Thermal control for crossbar-based input-queued switches

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

Thermal Control for Crossbar-based Input-Queued Switches

Andrea Bianco, Paolo Giaccone, Guido Masera, Marco Ricca
Dipartimento di Elettronica, Politecnico di Torino, Italy

***Abstract*—We consider an $N \times N$ input-queued switch based on a crossbar switching fabric implemented on a single chip. The thermal power produced by the crossbar chip grows as NR^3, where R is the maximum bit rate. Power dissipation is becoming more and more challenging, limiting the crossbar scalability for high performance switches.**

We propose to exploit Dynamic Voltage and Frequency Scaling (DVFS) techniques, quite commonly used in integrated circuit design, to control packet transmissions through each crosspoint of the switching fabric. Our thermal control operates independently of the packet scheduler and it is based on short-term traffic measurements. We propose a family of control algorithms to reduce the thermal power dissipation in non-overloaded conditions.

***Index Terms*—Input queued switch, energy, thermal control.**

I. Introduction

The aggregate bandwidth of high speed routers is growing fast, due to the increased traffic demand in the Internet. Usually, one or few switching fabrics are present in the core of the routers to switch all the data from the inputs to the outputs; each fabric is often implemented on a single integrated circuit. The hardware design of such fabrics is becoming more and more critical, because of the large pin count and the high bit rate. Indeed, if f is the maximum digital signal frequency, the power consumption of a single CMOS is proportional to f^3 [1]. In a $N \times N$ single-chip crossbar with N^2 crosspoint, each implemented through a combinatorial logic, we have $\theta(N^2)$ CMOSs (i.e., a fixed number for each crosspoint), and the total power consumption becomes proportional to R^3N, where R is the bit rate and N is the maximum number of data simultaneously flowing across the switching fabric.

Thermal dissipation is becoming a critical design issue, due to high integration level on a single chip, that implies very high power spatial density [2]. In integrated circuits, Dynamic Voltage and Frequency Scaling (DVFS) [1] is a classical technique used to control the power consumption. DVFS is based on the idea of jointly varying the power supply voltage and the peak signal frequency. A vast literature on DVFS techniques is available such as focusing on a single CMOS, on a CMOS cascade, and on a complete CPU.

In this paper we propose to use DVFS for the thermal control of a single-chip crossbar, analyzing the tradeoff between throughput (i.e., performance) and power consumption (i.e., thermal power to dissipate). The main idea is to exploit low load traffic conditions to extend packet duration by reducing bit voltage and frequency to control thermal power. Note that networks are typically provisioned for worst-case or peak-hour traffic. However, several measurements (see for example [3]) show that backbone utilization rarely exceeds 30%, thus suggesting that exploiting low traffic conditions can be a significant asset to reduce thermal power. We propose a set of algorithms for thermal control that operate on an estimated traffic matrix to assess the potential power gain that can be obtained exploiting DVFS. We take an idealized approach, i.e., we disregard the interaction with packet scheduling algorithms that select the packets to be transferred across the switching fabric. The system model is defined in Sec. II. Sec. III formalizes the optimal thermal control problem, describes its properties, and proposes a set of algorithms to solve it. Performance results in Sec. IV show the possible thermal gain of our approach. The main contributions of the paper: (i) definition of the thermal control problem for the crossbar; (ii) definition of the optimal algorithm and of simpler approximated algorithms; (iii) performance evaluation of such algorithms.

II. Problem Definition

We start by considering a single CMOS on which the combinatorial logic is based. Then, we examine the switching architecture to define the crossbar thermal control problem.

