

Joint delay and power control in single-server queueing systems

Original

Joint delay and power control in single-server queueing systems / Bianco, Andrea; Casu, MARIO ROBERTO; Giacccone, Paolo; Ricca, Marco. - ELETTRONICO. - (2013), pp. 50-55. (IEEE Online Conference on Green Communications (Greencom)October 2013) [10.1109/OnlineGreenCom.2013.6731028].

Availability:

This version is available at: 11583/2513850 since:

Publisher:

IEEE

Published

DOI:10.1109/OnlineGreenCom.2013.6731028

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Joint delay and power control in single-server queueing systems

Andrea Bianco, Mario R. Casu, Paolo Giaccone, Marco Ricca
Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy

Abstract—Many power-aware resource allocation problems in packet networks can be modeled as single-server queueing systems, in which the power consumption depends on the actual service rate. We consider the scenario in which the queue service rate is controlled to minimize server power consumption. We show that power control methods that tune the service rate by using the queue length or the arrival rate exhibit a non-monotonic curve of delay vs. load. This may lead to malfunctioning in end-to-end flow/congestion control protocols, which are based on the assumption that delays increase with increasing load. We propose a new policy, in which the service rate is changed while keeping almost flat the delay curve, which permits to achieve a close-to-optimal trade-off between power and delay.

I. INTRODUCTION

Reducing the power consumption of telecommunication networks and devices is crucial for a number of reasons: i) the need to increase the battery life in mobile devices, ii) to reduce the energy bill of telecommunication operators and service providers, and iii) to design eco-sustainable products. Furthermore, minimizing power is a key solution to increase hardware performance. Indeed, the higher the processing and switching rates, the higher the power dissipated in chips and their temperature. Thus, reducing power per operation leads to higher switching and processing rates given a finite, and often tight, power and thermal budget.

In packet networks (in a broad sense, including the case of on-chip networks), in which shared resources are modeled as servers accessed via a queueing system that models resource interplay, the speed at which packets are served can be controlled to obtain a target power-performance trade-off. The various techniques proposed to reach this goal, as discussed in Sec. II, share the simple idea of tuning the server speed according to the load: When the load is low, the server slows down its service speed to reduce power consumption at the price of longer delays, and, possibly, lower throughput. When the load is high, the service rate is increased to maximize throughput, thus requiring higher power consumption.

In this paper we focus on the power control of a single-server system, in which arriving packets are enqueued and served in First-In-First-Out (FIFO) order. Albeit simple, this toy system permits to derive general observations that hold also in complex, more realistic scenarios.

We classify the power control methods in two categories, queue-length-based and arrival-rate-based. In the former, when the packets waiting to be served are less/more than a pre-defined queue occupancy threshold, the server rate is reduced/increased to save power/to reduce delays. Similarly,

in the arrival-rate-based method, when the packet arrival rate is smaller/larger than the actual service rate, the server reduces/increases its service rate.

To estimate power consumption we fix our attention on the cubic power-load relation typical of hardware systems that use Dynamic Voltage and Frequency Scaling (DVFS), in which the supply voltage scales jointly with the clock frequency [1], [2]. However, the adopted methodology is general and can be applied to a large family of convex power models.

To the best of our knowledge, for the first time we show that tuning the service rate according to the input traffic load leads to an anomalous behavior in the delay-load curve, which becomes non-monotonic for both queue-length and arrival-rate based methods. This fact may negatively affect the performance of end-to-end flow/congestion control protocols, which often assume that delays increase with increasing network congestion. For example, the congestion avoidance algorithm of some versions of TCP (as TCP Westwood [3]) is based on the estimated instantaneous rate achieved by the flow, which is usually obtained by the number of received packets/ACKs over the Round Trip Time (RTT). Clearly, increasing delays imply a larger RTT; this fact is “seen” by the control algorithm as a congestion indicator and the transmission window will be likely decreased. In the case of non-monotonic delay-load curve, increasing delays could be also due to a smaller load/congestion, which must instead lead to a larger transmission window. As a consequence, the sender might decrease its rate when the congestion decreases, leading to a vicious circle which may stop the sender, at least theoretically.

To remove the non monotonic behavior, we propose a new control method that keeps the delay constant over a large load range, with a minor power penalty with respect to policies that minimize power consumption at the cost of unbounded delays.

