

Power-Aware Fat-Tree Networks Using On/Off Links

Marina Alonso¹, Salvador Coll², Vicente Santonja¹,
Juan-Miguel Martínez¹, Pedro López¹, and José Duato¹

¹Dept. of Computer Engineering

²Dept. of Electronic Engineering

Universidad Politécnica de Valencia

Camino de Vera s/n

46022 Valencia, Spain

{malonso,scoll,visan,jmmr,plopez,jduato}@upvnet.upv.es

Abstract. Nowadays, power consumption reduction techniques are being increasingly used in computer systems, and high-performance computing systems are not an exception. In particular, the power consumed by the interconnect circuitry has a non-negligible contribution to the total system budget. In this scenario, fat-tree interconnection networks are one of the most popular topologies. This topology is particularly well-suited for applying power consumption reduction techniques since it provides multiple alternative paths for each source/destination pair. In this paper, we present a mechanism that dynamically adjusts the available network bandwidth by switching links on and off, according to the traffic requirements. This mechanism provides significant reduction in power consumption while maintaining the original underlying routing algorithm, at the expense of slight latency increase for low loads.

1 Introduction

Nowadays power consumption reduction techniques are being increasingly used in computer systems, and high-performance computing systems are not an exception. Most of these systems are clusters (this is the architecture of 72.20% of the 500 systems listed in the November 2006 edition of the Top500 Supercomputers sites [1]). In particular, the power consumed by the interconnection network circuitry has a significant contribution to the total system budget. For example, the routers and links in a Mellanox server blade, consume about 37% of the total power budget [2]. In this scenario, fat-tree interconnection networks are one of the most popular topologies due to their high bisection bandwidth and ease of application mapping for arbitrary communication topologies [3]. But most applications have communication topology requirements that are far less than the total connectivity provided by fat-trees. Vetter and Mueller show that applications that scale most efficiently to large numbers of processors use point-to-point communications patterns where the average number of distinct destinations is relatively small [4]. This provides strong evidence that

many application communication topologies exercise a small fraction of the resources provided by fat-trees [5]. Moreover, traffic in an interconnection network exhibits large spatial and temporal variance, leading to inactivity periods at several links in the network [6]. On the other hand, fat-trees are particularly well-suited for applying power consumption reduction techniques since they provide multiple alternative paths for each source/destination pair. This paper shows that there is a chance to reduce power consumption by dynamically switching on/off links based on the traffic they support while running a set of applications.

Several power reduction techniques for interconnection networks have been proposed. Most of them are based on Dynamic Voltage Scaling (DVS). DVS was originally proposed, and now is widely deployed, for microprocessors. When applied to networks, this approach allows DVS links to work in a discrete range of frequencies and supply voltages, which leads to different levels of power consumption in response to their traffic utilization. The history-based DVS policy proposes to use past network utilization to predict future traffic, therefore tuning dynamically link frequency and voltage to reduce network power consumption [7]. Stine and Carter compare DVS with the use of adaptive routing in non DVS links, showing that, as long as the network provides enough bandwidth to meet the needs of the application, an adaptively-routed network can improve latency with the same power consumption [8]. DVS has significant drawbacks: it requires a sophisticated hardware mechanism to ensure correct link operation during scaling, it consumes significant CMOS area, and DVS links continue to consume power even while idle.

Other techniques are based on the use of on/off links that are selectively switched on and off according to their utilization [2,9,6]. In order to avoid deadlocks, adaptive routing algorithms must be used. Kim et al. also investigate hybrid techniques based on both DVS and on/off links. The idea is to shut down DVS links when traffic drops to very low levels [2].

In this paper, we present a new method to reduce power consumption in fat-trees based on the use of on/off links. The rest of the paper is organized as follows. Section 2 formalizes fat-tree network topology considering it a particular class of k -ary n -tree, including a description of packet routing. Section 3 describes the proposed power saving mechanism. Our proposal is evaluated using simulation in Section 4 and, finally, some conclusions are drawn in Section 5.

