STAFFING AND CONTROL OF INSTANT MESSAGING BASED CUSTOMER SERVICE CENTERS

Jiheng Zhang

(Joint work with Jun Luo)
Model and Motivation

Server Pool with multiple LPS servers

Instant Messaging based Services Centers / J. Zhang
Model and Motivation

Server Pool with multiple LPS servers

Customer contact centers via instant messaging.

Instant Messaging based Services Centers / J. Zhang
Servers work at different speed depending on how many customer in service.

![Chart showing service rate per half hour for different numbers of customers in service.](chart.png)
Servers work at different speed depending how many customer in service.

Classify the pool of $N$ homogeneous servers into “levels”.

- Level $k$: all servers serving $k$ customers.
The classical $\wedge$-model:
The classical $\wedge$-model:
Many-server Queues

- Puhalskii 2007, Mandelbaum, Massey & Reiman 1998, ...
- Perry & Whitt 2010 – now...

Averaging Principle

- Kurtz 1992, Hunt & Kurtz 1994, ...

LPS Queues

- Zhang, Dai and Zwart 2010, 2011
- Zhang & Zwart 2008, Gupta & Zhang 2011, ...
Notations

- Server pool $Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1}$
Notations

- Server pool $Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1}$
- Queue $Q(t) \in \mathbb{N}$
Notations

- Server pool $Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1}$
- Queue $Q(t) \in \mathbb{N}$

$$\sum_{k=0}^{K} Z_k(t) = N, \quad \text{and} \quad Q(t)(N - Z_K(t)) = 0.$$
Notations

- Server pool $Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1}$
- Queue $Q(t) \in \mathbb{N}$

$$\sum_{k=0}^{K} Z_k(t) = N, \quad \text{and} \quad Q(t)(N - Z_K(t)) = 0.$$ 

- Service

$$D_k(t) = S_k \left( \gamma_k \int_0^t Z_k(s)ds \right)$$
Notations

- Server pool $Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1}$
- Queue $Q(t) \in \mathbb{N}$

$$\sum_{k=0}^{K} Z_k(t) = N, \quad \text{and} \quad Q(t)(N - Z_K(t)) = 0.$$ 

- Service

$$D_k(t) = S_k \left( \gamma_k \int_{0}^{t} Z_k(s) ds \right)$$

Key assumption: exponential service time
Notations

- Server pool \( Z(t) = (Z_0(t), Z_1(t), \ldots, Z_K(t)) \in \mathbb{N}^{K+1} \)
- Queue \( Q(t) \in \mathbb{N} \)

\[
\sum_{k=0}^{K} Z_k(t) = N, \quad \text{and} \quad Q(t)(N - Z_K(t)) = 0.
\]

- Service

\[
D_k(t) = S_k \left( \gamma_k \int_0^t Z_k(s)ds \right)
\]

Key assumption: exponential service time

- The index (routing)

\[
i_*(t) = \min\{0 \leq k \leq K : Z_k(t) > 0\}.
\]
System Dynamics – some simulations

\[ \lambda^n = 400, \ N^n = 200, \ K = 6 \text{ and } \gamma = (1, 1.6, 1.8, 2.2, 2.3, 2.4). \]
System Dynamics

When will \( z_k(s) \) jump by 1
System Dynamics

When will $z_k(s)$ jump by 1

- $i_*(s-)=k-1$, and an arrival happens at $s$
When will $z_k(s)$ jump by 1

- $i^*(s-) = k - 1$, and an arrival happens at $s$
- a service completion from group $k + 1$ at $s$
System Dynamics

When will $z_k(s)$ jump by 1

- $i_*(s-) = k - 1$, and an arrival happens at $s$
- a service completion from group $k + 1$ at $s$

When will $z_k(s)$ jump by -1
System Dynamics

When will $z_k(s)$ jump by 1

- $i_*(s- ) = k - 1$, and an arrival happens at $s$
- a service completion from group $k + 1$ at $s$

When will $z_k(s)$ jump by -1

- $i_*(s- ) = k$, and an arrival happens at $s$
System Dynamics

When will $z_k(s)$ jump by 1

- $i_*(s-) = k - 1$, and an arrival happens at $s$
- a service completion from group $k + 1$ at $s$

