Analysis of a generic model for a bottleneck link in an integrated services communications network

Remco Litjens*
TNO ICT
Delft, The Netherlands
Phone +31 15 285 7184
Fax +31 15 285 7355
remco.litjens@tno.nl

Richard J. Boucherie
University of Twente
Enschede, The Netherlands
Phone +31 53 489 3432
Fax +31 53 489 3069
r.j.boucherie@utwente.nl

ABSTRACT

We develop and analyse a generic model for performance evaluation, parameter optimisation and dimensioning of a bottleneck link in an integrated services communications network. Possible application areas include IP, ATM and GSM/GPRS networks. The model enables analytical evaluation for a scenario of integrated speech, video and data services, selected for the fundamental differences in their service characteristics. While a speech call is assigned a single resource unit for its entire duration, both video and data calls can handle varying resource assignments. The principal distinction between these elastic call types, is that in case of video calls, a more generous resource assignment implies a better throughput and thus video quality, while for data calls the increased throughput implies a reduced transfer time. Markov chain analysis is applied to derive basic performance measures such as the expected resource utilization, service-specific blocking probabilities and the expected video and data QOS. Furthermore, analytical expressions are derived for the expected video and data QOS, conditional on the call duration or file size, respectively, and on the system state on arrival. As a potential application, these measures can be fed back to the caller as an indication of the expected QOS. A numerical study, focussing on a wireless access network, is included to demonstrate the merit of the presented generic model and performance analysis.

Keywords

Stochastic models, Markov models, performance modelling, communication systems and networks, wireless and mobile systems and networks.

1. INTRODUCTION

The incredible growth of data and multimedia communications both in wired (e.g. Internet) and wireless (e.g. Wifi) systems is undisputed, as well as the expectation of their convergence in the form of an integrated 'all IP' network handling all communications. Transmissions commonly experience high variation in transmission rates, due to concurrence with other traffic flows. This variation increases due to the inherit differences among the transmission types, such as voice, video and data, that each have their specific characteristics and resource requirements.

*Corresponding author.

Aside from the required technological network upgrades that enable the provisioning of the foreseen variety of services, it is essential to optimise link dimensioning and traffic management mechanisms to efficiently establish the desired service-specific Quality-Of-Service by means of the appropriate deployment of a.o. admission control, resource reservation and packet scheduling policies. Whereas such mechanisms were trivial or irrelevant in single service systems, in integrated services networks they are not only essential to avoid the loss of unsatisfied customers, but also offer an important means for differentiated service provisioning.

Contribution

The principal contribution of the present paper is the development and analysis of a *generic model* for performance evaluation, parameter optimisation and dimensioning in an integrated services communications network. The model enables analytical evaluation for a scenario of integrated speech, video and data services, with service-specific traffic characteristics and potentially offered in distinct priority classes. We note that the considered set of services cover the principal characteristics specifying the different traffic/QOS classes that are standardised for future integrated networks.

In the *performance analysis* presented to determine the QOS of the video and data service, the corresponding dynamics are modelled as queues in a random environment (see e.g. [12]). In our case the influence can be mutual, i.e. the arrival, service and departure process of all the different call types influence each other. Aside from basic performance measures such as service-specific call blocking probabilities, expected resource utilisation and expected video and data QOS, we also derive closed-form expressions for the expected video QOS (throughput) and data QOS (transfer time) conditional on the service requirement and the system state upon call arrival. A *numerical evaluation* is included in the example setting of a GSM/GPRS system to demonstrate the merit of the studied generic model and performance analysis.

Literature

There is a rich variety of models in which, for a single traffic type, the available service rate alternates between a positive value and complete absence of service, including unreliable servers, server vacations and service failures [6, 12, 14]. These models allow for closed-form solutions for many performance measures or structural decomposition results. When the service rate varies between several positive values explicit results are no longer available. Approximations for

Print 0-9553018-4-X, CD 0-9553018-3-1

transient analysis of single server queues with fluctuating service or arrival rate are studied in [3, 7, 11]. In particular, the mean queue length in a process with varying service rate is shown to exceed that in a process with a constant service rate (with the same mean). These papers consider only the case of a single customer type, and do not allow for generalisation to multiple types with priorities.

Analytical performance studies focusing on the impact of a random environment on the experienced QOS are rare. Based on general results for a queue in a random environment determined by a birth and death process [12], the conditional expected transfer time of data calls in an integrated system with stream and elastic traffic is analysed in [13] for an IP setting, and in [5, 9] for a wireless setting. An integrated system serving prioritised and best effort jobs is investigated in [2], presenting exact closed-form expressions and useful approximations for the expected sojourn times of prioritised and best effort jobs, respectively. The present paper generalises the frameworks mentioned above to also include a distinct service of the video type, for which the transmission time is fixed, but the perceived QOS is determined by the experienced throughput. With the inclusion of the video service, we present a tractable flow level integrated services model that covers all key service types.

Outline

Section 2 presents a generic model for a bottleneck link in an integrated services network, which is extensively analysed in Section 3. Considering a GSM/GPRS system as a possible application area for the generic model and analysis, Section 4 presents a set of illustrative numerical experiments. The proof of a key result is provided in the Appendix.

2. MODEL

This section defines the framework for the performance analysis of a bottleneck link in a communications network integrating speech, video, and data services. See Figure 1.

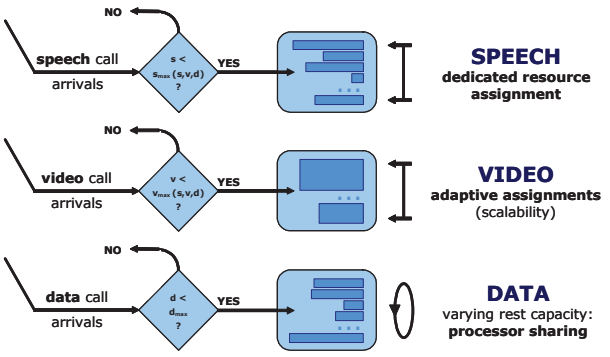

Figure 1: Illustration of the considered model of a bottleneck link in an integrated services communications network.

2.1 Call characteristics

The considered services have been selected for the fundamental differences in their characteristics.

Speech calls arrive according to a Poisson process with arrival intensity λ_{speech}, have an exponentially distributed duration with mean $1/\mu_{\text{speech}}$ and require a fixed assignment of one resource unit. The speech traffic load is given by $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$.

