Thrifty Interconnection Network for HPC Systems

Jian Li[1] Lixin Zhang[1] Charles Lefurgy[1] Richard Treumann[2] Wolfgang E. Denzel[3]

[1]IBM Austin Research Laboratory, Austin, TX 78758, USA

[2]IBM System and Technology Group, Poughkeepsie, NY 12601, USA

[3]IBM Zurich Research Laboratory, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

Email: [12]{jianli, zhangl, lefurgy, treumann}@us.ibm.com, [3]wde@zurich, ibm.com

ABSTRACT

We propose *Thrifty Interconnection Network (TIN)*, where the network links are activated and de-activated dynamically to save power with little or no overhead by using inherent system events to overlap the link activation or de-activation time. Our simulation results on a set of real world HPC workload traces show on average 35% network power reduction.

Categories & Subject Descriptors: C.2.1 [Computer Systems Organization]: Computer-communication networks – Network architecture and design

General Terms: Design, Experimentation, Performance

Keywords: High-performance computing, interconnection network, power, energy

Power is a critical problem in modern supercomputing systems [1]. A large-scale supercomputer, running High Performance Computing (HPC) workloads, can include hundreds of thousands of processing nodes connected via a large packet-switched interconnection network. A closer look into these systems reveals that the power consumption of interconnection links (including link controllers) constitutes not only a majority of the power of the switches, but also a substantial percentage of the total system power. For instance, the links in an IBM 8-port 12X switch can take 64% of the switch power. The power consumption of the interconnection network in HPC systems can contribute to around 30% the total system power [2]. Current high-speed links in the interconnection networks require continuous pulse transmission to keep both ends synchronized, even when no data is transmitting. Therefore, the average power consumption of such links is almost identical to their worst-case power consumption. Furthermore, network subsystem designers often over-provision the network capacity, with higher power consumption, to meet performance commitments and to avoid network congestion.

We observe that HPC workloads rarely operate all system elements at their maximum capacity simultaneously. For example, infrequent communication patterns exhibited by may HPC workloads allows provisioned network links to stay idle for most of the time. Fig. 1 illustrates a data communication path in an HPC system from a sender compute node (compute node 0) to a receiver compute node (compute node M).

Our approach to saving network power is based on an observation that there can be a significant delay from when a switch sees the first command of a data transfer to when it sends out the first data packet. We propose the *Thrifty Interconnection Network (TIN)*. TIN includes a hardware approach that overlaps inherent system events with link transition delay for significant network power reduction without noticeable network performance overhead. Fig. 2 illustrates a high-level flowchart of providing link services in the thrifty network. Fig. 3 illustrates the hardware support for the thrifty network link management. Note that our proposal differs from prior art in that we do not use prediction. It is also independent on the network topology.

TIN also includes a software extension that uses two software-initiated commands as hints to activate and release links. With this two commands, MPI programs or run time systems can be instrumented to identify longer MPI communication phase to reduce the number of link power state transitions.

We use an in-house simulator called *MARS (MPI Application Relay network Simulator)* to simulate large-scale HPC systems and evaluate our designs [3]. The components of MARS are shown in Fig. 4. MARS has been used in guiding high-level system design, projecting performance, and tuning MPI libraries and applications in the design of a next generation HPC system. We evaluate our techniques with a hierarchical direct interconnect architecture as illustrated in Fig. 5.

The simulated processor model is an abstraction of IBM POWER5 processor [4], with 2GHz cores and corresponding cache memory hierarchy. Commercial power-aware transceivers and switches are not available for experimental measurements. By consulting appropriate literature and industry data sheets, we instead carefully choose three Low-Power Modes, LPM1, LPM2 and LPM3, with increasing power reduction as well as link transition delay. We also consider an Ideal configuration where all idle network components can shut down and power up instantaneously. We use MPI traces of a few popular HPC workloads collected from IBM POWER systems deployed at various client sites to drive the simulations. More details of our design and evaluation can be found in [5].

Fig. 6 shows the average power consumption of these workloads in the network normalized to the original network power without the thrifty network support. On average, TIN achieves 35% network power reduction for these workloads.

We have also studied *Wavefront Power Shifting (WPS)*, which dynamically shifts the total power budget between the compute nodes and the interconnection network that connect them, of which the details are not shown due to limited space. Combining TIN and WPS, our simulation results show 3% system performance improvement and 6% system energy reduction. Further performance improvement is possible if the compute node frequency can speed up more than the assumed 10% and fully utilize the extra power budget reinvested from the thrifty network.

ICS'09, June 8–12, 2009, Yorktown Heights, New York, USA.

ACM 978-1-60558-498-0/09/06.PANELISTS POSITIONS

ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002. Details presented in this paper may be covered by existing patents or pending patent applications.

REFERENCES

[1] W. Tschudi. "LBNL and Government Data Center Programs." *SC'07: Panel Discussion on Power, Cooling and Energy Consumption for Petascale and Beyond*, 2007.

[2] P. M. Kogge. "Architectural Challenges at the Exascale Frontier (invited talk)." *In Simulating the Future: Using One Million Cores and Beyond*, 2008.

[3] W. E. Denzel, J. Li, P. Walker and Y. Jin. "A Framework for End-to-End Simulation of High-Performance Computing Systems." *In Intl. Conf. on Simulation Tools and Techniques for Communications, Networks and Systems (SimuTools)*, Marseille, France, March 2008.

[4] B. Sinharoy, R. N. Kalla, J. M. Tendler, R. J. Eickemeyer, and J. B. Joyner. "POWER5 System Microarchitecture." IBM J. Res. Dev., 49(4/5): 505-521, 2005

[5] J. Li, L. Zhang, C. Lefurgy and E. Elnozahy. "Wavefront Power Shifting via Thrifty Interconnection Networks." *IBM Austin Research Laboratory Technical Report*, 2009.

