An Analytical Model for Dynamic Inter-operator Resource Sharing in 4G Networks

Ahmet Cihat Toker, Fikret Sivrikaya, Nadim El Sayed, and Sahin Albayrak

Technische Universität Berlin,
Ernst-Reuter Platz 7 Berlin, Germany
{ahmet-cihat.toker,sahin.albayrak,fikret.sivrikaya,nadim.elsayed}
@dai-labor.de
http://www.dai-labor.de

Abstract. Realtime resource sharing between operators is an efficient approach to deal with the ever increasing wireless traffic. Simple and closed form formulas relating cooperation terms to the network performance measures are needed, so that the operators can take cooperate/not cooperate decisions, and track the benefits/losses of sharing. Analytical solutions to resource sharing the among access networks of a single operator exist. However the fact that operators will not share network internal data calls for solutions that are separable from each other. In this paper we provide such a closed form formulation, validate it with simulations and propose a simple negotiation mechanism between two operators utilizing this model. This simple model can be extended to model more complex interactions between operators, interaction between more than two operators, or can be used to evaluate long term cooperation policies.

Key words: resource sharing, 4G, performance modeling, cooperation, traffic engineering

1 Introduction

Telecommunication network management practices are strongly rooted in the monopolistic telecom operators. The liberalization of the operators has only changed the landscape in a way that there were multiple closed operators rather than one closed operator. This trend has had related implications on the network operator and the user side.

The operator networks are usually centrally managed, poorly integrated with outside components, and strictly isolated from external access. On the other hand the Internet was born out of the need for integrating networks. The exposure of users to the prolific Internet services means that similar service models will have to be provided by the next generation telecom networks. The clash between these two opposite approaches poses important challenges for network operators. This is due to the fundamental risk associated with their networks turning into mere bit-pipes. In order for future telecom networks to be economically viable, they should provide a similar user experience with Internet services, albeit in a more managed and reliable manner. Here lies the grand challenge of the so-called Telco 2.0 operators. The operators have to offer even more data intensive applications on their networks to make their operations profitable. This comes in a time, when the increasing data traffic is starting to hurt user experience, and pose itself as the biggest risk facing the operators.

There are three strategies that the network operators and broadband service providers can follow under these circumstances:

1. Competitive: Capacity expansion.
2. Cooperative: Employing untapped networking resources, such as community wireless networks or sharing new spectrum.
3. Cooperative: Establishing strategic partnerships with other operators to share the existing spectrum.

Clearly, capacity extension is a brute-force solution to the problem and can only be extended to the point when the investment costs drive access prices beyond market prices. The cooperation with community networks is also viable, under the condition that community network fees are below a threshold value [1]. This necessities a wide community participation, which would down the connection fees. Finally, the agreements between operators are off-line in nature and can only be reached after long legal and financial negotiations by the involved parties. Despite the difficulties associated with cooperative solutions, they represent a far more long term and sustainable solution compared to capacity extension.

We believe that the dynamic resource sharing between two licensed or virtual operators and cooperation between a licensed operator and a wireless community network are of similar nature, and provide the best solution to the operators capacity scarcity problem.

An important issue for the dynamic resource sharing is the ownership of digital identities of individual users. Current practices in the telecommunications area tie the users to a single operator even though the number of players in the market has long been growing. The users tend to manually combine their subscriptions to multiple operators in order to take simultaneous advantage of their different offers that are suited for a variety of services. *User-centric networking* is a new approach to the relation to the ownership of identity in next generation networking. In its most generic sense, the user-centric view in telecommunications considers that the users are free from subscription to any one network operator and are the real owners of their identities. This allows them to dynamically choose the most suitable transport infrastructure from the available network providers for their terminal and application requirements. An important result of this is the churn rates approaching session initiation rates. Therefore dynamic resource sharing is an important aspect of user-centric networking.

The PERIMETER project [2], funded by the European Union under the Framework Program 7 (FP7), has been investigating such user-centric networking paradigm for future telecommunication networks. The key innovation in PERIMETER project is the introduction of a distributed database implemented on user terminals, which stores the *Quality of Experience* (QoE) reports associated with a certain network generated by the users. The users utilize the reports of current and past users of networks to decide which network to choose. This database can be used to overcome important challenges for the introduction of dynamic resource sharing, which are:

1. Lack of analytical solutions to model load balancing,
2. Information asymmetry and lack of transparency between different operators.

