SANDIA REPORT

SAND2015-8717
Unlimited Release
Printed October 2015

Power Aware, Dynamic Provisioning of HPC Networks

Taylor Groves and Ryan Grant

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2015-8717
Unlimited Release
Printed October 2015

Power Aware, Dynamic Provisioning of HPC Networks

Taylor Groves and Ryan Grant
Center for Computing Research
Sandia National Laboratories
tgroves, regrant@sandia.gov

Abstract

Future exascale systems are under increased pressure to find power savings. The network, while it consumes a considerable amount of power is often left out of the picture when discussing total system power. Even when network power is being considered, the references are frequently a decade or older and rely on models that lack validation on modern interconnects. In this work we explore how dynamic mechanisms of an Infiniband network save power and at what granularity we can engage these features. We explore this within the context of the host controller adapter (HCA) on the node and for the fabric, i.e. switches, using three different mechanisms of dynamic link width, frequency and disabling of links for QLogic and Mellanox systems. Our results show that while there is some potential for modest power savings, real world systems need to improved responsiveness to adjustments in order to fully leverage these savings.

This page intentionally left blank.

Contents

List of Figures

List of Tables

This page intentionally left blank.

Chapter 1

Introduction

Current trends in HPC design point to future systems over-provisioned to meet the peak demands of applications – with a large number of processing elements and massive network capacity. Despite the increased capabilities of these systems, utilization of these components is generally sub optimal. Applications may not be able to take full advantage of the degree of parallelism offered by the system, or contention for resources creates barriers to forward progress. Similarly, networks are often subject to bursty patterns of traffic and a imbalanced distribution of traffic across links. Reducing wasted power and efficiently managing these resources is a significant challenge in exascale design. As of this writing, Infiniband switches with a typical radix of 36 ports utilize a power supply in the range of 100-300 watts. Often the network is overlooked as a consumer of system power, but at exascale the network may consume 10-20% of the total power budget. In traditional data centers, the proportion of power going to the network is expected to be even greater [2].

Recent work that explores the power and energy costs of dynamic network fabric references power estimates more than a decade old. Other work leverages literature on Energy Efficient Ethernet to make assumptions about the capabilities of Infiniband networks. These estimates are often derived from models which assume optimal environments and hardware that is not currently deployed in commercial network technologies. One of the goals of this report is to examine modern network hardware and take empirical measurements of power and fabric adjustment delays. With these measurements we can discuss where modern interconnects fall short of the requirements needed to save power in modern systems.

For the purposes of this report we focus on Infiniband network fabric, but future work may focus on other fabrics (including Ethernet and Cray networks). As of 2014 Infiniband is the most common supercomputing interconnect (illustrated in Figure 1.1). Infiniband provides massive bandwidth of 56 Gbit/s with microsecond latency. Remote Direct Memory Access (RDMA) is supported, providing communication with low CPU overhead. Performance of Infiniband networks is classified by the *frequency*, e.g. Single/Double/Quad/Full Data Rate (SDR/DDR/QDR/FDR), as well as the *width* (the number of aggregate links per port). Of particular importance to our work, Infiniband provides both adjustable link width and link operating frequency. Additionally, links may be completely disabled if there are redundant paths.

For this report we are interested in answering the following questions:

