Statistical Multiplexing, Stochastic Knapsacks and Admission Control

Ravi R. Mazumdar

Dept. of ECE

University of Waterloo

Collaborators: F. Guillemin (Orange Labs), N. Likhanov (Russian Academy of Sciences), and F. Theberge (U of Ottawa)

Keynote given at ITC 21 Paris, Sept 17, 2009

Overview of talk

- Motivation for large networks
- Stochastic knapsacks and the Multirate-Erlang loss model - scaling
- QoS for packet switched networks
- Main mathematical insights
- Results
- Scalability and connection acceptance control
- Networks
- Conclusions

Current trends

- Response times are going up. Too many users with too many high-bandwidth peer-to-peer connections: Internet is victim of its success!
- The *best-effort* paradigm is not attractive for real-time services – leads to lack of willingness to pay for services
- Increasing pressure to provide performance guarantees (Quality of Service (QoS))
- Future is uncertain but unlikely that the best-effort model is going to attract paying users

Hence, the network must evolve into one capable of providing QoS

QoS Issues

- QoS is not a new issue – Well studied in the context of ATM and 1000's of papers.
- *Best-effort* not suited to provide hard QoS - we must allocate resources
- Solutions must be simple and yield substantial efficiency gains over simple resource reservation based on peak requirements
- Solutions must be scalable

University of Waterloo

Classical telephone networks

Circuit-switched: a call is allocated one circuit which it holds for the (random) duration. Calls arrive as a Poisson process.

Main performance measure: blocking probability i.e., the probability that on arrival a call cannot find a free circuit.

Solution: Erlang's formula (1917)

$$E(\lambda, C) = \frac{\lambda^C}{C!}\left[\sum_{n=0}^{C} \frac{\lambda^n}{n!}\right]^{-1}$$

C = Total number of circuits

Mean holding time of a call: 1 unit

Stochastic Knapsack

Stochastic Knapsack problem:

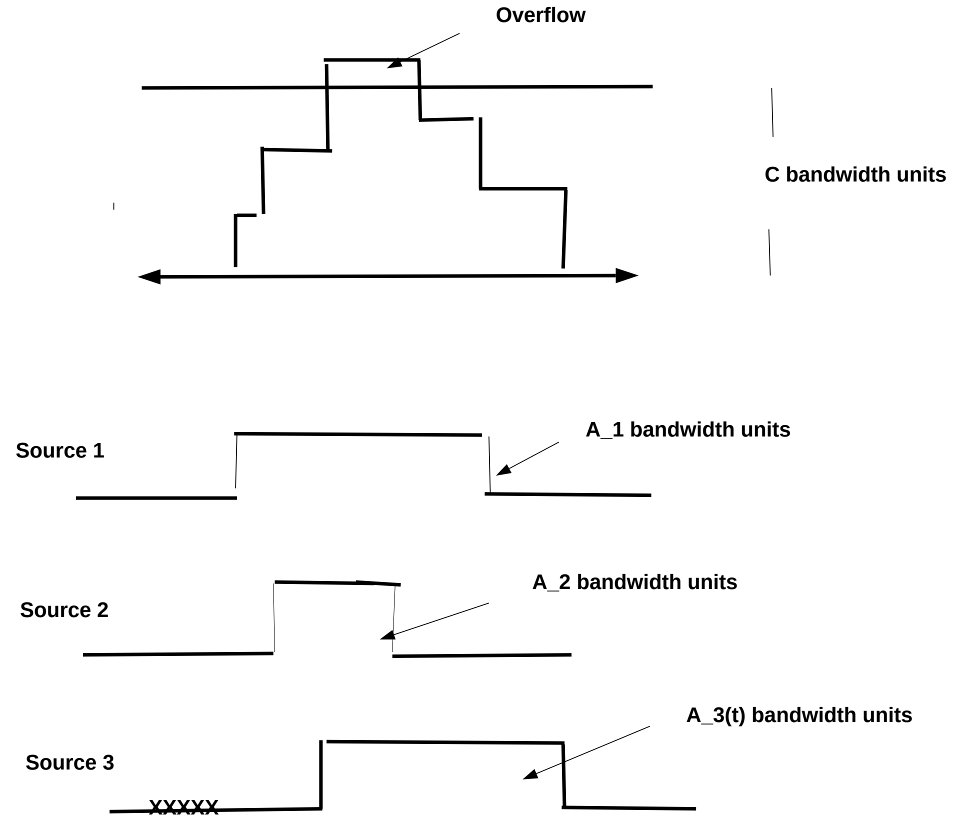
Given a knapsack (or container) of a given size and M objects of random size $\{S_k\}$.

