ITC19/ Performance Challenges for Efficient Next Generation Networks 1295-1304
LIANG X.J. and XIN Z.H.(Editors)
V.B. IVERSEN and KUO G.S.(Editors)
Beijing University of Posts and Telecommunications Press

Performance Analysis of Wireless Data Systems with a Finite Population of Mobile Users

Shuping Liu [1] and Jorma Virtamo [1]

Networking Laboratory
Helsinki University of Technology
P.O.Box 3000, FI-02015 TKK, Finland
{shuping.liu, jorma.virtamo}@tkk.fi

Abstract. In a recent paper [1] Bonald et al. explored the impact of mobility on the flow level performance of wireless data systems. In contrast to their model (referred to as the BBP model), which implicitly assumes an infinite user population, we present and study two finite user population models called the OCOF (One Customer One Flow) model and the OCMF (One Customer Multiple Flows) model. In the previous one, only one flow per user can be in progress at any time and consecutive flows are separated by a thinking time. In the latter model, each customer is assumed to generate flows as a Poisson process. As in [1], the performance is analyzed in two tractable regimes, quasi-stationary (QS) and fluid (FL) regimes, corresponding to infinitely slow and infinitely fast user motion, respectively. The flow throughput of the models in the two regimes is evaluated. The models are compared by a set of numerical examples. The results show that the performance of the BBP model is between the OCOF and OCMF models. The reasons for this are discussed. The numerical studies also demonstrate how the finite user population models approach the BBP model as the number of users grows.
Keywords: cellular system, wireless data network, user mobility

1 Introduction

The next generation wireless cellular systems are expected to support a wide variety of high-speed data applications, in addition to conventional voice services and current low-bandwidth data service such as short message. In general, modeling of wireless systems is complex due to many factors influencing the performance such as interference, widely varying channel quality among spatially distributed users, as well as over time because of fading effects. In addition, one has to take into account the ”elastic” nature of data traffic where the flows dynamically share the available link capacity and the performance experienced by the users is characterized by the average flow response time in the stochastic environment. Relatively few theoretical models on the flow-level performance have been published. Notable exceptions are the works by Bonald and Proutière [2, 3] and the work by Borst [4], where they have analyzed the performance of single and multi-cell wireless systems with randomly located but stationary user population generating flows as a spatio-temporal Poisson process.

In a recent joint paper [1] these authors study the impact of mobility on the user performance data performance assuming an adaptive link with distance dependent data rate. They conclude that, in general, the performance is difficult to analyze. However, in two limit regimes, quasi-stationary (QS) and fluid (FL), where customers move infinitely slowly and infinitely fast, the system is amenable to analysis and, moreover, the performance is insensitive to detailed traffic characteristics. Furthermore, they were able to show that under certain assumptions these two regimes provide bounds for the performance of a real system where the users move at finite speeds.

The model of [1] (hereafter referred to as the BPP model) implicitly assumes that the user population in the cell is very large, whence flows arrive to the system as a Poissonian process, each flow being destined to a random point within the cell independent of the other flows. In a real system, of course, the user population cannot be infinite. In this paper, we study how finite user population affects the results. To this end, we present two different models. The first one, the OCOF (One Customer One Flow) model, assumes that at most one flow per customer can be in progress at any time. A new flow can be generated only after the old flow has ended and a random thinking time has elapsed. In the second model, the OCMF (One Customer Multiple Flows) model, a customer can transmit multiple flows simultaneously. In particular we assume that the total Poissonian arrival rate λ of the BPP model is just divided between the n users present in the system, each of them sending or receiving flows at rate λ/n. In all these models, the flows that are active in the system are assumed to share equally the service time of the base station.

The main contribution of the work is the derivation of the formulae for the main performance measure, so-called flow throughput, for both of the models and in both limit regimes. In particular, for the OCOF model is presented by a two station multi-class closed network for which we make use of the BCMP theorem. It is noteworthy that also in the finite population case the performance in the QS and FL regimes is insensitive. The performance of different systems are compared by a set of numerical examples.

2 OCOF finite population model

2.1 Model Description

The system model we consider is basically the same as that introduced in [1], incorporating also some elements of the model in [2]. It consists of a single circular cell with radius r and the base station at the center of cell. Elastic data flows share dynamically the service capacity of the base station in a fair manner so that with a given number of concurrent flows each of them receives equal share of the service time of the base station. The feasible bit rate of a flow, denoted R, means the bit rate at which the data of the flow would be transmitted if the flow were the only one in the system, i.e. if the base station used all its time sending this flow. If at any given instant there are m contending flows in the system, then the actual transmission rate of a flow with feasible rate R is R/m.

