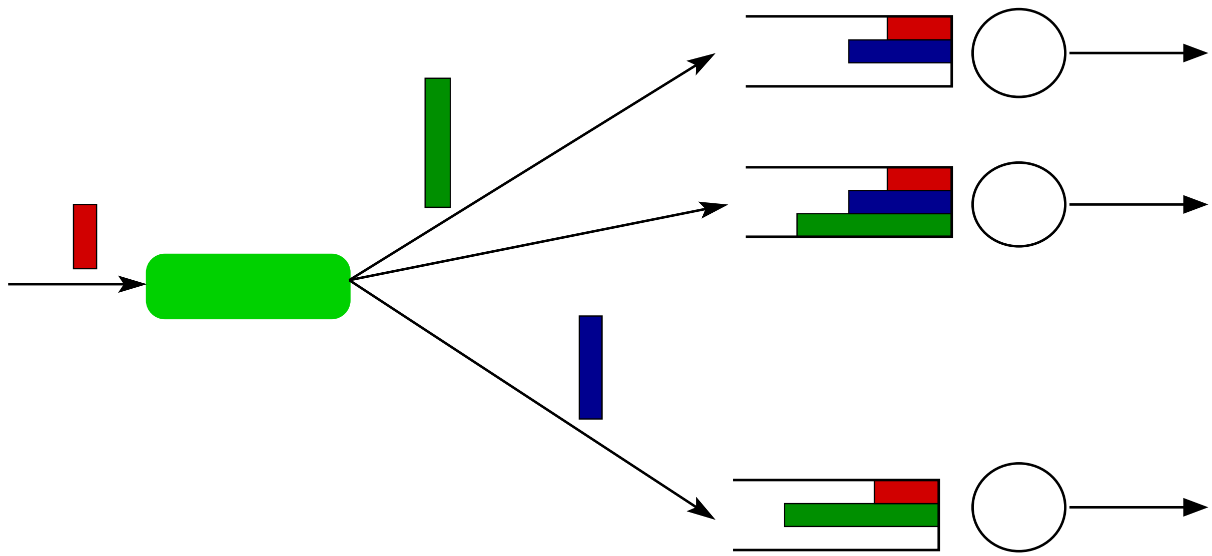
Asymptotics of insensitive load balancing with blocking phases

M. Jonckheere and B.J. Prabhu

UBA, Buenos Aires, Argentina
LAAS-CNRS, Toulouse, France

The load balancing problem



- Finite buffer size of θ at each server
- Knowledge of number of jobs at each server

Objective: minimize blocking probability

Join the Shortest Queue

- JSQ is optimal for general inter-arrival times and *exponential service times* (Hordijk and Koole (1990), Sparaggis *et al.* (1993)

Join the Shortest Queue

- JSQ is optimal for general inter-arrival times and *exponential service times* (Hordijk and Koole (1990), Sparaggis *et al.* (1993)

- Performance analysis is complicated
- How to dimension the system (number of servers, buffer size)?
- No results on general service times

- Similar optimality results for JSQ with infinite buffer: arbitrary arrival process, service time distribution with decreasing hazard rate

- counterexample of Whitt
- No easy way to compute performance

Asymptotic analysis: infinite buffer

- JSQ(d)
  - Pioneering work of Vdvenskaya *et al.* and Mitzenmacher (1996): introduced mean-field limits for exponential service times
  - Bramson *et al.* (2012): mean-field for FIFO and decreasing hazard rate

- JSQ
  - Graham (2000): mean field, exponential
  - Eschenfeldt and Gamarnik (2015): heavy-traffic, exponential

- JIQ
  - Stolyar (2015): mean-field optimality, exponential
  - Mukherjee *et al.* (2016) Halfin-Whitt and diffusion, exponential

Asymptotic analysis: finite buffer

- JSQ(d)
  - Xie *et al.* (2015): mean-field, exponential
  - Mukhopadhyay *et al.* (2015): mean-field, exponential, heterogeneous server speeds

Asymptotic analysis: finite buffer

- JSQ(d)
  - Xie *et al.* (2015): mean-field, exponential
  - Mukhopadhyay *et al.* (2015): mean-field, exponential, heterogeneous server speeds

- Mostly limited to exponential distribution
- Even then, mainly mean-field limits

Asymptotic analysis: finite buffer

- JSQ(d)
  - Xie *et al.* (2015): mean-field, exponential
  - Mukhopadhyay *et al.* (2015): mean-field, exponential, heterogeneous server speeds

- Mostly limited to exponential distribution
- Even then, mainly mean-field limits

- no simple formulas for performance measures $\Rightarrow$ no simple dimensioning rules

Insensitivity

- Erlang formula (1917) for blocking is insensitive to higher moments of the service time distribution.

Erlang formula = simple and robust dimensioning rule

Insensitivity

- Erlang formula (1917) for blocking is insensitive to higher moments of the service time distribution.

Erlang formula = simple and robust dimensioning rule

- $1970s$ onwards lots on interest in insensitive process-sharing networks: Muntz, Schassberger, Whittle, Kelly
- What are the requirements for a policy to be insensitive? Quasi-reversibility (partial balance equations)

Insensitivity

- Erlang formula (1917) for blocking is insensitive to higher moments of the service time distribution.

Erlang formula = simple and robust dimensioning rule

- $1970s$ onwards lots on interest in insensitive process-sharing networks: Muntz, Schassberger, Whittle, Kelly
- What are the requirements for a policy to be insensitive? Quasi-reversibility (partial balance equations)

+ Insensitivity $\Rightarrow$ robustness with respect to service time distribution
+ Closed-form stationary distribution $\Rightarrow$ formulae for performance measures

Insensitivity

- Erlang formula (1917) for blocking is insensitive to higher moments of the service time distribution.