A. Energy model for a single CMOS gate

The energy consumption of a CMOS gate is strongly dependent on the supply voltage V and it can be modeled as a sum of a dynamic energy component (due to electrical signal switching activity needed to transfer sequence of 0s and 1s) and a static energy component (due to leakage currents). We consider only the dynamic energy component, while we neglect the latter contribution. Indeed, leakage currents can be made negligible with a proper hardware design, whose discussion is out of the scope of this paper. The energy due to a bit transition (i.e., the switching activity) is a quadratic function of V according to the well known formula $E_{bit} = 0.5CV^2$, where C is the load capacitance. If we consider a 0-1 square wave signal with frequency f, the power consumption is $P = E_{bit} f \propto fV^2$; this value represents also the thermal energy to dissipate. The maximum allowed frequency is

$$f_{\max} \propto V \tag{1}$$

due to the delay needed to switch from one logic state to another [4]. Thereby, the power consumption for a CMOS operating at maximum frequency and voltage is proportional to f^3. DVFS techniques jointly reduce V and f to minimize power consumption, exploiting time periods in which the signal can be "slowed down" to a lower peak frequency.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

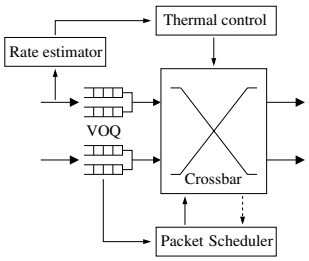

Fig. 1. Thermal control scheme

We consider a CMOS device operating at a voltage V between a minimum $V_{\min}$ and maximum $V_{\max}$. Within this range, we assume that bit transmission can occur at intermediate voltage levels. When operating at $V < V_{\max}$, thanks to (1), the signal peak frequency can be slowed down by a factor $\alpha = V_{\max}/V$ with respect to the maximum frequency allowed when using $V_{\max}$. Thus, α represents an expansion factor of the bit duration with respect to the bit duration when using $V_{\max}$. Furthermore, V should be larger than $V_{\min} > 0$, because of technological constraints that forbid to reduce too much the voltage level and of the impact of leakage currents, that would become not negligible anymore. Define $\beta = V_{\min}/V_{\max}$. Depending on the technology, $\beta = 0.5$ for a classical DVFS scheme or $\beta = 0.3$ in the case of an "extreme" DVFS scheme, according to [1]. By construction, $1 \leq \alpha \leq 1/\beta$.

B. Switching architecture

We consider an $N \times N$ input queued (IQ) switch, with virtual output queueing (VOQ), i.e. one queue VOQ_{ij} for each input i and output j pair. This IQ architecture allows high scalability in terms of line rate and number of ports, and the VOQ scheme is theoretically optimal from the performance point of view. To avoid dealing with data content at this abstract level, we assume that a data packet of length P is transmitted using P signal transitions: i.e., each packet is composed by a sequence of alternating 0 and 1. The maximum line rate for each port is $R_{\max}$, measured in [bit/s]; this value can be only reached for $V = V_{\max}$. The switching fabric is a $N \times N$ crossbar, with N^2 crosspoints and $\theta(N^2)$ CMOSs. The crosspoint from input i to output j is denoted as XP_{ij}.

III. Crossbar thermal control

In real switch implementation, a packet scheduler is responsible of selecting the set of packets to transfer simultaneously through the crossbar, satisfying the constraints that at most one packet is sent from each input and to each output. The scheduling decisions occur at a packet level, with a time granularity equal to the minimum packet duration. In the case of minimum Ethernet packet size and 10 Gbit/s line rates, a new scheduling decision must be taken every 50 ns. Given such a strict timing constraint, packet schedulers are implemented directly in hardware. A large literature is available on the design of low complexity and high performance packet schedulers for input queued switches [5]–[7].

Differently from the packet scheduler, the thermal control operates at a larger time scale, related to the milliseconds thermal constants of the materials employed to build the chip. As shown in Fig. 1, we propose a thermal control scheme decoupled from the packet scheduling decision, whose aim is to exploit DVFS at crosspoints to reduce the crossbar thermal power. Based on traffic measurements on the millisecond scale, the control sets the DVFS factor α_{ij} for the combinatorial logic present at XP_{ij}; each crosspoint is controlled independently. Note that, due to the relaxed timing constraints, the algorithm for thermal control can be implemented in software.

