GPRS session time distribution

Problem presented by

Tunde Williams

Motorola Research

Problem statement

GPRS is a recent standard for data communications with GSM mobile phones, which uses the same infrastructure as traditional voice calls, but which has different queueing and delay properties. The Study Group was asked to construct a model for the simultaneous transmission of voice and data and to determine what level of equipment is necessary to provide a required grade of service.

Study Group contributors

David Allwright (Smith Institute)
Sean Collins (Universty of Bristol)
Zhaohui Guo (Bao Steel)
Rob Hinch (University of Oxford)
Sam Howison (University of Oxford)
Robert Leese (Smith Institute)
Mark Muldoon (UMIST)
Chris Williams (University of Bristol)

Report prepared by

David Allwright (Smith Institute)
Rob Hinch (University of Oxford)

1 Problem description

1.1 Circulated description

GPRS is a recent standard for data communications with GSM mobile phones, which uses the same infrastructure as traditional voice calls, but which has different queueing and delay properties. Under the GSM standard, each local antenna has a number of *channels* (usually a multiple of 8). In the case of voice calls, if a mobile phone attempts to send a *block* (which represents a small segment of speech) to the local antenna, then several things can happen:

(a) The block is successfully transmitted through an available channel.

(b) There is an error in transmission and the local antenna does not successfully receive the block.

(c) There are no available channels on the local antenna.

In either case (b) or (c), the block is dropped. This is not a significant issue for voice calls since the loss of some blocks leads only to a gradual break-up of the call. If we assume that:

- voice calls are made to the local antenna as a Poisson process with a given rate,
- call times are exponentially distributed with a given mean, and
- a given number N of channels are available,

then the probability of a block being dropped can be worked out by the well-known *Erlang-B* formula. Hence the number of channels required to keep the block dropping rate below a given threshold can be calculated.

For transmitting *data*, the problem is more complex, because it is essential that all blocks get through: if there is a transmission error, then the block must be re-sent. (Thus if the block transmission error rate is β, then to send, say, 8kbytes in 20-byte blocks will take on average $400/(1-\beta)$ block transmissions.) Further, in the case where there is no transmission error for a given GPRS block, then either

(a) the antenna has one *or more* channels to allocate to that block, or

(b) all of the antenna's channels are busy, in which case the block is *queued*.

The total time perceived by the user is thus the transmission time plus the queueing time.

Let us now consider the channel allocation process in more detail, and assume for simplicity that the local antenna has 8 channels. Some calls (*e.g.* voice) will use only one channel, but for GPRS data transmission some users have phones that can use 2 or more (up to 8) channels simultaneously.

For example, suppose a user whose phone can use up to 4 channels makes a data transfer request, and that there are F free channels available. If $F = 0$ the call is queued; otherwise the call is allocated $\min(4, F)$ of the available channels. Compared

with using just 1 channel, this speeds up his transfer if $F \geq 2$, but it makes a considerable complication to the queueing process. The essence of the problem proposed to the Study Group is to develop the queueing theory for this kind of multichannel use, so that the distribution of queueing and transmission times can be calculated.

The assumptions to be made are certain mean call rates for the different kinds of call (parametrised by the number of data blocks) and the number of user channels, with each call stream assumed to be an independent Poisson process.

One could start by modelling the state of (say) an 8-channel antenna as a continuous-time Markov chain with a state space consisting of the partitions of all integers $n \leq 8$. Each partition of 8 would correspond to a state in which all the channels were in use, with each block of the partition corresponding to a single call, so that for instance $4+2+1+1$ would correspond to a state where one user had 4 channels, one had 2, and two each had 1 channel. Each partition of a number less than 8 would correspond to a state where some of the channels were free. The state space would therefore be of dimension $p(0) + p(1) + \ldots + p(8) = 1 + 1 + 2 + 3 + 5 + 7 + 11 + 15 + 22 = 67$ and it would not be too difficult to set up the transition state matrix, find the invariant measure *etc.*. If an antenna had $n > 8$ channels, the complexity of this procedure would grow like $p(n) \sim 1/(4n\sqrt{3})\exp(\pi\sqrt{2n/3})$, or alternatively if a call cannot be split between 8-channel groups, then, *e.g.*, a 16-channel antenna would have a state space of dimension 67^2 which is computationally tractable.

1.2 Study Group presentation

In Tunde Williams' initial presentation to the Study Group, he explained that the aim of the research is to be able to determine what level of equipment is necessary to provide a required grade of service. If voice calls arrive at a rate λ_V, then the number N of channels is determined so that the proportion of blocked voice calls is below some threshold (*e.g.* 2%). Data calls arrive at some rate λ_D, and have a distribution of sizes, generally of order tens of kilobytes. Each call is broken up into 20-byte blocks which are transmitted (subject to the block error rate β) or are queued if no free channel is available.

One distinctive point of the process is that *voice* calls have priority over *data*, so if a voice call arrives when all channels are occupied, some of them by data calls, then one of the data calls loses a channel to the new voice call.

The inputs to the process are

- The voice call arrival rate λ_V.
- The data call arrival rate λ_D.
- The mean voice call hold time (typically 120 s).
- The block error rate (β) distribution.
- The mean data call length (in bytes).
- The (maximum) number of bytes in a transmission block (typically 20).
- The time to transmit one block (typically 0.02 s).

The required outputs are

- The data call time probability distribution for each user type (considered as a sum of queueing and transmission times).
- The mean data call queue length.
- The average time in the queue for data calls.
- The average number of data calls in the system.
- The mean number of busy channels.

Motorola Research would like the model to be analytical as far as possible, and are happy with the following assumptions:

- The data block size is constant.
- Calls arrive according to a Poisson process.
- The voice call time is exponentially distributed.
- The data queue is potentially infinite. (In practice the buffer size can be provided as large as necessary.)
- Suppose a data call from a 4-channel caller is using 2 channels and 1 or 2 further channels become available. It can either be assumed that the call stays with 2 channels (constant allocation), or that it takes up the newly available ones (dynamic allocation).

2 Basics of the model

2.1 Voice calls

The model for voice calls is unambiguous: they arrive as a Poisson process with rate λ_V, and the call lengths are independent exponential random variables, with parameter which we call μ_V (so the mean call time is $1/\mu_V$). Since the voice calls take priority over data calls, the whole behaviour of the voice call process is standard: the number of voice calls in progress, say k, is a continuous time Markov chain with states $k = 0, 1, \ldots, N$ and transition rates represented by the diagram in Figure 1. (We explain our terminology

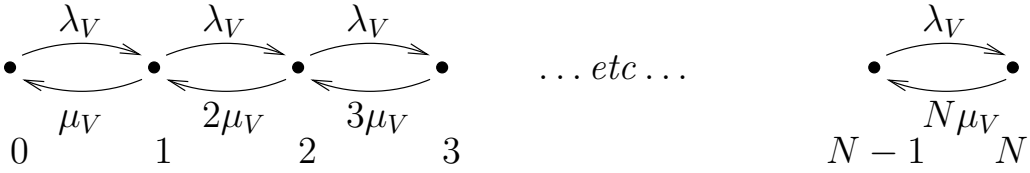

Figure 1: Transition rate diagram for voice calls.

and notation for continuous time Markov chains in the Appendix.) The steady state distribution $p_0, \ldots, p_N$ is the truncated Poisson distribution

$$p_k = \frac{\rho_V^k/k!}{\sum_{j=0}^{N} \rho_V^j/j!}, \tag{1}$$

where $\rho_V = \lambda_V/\mu_V$. A voice call arriving when the system is in state N is blocked, so the steady state blocking probability is just the probability p_N that the system is in state N. The mean number of voice calls in the system is $\overline{V} = \rho_V(1 - p_N)$. When p_N is small, the denominator of (1) is approximately $\exp(\rho_V)$, the distribution is approximately Poisson, and $\overline{V} \approx \rho_V$.

2.2 Data calls

We shall assume that data calls arrive as a Poisson process with rate λ_D. The Poisson process model is obviously questionable: often a request from a user, say to download a web page, will result in several data calls (for the text, the images, and other items needed for the full page). Hence the data call process is more clumpy than a Poisson process. One way to model this kind of clumpiness mathematically is to let the data call rate λ_D be itself a random process. However in this report we shall not consider that, but will stick to the simple case of a Poisson data call process.

We need some model of data call size, and the given information from Motorola Research only indicates a given *mean* data call size. Since the mean data call size is of order tens of kilobytes, much larger than the block size (20 bytes), we shall make the simplification of treating the data call size as a *continuous* random variable. Furthermore, again for simplicity we shall assume it is exponentially distributed with a mean $\overline{D}$ bytes. On a single perfect channel with block size b, and transmission time t_b for a single block, this makes the data call time exponentially distributed with parameter $b/(\overline{D}t_b)$. When we allow for the block error rate β, this is reduced to

$$\mu_D = \frac{(1-\beta)b}{\overline{D}t_b}, \tag{2}$$

corresponding to a mean transmission time of $(\overline{D}/b)t_b/(1-\beta)$. If a data call is serviced by r channels instead of a single channel, then the time taken will be exponential with parameter $r\mu_D$.

We must make some assumption about what happens when a data call begins transmission and then is pushed back into the queue by a voice call: do the previously sent blocks have to be retransmitted or not ? We shall here assume that they do *not* have to be retransmitted, chiefly because this simplifies the modelling, in that the time remaining for that data call is then still exponential with parameter μ_D (if it is sent on a single channel). (In the alternative case where the previously sent blocks *do* have to be retransmitted, calls that had got pushed back into the queue would have to be tagged in the model with a conditional distribution because they are already known to be of at least a certain length. It seems best for the initial studies to avoid this complication.)

2.3 Systems with voice and data

In general terms, our models for systems with both voice and data calls are going to be continuous-time Markov chains, with a state that specifies *both* the number of voice calls in the system, k, *and* the number of data calls in the system, l. Of course, with multichannel calls the state will be more complicated than that, but at a minimum the

states need to be indexed by a pair of integers (k, l). In the subsequent sections we consider first two extreme cases:

(A) Each data call can only use 1 channel.

(B) Each data call can use N channels, with dynamic allocation.

Then in section 6 we go on to deal with variable numbers of channels, in the case of constant allocation, setting up the transition rates and computing steady state distributions numerically.

3 Case A: Data calls use only 1 channel

As described above, we consider the case in which each data call uses a single channel, and we index the state space of the system by (k, l) where k and l are the numbers of voice calls and data calls in the system. We have $0 \leq k \leq N$, and a voice call arriving when $k = N$ is blocked. We have $l \geq 0$, and when $l \leq N - k$ all of the l data calls are being processed, but when $l > N - k$ then only $N - k$ data calls are being processed and the rest are queued. If $k < N$ and $l \geq N - k$ and a voice call arrives, then as described earlier, the voice call takes priority over one of the data calls, which is queued, and it is assumed that a queued data call will not need to resend data which has already been transmitted. The resulting model is a continuous time Markov chain with transition rates

$$\begin{array}{lcllll} (k,l) & \longrightarrow & (k+1,l) & \text{with rate} & \lambda_V, & \text{when} \quad k < N, \\ (k,l) & \longrightarrow & (k-1,l) & \text{with rate} & k\mu_V, & \\ (k,l) & \longrightarrow & (k,l+1) & \text{with rate} & \lambda_D, & \\ (k,l) & \longrightarrow & (k,l-1) & \text{with rate} & l\mu_D, & \text{when} \quad k+l \leq N, \\ & & & \text{and} & (N-k)\mu_D, & \text{when} \quad k+l > N. \end{array} \tag{3}$$

This Markov chain can easily be simulated using Monte Carlo simulations. We shall first present the exact balance equations that hold in the steady state, and then an asymptotic analysis which allows the approximate calculation of physically interesting quantities (*e.g.* the mean waiting time of data calls).