II. RELATED WORK

The power control problem in systems modeled as single server queues and with control policies based on the queue length and/or the arrival rate was previously investigated, but very rarely with focus on the delay behavior.

A queue-length-based control for DVFS applied to a multicore processor is proposed in [1], in which single-server queues model each core’s task queues. An emptying queue means that the core is running fast and is able to absorb its workload; a filling queue indicates that the core is not able to keep up with its assigned workload. The power control

is based on a standard Proportional-Integral (PI) feedback controller, which compensates the error between current queue size and target value by accelerating/decelerating the server. This scheme requires careful design and tuning of the control parameters and of the estimation procedures to achieve stability. In the field of Networks-on-Chip, a similar DVFS scheme has been used to control i) the power chip-wide, by using the size of the queues between various voltage domains [4], and ii) to control the power dissipated in on-chip routers using the size of input buffer queues [5]. The model that we present in Sec. IV refers to a version of the PI control, in which the average queue size is exactly equal to the target queue size.

For the power management of electronic systems, the authors of [6] propose a randomized policy for service-rate control based on the knowledge of the Markovian model describing the source behavior (i.e. the workload). As a consequence, the control is arrival-rate-based. The proposed policy achieves a good power-delay trade-off but it requires to solve a large size, nontrivial LP problem.

In a general context of communications between hardware components, the work in [7] focuses on DVFS applied to the interconnections modeled as a network of single-server queues. The authors compare different approaches to estimate the congestion information that feeds the rate algorithm. They also propose an alternative policy that combines arrival-rate-based and queue-length-based schemes and examines queueing delays. By comparing the state of the system with four target values (min/max queue occupancy/arrival rate), the service rate is adapted to minimize power consumption while keeping buffers occupancy small enough. The benefits of such scheme are evaluated through detailed hardware-level simulations, but the authors do not focus on the non-monotonicity of the delay with respect to the arrival rate.

Another example of an hybrid scheme, combining queue-length-based and arrival-rate-based approaches, was studied in [8] for the power management of generic data networks. The scheme takes into account also a setup penalty when changing the rate and a packet deadline (i.e. a maximum delay). Based on [9], which shows that keeping the service rate constant while satisfying a given time deadline is the minimum energy policy when arrival times are known offline, the proposed policy also tries to keep the service rate as constant as possible while meeting a given deadline, which can be violated only with small probability in case of unknown arrival times. The input of the power control are queue size and estimated arrival rate. The main idea is to increase the rate whenever the actual rate does not meet the deadline for the actual backlog, whereas the rate is decreased whenever the queue becomes empty.

A multi-class M/G/1 scenario is considered in [10] in a more theoretical perspective. The power is minimized while satisfying a maximum average delay. The proposed optimal policies are based on the knowledge of the arrival rate and the queueing delay for the packets in the queue.

As a final comment, note that similar approaches have been studied for M/G/1-PS (Processor Sharing) queueing systems, modeling the sharing of server resources. This is a very

relevant model for many applications, for which the rate control has been deeply investigated in the past [11].

III. POWER AND RATE CONTROL OF A SINGLE QUEUE

We consider a single server system in which the packet service time S computed by the power controller is defined as

$$S = \alpha T_{pkt} \quad (1)$$

where T_{pkt} is the minimum service time (obtained for maximum service speed) and $\alpha \in [1, \alpha_{\max}]$ is the time *expansion factor*, computed by the power control. Intuitively, α is the level of slow-down with respect to the maximum service rate and can be seen as the “laziness” to serve the packets. For $\alpha = 1$ the server is running at the fastest speed.

We can easily map this model in the DVFS scenario for a single server processor whose clock frequency is inversely proportional to the applied voltage V . Thus, α is the voltage reduction factor with respect to the maximum available voltage V_{\max} : $V = V_{\max}/\alpha$. α_{\max} depends on the adopted technology and it usually assumes values in the range 2-3 [12].

We assume that packets arrive according to a stationary Poisson process at rate $\hat{\lambda}_{pkt}$. To be admissible, $\hat{\lambda}_{pkt} < 1/T_{pkt}$. The normalized arrival rate $\lambda \in [0, 1)$ is $\lambda = \hat{\lambda}_{pkt} T_{pkt}$.

