LNMB Course
Asymptotic Methods in Queueing Theory

Lecture 4, March 4, 2013

Rudesindo Núñez-Queija (CWI/UvA), Sem Borst (TU/e)

http://www.win.tue.nl/~sem/AsQT/
Course overview

Four main topics

- Large deviations and tail asymptotics
  - Introduction, large deviations, large-buffer asymptotics for light-tailed queues
  - Many-sources asymptotics, large-buffer asymptotics for heavy-tailed queues
  - Large-buffer asymptotics for heavy-tailed queues, impact of service discipline, Processor Sharing and its variants

- Fluid limits
  - (In)stability proofs based on fluid limits

- Perturbation analysis and time scale separation

- Heavy-traffic approximations
Fluid limits

Let \( \{X(t)\}_{t \geq 0} \) be some continuous-time stochastic process

Specific scenario of interest is a Markov process representing the queue lengths of various classes or at various nodes in a queueing network (with Poisson arrivals and exponentially distributed service requirements)

In order to obtain fluid limits, the original stochastic process is scaled in both space and time

More specifically, we consider a sequence of processes \( \{X^{(R)}(t)\}_{t \geq 0} \), indexed by a sequence of positive integers \( R \)

- Each process is governed by similar statistical laws as the original process
- The initial states satisfy \( ||X^{(R)}(0)|| = R \), e.g., \( \sum_i X_i^{(R)}(0) = R \), and \( \frac{1}{R}X^{(R)}(0) \to Q(0) \) as \( R \to \infty \)
Fluid limits (cont’d)

The process \( \{\bar{X}^{(R)}(t)\}_{t \geq 0} \), with \( \bar{X}^{(R)}(t) = \frac{1}{R}X^{(R)}(Rt) \), is referred to as the fluid-scaled version of the process \( \{X^{(R)}(t)\}_{t \geq 0} \)

- movement in time is slowed down by factor \( R \)
- movement in space is reduced by factor \( R \)

Consequently, the dynamics of the process \( \{\bar{X}^{(R)}(t)\}_{t \geq 0} \) continue to be of order \( O(1) \)

Any (possibly random) weak limit \( \{\bar{X}(t)\}_{t \geq 0} \) of the sequence \( \{\bar{X}^{(R)}(t)\}_{t \geq 0} \) as \( R \to \infty \) is called a fluid limit
Fluid limits (cont’d)

Fluid limits may be interpreted as first-order approximations of the original stochastic process

For example, consider the queue length process \( \{Q(t)\}_{t \geq 0} \) of an M/M/1 queue with arrival rate \( \lambda \) and service rate \( \mu \)

It may be shown that the fluid limit \( \{\bar{Q}(t)\}_{t \geq 0} \) obeys the differential equation

\[
\frac{d}{dt} \bar{Q}(t) = \lambda - \mu
\]

as long as \( \bar{Q}(t) > 0 \), and thus follows a linear trajectory

\[
\bar{Q}(t) = \max\{\bar{Q}(0) + (\lambda - \mu)t, 0\},
\]

with \( \bar{Q}(0) = \lim_{R \to \infty} \frac{1}{R} Q^{(R)}(0) \)
Fluid limits (cont’d)

Note that the fluid limit \( \{\bar{Q}(t)\}_{t \geq 0} \) of the M/M/1 queue shows a dichotomy in long-term behavior:

- In case \( \lambda < \mu \), i.e., the system is stable in the sense that the queue length process \( \{Q(t)\}_{t \geq 0} \) is positive-recurrent, \( \{\bar{Q}(t)\}_{t \geq 0} \) reaches 0 and then remains there after a finite time \( \tau = \bar{Q}(0)/(\mu - \lambda) \).

