Aalto University
School of Electrical
Engineering

Performance-Energy Trade-off in Multi-Server Queueing Systems with Setup Delay

Samuli Aalto
Aalto University, Finland

LCCC Cloud Computing Workshop
7–9 May 2014
Lund, Sweden

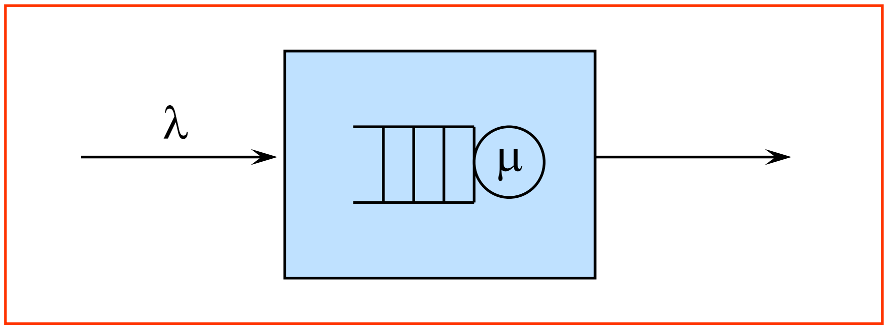
Co-operation with

Esa Hyytiä, Pasi Lassila, Misikir Gebrehiwot
(Aalto University)

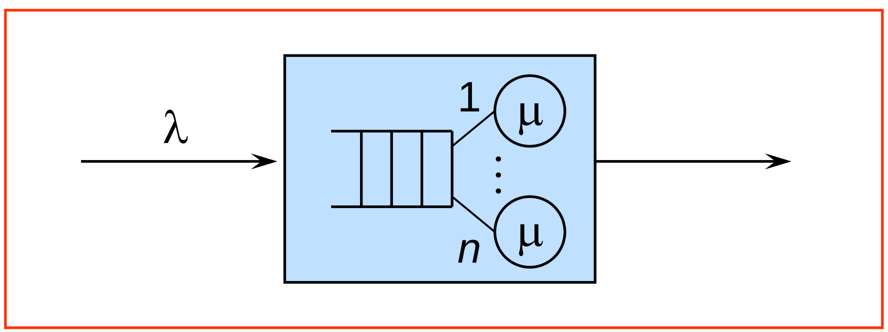
Rhonda Righter
(UC Berkeley)

Queueing models

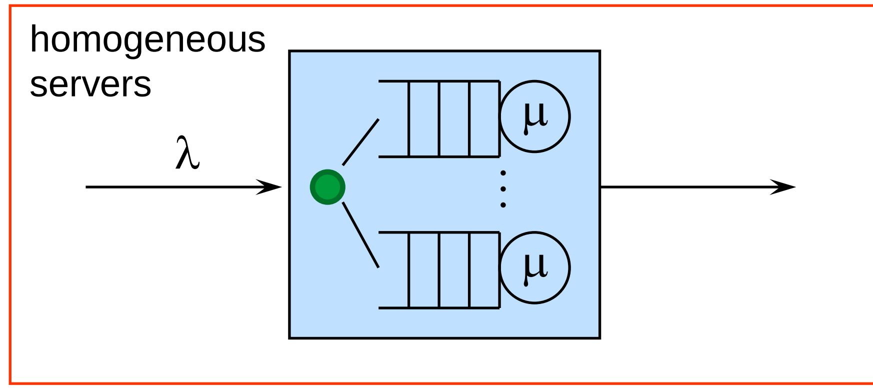
- Single-server queue (M/G/1)

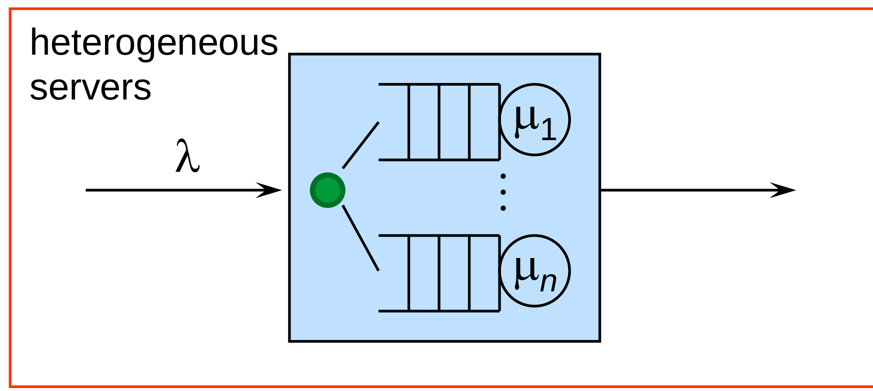


- Multi-server queue (M/M/*n*)



