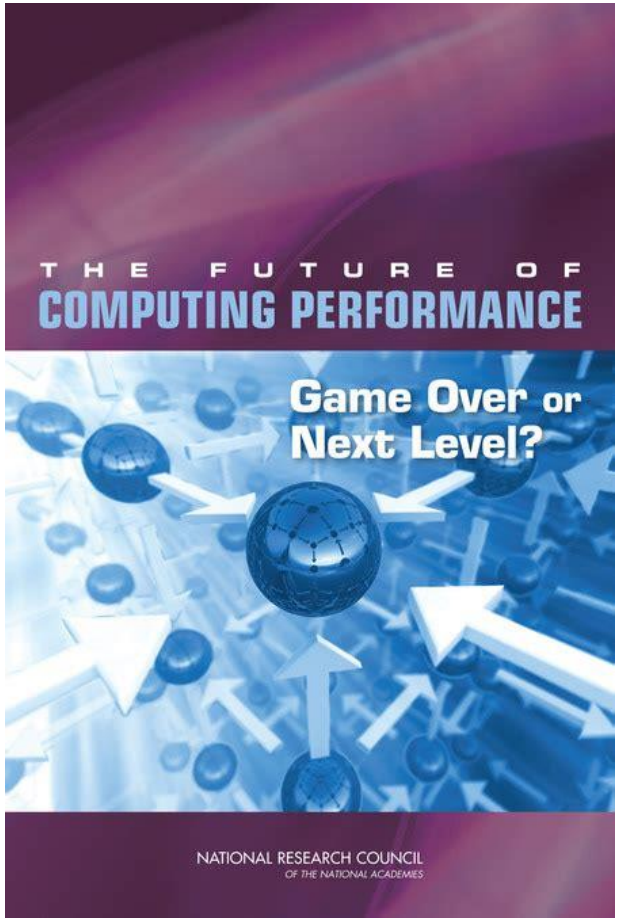
Exploring new hardware for data science

Gustavo Alonso
Systems Group
Department of Computer Science
ETH Zurich, Switzerland

