Modelling Resource Management for Multi-Class Traffic in Mobile Cellular Networks

Helmut Hlavacs, Günter Haring
Institute for Computer Science and Business Informatics,
University of Vienna, Lenaugasse 2/8,
Vienna, Austria
{ hlavacs | haring }@ani.univie.ac.at

Abhinav Kamra, Manish Bansal
Department of Computer Science and Engineering,
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-16, India.
{ abhinav | manish }@cse.iitd.ernet.in

Abstract

Future cellular mobile networks will be limited by the number of channels available in each cell. On the other hand, new broadband applications like video telephony will demand tight quality of service guarantees that must be met by the network at all times. Thus, advanced mechanisms for allocating these channels to incoming calls with different quality of service level will be of utmost importance. In this paper we introduce a new analytical model for cell channel allocation to multi-class traffic. Being based on Markov chains, the new model exploits the multi-class property and reduces the state space dramatically, thus enabling the solution of previously unsolvable problem classes. We additionally describe CECALL, a simulator implementing several different strategies for allocating cell channels to multi-class traffic, handoff pre-reservation and degradation of low-level call classes. The results of the analytical model are used for explaining important simulation results.

1. Introduction

Future cellular networks like UMTS will be able to transport both multimedia and data traffic at high bandwidths, enabling new applications like video telephony and streaming TV or video on demand from the Internet. However, the bandwidth of such a cellular network is limited by the available radio resources in each cell. These resources are typically defined by a number N of channels that may be used for transportation of digital data between the user and the cell base station. Channel multiplexing is performed by TDMA, FDMA or CDMA. The more channels are used by a single call, the more data can be transferred per second. Different applications will depend on different quality of service (QoS) requirements, resulting in multi-class traffic. Network users will then define their required QoS level by establishing contracts with the network. Violating these contracts may lead to dissatisfied users and potential loss of revenue. Hence it is crucial for network managers to provide sufficient resources and sophisticated resource allocation strategies in order to guarantee the agreed QoS contracts.

Before actually implementing such a network, it is advisable to model the relevant infrastructure architectures and perform analytical or simulation based evaluation. Analytical models often are based on steady-state analysis of Markov chains, queuing networks [5] or Petri nets [1], the two latter ones usually being again mapped on Markov chains. The drawbacks of these approaches often include unrealistic assumptions on the observed traffic and the often observed explosion of the Markov chain state space.

There are several ways for reducing the state space of large Markov chains, for example by reducing the model complexity, analyzing the structure of the Markov chain generator matrix [7] or aggregating closely related states into one macro-state [5, 8]. Depending on whether the Markov chain is exactly, ordinarily, or nearly lumpable, these aggregation techniques will yield exact solutions (for the whole chain or only the aggregated chain) or only approximations [9, 6].

In this paper an exact model for the channel allocation of multi-class traffic in cellular networks is described. By exploiting the multi-class domination property, the size of the state space can be reduced drastically, thus enabling the solution of previously unsolvable problems. We also describe the CEll Channel ALLocation simulator CECALL [11] for evaluating different network scenarios.

2. Related Work

Creating analytical models for channel allocation with QoS guarantees in cellular networks has been studied by

several authors [12, 15, 14, 10]. Here, single cells and groups of cells together with their interaction have been studied, both at the channel and packet level. In [4], models for multi-class traffic are described and solved with a numerical package for Petri nets. Generally, Markov chains modeling multi-class traffic will suffer from a large state space due to the need for using multi-dimensional Markov chains [2]. Thus, an alternative to analytical models is given by simulation [13, 11]. The drawbacks of simulation, however, are given by the long and error prone software development, the possibly long simulation runs and the difficult interpretation of the results [3].

3. Modelling Cellular Networks

The used model describes the call admission of a single cell A being part of a larger cellular radio network [11]. Calls originating in this cell may be of type *guaranteed* or *best effort*. Guaranteed calls represent normal voice or video calls enjoying high priority and QoS guarantees for their assigned QoS level. Best effort calls denote low priority data connections and are tolerant with respect to loosing once allocated channels down to a minimum number of required channels. Additionally, *handoff* calls denote guaranteed calls entering cell A from a neighboring cell. The system structure is shown in Figure 1.

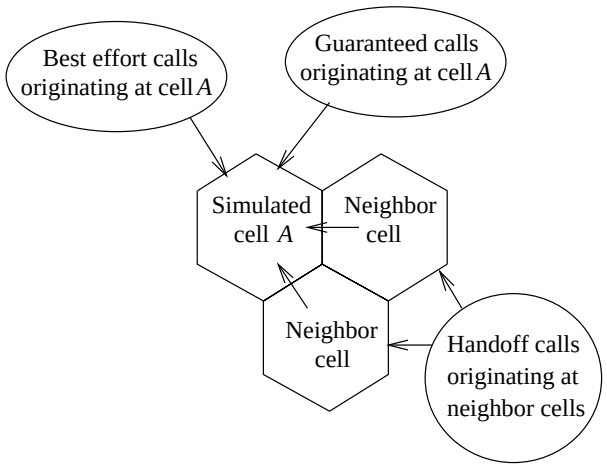

Figure 1. Model Structure.

The call arrival rate $\lambda = \lambda_b + \lambda_g + \lambda_h$ is the sum of the rates of the above mentioned call classes. While best effort (rate λ_b) and guaranteed (rate λ_g) calls enter the simulated cell directly (as this is their point of origin), handoff calls (rate λ_h) first enter a neighboring cell and immediately signal their presence to cell A. After a so-called *activation time a* has passed, handoff calls subsequently enter cell A.

