rather than based on implementations of the standard. Even though these authors employ the correct propagation models by 3GPP to compute the Signal to Interference plus Noise Ratio (SINR) of all links, and LTE-fitted Shannon-like mappings from SINR to rate [32] to compute their spectral efficiency, they lack an adequate treatment of the connection between spectral efficiency and rate. This depends on the number of spectrum resources allocated by the network to the relay links, and the aforementioned works assume that they are optimally balanced. Instead, following the allocation possibilities tabulated in the standard, we determine the end-user rates delivered much more accurately (and pessimistically), taking into account bottlenecks and imperfect allocation policies.

On the other hand, so far, LTE system-level simulation has been performed with several tools, including the TU Vienna LTE System Level Simulator [33], and modules for general purpose simulators such as ns-3 [34], but they have focused on one-hop cellular architectures and do not support relaying.

1.3.3 Contributions

Chapter 2 of this thesis is a survey and its contributions lie in the classification of literature rather than the creation of new models. Chapter 3 is an evaluation of LTE relaying and its contributions lie in the identification of practical problems in the standard that prevent it from achieving the benefits predicted by theory.

C1.1.1) **Compendium of Information-Theoretic Cooperative Diversity Models** There are many information theoretic models for cooperative diversity: in the capacity analysis of the RC, which takes the supremum over all types of relays, achievable rates and diversity order are not the same as in the analysis of specific types of relays such as Decode-and-Forward (DF) or Amplify-and-Forward (AF). Moreover, specific new types of relay operation are often designed with the sole purpose of improving diversity. Section 2.2 discusses the literature on this topic, the main areas of interest and methods to improve performance, and provides examples of theoretical relaying models.

C1.1.2) **Compendium of Physical (PHY) Layer Signal Processing Methods for Cooperative Diversity Exploitation** The PHY transmitter is responsible for encoding information in codewords and synthesizing them in electromagnetic symbols, and the receiver is responsible for detecting the transmitted symbol and figuring out which codeword was transmitted. In order to combine symbols from several channels in the receiver, some authors in cooperative diversity PHY literature have suggested to proceed at different levels: at the control plane, by selecting a best relay for each transmitter; at the channel coding plane, by distributing different bits of a Forward Error Correction (FEC) code among multiple transmitters; at the source coding plane, by mixing signals from different sources using NC; and at the physical plane, by sharing MIMO codes among devices forming a virtual antenna array. Section 2.3 discusses the literature on this topic, the properties of combination in each plane, and provides examples of these techniques.

C1.1.3) **Functional Decomposition of Medium Access Control (MAC) Protocols for Cooperative Diversity Exploitation** The MAC layer is responsible for arbitrating the access to the radio-electric medium between transmitter/receiver peers. Networking tasks such as discovering neighbors and multi-hop routes to deliver information to the final destination are usually included in wireless MAC designs as well. MAC protocols in the literature are often bottom-up designs, where research begins by defining the messages and procedures a protocol implements, followed by deriving the performance of that specification. This is undesirable for two reasons: first, there are too many protocols to choose from in order to introduce cooperation in a network. And, second, bottom-up thinking starts with the components and from them it derives system properties, but wireless standards should be designed top-down instead, starting with the desired system properties and determining the necessary building blocks. Section 2.4 constructs a *functional decomposition* of the cooperative MAC, defining a set of five abstract operations that any cooperative MAC needs to perform (neighborhood mapping, helper set design, cooperation analysis and decision, cooperator notification and agreement, and cooperative transmission design). These five abstract operations determine cooperation completely in a MAC protocol, and all the protocols in the literature implement them in different manners.

C1.1.4) **Compendium of MAC Protocols and Classification According to C**1.1.3 Section 2.4.1 lists the cooperative MAC algorithms found in the literature, and discusses their properties of interest and major contributions. Section 2.4.2 shows that the functional description (C1.1.3) allows a five-dimensional classification of all MAC protocols according to their implementation of each function.

C1.2.5) **Implementation of System-Level LTE-A Relay Simulation** A relaying extension to the Vienna LTE System Level Simulator was designed, implemented and tested. The details of the implementation are described in Appendix 3.A. A patch to modify the official simulator with this extension has been published under an open research license in the web `http://enigma.det.uvigo.es/~fgomez/`.

C1.2.6) **Design of a RN Interference Mitigation Technique in LTE-A** In the LTE-A standard all RN are at the same time either connected to the Base Station (BS) (relay phase) or to the users (access phase), but not to nodes of both types. The time-sharing factor between the two phases typically dedicates fewer resources to the relay phase. Therefore, quite often, RNs receive less data from the BS that what they could deliver to the users in the access phase. There is an excess of transmission resources dedicated to the access phase where RNs are transmitting according to the standard but they have no data to deliver. Therefore, network interference grows in vain by these transmissions. Sec. 3.4 studies the problem and provides a simple interference mitigation technique by adding a third phase where RNs remain silent when they have no more data to transmit, thus reducing interference. This mechanism is similar to Almost Blank SubFrames (ABSF) [35] used by LTE-A femtocells, but its necessity in RNs has not been reported so far.

C1.2.7) **Analysis and Failure Detection in LTE-A Proportional Fair Scheduling With RNs** The standard leaves open to the implementation the rules for eNBs to schedule users in time and frequency. A well known scheduler is the Proportional Fair (PF) algorithm, which guarantees eventual access to all users but, instead of using a sequential access order like Round Robin (RR), modifies the order of service so that users are swapped to better instantaneous realizations of their channel realizations (multi-user diversity). In our simulations, RNs produce substantially more rate gains under RR than PF schedulers, whereas the later are more desirable due to their higher baseline rates. Section 3.5 performs a differential equation analysis of PF schedulers modified to add a restriction that forces relays to be served in fixed frames. This analytical result proves that the flexibility of PF to reorder users is hampered by such static constraints, reducing multi-user diversity gains.

C1.2.8) **Discovery of New Tradeoffs in the Design of LTE-A RN Access-Time Balancing** Building on the analytical results, Section 3.5.3 analyzes the conflict between PF and RNs, shows that the typical configuration parameters in the standard are not good enough, predicts cases of poor performance, and suggests new configuration parameters that might mitigate those issues.

C1.2.9) **RN Admission Control Algorithm for LTE-A** Since RNs offer little gain, and sometimes even losses if they introduce more interference than benefits, it is desirable to modify RN attachment procedures to reject an attempt of joining the network from a RN if it is incapable of producing beneftis. Section 3.6 proposes a simple admission control algorithm for its use with the simulator.

1.4 Cognitive Radio and Cooperation

1.4.1 Description

In wireless communications the radio-electric medium is shared by all devices. The limited set of viable frequency bands for transmissions is a canonical example of a resource subject to the crisis of the commons. Traditionally, government agencies are in charge of spectrum management, typically issuing long term licenses to one single application and operator, except for a few official and amateur bands. This is the paradigm on which all current wireless networks have been built, from terrestrial television to cellular telephony. However, many applications do not use their licensed bands continuously or do not use them in all locations with the same intensity. In recent years, spectrum scarcity has sparkled interest in creating short-term dynamic management policies for reusing semi-unused bands, enabled by new radio technologies that rely on computer intelligence and active awareness of their surroundings to adapt frequency usage depending on the context. This new type of dynamic, intelligent and context-aware allocation of spectrum is called **Cognitive Radio** (CR) [36].

Current wireless systems are usually implemented in microwave (μWave) frequencies, which are very scarce. Therefore, the hopes for capacity improvement in the short term while using the same wireless technologies lie heavily in the improvement of spectrum management. Theoretically, the following three paradigms are distinguished:

- *Interweave CR*: An opportunist secondary device detects absence of transmissions by a primary and transmits in the (temporarily) empty bands.
- *Underlay CR*: A secondary transmits on top of the primary using special signals so that its interference is harmless.
- *Overlay CR*: A secondary transmits on top of the primary causing interference, but it also transmits a cooperative aid to the primary signal that repairs the damage.

However, the implementations of CR are not that clearly different. First, there are diverse implementations of interweave CR. In the *White Spaces* model, there is no need for permission from the primary and it is sufficient that the channel is empty to allow the secondary to transmit. On the contrary, in the *Spectrum Leasing* (SL) model, the secondary needs to detect an empty band and pay to obtain a temporary lease from the primary. Finally, the frontiers between the concepts in [36] fall apart when, instead of an economic retribution, the payment in SL is a cooperative retribution. This originates the technique of Cooperative Spectrum Leasing (CSL), which is both an interweave SL and a channel with cooperation similar to overlay CR.

The second part of this thesis studies the improvement of current wireless technologies through the simultaneous adoption of CR and cooperative communications. Chapter 4 studies the conversion of cooperative diversity gain to spectrum gain under CSL. Our theoretical analysis covers a novel cooperative diversity point of view with applications in cognitive radio. The contributions in this chapter compared to previous cooperative diversity analysis are twofold: First, it represents a new approach to cooperation incentives in the context of CR-type channel access conditions. And, secondly, the analytical result is a characterization of spectrum gains resulting from the application of cooperative diversity to spectrum leasing under time-varying channel conditions and for both instantaneous and ergodic mutual information (MI) decision-making strategies.

Chapter 5 discusses future evolutions and emerging needs during the life-cycle of 4G, which is expected to last at least until 2020. In this evolution a key aspect is the necessity to maintain retro-compatibility using the same scarce μWave spectrum from 800MHz to 2.4GHz. The analysis calculates potential improvements in spectrum usage by implementing CSL using LTE-A RNs. Although Chapter 3 establishes that the current implementation of relaying in LTE has several handicaps, future versions of the standard could potentially correct these problems, so this part of the thesis adopts theoretical RN models in that direction. In addition, this chapter also discusses the impact of emerging new applications, as the markets for cellular wireless networks keeps expanding.

The arrival of new applications not currently served by LTE will bring new scenarios that may foster the exploitation of RN gains. One of such applications are Peer-to-Peer (P2P) communications, that are already being introduced into the Second-Phase-LTE-A versions of the standard (LTE-B) by incorporating the Device-to-Device (D2D) mechanism of LTE-Direct. A second application that can dramatically change the usage of cellular networks is Machine-to-Machine (M2M) communication, particularly the Smart Grid (SG), shifting user profiles heavily: from thousands of users demanding several millions of bits per second with high mobility and variability, to millions of users demanding only a few thousand bits per second consistently and almost statically.

1.4.2 Prior Work

1.4.2.1 Cooperative Spectrum Leasing

Goldsmith et al. reviewed cognitive radio in [36], describing diverse information theoretical solutions to the problem of a secondary transmitter gaining access to the spectrum without affecting the spectrum rights of the owner, and identified different forms of cooperation between cognitive and non-cognitive users. According to their classification [36], the model in Chapter 4 belongs to category "overlay CR" and subtype "aware non-cognitive users".

Although the concept of *spectrum leasing* is known in cognitive radio, the analysis in this thesis follows an approach focused on computing the cooperative diversity social gains (that is, system-wise) in the CSL system through information theory, rather than evaluating CR negotiation between individuals to compute greedy gains. Cooperative leasing schemes were studied in [37] and [38], which proposed negotiating resource gains leading to performance improvement as an alternative to money transactions for spectrum leasing. However, in the model described in [37], each node makes greedy decisions on its own transmissions, which means that the secondary transmitter attempts to improve its information flow by reducing the power it allocates to collaboration unless the primary responds by pulling the cooperation away. This interaction is game theoretic and has a Nash equilibrium that is socially desirable. as the authors show that it is a global maximum.

The model presented here, in contrast, assumes that the primary transmitter, with full spectrum rights, directly chooses the proportion of resources to be allocated to each information flow (see Fig. 4.1), thereby controlling the fraction of resources that the secondary transmitter receives for transmission (Fig. 4.2). This is similar to traditional white-space-oriented cognitive radio in which secondary transmitters are only allowed to use free primary spectrum, but allowing the secondary to actively take actions to increase the amount of resources freed by the primary.

The analysis computes gains on the primary link by maximizing the average MI in the static case (statistical Channel State Information CSI) or its distribution in the dynamic case (full CSI). Other alternatives are possible. For example, [39] designed a spectrum leasing system that simply guarantees that the outage probability of the primary transmitter will not increase. And the fractional cooperation scheme in [40], even if it may seem similar, has significant differences. In fractional cooperation, the secondary transmitter only relays the minimum information necessary to achieve full diversity [41], while the remaining resources are released for secondary transmission.

In other words, fractional cooperation obtains resources by keeping cooperation to a minimum and using the rest of resources for secondary purposes, whereas the model in Chapter 4 uses cooperative transmission techniques to their full possibilities and, afterwards, extracts spectrum gains from the increased rate of the primary transmitter.

Regarding practical implementations of leasing mechanisms, [42] presents a cognitive spectrum leasing system with cooperation from the perspective of packet queuing and retransmission. In this system, the primary transmitter accepts packet acknowledgements (ACKs) from either the destination or the secondary transmitters, which can deliver primary packets. A throughput analysis of the different mechanisms is proposed. This implementation can be considered for any of the aforementioned models.

Finally, previous characterizations of the gain of DF relaying can be found in [43]. They characterize the probability that a two-hop DF relay channel requires a lower SNR than that required by a direct transmission. Chapter 4 starts from a similar formulation as a first step, but extends it several more steps by quantifying the proportion of resource gains. Also, the DF capacity model considered there derives from the RC as in [41], allowing a signal from the source to reach the receiver too; whereas [43] uses the more specific subclass of fully orthogonal DF with only S→R and R→D transmissions.

1.4.2.2 LTE RN Rate Gains

To model the possibility that practical implementation errors in LTE-A Release 10 RN will be solved shortly, this part of the thesis follows similar theoretical analysis methods as [27–31]. Spectral efficiency in a link is modeled, given its SINR, using the approximation in [32]. To compute the SINR values the parameters in [29] are considered. This work provides three path loss models and two scenarios are discussed: urban (Inter Station Distance [ISD]= 500 m) and suburban (ISD = 1732 m). The first model was chosen, because it is the most conservative and does not depend on ISD.

1.4.2.3 New Data Applications

The advent of each new cellular wireless communication generation brings exponential increments in the volume of the traffic to be carried, the number of devices and their density, and the heterogeneity of the applications they support. New wireless communications are embracing the Internet Protocol (IP), heading towards an all-IP design where any application may be carried by generic data-transport wireless services, whereas the traditional voice service is taken to an “over the top” carrier such as voice over IP (VoIP). In addition, the centralized architecture of traditional services is not suitable for new types of traffic such as P2P applications; a trend already acknowledged by LTE-B, in the form of direct D2D communications. In [7] the properties of P2P voice systems are described. This work also studies problems related with the implementation of a traditional cellular application (voice calls) through VoIP P2P methods.

New communication services not oriented to inter-personal exchanges is also a driver force of change that will fundamentally alter the needs of future networks. M2M communications within the *Internet of Things* (IoT) [44] demand wireless communications for billions of new devices with different needs than human users. Previous works such as [45] show the diversity of traffic and mobility patterns of human and M2M communications. For example, most M2M devices are less mobile than smartphones, and sometimes even completely static.

M2M and IoT integration in cellular networks has been a hot topic in the last years [45–47]. As previously said, Shafiq et al have studied the different traffic patterns of M2M and human communications, by analyzing traffic data from a tier-1 cellular network in the United States. Their results indicate that M2M devices are less mobile than smartphones and compete with them for network resources in co-located geographical regions. According to other works M2M applications may degrade human communications [46, 47]. Due to its colossal size, the particular type of M2M communications of SG must receive special attention in the development of future standards. The SG M2M network will be the result of the upcoming renovation of the electric grid to introduce state-of-the-art communications, computing, management and control technologies and bring energy services into the digital era. Utilities have raised many expectations in automation, integration of future energy sources and rapid-response automation mechanisms, while customers demand rich domestic applications for smart home management, satisfaction of their ecological concerns, and energy cost savings [48, 49].

SG communications requirements are studied in [5, 14, 15]. They are challenging because they differ from those in traditional data networks. The requirements include very tight and strict latency limits, massive numbers of network nodes, and high levels of integrity protection to ensure service of critical infrastructures. For this colossal digital upgrade effort to succeed, the reduction of spectrum costs and the improvement of communication reliability are of the utmost importance.

1.4.3 Contributions

C1.3.1) **Proposition of a Metric for CSL Social Utility** The analyses of CSL in the literature focus on the equilibrium of game-theoretic models. This provides information about how gains are shared once CSL is installed in networks, but it does not help to decide whether CSL should be implemented or not. On the contrary, even though the model in 4.2 is simpler, it is valuable because it focuses on the social gain by measuring the total spectrum gains.

C1.3.2) **Probability of CSL Gains** If cooperation can provide gains but only on rare occasions, the incentive to modify the hardware of many devices just in case one of them needs cooperation would be low. It is important to determine whether cooperation is beneficial frequently or just in rare situations. Section 4.3 analyzes the probability that there can be a rate increase through cooperation in a system where CSL is available.

C1.3.3) **Probability Density Function of Mutual Informations in CSL** The p.d.f. of mutual information characterizes achievable rates under direct and cooperative schemes. Sec.4.4 obtains the p.d.f.s and discusses their behavior with different system parameters, to motivate the observations in the model.

C1.3.4) **Analysis of CSL Gains With Long-term Static Decisions** In some cases the decision of cooperation can only be made once for many realizations of the fading channels. This is the case, for instance, if channels vary much faster than the response time of the decision scheme, or if instantaneous CSI is not available due to hardware constraints. Section 4.5 analyzes the ergodic spectrum gains of a CSL that chooses between cooperation and direct transmission once for many channel realizations.

C1.3.5) **Analysis of CSL Gains With Dynamic Decisions** If channels vary slower than the response time of the decision scheme and the hardware to obtain instantaneous CSI is affordable, the decision of cooperation can be made separately for every fading channel realization. Section 4.6 analyzes the p.d.f. of the instantaneous spectrum gain of a CSL that chooses between cooperation and direct transmission every time the channel changes.

C1.4.6) **Model and Gains of a Single CSL LTE-A RN** Section 5.2 proposes the adaptations that LTE-A RN models need to measure the spectrum gains in a CSL context, and Section 5.3 evaluates the gains of a single RN.

C1.4.7) **Gains of a Massive RN Population** The gain obtained in C1.4.1 is quite small, but it was also found that the size of the area served by a RN, compared to that of the cell, is negligible. Since CSL RNs are available for free, it is possible that a cell could admit many of these and aggregate their puny gains until the total is significant. Section 5.4 studies the gains in such a system for both regularly and randomly located RNs.

C1.4.8) **Rates in Application Examples** Since CSL devices will implement LTE hardware, it would be wasteful to install additional hardware in them and implement the cognitive application with a different transmission technology. This allows to estimate the rate of an application based on LTE CSL, by computing the spectrum gains of the cognitive RN and multiplying it by the spectral efficiency of the LTE physical layer. Section 5.5 provides two examples for two applications: short-range D2D communications for M2M or P2P communications and sensor data gathering in SG, with a data collector at the eNB.

1.5 Cooperation in Future Wireless Networks

1.5.1 Description

The first two parts of this thesis discuss implementations of cooperation a-posteriori, as introduced on standards that were not cooperative in the first place. This makes cooperative mechanisms subject to retro-compatibility constraints that limit performance, as demonstrated in Chapter 3. Nevertheless, in the long-term, the evolution of wireless communications will be defined by new architectures and standards, which, once engineered, may follow different design principles than their predecessors. It is reasonable to consider that future standards such as the fifth generation of mobile communications (5G) will come out with embedded support for cooperative communications. Consequently, in the third and last part of this thesis, instead of introducing cooperation in pre-existing technologies, we argue that in future standards it will be necessary to introduce multi-hop cooperation into the design philosophy of large wireless networks for them to achieve the highest throughput capacity.

The research community agrees on the need for far-reaching changes in cellular network architecture for the next generation of standards [50]. Rather than simply further developing μWave communications, an extension of cellular systems to the underutilized millimeter wave (mmWave) band emerges as a natural choice for next-generation 5G networks. This band was not included in previous cellular generations due to its significant path loss and expensive hardware, but recent innovations in multi-antenna processing and the decreasing manufacturing costs have led it to be reconsidered for 5G [51].

To meet the tremendous growth in demand for cellular wireless data, three new design approaches are widely-considered for the evolution of next-generation systems [50]:

- The use of very high frequencies, and specially mmWave bands above 6 GHz, where vast quantities of new spectrum are readily available [51–55]
- Massive MIMO systems [56, 57], where hundreds of antennas at the base station can be leveraged for high levels of spatial multiplexing
- Ultra dense deployments of small pico- and femtocells [58, 59] to increase the capacity per unit area.

Together, these technologies offer the potential of increases in capacity by orders of magnitude, and, if successful, may change the basic constraints that dictate network design today. This possibility leads to two basic questions: what is the fundamental capacity offered by these technologies and how can networks be designed to fully leverage their potential.

Chapter 6 provides a proof based on information theoretic scaling laws showing that cooperation is a necessity in next generation cellular networks to achieve the best capacity scaling. Chapter 7 contains the most recent works in this thesis in that regard, which are still in preparation for publication. These ongoing works develop more detailed analyses to characterize the necessities of the implementation of 5G beyond the realm of scaling laws. Chapter 7 discusses the research framework that will shape the design of the future 5G cellular standards. These standards should roll out after 2020, which means that the theoretical research backing them ought to be already in development.

1.5.2 Prior Work

1.5.2.1 Scaling Laws

The seminal work by Gupta and Kumar [60] showed that the feasible rate in a dense *ad-hoc* network scales as $R(n) \propto \Theta(\frac{1}{\sqrt{n}})$[1], where n is the number of nodes. Hence, the capacity per node decreases with network size. Ozgur, Lévêque and Tse introduced *hierarchical cooperation* (HC) [61], also for dense ad-hoc networks, achieving linear scaling (i.e. $R(n) = \Theta(1)$). Franceschetti, Migliore and Minero introduced an electromagnetic physical constraint [62]: previous results assumed that nodes were sufficiently separated to experience far-field propagation, but in reality increasing n in a constant area eventually exhausts the degrees of freedom of the electromagnetic field contained in a finite region, posing an ultimate physical limitation of $R(n) \leq \Theta(\frac{\log(n)^2}{\sqrt{n}})$. Ozgur, Johary, Tse and Lévêque argued that both results are compatible [63]: linear scaling is achievable

[1] The notation $f(n) \propto \Theta(g(n))$ means that there exist two positive real constants $c, c' \in \mathbb{R}^+$ and an index $n_0 \in \mathbb{N}$ such that $f(n)$ is lower and upper bounded by a scaled version of $g(n)$, $cg(n) < f(n) \leq c'g(n)\ \forall n > n_0$.

in a transitory regime with high, but finite, values of n, but ceases to hold when n is so high that the physical limit [62] is reached. In [63] they also replaced the traditional separate analysis of dense and extense networks with a generalized analysis of *operating regimes*, defining an arbitrary user density value and determining the threshold of that value for which the operating regime changes from dense-like to extense-like networks. More recently, the practical interest of hierarchical cooperation to achieve linear scaling was put into question in [64] by finding that the optimal number of layers in the cooperation hierarchy is small under practical limitations, well below the large number required to achieve linear scaling. There have been extensions of scaling laws of ad-hoc networks introducing cooperation, mobility, broadcast, infrastructure or wideband. See [65] for a comprehensive review.

Most literature on scaling laws follows ad-hoc network models, which are not adequate representations of a cellular network. Even though works like [66–68] have modeled *ad-hoc networks with infrastructure support*, in their analyses the networks are really ad-hoc ones and infrastructure is only employed as an intermediary for ad-hoc type communications where the data flows routed through an infrastructure are the same node-to-node flows of the ad-hoc network. The model in Chapter 6 follows a different approach taking into account that cellular networks do not require to transport data in the same manner as ad-hoc ones. Instead, each node sustains UL and DL data flows with the closest BS. We term the typical approach in previous literature the Ad-hoc-like Data Flow Model (ADFM) and the model presented here is called Cellular-like Data Flow Model (CDFM). The models have in common the presence of infrastructure with arbitrary scaling density, but they differ in the organization of the traffic. A key issue with ADFM is that it allows to fall back to pure ad-hoc protocols ignoring infrastructure when this is convenient, whereas CDFM always requires traffic to reach the infrastructure even if this creates bottlenecks limiting scaling [69]. Therefore ADFM is less accurate to analyze the scaling of cellular networks. This confusion between the concepts of “cellular networks” and “networks with infrastructure” means that many results that have used ADFM [67, 68, 70–72] may not necessarily have direct impact in the design of cellular systems.

1.5.2.2 Wideband Analysis

The main innovation of the scaling analysis method in Chapter 6 is evaluating the impact of very large bandwidths in capacity scaling. Most scaling analyses consider a constant finite bandwidth (B). However, in such setup links only become power limited with propagation distance, not with bandwidth, since a finite bandwidth may be sliced infinitesimally thin as the number of nodes grows. Another approach consists on letting $B \to \infty$ a priori for each constant value of n, and *then* let n grow in the resultant wideband-forced expression, as in [73]. Nevertheles, this makes the network to be always power-limited and does not provide insights on the interaction between bandwidth causes of power-limitation, network architecture and size. In Chapter 6 the goal is to find out what happens between these two extremes by letting B and n increase to infinity at the same time, following an arbitrary exponential relation with n:

$$\psi := \lim_{n,B\to\infty} \frac{\log B}{\log n}, \tag{1.1}$$

where the two cases in the literature correspond to $\psi = 0$ and $\psi = \infty$. By introducing this new parameter, bandwidth scaling becomes $B = B_0 n^{\psi}$, and it is possible to investigate how much bandwidth induces power-limitation in a large network as a function scaling with its number of nodes n.

The scaling results hold for different channel models, listed and discussed in Section 6.2, but the key model for mmWave 5G is the large-bandwidth fading channel without a-priori CSI, called a *non-coherent wideband channel* in information theory literature. Recent experimental measurements have demonstrated that mmWave outdoor links often rely on diffuse reflections with multiple NLOS paths [54, 74]. These diffuse reflections introduce new channel coefficients as bandwidth grows, so the wideband fading models described in [75, 76] apply.

In these models, it is possible to estimate the channel at limited bandwidths by dedicating a fraction of power overhead to pilot signals and still work with coherent receivers. Médard and Gallager [75] showed that for non-peaky[2] signaling, mutual information decreases to zero as channel estimation errors increase when finite power is spread over an infinite bandwidth (*overspreading*). Lozano and Porrat [76] argued that below a certain *critical bandwidth* (which grows linearly in power) the wideband non-coherent channel rate may be modeled as a combination of the capacity of a wideband coherent channel (a wideband channel with CSI) minus a “channel estimation” penalty, and that rate grows with the degrees of freedom. When available bandwidth exceeds critical badwidth, transmitters may only allocate transmissions to a fraction of the available bandwidth and still use the coherent-receiver-on-non-coherent-channel strategy, or switch to fully non-coherent receivers as in [77, 78]. Either method achieves a rate that scales linearly with SNR [79], although different signaling schemes might yield a different pre-scaling constant [76, 77, 80]. Therefore, our scaling results are valid for any signaling scheme with those critical bandwidth and overspreading effects.

[2] Signals with finite fourth moment, or equivalently a finite-power signal that does not result from averaging a flash pulse of infinite instantaneous power over time

Due to the fact that using non-peaky signaling schemes rate decreases to zero as bandwidth goes to infinity [75], it would seem better to approach C^{∞} on a wideband fading channel using peaky signaling schemes as in [77, 78, 81], but peaky signals have two drawbacks:

- The hardware implementation of a peaky signal is very difficult due to the fact that a signal with high (infinite) fourth moment is impossible to synthesize on devices with finite dynamic range or other nonlinearities.
- Moreover, peaky signals have very poor *spectral efficiency* (in nats/s/Hz). This efficiency is controlled by the second derivative of $C(B)$ at the point $\mathrm{SNR} = 0$. This was shown in [80], using a second order Taylor expansion, to be finite for AWGN and coherent fading channels but $-\infty$ for non-coherent fading channels, meaning that $C(B)$ converges to C^{∞} very slowly as $B \to \infty$.

However, peaky signaling as in these analyses is only compulsory if our requirement is to achieve C^{∞} when $B \to \infty$. Instead, non-peaky signals can be enough if it is sufficient to approach the linear-in-power capacity at some large—but finite—bandwidth, even though the rate decreases to zero if bandwidth grows further. Lozano and Porrat [76] showed that, for non-peaky signaling in the single-input single-output (SISO) channel with any type of fading, there is a transitory first stage when rate grows with B while power is not too spreaded, and then rate achieves a maximum value of

$$\mathrm{SNR}\,(1-\Delta)\,, \quad \lim_{L_{\mathrm{c}}\to\infty} \Delta = 0, \tag{1.2}$$

where Δ is a gap that vanishes with the increase of the *coherence length* (L_{c}) of the unknown channel and does not depend on SNR. The peak is achieved at some critical bandwidth B^{crit}, and then, when B keeps expanding, the rate starts decreasing to zero. They also provide upper and lower bounds of B^{crit} for Rayleigh fading. As $L_{\mathrm{c}} \to \infty$, estimating the channel becomes increasingly rewarding and the capacity of the noncoherent channel converges to the capacity of the coherent channel.

The role of L_{c} with non-coherent signals contradicts the abrupt change of the derivative in [80] with peaky signals, either wether the channel is perfectly known or unknown. The conflict is solved in [78, 81], by representing the capacity in a non-coherent Rayleigh fading channel as a polynomial with order $1+\alpha < 2$:

$$\frac{C(B)}{B} \simeq N_r \mathrm{SNR} - \frac{N_r(N_r+N_t)}{2N_t}\mathrm{SNR}^{1+\alpha} + o(B^{-(1+\alpha)}), \tag{1.3}$$

where N_t is the number of transmission antennas. The first term introduces the power constraint on capacity (C^{∞}). The second term is sub-quadratic $\mathrm{SNR}^{1+\alpha}$ ($\alpha \in (0,1)$), it vanishes with $B \to \infty$ and its speed of convergence determines the *spectral efficiency*. The third term captures the quickly-vanishing approximation error at large B. The parameter α converges to 1 as L_{c} grows, showing that also for peaky signals the non-coherent capacity converges to a coherent capacity.

1.5.2.3 Multi-hop Scheduling

As discussed on Chapter 3, LTE-A cellular systems have taken the first steps towards multi-hop cooperation. However, the implementation of cooperation in the current standard is intrinsically limited by numerous constraints including the scarce spectrum of micro-wave bands and the decisions carried on from the original design of LTE as a single-hop architecture. Most mmWave models up to date have assumed a static TDD frame structure like that of current TDD LTE, where all subframes are globally synchronized with BSs transmitting in one common set of DL time slots and UEs transmitting in the complementary UL set (see 3.2). Since eNBs typically transmit at a much greater power than the UEs, if adjacent cells are in a different UL/DL state, the UL signals may be overwhelmed by interference from the neighbor eNBs. Therefore, to prevent concurrent transmissions of devices with highly asymmetric transmission power, DL/UL transmission patterns are universal remain essentially static across the nework, without the possibility of adjustment for load balancing or changing channel conditions experienced by mobile nodes. Additionally, in the current RN specification, communication between the BS and the RN can occur only in specific subframes, which could result in severe wireless backhaul bottlenecks.

Such static synchronized duplexing may not be necessary and, in fact, may be particularly disadvantageous for mmWave systems and wireless systems that use high-gain directional antennas. In these systems, interference from transmitters can be isolated even if there are significant power disparities (as found in [74]).

Results for other types of wireless networks, such as ad-hoc networks, can help in the task of modeling a multi-hop 5G cellular system. A more flexible Dynamic Duplex (DD) allocation will become feasible, in which BSs, RNs and UEs will decide with topological awareness if they can be simultaneously active. This is a new paradigm with fundamental differences with traditional cellular networks. The study on directive MAC protocols in [82] highlights the need for more generic analytical models in multi-hop wireless networks with beam-forming and lack of fair protocols.

Kelly's work in [83, 84] introduced the Network Utility Maximization (NUM) approach. In the seminal work [85], the concept of imperfect scheduling was first introduced, permitting sub-optimal instantaneous schedules at each decision point without violating long-term throughput optimality (Pick and Compare algorithm, PaC). Subsequent extensions over the years introduced multi-hop structures [86, 87]; reduced time complexity of control signaling with or without trade-offs [88]; analyzed alternative QoS metrics beyond throughput and fairness, such as delay [89] or energy consumption [90]; considered fairness in heterogeneous networks with time-varying channels [91]; discussed different methods to model interference; dealt with optimal power allocation [92, 93]; studied the way reconfiguration delays affect network capacity [94], etc. We refer the reader to [95, 96] for surveys on these topics. With the exception of some recent studies that cover limited models, none of the previous works has considered beam-forming or mmWave communications. For instance [97] focused on video quality, did not take interference into account, and only provided a centralized solution.

Some authors have introduced a *pseudo-wired* hypothesis [98], claiming that mmWave wireless beams are so directive that interference should be negligible and independent links behave as fixed capacity wires. In this thesis this hypothesis is treated with extreme caution and put to test, rather than blindly followed.

1.5.3 Contributions

C1.5.1) **Throughput Capacity Scaling Laws for Wideband Cellular Networks** Section 6.3.1 obtains the throughput capacity scaling law of a cellular network with number of users, area, bandwidth, number of BS, and number of BS antennas. The protocol that achieves capacity scaling is cooperative multi-hop.

C1.5.2) **Critical Bandwidth Scaling Limit** The major change between operation regimes is the transition from capacity being limited by the degrees of freedom in the network to capacity being power limited. This mirrors the point-to-point result for capacity in wideband channels. The capacity scaling results (Section 6.3.1) show that bandwidth can only scale while contributing to the rate up to a threshold. Above that threshold, the network enters the power-limited regime and increasing bandwidth no longer allows to increase rate.

C1.5.3) **Only Cooperative Multi-hop Achieves Capacity Scaling in All Regimes** Section 6.3.2 studies the traditional single-hop transmission strategy in cellular networks, the use of dedicated RNs with lower density than users to implement multi-hop, and hierarchical cooperation in infrastructure networks. None of them can guarantee optimal capacity scaling for all scaling values of the network parameters, although there are degenerate cases, such as single-hop with a dedicated BS per user or RN multi-hop with as many RNs as users, where these protocols may be optimal.

C1.5.4) **Data Flow Models Are Critical for Accurate Scaling Analysis** Our result for cellular networks and prior ad-hoc network results coincide in that multi-hop is efficient in extense networks where short-distance communications are power-limited. However, for dense networks, direct transmission is optimal in narrowband cellular networks, which contradicts the fact that in dense ad-hoc networks direct transmission is suboptimal and hierarchical cooperation is the optimal strategy. This is due to the very different spatial traffic distributions in cellular and ad-hoc networks. In the former, nodes always communicate with the closest BS at hand, whereas in the latter nodes may communicate with any other node in the network. Consequently, direct communications between distant nodes in an ad-hoc network are long-range in a topological sense, because they cause interference to many neighbors in between, but they are short-range in a physical sense, because the network is dense and transmission distance is short. To illustrate the difference, Section 6.4.1 constructs a cellular hierarchical protocol and analyzes its operation in comparison to ad-hoc hierarchical cooperation. The analysis shows that cellular hierarchical cooperation is suboptimal because of the difference in spatial distribution of the network traffic, which is concentrated around the BS in the cellular case.

C1.6.5) **Unified Bandwidth Occupancy Framework for Peaky and Non-peaky Wideband Signals** Peaky and non-peaky signaling schemes have considered apart in non-coherent wideband fading channels because of their extremely different behaviors as bandwidth goes to infinity ($B \to \infty$). Peaky signals can achieve asymptotically the linear-in-power capacity of a wideband AWGN channel with the same SNR,

$$C^{\infty} = \lim_{B\to\infty} C(B) = \lim_{B\to\infty} BN_r\mathrm{SNR} = N_rP/N_0, \ \text{[nats/s]},$$

where P is the power, N_0 is the noise Power Spectral Density (PSD), N_r is the number of reception antennas, and $\mathrm{SNR} = P/(BN_0)$ is the SNR per degree of freedom at each reception antenna. On the other hand, non-peaky signals can only reach a peak rate at some finite *critical bandwidth* and then the rate falls to zero when bandwidth grows to infinity above the critical value. Section 7.2.1 describes a recent result of this thesis that unifies the theoretical study of peaky and non-peaky signaling, showing that they are nothing but corner points of a more fundamental trade-off in the *bandwidth occupancy* metric, that affects all types of signals.

C1.6.6) **Frequency Division Achieves Critical Bandwidth - Analysis of Multiuser Channels** 5G base stations will serve multiple users at once, relying on large bandwidth and numbers of antennas. This will be possible thanks to the small wavelengths and the corresponding miniaturization of antenna elements, exploiting multi-input multi-output (MIMO) multiuser techniques in the wideband regime. Section 7.2.2 describes a recent result that gives definition to the critical bandwidth of a multi-user channel, analyzes Multiple Access Channels (MAC) and Broadcast Channels (BC), and proves that the optimal strategy with a very large bandwidth is to separate users in orthogonal sub-bands.

C1.6.7) **Massive MIMO Mitigation of Overspreading in Wideband Channels** Even though the observation that as bandwidth grows capacity becomes power limited is classic, this holds in the traditional analysis of wideband channels, where there is an equivalence between B going to infinity, SNR going to zero, and capacity being power-limited. However, all the work in [75, 76, 78] assumes a fixed number of antennas, which leads to the ratio $\lim_{B\to\infty} \frac{N_\mathrm{r}}{B} = 0$, meaning that SNR $\to 0$ (low SNR regime). However, in combination with a second asymptotic behavior in the number of receive antennas $N_\mathrm{r} \to \infty$, their conclusions do not necessarily extend to Massive MIMO. Section 7.3 discusses the handicaps of implementing wideband transmissions with non-massive traditional peaky and non-peaky PHY schemes and shows that all of them have drawbacks, whereas by using a massive number of reception antennas it is possible to circumvent the critical bandwidth occupancy limitations and make rate grow with bandwidth to a larger extent.

C1.6.8) **Dynamic Resource Allocation in a Tree-topology mmWave Network with Multi-User MIMO** Section 7.4.1 describes a recent result in this thesis on the optimal resource allocation in a tree-topology mmWave network where nodes do not adjust to static TDD of LTE, but instead may choose their uplink and DL duplex ratios freely. The algorithm permits a node to transmit or receive to/from multiple neighbors at once, with orthogonal bandwidth and power allocations. This problem setup focuses on multi-user MIMO gains. However, a tree topology only lets users attach to one Access Point (AP) at a time, ignoring potential gains due to the diversity of traffic routes.

C1.6.9) **Dynamic Link Scheduling in an Arbitrary-Topology mmWave Network with Single-User MIMO** Sec. 7.4.2 describes a recent result in this thesis on the optimal link scheduling in a mesh-topology mmWave network where, again, nodes do not adjust to static TDD of LTE, but instead may choose their uplink and DL duplex ratios freely. The algorithm permits to attach a user to as many APs as it wants at once. This problem setup focuses on routing diversity gains. However, the problem of bandwidth allocation in a mesh topology is too complex and this algorithm can only let nodes transmit/receive to/from one neighbor at a time, ignoring potential gains of multi-user MIMO.

Part I

The Present: Cooperative Communications and Current Wireless Networks

Chapter 2

The Theory of Cooperative Communications: Cooperative Diversity for Wireless Networks

Contents

2.1 Introduction

Signal propagation in wireless channels varies in time due to effects such as fading, and so does instantaneous achievable rate. Thus, wireless systems typically include some degree of *diversity* so as to provide the receiver with several independent opportunities to receive the signal with a favorable channel realization. This increases the chances of a successful transmission and ultimately allows to increase the throughput of the network [99].

Many forms of diversity are possible depending on how different available channels or subchannels replicate the signal. *Time diversity* consists of transmitting replicas of the signal with enough delay in time to allow channel realizations to be uncorrelated. *Frequency diversity* relies analogously on different carriers in frequency-selective channels. *Space diversity* makes use of multiple antennas that are sufficiently spaced and transmit the same information, and *multi-user diversity* alters the scheduling order of resource allocations to exploit the fact that different users experience independent fading [99]. Among these diversity techniques, the *space diversity* of multi-antenna systems is particularly interesting since it is generated by spatial resources (i.e. antennas) that can be replicated if necessary, with no physical limitations except the budget. This can easily complement or even deprecate the basic forms of diversity that rely on physically scarce resources such as spectrum or time.

Wireless user devices tend to be constrained in size, complexity and power, limiting multiple-antenna spatial diversity. The *cooperative diversity* paradigm copes with this problem. It brings the advantages of multi-antenna space diversity to single antenna devices in a network, which cooperate and share their antennas to form virtual antenna arrays. This new approach has a great potential to improve the performance of current wireless networks. However, cooperative diversity is complex and its implementation is still in its early stages. In the first part of this thesis we describe the main levels in the analysis of cooperative communication mechanisms. The first level is the information-theoretical characterization of capacity in cooperative channels; the second level is the design and evaluation of physical layer (PHY) techniques to implement signals for a cooperative transmission; and the third level is the design and evaluation of MAC layer protocols to coordinate the negotiation and usage of those techniques in cooperative wireless networks.

When a single-antenna node in a network coordinates itself with others to form a virtual antenna array, it is possible to employ the techniques of MIMO spatial diversity [99]: each source associates itself to some *helpers* that first receive the transmission of the source and then relay the information. But as a result, one extra transmission is needed to send the information to the receiver and the number of transmissions is twice those of “local” MIMO. It must be noticed that the increase in cost is only due to the second stage, since the broadcast nature of wireless network allows simultaneous transmission to as many helpers as needed in the first stage. Furthermore, multi-hop transmission [100] with cooperative diversity may favor a better reception in each receiver along the path and, thus, improvements in range, rate or autonomy. These tradeoffs are analyzed in [99] and [101], which conclude that the strategy is profitable. For this reason, future wireless network designs should consider cooperation capabilities.

Cooperative diversity, as a technique to combat fading, should find its niche in the upcoming generations of mobile data networks, typically cellular architectures, improving throughput by means of spectrum reutilization. Consequently, the cooperative scenario requires new analyses as the introduction of additional transmissions increases interference [102, 103]. Figure 2.1 illustrates the mechanics of interference increase: The helper (H) receives the transmission of the source (S) and forwards it to the destination (D). Each of these three nodes have a certain range represented as a dotted circle. Neighbor nodes not related to the cooperative transmission are represented without a label. In this example, it can be seen that some neighbors are only in range of H, and the interference they are suffering would not have happened in a direct transmission.

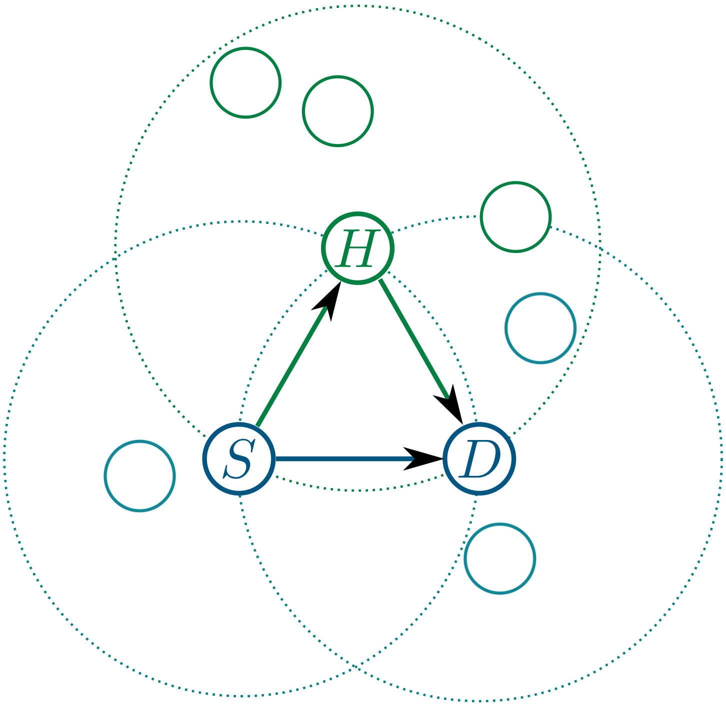


FIGURE 2.1: Interference is extended by cooperation.

Any wireless network affected by fading may benefit from cooperative diversity techniques. Since this benefit increases as more potential helpers conform the network, *dense sensor networks* [104, 105] represent a good application scenario for low complexity cooperative techniques. Virtual antenna arrays to construct "data highways" [61, 106] may also be interesting in *ad-hoc networks* [100, 107].

Cooperation can be exploited for other purposes than improving capacity. Cetinkaya and Orsun proposed a MAC protocol in which nodes cooperate to adapt their contention to the channel and improve fairness [104]. However, throughout this chapter we focus in cooperation as an enabler for better wireless capacity, not for user coordination at upper layers. The main difference is that cooperative diversity gains in wireless networks introduce changes in the signaling strategies, and in return cooperation takes an active part in improving the network throughput; whereas cooperative coordination does not affect the PHY layer and only modifies the user shares of the existing rate.

This chapter is structured in four sections. Section 2.2 reviews the background from the perspective of information theory, emphasizing the fact that a point-to-point solution cannot achieve the full capacity of a network [60]. Different theoretical models of achievable rate and capacity are reviewed and compared. Section 2.3 analyzes recent PHY architectures. A representative group of cooperative PHY techniques is discussed considering their relation to the information-theoretic models. Section 2.4 divides cooperative MAC protocol duties into five pseudo-services or tasks that all protocols perform, and creates a taxonomy of MAC protocols in the literature according to those tasks. Finally, section 2.5 summarizes the main ideas in the chapter.

2.2 Information Theoretic Model

2.2.1 Philosophy

Let us consider a four-node wireless network with two transmitters S_1 and S_2 and two destinations D_1 and D_2 as shown in Fig. 2.2. In the example, two sources S_1 and S_2 access the medium alternately by some means, such as, for instance, dividing every time interval T in two pre-assigned periodic symmetric Time Division Medium Access (TDMA) slots, where each source uses the channel half of the time to transmit to its destination as in Fig. 2.3. Assuming each link $i \in \{1, 2\}$ experiences a Rayleigh fading coefficient h_i that is slowly varying and flat, without cooperation, the link outage probability $P_o^{(i)}$ is defined as the probability that the signal-to-noise ratio seen by the receiver ($|h_i|^2$SNR, where SNR $= \frac{P}{N_0 B}$ is the signal-to-noise ratio at the transmitter) is lower than the minimum value necessary to sustain the transmission rate; or equivalently, the rate is above the mutual information (MI) of the point-to-point channel [41]:

$$P_o^{(i)} = P\left(R > \log_2(1 + |h_i|^2\text{SNR})\right) = P\left(|h|^2 < \frac{2^R - 1}{\text{SNR}}\right) \tag{2.1}$$

where R is the binary rate per Hertz, defined per slot.

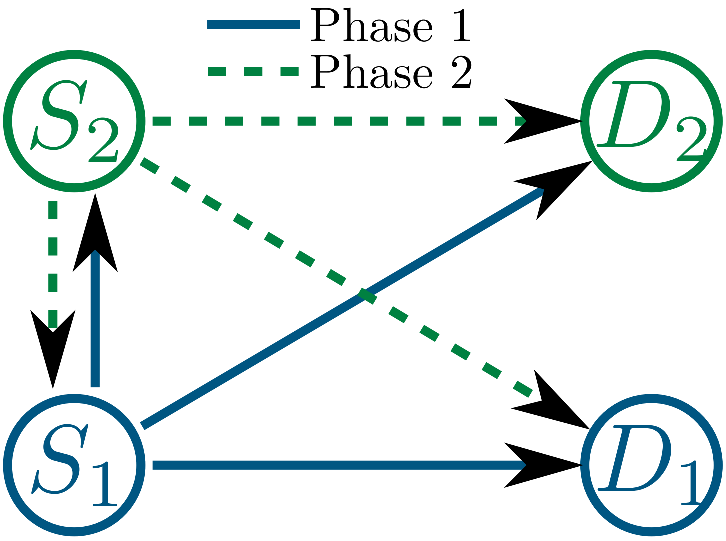


FIGURE 2.2: Example of wireless network where cooperation may be potentually useful.

Note that, due to the broadcast nature of the wireless medium, a node that is idle at a certain moment can overhear the transmissions of its peers. In the example, when one pair of nodes is idle, both may “overhear” the transmission of the other pair. In traditional wireless systems these received signals are simply discarded by the circuitry of a node when they do not address that node. However, since different nodes experience independent realizations of the fading phenomena, the success probabilities for each potential receiver are independent and, thus, it is likely than one of these neighbors will be able to decode the message before the intended destination. If we could recruit those neighbors as helpers to forward the message, this would create a form of diversity.

In the example, it is possible, with a simple modification -by halving the transmission rate- to allow a node to allocate half its transmission time to its own information and the other half to information relaying. Fig. 2.4 shows show the assignment interval T is now divided in four slots, encompassing two direct transmissions and two relay transmissions. The same information is transmitted twice -but once by each transmitter- and the rate is halved.

Using the forwarding policy above, and by decoding each transmission independently, the overall outage probability would already drop to the order of P_o^2, because a packet would only get lost if there is a simultaneous outage of two independent routes. But

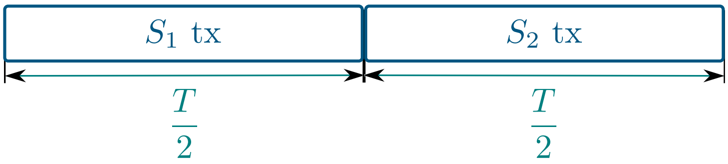


FIGURE 2.3: TDMA medium sharing without cooperation.

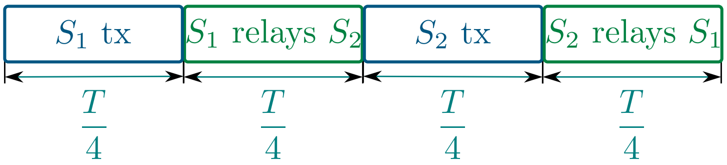


FIGURE 2.4: TDMA medium sharing with cooperation.

this model is just a toy example: link outages are assumed to be independent, packets with correlated information are encoded independently, the helper only retransmits information if it successfully decodes it, the receiver decodes each copy of the message independently, etc.

Taking information theory a little further and analyzing the mutual information in the whole four-node setup at a time [41], we can exploit other cross-layer techniques not available in the previous point-to-point mutual information characterization. Let us consider a wireless channel where the source transmits signal $x[n]$ in odd time slots, the relay receives

$$y_r[n] = h_{sr}x[n] + z[n]$$

and then processes it with some undetermined *relaying function*, with distribution $P(X_r = f(Y_r))$, to transmit $x_r[n]$ in even time slots, and finally the destination receives

$$y[n] = \begin{cases} h_{sd}x[n] + z[n] & \text{odd} \\ h_{rd}x_r[n] + z[n] & \text{even} \end{cases}$$

The mutual information in this scheme is given by the combined distribution of the signal that reaches the receiver through the direct link (Y) and the signal received through the two-hop link Y_x that is itself a function of the message from the source. This means that the destination can use information from both the received direct transmission and the relayed transmission to decode the data. Laneman et al [108] considered two classical relay algorithms depending on whether or not the relay/helper decoded the received signal: *decode-and-forward* (DF) and *amplify-and-forward* (AF), respectively. In cooperative AF, an optimum receiver accesses the information of two parallel noisy channels, one of which has a classical AF relay Y_{AF}.

$$\mathrm{I}\left(X;Y|X_r = AF(Y_r)\right) = \frac{1}{2}\log_2(1 + \mathrm{SNR}|h_{sd}|^2 + \frac{\mathrm{SNR}|h_{sr}|^2\mathrm{SNR}|h_{rd}|^2}{\mathrm{SNR}|h_{sr}|^2 + \mathrm{SNR}|h_{rd}|^2 + 1}) \quad (2.2)$$

Regarding DF, either the source-to-relay or the relay/source-to-destination links limit the maximum achievable rates.

$$\mathrm{I}\left(X;Y|X_r = DF(Y_r)\right) = \frac{1}{2}\min(\log_2(1+\mathrm{SNR}|h_{sr}|^2), \log_2(1+\mathrm{SNR}|h_{sd}|^2+\mathrm{SNR}|h_{rd}|^2)) \tag{2.3}$$

Based on classical relay techniques, Laneman et al suggested some improvements [41]:

- *Selection Relaying* is a control mechanism that can be used either with AF or DF. It detects the relay reception SNR and, if it is lower than a threshold, the relay stays silent and the source transmits again.
- *Incremental Relaying* is based on the assumption that most protocols implement some kind of ACK of the message. In this control scheme there is a single regular source transmission in first place and, if the helper receives the message but it misses the ACK by the destination, then the helper relays its copy of the signal to the destination.

2.2.2 Theoretical Relaying Model

In the general case, the capacity of a Relay Channel is given by the supremum over all probability distributions of the relaying operation $X_r = f(Y_r)$, over the max-flow min-cut constraint[109]

$$C = \sup_{P(X_r=f(Y_r))} \min(\mathrm{I}\left(X_r, X; Y\right), \mathrm{I}\left(X; Y, Y_r|X_r\right) \tag{2.4}$$

In the case of outage analysis, it is considered that wireless channels experience a random channel fading that remains constant for a long period. In this context the metric of interest is the probability that the channel realization cannot keep the rate, called outage probability, $P_o(\mathrm{SNR}_{norm})$ and expressed as as a function of a transmitter SNR normalized over the minimum SNR required by the chosen rate and protocol. The normalization allows to study diversity by separating the improvement due to diversity from the improvement due to the different spectral efficiency of each protocol, which alters the relation between rate R and SNR.

The *diversity order* of a system is defined as the exponent that asymptotically relates SNR increase to P_o decrease. This was formulated in [110] as:

$$d \equiv -\lim_{SNR\to\infty} \frac{\log(P_o)}{\log(SNR)} \tag{2.5}$$

A system with N_t transmitter antennae and N_r receiver antennae is said to provide *full diversity* when $d = N_\mathrm{t} \times N_\mathrm{r}$. In addition, the trade-off between diversity and normalized spectral efficiency -i.e. R normalized by the average rate sustainable for the average SNR- can be analyzed: *full diversity* is achievable when R_{norm} is zero as shown in Fig. 2.5. Incremental AF achieves the upper bound of a 2-antenna system. The normalized expression is:

$$R_{norm} := \frac{R}{\log_2(1 + \sigma^2_{s,d}\,\mathrm{SNR})} \tag{2.6}$$

where $\sigma^2_{s,d}$ is the channel variance. It may be identified as the improvement in R related to SNR, called *multiplexing gain* in [110]:

$$r \equiv \lim_{\mathrm{SNR}\to\infty} \frac{R}{\log(\mathrm{SNR})} \tag{2.7}$$

The diversity increase has some cost in spectral efficiency caused by repurposing a fraction of the resources, from transmitting additional data in parallel to retransmitting the same data to increase its protection. This is known as the *diversity versus multiplexing trade-off* ($d(r)$) [24], which is the main metric of interest for PHY procedures as explained in section 2.3.1.

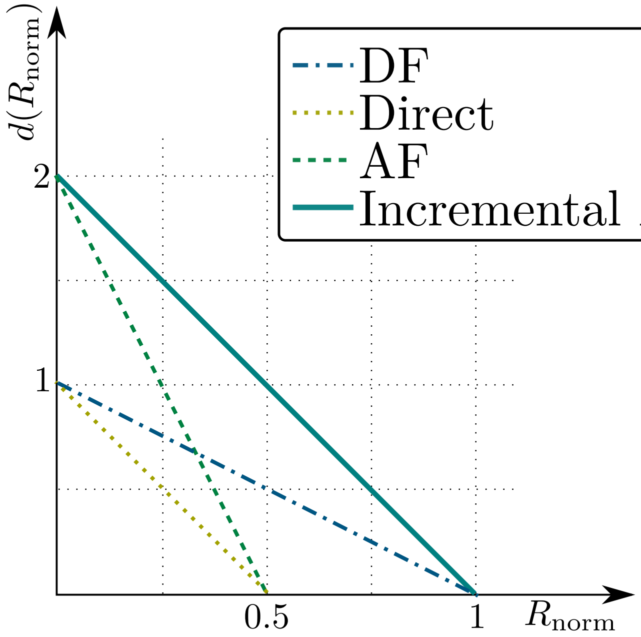


Figure 2.5: Diversity versus multiplexing tradeoff for direct transmission, AF, DF and incremental AF.

In [111], the authors proposed dropping the orthogonality assumption on source and relay nodes, in what they called the *Non-Orthogonal-AF Protocol* (NAF), allowing the source to join the transmission in the second slot. For DF, they proposed the *Dynamic DF Protocol* (DDF) without fixed time slots, in which, assuming incremental redundancy

Table 2.1: Orthogonality options for source and relay transmissions.

Phase	Option 1	Option 2	Option 3
I	S→H,D	S→H,D	S→H
II	S+H→D	H→D	S+H→D

codes, the relaying phase starts once there is enough information to decode (hence the term dynamic). Additionally, multi-user centralized schemes were outlined by analyzing the MAC channel (which is operated optimaly using NAF in the uplink) and BC channel (operated optimally using DDF in the downlink) architectures.

Contemporarily, Nabar, Bölcskei and Kneubühler designed space-time codes with three orthogonality options (table 2.1) [112].They arrived at the same conclusion: allowing the source to transmit during the second phase (Mode 1) increases overall performance. This is compatible with AF and DF.

The work in [110] extended that in [111], including the results of Nabar, Bölcskei and Kneubühler [112], whose *policy-1* used with AF (NBK-AF) is equivalent to NAF. It achieves the upper performance bound for the AF family, outperforming the orthogonal AF of Laneman, Tse and Wornell [41] (LTW-AF). The proposed DDF compared favorably to LTW-DF and, in almost all cases, to NBK-DF, as shown in Fig. 2.6. Note that DDF is the best overall option, and NBK-DF is slightly better in a small region. In addition, DDF outperforms the best AF strategy (NAF).

In [113], the authors focused on the differences between static and dynamic architectures. They provided two new schemes, *Extended Static DF* protocol (ESDF), allowing a fixed cooperation in time division, and *Extended Dynamic DF* (EDDF), which is dynamic. ESDF is superior to NAF and other static protocols, and EDDF is a modification of DDF with a slightly better trade-off.

Escalating from the simple two-user case to a full network with many users requires the study of systems with many helpers. In order to model the set of helpers of a transmission, we start by considering $\mathcal{H}$ as the set of all neighbors that can be potential relays. Many authors define the subset of relays that decode the message successfully $\mathcal{D}(\mathcal{H})$. This could lead us to think that, by forcing $\mathcal{D}(\mathcal{H}) = \mathcal{H}$ in our model, AF systems can be analyzed as a subfamily of DF. However this doesn't take into account that a transmission using DF from the nodes in $\mathcal{D}(\mathcal{H})$ does not amplify the noise they have received, whereas, in AF, all nodes in $\mathcal{H}$, in addition to always relaying, do amplify the noise. More generally, the surrounding nodes could potentially contribute to increase the receiver information even if they have do not achieve full decoding, encoding their received partial informations using different relaying functions, to be combined in a constructive way at the receiver.

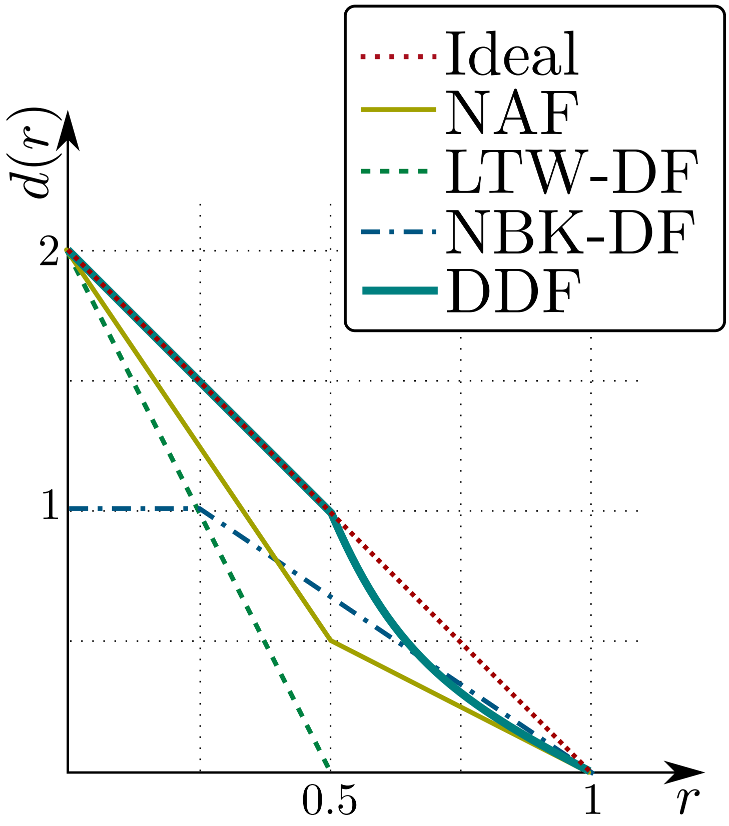


FIGURE 2.6: Diversity of NAF, LTW-DF, NBK-DF and DDF.

In [114], Laneman and Wornell studied systems with multiple nodes where all of them help each other. They proposed two cooperative schemes to divide time depending on how several helpers transmit concurrently:

1. Each user channel is divided into N time slots, where N is the number of transmitters. In each user's channel, all other users employ one of the slots to relay information as shown in Fig. 2.7. Since at each time slot a different relay transmits the same data, the rate R is divided by N.

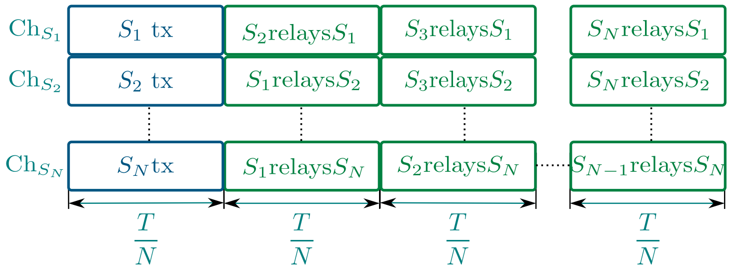


FIGURE 2.7: TDMA-based system with diversity order N and $R_{norm} = 1/N$.

2. The user channels only have two slots as in Fig. 2.8. In each user's channel, all other users that have succesfully decoded information relay it simultaneously in the second slot using Space-Time Coding (STC). The operator $\mathcal{D}(S_i)$ stands

for the subset of relays that receive correctly from S_i and perform the relaying actually. All relays transmit simultaneously so that information is only repeated twice, so that R is divided by two.

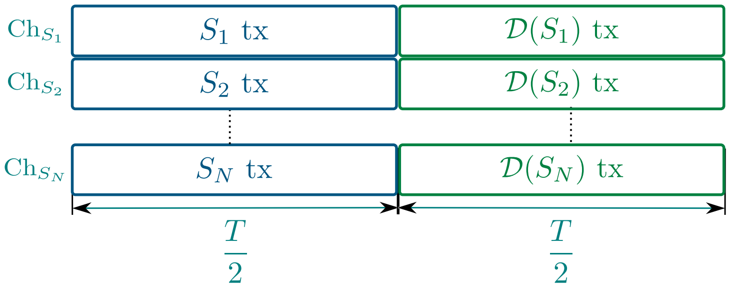


FIGURE 2.8: STC-based system with diversity order N and $R_{norm} = 1/2$.

Both schemes are valid both with AF and DF, as well as with their improvements in [41].

2.2.3 Open Issues

Cooperative diversity brings the advantages of multi-antenna transmission techniques to single-antenna nodes within a network. Despite the fact that it is necessary to split resources to allocate relaying transmissions, this technique can still offer important benefits. The study of cooperative diversity is similar to that of classical MIMO, using *diversity gain* and *multiplexing gain* as metrics, and taking into account that multiplexing gain is restricted to values below 1.

There are multiple alternatives from information theory to improve cooperative diversity transmission systems. Considering the basic approach of a single helper that relays source information, the first alternative is to increase the number of helpers. In addition, the orthogonal transmission constraint may be removed using coding techniques borrowed from multi-antenna transmitter design. Thus, it is possible to allow several helpers to transmit simultaneously, or to allow the source to transmit fresh information while previous information is being relayed. Time-division planning represents another degree of freedom to enhance cooperative diversity, either by associating the activation of the second phase to the failure of the first transmission (incremental relaying), or by tuning the length of each phase so that the channel split is optimum (dynamic protocols).

There are many open issues related to the theoretical aspects of cooperative diversity. We can classify them in three problems: generalization to non-trivial network topologies, relaxation of ideal assumptions, and design of advanced relaying functions to achieve (or approximate better) the theoretical trade-off bounds.

2.3 PHY Layer

2.3.1 Theoretical Overview

The PHY layer of wireless systems with cooperative diversity is usually modeled as a MIMO system with multiple hops. Without loss of generality, we will focus on the proposals that analyze two-hop schemes where one source performs Single-Input-Multiple-Output (SIMO) communication with multiple receivers (one or more helpers and a destination) followed by multiple transmitters (one or more helpers and the transmitter) that perform Multiple-Input-Single-Output (MISO) communication with the destination.

There are, in fact, many proposals for distributed MIMO that work with multiple clusters of helper users forming a virtual antenna array each, and forwarding traffic in array-to-array multi-hop routes. This would include a cluster of helpers around the transmitter, a different cluster around the receiver, and, optionally, intermediary clusters. These systems have three hops in total or more, if one counts the communication from the source to its cluster, one inter-cluster transmission at least, and the final hop to the destination from the nodes in its cluster. However, the intermediary hops may be studied by parts, considering at first each of the MISO links from the set of source helpers ($\mathcal{H}_\mathrm{s}$) towards each node in the set of destination helpers ($\mathcal{H}_\mathrm{d}$), separarely; and secondly the MISO link from the successful helper set $\mathcal{D}(\mathcal{H}_\mathrm{d})$ towards the destination. If $\mathcal{H}_\mathrm{s}$ and $\mathcal{H}_\mathrm{d}$ have N_s and N_d componets respectively, the final diversity order should reach $N_\mathrm{s} \times N_\mathrm{d}$.

One possible goal for designing MISO or SIMO transmission diversity techniques is *full diversity*:

Definition 2.1. In an N-antenna virtual array, **full diversity** is achieved if outage probability decreases asymptotically with $\lim\limits_{\mathrm{SNR}\to\infty} P_o \sim \mathrm{SNR}^{-N}$.

Other designs define their performance criteria by generalization the previous consideration with the well-known trade-off between diversity and multiplexing gain [110][115]. We copy the definition here for clarity.

Definition 2.2. The **multiplexing gain** r is defined as the asymptotic exponent of rate R growth with SNR

$$r \equiv \lim_{\mathrm{SNR}\to\infty} \frac{R}{log(\mathrm{SNR})} \tag{2.8}$$

Definition 2.3. The **diversity gain** d is defined as the asymptotic exponent of outage probability P_o decrease with SNR

$$d \equiv -\lim_{\mathrm{SNR}\to\infty} \frac{\log(P_o)}{\log(\mathrm{SNR})} \tag{2.9}$$

The following lemma results from the simplification of the trade-off for the particular case of a two-hop one-to-N-to-one cooperative system

Lemma 2.4. *In an N-antenna virtual array, the multiplexing gain and the diversity gain are complementary and upper bounded by $d(r) \leq N + 1 - r$*

Thus, the achievable curve of diversity gain versus multiplexing gain $d(r)$, and its closeness to the upper bound $d_u(r) = N + 1 - r$, determine another common performance metric for cooperative diversity systems. Figs. 2.5, 2.6 and 2.10 show examples by [41], [110] and [116].

In a full MIMO problem with N_{t} transmission antennas and N_{r} reception antennas, the general trade-off upper bound is a curve instead of a line and its corners are $r_{max} = N_{\mathrm{s}} + N_{\mathrm{d}}$ at $d = 0$ and $d_{max} = N_{\mathrm{s}} \times N_{\mathrm{d}}$ at $r = 0$, but in cooperative systems there are bottlenecks at the first and last hops so that all transmitter and receiver helpers decode the same information. Therefore the diversity and multiplexing gains of any hypothetical intermediate MIMO hop would be constrained by the bottlenecks and we can study the problem as a composition of MISO systems.

By definition, the full-diversity and tradeoff criteria are equivalent for a system with $r = 0$ and $d(r) = N$ but tradeoff optimality is more general. Therefore, two techniques that achieve full diversity achieve the point $d(0)$ of the optimal tradeoff but may not be equivalent or optimal in the remaining regions. Oh the other hand, techniques that do not achieve full diversity may be optimal in other regions of the tradeoff for $r > 0$, resulting in a solution that may be appealing for applications that desire to increase rate (multiplexing gain) as well as outage protection (diversity gain).

For example, Prasad and Varanasi [115], in their protocol, combined non-orthogonal STCs from [112], which are better for higher values of r, and orthogonal STCs from [114], which are better for smaller values of r. This allows to take the best choice between orthogonal and non orthogonal STCs for the desired value of r, and by definition outperforms both of its components. Internally, the difference between the first two protocols is the fact that the source cooperates in the second STC phase in [112], so the hybrid design is equivalent to the source deciding to transmit or not in the second phase based on whether it desires more rate gain (r) or more outage protection (d) for its data.

The trade-off determines the guideline for all practical applications to turn the increase in outage protection into improvements of other performance metrics, by increasing the rate for the same error probability guarantees instead of improving transmission protection at the same rate. This is specially interesting in links that have an error-correction mechanisms on top, where below a certain BER threshold errors cannot be corrected but above this threshold they will be corrected with high probability. Therefore, these systems do not benefit from any BER decrease above the threshold. Other alternatives to exploit the gains are:

- Power saving: by transmitting with lower power, or equivalently with lower SNR, relying on cooperation to achieve the same BER.
- Coverage extension: by extending the maximum distance (d) in the network while the BER requirements are still met.
- AMC boosting: in an *Adaptive Modulation and Coding* (AMC) system, cooperators may allow the AMC to shift to a faster modulation scheme.
- Bandwidth increase: Bandwidth can be increased by reducing the time between symbols $T_s = 1/B$, which in turn increases the rate and reduces the SNR.
- Error probability reduction: reducing the error probability that depends on fading.

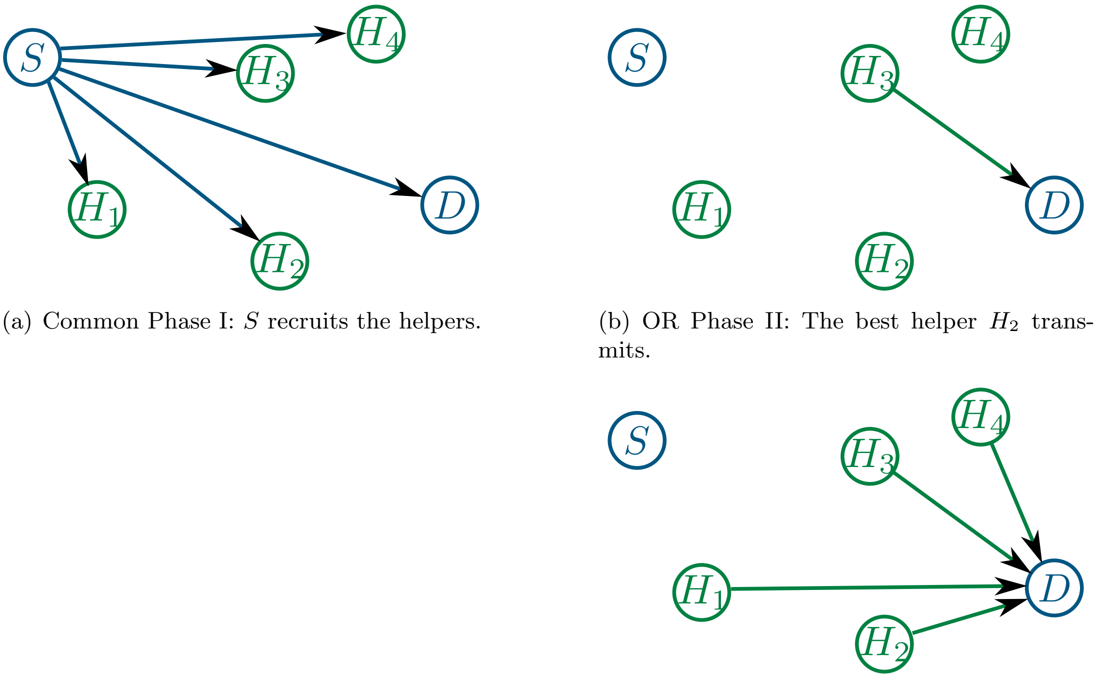
In practice, the combination of transmissions that arrive through different fading routes has been implemented in the PHY layer with different approaches, which we summarize in the following subsections.

2.3.2 The Best-Relay Approach

In [116], the authors proved that many helpers relaying information are not necessary to achieve a high diversity order. It is sufficient that they participate in an Opportunistic Relaying (OR) process for choosing the best relay, where only chosen ones transmit information. Fig. 2.9 shows the difference.