3.1 The balance equations

The stationary distribution of the chain, $p_{k,l}$, can be calculated by solving the full balance equations, which are a set of coupled difference equations

$$J_v\big(p_{k-1,l}, p_{k,l}, p_{k+1,l}\big) + J_d\big(p_{k,l-1}, p_{k,l}, p_{k,l+1}\big) = 0, \tag{4}$$

where

$$J_v = \begin{cases} -\lambda_V p_{k,l} + \mu_V (k+1) p_{k+1,l}, & k = 0, \\ \lambda_V \big(p_{k-1,l} - p_{k,l}\big) + \mu_V \big((k+1)p_{k+1,l} - k p_{k,l}\big), & 0 < k < N, \\ \lambda_V p_{k-1,l} - \mu_V k p_{k,l}, & k = N, \end{cases}$$

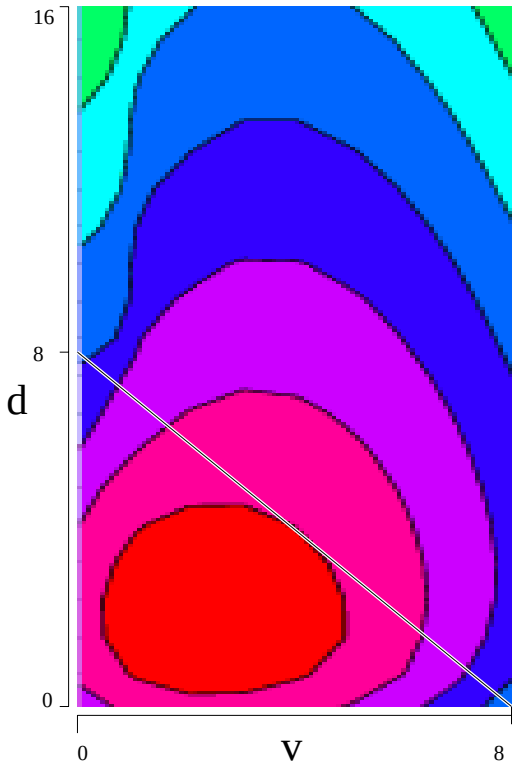

Figure 2: Stationary distribution of the model. The probability is denoted by colour (red signifies high probability, blue signifies low probability) and is a logarithmic scale. The straight line is $k+l=N$, and above the line some of the data calls will be queued, while below the line all the data calls are served. The model parameters used were: $N=8$, $\lambda_v=\lambda_V/N=0.3$, $\mu_v=\mu_V=1$, $\lambda_d=\lambda_D/N=0.3$ and $\mu_d=\mu_D=1$ (see equation (6)).

and

$$J_d=\begin{cases} -\lambda_D p_{k,l}+\mu_D(l+1)p_{k,l+1}, & l=0,k<N,\\ -\lambda_D p_{k,l}, & l=0,k=N,\\ \lambda_D\big(p_{k,l-1}-p_{k,l}\big)+\mu_D\big((l+1)p_{k,l+1}-lp_{k,l}\big), & l>0,k+l<N,\\ \lambda_D\big(p_{k,l-1}-p_{k,l}\big)+\mu_D(N-k)\big(p_{k,l+1}-p_{k,l}\big), & l>0,k+l\geq N. \end{cases}$$

Here J_v and J_d represent the total probability fluxes in the voice and data directions. The full balance equations can be solved numerically by finding the eigenvector with eigenvalue 1. A typical density plot for $p_{k,l}$ is shown in Figure 2. In the following section we shall consider their approximate solution in the limit of large N.

3.2 Large N limit

The mobile phone industry aims to provide sufficient coverage such that the probability of a voice call being blocked is less than 2%. Therefore we shall make the assumption that the probability of a voice call being blocked or a data call being queued is small. Additionally we shall assume that the number of channels N at each transmitter is large, and consider there to be a smooth function $p(v,d)$ of continuous variables v, d, which

agrees with the discrete values $p_{k,l}$ at the integer points. This then allows difference terms to be expanded about $p(v,d)$:

$$\begin{aligned} p_{v\pm1,d} &= p(v,d) \pm \frac{\partial p}{\partial v} + \frac{1}{2}\frac{\partial^2 p}{\partial v^2} + \text{higher order terms}, \\ p_{v,d\pm1} &= p(v,d) \pm \frac{\partial p}{\partial d} + \frac{1}{2}\frac{\partial^2 p}{\partial d^2} + \text{higher order terms}. \end{aligned} \tag{5}$$

We shall also make the definitions

$$\lambda_v = \lambda_V/N, \quad \lambda_d = \lambda_D/N, \quad \mu_v = \mu_V, \quad \mu_d = \mu_D, \tag{6}$$

and we consider the limit of large N in which the quantities λ_v, μ_v, λ_d, and μ_d are held fixed: this in effect means that as N increases, data calls and voice calls are considered as occupying fixed *proportions* of the capacity N. Inserting these expansions into the equations for J_v and J_d in the region $v + d < N$ yields, at leading order,

$$\begin{aligned} J_v &= (\mu_v v - N\lambda_v)\frac{\partial p}{\partial v} + (\mu_v v + N\lambda_v)\frac{1}{2}\frac{\partial^2 p}{\partial v^2} + \mu_v\left(p + \frac{\partial p}{\partial v} + \frac{1}{2}\frac{\partial^2 p}{\partial v^2}\right) + \text{h.o.t.}, \\ J_d &= (\mu_d d - N\lambda_d)\frac{\partial p}{\partial d} + (\mu_d d + N\lambda_d)\frac{1}{2}\frac{\partial^2 p}{\partial d^2} + \mu_d\left(p + \frac{\partial p}{\partial d} + \frac{1}{2}\frac{\partial^2 p}{\partial d^2}\right) + \text{h.o.t.}. \end{aligned} \tag{7}$$

In the limit $N \to \infty$ there is an inner region with width of order $\sqrt{N}$ centred on the maximum of the distribution. This can be shown by introducing the co-ordinates

$$v = N\rho_v + \sqrt{N\rho_v}\,x \quad \text{and} \quad d = N\rho_d + \sqrt{N\rho_d}\,y, \tag{8}$$

where

$$\rho_v = \frac{\lambda_v}{\mu_v} \quad \text{and} \quad \rho_d = \frac{\lambda_d}{\mu_d} \tag{9}$$

and the maximum of the distribution is at $(N\rho_v, N\rho_d)$. For this to lie in the region $v + d < N$ we must have $\rho_v + \rho_d < 1$. Inserting (8) into (7) yields

$$J_v = \mu_v\left(\frac{\partial^2 p}{\partial x^2} + x\frac{\partial p}{\partial x} + p\right) \quad \text{and} \quad J_d = \mu_d\left(\frac{\partial^2 p}{\partial y^2} + y\frac{\partial p}{\partial y} + p\right). \tag{10}$$

The full balance equation ($J_v + J_d = 0$) can now be solved using the separation of variables

$$p(x,y) = X(x)Y(y). \tag{11}$$

The separated full balance equations become

$$\frac{d^2X}{dx^2} + x\frac{dX}{dx} + (1+k)X = 0 \quad \text{and} \quad \frac{d^2Y}{dy^2} + y\frac{dY}{dy} + (1-k\alpha)Y = 0, \tag{12}$$

where k is the separation constant and $\alpha = \mu_v/\mu_d > 0$. The solutions for X and Y must decay in the limits that $x, y \to \pm\infty$ and be positive for all values of x and y. By considering the behaviour of equation (12) in these limits we find that

$$X \sim \exp\left(-\frac{x^2}{2}\right) \quad \text{and} \quad Y \sim \exp\left(-\frac{y^2}{2}\right), \tag{13}$$

(where $\sim$ is to be interpreted in a loose sense). Next, introduce the functions $f(x)$ and $g(x)$ by writing

$$X = \exp\left(-\frac{x^2}{2}\right) f(x) \quad \text{and} \quad Y = \exp\left(-\frac{y^2}{2}\right) g(x). \tag{14}$$

Inserting into the separated solutions (12), expanding and cancelling terms yields

$$f'' - xf' + kf = 0 \quad \text{and} \quad g'' - yg' - \alpha k g = 0, \tag{15}$$

where $' \equiv d/dx$. These equations can be solved by inserting a series solution

$$f = \sum_{j=0}^{\infty} a_j x^{j+s} \quad \text{and} \quad g = \sum_{j=0}^{\infty} b_j y^{j+t}. \tag{16}$$

Equating powers of x and y respectively we get $s = 0, 1$ and $t = 0, 1$, and

$$a_j = \frac{(j+s-2)-k}{(j+s)(j+s-1)} a_{j-2} \quad \text{and} \quad b_j = \frac{(j+t-2)+k\alpha}{(j+t)(j+t-1)} b_{j-2}. \tag{17}$$

By considering the large order terms of these sums we see that the series solution diverges like $\exp(x^2/2)$ and $\exp(y^2/2)$ respectively. This cannot be allowed to happen because the solution of $X(x)$ and $Y(y)$ must decay at large x and y. Therefore the series for a_j and b_j must terminate at finite j.[1] This will occur when $k = m$ and $k\alpha = -n$ where $m, n = \{0, 1, \ldots\}$, so

$$m\alpha = -n. \tag{18}$$

However, since $\alpha > 0$ this tells us that $m = n = k = 0$. This result is important because it tells us that both the partial balances $J_v = 0$ and $J_d = 0$ are satisfied. Additionally if $J_v = 0$ and $J_d = 0$ then the system will be in detailed balance where the net flux of probability between neighbouring sites is 0. Figure 3 shows plots of both the absolute and relative next neighbour fluxes of the stationary distribution when $N = 8$ (calculated numerically). The absolute fluxes on the left show an interesting structure: when $v < \overline{V}$ the number of data calls tends to fall, and when $v > \overline{V}$ it tends to rise. Thus there is an anticlockwise circulation in the figure, representing the typical cycle of fluctuations in data and voice calls.

In the region of maximum probability ($v \approx 2.4$ and $d \approx 2.4$; see Figure 3) the relative nearest neighbour fluxes are small, telling us that the system is in approximate detailed balance. The stationary distribution in the region of $(N\rho_v, N\rho_d)$ is then

$$p(x, y) \propto \exp\left(-\frac{x^2 + y^2}{2}\right), \tag{19}$$

which is the product of two independent Gaussians.

The analysis in this section has shown that in the limit of large N and in the region of the maximum of the probability density function, the voice and data calls can be treated

[1] This conclusion can alternatively be reached by transforming the original equation for X into the 1-dimensional quantum harmonic oscillator and using the known eigenvalues of that.