The reason for making a distinction between these call types stems from the fact that from the viewpoint of the cell, calls may either originate in it or enter from outside. As best effort calls are not able to pre-reserve channels, a distinction between the two origins is not necessary. On the other hand, handoff calls may pre-reserve channels during their activation time, thus limiting the available channels for other guaranteed calls. Therefore, a distinction must be made between guaranteed calls originating in cell A or handoff calls entering from outside. However, there is no distinction between call termination and changing the cell, in each case, the respective call will give back its allocated channels.

Each call entering cell A must be assigned a minimum number of channels from the N channels managed by A. Best effort calls additionally request more channels up to a certain maximum number, but can operate with any number of channels within the predefined range. If there are enough free channels, they are assigned to the newly arrived call and the call may proceed. After the call's *holding time* $h = 1/\mu$ has passed, the call terminates or leaves the cell and allocated channels are returned to the call admission control where they may be reassigned to other calls.

On the other hand, if there are not enough channels left for a new call, a call originating at cell A is said to be *blocked* and terminates immediately, and a handoff call being transferred from a neighbor cell is said to be *dropped*. In real life, a dropped handoff call experiences a sudden loss of connection when moving from one cell to another and is considered to be most inconvenient to the network user.

Several different strategies for channel assignment and best effort degradation have been implemented in the simulation tool CECALL (see Section 5). The basic strategies include:

- Complete Sharing (CS): No partitioning, no prioritization, no pre-reservation. All calls have equal access to the channels.
- Complete Partitioning (CP): The available N channels are partitioned such that for each of the three call classes, $N/3$ channels are available.
- Fractional Guard Channels (FG): A subset of size H of the channels is pre-reserved for handoff calls only. However, if all remaining $N - H$ channels are allocated and a new guaranteed call arrives, it is admitted with probability $0 \le p \le 1$. If $p = 0$ then this is equivalent to the strategy of guard channels [10].
- Dynamic Resource Partitioning (DRP): Arriving guaranteed and handoff calls may take away channels from best effort calls, *degrading* them, if their new number of channels does not fall below the required minimum (degradation phase), or *interrupting* them, if no such best effort call is found (interrupt phase).

Additionally, handoff calls currently being in a neighbor cell may passively reserve channels in cell *A*. These channels then can not be used for arriving guaranteed calls. However, they can be assigned to best effort calls for temporary use.

The remainder of this paper will be concerned with strategy DRP only. In this strategy two different call classes exist (guaranteed/handoff and best effort), where the guaranteed/handoff class dominates the best effort class in the sense that arriving guaranteed/handoff calls may take away channels from active best effort calls.

The model is based on three assumptions. Firstly, each handoff call being generated in a neighbor cell will eventually enter cell *A*. Secondly, a best effort call can only be degraded or interrupted, but can not be reassigned channels again. Thirdly, the call holding time does not change if a best effort call is degraded.

Typical performance measures important for network planners include:

1. Probability that guaranteed or best effort calls are blocked.
2. Probability that handoff calls are dropped.
3. Probability that best effort calls are degraded (and how often they are degraded).
4. Throughput for all call classes.

Generally, these measures will depend on the input parameters λ, a, h and N in a non-linear manner.

4. Analytical Evaluation

In this section, an analytical approach for evaluating the dynamics of the above described cell model are explained. However, the analytical model solves only a submodel of it. The activation time *a* is set to zero so that there can be no pre-reservation for handoff calls and both guaranteed and handoff calls are treated alike. Also, best effort calls can not allocate more channels than their minimum. This means that all call types allocate a certain number of channels at call start and keep them until call termination, implying that all best effort call degradations result in interruption of the respective calls.

4.1. General Solution

Solutions for the described model can be found by creating a continuous time Markov chain (CTMC) with state space

$$S_{t,u}^{N} = \left\{ (g_1,...,g_t,b_1,...,b_u) | \sum_{i=1}^{t} i g_i + \sum_{j=1}^{u} j b_j \le N \right\} \quad (1)$$

with $g_i \ge 0, b_j \ge 0$, where g_i denotes the number of guaranteed calls of type *i* currently in the cell, each using *i* channels, and b_j denotes the number of best effort calls of type *j* currently in the cell, each using *j* channels. The rate of transition $T_{s_1 \to s_2}$ of moving from an arbitrary state s_1 to another state s_2 is derived from the individual arrival and departure rates $\lambda_{g_i}, \lambda_{b_j}, \mu_{g_i}, \mu_{b_j}$, and the g_i and b_j of s_1.

Provided the Markov chain is in state $s_1 = (g_1,...,g_i,...,g_t,b_1,...,b_u)$ and a guaranteed call of type *i* arrives, the chain moves to state $s_2 = (g_1,...,g_i+1,...,g_t,b_1,...,b_u)$, if *i* channels are available. Otherwise, some of the best effort calls are interrupted (loosing all their allocated channels) and the Markov chain moves to state $s_3 = (g_1,...,g_i+1,...,g_t,b_1-k_1,b_2-k_2,...,b_u-k_u)$ according to the following strategy:

1. All currently free channels are assigned to the arriving call.
2. A minimum number of best effort calls will be interrupted.
3. In case, there are several choices for interrupting a minimum number of calls, the calls to be interrupted are chosen such that a minimum number of channels are additionally freed by interrupting them. For example, if three best effort call types exist $(u = 3)$ and one channel must be freed, and the system is in state $b_1 = 0, b_2 > 0$ and $b3 > 0$, then one call of type 2 is chosen for interruption, because this would additionally free one channel, whereas interrupting a type 3 call would additionally free 2 channels.

