Stationary distribution of large-scale queueing systems in Halfin-Whitt regime: Exponential bounds

A.Stolyar (Bell Labs)

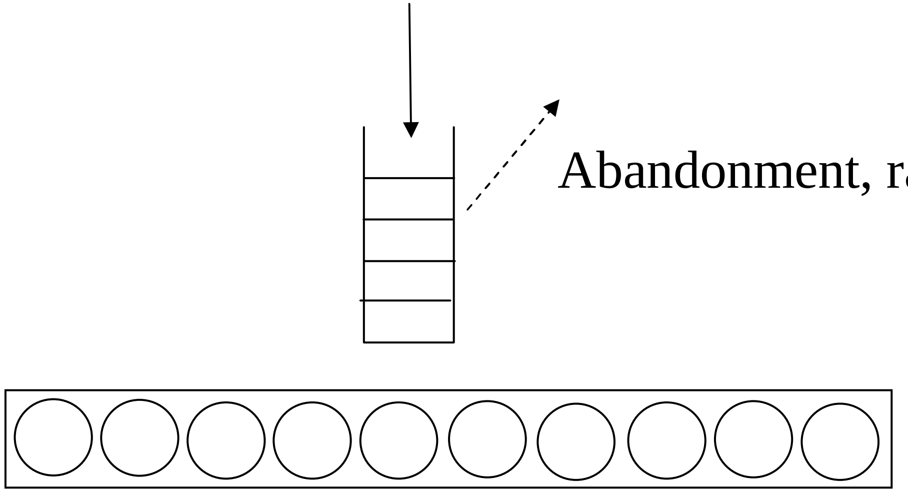
joint work(s) with D.Gamarnik (MIT) and E.Yudovina (Cambridge)

Stochastic Networks conference, DIMACS, October 2011

Outline

- Many-server systems:
  - Halfin-Whitt asymptotic regime
  - problem statement: stationary distribution bounds
  - motivation
- Multi-customer-class, single-server-pool model:
  - Results
  - Proof outline
- More general, multi-server-pool model, under natural load balancing:
  - Negative result in full generality
  - Positive result for a special case
- Conclusions

Basic many-server model. Halfin-Whitt regime



$$N^r = r + a\sqrt{r} = \text{load} + a\sqrt{r}$$

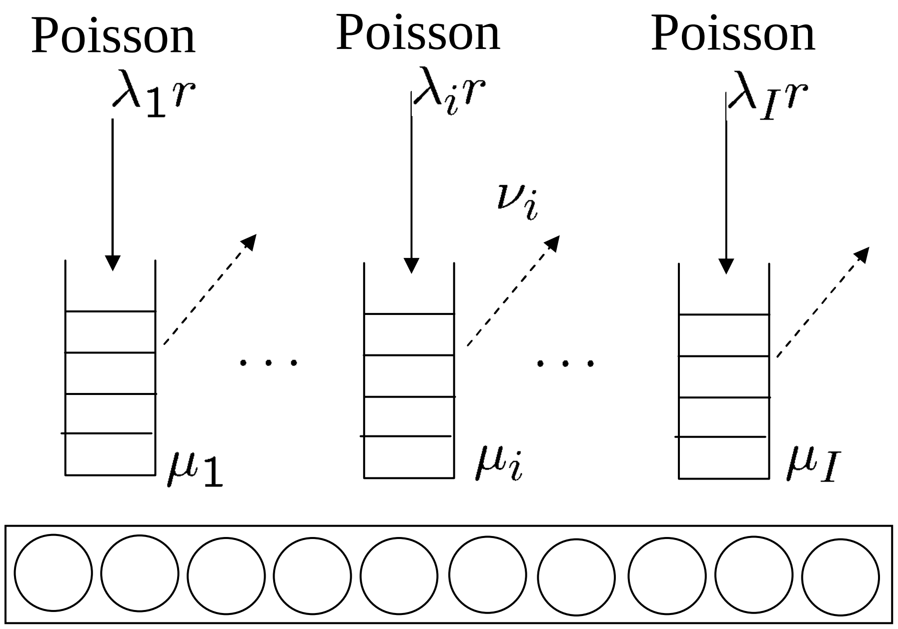
Number of customers in the system Z^r, birth-death process
=> Stationary distribution can be explicitly written and analyzed

