Queueing Systems with Many Servers: Null Controllability in Heavy Traffic

EE, Probability Seminar

December 13, 2005

Gennady Shaikhet

Joint work with R. Atar and A. Mandelbaum

Contents:

1. Critical Loading
2. Basic and Non-basic Activities
3. Asymptotic Null Controllability

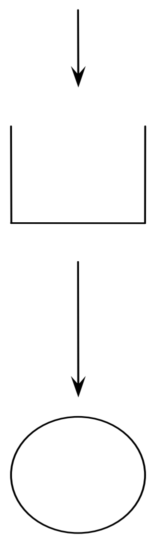
Part I

Critical Loading

Overview of Results

Background: Halfin - Whitt Theorem

Consider a queueing model $M/M/n$:



- One class of customers, one service station
- Arrivals: Poisson, rate λ
- Servers in station: n servers (i.i.d.)
- Service time: exponential, rate μ

Halfin - Whitt (cont.)

Consider a sequence of $M/M/n$ systems, indexed by n.

As $n \uparrow \infty$, assume $\mu^n \sim \mu, \quad \lambda^n \sim n\mu$.

As $n \uparrow \infty$, the system becomes <u>critically loaded</u>, i.e.:

utilization (fraction of server's busy time) : $\dfrac{\lambda^n}{n\mu^n} \to 1.$

For <u>diffusion approximations</u> assume

$$\frac{\lambda^n}{n\mu^n} \sim 1 - \frac{\beta}{\sqrt{n}}, \quad \text{for some } \beta > 0.$$

Let $X^n(t)$ = total number of customers at time t.

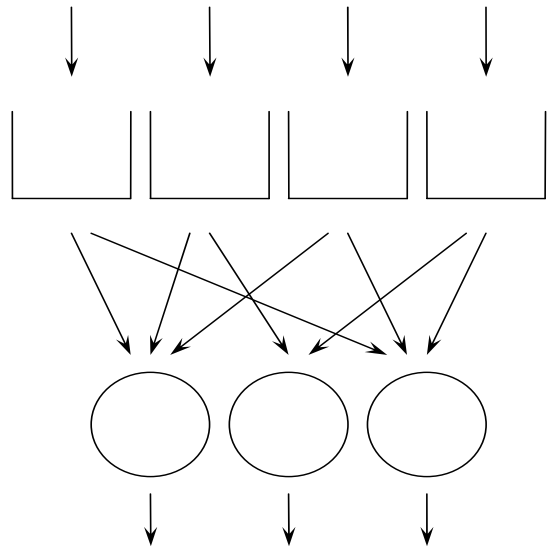
Define the centered and normalized process:

$$\hat{X}^n(t) = \frac{X^n(t) - n}{\sqrt{n}}, \quad t \geq 0.$$

Theorem (H. & W., 1981): *Assume* $\hat{X}^n(0) \xrightarrow{d} X(0)$. *Then* $\hat{X}^n \Rightarrow X$ *in* $D([0,\infty))$, *where* X *is a diffusion:*

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

Parallel Queueing Model



- Classes: $\mathcal{I} = \{1, ..., I\}$, $I \geq 1$
- Station: $\mathcal{J} = \{1, ..., J\}$, $J \geq 1$
- Arrivals: renewal processes A_i, rate λ_i, $i \in \mathcal{I}$
- Servers per station: N_j in station $j \in \mathcal{J}$
- Service: of class-i by type-j server, exponential, rate μ_{ij}

In order to describe the system completely, specify <u>control</u>:

Routing customers and *Scheduling servers*

Average Behavior. Fluid View

Consider a sequence of queueing systems indexed by $n \uparrow \infty$.

• First look at the average behavior (fluid view).

Assume $\lambda_i^n \sim n\lambda_i \qquad \mu_{ij}^n \sim \mu_{ij} \qquad N_j^n \sim n\nu_j.$

Consider the corresponding *fluid model*, where

• Arrivals and service processes are deterministic flows with the corresponding rates λ_i and μ_{ij}.

• ν_j - total server capacity of station j ("number of servers").

Optimize the fluid model, by statically allocating the incoming fluid among the service stations.

Choose ξ_{ij} - the fraction of ν_j, constantly dedicated to class i.