In case *i* channels can not be obtained even by interrupting all active best effort calls, the incoming guaranteed call is blocked and the state remains unchanged. For example, if $u = 2$ and a newly arriving guaranteed call requires five channels, but only three are free, then the following table shows which best effort calls are interrupted:

(b_1,b_2)	Interrupted best effort calls
(0,0) or (1,0)	None (arriving g call blocked)
$(b_1,0)_{b_1>1}$	Two best effort calls of type 1
$(b_1,b_2)_{b_2>0}$	One best effort call of type 2

The Markov chain would then be transformed into a system of linear equations (usually represented by a highly sparse matrix) reflecting the steady-state [5], with one unknown and one equation for each element of the state space. The solution of these equations then denotes the probability of being in any state of $S_{t,u}^{N}$. When denoting the size of $S_{t,u}^{N}$ by $L_{t,u}^{N}$ and using ordinary

solution techniques for the system of linear equations, then the size of the matrix (and thus the size of the memory necessary for storing this matrix) will be $O(L_{t,u}^N)^2$ and the number of necessary operations in order to obtain the result will be of order $O(L_{t,u}^N)^3$. However, by using appropriate sparse representations and special-purpose sparse solvers, this can be reduced significantly. Once, the solution of these equations is known, additional measures like the probability of guaranteed call blocking can be easily derived from them [4, 11].

Unfortunately, at high dimensions, the size of the state space will explode. By using fundamental theorems from discrete mathematics, the size of the state space can be computed using generating functions. A generating function $F(x)$ is equivalent to a series $A = \{a_n\}_{n>0}$ if $F(x) = \sum_{n>0} a_n x^n$ and the convergence radius of A is larger than 0. Due to the theorem of Taylor, given $F(x)$, the coefficients a_n of the series expansion of $F(x)$ are defined to be

$$a_n = \frac{F^{(n)}(x)}{n!}|_{x=0} \qquad (2)$$

where $F^{(n)}(x)$ denotes the n^{th} derivative of $F(x)$. When defining $S_{t,u}(x)$ to denote the generating function for the state space size of $S_{t,u}^N$ the following is easily derived:

$$S_{t,u}(x) = \frac{1}{1-x}\left(\prod_{i=1}^{t}(1-x)^i\right)^{-1}\left(\prod_{i=1}^{u}(1-x)^i\right)^{-1} \qquad (3)$$

Using (2) on (3), a_n will yield the size of $S_{t,u}^n$.

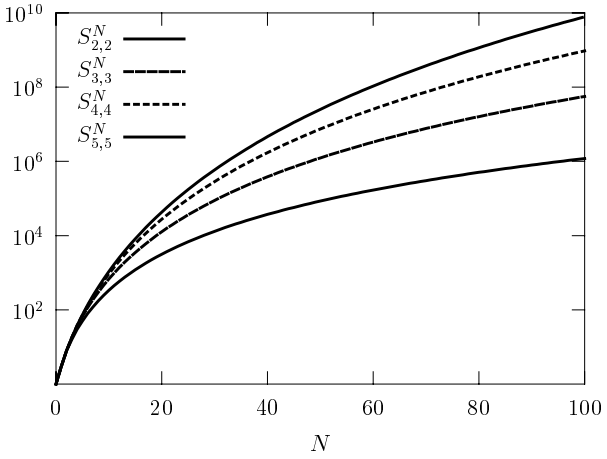

Figure 2. Size of the state space $S_{t,u}^N$.

Figures 2 and 3 show, how the size of the state space depends on t,u and N. Here S_t^N denotes a state space containing only guaranteed calls. It can be seen that even for moderate sizes of t,u and N, the state space for $S_{t,u}^N$ is far too large to be solved by any modern computer. Unfortunately, the Markov chain $S_{t,u}^N$ does not satisfy the exact lumpability condition [6] over any reasonable partition, and thus, traditional aggregation schemes can not be applied for reducing the state space.

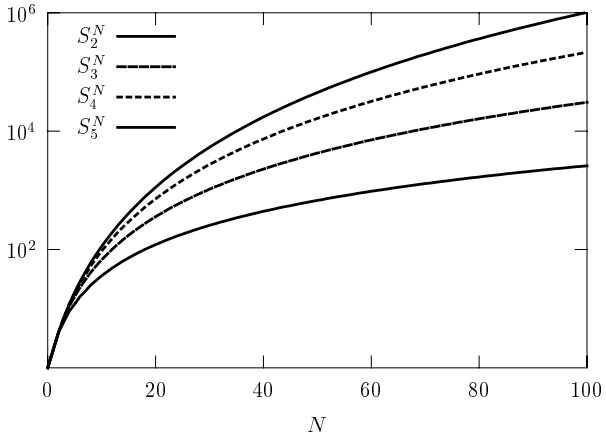

Figure 3. Size of the state space S_t^N.

4.2. Solution by Exact Aggregation

Figure 3 shows that the state space for S_t^N is of several orders of magnitude smaller than the one for $S_{t,u}^N$ (see Figure 2). Also it must be noted that there exists an analytical solution for the Markov chain representing S_t^N [2]:

$$P_{g_1,\ldots,g_t} = P_{0,\ldots,0}\prod_{k=1}^{t}\frac{a_k^{g_k}}{g_k!} \text{ and } \sum_{(g_1,\ldots,g_t)\subset S_t^N} P_{g_1,\ldots,g_t} = 1 \qquad (4)$$

where the sum goes over all possible states and the constants $a_k = \lambda_{g_k} / \mu_{g_k}$ define the load (in Erlang) of call type *k* imposed on the system. Further examination of the model shows that guaranteed calls have priority over best effort calls in the sense that they may take away channels from best effort calls if needed and available. Thus, from the viewpoint of guaranteed calls, best effort calls are invisible and have no influence on their behavior. Therefore, the probabilities of states of guaranteed calls can be computed independently of the best effort calls. The transitions of guaranteed calls occur as if all the channels are available to them. Thus, in order to determine the behavior of guaranteed calls, it is sufficient to solve the S_t^N chain using (4) instead of $S_{t,u}^N$.