Erlang formula = simple and robust dimensioning rule

- $1970s$ onwards lots on interest in insensitive process-sharing networks: Muntz, Schassberger, Whittle, Kelly
- What are the requirements for a policy to be insensitive? Quasi-reversibility (partial balance equations)

+ Insensitivity $\Rightarrow$ robustness with respect to service time distribution
+ Closed-form stationary distribution $\Rightarrow$ formulae for performance measures
- suboptimality

Insensitivity

- Erlang formula (1917) for blocking is insensitive to higher moments of the service time distribution.

Erlang formula = simple and robust dimensioning rule

- $1970s$ onwards lots on interest in insensitive process-sharing networks: Muntz, Schassberger, Whittle, Kelly
- What are the requirements for a policy to be insensitive? Quasi-reversibility (partial balance equations)

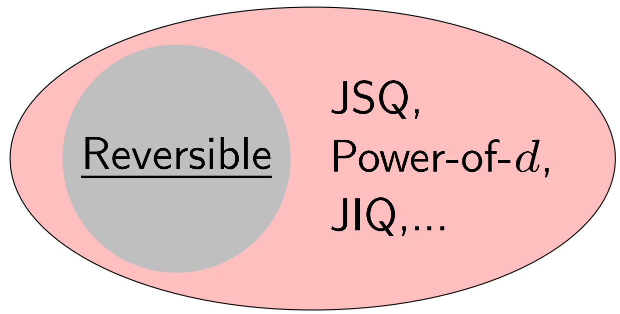
\+ Insensitivity $\Rightarrow$ robustness with respect to service time distribution

\+ Closed-form stationary distribution $\Rightarrow$ formulae for performance measures

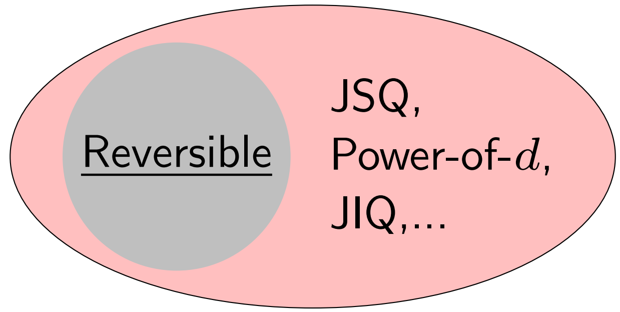
\- suboptimality

- Bonald and Proutire (2002): insensitive bandwidth-sharing networks

Insensitive load balancing



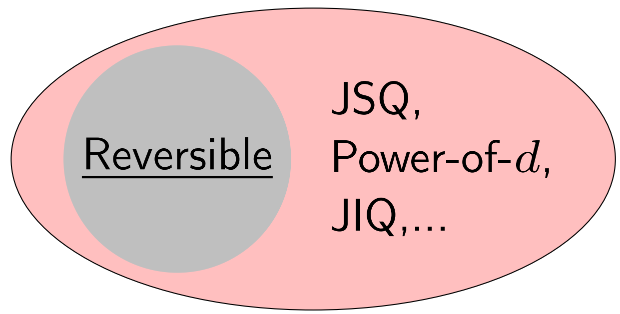
Insensitive load balancing



- Bonald, Proutière, Jonckheere (2004): optimal insensitive load balancing policy
  Route an arrival to server i with probability:

$$p_i(x_1, \dots, x_n) = \frac{\theta_i - x_i}{\sum_j \theta_j - x_j}.$$

Insensitive load balancing



- Bonald, Proutière, Jonckheere (2004): optimal insensitive load balancing policy
  Route an arrival to server i with probability:

$$p_i(x_1, \ldots, x_n) = \frac{\theta_i - x_i}{\sum_j \theta_j - x_j}.$$

+ Explicit stationary distribution for all job-size disitributions.
- Not very useful for $\theta = \infty$. Is equivalent to Bernoulli routing (Jonckheere (2006))