Diffusion scaling: $\hat{Z}^r = \frac{Z^r - r}{\sqrt{r}}$

Standard fact: *Convergence of stationary distributions,* $\hat{Z}^r \Rightarrow \hat{Z}$*, and moreover*

$$\limsup_r \exp(\theta|\hat{Z}^r|) < \infty.$$

Multiple customer classes, single server pool



$r \to \infty$

Exponential service times, different rates

$\sum_i \lambda_i/\mu_i = 1$

$$N^r = r + a\sqrt{r} = \text{load} + a\sqrt{r}$$

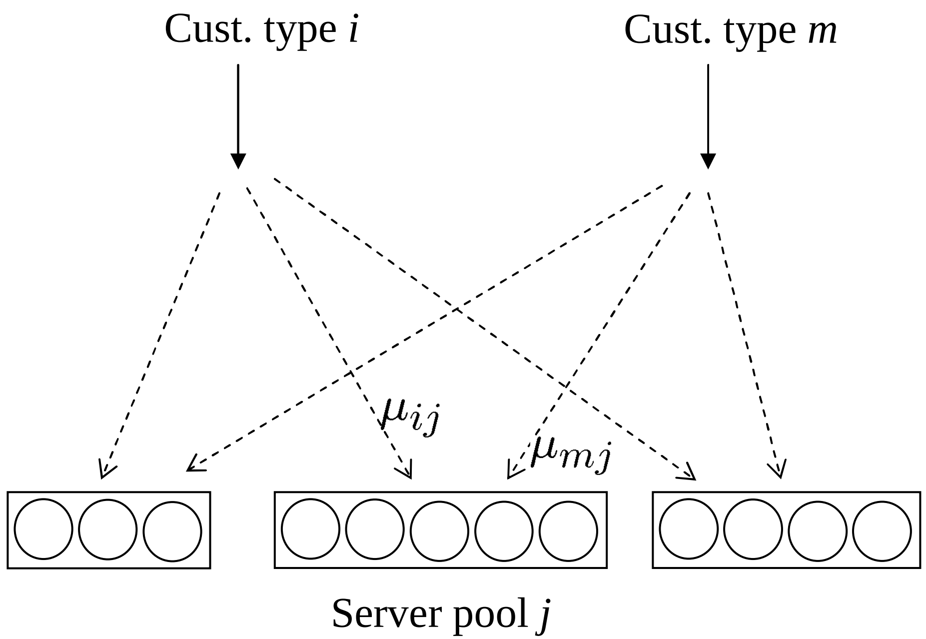
Arbitrary service/queuing discipline without idling => Stationary distribution exists

Diffusion scaled number in the system: $\hat{Z}_i^r = \frac{Z_i^r - (\lambda_i/\mu_i)r}{\sqrt{r}}$

Problem: Uniform in *r* bounds on the stationary distribution of $(\hat{Z}_i^r)$

Those imply bounds on the diffusion scaled queue lengths and server idleness as well

More general models. Motivation



- Several customer types, each has a flow of arrivals
- Several large server pools, homogeneous servers within each pool
- Pools are different and flexible:
  - Cust. service rate μ_{ij} depends on both cust. type *i* and server pool *j*
- Motivation:
  - Call centers: customers = calls; servers = agents
  - Health care systems (fashionable!): cust. = patients; servers = doctors, nurses, hospital beds, etc.
  - Large resource pools in cloud computing
- Problems:
  - Design and analysis of efficient real-time scheduling/routing controls
  - Diffision-scale tightness around desired operating point = Good performance

Multi-class, single-pool: main results

Theorem 1. *$\exists \theta > 0$ s.t. in stationary regime, uniformly on r and all non-idling disciplines:*

$$\limsup_{r\to\infty} E \exp(\theta \sum_i \hat{Z}_i^{r,+}) < \infty,$$

$$\limsup_{r\to\infty} E \exp(\theta \sum_i \hat{Z}_i^{r,-}) < \infty.$$

Corollary. *Stationary distributions are tight. There exists a limit in distribution:*

$$(\hat{Z}_i^r) \Rightarrow (\hat{Z}_i).$$

Theorem 2. *If $\nu_i > 0$, $\forall i$, there exists $\theta > 0$ s.t.*

$$E \exp\left(\theta(\sum_i \hat{Z}_i^+)^2\right) < \infty.$$

If $\nu_i \le \mu_i$, $\forall i$, there exists $\theta > 0$ s.t.

