Cloud Cover

Error-Resilient Server Ecosystems for Edge and Cloud Datacenters

Georgios Karakonstantis and Dimitrios S. Nikolopoulos, **Queen's University Belfast**

Dimitris Gizopoulos, **National and Kapodistrian University of Athens**

Pedro Trancoso and Yiannakis Sazeides, **University of Cyprus**

Christos D. Antonopoulos, **University of Thessaly**

Srikumar Venugopal, IBM Research

Shidhartha Das, ARM Research

The explosive growth of Internet-connected devices forming the Internet-of-Things and the flood of data they yield require new energy-efficient and error-resilient hardware and software server stacks for next-generation cloud and edge datacenters.

In the past few decades, aggressive miniaturization of semiconductor circuits has driven the explosive performance improvement of digital systems that have radically reshaped the way we work, entertain, and communicate. At the same time, new paradigms such as cloud and edge computing, as well as the Internet of Things, enable billions of devices to interconnect intelligently. These devices will generate huge volumes of data that must be processed and analyzed in centralized or decentralized datacenters located close to users. The analysis of these data could lead to new scientific discoveries and new applications that will improve our lives [IoT 2014].[1]

Such advances are at risk, however, because ongoing technology miniaturization appears to be ending, as identical nanoscale circuits are getting more likely to exhibit different behavior, operate slower or burn more power, even though that are designed using the same processes and architecture. Such variations in the behavior of identical circuits are caused by imperfections in the manufacturing process that are magnified as circuits are getting minute.

In fact, a situation where performance and power variations are not mitigated will result in many fabricated chips not meeting their intended performance and power specifications thus endangering the correct functionality of products that use them. Manufacturers try to deal with the huge performance and power variability in the fabricated chips and hide it from the software layers by adopting pessimistic safety timing margins and redundant error-correction schemes that are designed to counteract the worst possible scenario that may arise.[2][IEEE 2010]. In reality, such measures are extremely pessimistic since they are determined based on i) rare worst case operating conditions that are assumed at design time and ii) the capabilities of the worst performing chips are remarkably inferior as compared

to the vast majority of the identically manufactured circuits. The result of such pessimistic measures is that the majority of the chips are being constrained to operate at the low speed and high power of the worst case chips and not on the speed and power that they could really achieve .[3][ISSCC 2015] Put it in another way, few outliers dictate the way the entire population is treated.

The use of pessimistic timing margins—along with the fact that the supply voltage (the most efficient way to save power) cannot be scaled down easily anymore in nanoscale circuits since it makes circuits even more prone to failures has elevated the significance of energy efficiency even more [AHC 2011].

Reducing processors' power consumption could not only allow to meet the tight power budgets of many products but also let users improve performance by employing more resources or by operating the chips at a higher frequency [4][ISCA 2011]. This would be particularly important for servers, which will soon have to handle huge amounts of data that are going to be generated by the increasing number of interconnected devices, estimated to reach 24.3 exabytes per month in 2019 -.[5][CISCO 2017].

Improvements Require New Design Approaches

Substantially improving energy-efficiency requires new types of error-resilient server ecosystems that can handle the increased power and performance variability of the hardware components more intelligently than conventional pessimistic paradigm of a very large size can fit all.

Admitting variability as a fact of life rather than something that should be hidden away, is tantamount to transforming each identically manufactured processor and memory module to an individual one that is inherently different in terms of the performance that it can achieve (see Figure 1). The computing industry should see such heterogeneity not as a problem but as an opportunity to improve energy efficiency by avoiding to artificially constrain the performance of all chips based on few misbehaving outliers but rather allowing each chip to operate according to its true capabilities. Exploiting such heterogeneity requires shifting away from current approaches, and redesigning the hardware and system software of next generation servers.

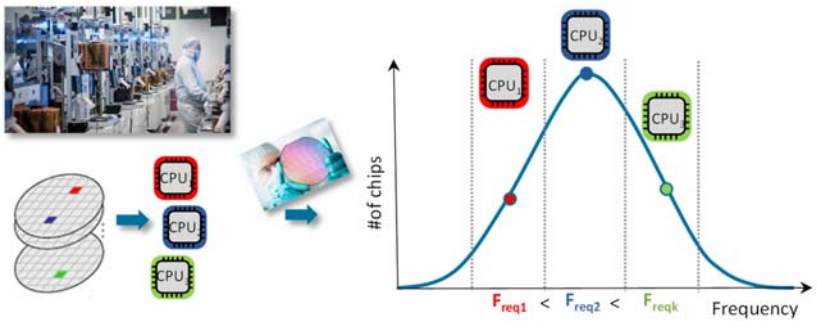

Figure 1. Identical chips, in this case Central Processing Units (CPUs) may have substantially different performance characteristics e.g. in terms of frequency of operation (Freq), even if they were designed and manufactured using the same processes.