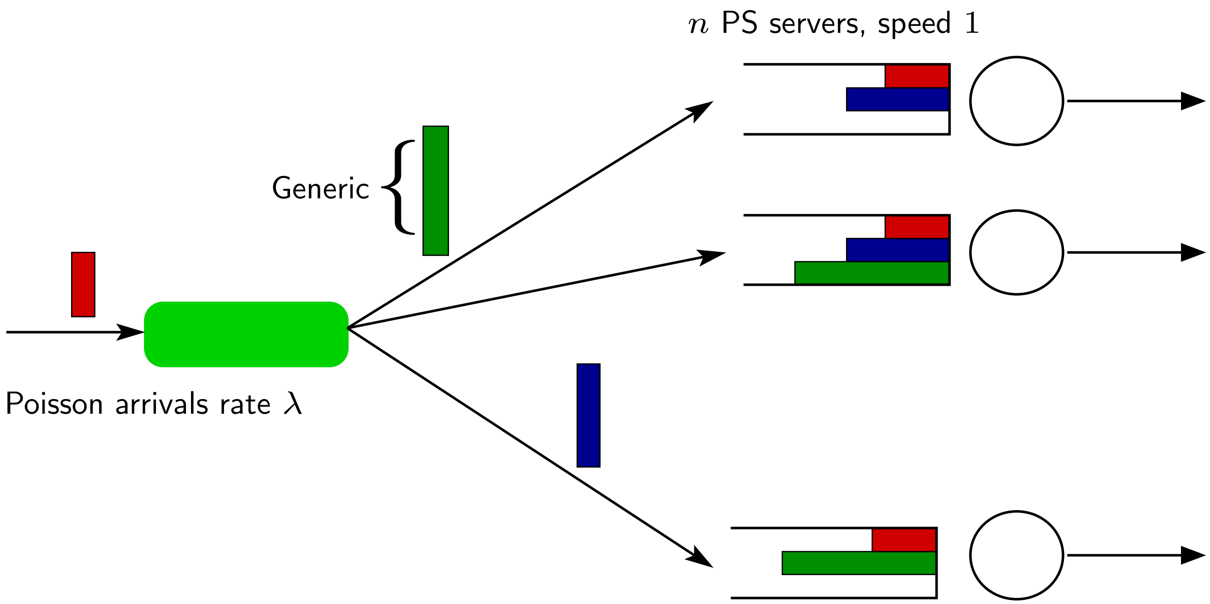
Objectives

- Performance measures in various asymptotic regimes
- Simple but non-trivial dimensioning rules

Objectives

- Performance measures in various asymptotic regimes
- Simple but non-trivial dimensioning rules
- Bounds for optimal policy
- Benchmarks for heuristics

Model



- Buffer size : θ at each server

Preliminaries

- Let $\mathbf{X}(t) = (X_i(t))_{i=1,\ldots n}$ be the number of tasks in server i at time t
- In state $\mathbf{x}$, a task is routed to server i with probability

$$\frac{\theta - x_i}{\sum_j (\theta - x_j)}. \tag{1}$$

- If the service times are i.i.d. exponential, then
  1. $\mathbf{X}(t)$ is a Markov process (birth-death) on $\mathbb{Z}_+^n$
  2. $\mathbf{X}(t)$ is reversible

Preliminaries

- Let $\mathbf{X}(t) = (X_i(t))_{i=1,\dots n}$ be the number of tasks in server i at time t
- In state $\mathbf{x}$, a task is routed to server i with probability

$$\frac{\theta - x_i}{\sum_j (\theta - x_j)}. \tag{1}$$

- If the service times are i.i.d. exponential, then
  1. $\mathbf{X}(t)$ is a Markov process (birth-death) on $\mathbb{Z}_+^n$
  2. $\mathbf{X}(t)$ is reversible

- $X(t)$ is insensitive to higher moments of the service time distribution.

Stationary distribution

- $\mathbf{X}(t)$ has closed-form stationary distribution

$$\pi(\mathbf{x}) = \frac{\Lambda(\mathbf{x})\Phi(\mathbf{x})}{\sum_{\mathbf{y}\in\mathbf{X}} \Phi(\mathbf{y})\Lambda(\mathbf{y})}, \tag{2}$$

with $\Phi(\mathbf{x}) = \prod_{i=1}^{n} \mu^{-x_i}$, and

$$\Lambda(\mathbf{x}) = \binom{|\theta - \mathbf{x}|}{\theta - \mathbf{x}} \lambda^{|\mathbf{x}|}, \tag{3}$$

where $\binom{|\theta-\mathbf{x}|}{\theta-\mathbf{x}} = \frac{|\theta-\mathbf{x}|!}{\prod_{i=1}^{n}(\theta-x_i)!}$ are the multinomial coefficients.

Stationary distribution

- $\mathbf{X}(t)$ has closed-form stationary distribution

$$\pi(\mathbf{x}) = \frac{\Lambda(\mathbf{x})\Phi(\mathbf{x})}{\sum_{\mathbf{y}\in\mathbf{X}} \Phi(\mathbf{y})\Lambda(\mathbf{y})}, \tag{2}$$

  with $\Phi(\mathbf{x}) = \prod_{i=1}^{n} \mu^{-x_i}$, and

$$\Lambda(\mathbf{x}) = \binom{|\theta - \mathbf{x}|}{\theta - \mathbf{x}} \lambda^{|\mathbf{x}|}, \tag{3}$$

  where $\binom{|\theta-\mathbf{x}|}{\theta-\mathbf{x}} = \frac{|\theta-\mathbf{x}|!}{\prod_{i=1}^{n}(\theta-x_i)!}$ are the multinomial coefficients.

- Blocking probability (apply PASTA): $\pi(\theta)$

Alternative representation

- Aggregate the servers according to the number of tasks.
- Let $\{S^{(n)}(t) \in \mathcal{S}\}_{t\geq 0}$ be the number of servers with i jobs at time t, with

$$\mathcal{S} = \{\mathbf{s} \in \{0, 1, \ldots, n\}^{\theta+1} : \sum_{i=0}^{\theta} s_i = n\}.$$

- Local arrival rate

$$\lambda_i(\mathbf{s}) = \lambda \frac{(\theta - i)s_i}{n\theta - \bar{s}}, \tag{4}$$

where $\bar{s} = \sum_{i=0}^{\theta} i s_i$.

