High-Performance Fluid Simulation using Multiple FPGAs with Bandwidth-Compressed Links

Kentaro Sano, Tomohiro Ueno, Daichi Tanaka, and Satoru Yamamoto
Tohoku University, Sendai, JAPAN
{kentah,ueno,tanaka,yamamoto}@caero.mech.tohoku.ac.jp

1. INTRODUCTION

Nowadays, higher performance per power is desired in super computing and big-data processing more than ever. One of the devices promising for both high performance and low power computing is FPGA (Field-Programmable Gate Arrays), which can be utilized to construct application-specific custom hardware for efficient processing with limited resources such as an external memory bandwidth. As reported in [1], dedicated hardware allows an FPGA to perform numerical simulation at a higher sustained performance and a lower power than those of GPU-based implementation. Furthermore, high-speed serial I/Os of FPGAs are very attractive in obtaining higher performance by directly clustering FPGAs. Recently, some large data centers have been installed with multiple FPGAs.

To scale computing performance with multiple FPGAs, a bandwidth of inter-FPGA communication is important. Although present FPGAs have a lot of serial transceiver channels with tens of Gbps, their total bandwidth is still insufficient for a bandwidth of external DDR memories. We rely on lossless data compression to address this problem. By applying our bandwidth compression hardware [2], we enhance communication bandwidth for floating-point data streams in computation. Contributions of this work are:

1. Enhancing the link bandwidth with data compression,
2. Evaluation with FPGA-based fluid simulation,
3. Performance estimation for 16 Stratix V FPGAs.

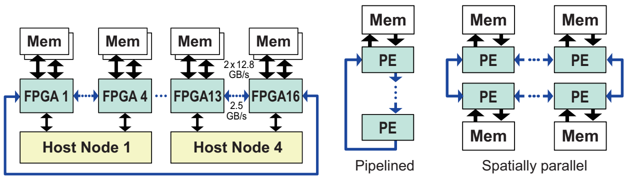
2. DESIGN AND RESULTS

We have developed an FPGA cluster with sixteen StratixV 5SGXA7 FPGAs. As shown in Fig.1a, the cluster is composed of four host nodes, each of which has four FPGA boards. Each FPGA board has four SFP+ ports with full duplex of 10.3125 Gbps. We form a ring-topology connection of the FPGAs by bunching up two SFP+ ports to form a link between two boards. With each FPGA, we implement four cascaded processing elements (PEs) for fluid simulation based on the lattice Boltzmann method (LBM). The four cascaded PEs operate as a computing pipeline at 200 MHz to take an input stream and output a data stream as computational results for four time steps. To use multiple FPGAs in parallel, we can apply either A) a pipelined or B) a spatially-parallel approach, as shown in Fig.1b. Although B) is good at an aggregate bandwidth of distributed memories, we are adopting A) due to its simplicity and localization of the computational data to a single memory.

H2RC Novemver 14, 2016, Salt Lake, UT, USA

ACM ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235



a. FPGA cluster with 1D ring. b. Available parallelism.

Figure 1: FPGA cluster and available parallelism.

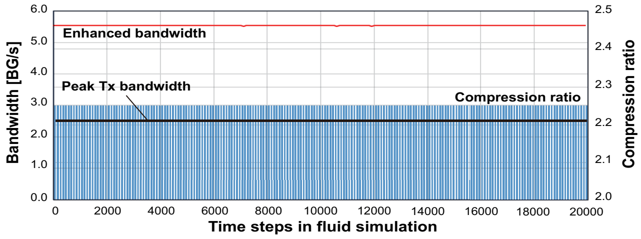


Figure 2: Peak Tx bandwidth, enhanced bandwidth, and compression ratio.

However, in the case of A), the bandwidth of the inter-FPGA link with two 10.3125 Gbps lanes, 2.5 GB/s, is insufficient to the bandwidth required by the computing pipeline, which is 8.0 GB/s. This limits the performance of the four PEs to 32.8 GF from their peak of 104 GF per FPGA. To solve this problem, we apply our prediction-based bandwidth compressor [2] to enhance the link's bandwidth. As Fig.2 shows, the compressor enhances it to about 5.5 GB/s with the compression ratio of 2.25, alleviating the bottleneck. With the results, we expect the improved performance of 74 GF/s per FPGA, and its linearly scaled performance of 1.18 TF/s with 16 FPGAs. Since there is still room for improvement of the compression ratio up to 4.0 with encoding optimization, we expect that the bandwidth-compressed link can achieve about 9 GB/s at most for 1.9 TF/s with 16 FPGAs.

3. REFERENCES

[1] K. Nagasu, K. Sano, F. Kono, and N. Nakasato. Performance and power evaluation of FPGA-based tsunami simulator using floating-point DSPs. Proceeding of COOL Chips XIX, April 2016.

[2] T. Ueno, R. Ito, K. Sano, and S. Yamamoto. Bandwidth compression of multiple numerical data streams for high performance custom computing. Proceedings of the International Conference on Application-specific Systems, Architectures and Processors, pages 190–191, June 2014.