- Parallel queues

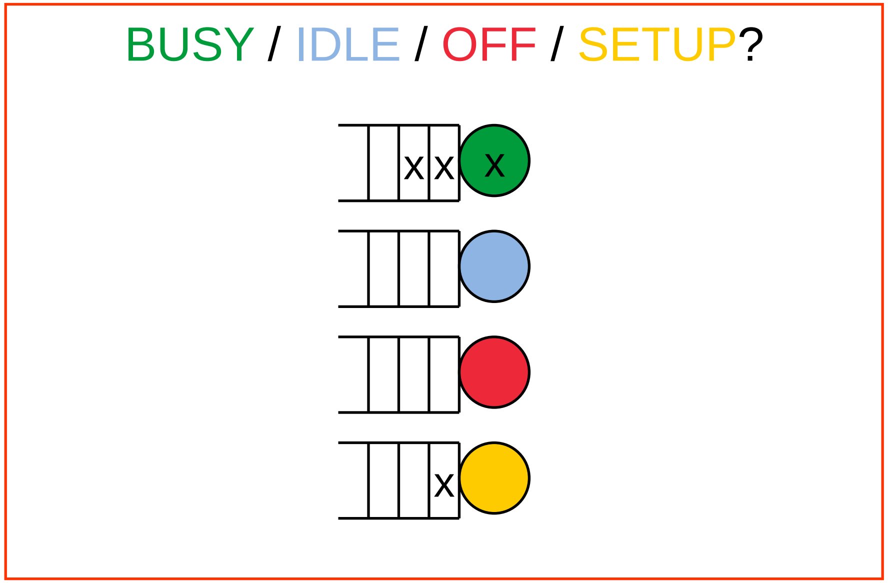




Performance-energy trade-off

- Energy saved by switching the server off when idle

- However, performance impaired, if switching the server back on takes time (setup delay)



Cost model

- Performance:

  E[T] = mean delay per job (in seconds)

  E[X] = mean number of jobs = $\lambda \cdot$E[T]

- Definition:
  delay = response time

- Energy:

  E[E] = mean energy per job (in joules)

  E[P] = mean power consumed = $\lambda \cdot$E[E]

  Power consumption levels:

$$0 = P_{\text{off}} < P_{\text{idle}} \le P_{\text{setup}} = P_{\text{busy}}$$

Objective function

- Energy-Response-time-Weighted-Sum (ERWS):

$$\mathrm{E}[T] + \mathrm{E}[E]/\beta$$

e.g. Wierman & al. (2009)

- Energy-Response-time-Product (ERP):

$$\mathrm{E}[T] \cdot \mathrm{E}[E]$$

e.g. Gandhi & al. (2010b)

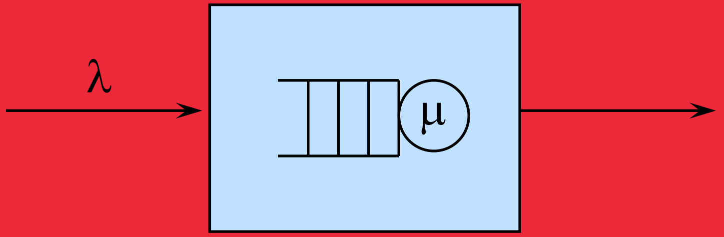
- General form:

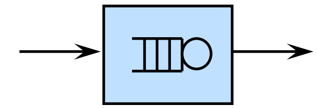
$$w_1 \cdot \mathrm{E}[T]^{t1} \cdot \mathrm{E}[E]^{e1} + w_2 \cdot \mathrm{E}[T]^{t2} \cdot \mathrm{E}[E]^{e2}$$

by Maccio & Down (2013)

  – ERWS:
    $w_1 = 1$, $t_1 = 1$, $e_1 = 0$
    $w_2 = 1/\beta$, $t_2 = 0$, $e_2 = 1$
  – ERP:
    $w_1 = 1$, $t_1 = 1$, $e_1 = 1$
    $w_2 = 0$, $t_2 = 0$, $e_2 = 0$

Part I
Single-server queue with setup delays





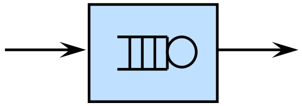
Optimal switching on/off policy

Maccio & Down (2013)

