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

Poisson, rate = \( r \) \( \quad r \to \infty \)

Abandonment, rate \( \nu \geq 0 \)

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

Number of customers in the system \( Z^r \), birth-death process

\( \Rightarrow \) 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

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

Arbitrary service/queuing discipline without idling \(\Rightarrow\) 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 $\mu_{ij}$ 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
  - Diffusion-scale tightness around desired operating point = Good performance
Multi-class, single-pool: main results

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

\[
\limsup_{r \to \infty} E \exp(\theta \sum_i \hat{Z}_{i}^{r,+}) < \infty,
\]

\[
\limsup_{r \to \infty} E \exp(\theta \sum_i \hat{Z}_{i}^{r,-}) < \infty.
\]

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

\[
(\hat{Z}_{i}^{r}) \Rightarrow (\check{Z}_{i}).
\]

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

\[
E \exp \left( \theta \left( \sum_i \hat{Z}_{i}^{+} \right)^2 \right) < \infty.
\]

If \( \nu_i \leq \mu_i, \ \forall i \), there exists \( \theta > 0 \) s.t.

\[
E \exp \left( \theta \left( \sum_i \hat{Z}_{i}^{-} \right)^2 \right) < \infty.
\]
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:

  \[
  (\tilde{Z}^r(t), \ t \geq 0) \Rightarrow (\tilde{Z}(t), \ t \geq 0),
  \]

  \[
  d\tilde{Z}(t) = C_1(\tilde{Z}(t))\tilde{Z}(t) + C_2dW(t).
  \]

- Not particularly useful for proving tightness: describes behavior if $\tilde{Z}^r(0) = O(1)$
  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): \[ \hat{\Phi}^r = \sum_i \frac{\hat{Z}_i^r}{\mu_i} \]

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

Diffusion scaled i-queue length: \[ \bar{Q}_i^r = Q_i^r / \sqrt{r} \]
\[ \sum_i \bar{Q}_i^r \equiv [\hat{Z}^r - a]^+ \]

System state, determines all other variables: \[ S \]

Markov process generator: \[ A\hat{\Phi}^r = -[\hat{Z}^r \land a] - \sum_i (\nu_i / \mu_i) \bar{Q}_i^r \]

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.

$$\hat{Z}_i^r \leq \hat{Z}_{i^*}^r \quad \forall i.$$ 

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, $\nu_i \leq \mu_i$, then we have an automatic lower bound:

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

Lemma [Poisson lower bound] If $\nu_i \leq \mu_i, \quad \forall i$, then for any fixed $\theta \geq 0$,

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

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

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

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

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

because

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

Important observation. For any fixed number \( b \),

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

Therefore,

\[
\hat{Z}^r \leq b \Rightarrow \hat{\Phi}^r \leq \sum_i \hat{Z}^{r,+}_i / \mu_i \leq b_1 \sum_i \hat{Z}^{r,-}_i + b_2
\]
Proof of Theorem 1 upper bound (cont.)

\[ A \exp(\theta \Phi^r) \leq \exp(\theta \Phi^r)[\theta A \Phi^r + c\theta^2] = \exp(\theta \Phi^r)[-\theta (\bar{Z}^r \wedge a) + c\theta^2] = \]

\[ = I\{\bar{Z}^r \geq a\} \exp(\theta \Phi^r)[-\theta a + c\theta^2] + I\{0 \leq \bar{Z}^r < a\} \exp(\theta \Phi^r)[c\theta^2] + I\{\bar{Z}^r < 0\} \exp(\theta \Phi^r)[-\theta \bar{Z}^r + c\theta^2] \]

\[ \Phi^r \leq b_1 \sum_i \bar{Z}_{i,-}^r + b_2 \]

\[ -\bar{Z}^r \leq \sum_i \bar{Z}_{i,-}^r \]

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

\[ EI\{\bar{Z}^r < 0\} \exp(\theta \Phi^r)[-\theta \bar{Z}^r + c\theta^2] \leq EI\{\bar{Z}^r < 0\}[c_1 \exp(\theta' \sum_i \bar{Z}_{i,-}^r) + c_2], \ \theta' > 0. \]

\[ EI\{\bar{Z}^r \geq a\} \exp(\theta \Phi^r) < c_3, \text{ uniformly in } r \]
Proof of Theorem 1 upper bound (cont.)

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

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

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

And we are done:

\[ E \exp(\theta \hat{\Phi}^r) < c_5, \text{ uniformly in } r. \]
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 \( \Rightarrow \) 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 \( \limsup_{r \to \infty} E \exp(-\theta \sum_i \hat{\Phi}_i^r) < \infty \) because

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

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

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

Therefore,

\[ \hat{Z}_i^r \geq b \Rightarrow -\hat{\Phi}_i^r \leq \sum_i \hat{Z}_i^{r,-} / \mu_i \leq b'_1 \sum_i \hat{Z}_i^{r,+} + b'_2 \]
Proof of Theorem 1 lower bound (cont.)

Write upper bound on \( A \exp(-\theta \hat{\Phi}) \)

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

\[
\{ \tilde{Z}^r \leq b \}, \quad \{ \tilde{Z}^r > b \}, \quad b < 0 \text{ fixed}
\]

Gives

\[
EI \{ \tilde{Z}^r \leq b \} \exp(-\theta \hat{\Phi}) < c_3, \quad \text{uniformly in } r
\]

and then

\[
E \exp(-\theta \hat{\Phi}) < c_5, \quad \text{uniformly in } r
\]
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):

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

This implies

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

In the proof of Theorem 2, we only have:

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

which is good enough to claim for the limit:

\[ E \exp(\theta [\Phi]^2) < C. \]

However,

\[ E \exp(\theta [\Phi^r]^2) = \infty, \quad \forall r. \]
Halfin-Whitt regime for a more general model

\[ r \to \infty \quad \text{scaling parameter} \]

\[ \lambda_i^r = \lambda_i r - \ell_i \sqrt{r} \]

\[ N_j^r = \beta_j r \]

Feasible activities \(((ij)\)-edges\) form a tree

Relation between parameters \(\lambda_i, \mu_{ij}, \beta_j\) is such that the (smallest possible) system utilization is \(1 - O(1/\sqrt{r})\) : next slide
"Fluid-scale" load balancing LP:

\[
\min_{\{\lambda_{ij}, \rho\}} \rho
\]

subject to

\[
\sum_i \lambda_{ij}/(\beta_j \mu_{ij}) \leq \mu, \quad \forall j,
\]

\[
\lambda_{ij} \geq 0,
\]

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

◆ 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:

\[
\tilde{Q}_i^r = Q_i^r / \sqrt{r}
\]

\[
\tilde{\psi}_{ij}^r = [\psi_{ij}^r - (\lambda_{ij}/\mu_{ij})r] / \sqrt{r}
\]
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 $((\bar{Q}_i^r), (\bar{\Psi}_{ij}^r))$ has uniform in $r$ exponential bounds.

Lyapinov function:

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

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

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