DESPLEGAMENT DELS CIRCUITS DEDICATS DE 1 I 10 GBPS ENTRE EL PIC I EL CERN

Memòria del Projecte Fi de Carrera
d'Enginyeria en Informàtica
realitzat per
Gerard Bernabeu
i dirigit per
Miquel Àngel Senar (DACSO - UAB)
Andreu Pacheco (PIC)

Bellaterra, 14 de juny de 2007

Taula de continguts

1 Capítol 1 - Introducció....4
1.1 Marc del projecte....4
1.2 Descripció del projecte....6
1.2.1 Propòsit, abast i objectius....6
1.2.2 Assumpcions, Restriccions i Riscs....8
1.2.3 Recursos....8
1.2.4 Metodologia del Projecte....8
1.2.5 Característiques dels entregables....10
1.3 Planificació....11
1.4 Organització de la memòria....12
2 Capítol 2 – Fonaments teòrics....13
2.1 El model OSI i TCP/IP....13
2.1.1 El model OSI....13
2.1.2 El model TCP/IP....14
2.2 Capa física....15
2.2.1 UTP....15
2.2.2 Fibra òptica....16
2.3 Capa d'enllaç de dades....17
2.3.1 Ethernet....17
2.3.2 Jumboframes....18
2.3.3 VLANs....18
2.4 Capa de xarxa....19
2.4.1 IP....20
2.4.2 ICMP....22
2.4.3 Encaminament IP....22
2.4.4 Sistemes Autònoms (AS)....23
2.4.5 BGP....24
2.5 Capa de transport....25
2.5.1 Ports....26
2.5.2 Protocol de Datagrames d'Usuari (UDP)....26
2.5.3 Protocol de Control la Transmissió (TCP)....27
2.6 Capa d'aplicació: eines de monitorització i mesura de rendiment....31
2.6.1 Eines de monitorització....31
2.6.2 Eines de mesura de rendiment....33
2.7 Altres tecnologies utilitzades....34
2.7.1 HSRP....34
2.7.2 Spanning-tree protocol....35
2.7.3 Etherchannel/Link Aggregation (802.3ad)....36
2.7.4 Bonding....37
2.7.5 Policy Route Map....37
3 Capítol 3 – Desplegament circuit dedicat a 1 Gbps....39
3.1 Estudi de solucions per a la integració dels servidors del PIC i la xarxa LHC-OPN sobre el

circuit dedicat d'1 Gbps..... 39
3.1.1 Descripció de la situació inicial..... 39
3.1.2 Especificacions del sistema objectiu..... 42
3.1.3 Possibles solucions..... 44
3.1.4 Comparativa de solucions..... 52
3.2 Pla d'implementació per a la integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps..... 53
3.2.1 Descripció de la solució escollida..... 53
3.2.2 Aspectes genèrics de la configuració..... 53
3.2.3 Detalls d'implementació: Opcions per a la solució “Dues IPs”..... 54
3.2.4 Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN.....61
3.2.5 Maqueta de la solució amb dues IPs..... 62
3.3 Informe de la execució del pla d'implementació sobre el circuit dedicat de 1 Gbps.....64
3.3.1 Modificacions respecte la planificació inicial..... 64
3.3.2 Execució del pla d'implementació..... 65
3.3.3 Simulacre de posta en marxa..... 66
3.3.4 Incidències i resolució de les mateixes..... 67
3.3.5 Rutes PIC-CERN..... 69
3.3.6 Proves de la connexió sobre el circuit dedicat d'1 Gbps..... 71
3.4 Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps.....77
3.4.1 Proves empíriques de rendiment i fiabilitat de la connexió..... 77
3.4.2 Anàlisi de la càrrega en els encaminadors i commutadors locals..... ..78
3.4.3 Anàlisi del trànsit de la nova xarxa..... 78
3.4.4 Anàlisi estadístic del rendiment de la connexió a mig termini..... 79
4 Capítol 4 – Desplegament circuit dedicat a 10 Gbps.....80
4.1 Estudi de solucions per a l'integració dels serveis del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps..... 80
4.1.1 Descripció de la situació inicial..... 80
4.1.2 Especificacions del sistema objectiu..... 83
4.1.3 Possibles solucions per al desplegament de la connexió i del circuit dedicat de 10 Gbps 84
4.1.4 Possibles solucions LAN: integració dels sistemes de gestió de dades del PIC amb la xarxa LHC-OPN..... 86
4.2 Pla d'implementació per a la integració dels serveis del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps..... 93
4.2.1 Descripció de la solució escollida..... 93
4.2.2 Sistema de certificació per al circuit dedicat de 10 Gbps..... ...94
4.2.3 Detalls d'implementació del receptor de dades T0-T1 de dCache..... 95
4.2.4 Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN.....99
4.3 Informe de la execució del pla d'implementació sobre el circuit de dedicat de 10 Gbps.....101
4.3.1 Modificacions respecte la planificació inicial..... 101
4.3.2 Execució del pla d'implementació..... 102
4.3.3 Incidències i resolució de les mateixes..... 104
4.3.4 Rutes PIC-CERN..... 106

4.3.5 Proves de la connexió sobre el circuit dedicat de 10 Gbps.....107
4.4 Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps.....113
4.4.1 Proves de rendiment i fiabilitat de la connexió.....113
4.4.2 Anàlisi de la càrrega en l'encaminador local.....115
4.4.3 Anàlisi del trànsit de la nova xarxa.....115
4.4.4 Anàlisi estadístic del rendiment de la connexió a mig termini.....115
4.5 Proposta per al desplegament d'una connexió redundant.....116
4.5.1 Descripció de la situació inicial.....116
4.5.2 Objectius.....116
4.5.3 Especificacions del sistema i de les aplicacions.....116
4.5.4 Viabilitat tècnica.....116
4.5.5 Viabilitat operativa.....117
4.5.6 Viabilitat econòmica.....117
4.5.7 Alternatives.....117
5 Capítol 5 – Conclusions.....118
6 Annex A: Planificació detallada del projecte.....120
7 Annex B: Script per a l'addició de rutes als servidors LHC-OPN.....122
8 Annex C: Proves de rendiment de la LAN sobre la maqueta solució amb dues IPs.....123
8.1 Proves de rendiment LAN unidireccional entre servidors del mateix switch.....124
8.2 Proves de rendiment LAN bidireccional entre servidors.....125
8.2.1 Proves de referència amb cable creuat.....125
8.2.2 Proves utilitzant switch Dell PowerConnect 5324.....127
8.3 Prova amb múltiples transmissions: lhcopn01->lhcopn02->lhcopn03->lhcopn01.....130
8.4 Conclusions de les proves de rendiment sobre la LAN.....130
9 Annex D: Detalls de la configuració de la maqueta local.....132
10 Annex E: Path MTU Black Hole Detection and Recovery.....137
11 Annex F: Sensor de Nagios i procediment d'alarma.....139
11.1 Procediment d'actuació.....139
11.2 Sensor de Nagios/Ganglia.....141
12 Annex G: pla d'actuació per al desplegament del circuit dedicat de 10 Gbps.....143
13 Annex H: Detalls d'implementació i proves del sistema de certificació pel circuit dedicat de 10 Gbps.....145
13.1 Configuració i us de vxargs.....145
13.2 Proves del sistema de certificació pel circuit dedicat de 10 Gbps.....146
13.3 Configuració del sistema de certificació.....149
14 Annex I: sensor de Nagios i procediment per a la monitorització de la connectivitat sobre la xarxa LHC-OPN.....151
15 Annex J: diagnosis del problema de connectivitat PIC-CERN sobre el circuit dedicat de 10 Gbps durant la certificació.....156
15.1 Bateria de proves per a la diagnosi.....156
Bibliografia.....160

1 Capítol 1 - Introducció

Aquest projecte consisteix en realitzar el disseny i desplegament d'una connexió entre el *Port d'Informació Científica* (PIC) i el *Consell Europeu per a la Recerca Nuclear* (CERN) sobre un circuit dedicat amb una velocitat de transferència de 10 Gbps. El projecte es du a terme en col·laboració amb l'Anella Científica, RedIRIS i el CERN.

Inicialment el desplegament de la connexió es realitzarà sobre un circuit dedicat de 1 Gbps i, un cop estigui disponible, sobre el circuit de 10 Gbps. El desenvolupament del projecte també implica el disseny dels plans d'actuació que han de permetre la integració de les noves connexions dins la xarxa i els serveis del PIC.

El PIC és un dels 12 centres *Tier-1* de processament de dades del *Large Hadron Collider Computing Grid Project* (LCG). A partir de 2008, el CERN enviarà una copia de les dades que adquireixi a través dels experiments del *Large Hadron Collider* (LHC) via la connexió sobre el circuit dedicat de 10 Gbps. Aquesta connexió s'ha d'implementar de forma transparent a través de diversos proveïdors, resultant en un enllaç directe entre el PIC i el CERN a través d'una lambda de la xarxa europea GÉANT.

En aquest primer capítol d'introducció al PFC es descriu el marc de desenvolupament del projecte i es defineixen els objectius i les característiques del mateix. Finalment hi ha una breu explicació de l'organització de la memòria.

1.1 Marc del projecte

Aquest Projecte de Final de Carrera (PFC) s'ha desenvolupat dins el PIC (seu a la figura *1.1*.1), un centre fundat el Juny del 2003, ubicat dins el campus de la UAB i finançat pel *Departament d'Educació i Universitats* (DeiU), el *Centro de Investigaciones Energéticas, Medioambientales y Tecnológica* (CIEMAT), la *Universitat Autònoma de Barcelona* (UAB) i l'*Institut de Física d'Altes Energies* (IFAE) amb l'objectiu de donar suport a comunitats científiques que treballen en projectes que necessiten recursos massius de processament i emmagatzematge de dades per a la computació distribuïda. Així el PIC es converteix en un centre d'excel·lència que permet a Espanya participar en projectes europeus que visen el desenvolupament de la infraestructura GRID internacional per a la ciència i la tecnologia.

Figura *1.1*.1: Edifici dels Serveis Informàtics al Campus de la UAB on es troba el PIC

El màxim òrgan de decisió de la política cientificotècnica del PIC és la Comissió Gestora, els seus membres són nomenats per les institucions fundadores del PIC. La presidència és rotativa i correspon en torns d'un any al CIEMAT i al DeiU.

El PIC és un centre d'R+D associat a la UAB i com a tal es troba dins de l'*Anella Científica*, xarxa d'àmbit autonòmic[1] gestionada pel CESCA, i de RedIRIS, la xarxa espanyola d'I+D.

Internament el PIC està estructurat en 5 equips de treball, cadascun d'ells amb un coordinador, i dirigit pel Dr. Manuel Delfino, director del centre. Per a la realització del PFC s'ha creat un nou equip de treball, format per dos grups:

- Grup de disseny: Codirigit pel director del PFC, Dr. Miquel Àngel Senar, i el tutor extern, Dr. Andreu Pacheco. En aquest grup és on s'analitza el problema i es dissenyen i discuteixen les diferents solucions plantejades per l'alumne del PFC, Gerard Bernabeu.

- Grup d'operacions: Coordinat pel Dr. Manuel Delfino, director del PIC i coordinador de l'Equip A. Dins aquest grup és on l'alumne del PFC s'encarrega de dissenyar les possibles solucions. La materialització de les solucions també és duta a terme per l'alumne del PFC, amb participacions puntuals dels equips de treball del PIC i, en casos on es requereix modificar la configuració de l'encaminador del PIC, amb l'administrador de la xarxa[2] del PIC.

En la figura *1.1*.2 es pot observar l'estructura organitzativa del PIC, així com les relacions establertes amb les diferents institucions per tal de poder dur a terme el desenvolupament del projecte.

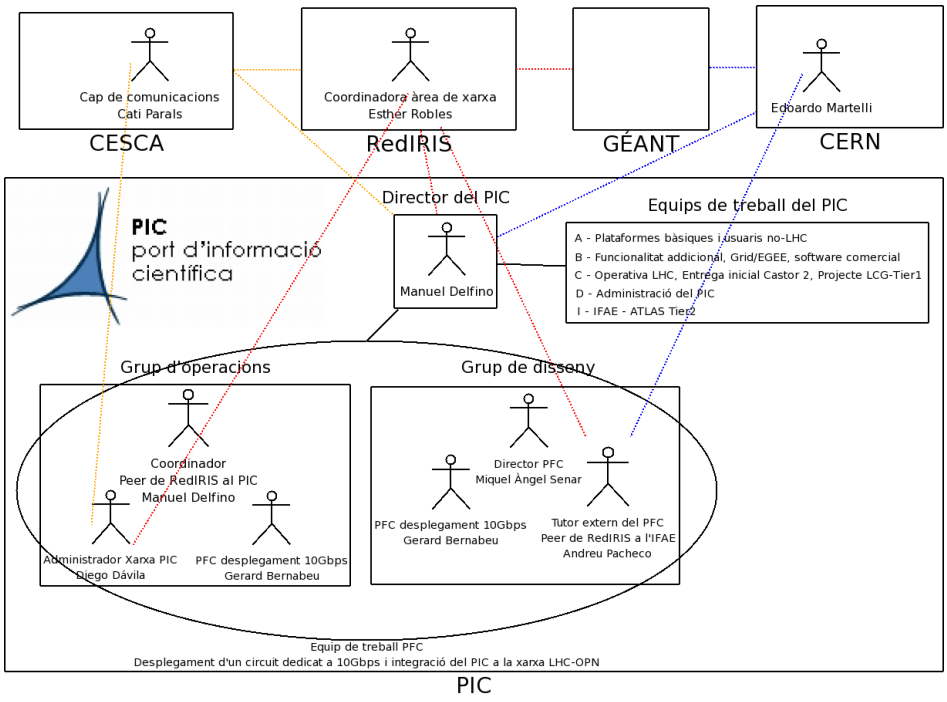


Figura *1.1*.2: Organigrama per al desenvolupament del PFC on es poden observar les vies formals de comunicació

1 *L'Anella Científica* és una xarxa de comunicacions creada per la *Fundació Catalana per a la Recerca i la Innovació* i gestionada pel CESCA que connecta universitats i centres d'investigació a Catalunya i els enllaça amb la xarxa nacional de RedIRIS.

2 L'adminsitrador de la xarxa (Diego Dávila) és el responsable exclusiu de la configuració de l'encaminador (*router*) Cisco 6509.

Les relacions que van a una “caixa” són aquelles que es fan amb la institució indicada, les altres relacions són entre persones, la majoria a nivell tècnic.

La realització al PIC d'aquest projecte ha implicat la redacció i presentació d'un seguit d'entregables on queda reflexat tot el procés d'anàlisi, disseny i implementació de forma breu i concisa, concentrant tota la informació útil a nivell executiu en el cos de l'entregable i descrivint els detalls tècnics en els annexes. El detall dels entregables presentats es pot trobar a la secció 1.2.4.

1.2 Descripció del projecte

1.2.1 Propòsit, abast i objectius

El propòsit d'aquest projecte és realitzar el desplegament inicial d'un circuit dedicat de 10 Gbps entre el PIC i el CERN, integrant-lo dins la xarxa LHC-OPN[3] i demostrant-ne la usabilitat empíricament.

Es pretén satisfer la necessitat de disposar d'una connexió privada, d'altes prestacions i disponibilitat entre el PIC i el CERN. Són fites d'aquest projecte assolir un alt nivell de rendiment i disponibilitat en la connexió a la xarxa LHC-OPN, proveint una metodologia per a l'adhesió a la xarxa LHC-OPN als servidors del PIC. Per tal d'aprofitar al màxim les noves característiques de la xarxa en els serveis relacionats amb la LHC-OPN, o en el GRID en si mateix, podria ser necessària la realització de plans addicionals.

Aquest projecte implica serveis del PIC relacionats amb la xarxa LHC-OPN com ara els sistemes de gestió de fitxers de dades Castor[4] i dCache[5]. La integració d'aquest projecte amb els serveis del PIC es durà a terme pels corresponents grups de treball del PIC mitjançant la metodologia provista pel projecte, prèviament consensuada.

Es realitzarà la connexió inicial sobre un circuit dedicat ja existent d'1 Gbps, migrant posteriorment a un nou circuit dedicat de 10 Gbps. Per a incrementar el nivell de disponibilitat del sistema final, es realitzarà un estudi de viabilitat per a la implementació d'una línia de backup per al circuit dedicat.

El projecte implica un procés de negociació amb diversos actors, que es realitzarà amb la col·laboració del PIC per tal d'assolir els objectius, els quals són:

1. Integrar el PIC a la LHC-OPN sobre el circuit dedicat de 1 Gbps
2. Integrar el PIC a la LHC-OPN sobre el circuit dedicat de 10 Gbps
3. Realitzar un estudi de viabilitat per a la línia de backup

3 La *Large Hadron Collider – Optical Private Network* és una xarxa dissenyada específicament per a donar suport als experiments de l'LHC.

4 CASTOR (CERN Advanced STORage manager) és un sistema d'emmagatzemament jeràrquic desenvolupat al CERN i utilitzat per a l'emmagatzemament de dades en disc i/o cinta.

5 dCache és un sistema d'emmagatzemament jeràrquic que sorgeix d'un esforç conjunt entre DESY (www.desy.de), FERMILAB (www.fnal.gov) i l'aportació d'altres centres implicats en el projecte LHC. Juntament amb Enstore, o una aplicació similar, és capaç de gestionar dades en robots de cintes.

El projecte es considerarà finalitzat un cop aprovats els diferents entregables per part de la direcció del PIC, quedant la xarxa amb un connexionat similar al mostrat en les figures *1.2.*1 i *1.2.*2

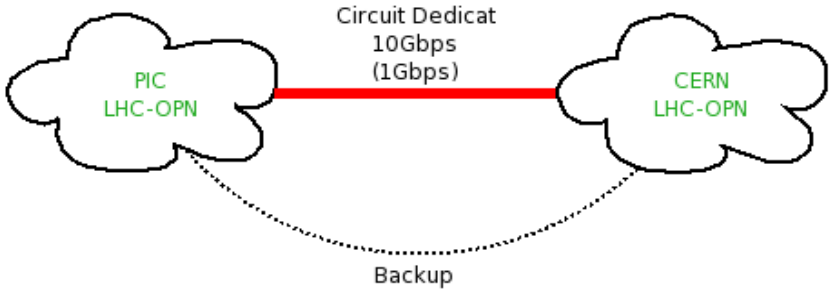


Figura *1.2.*1: Visió de la connexió a nivell 3

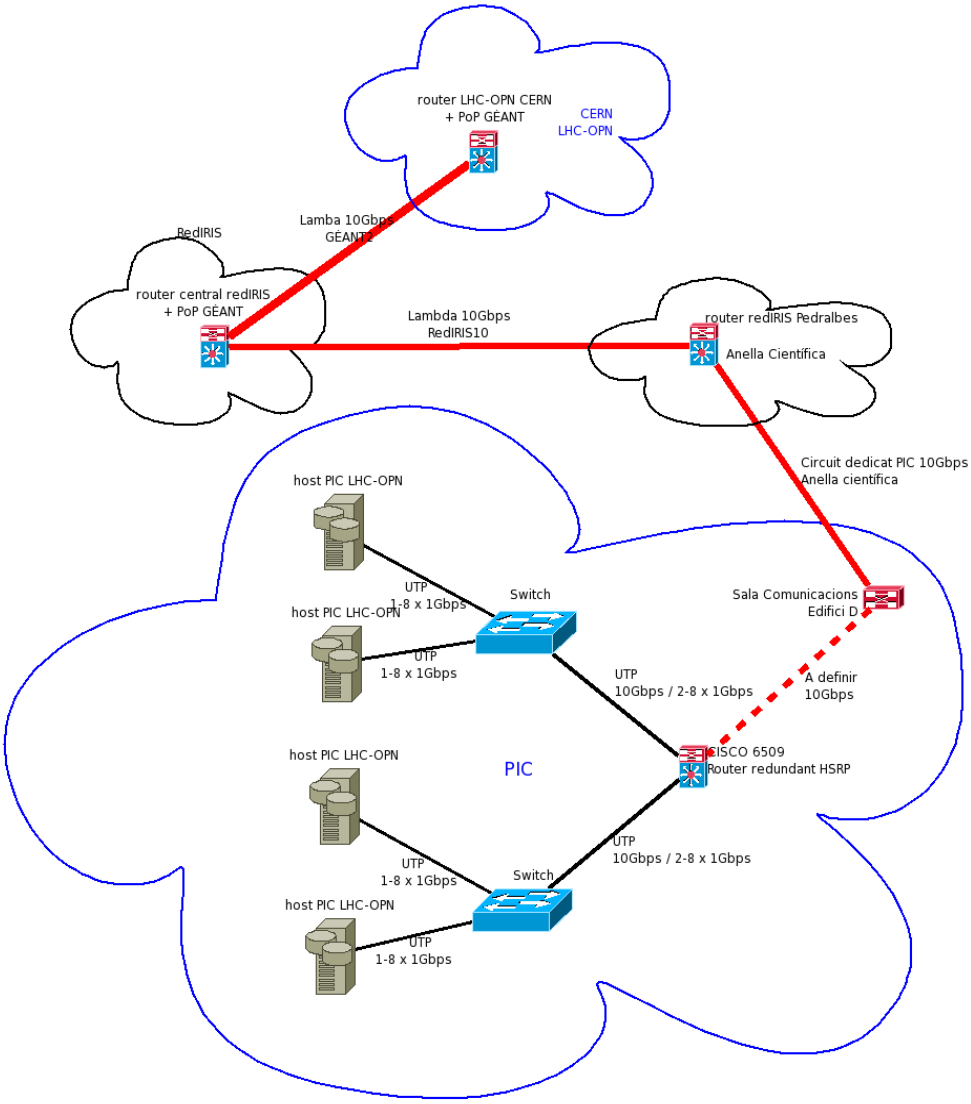


Figura *1.2.*2: Visió simplificada de la connexió a nivell 1 on es poden observar les cinc organitzacions implicades en el projecte: el PIC, l'Anella Científica, RedIRIS, GÉANT i el CERN.

1.2.2 Assumpcions, Restriccions i Riscs

S'assumeix que serà possible definir un mecanisme de decisió quan més d'una solució sigui possible i hi hagi discrepàncies sobre quina solució s'adopta. També s'assumeix la bona voluntat i col·laboració per part del CERN, GÉANT, RedIRIS, l'Anella Científica i el propi PIC.

Aquest projecte depen directament del bon funcionament de la LHC-OPN sobre GÉANT, del desplegament del projecte RedIris10 i de la instal·lació fins al PIC de la línia de 10 Gbps per part de l'Anella Científica, que crea la interconnexió física a 10 Gbps entre el PIC i RedIRIS.

Degut a les restriccions temporals del projecte, les dificultats en els processos de decisió i /o la no disponibilitat dels recursos necessaris en podria retrassar considerablement el desenvolupament. La connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps ha de ser operativa al llarg de febrer 2007, sobre el circuit dedicat de 10 Gbps ho ha de ser al llarg de l'abril 2007. Tanmateix l'estudi de viabilitat per a la línia de backup, i el projecte en si mateix, ha d'estar resolt abans del juny del 2007

Les tecnologies de xarxa a utilitzar seran els diferents estàndards publicats per l'IEEE 802[1] que estiguin disponibles en els diferents elements de la xarxa.

Per tal de poder assolir una velocitat de ~10 Gbps en les transferències CERN<->PIC, sota certes topologies de xarxa pot ser necessari dissenyar algun sistema d'agregació de línies en la LAN, la implementació del qual pot ser crítica i podria retrassar el desenvolupament del projecte.

1.2.3 Recursos

Els recursos necessaris per a la realització del projecte són:

- Encaminador/Switch redundant amb suficients ports GigabitEthernet i TenGigabitEthernet
- Recursos necessaris per a garantir una alta disponibilitat i un rendiment de la xarxa interna (LAN) en concordança amb la connexió LHC-OPN disponible.
- Ordinadors amb múltiples targes de xarxa GigabitEthernet o TenGigabitEthernet.
- Especificacions i disponibilitat de la línia d'1 Gbps amb connectivitat al CERN
- Especificacions i disponibilitat de la línia de 10 Gbps amb connectivitat al CERN
- Metodologia per al testeig de les línies des del CERN
- Metodologia per a l'adhesió a la LHC-OPN

1.2.4 Metodologia del Projecte

Previ a l'inici del qualsevol de les fases es va realitzar un procés d'anàlisi i estudi del marc de treball, indispensable per a comprendre el tramat de relacions entre les diferents entitats i les necessitats existents, així com per a l'adaptació al funcionament del centre.

Un cop finalitzat el procés inicial d'aprenentatge, el projecte s'ha dividit dues fases segons el desplegament sobre la connexió a 1 i 10 Gbps. En ambdues fases la metodologia seguida,

classificada per etapes, ha estat molt similar:

- en primer lloc s'ha realitzat una etapa d'anàlisi i recollida d'especificacions que, un cop finalitzada, ha donat pas a la generació d'una sèrie de possibles solucions.

- La segona etapa s'inicia amb la presentació formal i el posterior debat de les solucions plantejades, així com els seus avantatges i inconvenients. L'etapa finalitza quan s'arriba a una solució consensuada amb la directiva i els grups de treball del PIC, tenint en compte la posició i possibles repercussions amb els altres ens implicats en el desplegament de la connexió (l'Anella Científica, el CERN, etc.).

- Un cop obtinguda una solució i un pla de desplegament consensuats s'inicia l'etapa d'implementació i prova. En ambdues fases del PFC aquesta etapa ha inclòs la creació d'una maqueta capaç de demostrar la funcionalitat de la solució que s'ha decidit implementar.

- Cada fase finalitza amb el procés de certificació, on es demostra i documenta el correcte comportament de la connexió.

La metodologia anteriorment descrita de Anàlisi->Consens->Implementació->Certificació s'ha gestionat mitjançant la realització d'una sèrie d'entregables. Els entregables principals, en format d'informe, s'han entregat per mitjans electrònics i/o en paper a la directiva del PIC segons les següents dates:

Fase 1

- Finals de gener del 2007
  1. Estudi de solucions per a l'integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps – *Etapa d'anàlisi*
  2. Pla d'implementació de la solució consensuada sobre el circuit dedicat de 1 Gbps – *Etapa de decisió*
- A mitjans de febrer del 2007
  3. Informe de la execució del pla d'implementació sobre el circuit dedicat de 1 Gbps – *Etapa d'implementació*
  4. Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps – *Etapa de certificació*

Fase 2

- Abans de l'abril del 2007
  5. Estudi de solucions per a l'integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps – *Etapa d'anàlisi*
  6. Pla d'implementació de la solució consensuada sobre el circuit dedicat de 10 Gbps – *Etapa de decisió*
- Abans de la segona setmana de maig del 2007
  7. Informes d'acceptació dels diferents segments del circuit dedicat de 10 Gbps – *Etapa*

d'implementació

8. Informe de la execució del pla d'implementació sobre el circuit dedicat de 10 Gbps – *Etapa d'implementació*
9. Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps – *Etapa de certificació*

- Abans del final de Maig del 2007

10. Estudi de viabilitat sobre la connexió redundant per a la xarxa LHC-OPN – *Etapa d'anàlisi*

1.2.5 Característiques dels entregables

En els estudis de solucions per a l'integració dels servidors del PIC a la xarxa LHC-OPN (entregables 1 i 5) s'hi inclourà:

- Descripció de la situació inicial
- Especificacions del sistema i de les aplicacions
- Viabilitat tècnica, operativa i econòmica per a les diferents solucions alternatives
- Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN per a les diferents solucions alternatives

El pla d'implementació de la solució consensuada (entregables 2 i 6) inclourà:

- Descripció de la situació inicial
- Especificacions del sistema i de les aplicacions
- Recursos necessaris
- Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN

Els informes de la execució del pla d'implementació (entregables 3 i 8) inclouran:

- Descripció de la situació
- Incidències i resolució de les mateixes

La metodologia per a les certificacions (entregables 4 i 9) implicarà:

- Proves empíriques respecte al rendiment (throughput) i la fiabilitat (tassa d'errors) de les connexions
- Anàlisi de la càrrega en els encaminadors i commutadors locals pertanyents a la nova xarxa
- Anàlisi del trànsit de la nova xarxa
- Anàlisi estadístic del rendiment i la fiabilitat de la connexió a mig termini

Els informes d'acceptació dels diferents segments del circuit dedicat de 10 Gbps (entregable 7) contindran:

- Descripció de la situació
- Especificacions del segment de xarxa i aplicacions
- Proves de rendiment i fiabilitat del segment

Els continguts de l'estudi de viabilitat per a la connexió redundant (entregable 10) seran:

- Objectius de l'estudi
- Descripció de la situació inicial

- Especificacions del sistema i de les aplicacions
- Viabilitat tècnica
- Viabilitat operativa
- Viabilitat econòmica
- Alternatives

A diferència de la resta d'estudis, degut a restriccions temporals i de recursos del projecte, l'estudi de viabilitat per a la línia de backup no té continuïtat dins el marc del PFC.

1.3 Planificació

En la figura *1.3*.1 es pot observar un diagrama de *Gantt* amb la planificació del projecte simplificada.

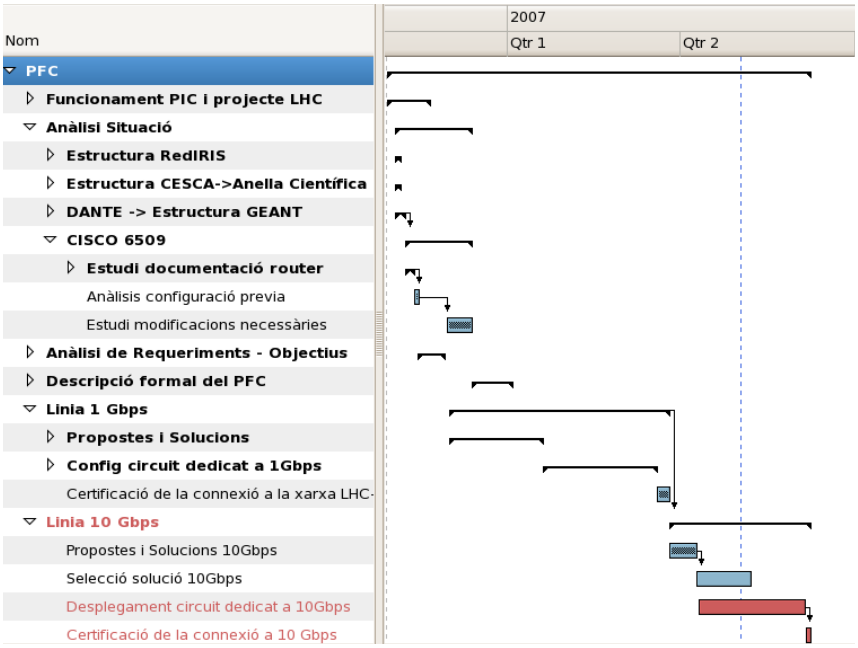


Figura *1.3*.1: Planificació simplificada del PFC on es poden observar les seves dues fases, línia de 1 i de 10 Gbps, finalitzant aquesta última al juny.

Cal dir que la planificació s'ha seguit amb força rigor fins a la finalització del desplegament sobre la línia d'1 Gbps (fase 1). Al llarg del desplegament de la línia de 10 Gbps (fase 2) s'ha seguit la planificació amb un retràs aproximat d'una setmana degut a complicacions diverses.

A causa de la complexitat de la segona fase del projecte i la necessitat d'una estreta coordinació amb diverses entitats, aquesta segona fase s'ha gestionat com si es tractés d'un projecte independent. La planificació del desplegament del circuit dedicat a 10 Gbps es pot observar en la figura *1.3*.2.

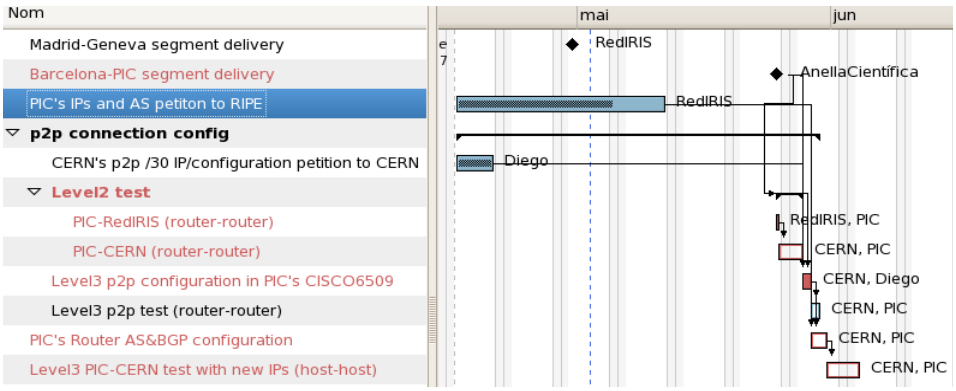


Figura *1.3.2*: Planificació del desplegament del circuit dedicat a 10 Gbps. L'entrega de la connexió física és clarament el camí crític

A l'annex A es pot trobar la planificació detallada de tot el projecte.

1.4 Organització de la memòria

La memòria d'aquest PFC s'estructura en cinc capítols dividits en seccions, les quals es troben organitzades en subseccions.

En el segon capítol s'expliquen els fonaments teòrics necessaris per entendre els conceptes tècnics que apareixen posteriorment en els capítols 3 i 4.

Els capítols 3 i 4 corresponen respectivament a les fases de disseny, implementació i proves de les dues fases del projecte. En el darrer capítol s'hi inclouen les conclusions i les possibles ampliacions.

La numeració dels capítols, seccions i subseccions segueix el model nºcapítol.nºsecció.nºsubsecció. Al llarg de la memòria hi ha petites ampliacions sobre conceptes relacionats amb el contingut de les subseccions, aquests apartats estan indicats amb un títol no numerat. La numeració de les figures és relativa a la secció, seguint l'estructura nºcapítol.nºsecció.nºfigura.

2 Capítol 2 – Fonaments teòrics

En aquest capítol es tractaran diversos conceptes i tecnologies relacionades amb el món de les xarxes de computadors.

Es donarà una visió general sobre com funcionen les xarxes (secció 2.1) i a continuació es detallarà la funcionalitat de cada capa de la xarxa a nivell tècnic, centrant-se en tecnologies, conceptes i aplicacions específiques rellevants per a la comprensió del projecte. El que es troba a continuació és una acurada selecció de fonaments teòrics relatius al disseny i la execució del projecte.

2.1 El model OSI i TCP/IP

En aquesta secció es donarà una visió general sobre com funcionen les xarxes. Es veurà el model de referència OSI que posteriorment es compararà amb TCP/IP, el model realment implementat.

2.1.1 El model OSI

OSI és el model de referència d'Interconnexió de Sistemes Oberts (OSI) llançat el 1984, és el model de xarxa descriptiu creat per ISO. El model OSI està dividit per nivells o capes que són:

7-**Aplicació**: aquesta capa ofereix a les aplicacions la possibilitat d'accedir als serveis de les demés capes i als protocols que utilitzen les aplicacions per a intercanviar dades. Dins aquesta capa trobem protocols com ara el correu electrònic (POP/SMTP), servidors de fitxers (FTP), web (HTTP), etc.

6-**Presentació**: s'encarrega de manegar les estructures de dades abstractes i realitzar les conversions de representació de dades necessàries per aconseguir una interpretació correcta encara que l'origen i el destí utilitzin representacions diferents.

5-**Sessió**: aquesta capa és la que s'encarrega de mantenir l'enllaç entre dues computadores que estan transmetent dades. En alguns escenaris aquesta capa és completament innecessària.

4-**Transport**: La seva funció bàsica és acceptar les dades enviades per les capes superiors, dividir-los en petites parts (si cal) i passar-los a la capa de xarxa, independitzant l'accés a la xarxa del suport físic utilitzat.

3-**Xarxa**: s'ocupa de fer que les dades arribin des de l'origen al destí, encara que ambdós no estiguin directament connectats. És la capa encarregada de trobar un camí mantenint una taula d'encaminament i travessant els equips que calgui per a portar les dades al destí.

2-**Enllaç de dades**: s'ocupa del direccionament físic, de la topologia de la xarxa, l'accés a la xarxa, la notificació d'errors, la distribució ordenada de trames i el control de flux.

1-**Física**: és l'encarregada de transmetre els bits d'informació a través del medi físic utilitzat per a la transmissió. S'ocupa de les propietats físiques, les característiques elèctriques dels diversos components i de la velocitat de transmissió. També d'aspectes mecànics de les connexions i terminals, incloent la interpretació de les senyals elèctriques/electromagnètiques.

Per a realitzar una comunicació cada capa passa la informació a la capa inferior, la qual encapsula la informació i fa el mateix fins a arribar a la capa 1. A la capa 1 és on es fa arribar les dades físicament al servidor[1]/dispositiu de xarxa[2] destí, on es realitza el procés invers. Així la capa *n* al destí rep exactament el mateix que ha enviat la capa *n* d'origen.

2.1.2 El model TCP/IP

Degut a la complexitat de OSI, el model implementat realment és TCP/IP v4/v6. La versió 4 de TCP/IP és la que actualment es troba funcionant a la majoria de xarxes, com ara la Internet. En la figura 2.1.1 es pot observar una relació entre les capes OSI i les capes TCP/IP, així com algunes de les aplicacions/implementacions més esteses per a les diferents capes TCP/IP.

<table>
<tr><th>OSI</th><th>TCP/IP</th></tr>
<tr><td>Aplicació</td><td rowspan="3">Aplicació
(DHCP, DNS, FTP, HTTP, ...)
PDU[3]=Missatge/Flux</td></tr>
<tr><td>Presentació</td></tr>
<tr><td>Sessió</td></tr>
<tr><td>Transport</td><td>Transport
(TCP, UDP, DCCP, ...)
PDU=Segment/DG d'usuari</td></tr>
<tr><td>Xarxa</td><td>Xarxa
(IPv4, IPv6, IGMP, ICMP, ARP, ...)
PDU=Paquet</td></tr>
<tr><td>Enllaç de dades</td><td>Enllaç de dades
(802.11, ATM, DTM, Ethernet, FDDI, Frame Relay, ...)
PDU=Trama</td></tr>
<tr><td>Física</td><td>Física
(Fibra òptica, UTP, ISDN, PLC, ...)
PDU=stream de bits</td></tr>
</table>

Figura 2.1.1: relació entre les capes OSI i les capes TCP/IP. Es pot observar com TCP/IP agrupa les 3 capes més altes d'OSI en una

1 En el context d'aquest PFC es considera servidor qualsevol ordinador/màquina connectat i que es comunica per la xarxa

2 Dispositiu e xarxa: qualsevol element de la xarxa que no és un servidor; encaminadors, switch, etc. Per a referir-nos a tots els servidors i/o elements de la xarxa es farà referència als membres de la xarxa.

3 PDUs (en anglès Protocol Data Units): és la unitat de dades amb la que es tracta en cada capa del protocol

En la figura 2.1.2 s'observa l'estructura per capes de TCP/IP i com, al llarg del flux de missatges entre les diferents capes, es van afegint/eliminant les diferents capçaleres.

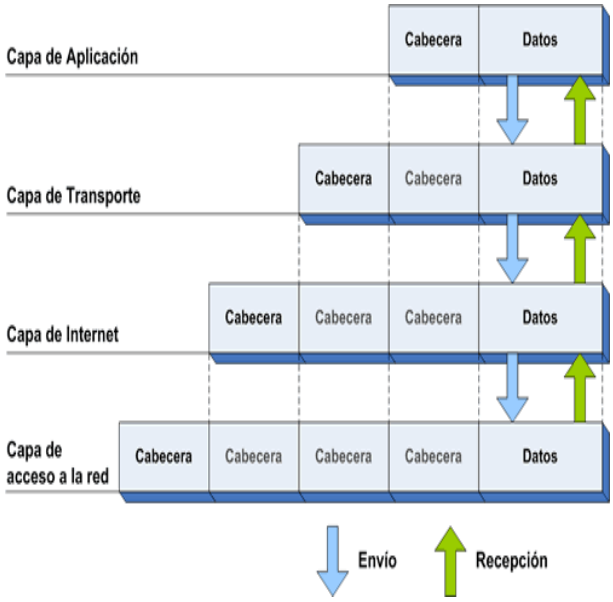


Figura 2.1.2: Flux de comunicació sobre TCP/IP on es veu clarament la separació per capes del protocol

2.2 Capa física

En aquest PFC les tecnologies de capa física utilitzades son cable de parell trenat (UTP) i fibra òptica. En ambdós casos la informació es restringirà als suports més estesos que permetin connexions d'enllaç de dades del tipus Ethernet (10/100/1000/10GE).

2.2.1 UTP

El cable de parell trenat UTP[2] (Unshielded Twisted Pair), com el de la figura 2.2.1, és el més estès en les xarxes LAN Ethernet actuals, proporcionant connexions a velocitats des dels 10Mbps fins als 10 Gbps a baix cost. Existeixen diferents categories (cat1,2,3,...,7), cada categoria estableix una sèrie de característiques i propietats. Per a xarxes GigabitEthernet és necessari utilitzar cable Cat6[3] o superior. També és possible disposar de connexions a 10 Gbps amb cables UTP, però es veuen limitades a distàncies molt curtes; actualment, amb les diferents tecnologies existents, de 15 (10GBASE-CX4) a 55 (10GBASE-T) metres.

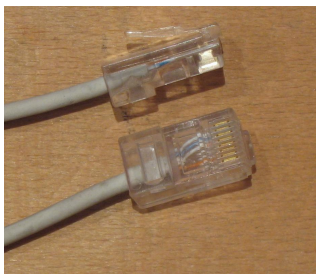


Figura 2.2.1: Connectors de 8 pins (RJ45) dels extrems d'un cable UTP per a Ethernet

2.2.2 Fibra òptica

La fibra òptica permet realitzar connexions d'alta velocitat a un cost moderat. Igual que sobre UTP, sobre fibra òptica trobem diferents estàndards amb diferents característiques i propietats. La varietat més coneguda és LAN PHY[4], que utilitza una velocitat de connexió de 10.3 Gbps i una codificació 64B/66B[4].

Dins de la varietat LAN PHY de 10 Gbps trobem diferents estàndards que permeten diferents longituds de connexió sobre diferents tipus de fibra. D'una banda hi ha 10GBASE-SR, dissenyat per a proporcionar 10 Gbps a curta distància (26-300m) sobre fibra multi-mode[5]. D'altra banda trobem l'estàndard 10GBASE-LR, una tecnologia òptica de llarga distància (Long Range) capaç de realitzar connexions a 10 Gbps sobre fibra mono-mode[6] de 1300nm a uns 10km de distància. Dins de la varietat LAN PHY trobem estàndards que poden arribar fins als 80km (10GBASE-ER).

També hi ha la varietat WAN PHY, que és una adaptació d'alguns dels estàndards de LAN PHY per tal que puguin funcionar sobre un canal SDH/SONET STS-192c/STM-64 a 9.953 Gbps, un suport físic per a connexions sobre fibra molt estès.

A part de la fibra és necessari disposar dels transmissors, comercialitzats normalment en forma de XENPAKs, com el mostrat en la figura 2.2.2. Degut a la íntima relació entre la fibra utilitzada i el transmissor, es pot trobar quasibé tants transmissors com estàndards de fibra.

Figura 2.2.2: Xenpak-10GB-LR com l'instal·lat al PIC per a la connexió a 10 Gbps de la segona fase del projecte

Entre la fibra i el transmissor encara ens falta un últim pas: el connector. Existeixen diversos tipus de connectors[7] de fibra òptica, tot i que els més usuals són LC (Lucent connector/Local connector) i SC (Subscriber Connector/Standard Connector), com el que es pot observar en la figura 2.2.3. A part de les diferències mecàniques, cada tipus de connector té les seves propietats respecte atenuació del senyal, resistència, etc.

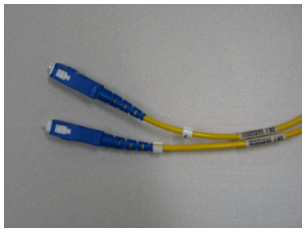


Figura 2.2.3: Connector tipus SC com l'utilitzat pel Xenpak de la figura 2.2.2

4 PHY és una abreujament comú per a *pyisical layer of OSI model*, és a dir la capa física.

5 Una fibra multi-mode permet múltiples modes de propagacio de llum i l'ús d'un canó làser de baixa intensitat; és simple de dissenyar i econòmic però té una distància de propagació força limitada.

6 Una fibra mono-mode només permet un mode de propagació de la llum i necessita un canó làser d'alta intensitat però permet distàncies connexions de gran distància.

7 Es pot trobar una bona classificació de tipus de connectors a http://en.wikipedia.org/wiki/Optical_fiber_connector

Quan es parla de connexions de fibra òptica sol aparèixer el terme fibra fosca (dark fiber[5]). Fibra fosca és un terme utilitzat per a referir-se a cables de fibra que estan instal·lats però resten inactius, sense llum ni l'electrònica necessària per al seu funcionament. És usual parlar de camins de fibra fosca degut a que les empreses que cablegen quan tracen un nou camí, per motius econòmics, solen passar força més fibres de les necessàries, així aquestes fibres passen a ser fibres fosques i resten disponibles per a un ús futur. Alhora de crear una connexió sobre un circuit de fibra fosca el client es posa en contacte amb una empresa gestora la qual, generalment, proveeix l'electrònica òptica necessària per a la regeneració del senyal (si cal) i és el client el qual aporta els dispositius necessaris per a transmetre la informació a través de la fibra, il·luminant-la.

Sobre un únic circuit de fibra fosca és possible crear múltiples enllaços (lambdes) a nivell físic mitjançant l'addició de sistemes WDM o DWDM[6] en els extrems de la fibra. Així doncs un cop es disposa d'un enllaç de 10 Gbps sobre fibra és possible afegir un sistema DWDM als extrems de la fibra per tal d'obtenir diversos enllaços a 10 Gbps, incrementant l'ample de banda ofert pel circuit de fibra.

2.3 Capa d'enllaç de dades

En el PFC la tecnologia d'enllaç de dades sobre la qual es treballa és Ethernet, estàndard àmpliament estès en tot tipus de xarxes. En aquesta secció es donarà una visió general sobre l'estàndard Ethernet, incidint en alguns conceptes específics com ara JumboFrames o les VLANs.

Els switch i concentradors (hubs) són els dispositius de xarxa que s'encarreguen de gestionar/comunicar els membres de la xarxa en aquesta capa, generalment les gammes estàndard d'aquests dispositius no són capaços d'entendre les dades de les capes superiors.

2.3.1 Ethernet

Ethernet/IEEE 802.3 és un protocol d'enllaç que proporciona una interfície unificada al medi de xarxa. Permet a un S.O. transmetre i rebre varis protocols del nivell de xarxa de forma simultània, no és orientat a connexió i no és fiable.

Components de l'estàndard d'Ethernet/IEE 802.3:

- Directives del nivell físic: tipus de cables/fibra, limitacions de cablatge i mètodes de senyalització.
- Mecanisme de control d'accés al medi -> CSMA/CD (Carrier Sense Multiple Access with Collision Detection - el que traduït és: Detecció de portadora amb accés múltiple). Serveix per evitar col·lisions en les transmissions.
- Format de trama: ordre i funcions dels bits transmesos, el format estàndard d'una trama Ethernet correspon al de la figura 2.3.1.

Preàmbul	**SOF**	**MAC destí**	**MAC origen**	**Tipus**	**Dades**	**FCS**
7 bytes	1 byte	6 bytes	6 bytes	2 bytes	46 a 1500 bytes	4 bytes

Figura 2.3.1: Format estàndard d'una trama Ethernet

Dels camps que formen una trama ethernet n'hi ha que són per a correcció d'errors, separació de trames, etc. Els camps més rellevants per al procés de comunicació són:

- Direcció MAC de destí: camp de 6 bytes (48 bits) que especifica la direcció MAC a qui s'envia la trama. Cada dispositiu Ethernet que es fabrica en el mon rep una adreça MAC única, els dos primers estan sempre a zero. Aquesta direcció de destí pot correspondre a una NIC (un únic membre de la xarxa), un grup multicast (alguns membres de al xarxa) o la direcció broadcast de la xarxa (tots els membres de la xarxa). Cada membre de la xarxa examina aquest camp per determinar si ha d'acceptar o descartar el paquet.

- Direcció d'origen: camp de 6 bytes (48 bits) que especifica la direcció MAC des de la que s'envia la trama. El membre de la xarxa que hagi d'acceptar la trama coneix a traves d'aquest camp la direcció d'origen amb la que intercanviar les dades.

- Tipus: camp de 2 bytes (16 bits) que identifica el protocol de xarxa d'alt nivell associat amb el paquet, o en el seu defecte la longitud del camp de dades.

- Dades: camp que conté la informació útil juntament amb les capçaleres de les capes superiors. Per defecte el camp té una longitud de 46 a 1500 bytes, existeixen però els anomenats JumboFrames (veure subsecció 2.3.2) o Jumbogrames que són trames de mida superior a 1500 bytes.

2.3.2 Jumboframes

Els avantatges que aporten els JumboFrames són una menor càrrega en els processadors dels elements de la xarxa (100%->60%) degut a que es redueix el número de paquets a generar/tractar, i un petit increment en el throughput[8] (5-10%) gràcies a la disminució de l'ample de banda destinat a les capçaleres.

Com que els JumboFrames no són obligatoris dins de l'estàndard Ethernet no tots els productes els suporten. Els dispositius que suporten Jumboframes ho fan amb mides que oscil·len sobre els 9000bytes, generalment és un paràmetre a configurar anomenat MTU (Maximum Transfer Unit) i que, per defecte a Ethernet, és 1500bytes.

2.3.3 VLANs

Així dons a Ethernet els membres de la xarxa es comuniquen entre si, directament, mitjançant l'adreça física de la seva NIC (Network Interface Controller); la MAC (Media Access Control address). Podríem pensar que qualsevol servidor connectat a un mateix switch forma part de la mateixa xarxa, dins la mateixa LAN (Local Area Network), però degut a l'existència de les VLAN (Virtual LAN) això no és sempre així.

Les VLAN permeten definir diferents xarxes de capa 2 (enllaç de dades) independents sobre la mateixa capa física. La versió estàndard de les VLAN es troba definida a l'IEEE 802.1q[7] i permet

8 Throughput és el terme utilitzat per a referir-nos al volum d'informació que flueix a través d'un sistema, per exemple 1 Gbps (un gigabit per segon)

separar una xarxa en diferents segments, tant a nivell administratiu dins el switch com en els propis elements de la xarxa, segons diferents paràmetres com ara servidor d'origen, tipus de trànsit, etc. És important saber que dins d'un mateix servidor o switch hi poden conviure diverses VLANs de forma simultània i transmetre's totes sobre el mateix medi físic, això s'anomena *trunking*[8]. En la figura 2.3.2 se'n pot observar un exemple.

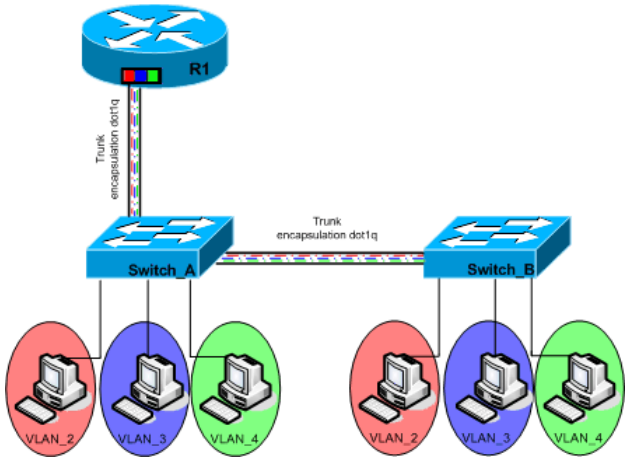


Figura 2.3.2: LAN separada en tres VLAN i amb els switch fent *trunking* per a crear un *trunk* (el canal per on passen les diferents VLANs)

Les VLAN aporten diverses avantatges com ara incrementar el número de dominis de broadcast però reduir-ne l'abast (reduint així el trànsit i incrementant la seguretat de la xarxa), redueixen l'esforç de crear i gestionar subxarxes, redueixen els requeriments de hardware (les xarxes es poden separar lògicament enlloc de físicament) i incrementen el control sobre els diferents tipus de xarxa.

2.4 Capa de xarxa

En les subseccions que es troben a continuació és dóna una visió general sobre l'essència del protocol IP: adreçament, resolució d'adreces i el format del paquet IP. Un cop vistes les característiques principals del protocol base de la capa de xarxa (IP) es fa un breu repàs a ICMP i s'explica com funciona l'encaminament sobre IP, que donarà pas a conceptes indispensables per a la segona fase del projecte: Sistemes Autònoms i BGP.

Els encaminadors (*routers*) són els dispositius de xarxa que s'encarreguen de gestionar/comunicar la xarxa en aquesta capa, generalment les gammes estàndard d'aquests dispositius també són capaces d'entendre algunes capçaleres de capes superiors, cosa que s'aprofita per a crear filtres/*firewalls*/ACLs[9] en els mateixos encaminadors per tal de controlar el trànsit de la xarxa. També és comú trobar switch de gamma alta capaços de realitzar tasques pròpies d'encaminadors, com per exemple la sèrie de switch Catalyst de Cisco Systems (veure figura 2.4.1).

9 Les sigles ACL corresponen a Access Control List, una llista de regles d'accés.

Figura 2.4.1: switch modular Cisco Catalyst 6509, com l'utilitzat al PIC

2.4.1 IP

En aquesta capa, treballant sobre IP, l'identificador únic d'un servidor en una xarxa és l'adreça IP, un enter de 32 bits que se sol expressar en nomenclatura *dotted quad*: 80.34.93.37. Degut a l'aparició de la Internet[10], i la necessitat de gestionar-la, es crearen les entitats de gestió de l'espai d'adreçament IP (RIPE a Europa, ARIN als EUA, etc.), i una classificació del mateix (esquema classful) que correspon a la taula de la figura 2.4.2.

Classe	**Representació binària**	**Dotted Quad**
A	0 | Net Id | Host[11]Id	de 1.0.0.0 a 126.0.0.0 2^7 (128) xarxes, amb 2^{24} (16M) servidors
B	10 | Net Id | HostId	de 128.1.0.0 a 191.255.0.0 2 ^{14}(16K) xarxes, amb 2^{16}(64K) servidors
C	110 | NetId | HostId	de 192.0.1.0 a 223.255.255.0 2^{21} (2M) xarxes, amb 2^8 (256) servidors
D	1110 | Multicast	de 224.0.0.0 a 239.255.255.255
E	1111 | Reservades	de 240.0.0.0 a 255.255.255.254

Figura 2.4.2: esquema de l'adreçament classful Ipv4 d'Internet. En l'àmbit d'aquest PFC el més normal és l'ús de classes C i/o B.

A part de seva adreça IP, un servidor necessita saber la màscara de xarxa per tal de delimitar l'abast de la subxarxa on es troba. Una màscara de xarxa se sol expressar com 255.255.255.0. Per a saber les IP que pertanyen a la mateixa subxarxa (l'abast de la xarxa) cal fer una AND binària amb l'adreça IP i la màscara de xarxa, els bits que queden a 0 són els que es poden variar per a representar l'espai d'adreçament de la subxarxa. Tanmateix, per tal de poder accedir a tots els servidors de la xarxa (fora de la seva subxarxa) és necessari que el servidor disposi d'accés a una porta d'enllaç (gateway o camí per defecte). Una porta d'enllaç és un encaminador (*router*) on el servidor enviarà els paquets on el destinatari no sigui una IP de la subxarxa del servidor. A la

10 Per definició, una internet és una xarxa de xarxes qualsevol (una empresa podria tenir una internet privada). La Internet (en majúscules) és la xarxa de xarxes pública i d'abast mundial que tots coneixem.

11 Host és el terme anglès per a referir-se a un servidor/màquina connectat i que es comunica per la xarxa

pràctica això significa que al comunicar-nos amb un servidor...

- ... de la mateixa subxarxa, la capa 3 indicarà a capa 2 que realitzi entrega directa al servidor, utilitzant així l'adreça MAC del servidor de destí. La porta d'enllaç no és necessària per a realitzar la comunicació.

- ... de fora de la subxarxa, la capa 3 indicarà a capa 2 que realitzi entrega directa a l'encaminador, utilitzant l'adreça MAC de l'encaminador, el qual rebrà el paquet i inspeccionarà el contingut de la capçalera d'IP (capa 3) per tal de saber on ha de reenviar-lo (*forwarding*).

Per a trobar l'adreça MAC a partir de la IP s'utilitza el protocol ARP (*Address Resolution Protocol*), a la inversa (MAC->IP) s'utilitza RARP (*Reverse ARP*). ARP funciona enviant una petició a l'adreça broadcast de capa 2 (MAC = ff ff ff ff ff ff) indicant l'adreça IP sobra la que fa la consulta, quan un element de la xarxa (p.e. el propi servidor) veu la petició i coneix la relació IP ->MAC, genera un missatge de resposta.

Cap al 1993 es va veure que l'esquema "*classful"* no oferia tota la flexibilitat necessària i es passà a l'esquema actual d'adreçament de la Internet, el "*classless"*. Aquest és un esquema basat en CIDRs (*Classless Inter-Domain Routing*), que són una espècie de subxarxes representades com IP/X on X és el número de bits de la IP que són fixes per a la xarxa. Així doncs un CIDR /28 seria equivalent a una màscara de xarxa 255.255.255.240, una subxarxa amb 16 IPs.

Ara que ja sabem com es poden comunicar diferents servidors via IP cal veure (figura 2.4.3) com és un paquet IP i quina informació conté.

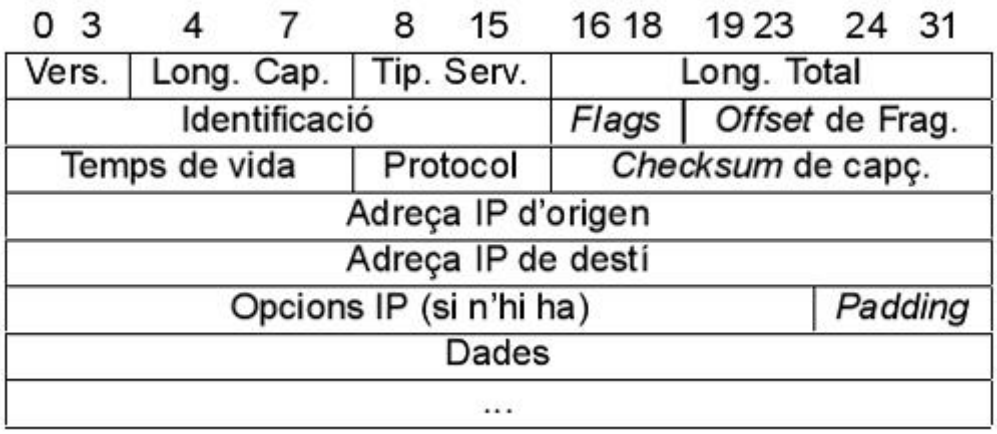


Figura 2.4.3: camps de la capçalera d'un paquet IP

Gràcies a la separació per capes de TCP/IP el format del paquet no depen del hardware subjacent. A continuació es descriuran alguns camps de la capçalera:

- Versió (4 bits): per indicar la versió del protocol que s'ha utilitzat per crear el paquet. Abans de processar-lo és obligat comprovar aquesta versió.
- Longitud capçalera (4 bits): per indicar la mida de la capçalera, en número de paraules de 32 bits.
- Longitud Total (16 bits): Longitud del datagrama IP sencer, comptant capçaleres i dades,

en bytes. La longitud de les dades s'obté restant-li la longitud de la capçalera. La longitud màxima d'un datagrama IP és 65535 Bytes = 64KBytes

- Tipus de Servei (8 bits): També anomenat TOS (Type of Service), indica cóm s'ha de tractar el datagrama (prioritats). No tots els encaminadors interpreten el TOS.
- Identificació (16 bits): identifica el datagrama. Tots els fragments d'un mateix datagrama tenen el mateix identificador. Les implementacions de IP solen utilitzar un comptador com a identificador únic.
- Offset del fragment: cada fragment de datagrama indica quina part del datagrama original porta. Aquest valor comença en 0 i està expressat en múltiples de 8 bytes.
- Flags: Són 3 bits, però només 2 són utilitzats pel control de la fragmentació: Bit de no fragmentació (es possible prohibir la fragmentació amb aquest bit) i el Bit de més fragments (està activat a tots els fragments a excepció de l'últim).

El servei d'aquesta capa no és orientat a connexió, o sigui, envia els paquets al següent node de la xarxa i se n'oblida. Per això IP és un protocol d'entrega no fiable, del millor intent (pressuposem que tot anirà bé) i sense connexió.

2.4.2 ICMP

IP és la base actual de les comunicacions per la Internet i com ja hem dit pot succeir que els paquets no arribin a la destinació. D'alguna manera hem d'esbrinar si ha fallat i per què, i aquí és on entra en acció el protocol ICMP (Internet Control Message Protocol).

ICMP serveix per a notificar alguns problemes de xarxa, sobretot centrat específicament en IP, com ara que el servidor destí estigui desconnectat, que algun dels encaminadors pel que ha passat el datagrama està congestionat o que el TTL[12] del paquet s'esgoti.

Aquest protocol de capa de xarxa, que al PIC (*Port d'Informació Científica*) funciona sobre Ethernet, també s'utilitza per a diagnosticar i *debugar* la xarxa amb paquets del tipus echo i echo reply (utilitzats per la utilitat ping), traceroute, per indicar que un paquet és massa gran i s'ha de fragmentar (cosa que IP fa a capa 3), etc.

2.4.3 Encaminament IP

Fins ara hem vist com es comuniquen els servidors dins d'una mateixa xarxa (o subxarxa), i que quan un paquet s'ha d'enviar a una altre xarxa, s'envia a l'encaminador. Però què fa l'encaminador (*router*) amb els paquets que són per un altre servidor i ell rep?