The relay selection mechanism proposed by Zheng and Tse is based on a MAC protocol with a contention mechanism implemented with a timer. This timer is initialized with a value proportional to the estimated channel qualities. This allows the best relay to gain access to the medium and become the only active one. Nevertheless, the disadvantage of contention-MAC strategies is that all potential helpers must lie within mutual range, to avoid the blind-node problem were several helpers may attempt to transmit simultaneously. If this happens, collision must be avoided somehow, for example with a collision detection followed by the destination selecting one of the relays with a Clear To Send (CTS) message. The proposed algorithm is explained in detail in section 2.4.1.4, along with many other MAC-PHY protocols that also select the helpers depending on

[!ht]



(a) Common Phase I: S recruits the helpers.

(b) OR Phase II: The best helper H_2 transmits.

(c) STC Phase II: All helpers transmit.

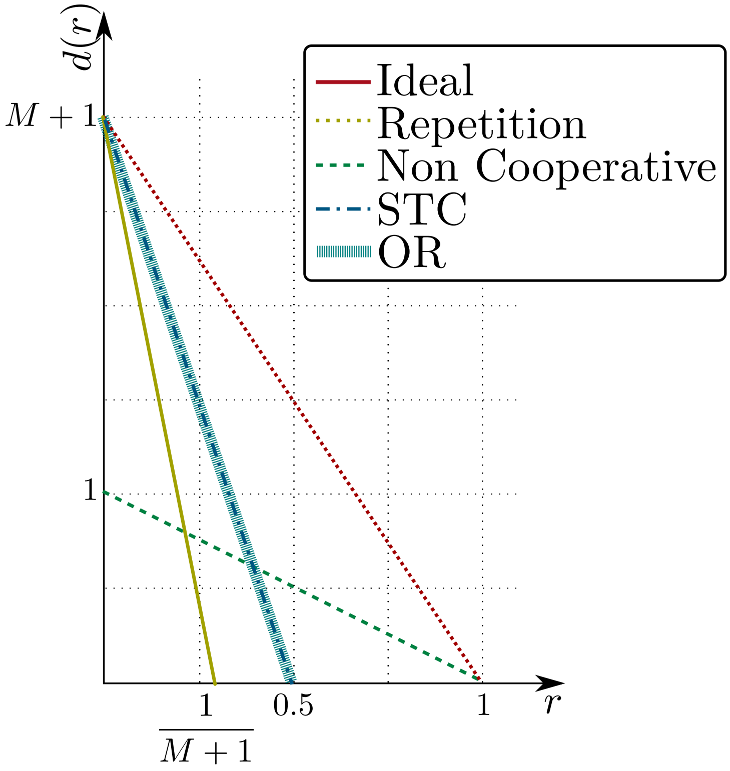
Figure 2.9: OR compared to STC

CSI using a variety of mechanisms to skip the disadvantages of a pure contention-based MAC.

When the best relay is selected, full diversity of order N can be achieved with only a single helper actually transmitting using plain repetition of the signal. This OR approach illustrates the fact that elaborate signal processing schemes such as STC in [114] are not mandatory to design cooperative systems with rate $R/2$ and diversity gain above 2.

Although their diversity versus multiplexing trade-off curves coincide ($d_{OR}(r) = d_{STC}(r)$) as shown by Fig. 2.10, this does not imply that the outage behavior of the two protocols is exactly the same (i.e. $P_o^{OR} \neq P_o^{STC}$). The metric $d(r)$ is asymptotic. This means that both outage probabilities tend asymptotically to $\text{SNR}^{-d_{STC}(r)}$ when SNR tends to infinity, but, for finite SNR, their behavior may differ and the trade-off would still be the same. At that finite SNR, we can easily verify that the outage probability of STC has to be less or equal than that of OR because, by definition, multiple replicas of the desired signal relayed by different helpers carry more information than the replica transmitted by the best helper alone.

Regarding the PHY, OR allows for simplicity. For simple point-to-point PHY signaling the system can still reach the benefits of full diversity asymptotically. But, as discussed



[!ht]

FIGURE 2.10: The diversity versus multiplexing trade-offs of OR and STC are equivalent.

above, this does not prevent the outage probability to be reduced even further in non-asymptotic scenarios. In the following sections, we discuss more elaborate PHY signal design techniques to improve the benefits obtained through cooperation.

2.3.3 Cooperative Channel Coding

Most classical systems include some kind of FEC codes at the PHY layer, and many of them are systematic, which means that each codeword begins with the original message bits followed by some redundancy. The analogy between systematic FEC and cooperative diversity is noticeable: in both cases additional information that increases the possibility of a successfull decoding is received after the original message. The comparison between cooperative diversity and *repetition coding* pointed out in [41] leads to what we call *cooperative channel coding.*

Srefanov and Erkip [117] elaborated on the idea of the source and the relay sharing a more complex coding architecture than repetition. In the scheme they proposed, the source message is protected by a 1/4 Convolutional Code (CC) that is carefully constructed so that its first two bits are in fact an independent 1/2 CC. The source *punctures* its encoded message, sending only the first two bits. The helper is expected to be able to receive the message correctly with a mere $\frac{1}{2}$ Viterbi decoder due to its proximity to the source, and finally it can compute the complementary code-bits that are missing in the original transmission and transmit them as relayed signal. Fig. 2.11 shows what this encoder would look like. The bits in a different color (from the complementary $\frac{1}{2}$ CC) are meant to be transmitted by the helper.

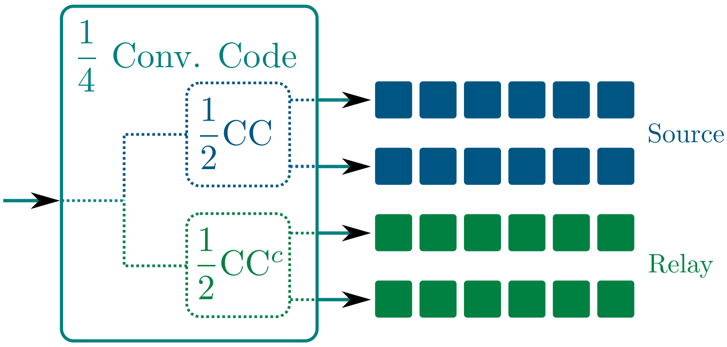


FIGURE 2.11: $\frac{1}{2}/\frac{1}{4}$ distributed CC.

If the relay failed to decode the message, the source could detect this by listening for a carrier (indicating an active transmission). And, if this carrier indicating that the relay is complying was not present, the source would start transmitting the remaining code-bits by itself. Moreover, the destination should store the two signals in the form of quantized soft symbols, merge the two symbol streams and perform full $\frac{1}{4}$ Viterbi decoding. Diversity is exploited by the Viterbi decoder because it performs better with higher values of energy per symbol E_x; and, therefore, when an independent channel provides half of the symbols, the probability that the decoder has to work with low values of E_x in all of them decreases, mitigating fading losses [118].

2.3.4 Cooperation through Network Coding

NC was originally proposed to merge correlated packets from multiple data streams in wired networks. In this type of coding, nodes that deliver related messages also forward compact summaries with their differences, typically linear combinations of multiple messages, to reduce the traffic load on the network. As an example, let us suppose that a destination can receive up to three packets from two different sources, and we wish to use the extra slot to increase reliability. In the NC scheme the destination receives in three consecutive transmission slots x_1, x_2, x_3 the binary sequences s_1, s_2 and the XOR combination $x_3 = s_1 \oplus s_2$. If at each slot the probability that the packet x is received with error is $P(\mathrm{e}_x)$, if only the two first slots are available the probability of **both** packets successfully arriving is $P_{\text{success,2slot}} = (1 - P(\mathrm{e}_x))^2$. However, if we add the third packet with the combination, the two messages can be recovered from the correct arrival of any two packets, and the success probability grows up to $P_{\text{success,3slot}} = (1 + 2P(\mathrm{e}_x))(1 - P(\mathrm{e}_x))^2 > P_{\text{success,2slot}}$ [119].

Note that, in this example, NC allows to achieve the same reliability as our first example in Fig.2.3, yet saving one transmission slot. So the achievable rate factor would be 2/3 instead of 1/2, while diversity gain is still 2.

Even though redundant information is also appended after systematic data, the main difference of this scheme with channel coding is that redundancy consists in combining information from several traffic flows. In addition, practical implementations also differ usually in that channel coding operates at the symbol level whereas NC operates at the packet level.

The XOR operation in the example above is only a simple implementation of the concept. In the following examples we illustrate more powerful techniques to construct network-coded redundancy packets of multiple traffic flows. The performance of the encoding schemes can be increased relying on advanced mathematic operations.

2.3.4.1 Binary Linear Combination

Xiao et al [120] criticized the cost in energy or time of independently-encoded transmissions for each cooperative relaying. They proposed to use a binary XOR function to build a combined message for multiple relaying actions, and to allocate all the channel resources to its transmission. They remark that this is a NC solution, but the implementation also makes internal use of a cooperative channel code. Let us suppose that some node S_1 starts with a local sequence of symbols $s_L^{S_1}(t)$ to be initially transmitted and another sequence that S_1 has already received from some neighbor S_2 and intends to relay $s_R^{S_1}(t) = s_L^{S_2}(t-1)$. Let G_L and G_R be the channel coding matrices used to encode the local and relayed streams of data, respectively. Then the use of the XOR operation to construct a network coded packet

$$s_L^{S_1}(t)G_L \oplus x_R^{S_1}(t)G_R$$

can be rewritten as a concatenation of the two encoders and the combination of the two data streams as as a code concatenation

$$[s_L^{S_1}(t)|s_R^{S_1}(t)] \begin{bmatrix} G_L \\ G_R \end{bmatrix}.$$

And, finally, neighbor S_2 proceeds similarly at the next time slot

$$[s_L^{S_2}(t+1)|s_R^{S_2}(t+1)]\begin{bmatrix} G_L \\ G_R \end{bmatrix}$$

with $s_R^{S_2}(t+1) = s_L^{S_1}(t)$. Thus, the receiver part can be implemented as a stream back-and-forth iterative XOR decoder that obtains the combined data of both transmission streams at the same time using NC and cooperative channel coding jointly.

2.3.4.2 Non-Binary Finite Field Linear Combination

The proposal presented in [121] and extended in [122] departs from the assumption that the relaying devices have multiple orthogonal channels for direct and relayed transmissions (like in TDMA cooperative diversity in our first example). For given transmission slots, the focus is on the composition of the optimal relay messages. We have shown that an XOR operation offers gains in a three time-slot system, but it becomes inefficient in systems with more time-slots. For example, for four time-slots the binary coding scheme could only produce the following four messages:

$$[s_1|s_2 \oplus s_1|s_2|s_1 \oplus s_2],$$

which outperforms the approach in [41], because both s_1 and s_2 can be recovered from three different combinations of two correct packets. However, this code is still suboptimal because no information can be recovered at all if packets 1 and 3 get lost.

The fundamental problem is that two equal redundancy packets, such as $s_2 \oplus s_1 = s_1 \oplus s_2$, cannot be combined in any meaningful way to extract their information. Therefore, linear encoding mechanisms that make use of several redundancy packets need to compute the redundancy in some space enabling several different linear combinations of the same packets. Xiao et al. demonstrated that the binary field is insufficient to achieve the full potential of cooperation. As a solution they proposed to generalize the encoding problem to any Galois field ($\mathcal{GF}(2^m)$) with more than two values.

As shown in Fig. 2.12, each source composes redundancy packets by combining its own message and the received message (s_1 and s_2) differently using a Gaulois Field NC (GFNC).

The four transmitted packets are

$$x_{L1} = s_1 \quad x_{L2} = s_2 \quad x_{R1} = s_1 \boxplus 2s_2 \quad x_{R2} = s_1 \boxplus s_2 \tag{2.10}$$

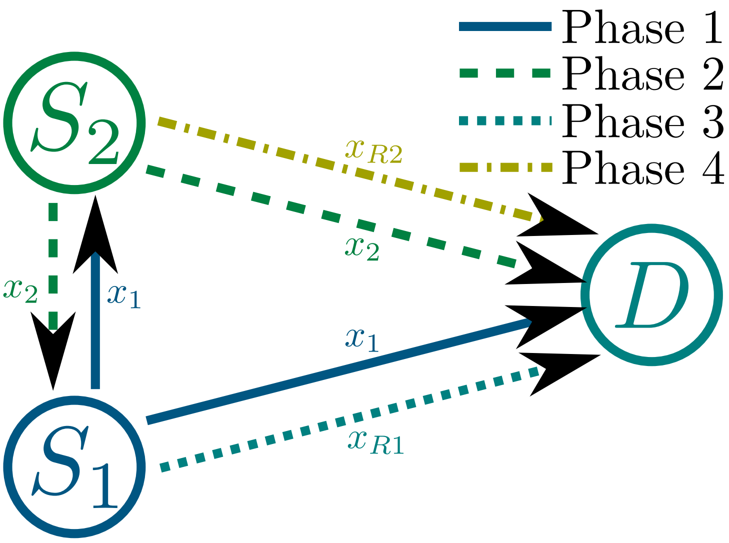


FIGURE 2.12: $\mathcal{GF}(2^2)$ network code for a 4-slot TDMA.

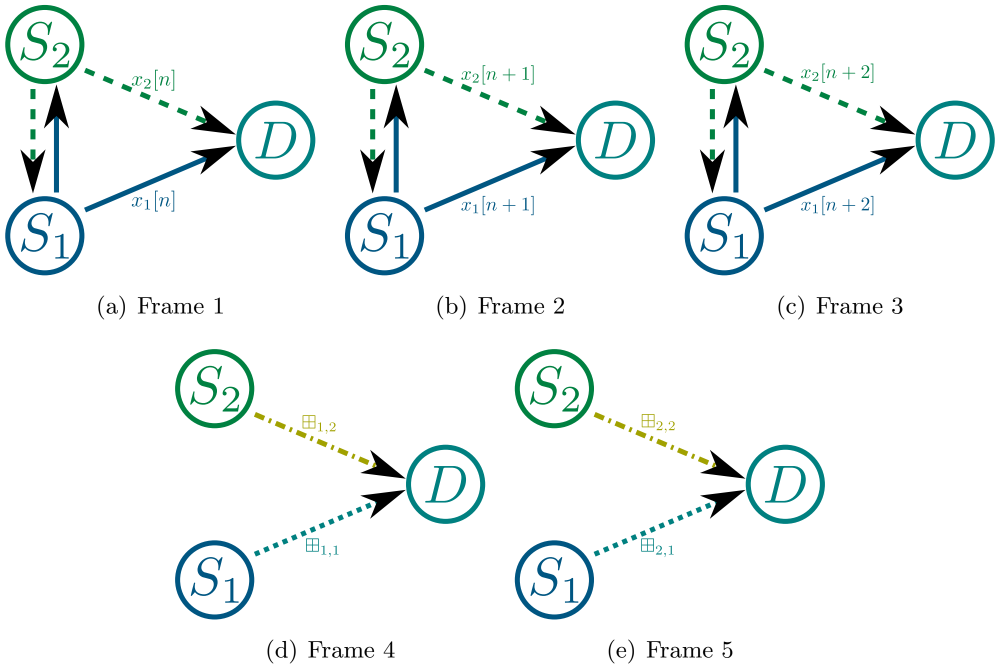
where, using a Galois field with four values, $\mathcal{GF}(4)$, we denote the sum and product on this field by $\boxplus$, and $\boxtimes$ respectively. The encoding operation can be written in matrix form as follows:

$$[x_{L1}|x_{L2}|x_{R1}|x_{R2}] = [s_1|s_2] \boxtimes \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 \end{pmatrix} \tag{2.11}$$

This type of encoding allows to recover one packet from the sum of two different redundancy packets ($x_{R1} \boxplus x_{R2} = s_2$) and the remaining packet afterwards ($x_{R1} \boxplus s_2 = s_1$). Therefore, successful decoding is possible in one more case. This fulfills the property that both original messages can be recovered by receiving any two packets, which was lost by the binary XOR function when we increased the number of slots to four.

In [121], one direct transmission and one relay transmission alternate as in the example above. In [122] an additional step extends coding to *super-frames*, which consist of an asymmetric proportion of alternating direct and relayed transmissions. Fig. 2.13 illustrates the mechanism for three data transmissions with two redundancy transmissions. In the first three frames the sources send six messages in total. Then, four different redundancy packets are computed and transmitted within the remaining two frames.

The matrix encoding representation for this case would result in a 6 × 10 matrix in a Galois field. This matrix can be optimized for encoding to minimize the lowest number of packets that always guarantee decoding. It is possible to establish an analogy between this problem and the problem of maximum distance separation in error correction block coding problems. The optimal solution is the well known family of Reed-Solomon codes, also known as maximum distance separable codes.



(a) Frame 1 (b) Frame 2 (c) Frame 3

(d) Frame 4 (e) Frame 5

FIGURE 2.13: $\mathcal{GF}$-based 6/10 network code for a 6-packet/10-slot super-frame.

2.3.4.3 Superposition Modulation

A principle shared by the schemes above is that encoding is performed on discrete alphabets, while links are assumed to deliver discrete symbols. However, there is also work on the joint encoding of multiple continuous symbols in MAC and BC channels, to form constructively superposed multiple-access constellations or hierarchical broadcast constellations. Larsson and Vojcic [123] discussed the application of this to cooperative communications. They proposed sharing the energy of the helper by simultaneous transmission of two hierarchical signals: the local signal originated at the helper node and the relayed signal from a neighbor source. The transmitted signal, in the discrete domain, would be synthesized as

$$x[n] = \sqrt{1-\gamma^2}x_L[n] + \gamma x_R[n],$$

in which $\gamma^2 < 0.5$ is the energy-sharing factor and the local $x_L[n]$ and relayed $x_R[n]$ signals are encoded independently. Even though this is not the typical discrete finite field operation in classic NC examples, the fact that the combined signal is a linear combination of the local and relayed parts allows to define this as an NC approach.

When each combined signal has its independent symbol constellation, the combined signal may be represented through a *composed constellation*. Fig. 2.14 shows an example where each source uses Binary Phase Shift Keying (BPSK) on the symbols of Table 2.2

and the combination becomes a case of hierarchical 4-level Pulse Amplitude Modulation (4-PAM) constellation on Table 2.3.

Table 2.2: Symbols on BPSK constellation

$s[n]$	$x[n]$
0	-1
1	1

By considering the composed constellation, and assigning to each point in it the corresponding sequence of bits from the joint data stream for the local and relayed data, it results that the continuous-domain and finite-field domain combinations are always interrelated.

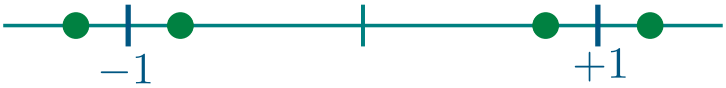


Figure 2.14: 4-PAM based on γ-superposed BPSK modulations

2.3.4.4 Complex Field NC (CFNC)

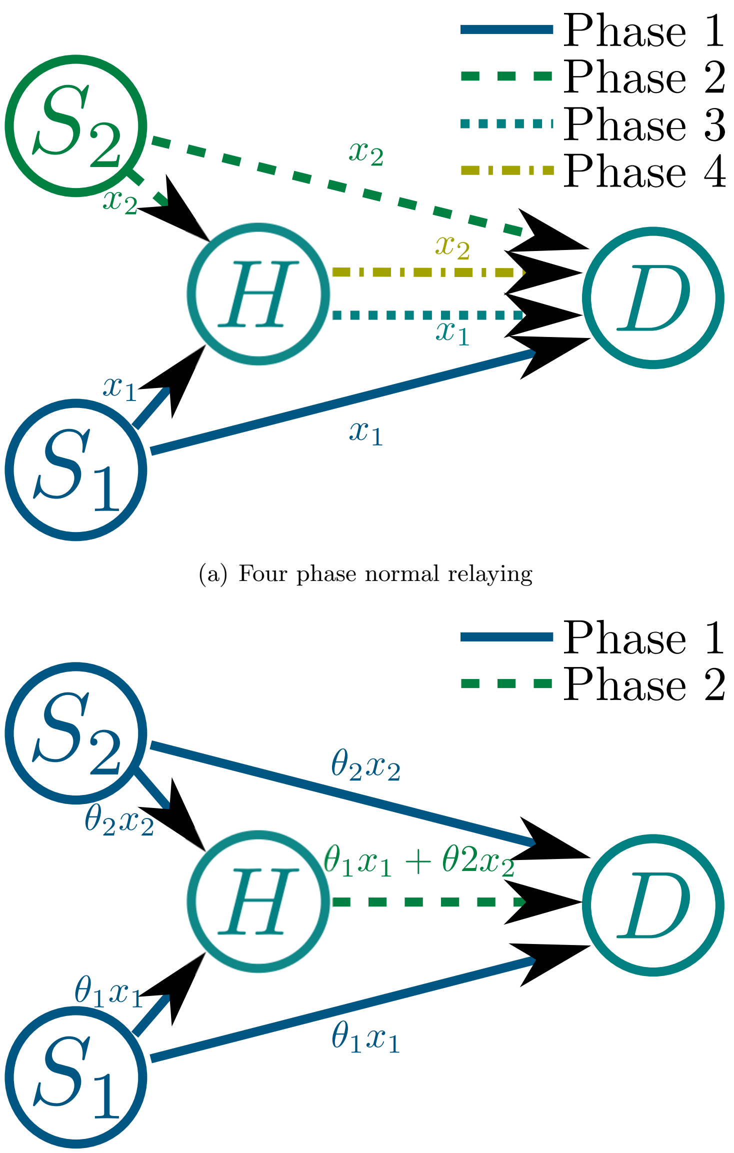
Building on the principle of PHY layer NC, Wang and Giannakis [124] proposed substituting the Gaulois field $\mathcal{GF}(2^m)$ with the continuous field of complex numbers $\mathbb{C}$ to bridge the gap between discrete codeword and continous symbol constellations. In this scheme the data sequences would be first translated into signals and then multiplyed by independent and orthogonal complex values $\theta_{i,j}$ to perform the NC operation.

$$(x_{C1}|x_{C2}|\dots) = (x_{L1}|x_{L2}) \begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \dots \\ \theta_{2,1} & \theta_{2,2} & \dots \end{pmatrix} \tag{2.12}$$

Fig 2.15 shows the savings in number of transmit phases by doing this. With sufficiently synchronized source transmissions, CFNC allows also to reduce the duration of the direct phase. This is because the sources can transmit the signal resulting from applying the corresponding row of the CFNC matrix to their local signal, and the signals separated in the complex plane are combined coherently in the air, forming the combined constellation. This is sometimes referred to, informally, as “XORs in the air”.

Table 2.3: Symbols on γ-4-PAM constellation

$s_L[n]$ $s_R[n]$	0	1
0	$-\sqrt{1-\gamma^2}-\gamma$	$-\sqrt{1-\gamma^2}+\gamma$
1	$+\sqrt{1-\gamma^2}-\gamma$	$+\sqrt{1-\gamma^2}+\gamma$



(a) Four phase normal relaying

(b) Two phase CFNC relaying

FIGURE 2.15: Comparison between plain relaying and complex NC to deliver four redundancy packets.

2.3.4.5 Dirty Paper Coding

Dirty paper coding is a technique based on the principle that a transmitter that knows a priori the interference that its receiver perceives can codify the information in such form that interference does not alter the decoding process. This is achieved using techniques such as nested lattice coding. In [125] it is proposed to use dirty paper coding to minimize the interference of the signals codified using CFNC or STC with each other. A CFNC relay does know the interferent signal perfectly, because it is the relay itself who adds it. Dirty paper coding could provide the means to generalize the aforementioned advantage of CFNC allowing multiple simultaneous transmissions in the direct phase.

2.3.5 Cooperation through Space-Time Coding

The STC transmission technique for multi-antenna systems mentioned in section 2.3.1 exploits related signals in separate antennas as spatial diversity. In brief, a block coding model consists of mapping a vector of L transmitted symbols $\mathbf{x}$ onto an $L \times N$ matrix $\mathcal{G}(\mathbf{x})$ whose columns correspond to the signal to be transmitted by each of the N antennae. Thus the discrete equivalent channel could be written by blocks as

$$\mathbf{y} = \mathcal{G}(\mathbf{x})\mathbf{h} + \mathbf{z} \tag{2.13}$$

Many methods have been proposed in the literature making use of different system parameters such as delay, gain, code, antenna, etc. In this section we present a non-exhaustive overview of the application of these techniques to cooperative transmissions. We consider all the STC techniques employed by the protocols in section 2.4. The reader may refer to [126, Section I.A *Related Work*] for a deeper review, to [127] for an approach to CSI-feedback relay phase and gain adaptation based on MIMO beam-forming, and to [128] for asynchronous code design.

2.3.5.1 Classical Space-Time Coding Techniques

The simplest coding adaptation would be to apply local STC techniques in the distributed cooperative transmission. This would simplify code design but imposes a large control overhead, as the members of the virtual antenna array must exchange information that would be locally available in a typical antenna array. Anghel et al [129] showed that space-time coding of classic multi-antenna systems performs adequately in distributed virtual array systems. They assumed that the relays know the code matrices and that each relay knows which column/antenna of the code each must utilize.

Cheng et al [130] analyzed an Orthogonal Space Time Block Code (OSTBC) base station with a helper relay. Fig. 2.16 represents a N_t-antenna source using OSTBC and a cooperative single antenna relay providing the destination with $N_\text{t} + 1$ realizations of the message. A diversity order of $N_\text{t}+1$ was demonstrated, thus showing that cooperative and local space diversities may benefit mutually.

Laneman and Wornell [114] elaborated on *deserter helpers* (appointed potential helpers that do not show up when they are expected to cooperate). In an OSTBC, removing a helper is equivalent to removing the corresponding column on the matrix of the code. Since the columns of the matrix are orthogonal, the shortened matrix is still orthogonal and carries all the information, so it serves as a punctured OSTBC. Nevertheless, a subset of the columns of an N-antenna OSTBC does not usually correspond to the optimum $(N-1)$-antenna OSTBC.

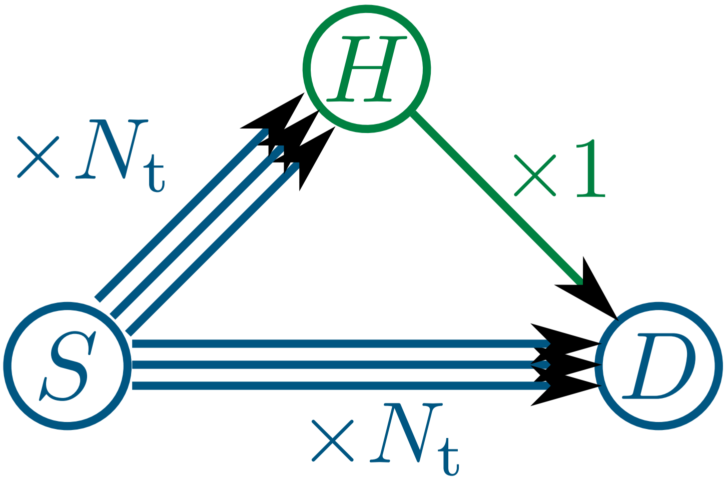


FIGURE 2.16: OSTBC base station with a single-antenna helper relay.

2.3.5.2 Distributed Space-Time Codes

Instead of using local STCs that rely on abundant local information, overloading the control protocol with so much coordination, some authors have worked on the design of STCs with less shared information, reviewing code design practices for the specific purpose of cooperative diversity. Some common study areas on Distributed STBC (DSTBC) design are:

- Constrained or unknown sets of available helpers and cooperation policies.
- Asynchronous helper tolerance: in classical MIMO architectures, all signals/antennae are synchronized.
- Notification of the code to the helpers, and their role within it.
- Helper desertion.
- Late-arriving helpers (those not present when the virtual array was first formed).
- Very large helper sets ($\mathcal{H}$), for which OSTBCs are unfeasible.

These problems cannot be solved separately, as there are trade-offs between them.

The initial approach with classic codes would rely on allocating one specific antenna of the code to each relay. Nevertheless, a single helper transmitting the wrong signal would spoil the whole code. Consequently a first step is using DF and CRCs to let only helpers that have received a correct packet collaborate. This means that the actual relay set is an aleatory subset $\mathcal{D}(\mathcal{H})$ of the set of helpers $\mathcal{H}$. In addition, the set of recruited helpers itself may be random as long as any node within recruiting range is allowed to join without further restrictions.

To avoid notifications, random antenna selection has been proposed. Although it is an elegant solution, the diversity order is reduced because the effective number of virtual antennas is only the number of columns of $\mathcal{G}(\mathbf{x})$ selected by at least one relay. This is distributed as the average of possible results in a multinomial distribution $(N_\mathrm{a}, N_\mathrm{h})$ (where N_a and N_h represent the number of antennas in the code and the number of helpers, respectively), which is by definition less than N_h preventing full-diversity.

The idea of using random choices is attractive, but the problem with antenna selection is the one-to-one mapping between relays and code columns. By rewriting the equivalent channel with an arbitrary transmitted matrix,

$$\mathbf{y} = \mathbf{S}\mathbf{h} + \mathbf{z}, \tag{2.14}$$

each column $\mathbf{s}_i$ corresponds to the signal transmitted by one relay but its design is not limited to $\mathbf{S} = \mathcal{G}(\mathbf{x})$. Instead, the symbols transmitted by each relay can be modified by rewriting $\mathbf{S} = \mathcal{G}(\mathbf{x}) \times \mathbf{C}$, where $\mathbf{C} \in \mathcal{M}^{N_\mathrm{a} \times N_\mathrm{h}}$ is a combination matrix that builds the transmission of the relays from the rows of $\mathcal{G}(\mathbf{x})$.

With this change the relay transmission and the code columns can be coupled in more elaborate manners than one-to-one when the i-th row of $\mathbf{C}$ has several non-zero elements. The i-th relay transmits a linear combination of signals of multiple code columns. This may be implemented as in Fig. 2.17 where after reception ($\mathbf{x}$) and codification ($\mathcal{G}(\mathbf{x})$), the whole STC signal for all N antennae is computed and the actual antenna is fed with a linear combination of them all ($\mathbf{s}_i = \mathcal{G}(\mathbf{x})\mathbf{c}_i$). This model allows gain tunning and/or antenna selection by simple adjustments in $\mathbf{C}$. Summing up, any arbitrary combination of antennae at the relay is possible, yielding the same properties as any local STC of choice except for a random degradation of the channel that we can parametrize to limit its effect, $\mathbf{y} = \mathcal{G}(\mathbf{x})\mathbf{h}' + \mathbf{z}, \mathbf{h}' = \mathbf{C}\mathbf{h}$.

Sirkeci-Mergen and Scaglione analyzed error probability and $\mathbf{C}$ matrix design criteria to achieve full diversity order [126], and indicated that the maximum achievable diversity order is the rank of the joint matrix, $\mathcal{G}(\mathbf{x})C$. This showed that $\mathbf{C}$ matrix design criteria are equivalent to classic full-rank criteria for space-time block codes applied to the joint matrix, achieving full diversity order $\min(N_\mathrm{h}, N_\mathrm{a})$. Yiu et al proposed that each node should have a *signature vector* $\mathbf{c_i}$ to define its unique combination of matrix columns [131]. Randomized DSTBCs (RDSTBC) with random $\mathbf{C}$ matrices were proposed in

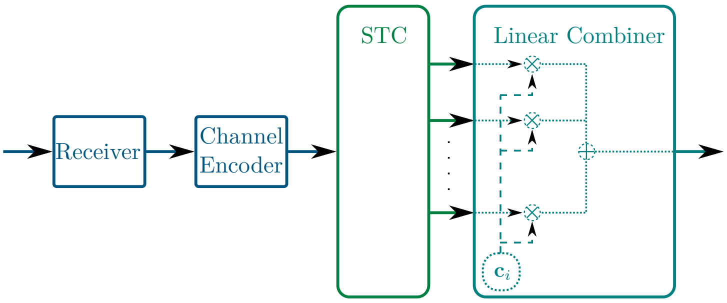


FIGURE 2.17: Architecture of a STC random combination relay.

[126] and [132], such that no signaling at all is required. Moreover, Sirkeci-Mergen and Scaglione also considered different random distributions [126] and together with Sharp they extended the system to account for asynchronous transmission in [132].

2.3.6 Open Issues

It is possible to construct PHY schemes that achieve in practice the benefits predicted by information theoretic results. Some techniques are completely novel, while others are based on preexistent signaling techniques. A few of them provide benefits with very simple implementations (best relay, binary NC), but most of them rely on elaborate encoding strategies to increase the benefits even further.

The relations between the different implementation schemes in the PHY layer are not clear because there is a known bound on diversity gain: the number of antennae or helpers in the system. Therefore, it is not clear how the combination of several of these techniques aggregates their benefits. Hybrid systems and comparisons between different PHY techniques must be taken into account when choosing a solution. In simultaneous transmissions such as CFNC or STC, the problem of imperfect synchronization between the nodes must be addressed either by robust signals or by synchronization measures.

In the *Best-Relay* approach some helpers may not need to participate. *Code Cooperation* shows that the two messages may result from operations more elaborate than a mere replication. *NC* solutions show that some extra benefit may be obtained if the relays include their own information, although the MAC layer becomes more complex. Finally, *Space-Time Coding* is the way classic MIMO systems transmit simultaneous replicas of the same signal. There is apparently no reason for these techniques to be mutually exclusive (say, for instance, we can combine OR and STC to propose a "Best STC Relay Group" scheme). It is also important to stress that some proposals depend directly on current PHY techniques for point-to point transmission that will continue to evolve, opening the opportunity to import those advancements to cooperation.

2.4 MAC Layer

Both telecommunications operators and end-users would reject cooperation in their wireless networks if they have to negotiate the terms every time the PHY employs cooperative diversity. Therefore, the role of the MAC layer to detect opportunities to cooperate, negotiate the necessary parameters and exchange control messages between the users is essential. In addition to cooperation control, the upper layers support other services such as error recovery, dynamic resource allocation optimization, mobility control and user attachment to the network.

As far as the design of traditional MAC protocols is concerned, the PHY is a lower-level service to stablish point-to-point links between two nodes that can be requested upon will. However, in cooperative diversity MAC, more than two nodes may take part in communication and this view is insufficient. Helpers must be recruited and negotiated, and the operating parameters of the cooperative PHY finely tuned to perform cooperation. As a result, cross-layer design is a must and there is a circular feedback loop in which the PHY provides the MAC with data delivery services and the MAC serves the PHY with radio resource control services.

2.4.1 Protocols Studied

We begin by listing an abundant sample of cooperative MAC protocols in the literature, omitting the details of their implementation. Instead, in section 2.4.2 we present functional decomposition of the MAC design problem in general, and discuss how each functional trait is implemented in all these protocols. In other words, our functional analysis allows a classification of the protocols in the literature or rapidly design the characteristics of new ones.

The following list must be interpreted only as illustrative and not an exhaustive survey.

2.4.1.1 CMAC

Chou and Ghosh proposed CMAC in [133]. It seeks the integration of cooperative diversity with extended IEEE 802.11g wireless local area networks. They worked on the assumption that helpers were already assigned and designed MAC layer signaling to coordinate the sources and the helpers. They also proposed an extension of the protocol for multiple relays named FCMAC. In it, messages are split in several blocks, one per helper. Each block is protected separately with an independent Reed-Solomon FEC and relayed by a different helper.

2.4.1.2 C-MAC

The scheme in [134], as others below, is based on the Distributed Coordination Function (DCF) of the IEEE 802.11 standard [105]. In this mechanism, originally devised to mitigate the "hidden node" problem, a source that sees the medium as available sends a small Request To Send (RTS) control packet to start the transmission, instead of transmitting the message straightaway. If the destination is not seeing the medium as available like the source did (meaning that there is a node transmitting, hidden to the source), then the RTS triggers a collision that lasts only for a minimal time. Otherwise, the destination sends a CTS control packet and the source starts transmitting the proper data packets.

C-MAC is a design of a cooperative MAC for ad-hoc networks with different nodes transmitting simultaneously without interference using Code Division Multiple Access (CDMA). It focuses on the support of multi-hop with energy savings through the selection of the best routes depending on instantaneous channel states, thus exploiting opportunistic cooperative diversity environments, estimating angles of arrival from the received signals to route packets in the (approximate) direction of neighbors that are likely to be close (if their angles of arrival are similar).

2.4.1.3 Relay-enabled DCF (rDCF)

Zhu and Cao [135] developed a *triangular handshake* mechanism that is an extension of DCF that adds helper messages. This allows to coordinate the communications between the source, the helper and the destination. Fig. 2.18 shows the messages exchanged. The source sends first a Relay Request To Send (RRTS) packet, which is received by the helper and the destination, and used to measure the channels that separate them from the source. The relay piggybacks the result of this measurement on a second RRTS packet, which also allows the destination to measure the H-D channel and discover the optimal strategy. The destination chooses either the direct route or the cooperative route and the rates that can be supported in each link. Finally, the destination encodes the optimal strategy in the Relay Clear To Send (RCTS) packet and sends it to the source. [136].

2.4.1.4 Opportunistic Relaying (OR)

The key idea of OR is to avoid the excessive complexity of STC and still attain the diversity of a multi-helper environment. For this purpose the technique is designed as a single-helper system where the best relay is selected [116]. It can be analytically

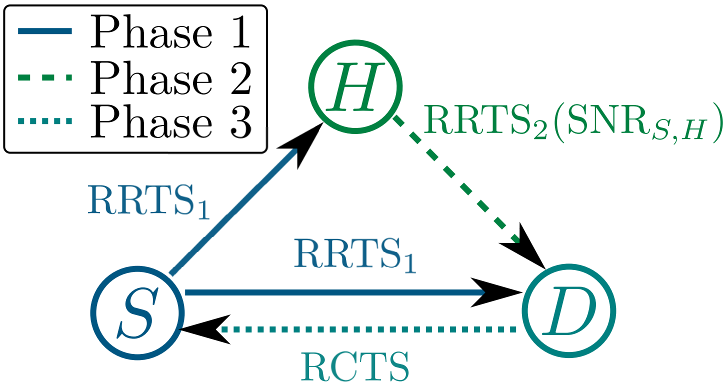


FIGURE 2.18: DCF-based triangular handshake.

proven that this guarantees full diversity-multiplexing trade-off. Fig. 2.9 illustrates the difference with STC.

The best helper is selected by using a DCF-based protocol in which all potential relays receive the RTS and compete for the medium. The medium contention is performed with a timer initialized with the inverse of the channel quality value. Therefore this timer expires first for the best next-hop, which wins access to the medium and transmits the CTS to the source. When a relay wins, it must forward the received information to the destination.

2.4.1.5 Power Aware Relay Selection (PARS)

Chen et al [137] studied power-aware relay selection strategies. By modifying the channel quality criteria of OR for relay selection, PARS selects relays using an *Optimal Power Allocation* (OPA) algorithm, which computes the channel access counters with a metric that combines achieved rate and power consumed. The source computes its own power cost function and competes with the relays, thus enabling the algorithm to choose whether to cooperate or not at the same time.

2.4.1.6 CD-MAC

Moh et al [138] seeked to improve link reliability, respecting the foundations of cooperative diversity theoretic models, rather than improving range or rate as the previously mentioned protocols. In their approach, also based on DCF, the hops in the direct route are protected by a DSTBC cooperative backup link each, which is only activated when the signal in that hop of the route is weak. The multi-hop route is assumed to be given and each step is independently protected against outage by cooperation. Fig. 2.19 illustrates the mechanism, where for the source (S) or intermediate node (I), in each step of the route, transmissions are overheard and stored by a neighboring helper. If it is necessary to introduce diversity due to poor link quality, the helper forwards the transmission to the next intermediate node or the destination (D).

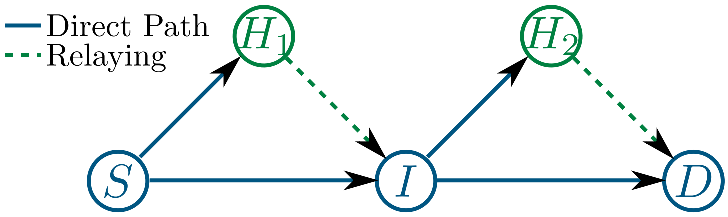


Figure 2.19: Cooperation in each link of a multi-hop route.

2.4.1.7 Cooperative Triple Busy Tone Multiple Access (CTBTMA)

In [139], Shan et al follow a radically different method for helper notification. In addition to exchanging DCF-type messages, they sense a number of dedicated tones for medium reservation. Their messages are based on DCF, but they add a "busy tone" signal that allows to reserve the medium by occupying it with a signal. The best helpers are selected, again, by contention, but in this case medium access is gained using *Helper Busy Tones* (HBT) as well. This type of reservation is inspired by non-cooperative busy-tone MAC equivalents such as [140].

2.4.1.8 Phoenix

This is the first protocol with NC in this list. It integrates NC in cooperative CSMA networks [141]. An Automatic Repeat reQuest (ARQ) mechanism that reacts to transmission errors was developed, called Cooperative CSMA. When a source fails, the destination transmits a negative acknowledgement (NACK) control packet so that the error is notified and the transmission repeated. If there is a potential helper neighbor that has overheard the failed transmission, has decoded correctly it, and has its own data to transmit to the destination; that neighbor may detect the NACK and respond to it, offering to perform the retransmission in the place of the source. A DCF triple-handshake-like mechanism permits to choose whether to accept the offer or not. If accepted, the neighbor transmits a packet with a network-coded XOR of the repeated source message and its own. The destination uses the previously stored message, despite its errors, as side information to revert the NC and decode the new packet. Then, the new packet is subtracted from the network coded reception and the distorted instances of the source packet combined for diversity decoding.

Table 2.4: Different coopMAC versions.

Ref.	Network	Cooperation	Control
[142]	WiFi	Best Relay	DCF-based
[142]	WiFi	Best Relay	DCF-retro-compatible
[102]	WiFi	Best Relay	DCF-based
[102]	WiFi	Source combining	DCF-based
[144]	WiFi	Source combining	DCF-based
[145]	WiFi	RDSTBC	DCF-based
[146]	WiFi	RDSTBC	DCF-based
[149]	WiMAX	RDSTBC	Modified 802.16-j OFDMA
[147]	WiFi	RDSTBC	DCF-based
[148]	WiFi	RDSTBC	DCF-based

2.4.1.9 CoopMAC family

By CoopMAC family we refer to a series of papers by related authors that improved a protocol with that name throughout the years [102, 107, 142–148]. CoopMAC is based, again, on IEEE 802.11 DCF, and its adaptation to centralized-scheduling-based IEEE 802.16 is called CoopMAX [149]. Chronologically, all proposals but CoopMAX consist of incremental improvements.

The initial CoopMAC was a pure DCF-based selection betwen one-hop or two-hop routing towards the wireless AP. To decide, the source considers the Adaptive Modulation and Coding (AMC) that can be used reliably in each possible link. Since the same data is delivered, the route that achieves a better effective end-to-end rate is that with the shortest total transmission time, defined for direct and cooperative routes as

$$\begin{aligned} T_{direct} &= \tau_d + \frac{L}{R_{sd}} \\ T_{coop} &= \tau_c + \frac{L}{R_{sr}} + \frac{L}{R_{rd}} \end{aligned}$$

In these equations, the minimum of the two transmission modes is chosen, τ_x is the initialization time of mode x, L is the packet length and R_y is the transmission rate of hop y. Fig. 2.20 illustrates the concept: since the helper node is closer to the source, it can receive the same packet much faster and forward it again also at a high rate; hence if the two rates are greater than twice the rate of the direct link, which is the minimum AMC mode at long distances, $2R_{min}$, then the end-to-end rate is better in cooperative mode.

We list all the variants in table 2.4. The improvements include the simplification of the control messages to make them fully backwards-compatible with standard devices, the introduction of signal combining between the source and helper transmissions, and multiple-relay transmissions using RDSTBC.

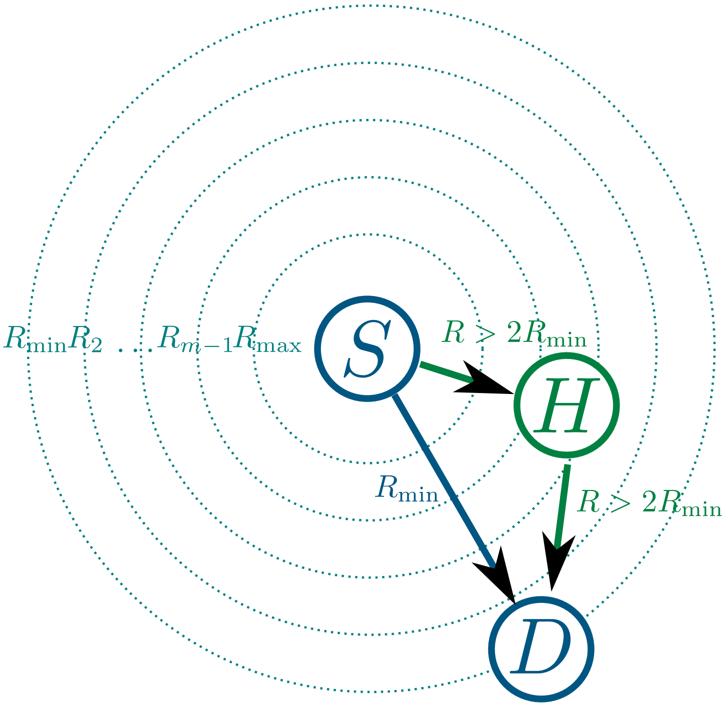


FIGURE 2.20: AMC rate concentric ranges and 2-hop exploitation.

2.4.1.10 Multi Hop Aware Cooperative Relaying (MHA-Coop-Relaying)

Adam et al [150] focused on the fact that a simple substitution of each direct link along a multi-hop path with a cooperative-diversity transmission, as in Fig. 2.19, only constructs a sequence of pairs of alternate SIMO and MISO transmissions, instead of providing a route with full cooperative MIMO transmissions in each stage. As a solution, their protocol differentiates two types of helpers: normal ones, which are only suitable for providing cooperative help once in a particular step along the route, and those that, in addition to the current hop, are also close to the destination of the following step along the route and can help in two consecutive steps. The relay selection mechanism, with contention, is modified to give a priority bonus to that second type of double helpers. When these are selected, in addition to cooperating in a particular hop in the path, the helpers that are close to their intermediate nodes keep enhancing the following hops along the route as illustrated in Fig. 2.21.

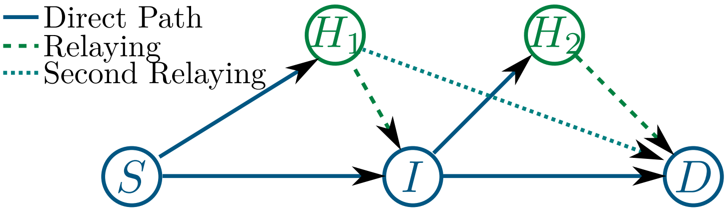


FIGURE 2.21: MHA enhancement of all links in a multi-hop route.

2.4.1.11 Distributed Cooperative MAC for Multi-hop Wireless Networks (DCMAC)

Shan et al [103] modified the best-relay selection contention mechanism in OR to incorporate the possibility of AMC helpers. This supports multi-rate best-relay selection.

2.4.1.12 FairMAC

Bocherer and Mathar [151] expressed their concern about the energy cost of cooperation. When a helper allows to achieve a higher throughput, the energy per transmitted bit increases. There is a trade-off between energy per transmitted bit and cooperative capacity. The proposal of Bocherer and Mathar, FairMAC, permits to select a cooperation factor $\alpha \in (0, 1)$ representing the ratio between assisted packets and source-transmitter packets. For $\alpha = 1$, all source packets receive help (for $\alpha = 0$, no source packets receive help). The ratio α is reached using transmitted and helped packet counters combined with triple-handshake mechanisms to notify the source whether cooperation will be employed or not.

2.4.2 Services Required for Cooperation

All the protocols in subsection 2.4.1 have diverse implementations with benefits and drawbacks. None of the solutions is completely superior in all scenarios. However, is is possible to define a basic set of five functionalities that the MAC layer must provide as a service for cooperative communications to be realized effectively. In this section, we provide a general view of those functionalities. They are represented in Fig. 2.22. Using them as defining traits it is possible to define a classification for the myriad of MAC protocols available in the literature. In addition, the functionality set provides a framework for the rapid characterization of new protocols.

Specifically, we present a classification of the protocols listed above comparing their implementations with the functionalities in Fig. 2.22. Tables 2.5 to 2.9 detail how each protocol implements each cooperation aspect.

2.4.2.1 Neighborhood Mapping

Neighborhood mapping is the service that provides cooperative MAC protocols with an image of their surroundings. It is typically implemented through a table of known neighbor nodes. The first important decision is the structure of this neighbor map. For

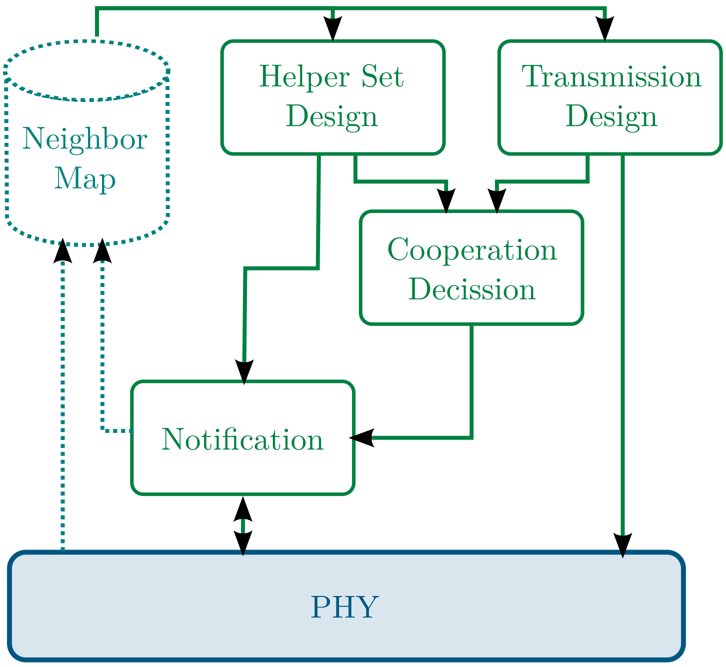


FIGURE 2.22: Main functionalities of a cooperative MAC protocol.

example, in some cases it contains metrics of estimated link qualities and cooperation capabilities.

Once the map is designed, it is necessary to define a process to add entries into it. There are many approaches with different levels of overhead. The most simplistic and passive approach is to annotate the existence of a neighbor when its transmissions are detected. The most costly and aggressive approach would be to poll actively the surroundings until all the neighbors have been detected. An intermediate approach would seek a compromise between polling overhead and probability of undetected neighbors. Table 2.5 lists the approaches in the protocols we have reviewed.

In some cooperative protocols, it is also necessary to exchange neighbor mapping information with other nodes in order to discover the tables of the neighbors and design the cooperative strategy. For example, in [135], the helper creates and distributes a *willing list*: a list of the source-destination pairs that it can assist.

Unfortunately, even if a node is sensed as being able to cooperate, it might refuse to do so. The proposal in [143] assumes optimistically that all known neighbors are collaborative and takes relaying faults as packet losses. Instead, an attribute on willingness to cooperate may me added to the map. It can be measured using announcement protocols, credit systems, game theory, etc.

Finally, the precision of the map will be affected by the aging of its entries, as pointed out in [146]. Cooperative diversity is devised to combat time variant fading, and the

Table 2.5: Cooparative MAC protocols implementation of Meighborhood Mapping.

Protocol	Neighborhood Mapping
CMAC	Not necessary
FCMAC	Not necessary
rDCF	Passive listening in helpers, active distribution of willing lists
C-MAC	Regular transmission of “hello” packets and estimation of angular position
OR	Not necessary
PARS	Not necessary
CD-MAC	Passive listening at the source
CTBTMA	Not necessary
Phoenix	Not necessary
coopMAC-I	Passive listening at the source
coopMAC-II	Passive listening at the source
c-coopMAC	Passive listening at the source
RcoopMAC	Passive listening at the source
coopMAX	Passive listening at the source with optional pilot signals for measurement
MHA-CR	Passive listening at the source
fairMAC	Passive listening at the source with pending ACK counter
DC-MAC	Not necessary

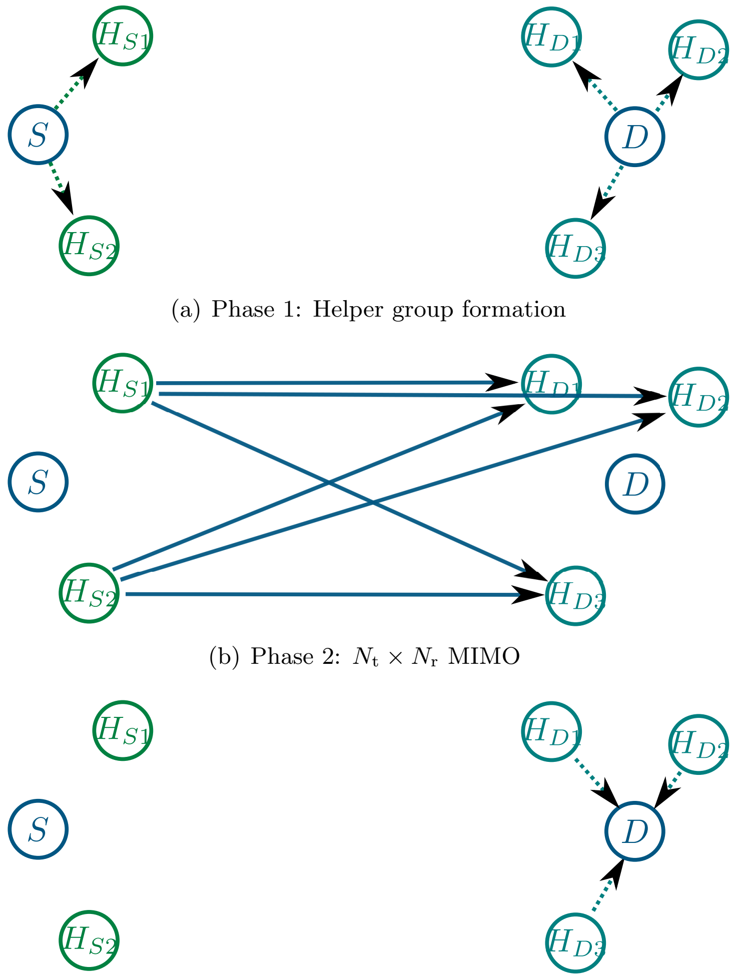
possibility to reach other nodes in the network is random. Bearing this in mind, neighbor table entries must be discarded as they get old and the neighbor discovery mechanism needs to track changes in network topology and channel fading. If fading is too variant or the node traffic patterns are too bursty, the passive hearing schemes will be ineffective. Conversely, if changes are slow, aggressive frequent polling strategies would result in unnecessary overhead.

Errors in the neighbor map imply differences between the map and the actual network, either caused by network changes after mapping of sensing errors. Thus any good design must include, besides of accurate computation of mapping parameters, some form of time-stamp to collect outdated table entries as garbage. Notification feedback must also be considered, to remove or correct table entries when a node fails to cooperate.

2.4.2.2 Helper Set Design

The network map lists the neighbors that can be individual helpers, but it does not specify which or how many of them should participate in the cooperative transmission for the best effect. The information in the neighbor map must be processed into a unique choice of the best helper set to reach each destination. There may be several options for cooperative signaling in the PHY, so the structure and size of the best helper set depends heavily on the performance criteria and the PHY layer. In this regard, best-relay systems order the list of helpers by link quality or any other performance metric, while systems with any form of signal combination must simultaneously consider the combination technique and the set of contributors in order to maximize performance. Table 2.6 shows how helpers are selected and grouped in the literature we have reviewed.

There are many alternative performance metrics for performance optimization of helper sets: average information rate [152], transmission time [143], covered distance [153], error probability [148], power consumption [154][151], power saving [137], battery lifespan [137], etc. Moreover, it is possible to design hybrid performance metrics that allow simultaneous optimization over several of these domains.



(a) Phase 1: Helper group formation

(b) Phase 2: $N_{\mathrm{t}} \times N_{\mathrm{r}}$ MIMO

(c) Phase 3: Data collection

FIGURE 2.23: A three-hop cooperative system with two helper sets.

In any case, there is no reason to specify a single helper set for each cooperative transmission. Let us consider the 3-hop proposal with a *transmitter helper group* and a *receiver helper group* of Fig. 2.23, used in [153]. First, both the source and the destination contact their own groups of local helpers to establish distributed transmission and reception.

Then, a full MIMO transmission is performed, and finally the destination collects data from the distributed reception group. This saves energy by shortening the ranges of the first and third hops, which have a lower diversity and improve transmission range to a lesser extent for a given increment in transmission power.

The result in [63] on scaling laws has shown that the protocols that combine multi-hop routing and cooperative MIMO techniques achieve the best throughput capacity scaling in ad-hoc wireless networks. More precisely, wireless networks are traditionally classified as *extensive* or *dense*. The main difference is that, in dense networks, communications are limited by interference and available degrees of freedom, while in extensive networks communications are limited by noise and available power. In the first case, hierarchical cooperative MIMO techniques are the best option, whereas in the latter multi-hop is recomendable. According to the authors there is a range of intermediate operating regimes between the two extremes, and in these intermediate regimes the optimal policy consists in dividing the network into cooperative MIMO clusters, cooperating locally and performing multi-hop routing between clusters globally. Consequently, the longer the distance in hops towards the destination, the more set-to-set hops need to be considered. Fig. 2.24 illustrastes and idealized path where each node is aided by a helper and intermediate nodes rely on their helpers both for distributed reception and transmission.

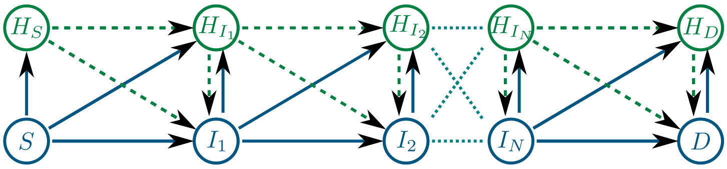


FIGURE 2.24: An ideal example of cooperative multi-hop route with diversity order $d = 2$ in all steps.

In general, the interaction between helper group formation and multi-hop switching for ad-hoc networks with cooperative diversity is an open research field [100]. The example in [152] is very illustrative but it relies on an excessively regular mesh topology. In [155] and [150], classical routing is used, and the hops are enhanced a posteriori. Fig. 2.25 illustrates how, in a pre-defined non-cooperative path, each intermediate node could be assisted by helpers (not shown in the figure) to skip a few steps along the route and reach some hop further ahead than the immediate non-cooperative hop. For this, every node I knows the full list of steps and it is allowed to replace direct forwarding to the next hop with cooperative transmission towards I nodes further ahead (or even to the final destination) in order to simplify the path. These are very good as initial approaches, but they do not guarantee an optimal solution.

Errors may occur when the designed set fails to assemble itself to form the virtual array, for example due to transmission errors in the first hop. Due to the strong dependence

Table 2.6: Cooparative MAC protocols implementation of Helper Set Design.

Protocol	Helper Set Design
CMAC	One random helper selected by contention
FCMAC	Random helper set selected by contention
rDCF	One optimal helper selected by the source
C-MAC	Optimal CDMA helper set selected by the source
OR	One optimal helper selected by contention
PARS	One optimal helper selected by contention
CD-MAC	One optimal helper selected by the source
CTBTMA	One optimal helper selected by contention
Phoenix	One optimal helper selected by contention
coopMAC-I	One optimal helper selected by the source
coopMAC-II	One optimal helper selected by the source
c-coopMAC	One optimal helper selected by the source
RcoopMAC	Optimal helper set selected by the source
coopMAX	Optimal helper set selected by the source
MHA-CR	One optimal helper selected by the source considering multi-hop double relaying.
fairMAC	One optimal helper selected by the source
DC-MAC	One optimal helper selected by contention

on the PHY layer of this function they should be designed jointly. For example, the coded cooperation PHY in [117] requires that, if the helper fails to decode the first part of the FEC, then the source should be notified and complete the remainder of the coded cooperation transmission by itself. On the other hand, RDSTBCs handle helper failures at the cost of some diversity loss, but they do not need of any recovery mechanism [146]. Note that, in principle, helper outage handling is embeded only in the PHY layer, but this is not true because simplified PHY-independent approaches also exist. For example, a process restart in case of failure might allow to reduce complexity.

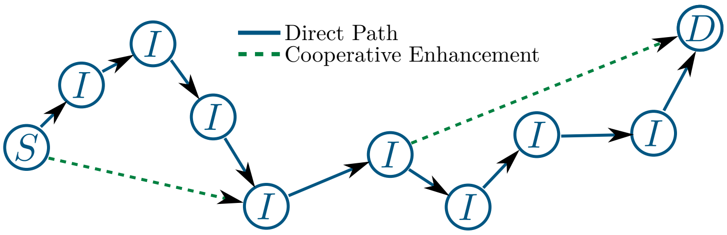


Figure 2.25: Cooperative enhancement of multi-hop routes.

2.4.2.3 Cooperation Analysis and Decision

It is necessary to evaluate the advantages and disadvantages of performing a cooperative transmission over a non-cooperative one. Clear decision schemes must define metrics for the *proficiency* and *cost* of each option. Unlike problem of helper set design, where the use of cooperation is assumed and the helper set is optimized, this function selects cooperation itself and the most convenient technique (OR, NC, STC, etc...) for it. It is possible to make the most informed decisions by a priori designing the optimal helper sets each type of cooperation technique would use. Or it is possible to reduce the workload of the helper set selection function by making a less-informed selection of the technique first, and then designing the helper set a posteriori just for the chosen technique (except for the case of the direct transmission, where it is unnecessary).

In addition, in [103] we can observe the need to distinguish between *proactive* and *reactive* cooperation decisions. The first starts at the point when a source wants to start a transmission, and decides in advance whether cooperation shall be taken into account. This gives the advantage of allowing the first source transmission to be encoded with cooparation-specific coding techniques. Reactive cooperation, on the other hand, always starts with a direct source transmission and cooperation is triggered only in case of primary link failure. This has the advantage of reduced cooperation overhead in scenarios where cooperation is seldom neccessary.

Unfortunately the design of a fair metric to compare dissimilar transmission mechanisms is difficult. For instance, STC may improve the rate of a single user compared to NC, but the latter may improve the rate of other users at the same time, yielding a higher network aggregate rate. If a MAC layer that can exploit the information theoretic limits of the wireless network is desired, a key factor would be the design of fair metrics to compare among many dissimilar PHY techniques.

As cooperative diversity was born to fight fading, a first performance metric would likely be outage probability. Fading analysis would require a good knowledge about the geometry of the environment, which would still be very expensive in a general architecture. A centralized topology with a complex BS is more favorable for outage analysis. Furthermore, a simplified network that is well conditioned for analytic treatment of outage, as in [41], represents a reasonably optimistic starting point to validate designs and further investigation.

A realistic protocol with limited resources cannot afford a complete analysis of outage probability every single time a transmission is started. Specially in scenarios with aggressively varying topologies. Practical MAC layers must proceed with limited knowledge. For this reason, performance metrics should be as generalist as possible to accept any available data. There are examples that employ expected packet loss [149], number of routes [152], achievable rate [144][138], elapsed time [143][107], and so on.

Cooperation analysis fails when the estimated performance becomes unsustainable. A helper might be unable to sustain the compromised behavior, even after being recognized as cooperator and receiving the data correctly in the first hop, due to changes in its

Table 2.7: Cooparative MAC protocols implementation of Cooperation Decision.

Protocol	Decision	Pro/reactive	Metric
CMAC	Cooperative retransmission	Reactive	Direct link failure
FCMAC	Cooperative retransmission	Reactive	Direct link failure
rDCF	Source informed choice	Proactive	Heuristic credit system
C-MAC	Iterative increments	Proactive	*Cooperation gain*
OR	Assumed in hypothesis	N/A	N/A
PARS	Access contention	Proactive	Power Cost
CD-MAC	Cooperative retransmission	Reactive	Direct link failure
CTBTMA	Relay offer announcement + Access contention	Proactive	Best Route
Phoenix	NACK	Reactive	Nest gain with NC
coopMAC-I	Source informed choice	Proactive	Min. transmission time
coopMAC-II	Source informed choice	Proactive	Min. transmission time
c-coopMAC	Source informed choice	Proactive	Min. transmission time
RcoopMAC	Source informed choice	Proactive	Min. transmission time
coopMAX	Source informed choice	Proactive	Min. transmission time
MHA-CR	Cooperative retransmission + Multi-hop-Aware	Reactive	Direct link failure
fairMAC	Source informed choice	Proactive	Min. transmission time + pending ACK counter
DC-MAC	Helper offers (hi packet)	Proactive	Best Helper

environment. In this case it should be necessary to decide wether continuing with the service even if it is less proficient than expected or sending a failure notification to the source to force a reinitialization. Analysis mishaps that lead to resource under-exploitation, like discarding a potential helper that later turns to be better, might also be covered by control mechanisms. For example, discarding a node that is able to cooperate might trigger a more aggressive announcement process.

Table 2.7 explains how, when and why the protocols we have reviewed switch from direct transmission to cooperative diversity.

2.4.2.4 Cooperator Notification and Agreement

The helpers need to be notified of the cooperative aspects of the transmission and the receiver should be aware of them as well. The notification is even more important when the cooperative mechanism requires the exchange of initialization values. Choosing the helper set without information on helper willingness is troublesome. In that case, either helper ACK is required [139] or helper cooperation is assumed to exist [146]. The latter is in some cases overoptimistic, unless MAC cooperation is mandated by standard. When the helpers do not return acknowledgements as a response to cooperation requests, the protocol must define the appropriate response policies.

Table 2.8: Cooparative MAC protocols implementation of Notification and Agreement.

Protocol	Helper Notification
CMAC	No ACK
FCMAC	No ACK / Reception of NACK
rDCF	DCF based triangular handshake
C-MAC	Multiple explicit messages based on DCF
OR	DCF with SNR-proportional contention
PARS	DCF with power-proportional contention
CD-MAC	Helper ID in data packets. C-DCF
CTBTMA	DCF with busy tone MAC and helper contention
Phoenix	NACK retransmission, DCF handshakes
coopMAC-I	DCF with ACK, helper indication on RTS
coopMAC-II	DCF without ACK, helper ID in data packets
c-coopMAC	DCF with ACK, helper indication on RTS
RcoopMAC	DCF without ACK, opportunistic, multiple helpers
coopMAX	Helper announcement/allocation to/by BS
MHA-CR	Helper ID in data packets. C-DCF
fairMAC	DCF with a pre-ACK message when packets are accepted and joint-ACK when delivered
DC-MAC	DCF with MCS-proportional contention

It is also possible to rely on the receiver to perform the agreement. This forces the necessity to choose whether the transmitter, the helpers, or both of them must send notifications to the receiver, and, if the receiver gives some feedback, which of them should receive and process it.

Additionally, it is necessary to decide which node is responsible of the final decisions in case of an anomaly (helpers that do not cooperate, lack of response, etc.). There are examples of protocols where those decisions are taken by the source [143], the receiver [136], the helper [135], etc. Some approaches divide the responsibility in multiple aspects between different peers, so that the best informed peer tackles each aspect [144].

Since the notification service might carry information for the other services, and even exception notifications are possible, error recovery is a critical component of this function. A suitable scheme of ACKs and timers should be designed to provide a sequence of actions leading to a backup strategy or a controlled stop (and possible retry) of the process.

Table 2.8 shows the notification mechanisms of the protocols in our review.

Table 2.9: Cooparative MAC protocols implementation of Transmission Design.

Protocol	Transmission Design
CMAC	Plain relaying
FCMAC	Plain relaying
rDCF	Plain relaying
C-MAC	Simultaneous CDMA relaying
OR	Not specified
PARS	Not specified
CD-MAC	OSTBC of source and relay
CTBTMA	Plain relaying with AMC boosting
Phoenix	CFNC and plain relaying
coopMAC-I	Plain relaying with AMC boosting
coopMAC-II	Plain relaying with AMC boosting
c-coopMAC	Source combination and AMC boosting
RcoopMAC	RDSTBC relaying with AMC boosting
coopMAX	RDSTBC relaying with AMC boosting
MHA-CR	DSTBC with multi-hop double relaying
fairMAC	Plain relaying during fraction of time
DC-MAC	Plain relaying with AMC boosting

2.4.2.5 Cooperative Transmission Design

The most versatile PHY mechanisms can adapt themselves to varying conditions such as the number of relays available, the transmission time, the channel coding, MCS, etc. The MAC layer should be capable of selecting these parameters to maximize the chances of meeting the system requirements. Of course, a one-size-fits-all solution, if it existed, would avoid the complexity of dynamic adaptation of the PHY configuration. Nevertheless, for strong adaptiveness and exhaustive exploitation of capacity limits, some form of on-the-fly design or selection of the transmission parameters becomes necessary. Table 2.9 summarizes the different approaches for cooperative transmission in the protocols we have reviewed.

2.4.3 Comments on Security and Fairness

Zhu and Cao [135] anticipated security issues. Assuming that helper nodes are in fact other users, malicious relays might steal, modify or forge messages from honest users. Such attacks are possible in all communication systems; hence cryptography, authentication and integrity checking techniques are extended in them. Even though cooperative diversity is a new playing ground for attackers, well-known protection techniques already in use should be effective. On the other hand, cooperative diversity services could be subject to some new forms of malice, in which greedy users would distort the coordination mechanism to their own benefit. These new security menaces need to be identified and counteracted. For example two new problems are considered in [135]:

- In a context where helpers annunciate their willingness to cooperate, attackers may maliciously report improved cooperation metrics to attract all the traffic to them, making it possible to perform attacks on said traffic.
- A malicious relay would intentionally drop the data received to perform a Denial of Service attack, and even forge the acknowledgements to make the loss unnoticeable.

Regarding fairness, adequate incentives are necessary for the network users to share their resources willingly. For each niche we discussed in the introduction, the scenario is slightly different. Ad-hoc dense sensor networks and rapid-deployment wireless networks can simply assume cooperation is natural since all nodes in the network are deployed by a single organization that seeks to maximize social benefit. For example, if deployed in a humanitarian catastrophe mission all nodes in the network would be controlled by the civil authorities in charge. The same holds for an industrial dense wireless sensor network where all nodes belong to the same company.

In cellular data networks, mobile devices are independent. However, they are controlled by few service providers, which improve their networks according to cost incentives. Those providers could offer by contract some cooperation incentive to end users by means of well designed policies of requirements and rewards.

Finally, in a completely arbitrary ad-hoc wireless data network, the problem of cooperation incentives is hard to solve, since the users are not grouped by providers or subject to service contracts. As a consequence, cooperation should be imposed by communication standards with fairness-forcing mechanisms, or by attractive end services.

2.4.4 On Interference in Cellular Networks

Surprisingly, the simulation results in [102] and [143] for coopMAC protocols reported that interfering signals are lower in cooperative scenarios. This contradicts the intuition that the interference area is extended by cooperation (Fig. 2.1). The explnation lies in the influence of the time dimension, and particularly the scheduling of different active links in different instants of time. Even though it is true that helper support does extend the area subject to interference, it also reduces interference duration into said area. In addition, the relay and the source do not transmit simultaneously, so each half of the interference influence area in Fig. 2.1 is only subject to interference half of the time. Therefore, there is no increase of the interference time-area product, just a *redistribution*.

In addition, many WiFi and cellular protocols are only concerned with inter-cell interference. In this type of interference, transmitters close to the edge of the cell experience most of the problem. In this case, by replacing direct transmissions by two-hops with similar time duration, one of them being close to the center of the cell, cooperation reduces effectively interference time at the cell edge, and increments it at the cell center (where it is easier to mitigate).

This reasoning could be extended to any other protocol with few modifications. It has to be considered that, if the nodes were allowed to transmit simultaneously, their interfered area-to-time product would indeed grow and, as a result, interfering power would effectively increase. Nevertheless, two transmissions provide the destination with twice as much energy to overcome interference and, consequently, the comparison would be unfair. We should perform interference measurements using SINR metrics instead of plain energy, time and area metrics. With this, the effect of interference extension would be approximately cancelled by the power increase and the basic results of coopMAC might be extensible to any other MAC protocol.

2.4.5 Open Issues

The MAC layer is of the utmost importance for feasible cooperative communications, as they allow to identify alternative ways of transmission within a networked context. In other words, the advantages of cooperative transmissions are only possible if the MAC layer is able to efficiently trace, classify and coordinate helpers at reasonable cost. For this purpose, most protocols in the literature rely on small control messages, such as the DCF of 802.11, or on a central controller, such as that of 802.16.

There is no clear winner among the studied protocols, because their features are better or worse depending on the application domain. In this regard, we consider more interesting to focus on five fundamental functions that cover the design of a cooperative MAC: *neighborhood mapping*, *helper set design*, *analysis and decision*, *notification and agreement*, and *transmission design*.

The MAC layer should manage some side effects that may arise from cooperative communications: High energy expenditure of nodes in “good” positions (since many peers use them as relays), new security concerns, and interference redistribution across cells.

Most MAC protocols achieve their goals by relying on previously existing protocols and tailoring them to cooperative diversity. Thus, it may be interesting to design new protocols aiming primarily at accomplishing theoretical bounds rather than at retro-compatibility.

Energy expenditure deserves a deeper analysis as the results of [151] show a trade-off between energy per bit and throughput increase, whereas the results in [143], taking into account idle energy consumption, show potential power savings due to a reduction of idle waiting time.

Integration with other network aspects, like routing, must be extended, as for example most models require source-destination signaling, preventing transmission towards nodes beyond direct range that could be reached by the cooperative signal. In general, routing and forwarding are heavily affected because node reachability depends on the cooperation environment.