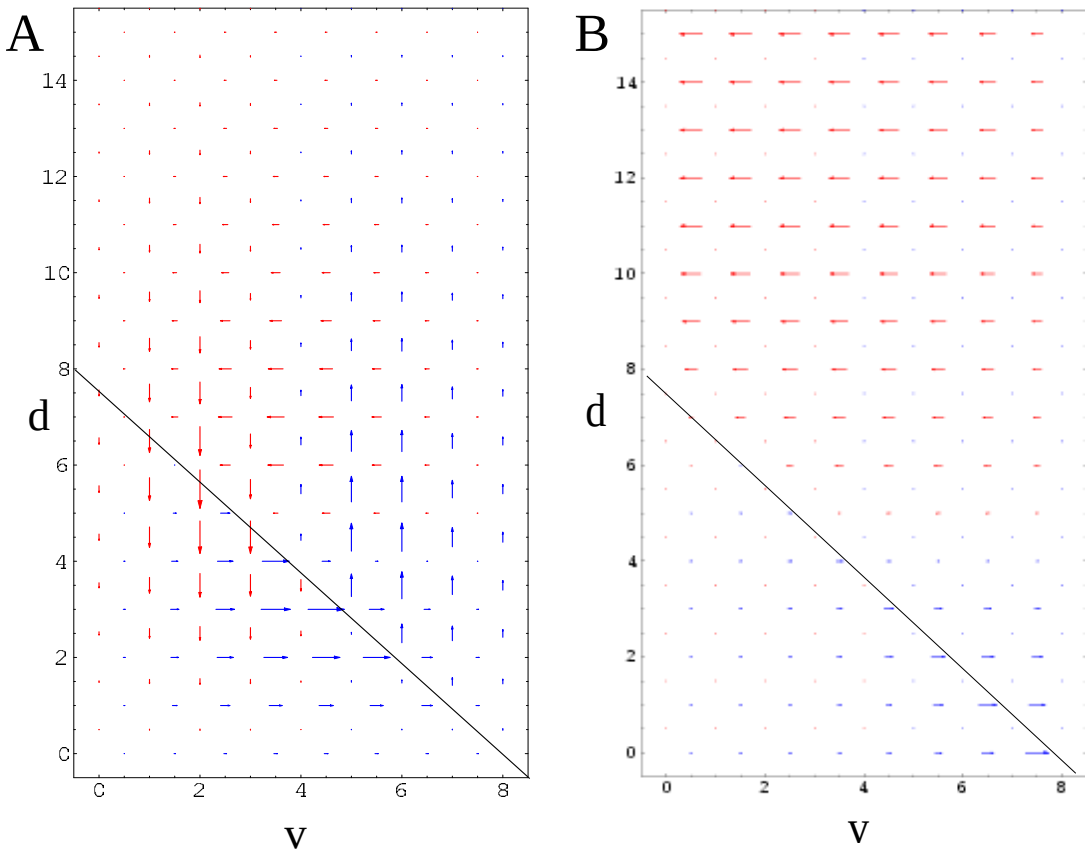

Figure 3: A. The absolute nearest neighbour fluxes of the stationary distribution. B. The relative nearest neighbour fluxes of the stationary distribution. The length of the arrow represents the size of the nearest neighbour flux. Note that in the region of maximum probability ($v \approx 2.4$ and $d \approx 2.4$, see Figure 2) the relative nearest neighbour fluxes are very small, telling us that the system is in approximate detailed balance at these points. The model parameters used were: $N = 8$, $\lambda_v = 0.3$, $\mu_v = 1$, $\lambda_d = 0.3$ and $\mu_d = 1$.

independently (to leading order) and that the boundary effects are not important. This suggests that in the limit of large N and when $v + d \leq N$, we can approximate $p_{v,d}$ by

$$p_{v,d} \approx p_v p_d, \tag{20}$$

where p_v and p_d are the probability density functions for the number of voice and data calls when only one type of call is allowed. In the limit of large N both p_v and p_d are approximate Poisson variables (equation (1)), so

$$\begin{aligned} p_{v,d} &\approx \frac{(N\rho_v)^v (N\rho_d)^d e^{-N(\rho_v+\rho_d)}}{d!v!} \\ &= \frac{(N(\rho_v+\rho_d))^{v+d} e^{-N(\rho_v+\rho_d)}}{(v+d)!} \begin{pmatrix} v+d \\ v \end{pmatrix} \left(\frac{\rho_v}{\rho_v+\rho_d}\right)^v \left(\frac{\rho_d}{\rho_v+\rho_d}\right)^d. \end{aligned} \tag{21}$$

Note that this can now be re-expressed as a Poisson variable multiplied by Binomial variable. Define $P(\text{Line})$ as the probability that all the channels are being used and no data calls are being queued (*i.e.* $v + d = N$). Then

$$P(\text{Line}) = \sum_{v=0}^{N} p_{v,N-v} \approx \frac{(N(\rho_v+\rho_d))^N e^{-N(\rho_v+\rho_d)}}{N!}. \tag{22}$$

The distribution for $p_{v,d}$ on the line $v + d = N$ can then be approximated using the Normal approximation of the Binomial distribution in the limit $N \to \infty$ and Stirling's formula to give

$$p_{v,N-v} \approx p_{\max} \exp\left(-\frac{N(\rho_v+\rho_d)^2}{2\rho_v\rho_d}\left(\frac{v}{N} - \frac{\rho_v}{\rho_v+\rho_d}\right)^2\right), \tag{23}$$

where

$$p_{\max} = (\rho_v+\rho_d)^{N+1} \frac{e^{N(1-\rho_v-\rho_d)}}{2\pi N\sqrt{\rho_v\rho_d}}. \tag{24}$$

Note that the maximum of this distribution occurs at (v^*, d^*), where

$$v^* = N\frac{\rho_v}{\rho_v+\rho_d} \quad \text{and} \quad d^* = N\frac{\rho_d}{\rho_v+\rho_d}, \tag{25}$$

and the width of the distribution is $O(\sqrt{N})$. This approximation for the distribution of $p_{v,N-v}$ can be compared with numerical results calculated using a Monte Carlo simulation (Figure 4). The analysis and simulation are in excellent agreement suggesting that the approximation that v and d can be treated separately is valid. The value of $p_{\max}$ can be compared with Monte Carlo simulations for a range of value of N (Figure 5). Note that even when $N = 8$ the analysis and simulation are in excellent agreement.

The final part of the analysis is to consider $p_{v,d}$ when $v + d > N$. This can be achieved by expanding the full balance equation (4) in the region of the maximum in the probability (v^*, d^*). The boundary condition is given by $p_{v,N-v}$ which has its maximum at $v = v^*$ and decays over a region of width $O(\sqrt{N})$ in v. This suggests we should consider new variables of the form (see Figure 6).

$$v = v^* + \sqrt{N}m \quad \text{and} \quad d = d^* - \sqrt{N}m + n. \tag{26}$$

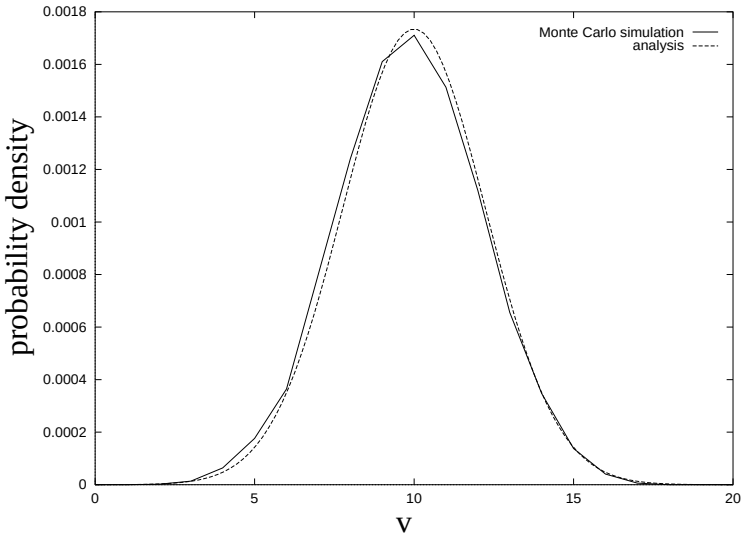

Figure 4: The probability density function $p_{v,d}$ on the line $v + d = N$ calculated using a Monte Carlo simulation and equation (23). The model parameters used were $N = 20$, $\lambda_v = 0.3$, $\mu_v = 1$, $\lambda_d = 0.3$ and $\mu_d = 1$.

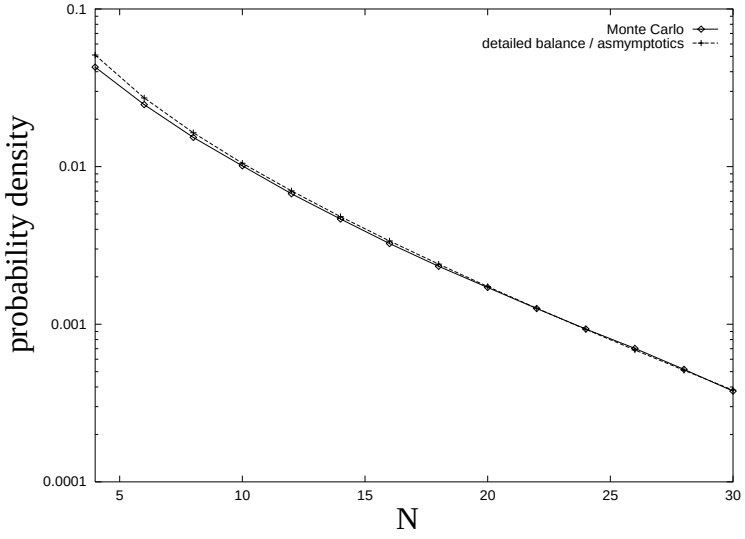

Figure 5: The maximum probability density $p_{\max}$ on the line $v + d = N$ calculated using a Monte Carlo simulation and equation (24). The model parameters used were $\lambda_v = 0.3$, $\mu_v = 1$, $\lambda_d = 0.3$ and $\mu_d = 1$. Note that even when $N = 8$ (the value of interest to Motorola) the analysis and simulation are in excellent agreement.

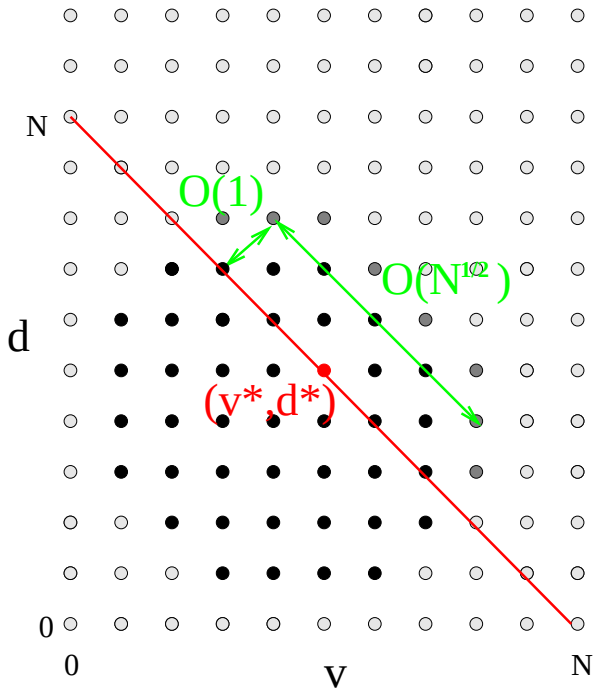

Figure 6: The probability density function in the region of (v^*, d^*). The probability density is indicated by the shade of the spot (high density = black; low density = white). The line $v + d = N$ is shown is red. The width of $p_{v,d}$ is $O(N^{1/2})$ in the direction of the line $v + d = N$, and $O(1)$ in the direction perpendicular to the line $v + d = N$.