- In case \( \lambda > \mu \), i.e., the system is unstable in the sense that the queue length process \( \{Q(t)\}_{t \geq 0} \) is transient, \( \{\bar{Q}(t)\}_{t \geq 0} \) grows at a linear rate \( \lambda - \mu \) forever and wanders off to infinity.
Fluid limits (cont’d)

The above-described dichotomy for the M/M/1 queue is representative of a much broader connection between the long-term behavior of the fluid limit and the (non-)ergodicity of the original stochastic process:

- If for all initial states $\tilde{X}(0)$ with $||\tilde{X}(0)|| = 1$, the fluid limit $\{\tilde{X}(t)\}_{t \geq 0}$ reaches state 0 and remains there after some finite time $\tau$, then the original stochastic process $\{X(t)\}_{t \geq 0}$ is positive-recurrent

- If the fluid limit $\{\tilde{X}(t)\}_{t \geq 0}$ exhibits ‘linear growth’, then the original stochastic process $\{X(t)\}_{t \geq 0}$ is transient
(Single-class) Jackson networks

Consider a (single-class) Jackson network with $N$ queues

- Customers arrive at node $i$ (from the external environment) as a Poisson process of rate $\lambda_i$
- Service requirements at node $i$ are exponentially distributed with parameter $\mu_i$
- After service completion at node $i$, customers proceed to node $j$ with probability $p_{ij}$, $j = 1, \ldots, N$, or leave the network (to the external environment) with probability $q_i = 1 - \sum_{j=1}^{N} p_{ij}$
Jackson networks (cont’d)

The throughputs $\nu_i$ (effective total arrival rates) at the various nodes are determined by the traffic equations (assuming the network is stable)

$$\nu_i = \lambda_i + \sum_{j=1}^{N} \nu_j p_{ji},$$

or in matrix-vector form,

$$\mathbf{\nu} = (\mathbf{I} - \mathbf{P}^T)^{-1}\mathbf{\lambda},$$

with $\lambda = (\lambda_i)_{i=1,\ldots,N}$, $\nu = (\nu_i)_{i=1,\ldots,N}$, $\mathbf{I}$ denoting the $N \times N$ identity matrix, and $\mathbf{P} = (p_{ij})_{i,j=1,\ldots,N}$ representing the matrix of routing probabilities.

It is known that $\nu_i < \mu_i$ for all $i = 1, \ldots, N$ is a sufficient and necessary condition for the network to be stable (with a product-form distribution for the joint queue length process)
Jackson networks: fluid-limit perspective on stability

It may be shown that the fluid limit \( \{\bar{Q}(t)\}_{t \geq 0} \) of the above-described Jackson network satisfies the equations

\[
\bar{Q}(t) = \bar{Q}(0) + \lambda t - (I - P^T)\mu\bar{T}(t) = \bar{Q}(0) + \lambda t - (I - P^T)\mu t + (I - P^T)\mu\bar{I}(t),
\]

\( \bar{Q}(t) \geq 0, \quad \bar{T}(0) = 0, \bar{T}(t) \) is non-decreasing, \quad \( \bar{I}(t) = et - \bar{T}(t) \) is non-decreasing,

\[
\int_{t=0}^{\infty} \bar{Q}(t)d\bar{I}(t) = 0,
\]

where \( \bar{T}_i(t) \) and \( \bar{I}_i(t) \) may be interpreted as the cumulative amounts of active and inactive time at node \( i \) up to time \( t \)

Or equivalently,

\[
\frac{d}{dt} \bar{Q}_i(t) = \lambda_i - \mu_i \frac{d}{dt} \bar{T}_i(t) + \sum_{j=1}^{N} \mu_j p_{ji} \frac{d}{dt} \bar{T}_j(t)
\]

and \( \bar{I}_i(t) \) cannot increase when \( \bar{Q}_i(t) > 0 \)
Define the diagonal matrix $M$ by $M_{ii} = 1/\mu_i$ for all $i = 1, \ldots, N$

Define the Lyapunov function

$$F(t) = e^T M (I - P^T)^{-1} \tilde{Q}(t),$$

representing a weighted sum of the queue lengths at the various nodes, and denote

$$\kappa = 1 - \max_{i=1,\ldots,N} \rho_i > 0,$$

with $\rho_i = \nu_i/\mu_i < 1$ for all $i = 1, \ldots, N$
Jackson networks: fluid-limit perspective on stability (cont’d)

Based on the equations for the fluid limit \( \{\bar{Q}(t)\}_{t \geq 0} \), it may be shown that

\[
\frac{d}{dt} F(t) \leq \max_{i=1, \ldots, N} -(1 - \rho_i) = -1 + \max_{i=1, \ldots, N} \rho_i = -\kappa < 0
\]

as long as \( \bar{Q}(t) \neq 0 \)

It follows that \( F(t) \) will hit 0 within a finite time \( \tau = F(0)/\kappa \), and subsequently remain 0

Hence, the fluid limit \( \{\bar{Q}(t)\}_{t \geq 0} \) must be in state 0 after time \( \tau \) as well, which implies that the original joint queue length process \( \{Q(t)\}_{t \geq 0} \) is positive-recurrent
Multi-class queueing networks

Consider a multi-class network with $K$ customer classes and $N$ queues

- Class-$k$ customers arrive at node $i$ (from the external environment) as a Poisson process of rate $\lambda_i^{(k)}$.