Another interesting field is the effect of network saturation as most models assume that the incoming flow of packets is unlimited and therefore any acceleration of the delivery process would increase throughput. Therefore, analysis of relaxed networks with plenty of resources should be investigated to determine how cooperation affects jitter, delay, etc.

2.5 Summary

Cooperative communications rely on cooperative diversity to reach the diversity degree of MIMO systems. However, unlike the latter, they must rely on smaller, single-antenna networked nodes. Based on recent analyses, the relation between transmission improvement and cooperation is evident and profitable even for sources that are conveniently placed for direct transmission. These nodes benefit from the faster average medium release time of their inefficient neighbors. The overall increase of throughput has a cost in energy that the peers should consider to cooperate.

In this chapter we have reviewed the information theoretical models that support cooperative diversity, the PHY techniques that can make these transmissions possible, and the MAC protocol techniques that enable the network to set up cooperation.

However, the information theoretical models are far from complete. Capacity is still unknown for many types of channels and further work on them is expected. In addition, research in theoretical relaying functions must continue to provide new perspectives.

The different existing PHY layer technologies offer a wealth of possible implementations. Some of these are mutually exclusive and cannot be employed at the same time. Future developments must take into account the evolution of the off-the-shelf non-cooperative transmission mechanisms PHY laters re based on, to absorb future enhancements. Simultaneous utilization of several PHY solutions in hybrid schemes must be evaluated to investigate how their gains combine when deployed cumulatively. The advent of completely new PHY technologies for cooperative diversity should not be discarded either.

The review of the MAC layer has also revealed the many proposals in the literature, especially in the form of cooperative add-ons over preexistent MAC networks. These represent excellent proofs of concept and offer effective short-term implementation methods, and therefore work in this direction is of great practical interest. However, more challenging research should be pursued, comparing multiple PHY support or even protocols that would switch PHY techniques if a globally optimal solution in time is not found. Integration with other fields of wireless technologies must also be considered.

For future MAC research, we have presented a functional decomposition that allows a complete characterization of the design problem and can also be used to create a taxonomy of available protocols.

Finally, cooperative communications are affected by the existence of too many entry points to the research problem. This is a direct consequence of cross-layer design: once layer frontiers are removed almost any preexistent layer-specific field of study, such as NC or STC, deserves discussion. A large space of possibilities is never a drawback by itself, but researchers on different knowledge fields are usually oblivious to the progress of one another. Again, let us take CFNC and RDSTBC as an example. It is obvious that these two apparently separated formulations have converged to increasingly similar conclusions and, if the research communities involved in them do not interact sufficiently, they are likely to waste resources in "reinventing the wheel". In the close future, convergence on cooperative transmission and a consensus on a common language for the topic are necessary. A consensus on a simple subset of good-for-now solutions to transfer to industry partners would also be welcome in order to prevent the new paradigm to drown in a sea of possibilities.

The content in this chapter is an extended and actualized version of a paper published in IEEE Communications Surveys & Tuturials [8].

Chapter 3

Current Cooperative Communications: Practical Multi-hop Relaying in LTE-A

Contents

3.1 Introduction

3GPP LTE-A aims at meeting the requirements published by International Telecommunications Union (ITU) for 4G standards, International Mobile Telecommunications Advanced (IMT-A) [156]. The Advanced version of LTE introduces several improvements in the standard to increase rate and meet these requirements. One of them is Carrier Aggregation (CA), consisting on the union of non-contiguous spectrum subbands to form a heterogeneous system with higher bandwidth. Another is Multiuser MIMO (MU-MIMO), in which the BS processes jointly the signals transmitted to/from multiple single-antenna users. The new standard also gives specifications for femtocells: small domestic APs designed for providing short-range telephone service using a home Internet connection as back-haul.

The introduction of RNs [157] in the standard, and the definition of their operation, has been conditioned by the context above. This makes possible to perform two-hop communications, but strongly limited by hardware constraints. Although this is still a timid first approach, the introduction of multi-hop communications in cellular networks [1] opens the possibility to analyze cooperation in the architecture.

For backwards compatibilty with UEs that were not prepared to deal with multi-hop transmission, LTE-A RNs are required to act like eNBs (or femtos) from the point of view of the UEs [17]. The eNB of the cell where the RN is located, called the Donor eNB (DeNB), is aware of the relay and behaves as a proxy forwarding to it all the control information that typical eNB interfaces receive from the rest of the network infrastructure. This proxying, as well as UE traffic forwarding, is tunneled through a physical wireless DeNB-RN connection implemented using the normal LTE-A eNB-UE channels. This requires the RN to behave temporarily as a UE from the point of view of the DeNB. Combining the required "perspective" from UEs and DeNBs, functionally, a RN is little more than a gateway device with a UE interface, a eNB interface, and the requirement to activate exclusively one of those two at any give time. Hereafter, we will call the RN-DeNB link the **relay** link; **access** links are those between UEs and RNs and **direct** links those between real UEs and DeNBs.

LTE-A networks high-performance features include intelligent UE scheduling as a function of instantaneous channel states; multi-antenna techniques; interference management; and admission control [156, 158]. As shown in this chapter, the introduction of relays may alter these features. In the literature, RN location optimization, which may

[1] Considering multiple hops in the same level of network hierarchy. There have been previous cases of single-hop cells with an out-of-band wireless backhaul that could be called "multi-hop wireless" in some sense, but in reality they did never allow to perform any multi-hop transmission in cellular bands towards UEs.

be limited by planning constraints to mitigate conflicts, is the common approach, but random location of RN or femtocells has been sometimes considered [27–29, 159].

There are conflicts related to interference management, scheduling and half-duplex relay operation that are largely ignored by LTE RN literature. This is probably due to the frequent flexibility assumptions in theoretical studies on relaying/cooperation, which are not achievable by the standard, or to the influence of single-hop cellular research. The latter focuses on low-level properties such as the SINR, since in single hop links it is sufficient to determine user rate, which may lead to the naive assumption that there is the same one-to-one relation in a multi-hop network.

In this chapter we analyze diverse conflicts that result from the inmature deployment of RN technology in LTE-A networks and propose solutions to mitigate them. These conflicts have been identified by implementing standard-compliant relay functionalities on top of the well-known Vienna LTE System Level Simulator [33]. As an original experimental approach, we model realistic RN operation, taking into account that transmitters are not active all the time due to half duplex RN operation. Implementing RN firmware is too complex, but, without loss of realism, we have rearranged existing elements (eNB, UE, scheduler) in a manner that -seen as a black box- emulates the behavior of a relay, rather than implementing a standalone brand-new component for the simulator. We also study analytically some of the issues observed in the interaction between the network and the RNs, to draw conclusions beyond simulation observation.

The main issues treated in this chapter are:

i) Additional time-varying interference management steps are needed. The LTE-A standard employs a fixed time division between eNB-RN and RN-UE transmissions that is global for all cells. On the opposite, in typical relay capacity or throughput research models these two phases are optimally balanced for each user. Thus, this granularity is impossible in LTE-A.

ii) There is a trade-off between the time-division constraints of relaying and multi-user diversity. The stricter the constraints on relay scheduling are, the less flexible schedulers are to exploit channel variation. This is because the schedulers should be reallocating user's transmission to the best realizations of their channels, but this becomes unfeasible if such allocations violate the relaying timing requirements.

iii) The standard contains a variety of parameters for relaying frame configuration, but not all cases of interest suggested by theory are covered.

iv) Even though literature either considers optimal relay location or random relay distributions, we introduce an intermediate approach based on admission control. This improves performance on the random-location approach by removing sites with a negative influence, while still allowing to model networks where relay positions cannot be fully controlled by operators.

The rest of this chapter is organized as follows: Section 3.2 describes relaying in LTE-A Release 10. Section 3.3 describes our model and the configuration of the Vienna simulator to implement it. Section 3.4 describes the problems with interference and our proposal to minimize their effects on relaying. Section 3.5 describes how scheduling is hampered by relaying time constraints. We study this problem analytically, discuss the effect of all relevant parameters involved, and suggest configuration values for these parameters to mitigate the problems. Section 3.6 describes our relay admission control algorithm, which can be used to discard relays that do not produce any gains. Finally, section 3.7 concludes the chapter.

3.2 Implementation of Relaying in the LTE-A Standard

LTE-A relays behave like any other eNB [17]. The DeNB is aware of relay presence and provides proxy functionality to its `X2` and `S1` interfaces towards the rest of the E-UTRAN. In addition, RNs have the `S1` interface –common to all dNBs– renamed `S11`, and the DeNB has a dedicated control interface `Un` for the specific management of attached RNs (Fig. 3.1). The RN is an Open Systems Interconnection (OSI) layer 3 device that performs networking and packet forwarding, just like IP. The RN features two physical interfaces, with UE and eNB functionalities, in the relay and access links, respectively [17, 160]. Information is forwarded using IP tunneling, as discussed in [161], where the LTE-A relays behave as IP “bridges”, and queue IP packets independently and transfer them between the UEs and the DeNB. RNs become attached to a DeNB using the `relay_attachment_procedure` [160].

LTE uses a centralized MAC protocol with 10ms *frames* and an Orthogonal Frequency Division Multiplexing (OFDM) / Multiple Access (OFDMA) PHY [162]. Each frame is divided in 10 *subframes* of 1ms. Each subframe consists on the transmission of 2 OFDM symbols ($T_s = 0.5$ms), and each symbol contains a number of carriers N_c that is a multiple of six. Each set of 6 contiguous subcarriers during two consecutive symbols on a subframe is called a Resource Block (RB). In TDD, some subframes are marked for DL and some for UL, and one special transition subframe must be inserted in between. In FDD, all subframes have a DL subset of RBs and an UL subset of RBs. In each subframe, and within each cell, the Radio Resource Control (RRC) of the eNB assigns different RBs to its different users. Network configuration pre-determines the DL or UL allocation of the RBs and the RRC cannot choose it.

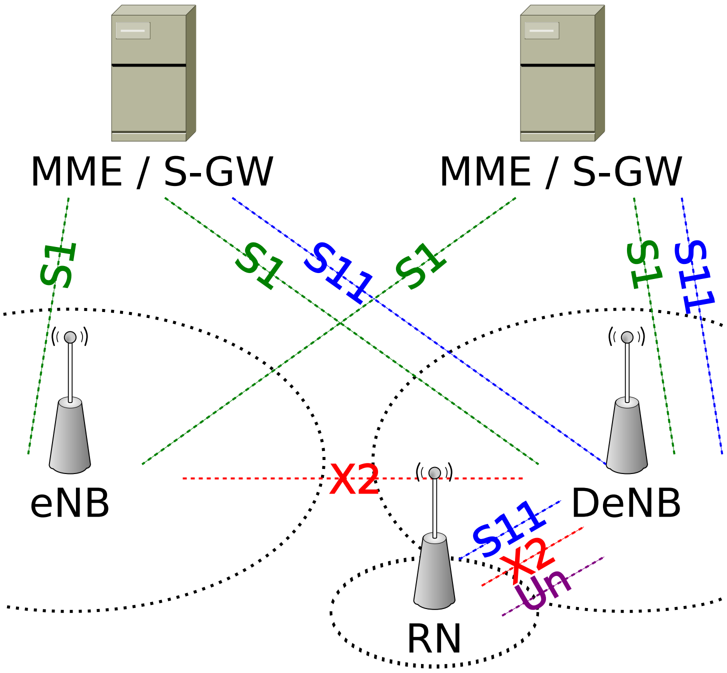


Figure 3.1: Relay architecture in LTE-A R10 (from [16, Fig. 4.7.2-1])

In some subframes, the RN, rather than behave as an eNB and allocate RBs to its UE (access link), is required to behave as a UE to which the DeNB allocates RBs (relay link). When the RN is in UE-like mode, the centralized MAC protocol of its eNB-like interface marks subframes with the Mobile Broadcast Multicast Services (MBMS) label for its associate UEs to ignore those subframes. This depends on configuration parameters `SubframeConfigurationFDD` and `SubframeConfigurationTDD`, for the different multiplexing modes. The FDD parameter is a binary mask that can mark any subframe, yielding eight possible resource partitions of relay and access links: $1/7, 2/6, , 3/5 \dots 6/2, 7/1$ with UL/DL symmetry. The TDD parameter allows the setups listed in [17, table 5.2-2].

Figure 3.2 illustrates the LTE MAC and PHY, the TDD special subframes, and the configuration of a subset of subframes for RNs to behave as UEs. In table 3.1 we list the total number of UL and DL subframes of each TDD configuration and the number of subframes allocated to RN-DeNB communications.

LTE relay performance has been assessed by several authors [27–31]. For example, Saleh et al [27][28] studied RN and Pico-eNB throughput gains for the worst 10th percentile of LTE-A network users, with a fairly realistic simulation setup. They considered a hexagonal lattice with three 120°sectors per eNB, helped by 5-12 small cells. From the resulting iso-performance curves (number of eNB per km^2 vs. number of small cells per km^2 with the same performance), they concluded that the relays must be cheaper than 1/30 of the cost of the eNBs for the approach to be more advantageous than a mere increase of eNB density.

Table 3.1: Number of LTE-A TDD subframes dedicated to direct and relay links (from [17, table 5.2-2]).

TDDD subframe config.	Uplink-DL setup	DL subframes	DL RN subframes	UL subframes	UL RN subframes
0	1	4	1	4	1
1	1	4	1	4	1
2	1	4	2	4	1
3	1	4	2	4	1
4	1	4	2	4	2
5	2	6	1	2	1
6	2	6	1	2	1
7	2	6	2	2	1
8	2	6	2	2	1
9	2	6	3	2	1
10	2	6	3	2	1
11	3	6	2	3	1
12	3	6	3	3	1
13	4	7	1	2	1
14	4	7	2	2	1
15	4	7	2	2	1
16	4	7	3	2	1
17	4	7	4	2	1
18	6	3	1	5	1

Even though some studies have dealt with Amplify-and-Forward relaying [31], the most frequent relaying model is non-orthogonal Decode and Forward. Within this model, the DeNB transmits directly to the UE in the RN-UE phase. We also consider this system, and, from the three 3GPP relay types in [29], *Type 1* inband relays, without antenna isolation between relay and access links.

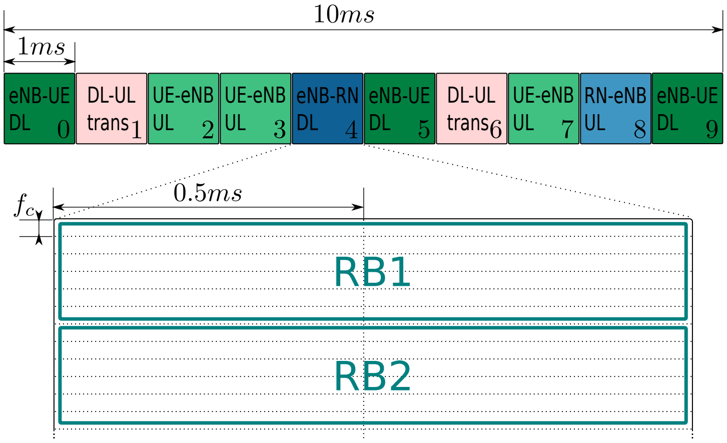


Figure 3.2: LTE-A TDD subframe with RN-DeNB communications following `Subframe Configuration TDD`= 1

A major characteristic of all these works is that RN communications are modeled according to theoretical characterizations rather than to the implementation above. For instance, most works use accurate 3GPP propagation models to compute the SINR of all links, and LTE-fitted Shannon-like mappings between SINR and rate [32] to compute the spectral efficiency ρ_x of each link x, but these results fail to model the correspondence between link rates and delivered end-user rate, due to the assumption that the resources dedicated to each link are allocated optimally per user.

Typically, a 1/2 time ratio between the relay link and the access link is not optimal in a two-hop scheme. For instance, if the relay link is better than the access link, it is beneficial to balance link traffic by allocating more time to the weakest ones, so that the RN does not have to drop packets that the DeNB delivers correctly, thus wasting resources. When the spectral efficiencies of relay and access links are known $(\rho_\mathrm{r}, \rho_\mathrm{a})$, and the resources are dedicated to the relay link for a fraction of time α, the end-to-end spectral efficiency is:

$$\rho_\mathrm{e2e} = \min(\alpha\rho_\mathrm{r}, (1-\alpha)\rho_\mathrm{a}). \tag{3.1}$$

And the time division is [28, 29]:

$$\alpha^* = \frac{\rho_\mathrm{a}}{\rho_\mathrm{r} + \rho_\mathrm{a}} \tag{3.2}$$

so that the end-to-end spectral efficiency is maximized at

$$\rho^*_\mathrm{e2e} = \frac{\rho_\mathrm{r}\rho_\mathrm{a}}{\rho_\mathrm{r} + \rho_\mathrm{a}} \tag{3.3}$$

However, the standard defines a fixed time-sharing framework that is applied to the whole system, the time-sharing α fraction must be a unique common parameter for all DeNB-RN-UE paths instead of taking arbitrary values for different user-specific paths, and the results in previous works are overoptimistic, as they ignore implementation constraints.

3.3 Simulation of the Relaying System

The Vienna LTE DL System Level Simulator is a software that calculates the main MAC and RRC operations for a sequence of LTE TDD subframes (referred as Transmission Time Intervals, TTIs). It defines scheduling granularity by explicitly computing RB allocation, whereas the granularity of traffic delivery/error measurement is defined by Transport Blocks (TBs) [158]. A TB is a packet of upper-layer data generated by a traffic generator entity at the source (eNB in DL) and received by the traffic sink at the destination (UE). The simulator implements different data structures to model schedulers, eNBs, UEs, path-loss models of terrain, etc. The model supports femtocell simulation but relay simulation was unsupported at the time this thesis was written.

The scheme we have developed is aimed at replicating at system level the behavior of LTE-A networks with relays, in order to study how interference, scheduling and data delivery are affected by the addition of these relays, with the realistic (novel) approach that some transmitters are inactive part of the time due to half-duplex RN operation, which is a global network parameter. As previously stated, we avoided implementing complex internal characteristics of RN firmware by rearranging existing elements of the simulator (eNB, UE, scheduler), in a manner that -seen as a black box- emulates how relays handle traffic. The alternative would have been to implement a standalone brand-new component but this would in turn have required to duplicate many lines of code already written in eNBs or UEs anyway (for transmission and reception purposes respectively).

Figure 3.3 illustrates the philosophy of the RN setup design. For each RN, we created a virtual eNB to act as transmitter in the access link. Normal UEs are created by the simulator and attached to the nearby eNBs using the default simulator procedures, which include virtual eNB connections if applicable. Finally, each RN takes each UE attached to its virtual eNB (relay UEs) and creates a clone of that UE (cloned UE), located at the same position of the RN and attached to its DeNB.

The DeNB delivers traffic both to its real UEs (direct UE) on the direct link (in all subframes) and to cloned ones on the relay link. Instead of acting as sinks, cloned UEs forward the data they receive to the output traffic source buffer of the virtual eNB of their RN. When the RN is supposed to act as UE (relay phase), the DeNB scheduler can select both direct and cloned UEs and the virtual eNB transmits empty frames. When the RN is supposed to act as an eNB (access phase), the DeNB scheduler can only select direct UEs and the virtual eNB can schedule relay UEs, which receive data from the RN that was previously forwarded by their clones. Thus, outside the grey box in Fig 3.3, the network behaves exactly as if a true RN was implemented, and no modification of simulator core components is required. The novelty of our implementation lies in the cloned UEs and their traffic-forwarding behavior, and in restricting the UEs that the schedulers are allowed to serve at certain given times. The virtual eNB at the relay is implemented with the original femtocell model of the official simulator distribution. Appendix 3.A contains the details of the simulator implementation.

Figure 3.4 shows the effect of inserting 200 randomly located RNs in the simulation. Comparing Figs. 3.4(a) and 3.4(b), SINR values improve in the neighborhood of the

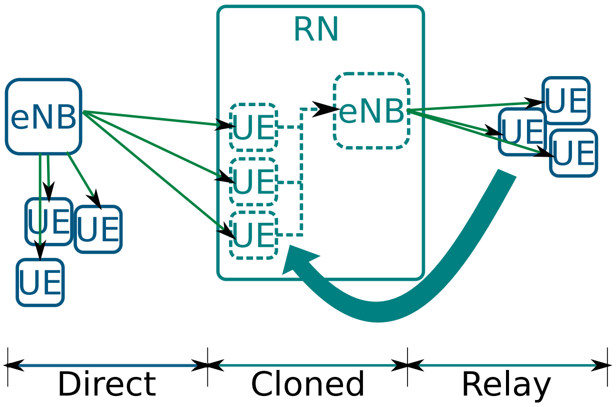


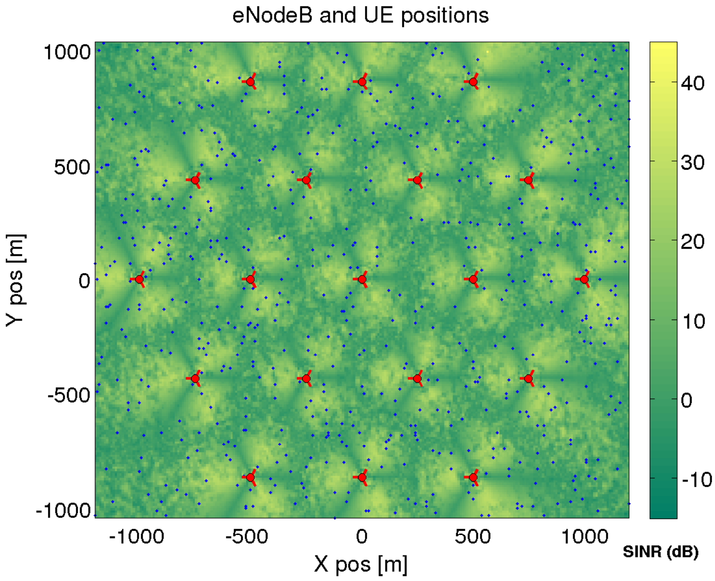
FIGURE 3.3: Design of a setup for the simulator that emulates the behavior of an RN using only existing components.

RNs but decrease around the eNBs. Average SINR was slightly worse in simulations with relays.

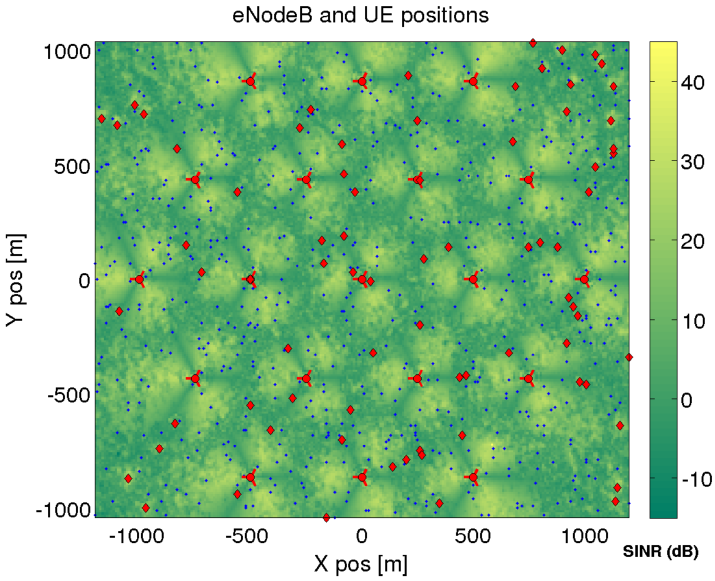
3.4 Interference and Half-Duplex Relay Operation

All cells in an TDD LTE network must be in the same phase at the same time, either DL or UL. Otherwise, if a eNB were receiving from its users while a neighboring eNB transmitted, the former would be unable to receive due to the great interfering power of the latter with the desired UE signals, because of the overwhelming power difference between UEs and eNBs. Following this same rationale, all RNs in the network are required to operate with the same time-division scheme as we discussed in section 3.2. This makes it impossible to achieve a correct balance of resources dedicated to relay and access links for all the RN-UE pairs at the same time. This decreases the effective spectral efficiency (3.1) and degrades relaying. Throughput balance is sacrificed to achieve a common timing across the network for the different operation phases to keep interference under control.

We have observed in our simulations that, when the network enters its relaying phase (subframes marked for RN connection with the DeNB), access links are transmission-free. From a radio perspective, at those moments the network should experience the same interference levels as if there were no relays at all (Fig. 3.4(a)). Next, when the network enters its access phase (subframes marked for the RNs to serve their users), even though DeNBs do not schedule relay links, they still transmit to their direct users with full power. In addition, RNs transmit to their UEs. From a radio perspective, this second phase has higher interfering power due to the presence of more transmitting entities in addition to all the previous ones (Fig, 3.4(b)).



(a) Without RN



(b) With RN

FIGURE 3.4: Comparison of simulation setups and SINR with and without RNs

The key question when introducing RNs in an LTE-A network is whether the rate improvement in UEs at bad locations compensates for this greater interference or not. We recall that the throughputs are imbalanced and therefore it is likely that (some) RNs will not have enough queued DL packets to transmit continuously during the entirety of the access phase. It is also likely that there will be less buffered data than the size in bits of all the RBs to be transmitted. Following a interpretation of the standard, the RN might fill the excess RBs with zeroes and continue transmitting, to maintain the timing of the MAC layer.

However, it is counterproductive to allow a transmission that increases interference and does not carry information. Note that `Subframe- ConfigurationTDD` in the standard mandates the number of subframes of the relaying phase (α), whereas the duration of the access phase is implied to be the rest ($1 - \alpha$). Thus, we implement a crude form of interference management by simply reducing the number of subframes that RNs actively use for the access phase ($< 1 - \alpha$) and introduce a third *direct phase* where RNs neither behave as UEs nor as eNBs. From the point of view of link balance, this represents a downgrading of the strongest link down to the rate of the weakest, which is fixed and can not be altered, as opposed to an optimal balancing which would require changing both link allocations. In the direct phase, the bonus subframes that result from reducing the access phase experience the interference levels of the scenario without RNs. The price RNs pay is that their schedulers have fewer access-phase subframes to allocate relay UEs in and to exploit multiuser-diversity with PF schedulers. We cover scheduling conflicts in more detail on section 3.5.

In our scheme, on each TTI the simulator calls the scheduler to manage a sub-list ofusers determined by the following types of subframes:

- *b-subframes:* All flows connected to the DeNB are subject to scheduling. This is the relaying phase and its duration **must** be implemented in compliance with the standard parameter.
- *u-subframes:* All the UEs, but not RN flows, are subject to scheduling. This is the behavior the subframes would have in the standard by default if they do not belong to the first type.
- *d-subframes:* This is the contribution in this thesis. These subframes are part of the access phase too but only direct UEs are passed to the scheduling routines. This type of subframe behavior in the access phase is neither mandated by the relaying part of the standard nor incompatible with it.

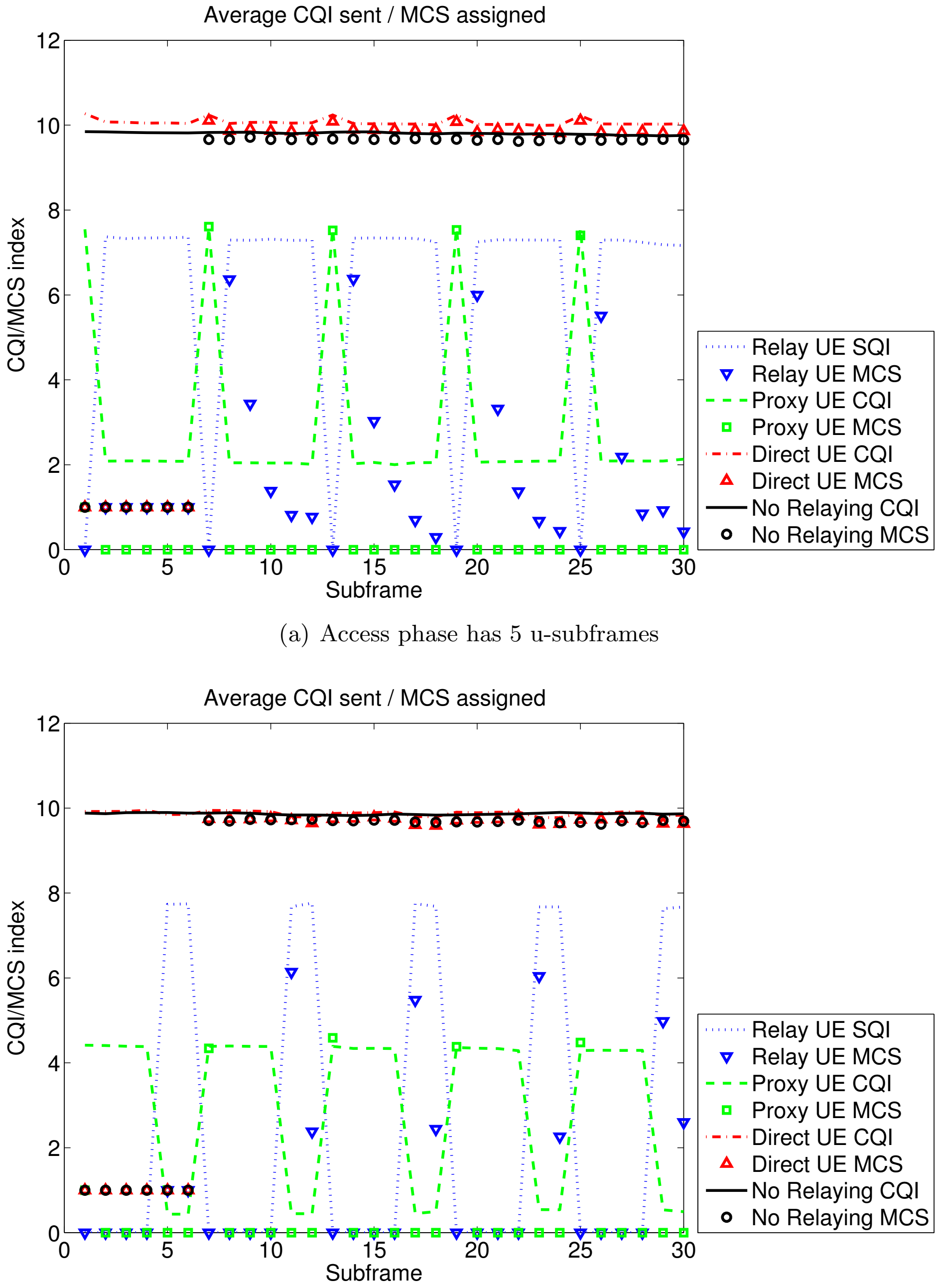
The *d-subframes*, where RNs do not transmit, are similar to the technique employed in the case of femtocells to reduce interference known as ABSF [35]. To implement our interference mitigation, we propose alternating u-subframes and b-subframes in the part of the frame that the standard leaves for RN to behave as eNBs (the access phase), balancing the number of frames of each type in search of good operation regimes with regard to interference and link balance.

Following the global subframe configuration that the standard imposes, the simulator will have a unique number of b-subframes for all cells in the network, unbalancing throughput. When there are more RBs in u-subframes than needed, some are left empty, but the receivers around the RN transmitting empty subframes are oblivious to this: as far as they know, the current subframe is an u-subframe and RNs are expected to be transmitting and increasing interference. In practice, RN interference expected by schedulers increases with the number of u-subframes even if the RNs have no data to deliver.

Therefore, we recommend adding a global parameter of number of RN-less d-subframes and maintaining a balance between the number of u- and d-subframes. For instance, by selecting `subframeConfigurationTDD`=6, according to the standard one out of six subframes is a b-subframe where the RN receives the DL from the DeNB. By omission, there would be five u-subframes where RNs transmit the DL to UEs, which is wasteful. Our simulation showed better average throughputs (with the same random seed and otherwise identical conditions) by letting RNs transmit only for 2 u-subframes followed by 3 d-subframes of RN silence. Note that the reason why d-subframes have to be introduced instead of b-subframes is that the standard does not allow to increase the number of b-subframes under the given configuration parameter.

In Fig 3.5 we illustrate with lines the Channel Quality Indicator (CQI) code that the UEs send to the eNB. This is a 4-bit number from 0 to 15 that quantifies the SINR and also indexes the MCS that can be used and the size in bits of the data transmitted per RB. When this number is small, the SINR is low and a more redundant channel code is applied per RB, which for a constant number of symbols per RB gives a proportional variation of the number of data bits transported. We represent with bullets the actual MCS that the eNB uses to transmit data in each frame. This means that the difference between the bullets and the lines is the amount of data that could have been delivered through the channel but was not sent because there were no packets in the transmit queue. If all RBs were fully occupied with data, the MCS should match the reported CSI, as it occurs in the simulation without relays (continuous black line) and also in the case of direct and proxy UE communications in the simulation with relays (dot-and-dashed red and dashed green lines, respectively). On the other hand, for relayed UE connections (dotted blue line) we see that there are subframes for which RN transmission queues are empty but RNs have still many open RBs to be allocated, so that actual data sent (bullets) fall below the level of potential data to be sent (line).

In Fig. 3.5(a) all subframes in the access phase are u-subframes. There is a periodical pattern of six subframes (`'buuuuu'`), starting with a relay phase with one b-subframe. Although relayed UEs report the same CQI throughout the access phase, data delivery is only scheduled in subframes 2 and 3, whereas in subframes 4 to 6 most RNs have already spent their buffered RN data. However, direct UEs continue reporting the same



(a) Access phase has 5 u-subframes

(b) Access phase has 3 d-subframes followed by 2 u-subframes

FIGURE 3.5: CQI in 5 simulation frames showing RN scheduling for interference mitigation (`SubframeConfigurationTDD`=6).

CQI throughout the entire access phase, which means that they experience interference by RN transmission while RNs are not actually delivering data.

In Fig. 3.5(b) we repeated the simulation removing the excess of RN-UE subframes (`'bddduu'`), allowing a single b-subframe followed by three d-subframes only with direct users, followed by two d-subframes with direct and relayed users. This maintains throughput and reduces interference.

3.5 Scheduling Gain Degradation with Static Relaying Intervals

3.5.1 Description

The LTE-A standard allows for UE traffic to be scheduled in time and frequency RBs, leaving scheduler implementation details to the vendors. There are different commercial products ranging from the simplest First-Come-First-Served (FCFS) schedulers to elaborate optimization algorithms to maximize the utility of specific UEs in specific RBs. Well-known scheduling algorithms featured in the Vienna simulator include Round-Robin (RR), maximal throughput, and PF scheduling.

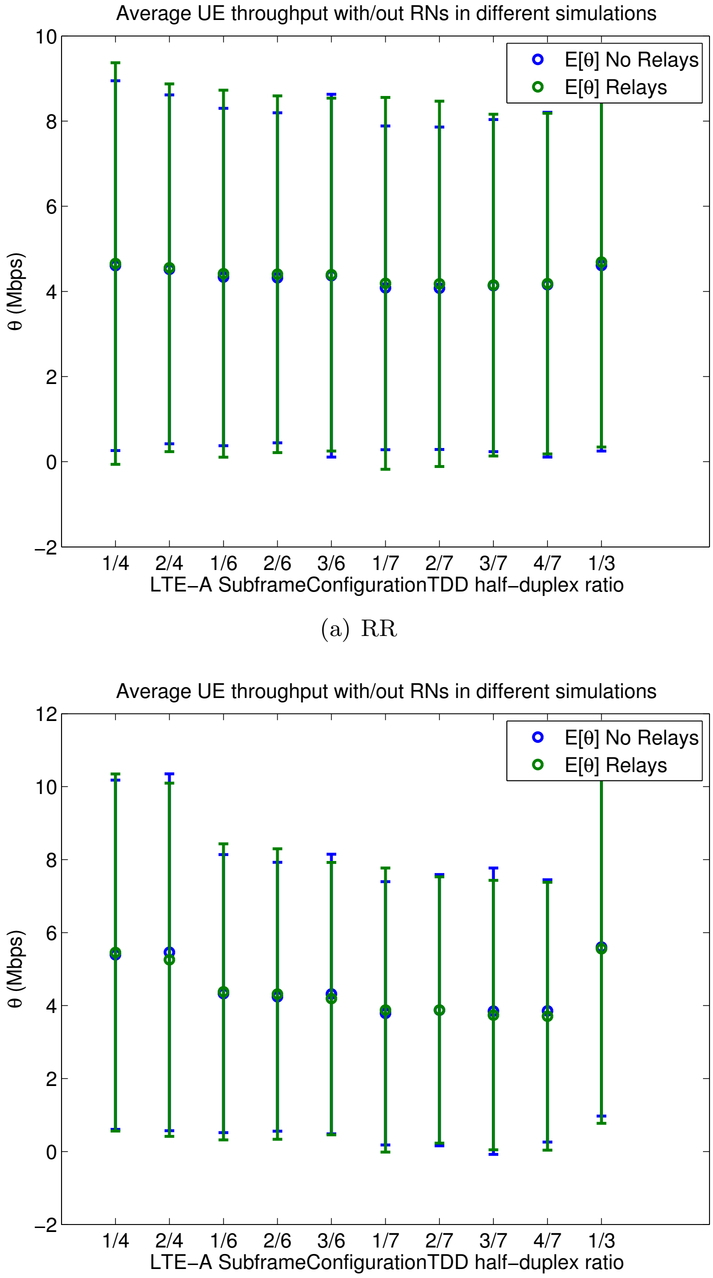
During simulation, we consistently observed that the performance improvements after adding relays were smaller with PF schedulers than with RR schedulers, and, in some cases, for PF schedulers and unconvenient values of `SubframeConfigurationTDD` adding extra RNs damaged UE average rates. Fig 3.6 shows the observed throughputs in many simulations, for all half-duplex factors (depending on `SubframeConfigurationTDD`), for the two most common schedulers, RR and PF. Even though the average rates for RNs using RR always improve a little in Fig. 3.6(a), the average rate improvement for PF is lower and sometimes it even decreases (cases 2/4 3/6 and 4/7), as shown in Fig. 3.6(b). In addition, the standard rate deviation with RNs augmented, meaning that the introduction of RNs with strict scheduling constraints leads to higher scheduler unfairness.

To interpret this, on the one hand we have that RR can be considered a statistically-neutral time multiplexing where the effect of adding a relay on the statistical distribution of the spectral channel efficiency of a single UE keeps unaltered for the proportional allocation of resources to that particular user by the scheduler. On the other hand, PF schedulers exploit multiuser diversity and, thus, the average user channel conditioned on its selection by the scheduler, is better than the unconditioned time-average of the same channel.

$$\mathrm{E}\left[|h_i|^2\right]|_{i\text{ scheduled by PF}} > \mathrm{E}\left[|h_i|^2\right] \tag{3.4}$$

However, when a user is attached to an RN, the scheduler cannot allocate relay link traffic arbitrarily any longer: it must take place in specific b-subframes of the LTE radio interfaces regardless of how a PF scheduler would have handled that traffic if it was free. In other words, half-duplex operation overrides scheduling decisions and causes a conflict between relaying and the ability of PF scheduling to exploit multiuser-diversity.

$$\mathrm{E}\left[|h_i|^2\right]|_{i\text{ scheduled by PF \& valid b-subframe}} < \mathrm{E}\left[|h_i|^2\right]|_{i\text{ scheduled by PF}} \tag{3.5}$$



(a) RR

(b) PF

Figure 3.6: Average throughput per user over 10 random topologies for 84 TTIs.

Figure 3.7 illustrates the conflict between PF scheduling and RN users due to the half-duplex relaying constraint for a two user scenario: a direct user (DU) and a relay user (RU). Given a direct UE and a RN, in a scenario in which the RN channel is better in the second subframe (rows 1 and 2), the ideal decision of an unconstrained PF scheduler would have been to serve the direct UE first (row 3). However, relaying constraints force this subframe to be occupied by the RN, so the constrained scheduler has to place the RN in a worse channel. A cascade effect moves the direct UE from its ideal channel to a worse allocation (row 4). Furthemore, the final allocation of the direct UE is *an* *even worse* channel, as the RN allocates the second hop of its traffic in the same RB, increasing interference (row 5).

It is erroneous to assume that these effects are reversed when rows 1 and 2 (channel qualities) are exchanged. In that case, it is true that both UEs stay in the RB that would be optimal without RNs. But this only means that the constrained PF has a certain probability, say p, of matching the unconstrained case. And for probability $1-p$ it is certainly worse. Therefore, since the instantaneous rate is the same with probability p and worse with probability $1-p$, the result is on average worse for any $p > 0$.

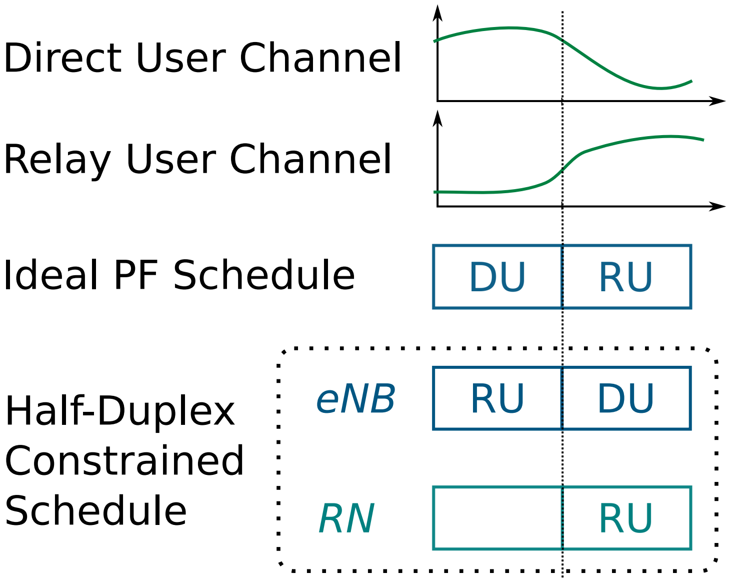


FIGURE 3.7: Example of PF scheduler and RN conflict.

3.5.2 Scheduler Analysis with Relays

Since PF scheduling is well studied throughout the literature, it is possible to leave the trivial example above and introduce a complete analytical study of the previous problem. The general convergence of PF schedulers was analyzed in [163]. For a single cell scheduler, the main result is that, for each user i, average throughputs $\overline{\theta}_i$ satisfy the first order Ordinary Differential Equation (ODE) system

$$\begin{array}{ccccc} \dot{\overline{\theta}}_1 & = & \mathrm{E}_{t:s[t]=1}\left[\theta_1[t]\right] & - & \overline{\theta}_1 \\ \dot{\overline{\theta}}_2 & = & \mathrm{E}_{t:s[t]=2}\left[\theta_2[t]\right] & - & \overline{\theta}_2 \\ \vdots & & \vdots & & \vdots \\ \dot{\overline{\theta}}_n & = & \mathrm{E}_{t:s[t]=n}\left[\theta_n[t]\right] & - & \overline{\theta}_n \end{array} \tag{3.6}$$

where $s[t] \in 1 \ldots n = \arg_i \max \frac{\theta_i[t]}{\overline{\theta}_i}$ is the proportional fair scheduling function that selects the user with the highest ratio between its instantaneously achievable throughput and its average.

Assuming that all UE channels experience a Rayleigh fading distribution h_i with parameter λ_i and that capacity is, in the low-SNR regime, approximately linear with SNR, the i-th user under PF achieves the average throughput

$$\begin{aligned} \overline{\theta}_{i,PF} &\simeq \overline{h}_i \\ &= \mathrm{E}_{h_i[t]/\overline{h}_i > h_j[t]/\overline{h}_j \forall j} \left[h_i[t]\right] \\ &= \frac{\sum_{j=1}^{n} \frac{1}{j}}{n} \frac{1}{\lambda_i}, \end{aligned} \tag{3.7}$$

where the first term represents a multiuser-diversity gain versus the rate for RR, $\overline{\theta}_{i,RR} = \frac{1}{\lambda_i}$. The integral to compute this average is formulated in appendix 3.B.1.

We provide a modified solution to (3.6) for the case where some users are relays and may not be scheduled in all time slots. We consider a single-cell DeNB with multiple RNs behaving as UEs and a PF scheduler operating over a total of $\tau_\mathrm{r} + \tau_\mathrm{a}$ RBs per frame. Users 1 to n_d are direct users, and users $n_\mathrm{d} + 1$ to $n_\mathrm{d} + n_\mathrm{r}$ are relayed users. For the first τ_r RBs, direct and relay flows may be scheduled, and the RN flow channels are independent. In the remaining τ_a RBs, only direct flows can be scheduled. In addition, during τ_r and τ_a, direct users may experience different channels, modeled with two independent variables with averages $\lambda_{i,r}$ and $\lambda_{i,a}$, due to the activation of the part of the relay links where RNs behave as eNBs during the access phase.

To reutilize the analysis in [163], instead of writing the relaying constraints in the scheduling function, we introduce them in an *effective instantaneous throughput distribution* that is passed to a normal PF scheduler. Consequently, the instantaneous throughputs of the flows in the DeNB scheduler are:

$$\theta_i[t] = \begin{cases} h_i & \{i \leq n_\mathrm{d}\} \cup \{t \mod (\tau_\mathrm{r}+\tau_\mathrm{a}) < \tau_\mathrm{r}\} \\ 0 & \{i > n_\mathrm{d}\} \cap \{t \mod (\tau_\mathrm{r}+\tau_\mathrm{a}) \geq \tau_\mathrm{r}\} \end{cases} \tag{3.8}$$

In the relay phase, allocating a direct transmission only achieves the benefit of that particular data delivery, whereas allocating a relay transmission achieves the additional benefit of ensuring that the relay will have data to deliver in the following access phase. Therefore, we consider an *incentivized* PF scheduler as in [164], with an incentive parameter $\beta > 1$ to prioritize RNs in the scheduler to enforce fairness.

In the first fraction of the frame, with relative duration $\alpha = \frac{\tau_\mathrm{r}}{\tau_\mathrm{r}+\tau_\mathrm{a}}$, all users can be allocated, while in the second, with relative duration $1-\alpha$, only direct users can access the channel. Thus, taking the averages separately in these two fractions, we obtain an ODE that merges the original PF scheduler in the first fraction of the time, and a new PF subsystem managing only a subset of the users in the second fraction.

$$\overline{\theta}_i = \begin{cases} \alpha \mathrm{E}_{h_i[t]/\overline{\theta}_i > b_j h_j[t]/\overline{\theta}_j \forall j}\left[h_i[t]\right] + (1-\alpha)\mathrm{E}_{h_i[t]/\overline{\theta}_i > h_j[t]/\overline{\theta}_j \forall j \leq n_\mathrm{d}}\left[h_i[t]\right] & i \leq n_\mathrm{d} \\ \alpha \mathrm{E}_{h_i[t]/\overline{\theta}_i > b_j h_j[t]/\overline{\theta}_j \forall j}\left[h_i[t]\right] & i > n_\mathrm{d} \end{cases} \tag{3.9}$$

Where scheduling is biased towards RNs by defining the weight value

$$b_j = \begin{cases} 1 & j \leq n_\mathrm{d} \\ \beta & n_\mathrm{d} < j \end{cases} \tag{3.10}$$

Consequently, at the end of the frame, direct flows will be more likely to have gained access to the channel than relay flows, an effect that can be compensated with incentive β. The resulting ODE system, formulated in Appendix 3.B.2, has two sets of equations: direct flow throughputs receive two contributions, whereas relay throughputs receive only one contribution. These three contributions are balanced by the choice of β. Next, we will discuss the effect on the system for the two-user case, with one user of each UE class (relayed or not). The conclusions can be easily generalized for more users of both classes.

Figure 3.8 shows the analytical solution for a two node scheduler and its simulation, for $\beta = 1$ (no relay compensation incentives), $\alpha = 0.5$ (relayed and access flows are multiplexed with a ratio of 1/2) and $\lambda_{1,a} = \lambda_{1,r} = \lambda_2 = 1$ (all Rayleigh channels have unit mean). It is important to compare the pair of rates that the scheme achieves with a RN $(0.79, 0.44) \times \frac{1}{\lambda}$ with the rates of a two-user system without relays, $(0.75, 0.75) \times \frac{1}{\lambda}$. The difference means that the RN flow loses almost 50% of its rate in exchange for a meagre 5% increment in direct UE rate.

If there are multiple users in each category, their throughput fractions depend both on their individual channel distributions λ_i and the relaying scheme. For example, Fig 3.9 illustrates a simulation with ten users (six direct and four relayed flows) where all flow average rates can take one of two values $\lambda_{1,a} = \lambda_{1,r} = \lambda_2 = 1$ or $\lambda_{1,a} = \lambda_{1,r} = \lambda_2 = 4$. Basically, four groups appear: direct UEs with good channels, direct UEs with bad channels, RNs with good channels, and RNs with bad channels. By symmetry, the throughputs in each group converge to the same values. Again, note that the multiuser diversity gains of the RN groups are severely hampered while the gain of the direct groups improves slightly.

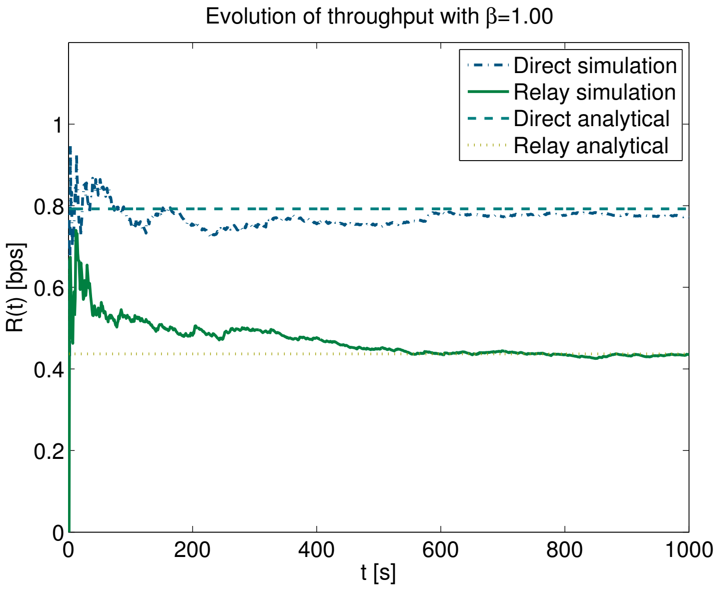


Figure 3.8: Evolution of the throughput of a two-user PF scheduler, with one user served through a relay.

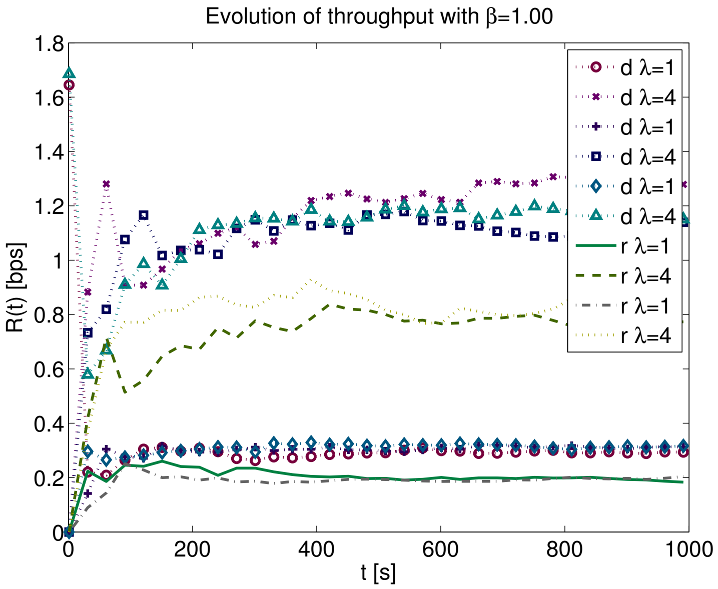


Figure 3.9: Evolution of the throughput of a ten-user PF scheduler, with four users served through a relay (simulation only).

We conclude that PF schedulers without incentives are biased towards direct flows. They hamper relayed flows heavily in exchange for small improvements in direct flows. The physical interpretation is that RNs are sometimes forced to renounce the most proficient transmission opportunities. On the opposite, direct UEs already gained their best channel realizations with a direct PF scheduler, and they can only gain access to worse resource blocks as a consequence of the addition of RNs, giving them poor gains that do not compensate, in terms of average user rate, the losses of the RNs.

3.5.3 Discussion of Balanced Parameters

3.5.3.1 Effect of an Incentive Parameter

We cannot correct the bias of the PF scheduler in the scenario with relays using an incentive. Figure 3.10 shows the evolution of the theoretical multiuser gain factors for the scenario with two users and one RN, as a function of the incentive β. The effect we observe is that, when we increase β, the rates converge to $(0.5, 0.5) \times \frac{1}{\lambda}$, which is merely the rate gain of the RR scheduler. Multi-user diversity gain is lost because it is increasingly likely that each user will be scheduled at fixed RBs, instead of dynamically searching for the best RB channel realizations.

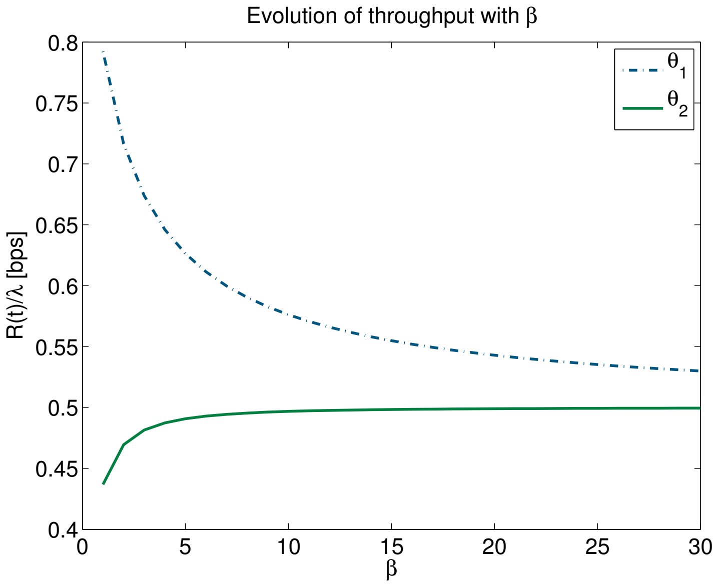


FIGURE 3.10: Evolution of the throughput of a two user PF scheduler, with one user served through a relay, as a function of β (analytical results).

The convergence to RR, however, is due to the presence of only two users. It is easy to generalize this conclusion to multiple users as follows:

Proposition 3.1. *As the incentive parameter β tends to infinity, a PF system with n_{d} direct flows and n_{r} relay flows converges to two independent PF systems that alternate in time: one with a PF scheduler with n_{r} users operating α of the time, and another one with a PF scheduler with n_{d} users operating $1-\alpha$ of the time.*

Therefore, even with incentives, the introduction of relays causes two types of losses: a multiuser diversity gain decrease from $n_{\mathrm{d}} + n_{\mathrm{r}}$ to $\min(n_{\mathrm{d}}, n_{\mathrm{r}})$, and a multiplexing loss from $\frac{1}{n_{\mathrm{d}}+n_{\mathrm{r}}}$ to $\min(\frac{\alpha}{n_{\mathrm{r}}}, n\frac{1-\alpha}{n_{\mathrm{d}}})$. We discuss this effect in Appendix 3.B.2.

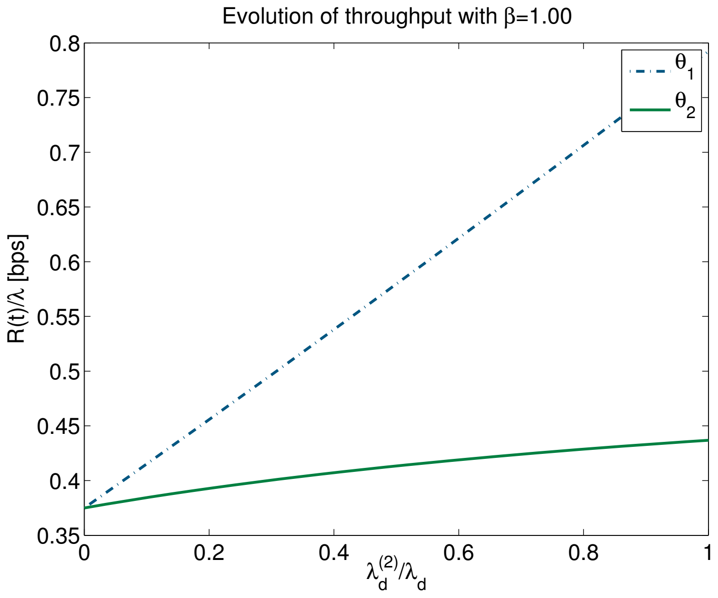


FIGURE 3.11: Evolution of the throughput of a two user PF scheduler, with one user served through a relay, as a function of $\frac{\lambda_{d,a}}{\lambda_{d,r}}$ (analytical results).

3.5.3.2 Effect of Interference

The throughputs of the relay scenario vary with the types of the users, but they are consistently proportional to user channel gains. In Fig. 3.11 we represent this relation for the two user case. We study the impact of the additional interference from other eNBs during the access phase because transmissions of RNs behaving as eNBs appear. We represent the interference as an equivalent drop in the mean channel gain during the access phase of half-duplex relaying. On the one hand, when the ratio between the channel gains is $\frac{\lambda_{d,a}}{\lambda_{d,r}} \to 0$, the second phase becomes less and less useful to direct users, until the scheduler converges again to a part-time pure PF scheduler with multiuser diversity $\alpha(0.75, 0.75)$. On the other hand, when the channels are equal $\frac{\lambda_{d,a}}{\lambda_{d,r}} = 1$, as if relay interference was completely avoided, the second phase contributes linearly to direct user throughput θ_d, and due to the reduction of the direct user demand of the first phase, there is also a noticeable increment in relayed user throughput θ_{r}.

This means that the lack of interference management we commented in the previous section is aggravated by the conflict between scheduling gain and relaying.

3.5.3.3 Effect of Half-Duplex Relaying Factor

PF scheduling is also affected by the time-sharing factor between the relay phase (RN-as-eNB) and the access phase (RN-as-UE), α, as illustrated in Fig 3.12. Obviously, the scheduler becomes increasingly biased as α shrinks. In the limit when $\alpha \to 1$, the PF scheduler is not biased, whereas when $\alpha \to 0$ it ends up serving only direct users.

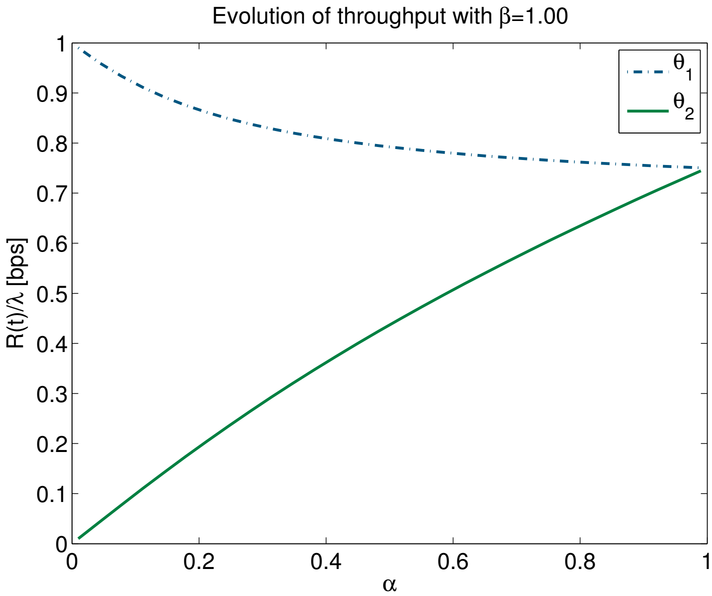


Figure 3.12: Evolution of the throughput of a two user PF scheduler, with one user served through a relay, as a function of α (analytical results).

As discussed above, relayed and access flow rates must be balanced to match inbound with outbound throughputs. If the relayed flow was higher, the relay would have to drop a significant number of packets, and if it was lower, the relay would have to transmit empty RBs. However, half-duplex timing values for TDD in the LTE-A standard are quite rigid, and the optimum α for (3.2) is not generally achievable.

Furthermore, there are two contradictory criteria in the standard. On the one hand, its `SubframeConfigurationTDD` values correspond to relays acting as receivers less than 50% of the time (small α). This implies, quite reasonably, that throughput balancing (3.2) should typically rely on better radio hardware in the DeNB-RN link than in the RN-UE.

On the other hand, PF is one of the preferred schedulers in LTE, but, as shown in Section 3.5.3.3, this family of schedulers only behaves fairly and achieves multi-user diversity gain for high values of α. This implies that scheduling would typically be heavily hampered in the most common relaying scenario.

3.5.3.4 Recomendations

Since choosing small α values to achieve link rate balancing would hamper interference and multi-user diversity, we propose setting α to relatively high values and complementing the system by selecting small values of β such that, for a fraction α of the channel, relay flows are scheduled to balance the average throughput as access flows, $\beta = \frac{\alpha\lambda_{\mathrm{r}}}{(1-\alpha)\lambda_{\mathrm{a}}}$. The purpose of this choice is to normalize the distribution of y βh_{r}, for it to offer an average $\frac{\alpha}{(1-\alpha)\lambda_{\mathrm{a}}}$ to the DeNB PF scheduler.

Note that we have not considered the RN scheduler. The RN access links can be analyzed as single-hop cells without full traffic generator buffers, achieving some given $\overline{\theta}_{\mathrm{a}}[t]$ independently of the DeNB, using any valid scheduler. Although the scheduler does not necessarily have to be a PF one, the reasons that make this scheduler popular for DeNBs also apply to RNs.

3.6 Relay Admission Control Strategies

In our simulation model, RNs are scattered across the terrain according to a uniform random spatial distribution. Operators may not always be free to place RN in the optimum spots, and relays purchased by customers or third-parties may need to be accommodated. Since randomly located RNs could appear at positions without any gains at all, it is necessary to develop admission-control strategies to avoid attaching those relays to the network. We aim at developing rules for the DeNB to decide to accept or not an RN during the `RN attachment procedure` described in [16]. To evaluate the advantage of accepting an RN into the network, we estimate the spectral efficiency of a user served by that RN instead of the DeNB.

Consider a simplified LTE single-user scenario as in Fig. 3.13. The DeNB is placed at the origin of coordinates $(0, 0)$. The DeNB allocates resources to serve a UE at position (x, y), and at a point $(x_{\mathrm{r}}, y_{\mathrm{r}})$ there is an RN, which may or may not be employed for transmission towards each UE. During the turn where this UE device is assigned the channel, scheduled by the eNB, we assume others are silent. We also asume that UE spatial distribution is homogeneous, so it is sufficient to approximate the achievable spectral efficiency for a single UE device at all points in the space. Achievable rates at each link depend on distance. The distance between the eNB and the UE is $r = \sqrt{x^2 + y^2}$, the distance between the eNB and the RN is $r_{\mathrm{r}} = \sqrt{x_{\mathrm{r}}^2 + y_{\mathrm{r}}^2}$, and the distance between the RN and the UE is $d = \sqrt{(x - x_{\mathrm{r}})^2 + (y - y_{\mathrm{r}})^2}$.

Let $\rho_{\mathrm{d}}(r)$ (b/s/Hz) be the achievable spectral efficiency of the direct link as a function of its distance. Similarly, let $\rho_{\mathrm{r}}(r_{\mathrm{r}})$ and $\rho_{\mathrm{a}}(d)$ be the respective achievable SEs of the

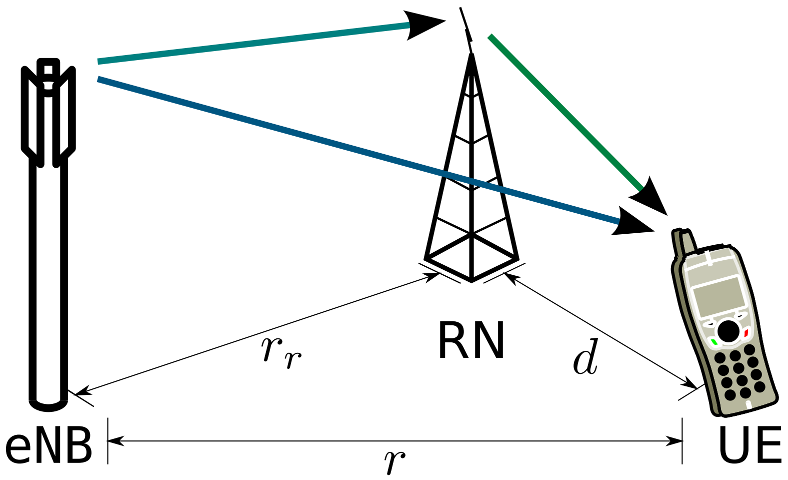


FIGURE 3.13: Simple LTE-A user DL options with a relay.

relay and access links. The optimal end-to-end spectral efficiency $\rho_{\mathrm{e2e}}(r_{\mathrm{r}}, d)$ is expressed in (3.3), and optimal UE atachement switching attains maximum spectral efficiency:

$$\rho_{\max}(r, r_{\mathrm{r}}, d) = \max(\rho_{\mathrm{d}}(r), \rho_{\mathrm{e2e}}(r_{\mathrm{r}}, d)) \tag{3.11}$$

In order to determine the spectral efficiency that is expected on a link, we employ the fitted Shannon curve for the AWGN case in [32]:

$$\begin{aligned} \rho_{\mathrm{LTE}}(\mathrm{SINR}) \simeq & \eta_W \log_2(1 + \eta_S^{-1}\mathrm{SINR}), \\ & \eta_W = 0.75, \eta_S = 1.25 \end{aligned} \tag{3.12}$$

were we consider a link budget as a function of path loss and a worst-case constant Signal to Interference Ratio (SIR)

$$\mathrm{SINR} = \frac{P_{\mathrm{t}} G_{\mathrm{t}} G_{\mathrm{r}} PL(r)}{BN_0(1 + f_r)(1 + \max \mathrm{SIR})} \tag{3.13}$$

where P_{t} represents the power at the transmitter and G_{t}, G_{r} the antenna gains; $BN_0(1+f_r)$ is the receiver noise power, computed multiplying the noise power spectral density, the noise bandwidth and the noise factor of the receiver; and $PL(r)$ is the path loss coefficient, which in practice grows with distance according to the simulator path-loss models. An approximate path-loss model for each link $x \in \{d, r, a\}$ is usually expressed as a function of the free-space loss PL_0, the path-loss exponent α_x (), and the distance r_x

$$PL(r) = r^{-\alpha} PL_0 \tag{3.14}$$

By combining expressions (3.12), (3.13), and (3.14), we can finally formulate the spectral efficiency ρ_x for each link x as a function of distance grouping all the constant link budget

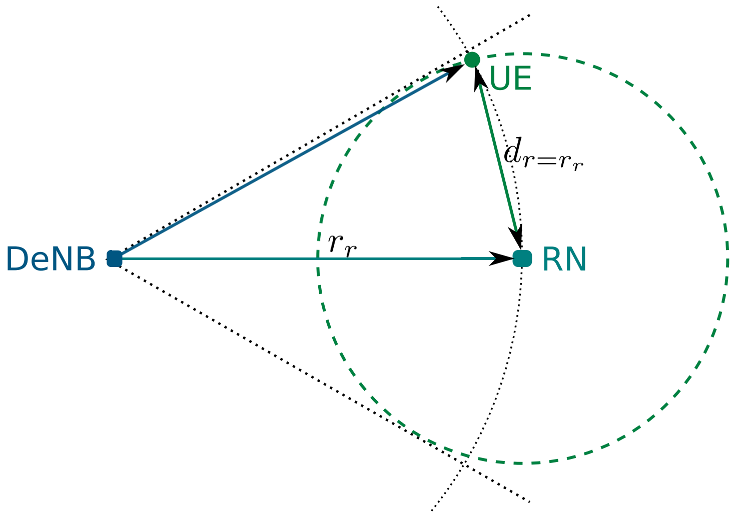


FIGURE 3.14: Geometry for the $r = r_\mathrm{r}$ point, which allows to calculate $d_{r=r_\mathrm{r}}$

parameters in a constant $\mathrm{SINR}_{0,x}$.

$$\rho_x(r_x) \simeq \eta_W \log_2(1 + \eta_S^{-1}\mathrm{SINR}_{0,x}|r_x|^{\alpha_x}), \tag{3.15}$$

Considering the point where the RN and the UE lie at equal distances from the DeNB, by forcing $r = r_\mathrm{r}$, only one of the three spectral efficiency components in (3.11) remains undetermined, $\rho_\mathrm{a}(d)$. Since the boundary of the relay serving area is the region where the two elements in the maximization (3.11) are equal, we can obtain the distance $d_{r=r_\mathrm{r}}$ that achieves this by inverting (3.15):

$$d_{r=r_\mathrm{r}} = \left(\frac{2^{\frac{1}{\eta_W}\frac{\rho_\mathrm{r}\rho_d}{\rho_\mathrm{r}-\rho_d}} - 1}{\eta_S^{-1}\mathrm{SINR}_{0,\mathrm{a}}} \right)^{-\frac{1}{\alpha_\mathrm{a}}} \tag{3.16}$$

as shown in Fig. 3.14.

Note that $\rho_\mathrm{r} - \rho_\mathrm{d} \neq 0$ at this precise distance due to the fact that the RNs have better reception than the UEs. This allows us to write a simple heuristic access control algorithm that rejects RNs with service areas that are too small or overlap with those of previously-attached RNs, as presented in Algorithm 1.

3.7 Summary

Relaying is an outstanding first step towards cooperative communications in future cellular networks, producing abundant gain as predicted in numerous research works. The 3GPP implementation standard for LTE-A is based on Layer-3 data forwarding and fixed

Algorithm 1 Third-party relay admission control rule

```
attachedRNs = vector of previously associated RNs
$d_{th}$=minimum range to accept;
RN = candidate RN
$\mathbf{p}_{\mathrm{r}}$=position of r;
$d_{RN}$=service range of RN;
if $d_{RN} > d_{th}$ then
  accept= true // default if range over threshold
  for all c $\in$ attachedRNs do
    $\mathbf{p}_c$=position of c;
    $d_c$=radius of c's attributed effect area;
    if $|\mathbf{p}_c - \mathbf{p}_{RN}| < d_{RN} + d_{\mathrm{r}}$ then
      accept=false // do not accept if overlap
      exit loop
    else
      Continue searching.
    end if
  end for
else
  accept=false
end if
```

parametric timing of the two relaying phases. We have shown that this rigid relaying implementation generates conflicts with other network mechanisms, such as scheduling, raising problems that have been ignored by theoretical work with flexible parameters.

The standard defines a fixed partitioning between relay and access links, unbalancing them, and underutilizing part of their capacities. We propose to apply interference management to concentrate transmission on a fraction of the total resources available to the underutilized link. This may mitigate the undesired interference increase due to relaying. Our analysis of a relay system with PF schedulers shows that interference with direct users during the access phase is also harmful for scheduling. There is a trade-off, because reducing the number of subframes with RN service reduces both interference and RN-UE transmission scheduling flexibility and, thus, multiuser-diversity gains.

DeNB schedulers may also experience other adverse effects. The wors effect would be a loss of multi-user diversity due to half-duplex relay operation, leading to suboptimal scheduling. We show that, even though schedulers can be modified to compensate for link unbalancing by introducing incentives, the cost is an even higher loss of multi-user diversity, to the point that the PF scheduler may be limited to RR performance at most. We propose seeking an intermediate point with incentives designed to balance DeNB-RN and RN-UE links at a scheduling level instead of at the subframe configuration level.

The set of relay configuration parameters in the standard is reasonable, but it seems to ignore some scenarios that may appear in a real deployment. Particularly, when potential multi-user diversity gain is high (for instance, when many relays and UEs are present), we have shown that it might be beneficial to grant more resources to relay channels than what is currently allowed.

RN location is a critical performance parameter. We have shown that it is possible to develop admission control strategies for randomly located RNs that allow to discard some of them when they are unlikely to contribute to a better network capacity.

The content in this chapter is an extended and actualized version of a paper published in the IEEE International Conference on Communications [12].

The relaying module for the Viena LTE System Level Simulator has been published under a free research license in `http://enigma.det.uvigo.es/~fgomez/` and featured in the official download page of the Vienna Simulator `http://www.nt.tuwien.ac.at/research/mobile-communications/lte-a-downlink-system-level-simulator/`

Appendix 3.A Implementation of Relaying Simulation

3.A.1 Overview of the Viena LTE System Level Simulator

3.A.1.1 LTE Network Elements Models

Some classes define the main functionality of LTE devices (depicted in blue in Fig. 3.15):

`eNodeB` Defines the *eNodeB* position and has a variable number of sectors. Does not serve *UEs* directly.

`eNodeB_sector` Defines *eNodeB*'s sectorial antennas, scheduling, etc, and serves *UEs*.

`UE` Performs reception, computes its own throughput, and it is basically is in charge of the final processes that yield the simulation results. It also has a `traffic_model` that queues packets in the attached `eNodeB_sector` according to different models.

