Analysis of Call Overflow:

Many-Server Approximations and Implications to Call-Center Outsourcing

Itai Gurvich and Ohad Perry

September, 2010

Related Literature

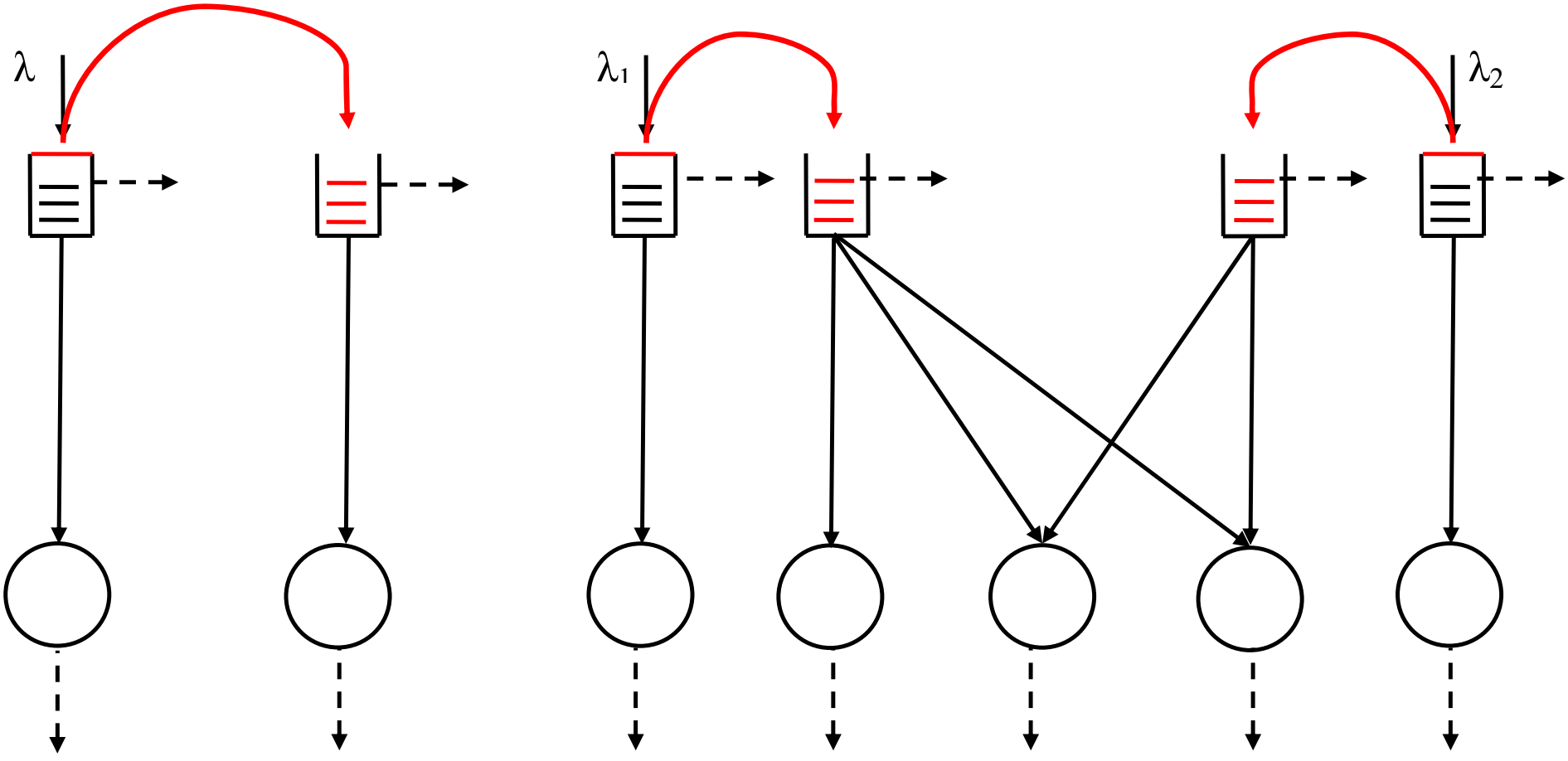
Blocking and overflow:

- Exact characterization: Van Doorn (’83);
- Approximations: Whitt (’83), Koole et. al (’00,’05);
- Heavy Traffic: Hunt and Kurtz (’94), Koçaga and Ward (’10), Pang et. al (’07), Whitt (’04);