- M/G/1-FIFO
  - Setup delay *D* generally distributed with mean $1/\gamma$
- Control parameters:
  - Delayed switch-off for an exponential time with mean $1/\alpha$
  - Server switched on after *k* new job arrivals
- Objective function: Gen. form

$$w_1 \mathrm{E}[T]^{t_1} \mathrm{E}[E]^{e_1} + w_2 \mathrm{E}[T]^{t_2} \mathrm{E}[E]^{e_2}$$

- Policies:
  - NEVEROFF: $\alpha = 0$
  - DELAYEDOFF: $0 < \alpha < \infty$
  - INSTANTOFF: $\alpha = \infty$
- Theorem:
  For ERWS objective function optimal policy is either NEVEROFF or INSTANTOFF
  - Similar result in Gandhi & al. (2010b) for ERP objective function



Optimal switching on/off policy

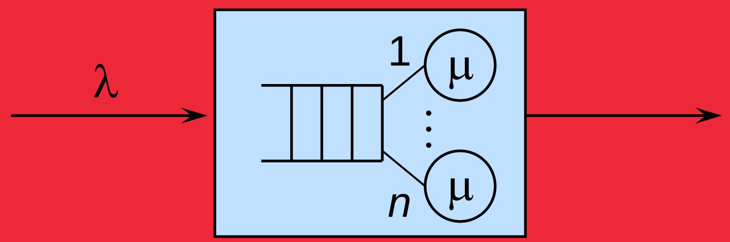
Gebrehiwot & al. (2014)

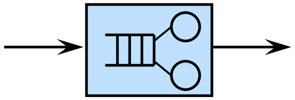
- M/G/1-FIFO
  - Setup delay *D* generally distributed with mean $1/\gamma$
- Control parameters:
  - Delayed switch-off for a gen. distributed time with mean $1/\alpha$
  - Server switched on after *k* new job arrivals
- Objective function: Gen. form

$$w_1 \mathrm{E}[T]^{t_1}\,\mathrm{E}[E]^{e_1} + w_2 \mathrm{E}[T]^{t_2}\,\mathrm{E}[E]^{e_2}$$

- Policies:
  - NEVEROFF: $\alpha = 0$
  - DELAYEDOFF: $0 < \alpha < \infty$
  - INSTANTOFF: $\alpha = \infty$
- Theorem:
  For gen. objective function optimal policy is either NEVEROFF or INSTANTOFF
  - NEVEROFF is better if P_{idle} is sufficiently small compared to P_{setup}

Part II
Multi-server queue with setup delays





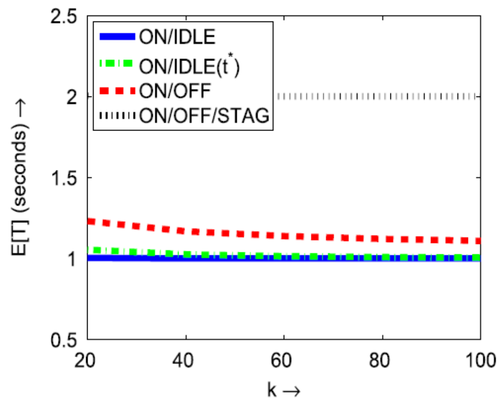
Analysis of server farms with setup delays

Gandhi & al. (2010a)

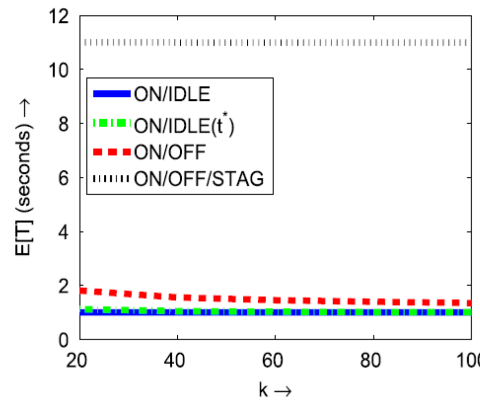
- M/M/*n*
  - Setup delay *D* exponentially distributed
- Objective function:
  Separately E[*T*] and E[*P*]
- Policies:
  - ON/IDLE = NEVEROFF
  - ON/OFF = INSTANTOFF
  - ON/OFF/STAG = INSTANTOFF with ”staggered bootup”
- Mixed policy:
  - ON/IDLE(*t*)
    switching idle server off only if nr of busy and idle servers > *t*
- Conclusions:
  - “Under high loads, turning servers off can result in higher power consumption and far higher response times.”
  - “As the size of the server farm is increased, the advantages of turning servers off increase.”

