Performance-Energy Trade-Off in Data Centers: Impact of Switching Delay

Xiaohua Lu
School of Electrical Engineering
Aalto University, Finland
Email: xiaohua.lu@aalto.fi

Samuli Aalto
School of Electrical Engineering
Aalto University, Finland
Email: samuli.aalto@aalto.fi

Pasi Lassila
School of Electrical Engineering
Aalto University, Finland
Email: pasi.lassila@aalto.fi

***Abstract*—We develop simple queuing models for a single node in a server farm and analytically study the impact of switching delay on the performance-energy trade-off. The objective is to compare how an optimized static speed scaling scheme performs against two (gated and linear) optimized dynamic speed scaling schemes, where the processor can be switched off when it is idle but the penalty is the switching delay (time to wake up the processor from the off state). In the gated scheme, the processor speed is zero when the server is idle and constant otherwise, and in the linear scheme the processing speed scales linearly with the number of jobs. Our results demonstrate that the switching delay can have a considerable impact on the optimal trade-off. The linear scheme is always better than the gated scheme and, when the switching delays are long, even the static scheme can be better. In practice, the trade-off is affected highly by the parameters and our models allow an explicit evaluation of the trade-off.**

I. Introduction

Data centers, made up of a large number of parallel servers, consume huge amounts of energy. Thus, their energy management has become a key issue [1], and a fundamental trade-off must be made between performance and energy.

Speed scaling [2]–[4] is a well-known technique that allows the server to slow down the service speed when the number of customers is low, and thus the server is running at a lower power and also saving energy. Based on the seminal work by George and Harrison [5], Wierman et al. [3] studied how to optimally scale speed to balance mean response time and mean energy consumption in a single-server queue. Surprisingly, they found that a simple gated scheme which switches the server off when the system is idle (thus consuming no energy) and uses an optimal static service rate otherwise provides almost as good results as the optimal dynamic speed scaling.

However, when the server is switched off, it typically takes time to wake it up again, which is called the setup or switching delay. Our purpose in this paper is to explore how switching delay affects the performance-energy trade-off in a single-server queue which allows speed scaling. As in [3] (and many other papers), this trade-off is characterized by an appropriately weighted sum of the mean number of jobs and the mean power consumption. Our baseline is a static scheme with an optimized fixed service rate (thus avoiding any switching delays), and we compare it with two dynamic speed-scaling schemes that consume less energy in certain states but, on the other hand, suffer from the switching delay. One is the gated scheme that switches the server off when the system is idle and uses an optimized fixed service rate otherwise (which is different from the service rate of the optimal static system), and the other is a more dynamic linear speed scaling scheme where the service rate is proportional to the number of jobs in the system. Even in the latter system we optimize the linear step so that we use the optimal linear scheme for comparison.

Our main observations are as follows: (i) when switching delay is taken into account, the simple gated scheme can typically be improved considerably by introducing (fully) dynamic linear speed scaling; (ii) on the other hand, for long switching delays, the optimal performance-energy trade-off is not achieved by dynamic speed scaling but with the baseline static scheme with a fixed optimized service rate. Interestingly, it also turns out that the optimal performance-energy trade-off of the optimized linear scheme that relies on explicit knowledge on traffic parameters can be matched quite well by a robust linear scheme, which does not require any a priori information about the traffic. In a practical setting, the parameters affect the trade-off considerably and our models allow an explicit quantitative analysis of the impact of the switching delay on the performance-energy trade-off.

A single-server queue with switching delay was originally analyzed by Welch [6], and later considered, e.g., in [7], [8]. Gandhi et al. [9] consider a multiserver queue with switching delay. However, their model (restricted to the single-server case) differs from ours, since they just compare a static system with a fixed service rate μ and the corresponding gated system that uses the same (and not the optimal) service rate μ when switched on.

