A Fair Policy for the Servers in the $G/GI/N$ Queue

Josh Reed

NYU
Stern School of Business

Joint work with Yair Shaki

Stochastic Networks Conference

June 20, 2012

Introduction

- Modern call centers employ 100's, if not 1000's of agents.

- Moreover, often times agents may have distinct skill sets.

Introduction

- For example, some of the agents staffed in a call center at a bank may be very good at opening a new account while others may be more skilled in handling cases of fraud.
- Some agents may even be trained to handle both of these types of customer service requests.
- Some agents might be fast at processing requests, while others might be slower.

Introduction

- A natural question which arises in these settings is how to decide who to route incoming customers to?
- Assume that there is a single customer class with several heterogeneous servers.
- Always route to the fastest available server?
- Perhaps use some sort of threshold rule?
- What should the proper objective be?
- Minimizing customer waiting times sounds reasonable but may not always be the best choice. Why not?

Introduction

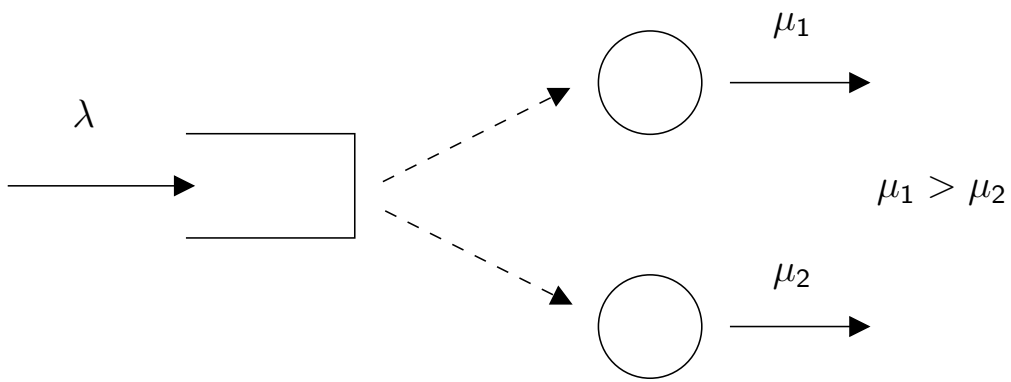
- Minimizing customer waiting times might be good for the customers arriving to the system but not so great for the servers themselves.
- This is especially true if the servers are human beings (as opposed to machines) as is the case for a telephone call center.
- In this talk, we will look how to develop policies which are efficient from the customers' point of view but are also "fair" to the servers in the system.

Outline of the Talk

- Introduction
- Literature Review
- Asymptotic Regime
- u-Greedy Policies
- Main Results
- Proof Techniques
- Future Work

Literature Review

- In (1984), Lin and Kumar considered the following system.



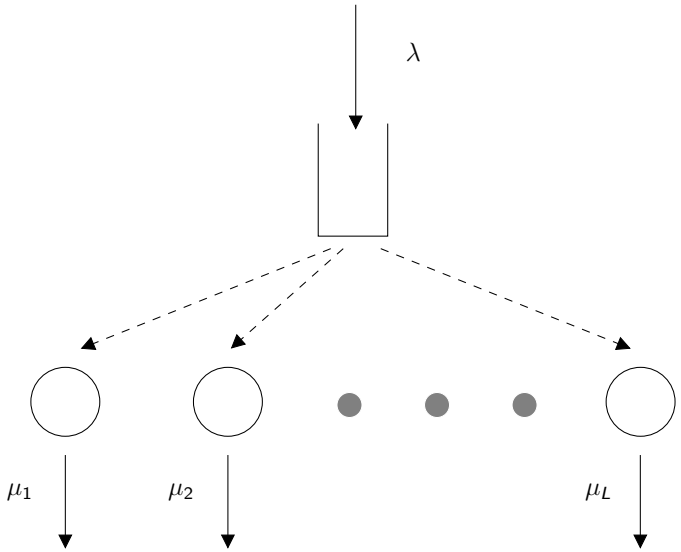
- There is only a single buffer and the decision is when a server becomes free, should you send a customer to it or not?

