Large scale queueing systems in the Quality/Efficiency (Halfin-Whitt) driven regime, and applications

David Gamarnik
MIT

33rd CONFERENCE ON THE MATHEMATICS OF OPERATIONS RESEARCH

January 2008

Joint work with
P. Momcilovic, U of Michigan,

Talk Outline

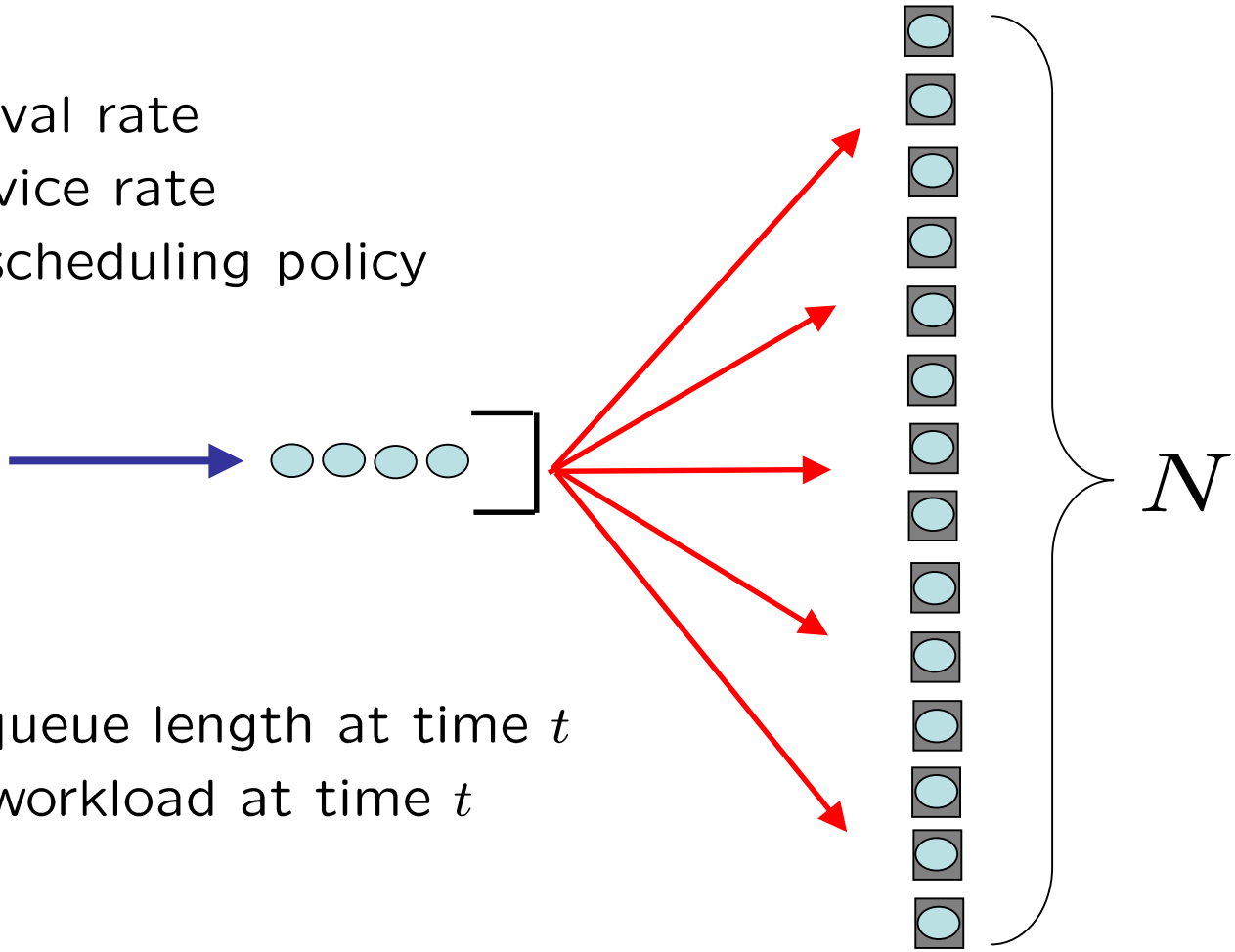
- GI/GI/N queueing model
- Applications to call/contact centers
- Challenge: non-Markovian systems
- **Results I**: Markov chain characterization of the queue length process
- **Result II**: Interchange of Heavy Traffic-Steady State limits. Tight decay rate.
- Discussion of methods: Lyapunov functions
- Further challenges