How many objects can we fit in the container with minimal left over volume?

In our context, a link of rate C and connections with differing bandwidth requests that arrive randomly and stay for a random time.

In our context find out the probability that an arriving connection cannot be accommodated.

A more complicated model: Multi-rate loss model



Source 3 denied admission in circuit switched system

Occupancy distribution

Let $\mathbf{n} = (n_1, n_2, ..., n_M)$ be the vector of the number of sources of each type being carried. Then the stationary distribution has a product form given by

$$P(n_1, n_2, ..., n_M) = \frac{1}{G} \prod_{k=1}^{M} \frac{\lambda_k^{n_k}}{n_k!}$$

for $\mathbf{n} \in S$ where:

$$S \doteq \{\mathbf{n} : n_k \in Z; \sum_{k=1}^{M} A_k n_k \leq C\}$$

and the normalization constant G is given by

$$G = \sum_{\mathbf{n} \in S} \prod_{k=1}^{M} \frac{\lambda_k^{n_k}}{n_k!}$$

A source of type k gets blocked if upon arrival less than A_k bandwidth units are available. Therefore the blocking probability for type k is given by

$$P_k = \frac{1}{G} \sum_{\mathbf{n} \in X_k} \prod_{i=1}^{M} \frac{\lambda_i^{n_i}}{n_i!} \; ; \quad k = 1, 2, ..., M$$

and

$$X_k = \{\mathbf{n} : \; C - A_k < \sum_{m=1}^{M} n_m A_m \leq C\}$$

When M, C are large this is extremely computationally intensive. Order of calculations $O(CM)$. Difficult if CM is large. Thus we seek approximations for P_k.

Turns out that when the system is large then we can actually obtain explicit closed form expressions that are remarkably.

Notion of a large system

The notion of a large system is obtained by scaling both the capacity and arrival rates by a factor N. Define $C(N) = NC$ and $\lambda_k(N) = N\lambda_k$. Note this notion extends to networks

In other words the *large* system can be seen as a N fold scaling of a nominal system where connections arrive at rate λ_k, require A_k units of bandwidth, and the server capacity is C.

Main results

Let $P_k(N)$ denote the blocking probability of class k in the scaled system. We have to re-define the regions $S(N), X_k(N)$ and the corresponding normalization factor $G(N)$.

We consider the following 3 cases:

(**Light Load**) $\sum_1^M \lambda_k A_k < C$

(**Critical load**) $\sum_1^M \lambda_k A_k = C$

(**Heavy load**) $\sum_1^M \lambda_k A_k > C$

Main results for the multi-rate loss system

- Light load

$$P_k(N) = \exp(\tau_C d\epsilon)\frac{\exp(-NI(C))(1-\exp(\tau_C A_k))}{\sqrt{2\pi N}\sigma(1-\exp(\tau_c d))}\left(1+O(\frac{1}{N})\right)$$

- Critical load

$$P_k(N) = \sqrt{\frac{2}{\pi N}}\frac{A_k}{\sigma}\left(1+O(\frac{1}{\sqrt{N}})\right)$$

- Heavy load

$$P_k(N) = (1-\exp(\tau_C A_k))(1+O(\frac{1}{N}))$$

The parameters $I(C), \tau_C, \epsilon, \sigma, \delta$ and d are defined as

- d is the GCD of $\{A_1, A_2, ..., A_M\}$
- $\epsilon = \frac{NC}{d} - int(\frac{NC}{d})$
- τ_C is the unique solution to $\sum_1^M \lambda_k A_k \exp(\tau_C A_k) = C$
- $I(C) = C\tau_C - \sum_1^M \lambda_k(\exp(\tau_C A_k) - 1)$
- $\sigma^2 = \sum_1^M \lambda_k A_k^2 \exp(\tau_C A_k)$

Networks more difficult due to dependencies between link flows.

However, if we study networks when they are large (in a scaled regime) we can explicitly compute the blocking along any route and moreover we can show:

- Independence of blocking (i.e., single-link computations) holds if error of the order $O(\frac{1}{N})$ is required under light-to-critical loading

$$\mathcal{B}(\mathcal{N})_r = 1 - \prod_{A_{j,r} \neq 0} (1 - B_j(N))$$

where B_j is the blocking formula for a single link j and $A_{j,r} = 1$ if route r uses link j and is 0 otherwise.

