Parallel Computing 48 (2015) 59–80

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Power consumption management in fat-tree interconnection networks

M. Alonso [a,*], S. Coll [b], J.M. Martínez [a], V. Santonja [a], P. López [a]

[a] Department of Computer Engineering, Universitat Politècnica de València, Valencia 46022, Spain
[b] Department of Electronic Engineering, Universitat Politècnica de València, Valencia 46022, Spain

ARTICLE INFO

Article history:
Received 9 June 2014
Revised 13 March 2015
Accepted 19 March 2015
Available online 9 April 2015

Keywords:
Power consumption
Interconnection network
Fat-tree

ABSTRACT

As higher communication bandwidth is required in current designs of high performance parallel computers, the amount of power consumed by the interconnection network also increases. Fat-tree is one of the most popular topologies in high performance interconnection networks aiming at low latency, efficient collective communication and scalability. We present a new methodology for managing power consumption of fat-tree interconnection networks. Our proposal is based on dynamically adjusting available network bandwidth according to traffic requirements. To meet this goal, we define a mechanism for managing the operating status of network links as a function of network load. Our main contributions include a complete definition of the mechanism and a tuning methodology based on its sensitivity and aggressiveness in terms of potential power savings. Results show that our proposal can provide significant power savings (up to 67% in a 4-ary 4-tree) with no changes in the underlying routing algorithm, with minimal impact on network performance. Experiments conducted on a 16-ary 3-tree topology provide up to 36% energy savings with performance degradation below 1.10%.

1. Introduction

For some time, the cluster architecture is becoming increasingly popular in the supercomputing arena. The list of the most powerful computers (Top500) shows that cluster-based machines represent more than 80% and the size of the clusters is progressively growing. The cluster architecture is also common in large Internet servers, from web servers to the latest cloud systems.

Traditionally, computer architects have made an important effort in order to obtain high computing power. However, for quite some time, it is equally important to consider energy consumption because power saving is becoming an important issue in the design of any device that uses electricity as power supply. It is also the case of clusters. As an example, the Top500 list was proposed in 1993 to classify machines based only on computational power. In 2003, the Green500 list highlights the importance of energy efficiency with the introduction of the MFLOPS/watt figure of merit, that measures the rate of computation that can be delivered by a computer for every watt of consumed power.

Power saving is a very important aspect in cluster design due to different reasons: the problems generated by high power dissipation, such as high temperatures, that can damage the machine, plus the additional complexity involved in heat removal. Moreover, the cost of the energy bill is getting higher and any strategy to reduce it is welcomed. Finally, all computer architects

* Corresponding author. Tel.: +34 963877007.
E-mail address: malonso@disca.upv.es (M. Alonso).

http://dx.doi.org/10.1016/j.parco.2015.03.007
0167-8191/

must have a commitment to the environment, making their designs more sustainable. Following a holistic approach, power saving techniques are applied in different aspects of computer design. With energy optimization techniques being applied in processors and memories, attention is moving towards the interconnection network.

Interconnects are significant contributors to the power consumed by the whole system. For example, the routers and links in a Mellanox server blade, consume about 37% of the total power budget [1]. Greenberg et al. [2] showed that the network consumes 10–20% of a data center total power. Abts et al. [3] indicate that a typical interconnect consumes 12% of the total system power at full load, and up to 50% when the system is not fully utilized. They also state that a typical Google cluster spends most of its time within the 10–50% utilization range. Furthermore, increasing need for higher bisection bandwidth and faster, more power-consuming links, will make the network power become a first-order contributor to the overall power budget [3].

1.1. Related work

Different power reduction techniques for interconnection networks have been proposed, either based on varying the bandwidth; e.g. DVS, or through on/off links.

Most of them are based on dynamic voltage scaling (DVS). DVS was originally proposed for microprocessors and now is widely deployed. 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 dynamically tuning link frequency and voltage [4]. 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 [5]. However, Abts et al. [3] propose dynamically adjusting link data rates on data center networks using the flattened butterfly topology. They set a target utilization for each link and set the link rate (by halving or doubling the current rate) according to the actual utilization. This work, however, does not allow tuning the responsiveness of the mechanism. Situations were low link utilization is due to network congestion are not considered by this proposal. In addition, DVS has significant drawbacks: it requires a sophisticated hardware mechanism to ensure correct link operation during scaling, it consumes significant CMOS area, and links continue to consume power even while idle.

Another area of research is based on the use of on/off links that are selectively switched on and off according to their utilization [1,6,7]. These proposals require the interconnection network to use complex adaptive routing algorithms in order to avoid deadlocks. Koibuchi et al. apply the same strategy to Ethernet-based clusters where link deactivation require the paths that go through it must be changed to avoid the set of deactivated links [8]. Their proposal needs a network reconfiguration every time a link is activated or deactivated. 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 [1]. However, their work requires modifications in the routing algorithm.

Fat-tree interconnection networks have become one of the most used topologies in supercomputers due to their high bisection bandwidth and ease of application mapping for arbitrary communication topologies [9]. 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 [10]. This provides strong evidence that many application communication topologies exercise a small fraction of the resources provided by fat-trees [11]. Moreover, traffic in an interconnection network exhibits large spatial and temporal variance, leading to inactivity periods at several links in the network [7]. 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 while running a set of applications. Significant advantages of the proposed method are that it works in a distributed way, with no need for a central controller, it keeps network connectivity and does not require changes in the routing algorithm.

ElasticTree [12] is a network-wide power manager for fat-tree networks which dynamically adjusts the set of active switches and links devised for satisfying changing data center loads. Their proposal is based on a central network controller that predicts the incoming traffic matrix based on historical traffic. Yi and Singh [13] add a merge network to a fat-tree for pushing traffic to the ”left” so additional switches can be put into low power modes. Their proposal requires additional hardware to be deployed in the network and they do not provide any performance results.

In 2010, the IEEE 802.3az Energy Efficient Ethernet (EEE) standard was published [14]. After analyzing various alternatives, the task force defined a framework for dynamic on/off link management. The standard allows network links to switch between “sleep” and “wake” modes on demand, to save energy. Switches that support EEE, targeting data centers, are already commercially available.

In this paper, we present a method to reduce power consumption in fat-tree networks based by dynamically managing on/off links. This method can be easily applied in EEE networks, as the additional functionality required to turn links on and off is already defined. Our proposal defines a strategy for deciding which links are turned on/off and when. We propose a distributed mechanism based on two link utilization thresholds that control link, powering up and down according to network traffic. This mechanism can be easily set to provide different levels of sensitivity to traffic changes (responsiveness) and diverse power saving strategies (aggressiveness). Another advantage of our proposal is it does not require changes in the routing algorithm nor additional switching elements.