When will $z_k(s)$ jump by -1

- $i_*(s-) = k$, and an arrival happens at $s$
- a service completion from group $k$ at $s$
System Dynamics

When will \( z_k(s) \) jump by 1

- \( i_*(s-) = k - 1 \), and an arrival happens at \( s \)
- A service completion from group \( k + 1 \) at \( s \)

When will \( z_k(s) \) jump by -1

- \( i_*(s-) = k \), and an arrival happens at \( s \)
- A service completion from group \( k \) at \( s \)

**Dynamic Equation for** \( Z_k \), \( 0 < k < K \)

\[
Z_k(t) = Z_k(0) + \int_0^t 1_{\{i_*(s-) = k-1\}} d\Lambda(s) + \int_0^t 1_{\{Q(s-) = 0\}} dD_{k+1}(s) - \int_0^t 1_{\{i_*(s-) = k\}} d\Lambda(s) - D_k(t)
\]
Dynamic Equation for $Z_0$

$$Z_0(t) = Z_0(0) - \int_0^t 1_{\{i^*(s-) = 0\}} d\Lambda(s) + D_1(t)$$
System Dynamics

**Dynamic Equation for \( Z_0 \)**

\[
Z_0(t) = Z_0(0) - \int_0^t 1\{i_*(s-)=0\} d\Lambda(s) + D_1(t)
\]

**Dynamic Equation for \( Z_K \)**

\[
Z_K(t) = Z_K(0) + \int_0^t 1\{i_*(s-)=K-1\} d\Lambda(s) - \int_0^t 1\{Q(s-)=0\} dD_K(s)
\]
System Dynamics

**Dynamic Equation for** $Z_0$

$$Z_0(t) = Z_0(0) - \int_0^t 1\{i_*(s-)=0\} d\Lambda(s) + D_1(t)$$

**Dynamic Equation for** $Z_K$

$$Z_K(t) = Z_K(0) + \int_0^t 1\{i_*(s-)=K-1\} d\Lambda(s) - \int_0^t 1\{Q(s-)=0\} dD_K(s)$$

**Dynamic Equation for** $Q$

$$Q(t) = Q(0) + \int_0^t 1\{i_*(s-)=K\} d\Lambda(s) - \int_0^t 1\{Q(s)>0\} dD_K(s)$$
Constant Arrival v.s. Time-Varying Arrival

Arrival process, $\Lambda(t)$, is assumed to be non-homogeneous Poisson process with rate $\lambda(t)$. 

**Arrival Rate/ 0.5 Hour on Monday**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>69.94</td>
</tr>
<tr>
<td>8</td>
<td>132</td>
</tr>
<tr>
<td>12</td>
<td>186.5</td>
</tr>
<tr>
<td>16</td>
<td>155.5</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>

**Arrival Rate/ 0.5 Hour on Tuesday**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>65.44</td>
</tr>
<tr>
<td>8</td>
<td>174.5</td>
</tr>
<tr>
<td>12</td>
<td>155.5</td>
</tr>
<tr>
<td>16</td>
<td>157.1</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>

**Arrival Rate/ 0.5 Hour on Wednesday**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>35.44</td>
</tr>
<tr>
<td>8</td>
<td>94.5</td>
</tr>
<tr>
<td>12</td>
<td>155.5</td>
</tr>
<tr>
<td>16</td>
<td>157.1</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>

**Arrival Rate/ 0.5 Hour on Thursday**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>66.31</td>
</tr>
<tr>
<td>8</td>
<td>132</td>
</tr>
<tr>
<td>12</td>
<td>155.5</td>
</tr>
<tr>
<td>16</td>
<td>157.1</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>

**Arrival Rate/ 0.5 Hour on Friday**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>58.31</td>
</tr>
<tr>
<td>8</td>
<td>132</td>
</tr>
<tr>
<td>12</td>
<td>155.5</td>
</tr>
<tr>
<td>16</td>
<td>157.1</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>

**Arrival Rate/ 0.5 Hour on Weekdays**

<table>
<thead>
<tr>
<th>Time of the Day</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>58.91</td>
</tr>
<tr>
<td>8</td>
<td>132</td>
</tr>
<tr>
<td>12</td>
<td>155.5</td>
</tr>
<tr>
<td>16</td>
<td>157.1</td>
</tr>
<tr>
<td>20</td>
<td>157.1</td>
</tr>
<tr>
<td>24</td>
<td>157.1</td>
</tr>
</tbody>
</table>
A Heavy Traffic Regime

Large number of servers to accommodate large demand.
A Heavy Traffic Regime

Large number of servers to accommodate large demand.