- Clearly, you should always send a customer to the faster server.

Literature Review

- But what about the slower server?
- Lin and Kumar proved that in order to minimize customer sojourn times you should only route to the slower server when the number of customers in the buffer is above a certain threshold.
- More difficult find solutions to when there are more than two servers.

The Inverted-V Model



- This is sometimes referred to as the "inverted-V" model.
- Because of the difficulty of handling multiple types of servers, many authors have turned to an asymptotic analysis.

The Asymptotic Regime

- Consider a sequence of systems indexed by the number of servers N which we let tend to ∞.
- The system with N servers has an arrival rate of λ^N.
- Each system has $L \geq 1$ server pools (fixed, does not change with N) and the number of servers in server pool l for $l = 1, ..., L$, is given by

$$N_l = \lfloor \nu_l N \rfloor,$$

where

$$\nu_1 + ... + \nu_L = 1.$$

- Service times in server pool l have a fixed distribution F_l with mean $1/\mu_l$.

The Asymptotic Regime

- The capacity of the system with N servers is approximately

$$\sum_{l=1}^{L} \lfloor \nu_l N_l \rfloor \mu_l.$$

- In the Halfin and Whitt regime, we assume that capacity is approximately matched with the incoming demand rate.

- In particular, letting

$$\beta^N = \frac{1}{\sqrt{N}} \left(\lambda^N - \sum_{l=1}^{L} \lfloor \nu_l N \rfloor \mu_l \right),$$

we assume that

$$\beta^N \to \beta < 0 \quad \text{as} \quad N \to \infty.$$

The Asymptotic Regime

- The previous convergence implies that

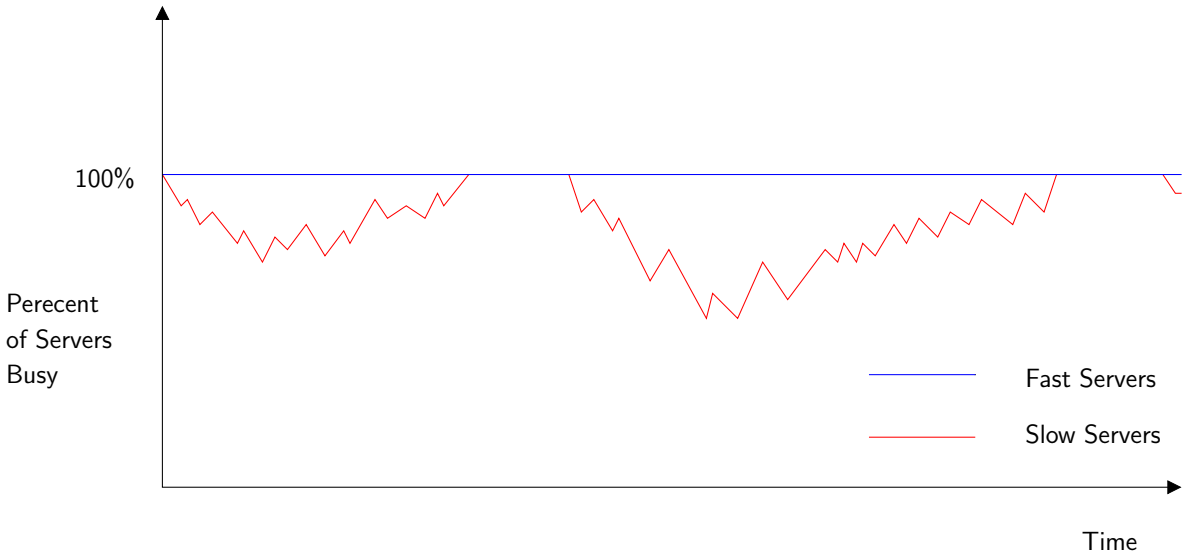
$$\sum_{l=1}^{L} \lfloor \nu_l N \rfloor \mu_l \quad = \quad \lambda^N - \beta\sqrt{N} + o(\sqrt{N}).$$

- In other words, the capacity of the system differs from the arrival rate by a $\mathcal{O}(\sqrt{N})$ term.

