Reconfigurable Computing Fabrics: Roles in Cloud Computing, Data Analytics, and Machine Learning

Ronald F. DeMara
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, Florida 32816-2362
E-mail: **ronald.demara@ucf.edu**

March 14, 2019

***Abstract* — A panel session was held during the 24th Annual IEEE Reconfigurable Architectures Workshop (RAW) on Monday, 29 May 2017. This document conveys this panelist's, Ronald F. DeMara, position statement in a concise format. Discussions have been scoped and leveled to provide a broad perspective of technologies related to the topic. Thus, this summary also discusses future directions in FPGAs within the domains of cloud-based processing, large-scale data analytics, and unsupervised machine learning methods.**

Keywords* — *FPGAs, Cloud Computing, IEEE Reconfigurable Architectures Workshop (RAW), Data Analytics, Technical Panel Discussion.

1.0 Introduction

In this paper, the primary focus is to identify the author's perspective regarding the roles in cloud computing, data analytics, and machine learning as a technical panelist at IEEE RAW 2017. The panel, and thus this summary document, considers two roles of reconfigurable computing in the cloud. A top-down perspective is based on developer's perspective of the demand for computational resources. A bottom-up perspective is rooted in Reconfigurable Computing (RC) fabric features that may be accessed via the cloud, either currently or in the future.

2.0 Roles of RC in Cloud

Two cases for utilizing Reconfigurable Computing (RC) within the Cloud are: 1) motivated by extending hardware capabilities upward to a cloud-abstraction, and 2) Motivated by extending programming models downward. These are enumerated as:

1) How can cloud computing applications leverage the unique features of reconfigurable fabrics?
→ bottom up: fabric features (which ones?) → enhance cloud (which ways?)

2) How can developers of RC products/services access vast ensembles of FPGAs that will reside in the cloud?
→ top down view: developer access → how to enable it? (MapReduce programming model? Or other model)

Each of these cases is depicted as a flow within Figure 1.

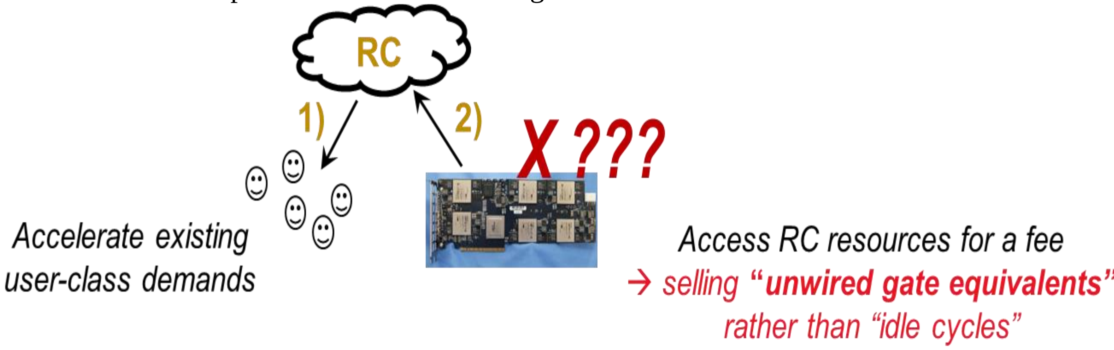

Figure 1: Two roles for reconfigurable computing within a cloud computing environment.

3.0 Microsoft's Project Catapult

Microsoft's Hyperscale Acceleration Goals: [1]

- "Post-CPU technologies in the cloud" – FPGA, GPU, custom ASIC
- Data Analytics and AI applications despite diminishing CPU improvements

Facilities and Results: [1]

- Deal with diminishing CPU improvements
- FPGA Accelerator based on Altera Stratix V D5 FPGA172k Adaptive Logic Modules
- Elastic reconfigurable acceleration fabric: individual to 1000s FPGAs
- 10x throughput for 1.3x expense and 1.1x energy increase: 40GOPS/W

Timeline: [1]

2012: Pilot of FPGA boards in 1,632 servers

2013: ROI demonstrated: reduced latency ranking; required servers by half

2016: "Configurable Cloud" architecture in nearly every new production server

2017: Project Catapult Academic Program: submit 1-page proposal to catapult@microsoft.com

Perspectives:

- Data centers/HPC exhibit a dichotomy/ tradeoff between energy and resilience [2][9][16]
- Autonomic sustainability for large fabric ensembles [3][4][5][7][15][20]
- RC can help both in ways that CPU/GPU cannot [3]
- What are those ways; what middleware needed? [6]
- Future fabric architecture [10][11][12][13][14]
- RC using emerging devices [8][17][18][19]