3.A.1.2 Scenario Models

The scenario is a $(-X_{\min}, -Y_{\min}) \dots (X_{\max}, Y_{\max})$ square area. Nodes are placed in any discrete point of a $N \times M$ grid matrix (depicted in green in Fig. 3.15). Some grid classes are:

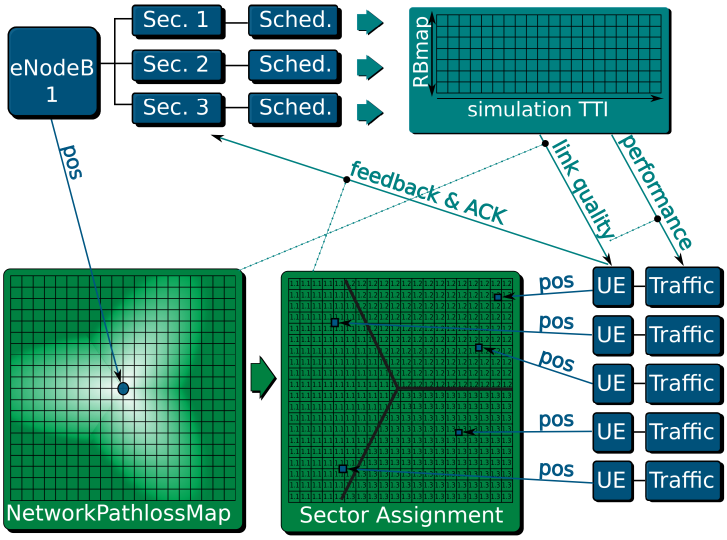


FIGURE 3.15: Simplified simulator architecture

`NetworkPathlossMap` Precomputed path loss from each `eNodeB_sector` to all $N \times M$ points in the simulation area.

`eNodeBs.pos` and `UEs.pos` The position of each *eNodeB* and *UE* as integer coordinates.

`sector_assignment` For each point on the map grid, the `eNodeB_sector` *cellID* assigned if a user takes that position.

3.A.1.3 System Operation Model

At each time t, RRC actions modify the data structures represented in teal in Fig. 3.15. These actions, in black in pseudocodes Alg 2 and Alg 3, include:

1. `UE.traffic_models`, which generate packets and queue them at their `eNodeB_sector` source.
2. `eNodeB_sectors`, which receive wireless channel feedback from `UEs`.
3. `eNodeB_sector`, which calls the `Scheduler` to allocate RBs in the `RB_grid`.
4. `UE`, which, according to the allocations and their attributes,
   - **link_quality_model:** Using `NetworkPathlossMap` and configuration, generates random instantaneous SINR.
   - **calculate_feedback:** Calculates feedback using `link_quality_model` results.
   - **link_performance_model:** Using `link_quality_model` results and traffic models, computes the Block Error Probability, stores or drops packets accordingly, and delivers ACK/NACK in sector feedback.

5. Updates `traffic_model` according to succesful deliveries.

Algorithm 2 Simulator operation. Changes for relaying (initialization)

```
//Initialization
Generate eNodeB locations and setup sectors;
Generate RN locations and setup sectors;
For each position $x, y$, compute path loss without relays from each sector $k \in$DeNB_sectors;
For each position $x, y$, compute sector assignment for relays; //max of SINR_all_sectors

Perform RN admission control; //optional
Assign RNs to DeNBs;
For each position $x, y$, compute path loss from each sector $k$;
For each position $x, y$, compute sector assignment; //max ~~SINR~~R_all_sectors
Generate UE positions;
Create traffic models and attach them to UEs;
for all UE $u$ do
    Attach $u$ to sector ID sector_assignment(UE(u).pos.X,UE(u).pos.Y));
end for
Create schedulers, attach them to sectors and provide UE and traffic_model pointers to them;
for all RN $r$ do
    for all UE $u$ attached to $r$ do
        Create a clone $c$ of $u$;
        Assign $r$ position to $c$;
        Change $c$'s antenna and noise model from UE radio to RN radio;
        Assign $u$'s traffic generator to $c$;
        Create a relay_traffic_generator, depending on $c$, and assign it to $u$;
        Attach $c$ to $r$'s DeNB;
    end for
end for
```

3.A.2 Relaying Modifications of the Simulator

3.A.2.1 Changes in Network Elements

New elements were created for the simulator to emulate a relaying scenario (Blue in Fig 3.16).

The RN is modeled with three components:

`RelayNode` site: It has the same purpose as `eNodeB` elements. In addition, it manages relay attachment (`DeNBID`) and traffic forwarding.

`eNodeB_Rsector`: It has the same functions as `eNodeB_sector` but with the restriction that some UEs may not be scheduled at some instants. A new string parameter

Algorithm 3 Simulator operation. Changes for relaying (main)

```
//Main loop
while simulation_clock.current_TTI<SIM_DURATION_TTI do
  ~~Move UEs and reassociate to sectors //handover~~
  for all sector $s$ do
    L=length(relay_service_sequence)
    Receive UEs feedback, with ~~optional~~ delay L;
    switch ($s$.relay_service_sequence(current_TTI mod L))
    case 'a':
      Schedule all UEs;//legacy behavior, not consistent with relays due to self-
      interference
    case 'd':
      Schedule direct UEs;
    case 'r':
      Schedule relay UEs;
    case 'p':
      Schedule cloned UEs;
    case 'u':
      Schedule direct and relay UEs;
    case 'b':
      Schedule direct and cloned UEs;
    end switch
  end for
  for all UE do
    compute link quality;
    send feedback;
    check if it was scheduled and generate link performance; //transmission errors,
    ACK, and traces
  end for
  for all relayed UE relay_traffic_generator do
    check acknowledged traffic of cloned UE of reference;
    generate packets with the same size towards the relayed UE;
  end for
end while
//End
Compute averages of traced parameters; /througput, SINR, etc...
Clean memory;
```

`relay_service_sequence` is added. This is a sequence of characters that specify the cyclic pattern according to which UEs are to be served in each TTI (table 3.2). For example `relay_service_sequence='ddbr'` means that two TTIs serve only direct UEs, followed by one that serves direct UEs and cloned UEs, followed by one that serves only relay UEs. With this method, the simulator can simulate any conceivable time-division scheme, including those non-standard ones.

`proxyUE`: The relaying module has two types of UEs: *real* UEs and *virtual* UEs; although they are implemented internally by the same class.

- A real UE may be attached to RNs (relayed UE) or DeNBs (direct UE) and models a user of the system.
- A cloned UE emulates the DeNB-RN link for its mirrored real UE. Every time a real UE attaches itself to a RN, the RN creates a cloned UE attached to its DeNB.

Regarding the element `eNodeB_Rsector.scheduler`, legacy schedulers may be used with this module, because the changes affect the sectors that call those schedulers. However, the scheduler is responsible for the semantic correctness of replacing calls of the form *schedule(UElist)* with calls of the form *schedule(chooseUEsubset(UElist,relay_service_sequence))*. The `PropFair_traffic` scheduler used in the tests was debugged for this purpose.

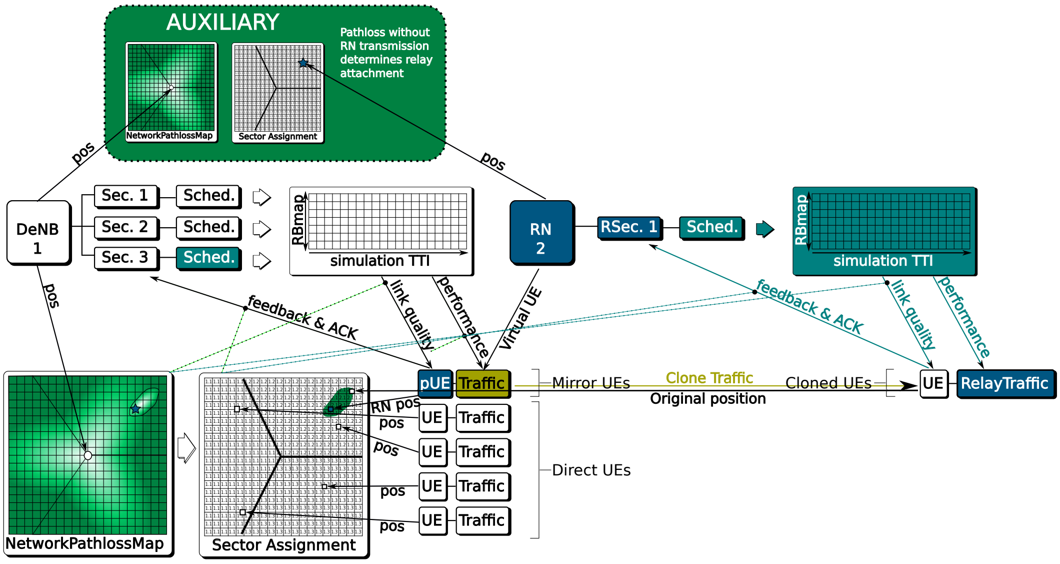


FIGURE 3.16: Changes in the simulator architecture for relaying.

TABLE 3.2: Possible values of the characters in `relay_service_sequence` for any TTI.

Value	
`'a'`	**BACKWARDS-COMPATIBLE BEHAVIOR** Serve all UEs. *Warning: There is no self-interference model in the relays. Therefore, the results of using this mode are not physically correct.*
`'d'`	Serve only “direct” UEs, connected directly to a main eNB
`'r'`	Serve only “relayed” UEs, connected to a RN
`'p'`	Serve only RNs' “proxy” UE connections, DeNB-RN
`'u'`	Serve all UEs, (d+r)
`'b'`	Serve all DeNB connections (d+p)

3.A.2.2 Changes in the Traffic Model

The LTE system level simulator has parametric traffic source models. DeNBs have the usual parametric traffic generators. RNs re-queue traffic received by cloned UEs in a `relay_traffic_generator` to the final destination (depicted in yellow in Fig. 3.16). For rate tracing, it suffices to check traces of the final destinations, discarding throughput metrics of cloned UEs.

3.A.2.3 Scenario Models

`NetworkPathlossMap` and `SectorAssignment`: RNs cannot attach themselves to other RNs. Therefore, the relaying module needs two instances of these variables. `NetworkPathlossMap_norelays` and `SectorAssignment_norelays` (depicted in green in Fig 3.16) are computed taking only DeNBs into account and they are used for RN-DeNB assignment only. The original homonimous data structures maintain their function but UE attachment is performed by a estimator of effective rate instead of based on highest SINR.

$$\begin{aligned} R_{RN} &\simeq \frac{\text{num. RN TTIs}}{\text{total num. TTIs}} \log_2(1 + SINR_{RN} \times \varepsilon_{SINR}) \times \varepsilon_{BW} \\ R_{DeNB} &\simeq \log_2(1 + SINR_{DeNB} \times \varepsilon_{SINR}) \times \varepsilon_{BW} \end{aligned} \tag{3.17}$$

where ε_{SINR} and ε_{BW} represent LTE inefficiency compared to Shannon's capacity [32].

RN admission control: After generating the RN positions and before attaching them to DeNBs, a procedure is called to remove from the list those RNs that do not grant an estimated effective throughput gain compared to the previous network state.

3.A.2.4 Operation

At each TTI t, in addition to the changes described above, operation structures (depicted in teal in Fig 3.16) include changes in feedback reception. The state of the channel changes at different points on the `relay_service_sequence`. Therefore, UEs delay transmitted feedback at equal instants of (`relay_service_sequence`$t \mod L$). Thus, the module introduces feedback delay. All the modifications of the simulation procedure are represented in color fonts in Alg. 2 and Alg. 3.

Appendix 3.B Average Throughput in PF Schedulers

3.B.1 Non-relaying eNB Scheduling Analysis

In [163], a set of users $\{i\}$, $i \in [1, n]$ experiencing Rayleigh fading is analyzed. The instantaneous channel of each user experiences an exponentially distributed instantaneous SNR $h_i \sim \mathrm{Exp}(\lambda_i)$ (the squared module of a Rayleigh distribution is an Exponential one). The instantaneously achievable rate per resource use is approximately

$$\theta_i[t] \simeq h_i \tag{3.18}$$

in the low-SNR regime.

By definition, the time average of the instantaneous rates of each user would converge to the constant value:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1,\{t:s[t]=i\}}^{T} (\theta_i[t]) \to \overline{\theta}_i, \tag{3.19}$$

and the scheduling function chooses, for each resource, the user that in that moment is experiencing the highest instantaneous rate normalized by its average

$$s[t] \in [1, n] = \arg_i \max \frac{\theta_i[t]}{\overline{\theta}_i}. \tag{3.20}$$

The evolution of the set of user rates follows the ODE (3.6). This dynamic system is self-stabilizing and converges to a permanent state if the ODE has a solution with zero derivatives, which means that the mean rate distribution converges to the time-average $\mathrm{E}_{t:s[t]=1}\left[\theta_i[t]\right] - \overline{\theta}_i = 0$ for each user.

We compute the mean of the rate distribution as a function of a given parameter $\overline{\theta}_i$, which is equivalent to computing the average channel state conditioned to the event that said channel is selected by the scheduler, as

$$
\begin{aligned}
\mathrm{E}_{t:s[t]=1}\left[\theta_i[t]|\overline{\theta}_i\right] &= \mathrm{E}_{h_i[t]/\overline{h}_i > h_j[t]/\overline{h}_j \forall j}\left[h_i[t]\right] \\
&= \int_0^\infty x f_{h_i[t]}(x) \prod_{j\neq i} F_{h_j[t]}\left(x\frac{\overline{h}_j}{\overline{h}_i}\right) dx \\
&= \int_0^\infty x\lambda_i e^{-\lambda_i x} \prod_{j\neq i} 1 - e^{-\lambda_j x \frac{\overline{h}_j}{\overline{h}_i}} dx \\
&= \int_0^\infty x\lambda_i e^{-x\sum_j \lambda_j \frac{\overline{h}_j}{\overline{h}_i}} dx \\
&= \frac{\lambda_i \overline{h}_i^2}{\left(\sum_j \lambda_j \overline{h}_j\right)^2}
\end{aligned}
\tag{3.21}
$$

For a small number of users, we can extend the expression completely, but this requires solving a different integral for each number of users. The integral for the two-user case is

$$
\int_0^\infty x\lambda_1 \left[e^{-x\lambda_1} - e^{\lambda_1 + \lambda_2 \frac{\overline{h}_1}{\overline{h}_2}}\right] dx = \frac{1}{\lambda_1} - \frac{\lambda_1 \overline{h}_1^2}{\left(\lambda_1 \overline{h}_1 + \lambda_2 \overline{h}_2\right)^2}
\tag{3.22}
$$

And taking this to the ODE leaves the throughputs as the solution of equation

$$
\begin{aligned}
0 &= \tfrac{1}{\lambda_1} - \tfrac{\lambda_1 \overline{\theta}_1^2}{\left(\lambda_1\overline{\theta}_1 + \lambda_2\overline{\theta}_2\right)^2} - \overline{\theta}_1 \\
0 &= \tfrac{1}{\lambda_2} - \tfrac{\lambda_2 \overline{\theta}_2^2}{\left(\lambda_1\overline{\theta}_1 + \lambda_2\overline{\theta}_2\right)^2} - \overline{\theta}_2
\end{aligned}
\tag{3.23}
$$

Other interesting simplification for any number of channels appears when all channels are independent and identically distributed (i.i.d.) $\lambda_j = \lambda \forall j$. The ODE can be rewritten as

$$
\begin{aligned}
\overline{\theta}_1^3 &= \tfrac{1}{\lambda\left(\sum_j \frac{1}{\overline{\theta}_j}\right)^2} \\
\overline{\theta}_2^3 &= \tfrac{1}{\lambda\left(\sum_j \frac{1}{\overline{\theta}_j}\right)^2} \\
\vdots & \quad \vdots \\
\overline{\theta}_n^3 &= \tfrac{1}{\lambda\left(\sum_j \frac{1}{\overline{\theta}_j}\right)^2}
\end{aligned}
,
\tag{3.24}
$$

producing the multiuser-diversity gain mentioned before $\overline{\theta}_i = \frac{1}{\left(\sum_j \frac{1}{\overline{\theta}_j}\right)^2}\frac{1}{\lambda}$. This gain is also present in the general solution for n non i.i.d. users given in [163].

3.B.2 Modification of the Analysis for eNBs with RNs

When there are n_d direct flows and n_r flows towards the relay, the solution of 3.6 for the scheduling of the first hop of a DeNB follows the ODE (3.9). The averages in the three terms can be computed analogously to [163] with the trivial inclusion of b_i weighted values in the scheduling rule and by differentiating λ_a and λ_r according to their definition.

For the two-user case we get

$$\begin{aligned}
\mathrm{E}_{t:s[t]=1}\left[\theta_i[t]|\overline{\theta}_i\right] =&\alpha\int_0^\infty x\lambda_{1,r}\left[e^{-x\lambda_{1,r}}-e^{\lambda_{1,r}+\lambda_{2,r}\frac{\overline{h}_2 b_1}{\overline{h}_1 b_2}}\right]dx\\
&+(1-\alpha)\int_0^\infty x\lambda_{1,a}\left[e^{-x\lambda_{1,r}}-e^{\lambda_{1,a}+\lambda_{2,a}\frac{\overline{h}_2 b_1}{\overline{h}_1 b_2}}\right]dx\\
=&\alpha\left[\frac{1}{\lambda_{1,r}}-\frac{(\overline{h}_1 b_1)^2\lambda_{1,r}}{\left(\lambda_{1,r}\overline{h}_1+\lambda_{2,r}\overline{h}_2\right)^2}\right]+(1-\alpha)\left[\frac{1}{\lambda_{1,a}}-\frac{(\overline{h}_1 b_1)^2\lambda_{1,a}}{\left(\lambda_{1,a}\overline{h}_1+\lambda_{2,a}\overline{h}_2\right)^2}\right]
\end{aligned} \tag{3.25}$$

Assuming that first user is direct and the second one is relayed, the exact solution with $b_1 = 1$, $b_2 = \beta$, and $\lambda_{2,a} = 0$ must satisfy the ODE as

$$\begin{aligned}
\overline{\theta}_1 &= \alpha\left[\frac{1}{\lambda_{1,r}}-\frac{\lambda_{1,r}\overline{\theta}_1^2}{\left(\lambda_{1,r}\overline{\theta}_1+\lambda_{2,r}\beta\overline{\theta}_2\right)^2}\right]+(1-\alpha)\left[\frac{1}{\lambda_{1,a}}-\overline{\theta}_1\right]\\
\overline{\theta}_2 &= \alpha\left[\frac{1}{\lambda_{2,r}}-\frac{(\overline{\theta}_2\beta)^2\lambda_{2,r}}{\left(\lambda_{1,r}\overline{\theta}_1+\lambda_{2,r}\beta\overline{\theta}_2\right)^2}\right]
\end{aligned} \tag{3.26}$$

Due to the additive term $+(1-\alpha)\left[\frac{1}{\lambda_{1,a}}-\overline{\theta}_1\right]$, this system has a solution that is not a trivial transformation of the original one in [163]. Therefore, we have solved this equation system to obtain the analytic values in 3.6 using numeric methods.

Looking at the general case with several users, the ODE would be

$$
\begin{array}{rcll}
\overline{\theta}^3_1 & = & \alpha\dfrac{\lambda_{1,r}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} & + \;(1-\alpha)\dfrac{\lambda_{1,a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\overline{\theta}^3_2 & = & \alpha\dfrac{\lambda_{2,r}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} & + \;(1-\alpha)\dfrac{\lambda_{2,a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\vdots & & \vdots & \qquad\qquad\vdots \\
\overline{\theta}^3_{n_\mathrm{d}} & = & \alpha\dfrac{\lambda_{n_\mathrm{d},r}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} & + \;(1-\alpha)\dfrac{\lambda_{n_\mathrm{d},a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\overline{\theta}^3_{n_\mathrm{d}+1} & = & \alpha\dfrac{\lambda_{n_\mathrm{d}+1}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} & \\
\overline{\theta}^3_{n_\mathrm{d}+2} & = & \alpha\dfrac{\lambda_{n_\mathrm{d}+2}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} & \\
\vdots & & \vdots & \\
\overline{\theta}^3_{n_\mathrm{d}+n_\mathrm{r}} & = & \alpha\dfrac{\lambda_{n_\mathrm{d}+n_\mathrm{r}}}{\left(\sum_{j=1}^{N_d+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} &
\end{array}
\tag{3.27}
$$

where b_i equals 1 for direct users and β for relay users. We have that, when the incentive becomes too large ($\beta \to \infty$), each equation of the ODE system can be approximated as a system that only takes into account users of the same type

$$
\lim_{\beta\to\infty}
\left\{
\begin{array}{c}
\overline{\theta}^3_1 \\
\overline{\theta}^3_2 \\
\vdots \\
\overline{\theta}^3_{n_\mathrm{d}} \\
\overline{\theta}^3_{n_\mathrm{d}+1} \\
\overline{\theta}^3_{n_\mathrm{d}+2} \\
\vdots \\
\overline{\theta}^3_{n_\mathrm{d}+n_\mathrm{r}}
\end{array}
\right\}
\simeq
\left\{
\begin{array}{c}
(1-\alpha)\dfrac{\lambda_{1,a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
(1-\alpha)\dfrac{\lambda_{2,a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\vdots \\
(1-\alpha)\dfrac{\lambda_{n_\mathrm{d},a}}{\left(\sum_{j=1}^{n_\mathrm{d}} \lambda_{j,a}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\alpha\dfrac{\lambda_{n_\mathrm{d}+1}}{\left(\sum_{j=n_\mathrm{d}+1}^{n_\mathrm{d}+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\alpha\dfrac{\lambda_{n_\mathrm{d}+2}}{\left(\sum_{j=n_\mathrm{d}+1}^{n_\mathrm{d}+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2} \\
\vdots \\
\alpha\dfrac{\lambda_{n_\mathrm{d}+n_\mathrm{r}}}{\left(\sum_{j=n_\mathrm{d}+1}^{n_\mathrm{d}+n_\mathrm{r}} \lambda_{j,r}\frac{b_j}{\overline{\theta}_j}\right)^2}
\end{array}
\right\}
\tag{3.28}
$$

Part II

The Near Future: Better Spectrum Usage in Current Wireless Networks through Cooperation

Chapter 4

Analysis of Cooperative Diversity in Cognitive Radio Spectrum Leasing

Contents

4.1 Introduction

We consider a channel model where there is a *primary* transmitter subject to random channel fading, with full spectrum rights available, and a *secondary* transmitter without any spectrum rights. In the earlier CR approaches, however, there is no interaction between primary and secondary signals. Secondary nodes simply detect and transmit in pre-existing white spaces of spectrum that belongs to the primary, independently of its transmission being efficient or not. Going a step further, Spectrum Leasing (SL) is a CR technique where the primary transmitter actively leases part of its resources to a secondary transmitter [37], typically in exchange for a payment.

Our CSL model considers a situation in which SL is implemented using cooperation as payment. From a point of view of CR, this is a type of SL where the secondary transmitter, willing to gain access, offers its help to induce a rate gain on the primary transmitter in exchange for a portion of the spectrum resource gains. By *rate gains* we refer to the extra capacity per resource that becomes achievable thanks to the improved efficiency with cooperation. And by *resource gains* we refer to the fraction of spectrum resources that the primary does not need for achieving its original rate any longer, after cooperation is established.

The first contribution of this chapter is the CSL performance evaluation philosophy itself. We characterize the system-level resource gain as a whole, but we do not consider the negotiation that takes place to split the gained spectrum between the primary and secondary. Unlike previous approaches in CR based on game theory and other areas that study the equilibrium in this sort of negotiations [37, 38], we are interested in testing the concept of CSL as a whole by measuring the total amount of spectrum gained -as a form of aggregate social benefit- and we do not concern ourselves by the dynamics of its distribution between individuals.

On the other hand, from the point of view of cooperative communications, this is a novel type of cooperative channel where one transmitter has full channel access rights and the other none. To the best of our knowledge, all previous studies of cooperative diversity rate gains assume that the channel is shared equally by all the nodes, which in practice may mean that all nodes are equal and implement the same fair MAC protocol [41, 108, 110, 112]. In order to study the cooperative diversity gain in CR we remove this assumption.

In our cooperative communications analysis a primary user has a poor channel gain and full rights on spectrum resources while a secondary user with better channel gain must provide a collaboration payment in order to gain access to the resources. The goal of the cooperation decision is to obtain net resource gains relative to a model without collaboration. This “asymmetric” channel allocation gives nodes in a network incentives to be cooperative and solves the problem we introduced on 2.4.3 that mechanisms are needed to induce users to cooperate willingly despite the power and computational costs. This model could even be used in non-truly CR networks as an ultimatum-type incentive mechanisms to encourage cooperation in the MAC of a normal wireless network. For this, it would suffice to introduce a “virtual” primary-secondary relationship in a cooperative MAC protocol. Even though we do not focus on the way in which gains are shared, we suggest that, for the purpose of creating incentives for cooperation in regular wireless networks, the resources should be divided in half between the two transmitters, as fairness is commonly a desirable network property.

Ultimate decision-making corresponds to the primary transmitter, which owns the spectrum. Typically, cooperative diversity transmission involves two phases. First, the source node attempts to transmit information towards the destination. Meanwhile, neighbor nodes simultaneously store the information they overhear. In the second phase, this stored information is relayed by one (or several) of those nodes. In our model the primary system is responsible for switching its transmission mode from direct mode to cooperative mode (cooperation decision) when the latter produces net resource gains, and for making part of these resources available to collaborators (cooperative transmission design) who are notified (neighbor notification).

To perform the primary decision, we formulate two decision-making rules depending on the degree of channel knowledge available to the primary transmitter: The first scheme assumes that the primary has only partial knowledge of the statistics of random fading distribution (statistical CSI) and/or can only make a long-term static leasing decision for all channel realizations. The second scheme assumes that the primary has perfect knowledge of the time-variant channel (instantaneous CSI) and makes matched time-varying decisions to lease a dynamic amount of resources. The instantaneous CSI knowledge scenario is only feasible in practice if the resource gains are worth the overhead. The analysis we perform assumes that a “genie” makes CSI knowledge available at zero overhead, and provides an upper bound on achievable resource gains in practical scenarios. If the channel varies too rapidly to be tracked, only the statistical (distribution) CSI strategy can be practically used.

The second main contribution of this chapter is an analysis of the conditions in which the secondary transmitter helps to gain spectrum resources, quantifying the gain for the two scenarios with different levels of CSI. Our analysis shows that MI increases in these scenarios in particular circumstances. In the first scenario, the average MI will only improve if the primary channel is suffering an important degradation. In the second scenario, MI distribution can always produce a gain with some probability that grows with the degradation of the primary channel.

The structure of this chapter is as follows: In section 4.2 we describe the CSL channel model. In section 4.3 we obtain analytical expressions for the probability of increasing MI by cooperation. This is equivalent to the probability of using cooperation and is relevant to show how often, if deployed, the cooperative mechanism would be effectively activated by the primary. In section 4.4 we analize MI as a random variable, and obtain its probability density function (p.d.f.). In section 4.5 we obtain the closed expressions for the average MI and apply them to compute the fractional resource gains achieved for the static decisions based on statistical CSI. In section 4.6 we analyze the p.d.f. of the time-varying fraction of resource gains in the instantaneous CSI knowledge scenario. Finally, on section 4.7 we conclude the chapter discussing the practical conditions for resource gains: for statistical channel knowledge to suffice, the primary link must be of severely low quality, whereas for instantaneous channel knowledge non-zero resource gains are achieved in any situation, but they are small for reasonably good channels and they only increase to significant amounts when the primary channel gets worse.

4.2 System Model

We consider a wireless link between a primary source S_p and a primary destination D_p, with a transmitted signal energy to noise ratio of $\mathrm{SNR} = P/BN_0$, through a narrow-band channel with random Gaussian fading with coefficient $h_\mathrm{sd} \sim \mathcal{N}(0, \sigma^2_\mathrm{sd})$. The transmission rate is originally limited by the MI of the direct channel [41]:

$$I_\mathrm{D} = \log_2(1 + |h_\mathrm{sd}|^2 \mathrm{SNR}) \tag{4.1}$$

Our goal is to study the effect of introducing a secondary transmitter S_s that wants extra resources to transmit its own information to its own destination. The result is the interference channel shown in figure 4.1. If S_s is not receiving leased resources, the primary transmission from S_p to D_p occupies the entire medium. S_s can overhear the primary transmission and in some cases enhance it by relaying information. In such a case, resource gains may be shared with a secondary information flow from S_s to D_s, as a reward for cooperation. S_p, the spectrum owner, controls the whole process.

From the point of view of a secondary, and taking into account all possible CR models of spectrum ownership and occupancy, the sequence of actions taken by S_s to gain spectrum resources would be as follows:

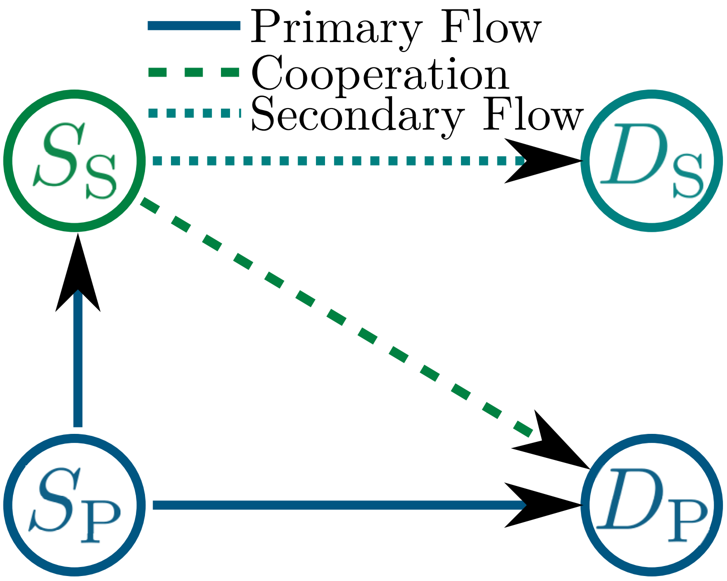


FIGURE 4.1: CSL Wireless network with four nodes.

- Firstly, if S_s has its own channel, it will use it, which is not reflected in our model.
- Secondly, it will try to find white spaces -i.e. licensed spectrum that owners do are not using in their direct transmissions- thereby avoiding the need to cooperate and waste energy in this process. This is not reflected in our model either.
- Finally, if it still needs more resources, or if none of the previous alternatives are available, S_s as a last option can recur to cooperation as described in our model. It will offer aid to inefficient primaries and, in such a case, cooperation may yield resource gains that the primaries will share with S_s.

Like in [41], we assume that wireless nodes are half duplex and that transmissions are not overlapped by assigning a channel slot to a single transmitter at a time. This places fewer homogeneity constraints, and allows cooperation between technologies that do not support overlapping. Our analytical results are obtained using the cooperative DF protocol [41] but more powerful relaying techniques would achieve better gains.

Two levels of channel knowledge are considered: knowledge of average fading (statistical CSI) and full knowledge of instantaneous CSI. The former is more adequate for channels that change faster than the response time of the channel estimation and decision mechanism. The latter is more suitable in case of a slower channel variation, where the channel estimation and decision mechanism can dynamically follow instantaneous channel states. Obviously, a heavy overhead of channel measurement may degrade performance in the second case. We have deliberately ignored the overhead and assumed that a "genie" provides instantaneous CSI to the decision process. Thus, our results set an upper bound on the benefits achievable in any practical system. This upper bound is useful to provide an estimation of the value of CSL that is independent of the MAC protocol.

The decision rules for statistical CSI (D_s) and instantaneous CSI (D_i) can be expressed as follows:

$$D_\mathrm{s} = \arg\max(\mathrm{E}\left[I_\mathrm{D}\right], \mathrm{E}\left[I_\mathrm{C}\right]) \tag{4.2}$$

$$D_\mathrm{i} = \arg\max(I_\mathrm{D}, I_\mathrm{C}) \tag{4.3}$$

where I_C is the MI achieved by the primary using a cooperative transmission mechanism. Node S_p takes all the decisions, as the owner of the spectrum. S_p must reconfigure its transmission mechanism and inform S_s of its decision. In cooperative diversity studies there are many proposals for mechanisms that allow channel measurement and signaling (Chapter 2). If channel variations are slow, one possibility would be to grant the nodes access to a shared database [165]. Faster varying channels can be estimated from control packets as in the three-way handshake mechanism described in [102]. If channel variation is too fast for any of the mechanisms available, the statistical CSI approach is the only practical alternative. Regarding the physical layer, two DF receivers are described in [102]: separate decode attempts for each version of the packet received and the application of Maximum Rate Combining (MRC) to exploit the information in the two signals received.

When the primary transmitter and the secondary follow different radio standards, exploiting cooperative diversity imposes certain compatibility requirements, as the secondary radio must be able to receive and retransmit S_p signals. However, this may not be a limiting assumption, since new paradigms such as Software Defined Radio (SDR) suggest that hardware reconfigurability will be easily achievable in future devices.

To illustrate the difference between CSL and the traditional approaches on cooperative diversity with equal spectrum rights, let us compare what would happen in a fair MAC protocol that arbitrates the accesses of S_p and S_s to the medium and what would happen in a CR approach.

4.2.1 Two nodes with equal rights: fair MAC protocol

If a fair MAC is used by S_p and S_s, the two nodes will receive an equal share of resources. The usual approache to capacity is the well-known MAC channel, but, as previously mentioned, analytical studies of cooperative diversity follow a simplified TDMA/FDMA protocol, such as that described in [41], where equal portions of channel resources are granted to each user and overlapping transmissions are not allowed. Thus the rate of each transmitter-receiver pair is limited by half the mutual information available in two parallel single-user channels. Using protocol x (x may refer to direct transmission or cooperative transmission with different protocols), each node i (s or p) is limited by:

$$R_i \leq \frac{I_{x,i}}{2} \tag{4.4}$$

where the division by 2 represents a fair medium sharing and $I_{x,i}$ is the achievable MI for node i under protocol x. For example, for cooperative diversity using one DF relay ($x = \mathrm{C} = \mathrm{DF}$), the MI achieved by each node i combining direct and relayed receptions is given by (2.3)

$$I_{\mathrm{DF}} = \frac{1}{2}\min(\log_2(1 + \mathrm{SNR}|h_{\mathrm{sr}}|^2), \log_2(1 + \mathrm{SNR}|h_{\mathrm{sd}}|^2 + \mathrm{SNR}|h_{\mathrm{rd}}|^2))$$

where h_{sr}, and h_{rd} are the fading coefficients of the source-relay and relay-destination channels, respectively.

4.2.2 Two nodes with different rights: Traditional cognitive radio and the model we propose

If nodes S_{p} and S_{s} do not use a fair TDMA/FDMA protocol, resource allocation may be asymmetric. In many commercial communications, a licensed operator can retain its frequency bands even if it is not using them fully. This may lead to a considerable waste of resources if operator transmission is inefficient.

In a typical CR scenario, S_{s} would use its knowledge of the primary transmission to access free portions of the spectrum (white spaces) or to transmit its signal overlapped with the primary transmission spectrum, using techniques that prevent or repair interference with the primary reception. Thus, traditional CR assumes that S_{p} does not actively reduce its spectrum usage. Our approach, in contrast, assumes that S_{p} can lease out spectrum resources that are not utilized efficiently in exchange for cooperation from S_{s}. This leads to net resource gains, which S_{p}, as the spectrum owner, controls.

To characterize resource sharing, let us define the following:

Definition 4.1. The **leasing fraction** $\alpha \in [0, 1]$ is a scalar representing the proportion of channel resources that S_{p} releases to S_{s}.

For example, in a TDMA scheme such as that shown in Fig. 4.2, αT would represent the interval for secondary information transmission and $(1 - \alpha)T$ the interval for primary information transmission. In OFDMA, α would be the proportion of carriers that S_{s} would borrow for the transmission of secondary information [38].

Concerning the resources that the primary keeps to itself, they are further divided according to the needs of the chosen cooperative protocol according to the different phases of the cooperative transmission assistance. In our analysis, using DF, on the first stage S_{p} transmits towards D_{p} and S_{s} overhears the signal. In the second stage of cooperation -still within the $1-\alpha$ fraction of the resources- S_{s} relays the primary information. Finally, S_{s} is granted a fraction αT of resource gains for it own secondary transmission.

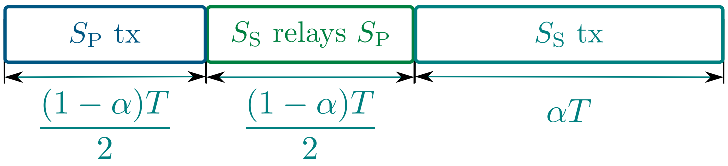


FIGURE 4.2: Time division scheme for resource leasing characterized by α.

From its knowledge of the channel, S_{p} selects the values of α. We are only interested in studying the limitations of accumulated cooperative gains, or potential values of α, without particularizing to any policy for distributing them. For this purpose we define and study the following parameters with influence on α:

Definition 4.2. The equal-rate leasing factor α_{eq} is the fraction of resources that the primary transmitter would need at any given instant to obtain with cooperation the same rates as it obtains using all the resources with direct transmission. It is computed as a function of the instantaneous MI values as follows

$$I_{\mathrm{D}} \leq (1-\alpha)I_{\mathrm{C}} \Rightarrow \alpha_{\mathrm{eq}} = (1 - \frac{I_{\mathrm{D}}}{I_{\mathrm{C}}}) \tag{4.5}$$

Note that (4.5) may yield negative values for α_{eq} when cooperation is not beneficial; i.e. the primary needs more resources with a cooperative protocol than with direct transmission. Thus, we define:

Definition 4.3. With instantaneous CSI knowledge, the maximum dynamic leasing fraction for each channel realization α_{max} is the maximum fraction of instantaneous resource gains for the resulting from decisions according to (4.3)

$$\alpha_{\mathrm{max}} = \max(0, \alpha_{\mathrm{eq}}) \tag{4.6}$$

On the other hand, for decisions based on statistical CSI, the fraction of resources gained must be computed with a static ergodic expression because the selected parameter α can only be static and track coarse long-term channel distribution variations. Thus,

Definition 4.4. With statistical CSI knowledge, the maximum static leasing fraction $\alpha_{\tilde{\max}}$ is the maximum fraction of resource gains of the mutual information averaged over all channel realizations, resulting from decisions according to (4.2)

$$\tilde{\alpha}_{\max} = \max\left(0, 1 - \frac{E\left[I_{\mathrm{D}}\right]}{E\left[I_{\mathrm{DF}}\right]}\right) \tag{4.7}$$

Expressions (4.6) and (4.7) determine the spectrum gains of cooperation for each level of decision-making. They are zero when cooperation does not produce resource gains. We have that when $P(\alpha_{\max} > 0) > 0$, with instantaneous CSI there is a potential gain that can be exploited by selecting dynamic values $\alpha \in [0, \alpha_{\max}]$. For the case of statistical CSI knowledge, the same is true for static allocation of values $\alpha \in [0, \tilde{\alpha}_{\max}]$.

Note that the decision rules (4.2) and (4.3) only determine whether or not cooperation is used, whereas the maximum leasing fractions (4.6) and (4.7) only determine the maximum amount of leased resources for the primary to maintain its original rate. However, the actual selected fraction of leased resources α is not specified by our model.

In CR, the choice of α within the margins we provide is an open problem that could be studied with economic or game theory models. However, from the perspective of cooperation incentive mechanisms for network standards using CSL, we can suggest different policies depending on the type of traffic of the primary transmitters. Next, we list our suggestions for the instantaneous CSI scenario, but $\alpha_{\max}$ can be replaced with $\tilde{\alpha}_{\max}$ to obtain the equivalent guidelines for statistical CSI:

- If S_{s} is a relay introduced on purpose by the network owner to enhance primary transmission, then $\alpha = 0$, since the primary transmitter receives all the benefits. This takes us back to the case of LTE-A RNs in chapter 3.
- If S_{p} sustains a *Constant Bit Rate* (CBR) service, such as a video broadcast, a rate increase is useless for the primary transmitter. In this case, S_{s} could be rewarded with all the recycled resources ($\alpha = \alpha_{\max}$).
- If both S_{p} and S_{s} are similar *Variable Bit Rate* (VBR) sources, they should share the released resources as fairly as possible. Usually, MAC protocols treat fairness in terms of channel access opportunities, ignoring the rate that each user achieves per channel access event [166]. This is an intelligent approach since, if throughputs were taken into account, giving priority to the best users would result in starvation. On the other hand, assigning more resources to the worst users seeking equal throughputs would heavily hamper the sum-rate. The philosophy of throughput-agnostic fairness is also desirable in our design. Since the goal of CR is to find

spectrum resources, we think it would be fair to allocate half the spectrum gains to each user. Therefore, we suggest setting $\alpha = \alpha_{\max}/2$, so that the primary transmitter is allocated its original capacity plus half the resource gains.

In the following sections we characterize $\alpha_{\max}$ and $\tilde{\alpha}_{\max}$ for different channel conditions. This provides insight into achievable gains, which are of interest regardless of how they are divided. In any case, given our assumptions, the less efficient S_{p} is, the more incentives S_{s} will have to cooperate.

4.3 Benefit Probability

To determine whether cooperation can be beneficial often or just in rare situations, we will analyze the probability that any rate increase can be achieved through cooperation. For this we consider the maximum MI that S_{p} can achieve by cooperating, having instantaneous CSI, when S_{s} does not receive any resources ($\alpha = 0$). Cooperative diversity will be activated according to (4.3), or, equivalently, in the event

$$I_{\mathrm{C}} > I_{\mathrm{D}} \tag{4.8}$$

Next, we characterize statistically the probability of this event for a single-helper network under Rayleigh fading, in which cooperation is based on the DF protocol. We use the exponential distribution of the squared absolute value of the fading coefficient

$$\begin{gathered} u_{ij} \equiv |h_{ij}|^2, \\ h \sim \mathcal{CN}(0, \sigma^2) \Rightarrow u \sim \mathrm{Exp}(\lambda), \lambda = \frac{1}{2\sigma^2}. \end{gathered} \tag{4.9}$$

Using the MI for the DF protocol (2.3), the event that triggers cooperation is

$$\begin{gathered} I_{\mathrm{DF}} > I_{\mathrm{D}} \Leftrightarrow \\ \frac{1}{2}\log_2(1 + \mathrm{SNR}\min(u_{\mathrm{sr}}, u_{\mathrm{sd}} + u_{\mathrm{rd}})) > \log_2(1 + \mathrm{SNR}u_{\mathrm{sd}}) \Leftrightarrow \\ \min(u_{\mathrm{sr}}, u_{\mathrm{sd}} + u_{\mathrm{rd}}) > u_{\mathrm{sd}}(2 + \mathrm{SNR}u_{\mathrm{sd}}) \end{gathered} \tag{4.10}$$

and we compute its probability in the following theorem

Theorem 4.5. *The probability that a CSL system with a single DF-relaying secondary, instantaneous CSI and Rayleigh fading channels relies on cooperation at any given fading* *realization is*

$$P(I_{\mathrm{DF}} > I_{\mathrm{D}}) = \frac{\sqrt{\pi}\lambda_{\mathrm{sd}}}{\sqrt{(\lambda_{\mathrm{rd}} + \lambda_{\mathrm{sr}})\mathrm{SNR}}} e^{\frac{\mu^2}{2}} Q(\mu), \quad \mu = \frac{(\lambda_{\mathrm{rd}} + 2\lambda_{\mathrm{sr}} + \lambda_{\mathrm{sd}})}{\sqrt{2(\lambda_{\mathrm{rd}} + \lambda_{\mathrm{sr}})\mathrm{SNR}}} \tag{4.11}$$

Proof. Appendix 4.A. □

Figs. 4.3, 4.4 and 4.5 show examples of different behaviors. Each shows the benefit probability for the DF relay channel in which the fading parameters of two of the three channels (source to destination (SD), source to relay (SR) and relay to destination (RD)) are fixed, $\sigma_{ij} = 1$, and the third channel has a fading parameter varying between 0.1 and 10 on the axis of ordinates.

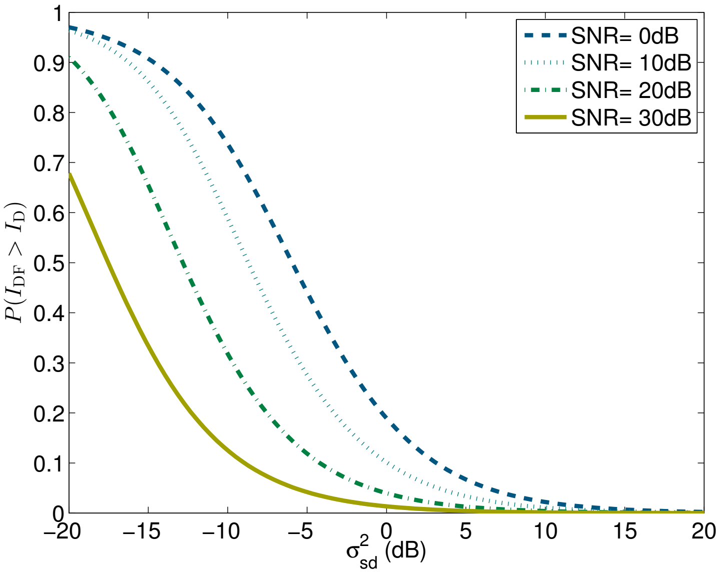


Figure 4.3: Probability of cooperation depending on SD channel statistics.

Fig. 4.3 shows that the probability of cooperation converges asymptotically to 1 as $\sigma^2_{\mathrm{sd}} \to 0$, representing a SD channel much worse than the other two. This is coherent with the fact that completely cutting the direct link would make communications impossible without cooperation. Asymptotic behavior of the function is in this case mostly dominated by the first term $\frac{\sqrt{\pi}\lambda_{\mathrm{sd}}}{\sqrt{(\lambda_{\mathrm{rd}}+\lambda_{\mathrm{sr}})}}$ (recall that $\lambda_{ij} = \frac{1}{2\sigma^2_{ij}}$). The physical interpretation is, naturally, that primary transmitters affected by poor channel conditions usually benefit greatly from cooperation.

In contrast, improvements in the other two channels do not necessarily yield resource gains. For the SR case, $P(I_{\mathrm{DF}} > I_{\mathrm{D}})$ converges to a value < 1 as $\sigma^2_{\mathrm{sr}} \to 0$. Fig. 4.4 shows that, if the source-to-relay channel is good, there may be potential gains, but this channel is less determinant than SD.

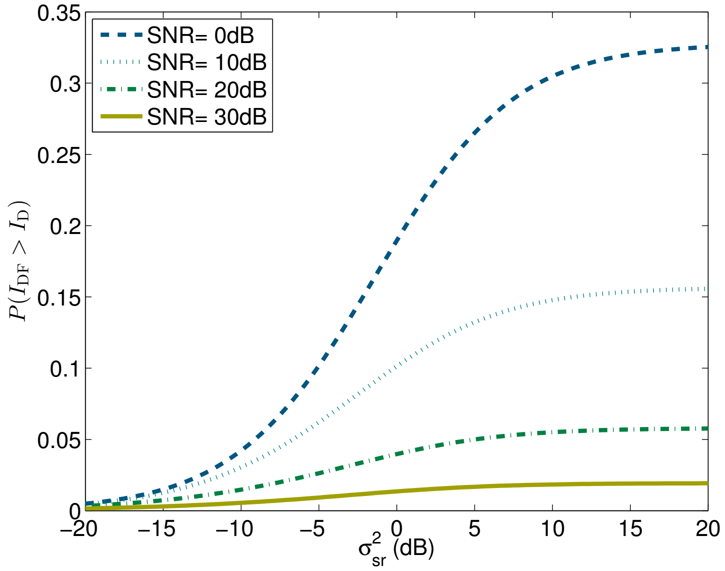


FIGURE 4.4: Probability of cooperation depending on SR channel statistics.

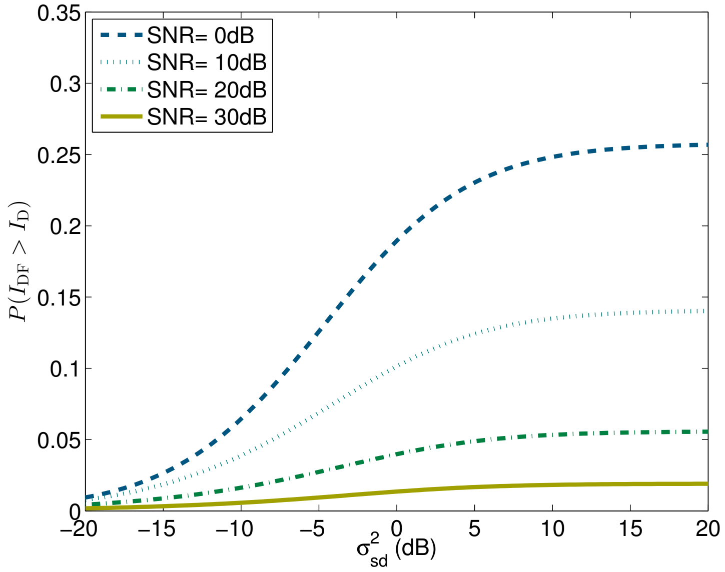


FIGURE 4.5: Probability of cooperation depending on RD channel statistics.

In the RD case, $P(I_{\mathrm{DF}} > I_{\mathrm{D}})$ converges to a value even smaller than $\lambda_{\mathrm{rd}} \to 0$. Fig. 4.4 shows that if the relay-to-destination channel is good, there may be potential gains, but this channel is the least determinant. The difference in asymptotic behavior between SR and RD channels seems to be due to the fact that the first has twice the weight of the second in the numerator of parameter $\mu = \frac{(\lambda_{\mathrm{rd}}+2\lambda_{\mathrm{sr}}+\lambda_{\mathrm{sd}})}{\sqrt{2(\lambda_{\mathrm{rd}}+\lambda_{\mathrm{sr}})\mathrm{SNR}}}$.

Moreover, the influence of transmitted signal energy to noise ratio SNR is appreciable (as suggested by the intuition that low SNR systems tend to benefit more from cooperative diversity). The fact that values of near one are achieved for any value of SNR when primary channels are weak makes the potential of this type of context very appealing.

The asymptotic observations confirm the assumption that the greatest value of CSL lies in facing inefficient spectrum usage by its primary owner.

4.4 Mutual Information Distribution

In the previous section we proved that cooperative diversity gains exist in CSL scenarios with poor primary channels, but we did not quantify those gains. In this section and the following we consider MI as a random variable dependent on channel state. We first obtain the p.d.f. of the MI in the two transmission modes (D and DF). As in section 4.3, we employ a three-node topology in the analysis below. The MI of each protocol X will be noted as I_{X}, its realizations as i_{X}, and the p.d.f.s as $f_{I_{\mathrm{X}}}(i_{\mathrm{X}})$.

Lemma 4.6. *The p.d.f.s of the MI of direct primary transmissions and DF cooperative transmissions are*

$$f_{I_{\mathrm{D}}}(i_{\mathrm{DF}}) = \frac{\ln(2)}{\mathrm{SNR}} 2^{i_{\mathrm{DF}}} e^{-\lambda_{\mathrm{sd}} \frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}} \tag{4.12}$$

and

$$f_{I_{\mathrm{DF}}}(i_{DF}) = \begin{cases} \frac{\ln(2)}{\mathrm{SNR}} 2^{i_{\mathrm{DF}}} \left(1 + (\lambda_{\mathrm{sd}}^2 + \lambda_{\mathrm{sd}}\lambda_{\mathrm{sr}}) \frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}\right) e^{-(\lambda_{\mathrm{sd}}+\lambda_{\mathrm{sr}}) \frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}} & \lambda_{\mathrm{sd}} = \lambda_{\mathrm{rd}} \\ \frac{\ln(2)}{\mathrm{SNR}} 2^{i_{\mathrm{DF}}} \frac{\left[(\lambda_{\mathrm{sd}}\lambda_{\mathrm{rd}}-\lambda_{\mathrm{sd}}\lambda_{\mathrm{sr}}) e^{-(\lambda_{\mathrm{rd}}+\lambda_{\mathrm{sr}}) \frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}} + (\lambda_{\mathrm{rd}}\lambda_{\mathrm{sr}}-\lambda_{\mathrm{sd}}\lambda_{\mathrm{rd}}) e^{-(\lambda_{\mathrm{sd}}+\lambda_{\mathrm{sr}}) \frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}}\right]}{\lambda_{\mathrm{rd}}-\lambda_{\mathrm{sd}}} & \lambda_{\mathrm{sd}} \neq \lambda_{\mathrm{rd}} \end{cases} \tag{4.13}$$

Proof. Appendix 4.B. □

Let us look closely at the shape of these p.d.f.s on Figs. 4.6, 4.7 and 4.8 by comparing the MI for the direct and DF modes covering three types of scenarios. All curves show channel realizations concentrated in "lobes" around the mean, which increases with SNR. They also show that the direct link lobes are wider than those of DF (whose lobes shift less to the right as SNR increases). This is due to the rate-halving nature of the DF protocol. In section 4.5 we further elaborate on the mean of the MI, which can be interpreted as the center of probability mass of the lobes. The thinner lobe of DF is directly related to its outage behavior: this protocol exhibits a larger reduction of outage probability for a given reduction in rate.

In the first example, all parameters σ_{ij} are set to 1, as would be the case in an equilateral triangular network with all links having the same average power. We see that the direct and DF p.d.f.s on Fig. 4.6 concentrate probability mass on lobes. For each SNR value, the corresponding lobe of direct transmission lies to the left (smaller rate) of the corresponding direct transmission p.d.f.s. Decisions based on statistical CSI would select

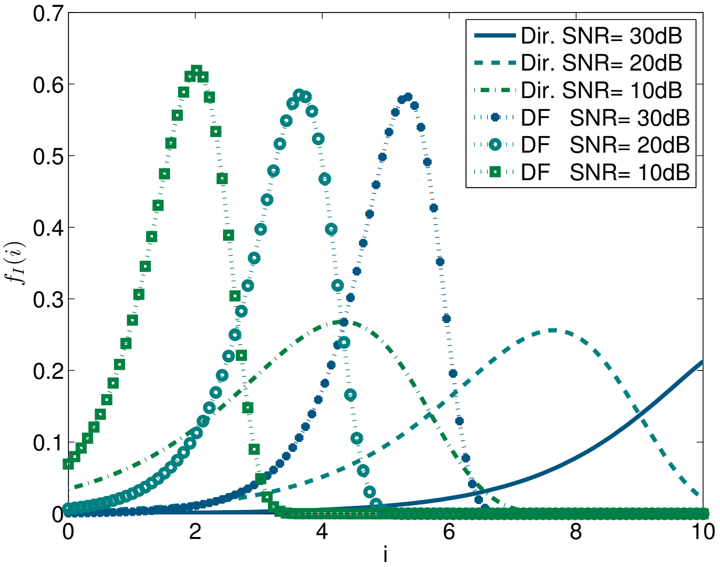


FIGURE 4.6: Mutual information p.d.f. of direct transmission and cooperative DF for an equilateral triangular example layout.

the direct mode in this scenario. Decisions based on instantaneous CSI would rarely opt for cooperative diversity when the direct channel is affected by deep instantaneous fading.

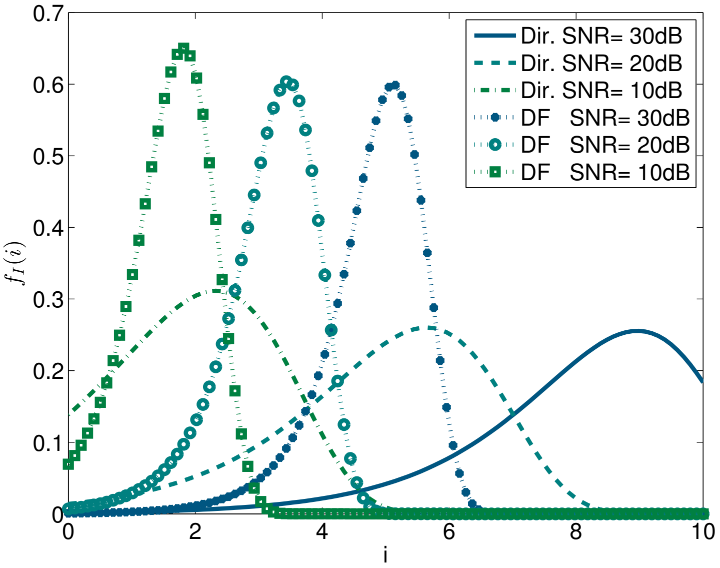


FIGURE 4.7: Mutual information p.d.f. of direct transmission and cooperative DF for a relay-in-the-middle example layout.

The second example of scenario features a layout with $\sigma_{SD} = 0.5\sigma_{SR} = 0.5\sigma_{RD}$ as would correspond to a relay placed in the middle between the source and destination, so that the channels are twice better than the direct channel (+3dB). We can see that the worse direct channel shifts the direct p.d.f.s on Fig. 4.7 to the left while barely altering those of DF. When the SNR is low, the p.d.f.s are highly overlapped. Thus, decisions based on instantaneous CSI would frequently switch between modes. DF lobe mass centers appear still to be slightly to the left compared to the respective direct transmission p.d.f.s. For higher SNRs the behavior is similar to that shown in Fig. 4.6.

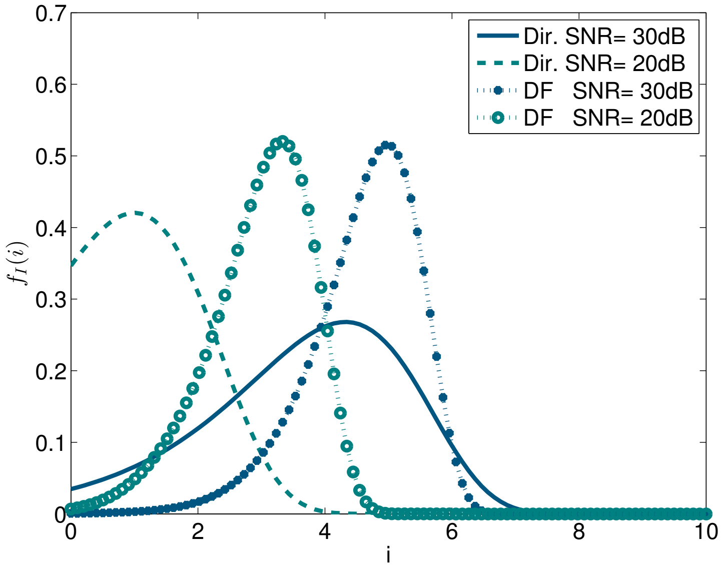


Figure 4.8: Mutual information p.d.f. of direct transmission and cooperative DF for an example of layout with obstacles in the primary transmitter link.

And, finally, the third example of scenario represents a case with a primary link that is much worse than the others, where σ_{SD} is ten times smaller than the corresponding parameters of the other links ($\sigma_{SD} = 0.1\sigma_{SR} = 0.1\sigma_{RD}$). This would be the case, for instance, when obstacles shadow the propagation of the direct transmission (about 20dB). The large degradation of the primary channel hampers direct transmission severely and degrades DF considerably, as shown in Fig. 4.8. The two DF lobes are shifted to the right of the corresponding direct transmission equivalents (we had to omit the 10dB case for the figure to be visible, because it concentrated so much probability mass around 0 that its peak value was much higher than in the other cases). Consequently, decisions based on statistical CSI will select the cooperative mode. Decisions based on instantaneous CSI will select DF in most cases, switching to direct mode on the rare occasions when only the source-to-destination channel is in good instantaneous shape despite its worse average. If we kept increasing SNR, the transition to the behavior shown in Figs. 4.6 and 4.7 would still be observed.

From these plots, we can state that DF MI is more stable against changes in the channel than direct transmission. DF does not produce the thick lower tails of direct transmission, but, on the other hand, it is harder for it to reach the peak MI values seen with direct transmission. The decision scheme based on instantaneous CSI exploits the higher MI values of direct transmission when they happen, and it is not hampered by the lower MI values because in those cases it opts for DF cooperation. For DF MI grows less with SNR (the lobe shifts to the right due to that growth) than for direct transmission, because low-SNR systems tend to benefit more from cooperation. At a higher SNR it becomes more unlikely that the direct channel will be so faded that DF cooperation will be useful. Moreover, the relative positions of the two functions (D and DF) as SNR changes depend on the channel parameter layout: the worse the SD channel is, the more likely cooperation is. Note that the DF lobes lie even to the right of the direct transmission lobes in Fig. 4.8. All these statements confirm the intuitive interpretation of the previous section: the more inefficient the primary transmitter is, the more spectrum can be recycled for secondary transmission at typical SNR values, although the effect is more evident at low SNRs.

4.5 Static Decision-Making With Statistic CSI

As we have observed in the discussion of the p.d.f.s of MI, the benefit of cooperation decisions based on statistical CSI depends on the relative positions of the lobes that contain most of the probability mass of the respective p.d.f.s. A decision scheme based on statistical CSI would be of interest if no instantaneous CSI is available (such as when channels change too fast for the measurement mechanism), if the implementation of the decision-making is static (the decision is taken once for all channel realizations) or in systems with complexity constraints. For the implementation of a decision guided by (4.2) it is necessary to compute the average MI of each protocol. If capacity-achieving protocols are employed for each channel this gives the ergodic capacity.

The following lemma explains how to compute the average capacities:

Theorem 4.7. *The average MIs of direct and DF transmissions under a leasing fraction α are given by*

$$\mathrm{E}\left[I_{\mathrm{D}}\right] = \frac{1}{\ln(2)} e^{\frac{\lambda_{\mathrm{sd}}}{\mathrm{SNR}}} \Gamma\left(0, \frac{\lambda_{\mathrm{sd}}}{\mathrm{SNR}}\right) \tag{4.14}$$

and

$$\mathrm{E}\left[(1-\alpha) I_{\mathrm{DF}}\right] = (1-\alpha)\left[P^{c}\mathrm{E}\left[I_{DF/\epsilon^{c}}\right] + P\mathrm{E}\left[I_{DF/\epsilon}\right]\right] \tag{4.15}$$

where the two terms in (4.15) *are computed using expressions* (4.31) *and* (4.34) *if* $\sigma_{\mathrm{sd}} \neq \sigma_{\mathrm{rd}}$ *or and* (4.32) *and* (4.35) *if* $\sigma_{\mathrm{sd}} = \sigma_{\mathrm{rd}}$.

Proof. Appendix 4.C □

Fig. 4.9 shows the average MI curves for an SNR variation of between 0 dB and 30 dB and setting $\alpha = 0$. The curves represent the three scenarios described in section

4.4: equilateral-triangular, relay-in-the-middle and a lossy primary link. A fourth intermediate loss scenario with $\sigma_{SD} = 0.25$ has been added, to provide a more smooth picture of the evolution of capacities. The standard deviation of source-to-destination channel fading is progressively increased to illustrate the performance loss: the average MI curves shift downwards as the channel worsens. The loss is more pronounced in the case of direct transmission.

Cooperation is useful if average MI of DF exceeds that of direct transmission. I_{D} decreases substantially with SD channel degradation, whereas I_{DF} experiences comparatively smaller reductions. Cooperation therefore takes place in the case of permanent impairments, such as primary link obstacles, rather than in the case of particular fading realizations.

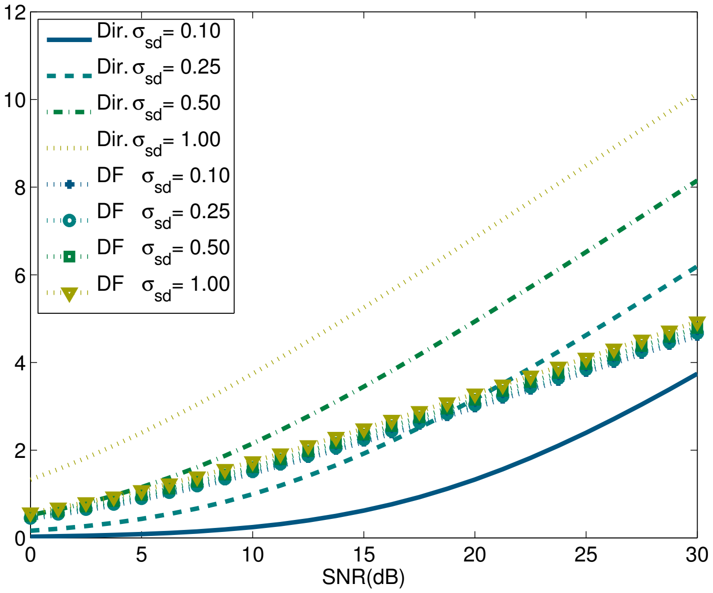


FIGURE 4.9: Average MI of cooperative DF and direct transmission versus SNR, for varying values of σ_{sd}.

It is important to recall that $1-\alpha$ is a constant factor in the derivation of the means, and that it will be selected as $\alpha \in [0, \tilde{\alpha}_{\max}]$. This interval is obtained using expression (4.7) together with (4.14) and (4.15). Fig. 4.10 shows the values of $\tilde{\alpha}_{\max}$ resulting from the cases in the previous examples. Cooperation is useful in all the regions where the maximum static leasing fraction is nonzero. For the SNR limit of $0dB$ in the $\sigma_{\mathrm{sd}} = 0.1$ curve, over 90% of the resources are released. This value decreases considerably with increasing SNR. Resource gains decrease rapidly with improvements in the direct channel, leading to a complete lack of cooperation for $\sigma_{\mathrm{sd}} = 0.5$

Like Fig. 4.7, Fig. 4.10 shows that the direct mode is used for the equilateral triangular and relay-in-the-middle scenarios (thick-dotted and solid lines), since in these cases $\tilde{\alpha}_{\max} = 0$. However, in the case with a worse source-to-destination channel ($\sigma_{SD} = 0.25$

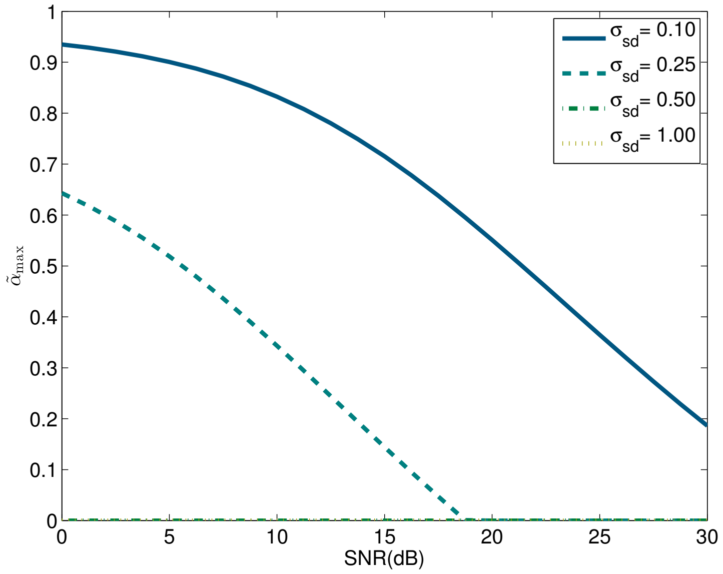


FIGURE 4.10: Maximum static leasing fraction ($\tilde{\alpha}_{\max}$) versus SNR, for varying values of σ_{sd}.

and $\sigma_{SD} = 0.1$) resource gains appear and increase rapidly. Note that, for low SNRs, $\tilde{\alpha}_{\max} = 0.64$ for $\sigma_{SD} = 0.25$, and it approaches 1 for $\sigma_{SD} = 0.1$ (93.4% resource gain). It is noteworthy that the resource gain seen in the case of statistical CSI is due to a primary link that is bad on average, rather than to temporary deep fading events. The low complexity approach of statistical CSI is more adequately labeled as a cooperative routing gain than a cooperative diversity gain, as it obtains resource gains from primary transmitters affected by semi-permanent problems, such as obstacles or deficient antenna orientations, combined with cooperative communications.

4.6 Dynamic Decision-Making with Instantaneous CSI

As we have shown, a major disadvantage of the statistic decision scheme is it tendency to two extremes: good primary channels are simply too good to obtain an average MI gain by always cooperating, and primary channels so bad that they can obtain great gains using static cooperation are unlikely to exist in cognitive scenarios in the first place. After all, if there is such a great potential gain by installing a permanent relay, the primary operator is likely to redesign its network to introduce such a component.

A more likely case to appeal for CSL application is the case where S_{p} decisions are supported by instant knowedge of the channel realizations for transmission. This scheme can be used in systems that can afford extra complexity for a better exploitation of resources. In order to implement decisor (4.3), inequality (4.10) determines the optimal choice:

$$D_{\mathrm{i}} = \arg\max(I_{\mathrm{D}}, I_{\mathrm{DF}}) = \begin{cases} DF & \text{if (4.10) is true} \\ D & \text{otherwise} \end{cases} \tag{4.16}$$

In this scenario, the value of the maximum dynamic leasing fraction α_{max} is different for each channel realization. We treat this value as a random variable (A_{max}) and study its behavior ($f_{A_{\mathrm{max}}}(\alpha_{\mathrm{max}})$). This p.d.f. is divided into two parts: it takes zero values when $I_{\mathrm{C}} \leq I_{\mathrm{D}}$ (cooperation is not beneficial), and positive values when $I_{\mathrm{C}} > I_{\mathrm{D}}$. In the second case, we define the auxiliar random variable A_{eq} and its p.d.f. $f_{A_{\mathrm{eq}}}(\alpha_{\mathrm{eq}})$

Lemma 4.8. *The p.d.f. of the maximum dynamic leasing fraction is*

$$f_{A_{\mathrm{max}}}(\alpha_{\mathrm{max}}) =(1 - P[I_{\mathrm{C}} > I_{\mathrm{D}}])\delta(\alpha_{\mathrm{max}}) + P[I_{\mathrm{C}} > I_{\mathrm{D}}]f_{A_{\mathrm{eq}}/A_{\mathrm{eq}}>0}(\alpha_{\mathrm{max}}) \tag{4.17}$$

Proof. The probability of each side for DF is given by Theorem 4.5 and in the non-zero case the function takes the normalized values of the distribution of positive non-zero values of α_{eq}. □

Unfortunately, for the DF protocol, the p.d.f. of α_{eq} cannot be obtained through simple analytic steps

Lemma 4.9. *The p.d.f.* $f_{A_{\mathrm{eq}}}(\alpha_{\mathrm{eq}})$ *is given by integral*

$$f_{A_{\mathrm{eq}}}(\alpha_{\mathrm{eq}}) = \int_0^{\infty} i_c f_{I_{\mathrm{D}},I_{\mathrm{C}}}((1-\alpha_{\mathrm{eq}})i_c, i_c)di_c \tag{4.18}$$

which for DF is expanded as shown in Appendix 4.D and cannot be solved analytically in general.

As we have obtained an almost-closed expression for $f_{A_{\mathrm{eq}}}(\alpha_{\mathrm{eq}})$, and the range of values of interest is small, numeric integration methods can be employed to evaluate the integral and provide some examples of $f_{A_{\mathrm{max}}}(\alpha_{\mathrm{max}})$. Fig. 4.11 shows some cases and, once again, confirms that CSL should be used when primaries are inefficient, in the sense the that probability mass of α_{max} is accumulated around 1 when $\sigma_{SD} \to 0$, meaning that cooperation is particularly beneficial for bad primaries.

The probability of cooperation is equivalent to $P(\alpha_{\mathrm{max}} > 0)$, and to evaluate aggregate dynamic gains along many channel realizations we can use numeric integration to evaluate $E\left[\alpha_{\mathrm{max}}\right]$ too (not to be confused with $\tilde{\alpha}_{\mathrm{max}}$). Table 4.1 provides the gains for both schemes in the four scenarios in section 4.5: equilateral triangular, relay-in-the-middle and two degrees of high losses in the primary link. For comparison, the last column

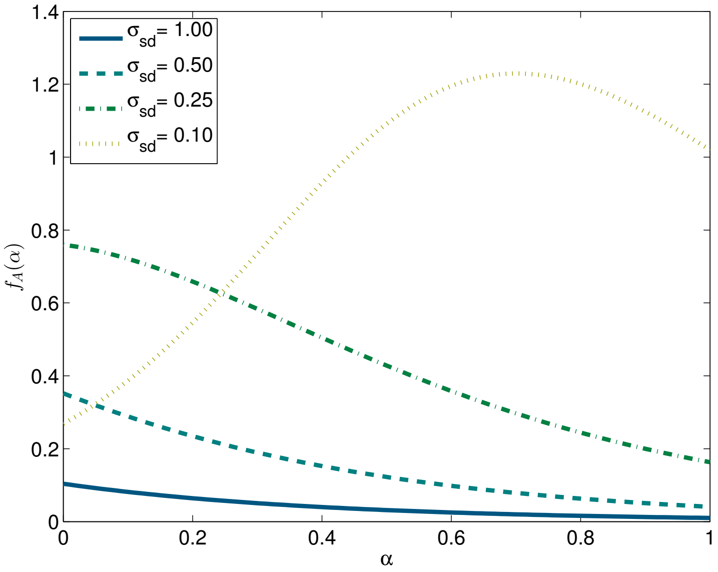


FIGURE 4.11: Distribution of positive α_{max} values in the four channel scenarios described in section 4.5.

TABLE 4.1: Probability of cooperation and average resource gains with instantaneous and statistical CSI

σ_{sd}	SNR	P_{coop}	$E\left[\alpha_{\text{max}}\right] \times 100\%$	$\tilde{\alpha}_{\text{max}} \times 100\%$
1	$0dB$	0.19	8.7%	0%
	$10dB$	0.10	4.0%	0%
	$20dB$	0.04	1.3%	0%
	$30dB$	0.01	0.3%	0%
0.5	$0dB$	0.51	27.0%	0%
	$10dB$	0.33	14.2%	0%
	$20dB$	0.15	4.9%	0%
	$30dB$	0.05	1.4%	0%
0.25	$0dB$	0.82	57.5%	64.3%
	$10dB$	0.72	39.2%	34.3%
	$20dB$	0.45	17.7%	0%
	$30dB$	0.19	5.2%	0%
0.1	$0dB$	0.97	93.1%	93.4%
	$10dB$	0.96	79.8%	83.2%
	$20dB$	0.91	53.5%	55.1%
	$30dB$	0.68	24.0%	18.6%

of the table shows again the benefits of static allocation using $\tilde{\alpha}_{\text{max}}$. Note that for instantaneous CSI some benefits persist for high SNR values and good primaries. In fact, decisions based on instantaneous CSI can always generate nonzero resource gains but, as the primary link improves, these gains decrease to a point at which the channel estimation overhead may be unacceptable.