The rest of the paper is organized as follows. In Section II, we give the system model, introduce the three speed scaling schemes, and specify the objective reflecting the trade-off between performance and energy. The performance analysis of the three schemes is presented in Section III, while Section IV specifies corresponding optimization problems to determine the optimal schemes. Based on numerical experiments, the three schemes are compared in Section V. Section VI concludes the paper.

II. System model and objectives

We model a single server (or computing node) in a server farm. The server processes jobs that arrive according to a Poisson process with rate λ. The arriving jobs have an intrinsic random size, denoted by Y, which is assumed to obey an exponential distribution with mean $\mathrm{E}[Y] = m$. Let $X(t)$ denote the number of jobs in the system at time t.

In order to optimize the energy usage, the server supports speed scaling, i.e., it may be operating at a lower rate when the number of jobs is small, and the rate can be increased as more jobs enter the system. Specifically, we consider three approaches. The baseline is the *static speed scaling*, where the server runs at a constant rate s. In the *gated speed scaling*, the server is switched off whenever the system becomes idle, otherwise the processor is running at a fixed constant rate s. This is a very primitive form of speed scaling that only takes into account the idle state of the system. Finally, a more advanced form of speed scaling is *linear speed scaling*, where the processor is running at a rate that is linear in the number of jobs, i.e., when there are n jobs in the system the server is running at rate $s_n = ns$, where s is the service rate of a single job. In all schemes, the speed parameter s can be optimized for a given cost function. Let $\mu = s/m$, which for the static and gated schemes is also equivalent to the job departure rate. For the linear scheme, when there are n jobs the job departure rate is correspondingly $n\mu$. Also, it is convenient to denote $r = \lambda m$ as the offered traffic.

In the gated and linear schemes, whenever the system becomes idle, i.e., $X(t) = 0$, the server goes into sleep state, where we assume that it is not consuming any energy. However, when the idle period ends with the arrival of a new job, the system cannot become fully operational immediately, but there will be a random delay D before the server can start processing the jobs. This additional delay for the first job in a busy period is what we refer to as the *switching delay*. Note that in the static scheme, there is no switching delay as we assume that the server is continuously operating at rate s. For tractability of our analytical models, we assume that also D is exponentially distributed with mean $\mathrm{E}[D] = 1/\delta$. Let the process $Z(t)$ track the switching delay state of the system, i.e., $Z(t) = 1$ if the system is in the switching delay state and $Z(t) = 0$ when it is operating normally.

The process $(X(t), Z(t))$ is clearly a Markov process. Thus, in the static and gated schemes the performance of the system is the same for any work-conserving discipline, such as FIFO or processor sharing. In the linear scheme, the system appears like an infinite server model (each job has its own server). In the static and gated schemes, the system is stable whenever $r < s$. The linear system is stable for any $r > 0$.

The objective in our paper is to analyze the performance-energy trade-off taking into account also the impact of the switching delay. Our cost model is the same as in [3]. Thus, when there are n jobs in the system and the server is operating at rate s, costs are accumulating at the rate $n + s^\alpha/\beta$ per time unit, where $\alpha > 1$ and β is the (adjustable) weight parameter converting power units to time units (called delay-aversion in [3]). The model assumes that the power used by the server is proportional to s^α when it is running at rate s. According to [3], the dynamic power use of real chips is well modeled by this kind of a polynomial model.

In our analysis, our objective function z is the *average energy-aware cost per unit time*, i.e.,

$$z = \mathrm{E}[X] + \mathrm{E}[S^\alpha]/\beta, \tag{1}$$

where S denotes the service rate which is random in the gated and linear schemes depending on the state of the system. Due to Little's result, the objective function (1) characterizes the trade-off between delay and energy in our system.[1] For a given speed scaling scheme, (1) depends on our speed parameter s and to achieve an optimal performance-energy trade-off (1) can be minimized with respect to s for any scheme separately.

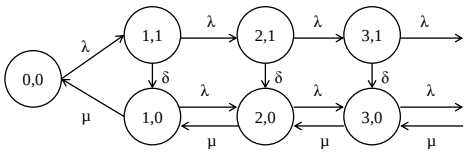

Fig. 1. State transition diagram for static and gated schemes.