Video calls arrive according to a Poisson process with arrival intensity λ_{video} and have an exponentially distributed duration with mean $1/\mu_{\text{video}}$. Video calls are modelled as continuous real-time streams that are *scalable* in the sense that the amount of assigned resources and thus the video quality is adaptive to the varying network load. Scaling is assumed to adhere to any resource reassignment ideally and instantaneously. Denote with r_{video} the fixed video information bit rate (in kbits/s) per assigned resource unit. As a minimum QOS requirement, each video call must be assigned no fewer than $\beta^{\min}_{\text{video}}$ resource units, corresponding to a bit rate of $r_{\text{video}}\,\beta^{\min}_{\text{video}}$ kbits/s. On the other hand, denote with $\beta^{\max}_{\text{video}} \geq \beta^{\min}_{\text{video}}$ the peak resource assignment for video calls, which may be dictated by the service or terminal characteristics. The video traffic load is defined as $\rho_{\text{video}} \equiv \beta^{\min}_{\text{video}}\lambda_{\text{video}}/\mu_{\text{video}}$.

Data calls arrive according to a Poisson process with arrival intensity λ_{data}. A data call is assumed to be the downlink transfer of a file with an exponentially distributed size

The data call size is expressed in units of the data information bit rate of r_{data} kbits/s per resource unit and has mean $1/\mu_{\text{data}}$, which corresponds to $r_{\text{data}}/\mu_{\text{data}}$ kbits. The data traffic load is given by $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$. Data calls are assumed to be *elastic* in the sense that they are delay tolerant and can handle varying resource assignments. In contrast to the video service, where the resource assignments do not affect the autonomously sampled call durations, the presence of a data call *is* affected by the resource assignment: the more generous the assignment, the shorted the transfer time. The number of resource units that can be assigned to a data call is limited to the service- or terminal-dictated maximum $\beta^{\max}_{\text{data}}$. We assume that for a given file there is at any time sufficient data available in the buffer feeding the considered communication link to be carried on the dynamically assigned resources. In integrated services models, the assumption of exponential data call sizes, required for analytical tractability, generally leads to some degree of *over*estimation of the expected transfer times, compared to scenarios with a greater call size variability, thus leading to conservative network dimensioning [10].

2.2 Call handling schemes

The considered bottleneck link, integrating speech, video and data services, is characterised by a capacity of C_{total} resource units. The proposed *resource sharing* scheme splits the pool of C_{total} resource units into three distinct subsets: C_{speech}, C_{video}, and C_{data} with $C_{\text{speech}} + C_{\text{video}} + C_{\text{data}} = C_{\text{total}}$. Based on these 'territories', we propose and consider the following resource sharing scheme, which establishes some form of capacity reservation for all service types, while still providing a high resource utilization through varying elastic call assignments.

C_{speech} resource units are reserved for speech calls with preemptive priority, i.e. video and data calls may use these resources whenever they are unused by the speech service, but must free them immediately once needed to support newly admitted speech calls. Furthermore, within these C_{speech} resource units, video calls are treated with strict preference over data calls. C_{video} resource units are shared by all call types with preference for speech and video calls. Video calls must downgrade their assignment (potentially

down to $\beta_{\text{video}}^{\min}$ resource units) only in support of newly admitted speech or video calls. Data calls may utilise the resources that cannot be assigned to the preferred speech or video calls. Note that it is in this resource pool that video calls must find their minimum assignment of $\beta_{\text{video}}^{\min}$ resource units, as resources grabbed elsewhere may have to be released again in favour of newly admitted calls. C_{data} resource units are reserved for data calls with preemptive priority, i.e. video calls can grab free resources in this pool due to their scalability property but only if such resources would otherwise be idle, i.e. if $\beta_{\text{data}}^{\max} d < C_{\text{data}}$, with d the number of present data calls. Speech calls are prohibited to use these resources.

The system evolution can be modelled as a continuous-time Markov chain $(S(t), V(t), D(t))_{t\geq 0}$ where $S(t)$, $V(t)$ and $D(t)$ are defined as the number of speech, video and data calls, respectively, that are present at time t. The system states are denoted (s, v, d) with state space $\mathbb{S}$. Using this notation, the service-specific *call admission control* and *resource assignment* schemes are defined as follows.

A *speech* call is blocked iff upon arrival no resource unit is, or can be made available to support the call, i.e. iff

$$s = s_{\max}(v) \equiv \left\lfloor C_{\text{speech}} + C_{\text{video}} - \beta_{\text{video}}^{\min} v \right\rfloor .$$

A *video* call is blocked iff upon arrival the minimum assignment of $\beta_{\text{video}}^{\min}$ resource units cannot be made available to support the call, i.e. iff

$$v = v_{\max}(s) \equiv \left\lfloor \frac{C_{\text{video}} - \max\{s - C_{\text{speech}}, 0\}}{\beta_{\text{video}}^{\min}} \right\rfloor ,$$

where $\max\{0, s - C_{\text{speech}}\}$ is the number of shared resource units that are in use by speech calls. The number of resource units available for video transfer is $\max\{C_{\text{data}} - \beta_{\text{data}}^{\max} d, 0\} + (C_{\text{video}} + C_{\text{speech}} - s)$. The first part of this expression indicates the resources that are available in the C_{data} pool, while the second part gives the available resources in the joint $C_{\text{video}} + C_{\text{speech}}$ pool. The available resources are then evenly distributed over the present video calls, effectively applying a Processor Sharing service (PS) discipline. The resource assignment function $\beta_{\text{video}}(s, v, d)$ prescribes the amount of resources that is assigned to each admitted video call in system state (s, v, d).

$$\beta_{\text{video}}(s, v, d) \equiv \min\{\beta_{\text{video}}^{\max},$$

$$\left. \frac{\max\{C_{\text{data}} - \beta_{\text{data}}^{\max} d, 0\} + (C_{\text{video}} + C_{\text{speech}} - s)}{v} \right\} .$$

It is readily verified that $\beta_{\text{video}}(s, v, d) \geq \beta_{\text{video}}^{\min}$ for all $(s, v, d) \in \mathbb{S}$. As no minimum assignment is assumed for *data* calls, admission control is simply based on an exogenously given maximum on the number of present data calls in the system, denoted $d_{\max}$. As for video calls, at any time, the resources that are available for the data service are fairly shared by all admitted data calls according to a PS service discipline:

$$\beta_{\text{data}}(s, v, d) \equiv \min\left\{ \beta_{\text{data}}^{\max}, \frac{C_{\text{total}} - s - \beta_{\text{video}}(s, v, d) v}{d} \right\} .$$

2.3 On the genericity of the model

A number of extensions and generalisations of the system model and, in particular, the call handling schemes can be made without complicating the performance analysis presented below, as long as model adjustments can be captured by admission control thresholds and resource assignment schemes that have the same general form as those given above. These generalisations have been consciously omitted here for clarity of presentation.

