Queueing Systems with Many Servers:

Null Controllability in Heavy Traffic

ORSIS, 2006

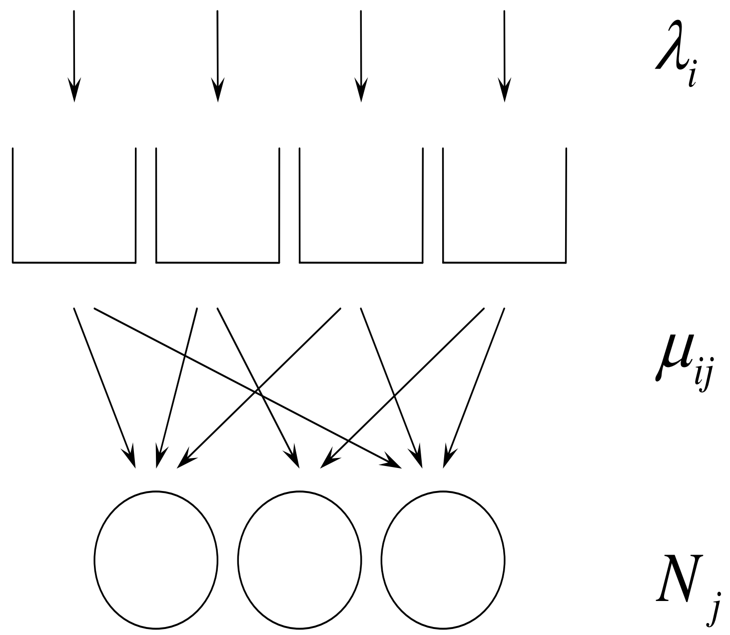
Gennady L. Shaikhet

`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}



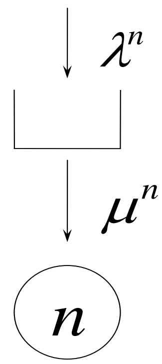
- Control: has to be specified to complete the description:
  Routing customers   Scheduling servers

Optimization Problem

- Given the cost, we face the stochastic control problem, which is impossible to solve in general (not Markov, etc..)
- 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.

One-to-One. Heavy Traffic

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

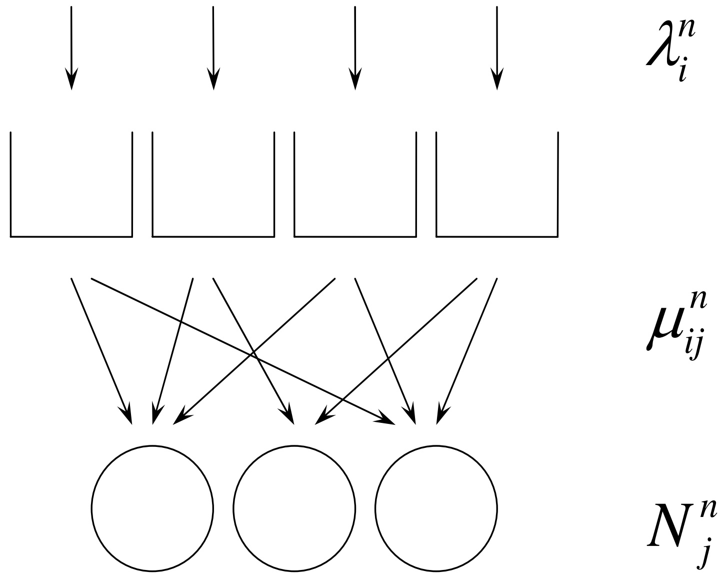


- Assume $\lambda^n = n\lambda + O(\sqrt{n}), \quad n\mu^n = n\mu + O(\sqrt{n})$
- Take $\lambda = \mu$. The system becomes critically loaded: utilization $= \frac{\lambda^n}{n\mu^n} \uparrow 1$.
- Expect fluctuations of order $O(\sqrt{n})$ around "average" = n.
- Define $X^n(t)$ = number of customers 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 the sequence of systems, indexed by $n \uparrow \infty$
- $\lambda_i^n = n\lambda_i + O(\sqrt{n})$
- $n\mu_{ij}^n = n\mu_{ij} + O(\sqrt{n})$
- $N_j^n = n\nu_j + O(\sqrt{n})$
- What is critically loaded ?

