
**Multiclass Many Server Diffusion Models:
Reduction to a One Dimensional
Control Problem**

ORSIS, 2007

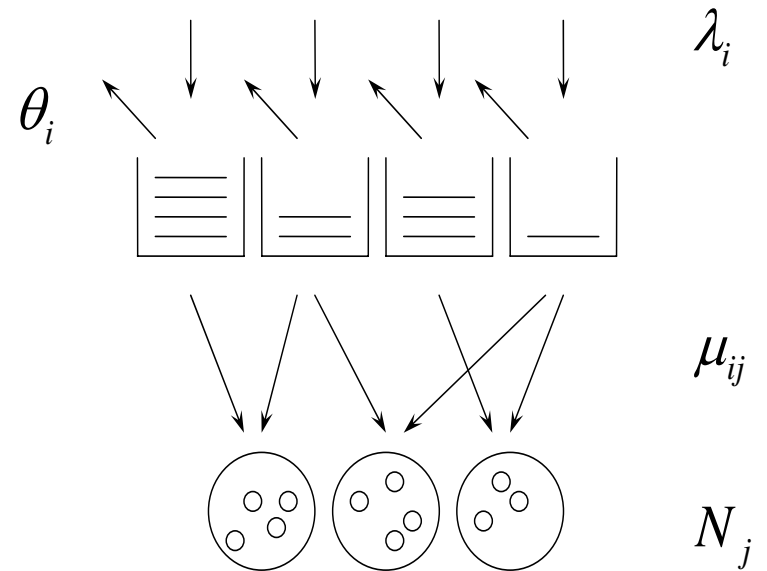
Gennady Shaikhet

Technion, Israel

with R. Atar and A. Mandelbaum

Queueing Model

- $I \geq 1$ customer classes
- $J \geq 1$ service stations
- Arrivals for class i :
renewal processes, rate λ_i
- Servers in station j :
 N_j (stat. identical)
- Service of class- i by server- j :
exponential, rate μ_{ij}
- Abandonments for class i :
exponential clocks, rate θ_i

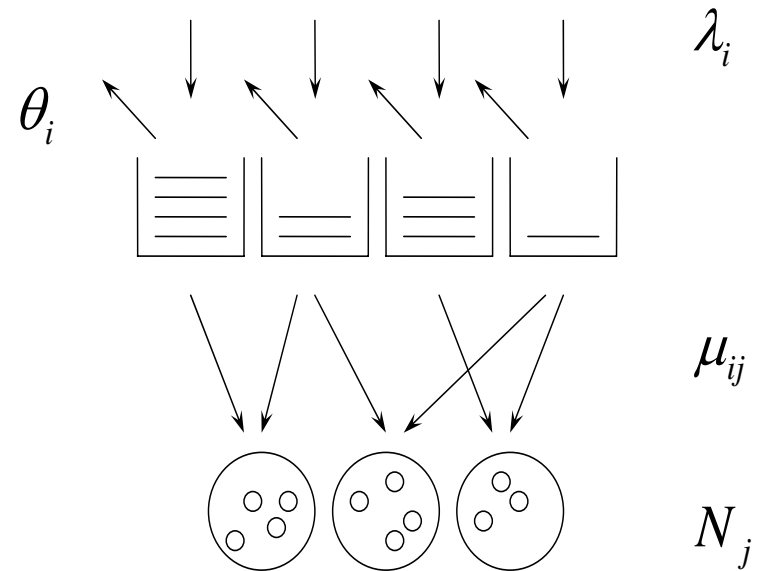


Queueing Model

- $I \geq 1$ customer classes
- $J \geq 1$ service stations
- Arrivals for class i :
renewal processes, rate λ_i
- Servers in station j :
 N_j (stat. identical)
- Service of class- i by server- j :
exponential, rate μ_{ij}
- Abandonments for class i :
exponential clocks, rate θ_i
- Control: has to be specified to complete the description:

Routing customers

Scheduling servers



Optimization Problem

- Given a cost, we face a stochastic control problem, which is **impossible** to solve in general (Not Markovian, high dimensional).

Optimization Problem

- Given a cost, we face a stochastic control problem, which is **impossible** to solve in general (Not Markovian, high dimensional).
- Consider the **heavy traffic** regime, in which the **number of servers** at each station and the **arrival rates grow without bound**, while keeping a **critically loaded system**.