Let $\hat{\alpha} = [\alpha_{ij}]$ be the $N \times N$ matrix with all the DVFS factors. Note that setting $\alpha_{ij} > 1$ implies that the forwarding rate at XP_{ij} is reduced and the packet transmission time is increased by a factor α_{ij}. This has two main consequences: i) an additional queueing delay in VOQ_{ij}, ii) the packet scheduler cannot serve any new packet from input i and to output j until XP_{ij} ends the packet transmission. This means that the packet scheduler should take into account thermal control expansion factor in packet scheduling. We disregard this issue in the paper, and we take an ideal fluid-based approach, looking only at rates of I/O flows, to understand which are the potential benefits that can be obtained in terms of reduced power consumption.

A. Problem definition

The traffic load on each link is measured on a time window which duration is in the order of thermal constants (ms). Let r_{ij} be the average rate [bit/s] for the traffic flows enqueued at VOQ_{ij}, and $R = [r_{ij}]$ the corresponding $N \times N$ traffic matrix. Let $S = [s_{ij}]$ be the normalized traffic matrix obtained by setting $s_{ij} = r_{ij}/R_{max}$, with $s_{ij} \in [0, 1]$. We assume that $s_{ij} > 0$ for any i and j. The average load of matrix S is defined as $\rho_{ave}(S) = (\sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij})/N$. The load at input i and at output j is $\rho_i^I(S) = \sum_{k=1}^{N} s_{ik}$ and $\rho_j^O(S) = \sum_{k=1}^{N} s_{kj}$ respectively. The maximum load of matrix S is defined as $\rho_{\max}(S) = \max\{\max_k\{\rho_k^I\}, \max_k\{\rho_k^O\}\}$. and it is said to be admissible iff $\rho_{\max}(S) \leq 1$. Obviously, $\rho_{ave}(S) \leq \rho_{\max}(S)$.

We now model the constraints related to the maximum time expansion allowed for the transmitted bits. Consider a generic time period T, for which a flow rate is equal to r_{ij}. The total duration of the bit transmissions is $T_1 = r_{ij}T/r_{\max}$ and the maximum bit expansion factor is $T/T_1 = r_{\max}/r_{ij}$, i.e. $\alpha_{ij} r_{ij} \leq r_{\max}$. At the same time, we have stricter constraints on the expansion factor imposed by the traffic load in T on input and output traffic relations: $\sum_{i=1}^{N} \alpha_{ik} r_{ik} \leq r_{\max}$, $\sum_{j=1}^{N} \alpha_{kj} r_{kj} \leq r_{\max}$, $\forall k \in \{1, \ldots, N\}$ which can be normalized as

$$\sum_{i=1}^{N} \alpha_{ik} s_{ik} \leq 1 \qquad \sum_{j=1}^{N} \alpha_{kj} s_{kj} \leq 1 \tag{2}$$

The power consumption of XP_{ij} is

$$P_{ij} = r_{ij} \left(\frac{V_{max}}{\alpha_{ij}} \right)^2 = s_{ij} r_{\max} \left(\frac{V_{max}}{\alpha_{ij}} \right)^2$$

neglecting all constants of proportionality. The total crossbar power consumption is the sum of the power contributions of all crosspoints:

$$P_{tot} = \sum_{i=1}^{N}\sum_{j=1}^{N} P_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{s_{ij}}{\alpha_{ij}^2} r_{\max} V_{\max}^2$$

Finally, the minimum thermal power problem (denoted as OPT-MTP) becomes: given a admissible S, find $\hat{\alpha}$ that minimizes the cost function f_P:

$$\min_{\hat{\alpha}} f_P(\hat{\alpha}) = \min_{\alpha_{ij}\in\mathbb{R}^+} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{s_{ij}}{\alpha_{ij}^2} \tag{3}$$

such that

$$\sum_{k=1}^{N} \alpha_{ik} s_{ik} \le 1 \qquad \forall i \tag{4}$$

$$\sum_{k=1}^{N} \alpha_{kj} s_{kj} \le 1 \qquad \forall j \tag{5}$$

$$\alpha_{ij} \in \mathcal{A} \tag{6}$$

where $\mathcal{A}$ is the set of all available voltage levels.

