A New Era of Hardware Microservices in the Cloud

Doug Burger

Microsoft Research, US

Abstract:

The Cloud is causing a major shift in both the business ecosystem and system infrastructures. The major hyperscale providers are building out highly-interconnected, worldwide computers at a scale that allows them to make significant first-party investments. This verticalization allows them to make cross-layer architectural changes more rapidly than would the old horizontal model. A second trend is the emergence of ultra-low latency requirements in the Cloud, moving storage, networking, and services from the millisecond to the microsecond regime. In this talk, I will describe how these architectural shifts are enabling the emergence of specialized hardware in datacenters, that enable services to be operated in the microsecond regime. On FPGAs, GPUs, and ASICs, these services can run with no CPU intervention, allowing much lower latencies and better cost structures than previously possible for key services such as deep learning. Over time this transition will enable a much broader collection of hardware IP to run at scale in the Cloud.

978-3-9815370-8-6/17/$31.00 ©2017 IEEE