We assume the radio interface is adaptive so that the feasible rate depends on the link conditions and, in particular, on the distance of the user from the base station. In a real system only a limited set of feasible rates R_k, $k = 1, 2, \ldots, K$, are available. A feasible rate R_k is obtained when the distance of the user from the base station is between r_{k-1} and r_k (with the convention $r_0 = 0$ and $r_K = r$). The cell is thus divided into annular rings with areas $a_k = \pi(r_k^2 - r_{k-1}^2)$ and feasible rates R_k as depicted in Figure 1.

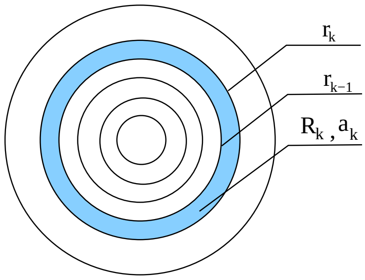


Fig. 1. Division of the cell into annular rings of a given feasible rate R_k

Our model differs from the previous work in that, whereas in [1] it is assumed that new flows are generated as a Poisson process at rate λ, implicitly implying that the customer population is infinite, we consider a finite population of n customers moving within the cell. In the OCOF (one customer one flow) model considered in this section we assume that each customer can transmit only one flow at a time. After the completion of the flow the user is assumed to go into a thinking phase, with mean duration $1/\nu$, after which a new flow is again transmitted, and so on. Our aim is to evaluate how much the finiteness of the customer population affects the performance results. For notational simplicity, we suppress the classification in [1] of customers by their traffic and mobility characteristics. If needed, it is fairly straight forward to make this generalization.

The focus of [1] was to assess the impact of user mobility on the flow-level performance. As mentioned in the Introduction, the authors were able to show that under fairly general assumptions the performance is bounded by the QS and FL limits. In the previous, the users move slowly in comparison to the characteristic time scale of the flow dynamics so that effectively they can be considered stationary. On a slower time scale, the users move to new positions, but assuming the users move independently and in such a way that the distribution of the user location in the cell is uniform, at any instant the quasi stationary system looks the same: new flows are always generated at a constant rate per unit time and unit area. In the fluid limit, on the contrary, the users move very fast so that, in effect, the feasible rate of each customer is a constant $R^{\text{fl}} = \mathrm{E}[R]$, where the expectation is over the distribution of the user location in the cell .

At both limits, the system of [1] reduces to a single PS queue. In the fluid limit, the mean service time of a flow is σ/R^{fl}. In the QS limit, the randomness of the service time of a single flow arises, independently, from two sources, the random size of the flow and the random feasible rate (because of the random location of the user). The mean service time then is $\sigma\mathrm{E}[1/R]$, where again the expectation is over the distribution of the user location in the cell. This may be written as σ/R^{qs} with $R^{\text{qs}} = 1/\mathrm{E}[1/R]$.

The main quantity of performance we consider is the mean response time (sojourn time) of a flow $\mathrm{E}[T]$ or, alternatively, the flow throughput γ defined as $\gamma = \sigma/\mathrm{E}[T]$. The analysis in [1] leads to the following results of a single PS queue:

$$\gamma^{\mathrm{qs}} = R^{\mathrm{qs}}(1-\rho^{\mathrm{qs}}), \qquad \text{with} \quad \rho^{\mathrm{qs}} = \lambda\sigma/R^{\mathrm{qs}},$$
$$\gamma^{\mathrm{fl}} = R^{\mathrm{fl}}(1-\rho^{\mathrm{fl}}), \qquad \text{with} \quad \rho^{\mathrm{fl}} = \lambda\sigma/R^{\mathrm{fl}},$$

where λ is the total arrival rate of flows in the cell.

We now extend the analysis of the quasi stationary and fluid limits to our finite population OCOF model.