Static allocation problem [Harrison & Lopez (1999)]: choose an allocation matrix (ξ_{ij}) and a scalar ρ to

$$\mathrm{Min}\Big\{\rho \;:\; \sum_{j\in\mathcal{J}} \mu_{ij}\,\nu_j\,\xi_{ij} = \lambda_i, \quad \sum_{i\in\mathcal{I}} \xi_{ij} \le \rho, \quad \xi_{ij} \ge 0\Big\}.$$

Heavy Traffic condition (HT):

There exists a unique optimal solution (ξ^*, ρ^*) to the linear program. Moreover, $\rho^* = 1$ and $\sum_{i\in\mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$.

Parameters (second order)

Static Allocation Problem gives us

$\psi^*_{ij} = \xi^*_{ij}\nu_j$ - the amount of fluid i in station j.

$x^*_i = \sum_j \psi^*_{ij}$ - the total amount of fluid i in process.

• For the original model expect the averages of corresponding quantities to be nx^* and $n\psi^*_{ij}$.

For the diffusion approximations assume

$$N^n_j = n\nu_j + o(\sqrt{n})$$

$$\lambda^n_i = n\lambda_i + \hat{\lambda}_i\sqrt{n} + o(\sqrt{n})$$

$$n\mu^n_{ij} = n\mu_{ij} + \hat{\mu}_{ij}\sqrt{n} + o(\sqrt{n})$$

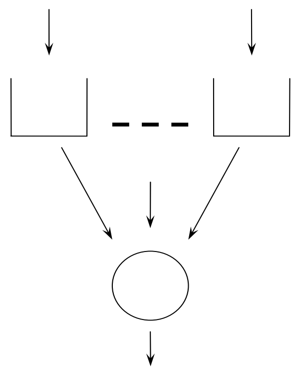
where λ, μ and ν satisfy HT conditions.

• For our queueing model expect stochastic fluctuations of $O(\sqrt{n})$ around the average.

• Meaningless to talk about diffusion optimization (order $O(\sqrt{n})$), without the optimally allocated fluid (order $O(n)$).

Diffusion Model: Many - to - One

Atar, Mandelbaum and Reiman (2004), Harrison and Zeevi (2004).



Define X_i^n = number of class-i customers in the system.

Perform centering *around fluid* and rescaling.

As $n \uparrow \infty$, the original model gives rise (under formal limits) to a controlled diffusion model:

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

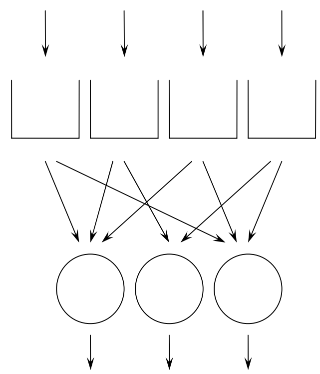
• Controlled process: $X \in \mathbb{R}^I$. Control: $U \in \mathbb{R}^I$.

• Given the cost, obtain stochastic control problem with drift control.

• Optimal control of the diffusion gives rise to *asymptotically* optimal scheduling for queueing model.

• Rigorous result is not weak convergence of the processes but convergence of optimal cost.

Diffusion Model: Many - to - Many



After scaling and centering about fluid, take formal limits as $n \to \infty$ to get surprisingly new kind of a diffusion model:

$$X(t) = X(0) + \sigma W(t) + \int_0^t b(X(s), U(s))ds + \sum_{c \in \mathcal{C}} m_c \eta_c(t).$$

- Controlled process: $X \in \mathbb{R}^I$. Control: (U, η).
- $\mathcal{C}$ - finite set. $m_c \in \mathbb{R}^I$ - constant vectors, depend on μ_{ij}.
- For each c, $\eta_c \in \mathbb{R}$ is nondecreasing with $\eta_c(0) \geq 0$.

• The "singular" term is due to our many servers limit. Do not confuse with the singular term in the classical heavy-traffic, which is due to a necessity to be constrained to a certain domain, hence, reflection.

• Before it was expected to get only drift control as in Atar (2005), Borkar (2005).

