Application-aware Collective Communication (extended abstract)

Jiayi Sheng Qingqing Xiong Chen Yang Martin C. Herbordt
Department of Electrical and Computer Engineering; Boston University, Boston, MA

***Abstract*—Preliminary results are presented of hardware support for collective communication that takes advantage of *a priori* routing information.**

***Keywords*-Collective Communication; FPGA Clusters**

I. Introduction

Collective operations including multicast and reduction appear in wide range of applications and often throttle performance; this trend is likely to worsen as the gap between communication and compute capability widens. Here we explore hardware support for collectives that takes advantage of *a priori* routing knowledge. We use a table-based switch; routing decisions are precomputed offline and then loaded into the network. These offline-routing methods are applicable to Molecular Dynamics (MD) and many other compute-intensive applications with predictable communication [1]. Previous work with FPGA [2], [3] and ASIC clusters [4] support collectives with softcores and specialized hardware.

II. Experimental Results

Our target system is the Novo-G#, an FPGA cluster with 128 Gidel Proce V boards [5]. Each board has an Altera Stratix V 5GSMD8 and is interconnected with six Multigigabit Transceivers (MGTs); each MGT has a bandwidth as 40Gbps and latency of around 175 ns [6]. All the boards run at 156.25MHz, which is the same frequency as the MGTs.

We now quantify the advantages of our hardware-based scheme by comparing it with unicast-based implementations. In Figure 1 we present typical results from a 10K particle MD simulation. The x axis represents r/c, which is proportional to the network diameter. Performance with increasing r/c is therefore a measure of strong scaling. We find that tree-based multicast has half of the latency of the unicast-based multicast and that this difference increases with r/c. Now looking at reduction, we see that tree-based reduction does not show much advantage when the number of nodes is small (or the number of particles large). As the number of nodes gets bigger, however, the tree-based reduction shows better latency until it has about three times the performance. The delay in benefit is because the tree-based reduction requires that the packet be allowed to be ejected only when all the related downstream packets have arrived; this introduces extra blocking in the network. There is no analogous problem in unicast-based reduction.

This work was supported in part by the NSF through award #CNS-1405695. Email: (jysheng|qx|cyang90|herbordt)@bu.edu

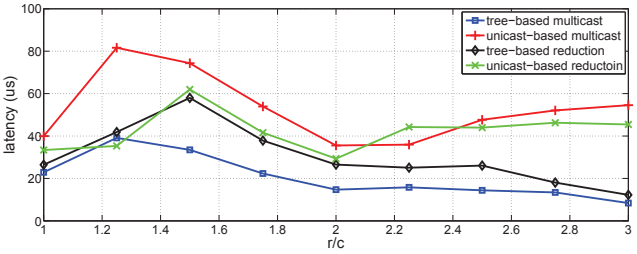

Figure 1. Comparison of latency between tree-based and unicast-based collective operations for a 10K particles MD simulation

Resource utilization is modest. Each switch contains seven routers which in turn each contains one multicast unit, one reduction unit, and a small amount of control logic. Tree-based reduction requires more memory than unicast-based because each particle on the same home node needs to have a different table index on each intermediate node.

III. Discussion and Future Work

The major contribution in this abstract is that our design and results demonstrate that a FPGA cluster is able to provide effective application-aware collective operations with trivial extra hardware cost. By precomputing the routing tables, we save latency and hardware costs over dynamic routing. The application-aware table-based routing scheme should also enable optimization of routing congestion offline; this is the focus of current work.

References

[1] J. Sheng, B. Humphries, H. Zhang, and M. Herbordt, "Design of 3D FFTs with FPGA Clusters," in *Proc. IEEE Conf. High Performance Extreme Computing*, 2014.

[2] A. Schmidt, W. Kritikos, S. Gao, and R. Sass, "An Evaluation of an Integrated On-chip/Off-chip Network for High-performance Reconfigurable Computing," *Int. J. Reconfig. Comput.*, vol. 2012, 2012.

[3] Y. Peng, M. Saldana, C. Madill, X. Zou, and P. Chow, "Benefits of Adding Hardware Support for Broadcast and Reduce Operations in MPSoC Applications," *ACM Trans. Reconfigurable Technology and Systems*, vol. 7, 2014.

[4] J. Grossman, B. Towles, B. Greskamp, and D. Shaw, "Filtering, reductions and synchronization in the Anton 2 network," in *Proc. Int. Par. and Distr. Processing Symp.*, 2015.

[5] A. George, M. Herbordt, H. Lam, A. Lawande, J. Sheng, and C. Yang, "Novo-G#: A Community Resource for Exploring Large-Scale Reconfigurable Computing Through Direct and Programmable Interconnects," in *Proc. IEEE Conf. High Performance Extreme Computing*, 2016.

[6] J. Sheng, C. Yang, and M. Herbordt, "Towards Low-Latency Communication on FPGA Clusters with 3D FFT Case Study," in *4th Int. Symp. Highly Efficient Accelerators and Reconfigurable Technologies*, 2015.