III. Performance analysis

In this section, we present the Markov models that characterize the system which are needed in order to quantify the performance-energy trade-off (1). As mentioned above, the process $(X(t), Z(t))$, representing the joint state of the queue length and the switching delay, is a Markov process. However, the Markov processes corresponding to the static, gated and linear speed scaling schemes have different properties.

A. Static speed scaling

In the static speed scaling, the switching delay does not affect the system as the processor is continuously operating at rate s. Under our assumptions the system then simply corresponds to the M/M/1 queue, for which the mean queue length is simply

$$\mathrm{E}[X] = \frac{\lambda}{\mu - \lambda} = \frac{r}{s - r}. \tag{2}$$

Static speed scaling represents the baseline approach for us, where the switching delay does not incur any additional delay cost, but for which the energy cost may still be large as the system can not benefit from switching off the processor.

B. Gated speed scaling

As discussed earlier, in the gated scheme, the server is operating at rate s whenever there are jobs in the system. Thus, in our Markov model the service rate is simply $\mu = s/m$. This gives rise to the two-dimensional Markov process as shown in Figure 1. During the time that the system is empty, no energy is consumed. However, when the idle period ends with a first arrival (occurring with rate λ), this leads the process to the upper part of the state space, where $Z(t) = 1$. During the switching delay, there can be more arrivals, each increasing the queue length. After the switching delay phase ends, which happens at rate δ from any state where $Z(t) = 1$, the system resumes normal operation where the jobs are processed at rate μ until the system eventually returns back to idle state. At some point, a new busy period begins with a new switching delay phase.

[1] Dividing (1) by λ gives us a weighted combination of the mean delay and the mean energy per job.

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

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

Correspondingly, for the states (i,1), $i = 1, 2, ...$, we have

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

By solving the above, we obtain

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

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 yields

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

It turns out that the equilibrium distribution can be solved explicitly. First we obtain

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

and using this, we can express $\pi_{i,0}$ and $\pi_{i,1}$ as

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

for $i = 1, ...$ With some algebraic manipulations, the mean queue length $\mathrm{E}[X]$ has an explicit solution that equals

$$\mathrm{E}[X] = \frac{\lambda}{\mu-\lambda} + \frac{\lambda}{\delta} = \frac{r}{s-r} + \frac{r}{m\delta}. \tag{4}$$

From the result above, we observe that the mean queue length of the system consists of two parts, $\lambda/(\mu-\lambda)$ and λ/δ. The first part, $\lambda/(\mu-\lambda)$, equals the mean queue length of the M/M/1 system without switching delay. The second part, λ/δ, is the mean number of arrivals during the switching delay. Thus, the impacts of the service time and the switching delay on the mean queue length are separated (which is a well-known result, see, e.g., [7]).

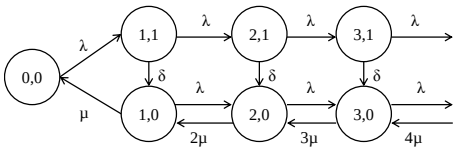

Fig. 2. State transition diagram for the linear speed scaling system.

C. Linear speed scaling

Next we consider the system with linear speed scaling. The behavior of the system is similar to that of the gated system in the previous section. The only difference is that when the system is not in the switching delay state, i.e., $Z(t) = 0$, due to the linear speed scaling scheme the departure rate given that there are n jobs in the system equals $n\mu$. This is illustrated in Figure 2. Thus, the system resembles an M/M/∞ queue.

To solve the equilibrium distribution we again apply the global balance equations. One can readily observe, compare Figures 1 and 2, that the balance equations for the states $(0,0)$ and $(i,1)$, $i = 1, \ldots,$ are identical to the ones of the gated system in the previous section. Thus, also the steady state probabilities $\pi_{i,1}$, $i = 1, \ldots$, have the same form as (3),

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

However, for state $(1,0)$, the global balance equation is

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

and for states (i,0), we have

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

Also, recall that the global balance equation for state $(0,0)$ yields $\pi_{1,0} = (\lambda/\mu)\pi_{0,0}$.

