Capacity of Multi-service Cellular Networks

with Transmission-Rate Control: A Queueing Analysis

Eitan Altman
Projet MISTRAL,INRIA
Sophia Antipolis, France
altman@sophia.inria.fr

MOBICOM, ATLANTA, Sept. 2002

Outline

- Introduction
- Uplink static model
- Dynamic statistical model
- Overview of results
- Extending the model
- References

1 Introduction

There exist two main approaches to define capacities in wireless networks:

•Pole capacity: the maximum of the "load" that the system can handle (e.g. maximum number of mobiles). This is a **Static Notion**.

•Erlang capacity: defined with respect to a rejection limit ϵ. It is the "arrival rate" of connections that the system can handle while ensuring outages or drop of calls with probability less than ϵ.

•We call this the Erlang(ϵ) capacity.

Introduced in telephony networks in the beginning of the last century. Computed in CDMA context in A. M. Viterbi and A. J. Viterbi. Erlang capacity of a power controlled CDMA system. *IEEE Journal of Selected Areas in Communications*, pages 892–900, 1993.

•Consider here NRT traffic (e.g. data).

•NRT traffic does not have strict throughput requirements. Therefore the "pole capacity" (in terms of number of connections) is unbounded, if we can assign sufficiently small throughputs.

•What is the impact of smaller throughputs on Erlang capacity?

•If we decrease NRT rate, we can accept more calls but the transmission times are longer (in contrast to RT applications). Impact on Erlang capacity not clear.

•How to define a statistical type capacity when we can control the transmission rate?

Possible answers:

•Define the best Erlang capacity, where we optimize over different transmission rates.
(1st part of the talk)

•Erlang capacity is defined with respect to a given ϵ. Is there a version of the capacity that does not depend on ϵ? (like Shannon capacity)?
(2nd part of the talk)

•Erlang capacity is related to REJECTION RATE.
If we can avoid rejections (by slowing transmission rates), one can define capacity related to OTHER QoS!!!
For example (ongoing work with Nidhi Hedge):

The arrival rates for which the expected delay is smaller than D

2 Uplink static Model

•Consider

- $K = \{1, ..., k\}$ classes of service
- $M(s)$= number of calls of type s, $\mathbf{M} = (M(1), ..., M(k))$.
- $R(s)$= transmission rate of class s
- The necessary received power at the BS for mobile of type s:

$$\frac{P(s)}{N + I_{own} + I_{other} - P(s)} = \tilde{\Delta}(s), \quad s = 1, ..., k. \tag{1}$$

where $\tilde{\Delta}(s) = \frac{E(s)}{WN_o} R(s), \quad I_{own} = \sum_{j=1}^{k} M(j)P(j), \quad I_{other} = i \times I_{own}$

(used when there is a constant number of mobiles).

•We can rewrite the power constraints as

$$\frac{P(s)}{N + I_{own} + I_{other}} = \Delta(s), \text{ where } \Delta(s) = \frac{\tilde{\Delta}(s)}{1 + \tilde{\Delta}(s)} \tag{2}$$

$$\Longleftrightarrow \tilde{\Delta}(s) = \frac{\Delta(s)}{1 - \Delta(s)},$$

$s = 1, ..., k.$

•Recall: $I_{own} + I_{other} = (1 + i) \sum_{j=1}^{k} M(j)P(j).$

•The solution is

$$P(s) = \frac{N\Delta(s)}{1-(1+i)\sum_{j=1}^{k} M(j)\Delta(j)}. \quad (3)$$

•The **Pole Capacity** is the polyhedron that makes the denominator vanish:

$$M^* = \{\mathbf{M} : 1 = (1+i)\sum_{j=1}^{k} M(j)\Delta(j)\}.$$

•A solution P is finite if and only if $\mathbf{M} < m$ (in the Pareto sense) for some $m \in M^*$.

•A finite solution P exists if and only if all its components are finite. Hence when accepting a call when capacity is reached, not only its own QOS will not be respected, but also all ongoing calls will suffer.