4.0 Amazon AWS FPGA Instances

Based on the Amazon AWS FPGA platform, several challenges are highlighted: [21]

- Security: Intellectual property concerns
- Access cost: pricing plan still evolving
- Programming model: what is needed

5.0 Perspectives

This panelist's and author's perspective would be ground-up coming from the device aspects: Thus, it is interesting to consider features that RC can offer. It is possible to identify four areas:

1) bloat-free/energy savings

2) autonomous adaptation to workloads at circuit level vs software means, which types of adaptation at what times

3) data centers and HPC there is a dichotomy and tradeoff between energy and resilience. RC can help both in unique ways that CPU/GPU cannot. What are those ways, and what middleware is needed?

4) new computing paradigms such as Spin Orbit Torque Based Neural FPGA or Cellular Neural Network can impact deep learning in future.

5.0 Recent Developments

Commercialization continues with Xilinx now actively advocating *Reconfigurable Acceleration in the Cloud.* Their website describes use of Xilinx UltraScale and UltraScale+ FPGAs for hardware and application developers of various cloud computing services. To utilize these, tutorials such as "Amazon EC2 F1 SDAccel Developer Lab" which go over as a self-paced tutorial the Amazon AWS SDAccel. It is delivered as a stepwise procedure how to employ Amazon EC2 F1 instances to accelerate applications. There is a virtual developer lab spanning connecting to F1 instances, accelerating F1 results, and optimizing via SDAccel [22] and general Custom Computing Machine (CCM) virtualizations have been proposed [23].

6.0 Conclusion

Popularization should continue and increase as FPGAs in the cloud grow more widely into teaching curriculum [24]. Final questions facing Roles of Reconfigurable Computing in Cloud Computing and Data Analytics, center around "What would you use such services for? … And most importantly … why." The answer to that question hints to us what will be needed in the future within this space.

References

[1] Microsoft Corporation, "Project Catapult Academic Program," Available at: http://aka.ms/catapult-academic

[2] N. Imran, R. Ashraf, and R. F. DeMara, "Evaluating Quality and Resilience of an Embedded Video Encoder against a Continuum of Energy Consumption," invited submission to *2014 Workshop on Suite of Embedded Applications and Kernels*, San Francisco, California, USA, June 1, 2014.

[3] N. Imran, R. Ashraf, and R. F. DeMara, "On-demand Fault Scrubbing Using Adaptive Modular Redundancy," in *Proceedings of the International Conference on Engineering of Reconfigurable Systems and Algorithms*, Las Vegas, Nevada, USA, July 22 – 25, 2013.

[4] R. Al-Haddad, R. Oreifej, R. A. Ashraf, and R. F. DeMara, "Sustainable Modular Adaptive Redundancy Technique Emphasizing Partial Reconfiguration for Reduced Power Consumption," *International Journal of Reconfigurable Computing*, Article ID 430808, June, 2011, pp 1 – 25, 2011.

[5] N. Imran and R. F. DeMara, "Distance-Ranked Fault Identification of Reconfigurable Hardware Bitstreams via Functional Input," *International Journal of Reconfigurable Computing*, vol. 2014, pp. 1 – 21, March 2014. DOI: 10.1155/2014/279673

[6] H. Tan and R. F. DeMara, "A Multi-layer FPGA Framework Supporting Autonomous Partial Runtime Reconfiguration," *IEEE Transactions on VLSI Systems*, Vol. 16, No. 5, May, 2008, pp. 504 – 516.

[7] R. A. Ashraf and R. F. DeMara, "Scalable FPGA Refurbishment using Netlist-driven Evolutionary Algorithms," *IEEE Transactions on Computers*, vol. 62, no. 8, pp. 1526 – 1541, August 2013. DOI:10.1109/TC.2013.58

[8] Ramtin Zand, Kerem Yunus Camsari, Ibrahim Ahmed, Steven D. Pyle, Chris H. Kim, Supriyo Datta, Ronald F. DeMara, "R-DBN: A Resistive Deep Belief Network Architecture Leveraging the Intrinsic Behavior of Probabilistic Devices," ArXiv e-prints, September 2017, e-print: 1710.00249, http://adsabs.harvard.edu/abs/2017arXiv171000249Z

[9] R. Oreifej, R. Al-Haddad, R. A. Ashraf, R. Zand, and R. F. DeMara, "Survivability Modeling and Resource Planning for Self-Repairing Reconfigurable Device Fabrics," *IEEE Transactions on Cybernetics*, in-press, accepted 23 August 2016.