Literature Review

- Armony (2005) was one of the first authors to consider the inverted-V model in the Halfin and Whitt asymptotic regime.
- She considered the fastest server first (FSF) routing policy.
- Incoming customers are routed to the fastest available server. If it happens to be the slowest server in the system, no problem.
- Armony showed that for exponentially distributed service time in each server pool, FSF asymptotically minimizes customer waiting times. In particular, no thresholds are needed.

Literature Review



- Unfortunately, under the FSF routing policy, the slow servers will be given $\mathcal{O}(1/\sqrt{N})$ idle time, while the fast server pool will never be allowed to idle.

Literature Review

- Tezcan (2011) considered H_2^* service time distributions which are a mixture of an exponential and a point mass at zero.
- Showed that a static priority policy is optimal but not necessarily FSF.

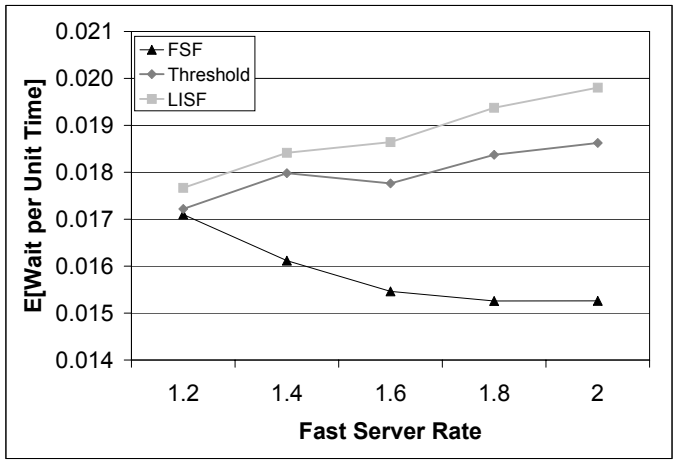
Literature Review

- Atar (2008) considered the longest idled served first (LISF) routing policy for the inverted-V model.
- LISF tends to be biased towards fast servers. They will finish serving customers more often and so will end up having longer cumulative idle times.
- Gurvich and Whitt propose (2009) Idleness Ratio (IR) routing which attempts to keeps the idle servers in fixed proportions. Performs similar to LISF asymptotically.
- Mandelbaum, Momcilovic and Tseytlin (2012) consider the Randomized Most-Idle (RMI) policy which uniformly at random picks an available server to route to next. Also performs similar to LISF asymptotically.

Literature Review

- Armony and Ward (2010) proposed a middle ground between minimizing customer waiting times and achieving server fairness.
- Their objective is to minimize customer waiting times subject to the steady state percentage of idle servers from each server pool being fixed constants.
- In Armony and Ward, it is shown that the asymptotically optimal policy is a FSF-excluding-pool-k policy. This policy operates similar to the FSF policy with the exception that server pool k is given the lowest priority.
- The server pool k varies depending on the number of customers in the system.

Literature Review



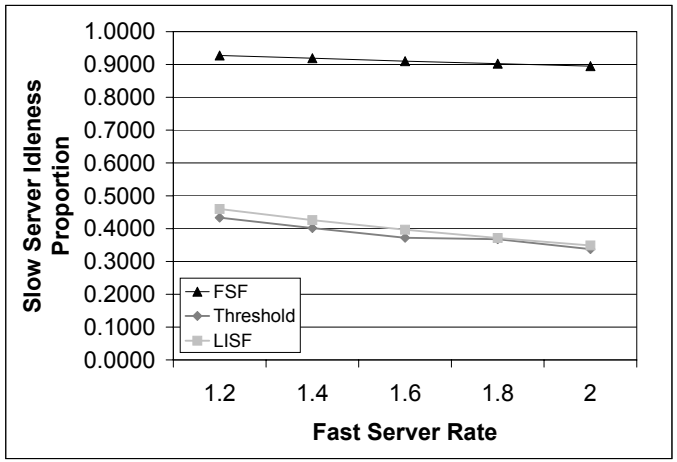


Figure : from Armony and Ward (2010)

Literature Review

- In (2011), Atar, Shaki and Shwartz modified LISF to longest cumulative idled served first (LIPF). This policy routes customers to the server pool with the longest cumulative idleness.
- Atar, Shaki and Shwartz showed that longest cumulative idled served first asymptotically equalizes cumulative idleness.
- They considered exponentially distributed service times in each server pool.
- However, empirical evidence suggests that service times at telephone call centers are not exponentially distributed.