- Consider a sequence of system indexed by \( n \)

\[
\frac{1}{n} N^n \to N, \quad \text{and} \quad \frac{1}{n} \lambda^n(t) \to \lambda(t), \quad \text{as} \ n \to \infty.
\]
A Heavy Traffic Regime

Large number of servers to accommodate large demand.

- Consider a sequence of system indexed by $n$

  \[
  \frac{1}{n} N^n \to N, \quad \text{and} \quad \frac{1}{n} \lambda^n(t) \to \lambda(t), \quad \text{as} \ n \to \infty.
  \]

- But each server's service rate $\gamma = (\gamma_1, \ldots, \gamma_K)$ is fixed.
A Heavy Traffic Regime

Large number of servers to accommodate large demand.

Consider a sequence of system indexed by $n$

$$\frac{1}{n}N^n \to N, \quad \text{and} \quad \frac{1}{n}\lambda^n(t) \to \lambda(t), \quad \text{as } n \to \infty.$$ 

But each server’s service rate $\gamma = (\gamma_1, \ldots, \gamma_K)$ is fixed.

Fluid Scaling:

$$\bar{Z}^n(t) = \frac{Z^n(t)}{n}, \quad \bar{Q}^n(t) = \frac{Q^n(t)}{n}.$$
Optimality

Utility function

$$\bar{C}_T^n(\bar{N}^n, K) = c\bar{N}^n + \frac{1}{T} \mathbb{E} \left[ \int_0^T h(\bar{Z}^n(s), \bar{Q}^n(s)) \, ds \right].$$
Optimality

Utility function

\[ \bar{C}_T^n(\bar{N}^n, K) = c\bar{N}^n + \frac{1}{T} \mathbb{E} \left[ \int_0^T h(\bar{Z}^n(s), \bar{Q}^n(s)) \, ds \right]. \]

Eg. linear cost \( h(z, q) = \frac{1}{\lambda} (q + \sum_k k z_k). \)
Optimality

Utility function

$$\bar{C}_T^n(\bar{N}^n, K) = c\bar{N}^n + \frac{1}{T} \mathbb{E} \left[ \int_0^T h(\bar{Z}^n(s), \bar{Q}^n(s)) \, ds \right].$$

Eg. linear cost $h(z, q) = \frac{1}{\lambda} (q + \sum_k k z_k)$.

Staffing $\{\bar{N}^*_n\}$ and control $K_*$ is asymptotically optimal if

$$\lim_{n \to \infty} \sup_{K_*} \bar{C}_T^n(\bar{N}^*_n, K_*) \leq \lim_{n \to \infty} \inf \bar{C}_T^n(\bar{N}^n, K).$$
Optimality

Utility function

\[
\bar{C}^n_T(\bar{N}^n, K) = c\bar{N}^n + \frac{1}{T} \mathbb{E} \left[ \int_0^T h(\bar{Z}^n(s), \bar{Q}^n(s)) \, ds \right].
\]

Eg. linear cost \( h(z, q) = \frac{1}{\lambda} (q + \sum_k k z_k) \).