Technical:

- Whitt (’91), Bassamboo et. al (’05), P’ and Whitt (’10a);
- Glynn and Whitt (’93), P’ and Whitt (’10b);

Call Centers with Overflow – 2 Examples

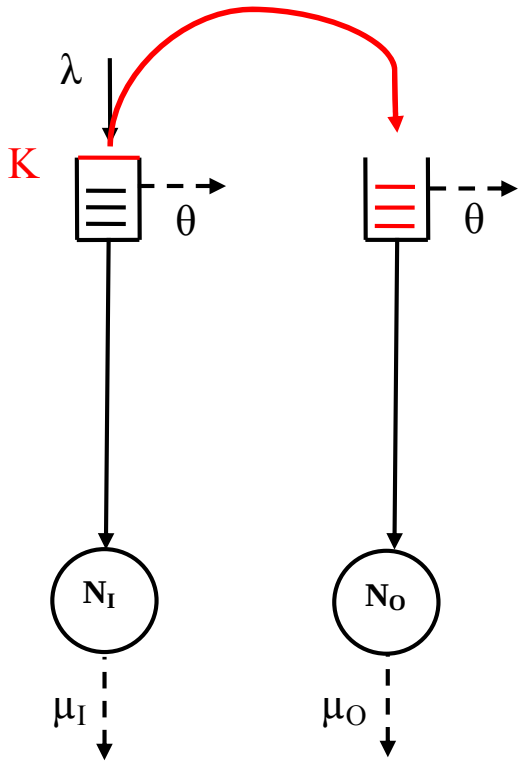


Basic Model

- $A(t)$ - number of arrivals to pool I by time t:

  $A(t)$ is a Poisson process with rate λ.

- $A_O(t)$ - number of overflowed calls by time t.
- $A_I(t) = A(t) - A_O(t)$ - arrivals **entering** pool I by t.
- $X_I(t), X_O(t)$ - total number in respective system at t.
- K - threshold in pool I ($K \geq 0$).

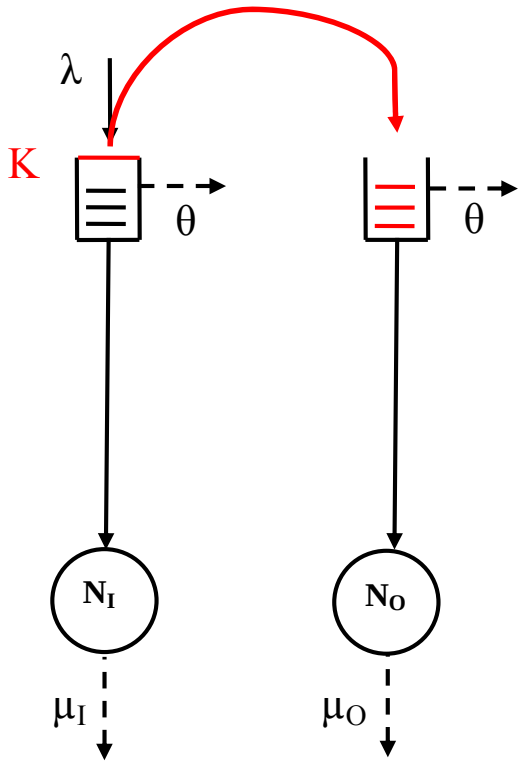


Basic Model

- $A(t)$ - number of arrivals to pool I by time t:

  $A(t)$ is a Poisson process with rate λ.

- $A_O(t)$ - number of overflowed calls by time t.
- $A_I(t) = A(t) - A_O(t)$ - arrivals **entering** pool I by t.
- $X_I(t), X_O(t)$ - total number in respective system at t.
- K - threshold in pool I ($K \geq 0$).



The two pools are dependent !

A Motivating Example

$C_s^I(N_I)$ and $C_s^O(N_O)$ are capacity cost functions for pools I and O, respectively.

w_k = waiting time of the k^{th} arriving customer by time T.

A Motivating Example

$C_s^I(N_I)$ and $C_s^O(N_O)$ are capacity cost functions for pools I and O, respectively.

w_k = waiting time of the k^{th} arriving customer by time T.