NUMERICAL RESULTS

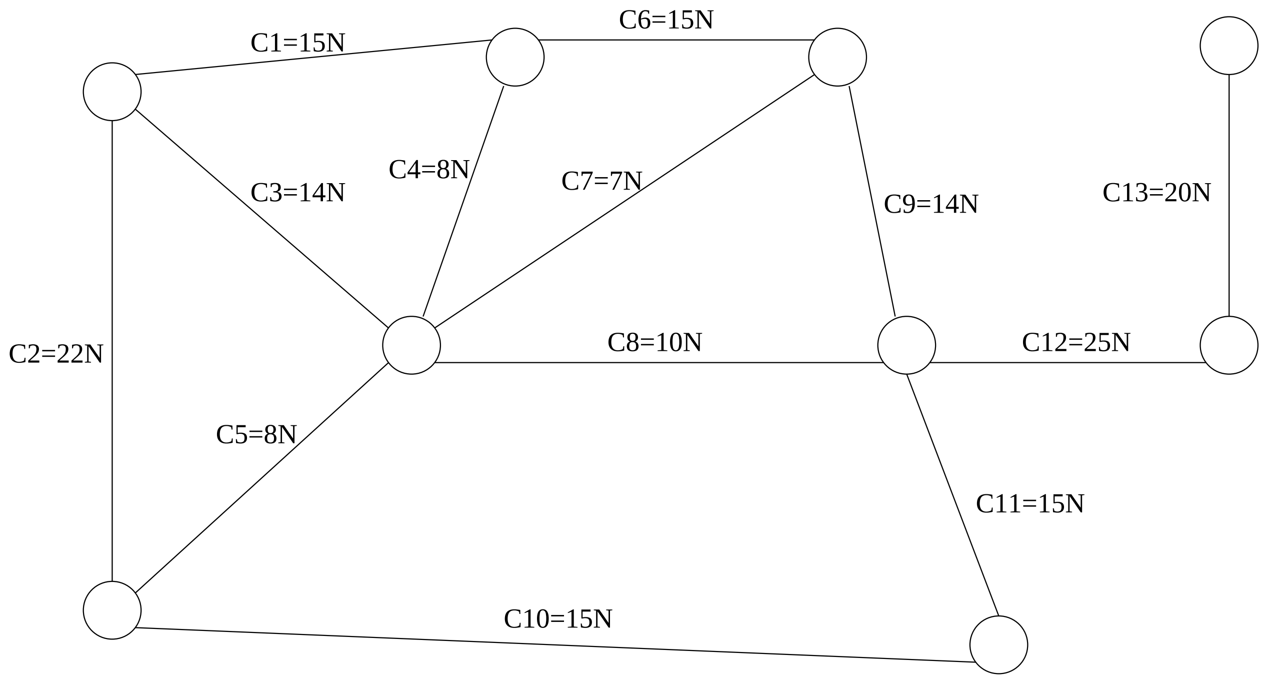


Figure 1: Typical network with scaling parameter N

A_r	r	ν_r	Simulation	Knapsack	Formula
1	1,2	4	(7.0-8.0)e-5	7.0e-5	7.23e-5
2	5,8,12	3	(4.0-4.3)e-4	4.4e-4	4.50e-4
3	6,9,12,13	2	(5.1-5.4)e-4	5.5e-4	5.57e-4
2	10,11,12,13	3	(3.5-3.7)e-4	3.9e-4	3.89e-4
1	3,4,6	6	$<$ 1e-5	$<$ 1e-5	2.01e-6
1	2,3,7,9	4	(7.0-8.0)e-5	7.0e-5	7.23e-5
2	8,11	1	(8.0-9.0)e-5	1.0e-4	1.11e-4
4	11,12,13	1	(7.3-7.7)e-4	8.1e-4	8.17e-4
2	1,2,10	3	(1.5-1.6)e-4	1.5e-4	1.54e-4
5	2	1	(4.0-4.3)e-4	4.1e-4	4.12e-4

Table 1: Blocking in large network with scaling N = 50. Note entries with $<$ cannot be estimated via simulation

University of Waterloo

Context of arriving session model

When sessions are streams or flows whose rate is variable (random) how do we determine its bandwidth?

The peak rate? Mean rate? Or is there a measure somewhere in between?

This has consequences in terms of allocating bandwidth and hence the total number of flows that can be accommodated by the server.

