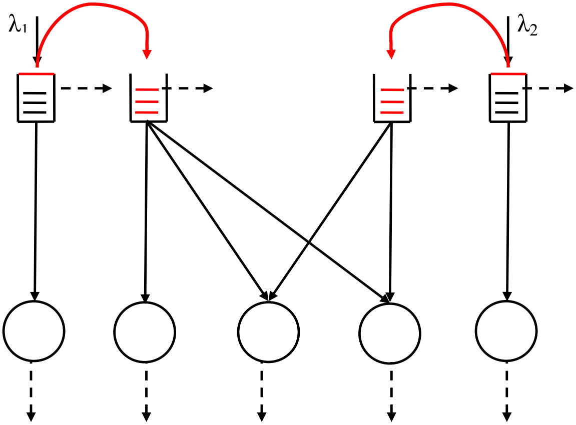
Overflow Networks: Approximations and Implications to Call-Center Outsourcing

Itai Gurvich (Northwestern University)

Joint work with Ohad Perry (CWI)

Call Centers with Overflow



Source of complexity: dependence

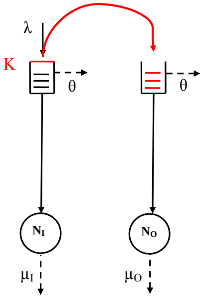
Motivating Example 1

$C_s^O(N_O)$ = capacity cost function for station O.

$W(t)$ = virtual waiting time at time t. (Similarly, $W_I(t)$ and $W_O(t)$)

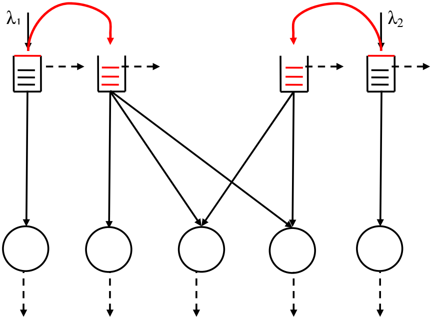
$$W(t) = W_I(t)\mathbb{1}\{X_I(t) < N_I + K_I\} + W_O(t)\mathbb{1}\{X_I(t) = N_I + K_I\}$$

$$\begin{aligned} \min_{N_O} \quad & C_s^O(N_O) \\ \text{s.t.} \quad & \mathbb{E}\left[f(W_O(t))\mathbb{1}\{X_I(t) = N_I + K_I\}\right] \leq \alpha \\ & N_O \in \mathbb{Z}_+, \end{aligned}$$



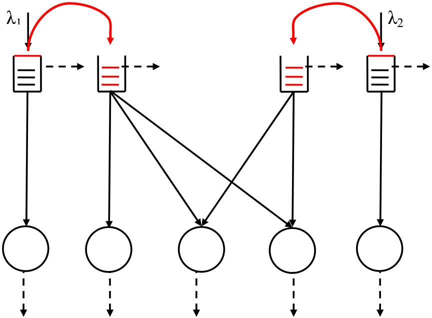
Motivating Example 2

$$\min \quad \sum_i \mathbb{E}[\int_0^T C_i(Q_i^\pi(s))ds]$$
$$\text{s.t.} \quad \pi \in \Pi.$$



Motivating Example 2

$$\begin{aligned} \min \quad & \sum_i \mathbb{E}[\int_0^T C_i(Q_i^\pi(s))ds] \\ \text{s.t.} \quad & \pi \in \Pi. \end{aligned}$$



- Optimal policy may benefit from state information on in-house

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);

Outsourcing and optimization:

- Gans and Zhou ('07); Chevalier et. al. ('03,'04);

Technical:

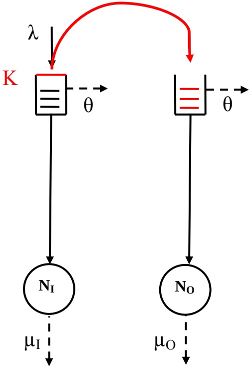
- Whitt ('91), Bassamboo et. al ('05), Perry and Whitt ('10a);
- Glynn and Whitt ('93), Perry and Whitt ('10b);

Basic Model

- $A(t)$ - number of arrivals to station 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** station I.
- $X_I(t), X_O(t)$ - total number in respective system at t.
- K_I - threshold in station I ($K_I \geq 0$).
- In isolation: station O is an $GI/M/N + M$ queue.