L'encaminador és un dispositiu que ha d'estar present a múltiples xarxes de forma simultània i, mitjançant una taula d'encaminament, reenvia els paquets d'una xarxa a una altre a través de les seves interfícies. Una interfície és una connexió que té un dispositiu amb una xarxa determinada, per a cada accés a una xarxa existeix una interfície, on es manté informació sobre la xarxa: el tipus, l'adreça física, l'MTU, etc. La interfície també mantindrà informació sobre l'accés a la xarxa

12 Temps de vida (Time To Live): camp de la capçalera IP que serveix per evitar bucles. Es decrementa a cada router i quan arriba a 0 el paquet es descarta.

(adreça IP, màscara de xarxa, estadístiques, etc.), i oferirà els serveis per accedir-hi.

Així doncs un paquet que ha d'anar de la xarxa A a la xarxa B és possible que hagi de traspassar múltiples xarxes intermediàries, via els seus respectius encaminadors, per arribar al destí.

Per tal de mantenir la citada taula d'encaminament els encaminadors utilitzen diversos protocols per a comunicar-se entre si: els protocols d'encaminament (*routing*). L'encaminament és l'elecció del camí que seguirà el datagrama a través de les possibles xarxes a les quals està connectat i el posterior enviament. Depenent de l'origen també es pot dir re-encaminament (forwarding) de paquets.

2.4.4 Sistemes Autònoms (AS)

Per tal d'entendre els diferents protocols d'encaminament cal saber que la Internet, vista "de lluny", està formada per Sistemes Autònoms (AS de l'anglès *Autonomous System*). Els AS són conjunts d'encaminadors controlats per una única autoritat administrativa (com ara un ISP o una gran empresa) i que utilitzen, dins el seu domini, un mateix protocol d'encaminament intern (IGP: *Interior Gateway Protocol*) per a la distribució i actualització de la informació d'encaminament.

Els AS, utilitzant el seu ASN (*Autonomous System Number*), es connecten entre si mitjançant encaminadors (*routers)* de frontera (RF) que utilitzen un mateix protocol d'encaminament (EGP: *Exterior Gateway Protocol*). A part del RF (p.e. ISP), dins d'un AS trobem encaminadors interns (RI, p.e. delegacions de l'ISP) i encaminadors locals (R, p.e. empreses a les que dóna accés l'ISP). Els RI serveixen per a segmentar internament l'AS, tenen taules dinàmiques on es reflexa l'estat actual de l'AS i una entrada per defecte que apunta al RF. Els R tenen taules d'encaminament estàtiques, que omple manualment l'administrador de la xarxa, també tenen una entrada per defecte que apunta al RI.

Dins dels AS també existeix una certa taxonomia:

- Sistema Autònom Extrem (SAE): Només encamina trànsit propi, pot estar connectat a un altre SAE però seria com una extensió.
- Sistema Autònom de Trànsit (SAT): Un SAT connecta diversos SAE en forma d'arbre, és com un SAE que encamina tot el trànsit (i no només el propi).
- Sistema Autònom Multihomed (SAM): és un AS que està connectat a més d'un SAE, però a diferència d'un SAT només encamina el trànsit propi. Podem veure-ho com un SAE amb més d'una connexió, per tal de simplificar el model generalment en els diagrames no es fa la diferència de nomenclatura SAE-SAM.
- Network Access Point (NAP[13]): serveixen de punts de connexió entre diversos AS, convertint l'arbre format pels SAT en un graf.

En la figura 2.4.4 es pot veure un exemple de la topologia de xarxa que es pot fer mitjançant els

13 A Espanya, podem trobar exemples de NAP amb CATNIX i ESPANIX

diferents tipus d'AS, es pot observar els SAE que hi apareixen detallats són en realitat SAM, però no s'indica per a simplificar el sistema.

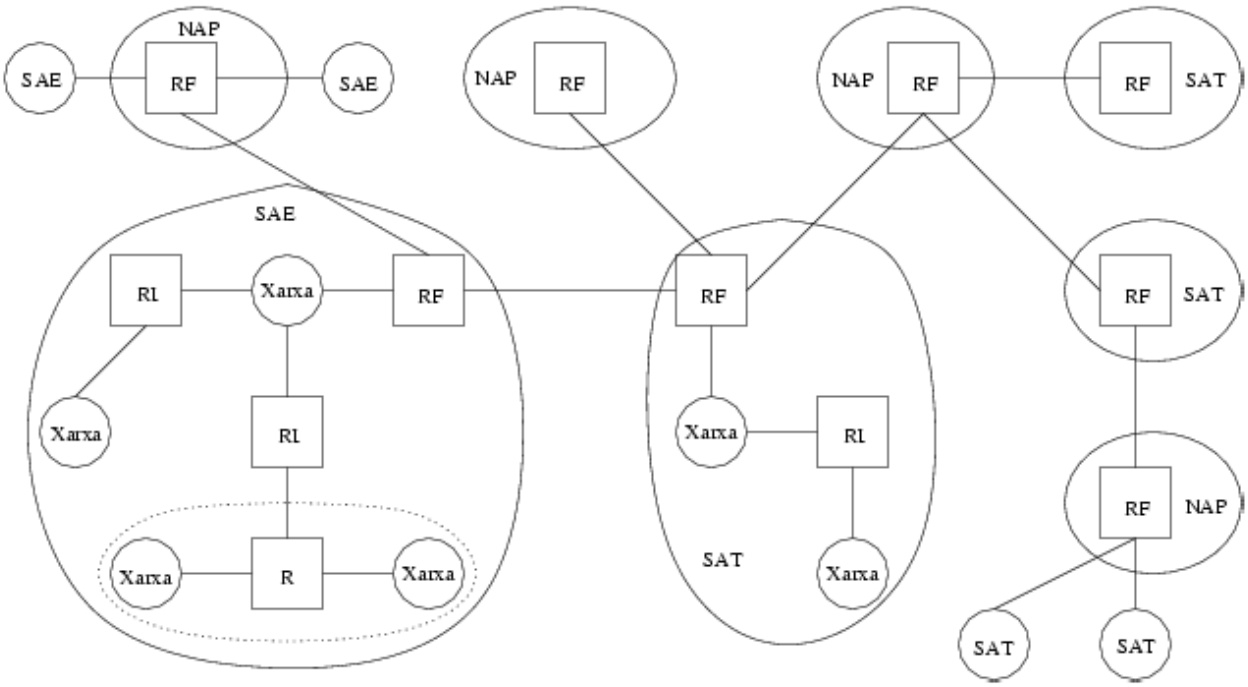


Figura 2.4.4: topologia formada per diferents tipus d'*Autonomous System*

2.4.5 BGP

Tal i com s'ha dit abans, per a la comunicació entre Sistemes Autònoms (AS) s'utilitza un EGP (*Exterior Gateway Protocol*); actualment a la Internet s'utilitza BGP versió 4, especificat a l'RFC[9] 4271. BGP no és un EGP de tipus vector de distàncies pur ni tampoc de tipus d'estat d'enllaços pur, però podem dir que BGP està més a prop del tipus vector de distàncies (familia Bellman-Ford).

El protocol es caracteritza perquè:

- Permet la comunicació entre diferents sistemes autònoms. Per això s'utilitza en els RF com a EGP (els RF també han de disposar d'un IGP per a l'intercanvi d'informació dins el seu SA)
- Coordina diferents Encaminadors de Frontera (RF). Així s'evita la necessitat de tenir un "Master"; tots els RF es coordinen mitjançant aquest protocol
- Propaga informació dels enllaços.
- Dóna informació de "següent salt", com els protocols de vector de distàncies. És una bona forma d'optimitzar el rendiment del sistema
- Transport fiable: utilitza TCP, així s'evita haver de realitzar el control de fiabilitat en el propi protocol. El problema és que al funcionar sobre TCP BGP es veu afectat per les vulnerabilitats d'aquest i és necessari establir primer una connexió punt a punt entre els diferents *encaminadors*.

- Ofereix informació dels camins.
- Modificacions incrementals: Només envia la informació sencera el primer cop, després només envia diferències (deltes). Així s'evita congestionar la xarxa innecessàriament.
- Suport per adreçament “classless” (adreces amb CIDR).
- Permet l’agregació de rutes.
- Suport de polítiques.
- S’utilitza autenticació

Les funcions bàsiques que ha de realitzar el protocol BGP són les següents:

1. Establiment del veïnatge
2. Intercanvi d’informació d'encaminament
3. Manteniment del veí/connexió (detecció de problemes en la connexió)

Com ja s'ha dit, BGP requereix utilitzar un encaminador que tingui configurades connexions punt a punt amb cadascun dels seus veïns, amb els quals s'intercanviarà informació de les rutes que cadascun conegui.

Així doncs necessitarem que la taula de cada RF tingui entrades del tipus “Xarxa Següent, encaminador, Camí“, on el camí contindrà informació de com arribar al destí. D’aquest manera es poden detectar bucles (AS repetits en el camí especificat). Gràcies a aquesta estructura també es poden aplicar polítiques d'encaminament: per exemple triar unes rutes o altres en funció del seu cost.

En aquest apartat no es donarà molta informació respecte als IGP, els protocols d'encaminament utilitzats dins el propi AS, ja que la topologia local del PIC no és complexa i el seu paper dins el PFC no és rellevant. De totes formes cal dir que, d'entre els múltiples IGP existents, al PIC s'utilitza EIGRP[10]: un protocol híbrid, propietari de Cisco Systems, que ofreix el millor dels algorismes de vector de distàncies i de l'estat de l'enllaç. Es poden trobar alternatives a EIGRP en OSPF o IGRP.

2.5 Capa de transport

El protocol IP ofereix un servei no fiable de comunicació, és a dir, els paquets poden arribar fora d'ordre, duplicats, o senzillament, no arribar. Així doncs en IP la fiabilitat depen de l'aplicació. Per a no dependre de les aplicacions existeixen altres protocols que van per sobre de IP, en capes superiors, i que s'ocupen del transport de datagrames/segments. Aquests protocols son protocols d'extrem a extrem:

- User Datagram Protocol (UDP) : Transferència de missatges no fiable entre aplicacions.
- Transmission Control Protocol (TCP): Flux fiable de bytes entre aplicacions.

Podríem pensar que si existeix un protocol fiable per la transmissió com és el cas de TCP, aleshores UDP no el necessitem per res. Això no és cert ja que hi ha aplicacions en les quals no ens

importa perdre alguns datagrames. Un exemple d'això és el Streaming de vídeo: si perdem algun frame de la imatge no ens importa massa, podem continuar reproduint el vídeo sense problemes.

2.5.1 Ports

Els destinataris finals dels missatges son les aplicacions i en una màquina normalment s'estan executant diverses aplicacions a la vegada. Així doncs, hem de trobar la forma d'identificar el programa al qual li volem enviar el datagrama, ja que amb l'adreça IP de la màquina destí no serà suficient. Per a solventar aquest problema apareixen els ports, que són punts de destí abstractes que pertanyen a un servidor concret. En TCP i UDP s'identifiquen amb un enter positiu (16 bits) que es troba en la capçalera de la capa de transport, així doncs els ports seràn independents en cada protocol de la capa de transport (port 1 TCP i port 1 UDP són diferents).

A la pràctica s'han definit dues categories de ports:

- Ports d'assignació restringida (Well Known Ports): Rang: [1..1023] . Són ports reservats per a una aplicació específica, i son comuns a tots els servidors. Fan falta privilegis especials per a poder-los assignar (en Linux privilegis de *root*). Exemples són el port 21/tcp per ftp, 22/tcp per al servei ssh, etc. En Unix trobem una relació port-servei en el fitxer */etc/services*

- Ports d'aplicacions de client: Rang: [1024..65535] . L'assignació és dinàmica, l'ús és lliure i no calen permisos especials (qualsevol usuari de la màquina pot utilitzar-los).

En la figura 2.5.1 es pot observar l'esquema d'utilització dels ports.

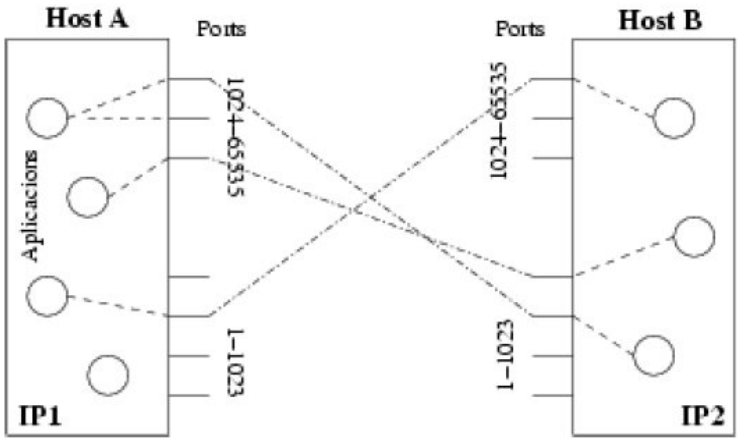


Figura 2.5.1: Esquema d'utilització dels ports TCP/UDP

2.5.2 Protocol de Datagrames d'Usuari (UDP)

El protocol UDP ens ofereix que donat un servidor de destinació i un port, li podem enviar missatges (datagrames) sabent que si arriben, arribaran a l'aplicació de destí que escolta el port al que volem enviar la informació.

UDP no és un protocol d'entrega fiable pels següents motius:

- A l'igual que IP, UDP no està orientat a connexió.
- No usem missatges de confirmació (ACK) quan es reben els datagrames (almenys no en aquesta capa).
- Els missatges poden arribar desordenats (donat que IP no garanteix ordenació) i UDP no s'encarrega d'ordenar-los.
- No hi ha cap mena de control sobre el flux d'informació entre els 2 extrems.

Tal i com es pot observar en la figura 2.5.2 els datagrames UDP són molt simples, només consten dels ports de destí i origen, de la longitud del missatge, del chekcsum i de les pròpies dades. La resta de dades necessàries es troben contingudes en les capçaleres de les capes inferiors, com ara la del paquet IP que transportarà el datagrama.

0-15	16-31
Port UDP origen	Port UDP destí
Longitud missatge	Checksum UDP
Dades	

Figura 2.5.2: esquema d'un datagrama UDP

2.5.3 Protocol de Control la Transmissió (TCP)

En el decurs del projecte apareixen múltiples referències a aspectes força concrets del protocol TCP. Donada la complexitat del protocol en aquesta subsecció només se n'explicaran els aspectes més rellevants. Per obtenir-ne més detalls s'aconsella consultar la bibliografia [11] i [12].

La propietat principal de TCP és que ens ofereix transmissions fiables sobre IP, altres característiques importants del protocol són:

- Orientació a flux: El flux que surt d'una aplicació d'origen arribada a l'aplicació destí exactament igual.
- Connexió: Implica fer un establiment de connexió, els extrems han d'estar d'acord amb establir la connexió.
- Transferència amb *buffer*: Les dades es mantenen en un *buffer*, les escriptures i lectures d'aquests *buffers* són blocants (escriptura blocada si està ple, lectura blocada si està buit). Si el receptor no dóna a l'abast el protocol farà que l'emissor aturi la transmissió temporalment.
- Flux no estructurat: L'estructura de les dades l'ha de proporcionar l'aplicació que les utilitzi.
- Connexió bidireccional (full duplex): El servei de flux ha de permetre transferències simultànies d'informació en tots dos sentits de la comunicació.

Fiabilitat en TCP

Per tal d'oferir fiabilitat TCP es basa en dos mecanismes:

A. **Confirmació positiva amb retransmissió.** El funcionament del mecanisme és el següent:

   I. L'Emissor envia un segment de dades

   II. El receptor rep les dades i envia un segment ACK (acknowledgment) com a resposta de confirmació a l'emissor.

   III. L'emissor espera per enviar el següent segment de dades fins rebre el missatge ACK del receptor.

   IV. Si l'emissor no rep l'ACK després d'un cert temps es reenvia el segment (es retransmet).

B. Finestra lliscant (sliding window). Es defineix una finestra com un subconjunt de segments consecutius que s'han d'enviar. Els segments de la finestra es poden enviar sense esperar rebre la confirmació del segment anterior, optimitzant substancialment el mecanisme anterior (confirmació positiva).

TCP utilitza un mecanisme de finestra lliscant especialitzat que permet modificar la mida de la finestra durant la transmissió i que obliga a l'emissor a mantenir tres apuntadors associats a la finestra: inici de finestra, bytes enviats i límit de finestra. En una seqüència de dades que espera ser enviada via TCP tindrem tres tipus de segments:

I. Enviats i confirmats: segments anteriors a l'apuntador d'inici de finestra.

II. Enviats però no confirmats: segments que estan dins la finestra, l'apuntador de bytes ja enviats indica l'últim d'aquests segments.

III. No enviables: encara no entren dins l'abast de la finestra, són segments posteriors a l'indicat per l'apuntador límit de finestra.

Els apuntadors d'inici i límit de finestra aniran avançant a mesura que es vagin rebent els missatges ACK del receptor i el de bytes enviats a mesura que l'emissor enviï els segments de dades. Així doncs, per tal que el mecanisme sigui útil i les connexions TCP eficients, serà molt important fixar una mida de finestra adequada a la velocitat de transferència i recepció de dades de la connexió. En versions superiors a la 2.6 del kernel de Linux hi ha una opció que permet que la mida de finestra s'autoajusti[13], dins d'uns límits, a les necessitats de la connexió.

Format del segment

En la figura 2.5.3 es mostra l'ordre i la mida dels diferents camps d'un segment TCP. Cal recordar que més endavant aquest segment serà encapsulat dins d'un paquet IP, que indicarà protocol nº6 i inclourà la informació necessària per a la transmissió del segment cap al receptor.

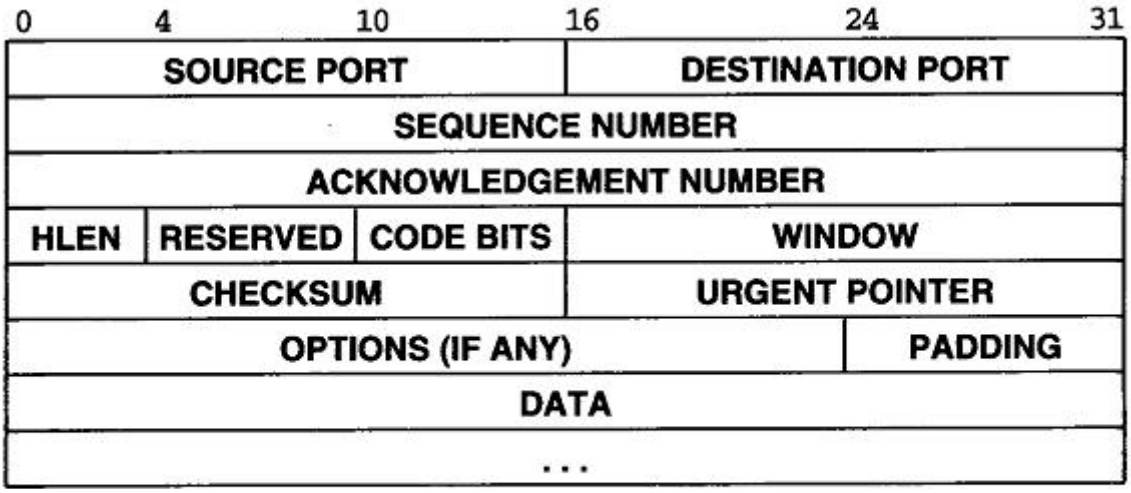


Figura 2.5.3: esquema d'un segment TCP

Els camps *SOURCE PORT* i *DESTINATION PORT* contenen els números de ports TCP origen i destí, *SEQUENCE NUMBER* identifica la posició del segment dins el flux de dades i l'utilitza el receptor per omplir el camp *ACKNOWLEDGEMENT NUMBER* alhora de confirmar la recepció (ACK). *HLEN* conté un enter que especifica la longitud del segment mesurat en múltiples de 32 bits, és necessari perquè el camp *OPTIONS* és de longitud variable.

El camp *WINDOW* anuncia la mida de la finestra. L'anunci de finestra pot ser diferent en tots els segments, permetent l'adaptació a les situacions del moment.

TCP pot transportar en un mateix segment dades, confirmacions (ACK), establiment/tancament de connexió, etc. Per això s'utilitza el camp *CODE BITS*, de 6 bits, on es pot indicar un valor a cada bit:

- URG: el segment és urgent i s'ha de processar tan bon punt es rebi
- ACK: el camp ACKNOWLEDGEMENT confirma un segment
- PSH: el segment requereix un push
- RST: reset/tancament abrupte de la connexió
- SYN: sincronització de números de seqüència / establiment de connexió
- FIN: fi de connexió (tancament no abrupte de la connexió)

El camp *CHECKSUM*, de 16 bits, és utilitzat com a CRC per a verificar l'integritat de les dades i de la pròpia capçalera TCP.

La màquina d'estats

Com en la majoria de protocols, el funcionament de TCP es pot representar mitjançant un autòmat. En aquest cas es tracta de l'autòmat d'onze estats de la figura 2.5.4, on l'estat inicial és CLOSED.

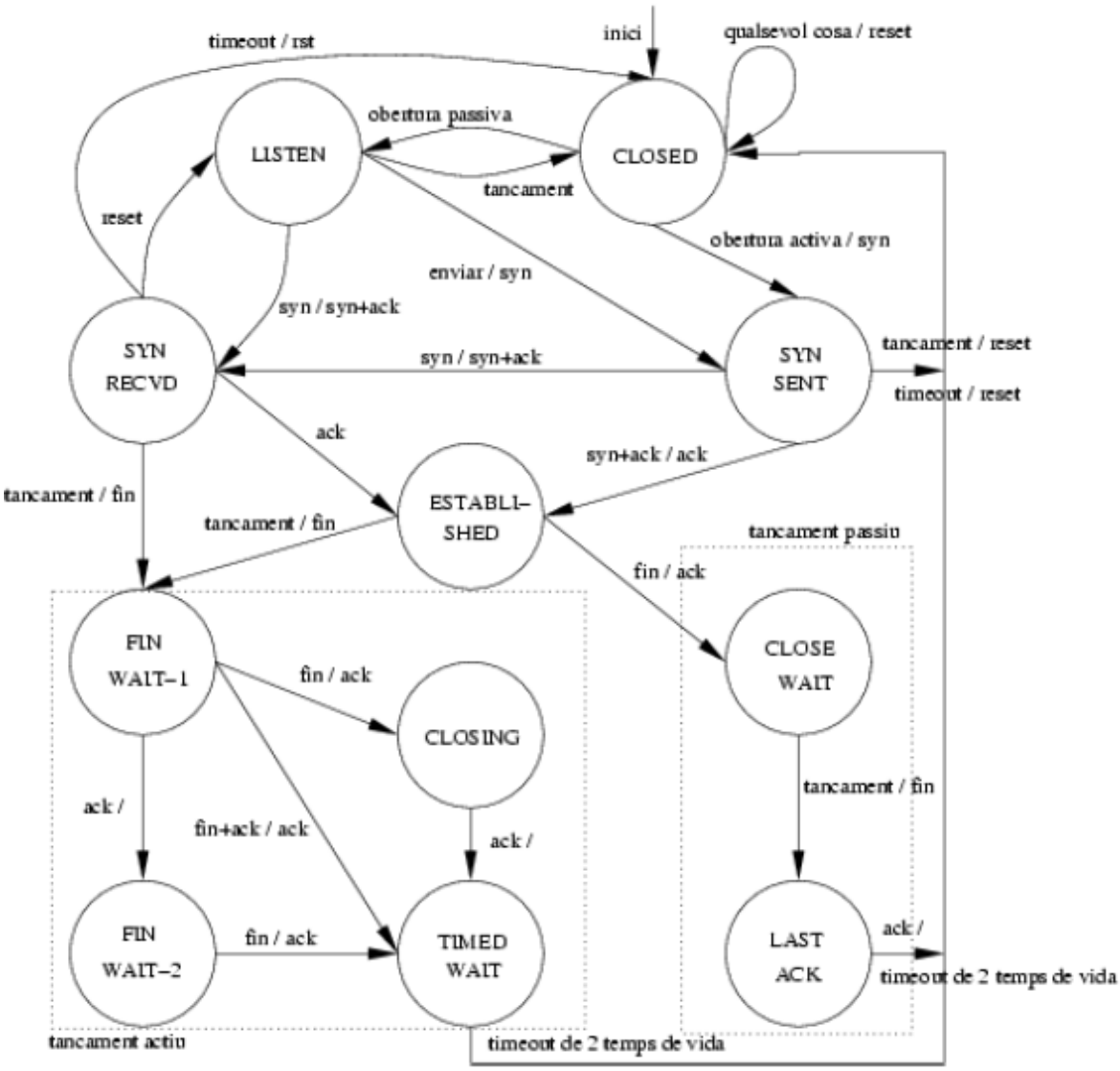


Figura 2.5.4: autòmat que representa el protocol TCP

2.6 Capa d'aplicació: eines de monitorització i mesura de rendiment

Hi ha infinitat d'aplicacions de capa d'aplicació com per exemple DNS, FTP, HTTP, IMAP, IRC, NFS, NNTP, NTP, POP3, SMB/CIFS, SMTP, SNMP, SSH, Telnet, SIP, etc. De fet pràcticament podríem dir que hi ha tants protocols com aplicacions diferents, per tal d'entendre aquesta afirmació és important veure que l'usuari normalment no interacciona directament amb el nivell d'aplicació, sinó que utilitza altres programes; un usuari no envia una petició "*HTTP/1.0 GET index.html*" per aconseguir una pàgina en HTML, sinó que utilitza un navegador web (firefox, IE, etc.).

En aquesta secció no es pretén tractar protocols de capa d'aplicació sinó donar una petita explicació de diferents programes o eines molt properes a aquests protocols, utilitzades per a la monitorització i mesura de rendiment de la xarxa sobre TCP/IP.

2.6.1 Eines de monitorització

Al llarg del desenvolupament del PFC s'han utilitzat diverses eines per a la monitorització del trànsit i les connexions des de la pròpia xarxa amb MRTG, Netflow Analyzer i altres aplicacions que consulten dades als switch i a l'encaminador per a generar estadístiques i gràfiques.

També s'han utilitzat eines per a la monitorització des dels servidors (Netstat, TCPdump, WireShark, ifconfig, etc). Donat que les eines per a la monitorització dels servidors és més complexa, en aquesta subsecció es donarà una vista ràpida de les utilitats més bàsiques, i alhora complexes, que són Netstat, TCPdump i Nmap.

Netstat

Aquesta utilitat serveix per a veure les connexions obertes en una màquina. Entre d'altres paràmetres mostra l'estat, les adreces i els ports de les connexions TCP i UDP. També mostra (veure figura 2.6.1) la quantitat de dades emmagatzemades en els buffers d'enviament/recepció de dades (la finestra de TCP).

```
[gerard@wl-gerard ]$ netstat
Active Internet connections (w/o servers)
Proto Recv-Q Send-Q Local Address               Foreign Address             State
tcp        0      0 localhost.localdomain:ipp   localhost.localdomain:35603 ESTABLISHED
tcp        1      0 wl-gerard.pic.es:35562      fpserv.fedoraproject.o:http CLOSE_WAIT
tcp        0      0 wl-gerard.pic.es:34241      ifae-s0.ifae.es:imap        ESTABLISHED
tcp        0      0 wl-gerard.pic.es:42346      fk-in-f164.google.com:http  ESTABLISHED
tcp        0      0 wl-gerard.pic.es:42345      fk-in-f164.google.com:http  ESTABLISHED
tcp        0      0 wl-gerard.pic.es:52121      ifae-s0.ifae.es:imap        ESTABLISHED
tcp        0      0 wl-gerard.pic.es:57151      fk-in-f104.google.com:http  ESTABLISHED
tcp        0      0 wl-gerard.pic.es:57152      fk-in-f104.google.com:http  ESTABLISHED
tcp        0      0 wl-gerard.pic.es:33697      mg-in-f18.google.com:http   ESTABLISHED
tcp        0      0 localhost.localdomain:35602 localhost.localdomain:ipp   TIME_WAIT
tcp    57352      0 localhost.localdomain:40331 localhost.localdomain:6498  ESTABLISHED
```

Figura 2.6.1: extracte de la sortida de netstat

Netstat és especialment útil alhora d'analitzar problemes amb tranferències sobre TCP.

TCPdump

Tcpdump mostra les capçaleres dels paquets que passen per una interfície de xarxa determinada, es pot veure un exemple de la seva sortida a la figura 2.6.2. És pot configurar de tal forma que només ens mostri un cert tipus de paquets, d'un orígen/destí determinats, etc. Per tal d'utilitzar-lo sense limitacions és necessari tenir permisos d'administrador (root) a la màquina. Es pot trobar l'aplicació i més informació a [14]

```
[root@wl-gerard octoshape]# tcpdump host google.es
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eth0, link-type EN10MB (Ethernet), capture size 96 bytes
14:57:05.464600 IP wl-gerard.pic.es.50094 > gv-in-f104.google.com.http: S
2349073111:2349073111(0) win 5840 <mss 1460>
14:57:05.543104 IP gv-in-f104.google.com.http > wl-gerard.pic.es.50094: S
1923158691:1923158691(0) ack 2349073112 win 8190 <mss 1460>
14:57:05.543225 IP wl-gerard.pic.es.50094 > gv-in-f104.google.com.http: . ack 1 win 5840
14:57:05.543459 IP wl-gerard.pic.es.50094 > gv-in-f104.google.com.http: P 1:528(527) ack 1
win 5840
14:57:05.621310 IP gv-in-f104.google.com.http > wl-gerard.pic.es.50094: . ack 528 win 6660
14:57:05.624548 IP gv-in-f104.google.com.http > wl-gerard.pic.es.50094: P 1:448(447) ack
528 win 6660
14:57:05.624576 IP wl-gerard.pic.es.50094 > gv-in-f104.google.com.http: . ack 448 win 6432
```

Figura 2.6.2: extracte de la sortida de tcpdump al realitzar una consulta a google.es

Per tal de poder capturar tot el trànsit que passa per la interfície, TCPdump la configura en mode promiscu: forçant al SO a no descartar cap paquet. Aquesta és una eina que funciona en mode consola, hi ha altres programes com ara WireShark que fan el mateix però sobre un entorn gràfic més amigable.

Nmap

Nmap (Network Mapper) és una eina per a la exploració de la xarxa i l'auditoria de seguretat. El que realment fa Nmap és escanejar (analitzar) ports de servidors mitjançant diversos tipus de paquet TCP/IP i així determinar quins serveis, S.O., firewalls, etc hi ha a la xarxa. En la figura 2.6.3 hi ha un exemple de la sortida generada al analitzar un servidor de la xarxa. Es pot trobar l'aplicació i més informació a [15]

```
[gerard@wl-gerard]$ nmap xxxx.pic.es

Starting Nmap 4.11 ( http://www.insecure.org/nmap/ ) at 2007-05-07 14:32 CEST
Interesting ports on services1.pic.es (193.146.196.x):
Not shown: 1667 closed ports
PORT      STATE SERVICE
22/tcp    open  ssh
53/tcp    open  domain
111/tcp   open  rpcbind
199/tcp   open  smux
705/tcp   open  unknown
724/tcp   open  unknown
868/tcp   open  unknown
873/tcp   open  rsync
907/tcp   open  unknown
1005/tcp  open  unknown
2049/tcp  open  nfs
32774/tcp open  sometimes-rpc11
32776/tcp open  sometimes-rpc15

Nmap finished: 1 IP address (1 host up) scanned in 6.731 seconds
```

Figura 2.6.3: Sortida generada al analitzar un servidor amb nmap

També és possible utilitzar Nmap per a consultar l'estat d'un servei (mirar si algú està escoltant al port o no), saber si un servidor està accessible dins la xarxa, etc.

2.6.2 Eines de mesura de rendiment

Hi ha una gran quantitat d'eines dedicades a la mesura de diferents paràmetres de rendiment. Un cop provades i comparades diverses utilitats (ttcp, nttcp, nuttcp, iperf, etc.), es va consultar la documentació d'altres centres *Tier-1* compromesos amb la xarxa LHC-OPN i vaig prendre la decisió d'utilitzar Iperf i Thrulay per a les proves del projecte.

Iperf

Iperf és una eina per a mesurar l'ample de banda màxim sobre TCP i UDP. Permet ajustar diversos paràmetres i característiques com ara la mida de la finestra TCP/buffers UDP, la mida màxima del segment (MSS de TCP), etc. Es pot trobar l'aplicació i més informació a [16]

Iperf informa de l'ample de banda utilitzat (TCP i UDP), el *delay jitter* (UDP) i els datagrames perduts (UDP).

També és possible utilitzar Iperf per a la generació d'un o varis flux de dades *bulk*, és a dir, generar soroll dins una xarxa o canal sobre UDP.

Aquesta ha estat una de les eines escollides alhora de fer les proves i la certificació de les connexions degut a que és l'estàndard "de facto" per a les proves de *throughput* dins la xarxa LHC-OPN.

Thrulay

Thrulay és una altre eina per a mesurar la capacitat d'una xarxa mitjançant l'enviament d'un flux TCP. Al igual que altres eines (com iperf, netperf, nuttcp, etc) thrulay informa periòdicament de l'ample de banda (throughput), la diferència amb les altres eines és que també informa del *round-trip delay*, és a dir, del temps d'enviament dels segments del flux de dades (quan triga un paquet des de l'origen al destí). A més la sortida de Thrulay és fàcil de tractar de forma automatitzada, per exemple amb gnuplot.

Al igual que Iperf, Thrulay també és capaç de realitzar proves enviant flux de dades sobre UDP. Es pot trobar l'aplicació i més informació a [17]

L'ús de Thrulay alhora de realitzar algunes proves es deu a que informa del round-trip delay i facilita la generació automàtica de gràfiques amb gnuplot, alhora que retorna uns valors quasibé idèntics als d'iperf.

2.7 Altres tecnologies utilitzades

En aquesta secció es tractaran tecnologies que pertanyen a alguna o vàries de les capes TCP/IP però que, pel seu grau d'especialització i rellevància dins el desenvolupament del PFC, és millor explicar separadament.

2.7.1 HSRP

Hot Standby Router Protocol (HSRP) és un protocol de redundància propietari de Cisco que fou dissenyat per a poder crear portes d'enllaç (d*efault gateway*) tolerants a fallides. HSRP està descrit en l'RFC 2281 i el seu homòleg estàndard de l'IETF, Virtual Router Redundancy Protocol (VRRP), a l'RFC 3768. Ambdues tecnologies són molt similars en quan a concepte, però no són compatibles entre si.

Tal i com es mostra en la figura 2.7.1 HSRP defineix una porta d'enllaç virtual (una interfície virtual). El protocol estableix una associació entre els encaminadors de la xarxa creant un encaminador (*router)* primari, el que té configurada una prioritat més alta, que actua com a encaminador virtual i és qui gestiona realment la xarxa com si la porta d'enllaç virtual fos realment una interfície pròpia. En el cas de que falli l'encaminador primari, el següent més prioritari començarà a gestionar la interfície virtual, evitant així qualsevol interrupció en la xarxa.

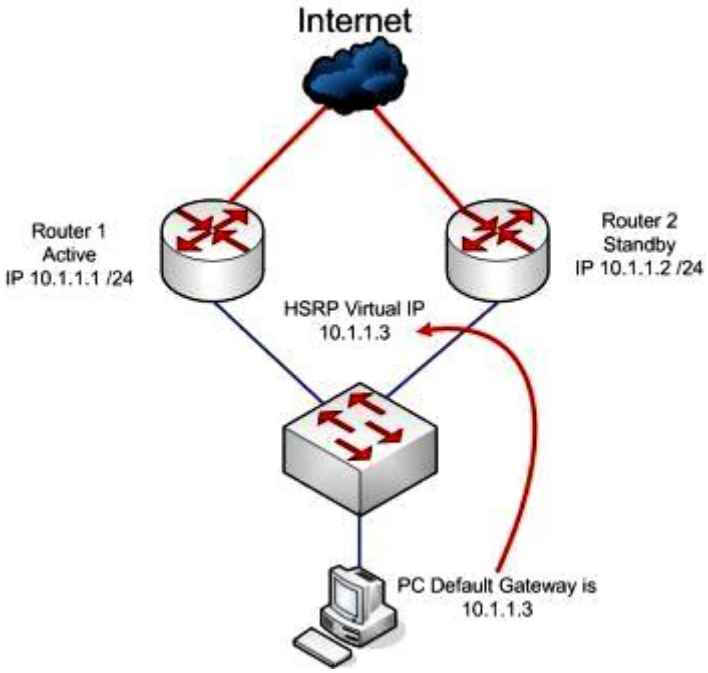


Figura 2.7.1: Exemple d'una porta d'enllaç redundada mitjançant HSRP

Cal notar que el protocol es limita a oferir redundància i en cap cas realitza tasques d'encaminament. Aquest és un protocol que té el seu marc d'actuació entre capa 2 i capa 3, per a obtenir més detalls del protocol HSRP consultar la bibliografia [18] i [19]

2.7.2 Spanning-tree protocol

Spanning Tree Protocol (STP) es un protocol de xarxa de la capa d'enllaç de dades (capa 2) que gestiona enllaços redundants (per exemple xarxes en anell). L'objectiu de STP és prevenir bucles infinits de repetició de dades en xarxes que presenten una configuració redundant.

STP és transparent als servidors dels usuaris. Hi ha dues versions d'STP que no són compatibles entre si: la original (DEC STP) y la estandarditzada per l'IEEE_802.1D.

Els bucles infinits apareixen quan hi ha rutes alternatives entre elements de la xarxa, com en la figura 2.7.2. Aquestes rutes alternatives són necessàries per a obtenir una xarxa fiable ja que al existir diversos enllaços es proporciona redundància i en el cas de que un enllaç falli, un altre pot seguir suportant el trànsit de la xarxa sense causar interrupcions.

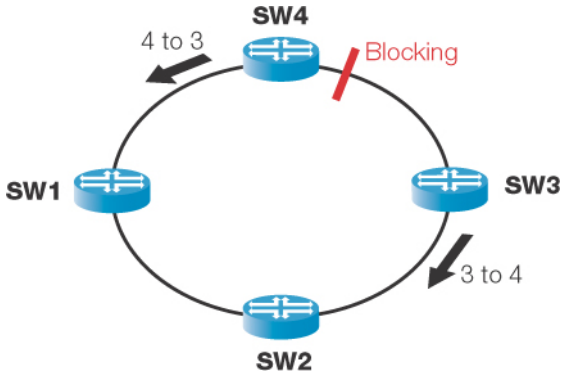


Figura 2.7.2: efecte de STP en una topologia d'anell

Per tal d'evitar els bucles infinits el que fa STP és permetre l'existència de cicles físics però crea una topologia lògica de la xarxa sense cicles, permetent en cada moment una única trajectòria activa entre dos dispositius de la xarxa i mantenint els camins redundants com a reserva per si el primari falla. L'elecció dels camins primaris es realitza en funció a una puntuació que es configura per a cada camí disponible, segons la seva prioritat.

Quan la topologia de la xarxa o la configuració de STP canvia, o si un segment de la xarxa redundant es considera inaccessible, l'algorisme convergeix i s'activa un nou comí. En el cas de que el protocol fallés seria possible que es crees un bucle a la xarxa, cosa que podria ocasionar la generació de trànsit infinit, col·lapsant la xarxa.

Existeixen diverses variants de l'Spanning Tree Protocol, degut principalment al temps que triga l'algorisme a convergir. Una de les principals variants és Rapid Spanning Tree Protocol (RSTP), els objectius del qual són:

- Disminuir el temps de convergència quan un enllaç falla de 30-60 segons a milisegons.
- Soportar xarxes exteses: d'un màxim de 256 ports interconectats en STP a 2048 connexions o 4096 ports en RSTP
- Retrocompatibilitat amb STP

2.7.3 Etherchannel/Link Aggregation (802.3ad)

Etherchannel i Link Aggregation (IEEE 802.3ad) són dues tecnologies independents però que persegueixen una mateixa finalitat: aconseguir incrementar l'ample de banda d'una connexió mitjançant l'ús de múltiples línies físiques (agregació). El marc d'actuació d'ambdues tecnologies es troba a la capa d'enllaç de dades, fent frontera amb la capa inferior (capa física).

Quan es crea un enllaç 802.3ad/Etherchannel el que es fa és crear una interfície virtual que agrega el trànsit generat en dues o més interfícies físiques, creant una topologia física com la de la figura 2.7.3. Així amb l'agregació de línies s'afegeix redundància i s'incrementa l'ample de banda de la connexió. És necessari configurar els elements en ambdues bandes de l'agregació de línies, que poden ser qualsevol tipus de dispositiu que suporti el protocol.

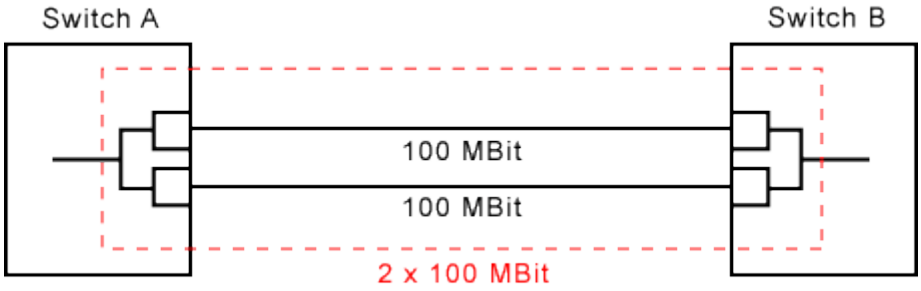


Figura 2.7.3: esquema físic d'agregació de dues línies de 100Mbps

Les diferències entre ambdós protocols d'agregació de línies radiquen en el fet que EtherChannel és un protocol propietari de Cisco i 802.3ad és un estàndard obert. En la taula de la figura 2.7.4 es poden observar les principals diferències entre els protocols.

EtherChannel	**802.3ad**
Requereix configuració directa al Switch o via PagP[14]	Requereix configuració direta al switch o via LACP[15]
Suporta diversos modes de distribució de la càrrega entre les línies de l'agregació	Suporta només un mode de distribució de la càrrega entre les línies de l'agregació

Figura 2.7.4: principals diferències entre EtherChannel i 802.3ad

La majoria d'encaminadors i switch Cisco suporten l'estàndard 802.3ad utilitzant LACP i amb un mode de distribució de càrrega EtherChannel idèntic al definit a 802.3ad, aconseguint així la compatibilitat entre ambdues tecnologies.

14 Port Aggregation Protocol: un protocol propietari de Cisco per a negociar l'agregació de línies
15 Link Aggregation Control Protocol: un protocol estàndard per a negociar l'agregació de línies 802.3ad

2.7.4 Bonding

Bonding és l'homòleg de 802.3ad per a linux i ve integrat com a mòdul en quasibé en totes les distribucions. De fet bonding és més complex que Etherchannel/802.3ad i permet diversos modes de funcionament:

0. Balance-rr (mode per defecte). Necessita suport per part del switch. Envia les dades segons una política *round-robin*. Proporciona tolerància a fallides i balanceig de càrrega (augment de l'ample de banda).

1. Active-Backup. No necessita suport per part del switch. Només una de les NIC associada a la interfície de bonding estarà activa en cada moment, en cas de pèrdua de connexió es desactiva i el trànsit comença a fluir per una altre NIC. Proporciona tolerància a fallides.

2. Balance-XOR. Necessita suport per part del switch. Transmet les dades per una o altre NIC segons una política de hash. Proporciona tolerància a fallides i balanceig de càrrega (augment de l'ample de banda).

3. Broadcast. Aquest mode necessita configuracions de capa 2 específiques (per exemple dues xarxes aïllades). Envia totes els dades per totes les NIC del bonding. Proporciona tolerància a fallides.

4. 802.3ad. Necessita suport per part del switch. Crea una agregació de línies dinàmica segons l'estàndard IEEE 802.3ad. Proporciona tolerància a fallides i balanceig de càrrega (augment de l'ample de banda).

5. Balance-tlb. No necessita suport per part del switch. El trànsit de sortida es balanceja segons la càrrega de cada NIC, tot el tràfic d'entrada va per la NIC primària. Proporciona tolerància a fallides i balanceig de càrrega (augment de l'ample de banda).

6. Balance-alb. No necessita suport per part del switch. Igual que tlb però el trànsit d'entrada és també balancejat. Proporciona tolerància a fallides i balanceig de càrrega (augment de l'ample de banda).

Que un mode de funcionament de bonding necessiti suport per part del switch implica que en aquest cal configurar-hi una agregació de línies del tipus 802.3ad o bé EtherChannel.

Es pot obtenir més informació sobre bonding a [20]

2.7.5 Policy Route Map

Els Policy Route Map, o simplement route-map, són l'eina bàsica per a definir encaminaments basats en polítiques diferents de l'encaminament estàndard, permetent alterar a voluntat el funcionament de l'encaminament IP estàndard. Mitjançant route-map és possible que el trànsit d'una interfície d'un servidor (una IP) surti per dos camins

diferents encara que vagi al mateix destí. El marc d'actuació d'aquesta tecnologia es troba en capa de xarxa (capa 3), lleugerament per sota de l'encaminament IP estàndard (per tal de poder-ne alterar el comportament).

El motiu pel qual aquest tipus de comportament pot ser desitjable és que ens podria interessar disposar d'una connexió cara, però amb molt poca latència, per algunes finalitats (per exemple VeuIP) i per a la resta de transferències utilitzar la connexió estàndard, de cost inferior, com es pot veure en l'exemple de la figura 2.7.5.

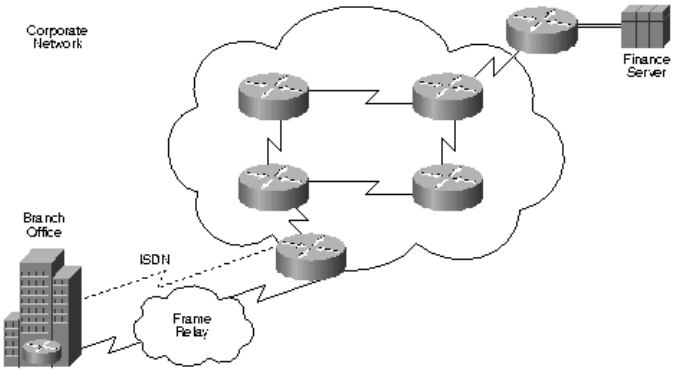


Figura 2.7.5: exemple de Policy Based Routing entre una conexió ISDN i Frame Relay

L'encaminament basat en polítiques també ens pot servir per a fer balanceig del trànsit a la xarxa entre dues connexions físiques independents de forma transparent als usuaris.

3 Capítol 3 – Desplegament circuit dedicat a 1 Gbps

En aquest capítol s'abordarà la primera fase del desenvolupament del projecte, que correspon al desplegament de la connexió PIC-CERN sobre el circuit dedicat a 1 Gbps. Tal i com s'ha explicat en el primer capítol, subsecció *Metodologia del Projecte*, el desplegament del circuit dedicat a 1 Gbps s'ha organitzat en quatre etapes:

- Etapa 1: recollida d'especificacions, anàlisi de necessitats i generació de possibles solucions. Aquesta fase es troba documentada en la secció primera d'aquest capítol, juntament amb una comparativa entre les solucions proposades.

- Etapa 2: consens d'una solució i generació d'un pla per al desplegament del circuit dedicat a 1 Gbps, descrit en la segona secció del capítol.

- Etapa 3: execució del pla, resumit en la tercera secció del capítol, on es detalla el procés i les diferents incidències ocorregudes al llarg del desplegament de la connexió.

- Etapa 4: certificació de la connexió. Aquesta fase es troba documentada en la quarta secció.

3.1 Estudi de solucions per a la integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps

La fita d'aquesta primera secció és descriure les solucions més adequades per tal d'integrar els servidors del PIC a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps, donant un nivell de detall suficient per a poder decidir quina és la solució a implementar.

En primer lloc es donarà una visió de la situació inicial del PIC, explicant d'on es parteix i on es vol arribar en aquesta primera fase del projecte. Un cop clars els marcs d'actuació es plantejaran i compararan tres possibles solucions per a la integració dels servidors del PIC i la xarxa LHC-OPN.

3.1.1 Descripció de la situació inicial

En el punt de partida del projecte, al PIC es disposa d'una connexió a 1 Gbps associada a la subxarxa 193.145.217.0/24 i proporcionada per RedIRIS, a on està connectada des d'equipament propietat de l'Anella Científica. La connexió passà les proves de rendiment, realitzades per RedIRIS, el 16 de maig del 2006 i resta en desús a l'espera d'ésser integrada a la xarxa del PIC.

En la figura 3.1.1 hi ha un diagrama de la connectivitat inicial (desembre 2006) a nivell 3 d'aquesta connexió dedicada.

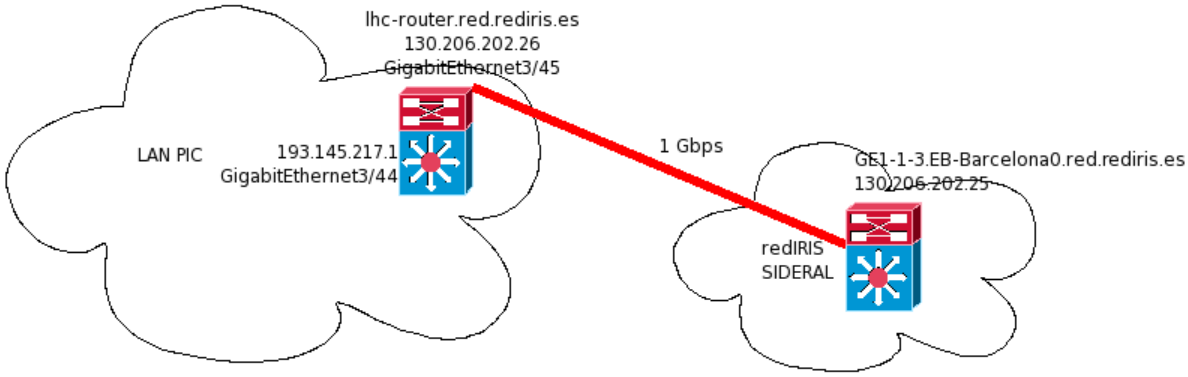


Figura 3.1.1: diagrama de la connectivitat inicial del circuit dedicat de 1 Gbps, és una connexió directa a la xarxa nacional de RedIRIS

A RedIRIS la xarxa està connectada al *Servicio de Interconexión de Redes de Area Local* (SIDERAL) i, per tant, a la Internet global. Al PIC a la xarxa 193.145.217.0 no hi ha cap servidor a excepció de l'encaminador Cisco6509, amb la IP 193.145.217.1. L'encaminador només deixa accedir a aquesta xarxa des del port GigabitEthernet3/44.

Actualment els servidors del PIC que pertanyen a la LHC-OPN s'hi connecten mitjançant la VLAN236, una connexió compartida a Internet. La connexió es realitza des de la xarxa 193.146.196.0/22, que pertany a la LHC-OPN, cosa que implica que estigui inclosa dins el *route-set* RS-LHCOPN de RIPE[1].

Dins del PIC tots els servidors (LHC-OPN[2] i no-LHC-OPN[3]) formen part de la mateixa xarxa (193.146.196.0/22) i VLAN (VLAN100). La xarxa no conté subxarxes, només hi ha un encaminador (Cisco6509) i és possible la comunicació a nivell2 entre tots els servidors del PIC. En aquest moment es disposa d'una xarxa GigabitEhternet en topologia d'estrella, per on passa tot el trànsit de la xarxa, tant LHC-OPN com no-LHC-OPN. En el següent apartat d'aquesta subsecció es pot trobar una anàlisi de la càrrega inicial de la xarxa.

En el diagrama de la figura 3.1.2 es representa la interconnexió dels servidors LHC-OPN del PIC i els flux de dades entre els servidors LHC-OPN i no-LHC-OPN inicials, previ a l'implementació d'alguna de les solucions proposades a continuació. Com que en cap de les solucions que es plantejaran s'afecta a les connexions dels servidors no-LHC-OPN amb la resta de xarxes (Internet, VPNs, etc), aquestes no seran mostrades en la seva totalitat.

1 Es pot consultar l'estat del *route-set* al cercador de RIPE: http://www.ripe.net/whois

2 Els servidors LHC-OPN són màquines que necessiten formar part i intercanviar dades directament amb la xarxa LHC-OPN i els seus membres.

3 Els servidors no-LHC-OPN són aquelles màquines complementàries als servidors LHC-OPN, és a dir, que no han d'intercanviar dades directament amb la xarxa LHC-OPN i els seus membres.

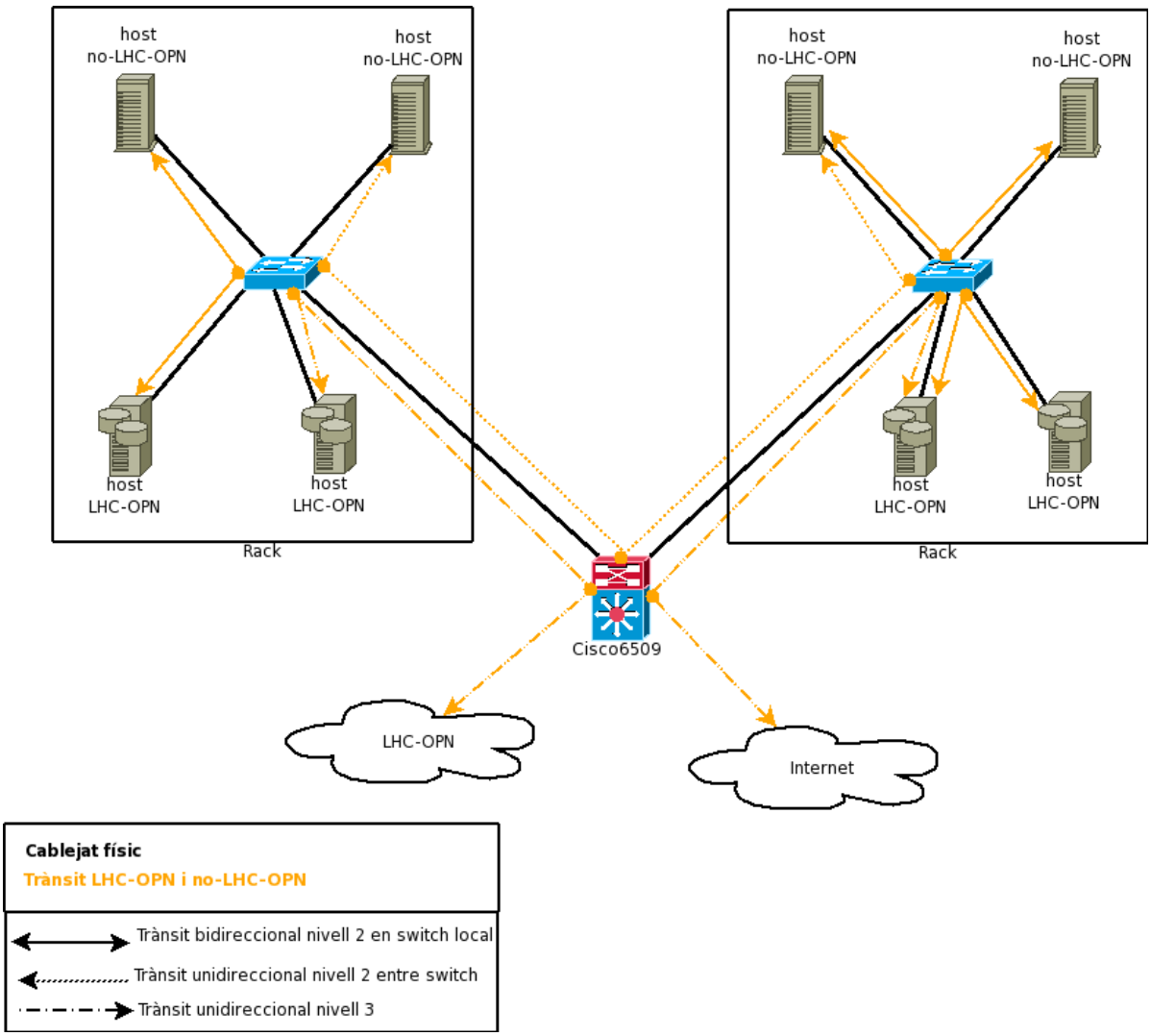
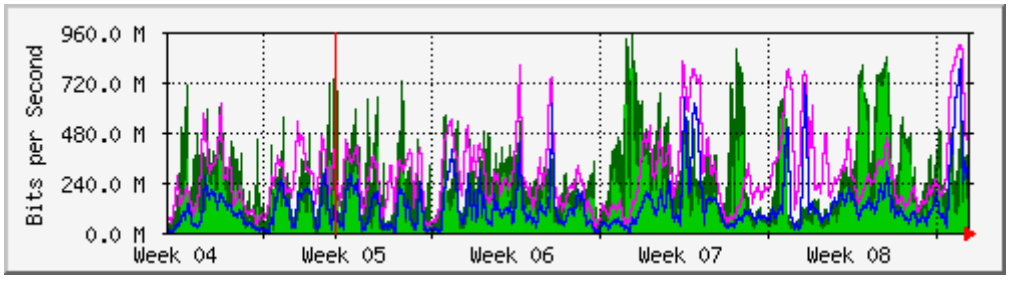


Figura 3.1.2: interconnexió dels servidors LHC-OPN del PIC i els flux de dades entre els servidors LHC-OPN i no-LHC-OPN inicials. Com es pot observar no hi ha cap distinció ni priorització entre tipus de trànsit diferent (LHC-OPN vs no-LHC-OPN)

Anàlisi de la càrrega inicial de la xarxa

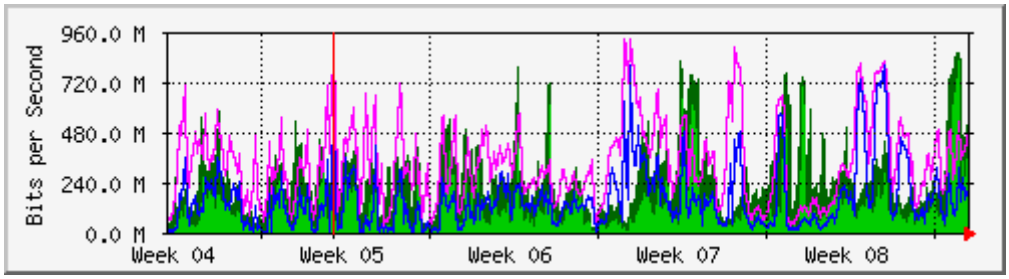
Analitzant la càrrega del Cisco6509, centre de l'estrella de la xarxa, podem veure que el trànsit destinat/originat a la VLAN100 (LAN del PIC, figura 3.1.3) té com a origen/destí principalment la VLAN236.



Màxima In:755.8 Mb/s (75.6%) Mitjana In:78.3 Mb/s (7.8%) Actual In:28.9 Mb/s (2.9%)
Màxima Out:782.5 Mb/s (78.2%) Mitjana Out:50.8 Mb/s (5.1%) Actual Out:73.1 Mb/s (7.3%)

Figura 3.1.3: Gràfica mensual d'entrada/sortida de la VLAN100, la mitjana d'ús és d'un 5-8% amb pics de fins al 96%

Actualment la VLAN236 (veure figura 3.1.4) té una capacitat màxima d'1 Gbps, del qual només se'n garanteix un 60% degut a que el canal es comparteix amb altres entitats com ara la UAB o l'IFAE. Amb la posta en marxa del circuit dedicat s'espera un rendiment lleugerament superior a l'actual, ja que serà d'ús intensiu i exclusiu LHC-OPN.



Màxima In:773.3 Mb/s (77.3%) Mitjana In:46.6 Mb/s (4.7%) Actual In:72.7 Mb/s (7.3%)
Màxima Out:726.8 Mb/s (72.7%) Mitjana Out:77.0 Mb/s (7.7%) Actual Out:30.3 Mb/s (3.0%)

Figura 3.1.4: Gràfica mensual d'entrada/sortida de la VLAN236, pràcticament complementària amb la gràfica 3.1.3

A nivell de la LAN, la nova càrrega és assumible amb la infraestructura actual només si es menysprea el trànsit aliè a la LHC-OPN o es realitza una repartició acurada, cap a diferents servidors de la xarxa, de l'ample de banda. En cas contrari no es pot garantir la disponibilitat d'1 Gbps, ja que és la capacitat màxima de la xarxa per a una connexió punt a punt dins la LAN actual del PIC.

Donat que la connexió dels switch fins al Cisco6509 és d'1 Gbps, per tal de disposar de 1 Gbps real en un servidor serà necessària la connexió dels servidors directament al Cisco6509 o bé dedicar el switch al servidor objectiu.

Per tal de disposar d'una velocitat superior per a una connexió punt a punt, permetent així la convivència amb trànsit no-LHC-OPN, és necessària la realització d'un estudi per a incrementar la velocitat de la LAN, actualitzant el HW o amb algun sistema d'agregació de línies (per a més informació veure subsecció 2.7.3).

3.1.2 Especificacions del sistema objectiu

En aquesta subsecció s'exposen les restriccions imposades i els objectius que es persegueixen amb el desplegament del circuit dedicat de 1 Gbps i la integració amb la xarxa LHC-OPN.

Cal tenir en compte que sobre el circuit dedicat d'1 Gbps ja es disposa d'una connexió operativa, a nivell 3, amb el CERN i, en conseqüència, amb la xarxa LHC-OPN. Així doncs aquesta primera fase del projecte es centra en la integració dels servidors del PIC a la nova xarxa LHC-OPN sobre la connexió ja existent en el circuit dedicat de 1 Gbps.