Model: G/G/N queueing system

λ - arrival rate

μ - service rate

FIFO scheduling policy



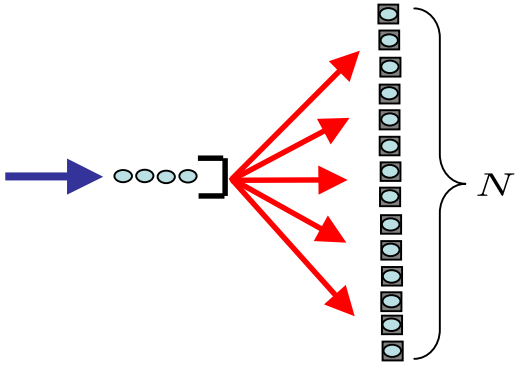
Q_t - queue length at time t

W_t - workload at time t

Model: G/G/N queueing system

Kiefer & Wolfowitz [56] – Steady state regime exists (stability) iff

$$\rho_N \triangleq \frac{\lambda}{N\mu} < 1.$$



… but no explicit formulas for Q_∞, W_∞ except

Erlang M/M/N (Erlang-C) formulas

Halfin-Whitt Theory (summary)

In M/M/N queueing system in steady-state

if

$$\rho = 1 - \frac{\beta}{\sqrt{N}}, \qquad \left(N \approx \lambda \mathbb{E}[S] + \beta\sqrt{\lambda \mathbb{E}[S]}\right)$$

Efficiency

$\rho \approx 1$

then

$$\lim_N \mathbb{P}(Q_\infty^N > 0) = \frac{1}{1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\frac{\beta^2}{2}}}$$

Quality

$\mathbb{P}(Q_\infty = 0) > 0$

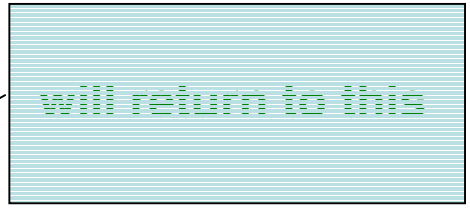
Also

$$\mathbb{P}\left(\frac{Q_\infty^N}{\sqrt{N}} > t | Q_\infty^N > 0\right) \to e^{-\beta t}$$

$$\mathbb{P}\left(\frac{I_\infty^N}{\sqrt{N}} > t | Q_\infty^N = 0\right) \to \frac{\Phi(\beta - t)}{\Phi(\beta)}$$

Furthermore:

- Extends to non-Poisson arrival processes: G/M/N



$$\mathbb{P}(\frac{Q_\infty^N}{\sqrt{N}} > t | Q_\infty^N > 0) \to e^{-\theta t}, \quad \theta = \frac{2\beta}{1+c_a^2}$$