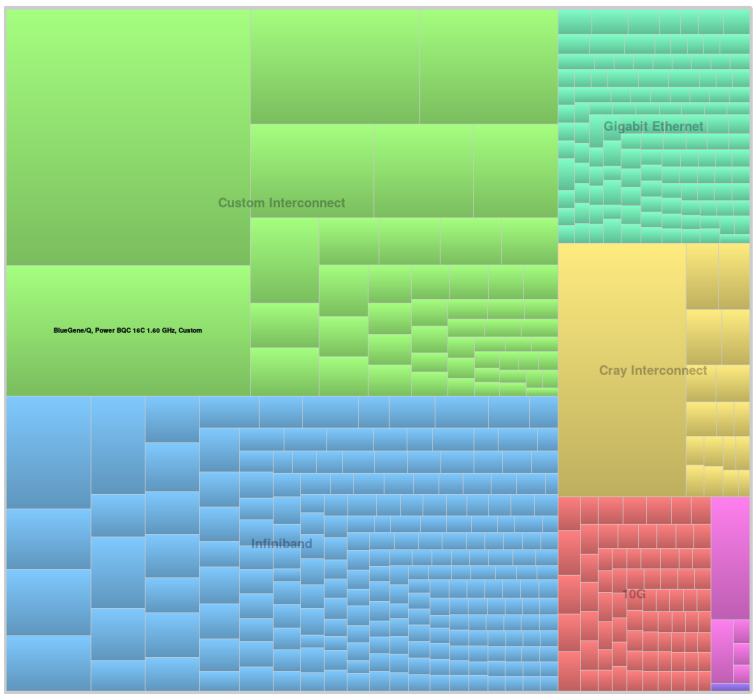

Figure 1.1. Treemap of Top500 interconnect families with box size based on Rmax flops. The top left group represents "custom" interconnects, including IBM, Cray, Tiahne, and the K computer. The bottom left cluster is strictly Infiniband networks. Figure taken from TOP500.org November 12, 2014.

- What is the delay in making an adjustment to link width, frequency, or completely disabling it?
- What are the potential power savings?

Chapter 2

Related Work

Response times to fabric adjustments Some related work has already examined the theoretical delays to make an adjustment to a network link. In [7], Kim states that the frequency change at run-time is accomplished by changing the input clock to the transmitter. In this scenario, the adaptive power supply will attempt to track the new frequency, while the receiver will attempt to regain lock to again receive data. Therefore, the link does not support data transfer during periods of frequency change. Two components determine the time to regain a lock. These components are the clock-matching phase-locked loop (PLL) and the adaptive supplies. Duarte et al. give PLL lock times a delay of 400ns [5], while lock times on the variable power supply may be higher [8].

In [16, 15], Totoni et al. propose the addition of hardware support for on/off link control. Furthermore, they suggest that this control should be managed by an adaptive runtime system. This work takes a binary approach of completely disabling/enabling links (with a zero cost delay). Simulated experiments suggest that a up to 20% of total system power may be saved as a result. One of the strengths of this work is a thorough analysis of the link utilization for a set of real applications using realistic topologies.

In [6] link adjustment downtime for Energy Efficient Ethernet is estimated as 10μs. In the absence of traffic, Recent products from Mellanox offer a technique called Width Reduction Power Saving (WRPS) [1]. This approach promises link width adjustments without experiencing a disconnection of the network. While, we have not validated the manufacturer's claims, if true, there would be significant opportunities for power savings.

The work of Saravanan et al. [12] is significant – in that it examines the performance of Energy Efficient Ethernet (EEE) in the domain of HPC. Their findings suggest that by default, EEE did not provide power savings, but given a reduced on/off transition delay there were overall power savings of 7.5%.

Power savings from fabric adjustments In 2003, Kim et al. [7] introduced a scheme for Dynamic Link Shutdown (DLS), which attempts to identify highly used links and shut down other links whose usage is below a threshold, without creating a disjoint network. In this work they develop detailed models of link and switch energy at 180nm and simulate energy consumption on a RISC based energy simulator.

An approach by Shang, Peh and Jha [13] explored how Dynamic Voltage Scaling (DVS) may predictively be applied to links, based on a weighted moving average of buffer and link utilization. This work constructs a multi-level DVS link model, where link frequencies range from of 0.2 to 1 GHz and 1V to 2.5V, respectively.

In 2004, Soteriou and Peh [14] state that the IBM Infiniband 12X LPE TX consumes between 0.26 and 0.3 watts of power, while the RX consumes between 0.17W to 0.2W. These values are later referenced by Dickov et al. who multiply the combined 0.5W by the link width.