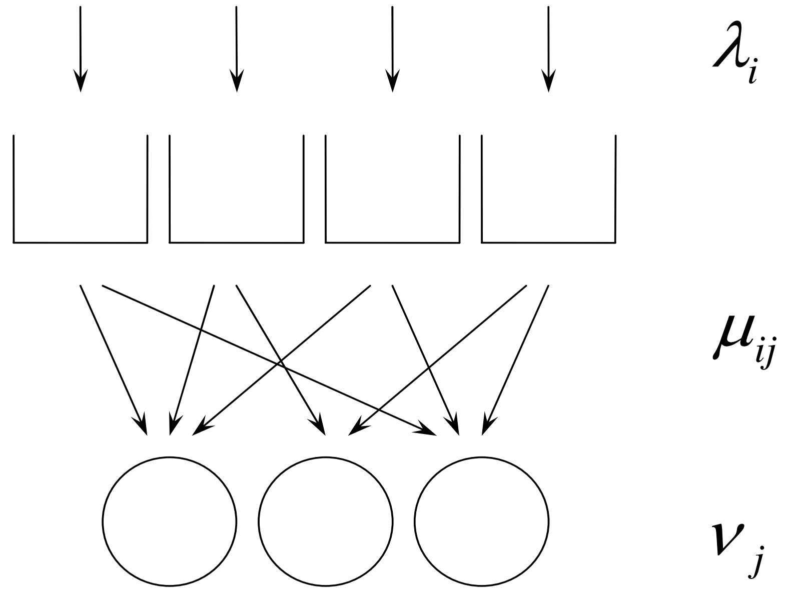


- All stations should be busy on the fluid (order n) level.
- Static fluid analysis is needed.

Critical loading. Fluid View

- Consider the corresponding fluid model, where

- Arrival and Service:
  deterministic, rates λ_i and μ_{ij}



- Server capacity of station j
  ν_j ("number of servers").

- All stations should be fully (though optimally) utilized.

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

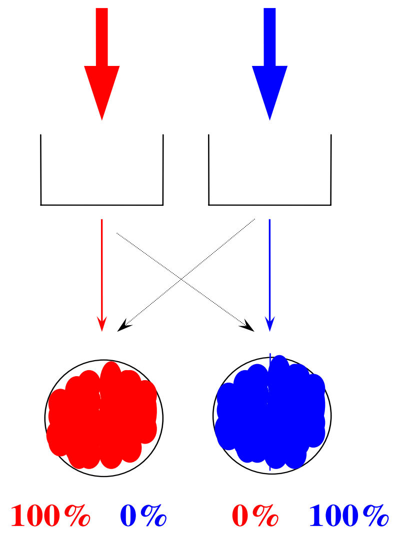
Many-to-Many. Fluid View

- Consider the following fluid system:

  $$\lambda_1 = 2, \quad \lambda_2 = 1$$
  $$\mu_{11} = 2, \quad \mu_{12} = 2$$
  $$\mu_{21} = 2, \quad \mu_{22} = 1$$
  $$\nu_1 = 1, \quad \nu_2 = 1$$

  Allocate the fluid as

  $$\xi_{11} = 1, \quad \xi_{12} = 0$$
  $$\xi_{21} = 0, \quad \xi_{22} = 1$$



  Both classes are processed to completion:

  $$\lambda_1 = 1 \cdot \mu_{11}, \quad \lambda_2 = 1 \cdot \mu_{22}$$

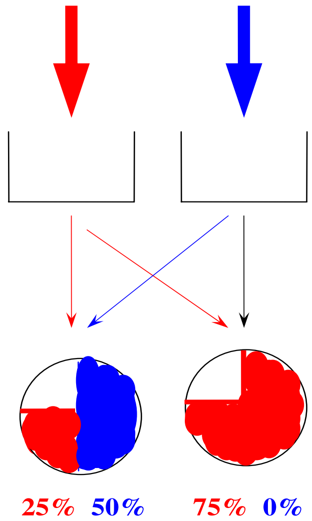
  Utilization of both stations $= 1$.