Staffing \( \{\bar{N}^n\} \) and control \( K^* \) is asymptotically optimal if

\[
\limsup_{n \to \infty} \bar{C}^n_T(\bar{N}^n, K^*) \leq \liminf_{n \to \infty} \bar{C}^n_T(\bar{N}^n, K).
\]

\[
\limsup_{T \to \infty} \limsup_{n \to \infty} \bar{C}^n_T(\bar{N}^n, K^*) \leq \limsup_{T \to \infty} \liminf_{n \to \infty} \bar{C}^n_T(\bar{N}^n, K).
\]
PROPOSITION

In the heavy traffic regime

\[ \tilde{C}_T^n(\tilde{N}^n, K) \rightarrow C_T(N, K), \quad \text{as } n \rightarrow \infty. \]
Optimal Staffing and Control - Finite Horizon

\[ \gamma = (2, 3, 2.7, 3.2), \quad c = 19, \quad h(z, q) = 1 \times \left( \sum_{k} k z_k + q \right) \]
PROPOSITION

Assume $\lambda(t) \equiv \lambda$ and $\gamma_k$ is increasing and $N > \lambda / \sup_{k \leq K_0} \gamma_k$. For any $K \leq K_0$

$$\lim_{T \to \infty} \lim_{n \to \infty} \bar{C}_T^n(\bar{N}^n, K) \to C(N), \quad \text{as } n \to \infty.$$
Optimal Staffing and Control - Infinite Horizon

(a) Linear Holding Cost

(b) Quadratic Holding Cost

Instant Messaging based Services Centers / J. Zhang
Optimal Staffing and Control - Infinite Horizon

Graphs showing the cost for different values of $N$ under linear and quadratic holding costs for fluid models with $n=50$, $n=100$, and $n=200$.

\[
\gamma = (1, 1.6, 1.9, 2.3, 2.6, 2.8)
\]
\[
\lambda = 1, \ c = 2, \text{ and } h(z, q) = 1 \times \left( \sum_{k} k z_k + q \right).
\]

Instant Messaging based Services Centers / J. Zhang
Optimal Staffing and Control - Infinite Horizon

\begin{align*}
\gamma &= (1, 1.8, 2.1, 2.2, 2.5, 2.7) \\
\lambda &= 1, \ c = 10, \text{ and } \ h(z, q) &= \left( \sum_k k z_k + q \right)^2.
\end{align*}

Instant Messaging based Services Centers / J. Zhang
Fluid Model

Let \((z, q)\) be the fluid counterpart of \((\bar{Z}^n, \bar{Q}^n)\).
Fluid Model

Let \((z, q)\) be the fluid counterpart of \((\bar{Z}^n, \bar{Q}^n)\).

\[
\mathcal{S} = \{(z, q) \in [0, N]^{K+1} \times \mathbb{R}_+ : \sum_{k=0}^{K} z_k = N \text{ and } q(N - z_K) = 0\}
\]
Fluid Model

Let \((z, q)\) be the fluid counterpart of \((\bar{Z}^n, \bar{Q}^n)\).

\[
S = \{(z, q) \in [0, N]^{K+1} \times \mathbb{R}_+ : \sum_{k=0}^{K} z_k = N \text{ and } q(N - z_K) = 0\}
\]

How do we deal with \(\int_0^t 1_{\{i_*(s-) = k\}} d\bar{\Lambda}^n(s)\)?
Let $(z, q)$ be the fluid counterpart of $(\bar{Z}^n, \bar{Q}^n)$.

$$S = \{(z, q) \in [0, N]^{K+1} \times \mathbb{R}_+ : \sum_{k=0}^{K} z_k = N \text{ and } q(N - z_K) = 0 \}$$

How do we deal with $\int_0^t 1_{\{i_*(s-) = k\}} d\bar{\Lambda}^n(s)$?

Introduce the mapping $f : \mathbb{R}^{K+2} \to [0, 1]^{K+1}$,

$$f_k(z, \lambda) = \begin{cases} \frac{\gamma k + 1 z k + 1}{\lambda} \wedge 1, & \text{if } k = I(z) - 1, \\ \left(1 - \frac{\gamma k z k}{\lambda}\right)^+, & \text{if } k = I(z), \\ 0, & \text{otherwise}, \end{cases}$$
Fluid Model

Let \((z, q)\) be the fluid counterpart of \((\bar{Z}^n, \bar{Q}^n)\).

\[ S = \{(z, q) \in [0, N]^{K+1} \times \mathbb{R}_+ : \sum_{k=0}^{K} z_k = N \text{ and } q(N - z_K) = 0\} \]

How do we deal with
\[ \int_0^t 1_{\{i_*(s-) = k\}} d\bar{\Lambda}^n(s) \,? \]