Interconnexió objectiu

El diagrama de la figura 3.1.5 representa de forma abstracta la interconnexió objectiu dels

servidors LHC-OPN del PIC. Com es pot observar l'objectiu principal és separar el trànsit LHC-OPN i no-LHC-OPN en tot moment, d'extrem a extrem, però permetent el trànsit local entre servidors LHC-OPN i no-LHC-OPN.

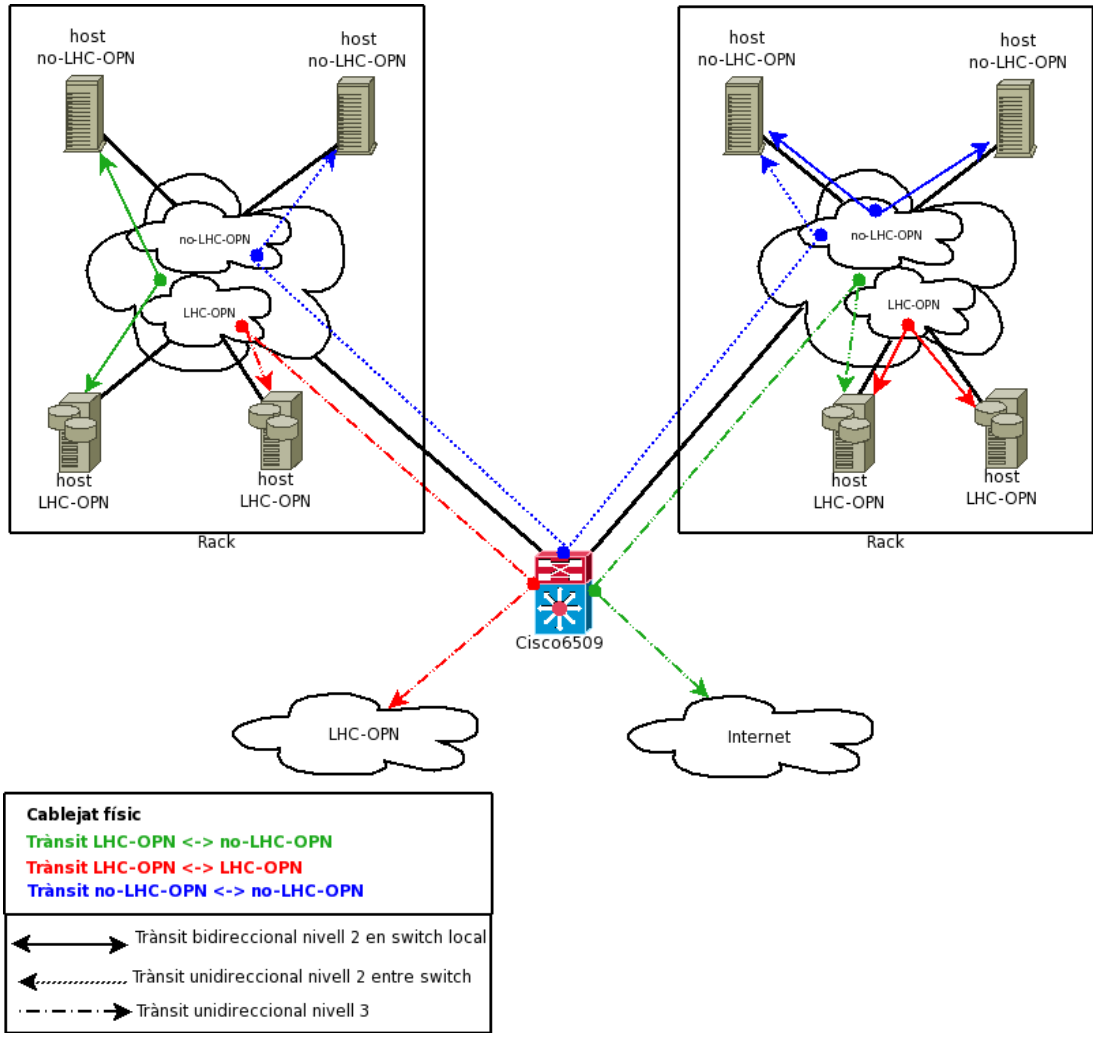


Figura 3.1.5: interconnexió dels servidors LHC-OPN i no-LHC-OPN objectiu, on es separa el trànsit LHC-OPN i no-LHC-OPN

Restriccions i Assumpcions

En la integració dels servidors del PIC a la xarxa LHC-OPN sobre el circuit dedicat s'ha de garantir el compliment dels requeriments imposats per la LHC-OPN en la segona versió (30/07/2005) del document *LHC Tier-0 to Tier-1 High-Level Network Architecture*[21]. A grans trets el document es pot resumir en els següents punts:

1. Ús dedicat al trànsit de *dades* i *informació de control* dins dels experiments LHC. Tier0-Tier1
2. Funciona sobre IPv4 (migració futura a IPv6)
3. S'ha de garantir una MTU mínima de 9000 bytes
4. No pot haver-hi firewalls dins la OPN, per tal d'evitar colls d'ampolla
5. S'ha de restringir al màxim la connectivitat de les màquines connectades a la xarxa LHC-

OPN mitjançant ACLs

6. Ús d'adreces públiques (no IP privades) en els servidors connectats a la xarxa LHC-OPN
7. Ús de pocs CIDRs (idealment 1, això implica adreçament contigu dels servidors LHC-OPN)
8. Adreces LHC-OPN (CIDRs) dedicades per al trànsit LHC-OPN
9. Cal notificar els CIDRs utilitzats per a LHC-OPN al T0 (CERN) per tal d'actualitzar *route-set* RS-LHCOPN
10. Per al BGP cal utilitzar l'AS del nostre NREN (redIRIS -> AS766) o bé un de propi
11. No es recomana l'ús de rutes estàtiques (en BGP)
12. Està prohibit usar rutes per defecte (*default gateway*) cap a la xarxa LHC-OPN
13. Els únics rangs vàlids per a la OPN són, estrictament, els indicats a http://www.ripe.net/perl/whois?&searchtext=rs-LHCOPN. Es recomana restringir l'accés de/als servidors via ACLs.
14. Es recomana la disponibilitat d'un camí alternatiu, per redundància i si és possible evitant la VLAN236 (per evitar interferències amb el trànsit convencional d'Internet i altres projectes)
15. L'inici del trànsit LHC-OPN de producció previst per al juny del 2007

Per al desenvolupament d'aquesta primera fase del projecte s'assumeix que per a la comunicació amb la xarxa LHC-OPN via el nou circuit dedicat a 1 Gbps s'utilitzarà la xarxa 193.145.217.0/24.

en el disseny de les solucions cal tenir en compte que la LAN (xarxa local) no pot constituir un coll d'ampolla per a la connexió d'1 Gbps. Tampoc es pot perjudicar el rendiment actual en les transferències entre els servidors de la xarxa.

3.1.3 Possibles solucions

En primer lloc cal dir que les principals conseqüències de les diferents solucions afecten principalment les comunicacions entre els servidors no-LHC-OPN <-> LHC-OPN, en ambdós sentits. També és important tenir en compte els recursos disponibles, que són essencialment:

- Switch redundant Cisco Catalyst 6509 amb 1x48 ports GigabitEthernet i 2x4 TenGigabitEthernet
- Dos o més servidors amb >=2 tarja de xarxa GigabitEthernet connectada a un switch dedicat o directament al Cisco6509
- Línia d'1 Gbps subministrada per redIRIS i incorporada a SIDERAL mitjançant 193.145.217.0/24, amb connectivitat al CERN

Es poden preveure alguns recursos que seran necessaris però que inicialment no estan disponibles, així doncs és important iniciar les gestions per tal d'obtenir:

- Metodologia per afegir servidors de la xarxa 193.145.217.0/24 al DNS
- Metodologia per al testeig de les línies des del CERN
- Metodologia per a l'adhesió a la LHC-OPN (dins el route-set RS-LHCOPN)

Un cop definit el marc d'actuació de la primera fase del projecte, els objectius i les restriccions es proposen tres solucions ("Dues IPs", "Una IP amb encaminament" i "Una IP amb entrega directa").

○ Solució "Dues IPs"

Descripció

Aquesta solució consisteix en afegir als servidors LHC-OPN una segona interfície amb una IP del rang LHC-OPN. Així el trànsit local es farà per entrega directa[4] i el trànsit LHC-OPN per una xarxa dedicada, via la NIC i el maquinari (Hardware) d'aquesta segona interfície.

Viabilitat tècnica i operativa

Els servidors que hagin d'accedir/pertànyer a ambdues xarxes tindran com a mínim dues IP i dos noms (hostname), un no-LHC-OPN i un de la LHC-OPN.

Per a la comunicació amb servidors no-LHC-OPN s'utilitzarà la interfície no-LHC-OPN, tant a nivell local com a la Internet, quedant la interfície LHC-OPN dedicada de forma exclusiva a la comunicació dins la xarxa LHC-OPN.

Aquesta solució implica canvis en l'encaminador i en els servidors que formen la LHC-OPN.

Adaptacions necessàries

Les accions a realitzar en l'encaminador seran:

- Configurar com a xarxa local la subxarxa LHC-OPN local (193.145.217.0/24)
- Crear la VLAN LHC-OPN (VLAN 222) o bé Acceptar/Afegir dins la VLAN100 a la LHC-OPN.
- Configurar el Cisco com a porta d'enllaç (*gateway)* per a la LHC-OPN
- Configurar les ACLs d'entrada i sortida necessàries per a la connexió LHC-OPN

En els servidors el canvi es pot fer de dues formes, les quals poden coexistir en servidors diferents de la mateixa xarxa:

1. Addició d'una NIC dedicada. Opció recomanada

2. Creació d'una interfície virtual sobre una NIC ja existent

En el cas de disposar de NIC dedicades, es recomana la separació física de la xarxa LHC-OPN i no-LHC-OPN des de l'encaminador fins al servidor, així com la creació d'una VLAN per a la xarxa LHC-OPN (VLAN 222), garantint d'aquesta manera un rendiment

4 Recordem que el fet que es realitzi entrega directa significa que si dos servidors estan en el mateix switch els paquets s'entregaran a nivell 2, mitjançant l'adreça MAC i sense la necessitat de passar per l'encaminador. Si els servidors estan en switch diferents els paquets seguiran la ruta determinada pel protocol a nivell 2, en el nostre cas hauran de passar pel switch central, que és el router Cisco6509, i anar al switch on hi ha el servidor de destí.

òptim i capacitats de monitorització per xarxa.

En el cas de servidors pertanyents a la xarxa no-LHC-OPN i LHC-OPN, amb aquesta solució, un únic servidor estaria utilitzant dues adreces IP públiques, tanmateix es preveu que el número de servidors d'aquest tipus sigui reduït.

Comunicació de servidors LHC-OPN amb no-LHC-OPN

L'accés als servidors LHC-OPN locals des de la xarxa no-LHC-OPN local és possible, a nivell 2, accedint a la interfície no-LHC-OPN del servidor. Altrament s'hi hauria d'accedir des d'un servidor de la LHC-OPN local.

La comunicació entre els servidors LHC-OPN locals i els servidors no-LHC-OPN externs (p.e. Internet) per part dels servidors LHC-OPN es realitza per la interfície no-LHC-OPN via la VLAN236. Si hi ha algun servidor que no necessita realitzar connexions fora de la LHC-OPN, aquest pot prescindir de la interfície no-LHC-OPN o bé, en cas d'ús esporàdic, tenir-la com a interfície virtual.

A la figura 3.1.6 es pot veure una possibilitat de les rutes dels flux de dades creats entre diferents servidors seguint aquesta solució.

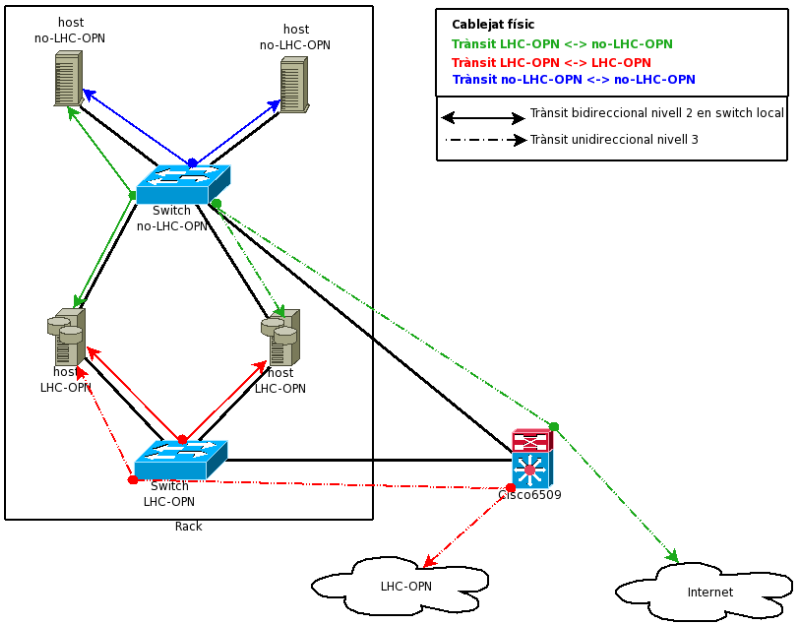


Figura 3.1.6: rutes dels flux de dades segons la solució “Dues IPs”. Cada interfície gestiona un tipus de trànsit, separant LHC-OPN i no-LHC-OPN

Viabilitat econòmica

- Donat que la majoria dels servidors del PIC disposen de dues targes GigabitEthernet, la solució proposada no suposa l'adquisició de nou Hardware, tot i que seria necessària la instal·lació de cablejat i, en alguns casos, d'alguna tarja GigabitEthernet i/o switch.

Metodologia per afegir un servidor a la LHC-OPN

1. a) Si NIC dedicada (xarxes físicament separades): Instal·lar NIC per a la LHC-OPN
   b) Sinó: Crear interfície virtual LHC-OPN
2. a) Si NIC dedicada i VLAN LHC-OPN (xarxes lògicament separades): Configurar VLAN LHC-OPN en switch
   b) Si interfície virtual i VLAN: Configurar tagging 802.1q en les interfícies del servidor
   c) Sinó: anar al següent pas
3. Afegir el servidor LHC-OPN al DNS (ambdues IP/hostname)
4. Configurar IP LHC-OPN a la interfície LHC-OPN del servidor
5. Afegir rutes LHC-OPN al servidor, aquest procés es pot automatitzar mitjançant l'script presentat en l'annex B
6. a) Si NIC dedicada: Connectar interfície LHC-OPN al switch
   b) Sinó: Activar interfície virtual LHC-OPN

○ Solució “Una IP amb encaminament”

Descripció

La solució d'una IP amb encaminament consisteix en canviar la IP dels servidors LHC-OPN, afegint-los la xarxa no-LHC-OPN local com a xarxa per a entrega directa i fent que l'encaminador gestioni les respostes dels servidors no-LHC-OPN, evitant així modificar la configuració de xarxa d'aquests últims.

Viabilitat tècnica i operativa

Aquesta solució implica canvis en l'encaminador i en els servidors que formin la xarxa LHC-OPN.

Adaptacions necessàries

Les accions a realitzar en l'encaminador seran:

- Configurar com a xarxa local la subxarxa LHC-OPN local
- Crear la VLAN LHC-OPN o bé Acceptar/Afegir dins la VLAN100 a la LHC-OPN. Permetent la intercomunicació entre VLANs
- Configurar com a gateway per a la LHC-OPN
- ACLs d'entrada i sortida per a la connexió LHC-OPN

Als servidors de la LHC-OPN se'ls canviarà la IP no-LHC-OPN per una de la xarxa LHC-OPN local (193.145.217.0/24), a més s'afegirà una regla d'encaminament (*routing)* per tal de fer entrega directa als servidors no-LHC-OPN: *route add -net 193.146.196.0/22 dev eth0*

És possible la optimització de la xarxa a nivell físic creant una infraestructura dedicada LHC-OPN des dels servidors LHC-OPN fins a l'encaminador. Per a realitzar una separació lògica via VLAN s'hauria de realitzar *tagging* en el propi servidor segons l'adreça de destí dels paquets, podent així establir prioritats segons VLAN en el switch.

Comunicació dels servidors LHC-OPN amb no-LHC-OPN

El trànsit dels servidors no-LHC-OPN locals cap als servidors LHC-OPN locals anirà a l'encaminador (*router*) i aquest el reenviarà (encaminament a nivell 3) cap al servidor corresponent. En el sentit contrari (LHC-OPN locals -> no-LHC-OPN locals) els paquets s'enviaran per entrega directa.

Per a la comunicació entre els servidors LHC-OPN locals i els servidors no-LHC-OPN externs (p.e. Internet) és necessari muntar un sistema de NAT que doni accés via la VLAN236. Això és degut a les restriccions 1, 5 i 8 de la LHC-OPN (3.1.2, apartat Restriccions i Assumpcions), que fan inviable aquesta comunicació mitjançant el circuit de 1 Gbps.

En el cas de que algun servidor es canviï de xarxa, que passi de LHC-OPN a no-LHC-OPN o viceversa, si hi ha algun servidor que ha de continuar accedint-hi caldrà tenir-ho en compte i rectificar-hi la IP/hostname.

A la figura 3.1.7 es pot veure les rutes dels flux de dades creats entre diferents servidors seguint aquesta solució, cal fixar-se en les connexions del servidor no-LHC-OPN cap al LHC-OPN (verd discontinu).

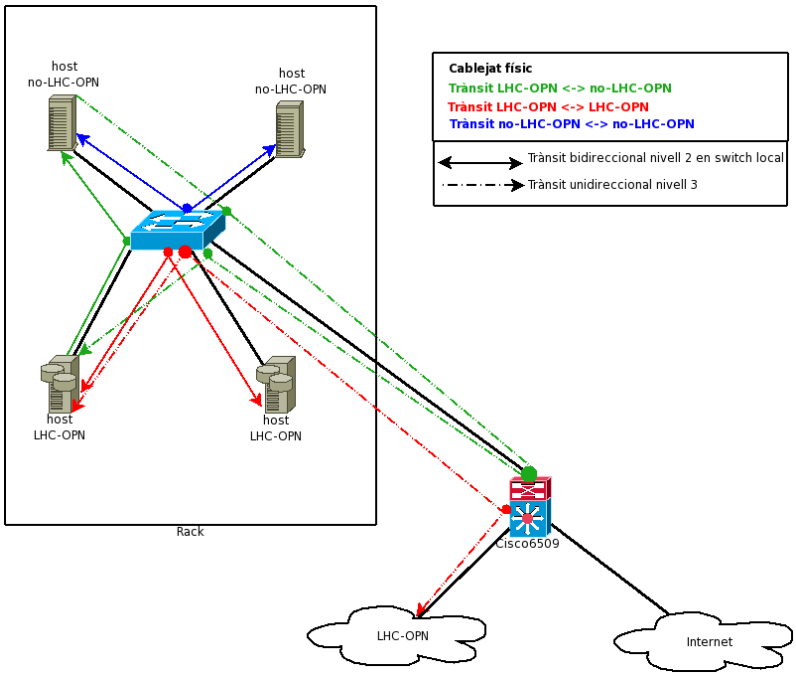


Figura 3.1.7: rutes dels flux de dades segons solució d'una IP amb encaminament. Es pot observar com l'encaminador cisco és que reb el trànsit no-LHC-OPN->LHC-OPN i el reenvia als servidors de la xarxa LHC-OPN locals.

Viabilitat econòmica

- Creació d'un sistema de NAT d'altes prestacions
- Increment de l'ample de banda de les connexions entre els switch i l'encaminador

Metodologia per afegir un servidor a la LHC-OPN

1. Afegir el servidor LHC-OPN al DNS
2. Configurar IP LHC-OPN a la interfície del servidor
3. a) Si VLAN (xarxes lògicament separades): Configurar tagging VLAN LHC-OPN/ VLAN100 en el servidor
   b) Sinó: anar al següent pas
4. Configurar ruta d'entrega directa a servidors no-LHC-OPN i default gateway via sistema NAT

○ Solució "Una IP amb entrega directa"

Descripció

Aquesta solució és similar a l'anterior (solució "Una IP amb encaminament", en l'apartat anterior), però permet realitzar tota la comunicació local entre servidors LHC-OPN i no-LHC-OPN a nivell 2, millorant la comunicació dels servidors que comparteixen switch.

Es proposa fer conviure els servidors LHC-OPN i no-LHC-OPN dins d'una mateixa xarxa, comunicant-se entre ells mitjançant entrega directa, exactament igual que si els diferents servidors estiguessin dins la mateixa xarxa lògica i física.

Viabilitat tècnica i operativa

Es configurarà tots els servidors que ho necessitin per tal que aquests vegin les subxarxes LHC-OPN (193.145.217.0/24) i no-LHC-OPN (193.146.196.0/22) com a xarxes locals d'entrega directa, comunicant-se a nivell 2.

Aquesta solució implica canvis en la configuració de l'encaminador i en tots els servidors del PIC que hagin d'accedir o ser accedits des de la LHC-OPN.

Adaptacions necessàries

Les accions a realitzar en l'encaminador seran:

- Configurar com a xarxa local la subxarxa LHC-OPN local
- Crear la VLAN LHC-OPN o bé Acceptar/Afegir dins la VLAN100 a la LHC-OPN. Permetent la intercomunicació entre VLANs
- Configurar com a gateway per a la LHC-OPN
- ACLs d'entrada i sortida per a la connexió LHC-OPN

Als servidors de la LHC-OPN se'ls canviarà la IP no-LHC-OPN per una de la xarxa LHC-

OPN local (193.145.217.0/24).

A tots els servidors del PIC que hagin d'accedir o ser accedits des de la LHC-OPN se'ls ha d'afegir una nova ruta, indicant al sistema que ha de fer entrega local a la xarxa amb la que comparteix medi físic:

Servidors no-LHC-OPN: *route add -net 193.145.217.0/24 dev eth0*

Servidors LHC-OPN: *route add -net 193.146.196.0/22 dev eth0*

És possible la optimització de la xarxa a nivell físic creant una infraestructura dedicada LHC-OPN des dels servidors LHC-OPN fins a l'encaminador. Per a realitzar una separació lògica via VLAN s'hauria de realitzar tagging en el propi servidor segons l'adreça de destí dels paquets, podent així establir prioritats segons VLAN en el switch.

Comunicació dels servidors LHC-OPN amb no-LHC-OPN

La comunicació entre els servidors locals LHC-OPN i no-LHC-OPN es farà per entrega directa. Per a la comunicació entre els servidors LHC-OPN locals i els servidors no-LHC-OPN externs (p.e. Internet), és necessari muntar un sistema de NAT que doni accés via la VLAN236. Això és degut a les restriccions 1, 5 i 8 de la xarxa LHC-OPN (apartat *Restriccions i Assumpcions* de 3.1.2), que fan inviable aquesta comunicació mitjançant el circuit de 1 Gbps.

En el cas de que algun servidor es canviï de xarxa, que passi de LHC-OPN a no-LHC-OPN o viceversa, si hi ha algun servidor que ha de continuar accedint-hi caldrà tenir-ho en compte i rectificar-hi la IP/hostname.

A la figura 3.1.8 es pot veure les rutes dels flux de dades creats entre diferents servidors seguint aquesta solució.

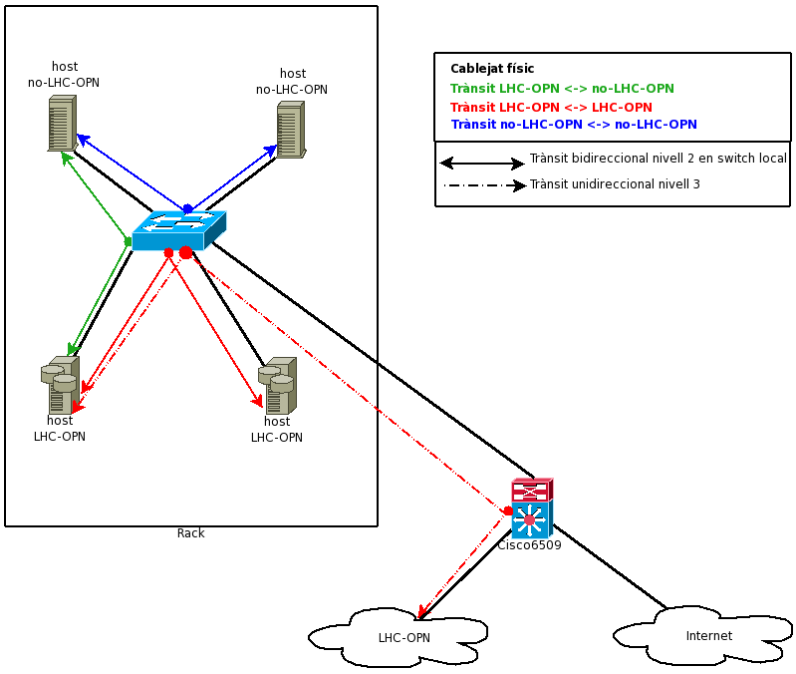


Figura 3.1.8: rutes dels flux de dades segons solució d'una IP amb entrega directa, a diferència de la solució anterior

en aquest cas el trànsit es realitza de forma local, sense necessitat d'encaminament.

Viabilitat econòmica

- Creació d'un sistema de NAT d'altes prestacions

Metodologia per afegir un servidor a la LHC-OPN

1. Afegir el servidor LHC-OPN al DNS
2. Configurar IP LHC-OPN a la interfície del servidor
3. a) Si VLAN (xarxes lògicament separades): Configurar tagging VLAN LHC-OPN/ VLAN100 en el servidor
   b) Sinó: anar al següent pas
4. Configurar ruta d'entrega directa a servidors no-LHC-OPN i porta d'enllaç via un sistema NAT

3.1.4 Comparativa de solucions

En la taula de la figura 3.1.9 es pot observar una comparativa de les diferents solucions. S'han marcat en verd les propietats guanyadores de cada solució, essent la solució “Dues IPs” la recomanada.

	Aprofitament de l'adreçament IP	***Escalabilitat ample de banda LHC-OPN***	***Separació física del trànsit entre servidors LHC-OPN i no-LHC-OPN***	***Separació lògica del trànsit LHC-OPN i NO-LHC-OPN***	***Comunicació entre servidors LHC-OPN <-> no-LHC-OPN en el mateix switch***
Dues IPs	NO (2 IP per servidor)	Switch dedicat	SI	VLAN en switch / servidor (si IF virtual)	Nivell 2 total
Una IP amb encaminament	SI	Switch compartit	NO	VLAN en servidor	Nivell 2 parcial (des de no-LHC-OPN sempre passa per l'encaminador)
Una IP amb entrega directa	SI	Switch compartit	NO	VLAN en servidor	Nivell 2 total

Figura 3.1.9: Taula comparativa de les solucions diferents solucions. La solució “Dues IPs” és la recomanada i que aporta més avantatges

Que el trànsit entre servidors LHC-OPN i no-LHC-OPN estigui separat i que els switch sigui dedicat és positiu perque permet independitzar ambdues xarxes i dóna més possibilitats alhora de créixer.

El fet de realitzar les comunicacions a nivell 2 significa que no és necessari passar per l'encaminador, cosa positiva que millora l'eficiència dels diferents recursos de la xarxa.

3.2 Pla d'implementació per a la integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps

L'objectiu d'aquesta segona secció és documentar la segona etapa de la primera fase del projecte, on es defineix un pla per a proporcionar una metodologia detallada que permeti la instal·lació/integració dels servidors del PIC a la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps.

En aquesta segona etapa es parteix de les solucions detallades en la secció anterior, les quals són presentades formalment davant una representació del PIC, constituïda pel director del centre i els coordinadors dels equips de treball.

Un cop presentades les diferents solucions, la direcció del centre decideix que per a la generació del pla se seguiran els principis definits en la solució recomanada, "Dues IPs", detallada en la subsecció 3.1.3 d'aquesta memòria. També es decideix que s'implementarà una maqueta de la instal·lació, seguint la metodologia que es definirà en aquesta secció, per tal de demostrar empíricament la validesa de la solució escollida.

En aquesta secció es resumiran els trets principals de la solució escollida i es definiran els aspectes de configuració genèrics requerits per a la implementació de la solució. Un cop definits els aspectes més generals de la configuració s'aprofundirà en els detalls de la implementació i es plantejarà una metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN. Finalment es presentarà el disseny de la maqueta a implementar.

3.2.1 Descripció de la solució escollida

Es pretén implementar la solució "Dues IPs", aquesta solució consisteix en afegir als servidors LHC-OPN locals una segona NIC[5] amb una adreça IP de la xarxa LHC-OPN local: 193.145.217.0/24. Així el tot trànsit local es farà per entrega directa via la NIC primària i el trànsit LHC-OPN pel circuit dedicat, via l'esmentada segona NIC.

Es poden trobar tots els detalls sobre la solució a la subsecció 3.1.3, Solució "Dues IPs".

3.2.2 Aspectes genèrics de la configuració

Per tal d'implementar la solució "Dues IPs" és necessari realitzar una sèrie d'adaptacions/configuracions generals en alguns elements de la xarxa actual per tal de preparar la LAN del PIC per a la integració amb la xarxa LHC-OPN.

Adaptacions generals en el Cisco6509

- Configurar com a xarxa local les adreces de la xarxa LHC-OPN local (193.145.217.0/24).
- Acceptar dins la VLAN100 les adreces de la xarxa local LHC-OPN.

5 Network Interface Card

- Configurar com a porta d'enllaç per a la xarxa local LHC-OPN, via el circuit dedicat d'1 Gbps.
- Afegir ACLs[6] d'entrada i sortida per al circuit dedicat d'1 Gbps; es restringirà l'entrada a tothom excepte als CIDR indicats a RS-LHCOPN [www.ripe.net], es restringirà la sortida a tothom excepte les adreces de la xarxa LHC-OPN local.
- S'habilitarà el JumboFrame (MTU 9000 bytes)
- Configurar el *tagging* VLAN als ports corresponents a switch/servidors de la xarxa LHC-OPN

Configuració general dels switch en racks LHC-OPN

- S'habilitarà el JumboFrame (MTU >= 9000 bytes)
- En els casos on el coll d'ampolla es trobi en la connexió switch<->encaminador (Cisco6509) està prevista la creació d'un EtherChannel[7]/802.3ad (agregació de línies) entre ambdós per tal d'augmentar l'ample de banda disponible.

3.2.3 Detalls d'implementació: Opcions per a la solució "Dues IPs"

Fins ara s'han definit les característiques i configuracions principals de la solució, alhora de realitzar la implementació cal entrar en detalls per tal d'eliminar ambigüitats i especificar clarament les diferents opcions d'implementació.

Al PIC la xarxa local (LAN) està completament implementada sobre tecnologia GigabitEthernet, cosa que permet assolir velocitats de fins a 1 Gbps amb un únic servidor, amb una única NIC. Donat que sobre el circuit dedicat pel qual es dissenya el pla d'implementació no és possible assolir velocitats superiors a 1 Gbps, la fita de les configuracions proposades a continuació és aportar, de la forma més simple possible, redundància i fiabilitat sobre la xarxa (LAN) GigabitEthernet.

A continuació es detallen tres opcions d'implementació. Cal tenir en compte que la solució "Dues IPs" garanteix que la xarxa no-LHC-OPN i LHC-OPN es mantenen separades de forma natural, a nivell 3 des del servidor fins a l'encaminador, ja que són dues xarxes diferents.

○ Opció "Gestió exclusiva per VLAN100"

Descripció

Tant el trànsit LHC-OPN com no-LHC-OPN van sobre la VLAN100, tal i com es fa ara: untagged[8] fins al Cisco, i en aquest es realitza el tagging[9] a VLAN100. Inicialment aquesta és la solució preferida, així que es detallarà especialment la configuració dels servidors.

---

6 Access Control List

7 EtherCannel és una tecnologia propietària de Cisco que permet l'agregació de línies. Hi ha una versió estàndard compatible, anomenada 802.3ad (LinkAggregation). Es pot trobar més informació a la subsecció "Etherchannel/Link Aggregation (802.3ad)" del capítol 2.

8 Untagged es refereix a que les trames Ethernet (nivell 2) no van marcades com a membres de cap VLAN

9 Tagging és el procés pel qual les trames Ethernet (nivell 2) són marcades com a membres d'una VLAN concreta

Avantatges

- No cal configuració específica VLAN en switch/servidors/Cisco

Inconvenients

- Monitorització més complexa ja que el trànsit LHC-OPN i no-LHC-OPN no queda separat per VLANs
- Al anar sobre la mateixa VLAN els paquets en broadcast (cap a 255.255.255.255) afecten a ambdues xarxes per igual.

Detalls Tècnics

Adaptacions específiques en el Cisco6509

- Acceptar IPs xarxa LHC-OPN local a VLAN100
- Permetre forwarding entre VLAN100 i LHC-OPN-in o bé afegir LHC-OPN-in a la VLAN 100.

Configuració específica dels switch en racks LHC-OPN

- No cal tagging ni trunking[10], configuració per defecte. A nivell de connexionat se'n definirà un com a LHC-OPN i l'altre com a no-LHC-OPN

Configuració de xarxa dels servidors no-LHC-OPN en racks LHC-OPN

a) sense switch redundant
   - La configuració actual ja és correcte: es connectaran al switch no-LHC-OPN

b) amb switch redundant
   - Es configurarà bonding[11] mode 1 (active-backup) entre les dues NIC GigabitEthernet: la NIC activa (primària) es connectarà al switch no-LHC-OPN, la NIC en standby es connectarà al switch LHC-OPN.

Configuració de xarxa dels servidors LHC-OPN

1. Afegir les dues IP i hostname[12] del servidor LHC-OPN al DNS
2. Si el servidor té 2 NIC

a) Amb flux de dades per diferents switch (veure figura 3.2.1)
   - Connectar NIC 1 (ethx) al switch LHC-OPN, aquesta serà la interfície LHC-OPN (no redundant)
   - Connectar NIC 2 (ethy) al switch no-LHC-OPN, aquesta serà la interfície no-

10 Trunking és utilitzat per a transportar més d'una VLAN per una mateixa connexió. Per a més informació consultar el capítol 2, secció 2.3.3

11 Es pot trobar més informació sobre bonding a http://linux-net.osdl.org/index.php/Bonding

12 nom de la màquina, s'utilitza per evitar memoritzar les adreces IP

LHC-OPN (no redundant)

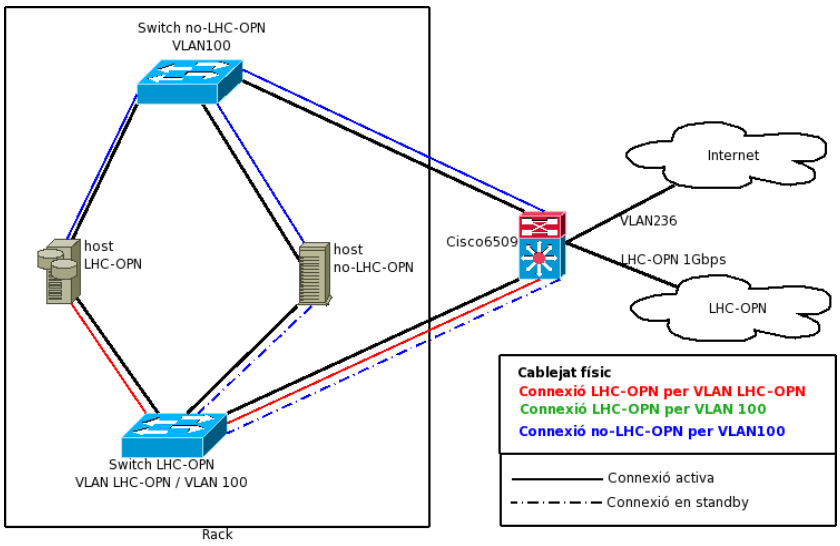


Figura 3.2.1: Flux de dades per a servidors amb 2 NIC dedicades, sense redundància per a la connexió LHC-OPN

b) Amb redundància (veure figura 3.2.2)

- Connectar NIC 1 (ethx) al switch LHC-OPN

  Connectar NIC 2 (ethy) al switch no-LHC-OPN
- Configurar bonding mode 1 (active-backup) entre NIC 1 (primari, actiu) i NIC 2 (standby) [ifcfg-ethx, ifcfg-ethy, ifcfg-bond0, modules.conf][13]
- bond0 serà la interfície LHC-OPN (redundant)
- Crear interfície virtual [ifcfg-bond0:1], aquesta serà la interfície no-LHC-OPN (redundant)

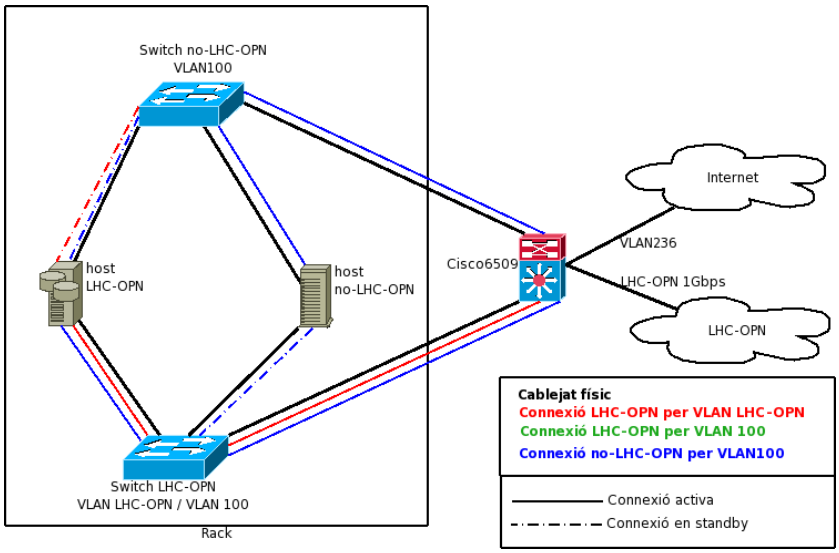


Figura 3.2.2: Flux de dades per a servidors amb 2 NIC i redundant per a les connexions LHC-OPN i no-LHC-OPN, que sempre comparteixen interfície física

Si el servidor té 3 NIC (veure figura 3.2.3)

- Connectar NIC 1 (ethx) al switch LHC-OPN

13 La llista de fitxers correspon als principals fitxers de configuració a modificar/crear en cada pas

Connectar NIC 2 (ethy) al switch no-LHC-OPN

Connectar NIC 3 (ethz) al switch no-LHC-OPN, aquesta serà la interfície no-LHC-OPN (no redundant)

- Configurar bonding mode 1 (active-backup) entre NIC 1 (primari, actiu) i NIC 2 (standby) [ifcfg-ethx, ifcfg-ethy, ifcfg-bond0, modules.conf]
- bond0 serà la interfície LHC-OPN (redundant)

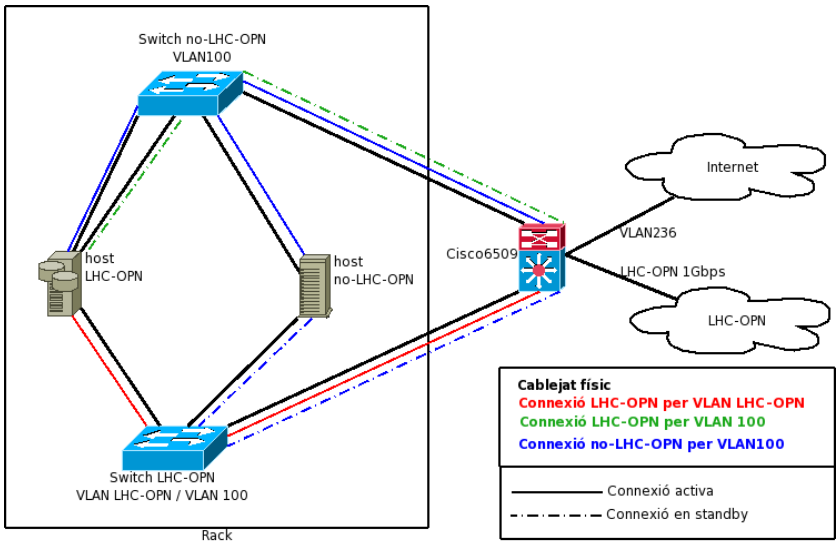


Figura 3.2.3: Flux de dades per a servidors amb 3 NIC, redundant la connexió LHC-OPN per la interfície no-LHC-OPN

Si el servidor té 4 NIC (veure figura 3.2.4)

- Connectar NIC 1 (etha) al switch LHC-OPN

  Connectar NIC 2 (ethb) al switch no-LHC-OPN

  Connectar NIC 3 (ethc) al switch no-LHC-OPN

  Connectar NIC 3 (ethd) al switch no-LHC-OPN
- Configurar bonding mode 1 (active-backup) entre NIC 1 (primari, actiu) i NIC 2
- Configurar bonding mode 1 (active-backup) entre NIC 3 (primari, actiu) i NIC 4 (standby) [ifcfg-ethc, ifcfg-ethd, ifcfg-bond1] (standby) [ifcfg-ethc, ifcfg-ethd, ifcfg-bond0, modules.conf]
- bond0 serà la interfície LHC-OPN (redundant)
- bond1 serà la interfície no-LHC-OPN (redundant)

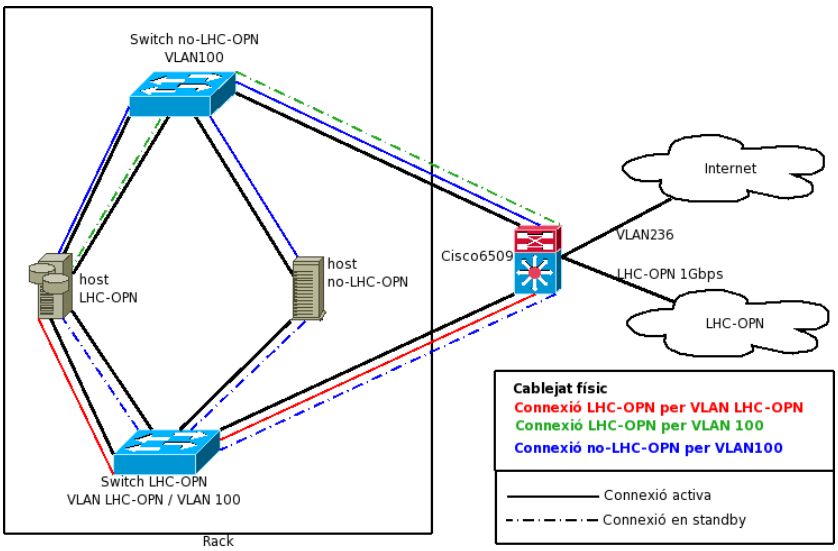


Figura 3.2.4: Flux de dades per a servidors amb 3 NIC, redundant totes les connexions per NIC independents

3. Configurar (IP,màscara de xarxa,MTU=9000bytes, nom del servidor, etc) LHC-OPN a la interfície LHC-OPN del servidor, preferiblement de forma estàtica.

4. Configurar (IP,màscara de xarxa, nom del servidor, etc) no-LHC-OPN a la interfície no-LHC-OPN del servidor, preferiblement de forma estàtica.

5. Configurar rutes

   - Afegir rutes LHC-OPN al servidor[14]

   - Configurar porta d'enllaç per defecte (*default gateway)* via interfície no-LHC-OPN

   En els servidors LHC-OPN caldrà realitzar certes modificacions als paràmetres del sistema per tal d'optimitzar les transferències PIC<->CERN, com ara per exemple la mida per per defecte i màxima de les finestres TCP, activar les extensions de TCP, etc. Així doncs s'hauran d'alterar paràmetres de /proc/sys/net/ipv4/tcp_*, bé directament (echo ...) o bé mitjançant sysctl.

○ Opció "Gestió disjunta per VLAN100 i VLAN LHC-OPN"

Descripció

Tot el trànsit LHC-OPN va sempre per la VLAN LHC-OPN.

Avantatges

- Monitorització senzilla segons VLAN.
- Tipus de trànsit clarament diferenciat a nivell VLAN, possibilitat de priorització
- Aïllament complet de les dues xarxes a nivell VLAN

14 L'addició de les rutes estàtiques LHC-OPN es pot realitzar mitjançant l'script presentat en l'annex B

Inconvenients

- Tagging VLAN en interfície no-LHC-OPN dels servidors LHC-OPN

Detalls Tècnics

Adaptacions específiques en el Cisco6509

- Crear VLAN LHC-OPN per a xarxa LHC-OPN local
- Afegir LHC-OPN-in dins VLAN LHC-OPN, per a evitar fer *forwarding* entre VLAN LHC-OPN i LHC-OPN-in
- Habilitar trunking per als ports on hi ha switch d'un rack LHC-OPN

Configuració específica dels switch LHC-OPN

- Ports on hi hagi interfície LHC-OPN fer tagging a VLAN LHC-OPN
- Ports on hi hagi interfície d'un servidor no-LHC-OPN fer *tagging* a VLAN 100

Configuració específica dels switch no-LHC-OPN en racks LHC-OPN

- Ports on hi hagi interfície d'un servidor no-LHC-OPN fer *tagging* a VLAN100
- Ports on hi hagi interfície no-LHC-OPN amb backup LHC-OPN d'un servidor LHC-OPN permetre trunking

Servidors LHC-OPN i no-LHC-OPN

- La configuració és la mateixa que a l'opció de "Gestió exclusiva per VLAN100" en ambdós casos, exceptuant el punt 2 en el cas de servidors LHC-OPN, ja que s'ha de tenir en compte l'existència de ports d'VLANs diferents en un mateix switch.
- En els servidors LHC-OPN serà necessari realitzar tagging VLAN.

○ Opció "Gestió per VLAN LHC-OPN i redundància via VLAN100"

Descripció

El trànsit de la interfície primària LHC-OPN va per la VLAN LHC-OPN.

Si hi ha algun problema i s'activa la interfície en standby el trànsit sortiria per VLAN100 i es realitzaria reenviament (*forwarding*) entre VLANs en el Cisco.

Avantatges

- Monitorització segons VLAN. Si es detecta trànsit VLAN 100 – VLAN LHC-OPN implica que hi ha un servidor LHC-OPN que funciona sobre la interfície de backup.
- Aïllament parcial (complet en mode d'operació normal) de les dues xarxes a nivell VLAN

Inconvenients

- Quan s'activi la línia de backup d'un servidor LHC-OPN el rendiment es veurà lleugerament més afectat que en les altres opcions.

Detalls Tècnics

Adaptacions específiques en el Cisco6509

- Crear VLAN LHC-OPN per a xarxa LHC-OPN local
- Afegir LHC-OPN-in dins VLAN LHC-OPN, per a evitar fer forwarding entre VLAN LHC-OPN i LHC-OPN-in
- Permetre forwarding entre VLAN 100 i LHC-OPN
- Habilitar trunking per als ports on hi ha switch LHC-OPN

Configuració específica dels switch LHC-OPN

- Ports on hi hagi interfície LHC-OPN fer tagging a VLAN LHC-OPN
- Ports on hi hagi interfície d'un servidor no-LHC-OPN fer tagging a VLAN 100

Configuració específica dels switch no-LHC-OPN en racks LHC-OPN

- No cal tagging ni trunking, configuració per defecte

Servidors LHC-OPN i no-LHC-OPN

- La configuració és la mateixa que a l'opció de “Gestió exclusiva per VLAN100” en ambdós tipus de servidors.

3.2.4 Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN

Per tal d'integrar els serveis del PIC a la xarxa LHC-OPN és necessari implementar alguna de les configuracions anteriorment descrites (subsecció 3.2.3), tant alhora de configurar un nou servidor com alhora d'integrar un servidor ja en producció a la xarxa LHC-OPN.

Els serveis que reben dades a través de la xarxa LHC-OPN, com ara Castor i dCache (veure figura 3.2.5), necessitaran disposar de servidors LHC-OPN. Com que els servidors LHC-OPN disposen de dos noms i dues IP cal tenir clar que per a la comunicació local entre dos servidors LHC-OPN caldrà utilitzar el nom/IP LHC-OPN mentre que per a les comunicacions entre un servidor LHC-OPN i un no-LHC-OPN s'haurà d'utilitzar el nom/IP no-LHC-OPN del servidor LHC-OPN.

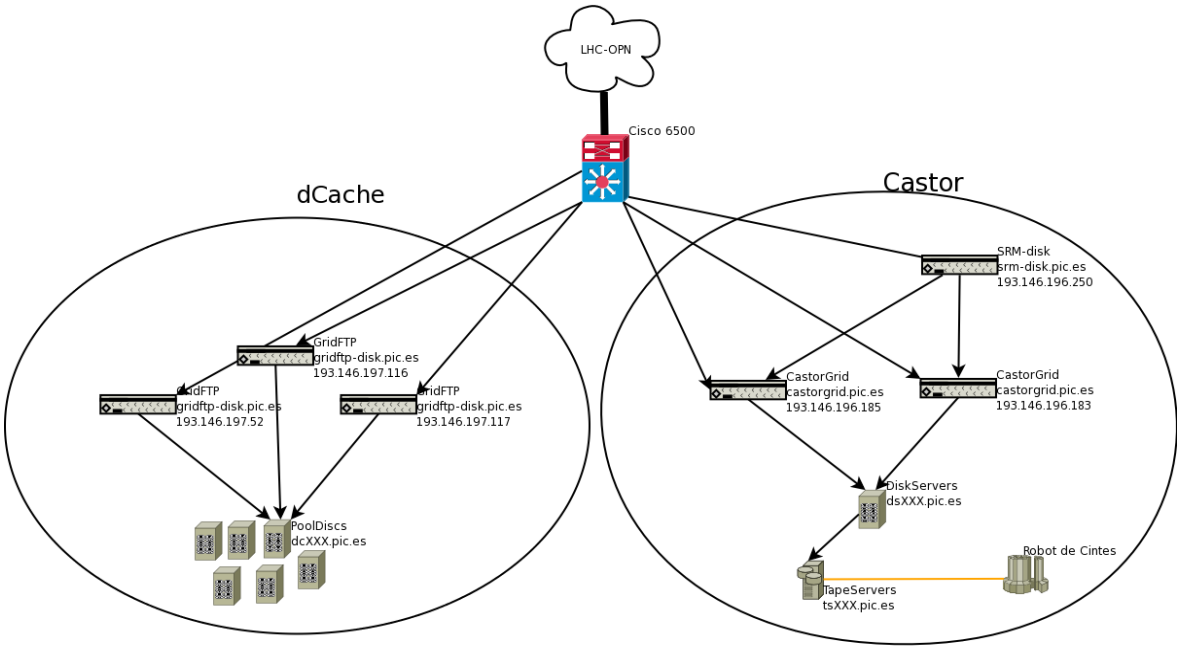


Figura 3.2.5: Esquema simplificat dels serveis d'emmagatzemament de dades dCache i CASTOR. El servidor SRM és només de control, els flux de dades que es reben des de la LHC-OPN van sempre als servidors GridFTP.

És important tenir en compte les comunicacions locals entre servidors que s'enviïn una gran quantitat de dades, així es recomana instal·lar-los en el mateix rack, compartint switch, per tal d'evitar colls d'ampolla en la comunicació entre switch via el cisco6509.

En els servidors LHC-OPN es pot obtenir redundància a diferents nivells, depenent dels recursos disponibles en cada servidor. Tanmateix s'aconsella l'ús d'adreces estàtiques per tal de simplificar el sistema i fer-lo més robust i independent.

3.2.5 Maqueta de la solució amb dues IPs

L'objectiu del disseny i la implementació de la maquesta és demostrar l'efectivitat de la solució escollida. S'utilitzarà la infraestructura provista per la maqueta per a les tasques de certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps.

La maqueta, dissenyada segons la solució amb dues IPs explicada anteriorment, ha d'incloure el sistema de *backup* en cas de fallida d'un switch o una targeta (NIC) GigabitEthenet.

Així doncs es construirà una petita reproducció de la xarxa del PIC per tal de poder provar tots els tipus de comunicació entre els diferents servidors, tant a nivell local com amb la xarxa LHC-OPN externa. Ha de ser possible mesurar el rendiment en els següents tipus de connexions:

1. LHC-OPN local <-> LHC-OPN local
2. LHC-OPN local <-> no-LHC-OPN local
3. LHC-OPN local <-> LHC-OPN extern

També ha de ser possible realitzar proves de disponibilitat sobre el sistema, demostrant així l'efectivitat dels sistemes de backup implementats.

A nivell físic la maqueta correspondrà a dos servidors LHC-OPN i un servidor no-LHC-OPN connectats a dos switch, els quals tindran una connexió directa a l'encaminador cisco 6509, tal i com es pot observar en la figura 3.2.6. A través del cisco els servidors sortiran a la xarxa WAN estàndard (Internet) via la VLAN236 o bé a la xarxa LHC-OPN, via el circuit dedicat a 1 Gbps.

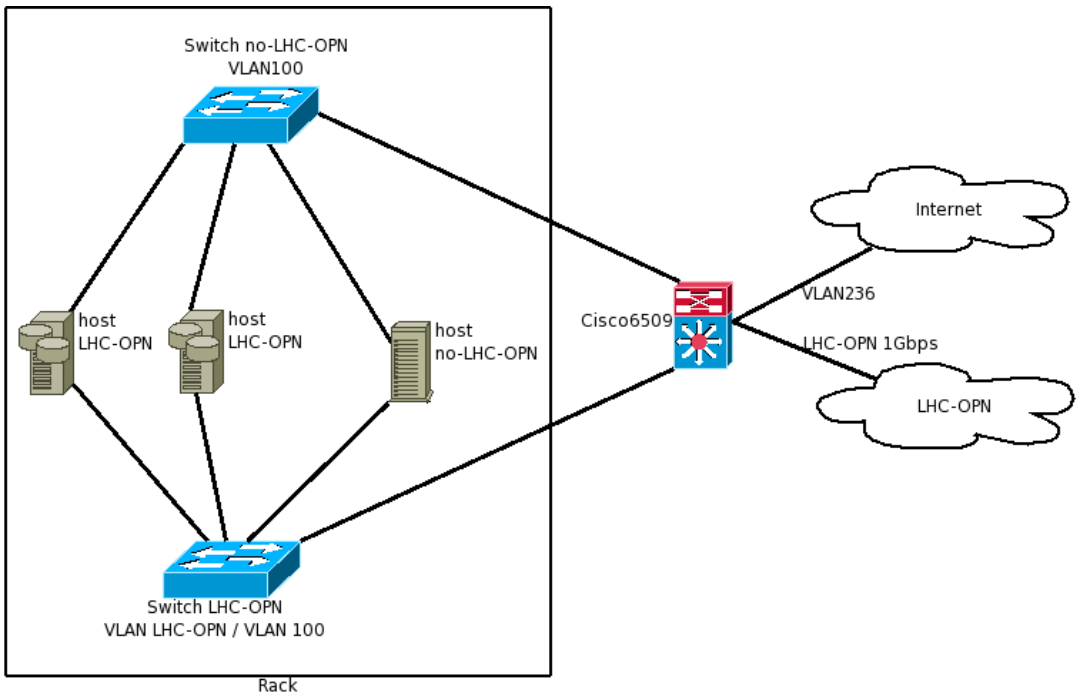


Figura 3.2.6: maqueta de la solució amb dues IPs sobre la qual s'han realitzat les proves i la certificació del circuit dedicat de 1 Gbps

Recursos necessaris Maqueta

Per tal de poder construir la maqueta i demostrar l'efectivitat de la solució, sota qualsevol de les opcions tractades en la subsecció 3.2.3, serà necessari disposar de:

Recursos disponibles

- Switch redundant Cisco Catalyst 6509 amb 1x48 ports GigabitEthernet i 2x4 TenGigabitEthernet
- Línia d'1 Gbps subministrada per redIRIS, incorporada a SIDERAL mitjançant 193.145.217.0/24, amb connectivitat al CERN

Recursos no disponibles

- 2 Switch Gigabit estàndard del PIC (PowerConnect 5224/5234), ambdós connectats directament al Cisco6509.
- Tres servidors amb >=2 tarja de xarxa GigabitEthernet. Els servidors han de poder gestionar el trànsit de test (generat per iperf o similar) per a la connexió d'1 Gbps. Els servidors estaran connectats amb una GigabitEthernet a cada switch.
- Metodologia per afegir servidors de la xarxa LHC-OPN local (193.145.217.0/24) al DNS
- Metodologia per al testeig de les línies des del CERN
- Metodologia per a l'adhesió a la LHC-OPN (RS-LHCOPN)

3.3 Informe de la execució del pla d'implementació sobre el circuit dedicat de 1 Gbps

En aquesta secció es documenta el procés d'execució del pla d'implementació, el qual es troba detallat en la segona secció d'aquest mateix capítol.

Al llarg d'aquesta secció es descriuran els canvis d'especificacions soferts respecte la planificació inicial i l'estat resultant de l'execució del pla d'implementació. També s'inclouen els resultats d'un simulacre de posta en marxa i les incidències ocorregudes al llarg del procés.

3.3.1 Modificacions respecte la planificació inicial

Per causes de força major finalment s'ha decidit que un cop la nova línia d'1 Gbps sigui plenament operativa s'hi traspassarà tot el trànsit generat pel PIC, que actualment va per la VLAN236.

Aquesta modificació als requeriments inicials del projecte suposa alguns canvis respecte el *Pla d'implementació per a la integració dels servidors del PIC i la xarxa LHC-OPN sobre el circuit dedicat d'1 Gbps*:

- Tot el trànsit s'enrutarà per la mateixa porta d'enllaç (*gateway*), el de la nova línia d'1 Gbps, alliberant completament de trànsit del PIC la VLAN236, ja que el medi físic utilitzat per aquesta és utilitzat per altres entitats. Així doncs a nivell de xarxa desapareixen les distincions entre els servidors LHC-OPN i no-LHC-OPN, fent innecessari l'ús de rutes/flux de dades independents per als diferents tipus de trànsit. No serà necessari realitzar cap canvi d'adreces als serveis del PIC i el rang que originalment estava dedicat per a la xarxa LHC-OPN (193.145.217.0/24) restarà en desús.

- El marcatge (tagging) VLAN es realitzarà internament en l'encaminador Cisco6509. Així doncs els switch/servidors treballaran sense VLANs (de forma *untagged)* en tota la xarxa (LAN) del PIC.

- L'entrada en producció amb JumboFrames queda retrassada fins al desplegament del circuit dedicat de 10 Gbps

Amb els nous requeriments del projecte no és possible garantir totes les especificacions detallades a l'apartat 3.1.2. Tot i això aquests canvis d'especificació només apliquen a la primera fase del projecte, és a dir, el desplegament de la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps. Per a la segona fase del projecte els objectius, restriccions i solucions descrites a 3.1.2 segueixen essent vàlides.

3.3.2 Execució del pla d'implementació

L'execució del pla d'implementació s'ha dividit en diverses tasques, la planificació temporal de les quals es pot observar a l'annex A.

A la taula de la figura 3.3.1 es mostra un resum de l'estat final de les tasques mes rellevants de l'execució del pla d'implementació per a la Fase 1 del projecte. La columna d'incidències fa referència a les incidències ocorregudes al llarg del desenvolupament de cada tasca i que es troben detallades en la subsecció 3.3.4.

[id] Tasca	**Observacions**	**Incidències**
[1] Proves de rendiment del switch	Els resultats es poden consultar a l'annex C *(Proves de rendiment de la LAN sobre la maqueta Solució "Dues IPs", Opció "Gestió exclusiva per VLAN100")*	
[2] Configurar maqueta local	Es poden consultar els detalls de la configuració a l'annex D	2
[3] Proves de rendiment sobre maqueta LAN	Proves finalitzades amb èxit, per als detalls re rendiment consultar l'annex C. Per als detalls de configuració consultar l'annex D.	
[4] Configuració Cisco6509	Es poden consultar els detalls de la configuració a l'annex D.	1,2,3
[5] Prova comunicació amb encaminadors en ruta	Els detalls es troben a la subsecció 3.3.5	1
[6] Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps	Certificació finalitzada amb èxit, consultar la secció 3.4 *Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps*	4,6
[7] Traspàs de tot el trànsit del PIC a la nova línia d'1 Gbps	Traspàs finalitzat amb èxit. Es pot trobar més informació als resultats del simulacre (subsecció 3.3.3) i a la subsecció de les incidències (3.3.4)	1,3

Figura 3.3.1: Taula resum de l'estat final de l'execució de les tasques de la Fase1

Resultat final de l'execució

L'execució del pla d'implementació sobre el circuit dedicat de 1 Gbps ha finalitzat correctament,

proveint una connexió entre el PIC i el CERN.

Un cop finalitzat el procés es disposa d'una línia dedicada d'1 Gbps entre el PIC i el CESCA. Entre el CESCA i RedIRIS, així com fins al CERN (via GÉANT) les comunicacions es realitzen per la infraestructura comú dels diferents NRENs[15].

El rendiment de la línia és adequat per a connexions sobre UDP, en les connexions TCP hi ha un límit per connexió i s'hi observa una certa variabilitat (veure incidència 4). A la subsecció 3.3.5 es pot trobar una mostra de les rutes i a la subsecció 3.3.6 el detall de les proves de rendiment de la nova línia.

3.3.3 Simulacre de posta en marxa

El dia 28/02/07 es realitzà un simulacre del traspàs del trànsit del PIC a la nova connexió de 1 Gbps. Concretament es va provar de traspassar, tal i com s'indica en les noves especificacions, tot el trànsit del PIC des de la VLAN236 a la nova línia d'1 Gbps mitjançant un route-map, resolent satisfactòriament la incidència 3.

El primer problema detectat és que les rutes eren asimètriques, degut a que des del CESCA es continuava enviant el trànsit cap al PIC via la VLAN236. Per tal de modificar aquest comportament, i que la xarxa funcioni correctament, cal avisar al CESCA amb antel·lació al traspàs definitiu del trànsit.

L'MTU de la connexió (incidència 1), causà problemes amb certs protocols com ara https (i.e. gmail.com). Configurant els servidors del pic amb un MTU <=1486 es va solucionar el problema, de forma provisional.