Among the feasible generalisations we mention the following: *(i)* the application of QOS differentiation between different classes of video and/or data calls; *(ii)* the introduction a service-specific FCFS access queue to hold calls that cannot be admitted immediately upon arrival; *(iii)* analysis of different resource sharing schemes; *(iv)* consideration of minimum resource assignments for data calls to ensure some minimum QOS; *(v)* restriction of (data or) video call assignments to be limited to a number of prefixed levels, e.g. 2, 4 or 8 resource units if this corresponds more accurately to an assumed scalable video coding algorithm (see e.g. [1]).

3. PERFORMANCE ANALYSIS

In this section, the system evolution of the model is formulated as a continuous-time Markov chain. Some basic performance measures and an extensive conditional analysis is presented for video and data calls, deriving the expected QOS as a function of the call duration/size and the system state upon call arrival.

3.1 Markov chain and equilibrium

The evolution of the system model can be described by an irreducible three-dimensional continuous-time Markov chain $(S(t), V(t), D(t))_{t\geq 0}$, with states denoted (s, v, d). The state space of the Markov chain is given by

$$\mathbb{S} \equiv \left\{ \ (s, v, d) : s \leq s_{\max}(v) \wedge v \leq v_{\max}(s) \wedge d \leq d_{\max} \ \right\} .$$

The speech, video and data call arrival rates are given by λ_{speech}, λ_{video} and λ_{data}, while the speech, video and data call departure rates in system state (s, v, d) are given by $s\mu_{\text{speech}}$, $v\mu_{\text{video}}$ and $\beta_{\text{data}}(s, v, d)d\mu_{\text{data}}$, respectively. Ordering $\mathbb{S}$ lexicographically in (s, v, d), the infinitesimal generator $\mathcal{Q}$ is of tridiagonal block structure, with above-diagonal blocks generating speech call arrivals, below-diago-nal blocks generating speech call terminations, and diagonal blocks generating video and data call arrival and termination events.

Since the considered finite state space Markov chain $(S(t), V(t), D(t))_{t\geq 0}$ is irreducible, a unique probability vector π exists that satisfies the system of global balance equations: $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$, with $\mathbf{0}$ the vector with all entries zero, and $\boldsymbol{\pi}$ lexicographically ordered in $(s, v, d) \in \mathbb{S}$.

3.2 Basic performance measures

From a system's perspective, the resource efficiency can be measured by the *expected resource utilization:*

$$\mathbf{U} \equiv C_{\text{total}}^{-1} \sum_{(s,v,d)\in\mathbb{S}} \pi(s, v, d) \begin{pmatrix} s + \beta_{\text{video}}(s, v, d) v \\ +\beta_{\text{data}}(s, v, d) d \end{pmatrix} .$$

The service-specific *blocking probabilities* are readily derived from the equilibrium distribution using using the PASTA property. The video QOS is expressed in the *expected video throughput*. As the measure indicates the expected *per-call* video throughput, we must condition on the presence of at least one video call, obtaining

$$\mathbf{R}_{\text{video}} \equiv r_{\text{video}} \left(\frac{\sum_{(s,v,d)\in\mathbb{S}^+_{\text{video}}} \pi(s, v, d) \beta_{\text{video}}(s, v, d)}{\sum_{(s,v,d)\in\mathbb{S}^+_{\text{video}}} \pi(s, v, d)} \right) ,$$

with $\mathbb{S}^+_{\rm video} \equiv \{(s,v,d) \in \mathbb{S} : v > 0\}$. The data QOS is expressed in the *expected transfer time* of a data call, which is readily obtained using Little's law:

$$\mathbf{T}_{\rm data} \equiv \frac{\sum_{(s,v,d)\in\mathbb{S}} \pi(s,v,d)d}{\lambda_{\rm data}(1-\mathbf{P}_{\rm data})}. \quad (1)$$

Another relevant measure characterizing the data QOS is the *expected data throughput,* which is given by

$$\mathbf{R}_{\rm data} \equiv r_{\rm data}\left(\frac{\sum_{(s,v,d)\in\mathbb{S}^+_{\rm data}} \pi(s,v,d)\beta_{\rm data}(s,v,d)}{\sum_{(s,v,d)\in\mathbb{S}^+_{\rm data}} \pi(s,v,d)}\right),$$

with $\mathbb{S}^+_{\rm data} \equiv \{(s,v,d) \in \mathbb{S} : d > 0\}$.

3.3 Conditional performance measures

3.3.1 Conditional analysis of the video QOS

While $\mathbf{R}_{\rm video}$ is a *time*-average video throughput measure, in this section a *call*-average throughput measure is determined, which is undeniably the most appropriate throughput measure from a call's perspective.

For each state $(s,v,d) \in \mathbb{S}^+_{\rm video}$ define $x_{s,v,d}(\tau)$ as the random number of bits (*transfer volume*) transmitted by an admitted video call of duration τ, arriving at a given system state (s,v,d), where v *includes* the new video call, and let $\widehat{x}_{s,v,d}(\tau) \equiv \mathbf{E}\{x_{s,v,d}(\tau)\}$ denote its expectation. Then the corresponding expected throughput is equal to $\widehat{x}_{s,v,d}(\tau)/\tau$, while the expected throughput of an *admitted* video call of duration τ is given by

$$\mathbf{R}^\star_{\rm video}(\tau) \equiv \frac{\sum_{(s,v,d)\in\mathbb{S}^+_{\rm video}} \pi(s,v-1,d)\widehat{x}_{s,v,d}(\tau)}{\tau\,(1-\mathbf{P}_{\rm video})} \quad (2)$$

where $\pi(s,v-1,d)/(1-\mathbf{P}_{\rm video})$ is the equilibrium probability that the system is in state $(s,v-1,d)$, conditioned on the admission of an arriving video call. Integrating $\mathbf{R}^\star_{\rm video}(\tau)$ over the probability density function of τ yields the expected (call-average) throughput of an admitted video call:

$$\mathbf{R}^\star_{\rm video} \equiv \int_{\tau=0}^{\infty} \mathbf{R}^\star_{\rm video}(\tau)\,\mu_{\rm video}\,\exp\{-\tau\,\mu_{\rm video}\}\,d\tau.$$

We stress that in general the *time*-average video throughput $\mathbf{R}_{\rm video}$ and the *call*-average video throughput $\mathbf{R}^\star_{\rm video}$ need not be the same. Refer to [8] for a more extensive comparison of throughput measures in PS models. It is noted that the values of $\widehat{x}_{s,v,d}(\tau)/\tau$, $(s,v,d) \in \mathbb{S}^+_{\rm video}$, may be at least as valuable as $\mathbf{R}^\star_{\rm video}(\tau)$ or $\mathbf{R}^\star_{\rm video}$, since, given the system state at arrival, the appropriate value can be fed back to the source as an indication of the expected service quality.