Alternative representation

- $S^{(n)}(t)$ is a continuous-time jump Markov process on $\mathcal{S}$ with transition rates

$$S^{(n)}(t) \to \begin{cases} S^{(n)}(t) + e_i - e_{i-1} & \text{at rate } \lambda_{i-1}(s), i \geq 1; \\ S^{(n)}(t) + e_i - e_{i+1} & \text{at rate } s_{i+1}, \end{cases} \tag{5}$$

Alternative representation

- $S^{(n)}(t)$ is a continuous-time jump Markov process on $\mathcal{S}$ with transition rates

$$S^{(n)}(t) \to \begin{cases} S^{(n)}(t) + e_i - e_{i-1} & \text{at rate } \lambda_{i-1}(s),\, i \geq 1; \\ S^{(n)}(t) + e_i - e_{i+1} & \text{at rate } s_{i+1}, \end{cases} \tag{5}$$

- $S^{(n)}(t)$ inherits the insensitivity property of $\mathbf{X}(t)$

Theorem 1. *Its stationary distribution is given by*

$$\pi^{(n)}(s) = \pi_0^{(n)} \frac{(n\theta - \bar{s})!}{(n\theta)!} \binom{n}{s} \prod_{k=0}^{\theta} \left(\frac{\theta!}{(\theta - k)!} (n\rho)^k \right)^{s_k}, \tag{6}$$

where $\rho = \lambda/n$ is the load per server, and $\pi_0^{(n)}$ is the probability of the state with all servers empty, that is, $\bar{s} = 0$ and $s = (n, 0, \ldots, 0)$.

Alternative representation

Proof. Check that $\pi^{(n)}(s)$ satisfies the local balance equations (sufficient condition)

Alternative representation

Proof. Check that $\pi^{(n)}(s)$ satisfies the local balance equations (sufficient condition)

Take two states s and $s + e_i - e_{i-1} \in \mathcal{S}$.

$$\frac{\pi^{(n)}(s + e_i - e_{i-1})}{\pi^{(n)}(s)} = \frac{\lambda(\theta - (i-1))s_{i-1}}{n\theta - \bar{s}} \frac{1}{(s_i + 1)}, \tag{7}$$

$$= \frac{\lambda_{i-1}(s)}{(s_i + 1)} \tag{8}$$

Alternative representation

Proof. Check that $\pi^{(n)}(s)$ satisfies the local balance equations (sufficient condition)

Take two states s and $s + e_i - e_{i-1} \in \mathcal{S}$.

$$\frac{\pi^{(n)}(s + e_i - e_{i-1})}{\pi^{(n)}(s)} = \frac{\lambda(\theta - (i-1))s_{i-1}}{n\theta - \bar{s}} \frac{1}{(s_i + 1)}, \tag{7}$$

$$= \frac{\lambda_{i-1}(s)}{(s_i + 1)} \tag{8}$$

$$(s_i + 1)\pi^{(n)}(s + e_i - e_{i-1}) = \pi^{(n)}(s)\lambda_{i-1}(s) \tag{9}$$

□

Alternative representation

Proof. Check that $\pi^{(n)}(s)$ satisfies the local balance equations (sufficient condition)

Take two states s and $s + e_i - e_{i-1} \in \mathcal{S}$.

$$\frac{\pi^{(n)}(s + e_i - e_{i-1})}{\pi^{(n)}(s)} = \frac{\lambda(\theta - (i-1))s_{i-1}}{n\theta - \bar{s}} \frac{1}{(s_i + 1)}, \tag{7}$$

$$= \frac{\lambda_{i-1}(s)}{(s_i + 1)} \tag{8}$$

$$(s_i + 1)\pi^{(n)}(s + e_i - e_{i-1}) = \pi^{(n)}(s)\lambda_{i-1}(s) \tag{9}$$

□

Corollary 1. *Using the PASTA property, the blocking probability is given by*

$$B_\theta^{(n)} = \pi_0^{(n)} \frac{(n\rho)^{n\theta}(\theta!)^n}{(n\theta)!}. \tag{10}$$

Special case: Erlang loss system

- For $\theta = 1$, we get the classical $M/M/n/n$ queue or the Erlang loss system.

$$\pi^{(n)}(s_0) = \frac{(n\rho)^{(n-s_0)}}{(n-s_0)!}\pi_0^{(n)}, \tag{11}$$

where

$$\pi_0^{(n)} = \sum_{k \le n} \frac{(n\rho)^{n-k}}{(n-k)!}, \tag{12}$$

Asymptotic analysis

1. Mean field limit
2. Large deviations
3. Halfin-Whitt limit
4. Moderate and small deviations

Mean-field limit

- Limit $n \to \infty$, for a fixed $\rho < 1$.

Mean-field limit

- Limit $n \to \infty$, for a fixed $\rho < 1$.

Theorem 2. *Let* $y(0) = \lim_{n\to\infty} \frac{S^{(n)}(0)}{n}$. *For exponentially distributed job-sizes, for all* t, $S^{(n)}(t)/n \to y(t)$, *in probability, with* y *the solution of:*

$$\frac{dy_j(t)}{dt} = \rho \frac{\theta - (j-1)}{\theta - \sum_k k y_k(t)} y_{j-1}(t) + y_{j+1}(t) \tag{13}$$

$$- \rho \frac{\theta - j}{\theta - \sum_k k y_k(t)} y_j(t) - y_j(t),\ 0 < j < \theta,$$

$$\frac{dy_\theta(t)}{dt} = \rho \frac{1}{\theta - \sum_k k y_k(t)} y_{\theta-1}(t) - y_\theta(t), \tag{14}$$

$$\frac{dy_0(t)}{dt} = y_1(t) - \rho \frac{\theta}{\theta - \sum_k k y_k(t)} y_0(t). \tag{15}$$

Mean-field limit : steady-state solution

- The stationary point of the differential equations is obtained upon taking $t \to \infty$.