2 Fat-Trees

A k -ary n -tree is composed of $N = k^n$ processing nodes and $S = nk^{n-1}$ k -ary switches. Every switch has $2k$ ports, k “up” links and k “down” links. Processing nodes are identified by $(p_0, p_1, \dots, p_{n-1})$ where $p_i \in \{0, 1, \dots, k-1\}$ for $0 \leq i \leq n-1$, and each switch is identified by $(w_0, w_1, \dots, w_{n-2}, l)$, where $w_i \in \{0, 1, \dots, k-1\}$ for $0 \leq i \leq n-2$ and $l \in \{0, 1, \dots, n-1\}$ is the level of the switch (0 is the root level).

- Two given switches, $(w_0, w_1, \dots, w_{n-2}, l)$ and $(w'_0, w'_1, \dots, w'_{n-2}, l')$ are connected if and only if $l' = l + 1$ and $w_i = w'_i$ for all $i \neq l$. The link connecting both switches is labeled with w'_l on the level l switch and with $k + w_l$ on the l' switch.
- There is a link between the switch $(w_0, w_1, \dots, w_{n-2}, l)$ and the processing node $(p_0, p_1, \dots, p_{n-1})$ if and only if $w_i = p_i \forall i \in \{0, \dots, n - 2\}$.

The labeling scheme shown in the previous definitions makes the k -ary n -tree a delta network: any path starting from a level 0 switch and leading to a given node p_0, p_1, \dots, p_{n-1} traverses the same sequence of links (p_0 at level 0, p_1 at level 1, \dots, p_{n-1} at level $n - 1$) [10]. An example of such labeling is shown in Figure 1, for a quaternary fat-tree of dimension 3 (64-node network), that is a 4-ary 3-tree.

Given a k -ary n -tree, the Minimal Tree (MT) is the subset of the tree composed of all the processing nodes, a subset of the communication switches, and the edges between them. A switch $(w_0, w_1, \dots, w_{n-2}, l)$ belongs to the MT if one of the following properties holds:

1. $l < n - 1$ and $w_i = 0 \forall i \in \{l, \dots, n - 2\}$.
2. $l = n - 1$ (all the switches in the level $n - 1$ belong to the MT).

Within these switches, all the “down” links and the “up” links with index k also belong to the Minimal Tree. The Minimal Tree of a quaternary fat tree of dimension 3 (4-ary 3-tree) is shown in Figure 1 using thicker lines.

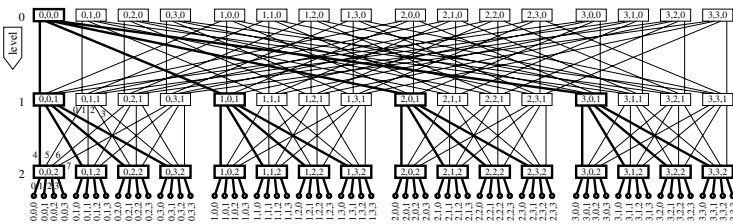


Fig. 1. 4-ary 3-tree node, switch and edge labels. The Minimal Tree is highlighted.

Minimal routing between any pair of processing nodes can be accomplished by sending the message to one of the nearest common ancestor switches and from there to the destination. Hence, each message experiences two routing phases: an ascending phase, from the processing node to a nearest common ancestor, followed by a descending phase. While the descending phase is necessarily deterministic, since there is a single path from a nearest common ancestor switch to the destination, there could be alternative routes to reach a nearest common ancestor. The availability of alternative routes makes it possible to randomly choose ascending links or even implementing an adaptive algorithm that makes a decision according to the local state of the switch, avoiding congested links.

3 Description of the On/Off Power Saving Mechanism

The proposed power saving mechanism is based on dynamically switching links on/off as a function of the required network throughput, and improves a preliminary version of this mechanism [11]. We consider bidirectional links that can be turned on/off in a given direction, either ascending or descending. Every switch in the network periodically measures outgoing traffic and controls the number of operating outgoing links depending on traffic variations. A subset of network links, which is defined as the Minimal Tree, fixing the maximum level of power saving, cannot be switched off in order to maintain the network connectivity.

In order to dynamically turn links on and off according to their utilization, the average utilization of all the “up” links in a switch, u_{up} , is periodically obtained. Two thresholds are defined to control the mechanism behavior: U_{off} is the turn off threshold, and U_{on} is the turn on threshold. If $u_{\text{up}} < U_{\text{off}}$ one of the “up” links is turned off, while when $u_{\text{up}} > U_{\text{on}}$, an inactive “up” link is turned on.