Defining $q_n(m) = p_{v,d}$, inserting the rescaled variables into the full balance equation (4) and expanding in the limit $N \to \infty$ gives

$$\begin{aligned} J_v &= O(\sqrt{N}), \\ J_d &= \frac{N\rho_d}{\rho_d + \rho_v}\big((\rho_v + \rho_d)(q_{n-1}(m) - q_n(m)) + (q_{n+1}(m) - q_n(m))\big) + O(\sqrt{N}). \end{aligned} \tag{27}$$

The boundary conditions are that $q_n(m)$ decays at large values of n and $q_0(m) = p_{v,N-v}$. Solving the full balance equations yields

$$q_n(m) = p_{\max} \exp\left(-\frac{m^2(\rho_v + \rho_d)^2}{2\rho_v\rho_d}\right)(\rho_v + \rho_d)^n . \tag{28}$$

3.3 Interesting quantities

In this section we use the stationary distribution (28) to calculate the quantities of interest to Motorola Research.

1. *New data calls being queued.* Define $P(\text{New queued})$ as the probability that a new data call will be blocked. This can be calculated from the stationary distribution

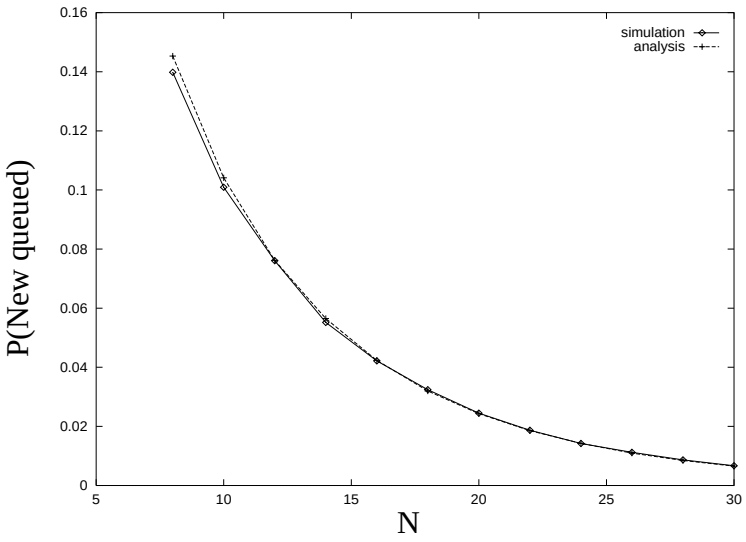

Figure 7: The probability of a new data call being blocked, P(New queued), as a function of the number of channels at the antenna. The graphs shows the value of P(New queued) obtained using a Monte Carlo simulation and expression (29). The other parameter values used were: $\lambda_v = 0.3$, $\mu_v = 1$, $\lambda_d = 0.3$ and $\mu_d = 1$.

by finding the probability that all the channels are being used (*i.e.* $v + d \geq N$).

$$\begin{aligned} P(\text{New queued}) &= \frac{1}{\sqrt{N}} \int_{-\infty}^{\infty} \sum_{n \geq 0} q_n(m)\, \mathrm{d}m, \\ &\approx \frac{e^{N(1-\rho_v-\rho_d)}(\rho_v + \rho_d)^N}{\sqrt{2\pi N}(1-\rho_v-\rho_d)}\ . \end{aligned} \tag{29}$$

This expression for P(New queued) can be compared with the value obtained from a Monte Carlo simulation of the model. Figure 7 shows the value of P(New queued) as a function of N with all other parameters fixed. The relative error of expression (29) compared with the numerical results is shown in Figure 8 and is small ($< 2\%$) even when $N = 8$. The GPRS standard is to have an antenna with 8 channels ($N = 8$). The probability that a new data call is queued is calculated as a function of the arrival rate of data calls λ_d (Figure 9). The asymptotic expression (29) for P(New queued) is a good approximation providing the value of P(New queued) is not too large ($> 25\%$).

2. *Expected number of data calls queued.* Define E(Data queued) as the expected number of data calls queued at the antenna. This is simply given by

$$\begin{aligned} E(\text{Data queued}) &= \frac{1}{\sqrt{N}} \int_{-\infty}^{\infty} \sum_{n \geq 0} n q_n(m)\, \mathrm{d}m, \\ &\approx \frac{e^{N(1-\rho_v-\rho_d)}(\rho_v + \rho_d)^N}{\sqrt{2\pi N}(1-\rho_v-\rho_d)^2}\ . \end{aligned} \tag{30}$$

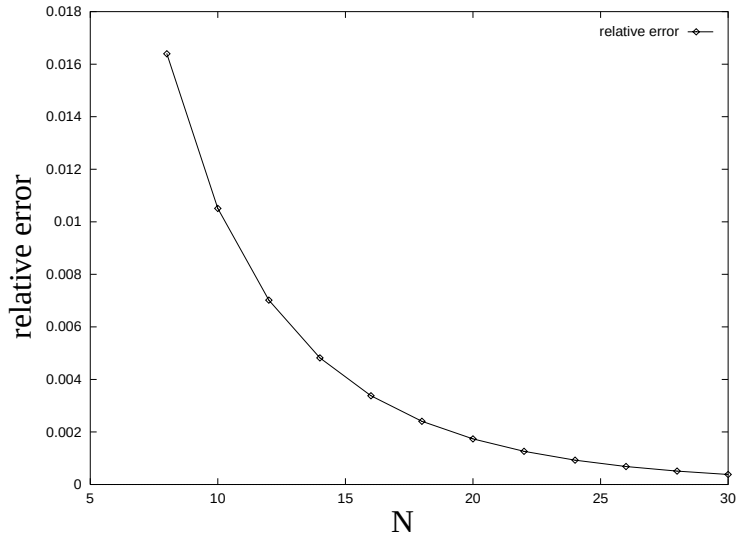

Figure 8: The difference between the numerically calculated value of P(New queued) and that obtained using expression (29). The graphs shows that even when $N = 8$ the relative error is less than 2%. The other parameter values used were: $\lambda_v = 0.3$, $\mu_v = 1$, $\lambda_d = 0.3$ and $\mu_v = 1$.

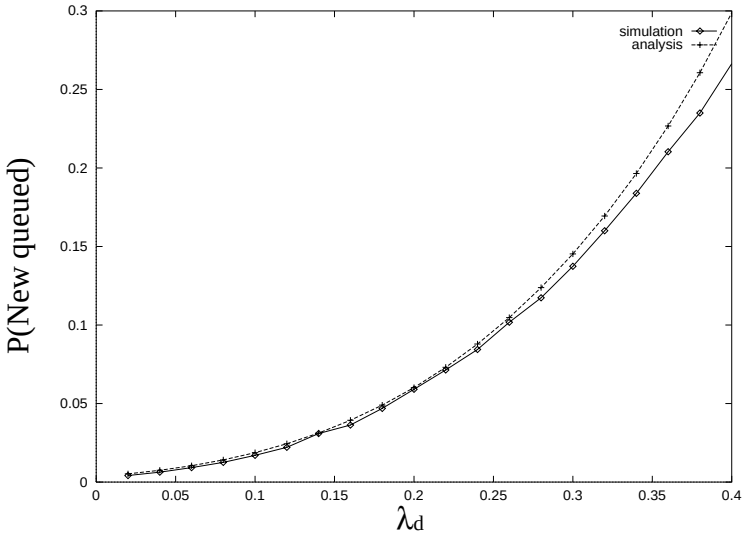

Figure 9: The probability of a new data call being blocked, P(New queued), as a function of the rate of arrival of data call, λ_d. The graphs shows the value of P(New queued) obtained using a Monte Carlo simulation and expression (29). The other parameter values used were: $N = 8$, $\lambda_v = 0.3$, $\mu_v = 1$ and $\mu_v = 1$.

3. *Proportion of data calls queued.* Define ϕ as the proportion of data calls queued. This is simply given by

$$\begin{aligned}\phi &= \frac{E(\text{Data queued})}{E(\text{Data calls})} \\ &\approx \frac{e^{N(1-\rho_v-\rho_d)}(\rho_v+\rho_d)^N}{\rho_d\sqrt{2\pi N}(1-\rho_v-\rho_d)^2}.\end{aligned} \tag{31}$$

4. *Mean queue time.* Define $E(\text{Queue time})$ as the expected amount of time a data call is queued. Since the length of a data call is independent of whether or not is queued, then

$$\begin{aligned}E(\text{Queue time}) &= \phi E(\text{Data length}) \\ &\approx \frac{e^{N(1-\rho_v-\rho_d)}(\rho_v+\rho_d)^N}{\mu_d\rho_d\sqrt{2\pi N}(1-\rho_v-\rho_d)^2}.\end{aligned} \tag{32}$$

5. *Maximum buffer required.* Define $E(B)$ as the expected amount of buffer required at the antenna to hold the queued data calls. Since the length of a data call is independent of whether it is being queued, then

$$E(B) = E(\text{Data queued})E(\text{Data length}). \tag{33}$$

Define $P(B > B_{\text{antenna}})$ as the probability that the buffer required exceeds the maximum buffer at the antenna. An upper bound can be put on this probability using Markov's inequality:

$$\begin{aligned}P(B > B_{\text{antenna}}) &\leq \frac{E(B)}{B_{\text{antenna}}} \\ &\leq \frac{e^{N(1-\rho_v-\rho_d)}(\rho_v+\rho_d)^N}{B_{\text{antenna}}\mu_d\sqrt{2\pi N}(1-\rho_v-\rho_d)^2}.\end{aligned} \tag{34}$$

4 Generalised large-N limit

In this section a general method of calculating the stationary distributions in the limit of large N is presented. The simple voice and single data call model will be analysed first to demonstrate the technique. The method can be extended to a model where the data calls can use multiple channels. The technique is similar to ray theory, which is used to calculate solutions of the Helmholtz equation in the limit of large wave number. The WKB ansatz is used to turn the difference equations of the full balance (4) into a nonlinear first-order PDE (in ray theory this would be the eikonal equation). Charpit's method is then used to find the characteristics. The maximum of the stationary distribution on the line $v + d = N$ is then found by solving a constrained maximisation problem.