In the following an explicit expression for the vector $\widehat{\mathbf{x}}(\tau) = \left(\widehat{x}_{s,v,d}(\tau),\ (s,v,d) \in \mathbb{S}^+_{\rm video}\right)$ is derived. To this end, we introduce the generator $\mathcal{Q}^\star_{\rm video}$, that is characterised by the presence of *one permanent video call* that never leaves the system, and shares in the available resources as if it were an ordinary video call. This permanent video call is the tagged call whose throughput is to be determined, while the behaviour of all other calls is unchanged. For all $(s,v,d) \in \mathbb{S}^+_{\rm video}$, the video call departure rates are modified as follows:

$$\mathcal{Q}^\star_{\rm video}\left((s,v,d);(s,v-1,d)\right) = (v-1)\,\mu_{\rm video}.$$

Let $\boldsymbol{\pi}^\star_{\rm video} \equiv (\pi^\star_{\rm video}(s,v,d),\ (s,v,d) \in \mathbb{S}^+_{\rm video})$ be the stationary distribution, i.e. $\boldsymbol{\pi}^\star_{\rm video}\mathcal{Q}^\star_{\rm video} = \mathbf{0}$. Let $\mathcal{B}_{\rm video} \equiv diag(\beta_{\rm video}(s,v,d),\ (s,v,d) \in \mathbb{S}^+_{\rm video})$ denote the diagonal matrix of average resource assignments, lexicographically ordered in (s,v,d). We may now formulate the following expression of the conditional expected transfer volume.

THEOREM 1. *Let* $\boldsymbol{\gamma}_{video} \equiv (\gamma_{video}(s,v,d),\ (s,v,d) \in \mathbb{S}^+_{video})$ *be the unique solution to*

$$\mathcal{Q}^\star_{video}\boldsymbol{\gamma}_{video} = (\boldsymbol{\pi}^\star_{video} r_{video}\mathcal{B}_{video}\mathbf{1})\,\mathbf{1} - r_{video}\mathcal{B}_{video}\mathbf{1}, \quad (3)$$

$$\boldsymbol{\pi}^\star\boldsymbol{\gamma}_{video} = \mathbf{0}. \quad (4)$$

Then the conditional expected throughput vector is given by

$$\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = (\boldsymbol{\pi}^\star_{video} r_{video}\mathcal{B}_{video}\mathbf{1})\,\mathbf{1} + \tau^{-1}\left[\mathcal{I} - \exp\{\tau\mathcal{Q}^\star_{video}\}\right]\boldsymbol{\gamma}_{video},$$

which asymptotically converges to

$$\lim_{\tau\to\infty}\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = (\boldsymbol{\pi}^\star_{video} r_{video}\mathcal{B}_{video}\mathbf{1})\,\mathbf{1}.$$

Hence the (conditional) expected (call-average) video throughput is asymptotically equal to the (conditional) expected (time-average) video throughput in a system with one permanent video call.

PROOF. See Appendix A. □

Observe that the asymptotic expression, given by $\widehat{\mathbf{x}}(\tau)/\tau = (\boldsymbol{\pi}^\star_{\rm video} r_{\rm video}\mathcal{B}_{\rm video}\mathbf{1})\,\mathbf{1} + \boldsymbol{\gamma}_{\rm video}/\tau$, is non-linear in τ, as will also be illustrated in Section 4. We further note that in a model without (noticeable) data traffic, i.e. if $\lambda_{\rm data} = 0$ and/or $C_{\rm data} = 0$, the asymptotic video throughput expression given above, holds not only for $\tau \to \infty$, but for any (finite) τ. Hence in this scenario the call-average video throughput is independent of the video call duration τ, and identical to the time-average throughput in a system with one permanent video call.

3.3.2 Conditional analysis of the data QOS

As demonstrated in Section 3.2, the expected transfer time $\mathbf{T}_{\rm data}$ of a data call is readily calculated from the equilibrium distribution $\pi(s,v,d)$, $(s,v,d) \in \mathbb{S}$. In this subsection, we determine $\mathbf{T}_{\rm data}(x)$, the expected transfer time of an admitted data call of size $x \geq 0$. Compared to $\mathbf{T}_{\rm data}$, the analysis of $\mathbf{T}_{\rm data}(x)$ is considerably more complicated. Since the conditional analysis is analogous to that presented in [9, 12, 13], we only state the main results here for completeness.

For each state $(s,v,d) \in \mathbb{S}^+_{\rm data}$ define $\tau_{s,v,d}(\tau)$ as the random transfer time of an admitted data call of size x, arriving at system state (s,v,d), where d *includes* the new data call, and let $\widehat{\tau}_{s,v,d}(x) \equiv \mathbf{E}\{\tau_{s,v,d}(x)\}$ denote its expectation. Then the expected transfer time of an *admitted* data call of size x is given by

$$\mathbf{T}_{\rm data}(x) \equiv \frac{\sum_{(s,v,d)\in\mathbb{S}^+_{\rm data}} \pi(s,v,d-1)\widehat{\tau}_{s,v,d}(x)}{1-\mathbf{P}_{\rm data}}. \quad (5)$$

The integral of $\mathbf{T}_{\rm data}(x)$ over all possible values of τ yields the expected (call-average) throughput $\mathbf{T}_{\rm data}$, see (1). As was noted for the video service, the obtained values of $\widehat{\tau}_{s,v,d}(x)$, $(s,v,d) \in \mathbb{S}^+_{\rm data}$, may be at least as valuable as $\mathbf{T}_{\rm data}(x)$ or $\mathbf{T}_{\rm data}$, since they can be fed back to the source as an indication of how long the transmission is expected to take.

In the following an explicit expression for the vector $\widehat{\boldsymbol{\tau}}(x) = (\widehat{\tau}_{s,v,d}(x),\ (s,v,d) \in \mathbb{S}^+_{\rm data})$, $x \in \mathbb{R}^+$, is derived. In a similar way as done in the video QOS analysis, for the data

transfer time analysis denote with $\mathcal{Q}^{\star}_{\rm data}$ the infinitesimal generator of the modified Markov chain, characterised by the presence of *one permanent data call*. Furthermore, let $\mathcal{B}_{\rm data} \equiv diag(\beta_{\rm data}(s,v,d),\ (s,v,d) \in \mathbb{S}^{+}_{\rm data})$ denote the diagonal matrix of average data resource assignments, lexicographically ordered in (s,v,d).