A Centralized Optimization Problem:

$$\begin{aligned} \min_{(N_I,N_O,K)} \quad & C_s^I(N_I) + C_s^O(N_O) \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{1}{A(T)}\sum_{k=1}^{A(T)} \mathbb{1}\left\{w_k > \tau\right\}\right] \leq \alpha, \\ & N_I, N_O, K \in \mathbb{Z}_+, \end{aligned}$$

A Motivating Example

$C_s^I(N_I)$ and $C_s^O(N_O)$ are capacity cost functions for pools I and O, respectively.

w_k = waiting time of the k^{th} arriving customer by time T.

A Centralized Optimization Problem:

$$\begin{aligned} \min_{(N_I,N_O,K)} \quad & C_s^I(N_I) + C_s^O(N_O) \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{1}{A(T)}\sum_{k=1}^{A(T)} \mathbb{1}\left\{w_k > \tau\right\}\right] \leq \alpha, \\ & N_I, N_O, K \in \mathbb{Z}_+, \end{aligned}$$

- Alternatively: constraints on the virtual (actual) waiting time $W(t)$.

A Motivating Example

$C_s^I(N_I)$ and $C_s^O(N_O)$ are capacity cost functions for pools I and O, respectively.

w_k = waiting time of the k^{th} arriving customer by time T.

A Centralized Optimization Problem:

$$\begin{aligned} \min_{(N_I,N_O,K)} \quad & C_s^I(N_I) + C_s^O(N_O) \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{1}{A(T)}\sum_{k=1}^{A(T)} \mathbb{1}\{w_k > \tau\}\right] \leq \alpha, \\ & N_I, N_O, K \in \mathbb{Z}_+, \end{aligned}$$

- Alternatively: constraints on the virtual (actual) waiting time $W(t)$.
- The centralized problem considers **all** customers, overflowed or not.

A Motivating Example

$C_s^I(N_I)$ and $C_s^O(N_O)$ are capacity cost functions for pools I and O, respectively.

w_k = waiting time of the k^{th} arriving customer by time T.

A Centralized Optimization Problem:

$$\begin{aligned} \min_{(N_I,N_O,K)} \quad & C_s^I(N_I) + C_s^O(N_O) \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{1}{A(T)}\sum_{k=1}^{A(T)} \mathbb{1}\left\{w_k > \tau\right\}\right] \leq \alpha, \\ & N_I, N_O, K \in \mathbb{Z}_+, \end{aligned}$$

- Alternatively: constraints on the virtual (actual) waiting time $W(t)$.
- The centralized problem considers **all** customers, overflowed or not.
- However, the two pools are operated by two distinct controllers!

Main Results

1. The two systems are asymptotically independent.
2. Simplify the complicated overflow process by a FCLT.

Main Results

1. The two systems are asymptotically independent.
2. Simplify the complicated overflow process by a FCLT.

Implications of (1):

both systems can optimize hierarchically (not jointly) to find optimal N_I, K and N_O.

Main Results

1. The two systems are asymptotically independent.
2. Simplify the complicated overflow process by a FCLT.

Implications of (1):

both systems can optimize hierarchically (not jointly) to find optimal N_I, K and N_O.

Implications of (2):

outsourcer can determine its staffing and routing, so that guaranteed QoS are met.

Asymptotic (Heavy Traffic) Analysis

We consider a sequence indexed by arrival rate λ, with $\lambda \to \infty$.

Assumption (Resource Pooling)

Non-negligible overflow: $\nu := \lim_{\lambda \to \infty} \dfrac{\mu_I N_I^\lambda + \theta K^\lambda}{\lambda} < 1$

Asymptotic (Heavy Traffic) Analysis

We consider a sequence indexed by arrival rate λ, with $\lambda \to \infty$.

Assumption (Resource Pooling)

Non-negligible overflow: $\nu := \lim_{\lambda\to\infty} \dfrac{\mu_I N_I^\lambda + \theta K^\lambda}{\lambda} < 1$

$$\hat{X}_I^\lambda(t) := \frac{X_I^\lambda(t) - N_I^\lambda - K^\lambda}{\sqrt{\lambda}}, \qquad \hat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K^\lambda)t}{\sqrt{\lambda}}.$$

Asymptotic (Heavy Traffic) Analysis

We consider a sequence indexed by arrival rate λ, with $\lambda \to \infty$.