2.2 OCOF model in the quasi stationary limit

In the quasi stationary analysis of the OCOF model, at the time scale of the flow dynamics the locations of the users are fixed. That is, the numbers of customers m_k in different annular rings can be considered fixed from the point of view of flow dynamics. We define the population vector $\mathbf{m} = (m_1, m_2, \ldots, m_K)$, which we alternatively call a constellation. In the sequel, we use for any vector the notation $|\mathbf{m}| = \sum_k m_k$. A feasible constellation satisfies $|\mathbf{m}| = n$. As detailed below, we can calculate the mean response time of a flow $\mathrm{E}[T\,|\,\mathbf{m}]$ conditioned on the population vector $\mathbf{m}$. Then, on the slower time scale we have to take into account that, after all, due to the mobility, the users do change their positions. This amounts to saying that in order to get the mean response time we have to average $\mathrm{E}[T\,|\,\mathbf{m}]$ over different constellations,

$$\mathrm{E}[T] = \sum_{\mathbf{m}} p(\mathbf{m})\mathrm{E}[T\,|\,\mathbf{m}], \tag{1}$$

where $p(\mathbf{m})$ is the probability of occurrence of the constellation $\mathbf{m}$. Note that in the infinite population model of [1] this kind of "outer averaging" or deconditioning is not needed since the infinite uniform population looks the same even after the motion. That is, all the constellations are identical.

Assuming the mobility of the users is such that the user location distribution is uniform the probability of finding a customer in annular ring k is $p_k = a_k/A$, where a_k is the area of ring k and A that of the whole cell (this can be easily extended to non-uniform distributions as well). As the users are assumed to move independently, the probability of a given constellation is obtained from the multinomial distribution [5],

$$p(\mathbf{m}) = \frac{n!}{m_1!m_2!\cdots m_K!}\, p_1^{m_1} p_2^{m_2} \cdots p_K^{m_K} \quad \text{for } |\mathbf{m}| = n. \tag{2}$$

For a fixed constellation $\mathbf{m}$, the system constitutes a K-class two-node closed network, shown in Figure 2. Node 1 is a multi-class PS queue and represents the users in active transmission phase. Node 0 represents the users in the thinking stage and is an IS (infinite server) queue. In node 1, a class-k customer has the service time $S_k = \sigma/R_k$. In node 0, all the customers have the same service rate ν.

The state of the system is defined by the state vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_K), \qquad \mathbf{x} \le \mathbf{m},$$

where x_k is the number of class-k customers in node 1. The number of customers in node 0 is then given by the vector $\mathbf{m} - \mathbf{x}$.

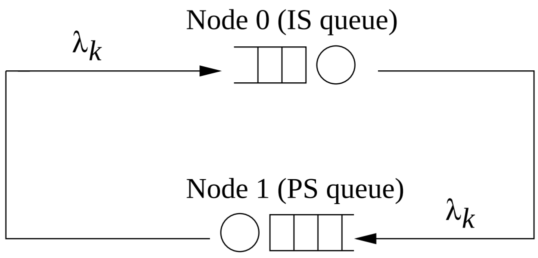


Fig. 2. The multi-class closed network model of the system with one IS node and one PS node

From the BCMP theorem [5] we obtain the stationary distribution of the whole network in constellation $\mathbf{m}$:

$$\pi_{\mathbf{m}}(\mathbf{x}) = G(\mathbf{m})^{-1}\chi_{\mathbf{m}}(\mathbf{x}),$$

where $\chi_{\mathbf{m}}(\mathbf{x})$ is the unnormalized distribution

$$\chi_{\mathbf{m}}(\mathbf{x}) = |\mathbf{x}|! \prod_{k=1}^{K} \frac{\rho_k^{x_k}}{x_k!(m_k - x_k)!}, \qquad \text{with} \quad \rho_k = \frac{\nu\sigma}{R_k},$$

and $G(\mathbf{m})$ is the normalization constant

$$G(\mathbf{m}) = \sum_{\mathbf{x}\le\mathbf{m}} \chi_{\mathbf{m}}(\mathbf{x}).$$

Note that for our closed network of two symmetric queues, the state distribution is insensitive to the flow size and thinking time distributions.

