Frequency Scaling in Multilevel Queues

IFIP Performance 2020

Maryam Elahi[1,3] Andrea Marin[2] Sabina Rossi[2] Carey Williamson[3]

[1] Mount Royal University, Canada
[2] Università Ca'Foscari Venezia, Italy
[3] University of Calgary, Canada

Talk overview

Introduction and Contribution

The queueing model and its solution

Case study

Conclusion

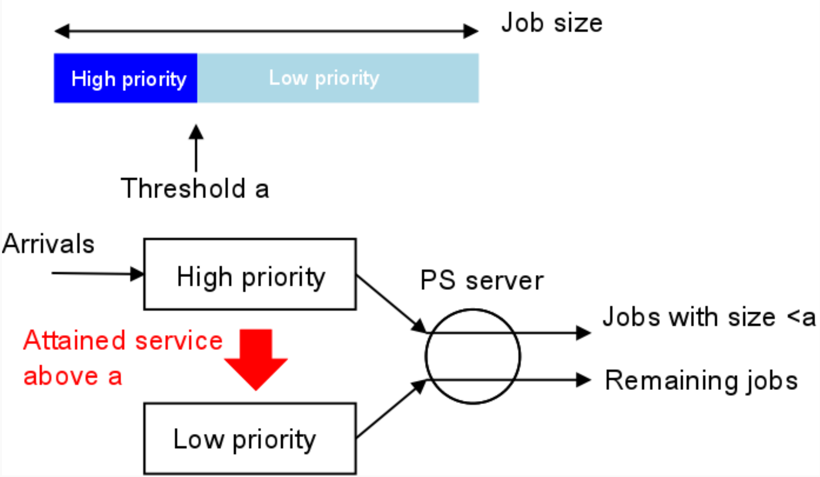
Introduction and Contribution

Size based scheduling

- **Goal:** serve first the smallest jobs to improve the expected response time
- **Assumptions:**
  - Do we know the job size at the arrival epoch? (**S**hortest **R**emaining **P**rocessing **T**ime)
  - Do we know the distribution of the job size? (Gittin's policy)
  - Is the job size distribution heavy tailed? (Least Attained Service (LAS), Multilevel queues)

How do multilevel queues work?

Example: two levels with **P**rocessor **S**haring (PS) server



Frequency scaling

- **Goal:** Energy saving
- **Main idea:** When there are few jobs in the system, we can reduce the processor speed
  - We increase the expected response time w.r.t. constant speed only for the *lucky* jobs
  - The service time is directly proportional to the service speed f
  - The power consumption depends on the service speed as f^{α}, were $2 \leq \alpha \leq 3$
- **Linear frequency scaling:** The server speed is proportional to the number of jobs in the system

Related work (short list)

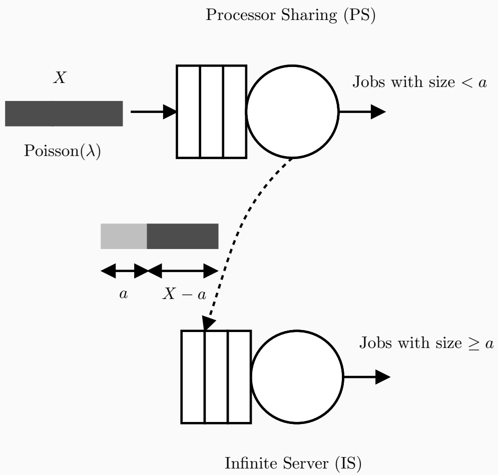
- **Policies independent of the received service**
  - (George, Harrison): FCFS queues and frequency scaling
  - (Wierman et al.): M/G/1/PS queues with frequency scaling
- **Policies with known job size**
  - (Bansal et al.; Andrew et al.): Worst case analysis of SRPT with frequency scaling
  - (Andrew at al.; Elahi, Williamson): Unfairness in SRPT with frequency scaling
  - (Lassila; Aalto): LAS with sleeping servers

Contribution

- We study a two-level queue with PS discipline and linear speed scaling on the low-priority jobs (PS+IS)
- We give a numerical solution of the queueing system and validate it with discrete event simulation
- We study the behaviour of the model with job size distributions obtained by monitoring TCP flows of a data centre