Constraining the Diffusion

Consider the diffusion model:

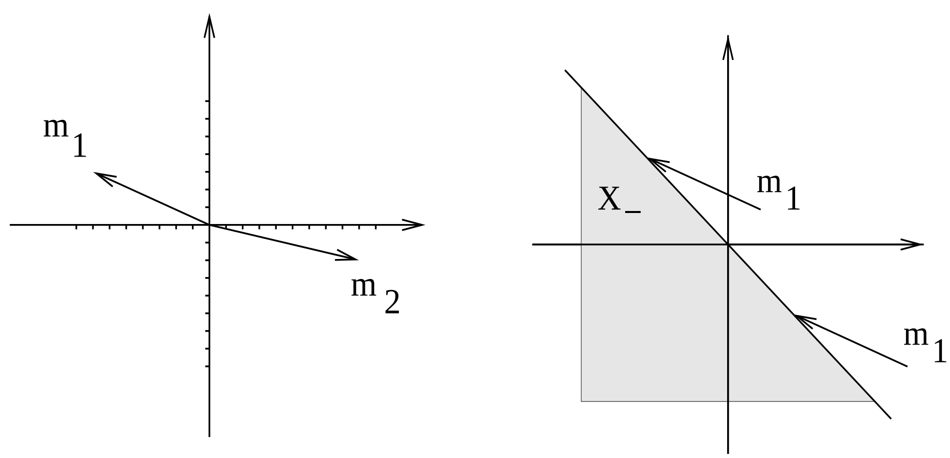
$$X(t) = X(0) + \sigma W(t) + \int_0^t b(X(s), U(s))ds + \sum_{c \in \mathcal{C}} m_c \eta_c(t).$$

The singular term η can restrict X to a certain closed domain.

It can happen that X can be restricted to a domain, corresponding to all queues being empty.

Here $X_- = \{x \in \mathbb{R}^I, \;\; e \cdot x \leq 0\}$.

Lemma 1 *Let $e \cdot m_c < 0$ for some c. Then there exists a control (U, η) under which $e \cdot X(t) \leq 0$, on $[0, \infty]$, P-a.s.*



Sketch of Proof: standard results on diffusion with oblique reflection.

Main Results (overview)

Consider a sequence of systems in heavy traffic.

Let $e \cdot m_c < 0$ for some c. Then the following effect happens (**null controllability**) :

For any finite interval $(0,T]$ there is a scheduling policy that keeps all queues empty, in the sense

$$\lim_{n\to\infty} P(\text{no queues in the } n^{th} \text{ system during } (0,T]) = 1.$$

We find the strategies for two different types of control

PREEMPTIVE REGIME:

a service to a customer can be interrupted and resumed at later time (possibly in a different station).

NON-PREEMPTIVE REGIME :

a service to a customer can not be interrupted before it is completed.

Critically loaded **system behaves like underloaded!!!**

Part II

Intuition and Explanations

Basic and Non-basic Activities

Basic and Non-basic Activities

<u>Heavy Traffic condition (HT)</u>:

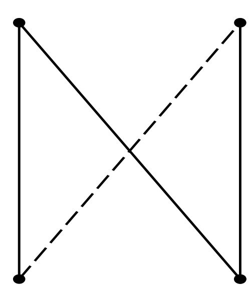
There exists a unique optimal solution (ξ^*, ρ^*) *to the linear program. Moreover,* $\rho^* = 1$ *and* $\sum_{i\in\mathcal{I}} \xi^*_{ij} = 1$, *for all* $j \in \mathcal{J}$.

- Basic activities ($\mathcal{BA}$): pairs (i,j), s.t. $\xi^*_{ij} > 0$.
- Non-basic activities ($\mathcal{NBA}$): pairs (i,j) s.t. $\xi^*_{ij} = 0$.

Example:

$$\nu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \lambda = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \mu = \begin{pmatrix} 3/4 & 1/3 \\ 8 & 4 \end{pmatrix}.$$

One checks that HT condition holds, with $\xi^* = \begin{pmatrix} 1 & 0.75 \\ 0 & 0.25 \end{pmatrix}$



The activity $(2,1)$ is non basic.

<u>Fact</u>: HT $\Rightarrow$ $\mathcal{BA}$ constitute a union of disjoint trees.

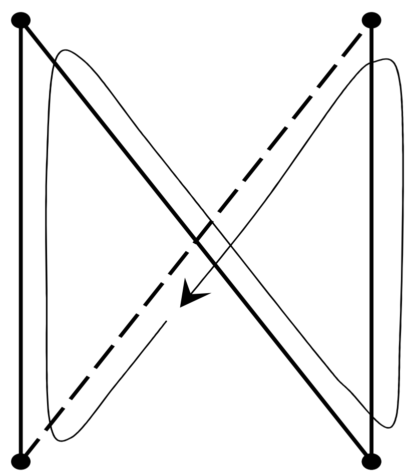
<u>Assumption</u>: the graph of $\mathcal{BA}$ is a tree.

Non-basic Activities (cont.)