QoS approaches

Peak rate based QoS provisioning

- *Problem*: Very poor network utilization

Deterministic QoS based on traffic shaping

- *Metrics*: Worst case delays, zero loss
- *Problem*: Low network utilization.

Statistical QoS

- *Metrics*: Average delay, packet loss probability, tail of delay distribution
- *Advantage*: High network utilization
- *Problem:* Difficult to characterize for small systems
- *Solution:* Can obtain very tight explicit formulae for large systems

QoS with Mean Delay: Motivation

Consider the following $M/G/1$ model where there are N sources that are transmitting at a Poisson rate of λ packets per second. The server serves at a rate of C bits per second. The packet sizes are variable and uniformly distributed in $[0, M]$ where M represents the maximum packet size in bits.

- Stability implies $N_{stab}\lambda\frac{M}{2} < C$ or $N < \frac{2C}{\lambda M}$.
- Peak rate implies: $N_{peak} \leq \frac{C}{\lambda M}$ or half as many.

Now suppose the mean delay constraint is D then from the Pollaczek-Khinchine formula:

$$N_{mult} \leq \frac{C}{\frac{\lambda M^2}{6CD} + \frac{\lambda M}{2}}$$

and hence $N_{peak} \leq N_{mult} \leq N_{stab}$

Now we see that if $C \to \infty$ the number $N_{mult} \to N_{stab}$ or in other words as the capacity increases the bandwidth associated with a connection goes towards its mean (the notion of statistical multiplexing)

Suppose there are J different types of sources: The quantity:

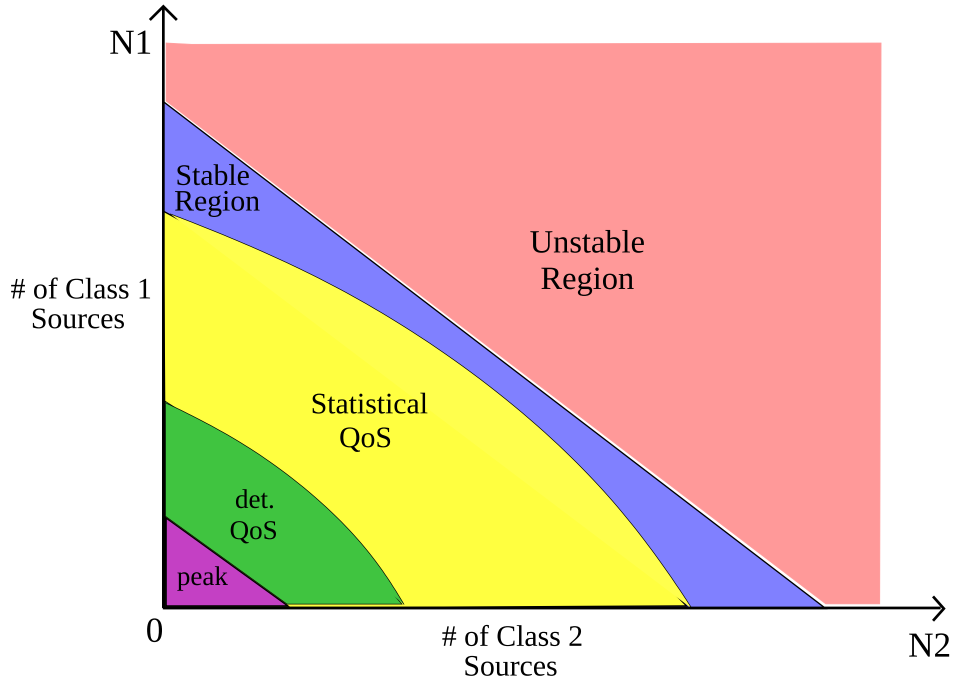
$$A_i = \frac{\lambda_i M_i^2}{6CD} + \frac{\lambda M_i}{2}$$

is what is referred to as the *effective bandwidth* and the rule for admission is:

$$\sum_{i=1}^{J} n_i A_i \leq C$$

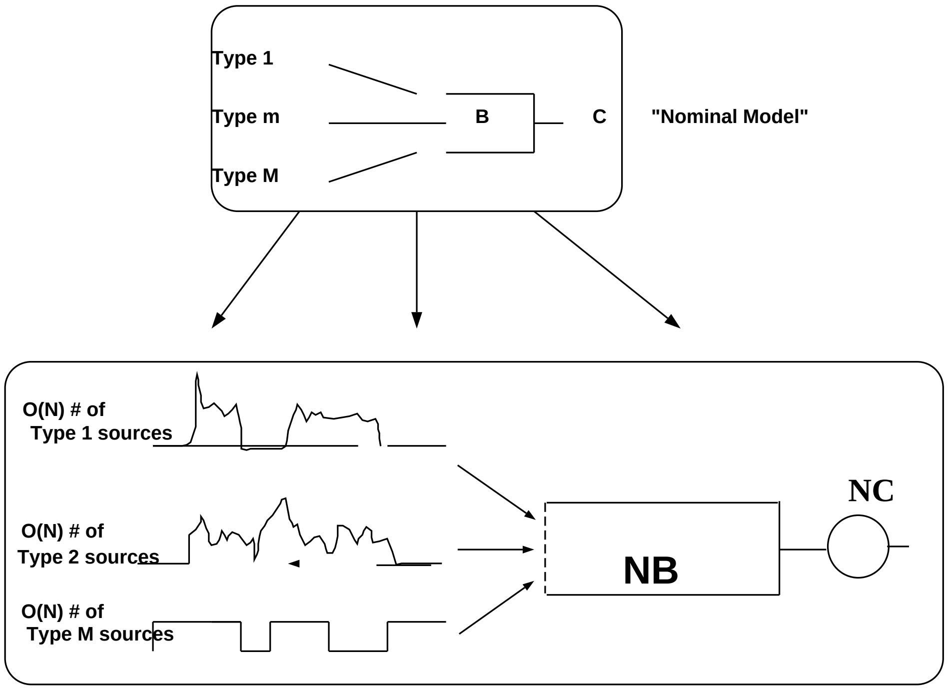
where n_i is the number of users of type i in the system, which looks like the condition for the multirate loss system.

Deterministic vs. statistical QoS



University of Waterloo

Multiplexer model- Overflows



Number of sources of type i = Nn_i

Total bits in [0,t] = $\sum$ x_i(0,t)

W_0 = sup { t $\geq$ 0: x(-t,0) - NCt} Buffer Content

Overflow prob. P(W_0 > NB)

Large Buffer Model : N-fold Scaling

Model

Discrete-time model for cell flow.

Total number of sources: N. Buffer size : NB Server speed = NC

Source assumptions: Independent, identical sources.

Server is assumed to be work conserving.

Source i emits $\lambda_{i,t}$ number of bits at time t.

Assumption: $E[\lambda_{i,t}] < C$ (stability assumption)

Let $X_i(0,t]$ denote the total number of bits emitted by source i in $(0,t]$.

$$X_i(0,t] = \sum_{j=1}^{t} \lambda_{i,j}$$

Assumption: $X_i(0,t]$ is a stationary, increment process.

Statistical QoS measures

- Loss Ratio (LR) (fraction of bits lost) is defined as:

$$LR = \frac{E[\sum_{t=1}^{T} (W_{t-1}^{(N)} + \lambda_t^{(N)} - N(C+B))^+]}{EX^{(N)}(0,T]}$$

  Note by stationarity, LR = Bit Loss Probability.

- Overflow probability or delay tail distribution (under FIFO)

$$\mathbf{P}\left(W^N > NB\right)$$

Bahadur-Rao Theorem

Let $\phi_t(h)$ denote the moment generating function of $X_i(0,t]$. Then uniformly in the argument $N(Ct+B)$:

$$\mathbf{P}\{X^{(N)}(0,t] \geq N(Ct+B)\} = \frac{e^{-NI_t(C,B)}}{\tau_t\sqrt{2\pi N\sigma_t^2}}\left(1+O(\frac{1}{N})\right)$$

where

- $$\begin{aligned} I_t(C,B) &= \sup_{\theta\geq 0}\{(Ct+B)\theta - \log\phi_t(\theta)\} \\ &= (Ct+B)\tau_t - \log(\phi_t(\tau_t)) \end{aligned}$$

- τ_t is the unique solution to

$$\frac{\phi_t'(\tau_t)}{\phi_t(\tau_t)} = Ct + B$$

-

$$\sigma_t^2 = \frac{\phi_t''(\tau_t)}{\phi_t(\tau_t)} - (Ct + B)^2$$

Idea is based on exponential measure change to set mean to $Ct + B$ and then use local Gaussian limit theorem exactly as for the loss system case.