Assumption (Resource Pooling)

Non-negligible overflow: $\nu := \lim_{\lambda\to\infty} \dfrac{\mu_I N_I^\lambda + \theta K^\lambda}{\lambda} < 1$

$$\hat{X}_I^\lambda(t) := \frac{X_I^\lambda(t) - N_I^\lambda - K^\lambda}{\sqrt{\lambda}}, \qquad \hat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K^\lambda)t}{\sqrt{\lambda}}.$$

Theorem

If $(\hat{X}^\lambda(0), \hat{A}_O^\lambda(0)) \Rightarrow (0,0)$, *then* $(\hat{X}^\lambda, \hat{A}_O^\lambda) \Rightarrow (0, \sigma B)$, *u.o.c., where* B *is a standard Brownian motion and* $\sigma^2 = 1 + \nu$.

Outline of the Proof

Proof for $\hat{X}_I^\lambda$:

$D^\lambda(t) := N_I^\lambda + K - X_I^\lambda(t)$ is "close" to a $M/M/1$ with arrival rate $\mu_I N_I^\lambda + \theta K^\lambda$ and service rate λ.

Outline of the Proof

Proof for $\hat{X}_I^\lambda$:

$D^\lambda(t) := N_I^\lambda + K - X_I^\lambda(t)$ is "close" to a $M/M/1$ with arrival rate $\mu_I N_I^\lambda + \theta K^\lambda$ and service rate λ.

- Let $Q_b(t)$ be $M/M/1$ with arrival rate ν and service rate 1. Then,

$$\{D^\lambda(t) : t \in [0,T) \approx \{Q_b(t) : t \in [0,\lambda T)\}.$$

Outline of the Proof

Proof for $\hat{X}_I^\lambda$:

$D^\lambda(t) := N_I^\lambda + K - X_I^\lambda(t)$ is "close" to a $M/M/1$ with arrival rate $\mu_I N_I^\lambda + \theta K^\lambda$ and service rate λ.

- Let $Q_b(t)$ be $M/M/1$ with arrival rate ν and service rate 1. Then,

$$\{D^\lambda(t) : t \in [0,T) \approx \{Q_b(t) : t \in [0,\lambda T)\}.$$

- apply established extreme-value theory for $M/M/1$.

Outline of the Proof

Proof for $\hat{X}_I^\lambda$:

$D^\lambda(t) := N_I^\lambda + K - X_I^\lambda(t)$ is "close" to a $M/M/1$ with arrival rate $\mu_I N_I^\lambda + \theta K^\lambda$ and service rate λ.

- Let $Q_b(t)$ be $M/M/1$ with arrival rate ν and service rate 1. Then,

$$\{D^\lambda(t) : t \in [0,T) \approx \{Q_b(t) : t \in [0,\lambda T)\}.$$

- apply established extreme-value theory for $M/M/1$.

Proof for $\hat{A}_O^\lambda$:

FCLT for the cumulative processes $\int_0^t \mathbb{1}\{D^\lambda(s) = 0\}ds$.

Outline of the Proof

Proof for $\hat{X}_I^\lambda$:

$D^\lambda(t) := N_I^\lambda + K - X_I^\lambda(t)$ is "close" to a $M/M/1$ with arrival rate $\mu_I N_I^\lambda + \theta K^\lambda$ and service rate λ.

- Let $Q_b(t)$ be $M/M/1$ with arrival rate ν and service rate 1. Then,

$$\{D^\lambda(t) : t \in [0,T) \approx \{Q_b(t) : t \in [0,\lambda T)\}.$$

- apply established extreme-value theory for $M/M/1$.

Proof for $\hat{A}_O^\lambda$:

FCLT for the cumulative processes $\int_0^t \mathbb{1}\{D^\lambda(s) = 0\}ds$.

D^λ completes $O(\lambda)$ cycles over $[0,t)$, for all $t > 0$.

Implications

- Approximating (complicated) overflow process with a simple process:

$$A_O^\lambda(t) \approx (\lambda - \mu_I N_I^\lambda - \theta K^\lambda)t + \sqrt{\lambda}\sigma B(t),$$

with the approximation being asymptotically exact.

Implications

- Approximating (complicated) overflow process with a simple process:

$$A_O^\lambda(t) \approx (\lambda - \mu_I N_I^\lambda - \theta K^\lambda)t + \sqrt{\lambda}\sigma B(t),$$

  with the approximation being asymptotically exact.

- Note that for each λ,

$$\lim_{t\to\infty} \frac{A_O^\lambda(t)}{t} = \lambda \mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K^\lambda\}.$$

Implications

- Approximating (complicated) overflow process with a simple process:

$$A_O^\lambda(t) \approx (\lambda - \mu_I N_I^\lambda - \theta K^\lambda)t + \sqrt{\lambda}\sigma B(t),$$

with the approximation being asymptotically exact.