Work by Li et al. [11] presented an approach called Network Power Shifting or NPS. In this work they assume that optical transceivers consume 3W each (citing Avago Technologies data sheets). The work assumes that 64% of nominal switch power is derived from the costs of the optical links.

In 2012, Laros et al. [10] provided insight on how statically scaling the CPU and network can provide power savings on Cray XT systems. In their results it was found that very few applications fully utilized the available bandwidth and that they could scale back the network to 50% of capability with very small execution time increases for the majority of applications. One of the significant contributions of this work is that the experiments utilized real systems, applications and power measurements rather than simulations.

Dickov et al. [4] simulated predictively reducing the link width of the network fabric by annotating the MPI layer. Their work references a Mellanox whitepaper [1], which claims a reduction of switch (SX6036) power to 43% of the nominal power, by reducing all links to 1X width.

Chapter 3

Evaluation of Dynamic Fabric

Link Adjustment Delay

In an effort to validate the estimates of previous studies, we have performed some preliminary experiments on the Teller system at Sandia National Labs. Teller nodes consists of a AMD A10-5800K (Piledriver) 3.8GHz Quad-core CPU and a Qlogic, QDR Infiniband Network. Teller has several nodes which utilize different off-loaded Mellanox cards, however these were not of use, since they only support 4X width. In these experiments we time the downtime of a link after the width or frequency has been changed. Specifically after changing the link width or frequency with the ibportstate command (on both the node and switch ports), we time the link while querying the state until it changes from *polling* into an *Active* state. The actual timing of the adjustment was accomplished using the *time* command on a script which continually polled link state and returned when the link state was Active.

Adjusting link frequency In our scripts we adjust the link frequency from 10Gbps to 2.5Gbps before issuing a reset command to the link, which we time. the time taken to reset the link was 4.697s. This downtime is significantly more than hardware necessitates, but it provides an upper bound for adjusting the links using a software defined approach.

Adjusting link width Multiple attempts to adjust the link width were unsuccessful and left the links in a polling state, where the switch and host HCA were unable to negotiate. The only way we were able to adjust the link width was by adjusting the link frequency to 2.5 Gbps frequency in advance on the host HCA, while leaving the switch HCA speed set to 7 (2.5, 5.0 or 10Gbps). Then we adjust the switch width to 1 (1x) while leaving the host HCA at 3 (1x or 4x). The time to complete the reset was again, long (4.439s).

Disabling the link As expected disabling the link takes very little time (0.045s). This is largely due to the fact that this is a local operation which does not require an interaction with the Infiniband Subnet Manager.

Table 3.1. Power Savings Per Switch Port (Mellanox)

Speed Gbps	Width	Decrease in Watts
10	4	0
5	4	0.24
2.5	4	0.45
2.5	1	1
Disabled	Disabled	1.29

Enabling the link Enabling the link on the QLogic switch from a disabled state took as much time as it took to perform a width or frequency adjustment (4.069s). On the Mellanox switch we were unable to re-enable the HCA through the ibportstate command once it was disabled.

Power Savings on the Switch

On the switch we performed three different operations in an attempt to find power savings. Using a WattsUp? power analyzer, watt meter, we took the average of 40 power measurements both before and after the link adjustment. The standard deviation across 40 measurements was under 0.05 watts in all cases. Our measurements did not show any change in power savings for the Qlogic 12300 series switch. Even in the disabled state there was no observable power savings, showing that this switch is not power optimized.

Following the negative results for the Qlogic Infiniband Network, we examined a smaller Mellanox cluster comprised of 2 nodes on a 36 port Mellanox MTX3600 switch. While, the time to adjust the links on this switch was similar to the time taken on the Qlogic switch, we observed small decreases to switch power consumption on a per-port basis. In Table 3 we show the recorded **per-port** power savings. That is, we show the power savings to the switch as the specified adjustment is made to each port. The Speed column in the table is specific to the parameter of ibportstate. For a given row of the table, if the width is unchanged and the speed is reduced, then there is an implicit decrease to the link frequency. For a switch with 36 ports the total maximum savings could be up to 36 times the value of the adjustment in Table 3. In the case of the MTX3600 switch we would expect the minimal power (if all links are disabled) to be 46 watts less than the switch running at full power. It should be noted that these results are for a 4X switch width and that greater widths (such as 12X) imply a potential for larger power savings.