- Service requirements of class-$k$ customers at node $i$ are exponentially distributed with parameter $\mu_i^{(k)}$.

- After service completion at node $i$, class-$k$ customers proceed to node $j$ and turn into class-$l$ customers with probability $p_{ij}^{(kl)}$, $j = 1, \ldots, N$, $l = 1, \ldots, K$, or leave the network (to the external environment) with probability $q_i^{(k)} = 1 - \sum_{j=1}^{N} \sum_{l=1}^{K} p_{ij}^{(kl)}$. 

Multi-class queueing networks (cont’d)

The throughputs $v_i^{(k)}$ of the various customer classes at the various nodes are determined by the traffic equations (assuming the network is stable)

$$v_i^{(k)} = \lambda_i^{(k)} + \sum_{j=1}^{N} \sum_{l=1}^{K} v_j^{(l)} p_{ji}^{(lk)},$$

or in matrix-vector form, $v = (I - P^T)^{-1} \lambda$, with $\lambda = (\lambda_i^{(k)})_{i=1,\ldots,N,k=1,\ldots,K}$, $v = (v_i^{(k)})_{i=1,\ldots,N,k=1,\ldots,K}$, $I$ denoting the $KN \times KN$ identity matrix, and $P = (p_{ij}^{(kl)})_{i,j=1,\ldots,N,k,l=1,\ldots,K}$

Define $\rho_i = \sum_{k=1}^{K} v_i^{(k)}/\mu_i^{(k)}$

Based on the analogy with single-class networks, it would seem plausible that $\rho_i < 1$ for all $i = 1, \ldots, N$ is a sufficient and necessary condition for the network to be stable...
Multi-class networks (cont’d)

Indeed, it may be shown that the fluid limit \( \{ \bar{Q}(t) \}_{t \geq 0} \) of the above-described multi-class network satisfies the equations

\[
\bar{Q}(t) = \bar{Q}(0) + \lambda t - (I - P^T)\mu \bar{T}(t)
= \bar{Q}(0) + \lambda t - (I - P^T)\mu t + (I - P^T)\mu \bar{I}(t),
\]

\( \bar{Q}(t) \geq 0, \)

\( \bar{T}(0) = 0, \bar{T}(t) \) is non-decreasing,

\( \bar{I}(t) = et - \bar{T}^{\text{tot}}(t) \) is non-decreasing,

\[
\int_{t=0}^{\infty} \bar{Q}^{\text{tot}}(t) \, d\bar{I}(t) = 0,
\]

where \( \bar{Q}^{\text{tot}}_i(t) = \sum_{k=1}^{K} \bar{Q}^{(k)}_i(t), \bar{T}^{\text{tot}}_i(t) = \sum_{k=1}^{K} \bar{T}^{(k)}_i(t), \) and \( \bar{T}^{(k)}_i(t) \) may be interpreted as the cumulative amount of time spent on service of class-\( k \) customers at node \( i \) up to time \( t \).
Multi-class networks (cont’d)

However, the necessary condition $\rho_i < 1$ for all $i = 1, \ldots, N$ is in general not sufficient for a multi-class network to be stable.

It is sufficient when the service disciplines at the various nodes are symmetric, e.g., Processor-Sharing (PS), Last-Come First-Served Preemptive (LCFS-P), even for generally distributed service requirements.

In that case the joint queue length process is known to have a product-form stationary distribution.

