Xiaohua Lu

Energy-aware Performance Analysis of Queueing Systems

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 01.08.2013

Thesis supervisor:

Prof. Samuli Aalto

Thesis instructor:

D.Sc. (Tech.) Pasi Lassila

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Xiaohua Lu

Title: Energy-aware Performance Analysis of Queueing Systems

Date: 01.08.2013 Language: English Number of pages:6+57

Department of Communications and Networking

Professorship: Teletraffic theory Code: S-38

Supervisor: Prof. Samuli Aalto

Instructor: D.Sc. (Tech.) Pasi Lassila

ICT systems, especially data centers, consume a significant amount of energy in our daily life. With the rapidly increasing number and size of data centers, energy management is becoming essential. Thus, it is beneficial if the used energy in data centers can be utilized more efficiently.

In this thesis, we analyze the energy-aware performance of queueing systems from the traffic point of view. The focus will be on using queueing theory to model and analyze a single processor in data centers. In data centers, the energy consumed by a processor depends on the processing speed. With higher speed, more energy is consumed, while with lower speed, the performance will be decreased. Thus, we consider the trade-off between the performance and energy consumption of processors. Based on this, we introduce a speed scaling method, which adjusts the processing speed of processors according to the traffic load of the queueing system. We mainly analyze and compare three optimized speed scaling methods, which are static, gated and linear speed scaling. In the gated and linear schemes, there is a switching delay when the processor is switched from the idle state to the busy state.

The results demonstrate that the switching delay has a great impact on the optimized trade-off. In our scenario, without switching delay, gated and linear schemes have the same performance, and they are better than the static scheme. With switching delay, however, the linear scheme is always better than the gated scheme. With a long switching delay, even the static scheme can be better. In practice, the trade-off of our model is highly affected by the parameters in the model.

Keywords: Speed Scaling, Energy-aware, Switching Delay, Queueing Theory

Preface

This thesis has been carried out in the Department of Communications and Networking, in Aalto University.

I would like to express my sincere gratitude to my supervisor, Professor Samuli Aalto, who has given me this precious opportunity to study such a valuable topic and has provided me much helpful guidance and many valuable comments for writing this thesis. His kindness and patience has helped me a lot for completing this thesis in a nice way.

Also, I would like to express my grateful thanks to my instructor, Dr. Pasi Lassila, who has provided me many useful comments about the content of my thesis. In addition, he has given me many good instructions on the usage of tools for my thesis.

I would like to thanks all my friends who helped and supported me during these two years in Finland. They have brought me lots happiness both in studying and living here.

Last but not least, I would like to express my deepest gratitude to my family. Thanks to my parents for always being there and supporting me. Thanks to my sister and brother for accompanying me despite the distance.

Otaniemi, 01.08.2013

Xiaohua Lu

Contents

List of Symbols

z	average energy cost per unit time
X	the number of jobs in the queueing system
$P(S)$	the power function when the processor runs at random speed S
β	parameter that relates the energy cost with delay
S	state space of a Markov process
I	parameter space of a Markov process
q_{ij}	state transition rate
q_i	rate of the process leaves state i
T_i	total time a process stays in state i
p_{ij}	probability that a process jumps from state i to state j
$\pi_i(t)$	time-dependent probability
π_i	limiting distribution
T_n	the n_{th} renewal sequence
τ_n	sum sequence of T_n
$N(t)$	renewal process
$Y_n(t)$	the n_{th} cycle of process $X(t)$
λ	arrival rate
μ	service rate
ρ	the actual traffic load
α_i	arrival time of customer i
A_i	inter-arrival time
I_n	the n_{th} idle period
B_n	the n_{th} busy period
C_n	the n_{th} busy cycle
N_n	the number of served customers in the n_{th} cycle
W_i	the waiting time of customer i
Y_i^w	the number of waiting customers that the arriving customer i sees
$R(\alpha_i-)$	the remaining service time seen by the arriving customer i
$X(t)$	the number of customers in the system at time t
$Z(t)$	indicates whether the system is in the switching delay state or not
δ	the rate that the state switches to switching delay state
s	the service speed
Y	the size of the job
m	the mean size of the job
r	the traffic load
γ	the energy-aware load

1 Introduction

1.1 Background

Energy consumption has increasingly become a vital issue. Whereas energy is essential to provide people with a better life, for example, energy consuming devices, such as computers and mobile phones help people to live more comfortably. Renewable energy, on the other hand, is finite and valuable; therefore, consuming a large amount of energy is disadvantageous to sustainable development and will aggravate the energy crisis. In addition, the emission of greenhouse gases, such as CO_2 in the process of energy consumption has worsened the global climate change [16]. Therefore, in order to maintain sustainable development and protect the environment, saving energy is essential.

A significant part of our daily consumed energy is comprised of ICT (Information and Communication Technology) systems. From the statistics given in [31], one can see that the relative proportion of electricity consumption increased from 8% in 2008 to 14% in 2020. However, a large part of the consumed energy is wasted because of inefficient usage [23]. For example, if the devices are in the unused state or the traffic load is low, retaining these devices in the normal working state may not be the best option. In this case, it may be appropriate if the devices are switched to sleeping mode. This situation leads to exploring energy efficiency methods to cope with the increasing challenge of energy efficiency problems in ICT systems.

The authors of [2] specified that data centers constitute a large part of the ICT systems. According to [1], a data center is a dedicated area that houses the servers and storages equipment of the communication systems. As has been noted in [2], data centers consume a large amount of energy in ICT systems. The energy consumed by a large industrial data center can be more than that in a small town [3]. In this fast developing information age, there is a large amount of new information emerging everyday, which leads to a tremendous increase of data centers and the consumption of energy. Consequently, it will be an important step if the available energy in ICT systems can be used efficiently, especially in data centers.

In this thesis, analyzing data centers from the traffic point of view will be the focus. Furthermore, the devices in data centers will be modeled as queueing systems. To be more precise, the processor of the device is considered as a server and the arriving jobs will be considered as customers.

1.2 Research problem

As mentioned in the preceding paragraph, queueing theory is utilized to model the devices with speed adjustable processors in data centers. Figure 1 illustrates the queueing model of a processor in a data center. Arriving jobs enter the data center. If there is already a job being served in the processor, then arriving jobs will wait in the queue. If not, they will enter the processor and be served. The main focus of this thesis is to analyze and compare different methods for increasing the energy efficiency of the processor by utilizing such kinds of queueing models.

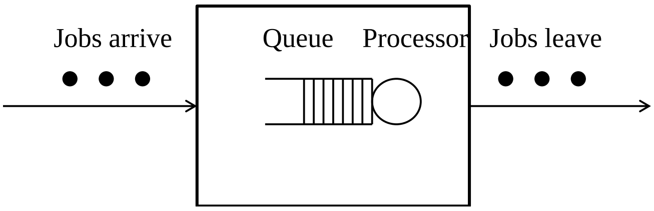


Figure 1: Queueing model for a processor (single server).

In data centers, the amount of energy consumed by a processor depends on the processing speed. In general, the higher the processing speed, the more energy is consumed. However, with lower processing speed, the performance of the queueing system will be decreased. Consequently, it is vital to focus on the trade-off between the performance and energy consumption of the processors. This method of adjusting the processing speed is called speed scaling.

In this thesis, the processing speed is assumed to take any positive value; therefore, processors can run at any speed whenever there is traffic in the system. One approach is to choose an optimal speed according to the average traffic load of the queueing system, which will balance the performance and energy consumption of the processors. This method always runs the processor at the optimal speed, regardless of the dynamic traffic load of the processor, and is called static speed scaling.

However, if there is no job to process in a processor, energy will be wasted when maintaining the processor running at the same static speed. Thus, it may be better if the processor is converted to the idle state, where the processing speed is zero, when there is no traffic. When there are new arriving jobs, the processor will be converted to the busy state from the idle state. But this process may result in a delay for processing the jobs, which is called switching delay. It also has an impact on the performance of the queueing system. It will make the total processing time of jobs longer. At the same time, this switching process will also consume energy. This second method is called gated static speed scaling.

Although gated static speed scaling is more energy efficient compared with the static one, as can be observed, it runs at a fixed speed whenever the processor is not empty. However, for a processor with fewer jobs, it may be more energy efficient if the processing speed can be dynamically adjusted according to the instantaneous traffic load. This will be the third method to be analyzed in this thesis, called dynamic speed scaling. For dynamic speed scaling, if the system is empty, then the processor is converted to the idle state. If not, the processing speed will be adjusted accordingly. Similarly, there may also be a switching delay when the processor converts from the idle state to the busy state.

There are different ways to adjust the speed dynamically. Since the number of jobs in the queue affects the traffic load and the performance of the queue, it is natural to adjust the processing speed according to the number of jobs in the queueing system. One simple method is to run the processor at a speed that is linear to the number of jobs in the system. This type of dynamic speed scaling will be studied in this thesis.

The main objectives of this thesis are as follows. First, switching delay is modeled and we will analyze its impact on the performance of different queueing systems. Then, the impact of speed scaling is analyzed without considering switching delay. Finally, speed scaling is combined with switching delay and the focus will be on their influence on the overall performance of the queueing systems. As the performance of different queueing models may be different, two different queueing models will be studied: the M/G/1 queue, which is related to static and gated static speed scaling, and the M/G/∞, which is related to linear speed scaling. Additionally, the way that jobs are served will also impact the performance of the queue, and this is determined by the scheduling policies. A typical scheduling policy in ICT systems (e.g., web servers, routers) is PS [39], and FIFO is utilized in some operating systems [40]. Therefore, these two scheduling policies will also be studied.

1.3 Outline

The thesis is organized as follows. In Section 2, we give a literature review about the energy-awareness in ICT systems including data centers. Several aspects of speed scaling in energy efficiency are reviewed in this section. The related theoretical background is introduced in Section 3. The main focus in this section is on Markov processes and regenerative processes. In addition, we introduce several related queueing models. Section 4 presents and compares the performance (in terms of the mean queue length) of different queueing models for the cases with and without switching delay. Then in Section 5, we conduct the energy-aware performance analysis. There are still two situations: without switching delay and with switching delay. By applying three different speed scaling methods, we analyze and compare the optimal average energy cost per time unit in several queueing systems. The numerical results are shown in Section 6, where we analyze the impact of different factors on the average energy cost. Finally, in Section 7, we make a brief summary about the whole thesis.

2 Energy-aware performance analysis of data centers

As mentioned in the introduction part, energy efficiency in data centers has gained much attention in recent years. One way to achieve it is by utilizing speed scaling methods from the traffic point of view. A large number of papers related to energy efficiency have emerged, and here some of them will be reviewed from the perspective of the issues discussed in the previous section.

2.1 Optimization formulations

There are mainly two types of analytic approaches related to the performance of queueing systems in the prior work: the stochastic analysis and the worst-case analysis. Stochastic analysis focuses on the average performance of a stochastic process, while worst-case analysis usually evaluates the performance by utilizing the *competitive ratio*, which is the ratio between the performance of a schedule or algorithm and the optimal performance [32]. Examples of stochastic analysis are in [39] and [19]. Paper [39] is about the optimization of energy cost in the M/G/1 queue with the PS scheduling policy. Stochastic analysis for queueing models with multiple servers is considered in [19]. This paper shows how to assign the available power to multiple servers in a server farm to get the optimal performance with PS scheduling policy. The worst-case analysis is presented in [41], [14], [4], [33], [27] and [29]. Competitive ratio of different models is studied in those papers.

Firstly, we give an overview of classification of the papers related to energy efficiency by utilizing speed scaling.

Speed scaling was first studied by Yao et al. in [41], and in the following years, a large number of papers related to speed scaling emerged. Among those papers, speed scaling is applied with different methods to achieve various goals, and they can be categorized into two classes according to the objectives. In the first class (e.g., [41], [14]), the objective is to minimize energy consumption of the system. There are deadlines for the jobs, and the target is to finish the jobs by their deadlines while trying to consume the least amount of energy. For the second class (e.g., [33]), optimizing the performance of the system is the main goal. The jobs of this class do not have deadlines, and the main focus is on the performance of the system, which can be measured by the mean response time of the jobs. In this class, there is usually a limited energy budget, so the target turns out to be minimizing the mean response time with the given energy budget.

These two kinds of goals are commonly used in many papers related to speed scaling. However, in actual ICT systems, there may be no deadlines for the jobs, and the energy budget is not necessarily fixed. So, it is natural to combine these two goals and try to find a trade-off between them (e.g., [20], [39] and [10]). In recent years, numerous such papers have appeared. These papers study the methods of finding a trade-off between the two goals mentioned above, to optimize the overall performance as well as the energy consumption. One common optimization formulation that is utilized is: $z = E[X] + E[P(S)]/\beta$, where z is the cost per unit time, X is the number of jobs (customers) in the system, $P(S)$ is the power function when the processor runs at a random speed S, and β is a parameter that relates the energy cost with delay. This optimization formulation is the weighted sum of the mean response time and mean consumed energy, and it has drawn much attention in many papers (e.g. [39], [6] and [30]).

Although this optimization formulation has drawn much attention in recent years, other models that measure performance metrics in a different way can also be used. For example, the energy model that is used in [17] and [21] is the product of the mean power consumption and the mean response time. For our analysis in this thesis, however, the weighted sum of the mean response time and the mean consumed energy model will be the focus.

2.2 Energy models for processors

Above is a review about the optimization formulations in speed scaling research areas, while in this subsection, the energy consumption models for processors will be examined.

First, the focus will be on a detailed review of the power function in the optimization formulation of [39], [6] and [30]. The traditional power function in the above model is: $P(s) = s^\alpha$, where $\alpha > 1$. This power function is additionally used in papers [41], [4] and [7]. However, in [10], the authors analyze speed scaling with an arbitrary power function. The authors explain why an arbitrary power function instead of the traditional one should be used. Generally, there are three motivations: (i) it might be beneficial for a processor to run at different speed ranges instead of only one speed range, which requires different circuitry as well as different power functions; (ii) there seems to be no specific reason to assume that the power function should be the form of s^α; (iii) from the perspective of data centers operators, they are more concerned about the cost of energy instead of the consumption of the energy.

For the traditional power function, there are some typical values of the parameter α for the purpose of numerical analysis. In [39], [41] and [6], the value of α is 2. Paper [6] points out that α is usually between 1 and 3, and for simple numerical analysis, α is usually taken as 2. However, in [10] the value of α is considered as 3. This is because of the cube-root rule, which states that the consumed power is proportional to the cube of the speed [13].

Apart from the expression of the power function, it is also important to consider the possible values of the speed of the processor. Many papers that are related to speed scaling assume that s can take any real value greater than or equal to zero ([39], [33]). In some other papers, the processing speed varies in a finite interval ([7], [27], [9]). In reality, there is only a limited discrete set of speeds available.

The papers mentioned above (except [10]) only consider an ideal situation, where the dynamic power is the main focus and the static power (leakage power) is ignored. Dynamic power is the power consumed by CMOS when switching the transistors while static power is the power used when there is no switching of the transistors ([10], [24]). In the design of the early days, dynamic power consumption was far more than static power consumption. In recent years, however, static power in the processors produced by some manufacturers have become non-negligible compared with dynamic power ([10], [24]). Thus it is more precise if the power function is modified as: $P(s) = s^{\alpha} + c$, where $\alpha > 1$ and c is a constant that represents the static power loss [10]. However, there are still very few papers that study the speed scaling by taking the static power into account.