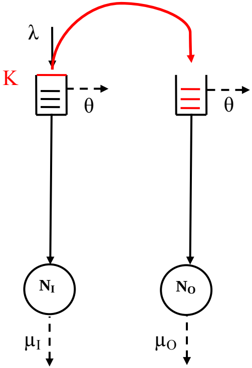
Main Results

A sequence of overflow networks in a many-server heavy-traffic regime

1. Functional Central Limit Theorem (FCLT) for Overflow Process
2. Pointwise Stationarity and Asymptotic Independence

Asymptotic (Heavy Traffic) Analysis

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



Main Assumption:

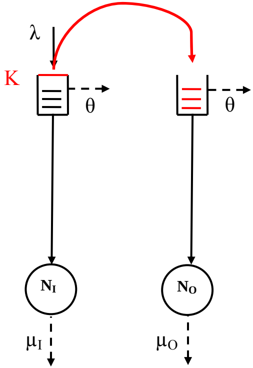
1. Non-negligible overflow:

$$\nu := \lim_{\lambda \to \infty} \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda} < 1$$

$1 - \nu$ interpreted as rough estimate for the (steady-state) blocking probability

Asymptotic (Heavy Traffic) Analysis

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



Main Assumption:

1. Non-negligible overflow:

$$\nu := \lim_{\lambda \to \infty} \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda} < 1$$

2. Sufficient capacity in station O:

$$N_O^\lambda = \frac{\lambda - \mu_I N_I^\lambda - \theta_I K_I^\lambda}{\mu_O} + o(\lambda)$$

$1 - \nu$ interpreted as rough estimate for the (steady-state) blocking probability

First Result

$$D_I^\lambda(t) = N_I^\lambda - K_I^\lambda - X_I^\lambda(t), \qquad \widehat{D}_I^\lambda(t) := \frac{D_I^\lambda(t)}{\sqrt{\lambda}},$$

$$\widehat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}}, \qquad \widehat{X}_O^\lambda(t) := \frac{X_O^\lambda(t) - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\mu_O}}{\sqrt{\lambda}}$$

First Result

$$D_I^\lambda(t) = N_I^\lambda - K_I^\lambda - X_I^\lambda(t), \qquad \widehat{D}_I^\lambda(t) := \frac{D_I^\lambda(t)}{\sqrt{\lambda}},$$

$$\widehat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}}, \qquad \widehat{X}_O^\lambda(t) := \frac{X_O^\lambda(t) - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\mu_O}}{\sqrt{\lambda}}$$

Theorem

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

First Result

$$D_I^\lambda(t) = N_I^\lambda - K_I^\lambda - X_I^\lambda(t), \qquad \widehat{D}_I^\lambda(t) := \frac{D_I^\lambda(t)}{\sqrt{\lambda}},$$

$$\widehat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}}, \qquad \widehat{X}_O^\lambda(t) := \frac{X_O^\lambda(t) - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\mu_O}}{\sqrt{\lambda}}$$

Theorem

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

Recall: $\nu := \lim_{\lambda\to\infty} \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda}$

Outline of the proof: $\widehat{D}_I^\lambda \Rightarrow 0$

$D_I^\lambda(t) := N_I^\lambda + K_I^\lambda - X_I^\lambda(t)$

- When close to threshold: $\uparrow$ rate $\approx \mu_I N_I^\lambda + \theta K_I^\lambda$, $\quad\downarrow$ rate $= \lambda$.
- Slowing $D_I^\lambda(t)$ down gives an $M/M/1$ with

$$\text{arrival rate } = \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda} \approx \nu < 1 \text{ and service rate } = \frac{\lambda}{\lambda} = 1$$

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

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

- From extreme-value theory for $M/M/1$: $\sup_{t \leq T} D_I^\lambda(t) = O(\log(\lambda T))$