In order to compute the corresponding best effort call state probabilities, the state space S_t^N is then collapsed

into an equivalent chain $G^N = \{(g) \mid g = 0,1,...,N\}$. Here, (g) represents all states of S_t^N where g channels are occupied. Using (4), the probability for being in state (g) is then given by

$$P_g = \sum_{g_1,...,g_t} P_{g_1,...,g_t}, \text{ where } \sum_{i=1}^{t} i g_i = g. \qquad (5)$$

Hence the exact solution of G^N is determined by (5). The corresponding transition rates from an arbitrary state (g) to possible successor states can also be derived as

$$\left.\begin{aligned} T_{g\to(g+i)} &= \lambda_{g_i} \\ T_{g\to(g-i)} &= \frac{1}{P_g} \sum_{g_1,...,g_t} \mu_{g_i} g_i P_{g_1,...,g_t} \mid \sum_{i=1}^{t} i g_i = g \end{aligned}\right\} \qquad (6)$$

where $i = 1,2,...,t$ and $(\mu_{g_i} g_i)$ denotes the departure rate of calls of type i when there are g_i such calls. The chain G^N is then used for constructing a larger chain

$$GS_u^N = \left\{ (g, b_1, b_2, ..., b_u) \mid g + \sum_{i=1}^{u} i b_i \le N \right\} \qquad (7)$$

with transition rates for transitions from state $(g, b_1, ..., b_u)$ to possible successor states described in Table 1. In this table, 'b' denotes best effort calls, 'g' denotes guaranteed calls, 'a' denotes a call arrival, and 'd' denotes a call departure. The transitions can only occur if the successor state belongs to GS_u^N. The transitions in the last row occur with best effort call interruption. In this case, the $k_i's$ are determined according to the interruption strategy described in Section 4.1. It must be noted that from the viewpoint of best effort calls, GS_u^N is equivalent to the original Markov chain $S_{t,u}^N$ since all state information of the guaranteed calls that can affect best effort calls is reflected by g, the number of channels occupied by guaranteed calls.

By solving the steady-state equations of the Markov chain GS_u^N, the probabilities $P_{g,b_1,...,b_u}$ for being in state $(g, b_1, ..., b_u)$ can be computed. It is worth noting that the state space of GS_u^N is much lower than the one of $S_{t,u}^N$ (see Figures 4 and 5).

The above-obtained results then can be used for computing the stationary probabilities of the original Markov chain $S_{t,u}^N$ exactly.

$$P_{g_1,...,g_t,b_1,...,b_u} = P_{g,b_1,...,b_u} \frac{P_{g_1,...,g_t}}{P_g}, \qquad (8)$$

where $g = \sum_{i=1}^{t} i g_i$ and P_g is derived from (5).

Table 1. Transition rates of GS_u^N.

Class	Type	a/d	Successor	Rate
b	1	a	$(g, b_1+1, ..., b_u)$	λ_{b_1}
b	2	a	$(g, b_1, b_2+1, ..., b_u)$	λ_{b_2}
…	…	…	…	…
b	u	a	$(g, b_1, b_2, ..., b_u+1)$	λ_{b_u}
b	1	d	$(g, b_1-1, ..., b_u)$	$\mu_{b_1} b_1$
b	2	d	$(g, b_1, b_2-1, ..., b_u)$	$\mu_{b_2} b_2$
…	…	…	…	…
b	u	d	$(g, b_1, b_2, ..., b_u-1)$	$\mu_{b_u} b_u$
g	1	a	$(g+1, b_1, ..., b_u)$	$T_{(g)\to(g+1)}$
g	2	a	$(g+2, b_1, ..., b_u)$	$T_{(g)\to(g+2)}$
…	…	…	…	…
g	t	a	$(g+t, b_1, ..., b_u)$	$T_{(g)\to(g+t)}$
g	1	d	$(g-1, b_1, ..., b_u)$	$T_{(g)\to(g-1)}$
g	2	d	$(g-2, b_1, ..., b_u)$	$T_{(g)\to(g-2)}$
…	…	…	…	…
g	t	d	$(g-t, b_1, ..., b_u)$	$T_{(g)\to(g-t)}$
g	i	a	$(g+i, b_1-k_1, ..., b_u-k_u)$	$T_{(g)\to(g+i)}$

4.3. Extension for Arbitrary Number of Traffic Classes

The above described procedure can even be applied for an arbitrary number of call classes $g, b, c, d, ...$, such that g calls dominate b calls, which in turn dominate c calls, and so on. This means that arriving g calls can take away channels from $..., d, c, b$ calls in that order, b calls can take away channels from $..., d, c$ calls in that order and so on. In this case, the state space $S_{t,u,v,w,...}^N$ to be solved is given by

$$S_{t,u,v,w,...}^N = \{(g_1,...,g_t, b_1,...,b_u, c_1,...,c_v, d_1,...,d_w,...)\}$$

$$\text{where } \sum_{i=1}^{t} i g_i + \sum_{i=1}^{u} i b_i + \sum_{i=1}^{v} i c_i + \sum_{i=1}^{w} i d_i + ... \le N$$

$$\text{and } g_i, b_i, c_i, d_i, ... \ge 0.$$

It is clear that even for a small number of traffic classes this state space can not be solved directly. However, by exploiting the property of successive domination of call classes and repeatedly using the above described aggregation method, the following sequence of Markov chains is created and solved:

$$S_t^N = \left\{(g_1,...,g_t) \mid \sum_{i=1}^{t} ig_i \le N\right\} \rightarrow$$

$$GS_u^N = \left\{(g,b_1,...,b_u) \mid g + \sum_{i=1}^{u} ib_i \le N\right\} \rightarrow$$

$$GBS_v^N = \left\{(gb,c_1,...,c_v) \mid gb + \sum_{i=1}^{v} ic_i \le N\right\} \rightarrow$$

$$GBCS_w^N = \left\{(gbc,d_1,...,d_w) \mid gbc + \sum_{i=1}^{w} id_i \le N\right\} \rightarrow$$

...