- Note that for each λ,

$$\lim_{t\to\infty} \frac{A_O^\lambda(t)}{t} = \lambda \mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K^\lambda\}.$$

We get the following local steady-state result:

$$\frac{A_O^\lambda(t)}{t} = \lambda \mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K^\lambda\} + O(\sqrt{\lambda}) \text{ for } \textbf{each } t > 0.$$

Implications Cont' – Simple Independence Result

Corollary (Independence in the Limit) (trivial!)

The limits $\hat{X}_I$ and $\hat{A}_O$ are independent.

Implications Cont' – Simple Independence Result

Corollary (Independence in the Limit) (trivial!)

The limits $\hat{X}_I$ and $\hat{A}_O$ are independent.

What does this independence mean for the pre-limits?

Implications Cont' – Simple Independence Result

Corollary (Independence in the Limit) (trivial!)

The limits $\hat{X}_I$ and $\hat{A}_O$ are independent.

What does this independence mean for the pre-limits? ... not much...

Implications Cont' – Simple Independence Result

Corollary (Independence in the Limit) (trivial!)

The limits $\hat{X}_I$ and $\hat{A}_O$ are independent.

What does this independence mean for the pre-limits? ... not much...

$$\mathbb{P}\{W^\lambda(t) > \tau\} = \mathbb{P}\{W^\lambda(t) > \tau, X_I^\lambda(t) < N_I^\lambda + K^\lambda\}$$
$$+ \mathbb{P}\{W^\lambda(t) > \tau, X_I^\lambda(t) = N_I^\lambda + K^\lambda\}.$$

Implications Cont' – Simple Independence Result

Corollary (Independence in the Limit) (trivial!)

The limits $\hat{X}_I$ and $\hat{A}_O$ are independent.

What does this independence mean for the pre-limits? ... not much...

$$\mathbb{P}\{W^\lambda(t) > \tau\} = \mathbb{P}\{W^\lambda(t) > \tau, X_I^\lambda(t) < N_I^\lambda + K^\lambda\}$$
$$+ \mathbb{P}\{W^\lambda(t) > \tau, X_I^\lambda(t) = N_I^\lambda + K^\lambda\}.$$

Need to consider $N_I^\lambda + K^\lambda - X_I^\lambda(t)$ in its natural scale (order $O(1)$).

Independence of the Limits

Independence of limits does not “carry over” to the pre-limits.

Independence of the Limits

Independence of limits does not "carry over" to the pre-limits.

Example:

$$Y^\lambda := \begin{cases} 1/\sqrt{\lambda}, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases} \qquad X^\lambda := \begin{cases} 1, & \text{if } Y^\lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

Independence of the Limits

Independence of limits does not "carry over" to the pre-limits.

Example:

$$Y^\lambda := \begin{cases} 1/\sqrt{\lambda}, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases} \qquad X^\lambda := \begin{cases} 1, & \text{if } Y^\lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$(Y^\lambda, X^\lambda) \Rightarrow (0, X), \quad \text{where } X = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$$

Independence of the Limits

Independence of limits does not "carry over" to the pre-limits.

Example:

$$Y^\lambda := \begin{cases} 1/\sqrt{\lambda}, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases} \qquad X^\lambda := \begin{cases} 1, & \text{if } Y^\lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$(Y^\lambda, X^\lambda) \Rightarrow (0, X), \qquad \text{where } X = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$$

Trivially, the limits 0 and X are independent.

Independence of the Limits

Independence of limits does not "carry over" to the pre-limits.

Example:

$$Y^\lambda := \begin{cases} 1/\sqrt{\lambda}, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases} \qquad X^\lambda := \begin{cases} 1, & \text{if } Y^\lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$(Y^\lambda, X^\lambda) \Rightarrow (0, X), \qquad \text{where } X = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$$

Trivially, the limits 0 and X are independent. However,

$$1/2 = \mathbb{P}\{X^\lambda > 0, Y^\lambda > 0\} \neq \mathbb{P}\{X^\lambda > 0\}\mathbb{P}\{Y^\lambda > 0\} = 1/4,$$

for all λ, no matter how large.

Asymptotic Independence

We want the dependency to "fade away" as λ grows.