2.3 Queueing theoretic approaches

Since this thesis will model processors in data centers using queueing models, we also give an overview of queueing issues related to energy efficiency.

Papers [20] and [39] both analyzed speed scaling methods from the aspect of queueing system. Furthermore, they utilized the optimization formulation we mentioned in Section 2.1 to study energy efficiency. In paper [20], the authors derived the optimal solution of energy-performance trade-off for dynamic speed scaling. Then, the authors of [39] analyzed optimal dynamic speed scaling based on [20]. In addition, they analyzed the optimal service speed of static and gated static speed scaling to balance the mean response time and the mean energy usage.

As mentioned earlier, one important issue for the performance in queueing models is the scheduling policies; therefore, it is important to decide which scheduling policy is used. In [39], the scheduling policy is PS, while SRPT and HDF (Highest Density First) scheduling algorithms are studied in [10]. Paper [11] defines HDF as a policy where the server always serves the jobs with the highest density, and the highest density implies that the ratio of the weight of a job to its size is the biggest. HDF is the same as SPT if the weights are the same. In [6], the optimality of speed scaling related to SRPT, PS and FIFO scheduling is studied and discussed.

Apart from scheduling policies in queueing models, the number of servers in the queueing system is also an important study issue. Most of the previous papers we mentioned work mainly on the single-server systems (except [6], [30] and [7]) and study the interaction between speed scaling and scheduling. However, in practical applications, there are usually multiple servers. Thus, it is also important to analyze the energy efficiency with respect to multiple servers. Several such kinds of papers have appeared in recent years. Paper [17] studied the optimality of policies for server farm management. In [5] and [15], load balancing is explored under the situation that there are multiple servers. Optimal load balancing in processor sharing systems for multiple PS servers is discussed in [5]. In [15], the focus is on the interaction between the gated-static speed scaling and load balancing for multiple servers. The dynamic dispatching problem in parallel queues (which can be seen as multiple servers) is considered in [30].

Since the papers mentioned above studied speed scaling under ideal conditions, they did not take switching delay into account. However, the cost of switching delay can be very significant. Thus, paper [18] analyzed the M/M/k-FIFO queue with exponentially distributed setup time (switching delay) and studied the costs of setup time.

There are some other studies on the performance of the queueing system considering switching delay, such as [38], [28], and [12]. In [38], a generalized M/G/1-FIFO queue is considered, and two special cases are discussed with respect to the delay. It is assumed that the distribution of the switching delay is arbitrary and does not depend on the service time. Only the second special case is about switching delay, where the server is idle when the first customer arrives, and he has to wait for some random delay so that the server turns to the busy state and serves him. The author examined the transient and asymptotic behavior of M/G/1 queue with switching delay. Paper [28] also analyzed the M/G/1-FIFO queue. One case considered in this paper is that there is switching delay when starting an idle server. The authors concluded that the delay distribution in this case is consisted of two parts. The first part is the delay in the M/G/1-FIFO queue without switching delay. The second part is related to the switching delay. In [12], the author discussed the impact of switching delay and service disciplines on the performance of the M/G/1 queue. The server has a vacation time after it has served a certain number of customers, and if the queue is not empty when the server comes back, there will be an extra delay before the server can serve the customers, which is also the switching delay.

3 Theoretical background

After examining literature on energy-aware performance of data centers, now we focus on the theoretical background for the analysis of queueing models in later sections. Processors in data centers are modeled as queueing models, where the arriving jobs can be modeled as a stochastic process. Here two common stochastic processes will be presented: Markov processes and regenerative processes. For Markov processes, the global balance equations (GBE) will be introduced, and two types of Markov processes, irreducible Markov processes and birth-death processes, will be discussed. As for the following subsection on regenerative processes, renewal sequences and processes as well as Wald's equation will be presented. Afterwards, the definition of a regenerative process will be given.

More information related to this section can be found in [34], [26], [35], [8], [22], [36] and [25].

3.1 Markov processes

A stochastic process is a set of random variables $X(t)$, where t is often used to refer to a time parameter that indexes the random variables. The possible values of $X(t)$ constitute the *state space* S of this stochastic process, and the possible values of parameter t compose the *parameter space* I. A Markov process is a stochastic process that satisfies the Markov property, which declares that the future state of a process is only dependent on its current state, not on its past state.

Let $X(t)$ be a Markov process with state space S. The *state transition rate* q_{ij} is the rate that the process leaves from state i to state j,

$$q_{ij} \geq 0, \quad j \neq i \quad \text{and} \quad i, j \in S.$$

Let q_i denote the *rate* that the process leaves state i,

$$q_i = \sum_{j \neq i} q_{ij}.$$

Let T_i denote the total time that $X(t)$ stays in state i. According to the Markov property, they are independent random variables that follow the $\text{Exp}(q_i)$ distribution. In addition, the *probability* that $X(t)$ jumps from state i to state j is

$$p_{ij} = q_{ij}/q_i, \quad j \neq i.$$

For Markov processes, the long term behavior is usually of special interest. Thus, we introduce the limiting distribution and equilibrium distribution, which are related to the long term behavior of Markov processes.

Define $\pi_i(t)$ as the *time-dependent probability*, which is the probability that a process is in state i at time t,

$$\pi_i(t) = P\{X(t) = i\}.$$

Let π_i denote the *limiting distribution* defined by

$$\pi_i = \text{lim}_{t\to\infty}\pi_i(t) = \text{lim}_{t\to\infty}P\{X(t) = i\}.$$

This limiting distribution satisfies

$$\sum_i \pi_i q_{ij} = 0, \quad \text{for all} \quad j \in S,$$

which are called the *global balance equations* (GBE). These equations can be also written as

$$\sum_{j\neq i} \pi_i q_{ij} = \sum_{j\neq i} \pi_j q_{ji}.$$

If for each $i \in S$, π_i satisfies the global balance equations and also the *normalization condition*, which is

$$\sum_{i\in S} \pi_i = 1,$$

then it is the *equilibrium distribution.*

3.1.1 Irreducible Markov processes

Irreducible Markov processes are special types of Markov processes. Before introducing these processes, we first give two definitions related to them. If the process can transfer from state i to state j in the state transition diagram, which means the probability that state i goes to state j is positive, then state j is *accessible* from state i ($i \to j$). Moreover, if $i \to j$ and $j \to i$, then state i and state j *communicate.*

For a Markov process, if all the states in its state space S communicate with each other, then it is an *irreducible* Markov process.

The equilibrium distribution for an irreducible Markov process is unique (whenever it exists).

3.1.2 Birth-death process

In subsection 3.1.1, a brief introduction to irreducible Markov processes is given. Now the focus will be on birth-death processes for which $S \subset \{0, 1, 2, ...\}$. For a Markov process, if the state transitions happen only between the neighbouring states, then it is a *birth-death* process.

Birth-death processes are important in queueing theory. Common queueing models such as M/M/1, M/M/∞ (we will explain them in later sections) are birth-death processes. In queueing theory, if a process is modeled as a birth-death process, then the state of the process is the number of customers (jobs), and the arrivals of customers and the departures of customers correspond to "births" and "deaths", respectively. In birth-death processes, if the current state is i, then there are only two possibilities for the next state: state $i-1$ or state $i+1$. Parameter λ_i is defined as the birth rate, which is the transition rate from state i to state $i+1$, and μ_i is the death rate, which refers to the transition rate from state i to state $i-1$.

According to the definition of irreducibility given in subsection 3.1.1, it is obvious that a birth-death process is irreducible when the birth rates $\lambda_i > 0$ and the death rates $\mu_i > 0$ for all i. There are two types of irreducible birth-death processes: an infinite-state irreducible birth-death process and a finite-state irreducible birth-death process. Figure 2 shows the state transition diagram of an infinite-state irreducible birth-death process.

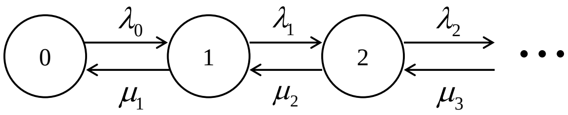


Figure 2: Infinite-state irreducible birth-death process.

Instead of the global balance equations, in birth-death processes, the so-called local balance equations (LBE) may be applied, which are:

$$\pi_i q_{ij} = \pi_j q_{ji}, \quad \text{for all} \quad i, j \in S.$$

Similarly, the normalization condition applies to birth-death processes.

Next, we will derive the equilibrium distribution of the infinite-state irreducible birth-death processes. By using the local balance equations, we have

$$\pi_i \lambda_i = \pi_{i+1} \mu_{i+1},$$

which gives

$$\pi_{i+1} = \frac{\lambda_i}{\mu_{i+1}} \pi_i.$$

By utilizing this equation recursively, we get

$$\begin{aligned} \pi_i &= \frac{\lambda_{i-1}}{\mu_i} \pi_{i-1} \\ &= \frac{\lambda_{i-1}}{\mu_i} \frac{\lambda_{i-2}}{\mu_{i-1}} \pi_{i-2} \\ &= \cdots\cdots \\ &= \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}. \end{aligned} \tag{1}$$

Then one substitutes (1) to the normalization condition,

$$\sum_{i\in S}\pi_i = \pi_0 \sum_{i\in S}\prod_{j=1}^{i}\frac{\lambda_{j-1}}{\mu_j} = 1.$$

When $\sum_{i\in S}\prod_{j=1}^{i}\frac{\lambda_{j-1}}{\mu_j} < \infty$, the equilibrium distribution exists. Then, due to the normalization condition, we have

$$\pi_0 = (1+\sum_{i=1}^{\infty}\prod_{j=1}^{i}\frac{\lambda_{j-1}}{\mu_j})^{-1}.$$

3.2 Regenerative processes

In addition to Markov processes, there is another type of stochastic processes that is useful in queueing theory: regenerative processes. A regenerative process has the property that at certain time points, it will start itself again probabilistically. In this section, basics of regenerative processes and renewal processes will be presented.

3.2.1 Renewal sequences and processes

First, the definition of renewal sequences and renewal processes will be given. Let T_n be a sequence of nonnegative and IID random variables and define

$$\begin{cases}\tau_0 = 0,\\ \tau_n = T_1 + ... + T_n.\end{cases}$$

Thus, τ_n is the sequence that sums up the first n random variables of the renewal sequence T_n, called the *sum sequence* of T_n.

Now, define a counter process $N(t)$ by

$$\begin{cases}N(0) = 0,\\ N(t) = \sum_{n=1}^{\infty} 1_{\{\tau_n \le t\}}.\end{cases}$$

Then $N(t)$ is the corresponding renewal process. Figure 3 illustrates the renewal sequence as well as the associated renewal processes.

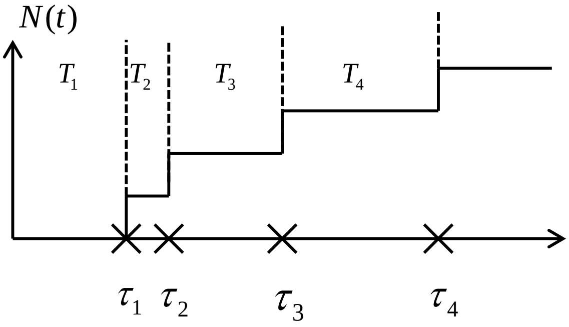


Figure 3: Renewal sequence and associated renewal processes.

Below we give an important equation in renewal processes: Wald's equation. However, before that, we define the concept of stopping time, which will be used in the definition of Wald's equation.

Denote by T_n a renewal sequence, and let N be a random variable. If the occurrence of event $\{N = n\}$ depends only on $T_1, T_2 ... T_n$, then N is called a *stopping time.*

If the expectation of the renewal sequence T_n, $E[T]$, as well as the expectation of the stopping time of the sequence N, $E[N]$, are finite, then

$$E[\sum_{n=1}^{N} T_n] = E[N]E[T]. \tag{2}$$

Equation (2) is referred to as *Wald's equation.* According to the definition above, this equation can be applied only when N is a stopping time for the renewal sequence.

3.2.2 Regenerative processes

Let $X(t)$ be a stochastic process with $X(t) \geqslant 0$ and T_n a renewal sequence. Process $Y_n(t)$ is the nth *cycle* of process $X(t)$ if

$$Y_n(t) = X(\tau_{n-1} + t)1_{\{\tau_{n-1} \leqslant t < \tau_n\}}, \tag{3}$$

where τ_n refers to the sum sequence of the renewal sequence T_n.

If the cycles $Y_n(t)$ are IID, then process $X(t)$ is a *regenerative process* with respect to the renewal sequence T_n. Figure 4 gives an example of a regenerative process. This figure defines that an idle period starts a cycle and the next idle period ends this cycle. This kind of a regenerative process will be utilized in the following sections related to the M/G/1-FIFO queue.

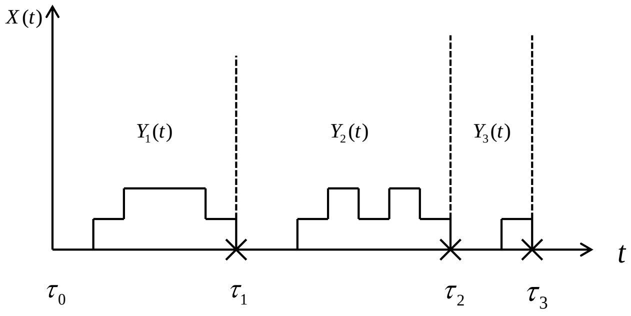


Figure 4: Regenerative process.

If $X(t)$ is a regenerative process, and T_n is the renewal sequence that has a non-lattice distribution with a finite mean, $E[T] = E[T_n] < \infty$, and $f(x)$ is a non-negative function defined on $[0, \infty)$, then

$$\lim_{t\to\infty} E[f\,(X(t))] = \frac{E[\int_0^T f\,(X(t))\,dt]}{E[T]}. \tag{4}$$

This equation will be used in the analysis of M/G/1-FIFO queueing models.

4 Classical performance analysis

The preceding three sections present the background on energy consumption in data centers, literature related to energy-aware performance in data centers, as well as theoretical background on stochastic processes. In this section, we will combine stochastic processes with queueing theory. First we study the characteristics of different queueing models without switching delay. Then we will analyze the properties of those queueing models with switching delay. At last, we will examine the performance of these queueing models with two different switching delay distributions.

More information related to this section can be found in [26], [8] and [37].

4.1 Queueing analysis without switching delay

As specified before, the ideal situation of queueing analysis does not consider switching delay. Therefore, we first focus on this ideal situation.

Let $X(t)$ denote the queueing process at time t and X the corresponding steady-state variable. Denote the arrival rate by λ, and the service rate by μ. Then we define the traffic load $\rho = \lambda/\mu$. Let S_i be the service time of customer i and S a generic service time. In addition, denote the arrival time of customer i by α_i. We define the inter-arrival time as $A_i = \alpha_{i+1} - \alpha_i$.

First, we discuss the single server queue, i.e., the M/G/1 queue. In the M/G/1 queue, the number of customers is infinite and customers are independent. In addition, there are ∞ customer places in the system. The inter-arrival times of customers are IID and exponentially distributed with mean of $1/\lambda$. Service times are IID and generally distributed with mean of $1/\mu$. Figure 5 shows the M/G/1 queue model. If the service times are IID and exponentially distributed, then this queue is an M/M/1 queue.