Definition: Let $\mathcal{M}$ be the subset of N^k for which $1 > (1+i)\sum_{j=1}^{k} M(j)\Delta(j)$, and let

$$\eta = \max_{m \in \mathcal{M}} (1+i) \sum_{j=1}^{k} m(j)\Delta(j). \tag{4}$$

We define the integer capacity $\mathcal{M}_B$ of the system as the boundary of $\mathcal{M}$ for which adding a call results in an infinite power.
It is the set of $\mathbf{M}$ for which

$$\eta = (1+i) \sum_{j=1}^{k} M(j)\Delta(j).$$

Definition: The blocking set $\mathcal{M}_B^j$ of class $j \in K$ is the subset of $\mathcal{M}$ for which another call of type j cannot be accepted:
$m \in \mathcal{M}_B^j$ iff $m \in \mathcal{M}$ and $m + e_j \notin \mathcal{M}$, where e_j is the unit vector in direction j.

3 Dynamic statistical model

•**Standard assumptions:**

- Arrival of class s calls: Poisson process with parameter λ_s.
- Duration of calls of class s: exponentially distributed with parameter μ_s.
- Load of class s: $\rho(s) := \lambda_s/\mu_s$.

•The process of ongoing calls is a Markov chain with a stationary distribution $\pi_\rho(\mathbf{M})$.

•**Definition:** The Erlang capacity $EC(\epsilon)$ is the set of vectors $\rho = (\rho(1), ..., \rho(k))$ s.t.

$$P_B(\rho) \leq \epsilon.$$

It is a set!

•**Theorem:** The steady state probabilities of the Markov chain are

$$\pi_\rho(\mathbf{M}) = \frac{1}{G_\rho} \prod_{s=1}^{k} \frac{\rho(s)^{M(s)}}{M(s)!}, \quad \mathbf{M} \in \mathcal{M}, \text{ where } G_\rho = \sum_{m \in \mathcal{M}} \prod_{s=1}^{k} \frac{\rho(s)^{m(s)}}{m(s)!}. \quad (5)$$

The probability $P_B^s(\rho)$ that a class s class be blocked and the average blocking probability $P_B(\rho)$ are

$$P_B^s(\rho) = \sum_{m \in \mathcal{M}_B^s} \pi_\rho(m), \quad P_B(\rho) = \sum_{s=1}^{k} \frac{\lambda_s}{\lambda} P_B^s.$$

•**Goal:** Study the behavior of these probabilities and the Erlang capacity as a function of the transmission rate while keeping the same processes of arrival and size of files.

4 Overview of the Results

•Homogeneous case, single cell:
If we slow down the transmission rate by a factor a,

- The blocking probability decreases,
- The Erlang capacity increases,
- As $a \to \infty$, the blocking probability tends to zero if $\rho < \mathcal{M}_B$.

- Example:
  - Consider a high speed application: $R = 160KB/sec$ (1.28 Mbps).
  - Let $\Delta = 0.199$; the system is dimensioned to accept at most 5 simultaneous calls.
  - The average quantity of transmitted information per call is 10KB (this is the measured average value over the Internet, see Sikdar et al, Perf. Eval, 2001).
  - Hence the average call duration is $\mu^{-1} = \frac{10KB}{160KB/sec} = 62.4msec$, et $\mu = 16.03$.
  - Assume that we wish to have a blocking probability of less than 1%. Then the Erlang Capacity is $\rho = 1.361$. We can accept $\lambda = \rho\mu = 21.8$ calls per sec.

Slowing factor a	1	2	3	4	5	6	20
Δ	0.199	0.110	0.0764	0.0583	0.0473	0.0398	0.0122
$\tilde{\Delta}$	0.248	0.124	0.0827	0.0620	0.0497	0.0414	0.0124
$\mathcal{M}_{B,a}$	5	9	13	17	21	25	81
Erl. Cap. EC(1%)	1.361	1.891	2.202	2.413	2.568	2.688	3.315
Arrival rate λ	21.8	30.3	35.29	38.67	41.15	43.08	53.13
call duration (msec)	62.4	124.8	187.2	249.6	312	374	1248
Gain in %	0	43.7	67.4	83.4	95	104	144

Table 1: Gain in Erlang capacity by slowing transmission rates by a factor of a

•**Significance of results:** if we propose to subscribers services that differ in offered throughput, we can determine how to price per volume of traffic as a function of the *effective resources* that each service consumes.

•Example: We see that we double the capacity when dividing throughput by five.

- Hence a service five times faster should cost the double per traffic volume.
- Other factors that depend on the throughput may influence the pricing such as the equipment.