$$E \exp\left(\theta(\sum_i \hat{Z}_i^-)^2\right) < \infty.$$

Quick comments

- If discipline is FIFO, we have a *M/PH/N* single-class system. Tightness results for *GI/GI/N* by Gamarnik-Goldberg'2011

- For some specific disciplines (e.g. priority, queue balancing), it is possible to obtain process convergence to a diffusion limit:

$$(\hat{Z}^r(t),\ t \geq 0) \Rightarrow (\hat{Z}(t),\ t \geq 0),$$

$$d\hat{Z}(t) = C_1(\hat{Z}(t))\hat{Z}(t) + C_2 dW(t).$$

- Not particularly useful for proving tightness: describes behavior if $\hat{Z}^r(0) = O(1)$ But that's exactly what needs to be proved for steady-state

- It looks like you cannot avoid discipline-specific analysis of system dynamics, e.g. discipline-specific Lyapunov functions

Key difficulty with using workload as Lyapunov function

Diffusion scaled workload (expected unfinished work): $\hat{\Phi}^r = \sum_i \hat{Z}_i^r / \mu_i$

Diffusion scaled total number in the system: $\hat{Z}^r = \sum_i \hat{Z}_i^r$

Diffusion scaled i-queue length: $\hat{Q}_i^r = Q_i^r / \sqrt{r}$

$$\sum_i \hat{Q}_i^r \equiv [\hat{Z}^r - a]^+$$

System state, determines all other variables: S

Markov process generator: $AF = AF(S)$

$$A\hat{\Phi}^r = -[\hat{Z}^r \wedge a] - \sum_i (\nu_i / \mu_i) \hat{Q}_i^r$$

Due to class-dependence of service rates, it is quite possible for workload to be large positive, and yet have positive drift

Monotonicity

If we reduce abandonment rates, we can construct a non-idling discipline with larger number of customers:

Lemma [Monotonicity] *Consider a modified set of abandonment rates* $\nu_i^* \leq \nu_i$. *Then, for the modified system, there exists another non-idling discipline, s.t.*

$$\widehat{Z}_i^r \leq \widehat{Z}_i^{*,r}, \quad \forall i.$$

Does not work in opposite direction: cannot claim that by increasing abandonment rates we can have a discipline with smaller number of customers.

Poisson lower bound

If abandonment rates do not exceed service rates, $\nu_i \le \mu_i$, then we have automatic lower bound:

$$Z_i^r \ge \text{Poisson}(\lambda_i r/\mu_i), \quad \forall i$$

Lemma [Poisson lower bound] *If* $\nu_i \le \mu_i, \ \forall i,$ *then for any fixed* $\theta \ge 0$,

$$\limsup_{r\to\infty} E \exp(\theta \sum_i \hat{Z}_i^{r,-}) < \infty.$$

By monotonicity, to obtain bound on $\sum_i \hat{Z}_i^{r,+}$ it suffices to assume zero abandonment rates, so that the above lemma holds.

Proof of Theorem 1 upper bound

Need to show $\limsup_{r\to\infty} E\exp(\theta\sum_i \hat{Z}_i^{r,+}) < \infty$

for some $\theta>0$, assuming zero abandonment rates, and thus

$$\limsup_{r\to\infty} E\exp(\theta\sum_i \hat{Z}_i^{r,-}) < \infty$$

Suffices to show $\limsup_{r\to\infty} E\exp(\theta\sum_i \hat{\Phi}^r) < \infty$ because

$$\sum_i \hat{Z}_i^{r,+}/\mu_i = \hat{\Phi}^r + \sum_i \hat{Z}_i^{r,-}/\mu_i$$

Important observation. For any fixed number b,

$$\hat{Z}^r \equiv \sum_i \hat{Z}_i^{r,+} - \sum_i \hat{Z}_i^{r,-} \le b \;\Rightarrow\; \sum_i \hat{Z}_i^{r,+} \le b + \sum_i \hat{Z}_i^{r,-}$$

