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.](image)

**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].

6.0 Conclusion
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