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

$$\begin{aligned} \pi_{i,0} &= \frac{\lambda^{i-1}\left(\sum_{n=0}^{i-1}(\lambda+\delta)^{i-1-n}\mu^n n!\right)}{i!\,(\lambda+\delta)^{i-1}\mu^{i-1}}\pi_{1,0} \\ &= \sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n \pi_{0,0}. \end{aligned} \tag{5}$$

The normalization condition can be expressed as

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

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

$$\pi_{0,0} = \left(1 + \frac{\lambda}{\delta} + \sum_{i=1}^{\infty}\sum_{n=0}^{i-1}\frac{n!}{i!}\left(\frac{\lambda}{\mu}\right)^{i-n}\left(\frac{\lambda}{\lambda+\delta}\right)^n\right)^{-1}. \tag{6}$$

Even though the equilibrium distribution has an explicit form, the mean queue length $\mathrm{E}[X]$ can only be numerically evaluated by truncation from

$$\mathrm{E}[X] = \sum_{i=0}^{\infty} i \cdot (\pi_{i,0} + \pi_{i,1}). \tag{7}$$

IV. Energy-aware delay optimization

In this section, we derive the optimization problems corresponding to the three different speed scaling schemes. In each case, to optimize the performance-energy trade-off (1) we have a separate non-linear optimization problem over our speed parameter s.

A. Static speed scaling

In the static scheme, the mean number of jobs is simply $\mathrm{E}[X] = r/(s-r)$, see (2) and the average energy-aware cost per unit time in the static system is

$$z_{\text{static}}(s) = \mathrm{E}[X] + \frac{s^\alpha}{\beta} = \frac{r}{s-r} + \frac{s^\alpha}{\beta}. \tag{8}$$

The equation (8) can be minimized with respect to s. It is easy to see that the function has a unique minimum when $s > r$. However, the optimization can not be done explicitly and needs to solved numerically. The optimal value of (8) is denoted by z^*_{static}.

B. Gated speed scaling

In the gated scheme, the system consumes energy only when there are jobs in the system and the system is not in the switching delay state.[2] Thus, the objective function in the gated scheme is given by

$$z_{gated}(s) = \mathrm{E}[X] + \frac{s^\alpha}{\beta}\mathrm{P}\{X > 0, Z = 0\}.$$

The probability $\mathrm{P}\{X > 0, Z = 0\}$ can be readily obtained from the explicit expressions in Section III-B for the equilibrium distribution

$$\mathrm{P}\{X > 0, Z = 0\} = 1 - \sum_{i=0}^{\infty} \pi_{i,1} = \lambda/\mu.$$

Combining this with (4), the objective function becomes

$$\begin{aligned} z_{\text{gated}}(s) &= \frac{\lambda}{\mu-\lambda} + \frac{\lambda}{\delta} + \frac{\lambda}{\mu}\frac{s^\alpha}{\beta} \\ &= \frac{r}{s-r} + \frac{r}{m\delta} + r\frac{s^{\alpha-1}}{\beta}. \end{aligned} \tag{9}$$

Again the mean energy-aware cost rate (9) has a unique minimum with respect to s when $s > r$. In the case where $\alpha = 2$, the solution can be obtained explicitly. The optimal value is achieved with $s = r + \sqrt{\beta}$ and the optimal value of the performance-energy trade-off (9) is

$$z^*_{\text{gated}} = \frac{r^2}{\beta} + \frac{2r}{\sqrt{\beta}} + \frac{r}{m\delta}.$$

We note that the optimal speed $s = r + \sqrt{\beta}$ is the same as in the gated system without switching delay (see, e.g., [3]), which is clearly due to the separation property discussed in Section III-B. Also, this optimal solution for the speed in the gated scheme is always greater than the corresponding optimal speed in the static system.