4.1 Eikonal-type equation

The WKB ansatz will now be used to turn the full balance difference equations into a non-linear PDE. Define the rescaled variables

$$x = \frac{v}{N} \quad \text{and} \quad y = \frac{d}{N}. \tag{35}$$

The WKB ansatz is now used to look for a solution of the full balance equation of the form

$$p_{v,d} = R(x,y)\exp\big(Nu(x,y)\big). \tag{36}$$

The difference terms in the limit $N \to \infty$ are

$$\begin{aligned} p_{v\pm1,d} &= p_{v,d}\exp\left(\pm\frac{\partial u}{\partial x}\right)\left(1 \pm \frac{1}{N}\frac{\partial R}{\partial x} + \frac{1}{2N}\frac{\partial^2 u}{\partial x^2} + O(1/N^2)\right) \\ p_{v,d\pm1} &= p_{v,d}\exp\left(\pm\frac{\partial u}{\partial y}\right)\left(1 \pm \frac{1}{N}\frac{\partial R}{\partial y} + \frac{1}{2N}\frac{\partial^2 u}{\partial y^2} + O(1/N^2)\right). \end{aligned} \tag{37}$$

For notational convenience we shall introduce the standard shorthand $p = \partial u/\partial x$ and $q = \partial u/\partial y$. Insert these expansions into the full balance difference equation (4) when $v + d \le N$ and expand in the limit $N \to \infty$. The leading order terms are

$$\rho_v(e^{-p} - 1) + x(e^p - 1) + \alpha\big\{\rho_d(e^{-q} - 1) + y(e^q - 1)\big\} + O(1/N) = 0. \tag{38}$$

This equation is analogous to the eikonal equation in ray theory. The solution can be calculated along characteristics by using Charpit's method. Define t as the parameter along the characteristic, and s as the parameter of the initial data. Initial data is given on the curve (this is discussed in the next section)

$$\begin{gathered} x(s,0) = x_0(s), \quad y(s,0) = y_0(s), \quad p(s,0) = p_0(s), \\ q(s,0) = q_0(s), \quad u(s,0) = u_0(s). \end{gathered} \tag{39}$$

Charpit's method then tells us that the solution along the characteristics is found by solving the differential equations

$$\begin{aligned} \dot{x} &= xe^p - \rho_v e^{-p}, \\ \dot{y} &= \alpha\big(ye^q - \rho_d e^{-q}\big), \\ \dot{p} &= 1 - e^p, \\ \dot{q} &= \alpha\big(1 - e^q\big), \\ \dot{u} &= p\big(xe^p - \rho_v e^{-p}\big) + \alpha q\big(1 - e^{-p}\big). \end{aligned} \tag{40}$$

These equations can be solved sequentially. First solve for p and q

$$\begin{aligned} p &= -\ln\big[(e^{-p_0} - 1)e^{-t} + 1\big], \\ q &= -\ln\big[(e^{-q_0} - 1)e^{-\alpha t} + 1\big]. \end{aligned} \tag{41}$$

Inserting these into the equations for $\dot{x}$ and $\dot{y}$, then solving yields

$$\begin{aligned} x &= \big((e^t - 1) + e^{-p_0}\big)\big(\rho_v(e^{-t} - 1) + x_0 e^{p_o}\big), \\ y &= \big((e^{\alpha t} - 1) + e^{-q_0}\big)\big(\rho_d(e^{-\alpha t} - 1) + y_0 e^{q_o}\big). \end{aligned} \tag{42}$$

Finally the equation for $\dot{u}$ can be integrated to yield

$$\begin{aligned} u = u_0 &+ \ln\big[(e^{-p_0}-1)e^{-t}+1\big]\big(-x_0e^{t+p_0}+\rho_v e^t+\rho_v e^{-t}(1-e^{-p_0})\big)+t\rho_v-p_0x_0 \\ &+\rho_v(1+e^{-t})(e^{-p_0}-1)+\ln\big[(e^{-p_0}-1)+e^t\big]\big(-2\rho_v+x_0e^{p_0}-x_0+\rho_v e^{-p_0}\big) \\ &+\ln\big[(e^{-q_0}-1)e^{-\alpha t}+1\big]\big(-y_0e^{\alpha t+q_0}+\rho_d e^{\alpha t}+\rho_d e^{-\alpha t}(1-e^{-q_0})\big)+\alpha t\rho_d-q_0y_0 \\ &+\rho_d(1+e^{-\alpha t})(e^{-q_0}-1)+\ln\big[(e^{-q_0}-1)+e^{\alpha t}\big]\big(-2\rho_d+y_0e^{q_0}-y_0+\rho_d e^{-q_0}\big). \end{aligned} \tag{43}$$

This expression for u will be used to find the maximum of the stationary distribution on the line $x+y=1$.

4.2 Inner region

Next we find the solution on a boundary curve. The obvious choice would be the lines $(0,y)$, $(1,y)$ and $(x,0)$. However from the analysis in Section 3.2, $p_{v,d}$ (so therefore $u(x,y)$) has a maximum at (ρ_v,ρ_d). The calculation showed that there was an inner region of width $1/\sqrt{N}$ in x and y close to the maximum (ρ_v,ρ_d). A simple rescaling allowed us to calculate an asymptotic approximation of the solution in this region. Using the definition of x and y used in Section 4.1, the asymptotic solution in the inner–region is

$$p(x,y) \approx \exp\left(-N\left(\frac{x^2}{2\rho_v}+\frac{y^2}{2\rho_d}\right)\right). \tag{44}$$

The characteristics will intersect at the point (ρ_v,ρ_d). A similar intersection of characteristics at a point occurs in ray theory at either a source or a sink. The boundary curve we shall use is a small ellipse centered on the source of radius $\epsilon \ll 1$. The boundary data is then

$$\begin{aligned} x_0(s) &= \rho_v+\epsilon\sqrt{\rho_v}\cos(s), \\ y_0(s) &= \rho_d+\epsilon\sqrt{\rho_d}\sin(s), \\ p_0(s) &= -\epsilon\cos(s)/\sqrt{\rho_v} \\ q_0(s) &= -\epsilon\sin(s)/\sqrt{\rho_d} \\ u_0(s) &= -\epsilon^2. \end{aligned} \tag{45}$$

The characteristic directions can now be calculated.

4.3 Characteristic directions

The shape and direction of the characteristics depends on the size of the parameter α. In the following analysis we shall consider the two cases when $\alpha > 1$ and when $\alpha = 1$. The case when $\alpha < 1$ is omitted, but is essentially the same as the calculation when $\alpha > 1$

4.3.1 $\alpha = 1$

Define $T = t + \ln(\epsilon)$, which is a translation of the characteristic variable. Inserting (45) into (42) and expanding in the limit $\epsilon \to 0$ yields

$$\begin{aligned} x &= \rho_v+\sqrt{\rho_v}\cos(s)e^{-T}, \\ y &= \rho_d+\sqrt{\rho_d}\sin(s)e^{-T}. \end{aligned} \tag{46}$$

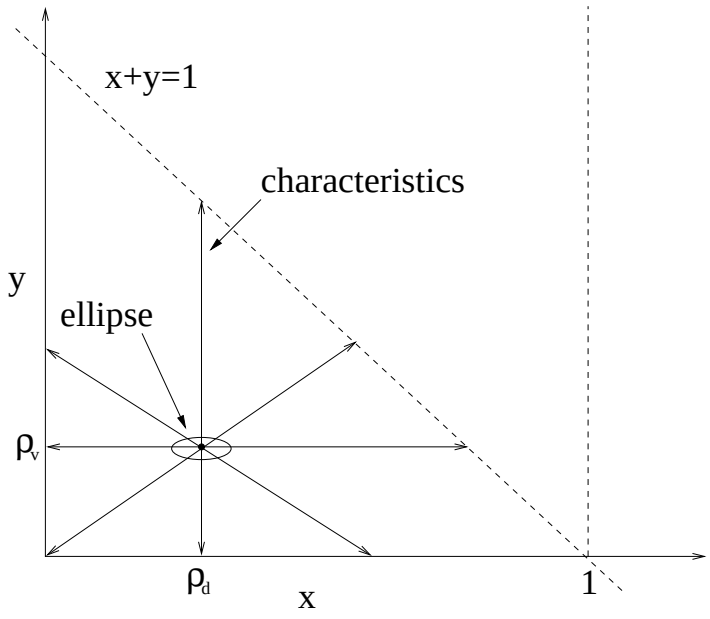

Figure 10: The characteristics originating from a small ellipse centred on (ρ_v, ρ_d) when $\alpha = 1$. The characteristics are straight lines.

Therefore the equation for the characteristics is given by

$$(y - \rho_d) = \tan(s)\sqrt{\frac{\rho_d}{\rho_v}}(x - \rho_v). \tag{47}$$

This tells us that the characteristics are straight lines radiating from (ρ_v, ρ_d). They are drawn in Figure 10.

4.3.2 $\alpha > 1$

The shape of the characteristics is different when $\alpha \neq 1$. They are now curved and originate from the ends of the ellipse where $s \approx 0$ and $s \approx \pi$ (Figure 11). The characteristics that intersect with the line $x + y = 1$ originate from the end where $s \approx 0$. Define $S = \epsilon^{1-\alpha}s$, which rescales the parameter s into a small region close to $s \approx 0$. Inserting (45) into (42) and expanding in the limit $\epsilon \to 0$ yields

$$\begin{aligned} x &= \rho_v + \sqrt{\rho_v}e^{-T}, \\ y &= \rho_d + \sqrt{\rho_d}Se^{-\alpha T}. \end{aligned} \tag{48}$$

Therefore the equation for the characteristics is

$$(y - \rho_d) = \sqrt{\rho_d}S\left(\frac{x - \rho_v}{\sqrt{\rho_v}}\right)^{\alpha}. \tag{49}$$

So the characteristics are α-power curves originating from the end of the ellipse (Figure 11).

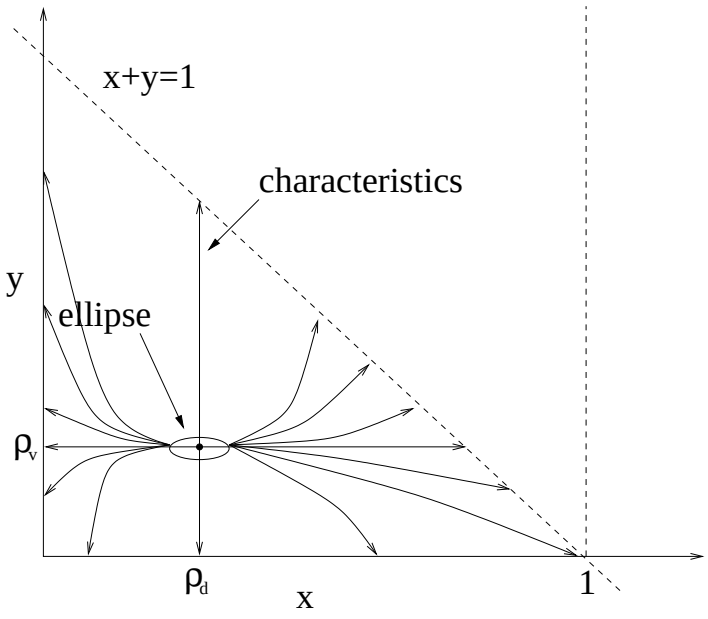

Figure 11: The characteristics originating from a small ellipse centred on (ρ_v, ρ_d) when $\alpha > 1$. The characteristics are α-power curves and originate from the two ends of the ellipse.

4.4 Maximisation of $u(x, y)$ on $x + y = 1$.

One of the quantities that we are interested in calculating is the maximum of the distribution $p_{v,d}$ on the line $v + d = N$ (see equations (24) and (25)). From the previous section we have x, y and u expressed in terms of s and t. Therefore the problem of maximising $u(x, y)$ on the line $x + y = 1$ is a constrained maximisation problem which is easily solved using Lagrange multipliers. Again the analysis depends on the size of α.