Therefore,

$$\hat{Z}^r \le b \;\Rightarrow\; \hat{\Phi}^r \le \sum_i \hat{Z}_i^{r,+}/\mu_i \le b_1 \sum_i \hat{Z}_i^{r,-} + b_2$$

Proof of Theorem 1 upper bound (cont.)

$$A\exp(\theta\hat{\Phi}^r) \le \exp(\theta\hat{\Phi}^r)\left[\theta A\hat{\Phi}^r + c\theta^2\right] = \exp(\theta\hat{\Phi}^r)\left[-\theta(\hat{Z}^r \wedge a) + c\theta^2\right] =$$

negative for small $\theta > 0$

$$I\{\hat{Z}^r \ge a\}\exp(\theta\hat{\Phi}^r)\left[-\theta a + c\theta^2\right]$$

$$+I\{0 \le \hat{Z}^r < a\}\exp(\theta\hat{\Phi}^r)\left[c\theta^2\right]$$

$$+I\{\hat{Z}^r < 0\}\exp(\theta\hat{\Phi}^r)\left[-\theta\hat{Z}^r + c\theta^2\right]$$

$$\hat{\Phi}^r \le b_1\sum_i \hat{Z}_i^{r,-} + b_2$$

$$-\hat{Z}^r \le \sum_i \hat{Z}_i^{r,-}$$

Take expectation w.r.t. stationary distribution, use $EA\exp(\theta\hat{\Phi}^r) = 0$ and (e.g.)

$$EI\{\hat{Z}^r < 0\}\exp(\theta\hat{\Phi}^r)\left[-\theta\hat{Z}^r + c\theta^2\right] \le EI\{\hat{Z}^r < 0\}\left[c_1\exp(\theta'\sum \hat{Z}_i^{r,-}) + c_2\right],\ \ \theta' > 0.$$

$$EI\{\hat{Z}^r \ge a\}\exp(\theta\hat{\Phi}^r) < c_3, \quad \text{uniformly in} \quad r$$

Proof of Theorem 1 upper bound (cont.)

$$EI\{\hat{Z}^r \geq a\} \exp(\theta \hat{\Phi}^r) < c_3, \quad \text{uniformly in} \quad r$$

Using again the upper bound on $\hat{Z}^r$,

$$EI\{\hat{Z}^r < a\} \exp(\theta \hat{\Phi}^r) < c_4, \quad \text{uniformly in} \quad r$$

And we are done:

$$E \exp(\theta \hat{\Phi}^r) < c_5, \quad \text{uniformly in} \quad r.$$

Proof of Theorem 1 lower bound

Need to show $\limsup_{r\to\infty} E \exp(\theta \sum_i \hat{Z}_i^{r,-}) < \infty$

for some $\theta > 0$. Not automatic, because $\nu_i > \mu_i$ is possible => No Poisson lower bound. However, we already do have the upper bound:

$$\limsup_{r\to\infty} E \exp(\theta \sum_i \hat{Z}_i^{r,+}) < \infty$$

Suffices to show $\limsup_{r\to\infty} E \exp(-\theta \sum_i \hat{\Phi}^r) < \infty$ because

$$\sum_i \hat{Z}_i^{r,-}/\mu_i = -\hat{\Phi}^r + \sum_i \hat{Z}_i^{r,+}/\mu_i$$

Important observation (in opposite direction). For any fixed number b,

$$\hat{Z}^r \equiv \sum_i \hat{Z}_i^{r,+} - \sum_i \hat{Z}_i^{r,-} \geq b \ \Rightarrow \ \sum_i \hat{Z}_i^{r,-} \leq -b + \sum_i \hat{Z}_i^{r,+}$$

Therefore,

$$\hat{Z}^r \geq b \ \Rightarrow \ -\hat{\Phi}^r \leq \sum_i \hat{Z}_i^{r,-}/\mu_i \leq b_1' \sum_i \hat{Z}_i^{r,+} + b_2'$$

Proof of Theorem 1 lower bound (cont.)