Theorem 3. *For $0 < \rho \leq 1$, the unique steady-state solution of the system of equations* (13)–(15) *is given by*

$$\hat{p}_j = \left(\frac{\theta - \hat{c}}{\rho}\right)^{\theta - j} \frac{1}{(\theta - j)!} \hat{p}_\theta, \tag{16}$$

$$\text{with } \hat{p}_\theta = \frac{1}{\sum_{k=0}^{\theta} \left(\frac{\theta - \hat{c}}{\rho}\right)^k \frac{1}{k!}}. \tag{17}$$

where

$$\hat{c} = \theta - \rho \zeta_\theta^{-1}(1 - \rho), \tag{18}$$

with ζ_θ^{-1} as the inverse function of the Erlang blocking viewed as a function of the traffic intensity for a fixed buffer depth θ.

If $\rho > 1$, the unique solution is $\hat{c} = \theta$, $\hat{p}_j = 0$, for $j \leq \theta - 1$ and $\hat{p}_\theta = 1$.

Mean-field limit : interchange of limits

- Does an interchange of the order of limits lead to the same limit?

$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{S^{(n)}(t)}{n} = \lim_{n\to\infty}\lim_{t\to\infty}\frac{S^{(n)}(t)}{n}? \tag{19}$$

Mean-field limit : interchange of limits

- Does an interchange of the order of limits lead to the same limit?

$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{S^{(n)}(t)}{n} = \lim_{n\to\infty}\lim_{t\to\infty}\frac{S^{(n)}(t)}{n}? \tag{19}$$

Proposition 1. *For $\rho < 1$, $\pi^{(n)}$ converges point wise to $\hat{p}$ when n and t converge to infinity.*

Proof. A corollary of Le Boudec's result for reversible Markov process. □

Remark 1. *By insensitivity, $\hat{p}$ is the limiting distribution of $\pi^{(n)}$ independent of the specific job-size distribution*

Mean-field limit : blocking probability

- A lower bound on the blocking probability

Proposition 2. *For* $\theta > 0$*, the blocking probability of any non-anticipating and size-unaware load balancing policy is greater than* $\max(0, 1 - \rho^{-1})$*.*

Proof. Cannot do better than the system with all the buffer and server capacity pooled. □

Mean-field limit : blocking probability

- A lower bound on the blocking probability

Proposition 2. *For $\theta > 0$, the blocking probability of any non-anticipating and size-unaware load balancing policy is greater than* $\max(0, 1-\rho^{-1})$.

Proof. Cannot do better than the system with all the buffer and server capacity pooled. □

- Blocking probability of the insensitive policy

Proposition 3. *The limiting blocking probability of the insensitive load balancing policy is given by*

$$B_\theta = \begin{cases} 0 & \textit{if } \rho < 1; \\ 1-\rho^{-1} & \textit{otherwise.} \end{cases} \tag{20}$$

Mean-field limit : blocking probability

- A lower bound on the blocking probability

  Proposition 2. *For $\theta > 0$, the blocking probability of any non-anticipating and size-unaware load balancing policy is greater than* $\max(0, 1-\rho^{-1})$.

  Proof. Cannot do better than the system with all the buffer and server capacity pooled. □

- Blocking probability of the insensitive policy

  Proposition 3. *The limiting blocking probability of the insensitive load balancing policy is given by*

$$B_\theta = \begin{cases} 0 & \text{if } \rho < 1; \\ 1 - \rho^{-1} & \text{otherwise.} \end{cases} \qquad (20)$$

- Insensitive policy is globally optimal in the mean-field limit

Mean-field limit : blocking probability

- A lower bound on the blocking probability

  Proposition 2. *For $\theta > 0$, the blocking probability of any non-anticipating and size-unaware load balancing policy is greater than* $\max(0, 1 - \rho^{-1})$.

  Proof. Cannot do better than the system with all the buffer and server capacity pooled. □

- Blocking probability of the insensitive policy

  Proposition 3. *The limiting blocking probability of the insensitive load balancing policy is given by*

$$B_\theta = \begin{cases} 0 & \text{if } \rho < 1; \\ 1 - \rho^{-1} & \text{otherwise.} \end{cases} \tag{20}$$

- Insensitive policy is globally optimal in the mean-field limit
- Any empty space filling policy will achieve this...

Asymptotic analysis

1. Mean field limit

2. Large deviations

3. Halfin-Whitt limit

4. Moderate and small deviations)

Large deviations

- Let $\mathcal{P}_c = \{q \in \mathbb{R}_+^\theta : \sum_{i=0}^\theta q_i = 1 \text{ and } \sum_{i=0}^\theta i q_i = c\}$
- Define $p \in \mathcal{P}_c$ by

$$p_k(c) := \frac{1}{(\theta - k)!} \left(\frac{\theta - c}{\rho}\right)^{\theta - k} \frac{1}{\psi(c)}. \tag{21}$$

where

$$\psi(c) = \sum_{k=0}^{\theta} \frac{1}{k!} \left(\frac{\theta - c}{\rho}\right)^{k}, \tag{22}$$

Large deviations

- Let $\mathcal{P}_c = \{q \in \mathbb{R}_+^\theta : \sum_{i=0}^\theta q_i = 1 \text{ and } \sum_{i=0}^\theta iq_i = c\}$
- Define $p \in \mathcal{P}_c$ by

$$p_k(c) := \frac{1}{(\theta - k)!} \left(\frac{\theta - c}{\rho}\right)^{\theta - k} \frac{1}{\psi(c)}. \tag{21}$$

where

$$\psi(c) = \sum_{k=0}^{\theta} \frac{1}{k!} \left(\frac{\theta - c}{\rho}\right)^{k}, \tag{22}$$

- Note that $p(\hat{c})$ is the steady-state solution of the mean-field limit.