[2] This is an optimistic assumption from the gated system point of view. The other extreme is to assume that the energy consumption is at its maximum in the switching delay phase, see, e.g., [9].

C. Linear speed scaling

As mentioned earlier, in the linear system, the server speed is proportional to the queue length, which means $s_n = ns$ when there are n jobs in the system. Similarly to the gated system, no energy is consumed when the system is idle or in the switching delay state. Thus, the mean energy-aware cost per unit time can be written as

$$z_{\text{linear}}(s) = \mathrm{E}[X] + \frac{s^\alpha \mathrm{E}[1_{\{X>0,Z=0\}}X^\alpha]}{\beta}, \tag{10}$$

where $\mathrm{E}[X]$ is given by (7) and $\mathrm{E}[1_{\{X>0,Z=0\}}X^\alpha]$ is obtained from

$$\mathrm{E}[1_{\{X>0,Z=0\}}X^\alpha] = \sum_{i=1}^{\infty} i^\alpha \pi_{i,0}.$$

The quantities $\mathrm{E}[X]$ and $\mathrm{E}[1_{\{X>0,Z=0\}}X^\alpha]$ are functions of the speed parameter s and again the energy-aware cost rate (10) can be optimized with respect to it. Although not perhaps easy to show rigorously, it is intuitively clear that also in the linear system there is a well-defined unique minimum for s (since $\mathrm{E}[X]$ is monotonously decreasing with s, $\mathrm{E}[X] = \infty$ when $s = 0$ and $s^\alpha \mathrm{E}[1_{\{X>0,Z=0\}}X^\alpha]$ is monotonously increasing). The minimization can be performed numerically to obtain the optimal value z^*_{linear} for (10). In the numerical results, we refer to the above as the *optimized linear scheme*.

Finally we note that, as given in [3], without any switching delay and assuming $\alpha = 2$, the optimal value for s in the linear system is $s = \sqrt{\beta}$ resulting in the mean energy-aware cost rate

$$\frac{r^2}{\beta} + \frac{2r}{\sqrt{\beta}},$$

which is exactly the same as that of the optimal gated system. Thus, without switching delay and for $\alpha = 2$, the optimal linear system and the optimal gated system perform equally. In this case, as noted in Section I, even the optimal dynamic speed scaling does not perform much better.

However, even in the case with switching delay, setting $s = \sqrt{\beta}$ may perform reasonably. This has the advantage that the scheme does not assume any a priori knowledge of the traffic statistics and is robust in this sense. In the numerical examples, we also consider such a scheme and we refer to it as the *robust linear scheme*.

D. Generalized model for the switching delay

In the above formulations, the switching delay was considered as an independent parameter from the other parameters of the system. However, it is also possible in our modeling approach to generalize this so that the mean switching delay is still exponentially distributed but the mean value can depend in an arbitrary way on the speed of the processor s. In the optimization formulations for the gated (9) and linear (10) systems, this simply means that the parameter δ is replaced by a function depending on s, say $\delta(s)$. For example, it can be reasonable to consider a situation where the mean switching delay $1/\delta(s)$ is linearly proportional to the processing speed s, i.e., $\delta(s) = c/s$, where c is an appropriate scaling constant.

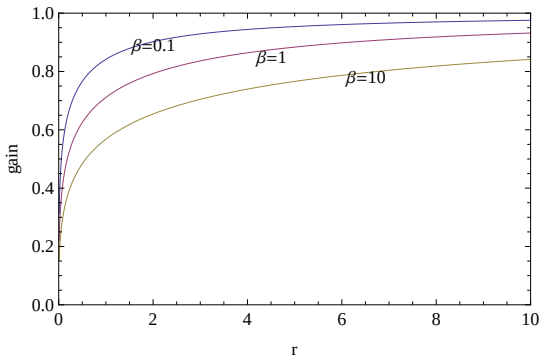

Fig. 3. Gain from the dynamic speed scaling schemes for different delay-aversions β in a system without any switching delay.