Analysis of server farms with setup delays

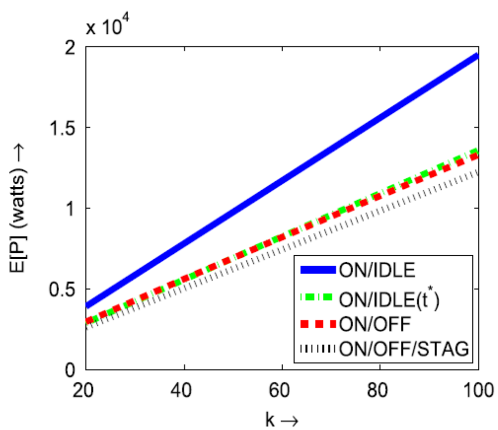
Gandhi & al. (2010a)



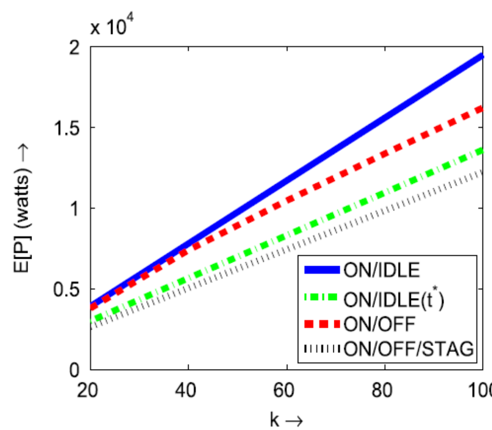
(a) **E** [T] for low setup time ($\alpha = 1\ \mathrm{s}^{-1}$).



(b) **E** [T] for high setup time ($\alpha = 0.1\ \mathrm{s}^{-1}$).



(c) **E** [P] for low setup time ($\alpha = 1\ \mathrm{s}^{-1}$).



(d) **E** [P] for high setup time ($\alpha = 0.1\ \mathrm{s}^{-1}$).

Optimization of server farms with setup delay

Gandhi & al. (2010b)

- M/M/n
  - Setup delay deterministic
  - Additional sleep states S with

$$0 = P_{\mathrm{off}} < P_{\mathrm{sleep}} < P_{\mathrm{idle}}$$

and deterministic (setup) delays

$$0 = d_{\mathrm{idle}} < d_{\mathrm{sleep}} < d_{\mathrm{off}}$$

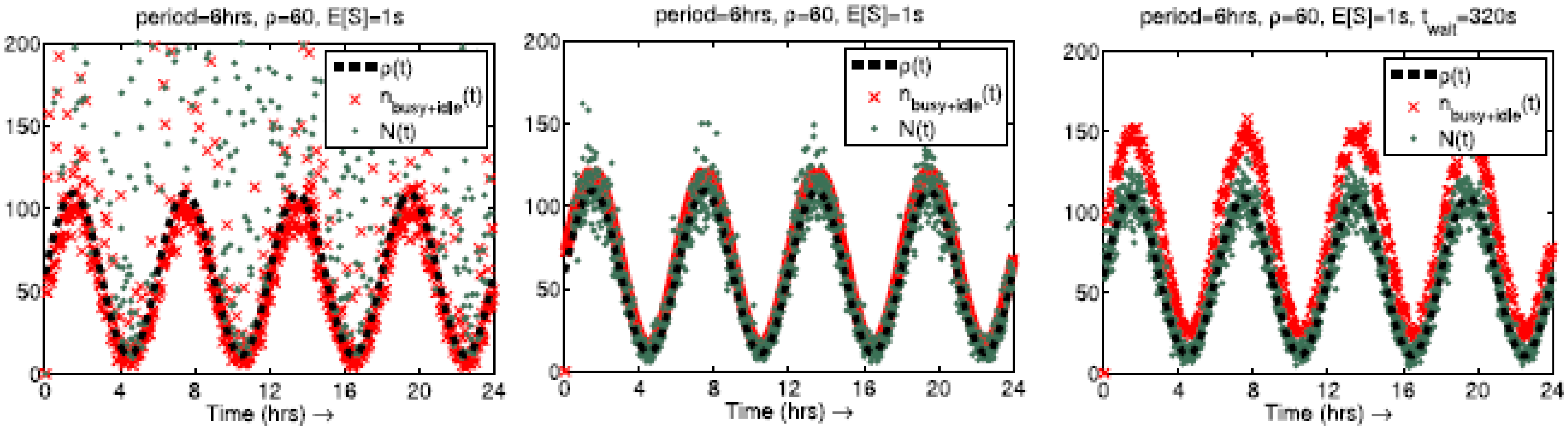
- Objective function: ERP

$$\mathrm{E}[T] \cdot \mathrm{E}[P]$$

- Policies:
  - NEVEROFF
  - INSTANTOFF
  - SLEEP(S)
  - Probabilistic and other
- Theorem:
  For $n = 1$, optimal static control is either NEVEROFF, INSTANTOFF or SLEEP(S)
- Robust policy:
  - DELAYEDOFF with MRB (Most Recent Busy)