We consider the case of *static* policies, in which S is fixed for a given λ . This choice permits to simply build policy models. Static policies provide a bound to the performance of the corresponding *dynamic* policies that react to instantaneous changes in the arrival rate and/or the queue size by dynamically changing S . Indeed, it can be shown that under stationary traffic assumption, keeping S constant is better than changing it while keeping the same average value, both in terms of average delay and power. More precisely, the power for a static policy is lower than the corresponding dynamic policy power due the convexity of the power vs server rate function. This can be formally proved exploiting Jensen inequality (following the same reasoning of the proof of Lemma 1 in [9]). Following standard arguments in queueing theory, it can be also proved that the average delay for a static policy is lower, thanks to the lower (i.e. null) variance in the service rate.

Since the power policy is static, we can consider a fixed S and exploit the Pollaczek–Khinchine formula of M/G/1 for fixed service time (i.e., the corresponding queueing system becomes an M/D/1) to evaluate the average delay W , normalized with respect to T_{pkt} , as:

$$W = \frac{\lambda \alpha^2}{2(1 - \lambda \alpha)} + \alpha \quad (2)$$

To achieve bounded delays and maximize throughput, the power controller cannot reduce the service rate below the arrival rate: $\lambda < 1/\alpha$. If we define the *utilization factor* of the queue as $\rho = \lambda \alpha$, this condition is equivalent to impose $\rho < 1$. This results in the following final constraint:

$$1 \leq \alpha \leq \min \left(\frac{1}{\lambda}, \alpha_{\max} \right) \quad (3)$$

We assume that power consumption can be modeled as

$$P = \frac{\lambda}{\alpha^2} \quad (4)$$

This well-known model is motivated by a DVFS scenario since it captures the dynamic power of CMOS gates powered at voltage $V = V_{\max}/\alpha$, as shown for example in [2]. We omit in (4) all the constant factors so as to normalize $P \in [0, 1]$, as they do not affect the relative behavior of the control policies.

In the following, we discuss three control policies to choose α : (i) one that achieves the minimum power but with large delays/queue length, (ii) one that fixes a given queue utilization, and (iii) one that sets a given queue length. For the sake of comparison, we also define the No-power-Control (NC) policy as the one that always sets $\alpha_{NC} = 1$, for any $\lambda \in [0, 1]$.

A. Minimum Power (MP) policy

The minimum power in (4) is obtained by the maximum value of α subject to (3). This implies that the optimal Minimum Power (MP) policy is

$$\alpha_{MP} = \begin{cases} 1/\lambda & \text{for } \lambda \in [1/\alpha_{\max}, 1) \\ \alpha_{\max} & \text{for } \lambda \in [0, \rho_v/\alpha_{\max}) \end{cases} \quad (5)$$

This policy corresponds to force the queue to run, for any $\lambda \geq 1/\alpha_{\max}$, at an operating point corresponding to $\rho = 1$, for which the average delay and queue size are infinite. By also exploiting (2), it is easy to observe:

Property 1: The average delay is

$$W = \begin{cases} \infty & \text{for } \lambda \in [1/\alpha_{\max}, 1) \\ \frac{\lambda \alpha_{\max}^2}{2(1 - \lambda \alpha_{\max})} + \alpha_{\max} & \text{for } \lambda \in [0, 1/\alpha_{\max}) \end{cases} \quad (6)$$

The corresponding power is:

$$P = \begin{cases} \lambda^3 & \text{for } \lambda \in [1/\alpha_{\max}, 1) \\ \frac{\lambda}{\alpha_{\max}^2} & \text{for } \lambda \in [0, 1/\alpha_{\max}) \end{cases} \quad (7)$$

B. Fixed Utilization (FU) policy

To avoid infinite delays in (6), we propose to modify the MP policy. Since we must enforce $\rho < 1$ to obtain finite delays, we introduce the parameter $\rho_v \in (0, 1)$, defined as “virtual utilization factor”, defining a Fixed Utilization policy, denoted as FU- ρ_v . The policy expansion factor is:

$$\alpha_{FU} = \begin{cases} 1 & \text{for } \lambda \in [\rho_v, 1] \\ \rho_v/\lambda & \text{for } \lambda \in [\rho_v/\alpha_{\max}, \rho_v) \\ \alpha_{\max} & \text{for } \lambda \in [0, \rho_v/\alpha_{\max}) \end{cases} \quad (8)$$

When $\rho_v \rightarrow 1$, FU corresponds to the optimal minimum power policy, whereas when $\rho_v \rightarrow 0$ FU behaves as NC.