Definition

$\{X^\lambda : \lambda \geq 1\}$ *and* $\{Y^\lambda : \lambda \geq 1\}$ *are asymptotically independent if*

$$\mathbb{P}\{X^\lambda > x, Y^\lambda > y\} = \mathbb{P}\{X^\lambda > x\}\mathbb{P}\{Y^\lambda > y\} + o(1).$$

Asymptotic Independence

We want the dependency to "fade away" as λ grows.

Definition

$\{X^\lambda : \lambda \geq 1\}$ *and* $\{Y^\lambda : \lambda \geq 1\}$ *are asymptotically independent if*

$$\mathbb{P}\{X^\lambda > x, Y^\lambda > y\} = \mathbb{P}\{X^\lambda > x\}\mathbb{P}\{Y^\lambda > y\} + o(1).$$

We can generalize the easy corollary (independence of the limits):

Asymptotic Independence

We want the dependency to "fade away" as λ grows.

Definition

$\{X^\lambda : \lambda \geq 1\}$ *and* $\{Y^\lambda : \lambda \geq 1\}$ *are asymptotically independent if*

$$\mathbb{P}\{X^\lambda > x, Y^\lambda > y\} = \mathbb{P}\{X^\lambda > x\}\mathbb{P}\{Y^\lambda > y\} + o(1).$$

We can generalize the easy corollary (independence of the limits):

Theorem (asymptotic independence) (not trivial)

$X_I^\lambda(t)$ *is asymptotically independent of* $\hat{A}_O^\lambda(t)$.

Asymptotic Independence

We want the dependency to "fade away" as λ grows.

Definition

$\{X^\lambda : \lambda \geq 1\}$ *and* $\{Y^\lambda : \lambda \geq 1\}$ *are asymptotically independent if*

$$\mathbb{P}\{X^\lambda > x, Y^\lambda > y\} = \mathbb{P}\{X^\lambda > x\}\mathbb{P}\{Y^\lambda > y\} + o(1).$$

We can generalize the easy corollary (independence of the limits):

Theorem (asymptotic independence) (not trivial)

$X_I^\lambda(t)$ *is asymptotically independent of* $\hat{A}_O^\lambda(t)$.

Note that $\hat{A}_O$ is scaled, but X_I^λ is not (requires refined analysis).

Asymptotic Independence

We want the dependency to "fade away" as λ grows.

Definition

$\{X^\lambda : \lambda \geq 1\}$ *and* $\{Y^\lambda : \lambda \geq 1\}$ *are asymptotically independent if*

$$\mathbb{P}\{X^\lambda > x, Y^\lambda > y\} = \mathbb{P}\{X^\lambda > x\}\mathbb{P}\{Y^\lambda > y\} + o(1).$$

We can generalize the easy corollary (independence of the limits):

Theorem (asymptotic independence) (not trivial)

$X_I^\lambda(t)$ *is asymptotically independent of* $\hat{A}_O^\lambda(t)$.

Note that $\hat{A}_O$ is scaled, but X_I^λ is not (requires refined analysis).

Main difficulty: establishing HT limits when X_I^λ unscaled.

Main Idea of the Proof

Showing asymptotic independence of the sequence via asymptotic independence of a process.

Main Idea of the Proof

Showing asymptotic independence of the sequence via asymptotic independence of a process.

The relevant state of $X_I^\lambda(t)$ with respect to $A_O^\lambda(t)$ is the availability process

$$D^\lambda(t) := N_I^\lambda + K^\lambda - X_I^\lambda(t). \qquad (O(1) \text{ process!})$$

Main Idea of the Proof

Showing asymptotic independence of the sequence via asymptotic independence of a process.

The relevant state of $X_I^\lambda(t)$ with respect to $A_O^\lambda(t)$ is the availability process

$$D^\lambda(t) := N_I^\lambda + K^\lambda - X_I^\lambda(t). \qquad (O(1) \text{ process!})$$

(*) Recall that $\{D^\lambda(s) : t \le s \le t+\epsilon\} \approx \{Q_b(s) : t \le s \le t+\lambda\epsilon\}$ for λ large, with Q_b denoting a $M/M/1$.

Main Idea of the Proof

Showing asymptotic independence of the sequence via asymptotic independence of a process.