Similar as in (8), the sequence of solutions can then be combined in reverse order to compute the solution for the original Markov chain $S_{t,u,v,w,...}^N$, and arbitrary partial intermediate solutions if needed. It follows that instead of solving the Markov chain $S_{t,u,v,w,...}^N$ of $t+u+v+w+...$ variables, the largest Markov chain, which needs to be solved by the above aggregation method, depends only on

$$\max\{t,1+\max\{u,v,w,...\}\} \qquad (9)$$

variables. Figures 4 and 5 show the ratio of the size of the original state spaces to the size of the maximum state spaces to be solved in the above aggregation method (9).

5. The Simulator CECALL

As already described in Section 3, in the current version of CECALL, several different strategies for managing the call access control have been implemented. The remainder of this section will focus on the strategy DRP and its different versions.

In DRP, newly arriving guaranteed and handoff calls can take away channels from already running best effort calls. If due to such a channel loss the number of allocated channels of a best effort call drops below its channel minimum, the call is said to be interrupted and terminates, thus deallocating all its still allocated channels. For finding the next channel to be taken away from a best effort call, all DRP versions implement a two-phase approach. In the first phase, channels are taken away only from those best effort calls that can spare channels without being interrupted. If no such calls exist, DRP enters the second phase, where best effort calls are chosen to be interrupted.

Additionally, if a handoff call is created, i.e., it arrives at a neighbor cell, it tries to pre-reserve channels in cell *A*. This procedure, however, is carried out only if enough channels exist that can be reserved. Otherwise, the handoff call will not reserve any channels and will enter cell *A* like a normal guaranteed call. If at this point in time there are still not enough either free or unreserved best effort channels, the handoff call terminates and is counted as being dropped. The channel reservation again follows a two-phase scheme by first reserving only unused and unreserved channels, and, if no such channels are available, by secondly choosing channels currently being used by best effort calls, which have not been pre-reserved by other handoff calls so far.

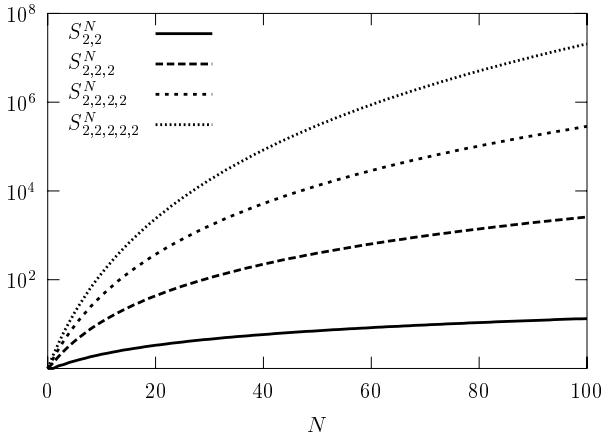

Figure 4. State space saving factor when using a state space with maximum dimension (9) instead of taking $S_{t,u,v,w,...}^N$. For each call class, calls may use either one or two channels.

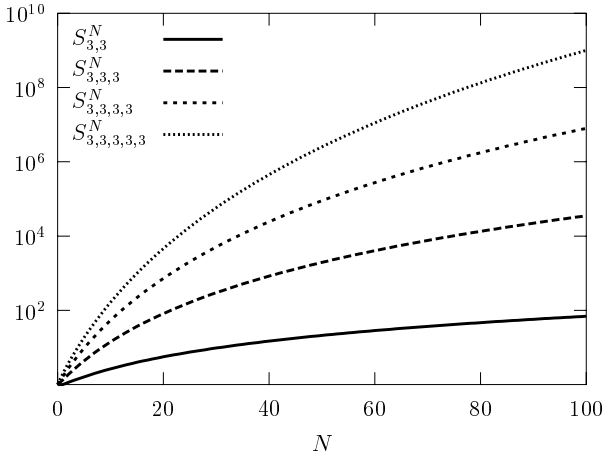

Figure 5. State space saving factor when using a state space with maximum dimension (9) instead of taking $S_{t,u,v,w,...}^N$. For each call class, calls may use either one, two or three channels.

Each channel thus may be marked as reserved and newly arriving guaranteed and handoff calls may not allocate or reserve such a marked channel. However, if a

best effort call does not find enough free channels it may temporarily use reserved channels, risking to be interrupted as soon as the reserving handoff call arrives at cell A.

The investigated versions of DRP differ in how best effort calls are chosen next to be either degraded or interrupted. Also, different versions for taking away only one or all channels that can be spared (with names starting with "DRPA") exist. In the standard DRP version, all best effort calls are put into a linear list and newly arriving calls are added to the list end. At all times, a list pointer called *cursor* points at the next best effort call to be degraded or interrupted. In the degradation phase, DRP starts at the cursor and searches for calls that can spare channels without being interrupted. If a call is found, it is degraded by one channel and the cursor is set to the call's successor or the list start in case the list end was reached. If the list is run through once without finding a call being able to spare a channel, DRP first interrupts the call being pointed at by the cursor, then its successors. Note that this strategy slightly differs from the interruption strategy used in the analytical solution presented in Section 4.1. The reason for this is that additional information like the order in a list of calls can not be used in the analytical approach.

In DRP_LT, the best effort calls are ordered according to their lifetime. When DRP_LT enters the degradation phase, it will first degrade the youngest call, then the second youngest and so on. Likewise, in the interrupt phase, first the youngest call will be interrupted, then the second youngest, and so on. This strategy is based on the assumption that interrupting a long-lasting call leads to higher customer dissatisfaction than interrupting a young one. Also, it is reasonable to assume that older calls are more likely to reach their call end sooner than younger calls.