In case $C_{\rm data} = 0$, it may occur that no resources are available to the data calls, i.e. $\beta_{\rm data}\,(s,v,d) = 0$ for some states $(s,v,d) \in \mathbb{S}^{+}_{\rm data}$, depending on the model parameters. Partition $\mathbb{S}^{+}_{\rm data}$ into $\mathbb{S}^{+}_{\rm data,0} \equiv \{(s,v,d) \in \mathbb{S}^{+}_{\rm data} : \beta_{\rm data}\,(s,v,d) = 0\}$ and its complement $\mathbb{S}^{+}_{\rm data,+} \equiv \mathbb{S}^{+}_{\rm data} \backslash \mathbb{S}^{+}_{\rm data,0}$, and accordingly partition

$$\mathcal{Q}^{\star}_{\rm data} = \begin{bmatrix} \mathcal{Q}^{\star}_{++} & \mathcal{Q}^{\star}_{+0} \\ \mathcal{Q}^{\star}_{0+} & \mathcal{Q}^{\star}_{00} \end{bmatrix},\ \mathcal{B}_{\rm data} = \begin{bmatrix} \mathcal{B}_{+} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{bmatrix},$$

Let $\boldsymbol{\pi}^{\star}_{\rm data} \equiv (\pi^{\star}_{\rm data}(s,v,d),\ (s,v,d) \in \mathbb{S}^{+}_{\rm data})$ be the stationary distribution, i.e. $\boldsymbol{\pi}^{\star}_{\rm data}\mathcal{Q}^{\star}_{\rm data} = \mathbf{0}$, and apply the partitionining $\boldsymbol{\pi}^{\star}_{\rm data} = (\boldsymbol{\pi}^{\star}_{\rm data,0},\ \boldsymbol{\pi}^{\star}_{\rm data,+})$. For the general case where $\mathbb{S}^{+}_{\rm data,0} \neq \emptyset$, Theorem 2 presents analytical expressions for the conditional expected transfer times $\widehat{\boldsymbol{\tau}}(x)$.

THEOREM 2. *Let* $\boldsymbol{\gamma}_{data} \equiv (\gamma_{data}\,(s,v,d),\ (s,v,d) \in \mathbb{S}^{+}_{data,+})$ *uniquely solve the system of linear equations*

$$\mathcal{B}_{+}^{-1}\left(\mathcal{Q}^{\star}_{++} + \mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathcal{Q}^{\star}_{0+}\right)\boldsymbol{\gamma}_{data} = \frac{1}{\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\mathbf{1}}\mathbf{1} - \mathcal{B}_{+}^{-1}\left(\mathcal{I} + \mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\right)\mathbf{1},$$

$$\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\boldsymbol{\gamma}_{data} = 1.$$

The solution for $\widehat{\boldsymbol{\tau}}(x) = (\widehat{\boldsymbol{\tau}}_0(x), \widehat{\boldsymbol{\tau}}_{+}(x))$ *is then given by*

$$\widehat{\boldsymbol{\tau}}_0(x) = \left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\left\{\mathbf{1} + \mathcal{Q}^{\star}_{0+}\widehat{\boldsymbol{\tau}}_{+}(x)\right\},$$

$$\widehat{\boldsymbol{\tau}}_{+}(x) = \frac{x}{\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}_{+}^{-1}\left(\mathcal{Q}^{\star}_{++} + \mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathcal{Q}^{\star}_{0+}\right)\right\}\right]\boldsymbol{\gamma}_{data},$$

while the asymptotic expressions are given by

$$\lim_{x\to\infty}\left\{\widehat{\boldsymbol{\tau}}_0(x) - \frac{x}{\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\mathbf{1}}\mathbf{1}\right\} = \left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathbf{1} + \left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathcal{Q}^{\star}_{0+}\boldsymbol{\gamma}_{data}$$

and

$$\lim_{x\to\infty}\left\{\widehat{\boldsymbol{\tau}}_{+}(x) - \frac{x}{\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\mathbf{1}}\mathbf{1}\right\} = \boldsymbol{\gamma}_{data}, \qquad (6)$$

indicating that for large data calls the expected transfer time is approximately linear in the size (fairness).

PROOF. The proof is a rather straightforward extension of Corollary 5.2 in [12] and therefore omitted. □

4. NUMERICAL RESULTS

This section presents a brief numerical study in order to demonstrate the merit of the considered model and performance analysis. Concentrating on a single cell in a GSM/GPRS radio access network as a typical example setting, Table 1 below gives an overview of all model parameters. Some comments regarding these parameters are made below.

SYSTEM PARAMETERS		CALL CHARACTERISTICS	
$C_{\rm total}$	21 channels	$\mu^{-1}_{\rm speech}$	50 seconds
$C_{\rm speech}$	12 channels	$\mu^{-1}_{\rm video}$	50 seconds
$C_{\rm video}$	6 channels	$\mu^{-1}_{\rm data}$	35.3591 seconds
$C_{\rm data}$	3 channels	$\beta^{\rm min}_{\rm video}$	2 channels
$\beta^{\rm max}_{\rm video}$	4 channels	$r_{\rm video}$	13.40 kbits/s
$\beta^{\rm max}_{\rm data}$	4 channels	$r_{\rm data}$	9.05 kbits/s
Q_t	10 data calls	$\varphi_{\rm speech}$	0.356625
		$\varphi_{\rm video}$	0.015547
		$\varphi_{\rm data}$	0.627828

Table 1: *Numerical results: parameter settings.*

The capacity of the considered GSM/ GPRS cell is prefixed by a typical assignment of 3 frequencies, which according to GSM's FD/TDMA technology provides $3 \times 8 = 24$ physical channels. Assigning 3 channels for control signalling purposes, this leaves $C_{\rm total} = 21$ traffic channels. As a typical GPRS terminal is characterised by a multichannel capability of four traffic channels, the upper bounds on the resource assignment is given by $\beta^{\rm max}_{\rm video} = \beta^{\rm max}_{\rm data} = 4$.

The average speech and video call holding time are both equal to 50 seconds. Assuming an average file size of 320 kbits and GPRS coding scheme CS-1 for maximum error correction potential, $r_{\rm data} = 9.05$ kb/s and hence the normalised data call size has an average of $320/9.05 = 35.3591$ transmission seconds, given a single dedicated channel assignment. For video calls, the less protective coding scheme CS-2 is assumed, which provides a channel rate of $r_{\rm video} = 13.4$ kb/s. The service mix is defined by the arrival fractions $\varphi_{\rm speech}$, $\varphi_{\rm video}$ and $\varphi_{\rm data}$, with $\varphi_{\rm speech} + \varphi_{\rm video} + \varphi_{\rm data} = 1$, which are determined as follows. Considering the given channel pool partitioning and channel sharing schemes, we determine for each service individually the maximum arrival rate such that the blocking probability is no more than 1%, assuming an otherwise empty system. This exercise yields $\lambda_{\rm speech} = 0.20874$, $\lambda_{\rm video} = 0.0091$ and $\lambda_{\rm data} = 0.36748$, which in turn yields the given relative arrival fractions.