Write upper bound on $A \exp(-\theta \widehat{\Phi}^r)$

Take expectation w.r.t. stationary distribution, and break down the RHS into cases

$$\{\widehat{Z}^r \le b\}, \quad \{\widehat{Z}^r > b\}, \quad b < 0 \text{ fixed}$$

Gives

$$EI\{\widehat{Z}^r \le b\} \exp(-\theta \widehat{\Phi}^r) < c_3, \quad \text{uniformly in } r$$

and then

$$E \exp(-\theta \widehat{\Phi}^r) < c_5, \quad \text{uniformly in } r$$

On the proof of Theorem 2

In the proof of Theorem 1, need to be careful with the domain of generator, so as an intermediate step we show (for the upper bound):

$$EI\{\widehat{\Phi}^r \leq k\} \exp(\theta \widehat{\Phi}^r) < C, \quad \text{uniformly in} \quad r, \ k.$$

This implies

$$E \exp(\theta \widehat{\Phi}^r) < C, \quad \text{uniformly in} \quad r$$

In the proof of Theorem 2, we only have:

$$\limsup_r EI\{\widehat{\Phi}^r \leq k\} \exp(\theta [\widehat{\Phi}^r]^2) < C, \quad \forall k.$$

which is good enough to claim for the limit:

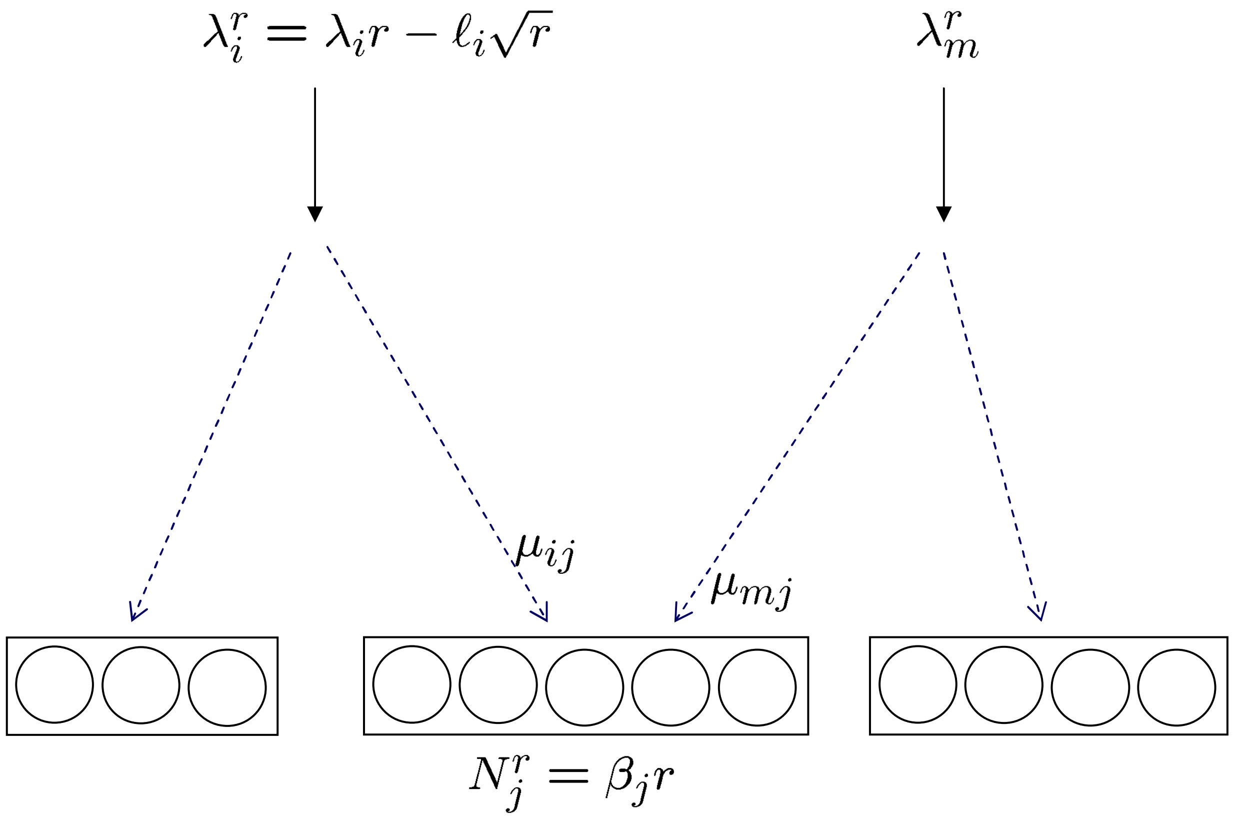
$$E \exp(\theta [\widehat{\Phi}]^2) < C.$$