V. Numerical results

We consider two ways to model the effect of the switching delay. In the first case, we assume that the mean switching delay is a completely external parameter, independent of everything. For the other case, the mean switching delay is assumed to be linearly proportional to the server rate parameter s, which models that the higher the processing rate of the server the longer the switching delay. For all our numerical studies, we assume that $m = 1$ and $\alpha = 2$. As in [3], $\alpha = 2$ is used to provide insight and allows us to compare the results.

However, to make a comparison with earlier results, we start with the system without any switching delay. The results are illustrated in Figure 3, where we have plotted the ratio $z^*_{\text{gated}}/z^*_{\text{static}}$ between the mean energy-aware costs in the optimal static and the optimal gated system as a function of the offered load r for different delay-aversions β. As mentioned in the previous section, there is no difference in the mean energy-aware costs between the optimal gated scheme and the optimal linear scheme in this case. Thus, the curves in the figure represent, as well, the ratio $z^*_{\text{linear}}/z^*_{\text{static}}$. As can be seen from the figure, dynamic speed scaling reduces the mean energy-aware cost most with light load, but with heavy load, when the system is idle only rarely, the gain disappears as the ratio approaches 1. On the other hand, we see that the gain is improving with higher values of delay-aversion β, i.e., when the performance part has higher relative weight. Still, with any traffic load and for any β, the dynamic speed scaling schemes perform consistently better than the optimal static scheme when there is no switching delay.

Next we consider the case where the mean switching delay is positive and independent of everything else. The results for two different mean switching delay values, $\mathrm{E}[D] = 1/\delta = 1, 10$, are illustrated in Figure 4, where we have plotted the ratios $z^*_{\text{gated}}/z^*_{\text{static}}$ and $z^*_{\text{linear}}/z^*_{\text{static}}$, which are now separate, as a function of the offered load r for different delay-aversions β. As can be seen from the figure, the ratios $z^*_{\text{gated}}/z^*_{\text{static}}$ and $z^*_{\text{linear}}/z^*_{\text{static}}$ are larger with switching delay $d = 10$ than that with $d = 1$. The performance of the linear scheme is always better than that for the gated scheme. With the same value of d, the ratios increase as the delay-aversion β increases. For the system with switching delay $d = 10$, the ratios are almost always larger than 1, which means that the performance of dynamic speed scaling schemes is not good with long switching delay regardless of the traffic load. When $d = 1$