The queueing model and its solution

Graphical representation



Note: The IS system works only when the PS is idle

Job size distribution

- We use **Generalized Hyperexponential** (GH) distributions

$$f_X(x) = \sum_{k=1}^{K} p_k \mu_k e^{-p_k \mu_k}$$

where $\sum_{k=1}^{K} p_k = 1$, $p_k \in \mathbb{R}$, $\mu_k \in \mathbb{R}^+$, $f_X(x) > 0$ for all $x \in \mathbb{R}^+$

- GH distributions are dense in the domain of the distributions
  - They can approximate any distribution arbitrary well

Analysis of the queue: sketch

- We can see the system consisting of two queues:
  - **High priority** one which is M/G/1/PS whose job sizes are truncated at a
  - **Low priority** one which works during the idle periods of the PS which is a $\mathrm{M}^B/\mathrm{G}/\infty$ queue
    - The arrival process is Poisson with intensity λ
    - The batch size is the number of jobs that crossed the threshold during a busy period of the PS level
- The **generating function** of the batch size distribution has not an explicit form but has a characteristic equation (Kleinrock)
- The solution of the IS queue requires the **distribution** of the batch size

Computation of the batch size distribution

- We invert the generating function with the **Lattice-Poisson algorithm** by Abate and Whitt
- The evaluation of the generating function is obtained with a **fixed point algorithm** whose convergence is proved by resorting to Banach's contraction mapping fixed point theorem
- The accuracy of the numerical procedure **is validated** in low and heavy-load by comparing the first two moments of the distribution (which can be computed explicitly for GH distributions from the characteristic equation) with those obtained by the numerical inversion

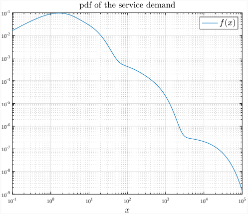
Computation of the power consumption

- We resort to the literature for the power consumed by the PS queue
- We provide a numerical solution for the IS queue and integer values of the exponent α
- The power consumption is derived from the second ($\alpha = 2$) and third ($\alpha = 3$) moments of the occupancy distribution in the IS queue

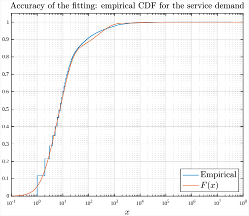
Case study

Dataset

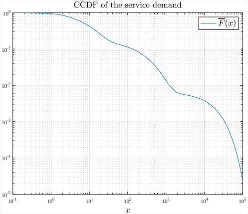
- **TCP flows** monitored at the data centre of the Università Ca' Foscari Venezia in November 2019
- **Fitting with PH-Fit** into an acyclic phase-type distribution
- **Transformation** of the acyclic phase-type distribution into a GH



(a) Probability density function in log-log scale.



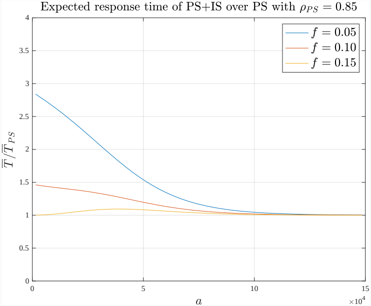
(b) Empirical and analytical cumulative density function in log-linear scale.



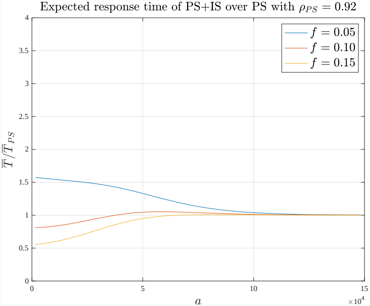
(c) Complementary cumulative density function in log-log scale.

PS+IS vs. PS: Comparison of the expected response time

- PS queue has speed 1 and IS has speed $f < 1$

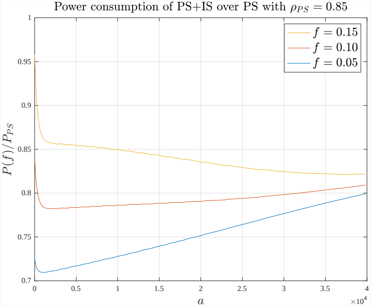


(a) Expected response time: $\rho_{PS} = 0.85$.