But for First-Come First-Served (FCFS) and various other service disciplines, the network may be unstable, even if $\rho_i < 1$ for all $i = 1, \ldots, N$. 
Multi-class networks (cont’d)

As an illustrative example, consider a network with $N = 2$ nodes and $K = 6$ classes

- Class 1 customers arrive at node 1 as a Poisson process of rate $\lambda_1^{(1)} = 1$, with mean service requirement $m_1^{(1)} = 0.001$

- After service completion at node 1, class-1 customers proceed to node 2 and turn into class-2 customers, with mean service requirement $m_2 = 0.897$

- After service completion at node 2, class-$k$ customers return to node 2 as class-$(k + 1)$ customers, with mean service requirement $m_{k+1} = 0.001$, $k = 2, 3, 4$

- After service completion at node 2, class-5 customers proceed to node 1 and turn into class-6 customers, with mean service requirement $m_6 = 0.899$

- After service completion at node 1, class-6 customers leave the network
Multi-class networks (cont’d)

Note that $\lambda_2^{(2)} = \lambda_2^{(3)} = \lambda_2^{(4)} = \lambda_2^{(5)} = 1$ and $\lambda_1^{(6)} = 1$, so that the nominal loads at the two nodes in the above-described network are $\rho_1 = m_1 + m_6 = 0.9 < 1$ and $\rho_2 = m_2 + m_3 + m_4 + m_5 = 0.9 < 1$

Yet, the network is unstable

The queue length process exhibits oscillatory behavior, with an increasing amplitude

The two nodes suffer starvation in an alternating fashion, with one server being idle and wasting its capacity and the other server dealing with a huge queue most of the time, which is the root cause of the instability
Bandwidth-sharing networks: Two-node three-class example

Network consists of two unit-capacity links (resources, servers) and three classes of users (flows, customers):

- Class-\(i\) users require service from link \(i\) only, \(i = 1, 2\)
- Class-0 users require same service rate from links 1 and 2 simultaneously
Bandwidth-sharing networks: Two-node three-class example (cont’d)

Assume population of users is $(n_0, n_1, n_2) = (1, 1, 1)$.

How to allocate service rates $r_0, r_1, r_2$ subject to link capacity constraints $r_0 + r_i \leq 1$, $i = 1, 2$?

- **Option 1:** $r_0 = r_1 = r_2 = \frac{1}{2}$
  Rate-fair, but not resource-fair: user $0$ receives twice as much service resources than users $1$ and $2$, as it is served by both links

- **Option 2:** $r_0 = \frac{1}{3}$, $r_1 = r_2 = \frac{2}{3}$
  Resource-fair, but not rate-efficient: higher aggregate rate can be achieved by allocating all service resources to users $1$ and $2$

- **Option 3:** $r_0 = 0$, $r_1 = r_2 = 1$
  Efficient (or so it seems), but unfair to user $0$!
Bandwidth-sharing networks: Model description

Bandwidth-sharing (BS) networks

- Network consists of several links (resources, servers) indexed by set $\mathcal{L} = \{1, \ldots, L\}$
- Links shared by several classes of users (flows, customers) indexed by set $\mathcal{K} = \{1, \ldots, K\}$
- Class-$k$ users require simultaneous service from subset of links $\mathcal{L}_k \subseteq \mathcal{L}$
Bandwidth-sharing networks: Rate allocation problem

Find rate allocation \((r_k)_{k \in K}\) that solves utility maximization problem:

\[
\begin{align*}
\max_{k \in K} & \quad \sum_{k \in K} n_k U_k(r_k) \\
\text{subject to} & \quad \sum_{k \in K_l} n_k r_k \leq C_l & l \in L \\
& \quad r_k \geq 0 & k \in K,
\end{align*}
\]

with

- \(U_k(\cdot)\): concave utility function
- \(n_k\): number of class-\(k\) users
- \(C_l\): capacity of link \(l\)
- \(K_l = \{k : l \in L_k\}\): classes that require capacity from link \(l\)
Bandwidth-sharing networks: Rate allocation problem (cont’d)

Family of $\alpha$-fair rate allocation policies:

$$U(r) = \frac{r^{1-\alpha}}{1-\alpha}$$

- $\alpha \downarrow 0$: maximum throughput
- ‘$\alpha = 1$’: proportional fairness
- $\alpha = 2$: ‘TCP’
- $\alpha \rightarrow \infty$: max-min fairness
Bandwidth-sharing networks: Two-node three-class example (cont’d)

- $\alpha = 0$: $\max r_0 + r_1 + r_2$: $r_0 = 0, r_1 = r_2 = 1$
- $\alpha = 1$: $\max \log(r_0) + \log(r_1) + \log(r_2)$: $r_0 = \frac{1}{3}, r_1 = r_2 = \frac{2}{3}$
- $\alpha \to \infty$: $\max \min\{r_0, r_1, r_2\}$: $r_0 = r_1 = r_2 = \frac{1}{2}$
Bandwidth-sharing networks: Model description (cont’d)