Even though the development of a MAC protocol is beyond the aim of this thesis, we can give some guidelines by looking at this table. For a given protocol, these analytical results should be checked against an estimation of its overhead. If the gain in the table minus the overhead results in resource losses, it can be concluded that the protocol does not benefit from CSL.

For example, in the relay-in-the-middle channel setting ($\sigma_{SD} = 0.5$), for transmissions at $\mathrm{SNR} = 10dB$, there are resource gains of $\sim 15\%$. Depending on the target MAC, if protocol overhead is below 15%, instantaneous CSI could be employed, whereas decisions based on statistical CSI would discourage cooperation in such a case. In the extreme scenario with $\sigma_{SD} = 0.1, \mathrm{SNR} = 0dB$, where decisions based on statistical CSI are effective, instantaneous CSI-assisted decisions also achieve resource gains above 90%.

From the point of view of relative channel power the approach with instantaneous CSI is suitable in more cases than the statistical CSI approach. It can provide similarly large gains from primary transmitters with permanent problems, and it can also extract some resource gains from the temporary degradation of good primary channels. This last behavior is akin to the philosophies of cognitive radio and cooperative diversity, which are oriented to take advantage on varying channel conditions.

4.7 Summary

The aim of cooperative diversity is to improve the rate stability of radio communications without the need for bulky antennas in portable devices. Peer cooperation in spectrum domains with heterogeneous priorities may boost cooperative diversity as an enabler for spectrum leasing in cognitive radio.

Previous CSL models have focused on negotiation or competition to achieve a sharing policy for the resource gains that satisfies all the parts. We focus instead on an alternate point of view to determine the social resource gains achievable by cooperation. Our model represents a scenario where secondary transmitters naturally tend to help inefficient spectrum owners (i.e. primary transmitters) to mutual benefit. Since we study the problem from the point of view of aggregate spectrum gains, as a social gain, game-theoretic strategies are unnecessary to quantify the appeal of CSL for future systems.

We have reviewed literature on physical and MAC layers for cooperative diversity that can be easily adapted to spectrum leasing according to the requirements of our model.

Two decision-making schemes are considered: a statistical CSI scheme and an instantaneous CSI scheme. We distinguish between the concepts of rate gains and resource gains and provide an analytical characterization of cooperative diversity gain. We then discuss the channel conditions in which the MI of a cooperative channel improves that of the direct channel for the two CSI schemes.

We have also formulated the statistical distribution of MI and used this to derive the average MI gain of the statistical CSI scheme and the p.d.f. of resource gains for the instantaneous CSI scheme. Even though there are no resource gains in the first scheme when the primary channel is good, these gains increase considerably when the primary channel is bad. The second scheme produces gains in all setups, and these increase progressively as primary channels get worse.

The resource gains resulting from the statistical CSI scheme are mostly related to persistent impairments, such as those due to obstacles or poor installations, whereas those resulting from the instantaneous CSI scheme depend both on permanent impairments and temporary fading. Both schemes provide large resource gains when the primary channel is bad. For many applications, statistical CSI would be sufficient because it would require lower signaling overhead than instantaneous CSI, and yet offer similar gains. However, instantaneous CSI can offer moderate gains for many more channel distribution scenarios, and thus its utilization by some MAC protocols should not be ruled out. We conclude that cooperative communications have a potential for use in spectrum reutilization in future cognitive wireless networks.

The content in this chapter is an extended and actualized version of a paper published in IEEE Transactions on Wireless Communications [6].

Appendix 4.A Calculation of Benefit Probability

$$\begin{aligned}
P(I_{\mathrm{DF}} > I_{\mathrm{D}}) &= P(\min(u_{\mathrm{sr}}, u_{\mathrm{rd}} + u_{\mathrm{sd}}) > u_{\mathrm{sd}}(2 + \mathrm{SNR}u_{\mathrm{sd}})) \\
&= \int_{-\infty}^{\infty} P(u_{\mathrm{rd}} > (x + \mathrm{SNR}x^2)) \times P(u_{\mathrm{sr}} > (2x + \mathrm{SNR}x^2)) f_{u_{\mathrm{sd}}}(x) dx \\
&= \lambda_{\mathrm{sd}} \int_{0}^{\infty} e^{-(\lambda_{\mathrm{rd}}+2\lambda_{\mathrm{sr}}+\lambda_{\mathrm{sd}})x+(\lambda_{\mathrm{rd}}+\lambda_{\mathrm{sr}})\mathrm{SNR}x^2)} dx
\end{aligned} \tag{4.19}$$

The following integration pattern is used:

$$\int_0^{\infty} e^{-(C_1 x + C_2 x^2)} dx = \frac{\sqrt{\pi} e^{\frac{\mu^2}{2}} Q(\mu)}{\sqrt{C_2}}, \quad \mu = \frac{C_1}{\sqrt{(2C_2)}} \tag{4.20}$$

where $Q(x) = \int_x^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$ is the upper tail integral of the normal distribution. Replacing $C_1 = \lambda_{\mathrm{rd}} + 2\lambda_{\mathrm{sr}} + \lambda_{\mathrm{sd}}$ and $C_2 = (\lambda_{\mathrm{rd}} + \lambda_{\mathrm{sr}})\mathrm{SNR}$ gives (4.11).

Appendix 4.B Mutual Information p.d.f.

For direct transmission, we use:

$$f_{I_{\mathrm{D}}}(i_{\mathrm{D}}) = \frac{\partial P(I_{\mathrm{D}} < i_{\mathrm{D}})}{\partial i_{\mathrm{D}}} = \frac{\ln(2)}{\mathrm{SNR}} 2^{i_{\mathrm{D}}} f_{u_{\mathrm{sd}}}(\frac{2^{i_{\mathrm{D}}}-1}{\mathrm{SNR}}) \tag{4.21}$$

For DF, we start by avoiding the minimization in the logarithm by using the partition property [167]. We define the event $\epsilon \equiv u_{\mathrm{sr}} > u_{\mathrm{sd}} + u_{\mathrm{rd}}$ and ϵ^c representing its complementary event. For compactness, let $P = P[\epsilon]$ and $P^c = 1 - P$.

$$f_{I_{\mathrm{DF}}}(i_{DF}) = P f_{I_{\mathrm{DF}}/\epsilon}(i_{DF}) + P^c f_{I_{\mathrm{DF}}/\epsilon^c}(i_{DF}), \tag{4.22}$$

then we replace the mutual information in the two terms by their channel gain equivalents

$$\begin{aligned} f_{I_{\mathrm{DF}}/\epsilon^c}(i) &= f_{u_{\mathrm{sr}}/\epsilon^c}\left(\frac{2^{2i}-1}{\mathrm{SNR}}\right)\frac{\partial \frac{2^{2i}-1}{\mathrm{SNR}}}{\partial i} \\ f_{I_{\mathrm{DF}}/\epsilon}(i) &= f_{u_{\mathrm{sd}}+u_{\mathrm{rd}}/\epsilon}\left(\frac{2^{2i}-1}{\mathrm{SNR}}\right)\frac{\partial \frac{2^{2i}-1}{\mathrm{SNR}}}{\partial i}, \end{aligned} \tag{4.23}$$

Finally we obtain the conditional channel p.d.f.s sing the Bayes theorem [167].

$$\begin{aligned} f_{u_{\mathrm{sd}}+u_{\mathrm{rd}}/\epsilon}(x) &= \frac{1 - F_{u_{\mathrm{sr}}}(x)}{P} f_{u_{\mathrm{sd}}+u_{\mathrm{rd}}}(x) \\ f_{u_{\mathrm{sr}}/\epsilon^c}(x) &= \frac{1 - F_{u_{\mathrm{sd}}+u_{\mathrm{rd}}}(x)}{P^c} f_{u_{\mathrm{sr}}}(x). \end{aligned} \tag{4.24}$$

It is not necessary to obtain the probabilities of ϵ because they appear in the denominator here and therefore are canceled out when we replace everything back in the weighted sum. Making the substitution $x = (\frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}})$:

$$\begin{aligned} f_{I_{\mathrm{DF}}}(i_{DF}) = \frac{\ln(2)}{\mathrm{SNR}} 2^{i_{\mathrm{DF}}} \Bigg[&\left(1 - F_{u_{\mathrm{sr}}}(\frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}})\right) f_{u_{\mathrm{sd}}+u_{\mathrm{rd}}}(\frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}) \\ &+ \left(1 - F_{u_{\mathrm{sd}}+u_{\mathrm{rd}}}(\frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}})\right) f_{u_{\mathrm{sr}}}(\frac{2^{i_{\mathrm{DF}}}-1}{\mathrm{SNR}}) \Bigg] \end{aligned} \tag{4.25}$$

The p.d.f. and Cummulative probability Density Function (C.D.F.) of the sum of two exponentials can be obtained through the convolution of two exponential p.d.f.s [167]

$$f_{u_{\mathrm{sd}}+u_{\mathrm{rd}}}(x) = \begin{cases} \lambda_{\mathrm{sd}}^2 x e^{-\lambda_{\mathrm{sd}} x} & \lambda_{\mathrm{sd}} = \lambda_{\mathrm{rd}} \\ \frac{\lambda_{\mathrm{sd}}\lambda_{\mathrm{rd}}}{\lambda_{\mathrm{rd}}-\lambda_{\mathrm{sd}}}\left[e^{-\lambda_{\mathrm{rd}} x} - e^{-\lambda_{\mathrm{sd}} x}\right] & \lambda_{\mathrm{sd}} \neq \lambda_{\mathrm{rd}} \end{cases} \tag{4.26}$$

$$F_{u_{\rm sd}+u_{\rm rd}}(x) = \begin{cases} 1 - e^{-\lambda_{\rm sd}x}(1+\lambda_{\rm sd}x) & , \lambda_{\rm sd} = \lambda_{\rm rd} \\ 1 - \frac{\left[\lambda_{\rm rd}e^{-\lambda_{\rm sd}x} - \lambda_{\rm sd}e^{-\lambda_{\rm rd}x}\right]}{\lambda_{\rm rd}-\lambda_{\rm sd}} & , \lambda_{\rm sd} \neq \lambda_{\rm rd} \end{cases} \tag{4.27}$$

Combining everything and simplifying the expression gives (4.13).

Appendix 4.C Average Mutual Information

The average MI for the direct transmission is:

$$\begin{aligned} \mathrm{E}\left[I_{\rm D}\right] &= \int_{-\infty}^{\infty} \log_2(1+\mathrm{SNR}x) f_{u_{\rm sd}}(x)dx \\ &= \frac{1}{\ln(2)} \int_0^{\infty} \ln(1+\mathrm{SNR}x)\lambda_{\rm sd}e^{-\lambda_{\rm sd}x}dx \end{aligned} \tag{4.28}$$

Integrating by parts it can be compacted to an incomplete gamma function:

$$\mathrm{E}\left[I_{\rm D}\right] = \frac{1}{\ln(2)} e^{\frac{\lambda_{\rm sd}}{\rm SNR}} \Gamma\left(0, \frac{\lambda_{\rm sd}}{\rm SNR}\right) \tag{4.29}$$

And for the cooperative DF protocol, the expression is:

$$E[(1-\alpha)I_{DF}] = (1-\alpha)\left[P^c E[I_{DF/\epsilon^c}] + PE[I_{DF/\epsilon}]\right]$$

We calculate the two parts of the partitioned expression as follows:

4.C.1 First Term

The first term of the MI is:

$$P^c \mathrm{E}\left[I_{DF/\epsilon^c}\right] = P^c \int_0^{\infty} \frac{1}{2} \log_2(1+\mathrm{SNR}x) f_{sr/\epsilon^c}(x)dx \tag{4.30}$$

Solving it for $\lambda_{\rm sd} \neq \lambda_{\rm rd}$ integrating by parts, the expression can be written as:

$$\frac{\lambda_{\rm sr}}{2\ln(2)(\lambda_{\rm rd}-\lambda_{\rm sd})} \times \left[\frac{\lambda_{\rm rd} G\left(\frac{(\lambda_{\rm sr}+\lambda_{\rm sd})}{\rm SNR}\right)}{\lambda_{\rm sr}+\lambda_{\rm sd}} - \frac{\lambda_{\rm sd} G\left(\frac{(\lambda_{\rm sr}+\lambda_{\rm rd})}{\rm SNR}\right)}{\lambda_{\rm sr}+\lambda_{\rm rd}} \right] \tag{4.31}$$

with $G(x) = e^x\Gamma(0,x)$.

On the other hand, for $\lambda_{\rm sd} = \lambda_{\rm rd} = \lambda$

$$\frac{\lambda_{\rm sr}}{2\ln(2)(\lambda_{\rm sr}+\lambda)} \times \left[\left(\frac{\lambda}{\lambda+\lambda_{\rm sr}} + 1 - \frac{\lambda}{\rm SNR} \right) G\left(\frac{(\lambda_{\rm sr}+\lambda)}{\rm SNR}\right) + \frac{\lambda}{(\lambda_{\rm sr}+\lambda)} \right] \tag{4.32}$$

4.C.2 Second Term

The second term of the MI is:

$$\begin{aligned} P\mathrm{E}\left[I_{DF/\epsilon}\right] &= P\int_{-\infty}^{\infty}\frac{1}{2}\log_2(1+\mathrm{SNR}x)f_{u_{\mathrm{rd}}+u_{\mathrm{sd}}/\epsilon}(x)dx \\ &= \frac{1}{2\ln(2)}\int_{-\infty}^{\infty}\ln(1+\mathrm{SNR}x)[1-F_{u_{\mathrm{sr}}}(x)]f_{u_{\mathrm{rd}}+u_{\mathrm{sd}}}(x)dx \end{aligned} \tag{4.33}$$

When $\lambda_{\mathrm{sd}} \neq \lambda_{\mathrm{rd}}$, the integral is very similar to the first term:

$$\frac{\lambda_{\mathrm{sd}}\lambda_{\mathrm{rd}}}{2\ln(2)(\lambda_{\mathrm{rd}}-\lambda_{\mathrm{sd}})}\times\left[\frac{G\left(\frac{(\lambda_{\mathrm{sr}}+\lambda_{\mathrm{sd}})}{\mathrm{SNR}}\right)}{\lambda_{\mathrm{sr}}+\lambda_{\mathrm{sd}}}-\frac{G\left(\frac{(\lambda_{\mathrm{sr}}+\lambda_{\mathrm{rd}})}{\mathrm{SNR}}\right)}{\lambda_{\mathrm{sr}}+\lambda_{\mathrm{rd}}}\right] \tag{4.34}$$

Otherwise, for $\lambda_{\mathrm{sd}} = \lambda_{\mathrm{rd}} = \lambda$, with similar integration:

$$\frac{\lambda}{2\ln(2)(\lambda+\lambda_{\mathrm{sr}})}\times\left[\left(\frac{\lambda}{(\lambda+\lambda_{\mathrm{sr}})}-\frac{\lambda}{\mathrm{SNR}}\right)G\left(\frac{(\lambda_{\mathrm{sr}}+\lambda)}{\mathrm{SNR}}\right)+\frac{\lambda}{(\lambda+\lambda_{\mathrm{sr}})}\right] \tag{4.35}$$

Appendix 4.D P.d.f. of Equivalent Leasing Fraction α_{eq}

We begin by using (4.5) in the chain rule

$$f_{\alpha_{\mathrm{eq}}}(x) = \frac{\partial P[\alpha_{\mathrm{eq}} < x]}{\partial x} = f_{\frac{I_D}{I_C}}(1-x)$$

The p.d.f. of the division of random variables is [167]

$$f_{\frac{I_D}{I_C}}(z) = \int_0^{\infty} i_c f_{I_D,I_C}(zi_c, i_c)di_c$$

I_D and I_C are mutually dependent, because the component of I_C that depends on the direct link fading will have the same behavior. Taking DF as cooperative protocol, the joint p.d.f. $f_{I_D,I_{DF}}(i_D, i_{DF})$ can be obtained by splitting, using again the event ϵ to apply

$$f_{I_D,I_{DF}}(i_D, i_{DF}) = Pf_{I_D,I_{DF}/\epsilon}(i_D, i_{DF}) + P^c f_{I_D,I_{DF}/\epsilon^c}(i_D, i_{DF})$$

Where the Bayes theorem leads again to the cancellation of the multiplications by P and P^c. We define the two-dimensional transformations $T_1(x,y) = (\frac{2^x-1}{\mathrm{SNR}}, \frac{2^{2y}-1}{\mathrm{SNR}} - \frac{2^x-1}{\mathrm{SNR}})$ and $T_2(x,y) = (\frac{2^x-1}{\mathrm{SNR}}, \frac{2^{2y}-1}{\mathrm{SNR}})$, and construct the p.d.f.s of interest by multiplying the determinant of the Jacobian matrices of the transformation with the original p.d.f [167]:

$$\begin{aligned}
f_{I_D,I_{DF}/\epsilon}(i_D,i_{DF}) =& f_{u_{\text{sd}},u_{\text{rd}}/\epsilon}(T_1(i_D,i_{DF})) \times \det(J[T_1(x,y)]) \\
f_{I_D,I_{DF}/\epsilon^c}(i_D,i_{DF}) =& f_{u_{\text{sd}},u_{\text{sr}}/\epsilon^c}(T_2(i_D,i_{DF})) \times \det(J[T_2(i_D,i_{DF})]) \\
\det(J[T_1(x,y)]) =& \left(\frac{\ln(2)}{\text{SNR}}\right)^2 2^{(x+2y+1)} \\
\det(J[T_2(x,y)]) =& \left(\frac{\ln(2)}{\text{SNR}}\right)^2 2^{(x+2y+1)}
\end{aligned}$$

As u_{sd}, u_{sr} and u_{rd} are independent, we can finally decompose $f_{u_{\text{sd}},u_{\text{rd}}}(x,y) = f_{u_{\text{sd}}}(x)f_{u_{\text{rd}}}(y)$ and likewise for the other p.d.f.s. Taking this into account and using the Bayes rule, the conditional distributions are:

$$\begin{aligned}
Pf_{u_{\text{sd}},u_{\text{rd}}/\epsilon}(x,y-x) =& [1-F_{u_{\text{sr}}}(y)]f_{u_{\text{sd}}}(x)f_{u_{\text{rd}}}(y-x) \\
P^c f_{u_{\text{sd}},u_{\text{sr}}/\epsilon^c}(x,y) =& [1-F_{u_{\text{rd}}}(y-x)]f_{u_{\text{sd}}}(x)f_{u_{\text{sr}}}(y)
\end{aligned}$$

The exponential distributions hold only for positive values. Most coefficients in the transformation are increasing functions of I_{DF}, which is always positive, but there are two points in the previous expressions where a subtraction appears, leading to a negative outlier which needs careful integration limits.

$$1 - F_{u_{\text{rd}}}(y-x) = \begin{cases} e^{-\lambda_{\text{rd}}(y-x)} & y > x \\ 1 & y < x \end{cases}$$

$$f_{u_{\text{rd}}}(y-x) = \begin{cases} \lambda_{\text{rd}} e^{-\lambda_{\text{rd}}(y-x)} & y > x \\ 0 & y < x \end{cases}$$

In our particular case $y - x = (2^{2I_{DF}} - 2^{(1-\alpha_{\text{eq}})i_{DF}})/\text{SNR}$ and therefore the second expressions must be used for values below $\alpha_{\text{eq}} < -1$.

Substituting recursively in the expressions above up to (4.18), we obtain a sum of exponential integrals with polynomial exponents of order $2 - \alpha_{\text{eq}}$ that only has a closed solution for some integer cases of the exponent. Therefore, the calculation must stop at this point and the evaluation of the p.d.f. is performed with numerical integration techniques.

Chapter 5

Practical Cooperative Spectrum Leasing: Third-Party Relaying in M2M LTE-A

Contents

5.1 Introduction

The first part of this thesis we discussed the current state of the art of cooperative communications, and particularly in Chapter 3 we described the implementation of multi-hop communications through relays in LTE-A. As we commented, the new generation of cellular data networks will achieve the IMT-A requirements for 4G standards by relying on a series of improvements such as CA, MU-MIMO, femtocells, and RNs.

However, in the current 3GPP standard cooperation appears in its simplest multi-hop form and it is still very limited. The standard does not account for the possibility of cooperation between UE devices, cooperation of multiple RNs to form distributed virtual antenna arrays, or CR techniques for third-party secondary devices to cooperate with the network.

Work on cooperation between UEs and the formation of distributed virtual antenna arrays are partially under way in LTE-B by means of LTE-Direct and multiple-location, single-controller-entity groups of antennas named Remote Radio Heads (RRH). However, CR, as an enabler to improve spectral efficiency, and specially CSL as discussed in Chapter 4, are not considered yet in upcoming standard drafts.

The advent of each new cellular wireless communication generation brings exponential increments in the volume of carried traffic, the number of devices and their density, and the heterogeneity of applications as well. Wireless communications are heading towards all-IP designs where any application may be supported by generic transport protocols, whereas the traditional voice service is taken to “over the top” carriers such as voice over IP (VoIP). In addition, the centralized architecture of traditional services is not suitable for new types of traffic as Peer-to-Peer (P2P) exchanges (a trend which is already being acknowledged by the development of LTE-Direct).

New communication services not oriented to inter-personal exchanges are also a driver force that will fundamentally alter the requirements of future networks. For example, Machine-to-Machine (M2M) communications, within the *Internet of Things* (IoT) [44], involve for billions of new devices with different needs than human users. The analysis in [45] shows the differences of traffic and mobility patterns of human and M2M communications. For example, most M2M devices are less mobile than smartphones, and sometimes even completely static.

Within this diversity of future possibilitys, in this chapter we assess the contribution of cooperative communications to the future evolution of current wireless standards. Particularly, we are interested in discussing whether it is possible to implement the concept of CSL on top of the current cellular standard, as discussed in the previous chapter, in a straightforward way. We evaluate a model where a third-party —which could potentially be some secondary company trying to set up P2P, M2M or SG services using CR—, employing a standard LTE-A RN device with the current version of the standard, offers its aid to a traditional LTE-A network in exchange for a fraction of its spectrum resources, following CSL principles.

We are interested in determining whether such third-party CSL RNs are worth the implementation effort with current LTE technology. For this we evaluate the increment in spectral efficiency on a LTE-A cell due to the introduction of relays, and calculate how many leased spectral resources the third-party RNs can receive as a reward. To account for future improvement, we assume that the implementation-related problems of LTE RNs reported on Chapter 3 can be corrected and that the performance of RNs will grow to approach the theoretical limits in [28, 29, 32].

Our results show that even if these near-optimal RNs are considered, a single CSL RN can only obtain a limited resource gain, mostly due to the fact that a single RN only serves a small area of the cell. However, using the admission-control algorithm we designed in 3.6, it is possible to take advantage of large RN populations, from whose many small contributions the LTE-A network obtains cumulative gains. The improvement is moderate, but we show it is sufficient to support future low-rate applications such as M2M. Due to the fact that, in this scenario, CSL requires massive numbers of RNs to be of interest, and even then it is only relevant for new M2M-like servicess, we conclude that in current LTE-A networks will remain for now a supporting technology for niche application. It will still be necessary to create new wireless standards to lay out the conditions for a CSL-assisted human communications.

The rest of this chapter is organized as follows: In Section 5.2 we describe our LTE-A third-party relaying scheme. In Section 5.3 we compute the RB gain of a single third-party CSL relay. Section 5.4 extends the analysis for multiple relays coexisting within the same cell. In Section 5.5 we propose some use-cases. Finally, Section 5.6 concludes the capter.

5.2 A Model for Third-party RNs Operation in LTE

We consider the LTE-A relaying described in Chapter 3 is utilized to implement CSL as describes in Chapter 4. In this scenario:

- A primary DeNB cell holds the spectrum rights, but in some cases UE sare located far away or in deep shadowed areas and the spectral efficiency of the direct link is low.
- An unlicensed wireless device, implemented according to the standard RN specification, experiences more favorable propagation conditions.

- The DeNB and the candidate RN negotiate during the `relayAttachmentProcedure`. If there are CSL resource gains, a fraction of saved resources is released to the secondary for its own use as a reward. The leasing is implemented as a DeNB authorization to the RN to employ a certain set of RBs in each LTE subframe for its own goals.

The theoretical framework of CSL is covered in Chapter 4. We quantify the number of RBs that a eNB needs to serve a specific UE, and compare it with the number of RBs it would need to provide the same rate using a DeNB-NR-UE two-hop route. The difference between these two quantities consists of the RB savings in the LTE-A cell due to the introduction of RNs. As shown in our analysis, these RB gains in the downlink of a LTE-A cell are proportional to the amount of spectrum the secondaries are rewarded with. Link spectral efficiency is evaluated using the approximation in [32], based on an adjusted Shannon capacity formula. Previous authors have sutudied Amplify-and-Forward relaying [31] or non-orthogonal Decode and Forward (DF) Relays, but in this chapter we focus on *orthogonal DF* relays for simplicity.

Our scenario comprises a primary three-link communication as in Fig. 5.1 plus RN secondary transmissions in some leased RBs. Topologically, we consider a simplified LTE-A circular cell of radius R_{MAX} with the eNB at the origin of coordinates $(0,0)$. The eNB allocates resources to serve each UE at position (x,y), and at any point $(x_\mathrm{r}, y_\mathrm{r})$ there is a RN, which may or may not be employed for transmission towards each UE. UE devices access the channel in turns, scheduled by the eNB, and we assume their spatial distribution is homogeneous, so it is sufficient to study achievable spectral efficiency for one UE device at all points in the space. Achievable rates at each link depend on distance. The eNB-UE distance is $r = \sqrt{x^2+y^2}$, the eNB-RN distance is $r_\mathrm{r} = \sqrt{x_\mathrm{r}^2 + y_\mathrm{r}^2}$ and the RN-UE distance is $d = \sqrt{(x-x_\mathrm{r})^2 + (y-y_\mathrm{r})^2}$.

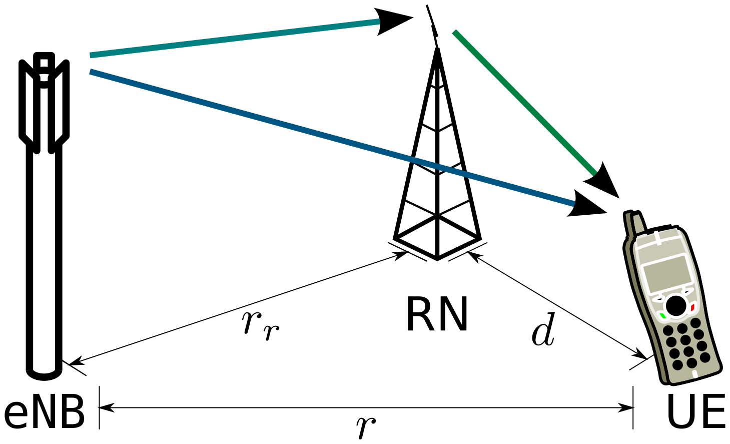


FIGURE 5.1: LTE-A user downlink with a relay.

For simplicity, we omit some key issues of cellular networks from Chapter 3, such as scheduling, realistic user distribution and interference management. This assumption is of course unrealistic, as for instance cellular networks are well known to be interference-limited rather than noise-limited. However, we consider that those factors are secondary for the purpose of a first rought approximation of the potential of CSL in LTE-A. Moreover, the behavior of mobility and interference depends on the specific applications for where the CSL RN is applied, and they may change strongly between P2P, M2M, SG...

Unlike Chapter 3 and the model in [29], our model in this chapter does not include an interference component. In fact, the previously described cell setup has only one eNB transmitting in DL and an orthogonal RN, and therefore there are no interfering entities. We made such a strong simplification of the problem setup to obtain a fast analysis of the potential of CSL in LTE-A. In this chapter, we are only interested in describing upper-layer decision schemes that lead to cognitive RB saving, enumerating tools within the standard to implement CSL, and obtaining a rough estimation of the saving to understand the essential CSL possibilities.

When we evaluate the improvement in spectral efficiency by introducing a RN, which depends on SNR and, ultimately, on distance, the addition of interference from other cells, treated as noise, would only degrade the SNR in all regions of space and the relative difference between BS coverage and RN coverage would be, at best, scaled. The incorporation of additional interference due to the introduction of relays and non-orthogonal transmission would just degrade channels even further. The overall result can only be a poorer gain in the range of the results presented here; and thus a complete analysis would be more confusing to achieve, at best, a similar conclusion. Instead we prefer to focus on discussing the upper-layer decision-making that leads to CSL RB gains and a rough estimation of how they are shared with the secondary. The readers interested in a profound analysis of practical implementations of LTE-A relaying may refer to Chapter 3 and references therein.

In the scenario under study it is necessary to modify the LTE-A RN architecture to deal with RNs owned by third-parties. The LTE-A standard specifies the `RN attach procedure` for a DeNB [16]. The same procedure should be adequate for third-party relay attachment, with the addition of appropriate admission control criteria (i.e. a DeNB must reject relay attachment if the third party RN is not authorized or does not return any gain). The standard leaves the criteria for RNs to start serving UEs open. Since RNs behave like eNBs, UE transitions to RN service are implented through handover procedures. We consider that the DeNB has sufficient knowledge to consider CSL principles in this decision, since it acts as a proxy of the `S1` and `X2` interfaces of the

RNs, and thus has extensive information about the RN and its UEs. Since CSL seeks to free resources, the implementation must consist in the DeNB activating a `X2`-interface handover procedure towards the RN for those UEs that would use fewer RBs if served through the relay, thus producing net resource gains.

The criterion to use the RN is the following: let $\rho_{\mathrm{d}}(r)$ (b/s/Hz) be the achievable spectral efficiency of the direct link. Similarly, let $\rho_{\mathrm{r}}(r_{\mathrm{r}})$ and $\rho_a(d)$ be the respective achievable spectral efficiencies of the relay and access links. The modified Shannon equation of [32] produces a good fit to compute the correspondence between SINR values and spectral efficiency (3.12).

The SNR of each link is modeled with the link budget:

$$SNR(dB) = P_{\mathrm{r}}(dB) - PL(dB) - NP_{\mathrm{r}}(dB) \tag{5.1}$$

where $P_{\mathrm{r}} = P_t + G_t + G_{\mathrm{r}}$ considers the power at the transmitter and the antenna gains (in dB), and $NP_{\mathrm{r}} = NP_0 + 10log_{10}(B) + NF$ is the noise power received, computed from the noise power spectral density, the noise bandwidth and the noise figure of the receiver. PL is the path loss, which grows with distance. These parameters are summarized in Table 5.1.

Table 5.1: Propagation parameters

Parameter	Value	Unit
BS transmit power P_{BS}	46	dBm
RN transmit power P_{RN}	30	dBm
BS-UE path loss PL_{BS-UE}	$128.1 + 37.6\log_{10}(R)$	dB(Km)
BS-RN path loss PL_{BS-UE}	$124.1 + 37.6\log_{10}(R)$	dB(Km)
RN-UE path loss PL_{BS-UE}	$140.7 + 36.7\log_{10}(R)$	dB(Km)
BS tx. antenna gain $G_{t,BS}$	14	dBi
RN tx. antenna gain $G_{t,RN}$	5	dBi
RN rx. antenna gain $G_{r,RN}$	7	dBi
UE rx. antenna gain $G_{r,UE}$	0	dBi

In [29] two scenarios are discussed: urban (Inter Station Distance [ISD]= 500 m) and suburban (ISD = 1732 m). We employ its simplest path-loss model, represented in table 5.1, for several reasons: it is a conservative worst-case approach, it has no ISD or interference assumptions, and it is highly analytically tractable. We simply consider a single cell operating in a noise-limited scheme, so it makes more sense to use a path loss model without dependence on the ISD parameter. BSs and RNs are placed outdoors and UEs are placed indoors, and all links are Non-Line-Of-Sight (NLOS). The path-loss model has an additional margin to account for shadowing and rapid fading.

Combining (3.12) and (5.1) we can finally formulate ρ_x as a function of distance r_x for each link $x \in \{\mathrm{d}, \mathrm{r}, \mathrm{a}\}$, grouping all link budget parameters in two constants: $\mathrm{SNR}_{0,x}$, containing received power, noise and the constant component of PL; and α_x, the path loss exponent with distance.

$$\rho_x(r_x) \simeq \eta_W \log_2(1 + \eta_S^{-1} \mathrm{SNR}_{0,x} |r_x|^{\alpha_x}) \tag{5.2}$$

Note that (3.15) in Chapter 3, which is identical to (5.2), was designed employing a different link budget model. But, due to the facts that in both cases we applied the Shannon approximation (3.12), and that all constants in the link budget are hidden in constant $\mathrm{SNR}_{0,x}$, both spectral efficiency calculations are identical except for the tipical values of their constants.

From [28] it is known that the end-to-end spectral efficiency of a half-duplex optimally time-multiplexed DF Type 1 RN is (3.3) and, thus, optimal switching between the two modes results in $\rho_{\max}$ as defined in (3.11). We can compute the spectral efficiency gain given each pair of UE and RN locations as:

$$\Delta\rho(r, r_\mathrm{r}, d) = \rho_{\max} - \rho_\mathrm{d} = \max(\rho_{\mathrm{e2e}}(r, r_\mathrm{r}, d) - \rho_\mathrm{d}(r), 0) \tag{5.3}$$

Each link x (direct, access or relay) experiences a different path loss environment and, since the relay is a OSI Layer 3 device, each path loss induces in each link an independently achieved transmission rate (R_x). That achieved rate is the product of the spectral efficiency (ρ_x) by the number of assigned resources (n_x) at each link.

$$R_x = \rho_x n_x \tag{5.4}$$

To formulate the RB gains that result from the definition of spectral efficiency, we asume that the "initial condition" is that the direct link achieves R_d with efficiency ρ_d when it is assigned n_d. If a third party RN is activated, we assume that end-to-end data delivery of the access and relay links are jointly assisted with a set of resources $n_{\mathrm{e2e}} = n_\mathrm{r} + n_\mathrm{a}$ and, with them, the two-hop chains achieve the optimal combined spectral efficiency calculated in (3.3). Therefore, a saving in RBs between n_d and n_{e2e} is obtained if ρ increases.

We have approximated the spectral efficiency of LTE with the shannon curve, but this may be misleading: any SNR value may lead to a nonzero value in spectral efficiency where the system seems to "work". However, the limited set of MCS of LTE places an upper/lower bound on the spectral efficiency that is actually achievable. Above the maximum value, spectral efficiency does not keep growing with SNR, and, below the minimum value, spectral efficiency is zero as no communication is possible.

We can introduce an additional quantification step through the standard MCS of LTE before the rates on each link are evaluated.

$$Q(\rho(\mathrm{SINR})) = \max_{m} \left\{ m \in \mathcal{MCS}_{\mathrm{LTE}} / m < \eta_W \log_2(1 + \eta_S^{-1}\mathrm{SINR}) \right\} \tag{5.5}$$

Expressions (5.2) and (5.5) allow the computation of the simulation parameter R_{MAX}. Since there is a minimum spectral efficiency in LTE-A [160] ($\min(\rho) = 0.1523$, corresponding to the most protected MCS), we can identify the distance at which this spectral efficiency is achieved as the cell boundary:

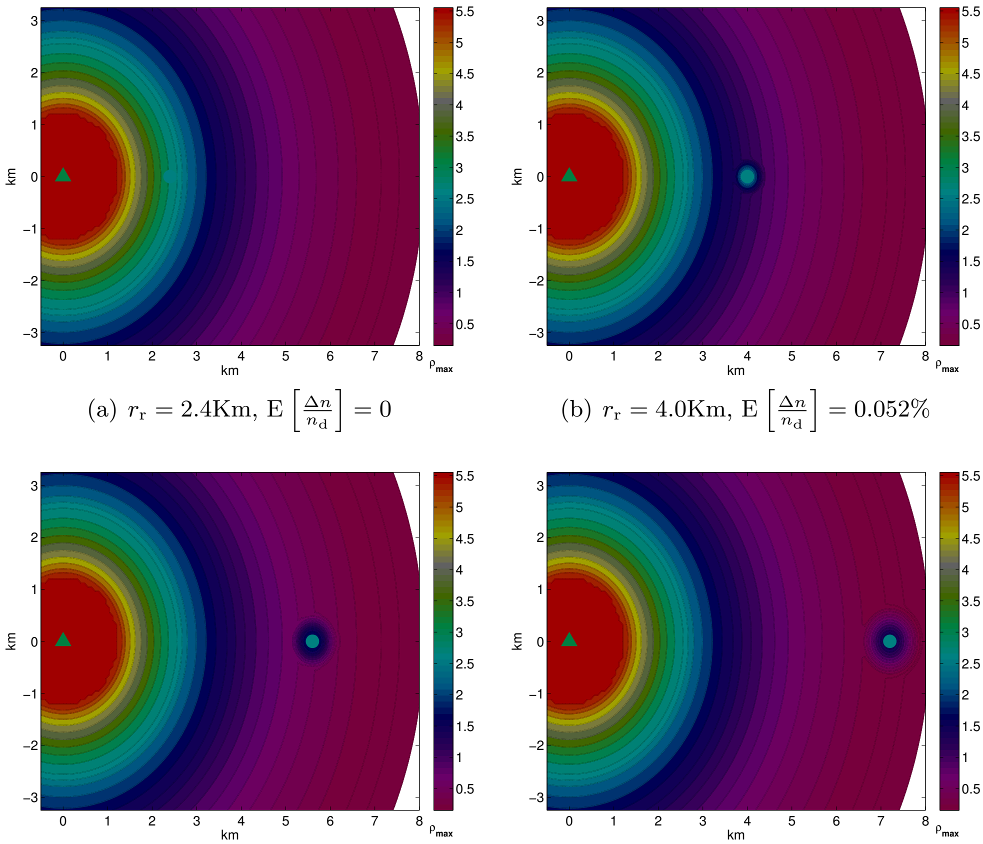
$$R_{MAX} = r / \{\rho_d(r) = \min(\rho)\} = 8.08 \text{ km} \tag{5.6}$$

5.3 Gains with a Single Relay

In this section we discuss the RB gain of a single RN and the effect on gains of multiple channel parameters such as MCS, fading, etc. Since we have only considered a simplified scenario with a single eNB, a RN and a UE, the model has circular symmetry. That is, the outcome is symmetric with respect to any joint rotation of the UE and the RN dispositions around the eNB. It is therefore sufficient to study the spectral efficiency variation in all points in space when the distance between the eNB and the RN varies. We illustrate this in Fig. 5.2 for four different values of r_{r}. These figures were obtained without MCS quantification.

We observe that the effect of a relay on the achievable spectral efficiency in the plane is a small "island" of improvement arount the RN. The eNB (green triangle) has a longer range due to its much higher transmitter power. When the relay is located in areas with a high ρ_{d}, there is practically no gain, but when the RN is placed further away from the eNB, where ρ_{d} is low, the better antennas and the regeneration capabilities of the RN allow a large surrounding area where ρ_{e2e} is high.

Finally, for a given third-party RN located at r_{r} that is a candidate for CSL, we can compute the instantaneous $\Delta\rho$ for each position of the UEs (instantaneous CSL) or the average spectral efficiency gain across all UE positions (average CSL). We next compute the amount of RBs saved by the eNB using the RN that is equivalent to $\Delta\rho$. And since we wish to apply our simple CSL approach, we choose the RN to be rewarded with half the RB gains as suggested in Chapter 4.



(a) $r_{\mathrm{r}} = 2.4\mathrm{Km}$, $\mathrm{E}\left[\frac{\Delta n}{n_{\mathrm{d}}}\right] = 0$

(b) $r_{\mathrm{r}} = 4.0\mathrm{Km}$, $\mathrm{E}\left[\frac{\Delta n}{n_{\mathrm{d}}}\right] = 0.052\%$

(c) $r_{\mathrm{r}} = 5.6\mathrm{Km}$, $\mathrm{E}\left[\frac{\Delta n}{n_{\mathrm{d}}}\right] = 0.192\%$

(d) $r_{\mathrm{r}} = 7.2\mathrm{Km}$, $\mathrm{E}\left[\frac{\Delta n}{n_{\mathrm{d}}}\right] = 0.404\%$

FIGURE 5.2: Spatial distribution of $\rho_{\max}$ with a RN at different distances.

RB savings can be calculated as follows: let R_{d} bps be the throughput that was originally experienced by the UE connected to the eNB before the RN took over. Then, the original number of RBs per second needed for that transmission is $n_{\mathrm{d}} = \frac{R_{\mathrm{d}}}{\rho_{\mathrm{d}}}$, and the new number of RBs needed is $n_{\mathrm{e}} = \frac{R_{\mathrm{d}}}{\rho_{\max}}$. Therefore the fraction of gained resources is

$$\frac{-\Delta n}{n_{\mathrm{d}}} = \frac{n_{\mathrm{d}} - n_{\max}}{n_{\mathrm{d}}} = \frac{\Delta\rho}{\rho_{\max}} = \frac{\Delta\rho}{\rho_{\mathrm{d}} + \Delta\rho}. \tag{5.7}$$

One way to implement leasing at the LTE-A physical layer, in the same way as the RN reserves OFDMA symbols in a subframe to receive the eNB-RN transmission [17], would be to mark the leased RBs as Multicast Broadcast Single Frequency Network (MBSFN) frames in the primary LTE-A system to make the UE aware that there is no eNB transmission in them, as the RN does with its UE when it enters receive mode. Hence, a possible solution for spectrum leasing is Algorithm 4.

It is important to remark that the DeNB, as a proxy of interfaces `X2` and `S1` of the RNs, knows detailed information of the UE attached to the RN that it would probably not know about UEs attached to other regular eNBs. To transfer UEs between the DeNB

Algorithm 4 Third-party relay spectrum leasing

```
nRBs = 0
for all u ∈ RN.attachedUE do
    RBgain=sizeOfDirectScheduling(u)-sizeOfRelayScheduling(u,RN);
    nRBs+=RBgain/2;
end for
eNB.leaseRBS(RN,nRBs)
```

and the RN, the usual LTE-A eNB-initiated handover procedure described in [16] can be employed using the requirement that (5.3) is greater than zero as a handover trigger, evaluating it from channel measurements reported by the UE.

Expression (5.3) allows to measure the gain provided by a particular relay, serving a particular UE, allocating all the resources of the LTE-A cell to that user. However, it is likely that the RNs will follow a much slower birth and death process than that of the UEs, and will be less mobile. Consequently, it is more interesting to develop a metric for the *average* spectral efficiency gain provided by a relay over all the possible UE locations it may serve. Let $f(r,\theta)$ be the UE location p.d.f, in polar coordinates.

$$\mathrm{E}_{f(r,\theta)}\left[\Delta\rho_{\max}(r_{\mathrm{r}})\right] = \int_0^{R_{MAX}} \int_0^{2\pi} \Delta\rho_{\max}(r, r_{\mathrm{r}}, d(r, r_{\mathrm{r}}, \theta)) f(r,\theta) r d\theta dr \tag{5.8}$$

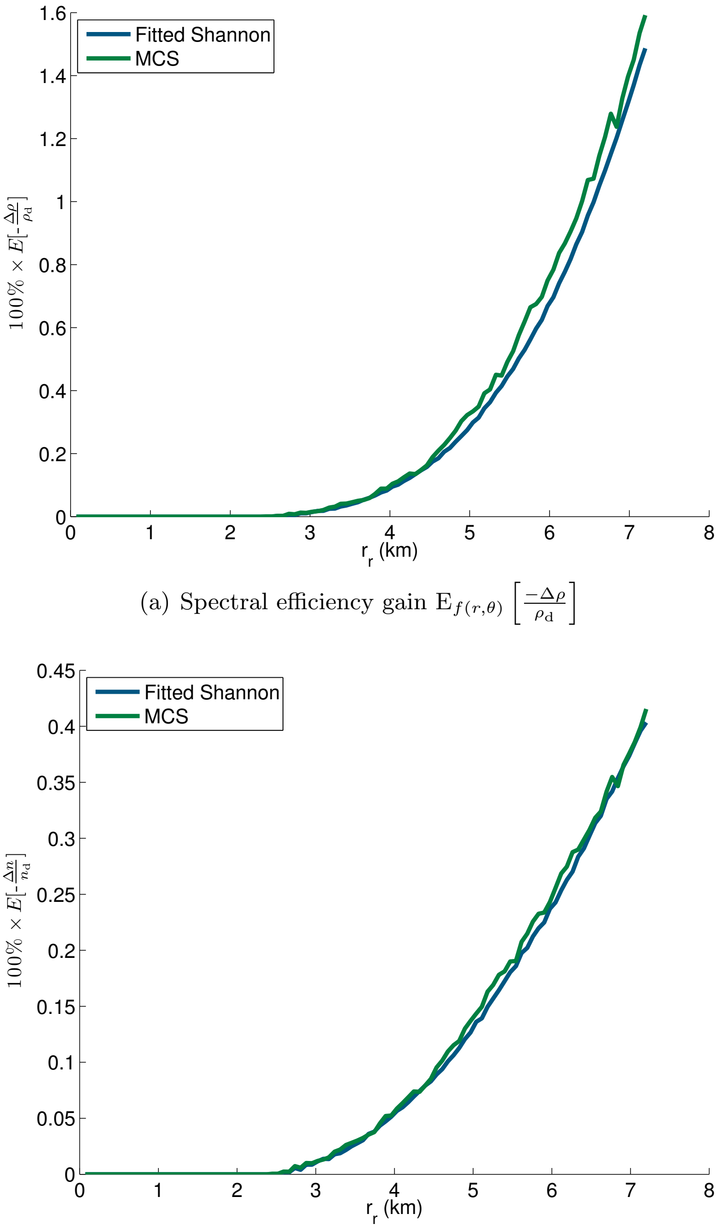
To obtain a simple representation of (5.8), we discretize locations into a grid with $N_{\mathrm{x}} \times N_{\mathrm{y}}$ coordinates and estimate the integral using the following discrete calculation:

$$\mathrm{E}_{f(r,\theta)}\left[\frac{-\Delta n}{n_d}\right] \simeq \frac{1}{N_x N_y} \sum_{x=-R_{MAX}}^{R_{MAX}} \sum_{y=-\sqrt{R_{MAX}^2-x^2}}^{\sqrt{R_{MAX}^2-x^2}} \frac{-\Delta n(x,y)}{n_d(x,y)} \tag{5.9}$$

Whose solution depends on distance between the DeNB and the RN, and it is equivalent to the spatial average of Fig. 5.2 across RN positions. We illustrate the result as a function of r_{r} on Fig. 5.3.

Resource gains lie in the range of $0.1 - 0.4\%$, which may seem too low, but this depends on the intended application. For example, from the point of view of the third-party RN, $\frac{0.1}{2}\%$ of the capacity of the whole LTE-A cell, which peaks to 1 Gbps, is obtained, yielding a practical leased throughput of up to 0.5 Mbps, which more than doubles the peak throughput that could be achieved with typical M2M technologies such as ZigBee or Dash7.

In Fig. 5.3 we can observe that quantification with standard MCS (5.5) follows approximately the same result as a continuous fitted Shannon curve (3.12). RNs very close



(a) Spectral efficiency gain $\mathrm{E}_{f(r,\theta)}\left[\frac{-\Delta\rho}{\rho_\mathrm{d}}\right]$

(b) RB gain $\mathrm{E}_{f(r,\theta)}\left[\frac{-\Delta n}{n_\mathrm{d}}\right]$

FIGURE 5.3: Spatial average of gains versus location of the relay (r_r)

to the DeNB offer strictly zero gains instead of little non-zero values, and gains of distant RNs DeNB are a little bit higher because quantization granularity is lower for long distance (i.e. direct) links.

Even though the gain of a single relay is small, this is because only one relay, rather than the usual amount of a dozen, is considered [27]. Since relays are free for the operator, the DeNB is free of aggregating relays until a significant total gain is achieved. We discuss this in the next section.

5.3.1 Multi-User Diversity with Fading

As we saw in the previous chapters 3 and 4, multi-user diversity is a promising enabler for rate increase. In our channel model, channels are static and a worst-case margin is added to path loss as indicated in [27]. If the channels suffer Rayleigh fading, an additional temporal dimension should be added to the integral in (5.8). Due to the diversity gain of the relay channel, this temporal dimension would allow to enhance the rate even further. We have simulated this scenario and compared the results in Fig. 5.3 (again, no MCS considered).

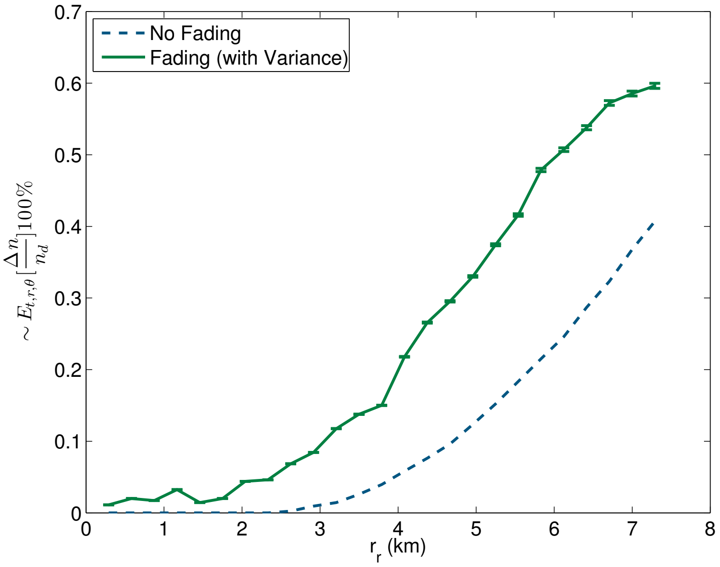


FIGURE 5.4: Temporal-spatial average of RB gain ($\mathrm{E}_t\left[\mathrm{E}_{f(r,\theta)}\left[\frac{-\Delta n}{n_d}\right]\right]$) versus relay location (r_r) averaged over 5000 fading realizations.

5.4 Gains with a Massive Relay Population

A single relay provides about 0.1% RB gain, but looking at Fig. 5.2 it appears that many relays could fit into the range of the main cell without coverage overlaps. In addition, if there is no relation between RN service areas, the aggregate gain of multiple RNs is the sum of their individual gains. The question is how many relays can fit into the cell.

Previous literature on LTE-A relaying typically consider optimally located relays, since they focus on operator RN planning, as in the case of eNBs. In a third-party relaying scenario, this assumption is not possible. Instead, an analysis of third-party relay CSL must expect relays to appear anywhere.

5.4.1 Regularly placed RNs

We begin our analysis looking at the gains for RNs distributed in an unrealistic regular fashion. On 3.6 we defined the *service area* of a RN as the circular region around the RN whose radius is equal to the RNs *service distance*, calculated using (3.16). By using these parameters we design a circular distribution of RNs to cover all the cell, illustrated in Fig 5.5, in a series of concentric rings, where the angular separation of the RNs in the same ring and the distance between rings is obtained through the calculation of the service distances. The radius of each ring j is:

$$r_{\mathrm{r}}(j) = \begin{cases} R_{MAX} & j = 1 \\ r_{\mathrm{r}}(j-1) - 2d(j-1) & j > 1 \end{cases} \tag{5.10}$$

where $d(j-1)$ is the diameter of the influence area of a relay of the previous ring.

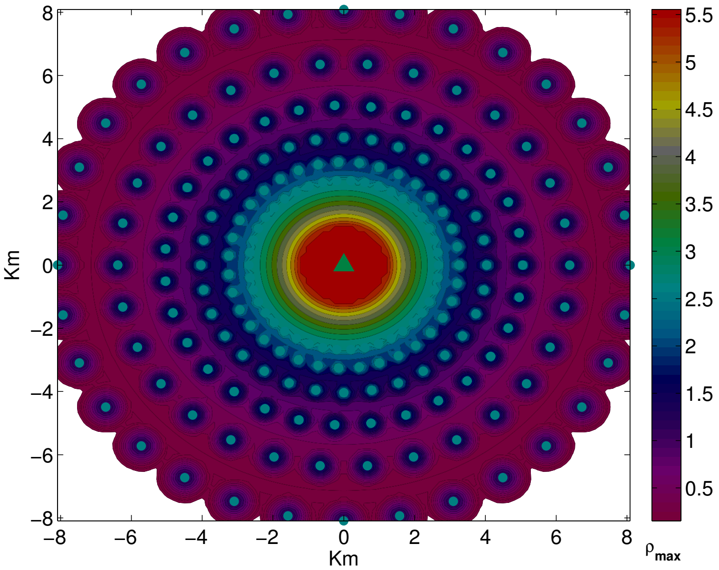


FIGURE 5.5: Spectral efficiency with 202 RNs located in regular rings.

The RB gain resulting of the RN distribution in Fig. 5.5 is $\mathrm{E}_{f(r,\theta)}\left[\frac{-\Delta n}{n_{\mathrm{d}}}\right] = 20\%$. We take this result as a best-case scenario, which means that LTE CSL spectrum gains are going to be moderate at best, as a benchmark for the random location case in the next section.

5.4.2 Randomly placed RNs

We now calculate the RB gain in a realistic CSL scenario where RNs pop up in random locations and are processed by the admission control algorithm in Section 3.6. The algorithm assumes that there are k already accepted RNs and defines an admission control (AC) rule for candidate RN $k+1$ based on the expected RB gain. In order to implement the AC, it suffices to include the desired rules in the `RN attachment procedure` in [16].

For test purposes, we follow a simple greedy AC method: a relay is admitted if it creates gain and its influence area does not overlap with that of other relay. The implementation of the AC rule in Algorithm 1 considers the service area of a RN both as an indicator of its possible gains and as a rejection mechanism if it interferes with the service are of one of the already accepted k RNs.

Fig. 5.6 shows how the algorithm copes with the consecutive arrival of 300 candidate RNs. In total, 95 out of 300 of the RNs are accepted, providing an aggregate spectral efficiency gain of 12.95%. It can be seen that the AC is quite effective in avoiding RN conflict, as no overlapping occurs between the service areas.

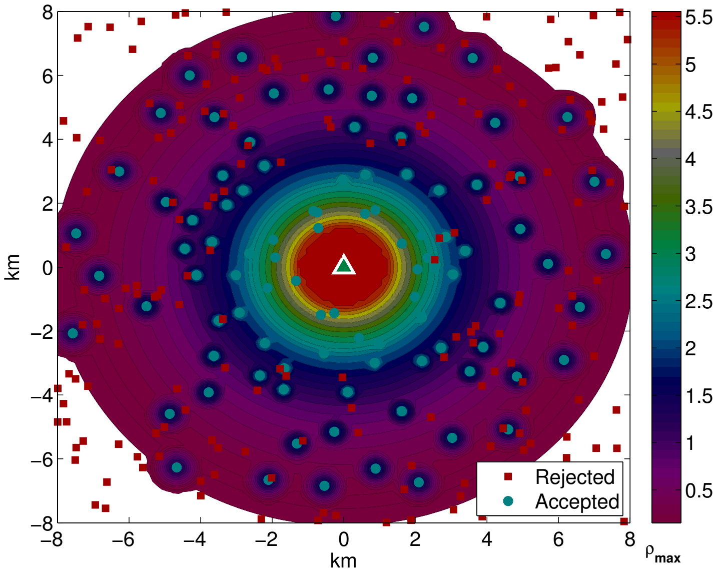


FIGURE 5.6: Spatial distribution of spectral efficiency with 300 uniformly distributed candidate RNs.

We evaluated the performance of the AC algorithm across 100 realizations of the RN location distribution and present the number of accepted RNs versus the number of relay candidates offered in Fig. 5.7. The error bars indicate the variance across realizations, which is quite low. Fig. 5.8 shows the RB gain of the cell in the same experiments.

These figures show that average RB gain $E\left[\frac{-\Delta n}{n_d}\right]$ grows rapidly with the initial increase in the number of candidate RNs, yielding a gain of over 15% RBs. Of course, this does not mean that optimization is irrelevant for future improvements, but that a simple AC mechanism such as the one we have presented suffices to demonstrate that CSL is interesting in some cases. Another possible improvement is related to the fact that the AC mechanism proposed accepts any relay that generates a gain, no matter how

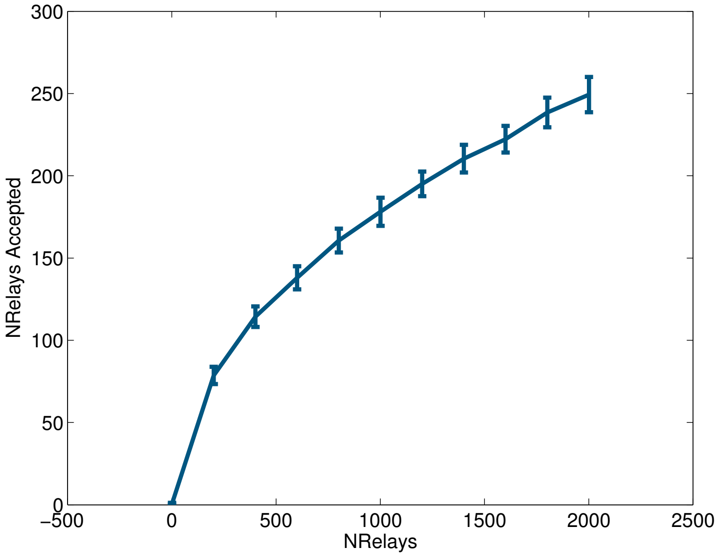


FIGURE 5.7: Number of accepted RNs vs. number of candidates RNs.

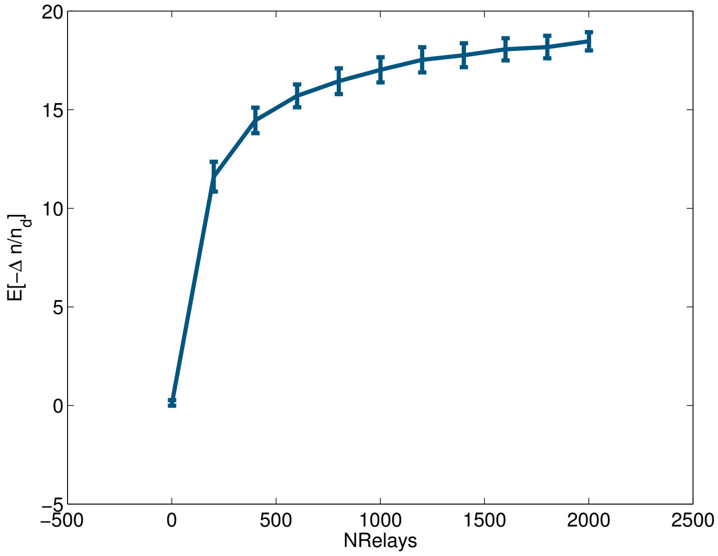


FIGURE 5.8: $\mathrm{E}_{f(r,\theta)}\left[\frac{\Delta n}{n_d}\right]$ vs. number of candidates RNs.

small that gain is. Since UE transfer between RNs and the DeNB is based on the handover procedure in [16], if the model includes user mobility overly frequent handover procedures might eat up the resources gained by accepting a relay. This can be easily corrected by setting a gain threshold requirement for candidate RNs. Another solution may be setting an AC mechanism for UEs trying to camp in a RN based on mobility. An UE with a low handover record would be preferable to the RN. This policy is well suited for M2M scenarios with static devices.

5.5 Examples of Leasing Scenarios

So far, following the philosophy we established in Chapter 4, we have focused on the global RB gains due to leasing (Fig. 5.8), which can be viewed as a social gain and is the same for any CSL scenario, and we have omitted how the primary and the secondary share those resources, since this depends on the final application. This is because, obviously, the benefit that the secondary obtains from the spectrum leased depends greatly on the intended use of the RBs.

However, to provide some examples, it is necessary to discuss some applications, for which the RB division and the spectral efficiency of the secondary service must be taken into account. To perform CSL, the secondary needs to implement the functionality of a standard LTE-A RN. LTE has a high spectral efficiency, and the secondary operator will desire to reduce costs, and therefore it is very likely that the CR applications will also rely on LTE-derived waveforms with similar spectral efficiency.

If we assume that the secondary application employs LTE transceivers, the utility value of their signals should not be expected to differ greatly from the value of cellular LTE traffic. If the secondary traffic had much more value, its operator would for sure purchase cellular connections, and if the secondary traffic had much less value, it would probably be carried with cheaper transmission technologies. With this in mind, we retake our basic approximation to resource partition in Chapter 4: splitting the RB gains in half between primary and secondary.

With the assumptions above, we provide the following examples:

5.5.1 Example 1: CSL Private Neighborhood Area Network for Smart Grid Last Mile

The term Smart Grid Last Mile (SGLM) refers to the communications between a power grid customer (house, company...) Energy Services Interface (ESI) and a SG Distribution Access Point (DAP) nearby. This device aggregates traffic from different households that are connected to the same branch in the last segment of the energy distribution network. The underlying network is called the Neighborhood Area Network (NAN) [168] and the number of nodes can theoretically scale up to tens of thousands [49].

In [15] we studied the implementation of SGLM NAN using WiMaX. In [5] we generalized the study to an arbitrary SGLM traffic model that can be applied to test the validity of any communications technology for SGLM. Moreover, in [14] we presented a practical implementation of the SGLM communications test. The main features of a SGLM application are:

- Very tight latency and reliability requirements.
- Very low and regular rate per device.
- Massive number of static users.
- Metropolitan range.

Even though LTE and WiMax operate at neighborhood scale, the former has better mechanisms to manage delay and high number of devices.

From the point of view of CSL, this type of scenario is representative of a situation where a massive number of remote sensors (ESI of each home) are controlled by a concentrator (ESI concentrator or gateway) that is a privately-owned third party LTE-A RN (one data concentrator per building or per block), and the third party application consists of exchanging data between the RNs and a centralized server for each neighborhood (the DAP, which can be physically installed near the DeNB). Moreover, if two ESI devices are too close to act as non-overlapping RNs and only one is accepted by the AC, the other could switch to sensor functionality.

We now study the rates that such a system would obtain. The RNs offer their help to the DeNB using CSL, and they employ the leased RBs afterwards to send their own signals to the DeNB as well, to implement a LTE backhaul towards the AP. Let us define $\mathcal{RN}$ as the set of accepted RNs. We can compute the average spectral efficiency that is leased to each RN (ρ_{NAN}) by multiplying the spectral efficiency of the relay links by the number of RBs leased to it.

$$\rho_{\mathrm{NAN}} = \sum_{r \in \mathcal{RN}} \rho_{\mathrm{r}}(x_{\mathrm{r}}, y_{\mathrm{r}}) \times \mathrm{E}_{(x - x_{\mathrm{r}}, y - y_{\mathrm{r}}) \leq d_{\max}} \left[\frac{\Delta n(x, y)}{n_d(x, y)} \right] \tag{5.11}$$

Fig. 5.9 shows that when the number of candidate RNs is very high the leasing gains approach ∼0.25 b/s/Hz. On a typical 20 MHz LTE channel this becomes a sum rate of 5 Mbps leased to the private NAN. This may seem a small rate to a reader that is used to the typical traffic volumes of LTE, but in the context of SGLM it is a significative improvement over the WiMaX [15] or Power Line Communication (PLC) alternatives.

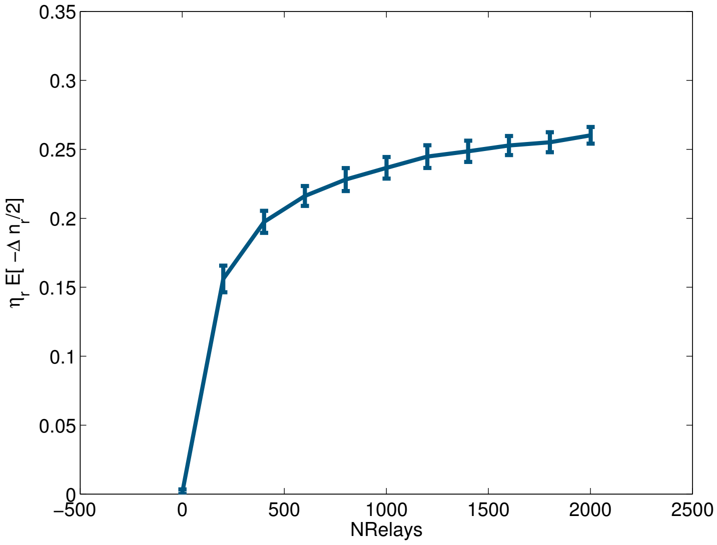


FIGURE 5.9: Average and typical deviation of leased spectral efficiency in a NAN (in b/s/Hz).

5.5.2 Example 2: CSL Private Local Area Network for P2P or M2M Applications

P2P services are often based on overlay networks, in which resource management is a key challenge [7]. The peers that participate in a P2P overlay not only act as conventional users, but also collectively play the normal role of a central server. In such a system, the content servers (where information is placed), proxy servers (intermediary nodes) and message routing functions of the traditional services are replaced by a distributed P2P overlay.

In [7] we studied problems related with the implementation of a traditional cellular application (voice calls) through VoIP P2P communications. One characteristic of P2P applications is local traffic clustering between neighbor devices. On the other hand, typical M2M applications generate local traffic, such as the case of SG home area networks (HAN) [5, 14], domotics, and sensor and actuator networks in industry. LTE-A RNs have the necessary hardware to form small cells to serve devices in a Local Area Networks (LAN) with short-range communications.

From the point of view of CSL, this type of networks would be an example of a situation where a massive number of LTE-enabled local devices (P2P users or M2M sensors) are organized in clusters with intra-cluster LAN communications. Each LAN would be managed by a a privately-owned third-party RN (group coordinator in P2P or data concentrator in M2M) and the third party application would consist of data exchange between the local nodes in each LAN and the RN. Moreover, if two devices were too close to act as non-overlapping RNs and only one was accepted by the AC, then they could reach each other, and the second could merge into the cluster of the first.

We now study the rates of such system, in which the RNs offer their help to the DeNB using CSL, and employ the leased RBs to exchange their own signals with UE-like devices by forming a small LTE cell LAN. Let us define $\mathcal{RN}$ as the set of accepted RNs, $\mathcal{U}(r)$ as the set of users connected to $r \in \mathcal{RN}$, and (x_u, y_u) as the location of each user $u \in \mathcal{U}(r)$. The average spectral efficiency leased to such cells is the average spectral efficiency of the nodes in the RN cell multiplied by the number of RBs leased to it:

$$\rho_{\mathrm{LAN}}(r) = \frac{1}{|\mathcal{U}(r)|} \sum_{u \in \mathcal{U}(r)} \mathrm{E}_{(x_u,y_u)} \left[\rho_{\dashv}(x_u - x_r, y_u - y_r)\right] \times \mathrm{E}_{(x-x_\mathrm{r},y-y_\mathrm{r}) \leq d_{\max}} \left[\frac{\Delta n(x,y)}{n_d(x,y)}\right] \forall r \tag{5.12}$$

Fig. 5.10 shows that, when the number of candidate RNs is very large, the leasing gains approach ∼0.5 b/s/Hz. On a typical 20 MHz LTE channel this means a sum rate of 10 Mbps leased to **each** private LAN. Again, this may seem a small rate, in the context of VoIP or M2M it is a significant capacity that may serve plenty of devices [7].

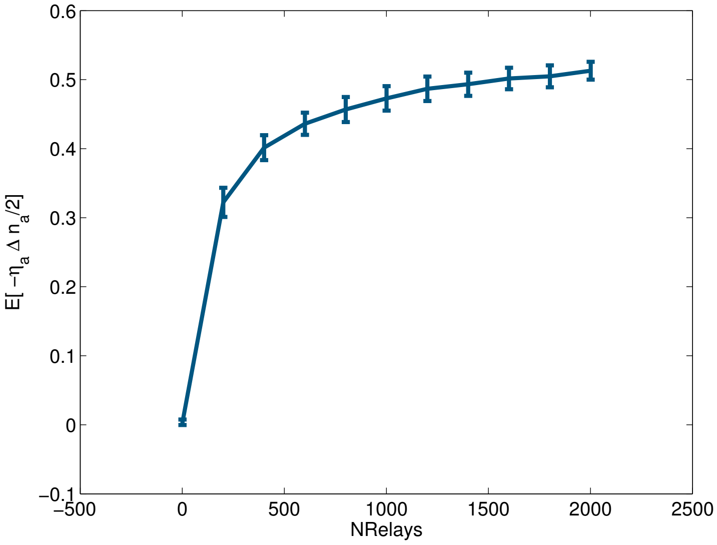


Figure 5.10: Average and typical deviation of leased spectral efficiency in a LAN (in b/s/Hz).

5.6 Summary

It is technically possible to implement CSL secondary transmitters as standard-compliant third-party relays for LTE-A networks. Unfortunately, when one RN is introduced, it only achieves tiny increases in spectral efficiency and RB savings in LTE-A cell downlink. However, the RN only serves a small portion of the surface of the cell and, since cognitive cooperative relays are free and their gains cumulative, a DeNB may benefit from the aggregation of many RNs with little limitation, adding up the gains of many spatially separated candidates.

The transfer of UEs to be served by RNs is a handover procedure, and its load can be mitigated by setting a threshold for the RNs to admit only UEs with low mobility. However, the AC system we have designed is based on RN service areas, and since small areas may lead to frequent handovers we can set a minimum area threshold for RN AC to restrict CSL to RNs with large service areas.

The gains of CSL with current cellular technology do not seem high enough to justify a generalization of the technique. It is likely to remain limited to niche solutions such as M2M until future versions of the LTE-A standard create the conditions for a broader use. The low gains are mainly due to the implementation of RNs as level 3 OSI entities. Non-standard level 2 AF, non-orthogonal relays and the more advanced cooperative signal processing techniques described in Chapter 2, if introduced in future standards, would increase CSL gains.

The content in this chapter is an extended and actualized version of a paper published in the 20th European Wireless conference (EW2014) [13].

SGLM applications and their traffic models, taken as a CSL use case, were studied as part of this doctoral work and published in three papers in the 2012 IEEE Innovative Smart Grid Technology Europe conference (ISGT Europe) [15], the IEEE Transactions on Smart Grid [5], and the 2014 IEEE International Energy Conference (Energycon) [14].

Part III

The Distant Future: Cooperation in Next Generation Wireless Network Technologies

Chapter 6

Cooperative Multi-hop Achieves The Capacity Scaling of Future Cellular Networks

Contents

6.1 Introduction

All future lines of evolution of cellular networks –mmW transmissions, massive MIMO and ultra-dense deployment– are, from an information theoretic perspective, various ways to increase the fundamental degrees of freedom of the network: bandwidth, antennas and area. To evaluate the potential value of these technologies, this chapter characterizes the fundamental capacity of cellular networks under parametric scaling of these dimensions. Our analysis follows the lines of the classic result by Gupta and Kumar [60], applied to cellular infrastructure networks rather than *ad hoc* networks without infrastructure. Specifically, we consider a large cellular network with n mobile nodes, with various scalings with n of the key parameters such as bandwidth, number of antennas and BSs, and area.

Our main results determine the capacity scaling by finding identically-scaling lower and upper bounds on the achievable rate. The upper bound is a cut-set bound in which all BSs and nodes cooperate from their respective communication sides forming a massive point-to-point MIMO system. The lower bound is found by considering a simple cooperative Infrastructure Multi-Hop (IMH) protocol where transmissions are relayed to the closest BSs via cooperative mobile nodes within the same cell.

The results show that massive spectrum availability, massive number of antennas, and ultra dense cell deployments push next-generation cellular networks into a new regime where the fundamental degrees of freedom are plentiful. The network capacity has a fundamental bandwidth scaling limit, beyond which the network becomes power-limited and capacity does not grow with additional bandwidth. The bandwidth scaling limit is a function of the worst-case received power. The IMH protocol improves this limit, achieving full capacity in all regimes. In contrast, current protocols relying on direct transmissions between the UEs and the BS do not achieve the maximum capacity scaling except in the special case when the density of BSs is taken to impractical extremes. The details in the analysis yield important and surprising findings:

- *Critical bandwidth scaling limit:* It is well-known that point-to-point links in wideband regimes become limited by power rather than bandwidth, particularly with channel fading and when CSI is not available [75, 76]. This work identifies a critical bandwidth scaling at the *network* level that defines the maximum bandwidth useful with any protocol. Below the critical point an increase in bandwidth improves capacity, whereas above this value, the network becomes power-limited and additional bandwidth no longer improves the overall rate.

- *Benefits of increased cell density:* Interestingly, the capacity scaling is related to BS density but the critical bandwidth scaling is related to the number of BS antennas and inter-node distance, but not to cell size. Therefore, at least in a simple rate scaling analysis, network capacity can be increased arbitrarily with higher cell densities, whereas increasing bandwidth or number of antennas will not improve network capacity after some point. Moreover, no scaling parameter seems affected by interference alignment [169], which is inherently present in the upper bound but not in the lower. Therefore, this technique, despite being potentially important in modern systems, is only present as a constant factor in the analyisis, without affecting scaling.