Property 1: OPT-MTP is an integer convex non-linear optimization problem.

Following a standard methodology, we start to relax OPT-MTP to continuous variables; this defines the following problem, denoted as CONT-MTP: minimize (3) subject to (4) and (5); (6) is substituted by $\alpha_{ij} \ge 1 \ \forall i,j$ that corresponds to a DVFS scheme in which any voltage between 0 and $V_{\max}$ is allowed. Let $\hat{\alpha}_{\text{OPT-MTP}}$ be the optimal solution of OPT-MTP. Let $\hat{\alpha}_{\text{CONT-MTP}}$ be the optimal solution of CONT-MTP.

Property 2: $f_P(\hat{\alpha}_{\text{CONT-MTP}}) \le f_P(\hat{\alpha}_{\text{OPT-MTP}})$

i.e. $\hat{\alpha}_{\text{CONT-MTP}}$ as a lower bound on the thermal power cost.

Theorem 1: CONT-MTP is equivalent to

$$\min_{\hat{\alpha}} f_P(\hat{\alpha})$$

$$\sum_{k=1}^{N} \alpha_{ik} s_{ik} = 1 \qquad \forall i \tag{7}$$

$$\sum_{k=1}^{N} \alpha_{kj} s_{kj} = 1 \qquad \forall j \tag{8}$$

$$\alpha_{ij} \ge 1 \qquad \forall i,j \tag{9}$$

The proof of Theorem 1 is omitted for the sake of space. Note that exactly one of the constraints in (7)-(8) is linearly dependent from the other, and it can be omitted.

A non-negative matrix $H \in \mathbb{R}^{N\times N}$ is said to be ρ-double-stochastic if $\rho_i^I = \rho$ for any i and $\rho_j^O = \rho$ for any j. In this case, $\rho_{ave}(H) = \rho_{\max}(H) = \rho$. A 1-double-stochastic matrix is called double-stochastic matrix.

A non-negative matrix $H \in \mathbb{R}^{N\times N}$ is said to be ρ-sub-stochastic if $\rho_{\max}(H) = \rho$. In this case, $\rho_{ave}(H) \le \rho_{\max}(H)$.

Thanks to Theorem 1, CONT-MTP has the following explanation: given a ρ-sub-stochastic matrix S, find a double-stochastic matrix $\hat{S} = [\hat{s}_{ij}]$ such that the set of $\alpha_{ij} = \hat{s}_{ij}/s_{ij}$ minimizes (3). Hence, the problem consists of augmenting S such that it becomes double-stochastic.

In the following specific case, we can analytically compute the optimal solution:

Theorem 2: Given a ρ-double-stochastic matrix S, the optimal solution for CONT-MTP is $\hat{\alpha}_{ij} = 1/\rho$, for any i,j. The corresponding power consumption is $f_P(\hat{\alpha}_{\text{CONT-MTP}}) = N\rho^3$.

The proof is based on the use of the Lagrange multipliers and on the Taylor's Theorem for multivariate functions, and it is omitted here due to lack of space. Furthermore, we can get an important intuition from the above Theorem, that will drive the design of approximated algorithms for the CONT-MTP problem: In the optimal solution, all the α_{ij} are expanded proportionally by the same factor.

When considering also $V_{\min}$, the expansion ratio is limited by: $\alpha_{ij} \le 1/\beta$. The optimal solution becomes $\alpha_{ij} = \min(1/\rho_{\max}(S), 1/\beta) \ \ \forall i,j$ and the corresponding optimal solution for CONT-MTP becomes:

$$f_P(\hat{\alpha}_{\text{CONT-MTP}}) = \begin{cases} N\rho_{\max}(S)\beta^2 & \text{if } \rho_{ave}(S) < \beta \\ N(\rho_{\max}(S))^3 & \text{if } \rho_{ave}(S) \ge \beta \end{cases} \tag{10}$$

According to (10), β is the value of "critical load" above which DVFS is not able to expand the bit duration due to the constraints imposed by the traffic load in (2). Recall that, in practical applications, $\beta \in [0.3, 0.5]$.