The Hardware Era

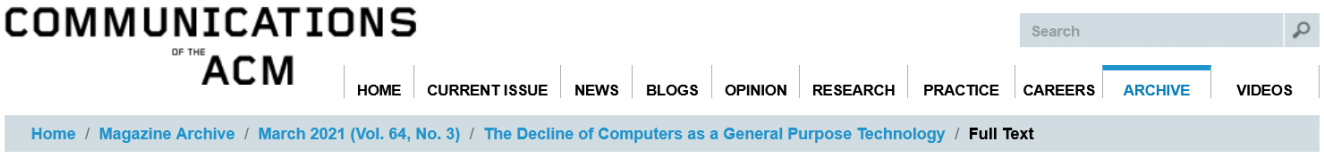
Not a new concept …

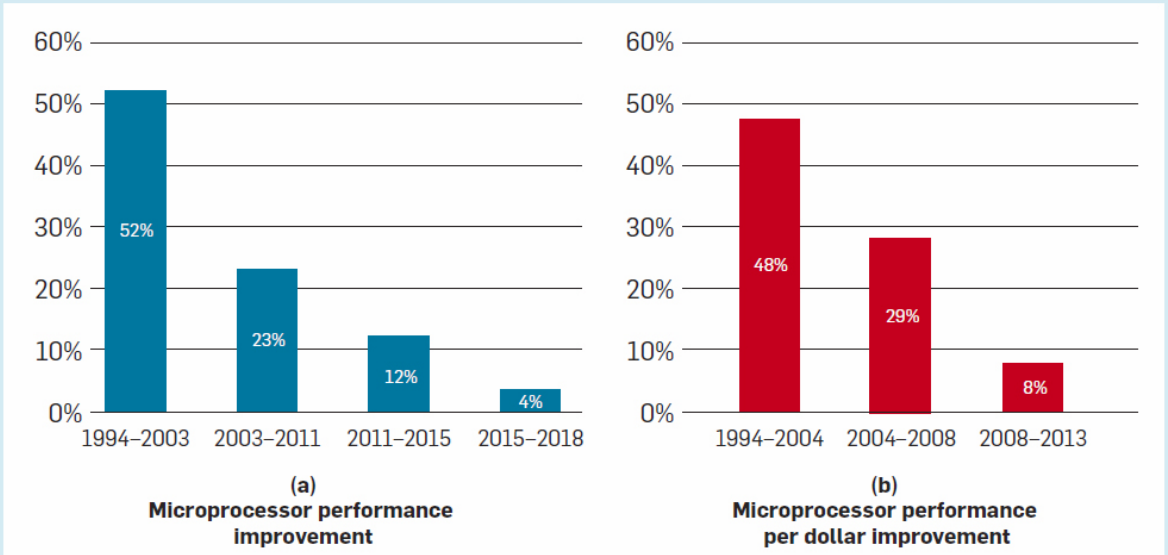


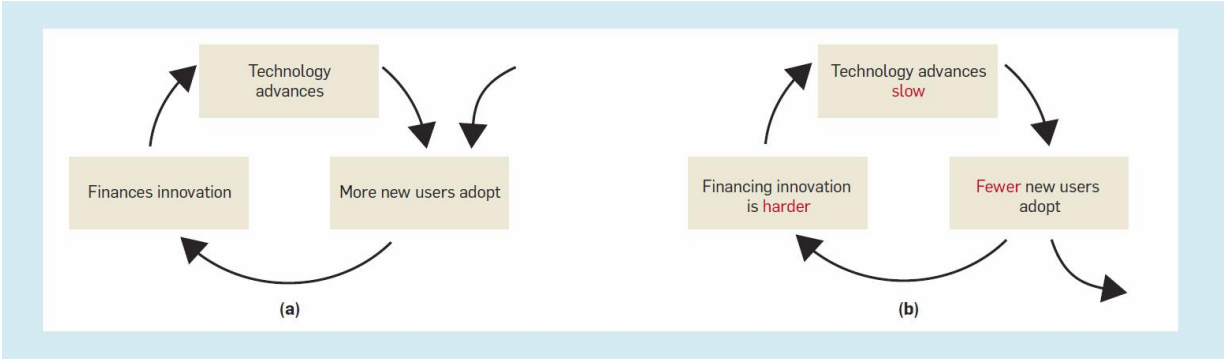
- 2011 Report
- Exponential growth for several decades
- Exponential growth no longer possible
- Switch to multicore and parallelism
  - Energy consumption becomes an issue
  - Multicore introduces parallelism that we do not know how to exploit well
- Situation will not change in near future
- Alternative is specialization
- Either somebody comes up with a new great invention or there is a problem