Optimization Problem

- Given a cost, we face a stochastic control problem, which is **impossible** to solve in general (Not Markovian, high dimensional).
- Consider the **heavy traffic** regime, in which the **number of servers** at each station and the **arrival rates grow without bound**, while keeping a **critically loaded system**.
- Expect the system to be always **busy**, but **stable**, on the Law-of-Large-Numbers level - **fluid level**.

Optimization Problem

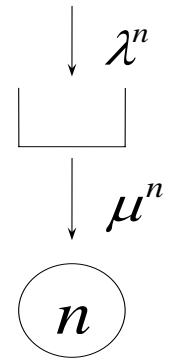
- Given a cost, we face a stochastic control problem, which is **impossible** to solve in general (Not Markovian, high dimensional).
- Consider the **heavy traffic** regime, in which the **number of servers** at each station and the **arrival rates grow without bound**, while keeping a **critically loaded system**.
- Expect the system to be always **busy**, but **stable**, on the Law-of-Large-Numbers level - **fluid level**.
- with stochastic fluctuations around the fluid - **diffusion level**.

Optimization Problem

- Given a cost, we face a stochastic control problem, which is **impossible** to solve in general (Not Markovian, high dimensional).
- Consider the **heavy traffic** regime, in which the **number of servers** at each station and the **arrival rates grow without bound**, while keeping a **critically loaded system**.
- Expect the system to be always **busy**, but **stable**, on the Law-of-Large-Numbers level - **fluid level**.
- with stochastic fluctuations around the fluid - **diffusion level**.
- Then to control the system dynamically on the diffusion level.

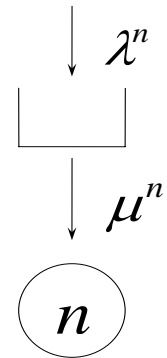
Single buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.



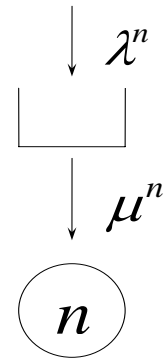
Single buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$



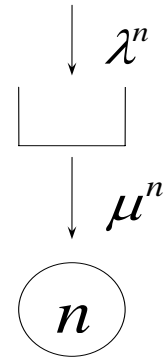
Single buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes **critically loaded**:
utilization = $\frac{\lambda^n}{n\mu^n} \uparrow 1$.



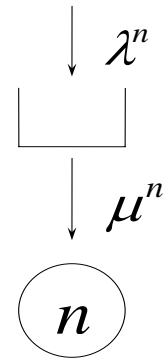
Single buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes **critically loaded**:
utilization = $\frac{\lambda^n}{n\mu^n} \uparrow 1$.
- Expect **fluctuations** of order $O(\sqrt{n})$ around "average" = n .



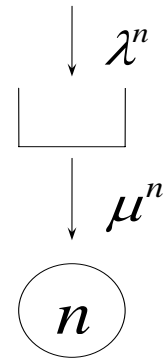
Singe buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes **critically loaded**:
utilization = $\frac{\lambda^n}{n\mu^n} \uparrow 1$.
- Expect **fluctuations** of order $O(\sqrt{n})$ around "average" = n .
- Define $X^n(t)$ = number of customers in the system at time $t \geq 0$



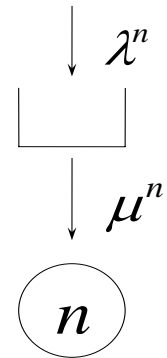
Singe buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes **critically loaded**:
utilization = $\frac{\lambda^n}{n\mu^n} \uparrow 1$.
- Expect **fluctuations** of order $O(\sqrt{n})$ around "**average**" = n .
- Define $X^n(t)$ = number of customers in the system at time $t \geq 0$
- Introduce **centered** and **rescaled** process $\hat{X}^n(t) = \frac{X^n(t) - n}{\sqrt{n}}$.



Singe buffer - single pool. Heavy Traffic

- Consider a sequence of $M/M/n$ models, indexed by $n \uparrow \infty$.
- Assume $\lambda^n = n\lambda + O(\sqrt{n})$, $n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes **critically loaded**:
utilization = $\frac{\lambda^n}{n\mu^n} \uparrow 1$.
- Expect **fluctuations** of order $O(\sqrt{n})$ around "**average**" = n .
- Define $X^n(t)$ = number of customers in the system at time $t \geq 0$
- Introduce **centered** and **rescaled** process $\hat{X}^n(t) = \frac{X^n(t) - n}{\sqrt{n}}$.
- **Thm.(Halfin - Whitt, 1981)**: \hat{X}^n converges weakly to a diffusion.