Introduce the mapping \(f : \mathbb{R}^{K+2} \rightarrow [0, 1]^{K+1}\),

\[ f_k(z, \lambda) = \begin{cases} \frac{\gamma_k z_k + 1}{\lambda} \wedge 1, & \text{if } k = I(z) - 1, \\ \left(1 - \frac{\gamma_k z_k}{\lambda}\right)^+, & \text{if } k = I(z), \\ 0, & \text{otherwise}, \end{cases} \]

where \(I(z) = \min\{0 \leq k \leq K : z_k > 0\}\).
Fluid Model

Divide the space $\mathbb{S}$ into two regions

$$
\mathbb{S}_+ = \left\{ (z_0, \ldots, z_K, q) \in \mathbb{S} : q > 0 \right\},
$$
$$
\mathbb{S}_0 = \left\{ (z_0, \ldots, z_K, q) \in \mathbb{S} : q = 0 \right\}.
$$

The fluid model can be defined by ODE’s in the two regions.
Divide the space $\mathcal{S}$ into two regions

$$
\begin{align*}
\mathcal{S}_+ &= \left\{ (z_0, \ldots, z_K, q) \in \mathcal{S} : q > 0 \right\}, \\
\mathcal{S}_0 &= \left\{ (z_0, \ldots, z_K, q) \in \mathcal{S} : q = 0 \right\}.
\end{align*}
$$

The fluid model can be defined by ODE’s in the two regions.

On $\mathcal{S}_+$

$$
\begin{align*}
z'_k(t) &= 0, \quad 0 \leq k \leq K, \\
q'(t) &= \lambda(t) - \gamma K N.
\end{align*}
$$
Fluid Model

on $\mathbb{S}_0$

\[
\begin{align*}
    z'_0(t) &= -f_0(z(t), \lambda(t))\lambda(t) + \gamma_1 z_1(t), \\
    z'_k(t) &= f_{k-1}(z(t), \lambda(t))\lambda(t) + \gamma_{k+1} z_{k+1}(t) \\
        &\quad - f_k(z(t), \lambda(t))\lambda(t) - \gamma_k z_k(t), \quad 0 < k < K, \\
    z'_K(t) &= f_{K-1}(z(t), \lambda(t))\lambda(t) - \gamma_K z_K(t), \\
    q'(t) &= f_K(z(t), \lambda(t))\lambda(t).
\end{align*}
\]
Fluid Model

on $S_0$

\[
\begin{align*}
    z'_0(t) &= -f_0(z(t), \lambda(t))\lambda(t) + \gamma_1 z_1(t), \\
    z'_k(t) &= f_{k-1}(z(t), \lambda(t))\lambda(t) + \gamma_{k+1} z_{k+1}(t) \\
    &\quad - f_k(z(t), \lambda(t))\lambda(t) - \gamma_k z_k(t), \quad 0 < k < K, \\
    z'_K(t) &= f_{K-1}(z(t), \lambda(t))\lambda(t) - \gamma_K z_K(t), \\
    q'(t) &= f_K(z(t), \lambda(t))\lambda(t).
\end{align*}
\]

The ODEs can be written into a vector form,

\[
(z', q') = \Psi(t, z, q).
\]
Theorem (Existence and Uniqueness)

Assume that $\lambda(t)$ is a piece-wise continuous function. There exists a unique solution to the fluid model, i.e., the ODEs, with initial condition $(z(0), q(0)) \in S$. 
Existence and Uniqueness of Fluid Model Solution

**Theorem (Existence and Uniqueness)**

Assume that $\lambda(t)$ is a piece-wise continuous function. There exists a unique solution to the fluid model, i.e., the ODEs, with initial condition $(z(0), q(0)) \in \mathbb{S}$.

For the solution $(z, q)$, define the associated *fluid cost* as

$$C_T(N, K) = cN + \frac{1}{T} \int_0^T h(z(s), q(s)) ds.$$
When the arrival rate is constant, we can study the “steady state” of the fluid model.
Invariant State of Fluid Model

When the arrival rate is constant, we can study the “steady state” of the fluid model.