General purpose computing

Slow improvements lead to specialization

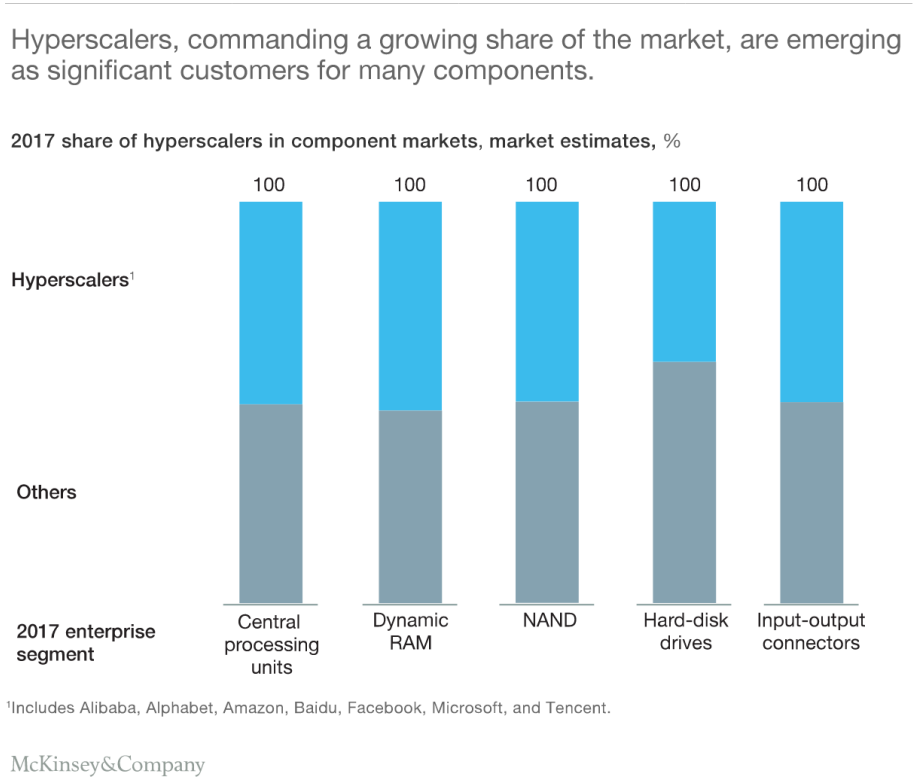






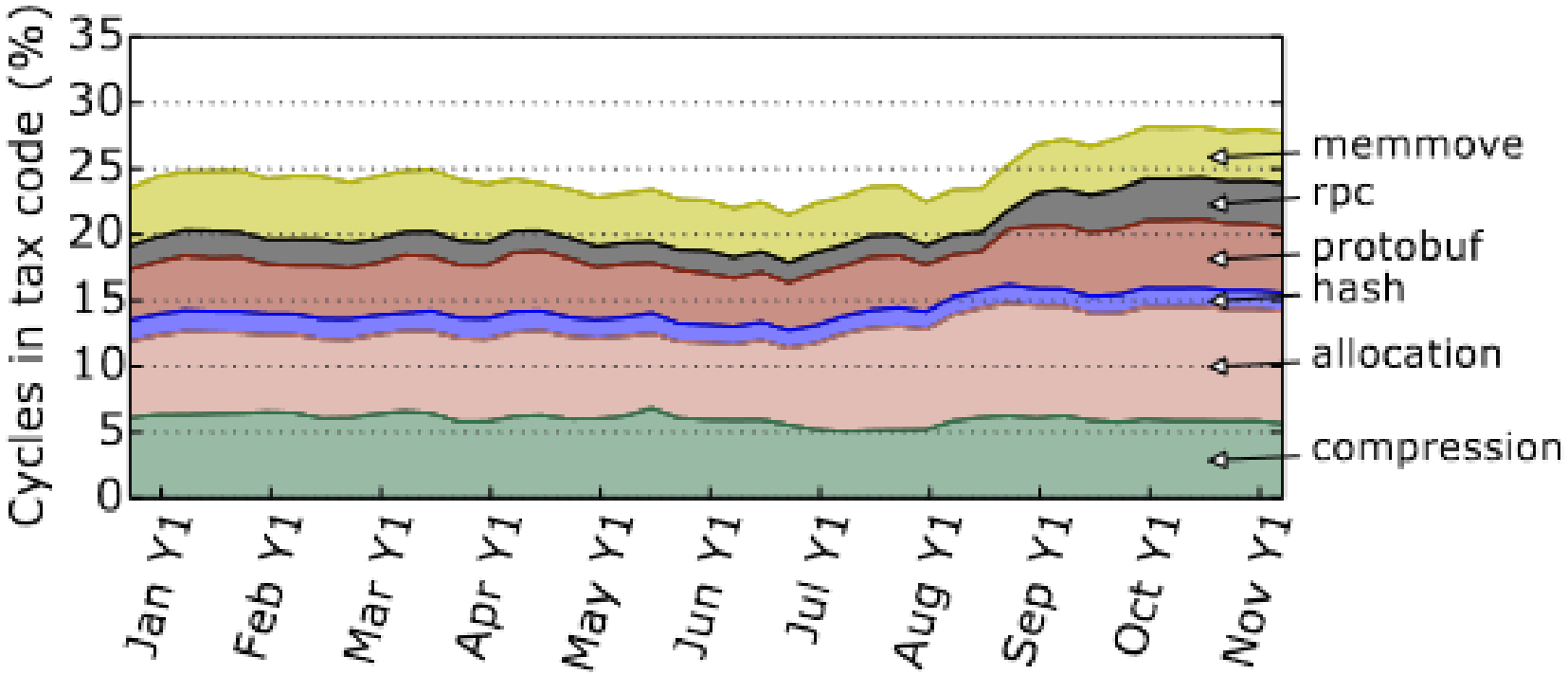
Driving specialization

- The cloud is the big game changer:
  - New business model
  - Economies of scale
  - Very large workloads
- Every hyper scaler is its own “Killer App”
  - The scale makes many things feasible
  - The gains have a very large multiplier



https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/how-high-tech-suppliers-are-responding-to-the-hyperscaler-opportunity

The Data Center Tax



Profiling a warehouse-scale computer, ISCA 2015

Data Compression (Microsoft Zipline/Corsica)