The power saving mechanism controls the network links state according to the following general rules:

- Up links: the utilization of the up links of a given switch is used to decide whether to turn on or turn off up links. This decision propagates upward to guarantee at least one path to level 0 switches (this is required to provide a path to every destination), and downward to guarantee descending routes to processors (for the same reason).
- Down links: these links are turned on/off all at once. Down links at a given switch are turned off when that switch cannot receive descending traffic, that is when all its input links (ascending and descending) have been turned off. Those links will be turned on again when some input link is turned on.
- The underlying routing algorithm need no changes since the mechanism is limited to those switches and links that do not belong to the MT (see Section 2), and the MT provides the minimum paths needed to maintain all the processing nodes connected.

A detailed description of the mechanism can be found in previous work [11].

The mechanism performance can be tuned by setting the values of the thresholds (U_{on} and U_{off}) used to control link on/off switching. Their effect can be analyzed by considering that threshold average indicates the mechanism aggressiveness while threshold difference (or hysteresis band) indicates mechanism responsiveness. Aggressiveness is related to the maximum power saving requested to the mechanism. Responsiveness refers to its ability to follow load changes.

The range of possible threshold values is conditioned by several limiting factors [12] that are summarized in the map of possible thresholds shown in Figure 2. This diagram is represented as a function of threshold difference and average. Any point inside the shaded region provides a valid configuration, with different responsiveness and aggressiveness as indicated with the bars depicted together with the axis in the graph.

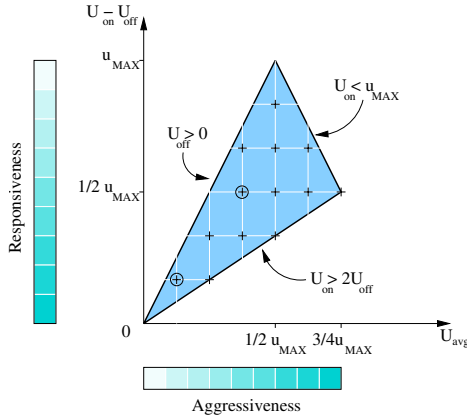


Fig. 2. Map of possible thresholds

3.1 Static Thresholds

Our mechanism uses static thresholds, as their value is constant regardless of the number of links that are on or off. Nevertheless, the mechanism generates an effect of increase (decrease) in the utilization of the available links as long as some of them are turned off (on). This effect could also be analyzed as whether the thresholds were reduced proportionally to the available throughput and the load was calculated relative to the full throughput. We refer those thresholds to as the effective thresholds. The effective threshold values as a function of the number of outgoing links in ascending direction for a 4-ary n -tree are indicated in Table 1. Factor column shows the fraction of available outgoing throughput used to calculate the effective thresholds. According to that, $\frac{3}{4}U_{on}$ can be considered as the minimum utilization for having all links active, while $\frac{2}{4}U_{off}$ is the maximum utilization that guarantees maximum power saving for a particular switch.

Table 1. Static threshold values according to the available links for a 4-ary switch

On Up Links	Static thresholds		Factor	Effective thresholds	
	On	Off		On	Off
4	U_{on}	U_{off}	1	U_{on}	U_{off}
3	U_{on}	U_{off}	$3/4$	$\frac{3}{4}U_{on}$	$\frac{3}{4}U_{off}$
2	U_{on}	U_{off}	$2/4$	$\frac{2}{4}U_{on}$	$\frac{2}{4}U_{off}$
1	U_{on}	U_{off}	$1/4$	$\frac{1}{4}U_{on}$	$\frac{1}{4}U_{off}$

Figure 3(a) shows the switch state (given by the number of ascending active links) versus the load traversing the switch in ascending direction for a 4-ary fat-tree. The state with all the links in the off state is not shown, since the last

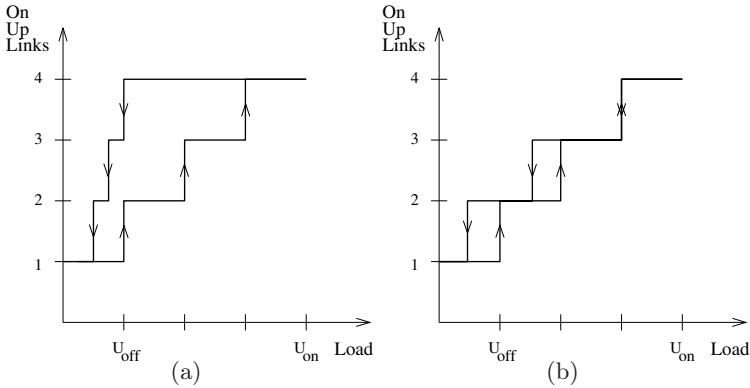


Fig. 3. Switch state as a function of traffic with (a) static and (b) dynamic thresholds

ascending link is turned off (except for the MT) when all the incoming links in the ascending direction have been already turned off.

