Frequency scaling in multilevel queues

Maryam Elahi
Mount Royal University,
Calgary, AB, Canada
melahi@mtroyal.ca

Andrea Marin
Università Ca' Foscari
Venezia, Venice, Italy
marin@unive.it

Sabina Rossi
Università Ca' Foscari
Venezia, Venice, Italy
sabina.rossi@unive.it

Carey Williamson
University of Calgary, Calgary,
AB, Canada
carey@cpsc.ucalgary.ca

ABSTRACT

In this paper, we study a variant of PS+PS multilevel scheduling, which we call the PS+IS queue. Specifically, we use Processor Sharing (PS) at both queues, but with linear frequency scaling on the second queue, so that the latter behaves like an Infinite Server (IS) queue. The goals of the system are low response times for small jobs in the first queue, and reduced power consumption for large jobs in the second queue. The novelty of our model includes the frequency scaling at the second queue, and the batch arrival process at the second queue induced by the busy period structure of the first queue which has strictly higher priority. We derive a numerical solution for the PS+IS queueing system in steady-state, and then study its properties under workloads obtained from fitting of TCP flow traces. The simulation results confirm the efficacy of the PS+IS model. This article is an introduction to our longer paper in [5].

Categories and Subject Descriptors

A.0 [**General Literature**]: General—*Performance*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Queueing Theory*; F.2 [**Theory of Computation**]: Analysis of algorithms and problem complexity—*Scheduling algorithms*

Keywords

Frequency scaling, Size-based Scheduling, Energy consumption, Queueing systems

1. BACKGROUND AND MOTIVATION

In scheduling, exploiting the knowledge of job sizes can significantly improve system performance. For example, it is well known that the Shortest Remaining Processing Time (SRPT) scheduling is optimal in terms of minimization of the expected response time in single server queues [12] and has asymptotically optimal mean response time in heavy-traffic regime for multiserver queues [7]. Dispatching policies may also take advantage of the knowledge of the job sizes [6] or by their statistical prediction [3]. However, in many cases, the scheduler is unaware of the job size, and knows only the amount of service received by a job up to a certain epoch.

In this paper, we study a preempt-resume single server with variable service rates that is not aware of the job sizes at their arrival instants. When the service rate is constant in these cases, multi-level scheduling is a well-known technique to reduce the mean response time in systems with highly variable job sizes. It can provide much lower response time than the Processor Sharing (PS) discipline, especially for heavy-tailed service time distributions [1, 2, 8, 10].

Unlike SRPT discipline, multilevel scheduling does not need to know job sizes in advance. Furthermore, it is much easier to implement than Least Attained Service (LAS) that is known to provide a significant reduction of the expected response time in practically relevant scenarios [11]. LAS may be seen as a multilevel queue with PS scheduling and infinitely many levels.

One classic example of multilevel scheduling is the PS+PS queue, which has two levels. All jobs arrive initially at the first queue, and are served at high-priority according to the PS discipline. These jobs are all served concurrently, and "small" jobs complete quickly. However, if a job's service requirements exceed a specified threshold a, then that job is deemed to be a "large" job, and is demoted to the second queue. The latter queue runs at lower priority, and also uses the PS discipline to serve all of its jobs. Note, however, that jobs in the second queue are served only when there are no high-priority jobs in the system (i.e., in the first queue). This type of queue has been first investigated by Kleinrock in [9] but several further results have been derived more recently. Among these, in [1, 2] the authors show that if the distribution of the service demand has decreasing hazard rate, then the PS+PS discipline has at most the same expected response time as the simple PS queue for any choice of the threshold.

2. CONTRIBUTIONS OF THIS PAPER

In this paper, we study a variant of PS+PS multilevel scheduling in which frequency scaling is applied on the second (low-priority) queue. Specifically, we consider a linear frequency scaling policy, in which the service rate of the second queue increases linearly based on the number of jobs being served. That is, the second queue behaves like a classic Infinite Server (IS) system, with PS scheduling, and a per-job service rate f, which is (much) smaller than the unitary aggregate service rate of the first queue. Furthermore, the second queue is served only when the first queue is empty.

We call our model the PS+IS queue. Our design goals include low response times for the small jobs handled by the first queue, and reduced power consumption for the large jobs handled by the second queue. The novelty of our model arises from both the frequency scaling of the second queue, and the batch arrival process induced at the second queue by the busy period structure of the first queue. Furthermore, the threshold parameter a influences the performance tradeoffs observed between small and large jobs.

The contributions of our paper are two-fold. First, we provide a numerical solution of the stationary queueing system for PS+IS multilevel scheduling with frequency scaling. We assume that the service demand has a generalized hyperexponential (GH) distribution. These distributions are very expressive and can approximate any distribution with arbitrary accuracy [4]. Second, we analyze the properties of PS+IS under empirical workloads. These are obtained from a fitting procedure from real-world datasets of TCP flow sizes.