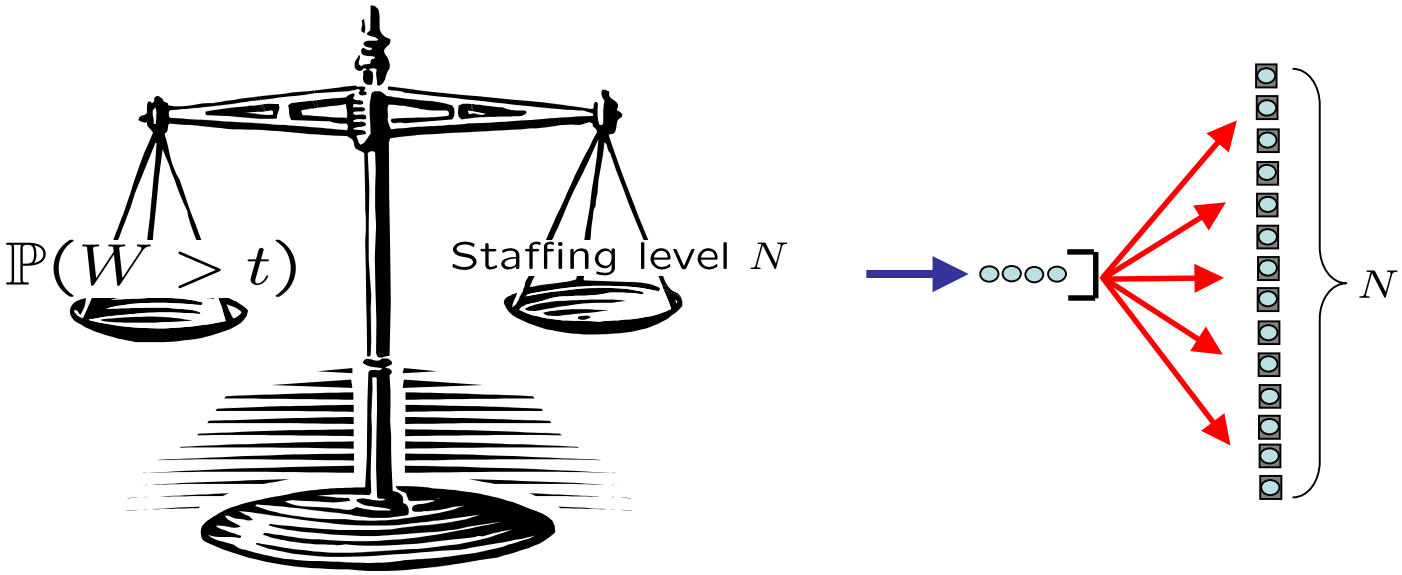
$$\mathbb{P}(\frac{I_\infty^N}{\sqrt{N}} > t | Q_\infty^N = 0) \to \frac{\Phi(\beta - t)}{\Phi(\beta)}$$

- Extends to phase-type service times: **Puhalskii & Reiman** [2000]

- Diffusion approximations:

$$\left(\frac{Q_t^N}{\sqrt{N}}, \frac{I_t^N}{\sqrt{N}}\right) \Rightarrow (\hat{Q}_t, \hat{I}_t)$$

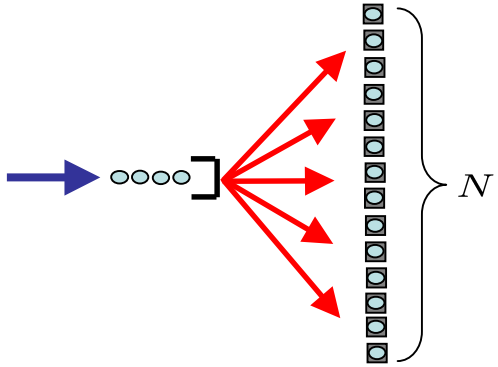
Motivation: *Call Centers*



Tradeoff between utilization ρ (cost of staffing N) and

performance $\mathbb{P}(\text{Wait} > x)$

Motivation: *Call Centers*



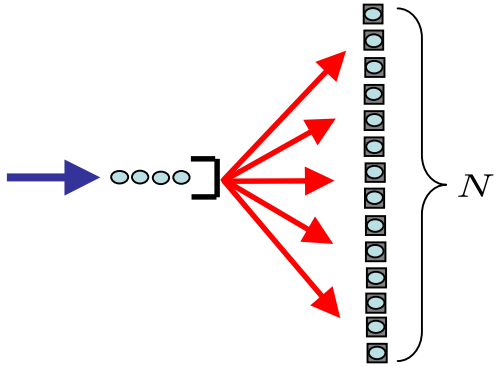
Surveys:

M. Armony: http://www.stern.nyu.edu/om/faculty/armony/research/CallCenterSurvey.pdf
A. Mandelbaum: http://iew3.technion.ac.il/serveng/References/references.html
W. Whitt: http://www.ieor.columbia.edu/~ww2040/CallCenterF04/call.html

Papers on queueing models of call centers by

Atar, Armony, Baron, Bassamboo, Brown, Dai, Green, Gans, Garnett, Gurvich, Harrison, Jelenkovic, Jennings, Halfin, Kolesar, Kumar, Maglaras, Mandelbaum, Massey, Momcilovic, Randhawa, Reed, Reiman, Sakov, Shen, Shimkin, Tezcan, Whitt, Zeevi, Zeltin, Zhao, Zohar

Motivation: *Call Centers*



Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn, Zhao [2002]

Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective.