La resolució inversa de les adreces de la classe C (193.145.217.0/24), que originàriament havia de representar la xarxa LHC-OPN del PIC, encara no es troba delegada al servidor DNS del PIC, es pot trobar més informació a la incidència 5.

Degut a problemes amb el sistema de monitorització MRTG del PIC no es disposa de dades respecte la càrrega de l'encaminador durant el simulacre.

15 National Research and Education Network. A Espanya l'NREN és RedIRIS, representat a Catalunya per l'Anella Científica.

3.3.4 Incidències i resolució de les mateixes

En aquesta subsecció es detallen diverses incidències ocorregudes al llarg de l'execució del pla d'implementació per al desplegament del circuit dedicat a 1 Gbps.

Les incidències resoltes es mostren en verd, les pendents de resoldre al tancament de la fase 1 del projecte resten ressaltades en vermell.

1. **INCIDÈNCIA**: L'MTU informat per l'encaminador 130.206.202.26 és de 1486 i els paquets amb MTU>1494bytes es perden sense generar cap tipus de resposta, generant un problema anomenat *Black Hole Detection* que es pot trobar descrit en l'RFC 2923. Es pot trobar més informació sobre aquest cas concret a l'annex E

   RESOLUCIÓ: Cal ampliar l'MTU de l'encaminador 130.206.202.26 a 1500 (standard Ethernet) o a 9000 bytes (standard de facto JumboFrames). Un cop estigui l'incidència solucionada es podrà procedir amb la confirmació el correcte funcionament de la resta de la ruta, aparentment correcte. Durant la diagnosis de la incidència s'ha localitzat i resolt la incidència 2 (07/03/07).

   Finalment, degut a problemes de coordinació entre els responsables de les diferents institucions, el problema s'ha resolt durant el traspàs del tot el trànsit a la línia d'1 Gbps (14/3/07), de la mateixa forma que el problema de l'MTU en la incidència 2; indicant explícitament l'MTU dels ports implicats en la connexió punt a punt amb l'encaminador de RedIRIS ubicat a l'Anella Científica.

2. INCIDÈNCIA: L'MTU mínim entre servidors LHC-OPN i no-LHC-OPN locals (p.e. 193.145.217.3 i ui08.pic.es) és inferior a 1500 bytes. A més les rutes d'entrada/sortida als servidors de la maqueta LHC-OPN són assimètriques; d'entrada ui08->193.145.217.1->193.145.217.3 i la de sortida és passant pel CESCA: 193.145.217.3->193.145.217.1->130.206.202.25->...->84.88.19.11->ui08.pic.es

   RESOLUCIÓ: S'ha especificat l'MTU correcte a totes les interfícies de l'encaminador afectades[16] i, per a solventar el problema de les rutes, s'ha corregit el route-map (de la incidència 3). Al prendre ambdues accions ja és possible realitzar transferències locals (LAN) amb MTU=1500 bytes; paquets TCP amb MSS En les comunicacions amb el CERN, encara a causa de la incidència 1, només és possible passar d'un MTU màxim de 1486 a 1494 bytes (07/03/07). Recordem que l'objectiu és utilitzar trames ethernet estàndard, amb un MTU de 1500 bytes.

3. **INCIDÈNCIA**: Problemes en la configuració de la nova connexió, en l'encaminador, alhora de definir les noves rutes de sortida.

   RESOLUCIÓ: S'ha implementat mitjançant un route-map, continua en fase de

16 Els no membres de la VLAN100, com ara el 3/44, agafaven per defecte un MTU inferior a 1500, s'ha hagut d'afegir la comanda mtu 1500 a la configuració del router Cisco per tal que l'MTU del port fos el correcte.

proves/configuració fins que no s'enllesteixi la integració de la maqueta dins la VLAN100 i/o en una nova VLAN 222. Finalment (14/03/07), al passar la nova línia a porta d'enllaç per defecte en l'encaminador, el route-map ha deixat de ser necessari i es realitza un encaminament estàndard, es pot trobar la configuració de l'encaminador a l'annex D.

4. **INCIDÈNCIA:** Variabilitat en les connexions PIC<->CERN sobre TCP.

   RESOLUCIÓ: Tot sembla indicar que hi ha algun mecanisme de limitació de throughput per a les connexions TCP, com ara cues RED[22] en algun encaminador, o un firewall que limita les connexions TCP (UDP funciona correctament).

   Per tal de localitzar la font del problema es proposa seguir dues línies d'investigació en paral·lel:

   1. Demanar un servidor per a fer proves amb 1 Gbps a RedIRIS, així ens assegurem que podem assolir la velocitat contractada dins el nostre ISP, descartant problemes de rendiment a causa de la configuració de l'encaminador o la connexió punt a punt d'aquest fins a l'anella (on hi ha incidència 1).

   2. Contactar amb el CERN per assegurar-nos de que chapuza.cern.ch disposa realment d'una connexió dedicada de, com a mínim 1 Gbps, sense firewalls ni possibles colls d'ampolla per a TCP. Aniria força bé tenir un esquema de la xarxa del CERN des de chapuza.cern.ch fins a GÉANT.

   Sobre la connexió actual la ruta és compartida des de l'encaminador de RedIRIS al CESCA fins al CERN i les connexions TCP són controlades/limitades de forma automàtica, així doncs aquesta incidència queda tancada, considerant que una velocitat d'uns 150-300 Mbps (20-40MB/seg) per connexió TCP és suficient i, donat que amb més connexions és possible assolir velocitats molt properes a 1 Gbps, la direcció està d'acord amb la limitació detectada.

5. **INCIDÈNCIA**[17]**:** La resolució DNS inversa no funciona, probablement no s'ha finalitzat correctament el procés de registre i delegació d'autoritat per a la classe C 193.145.217.0. En la figura 3.3.2 es mostra el resultat de la resolució DNS directa i inversa:

```
[chapuza] / > host lhcopn03.pic.es
lhcopn03.pic.es has address 193.145.217.3
lhcopn03.pic.es has address 193.146.197.155
[chapuza] / > host 193.145.217.3
Host 3.217.145.193.in-addr.arpa not found: 3(NXDOMAIN)
```

Figura 3.3.2: resultat de la resolució DNS directa i inversa

   RESOLUCIÓ: (07/03/07) cal fer la sol·licitud mitjançant el formulari corresponent, que es pot trobar a ftp://ftp.rediris.es/rediris/nic/iris-nic-in-addr.txt. Aquest formulari ha de ser enviat a RedIRIS pel contacte tècnic (UAB) de la classe C. S'està gestionant mitjançant el sistema de tickets interns, ticket 2614[18]. NOTA: a data 22/04/07 es constata que el responsable del servei DNS ja ha resolta la incidència.

17 La resolució d'aquesta incidència ha estat posterior a la finalització de la primera fase del projecte.

18 http://www.tickets.pic.es/?module=issues&action=view&issueid=2614&gid=12

6. **INCIDÈNCIA**: S'ha detectat que des del 8/3/07 algunes connexions TCP PIC->CERN (entre el 20 i el 50%) fallen donant error de timeout tant des de la VLAN236 com des de la nova línia d'1 Gbps.

   RESOLUCIÓ: mitjançant tcpdump a màquines locals del PIC s'ha vist que des del CERN no es respon a alguns paquets SYN (inici de connexió TCP). Sembla ser que el problema es repeteix en d'altres T1, fent sospitar d'un possible problema al CERN.

   S'ha creat un sensor de Nagios (script) per a monitoritzar la creació de connexions a servidors, així com un procediment d'actuació per quan el sensor faci saltar l'alarma, es troba a l'annex F.

   Finalment hi havia un problema de maquinari (HW) al CERN: una NIC de 10 Gbps d'un dels dos encaminadors LHC-OPN del CERN que estava connectat a la xarxa de GÉANT donava errors de paritat, un cop s'ha reiniciat (7:30 13/03/07) s'ha tornat a la normalitat, tal i com es mostra a la figura 3.3.3

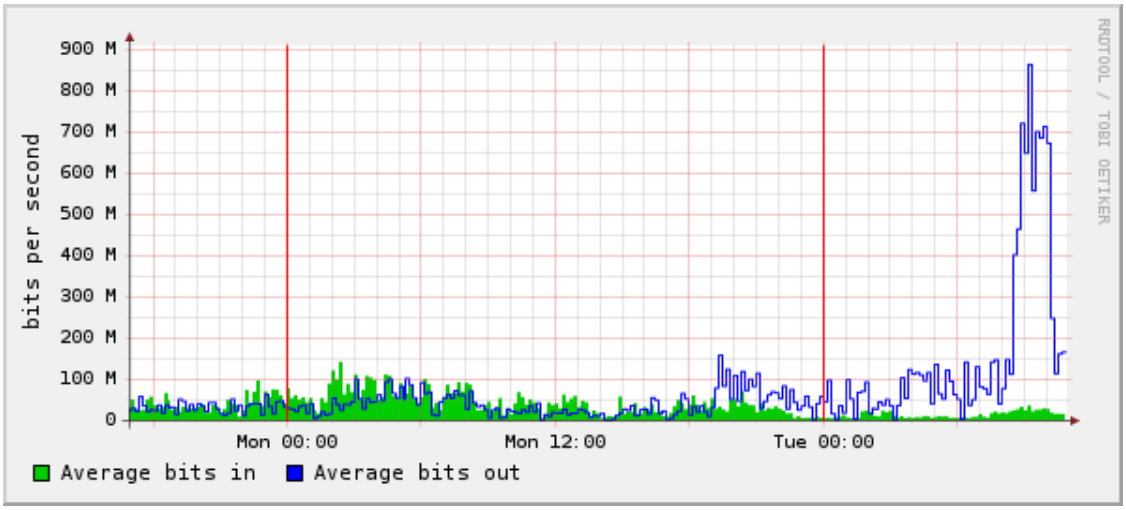


Figura 3.3.3: Gràfica de l'ús de la xarxa durant el període de fallida i reparació de la NIC defectuosa. Mentre la interfície fallava l'ample de banda utilitzat de la connexió de 10 Gbps era inferior als 200Mbps, un cop reparada hi ha un pic a 900Mbps i a continuació la recuperació és gradual degut als mecanismes de control de la congestió de TCP.

3.3.5 Rutes PIC-CERN

A continuació es mostra el detall de les rutes que segueixen els paquets del PIC al CERN en ambdues connexions del PIC. En el traceroute de la figura 3.3.4 es mostra la ruta via el circuit dedicat a 1 Gbps amb les IPs de la LHC-OPN, i en el traceroute de la figura 3.3.5 es mostra la ruta via la VLAN236 des d'una IP estàndard del PIC.

```
[root@lhcopn03 root]# traceroute chapuza.cern.ch
traceroute to chapuza.cern.ch (128.142.208.4), 30 hops max, 38 byte packets
 1  193.145.217.1 (193.145.217.1)  1.081 ms  0.703 ms  0.615 ms
 2  GE1-1-3.EB-Barcelona0.red.rediris.es (130.206.202.25)  0.861 ms  0.722 ms  0.619 ms
 3  CAT.XE6-0-0.EB-IRIS2.red.rediris.es (130.206.250.25)  14.846 ms  14.837 ms  14.857 ms
 4  SO0-0-0.EB-IRIS4.red.rediris.es (130.206.240.2)  21.349 ms  14.969 ms  14.982 ms
 5  rediris.rt1.mad.es.geant2.net (62.40.124.53)  14.979 ms  14.969 ms  14.982 ms
 6  so-7-2-0.rt1.gen.ch.geant2.net (62.40.112.25)  37.092 ms  36.954 ms  36.970 ms
 7  swiCE2-10GE-1-1.switch.ch (62.40.124.22)  37.086 ms  37.081 ms  37.093 ms
 8  e513-e-rci76-1-swice2.cern.ch (192.65.184.222)  37.092 ms  37.080 ms  37.093 ms
 9  l513-c-rftec-1-be8.cern.ch (192.16.166.129)  42.589 ms  37.207 ms  37.217 ms
```

```
10  * * *
11  * * *
12  chapuza.cern.ch (128.142.208.4)  37.364 ms  37.324 ms  37.343 ms
```

Figura 3.3.4: ruta PIC-CERN via el circuit dedicat d'1 Gbps amb IPs LHC-OPN. Un cop surt del PIC la connexió va directament a un encaminador de RedIRIS

La part de la ruta PIC-CERN comú en les dues connexions es troba ressaltada en negreta.

```
[root@ui08 root]# traceroute chapuza.cern.ch
traceroute to chapuza.cern.ch (128.142.208.4), 30 hops max, 38 byte packets
 1  cisco1.pic.org.es (193.146.196.130)  0.856 ms  0.620 ms  0.546 ms
 2  anella-ifae.cesca.es (84.88.19.9)  1.162 ms  1.246 ms  1.124 ms
 3  AE0.EB-Barcelona0.red.rediris.es (130.206.202.1)  1.365 ms  1.287 ms  1.448 ms
 4  CAT.XE6-0-0.EB-IRIS2.red.rediris.es (130.206.250.25)  15.408 ms  15.416 ms  15.420 ms
 5  SO0-0-0.EB-IRIS4.red.rediris.es (130.206.240.2)  15.537 ms  23.529 ms  15.581 ms
 6  rediris.rt1.mad.es.geant2.net (62.40.124.53)  15.592 ms  15.694 ms  15.652 ms
 7  so-7-2-0.rt1.gen.ch.geant2.net (62.40.112.25)  37.631 ms  37.749 ms  38.408 ms
 8  swiCE2-10GE-1-1.switch.ch (62.40.124.22)  37.956 ms  43.840 ms  38.065 ms
 9  e513-e-rci76-1-swice2.cern.ch (192.65.184.222)  37.597 ms  37.537 ms  37.562 ms
10  l513-c-rftec-1-be8.cern.ch (192.16.166.129)  45.177 ms  37.741 ms  50.894 ms
11  * * *
12  * * *
13  chapuza.cern.ch (128.142.208.4)  38.106 ms  37.899 ms  37.902 ms
```

Figura 3.3.5: ruta PIC-CERN via la VLAN236 amb IPs estàndard del PIC, a partir del primer encaminador de RedIRIS la ruta és idèntica a la de la figura 3.3.4

És important veure que la VLAN236 és una connexió amb l'Anella Científica i, en canvi, les IPs LHC-OPN associades al circuit dedicat d'1 Gbps connecten directament amb RedIRIS. Es pot observar la ruta convergeix a les instal·lacions del CESCA, en un encaminador (*router*) de RedIRIS, i a partir d'aquí és comú fins al servidor de destí.

Tal i com es mostra en la figura 3.3.6, amb MTU estàndard Ethernet (1500bytes) el funcionament de la xarxa sobre el circuit dedicat de 1 Gbps és correcte. Amb MTU de 9000 bytes la prova no s'ha pogut realitzar degut al retràs de l'entrada en producció de JumboFrames. Tal i com es mostra en les proves de la figura 3.3.6 el camí és correcte fins a l'encaminador de RedIRIS ubicat al CESCA.

```
Hop IP-Len    Type/Code  Host/Msg
=== ====== ============  ========
  1   9000    TTLX       bar-kirana-ge-0-2-0-0.3rox.net
  2   9000    TTLX       192.88.115.174
  3   9000    TTLX       wash-psc10G.layer3.nlr.net
  4   9000    TTLX       newy-wash-98.layer3.nlr.net
  5   9000    TTLX       216.24.184.86
  6   9000    TTLX       so-7-0-0.rt1.ams.nl.geant2.net
  7   9000    TTLX       so-6-2-0.rt1.fra.de.geant2.net
  8   9000    TTLX       so-6-2-0.rt1.gen.ch.geant2.net
  9   9000    TTLX       so-7-0-0.rt1.mad.es.geant2.net
 10   9000    TTLX       rediris-gw.rt1.mad.es.geant2.net
 11   9000    TTLX       SO1-1-0.EB-IRIS2.red.rediris.es
 12   9000    TTLX       NAC.XE0-1-0.EB-Barcelona0.red.rediris.es
 13   9000   NOREP       Possbile black hole; Repeating: 1
 13   9000   NOREP       Possbile black hole; Repeating: 2
 13   9000   NOREP       Possbile black hole; Repeating: 3
 13   9000   NOREP       Black hole
---------------- Probing far (smallest MTU possible)
 13     68    TTLX       lhc-router.red.rediris.es
---------------- Probing up
 13     68    TTLX       lhc-router.red.rediris.es
 13    296    TTLX       lhc-router.red.rediris.es
 13    508    TTLX       lhc-router.red.rediris.es
 13   1006    TTLX       lhc-router.red.rediris.es
 13   1492    TTLX       lhc-router.red.rediris.es
 13   1500    TTLX       lhc-router.red.rediris.es
 13   2002   NOREP       Possbile black hole; Repeating: 1
 13   2002   NOREP       Possbile black hole; Repeating: 2
 13   2002   NOREP       Possbile black hole; Repeating: 3
 13   2002   NOREP       Black hole
---------------- Probing up Done
```

```
---------------- Probing far Done
 14   1500 UnReach/PORT  193.145.217.3
Host: 193.145.217.3; Path MTU: 1500 (Max requested: 9000)
```

Figura 3.3.6: Proves de l'MTU del circuit dedicat de 1 Gbps. Hi ha més informació sobre el Black Hole del salt 13 a l'annex E.

La prova de la figura 3.3.6 ha estat realitzada mitjançant una utilitat del *Pittsburgh Supercomputing Center*[19]. Es pot observar que la ruta des de GÉANT (hop 10) fins a l'encaminador de RedIRIS al CESCA (hop 12) funciona correctament amb MTU de 9000bytes. Donat que el funcionament des de GÉANT fins al CERN (T0) i la resta de T1 és responsabilitat de cada T1 i el T0 no cal diagnostica la xarxa més enllà de la connexió amb GÉANT.

Dins la LAN, amb les proves de la maqueta s'ha provat que el funcionament de JumboFrames (MTU=9000) és correcte, en conseqüència per a poder utilitzar JumboFrames només falta per incrementar l'MTU en la connexió RedIRIS (CESCA)<->PIC.

3.3.6 Proves de la connexió sobre el circuit dedicat d'1 Gbps

En aquesta subsecció es detallen les diferents proves de latència, rendiment i fiabilitat realitzades amb la connexió al CERN sobre el circuit dedicat de 1 Gbps.

Per a la realització de les proves s'ha disposat d'un servidor de test al CERN (chapuza.cern.ch) amb una connexió directa a 1 Gbps. Des del PIC les proves s'han realitzat amb les màquines de la maqueta (veure subsecció 3.2.5).

Cal tenir en compte que totes les proves mostrades a continuació es realitzaren abans de que la connexió fos certificada i entrés en producció. Un cop amb la connexió certificada i en producció i totes les incidències resoltes es realitzà un petit resum de les proves, donant resultats idèntics (a excepció dels problemes amb MTU superior a 1494 bytes detallats en les incidències 1 i 2).

○ Detall de les proves de RTT (Round Trip Time)

En les figures 3.3.7 i 3.3.8 es mostren les gràfiques generades a partir de diferents proves amb la utilitat *ping,* mantenint la línia sense trànsit actiu. El paràmetre variable és la mida del paquet ICMP de echo. En les primeres posicions de l'eix *Packet Size* (mida del paquet) apareix (DF), que significa *Don't Fragment*, això serveix per evitar que cap màquina o encaminador fragmenti el paquet.

19 http://kirana.psc.edu:6880/mtudisc?dest=193.145.217.3&maxmtu=9000

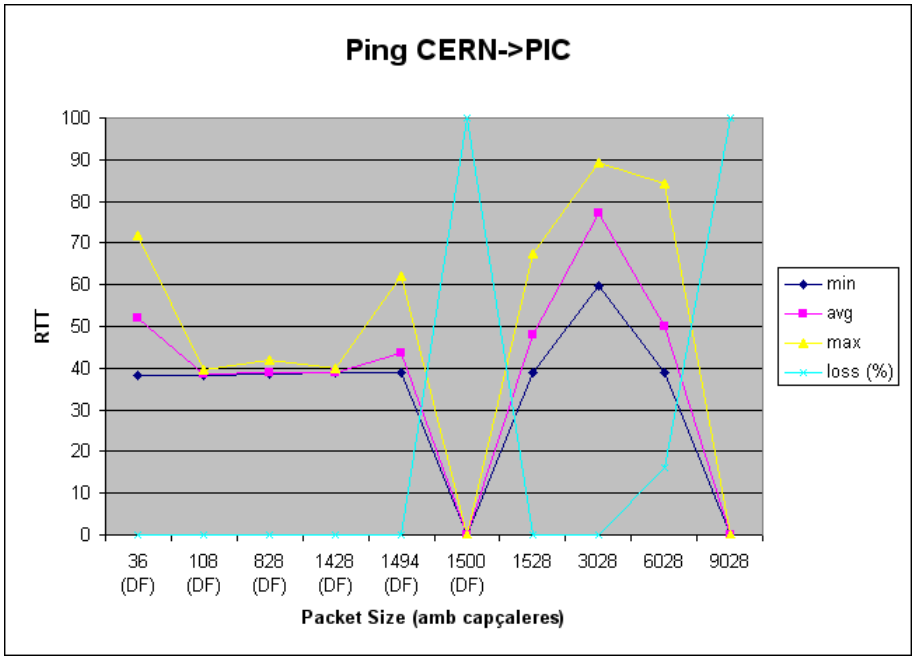


Figura 3.3.7: RTT de la connexió CERN->PIC en funció de la mida del paquet

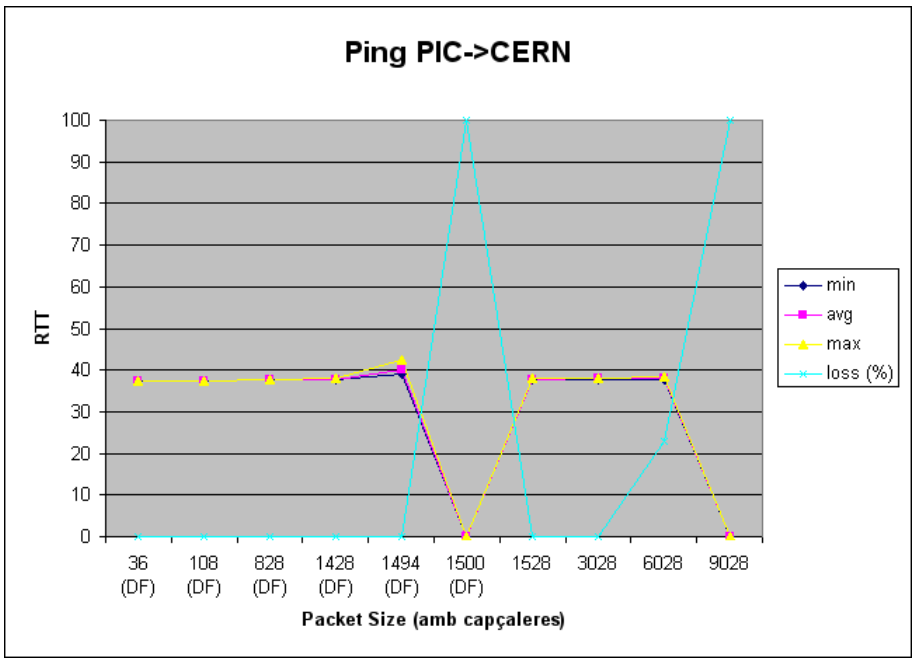


Figura 3.3.8: RTT de la connexió PIC->CERN en funció de la mida del paquet

Es pot observar més estabilitat en l'RTT dels paquets que van en sentit PIC->CERN, així com el notable increment de l'RTT al activar la desfragmentació en direcció CERN->PIC, fent evident la menor disponibilitat de recursos des del CERN (chapuza.cern.ch). El misstge retornat per ping respecte les pèrdues (loss) per a paquets de 6028 i 9028 bytes és *Frag reassembly time exceeded* en ambdós casos, és a dir, que el servidor no ha pogut desfragmentar el paquet a temps.

També cal notar la pèrdua de paquets causada quan l'MTU resultant dels missatges ICMP és >1494. En sentit CERN->PIC ping ens retorna el missatge *Frag needed and DF set (mtu=1486)*, en canvi en sentit PIC->CERN ping simplement no rep cap tipus de resposta, degut al "*forat*

negre", relacionat amb l'incidència 1 i explicat en l'annex E.

○ Detall de les proves d'ample de banda (throughput)

A continuació es mostraran els detalls de les proves de rendiment TCP i UDP.

En forma de resum es pot dir que per a connexions UDP s'han obtingut throughputs de ~800mbps amb <2% de pèrdues. Les connexions TCP hi ha algun punt de la xarxa on són limitades, en ambdós sentits, probablement això és degut a que la infraestructura de xarxa des del CESCA fins al CERN és compartida amb la resta de centres educatius i de recerca de l'estat. La velocitat mitjana de les connexions TCP és d'uns 200mbps, en ambdós sentits, tot i això s'ha arribat a aconseguir velocitats de fins a ~500mbps.

A continuació hi ha el detall de les proves TCP i UDP.

Proves TCP

Després de diverses proves de rendiment s'ha arribat a la conclusió de que sobre TCP el throughput generat i mostrat per *iperf version 1.7.0 (13 Mar 2003) pthreads* és pràcticament idèntic al de la última versió: *iperf version 2.0.2 (03 May 2005) pthreads*. Segons les estadístiques extretes del Cisco6509 per NetFlow Analyzer 5 ambdues versions d'iperf generen realment el trànsit que mostren. En les següents proves TCP s'utilitzarà la versió 1.7.0, ja que és la versió considerada més estable pels responsables del manteniment de *Scientific Linux Cern (SLC)*.

En la gràfica de la figura 3.3.9 es mostra l'efecte del canvi de la mida de finestra TCP en una transferència PIC->CERN sobre el circuit dedicat d'1 Gbps.

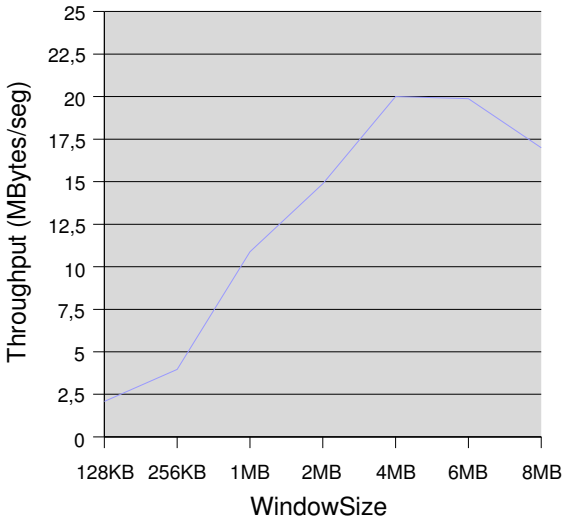


Figura 3.3.9: efecte del canvi de la mida de la finestra en transferències PIC->CERN, la velocitat de transferència òptima s'aconsegueix amb una finestra de 4 Mbytes.

Com es pot veure la velocitat màxima en la transferència de test és d'uns 20MBytes/seg, és a dir, 160Mbps. El llindar on la mida de finestra deixa de ser el factor limitant és sobre els 4Mb, i a partir dels 6 el rendiment comença a decaure. Aquest comportament és degut a la gestió dels sockets que

realitza el SO i, sobretot, a la limitació de velocitat que hi ha sobre les connexions TCP en la línia d'1 Gbps. Cal veure la limitació com una eina efectiva que evita que una única màquina pugui acaparar tot l'ample de banda amb una sola connexió.

A continuació es realitzaran proves unidireccionals seguides de bidireccionals per a mostrar l'impacte de realitzar connexions en ambdós sentits, tant el client com el servidor estan configurats amb una finestra de 6MB.

Proves PIC->CERN – figura 3.3.10

1. Connexió unidireccional lhcopn03->chapuza
2. Connexió bidireccional

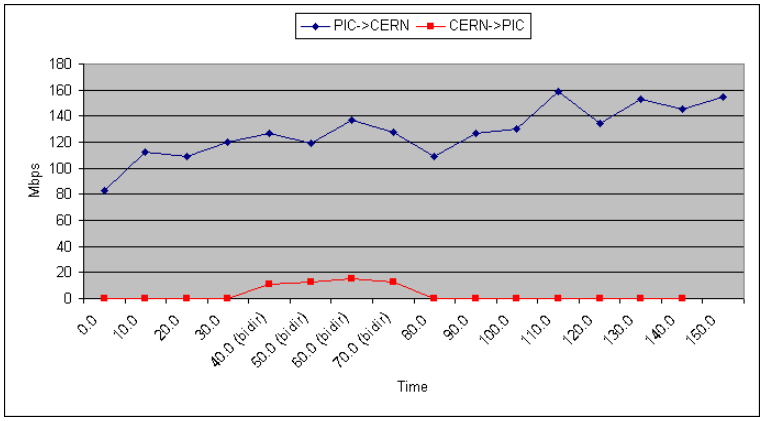


Figura 3.3.10: proves unidireccionals i bidireccionals TCP PIC->CERN, en Mbps (8*Mbytes) Es pot observar el domini de les transferències PIC->CERN a causa de la degradació de rendiment del servidor de test del CERN

Proves CERN->PIC – figura 3.3.11

1. Connexió unidireccional chapuza->lhcopn03
2. Connexió bidireccional

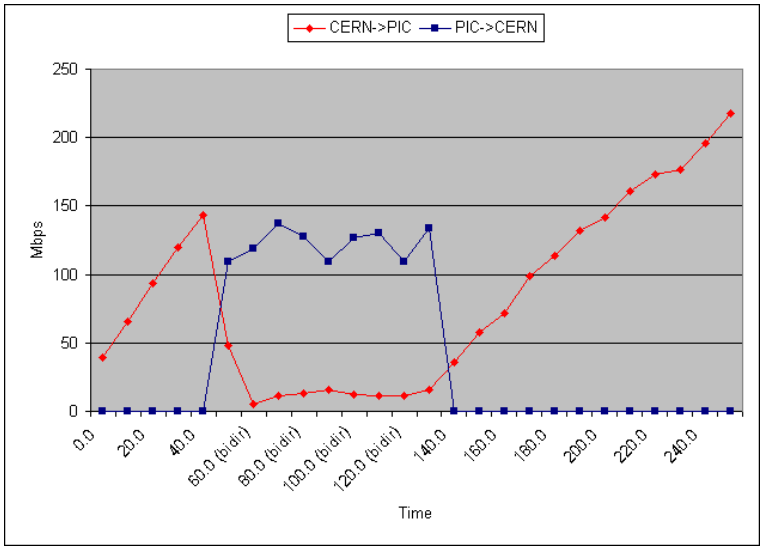


Figura 3.3.11: proves unidireccionals i bidireccionals TCP CERN->PIC. Novament el rendiment del servidor del CERN es veu degradat al iniciar transferències bidireccionals

Tal i com es pot observar en les gràfiques anteriors (3.3.10 i 3.3.11) hi ha algun problema en algun

punt de la xarxa que limita les connexions TCP dels servidors, en ambdós sentits. La velocitat mitjana de les connexions és d'uns 150mbps, en ambdós sentits, tot i això s'ha arribat a aconseguir velocitats de fins a ~500mbps. Quan la connexió és bidireccional les transferències TCP CERN->PIC es veuen especialment afectades, principalment degut al rendiment del servidor de test ofert pel CERN.

Amb un únic servidor extern de proves es fa difícil on es troba el coll d'ampolla, hi ha instants en els quals les connexions PIC->CERN són més lentes i d'altres en les que passa al revés.

Proves UDP

Després de diverses proves de rendiment s'ha arribat a la conclusió de que sobre UDP el throughput generat i mostrat per *iperf version 1.7.0 (13 Mar 2003) pthreads* és superior al de *iperf version 2.0.2 (03 May 2005) pthreads*. Segons les estadístiques extretes del Cisco6509 per NetFlow Analyzer 5 ambdues versions d'iperf generen realment el trànsit que mostren. En les següents proves UDP s'utilitzarà la versió 1.7.0, ja que és la més estable i que ofereix millor rendiment.

Així com les proves TCP mostraven força variabilitat, les proves sobre UDP estables i donen resultats similars a diferents hores i dies.

Proves PIC->CERN

En gràfica de la figura 3.3.12 primer s'estableix una connexió PIC->CERN i després una CERN->PIC que finalitza abans d'acabar la primera.

1. Connexió unidireccional lhcopn03->chapuza [*iperf -c chapuza.cern.ch -l1440 -t1000 -b1g*]
2. Connexió bidireccional

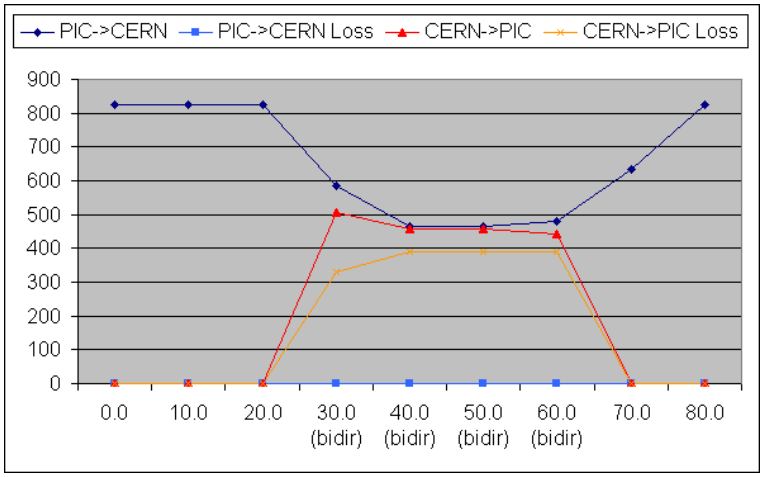


Figura 3.3.12: proves unidireccionals i bidireccionals UDP PIC->CERN, en aquest cas el coll d'ampolla es troba en el switch, el qual no és capaç de gestionar més d'1 Gbps FD (500+500Mbps)

Proves CERN->PIC

En la gràfica de la figura 3.3.13 primer s'estableix una connexió CERN->PIC i després una

PIC->CERN que finalitza abans d'acabar la primera.

1. Connexió unidireccional chapuza->lhcopn03 [*iperf -c 193.145.217.3 -l1440 -t1000 -b1g*]
2. Connexió bidireccional

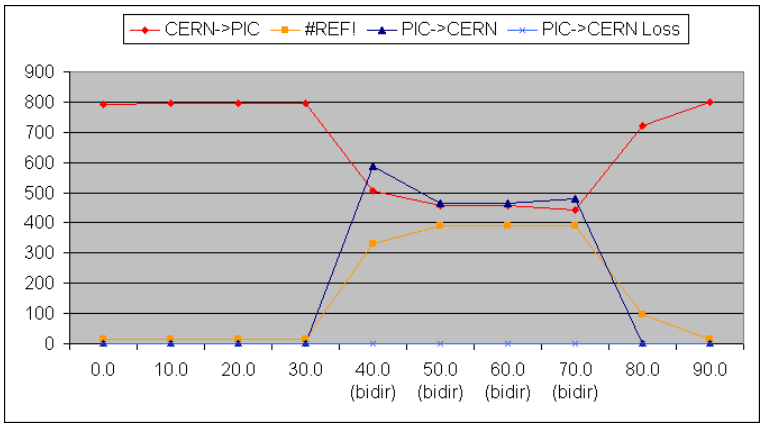


Figura 3.3.13: proves unidireccionals i bidireccionals UDP CERN->PIC, novament s'evidencien els problemes des switch Dell locals

Es pot concloure que la xarxa funciona bé per a trànsit UDP unidireccional, assolint velocitats de ~800mbps tant d'entrada com de sortida.

En el transit bidireccional es detecta una baixada de rendiment que fa que les connexions siguin de ~500mbps en cada direcció. En aquest cas el coll d'ampolla son els servidors i switch locals, en les transferències bidireccionals dins la LAN (el mateix switch) es detecta una baixada de rendiment similar.

Actualment no es pot confirmar l'existència o no d'un altre coll d'ampolla en les connexions PIC<->CERN degut a que no es disposa de diversos servidors, o d'un més potent, en el CERN per a fer més proves.

3.4 Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps

En aquesta secció, que tanca el tercer capítol i la primera fase del projecte, es documenta la certificació de la connexió sobre el circuit dedicat de 1 Gbps.

Part del procés de certificació de la connexió s'ha realitzat al llarg del desplegament del pla, en la secció anterior.

En aquesta secció es fa un petit resum de les proves empíriques de rendiment i fiabilitat de la connexió i, a continuació, es realitza un petit anàlisi de la càrrega dels encaminadors i switch locals. Tal i com s'indica en el primer capítol, el procés de certificació també inclou una anàlisi del trànsit de la nova xarxa i estadístiques del rendiment de la connexió a mig termini.

3.4.1 Proves empíriques de rendiment i fiabilitat de la connexió

En la subsecció 3.3.6 es pot trobar una explicació detallada de les proves de rendiment realitzades.

La capacitat de la connexió queda certificada al demostrar empíricament que per a connexions UDP és possible obtindre throughputs de ~800mbps amb <2% de pèrdues, on l'element limitador és el servidor del CERN. Al PIC actuen com a limitadors de velocitat els switch Dell PowerConnect 5324 utilitzats, estesos a tota la infraestructura de la xarxa local, els quals teòricament disposen de connexions a 1 Gbps Full Duplex però realment mostren un rendiment de connexions 1 Gbps Half Duplex[20].

Es pot afirmar que la connexió ofereix velocitats ~1 Gbps en ambdós sentits simultàniament (Full Duplex).

Respecte a TCP cal dir que la mida de la finestra (windowsize) utilitzada en les proves ha estat calculada, per tal d'evitar colls d'ampolla causats per la finestra de TCP, utilitzant la fórmula[23]:

WindowSize = bandwidth*rtt

Tal com es pot observar a la gràfica l'RTT de la connexió en desús és d'uns 40ms, així doncs tindrem que la mida de finestra òptima seria 1000mbps*40ms = 1000e6bps*0,04s = 40000000 bits **~= 4,8 Mbyte**

Les connexions TCP hi ha algun punt de la xarxa on són limitades, en ambdós sentits. Això és degut a que la infraestructura de xarxa des del CESCA fins al CERN és compartida amb la resta de centres educatius i de recerca de l'estat. La velocitat mitjana de les connexions és d'uns 200Mbps, en ambdós sentits, tot i això s'ha arribat a aconseguir velocitats de fins a ~500Mbps.

20 En 1 Gbps Half Duplex la suma entre ambdues direccions és 1 Gbps, en canvi amb 1 Gbps Full Duplex hi ha 1 Gbps d'ample de banda disponible per sentit de la connexió.

Com es pot observar a la figura 3.4.1 la velocitat màxima en la transferència de test és d'uns 20MBytes/seg, és a dir, 160Mbps. El llindar on la mida de finestra deixa de ser el factor limitant és sobre els 4MB (MegaBytes), i a partir dels 6MB el rendiment comença a decaure (a causa de la gestió TCP del SO dels servidors).

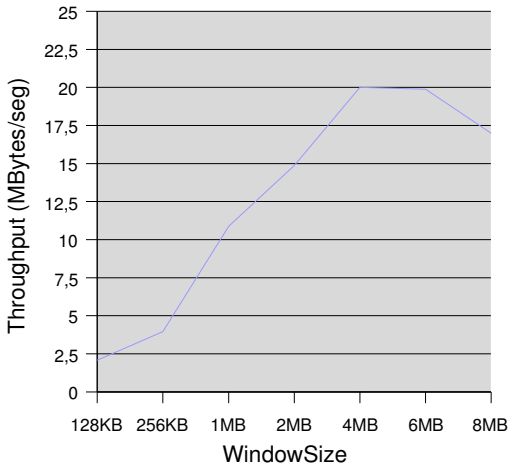


Figura 3.4.1: efecte del canvi de la mida de la finestra en transferències PIC->CERN

Aquest comportament és degut a la limitació de velocitat que hi ha sobre les connexions TCP en la línia d'1 Gbps i a que, al transferir dades entre els servidors, l'RTT dels paquets del flux de dades de la transferència s'incrementa fins a uns 300ms (mesurat amb Thrulay).

3.4.2 Anàlisi de la càrrega en els encaminadors i commutadors locals

Els switch Dell PowerConnect 5324 utilitzats en la LAN del PIC no permeten monitoritzar l'ús de la CPU, de totes formes amb la nova línia d'1 Gbps la càrrega d'aquests és molt similar a l'anterior.

La càrrega del Cisco6509 tampoc s'ha vist afectada per l'entrada en producció de la nova línia d'1 Gbps i manté l'ús de la CPU de la tarja supervisora per sota del 10%.

3.4.3 Anàlisi del trànsit de la nova xarxa

Degut a que finalment per la nova connexió s'hi encamina tot el trànsit del PIC, aquest és molt variat tot i que bàsicament té com a origen destí el CERN (T0) o algun dels T1.

Amb la incorporació de la nova classe C a la xarxa (193.145.217.0/24) s'ha detectat un increment en els escanejos de ports, els quals són generalment utilitzats per a detectar màquines vulnerables a la xarxa. Els diversos escanejos són realitzats amb nmap o eines similars, realitzant intents de connexió UDP, TCP i *echo* ICMP (ping). La majoria d'IPs corresponen a connexions domèstiques (ADSL o similars) de diferents ISP, com ara BSN-XX-89-251.dial-up.dsl.siol.net, XXX.Red-83-50-203.dynamicIP.rima-tde.net, modemcableXXX.234-70-69.mc.videotron.ca, etc.

3.4.4 Anàlisi estadístic del rendiment de la connexió a mig termini

A la figura 3.4.2 es mostren estadístiques de l'ús de la connexió (port g3/45) extretes del Cisco6509 amb NetFlow Analyzer durant 70hores

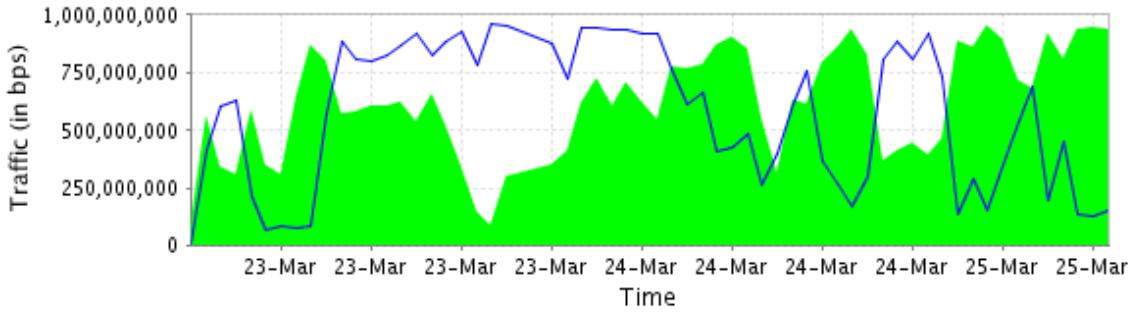
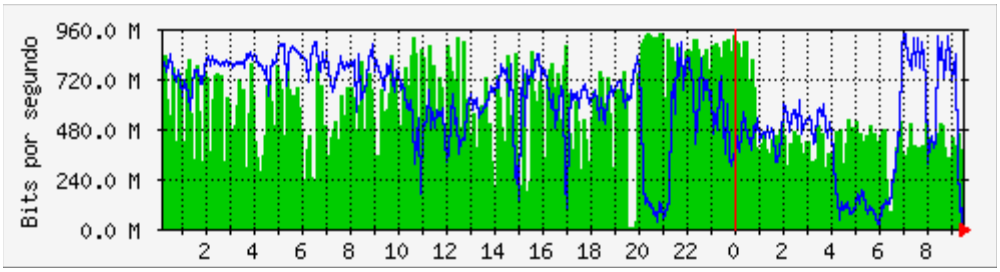


Figura 3.4.2: estadístiques locals de l'ús de la connexió sobre el circuit dedicat d'1 Gbps durant 70 hores. Pel comportament de les aplicacions és difícil veure transferències bidireccionals d'alta velocitat, tot i això hi ha algun pic de fins a 750Mbps bidireccional.

A la figura 3.4.3 es mostren les estadístiques de l'ús de la connexió d'1 Gbps durant 24 hores extretes de la monitorització de RedIRIS[21]



Max Entrant:945.0 Mbps (94.5%) Promig Entrant:535.5 Mbps (53.6%) Actual Entrant:390.0 Mbps (39.0%)
Max Sortint:946.1 Mbps (94.6%) Promig Sortint:592.9 Mbps (59.3%) Actual Sortint:165.5 Mbps (16.6%)

Figura 3.4.3: estadístiques RedIRIS de l'ús de la connexió sobre el circuit dedicat d'1 Gbps al llarg de 24h, amb pics superiors a 900Mbps bidireccionals

Com es pot observar s'assoleixen velocitats sostingudes properes a 1 Gbps, en ambdós sentits. Així doncs, a 27 de març del 2007, la nova connexió a 1 Gbps queda certificada com a connexió a 1 Gbps Full Duplex, amb un MTU de 1500bytes

21 http://www.rediris.es/red/stats/EB-Barcelona1/lhc.html

4 Capítol 4 – Desplegament circuit dedicat a 10 Gbps

En aquest capítol s'abordarà la segona fase del desenvolupament del projecte, que correspon al desplegament de la connexió a la xarxa LHC-OPN sobre el circuit dedicat a 10 Gbps.

Tal i com s'ha explicat en el primer capítol, i de la mateixa forma que en el capítol anterior, el desplegament de la connexió sobre el circuit dedicat a 10 Gbps s'ha organitzat en quatre etapes. En les quatre primeres seccions del capítol es detallen les diferents etapes del desplegament i, finalment, en la cinquena secció s'estudia la viabilitat d'una proposta per al desplegament d'una connexió redundant[1].

4.1 Estudi de solucions per a l'integració dels serveis del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps

En aquesta primera secció és pretén determinar i descriure les solucions més adequades per tal d'obtenir una connexió entre el PIC i el CERN sobre el circuit dedicat de 10 Gbps, així com una infraestructura capaç de certificar-la. També es proposaran diverses solucions per a la integració dels serveis de gestió de dades del PIC amb la nova connexió a la xarxa LHC-OPN.

En primer lloc s'analitzarà el punt de partida d'aquesta segona fase del PFC, determinant les especificacions i restriccions del sistema. Un cop clar el marc d'actuació es proposarà una solució per a resoldre els problemes plantejats amb el desplegament de la connexió i del circuit dedicat de 10 Gbps, plantejant algunes opcions per a la futura certificació de la connexió. Finalment, en la última subsecció, es presenten diverses solucions per a la integració dels sistemes de gestió de dades del PIC, Castor[2] i dCache[3], amb la xarxa LHC-OPN.

4.1.1 Descripció de la situació inicial

Des de l'inici d'aquesta segona fase (març 2007) i fins a mitjans de l'estiu del 2007, al PIC es disposa de dues connexions WAN operatives:

- VLAN236: connexió al node de l'Anella Científica per on, des del final de la primera fase, no es transmet trànsit del PIC
- Línia dedicada d'1 Gbps: connexió punt a punt (p2p) amb RedIRIS, posada en marxa a la primera fase del projecte, per on actualment s'encamina tot el trànsit del PIC

1 Amb una connexió redundant em refereixo a un circuit de *backup* que doni redundància al circuit dedicat de 10 Gbps primari

2 CASTOR (CERN Advanced STORage manager) és un sistema d'emmagatzemament jeràrquic desenvolupat al CERN i utilitzat per a l'emmagatzemament de dades en disc i/o cinta.

3 dCache és un sistema d'emmagatzemament jeràrquic que sorgeix d'un esforç conjunt entre DESY (www.desy.de), FERMILAB (www.fnal.gov) i l'aportació d'altres centres implicats en el projecte LHC. Juntament amb Enstore, o una aplicació similar, és capaç de gestionar dades en robots de cintes.

Des del gener del 2007 s'està realitzant el desplegament a nivell 1(capa física) del circuit dedicat de 10 Gbps, sobre fibra òptica monomode. L'estatus, per segments, del desplegament del circuit dedicat de fibra òptica PIC-CERN a l'inici d'aquesta fase és:

[ID] Segment	Estatus	Observacions
[1] Bellaterra (PIC) – Pedralbes (Anella)	20%	10/01/07 CESCA confirma la comanda i un termini de lliurament de 18 setmanes. Data estimada d'entrega: 25/05/07
[2] Pedralbes (CESCA) – MADRID (RedIRIS)	100%	Forma part del desplegament de RedIRIS10[4] i està completament desplegat
[3] Madrid (RedIRIS) – Ginebra (GÉANT)	50%	RedIRIS ja ha fet la petició formal a GÉANT. Data estimada d'entrega: 30/04/07
[4] Ginebra (GÉANT) - CERN	100%	Ja es troba operatiu, forma part de la infraestructura del CERN

El diagrama de la figura 4.1.1 representa la topologia física de la connexió, dividida per segments. El material que ha d'adquirir el PIC està indicat en verd, en vermell el de l'Anella Científica i en blau el d'altres entitats.

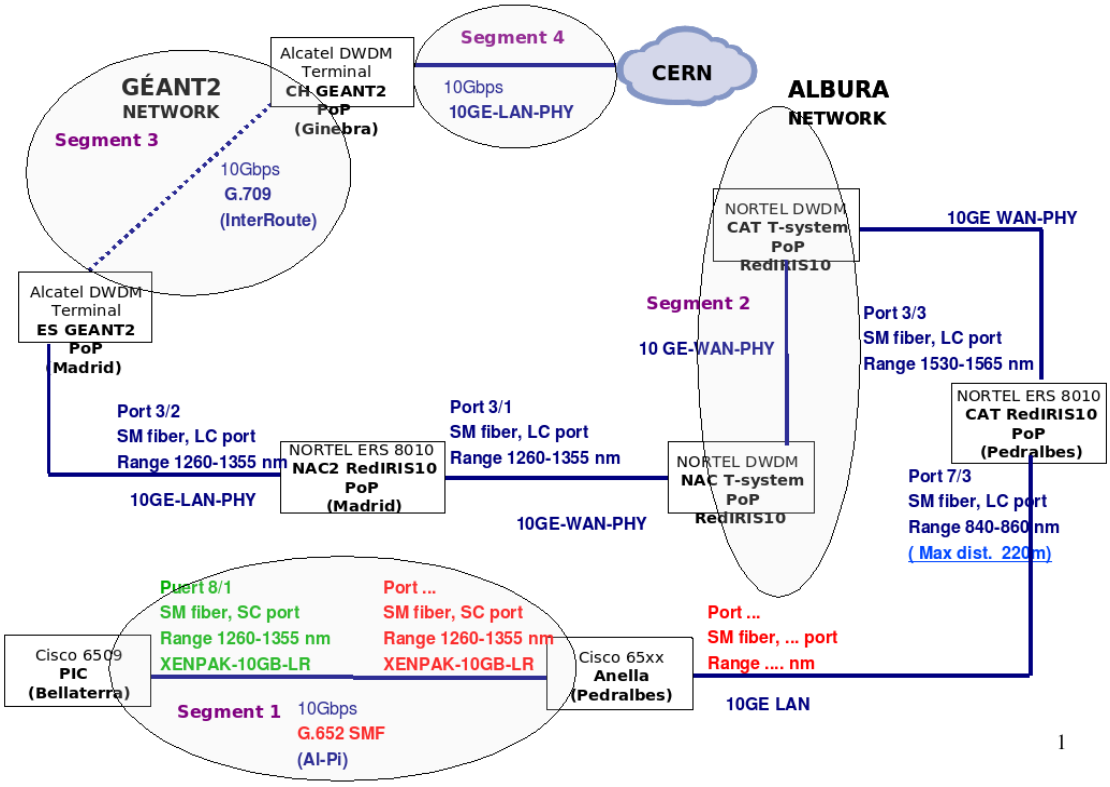


Figura 4.1.1: topologia física de la connexió. En verd el material a adquirir pel PIC, en vermell el que ha d'adquirir l'Anella Científica. La resta de components ja estan instal·lats (en blau)

4 Projecte de RedIRIS per a desplegar una xarxa d'altes prestacions a nivell nacional http://www.rediris.es/red/RedIRIS10/

L'extrem del segment 1 corresponent al PIC que es veu en la figura 4.1.1 és en realitat un Rack d'Al-Pi situat a la sala de comunicacions de la UAB. Des d'aquest rack fins a l'encaminador Cisco6509, al rack 11 del PIC, s'ha dissenyat el connexionat mostrat a la figura 4.1.2.

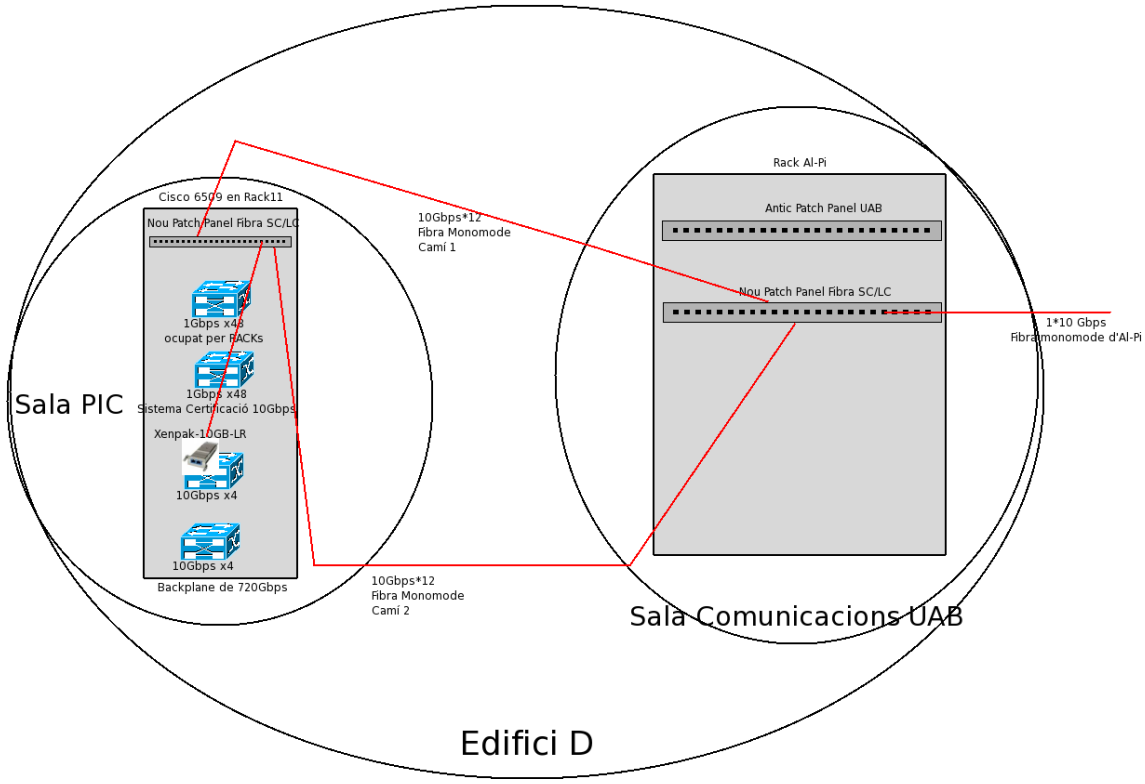


Figura 4.1.2: connexionat en fibra monomode del rack 10 del PIC al rack d'Al-Pi. Per obtenir redundància els camins són diferents

Tal i com es veu en el diagrama (figura 4.1.2) s'instal·laran dos cables independents, de 12 fibres monomode cadascun. Els cables seguiran camins independents des del tauler de connexions[5] del PIC fins al tauler de connexions d'Al-Pi. Per al circuit dedicat de 10 Gbps només s'utilitzarà una fibra d'un dels dos cables, la resta de la instal·lació és realitza per redundància i en previsió de l'addició de noves connexions.

La instal·lació la durà a terme l'empresa SEDER, i la data de finalització prevista és a finals d'Abril 2007.

Dins el PIC els servidors (LHC-OPN i no-LHC-OPN) es troben en la situació resultant del desplegament del circuit dedicat d'1 Gbps (veure capítol 3); tots els servidors en producció formen part de la mateixa xarxa (193.146.196.0/22) i VLAN (VLAN100), totes les transferències es realitzen mitjançant la connexió a RedIRIS sobre el circuit dedicat de 1 Gbps.

5 El tauler de connexions (*Patch Panel*) és un panell instal·lat especialment per a la connexió de fibres monomode amb connectors SC.

4.1.2 Especificacions del sistema objectiu

La interconnexió objectiu dels servidors es manté igual que a la planificació inicial per al circuit dedicat de 1 Gbps (capítol 3, figura 3.1.5).

Durant l'*ATLAS meeting at PIC*[6] Esther Robles, coordinadora de l'àrea de xarxa de RedIRIS, clarificà les especificacions del nou circuit dedicat de 10 Gbps indicant que:

- Es realitzarà una connexió a nivell 2 (capa d'enllaç) des del PIC fins al CERN, passant per l'Anella Científica, RedIRIS i GÉANT, tal i com es mostra en la figura 4.1.4.

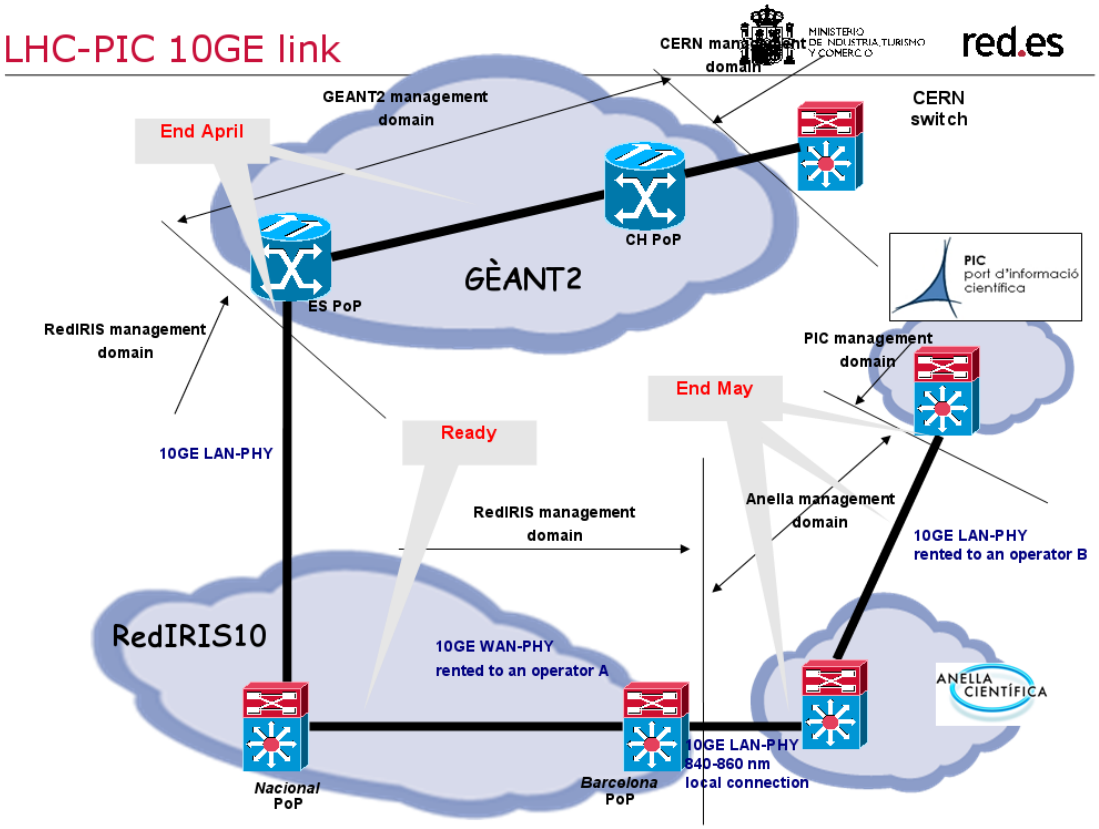


Figura 4.1.4 - extreta de l'exposició d'Esther Robles: connexionat a nivell físic i dominis de gestió per al circuit dedicat de 10 Gbps. Amb el domini de GÉANT només hi pot interactuar RedIRIS.

- Entre el CERN i el PIC es crearà una connexió punt a punt (p2p) de nivell 3 (capa de xarxa).

- Sobre la connexió punt a punt creada s'establirà una sessió BGP amb el CERN, per tal de fer-ho possible serà necessari que el PIC disposi d'un Sistema Autnònom (AS) amb un Autonomous System Number (ASN) públic.

6 Visita d'ATLAS al PIC durant els dies 11,12 i 13 d'abril, on assistí Esther Robles, per a més informació es pot consultar l'agenda i les transparències de les presentacions a http://indico.cern.ch/conferenceDisplay.py?confId=14041

Restriccions i Assumpcions

S'ha de garantir el compliment dels requeriments imposats per la LHC-OPN en el document *LHC Tier-0 to Tier-1 High-Level Network Architecture*, resumits l'apartat “Restriccions i Assumpcions” de 3.1.2.

Per tal d'evitar que la LAN (xarxa local) sigui el coll d'ampolla de la connexió serà necessari dissenyar solucions específiques per a les aplicacions i/o serveis que demandin grans amples de banda procedents de la xarxa LHC-OPN, com per exemple els sistemes de gestió de dades Castor o dCache.

S'assumeix que per a la comunicació amb la LHC-OPN via el nou circuit dedicat a 10 Gbps s'utilitzarà exclusivament la xarxa 193.145.217.0/24 o bé una nova xarxa assignada per RIPE (www.ripe.net), associada al nou Sistema Autònom (AS).

4.1.3 Possibles solucions per al desplegament de la connexió i del circuit dedicat de 10 Gbps

Degut al retràs en la obtenció de les especificacions finals de la nova connexió a 10 Gbps és possible que les dates d'entrega previstes inicialment per al desplegament del circuit de 10 Gbps es vegin lleugerament alterades.