Now we consider a relaxed version of the CONT-MTP problem, denoted as MISO-MTP, in which we remove the expansion constraints (7) on each input.

Theorem 3: Optimal solution of MISO-MTP is given by $\alpha_{ij} = 1/\rho_j^O(S)$. The related power cost is: $f_P(\hat{\alpha}_{\text{MISO-MTP}}) = \sum_j (\rho_j^O(S))^3$. The proof is omitted for absence of space and it is based on the definition of Lagrange function.

Property 3: $f_P(\hat{\alpha}_{\text{MISO-MTP}}) \le f_P(\hat{\alpha}_{\text{CONT-MTP}})$

i.e. MISO-MTP provides a lower bound, immediate to compute, for CONT-MTP and OPT-MTP.

A feasible, but not optimal, solution for OPT-MTP is when no DVFS scheme is adopted, i.e. $\alpha_{ij} = 1$ for all i,j. We define this scheme as NODVFS and the corresponding solution as $\hat{\alpha}_{\text{NODVFS}}$. The power cost f_P becomes

$$f_P(\hat{\alpha}_{\text{NODVFS}}) = \sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij} = N\rho_{ave}(S) \tag{11}$$

denoting a linear relationship between the average load on S and the total power consumption.

Property 4: $f_P(\hat{\alpha}_{\text{OPT-MTP}}) \le f_P(\hat{\alpha}_{\text{NODVFS}})$

This permits to use $f_P(\hat{\alpha}_{\text{NODVFS}})$ as a loose upper bound on the performance for OPT-MTP.

In summary, the solution to the CONT-MTP problem, assuming that any voltage level between $V_{\min}$ and $V_{\max}$ can be used, provides a lower bound for the thermal gain of the OPT-MTP problem deals with a finite number of voltage levels; the optimal solution to CONT-MTP is immediate only for ρ-double stochastic matrices.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

B. Thermal control algorithm

To solve OPT-MTP for any matrix we propose to: i) solve the corresponding CONT-MTP problem, ii) approximate each α_{ij} to the smaller voltage value available in $\mathcal{A}$. If α_{ij} is the solution for CONT-MTP, then use for OPT-MTP: $\alpha'_{ij} = \max\{\alpha \in \mathcal{A} \mid \alpha \leq \alpha_{ij}\}$. Note that, by construction, the set of α'_{ij} defines an admissible solution for OPT-MTP.

To solve CONT-MTP, we adopt two approaches:

- quasi-optimal algorithm (denoted as OPT), obtained by adopting the logarithmic barrier method for convex problems [8]. It provides an ϵ-approximation of the optimal solution, where ϵ is an input parameter, with enough large number of iterations. It converges quite slowly in our scenarios. Thus, we adopt it as a reference case for the optimal solution only.
- two-steps algorithm: we augment S to a double stochastic $\hat{S}$ according to three algorithms: AUGM-1, AUGM-MAX or AUGM-SORT. Then, we compute $\alpha_{ij} = \hat{s}_{ij}/s_{ij}$.

The three algorithms to augment S to a double-stochastic $\hat{S}$ are based on the MATRIX-INCREASE algorithm, described in the pseudo-code below.

MATRIX-INCREASE Algorithm
Input: $N \times N$ matrix $S = [s_{ij}]$, $\{\rho_i^I\}_{i=1}^N$, $\{\rho_j^O\}_{j=1}^N$, ρ_T, Ω^I, Ω^O.
Output: $N \times N$ matrix $\Delta = [\delta_{ij}]$
1. $\delta_{ij} = 0$ for any $1 \leq i, j \leq N$
2. $\Omega^{IO} = \{(i,j) : i \in \Omega^I, j \in \Omega^O\}$
3. **repeat** until no choice is anymore available
4. **choose** any $(i,j) \in \Omega$ such $\max\{\rho_i^I, \rho_j^O\} < \rho_T$
5. $\delta_{ij} = \min\{\rho_T - \rho_i^I, \rho_T - \rho_j^O\}$
6. $\rho_i^I = \rho_i^I + \delta_{ij}, \quad \rho_j^O = \rho_j^O + \delta_{ij}$