Large deviations

Theorem 4. *For $\rho < 1$, and $q \in \mathcal{P}_c$,*

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{\pi^{(n)}(q;c)}{\pi^{(n)}(p;\hat{c})}\right) = (c-\hat{c}) + \log\left(\frac{\psi(c)}{\psi(\hat{c})}\right) - D_{KL}(q(c)\|p(c)), \quad (23)$$

where D_{KL} is the Kullback-Liebler divergence.

- exponential decay in n in the probability of observing any distribution other than $p(\hat{c})$.

Large deviations

Theorem 4. *For* $\rho < 1$, *and* $q \in \mathcal{P}_c$,

$$\lim_{n\to\infty} \frac{1}{n} \log \left(\frac{\pi^{(n)}(q;c)}{\pi^{(n)}(p;\hat{c})} \right) = (c - \hat{c}) + \log \left(\frac{\psi(c)}{\psi(\hat{c})} \right) - D_{KL}(q(c) \| p(c)), \qquad (23)$$

where D_{KL} *is the Kullback-Liebler divergence.*

- exponential decay in n in the probability of observing any distribution other than $p(\hat{c})$.
- $p(c)$ is the most likely distribution that is observed conditioned on c.

Large deviations: blocking probability

Theorem 5. *For* $\rho \in (0,1)$,

$$\lim_{n\to\infty} B_\theta^{(n)} \exp(nR(\gamma_{\theta,\rho})) \left(\frac{2\pi n}{\alpha_{\theta,\rho}}\right)^{1/2} = 1. \tag{24}$$

where

$$R(t) = \log\left(\sum_{k=0}^{\theta} \frac{t^k}{k!}\right) - \rho t, \quad \gamma_{\theta,\rho} = \arg\max_{t\in(0,\infty)} R(t) = \frac{\theta - \hat{c}}{\rho}, \tag{25}$$

$$\alpha_{\theta,\rho} = \frac{(1-\rho)}{\rho}\left(\frac{\theta}{\rho\gamma_{\theta,\rho}} - 1\right). \tag{26}$$

Large deviations: blocking probability

Theorem 5. *For* $\rho \in (0,1)$,

$$\lim_{n\to\infty} B_\theta^{(n)} \exp(nR(\gamma_{\theta,\rho})) \left(\frac{2\pi n}{\alpha_{\theta,\rho}}\right)^{1/2} = 1. \tag{24}$$

where

$$R(t) = \log\left(\sum_{k=0}^{\theta} \frac{t^k}{k!}\right) - \rho t, \quad \gamma_{\theta,\rho} = \arg\max_{t\in(0,\infty)} R(t) = \frac{\theta - \hat{c}}{\rho}, \tag{25}$$

$$\alpha_{\theta,\rho} = \frac{(1-\rho)}{\rho}\left(\frac{\theta}{\rho\gamma_{\theta,\rho}} - 1\right). \tag{26}$$

Corollary 2. *For* $\theta = 1$, $\gamma_{\theta,\rho} = \frac{1-\rho}{\rho}^{-1}$ *and* $\alpha_{\theta,\rho} = 1$. *Thus,*

$$B_1^{(n)} \sim e^{n(1-\rho)}\rho^n(2\pi n)^{-1/2}. \tag{27}$$

Asymptotic analysis

1. Mean field limit

2. Large deviations

3. Halfin-Whitt-Jagerman limit

4. Moderate and small deviations

Halfin-Whitt-Jagerman limit

- Arrival rate $\lambda \uparrow \infty$. How should the number of servers scale?

$$n = \rho^{-1}\lambda$$

Halfin-Whitt-Jagerman limit

- Arrival rate $\lambda \uparrow \infty$. How should the number of servers scale?

$$n = \rho^{-1}\lambda$$

$\underline{\rho < 1}$

+ High quality: $B_\theta^{(n)} \sim e^{-Cn}$
- Low efficiency (low server utilisation): $n(1-\hat{p}_0)$ servers empty

$\underline{\rho > 1}$

- Low quality: $B_\theta^{(n)} \sim 1 - \rho^{-1}$
+ High efficiency: utilisation ~ 1

Halfin-Whitt-Jagerman limit

- Arrival rate $\lambda \uparrow \infty$. How should the number of servers scale?

$$n = \rho^{-1}\lambda$$

$\underline{\rho < 1}$

+ High quality: $B_\theta^{(n)} \sim e^{-Cn}$
- Low efficiency (low server utilisation): $n(1 - \hat{p}_0)$ servers empty

$\underline{\rho > 1}$

- Low quality: $B_\theta^{(n)} \sim 1 - \rho^{-1}$
+ High efficiency: utilisation ~ 1

- For $\theta = 1$, Quality and Efficiency Driven regime (H-W, Jagerman):

$$n = \lambda + \alpha\sqrt{\lambda}$$ Square-root staffing rule

- Good quality: $B_1^{(n)} \sim n^{-1/2}$; Good efficiency: server utilization ~ 1

Halfin-Whitt-Jagerman limit

- How high we can push ρ and still have asymptotically negligible blocking probability?

Halfin-Whitt-Jagerman limit

- How high we can push ρ and still have asymptotically negligible blocking probability?