4.1 Basic performance measures

In the first set of experiments we show the impact of the traffic load on the different basic performance measures. The traffic load is varied via the aggregate arrival rate $\lambda_{\rm speech} + \lambda_{\rm video} + \lambda_{\rm data}$, while fixing the ratio $\lambda_{\rm speech} : \lambda_{\rm video} : \lambda_{\rm data}$ to $\varphi_{\rm speech} : \varphi_{\rm video} : \varphi_{\rm data}$.

The results are depicted in the upper pair of charts in Figure 2. Not surprisingly, in the left chart the channel utilisation and service-specific blocking probabilities are increasing in the traffic load. Observe that, although the same target blocking probability of 1% was considered to determine the default arrival rates, the obtained blocking probabilities at the default traffic load are not only significantly larger than 1%, due to presence of other, competing traffic, but also significantly different, due to the distinct policies in the territories. The right chart depicts the time-average video and data throughput measures, the expected data transfer time, as well as derived (asymptotic) approximation for the call-average video throughput. Observe that both video throughput curves are very similar. For low traffic loads, the QOS curves are flat at the best achievable levels, imposed by the limitations imposed by $\beta^{\rm max}_{\rm video,data}$, while under heavier traffic all curves show monotonously worsening QOS

levels. Observe that the video QOS remains well above its minimum guarantee of $r_{\text{video}}\beta^{\min}_{\text{video}} = 26.8$ kb/s, indicating that the system is not really congested yet, from the video service's perspective.

4.2 Conditional performance measures

The middle pair of charts in Figure 2 show the conditional expected video QOS. The 3D chart on the left depicts the expected video throughput $\widehat{x}_{s,v,0}(\tau)/\tau$ experienced by a tagged video call of average duration ($\tau = 50$) as a function of s and v ($d = 0$). Note the domain $\{(s, v, 0) : s + \beta^{\min}_{\text{video}}\,(v + 1) \leq C_{\text{speech}} + C_{\text{video}} = 18$ and $v+1 \leq C_{\text{video}}/\beta^{\min}_{\text{video}} = 3\}$. As expected, $\widehat{x}_{s,v,0}(\tau)/\tau$ is decreasing in both s and v, while it is most sensitive to a change in the number of *video* calls, since an additional video call claims $\beta^{\min}_{\text{video}} = 2$ times as many traffic channels as an additional speech call. The chart on the right demonstrates the conditional video QOS conditioned only on the video call duration, as well as the corresponding derived asymptote, which appears to provide a very tight approximation, and both an upper and lower bound, corresponding with the best- ($\widehat{x}_{0,1,0}(\tau)/\tau$) and worst-case curves ($\widehat{x}_{12,3,10}(\tau)/\tau$) of the conditional expected QOS, given an empty or full system upon arrival of the considered call, respectively (see dashed curves).

Similarly, the lower pair of charts in Figure 3 show the conditional expected data QOS. The 3D chart on the left depicts the expected transfer time $\widehat{\tau}_{s,0,d}(x)$ experienced by a tagged data call of average size ($x = 320/9.05 = 35.3591$) as a function of s and d ($v = 0$). Theorem 2 has been applied to obtain the conditional expected data call transfer times. The right chart shows the conditional expected transfer time, conditioned only on the (normalised) data call size, accompanied by the derived asymptote and the upper/lower bounds, given by with $\widehat{\tau}_{12,3,10}(x)$ and $\widehat{\tau}_{0,0,1}(x)$, respectively.

5. CONCLUDING REMARKS

We have developed and analysed a generic model for performance evaluation, parameter optimisation and dimensioning of a bottleneck in an integrated services communication network. Markov chain analysis applying both dedicated resource assignments and Processor Sharing-type service disciplines, has been applied to derive a variety of performance measures, including exact expressions for the expected video and data QOS, conditional on the call duration or file size, respectively, and on the system state of arrival.

A number of extensions and generalisations of the system model and, in particular, the call handling schemes can be made without complicating the performance analysis presented below, e.g. *(i)* the application of QOS differentiation between different classes of video and/or data calls; *(ii)* the introduction a service-specific FCFS access queue to hold calls that cannot be admitted immediately upon arrival; *(iii)* analysis of different resource sharing schemes; *(iv)* restriction of (data or) video call assignments to be limited to certain prefixed levels, if this corresponds more accurately to an assumed scalable video coding algorithm (see e.g. [1]).

6. REFERENCES

[1] L. Begain, "Scalable multimedia services in GSM-based networks: an analytical approach," *Proc. of the ITC specialist seminar on Mobile systems and mobility*, Lillehammer, Norway, 2000.

[2] J.L. van den Berg, R. van der Mei, B. Gijsen, M. Pikaart and R. Vranken, "Processing times for transaction servers with quality of service differentiation", *Proc. of the 11th GI/ITG conference on Measuring, modelling and evaluation of computer and communications systems*, Aachen, Germany, 2001.

[3] S.K. Cheung, R. Nunez-Queija and R.J. Boucherie, "Effective load and adjusted stability in queues with fluctuating service rates", *Internal report*, Universiteit Twente, The Netherlands, 2006

[4] E.A. Coddington and N. Levinson, *"Theory of ordinary differential equations"*, McGraw-Hill, USA, 1955.

[5] G. Fodor and M. Telek, "Performance analysis of the uplink of a CDMA cell supporting elastic services, *Proc. of Networking '05*, Waterloo, Canada, 2005.

[6] S. Fuhrmann and R. Cooper, "Stochastic decompositions in the M/G/1 queue with generalized vacations", *Operations research*, 33 (5), 1985.

[7] V. Gupta, M. Harchol-Balter, A.S. Wolf and U. Yechiali, "Fundamental characteristics of queues with fluctuating load", *Proc. of Sigmetrics/ Performance '06*, Saint Malo, France, 2006.

[8] R. Litjens, J.L. van den Berg and R.J. Boucherie, "Throughput measures for processor sharing models," *submitted*, 2003.

[9] R. Litjens and R.J. Boucherie, " Performance analysis of fair channel sharing policies in an integrated cellular voice/data network," *Telecommunications systems*, 19 (2), 2002.

[10] R. Litjens and R.J. Boucherie, "Elastic calls in an integrated services network: the greater the variability the better the Quality-of-Service," *Performance evaluation*, 52 (4), 2003.

[11] W.A. Massey and W. Whitt, "Uniform acceleration expansions for Markov chains with time-varying rates", *Annals of applied probability*, 8 (4), 1997.

[12] R. Núñez Queija, "Sojourn times in non-homogeneous QBD processes with processor sharing," *Stochastic models*, 17, 2001.

[13] R. Núñez Queija, J.L. van den Berg, and M.R.H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic," *Proc. of ITC 16*, Edinburgh, Scotland, 1999.