Optimization of server farms with setup delay

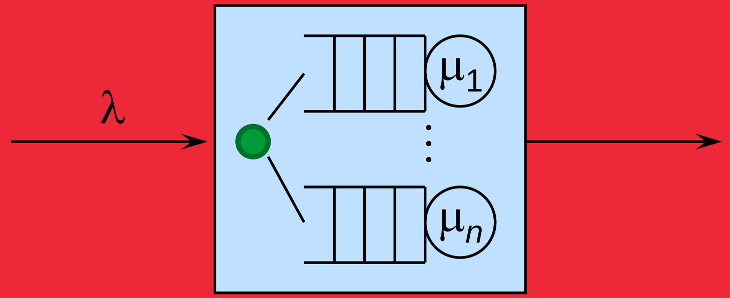
Gandhi & al. (2010b)

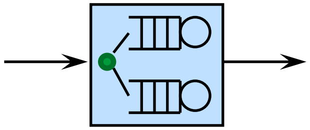


(a) INSTANTOFF. (b) LOOKAHEAD. (c) DELAYEDOFF.

Fig. 4. Dynamic capacity provisioning capabilities of INSTANTOFF, LOOKAHEAD and DELAYEDOFF. The dashed line denotes the load at time t, $\rho(t)$, the crosses denotes the number of servers that are busy or idle at time t, $n_{busy+idle}(t)$, and the dots represent the number of jobs in the system at time t, $N(t)$.

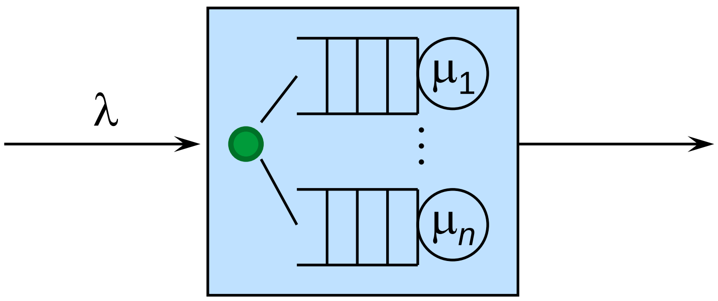
Part III
Parallel queues with setup delays



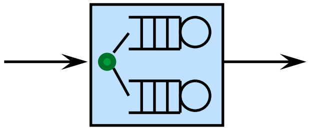


Dispatching problem

- Dispatching
  = Task assignment = Routing
  - Random job arrivals with random service requirements
  - Dispatching decision made upon the arrival



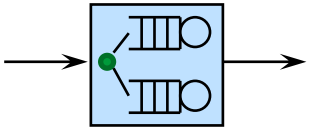
- Our setting: M/G/.
  - Poisson arrivals
  - generally distributed job sizes
  - heterogeneous servers with FIFO queueing discipline (NEVEROFF or INSTANTOFF)



Static dispatching policies

- RND =
  Bernoulli splitting
  – choose the queue pure randomly
  – no size nor state information needed

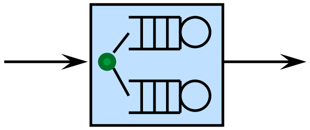
- SITA =
  Size Interval Task Assignment
  – choose the queue with similar jobs
  – based on the size of
  the arriving job, but
  no state information needed
  – Harchol-Balter et al. (1999)



MDP approach

- Any static policy (RND, SITA) results in parallel M/G/1 queues
- Fix the static policy and determine relative values for all these parallel M/G/1 queues
- Dispatch the arriving job to the queue that minimizes the mean additional costs
- As the result, you get a better dynamic dispatching policy
- This is called First Policy Iteration (FPI) in the MDP theory
- Applicable for the ERWS objective function

Relative values



- Definition:
  For a fixed policy resulting in a stable system, the value function $v(x)$ gives the expected difference in the infinite horizon cumulative costs between
  - the system initially in state x, and
  - the system initially in equilibrium