However,

$$E \exp(\theta [\widehat{\Phi}^r]^2) = \infty, \quad \forall r.$$

Halfin-Whitt regime for a more general model

$r \to \infty$ scaling parameter



- Feasible activities (*(ij)*-edges) form a tree
- Relation between parameters λ_i, μ_{ij}, β_j is such that the (smallest possible) system utilization is $1 - O(1/\sqrt{r})$: next slide

Halfin-Whitt regime for a general model

"Fluid-scale" load balancing LP:

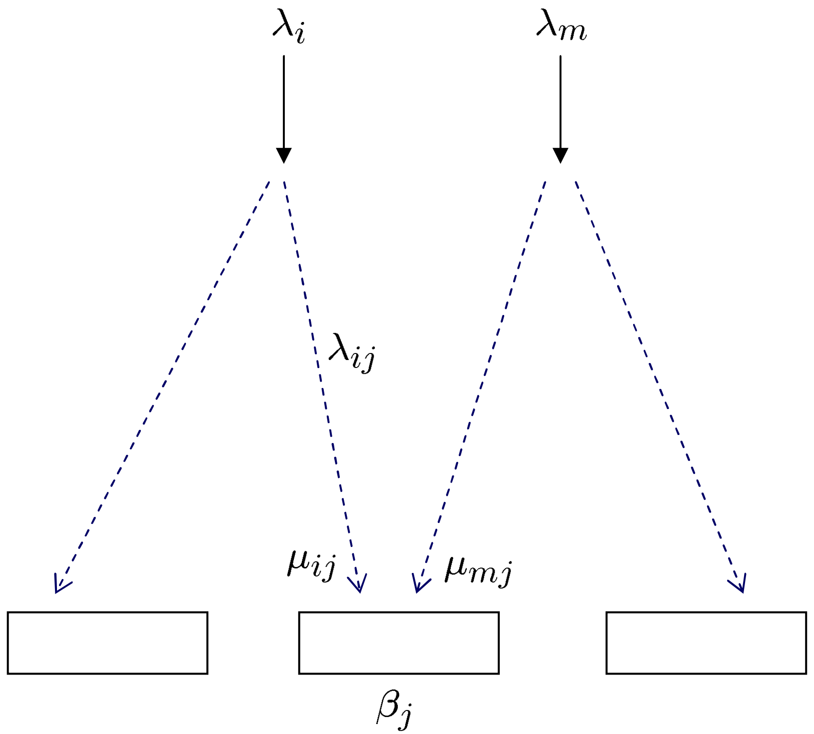
$$\min_{\{\lambda_{ij}\},\rho} \rho$$

subject to

$$\sum_i \lambda_{ij}/(\beta_j \mu_{ij}) \le \rho, \quad \forall j,$$

$$\lambda_{ij} \ge 0,$$

$$\sum_j \lambda_{ij} = \lambda_i, \quad \forall i.$$



- Optimal LP solution is such that $\rho=1$ and $\lambda_{ij} > 0$ for all edges *(ij)*
- Implies that on fluid-scale perfect load balancing is achievable

Equilibrium point. Diffusion scaling

Desired operating point ("equilibrium" point):

- Zero queues: $Q_i^r = 0$
- Perfect load balancing: Number of i-customers occupying j-servers