Energy Proportionality Literature from the last decade [14] is often cited stating that modern interconnects are not energy proportional. That is, they do not reduce their power utilization significantly as a function of the amount of data in transit. While this is an accepted claim in literature, we wanted to verify it in our experiments. For this validation, we sent a 2.7 GBps flow of data across the network using the ib_write_bw benchmark, while measuring switch power. We then compared this to idle switch power. In each case we

Table 3.2. Power Savings Per HCA

Speed Gbps Speed Gbps	Width Width	Decrease in Watts (Qlogic)
10	4	0
2.5	1	0.57

took 40 power measurements over 40 seconds (1 per second). The average idle power was 90.59 watts for the MTX3600, while under load it went up slightly to 90.69. In both cases the standard deviation was 0.03 watts. The measured difference between an idle state and state of high utilization was 0.10 watts. Such a small difference in power consumption for such a large difference in switch utilization suggests that modern fabric is still not power proportional.

Power Savings on the Node

Using PowerInsight [9], we measured the power of the HCA's on the Teller system as we adjusted link width and frequency. PowerInsight takes power measurements of a node completely out of band and utilizes a separate network from the main Infiniband network. This allows us to take power measurements which do not perturb the system. PowerInsight reported an average savings of 0.57 watts (6.83 to 6.26) for the 12 watt rail, when going from a 4X-10 Gbps link to a 1X-2.5 Gbps link, with a standard deviation of 0.16 watts. On the 3.3 watt rail there was no significant difference in measured power. For the experiments on the node we did not consider completely disabling the HCA. While disabling a redundant link in the fabric may be a feasible approach on future systems, it is not acceptable to disconnect the node completely from the network. In future work we would like to explore additional HCA's such as those from Mellanox. On the Teller cluster there are a limited number of Mellanox HCA's that we wanted to evaluate in addition to the QLogic HCA's. However, when we attempted to change the width or speed of the Mellanox HCA's they were locked in a polling/initialization stage. This could be due to incompatibilities with the QLogic switches.

This page intentionally left blank.

Chapter 4

Conclusions and Future Work

In summary, current network fabric has much too long a delay to apply the rapid adjustments to the network that we see in more recent work [4]. While new technology may change this, we have yet to evaluate the latest offerings from vendors. The incorporation of other techniques such as adaptive routing may create opportunities to find power savings in the network, even with lengthy delays to adjust links. The adjustment delays observed do support static changes to the network on a coarser (per-application) level [10]. However, this assumes that the system workload is not split across a multitude of jobs and applications simultaneously.

The power savings of a dynamic fabric appear to be closer in line with estimates cited in existing literature [3, 4, 1], though QLogic appears to be an exception (offering no power savings for operating at reduced link capability). While [4] assumes link savings of 0.5W per link width, we only found a reduction of 0.18W per lane reduction. While this is a modest power savings, this value is supplemented by the additional 0.19W per lane saved by the HCA.

In future work we would like to:

- try newer firmware for Mellanox switches, which might decrease the delay to adjust fabric links.
- examine other interconnects such as IBM and Cray networks.
- evaluation the incorporation of dynamic fabric with adaptive routing.

This page intentionally left blank.

References

[1] Power savings features in Mellanox products. `http://www.mellanox.com/pdf/whitepapers/WP_ECONET.pdf`. Mellanox, Sunnyvale, CA, USA, accessed: 2014-09-22.