- Definition:
  For a fixed policy resulting in a stable system, the relative value $v(x) - v(0)$ gives the expected difference in the infinite horizon cumulative costs between
  - the system initially in state x, and
  - the system initially in state 0

Size-aware M/G/1 queue without setup delays

Hyytiä et al. (2012)

- State description:

$$u = \Delta_1 + \ldots + \Delta_n$$

  - Δ_i = remaining service time of job *i*
  - *u* = backlog = unfinished work

- Mean values:

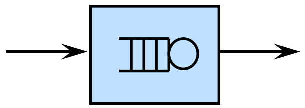
$$E[T] = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}$$

$$E[P] = (1-\rho) P_{\text{idle}} + \rho \cdot P_{\text{busy}}$$

- Result:
  Size-aware relative values

$$v_T(u) - v_T(0) = \frac{\lambda u^2}{2(1-\rho)}$$

$$v_P(u) - v_P(0) = u \cdot (P_{\text{busy}} - P_{\text{idle}})$$



Size-aware M/G/1 queue with setup delays

Hyytiä et al. (2014a)

- State description:

$$u = \Delta_0 + \Delta_1 + \ldots + \Delta_n$$

  - Δ_i = remaining service time of job i
  - Δ_0 = remaining setup delay
  - u = virtual backlog

- Assume:
  Deterministic setup delay d and

$$P_{\text{setup}} = P_{\text{busy}}$$

- Mean values:

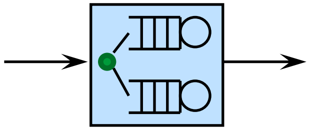
$$E[T] = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)} + \frac{d(2+\lambda d)}{2(1+\lambda d)}$$

$$E[P] = \frac{\rho + \lambda d}{1+\lambda d} \cdot P_{\text{busy}}$$

- Result:
  Size-aware relative values

$$v_T(u) - v_T(0) = \frac{\lambda u^2}{2(1-\rho)} - \frac{\lambda u d(2+\lambda d)}{2(1-\rho)(1+\lambda d)}$$

$$v_P(u) - v_P(0) = \frac{u}{1+\lambda d} \cdot P_{\text{busy}}$$



FPI policy

Hyytiä et al. (2012, 2014a)

- For NEVEROFF servers:

  Dispatch the job with service time x to queue i minimizing the mean additional costs:

$$a_T(u,x,i) = u + x + v_T(u+x,i) - v_T(u,i)$$

$$a_P(u,x,i) = v_P(u+x,i) - v_P(u,i)$$

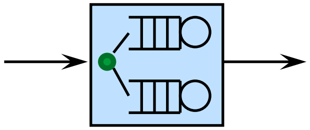
- For INSTANTOFF servers:

  Dispatch the job with service time x to queue i minimizing the mean additional costs:

$$a_T(u,x,i) = u + x + d_i \cdot 1(u=0) + v_T(u+x+d_i \cdot 1(u=0),i) - v_T(u,i)$$

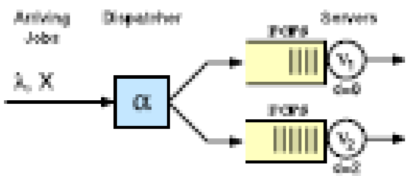
$$a_P(u,x,i) = v_P(u+x+d_i \cdot 1(u=0),i) - v_P(u,i)$$