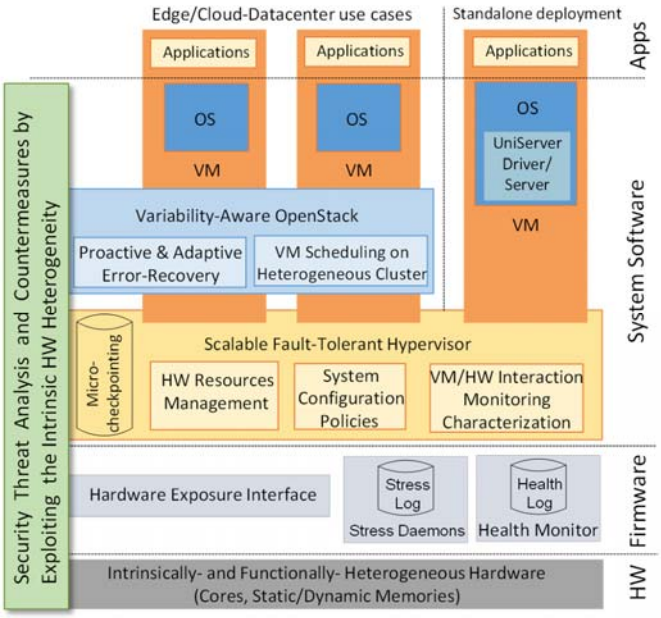

Figure 2. The envisioned server ecosystem spans all layers of the systems stack end enhances it with technologies for monitoring the hardware 'health', while exploiting operation of various edge and cloud applications at non-conventional pessimistic points.

Server Ecosystem

Making the most of heterogeneity requires automated firmware-level procedures to expose each processor's and memory resource's capabilities, even as those capabilities change over time. This requires the embedding of diagnostic and health-monitoring daemons' in the firmware of any server that evaluates hardware components' operations during the product lifetime (see Figure 2). These 'daemons'/monitors would access on-chip sensors and error-detection circuitry to collect, mine, and analyze various parameters, such as correctable and uncorrectable errors, performance counters, system crashes and hangs, and thermal and power behavior. This would be similar to the procedures that the machine-check architecture in x86 systems has adopted (www.mcelog.org).

The new system would use an enhanced hardware exposure interface (HEI) to communicate the hardware-related data it collects to the software stack. The latter would identify energy-efficient voltage, frequency, and refresh-rate states for processors and memory sub-systems.

We should also rethink the design of all system-software layers—including hypervisors and resource management frameworks such as OpenStack—used in today's datacenters. These layers should be prepared to operate hardware aggressively close to its performance and power limits. Hypervisors, for example, can use this capability to allocate processor and memory resources with different reliability, power and performance efficiency characteristics to virtual machines, so that they improve performance, while reducing energy consumption on a server. Cloud management frameworks can leverage the same capability to provide better Quality of Service to users and applications. On the other hand,

hardware operation outside its normal safety margins may introduce critical errors in system software. On a standard server such errors would immediately bring the server down. New approaches to system software resilience are needed to avoid or mitigate such errors and sustain high server availability. At the datacenter level, this implies avoiding degrading QoS of a customer workload in the event of a server crash, and adopting mechanisms for speeding up the recovery from any server crash.

.

Empowering Internet Evolution

Manufacturers could integrate our proposed ecosystem into standard high-end servers and the newly introduced microserver platforms. Microservers don't perform as well as mainstream servers yet, but they can service many types of application requests at a "just right" performance and with significantly less power consumption. Integrating our envisioned software and hardware ecosystem into servers would help power next-generation datacenters in the cloud and at the network edge, where energy efficiency is particularly critical for minimizing power supply, cooling, and maintenance costs.

Energy-efficient servers would help create a more sustainable Internet. Presently, most Internet processing and storage takes place in the cloud, in massive centralized datacenters that contain tens of thousands of servers, consume as much electricity as a small city, and utilize expensive cooling mechanisms. This won't be practical in the IoT era due the limited network capacity of the current Internet infrastructure that won't be able to accommodate the exabytes of data that are soon going to be generated by all the internet-connected devices. However, using these typical centralized datacenters along with new de-centralized datacenters at the network edge, closer to users, could help limit the load put on the Internet infrastructure by allowing pre-processing of the data and selective forwarding of some of them to the Cloud. Such a new paradigm, referred to as Edge or Fog computing is believed to be more viable than the current Cloud paradigm and is being promoted by major companies such as Cisco, Huawei, IBM, and Intel for transforming the next-generation Internet.[5][IOTJ 2016].

Finally, edge resources' ability to provide all necessary services within a home or small business improves privacy because the data they carry doesn't have to travel through the public network or reside in third-party datacenters.

Realizing our proposed error-resilient, energy-efficient ecosystem faces many challenges, in part because it requires the design of new technologies and the adoption of a system operation philosophy that departs from the current pessimistic one.