Main result for overflow probabilities

Hypotheses

H1: $\exists$ a unique $t_0 < \infty$ such that:

$$I_{t_0}(C,B) = \min_{t \geq 1} I_t(C,B) > 0$$

H2

$$\liminf_{t \to \infty} \frac{I_t(C,B)}{\log t} > 0$$

(this is satisfied by "self-similar" sources)

Then as $N \to \infty$, uniformly in NB

$$\mathbf{P}\{Y^{(N)} > NB\} = \frac{e^{-NI_{t_0}(C,B)}}{\tau_{t_0}\sqrt{2\pi\sigma_{t_0}^2 N}} \left(1 + O(\frac{1}{N})\right)$$

Loss probability

Under hypotheses H1 and H2, as $N \to \infty$

$$LR = \frac{e^{-NI_{t_0}(C,B)}}{\tau_{t_0}^2 C\rho\sqrt{2\pi N\sigma_{t_0}^2 N^3}}\left(1 + O(\frac{1}{N})\right)$$

where $\rho = \frac{E[\lambda_{t,1}]}{C}$ is the average load.

Note: Constant is of order $O(N^{-\frac{3}{2}})$ implying for large systems $N \sim 100 - 1000$ only considering exponential (as is done in many studies) gives LR two orders of magnitude off – i.e., if we design for 10^{-9} using only exponential then actual performance is 10^{-11} – conservative.

Simulation results

Deterministic ON-OFF Sources $\lambda_0 = \lambda_1 = 25$, and $\lambda_t = 0; t = 2, 3, ...49$.

These sources are periodic with period 50. The sources are randomly phase shifted in $[0, 49]$

$C = 2.5N$.

N	Simulation (90% confidence)	Formula
50	(-2.2915, -2.2684)	-2.1106
75	(-3.0144, -2.9625)	-2.9468
100	(-3.7310, -3.6428)	-3.7063
150	(-5.2031, -4.8751)	-5.1145

Table 2: Loss probabilities in finite buffers

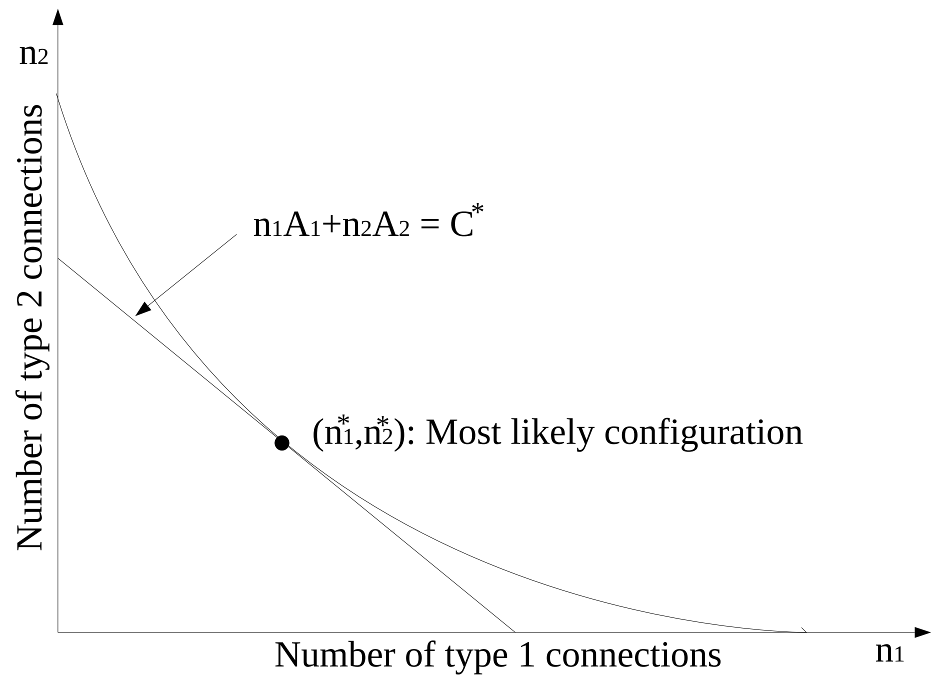
Engineering insights

- Statistical multiplexing gains are obtained whether sources are conventional or *self-similar* when many sources are multiplexed (exponentially decreasing in N)
- The parameter t_0 is called the critical time scale of a source. It is the most likely time scale for buffer overflow. Also it defines the time interval over which we need to measure source statistics.