$$X(t) = X(0) + \int_0^t b(X(s))ds + \sigma W(t).$$

Many-to-Many. Heavy Traffic

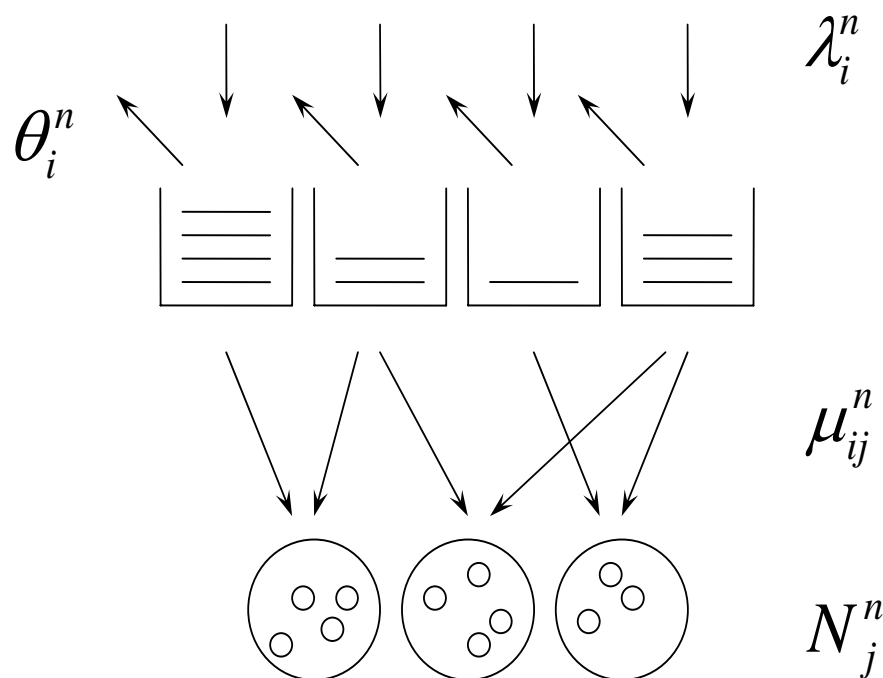
- Consider a sequence of systems, indexed by $n \uparrow \infty$

- $\lambda_i^n = n\lambda_i + O(\sqrt{n})$

- $\theta_i^n \equiv \theta_i$

- $n\mu_{ij}^n = n\mu_{ij} + O(\sqrt{n})$

- $N_j^n = n\nu_j + O(\sqrt{n})$



Many-to-Many. Heavy Traffic

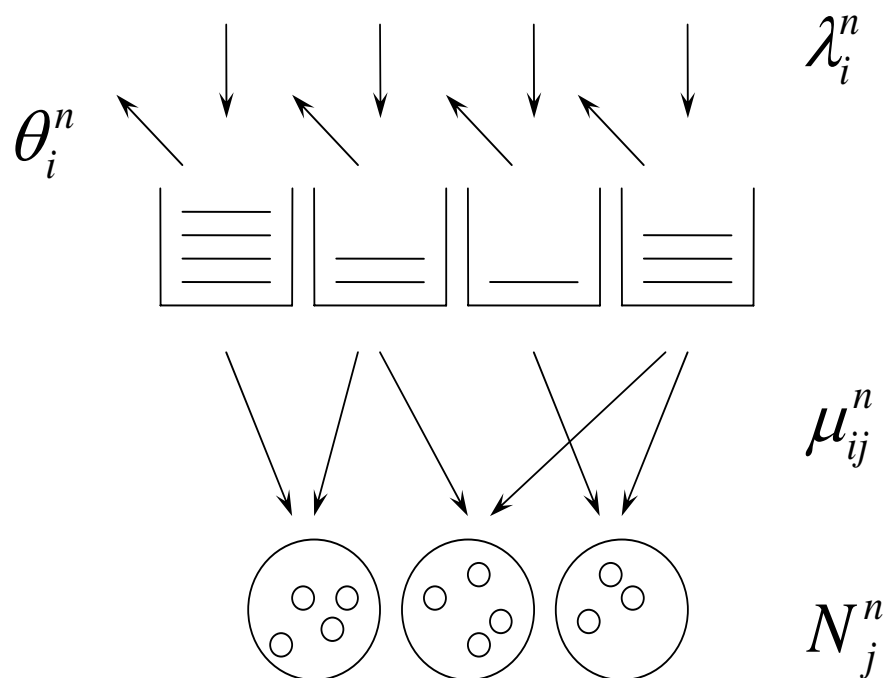
- Consider a sequence of systems, indexed by $n \uparrow \infty$

- $\lambda_i^n = n\lambda_i + O(\sqrt{n})$

- $\theta_i^n \equiv \theta_i$

- $n\mu_{ij}^n = n\mu_{ij} + O(\sqrt{n})$

- $N_j^n = n\nu_j + O(\sqrt{n})$



- All stations should be **busy** on the **fluid (order n) level**.