The dynamic nature of the problem requires analytical solutions available to the operator networks to take realtime cooperate /do not cooperate decisions. The availability of the open quality of experience database to the operators alleviates the information asymmetry problem. The operators can exploit this database in order to gather information about the other operators, with whom they may cooperate. With this information at hand, the operators need analytical models that can be evaluated in real-time in order to take the cooperation decisions. These analytical models should be separable, meaning that the operators should be able to verify them by using information available to both of them.

In this paper we present such an analytical model its verification based on discrete event simulations. In the next section we introduce the requirements and motivations of the analytical model. We formulate the problem formally in Section 3 and compare it to the state of the art in Section 4. We propose the analytical model in Section 5 and apply it to a single class resource sharing scenario in Section 6. We provide a simple negotiation mechanism for the operators in Section 7 and conclude with the evaluation of our model with simulation in Section 8.

2 Motivation and Requirements

The problem we are addressing is the minimization or avoidance of possible degradation in user perceived quality of experience in an access network as the number of users increases in an open user-centric network environment. The delay a session request experiences is a common performance measure that can be used to handle a variety of service types. Therefore, we choose the delay as the performance metric of dynamic resource sharing mechanisms.

The method with which the avoidance or minimization is achieved is by borrowing network layer resources from an access network that belong to another operator (community, virtual or real operator). In a user-centric environment, the operators have to find additional resources, not to degrade the QoE, otherwise the users will be moving away to alternative operators. What would be the incentives for the donor operator to lend some of its resources to the borrower? A quick answer would be that if the donor operator is under-utilized at that particular point of time, then it could increase its utilization to a point where it still can serve its current users, thereby increasing its revenues. However, the challenge of user centricity comes from the fact that users can instantaneously decide on the operators they choose. The donor operator may choose to ignore the borrowing operator, in an attempt to drive the QoE in the borrowing network down, and gain more users. Therefore the dynamics of the resource sharing between two operators become strategy dependent, and not trivial.

The aforementioned problem is not specific to the dynamic resource sharing in user-centric networking. As Dohler discusses in [3], the problem is not only technological, but also strategic. We adhere to Dohler's approach, in which he proposes that the success of any cooperative solution to any communications problem is more possible if the cooperation decisions are taken by software agents, and the benefits of these decisions are clear to the owners of these agents. Therefore we derive a simple and intuitive formula for the relation between the average delay and cooperation parameters.

Another important requirement to the model is the separability. Under separability we mean the following. Generally, the state space describing two independent systems interacting with each other is two dimensional. The solution of the probability distribution of such a distribution requires the knowledge of the states of the independent systems. This is not possible for two cooperating operators, since the operators will be reluctant to share their operation information with each other. If the performance metrics can be calculated by openly available information, without requiring the knowledge about the operative status of the other operators we call such a model separable.

Final requirement on the modeling approach is the capability to efficiently model heterogenous wireless networks. It is foreseen that the 4G networks will be composed of heterogenous wireless access technologies. Therefore the model should be able to accommodate a variety of them.

3 Problem Formulation

i	operator index, $i \epsilon A, B$.
P_i	Probability that users prefer Operator i.
$P_{b,i}$	Blocking probability in Operator i.
$P_{T,i}$	Transfer probability in Operator i.
$D_{\max,i}$	Maximum allowed average delay in Operator i.
$1/\lambda$	External average inter-arrival time ($secs$).
$1/\mu$	Average request size($bits$).

Table 1: Model variables.

In this section we define formally the abstraction level we employ in modeling the problem of dynamic resource sharing in user-centric networking. We consider a location where the users have two wireless operators to choose from, operator A and operator B. The users generate requests with exponential inter arrival times of mean $1/\lambda$, which have sizes that are also exponentially distributed with mean $1/\mu$. Since these users are not in contractual agreements with the operators, they can choose either one of the operators with probabilities P_A and P_B. The operators utilize a call admission control (CAC) mechanism that block incoming requests with probability P_b, or transfer to the other operator with a probability P_T. The operators employ these techniques in order to provide a maximum delay guarantee to the users given by $D_{\max}$. The variables of the model are summarized in Table 1.