From (8) it is possible to highlight three regimes:

- *high load* (when $\lambda > \rho_v$) in which the service rate is maximum and the power control is not effective;
- *low load* (when $\lambda < 1/\alpha_{\max}$) in which the service rate is the minimum allowed;

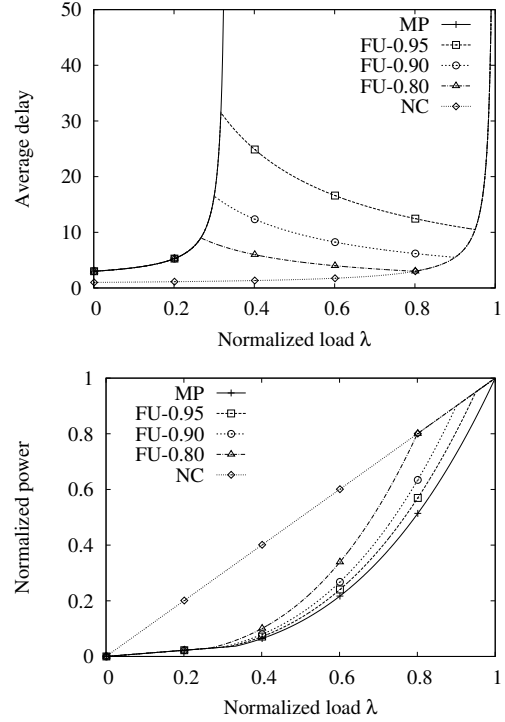


Fig. 1: Performance of FU- ρ_v policy when $\alpha_{\max} = 3$

- *medium load*, in which the service rate is controlled based on the arrival rate λ .

By again exploiting (2), it is easy to show:

Property 2: The FU policy defined via (8) obtains the maximum throughput for any admissible $\lambda \in [0, 1]$, given the finiteness of its average delay:

$$W = \begin{cases} \frac{\lambda}{2(1 - \lambda)} + 1 & \text{for } \lambda \in [\rho_v, 1) \quad (9a) \\ \frac{\rho_v^2}{2\lambda(1 - \rho_v)} + \frac{\rho_v}{\lambda} & \text{for } \lambda \in [\frac{\rho_v}{\alpha_{\max}}, \rho_v) \quad (9b) \\ \frac{\lambda \alpha_{\max}^2}{2(1 - \lambda \alpha_{\max})} + \alpha_{\max} & \text{for } \lambda \in [0, \frac{\rho_v}{\alpha_{\max}}) \quad (9c) \end{cases}$$

The corresponding power is:

$$P = \begin{cases} \lambda & \text{for } \lambda \in [\rho_v, 1] \\ \frac{\lambda^3}{\rho_v^2} & \text{for } \lambda \in [\rho_v/\alpha_{\max}, \rho_v) \\ \frac{\lambda}{\alpha_{\max}^2} & \text{for } \lambda \in [0, \rho_v/\alpha_{\max}) \end{cases}$$

Fig. 1 shows¹ average delay (measured as multiple of T_{pkt}) and normalized power as a function of arrival rate λ , for different values of control parameter ρ_v . For comparison purposes, we also report the results obtained by MP and NC.

As expected, MP delays are unbounded when $\lambda > 1/\alpha_{\max}$, whereas the power is minimum and corresponds to the cubic

¹All the curves in the following graphs (except the last two ones in the paper) are continuous, points simply help distinguish more easily the curves.