- *Optimality of infrastructure multi-hop and benefits over single-hop communication:* We have found that a simple IMH protocol is sufficient to obtain the optimal capacity scaling in all regimes. In the IMH protocol, each BS cell is divided into sub-cells and transmissions to and from the BS are relayed via mobile nodes within each sub-cell. In contrast, Infrastructure Single Hop (ISH), with direct UE-BS transmissions within the cell, is strictly worse than IHM in regimes with wide bandwidths or large numbers of antennas. The reason is that ISH relies on longer communication distances that become power-limited and hit a protocol-specific bandwidth scaling limitation earlier than IMH. On the other hand, in regimes with narrower bandwidths, ISH is sufficient to achieve the optimal scaling.

  This result suggests that, in today's networks, which are fundamentally bandwidth-limited, the current model of primarily using only direct transmissions between the UEs and the BS is sufficient for achieving the maximum capacity. As a result, our observation in Chapter 3 that LTE-A RNs do not offer great gains is theoretically justified. However, in future networks with much larger bandwidths or much more antenna degrees of freedom, multi-hop communication will become necessary to fully attain the network capacity. This reinforces our observation in Chapter 5 that a posteriori introduction of cooperative mechanisms in current networks will deliver small gains outside application niches with massive numbers of devices until the advent of new standards designed from scratch to exploit cooperative diversity gains.

- *Hierarchical cooperation is not necessary:* The upper-bound result implies that hierarchical cooperation cannot improve the scaling achievable with IMH, nor is it necessary to achieve capacity scaling in cellular networks. This contradicts the case for dense ad-hoc networks in [63]. Our cellular result and ad-hoc network regimes coincide in that multi-hop is efficient in extense networks where short-distance communications are power-limited. However, for dense networks, direct transmission is optimal in narrow-band cellular networks, which contradicts the results that in dense ad-hoc networks direct transmission is suboptimal and hierarchical cooperation is the optimal strategy. We postulate that this difference is due to the fact that cellular and ad-hoc networks have very different spatial traffic distributions. In the former, nodes always communicate with the closest BS at hand, whereas in the later nodes may communicate with any other node in the network. Therefore, some direct-communications between nodes far apart are long-range in a topological sense, interfering with many neighbors in between, but they are short-range in a physical sense, because the network is dense and transmission distance is short. The combination of both aspects yield unbounded interference, whereas interference in cellular networks is bounded. To illustrate the difference, we construct an hypothetical Infrastructure Hierarchical Cooperation (IHC) protocol, even though the upper bound already has proven that IHC cannot outperform IMH. We perform a detailed analysis of the operation of this hierarchical protocol in comparison to IMH and we discuss the difference with the ad-hoc HC analysis in [63]. The analysis shows not only that IHC cannot outperform IMH, but also that it is suboptimal in many cases. The comparison with the ad-hoc analysis shows that the limiting constraint is indeed the difference in spatial distribution of the network traffic, which is concentrated around the BS in the cellular case.

- *Practical implementation with infrastructure relays:* It may be difficult to support mobility, association and handover in a user-dependent IMH architecture. In addition, using subscriber devices for relaying traffic is troublesome nowadays. A more practical model consists on two levels of APs, some with wired backhauling (BSs) and some with wireless backhauling (RNs). We show that the resulting scheme, which we call Infrastructure Relay-multi-Hop (IRH), scales as ISH when RNs density is not higher than BS, and as IMH when RN and UE densities coincide. The scheme has the added advantage that it allows to build any intermediate model between ISH and IMH selecting RN density. Mobile nodes perform single hop communications towards their APs so the inherent difficulties of mobile multi-hop wireless routing and AP selection dissapear, improving resilience against node mobility.

6.2 System Model

NOTE: *In this thesis we use B to denote transmission bandwidth, but in scaling law analysis the usual notation is W. Moreover, β is also usually employed in this field to represent the scaling exponent of the number of BSs. In order to avoid confusion with the rest of the scaling laws literature, in this chapter we drop our own notation and use W to represent transmission bandwidth.*

6.2.1 Channels, Signals and Link Rates

In our analysis of achievable rate we model a continuous-time bandlimited channel in an orthogonal signal space, where a channel with low-pass bandwidth $W/2$ sampled at the Nyquist frequency has W complex-valued coefficients per second.

We index these channel samples by t. Our analysis applies to the following three common models:

- In the AWGN channel, the signal at each receiver is the sum of the signals transmitted towards it, other signals transmitted at the same time (interference), and an additive noise source. Lett $\mathcal{D}$ be the set of desired transmitters and $\mathcal{I}$ the set of interferers. For each receiver u we have

$$y_{u,\text{AWGN}}[t] = \sum_{d\in\mathcal{D}} x_d[t] + \sum_{i\in\mathcal{I}} x_i[t] + z[t] \tag{6.1}$$

  In this model a point to point link rate is limited by $\mathrm{I}(X;Y)$.

- In a Coherent Fading Channel (CFC), or fading channel with *a priori* CSI, each transmitter-receiver pair of antennas experiences a random gain between them. Let $N_\mathrm{t} \times N_\mathrm{r}$ be the dimensions of the channel matrices. The resulting MIMO channel is

$$\mathbf{y}_{u,\text{Fading}}[t] = \sum_{d\in\mathcal{D}} \mathbf{H}_{d,u}\mathbf{x}_d[t] + \sum_{i\in\mathcal{I}} \mathbf{H}_{i,u}\mathbf{x}_i[t] + \mathbf{z}[t] \tag{6.2}$$

  and since devices know the channel a priori, a point to point link rate would be limited by $\mathrm{I}(X;Y|H)$. Here, when CSI is available at the receiver (CSI-R), it is used for optimal decoding (for example, inverting the channel matrix). If

CSI is also available at the transmitter (CSI-T), it allows to maximize the mutual information over the set of possible encoding distributions of the transmitted vector $\mathbf{x}$ (for example, using transmitter beamforming or spatial multiplexing).

- A Non-Coherent Fading Channel (NFC), or fading channel without a-priori CSI (in both the transmitter and the receiver), would experience the same channel above its rate would be limited by channel uncertainty, $\mathrm{I}(X;Y) = \mathrm{I}(X,H;Y) - \mathrm{I}(H;Y|X)$. Note that lack of a-priori CSI does not necessarily imply that channel information cannot be used in the transmitter or receiver, it only implies that to use channel states they must first be estimated a posteriori from the received signals (i.e. non-coherent channel $\not\Rightarrow$ non-coherent receiver). Sometimes, transmitters can also get to know the CSI estimated at the receiver at no cost due to reciprocity of TDD systems. Otherwise CSI-T comes at the cost of using feedback links.

The *coherence time* (T_c) of the channel is the time it takes the channel to change, and its *coherence bandwidth* (B_c) is the minimum separation in Hertz between two coefficients of the channel frequency response for them to be i.i.d. The *coherence length* $L_\mathrm{c} = B_\mathrm{c}T_\mathrm{c}$ determines the overhead of channel estimation in NFC and as the first becomes longer the later becomes negligible. Therefore, the difference between the NFC and the CFC depends in reality on the degree of channel variation, so that it is desirable to formulate scaling results that are valid for both models.

We first argue that, as far as our analysis is concerned, the three types of point-to-point channels are subject to equivalent scaling laws as a function of the received power, bandwidth and antennas allocated to their transmission. We bridge the gap between the point-to-point models and information-theoretic channel models relevant to network models (MAC, BC, RC...) by arguing that in terms of scaling these do not differ from simple time-division or frequency-division protocols. We use Frequency Division Multiple Access/Multiplexing (FDMA/FDM) protocols to calculate allocated power and bandwidth per user in the network scaling with number of nodes, network area, total bandwidth and number of BS antennas. After establishing the common scaling for all channels, we focus on the NFC to develop the rest of the achievable rate models, because it corresponds to the worst case out of the three channel models.

Our upper bound results are based on a point-to-point global MIMO representation of the network and overestimate the potential rate achievable with perfect interference processing. On the other hand, in our achievable schemes, and excepting some cases where we specify a certain type of multi-user processing, we will assume that receiver nodes treat interference as additional Gaussian noise. Mostly, this will be the case for out-of-cell interference in the cellular network or out-of-sub-cell interference when cells are further subdivided. We also assume that the transmissions that cause this interference are allocated uniformly in frequency. Therefore, a receiver u in a link with assigned bandwidth W_u will experience the proportional fraction of the total interference power spread across the total available bandwidth W. We say receiver u experiences an equivalent AWGN process with PSD given by

$$N_{\mathrm{I}} = \frac{\sum_{i \in \mathcal{I}} P_i}{W} + N_0, \tag{6.3}$$

where N_0 is the thermal noise PSD, W is the total bandwidth, $\mathcal{I}$ is the interferer set that depends on the nodes that are active with each protocol and P_i the interfering power from each node in the set.

This definition provides a valid approximation for all the channels described above and it covers our two approaches to model interference: non-empty $\mathcal{I}$ for the achievable schemes and empty $\mathcal{I}$ for the upper bound. When $\lim\limits_{n \to \infty} \frac{\sum_{i \in \mathcal{I}} P_{\mathrm{I}}}{W} = \infty$, N_I is asymptotically dominated by the interference. When the limit tends to 0, N_I is asymptotically dominated by thermal noise. The outcome of this limit can affect link rate scaling given the resources allocated to a link. Lemma 6.1 shows the scaling of a point-to-point channel with the $\mathcal{I}$ PSD defined above.

Lemma 6.1. *A point to point link serving a node u in a rich scattering environment*[1]*, with received power $P_{r_{\mathrm{u}}}$ and allocated bandwidth W_{u}, with ℓ_t transmission antennas and ℓ_r reception antennas, achieves scaling*

$$R_{\mathrm{u}} = \begin{cases} \Theta\left(W_{\mathrm{u}} \min(\ell_t, \ell_r)\right) & W_{\mathrm{u}} < W_{\mathrm{u}}^* \\ \Theta\left(\frac{\ell_r}{\min(\ell_t,\ell_r)} \frac{P_{r_{\mathrm{u}}}}{N_{\mathrm{I}}}\right) & W_{\mathrm{u}} \geq W_{\mathrm{u}}^* \end{cases} \tag{6.4}$$

for $W_{\mathrm{u}}^* = \Theta(\frac{\ell_r P_{r_{\mathrm{u}}}}{\min(\ell_t,\ell_r) W_{\mathrm{u}} N_{\mathrm{I}}})$ and for any of the three channel models described above

Proof. Appendix 6.A □

Lemma 6.1 applies to point to point channels, but cellular transmissions employ usually MAC and broadcast channels. The following lemmas prove that, although the exact capacity of these channels may not be achievable, its optimum scaling is obtained using simpler FDMA/FDM. Without loss of generality, we focus on the NFC model.

Lemma 6.2. *The capacity region of an n-user MAC channel and its rate region using FDMA have the same scaling when $\ell_r = o(n\ell_t)$.*

[1] A scattering such that the distribution of fading coefficients makes the channel matrix full-rank w.h.p.

Proof. In the power-limited case this is trivial. Otherwise, the sum-rate is limited by $\mathrm{I}(X_1 \dots X_n; Y)$, which scales as the degrees of freedom $W \min(n\ell_t, \ell_r)$ and is an outer bound of the capacity region. On the other hand, the aggregate rate of FDMA is an inner bound of the capacity region that scales as $\sum_{u=1}^{n} W_{\mathrm{u}} \min(\ell_t, \ell_r) \leq W \min(\ell_t, \ell_r)$. When $\ell_r = o(n\ell_t)$, for a sufficiently high n, both scalings converge. □

Lemma 6.3. *The capacity region of an n-user BC channel and its rate region using FDM have the same scaling when $\ell_t = o(n\ell_r)$.*

Proof. Same reasoning as above inverting the role of transmitters and receivers. □

6.2.2 Network Scaling

We consider the sequence of cellular wireless networks indexed by n, where n is the number of single-antenna nodes randomly distributed across area A with uniform probability. The network is supported by m BSs, with ℓ antennas each, and communication takes place over an increasing bandwidth W. In the Infrastructure Relay Hop (IRH) protocol defined below, there are also $k > m$ of wireless-backhauled relaying devices (RNs). The BSs are assumed to be placed according to a regular hexagonal layout, and the RNs, when present, are placed according to an hexagonal layout within each cell. The transmission power constraints of the nodes, the BSs and the RNs are P, P_{BS} and P_{RN}, respectively. The network is organized regularly in cells around each BSs with radius r_{cell} as in Fig. 6.1. Signal attenuates with distance according to path-loss exponent α and channels experience random small-scale fading (not necessarily Rayleigh, as power-limitation can occur for any distribution [75]). Channel state information is not a priori available to the terminals. The DL from the BS to the nodes and the UL from the nodes to the BS are implemented in alternate TDD frames. This imposes a $\frac{1}{2}$ penalty in rate but does not alter the scaling laws.

The rate achievable by any individual user is a random variable, and the capacity region of the network is a n-dimensional figure. The following definitions, adapted from [60], serve the purpose of defining a unidimensional deterministic characterization of the capacity region so that we can study its scaling. First, the definition of feasible rate reduces the problem to a random unidimensional variable.

Definition 6.4. A DL (UL) rate of $R_{\mathrm{DL}}(n)$ ($R_{\mathrm{UL}}(n)$) bits per second per node is *feasible* in a cellular network if all nodes can receive from (transmit to) the BS at least $R_{\mathrm{DL}}(n)$ ($R_{\mathrm{UL}}(n)$) bits per second.

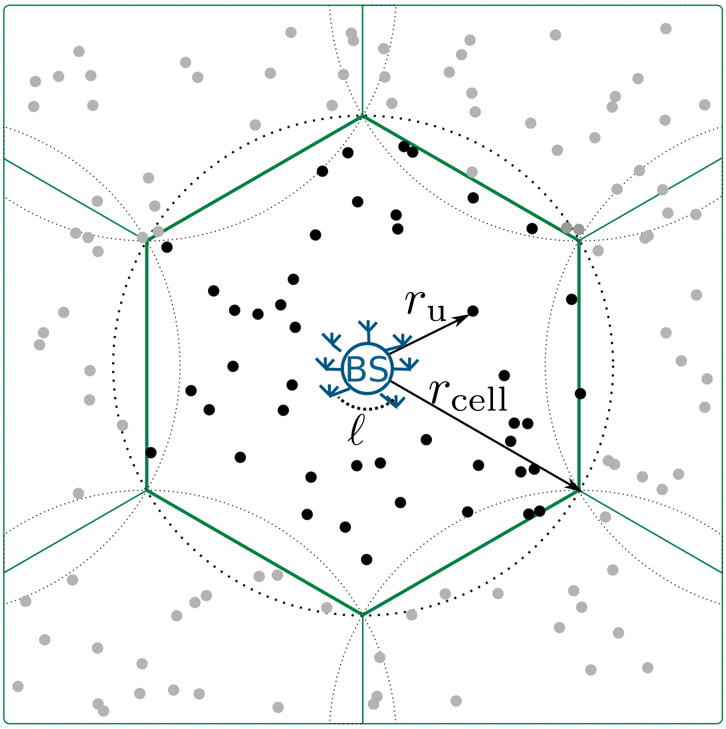


FIGURE 6.1: Network model. Only one cell and its neighbors are shown.

with this we are effectively measuring the scale of the n-dimensional arbitrarily-shaped region through the size of a n-dimensional hypercube contained in it. Secondly, we provide a definition of capacity scaling that avoids randomness.

Definition 6.5. We say that the DL (UL) *feasible rate capacity* of a set of random cellular networks $C_{\mathrm{DL}}(n)$ ($C_{\mathrm{UL}}(n)$) is of the order of $\Theta(f(n))$ bits per second per node if there are deterministic constants $c_1 < c_2$ such that:

$$\lim_{n\to\infty} P\left(R(n) = c_1 f(n) \text{ is feasible}\right) = 1 \tag{6.5}$$

$$\lim_{n\to\infty} P\left(R(n) = c_2 f(n) \text{ is feasible}\right) < 1 \tag{6.6}$$

Note that by using these definitions we are binding the scaling of a random rate region to the rate exponent that

1. can be offered simultaneously to all the nodes.
2. can be sustained almost in any realization of the distribution of node locations.

Its geometric interpretation is measuring the size of a random arbitrarily-shaped n-dimensional region through the largest deterministic n-dimensional hypercubes that with probability 1 fit into it. When adapting the definitions from [60], we have modified naming slightly to introduce cellular terminology for the sake of clarity.

We study the scaling of the feasible rate capacity $C(n)$ as $n \to \infty$ by finding an upper bound to the feasible rate with high probability, and specific protocols that can guarantee specific feasible rates. When these two elements coincide, capacity scaling is fully characterized. Table 6.1 defines the scaling relation between n and the different network parameters. The exponents of the number of BSs and BS antennas are taken from [67]. The constraint $\beta + \gamma \leq 1$ ensures that the number of infrastructure antennas per node does not grow without bounds. The scaling of the network area is as proposed by [63] to model a continuum of operating regimes between *dense* ($\nu = 0$) and *extended* ($\nu = 1$) networks. We introduced the bandwidth scaling exponent ψ, which satisfies that, for $\psi < 1$, bandwidth per node decreases as the number of nodes increases, while $\psi > 1$ represents asymptotically infinite bandwidth per node. Finally, we also introduce the scaling exponent $\rho \geq \beta$ of the number of RNs for the IRH protocol, that is based on fixed wireless relays, as previously described.

6.2.3 Protocols

6.2.3.1 Infrastructure Single-Hop

In the ISH protocol, BSs transmit directly to all destination nodes in the DL and all nodes transmit directly to the BSs in the UL. The signals that propagate between different cells are considered interference. There are $\frac{n}{m}$ nodes uniformly distributed within each cell. The ℓ BS antennas support multi-user MIMO (MU-MIMO) transmission, so that a BS can transmit or receive ℓ spatially separated streams at the same time.

To implement DL MU-MIMO precoding, the transmitter needs CSI (CSI-T). In systems with channel symmetry the DL transmitter can estimate the channel as a receiver during the UL phase, and vice-versa. If the channel is asymmetric, a feedback channel is needed, but this requirement is beyond the goal of our analysis. The BS can transmit ℓ orthogonal spatial streams and in each of them $\frac{n}{m\ell}$ nodes are further separated using FDMA in orthogonal bandwidths, so that $W_\mathrm{u} = W\ell\frac{m}{n}$ corresponds to user node u.

Table 6.1: Scaling Exponents of Network Parameters

Exponent	Range	Parameter (vs. no. of nodes n)
ψ	$[0, \infty)$	Bandwidth $W = W_0 n^{\psi}$
ν	$[0, 1]$	Area $A = A_0 n^{\nu}$
β	$[0, 1]$	No. of BSs $m = m_0 n^{\beta}$
γ	$[0, 1-\beta]$	No. of BS antennas $\ell = \ell_0 n^{\gamma}$
ρ	$[\beta, 1]$	No. of RNs $k = k_0 n^{\rho}$

This allocation is always possible because $\gamma < 1-\beta$, so nodes always outnumber antennas if n is sufficiently large. Also, the rank of the multi-user channel matrix is at least ℓ with high probability when nodes are separated at least a quarter of wavelength and far-field assumptions hold (i.e. we do not take the electromagnetic limitation of [62] into account). The BS transmits with power allocation $P_{\mathrm{u}} = P_{\mathrm{BS}}\frac{m}{n}$ per node.

Remark 6.6. As indicated by lemmas 6.2 and 6.3, orthogonal bandwidth allocation in each ISH stream with MU-MIMO is sufficient to attain the best scaling for ISH.

6.2.3.2 Infrastructure Multi-Hop

In the IMH protocol, each cell is subdivided regularly into smaller regions of area A_{r} called *routing sub-cells*, and information is forwarded from the BS via multi-hop communication using a node in each routing sub-cell as a relay as shown in Fig 6.2. For multi-hopping, the routing cells must contain at least one node with high probability, i.e. $A_{\mathrm{r}} > \frac{A}{m}\frac{2\log(\frac{n}{m})}{\frac{n}{m}}$ [67]. The BS uses MU-MIMO to start up to ℓ routing paths per transmission opportunity at the same time. For the remaining sub-cells, any single node forwards the data of a single path. Each hop covers distance d bounded by sub-cell radius (r_{subcell}), $d \leq 4r_{\mathrm{subcell}} \propto \sqrt{A_{\mathrm{r}}}$. Sub-cells alternate in becoming active using a non-scaling (i.e. constant) time or frequency division scheduling to avoid collisions and satisfy the half-duplex constraint.

6.2.3.3 Infrastructure Hierarchical Cooperation

The ad-hoc Hierarchical Cooperation (HC) protocol in [61] proceeds in three phases, dividing users in clusters of M users each. The hierarchical component arises from the fact that, according to the ad-hoc analysis [61], when the system employs a pre-existing protocol with scaling $\Theta(n^b)$ for intra-cluster user communications, the overall three-phase scheme scales better (with scaling exponent $\frac{1}{2-b} > b$) than the intra-subcell protocol employed. Hence, by recursive stacking of hierarchical layers and applying further subdivisions of subcells, a scaling exponent arbitrarily close to the unit $\Theta(n^{1-\epsilon})$ is achieved with a sufficient number of layers.

For cellular communication between nodes and BSs, we propose the equivalent Infrastructure Hierarchical Cooperation (IHC). We limit our design to one layer of the hierarchy and, unlike the ad-hoc case, we show that in a cellular network this "building block" cannot improve the given scaling of the underlying subordinate protocol. Therefore, it does not make sense to apply an IHC protocol to cellular communications.

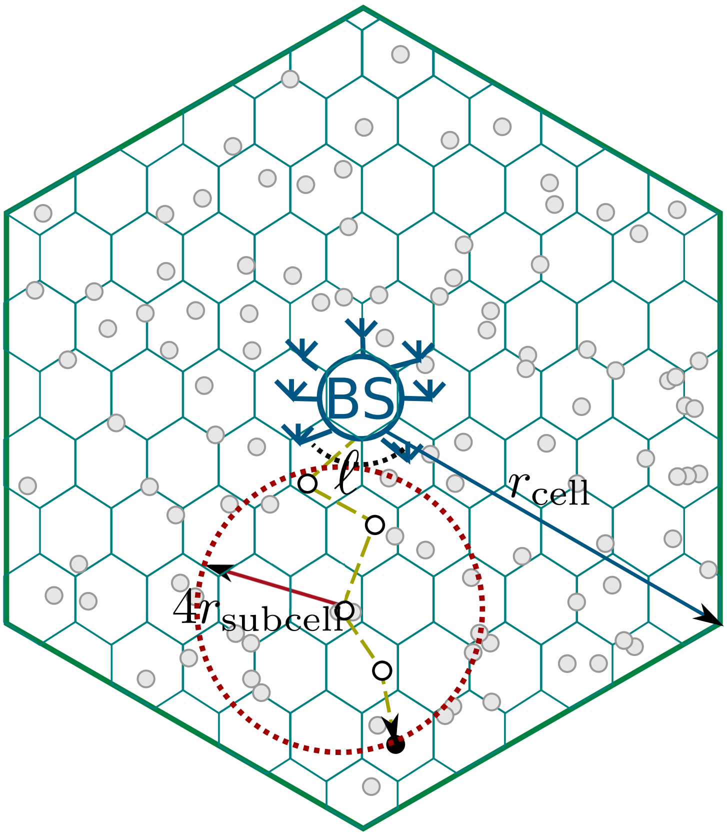


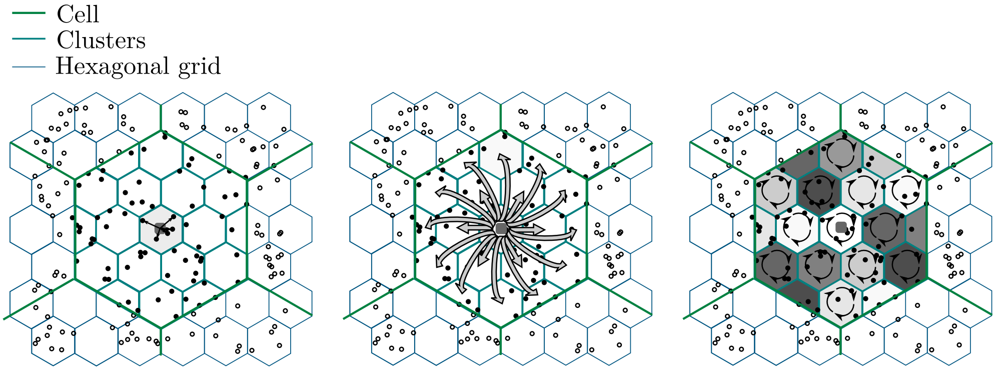
Figure 6.2: Routing in IMH. Only one cell is shown.

We begin by dividing the network area regularly in clusters of users, or μ-cells, with area $A_c = A\frac{M}{n}$, with $M \pm \delta$ users each w.h.p. The upper layer of the protocol has three phases that differ slightly from those of the ad-hoc case [61]. The DL scheme is detailed next. The definition of the UL scheme is analogous.

1. In the first phase each BS divides its data into M fragments and delivers a different one to each of its M neighbors in the BS subcell. A total of $\Theta(Mn^{1-\beta})$ bits of data have to be delivered, unlike in the ad-hoc case, where all subcells transmitted M messages and conveyed them to different destinations. Therefore the underlying protocol operates as a DL cellular system.

2. In the second phase, each subcell performs a $M \times M$ distributed MIMO transmission to the destination subcell. These transmissions must take place in each cell using cooperative single- or multi-hop transmissions in orthogonal time allocations. Unlike the ad-hoc case, in the cellular case the source subcells are not uniformly distributed: they are always those placed at the center of each cell, formed by the users that are closest to the corresponding BS.

3. In the third phase all nodes within the destination subcell perform quantize-and-forward relaying, conveying their observations of the **received signal** to the destination node. The destination node, after gathering all the observations across the virtual antenna array, performs a $M \times M$ decoding process to extract the messages. Since information is exchanged between any pair of nodes, the underlying protocol operates as an ad-hoc system.



(a) Phase 1: BS subcell uses a DL protocol.

(b) Phase 2: Cooperative intra-sub-cell DL.

(c) Phase 3: All subcell suse an ad-hoc protocol.

FIGURE 6.3: Three-phase protocol constituting one layer of the hierarchical scheme.

6.2.3.4 Infrastructure Relay-Multi-Hop

In the *Infrastructure Relay-Multi-Hop* (IRH) protocol, the network area is divided regularly into $m + k$ μcells, smaller than a cell. BSs and RNs are distributed regularly and each of them is in charge of its own μcell. Time is divided in *access* and *interconnection* phases, with relative durations $\tau_\mathrm{a} \in [0, 1]$ and $1 - \tau_\mathrm{a}$.

- In the *Access Phase*, for a fraction $\tau_\mathrm{a} \in [0, 1]$ of the time, **at each μcell**, all APs (RNs and BSs), exchange data with the user nodes using an ISH protocol. Signals that propagate between different μcells are treated as interference. There are $\frac{n}{m+k} \pm \delta$ nodes within each μcell w.h.p. Unlike BSs, RNs do not have ℓ antennas and therefore feasible rates in RN μcells are lower, so that they become bottlenecks. BSs allocate single-antenna transmissions to user nodes u on orthogonal bandwidths $W_\mathrm{u} = W\frac{m+k}{n}$. Their multi-antenna resources are partially underused, but the loss is negligible as $k \gg m$. BSs and RNs allocate power according to $P_\mathrm{u} = P_\mathrm{BS}\frac{m+k}{n}$ and $P_\mathrm{u} = P_\mathrm{RN}\frac{m+k}{n}$ splits, respectively.
- In the *Interconnection Phase*, for a fraction $1 - \tau_\mathrm{a}$ of the time, BSs exchange data with RNs using the IMH protocol. Each μcell with area $A_\mathrm{r} \sim n^{\nu-\rho}$ becomes the *routing sub-cell* of IMH, and information is forwarded from the BS via multi-hop

communication using the single RN in each routing sub-cell as relay as shown in Fig 6.4. The BS starts up to ℓ routing paths per transmission opportunity at the same time using MU-MIMO. Each hop covers distance d of exactly two μcell radiuses ($2r_{\mu\text{cell}} \propto \sqrt{A_\text{r}}$). μCells become active alternately following a non-scaling (i.e. constant) time or frequency division scheduling to avoid collisions and satisfy the half-duplex constraint.

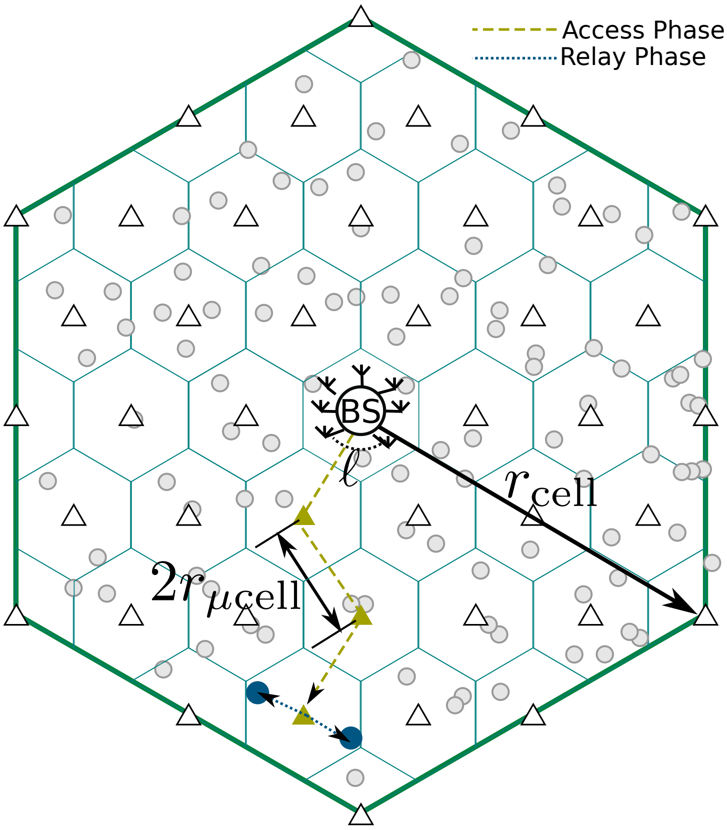


FIGURE 6.4: The two phases of IRH. In the Access Phase IMH routing is used across the cell. In the Relay Phase direct ISH tranmission is used within each μcell.

Remark 6.7. IRH phases employ protocols with different scalings. If the Access Phase has lower scaling exponent, the optimal partition is $\lim_{n\to\infty} \tau^* = 1$. If the Access Phase has higher scaling exponent, the optimal partition is $\lim_{n\to\infty} \tau^* = 0$. If they both have the same exponent, the optimal partition is irrelevant for scaling. In any of these three cases, the resulting scaling is the minimum of the two protocols.

Note that in IRH we model RNs as IMH nodes, without a scaling parameter for their number of antennas. Since in the close future it is likely that BS will still have many more antennas than nodes or RNs, we leave for future work the study of the scaling of a heterogeneous network with massive RN MIMO capability. Such study would first require the generalization of the IMH analysis for user nodes with massive MIMO capability, which is nowadays unrealistic due to size constraints.

6.2.4 Cut-set Bound

The protocols above treat interference as noise and deal with the worst-case transmission distances due to uniform user distribution in space. Conversely, the upper bound we employ to confine the scaling of capacity considers the most favorable scenario that still occurs with probability one in the limit as $n \to \infty$. We employ a cut-set bound that separates space in two regions, represented in Fig. 6.5.

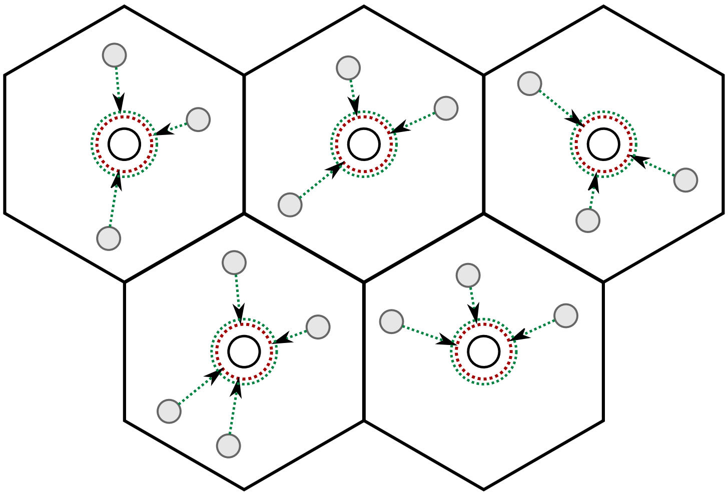


FIGURE 6.5: Cut set bound (red) and best-case transmission distance (green).

The first region is defined as the union of all the circles with radius $n^{\frac{1-\nu}{2}}$ around each BS location (represented in red). This region contains all BSs by definition and, as the probability that one user lies at a distance closer than $n^{\frac{1-\nu}{2}}$ to the BS tends to zero as $n \to \infty$. Each of these circles contain zero user nodes w.h.p. The second region is defined as the complementary of the first and contains most of the area of all the cells and all user nodes.

Besides of defining this cut of the network, when the communications become power-limited we consider the best-case transmission distances by supposing that, instead of at their actual spatial distribution, all nodes are located as close to the border as possible while still being power limited. This distance, represented in green in Fig. 6.5, is at least the radius of the cut $n^{\frac{1-\nu}{2}}$. Moreover, we consider this distance is observed towards **all** BSs, not only to the closest BS. This takes a little imagination effort, as one has to imagine that each node exists at the same time in various locations in each of the green circles in Fig. 6.5.

Finally, we consider best-case interference processing as well, by assuming that all BSs can cooperate perfectly at one side of the cut, and that all nodes can cooperate perfectly at the other. This converts the communication problem into an equivalent massive

MIMO point-to-point channel where all BSs and nodes put together all their antennas at both cut sides.

6.3 Capacity Scaling Laws in Future Cellular Networks

6.3.1 Capacity

We first present our result for the scaling of *feasible rate capacity.* We remark that this scaling is deterministic and defined at the frontier between the rates that are feasible w.h.p. and the rates that are feasible only with a probability less than one.

Theorem 6.8. *The rate from the BS to each node is upper bounded by a function with scaling*

$$R_{\mathrm{DL}(n)} \leq \Theta\left(n^{\beta-1+\min\left(\psi+\gamma,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.7}$$

and the rate from a node to the BS is upper bounded by a function with scaling

$$R_{\mathrm{UL}(n)} \leq \Theta\left(n^{\min\left(\psi+\beta+\gamma-1,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.8}$$

with probability $\lim\limits_{n\to\infty} P \to 1$.

Proof. Section 6.B.1 □

Theorem 6.9. *IMH DL feasible rate per node scales as*

$$R_{\mathrm{IMH-DL}}(n) \sim \Theta\left(n^{\beta-1+\min\left(\psi+\gamma,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.9}$$

and IMH UL feasible rate per node scales as

$$R_{\mathrm{IMH-UL}}(n) \sim \Theta\left(n^{\Theta(n^{\beta+\gamma-1+\min(\psi,\frac{\alpha}{2}(1-\nu))})}\right) \tag{6.10}$$

Proof. Section 6.B.2 □

Theorem 6.10. *The feasible rate capacity from the BS to each node scales as*

$$C_{\mathrm{DL}(n)} = \Theta\left(n^{\beta-1+\min\left(\psi+\gamma,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.11}$$

and when $\beta+\gamma=1$ *the capacity from a node to the BS scales as*

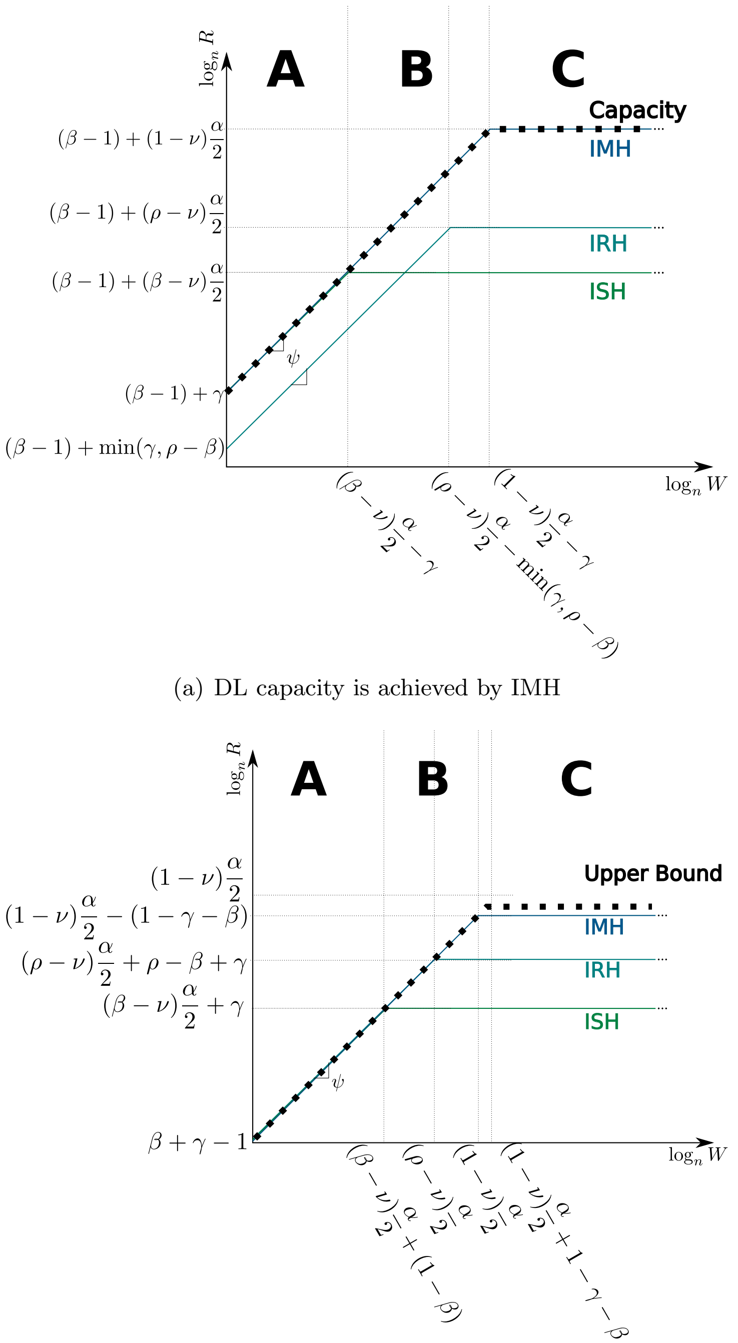
$$C_{\mathrm{UL}(n)} = \Theta\left(n^{\min\left(\psi,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.12}$$

Proof. Geometrically, Theorem 6.9 means that there exists a n-dimensional hypercube contained in the capacity region with probability one that scales as stated. Therefore, the given scaling exponent satisfies the "achievability" part of our definition of capacity scaling (6.5). Respectively, Theorem 6.8 means that there is no n-dimensional hypercube that contains the capacity region with probability one and scales larger than stated. Thus, the given scaling exponent satisfies the "there is no larger" part of our definition of capacity scaling (6.6). In the DL, both values coincide and consequenty the given exponent is the feasible rate capacity scaling by definition. In the UL, the two values differ only in an amount $\beta + \gamma - 1$ so they also coincide by adding the constraint $\beta + \gamma = 1$. □

Corollary 6.11. *IMH achieves capacity scaling in all DL regimes, and in all UL regimes when* $\beta + \gamma = 1$.

In the DL, when $\psi + \gamma < \frac{\alpha}{2}(1 - \nu)$, effective noise PSD ($N_{\mathrm{I}}$ in (6.3)) becomes asymptotically dominated by interference in BS-node links within the BS routing subcell. The capacity of the network is limited by the degrees of freedom, determined by the number of BS, the number of transmission antennas and the bandwidth. Conversely, when $\psi + \gamma > (1 - \nu)\frac{\alpha}{2}$, in IMH N_{I} is dominated by noise. Capacity is power-limited and received power is determined by path-loss, inter-node distances and BS density, but not by BS-node distance, bandwidth or number of transmission antennas. In the UL, when $\beta + \gamma = 1$, the same holds by taking the scaling of the number of antennas γ out of the threshold. The scaling $\beta + \gamma < 1$ represents a network where, asymptotically, each BS antenna serves infinite users, and even though it can be formulated in theory it has less practical interest than the case of $\beta + \gamma = 1$.

Fig. 6.6(a) illustrates the scaling exponents of capacity, IMH and other protocols in the DL case. The horizontal axis is the sum of the exponents of bandwidth and number of transmission antennas, $\psi + \gamma$, which represents the scaling of the degrees of freedom of BS signals. The vertical axis represents the exponent of the feasible node rate $\log(R(n))$. When bandwidth scales above the threshold $\frac{\alpha}{2}(1 - \nu) - \gamma$, the system is underpowered and capacity cannot grow with the degrees of freedom. Figure 6.6(b) illustrates the same for the UL case, where the number of BS antennas is removed from the critical bandwidth scaling. In the figure it is possible to see the gap between IMH and the upper bound if $\beta + \gamma < 1$. However, this relationship implies that the infrastructure antennas per user vanish and, instead, it is is usually considered as an equality in the literature [66–68]. The other protocols illustrated in Figures 6.6(a) and 6.6(b) are discussed later.



(a) DL capacity is achieved by IMH

(b) UL capacity lies in a small gap between the upper bound and IMH capacity.

FIGURE 6.6: Scaling exponents for capacity upper bounds, IMH, ISH and IRH.

6.3.2 Protocols that are not Guaranteed to Achieve Capacity

We now present our results for the feasible rate of practical protocols. Hereafter, scalings are deterministic by assuming worst-case transmission distances to obtain rates that are always feasible with probability 1 even for a finite number of nodes.

Theorem 6.12. *ISH DL feasible rate per node scales as*

$$R_{\text{ISH}-\text{DL}}(n) \sim \Theta\left(n^{\beta-1+\min\left(\psi+\gamma,(\beta-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.13}$$

and ISH UL feasible rate per node scales as

$$R_{\text{ISH}-\text{UL}}(n) \sim \Theta\left(n^{\beta-1+\gamma+\min\left(\psi,(\beta-\nu)\frac{\alpha}{2}+(1-\beta)\right)}\right) \tag{6.14}$$

Proof. Subsection 6.C.1. □

In the ISH DL, when $\psi + \gamma < (\beta - \nu)\frac{\alpha}{2}$, effective noise PSD ($N_{\text{I}}$ in (6.3)) becomes asymptotically dominated by interference. The feasible rate of the network is limited by the degrees of freedom, determined by the number of BSs, the number of transmission antennas and the bandwidth. Conversely, when $\psi + \gamma > (\beta - \nu)\frac{\alpha}{2}$, N_{I} is asymptotically dominated by noise and rate is power-limited. Reception power is determined by path-loss, node-BS distances and BS density, but not by inter-node distance, bandwidth or number of transmission antennas. In the ISH UL the same holds except for the role of the BS antennas, which allow for reception gain and do not affect critical bandwidth.

Figures 6.6(a) and 6.6(b) illustrate that the main difference between ISH and IMH is that the role of inter-node distance $(1 - \nu)\frac{\alpha}{2}$ is replaced by BS-node distances $(\beta - \nu)\frac{\alpha}{2}$ in the power-limited regime. This results in a power loss for ISH that, in turn, decreases its maximum bandwidth scaling, accelerating the change from the degrees-of-freedom-limited regime to the power-limited regime in case of direct transmission with large bandwidths. Note that:

- For $\psi + \gamma < \frac{\alpha}{2}(\beta - \nu)$, IMH and ISH are both degrees-of-freedom-limited. The rates of the two protocols do not differ in terms of scaling.
- For $\psi + \gamma > \frac{\alpha}{2}(\beta - \nu)$, IMH outperforms ISH. In the range $\frac{\alpha}{2}(\beta - \nu) < \psi + \gamma < \frac{\alpha}{2}(1-\nu)$ IMH remains degrees-of-freedom-limited, but ISH is power-limited. Unlike for ISH, IMH rate still grows with the bandwidth exponent.
- For $\psi + \gamma > \frac{\alpha}{2}(1 - \nu)$, both IMH and ISH are power limited. IMH rate ceases to grow with the bandwidth exponent but its exponent is larger than that of ISH, since IMH benefits from a power gain due to shorter transmission ranges.

In the degenerate case $\beta = 1$ every node can have a non-scaling dedicated channel to a BS, scaling is trivially linear and ISH direct transmissions achieve the optimal scaling. However, this implies that MIMO transmission gains cannot scale ($\gamma \leq 1 - \beta = 0$). For $\beta < 1$ (i.e. increasing number of users per BS), feasible IMH rate scaling outperforms that of ISH for large bandwidth scaling exponents (ψ) and direct transmissions are only efficient for small bandwidths.

Theorem 6.13. *It is not possible to construct a layer of a cellular IHC protocol that, using IMH as subordinate intra-cluster protocol, improves IMH performance.*

Proof. The IHC protocol can not surpass the upper bound achieved by IMH. So the IHC protocol performs, at best, as the IMH protocol. □

Although the result of this theorem stems directly from the upper bound in the proof of capacity scaling, there is great value in understanding *why* hierarchical cooperation fails in a cellular network with the right traffic model (CDFM). We thus analyzed IHC focusing on the root causes for hierarchical cooperation to fail in the cellular case. Section 6.4.1 describes the construction and analysis of an IHC protocol and demonstrates in lemmas 6.15 through 6.19 that the root cause of the lack of benefits from hierarchical cooperation in a cellular scenario is the CDFM model, which, compared to ADFM, concentrates the spatial distribution of traffic in short links between nodes and their closest BS, instead of uniformly distributing traffic across the whole network.

Theorem 6.14. *IRH DL feasible rate scales as*

$$R(n) \leq \Theta\left(n^{\beta-1+\min\left(\psi+\min(\gamma,\rho-\beta),(\rho-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.15}$$

and IRH UL rate scales as

$$R(n) \leq \Theta\left(n^{\beta-1+\min\left(\psi+\min(\gamma,\rho-\beta),(\rho-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.16}$$

Proof. Subsection 6.C.2. □

IRH can be as suboptimal as ISH when $\rho = \beta$, because theorem 6.14 reduces its scaling to that of ISH without multiple BS antennas ($\gamma = 0$). In this case a single RN with one antenna replaces each BS and the scaling of everything else is the same, so the introduction of RNs is of little help. On the other hand, ISH can be as optimal as IMH when $\rho = 1$, as theorem 6.14 allows ISH to achieve the scaling of IMH. In this case each user has a dedicated RN nearby to serve its data. This has strong advantages from a practical point of view over the implementation of user cooperation: it enables to attain IMH feasible rate scaling with *fixed* relays, without imposing users the battery drain and hardware complexity of a mobile multi-hop implementation.

Even though ISH turns out to be as good as IMH for $\beta = 1$, it is likely that there are more limitations in the deployment of wired backhauling to increase β than that to increase

ρ. Moreover, if $\rho = 1$, then the term $\min(\gamma, \rho - \beta)$ becomes $\min(\gamma, 1 - \beta) = \gamma$, and IRH achieves the same feasible rate scaling as IMH without the need of user collaboration, respecting the constraints on BS density. Figures 6.6(a) and 6.6(b) illustrate the scalings. A significant intermediate point is given by the effect of the term $\min(\gamma, \rho - \beta)$ in IRH DL in the intermediate case $\rho < 1$. In this case IRH is strictly worse than ISH for small bandwidths, which is coherent with the struggle of relaying implementations so far to achieve significant gains as we showed on Chapter 3, but the situation is the opposite for ψ large enough. In that case relaying outperforms direct transmissions thanks to the large available bandwidth.

6.4 Why IHC Fails: the Importance of the Traffic Model

6.4.1 Analysis of the IHC Protocol

To model the IHC protocol, we consider first that every cell in the network has $\Theta(n^{1-\beta})$ users, an area $\Theta(n^{\nu})$ and a user density of $\Theta(n^{1-\nu})$. Let us divide each cell into $\frac{n^{1-\beta}}{M}$ regular subcells. Each subcell has $M \pm \delta$ users with $\delta \to 0$ w.h.p. As in [61], for simplicity we develop the analysis with cells and subcells comprising exactly $n^{1-\beta}$ and M users respectively. The difference with the actual model disappears as n tends to infinity. Note that a value for M is yet to be selected.

The aggregate network sum-rate can be expressed as the total amount of data $D(n)$ transported to the destinations divided by the total time spent in the three phases of the protocol τ_1, τ_2 and τ_3.

$$T(n) = \frac{D(n)}{\tau_1(n) + \tau_2(n) + \tau_3(n)} \tag{6.17}$$

- The total amount of data is M messages per user (or M bits if we normalize the time frames by message lengths) times the number of users n.

$$D(n) = Mn \tag{6.18}$$

- The first phase carries $\Theta(n^{1-\beta}M)$ bits from each BS to the M nodes in its subcell using a pre-existing DL protocol with rate $\Theta(n^{b_m})$. Since this protocol is adjusted from a larger cell to a smaller subcell with only M nodes and one BS, we must first obtain its scaling in the new context

$$\begin{aligned} b'_m &= b_m\left(\alpha, \beta', \psi', \nu'\right)\right), \\ \beta' &= 0 \\ \psi' &= \psi \log_M n \\ \nu' &= 1 + (\nu - 1)\log_M n, \end{aligned} \tag{6.19}$$

and then we compute the duration of this phase

$$\tau_1(n) = n^{1-\beta} M^{1-b'_m} \tag{6.20}$$

The gain obtained by replacing a cell-wise DL system with an intra-subcell DL phase is due to the division of the network into subcells of area $A_c \propto M/nA$, and therefore the hierarchical DL phase has much shorter range and achieves new scaling parameters $\log_M W = \psi' = \psi \log_M(n)$, $\log_M \frac{M}{n} m = \beta' = 0$, $\log_M A_c = \nu' = 1 + (\nu - 1)\log_M(n)$, increasing SNR but also bandwidth-per-user compared to a direct DL scheme. This could increase the available power in the power-limited regime, or enable the exploitation of additional degrees of freedom by extending the operation of the degrees-of-freedom-limited regime.

However, we are interested in starting the hierarchy with our best non-hierarchical protocol at hand, IMH [11], so in practice we should use $b_m = \beta + \min(\psi, (1-\nu)\frac{\alpha}{2})$ and after applying the new parameters this gives $b'_m = \log_M n \min(\psi, (1-\nu)\frac{\alpha}{2})$. It follows that $b'_m = b_m \log_M(n)$ and, since $\log_M(n) < 1$, $b'_m \leq b_m$ and therefore the bandwidth excess dominates the power gain.

- The second phase needs to carry $\Theta(Mn^{1-\beta})$ bits per cell. We analyze this phase as if the user cluster in each subcell was a virtual node, and a multi-hop cellular protocol was applied to obtain the rates between clusters. The feasible rate in each of these $M \times M$ transmissions scales at least as $\Theta(\min(Mn^{\psi}, n^{\frac{\alpha}{2}(1-\nu)}))$, following a similar analysis as in [11] with $\ell = M$. An operation rate of $\Theta(\min(Mn^{\psi}, n^{\frac{\alpha}{2}(1-\nu)}))$ bits per channel use is achievable.

$$\tau_2(n) = \begin{cases} n^{1-\psi-\beta} & \psi < \frac{\alpha}{2}(1-\nu) - \log_n M \qquad (6.21a) \\ Mn^{1-\frac{\alpha}{2}(1-\nu)-\beta} & \psi > \frac{\alpha}{2}(1-\nu) - \log_n M \qquad (6.21b) \end{cases}$$

- In the third phase the observed signal is quantized at each receiver with Q bits, for a total of M received messages per cell and M different receivers each. There is no centralized coordination, an ad-hoc protocol with scaling $\Theta(M^{b_a})$ is applied.

  In this phase, ad-hoc hierarchical, bursty-hierarchical and multi-hop protocols may be optimal depending on the reception power available between nodes. The spectral density of received power scales with inter-node distance and bandwidth $\Theta(n^{(1-\nu)\frac{\alpha}{2}-\psi})$, while the HC protocol has a power saving of n^{-1}. Hierarchical cooperation is therefore directly usable when $(1-\nu)\frac{\alpha}{2}-\psi > -1$, and usable a fraction $\Theta(n^{1+(1-\nu)\frac{\alpha}{2}-\psi})$ of the time otherwise. Non-hierarchical multi-hop achieves scaling $\Theta(n^{\frac{1}{2}+\frac{\alpha}{2}(1-\nu)})$. Assuming that h layers of the HC protocol are available, the rate in this phase is

$$b_a(h) = \begin{cases} \dfrac{h}{h+1} & \psi < 1+(1-\nu)\frac{\alpha}{2} \quad (6.22a) \\ \dfrac{h}{h+1} + 1 + (1-\nu)\dfrac{\alpha}{2} - \psi & \frac{3}{2} > \psi > 1+(1-\nu)\frac{\alpha}{2} \quad (6.22b) \\ \dfrac{1}{2} + (1-\nu)\dfrac{\alpha}{2} & \psi > \frac{3}{2} \quad (6.22c) \end{cases}$$

Note that $\nu > \frac{\alpha-1}{\alpha}$ is necessary for the intermediate regime to exist.

The result is

$$\tau_3(n) = QM^{2-b_a} \tag{6.23}$$

We now analize the six possible combinations that arise from the three possible values of τ_3 with as many variants of b_a and the two possible values for τ_2, where cell-wise cooperative MIMO communications are interference dominated or power limited.

6.4.1.1 Case (6.21a) and (6.22a)

We first consider the case when the subordinate ad-hoc hierarchical protocol can achieve unit scaling $\lim_{h\to\infty} b_a(h) = \lim_{h\to\infty} \frac{h}{h+1} = 1$, when $\psi \leq 1+(1-\nu)\frac{\alpha}{2}$, and inter sub-cell MIMO communications are limited by the degrees of freedom, giving $\psi < \frac{\alpha}{2}(\beta-\nu) - \log_n M$. Using $b_a = 1$ we compute the optimal M^* and the resulting feasible rate scaling.

Lemma 6.15. *A hierarchical BC DL in regime* (6.21a) *and* (6.22a) *performs optimally when $M^* = n$. This makes the sum rate scaling $T(n)$:*

$$T(n) \sim \Theta\left(n^{\min(\beta+b'_m,1)}\right). \tag{6.24}$$

And since we can use IMH as a subordinate protocol ($b'_m = \min(\psi,(1-\nu)\frac{\alpha}{2})$ in this case), the hierarchical protocol never outperforms IMH because the respective scalings satisfy

$$b_m = \beta + \min(\psi,(1-\nu)\frac{\alpha}{2}) \geq \min(\beta+\min(\psi,(1-\nu)\frac{\alpha}{2}),1) \tag{6.25}$$

Proof. Appendix 6.D.1 □

6.4.1.2 Case (6.21b) and (6.22a)

In this case $b_a(h) = 1$ when $\psi \leq 1 + (1-\nu)\frac{\alpha}{2}$, and inter sub-cell MIMO communications are limited by power, giving $\psi > \frac{\alpha}{2}(1-\nu) - \log_n M$. The two constraints are always compatible because $M \geq 1$ (for non-empty subcells), so $\log_n M \geq 0$. We compute the optimal M^*.

Lemma 6.16. *A hierarchical BC DL in regime* (6.21b) *and* (6.22a) *performs optimally when $M^* = n$. This makes the rate scaling $T(n)$:*

$$T(n) \sim \Theta\left(n^{\min(\beta+b'_m,1)}\right). \tag{6.26}$$

And, since we can use IMH as a subordinate protocol ($b'_m = \min(\psi,(1-\nu)\frac{\alpha}{2})$ in this case), the hierarchical protocol never outperforms IMH because the respective scalings satisfy

$$b_m = \beta + \min(\psi,(1-\nu)\frac{\alpha}{2}) \geq \min(\beta + \min(\psi,(1-\nu)\frac{\alpha}{2}),1) \tag{6.27}$$

Proof. The proof is similar to appendix 6.D.1 in structure but the optimization over M is easier in this regime. □

6.4.1.3 Case (6.21a) and (6.22b)

This case would require $\psi > 1 + (1-\nu)\frac{\alpha}{2}$ and $\psi < (1-\nu)\frac{\alpha}{2} - \log_n M$ by the definition of cases (6.21a) and (6.22b). These cannot hold together because $\log_n M \geq 0$.

6.4.1.4 Case (6.21b) and (6.22b)

This case requires $\psi > 1+(1-\nu)\frac{\alpha}{2}$ and $\psi > (1-\nu)\frac{\alpha}{2} - \log_n M$, where the first constraint implies the second. Using $b_a = 2 + (1-\nu)\frac{\alpha}{2} - \psi$:

Lemma 6.17. *A hierarchical BC DL in case* (6.21b) *and* (6.22b) *performs optimally when $M^* = \frac{b'_m}{Q[\psi-1-(1-\nu)\frac{\alpha}{2}]} n^{\frac{1-\beta}{\psi-1-(1-\nu)\frac{\alpha}{2}+b'_m}}$. This makes the rate scaling $T(n)$:*

$$\begin{aligned} T(n) &= \Theta\left(n^{\beta+\min(\chi(b'_m),(1-\nu)\frac{\alpha}{2})}\right), \\ \chi(b'_m) &= (1-\beta)\frac{b'_m}{\psi - 1 - (1-\nu)\frac{\alpha}{2} + b'_m} \end{aligned} \tag{6.28}$$

And, since in this regime $\chi(b'_m) < 1 < \psi$, the hierarchical protocol never outperforms IMH because the respective scalings satisfy:

$$b_m = \beta + \min(\psi,(1-\nu)\frac{\alpha}{2}) \geq \beta + \min(\chi(b'_m),(1-\nu)\frac{\alpha}{2}) \tag{6.29}$$

Proof. Appendix 6.D.2 □

6.4.1.5 Case (6.21a) and (6.22c)

This case requires $\psi > \frac{3}{2}$ and $\psi < (1-\nu)\frac{\alpha}{2} - \log_n M$ by the definition of cases (6.21a) and (6.22c). The ad-hoc scaling is $b_a = \frac{1}{2} + (1-\nu)\frac{\alpha}{2}$.

Lemma 6.18. *A hierarchical BC DL in regime* (6.21a) *and* (6.22c) *performs optimally when $M^* = n$. This makes the rate scaling $T(n)$:*

$$T(n) \sim \Theta\left(n^{\min(\beta+b'_m, 1+\beta+\psi, \frac{1}{2}+(1-\nu)\frac{\alpha}{2})}\right). \tag{6.30}$$

and, since we can use IMH as a subordinate protocol ($b'_m = \min(\psi, (1-\nu)\frac{\alpha}{2})$ in this case), and cases (6.21a) *and* (6.22c) *require $\frac{1}{2} + (1-\nu)\frac{\alpha}{2} < 1$, the hierarchical protocol never outperforms IMH because the respective scalings satisfy*

$$b_m = \beta + \min(\psi, (1-\nu)\frac{\alpha}{2}) \geq \min(\beta + \min(\psi, (1-\nu)\frac{\alpha}{2}), \frac{1}{2} + (1-\nu)\frac{\alpha}{2}) \tag{6.31}$$

Proof. Appendix 6.D.3 □

6.4.1.6 Case (6.21b) and (6.22c)

This mode requires $\psi > \frac{3}{2}$ and $\psi > (\beta-\nu)\frac{\alpha}{2} - \log_n M$.

Lemma 6.19. *A hierarchical BC DL in regime* (6.21b) *and* (6.22c) *performs optimally when*

$$M^* = \begin{cases} n & (1-\nu)\alpha > 1 \\ \frac{b'_m}{Q\left[\frac{1}{2}-(1-\nu)\frac{\alpha}{2}\right]} n^{\frac{1-\beta}{\frac{1}{2}-(1-\nu)\frac{\alpha}{2}+b'_m}} & (1-\nu)\alpha < 1 \end{cases} \tag{6.32}$$

making the rate scaling $T(n)$:

$$T(n) \sim \begin{cases} \Theta\left(n^{\min(\beta+b'_m, 1+\beta+(1-\nu)\frac{\alpha}{2}, \frac{1}{2}+(1-\nu)\frac{\alpha}{2}}\right) & (1-\nu)\alpha > 1 \\ \Theta\left(n^{\beta+\min(\chi(b'_m),(1-\nu)\frac{\alpha}{2})}\right) & (1-\nu)\alpha < 1 \\ \quad \chi(b'_m) = (1-\beta)\left[\frac{b'_m}{\frac{1}{2}-(1-\nu)\frac{\alpha}{2}+b'_m})\right] \end{cases} \tag{6.33}$$

and, since we can use IMH as a subordinate protocol ($b'_m = \min(\psi, (1-\nu)\frac{\alpha}{2})$ in this case), $\chi(b'_m) < 1 < \frac{3}{2} < \psi$, the hierarchical protocol never outperforms IMH because the respective scalings satisfy:

$$b_m = \beta + \min(\psi, (1-\nu)\frac{\alpha}{2}) \geq \begin{cases} \min(\beta + \min(\psi, (1-\nu)\frac{\alpha}{2}), \frac{1}{2} + (1-\nu)\frac{\alpha}{2}) \\ \beta + \min(\chi(b'_m), (1-\nu)\frac{\alpha}{2}) \end{cases} \tag{6.34}$$

Proof. Appendix 6.D.4 □

6.4.2 Spatial Interpretation

The main difference between the IHC and HC protocols [61] is the nature of data flows. In the CDFM model, flows originate from BSs instead of from nodes. In a spatial interpretation, this means that data sources are concentrated in fixed points in space, rather than uniformly distributed across the whole network. In ADFM, if the space is divided in small subcells of size M, the data demand is also divided between all those subcells, proportionally to subcell area M. Therefore, the smaller the divisions are, the less data each subcell needs to handle, allowing for the asymptotic properties of the HC protocol when infinite cooperative layers are stacked.

On the other hand, in CDFM, dividing cells in subcells of size M will create $\frac{n-n^{\beta}}{M}$ subcells without any a-priori traffic load and $\frac{n^{\beta}}{M}$ subcells with the fixed traffic load of n^{β} data flows from the full cell, regardless of the division size M. Let us take the equations for the case with (6.21a) and (6.22a) as an example. In HC, the term M^{2-b_a} appears in the analogous expression to $n^{1-\beta}M^{1-b'_m}$ for IHC in (6.60). When we take the derivative of this in lemma 6.15, a non-negative term of the form $b'_m M^{1-b'_m} \times \dots$ appears that forces the whole derivative to be positive and ultimately leads to $M^* = n$. In the analogous case for HC this step of the optimization would produce a negative coefficient $(b_a - 1)M^{2-b_a}$ that would allow the derivative to be zero once in the range $0 < M^* < n$. In other words, the change in the data flow demand from M^{2-b_a} to $n^{1-\beta}M^{1-b'_m}$ per subcell modifies the fundamental dynamics of rate evolution with hierarchical cooperation as a function of the cell to subcell ratio, converting a convex function into a monotonous one. This fundamental modification by changing the data flow model converts a few subcells into bottlenecks that must carry the traffic of the whole cell regardless of sub-cell size. The regime is limited by the relation between the number of degrees of freedom available and the traffic load. In ADFM, the shrinking traffic demand with M allows to reduce subcell size, whereas in CDFM the constant traffic regardless of M in the subcell of the BS makes including as many users as possible in that subcell the best strategy, turning that subcell into the whole cell again.

In the regimes corresponding to the other cases, the dynamics of the solution are affected in a similar manner. Finally, we can state that, in CFDM, the concentration of traffic flows around a particular point of each cell is a major difference with ADFM that makes hierarchical cooperation worthless in a cellular network.

6.5 Summary

As cellular networks evolve the number of communicating devices will grow, increasing node density. There is a trend towards increasing cell density as well, which is limited by BS back-hauling constraints, so that wireless RNs may be necessary for the densities of APs and user nodes to be comparable. The number of antennas per BS will also grow. And broad new regions of spectrum will be made available in millimeter wave bands, as well as via opportunistic access to lower-frequency bands through cognitive radio. For all these reasons, next-generation cellular capacity will not evolve as traditional cellular systems, probably entering regimes with plenty of degrees of freedom, where power limitation will play a more significant role in network capacity.

In this chapter we have obtained the **capacity scaling** of cellular wireless networks. The result is a versatile model comprising the scaling of area, BS density, number of antennas per BS, and total available bandwidth. Capacity scaling laws have been obtained by finding coincident scaling exponents for upper and lower bounds, the first being an idealized cooperative MIMO channel involving all transmitters and receivers, and the second being the feasible rate under the IMH protocol. Capacity is limited by a *critical scaling exponent of bandwidth*, influenced by received power and, ultimately, by the other factors that scale: inter-node distance, number of antennas, number of user per cell, etc. Below the critical bandwidth, capacity scales with the degrees of freedom, but above this threshold increasing bandwidth no longer allows to increase the rate per node. This critical threshold coincides exactly with the bandwidth where power is so sufficiently spread that links are mostly influenced by thermal noise, rather than by interference, and the network changes from being limited by degrees-of-freedom to being power-limited. All the studied protocols have a protocol-specific critical bandwidth scaling and, even though for a specific value on the scaling parameters all protocols can achieve capacity, only the IMH protocol achieves capacity scaling in the general case for all the range of values. This protocol is therefore necessary to exploit bandwidth exploitation in future networks to its greatest extent.

In the trivial case when BSs are as dense as nodes and they have few antennas, the ISH protocol achieves capacity scaling along the whole bandwidth scaling range. However, if either the BSs cannot be deployed as densely as the nodes, or they perform massive MIMO, ISH has a much smaller critical bandwidth scaling than IMH. For small bandwidth scaling, ISH still achieves the same scaling as IMH, but for large bandwidth scaling ISH is suboptimal and short-range IMH is the only strategy that achieves capacity scaling.

We have proven that the traffic flow model can drastically alter capacity scaling results in wireless networks with infrastructure, and even determine which protocols achieve capacity scaling and which do not. We have developed a model for cellular wireless networks, CDFM, that is more accurate than the traditional ad-hoc-network-with-infrastructure-support ADFM model in the literature. This means that the results on some scaling analyses for infrastructure networks, based on ADFM instead of CDFM, may not be applicable to the engineering of future 5G networks. In the particular case of a dense network with small bandwidth scaling, where not only IMH but even ISH achieves capacity scaling, there is no possibility to use hierarchical cooperation to increase rate. This is surprising, because it contradicts previous results for ad-hoc networks: in dense ad-hoc networks hierarchical cooperation has been shown to achieve linear scaling $\Theta(n)$ and improve rate of direct transmissions significantly. To understand this, we constructed a theoretical IHC protocol, showing that the reason for this difference is due to the highly different traffic models of ad-hoc and cellular transmissions. On the one hand, in ad-hoc networks direct transmission flows may be directed to any random point in the network area, even across the entire network area, whereas in cellular networks direct transmissions are always directed to the closest BS (a much shorter maximum transmission distance). This makes direct transmission inefficient in ad-hoc networks, but not in cellular networks with small bandwidth scaling. On the other hand, the IHC protocol experiences bottlenecks that do not exist in the ad-hoc HC scenario. Suming up, when applying CDFM in cellular networks instead of ADFM, due to their different traffic spatial distributions, direct transmission improves and hierarchical cooperation worsens, altering scaling results fundamentally.

The IRH architecture represents a trade-off between scaling and practical issues relying on wireless-backhauled RNs. First, it allows to free mobile user nodes from the burden of multi-hop commmunications, which are based on auxiliar infrastructure devices. Second, the proximity between user nodes and RNs allows to overcome the limitations of ISH. The IRH protocol can achieve full capacity scaling if RNs are as dense as user nodes. And, since RNs are denser than wired BSs, IRH obtains a higher critical bandwidth scaling than ISH and outperforms that protocol for high bandwidth. Unfortunately, as RNs will not have as many antennas as BSs in the close future, the IRH DL has low-bandwidth due to the limitation in degrees of freedom.

The content in this chapter is an extended version of a paper in the IEEE International Symposium on Information Theory (ISIT) [11] and its extension to a journal paper, which is currently in preparation for submission [2].

Appendix 6.A Point-to-Point Link Scaling

6.A.1 AWGN

$\mathrm{I}(X;Y)$ is simply

$$W_\mathrm{u} \log(1 + \frac{P_{r_\mathrm{u}}}{W_\mathrm{u} N_\mathrm{I}}) \tag{6.35}$$

There are no MIMO antennas in this model, i.e. $\ell_t = \ell_r = 1$ and, regardless of whether $N_\mathrm{I} \sim N_0$ or $N_\mathrm{I} \sim \frac{\sum_{i \in \mathcal{I}} P_\mathrm{I}}{W}$, W^* indicates whether $\frac{P_{r_\mathrm{u}}}{W_\mathrm{u} N_\mathrm{I}}$ scales to 0 or ∞. In the first case expression (6.35) becomes $\frac{P_{r_\mathrm{u}}}{N_\mathrm{I}}$ and in the second case it becomes W_u multiplied by a constant, thus satisfying scaling (6.4).

6.A.2 CFC

$\mathrm{I}(X;Y|H) =$

$$\mathrm{E}_\mathbf{H}\left[\frac{W_\mathrm{u}}{B_\mathrm{c} T_\mathrm{c}} \max_{\mathbf{Q}_x} \log\det\left(\mathbf{I} + \mathbf{H}^H \mathbf{Q}_x \mathbf{H} (W_\mathrm{u} N_\mathrm{I} \mathbf{I})^{-1}\right)\right] \tag{6.36}$$

where $\frac{W_\mathrm{u}}{B_\mathrm{c} T_\mathrm{c}}$ is the number of i.i.d. channel sub-bands per second and $\mathbf{Q}_x$ is the cross-correlation matrix of X. Using Hadamard's equality for positive definite matrices we can upper bound it as

$$\mathrm{I}(X;Y|H) \leq \mathrm{E}_{\boldsymbol{\lambda}}\left[\max_{\mathbf{Q}_x} \frac{W_\mathrm{u}}{B_\mathrm{c} T_\mathrm{c}} \sum_{h=1}^{\min(\ell_\mathrm{t},\ell_\mathrm{r})} \log\left(1 + \frac{P}{W_\mathrm{u} N_\mathrm{I}} \lambda_h \mathbf{v}_h^H \mathbf{Q}_x \mathbf{v}_h\right)\right] \tag{6.37}$$

where $\boldsymbol{\lambda}$ contains the normalized eigenvalues λ_h of the squared channel realization $\mathbf{H}^H\mathbf{H}$ and $\mathbf{v}_h$ are the corresponding eigenvectors. The projection of signal covariance on the eigenvalues $\mathbf{v}_h^H \mathbf{Q}_x \mathbf{v}_h$ converts the maximization problem in a power allocation problem. By upper bounding all eigenvalues with the maximum, the optimum corresponds to the equal allocation across all dimensions[2]:

$$\leq \mathrm{E}_{\lambda_\mathrm{max}}\left[\frac{W_\mathrm{u}}{B_\mathrm{c} T_\mathrm{c}} \min(\ell_\mathrm{t},\ell_\mathrm{r}) \log\left(1 + \frac{P}{W_\mathrm{u} N_\mathrm{I} \min(\ell_\mathrm{t},\ell_\mathrm{r})} \lambda_\mathrm{max}\right)\right] \tag{6.38}$$

On the other hand, by taking the maximization out of the average by applying Jensen's inequality, we obtain an ergodic rate with a fixed encoding covariance that is maximized by equal power allocation. Applying Hadamard's equality and lower bounding with the

[2] $\mathbf{Q}_x = (\mathbf{V}\mathbf{V}^H)^{-1}$ for square matrices, or the rectangular equivalent to projected energy only in the subspace of non-zero eigenvalues

minimum non-zero eigenvalue:

$$\begin{aligned}\mathrm{I}\left(X;Y|H\right) &\geq \max_{\mathbf{Q}_x} \mathrm{E}_{\mathbf{H}}\left[\frac{W_\mathrm{u}}{B_\mathrm{c}T_\mathrm{c}}\log\det\left(\mathbf{I}+\mathbf{H}^H\mathbf{Q}_x\mathbf{H}(W_\mathrm{u}N_\mathrm{I}\mathbf{I})^{-1}\right)\right]\\ &= \mathrm{E}_{\boldsymbol{\lambda}}\left[\frac{W_\mathrm{u}}{B_\mathrm{c}T_\mathrm{c}}\sum_{h=1}^{\min(\ell_\mathrm{t},\ell_\mathrm{r})}\log\left(1+\frac{P}{W_\mathrm{u}N_\mathrm{I}\min(\ell_\mathrm{t},\ell_\mathrm{r})}\lambda_h\right)\right]\\ &\geq \mathrm{E}_{\lambda_{\min}}\left[\frac{W_\mathrm{u}}{B_\mathrm{c}T_\mathrm{c}}\min(\ell_\mathrm{t},\ell_\mathrm{r})\log\left(1+\frac{P}{W_\mathrm{u}N_\mathrm{I}\min(\ell_\mathrm{t},\ell_\mathrm{r})}\lambda_{\min}\right)\right]\end{aligned} \tag{6.39}$$

Again, W^* indicates whether $\frac{P_{r_\mathrm{u}}}{W_\mathrm{u}N_\mathrm{I}}$ scales to 0 or ∞. In the first case both upper and lower bounds become $\frac{P_{r_\mathrm{u}}}{N_\mathrm{I}}$ and, in the second case, $W_\mathrm{u}\min(\ell_\mathrm{t},\ell_\mathrm{r})\times\mathrm{const}$, thus satisfying scaling (6.4).