[2] Dennis Abts, Michael R Marty, Philip M Wells, Peter Klausler, and Hong Liu. Energy proportional datacenter networks. In *ACM SIGARCH Computer Architecture News*, volume 38, pages 338–347. ACM, 2010.

[3] Marina Alonso, Salvador Coll, J-M Martinez, Vicente Santonja, Pedro López, and José Duato. Dynamic power saving in fat-tree interconnection networks using on/off links. In *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*, pages 8–pp. IEEE, 2006.

[4] Branimir Dickov, Miquel Pericas, Paul Carpenter, Nacho Navarro, and Eduard Ayguadé. Software-managed power reduction in Infiniband links. In *Proceeding of the 2014 Conference on Parallell Processing, International, ICPP'14*. IEEE Computer Society, 2014.

[5] David Duarte, Yuh-Fang Tsai, Narayanan Vijaykrishnan, and Mary Jane Irwin. Evaluating run-time techniques for leakage power reduction. In *Proceedings of the 2002 Asia and South Pacific Design Automation Conference*, page 31. IEEE Computer Society, 2002.

[6] Torsten Hoefler. Software and hardware techniques for power-efficient hpc networking. *Computing in Science & Engineering*, 12(6):30–37, 2010.

[7] Eun Jung Kim, Ki Hwan Yum, Greg M Link, Narayanan Vijaykrishnan, M Kandemir, Mary Jane Irwin, M Yousif, and Chita R Das. Energy optimization techniques in cluster interconnects. In *Proceedings of the 2003 international symposium on Low power electronics and design*, pages 459–464. ACM, 2003.

[8] Jaeha Kim and M.A Horowitz. Adaptive supply serial links with sub-1-v operation and per-pin clock recovery. *Solid-State Circuits, IEEE Journal of*, 37(11):1403–1413, Nov 2002.

[9] James H Laros, Pavel Pokorny, and David DeBonis. Powerinsight-a commodity power measurement capability. In *Green Computing Conference (IGCC), 2013 International*, pages 1–6. IEEE, 2013.

[10] James H Laros III, Kevin T Pedretti, Suzanne M Kelly, Wei Shu, and Courtenay T Vaughan. Energy based performance tuning for large scale high performance computing systems. In *Proceedings of the 2012 Symposium on High Performance Computing*, page 6. Society for Computer Simulation International, 2012.

[11] Jian Li, Wei Huang, Charles Lefurgy, Lixin Zhang, Wolfgang E Denzel, Richard R Treumann, and Kun Wang. Power shifting in thrifty interconnection network. In *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*, pages 156–167. IEEE, 2011.

[12] Karthikeyan P Saravanan, Paul M Carpenter, and Alex Ramirez. Power/performance evaluation of energy efficient ethernet (eee) for high performance computing. In *Performance Analysis of Systems and Software (ISPASS), 2013 IEEE International Symposium on*, pages 205–214. IEEE, 2013.

[13] Li Shang, Li-Shiuan Peh, and Niraj K Jha. Dynamic voltage scaling with links for power optimization of interconnection networks. In *Proceedings. The Ninth International Symposium on High-Performance Computer Architecture, 2003. HPCA-9 2003.*, pages 91–102. IEEE, 2003.

[14] Vassos Soteriou and Li-Shiuan Peh. Design-space exploration of power-aware on/off interconnection networks. In *Computer Design: VLSI in Computers and Processors, 2004. ICCD 2004. Proceedings. IEEE International Conference on*, pages 510–517. IEEE, 2004.

[15] Ehsan Totoni, Nikhil Jain, and Laxmikant V Kale. Power management of extreme-scale networks with on/off links in runtime systems. 2013.

[16] Ehsan Totoni, Nikhil Jain, and Laxmikant V Kale. Toward runtime power management of exascale networks by on/off control of links. In *2013 IEEE 27th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)*, pages 915–922. IEEE, 2013.

DISTRIBUTION:

1 MS 0899 Technical Library, 9536 (electronic copy)

This page intentionally left blank.


Sandia National Laboratories