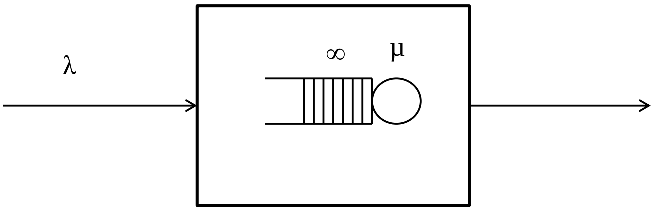


Figure 5: M/G/1 queue model diagram.

However, if the number of servers is infinite, then it is an M/G/∞ queue. Figure 6 illustrates an M/G/∞ queue model. Similarly, if service times are IID and exponentially distributed, then it is an M/M/∞ queue.

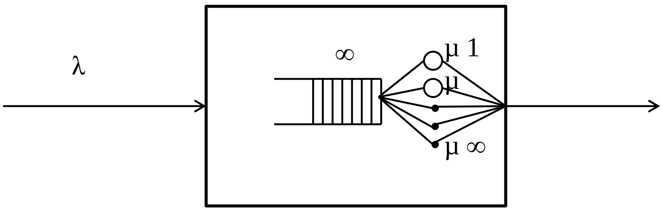


Figure 6: M/G/∞ queue model diagram.

4.1.1 M/M/1

After introducing two different queueing models briefly, in this section, we will first analyze the performance of the M/M/1 queue without switching delay. Recall that $X(t)$ is the queue length of the system at time t, which, in this case, is a Markov process with the state transition diagram given in Figure 7.

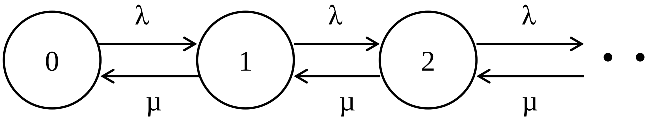


Figure 7: M/M/1 state transition diagram.

According to Section 3.1.2, $X(t)$ is an irreducible birth-death process with an infinite state space $\{0, 1, 2, ...\}$.

Recall that π_i is the equilibrium distribution of the Markov process $X(t)$. According to the state transition diagram, we apply the local balance equations (LBE):

$$\pi_i \lambda = \pi_{i+1} \mu,$$

from which we get

$$\pi_{i+1} = \frac{\lambda}{\mu} \pi_i = \rho \pi_i. \tag{5}$$

By utilizing (5) recursively, we have

$$\begin{aligned} \pi_i &= \rho \pi_{i-1} \\ &= \rho^2 \pi_{i-2} \\ &= ... \\ &= \rho^i \pi_0, \quad i = 0, 1, 2, ... \end{aligned} \tag{6}$$

Substituting (6) to the normalization condition gives

$$\sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \rho^i = 1. \tag{7}$$

It follows that

$$\pi_0 = \left(\sum_{i=0}^{\infty} \rho^i \right)^{-1} = \left(\frac{1}{1-\rho} \right)^{-1} = 1 - \rho, \quad \text{if } \rho{<}1. \tag{8}$$

Combining (6) and (8) gives

$$\pi_i = \rho^i (1 - \rho), \quad i = 0, 1, 2, ...$$

Thus, for a stable system ($\rho < 1$), the equilibrium distribution exists and is a geometric distribution such that $X \sim \text{Geom}(\rho)$. Since π_i is the steady-state probability that the process is in state i, we have

$$P\{X = i\} = \pi_i = \rho^i (1 - \rho), \quad i = 0, 1, 2, ... \tag{9}$$

Equation (9) indicates that the probability that a Markov process is in state i is only related to the traffic load of the system.

For the mean value and variance of X, we get

$$E[X] = \frac{\rho}{1-\rho}, \quad \text{and} \quad D^2[X] = \frac{\rho}{(1-\rho)^2}. \tag{10}$$

4.1.2 M/G/1-FIFO

Since service disciplines affect the performance of queueing models, in this subsection, we will discuss the service disciplines in the M/G/1 queue. The first service discipline we will examine is FIFO. This discipline states that the arriving customers are served based on the arrival order.

In the M/G/1-FIFO queue, the time duration that there is no customer being served is called *idle period*, and the time duration that there are customers being served is called *busy period*. One idle period and the following busy period is a *busy cycle*.

Let I denote the idle period, B the busy period, C the busy cycle, and N the number of served customers in the busy cycle. In addition, the corresponding symbols for the nth cycle will be I_n, B_n, C_n and N_n, respectively. According to Section 3.2.2, the busy cycles C_n in the M/G/1-FIFO are regenerative. Figure 8 shows the corresponding busy period, idle period and busy cycle.

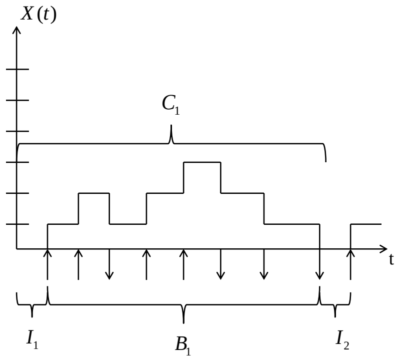


Figure 8: Busy cycle in the M/G/1-FIFO queue.

Now, we will analyze the performance of the M/G/1 queue under the FIFO service discipline. First we calculate $E[I]$, $E[B]$, $E[C]$, $E[N]$.

Since the inter-arrival times are exponentially distributed in the M/G/1-FIFO queue, according to the memoryless property for exponential distribution, the idle period has the same distribution as the inter-arrival times, and thus

$$E[I] = \frac{1}{\lambda}. \tag{11}$$

In addition, the first busy period B_1 is given by

$$B_1 = \sum_{i=1}^{N_1} S_i, \tag{12}$$

where N_1 is the number of customers served in the first busy period taking values in {i=1, 2,...}, and S_i is the service time of customer i. The occurrence of event $\{N_1 = 1\} = \{S_1 < A_1\}$ is totally determined by the pair (S_1, A_1). Also, event $\{N_1 = 2\} = \{S_1 \geqslant A_1, S_1 + S_2 < A_1 + A_2\}$ is only dependent on the pairs (S_1, A_1) and (S_2, A_2). Thus N_1 is a stopping time with respect to the IID sequence (S_i, A_i). Note that S_1, S_2,... are IID random variables with mean $E[S]$. According to Wald's equation, we have

$$\begin{aligned} E[B] &= E[B_1] \\ &= E\left[\sum_{i=1}^{N_1} S_i\right] \\ &= E[N]E[S]. \end{aligned} \tag{13}$$

As the definition of the busy cycle gives

$$C_1 = B_1 + I_1, \tag{14}$$

we have

$$E[C] = E[C_1] = E[B_1] + E[I_1] = E[B] + E[I]. \tag{15}$$

By definition, we have

$$B_1 + I_2 = \sum_{i=1}^{N_1} A_i. \tag{16}$$

By Wald's equation, we have

$$E[C] = E[B] + E[I] = E[B_1] + E[I_2] = E[N]E[A] = E[N]E[I]. \tag{17}$$

From the equations above, we get

$$E[N] = \frac{E[I]}{E[I] - E[S]} = \frac{1}{1-\rho}, \tag{18}$$

$$E[C] = E[N]E[I] = \frac{1}{1-\rho}\frac{1}{\lambda}. \tag{19}$$

Denote the remaining service time of the customer in service at time t by $R(t)$. Figure 9 illustrates the remaining service time process. Since the busy cycles C_n are regenerative, the remaining service time process $R(t)$ is also regenerative in each busy cycle. The busy cycles C_n are IID with non-lattice distributions, and then from (4), we have

$$E[R] = \frac{E[\int_0^C R(t)dt]}{E[C]}. \tag{20}$$

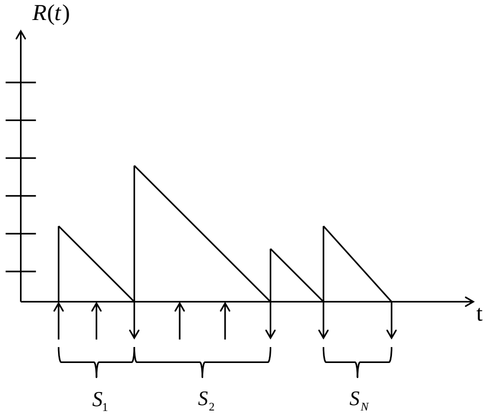


Figure 9: Remaining service time process.

From Figure 9, we get

$$\int_0^C R(t)dt = \sum_{i=1}^{N} \frac{1}{2} S_i^2. \tag{21}$$

By Wald's equation,

$$\begin{aligned} E\left[\int_0^C R(t)dt\right] &= \frac{1}{2}E[N]E\left[S^2\right] \\ &= \frac{E\left[S^2\right]}{2(1-\rho)}. \end{aligned} \tag{22}$$

By substituting (19) and (22) to (20), we get the mean steady-state remaining service time

$$\begin{aligned} E[R] &= \frac{E\left[\int_0^C R(t)dt\right]}{E[C]} \\ &= \frac{\lambda}{2}E\left[S^2\right]. \end{aligned} \tag{23}$$

Let W_i denote the waiting time of customer i. It satisfies

$$W_i = \sum_{j=1}^{Y_i^w} S_{i-j} + R(\alpha_i-), \tag{24}$$

where Y_i^w refers to the number of waiting customers that the arriving customer i sees, $R(\alpha_i-)$ refers to the remaining service time seen by the arriving customer i. Random variable Y_i^w is independent of the service times S_{i-1},...,$S_{i-Y_i^w}$. By utilizing the conditioning rule, we get

$$\begin{aligned} E[W_i] &= E\left[E\left[\sum_{j=1}^{Y_i^w} S_{i-j}|Y_i^w\right]\right] + E\left[R(\alpha_i-)\right] \\ &= E\left[Y_i^w E\left[S\right]\right] + E\left[R(\alpha_i-)\right] \\ &= E\left[Y_i^w\right] E[S] + E\left[R(\alpha_i-)\right]. \end{aligned} \tag{25}$$

According to PASTA, $\lim_{i\to\infty} Y_i^w$ has the same distribution as $X^w := \lim_{t\to\infty} X(t) - 1_{\{X(t)>0\}}$, and $\lim_{i\to\infty} R(\alpha_i-)$ is distributed as $R := \lim_{i\to\infty} R(t)$. Thus, we obtain

$$\begin{aligned} E[W] &= E[X^w]E[S] + E[R] \\ &= \lambda E[W]E[S] + E[R]. \end{aligned} \tag{26}$$

By combining (23) and (26), we have

$$\begin{aligned} E[W] &= \frac{E[R]}{1-\rho} \\ &= \frac{\lambda E\left[S^2\right]}{2(1-\rho)}. \end{aligned} \tag{27}$$

So, the mean steady-state sojourn time is

$$\begin{aligned} E[T] &= E[S] + E[W] \\ &= E[S] + \frac{\lambda E\left[S^2\right]}{2(1-\rho)}. \end{aligned} \tag{28}$$

Applying Little's formula, we obtain the mean steady-state queue length

$$\begin{aligned} E[X] &= \lambda E[T] \\ &= \rho + \frac{\lambda^2 E\left[S^2\right]}{2(1-\rho)}. \end{aligned} \tag{29}$$

Equations (27), (28) and (29) are known as the *Pollaczek-Khinchin* mean value formulas.

From equation (29) we see that, the mean queue length in the M/G/1-FIFO queue is affected by the service time distribution.

4.1.3 M/G/1-PS

Apart from FIFO, there are other service disciplines that can be used to model data centers, such as PS. For the PS service discipline, the customers are served simultaneously, and the service capacity is shared evenly among all customers.

According to [25], PS is a work-conserving discipline and the steady-state queue length distribution of the PS discipline is insensitive to the service time distribution. Thus, the steady-state queue length distribution of the M/G/1-PS queue is the same as that in the M/M/1-PS queue. By utilizing the result in Section 4.1.1, we have

$$P\{X = i\} = \rho^i(1-\rho), \quad i = 0, 1, 2, ... \tag{30}$$

The mean value and variance of the queue length can be obtained by utilizing the queue length distribution

$$E[X] = \frac{\rho}{1-\rho}, \quad \text{and} \quad D^2[X] = \frac{\rho}{(1-\rho)^2}.$$

4.1.4 M/M/∞

The previous parts in this section discussed the queueing models with a single server. In this subsection, we will consider an M/M/∞ queue, which has an infinite number of servers. Recall that the queue length of the system at time t is denoted by $X(t)$, and it is a Markov process with the state transition diagram given in Figure 10,

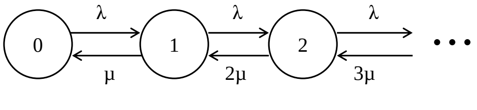


Figure 10: M/M/∞ state transition diagram.

Clearly, $X(t)$ is an irreducible birth-death process with an infinite state space $\{0, 1, 2, ...\}$. Therefore, we can apply the local balance equations,

$$\pi_i \lambda = \pi_{i+1}(i+1)\mu.$$

Then we get

$$\pi_{i+1} = \frac{\lambda}{(i+1)\mu}\pi_i = \frac{\rho}{i+1}\pi_i. \tag{31}$$

From (31), we derive,

$$\begin{aligned} \pi_i &= \frac{\rho}{i}\pi_{i-1} \\ &= \frac{\rho}{i} \cdot \frac{\rho}{i-1}\pi_{i-2} \\ &= ... \\ &= \frac{\rho^i}{i!}\pi_0, \quad i = 0, 1, 2, ... \end{aligned}$$

Combining it with the normalization condition gives

$$\sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \frac{\rho^i}{i!} = 1,$$

resulting in

$$\pi_0 = \left(\sum_{i=0}^{\infty} \frac{\rho^i}{i!} \right)^{-1} = (e^{\rho})^{-1} = e^{-\rho}.$$

Thus, the equilibrium distribution of $X(t)$ is a Poisson distribution with $X \sim$ Poisson(ρ), and the probability that the process stays in state i is

$$P\{X = i\} = \pi_i = \frac{\rho^i}{i!}e^{-\rho}, \quad i = 0, 1, 2, ...,$$

with the mean value and variance given by

$$E[X] = \rho, \quad \text{and} \quad D^2[X] = \rho.$$

Then, we obtain the second moment of the queue length

$$E[X^2] = E[X]^2 + D^2[X] = \rho^2 + \rho.$$

According to [25], the steady-state queue length distribution for the M/M/∞ queue is insensitive to the service time distribution. Thus, these results can be applied to the M/G/∞ queue.

4.2 Queueing analysis with switching delay

Having examined different queueing models without switching delay, now we move on to the analysis of those queueing models with switching delay. If the system is empty when the customer arrives, there is a random delay to switch the system to the busy state. This delay is called switching delay. It may influence the performance of queueing models. Denote the switching delay by a random variable D. We assume that switching delays are independent (of everything) and identically distributed.

4.2.1 M/M/1 with exponential switching delay

In the M/M/1 queue with switching delay, we need to consider a two-dimensional Markov process $(X(t), Z(t))$, where $X(t)$ is the number of customers in the system at time t, and $Z(t)$ indicates whether the system is in the switching delay state or not at time t. When the system is in the switching delay state, $Z(t) = 1$, otherwise, $Z(t) = 0$. Here we assume the switching delay D is exponentially distributed with mean $d = \frac{1}{\delta}$.