Assumption: $\gamma_k$ is increasing in $k$. 
Invariant State of Fluid Model

When the arrival rate is constant, we can study the “steady state” of the fluid model.

Assumption: $\gamma_k$ is increasing in $k$.

Then there exists some $k^*$ such that

\[
\begin{align*}
\lambda &= \gamma_{k^*+1}z_{k^*+1} + \gamma_{k^*}z_{k^*}, \\
N &= z_{k^*+1} + z_{k^*}.
\end{align*}
\]
Invariant State of Fluid Model

When the arrival rate is constant, we can study the “steady state” of the fluid model.

Assumption: \( \gamma_k \) is increasing in \( k \).

Then there exists some \( k^* \) such that

\[
\begin{align*}
\lambda &= \gamma_{k^*+1}z_{k^*+1} + \gamma_{k^*}z_{k^*}, \\
N &= z_{k^*+1} + z_{k^*}.
\end{align*}
\]

This implies that

\[
\begin{align*}
z_{k^*} &= \frac{\gamma_{k^*+1}N - \lambda}{\gamma_{k^*+1} - \gamma_{k^*}}, \quad z_{k^*+1} = \frac{\lambda - \gamma_{k^*}N}{\gamma_{k^*+1} - \gamma_{k^*}}.
\end{align*}
\]
When the arrival rate is constant, we can study the “steady state” of the fluid model.

Assumption: $\gamma_k$ is increasing in $k$.

Then there exists some $k^*$ such that

$$
\begin{align*}
\lambda &= \gamma_{k^*+1} z_{k^*+1} + \gamma_{k^*} z_{k^*}, \\
N &= z_{k^*+1} + z_{k^*}.
\end{align*}
$$

This implies that

$$
\begin{align*}
z_{k^*} &= \frac{\gamma_{k^*+1} N - \lambda}{\gamma_{k^*+1} - \gamma_{k^*}}, \\
z_{k^*+1} &= \frac{\lambda - \gamma_{k^*} N}{\gamma_{k^*+1} - \gamma_{k^*}}.
\end{align*}
$$

Define it to be the invariant state $\mathcal{I}$. 
**Theorem (Convergence to Invariant State)**

Given any initial value \((z(0), q(0)) \in \mathcal{S}\) the fluid model converges, fast, to the invariant state \(\tilde{z}(N)\).
Theorem (Convergence to Invariant State)

Given any initial value \((z(0), q(0)) \in \mathbb{S}\) the fluid model converges, fast, to the invariant state \(\tilde{z}(N)\).

Based on this proposition, it is easy to see that the fluid cost

\[
\lim_{T \to \infty} C_T(N, K) \rightarrow cN + h(\tilde{z}(N), 0) \triangleq C(N).
\]
Theorem (FWLLN)

If \((\bar{Z}^n(0), \bar{Q}^n(0)) \Rightarrow (z_0, q_0)\), then \((\bar{Z}^n, \bar{Q}^n)\) converges weakly in the heavy traffic regime to the fluid model solution \((z, q)\) with \((z(0), q(0)) = (z_0, q_0)\).
Simulated Stochastic Model and the Fluid Model

\[ \gamma = (1, 1, 6, 1.8, 2.2), \quad K = 4 \text{ and } \lambda(t) = 2 + 1 \sin(t). \]

\[ \lambda^n(t) = n\lambda(t), \quad n = 50, 100, 200. \]
\[ \gamma = (1, 1.6, 1.8, 2.2, 2.3, 2.4), \quad K = 6, \quad n = 200, \quad N_n = 200, \quad \lambda^n = 390. \]
\[ \tilde{z} = (0, 0, 0, 5/8, 3/8, 0, 0) \quad \text{and} \quad \tilde{q} = 0. \]
## Simulation with general service time distributions