[14] H. Takagi, *"Queueing analysis, vacations and priority systems"*, part 1, vol. 1, Elsevier Science Publishers, The Netherlands, 1991.

[15] H.C. Tijms, *"Stochastic modelling and analysis: a computational approach,"* Wiley, England, 1986.

APPENDIX

A. PROOF OF THEOREM 1

Define the Laplace-Stieltjes transform of the distribution of $x_{s,v,d}(\tau)$ by $X_{s,v,d}(\zeta,\tau) \equiv \mathbf{E}\left\{\exp\left\{-\zeta\, x_{s,v,d}(\tau)\right\}\right\}$, for $\text{Re}(\zeta) \geq 0$, $(s, v, d) \in \mathbb{S}^{+}_{\text{video}}$, and let $\mathbf{X}(\zeta,\tau) \equiv (X_{s,v,d}(\zeta,\tau),$ $(s, v, d) \in \mathbb{S}^{+}_{\text{video}})$ be lexicographically ordered.

LEMMA 1. *For $\tau \geq 0$ and $\text{Re}(\zeta) \geq 0$, $\mathbf{X}(\zeta,\tau)$ satisfies the*

following differential equation and initial condition:

$$\frac{\partial}{\partial\tau}\mathbf{X}(\zeta,\tau) = \left(\mathcal{Q}^{\star}_{video} - \zeta r_{video}\mathcal{B}_{video}\right)\mathbf{X}(\zeta,\tau), \quad (7)$$

$$\mathbf{X}(\zeta,0) = \mathbf{1}, \quad (8)$$

and hence the unique solution is given by

$$\mathbf{X}(\zeta,\tau) = \exp\left\{\tau\left(\mathcal{Q}^{\star}_{video} - \zeta r_{video}\mathcal{B}_{video}\right)\right\}\mathbf{1}. \quad (9)$$

PROOF. Consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the tagged video call cannot terminate within this time. Condition on all the possible events occurring in this interval, starting out in state $(s,v,d) \in \mathbb{S}^{+}_{\text{video}}$ (for notational convenience and readability, the boundary constraints are not explicitly considered):

$$X_{s,v,d}(\zeta,\tau) \equiv \mathbf{E}\left\{\exp\left\{-\zeta\, x_{s,v,d}(\tau)\right\}\right\}$$

$$\begin{aligned}
= \;& \lambda_{\text{speech}}\Delta X_{s+1,v,d}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s+1,v,d)\,O(\Delta)\end{array}\right)\right] \\
&+ s\mu_{\text{speech}}\Delta X_{s-1,v,d}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s-1,v,d)\,O(\Delta)\end{array}\right)\right] \\
&+ \lambda_{\text{video}}\Delta X_{s,v+1,d}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s,v+1,d)\,O(\Delta)\end{array}\right)\right] \\
&+ (v-1)\,\mu_{\text{video}}\Delta X_{s,v-1,d}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s,v-1,d)\,O(\Delta)\end{array}\right)\right] \\
&+ \lambda_{\text{data}}\Delta X_{s,v,d+1}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s,v,d+1)\,O(\Delta)\end{array}\right)\right] \\
&+ d\beta_{\text{data}}(s,v,d)\,\mu_{\text{data}}\Delta X_{s,v,d-1}(\zeta,\tau-\Delta) \\
&\quad \times \exp\left[-\zeta r_{\text{video}}\left(\begin{array}{c}\beta_{\text{video}}(s,v,d)\left(\Delta - O(\Delta)\right)\\ +\beta_{\text{video}}(s,v,d-1)\,O(\Delta)\end{array}\right)\right] \\
&+ \left(\begin{array}{c}-\lambda_{\text{speech}}\Delta - s\mu_{\text{speech}}\Delta - \lambda_{\text{video}}\Delta - \\ (v-1)\,\mu_{\text{video}}\Delta - \lambda_{\text{data}}\Delta - d\beta_{\text{data}}(s,v,d)\mu_{\text{data}}\Delta\end{array}\right) \\
&\quad \times X_{s,v,d}(\zeta,\tau-\Delta)\exp\left[-\zeta r_{\text{video}}\beta_{\text{video}}(s,v,d)\,\Delta\right] \\
&+ \left(\begin{array}{c}1-\zeta r_{\text{video}}\beta_{\text{video}}(s,v,d)\,\Delta \\ +\sum_{j=2}^{\infty}\frac{(-\zeta r_{\text{video}}\beta_{\text{video}}(s,v,d)\Delta)^{j}}{j!}\end{array}\right) X_{s,v,d}(\zeta,\tau-\Delta) \\
&+ o(\Delta).
\end{aligned}$$

Rearranging terms, letting $\Delta \downarrow 0$ and writing the resulting system of differential equations in matrix notation yields expression (7). Initial condition (8) simply reflects the fact that the transfer volume $x_{s,v,d}(0)$ of a video call with a duration of zero seconds equals zero bits. In order to prove that the system of differential equations (7) with initial condition (8) has a *unique* solution, note that it is a system of the form $\frac{\partial}{\partial\tau}\mathbf{X}(\zeta,\tau) = A\,\mathbf{X}(\zeta,\tau) \equiv f(\mathbf{X}(\zeta,\tau))$ where f is a linear function with continuous partial derivatives with respect to the entries of its argument vector. The existence and uniqueness of a solution $\mathbf{X}(\zeta,\tau)$ for every initial vector, immediately follows from e.g. [4, Chapter 1, Section 8]. To conclude the proof, it is readily verified that the claimed solution (9) indeed satisfies the system of differential equations (7) with initial condition (8). □

Using closed-form expression (9) for the Laplace-Stieltjes transform of the distribution of $x_{s,v,d}(\tau)$, Theorem 1 follows as a corollary to Lemma 1, as proven below.

PROOF. The existence of a vector $\boldsymbol{\gamma}_{\text{video}}$ that satisfies (3) and its uniqueness up to a translation along the vector $\mathbf{1}$, are guaranteed by results in Markov decision theory. Interpreting $\boldsymbol{\gamma}_{\text{video}}$ as the vector of relative values in a Markov reward chain governed by the generator $\mathcal{Q}^{\star}_{\text{video}}$ and with immediate cost vector $\frac{1}{\eta}\left(r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right)$ where η is the maximum rate of change in the Markov chain, and understanding that the long-term average costs are zero, $\frac{1}{\eta}\boldsymbol{\pi}^{\star}_{\text{video}}\left(r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right) = 0$, e.g. [15, Theorem 3.1, page 167] can be directly applied after uniformization of the continuous-time Markov chain. Hence in (3) a single degree of freedom exists in choosing $\boldsymbol{\gamma}_{\text{video}}$, which is used to normalise $\boldsymbol{\gamma}_{\text{video}}$ as in (4).