The state transition diagram of the Markov process $(X(t), Z(t))$ is given in Figure 11,

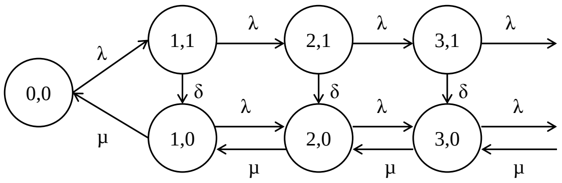


Figure 11: M/M/1 state transition diagram with switching delay.

Denote by $\pi_{i,j}$ the equilibrium distribution of the M/M/1 queue with switching delay, $i = 0, 1, 2, ...,$ $j = 0, 1$. According to the global balance equations, for state (0,0), we have

$$\pi_{1,0}\mu = \pi_{0,0}\lambda.$$

For states (i,1), $i = 1, 2, ...,$ we have

$$\begin{cases} \pi_{0,0}\lambda & = \pi_{1,1}(\lambda + \delta), \\ \pi_{1,1}\lambda & = \pi_{2,1}(\lambda + \delta), \\ & . \\ & . \\ & . \\ \pi_{i-1,1}\lambda & = \pi_{i,1}(\lambda + \delta). \end{cases}$$

By solving the equations above, we get

$$\begin{aligned}\pi_{i,1} &= \frac{\lambda}{\lambda+\delta}\pi_{i-1,1} \\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^2 \pi_{i-2,1} \\ &= \ldots\ldots \\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^{i-1} \pi_{1,1}, \\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^{i} \pi_{0,0}, \quad i = 1, 2, 3, ... \end{aligned} \tag{32}$$

For state (1,0), the global balance equation is

$$\pi_{1,1}\delta + \pi_{2,0}\mu = \pi_{1,0}(\lambda + \mu),$$

and for states (i,0), it reads as

$$\pi_{i,1}\delta + \pi_{i+1,0}\mu + \pi_{i-1,0}\lambda = \pi_{i,0}(\lambda + \mu), \quad i = 2, 3, ...$$

In addition, the normalization condition is as follows:

$$\pi_{0,0} + \sum_{i=1}^{\infty}(\pi_{i,0} + \pi_{i,1}) = 1.$$

After solving the equations above, we get

$$\pi_{0,0} = \frac{\delta}{\lambda+\delta}\left(1 - \frac{\lambda}{\mu}\right),$$

$$\pi_{i,1} = -\frac{\delta(\lambda-\mu)}{\mu}\frac{\lambda^i}{(\lambda+\delta)^{i+1}}, \quad i = 1, 2, ...,$$

and

$$\pi_{i,0} = \frac{\delta(\lambda-\mu)}{\mu(\delta+\lambda-\mu)}\left(\left(\frac{\lambda}{\lambda+\delta}\right)^i - \left(\frac{\lambda}{\mu}\right)^i\right), \quad i = 1, 2, ...$$

With some manipulations, we can get the expectation of the queue length

$$E[X] = \frac{\lambda}{\mu-\lambda} + \frac{\lambda}{\delta} = \frac{\rho}{1-\rho} + \frac{\lambda}{\delta}. \tag{33}$$

From the result above, we observe that the mean queue length of the system consists of two parts, $\rho/(1-\rho)$ and λ/δ. The first part, $\rho/(1-\rho)$, equals to the mean queue length of the M/M/1 system without switching delay. The second part, λ/δ, is related to the mean of the switching delay. Thus, the impacts of the service time and the switching delay on the mean queue length are separated.

By Little's formula, the mean delay is

$$\begin{aligned} E[T] &= \frac{E[X]}{\lambda} \\ &= \frac{1}{\mu-\lambda}+\frac{1}{\delta}. \end{aligned} \tag{34}$$

The first part in (34), $1/(\mu-\lambda)$, is the same as the mean delay in the M/M/1 queue without switching delay. The second part, $1/\delta$, equals to the mean of the switching delay D.

Note that for solving the equations above, the symbolic computing capabilities of Mathematica were utilized.

4.2.2 M/M/∞ with exponential switching delay

Having analyzed the performance of the M/M/1 queue with exponential switching delay, now we will study the impact of an exponential switching delay on the M/M/∞ queue.

In the M/M/∞ queue with switching delay, a similar two-dimensional Markov process $(X(t), Z(t))$ is used, but the state transition diagram is different, which is shown in Figure 12.

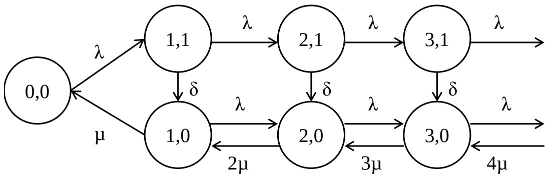


Figure 12: M/M/∞ state transition diagram with switching delay.

Again, by applying the global balance equation for state (0,0), we get

$$\pi_{1,0}\mu = \pi_{0,0}\lambda. \tag{35}$$

For states (i,1), $i = 1, 2, ...$, we have

$$\begin{cases} \pi_{0,0}\lambda &= \pi_{1,1}(\lambda+\delta), \\ \pi_{1,1}\lambda &= \pi_{2,1}(\lambda+\delta), \\ & . \\ & . \\ & . \\ \pi_{i-1,1}\lambda &= \pi_{i,1}(\lambda+\delta). \end{cases} \tag{36}$$

By iteratively applying (36), we get

$$\begin{aligned}\pi_{i,1} &= \frac{\lambda}{\lambda+\delta}\pi_{i-1,1}\\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^2\pi_{i-2,1}\\ &= \ldots\ldots\\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^{i-1}\pi_{1,1},\\ &= \left(\frac{\lambda}{\lambda+\delta}\right)^{i}\pi_{0,0}, \quad i=1,2,3,... \end{aligned} \tag{37}$$

For state (1,0), the global balance equation is

$$\pi_{1,1}\delta + \pi_{2,0}2\mu = \pi_{1,0}(\lambda+\mu), \tag{38}$$

and for states (i,0), we have

$$\pi_{i,1}\delta + \pi_{i+1,0}(i+1)\mu + \pi_{i-1,0}\lambda = \pi_{i,0}(\lambda+i\mu), \quad i=2,3,... \tag{39}$$

By substituting (37) to (39), we get

$$(\frac{\lambda}{\lambda+\delta})^i\pi_{0,0}\delta + \pi_{i+1,0}(i+1)\mu + \pi_{i-1,0}\lambda = \pi_{i,0}(\lambda+i\mu), \quad i=2,3,... \tag{40}$$

Then substituting (35) to (40) yields

$$(\frac{\lambda}{\lambda+\delta})^i\pi_{1,0}\frac{\mu}{\lambda}\delta + \pi_{i+1,0}(i+1)\mu + \pi_{i-1,0}\lambda = \pi_{i,0}(\lambda+i\mu), \quad i=2,3,...$$

By evaluating the global balance equations for higher values of i, we deduce that

$$\begin{aligned}\pi_{i,0} &= \frac{\lambda^{i-1}\left(\sum_{n=0}^{i-1}(\lambda+\delta)^{i-1-n}\mu^n n!\right)}{i!\,(\lambda+\delta)^{i-1}\mu^{i-1}}\pi_{1,0}\\ &= \sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-1-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n\pi_{1,0}\\ &= \sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n\pi_{0,0}. \end{aligned} \tag{41}$$

The normalization condition is

$$\pi_{0,0} + \sum_{i=1}^{\infty}(\pi_{i,0}+\pi_{i,1}) = 1.$$

By substituting $\pi_{i,0}$ and $\pi_{i,1}$ to the normalization condition, we have

$$\pi_{0,0} + \sum_{i=1}^{\infty}\left(\left(\frac{\lambda}{\lambda+\delta}\right)^i\pi_{0,0} + \sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n\pi_{0,0}\right) = 1.$$

Thus,

$$\pi_{0,0}\left(1+\frac{\lambda}{\delta}+\sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^{n}\right)=1,$$

and finally, we get

$$\pi_{0,0}=\left(1+\frac{\lambda}{\delta}+\sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^{n}\right)^{-1}. \tag{42}$$

Asymptotic result

Above we consider the general case of the M/M/∞ queue with switching delay, where the service speed and the service rate is finite. Here, we will give the asymptotic result when μ goes to infinity.

If the service rate μ goes to infinity, the state transition diagram simplifies considerably as shown in Figure 13.

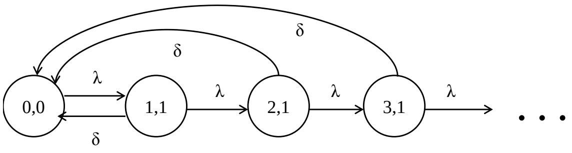


Figure 13: M/M/∞ state transition diagram with switching delay($\mu \to \infty$).

By applying the global balance equation, for state (0,0) we get

$$(\pi_{1,1}+\pi_{2,1}+...+\pi_{i,1})\delta=\pi_{0,0}\lambda, \tag{43}$$

which equals

$$(1-\pi_{0,0})\delta=\pi_{0,0}\lambda. \tag{44}$$

Thus, we get

$$\pi_{0,0}=\frac{\delta}{\lambda+\delta}. \tag{45}$$

For state (i,1), $i=1,2,...$, the equations are the same as (36) and (37). Thus, we have

$$\begin{aligned}\pi_{i,1} &= \left(\frac{\lambda}{\lambda+\delta}\right)^{i}\pi_{0,0}\\ &= \frac{\delta\lambda^{i}}{(\lambda+\delta)^{i+1}}, \quad i=1,2,3,...\end{aligned} \tag{46}$$

So, the asymptotic distribution for the M/M/∞ queue with switching delay (with $\mu \to \infty$) is geometric with parameter $\frac{\lambda}{\lambda+\delta}$.

Then, the mean queue length of the system

$$\begin{aligned} E[X] &= \sum_{i}^{\infty} i\pi_{i,1} \\ &= \sum_{i}^{\infty} i\frac{\delta\lambda^i}{(\lambda+\delta)^{i+1}} \\ &= \frac{\lambda}{\delta}. \qquad (47) \end{aligned}$$

Note that the asymptotic value for the mean queue length is only related to λ and δ.

4.2.3 M/G/1-FIFO with switching delay

In the M/G/1-FIFO queue with switching delay, we use the same symbols as in the M/G/1-FIFO queue without switching delay. In addition, define D as the switching delay. As said in Section 4.1.2, the busy cycles in the M/G/1-FIFO queue without switching delay are regenerative. This still applies to the M/G/1-FIFO queue with switching delay. Figure 14 illustrates the first busy cycle. From Figure 14, we see that the switching delay D will affect the busy periods and busy cycles. Moreover, it will have an impact on the number of customers served in the busy cycle.

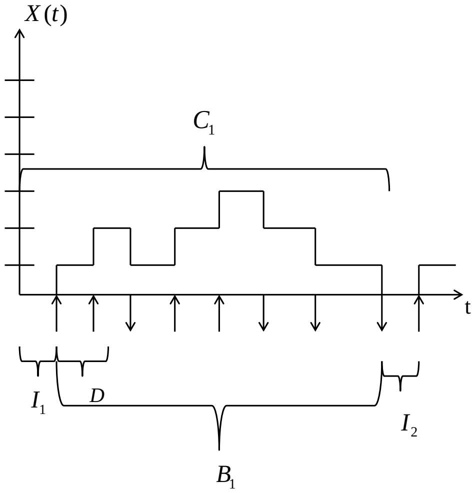


Figure 14: Busy cycle in the M/G/1-FIFO queue with switching delay.

Similarly as before, we first calculate $E[I]$, $E[B]$, $E[C]$, $E[N]$. The inter-arrival times are exponentially distributed in the M/G/1-FIFO queue. Due to the memoryless property of exponential distribution, the idle period has the same distribution as the inter-arrival times. The switching delay D will not influence the mean of the idle period, so

$$E[I] = \frac{1}{\lambda}. \tag{48}$$

Due to the effect of the switching delay, the first busy period B_1 becomes as:

$$B_1 = \sum_{i=1}^{N_1} S_i + D. \tag{49}$$

Similarly as in Section 4.1.2, event $\{N_1 = 1\} = \{D + S_1 < A_1\}$ depends only on the pair (S_1, A_1) and the switching delay D. Moreover, event $\{N_1 = 2\} = \{D + S_1 \geqslant A_1, D + S_1 + S_2 < A_1 + A_2\}$ is totally determined by the pairs (S_1, A_1), (S_2, A_2) and the switching delay D. Thus, N_1 is a stopping time with respect to D and IID sequence (S_i, A_i). Then, according to Wald's equation, we get

$$\begin{aligned} E[B] &= E[B_1] \\ &= E\left[\sum_{i=1}^{N_1} S_i + D\right] \\ &= E[N]E[S] + E[D]. \end{aligned} \tag{50}$$

By definition,

$$E[C] = E[C_1] = E[B_1] + E[I_1] = E[B] + E[I], \tag{51}$$

and

$$B_1 + I_2 = \sum_{i=1}^{N_1} A_i. \tag{52}$$

By Wald's equation, we get

$$E[C] = E[B] + E[I] = E[B_1] + E[I_2] = E[N]E[A] = E[N]E[I]. \tag{53}$$

From the equations above, we get

$$E[N] = \frac{E[D] + E[I]}{E[I] - E[S]} = \frac{1 + \lambda E[D]}{1 - \rho}, \tag{54}$$

$$E[C] = E[N]E[I] = \frac{1 + \lambda E[D]}{1 - \rho}\frac{1}{\lambda}. \tag{55}$$

Similarly as in Section 4.1.2, the waiting time of customer i satisfies

$$W_i = \sum_{j=1}^{Y_i^w} S_{i-j} + R(\alpha_i -), \tag{56}$$

and it follows that

$$E[W] = \frac{E[R]}{1 - \rho}.$$

The remaining service time process $R(t)$ is regenerative in each busy cycle so that

$$E[R] = \frac{E\left[\int_0^C R(t)dt\right]}{E[C]}. \tag{57}$$

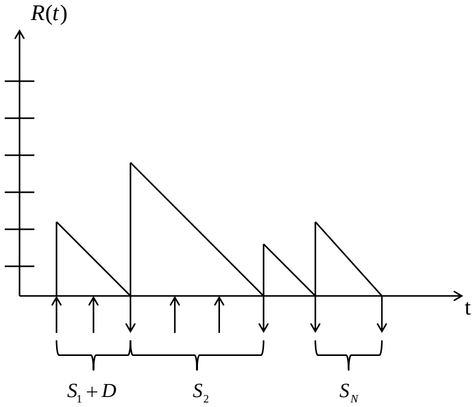


Figure 15: Remaining service time process in the M/G/1-FIFO queue with switching delay.