The relevant state of $X_I^\lambda(t)$ with respect to $A_O^\lambda(t)$ is the availability process

$$D^\lambda(t) := N_I^\lambda + K^\lambda - X_I^\lambda(t). \qquad (O(1) \text{ process!})$$

(*) Recall that $\{D^\lambda(s) : t \le s \le t+\epsilon\} \approx \{Q_b(s) : t \le s \le t+\lambda\epsilon\}$ for λ large, with Q_b denoting a $M/M/1$.

(**) $Q_b(t+\lambda\epsilon) \Rightarrow Q_b(\infty)$ as $\lambda \to \infty$ for all $\epsilon > 0$.

Main Idea of the Proof

Showing asymptotic independence of the sequence via asymptotic independence of a process.

The relevant state of $X_I^\lambda(t)$ with respect to $A_O^\lambda(t)$ is the availability process

$$D^\lambda(t) := N_I^\lambda + K^\lambda - X_I^\lambda(t). \qquad (O(1) \text{ process!})$$

(*) Recall that $\{D^\lambda(s) : t \le s \le t+\epsilon\} \approx \{Q_b(s) : t \le s \le t+\lambda\epsilon\}$ for λ large, with Q_b denoting a $M/M/1$.

(**) $Q_b(t+\lambda\epsilon) \Rightarrow Q_b(\infty)$ as $\lambda \to \infty$ for all $\epsilon > 0$.

Proof follows since the steady state $Q_b(\infty)$ is independent of $Q_b(t)$.

A Pointwise Averaging Principle (AP)

The following pointwise AP "follows" from (*) and (**):

Theorem (pointwise AP)

$D^{\lambda}(t) \Rightarrow Q_b(\infty)$ *in* $\mathbb{R}$ *as* $\lambda \to \infty$.

A Pointwise Averaging Principle (AP)

The following pointwise AP "follows" from (*) and (**):

Theorem (pointwise AP)

$D^{\lambda}(t) \Rightarrow Q_b(\infty)$ *in* $\mathbb{R}$ *as* $\lambda \to \infty$.

Implications of Asymptotic Independence and AP to waiting times

A Pointwise Averaging Principle (AP)

The following pointwise AP "follows" from (*) and (**):

Theorem (pointwise AP)

$D^\lambda(t) \Rightarrow Q_b(\infty)$ *in* $\mathbb{R}$ *as* $\lambda \to \infty$.

Implications of Asymptotic Independence and AP to waiting times

- Let W_I^λ and W_O^λ be virtual waiting times in respective pool.

A Pointwise Averaging Principle (AP)

The following pointwise AP "follows" from (*) and (**):

Theorem (pointwise AP)

$D^\lambda(t) \Rightarrow Q_b(\infty)$ *in* $\mathbb{R}$ *as* $\lambda \to \infty$.

Implications of Asymptotic Independence and AP to waiting times

- Let W_I^λ and W_O^λ be virtual waiting times in respective pool.
- Let $p_b^\lambda := \mathbb{P}\{X_I^\lambda(t) = N_I^\lambda + K^\lambda\} \approx \mathbb{P}\{Q_b(\infty) = 0\} = 1 - \nu$.

A Pointwise Averaging Principle (AP)

The following pointwise AP "follows" from (*) and (**):

Theorem (pointwise AP)

$D^\lambda(t) \Rightarrow Q_b(\infty)$ *in* $\mathbb{R}$ *as* $\lambda \to \infty$.

Implications of Asymptotic Independence and AP to waiting times

- Let W_I^λ and W_O^λ be virtual waiting times in respective pool.
- Let $p_b^\lambda := \mathbb{P}\{X_I^\lambda(t) = N_I^\lambda + K^\lambda\} \approx \mathbb{P}\{Q_b(\infty) = 0\} = 1 - \nu$.

$$\begin{aligned}\mathbb{P}\{W^\lambda(t) > \tau\} &= \mathbb{P}\{W_I^\lambda(t) > \tau \mid X_I^\lambda(t) < N_I^\lambda + K^\lambda\}\mathbb{P}\{X_I^\lambda(t) < N_I^\lambda + K^\lambda\} \\ &\quad + \mathbb{P}\{W_O^\lambda(t) > \tau \mid X_I^\lambda(t) = N_I^\lambda + K^\lambda\}\mathbb{P}\{X_I^\lambda(t) = N_I^\lambda + K^\lambda\} \\ &= \mathbb{P}\{W_I^\lambda(t) > \tau\}(1 - p_b^\lambda) + \mathbb{P}\{W_O^\lambda(t) > \tau\}p_b^\lambda + o(1)\end{aligned}$$