4.4.1 $\alpha > 1$

First insert the rescaled variable S into the boundary data (45) and expand in the limit $\epsilon \to 0$.

$$\begin{aligned} x_0 &\approx \rho_v + \epsilon\sqrt{\rho_v} \\ y_0 &\approx \rho_d + \epsilon^\alpha \sqrt{\rho_d} S \\ p_0 &\approx -\epsilon/\sqrt{\rho_v} \\ q_0 &\approx -\epsilon^\alpha S/\sqrt{\rho_d} \\ u_0 &= -\epsilon^2. \end{aligned} \tag{50}$$

Inserting these expansions and the rescaled variable T into the expression (43) for $u(s, t)$, then expanding in the limit $\epsilon \to 0$ yields

$$u \approx \rho_v\big\{-(1+w)\ln(1+w) + w\big\} + \rho_d\big\{-(1+w^\alpha v)\ln(1+w^\alpha v) + w^\alpha v\big\}, \tag{51}$$

where $w = e^{-T}/\sqrt{\rho_v}$ and $v = s\rho_v^{\alpha/2}/\sqrt{\rho_d}$. The constraint $x + y = 1$ in terms of w and v is

$$\rho_v(1 + w) + \rho_d(1 + w^\alpha v) = 1. \tag{52}$$

The constrained maximisation can now be performed using Lagrange multipliers. The maximum occurs at

$$w = vw^\alpha = \frac{1}{\rho_v + \rho_d} - 1. \tag{53}$$

Back substituting yields

$$\begin{aligned} x^* &= \frac{\rho_v}{\rho_v + \rho_d}, \\ y^* &= \frac{\rho_d}{\rho_v + \rho_d}, \\ u^* &= \ln(\rho_v + \rho_d) + 1 - \rho_v - \rho_d. \end{aligned} \tag{54}$$

The expressions for x^*, y^* and u^* are identical to those found in Section 3.2 (equations (24) and (25)). Notice that these expressions are independent of α.

4.4.2 $\alpha = 1$

Next we shall consider the case when $\alpha = 1$. The boundary data in the limit $\epsilon \to 0$ is given by (45). The value of T where the characteristics intersect the line $x + y = 1$ can be found by solving (46):

$$T = -\ln\left[\frac{1 - \rho_v - \rho_d}{\sqrt{\rho_v}\cos(s) + \sqrt{\rho_d}\sin(s)}\right]. \tag{55}$$

To find the value of $u(s,t)$ on the line $x + y = 1$ we insert this value of T into the expression (43). Expanding in the limit $\epsilon \to 0$ and re–arranging yields

$$\begin{aligned} u \approx &- \frac{1 + \rho_d z - \rho_v}{1 + z}\ln\left[\frac{1 + \rho_d z - \rho_v}{\rho_d(1 + z)}\right] - \frac{\rho_v + z - \rho_d z}{1 + z}\ln\left[\frac{\rho_v + z - \rho_d z}{\rho_v(1 + z)}\right] \\ &+ 1 - \rho_v - \rho_d, \end{aligned} \tag{56}$$

where $z = \sqrt{\rho_v/\rho_d}\cot(s)$. The final part of the analysis is to maximise this with respect to z, which gives

$$z = \frac{\rho_v}{\rho_d}, \tag{57}$$

and back–substituting yields

$$\begin{aligned} x^* &= \frac{\rho_v}{\rho_v + \rho_d}, \\ y^* &= \frac{\rho_d}{\rho_v + \rho_d}, \\ u^* &= \ln(\rho_v + \rho_d) + 1 - \rho_v - \rho_d. \end{aligned} \tag{58}$$

The expressions for x^*, y^* and u^* are identical to those found in Section 3.2 (equations (24) and (25)). Note that these equations for the maximum are the same as (54) for the case when $\alpha > 1$.

4.5 Region above line $x + y = 1$

The results of the ray theory approach to the problem yielded the same results below the line $x + y = 1$ as those obtained when we assumed the system was in detailed balance. In Section 3.2, the solution above the line $x + y = 1$ could only be found in a narrow boundary layer. The advantage of the ray theory method is that it can be extended to calculate the stationary distribution in the whole region where $x+y > 1$. While it is not necessary to do this calculation to obtain the results of Section 3.3, it will be necessary to do similar calculations when we consider more general queues. In this section we shall briefly derive the eikonal-type equation and solve it for the region where $x + y > 1$. The eikonal-type equation is found by inserting the WKB ansatz (36) into the full balance equations (4) and taking the limit $N \to \infty$. This gives

$$\rho_v(e^{-p} - 1) + x(e^p - 1) + \alpha\big\{\rho_d(e^{-q} - 1) + (1 - x)(e^q - 1)\big\} = 0. \tag{59}$$

The solution of this equation can be solved in terms of characteristics by using Charpit's method. The solution along the characteristics then satisfies

$$\begin{aligned}
\dot{x} &= xe^p - \rho_v e^{-p},\\
\dot{y} &= \alpha\big((1 - x)e^q - \rho_d e^{-q}\big),\\
\dot{p} &= 1 - e^p + \alpha(e^q - 1),\\
\dot{q} &= 0,\\
\dot{u} &= p\big(xe^p - \rho_v e^{-p}\big) + \alpha q\big((1 - x)e^q - \rho_d e^{-q}\big).
\end{aligned} \tag{60}$$

Initial data is given on the boundary curve $x+y = 1$ which we will label $(x_1, y_1, p_1, q_1, u_1)$. These equations can now be solved sequentially. The equation for q is solved first to give $q = q_1$. Inserting this into the equation for p and solving, yields

$$p = \ln\big[a\big] - \ln\big[(ae^{-p_1} - 1)e^{-at} + 1\big], \tag{61}$$

where $a = \alpha(e^{q_0} - 1) + 1$. Next the equations for x and y can be solved to give

$$\begin{aligned}
x &= \frac{1}{a^2}\big((ae^{-p_1} - 1) + e^{at}\big)\big(\rho_v(e^{-at} - 1) + ax_1e^{p_1}\big),\\
y &= y_1 - \rho_d e^{-q_1}\alpha t - \alpha e^{q_0}\left\{\frac{ax_1e^{p_1} - \rho_v}{a^3}e^{at} - \frac{\rho_v(ae^{-p_1} - 1)e^{-at}}{a^3}\right.\\
&\quad \left. + \left(-1 + x_1 + \frac{2\rho_v}{a^2} - \frac{x_1e^{p_1}}{a} - \frac{e^{-p_1}\rho_v}{a}\right)t - \frac{x_1e^{p_1}}{a^2} + \frac{\rho_v e^{-p_1}}{a^2}\right\}.
\end{aligned} \tag{62}$$

Finally the equation for u can be integrated to give

$$\begin{aligned}
u &= u_1 - p_1x_1 - t(ax_1e^{p_1} - \rho_v)\left(e^{-p_1} - \frac{1}{a}\right) + \frac{\rho_v(e^{-at} - 1)}{a}\left(e^{-p_0} - \frac{1}{a}\right)\\
&\quad + \big(\ln[a] - \ln\big[(ae^{-p_1} - 1)e^{-at} + 1\big]\big)\left(e^{-p_1} - \frac{1}{a} + \frac{e^{at}}{a}\right)\left(x_1e^{p_1} - \frac{\rho_v}{a} + \frac{\rho_v e^{-at}}{a}\right)\\
&\quad - q_1\rho_d e^{-q_1}\alpha t - q_1\alpha e^{q_1}\left\{\frac{(ax_1e^{p_1})e^{at}}{a^3} + \left(-1 + x_1 + \frac{2\rho_v}{a^2} - \frac{x_1e^{p_1}}{a} - \frac{e^{-p_1}\rho_v}{a}\right)t\right.\\
&\quad \left. - \frac{\rho_v(ae^{-p_1} - 1)e^{-at}}{a^3} - \frac{x_1e^{p_1}}{a^2} + \frac{\rho_v e^{-p_0}}{a^2}\right\}.
\end{aligned} \tag{63}$$

4.6 One voice and two data call types

The ray theory method used in the previous section can be used to consider more complicated models using multiple data caller types. Consider two types of data caller: one type can use only one channel, and the other can use two channels. Define v as the number of voice callers at the station, d_1 as the number of data callers who can use only one channel, and d_2 as the number of data callers who can use two channels. The rules for channel allocation are as follows:

1. Voice calls always take preference. If all channels are in use than the voice callers are dropped.

2. If $v + d_1 + 2d_2 \leq N$ then all data calls are assigned their maximum number of channels.

3. If $N - d_2 < v + d_1 + d_2 \leq N$ then all data callers are assigned at least one channel, and $N - v - d_1 - d_2$ callers are assigned 2 data channels.

4. If $v + d_1 + d_2 > N$ then $N - v$ data callers are assigned channels and the rest are queued.

The calls are assumed to arrive via a Poisson process with rate $N\lambda_v$, $N\lambda_{d_1}$ and $N\lambda_{d_2}$ respectively. The call times are distributed exponentially with means $1/\mu_v$, $1/\mu_{d_1}$ and $1/\mu_{d_2}$ respectively. The call time for d_2 callers assumes that the data call is using two channels. If two channels are used it is assumed that they take half the time.

The state space is now partitioned by planes ($v + d_1 + 2d_2 = N$ and $v + d_1 + d_2 = N$), with different transition rules in each region. To calculate quantities such as the probability that a data call is queued, it is necessary to calculate the stationary distribution above the plane $v+d_1+d_2 = N$. The boundary curve will now be an ellipsoid centred on the maximum of the stationary distribution. To calculate the stationary distribution above the plane $v + d_1 + d_2 = N$, it is necessary to calculate along a ray originating from the ellipsoid which will then be refracted at the plane $v + d_1 + 2d_2 = N$ and again at the plane $v + d_1 + d_2 = N$.

5 Case B: Data calls can use N channels, with dynamic allocation

We now move on to consider the case in which all users have at least N channels available on their phones and there is dynamic allocation. This is attractive to deal with because we can assume that *only one* data call is transmitted at a time, and that it uses *all* the available channels, contracting and expanding as k (the number of voice calls) goes up and down. So if the state of the system is (k, l) with again l being the number of data calls in the system, then for $l \geq 1$ there is 1 data call being transmitted and $l - 1$ in the queue. The transition rates then are as indicated on the diagram in Figure 12, representing:

1. Voice call arrival (rate λ_V): if $k < N$ this takes k to $k + 1$.

2. Data call arrival (rate λ_D): this takes l to $l+1$.

3. Voice call finish (rate $k\mu_V$, where $k > 0$): this takes k to $k-1$.

4. Data call finish: if $k < N$ and $l > 0$ then the data call being transmitted is using $N-k$ channels, so its finishing rate is $(N-k)\mu_D$, and, on finishing, l goes to $l-1$.