Literature Review

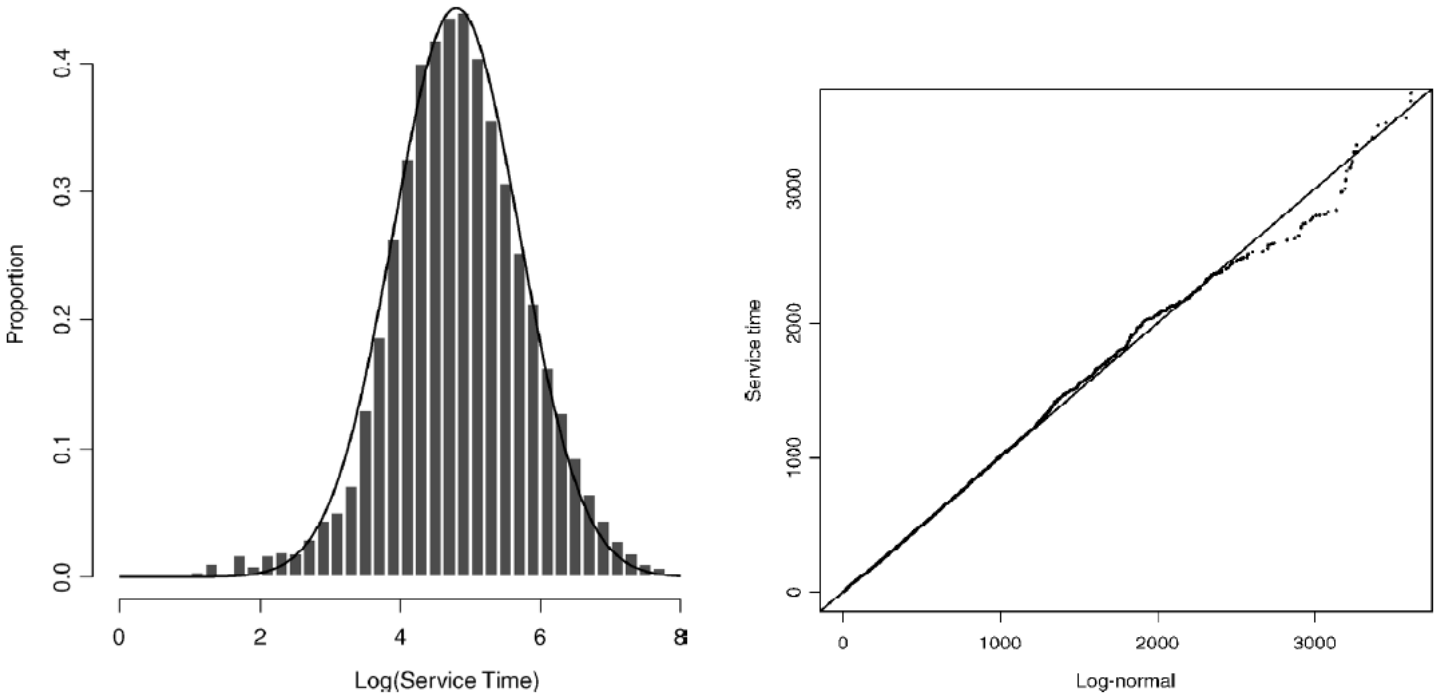


Figure 3. Histogram (a) and Q–Q Plot (b) of log(service time), November–December.