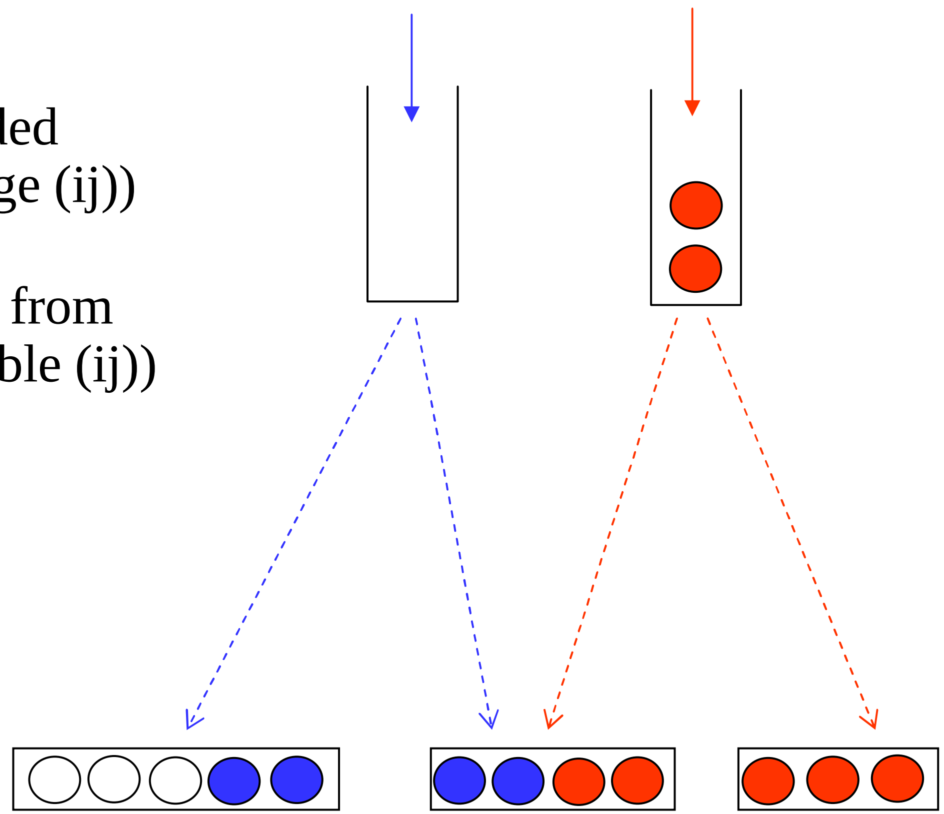
$$\Psi_{ij}^r = (\lambda_{ij}/\mu_{ij})r$$

Diffusion scaling: $\hat{Q}_i^r = Q_i^r/\sqrt{r}$

$$\hat{\Psi}_{ij}^r = [\Psi_{ij}^r - (\lambda_{ij}/\mu_{ij})r]/\sqrt{r}$$

Natural load balancing strategy

- Customer routing: Go to least loaded server pool (along an available edge (ij))
- Server scheduling: Take customer from the longest queue (along an available (ij))



- No need to know any parameters => Very desirable feature
- Intuitively, should work just fine
- Unfortunately, can be unstable around the equilibrium point => No tightness of diffusion-scaled stationary distributions

Natural load balancing strategy: $\mu_{ij}=\mu_j$ case

Theorem 3. *Suppose $\mu_{ij} = \mu_j$ and all $\mu_i = 0$. Then the sequence of stationary distributions of $\Big((\widehat{Q}_i^r), (\widehat{\Psi}_{ij}^r)\Big)$ has uniform in r exponential bounds.*

Lyapinov function:

$$\mathcal{L}^r = \sum_i \exp(\theta \widehat{Q}_i^r) + \sum_j \beta_j \exp(\theta \widehat{\Psi}_j^r / \beta_j)$$

for $\theta > 0$ and $\theta < 0$,

$$\widehat{\Psi}_j^r = \sum_i \widehat{\Psi}_{ij}^r \leq 0.$$

Discussion

- Showing diffusion-scale tightness/bounds in many-server models is challenging
- For multi-customer-class, single-server-pool model:
  - Prove exponential bounds, uniform w.r.t. scale parameter and all non-idling disciplines
  - Sub-Gaussian tail result for a weak limit of stationary distribution, given positive abandonment rates
  - Using lower bounds to obtain upper bounds, and vice versa:
    - » may be of more general use
    - » in our case, enables use of workload as Lyapunov function
- For more general, multi-server-pool model:
  - Prove uniform exponential bounds for natural load balancing, in the special case of server-only dependent service rates
- Many more challenges remain ...