- Critically loaded system?

Many-to-Many. Fluid View

- Consider the following fluid system:

  $$\lambda_1 = 2, \quad \lambda_2 = 1$$
  $$\mu_{11} = 2, \quad \mu_{12} = 2$$
  $$\mu_{21} = 2, \quad \mu_{22} = 1$$
  $$\nu_1 = 1, \quad \nu_2 = 1$$

  Reallocate the fluid as

  $$\xi_{11} = 0.25, \quad \xi_{12} = 0.75$$
  $$\xi_{21} = 0.5, \quad \xi_{22} = 0$$

  Both classes are processed to completion:

  $$\lambda_1 = 0.25 \cdot \mu_{11} + 0.75 \cdot \mu_{12}, \quad \lambda_2 = 0.5 \cdot \mu_{21}$$

  Utilization of both stations $= 0.75$

- The system is NOT critically loaded!

Heavy Traffic. Fluid View

- Some LP has to be formulated...
- (ξ_{ij}) - allocation matrix,
- ρ - utilization of the busiest station
- Static allocation problem [Harrison $\&$ Lopez (1999)]:
  choose (ξ_{ij}) and ρ to

$$\min\Big\{\rho\ :\ \sum_j \mu_{ij}\,\nu_j\,\xi_{ij} = \lambda_i, \quad \sum_i \xi_{ij} \le \rho, \quad \xi_{ij} \ge 0\Big\}.$$

- Heavy Traffic condition:
  There exists a unique optimal solution (ξ^*, ρ^*) to the linear program. Moreover, $\rho^* = 1$ and $\sum_i \xi^*_{ij} = 1$ for all j.

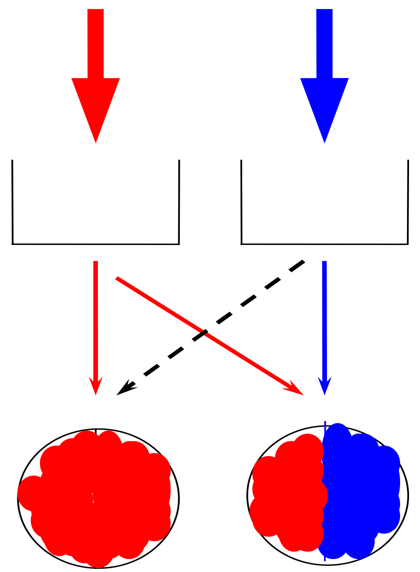
Heavy Traffic. Fluid View

- An example of critically loaded system:

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

$$\mu_{11} = 4, \quad \mu_{12} = 7$$
$$\mu_{21} = 2, \quad \mu_{22} = 4$$

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



Heavy Traffic allocation:

$$\xi^*_{11} = 1, \; \xi^*_{12} = 0.5$$
$$\xi^*_{21} = 0, \; \xi^*_{22} = 0.5$$

- Any reallocation will cause some of the classes to explode.