Figure : Picture from Brown et. al (2005)

Main Results

- Our goal in this work is to extend the results of Atar, Shaki and Shwartz to general service time distributions.
- In the process, we develop a new technique for the asymptotic analysis of many server queues with general service time distributions.
- This technique is based off of a simple conservation of flow identity and appears to be promising for analyzing a wide variety of routing policies for the inverted-V model such as FSF or LISF.
- It can also be used in the analysis of networks of many server queues.

u-GREEDY ROUTING POLICIES

- For each $l = 1, ..., L$, let $I_l(t)$ be the total number of idle servers in server pool l at time t and set

$$J_l(t) \quad = \quad \int_0^t I_l(s)ds$$

to be the total cumulative idleness of server pool l up until time t.

- Next, let $u = (u_1, ..., u_L)$ be a vector of target weights such that $0 < u_l < 1$ and

$$u_1 + ... + u_L \quad = \quad 1.$$

- The u-greedy routing policy then works in the following way.

u-GREEDY ROUTING POLICIES

1. If a server becomes free and there are still customers waiting in the queue, then the server selects the customer in the queue who has waited the longest to serve next. Thus, the policy is non-idling.

2. If a customer arrives to the system at time t and there are multiple pools with idle servers, then the customer is routed to the server pool with the largest value of $(1/u_l)J_l(t)$. In other words, incoming customers are routed to the server pool with the longest weighted cumulative idleness.

u-GREEDY ROUTING POLICIES

- The target weight vector $u = (u_1, ..., u_L)$ is chosen such that asymptotically the fraction of total cumulative idleness coming from server pool l is u_l.

- A desirable feature of u-greedy policies is that they do not require previous knowledge of the system parameters. They are so-called "blind".

Main Results

- Let
  - $A(t)$ = number of customers who have arrived to the system by time t
  - $Q(t)$ = number of customers in the queue at time t
  - $Z_l(t)$ = number of customers in service in server pool l at time t
- Next, define the corresponding fluid scaled quantities

$$\bar{A}^N(t) = \frac{A^N(t)}{N}, \quad \bar{Q}^N(t) = \frac{Q^N(t)}{N} \quad \text{and} \quad \bar{Z}_l^N(t) = \frac{Z_l^N(t)}{N}.$$

Main Results

- Also, let $E_l(t)$ be the number of customers who have entered service in server pool l by time t and define

$$\bar{E}_l^N(t) \quad = \quad \frac{E_l^N(t)}{N}.$$

- Finally, set

$$\bar{Z}^N(0) \quad = \quad (\bar{Z}_1^N(0), ..., \bar{Z}_L^N(0)).$$

- Then, we have the following fluid limit result.

Theorem 1

Suppose that

$$(\bar{A}^N, \bar{Q}^N(0), \bar{Z}^N(0)) \quad \Rightarrow \quad (\lambda e, 0, (\nu_1, ..., \nu_L)) \quad \text{as} \quad N \to \infty,$$

and that those customers in service in server pool l at time zero have i.i.d. residual service times equal to the equilibrium distribution F_l^e associated with F_l.

Then, under any non-idling routing policy, $\bar{E}_l^N \Rightarrow \nu_l \mu_l e$ as $N \to \infty$.

Main Results

- Theorem 1 states that under certain desirable initial conditions, the rate of customers entering service in each service pool is constant.
- It is interesting on its own but it is also useful in proving our second order results.

Main Results

- Define the diffusion scaled quantities

$$\tilde{A}^N(t) = \frac{A^N(t) - \lambda^N t}{\sqrt{N}}, \qquad \tilde{Q}^N(t) = \frac{Q^N(t)}{\sqrt{N}}$$

and

$$\tilde{Z}_l^N(t) = \frac{Z_l^N(t) - \lfloor \nu_l N \rfloor}{\sqrt{N}}.$$

- Also, let

$$\tilde{Z}^N(0) = (\tilde{Z}_1^N(0), ..., \tilde{Z}_L^N(0)).$$

Main Results

- Let us also define

$$\tilde{J}_l^N(t) = \frac{1}{\sqrt{N}} \int_0^t J_l^N(s)ds$$

to be the normalized cumulative idleness of server pool l and let

$$\tilde{J}^N(t) = \sum_{l=1}^{L} \tilde{J}_l^N(t)$$

be the normalized cumulative idleness of the system.