The MATRIX-INCREASE algorithms inputs are i) a sub-stochastic matrix S, whose corresponding row ρ_i^I and column ρ_j^O sums are pre-computed; ii) a target load value ρ_T such that $\rho_T \leq \max_k\{\rho_k^I\}$ and $\rho_T \leq \max_k\{\rho_k^O\}$, and iii) a set of inputs Ω^I and a set of outputs Ω^O. The algorithm returns a matrix $\Delta = [\delta_{ij}]$ with the largest possible elements such that: (i) only the elements δ_{ij} corresponding to rows and columns present in both Ω^I and Ω^O may be > 0; (ii) the maximum row and column sum is ρ_T, i.e.

$$\sum_{k=1}^{N} s_{ik} + \delta_{ik} \leq \rho_T \text{ for any } i \in \Omega^I$$

$$\sum_{k=1}^{N} s_{kj} + \delta_{kj} \leq \rho_T \text{ for any } j \in \Omega^O$$

The algorithm operates only on a sub-matrix corresponding to the rows in Ω^I and the columns in Ω^O. It chooses a sequence of elements in such sub-matrix for which both row and column sum to less than ρ_T. Then, each element in the sub-matrix is augmented as much as possible to reach ρ_T. Note that the maximum number of iterations in steps 3-6 is upper bounded by $2N$. Having defined INCREASE-MATRIX, we describe the algorithms that we propose to augment S to a double-stochastic $\hat{S}$:

- AUGM-1: i) compute ρ_i^I and ρ_j^O for any i, j; ii) run INCREASE-MATRIX on S, ρ_i^I, ρ_j^O, $\rho_T = 1$, $\Omega^I = \Omega^O = \{1, \ldots, N\}$; iii) compute $\hat{s}_{ij} = s_{ij} + \delta_{ij}$. This algorithm is a classical iterative algorithm (see Sec. II.A of [9]) to augment a sub-stochastic matrix to a double-stochastic one. The complexity is $O(N^2)$, due to steps i) and iii).
- AUGM-MAX: i) for any i, j compute ρ_i^I, ρ_j^O and then lastly $\rho_{\max}(S)$; ii) run INCREASE-MATRIX on S, ρ_i^I, ρ_j^O, $\rho_T = \rho_{\max}(S)$, $\Omega^I = \Omega^O = \{1, \ldots, N\}$; iii) compute $\hat{s}_{ij} = s_{ij} + \delta_{ij} + (1 - \rho_{\max}(S))/N$. The complexity is $O(N^2)$, due to steps i) and iv).
- AUGM-SORT: i) initialize $\hat{S}$ as S by setting $\hat{s_{ij}} = s_{ij}$ for any i and j; ii) compute ρ_i^I and ρ_j^O for any i and j; iii) sort ρ_i^I and ρ_j^O in increasing order; the induced sequence of inputs is described by $I_{(k)}$ defined as the k-th input and by $O_{(k)}$ defined as the k-th output; iv) let $\Omega^I = \Omega^O = \emptyset$. Iterate, for k from 1 to N, the following procedure: iv.a) $\Omega^I = \Omega^I \cup I_{(k)}$, i.e. compute the set of the k inputs with the smallest row sums; iv.b) $\Omega^O = \Omega^O \cup O_{(k)}$, i.e. compute the set of the k outputs with the smallest column sums; iv.c) run INCREASE-MATRIX on S, ρ_i^I, ρ_j^O, Ω^I, Ω^O and $\rho_T = \max\{\rho_{I_{(k)}}, \rho_{O_{(k)}}\}$, i.e. ρ_T is the maximum load for first k inputs and outputs of S; iv.d) update $\hat{s}_{ij} = \hat{s}_{ij} + \delta_{ij}$ for any i and j, and continue with a new iteration; v) compute $\hat{s}_{ij} = \hat{s}_{ij} + (1 - \rho_{\max}(S))/N$. The complexity is $O(N^2)$. In iv) this complexity is achieved by optimizing the data structure to choose an $(i,j) \in \Omega^{IO}$ in INCREASE-MATRIX and by initializing only once δ_{ij}.