User dynamics:

- Class-$k$ users arrive as Poisson process of rate $\lambda_k$
- Class-$k$ users have generally distributed service requirements $B_k$
- Class-$k$ traffic intensity is $\rho_k := \lambda_k \mathbb{E}\{B_k\}$
- Whenever population of users is $(n_1, n_2, \ldots, n_K)$, service rates are given by solution to corresponding $\alpha$-fair utility maximization problem
Bandwidth-sharing networks: Stability cond’n

Stability condition: \( \sum_{k \in \mathcal{K}_l} \rho_k < C_l \) for all \( l \in \mathcal{L} \)

- Evidently necessary
- Sufficient in case of \( \alpha \)-fair policies with \( \alpha > 0 \) and exponentially distributed service requirements (proof sketch based on fluid limits provided later)
- Widely believed to hold for generally distributed service requirements, but proof has remained elusive
- Assumption \( \alpha > 0 \) is essential!
Bandwidth-sharing networks: Two-node three-class example (cont’d)

In case $\alpha = 0$, classes 1 and 2 receive strict priority over class 0, and thus

$$\mathbb{P}\{N_1 = n_1, N_2 = n_2\} = \mathbb{P}\{N_1 = n_1\}\mathbb{P}\{N_2 = n_2\} = (1 - \rho_1)^n_1 (1 - \rho_2)^n_2.$$

In particular, fraction of time that is available for service of class 0 is

$$\tau_0 = \mathbb{P}\{N_1 = 0, N_2 = 0\} = (1 - \rho_1)(1 - \rho_2).$$
Bandwidth-sharing networks: Two-node three-class example (cont’d)

In order for class 0 to be stable, we need $\rho_0 < \tau_0$, i.e.,

$$\rho_0 < (1 - \rho_1)(1 - \rho_2),$$

which is more stringent than sufficient condition in case $\alpha > 0$:

$$\rho_0 < \min\{1 - \rho_1, 1 - \rho_2\}.$$

Choosing $\alpha = 0$ and maximizing aggregate instantaneous rate at all times by allocating all service resources to classes 1 and 2 seems a priori reasonable, if not optimal. Implementing such priority policies in networks with user dynamics can however be recipe for severe trouble!
Bandwidth-sharing networks: Stability proof

We will sketch a stability proof based on fluid limits for a scenario with an $\alpha$-fair policy, $\alpha > 0$, and exponentially distributed service requirements.

It may be shown that the fluid limit $\{\bar{Q}(t)\}_{t \geq 0}$ satisfies the equations

$$\frac{d}{dt} \bar{Q}_k(t) = \lambda_k - \mu_k R_k(\bar{Q}(t)),$$

as long as $\bar{Q}_k(t) > 0$, where $R(q) = (R_k(q))_{k \in \mathcal{K}}$ is the optimal solution to the rate allocation problem

$$\max \quad G(R_1, \ldots, R_k) = \sum_{k \in \mathcal{K}} q_k^\alpha \frac{R_k^{1-\alpha}}{1-\alpha}$$

subject to

$$\sum_{k \in \mathcal{K}_l} R_k \leq C_l \quad l \in \mathcal{L}$$

$$R_k \geq 0 \quad k \in \mathcal{K}$$
Bandwidth-sharing networks: Stability proof (cont’d)

Define the Lyapunov function $F(t) = H(\bar{Q}(t))$, with

$$H(q) = \sum_{k \in K} \frac{1}{\mu_k} \frac{\rho_k^{-\alpha} q_k^{\alpha+1}}{\alpha + 1},$$

then

$$\frac{d}{dt} F(t) = \sum_{k \in K} \frac{\partial H(q)}{\partial q_k} \bigg|_{q_k = \bar{Q}_k(t)} \frac{d}{dt} \bar{Q}_k(t)$$

$$= \sum_{k \in K} \frac{1}{\mu_k} \rho_k^{-\alpha} (\bar{Q}_k(t))^{\alpha} (\lambda_k - \mu_k R_k(\bar{Q}(t)))$$

$$= \sum_{k \in K} \rho_k^{-\alpha} (\bar{Q}_k(t))^{\alpha} (\rho_k - R_k(\bar{Q}(t)))$$

$$(1)$$
Bandwidth-sharing networks: Stability proof (cont’d)

The inequalities \( \sum_{k \in K_l} \rho_k < C_l \) for all \( l = 1, \ldots, L \) imply that there exists an \( \epsilon > 0 \) such that \( \sum_{k \in K_l} (1 + \epsilon) \rho_k < C_l \) for all \( l = 1, \ldots, L \).