Engineering implications: Large number of sources actually helps in the context of buffer design providing multiplexing gains irrespective of the type of sources i.e."self-similarity" and long-range dependence do not matter in the core of the network.

Connection acceptance control

To develop CAC we need to estimate the bandwidth of a connection i.e. the $\{A_k\}$ in the multi-rate model. How do we define it?



C^* = Effective capacity
A_i = Effective bandwidth of type *i*

Acceptance Region

Suppose the QoS for loss is ε.

Define:

$$\Omega_\varepsilon = \{\{n_i\}_{1=1}^{M} : P_L \leq \varepsilon\}$$

Then Ω_ε is the acceptance region.

Define the boundary configurations

$$\partial\Omega_\varepsilon = \{\mathbf{n} : P_L = \varepsilon\}$$

Acceptance region- contd.

Once we have Ω_ε we can study some properties.

Coordinate convexity: Let **S** be a set of possible configurations. Then **S** is said to be coordinate convex if for $\mathbf{n} \in S$, the vector $\mathbf{n} - e_k \in S$ for all $n_k > 0$ and $k = 1, 2, \ldots, M$.

- **Fact 1:** The set Ω_ε is co-ordinate convex under the "true" loss probability.
- **Fact 2:** The set Ω_ε is co-ordinate convex under $P_L(.)$ (approximation) for large N.

Ramifications: Co-ordinate convexity implies that the equilibrium distribution of the configurations is given by a "*product-form*".

i.e.

$$\Pi(\mathbf{m}) = \frac{1}{G} \prod \frac{(N\lambda_i)^{m_i}}{m_i!}$$

where G is the normalizing constant given by:

$$G = \sum_{\mathbf{m} \in \Omega_\varepsilon} \frac{(N\lambda_i)^{m_i}}{m_i!}$$

Most likely loss state

Definition: The configuration $\mathbf{m}^* \in \partial\Omega_\varepsilon$ which maximizes $\Pi(\mathbf{m})$ is called the most likely loss state.

Properties:

- $\mathbf{m}^*$ is unique
- Let $\mathbf{m}$ be any other state in $\partial\Omega_\varepsilon$. Then:

$$\frac{\Pi(\mathbf{m})}{\Pi(\mathbf{m}^*)} \sim O(e^{-N})$$

Implications: loss is concentrated at $\mathbf{m}^*$

Effective rate

Idea is to associate a bandwidth assignment to a call such that if admitted the call will satisfy the QoS and we can use the multi-rate loss model for blocking.

Questions:

- What is the effective rate?
- What are the properties?
- What is the coupling between loss and the effective rate?

Effective rates?

Effective rates= Effective Bandwidth idea due to Hui and Kelly.

The idea is to replace the burstiness of traffic flow by an equivalent bandwidth requirement.

- Effective bandwidth is defined as $A_i = \frac{\Gamma_{i,t}(\theta)}{\theta}$ where $\Gamma_t(\theta) = \log M_{i,t}(\theta)$
- $r_{i,min} \leq A_i \leq r_{i,peak}$
- $A_i \to r_{i,mean}$ as the number of sources becomes large.

Effective rates

Having identified $\mathbf{m}^*$ let us compute it explicitly for our model where we replace C by $C + \frac{B}{t_0}$.

$$m_j^* = N\lambda_j(\phi_j(\tau_c))^{\mathbf{y}} \exp\{\frac{\mathbf{y}}{N\Gamma^2}\left[(1 + \frac{2}{e^{\tau_c} - 1})\frac{\phi_j'(\tau_c)}{\phi_j(\tau_c)}\right]\}$$

where $\mathbf{y}$ is a Lagrange multiplier (for constraint satisfaction) and τ_c satisfies:

$$\sum_{i=1}^{M} m_i^* \frac{\phi_i'(\tau_c)}{\phi_i(\tau_c)} = C$$

We have $(M+2)$ unknowns and (M+2) equations to solve for the unknowns $\tau_c, \mathbf{y}, m_i^*$.

Effective rates (contd.)

Taking the gradient of $P(loss)$ at the most likely state gives:

$$a_j = \ln(\phi_j(\tau_c)) + \frac{1}{N\Gamma^2}\left(1 + \frac{2}{e^{\tau_c} - 1}\right)\frac{\phi_j'(\tau_c)}{\phi_j(\tau_c)}$$

Define:

$$A_j = \frac{a_j}{a_{\min}}$$

Then A_j denotes the slope of the hyperplane at m^* (normalized to the minimum of a_j). This is nothing but the sensitivity of the loss probability and therefore the natural definition of the *effective rate* of the connection.