One of the conclusions: service times have log-normal distribution.

General service time distributions

- Deterministic service times:
  Jelenkovic, Mandelbaum & Momcilovic [2004]

- G/G/N virtual waiting times process limits, discrete service times:
  Mandelbaum & Momcilovic [2005]

$$\frac{W_t}{\sqrt{N}} \Rightarrow \hat{W}_t$$

- G/G/N process limits:
  Reed [2006]

Principal questions for this work:

- What about G/G/N in **steady state**?
- Does $Q^N_\infty \approx \sqrt{N}$ hold in general?

Assumption: service times have a *discrete* probability distribution.

$$\mathbb{P}(S = sk) = p_k, \quad k = 1, 2, \ldots, K$$

Summary of results:

- Markov chain characterization of the limiting process $\lim_{N\to\infty} \frac{Q^N(t)}{\sqrt{N}} = Q(t)$
- Tight decay rate for the limiting queue length in steady-state

Consider the following Markov chain on $(Q_t, \boldsymbol{L}_t) \in \mathbb{R}^{K+1}$

$$\boldsymbol{L}_{t+1} = \mathcal{T}[\boldsymbol{L}_t] + \boldsymbol{X} + \Big((Q_t + \sqrt{\mu}\sigma X) \wedge (\beta - \sum_{j\le 2} L_{j,t}\Big)p,$$

$$Q_{t+1} = \Big(\sqrt{\mu}\sigma X + Q_t + \sum_{j\le 2} L_{j,t} - \beta\Big)^+,$$

where

$$X \stackrel{d}{=} N(0,1)$$

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{N}(0,\boldsymbol{\Sigma}), \qquad \Sigma_{ij} = \begin{cases} (1-p_i)p_i, & 1 \le i = j \le K, \\ -p_i p_j, & 1 \le i \ne j \le K; \end{cases}$$

$$\mathcal{T}(l_1,\dots,l_K) = (l_2,\dots,l_{K-1},0)$$

$\boldsymbol{L}_t^N = (L_{1,t}^N, \ldots, L_{K,t}^N)$, **- centered work in progress**

Theorem I. [Transient] Suppose

Discrete service times

$$\mathbb{P}(S = k) = p_k, \quad k = 1, \ldots, K,$$

$$\lambda_N = \mu N - \beta\mu\sqrt{N}$$

Halfin-Whitt regime

$$\left(\frac{Q_0^N}{\sqrt{N}}, \frac{\boldsymbol{L}_0^N}{\sqrt{N}}\right) \Rightarrow (Q_0, \boldsymbol{L}_0).$$

Initialization

Then $$\left(\frac{Q_t^N}{\sqrt{N}}, \frac{\boldsymbol{L}_t^N}{\sqrt{N}}\right) \Rightarrow (Q_t, \boldsymbol{L}_t), \quad t = 1, 2, \ldots \ .$$

Theorem II. [Steady-state]

- $(Q_t, \boldsymbol{L}_t)$ has a *unique* stationary distribution $(Q_\infty, \boldsymbol{L}_\infty)$.

- $\limsup_{N\to\infty} \mathbb{E}\Big[\exp\Big(\frac{\theta Q_\infty^N}{\sqrt{N}}\Big)\Big] < \infty, \quad \text{for some } \theta > 0.$

- $\Big(\frac{Q_\infty^N}{\sqrt{N}}, \frac{\boldsymbol{L}_\infty^N}{\sqrt{N}}\Big) \Rightarrow (Q_\infty, \boldsymbol{L}_\infty).$