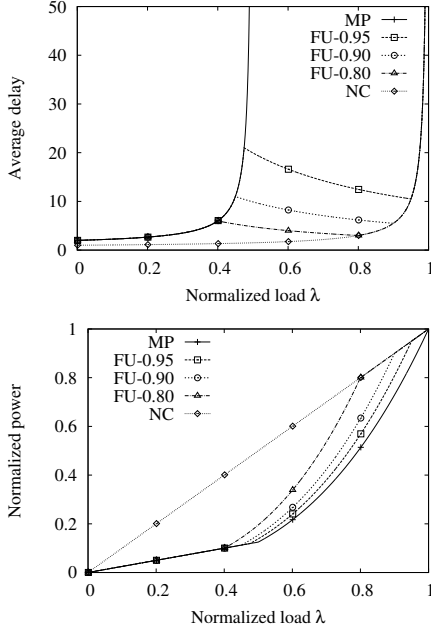


Fig. 2: Performance of FU- ρ_v policy when $\alpha_{\max} = 2$

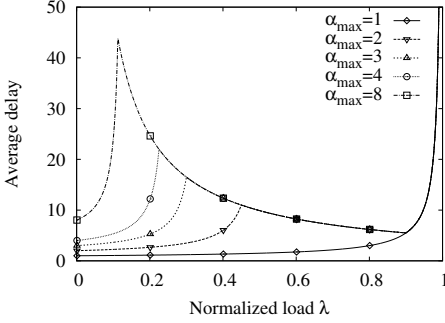


Fig. 3: Delay for FU-0.90 and different values of α_{\max} .

function in (7). Conversely, NC delays correspond to those of a standard M/D/1 queue and power grows linearly with λ .

More interestingly, under the FU policy the delays show a non-monotonic behavior as a function of the load, with a local maximum for $\lambda = 1/\alpha_{\max}$. This behavior is due to the fact that, in the medium-load regime, when λ decreases, the service time must increase to keep the same utilization factor ρ_v , since $\lambda\alpha = \rho_v$. Similar results are reported in Fig 2, which refers to the case of $\alpha_{\max} = 2$. Delays start decreasing for a different load value, but the curve shows the same trend.

Fig. 3 shows the effect on delays of different values of α_{\max} for the FU policy. The non-monotonic behavior is more evident for larger values of α_{\max} . Note that if α could increase without any bound (i.e., $\alpha_{\max} \rightarrow \infty$), then the corresponding delay would tend to infinity for $\lambda \rightarrow 0$.

C. Fixed Queue (FQ) policy

Another approach to cope with infinite queue lengths in MP is a power control based on the queue size. For this policy,

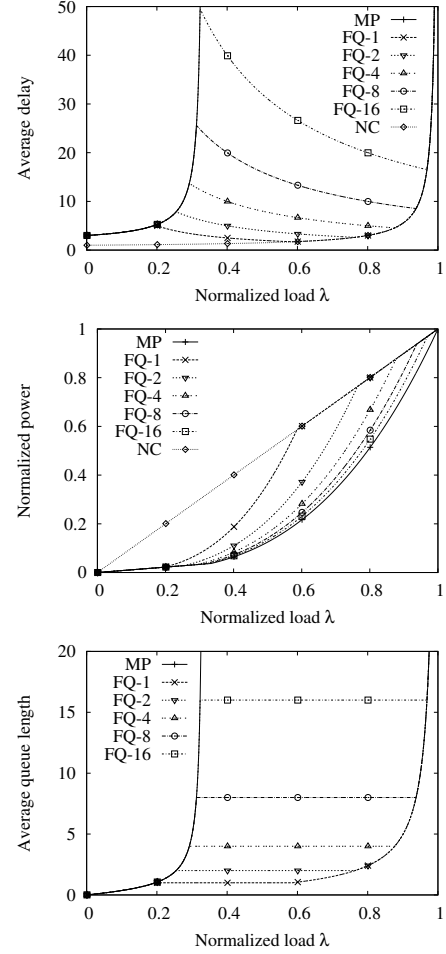


Fig. 4: Performance of FQ- L policies when $\alpha_{\max} = 3$.

FQ, α is chosen so that the average queue size equals a target value L . By using Little's law to express the average queue size, we can set $W(\lambda)\lambda = L$ and, thanks to (2), we have

$$\frac{\lambda^2 \alpha^2}{2(1 - \lambda\alpha)} + \alpha\lambda = L$$

Solving the above equation and considering the constraints in (3), we can define the FQ policy as follows:

$$\alpha_{FQ} = \begin{cases} 1 & \text{for } \lambda \in [\hat{L}, 1) \\ \hat{L}/\lambda & \text{for } \lambda \in [\hat{L}/\alpha_{\max}, \hat{L}) \\ \alpha_{\max} & \text{for } \lambda \in [0, \hat{L}/\alpha_{\max}) \end{cases} \quad (10)$$

where the new control parameter $\hat{L} = L + 1 - \sqrt{L^2 + 1}$, and $\hat{L} \in (0, 1)$. Even though this policy targets a fixed queue size L , is still arrival-rate based because it requires to know λ .