Single cell, heterogeneous case:

•Assume that the distribution of the size (number of packets) of class s call is exponentially distributed with parameter $\zeta(s)$

Let $\nu(s) = \lambda(s)/\zeta(s)$,

Let

$$\delta(s) = \frac{E(s)}{WN_o} \quad \text{and hence } \tilde{\Delta}(s) = R(s)\delta(s) \text{ and } \Delta(s) = \frac{R(s)\delta(s)}{1 + R(s)\delta(s)},$$

•If we slow down the throughputs by a factor a, as $a \to \infty$, the blocking probability tends to zero if

$$\sum_{s=1}^{k} \nu(s)\delta(s) < 1.$$

•If $\sum_{s=1}^{k} \nu(s)\delta(s) > 1$ then for any $\mathbf{R}$,

$$\sum_{s} P_B^s \delta(s)\nu(s) > \sum_{s} \delta(s)\nu(s) - 1 > 0.$$

Other results:

•A single cell with both RT and NRT traffic:

We compute the limit blocking probabilities and the capacity when slowing down only the NRT traffic.

•The multiclass and multicell case

- We use a fixed-point approach to compute the blocking probabilities It is based on replacing $I_{other} = i \times I_{own}$ by

$$I_{other} = iE[I_{own}].$$

- We compute the limit capacity as we slow down the transmission throughputs.

Fixed point approach

•We obtain

$$P(s) = \frac{N\Delta(s)}{1 - \sum_{j=1}^{k} M(j)\Delta(j) - Q} \tag{6}$$

where $Q = i\sum_{j=1}^{k} E[M(j)]\Delta(j)$

•For each q (possibly different than $i\sum_{j=1}^{k} E[M(j)]\Delta(j)$), we obtain the probability distribution of $M(s), s = 1, ..., k$ as before:

$$\pi_\rho(\mathbf{M}) = \frac{1}{G_\rho}\prod_{s=1}^{k}\frac{\rho(s)^{M(s)}}{M(s)!}, \quad \mathbf{M} \in \mathcal{M}(q), \qquad \text{where } G_\rho = \sum_{m\in\mathcal{M}(q)}\prod_{s=1}^{k}\frac{\rho(s)^{m(s)}}{m(s)!}.$$

and where

$$\begin{aligned}
\mathcal{M}(q) &= \Big\{(m(1), ..., m(k)) : \sum_{j=1}^{k}\Delta(j)m(j) < 1 - q\Big\}, \\
\mathcal{M}_B^s(q) &= \Big\{(m(1), ..., m(k)) \in \mathcal{M} : \\
&\qquad \sum_{j=1}^{k}\Delta(j)m(j) \geq 1 - q - \Delta(s)\Big\}.
\end{aligned}$$

•Define

$$F(q) = i \sum_{j=1}^{k} E_q[M(j)]\Delta(j).$$

Then Q is the solution of the fixed point equation:

$$q = F(q). \tag{7}$$

•$F(q)$ is piecewize constant, and has thus discontinuities. Hence (7) need not have a solution. However, the set of values of i for which a solution to (7) does not exist has Lebesgue measure zero. A slight change in the value of i will yield a solution.

•$F(q)$ is nonincreasing in q which implies uniqueness of the solution to (7).

5 Extending the model

•Adding activity factors $\alpha(s)$ (between 0 and 1).
Approximation (Koo et al, 99): $I_{own} = \sum_{j=1}^{k} M(j)P(j)$ is replaced by
$I_{own} = \sum_{j=1}^{k} \alpha(j)M(j)P(j)$.

•Imperfect power control:
multiply the $\Delta(j)$'s by some constants that depend on the standard deviation of the received signal to interference ratio.

•Exact analysis is not tractable, see [Viterbi+Viterbi, 93].

Dynamic throughput

•NRT share the bandwidth leftover from RT traffic.

•RT bandwidth can be reduced when the load is high.
E.g., UMTS will use the Adaptive Multi-Rate (AMR) codec that offers 8 transmission rates of voice between 4.75 kbps to 12.2 kbps.

•Ongoing work with Nidhi Hedge

6 References

•I. Koo, J.H. Ahn, J. A. Lee and K. Kim, "Analysis of Erlang capacity for the multimedia DS-CDMA systems", IEICE trans. Fundamentals, May 1999.
(Includes Erlang capacity formula for multiservice systems using product form M/M/S/S type formulae).

•M. Meo, E. Viterbo, "Performance of wideband CDMA systems supporting multimedia traffic" IEEE Com Let, June 2001.
(Includes Erlang capacity formula for multiservice systems using product form M/M/S/S type formulae).