Heavy Traffic. Fluid View

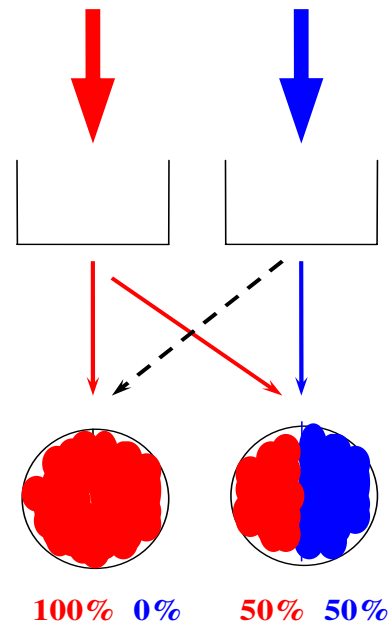
- An example of critically loaded system:

$$\lambda_1 = 7.5, \quad \lambda_2 = 2$$

$$\mu_{11} = 4, \quad \mu_{12} = 7$$

$$\mu_{21} = 2, \quad \mu_{22} = 4$$

$$\nu_1 = 1, \quad \nu_2 = 1$$



Static fluid allocation:

$$\psi_{11}^* = 1, \quad \psi_{12}^* = 0.5, \quad \Rightarrow \quad \lambda_1 = 1 \cdot \mu_{11} + 0.5 \cdot \mu_{12}$$

$$\psi_{21}^* = 0, \quad \psi_{22}^* = 0.5, \quad \Rightarrow \quad \lambda_2 = 0 \cdot \mu_{21} + 0.5 \cdot \mu_{22}$$

$$x_1^* = \psi_{11}^* + \psi_{12}^* = 1.5$$

$$x_2^* = \psi_{21}^* + \psi_{22}^* = 0.5$$

Diffusion Scaling

Introduce the processes:

$X_i^n(t)$ = number of class- i customers in the system at time t ,

$Y_i^n(t)$ = number of class- i customers in the queue at time t ,

$Z_j^n(t)$ = number of idle servers in station j at time t ,

$\Psi_{ij}^n(t)$ = number of class- i customers in service in station j at time t ,

Scale them around the **static fluid**: ψ_{ij}^* and x_i^* :

$$\hat{X}_i^n(t) = n^{-1/2}(X_i^n(t) - nx_i^*), \quad \hat{\Psi}_{ij}^n(t) = n^{-1/2}(\Psi_{ij}^n(t) - n\psi_{ij}^*).$$

$$\hat{Y}_i^n(t) = n^{-1/2}Y_i^n(t), \quad \hat{Z}_i^n(t) = n^{-1/2}Z_i^n(t).$$

Diffusion Model

Take the formal weak limits to get a **diffusion model** \mathcal{M} :

\mathcal{M} - a set of processes (X, Y, Z, Ψ) , satisfying:

$$X_i(t) = x_i + W_i(t) - \sum_{j \in \mathcal{J}} \mu_{ij} \int_0^t \Psi_{ij}(s) ds - \theta_i \int_0^t Y_i(s) ds, \quad i \in \mathcal{I}.$$

$$\sum_{j \in \mathcal{J}} \Psi_{ij}(t) = X_i(t) - Y_i(t), \quad i \in \mathcal{I},$$

$$\sum_{i \in \mathcal{I}} \Psi_{ij}(t) = -Z_j(t), \quad j \in \mathcal{J}.$$

$$Y_i(t) \geq 0, \quad Z_j(t) \geq 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad t \geq 0.$$

Pool-Dependent Service Rates

Let $\mu_{ij} = \mu_j$.