- Also, for $\varepsilon > 0$, define $\gamma^N(\varepsilon) = \inf\{t \geq 0 : \tilde{J}^N(t) > \varepsilon\}$.

- Our next two results are the following.

Theorem 2

Suppose that F_l is continuous with a finite second moment, that

$$(\tilde{A}^N, \tilde{Q}^N(0), \tilde{Z}^N(0)) \quad \Rightarrow \quad (\tilde{A}, \tilde{Q}^N, \tilde{Z}^N) \quad \text{as} \quad N \to \infty,$$

and that customers in service in server pool l at time zero have i.i.d. residual service times equal to the equilibrium distribution F_l^e.

Then, the sequence $\{\gamma^N(\varepsilon), N \geq 0\}$ is tight, and one has that for each $t \geq 0$,

$$\liminf_{N\to\infty} \mathbb{P}\left(\gamma^N(\varepsilon) < \infty \text{ and } \max_{l\in L}\left|\frac{\tilde{J}_l^N(\gamma^N(\varepsilon)+t)}{\tilde{J}^N(\gamma^N(\varepsilon)+t)} - u_l\right| \leq \varepsilon\right) \quad \geq \quad 1-\varepsilon.$$

Main Results

- We also may prove the following result.

Theorem 3

Suppose the same conditions as in Theorem 2. Then, for every $\varepsilon > 0$ and every $T \geq 0$,

$$\lim_{N\to\infty} \mathbb{P}\left(\max_{i,j\in L, i\neq j} \sup_{s\in[0,T]} \left| (1/u_i)\tilde{J}_i^N(s) - (1/u_j)\tilde{J}_j^N(s) \right| \geq \varepsilon \right) = 0.$$

Main Results

- Our third main result provides a second order process level limit for the total number of customers in the system.

- Let $X^N(t)$ be the total number of customers in the Nth system at time t and let

$$\tilde{X}^N(t) \quad = \quad \frac{X^N(t) - N}{\sqrt{N}}.$$

- Also, let M_l be the renewal function associated with the service time distribution F_l.

- We then have the following.

Theorem 4

Assume the same conditions as in Theorems 2 and 3 and in addition assume that F_l has a continuous density. Then,

$$\tilde{X}^N \quad \Rightarrow \quad \tilde{X} \quad \text{as} \quad N \to \infty,$$

where $\tilde{X}$ is the unique, strong solution to

$$\begin{aligned} \tilde{X}(t) &= \tilde{X}(0) + \tilde{A}(t) - \beta t \\ &\quad - \sum_{l=1}^{L} \tilde{S}_l(t) - \int_0^t \min\{0, \tilde{X}(t-s)\} d\left(\sum_{l=1}^{L} u_l M_l(s) \right). \end{aligned}$$

Main Results

- For each $l = 1, ..., L$, the process $\tilde{S}_l$ in Theorem 4 is a centered, Gaussian process with covariance function

$$\begin{aligned} & E[\tilde{S}_l(t)\tilde{S}_l(t+\delta)] \\ = \; & 2\nu_l \int_0^t (M_l(u) - \mu_l u + .5)du \\ & + \nu_l \mu_l^3 \int_0^t \int_0^\delta M_l(t-a) M_l(\delta - b) dF_l(a+b), \end{aligned}$$

  for $t, \delta \geq 0$.

- In the case of exponentially distributed service times at each service pool, one has that $M_l(t) = \mu_l t$ and $\tilde{S}_l$ is a Brownian motion. It is then straightforward to verify that the limit process $\tilde{X}$ of Theorem 4 reduces to the diffusion process obtained in Theorem 4.1 of Atar, Shaki and Shwartz.

Proof Techniques

- The general proof technique for the above results begins with a conservation of flow identity.
- This is similar to the approach used to prove conventional heavy traffic limit theorems.
- However, there are significant difficulties which arise in the many-server setting.
- In order to illustrate the general technique, we consider a system with a single server pool and assume that all of the servers are busy serving customers at time 0 and that there are no customers in the queue at time 0.