The UniServer Consortium (www.uniserver2020.eu)—consisting of academic institutions and leading companies such as AppliedMicro Circuits, ARM, and IBM—is working toward such a vision. Its goal is development of a universal system architecture and software ecosystem for servers used for cloud- and edge-based datacenters. The European Community's Horizon 2020 research program is funding UniServer.

The consortium is already implementing our proposed ecosystem in a state-of-the art X-Gene2 eight-core, ARMv8-based microserver with 28nm feature sizes. The initial characterization of the processing cores of that server shows that there is a significant safety

margin in the supply voltage used for operating each of the cores . Results show that the supply-voltage in some 'good' cores can be reduced by 10 percent below the nominal that is actually being enforced to all cores by the manufacturer. Allowing operation of the 'good' cores at such reduced supply-voltage could lead up-to 38 percent power savings. [MICRO 2017]

Similarly promising is the characterization of the DRAM memories that are used on board the mentioned ARMv8-based microserver by 43x times and 5 percent, respectively. Such reductions can lead to an average power savings of more than 22 percent across a range of benchmarks. [IOLTS 2017]

References

1. [IoT 2014] H. Bauer, M Patel, and J. Veira, "The Internet of Things: Sizing up the Opportunity," online report, McKinsey & Co., December 2014; www.mckinsey.com/industries/semiconductors/our-insights/the-internet-of-things-sizing-up-the-opportunity
2. [CISCO 20174… s/b 5] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, online white paper, Cisco Systems, March 2017; www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html
3. [AHC 2011] H. Wong et al, "Implications of Historical Trends in the Electrical Efficiency of Computing," in *IEEE Annals of the History of Computing*, vol. 33, no. 3, 2011, pp. 46–54.
4. [ISCA 2011 s/stay 4] H. Esmaeilzadeh et al, "Dark Silicon and the End of Multicore Scaling," *Proc. 38th IEEE Int'l Symp. Computer Architecture (ISCA 11)*, 2011, pp. 365–376.
5. [IEEE 2010 s/b 2] S. Ghosh and K. Roy, "Parameter Variation Tolerance and Error Resiliency: New Design Paradigm for the Nanoscale Era," *Proc. IEEE*, volume 98, no. 10, 2010, pp. 1718–1751.
6. [ISSCC 2015 s/b 3] P.N. Whatmough et al, "An All-Digital Power-Delivery Monitor for Analysis of a 28nm Dual-Core ARM Cortex-A57 cluster," *Proc. 2015 IEEE Solid-State Circuits Conf.* (ISSCC 15), 2015, pp. 1–3.
7. [IOTJ 2016 s/b 5] W. Shi et al, "Edge Computing: Vision and Challenges," *IEEE Internet of Things J.*, vol 3, no. 5, 2016, pp. 637–646.
8. [MICRO 2017] G. Papadimitriou et al, "Harnessing Voltage Margins for Energy Efficiency in Multicore CPUs," *Proc. 50th IEEE/ACM Int'l Symp. Microarchitecture* (MICRO 17), 2017, pp. 503–516.
9. [IOLTS 2017] K. Tovletoglou, D. Nikolopoulos, and G. Karakonstantis, "Relaxing DRAM Refresh Rate through Access Pattern Scheduling: A Case Study on Stencil-Based Algorithms", *Proc. 23rd IEEE Int'l Symp. Online Testing and Robust System Design* (IOLTS 17), 2017, pp. 45-50. .

Georgios Karakonstantis is an assistant professor in the School of Electronics, Electrical Engineering, and Computer Science (EEECS) at Queen's University Belfast and the scientific coordinator of the UniServer project. Contact him at g.karakonstantis@qub.ac.uk.

Dimitrios S. Nikolopoulos is a professor and the Head of the School of EEECS at Queen's University Belfast. Contact him at d.nikolopoulos@qub.ac.uk.

Dimitris Gizopoulos is a professor at the Department of Informatics and Telecommunications at the National and Kapodistrian University of Athens, where he leads the Computer Architecture Laboratory. Contact him at dgizop@di.uoa.gr.

Pedro Trancoso is an associate professor in the University of Cyprus' Department of Computer Science. Contact him at pedro@cs.ucy.ac.cy.

Yiannakis Sazeides is an associate professor in the University of Cyprus' Department of Computer Science. Contact him at yanos@cs.ucy.ac.cy.

Christos D. Antonopoulos is an assistant professor in the University of Thessaly's Electrical and Computer Engineering Department. Contact him at cda@uth.gr .

Srikumar Venugopal is a Research Scientist in IBM Research – Ireland. Contact him at srikumarv@ie.ibm.com

Shidhartha Das is a Principal Research Engineer at ARM Research, Cambridge, UK and Royal Academy of Engineering Visiting Professor at Newcastle University, UK. Contact him at sdas@arm.com