The results of our numerical analysis confirm the efficacy of the PS+IS model. When the threshold a is small, there is a pronounced response time advantage for small jobs, since they are minimally impacted by the short residency time of large jobs in the first queue. However, the occupancy of the second queue grows, since more of the jobs are deemed to be large. This in turn increases the service rate and power consumption of the second queue. For larger values of the threshold a, the power consumption of the second queue is reduced, since there are fewer large jobs. However, the smallest jobs have higher response times in the first queue because of increased competition from medium-to-large jobs. Furthermore, the response times for large jobs are high, since the second queue is not served as often.

Simulation results are presented for metrics that are not easily obtainable via numerical analysis, including the distribution of occupancy for the low-priority queue, and the ratio of time that the system is busy, i.e., system utilization. The results indicate that when the PS+IS system is configured with the right threshold and frequency scaling factor, it maintains good average response time while having a high utilization that is close to saturation. This is due to the fact that the system dynamically scales to run fast enough to process the ambient workload, but no faster.

3. CONCLUSION

Multilevel systems, including this one, have practical importance in the cases of heavy tailed service demand distributions, i.e., when a job whose attained service is larger than a threshold a will very likely require a large residual service demand. In these contexts, it is often the case that the large jobs can be seen as *background* work and their service requirements are delay tolerant. In contrast, small jobs usually correspond to interactive activities that have stricter requirements. Intuitively, PS+IS systems work on large jobs during what would be the idle periods of simple PS systems with the same workload. In this way, the processor keeps working at low speed instead of staying idle, and we obtain benefits in terms of power consumption. Indeed, the idle periods of PS+IS are very small as shown by the simulations.

The outcomes of our experiments show that model-driven configuration of the PS+IS queue is crucial for its success in realistic scenarios. The model is useful for determining proper settings, and avoiding poor settings that can compromise system performance. When properly configured, the PS+IS queue can provide advantages both in response time and in power consumption compared to either a PS queue or an IS queue. Future works include the optimization of the system on three parameters: the speed at the high-priority PS queue, the base speed at the low-priority IS queue, and the threshold a.

4. REFERENCES

[1] S. Aalto and U. Ayesta. Mean delay analysis of multi level processor sharing disciplines. In *Proc. IEEE of the 25th Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM)*, 2006.

[2] S. Aalto, U. Ayesta, and E. Nyberg-Oksanen. Two-level processor-sharing scheduling disciplines: Mean delay analysis. *ACM SIGMETRICS Perf. Eval. Review*, 32(1):97–105, 2004. Proc. of ACM SIGMETRICS/Performance.

[3] E. Bachmat, J. Doncel, and H. Sarfati. Performance and stability analysis of the task assignment based on guessing size routing policy. In *Proc. of Int. Symp. on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 1–13, 2019.

[4] R. Botta, C. Harris, and W. Marchal. Characterizations of generalized hyperexponential disribution functions. *Communications in Statistics. Stochastic Models*, 3(1):115–148, 1987.

[5] M. Elahi, A. Marin, S. Rossi, and C. Williamson. Frequency scaling in multilevel queues. *Performance Evaluation*, 143:102140, 2020.

[6] I. Grosof, A. Scully, and M. Harchol-Balter. Load balancing guardrails: Keeping your heavy traffic on the road to low response times. *Proc. of the ACM SIGMETRICS on Measurement and Analysis of Computing Systems (POMACS)*, 3(2):42:1–42:31, 2019.

[7] I. Grosof, Z. Scully, and M. Harchol-Balter. SRPT for Multiserver Systems. *Performance Evaluation*, 127-128:154–175, 2018. Proc. of IFIP Performance.

[8] L. Guo and I. Matta. Scheduling flows with unknown sizes: Approximate analysis. *ACM SIGMETRICS Perf. Eval. Review*, 30(1):276–277, 2002. Proc. of ACM SIGMETRICS.

[9] L. Kleinrock. *Queueing Systems*, volume II: Computer Applications. Wiley Interscience, 1976.

[10] A. Marin, S. Rossi, M. Sottana, and C. Zen. Theoretical and experimental evaluation of the two-level processor sharing discipline for TCP flows. In *Proc. of Int. Symp. on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 94–106, 2019.

[11] I. Rai, G. Urvoy-Keller, M. Vernon, and E. Biersack. Performance analysis of LAS-based scheduling disciplines in a packet switched network. In *Proc. of the ACM SIGMETRICS Int. Conf. on Measurement and modelling of computer systems*, pages 106–117, 2004.

[12] L. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:678–690, 1968.