Finally, for the degradation phase, strategy DRP_RL orders the best effort calls according to the relative number of channels they can spare. If a_b denotes the number of currently allocated and m_b denotes the minimum number of channels of best effort call b, then the calls are ordered according to $(a_b - m_b)/m_b$ descending.

Table 2. Degradation strategies.

Strategy	Order Phase 1	Taken away in Phase 1	Order Phase 2
DRP	List	1	List
DRPA	List	All Spare	List
DRP_LT	Lifetime	1	Lifetime
DRPA_LT	Lifetime	All Spare	Lifetime
DRP_RL	Spare	1	Minimum
DRPA_RL	Spare	All Spare	Minimum

For the interrupt phase, this time the call with the maximum number of still allocated channels (=min) is chosen to be interrupted next. This is done in the hope that fewer calls will be interrupted, if first the ones with more still allocated channels are terminated. Table 2 shows the investigated degradation strategies.

6. Numerical Results

6.1. Example

In order to illustrate the aggregation method described in Section 4.2, its application is demonstrated on a simple example. Let the state space be

$$S_{2,2}^2 = \{(g_1, g_2, b_1, b_2) \mid g_1 + 2g_2 + b_1 + 2b_2 \le 2\}.$$

The size of this state space is eight, the individual states being (0,0,0,0), (0,0,0,1), (0,0,1,0), (0,0,2,0), (0,1,0,0), (1,0,0,0), (1,0,1,0), and (2,0,0,0). Furthermore, the call arrival and departure rates are set as

$$\lambda_{g_1} = \lambda_{g_2} = \lambda_{b_1} = \lambda_{b_2} = \frac{1}{2}\mu_{g_1} = \frac{1}{2}\mu_{g_2} = \frac{1}{2}\mu_{b_1} = \frac{1}{2}\mu_{b_2}.$$

Step 1: Solve S_2^2

By following our approach, first a new Markov chain of four states $S_2^2 = \{(g_1, g_2) \mid g_1 + 2g_2 \le 2, g_1, g_2 \ge 0\}$ is created. Using (4), the appropriate steady-state probabilities are computed to be $P_{0,0} = 8/17, P_{1,0} = 4/17, P_{0,1} = 4/17$ and $P_{2,0} = 1/17$.

Step 2: Aggregate into G^2

In the next step, the Markov chain S_2^2 is collapsed into $G^2 = \{(g) \mid g \le 2, g \ge 0\}$ and using (5), its steady-state probabilities are computed to be $P_0 = 8/17, P_1 = 4/17$ and $P_2 = 5/17$. Also, according to (6) the transition rates are calculated to be

$$T_{0\to1} = \lambda_{g_1} \qquad T_{1\to0} = \mu_{g_1}\frac{P_{1,0}}{P_1} = \mu_{g_1}$$

$$T_{1\to2} = \lambda_{g_1} \qquad T_{2\to1} = 2\mu_{g_1}\frac{P_{2,0}}{P_2} = \frac{2}{5}\mu_{g_1}$$

$$T_{0\to2} = \lambda_{g_2} \qquad T_{2\to0} = \mu_{g_2}\frac{P_{0,1}}{P_2} = \frac{4}{5}\mu_{g_2}$$

Step 3: Solve GS_2^2

Now the Markov chain $GS_2^2 = \{(g, b_1, b_2) \mid g + b_1 + 2b_2 \le 2\}$ defined in Table 3, with transition rates which are computed according to Table 1, is created and solved.

Table 3. Definition of GS_2^2.

From	To	Rate	From	To	Rate
(0,0,0)	(0,1,0)	λ_{b_1}	(0,2,0)	(0,1,0)	$2\mu_{b_1}$
(0,0,0)	(0,0,1)	λ_{b_2}	(0,2,0)	(1,1,0)	$*\,T_{(0)\to(1)}$
(0,0,0)	(1,0,0)	$T_{(0)\to(1)}$	(0,2,0)	(2,0,0)	$*\,T_{(0)\to(2)}$
(0,0,0)	(2,0,0)	$T_{(0)\to(2)}$	(1,0,0)	(0,0,0)	$T_{(0)\to(1)}$
(0,1,0)	(0,0,0)	μ_{b_1}	(1,0,0)	(1,1,0)	λ_{b_1}
(0,1,0)	(0,2,0)	λ_{b_1}	(1,0,0)	(2,0,0)	$T_{(1)\to(2)}$
(0,1,0)	(1,1,0)	$T_{(0)\to(1)}$	(1,1,0)	(0,1,0)	$T_{(1)\to(0)}$
(0,1,0)	(2,0,0)	$T_{(0)\to(2)}$	(1,1,0)	(1,0,0)	μ_{b_1}
(0,0,1)	(0,0,0)	μ_{b_2}	(1,1,0)	(2,0,0)	$*\,T_{(1)\to(2)}$
(0,0,1)	(1,0,0)	$*\,T_{(0)\to(1)}$	(2,0,0)	(0,0,0)	$T_{(2)\to(0)}$
(0,0,1)	(2,0,0)	$*\,T_{(0)\to(2)}$	(2,0,0)	(1,0,0)	$T_{(2)\to(1)}$

The transitions marked with a '*' involve interruption of one or more best effort calls. The computed steady-state probabilities are

$$P_{0,0,0} = 2160/7463 \qquad P_{1,0,0} = 1328/7463$$
$$P_{0,0,1} = 540/7463 \qquad P_{1,1,0} = 428/7463$$
$$P_{0,1,0} = 696/7463 \qquad P_{0,0,0} = 2160/7463$$
$$P_{0,2,0} = 116/7463$$

Step 4: Obtain $S_{2,2}^2$

Using (8), the steady-state probabilities of the original Markov chain $S_{2,2}^2$ are computed to be

$$P_{0,0,0,0} = 2160/7463 \qquad P_{1,0,0,0} = 1328/7463$$
$$P_{0,0,0,1} = 540/7463 \qquad P_{1,0,1,0} = 428/7463$$
$$P_{0,1,0} = 696/7463 \qquad P_{0,1,0,0} = 4/17$$
$$P_{0,0,2,0} = 116/7463 \qquad P_{2,0,0,0} = 1/17$$

It can easily be verified that this solution is the same that would be obtained by solving the original Markov chain $S_{2,2}^2$ directly.