(We use notation such as $\lambda_V[k < N]$ to indicate that the term λ_V is only present if $k < N$. Formally, $[k < N]$ is an indicator function taking the value 1 if $k < N$ and 0 if not.) To find the steady state distribution $p_{k,l}$ there are broadly speaking three

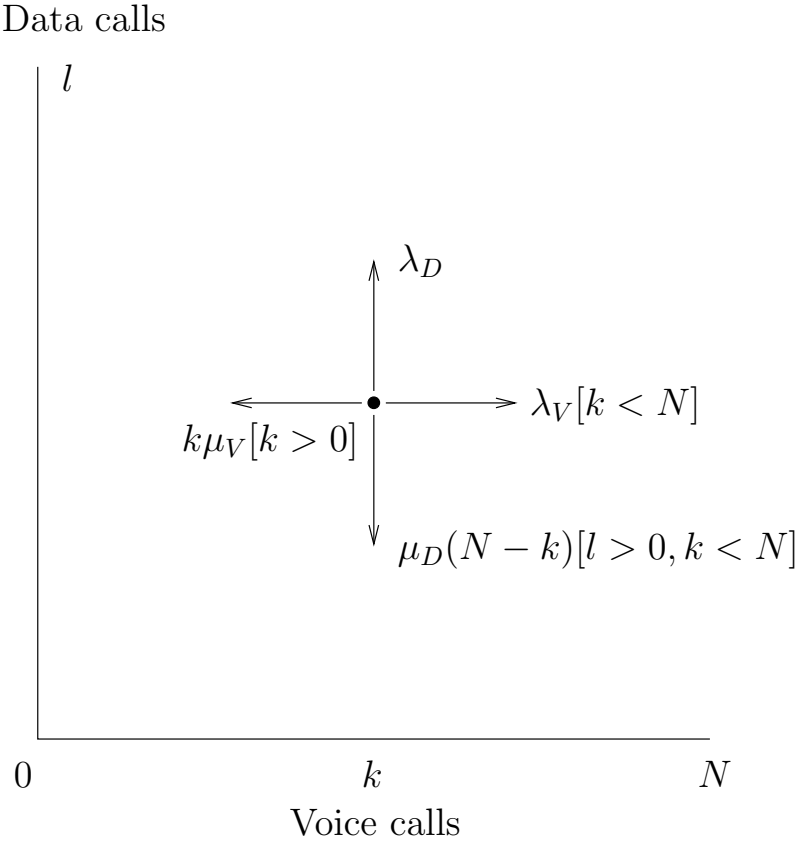

Figure 12: Transition rate diagram for Case B.

approaches:

1. Monte Carlo simulation;

2. solving the balance equations numerically;

3. analytical approach using generating functions.

We shall here develop the third of these, which begins by writing down the balance equations:

$$\begin{aligned} p_{k,l}(\lambda_V[k < N] + \lambda_D + k\mu_V + (N-k)\mu_D[l > 0]) = \\ p_{k-1,l}\lambda_V[k > 0] + p_{k,l-1}\lambda_D[l > 0] + \\ p_{k+1,l}(k+1)\mu_V[k < N] + p_{k,l+1}(N-k)\mu_D \end{aligned} \tag{64}$$

for $0 < k < N$ and $l > 0$. Now we introduce the generating functions

$$F_k(z) = \sum_{l \geq 0} p_{k,l} z^l \,, \tag{65}$$

which are infinite series. Since the $p_{k,l}$ are nonnegative and $\sum_{l \geq 0} p_{k,l}$ is convergent, the series (65) is certainly convergent for $|z| \leq 1$. Multiplying the balance equations (64) by z^l and summing over $l \geq 0$ we obtain

$$\begin{aligned} F_k(z)\lambda_V[k < N] + F_k(z)\lambda_D + F_k(z)k\mu_V + (F_k(z) - F_k(0))(N-k)\mu_D = \\ F_{k-1}(z)\lambda_V[k > 0] + F_k(z)z\lambda_D + F_{k+1}(z)(k+1)\mu_V[k < N] + \\ ((F_k(z) - F_k(0))/z)(N-k)\mu_D. \end{aligned} \tag{66}$$

This we can write as a tridiagonal system of equations for the $F_k(z)$:

$$\begin{pmatrix} b_0 & c_0 & 0 & 0 & \dots & \dots & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & \dots & \dots & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & 0 & \dots & 0 & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ 0 & 0 & \dots & 0 & a_{N-2} & b_{N-2} & c_{N-2} & 0 \\ 0 & 0 & \dots & \dots & 0 & a_{N-1} & b_{N-1} & c_{N-1} \\ 0 & 0 & \dots & \dots & 0 & 0 & a_N & b_N \end{pmatrix} \begin{pmatrix} F_0(z) \\ F_1(z) \\ F_2(z) \\ \vdots \\ \vdots \\ F_{N-2}(z) \\ F_{N-1}(z) \\ F_N(z) \end{pmatrix} = \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ \vdots \\ r_{N-2} \\ r_{N-1} \\ r_N \end{pmatrix}, \tag{67}$$

where the entries are

$$a_k = -\lambda_V[k < N], \tag{68}$$

$$b_k = \lambda_V[k < N] + \lambda_D(1-z) + k\mu_V + (N-k)\mu_D(1-1/z), \tag{69}$$

$$c_k = -\mu_V(k+1)[k < N], \tag{70}$$

$$r_k = (N-k)\mu_D(1-1/z)F_k(0). \tag{71}$$

The right-hand side involves the quantities $F_k(0) = p_{k,0}$ for $k = 0, \dots, N-1$, which are not known *a priori* . The additional fact that has to be used to find them is that the functions $F_k(z)$ are analytic in the region $|z| < 1$. However, the determinant of the matrix on the left of (67) has $N-1$ zeros in $|z| < 1$ if $\lambda_D/\mu_D < N - \overline{V}$, and N if that inequality is reversed. In the first of these cases (which is when the queue does not build up indefinitely) let those zeros be $z_1, \dots, z_{N-1}$. Then the condition that $F_N(z)$ does not become singular at z_i gives $N-1$ linear constraints on the values $F_0(0), \dots, F_{N-1}(0)$. When those conditions are satisfied, the tridiagonal form of (67) ensures that *each* $F_k(z)$ then has no singularity at z_i. Then finally the condition that $\sum_{k=0}^{N} F_k(1) = 1$ is needed to normalize the probabilities to sum to 1 and completely determine the solution. For $N = 2$ the procedure can be carried through by pencil-and-paper: for $N = 8$ the roots z_i would need to be computed numerically but it would not be difficult to write a program that did this and computed the steady state distribution $p_{k,l}$ from which everything else can be found. For instance, the mean service time for a data call will be $(\sum_{k,l} l p_{k,l})/\lambda_D$.

6 Multichannel data calls

When data calls can use a number of channels dependent on the capacity of the user's phone and on the current state of the system, a more complicated structure is formed. We shall let the state of the system be denoted by $(k, l_1, l_2, \ldots, l_N, q)$ where k is the number of voice calls in progress, l_r is the number of r-channel data calls being transmitted, and q is the number of data calls in the queue. The number of channels in use will be denoted by $U = k + \sum_0^N r l_r \leq N$, and either $q = 0$ or $U = N$. The parameters defining the system must now also include some parameters f_1, f_2, $\ldots$, f_N where f_r is the proportion of users whose phone capacity is exactly r channels.

If we let $\mathbf{l} = (l_1, l_2, \ldots, l_N)$ then counting the number of states in the system depends on counting the number of vectors $\mathbf{l}$ of nonnegative integers such that $\sum r l_r = n$ (where $0 \leq n \leq N$). This number is denoted by $p(n)$ and is the classical number of unrestricted partitions of n, which was first introduced and studied by Euler; a suitable modern reference is [2, Ch.19]. The rate of growth of $p(n)$ for large n was first found by Hardy and Ramanujan in [1] and is

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right). \tag{72}$$

In fact they found a (more complicated) asymptotic series from which the *exact* value of $p(n)$ can be computed, and Rademacher subsequently simplified the approach. The number of $(k, \mathbf{l})$ such that $U = n$ is

$$p_1(n) = p(0) + p(1) + \ldots + p(n), \tag{73}$$

since for each value of k from 0 to n, the vector $\mathbf{l}$ can run over all partitions of $n - k$. So the number of states of our queueing system with $q = 0$ is the sum $p_2(N)$ where

$$p_2(n) = p_1(0) + p_1(1) + \ldots + p_1(n), \tag{74}$$

since U can take any value from 0 to N. The number of states with $q > 0$, is $p_1(N)Q$ (where we suppose the maximum queue size is Q) since for $q > 0$ the channels must all be in use so $U = N$. The total number of states is therefore $N_0 = p_2(N) + p_1(N)Q$. (It seems these cumulative partition counting functions $p_1(n)$ and $p_2(n)$ have not been studied in the number theory literature, and no exact asymptotic formulae are known.)

If $N = 8$, $N_0 = 187 + 67Q$, which is rather difficult to illustrate, so Table 1 shows the states if $N = 4$, in which case $N_0 = 26 + 12Q$.

The state transitions that can be made now are of the following types:

1. Voice call arrives, rate λ_V: if $U < N$ then the voice call simply goes into one of the free channels, so $k \mapsto k + 1$. If $k = N$ then all channels are already used for voice calls, and the call is blocked. The more complicated case is where all channels are in use but at least one of them is carrying a data call, so the rule that voice takes priority has to be applied. If some *multichannel* data calls are in progress then we shall assume that one of those currently using the maximum number of channels has one channel removed from it to give to the new voice call. If all the data calls

No queue ($q = 0$)	0 voice	0
		1
		2
		11
		3
		21
		111
		4
		31
		22
		211
		1111
	1 voice	V
		V1
		V2
		V11
		V3
		V21
		V111
	2 voice	VV
		VV1
		VV2
		VV11
	3 voice	VVV
		VVV1
	4 voice	VVVV
Queue ($1 \leq q \leq Q$)	0 voice	4
		31
		22
		211
		1111
	1 voice	V3
		V21
		V111
	2 voice	VV2
		VV11
	3 voice	VVV1
	4 voice	VVVV

Table 1: Table of states for $N = 4$: VV2, for instance, denotes two voice calls and a 2-channel data call. The lower part of the table is repeated for each $q = 1, \ldots, Q$.

in progress are *single* channel then we assume that one of them is displaced back into the queue. In symbols, if $k < U = N$, find the maximal r such that $l_r > 0$: if $r \geq 2$ then we make the transition $k \mapsto k + 1$, $l_r \mapsto l_r - 1$, $l_{r-1} \mapsto l_{r-1} + 1$; if $r = 1$

then $k \mapsto k+1$, $l_1 \mapsto l_1 - 1$, $q \mapsto q+1$.

2. Data call arrives, rate λ_D, from a user who will have r channels available with probability f_r. If all channels are in use, $(U = N)$ the call is queued, $q \mapsto q+1$. If there are $F = N - U > 0$ free channels available, the new call is assigned $\max(F, r)$ of them, so for $r < F$, $l_r \mapsto l_r + 1$ with rate $\lambda_D f_r$, while at $r = F$, $l_F \mapsto l_F + 1$ with rate $\lambda_D \sum_F^N f_r$.

3. Voice call finishes, rate $k\mu_V$. If there are no data calls in the queue, then we simply have $k \mapsto k-1$. But if $q > 0$ then a data call from the queue will begin using the vacated channel, so we then have $k \mapsto k-1$, $l_1 \mapsto l_1 + 1$, $q \mapsto q-1$.

4. r-channel data call finishes, rate $r l_r \mu_D$. When r channels become free, calls from the data queue (if any) are allocated to the free channels successively until either the queue is empty or all those r channels are used again. (It is more difficult to write down explicit expressions for the probabilities, but the method is clear enough.)