Figure 15 shows the remaining service time process. We can observe from Figure 15 that

$$\int_0^C R(t)dt = \sum_{i=2}^{N} \frac{S_i^2}{2} + \frac{(D+S_1)^2}{2} = \sum_{i=1}^{N} \frac{S_i^2}{2} + S_1 D + \frac{D^2}{2}. \tag{58}$$

Thus, by Wald's equation

$$\begin{aligned} E\left[\int_0^C R(t)dt\right] &= \frac{E[N]E\left[S^2\right]}{2} + E[S]E[D] + \frac{E\left[D^2\right]}{2} \\ &= \frac{E\left[S^2\right](1+\lambda E[D])}{2(1-\rho)} + E[S]E[D] + \frac{E\left[D^2\right]}{2}. \end{aligned} \tag{59}$$

By substituting (59) to (57), we get

$$\begin{aligned} E[R] &= \frac{E\left[\int_0^C R(t)dt\right]}{E[C]} \\ &= \frac{1}{2}\lambda E\left[S^2\right] + \frac{E[D]\rho(1-\rho)}{1+\lambda E[D]} + \frac{\lambda E\left[D^2\right](1-\rho)}{2(1+\lambda E[D])}. \end{aligned} \tag{60}$$

Thus, we get

$$\begin{aligned} E[W] &= \frac{E[R]}{1-\rho} \\ &= \frac{\lambda E\left[S^2\right]}{2(1-\rho)} + \frac{E[D]\rho}{1+\lambda E[D]} + \frac{\lambda E\left[D^2\right]}{2(1+\lambda E[D])}. \end{aligned} \tag{61}$$

Let $\tilde{S}$ denote the effective service time of a customer defined by

$$\tilde{S} = S + D \cdot 1_{\{\text{the customer is the first one in a busy period}\}}. \tag{62}$$

According to Little's formula:

$$E[X] = \lambda \left(E[\tilde{S}] + E[W] \right), \tag{63}$$

where

$$\begin{aligned} E[\tilde{S}] &= E[S] + E[D]P\{\text{the customer is the first one in a busy period}\} \\ &= E[S] + E[D]\frac{1}{E[N]} \\ &= E[S] + E[D]\frac{1-\rho}{1+\lambda E[D]}. \end{aligned} \tag{64}$$

Thus, we obtain

$$E[X] = \lambda \left(E[S] + \frac{\lambda E\left[S^2\right]}{2(1-\rho)} + \frac{E[D]}{1+\lambda E[D]} + \frac{\lambda E\left[D^2\right]}{2(1+\lambda E[D])} \right). \tag{65}$$

From (65), we can see that in the M/G/1-FIFO queue with switching delay, the mean queue length is sensitive to the service time distribution. Another important observation is that the impacts of service time S and switching delay D on the mean queue length are separated.

Again, applying Little's formula, we get the mean delay is

$$\begin{aligned} E[T] &= \frac{E[X]}{\lambda} \\ &= E[S] + \frac{\lambda E\left[S^2\right]}{2(1-\rho)} + \frac{E[D]}{1+\lambda E[D]} + \frac{\lambda E\left[D^2\right]}{2(1+\lambda E[D])}. \end{aligned} \tag{66}$$

The first part in (66), $E[S] + \lambda E\left[S^2\right]/2(1-\rho)$, is the same as the mean delay in the M/G/1-FIFO queue. The second part, $E[D]/(1+\lambda E[D]) + \lambda E\left[D^2\right]/2(1+\lambda E[D])$, is related to the switching delay D.

4.2.4 Examples of M/M/1-FIFO with switching delay

Section 4.2.3 studies the M/G/1-FIFO queue with switching delay. Now in this part, we will examine the effect of switching delay with different distributions. We will present two special cases for the M/M/1-FIFO queue, where $S \sim \text{Exp}(\mu)$, $E[S] = 1/\mu$, and $E[S^2] = 2/\mu^2$. Then, according to (65), we have

$$\begin{aligned} E[X] &= \lambda \left(\frac{1}{\mu} + \frac{\lambda \frac{2}{\mu^2}}{2\,(1-\rho)} + \frac{E[D]}{1+\lambda E[D]} + \frac{\lambda E[D^2]}{2\,(1+\lambda E[D])} \right) \\ &= \rho + \frac{\rho^2}{1-\rho} + \frac{\lambda E[D]}{1+\lambda E[D]} + \frac{\lambda^2 E[D^2]}{2\,(1+\lambda E[D])} \\ &= \frac{\rho}{1-\rho} + \frac{\lambda E[D]}{1+\lambda E[D]} + \frac{\lambda^2 E[D^2]}{2\,(1+\lambda E[D])}. \end{aligned} \tag{67}$$

1. If the switching delay is deterministic and the mean value is d, then $E[D] = d$, $E[D^2] = d^2$, the mean queue length

$$\begin{aligned} E[X] &= \frac{\rho}{1-\rho} + \frac{\lambda d}{1+\lambda d} + \frac{\lambda^2 d^2}{2(1+\lambda d)} \\ &= \frac{\rho}{1-\rho} + \frac{\lambda d}{1+\lambda d}\left(1 + \frac{\lambda d}{2}\right). \end{aligned} \tag{68}$$

2. If the switching delay is exponentially distributed, $D \sim \mathrm{Exp}(\delta)$ with $d = \frac{1}{\delta}$, then $E[D] = \frac{1}{\delta} = d$, $E[D^2] = \frac{2}{\delta^2} = 2d^2$,

$$\begin{aligned} E[X] &= \frac{\rho}{1-\rho} + \frac{\lambda d}{1+\lambda d} + \frac{\lambda^2 d^2}{1+\lambda d} \\ &= \frac{\rho}{1-\rho} + \frac{\lambda d(1+\lambda d)}{1+\lambda d} \\ &= \frac{\rho}{1-\rho} + \lambda d. \end{aligned} \tag{69}$$

   This result corresponds to the mean queue length of (33) in Section 4.2.1, where the switching delay is also exponentially distributed.

Since in the M/M/1 queue, the results are insensitive to the service disciplines, these results for the M/M/1-FIFO queue with switching delay are also valid for the M/M/1-PS queue with switching delay.

5 Energy-aware performance analysis

Previous section analyzes the queueing models in a classic way, which does not take into account the energy-aware cost of those queueing systems. In this section, however, we will consider the average energy-aware cost in queueing models with three different speed scaling methods, which are the static speed scaling, the gated static speed scaling and the linear speed scaling. Recall from Section 1.2, for static speed scaling, the processor always runs at a constant optimal speed. For the gated static speed scaling, if there is no job in the system, the service speed is zero. Otherwise, the processor also runs at a constant optimal speed. The linear speed scaling adjusts the processing speed linearly to the number of jobs in the system.

Similarly as in Section 4, we will first analyze the energy-aware performance of different queueing models without switching delay. Then, we take switching delay into account and conduct a similar analysis.

5.1 Energy-aware schemes without switching delay

We first consider the situation without switching delay. Recall that λ is the arrival rate, μ is the service rate, and the traffic load $\rho = \lambda/\mu$. Let S refer to a generic service time. Denote by s the service speed, and by Y the size of a job with $E[Y] = m$. It follows that $S = Y/s$, and $E[S] = E[Y]/s = m/s = 1/\mu$. Define $r = \lambda m$. Thus, we have $\rho = \lambda/\mu = \lambda m/s = r/s$.

Recall that we define $X(t)$ as the queueing process at time t and X as the corresponding steady-state variable. As mentioned in Sections 2.1 and 2.2, the energy-aware cost model is as follows: with a fixed service speed s, costs are accumulated with rate $X(t) + s^\alpha/\beta$, where $\alpha > 1$ and β is the parameter converting power units to time units. Denote by z the average energy-aware cost per unit time.

The average energy-aware cost per unit time in the static system is clearly

$$z_{static} = E[X] + \frac{s^\alpha}{\beta}. \tag{70}$$

In the gated system, we need to consider the probability that the system is busy, thus

$$\begin{aligned} z_{gated} &= E[X] + \frac{E[1_{\{X>0\}} s^\alpha]}{\beta} \\ &= E[X] + \frac{s^\alpha}{\beta} P\{X > 0\}. \end{aligned} \tag{71}$$

For the linear system, the server speed is linear with respect to the queue length, which means $s_n = ns$ when there are n jobs in the system. Thus, the average cost per unit time can be written as

$$\begin{aligned} z_{linear} &= E[X] + \frac{E\left[(s_X)^\alpha\right]}{\beta} \\ &= E[X] + \frac{E\left[(sX)^\alpha\right]}{\beta}. \end{aligned} \tag{72}$$

Define the energy-aware load as $\gamma = r/\beta^{1/\alpha}$, where parameters α, β will affect the workload.

5.1.1 M/G/1-PS queue with static speed scaling

In the static system, an optimal constant service speed is chosen to minimize the average energy cost. According to Section 4.1.3, the mean queue length in the M/G/1-PS queue $E[X] = \rho/(1-\rho)$. Thus, the average cost per unit time in the static system can be written as

$$\begin{aligned} z_{static} &= \frac{\rho}{1-\rho} + \frac{s^\alpha}{\beta} \\ &= \frac{\frac{r}{s}}{1-\frac{r}{s}} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s-r} + \frac{s^\alpha}{\beta}. \end{aligned} \tag{73}$$

Equation (73) shows that, for the M/G/1-PS queue, the average cost is insensitive to the service time distribution.

We obtain the optimal solution of (73) by taking the derivative with respect to s and solving

$$\frac{\alpha s^{-1+\alpha}}{\beta} - \frac{r}{(s-r)^2} = 0 \quad \text{and} \quad s > r. \tag{74}$$

This condition can be written as

$$\alpha s^{-1+\alpha}(s-r)^2 = r\beta \quad \text{and} \quad s > r. \tag{75}$$

When $\alpha = 2$, (75) becomes

$$2s(s-r)^2 = r\beta. \tag{76}$$

We can prove that this equation has only one root when $s > r$. Denote $f(s) = 2s(s-r)^2 - r\beta$. Take the derivative of $f(s)$, $f'(s) = 4s(s-r) + 2(s-r)^2$. This derivative function is positive when $s > r$. So $f(s)$ is an increasing function in the range $[r, \infty)$. In addition, $f(r) = -\beta r < 0$, and $f(s) \to \infty$ as $s \to \infty$. Thus, $f(s)$ has only one root in $[r, \infty)$.

Now we prove that z_{static} is only a function of γ when $\alpha = 2$. If $\alpha = 2$, $\gamma = r/\sqrt{\beta}$. By (73), we get

$$\begin{aligned} z_{static} &= \frac{\frac{r}{\sqrt{\beta}}}{\frac{s}{\sqrt{\beta}} - \frac{r}{\sqrt{\beta}}} + \left(\frac{s}{\sqrt{\beta}}\right)^2 \\ &= \frac{\gamma}{\tilde{s} - \gamma} + \tilde{s}^2, \end{aligned} \tag{77}$$

where $\tilde{s} = \frac{s}{\sqrt{\beta}}$.

In addition, (76) can be written as

$$\frac{2s}{\sqrt{\beta}} \left(\frac{s}{\sqrt{\beta}} - \frac{r}{\sqrt{\beta}}\right)^2 = \frac{r}{\sqrt{\beta}}. \tag{78}$$

Thus, $2\tilde{s}\left(\tilde{s} - \gamma\right)^2 = \gamma$. Clearly, z_{static} is only related to γ.

5.1.2 M/G/1-FIFO queue with static speed scaling

In this section, we examine the average energy-aware cost in the M/G/1-FIFO queue with static speed scaling.

From (29) in Section 4.1.2, we have $E[X] = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}$ for the M/G/1-FIFO queue. Thus, the average cost per unit time is

$$\begin{aligned} z_{static} &= \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s} + \frac{\lambda^2 E[S^2]}{2\left(1 - \frac{r}{s}\right)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s} + \frac{s\lambda^2 \frac{E[Y^2]}{s^2}}{2(s-r)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s} + \frac{\lambda^2 E[Y^2]}{2s(s-r)} + \frac{s^\alpha}{\beta}. \end{aligned} \tag{79}$$

In Section 5.1.1, we observed that the average cost in the M/G/1-PS queue is insensitive to the service time distribution. For the M/G/1-FIFO queue, however, the average cost is affected by the service time distribution.

Because $S = Y/s$, where s is constant, the service time S and the size of the job Y have the same distribution type. If Y is exponentially distributed, $Y \sim \text{Exp}(\frac{1}{m})$

with $E[Y] = m$ and $E[Y^2] = 2m^2$, then

$$\begin{aligned} z_{static} &= \frac{r}{s} + \frac{\lambda^2 m^2}{s(s-r)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{s(s-r)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s-r} + \frac{s^\alpha}{\beta}. \end{aligned} \tag{80}$$

We observe that the average cost for the M/M/1-FIFO queue in static system is the same as that for M/G/1-PS queue.

If Y is deterministic, $Y \sim \text{Det}(m)$, then, $E[Y] = m$ and $E[Y^2] = m^2$, and we have

$$\begin{aligned} z_{static} &= \frac{r}{s} + \frac{\lambda^2 m^2}{2s(s-r)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{2s(s-r)} + \frac{s^\alpha}{\beta} \\ &= \frac{r}{s-r} - \frac{r^2}{2s(s-r)} + \frac{s^\alpha}{\beta}. \end{aligned} \tag{81}$$

By comparing (80) with (81), we find out that the average cost for the M/M/1-FIFO queue in the static system is larger than that for the M/D/1-FIFO queue.

Similarly as in Section 5.1.1, we optimize (81) and get the following conditions for the optimal speed s:

$$\frac{r}{2}\left(\frac{1}{(s-r)^2} + \frac{1}{s^2}\right) = \frac{\alpha s^{-1+\alpha}}{\beta} \quad \text{and} \quad s > r. \tag{82}$$

When $\alpha = 2$, (82) can be written as

$$\frac{r}{2}\left(\frac{1}{(s-r)^2} + \frac{1}{s^2}\right) = \frac{2s}{\beta} \quad \text{and} \quad s > r. \tag{83}$$

By substituting (83) to (81), we can get the optimal average energy-aware cost for the M/D/1-FIFO queue.

5.1.3 M/G/1-PS queue with gated static speed scaling

After studying the average energy-aware cost in static systems, now we analyze the average cost in gated systems. For the gated system, the energy-aware cost is incurred only when the server is busy. According to (30) in Section 4.1.3, for the

M/G/1-PS queue, the probability that the system is busy is $P\{X > 0\} = \rho$. Thus, the average cost per unit time in the gated system can be written as

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \frac{s^\alpha}{\beta}\rho \\ &= \frac{\frac{r}{s}}{1-\frac{r}{s}} + \frac{s^\alpha}{\beta}\frac{r}{s} \\ &= \frac{r}{s-r} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{84}$$

Equation (84) shows that the average cost in the gated static system for the M/G/1-PS queue is also insensitive to the distribution of the service time.

Similarly as in Section 5.1.1, we take the derivative of (84). The optimal value of s clearly satisfies

$$\frac{r(-1+\alpha)s^{-2+\alpha}}{\beta} - \frac{r}{(s-r)^2} = 0 \quad \text{and} \quad s > r. \tag{85}$$

This can be expressed as

$$\beta = s^{\alpha-2}(-1+\alpha)(s-r)^2 \quad \text{and} \quad s > r. \tag{86}$$

When $\alpha = 2$, the equation becomes $s = r + \sqrt{\beta}$. We substitute it to (84), and obtain

$$\begin{aligned} z_{gated} &= \frac{r}{s-r} + r\frac{s}{\beta} \\ &= \frac{r}{\sqrt{\beta}} + r\frac{r+\sqrt{\beta}}{\beta} \\ &= 2\frac{r}{\sqrt{\beta}} + \frac{r^2}{\beta} \\ &= 2\gamma + \gamma^2 \\ &= \gamma(2+\gamma). \end{aligned} \tag{87}$$

Obviously, it is only a function of γ.