6.2. Case Study

In this Section results from using the analytical approach are compared to results obtained by simulation using the simulator CECALL. The effect of varying the traffic load and available channels is considered. For obtaining the simulation results, for each parameter set, ten simulation run replications were carried out in order to additionally compute 95% confidence intervals for all results. Each replication ran over at least 200,000 virtual seconds. The lengths of the 95% confidence intervals were at the order of magnitude of 1-5% of the obtained means.

Figures 6 to 9 compare the analytical results with the simulation results. Here, the arrival rate λ is divided equally between the guaranteed/handoff and best effort calls. Within each call class, two call types exist, one using one channel, the other one using two. Thus, $\lambda/4 = \lambda_{g_1} = \lambda_{g_2} = \lambda_{b_1} = \lambda_{b_2}$. Also, the holding time h is the same for all call types.

Generally, it can be seen that the results concerning only guaranteed/handoff calls are exactly the same for the analytical approach as well as for simulation. Results, where best effort calls are also reflected, show a small deviation due to the fact that the degradation strategies vary slightly between the analytical model and strategy DRP used by CECALL.

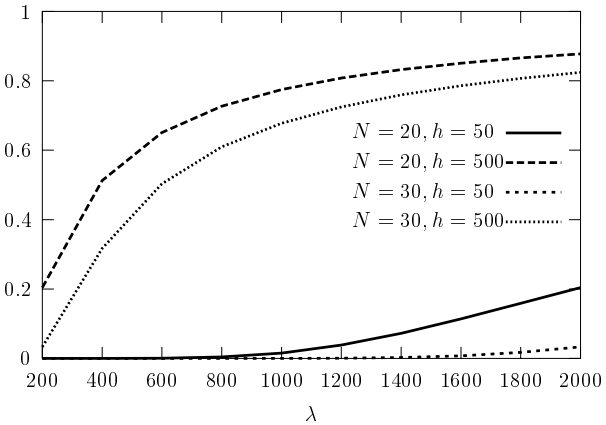

Figure 6. Probability for a guaranteed/handoff call being blocked. Analytical results are represented by a fat line, simulation results by a thin line. In this case, no difference between them can be observed.

The probability for guaranteed/handoff call blocking is shown in Figure 6. It can be observed that as λ increases, more and more guaranteed/handoff calls will be blocked. Also, the effect of having more channels can be seen. As the load increases, the advantage of having more channels is nullified. The curves for $h = 50$ will show the same behavior for higher values of λ (not shown here).

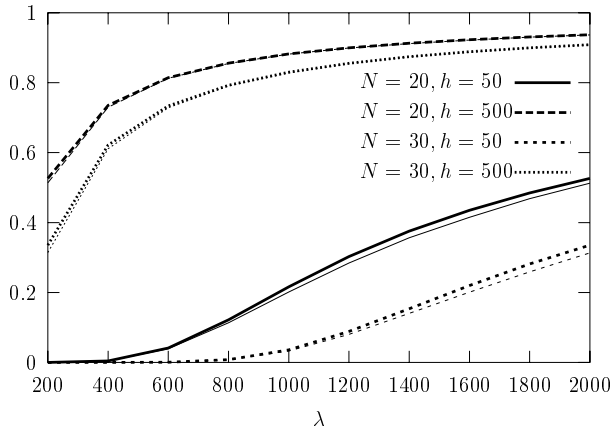

Figure 7. Probability for a best effort call being blocked. Analytical results are represented by a fat line, simulation results by a thin line.

Figure 7 shows the probabilities for best effort calls being blocked. The observed behavior is similar to that in Figure 6, however in this case the blocking probabilities are much higher because guaranteed/handoff calls enjoy priority over best effort calls.

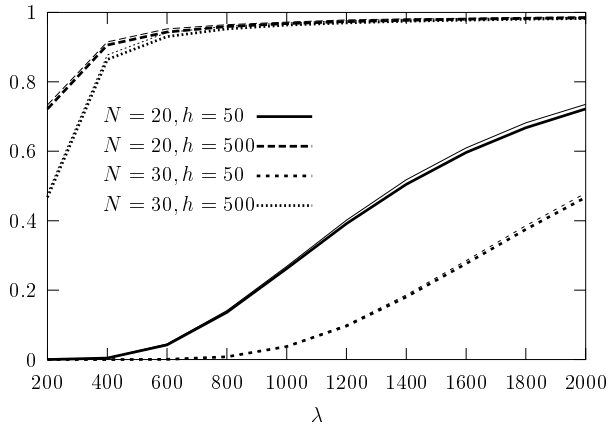

Figure 8. Probability for a best effort call being interrupted. Analytical results are represented by a fat line, simulation results by a thin line.

Figure 8 shows the probabilities for best effort calls being interrupted. This is the ratio of interrupted best effort calls to the total number of admitted (not blocked) best effort calls. At high load, the advantage of having more channels is quickly nullified. At lower load, having more channels offers a significant advantage for best effort calls.

Figure 9 shows the probability that all channels are used (P[N]) and that all channels are used by guaranteed/handoff calls only (P[N,0]). If all call types used one channel each, the difference of these curves, denoting the fact that all channels are used, but some are used by best effort calls, would represent the probability for best effort calls being interrupted. It can be seen that, contradicting the expected behavior that this probability should increase monotonically with the load, the reverse is the case and for higher load, it drops monotonically. This can be explained by the fact that at high loads, less best effort calls will be admitted and even lesser will be interrupted.