Define:

$$C^* = \sum_{i=1}^{M} A_j m_j^*$$

Then C^* denotes the effective capacity of the VP.

The interpretation: for statistical multiplexing C^* corresponds to C to be able to use the linear decision rule since C^* defines the hyperplane:

$$T_\varepsilon = \{\mathbf{m} : \sum_{i=1}^{M} A_i m_i = C^*$$

CAC Procedure

- Compute m_j^* and A_j for each connection.
- If $A_{incoming} + \sum_{ongoing} A_i n_i < C$ accept request.

  Else reject request

Properties of effective rates in large systems

Let us keep ε fixed and see some properties as N increases

- $\mathbf{m}(\mathbf{N})$ converges to m^0 such that $\sum_{i=1}^{M} m_i^0 r_i = C$ where r_i is the mean rate of source i.
- $A_j(N)$ converges to $\frac{r_i}{r_{\min}}$
- Hyperplane is exact in the limit i.e. $T_\varepsilon = \partial\Omega_\varepsilon$. This implies tat the boundary of the acceptance region coincides with the boundary of the stability region.

Example

Consider multiplexing two classes of ON-OFF sources. $C = 2000$, $N = 100$. Source 1: $\lambda_1 = 14$, $p_1 = 0.275$, $Peak_1 = 2$ Source 2: $\lambda_2 = 14$, $p_2 = 0.8$ and $Peak_2 = 1$.

From which we obtain: $A_1 = 1.0$, $A_2 = 1.385$ and $C^* = 3384.7$

To check that our rate or bandwidth assignment is right the multi-rate blocking rate formula must give consistent results.

Technique	Class 1 blocking	Class 2 blocking
Simul. (95 % conf. int.)	.00427-.00501	.00631-.00724
Theorem	.00479	.00661

Table 3: Connection blocking probabilities

This procedure defines an acceptance region of the form $\sum_j A_j n_j \leq NC^*$ The table below indicates simulation results the loss probability for a region that is designed for loss of order of 10^{-4}.

Number of Class 1 calls	Number of Class 2 calls	Base 10 logarithm of 95% conf. int. for loss
500	2083	(-4.13,-4.03)
1000	1722	(-4.19,-4.11)
1416	1422	(-4.16,-4.09)
1500	1361	(-4.30,-4.23)
2000	1000	(-4.25,-4.17)

Table 4: Packet loss values

Concluding remarks

- Mathematical analysis of large communication networks can provide many insights
- Identifying features such as critical time scales have important measurement implications
- In large systems source characteristics (long-range dependence etc.) do not affect behavior
- Extremely accurate formulae for dimensioning and allocating resources
- Large networks are in fact easier to analyse, even end-to-end!
- There is no single mathematical tool but large deviations and Palm theory play a key role
- Important new concepts such as effective bandwidths have emerged
- Thousands of long simulations needed to obtain the same knowledge

References

P. Gazdzicki, I. Lambadaris, R.R. Mazumdar, *Blocking probabilities for large multirate Erlang loss systems*, Adv.Appl.Prob. 25, 1993 pp. 997-1009

A. Simonian, F. Théberge, J. Roberts, and R. R. Mazumdar, ;*Asymptotic estimates for blocking probabilities in a large multi-rateloss network*, Advances in Applied Probability, Vol. 29, No. 3, 1997, pp. 806-829.

D.Mitra, J.Morrison, *Erlang capacity and uniform approximations for shared unbuffered resources*, IEEE/ACM Transactions on Networking, vol.2, N.6, 1994, pp.558-570

J.Y. Hui, *Resource allocation in broadband networks,* IEEE Journal of Selected Areas in Communications, 1989, 6:1598–1608. (Original idea on Effective Bandwidths)

F.P. Kelly, *Notes on effective bandwidths*, in Stochastic Networks, Kelly, F.P., Zachary, S., and Zeidins, I., editors, Oxford University Press, 1996

N. B. Likhanov, R. R. Mazumdar, and F. Theberge Providing QoS in Large Networks: Statistical Multiplexing and Admission Control In E. Boukas and R. Malhame , editors, *Analysis, Control and Optimization of Complex Dynamic Systems*, Springer 2005, pp 137-168.

R. R. Mazumdar, *Performance Modeling, Loss Networks and Statistical Multiplexing*, to be published by Morgan and Claypool, San Francisco, Dec. 2009.