5.1.4 M/G/1-FIFO queue with gated static speed scaling

The previous section analyzed the average energy-aware cost in the gated static system for the M/G/1-PS queue. Here we examine the energy performance for the M/G/1-FIFO queue.

As already mentioned, for the M/G/1-FIFO queue, the mean queue length $E[X] =$

$\rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}$. Thus, we obtain

$$\begin{aligned} z_{gated} &= \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)} + \frac{s^\alpha}{\beta}\rho \\ &= \frac{r}{s} + \frac{s\lambda^2 \frac{E[Y^2]}{s^2}}{2(s-r)} + \frac{s^\alpha}{\beta}\frac{r}{s} \\ &= \frac{r}{s} + \frac{\lambda^2 E[Y^2]}{2s(s-r)} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{88}$$

Obviously, the average cost in the gated system for the M/G/1-FIFO queue also depends on the service time distribution.

For the M/M/1-FIFO queue where Y is exponentially distributed, $Y \sim \text{Exp}(\frac{1}{m})$, $E[Y] = m$, and $E[Y^2] = 2m^2$, we have

$$\begin{aligned} z_{gated} &= \frac{r}{s} + \frac{\lambda^2 m^2}{s(s-r)} + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{s(s-r)} + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{89}$$

We also observe that the average cost for the M/M/1-FIFO queue and M/G/1-PS queue in the gated system are the same.

If Y is deterministic with $Y \sim \text{Det}(m)$, $E[Y] = m$, and $E[Y^2] = m^2$, then

$$\begin{aligned} z_{gated} &= \frac{r}{s} + \frac{\lambda^2 m^2}{2s(s-r)} + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{2s(s-r)} + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} - \frac{r^2}{2s(s-r)} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{90}$$

Obviously, the average cost for the M/M/1-FIFO queue in the gated system is larger than that for the M/D/1-FIFO queue.

Again, we conclude that the average cost for the M/G/1-PS queue in the gated system is insensitive to the service time distribution, while for the M/G/1-FIFO queue, the service time distribution will affect the average cost.

By taking the derivative of (90) with respect to s, we get the following conditions for the optimal speed s:

$$\frac{r}{2}\left(\frac{1}{s^2}+\frac{1}{(s-r)^2}\right)=\frac{r(-1+\alpha)s^{\alpha-2}}{\beta} \quad \text{and} \quad s>r. \tag{91}$$

When $\alpha = 2$, the conditions can be written as

$$\frac{1}{s^2}+\frac{1}{(s-r)^2}=\frac{2}{\beta} \quad \text{and} \quad s>r. \tag{92}$$

By solving (92), we get

$$s=\frac{1}{2}\left(r+\sqrt{r^2+2\left(\beta+\sqrt{\beta(2r^2+\beta)}\right)}\right). \tag{93}$$

Again, we get the optimal average energy-aware cost for the M/D/1-FIFO queue by substituting (93) to (90).

5.1.5 Linear system

Since the server speed is a linear function of the queue length in the linear system, it can be viewed as an M/G/∞ model. According to (72), the average energy-aware cost is

$$z_{linear}=E[X]+\frac{E\left[(sX)^{\alpha}\right]}{\beta}$$

When $\alpha = 2$, the average cost per unit time can be written as

$$\begin{aligned} z_{linear} &= E[X]+\frac{E\left[(sX)^2\right]}{\beta} \\ &= E[X]+\frac{s^2E\left[X^2\right]}{\beta}. \end{aligned} \tag{94}$$

Note that here s refers to s_1, which represents the service speed of the system when there is only one job in the queue.

According to Section 4.1.4, the mean queue length in the M/G/∞ queue $E[X]=\rho$. The second moment of the queue length $E[X^2]=\rho^2+\rho$. Thus, we get

$$\begin{aligned} z_{linear} &= \rho+\frac{s^2(\rho+\rho^2)}{\beta} \\ &= \frac{r}{s}+\frac{s^2\left(\frac{r}{s}+\frac{r^2}{s^2}\right)}{\beta} \\ &= \frac{r}{s}+\frac{rs+r^2}{\beta} \end{aligned} \tag{95}$$

The optimal speed unit s satisfies $\frac{-r}{s^2}+\frac{r}{\beta}$=0, which implies that $s=\sqrt{\beta}$. Note that the optimal s is independent of r. In addition, we get

$$\begin{aligned} z_{linear} &= \frac{r^2}{\beta}+2\frac{r}{\sqrt{\beta}} \\ &= \gamma^2+2\gamma. \end{aligned} \tag{96}$$

We see that when $\alpha=2$, the optimal average cost in the linear system is the same as that in the gated system with the PS discipline.

5.2 Energy-aware schemes with switching delay

After examining the average energy-aware cost in different queueing models without switching delay, in this section, we will analyze the energy-aware cost with switching delay. The M/M/1 queue and the M/M/∞ queue with switching delay will be considered. Meanwhile, the average cost with different switching delay distributions will be calculated. Since in static system the service speed is always nonzero regardless of the traffic load of the system, there is no switching delay for the static system. We only consider gated and linear systems with switching delay.

Recall from Section 4.2.1 that $X(t)$ refers to the number of customers in the system at time t, and $Z(t)$ indicates whether the system is in the switching delay state or not at time t.

Similarly as in Section 5.1, we first give the expression of the average cost for the gated and static systems with switching delay.

For the gated system, we assume that the switching delay will not consume energy, thus

$$\begin{aligned} z_{gated} &= E[X]+E[1_{\{X>0,Z=0\}}\frac{s^\alpha}{\beta}] \\ &= E[X]+\frac{s^\alpha}{\beta}E[1_{\{X>0,Z=0\}}] \\ &= E[X]+\frac{s^\alpha}{\beta}P\{X>0,Z=0\}. \end{aligned} \tag{97}$$

For the linear system, we also assume that the switching delay does not consume energy. So, the average cost is

$$z_{linear}=E[X]+\frac{E\left[1_{\{X>0,Z=0\}}\left(sX\right)^\alpha\right]}{\beta}. \tag{98}$$

5.2.1 M/M/1 queue with gated speed scaling

When considering switching delay in the gated system, we also need to think about whether switching delay will cost energy or not. If switching delay does not cost energy, then the energy cost is incurred only during the time the sever is busy and the system is not in the switching delay state, i.e. $X > 0$ and $Z = 0$. According to Section 4.2.3, the system is consuming energy with probability

$$\begin{aligned} P\{X > 0, Z = 0\} &= \frac{E[B] - E[D]}{E[C]} \\ &= \frac{E[N]E[S]}{E[C]} \\ &= \frac{E[S]}{E[I]} \\ &= \lambda E[S] \\ &= \rho. \end{aligned} \tag{99}$$

In this thesis, we only consider the case that the switching delay does not consume energy.

For the gated M/M/1 queue with switching delay, we have

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \frac{\lambda E[D]}{1+\lambda E[D]} + \frac{\lambda^2 E[D^2]}{2(1+\lambda E[D])} + \frac{s^\alpha}{\beta}\rho \\ &= \frac{\rho}{1-\rho} + \frac{\lambda E[D]}{1+\lambda E[D]} + \frac{\lambda^2 E[D^2]}{2(1+\lambda E[D])} + \frac{s^\alpha}{\beta}\frac{r}{s} \\ &= \frac{\rho}{1-\rho} + \frac{\lambda E[D]}{1+\lambda E[D]} + \frac{\lambda^2 E[D^2]}{2(1+\lambda E[D])} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{100}$$

Since we assume switching delay does not consume energy, the impact of switching delay is only on the mean queue length. In the following, we analyze the impact of switching delay distribution on the average cost.

If $D \sim \text{Exp}(\delta)$ with $E[D] = \frac{1}{\delta} = d$, then, we have

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \lambda d + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} + \frac{rd}{m} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{101}$$

If $D \sim \text{Exp}(1/(ks))$ with $E[D] = ks$, then

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \lambda ks + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} + \frac{rks}{m} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{102}$$

If $D \sim \text{Det}(d)$ with $E[D] = d$, we have

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \frac{\lambda d}{1+\lambda d}\left(1 + \frac{\lambda d}{2}\right) + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} + \frac{rd}{m+rd}\left(1 + \frac{rd}{2m}\right) + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{103}$$

If $D \sim \text{Det}(ks)$ with $E[D] = ks$ and k is a constant, then

$$\begin{aligned} z_{gated} &= \frac{\rho}{1-\rho} + \frac{\lambda ks}{1+\lambda ks}\left(1+\frac{\lambda ks}{2}\right) + r\frac{s^{\alpha-1}}{\beta} \\ &= \frac{r}{s-r} + \frac{rks}{m+rks}\left(1+\frac{rks}{2m}\right) + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{104}$$

5.2.2 M/G/1-FIFO queue with gated speed scaling

If the service time is generally distributed, then the average energy-aware cost in the gated M/G/1-FIFO queue with switching delay is

$$\begin{aligned} z_{gated} &= \lambda\left(E[S] + \frac{\lambda E[S^2]}{2(1-\rho)} + \frac{E[D]}{1+\lambda E[D]} + \frac{\lambda E[D^2]}{2\left(1+\lambda E[D]\right)}\right) + \frac{s^\alpha}{\beta}P\{X>0, Z=0\} \\ &= \lambda\left(E[S] + \frac{\lambda E[S^2]}{2(1-\rho)} + \frac{E[D]}{1+\lambda E[D]} + \frac{\lambda E[D^2]}{2\left(1+\lambda E[D]\right)}\right) + \frac{s^\alpha}{\beta}\rho \\ &= \frac{r}{m}\left(\frac{E[Y]}{s} + \frac{rE[Y^2]}{2ms(s-r)} + \frac{mE[D]}{m+rE[D]} + \frac{rE[D^2]}{2(m+rE[D])}\right) + \frac{s^{\alpha-1}r}{\beta}. \end{aligned} \tag{105}$$

which is also affected by the distribution of the service time and the switching delay.

Note that for the switching delay distributed as $D \sim \text{Exp}(\delta)$ and $D \sim \text{Det}(d)$, the optimal s is the same as that for the system without switching delay(determined by $s = r + \sqrt{\beta}$).

If the service time is deterministic distributed, we have $E[Y] = m$, $E[Y^2] = m^2$. When $\alpha = 2$, the average cost for the M/D/1-FIFO queue is

$$z_{gated} = \frac{r}{m}\left(\frac{m}{s} + \frac{rm}{2s(s-r)} + \frac{mE[D]}{m+rE[D]} + \frac{rE[D^2]}{2(m+rE[D])}\right) + \frac{sr}{\beta}. \tag{106}$$

If $D \sim \text{Det}(d)$ with $E[D] = d$, we have

$$\begin{aligned} z_{gated} &= \frac{r}{m}\left(\frac{m}{s} + \frac{rm}{2s(s-r)} + \frac{md}{m+rd} + \frac{rd^2}{2(m+rd)}\right) + \frac{sr}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{2s(s-r)} + \frac{rd}{m+rd} + \frac{r^2d^2}{2m(m+rd)} + \frac{sr}{\beta} \\ &= \frac{r}{s-r} - \frac{r^2}{2s(s-r)} + \frac{rd}{m+rd} + \frac{r^2d^2}{2m(m+rd)} + \frac{sr}{\beta}. \end{aligned} \tag{107}$$

Obviously, the impact of switching delay and the service time are also separated. Thus, the optimal s satisfies (93) in Section 5.1.4.

If $D \sim \text{Det}(ks)$ with $E[D] = ks$, then

$$\begin{aligned} z_{gated} &= \frac{r}{m}\left(\frac{m}{s} + \frac{rm}{2s(s-r)} + \frac{mks}{m+rks} + \frac{rk^2s^2}{2(m+rks)}\right) + \frac{sr}{\beta} \\ &= \frac{r}{s} + \frac{r^2}{2s(s-r)} + \frac{rks}{m+rks} + \frac{r^2k^2s^2}{2m(m+rks)} + \frac{sr}{\beta} \\ &= \frac{r}{s-r} - \frac{r^2}{2s(s-r)} + \frac{rks}{m+rks} + \frac{r^2k^2s^2}{2m(m+rks)} + \frac{sr}{\beta}. \end{aligned} \tag{108}$$

For this case, the switching delay is dependent on the service time. Thus, the optimal service speed s is different from the previous one.

5.2.3 Linear system

In the linear system, we also only consider the case where the switching delay does not consume energy. Then, energy is consumed only during the time the server is busy and the system is not in the switching delay state.

Assume the switching delay is exponentially distributed. Then according to Section 4.2.2, we have

$$\pi_{i,1} = \left(\frac{\lambda}{\lambda+\delta}\right)^i \pi_{0,0}, \quad i = 1, 2, 3, ...,$$

$$\pi_{i,0} = \sum_{n=0}^{i-1} \frac{n!}{i!} \left(\frac{\lambda}{\mu}\right)^{i-n} \left(\frac{\lambda}{\lambda+\delta}\right)^n \pi_{0,0},$$

$$\pi_{0,0} = \left(1 + \frac{\lambda}{\delta} + \sum_{i=1}^{\infty}\sum_{n=0}^{i-1} \frac{n!}{i!} \left(\frac{\lambda}{\mu}\right)^{i-n} \left(\frac{\lambda}{\lambda+\delta}\right)^n\right)^{-1}.$$

The mean queue length of the linear system with switching delay is

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} i\left(\pi_{i,0} + \pi_{i,1}\right) \\ &= \frac{\sum_{i=1}^{\infty}\left(\sum_{n=0}^{i-1} \frac{n!}{(i-1)!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n + i\left(\frac{\lambda}{\lambda+\delta}\right)^i\right)}{1 + \frac{\lambda}{\delta} + \sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n}. \end{aligned} \tag{109}$$

For the convenience of numerical evaluation, we express the mean queue length in the form of the service speed and mean job size. Recall that $\mu = s/m$, thus

$$E[X] = \frac{\sum_{i=1}^{\infty}\left(\sum_{n=0}^{i-1} \frac{n!}{(i-1)!}\left(\frac{\lambda m}{s}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n + i\left(\frac{\lambda}{\lambda+\delta}\right)^i\right)}{1 + \frac{\lambda}{\delta} + \sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda m}{s}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n}. \tag{110}$$

Numerical evaluation

For numerical analysis, we assume $\alpha = 2$. Then, by (98) in Section 5.2, the average cost for the linear system

$$\begin{aligned} z_{linear} &= E[X] + \frac{E[1_{\{X>0,Z=0\}}(sX)^2]}{\beta} \\ &= E[X] + \frac{E[1_{\{X>0,Z=0\}}s^2X^2]}{\beta} \\ &= E[X] + \frac{s^2E[1_{\{X>0,Z=0\}}X^2]}{\beta}. \end{aligned}$$

The calculation of $E[1_{\{X>0,Z=0\}}X^2]$ is as follows,

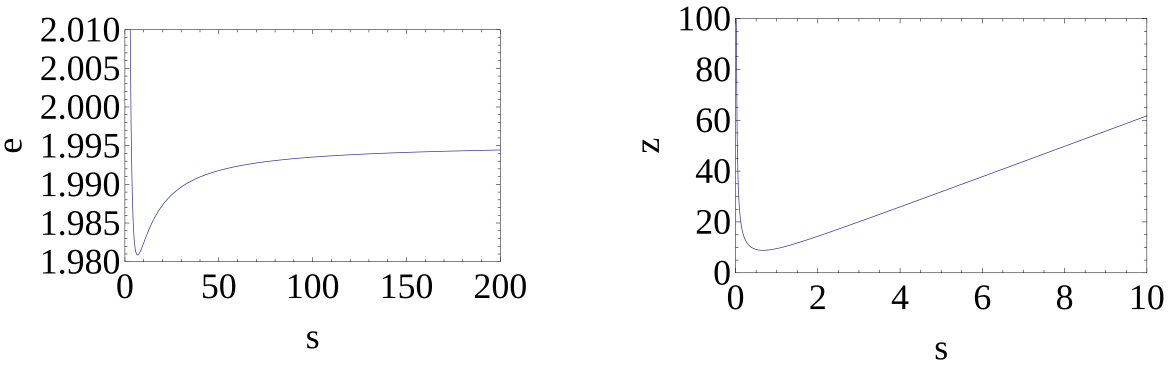
$$\begin{aligned} E[1_{\{X>0,Z=0\}}X^2] &= \sum_{i=1}^{\infty} i^2\pi_{i,0} \\ &= \sum_{i=1}^{\infty} \frac{i^2\left(\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda m}{s}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n\right)}{1+\frac{\lambda}{\delta}+\sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda m}{s}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n}. \end{aligned}$$

Since it is impossible to take the value of i to infinity for numerical evaluation, we define i_{max} as the maximum value that is suitable for the numerical evaluation. First, we consider the numerical evaluation of the mean queue length. According to the calculation of Mathematica, when i_{max} is 20 and 100, the numerical value of the mean queue length is practically the same for sufficiently large s. Thus, we take $i_{max} = 20$ for the numerical evaluation. In addition, let $\lambda = 2$, $\delta = 1$ and $m = 1$. From Figure 16a we see that the mean queue length first goes down as s increases, but when it reaches the minimum point, it begins to increase and asymptotically approaches a constant value λ/δ. This result corresponds to (47) in Section 4.2.2. Numerically, we find that when $s = 6.76851$, the minimum queue length equals 1.98087.