Waiting Times and Asymptotic ASTA

w_k^λ, $w_{I,k}^\lambda$, $w_{O,k}^\lambda$ - waiting time of k^{th} arrival to respective pool.

f is a continuous and bounded function or, e.g., $f(x) := \mathbb{1}\{x > \tau\}$.

$$\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)} f(w_k^\lambda)\right] = (1-p_b^\lambda)\mathbb{E}\left[\frac{1}{A_I^\lambda(t)}\sum_{k=1}^{A_I^\lambda(t)} f(w_{I,k}^\lambda)\right] + p_b^\lambda \mathbb{E}\left[\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(w_{O,k}^\lambda)\right] + o(1).$$

Waiting Times and Asymptotic ASTA

w_k^λ, $w_{I,k}^\lambda$, $w_{O,k}^\lambda$ - waiting time of k^{th} arrival to respective pool.

f is a continuous and bounded function or, e.g., $f(x) := \mathbb{1}\{x > \tau\}$.

$$\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)} f(w_k^\lambda)\right] = (1-p_b^\lambda)\mathbb{E}\left[\frac{1}{A_I^\lambda(t)}\sum_{k=1}^{A_I^\lambda(t)} f(w_{I,k}^\lambda)\right] + p_b^\lambda \mathbb{E}\left[\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)} f(w_{O,k}^\lambda)\right] + o(1).$$

Theorem (asymptotic finite-horizon ASTA)

For all $t > 0$,

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)} f(w_k^\lambda)\right] = \nu\frac{1}{t}\int_0^t \mathbb{E}\left[f(\widehat{W}_I(s))\right] ds + (1-\nu)\frac{1}{t}\int_0^t \mathbb{E}\left[f(\widehat{W}_O(s))ds\right].$$

where $\widehat{W}_O(t)$ is the diffusion limit of the virtual waiting-time process in the $GI/M/N+M$ queue and $\widehat{W}_I(t) = \bar{K}$.

Generalizing to SBR

- Several overflow processes.
- Possibly several generalist pools.

Generalizing to SBR

- Several overflow processes.
- Possibly several generalist pools.
- If asymptotic independence holds, then we can treat outsourcer as independent system with renewal arrivals.

Generalizing to SBR

- Several overflow processes.
- Possibly several generalist pools.
- If asymptotic independence holds, then we can treat outsourcer as independent system with renewal arrivals.
- Technical assumption for HT limits: all queues are C-tight.

Generalizing to SBR

- Several overflow processes.
- Possibly several generalist pools.
- If asymptotic independence holds, then we can treat outsourcer as independent system with renewal arrivals.
- Technical assumption for HT limits: all queues are C-tight.
- In particular, if continuous limits exist, e.g., QIR controls in Gurvich and Whitt (07).

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.
- Under a resource pooling condition our heavy traffic analysis:

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.
- Under a resource pooling condition our heavy traffic analysis:
  - provides simple approximations for the overflow renewal process, which are asymptotically correct.
  - proves that $M/M/N_I/K+M$ is asymptotically independent of $G/M/N_O+M$.

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.
- Under a resource pooling condition our heavy traffic analysis:
  - provides simple approximations for the overflow renewal process, which are asymptotically correct.
  - proves that $M/M/N_I/K+M$ is asymptotically independent of $G/M/N_O+M$.
- Proofs build on a separation of time scales and a resulting pointwise AP.

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.
- Under a resource pooling condition our heavy traffic analysis:
  - provides simple approximations for the overflow renewal process, which are asymptotically correct.
  - proves that $M/M/N_I/K+M$ is asymptotically independent of $G/M/N_O+M$.
- Proofs build on a separation of time scales and a resulting pointwise AP.
- Results are applied to waiting times and virtual waiting times.

Summary

- Motivated by an outsourcing problem, we considered an overflow system: from $M/M/N_I/K+M$ to $G/M/N_O+M$.
- Under a resource pooling condition our heavy traffic analysis:
  - provides simple approximations for the overflow renewal process, which are asymptotically correct.
  - proves that $M/M/N_I/K+M$ is asymptotically independent of $G/M/N_O+M$.
- Proofs build on a separation of time scales and a resulting pointwise AP.
- Results are applied to waiting times and virtual waiting times.
- Generalized to more complicated systems (if queues are C-tight).

Thank You