- A subset $\widetilde{\mathcal{M}}$ of \mathcal{M} is called a **reduction of \mathcal{M}** if for any $(X, Y, Z, \Psi) \in \mathcal{M}$ there exists $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \widetilde{\mathcal{M}}$, such that for any constant $c \in R_+^I$, a.s. holds:

$$c \cdot \widetilde{Y}(t) \leq c \cdot Y(t) \text{ for all } t \geq 0.$$

Pool-Dependent Service Rates

Let $\mu_{ij} = \mu_j$.

- A subset $\widetilde{\mathcal{M}}$ of \mathcal{M} is called a **reduction of \mathcal{M}** if for any $(X, Y, Z, \Psi) \in \mathcal{M}$ there exists $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \widetilde{\mathcal{M}}$, such that for any constant $c \in R_+^I$, a.s. holds:

$$c \cdot \widetilde{Y}(t) \leq c \cdot Y(t) \text{ for all } t \geq 0.$$

- **Theorem:** a subset $\widetilde{\mathcal{M}}$ of \mathcal{M} , satisfying, in addition,

$$Z(t) = Z_e(t)e_{j_0} \text{ for } j_0 = \arg \min_j \{\mu_j\}$$

is a reduction of \mathcal{M} .

We use $x_e = \sum_i x_i$ and e_i - a unit coordinate vector

Pool-Dependent Service Rates

- Consider the family \mathcal{O} of processes $(\check{X}, u) \in R \times U$, where

$$\check{X}(t) = x_e + W_e(t) + \mu_{j_0} \int_0^t \check{X}^-(s) ds - \int_0^t [\theta \cdot u(s)] \check{X}^+(s) ds, \quad t \geq 0;$$

$$U = \{u \in R^I : u_i \geq 0, u_e = 1\}.$$

Pool-Dependent Service Rates

- Consider the family \mathcal{O} of processes $(\check{X}, u) \in R \times U$, where

$$\check{X}(t) = x_e + W_e(t) + \mu_{j_0} \int_0^t \check{X}^-(s) ds - \int_0^t [\theta \cdot u(s)] \check{X}^+(s) ds, \quad t \geq 0;$$

$$U = \{u \in R^I : u_i \geq 0, u_e = 1\}.$$

- Theorem:** Let $\mathcal{M}^{\mathcal{O}}$ be a set of processes (X, Y, Z, Ψ) such that a.s. for all $t \geq 0$

$$X(t) = x + W(t) + \int_0^t H(X(s), u(s)) ds, \quad X \in R^I$$

$$Y(t) = X_e^+(t)u(t), \quad Z(t) = X_e^-(t)e_{j_0} \quad \Psi_{ij}(t) = G_{ij}(X(t) - Y(t), -Z(t))$$

where u is such that $(\check{X}, u) \in \mathcal{O}$; (H and G are known functions). Then

- $\mathcal{M}^{\mathcal{O}} = \widetilde{\mathcal{M}}$.
- For any $(\check{X}, u) \in \mathcal{O}$ and a corresponding $(X, Y, Z, \Psi) \in \mathcal{M}^{\mathcal{O}}$ we have a.s. $X_e = \check{X}$.

Reduction of a Control Problem

- A multi-dimensional control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} L[Y(t)] dt, \quad x \in R^I.$$

$$X(t) = x + W(t) + \int_0^t H(X(s), u(s)) ds$$

$$Y(t) = X_e^+(t)u(t), \quad Z(t) = X_e^-(t)e_{j_0} \quad \Psi_{ij}(t) = G_{ij}(X(t) - Y(t), -Z(t))$$

Reduction of a Control Problem

- A **multi-dimensional** control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} L[Y(t)] dt, \quad x \in R^I.$$

$$X(t) = x + W(t) + \int_0^t H(X(s), u(s)) ds$$

$$Y(t) = X_e^+(t)u(t), \quad Z(t) = X_e^-(t)e_{j_0} \quad \Psi_{ij}(t) = G_{ij}(X(t) - Y(t), -Z(t))$$

- is equivalent to a **1-dimensional** control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} L[\check{X}^+(t)u(t)] dt, \quad x \in R.$$

$$\check{X}(t) = x_e + W_e(t) + \mu_{j_0} \int_0^t \check{X}^-(s) ds - \int_0^t [\theta \cdot u(s)] \check{X}^+(s) ds, \quad t \geq 0;$$

1 - dim Diffusion Control Problem

- Consider the following stochastic control problem:

$$\inf_{\Pi} E_x^{\pi} \int_0^{\infty} e^{-\gamma t} [c \cdot u(t)] X^+(t) dt, \quad x \in R.$$

$$X(t) = x_e + W_e(t) + \mu_{min} \int_0^t X^-(s) ds - \int_0^t [\theta \cdot u(s)] X^+(s) ds, \quad t \geq 0;$$