Example (cont.):

For $\varepsilon > 0$ small enough make a slight perturbation to the static allocation ξ^*_{ij} by introducing

$$\widetilde{\xi}_{ij} = \xi^*_{ij} + \delta_{ij}, \quad \text{where } \delta = \begin{pmatrix} -\varepsilon & \varepsilon \\ \varepsilon & -\varepsilon \end{pmatrix}.$$



$$\text{Total processing rate} = \sum_{ij} \mu_{ij} \widetilde{\xi}_{ij}$$
$$= \xi^*_{11}\mu_{11} + \xi^*_{12}\mu_{12} + \xi^*_{21}\mu_{21} + \xi^*_{22}\mu_{22} + \Delta$$

where $\Delta := (8 - 3/4 + 1/3 - 4)\,\varepsilon > 0$

- The fluid "mass" was cyclically transferred.
- **The throughput was increased!**

When do m_C's arise? Simple Cycles

Denote the set of simple cycles:

$\mathcal{C} = \{$cycles, for which exactly one edge belongs to $\mathcal{NBA}\}$.

NOTE: *one-to-one correspondence between $\mathcal{NBA}$ and $\mathcal{C}$.*

For each cycle $c \in \mathcal{C}$ corresponds a unique vector m_c, s.t.

$$\text{Throughput increased} \quad \Leftrightarrow \quad \text{exists } c \in \mathcal{C} \ \text{ with } \ e \cdot m_c < 0$$

APPLYING THE ABOVE FOR ORIGINAL MODEL

Cyclically transfer the "mass" by putting large amount of customers into non-basic activity, *for a certain period of time.*

As a result, the *rate* of service completions will increase.

NOTE: Effect of above will not be seen in one - server stations.

Part III

Formal Results and Discussions

Asymptotic Null Controllability

Reminder

<u>Main Assumption</u>: there exists $c \in \mathcal{C}$, with $e \cdot m_c < 0$.

Let Y_i^n = number of class-i customers in the queue.

<u>Preemptive regime:</u>

Theorem 1 *Assume that Main Assumption holds and let $0 < \varepsilon < T < \infty$ be given. Then there exist a preemptive strategy under which*

$$\lim_{n \to \infty} P(Y^n(t) = 0 \textit{ for all } t \in [\varepsilon, T]) = 1. \qquad (1)$$

<u>Non-preemptive regime:</u>

Theorem 2 *Assume $I = J = 2$, let Main Assumption hold and let $0 < \varepsilon < T < \infty$ be given. Then there exist non-preemptive strategy under which (1) holds.*

Discussion: Preemptive Regime

The nature of *preemption* makes it possible to mimic the reflection mechanism from the diffusion model.

Denote
$$D_1 = \{\xi \in \mathbb{R}^I : e \cdot \xi < -1\}.$$

Fix throughout a simple cycle c_0 for which $e \cdot m_{c_0} < 0$.

Let $\Psi^n_{c_0}$ = number of customers in the non-basic activity, corresponding to cycle c_0. Set

$$\Psi^n_{c_0}(t) = \begin{cases} 0, & \hat{X}^n(t) \in D_1 \\ n^{5/8}, & \hat{X}^n(t) \in D_1^c \end{cases} \qquad t \geq 0,$$

and $\Psi^n_c(t) = 0$ for all $c \neq c_0$, $t \geq 0$.

• The resulting scaled process $\hat{X}^n$ is tight. This enables to obtain weak convergence of diffusion-level processes.

• The fluctuations are of order $O(n^{1/2})$ around fluid. This is a reason for increasing the non-basic population to a greater power of n, thus to move the system quickly away from the forbidden region.

Discussion: Non-Preemptive Regime

Much more difficult!!!

It is *impossible* to follow the reflection mechanism, because instantaneous changes in the population are impossible

Intuition:

Assume that K servers are currently working in activity (i, j). Let $K \to \infty$.

1. For any $\alpha < 1$, it will take infinitesimal $O(K^{\alpha-1}) = o(1)$ time to serve $O(K^{\alpha})$ customers.
2. But it will take $O(1)$ time to serve $O(K)$ customers!

• In fact, the diffusion model does not describe correctly the asymptotic problem.

• Completely different treatment needed. However, one still uses mass transfer along the cycle c.

• Special routing keeps the population in the non-basic activity of order $O(n^{5/8})$ all the time.

• The resulting scaled process $\hat{X}^n$ is not tight. Thus, stability only for a finite interval.

Present and Future Research

- Null Controllability of Fluid Queue.
- Applications to Staffing.
- General Service Times.
- PDE Analysis (Viscosity Solutions).