It should be noted that we are making a simplifying approximation here, in that when a data call is pushed back into the queue, we do not keep track of how many channels its user had: effectively when it begins to be processed again we are treating it as a new call, as if the number of user channels were reassigned randomly with the probabilities f_r.

Results from this process can be computed numerically: a suitable maximum queue size Q has to be chosen, and then the state transition matrix set up according to the above rules, and then the steady state probability distribution found. This has been done and some results computed for the case $\lambda_V = \lambda_D = 1.5$, $\mu_V = \mu_D = 0.5$, $N = 8$, with various user-capacity distributions f_r, and results for the mean data call time are shown in Table 2. (If we treat the time unit as minutes, then $\mu_V = 0.5$ represents a mean voice call time of 2 minutes, $\lambda_V/\mu_V = 3$ represents a mean of 3 voice calls in the system at a time (apart from blocking) and we set similar parameters for the data. The maximum queue size Q was taken as 50.) The decrease in mean call time as user capacity is set to 1, 2, 4 or 8 channels can be seen, and also a case where there is a mixed population of different user capacities. The mean data call time of 2.6434 in the first line of Table 2 — where all users have only 1 channel available — agrees with that calculated by the methods described for Case A in Section 3, as of course it should. However the case where all users have 8 channels is somewhat different from Case B considered in Section 5, because here there is not dynamic allocation whereas in Case B there was. So in Case B the single data call being transmitted used all the available channels until it was finished. But here, a data call sticks with the number of channels it started with, possibly losing one occasionally to a voice call that takes priority, and any other channels that become free are allocated to another data call from the queue. The table shows that the mean data call time in this allocation system is 1.8051. However, the analysis methods for Case B in Section 5 give a mean data call time of 1.3271 when dynamic allocation is used — a significant improvement.

$r =$	1	2	3	4	5	6	7	8	mean time
$f_r =$	1	0	0	0	0	0	0	0	2.6434
	0	1	0	0	0	0	0	0	2.1069
	0	0	0	1	0	0	0	0	1.8799
	0	0	0	0	0	0	0	1	1.8051
	0.1	0.2	0.3	0.4	0	0	0	0	2.0297

Table 2: Mean data call time for some different distributions of user capacity.

7 Additional remarks

In addition to the approaches taken here, it was noted at the Study Group that work by C. Knessl may be worth looking into. Furthermore, an approach similar in style to that of Section 3, treating N as large, was suggested, along the following lines. Let the cumulative number of voice calls be N_V (treated as a continuous random variable), obeying $dN_V = I_V dt$, where the intensity of voice calls I_V obeys a stochastic differential equation $dI_V = (\ldots)dt + (\ldots)dW$, where W is Brownian motion and the coefficients $(\ldots)$ are chosen to give representative variations of the rate and to ensure that I_V remains positive. A similar model would be taken for the incidence of data calls. Then the antenna channels would be regarded as a continuous resource, of which time-varying fractions ϕ_V, ϕ_D, ϕ_F (summing to 1) are transmitting voice calls, transmitting data calls, and free. Differential equations would need to be developed to model these, incorporating elements that represent the priority of voice over data, *etc.*.

8 Conclusions

The GPRS service system for voice and data calls has been modelled as a continuous time Markov chain in various circumstances. When each user has only 1 channel available for data calls, the analysis presented in Section 3 gives ways of calculating the relevant quantities, such as mean service time. These methods are based on the idea of treating the total number of channels N as large and making appropriate approximations, and it is shown that the approximations are good when $N = 8$, the typical value in practice. Then Section 4 shows how that method can be extended to more complex queues, *e.g.* where users have more than 1 channel available for data calls. Then Section 5 deals with the case where all users have N channels available for data calls, and there is dynamic allocation of channels. This can be dealt with analytically by the method of generating functions. Finally in Section 6 we consider the case where users' phones can have any number of channels available for data calls, and show how to set up the transition rate matrix for this. The steady state distribution can then be found computationally, for any distribution of data call capacity in the user population. It is shown how the mean service time for data calls drops as the number of data channels available to users rises. Also the decrease of mean service time for data calls when *dynamic* channel allocation is used has been shown.

A Service time distributions

When a Markov chain is used to model a voice and data queueing system, we need some method to compute the service time distribution for data calls. We can think of this time as obtained in the following way:

1. The data call arrives, and takes the system to some state i.
2. Imagine that there are no later data arrivals, but voice arrivals continue. We then still have a Markov chain, but the data arrival rate constant λ_D has been set to 0.
3. In this system we want to know the distribution of the time T_{iA} to travel from state i to the set of states A in which there are no data calls in the system, so the data call of interest has just cleared.

These times T_{iA} will have complicated distributions: although the residence time in any single state is exponentially distributed, T_{iA} is a sum of a random number of such times, with different parameters, depending on the path followed from i to A. The way to obtain the distribution of T_{iA} is to consider the first step from i: either it is directly to a state in the set A, or it is to some other state j. Hence

$$T_{iA} = \begin{cases} R_i & \text{with probability } q_{iA}/q_i \\ R_i + T_{jA} & \text{with probability } q_{ij}/q_i \end{cases} . \tag{75}$$

Here R_i is the residence time in state i, q_{ij}/q_i is the probability that the next state after i is j, and $q_{iA} = \sum_{a\in A} q_{ia}$. We can deal with this by taking the Laplace transform $X_i(s)$ of the probability density function of T_{iA},

$$X_i(s) = \mathbb{E}\mathrm{x}(\exp(-sT_{iA})) = \mathbb{E}\mathrm{x}(\exp(-sR_i)) \left(\frac{q_{iA}}{q_i} + \sum_{j\neq i, j\notin A} \frac{q_{ij}}{q_i} X_j(s) \right) . \tag{76}$$

Since R_i is exponential, $\mathbb{E}\mathrm{x}(\exp(-sR_i)) = q_i/(q_i + s)$ and we simply have

$$(q_i + s)X_i(s) = q_{iA} + \sum_{j\neq i, j\notin A} q_{ij} X_j(s). \tag{77}$$

These linear equations allow the $X_i(s)$ to be found. Numerical computations of quantities of interest then involve inversions of these Laplace transforms, which is generally difficult. Nevertheless, if some integral properties of the distribution, such as tail probabilities, are required, then it is possible to compute those integrals inside the Laplace transform, which generally makes the inversion better conditioned. In some situations this can allow analytical estimation by saddle point methods, but we have not looked into whether that will be possible in this case.

B Continuous time Markov chains: notation and methods

We here describe in outline the basic theory and notation that we use for Markov chains. A Markov chain is a stochastic process whose future transition probabilities depend only

on the current state, not on the past history. We shall be thinking of chains with discrete states, which we represent here by i, j *etc.*, with continuous time t. The Markov chain is defined by the *transition rates*

$$q_{ij} = \lim_{\delta t \to 0} \left(\frac{\mathbb{P}\text{r}(\text{state } j \text{ at time } t + \delta t \mid \text{state } i \text{ at time } t)}{\delta t} \right) \geq 0, \quad (j \neq i). \tag{78}$$

If the system starts in state i then after a small time δt, the probability that it is in state $j \neq i$ is $q_{ij}\delta t + o(\delta t)$, and so the probability that it is still in state i is $1 - q_i \delta t + o(\delta t)$, where

$$q_i = \sum_{j \neq i} q_{ij}. \tag{79}$$

So the residence time in state i is exponentially distributed with parameter q_i. Provided that a certain regularity condition is satisfied (which it is in all cases of interest to us) and provided that all the states are accessible from each other, then there is a unique steady state distribution p_i such that

$$p_i q_i = \sum_{j \neq i} p_j q_{ji}, \tag{80}$$

representing the fact that the rate of flow of probability out of state i is equal to the rate of flow into it from all states $j \neq i$. The equations (80) are linear and homogeneous, and the solution to them has to be normalized to have $\sum p_i = 1$ to make it a probability distribution.

B.1 Numerical methods

There is of course hardly a shortage of numerical algorithms for solving systems of linear equations such as (80) or (77). However we just wish to point out here a method that has particular advantages in terms of accuracy and speed for systems like

$$(a_i + \sum_{j=1}^{n}{}' b_{ij})x_i - \sum_{j=1}^{n}{}' b_{ji}x_j = c_i \tag{81}$$

where the a_i, b_{ij} and c_i are all nonnegative and where $\sum'$ denotes that the term $j = i$ is omitted. Gaussian elimination writes

$$x_1 = \frac{c_1 + \sum_2^n b_{j1}x_j}{\hat{a}_1}, \quad \hat{a}_1 = a_1 + \sum_2^n b_{1j}, \tag{82}$$

and substituting this into (81) for $i \geq 2$ results in the form

$$(a_i^* + \sum_{j=2}^{n}{}' b_{ij}^*)x_i - \sum_{j=2}^{n}{}' b_{ji}^* x_j = c_i^* \tag{83}$$

where

$$c_i^* = c_i + b_{1i}c_1/\hat{a}_1, \quad b_{ji}^* = b_{ji} + b_{j1}b_{1i}/\hat{a}_1, \quad a_i^* = a_i + b_{i1}a_1/\hat{a}_1. \tag{84}$$

We therefore reduce to a system of the same form but in the variables $x_2, \ldots, x_n$ and can proceed to eliminate x_2, *etc.*. After these reductions we calculate x_n, and then work backwards by equations like (82) to find all the x_i. The advantage of the scheme numerically is that the equations (84) and (82) for the new coefficients and the back substitutions only involve addition, multiplication and division of *positive* quantities, no subtraction, so the method takes full advantage of the well-conditioned nature of the equations. The advantage of the scheme in efficiency depends on ordering the variables so that sparsity properties are maintained. For instance, if the coordinates $1, 2, \ldots, n$ are grouped into blocks $I_1, I_2, \ldots, I_k$ such that $b_{ij} > 0$ only for coordinates i, j in the same or adjacent blocks, then this property is maintained during the whole of the elimination process. Consequently the potentially time-consuming step in the algorithm, which is the calculation of the b^*_{ji} is not of order n^3, but n^2 times the size of the largest block.

For Markov chains, the balance equations for the steady state probabilities (80) are of this form with $a_i = 0$, $b_{ij} = q_{ij}$, and $c_i = 0$. Consequently the a^*_i and c^*_i remain 0 throughout and need not be considered. When the variables x_1, ..., x_{n-2} have been eliminated, the remaining equation just fixes the ratio of x_{n-1} to x_n. So by setting $x_n = 1$ one can back-substitute to find $x_{n-1}, \ldots, x_1$, and then rescale them all to make $\sum x_i = 1$ as required. For the time distribution equations (77), they are in exactly the right form if s is real and positive, and for $\mathrm{Re}(s) \geq 0$ they retain many of the same properties. To take advantage of the sparsity in cases A and B considered in this report, take I_l to be the states with exactly l data calls in the system, and the size of the largest block is then N.

References

[1] G. H. Hardy, S. Ramanujan. Asymptotic formulae in combinatory analysis. Proceedings of the London Mathematical Society, 2, XVII, 75–115 (1918).

[2] G. H. Hardy, E. M. Wright. An introduction to the theory of numbers. Clarendon Press, Oxford. 4th edition 1960.