In Fig. 4 we show the average delay, power and average queue length for different target queue L and $\alpha_{\max} = 3$. We also report the power achieved by MP and NC, as a reference for the minimum and maximum possible power value, respectively. As expected, the FQ policy is able to guarantee a fixed average queue length under medium load. Furthermore, to achieve small L , the server rate must be

large enough: this translates to small delays but high power consumption. For large L , the server rate can be further lowered. Similar results are observed for other values of α_{\max} .

Note that FQ can be seen as the static version of a dynamic policy that varies the service rate using a formal control technique [1]. As noted at the beginning of Sec. III, the static policy outperforms the dynamic one under our assumptions.

It is easy to note a similarity between the (8) and (10), from which stems the equivalence between FU and FQ policies:

Property 3: A FQ- L policy with target queue size L is equivalent to a PM- ρ_v policy with virtual utilization factor ρ_v if any of the following conditions hold:

$$\rho_v = L + 1 - \sqrt{L^2 + 1} \Leftrightarrow L = \frac{\rho_v^2}{2(1 - \rho_v)} + \rho_v \quad (11)$$

Indeed, (9b) shows that for the FU policy the average queue size (computed as λW) is also constant for medium load. As a consequence of this equivalence, the same power-delay tradeoff is achieved by the two policies.

IV. POLICY WITH CONTROLLED DELAYS

It is possible to avoid the non-monotonic delays behavior by a careful choice of the expansion factor. We propose a new policy, denoted as Fixed Delay (FD), in which we impose that the delay for medium load is fixed. For a fair comparison with the previous policies, we set such fixed value equal to the delay $W'(\rho_v)$ obtained for FU- ρ_v at the specific load $\lambda = \rho_v$:

$$W(\lambda) = W'(\rho_v) \quad \text{for } \lambda \leq \rho_v$$

We can now leverage (2) and the fact that it must be $\alpha = 1$ for $\lambda = \rho_v$, to impose:

$$\frac{\lambda \alpha_{FD}^2}{2(1 - \lambda \alpha_{FD})} + \alpha_{FD} = \frac{\rho_v}{2(1 - \rho_v)} + 1$$

By solving the equation, we obtain the expansion factor α_{FD} for medium load. Hence, FD policy is defined as follows:

$$\alpha_{FD} = \begin{cases} 1 & \text{for } \lambda \in [\rho_v, 1] \\ \frac{-b + \sqrt{b^2 - 4ac}}{2a} & \text{for } \lambda \in [\rho^*, \rho_v) \\ \alpha_{\max} & \text{for } \lambda \in [0, \rho^*) \end{cases} \quad (12)$$

where $a = -\lambda(1 - \rho_v)$, $b = 2(1 + \lambda)(1 - \rho_v) + \lambda \rho_v$, $c = \rho_v - 2$ and ρ^* can be computed by imposing $\alpha_{FD} = \alpha_{\max}$. Observe that the FD policy is arrival-rate based, as the previous ones.

Fig. 5 shows the performance of the FD policy, for different values of ρ_v . As expected, the average delay is constant for medium load, whereas the power shows the same qualitative behavior of FU (and also FQ). To better highlight the differences, Fig. 6 compares the performance of FU and FD for the same value of ρ_v . By construction, the delay of FD for medium load is the same of FU at load $\lambda = \rho_v$. In the bottom graph, we plot the power ratio between FD and NC, which helps understanding the power reduction with respect to the case without power control. Since $\alpha_{\max} = 3$, the maximum power gain is $\alpha_{\max}^2 = 9$ for low load, corresponding to 11% relative power. For larger loads, the relative power tends to 100% as