<table>
<thead>
<tr>
<th>Performance</th>
<th>Exponential</th>
<th>Erlang-2</th>
<th>LN(1, 4)</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>0.00004</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>±0.0008</td>
<td>±0.0007</td>
<td>±0.0010</td>
<td>–</td>
</tr>
<tr>
<td>Level 1</td>
<td>0.0088</td>
<td>0.0084</td>
<td>0.0102</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>±0.0032</td>
<td>±0.0024</td>
<td>±0.0059</td>
<td>–</td>
</tr>
<tr>
<td>Level 2</td>
<td>1.7325</td>
<td>1.7174</td>
<td>1.7553</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>±0.0201</td>
<td>±0.0154</td>
<td>±0.0310</td>
<td>–</td>
</tr>
<tr>
<td>Level 3</td>
<td>122.2821</td>
<td>122.2991</td>
<td>122.3772</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>±0.3716</td>
<td>±0.2532</td>
<td>±0.4488</td>
<td>–</td>
</tr>
<tr>
<td>Level 4</td>
<td>75.9753</td>
<td>75.9740</td>
<td>75.8561</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>±0.3837</td>
<td>±0.2649</td>
<td>±0.4683</td>
<td>–</td>
</tr>
<tr>
<td>Level 5</td>
<td>0.0010</td>
<td>0.0006</td>
<td>0.0007</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>±0.0007</td>
<td>±0.0003</td>
<td>±0.0006</td>
<td>–</td>
</tr>
<tr>
<td>Level 6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sojourn Time</td>
<td>1.7287</td>
<td>1.7287</td>
<td>1.7283</td>
<td>1.7308</td>
</tr>
<tr>
<td></td>
<td>±0.0007</td>
<td>±0.0004</td>
<td>±0.0011</td>
<td>–</td>
</tr>
</tbody>
</table>
The Averaging Principle

On a small interval \([t, t + \delta]\),
The Averaging Principle

On a small interval \([t, t + \delta]\),

- the number of arrivals routed to level \(k\)

\[
\frac{1}{\delta} \int_{t}^{t+\delta} 1_{\{i^*_n(s-)=k\}} d\bar{\Lambda}^n(s)
\]
The Averaging Principle

On a small interval \([t, t + \delta]\),

- the number of arrivals routed to level \(k\)
  \[
  \frac{1}{\delta} \int_{t}^{t+\delta} 1\{i_n^*(s-) = k\} d\bar{\Lambda}^n(s)
  \]

- the amount of fluid routed to level \(k\)
  \[
  f_k(z(t), \lambda(t)) \lambda(t)
  \]
The Averaging Principle

On a small interval \([t, t + \delta]\),

- the number of arrivals routed to level \(k\)
  \[
  \frac{1}{\delta} \int_{t}^{t+\delta} 1\{i^*_n(s^-) = k\} d\bar{\Lambda}^n(s)
  \]

- the amount of fluid routed to level \(k\)
  \[
  f_k(z(t), \lambda(t)) \lambda(t)
  \]

The averaging principle

\[
\lim_{\delta \to 0} \lim_{n \to \infty} \frac{1}{\delta} \int_{t}^{t+\delta} 1\{i^*_n(s^-) = k\} d\bar{\Lambda}^n(s) = f_k(z(t), \lambda(t)) \lambda(t).
\]
The Averaging Principle

The interplay here between the $i^n_*(t)$ and $\bar{Z}^n(t)$:

- The process $\bar{Z}^n(t)$ evolves slowly and determines the transition rates for $i^n_*(t)$.
- The process $i^n_*(t)$ evolves quickly and its “steady state” determines the evolution of $\bar{Z}^n(t)$. 
The Averaging Principle

The interplay here between the $i_n^*(t)$ and $\bar{Z}^n(t)$:

- The process $\bar{Z}^n(t)$ evolves slowly and determines the transition rates for $i_n^*(t)$.
- The process $i_n^*(t)$ evolves quickly and its “steady state” determines the evolution of $\bar{Z}^n(t)$.

To see this intuitively,

$$\frac{1}{\delta} \int_t^{t+\delta} 1\{i_n^*(s-) = k\} \lambda ds = \frac{1}{n\delta} \int_0^{n\delta} 1\{i_n^*(t+\frac{s}{n} -) = k\} \lambda ds.$$

When $n$ becomes large, what determines the above integral is actually the “steady state” of the process $i_n^*(t + \frac{s}{n})$. 
Define the random measure $\nu^n$ by

$$
\nu^n([0, t] \times A) = \int_0^t 1\{Z^n(s-) \in A\} ds,
$$

for any $t > 0$ and subset $A \subset \bar{Z}_K^+$. 