From the normalization constant $G(\mathbf{m})$ we can derive the mean response times of flows in different classes. By Little's formula we have for a given constellation $\mathbf{m}$

$$\mathrm{E}\left[T_k \,|\, \mathbf{m}\right] = \mathrm{E}\left[X_k \,|\, \mathbf{m}\right] / \lambda_k(\mathbf{m}),$$

where $\mathrm{E}\left[X_k \,|\, \mathbf{m}\right]$ is the mean number of class-k flows in node 1 and $\lambda_k(\mathbf{m})$ is the rate of circulation of class-k customers in the network. One easily finds

$$\mathrm{E}\left[X_k \,|\, \mathbf{m}\right] = \frac{\frac{\partial}{\partial\rho_k}G(\mathbf{m})}{G(\mathbf{m})}\,\rho_k = m_k - \frac{G(\mathbf{m}-\mathbf{e}_k)}{G(\mathbf{m})},$$

where $\mathbf{e}_k$ is a K-vector with one in the kth position and zero elsewhere. Further, for the rate parameters we have

$$\lambda_k(\mathbf{m}) = \left(m_k - \mathrm{E}\left[X_k \,|\, \mathbf{m}\right]\right)\nu = \frac{G(\mathbf{m}-\mathbf{e}_k)}{G(\mathbf{m})}\,\nu,$$

where the expression in the parentheses is recognized to be the mean number of class-k customers in thinking stage. Thus we obtain

$$\mathrm{E}\left[T_k \,|\, \mathbf{m}\right] = \frac{\frac{\partial}{\partial\rho_k}G(\mathbf{m})}{G(\mathbf{m}-\mathbf{e}_k)}\,\frac{\sigma}{R_k} = \left(m_k\,\frac{G(\mathbf{m})}{G(\mathbf{m}-\mathbf{e}_k)} - 1\right)\frac{1}{\nu}.$$

The mean flow response time in constellation $\mathbf{m}$ is

$$\mathrm{E}\,[T\,|\,\mathbf{m}] = \frac{1}{\sum_k \lambda_k(\mathbf{m})} \sum_k \lambda_k(\mathbf{m}) \mathrm{E}\,[T_k\,|\,\mathbf{m}] = \frac{\sum_k \mathrm{E}\,[X_k\,|\,\mathbf{m}]}{n - \sum_k \mathrm{E}\,[X_k\,|\,\mathbf{m}]} \frac{1}{\nu}.$$

This, together with (2), can be used in (1) to obtain the overall mean response time $\mathrm{E}\,[T]$ averaged over all constellations.

Similarly we can calculate the mean circulation rate of customers (i.e. the carried traffic) averaged over all constellations

$$\lambda = \sum_{\mathbf{m}} p(\mathbf{m}) \sum_k \lambda_k(\mathbf{m}) = \sum_{\mathbf{m}} p(\mathbf{m})(n - \sum_k \mathrm{E}\,[X_k\,|\,\mathbf{m}])\nu.$$

Note that λ approaches a finite limit $\frac{1}{\sigma}\sum_k p_k R_k$ when $\nu \to \infty$. This is because, for large ν, all n customers reside in node 1 and each of them obtains the share $1/n$ of the feasible rate, whence $\lambda_k(\mathbf{m}) = \frac{m_k}{n}\frac{R_k}{\sigma}$. The result then follows since $\frac{1}{n}\sum_{\mathbf{m}} p(\mathbf{m}) m_k = p_k$.

2.3 OCOF model in the fluid limit

In the fluid regime, all users sample different positions very fast in comparison with the time scale of flow dynamics and therefore effectively experience a common constant feasible rate $\bar{R}$,

$$\bar{R} = \sum_{k=1}^{K} p_k R_k, \tag{3}$$

where $p_k = a_k/A$, as before. Thus, the system reduces to a single class, two-station closed network, and the results can be obtained as a special case from the previous ones. In particular, we have

$$\mathrm{E}\,[T] = \frac{\frac{\partial}{\partial\rho}G(n)}{G(n-1)} \frac{\sigma}{\bar{R}}, \qquad \lambda = \frac{G(n-1)}{G(n)}\,\nu,$$

where

$$G(n) = \sum_{x=0}^{n} \frac{\rho^x}{(n-x)!}, \qquad \text{with} \quad \rho = \frac{\nu\sigma}{\bar{R}}.$$

Trivially, the carried traffic cannot exceed the limit $\lambda = \bar{R}/\sigma$, obtained when $\nu \to \infty$. This is the same upper limit for carried traffic we found above for the QS regime. Interestingly, this also equals the stability limit in the FL regime of both the BBP model and the OCMF model (to be discussed below).