AUGM-1 is a classical way to augment a matrix, but has the disadvantage that it increases a selected element to set the sum of the corresponding row or column equal to one. As such, a non uniform increase in matrix element is obtained. AUGM-MAX is a simple variant of AUGM-1 in which the matrix is augmented to reach exactly $\rho_{\max}$ for all rows and columns. Then, all matrix element are proportionally augmented until the matrix becomes double stochastic. This approach is based on the intuition derived from Theorem 2. Unfortunately, if even a single row or column sums to 1, AUGM-MAX degenerates into AUGM-1. Finally, AUGM-SORT order rows and columns in the original matrix so as to define a set of sub-matrices of different size. Each $k \times k$ sub-matrix includes a sub-matrix of size $(k-1) \times (k-1)$ of smaller load. Elements in each sub-matrix are augmented to reach the maximum admissible value in the sub-matrix, starting from the sub-matrix of smallest size.

IV. PERFORMANCE EVALUATION

To compare the DVFS schemes, we define the *relative power* $\eta(\hat{\alpha})$ of a DVFS solution $\hat{\alpha}$, relative to NODVFS, as:

$$\eta(\hat{\alpha}) = \frac{f_P(\hat{\alpha})}{f_P(\hat{\alpha}_{\text{NODVFS}})} = \frac{f_P(\hat{\alpha})}{N\rho_{ave}(S)} \tag{12}$$

Roughly speaking, $\eta(\hat{\alpha})$ is the thermal reduction factor compared to NODVFS. Since $\eta(\hat{\alpha}) \in [0,1]$, the closer $\eta(\hat{\alpha})$ to zero, the larger the scheme gain with respect to NODVFS.

A. Power consumption under ρ-double-stochastic matrices

According to Theorem 2, the optimal solution for CONT-MTP is expressed by (10). Fig. 2 shows the power consumption per port $f_P(\hat{\alpha})/N$ vs. the average load for CONT-MTP

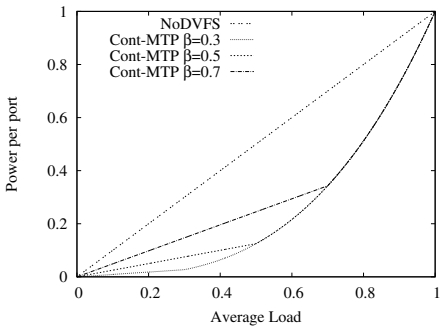

Fig. 2. Optimal solution for continuous DVFS, ρ-double-stochastic matrices.

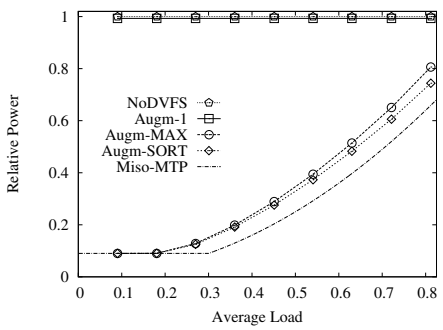

Fig. 3. Relative power vs. average load for $N = 256$ and continuous DVFS.

optimal solution and $\beta \in \{0.3, 0.5, 0.7\}$; NoDVFS shows a linear growing according to (11). For small loads, the DVFS scheme is really efficient, and the power reduction can be as high as 50%. For larger loads, the DVFS gain decreases, becoming negligible in highly loaded conditions, because bit expansion is not allowed due to the high traffic load.

In the case of a finite set of voltage levels, it can be shown that with only 4 voltage levels, the thermal gain is very close to the thermal gain that could be obtained with infinite voltage levels, i.e., the practical solution is not far from the ideal one.