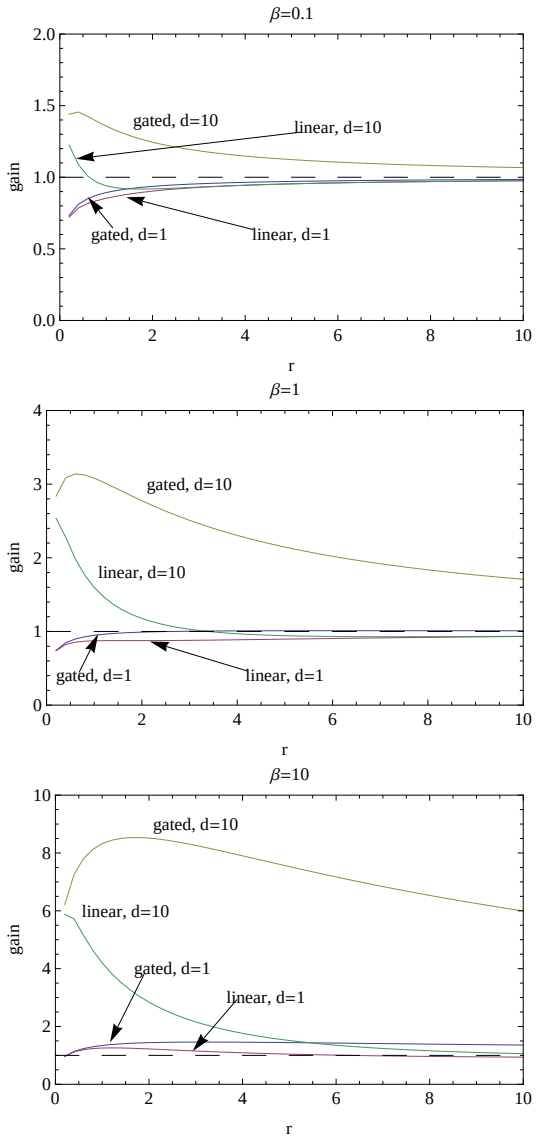

Fig. 4. Gain from the dynamic speed scaling schemes for different delay-aversions β in a system with an independent mean switching delay $\mathrm{E}[D] = d$, where $d = 1, 10$. Yellow curve: $z^*_{\text{gated}}/z^*_{\text{static}}$ for $d = 10$. Green curve: $z^*_{\text{linear}}/z^*_{\text{static}}$ for $d = 10$. Blue curve: $z^*_{\text{gated}}/z^*_{\text{static}}$ for $d = 1$. Red curve: $z^*_{\text{linear}}/z^*_{\text{static}}$ for $d = 1$.

and $\beta = 0.1$ the dynamic schemes are consistently better than the static scheme. However, for $\beta = 1$, when the traffic load is light, the ratios are less than 1 (dynamic schemes are better), but when $\beta = 10$, the ratios are larger than 1 even when the traffic load is light (static scheme is better). The optimal speed of the static scheme, s^*_{static} satisfies $r < s^*_{\text{static}} < r + \sqrt{\beta}$. Thus, as r grows the ratio of the speed in the optimized gated and static schemes approaches 1, implying that the ratio of the costs $z^*_{\text{gated}}/z^*_{\text{static}}$ approaches 1. However, for the linear scheme this may not hold.

Finally we consider the case where the mean switching delay is positive and proportional to speed s, $E[D] = s/c$, where we have assumed that $c = 1$. In this case, we also experiment with the robust linear scheme as introduced in Section IV.C and the corresponding performance of the scheme

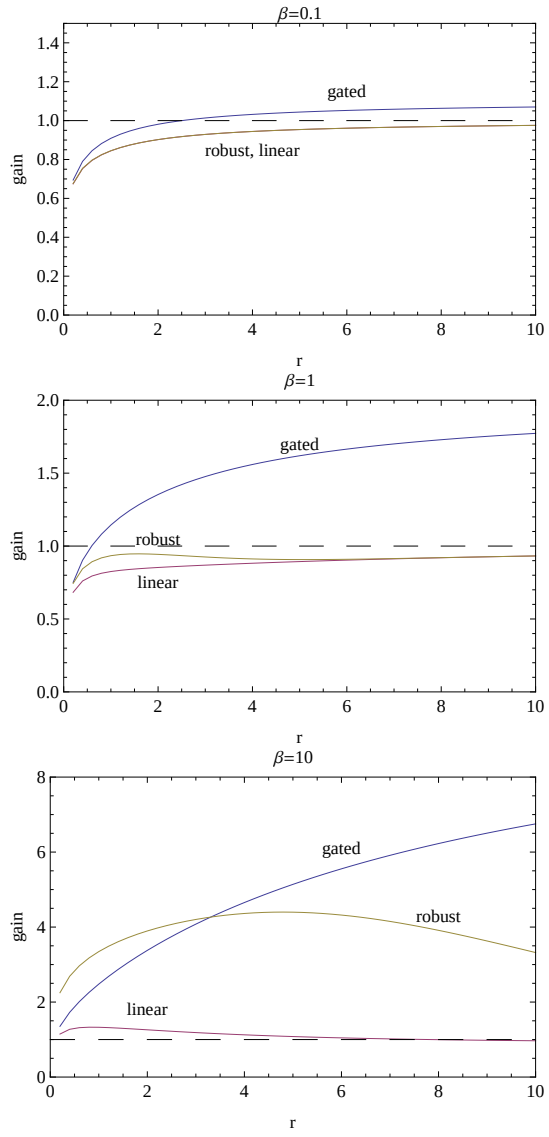

Fig. 5. Gain with dependent switching delay. Gain from the dynamic speed scaling schemes for different delay-aversions β in a system with a dependent mean switching delay $\mathrm{E}[D] = c/s$, where $c = 1$. Blue curve: $z^*_{\text{gated}}/z^*_{\text{static}}$. Yellow curve: $z^*_{\text{robust}}/z^*_{\text{static}}$. Red curve: $z^*_{\text{linear}}/z^*_{\text{static}}$.