3 OCMF finite population model

In the previous section we evaluated the response time for the OCOF model, where only one flow per customer can be active at any time. However, sometimes a customer could receive or generate new flows during the duration of the old flow. For this purpose, we introduce the OCMF (one customer, multiple flows) finite population model. In all other respects the model is the same as the OCOF model but now we assume that each of the n customers generates flows according to a Poisson process with intensity λ/n so that the total flow arrival rate in the cell is λ. There is no thinking time in this model and in some sense it is an intermediate form between the original infinite population model of [1] and the OCOF finite population model.

3.1 OCMF model in the quasi stationary limit

For a fixed constellation $\mathbf{m}$, the system reduces to an ordinary M/G/1-PS queue with the mean service time and load

$$\bar{S}(\mathbf{m}) = \frac{1}{n}\sum_{k=1}^{K} m_k S_k, \qquad \rho(\mathbf{m}) = \lambda \bar{S}(\mathbf{m}),$$

where, as before, $S_k = \sigma/R_k$. Thus the conditional response time is given by

$$\mathrm{E}\,[T\,|\,\mathbf{m}] = \frac{\bar{S}(\mathbf{m})}{1-\rho(\mathbf{m})},$$

and the overall mean response time averaged over all constellations is again given by (1).

Because in the OCMF model the data flows arrive as a Poisson process, it is necessary to consider the condition for the stability of the system (whereas the system in the OCOF model is self-regulating and always stable). In order for the QS approximation to make sense the system has to be stable, $\rho(\mathbf{m}) < 1$, for every constellation $\mathbf{m}$. The worst case is that all the customers are in the outermost ring with the lowest feasible rate R_K. This leads to the condition $\lambda\sigma/R_K \le 1$.

3.2 OCMF model in the fluid limit

In the fluid regime, all the customers have the same feasible rate $\bar{R}$ of eq. (3), whence the mean service time is $\bar{S} = \sigma/\bar{R}$. So the mean response time is:

$$\mathrm{E}\,[T] = \frac{\sigma}{(1-\rho)\bar{R}}, \qquad \text{with} \quad \rho = \frac{\lambda\sigma}{\bar{R}}.$$

The ordinary stability condition reads: $\lambda\sigma/\bar{R} \le 1$. It is easy to see that, in fact, in this case of fluid regime our OCMF model is equivalent to the original model of [1]; it does not matter whether a Poisson rate λ of flows with a constant service rate originates from a finite or infinite population of customers.

4 Numerical results

For numerical comparisons we consider a cell with radius $r = 15$ km divided into $K = 10$ annular rings, and a finite population of $n = 5$ users transmitting flows of mean size $\sigma = 1$ Mbit. The outer radii r_k and the feasible rates R_k of the rings are given in Table 1.

Table 1. Parameters of the studied system

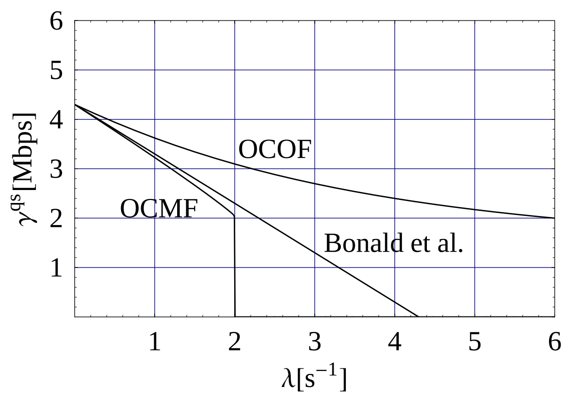
k	1	2	3	4	5	6	7	8	9	10
r_k/km	6	7	8	9	10	11	12	13	14	15
R_k/Mbps	8.0	7.5	7.0	6.5	6.0	5.5	5.0	4.0	3.0	2.0

As noted above, the system described by the OCOF model is always stable since in a closed network the number of flows is inherently limited. In the models with Poisson arrivals (i.e. the BBP [1] and OCMF models), however, the total arrival rate λ has to be limited to values less than a critical value λ_{cr} in order to guarantee stability, where

$$\lambda_{\mathrm{cr}} = \begin{cases} \frac{1}{\sigma}\bar{R} = \frac{1}{\sigma}\sum_k p_k R_k = 5.26\ \mathrm{s}^{-1} & \text{FL regime (both models)} \\ \frac{1}{\sigma}R^{\mathrm{q}s} = \frac{1}{\sigma}\left(\sum_k \frac{p_k}{R_k}\right)^{-1} = 4.30\ \mathrm{s}^{-1} & \text{QS regime, BBP model} \\ \frac{1}{\sigma}R_K = \frac{1}{\sigma}\min_k R_k = 2.00\ \mathrm{s}^{-1} & \text{QS regime, OCMF model} \end{cases}$$