The reduction in available throughput when reducing the active outgoing links generates an effect that can be viewed as a reduction in the hysteresis band (increase on responsiveness) of the mechanism. This is a positive effect since the network is more sensitive to congestion when the fraction of active links is reduced. In that situations, having a higher responsiveness will increase the mechanism agility to react against small changes in traffic, providing additional active links if needed. Otherwise a low responsiveness could lead to network congestion during limited periods of time, with a significant increase in latency.

An important limitation of the static thresholds is that U_{on} must be higher than $2 \cdot U_{\text{off}}$. This makes the threshold average to be low for situations with several active links, and hence making the mechanism less aggressive, which means reducing the margin for power saving. Moreover, this condition precludes the use of a mechanism that is both very aggressive and very responsive (see Figure 2).

3.2 Dynamic Thresholds

As an alternative, we have devised a version of the mechanism which is based on dynamic thresholds. The main objective is trying to divide the full range of network utilization in as many slots as indicated by the network arity. On average every switch distributes the ascending traffic among k links, when the network is fully active. If the traffic decreases $1/k$ of the nominal traffic requiring the full switch throughput it seems reasonable to reduce the available throughput by exactly the same amount, thus turning off one link.

The dynamic implementation is based on a fixed “on” threshold, U_{on} , and a dynamic version of the “off” threshold, U_{off} , that depends on the number of active outgoing links according to the following expression:

$$U_{\text{off}_i} = \frac{U_{\text{on}} \cdot (i - 1)}{k}$$

i being the number of active outgoing links in the ascending direction.

Considering a 4-ary n -tree, the set of dynamic thresholds together with the effective thresholds is indicated in Table 2, with the corresponding state transition diagram shown in Figure 3(b).

The dynamic implementation of thresholds provides significant power saving improvements, since the fraction of switch utilization where some links are turned off increases with respect to the static version. This result is validated in Section 4. The overlap among the transitions between 4 and 3 on links is not a problem (multiple alternative transitions due to traffic oscillations) since the mechanism operation is based on the average switch utilization during fixed length periods (Section 4.2) which has the effect of filtering quick traffic changes.

Table 2. Dynamic threshold values according to the available links for a 4-ary switch

	Dynamic thresholds			Effective thresholds	
On Up Links	On	Off	Factor	On	Off
4	n.a.	$\frac{3}{4}U_{\text{on}}$	1	n.a.	$\frac{3}{4}U_{\text{on}}$
3	U_{on}	$\frac{2}{4}U_{\text{on}}$	$3/4$	$\frac{3}{4}U_{\text{on}}$	$\frac{6}{16}U_{\text{on}}$
2	U_{on}	$\frac{1}{4}U_{\text{on}}$	$2/4$	$\frac{2}{4}U_{\text{on}}$	$\frac{2}{16}U_{\text{on}}$
1	U_{on}	0	$1/4$	$\frac{1}{4}U_{\text{on}}$	0

4 Performance Evaluation

In this section, we study, using simulation, the impact of the power reduction mechanism on latency. The metrics used in this study are the average latency of a message (measured from generation to delivery time) and the relative power consumption of the links as compared with the default system.

Two types of graphs are presented: the first shows the relative power consumption of the network links as a function of the injected traffic. These graphs includes two separate curves corresponding to the network behavior when an increasing or decreasing load is applied. Assuming that uniform traffic is used, this representation provides a view of the power consumption hysteresis band for the whole network. Note that the graph corresponding to the increasing workload should be read from left to right, while the one corresponding to decreasing traffic should be read from right to left. The second graph type shows the latency evolution for the same range of injected traffic, including results with the network links fully operational.