  Theorem 6. *For* $a \in (-\infty, \infty)$, *let*

$$n\rho = n + an^{1/(\theta+1)}. \tag{28}$$

  Then,

$$\lim_{n\to\infty} B_\theta^{(n)} n^{\theta/(\theta+1)} \int_0^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du = 1. \tag{29}$$

- $\rho = 1 + an^{-\theta/(\theta+1)}$

Halfin-Whitt-Jagerman limit: observations

Corollary 3. *If* $\rho = 1$*:*

$$B_\theta^{(n)} \sim \frac{(\theta+1)!^{\frac{1}{\theta+1}}}{\theta+1} \Gamma\Big(\frac{1}{\theta+1}\Big) n^{-\theta/(\theta+1)}, \tag{30}$$

where Γ *is the Gamma function.*

Halfin-Whitt-Jagerman limit: observations

Corollary 3. *If* $\rho = 1$*:*

$$B_\theta^{(n)} \sim \frac{(\theta+1)!^{\frac{1}{\theta+1}}}{\theta+1}\Gamma\Big(\frac{1}{\theta+1}\Big)n^{-\theta/(\theta+1)}, \qquad (30)$$

where Γ *is the Gamma function.*

Corollary 4.

$$B_1^{(n)} \sim (0.5\pi n)^{-1/2}. \qquad (31)$$

- Order of decay increases with θ: $n^{-1/2}$ for $\theta = 1$ and n^{-1} for $\theta = \infty$
- Higher the θ, closer ρ can be to 1 for the same blocking probability

Trichotomy of ILB

$\rho < 1$	Critical regime $\rho_n = 1 + an^{-\frac{\theta}{\theta+1}}$	$\rho > 1$
Blocking $\sim e^{-C(\theta)n}$	Blocking $\sim n^{\frac{-\theta}{\theta+1}}$	Blocking $= 1 - \rho^{-1}$

- $\rho < 1$, the blocking is exponential small in n (Large deviations)
- Generalized HWJ:

$$\rho_n = 1 + an^{-\frac{\theta}{\theta+1}}.$$

- $\rho > 1$, the blocking is constant

Asymptotic analysis

1. Mean field limit
2. Large deviations
3. Halfin-Whitt-Jagerman limit
4. Moderate and small deviations

Moderate deviations

Theorem 7 (Central limit). *For* $\rho < 1$,

$$\frac{1}{\sqrt{n}}\left(\left(S^{(n)}(\infty)\right)_{0\leq i<\theta} - n(\hat{p})_{0\leq i<\theta}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\Sigma), \tag{32}$$

where

$$\begin{aligned}\Sigma^{-1} &= \psi(1,1,\dots,1)\cdot(1,1,\dots,1)^{\top} \\ &- \left(\frac{1}{\theta-\hat{c}}\right)(\theta,\theta-1,\dots,1)\cdot(\theta,\theta-1,\dots,1)^{\top} \\ &+ \begin{pmatrix} 1/\hat{p}_0 & 0 & \dots & 0 \\ 0 & 1/\hat{p}_1 & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\hat{p}_{\theta-1} \end{pmatrix}\end{aligned} \tag{33}$$

Moderate deviations

- Define

$$\widehat{\Phi}_\theta(z;a) = \int_z^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du. \tag{34}$$

Theorem 8. *For* $\rho = 1$ *and* $z \in \mathbb{R}_+$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_{\theta-1}^{(n)}(\infty)}{n^{\theta/(\theta+1)}} > z\right) = \frac{\widehat{\Phi}_\theta(z;0)}{\widehat{\Phi}_\theta(0;0)}, \tag{35}$$

Moderate deviations

- Define

$$\widehat{\Phi}_\theta(z;a) = \int_z^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du. \tag{34}$$

Theorem 8. *For* $\rho = 1$ *and* $z \in \mathbb{R}_+$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_{\theta-1}^{(n)}(\infty)}{n^{\theta/(\theta+1)}} > z\right) = \frac{\widehat{\Phi}_\theta(z;0)}{\widehat{\Phi}_\theta(0;0)}, \tag{35}$$

- Variations are visible only in θ and $\theta - 1$.

Moderate deviations

- Define
$$\widehat{\Phi}_\theta(z;a) = \int_z^{\infty} \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du. \tag{34}$$
Theorem 8. *For* $\rho = 1$ *and* $z \in \mathbb{R}_+$,
$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_{\theta-1}^{(n)}(\infty)}{n^{\theta/(\theta+1)}} > z\right) = \frac{\widehat{\Phi}_\theta(z;0)}{\widehat{\Phi}_\theta(0;0)}, \tag{35}$$

- Variations are visible only in θ and $\theta - 1$.
- Number of servers having i jobs $O(n^{(i+1)/(\theta+1)})$.

Small deviations

Theorem 9. *For* $\rho > 1$,

$$S_{\theta-1}^{(n)}(\infty) \xrightarrow[n\to\infty]{d} Geo(\rho^{-1}), \tag{36}$$

and the blocking probability is

$$B_{\theta}^{(n)} \sim 1 - \rho^{-1}. \tag{37}$$

- Deviations are of constant size, and happen in θ and $\theta - 1$.