Random Measure and Martingale Representation
Define the random measure $\nu^n$ by

$$
\nu^n([0, t] \times A) = \int_0^t 1\{Z^n(s-) \in A\} ds,
$$

for any $t > 0$ and subset $A \subset \mathbb{Z}_+^K$.

Let $A_k = \{z \in \mathbb{Z}_+^K : z_k > 0 \text{ and } z_j = 0, j < k\}$, then

$$
1\{i^*_n(t-)=k\} = 1\{Z^n(t-) \in A_k\}.
$$
Random Measure and Martingale Representation

Define the random measure $\nu^n$ by

$$\nu^n([0, t] \times A) = \int_0^t 1\{Z^n(s-) \in A\} ds,$$

for any $t > 0$ and subset $A \subset \bar{\mathbb{Z}}^K_+$. Let $A_k = \{z \in \bar{\mathbb{Z}}^K : z_k > 0$ and $z_j = 0, j < k\}$, then

$$1\{i^*_n(t-) = k\} = 1\{Z^n(t-) \in A_k\}.$$

Martingale Representation

$$\bar{M}^n_a(t) = \bar{\Lambda}^n(t) - \int_0^t \bar{\lambda}^n(s) ds,$$

$$\bar{M}^n_k(t) = \frac{1}{n} \left( S^n_k \left( \gamma_k \int_0^t Z^n_k(s) ds \right) - \gamma_k \int_0^t Z^n_k(s) ds \right), \quad k = 1, \ldots, K.$$
\[
\bar{Z}_k^n(t) = \bar{Z}_k^n(0) + \int_0^t 1\{Z^n(s-) \in A_{k-1}\} d\bar{M}_a^n(s) - \int_0^t 1\{Z^n(s-) \in A_k\} d\bar{M}_a^n(s)
\]
\[
- \bar{M}_k^n(t) + \int_0^t 1\{\bar{Q}^n(s-) = 0\} d\bar{M}_{k+1}^n(s)
\]
\[
+ \int_{[0,t] \times A_{k-1}} \bar{\lambda}^n(s) \nu^n(ds \times dy) - \int_{[0,t] \times A_k} \bar{\lambda}^n(s) \nu^n(ds \times dy)
\]
\[
- \gamma_k \int_0^t \bar{Z}_k^n(s) ds + \gamma_{k+1} \int_0^t 1\{\bar{Q}^n(s-) = 0\} \bar{Z}_{k+1}^n(s) ds, \quad 0 < k < K,
\]
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits

$$(ar{Z}^n(t), \bar{Q}^n(t)) \xrightarrow{t \to \infty} (\bar{Z}^\infty, \bar{Z}^\infty)$$
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits

\[(\bar{Z}^n(t), \bar{Q}^n(t)) \rightarrow (\bar{Z}^\infty, \bar{Z}^\infty) \quad \text{as} \quad t \rightarrow \infty, n \rightarrow \infty\]

\[(z(t), q(t)) \rightarrow (z^\infty, q^\infty) \quad \text{as} \quad t \rightarrow \infty, n \rightarrow \infty\]
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits

\[
\left( \bar{Z}^n(t), \bar{Q}^n(t) \right) \quad t \to \infty \quad \Rightarrow \quad \left( \bar{Z}^n, \bar{Z}^n \right)
\]

\[
(n \to \infty) \quad \Rightarrow \quad (z^n(t), q^n(t)) \quad t \to \infty \quad \Rightarrow \quad (z^n, q^n)
\]

Instant Messaging based Services Centers / J. Zhang
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits

\[
(\bar{Z}^n(t), \bar{Q}^n(t)) \xrightarrow{n \to \infty} (z(\infty), q(\infty)) \quad t \to \infty
\]

\[
(\bar{Z}_\infty^n, \bar{Z}_\infty^n) \xrightarrow{n \to \infty} (z_\infty, q_\infty)
\]
Future Research

- Customer Abandonment
- Inefficient levels
- Interchange of Limits
- Dynamic control on *finite* horizon with *time-varying* arrival
- Diffusion approximation
Questions?
Thank you!