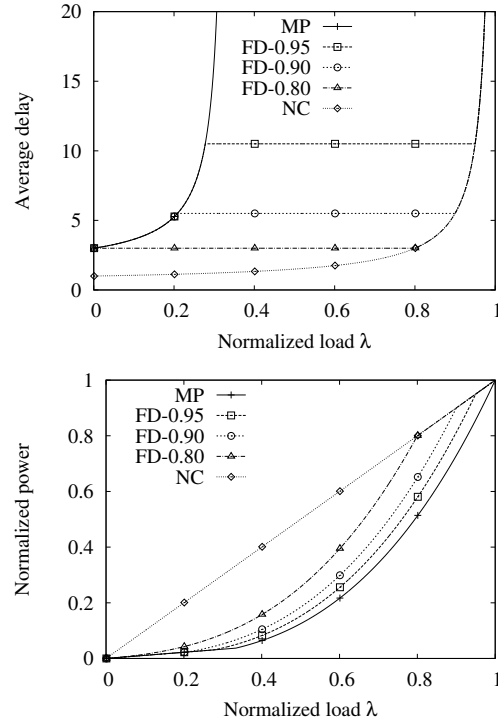


Fig. 5: Performance of FD- ρ_v policy when $\alpha_{\max} = 3$

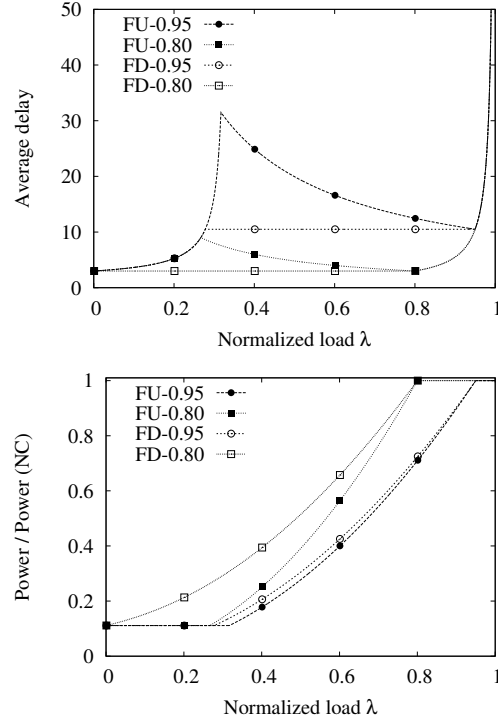


Fig. 6: FU- ρ_v policies vs. FD- ρ_v policies, $\alpha_{\max} = 3$

the power control becomes less and less effective. For medium load, the difference between the two policies for $\rho_v = 0.95$ is small (the power of FD is 15% larger than FD for $\lambda = 0.5$), but FD experiences smaller and monotonic delays (the delay

of FU is 65% larger than FD for $\lambda = 0.5$). Furthermore, for smaller values of ρ_v , the difference becomes larger: e.g., for $\lambda = 0.5$ and $\rho_v = 0.8$, the power of FD is 35% larger than FU, whereas the delay of FU is 60% larger than FD.

A. FD policy for a generic queue

The FD policy requires the knowledge of the analytic formula (2) relating the average delay to the load in the corresponding queueing system. In practical cases, this formula is not available, but this lack can be compensated by the empirical knowledge of $W(\lambda, \alpha)$ that can be obtained by profiling the delay for a large enough set of values of (λ, α) .

As an example, we consider a finite M/G/1/K queue, for which $W(\lambda, \alpha)$ can only be numerically computed. To profile queueing delays, we evaluate the average delay $\hat{W}(\lambda, \alpha)$ of the queue with steps $\Delta\lambda = 0.05$ and $\Delta\alpha = 0.01$. Then we set $\hat{W}(\lambda) = \hat{W}(\rho_v)$ as in the original FD policy, and compute the required value of α . To show the feasibility of the approach, we use an ad-hoc C++ queue simulator to simulate the packet arrival process in an M/D/1/K queue and to evaluate delay and power. Graphs² in Fig. 7 exhibit the same qualitative behavior of the FD policy adopted for the M/D/1 queue, proving that our approach is feasible also without analytic formulas.

According to the M/D/1/K model, which implies queue finiteness and Poisson arrival process, $\lambda = 1$ is not enough to saturate the queue, and the corresponding average queue size and delay tend to 5, i.e., half the maximum queue size.

V. CONCLUSIONS

We considered a single server queueing system in which the server rate is controlled to minimize the power consumption. We showed that the minimum power consumption is obtained only at the cost of unbounded average delays, and that two possible policies that exhibits finite delays while targeting either a fixed utilization (FU) or a fixed length of the queue (FQ) show a non monotonic delay/load curve. To overcome the possible drawbacks of such non-monotonic behavior, we proposed the Fixed-Delay (FD) policy, based on the arrival rate estimation, which achieves a fixed delay for a wide load range, with a slight power penalty if compared to FU and FQ.