Now, we evaluate the mean energy-aware cost numerically. From Section 4.2.2, we can argue that

$$E[1_{\{X>0,Z=0\}}X^2] \to 0, \tag{111}$$

when s goes infinity. However, it seems that as s goes to infinity, the average cost $E[X] + \frac{s^2E[1_{\{X>0,Z=0\}}X^2]}{\beta}$ goes to infinity. By taking $i_{max} = 100$, $\lambda = 2$, $\delta = 1$, $m = 1$ and $\beta = 1$, we get Figure 16b, which demonstrates that the mean energy-aware cost in the linear system goes to infinity as s goes to infinity. The minimum average cost is 8.83805, when s is 0.669966.



(a) Mean queue length

(b) Average energy-aware cost

Figure 16: Numerical evaluation for the M/M/∞ queue.

6 Numerical results

In this section, we will give numerical results related to the analysis in Section 5. For convenience, we assume $\alpha = 2$.

6.1 M/G/1-PS queue without switching delay

Firstly, we study the M/G/1-PS queue for the three different speed scaling schemes without switching delay. As mentioned in Section 5.1.5, the optimal average energy-aware costs in the gated static system and linear system for the M/G/1-PS queue are the same. Thus, we only consider the static system and the gated static system here.

We first plot the average energy-aware cost per time unit, z, versus r with different β values, $\beta = 0.1, 1, 10$, which is shown in Figure 17. According to Figure 17, with the same r value, when β increases, which means the relative weight of the performance part increases, the average cost decreases. Obviously, the optimal average energy-aware cost in the static system is larger than that in the gated static system regardless of the traffic load r.

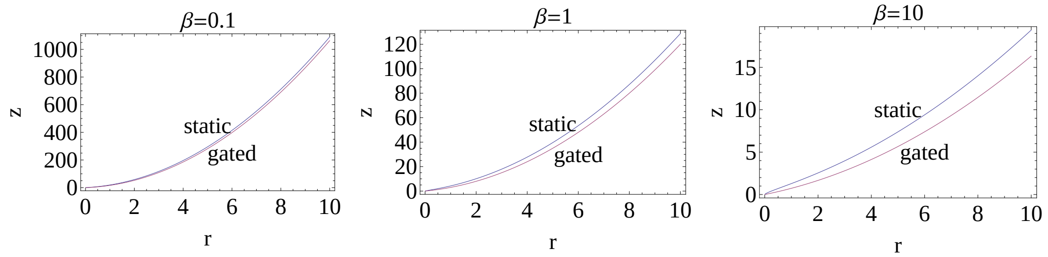


Figure 17: Average energy-aware cost per time unit z versus r for the M/G/1-PS queue.

Next, we plot the ratios $\frac{z_{static}}{r}$ and $\frac{z_{gated}}{r}$ as a function of r with β equal to 0.1,1,10, respectively. As z is the average cost per time unit, r is measured in cycles per time unit, thus, the ratio $\frac{z}{r}$ is the average cost per cycle. The result is presented in Figure 18.

From Figure 18, we can see that with the same β value, when the traffic load r increases, the ratio increases linearly in the gated system. This can be verified by applying (84), which implies that $\frac{z_{gated}}{r} = \frac{1}{s-r} + \frac{s}{\beta}$. Since the optimal value of s satisfies $s = r + \sqrt{\beta}$, and thus, $\frac{z_{gated}}{r} = \frac{2}{\sqrt{\beta}} + \frac{r}{\beta}$. This equation is clearly a linear function of r.

For the static system, there is a decrease at first, and then the ratio increases as r increases. Thus, there is a minimum optimal value r for the ratio z/r. We can also explain this phenomenon by utilizing (73). For any fixed s, we have $\frac{z_{static}}{r} = \frac{1}{s-r} + \frac{s^2}{r\beta}$. By taking roots for the derivative of $\frac{z_{static}}{r}$, we get two solutions: $r = \frac{s^2}{s\pm\sqrt{\beta}}$. Since

$s > r > 0$ for the optimal values of s, if $r = \frac{s^2}{s-\sqrt{\beta}}$, we have $0 < -\sqrt{\beta}$. This is impossible and thus there is only one root for the derivative of z_{static}/r from the range $[0, \infty)$. When $0 < r < \frac{s^2}{s+\sqrt{\beta}}$, the derivative of z_{static}/r is smaller than zero, and when $r > \frac{s^2}{s+\sqrt{\beta}}$, the derivative of z_{static}/r is greater than zero. As $r \to 0$, $z_{static}/r \to \infty$, and thus, the ratio z_{static}/r decreases when $0 < r < \frac{s^2}{s+\sqrt{\beta}}$, and it increases when $r > \frac{s^2}{s+\sqrt{\beta}}$.

In order to compare the gain of the average cost in the gated system, we plot the ratio $\frac{z_{gated}}{z_{static}}$ as a function of r, which is shown in Figure 19. As can be seen in Figure 19, the gated system performs better with light traffic load. As the traffic load increases, the ratio approaches 1 and the gain disappears. In addition, with larger β value, the relative weight of the performance part is higher, and the ratio $\frac{z_{gated}}{z_{static}}$ is smaller.

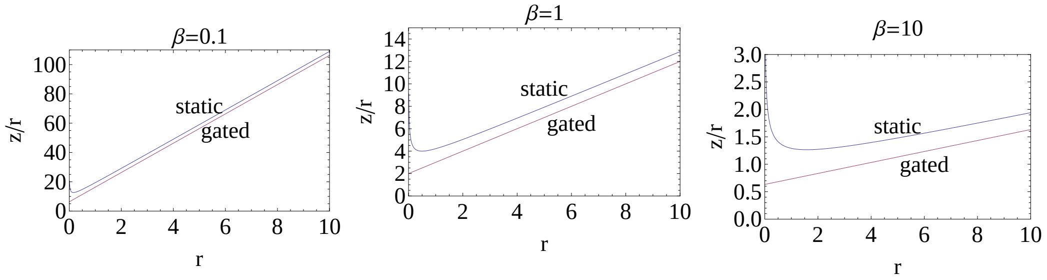


Figure 18: Ratio z/r versus r for the M/G/1-PS queue.

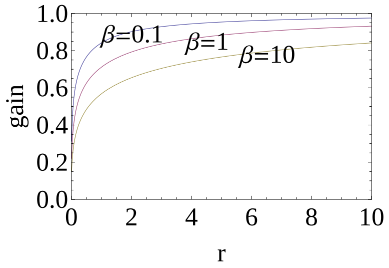


Figure 19: Gain z_{gated}/z_{static} versus r with different β values for the M/G/1-PS queue.

6.2 M/D/1-FIFO queue without switching delay

In Section 6.1, we studied the average energy-aware cost for the M/G/1-PS queue without switching delay. In this section, however, we will analyze the energy-aware performance for the M/D/1-FIFO queue without switching delay. Similarly as in Section 6.1, we plot the average cost z, the ratio z/r as well as the ratio z_{gated}/z_{static}

versus the traffic load r. The results are shown in Figures 20, 21 and 22, respectively.

By comparing the results with those in Section 6.1, we can find out that the average energy-aware cost for the M/D/1-FIFO queue without switching delay behaves similarly to that in the M/G/1-PS queue without switching delay.

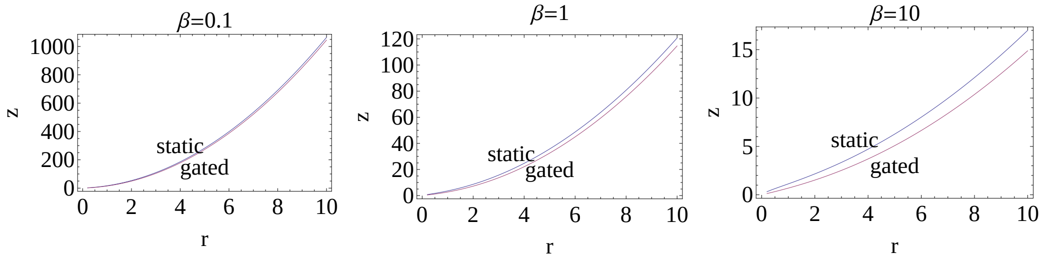


Figure 20: Average energy-aware cost per time unit z versus r for the M/D/1-FIFO queue.

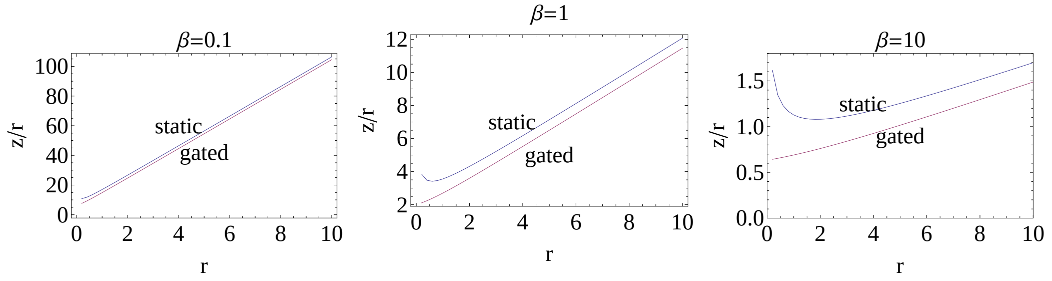


Figure 21: Ratio z/r versus r for the M/D/1-FIFO queue.

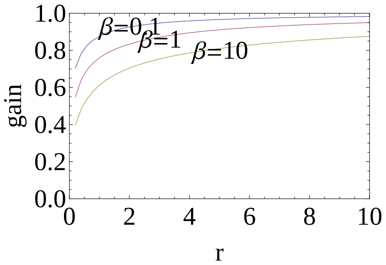


Figure 22: Gain z_{gated}/z_{static} versus r with different β values for the M/D/1-FIFO queue.

6.3 M/M/1 and M/M/∞ queues with switching delay

In the previous two sections, we analyzed the average energy-aware cost without switching delay. In this section, we will study the impact of switching delay on the average energy-aware cost. As mentioned in 5.2, there is no switching delay for the static system. Thus, we consider the ratio of the average cost of the gated system with switching delay and that of the static system without switching delay, as well as that for the linear system with switching delay and the static system without switching delay.

Here, we only study two different cases. The first case is that the switching delay is exponentially distributed with $D \sim Exp(\delta)$, where $\delta = \frac{1}{d}$ and d is independent of the service speed s. The second case is that the switching delay is exponentially distributed with $D \sim Exp(1/s)$. For the first case, we take two different d values, $d = 1, d = 10$, to compare the impact of switching delay. The results are shown in Figures 23 and 24.

As can be seen from Figure 23, with the same value of β, the ratios z_{gated}/z_{static} and z_{linear}/z_{static} are larger with switching delay $d = 10$ than that with $d = 1$. The performance of the linear system is always better than that for the gated system. With the same value of d, the ratios increase as β increases. For the system with switching delay $d = 10$, the ratios are always larger than 1, which means that the performance of the gated and linear system is not good with long switching delay regardless of the traffic load. In fact, the ratios approach 0 when the traffic load approaches 0. This is because when $r \to 0$, there are typically no customers to be served, no switching delay and the system is not using energy. However, the static system always runs at some fixed rate for any positive value of load. Thus, the ratio of the average costs for the gated and linear system approach 0 when $r \to 0$. When the traffic load is light, for $\beta = 0.1$ and $\beta = 1$, the ratios are less than 1, but when $\beta = 10$, the ratios are larger than 1 even when the traffic load is light.

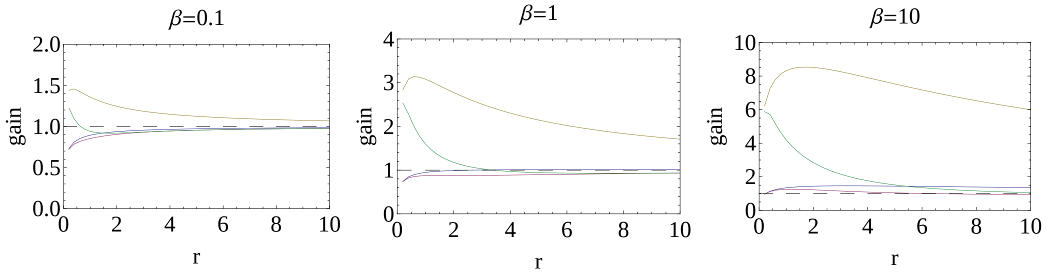


Figure 23: Gain z_{gated}/z_{static} and z_{linear}/z_{static} versus r with different β values for M/M/1 and M/M/∞ queues, $E[D] = d$. Yellow curve: z_{gated}/z_{static} for $d = 10$. Green curve: z_{linear}/z_{static} for $d = 10$. Blue curve: z_{gated}/z_{static} for $d = 1$. Red curve: z_{linear}/z_{static} for $d = 1$.

Finally we consider the case where the mean switching delay is positive and proportional to speed s, $E[D] = ks$, where we have assumed that $k = 1$. As can be observed from Figure 24, the performance of linear system is better than that in the gated system. With light traffic load, when β is small, the energy part has higher relative weight, the linear system performs better than the static system. With

$\beta = 10$, however, the performance of the linear system is worse than that of the static system. For the linear system, the ratio is approaching 1 as the traffic load becomes heavy. This is because when the traffic load is heavy, there is small chance that the system is switched to the idle state, thus, switching delay has little affect on the performance of the system.