Outline of the proof: $\widehat{A}_O^\lambda \Rightarrow \sigma B$

- $A_O^\lambda(t) = \int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} dA^\lambda(s)$
- D_I^λ completes $O(\lambda)$ cycles over any time interval $[s, t]$
- Functional limits for the cumulative processes

$$\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds \Rightarrow (1 - \nu)t$$

$$\sqrt{\lambda}\left(\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds - (1 - \nu)t\right) \Rightarrow \tilde{\sigma} B(t)$$

- Functional limit for $\widehat{A}_O^\lambda$ follows from that for $\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds$

Outline of the proof: $\widehat{A}_O^\lambda \Rightarrow \sigma B$

- $A_O^\lambda(t) = \int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} dA^\lambda(s)$
- D_I^λ completes $O(\lambda)$ cycles over any time interval $[s, t]$
- Functional limits for the cumulative processes

$$\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds \Rightarrow (1 - \nu)t$$

$$\sqrt{\lambda}\left(\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds - (1 - \nu)t\right) \Rightarrow \tilde{\sigma} B(t)$$

- Functional limit for $\widehat{A}_O^\lambda$ follows from that for $\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\} ds$

An Averaging Principle (AP):

$\widehat{A}_O^\lambda$ is "driven" by a process that moves at a different time scale

Implications

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

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

  with the approximation being asymptotically exact.

- $A_O^\lambda(t)$ is close to a renewal process with mean "inter-arrival" time $(\lambda - \mu_I N_I^\lambda - \theta_I K_I^\lambda)^{-1}$ and squared coefficient of variation (SCV)

$$\frac{\lambda\sigma^2}{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda} \approx \frac{\sigma^2}{(1-\nu)} \geq 1.$$

Implications

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

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

  with the approximation being asymptotically exact.

- $A_O^\lambda(t)$ is close to a renewal process with mean "inter-arrival" time $(\lambda - \mu_I N_I^\lambda - \theta_I K_I^\lambda)^{-1}$ and squared coefficient of variation (SCV)

$$\frac{\lambda\sigma^2}{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda} \approx \frac{\sigma^2}{(1-\nu)} \geq 1.$$

Simpler than the original overflow renewal process.

Implications Cont.

- In isolation, station O is a $GI/M/N+M$ queue.
- Using overflow convergence and known results for $GI/M/N+M$,

$$(\widehat{D}_I^\lambda, \widehat{X}_O^\lambda) \Rightarrow (0, \widehat{X}_O)$$

- Asymptotic independence in a trivialized sense

Implications Cont.

- In isolation, station O is a $GI/M/N+M$ queue.
- Using overflow convergence and known results for $GI/M/N+M$,

$$(\widehat{D}_I^\lambda, \widehat{X}_O^\lambda) \Rightarrow (0, \widehat{X}_O)$$

- Asymptotic independence in a trivialized sense

What does this limit imply for joint distributions? ... not much...

$$\mathbb{E}[W^\lambda(t)] = \mathbb{E}[W_I^\lambda(t)\mathbb{1}\{X_I^\lambda(t) < N_I^\lambda + K_I^\lambda\}] \\ + \mathbb{E}[W_O^\lambda(t)\mathbb{1}\{X_I^\lambda(t) = N_I^\lambda + K_I^\lambda\}].$$

Independence of the Limits

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

Example:

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

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

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

"Natural scale" of station I = constant

"Natural scale" of station $O = \sqrt{\lambda}$

Theorem (asymptotic independence)

D_I^λ is asymptotically independent of $\widehat{X}_O^\lambda$, i.e, for all $t > 0$,

$$\mathbb{P}\left\{D_I^\lambda(t) \geq x, \widehat{X}_O^\lambda(t) \geq y\right\} = \mathbb{P}\left\{D_I^\lambda(t) \geq x\right\} \mathbb{P}\left\{\widehat{X}_O^\lambda(t) \geq y\right\} + o(1)$$

Asymptotic Independence

"Natural scale" of station I = constant

"Natural scale" of station O = $\sqrt{\lambda}$

Theorem (asymptotic independence)

D_I^λ is asymptotically independent of $\widehat{X}_O^\lambda$, i.e, for all $t > 0$,

$$\mathbb{P}\left\{D_I^\lambda(t) \geq x, \widehat{X}_O^\lambda(t) \geq y\right\} = \mathbb{P}\left\{D_I^\lambda(t) \geq x\right\} \mathbb{P}\left\{\widehat{X}_O^\lambda(t) \geq y\right\} + o(1)$$

Note that $\widehat{X}_O^\lambda$ is scaled, but D_I^λ is not (requires refined analysis).