6.A.3 NFC

W^* is always at most the *critical bandwidth* defined in [76] and generalized in Sec.7.2.1. When $W_\mathrm{u} > W^*$, the transmitted signal has to concentrate its transmitted power in the critical bandwidth as in [76] or use peaky signaling as in [78], in both cases achieving $\frac{P_{r_\mathrm{u}}}{N_\mathrm{I}}$. When $W_\mathrm{u} < W^*$, the limit of N_I is dominated by interference and does not grow with W_u. In that case the SINR converges to a constant and the mutual information $\mathrm{I}\left(X;Y\right)$ scales approximately as $W_\mathrm{u}\min(\ell_\mathrm{t},\ell_\mathrm{r})$ multiplied by a constant logarithm.

Appendix 6.B Capacity Scaling Analysis

6.B.1 Proof of Theorem 6.8

We upper bound the rate from all BS to all nodes (in the DL) or from all nodes to all BS (in the UL). We focus on the DL case first, obtain the total sum rate $T_\mathrm{DL}(n) = nR_\mathrm{DL}(n)$, and then divide it by n. The MIMO channel from the BS to the nodes can be upper bounded by the capacity of a multi-antenna single-transmitter single-receiver system, as if all BS and nodes cooperated to transmit and receive respectively. The total transmission power is mP_BS and MIMO dimensions are $N_\mathrm{t} = m\ell \leq n = N_\mathrm{r}$. With independent block fading realizations that change every coherence interval T_c, independent codewords are transmitted in each block with bandwidth W and duration T_c. This means that a discrete Fourier transform with $K = \lceil W/T_\mathrm{c}\rceil$ coefficients can represent any valid codeword in one transmission antenna, and therefore the whole MIMO codeword can be represented by a $N_\mathrm{t}K$-dimension vector, $\mathbf{x} = (\mathbf{x}^{(1)}[0]\dots\mathbf{x}^{(N_\mathrm{t})}[0]\dots\mathbf{x}^{(N_\mathrm{t})}[K-1])^T$ satisfying $|\mathbf{x}|^2 \leq mP_\mathrm{BS}$. We represent macroscopic channel gains between each pair of antenna locations (b, u) as $\sqrt{g_{n_\mathrm{t},n_\mathrm{r}}}$, in the diagonal matrix $\mathbf{D}_g$, and $\mathbf{Q}_x$ is the normalized covariance matrix of the distribution of $\mathbf{x}$.

Expression (6.40) shows the mutual information resulting from this model:

$$\begin{aligned}
T_{\mathrm{DL}}(n) &\leq \frac{1}{T_\mathrm{c}} \max_{\mathbf{Q}_x} \mathrm{E}_{\mathbf{H},\mathbf{x}} \left[\log\det\left(\mathbf{I}_{nN_\mathrm{r}K} + \frac{P_{\mathrm{BS}}}{KN_\mathrm{t}N_0} \mathbf{H}^H [\mathbf{I}_K \bullet \mathbf{D}_g]^H \mathbf{Q}_x [\mathbf{I}_K \bullet \mathbf{D}_g] \mathbf{H} \right) \right] \\
&\leq \frac{1}{T_\mathrm{c}} \max_{\mathbf{Q}_x} \sum_{k=1}^{K} \mathrm{E}_{\mathbf{H}[k],\mathbf{x}[k]} \left[\log\det\left(\mathbf{I}_{nN_\mathrm{r}} + \frac{mP_{\mathrm{BS}}}{KN_\mathrm{t}N_0} \mathbf{H}[k]^H \mathbf{D}_g^H \mathbf{Q}_x \mathbf{D}_g \mathbf{H}[k] \right) \right] \\
&\leq W \sum_{u=1}^{n} \mathrm{E}_{\mathbf{H}[k],\mathbf{x}[k]} \left[\log\left(1 + \frac{mP_{\mathrm{BS}}}{KN_\mathrm{t}N_0} \max(\lambda_{\mathbf{H}^H\mathbf{H}}) \max_b g_{b.u} \right) \right]
\end{aligned} \tag{6.40}$$

where the first inequality is the capacity of the MIMO channel composed by all BSs and all nodes. Note that $\mathbf{H}$ is composed of the single-coefficient random fading at each discrete frequency and between each antenna pair, whereas the Kroneker product $(\mathbf{D} \bullet \mathbf{I}_K)$ is the macroscopic gain between each pair of antennas repeated K times for all frequencies. The second inequality uses the generalized Hadamard inequality to separate the encoding contribution of each carrier, but note that spectral diversity is maintained by preserving optimization in the domain of $\mathbf{Q}_x$. Finally, the third identity consists on upper bounding all path gains from each BS b to each user u by the largest $g_{b,u} \leq \max_b g_{b,u}$ (i.e. assuming that all BS are as close to the user as the closest BS), and upper bounding all the eigenvalues of the matrix $\mathbf{H}^H\mathbf{H}$ by the largest one.

If we merely assumed SNR → 0 to replace the logarithm with the identity, at this point we would obtain an upper bound that would be tight in the power-limited regime, but loose in the interference-dominated regime.

To build a tighter bound, we have to divide the set of users $\mathcal{N}$ in two groups: $\mathcal{V}$ contains the users at a distance less or equal than $r_\mathrm{v} \leq \Theta((K\ell)^{-\frac{1}{\alpha}})$ from the closest BS, resulting in a large SNR, and the complementary set $\mathcal{V}^c = \mathcal{N} \setminus \mathcal{V}$ contains those that are farther away resulting in low SNR regime. The information contributed by the first group is limited by degrees of freedom, whereas the information contributed by the second is power limited. There will be three regimes, represented in Fig. 6.7. Regimes A and C are simple to characterize, whereas regime B requires some discussion.

Regime A

Note that cell radius is $r_{\mathrm{cell}} = \Theta(n^{\frac{(\beta-\nu)}{2}})$. Therefore if $\psi + \gamma < \frac{\alpha}{2}(\beta - \nu)$, $\mathcal{V}^c$ is asymptotically empty, and the contributions received at all node locations are degrees-of-freedom-limited. Then all the terms in the sum for each user in (6.40) scale with $\Theta(Wm\ell)$.

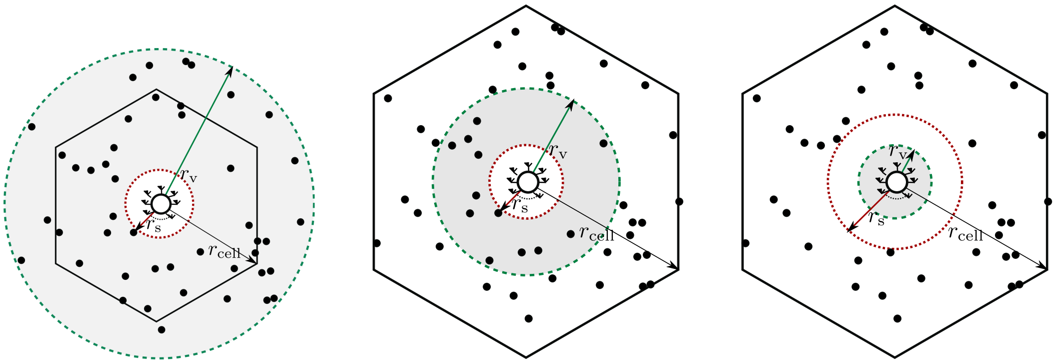
Regime C

Note that, with high probability, there are no UEs less than distance $r_{\mathrm{s}} \leq \Theta(n^{\frac{1-\nu}{2}})$ away from the closest BS. Therefore, if $\psi + \gamma > \frac{\alpha}{2}(1-\nu)$, $\mathcal{V}$ is empty and the contributions received at all node locations are power-limited. Since the best case is a transmit distance of $\Theta(n^{\frac{1-\nu}{2}})$, we can upper bound the scaling of all the terms in expression (6.40) with $\Theta(m r_{\mathrm{s}}^{\alpha})$.

Regime B

For $\frac{\alpha}{2}(\beta-\nu) < \psi+\gamma < \frac{\alpha}{2}(1-\nu)$ both regions are populated. Therefore the scaling of our upper bound is limited by the power transfer between the BSs and the set of node locations that are far from those BSs, or by the degrees of freedom between the BSs and the nodes close to them, whichever dominates. Sets are occupied by a fraction $|\mathcal{V}| = \frac{n^{\frac{\psi+\gamma}{\alpha}} - n^{\frac{1-\nu}{2}}}{n^{\frac{\beta-\nu}{2}}} = \Theta(0)$ and $|\mathcal{V}^c| = \frac{n^{\frac{\beta-\nu}{2}} - n^{\frac{\psi+\gamma}{\alpha}}}{n^{\frac{\beta-\nu}{2}}} = \Theta(1)$ of the nodes. Therefore, to analyze the scaling of this regime it suffices to assume that $\mathcal{V}$ is asymptotically empty as well. Then we upper bound the users in sum (6.40) that belong to the set $\mathcal{V}^c$ by considering the best-case scaling of power: at the smallest transmission distance *within this set* r_{v}. The remaining terms in sum (6.40) scale with the degrees of freedom. The final scaling of the overall sum is $\Theta(m r_{\mathrm{v}}^{\alpha})$.

Examining expression (6.40) in the three regimes leads to



(a) Regime A: All nodes belong to the high-SNR region. Capacity is degrees-of-freedom limited.

(b) Regime B: The high-SNR region contains a progressively vanishing, nonzero number of nodes. Capacity is power limited at a distance depending on r_{v}

(c) Regime C: The high-SNR region does not contain any node at all. Capacity is power limited at a distance depending on the minimum BS-node distance r_{s}

Figure 6.7: The three regimes of capacity scaling according to the frontier region between high SNR and low SNR.

$$T_{\mathrm{DL}}(n) = \begin{cases} \Theta\left(n^{\beta+\psi+\gamma}\right) & \psi+\gamma < \frac{\alpha}{2}(\beta-\nu) \\ \Theta\left(n^{\beta+\psi+\gamma}\right) & \frac{\alpha}{2}(\beta-\nu) < \psi+\gamma < \frac{\alpha}{2}(1-\nu) \\ \Theta\left(n^{\beta+\frac{\alpha}{2}(1-\nu)}\right) & \psi+\gamma > \frac{\alpha}{2}(1-\nu) \end{cases} \tag{6.41}$$

Taking $R_{\mathrm{DL}}(n) = \Theta(\frac{T_{\mathrm{DL}}(n)}{n})$ completes the proof for the DL case. In the analogous analysis for the UL case the total available power mP_{BS} is replaced with nP, the number of transmission antennas $m\ell$ with n and vice-versa for the reception antennas. The power gain extends the region where the degrees of freedom can be exploited, $r_{\mathrm{v}} \leq \Theta((\frac{m}{n}W\ell)^{-\frac{1}{\alpha}})$.

$$T_{\mathrm{UL}}(n) = \begin{cases} \Theta\left(n^{\beta+\gamma+\psi}\right) & \psi+\gamma+\beta-1 < \frac{\alpha}{2}(\beta-\nu) \\ \Theta\left(n^{\beta+\gamma+\psi}\right) & \frac{\alpha}{2}(\beta-\nu) < \psi+\gamma+\beta-1 < \frac{\alpha}{2}(1-\nu) \\ \Theta\left(n^{1+\frac{\alpha}{2}(1-\nu)}\right) & \psi+\gamma+\beta-1 > \frac{\alpha}{2}(1-\nu) \end{cases} \tag{6.42}$$

Finally, we get that these bounds are satisfied with probability $\lim_{n\to\infty} P = 1$ because there is zero probability that either a disc with radius $\Theta(n^{(\frac{\alpha}{2}(1-\nu))})$ around the BSs is non-empty or there are cell transmissions with more than $\Theta\left(n^{\gamma+\psi}\right)$ degrees of freedom.

6.B.2 Proof of Theorem 6.9

We first present the proof for the DL, which relies on lemmas 6.20 and 6.22. The first lemma analyzes the rate of each node through the first hop of the multi-hop route that originates from the BS.

Lemma 6.20. *In an IMH DL, the rate in the first link of each route from the BS to its neighboring nodes scales as*

$$R_{\mathrm{BS}\to u}(n) = \Theta(n^{\beta-1+\min(\gamma+\psi,\frac{\alpha}{2}(1-\nu))}). \tag{6.43}$$

Proof. The proof relies on the following two ideas:

1. The exponent of N_{I} for the IMH DL is

$$N_{\mathrm{I,IMH-DL}} \sim \Theta\left(n^{\max\left(\frac{\alpha}{2}(1-\nu)-\psi-\gamma,0\right)}\right) \tag{6.44}$$

2. This link becomes degrees-of-freedom limited when N_{I} is dominated by interference, and power limited when it is dominated by noise. In other words, as $n \to \infty$,

when N_{I} is dominated by interference the probability of $W > W^*$ in (6.4) tends to zero for all links, and when it is dominated by noise the probability of $W > W^*$ in (6.4) for all links tends to one.

To demonstrate the first point, we consider for each receiver u the interferer set $\mathcal{I}_{\mathrm{IMH-DL}}$ containing all nodes and BSs that transmit at the same time as the transmitter to u. We get

$$N_{\mathrm{I,IMH-DL}} = \frac{1}{W\ell} \sum_{i \in \mathcal{I}_{\mathrm{IMH}}} r_{i,u}^{-\alpha} P_i + N_0. \tag{6.45}$$

where P_i scales as a constant that is either P_{BS} or P, and only m out of every n subcells have the power of a BS. Note that we can find a lower bound of $r_{i,u}$, the distance between user u and transmitter i, as the distance between i and the border of the subcell containing u. In the hexagonal tessellation of the plane there are $6k$ subcells that form a ring at exactly $2k-1$ subcell radii from the subcell of u. The network is finite and a maximum k exists, but we can get rid of border effects by extending the sum of these interfering rings by letting $k \to \infty$. Also, only 1 out of every 7 neighbor subcells transmits at the same time due to a constant time-division, and we define the equivalent constant power $\overline{P} = \left(\frac{m}{n} P_{\mathrm{BS}} + \frac{n-m}{n} P\right)$

$$\begin{aligned}
\sum_{i \in \mathcal{I}_{\mathrm{IMH}}} P_i r_{i,u}^{-\alpha} &\leq \sum_{k=1}^{\infty} \sum_{i \in \mathcal{I}_k} P_i \left(r_{\mathrm{cell}}(2k-1)\right)^{-\alpha} \\
&\leq \left(\frac{m}{n} P_{\mathrm{BS}} + \frac{n-m}{n} P\right) r_{\mathrm{cell}}^{-\alpha} \sum_{k=1}^{\infty} |\mathcal{I}_k| (2k-1)^{-\alpha} \\
&\leq \overline{P} r_{\mathrm{cell}}^{-\alpha} \sum_{k=1}^{\infty} \frac{1}{7} (6k)(2k)^{-\alpha} \\
&= \frac{6}{7} \overline{P} (2r_{\mathrm{cell}})^{-\alpha} \sum_{k=1}^{\infty} k^{1-\alpha} \\
&\leq \frac{6}{7} \overline{P} (2r_{\mathrm{cell}})^{-\alpha} \zeta(\alpha - 1)
\end{aligned} \tag{6.46}$$

where $\zeta(\alpha-1)$ is the Riemann Zeta function evaluated in $\alpha - 1$, which is just some constant for any fixed $\alpha > 2$. This shows that interference power scales as subcell radius $n^{\frac{\alpha}{2}(1-\nu)}$, where the noise PSD is constant, so (6.45) scales as (6.44).

To prove the second point we consider that the BS-node feasible rate can be analyzed using the definition of critical bandwidth W_{crit} in [76]. We formulate the critical bandwidth for each user u as a function of its distance to the transmitting BS, r_{u}, and the transmission power allocated to the user $P_{\mathrm{u}} = \frac{P_{\mathrm{BS}}}{\ell}$.

$$W_{\mathrm{crit}}(r_{\mathrm{u}}) \propto \frac{P_{\mathrm{BS}}}{\ell N_I} (r_{\mathrm{u}})^{-\alpha} \tag{6.47}$$

Instead of comparing the bandwidth limitation $W_{\text{crit}}(r_{\text{u}})$ with the actual bandwidth W_{u} for one user at a time, we compute the *critical distance* r^* from the BS that defines the border between the region where users cannot see overspread transmissions and the region where they observe overspread transmissions with the allocated power and bandwidth.

Definition 6.21. The *critical distance* is

$$r^* : W_{\text{u}} = W_{\text{crit}}(r^*) \propto \frac{P_{\text{u}}}{\frac{W_{\text{u}}}{W} N_0}(r^*)^{-\alpha} \tag{6.48}$$

The critical distance of the first IMH hop scales as

$$r^*_{\text{IMH}} \sim \Theta(n^{\frac{(-\psi-\gamma)}{\alpha}}). \tag{6.49}$$

BS transmitters divide power across all available antennas and allocated bandwidth (the minimum of available and system-wide critical badwidths). We get

$$SINR_{\text{IMH}} \geq \frac{r_{\text{cell}}^{-\alpha} P_{\text{BS}}}{\ell W N_{I,IMH}} \tag{6.50}$$

The users that receive BS transmissions in the first hop are located in a circle with radius $1.5r_{\text{subcell}}$, which we divide in two areas: an inner circle containing nodes closer than the critical distance and an outer ring containing the nodes beyond the critical distance.

The fraction of nodes inside the inner circle is

$$f_{\text{IMH}} = \min\left(\frac{2\pi(r^*_{\text{IMH}})^2}{(1.5)^2 A_0 n^{\nu-1}}, 1\right) \propto n^{\frac{-2(\gamma+\psi)}{\alpha}+(1-\nu)}, \tag{6.51}$$

and this converges to one as $n \to \infty$ when $\frac{-2(\gamma+\psi)}{\alpha} + (1-\nu) > 0$ (interference-dominated case), and to zero otherwise (noise-dominated case).

The feasible rates in each regime are given by Lemma 6.1: $\Theta(W_{\text{u}}) = W$ in the degrees-of-freedom-limited case and $\Theta(\frac{P_{r_{\text{u}}}}{N_{\text{I}}}) \propto n^{\frac{\alpha}{2}(1-\nu)-\gamma}$ in the power-limited case. This gives the rate between the BS and each of its ℓ neighbors nearby, since there are a total of $\frac{n}{m}$ routes per the cell, and the rate of each first-hop link is time-shared by $\frac{n}{m\ell}$ routes. Putting everything together we get $R_{\text{BS}\to u}(n) = \frac{m\ell}{n}\min(W, \frac{P_{\text{u}} n^{-\frac{\alpha}{2}}}{N_{\text{u}} N_0})$ which may be rewritten as lemma 6.20. □

Lemma 6.22. *In the IMH DL, the rate in the second and subsequent links of each route from the BS to its neighboring nodes scales as*

$$R_{\text{BS}\to u}(n) = \Theta(n^{\beta-1+\gamma+\min(\psi,\frac{\alpha}{2}(1-\nu))}) \tag{6.52}$$

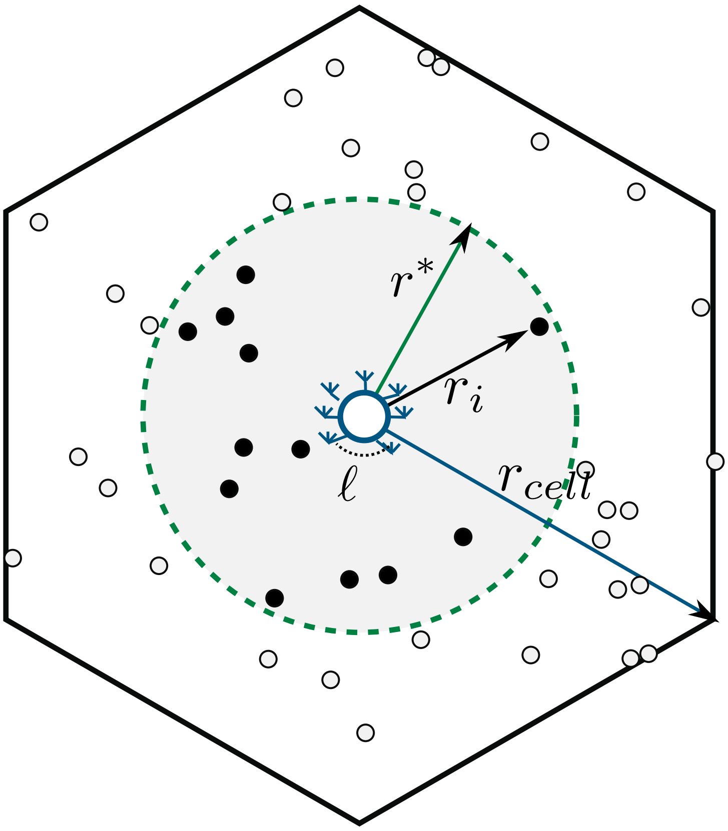


FIGURE 6.8: The two areas of coverage of the network.

Proof. The two points of the proof of lemma 6.20 can be applied. The first point can be applied directly, and to apply the second it suffices to change the allocated power to $n^{\frac{\alpha}{2}(1-\nu)}$. □

By combining the analysis of different hops we can distinguish three regimes:

- $\psi + \gamma < \frac{\alpha}{2}(1 - \nu)$, where all links are degrees-of-freedom-limited.
- $\psi < \frac{\alpha}{2}(1 - \nu) < \psi + \gamma$, where first-hop links are power-limited, and the following links are degrees-of-freedom limited. The percentage of nodes that experience $W_\mathrm{u} < W_\mathrm{crit}$ in the first hop tends to zero and, when the nodes transmit, overspreading affects the allocation of the ℓW antennas and the bandwidth resources at the BS, but not the usage of bandwidth W.
- $\psi > \frac{\alpha}{2}(1 - \nu)$, where all links are power-limited. The percentage of nodes that see $W_\mathrm{u} < W_\mathrm{crit}$ in the second and the following hops converges to zero and overspreading affects all bandwidth usages.

By comparing the rates in all regimes, it is shown that the bottleneck always takes place at the first hop. Combining them we prove Theorem 6.9 for the DL. The proof for the

UL follows the same lines, but, in that case, the bottleneck is in the last hop of each route, and multiple users transmitting towards the BS provide a small power gain.

$$R_{u\to\mathrm{BS}}(n) = \Theta(n^{\beta+\gamma-1+\min(\psi,\frac{\alpha}{2}(1-\nu))}) \tag{6.53}$$

Appendix 6.C Analysis of Other Protocols

6.C.1 Proof of Theorem 6.12

The proof for the DL relies again on the following two ideas

1. The exponent of N_I for ISH is

$$N_{\mathrm{I,ISH}} \sim \Theta\left(n^{\max\left(\frac{\alpha}{2}(\beta-\nu)-\psi-\gamma,0\right)}\right) \tag{6.54}$$

2. In ISH all links become asymptotically degrees-of-freedom limited when N_I is domninated by interference and power limited when it is dominated by noise. In other words, as $n\to\infty$, when N_I is dominated by interference the probability of overspreading tends to zero, and when it is dominated by noise the probability of overspreading tends to one.

To prove each point it is sufficient to modify the arguments in the proof of lemma 6.20, replacing the product of the subcell radius by 1.5 with the cell radius, replacing the number of BS antennas ℓ with the total number of nodes in the cell $\frac{n}{m}$, and letting the power and bandwidth allocation per user be $P_\mathrm{u} = \frac{m}{n}P_\mathrm{BS}$ and $W_\mathrm{u} = \frac{m\ell}{n}W$, respectively.

The feasible rates in each regime are given by Lemma 6.1: $\Theta(W_\mathrm{u}) = \frac{m\ell}{n}W$ in the degrees-of-freedom-limited case and $\Theta(\frac{P_{r_\mathrm{u}}}{N_\mathrm{I}}) \propto n^{\beta-1+\frac{\alpha}{2}(\beta-\nu)}$ in the power-limited case. Putting everything together Theorem 6.12 holds for the DL. An equivalent analysis can be applied to the UL case, considering that the power that is available at the transmitters increases by a factor $\frac{n}{m}$, yielding expression (6.14).

6.C.2 Proof of Theorem 6.14

The IRH protocol is similar to an IMH protocol where the BS acts as the infrastructure and the RNs act as the users (IMH-R), and it is similar to an ISH protocol where the RNs act as the infrastructure and the nodes act as the users (ISH-R). The aggregate network rate of the first scales as

$$\begin{aligned} T_{\text{IMH-R}}(k) &= \Theta\left(k^{\beta'+\min\left(\psi'+\gamma',(1-\nu')\frac{\alpha}{2}\right)}\right) \\ \beta' &= \frac{\beta}{\rho} \\ \gamma' &= \frac{\gamma}{\rho} \\ \psi' &= \frac{\psi}{\rho} \end{aligned} \tag{6.55}$$

where the new exponents appear because the number of BSs, the number of antennas and the bandwidth in the network do not change, and therefore when we replace the user index with $k < n$ we must adjust all the exponents.

The aggregate network rate of the second phase scales as

$$T_{\text{ISH-R}}(n) = \Theta\left(n^{\rho+\min\left(\psi,(\rho-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.56}$$

where the main difference with a normal ISH is the denser infrastructure ($\rho > \beta$) and the lack of multiple BS antennas (γ).

By defining two trivial cuts of the network, the scaling of the effective sum-rate seen by the end users cannot exceed the minimum of the scalings of each protocol rewritten as a function of n. By replacing $k = n^\rho$ in the first and taking the minimum:

$$T_{\text{IRH}}(n) = \Theta\left(\min \tau n^{\rho+\min\left(\psi,(\rho-\nu)\frac{\alpha}{2}\right)}, (1-\tau)n^{\beta+\min\left(\psi+\gamma,(\rho-\nu)\frac{\alpha}{2}\right)}\right) \tag{6.57}$$

The optimum time-sharing factor τ^* allows to balance the rate of the two protocols

$$\tau^* = \frac{T_{\text{IMH-R}}(n)}{T_{\text{ISH-R}}(n) + T_{\text{IRH}}(n)} \tag{6.58}$$

and converges asymptotically to

$$\lim_{n\to} \tau^* = \begin{cases} 0 & \rho + \min\left(\psi, (\rho-\nu)\frac{\alpha}{2}\right) < \beta + \min\left(\psi+\gamma, (\rho-\nu)\frac{\alpha}{2}\right) \\ c_\tau & \rho + \min\left(\psi, (\rho-\nu)\frac{\alpha}{2}\right) = \beta + \min\left(\psi+\gamma, (\rho-\nu)\frac{\alpha}{2}\right) \\ 1 & \rho + \min\left(\psi, (\rho-\nu)\frac{\alpha}{2}\right) < \beta + \min\left(\psi+\gamma, (\rho-\nu)\frac{\alpha}{2}\right) \end{cases} \tag{6.59}$$

where $c_\tau \in [0, 1]$ is a constant that is irrelevant for scaling because it only appears when the two protocols have the same exponent.

Putting everything together and dividing by the number of nodes, the feasible rate per node scales as Theorem 6.14 states.

Appendix 6.D Analysis of IHC

6.D.1 Case (6.21a) and (6.22a)

We consider $\psi \leq 1 + (1-\nu)\frac{\alpha}{2}$, and $\psi < \frac{\alpha}{2}(\beta - \nu) - \log_n M$.

Using $b_a = 1$ we compute the optimal M^* and the resulting feasible rate scaling. Choosing any arbitrary subcell size M, $b_a = 1$, the scalings in this regime and any arbitrary $b'_m \geq 0$ we have

$$T(n) = \frac{nM}{n^{1-\beta} M^{1-b'_m} + n^{1-\psi-\beta} + QM} \tag{6.60}$$

To find the optimal M^* we optimize it over any $M \in [1, n]$ because empty or negative subcells or subcells that are larger than the network do not make any sense.

$$M^* = \arg\max_M \frac{nM}{n^{1-\beta} M^{1-b'_m} + n^{1-\psi-\beta} + QM} \tag{6.61}$$

By calcluating the partial $T(n)$ derivative, we get

$$\frac{\partial T(n)}{\partial M} = \frac{n^{2-\psi-\beta} + b'_m n^{2-\beta} M^{1-b'_m}}{(n^{1-\beta} M^{1-b'_m} + n^{1-\psi-\beta} + QM)^2} > 0 \forall M \tag{6.62}$$

This means $M^* = n$, as we proposed for optimal subcell size. With this subcell size, the first phase is identical to a non-hierarchical BC protocol for the full cell, and the hierarchical BC in regime AD that makes use of that subordinate protocol may only scale like it or worse.

6.D.2 Case (6.21b) and (6.22b)

This case requires $\psi > 1 + (1-\nu)\frac{\alpha}{2}$ and $\psi > (1-\nu)\frac{\alpha}{2} - \log_n M$, where the first constraint implies the second. Using $b_a = 2 + (1-\nu)\frac{\alpha}{2} - \psi$ we get the following:

$$T(n) = \frac{nM}{n^{1-\beta} M^{1-b'_m} + M n^{1-\frac{\alpha}{2}(1-\nu)-\beta} + QM^{\psi-(1-\nu)\frac{\alpha}{2}}} \tag{6.63}$$

To optimize M^* we make the numerator of its derivative equal to zero:

$$0 = -n(-b'_m Q n^{1-\beta} M^{-b'_m-1} + M^{\psi-(1-\nu)\frac{\alpha}{2}-2}(\psi - 1 - (1-\nu)\frac{\alpha}{2})) \tag{6.64}$$

Using the hypothesis for this regime $\psi > 1 + (1-\nu)\frac{\alpha}{2}$,

$$M^* = \frac{b'_m}{Q[\psi - 1 - (1-\nu)\frac{\alpha}{2}]} n^{\frac{1-\beta}{\psi - 1 - (1-\nu)\frac{\alpha}{2} + b'_m}} \tag{6.65}$$

Therefore the rate scales as

$$T(n) = \Theta\left(n^{\beta + \min((1-\beta)\frac{b'_m}{\psi - 1 - (1-\nu)\frac{\alpha}{2} + b'_m}, (1-\nu)\frac{\alpha}{2})} \right) \tag{6.66}$$

Note that in this regime $\psi > 1 + (1-\nu)\frac{\alpha}{2}$, so that $\left[\frac{b'_m}{\psi-1-(1-\nu)\frac{\alpha}{2}+b'_m})\right]$ is a monotonous function of b'_m bounded between zero and one. Let us refer to this value simply as $\chi(b'_m) = (1-\beta)\left[\frac{b'_m}{\psi-1-(1-\nu)\frac{\alpha}{2}+b'_m})\right] \in [0,1]$. Therefore we have

$$\begin{aligned} \psi - \chi(b'_m) &\geq 1 + (1-\nu)\frac{\alpha}{2} - \chi(b'_m) \\ &\geq (1-\nu)\frac{\alpha}{2} \\ &\geq 0 \end{aligned} \tag{6.67}$$

and $\beta + \min(\chi(b'_m), (1-\nu)\frac{\alpha}{2}) \leq \beta + \min(\psi, (1-\nu)\frac{\alpha}{2})$, and thus the hierarchical protocol in this regime never outperforms IMH.

6.D.3 Case (6.21a) and (6.22c)

This case requires $\psi > \frac{3}{2}$ and $\psi < (1-\nu)\frac{\alpha}{2} - \log_n M$ by definition of case (6.21a) and (6.22c). The ad-hoc scaling is $b_a = \frac{1}{2} + (1-\nu)\frac{\alpha}{2}$.

$$T(n) = \frac{nM}{n^{1-\beta}M^{1-b'_m} + n^{1-\psi-\beta} + QM^{2-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}} \tag{6.68}$$

$$\frac{\partial T(n)}{\partial M} = \frac{n^{2-\psi-\beta} + b'_m n^{1-\beta}M^{1-b'_m} + Q\left[(1-\nu)\frac{\alpha}{2} - \frac{1}{2}\right]M^{2-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}}{\left(n^{1-\beta}M^{1-b'_m} + n^{1-\psi-\beta} + QM^{2-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}\right)^2} \tag{6.69}$$

Using the assumption $\frac{3}{2} < \psi < (\beta-\nu)\frac{\alpha}{2} - \log_n M < (1-\nu)\frac{\alpha}{2}$, the derivative is zero or strictly positive for all M, and therefore we are in the same situation as in lemma 6.15.

6.D.4 Case (6.21b) and (6.22c)

This mode requires $\psi > \frac{3}{2}$ and $\psi > (\beta - \nu)\frac{\alpha}{2} - \log_n M$.

$$T(n) = \frac{n}{n^{-\beta} M^{1-b'_m} + n^{1-(1-\nu)\frac{\alpha}{2}-\beta} + QM^{1-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}} \tag{6.70}$$

$$\frac{\partial T(n)}{\partial M} = \frac{b'_m n^{1-\beta} M^{-b'_m-1} + Q\left[(1-\nu)\frac{\alpha}{2} - \frac{1}{2}\right] M^{-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}}{\left(n^{-\beta} M^{1-b'_m} + n^{1-(1-\nu)\frac{\alpha}{2}} + QM^{1-\left[\frac{1}{2}+(1-\nu)\frac{\alpha}{2}\right]}\right)^2} \tag{6.71}$$

If $(1-\nu)\alpha \geq 1$, the derivative is always zero or strictly positive and we are again the situation in lemma 6.15. In the opposite case the derivative is very similar as that in subsection 6.4.1.4, resulting in a scaling

$$T(n) = \Theta\left(n^{\beta+\min((1-\beta)\left[\frac{b'_m}{\frac{1}{2}-(1-\nu)\frac{\alpha}{2}+b'_m})\right],(1-\nu)\frac{\alpha}{2})}\right) \tag{6.72}$$

which, with similar arguments as those in subsection 6.4.1.4, shows that the hierarchical protocol never outperforms IMH.

Chapter 7

Challenges of Cooperative Fifth Generation Cellular Networks

Contents

7.1 Introduction

The scaling result in Chapter 6 indicates that cooperation must be an integral part of 5G cellular design from its onset, but the analysis is too general and approximate to provide insights in the desirable details of implementation of such cooperation. The last chapter of this thesis introduces recent research, still in preparation for publication, on different aspects of the practical implementation of cooperative multi-hop in mmW 5G cellular networks.

mmW 5G networks will have short-range cells limited by transmission distances of 100-200 meters. This follows the trend of small-cell LTE-A systems. Such massive cell density will require cost prohibitive back-hauling in any frequency. For this reason, both small cell LTE-A and mmW cellular systems will have to rely on multi-hop relaying to improve performance as we anticipated in Chapter 6. Furthermore, efficient mmW transmissions will only be feasible if supported by high-dimensional antenna arrays capable of highly directive beam-forming [170], since otherwise even the closest nodes will not see each other.

First, in Section 7.2 we explore in detail the information-theoretic operational significance of a very large bandwidth. It is necessary to go beyond the results of lemmas 6.1, 6.2, 6.3 because, even though these are sufficient to characterize the scaling laws of wireless links and multi-user cells, they do not explain how to build or operate capacity-achieving practical communications. We first deal with the point-to-point case, showing that non-coherent fading channels with a large bandwidth can obtain, if bandwidth is below a critical threshold, rates similar to pilot-assisted or coherent schemes. When the bandwidth is above the critical threshold, we confirm that the trend in lemma 6.1 is not only true in a scaling sense, but exactly as well: Any signaling scheme with bandwidth exceeding the threshold achieves, regardless of peakyness, a linear-in-power maximal capacity that is proportional to the ratio $\frac{P}{N_0}$. Going one step further, we explore the critical parameters of multi-user channels (MAC and BC). We show that the critical bandwidth of these is **exactly** the union, in orthogonal bands, of the critical bandwidths of the users. Simultaneous transmissions with joint signal processing may be optimal in AWGN and coherent channels, but in non-coherent channels with very large bandwidth channel uncertainty affects more severely that type of signaling. In fact, orthogonal schemes are not only a low-complexity alternative in that case, but in fact the alternative that offers the highest rates. This strengthens lemmas 6.2 and 6.3.

Second, in Section 7.3 we discuss the literature about practical signaling schemes in the PHY layer that can be selected to implement a multi-hop scheme and discuss possible divergences with the state of the art we introduced in Section 2.3. We distinguish three types of strategies: the first consists in attempting to use traditional pilot-assisted coherent schemes, and detect and avoid transmissions above the critical bandwidths so that rate of the scheme does not fall apart due to overspreading. Unfortunately, pilot-assisted schemes do not necessarily achieve the theoretical bounds of Section 7.2. The second scheme we consider is the use of non-coherent signaling schemes. However, even though these are asymptotically optimal in infinite bandwidths, they do not work well in large finite bandwidths because of their low spectral efficiency. The more promising option is the combination of results on wideband and massive MIMO in the literature. In a coherent channel, coherent signal combination from massive antennas would produce a power gain that prevents SNR going to zero, as in the wideband SISO channel. Our main finding is that a similar gain is possible in non-coherent schemes, albeit with lower quantitative benefits, due to the diversity from multiple replicas of the received signal, even if not coherently combined.

Third, in Section 7.4 we study the requirements of MAC protocols for the 5G cellular framework and compare them with multi-hop implementations according to current cellular standards as discussed in Section 3.2. We propose two different, albeit affine, open problems: On the one hand, we discuss multi-hop scheduling in a MAC protocol with optimized allocation of bandwidth and power to simultaneous links in a mmWave cell. For that prupose, we have to assume that users can only attach themselves to one AP in the network infrastructure and RNs do not communicate with each other. This leads to a network tree topology k whose recursive structure is essential to solve the optimization problem in short time. On the other hand, we discuss multi-hop scheduling in a MAC with arbitrary mesh topology that permits multiple attachments of single devices to several APs and communications between RNs. This allows for greater routing and traffic balancing gains, but increases complexity considerably. To study the problem realistically, in this analysis we assume that devices cannot have multiple *simultaneous* active links, removing the need to optimize power and bandwidth allocation in our utility-cost functions.

Finally, Section 7.5 is a summary of this chapter.

7.2 Information Theoretic Framework

In this section we formulate detailed models for the point-to-point and multi-user channels whose scaling we considered in Section 6.2.1. These models provide the information-theoretic guidelines for the implementation of cooperative multi-hop communication schemes as introduced in Chapter 6.

7.2.1 Unified Capacity of Wideband Peaky and Non-Peaky Signaling

Peaky and non-peaky signaling schemes have been traditionally considered fundamentally different in non-coherent wideband fading channels because of their extremely different behaviors as bandwidth goes to infinity ($B \to \infty$). Peaky signals can achieve asymptotically the linear-in-power capacity of a wideband AWGN channel with the same SNR,

$$C^{\infty} = \lim_{B\to\infty} C(B) = \lim_{B\to\infty} BN_r\mathrm{SNR} = N_rP/N_0,\ \text{[nats/s]},$$

where P is the power, N_0 is the noise Power Spectral Density (PSD), N_r is the number of reception antennas, and $\mathrm{SNR} = P/(BN_0)$ is the SNR per degree of freedom at each reception antenna. On the other hand, non-peaky signals can only reach a peak rate at some finite *critical bandwidth*, and then the rate falls to zero when bandwidth grows to infinity above the critical value.

In a recent work submitted to IEEE Transactions on Information Theory [4], we show that this traditional distinction is in fact due to the limited attention paid so far to the product of the bandwidth by the fraction of time it is in use. We call this product *Bandwidth Occupancy*, and it measures average bandwidth usage over time.

It holds that both peaky and non-peaky signaling can approach the wideband capacity limit when there is plenty of bandwidth, but it is not immediately clear how the power-limited rate (1.2) in [76] (developed for SISO) is related to the polynomial near-power-limited rate (1.3) in [78] (developed for MIMO). Besides, there are no clear criteria to choose between these two seemingly distinct schemes.

In this work we unify the theoretical study of peaky and non-peaky signaling, showing that they are merely corner points of a more fundamental trade-off that affects all types of signals. The analyses in [76, 78] are simply two wildly different methods of representing the same physical reality.

Our analysis method is a generalization of those in [76] by adding MIMO and arbitrary levels of signal peakiness through a low duty-cycle transmission, where the peakiness parameter $\delta \in (0, 1]$ defines the fraction of time the transmitter is active. The analysis follows four steps, represented in Fig. 7.1.

1. Find a bell-shaped lower bound $R^{\mathrm{LB}}(B) \leq \mathrm{I}(X;Y)$;
2. Determine the unique maximum of $R^{\mathrm{LB}}(B)$, $R^{\mathrm{LB}*}(B^*)$;
3. Find a bell-shaped upper bound $R^{\mathrm{UB}}(B) \geq \mathrm{I}(X;Y)$;
4. Determine the two bandwidth values B^+ and B^- such that $B^- \leq B^{\mathrm{crit}} \leq B^+$ and $R^{\mathrm{UB}}(B^{\pm}) = R^{\mathrm{LB}*}(B^*)$.

The result of [76] is that capacity in a noncoherent fading channel only grows with bandwidth below a *critical bandwidth* B^{crit} which falls into the range $[B^-, B^+]$. A system operating with insufficient bandwidth $B{<}B^{\mathrm{crit}}$ is less efficient in converting available energy into data rate due to the convexity of the logarithm function w.r.t. the SNR, and the achievable rate grows with increasing bandwidth.

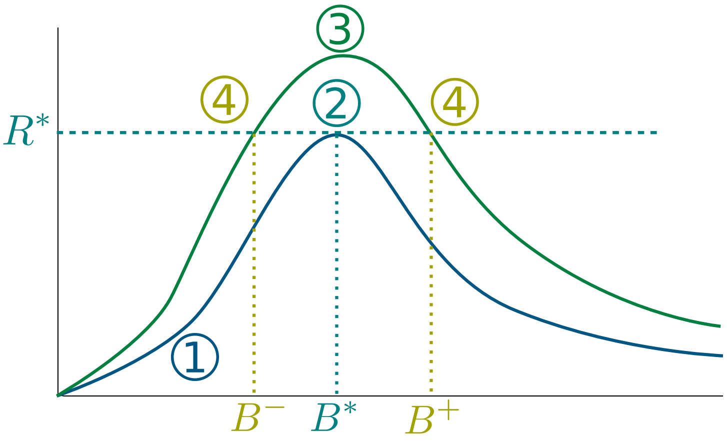


FIGURE 7.1: The four-step approach to determine the range of critical bandwidth occupancy.

Our contribution is a generalization of this argument to arbitrary levels of signal peakiness δ. We show that the capacity is a function of the quantity that we call *bandwidth occupancy* (δB), and we prove that the capacity $C(\delta B)$ experiences a critical value $(\delta B)^{\text{crit}}$. Moreover, the maximum rate, at the critical value $(\delta B)^{\text{crit}}$, is lower bounded by

$$R^{\text{LB}}((\delta B)^{\text{crit}}) = N_{\text{r}} \frac{P}{N_0} (1 - \Delta) \,, \tag{7.1}$$

with the same Δ for all levels of peakiness δ. Therefore, it is possible to approach C^{∞} with the same capacity gap at the same convergence speed with any signaling scheme within the family using a bandwidth $B \geq B^{\text{crit}}$ together with the peakiness parameter $\delta \simeq \frac{B^{\text{crit}}}{B}$ as represented in Fig. 7.2.

Using the relation between the main sublinear exponent α and the peakiness parameter δ=SNR$^{1-\alpha}$ as in [78], we show that Δ∼SNR$^{\alpha}$ at $(\delta B)^{\text{crit}}$. This is, the multiplicative capacity gap Δ in [76] and the sub-linear polynomial approximation SNR$^{\alpha}$ in [78] represent the same physical reality.

7.2.1.1 Peaky Signal Model

We consider a rich scattering, frequency selective, block fading, $N_{\text{t}} \times N_{\text{r}}$ MIMO wideband channel with an impulse response $h(t)^{(u,v)}$ between antennas (u, v). For compactness we assume that all channels experience a coherence time T_{c} and a delay spread D and the channel frequency response becomes uncorrelated for frequencies separated more than one coherence bandwidth $B_{\text{c}} = 1/D$. We focus only on the frequency signaling scheme

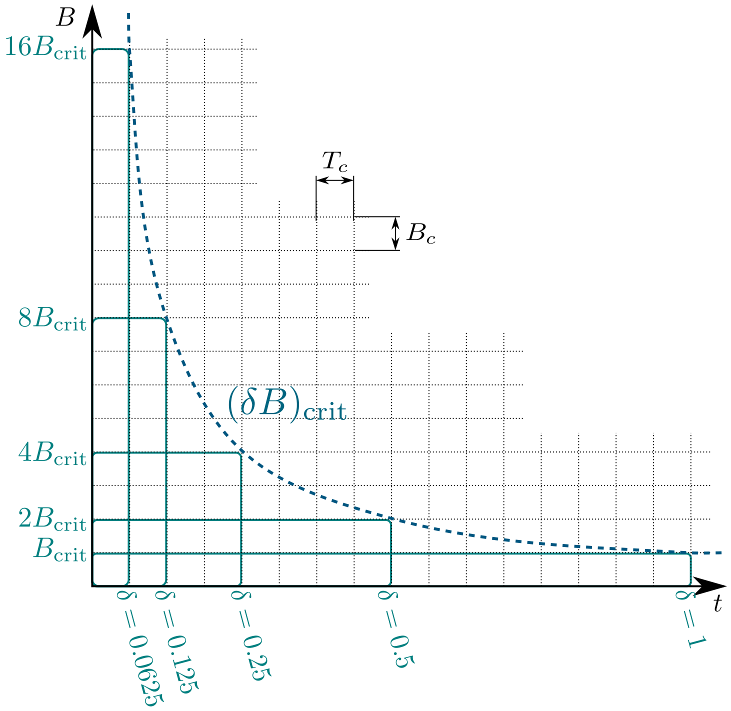


FIGURE 7.2: All transmission strategies that have the same product δB achieve the same polynomial approximation of C^{∞}.

since it is known that differences between frequency and time signaling only affect the scaling with bandwidth in its vanishing higher order terms [76].

Our model represents a signaling scheme where every T_{c}, the transmitted signal $x^{(u)}[n]$ with bandwidth $B/2$ carries $K{=}BT_{\text{c}}$ complex samples on antenna $u \in [0, N_{\text{t}}{-}1]$. Taking a K-point DFT, the transmitted codeword is uniquely defined by the $N_{\text{r}}K \times 1$ vector $\mathbf{x}$ that satisfies the average power

$$\frac{1}{KN_{\text{t}}}\text{E}\left[|\mathbf{x}|^2\right] \leq PT_{\text{c}}.$$

For $i{=}kN_{\text{t}}{+}u$, the i-th coefficient of $\mathbf{x}$, denoted as $x^{(i)}$, corresponds to the transmitted signal on antenna u with DFT index $k \in \{0, 1, \ldots, K{-}1\}$. For each pair of antennas (u, v), the discrete samples of the channel $h^{(u,v)}[n]$ have $M{=}BD$ i.i.d. coefficients, with $M/K{=}D/T_{\text{c}}{=}\frac{1}{B_{\text{c}}T_{\text{c}}}$. After applying a K-point DFT to each discrete channel sequence $h^{(u,v)}[n]$, we define a block-diagonal matrix $\mathbf{H}$ with K blocks of size $N_{\text{r}} \times N_{\text{t}}$ matrices,

$$\mathbf{H} = \left(\begin{array}{c|c|c} \mathbf{H}[0] & \ldots & \mathbf{0} \\ \hline \vdots & \ddots & \vdots \\ \hline \mathbf{0} & \ldots & \mathbf{H}[K-1] \end{array}\right), \tag{7.2}$$

where $\mathbf{H}[k]$ contains in its component (u, v) the k-th DFT coefficient of $h^{(u,v)}[n]$. Each channel only has M i.i.d. coefficients and any two blocks $\mathbf{H}[k]$ and $\mathbf{H}[k']$ are correlated if $|k - k'| < B_\mathrm{c}T_\mathrm{c}$ and independent otherwise. We also define the average gain of the n-th channel coefficient $g_n^{(u,v)} = \mathrm{E}\left[|h^{(u,v)}[n]|^2\right]$ satisfying $\sum_{n=0}^{M-1} g_n^{(u,v)} = 1$.

Assuming $D \ll T_\mathrm{c}$, there is no inter-symbol interference and the signal received on each fading realization, T_c, depends only on the state of the channel and the signal transmitted during the same realization. Taking a K-point DFT of the received signal we can represent the system model as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{7.3}$$

where $\mathbf{y}$ is a $N_\mathrm{r}K{\times}1$ vector whose i-th element $y^{(i)}$, $i{=}kN_\mathrm{r}{+}v$, corresponds to the signal received on antenna $v{\in}[0, N_\mathrm{r}{-}1]$ with DFT coefficient index k, where the $N_\mathrm{r}K \times 1$ noise vector $\mathbf{z}$ follows a Gaussian distribution $\mathcal{CN}(0, \mathbf{I}_{N_\mathrm{r}K}N_0T_\mathrm{c})$ (with PSD N_0).

Some authors, such as in [78], use a different system model with fewer frequency coefficients, where all bins experience independent fading, but in that model encoding block duration is shorter than T_c, and it has the requirement to perform encoding across multiple channel blocks. It is possible to prove that both models are equivalent using concepts of Single-Carrier-OFDM modulations.

The classic example of a peaky signal distribution is the on/off distribution. To make our signaling scheme peaky we only activate a fraction δ of the encoding symbols,

$$P_r(|\mathbf{x}|^2 = 0) = 1 - \delta. \tag{7.4}$$

This converts the system into the time-alternation of an arbitrarily distributed scheme for a fraction δ of the time, achieving a rate $R(\delta)$ with power gain $P' = \frac{P}{\delta}$, and an idle stage for a fraction $1 - \delta$ of the time. When $\frac{1-\delta}{\delta} > D/T_\mathrm{c}$ the idle stage serves also as a sort of “zero-padding prefix” that enforces our approximation that there is no ISI. For a random signal x drawn from a random variable $X \sim p(x)$, we will refer to its *kurtosis*

$$\kappa(X) = \frac{\mathrm{E}_X\left[|x|^4\right]}{\mathrm{E}_X\left[|x|^2\right]^2}, \tag{7.5}$$

to measure the *peakiness* of the random distribution. Note that when a signal $\mathbf{x}$ is zero a fraction $1{-}\delta$ of the time, its kurtosis can be written as a function of the kurtosis of the distribution of non-zero elements, $\kappa(\mathbf{x}) = \frac{\kappa(\mathbf{x}\neq 0)}{\delta}$, and therefore our measurements of peakiness using the on/off ratio δ and the kurtosis statistic κ are coherent with each other.

7.2.1.2 Lower Bound

Lemma 7.1. *The achievable rate in a wideband noncoherent fading channel is lower bounded by*

$$\begin{aligned} R^{\mathrm{LB}}(\delta B) =& \frac{PN_r}{N_0}\left[1-\frac{P(\kappa-2+N_t+N_r)}{2\delta BN_tN_0}\right] \\ & -\delta\frac{BN_tN_r}{B_\mathrm{c}T_\mathrm{c}}\log(1+\frac{P}{\delta BN_tN_0}B_\mathrm{c}T_\mathrm{c}), \end{aligned} \tag{7.6}$$

where κ is the kurtosis *of the channel.*

Proof. Generalization of [76, Eq. 40], with three key steps:

- Use $\frac{1}{T_\mathrm{c}}\mathrm{I}\left(X;Y\right)=\frac{1}{T_\mathrm{c}}\mathrm{I}\left(X,H;Y\right)-\frac{1}{T_\mathrm{c}}\mathrm{I}\left(H;Y|X\right)$;
- Lower bound $\frac{1}{T_\mathrm{c}}\mathrm{I}\left(X,H;Y\right)\geq\frac{1}{T_\mathrm{c}}\mathrm{I}\left(X;Y|H\right)$;
- Use $\log\det(\mathbf{I}+\mathbf{A}^H\mathbf{A})\geq\mathrm{tr}(\mathbf{A}^H\mathbf{A})-\mathrm{tr}((\mathbf{A}^H\mathbf{A})^2)/2$.

□

7.2.1.3 Maximum Rate

Lemma 7.2. *$R^{\mathrm{LB}}(\delta B)$ is maximized at $R^{\mathrm{LB}}((\delta B)^*)$ with*

$$(\delta B)^*\simeq\frac{P}{N_0N_t}\sqrt{\frac{B_\mathrm{c}T_\mathrm{c}}{\log B_\mathrm{c}T_\mathrm{c}}(\kappa-2+N_t+N_r)}, \tag{7.7}$$

and

$$R^{\mathrm{LB}}((\delta B)^*)\geq\frac{PN_r}{N_0}\left[1-\sqrt{\frac{\log B_\mathrm{c}T_\mathrm{c}}{B_\mathrm{c}T_\mathrm{c}}(\kappa-2+N_t+N_r)\log\pi}\right]. \tag{7.8}$$

Proof. Generalization of [76, Eq. 55 and 60]. Maximization of (7.6) with respect to the joint variable $(B\delta)$. □

Below the optimal bandwidth occupancy $(\delta B)^*$, the third term of (7.6) is smaller in absolute value than the second. Replacing the third by the second and substituting $\delta=\mathrm{SNR}^{1-\alpha}$ gives the following corollary on sufficient conditions.

Corollary 7.3. *If $\delta(B)B\leq(\delta B)^*$, the achievable rate is lower bounded by*

$$R^{\mathrm{LB}}(\delta B)\geq\frac{PN_r}{N_0}\left[1-\left(\frac{P}{BN_0}\right)^\alpha\frac{(\kappa-2+N_t+N_r)}{N_t}\right]. \tag{7.9}$$

7.2.1.4 Upper Bound

Lemma 7.4. *The achievable rate of flash signaling in a wideband noncoherent Rayleigh fading channel is upper bounded by*

$$\begin{aligned}R^{\mathrm{UB}}(\delta B) = \frac{PN_r}{N_0}\Bigg[1 - \frac{P}{2\delta B N_0} \\ - \frac{\delta B N_t N_0}{P B_{\mathrm{c}} T_{\mathrm{c}}} \mathrm{E}_H \left[\log(1+\frac{P}{\delta N_t B N_0} B_{\mathrm{c}} T_{\mathrm{c}} g_{\min} \psi)\right]\Bigg] + o(\frac{1}{W}),\end{aligned} \tag{7.10}$$

where $g_{\min} = \min_{m,u,v} |h^{(u,v)}[m]|^2$ *is the minimum non-zero square channel gain among all delays and antenna pairs, and* $\psi = \frac{\lambda_*}{K}$ *is the eigenvalue, normalized by* K, *of some "pilot signal" matrix* $\Xi\Xi^H$ *that provides the minimum* $\mathrm{E}_H\left[\log(1+\frac{P}{\delta W N_0} B_{\mathrm{c}} T_{\mathrm{c}} g_{\min} \lambda_m(\Xi\Xi^H)/K)\right]$ *for all eigenvalues indexed by* m.

Proof. Generalization of [76, Eq. 72]. Define matrix Ξ as a "pilot" signal derived from $\mathbf{x}$, define the normalized unit-mean random distribution of its K eigenvalues $\psi_{k,\ell} = \frac{\lambda\{\Xi^H\Xi\}_{k,\ell}}{K}$ and replace all k, ℓ components with the one that minimizes $\mathrm{I}(H; Y|\Xi)$. □

Remark 7.5. The minimization with regard to $\psi_{k,\ell}$ does not affect our analysis in the fourth and last step.

7.2.1.5 Critical Bandwidth Occupancy

Lemma 7.6. *In a wideband noncoherent Rayleigh fading channel, the maximum rate in* (7.8) *is achievable at a critical bandwidth occupancy* $(\delta B)_{\mathrm{crit}}$ *that lies in the range*

$$(\delta B)^- \leq (\delta B)_{\mathrm{crit}} \leq (\delta B)^+, \tag{7.11}$$

where

$$\begin{aligned}(\delta B)^- &= \frac{P}{N_0} \frac{1}{2\sqrt{(N_t + N_r)\log \pi}} \sqrt{\frac{B_{\mathrm{c}} T_{\mathrm{c}}}{\log B_{\mathrm{c}} T_{\mathrm{c}}}}, \\ (\delta B)^+ &= \frac{P}{N_0} 2\sqrt{\frac{(N_t + N_r)}{N_t^2}\log \pi} \sqrt{\frac{B_{\mathrm{c}} T_{\mathrm{c}}}{\log B_{\mathrm{c}} T_{\mathrm{c}}}}.\end{aligned} \tag{7.12}$$

Proof. Define the pair of solutions $(\delta B)^-$ and $(\delta B)^+$ such that

$$\frac{P}{(\delta B)^{\pm} N_0} = \sqrt{\Omega \frac{\log B_{\mathrm{c}} T_{\mathrm{c}}}{B_{\mathrm{c}} T_{\mathrm{c}}}} + o\left(\sqrt{\frac{\log B_{\mathrm{c}} T_{\mathrm{c}}}{B_{\mathrm{c}} T_{\mathrm{c}}}}\right), \tag{7.13}$$

and, as generalization of [76, Eq. 79 and 80], solve for Ω the equality $R^{\mathrm{UB}}(\delta B)^{\pm} = R^{\mathrm{LB}}(\delta B)^{*} + o(\frac{1}{B_{\mathrm{c}}T_{\mathrm{c}}})$. □

Above the critical bandwidth occupancy $(\delta B)_{\mathrm{crit}}$, the third term of (7.6) is greater in absolute value than the second. This means that the capacity is lower than expression (7.9), which leads to the following corollary on necessary conditions.

Corollary 7.7. *In case of Rayleigh fading (κ=2), if $\delta(B) = \mathrm{SNR}^{1-\alpha}$ and*

$$R(\delta(B)B) \geq \frac{PN_r}{N_0}\left[1 - \left(\frac{P}{BN_0}\right)^{\alpha}\frac{(N_t+N_r)}{N_t}\right], \tag{7.14}$$

then the bandwidth occupancy satisfies $\delta(B)B < (\delta B)^{+}$.

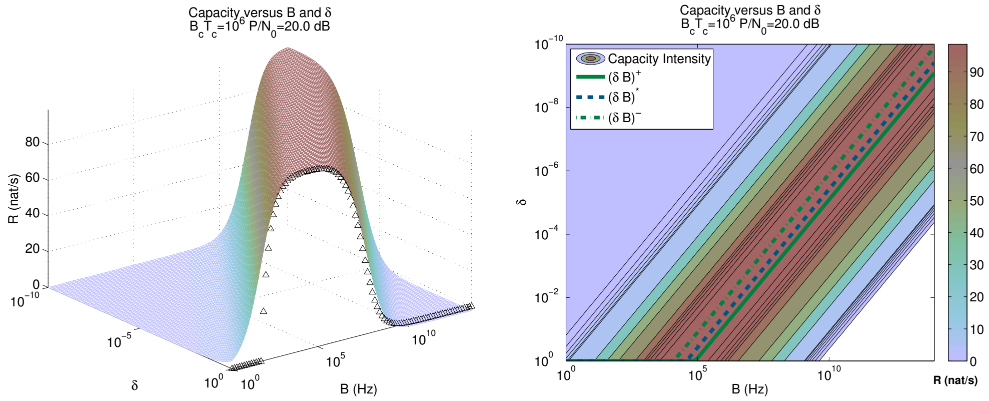
Note that both our capacity lower/upper bounds (7.6) and (7.10) derive from $\mathrm{I}\left(X;Y\right)$=$\mathrm{I}\left(X,H;Y\right)$–$\mathrm{I}\left(H;Y|X\right)$, which leads to the following capacity expression

$$\begin{aligned}&\frac{\delta}{T_{\mathrm{c}}}\left[\Theta(K)\log(1{+}\Theta(\frac{P/\delta}{N_0B})) - \Theta(M)\log(1{+}\Theta(\frac{P/\delta}{N_0B}\frac{K}{M}))\right]\\ &{=}\Theta\left(\delta B\right)\log(1{+}\Theta(\frac{P/\delta}{N_0B})) - \Theta(\frac{\delta B}{B_{\mathrm{c}}T_{\mathrm{c}}})\log(1{+}\Theta(\frac{PB_{\mathrm{c}}T_{\mathrm{c}}}{N_0\delta B})),\end{aligned}$$

where the equality is due to the substitution of K=BT_{c} and M=BD. The first term corresponds to the capacity in the wideband regime, and the second term represents the penalty from channel uncertainty. According to our derived channel model, during a period of coherence time T_{c}, for each spatial dimension we have K i.i.d. input symbols and M i.i.d. channel coefficients. The penalty term resembles a "channel estimation" setup where M unknown channel coefficients are inferred based on K training symbols, resulting in a "power gain" of $\frac{K}{M} = B_{\mathrm{c}}T_{\mathrm{c}}$. As bandwidth B grows, both the number of parallel channels and the number of independent channel coefficients grow linearly with B, but the growth ratio is T_{c} for the former and D for the latter. That is, the penalty term grows $B_{\mathrm{c}}T_{\mathrm{c}}$ times slower than the first term. Since there is also a "power gain" of $B_{\mathrm{c}}T_{\mathrm{c}}$ in "channel estimation", the penalty term "catches up" with the first by an additional factor $\log(B_{\mathrm{c}}T_{\mathrm{c}})$. This explains the origin of $\sqrt{B_{\mathrm{c}}T_{\mathrm{c}}}$ and $\sqrt{\frac{1}{\log B_{\mathrm{c}}T_{\mathrm{c}}}}$ in the critical bandwidth occupancy proved in Lemma 7.6.

In Fig. 7.3(a) we represent the upper bound on capacity as a field over the 2D plane (B,δ), and in the vertical cut for $\delta = 1$ we also represent the lower bound using triangular bullets to illustrate the relation with Fig. 7.1. On the B axis, we can see that for fixed values of δ the capacity as a function of bandwidth is bell-shaped, grows at small bandwidth, reaches a maximum and then decreases to zero. Fig. 7.3(b) provides a better perspective of the value of capacity upper bounds as a function of the bandwidth occupancy, where the optimal $(\delta B)^{*}$ that maximizes the capacity lower bound R^{LB}

and the range $[(\delta B)^-, (\delta B)^+]$ for the critical bandwidth occupancy $(\delta B)^{\text{crit}}$ are also plotted. For bandwidth occupancy close to $(\delta B)^{\text{crit}}$, capacity is nearly power-limited. For different level of peakiness δ, the peak values of capacity are the same but appear at different values of bandwidth B, and in fact all points with identical value δB have the same lower/upper bounds. Our analysis retakes the previous result for non-peaky signals by selecting $\delta = 1$, producing a finite critical bandwidth. It also captures the classical results for infinite-fourth-moment signals by making $\delta \to 0$, which takes the critical bandwidth occupancy point further into higher bandwidths following $\lim_{\delta\to 0} \frac{(B\delta)^{\text{crit}}}{\delta} = \infty$.



(a) Capacity upper bound for (δ, B) and low bound for δ=1.

(b) Contour plot w.r.t. (δ, B) and levels of bandwidth occupancy.

Figure 7.3: Capacity upper bound and critical bandwidth on plane (δ, B).

7.2.2 Optimal Bandwidth Allocation in Multiuser Wideband Systems

In a recent work in preparation to be submitted as a letter , we extend the notion of critical bandwidth analysis to multiuser MIMO channels. In particular, the MIMO Multiple Access Channel (MAC) modeling UL and the MIMO Broadcast Channel (BC) modeling the DL.

This analysis contains the following contributions:

i) We introduce a definition of **multi-user critical bandwidth** generalizing the idea of [76] to n-dimensional capacity regions.

ii) We analyze the critical bandwidth of the MAC channel and proof it is bounded by two scenarios:

ii-a) An upper bound of the MAC critical bandwidth is obtained by allocating transmissions of different users in orthogonal bands using Frequency Division Multiple Access (FDMA).

ii-b) A proof that orthogonal transmissions are necessary in this regime is obtained by showing that superposed transmissions with receiver Successive Interference Cancelling (SIC) -known to offer a larger capacity than FDMA with small bandwidth- have lower critical bandwidth and maximal sum-rate.

iii) We argue that the conclusions about orthogonality of user signals hold in the BC channel by defining similarities between MAC and BC channel analyses that permit to use the same upper bounds and obtain modified lower bounds using the classic result of MAC-BC duality with CSI.

7.2.2.1 MAC System Model

We consider a MAC channel with N MIMO transmitters with N_t transmission antennas each delivering information to a common destination with N_r reception antennas, using a total bandwidth B. Each user i reaches the destination with power P_i. We denote by $\{\mathcal{K}_i\}_{i=1}^N$ the set of subcarriers (DFT coefficients) allocated to each of the N users of a multi-user channel

The discrete signals in the system are:

$$\mathbf{y} = \sum_{i=1}^{N} \mathbf{H}_i \mathbf{D}_{\mathcal{K}_i} \mathbf{x}_i + \mathbf{n}, \tag{7.15}$$

where $\mathbf{D}_{\mathcal{K}_i}$ is a diagonal matrix with ones in the DFT coefficients contained in $\mathcal{K}_i$ and zeros elsewhere, channel matrices are normalized, and each user's effective transmitted signal $\mathbf{D}_{\mathcal{K}_i}\mathbf{x}_i$ is subject to the power constraint. This model allows arbitrary bandwidth allocation for each user but we focus on the following two extreme cases:

- **SIC:** All users transmit over the whole spectrum $\mathcal{K}_i = [0 \dots K] \forall i$. Since transmissions are overlapped, the receiver must apply SIC to decode the user transmissions. This strategy is known to be optimal in general [171].
- **FDMA:** Each user is allocated a part of the spectrum $\mathcal{K}_i$ that is orthogonal to the other user parts with $\mathcal{K}_i \cap \mathcal{K}_j = \emptyset \forall i \neq j$. This strategy is in general suboptimal, but performs close to optimal for low SNRs [171].

7.2.2.2 BC System Model

We consider a BC channel with a common source with N_t transmission antennas delivering information to N MIMO receivers with N_r reception antennas each, using a total bandwidth B. The source has a power budget P.

Each receiver i observes

$$\mathbf{y}_i = \sqrt{G_i}\mathbf{H}_i\left(\sum_{j=1}^{N}\sqrt{\alpha_j}\mathbf{D}_{\mathcal{K}_j}\mathbf{x}_j\right) + \mathbf{n} \tag{7.16}$$

where G_i is the large scale path loss from the BC transmitter to i, parameter α_j such that $\sum_{j=1}^{N}\alpha_j = 1$ is the fraction of power P that the source assigns to the signal for receiver i and matrix $\mathbf{D}_{\mathcal{K}_i}$ represents the DFT coefficient allocation dedicated to the same receiver.

It is possible to analyze the MIMO BC channel with perfect CSI at the transmitter and the receivers using the duality of the corresponding MIMO MAC channel. To build a critical bandwidth analysis for the MIMO BC without CSI, we argue that is is possible to stablish a relation with the MIMO MAC without CSI through the use of a CSI-independent upper bound and by applying the duality of the channels with CSI to compute a lower bound.

7.2.2.3 Definition of Metrics

The following definitions will be used to extend the notion of critical bandwidth in [76] to multiuser channels (MAC and BC).

The *achievable rate region subject to the total bandwidth B of multi-user channel scheme* S will be denoted by $\mathcal{R}_S(B)$.

Definition 7.8. The *multi-user critical bandwidth*, B_S^{crit} of a communication scheme S is the bandwidth beyond which increasing the bandwidth does not extend the rate region under scheme S.

$$B_S^{\text{crit}} = \inf\left\{B : \mathcal{R}(B) = \mathcal{R}(B'), \forall B' > B\right\} \tag{7.17}$$

We say that a multi-user scheme S is **overspread** when it is assigned more bandwidth than the B_S^{crit}.

7.2.2.4 FDMA MAC

The rate region using FDMA is simply the union over all possible bandwidth allocations of the combination of independent restrictions on each user:

$$\mathcal{R}_{\text{FDMA}}(B) = \bigcup_{\{\mathcal{K}_i\}_{i=1}^{N}}\left\{\{\mathcal{R}_i\}_{i=1}^{N} : R_i \leq \sup_{f(X_i)} \mathrm{I}\left(X_i; Y|\{X_j\}_{j\neq i}\right)\right\} \tag{7.18}$$

In the orthogonal band of each user, the rate as a function of bandwidth $R_i(B_i)$ grows independently. Therefore

Remark 7.9. FDMA has multi-user critical bandwidth

$$B_{\mathrm{FDMA}}^{\mathrm{crit}} = \sum_{i=1}^{N} B_i^{\mathrm{crit}} \tag{7.19}$$

Remark 7.10. Replacing $\sum_{i=1}^{N} P_i$ in lemma 7.6 determines $B_{\mathrm{FDMA}}^{\mathrm{crit}}$.

Lemma 7.11. *Any fading FDMA sum-rate at* $B_{\mathrm{FDMA}}^{\mathrm{crit}}$ *is*

$$\sum_{i=1}^{N} R_{i,\mathrm{FDMA}}(B_i^{\mathrm{crit}}) \geq \log_2(e)\frac{\sum_{i=1}^{N} P_i}{N_0} N_{\mathrm{r}} \left[1 - \sqrt{(\kappa(H) - 2 + N_{\mathrm{r}} + N_{\mathrm{t}})\log(\pi)\frac{1+\log(B_{\mathrm{c}}T_{\mathrm{c}})}{B_{\mathrm{c}}T_{\mathrm{c}}}}\right] \tag{7.20}$$

Proof. Lemma 7.2 on orthogonal bands. □

7.2.2.5 SIC MAC

Let $p(i) \in \mathcal{P}(N)$ denote the i-th element of a permutation p in the set $\mathcal{P}(N)$ of all permutations of the sequence $1, 2, \ldots, N$. Each permutation represents one possible SIC decoding order. The rate region of a MAC channel using superposed transmission and SIC is given by expression (7.21)

$$\mathcal{C}_{\mathrm{SIC}}(B) = \sup_{f(\{X_i\}_{i=1}^N)} \bigcup_{p \in \mathcal{P}(N)} \left\{ \{\mathcal{R}_i\}_{i=1}^N : \begin{array}{l} R_{p(i)} \leq \mathrm{I}\left(X_{p(i)}; Y | X_{p(0)} \ldots X_{p(i-1)}\right) \\ R_{p(i)} \leq \mathrm{I}\left(X_{p(i)} \ldots X_{p(N)}; Y | X_{p(0)} \ldots X_{p(i-1)}\right) - \sum\limits_{j=p(i+1)}^{p(N)} R_j \end{array} \right\} \tag{7.21}$$

evaluated over all decoding orders, where the rate of each link is limited both by its individual mutual information and the sum-rate limitation for that link and all those that are decoded afterwards.

In FDMA the results are a mere aggregation with $\sum_{i=1}^{N} P_i$ of lemmas 7.2 and 7.6, but this is not true for SIC. We define

$$\Upsilon = \frac{\left((\sum_{i=1}^{N} P_i^2)(\kappa - 2 + N_{\mathrm{t}} + N_{\mathrm{r}}) + 2N_{\mathrm{t}} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} P_i P_j\right)}{(\sum P_i)^2} \leq (\kappa - 2 + N_{\mathrm{r}} + N_{\mathrm{t}}) \tag{7.22}$$

which may be interpreted as an effective number of antennas that decreases when users have different power budgets.

Lemma 7.12. *SIC has multi-user critical bandwidth bounded by*

$$B_{\mathrm{SIC}}^{-} \leq B_{\mathrm{SIC}}^{\mathrm{crit}} \leq B_{\mathrm{SIC}}^{+}, \tag{7.23}$$

where

$$\begin{aligned} B_{\mathrm{SIC}}^{-} &= \frac{P}{N_0} \frac{1}{2\sqrt{\Upsilon \log \pi}} \sqrt{\frac{B_c T_c}{\log(B_c T_c)}}, \\ B_{\mathrm{SIC}}^{+} &= \frac{P}{N_0} 2\sqrt{\frac{\Upsilon}{N_{\mathrm{t}}^2} \log \pi} \sqrt{\frac{B_c T_c}{\log(B_c T_c)}}. \end{aligned} \tag{7.24}$$

Proof. Adapting lemma 7.6 using Υ, the total sum rate of limitation terms in (7.21) ceases to grow at this bandwidth. □

Lemma 7.13. *The sum of the rates using SIC at the critical bandwidth of this protocol is smaller than the sum rate achievable with FDMA.*

$$\sum_{i=1}^{N} R_{i,\mathrm{SIC}}(B_{\mathrm{SIC}}^{\mathrm{crit}}) < \sum_{i=1}^{N} R_{i,\mathrm{FDMA}}(B_{i}^{\mathrm{crit}}) \tag{7.25}$$

Proof. An upper bound on the sum rate with SIC can be obtained modifying lemma 7.2 with the total power $\sum_{i=1}^{N} P_i$ and the effective number of antennas Υ. By the definition of Υ, this is smaller than the lower bound for FDMA in lemma 7.11. □

Theorem 7.14. *To achieve all points in the capacity region of a MAC for every value of B, SIC and FDMA must alternate. Superposed transmissions must be used at bandwidths $B < B_{\mathrm{SIC}}^{\mathrm{crit}}$ and orthogonal transmissions at bandwidths $B > B_{\mathrm{FDMA}}^{\mathrm{crit}}$.*

Proof. At $B < B_{\mathrm{SIC}}^{\mathrm{crit}}$, the sum-rate of SIC is higher. At $B = B_{\mathrm{SIC}}^{\mathrm{crit}} < B_{\mathrm{FDMA}}^{\mathrm{crit}}$, the sum-rate of SIC stops growing but FDMA continues to grow (lemma 7.12), eventually outperforming SIC at $B = B_{\mathrm{FDMA}}^{\mathrm{crit}} > B_{\mathrm{SIC}}^{\mathrm{crit}}$ (lemma 7.13). □

Remark 7.15. There exists an intermediate region $B_{\mathrm{SIC}}^{\mathrm{crit}} < B < B_{\mathrm{FDMA}}^{\mathrm{crit}}$ where our results are inconclusive about the best strategy to choose. A combination of both protocols might be necessary.