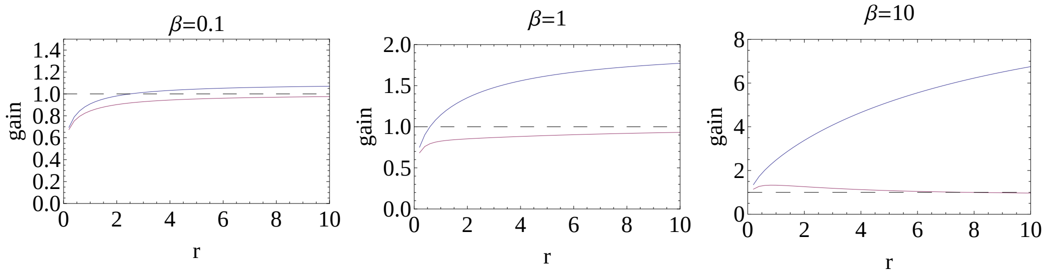


Figure 24: Gain z_{gated}/z_{static} and z_{linear}/z_{static} versus r with different β values for M/M/1 and M/M/∞ queues, $E[D] = s$. Blue curve: z_{gated}/z_{static}. Red curve: z_{linear}/z_{static}.

6.4 M/D/1-FIFO queue with switching delay

In the previous section, we studied the average cost for queueing systems with exponentially distributed service time. In this section, we will consider service time that has deterministic distribution. We consider two situations. For the first situation, the switching delay is independent of the service time, and for the second case, the switching delay is linear to the service time. Similarly as in Section 6.3, we plot the ratio z_{gated}/z_{static} to analyze the impact of switching delay on the average cost. The results are shown in Figures 25 and 26.

As can be observed from Figures 25 and 26, the average energy-aware cost for the M/D/1-FIFO queue with switching delay behaves similarly to that in the M/M/1 queue with switching delay, which is analyzed in detail in Section 6.3.

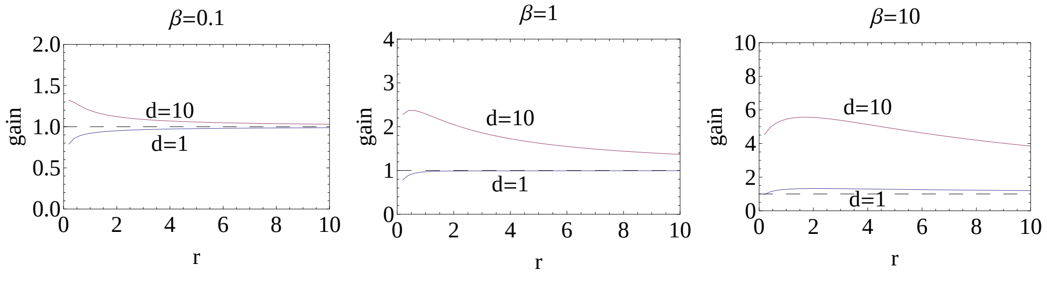


Figure 25: Gain z_{gated}/z_{static} versus r with different β values for the M/D/1-FIFO queue with switching delay, $E[D] = d$.

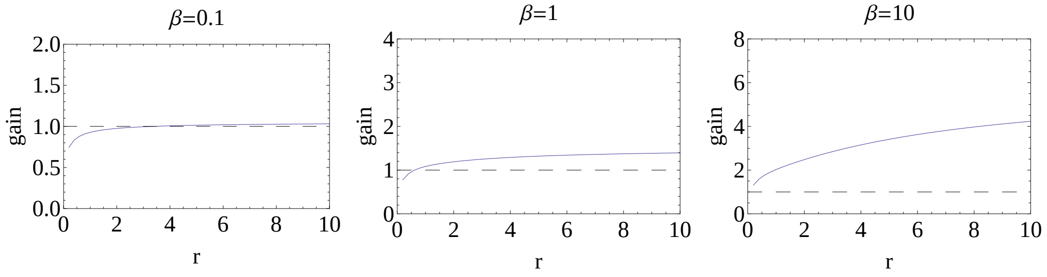


Figure 26: Gain z_{gated}/z_{static} versus r with different β values for the M/D/1-FIFO queue with switching delay, $E[D] = s$.

7 Conclusions and summary

In this thesis, we have analyzed the energy-aware performance of queueing systems. The focus has been on using queueing theory to model and analyze a single processor in a data center.

In order to find a trade-off between the performance and energy consumption of processors in data centers, we introduced three speed scaling methods, which adjust the processing speed of processors according to the traffic load of the queueing systems. They are static speed scaling, gated speed scaling and linear speed scaling. In gated and linear systems, switching delay is considered. The impact of switching delay on the queueing system is another focus in this thesis.

To analyze the trade-off, we introduced an energy-aware cost model which is a weighted sum of the mean queue length and the energy consumption of a queueing system. Based on this, we first analyze the queueing systems in a classical way, which does not consider the energy consumption and only study the system performance (which can be measured by the mean queue length). Both for the case without switching delay and with switching delay, we analyzed the M/M/1 queue, the M/G/1 queue and the M/M/∞ queue. Without switching delay, we found out that the performance of the system for the M/G/1-PS queue is the same as that for the M/M/1 queue. For the M/G/1-FIFO queue, however, the performance is sensitive to the service time distribution. When switching delay is taken into consideration, the situation is more complicated. The distribution of the switching delay also has impact on the system performance. We mainly considered exponential distribution and deterministic distribution.

After this, we analyzed the energy-aware performance of the queueing systems. Similarly, we considered the cases without switching delay and with switching delay. For queueing systems without switching delay, we gave an analysis of the static, gated and linear systems. For static and gated systems, we studied the energy-aware performance for the M/G/1-PS queue, the M/G/1-FIFO queue and the M/D/1-FIFO queue. For linear systems, the average energy-aware cost was analyzed for the M/M/∞ queue. For the case with switching delay, we only considered gated and linear systems. The reason is that in the static system, the processor always runs at an optimal service speed regardless of the traffic load. Thus, there is no switching delay in the static system. The M/M/1 queue and the M/M/∞ queue were the main study objects in this part.

In the numerical results section, we mainly studied the average energy-aware cost per unit time and average energy-aware cost per cycle versus the traffic load. For the case without switching delay, the focus is on analyzing the M/G/1-PS queue, the M/G/1-FIFO queue and the M/M/∞ queue. The results showed that without switching delay, the performance of the gated system and linear system are better than that for the static system. When parameter β increases, the relative weight of the performance part is higher, and the performance of the gated and linear system becomes better. For the case with switching delay, we focused on studying the M/M/1 queue, the M/D/1-FIFO queue and the M/M/∞ queue. Two different switching delay distributions are considered. One is independent of the mean service time and the other one is dependent on the mean service time. We concluded that parameter β has large impact on the average energy cost for different queueing systems. For the switching delay independent of the mean service time, we found out that with larger switching delay, the performance of the gated and linear system is worse than that with lower switching delay. With larger switching delay, as the traffic load becomes heavier, the performance improves. For the case where switching delay depends on the service time, the linear system performs better than the gated system. When β increases, the performance of both the systems become worse.

In summary, we concluded that without switching delay, the gated system and the linear system perform better than the static system, the average energy cost for the gated and the linear system are the same when parameter $\alpha = 2$. With switching delay, however, there is a large difference between them: the traffic load, the parameter β and the value of the switching delay all have impact on the overall performance of the queueing system. In addition, the distribution of the switching delay also affects the average energy-aware cost.

In the future, more work related to energy-aware performance of queueing systems can be done. We mainly analyzed three speed scaling methods in this thesis, thus, we can study other speed scaling methods in the future. Apart from this, this thesis focused on the FIFO and PS service disciplines, but we can conduct the similar analysis on some other service disciplines. In addition, we assumed that switching delay does not consume energy itself, so the case that switching delay consumes energy can be considered in the future.

References

[1] Data centres. `http://www.interxion.com/data-centres/`.

[2] Smart 2020: Enabling the low carbon economy in the information age. `http://www.smart2020.org/_assets/files/03_Smart2020Report_lo_res.pdf`, 2008.

[3] Power, pollution and the Internet. `http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html`, 2012.

[4] Susanne Albers and Hiroshi Fujiwara. Energy-efficient algorithms for flow time minimization. *ACM Trans. Algorithms*, 3(4), November 2007.

[5] Eitan Altman, Urtzi Ayesta, and Balakrishna Prabhu. Load balancing in processor sharing systems. In *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools*, ValueTools '08, pages 12:1–12:10, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[6] Lachlan L.H. Andrew, Minghong Lin, and Adam Wierman. Optimality, fairness, and robustness in speed scaling designs. *SIGMETRICS Perform. Eval. Rev.*, 38(1):37–48, June 2010.

[7] M. Andrews, S. Antonakopoulos, and L. Zhang. Energy-aware scheduling algorithms for network stability. In *INFOCOM, 2011 Proceedings IEEE*, pages 1359 –1367, april 2011.

[8] S. Asmussen. *Applied Probability and Queues.* Wiley, 1987.

[9] Nikhil Bansal, Ho-Leung Chan, Tak-Wah Lam, and Lap-Kei Lee. Scheduling for speed bounded processors. In *Proceedings of the 35th international colloquium on Automata, Languages and Programming, Part I*, ICALP '08, pages 409–420, Berlin, Heidelberg, 2008. Springer-Verlag.

[10] Nikhil Bansal, Ho-Leung Chan, and Kirk Pruhs. Speed scaling with an arbitrary power function. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 693–701, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[11] Nikhil Bansal, Kirk Pruhs, and Cliff Stein. Speed scaling for weighted flow time. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 805–813, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[12] Wolfgang Bischof. Analysis of M/G/1-queues with setup times and vacations under six different service disciplines. *Queueing Syst. Theory Appl.*, 39(4):265–301, December 2001.

[13] D.M. Brooks, P. Bose, S.E. Schuster, H. Jacobson, P.N. Kudva, A. Buyuktosunoglu, J. Wellman, V. Zyuban, M. Gupta, and P.W. Cook. Power-aware microarchitecture: design and modeling challenges for next-generation microprocessors. *Micro, IEEE*, 20(6):26 –44, nov/dec 2000.

[14] Ho-Leung Chan, Wun-Tat Chan, Tak-Wah Lam, Lap-Kei Lee, Kin-Sum Mak, and Prudence W. H. Wong. Energy efficient online deadline scheduling. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 795–804, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[15] Lijun Chen, Na Li, and Steven H. Low. On the interaction between load balancing and speed scaling. Information Theory and Applications Workshop, 2011.

[16] G. Fettweis and E. Zimmermann. ICT energy consumption-trends and challenges. In *The 11th International Symposium on Wireless Personal Multimedia Communications*, 2008.

[17] Anshul Gandhi, Varun Gupta, Mor Harchol-Balter, and Michael A. Kozuch. Optimality analysis of energy-performance trade-off for server farm management. *Perform. Eval.*, 67(11):1155–1171, November 2010.

[18] Anshul Gandhi, Mor Harchol-Balter, and Ivo Adan. Server farms with setup costs. *Perform. Eval.*, 67(11):1123–1138, November 2010.

[19] Anshul Gandhi, Mor Harchol-Balter, Rajarshi Das, and Charles Lefurgy. Optimal power allocation in server farms. In *Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, SIGMETRICS '09, pages 157–168, New York, NY, USA, 2009. ACM.

[20] Jennifer M. George and J. Michael Harrison. Dynamic control of a queue with adjustable service rate. *Oper. Res.*, 49(5):720–731, September 2001.

[21] R. Gonzalez and M. Horowitz. Energy dissipation in general purpose microprocessors. *Solid-State Circuits, IEEE Journal of*, 31(9):1277 –1284, sep 1996.

[22] Donald Gross, John F. Shortle, James M. Thompson, and Carl M. Harris. *Fundamentals of Queueing Theory.* Wiley-Interscience, New York, NY, USA, 4th edition, 2008.

[23] Maruti Gupta and Suresh Singh. Greening of the Internet. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '03, pages 19–26, New York, NY, USA, 2003. ACM.

[24] Stefanos Kaxiras and Margaret Martonosi. *Computer Architecture Techniques for Power-Efficiency.* Morgan and Claypool Publishers, 1st edition, 2008.

[25] F. P. Kelly. *Reversibility and Stochastic Networks.* Cambridge University Press, New York, NY, USA, 2011.

[26] Leonard Kleinrock. *Queueing Systems. Volume I: Theory.* Wiley-Interscience, 1975.

[27] Tak-Wah Lam, Lap-Kei Lee, Isaac K. To, and Prudence W. Wong. Speed scaling functions for flow time scheduling based on active job count. In *Proceedings of the 16th annual European symposium on Algorithms*, ESA '08, pages 647–659, Berlin, Heidelberg, 2008. Springer-Verlag.

[28] Hanoch Levy and Leonard Kleinrock. A queue with starter and a queue with vacations: delay analysis by decomposition. *Oper. Res.*, 34(3):426–436, June 1986.

[29] Minghong Lin, A. Wierman, L.L.H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *INFOCOM, 2011 Proceedings IEEE*, pages 1098 –1106, april 2011.

[30] A. Penttinen, E. Hyytiä, and S. Aalto. Energy-aware dispatching in parallel queues with on-off energy consumption. In *Performance Computing and Communications Conference (IPCCC), 2011 IEEE 30th International*, pages 1 –8, nov. 2011.

[31] M. Pickavet, W. Vereecken, S. Demeyer, P. Audenaert, B. Vermeulen, C. Develder, D. Colle, B. Dhoedt, and P. Demeester. Worldwide energy needs for ICT: The rise of power-aware networking. In *Advanced Networks and Telecommunication Systems, 2008. ANTS '08. 2nd International Symposium on*, pages 1 –3, dec. 2008.

[32] Kirk Pruhs. Competitive online scheduling for server systems. *SIGMETRICS Perform. Eval. Rev.*, 34(4):52–58, March 2007.

[33] Kirk Pruhs, Patchrawat Uthaisombut, and Gerhard Woeginger. Getting the best response for your erg. *ACM Trans. Algorithms*, 4(3):38:1–38:17, July 2008.

[34] S. M. Ross. *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco, 1970.

[35] Sheldon M. Ross. *Stochastic processes.* Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, New York, Chichester, 1983.

[36] Richard Serfozo. *Basics of Applied Stochastic Processes.* Springer, 2008.

[37] M. Feuillet T. Bonald. *Network performance analysis.* Wiley, 2011.

[38] Peter D. Welch. On a generalized M/G/1 queuing process in which the first customer of each busy period receives exceptional service. *Operations Research*, 12(5):736–752, 1964.

[39] A. Wierman, L.L.H. Andrew, and Ao Tang. Power-aware speed scaling in processor sharing systems. In *INFOCOM 2009, IEEE*, pages 2007 –2015, april 2009.

[40] J. A. Williams, N. W. Bergmann, and X. Xie. Fifo communication models in operating systems for reconfigurable computing. In *Proceedings of the 13th Annual IEEE Symposium on Field-Programmable Custom Computing Machines*, FCCM '05, pages 277–278, Washington, DC, USA, 2005. IEEE Computer Society.

[41] F. Yao, A. Demers, and S. Shenker. A scheduling model for reduced CPU energy. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 374 –382, oct 1995.