Exponential tightness

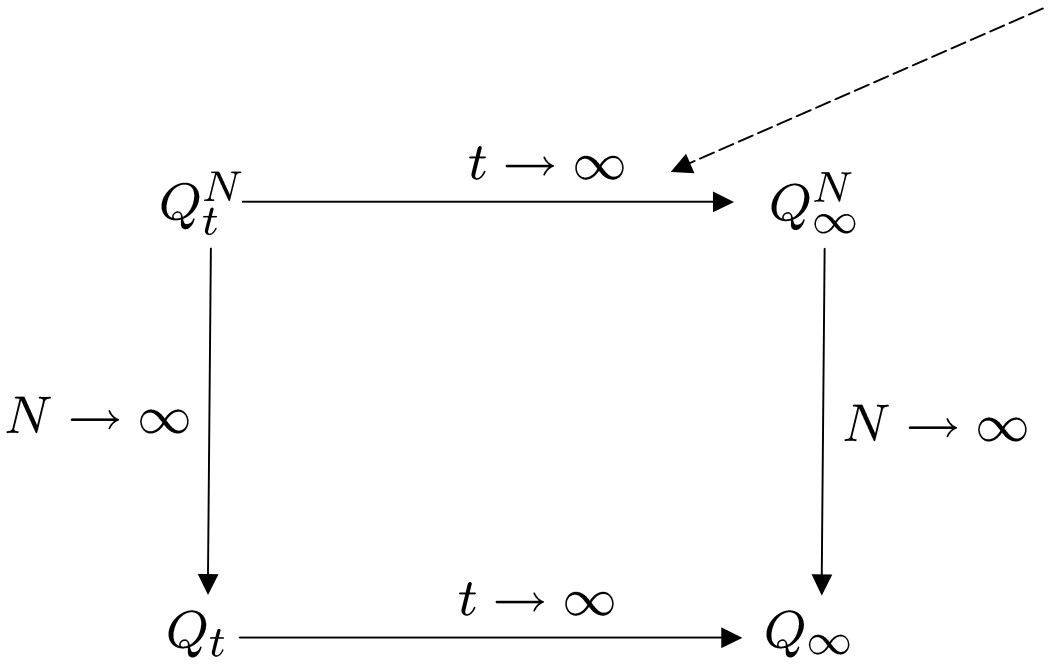
Interchange of limits

$$Q_\infty^N \approx \sqrt{N}$$

Confirmed

Interchange of limits

Kiefer & Wolfowitz



Theorem II. [Steady-state] Moreover (exact decay rate)

$$\mathbb{P}(Q_\infty > l) \approx \exp\left(-\frac{2\beta l}{c_a^2 + c_s^2}\right), \quad \text{for large } l.$$

Same decay rate as the conventional heavy-traffic model of G/G/N!

Illustration

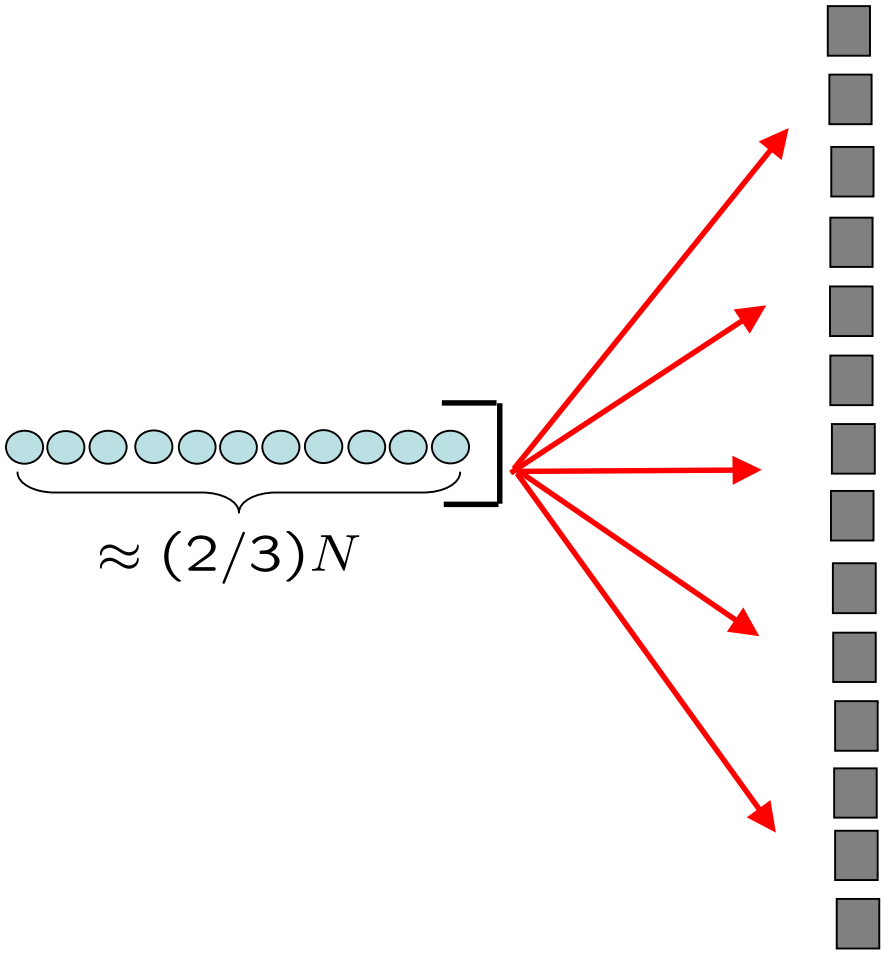
- Service time $S = \begin{cases} 1, & w.p. \quad 0.5; \\ 2, & w.p. \quad 0.5. \end{cases}$