7.2.2.6 Dual Critical Bandwidths for BC

In a-priori CSI models, the capacity region of a BC is typically expressed as the convex hull among all possible power allocations $\{\alpha_i\}_{i=1^N}$ where for each power allocation the rate region corresponds to a dual MAC channel that defines some points of the capacity region border.

$$\mathcal{C}_{\mathrm{BC,CSI-T}} = \bigcup_{\{\alpha_i\}_{i=1}^{N}} \mathcal{C}_{\mathrm{MAC,CSI-T}}(P_i = \alpha_i G_i P) \tag{7.26}$$

Both in [76] and in our lemmas the chain rule $\mathrm{I}\left(X;Y\right) = \mathrm{I}\left(X,H;Y\right) - \mathrm{I}\left(H;Y|X\right)$ and the lower bound $\mathrm{I}\left(X,H;Y\right) > \mathrm{I}\left(X;Y|H\right)$ are applied. Lower bounds originate the max sum-rate results, and Lemma 7.13 and upper bounds help to determine the critical bandwidth. For the BC channel, the following considerations allow a similar analysis:

- First, we consider the complete family of rate limitations composed of each limitation of the partial sum rate of an arbitrary subset of the users $\mathcal{A}$. Individual link limitations can be defined using single-element sets.

$$\sum_{i\in\mathcal{A}} R_i \leq \mathrm{I}\left(X; \bigcup_{i\in\mathcal{A}}\{Y_i\}\right) \forall \mathcal{A} \subset \{1\dots N\}$$

- For any $\mathcal{A}$, applying the chain rule, the first term $\mathrm{I}\left(X, \bigcup_{i\in\mathcal{A}}\{H_i\}; \bigcup_{i\in\mathcal{A}}\{Y_i\}\right)$ is upper bounded by the entropy of a series of i.i.d. Gaussian channel outputs $\tilde{Y}_i \sim \mathcal{CN}(0, 1 + \frac{\mathrm{E}\left[|\mathbf{H}_i|^2 P\right]}{BN_0})$. For each receiver, outputs $\tilde{Y}_i$ are i.i.d. and $\mathcal{A}$ is irrelevant to them, so any upper bound is merely the union of upper bounds on the individual links.

- By lower bounding the first term with $\mathrm{I}\left(X; \bigcup_{i\in\mathcal{A}}\{Y_i\}|H\right)$, the term satisfies the MAC-BC duality with CSI-R.

- Both in the upper and lower bounds, the second term of the chain rule $-\mathrm{I}\left(\bigcup_{i\in\mathcal{A}}\{H_i\}; \bigcup_{i\in\mathcal{A}}\{Y_i\}|X\right)$ applied to any of the rate restrictions above is simply the sum of individual contributions at each user $\sum_{i\in\mathcal{A}} \mathrm{I}\left(H_i;Y_i|X\right)$. Therefore the terms that penalize channel uncertainty are also reduced to the union of the individual link limitations.

- Finally, we have a family of upper bounds defined by $\mathcal{A}$ as a linear combination of single-link bounds

$$\mathrm{I}\left(X; \bigcup_{i\in\mathcal{A}}\{Y_i\}\right) \leq \sum_{i\in\mathcal{A}} H(\tilde{Y}_i) - \mathrm{I}\left(H_i;Y_i|X\right)$$

  and the following family of lower bounds that satisfy the BC-MAC duality with CSI-R minus a penalty term defined by $\mathcal{A}$ as a linear combination of single-link penalizations

$$\mathrm{I}\left(X; \bigcup_{i\in\mathcal{A}}\{Y_i\}\right) \geq \mathrm{I}\left(X; \bigcup_{i\in\mathcal{A}}\{Y_i\}|H\right) - \sum_{i\in\mathcal{A}} \mathrm{I}\left(H_i;Y_i|X\right)$$

Using the transformations above, the inner bound of the capacity region of a MAC channel can be transformed into a BC channel using mere power allocations, as typically done in the CSI-R case. And the outer bound of the capacity region merely remains the same regardless of which sets $\mathcal{A}$ are in reality relevant bound the capacity region. Taking the convex hull over all power allocations thus provides a result that is equivalent to Theorem 7.14: in the BC channel, orthogonal transmissions must be used below some critical bandwidth, and Frequency Division Multiplexing (FDM) must be used above the critical point.

It is worth mentioning that these dual bounds of a BC channel without CSI and those of a MAC channel without CSI differ in the penalization term, which is $\sum_{i\in\mathcal{A}} \mathrm{I}\left(H_i; Y_i|X\right)$ in BC and $\sum_{i\in\mathcal{A}} \mathrm{I}\left(H_i; Y_i|X_i\right)$ in MAC. This is because BC receivers estimate the channel from a common "pilot" with all the power of the system, whereas in MAC channels the receiver needs to estimate each channel from an different independent "pilot" signal transmitted only with the power of one user.

7.3 PHY Layer

Wideband massive MIMO is a recurring approach in 5G proposals as massive antenna array beamforming compensates the path loss suffered by high frequency. Nowaday there are prototype systems in industry and academia [51].

However it is not immediate how to handle high bandwidth and large antenna arrays at the same time. The potential gains from wideband massive antenna deployments are easier to derive in the limit case when the number of transmission and/or reception antennas and bandwidth tend to infinity. But even if noise and channel fading effects become negligible for an infinite number of antennas, they can contribute a significant error if that number is finite. The question of how large the system needs to be for the asymptotic analysis to hold remains a pertinent question in many application scenarios. In this section we discuss these issues for the design of three different PHY architectures.

7.3.1 Band-Limited Non-peaky Signaling

The first PHY architecture we consider consists in the combination of conventional PHY techniques with detection and allocation of resources constrained to their maximal values. As we showed in Section 7.2, for non-peaky signaling there is a critical bandwidth that can be calculated reasonably. Afterwards, it is sufficient to set those maximal values as thresholds and fit system transmissions to the critical values. For example, in an

OFDMA technology like LTE, a rectangular mask with variable width adjusted to the critical bandwidth may be used to limit the number of RBs dedicated to each user in the allocation process we described in Section 3.2.

It should be noticed, however, that the "achievable rate" discussed in 7.2 is obtained by maximizing a lower bound of the non-peaky signaling rate that considers the rate by coherent non-peaky signaling minus a rate penalty of **ideal** channel estimation. Therefore, the actual performance of a non-peaky signaling system with a practical channel estimation technique [172–174] may be degraded compared to the theoretical bounds we have obtained. In addition, traditional OFDM schemes relying on channel estimation usually require synchronous communication, but some new paradigms such as M2M often require asynchronous operation [175]

7.3.2 Wideband Peaky Signaling

Noncoherent transceivers are based on direct estimation of the transmitted signal without channel knowledge. As the phase information is usually lost, the amplitude information of the signal carries the information. Peaky signals are intrinsically suitable for the implementation of non-coherent receiver architectures, and vice-versa. A few examples of non-coherent peaky schemes are:

- On/off signaling with a vanishingly small "on" phase. This is the most basic non-coherent scheme but it has very poor spectral efficiency, requiring about 618% the bandwidth of the AWGN channel for the same rate [80].
- Pulse Position Modulation (PPM). This is basically a multi-level equivalent of the on/off signaling scheme where information is encoded in the instant when the "on" phase takes place. In this case it is easy to determine that the poor spectral efficiency is due to the fact that, to encode n bits, pulse duration must be very long to make room for 2^n distinct pulse positions.
- Frequency Shift Keying (FSK). This scheme uses a single tone within a collection of possible carriers to encode information and it can achieve capacity [79]. From a theoretical point of view, it is merely the Fourier-transform dual of PPM. However, from the point of view of implementation, it is highly advantageous by permitting to implement signals with constrained peak power. It still has poor spectral efficiency as it is also a $n \to 2^n$ projection.
- Multi-tone FSK (m-FSK). The spectral efficiency of FSK improves by allowing to activate a number of subcarriers $\binom{M}{Q}$ at a time, as in [77]. This greatly increases

the number of bits that may be encoded on a single FSK symbol. As $B \to \infty$, however, the ratio $\frac{Q}{M}$ goes to zero (but spectral efficiency is better than with PPM).

- Flash m-FSK. In [77], the m-FSK scheme is combined with an on/off cycle $\delta \to 0$ that allows to maintain spectral efficiency. In fact, this signaling achieves the subquadratic polynomial result with $\delta = \mathrm{SNR}^{1-\alpha}$ in [78, 81]. This shows the value of our analysis in Section 7.2, where we showed that this type of polynomial capacity is actually achieved with the same asymptotic behaviour by any signal with the same bandwidth occupancy (δB). This result manifests itself in the interplay between δ and the maximum Q in flash m-FSK signals [78, 81].

7.3.3 Wideband Massive MIMO Signaling

In the traditional analysis of wideband channels, there is an equivalence between B going to infinity, SNR going to zero, and capacity being power-limited. However, the work in [75, 76, 78] assumed a fixed number of antennas, which leads to $\lim_{B\to\infty} \frac{N_\mathrm{r}}{B} = 0$, meaning that SNR $\to 0$ (low SNR regime). However, if we introduce an asymptote in $N_\mathrm{r} \to \infty$, the results in that work do not necessarily hold for Massive MIMO.

Particularly, it is straightforward that a MIMO receiver with full CSI-R would use coherent combining with gain $\times N_\mathrm{r}$ to make $\mathrm{SNR} \propto \frac{N_\mathrm{r}}{B}$. Hence, the coherent model would be in the high-SNR regime if $\lim_{B,N_\mathrm{r}\to\infty} \frac{N_\mathrm{r}}{B} = \infty$. However, in a system without CSI-R, there is no such gain and the SNR is, by definition, low.

And, however, Manolakos et al in [176, 177] explored energy-based noncoherent narrowband massive SIMO wireless system where only the large-scale channel and noise statistics are known to the transmitter and the receiver. They presented optimized constellation designs for which the number of receive antennas is realistic for current technology. Their results show that very simple amplitude modulation designs could take advantage of large sets of reception antennas to reduce the error exponent. The conclusion is that, even though non-coherent schemes do not have a combining gain, their **diversity gain** still reduces error probability and makes rate grow with N_r

Therefore, the question that must be answered is what is the value of a massive number of receive antennas in a wideband non-coherent scheme. In a conference work accepted in IEEE Information Theory Workshop (ITW) 2015 [10] we consider a wideband massive SIMO non-coherent fading channel, motivated by the emergence of massive MIMO systems [178], and study the effect of the joint scaling of bandwidth and number of antennas on the achievable rates.

We consider a SIMO system (over a MIMO system) partly due to the simplicity of presentation and partly due to the fact that the analysis for SIMO systems generalizes very easily to MIMO systems with a finite number of transmit antennas. We investigate the capacity scaling of SIMO systems with simultaneously large receive antenna arrays and bandwidth. We also provide a practical encoding scheme that achieves capacity scaling.

Our results show that when the number of antennas grows much faster than bandwidth, there is a new wideband operating regime where, despite of low SNR, a different type of bandwidth-limited scaling of capacity takes place. To the best of our knowledge, this regime has never been previously discussed in the literature, as traditional wideband analysis always assumes an equivalence between the concepts of B going to infinity, SNR going to zero, and capacity scaling being power-limited [75, 76, 78].

Our analysis takes into account the joint scaling of the number of receive antennas and the bandwidth and suggests that there exists a *critical scaling* of the bandwidth that is proportional to the square root of the number of receiver antennas. This scaling is characterized by the following:

- When the bandwidth is smaller than the critical value in a scaling law sense ($B \leq o(N_{\mathrm{r}}^{\frac{1}{2}})$), the achievable rate is bandwidth-limited and grows with bandwidth. Surprisingly, a new type of regime where SNR is low but capacity grows with bandwidth is obtained. With this scaling, rates can be achieved in practice using a multi-tone generalization of the narrowband scheme in [176, 179].
- When bandwidth scales faster than the critical value ($B \geq \Theta(N_{\mathrm{r}}^{\frac{1}{2}+\alpha})$ for some $\alpha > 0$), additional bandwidth does not help to increase the achievable rates. This is similar to the problem of overspreading, previously reported for fixed N_{r}. Optimal achievable rates are obtained by restricting transmission to a $o\left(n^{\frac{1}{2}}\right)$ subset of the bandwidth.

The fundamental practical interpretation of these results is that, for a sufficiently large number of reception antennas N_{r}, the cooperative PHY technique may use all the bandwidth by encoding information in the amplitude of transmitted energy, relying on spatial diversity for decoding [176, 179]. Nevertheless, when the number of reception antennas is not large enough the critical bandwidth occupancy limitations we identified in Section 7.2 hold and non-coherent signaling schemes must use a subsect of the bandwidth and encode information in the frequency carriers with a fixed transmitted energy, as in [77, 78, 81].

7.4 MAC Layer

7.4.1 Resource Allocation in Dynamic Duplex Tree mmWave Network Topologies

In a collaboration with doctoral student Russell Ford [9], we studied DD in a cellular network where the topology keeps the tree-like topology of the LTE-A RN specification: Individual UEs and RNs only attach to one AP at a time, and such device interconection forms one tree per cell with the DeNB in its root ("top"). The novelty in our scheme is that we consider a duplexing policy where individual links can select their own transmission-reception duplexing pattern. Specifically, we consider a system whose subframes are synchronized network-wide, but the transmission/reception selections at each subframe can be made on a link-by-link basis (a feature uniquely practical in the mmWave range due to directional isolation). This flexibility allows to optimize the duplexing pattern dynamically according to current traffic load and channel conditions. In addition, duplexing can be adapted to local topological constraints. This adaptation is particularly valuable, since the number of hops and their capacity are likely to vary significantly due to different cell sizes, propagation obstacles and availability and quality of wired backhaul.

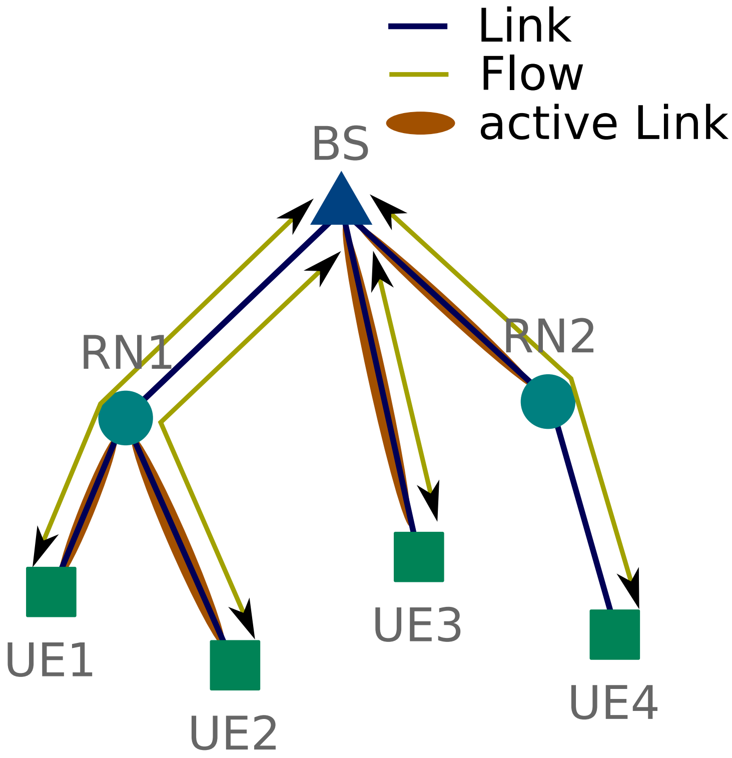


Figure 7.4: Tree-like topology and scheduling in a mmWave cell.

We consider an OFDMA-TDD tree cellular network with directional smart antennas and multi-hop transmission, and in-band relaying as shown in Fig. 7.4, where a BS with wireline backhaul provides a root from which connections to N_u UEs via N_r DF RNs depart. The RNs behave as self-backhauled BSs and are essentially indistinguishable in operation from the wired BS from the UEs. For the sake of clarity, we assume a two-hop network, although the methods and algorithms we develop apply to tree networks with an arbitrary number of hops. Each user is associated with either the BS (i.e. direct-link) or a singe RN, completing the tree topology.

Users are scheduled according to a series of frames of period T_f, which are further subdivided into N_sf subframes of period $T_\mathrm{sf} = \frac{T_\mathrm{f}}{N_\mathrm{sf}}$. From the perspective of each node, each subframe can be designated for DL or UL transmission indicating whether the transmissions in that subframe go towards the BS tree root or vice-versa. A subframe can also be muted (i.e. unused). Within each subframe, we assume that Orthogonal Frequency Division Multiple Access (OFDMA) is employed, allowing multiple users to be allocated orthogonal frequency resources within the same subframe.

In contrast to the semi-static and globally synchronized TDD configurations supported by relay-enhanced TD-LTE networks, we allow each individual BS and RN to dynamically select the transmission mode of each subframe. This dynamic is coordinated by the BS in a centralized fashion through control messaging, although the operation of the specific MAC protocol is beyond the scope of this thesis.

Users have one DL flow and one UL flow, but the model can be easily extended to multiple differentiated types of traffic per user. When a link is active and operating in the UL regime, the receiver is higher up the tree than the transmitter and it sees a MAC channel. The rate between the UL transmitter (UE or RN) and the receiver (RN or BS) depends on the total power of the transmitter and the bandwidth allocated by the receiver to each of its users. Conversely, when a link is active in the DL regime, the transmitter is higher up in the tree and sees a BC channel, and the rate depends on the simultaneous allocation of bandwidth and power by the transmitter to each of its BC users.

We consider the problem of maximizing the sum utility function of all flows, optimized over all possible UL/DL designations of each subframe on each node (which we refer to as *scheduling*), and over all possible bandwidth allocations and power allocations given a fixed scheduling.

The problem is a binary mixed-integer non-convex optimization. A number of algorithms and heuristics have been proposed for OFDMA-TDD scheduling, many of which employ

Table 7.1: UE mean and cell-edge (worst 5%) rates for static TDD and DD

		Mean		5-percentile	
		DL	UL	DL	UL
2 RN	TDD	11.86	11.60	2.55	3.12
	DD	28.91	27.82	5.67	3.85
4 RN	TDD	8.51	8.88	3.32	3.61
	DD	37.08	38.11	8.55	5.64

mixed-integer methods over the integral search space formed by the subcarrier and time slot indices.

However, two key aspects of our problem allow to reduce the size of the search space. Firstly, subframe allocation can be performed individually for each BS and RN without regard to interference with other nodes due to the high spatial isolation of mmWave transmissions, which make inter-cell interference negligible. Secondly, the hierarchical structure of the tree network allows to formulate the optimization as a recursive operation, from the BS down to the UEs, and reassign resources to links to improve utility iteratively.

We performed simulations of the algorithm in realistic deployments with two and four RNs per cell, and achieved a four-fold average rate gain and two-fold cell-edge rate gain versus static TDD allocations, even under fair uniform traffic. Table 7.1 shows the results.

7.4.2 Link Scheduling in mmWave Networks with Dynamic Duplex Arbitrary Topology

The previous result suggests promising rate improvements through the relaxation of duplexing requirements from TDD to DD, but it is still strongly tied to the LTE-A relaying architecture due to the assumption of a tree topology (that is, a single attachment point per UE). However, the increase in scheduling flexibility brought by highly directive antennas does not have to stop at dynamic duplex in tree topologies, as the spatial isolation between transmissions has made it possible to attach each device alternately to multiple APs. In this section we focus on the generalization of the wireless network topology of 5G systems to an arbitrary mesh, and the contributions of an improved scheduling protocol to the higher layers. Note that, even if in some scenarios a tree topology is indeed optimal, the solution of the generalized problem should lead to the same conclusions as above.

The previous section is more in line with previous research [50, 180, 181] that analyzes massive MIMO, cognitive radio, mmWave and relaying and has focused on the properties of the physical layer, to determine the best allocation of available radio resources in a simple tree topology. Since mmWave research has paid little attention to mesh networking in the past, we import results from other fields such as ad-hoc networks, to help in the design of a multi-hop 5G cellular systems with even higher capacities through the generalization of the topology to an arbitrary mesh.

In a collaboration with doctoral student Juan García Rois [3], currently being revised for publication[1] , we analyze the scheduling problem for a 5G cellular mesh network (Fig. 7.5) introducing the recent physical layer models in [51, 74, 170, 182, 183] in an architecture combining the flexibility of DD with optimal control of multi-hop wireless mesh networks as in [87]. Our architecture features multi-hop communications with multiple-attachment relays, full operation at mmW bands, directive-only antenna configurations (no omnidirectional control signals), DD scheduling, half-duplex transmission/reception separation, and single-link-per-device activation constraints. Due to the high complexity of arbitrary topology scheduling, we do not consider that a node can transmit in multiple links at the same time, avoiding the burden of bandwidth and power allocation.

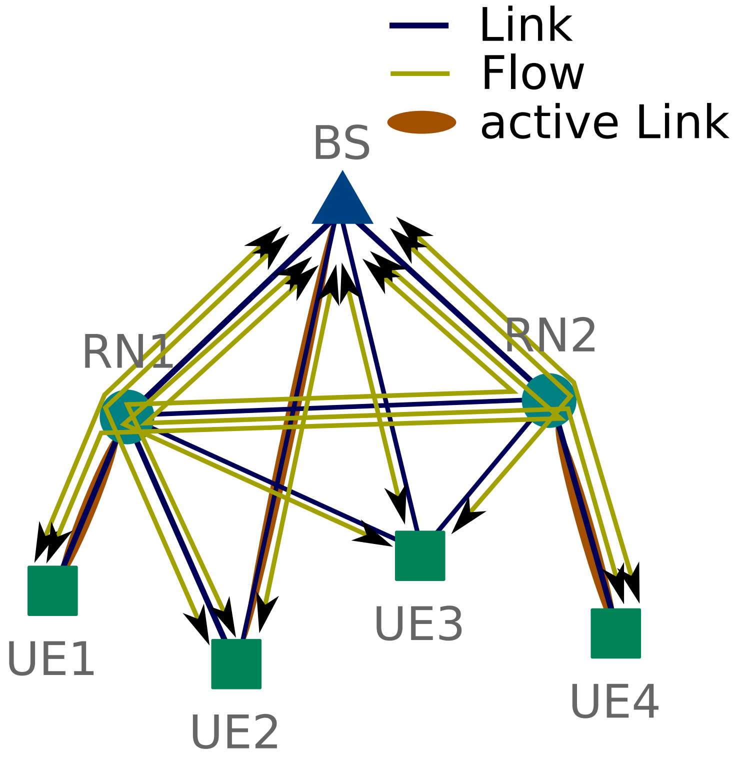


Figure 7.5: Arbitrary graph topology and scheduling in a mmWave cell.

[1] The PhD candidate in this thesis contributed in the design of the mmWave connectivity and link capacity models, and their implementation in the simulator.

This work presents an analytical model for the problem of joint scheduling and congestion control within the well known framework of Network Utility Maximization (NUM), which allows to evaluate the capacity region of the new mmWave cellular architectures. The main contribution is the combination of mmWave characteristics and classic results in multi-hop scheduling. We propose a wireless architecture for mmWave operation consisting in a multi-hop backhauling mesh structure. It is possible to connect a user UE with several RNs (not simultaneously) or the base station (BS), and RNs and BSs with any other node. Two fundamental constraints are placed on mmWave links: *half-duplex* transmissions, such that nodes cannot transmit and receive at the same time, and *one-to-one* communications, so that a node cannot use several links at a time, since analog beamforming is limited to one direction at a time [184]. These constraints are likely to define the first generation of mmWave devices.

Furthermore, we assume a complete DD TDD operation, without any restriction in terms of UL/DL simultaneous transmissions, thanks to the spatial isolation in mmWave bands. This is a fundamental difference with current cellular systems. We refer to the set of transmitter/receiver pairs chosen to be active within a certain frame as a *schedule*.

As previously said, our architecture follows the most recent results in the literature, such as [51, 170, 182]. For this reason me call it an Actual Interference (AI) model. It is graph-based, yet it accounts for Signal to Interference plus Noise Ratio (SINR) and is specially suitable for mmWave networks.

Our main contribution is the maximum achievable capacity of the proposed mmWave cellular system architecture under fairness requirements. In order to maximize the set of feasible UL/DL rates per user and enforce fairness we follow the NUM framework. Given that the AI model considers several layers of the protocol stack, including the Physical (PHY), Medium Access Control (MAC), network, and transport layers, the problem is decomposed in a cross-layer fashion. This allows us to approach the classic Maximum Back-Pressure optimal solution of queue size (weighted by variable link capacities) using the also classic approximation of the random Pick-and-Compare (PaC) scheduling algorithm with a Dual Congestion Control (DCC) mechanism. To show that throughput-utility optimality is indeed achieved in the AI model, we adapt the proof in [87] to the case when link capacities experience schedule-dependent interference.

Considering recent work suggesting that mmWave interference is negligible [98, 185, 186], we also define two simplified, but still *topology-aware*, interference schemes: Interference Free (IF) and Worst-case Interference (WI), to test the validity of such assumption. These schemes provide upper and lower bounds of the capacity achieved by the AI model but are computationally simpler thanks to the assumption of static link capacities. Even though we have found some problem instances where cross-interference by simultaneously selected links cannot be ignored, throughput-utility optimal management of dynamic duplex transmissions allows the simplified IF model to provide a very tight upper bound of AI perfoirmance. We interpret that this occurs because our optimal management algorithm avoids the scheduling of links with significant cross-talk, performing an implicit interference avoidance. This observation suggests that the IF model allows a realistic evaluation of system capacity in mmWave cellular networks as long as we can guarantee that the real network is operated with optimal control. This is a fundamental difference with traditional cellular and wireless systems, where a similar IF model would provide much loose bounds. A byproduct of this result is that a significant part of wireless ad-hoc graph theory may be applied to the new mmWave cellular paradigms, whereas traditionally hot research areas on μWave cellular networks such as interference cancellation or power control will be much less relevant.

All these results have been validated by numerical simulation, using mmWave channel and beamforming models derived from real-world data [74].

7.5 Summary

Implementation practical signaling and networking solutions for 5G cellular networks to achieve the optimal scalings predicted in Chapter 6 is not trivial. There is still an ample field of research.

From the point of view of channel capacity, the significance of very large bandwidths for operational purposes needs to be better understood. Until very recently, the common belief was that the overspreading problem in a point to point non-coherent fading channel could be averted through the implementation of increasing signal peakiness. If this was true, the transition from non-peaky schemes with bandwidth limitations to peaky schemes capable of unlimited bandwidth exploitation would be merely a technological matter. Unfortunately, we have proven that this is not the case: By defining the *bandwidth occupancy* metric, which measures average bandwidth usage over time, it can be shown that peaky and non-peaky signals are in fact treated by the channel in the same way. There is a bandwidth occupancy limit for any signal, and exceeding that limit does not provide any gain.

Our model shows that, since spectrum is a scarce resource, from the perspective of many simultaneous independent point to point channels there is little difference between a classic implementation in adjacent dedicated narrow bands or one in a single wider band which systems occupy alternately as in cognitive radio (See Chapter 4). But unfortunately, the interplay between multiple-user channels and massive bandwidths is more complex than that. Even if it is true that, for a very large bandwidth, MAC and BC channel users must be allocated in orthogonal bands using FDMA/FDD, our results also show that simultaneous transmissions are compulsory below the critical bandwidth of joint signal processing techniques. Moreover, the problem becomes even harder in an intermediate bandwidth regime, above the bandwidth where joint signaling is effective but below the bandwidth where orthogonal signaling is optimal. In that intermediate regime, hybrid strategies should be explored.

The implementation of signaling schemes in the PHY layer may follow the trend of coherent schemes relying on channel estimation. Nevertheless, practical pilot-assisted schemes may not suffice to achieve capacity. Non-coherent signaling schemes may help, but, due to their reduced spectral efficiency, they are only valuable very close to the critical bandwidth or above it, as in the case of flash peaky signals. A promising alternative to circumvent the problem of overspreading is the combination of wideband and massive MIMO, as we have shown that a sufficiently high number of receive antennas provides gains even in non-coherent receivers. However, unlike the coherent case where the antenna combination gain increases SNR to avoid the low-SNR regimes, in non-coherent systems reception antennas can only produce diversity gain (with lower scaling by a factor of $\frac{1}{2}$ in the exponent). This creates a new type of operating regime in the meeting point of wideband and massive MIMO, where capacity is degrees-of-freedom limited although SNR goes to zero.

Finally, the design of mmWave 5G MAC protocols to implement cooperative multi-hop schemes requires dramatic changes in comparison with traditional system design. Replacing omnidirectional transmissions with highly-directive beams makes signaling between nodes more difficult, but it is advantageous through increased spatial isolation and scheduling flexibility. First, it is possible to relax the duplexing requirements of traditional cellular protocols, so that UL and DL traffic scheduling is more flexible, even for the traditional single-point attachment requirement. Second, it is possible to introduce more advanced routing strategies by allowing devices to attach themselves to (i.e. deliver traffic through) multiple infrastructure APs at the same time, and allowing RNs to communicate with each other. The scheduling problems that arise from such layouts are extremely complex, since they rhave multiple dimensions including the allocation of power and bandwidth to simultaneous transmissions, the selection of simultaneouly active links, the accurate modeling of interference as a time-variant phenomenon depending on the current set of active links, etc. Mixed-integer optimization methods make possible to solve the problem by exploiting its recursive tree structure, whereas different practices from other fields of research such as ad-hoc networks allows to find optimal schedules, at least in a stochastic sense.

The contents of this chapter include recent work of this PhD that is currently in revision phase or still unpublished. Related publications that are available are:

- A journal paper submitted to to IEEE Transactions on Information Theory [4].
- A conference paper accepted in the IEEE International Information Theory Workshop (ITW 2015) [10].
- A journal paper submitted to IEEE Transactions on Wireless Communications [3].

Chapter 8

Conclusions

Contents

This thesis studies the present and future role of cooperative communications in wireless networks. We have identified three different stages for the deployment of cooperative communications in wireless networks, and particularly in cellular data networks, and discussed each stage in a part of this thesis. The first part corresponds to the present, when the theory for cooperative channels is relatively developed but its implementation in hardware is seemingly inexistent save for a few initiatives such as the static relays in the last version of the cellular standard. The second part corresponds to the near future (today-2020), during the duration of the life cycle of the current cellular generation. In itretro-compatibility will be a critical requirement and improvements of the state will consist in methods for improved usage of the same frequencies (μWave) and architectures (static duplex and tree topology). The last part corresponds to the distant future ($\gtrsim$2020), once the current generation will be exhausted. Drastic changes in design philosophy will be more plausible, allowing to expand cellular networks into new frequency bands (mmWave) with new topologies (dynamic duplex and mesh topology).

8.1 The Present

In the part dedicated to the present, in Chapter 2 we surveyed the theoretical research in cooperative diversity gains and in Chapter 3 we studied by simulation the performance of the relaying specification for LTE-A.

8.1.1 The Theory of Cooperative Communications

Regarding the theoretical study of cooperative diversity, we distinguished three work areas from a deep review of the literatue:

- At the information-theoretic level, the capacity of cooperative transmission is studied through models derived from the relay channel. A few key aspects to consider are the *relay function* $P(X_r = f(Y_r))$ that relates the distribution of the relay transmission to its input (examples are AF, DF ...), the time division of source and relay transmission (static versus dynamic), the orthogonality transmissions (source is silent or not during relay transmission), and the possibility of multiple encoding layers (as in Enhanced-DDF, which adds a third underlying transmission to DDF).
- At the PHY level, the combined processing of multiple cooperative channels can take place at different points of the radio chain: at the control level, by selecting the best relay and transmitting a single signal; at the level of channel coding through cooperative distributed concatenated codes; at the level of source coding through the combination of multiple data flows with network coding; or at the level of MIMO processing relying on distributed multi-antenna codes to form a virtual array.
- At the MAC level, we formulated that any cooperative MAC protocols must perform five different fundamental operations: Neighborhood mapping, design of optimal relay sets (with the marginal case of single best-relay selection), selection of cooperative or direct communications, notifications to the helpers and agreements with them, and design of cooperative transmission signaling schemes.

8.1.2 Practical Cooperation: Relaying in LTE-A

Regarding the performance of practical relaying in LTE-A, our system-level simulation revealed problems in upper OSI layers (scheduling, queuing) that previous LTE-A reseach had ignored.

- The use of a global parameter, common to all cells and relays, to determine access and relay time-sharing, negates the possibility of balancing links on a per-user basis. Consequently, almost all links operate far from their optimal point creating throughput bottlenecks.

- With RNs interference increases, as expected. But, regrettably, since links are not correctly balanced, this increase in interference occurs in vain, because even though RNs are transmitting no data is actually being delivered.

- Forcing RN operation in certain subframes hampers the flexibility required by the schedulers that exploit multi-user diversity, such as PF. If this happens, even if RNs create power gains in the PHY layer, these gains are canceled by the MAC layer, which destroys scheduling gains.

- Incentive parameters and improvement of balancing parameters in PF schedulers can, at most, mitigate the losses, never remove them completely. The worst case scenario is when PF turns into an unnecessarily complex equivalent of RR.

- The size of the coverage area of RNs is small, sometimes negligible. Admission Control should reject RNs that produce no benefits during the `relay_attachment_procedure`.

8.2 The Near Future

In the part dedicated to the near future, we focus on the severe scarcity of the μWave frequencies that current technologies employ and on the need of wireless networks to be retro-compatible in the short term. In Chapter 4, CR, and specially CSL, are identified as the main drivers of the introduction of cooperative communications in current standards. In addition, in Chapter 5 we discuss emergent communication paradigms with CSL potential such as P2P or M2M, whose application niches may benefit from LTE-A RN gains, unlike personal communications.

8.2.1 Analysis of Cooperative Spectrum Leasing

Regarding the study of CSL gains, we focus on the volume of aggregate social gains, unlike other CR works that focus on the equilibrium of game-theoretic competition between greedy individuals.

- We model the (social) spectrum gain as the amount of spectrum that a primary can release thanks to cooperation while still achieving the rate it sustained with direct transmission using the full spectrum.

- We found that the probability of spectrum gain converges to 1 when the primary has a bad channel distribution compared to the channels of the secondary, confirming that the desirable scenario for CSL is an inefficient primary system.
- The ergodic spectrum gains taking static long-term decisions experience a tipping point: if the primary is good enough, the gain is zero, but if the primary is bad, gains rapidly grow to 90%.
- The distribution of spectrum gains taking dynamic instantaneous decisions is smooth: for bad primaries the probability of a high spectrum gain is also high, but even, when the primary experiences a good channel in average, a CSL system may proceed during the (few) deep fading events of the primary to cooperate temporarily and obtain a small spectrum gain.

8.2.2 Opportunities for Cooperative Spectrum Leasing in LTE-A

Regarding the implementation of CSL in LTE-A and its potential applications, we found that the gains are moderate

- A single RN produces very tiny gains, in the order of a thousandth of the spectrum. The coverage of a RN is also very small.
- Since a cognitive helper is free for the cell, the tiny individual gains could be aggregated over hundreds of RNs until they are valuable. Using either regularly distributed or random RNs with admission control, the spectrum gains are still 20% only.
- The CSL secondaries employ LTE hardware to perform cooperation. Therefore, we can estimate cognitive rates informally by multiplying LTE spectral efficiency by half the spectrum gains.
- The result of this estimation is in the order of tens of Mbps. This is definitely not enough for future personal communications, but it is competitive with M2M technologies such as ZigBee (250Kbps). This confirms that LTE-A RNs, even with massive CSL support, will most likely be limited to niche applications

8.3 The Distant Future

In the part dedicated to the distant future, we focus on the drastic re-engineering of cellular networks that will take place when future communications will move into mmWave bands. Spectrum will no longer be so scarce, permitting transmissions with larger bandwidth; cell radiuses will decrease dramatically, fostering spatial multiplexing gains, and the number of antennas per integrated circuit area will increase as the wavelengths shrink. In Chapter 6 we derive the throughput capacity scaling laws of cellular networks in this future context, and prove that cooperative multihop is necessary to achieve optimal capacity scaling. In Chapter 7 we discuss qualitatively the information theoretic and PHY challenges of the new paradigm.

8.3.1 Capacity Scaling of 5G Cellular Networks

We have obtained the scaling of upper bounds on capacity and achievable rates for several protocols in cellular networks as a function of number of users, area, bandwidth, number of BS, and number of BS antennas. The results allow us to state the following.

- The capacity of a network with a scaling number of nodes experiences a "critical bandwidth scaling", in the same manner as a point-to-point link experiences a transition from a bandwidth-limited capacity in the narrowband regime to a power-limited capacity in the wideband regime. When the network is in the bandwidth-limited scaling regime, adding bandwidth increases rate. However, if bandwidth scales too much the network becomes power limited and additional spectrum is of no help in a scaling law sense.
- Different protocols experience diferent bandwidth scaling limit thresholds. Cooperative inter-user multi-hop is the best, since it guarantees achieving the optimal scaling for any value of the parameters. The protocol-specific limits depend on typical transmission distances, and thus, a multi-hop transmission implemented with fewer static relay nodes than users is worse than full user cooperation in terms of bandwidth exploitation, but it is still better than a protocol that simply issues direct BS-user transmissions as it is the norm in traditional cellular architectures.
- The traffic model has a deep influence in capacity scaling. In ad-hoc networks with dense user distributions (small area scaling), hierarchical cooperation is optimal. However, according to our result for cellular networks, an equivalent hierarchical infrastructure protocol is suboptimal. This is because in ad-hoc networks direct transmission flows may be directed at any random point in the network area, even across the entire network area, whereas in cellular networks direct transmissions are always directed to the closest BS (a much shorter maximum transmission distance). Direct transmission is thus inefficient in ad-hoc networks, but not in cellular networks with small bandwidth scaling. This means that the results on

some scaling models for infrastructure-assisted networks that still consider ad-hoc-like traffic, may not apply to future 5G networks.

8.3.2 Implementation Challenges of Cooperative 5G Communications

- At the information-theoretic level, the capacity of channels with large bandwidth needs to be better understood. Even though there is abundant literature covering point-to-point channels in the limit as $B \to \infty$, the ramifications of the problem once additional concepts are taken into account must be tackled. We have unified the so far seemingly distinct capacity-achieving peaky signaling and critical bandwidth with non-peaky signaling. Remarkably, the result is that signaling peakiness is not relevant for correct physical quantities. We have also generalized the concept of critical bandwidth to multi-user channels and showed that, in addition to the expected phenomena that superposed transmissions are better at small bandwidth and orthogonal transmissions are better at large bandwidth, there is an intermediate regime where bandwidth is too large for total superposition but too small for total orthogonality. The little attention in the mmWave literature to these two important issues is surprising.

- At the PHY level, there are multiple options to implement mmWave signaling . We can distinguish between the strategies that deal with the critical bandwidth occupancy limitation we saw in theory, and strategies that employ additional hardware (namely, massive MIMO) to try and circumvent it. The traditional signaling schemes already in use in LTE-A, which are smooth non-peaky signals, belomg to the first type; and the traditional non-coherent modulations such as FSK and their modern improvements such as m-FSK belong to the second. In the latter, we have analyzed the cases in which a sufficiently large number of reception antennas avoids the critical bandwidth limitation and allows to implement non-coherent channels were capacity grows with bandwidth without bounds. However, the high hardware cost -the number of reception antennas must exceed the square of the bandwidth, $\Theta(B^2)$- may not be affordable in the medium term.

- At the MAC level, highly-directive mmWave antennas bring deep paradigm changes. Due to the spatial isolation of interference, BS and RNs in different cells no longer need to perform DL-UL synchronized, and thus duplexing will be a dynamic decision per node that may be adjusted to optimize performance. In addition, the possibility to communicate with other access points nearby without interfering the closest ones makes multi-attachment more appealing. A completely general scheduling problem is rather complex, but we have studied the gains of DD with

MU-MIMO in a classical tree topology and with single user MIMO in a multi-attachment mesh topology, and showed that both strategies may dramatically increase the rates of a mmWave cellular network.

Bibliography

[1] Cisco Systems, "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018," 2013. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

[2] F. Gómez-Cuba, S. Rangan, E. Erkip, and F. J. González-Castaño, "Capacity Scaling of Cellular Networks with Arbitrary Infrastructure Density and Bandwidth," *In preparation for submission*, pp. 1–19.

[3] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castaño, J. C. Burguillo-Rial, S. Rangan, and B. Lorenzo, "On the Analysis of Scheduling in Dynamic Duplex Multi-Hop mmWave Cellular Systems," *Submitted to IEEE Transactions on Wireless Communications*, pp. 1–32.

[4] F. Gómez-Cuba, J. Du, M. Médard, and E. Erkip, "Unified Capacity Limit of Non-coherent Wideband Fading Channels," *submitted to IEEE Transactions on Information Theory*, p. 14. [Online]. Available: http://arxiv.org/abs/1501.04905

[5] F. Gómez-Cuba, R. Asorey-Cacheda, and F. J. González-Castaño, "Smart Grid Last-Mile Communications Model and Its Application to the Study of Leased Broadband Wired-Access," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 5–12, Mar. 2013.

[6] F. Gómez-Cuba, R. Asorey-Cacheda, F. J. González-Castaño, and H. Huang, "Application of Cooperative Diversity to Cognitive Radio Leasing: Model and Analytical Characterization of Resource Gains," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 40–49, Jan. 2013.

[7] P. Sendín-Raña, F. J. González-Castaño, F. Gómez-Cuba, R. Asorey-Cacheda, and J. M. Pousada-Carballo, "Improving Management Performance of P2PSIP for Mobile Sensing in Wireless Overlays," *Sensors (Basel, Switzerland)*, vol. 13, no. 11, pp. 15 364–84, Jan. 2013.

[8] F. Gómez-Cuba, R. Asorey-Cacheda, and F. J. González-Castaño, “A Survey on Cooperative Diversity for Wireless Networks,” *IEEE Communications Surveys & Tutorials*, vol. 14, no. 3, pp. 822–835, 2011.

[9] R. Ford, F. Gómez-Cuba, M. Mezzavilla, and S. Rangan, “Dynamic Time-domain Duplexing for Self-backhauled Millimeter Wave Cellular Networks,” in *Accepted in IEEE International Conference on Communications (ICC) Workshop on Next Generation Backhaul/Fronthaul Networks (BackNets)*, 2015.

[10] M. Chowdhury, A. Manolakos, F. Gómez-Cuba, E. Erkip, and A. J. Goldsmith, “Capacity Scaling in Noncoherent Wideband Massive SIMO Systems,” in *IEEE Information Theory Workshop (ITW) [Accepted for publication]*, 2015.

[11] F. Gómez-Cuba, S. Rangan, and E. Erkip, “Scaling Laws for Infrastructure Single and Multihop Wireless Networks in Wideband Regimes,” in *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[12] F. Gómez-Cuba and F. J. González-Castaño, “Improving Third-Party Relaying for LTE-A: A Realistic Simulation Approach,” in *IEEE International Conference on Communications (ICC)*, 2014.

[13] F. Gómez-Cuba, F. J. González-Castaño, and J. Muñoz Castañer, “Is Cooperative Spectrum Leasing by Third-Party Relays Advantageous in Next-Generation Cellular Networks?” in *20th European Wireless Conference (EW)*, 2014.

[14] F. Gómez-Cuba, F. J. González-Castaño, and C. P. Pérez-Garrido, “Practical Smart Grid Traffic Management in Leased Internet Access Networks,” in *IEEE International Energy Conference (ENERGYCON)*, 2014.

[15] F. Gómez-Cuba, R. Asorey-Cacheda, and F. J. González-Castaño, “WiMAX for Smart Grid Last-Mile Communications: TOS Traffic Mapping and Performance Assessment,” in *IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Oct. 2012.

[16] 3GPP, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (3GPP TS 36.300 version 10.7.0 Release 10),” *ETSI TS 136 300*, no. 10.7.0, pp. 1–204, 2012.

[17] ——, “Universal Mobile Telecommunications System (UMTS); Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer for relaying operation,” *ETSI TS 136 216*, no. 10.3.1, pp. 1–18, 2011.

[18] A. J. Goldsmith, "To Infinity and Beyond: New Frontiers in Wireless Information Theory (Plennary Talk)," 2014.

[19] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. A. Valenzuela, "Is the PHY Layer Dead?" *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[20] R. Vaze and R. W. Heath, "On the Capacity and Diversity-Multiplexing Tradeoff of the Two-Way Relay Channel," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4219–4234, Jul. 2011.

[21] K. J. R. Liu, A. K. Sadek, W. Su, and A. Kwasinski, *Cooperative Communications and Networking.* Cambridge University Press, 2008.

[22] M. Dohler and Y. Li, *Cooperative Communications: Hardware, Channel and PHY.* Wiley, 2010.

[23] Y.-W. P. Hong, W.-J. Huang, and C.-C. J. Kuo, *Cooperative Communications and Networking. Technologies and System Design.* Springer, 2010.

[24] L. Zheng and D. N. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels," *IEEE Transactions on Information Theory*, pp. 1–50, 2003.

[25] T. Cover and A. Gamal, "Capacity Theorems for the Relay Channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sep. 1979.

[26] J. N. Laneman, "Cooperative Diversity in Wireless Networks: Algorithms and Architectures," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[27] A. B. Saleh, S. Redana, B. Raaf, and J. Hämäläinen, "Comparison of Relay and Pico eNB Deployments in LTE-Advanced." in *IEEE Vehicular Technology Conference (VTC Fall)*, 2009.

[28] A. B. Saleh, B. Raaf, S. Redana, T. Riihonen, J. Hämäläinen, and R. Wichman, "Performance of Amplify and Forward and Decode and Forward Relays in LTE-Advanced." in *IEEE Vehicular Technology Conference (VTC Fall)*, 2009.

[29] A. B. Saleh, S. Redana, J. Hämäläinen, and B. Raaf, "On the Coverage Extension and Capacity Enhancement of Inband Relay Deployments in LTE-Advanced Networks," *Journal of Electrical and Computer Engineering*, 2010.

[30] K. Doppler, S. Redana, M. Wódczak, P. Rost, and R. Wichman, "Dynamic Resource Assignment and Cooperative Relaying in Cellular Networks: Concept and Performance Assessment," *EURASIP Journal on Wireless Communications and Networking*, 2009.

[31] L. Rong, S. E. Elayoubi, and O. B. Haddada, "Impact of Relays on LTE-Advanced Performance." in *IEEE International Conference on Communications (ICC)*, 2010.

[32] P. E. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity Compared to the Shannon Bound." in *IEEE Vehicular Technology Conference (VTC Spring)*, no. 1, 2007.

[33] J. C. Ikuno, M. Wrulich, and M. Rupp, "System Level Simulation of LTE Networks," in *IEEE Vehicular Technology Conference (VTC Spring)*, Taipei, Taiwan, May 2010.

[34] "ns-3 LTE Module." [Online]. Available: http://www.nsnam.org/docs/models/html/lte.html

[35] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On Number of Almost Blank Subframes in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 5061–5073, Oct. 2013.

[36] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking Spectrum Gridlock with Cognitive Radios: An Information Theoretic Perspective," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.

[37] O. Simeone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz, "Spectrum Leasing to Cooperating Secondary Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 203–213, Jan. 2008.

[38] S. K. Jayaweera, M. Bkassiny, and K. A. Avery, "Asymmetric Cooperative Communications Based Spectrum Leasing via Auctions in Cognitive Radio Networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2716–2724, Aug. 2011.

[39] Y. Han, A. Pandharipande, and S. H. Ting, "Cooperative Decode-and-Forward Relaying for Secondary Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 4945–4950, Oct. 2009.

[40] J. P. K. Chu, R. S. Adve, and A. W. Eckford, "Fractional Cooperation using Coded Demodulate-and-Forward," in *IEEE Global Telecommunications Conference (GLOBECOM)*. IEEE, Nov. 2007.

[41] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[42] I. Krikidis, J. N. Laneman, J. S. Thompson, and S. Mclaughlin, "Protocol Design and Throughput Analysis for Multi-User Cognitive Cooperative Systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4740–4751, Sep. 2009.

[43] I.-H. Lee and D. Kim, "Probability of SNR Gain by Dual-Hop Relaying over Single-Hop Transmission in SISO Rayleigh Fading Channels," *IEEE Communications Letters*, vol. 12, no. 10, pp. 734–736, Oct. 2008.

[44] 3GPP, "LTE; Service requirements for Machine-Type Communications (MTC); Stage 1 (3GPP TS 22.368 version 11.6.0 Release 11)," *ETSI TS 136 421*, no. 11.6.0, pp. 0–20, 2012.

[45] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1–1, 2013.

[46] Shao-Yu Lien, Kwang-Cheng Chen, and Yonghua Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[47] C. Ide, B. Dusza, M. Putzke, C. Muller, and C. Wietfeld, "Influence of M2M Communication on the Physical Resource Utilization of LTE," in *IEEE Wireless Telecommunications Symposium*, Apr. 2012.

[48] R. E. Brown, "Impact of Smart Grid on Distribution System Design," in *IEEE PES General Meeting*. Ieee, 2008.

[49] Z. Fan, G. Kalogridis, C. Efthymiou, M. Sooriyabandara, M. Serizawa, and J. McGeehan, "The New Frontier of Communications Research: Smart Grid and Smart Metering," in *ACM International Conference on Energy-Efficient Computing and Networking (e-Energy)*, ser. e-Energy '10, 2010.

[50] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[51] S. Rangan, T. T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[52] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Communications Magazine*, vol. 49, no. Jun., pp. 101–107, 2011.

[53] P. Pietraski, D. Britz, A. Roy, R. Pragada, and G. Charlton, "Millimeter Wave and Terahertz Communications: Feasibility and Challenges," *ZTE Communications*, vol. 10, no. 4, pp. 3–12, 2012.

[54] Y. Azar, G. Wong, and K. Wang, "28 GHZ Propagation Measurements for Outdoor Cellular Communications Using Steerable Beam Antennas in New York City," in *IEEE International Conference on Communications (ICC)*, 2013.

[55] T. S. Rappaport, R. W. Heath Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications.* Pearson Education, 2014.

[56] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. Feb., pp. 160–171, 2013.

[57] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems," *IEEE Communications Magazine*, vol. 52, no. Feb., pp. 186–195, 2014.

[58] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell Networks: A Survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, 2008.

[59] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[60] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Transactions on Information Theory*, vol. 49, no. 11, p. 3117, 2003.

[61] A. Ozgur, O. Lévêque, and D. Tse, "Hierarchical Cooperation Achieves Linear Capacity Scaling in Ad Hoc Networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3549–3572, 2007.

[62] M. Franceschetti, M. D. Migliore, and P. Minero, "The Capacity of Wireless Networks: Information-Theoretic and Physical Limits," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3413–3424, Aug. 2009.

[63] A. Ozgur, R. Johari, D. N. C. Tse, and O. Lévêque, "Information-Theoretic Operating Regimes of Large Wireless Networks," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 427–437, Jan. 2009.

[64] S. S.-n. Hong and G. Caire, "Demystifying the Scaling Laws of Dense Wireless Networks: No Linear Scaling in Practice," in *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[65] N. Lu and X. S. Shen, "Scaling Laws for Throughput Capacity and Delay in Wireless Networks — A Survey," *IEEE Communications Surveys & Tutorials*, pp. 1–16, 2013.

[66] U. C. Kozat and L. Tassiulas, "Throughput Capacity of Random Ad Hoc Networks with Infrastructure Support," in *International conference on Mobile computing and networking*, 2003.

[67] W.-y. Shin, S.-W. Jeon, N. Devroye, M. H. Vu, S.-y. Chung, Y. H. Lee, and V. Tarokh, "Improved Capacity Scaling in Wireless Networks With Infrastructure," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5088–5102, Aug. 2011.

[68] C. Yeong, W.-y. Shin, and C. Jeong, "Ad Hoc Networking With Cost-Effective Infrastructure: Generalized Capacity Scaling," in *IEEE International Symposium on Information Theory (ISIT)*, no. July, 2014, pp. 1–27.

[69] L. Zeger and M. Médard, "On Scalability of Wireless Networks: A Practical Primer for Large Scale Cooperation," *arXiv preprint arXiv:1402.1761*, pp. 1–7, 2014. [Online]. Available: http://arxiv.org/abs/1402.1761

[70] H. H. S. Dhillon and G. Caire, "Information Theoretic Upper Bound on the Capacity of Wireless Backhaul Networks," in *IEEE International Symposium on Information Theory (ISIT)*, Jun. 2014.

[71] H. S. Dhillon and G. Caire, "Scalability of Line-of-Sight Massive MIMO Mesh Networks for Wireless Backhaul," in *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[72] J. Yoon, W.-y. Shin, and S.-W. Jeon, "Elastic Routing in Wireless Networks With Directional Antennas," in *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[73] R. Negi and A. Rajeswaran, "Capacity of Power Constrained Ad-Hoc Networks," in *IEEE INFOCOM*, 2004.

[74] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. Jun., pp. 1164–1179, Apr. 2013.

[75] M. Médard and R. G. Gallager, "Bandwidth Scaling for Fading Multipath Channels," *IEEE Transactions on Information Theory*, vol. 48, no. Apr., pp. 840–852, 2002.

[76] A. Lozano and D. Porrat, “Non-Peaky Signals in Wideband Fading Channels: Achievable Bit Rates and Optimal Bandwidth.” *IEEE Transactions on Wireless Communications*, vol. 11, pp. 246–257, 2012.

[77] M. Medard, “On Approaching Wideband Capacity Using Multitone FSK,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 9, pp. 1830–1838, Sep. 2005.

[78] S. Ray, M. Medard, and L. Zheng, “On Noncoherent MIMO Channels in the Wideband Regime: Capacity and Reliability,” *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 1983–2009, Jun. 2007.

[79] I. E. Telatar and D. N. Tse, “Capacity and Mutual Information of Wideband Multipath Fading Channels,” *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1384–1400, Jul. 2000.

[80] S. Verdú, “Spectral Efficiency in the Wideband Regime,” *IEEE Transactions on Information Theory*, vol. 48, no. Jun., pp. 1319–1343, 2002.

[81] L. Zheng, D. N. C. Tse, and M. Medard, “Channel Coherence in the Low-SNR Regime,” *IEEE Transactions on Information Theory*, vol. 53, no. Mar., pp. 976–997, 2007.

[82] O. Bazan and M. Jaseemuddin, “On the Design of Opportunistic MAC Protocols for Multihop Wireless Networks with Beamforming Antennas,” *IEEE Transactions on Mobile Computing*, vol. 14, no. 2, pp. 216–239, Mar. 2012.

[83] F. P. Kelly, “Charging and Rate Control for Elastic Rraffic,” *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[84] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, “Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability,” *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[85] L. Tassiulas, “Linear Complexity Algorithms for Maximum Throughput in Radio Networks and Input Queued Switches,” in *IEEE INFOCOM*, vol. 2, 1998.

[86] A. Eryilmaz, A. Ozdaglar, and E. Modiano, “Polynomial Complexity Algorithms for Full Utilization of Multi-Hop Wireless Networks,” in *IEEE INFOCOM*, 2007.

[87] A. Eryilmaz, A. Ozdaglar, D. Shah, and E. Modiano, “Distributed Cross-Layer Algorithms for the Optimal Control of Multihop Wireless Networks,” *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, pp. 638–651, 2010.

[88] Y. Yi, A. Proutière, and M. Chiang, "Complexity in Wireless Scheduling: Impact and Tradeoffs," in *ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2008.

[89] A. Zhou, M. Liu, Z. Li, and E. Dutkiewicz, "Cross-Layer Design for Proportional Delay Differentiation and Network Utility Maximization in Multi-Hop Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1446–1455, Apr. 2012.

[90] L. Lin, X. Lin, and N. B. Shroff, "Low-Complexity and Distributed Energy Minimization in Multi-hop Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, pp. 501–514, 2010.

[91] M. J. Neely, E. Modiano, and C.-p. Li, "Fairness and Optimal Stochastic Control for Heterogeneous Networks," in *IEEE INFOCOM*, vol. 3, no. MARCH, 2005.

[92] H.-W. Lee, E. Modiano, and L. B. Le, "Distributed Throughput Maximization in Wireless Networks via Random Power Allocation," *IEEE Transactions on Mobile Computing*, vol. 11, no. 4, 2012.

[93] H. Ju, B. Liang, J. Li, and X. Yang, "Dynamic Power Allocation for Throughput Utility Maximization in Interference-Limited Networks," *IEEE Wireless Communications Letters*, vol. 2, no. 1, pp. 22–25, 2013.

[94] G. D. Celik and E. Modiano, "Scheduling in Networks with Time-Varying Channels and Reconfiguration Delay," in *IEEE INFOCOM*, 2012.

[95] N. Shroff and R. Srikant, "A Tutorial on Cross-Layer Optimization in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[96] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems.* Morgan & Claypool, 2010.

[97] J. Kim and A. F. Molisch, "Quality-Aware Millimeter-Wave Device-to-Device Multi-Hop Routing for 5G Cellular Networks," in *IEEE International Conference on Communications (ICC)*, 2014.

[98] R. Mudumbai, S. K. Singh, and U. Madhow, "Medium Access Control for 60 GHz Outdoor Mesh Networks with Highly Directional Links," in *IEEE INFOCOM*, Apr. 2009.

[99] A. Sendonaris, E. Erkip, and B. Aazhang, "User Cooperation Diversity—Part I: System Description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[100] P. H. Pathak and R. Dutta, “A Survey of Network Design Problems and Joint Design Approaches in Wireless Mesh Networks,” *IEEE Communications Surveys & Tutorials*, pp. 1–33, 2010.

[101] A. Sendonaris, E. Erkip, and B. Aazhang, “User Cooperation Diversity—Part II: Implementation Aspects and Performance Analysis,” *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.

[102] P. Liu, Z. Tao, Z. Lin, E. Erkip, and S. Panwar, “Cooperative Wireless Communications: A Cross-Layer Approach,” *IEEE Wireless Communications*, vol. 13, no. August, pp. 84–92, 2006.

[103] H. Shan, W. Zhuang, and Z. Wang, “Distributed Cooperative MAC for Multihop Wireless Networks,” *IEEE Communications Magazine*, vol. 47, no. 2, pp. 126–133, 2009.

[104] C. Cetinkaya and F. Orsun, “Cooperative Medium Access Protocol for Dense Wireless Networks,” in *The Third Annual Mediterranean Ad Hoc Networking Workshop (Med Hoc Net)*, 2004.

[105] A. Bachir, M. Dohler, T. Watteyne, and K. K. Leung, “MAC Essentials for Wireless Sensor Networks,” *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 222–248, 2010.

[106] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran, “Closing the Gap in the Capacity of Wireless Networks Via Percolation Theory,” *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 1009–1018, Mar. 2007.

[107] T. Korakis, Z. Tao, Y. Slutskiy, and S. Panwar, “A Cooperative MAC Protocol for Ad Hoc Wireless Networks,” Mitsubishi Electric Research Laboratories, Tech. Rep., Apr. 2007.

[108] J. N. Laneman, G. W. Wornell, and D. N. C. Tse, “An Efficient Protocol for Realizing Cooperative Diversity in Wireless Networks,” in *IEEE International Symposium on Information Theory (ISIT)*, 2001.

[109] T. M. Cover and J. a. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York, USA: John Wiley & Sons, Inc., 1991, vol. 6.

[110] K. Azarian, H. E. Gamal, and P. Schniter, “On the Achievable Diversity-Multiplexing Tradeoff in Half-Duplex Cooperative Channels,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4152–4172, 2005.

[111] K. Azarian, H. El Gamal, and P. Schniter, "Achievable Diversity-vs-Multiplexing Tradeoffs in Half-Duplex Cooperative Channels," in *IEEE Information Theory Workshop (ITW)*, Oct. 2004.

[112] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler, "Fading Relay Channels: Performance Limits and Space–Time Signal Design," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1099–1109, Aug. 2004.

[113] N. Prasad and M. K. Varanasi, "High Performance Static and Dynamic Cooperative Communication Protocols for the Half Duplex Fading Relay Channel," in *IEEE Global Telecommunications Conference (GLOBECOM)*, Jan. 2006.

[114] J. N. Laneman and G. W. Wornell, "Distributed Space-Time-Coded Protocols for Exploiting Cooperative Diversity in Wireless Networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, 2003.

[115] N. Prasad and M. K. Varanasi, "Diversity and Multiplexing Tradeoff Bounds for Cooperative Diversity Protocols," in *IEEE International Symposium on Information Theory (ISIT)*, 2005.

[116] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A Simple Cooperative Diversity Method Mased on Network Path Selection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[117] A. Stefanov and E. Erkip, "Cooperative Coding for Wireless Networks," *IEEE Transactions on Communications*, vol. 52, no. 9, pp. 1470–1476, 2004.

[118] T. E. Hunter and A. Nosratinia, "Diversity through coded cooperation," *IEEE Transactions on Wireless Communications*, vol. 5, no. 2, pp. 283–289, Feb. 2006.

[119] Y. Chen, S. Kishore, and J. Li, "Wireless Diversity Through Network Coding," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2006.

[120] L. Xiao, T. E. Fuja, J. Kliewer, D. J. C. Jr, and D. Costello, "A Network Coding Approach to Cooperative Diversity," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3714–3722, Oct. 2007.

[121] M. Xiao and M. Skoglund, "M-user Cooperative Wireless Communications Based on Nonbinary Network Codes," in *IEEE Information Theory Workshop (ITW)*, 2009.

[122] J. Rebelatto, B. Uchoa-Filho, Y. Li, and B. Vucetic, "Multiuser Cooperative Diversity Through Network Coding Based on Classical Coding Theory," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 916–926, 2012.

[123] E. G. Larsson and B. R. Vojcic, “Cooperative Transmit Diversity Based on Superposition Modulation,” *IEEE Communications Letters*, vol. 9, no. 9, pp. 778–780, Sep. 2005.

[124] T. Wang and G. G. B. Giannakis, “Complex Field Network Coding for Multiuser Cooperative Communications,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 3, pp. 561–571, Apr. 2008.

[125] N. Fawaz, D. Gesbert, and M. Debbah, “When Network Coding and Dirty Paper Coding Meet in a Cooperative Ad Hoc Network,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 5-2, pp. 1862–1867, May 2008.

[126] B. Sirkeci-Mergen and A. Scaglione, “Randomized space-time coding for distributed cooperative communication,” *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5003–5017, Oct. 2007.

[127] E. Koyuncu, Y. Jing, and H. Jafarkhani, “Distributed Beamforming in Wireless Relay Networks with Quantized Feedback,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1429–1439, 2008.

[128] N. Wu and H. Gharavi, “Asynchronous Cooperative MIMO Systems Using a Linear Dispersion Structure,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 2, pp. 779–787, Feb. 2010.

[129] P. A. Anghel, G. Leus, and M. Kavehl, “Multi-user Space-Time Coding in Cooperative Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.

[130] S. Chen, W. Wang, X. Zhang, and Z. Sun, “Performance Analysis of OSTBC Transmission in Amplify-and-Forward Cooperative Relay Networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 1, pp. 105–113, Jan. 2010.

[131] S. Yiu, R. Schober, and L. Lampe, “Distributed Space-Time Block Coding,” *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1195–1206, 2006.

[132] M. Sharp, A. Scaglione, and B. Sirkeci-Mergen, “Randomized Cooperation in Asynchronous Dispersive Links,” *IEEE Transactions on Communications*, vol. 57, no. 1, pp. 64–68, Jan. 2009.

[133] C. T. Chou and M. Ghosh, “Cooperative Communication MAC (CMAC)-a New MAC Protocol for Next Generation Wireless LANs,” in *International Conference on Wireless Networks, Communications and Mobile Computing (IWCMC)*, 2005.

[134] A. Azgin, Y. Altunbasak, and G. AlRegib, "Cooperative MAC and Routing Protocols for Wireless Ad Hoc Networks," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2006.

[135] H. Zhu and G. Cao, "rDCF: A Relay-Enabled Medium Access Control Protocol for Wireless Ad Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 9, pp. 1201–1214, Sep. 2006.

[136] G. Holland, N. Vaidya, and P. Bahl, "A Rate-adaptive MAC Protocol for Multihop Wireless Networks," in *International Conference on Mobile Computing and Networking (mobiCom)*, New York, New York, USA, 2001.

[137] Y. Chen, G. Yu, P. Qiu, and Z. Zhang, "Power-Aware Cooperative Relay Selection Strategies in Wireless Ad Hoc Networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2006.

[138] S. Moh, C. Yu, S.-M. M. Park, H.-N. N. Kim, and J. Park, "CD-MAC: Cooperative Diversity MAC for Robust Communication in Wireless Ad Hoc Networks," in *IEEE International Conference on Communications (ICC)*, Jun. 2007.

[139] H. Shan, P. Wang, W. Zhuang, and Z. Wang, "Cross-Layer Cooperative Triple Busy Tone Multiple Access for Wireless Networks," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[140] Z. J. Haas and J. Deng, "Dual Busy Tone Multiple Access (DBTMA)-A Multiple Access Control Scheme for Ad Hoc Networks," *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 975–985, 2002.

[141] E. Fasolo, A. Munari, F. Rossetto, and M. Zorzi, "Phoenix: A Hybrid Cooperative-Network Coding Protocol for Fast Failure Recovery in Ad Hoc Networks," in *IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Jun. 2008.

[142] P. Liu, Z. Tao, and S. Panwar, "A cooperative MAC protocol for wireless local area networks," in *IEEE International Conference on Communications (ICC)*, 2005.

[143] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. Panwar, "CoopMAC: A Cooperative MAC for Wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 340–354, Feb. 2007.

[144] F. Liu, T. Korakis, Z. Tao, and S. Panwar, "A MAC-PHY Cross-Layer Protocol for Ad Hoc Wireless Networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2008.

[145] F. Verde, T. Korakis, E. Erkip, and A. Scaglione, "On Avoiding Collisions and Promoting Cooperation: Catching Two Birds with One Stone," in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2008.

[146] P. Liu, Y. Liu, T. Korakis, A. Scaglione, E. Erkip, and S. S. Panwar, "Cooperative MAC for Rate Adaptive Randomized Distributed Space-Time Coding," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[147] P. Liu, C. Nie, E. Erkip, and S. Panwar, "Robust Cooperative Relaying in a Wireless LAN: Cross-Layer Design and Performance Analysis," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[148] F. Verde, T. Korakis, E. Erkip, and A. Scaglione, "A Simple Recruitment Scheme of Multiple Nodes for Cooperative MAC," *IEEE Transactions on Communications*, vol. 58, no. 9, pp. 2667–2682, Sep. 2010.

[149] C. Nie, P. Liu, T. Korakis, E. Erkip, and S. S. Panwar, "CoopMAX: A Cooperative MAC with Randomized Distributed Space-Time Coding for an IEEE 802.16 Network," in *IEEE International Conference on Communications (ICC)*, 2009.

[150] H. Adam, C. Bettstetter, and S. M. Senouci, "Multi-Hop-Aware Cooperative Relaying," in *IEEE 69th Vehicular Technology Conference (VTC Spring)*, Apr. 2009.

[151] G. Bocherer and R. Mathar, "On the Throughput/Bit-Cost Tradeoff in CSMA Based Cooperative Networks," in *International Conference on Source and Channel Coding (SCC)*, 2010.

[152] H. Gharavi, B. Hu, and N. Wu, "A Design Framework for High-Density Wireless Ad-Hoc Networks Achieving Cooperative Diversity," in *IEEE International Conference on Communications (ICC)*, May 2010.

[153] H.-Y. Y. Shen, H. Yang, B. Sikdar, and S. Kalyanaraman, "A Distributed System for Cooperative MIMO Transmissions," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[154] X. Bai, D. Liu, G. Yue, and H. Wu, "Joint Relay Selection and Power Allocation in Cooperative-Diversity System," in *IEEE International Conference on Communications and Mobile Computing (CMC)*, Apr. 2010.

[155] G. Jakllari, S. V. Krishnamurthy, M. Faloutsos, P. V. Krishnamurthy, and O. Ercetin, "A Framework for Distributed Spatio-Temporal Communications in Mobile Ad hoc Networks," in *IEEE INFOCOM*, 2006.

[156] P. E. Mogensen, T. Koivisto, K. I. Pedersen, I. Z. Kovács, B. Raaf, K. Pajukoski, and M. J. Rinne, "LTE-Advanced: The Path towards Gigabit/s in Wireless Mobile Communications," in *IEEE International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE).* IEEE, 2009.

[157] K. Loa, C.-C. Wu, S.-t. Sheu, Y. Yuan, M. Chion, D. Huo, and L. Xu, "IMT-Advanced Relay Standards [WiMAX/LTE Update]." *IEEE Communications Magazine*, vol. 48, no. 8, pp. 40–48, 2010.

[158] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description (3GPP TS 36.201 version 10.0.0 Release 10)," *ETSI TS 136 201*, vol. 0, no. 10.0.0, pp. 0–14, 2011.

[159] A. Karaer, O. Bulakci, S. Redana, and J. Hämäläinen, "Uplink Performance Optimization in Relay Enhanced LTE-Advanced Networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2009.

[160] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 8.8.0 Release 8)," *ETSI TS 136 213*, vol. 0, no. 8.8.8, pp. 1–79, 2009.

[161] X. Huang, F. Ulupinar, P. Agashe, D. Ho, and G. Bao, "LTE Relay Architecture and Its Upper Layer Solutions." in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[162] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Long Term Evolution (LTE) physical layer; General description (3GPP TS 36.201 version 8.3.0 Release 8)," *ETSI TR 136 201*, no. 8.3.0, pp. 1–15, 2009.

[163] H. J. Kushner and P. A. Whiting, "Convergence of Proportional-Fair Sharing Algorithms Under General Conditions," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.

[164] H.-Y. Wei and R. Gitlin, "Incentive Scheduling for Cooperative Relay in WWAN/WLAN Two-Hop-Relay Network," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2005.

[165] "IEEE Standard for Information Technology–Telecommunications and information exchange between systems Wireless Regional Area Networks (WRAN)–Specific requirements Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specif," *IEEE Std 802.22-2011*, pp. 1–680, 2011.

[166] "IEEE Draft Standard for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specif," *IEEE Draft P802.11-REVmb/D9.0, May 2011 (Revision of IEEE Std 802.11-2007, as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, IEEE Std 802.11w-2009 and IEEE Std 802.11n-2009)*, pp. 1–2742, 2011.

[167] A. Papoulis, S. U. Pillai, and S. Unnikrishna, *Probability, random variables, and stochastic processes.* McGraw-Hill New York, 2002, vol. 73660116.

[168] "IEEE Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation with the Electric Power System (EPS), End-Use Applications, and Loads," *IEEE Std 2030*, no. September, pp. 1–126, Feb. 2011.

[169] V. R. Cadambe and S. A. Jafar, "Interference Alignment and Degrees of Freedom of the K-user Interference Channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[170] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, 2014.

[171] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge university press, 2005.

[172] A. Lozano, "Interplay of Spectral Efficiency, Power and Doppler Spectrum for Reference-Signal-Assisted Wireless Communication," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5020–5029, Dec. 2008.

[173] N. Jindal and A. Lozano, "A Unified Treatment of Optimum Pilot Overhead in Multipath Fading Channels," *IEEE Transactions on Communications*, vol. 58, no. Oct., pp. 2939–2948, Oct. 2010.

[174] P. Cheng, Z. Chen, Y. Rui, Y. J. Guo, L. Gui, M. Tao, and Q. T. Zhang, "Channel Estimation for OFDM Systems over Doubly Selective Channels: A Distributed Compressive Sensing Based Approach," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4173–4185, Oct. 2013.

[175] N. Michailow, M. Matthe, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3045–3061, Sep. 2014.

[176] A. Manolakos, M. Chowdhury, and A. J. Goldsmith, "Constellation Design in Noncoherent Massive SIMO Systems," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2014.

[177] ——, "Constellation Design in an Energy-based Noncoherent Massive SIMO System," *Submitted to IEEE Transactions on Wireless Communications*, 2014.

[178] T. L. Marzetta, G. Caire, M. Debbah, I. Chih-Lin, and S. K. Mohammed, "Special Issue on Massive MIMO," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 333–337, 2013.

[179] M. Chowdhury, A. Manolakos, and A. J. Goldsmith, "Design and Performance of Non coherent Massive SIMO Systems," in *IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2014.

[180] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, 2014.

[181] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[182] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, and S. S. Panwar, "Directional Cell Search for Millimeter Wave Cellular Systems," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Apr. 2014.

[183] M. R. Akdeniz, Y. Liu, S. Rangan, and E. Erkip, "Millimeter Wave Picocellular System Evaluation for Urban Deployments," in *IEEE Global Telecommunications Conference (GLOBECOM)*, Apr. 2013.

[184] S. Sun, T. Rappaport, R. Heath, A. Nix, and S. Rangan, "Mimo for Millimeter-Wave Wireless Communications: Beamforming, Spatial Multiplexing, or Both?" *IEEE Communications Magazine*, no. December, pp. 110–121, 2014.

[185] S. Singh, R. Mudumbai, and U. Madhow, "Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1513–1527, 2011.

[186] T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, "38 GHz and 60 GHz Angle-Dependent Propagation for Cellular & Peer-to-Peer Wireless Communications," in *IEEE International Conference on Communications (ICC)*, 2012.