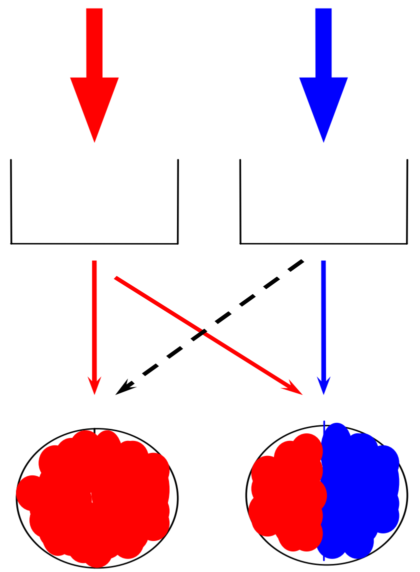
Basic and non-basic activities

Activities: pairs (i, j), with $\mu_{ij} > 0$

Activities can be:

basic (BA), if $\xi^*_{ij} > 0$

non-basic, if $\xi^*_{ij} = 0$



In the example :

basic : $(1, 1)$, $(1, 2)$, $(2, 2)$

non-basic : $(2, 1)$

- Fact: HT implies that BA is a union of disjoint trees.
- Assume that it is one tree.

Stochastic Control of Diffusions

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

Perform centering about static fluid and rescaling.

Assume no usage of non-basics. Then take formal weak limits as $n \to \infty$, to get a controlled diffusion:

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

- Given the cost, obtain drift control problem.
- Works of Atar (2005), (2006), Harrison and Zeevi (2004), Atar, Mandelbaum and Reiman (2004)
- Optimal control of the diffusion gives rise to asymptotically optimal scheduling for original queueing model.

Stochastic Control of Diffusions

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

Perform centering about static fluid and rescaling.

Use non-basics:

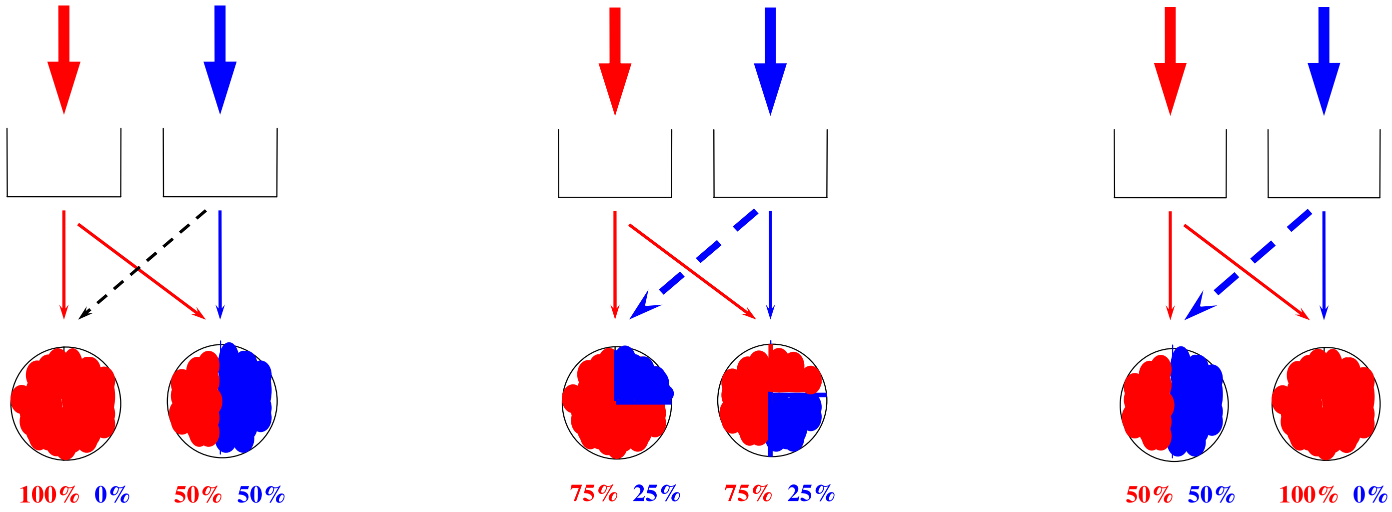
$$X(t) = X(0) + \sigma W(t) + \int_0^t b(X(s), U(s))ds + \sum_{c \in \mathcal{C}} m_c \eta_c(t)$$

Controlled diffusion with a singular control.