In comparing a finite population model with a model with Poisson arrivals, it is not completely obvious how the parameters should be selected in order to put different models on the same footing. Comparing the BBP model with the OCMF flow model is straightforward: the parameter λ, representing both the offered and carried traffic in these models, should be set equal in the models. In the OCOF model, however, the carried traffic is only indirectly determined by the internal parameter ν. Here we choose to simply set the *intended* traffic $n\nu$ equal to λ of the other models, i.e. $\nu = \lambda/n$

The quantity by which we assess the performance is the flow throughput $\gamma = \sigma/\mathrm{E}[T]$, i.e. essentially the inverse of the flow response time. It is easy to see, that in the low load limit $\lambda \to 0$ all the three systems are identical. In this limit, the customers of in the OCOF model are almost always in the thinking phase, and node 0 feeds queue 1 with a Poisson stream of flows with intensity equalling the intended traffic.



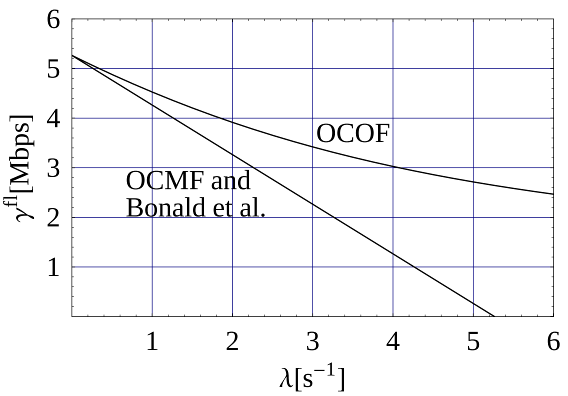


Fig. 3. Comparison of the flow throughputs of different models as a function of λ in the QS (left) and FL (right) regimes.

For finite values of λ, however, the models differ. The result of the throughput comparison of the three models is shown in Figure 3 in the QS (left) and FL (right) regimes. We notice that the OCOF model has always the highest flow throughput. There are two reasons for this result. First, in this comparison, as explained above, we have set the intended loads of different models equal. In the OCMF and BBP models the intended load is identical to the carried load, but in the OCOF model the carried load is smaller. This is shown in Figure 4 (left graph), where the carried load of the OCOF model is plotted against the intended load, in both QS and FL regimes. When fewer flows are carried, this is clearly favourable for the response time of those actually carried. This alone, however, does not wholly explain the difference. We have also made a comparison, not shown here, where the intensities of the carried traffic in the three models are enforced to be equal (by a proper choice of the parameter ν). This brings the curve for the OCOF model somewhat lower, and notably to go rather sharply to zero at the value $\frac{1}{\sigma}\sum_k p_k R_k = 5.26\ \mathrm{s}^{-1}$ representing the highest possible value of the carried traffic as discussed in subsections 2.2 and 2.3. Still the throughput for the OCOF model remains above the others.

The second reason is that, while in the OCMF model and the BBP model the spatial distribution of the traffic is fixed in advance, the self-regulation of the traffic in the OCOF model leads to most of the traffic going to where it is easiest to get through, i.e. in the central region close to the base station. Thus, a larger proportion of the flows experience a shorter response time.

In comparing the OCMF and BBP models, we have already noted that in the FL regime they are identical. In the QS regime the differ slightly for small values of λ. This difference in a way represents the pure effect of the finiteness of the user population: all the flows go through a finite set of discrete points in the disc (as opposed to being randomly located). These points may be unfavourably located with the overall effect of lowering the throughput. This effect is extreme at the point where the most unfavourable configuration, where all the users are located at the outermost ring, becomes unstable. Though very rare, this event makes the whole system unstable. This happens, in our example at $\lambda = 2\ \mathrm{s}^{-1}$, where the throughput drops rapidly to zero. The effect of the finiteness of the user population is also present in the results for the OCOF model, but