- Arrival rate $\lambda_N = (2/3)N - \beta\sqrt{N}$

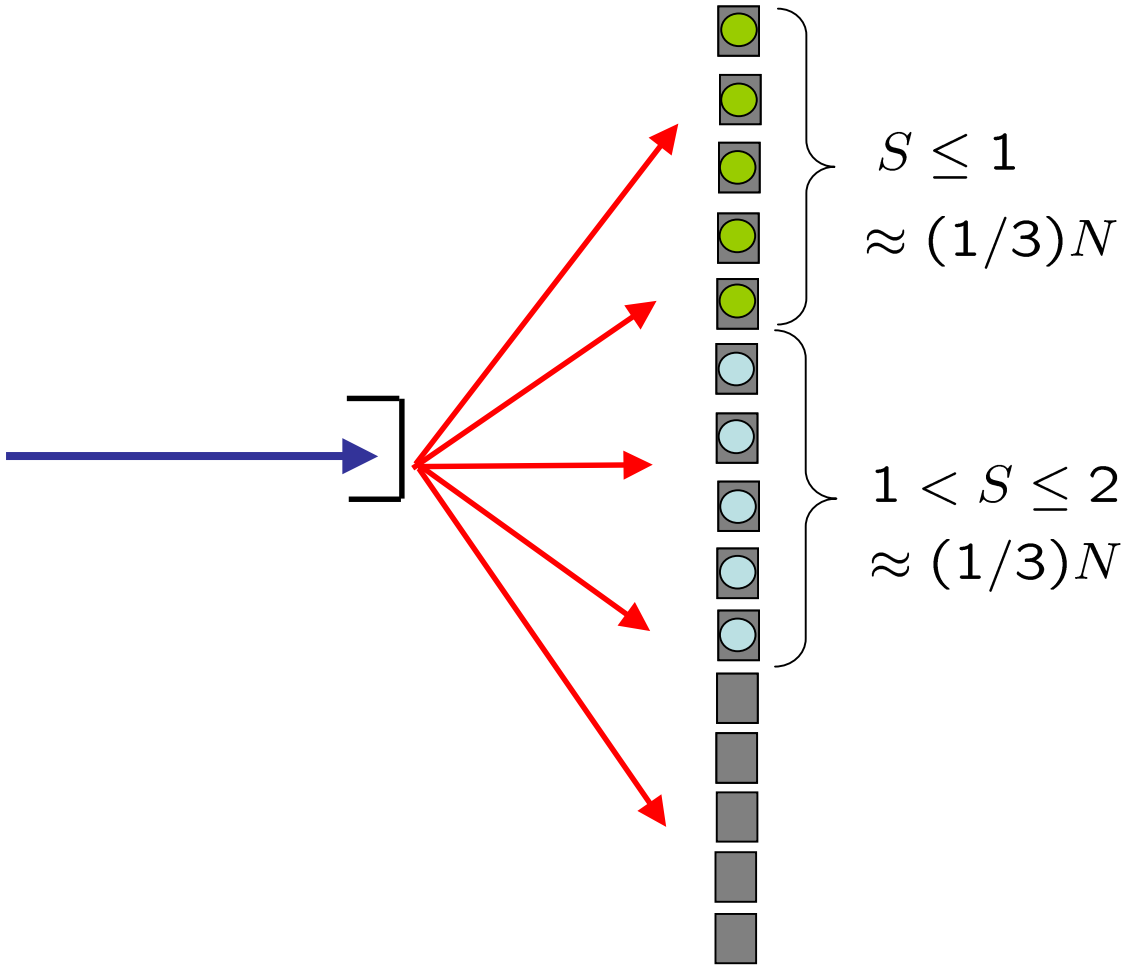
  $Z_{1,t}$ – number of calls with remaining service time < 1