B. Power consumption under ρ-sub-stochastic matrices

For general traffic matrices, simple optimal solutions are not available. We consider the family of *uniform* traffic matrices generated as follows. Given $\rho_{\max}(S) \in (0, 1]$, generate N^2 random variables u_{ij} uniformly distributed on the interval $(0, 1]$. Then, compute $\rho_{\max}(U)$ and derive each element of S as $s_{ij} = \rho_{\max}(S)u_{ij}/\rho_{\max}(U)$. Using this construction, it can be shown that the average load follows the law $\rho_{ave}(S) = \rho_{\max}(S)/(1 + \Theta(\sqrt{\log(N)/N}))$ for enough large N.

We compare the algorithms proposed in Sect. III-B for CONT-MTP, we have observed that it is a good approximation of OPT-MTP even when few voltage levels are available. The solution to MISO-MTP is easily obtainable, but it provides potential thermal gain that could not be reached in practice, given that the constraints among different output are neglected. The OPT algorithm is a tight bound but its computational complexity is huge. We set $\beta = 0.3$ and we do not report results for other values of β, being qualitatively similar.

To understand the algorithm scalability, we consider larger switching fabric size. We do not report thermal gain result for smaller size switching fabric being quite similar to those observed. Due to its slow convergence, especially at higher average load values, we do not present results for OPT in this scenario. Results for the CONT-MTP problem are presented in Fig. 3 that shown the relative power vs. the average load, for $N = 256$. Note that the maximum average load in the abscissa is limited due to the procedure to compute S. We have observed that the thermal gain for higher load is negligible.

For $\rho_{ave}(S)$ increasing, $\eta(\hat{\alpha}_{\text{MISO-MTP}})$ shows a quadratic growth, as expected by combining (10) with (12). More interestingly, $\eta(\hat{\alpha}_{\text{AUGM-SORT}})$ and $\eta(\hat{\alpha}_{\text{AUGM-MAX}})$ show a similar growth. As a consequence, the algorithms AUGM-SORT and AUGM-MAX provide performance close to the lower bound MISO-MTP. Regardless of the considered load $\eta(\hat{\alpha}_{\text{AUGM-1}})$ overlaps $\eta(\hat{\alpha}_{\text{NODVFS}})$ to confirm the inability of AUGM-1 to exploit the DVFS gain.

V. CONCLUSIONS

We discussed the potential thermal gains that DVFS techniques can provide when controlling a crossbar used as a switching fabric in an input-queued switch. We took an idealized approach, disregarding the details related to packet scheduling, looking at flow rates. Thus, DVFS schemes can be efficiently used to reduce power consumption especially at low average load regardless of the switch size. The proposed algorithms are computationally simple and obtain performance gain close to those of more complex, optimal algorithms.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's 7^{th} Framework Programme under grant agreement 247674 (STRONGEST).

REFERENCES

[1] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "The limit of dynamic voltage scaling and insomniac dynamic voltage scaling," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 13, no. 11, pp. 1239–1252, nov. 2005.

[2] F. J. Pollack, "New microarchitecture challenges in the coming generations of cmos process technologies," *Microarchitecture, IEEE/ACM International Symposium on*, 1999.

[3] https://research.sprintlabs.com/packstat/packetoverview.php.

[4] M. Flynn and P. Hung, "Microprocessor design issues: thoughts on the road ahead," *Micro, IEEE*, vol. 25, no. 3, pp. 16–31, May-June 2005.

[5] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Transactions on Communications*, pp. 1260–302, 1999.

[6] P. Giaccone, B. Prabhakar, and D. Shah, "Randomized scheduling algorithms for high-aggregate bandwidth switches," *IEEE Journal on Selected Areas in Communications, High-performance electronic switches/routers for high-speed internet*, vol. 21, pp. 546–559, May 2003.

[7] H. J. Chao and B. Liu, *High Performance Switches and Routers*. Wiley-IEEE Press, 2007.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[9] C.-S. Chang, W.-J. Chen, and H.-Y. Huang, "Birkhoff-von neumann input buffered crossbar switches," in *IEEE INFOCOM*, vol. 3, March 2000, pp. 1614–1623.