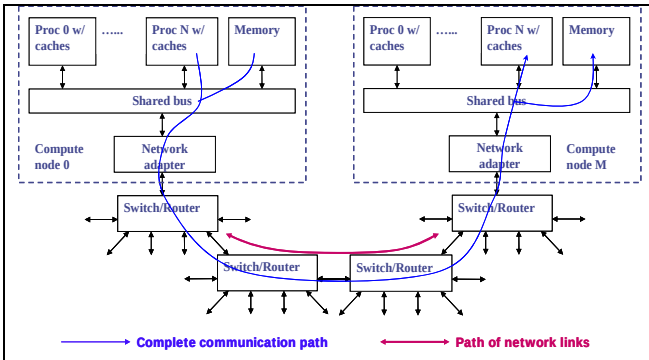

Fig. 1: A data transmission path in an interconnected system.

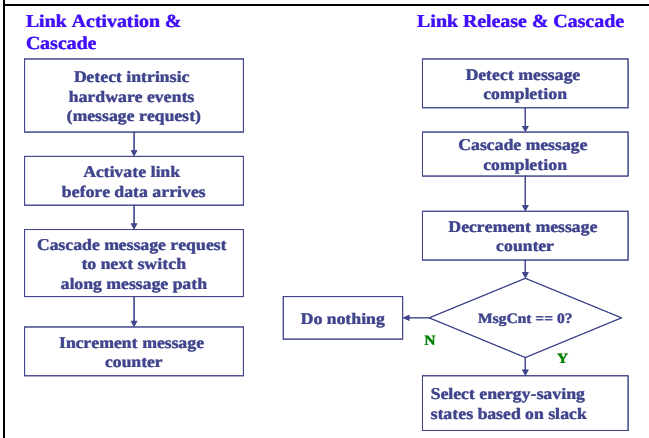

Fig. 2: Link policy for activation (left) and release (right).

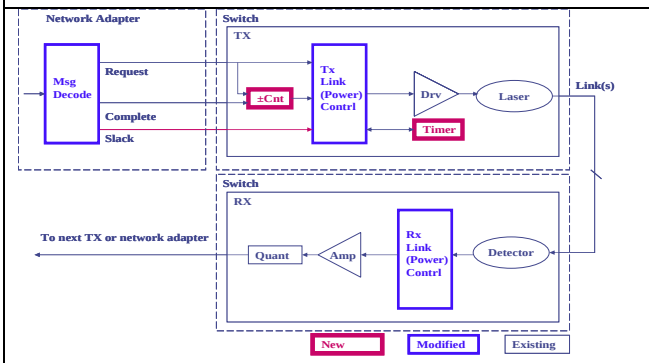

Fig. 3: Hardware support for link management.

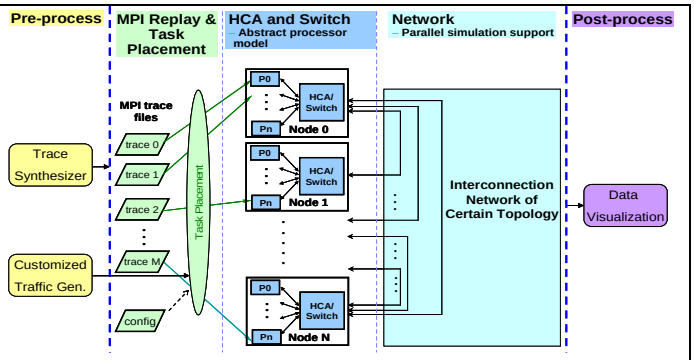

Fig. 4: An end-to-end simulation framework for large-scale systems.

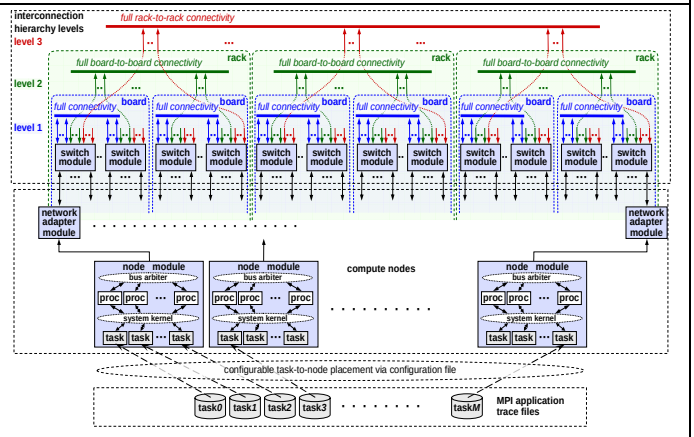

Fig. 5: A hierarchical direct interconnect architecture.

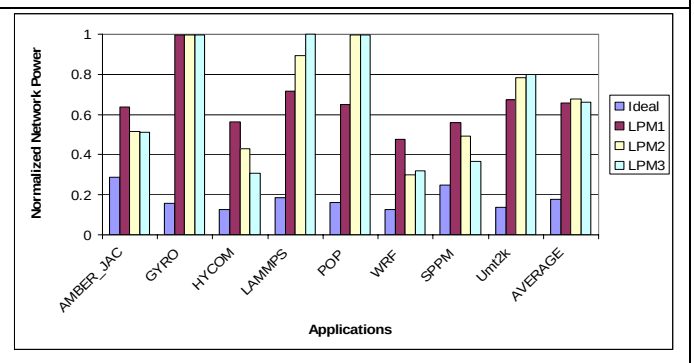

Fig. 6: Average power of the thrifty interconnection network.