This observation is important to understand the following simulation results obtained for a different scenario. Here, $N = 250, h = 400$ and the handoff activation time is 50 [11]. Furthermore, guaranteed/ handoff calls may use up to three and best effort calls up to eight channels. Figure 10 shows how the mean number of degradations occurring for each admitted best effort call depends on the used degradation strategy. Again, this number first rises but will continuously decrease for higher loads.

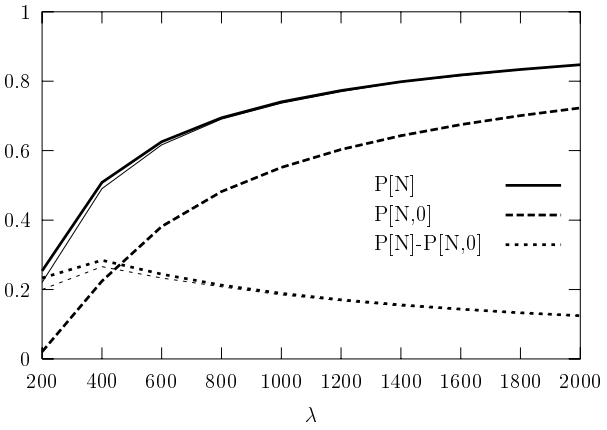

Figure 9. Probability of all channels being occupied (P[N]), probability that all channels are occupied by guaranteed/handoff calls only (P[N,0]).

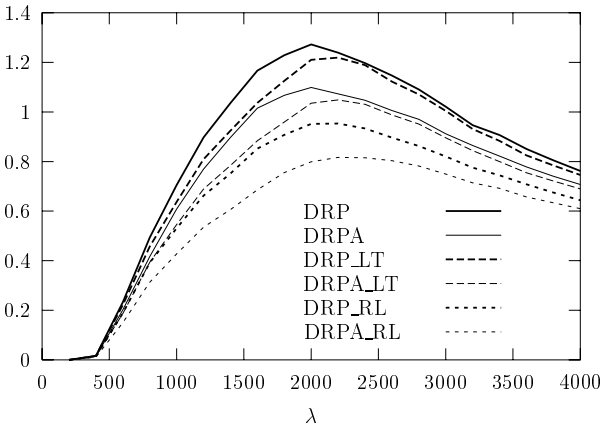

Figure 10. Number of degradations of best effort calls per admitted best effort call, caused by arriving guaranteed/handoff calls for different degradation strategies.

7. Conclusion

In this paper a new exact aggregation strategy for multi-dimensional Markov chains representing the channel allocation in cellular networks for multi-class traffic has been presented. Although the Markov chains representing the original state space are not exactly (but only ordinarily) lumpable, rendering the use of traditional aggregation schemes approximate only, this strategy will obtain exact results by a different aggregation approach.

The approach exploits the successive domination property of multi-class traffic in cellular networks, as introduced in [11]. Using this property, the analysis of many traffic classes can be reduced to the successive analysis of individual classes. This way the state spaces of the analyzed Markov chains (and therefore the sizes of the systems of linear equations to be solved) are drastically reduced. Also, various types of intermediate results can be derived with even lesser effort. This procedure thus enables the analytical evaluation of large sized multi-class cellular networks which were previously analytically intractable.

The presented results have been verified by simulation with the simulation tool CECALL. By using an analytical model, the interpretation of obscure simulation results became possible and hence similar techniques can be applied in other cases.

8. References

[1] M. Ajmone Marsan, G. Balbone, G. Conte, S. Donatelli, and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets.* John Wiley & Sons, New York, 1995.

[2] H. Akimaru and K. Kawashima. *Teletraffic,* Springer, Berlin New York, 1999.

[3] J. Banks, editor. *Handbook of Simulation,* John Wiley & Sons, New York, 1998.

[4] K. Begain, G. Bolch, and M. Telek. "Scalable schemes for call admission and handover in cellular networks with multiple services". *Wireless Personal Communications* 15 (2000), pp. 125-144.

[5] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. *Queuing Networks and Markov Chains*. John Wiley & Sons, New York, 1998.

[6] P. Buchholz. *Exact and ordinary lumpability in finite markov chains*. Journal of Applied Probability 31 (1994), pp. 59-75.

[7] P. Buchholz. *Structured analysis approaches for large markov chains*. Applied Numerical Mathematics, 31-4 (1999), pp. 375-404.

[8] P.J. Courtois. *Decomposability. Queuing and Computer System Analysis*. Academic Press, New York, 1977.

[9] P.J. Courtois and P. Semal. *Bounds for the positive eigenvectors of nonnegative metrices and for their approximations by decomposition*. Journal of the ACM, 31-4 (1984), pp. 804-825.

[10] G. Haring, R. Marie, R. Puigjaner, and K. Trivedi. *Loss formulae and their application to optimization for cellular networks*. IEEE transactions on Vehicular Systems 50-3 (2001), pp. 664-673.

[11] H. Hlavacs and G. Haring. *On degrading best effort calls in future cellular mobile networks*. In International Symposium on 3rd Generation Infrastructure and Services (3GIS), 2001.

[12] C.-J. Ho and C.-T. Lea. *Improving call admission policies in wireless networks*. Wireless Networks 5 (1999), pp. 257-265.

[13] A. Mahmoodian and G. Haring. *Mobile rsvp with dynamic resource sharing*. In Proceedings of Wireless Communications and Networking Conference 2000, Chicago, September 2000.

[14] M. Naghshineh and A.S. Acampora. *Design and control of micro-cellular networks with qos provisioning for data traffic*. Wireless Networks 3 (1997), 249-256.

[15] R. Ramjee, D. Towsley, and R. Nagarajan. *On optimal call admission control in cellular networks*. Wireless Networks, 3 (1997), pp. 29-41.