Considering the requirements we listed in Section 2, we have decided to use Queueing Networks [4] as the modeling framework. Queueing networks are a generalization of the classical queueing systems. They are concerned with the performance modeling of systems that are composed of interconnected queueing stations. The main goal of the queueing networks is the derivation of the joint probability distribution of the number of jobs in each service station. One of the most important results of Queueing Networks is the Baskett, Chandy, Muntz and Palacios (BCMP) theorem, named after the researchers who jointly developed the theorem in [5]. The theorem states that the joint probability distribution of a queueing network can be written as the product of marginal probability distributions of the individual service stations. Furthermore, each service stations behaves like a traditional M/M/1 queue, with a modified input traffic rate, reflecting the network topology. This formulation satisfies the separability and simplicity requirements.

Processor Sharing (PS) is a queueing service discipline first analyzed by Kleinrock in [6]. It is an idealized version of the Round Robin (RR) service discipline, in which the service quanta that each job in the queue receives is infinitesimal. In the limit the server operates in a manner that each job receives a service rate equivalent to the overall server capacity divided by the number of jobs in the queue. We have chosen to model individual radio access networks by PS discipline. The intuition that the wireless access network can be modeled as a service station that simultaneously serve the active users can be traced back to the work of Telatar [7]. Furthermore, it has been shown that the Weighted Fair Queueing service discipline, used widely in radio network base stations, approximates processor sharing when the packet size is small compared to the session size [8].

4 State of The Art

Historically, queueing networks have provided very attractive models for a wide variety of communication networks and applications running on these networks. Early applications include [9] Conway's queueing model for the performance evaluation of Signalling System 7 (SS7).

One of the most active application of queueing networks to the 4G networking has been the work of Kouvastos et al. Kouvastos first employed a queueing network model to analyze the performance of ATM Asynchronous Transfer Mode switches developed for the ISDN Integrated Services Digital Networkss [10]. The challenge he addressed was the development of an analytical performance model for the switches that can be used during the design and development phase. The model regarded the switching matrix at the core of the switch as a medium that had to be shared among flows of different service types.

The author applied the same queueing model first to performance analysis of GSM/GPRS base stations [11], and subsequently to the performance analysis of hypothetical 4G base stations accommodating different services [12] and [13]. A hypothetical 4G cell is modeled as a combination of three service centers. Voice calls are handled by a classical Erlang loss system, the data calls are handled by a PS (Processor Sharing) system and finally streaming calls are handled by a FCFS (First Come First Serve) system. These service centers exchange resources among themselves according to the state of the cell, which consists of $\mathbf{n} = (n_1, n_2, n_3)$ where n_1, n_2, n_3 are the number of calls in the respective service centers. The state space is three dimensional and therefore not analytically tractable. The authors make use of the maximum entropy principle to find a product form approximation that yields a closed form solution. Maximum entropy principle was introduced by Jaynes in 1960s [14], and can be seen as an equivalence principle between statistical and information theoretical entropy definitions. It states that given a constraint on the mean values of a family of probability distributions that may describe a physical process, the distribution that maximizes the entropy is the least biased estimate of the probability distribution. The constraints on the mean values under investigation are:

$$\lambda_i(1 - p_i) = \mu_i U_i, \quad i = 1, 2, 3 \tag{1}$$

where λ_i is the overall flow into a server center, p_i is the blocking probability, μ_i is the service rate and U_i is the utilization of the service center associated with the service types described before. This equality is a reformulation of global balance conditions, equating the flows into and out of a state. The goal is to find the probability distribution of $\mathbf{n}$. Maximum entropy method involves solving the maximization problem with the method of Lagrange multipliers. These Lagrange multipliers are then used to formulate equivalent flows into the individual service centers, which can then be analyzed independently. Similar to this work, we use PS service stations to model heterogeneous access networks. Our model not only takes into account multiple access networks, but also different operators. Furthermore, we use the BCMP method to provide exact solution to the queueing model. Our solution is also computationally more efficient than this, as it does not include any recursive solutions.

Fukushima et al. present an innovative application of queueing networks to the wireless communications domain in [15]. Specifically, they employ queueing networks to model the interplay between user mobility and bursty nature of packet traffic in wireless systems. The mobility of the users between cells and their service requests evolve with different timescales. The authors utilize queueing networks to model these two aspects jointly. In their formulation each cell has two service centers. The first service center, which is an infinite server (IS) center, models user mobility, where users move from one infinite server to the other based on routing probabilities obtained from an external mobility model. The second server is a PS center, which models the sharing of base station transmission capacity among the service requests of users in the cell. The service demand at the second service station is a function of the state of the first service station. It can be said that the authors use a queueing network with state-dependent routing. The state dependency expresses itself as non-linear traffic equations, which can be solved using fixed point iterative methods. The authors use this model to analyze a hierarchical WLAN-cellular integration, with static *"WLAN first"*policy, similar to [16].

