Strong Scaling of MD Simulations with FPGA-Centric Clouds and Clusters (Extended Abstract)*

Jiayi Sheng Chen Yang Ahmed Sanaullah Qingqing Xiong Martin C. Herbordt
Dept. of Electrical and Computer Engineering, Boston University

ABSTRACT

We combine extensive preliminary results with analytical models to show that FPGA-centric clouds and clusters can fill an important niche for Molecular Dynamics simulations between GPU/CPU clusters and ASICs.

Keywords

FPGAs; Molecular Dynamics

1. EXTENDED ABSTRACT

In previous work we have shown that Molecular Dynamics simulations on FPGAs are highly competitive per chip with other COTS architectures [1]. Since then, there has been an emergence of FPGA-centric parallel systems with direct and programmable interconnects (DPI), e.g., the 128-node Novo-G# cluster [2] and the large-scale deployment of FPGAs in the Cloud by Microsoft. DPI facilitates highly efficient communication through a number of mechanisms. These include bypassing layers of hardware and software and enabling intelligent communication through, e.g., compute-in-the-network and statically scheduled routing. We have demonstrated DPI in two domains: the 3D FFT [3] and generic Collectives [4].

In recent work we have also ported the 3D FFT to a large-scale FPGA-centric cloud. These results, together with those for the Novo-G# and reference architectures, are shown in Table 1. The critical result here is that while the FPGAs in the Cloud have substantially higher communication latency than the Novo-G#, as well as lower bandwidth per node, the performance is still sufficient to enable high performance Molecular Dynamics. Results from our analytical model, calibrated with the results cited, are shown in Figure 1.

Note that the projected results fall between the 1μs/day that is the typical maximum of COTS clusters and the 100μs/day achieved by Anton 2. The significance is that FPGA-based systems can potentially be cost-effective even in traditional HPC domains.

Table 1: 3D FFT – times in μs.

FFT Size	Cloud	Novo-G#	CPU	GPU	ASIC
16^3	21.5	3.86	22	25	NA
32^3	29.3	5.30	55	29	4
64^3	47.4	9.3	288	92	13
128^3	104.5	101.1	2000	NA	NA

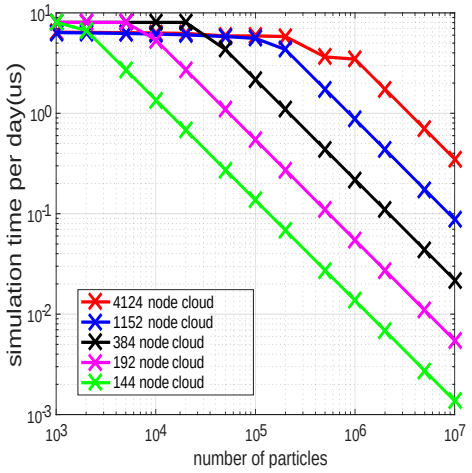

Figure 1: Simulated time per day as a function of # of particles for clouds with various #s of nodes.

2. REFERENCES

[1] M. Chiu and M. Herbordt. Molecular dynamics simulations on high performance reconfigurable computing systems. *ACM Trans. Reconfigurable Tech. and Sys.*, 3(4):1–37, 2010.

[2] 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 *IEEE High Perf. Extreme Computing Conf.*, 2016.

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

[4] J. Sheng, C. Yang, and M. Herbordt. Collective Communication on FPGA Clusters with Static Scheduling. In *Proc. Highly Efficient and Reconfigurable Technologies*, 2016.

*This work was supported in part by the NSF through awards #CNS-1405695 and #CCF-1618303/7960, and by a grant from Microsoft Research. Email: herbordt@bu.edu

H2RC'16 Salt Lake City, Utah USA

DOI: 10.1145/1235