A data 16/4/07[7] no es disposa dels detalls tècnics de la connexió punt a punt amb el CERN ni d'un AS propi per a poder fer BGP dins la LHC-OPN. RedIRIS i l'Anella Científica s'han ofert, com a membres de RIPE, a gestionar l'alta del nou AS per al PIC.

Per tal de poder realitzar el procés amb èxit serà necessari coordinar activament les tasques entre el PIC, RedIRIS i el CERN. Amb aquest objectiu s'ha dissenyat el pla d'actuació per al desplegament del circuit dedicat de 10 Gbps que es troba a l'annex G.

Per tal d'agilitzar el procés de transferència d'informació amb al resta d'entitats (RedIRIS, CERN, etc.) el pla (que es troba en l'annex G) s'ha redactat en anglès i hi inclou les dates d'entrega dels segments del circuit dedicat pendents (Madrid-Ginebra i Barcelona-PIC), la petició d'IPs i el Sistema Autònom a RIPE, la petició dels paràmetres de la connexió punt a punt amb el CERN i la seva posterior configuració, la configuració de l'AS (BGP) i tot un seguit de proves tant a capa 2 (sobre la connexió punt a punt) com a capa 3 (amb l'encaminament de BGP). Per a veure el diagrama de Gantt amb la planificació completa consultar l'annex G.

Certificació del circuit dedicat a 10 Gbps

Un cop sigui operatiu, i abans d'entrar en producció al PIC, el circuit dedicat de 10 Gbps s'ha de certificar. Amb aquest objectiu s'ha dissenyat la solució de certificació mostrada a la figura 4.1.5, plantejant tres possibles escenaris al CERN.

Al PIC l'arquitectura ve condicionada pels recursos disponibles (N servidors amb NIC

7 Durant la primera etapa de la segona fase del PFC

GigabitEthernet) i la naturalesa de la certificació: s'intenta certificar les característiques del circuit dedicat de la forma més realista possible, és a dir, generant múltiples flux de dades tal i com faran les aplicacions quan el circuit entri en producció.

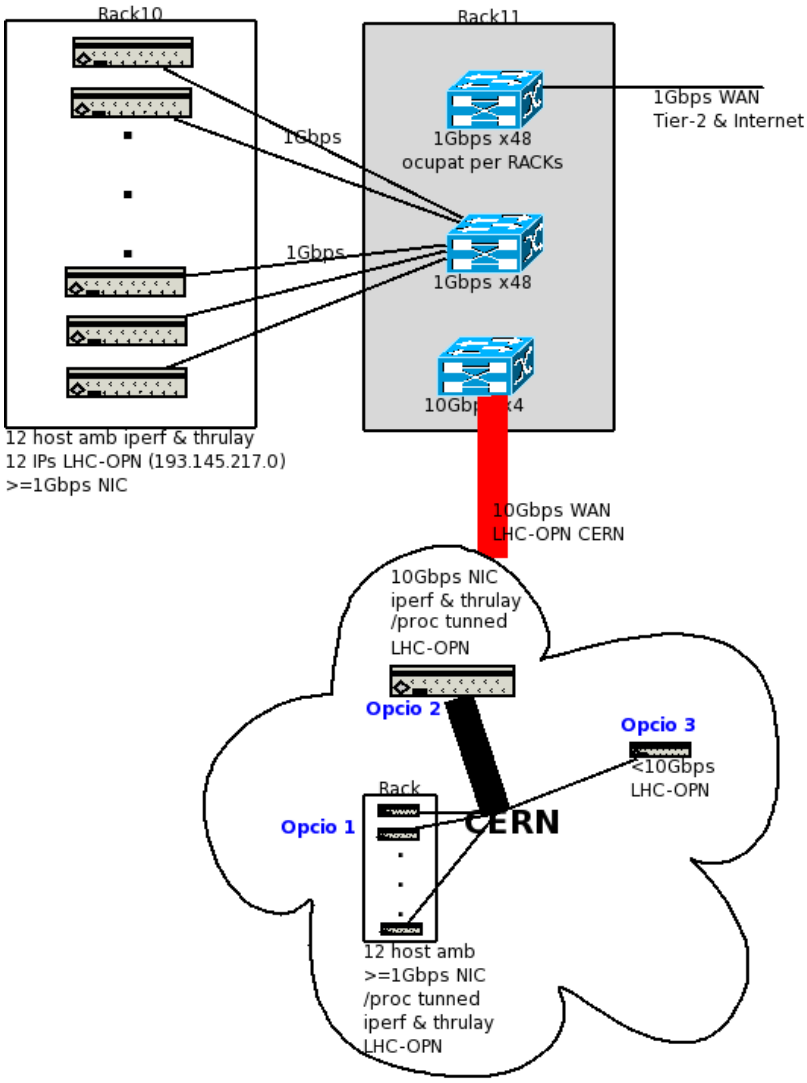


Figura 4.1.5: arquitectura de certificació per al circuit dedicat de 10 Gbps, amb tres alternatives de certificació al CERN, la preferida és l'opció 1.

Com es pot observar en la figura 4.1.5, prèviament a una consulta oficial, al CERN s'hi han plantejat tres possibles escenaris. De més a menys pràctic, des del punt de vista del PIC, s'han plantejat les següents opcions:

- Opció 1: 12 servidors dedicats amb NICs GigabitEthernet al CERN
  - Es necessita que el CERN col·labori de forma activa
  - + Prova real amb M flows entre 12 servidors 1-1 i M*(1-N)
  - + Proves TCP i UDP Bidireccionals
  - + Monitorització 100% (throughput, pèrdua de paquets, càrrega en encaminadors)
  - + Es pot simular un tipus de càrrega similar a la real.

- Opció 2: 1 servidor dedicat amb una NIC TenGigabitEthernet al CERN
    - Es necessita que el CERN col·labori activament
    - - Prova fictícia amb M flows N-1
    - - servidor/NIC en CERN pot limitar
    - + Proves TCP i UDP Bidireccionals
    - + Monitorització 100% (throughput, pèrdua de paquets, càrrega en encaminadors)
- Opció 3: Generació de soroll contra un servidor “víctima”
    - Amb aquest tipus de certificació no cal que el CERN col·labori de forma activa
    - - Només es pot provar UDP en una direcció (PIC->CERN)
    - - La monitorització s'ha de realitzar des dels switch i/o encaminadors
    - La prova és similar a un atac DoS contra un servidor “víctima”, el qual quedarà saturat durant el temps que duri la certificació

Al PIC, tal i com s'indica a la figura 4.1.5, s'instal·laran 12 servidors amb NIC GigabitEthernet connectats directament a l'encaminador Cisco6509. En els servidors, amb SLC3[8], s'hi instal·larà iperf i thrulay, que seran executats de forma simultània i paral·lela en tots els servidors mitjançant vxargs[9]. Així s'aconseguirà crear un flux d'entre 0-12 Gbps, segons es desitgi, suficient per a saturar la línia de 10 Gbps i provar la connexió en situacions properes a la realitat prevista.

Per tal de poder realitzar les proves cal contactar amb el CERN (Edoardo Martelli) i acordar la metodologia de certificació amb una de les alternatives proposades o bé amb una que ells proposin.

4.1.4 Possibles solucions LAN: integració dels sistemes de gestió de dades del PIC amb la xarxa LHC-OPN

Les solucions per a la integració dels servidors LHC-OPN via el circuit dedicat proposades per al desplegament del circuit dedicat a 1 Gbps (capítol 3) segueixen essent vàlides.

Donat l'èxit de la solució implementada sobre la connexió d'1 Gbps (Solució “Dues IPs”, opció de “Gestió exclusiva per VLAN100”), es mantindrà el mateix model d'integració dels servidors. En aquesta secció es plantejaran solucions específiques que permetin als sistemes de gestió de dades Castor i dCache aprofitar l'ample de banda de 10 Gbps del nou circuit dedicat.

Properament el PIC s'ha de decidir per l'ús de CASTOR2 o dCache i definir exactament el model arquitectural de producció que s'utilitzarà, així doncs és molt probable que les arquitectures en base a les quals s'han dissenyat les solucions que es presenten a continuació pateixin alguna variació en la implementació final.

En les solucions que es mostren a continuació hi ha casos en els quals s'assigna dues IP a un servidor, segons els experts consultats això pot presentar problemes per al sistema d'autenticació via certificats dels serveis dCache i CASTOR, així com als GridFTP del Globus toolkit. Per tal de

8 Scientific Linux CERN 3, més informació a http://linux.web.cern.ch/linux/scientific3/

9 Es pot trobar més informació sobre vxargs a http://dharma.cis.upenn.edu/planetlab/vxargs/

solventar els problemes esmentats es pot muntar un servidor DNS dedicat per a certs servidors locals, “enganyant” així als serveis i forçant la validesa dels certificats. Es pot trobar més informació repecte aquesta pràctica al web http://hep.kbfi.ee/index.php/IT/DCacheOnMultipleIFs i a [24]

Per tal d'entendre els diagrames cal saber que la majoria dels servidors del PIC disposen de 2 NIC GigabitEthernet. Addicionalment es disposa d'alguns servidors SUN Fire X4500 que seran utilitzats com a DiskSErvers (en CASTOR) o PoolDiscs (en dCache), aquests servidors disposen de 4NIC GigabitEthernet cadascun. Al PIC hi ha disponibles de dos switch 3com 3870 apilables (*stackables*), aquests switch seran utilitzats en el disseny inicial de les solucions.

En els diagrames de les solucions plantejades es pot observar que entre l'*stack*[10] 3Com i el Cisco6509 hi ha 8 connexions puntejades (connexions dins la xarxa PIC), les quals es troben encerclades en els extrems indicant que formen part d'una agregació de línies 802.3ad. L'objectiu d'aquesta connexió és interconnectar tots els servidors dins la xarxa del PIC ubicats a l'*stack* 3com amb la resta de la xarxa. Cal saber que hi ha múltiples agregacions de línies, algunes on es detalla cada línia individual i d'altres on només es mostra una connexió i hi posa Nx1Gbps, cosa que significa una agregació de N línies d'1 Gbps.

10 Un *stack* de switch es forma quan s'apila més d'un switch i aquests es comporten com si d'un únic switch es tractés.

CASTOR

Arquitectura

En la figura 4.1.6 es poden observar els flux de dades implicats en els tres casos d'ús de CASTOR.

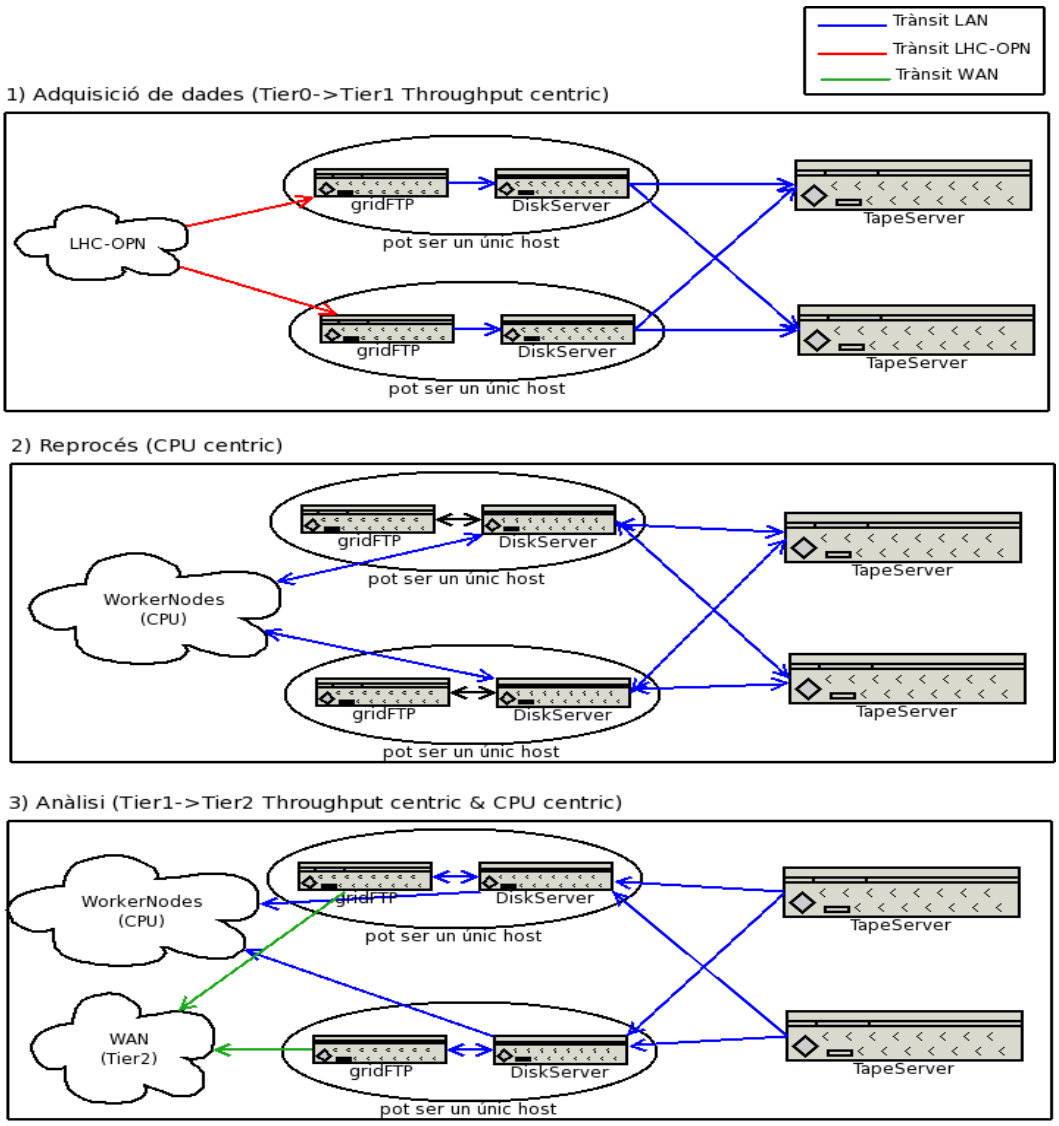


Figura 4.1.6: flux de dades dels casos d'us de CASTOR. L'adquisició de dades és el cas prioritari i més conflictiu a causa que és l'únic on actua la xarxa LHC-OPN

Donada la naturalesa dels diferents processos és poc probable que aquests hagin de conviure, tot i això la possibilitat existeix. També cal tenir en compte que l'adquisició de dades (cas 1) és prioritària i s'hi requereix una disponibilitat del 99%.

Solució 1: limitació DiskServer<=>WorkerNodes

Aquesta solució imposa un límit màxim de 8Gbps (Full duplex) per a les transferències DiskServer<=>WorkerNodes del reprocés i l'anàlisi local. En aquesta solució el problema dels certificats esmentat anteriorment queda reduït als GridFTP, on està completament resolt.

A continuació es mostren dos diagrames, segons si CASTOR agrupa (figura 4.1.7) GridFTP i DiskServer en un mateix servidor o no (figura 4.1.8). Aquesta diferència es planteja degut a que actualment encara no està definida l'arquitectura definitiva de CASTOR i ambdós escenaris són possibles.

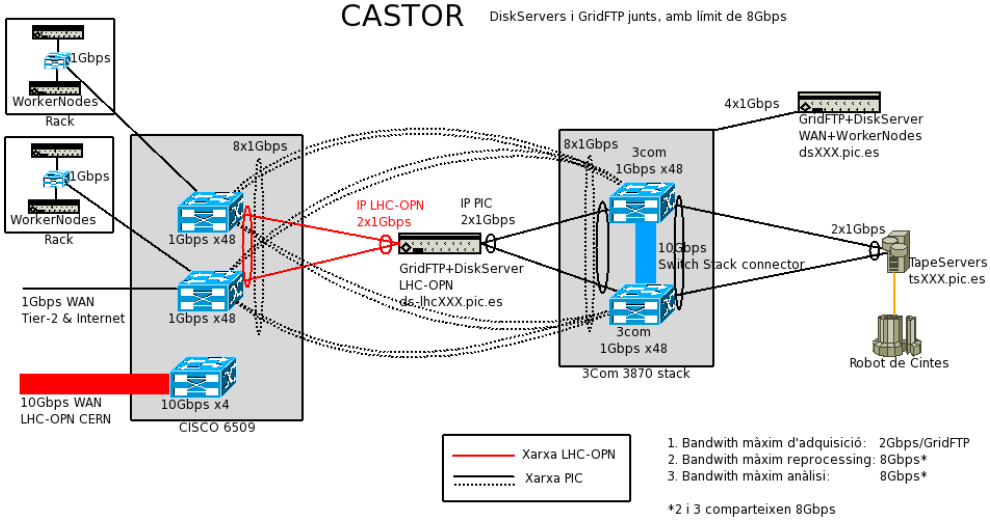


Figura 4.1.7: GridFTP i DiskServers agrupats en un mateix servidor amb limitació de 8 Gbps, on només els GridFTP tenen dues xarxes.

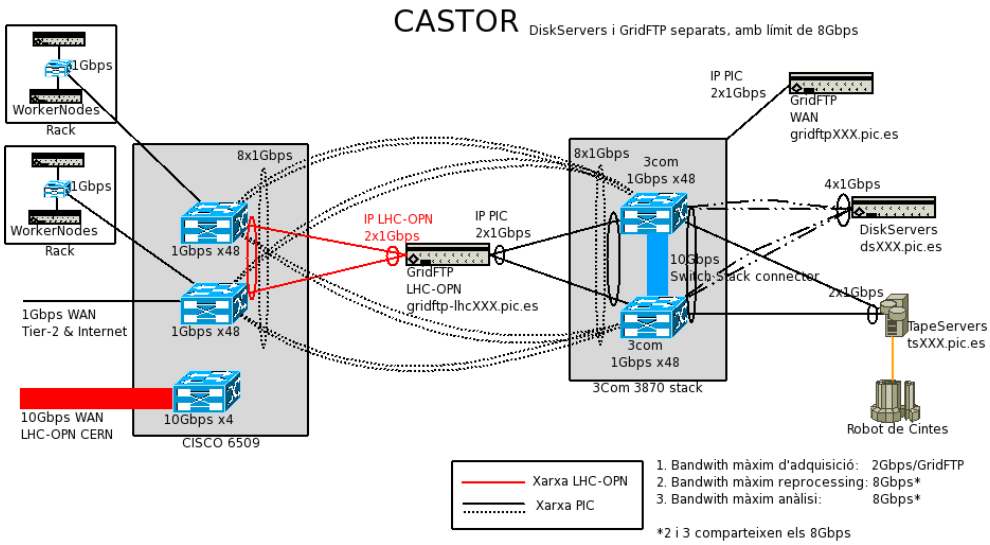


Figura 4.1.8: GridFTP i DiskServers en servidors diferents amb limitació de 8 Gbps, on només els GridFTP tenen dues xarxes.

El límit de 8Gbps indicat ens els diagrames és degut a que no és possible realitzar agregació de línies (802.3ad) amb més de 8 ports. En els diagrames l'agregació de línies (EtherChannel) es mostra com una sèrie de connexions encerclades en els extrems i/o una

etiqueta que indica *Nºlínies*x*Velocitat/línia*.

Solució 2: DiskServer amb dues IPs

Aquesta solució és una evolució de la solució anterior que elimina el límit dels 8Gbps en les transferències del reprocessat i d'anàlisi amb els WorkerNodes locals. S'afegeix un grau més de complexitat que permet que la solució escali a mesura que s'hi afegeixen nous GridFTP i DiskServer, incrementant l'ús de ports en el Cisco i en el 3Com 3870 *stack*. Com a contrapartida tenim que amb aquesta solució es podria presentar el problema dels certificats esmentat anteriorment.

A continuació es mostren dos diagrames, segons si CASTOR agrupa (figura 4.1.9) GridFTP i DiskServer en un mateix servidor o no (figura 4.1.10).

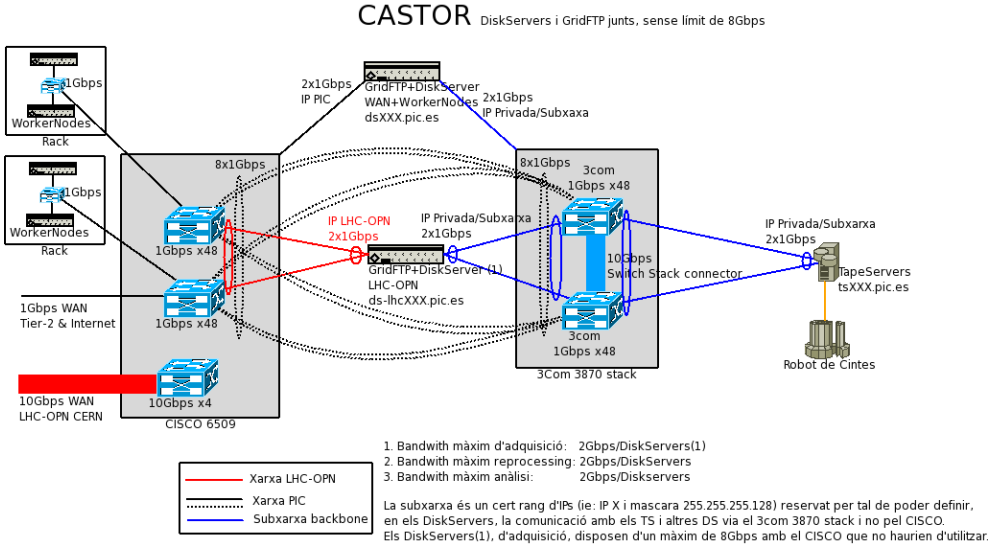


Figura 4.1.9: GridFTP i DiskServers agrupats en un mateix servidor, amb dues Ips. Es necessita establir una nova xarxa per a CASTOR

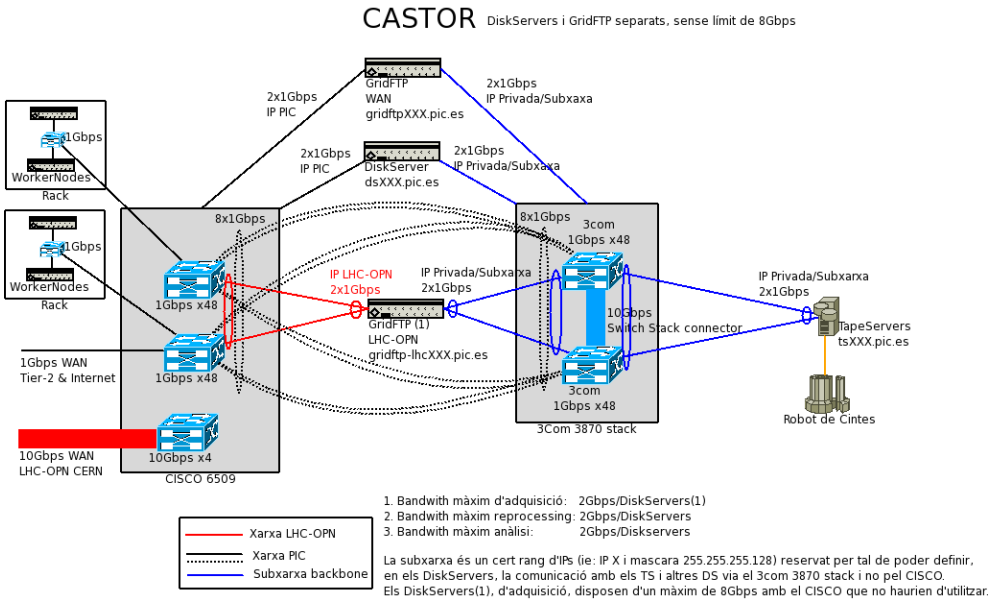


Figura 4.1.10: GridFTP i DiskServers en servidors diferents, amb dues IPs. Es necessita establir una nova xarxa per a CASTOR

dCache

Arquitectura

En la figura 4.1.11 es poden observar els flux de dades implicats en els tres casos d'ús de dCache.

Al igual que en el cas de CASTOR, donada la naturalesa dels diferents processos és poc probable que aquests hagin de conviure, tot i això la possibilitat existeix i cal tenir en compte que en l'adquisició de dades (cas 1) és el més important i s'hi requereix una disponibilitat del 99%.

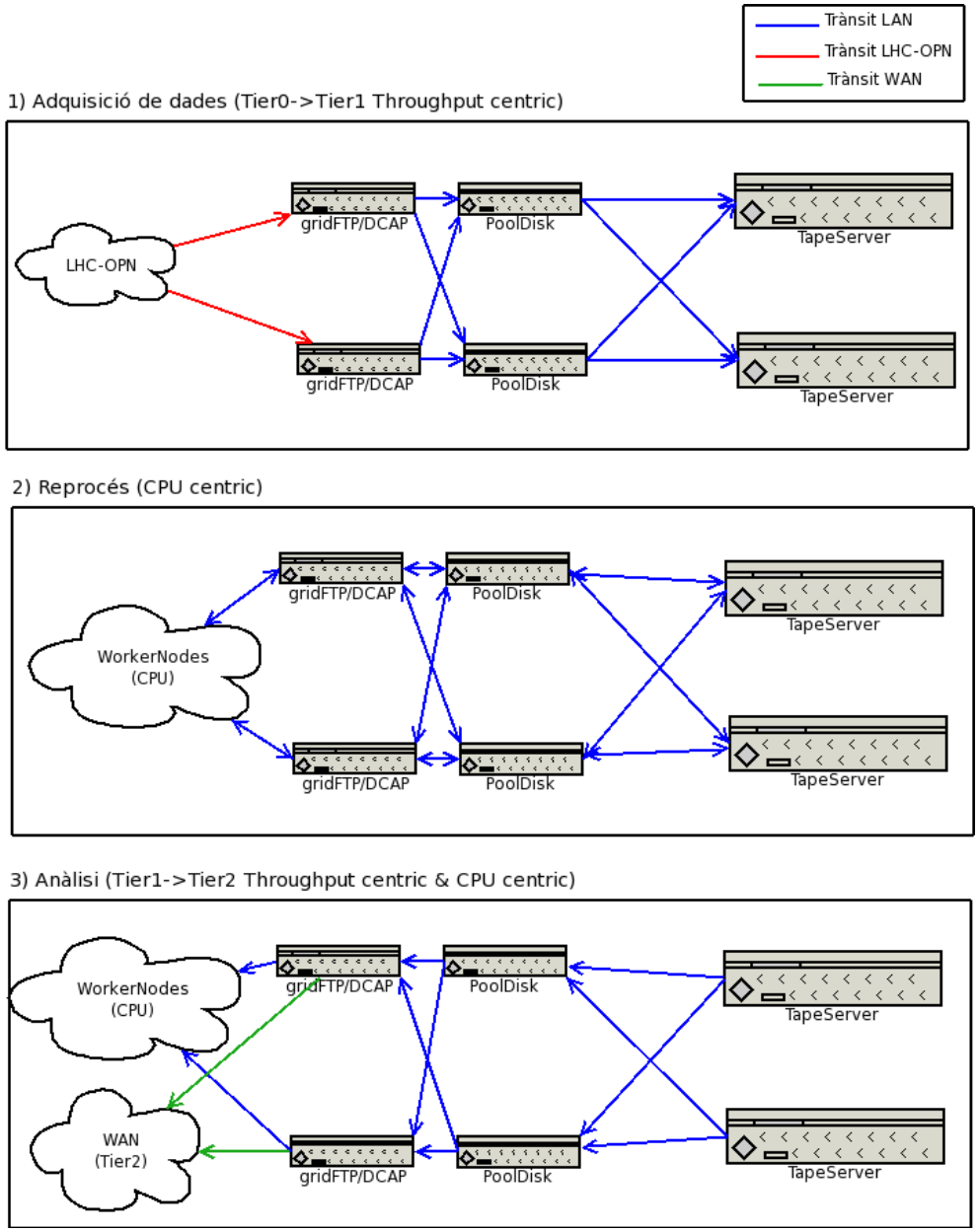


Figura 4.1.11: flux de dades dels casos d'us de dCache. El cas prioritari és el d'adquisició de dades

Solució 1: limitació DCAP<=>WorkerNodes

Aquesta solució imposa un límit màxim de 8Gbps (Full duplex) per a les transferències DiskServer<=>WorkerNodes del reprocés i l'anàlisi local. El problema dels certificats només apareix en els GridFTP connectats a la xarxa LHC-OPN.

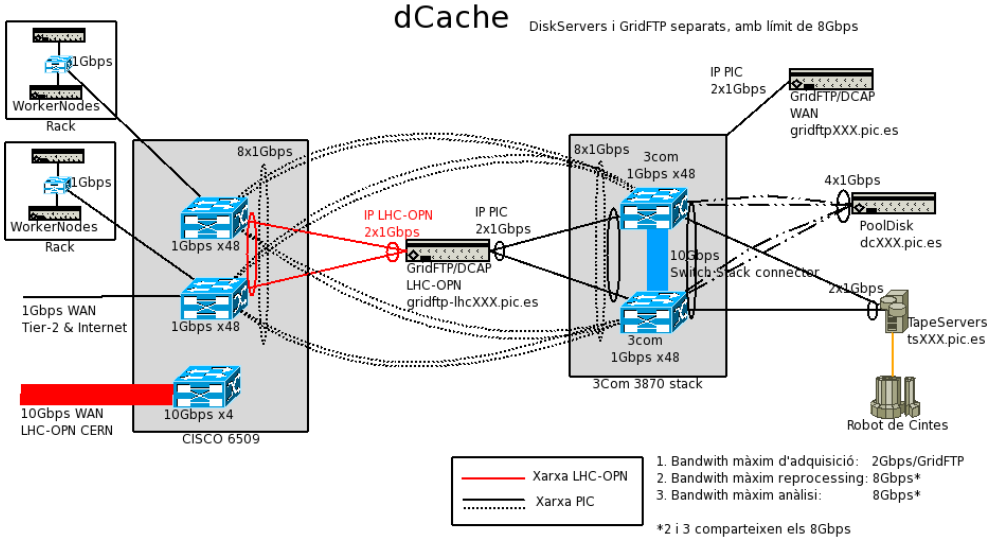


Figura 4.1.12: dCache amb limitació de 8Gbps. Només s'utilitza la xarxa del PIC i, en els GridFTP, també la LHC-OPN

Solució 2: DCAP amb dues IPs

En aquesta solució, malgrat són necessaris més ports en el Cisco6509, s'elimina el límit dels 8gbps de la solució anterior. El problema dels certificats apareix en tots els GridFTP i servidors DCAP.

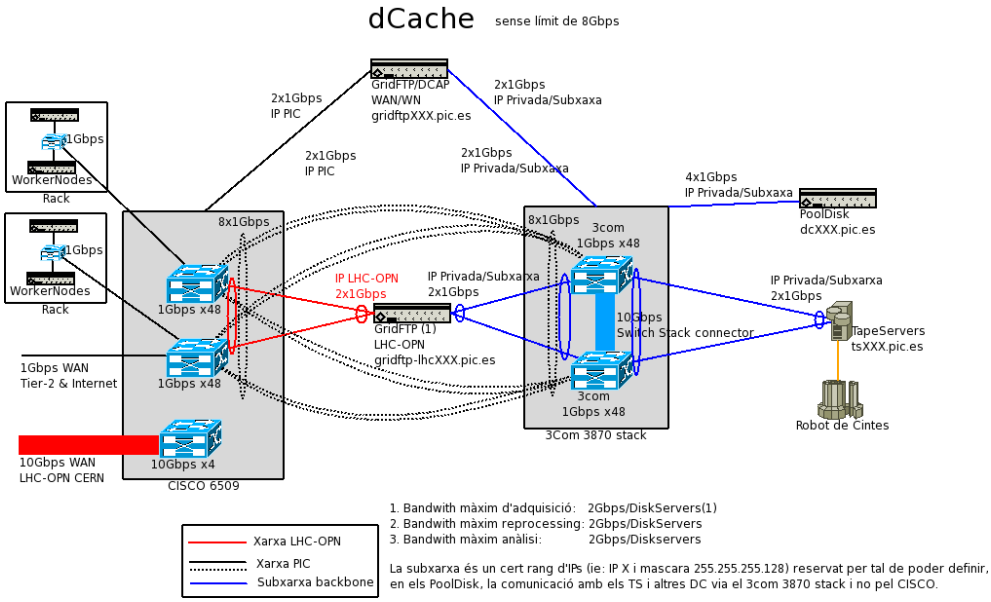


Figura 4.1.12: dCache amb servidor DCAP de dues IPs. Es necessita establir una nova xarxa per a la gestió interna de dCache

4.2 Pla d'implementació per a la integració dels serveis del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps

L'objectiu d'aquesta segona secció és documentar la segona etapa de la segona fase del projecte, on es defineix un pla per a la integració dels serveis del PIC i la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps.

En aquesta segona etapa es parteix de les solucions detallades en la secció anterior, tant per al desplegament de la connexió del circuit dedicat (veure 4.1.3) com per a la integració dels sistemes de gestió de dades del PIC a la xarxa LHC-OPN (consultar 4.1.4).

En aquesta secció es resumiran els trets principals de les solucions escollides, incloent el disseny del sistema de certificació per al circuit dedicat de 10 Gbps. Un cop definida la solució escollida s'aprofundirà en els detalls de la implementació i es plantejarà una metodologia per a la integració dels serveis de gestió de dades del PIC (dCache o CASTOR2) a la xarxa LHC-OPN.

4.2.1 Descripció de la solució escollida

Amb la direcció del centre s'ha acordat que per al desplegament circuit dedicat de 10 Gbps es seguirà la planificació proposada en l'estudi de solucions, descrita en l'annex G. L'execució del pla es durà a terme notificant i coordinant les accions amb l'Anella Científica, RedIRIS i el CERN, on el contacte oficial per a serà l'administrador de la xarxa del PIC, Diego Dávila. Tot i això, a excepció de la configuració de l'encaminador cisco6509 del PIC, les accions tècniques seran dutes a terme per l'alumne del PFC, Gerard Bernabeu.

S'ha acordat que l''Anella Científica, en nom de la UAB, serà qui gestioni la petició d'IPs i del Sistema Autònom a RIPE. Seguint el consell de l'Anella Científica, i en concordança amb els plans de futur del PIC, es realitzarà la petició d'una classe B (/22) d'adreces IPv4 del tipus PI[11] associada a un AS públic.

Per a la certificació del circuit dedicat i la connexió LHC-OPN a 10 Gbps el CERN ha posat a disposició del PIC un servidor amb una connexió a la xarxa LHC-OPN de 10 Gbps. L'arquitectura del sistema de certificació està detallada a la subsecció 4.2.2.

Finalment el sistema de gestió de dades del PIC escollit és dCache. La solució per a la integració dels sistemes de gestió de dades del PIC amb la xarxa LHC-OPN sobre dCache preferida és la segona proposada en la secció anterior: “DCAP amb dues IPs” (veure subsecció 4.1.4). També s'ha decidit la creació d'una maqueta, detallada a la subsecció 4.2.3, per a provar la implementabilitat de la solució “DCAP amb dues IPs”.

Un cop decidit l'ús de dCache i refinades les especificacions de la seva implementació s'ha generat un nou disseny per a la integració del servei amb la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps. El sistema de gestió de dades dCache dissenyat, encarregat de gestionar les dades provinents

11 L'adreçament PI (*Provider Independent)* permet l'associació a de les IPs a qualsevol Sistema Autònom, independentment del proveïdor d'accés a la Internet.

de la xarxa LHC-OPN, s'anomena “Receptor de dades T0-T1” i està detallat a la subsecció 4.2.3. És important esmentar que en el marc del PFC del “Receptor de dades T0-T1” només se'n plantejarà el disseny, la implementació no està inclosa dins l'abast d'aquest projecte i es durà a terme pels diferents equips de treball del PIC (veure capítol 1).

4.2.2 Sistema de certificació per al circuit dedicat de 10 Gbps

Per a la certificació del circuit dedicat PIC-CERN de 10 Gbps, el CERN ha proporcionat un servidor (*hufsa.cern.ch*) Intel(R) XEON(TM) CPU 2.40GHz amb una NIC de 10 Gbps i *iperf* versió *2.0.2 (03 May 2005) pthreads* instal·lat. Així doncs la certificació es realitzarà segons la opció 2 (1 servidors dedicat amb una NIC TenGigabitEthernet al CERN) proposada a l'apartat “*Certificació del circuit dedicat a 10 Gbps*” de la subsecció 4.1.3.

Donada la possibilitat de patir retrassos en la obtenció i configuració de l'AS i les IPs associades a la connexió sobre el circuit dedicat de 10 Gbps i per tal de poder realitzar la certificació del circuit tan bon punt estigui disponible, el CERN ha cedit temporalment al PIC un petit CIDR (192.16.166.240/28), amb capacitat per a 16 IPs, que s'utilitzarà per al direccionament dels servidors del PIC del sistema de certificació dissenyat.

En la figura 4.2.1 es pot observar un diagrama del sistema de certificació del circuit dedicat de 10 Gbps a implementar, incloent la infraestructura al PIC i al CERN. Donat que les IPs utilitzades al PIC són de la xarxa del CERN, les proves de certificació del circuit dedicat de 10 Gbps es podran realitzar tan bon punt es disposi de la connexió punt a punt PIC-CERN operativa.

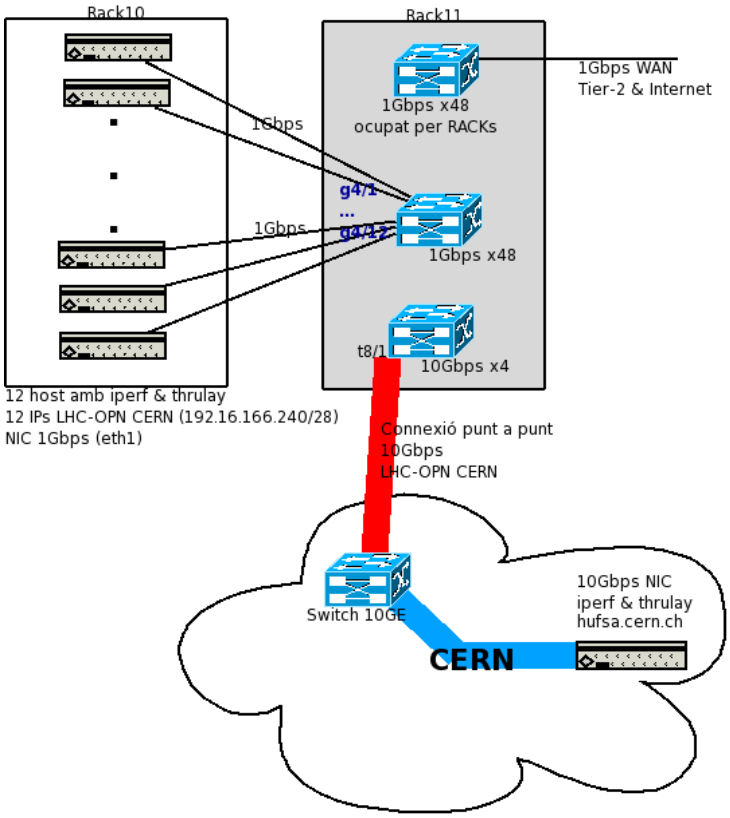


Figura 4.2.1: sistema de certificació per al circuit dedicat de 10 Gbps. Al PIC es disposa de 12 servidors amb iperf que enviaran 12 flux de dades TCP i/o UDP a ~1 Gbps cap al servidor amb NIC de 10 Gbps del CERN.

Per a la realització de la certificació del circuit dedicat es realitzaran proves unidireccionals en ambdós sentits (PIC-CERN i CERN-PIC) i bidireccionals.

Per als test unidireccionals PIC-CERN es crearan dotze flux de dades TCP i/o UDP d'aproximadament 1 Gbps cadascun, fent un total d'uns 10-12 gbps i saturant la connexió en sentit PIC-CERN. En sentit contrari (CERN-PIC) es realitzarà una connexió des del servidor del CERN a cada servidor del PIC (fins un total de 12 connexions), és possible no poder certificar el circuit a 10 Gbps en sentit CERN-PIC degut al rendiment del servidor del CERN (*hufsa.cern.ch*). Finalment, per a la certificació bidireccional es realitzaran els dos procediments citats anteriorment de forma simultània, novament és possible que la potència del servidor del CERN sigui insuficient i no es pugui generar un flux de dades de 10 Gbps en sentit CERN-PIC. En el cas de no disposar de capacitat de generació de trànsit suficient al CERN s'utilitzaran rutes estàtiques en els encaminadors per a crear un bucle entre la connexió punt a punt sobre el circuit dedicat de 10 Gbps, multiplicant així el trànsit generat pel valor inicial del TTL dels paquets.

Per tal de poder realitzar les proves correctament, els 12 servidors del sistema de certificació del PIC han d'estar sincronitzats. Per a tal efecte s'instal·larà *vxargs*[12] en tots els servidors, un programa que permet l'execució paral·lela de comandes, en aquest cas *iperf*.

4.2.3 Detalls d'implementació del receptor de dades T0-T1 de dCache

Finalment s'ha decidit que la integració del sistema de gestió de dades dCache del PIC amb la xarxa LHC-OPN només afecti al flux d'adquisició de dades Tier0->Tier1 (veure figura 4.2.2), creant així l'anomenat "Receptor de dades T0-T1", responsable exclusiu del procés d'adquisició de dades a la xarxa LHC-OPN.

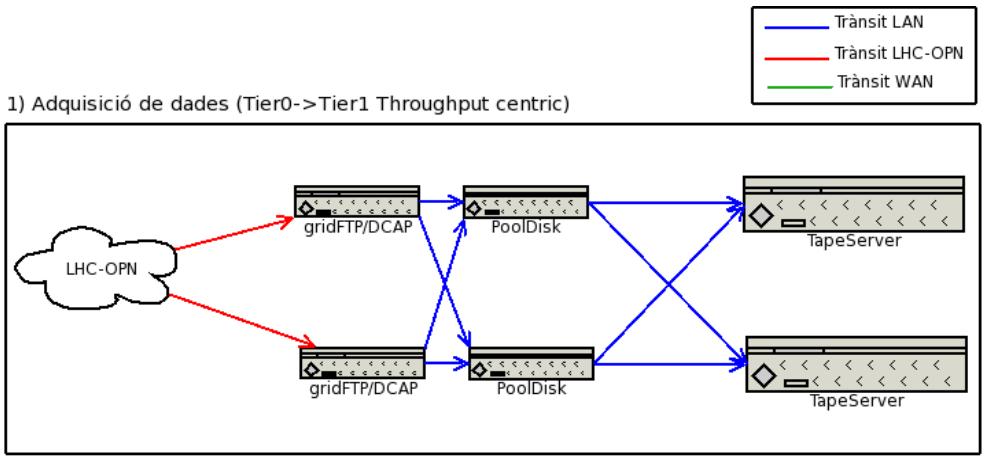


Figura 4.2.2: flux d'adquisició de dades Tier0->Tier1 que ha de suportar el sistema de gestió de fitxers de dades dCache. Totes les dades vindran de la xarxa LHC-OPN cap als gridFTP, els quals emmagatzemaran les dades als PoolDisk, d'on es migraran a cinta mitjançant els TapeServer.

L'arquitectura de dCache també ha patit algunes modificacions respecte les solucions plantejades a la subsecció 4.1.4, tot i això la solució 1 "limitació DCAP<=>WorkerNodes" s'aproxima força a l'arquitectura requerida pel Receptor de dades T0-T1. La modificació principal respecte la solució

12 Per a més informació sobre vxargs consultar el web http://dharma.cis.upenn.edu/planetlab/vxargs/

proposada en la secció anterior és l'eliminació dels servidors GridFTP/DCAP que donen servei a la WAN (xarxa no-LHC-OPN) i als WorkerNodes de l'*stack* 3com redundant, on si que es connecten els GridFTP receptors de dades de la xarxa LHC-OPN.

Amb la nova arquitectura de dCache, els servidors GridFTP/DCAP que donen servei als servidors no-LHC-OPN es connectaran a la xarxa estàndard del PIC mitjançant switch distribuïts pels racks del PIC, com si de servidors estàndards es tractés. Per al "Receptor de dades T0-T1" es realitzarà un muntatge especial mitjançant un *stack* 3com de switch 5500G[13], el qual servirà de nucli de xarxa per al sistema dCache. En la figura 4.2.3 es mostra un diagrama del redisseny de la solució, amb una simplificació del "Receptor de dades T1-T0".

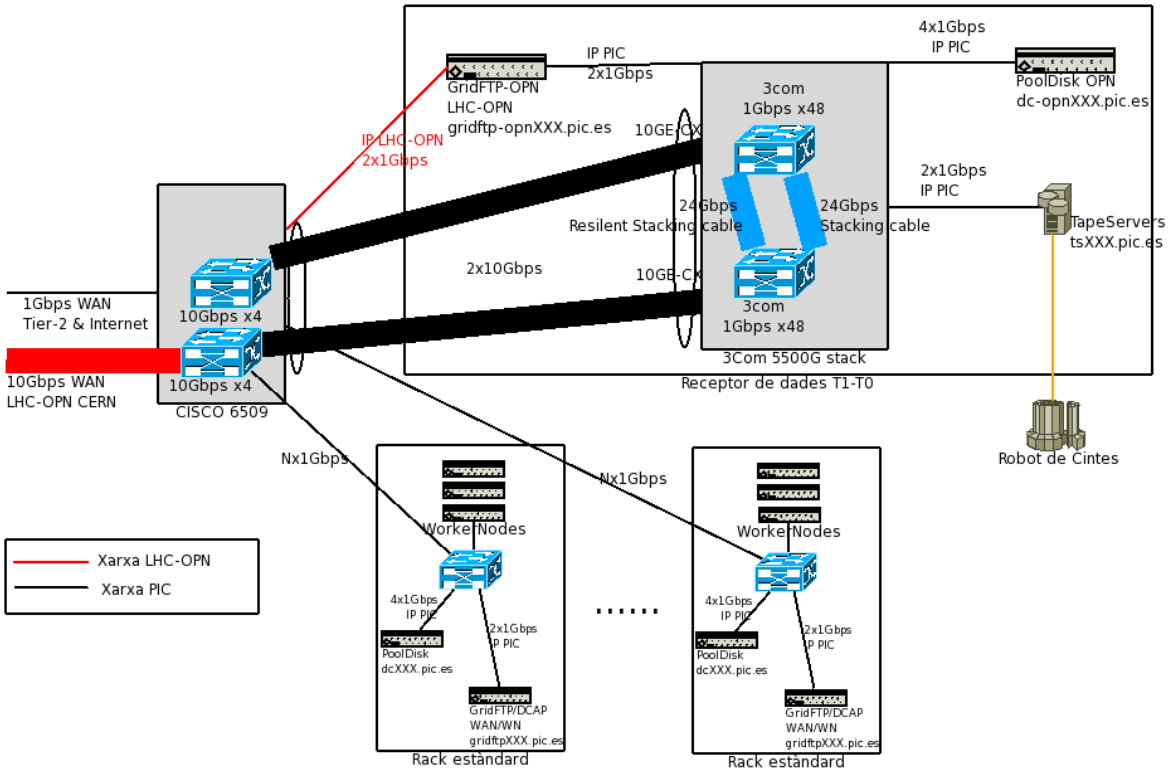


Figura 4.2.3: integració del sistema dCache amb la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps. Les dades provinents de la xarxa LHC-OPN (en vermell) entraran a l'encaminador del PIC mitjançant el circuit dedicat de 10 Gbps cap als servidors GridFTP-OPN, mitjançant la seva interfície LHC-OPN. Els servidors GridFTP-OPN processaran les dades i, mitjançant un protocol intern de dCache, emmagatzemaran les dades als PoolDisk OPN que posteriorment les migraran a cinta mitjançant els TapeServers. Així doncs les dades només passaran un cop pel cisco6509 per anar las GridFTP-OPN i després totes les gestions de dCache relacionades amb la recepció de dades LHC-OPN es realitzaran entre servidors connectats a l'*stack* d'alta disponibilitat 3com 5500G.

Per entendre el disseny detallat del receptor de dades T0-T1 cal tenir en compte els requeriments imposats:

1. La disponibilitat del sistema ha de ser del 99% (màxim de 4 dies de fallida en un any, excloent les aturades programades)
2. El sistma tindrà una càrrega de recepció de dades de 16 TBytes diaris

13 Inicialment el disseny contemplava un *stack* amb uns switch 3com 3870 disponibles al PIC, però degut a la necessitat de més ample de banda entre els membres de l'*stack* i als requeriments de fiabilitat s'ha optat per l'adquisició del model superior 5500G-EI, amb tecnologia XRN.

3. Els servidors PoolDisk seran Thumper (SUN Fire X4500), amb una capacitat de 16TB nets per servidor. La capacitat màxima de xarxa és d'uns 470 Mbyte/segon agregant les 4 NIC GigabitEthernet (802.3ad).
4. El rendiment dels servidors GridFTP s'estima en uns 40 Mbyte/segon
5. Un TapeServer processa com a mínim 30 Mbytes/segon (es disposarà d'un mínim de 12 TapeServer)
6. S'ha de disposar de suficient espai als PoolDisk per suportar una fallida de sis dies en l'escriptura a cinta.

La sisena restricció implica disposar de 16*6 TBytes de disc en els PoolDisc, és a dir, 6 Thumpers. Una recepció de dades de 16 TBytes diaris significa, en el cas d'obtenir un flux de dades constant durant les 24h, una càrrega aproximada[14] de 200 MB/segon (1,6 Gbps), que balancejada correctament representa uns 34MB/s (0,27Gbps) per PoolDisk (en el cas de tenir-ne 6 operatius).

La càrrega de 200MB/segon es tradueix en un mínim de 5 GridFTP i uns 7 TapeServer operatius, dedicats per al receptor de dades T0-T1. Donada la primera restricció (99% de disponibilitat) cal disposar d'una arquitectura redundant, cosa que implica sobredimensionar la solució i disposar de més GridFTP i TapeServer. Donada la restricció 6, els PoolDisk ja es troben sobredimensionats, tot i això la solució dissenyada permet doblar la quantitat de servidors PoolDisk (de 6 a 12), permetent 6 o més dies d'espai de disc en cas de fallida dels servidors d'un dels racks del receptor de dades T0-T1 (per exemple per problemes elèctrics en un rack).

La solució dissenyada per al "Receptor de dades T0-T1" correspon al diagrama de la figura 4.2.4 i disposa de capacitat per a 18 GridFTP, 6[15] PoolDisk (Thumper=Sun fire X4500) i 18 TapeServer, oferint 3 GridFTP i TapeServer per PoolDisk. Cal tenir en compte que els racks disposen de 35 U de capacitat neta[16] i que, amb la solució completament desplegada la ocupació és:

- Part alta: 14/17 U
- Part baixa: 2/4 Thumper (8/16U) + 1/1U (switch 3com 5500G)

14 16TB = 16384Gbytes/dia * 1dia/86400s * 8bits/1byte =~ 1,6 Gbps = 200MBytes/s

15 Ampliable a un màxim de 12 Thumper entre els 3 racks, revisant la instal·lació elèctrica dels racks afectats

16 Un cop descomptades les U utilitzades per als passacables.

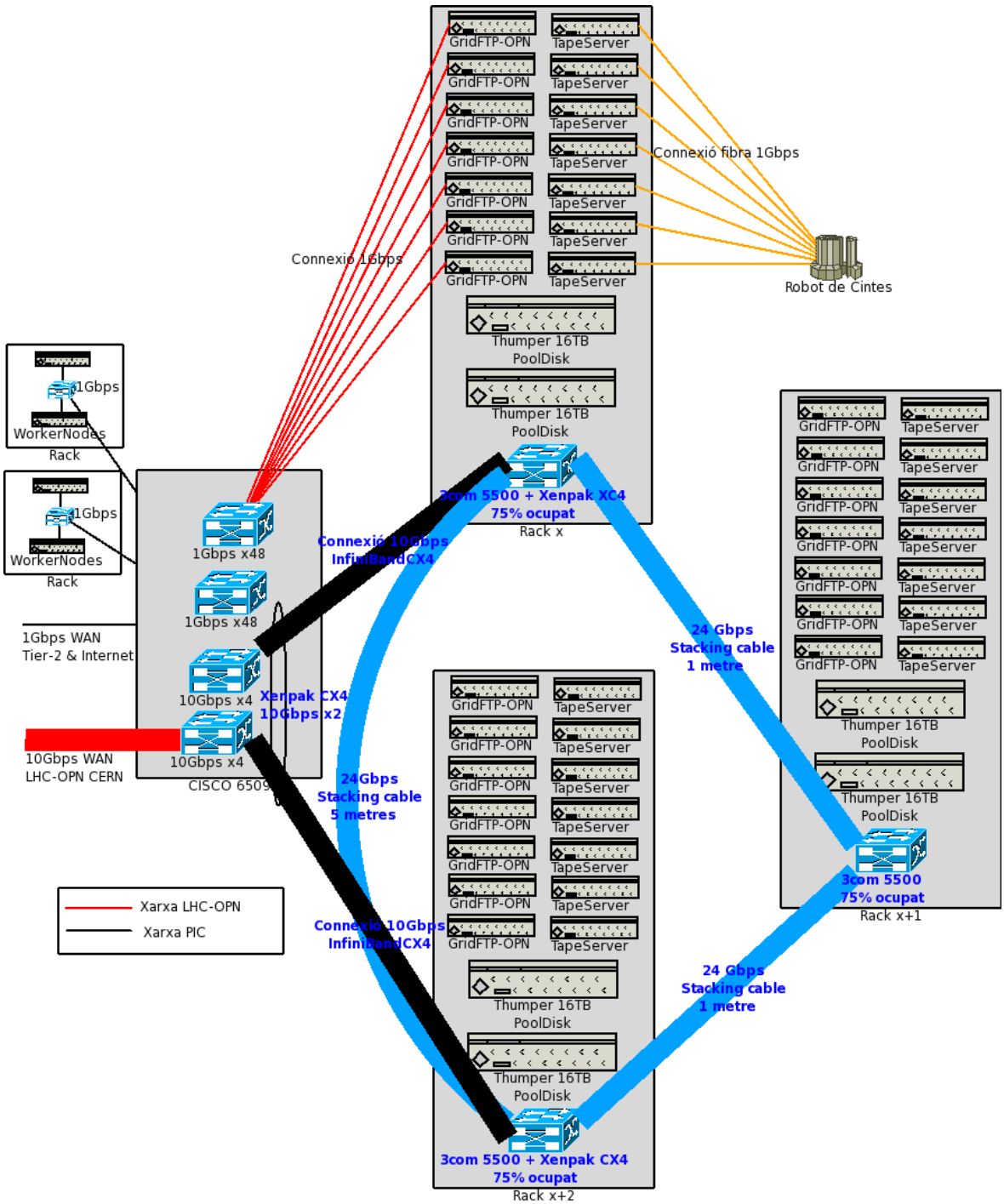


Figura 4.2.4: Solució dissenyada per al "Receptor de dades T0-T1". El nucli de xarxa de la solució està compost per un *stack* de switch 3com 5500G-EI amb tecnologia XRN, que ofereix una interconnexió entre switch a 48Gbps (Full Duplex).

La connexió de 2x10 Gbps entre l'*stack* 3com i l'encaminador cisco6509 és utilitzada pels TapeServers de dCache durant els casos d'us de reprocés i anàlisi de dades emmagatzemades en cinta. Les dades extretes pels TapeServer seran enviades als PoolDisk no-LHC-OPN, distribuïts per la resta de la xarxa del PIC.

Per a poder monitoritzar el bon funcionament de la xarxa LHC-OPN sobre el circuit dedicat de 10

Gbps es desenvoluparà un sensor de Nagios i el seu corresponent procediment, on s'establirà un protocol d'actuació.

Per a demostrar la validesa funcional de la solució proposada s'implementarà la maqueta del receptor de dades de la figura 4.2.5. Tal i com es pot observar en el diagrama, en la maqueta no s'hi ha inclòs els TapeServer (per motius de disponibilitat). Aquest fet no representa un problema ja que els únics servidors del receptor de dades que necessiten una configuració diferent a la que disposa l'actual dCache en producció són els servidors GridFTP-OPN, pel fet de disposar de dues interfícies.

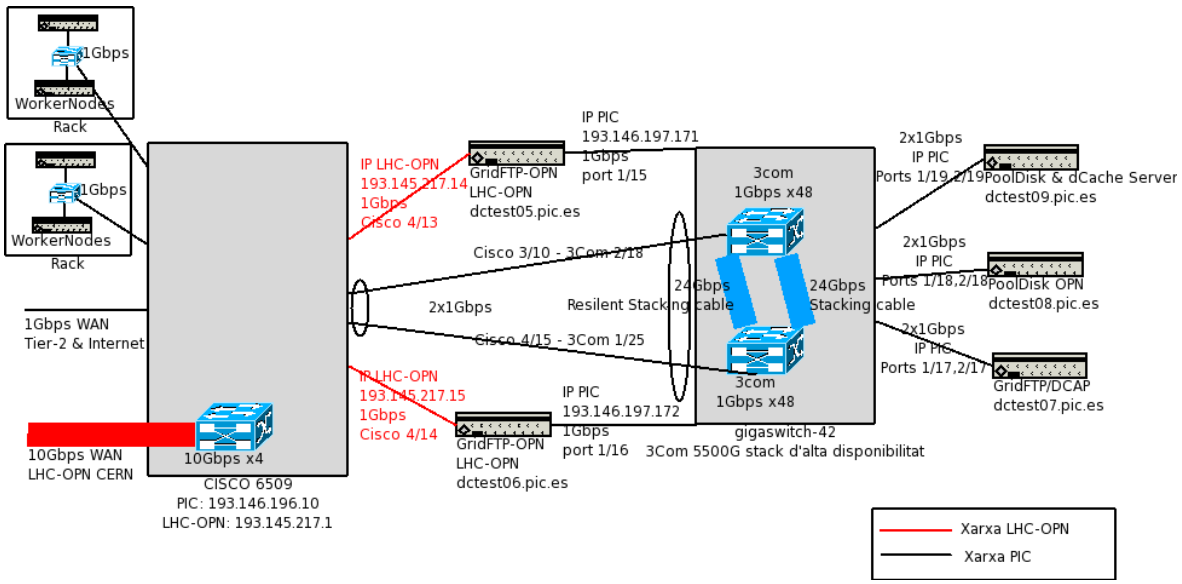


Figura 4.2.5: Maqueta del "Receptor de dades T0-T1". La principal diferència de configuració entre la solució de dCache proposada i la configuració estàndard es troba en els servidors GridFTP, que han de disposar de dues interfícies.

4.2.4 Metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN

Per tal de poder dur a terme la integració de dCache (implementació del receptor de dades T0-T1) cal adquirir tot el material necessari. A la taula de la figura 4.2.6 es mostren els dispositius de xarxa necessaris, la resta de necessitats per al muntatge dels servidors seran establertes pels responsables del desplegament de dCache.

Component	**Unitats**
3Com® Switch 5500G-EI	3
3Com® 5500G-EI Stacking Cable (2*1,5m +1*5m)	3
3Com® Switch 5500G-EI 1-Port 10G Module	2
3Com® 10GBASE-CX4 XENPAK	2
Cable CX4 estàndard de 7-15 metres[17]	2
Cisco 10GBASE-CX4 XENPAK	2

Figura 4.2.6: Taula amb tot el material necessari per al desplegament de la xarxa per al Receptor de dades T0-T1

17 La longitud del cable dependrà de la ubicació dels switch respecte al cisco6509.

Per tal d'integrar el servei dCache a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps mitjançant el "Receptor de dades T0-T1" detallat en la subsecció anterior cal seguir el següent procediment:

1.Preparació dels racks on s'allotjarà el "Receptor de dades T0-T1". Cal obtenir tres racks contigus des dels quals es pugui arribar al cisco6509 amb un cable de coure del tipus CX4 (màxim 15 metres). Aquests racks s'han d'alliberar completament i, si és possible, reforçar-ne la potència elèctrica disponible.

2. Muntar la xarxa del "Receptor de dades T0-T1"

a) Inserir i configurar dins la VLAN100 del cisco6509 els XENPAK 10GBASE-CX4
b) Traçar dos cables CX4 des del cisco6509 a dos racks del receptor de dades.
c) Traçar N cables Cat.6 amb connectors RJ45 des del switch a cada rack del receptor, on N és el número de servidors GridFTP-OPN continguts en cada rack.
d) Muntar el "*3Com® Switch 5500G-EI 1-Port 10G Module*" i "*3Com® 10GBASE-CX4 XENPAK"* en dos dels switch 3com 5500G-EI.
e) Instal·lar un switch 3com5500G-EI en cada rack del receptor de dades i interconnectar-los entre si mitjançant "*3Com® 5500G-EI Stacking Cable (2*1,5m +1*5m)*".
f) Habilitar LACP[18], mode actiu, en tots els ports de l'*stack* 3com.
g) Habilitar LACP, mode actiu, en els ports 10GE del Cisco6509 on es connectaran els cables CX4.
h) Connectar els cables CX4 en el Cisco6509 i en els switch 3com 5500G-EI preparats amb el mòdul de 10 Gbps i el XENPAK.

3. Muntar els servidors del receptor de dadesT0-T1.

i) Muntar els servidors PoolDisk, creant una interfície (no-LHC-OPN) de bonding del tipus agregació de línies 802.3ad (LACP).
j) Muntar els servidors GridFTP, creant dues interfícies (LHC-OPN i no-LHC-OPN) de bonding del tipus Active-Backup[19].
k) Muntar els servidors TapeServer, creant una interfície (no-LHC-OPN) de bonding del tipus Active-Backup (el màxim d'escriptura a cinta és 80MB/segon~= 0,64Gbps)

Un cop el "Receptor de dades T0-T1" passi la fase de test es podrà posar en producció i aturar la recepció de dades per la connexió general del PIC, de 1 Gbps.

Cal recordar que la implementació del "Receptor de dades T0-T1" no està inclosa dins l'abast del PFC, es durà a terme pels equips de treball del PIC.