The vector of conditional expected transfer volumes $\widehat{\mathbf{x}}(\tau)$ is then obtained by taking the derivative of $\mathbf{X}(\zeta,\tau)$ with respect to ζ, and subsequently setting $\zeta = 0$.

$$\begin{aligned}
\widehat{\mathbf{x}}(\tau) &= -\frac{\partial}{\partial\zeta}\mathbf{X}(\zeta,\tau)\Big|_{\zeta=0} \\
&= -\frac{\partial}{\partial\zeta}\sum_{k=0}^{\infty}\frac{\left(\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right) + \left(-\zeta\,\tau r_{\text{video}}\mathcal{B}_{\text{video}}\right)\right)^{k}}{k!}\mathbf{1}\Big|_{\zeta=0} \\
&= -\left(\sum_{k=1}^{\infty}\sum_{i=0}^{k-1}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{v}}\right)^{k-i-1}\left(-\tau r_{\text{v}}\mathcal{B}_{\text{v}}\right)\left(\tau\mathcal{Q}^{\star}_{\text{v}}\right)^{i}}{k!}\right)\mathbf{1} \\
&= \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \\
&\quad \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\left[\begin{array}{c}\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} - \\ \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\end{array}\right] \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} - \\
&\quad \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau\mathcal{Q}^{\star}_{\text{video}}\boldsymbol{\gamma}_{\text{video}} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{v}}r_{\text{v}}\mathcal{B}_{\text{v}}\mathbf{1}\right)\mathbf{1} + \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{v}}\right\}\right]\boldsymbol{\gamma}_{\text{v}}
\end{aligned}$$

where after the third equality sign only those matrix cross-products appear that remain after differentiating the terms in the preceding expression, and setting ζ to 0. The subsequent equality sign uses $\mathcal{Q}^{\star}_{\text{video}}\mathbf{1} = 0$, so that all terms with $i > 0$ disappear. A similar argument is used to obtain the fifth equality. Equation (3) is used for the sixth equality.

With regards to the asymptotic expressions, note that since $\mathcal{Q}^{\star}_{\text{video}}$ is the generator of an irreducible finite state space Markov chain, with equilibrium distribution vector $\boldsymbol{\pi}^{\star}_{\text{video}}$, it holds that $\lim_{\tau\to\infty}\exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\} = \mathbf{1}\,\boldsymbol{\pi}^{\star}_{\text{video}}$, and thus $\lim_{\tau\to\infty}\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}} = \boldsymbol{\gamma}_{\text{video}}$, using (4), while $\lim_{\tau\to\infty}\tau^{-1}\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}} = \mathbf{0}$. □

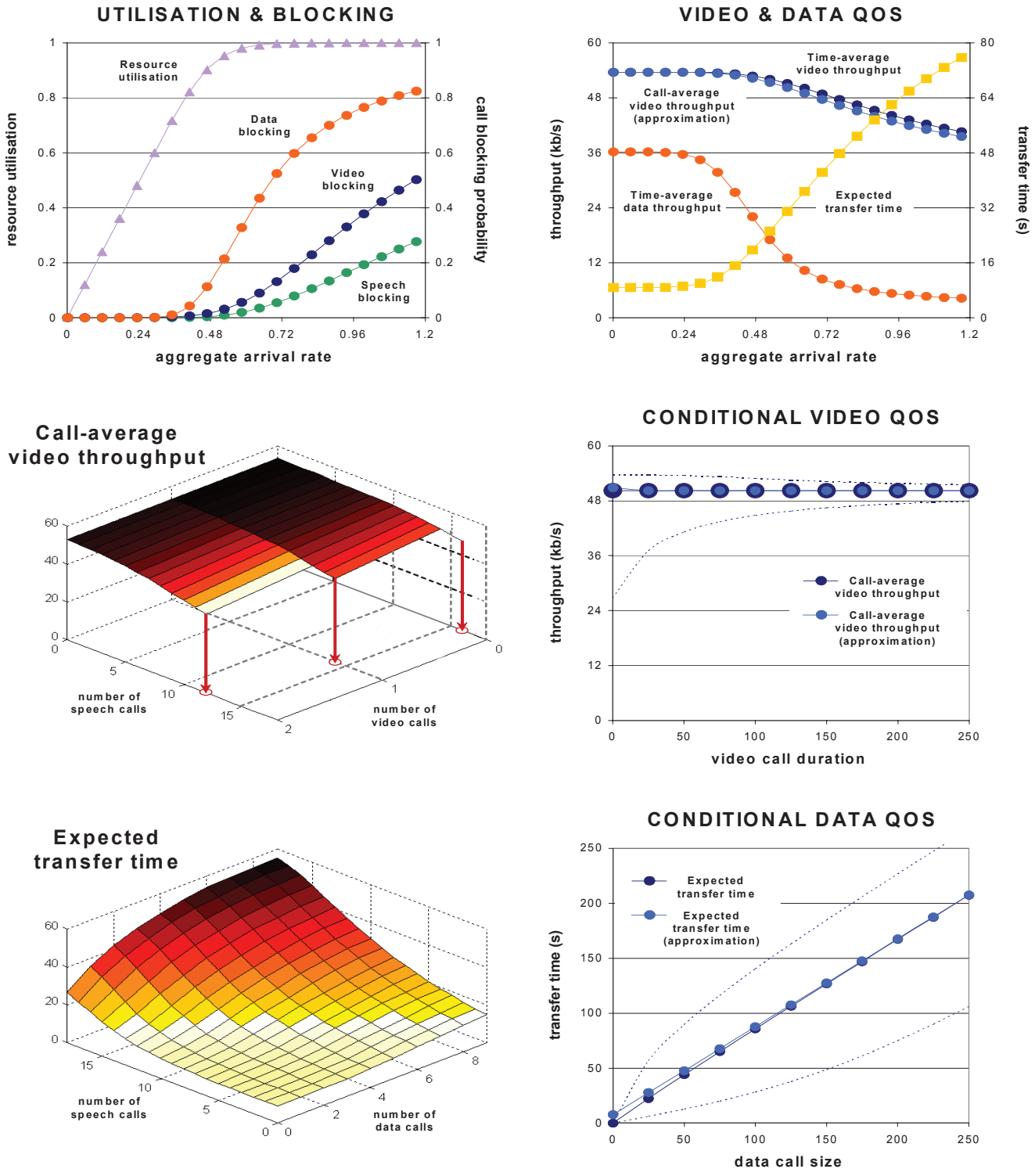

Figure 2: Results of some illustrative numerical experiments. For a given scenario, the upper two charts depict some basic performance measures as a function of the aggregate traffic load. The middle and lower pairs of charts concentrate on some conditional video and data QOS measures, respectively.