Proof Techniques

- Assume that we have N servers.
- Let $A(t)$ be the number of customers that have arrived to the system by time t and assume that arrivals occur at rate λ.
- Let $S_n(t)$ be the number of customers served by server $n = 1, ..., N$, in its first t units of processing time.
- Let $B_n(t)$ be the cumulative busy time of server $n = 1, ..., N$, up until time t.
- Let $X(t)$ be the number of customers in the system at time t.

Proof Techniques

- Using a simple conservation of flow identity, it is straightforward to write that

$$X(t) \quad = \quad N + A(t) - \sum_{n=1}^{N} S_n(B_n(t)).$$

- The next step in conventional heavy traffic analysis is to center the arrival and departure processes.

- As usual, we center the arrival process by λt and so we write

$$\hat{A}(t) \quad = \quad A(t) - \lambda t.$$

- But what about the departure process $\sum_{n=1}^{N} S_n(B_n(t))$?

Proof Techniques

- Let M be the renewal function associated with the service time distribution F
- Next, let $Q_n(t) = 1$ if server n is busy serving a customer at time t and let $Q_n(t) = 0$ otherwise.
- Then, we define the centered departure process from server n by setting

$$\hat{S}_n(t) \quad = \quad S_n(B_n(t)) - \left(\mu t - \int_0^t (1 - Q_n(t-s)) dM(s) \right).$$

Proof Techniques

- It may be rigorously shown that $E[\hat{S}_n(t)] = 0$.
- Now, using this choice of centering and our conservation of flow identity we may write

$$\begin{aligned} X(t) - N &= \hat{A}(t) - \sum_{n=1}^{N} \hat{S}_n(t) + (\lambda - N\mu)t \\ &\quad + \sum_{n=1}^{N} \int_0^t (1 - Q_n(t-s)) dM(s). \end{aligned}$$

- Then, by the non-idling condition, we obtain that

$$\begin{aligned} & \sum_{n=1}^{N} \int_0^t (1 - Q_n(t-s)) dM(s) \\ = & \int_0^t \sum_{n=1}^{N} (1 - Q_n(t-s)) dM(s) \\ = & -\int_0^t \min((X(t-s) - N), 0) dM(s). \end{aligned}$$

Proof Techniques

- Substituting this expression into our centered version of the queue length equation, we then obtain that

$$\begin{aligned} X(t) - N &= \hat{A}(t) - \sum_{n=1}^{N} \hat{S}_n(t) + (\lambda - N\mu)t \\ &\quad - \int_0^t \min((X(t-s) - N), 0) dM(s). \end{aligned}$$

- Now note that $X(t) - N$ appears on both sides of the above and so one may view $X(t) - N$ as the solution to a convolution equation.

- Indeed, we have the following result.

Lemma 5

Let M be the renewal function associated with a distribution function F and for each $z \in D([0,\infty),\mathbb{R})$, let x be the solution to

$$x(t) \quad = \quad z(t) - \int_0^t \min(x(t-s),0)dM(s), \quad t \geq 0.$$

Then, x is unique. Moreover, the mapping

$$\Phi : D([0,\infty),\mathbb{R}) \mapsto D([0,\infty),\mathbb{R})$$

such that $x = \Phi(z)$ is Lipschitz continuous with respect to the topology of uniform convergence over compact sets and measurable with respect to the Skorokhod J_1 topology.

Proof Techniques

- Thus, after proper scaling, one may write

$$\tilde{X}^N \quad = \quad \Phi(\tilde{A}^N - \tilde{S}^N + N^{-1/2}(\lambda^N - N\mu)e).$$

- Moreover, it may be shown that

$$\tilde{A}^N - \tilde{S}^N + N^{-1/2}(\lambda^N - N\mu)e \quad \Rightarrow \quad \tilde{A} - \tilde{S} + \beta e.$$

- The main result now follows by the representation above and an application of the continuous mapping theorem.

Future Research

- The general proof technique can be used to analyze other routing policies for the inverted-V model such as FSF, LISF, RMI, IR or perhaps threshold policies.
- It could also be used to analyze networks of many server queues, for instance queues in tandem.

Thank You

THANK YOU!