18 Link Aggregation Control Protocol: un protocol estàndard per a negociar l'agregació de línies 802.3ad, per a més informació consultar subsecció 2.7.3

19 En el cas de que en un futur es detectés que l'ample de banda utilitzat pels servidors s'aproxima a 1 Gbps el tipus de bonding es canviaria per un que agregués el trànsit, com ara LACP, round-robin, etc.

4.3 Informe de la execució del pla d'implementació sobre el circuit de dedicat de 10 Gbps

En aquesta secció es documenta el procés d'execució del pla d'implementació detallat en la secció anterior (secció 4.2).

Es descriuran els canvis d'especificacions soferts respecte la planificació inicial, les incidències ocorregudes i l'estat resultant de l'execució del pla d'implementació, incloent els resultats de la implementació del sistema de certificació i de la maqueta que prova la validesa del Receptor de dades T0-T1, presentat en la subsecció 4.2.3.

4.3.1 Modificacions respecte la planificació inicial

Degut als tràmits burocràtics, la petició de les IPs *Provider Independent* i del Sistema Autònom a RIPE ha patit força retrassos. Com a resultat d'aquest fet el desplegament de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps amb l'AS propi del PIC ha quedat ajornat a dates posteriors a la finalització del projecte (probablement es realitzarà al llarg del mes d'agost).

Per tal de no afectar l'objectiu principal d'aquesta segona fase del PFC, que és certificar el circuit dedicat de 10 Gbps PIC-CERN, s'ha modificat la planificació inicial per tal de realitzar les proves de servidor a servidor amb IPs cedides pel CERN, evitant així la necessitat de disposar de l'AS i les IPs pròpies del PIC.

En la figura 4.3.1 es pot observar com la planificació inicial (consultar annex G), a part de per les IPs i l'AS, també s'ha vist afectada pels retrassos en l'entrega del circuit dedicat PIC-CESCA per part d'Al-Pi, el proveïdor d'aquest segment del circuit dedicat de 10 Gbps.

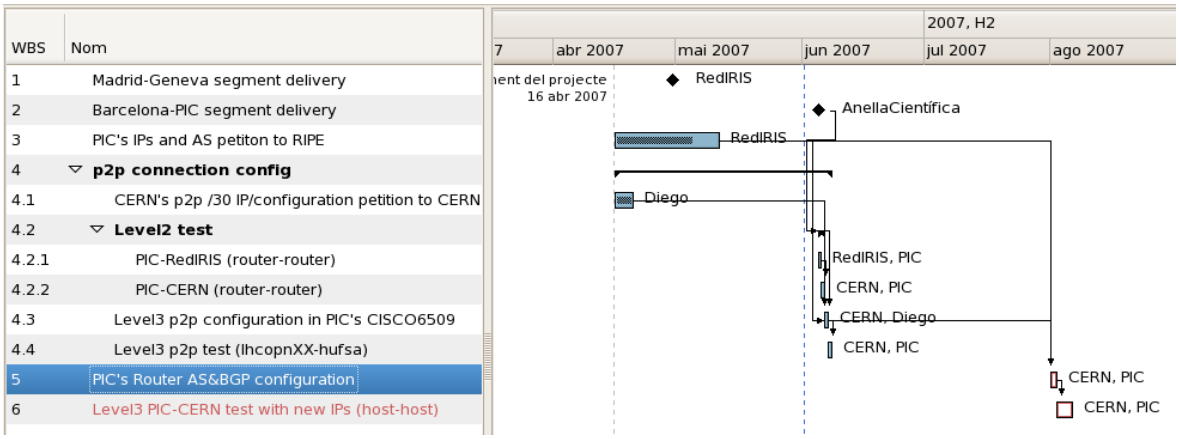


Figura 4.3.1: Planificació corregida del desplegament del circuit dedicat de 10 Gbps. Es pot observar com la configuració de l'AS (BGP) s'ha retrassat fins al mes d'agost, poterior a la finalització del PFC. Cal tenir en compte que el punt 3 (*PIC's IPs and AS petition to RIPE*) només inclou la sol·licitud de les adreces i l'AS, la ressolució d'aquesta depen exclusivament de RIPE.

4.3.2 Execució del pla d'implementació

L'execució del pla d'implementació per al desplegament del circuit dedicat a 10 Gbps s'ha dividit en diverses tasques, organitzades en dos grups: el desplegament del circuit dedicat (veure taula de la figura 4.3.2) i la integració dels serveis del PIC a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps (veure taula de la figura 4.3.3).

[id] Tasca	**Observacions**	**Incidències**
[1] Implementació i prova del sistema de certificació per al circuit dedicat de 10 Gbps	S'ha realitzat amb èxit la implementació, les proves de baixa velocitat via el circuit dedicat d'1 Gbps i les proves d'alta velocitat dins la LAN. Els detalls es poden consultar a l'annex H.	
[2] Instal·lació i certificació de la connexió redundant de fibra òptica monomode del rack d'Al-Pi al rack 10 del PIC	La instal·lació s'ha dut a terme segons l'explicat a la subsecció 4.1.1. En el panell de connexions les fibres 1 a 12 són d'uns 40,7 metres de longitud, les fibres 13 a 24 d'uns 38,3 metres. Tots els cables han estat certificats segons *ISO 11801:2002* a 1310 nm i 1550 nm, amb una pèrdua d'entre 0,06dB (fibra 14) i 1,25 dB (fibra 12).	1
[3] Desplegament del segment Madrid (RedIRIS) – Ginebra (GÉANT)	El dia 28/05/07 Esther Robles de RedIRIS confirma la finalització del desplegament del segment, disposant ja de connectivitat des del CERN fins a l'Anella Científica (Pedralbes, Barcelona)	
[4] Desplegament del segment Bellaterra (PIC) – Pedralbes (Anella)	A data 5/06/07 Al-Pi finalitza el desplegament del segment, instal·lant a la sala de comunicacions de la UAB un equip DWDM Nortel Optera Metro 5200, dissenyat per a xarxes metropolitanes i capaç de generar 32 longituds d'ona (lambdes) de fins a 10 Gbps d'ample de banda. Per al PIC es proveeix una única connexió de 10 Gbps mitjançant una fibra òptica monomode al nou panell de connexions (tasca 2). Es realitzen proves d'ample de banda entre els encaminadors del segment de fins a 1 Gbps (veure incidència 3)	2, 3
[5] Configuració de la connexió Punt a Punt amb el CERN (configuració del cisco6509 del PIC)	A data 6/06/07 finalitza la configuració de la connexió punt a punt amb el CERN, sobre les IPs 192.16.166.56/30 (.57 al CERN i .58 al PIC). Per al tram PIC-Anella Científica s'ha definit un *trunk*[20] per on la vlan287 transporta les dades de	

	la xarxa LHC-OPN. A l'Anella Científica es desetiqueta la vlan287 i les dades s'envien pel circuit dedicat fins al CERN, mitjançant l'equipament òptic de RedIRIS10, GÉANT2 i el CERN.	
[6] Prova de comunicació entre els encaminadors	Les rutes són correctes, els detalls es troben a la subsecció 4.3.4	
[7] Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps.	Amb les proves de certificació s'ha aconseguit certificar la connexió, de ~30ms de latència, a 9,2 Gbps. Més informació a la subsecció 4.3.5 i a la secció 4.4. Durant la realització de les proves de rendiment a mig termini s'ha detectat la incidència 4 (consultar subsecció 4.3.3). La certificació bidireccional amb transferències a ~10 Gbps no s'ha pogut realitzar a temps a causa de la incidència 5.	4,5

Figura 4.3.2: Taula resum de l'estat final de l'execució de les tasques de la Fase 2 relacionades amb el desplegament del circuit dedicat

[id] Tasca	**Observacions**	**Incidències**
[8] Configuració de la xarxa per a la maqueta del Receptor de dades T0-T1	La configuració s'ha realitzat sobre un *stack* de switch 3com 3870 i, gràcies a un préstec de 3com, sobre un *stack* de 3com 5500G-EI amb tecnologia XNR. La configuració s'ha realitzat segons el disseny de la subsecció 4.2.3 i seguint parcialment la metodologia de 4.2.4 (en la maqueta no es disposa de connexions a 10 Gbps), quedant la configuració de la xarxa de la maqueta com es mostra a la figura 4.3.4 També s'ha comprovat el rendiment de xarxa dels Thumper (SUN FIRE X4500) sota Solaris10 amb bonding[21], aconseguint velocitats sostingudes de 3,8 Gbps (475 Mbyte/segon).	6
[9] Comprovació de la implementabilitat de dCache sobre la maqueta	L'equip de treball responsable de dCache al PIC ha comprovat la implementabilitat de dCache sobre la maqueta del Receptor de dades T0-T1, essent les principals	

20 Un *trunk* és un canal (o connexió) per on poden passar una o més VLANs. Per a més informació sobre VLANs i *trunking* consultar la subsecció 2.3.3, en el capítol de fonaments teòrics.

21 En solaris s'anomena Trunking, a la documentació oficial de Solaris10 es pot trobar una guia per a la seva configuració: http://docs.sun.com/app/docs/doc/816-4554/6maoq01ne?a=view

del Receptor de dades T0-T1	modificacions necessàries les indicades a la secció 3 "*GridFTP with Pools in a Private Subnet*" del capítol 21 "*Complex Network Configuration*" del manual de dCache[25].	
[10] Desenvolupament del sensor Nagios de connectivitat amb la xarxa LHC-OPN i el seu corresponent procediment.	S'ha dissenyat i desenvolupat un sensor de Nagios que comprova la connectivitat amb el servei SRM dels diferents Tier1 i el Tier0, els detalls es troben a l'annex I.	

Figura 4.3.3: Taula resum de l'estat final de l'execució de les tasques de la Fase 2 relacionades amb la integració dels serveis del PIC a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps

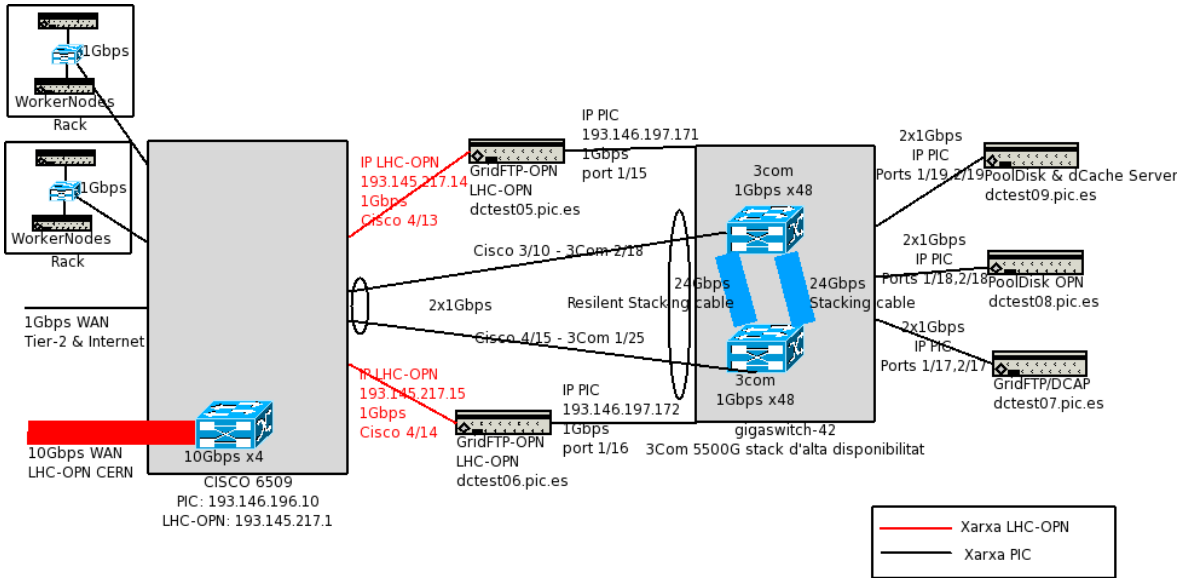


Figura 4.3.4: Maqueta implementada del Receptor de dades amb dCache. La única configuració específica del disseny es troba en els GridFTP, la resta de servidors tenen una configuració estàndard.

4.3.3 Incidències i resolució de les mateixes

En aquesta subsecció es detallen diverses incidències ocorregudes al llarg de la execució del pla d'implementació sobre el circuit dedicat de 10 Gbps.

Les incidències resoltes es mostren en verd, les pendents de resoldre al tancament de la fase 2 del projecte resten ressaltades en vermell.

1. **INCIDÈNCIA:** Retrassos d'un parell de setmanes en l'entrega dels cables de fibra amb connectors SC que connecten el Tauler de connexions de fibra al Xenpak del cisco6509.

   RESOLUCIÓ: S'ha insistit al tècnic de SEDER per tal que vingues al PIC a realitzar l'entrega dels connectors i l'entrega formal de les connexions certificades.

2. **INCIDÈNCIA:** El dia 1/06/07 (divendres) els tècnics d'Al-Pi venen al PIC per entregar el segment PIC-Anella Científica del circuit dedicat, però els seus equips Nortel Optera Metro 5200 tenen el led CRITICAL encès. Els tècnics informen de que la configuració a

l'extrem del PIC pel que respecta a l'enllaç físic (capa 1) és correcta.

RESSOLUCIÓ: Segons els tècnics d'Al-Pi el motiu de la falta de connectivitat del segment es deu a una configuració incorrecta dels seus equips ubicats a l'Anella Científica. El dia 5/06/07 (primer dia laborable posterior al 1/06/07) Al-Pi entrega el segment, obtenint així connectivitat amb l'Anella Científica. Cal recordar que en aquest punt del desplegament a l'Anella Científica ja s'ha provat el segment Anella Científica – CERN, només resta pendent la certificació del segment PIC-Anella Científica i que a l'Anella Científica es realitzi la connexió, a capa 2, entre ambdós segments.

3. **INCIDÈNCIA:** Durant les proves d'ample de banda del segment PIC-Anella Científica es satura l'encaminador cisco6509 del PIC. Les proves són realitzades pels administradors dels encaminadors mitjançant rutes estàtiques i l'addició de paquets de control amb un TTL alt que va rebotant entre els dos encaminadors gràcies a les rutes estàtiques configurades.

   RESOLUCIÓ: La prova provoca una saturació de la tarja supervisora del cisco6509 del PIC, causant l'anulació de la prova d'ample de banda del segment PIC-Anella Científica. Es decideix fer la prova d'ample de banda PIC-CERN i seguir de prop el trànsit pel segment PIC-Anella Científica per detectar possibles pèrdues de dades. Amb posterioritat s'ha detectat que el problema que causava la saturació de la tarja supervisora del cisco6509 del PIC era que la ruta estàtica no estava definida correctament i la supervisora es col·lapasava al intentar generar la resposta al trànsit generat des de l'Anella Científica (al generar els paquets ICMP de "*host unreachable*").

4. **INCIDÈNCIA:** durant les proves de certificació es detecta un tall de connectivitat en la connexió punt a punt amb el CERN a les 21h del 6/06/07.

   RESOLUCIÓ: Es reporta la incidència al NOC del CERN a les 23:55 del 6/06/07 i a les 9:36 del 7/06/07 David Gutiérrez, del NOC del CERN, informa que GÉANT ha obert una incidència en referència al problema de connectivitat.

   En la incidència de GÉANT (ticket 1042462) s'explica que Interoute (proveïdor de la fibra fosca de GÉANT) està experimentant problemes entre la regió de Meysse (França) i Ginebra, éssent difícil determinar el punt exacte degut a la freqüent falta de supervisió (segons GÉANT) entre els equips de la lambda Ginebra-Madrid. Això causa la falta de connectivitat en el segment RedIRIS(Madrid)-CERN i, en conseqüència, entre el PIC i el CERN sobre el circuit dedicat de 10 Gbps. Al llarg del matí del 7/06/07 GÉANT reestableix la connectivitat.

5. INCIDÈNCIA: al iniciar les proves de certificació amb transferències bidireccionals d'alta velocitat es perd la connectivitat PIC-CERN sobre el circuit dedicat de 10 Gbps.

   RESOLUCIÓ: La falta de connectivitat és causada per un error en una interfície del PoP de RedIRIS10 a Barcelona (un Nortel ERS 8010). Al reiniciar la interfície corresponent al circuit dedicat de 10 Gbps PIC-CERN la connexió es reestableix. La diagnosis i possible solució d'aquest problema correspon a RedIRIS, de totes formes se'n pot trobar una anàlisi

a l'annex J. Cal remarcar que la incidència només afecta quan s'intenten realitzar les proves de certificació amb transferències bidireccionals d'alta velocitat degut a les particularitats d'aquestes (explicat en l'annex J).

A data 14/06/07 RedIRIS ha corregit el problema en el switch Nortel de Barcelona i el trànsit pot fluir bidireccionalment pel bucle entre l'encaminador del PIC i del CERN correctament.

6. **INCIDÈNCIA:** Durant la configuració de l'agregació de línies en l'encaminador Cisco 6509 es detecten alguns talls de connectivitat a la xarxa del PIC. Els talls de connectivitat tenen el seu origen en la creació d'un bucle a la xarxa, causat per la configuració de l'agregació de línies en el Cisco 6509.

   RESOLUCIÓ: La gestió i ressolució de la incidència es realitza mitjançant el ticket 3022. Finalment es detecta que en el Cisco 6509 s'estava configurant un EtherChannel mitjançant PAGP enlloc de l'agregació de línies estàndard IEEE 802.3ad (amb LACP) que s'indica en la metodologia per a la integració dels serveis del PIC a la xarxa LHC-OPN (veure punt 2 de la metodologia a la subsecció 4.2.4).

4.3.4 Rutes PIC-CERN

A continuació es mostra el detall de la ruta que segueixen els paquets del PIC al CERN via el circuit dedicat de 10 Gbps, sobre la connexió punt a punt i amb IPs del CERN. En el *traceroute* de la figura 4.3.5 es mostra la ruta via el circuit dedicat de 10 Gbps des del PIC, i en el *traceroute* de la figura 4.3.6 es mostra la ruta des del CERN.

La ruta PIC-CERN és simètrica i amb una latència (RTT) d'uns 32 ms, en ambdós sentits.

```
[root@lhcopn03 root]# traceroute hufsa.cern.ch
traceroute to hufsa.cern.ch (128.142.208.6), 30 hops max, 38 byte packets
 1  192.16.166.241 (192.16.166.241)  1.070 ms  0.801 ms  0.740 ms
 2  l513-c-rftec-1-be12.cern.ch (192.16.166.57)  36.841 ms  35.543 ms  36.095 ms
 3  * * *
 4  hufsa.cern.ch (128.142.208.6)  31.826 ms  31.795 ms  31.847 ms
```

Figura 4.3.5: ruta PIC-CERN via el circuit dedicat de 10 Gbps amb IPs del CERN. Un cop surt del PIC la connexió va directament a l'encaminador del CERN; 192.16.166.57 és l'extrem del CERN de la connexió punt a punt. Això és gràcies a que des del PIC fins al CERN es disposa d'un enllaç de capa 2 (ethernet).

```
[pictest@hufsa ]# traceroute 192.16.166.244 #(lhcopn03, interfície Lhc-OPN IP del CERN)
traceroute to 192.16.166.244 (192.16.166.244), 30 hops max, 40 byte packets
 1  l513-c-rftec-2-de (128.142.208.1)  7.529 ms  0.288 ms  0.260 ms
 2  l513-c-rftec-1-tl1 (194.12.139.1)  2.150 ms  0.411 ms  0.372 ms
 3  * * *
 4  192.16.166.244 (192.16.166.244)  33.075 ms  31.850 ms  31.842 ms
```

Figura 4.3.6: ruta CERN-PIC via el circuit dedicat de 10 Gbps amb IPs del CERN. La ruta és simètrica tot i que degut a filtres i a que l'enllaç és una connexió punt a punt els salts de la ruta vists per *traceroute* són diferents en cada sentit.

Tal i com es mostra en la figura 4.3.7 la connectivitat de la xarxa sobre el circuit dedicat de 10 Gbps és correcte amb MTU estàndard Ethernet (1500bytes) i amb JumboFrames (MTU=9000).

```
[root@lhcopn02 root]# tracepath hufsa.cern.ch #MTU=1500
 1:  192.16.166.243 (192.16.166.243)                        0.147ms pmtu 1500
 1:  192.16.166.241 (192.16.166.241)                        2.378ms
 2:  l513-c-rftec-1-be12.cern.ch (192.16.166.57)           32.250ms
 3:  no reply
```

```
 4:  hufsa.cern.ch (128.142.208.6)                          31.891ms reached
     Resume: pmtu 1500 hops 4 back 4

hufsa# tracepath 192.16.166.242 #MTU=9000
 1:  hufsa.cern.ch (128.142.208.6)                           0.174ms pmtu 9000
 1:  l513-c-rftec-2-de.cern.ch (128.142.208.1)               5.165ms
 2:  l513-c-rftec-1-tl1.cern.ch (194.12.139.1)               3.691ms
 3:  no reply
 4:  192.16.166.242 (192.16.166.242)                        43.406ms reached
     Resume: pmtu 9000 hops 4 back 4
```

Figura 4.3.7: *tracepath* des de 192.16.166.243 (lhcopn02.pic.es) cap a hufsa.cern.ch amb MTU=1500 i tracepath des de hufsa.cern.ch a192.16.166.242 (lhcopn01) amb MTU=9000. El funcionament és correcte en ambos casos.

Cal notar que el fet de deixar d'utilitzar adreces IP del CERN i utilitzar les pròpies no modificarà substancialment les característiques de la ruta com ara l'rtt o el número de salts.

4.3.5 Proves de la connexió sobre el circuit dedicat de 10 Gbps

En aquesta subsecció es detallen les diferents proves de latència, rendiment i fiabilitat realitzades amb la connexió PIC-CERN sobre el circuit dedicat de 10 Gbps.

Per a la realització de les proves s'ha disposat d'un servidor de test al CERN (*hufsa.cern.ch*) amb una connexió directa a 10 Gbps. Des del PIC les proves s'han realitzat des de 12 host dedicats amb connexió directa al Cisco6509, veure la subsecció 4.2.2 per als detalls del sistema de certificació del circuit dedicat de 10 Gbps i l'annex H, secció 3, per consultar la configuració dels diferents servidors i encaminadors.

Cal tenir en compte que totes les proves mostrades a continuació s'han dut a terme abans de que la connexió fos certificada i entrés en producció, utilitzant IPs del CERN. També és important saber que el servidor del CERN (*hufsa.cern.ch*) tot i disposar d'una NIC de 10 Gbps, per limitacions arquitecturals del maquinari del servidor, no és capaç de gestionar més de 5 Gbps.

○ Detall de les proves de RTT (Round Trip Time)

En les figures 4.3.8 i 4.3.9 es mostren les gràfiques generades a partir de diferents proves amb la utilitat *ping,* mantenint la línia sense trànsit. El paràmetre variable és la mida del paquet ICMP de *echo*. Les proves s'han realitzat amb el bit de no fragmentació (DF) activat per evitar que cap servidor fragmenti els paquets de *echo* ICMP.

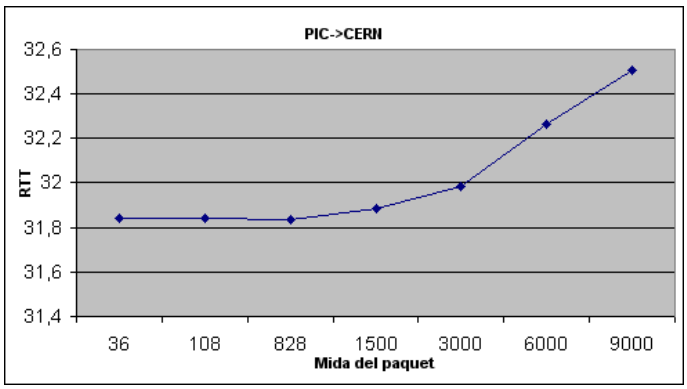


Figura 4.3.8: RTT (en ms) de la connexió PIC->CERN en funció de la mida del paquet (en bytes). Com es pot observar les variacions són absolutament menyspreables (<1ms)

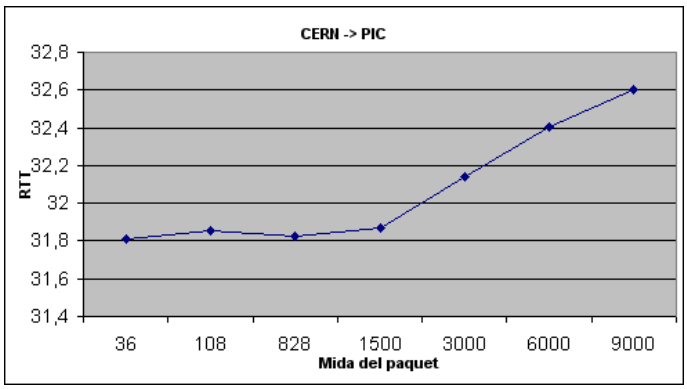


Figura 4.3.9: RTT (en ms) de la connexió CERN->PIC en funció de la mida del paquet (en bytes). En aquesta direcció les variacions també són menyspreables (<1ms)

Es pot afirmar que la latència (RTT) de la connexió és d'uns 32-33 ms entre el PIC i el CERN sobre el circuit dedicat de 10 Gbps, una latència que podem considerar com a bona. El petit increment de l'RTT que es veu al augmentar la mida del paquet és absolutament normal, ja que hi ha més bytes per transferir en cada paquet enviat.

○ Detall de les proves d'ample de banda (throughput)

A continuació es mostraran els detalls d'algunes de les proves de rendiment TCP i UDP realitzades. Totes les proves s'han realitzat amb JumboFrames (MTU=9000), tal i com ha de funcionar la xarxa LHC-OPN quan entri en producció. Cal recordar que, tal i com es demostra a l'annex C, l'ús de JumboFrames no aporta cap problema a la xarxa sinó que en millora lleugerament el rendiment.

En forma de resum es pot dir que per a connexions UDP s'han obtingut velocitats de transferència sostinguda de 9,2 Gbps. Sobre TCP les proves s'han vist limitades pel servidor del CERN (*hufsa.cern.ch*), incapaç de superar els 5 Gbps.

Proves UDP

Pera a la realització de les proves UDP s'han utilitzat les versions 1.7.0 i 2.2.0 d'iperf, obtenint resultats molt similars amb ambdues versions. Degut a mecanismes de control de flux (*flow control*) aplicats a la NIC del servidor del CERN (*hufsa*) el trànsit UDP entrant al servidor, i l'acceptat per la interfície del switch on està connectat, és limitat (a l'annex J, secció 15.1 es pot observar el comportament del mecanisme del control de flux).

Proves PIC->CERN

En primer lloc, per tal de provar la connectivitat del circuit dedicat, s'ha dut a terme una prova mitjançant la generació d'un flux de dades UDP de “soroll”, sense que al CERN ningú el rebés. Els resultats corresponen a la gràfica de la figura 4.3.10.

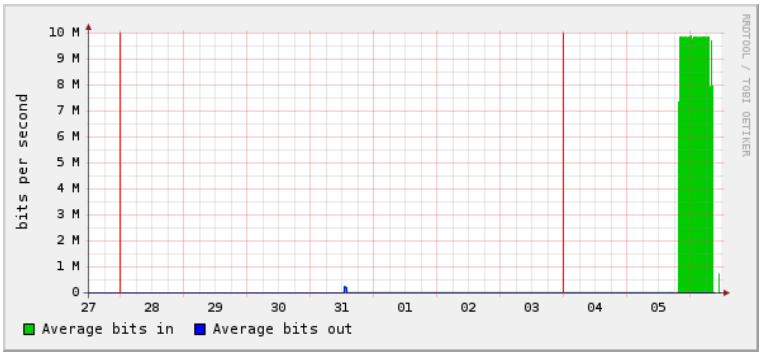


Figura 4.3.10: gràfica generada pel CERN de la interfície del circuit dedicat PIC-CERN a 10 Gbps, corresponent a la prova de connectivitat de 100.000 segons de duració.

Un cop demostrada la connectivitat sobre el circuit dedicat es preparen els servidors ieprf per a iniciar les per a les proves de rendiment de la connexió (realitzades entre les 15 i les 16 hores). A la gràfica de la figura 4.3.11 es pot observar el trànsit generat pel sistema de certificació que surt de l'encaminador del PIC en direcció al CERN, a la figura 4.3.12 el trànsit rebut a l'Anella Científica i, finalment, a la figura 4.3.13 el trànsit rebut a l'encaminador del CERN.

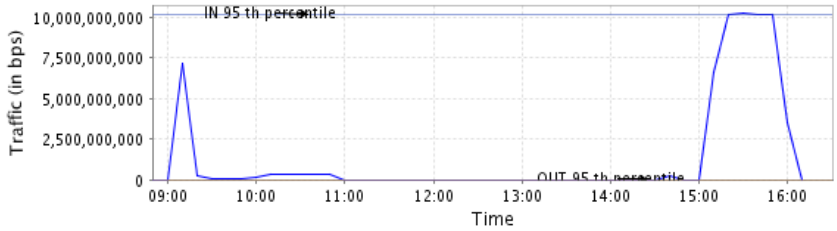


Figura 4.3.11: Monitorització del cisco6509 del PIC del trànsit PIC->CERN generat pel sistema de certificació. La prova correspon al trànsit generat de 15:00 a 16:00.

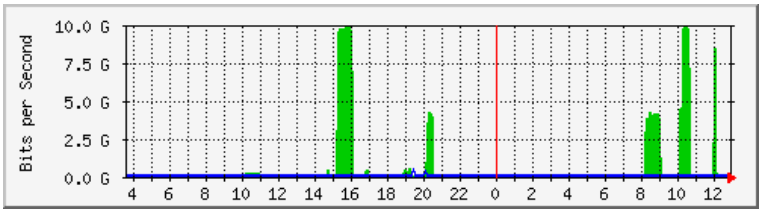


Figura 4.3.12: Monitorització de l'encaminador de l'Anella del trànsit PIC->CERN. La prova correspon al trànsit generat de 15:00 a 16:00, el trànsit posterior és d'altres proves posteriors (TCP i/o UDP).

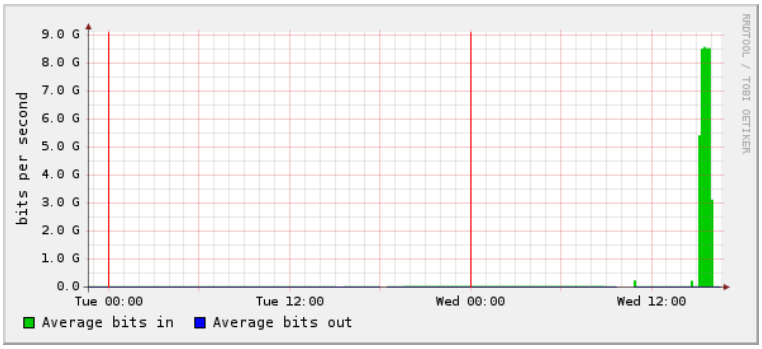


Figura 4.3.13: Monitorització de la interfície del circuit dedicat PIC-CERN en l'encaminador del CERN.

Cal dir que les gràfiques de les figures anteriors es dibuixen utilitzant promitjos, raó per la qual algunes transferències semblen tenir arrancades i parades graduals. Així mateix

l'exactitud no és del 100%, però si orientativa. La conclusió extreta és que al PIC es generen poc més de 10 Gbps, dels quals es pot afirmar que un 98-99% arriba a l'Anella Científica. Al CERN arriben aproximadament 9,2 Gbps, tal i com es mostra en les estadístiques extretes de la interfície de la figura 4.3.14. Així doncs, segons les dades aportades anteriorment, es pot dir que en condicions de saturació (10 Gbps) hi ha una pèrdua del ~7% de paquets, assolint una ocupació del circuit dedicat del 93,49% sobre el límit teòric (10 Gbps).

```
L513-C-RFTEC-1#sh inter te3/2
TenGigabitEthernet 3/2 is up, line protocol is up
Description: ----> PIC Spain: primary <#S513-C-BE12-TenGigabitEthernet 3/2
Hardware is Force10Eth, address is 00:01:e8:18:f0:d4
    Current address is 00:01:e8:18:f0:d4
Pluggable media present, XFP type is 10GBASE-LR.
    Medium is MultiRate, Wavelength is 1310.00nm
    XFP receive power reading is -4.5395
Interface index is 135069755
Internet address is not set
MTU 9216 bytes, IP MTU 9000 bytes
LineSpeed 10000 Mbit
ARP type: ARPA, ARP Timeout 04:00:00
Last clearing of "show interface" counters 00:00:54
Queueing strategy: fifo
Input Statistics:
     7213534 packets, 63137570255 bytes
     0 Vlans
     0 64-byte pkts, 210781 over 64-byte pkts, 0 over 127-byte pkts
     0 over 255-byte pkts, 0 over 511-byte pkts, 7002754 over 1023-byte pkts
     1 Multicasts, 0 Broadcasts
     0 runts, 0 giants, 0 throttles
     0 CRC, 0 overrun, 0 discarded
Output Statistics:
     1 packets, 102 bytes, 0 underruns
     0 64-byte pkts, 1 over 64-byte pkts, 0 over 127-byte pkts
     0 over 255-byte pkts, 0 over 511-byte pkts, 0 over 1023-byte pkts
     0 Multicasts, 0 Broadcasts, 1 Unicasts
     0 Vlans, 0 throttles, 0 discarded, 0 collisions
Rate info (interval 30 seconds):
     Input 9328.18 Mbits/sec,     133224 packets/sec, 93.49% of line-rate
     Output 00.00 Mbits/sec,          0 packets/sec, 0.00% of line-rate
Time since last interface status change: 1d13h28m
```

Figura 4.3.14: estadístiques de l'encaminador Force10 del CERN corresponent a la interfície del circuit dedicat a 10 Gbps PIC-CERN.

Degut a la falta d'accés a la monitorització en punts intermitjos del circuit dedicat (per exemple a RedIRIS) es fa difícil esbrinar els punts de la xarxa on es perden paquets. De totes formes queda palès que la part d'instal·lació responsabilitat del PIC té un funcionament correcte, amb una connexió certificada a 10 Gbps (des del PIC fins a l'enllaç amb l'Anella Científica).

Proves CERN->PIC

Degut a les incidències documentades en la subsecció 4.3.3 no s'han pogut realitzar totes les proves planificades, això afecta a les proves en direcció CERN->PIC. Tan bon punt les incidències siguin resoltes serà possible realitzar les proves de transferències bidireccional que han de servir per a demostrar i certificar la capacitat de transferir dades bidireccionalment entre el PIC i CERN a una velocitat de ~10 Gbps i, en conseqüència, provar la capacitat de la connexió CERN->PIC.

Les proves realitzades des del servidor del CERN *hufsa.cern.ch* revelen un comportament normal de la connexió, essent el màxim aconseguit de ~3Gbps, degut a les limitacions del

servidor *hufsa*, el qual es saturava al generar tal quantitat de dades (mitjançant iperf).

Un cop comprovat el rendiment de la connexió en direcció PIC->CERN, i esperant un comportament idèntic o molt similar en el sentit contrari, es pot concloure que el funcionament del circuit dedicat és correcte. De totes formes no és possible donar el circuit dedicat de 10 Gbps com a certificat fins a la finalització de totes les proves planificades.

Proves TCP

Pera a la realització de les proves s'han utilitzat les versions 1.7.0 i 2.0.2, donant resultats idèntics. No s'ha pogut utilitzar *Thrulay* degut a la no disponibilitat d'aquest en el servidor hufsa del CERN.

En aquesta ocasió la generació de trànsit bidireccional no té sentit degut a que, com ja s'ha dit, el servidor hufsa.cern.ch no és capaç de gestionar més de 5 Gbps i, en conseqüència, no es pot realitzar una proba bidireccional de més de 2,5Gbps per direcció (2,5PIC->CERN + 2,5 CERN->PIC). De fet per aquesta mateixa limitació no es poden esperar transferències per sobre dels 5 Gbps en aquestes proves i, en situacions prop d'aquest límit, les variacions (de fins a ~1 Gbps) en els resultats de repeticions d'una prova són habituals (per exemple, la mateixa prova repetida en condicions idèntiques 5 vegades oscil·la entre els 4,2 i els 4,9 Gbps).

Proves PIC->CERN

Amb els paràmetres estàndard de SLC3 (finestra de 85.3Kbyte sense dimensionament automàtic[22]) la velocitat màxima (i sostinguda) de transferència és de 28,6-29,4 Mbps. Al incrementar el número de transferències simultànies (opció -P de iperf) la velocitat per transferència per flux de dades TCP es comença a degradar:

- 5 transferències simultànies -> 28-29 Mbps/transferència (total de ~145Mbps)
- 10 transferències simultànies -> 21-27 Mbps/transferència (total de ~270Mbps)
- 15 transferències simultànies -> 20-26Mbps/transferència (total de ~370Mbps)

A partir de 20 transferències simultànies ja no es percep cap millora en la velocitat de transferència total per servidor. Així doncs, per tal de poder aprofitar l'ample de banda ofert pel circuit dedicat a 10 Gbps amb transferències TCP és necessari optimitzar els paràmetres TCP dels servidors. En l'annex H, secció tercera, es detallen les modificacions de paràmetres realitzades.

Un cop amb els paràmetres optimitzats s'han realitzat diverses proves modificant la mida de la finestra, en la figura 4.3.15 es poden observar els resultats obtinguts pel sistema de certificació segons la mida de finestra TCP utilitzada i en la figura 4.3.16 la velocitat màxima per flux de dades. En tot moment cal tenir en compte que aquestes proves es troben limitades per la potència del servidor que rep les dades al CERN que, tal i com ens han avisat des del CERN, no pot superar els 5 Gbps.

22 Com a dimensionament automàtic em refereixo a *autotuning* de la finestra TCP. Amb kernel 2.6 (com ara SLC4) aquesta opció ve activada per defecte, fent que la mida de la finestra es dimensioni automàticament des d'uns 8Kbytes fins a 4Mbytes (paràmetres per defecte).

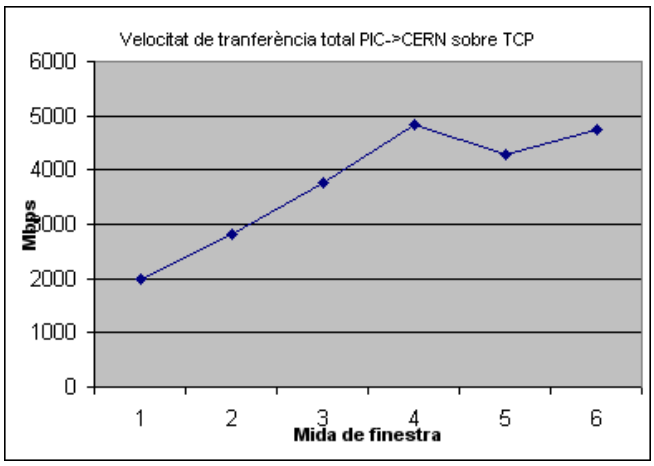


Figura 4.3.15: proves unidireccionals TCP PIC->CERN, en Mbps (8*Mbytes/segon). Resultats obtinguts pel sistema de certificació, en funció la mida de finestra TCP utilitzada, a partir de 4MBytes de finestra els resultats ja no milloren.

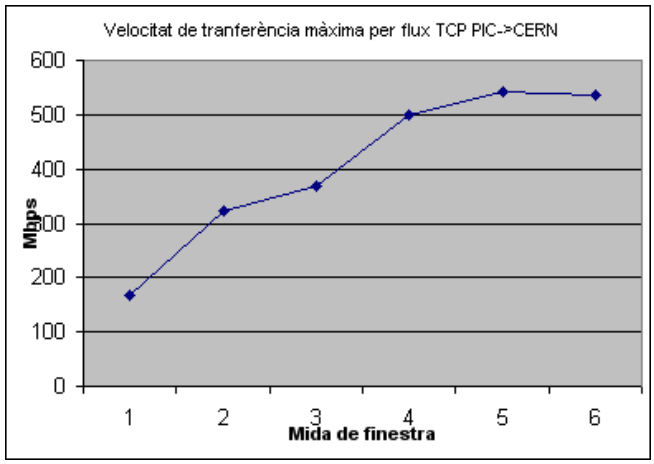


Figura 4.3.16: velocitat màxima d'un únic flux de dades en les proves unidireccionals TCP PIC->CERN, en Mbps (8*Mbytes/segon). Resultats obtinguts amb el sistema de certificació, en funció de la mida de finestra TCP utilitzada. Cal mencionar que amb una finestra de 4Mbytes i dos flux de dades des d'un servidor s'han aconseguit velocitats de fins a 769Mbps.

Tot i les limitacions en les proves, i tenint en compte que com més petita sigui la mida de la finestra millor per a la gestió de la memòria dels servidors, la finestra òptima recomanada per a obtenir velocitats de transferència elevades amb un únic flux de dades TCP és 4 Mbytes.

Proves CERN->PIC

En les proves realitzades des del servidor del CERN (*hufsa.cern.ch*) s'ha aconseguit generar flux de dades d'uns 2700 Mbps, essent l'element limitant el propi servidor *hufsa*.

Degut a la manca de recursos per a realitzar les proves en el CERN no s'han pogut realitzar proves sobre TCP en en sentit CERN->PIC amb un ample de banda significatiu. El mateix escenari s'ha repetit al intentar realitzar transferències bidireccionals, obtenint uns 2800Mbps sumant el trànsit TCP d'ambdues direccions (~2Gbps PIC->CERN i ~800 Mbps CERN->PIC).

4.4 Certificació de la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps

En aquesta secció, que tanca el quart capítol i el desenvolupament del projecte en si mateix, es documenta la certificació de la connexió sobre el circuit dedicat de 10 Gbps.

Part del procés de certificació de la connexió s'ha realitzat al llarg del desplegament del pla, en la secció anterior.

A continuació es pot trobar un petit resum de les proves empíriques de rendiment i fiabilitat de la connexió i, a continuació, es realitza un petit anàlisi de la càrrega de l'encaminador local. Tal i com s'indica en el primer capítol, el procés de certificació també inclou una anàlisi del trànsit de la nova xarxa i estadístiques del rendiment de la connexió a mig termini.

4.4.1 Proves de rendiment i fiabilitat de la connexió

La capacitat de la connexió en direcció PIC->CERN queda certificada al demostrar que per a connexions UDP és possible obtenir transferències de fins a 9,2 Gbps, el coll d'ampolla de la connexió es troba entre l'Anella Científica i el CERN.

En sentit PIC->CERN la connexió s'ha pogut certificar a una velocitat de 9,2 Gbps i en sentit CERN->PIC amb una velocitat de fins a 6 Gbps. El coll d'ampolla s'ha localitzat entre l'Anella Científica i l'equip Nortel de RedIRIS a Barcelona (veure figures 4.4.1, 4.4.2, 4.4.3, 4.4.4).

```
cisco-6500#sh interfaces tenGigabitEthernet 8/1
TenGigabitEthernet8/1 is up, line protocol is up (connected)
  Hardware is C6k 10000Mb 802.3, address is 0018.7383.5a9c (bia 0018.7383.5a9c)
  Description: 10 Gbps LHC-OPN
  MTU 9216 bytes, BW 10000000 Kbit, DLY 10 usec,
     reliability 255/255, txload 253/255, rxload 154/255
  Encapsulation ARPA, loopback not set
  Keepalive set (10 sec)
  Full-duplex, 10Gb/s
  input flow-control is off, output flow-control is off
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input 1w1d, output 1w1d, output hang never
  Last clearing of "show interface" counters 10w1d
  Input queue: 0/4096/17382801/0 (size/max/drops/flushes); Total output drops: 1596917838
// es per la saturació creada pel sistema de certificació que genera ~11 Gbps
  Queueing strategy: fifo
  Output queue: 0/300 (size/max)
  5 minute input rate 6057180000 bits/sec, 219413 packets/sec  //6 Gbps
  5 minute output rate 9935709000 bits/sec, 401907 packets/sec //9,93 Gbps
     3323149259 packets input, 10018831741211 bytes, 0 no buffer
     Received 95931339 broadcasts (95921233 multicasts)
     0 runts, 0 giants, 0 throttles
     4 input errors, 0 CRC, 0 frame, 17382797 overrun, 0 ignored
     0 watchdog, 0 multicast, 0 pause input
     0 input packets with dribble condition detected
     12068930589 packets output, 31428476700902 bytes, 0 underruns
     0 output errors, 0 collisions, 5 interface resets
     0 babbles, 0 late collision, 0 deferred
     0 lost carrier, 0 no carrier, 0 PAUSE output
     0 output buffer failures, 0 output buffers swapped out
```

Figura 4.4.1: estadístiques de la interfície de 10 Gbps del circuit dedicat PIC-CERN a l'encaminador del PIC el dia 14/06/07 al matí (9:30). Es pot observar com entre el PIC i el CERN s'envien 9,9 Gbps i se'n reben 6 Gbps.

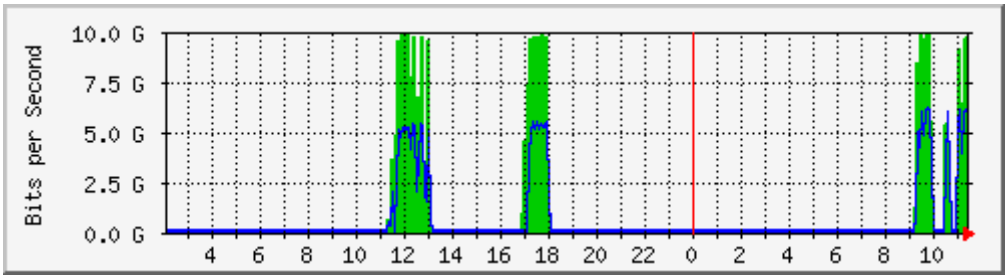


Figura 4.4.2: estadístiques de la interfície de 10 Gbps del circuit dedicat PIC-CERN al switch de l'Anella Científica (el primer punt del camí després del PIC). Es pot observar com del PIC al CERN s'envien quasibé 10 Gbps i se'n reben 6 Gbps. No hi ha diferències notables amb el que es percep al PIC a les 9:30 del 14/06/07.

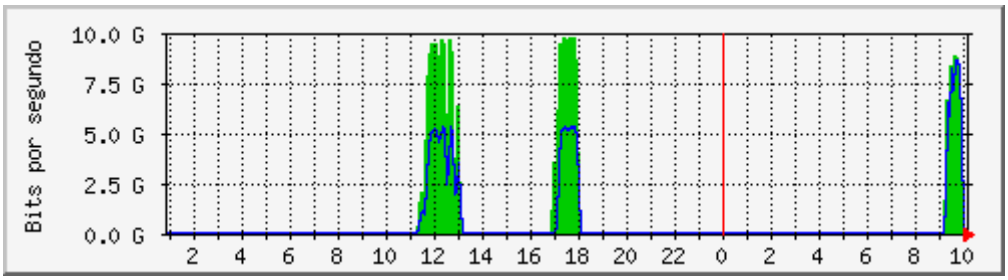


Figura 4.4.3: estadístiques de la interfície de 10 Gbps del circuit dedicat PIC-CERN al switch de RedIRIS, la gràfica és dues hores anterior a la de la figura 4.4.2. Es pot observar com entre el PIC i el CERN s'envien i es reben uns 8 Gbps. En la connexió entre l'Anella Científica i RedIRIS hi ha un coll d'ampolla a 8 Gbps per a l'enviament cap al CERN i de 6 Gbps cap al PIC.

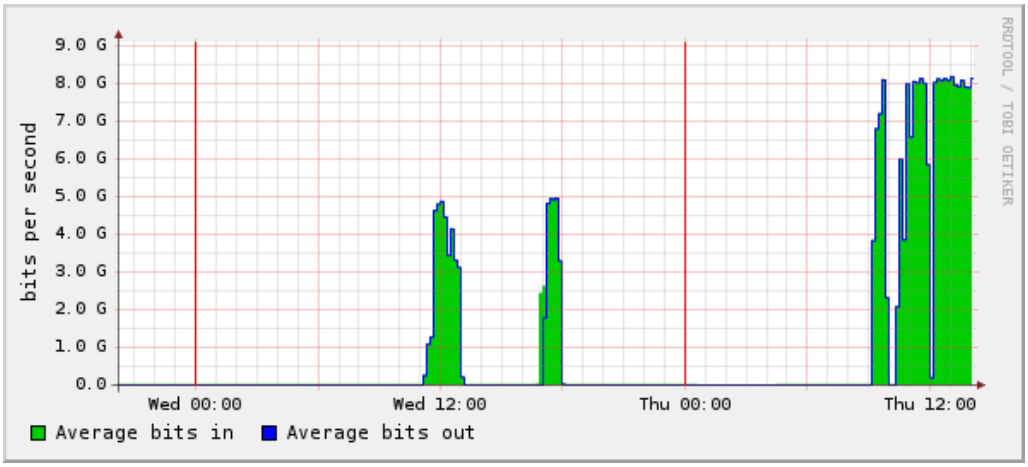


Figura 4.4.4: estadístiques de la interfície de 10 Gbps del circuit dedicat PIC-CERN a l'encaminador del CERN. Es pot observar com entre el PIC i el CERN s'envien i es reben uns 8 Gbps, igual que en la gràfica de la figura 4.4.3.

Respecte a les transferències sobre TCP cal dir que la mida òptima de la finestra TCP (windowsize) ha estat calculada, per tal d'evitar colls d'ampolla causats per la finestra de TCP, de forma empírica i segons la fórmula *WindowSize = bandwidth*rtt*. Així doncs, donat que l'RTT de la connexió és d'uns 32 ms, la mida de finestra òptima, tenint en compte que els servidors del PIC disposen de NIC de 1 Gbps, és 1000mbps*32ms = 10000e6bps*0,032s = 32000000 bits ~= **3,8 Mbyte**. Durant les proves realitzades en la secció anterior s'ha demostrat que l'ample de banda màxim per una única connexió es pot assolir amb una finestra de 4MBytes o més.

En la subsecció 4.3.5 es pot trobar una explicació detallada de les proves de rendiment realitzades.

4.4.2 Anàlisi de la càrrega en l'encaminador local

S'ha analitzat detalladament la càrrega que suposa al Cisco6509 del PIC la nova connexió sobre el circuit dedicat de 10 Gbps. La conclusió és que la tarja supervisora no pateix cap càrrega addicional i manté el seu percentatge d'ús entre el 5 i el 10%, gràcies a les acceleradores instal·lades en els diferents components del xassís de l'encaminador.

4.4.3 Anàlisi del trànsit de la nova xarxa

No s'ha detectat cap paquet que no s'hagi generat voluntàriament pel PIC o el CERN.

Donat que es tracta d'una connexió de capa 3 punt a punt entre l'encaminador del CERN i l'encaminador del PIC és natural que no es detecti a la xarxa trànsit que no sigui generat de forma expresa.

4.4.4 Anàlisi estadístic del rendiment de la connexió a mig termini

A causa de les diverses incidències ocorregudes al llarg del desplegament de la connexió sobre el circuit dedicat de 10 Gbps PIC-CERN, no s'ha disposat del temps necessari per a incloure en aquesta memòria una anàlisi del rendiment de la connexió a mig termini.

4.5 Proposta per al desplegament d'una connexió redundant

En aquest capítol es defineix una proposta per al desplegament d'una connexió redundant al circuit dedicat de 10 Gbps.

En les següents seccions es descriu el punt de partida i es detallen els objectius de la proposta, així com les especificacions del sistema a implementar i la seva aplicació. En les tres últimes seccions es realitza una anàlisi de la viabilitat tècnica, operativa i econòmica de la solució plantejada.

4.5.1 Descripció de la situació inicial

Un cop finalitzat el desplegament del circuit dedicat PIC-CERN de 10 Gbps, al PIC es disposa d'una única connexió no redundant a la xarxa LHC-OPN amb un ample de banda de 10 Gbps i una connexió de 1 Gbps a la xarxa general (la Internet).

4.5.2 Objectius

Es pretén analitzar una solució que permeti al PIC disposar d'un camí redundant per a la connexió a la xarxa LHC-OPN per tal que, en cas de fallida del circuit dedicat de 10 Gbps existent, el PIC mai no es quedi desconnectat de la xarxa LHC-OPN.

El plantejament inicial del CERN és que la connexió redundant del PIC es realitzi via el centre Tier-1 IN2P3, a Lyon (França).

4.5.3 Especificacions del sistema i de les aplicacions

Cal que tots els segments de la connexió redundant segueixin un camí físic diferent del seguit pel circuit primari i es recomana que el circuit redundant sigui un circuit dedicat i disposi d'un ample de banda similar al del circuit primari.

És necessari que en cas de fallida del circuit primari les transferències s'encaminin automàticament per la connexió del circuit redundant. El canvi en l'encaminament ha de ser transparent pels serveis i servidors que utilitzin la xarxa LHC-OPN.

En cas de no establir la relació de veïnatge directament amb el CERN, cal establir-la amb algun Tier-1 i aquest ha de reanunciar el Sistema Autònom del PIC al CERN, mitjançant BGP. Cal saber que aquest és un comportament que el grup de desenvolupament de la xarxa LHC-OPN ja ha previst per a casos de fallida de la connexió primària d'algun centre Tier-1.

4.5.4 Viabilitat tècnica

Tècnicament cal tenir en compte que en el plantejament inicial de la xarxa LHC-OPN s'hi inclouen Sistemes Autònoms encaminats mitjançant BGP per tal de poder gestionar connexions redundants. Així doncs disposar d'una segona connexió a la xarxa LHC-OPN significa afegir una nova relació

de veïnatge BGP amb el CERN i/o un centre Tier-1, convertint l'AS del PIC en un Sistema Autònom *multihomed*, és a dir, amb més d'una connexió a l'exterior.

Per tal de tenir el rol de connexió redundant, la nova relació de veïnatge de la connexió hauria de disposar d'una prioritat inferior respecte la connexió primària.

En el cas d'establir la connexió redundant amb un centre Tier-1 es recomana utilitzar el circuit dedicat com a connexió redundant comú, és a dir, que la connexió pugui ser utilitzada tant pel centre Tier-1 com pel PIC si algun dels dos pateix una fallida de la connexió primària amb el CERN. També cal tenir en compte que una connexió directa amb un centre Tier-1 es pot utilitzar per a realitzar les comunicacions Tier-1<->Tier-1 sense la necessitat d'utilitzar la connexió primària.

4.5.5 Viabilitat operativa

A nivell operatiu és necessari disposar d'una connexió redundant que permeti una alternativa a la connexió primària en situacions de fallida o manteniment.

En aquest cas la política operativa es pot implementar en els encaminadors de la xarxa, mitjançant el protocol BGP, per tal de poder donar una resposta automatitzada en cas de la detecció de problemes.

4.5.6 Viabilitat econòmica

El cost principal de la solució rau en el manteniment del circuit dedicat redundant. Cal tenir en compte que disposar d'un circuit dedicat a 10 Gbps PIC-CERN o PIC-IN2P3 (el Tier-1 més pròxim al PIC) que utilitzi un camí físic independent del camí del circuit primari és tant o més costós que el desplegament i manteniment de la connexió via el circuit dedicat primari.

4.5.7 Alternatives

Una alternativa que proposa l''Anella Científica és que enlloc d'utilitzar un circuit dedicat a 10 Gbps s'utilitzi la connexió de la UAB (de la mateixa forma que es feia amb l'antiga VLAN236).

El problema d'aquesta alternativa és que en cas de fallida de la connexió primària es passaria d'una connexió dedicada de 10 Gbps a compartir una connexió d'1 Gbps, on gran part del camí físic és comú i, per tant, és molt possible que també es vegi afectat per una possible avaria. D'altra banda, en aquesta alternativa els costs són sensiblement inferiors.

5 Capítol 5 – Conclusions

En aquest capítol es detallaran els objectius assolits i les línies de continuïtat del PFC.

Aquest PFC s'ha dividit en dues fases, en la primera s'ha dissenyat i dut a terme la posta en producció d'una connexió a la Internet sobre un circuit dedicat de 1 Gbps. En la segona fase s'ha realitzat el disseny i desplegament de la connexió a la xarxa LHC-OPN sobre un circuit dedicat de 10 Gbps.

Al llarg del desenvolupament del projecte s'han gestionat i resolt diverses incidències tècniques. No obstant això, la major part dels endarreriments patits es deuen a tràmits burocràtics i a la polititzación del món de les xarxes.

Durant la realització del PFC s'han assolit satisfactòriament els objectius principals del projecte, que es poden resumir en:

- Certificació i posada en producció del circuit dedicat de 1 Gbps.

  S'ha dut a terme durant la primera fase del projecte (fins al març del 2007), assolint els objectius de fiabilitat, ample de banda i connectivitat. Durant la primera fase del projecte també s'ha dissenyat i demostrat la validesa d'una metodologia per a la integració dels servidors del PIC dins la xarxa LHC-OPN, detectant i trobant la solució a algunes deficiències en els switch i en la configuració d'alguns servidors utilitzats al PIC.

- Certificació i posada en producció amb IPs del CERN del circuit dedicat de 10 Gbps.

  S'ha dut a terme durant la segona fase del projecte (del març fins al juny del 2007). Degut a l'endarreriment en la obtenció del Sistema Autònom (AS) i les adreces IP del PIC la posada en producció s'ha realitzat amb IPs del CERN. S'ha certificat la connexió amb una velocitat de 9,2 Gbps en sentit PIC->CERN i de 6 Gbps en sentit contrari[1]. Per a la certificació del circuit dedicat s'ha dissenyat i implementat amb èxit un sistema de certificació capaç d'assolir velocitats de 11 Gbps sostinguts.

  Durant la segona fase del PFC també s'ha definit i analitzat una proposta per a la implementació d'una connexió redundant al circuit dedicat de 10 Gbps.

- Disseny de l'arquitectura per al sistema de recepció de dades Tier0-Tier1 sobre el circuit dedicat de 10 Gbps. S'ha dut a terme durant la segona fase del projecte de forma addicional als objectius fixats pel PFC

1 Els 6 Gbps de rendiment en transferències del CERN al PIC es deuen a un coll d'ampolla en el segment RedIRIS - Anella Científica. El circuit ha de ser capaç de transmetre dades bidireccionalment a 10 Gbps per cada sentit, el coll d'ampolla de 6 Gbps és degut a causes alienes a la realització del PFC, que han de ser resoltes per RedIRIS i l'Anella Científica.