In [16]the authors provide an analysis of a hierarchically integrated WLAN and cellular network by employing queueing networks. In a hierarchical integration architecture [17], the WLAN cells are used as high-speed hot spots, and are carefully positioned as an overlay on the cellular infrastructure. The natural question that arises in such an architecture is when the overlay will be used. Assuming that the network operator has the final say on this decision, the authors employ the *"WLAN first"* resource allocation policy. In this policy the calls are admitted to the WLAN cell as long as the capacity is not reached. The calls are then admitted to the underlay cellular cell, once the capacity is reached. The authors assume that the data calls can be modeled by discrete bandwidth units they fill in different subsystems. With this assumption, they are able to model the individual overlay and underlay cells as classical Erlang loss-systems. The availability of a static allocation policy, the interaction between different cells are modeled by static and additive traffic flows. The authors mention that the overall system is represented by a multi-dimensional Markov chain, but do not propose a solution for the global equations. Subsequently, no relation between the marginal probabilities and the global probabilities are presented. Based on the marginal probabilities, the authors are able to derive closed-form blocking probabilities. Both of these works consider a single operator owning the heterogeneous access networks.

Early applications of BCMP theorem to data communication networks include the analysis of general sliding window type flow control, such as the one used in TCP by Reiser [18]. Recent application fields include the modeling of multi-tier Internet applications and the operational optimization of data centers hosting these services. A typical Internet application is provisioned by a multi-tier arrangement of servers. In the front end is the load balancer that routes the application requests to replicated first tier web servers. The first tier web servers route the session requests to the second tier application servers that host the applications are replicated. Finally, the application servers use the third tier data base servers to compose the applications. In [19] Urgaonkar models this architecture via a single class closed BCMP network and present analytic solutions for the average delay. Data centers that host these applications on identical server clusters can also be modeled as closed BCMP networks. In this case the server clusters become the service centers, and the different applications represent the different applications. The authors use such an analytical model to optimize the energy use and scaling of data servers in [20]. To the best of our knowledge BCMP networks have not been applied in modeling 4G networks.

5 Analytical Model

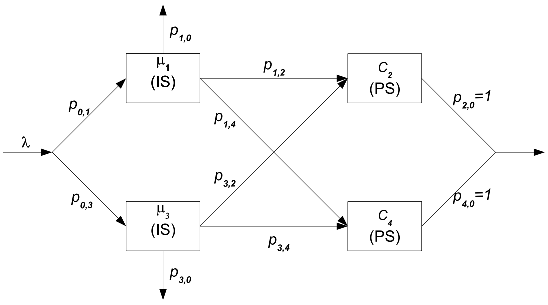


Fig. 1: The queueing network model.

In order to develop a tractable, closed form, and separable solution for the delay performance metric we use the BCMP theorem on the queueing network depicted in Figure 1. Each wireless operator is modeled by a tandem of queues. The first stage IS queue, represents the CAC decision making. The traffic exiting the first stage queue enters the PS queue which jointly models the shared air interface and access router that connects the base station to the backbone.

The first step of applying the BCMP model is the solution of the traffic equations. For open networks they can be expressed as:

$$\lambda_i = \lambda \cdot p_{0,i} + \sum_{j=1}^{4} \lambda_j \cdot p_{j,i}, \quad i \in \{1, 2, 3, 4\}. \tag{2}$$

λ is the external arrival rate, and $p_{i,j}$ represent the routing probability between service stations i and j where the subscript 0 denotes the external environment. By making the relevant substitutions we obtain the expressions for the input rates of the different service stations as:

$$\begin{aligned}
\lambda_1 &= \lambda \cdot p_{0,1}.\\
\lambda_2 &= \lambda \cdot p_{0,1} \cdot p_{1,2} + \lambda \cdot p_{0,3} \cdot p_{3,2}.\\
\lambda_3 &= \lambda \cdot p_{0,3}.\\
\lambda_4 &= \lambda \cdot p_{0,3} \cdot p_{3,4} + \lambda \cdot p_{0,1} \cdot p_{1,4}.
\end{aligned} \tag{3}$$