  $Z_{2,t}$ – number of calls with remaining service time in $[1, 2)$

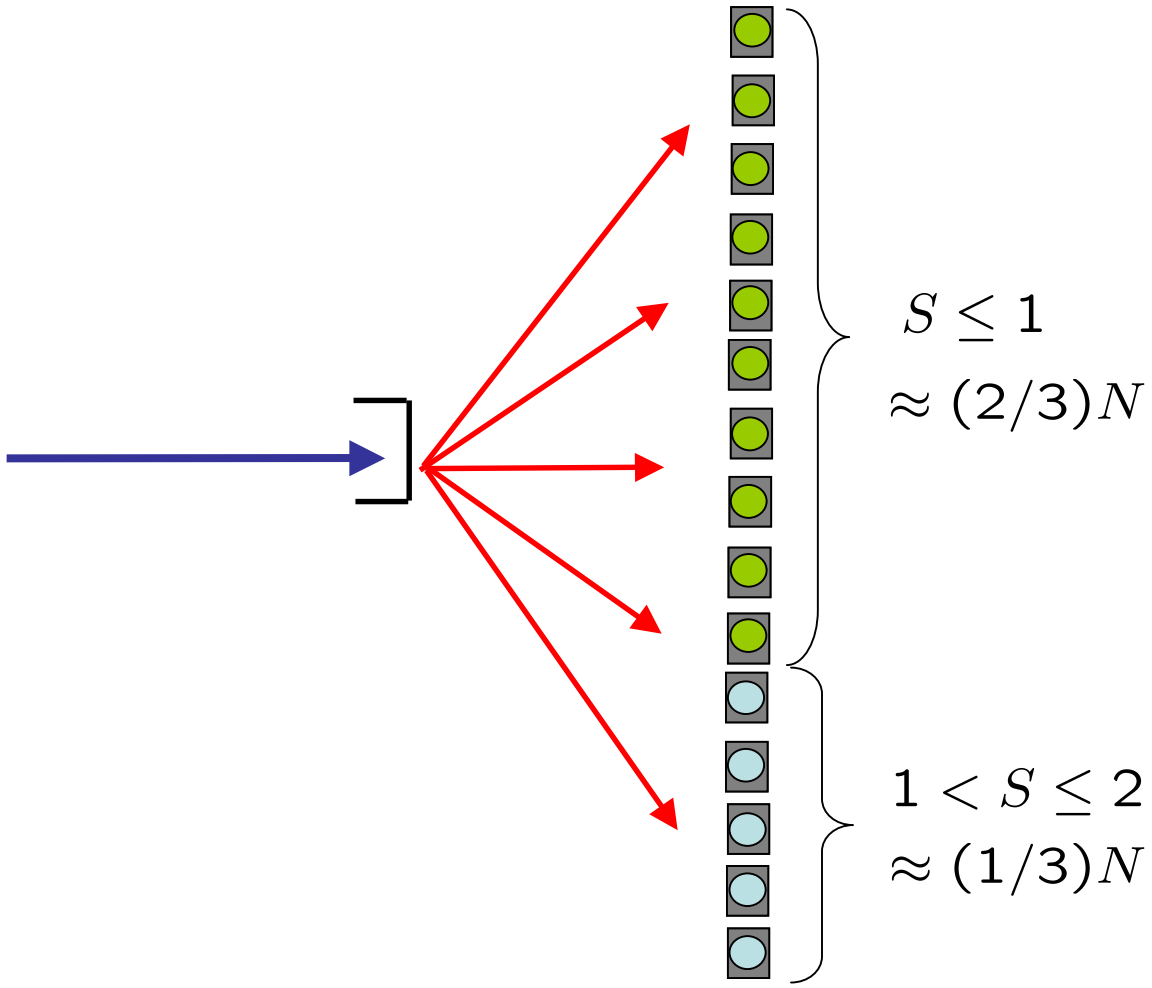
Illustration



Illustration



Illustration



Illustration

Scaling N does not reveal anything. Look at Gaussian scaling:

$$\widehat{A}_{t-1,t} = \frac{A_{t-1,t} - (2/3)N}{\sqrt{N}}$$

$$L_{1,t} = \frac{Z_{1,t} - (2/3)N}{\sqrt{N}}$$

$$L_{2,t} = \frac{Z_{1,t} - (1/3)N}{\sqrt{N}}$$

Illustration

- Case I $Z_{1,t} > Q_t + A_{t,t+1}$

Then

$$Z_{2,t+1} \approx \frac{1}{2}(Q_t + A_{t,t+1})$$
$$Z_{1,t+1} \approx Z_{2,t} + \frac{1}{2}(Q_t + A_{t,t+1})$$
$$Q_{1,t+1} = 0.$$

Subtract the means

$$L_{2,t+1} \approx \frac{1}{2}(Q_t + A_{t,t+1}) - \frac{1}{3}N$$
$$L_{1,t+1} \approx L_{2,t} + \frac{1}{2}(Q_t + A_{t,t+1}) - \frac{2}{3}N$$
$$Q_{1,t+1} = 0.$$

- Case II similar …

Illustration

- Case II similar …

Putting things together, the Markov chain dynamics is obtained

$$\boldsymbol{L}_{t+1} = \mathcal{T}[\boldsymbol{L}_t] + \boldsymbol{X} + \Big((Q_t + \sqrt{\mu}\sigma X) \wedge (\beta + L_{1,t} - \|\boldsymbol{L}_t\|)\Big)p,$$
$$Q_{t+1} = \Big(\sqrt{\mu}\sigma X + Q_t + \|\boldsymbol{L}_t\| - L_{1,t} - \beta\Big)^+,$$

Dynamical system is hard to analyze even without stochastic fluctuations.

$$\boldsymbol{L}_{t+1} = \mathcal{T}[\boldsymbol{L}_t] + \Big(Q_t \wedge (\beta + L_{1,t} - \|\boldsymbol{L}_t\|)\Big)p,$$
$$Q_{t+1} = \Big(Q_t + \|\boldsymbol{L}_t\| - L_{1,t} - \beta\Big)^+,$$

Steady-State

Proposition. The process $Y_t = Q_t + \sum_j \boldsymbol{L}_{t,j},\ Q_t = (Y_t - \beta)^+$ satisfies

$$Y_t = \sum_k p_k (Y_{t-k} - \beta)^+ + \sum_k \alpha_k Z_{t-k},$$

Z_t **-** stationary Gaussian process

When Y_t is "large", the drift is $\approx -\beta$

Steady-State

Theorem. $\Phi(y,z) = \exp(\theta \sum_k p_k^* y_k + \theta \sum_k \alpha_k^* z_k), \qquad (v_k^* = \sum_{j \geq k} v_j).$

is a (geometric) Lyapunov function.

$$\mathbb{E}\Big[\Phi(Y_{t+1}, Z_{t+1}) \Big| Y_t = y, Z_t = z\Big] < (1-\gamma)\Phi(y,z)$$

Lyapunov function argument is used to show that

- $(\boldsymbol{L}_t, Q_t)$ has a stationary distribution $(Q_\infty, \boldsymbol{L}_\infty)$.

- $\left(\frac{Q_\infty^N}{\sqrt{N}}, \frac{\boldsymbol{L}_\infty^N}{\sqrt{N}}\right) \Rightarrow (Q_\infty, \boldsymbol{L}_\infty)$.

- $\limsup_{N\to\infty} \mathbb{E}\left[\exp\left(\frac{\theta Q_\infty^N}{\sqrt{N}}\right)\right] < \infty,$ for some $\theta > 0$.

Summary, Challenges and ongoing work:

- G/G/N in Halfin-Whitt regime with discrete service time distribution. Process level and steady-state results.
- Tight decay rate for the queue length in steady-state.
- **Challenge:** non-discrete service times.
- In progress: relaxation in M/M/N in Halfin-Whitt (joint work with D. Goldberg)