D'altre banda, les línies de continuïtat planificades per aquest Projecte Final de Carrera són:

- **Juny - Juliol 2007**
  - Comprensió i resolució del coll d'ampolla de 6 Gbps localitzat en el segment RedIRIS-Anella Científica del circuit dedicat de 10 Gbps per a transferències del CERN al PIC.
- **Agost 2007**
  - Configuració del Sistema Autònom del PIC mitjançant BGP en l'encaminador Cisco 6509, aplicant als servidors del PIC la configuració necessària (recollida en els capítols 3 i 4 d'aquest PFC) per tal que els serveis puguin utilitzar la nova connexió (per als servidors implica canvis de direccionament i encaminament IP).
  - Implementació del sistema de recepció de dades Tier0-Tier1 sobre el circuit dedicat de 10 Gbps, dissenyat en el capítol 4.
- **Setembre 2007**
  - Entrada en producció del sistema de recepció de dades Tier0-Tier1 sobre el circuit dedicat de 10 Gbps.
  - Donat que la càrrega prevista per al circuit dedicat és inferior als 8 Gbps es proposa la creació de dues VLAN sobre el segment PIC-Anella Científica per tal de dividir la línia de 10 Gbps en dues (per exemple 8+2 Gbps) i establir la connexió del PIC a la Internet, incrementant així l'ample de banda disponible entre el PIC i la Internet.
- **4t trimestre 2007**
  - Implementació d'un sensor per a la detecció de modificacions en el route-set RS-LHCOPN de ripe.net
  - Optimització dels paràmetres de xarxa en el nucli del S.O. dels servidors de la xarxa LHC-OPN (mida de finestra, *timeout*, etc.). Així es milloraria el rendiment de les transferències i, en conseqüència, dels servidors.
- **Al llarg del 2008**
  - Migració de totes les adreces IP del PIC a IPs del nou Sistema Autònom. Això implica establir una relació de veïnatge amb l'Anella Científica i/o RedIRIS, però permetria simplificar la gestió interna de l'adreçament i afegiria un grau més de llibertat a les polítiques d'encaminament del PIC.
  - Perfilar a nivell de xarxa les aplicacions que utilitzen el circuit dedicat per tal de restringir al màxim l'accés dels servidors a la xarxa mitjançant ACLs en l'encaminador Cisco6509.
  - Monitorització activa i automatitzada del trànsit de la xarxa per tal de poder detectar fallides de serveis/servidors i/o comportaments anòmals (per exemple intrusions)

6 Annex A: Planificació detallada del projecte

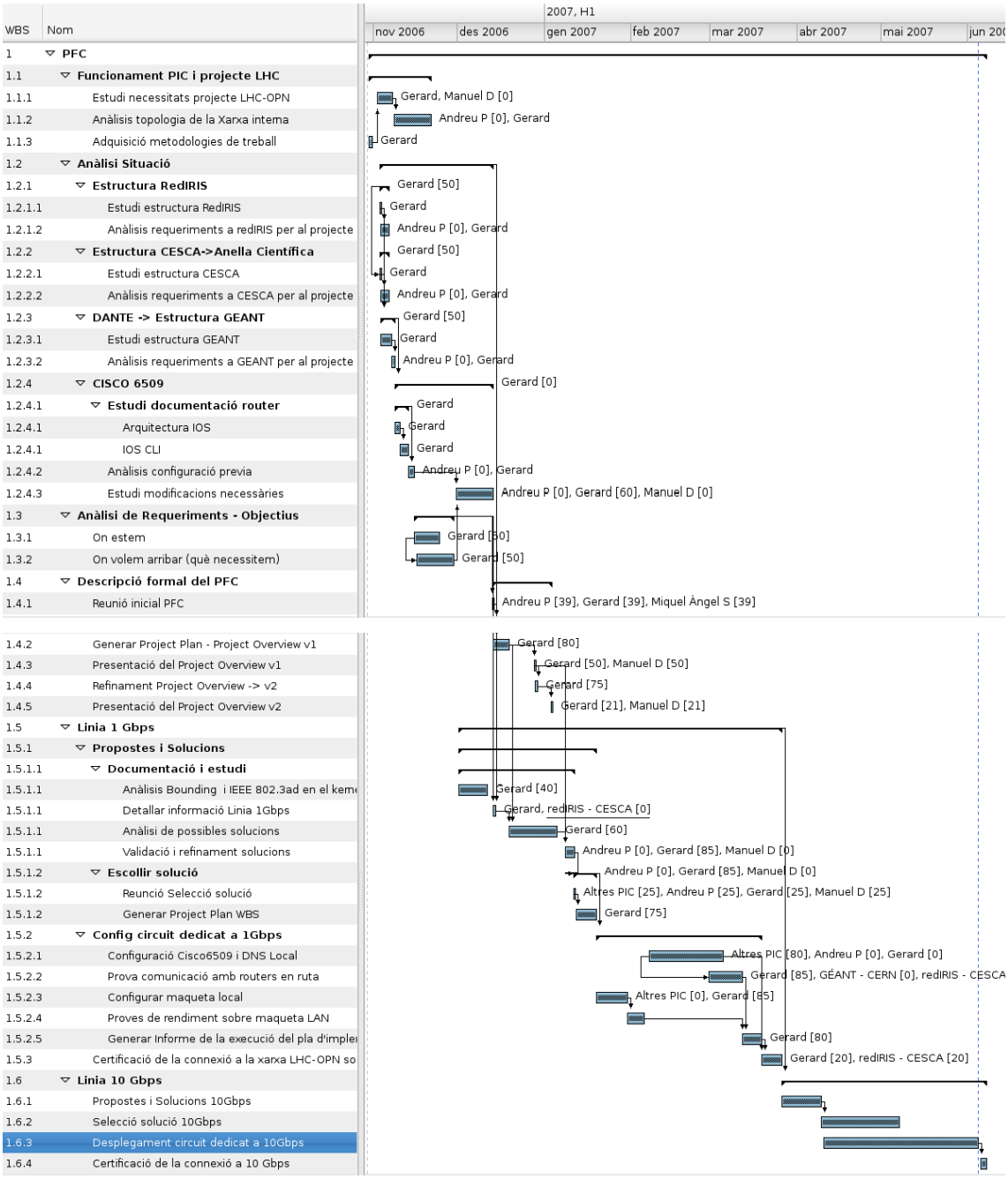


Figura 6.0.1: Planificació completa desglossada del PFC. En el diagrama de Gantt es troba detallat el procés d'adaptació al PIC, l'anàlisi de la situació prèvia, la descripció formal del projecte i la primera fase (desplegament del circuit dedicat de 1 Gbps). El desplegament del circuit dedicat a 10 Gbps, pertanyent a la segona fase del projecte, es troba detallat en el diagrama de Gantt de la figura 6.0.2

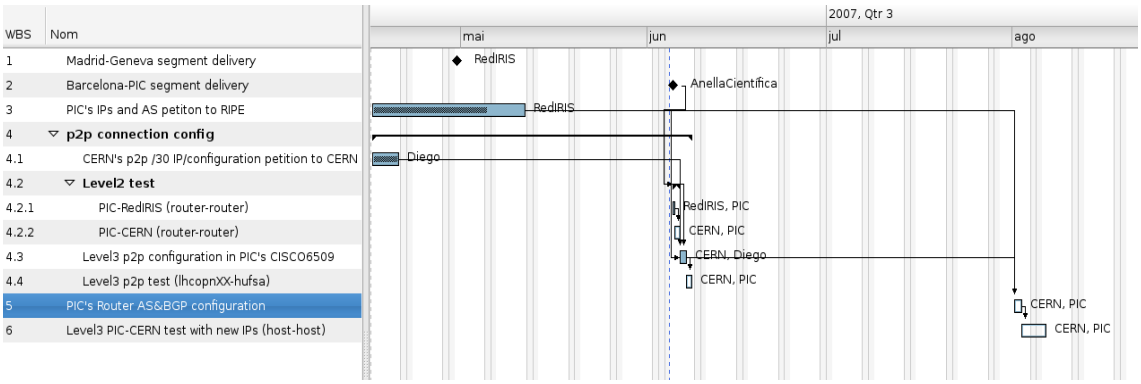


Figura 6.0.1: Planificació del desplegament del circuit dedicat de 10 Gbps (fase 2)

7 Annex B: Script per a l'addició de rutes als servidors LHC-OPN

Aquest script està dissenyat per tal d'afegir/eliminar de forma automàtica les rutes estàtiques necessàries als servidors LHC-OPN, segons la solució de dues IPs de la subsecció 3.1

Per a afegir rutes estàtiques en Linux es pot fer servir la comanda

/sbin/route add -net <XarxaDestí> gw <Gateway>

on <Gateway> serà l'encaminador LHC-OPN i <XarxaDestí> les diferents xarxes de RS-LHCOPN, que es poden consultar (actualitzades) mitjançant:

whois -h whois.ripe.net RS-LHCOPN | grep members | awk '{print $2 }'

L'script per a afegir/eliminar automàticament totes les rutes del *route-set* de RIPE RS-LHCOPN via un gateway és:

```
#!/bin/csh
##########################################
# Versió 1
# Autor: Gerard B.A.
# Agafa les adreces d'una query de RIPE i afegeix rutes estàtiques via el gateway definit
#Si es posa "del" com a primer paràmetre enlloc d'afegir les rutes les elimina
##########################################

#--> Xarxes a excloure de la LHC-OPN (que estiguin dins de RS-LHCOPN) <--
set xarxesExcepcio=(193.145.217.0/24 193.146.196.0/22)

#-----------------------------------------------#
set Gateway=gw.pic.es           # gateway LHC-OPN
set wServer=whois.ripe.net      # whois server
set wQuery=RS-LHCOPN            # whois query
#-----------------------------------------------#

set accio=add
if ($1 == "del") set accio=del

foreach ipLHCOPN (`whois -h $wServer $wQuery | grep members | awk '{print $2 }'`)
        set excloure = 0

        foreach ipExcepcio ($xarxesExcepcio)
                if ($ipExcepcio =~ $ipLHCOPN) set excloure = 1
        end

        if ($excloure == 0) /sbin/route $accio -net $ipLHCOPN gw $Gateway

end
```

8 Annex C: Proves de rendiment de la *LAN sobre la maqueta solució amb dues IPs*

Aquestes proves corresponen a la comunicació entres dos servidors, un LHC-OPN (2 IPs) i un no-LHC-OPN (1 IP) connectats al mateix switch sobre la maqueta (veure figura 8.0.1) resultat d'implementar la solució A sense divisió per VLANs seguint l'opció de "Gestió exclusiva per VLAN100", és a dir, a nivell de l'encaminador tot va per la VLAN100, a nivell del switch el trànsit és *untagged*.

Els servidors utilitzats per a les proves són tres Dell PowerEdge750 amb dues GigabitEthernet cadascun, els switch són Dell PowerConnect 5324.

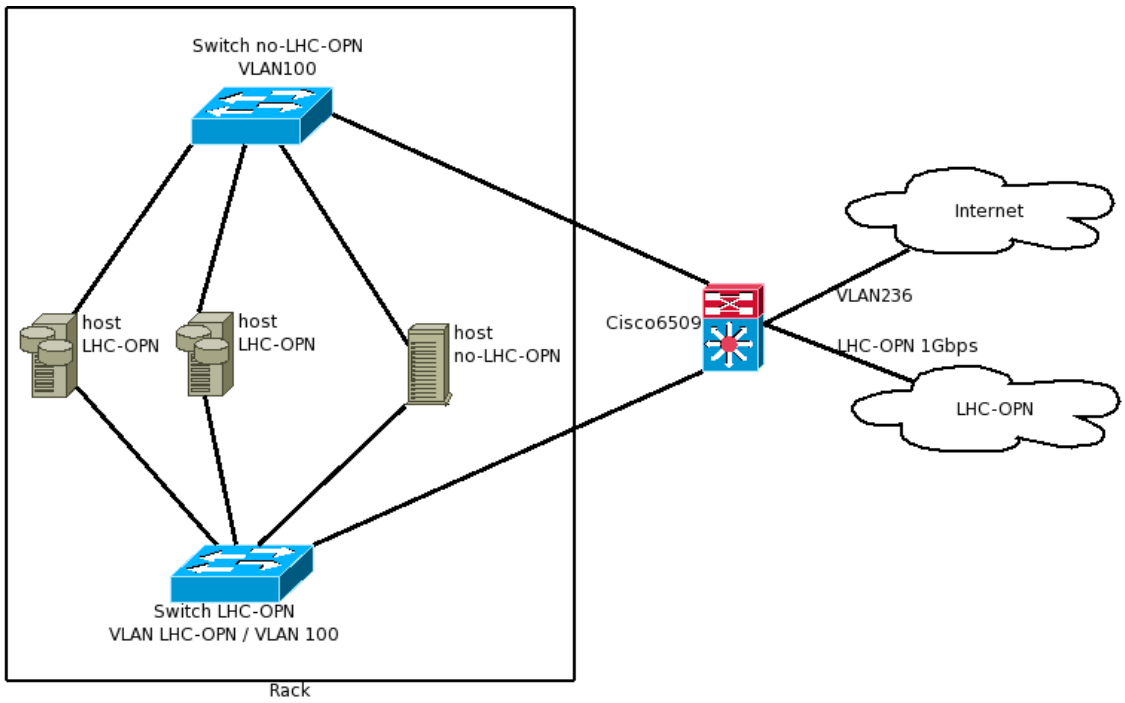


Figura 8.0.1: Maqueta implementada per a la primera fase del projecte, amb host LHC-OPN i no-LHC-OPN

La mida de la finestra (windowsize) utilitzada en les proves ha estat calculada, per tal d'evitar crear colls d'ampolla, utilitzant la fórmula[1]:

WindowSize = bandwidth*rtt

En aquestes proves de la LAN l'RTT[2] és inferior a 1ms, tal i com es veu en la prova de ping de la figura 8.0.2.

```
77 packets transmitted, 77 received, 0% packet loss, time 76005ms
rtt min/avg/max/mdev = 0.066/0.152/0.264/0.048 ms
```

Figura 8.0.2: prova de *ping* entre dos servidors de la maqueta (LAN), amb RTT inferiors a 1ms

Així doncs tindrem que la mida òptima de la finestra seria 1000mbps*1ms = 1000e6*1e-3 = 1000000

1 Es pot trobar una explicació més extensa a http://www.nps-llc.com/frequently_asked_questions.htm#TCPWindowsize1

2 El Round Trip Time és el temps que triga un paquet en arribar d'un extrem a l'altre de la connexió i tornar.

bits ~= **125 kbyte**

Per a un timeout de 50ms la finestra hauria de ser de 50.000.000 bits, uns 6 MBytes.

Com que la mida de la finestra depen directament de l'rtt, i en les proves actuals és despreciable, per a afegir més realisme s'utilitzarà el windowsize utilitzat actualment en els GridFTP[3] (2mb) del PIC. A més s'ha demostrat teòrica i empíricament (veure fibura 8.0.3) que el canvi de 125kbyte a 2Mbyte no decrementa el throughput de la xarxa[4].

```
[root@lhcopn03 root]# iperf -c lhcopn02 -w 1Mb
------------------------------------------------------------
Client connecting to lhcopn02, TCP port 5001
TCP window size: 2.00 MByte (WARNING: requested 1.00 MByte)
------------------------------------------------------------
[  3] local 193.146.197.155 port 32775 connected with 193.146.197.154 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec  1.10 GBytes   941 Mbits/sec

[root@lhcopn03 root]# iperf -c lhcopn02 -w 65kb
------------------------------------------------------------
Client connecting to lhcopn02, TCP port 5001
TCP window size:  127 KByte (WARNING: requested 63.5 KByte)
------------------------------------------------------------
[  3] local 193.146.197.155 port 32777 connected with 193.146.197.154 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec  1.10 GBytes   941 Mbits/sec
```

Figura 8.0.3: demostració del manteniment en el rendiment al augmentar la mida de la finestra de 127Kb a 2MB.

8.1 Proves de rendiment LAN unidireccional entre servidors del mateix switch

En les comparatives no apareix l'ús de CPU en els switch perquè els Dell PowerConnect 5324 utilitzats no en permeten la monitorització[5]. Els switch Dell utilitzats tenen una opció per activar específicament l'ús de JumboFrames en el switch, si aquesta opció no s'activa els paquets de 9000bytes són descartats sense generar cap tipus de missatge d'error (ICMP o altres), creant un *MTU Black Hole* similar al documentat en l'annex E.

Per a la realització de les proves d'rtt s'ha utilitzat ping (count=50), per al throughput s'ha utilitzat la versió 1.7.0 d'Iperf[6] (per defecte en el repositori *yum* de *Scientific Linux Cern 3*) i Thrulay[7] (en les proves amb transferències bidireccionals). En les proves actuals no s'ha estimat necessària la utilització de BWCTL[8] però no es descarta el seu ús en proves posteriors.

A la taula de la figura 8.1.1 es mostra un resum de les proves realitzades en funció de l'activació o no de *bonding* i/o JumboFrames.

3 Els GridFTP són els servidors encarregats de la recepció de dades via la LHC-OPN per als serveis Castor i dCache

4 El que Linux fa al incrementar la finestra de TCP és reservar més memòria per tal de poder mantenir més paquets dins la finestra que s'està enviant.

5 http://www.dellcommunity.com/supportforums/board/message?board.id=pc_managed&message.id=7657&c=us&l=en&cs=&s=gen

6 http://dast.nlanr.net/Projects/Iperf/

7 http://shlang.com/thrulay/ Els resultats s'han contrastat amb resultats d'Iperf i amb les eines de monitorització del propi switch, donant valors molt similars.

8 http://e2epi.internet2.edu/bwctl/architecture.html

Configuració Xarxa		Paràmetres TCP/IP (iperf)			Resultats (iperf)			
Bonding	JumboFrame	WindowSize TCP (-w)/ DatagramSize UDP (-l)	Temps (-t)	Bandwith UDP (-b)	Paquets perduts/total UDP	Jitter UDP, en ms	Throughput, en Mbps (i % respecte 1 Gbps)	RTT min/promig/max/ max desv en ms (amb ping)
NO	NO	2 Mbytes	10	-	-	-	941 (94,1%)	0.060/0.154/0.463/ 0.066
		2 Mbytes	100	-	-	-	**941 (94,1%)**	
		1470 bytes	10	1 Gbps	0/813836	0.027	957 (95,7%)	
		1470 bytes	100	1 Gbps	0/813805	0.018	957 (95,7%)	
	SI	2 Mbytes	10	-	-	-	990 (99,0%)	0.045/0.147/0.240/ 0.048
		2 Mbytes	100	-	-	-	**990 (99,0%)**	
		1470 bytes	10	1 Gbps	0/138649	0.061	993 (99,3%)	
		1470 bytes	100	1 Gbps	0/138643	0.090	993 (99,3%)	
SI	SI	2 Mbytes	10	-	-	-	990 (99,0%)	0.081/0.161/0.238/ 0.040
		2 Mbytes	100	-	-	-	**990 (99,0%)**	
		1470 bytes	10	1 Gbps	0/138790	0.059	993 (99,3%)	
		1470 bytes	100	1 Gbps	0/1386654	0.054	**993 (99,3%)**	

Figura 8.1.1: Taula resum de les proves de rendiment en transferències unidireccionals. En negreta ressaltats els casos més representatius, on es pot observar que el millor cas és amb JumboFrames activat, on s'aconsegueixen rendiments del 99% TCP i 99,3% UDP. Es constata que l'activació o no de Bonding (en mode de backup) no afecta al rendiment de la xarxa. Les proves s'han realitzat en la VLAN100.

8.2 Proves de rendiment LAN bidireccional entre servidors

Un cop comprovat el funcionament en transmissions unidireccionals s'ha provat el comportament dels servidors i els switch davant de connexions bidireccionals.

Les proves mostrades a continuació s'han realitzat mitjançant connexions TCP amb Thrulay[9], donant resultats pràcticament idèntics a iperf però en un format més òptim per a gnuplot. Els detalls de les proves es poden trobar al document *taula rendiment maqueta sobre LAN - detall proves.*

8.2.1 Proves de referència amb cable creuat

En les proves amb cable creuat els resultats son els mateixos amb bonding activat i desactivat; es pot observar que la velocitat, quan només hi ha transferències en un sentit, amb MTU=1500 és d'uns 950Mbps i amb JumboFrames (MTU=9000) d'uns 990Mbps.

Tan bon punt s'inicia la transmissió en sentit contrari la velocitat de la connexió baixa, i es mantenen les connexions en ambdós sentits sostingudes a uns 700Mbps en el cas de MTU=1500 (veure figura 8.2.1) , essent les velocitats uns 20-50Mbps inferiors per a MTU=9000 (veure figura 8.2.2).

9 http://shlang.com/thrulay/ Els resultats s'han contrastat amb resultats d'Iperf i amb les eines de monitorització del propi switch, donant valors molt similars.

Cal notar que la suma d'ambdues transferències supera el llindar d'1 Gbps, però no arriba en cap cas als 2Gbps teòrics (full duplex). El màxim throughput que es pot observar és d'uns 1400Mbps (700+700).

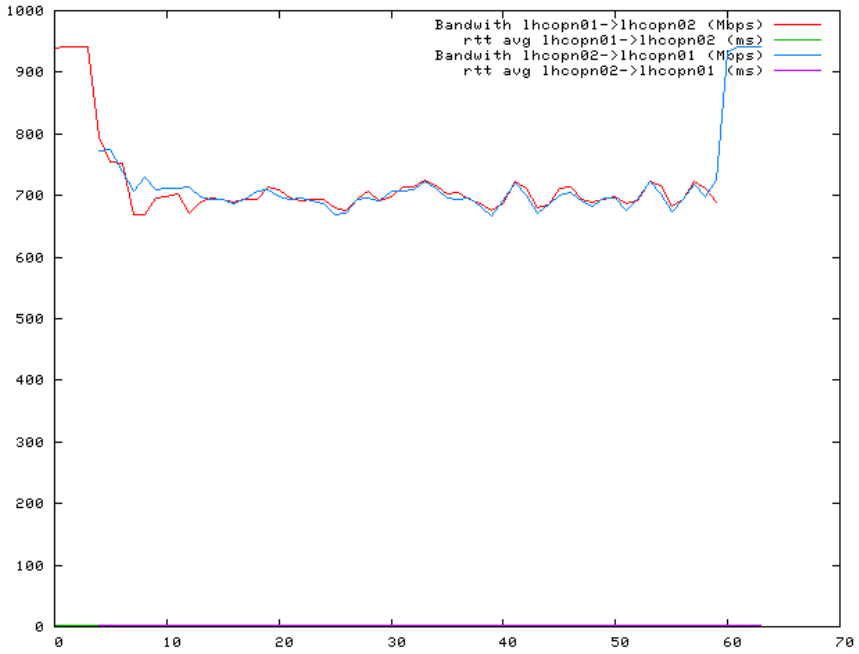


Figura 8.2.1: Cable creuat, MTU=1500 en els servidors (resta de la configuració per defecte en els servidors), cas òptim sense JumboFrames

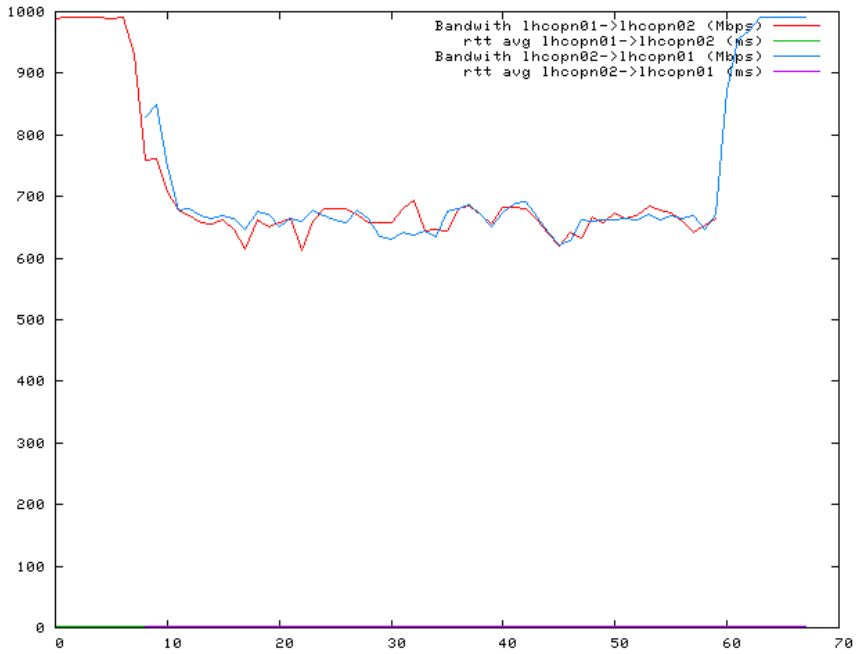


Figura 8.2.2: Cable creuat, MTU=9000 en els servidors (resta de la configuració per defecte en els servidors), cas òptim amb JumboFrames

Al utilitzar un cable creuat per al connexionat evitem qualsevol interferència introduïda per switch/encaminadors, aconseguint així el throughput màxim entre ambdós servidors. Així doncs les proves amb cable creuat serveixen com a punt de referència per a les que es realitzaran al subapartat 8.2.2 mitjançant un switch.

8.2.2 Proves utilitzant switch Dell PowerConnect 5324

Al inserir un switch en el sistema i amb un MTU=1500 en els servidors, i sense JumboFrames activat en el switch, l'activació (veure figura 8.2.3) o no (veure figura 8.2.4) del FlowControl no afecta a l'ample de banda disponible per a les connexions TCP.

És important observar el notable descens de l'ample de banda disponible quan les transferències es realitzen en ambdós sentits, que és pràcticament la meitat del disponible amb transferències en un únic sentit.

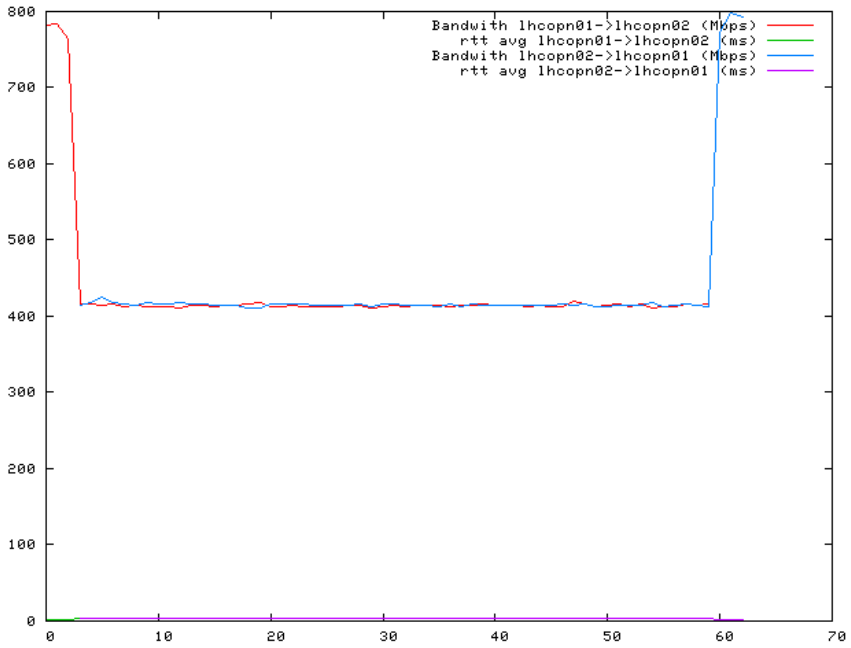


Figura 8.2.3: passant pel Switch, no JumboFrame, si Flow Control, MTU=1500 en els servidors

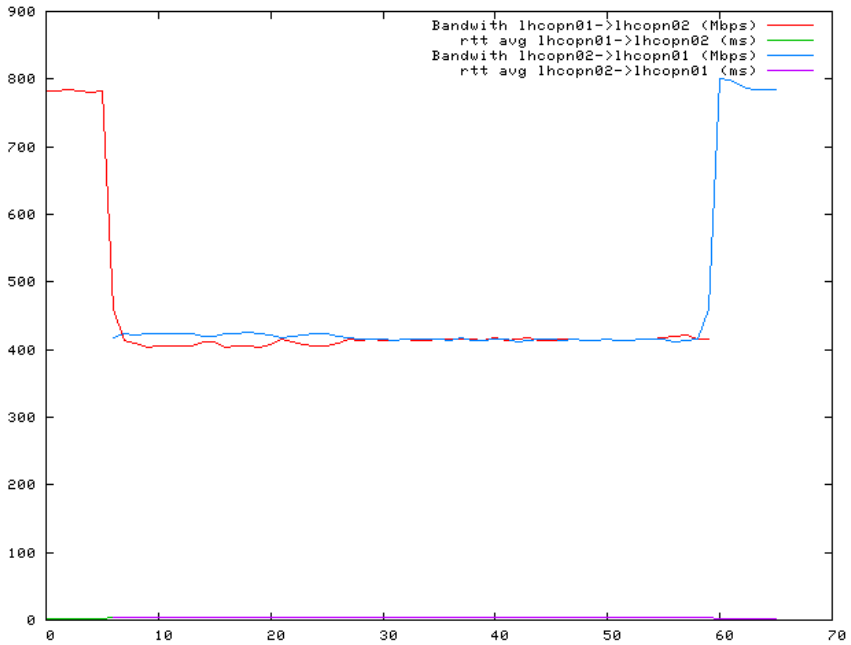


Figura 8.2.4: passant pel Switch, no JumboFrame, no Flow Control, MTU=1500 en els servidors

Al activar JumboFrames en el switch les transferències amb MTU=1500 (veure figura 8.2.5) en els servidors han mantingut el comportament anteriorment observat. Al posar MTU=9000 (veure figura 8.2.6) en els servidors les transferències en un únic sentit han millorat fins als ~950Mbps (+150mbps), aconseguint velocitats properes al màxim observat amb el cable creuat. En canvi les transferències en ambdós sentits continuen mostrant el mateix comportament; ~400Mbps per connexió.

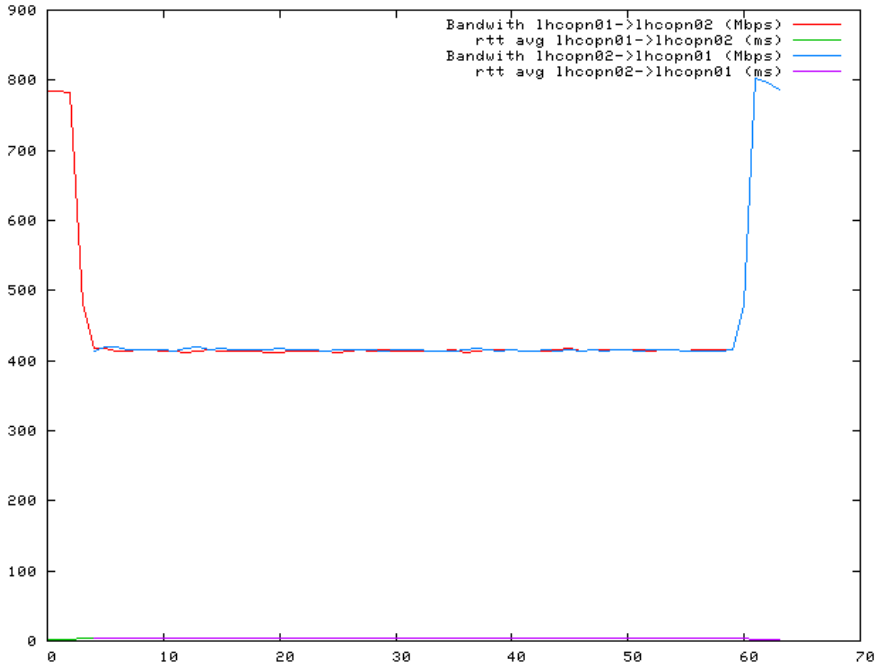


Figura 8.2.5: passant pel Switch, si JumboFrame, no Flow Control, MTU=1500 en els servidors

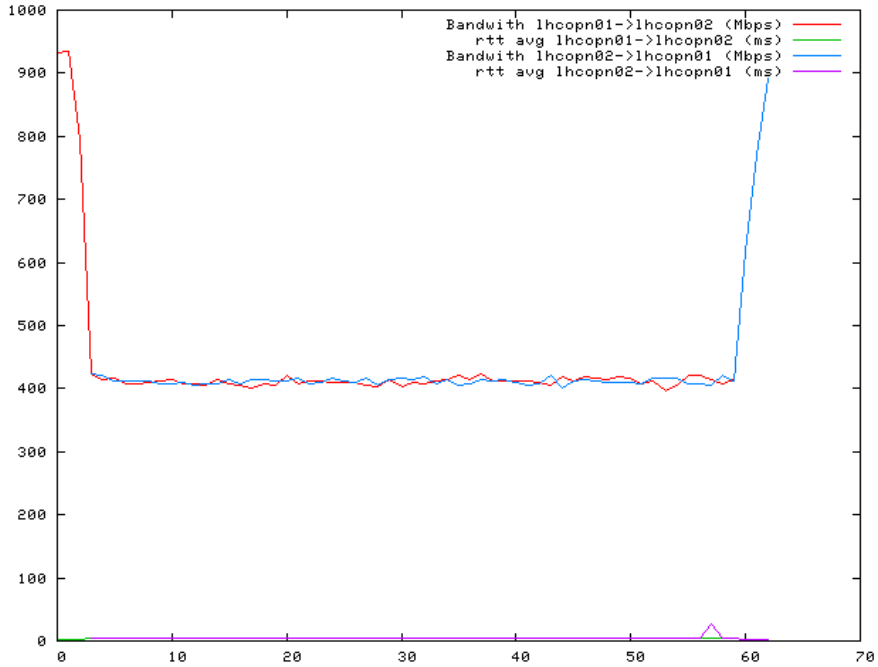


Figura 8.2.6: passant pel Switch, si JumboFrame, no Flow Control, MTU=9000 en els servidors

Tal i com es mostra a la gràfica de la figura 8.2.7 al activar el Flow Control, amb JumboFrames activat, el comportament es manté per a MTU=1500: ~800mbps individual i ~400mbps en

transferències simultànies en ambdós sentits. Al posar MTU=9000 (figura 8.2.8) en els servidors el throughput de les transferències es desestabilitza i baixa dels ~950mbps a ~500mbps en el cas d'un únic sentit i a un rang de 50-400mbps en transferències en ambdós sentits.

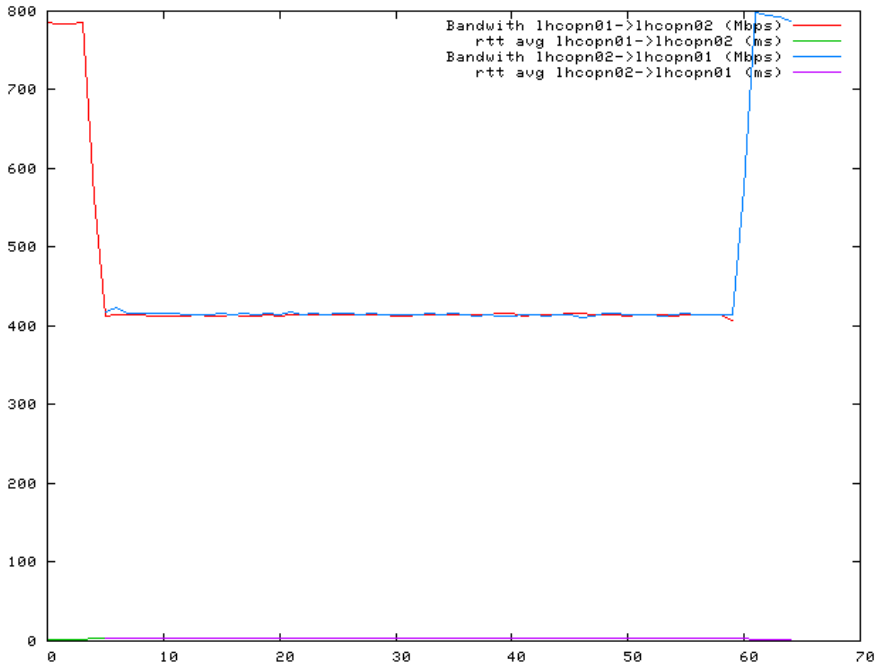


Figura 8.2.7: passant pel Switch, si JumboFrame, si Flow Control, MTU=1500 en els servidors

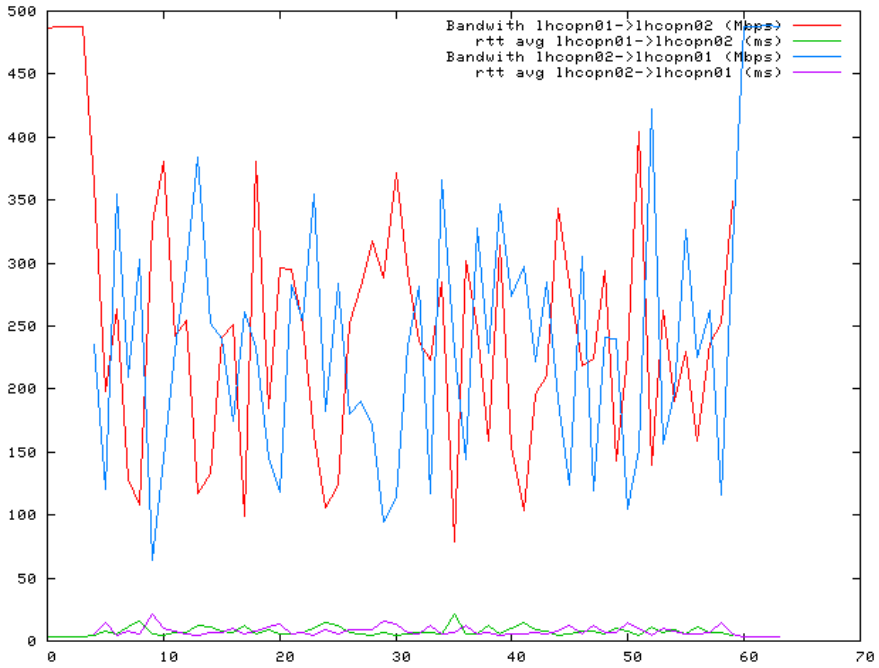


Figura 8.2.8: Passant pel Switch, si JumboFrame, si Flow Control, MTU=9000 en els servidors

8.3 Prova amb múltiples transmissions: lhcopn01->lhcopn02->lhcopn03->lhcopn01

En la gràfica de la figura 8.3.1 es mostren diferents transferències realitzades de forma gradual formant un cercle. Es pot observar que la limitació es troba en l'ample de banda disponible "per port" i no a nivell del *backend* del switch, així podem dir que el comportament dels ports GigabitEthernet del Dell PowerConnect5324 funcionen com si fóssin halfduplex, és a dir, amb un throughput màxim de 1 Gbps sumant el throughput en ambdós sentits.

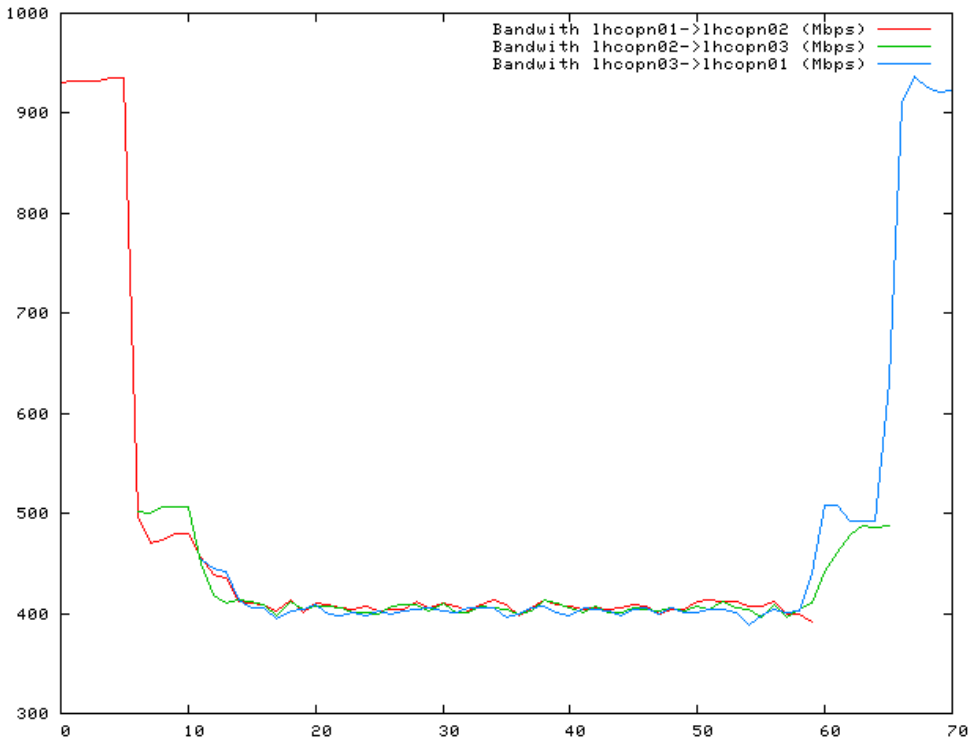


Figura 8.3.1: passant pel switch, si JumboFrames, no Flow Control, MTU=9000 en els servidors

8.4 Conclusions de les proves de rendiment sobre la LAN

Al llarg de la implementació de la maqueta s'ha trobat que, degut a que el mòdul de bonding s'apropia del control de les NIC, és impossible crear NICs virtuals que permetin un segon bonding amb característiques pròpies, així s'han hagut de desestimar algunes propostes inicials que pretenien aconseguir dues interfícies de bonding redundants amb només dues NIC. Finalment per tal de d'obtenir dues interfícies de bonding amb característiques independents és necessari disposar d'un mínim de 4 NIC, dues per cada interfície de bonding.

L'activació del JumboFrame (MTU màxim de 9000 bytes en switch i servidors) comporta una millora d'un 4,9% en el throughput de les connexions TCP i d'un 4% en les connexions UDP.

També és important és la descàrrega que l'activació de JumboFrame suposa per a la CPU dels servidors, que passa de l'10,9% al 8,4% en el cas de connexions TCP i del 62,7% al 56,8% en el cas d'UDP. Podem suposar que l'ús de CPU en els switch també ha decrementat, ja que al tenir un

MTU més gran, per a la mateixa quantitat de bytes no s'han de gestionar tants paquets.

Respecte a l'RTT les variacions són mínimes.

L'activació de Bonding+JumboFrames no fa variar el throughput respecte a l'obtingut només amb JumboFrames, en canvi l'ús de la CPU es duplica i passa a ser del 20%, en el cas d'una connexió TCP.

Respecte a les proves de transferències bidireccionals, òbviament la configuració que millor rendiment ofereix és amb el cable creuat, però només serveix com a referència. Així doncs la millor opció, passant pel switch Dell PowerConnect 5324, és:

- JumboFrame activat
- Flow Control desactivat
- MTU=9000 en els servidors

És important tenir en compte la limitació dels switch Dell PowerConnect 5324 que, tot i especificar que són switch 1 Gbps Full Duplex, es comporten com switch 1 Gbps Half Duplex.

Els detalls de la configuració de la maqueta definitiva es poden consultar a l'annex D

9 Annex D: Detalls de la configuració de la maqueta local

A continuació es mostren els detalls de la configuració dels servidors i dispositius de xarxa que formen part de la maquetaimplementada per a provar la validesa de les solució per a la integració de la xarxa LHC-OPN sobre el circuit dedicat de 1 Gbps. A la figura 9.0.1 es mostra la versió final de la maqueta, utilitzant bonding i amb JumboFrames activat.

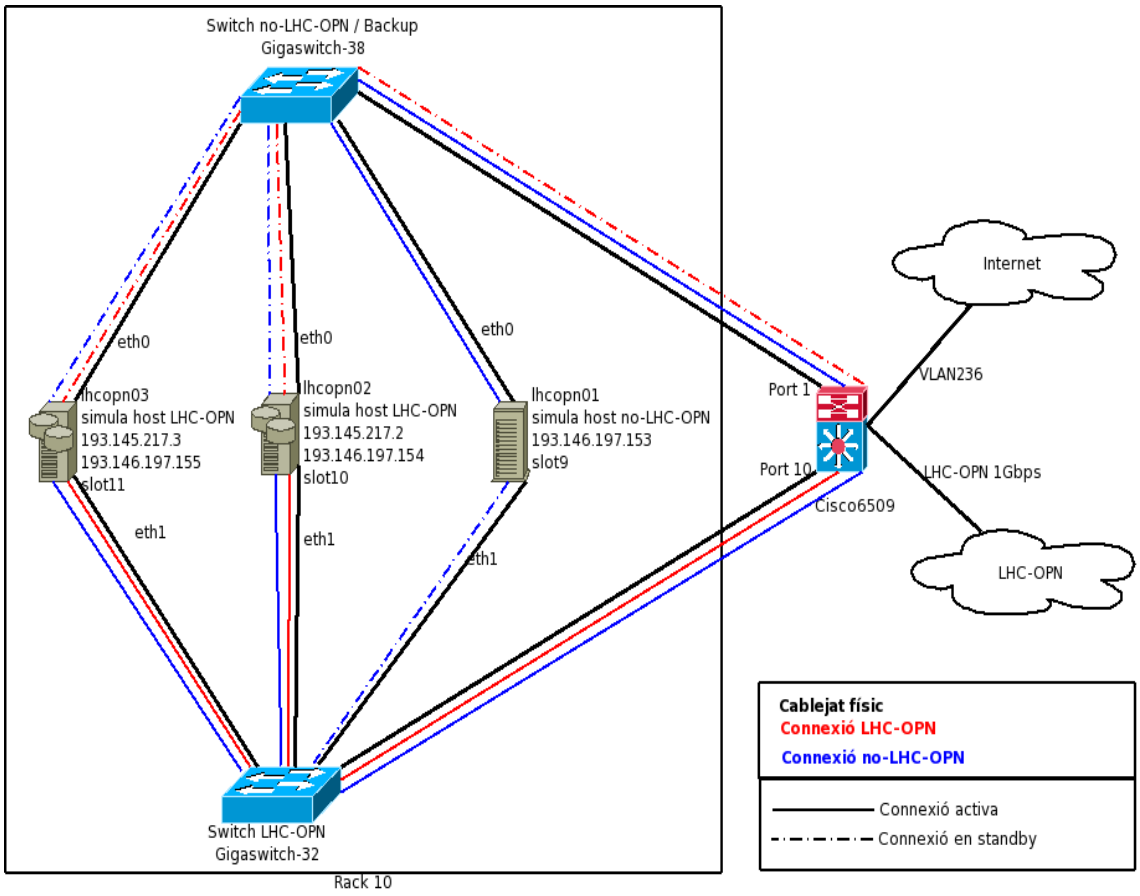


Figura 9.0.1: Maqueta per a la integració dels servidors sobre la connexió de 1 Gbps

Configuració de l'encaminador Cisco6509 (extracte)

```
vlan 100
 name PIC
!
interface GigabitEthernet3/41
description Connexio VLAN100 PIC
switchport
switchport access vlan 100
switchport mode access
mtu 9216
no ip address
logging event link-status
no cdp enable
!
interface GigabitEthernet3/44
description Connexio VLAN100 PIC
switchport
```

```
switchport access vlan 100
switchport mode access
mtu 9216
no ip address
logging event link-status
no cdp enable
!
interface GigabitEthernet3/45
description Connexio LHC
bandwidth 1000000
ip address 130.206.202.26 255.255.255.252
ip access-group 151 in
no ip redirects
ip route-cache flow
logging event link-status
no cdp enable
!
interface Vlan100
description CONNEXIO PIC
mtu 9216
ip address 193.145.217.1 255.255.255.0 secondary
ip address 193.146.196.130 255.255.252.0
ip access-group 103 out
no ip redirects
ip flow ingress
ip policy route-map IFAE
logging event link-status
mls netflow sampling
standby 2 ip 193.146.196.10
standby 2 priority 110
standby 2 preempt
standby 2 track GigabitEthernet3/47 50
!
ip classless
ip default-network 0.0.0.0
ip route 0.0.0.0 0.0.0.0 130.206.202.25
```

Configuració dels switch Dell PowerConnect5324

```
no spanning-tree
port jumbo-frame
interface vlan 1
ip address 193.146.199.X 255.255.252.0
ip address 193.145.217.X 255.255.255.0
exit
ip default-gateway 193.146.196.10
no qos
hostname gigaswitch-32
management access-list HTTPS/SSH
permit service https
permit service ssh
deny
exit
management access-class HTTPS/SSH
aaa authentication login default local
username admin password 2c22dc1f4f62dcfe69c3702ca51d4db9 level 15 encrypted
username operador password b9893b42ec4a597eaaa307690b63940d level 7 encrypted
ip ssh server
snmp-server community public  193.146.196.4
snmp-server community picsnmp  193.146.196.51
snmp-server community picsnmp  193.146.197.121
snmp-server community picsnmp  193.146.197.144
no ip http server
ip https server
clock source sntp
sntp unicast client poll
sntp server 130.206.3.166
sntp server 192.101.162.68
ip domain-name pic.es
ip name-server  193.146.196.3
```

Per a configurar els switch Dell PowerConnect 5324 cal fer login amb l'usuari admin, executar la

comanda *configure*, posar la configuració anterior, executar la comanda *end*, guardar la configuració amb *copy run start* i, finalment, reiniciar el switch mitjançant *reload*.

Configuració dels servidors LHC-OPN (lhcopn03)

cat /etc/modules.conf

```
alias eth0 e1000
alias eth1 e1000
alias scsi_hostadapter ata_piix
alias usb-controller usb-uhci
alias usb-controller1 ehci-hcd
###Inici BONDING###
#  no-LHC-OPN
add above bonding e1000
alias bond0 bonding
options bond0 mode=1 miimon=100 primary=eth1
###Fi BONDING###
```

cat /etc/rc.local

```
#!/bin/sh
#
# This script will be executed *after* all the other init scripts.
# You can put your own initialization stuff in here if you don't
# want to do the full Sys V style init stuff.
touch /var/lock/subsys/local

#BONDING
/etc/init.d/network restart
route del -net 193.145.217.0/24 dev eth0
route del -net 193.145.217.0/24 dev eth1
#FI BONDING
#És necessari modificar alguns paràmetres del kernel 2.4 de SLC3 a /proc per tal
de:
#Ampliar límits en els tamanys de finestra (min/inicial/max)
echo "4096 87380 128388607" > /proc/sys/net/ipv4/tcp_rmem
echo "4096 65530 128388607" > /proc/sys/net/ipv4/tcp_wmem
echo 128388607 > /proc/sys/net/core/wmem_max
echo 128388607 > /proc/sys/net/core/rmem_max
#Habilitar les opcions avançades standard de TCP
echo 1 > /proc/sys/net/ipv4/tcp_timestamps
echo 1 > /proc/sys/net/ipv4/tcp_window_scaling
echo 1 > /proc/sys/net/ipv4/tcp_sack
#NOTA: això també es pot realitzar amb sysctl, modificant /etc/sysctl.conf per
afegir els valors anteriors
```

cat /etc/sysconfig/network-scripts/ifcfg-bond0

```
DEVICE=bond0
USERCTL=no
ONBOOT=yes
IPADDR=193.145.217.3
NETMASK=255.255.255.0
NETWORK=193.145.217.0
BROADCAST=193.145.217.255
MTU=9000
BOOTPROTO=none
```

cat /etc/sysconfig/network-scripts/ifcfg-bond0:1

```
DEVICE=bond0:1
USERCTL=no
ONBOOT=yes
IPADDR=193.146.197.155
NETMASK=255.255.252.0
MTU=9000
```

cat /etc/sysconfig/network-scripts/ifcfg-eth0

```
DEVICE=eth0
USERCTL=no
ONBOOT=yes
MASTER=bond0
SLAVE=yes
BOOTPROTO=none
```

cat /etc/sysconfig/network-scripts/ifcfg-eth1

```
DEVICE=eth1
USERCTL=no
ONBOOT=yes
MASTER=bond0
SLAVE=yes
BOOTPROTO=none
```

Configuració dels servidors no-LHC-OPN (lhcopn01)

cat /etc/modules.conf

```
alias eth0 e1000
alias eth1 e1000
alias scsi_hostadapter ata_piix
alias usb-controller usb-uhci
alias usb-controller1 ehci-hcd
###Inici Bonding###
#probeall bond0 eth0 eth1 bonding
add above bonding e1000
alias bond0 bonding
options bond0 mode=1 miimon=100 primary=eth0
###Fi Bonding###
```

cat /etc/rc.local

```
#!/bin/sh
#
# This script will be executed *after* all the other init scripts.
# You can put your own initialization stuff in here if you don't
# want to do the full Sys V style init stuff.

touch /var/lock/subsys/local
#BONDING
/etc/init.d/network restart
route del -net 193.146.196.0/22 dev eth0
route del -net 193.146.196.0/22 dev eth1
#FI BONDING
#És necessari modificar alguns paràmetres del kernel 2.4 de SLC3 a /proc per tal
de:
#Ampliar límits en els tamanys de finestra (min/inicial/max)
echo "4096 87380 128388607" > /proc/sys/net/ipv4/tcp_rmem
echo "4096 65530 128388607" > /proc/sys/net/ipv4/tcp_wmem
echo 128388607 > /proc/sys/net/core/wmem_max
echo 128388607 > /proc/sys/net/core/rmem_max
#Habilitar les opcions avançades standard de TCP
echo 1 > /proc/sys/net/ipv4/tcp_timestamps
echo 1 > /proc/sys/net/ipv4/tcp_window_scaling
echo 1 > /proc/sys/net/ipv4/tcp_sack
#NOTA: això també es pot realitzar amb sysctl, modificant /etc/sysctl.conf per
afegir els valors anteriors
```

cat /etc/sysconfig/network-scripts/ifcfg-bond0

```
DEVICE=bond0
USERCTL=no
ONBOOT=yes
IPADDR=193.146.197.153
NETMASK=255.255.252.0
BOOTPROTO=n
```

cat /etc/sysconfig/network-scripts/ifcfg-eth0

```
DEVICE=eth0
USERCTL=no
ONBOOT=yes
MASTER=bond0
SLAVE=yes
BOOTPROTO=none
```

cat /etc/sysconfig/network-scripts/ifcfg-eth1

```
DEVICE=eth1
USERCTL=no
ONBOOT=yes
MASTER=bond0
SLAVE=yes
BOOTPROTO=none
```

10 Annex E: Path MTU Black Hole Detection and Recovery

En una xarxa poden aparèixer escenaris on les màquines queden literalment *sordes*, a causa de que no reben els paquets de retorn ICMP que indiquen que s'ha de reduir la mida del paquet. Als equips on es filtren, o no generen, els paquets ICMP se'ls anomena "forats negres", ja que es *mengen* els paquets de control que permeten que el descobriment de l'MTU funcioni.

A continuació es mostra una traça extreta amb tcpdump[1] on s'evidencia el *forat negre* (pmtublackhole) que apareix a 130.206.202.26 (lhc-router.red.rediris.es) quan l'MTU és de 1500bytes als dos costats de la connexió, el problema està àmpliament documentat[26]. Ambdues traces estan extretes des de lhcopn03, on es disposa de permisos per executar *tcpdump*:

Connexió TCP chapuza->lhcopn03

```
11:40:33.753218 chapuza.cern.ch.38225 > 193.145.217.3.5001: S 727424485:727424485(0) win
5840 <mss 1460,sackOK,timestamp 1745003986 0,nop,wscale 0> (DF)
```

S'estableix la connexió chapuza->lhcopn03 (SYN), indicant un mss de 1460 (=>MTU de 1500), el flag Don't Fragment (DF) està activat

```
11:40:33.753254 193.145.217.3.5001 > chapuza.cern.ch.38225: S 873981383:873981383(0) ack
727424486 win 5792 <mss 1460,sackOK,timestamp 85852662 1745003986,nop,wscale 0> (DF)
```

S'estableix la connexió lhcopn03->chapuza (SYN) i es confirma la connexió del paquet anterior (ACK), indicant un mss de 1460 (=>MTU de 1500), el flag Don't Fragment (DF) està activat

```
11:40:33.791819 chapuza.cern.ch.38225 > 193.145.217.3.5001: . ack 1 win 5840
<nop,nop,timestamp 1745003989 85852662> (DF)
```

chapuza confirma la connexió lhcopn03->chapuza

```
11:40:33.791834 chapuza.cern.ch.38225 > 193.145.217.3.5001: P 1:25(24) ack 1 win 5840
<nop,nop,timestamp 1745003989 85852662> (DF)
```

chapuza envia inicialment un paquet amb 24bytes de dades

```
11:40:33.791840 193.145.217.3.5001 > chapuza.cern.ch.38225: . ack 25 win 5792
<nop,nop,timestamp 85852666 1745003989> (DF)
```

lhcopn03 confirma la recepció dels primers 24 bytes de dades (ACK 25)

```
11:41:41.365527 193.145.217.3.5001 > chapuza.cern.ch.38225: F 1:1(0) ack 25 win 5792
<nop,nop,timestamp 85859423 1745003989> (DF)
```

Han passat 8 segons i no s'ha rebut res més, això és degut a que chapuza ha estat enviant paquets de mss=1460, els quals s'han perdut al forat negre. Des de lhcopn03 es cancel·la manualment la connexió (FIN).

```
11:41:41.413895 chapuza.cern.ch.38225 > 193.145.217.3.5001: . ack 2 win 5840
<nop,nop,timestamp 1745010752 85859423> (DF)
```

chapuza confirma (ACK) el FIN rebut

Tal i com es pot observar en les traces mostrades des des de lhcopn03.pic.es no es reben paquets de dades provinents de chapuza.cern.ch i, finalment, la connexió és cancel·lada (manualment) des de lhcopn03.