One can calculate the utilizations of individual service stations by using the service rates of the individual server stations, $\rho_i = \frac{\lambda_i}{\mu_i}$. Once these modified utilizations are computed, the joint probability distribution can be written as:

$$\pi(n_1, n_2, n_3, n_4) = \prod_{i=1,3} e^{-\rho_i} \frac{\rho_i^{n_i}}{n_i!} \cdot \prod_{i=2,4} (1 - \rho_i)\rho_i^{n_i}. \tag{4}$$

Let us substitute the routing probabilities of the BCMP model with the system parameters we defined in Section 3. The indices $i = 1, 2$ describe the operator A and $i = 3, 4$ the operator B. $p_{0,1}$ and $p_{0,3}$ correspond to the users operator preferences, P_A and P_B respectively. $p_{3,2}$ and $p_{1,4}$ are the transfer ratios of the operators, $P_{T,A}$ and $P_{T,B}$. The CAC mechanisms reject calls with probabilities $P_{b,A}$ and $P_{b,B}$, hence we have $P_{1,0} = P_{b,A}$ and $P_{3,0} = P_{b,B}$. By using total probability principle we obtain $P_{1,2} = 1 - P_{T,A} - P_{b,A}$ and $P_{3,4} = 1 - P_{T,B} - P_{b,B}$. Let us analyze the utilization of the PS part of the individual operators. Given that the users generate requests whose sizes are distributed according to an exponential distribution of average μ bits and the operators access networks have a capacity of C_A and C_B bits per second, we have $\mu_1 = \mu \cdot C_A$ and $\mu_2 = \mu \cdot C_B$. Thus we have:

$$\begin{aligned}
\rho_{PS,A} &= (P_A \cdot (1 - P_{T,A} - P_{b,A}) + P_B \cdot P_{T,B}) \cdot \tfrac{\lambda}{\mu C_A}.\\
\rho_{PS,B} &= (P_B \cdot (1 - P_{T,B} - P_{b,B}) + P_A \cdot P_{T,A}) \cdot \tfrac{\lambda}{\mu C_B}.
\end{aligned} \tag{5}$$

These equations show an intuitive and linear relationship between operator utilizations and transfer and blocking probabilities. Specifically, an operator may increase its utilization by accepting additional traffic from the other operator, or may decrease its utilization by transferring traffic to the other operator or by increasing the CAC level. When increasing the utilization, the condition $\rho_{PS,A}, \rho_{PS,B} < 1$ should be considered to guarantee stability. These utilizations can be used to find the expected delay conditioned on the request size x given in bits at the PS side of the operators. These are given by:

$$
\begin{aligned}
D_{PS,A}(x) &= \frac{x/C_A}{1-\rho_{PS,A}}. \\
D_{PS,B}(x) &= \frac{x/C_B}{1-\rho_{PS,B}}.
\end{aligned}
\tag{6}
$$

6 Application of the Model to Single Class Resource Sharing

In a single service class scenario, the sharing of resources to avoid overload situations becomes mutually exclusive with the under-utilization situations. This means that the borrowing operator will not donate resources, and the donor operator will not borrow resources from each other. Let us arbitrarily assign operator A to be the donor operator, and operator B to be the borrowing operator. The borrower operator A borrows resources and sends a given portion of traffic to the donor operator B, who accepts additional traffic. Furthermore, let us assume the average service demand x is fixed.

In this case the delays become a function of the transfer probability P_T and the operator preferences of the users P_A, P_B. This means that the exchange of resources is one way, i.e. $P_{T,A} = 0$. The problem is characterized the relative operator preferences of the users P_A, P_B; the delay thresholds $D_{\max,A}, D_{\max,B}$; and the CAC probabilities $P_{b,A}$ and $P_{b,B}$.