1 - dim Diffusion Control Problem

- Consider the following stochastic control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} [c \cdot u(t)] X^+(t) dt, \quad x \in R.$$

$$X(t) = x_e + W_e(t) + \mu_{min} \int_0^t X^-(s) ds - \int_0^t [\theta \cdot u(s)] X^+(s) ds, \quad t \geq 0;$$

- Assume that $\theta_i = \theta$ for all i . Then, for $i_0 = \arg \min_i \{c_i\}$,

$$u_{opt}(t) \equiv e_{i_0}$$

1 - dim Diffusion Control Problem

- Consider the following stochastic control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} [c \cdot u(t)] X^+(t) dt, \quad x \in R.$$

$$X(t) = x_e + W_e(t) + \mu_{min} \int_0^t X^-(s) ds - \int_0^t [\theta \cdot u(s)] X^+(s) ds, \quad t \geq 0;$$

- Assume that $\theta_i = \theta$ for all i . Then, for $i_0 = \arg \min_i \{c_i\}$,

$$u_{opt}(t) \equiv e_{i_0}$$

- Assume that $\theta_1 \leq \theta_2 \leq \dots \leq \theta_I$ and $c_1 \geq c_2 \geq \dots \geq c_I$. Then

$$u_{opt}(t) \equiv e_I$$

A Connection to Prelimit Model

Conjecture: the following is an asymptotically optimal policy:

- **Routing:** each arriving customer, if not queued, is **directed to the fastest server available**, otherwise stays in queue. The system "tends" to have idle servers only in station j_0 (recall $Z_{j_0} = Z_e$)

A Connection to Prelimit Model

Conjecture: the following is an asymptotically optimal policy:

- **Routing:** each arriving customer, if not queued, is **directed to the fastest server available**, otherwise stays in queue. The system "tends" to have idle servers only in station j_0 (recall $Z_{j_0} = Z_e$)
- **Scheduling:** a newly available agent, that can serve class i_0 , **will accept a waiting i_0 -class customer, only if no other classes are waiting** for him. All other situations are resolved arbitrarily. In other words, class i_0 always has the lowest priority, and thus the system will seek to have waiting customers only in class i_0 (recall $Y_{i_0} = Y_e$).

1 - dim Diffusion Control Problem

- Consider the following stochastic control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} \sum_i C_i(u_i(t)X^+(t)) dt, \quad x \in R.$$

$$X(t) = x_e + W_e(t) + \mu_{min} \int_0^t X^-(s) ds - \int_0^t [\theta \cdot u(s)] X^+(s) ds, \quad t \geq 0;$$

$C_i(\cdot)$ are strictly increasing smooth convex functions.

1 - dim Diffusion Control Problem

- Consider the following stochastic control problem:

$$\inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} \sum_i C_i(u_i(t)X^+(t)) dt, \quad x \in R.$$

$$X(t) = x_e + W_e(t) + \mu_{min} \int_0^t X^-(s) ds - \int_0^t [\theta \cdot u(s)] X^+(s) ds, \quad t \geq 0;$$

$C_i(\cdot)$ are strictly increasing smooth convex functions.

- Then optimal controls are characterized by

$$C'_1(u_1^o(t)X^+(t)) = C'_2(u_2^o(t)X^+(t)) = \dots = C'_I(u_I^o(t)X^+(t))$$

A Connection to Prelimit Model

Conjecture: the following is an asymptotically optimal policy:

- **Routing:** each arriving customer, if not queued, is **directed to the fastest server available**, otherwise stays in queue. The system "tends" to have idle servers only in station j_0 (recall $Z_{j_0} = Z_e$)

A Connection to Prelimit Model

Conjecture: the following is an asymptotically optimal policy:

- **Routing:** each arriving customer, if not queued, is **directed to the fastest server available**, otherwise stays in queue. The system "tends" to have idle servers only in station j_0 (recall $Z_{j_0} = Z_e$)
- **Scheduling:** at any time t , among all the waiting customers that are available for him, a newly available agent at station j will **serve a customer from class $i^* = \arg \max_{i \sim j} \{C'_i(Y_i^n(t))\}$** . The system will seek to achieve approximate equality in

$$C'_1(Y_1^n(t)) = C'_2(Y_2^n(t)) = \dots = C'_I(Y_I^n(t))$$