Because of concavity of the function \( G(\cdot) \) and optimality of \( R(q) = (R_k(q))_{k \in K}, \) we have

\[
G'(r)(r - R(q)) \leq 0
\]

for all \( r \) that satisfy the capacity constraints \( \sum_{k \in K_l} R_k \leq C_l, l \in \mathcal{L}, \) and in particular for \( r = (1 + \epsilon)(\rho_k)_{k \in K}, \) yielding

\[
\sum_{k \in K} \rho_k^{-\alpha} q_k^\alpha ((1 + \epsilon) \rho_k - R_k(q)) \leq 0 \tag{2}
\]
Bandwidth-sharing networks: Stability proof (cont’d)

Substitution of (2) into (1) yields

\[
\frac{d}{dt} F(t) \leq -\epsilon \sum_{k \in \mathcal{K}} \rho_k^{1-\alpha} (\bar{Q}_k(t))^\alpha
\]

It may then be shown that there exists \( \tau < \infty \) such that \( F(t) = 0 \), and hence \( \bar{Q}(t) = 0 \) for all \( t \geq \tau \), which implies that the original stochastic process \( \{Q(t)\}_{t \geq 0} \) is positive-recurrent
Bandwidth-sharing networks: Stationary distr’n

While stability condition is simple and intuitive, joint distribution of numbers of users of various classes \((N_1, N_2, \ldots, N_K)\) does not seem to be tractable in general.

Even mean numbers of users and mean sojourn times do not seem tractable for general topologies and \(\alpha\)-values.

Only significant exception to the rule is for Proportional Fair rate allocation \((\alpha = 1)\) and linear hypercube topologies.
Bandwidth-sharing networks: Linear topologies

Network consists of $L$ unit-capacity links and $L + 1$ classes of users:

- Class-$l$ users require service from link $l$ only, $l = 1, \ldots, L$
- Class-0 users require same service rate from all $L$ links simultaneously
Linear topologies with Proportional Fair rate allocation

When population of users is \((n_0, n_1, \ldots, n_L)\), weighted \(\alpha\)-fair rate allocation is:

\[ n_0 r_0 = \frac{(w_0 n_0^\alpha)^{1/\alpha}}{(w_0 n_0^\alpha)^{1/\alpha} + \left(\sum_{l=1}^{L} w_l n_l^\alpha\right)^{1/\alpha}}, \]

and

\[ n_l r_l = 1 - n_0 r_0, \quad n_l \geq 1, \quad l = 1, \ldots, L, \]

so Proportional Fair rate allocation (\(\alpha = 1, w_k \equiv 1\)) is:

\[ r_0 = \frac{1}{\sum_{k=0}^{L} n_k}, \quad n_0 \geq 1, \]

and

\[ r_l = \frac{1}{n_l} \frac{\sum_{k=1}^{L} n_k}{n_l \sum_{k=0}^{L} n_k}, \quad n_l \geq 1, \quad l = 1, \ldots, L. \]
Linear topologies with Proportional Fair rate allocation (cont’d)

$$
\mathbb{P}\{N_0 = n_0, N_1 = n_1, \ldots, N_L = n_L\} = Z \left( \sum_{k=0}^{L} \frac{n_k}{n_0} \right) \prod_{k=0}^{L} \rho_k^{n_k},
$$

where normalization constant is

$$
Z = \frac{\prod_{l=1}^{L} (1 - \rho_0 - \rho_l)}{(1 - \rho_0)^{L-1}}.
$$

Mean number of class-0 users is

$$
\mathbb{E}\{N_0\} = \frac{\rho_0}{1 - \rho_0} \left( 1 + \sum_{l=1}^{L} \frac{\rho_0}{1 - \rho_0 - \rho_l} \right),
$$

while mean number of class-\(l\) users, \(l = 1, \ldots, L\), is

$$
\mathbb{E}\{N_l\} = \frac{\rho_l}{1 - \rho_0 - \rho_l}.
$$
References

References (cont’d)