As future work, we plan to investigate the interaction between these policies and end-to-end congestion/flow control schemes, in which non-monotonic delays may negatively affect performance. We will also consider the practical case in which arrival rates must be estimated in real time. This study will permit evaluating the performance degradation of dynamic policies with respect to the static policies considered in this work.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 257740 (Network of Excellence "TREND").

²In these figures only the points shown in the graphs have been simulated.

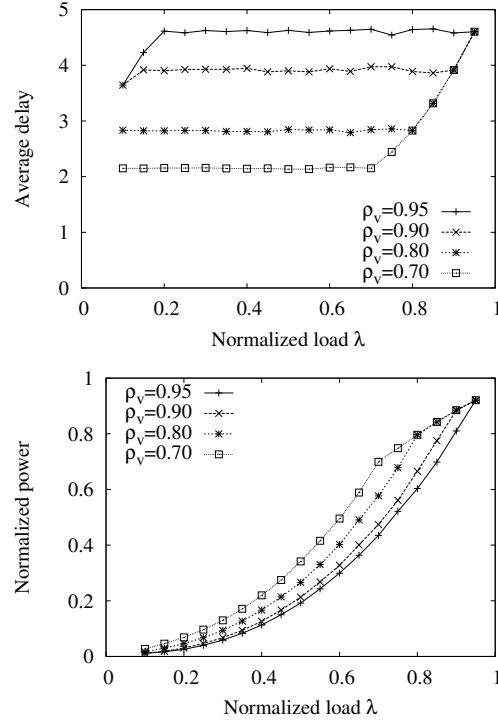


Fig. 7: Performance of FD policy in a simulated M/D/1/K queue with $K = 10$ and $\alpha_{\max} = 3$

REFERENCES

- [1] Q. Wu, P. Juang, M. Martonosi, L.-S. Peh, and D. Clark, "Formal control techniques for power-performance management," *Micro, IEEE*, vol. 25, no. 5, pp. 52–62, 2005.
- [2] A. Bianco, P. Giaccone, M. R. Casu, and N. Li, "Exploiting space diversity and dynamic voltage frequency scaling in multiplane network-on-chips," in *Proc. IEEE GLOBECOM*, 2012, pp. 3080–3085.
- [3] C. Casetti, M. Gerla, S. Mascolo, M. Sanadidi, and R. Wang, "TCP Westwood: end-to-end congestion control for wired/wireless networks," *Wireless Networks*, vol. 8, no. 5, pp. 467–479, 2002.
- [4] U. Ogras, R. Marculescu, D. Marculescu, and E.-G. Jung, "Design and management of voltage-frequency island partitioned networks-on-chip," *IEEE Trans. VLSI Syst.*, vol. 17, no. 3, pp. 330–341, 2009.
- [5] M. K. Yadav, M. R. Casu, and M. Zamboni, "A simple DVFS controller for a NoC switch," in *Proc. PRIME*, 2012, pp. 131–134.
- [6] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management," *IEEE Trans. VLSI Syst.*, vol. 8, no. 3, pp. 299–316, 2000.
- [7] L. Shang, L.-S. Peh, and N. K. Jha, "Dynamic voltage scaling with links for power optimization of interconnection networks," in *Proc. HPCA*, IEEE, 2003, pp. 91–102.
- [8] S. Nedeveschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing network energy consumption via sleeping and rate-adaptation," in *Proc. of USENIX NSDI Symp.*, no. 14, 2008.
- [9] M. Lin and Y. Ganjali, "Power-efficient rate scheduling in wireless links using computational geometric algorithms," in *Proc. IWCMC*, ACM, 2006, pp. 1253–1258.
- [10] C.-p. Li and M. J. Neely, "Delay and rate-optimal control in a multi-class priority queue with adjustable service rates," in *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 2976–2980.
- [11] T. V. Dinh, L. L. H. Andrew, and Y. Nazarathy, "Architecture and robustness tradeoffs in speed-scaled queues with application to energy management," *International Journal of Systems Science*, pp. 1–12, 2013.
- [12] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. DAC*, 2004, pp. 868–873.