is denoted by z^*_{robust}. The results are illustrated in Figure 5, where we have plotted the ratios $z^*_{\text{gated}}/z^*_{\text{static}}$, $z^*_{\text{robust}}/z^*_{\text{static}}$ $z^*_{\text{linear}}/z^*_{\text{static}}$ as a function of the offered load r again for different delay-aversions β. From the figure, we observe that the performance of the (optimized) linear scheme is better than that in the gated scheme. With light traffic load, when the delay-aversion β is small, i.e., the energy part has higher relative weight, the dynamic speed scaling schemes perform better than the static scheme. With $\beta = 10$, however, the gain disappears, but the optimized linear scheme is still not far from the static one. The robust scheme is also interestingly performing nearly as good as the optimized linear scheme for $\beta = 0.1$ and 1 (for $\beta = 0.1$ the results are numerically identical). However, when $\beta = 10$ the robust scheme is not able to match the performance of the optimized linear scheme.

VI. Conclusions

We considered the impact of switching delay on the performance-energy trade-off. The objective was to gain insight on how the optimized static speed scaling scheme performs against two optimized dynamic speed scaling schemes, the gated and the linear schemes, where the processor is switched off during idle state (to save energy) but that have an additional delay cost, the switching delay, when turned active again. Overall, our numerical results showed that, while in the system with no switching delay the dynamic schemes always yield a gain over the static scheme, the switching delay changes the situation dramatically. Indeed, the gated scheme especially can perform very badly, recall Figures 4 and 5 for $\beta = 1$ and 10. The optimized linear scheme, on the other hand, is much less sensitive to the switching delay. Finally, the trade-off is in practice very much affected by the parameters.

Future research topics include considering the case that energy is consumed already during the switching delay. Also, the impact of other service time and switching delay distributions could be considered. Finally, studying the energy-performance trade-off in a multi-server setting is of interest. This leads to a dispatching problem that has been analyzed without the switching delay for the gated scheme in [10].

Acknowledgement

This research has been partially supported by the TOP-Energy project funded by the Academy of Finland.

References

[1] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.

[2] 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*, Oct. 1995, pp. 374 –382.

[3] A. Wierman, L. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems," in *INFOCOM 2009, IEEE*, Apr. 2009, pp. 2007 –2015.

[4] L. L. Andrew, M. Lin, and A. Wierman, "Optimality, fairness, and robustness in speed scaling designs," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 1, pp. 37–48, Jun. 2010.

[5] J. M. George and J. M. Harrison, "Dynamic control of a queue with adjustable service rate," *Oper. Res.*, vol. 49, no. 5, pp. 720–731, Sep. 2001.

[6] P. D. Welch, "On a generalized M/G/1 queuing process in which the first customer of each busy period receives exceptional service," *Operations Research*, vol. 12, no. 5, pp. 736–752, 1964.

[7] H. Levy and L. Kleinrock, "A queue with starter and a queue with vacations: delay analysis by decomposition," *Oper. Res.*, vol. 34, no. 3, pp. 426–436, Jun. 1986.

[8] W. Bischof, "Analysis of M/G/1-queues with setup times and vacations under six different service disciplines," *Queueing Syst. Theory Appl.*, vol. 39, no. 4, pp. 265–301, Dec. 2001.

[9] A. Gandhi, M. Harchol-Balter, and I. Adan, "Server farms with setup costs," *Perform. Eval.*, vol. 67, no. 11, pp. 1123–1138, Nov. 2010.

[10] 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*, Nov. 2011, pp. 1 –8.