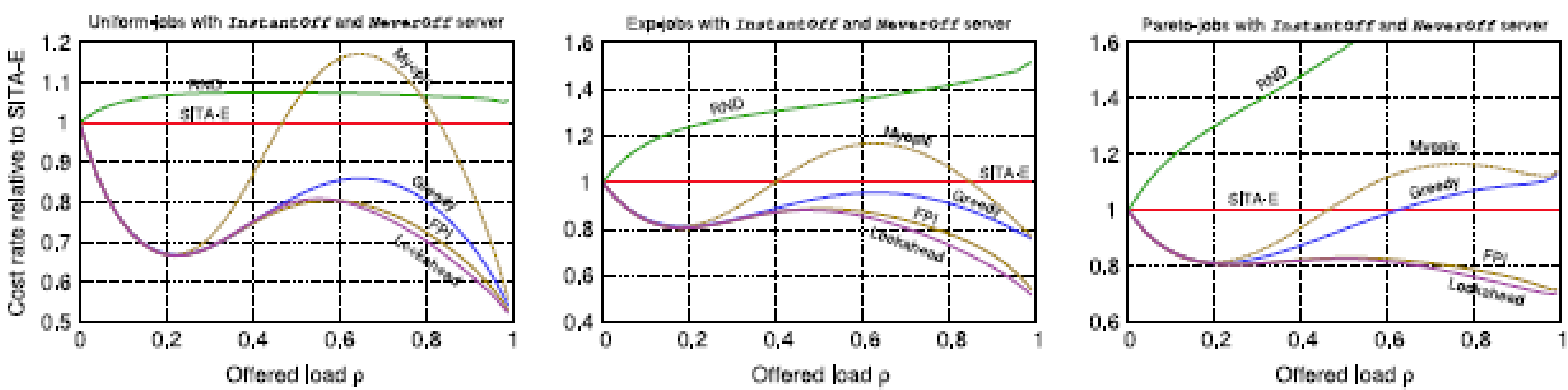
Numerical results



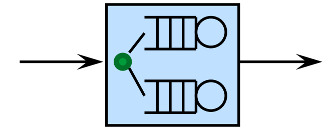
Hyytiä et al. (2014a)

Table 2
Two-server system.

	Server 1:	$\nu_1 = 1$	$e_1 = 1$		`NeverOff`
	Server 2:	$\nu_2 = 1$	$e_2 = 1$	$d_2 = 2$	`InstantOff`



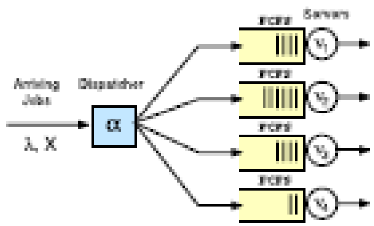
(a) Uniform.   (b) Exponential.   (c) Truncated Pareto.

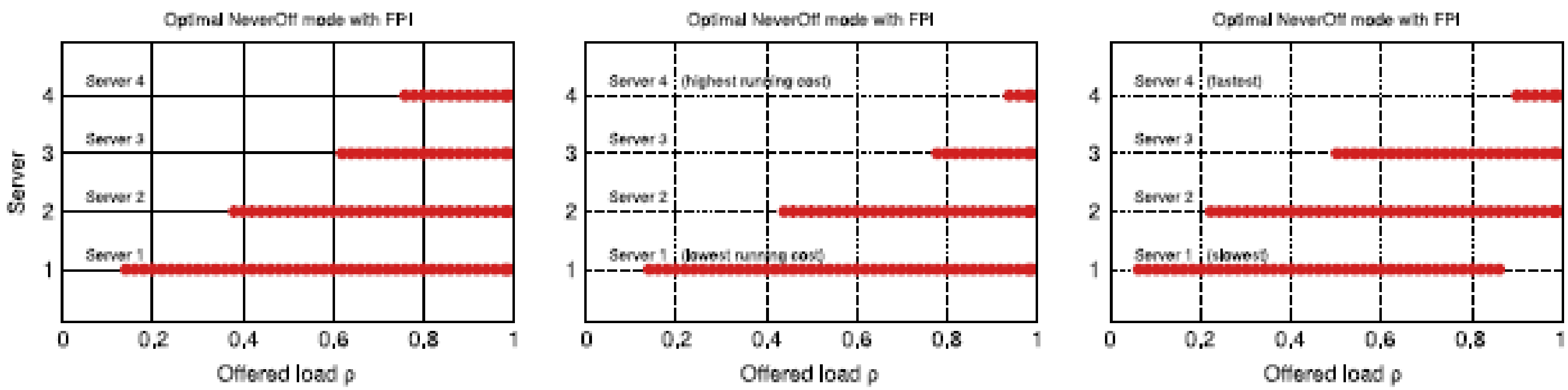


Numerical results

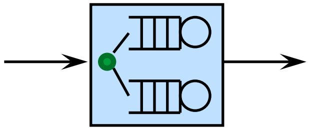
Hyytiä et al. (2014a)

Table 3
Four-server systems.

	Parameter		(a) Identical	(b) Linear e	(c) Squared e
	Service rates	$\nu_1, \ldots, \nu_4$:	1, 1, 1, 1	1, 1, 1, 1	1, 2, 3, 4
	Running costs	$e_1, \ldots, e_4$:	1, 1, 1, 1	1, 2, 3, 4	1, 2, 9, 16
	Switching delay	$d_1, \ldots, d_4$:	1, 1, 1, 1	1, 1, 1, 1	1, 1, 1, 1



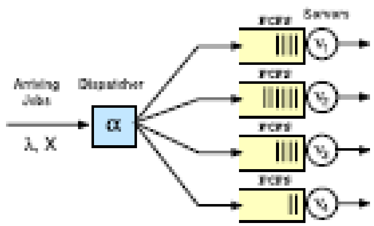
(a) Identical servers. (b) Linear running cost. (c) Squared power consumption.