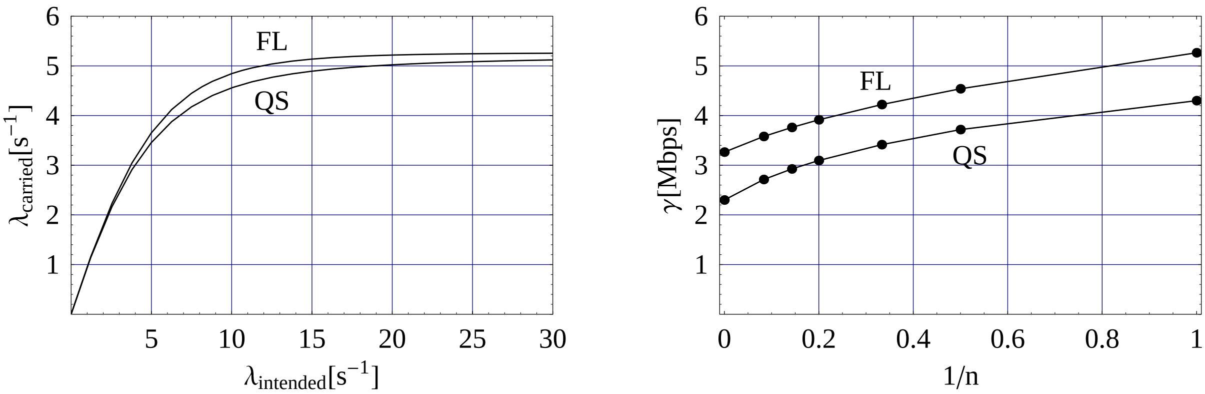


Fig. 4. Left: Carried traffic as a function of the intended traffic $n\nu$ for the OCOF model in the QS and FL regimes. Right: Convergence of the flow throughput γ of the OCOF model in the QS and FL regimes as a function of $1/n$ for $n = 1, 2, 3, 5, 7, 12$ with $\lambda = 2\ \mathrm{s}^{-1}$. The points for $1/n = 0$, in respective regimes, are from the BBP model.

the system in this case is never unstable and the effect is overshadowed by the two other effects discussed above.

Finally, the graph on the right in Figure 4 shows how the OCOF model tends to the infinite population BBP model as the number of users n grows (with the intended traffic $\lambda = 2\ \mathrm{s}^{-1}$). We see that already for a relatively small population, of the order of 10, the performance is rather close to that of the infinite population case. Obviously, the BBP model also represent a limit case of the OCMF model.

5 Conclusions

We have introduced and explored two finite population models for mobile users sending data traffic in a cell of a wireless system. These models extend the work by Bonald et al. [1] where flows arrive as a spatio-temporal Poisson process. In the OCOF model, only one flow is allowed for each customer at any given time, and consecutive flows sent by a customer are separated by a thinking time. In the OCMF model, each customer sends flows according to a Poisson process: the total traffic is just split between the customers.

As already noted in [1], the performance of the system is analytically tractable only in the limits of very slow or very fast mobility, called the QS and FL regimes, and, moreover, is insensitive in these two regimes, i.e. independent of detailed traffic characteristics. The main contribution of this paper is the derivation of the flow response time or the flow throughput for the OCOF model in these regimes; for the OCMF model the results are rather trivial.

The flow throughput results of the finite population OCOF and OCMF models are compared with those of the BBP model. It is found that the throughput of the BBP model is always between those of the two other models. In the QS regime the throughput of the OCMF model is slightly worse than that of the BBP model as a result of a pure finite population effect. The throughput of the OCOF model, however, is better than that of the others. This is due to the self-regulating nature of a model with thinking times, leading to smaller carried load and concentration of the traffic to the high throughput region of the cell.

References

1. T. Bonald, S. Borst, and A. Proutière, *How mobility impacts the flow-level performance of wireless data systems*, Proc. of IEEE Infocom 2004 (2004) 1872–1881.
2. T. Bonald, and A. Proutière, *Wireless downlink data channels: User performance and cell dimensioning* Proc. of Mobicom 2003 (2003) 339–352.
3. T. Bonald, S. Borst, N. Hedge, and A. Proutière, *Wireless data performance in multi-cell scenarios*, Proc. of Sigmetrics/Performance 2004 (2004) 378–388.
4. S. Borst, *User-level performance of channel-aware scheduling algorithms in wireless data networks*, Proc. of IEEE Infocom 2003 (2003) 321-331.
5. P. G. Harrison and N. M. Patel, *Performance Modelling of Communication Networks and Computer Architectures*, Wokingham Addison-Wesley (1994) 259–276.