In a resource sharing scenario, the donor operator has to find the amount of traffic it can accept, equivalently the amount of resources it can donate, without increasing the expected delay of the already accepted users above a threshold $D_{\max,A}$, described by (7). The donor operator finds the maximum P_T value that satisfies this condition described by which we term $P_{T,D}$:

$$
D_A(P_{T,D}) = D_{PS_A}(P_{T,D}) = D_{\max,A}. \tag{7}
$$

On the other side the borrowing operator has to calculate how much traffic it should transfer, or equivalently how much resources it should borrow, in order to reduce the expected delay in its access network to a threshold. However, it also has to consider the increase in delay it induces on the donor operators. We are considering a *seamless* scenario, in which the transferred users are not aware of the fact that their session request is served by an alternative operator. The borrowing operator should not load the donor operator excessively, since this excessive loading would increase the delays experienced by the transferred users, who would associate this with the borrowing operator.

There are two approaches for considering this aspect. One can consider the maximum of the delays in the two networks as in (8).

$$
\max\{D_{PS_A}(P_{T,R}), D_{PS_B}(P_{T,R})\} \le D_{\max,B}(x). \tag{8}
$$

An alternative is to consider the average delay as experienced by the users of the borrowing operator, as in (9).

$$E\{D_B(P_{T,R})\} = (1-P_{T,R})\cdot D_{PS_B}(P_{T,R})+P_{T,R}\cdot D_{PS_A}(P_{T,R}) \leq D_{\max,B}(x). \quad (9)$$

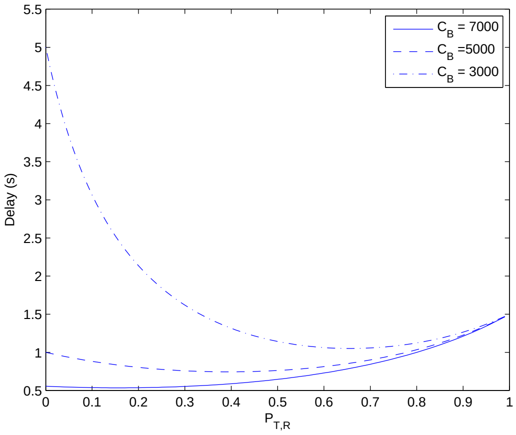


Fig. 2: The variation of average delay versus transfer probability for $C_A = 5000 bps, D_{A,init} = 0.6$.

In both cases, the borrowing operator chooses the transfer probability value $P_{T,R}$ that minimizes the delay. We choose the second option, as it is more fair to the users of the borrowing operator. In order to calculate $E\{D_B(P_{T,R})\}$ operator B has to be able to calculate the utilization of operator A, which requires the knowledge of $P_{b,A}$. This is an private information that is not available to operator B. However this can be solved by utilizing the open user experience database, proposed by PERIMETER. We assume operator B is able to gather the initial average delay in operator A, $D_{A,init}$, by exploiting the database. From this value it can calculate initial value of the utilization via (6). Employing (5) allows us to formulate the utilization of operator A after resource sharing, by consulting only to the publicly available average delay. Using this value one is able to write down the expression for $E\{D_B(P_{T,R})\}$:

$$E\{D_B(P_{T,R})\} = \frac{\frac{x}{C_A}\cdot P_{T,R}}{\frac{x}{C_A D_{A,init}} - \frac{P_B \lambda}{\mu C_A}\cdot P_{T,R}} + \frac{\frac{x}{C_B}\cdot(1-P_{T,R})}{1-P_B(1-P_{b,B}-P_{T,R})\frac{\lambda}{\mu C_B}}. \quad (10)$$

Equation 10 is plotted for varying borrowing operator capacities in Figure 2. It can be observed that the amount of reduction in the average delay is directly proportional with the capacity of the donor operator. For all cases the delay is a rational function of transfer probability with global minima.

7 A Simple Negotiation Mechanism for Single Class Resource Sharing

A simple mechanism for the interaction between the operators based on these values works as follows. The borrowing operator calculates the $P_{T,R}$ value that minimizes (10) and communicates this to the the donor operator. Donor operator checks if this value increases its average delay above the limit. If $P_{T,R}$ is acceptable for the donor operator, it replies with the same probability, which corresponds to an agreement. If not, the donor operator replies with the maximum $P_{T,D}$ that it can handle calculated from (7). An agreement is reached on this value. If the final agreed P_T is not enough for the borrower operator to reach its delay goals, it has to increase its blocking probability in order to meet its delay goals.