4.1 Network and Traffic Model

Our simulator models a wormhole switching network at the flit level [13]. The network is composed of switch nodes and processor nodes. The switches contain

a routing control unit, a crossbar and as many physical links as indicated by the network arity. Physical links are split into three virtual channels, with capacity for four flits. The results have been obtained for quaternary fat-trees of dimension 4 (256 nodes).

The network load is defined by the message generation rate at each node, the message size, and the destination of each message. Initially only the links in the minimal tree are active. For each test, the generation rate is kept constant at 0.01 flits/cycle/node during 60000 cycles, increased to its maximum value (0.60 flits/cycle/node) with constant slope during 120000 cycles, maintained at the maximum rate during 120000 cycles and decreased to the minimum load, again in 120000 cycles. A synthetic workload based on the uniform distribution is used. All the nodes in the network have the same behavior: the message inter arrival time is generated according to the workload type, the message length is fixed to 16 flits and, the destination node for each message is chosen among all the nodes in the network (except the source node) with the same probability.

We use a uniform distribution for the destinations because it corresponds to the worst case, as this workload produces a sustained traffic all over the network. If we had selected a workload that exhibits locality, some parts of the network would not receive any traffic. In that case, our mechanism would obtain much better results by permanently keeping the links in this area switched off.

4.2 Parameters of the Proposed Mechanism

As explained in Section 3, at a given time the operational level of a link depends on its utilization. The dynamics of the model is driven by the off threshold U_{off} and the on threshold U_{on} . In the following figures, we explore part of the design space by selecting different values of U_{off} and U_{on} in order to achieve different goals of responsiveness and aggressiveness for the power saving mechanism. The complete set of tested configurations is given by the cross (+) signs on the map of possible thresholds shown on Figure 2.

On the other hand, a link cannot be instantaneously turned on, but it requires a time T_{on} . Turning off a link also needs some time T_{off} to decrease the circuit voltage level to zero. When a link is turned off, we assume that it becomes immediately unavailable but it continues consuming power until T_{off} cycles have elapsed. Similarly, when a link is turned on, the new link is available to messages after T_{on} cycles, but power consumption increases at once. Based on the values reported by Kim et al., we have used $T_{\text{on}} = T_{\text{off}} = 1000$ clock cycles [14,9]. The state of the network is periodically checked to decide if it is necessary to turn on or off any link. We use a period greater than T_{on} and T_{off} in order to allow network stabilization after the changes. Specifically, the check period used is 2000 clock cycles.

4.3 Results with Static Thresholds

The power consumption and the latency results for three selected points from Figure 2 (highlighted with a circle) are presented below.

Figures 4(a) and 4(b) show results for the static thresholds $[U_{on}, U_{off}] = [0.030, 0.150]$. The power curves are very similar to those predicted by Figure 3, since they are the result of the overlapped effect of all network switches. These curves clearly show four different and stable network states. Each one of them is given by the average network performance of network states where the switches directly connected to processors have one, two, three or four active ascending links. This is also due in part to the fact that the reported experiments have been performed with uniform traffic to show the average network behavior. As expected, the network is at 100% power consumption with ascending traffic when the injected traffic surpasses $\frac{3}{4}U_{on}$ ($\frac{3}{4}U_{on} = 0.11$ for this test) for the 4-ary 4-tree topology.