- For each c, η_c is nondecreasing with $\eta_c(0) \geq 0$.
- $\mathcal{C}$ - finite set. m_c - constant vectors, depend on μ_{ij}.
- What is the reason for a singular component?...

Reallocation on Fluid Level

- Consider the following massive (order n) customers transfers:



- Performed instantaneously, such transfers may result in abrupt change of a total service rate.

- Such transfers require existence of cycles.

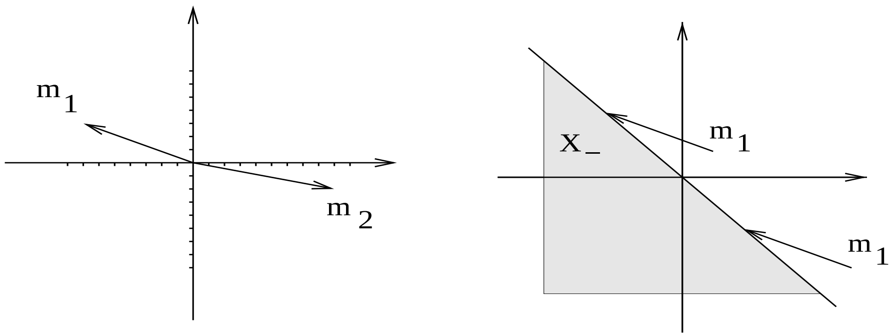
- Cycles are only due to non-basic activities, since basic activities constitute a tree (as known, tree does not contain cycles).

Effect of Singular Component

- Consider a singular controlled diffusion

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

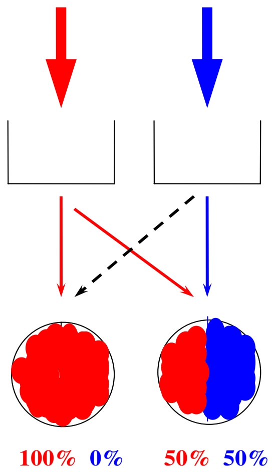
- The singular term η can restrict X to a certain closed domain.
- It can happen that X can be restricted to a domain, corresponding to all queues being empty.

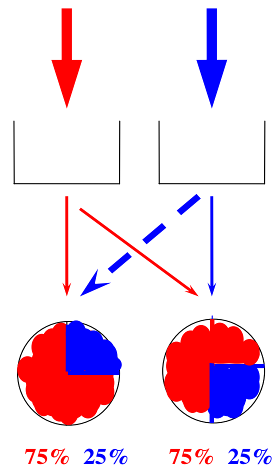


- It happens if $e \cdot m_c < 0$ for some c. Assume this condition!

Changing the Fluid Throughput

$$\underline{\lambda_1 = 7.5, \lambda_2 = 2, \mu_{11} = 4, \mu_{12} = 7, \mu_{21} = 2, \mu_{22} = 4}$$





Total incoming rate:

$$7.5 + 2 = 9.5$$

Total processing rate:

$$4 \cdot 1 + 7 \cdot 0.5 + 4 \cdot 0.5 = 9.5$$

(Total) output equals to input.

Total incoming rate:

$$7.5 + 2 = 9.5$$

Total processing rate:

$$4 \cdot 0.75 + 7 \cdot 0.75 + 2 \cdot 0.25 + 4 \cdot 0.25 = 9.75.$$

(Total) output is greater than input.

Back to Original (prelimit) Model

- Goal: Find a policy,

  that asymptotically (large n) achieves empty queues.

- For two types of control policies:

- Preemptive regime:

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

- Non-preemptive regime:

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

Asymptotic Null Controllability

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

- **Theorem:** There exist a sequence of policies (n-dependent), s.t. for any given $0 < \varepsilon < T < \infty$,

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

- All policies are constructed explicitly!
- The system is critically loaded:
  (any increase in λ_i explodes the system), but...
- behaves like an underloaded (empty queues).