Main Idea of the Proof

Showing asymptotic independence of the sequence in n via asymptotic independence of the process in t

Main Idea of the Proof

Showing asymptotic independence of the sequence in n via asymptotic independence of the process in t

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

Main Idea of the Proof

Showing asymptotic independence of the sequence in n via asymptotic independence of the process in t

(*) Recall that $\{D_I^\lambda(s) : t \leq s \leq t + \epsilon\} \approx \{Q_b(s) : t \leq s \leq t + \lambda\epsilon\}$ 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 in n via asymptotic independence of the process in t

(*) Recall that $\{D_I^\lambda(s) : t \leq s \leq t + \epsilon\} \approx \{Q_b(s) : t \leq s \leq t + \lambda\epsilon\}$ 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 $\epsilon > 0$.

Steady state $Q_b(\infty)$ is independent of $Q_b(t)$.

Main Idea of the Proof

Showing asymptotic independence of the sequence in n via asymptotic independence of the process in t

(*) Recall that $\{D_I^{\lambda}(s) : t \leq s \leq t+\epsilon\} \approx \{Q_b(s) : t \leq s \leq t+\lambda\epsilon\}$ 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 $\epsilon > 0$.

Steady state $Q_b(\infty)$ is independent of $Q_b(t)$.

$\widehat{X}_O^{\lambda}$ has a continuous limit hence hardly changes within ϵ,

$$\widehat{X}_O^{\lambda}(t+\epsilon) \approx \widehat{X}_O^{\lambda}(t).$$

Pointwise Stationarity

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

Theorem (pointwise stationarity)

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

Pointwise Stationarity

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

Theorem (pointwise stationarity)

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

Pointwise stationarity and asymptotic independence allow us to obtain performance metrics by treating

- Station I as a stationary $M/M/N/K+M$ queue
- Station O as a $GI/M/N+M$ queue (that is independent of station I)

$$\mathbb{P}\{D_I^\lambda(t) \geq d, X_O^\lambda(t) \geq q\} = \mathbb{P}\{D_I^\lambda(\infty) \geq d\}\mathbb{P}\{X_O^\lambda(t) \geq q\}$$

- Overflow approximation simplifies analysis of station O

Waiting Times and Asymptotic ASTA

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

f is a continuous and bounded function or of the form $f(x) := \mathbb{1}\{x > \tau\}$.

Theorem (asymptotic finite-horizon ASTA)

For all $t > 0$,

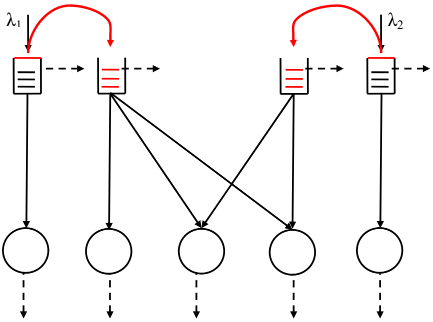
$$\lim_{\lambda\to\infty} \mathbb{E}\left[\frac{1}{A^\lambda(t)} \sum_{k=1}^{A^\lambda(t)} f(w_k^\lambda)\right] = \nu \frac{1}{t}\int_0^t \mathbb{E}\left[f(\widehat{W}_I(s))\right] ds + (1-\nu)\frac{1}{t}\int_0^t \mathbb{E}\left[f(\widehat{W}_O(s)) ds\right].$$

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

Generalizing to multiple classes

$$\min \quad \sum_i \mathbb{E}[\int_0^T C_i(Q_i^\pi(s))ds]$$

$$\text{s.t.} \quad \pi \in \Pi.$$



- Theorem: "Benefit from in-house state information is marginal."

Summary

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

Questions?