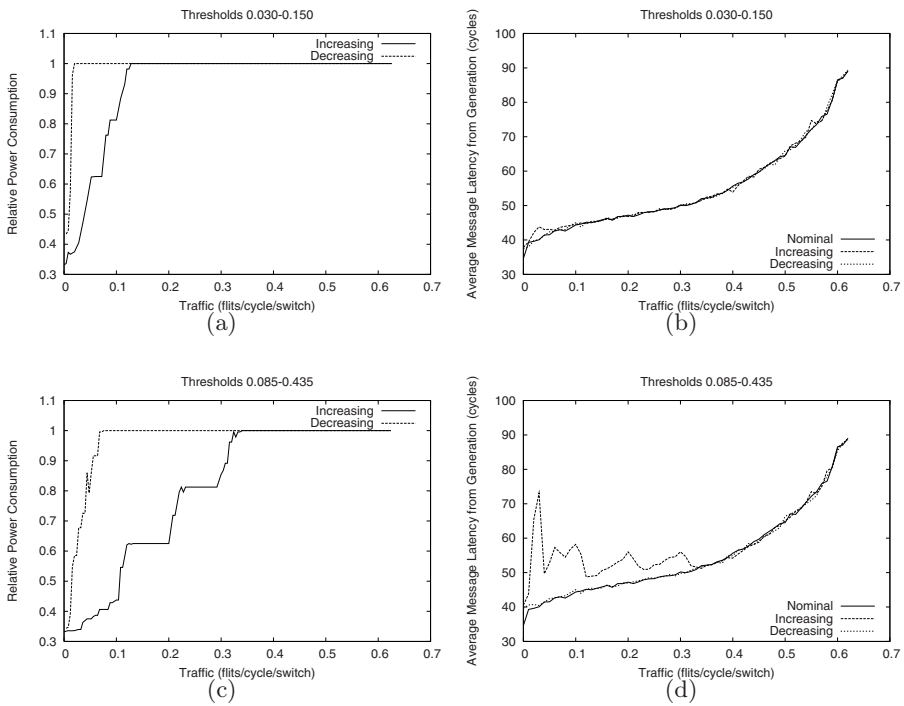


Fig. 4. Power and latency versus traffic with static thresholds

Figures 4(c) and 4(d) show results for the static thresholds $[U_{on}, U_{off}] = [0.085, 0.435]$, that provide a more aggressive power reduction mechanism. For this reason, the switches remain partially disconnected for higher loads and, as a consequence, the latency penalty increases. As can be seen the latency obtained for increasing traffic experiences several peaks for low traffics, which are due to the availability of a fraction of the total network throughput until the turning on of additional link allows higher loads. The latency curve with increasing traffic

denotes that the effective U_{on} thresholds are very close to the maximum traffic each one of the stable network states can deliver.

4.4 Results with Dynamic Thresholds

In this section, we report the results obtained for the selected points when using the dynamic thresholds version of the mechanism. As can be seen in Figure 5, the hysteresis band width reduction provides additional power reductions at no additional latency cost.

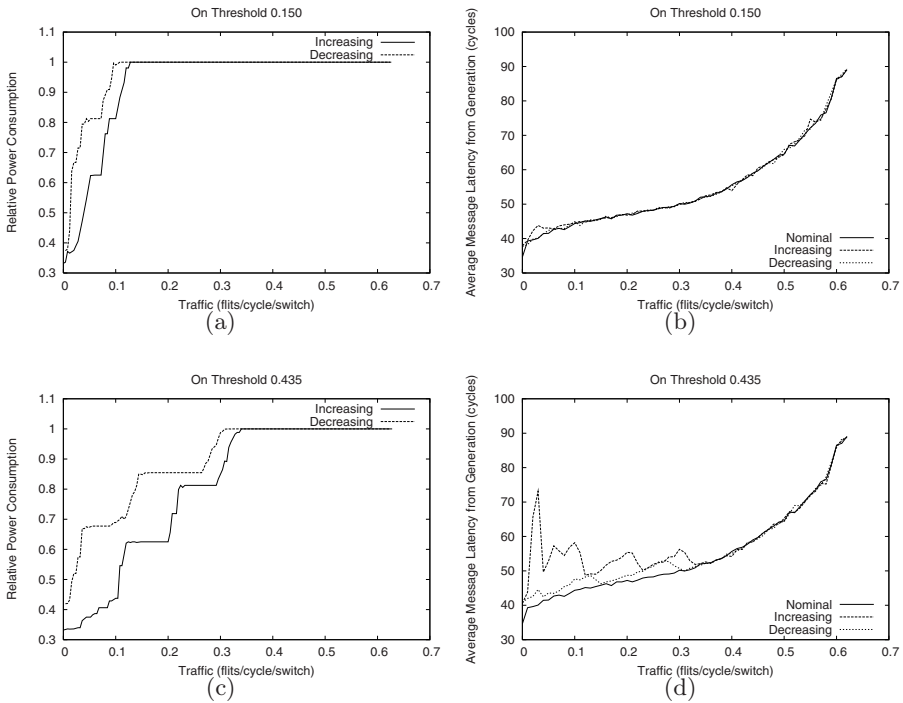


Fig. 5. Power and latency versus traffic with dynamic thresholds

The main benefits of the dynamic thresholds performance are based on the displacement of the decreasing power consumption traffic curve closer to the increasing one. The relative power saving increases dynamic thresholds provide range between 50% and 30% for the tests reported in this paper. It is important to note that the increasing traffic power consumption curves make state transitions for higher loads; hence the network experiences higher latency penalties than for decreasing traffic, as shown in the latency graphs.