Outline

- Results for finite systems
- Asymptotic analysis
- Numerical results
- Open problems

Numerical results

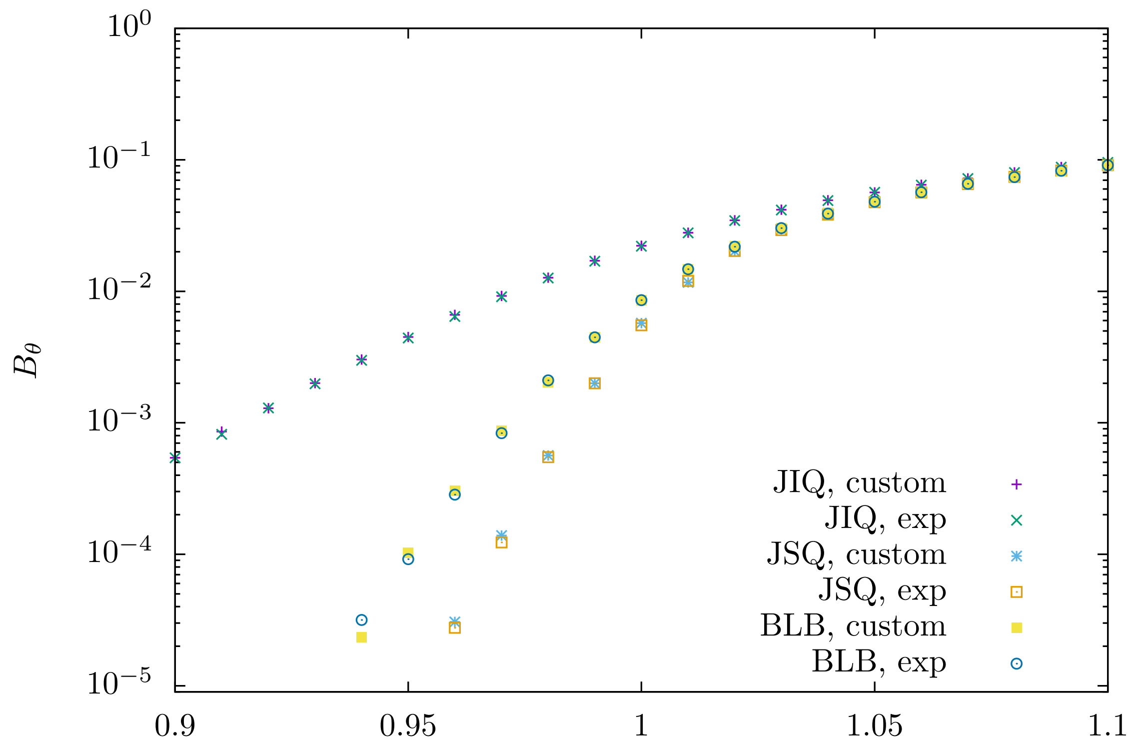


Figure 1: Comparison of the blocking probability for different load balancing policies. Number of servers is 20. Buffer size is 10.

Numerical results

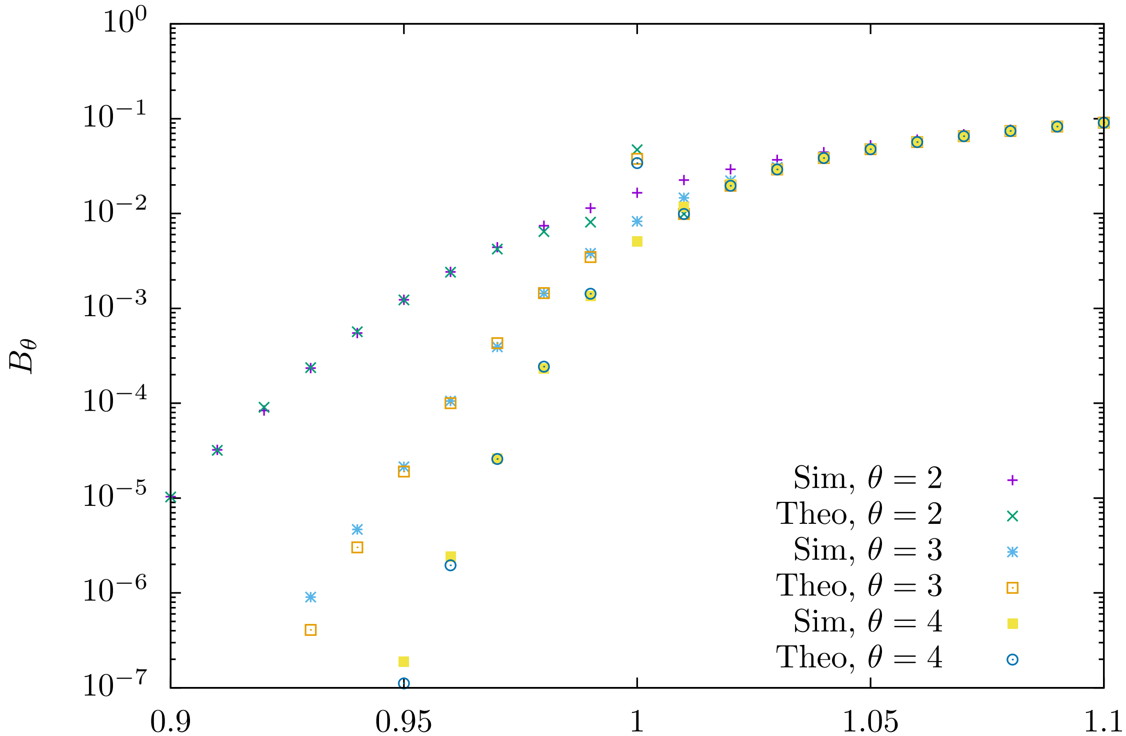


Figure 2: Comparison of the blocking probability computed from Theorems 5 and 9 with that obtained from simulations. Number of servers is 200.

Outline

- Results for finite systems
- Asymptotic analysis
- Numerical results
- Open problems

Open problems

- Is the HWJ scaling optimal?
- How does the optimality gaps for specific families of jobs-size distributions?
- Can similar results be established for sensitive policies like JSQ(d) or JIQ?
- Similar results for infinite buffer systems