Connexió TCP lhcopn03->chapuza

```
11:35:47.471365 193.145.217.3.33026 > chapuza.cern.ch.5001: S 570121783:570121783(0) win
5840 <mss 1460,sackOK,timestamp 85824034 0,nop,wscale 0> (DF)
```

S'estableix la connexió lhcopn03-> chapuza (SYN), indicant un mss de 1460 (=>MTU de 1500), el flag Don't Fragment (DF) està activat

1 Des de lhcopn03 ja que en chapuza.cern.ch no hi ha permisos suficients per executar *tcpdump -i eth2*

```
11:35:47.509908 chapuza.cern.ch.5001 > 193.145.217.3.33026: S 419772170:419772170(0) ack
570121784 win 5792 <mss 1460,sackOK,timestamp 1744975362 85824034,nop,wscale 0> (DF)
```

S'estableix la connexió lhcopn03-> chapuza (SYN) i es confirma la connexió del paquet anterior (ACK), indicant un mss de 1460 (=>MTU de 1500), el flag Don't Fragment (DF) està activat

```
11:35:47.509921 193.145.217.3.33026 > chapuza.cern.ch.5001: . ack 1 win 5840
<nop,nop,timestamp 85824037 1744975362> (DF)
```

lhconp03 confirma la connexió chapuza-> lhcopn03

```
11:35:47.509966 193.145.217.3.33026 > chapuza.cern.ch.5001: P 1:25(24) ack 1 win 5840
<nop,nop,timestamp 85824037 1744975362> (DF)
```

lhcopn03 envia inicialment un paquet (1) amb 24bytes de dades

```
11:35:47.510000 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824037 1744975362> (DF)
```

lhcopn03 envia un paquet (2) amb 1448bytes de dades (mida total del missatge=1448+40=1488)

```
11:35:47.549010 chapuza.cern.ch.5001 > 193.145.217.3.33026: . ack 25 win 5792
<nop,nop,timestamp 1744975366 85824037> (DF)
```

chapuza confirma la recepció dels primers 24 bytes de dades (ACK 25)

```
11:35:47.549017 193.145.217.3.33026 > chapuza.cern.ch.5001: . 1473:2921(1448) ack 1 win
5840 <nop,nop,timestamp 85824041 1744975366> (DF)
```

lhcopn03 envia un paquet (3) amb 1448bytes de dades

```
11:35:47.549021 193.145.217.3.33026 > chapuza.cern.ch.5001: P 2921:4369(1448) ack 1 win
5840 <nop,nop,timestamp 85824041 1744975366> (DF)
```

lhcopn03 envia un paquet (4) amb 1448bytes de dades

```
11:35:47.771347 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824064 1744975366> (DF)
11:35:48.231346 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824110 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

```
11:35:49.151346 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824202 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

```
11:35:50.991348 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824386 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

```
11:35:54.671348 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85824754 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

```
11:36:02.031349 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85825490 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

```
11:36:16.751351 193.145.217.3.33026 > chapuza.cern.ch.5001: . 25:1473(1448) ack 1 win 5840
<nop,nop,timestamp 85826962 1744975366> (DF)
```

lhcopn03 reenvia el paquet (2), de 1448bytes de dades (ha arribat al timeout sense rebre ACK i es retransmet)

....

Tal i com es pot observar no es rep la confirmació (ACK) dels paquets amb MTU 1500 (tots excepte el primer en el qual la mida de la trama és inferior: 24[dades]+40[capçaleres IP+TCP]). Aquest comportament de TCP ens revela que els paquets no s'estan reben a chapuza.cern.ch

Amb una metodologia de prova i error mitjançant l'enviament de paquets ICMP (ping -s), UDP (iperf -l) i TCP (modificant MTU per alterar l'MSS) des del PIC s'ha trobat que l'MTU màxim admès al *forat negre* és de 1494 bytes.

Al realitzar connexions des del CERN no apareix el *forat negre* que “es veu” des del PIC, en aquest cas si que es segueix el protocol descrit a l'RFC1191 (Path MTU discover), anunciant un MTU de 1486bytes.

11 Annex F: Sensor de Nagios i procediment d'alarma

Per tal de poder detectar errors en l'establiment de connexions TCP PIC->CERN s'ha dissenyat un sensor per a *Nagios* i *Ganglia*, dos sistemes de monitorització utilitzats al PIC.

També s'ha establert un procediment d'actuació per a quan es generen les alarmes des de *Nagios*:

11.1 Procediment d'actuació

El procediment d'actuació que es mostra a continuació es troba publicat al WIKI intern del PIC, a on s'enllaça quan salta l'alarma corresponent.

Esta alarma significa que hay un problema de conexión TCP con un servidor externo.

Output del error: Fallos de connexion SYN con **castorgrid.cern.ch**: 5/5 Fallos descubrimiento PING: 5/5
El problema puede ser tanto a nivel local (servidor que hace la prueba) como remoto (servidor al que se intenta la conexión), así como un problema de red general (ie red GÉANT) o concreto (switch del servidor).
Para intentar localizar el problema se seguiran los pasos a continuación:

Se seguirà el procediment estàndard per a la resolució de fallides de xarxa:
https://www.wiki.pic.es/index.php/Procediment_en_cas_de_problemes_de_xarxa

Addicionalment, per ajudar a diagnosticar el problema es pot seguir el següent procediment:

Executar les comandes

```
traceroute  -I castorgrid.cern.ch -A                 #ICMP - Ruta OK
traceroute  -I castorgrid.cern.ch 1472 -A -F         #ICMP - MTU 1500 OK
traceroute  -I castorgrid.cern.ch 2000 -A            #ICMP - Fragmentació OK
traceroute  -T -p2811 castorgrid.cern.ch -A          #TCP OK
```

Cal examinar el resultat dels diferents traceroute que, si tot anés bé, hauria de ser similar al que es mostra a continuació:

traceroute -I castorgrid.cern.ch -A #ICMP - Ruta OK
traceroute -I castorgrid.cern.ch 1472 -A -F #ICMP - MTU 1500 OK

```
traceroute to castorgrid.cern.ch (128.142.175.74), 30 hops max, 40 byte packets
 1  cisco1.pic.org.es (193.146.196.130) [AS766]  1.341 ms  1.437 ms  1.535 ms
 2  anella-ifae.cesca.es (84.88.19.9) [AS13041]  1.017 ms  1.135 ms  1.129 ms
 3  AE0.EB-Barcelona0.red.rediris.es (130.206.202.1) [AS766]  1.233 ms  1.233 ms  1.341 ms
 4  * * *
 5  SO0-0-0.EB-IRIS4.red.rediris.es (130.206.240.2) [AS766]  16.174 ms  16.286 ms  16.285 ms
 6  rediris.rt1.mad.es.geant2.net (62.40.124.53) [AS20965]  15.541 ms  15.467 ms  15.461 ms
 7  so-7-2-0.rt1.gen.ch.geant2.net (62.40.112.25) [AS20965]  37.439 ms  37.449 ms  37.840 ms
 8  * * *
 9  e513-e-rci76-1-swice2.cern.ch (192.65.184.222) [AS513]  37.579 ms  37.550 ms  37.625 ms
10  l513-c-rftec-1-be8.cern.ch (192.16.166.129) [AS513]  37.616 ms  37.607 ms  37.579 ms
11  * * *
12  * * *
13  castorgrid03.cern.ch (128.142.175.74) [AS513]  37.823 ms  37.929 ms  37.819 ms
```

traceroute -I castorgrid.cern.ch 2000 -A #ICMP - Fragmentació OK

```
traceroute to castorgrid.cern.ch (128.142.175.73), 30 hops max, 2000 byte packets
 1  cisco1.pic.org.es (193.146.196.130) [AS766]  2.204 ms  2.278 ms  2.379 ms
 2  anella-ifae.cesca.es (84.88.19.9) [AS13041]  2.374 ms  2.585 ms  2.814 ms
 3  AE0.EB-Barcelona0.red.rediris.es (130.206.202.1) [AS766]  2.803 ms  3.032 ms  3.142 ms
 4  * * *
 5  * * *
 6  * rediris.rt1.mad.es.geant2.net (62.40.124.53) [AS20965]  17.470 ms  17.526 ms
 7  so-7-2-0.rt1.gen.ch.geant2.net (62.40.112.25) [AS20965]  38.484 ms * *
 8  swiCE2-10GE-1-1.switch.ch (62.40.124.22) [AS20965]  38.692 ms  38.719 ms  38.678 ms
 9  e513-e-rci76-1-swice2.cern.ch (192.65.184.222) [AS513]  38.599 ms * *
10  * * l513-c-rftec-1-be8.cern.ch (192.16.166.129) [AS513]  39.423 ms
11  * * *
12  * * *
.....
X  castorgrid02.cern.ch (128.142.175.73) [AS513]  38.853 ms  38.868 ms *
```

traceroute -T -p2811 castorgrid.cern.ch -A #TCP OK

```
traceroute to castorgrid.cern.ch (128.142.175.73), 30 hops max, 40 byte packets
 1  cisco1.pic.org.es (193.146.196.130) [AS766]  1.214 ms  1.262 ms  1.264 ms
 2  anella-ifae.cesca.es (84.88.19.9) [AS13041]  1.095 ms  1.196 ms  1.179 ms
 3  AE0.EB-Barcelona0.red.rediris.es (130.206.202.1) [AS766]  1.230 ms  1.332 ms  1.317 ms
 4  CAT.XE6-0-0.EB-IRIS2.red.rediris.es (130.206.250.25) [AS766]  15.650 ms  15.659 ms  15.639 ms
 5  SO0-0-0.EB-IRIS4.red.rediris.es (130.206.240.2) [AS766]  15.598 ms  15.596 ms  15.574 ms
 6  rediris.rt1.mad.es.geant2.net (62.40.124.53) [AS20965]  15.546 ms  15.586 ms  15.709 ms
 7  so-7-2-0.rt1.gen.ch.geant2.net (62.40.112.25) [AS20965]  37.717 ms  37.686 ms  37.644 ms
 8  swiCE2-10GE-1-1.switch.ch (62.40.124.22) [AS20965]  37.808 ms  41.592 ms  41.459 ms
 9  e513-e-rci76-1-swice2.cern.ch (192.65.184.222) [AS513]  37.630 ms  96.029 ms  95.942 ms
10  l513-c-rftec-1-be8.cern.ch (192.16.166.129) [AS513]  44.459 ms  37.749 ms  37.539 ms
11  * * *
12  * * *
13  castorgrid02.cern.ch (128.142.175.73) [AS513]  370.623 ms  371.403 ms  371.326 ms
```

Alhora d'analitzar els resultats cal tenir en compte que el detall de les rutes (nom/ip dels encaminadors) pot canviar, les mostrades anteriorment corresponen al dia 13/03/07.

Davant d'una alarma de Warning el més probable és que els errors siguin intermitents, així que es recomana realitzar les comprovacions diverses vegades per a poder detectar la font del problema.

En el cas de que els traceroute sobre TCP mostrin que no és possible establir la ruta fins a:

- RedIRIS (x.rediris.es, [AS766]): revisar l'estatus de les connexions del router, de les interfícies al CERN i posar-se en contacte amb RedIRIS.
- GÉANT (geant2.net, [AS20965]): revisar l'estatus de les connexions del router, de les interfícies al CERN i posar-se en contacte amb RedIRIS i GÉANT
- CERN (x.cern.ch, [AS513]): revisar l'estatus de les connexions del router, de les interfícies al CERN i posar-se en contacte amb RedIRIS, GÉANT i el CERN

Al posar-se en contacte amb les diferents entitats pot ser útil enviar-los les gràfiques provistes per mrtg/cacti, així com els resultats dels traceroute anteriors i d'altres logs d'error disponibles (tests de SAM, etc).

Punts de contacte

- Status de les connexions del router local
  - mrtg.pic.es
  - cacti.pic.es
- RedIRIS, GÉANT, CERN
  - https://twiki.cern.ch/twiki/bin/view/LHCOPN/LHCopnOperations#AnchorPIC
  - Status de les interfícies al CERN: http://lhcopn.web.cern.ch/lhcopn/lhcopn-interfaces.html

Històric d'alertes

- Del 8/03/07 al 13/03/07 l'alarma s'activava alternativament d'estat warning a critical, els traceroute funcionaven, donant resultats intermitents un cop al CERN ([AS513]). Finalment hi havia un problema al CERN: una NIC de 10 Gbps d'un

dels dos encaminadors LHC-OPN del CERN donava errors de paritat, un cop s'ha reiniciat (7:30 13/03/07) s'ha tornat a la normalitat

11.2 Sensor de Nagios/Ganglia

A continuació es mostra l'script dissenyat pet a poder diagnosticar problemes alhora de realitzar connexions TCP (incidència 6).

L'script necessita ser executat amb permisos de root i utilitza nmap, s'ha comprovat el seu correcte funcionament amb Nmap versió 4.11

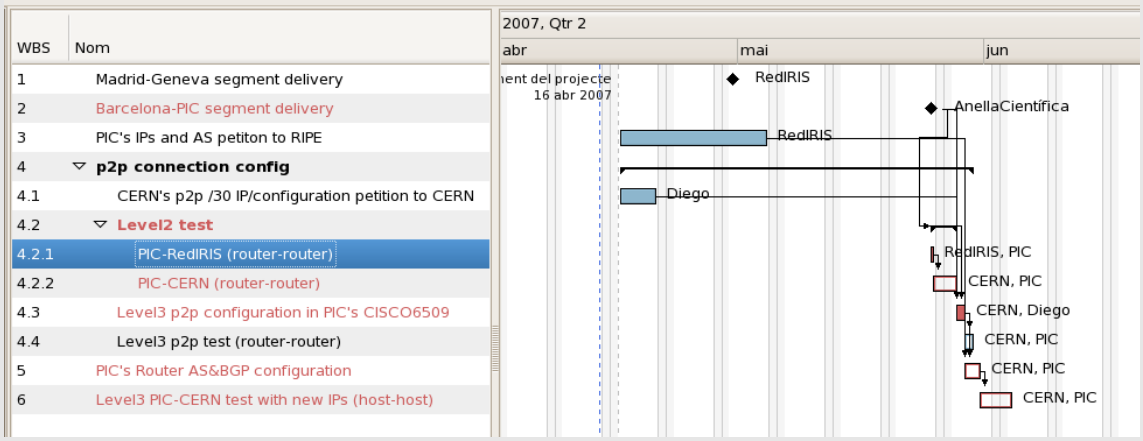
```
#!/bin/sh
#
#Se comprueba:
# *La exitencia del servidor mediante: respuesta a ping (ICMP echo request)
# *La habilidad de realizar conexiones con un servidor remoto en un puerto determinado
#
#Retorna el número de peticiones de conexión TCP fallidas sobre 10 y de peticiones
fallidas con descubrimiento PING (sobre 5)
#
#IMPORTANTE: cuando els descubrimiento PING no funciona la petici o  de conexión TCP no se
realiza. En el caso de que se indique el parámetro -c las connexiones a puertos cerrados
se cuentan como correctas si siguen el protocolo (se manda un SYN y se recibe un RST)
#
#Versión       Autor
#---------    ------
#   1           GBA
#   1.1         GBA
#
#***********************************************
#Se debe llamar como:
# AlarmaSYN.sh [-c] [-v] [-h] servidor puerto
#Retorna:
# TotalFallosSYN(/10 o /5 si ping no KO) FallosPING (/5)
#***********************************************
#Parámetros:
# -c: se realiza la prueba contando como buenas las conexiones a puertos cerrados (closed)
# -v: verbose
# -h: muestra la forma de uso
########################GENERATING/PARSING PARAMETERS###################################
verbose="0"
mode="open"
file=`date +SYNDetector%y%m%d%H.txt`

while getopts cvh o
do
 case "$o" in
    c)  mode="closed";;
    v)  verbose="yes";;
    h|?)  echo "Usage: $0  [-c] [-v] [-h] host port"
       echo "-c: se realiza la prueba contando como buenas las conexiones a puertos
cerrados (closed)"
       echo "-v: verbose"
       echo "-h: muestra la forma de uso"
       exit -1
       ;;
 esac
done

i=1
while [ $i -lt $OPTIND ]; do
       i=$(($i+1))
       shift
done

host=$1
port=$2
```

```
#######################################ACTIONS#############################################
echo "Descubrimiento con PING" > $file
for i in 1 2 3 4 5; do
    echo Intento nº$i >> $file
    nmap -sS $host -p $port -o $file --append_output > /dev/null 2>&1
done
fallosPING=`grep -c '0 hosts up' $file`

echo "Sin descubrimiento" >> $file
for i in 6 7 8 9 10; do
    echo Intento nº$i >> $file
    nmap -sS $host -p $port -o $file --append_output -P0 > /dev/null 2>&1
done
#######################################OUTPUT
GENERATION#################################
if [ "$mode" = "closed" ]; then
       fallos=$((10 - `grep -c ${mode} $file` - `grep -c 'open' $file`))
else
       fallos=$((10 - `grep -c ${mode} $file`))
fi

if [ "$verbose" = "yes" ]; then
       cat $file
fi

rm -f $file

if [ "$fallosPING" = "5" ]; then
       pingKO="/5"
else
       pingKO="/10"
fi
#################################OUTPUT GANGLIA#######################################
#echo Fallos puros de connexión SYN: $(($fallos - $fallosPING))$pingKO "Debugging: "
$fallos"/10" $fallosPING"/5"
#exit $(($fallos - $fallosPING))
#################################OUTPUT NAGIOS########################################
echo Fallos de connexión SYN: $(($fallos - $fallosPING))$pingKO "   Fallos descubrimiento
PING: " $fallosPING"/5"

if [ "$(($fallos - $fallosPING))" -lt "2" ]; then
       outputNagios="0"
else
       if [ "$(($fallos - $fallosPING))" -lt "5" ]; then
              outputNagios="1"
       else
              outputNagios="2"
       fi
fi
exit $outputNagios
#################################END SCRIPT#########################################
```

12 Annex G: pla d'actuació per al desplegament del circuit dedicat de 10 Gbps

Company: PIC
Manager: Gerard B.A.
Start: April 16, 2007
Finish: June 4, 2007
Report Date: April 13, 2007

Gantt Chart



Tasks

WBS	Name	Start	Finish	Work	Notes
1	Madrid-Geneva segment delivery	Apr 30	Apr 30		
2	Barcelona-PIC segment delivery	May 25	May 25		
3	PIC's IPs and AS petiton to RIPE	Apr 16	May 4	15d	http://www.ripe.net/rs/ipv4/ http://www.ripe.net/rs/as/
4	p2p connection config	Apr 16	May 30	13d	
4.1	CERN's p2p /30 IP/configuration petition to CERN	Apr 16	Apr 20	5d	
4.2	Level2 test	May 25	May 28	4d	ARP or similar
4.2.1	PIC-RedIRIS (router-router)	May 25	May 25	2d	
4.2.2	PIC-CERN (router-router)	May 25	May 28	2d	
4.3	Level3 p2p configuration in PIC's CISCO6509	May 28	May 29	2d	
4.4	Level3 p2p test (router-router)	May 29	May 30	2d	Router-Router ping with CERN's IPs
5	PIC's Router AS&BGP configuration	May 29	May 31	4d	

6	Level3 PIC-CERN test with new IPs (host-host)	May 31	Jun 4	4d	host to host

Resources

Name	Short name	Group
Anella Científica & Al-Pi (fibre deployer)	AnellaCientífica	
CERN LHC-OPN Team	CERN	
PIC's Network Administrator (Diego Dávila)	Diego	PIC
GÉANT	RedIRIS	
PIC LHC-OPN Team (Gerard/Diego)	PIC	PIC

13 Annex H: Detalls d'implementació i proves del sistema de certificació pel circuit dedicat de 10 Gbps

Per a la execució en paral·lel d'aplicacions en tots els servidors del PIC membres del sistema de certificació pel circuit dedicat de 10 Gbps s'ha utilitzat la utilitat vxargs (veure secció 13.1). A la secció 13.2 hi ha les proves del sistema de certificació, realitzades sobre la connexió a la xarxa LHC-OPN sobre el circuit d'1 Gbps. La configuració definitiva del sistema de certificació està detallada a la secció 13.3.

13.1 Configuració i us de vxargs

Els dotze servidors (lhcopn01 a lhcopn12) membres del sisetma de certifcació s'han configurat per a poder ordenar l'execució de comandes des de lhcopn01 mitjançant l'ordre

*exec_vxargs.sh hostsOPN.txt "**comanda a executar en paral·lel**"*

On *hostOPN.txt* és un fitxer amb les adreces dels 12 servidors, tal i com es mostra a la figura 13.1.1.

```
[root@lhcopn01 vxargs]# cat hostsOPN.txt
193.145.217.2 #lhcopn01
193.145.217.3 #lhcopn02
193.145.217.4 #lhcopn03
193.145.217.5 #lhcopn04
193.145.217.6 #lhcopn05
193.145.217.7 #lhcopn06
193.145.217.8 #lhcopn07
193.145.217.9 #lhcopn08
193.145.217.10 #lhcopn09
193.145.217.11 #lhcopn10
193.145.217.12 #lhcopn11
193.145.217.13 #lhcopn12
```

Figura 13.1.1: fitxer hostsOPN.txt utilitzat per vxargs i que conté les adreces IP de la interfícies LHC-OPN dels host lhcopn01 al lhcopn12

Per tal de poder executar comandes des de lhcopn01 en tots els servidors lhcopnXX cal seguir el següent procediment:

1. Habilitar la versió 2 de ssh a lhcopn01 (*/etc/ssh/ssh_config*)
2. Generar la clau ssh en lhcopn01 mitjançant l'execució de ssh-keygen com a root, sense indicar password.
3. A cada servidor lhcopn (lhcopn01 a lhcopn12)

   Copiar la clau pública de root@lhcopn01 com a clau autoritzada (*scp /root/.ssh/id_rsa.pub lhcopnXX:/root/.ssh/authorized_keys*)

13.2 Proves del sistema de certificació pel circuit dedicat de 10 Gbps

Per tal de comprovar el funcionament de vxargs s'intenta transferir des dels 12 servidors de forma simultània cap a wl-gerard.pic.es, un ordinador d'escriptori des d'on s'està executant el servidor iperf. en la figura 13.2.1 es pot observar l'execució de la comanda des de lhcopn01, a la figura 13.2.2 es mostra la resposta en el servidor iperf de l'ordinador wl-gerard.

```
[root@lhcopn01 vxargs]# ./exec_vxargs.sh hostsOPN.txt "iperf -c wl-gerard.pic.es"
exit code 0: 12 job(s)
total number of jobs: 12
```

Figura 13.2.1: execució de la comanda des de lhcopn01. Com es pot observar tots els processos s'executen correctament, ja que retornen 0.

```
[gerard@wl-gerard ~]$ iperf -s
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 85.3 KByte (default)
------------------------------------------------------------
[  4] local 193.146.196.233 port 5001 connected with 193.146.197.158 port 32802
[  5] local 193.146.196.233 port 5001 connected with 193.146.197.206 port 33010
[  6] local 193.146.196.233 port 5001 connected with 193.146.197.155 port 32780
[  7] local 193.146.196.233 port 5001 connected with 193.146.197.210 port 32909
[  8] local 193.146.196.233 port 5001 connected with 193.146.197.207 port 33001
[  9] local 193.146.196.233 port 5001 connected with 193.146.197.211 port 32909
[ 10] local 193.146.196.233 port 5001 connected with 193.146.197.159 port 33014
[ 11] local 193.146.196.233 port 5001 connected with 193.146.197.153 port 33072
[ 12] local 193.146.196.233 port 5001 connected with 193.146.197.209 port 32998
[ 13] local 193.146.196.233 port 5001 connected with 193.146.197.154 port 32783
[ 14] local 193.146.196.233 port 5001 connected with 193.146.197.208 port 32998
[ 15] local 193.146.196.233 port 5001 connected with 193.146.197.212 port 32907
[  9]  0.0-10.1 sec  13.6 MBytes  11.3 Mbits/sec
[ 10]  0.0-10.1 sec  11.3 MBytes  9.42 Mbits/sec
[  8]  0.0-10.2 sec  14.2 MBytes  11.7 Mbits/sec
[  6]  0.0-10.3 sec  11.4 MBytes  9.31 Mbits/sec
[ 12]  0.0- 7.2 sec  6.13 MBytes  7.13 Mbits/sec
[  7]  0.0-10.3 sec  13.3 MBytes  10.8 Mbits/sec
[  5]  0.0-10.6 sec  12.5 MBytes  9.90 Mbits/sec
[  4]  0.0-11.1 sec  10.5 MBytes  7.94 Mbits/sec
[ 13]  0.0- 9.9 sec  20.4 MBytes  17.3 Mbits/sec
[ 15]  0.0- 4.0 sec  9.85 MBytes  20.6 Mbits/sec
[ 14]  0.0-10.2 sec  14.5 MBytes  11.9 Mbits/sec
[ 11]  0.0-18.0 sec  16.0 KBytes  7.28 Kbits/sec
```

Figura 13.2.2: resposta de dotze transferències iperf en el servidor iperf de l'ordinador d'escriptori wl-gerard.pic.es. Es pot observar que, degut al poc ample de banda disponible en wl-gerard.pic.es (100Mbps) en comparació amb l'ample de banda del sistema de certificació (12 Gbps), les tranferències s'executen amb moltes desigualtats, observant variacions en temps de tranferència i velocitat.

Un cop vist que el funcionament de vxargs és correcte s'han realitzat dues proves per a comprovar la capacitat de generació de trànsit del sistema de certificació. Les proves s'han dut a terme simulant l'arquitectura de certificació acordada amb el CERN, és a dir, intentant generar trànsit des de N servidors a 1.

En aquest cas no es disposa d'un servidor amb una NIC de 10 Gbps, és de 1 Gbps, per tant la monitorització s'ha de realitzar des de l'encaminador cisco 6509.

La prova es realitza generant trànsit UDP des de la interfície LHC-OPN dels servidors lhcopnXX (excepte lhcopn01) cap a la interfície no-LHC-OPN del servidor lhcopn01, monitoritzant des del cisco6509 el trànsit entre la VLAN222 (intefície LHC-OPN dels servidors) i la VLAN100 (interfície no-LHC-OPN de lhcopn01).

Per a la realització de la prova s'ha executat "*iperf -su*" (servidor UDP de iperf) en un terminal de lhcopn01 i la crida dels clients "*./exec_vxargs.sh hostsOPN.txt "ifdown eth0; route add default gw 193.145.217.1; iperf -c 193.146.197.153 -b1g -t 900""* en un segon terminal. Com es pot observar en la comanda, abans de cridar al client iperf s'apaga la interfície no-LHC-OPN (eth1) i es defineixz la porta d'enllaç per defecte via la interfície LHC-OPN de tots els host excepte lhcopn01.

La prova es realitza dues vegades, una amb 8 (figures 13.2.3 i 13.2.4) i l'altre amb 11 (figures 13.2.5 i 13.2.6) servidors. En ambdós casos l'ús de CPU en la placa supervisora del cisco6509 es manté en nivells estàndard (entre el 6 i el 8%)

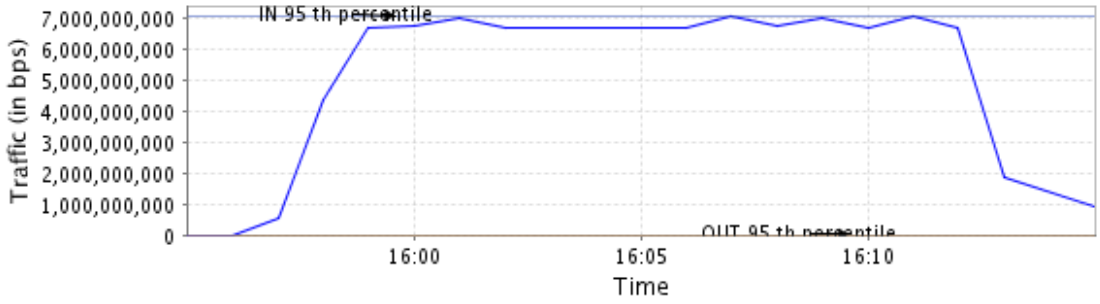


Figura 13.2.3: es mostra la VLAN22 amb el trànsit genrat pel test amb 8 servidors, generant un trànsit total sostingut d'uns 7 Gbps durant els 900 segons de la prova.

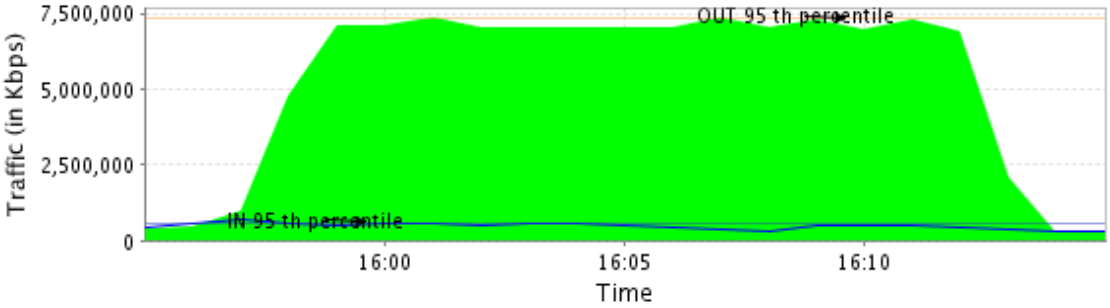


Figura 13.2.4: es mostra la VLAN100 amb el trànsit normal (uns 500Mbps) i el test amb 8 servidors (uns 7 Gbps), generant un trànsit total sostingut d'uns 7,5 Gbps durant els 900 segons de la prova. Des de les estadístiques del cisco6509 no es detecta pèrdua de paquets entre les intefícies vlan100 i vlan222 (si que se'n detecta en la intefície no-LHC-OPN del servidor lhcopn01).

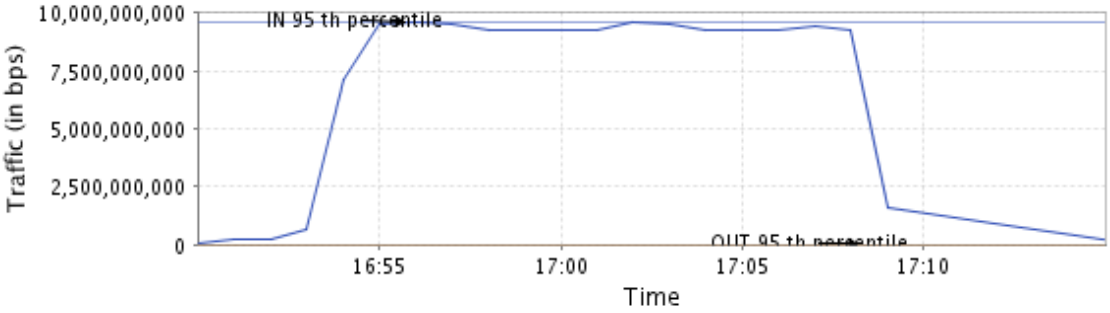


Figura 13.2.5: es mostra la VLAN22 amb el trànsit genrat pel test amb 10 servidors, generant un trànsit total sostingut d'uns 9,5 Gbps durant els 900 segons de la prova.

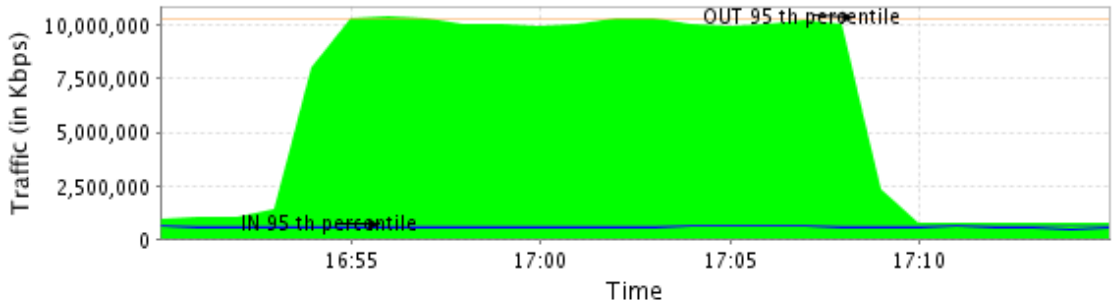


Figura 13.2.6: es mostra la VLAN100 amb el trànsit normal (uns 500Mbps) i el test amb 8 servidors (uns 9,5 Gbps), generant un trànsit total sostingut una mica per sobre els 10 Gbps durant els 900 segons de la prova. Des de les estadístiques del cisco6509 no es detecta pèrdua de paquets entre les intefícies vlan100 i vlan222 (si que se'n detecta en la intefície no-LHC-OPN del servidor lhcopn01).

La prova definitiva per a comprovar la validesa del sistema de certificació pel circuit dedicat de 10 Gbps s'ha realitzat amb el servidor que el CERN ha preparat per a la realització de la prova amb una NIC de 10 Gbps (*hufsa.cern.ch)*. En aquest cas la prova s'ha realitzat sobre el circuit d'1 Gbps, realitzant transferències UDP amb ample de banda limitat a 1Mbps, per evitar interferir amb la operació normal de la connexió (que es troba en producció). En la figura 13.2.7 es pot observar el resultat de la prova des del servidor iperf executat a hufsa.cern.ch.

```
pictest@hufsa:~$ iperf -su
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 1470 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 193.145.217.2 port 32985
[  4] local 128.142.208.6 port 5001 connected with 193.145.217.4 port 32779
[  5] local 128.142.208.6 port 5001 connected with 193.146.197.206 port 32788
[  6] local 128.142.208.6 port 5001 connected with 193.146.197.159 port 32786
[  7] local 128.142.208.6 port 5001 connected with 193.146.197.209 port 32786
[  8] local 128.142.208.6 port 5001 connected with 193.146.197.210 port 32785
[  9] local 128.142.208.6 port 5001 connected with 193.145.217.3 port 32775
[ 10] local 128.142.208.6 port 5001 connected with 193.146.197.207 port 32781
[ 11] local 128.142.208.6 port 5001 connected with 193.146.197.158 port 32790
[ 12] local 128.142.208.6 port 5001 connected with 193.146.197.211 port 32788
[ 13] local 128.142.208.6 port 5001 connected with 193.146.197.208 port 32789
[ 14] local 128.142.208.6 port 5001 connected with 193.146.197.212 port 32788
[  3]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.066 ms    0/  852 (0%)
[  4]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.022 ms    0/  852 (0%)
[  5]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.052 ms    0/  852 (0%)
[  6]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.045 ms    0/  852 (0%)
[  7]  0.0-10.0 sec  1.19 MBytes  1.00 Mbits/sec  0.042 ms    0/  852 (0%)
[  8]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.050 ms    0/  852 (0%)
[  9]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.044 ms    0/  852 (0%)
[ 10]  0.0-10.0 sec  1.19 MBytes  1.00 Mbits/sec  0.046 ms    0/  852 (0%)
[ 11]  0.0-10.0 sec  1.19 MBytes  1.00 Mbits/sec  0.051 ms    0/  852 (0%)
[ 12]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.052 ms    0/  852 (0%)
[ 13]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.029 ms    0/  852 (0%)
[ 14]  0.0-10.0 sec  1.19 MBytes  1000 Kbits/sec  0.029 ms    0/  852 (0%)
```

Figura 13.2.7: es pot observar com la prova s'ha realitzat correctament, generant 12 flux de dades d'1 Mbps sobre UDP, amb una duració de 10 segons i una pèrdua del 0%.

Un cop demostrada la capacitat de generació de trànsit del sistema de certificació i havent provat el sistema de certificació real a baixa velocitat es pot considerar vàlid el sistema de certificació implementat.

13.3 Configuració del sistema de certificació

Per a la realització de les proves de certificació s'han utilitzat adreces IP del CERN, el canvi d'IPs en els servidors del sistema de certificació s'ha realitzat mitjançant les comandes mostrades en la figura 13.3.1.

```
ifconfig eth1 192.16.166.24X/28; #on X=Y+1 i Y surt de lhcopn0Y
route del default gw gw; #Eliminem porta d'enllaç via xarxa PIC
route del default gw 193.145.217.1; #Eliminem porta d'enllaç via xarxa LHC-OPN del PIC
route add default gw 192.16.166.241 #Afegim nova porta d'enllaç via adreces cedides pel
CERN
```

Figura 13.3.1: comandes per al canvi de les IPs i les rutes en els servidors del sistema de certificació.

Al PIC els servidors utilitzats pel sistema de certificació són Dell PowerEdge750, amb SLC3 (Scientific Linux Cern 3) i *Kernel* versió 2.4, és a dir, sense *autotunning* de la mida de finestra de TCP. Per tal de millorar-ne el rendiment, a aquests servidors se'ls ha ampliat la mida màxima de la finestra, tal i com es va fer per a les proves i la certificació del circuit dedicat d'1 Gbps (més informació a l'annex D). Per tal de realitzar els canvis simultàniament a tots els servidors del sistema de certificació s'ha utilitzat *vxargs*, tal i com es mostra en la figura 13.3.2..

```
[root@lhcopn01 vxargs]# ./exec_vxargs.sh hostsPIC.txt 'echo "4096 87380 128388607" >
/proc/sys/net/ipv4/tcp_rmem'
exit code 0: 12 job(s)
total number of jobs: 12
[root@lhcopn01 vxargs]# ./exec_vxargs.sh hostsPIC.txt 'echo 128388607 >
/proc/sys/net/core/wmem_max'
exit code 0: 12 job(s)
total number of jobs: 12
[root@lhcopn01 vxargs]# ./exec_vxargs.sh hostsPIC.txt 'echo "4096 65530 128388607" >
/proc/sys/net/ipv4/tcp_wmem '
exit code 0: 12 job(s)
total number of jobs: 12
[root@lhcopn01 vxargs]# ./exec_vxargs.sh hostsPIC.txt 'echo 128388607 >
/proc/sys/net/core/rmem_max'
exit code 0: 12 job(s)
total number of jobs: 12
```

Figura 13.3.2: comandes per al canvi dels paràmetres de TCP al kernel 2.4 dels servidors del sistema de certificació.

Al CERN el servidor utilitzat per a la realització de les proves (*hufsa.cern.ch*) disposa d'un Kernel versió 2.6, amb *autotunning* i amb la mida de finestra màxima suficientment ampliada.

Pel que fa al router del PIC, la configuració que permet la comunicació mitjançant IPs del CERN sobre la connexió punt a punt es troba en la figura 13.3.3 (extracte de la configuració completa).

```
!
vlan 222
 name LHC-OPN-test
!
vlan 287
 name LHCOPN
!
!
interface GigabitEthernet4/1
 description Connexio VLAN222 LHC-OPN
 switchport
 switchport access vlan 222
 switchport mode access
 no ip address
 logging event link-status
 no cdp enable
 spanning-tree portfast
!
interface GigabitEthernet4/2
```

```
 description Connexio VLAN222 LHC-OPN
 switchport
 switchport access vlan 222
 switchport mode access
 no ip address
 logging event link-status
 no cdp enable
 spanning-tree portfast
!
.............. !Les 12 interfícies del sistema de certificació
!
interface TenGigabitEthernet8/1
 description 10 Gbps LHC-OPN
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 287
 switchport trunk pruning vlan none
 switchport mode trunk
 mtu 9216
 no ip address
 logging event link-status
 no cdp enable
!
!
interface Vlan222
 description LHC-OPN vlan222
 mtu 9216
 ip address 192.16.166.241 255.255.255.240
 ip helper-address 193.146.197.16
 no ip redirects
 ip route-cache flow
 ip policy route-map LHC-OPN
 logging event link-status
!
interface Vlan287
 description 10 Gbps VLAN 287
 mtu 9216
 ip address 192.16.166.58 255.255.255.252
 no ip redirects
 ip route-cache flow
 logging event link-status
!
access-list 188 permit ip 192.16.166.240 0.0.0.15 any
access-list 188 permit icmp 192.16.166.240 0.0.0.15 any
access-list 188 deny   ip any any
access-list 188 deny   icmp any a
!
route-map LHC-OPN permit 10
 match ip address 188
 set ip next-hop 192.16.166.57
!
```

Figura 13.3.3: extracte de la configuració de l'encaminador cisco6509 amb la connexió a la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps operatiu.

14 Annex I: sensor de Nagios i procediment per a la monitorització de la connectivitat sobre la xarxa LHC-OPN

Per a la monitorització de la xarxa LHC-OPN sobre el circuit dedicat de 10 Gbps s'ha dissenyat un sensor (veure figura 14.0.1) de Nagios que comprova la connectivitat amb el servei SRM dels diferents Tier1 i amb el Tier0 (el CERN).

El que realment fa el sensor dissenyat és comprovar l'estat d'un port en un servidor remot. En el cas de la monitorització de la xarxa LHC-OPN els servidors escollits són els servidors SRM, encarregats de controlar les tranferències de dades entre centres, de tots els centres integrants de la xarxa LHC-OPN.

També s'ha redactat un procediment d'actuació a seguir en cas de que el sensor generi una alarma. El procediment d'actuació i alguns comentaris del codi del sensor s'han realitzat en castellà per tal de facilitar l'intercanvi d'informació dins el PIC.

```
#!/bin/sh
#
#Comprova els N host indicats a continuació en el port definit.
#Els fallos de connexió es valoren segons el pes assignat a cada prova (nºfallos*pes).
#Finalment es calcula un total sobre 100 (%).
#
#La sortida a NAGIOS és:
#falla 0% ->OK
#falla menys del 30% -> OK amb missatge
#falla més del 30% i menys 75%-> WARNING
#falla més del 75% -> CRITICAL
#TOTES les proves fallen -> CRITICAL
#
################################################################################
################################################################################

MaxOK="150"
MaxWarning="500"

##################DATABASE##################
#hosts[n]="dns port pes"
hosts[0]="srm.cern.ch 8443 20"
hosts[1]="ccsrm.in2p3.fr 8443 9"
hosts[2]="dcache.gridpp.rl.ac.uk 8443 9"
hosts[3]="castorsrm.cr.cnaf.infn.it  8443 9"
hosts[4]="srm.triumf.ca  8443 9"
hosts[5]="gridka-dCache.fzk.de  8443 9"
hosts[6]="srm.grid.sara.nl 8443 9"
hosts[7]="srm.ndgf.org  8443 9"
hosts[8]="dcsrm.usatlas.bnl.gov  8443 9"
hosts[9]="cmssrm.fnal.gov  8443 9"

numhosts=$((${#hosts[*]}-1))
################################################################################
################################################################################

check_tcp_syn()
{
        #Se comprueba:
        # *La exitencia del host mediante: respuesta a ping (ICMP echo request)
```

```
        # *La habilidad de realizar conexiones con un host remoto en un puerto determinado
        #
        #Retorna el número de peticiones de conexión TCP fallidas sobre 10 y de peticiones
fallidas con descubrimiento PING (sobre 5)
        #
        #IMPORTANTE: cuando els descubrimiento PING no funciona la petición de conexión
TCP no se realiza. En el caso de que se indique el parámetro -c las connexiones a puertos
cerrados se cuentan como correctas si siguen el protocolo (se manda un SYN y se recibe un
RST)
        #
        #Versión       Autor
        #---------    ------
        #   1           GBA
        #   1.1         GBA
        #
        #***********************************************
        #Se debe llamar como:
        # AlarmaSYN.sh [-c] [-v] [-h] host puerto
        #Retorna:
        # TotalFallosSYN(/10)
        #***********************************************
        #Parámetros:
        # -c: se realiza la prueba contando como buenas las conexiones a puertos cerrados
(closed)
        # -v: verbose
        # -h: muestra la forma de uso

        ##########################GENERATING/PARSING PARAMETERS##############################
        verbose="0"
        mode="open"
        file=`date +SYNDetector%y%m%d%H.txt`

        while getopts cvh o
        do
         case "$o" in
            c)  mode="closed";;
            v)  verbose="yes";;
            h|?)  echo "Usage: $0  [-c] [-v] [-h] host port"
                  echo "-c: se realiza la prueba contando como buenas las conexiones a
puertos cerrados (closed)"
                  echo "-v: verbose"
                  echo "-h: muestra la forma de uso"
                  exit -1
                  ;;
         esac
        done

        i=1
        while [ $i -lt $OPTIND ]; do
                i=$(($i+1))
                shift
        done

        host=$1
        port=$2

        ########################################ACTIONS#####################################

        for i in `seq 1 10`; do
            echo Intento nº$i >> $file
            nmap -sS $host -p $port -o $file --append_output -P0 --host_timeout 3000 --
initial_rtt_timeout 700 > /dev/null 2>&1
        done

        ########################################OUTPUT GENERATION ########################

        if [ "$mode" = "closed" ]; then
                fallos=$((10 - `grep -c ${mode} $file` - `grep -c 'open' $file`))
        else
                fallos=$((10 - `grep -c ${mode} $file`))
        fi

        if [ "$verbose" = "yes" ]; then
                cat $file
        fi
```

```
        rm -f $file

        #echo Fallos de connexion SYN con ${host}: $fallos "/10"

        return $fallos
}


##################PONDERACIO###################

fallentots=1 #1=si, 0=no
totalfallos=0
totalfalloponderat=0
hostKO=""
for i in `seq 0 $numhosts`; do
        host=`echo ${hosts[${i}]} | awk '{ print $1}'`
        port=`echo ${hosts[${i}]} | awk '{ print $2}'`
        pes=`echo ${hosts[${i}]} | awk '{ print $3}'`

        check_tcp_syn $host $port
        fallos=$?
        totalfallos=$(( $totalfallos + $fallos ))

        falloponderat=$(( ${fallos} * ${pes} ))
        totalfalloponderat=$(( $totalfalloponderat + $falloponderat ))

        if [ "$fallos" -gt "0" ]; then #si falla menys del 20%
                hostKO="${host}:${port}-${fallos}fallos, ${hostKO}"
        else
                fallentots=0 #no falla i, per tant, no fallen tots
        fi
done


###################################OUTPUT NAGIOS#########################################

if [ "$fallentots" = "1" ]; then
        outputNagios="2" #totes les proves fallen més del 20% ->CRITICAL
        echo Todas las pruebas de conexion TCP fallan MAS del 20%: Fallos de conexion TCP
$(($totalfalloponderat / 10))%. Desglose de fallos "("$totalfallos")" con $hostKO
else
        if [ "$totalfalloponderat" -lt $MaxOK ]; then #falla menys del 30% -> OK amb
missatge
                outputNagios="0"
                if [ "$totalfalloponderat" -eq "0" ]; then
                        echo OK: Ningun fallo de conexion
                else
                        echo Fallos de conexion TCP $(($totalfalloponderat / 10))%.
Desglose de fallos "("$totalfallos")" con $hostKO
                fi
        else
                if [ "$totalfalloponderat" -lt $MaxWarning ]; then #falla més del 30% i
menys 75%-> WARNING
                        outputNagios="1"
                        echo Fallos de conexion TCP $(($totalfalloponderat / 10))%.
Desglose de fallos "("$totalfallos")" con $hostKO
                else
                        outputNagios="2" #falla més del 75% -> CRITICAL
                        echo Fallos de conexion TCP $(($totalfalloponderat / 10))%.
Desglose de fallos "("$totalfallos")" con$hostKO
                fi
        fi
fi
exit $outputNagios
```

Figura 14.0.1: codi del sensor de Nagios per a la monitorització de la connectivitat amb el servidors SRM del centres integrants de la xarxa LHC-OPN.

Procediment d'actuació

*Esta alarma significa que hay un problema de conexión TCP con diversos **host**, atualmente los servidores de SRM. Esta alarma puntua según los fallos encontrados en 10 intentos de conexión a*

distintos host.

El funcionamiento/salida de la alarma es

1) **#falla 0% ->OK** No es necesario tomar ninguna acción

2) **#falla menos del 20% -> OK con mensaje** Probablemente los host/servicios indicados en el mensaje están en periodo de mantenimiento, consultar el estado de los servicios implicados en http://cic.gridops.org/

3) **#falla mas del 20% y menos 50%-> WARNING** Consultar el estado de los servicios fallidos en http://cic.gridops.org/. Si estos deben estar en producción reportar el incidente según el procedimiento estándard

4) **#falla mas del 50% -> CRITICAL** Consultar el estado de los servicios fallidos en http://cic.gridops.org/. Si estos deben estar en producción reportar el incidente según el procedimiento estándard. Consultar el estado del router en netflow.pic.es:8080 y mirar el estado de los servicios LHC-OPN del PIC http://www.rediris.es/red/stats/EB-Barcelona1/lhc.html?

5) **#TODAS las pruebas fallan -> CRITICAL** Significa que todos los host fallan alguna vez intentando establecer conexión con ellos, probablemente se trata de un fallo del host que ejecuta la alarma o del router/conexión local. Consultar el estado del router en netflow.pic.es:8080 y mirar el estado de los servicios LHC-OPN del PIC

NOTA IMPORTANTE: Es posible que a veces Nagios reporte 'Service Check Timed Out', generalmente esto significa que hay problemas de conectividad graves.

Cuando se detecte un fallo se seguirá el procedimiento estándard para la resolución de fallidas de red: https://www.wiki.pic.es/index.php/Procediment_en_cas_de_problemes_de_xarxa

Respcto a la puntuación cabe destacar que ha sido diseñada para que tenga el siguiente comportamiento:

- si falla el CERN->WARNING
- Si falla un único T1 -> OK con mensaje que avisa del fallo
- Si fallan dos T1 -> de forma simultània da WARNING y si falla mas de 5 da CRITICAL.
- En el caso de que no fallen todas las conexiones peri si algunas lo más probable es que el estado varie entre WARNING y CRITICAL constantemente.

En caso de no encontrar una solución, para obtener mas información se puede consultar aquí (link al procediment de l'annx F).

Historico de alertas

- Del 8/03/07 al 13/03/07 la alarma se activaba alternativamente de warning a critical, los traceroute funcionaban, dando resultados intermitentes al alcanzar al CERN ([AS513]). Finalmente havia un problema en el CERN: una NIC de 10 Gbps de uno de los dos routers LHC-OPN del CERN daba errores de paridad, en cuanto se ha reiniciado (7:30 13/03/07) se ha vuelto a la normalidad.
- El 2/04/07 se ha activado a las 8:00 debido a un scheduled downtime del CERN en el que se han apagado los servicios castorgrid. Se puede ber como fallan las conexionse TCP al hacer un *nmap -sS castorgrid.cern.ch -p2811*
- Se ha modificado el SCRIPT para que compruebe tanto el T0 como algunos T1, ahora los errores serán más explicativos.
- El 07/05/07 ha saltado alarma CRITICAL con 'Service Check Timed Out' debido a un fallo en el DNS primario del PIC. Ha dado timeout debido a que el servidor DNS tardaba mucho en devolver un mensaje de error y esto ha relantecido enormemente todos los test del sensor.
- El 21/05/07 ha saltado alarma CRITICAL debido a un error con el DNS (si no puede resolver IP no puede contactar). Gracias a correcciones en el script ya no da "Service Check Timed Out" sino el error real de no-conexión: "Todas las pruebas de conexion TCP fallan MAS del 20%: Fallos de conexion TCP 101%. Desglose de fallos (100) con cmssrm.fnal.gov:8443-10fallos, dcsrm.usatlas.bnl.gov:8443-10fallos, srm.ndgf.org:8443-10fallos, srm.grid.sara.nl:8443-10fallos, gridka-dCache.fzk.de:8443-10fallos, srm.triumf.ca:8443-10fallos, cast"

15 Annex J: diagnosis del problema de connectivitat PIC-CERN sobre el circuit dedicat de 10 Gbps durant la certificació

Tal i com s'explica en la incidència 5 de la subsecció 4.3.3, durant les proves de certificació amb transferències bidireccionals i amples de banda de ~10 Gbps la connexió punt a punt PIC-CERN es perdia.

Per tal de realitzar les proves esmentades, en els encaminadors de la connexió p2p PIC-CERN s'han creat rutes estàtiques per tal que el trànsit dirigit a l'adreça IP 192.16.166.200 es reenvii per la connexió p2p d'un extrem a l'altre creant un bucle, que dura fins a l'extinció del TTL (*Time To Live*) dels paquets.

Per tal de *debugar* el circuit dedicat i localitzar el motiu de la incidència s'han realitzat una bateria de proves (veure 15.1, en anglès per a facilitar-ne la difusió) que ha ajudat a entendre el problema detectat pels tècnics de RedIRIS en el PoP de RedIRIS10 a Barcelona (un Nortel ERS 8010).

Donat que les transferències "estàndard" mostren un comportament estrany i que, just al iniciar les transferències a la IP amb la que es genera el bucle (192.16.166.200) es produeix el tall de connectivitat, es dedueix que el problema es deu a alguna opció de protecció davant de bucles de l'equip Nortel de RedIRIS.

Un cop notificada tota la informació a RedIRIS es resta en espera de que la incidència sigui reparada per tal de poder finalitzar les proves de certificació amb la prova de tranferència bidireccional d'alta velocitat.

15.1 Bateria de proves per a la diagnosi

Testing 192.16.166.240/28 -> hufsa.cern.ch

UDP 100MBps*12 - p2p ping: OK

```
pictest@hufsa:~$ iperf -su -i20 -l9000
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40092
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[  3]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.032 ms    0/138890 (0%)
[  4]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.033 ms    0/138890 (0%)
[  5]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.034 ms    0/138890 (0%)
```

```
[  6]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.015 ms    0/138890 (0%)
[  7]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.022 ms    0/138890 (0%)
[  8]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.073 ms    0/138890 (0%)
[  9]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.011 ms    0/138890 (0%)
[ 10]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.045 ms    0/138890 (0%)
[ 11]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.035 ms    0/138890 (0%)
[ 12]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.014 ms    0/138890 (0%)
[ 13]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.014 ms    0/138890 (0%)
[ 14]  0.0-100.0 sec  1.16 GBytes    100 Mbits/sec  0.008 ms    0/138890 (0%)
```

UDP 150MBps*12 - p2p ping: OK

```
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40209
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
[  3]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.105 ms    0/125002 (0%)
[  4]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.078 ms    0/125002 (0%)
[  5]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.043 ms    0/125002 (0%)
[  6]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.048 ms    0/124906 (0%)
[  7]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.018 ms    0/125002 (0%)
[  8]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.046 ms    0/125002 (0%)
[  9]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.043 ms    0/125002 (0%)
[ 10]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.060 ms    0/125002 (0%)
[ 11]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.012 ms    0/125002 (0%)
[ 12]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.007 ms    0/125002 (0%)
[ 13]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.005 ms    0/125002 (0%)
[ 14]  0.0-60.0 sec  1.05 GBytes    150 Mbits/sec  0.007 ms    0/125002 (0%)
```

UDP 300MBps*12 - p2p ping: OK

```
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40270
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[  3]  0.0-60.0 sec  2.01 GBytes    288 Mbits/sec  0.110 ms 9541/249636 (3.8%)
[  4]  0.0-60.0 sec  2.01 GBytes    287 Mbits/sec  0.200 ms 10397/250002 (4.2%)
[  5]  0.0-60.0 sec  2.00 GBytes    286 Mbits/sec  0.365 ms 11467/250002 (4.6%)
[  6]  0.0-60.0 sec  2.01 GBytes    287 Mbits/sec  0.272 ms 10493/250002 (4.2%)
[  7]  0.0-60.0 sec  2.01 GBytes    288 Mbits/sec  0.168 ms 10096/250002 (4%)
[  8]  0.0-60.0 sec  2.01 GBytes    287 Mbits/sec  0.088 ms 10765/250002 (4.3%)
[  9]  0.0-60.0 sec  2.01 GBytes    287 Mbits/sec  0.030 ms 10638/250002 (4.3%)
[ 10]  0.0-60.0 sec  2.01 GBytes    287 Mbits/sec  0.069 ms 10494/250002 (4.2%)
[ 11]  0.0-60.0 sec  2.01 GBytes    288 Mbits/sec  0.058 ms 10020/250002 (4%)
[ 12]  0.0-60.0 sec  2.00 GBytes    287 Mbits/sec  0.021 ms 10877/250002 (4.4%)
[ 13]  0.0-60.0 sec  2.01 GBytes    288 Mbits/sec  0.018 ms 9992/250002 (4%)
[ 14]  0.0-60.0 sec  2.01 GBytes    288 Mbits/sec  0.003 ms 10218/250002 (4.1%)
```

UDP 600MBps*12 - p2p ping: OK

```
------------------------------------------------------------
Server listening on UDP port 5001
```

```
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40390
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[  3]  0.0-60.0 sec  1.79 GBytes    256 Mbits/sec  0.710 ms 285174/498631 (57%)
[  5]  0.0-60.0 sec  1.81 GBytes    259 Mbits/sec  0.267 ms 284245/500002 (57%)
[  6]  0.0-60.0 sec  1.79 GBytes    256 Mbits/sec  0.026 ms 286923/500002 (57%)
[  7]  0.0-60.0 sec  1.81 GBytes    260 Mbits/sec  0.188 ms 283601/500002 (57%)
[  8]  0.0-60.0 sec  1.82 GBytes    260 Mbits/sec  0.225 ms 282974/500002 (57%)
[  9]  0.0-60.0 sec  1.81 GBytes    259 Mbits/sec  0.067 ms 283663/500002 (57%)
[ 10]  0.0-60.0 sec  1.81 GBytes    260 Mbits/sec  0.058 ms 283483/500002 (57%)
[ 11]  0.0-60.0 sec  1.80 GBytes    257 Mbits/sec  0.097 ms 285535/500002 (57%)
[ 12]  0.0-60.0 sec  1.81 GBytes    259 Mbits/sec  0.024 ms 284023/500002 (57%)
[ 13]  0.0-60.0 sec  1.82 GBytes    261 Mbits/sec  0.011 ms 282656/500002 (57%)
[  4]  0.0-60.3 sec  1.81 GBytes    258 Mbits/sec  13.649 ms 284287/500002 (57%)
[ 14]  0.0-60.0 sec  1.80 GBytes    258 Mbits/sec  0.009 ms 284859/500002 (57%)
```

UDP 800MBps*12 - p2p ping: Hardly all lost (1/60 recieved) – Bottleneck is in *hufsa*, P2P traffic is OK

```
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40541
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[  6]  0.0-60.0 sec    224 MBytes  31.3 Mbits/sec  0.049 ms 640556/666666 (96%)
[  7]  0.0-60.0 sec    185 MBytes  25.8 Mbits/sec  0.059 ms 645122/666666 (97%)
[  9]  0.0-60.0 sec    189 MBytes  26.4 Mbits/sec  0.030 ms 644632/666663 (97%)
[ 10]  0.0-60.0 sec    252 MBytes  35.2 Mbits/sec  0.030 ms 637300/666663 (96%)
[ 11]  0.0-60.0 sec    192 MBytes  26.8 Mbits/sec  0.088 ms 644308/666664 (97%)
[ 12]  0.0-60.0 sec    141 MBytes  19.7 Mbits/sec  0.129 ms 650270/666665 (98%)
[ 13]  0.0-60.0 sec    173 MBytes  24.2 Mbits/sec  0.101 ms 646494/666668 (97%)
[ 14]  0.0-60.0 sec    208 MBytes  29.1 Mbits/sec  0.020 ms 642413/666666 (96%)
[  3]  0.0-60.2 sec    247 MBytes  34.3 Mbits/sec  14.642 ms 637940/666664 (96%)
[  4]  0.0-60.3 sec    132 MBytes  18.4 Mbits/sec  14.201 ms 649659/665040 (98%)
[  5]  0.0-60.3 sec    163 MBytes  22.7 Mbits/sec  13.697 ms 647708/666663 (97%)
[  8]  0.0-60.3 sec    205 MBytes  28.5 Mbits/sec  13.827 ms 642777/666664 (96%)
```

UDP 1G*12 - p2p ping: Hardly all lost (1/60 recieved) – Bottleneck is in *hufsa*, P2P traffic is OK

```
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32880
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32807
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32881
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32804
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32872
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32864
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 40736
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32867
```

```
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32869
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32874
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32858
[  8]  0.0-60.0 sec  25.1 MBytes  3.50 Mbits/sec  0.163 ms 756341/759260 (1e+02%)
[  9]  0.0-60.0 sec  23.8 MBytes  3.32 Mbits/sec  0.060 ms 755525/758294 (1e+02%)
[ 11]  0.0-60.0 sec  26.5 MBytes  3.71 Mbits/sec  0.125 ms 756478/759568 (1e+02%)
[ 13]  0.0-60.0 sec  26.2 MBytes  3.67 Mbits/sec  0.070 ms 756485/759541 (1e+02%)
[ 14]  0.0-60.0 sec  27.0 MBytes  3.77 Mbits/sec  0.100 ms 755863/759006 (1e+02%)
[  3]  0.0-60.3 sec  31.5 MBytes  4.39 Mbits/sec  15.243 ms 755668/759338 (1e+02%)
[  4]  0.0-60.2 sec  30.1 MBytes  4.19 Mbits/sec  14.208 ms 754890/758397 (1e+02%)
[  5]  0.0-60.2 sec  29.9 MBytes  4.17 Mbits/sec  14.291 ms 755065/758551 (1e+02%)
[  6]  0.0-60.3 sec  27.9 MBytes  3.89 Mbits/sec  14.232 ms 755484/758740 (1e+02%)
[  7]  0.0-60.3 sec  25.1 MBytes  3.50 Mbits/sec  14.150 ms 756246/759176 (1e+02%)
[ 10]  0.0-60.2 sec  23.4 MBytes  3.25 Mbits/sec  13.738 ms 756500/759221 (1e+02%)
[ 12]  0.0-60.2 sec  22.7 MBytes  3.17 Mbits/sec  15.349 ms 756438/759088 (1e+02%)
```

Testing 192.16.166.240/28 -> 192.16.166.200

UDP 100MBps*12 - p2p ping: KO - Line goes down in a few seconds

```
------------------------------------------------------------
Server listening on UDP port 5001
Receiving 9000 byte datagrams
UDP buffer size:   107 KByte (default)
------------------------------------------------------------
[  3] local 128.142.208.6 port 5001 connected with 192.16.166.242 port 41399
[  4] local 128.142.208.6 port 5001 connected with 192.16.166.246 port 32873
[  5] local 128.142.208.6 port 5001 connected with 192.16.166.247 port 32866
[  6] local 128.142.208.6 port 5001 connected with 192.16.166.250 port 32881
[  7] local 128.142.208.6 port 5001 connected with 192.16.166.244 port 32805
[  8] local 128.142.208.6 port 5001 connected with 192.16.166.245 port 32876
[  9] local 128.142.208.6 port 5001 connected with 192.16.166.243 port 32808
[ 10] local 128.142.208.6 port 5001 connected with 192.16.166.251 port 32883
[ 11] local 128.142.208.6 port 5001 connected with 192.16.166.248 port 32869
[ 12] local 128.142.208.6 port 5001 connected with 192.16.166.254 port 32859
[ 13] local 128.142.208.6 port 5001 connected with 192.16.166.252 port 32871
[ 14] local 128.142.208.6 port 5001 connected with 192.16.166.249 port 32876
Waiting for server threads to complete. Interrupt again to force quit.
```

UDP 150MBps*12

can't test

UDP 300MBps*12

can't test

UDP 600MBps*12

can't test

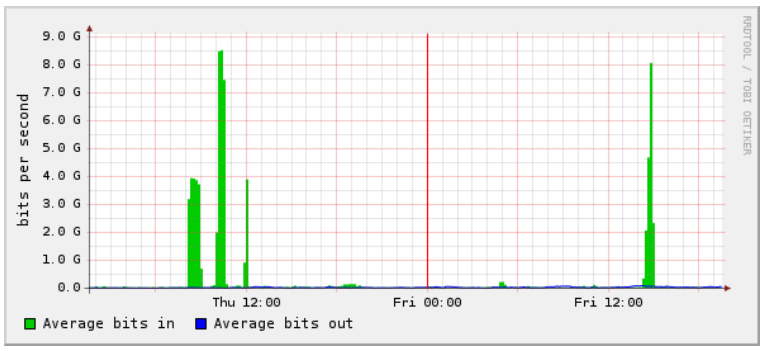
UDP 800MBps*12

can't test

UDP 1G*12

can't test

Statistics for the PIC-CERN's p2p interface at CERN:



Bibliografia

[1] “IEEE 802 LAN/MAN Standards Committee”, http://www.ieee802.org/ (15/05/2007)

[2] “Unshielded twisted pair”, http://en.wikipedia.org/wiki/Unshielded_twisted_pair (15/05/2007)

[3] “Category 6 Cabling Overview, FAQs and Whitepapers”, http://www.tiaonline.org/standards/technology/cat6/faq.cfm (15/05/2007)

[4] Sowmya S. Luckoor , “Introduction to 10 Gigabit 64b/66b (Clause 49)”, 22/10/2001

[5] Bill St Arnaud, “Frequently Asked Questions about Customer Owned Dark Fiber, Condominium Fiber, Community and Municipal Fiber Networks”, 31/03/2002, http://www.canarie.ca/canet4/library/customer/frequentlyaskedquestionsaboutdarkfiber.pdf, (28/05/07)

[6] “Introducing DWDM”, http://www.cisco.com/univercd/cc/td/doc/product/mels/dwdm/dwdm_fns.htm (28/05/2007)

[7] “IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks”, IEEE Std 802.1q, 19/05/2006

[8] “Catalyst 6500 Series Switch Cisco IOS Software Configuration Guide”, Capítol 7.2, Cisco Press, 2003

[9] “IETF RFC Page”, http://www.ietf.org/rfc.html (15/05/2007)

[10]Ivan Pepelnjakf. “EIGRP Network Design Solutions”, Cisco Press, 2000

[11]Matthew G. Naugle , “Illustrated TCP/IP”, Wiley Computer Publishing, John Wiley & sons, Inc. ,1998

[12]“Pat Eyler, “Networking Linux®: A Practical Guide to TCP/IP”, Que, 21/3/2001

[13]“Enabling High Performance Data Transfers”, http://www.psc.edu/networking/projects/tcptune/ (15/05/2007)

[14]“TCPDUMP Public repository”, http://www.tcpdump.org/ (15/05/2007)

[15]“Nmap – Free Security Scanner For Network Exploration & Security Audits”, http://insecure.org/nmap/ (15/05/2007)

[16]“The TCP/UDP Bandwidth Measurement Tool”, http://dast.nlanr.net/Projects/Iperf/ (15/05/2007)

[17]“Thrulay, network capacity tester” http://shlang.com/thrulay/ (15/05/2007)

[18]“Hot Standby Router Protocol Features and Functionality”, http://www.cisco.com/en/US/tech/tk648/tk362/technologies_tech_note09186a0080094a91.shtml

[19]“Hot Standby Routing Protocol by Peter J. Welcher”, http://www.netcraftsmen.net/welcher/papers/hsrp.htm

[20]Documentació oficial de Linux, apartat “Ehternet Bonding Driver”

[21]“The LHCOPN architecture document”, https://twiki.cern.ch/twiki/bin/view/LHCOPN/LHCopnArchitecture (16/05/2007)

[22]“RED (Random Early Detection) Queue Management”, http://www.icir.org/floyd/red.html (21/05/2007)

[23]“TCP over WAN Performance Tuning and Troubleshooting”, http://shlang.com/writing/tcp-perf.html (21/05/2007)

[24] Mathias de Riese, Patrick Fuhrmann, Tigran Mkrtchyan, Michael Ernst, Alex Kulyavtsev, Vladimir Podstavkov, Martin Radicke, Neha Sharma, Dmitry Litvintsev, Timur Perelmutov,“dCache, the Book”, http://www.dcache.org/manuals/Book/ (02/06/07)

[25] Mathias de Riese, Patrick Fuhrmann, Tigran Mkrtchyan, Michael Ernst, Alex Kulyavtsev, Vladimir Podstavkov, Martin Radicke, Neha Sharma, Dmitry Litvintsev, Timur Perelmutov,“dCache, the Book”, http://www.dcache.org/manuals/Book/ (02/06/07)

[26]K. Lahey, “TCP Problems with Path MTU Discovery”, rfc2923, 2000

Català

Aquest projecte consisteix en realitzar el disseny i desplegament d'una connexió entre el *Port d'Informació Científica* (PIC) i el *Consell Europeu per a la Recerca Nuclear* (CERN) sobre un circuit dedicat amb una velocitat de transferència de 10 Gbps. En una primera fase el desplegament de la connexió es realitza sobre un circuit dedicat de 1 Gbps.

El projecte implica la certificació dels circuits dedicats de 1 i 10 Gbps i el disseny dels plans d'actuació que han de permetre la integració de les noves connexions dins la xarxa i els serveis del PIC.

Castellà

Este proyecto consiste en realizar el diseño y despliegue de una conexión entre el *Port d'Informació Científica* (PIC) y el *Consejo Europeo para la Investigación Nuclear* (CERN) sobre un circuito dedicado con una velocidad de transferencia de 10 Gbps. En una primera fase el despliegue de la conexión se realiza sobre un circuito dedicado de 1 Gbps.

El proyecto implica la certificación de los circuitos dedicados de 1 y 10 Gbps y el diseño de los planes de actuación que deben permitir la integración de las nuevas conexiones dentro de la red y los servicios del PIC.

Anglès

This project consists in designing and deploying a connection between the *Port d'Informació Científica* (PIC) and the *European Organization for Nuclear Research* (CERN) on a 10 Gbps dedicated circuit. In a first phase the deployment of the connection is made on a 1 Gbps dedicated circuit.

The project implies the certification of the 1 and 10 Gbps dedicated circuits and the design of the plans for the integration of the new connections within PIC's network and services.