The rest of the experiments corresponding to the points in the area shown in Figure 2 confirms the results reported in this paper. In particular, more aggressive thresholds introduce significant latency penalties. In some cases, latency

for very low loads becomes similar to the nominal latency obtained with the maximum traffic rate. Therefore, these very aggressive thresholds may not be interesting.

5 Conclusions

In this paper, we have presented a novel technique to reduce power consumption in fat-tree interconnection networks. Two important contributions of our mechanism are its simple implementation and the fact that the underlying routing algorithm does not need to be modified. The mechanism can be set to provide different levels of sensitivity to traffic variations. Moreover, power reduction policies with different levels of aggressiveness can be set, too. Hence, different ratios of power saving versus performance penalty can be obtained. Another significant contribution is the improvement of our original mechanism implementation by defining a dynamic behavior of the thresholds that control the mechanism operation. The dynamic version of the mechanism significantly outperforms the static approach at no additional performance cost.

Our results, obtained by simulation on 4-ary 4-tree networks, show that significant power savings can be obtained with moderate latency penalty, by selecting a conservative power reduction policy. Additional power savings can be obtained as well by further stressing the network at the cost of increasing latency. As future work we will further explore our proposal with realistic communication traffic patterns and we will analyze the eventual local congestion that may arise when many links are in the off state in order to improve network power-performance.

References

1. TOP500 Supercomputer Sites (2007), <http://www.top500.org>
2. Kim, E.J., et al.: Energy optimization techniques in cluster interconnects. In: ISLPED 2003, pp. 459–464 (2003)
3. Petrini, F., Vanneschi, M.: Performance analysis of wormhole routed k-ary n-trees. *Int. Journal on Foundations of Computer Science* 9(2), 157–177 (1998)
4. Vetter, J., Mueller, F.: Communication characteristics of large-scale scientific applications for contemporary cluster architectures. In: IPDPS (2002)
5. Shalf, J., Kamil, S., Oliker, L., Skinner, D.: Analyzing ultrascale application communication requirements for a reconfigurable hybrid interconnect. In: Proceedings of the ACM/IEEE SC 2005 Conference, SuperComputing 2005 (2005)
6. Soteriou, V., Peh, L.S.: Design-space exploration of power-aware on/off interconnection networks. In: ICCD 2004, San Jose, pp. 510–517 (2004)
7. Shang, L., Peh, L.S., Jha, N.K.: Dynamic voltage scaling with links for power optimization of interconnection networks. In: Proceedings of the 9th Int. Symposium on High-Performance Computer Architecture (HPCA-9), Anaheim, CA, pp. 79–90 (2003)
8. Stine, J.M., Carter, N.P.: Comparing adaptive routing and dynamic voltage scaling for link power reduction. *Computer Architecture Letters* (2004)

9. Soteriou, V., Peh, L.S.: Dynamic Power Management for Power Optimization of Interconnection Networks Using On/Off Links. In: Hot Interconnects 11, Stanford University, Palo Alto CA (2003)
10. Duato, J., Yalamanchili, S., Ni, L.: Interconnection Networks: an Engineering Approach. Morgan Kaufmann, San Francisco (2002)
11. Alonso, M., Coll, S., Martínez, J.M., Santonja, V., López, P.: Dynamic power saving in fat-tree interconnection networks using on/off links. In: HPPAC 2006, Rhodes Island, Greece, IEEE Computer Society, Los Alamitos (2006)
12. Alonso, M., Martínez, J.M., Santonja, V., López, P.: Reducing Power Consumption in Interconnection Networks by Dynamically Adjusting Link Width. In: Danelutto, M., Vanneschi, M., Laforenza, D. (eds.) Euro-Par 2004. LNCS, vol. 3149, pp. 882–890. Springer, Heidelberg (2004)
13. Duato, J.: A New Theory of Deadlock-Free Adaptive Routing in Wormhole Networks. IEEE Transactions on Parallel and Distributed Systems 4(12), 1320–1331 (1993)
14. Kim, J., Horowitz, M.A.: Adaptive Supply Serial Links with sub-1V Operation and Per-pin Clock Recovery. IEEE Journal of Solid State Circuits 1403–1413 (2002)