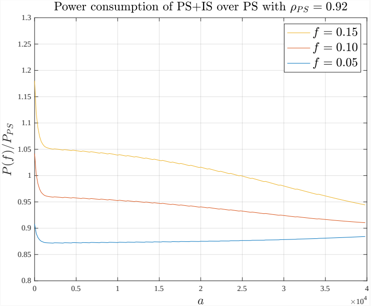


(b) Expected response time: $\rho_{PS} = 0.92$.

PS+IS vs. PS: Comparison of the power consumption

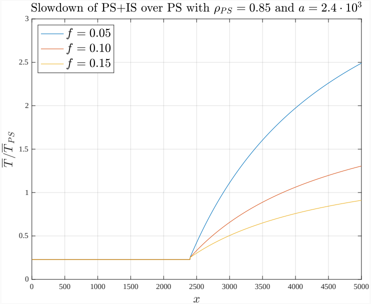


(a) Power consumption: $\rho_{PS} = 0.85$ when $0 \leq a \leq 4 \cdot 10^4$.

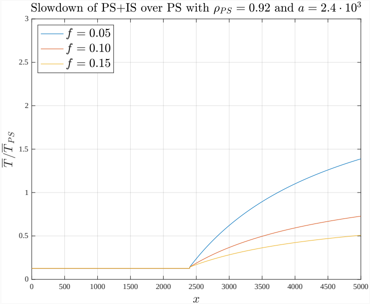


(b) Power consumption: $\rho_{PS} = 0.92$ when $0 \leq a \leq 4 \cdot 10^4$.

PS+IS vs. PS: Slowdown

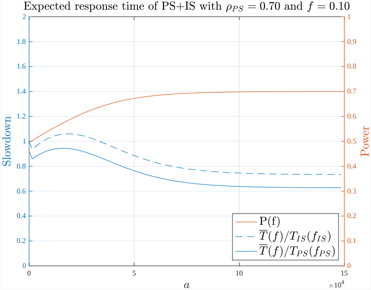


(a) Slowdown of PS+IS conditioned to the job size x with $\rho_{PS} = 0.85$.

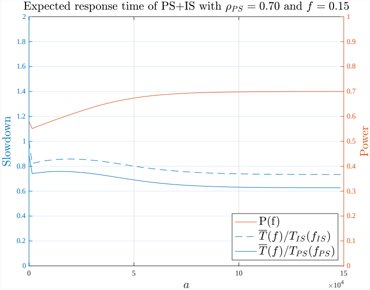


(b) Slowdown of PS+IS conditioned to the job size x with $\rho_{PS} = 0.92$.

PS+IS vs. PS: Comparison of the expected response times with same power consumption



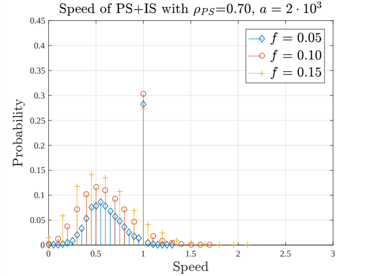
(a) Comparison of the expected response time with $\rho_{PS} = 0.70$ and $f = 0.10$.



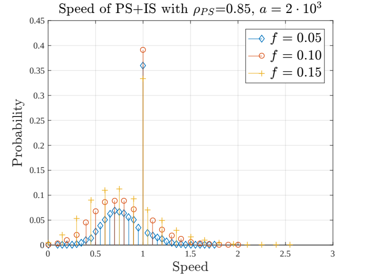
(b) Comparison of the expected response time with $\rho_{PS} = 0.70$ and $f = 0.15$.

Simulation

- Simulation has been used to cross validate the numerical results
- Simulation allows the investigation of other characteristics of the system such as the distribution of the system speed



(a) System speed: $\rho_{PS} = 0.70$.



(b) System speed: $\rho_{PS} = 0.85$.

Conclusion

Conclusion

- We have introduced a two-level queueing system (PS+IS) with linear speed scaling for the low-priority level
- A numerical solution procedure has been proposed and its accuracy has been validated with discrete event simulation
- Experiments on real-world job size distributions have been carried out
- We showed that the model-driven configuration of the PS+IS system is crucial for obtaining the benefits of the speed scaling without compromising the slowdown of the system too much