8 Simulation Results

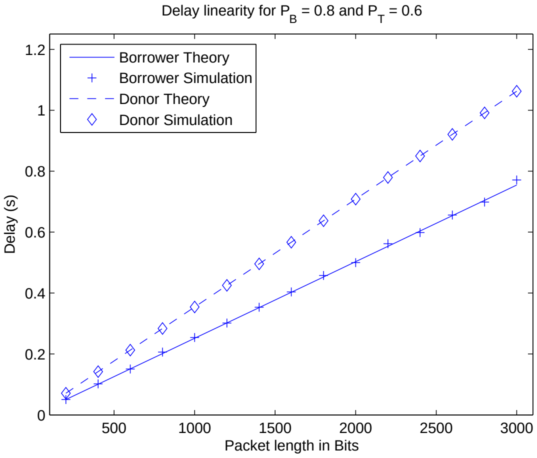
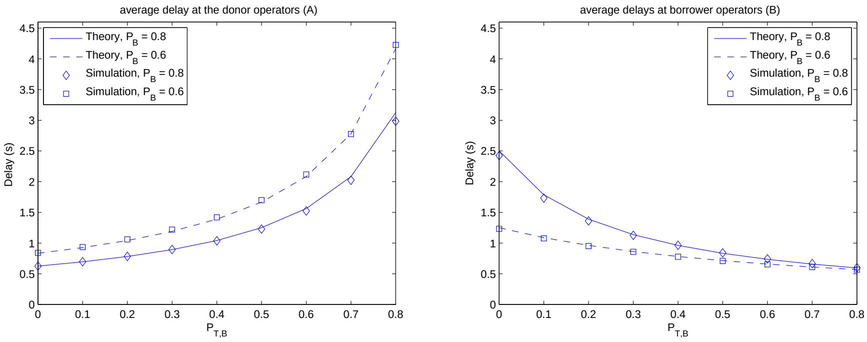


Fig. 3: Linearity of delay with packet size.

We simulated the system presented in Fig. 2 using OPNET Modeler 15.0 [21]. Users initiate session requests with an inter-arrival time of 0.5 seconds. These requests go according to the user preferences to the respective operator. At the operator the IS server is modeled as a decision entity without any queuing or delay, which forwards the session requests either to the operator's own PS-server or to the other operator's server, according to the transfer probability. The PS-server is modeled as a queue, which handles the session requests with the PS-queuing discipline. We implemented the PS queuing discipline in three steps.

First we modeled the queue of session requests simply as a set of those, where each one has his own estimated finish time. Second we defined the two events that can change the state of the PS server, namely arrivals and departures. Third we implemented the changes each event induces on the session requests and their respective estimated finish times as follows. On one hand an arrival adds an entry to the queue and prolongs the remaining estimated finished times of the elements of the queue. On the other hand, the minimum estimated finish time is the time at which a departure takes place, causing the correspondent session request to get removed from the queue on one side, and shortening the remaining estimated finish times for the other elements of the queue. After simulating ten hours of the system behavior we were able to collect the results depicted in Fig. 3 and 4.

According to [22] the PS-queuing discipline has one distinguishing property, that is, the delays are linear to the service demands for a given utilization as in Equations (6). In order to verify the correctness of our PS implementation, we checked if the developed model possess this property.We simulated the system with $\frac{1}{\lambda} = 0.5$, the average demand $\frac{1}{\mu}$ set to 2500, the server capacities for operator A and B set to 5000, the user preference of B set to $P_B = 0.8$ and the transfer probability $P_{TB} = 0.6$. The results depicted in Fig. 3, which verifies that our OPNET models conformance with PS discipline.



(a) Donor

(b) Borrower

Fig. 4: Comparison of analytical results with simulation.

We ran the simulations for the varying values of transfer probability and gathered the average delay in individual networks. In Fig. 4 we verify our theoretical results by comparing compare the simulation values to the theoretical results derived in (6). It can be observed that increasing transfer ratio reduces the delay in a diminishing manner, that is the effectiveness of dynamic resource sharing reduces with higher transfer probabilities. Furthermore, increasing the transfer ratio above the diminishing return boundary, the vicinity of $P_{T,B} = 0.4$

in this case, increases the donor operator delay substantially. This justifies the use of average delay over the operators defined in (10) as the decision criterium. For the value of $P_{T,B} = 0.4$, we are able to reduce borrower delay from 2.5 to 1.25 seconds, while increasing the donor delay from 0.6 to 0.8 seconds. We believe that this demonstrates the viability of dynamic resource sharing.