[10] R. F. DeMara, M. Platzner, and M. Ottavi, "Guest Editorial: IEEE Transactions on Computers and IEEE Transactions on Emerging Topics in Computing Joint Special Section on Innovation in Reconfigurable Computing Fabrics from Devices to Architectures," *IEEE Transactions on Computers*, Vol. 66, No. 6, pp. 927 – 929, June 2017. DOI: 10.1109/TC.2017.2689118

[11] R. F. DeMara, M. Platzner, and M. Ottavi, "Guest Editorial: IEEE Transactions on Computers and IEEE Transactions on Emerging Topics in Computing Joint Special Section on Innovation in Reconfigurable Computing Fabrics from Devices to Architectures," *IEEE Transactions on Emerging Topics in Computing*, Vol. 5, No. 2, pp. 207 – 209, April - June 2017. DOI: 10.1109/TETC.2016.2641599

[12] R. F. DeMara, A. Roohi, R. Zand, S. D. Pyle, "Heterogeneous Technology Configurable Fabrics for Field Programmable Co-design of CMOS and Spin-based Devices," in *Proceedings of IEEE International Conference on Rebooting Computing (ICRC-2017)*, Washington, DC, USA, November 8 – 9, 2017.

[13] Y. Bai, S. Hu, R. F. DeMara, and M. Lin, "A Spin-Orbit Torque based Cellular Neural Network (CNN) Architecture," in *Proceedings of 27th IEEE/ACM Great Lakes Symposium on VLSI (GLSVLSI-2017)*, Banff, Alberta, Canada, May 10 – 12, 2017.

[14] R. F. DeMara, "Heterogeneous Technology Configurable Fabrics: Leveraging Reconfiguration as a Pathway Towards Emerging Devices," *in Proceedings of IEEE Reconfigurable Architectures Workshop (RAW-2017)*, Orlando, FL, USA, May 29, 2017.

[15] A. Al-Zahrani and R. F. DeMara, "Hypergraph-Cover Diversity for Maximally-Resilient Reconfigurable Systems," in *Proceedings of 12th IEEE International Conference on Embedded Software and Systems (ICESS-2015)*, pp. 1086 – 1092, New York, NY, USA, August 24 – 26, 2015.

[16] F. Alghareb, R. A. Ashraf, and R. F. DeMara, "Designing and Evaluating Redundancy-based Soft Error Masking on a Continuum of Energy versus Robustness," accepted to *IEEE Transactions on Sustainable Computing* on 20 September 2017.

[17] M. Krishna, R. Zand, A. Roohi, and R. F. DeMara, "Heterogeneous Energy-Sparing Reconfigurable Logic: Spin-based Storage and CNFET-based Multiplexing," *IET Circuits, Devices, and Systems, (IEEE-indexed)*, Vol. PP, No. PP, pp. 1-8, Accepted 11 January 2017.

[18] R. Zand, A. Roohi, D. Fan and R. F. DeMara, “Energy-Efficient Nonvolatile Reconfigurable Logic using Spin Hall Effect-based Lookup Tables,” *IEEE Transactions on Nanotechnology*, Vol. 16, No. 1, pp. 32 - 43, January 2017.

[19] R. Zand, A. Roohi, S. Salehi, and R. F. DeMara, “Scalable Adaptive Spintronic Reconfigurable Logic using Area-Matched MTJ Design,” *IEEE Transactions on Circuits and Systems II*, Vol. 63, No. 7, pp. 678-682, July 2016.

[20] A. Al-Zahrani and R. F. DeMara, “Fast Online Diagnosis and Recovery of Reconfigurable Logic Fabrics using Design Disjunction,” *IEEE Transactions on Computers*, Vol. 65, No. 10, pp. 3055-3069, October 2016.

[21] Frederic Lardinois, “AWS announces FPGA instances for its EC2 cloud computing services,” TechCrunch, November 30, 2016, https://techcrunch.com/2016/11/30/aws-announces-fpga-instances-for-its-ec2-cloud-computing-service/

[22] Xilinx Inc. “Reconfigurable Acceleration in the Cloud” available at: https://www.xilinx.com/products/design-tools/cloud-based-acceleration.html accessed 25 November 2018.

[23] Amran A. Al-Aghbari, and Muhammad ES Elrabaa. "Cloud-Based FPGA Custom Computing Machines for Streaming Applications." *IEEE Access* (2019).

[24] Ke Zhang, Yisong Chang, Mingyu Chen, Yungang Bao, and Zhiwei Xu. "Computer Organization and Design Course with FPGA Cloud." In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pp. 927-933. ACM, 2019.