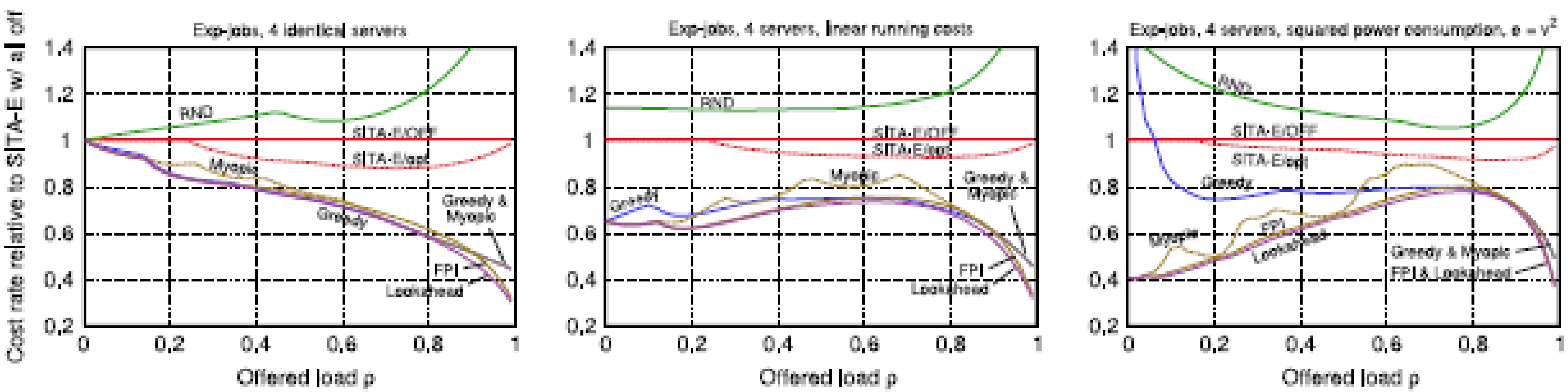
Numerical results

Hyytiä et al. (2014a)

Table 3
Four-server systems.



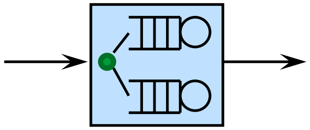
Parameter		(a) Identical	(b) Linear e	(c) Squared e
Service rates	$\nu_1, \ldots, \nu_4$:	1, 1, 1, 1	1, 1, 1, 1	1, 2, 3, 4
Running costs	$e_1, \ldots, e_4$:	1, 1, 1, 1	1, 2, 3, 4	1, 2, 9, 16
Switching delay	$d_1, \ldots, d_4$:	1, 1, 1, 1	1, 1, 1, 1	1, 1, 1, 1



(a) Identical servers.

(b) Linear running cost.

(c) Squared power consumption.



Other queueing disciplines

Hyytiä et al. (2014b)

- LIFO in the M/G/. setting with setup delays
- PS in the M/D/. setting with setup delays

- But it is another story …

References

- Harchol-Balter, Crovella & Murta (1999)
  On Choosing a Task Assignment Policy for a Distributed Server System, *JPDC*
- Wierman, Andrew & Tang (2009)
  Power-aware speed scaling in processor sharing systems, in *IEEE INFOCOM*
- Gandhi, Harchol-Balter & Adan (2010a)
  Server farms with setup costs, *PEVA*
- Gandhi, Gupta, Harchol-Balter & Kozuch (2010b)
  Optimality analysis of energy-performance trade-off for server farm management, *PEVA*
- Hyytiä, Penttinen & Aalto (2012)
  Size- and state-aware dispatching problem with queue-specific job sizes, *EJOR*
- Maccio & Down (2013)
  On optimal policies for energy-aware servers, in *MASCOTS*
- Hyytiä, Righter & Aalto (2014a)
  Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure, to appear in *PEVA*
- Hyytiä, Righter & Aalto (2014b)
  Energy-aware job assignment in server farms with setup delays under LCFS and PS, accepted to *ITC*
- Gebrehiwot, Lassila & Aalto (2014)
  Energy-aware queueing models and controls for server farms, ongoing work

The End