9 Conclusion and Future Work

In this paper we have presented a simple and separable analytical model for the dynamic resource sharing between operators. The simplicity of the model allows operators to take real-time decisions to cooperate or not to cooperate with other operators in the same geographic area, where coverage is shared. The separability of our solution allows each operator to base their decisions on openly available data. We have not exploited the separable solution to its fullest extent and have presented a solution based on the mean value of delays experienced by end users. We have shown analytically and using simulations, that the average expected delay can be reduced with dynamic resource sharing.

We will extend this work in order to exploit the separable solution to delay metric. Specifically, we will propose a solution in which operators take their decisions not based on average delay values, but on probabilities that delay will exceed the thresholds. Furthermore, we intend to extend this model to include multiple service classes. In this scenario a simultaneous borrowing and lending of resources will be of interest.

References

1. Manshaei, M.H., Marbach, P., Hubaux, J.P.: Evolution and market share of wireless community networks. (June 2009) 508–514
2. Toker, A.C., Cleary, F., Fiedler, M., Ridel, L., Yavuz, B.: Perimeter: Privacy-preserving contract-less, user centric, seamless roaming for always best connected future internet. In: Proceedings of 22th World Wireless Research Forum. (2009)
3. Dohler, M., Meddour, D.E., Senouci, S.M., Saadani, A.: Cooperation in 4g - hype or ripe? Technology and Society Magazine, IEEE **27**(1) (March 2008) 13–17
4. Bolch, G., Greiner, S., de Meer, H., Trivedi, S.: Queueing Networks and Markov Chains: Modeling and Performance Evaluation. 2 edn. Wiley (2006)
5. Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G.: Open, closed, and mixed networks of queues with different classes of customers. J. ACM **22**(2) (1975) 248–260
6. Kleinrock, L.: Time-shared systems: a theoretical treatment. J. ACM **14**(2) (1967) 242–261
7. Telatar, I.E., Gallager, R.G.: Combining queueing theory with information theory for multiaccess. Selected Areas in Communications, IEEE Journal on **13**(6) (1995) 963–969
8. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the single-node case. Networking, IEEE/ACM Transactions on **1**(3) (August 2002) 344–357

9. Queueing network modeling of signaling system No.7. In: Global Telecommunications Conference, 1990, and Exhibition. 'Communications: Connecting the Future', GLOBECOM '90., IEEE. (December 1990)
10. Kouvatsos, D.: Performance modelling and cost-effective analysis of multiservice integrated networks. **9**(3) (August 2002) 127–135
11. Kouvatsos, D.D., Awan, I., Al-Begain, K.: Performance modelling of gprs with bursty multiclass traffic. **150**(2) (May 2003) 75–85
12. Performance Modelling of a Wireless 4G Cell under GPS Scheme. In: Applications and the Internet Workshops, 2005. Saint Workshops 2005. The 2005 Symposium on. (January 2005)
13. Performance modeling of a wireless 4G cell under a GPS scheme with hand off. In: Next Generation Internet Networks, 2005. (April 2005)
14. Jaynes, E.T.: Prior probabilities. Systems Science and Cybernetics, IEEE Transactions on **4**(3) (February 2007) 227–241
15. Impact of mobility on traffic distribution in seamless interworking environments. In: Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th. Volume 6. (2004)
16. Modelling Cellular/Wireless LAN Integrated Systems with Multi-Rate Traffic Using Queueing Network. In: Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on. (October 2008)
17. Salkintzis, A.K.: Interworking techniques and architectures for wlan/3g integration toward 4g mobile data networks. Wireless Communications, IEEE [see also IEEE Personal Communications] **11**(3) (June 2004) 50–61
18. Reiser, M.: A queueing network analysis of computer communication networks with window flow control. **27**(8) (January 2003) 1199–1209
19. Urgaonkar, B., Pacifici, G., Shenoy, P., Spreitzer, M., Tantawi, A.: Analytic modeling of multitier internet applications. ACM Trans. Web **1**(1) (May 2007) 2+
20. Chen, Y., Das, A., Qin, W., Sivasubramaniam, A., Wang, Q., Gautam, N.: Managing server energy and operational costs in hosting centers. In: SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, New York, NY, USA, ACM (2005) 303–314
21. Opnet: Opnet official website. [Online] http://www.opnet.com
22. Kleinrock, L.: Queueing Systems, Volume II: Computer Applications. Wiley Interscience, New York (1976)