https://azure.microsoft.com/en-us/blog/improved-cloud-service-performance-through-asic-acceleration/

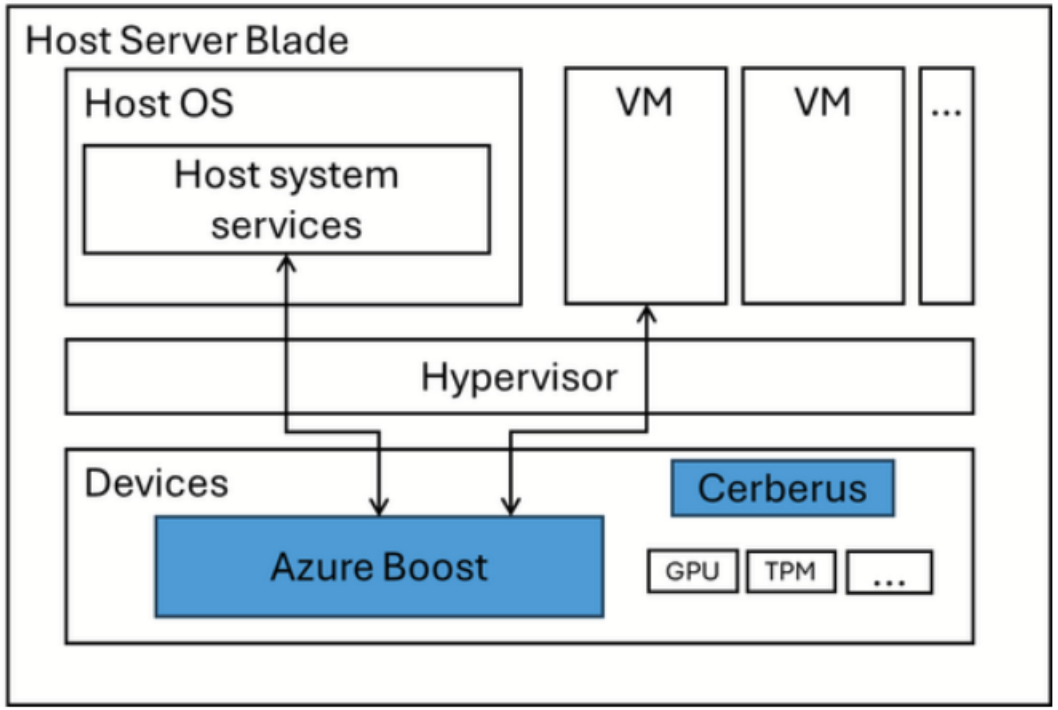
Microsoft Azure Boost

Security architecture components

Designed to enhance Azure workload security, Azure Boost includes the following security components:

- **An independent hardware root of trust** - Cerberus fulfils NIST 800-193 certification.
- **Azure Boost system on chip (SoC)** – dedicated, Linux based system conducting management operations for the control plane.
- **Configurable field-programable gate array (FPGA)** – programable network and storage acceleration capabilities for the data plane.

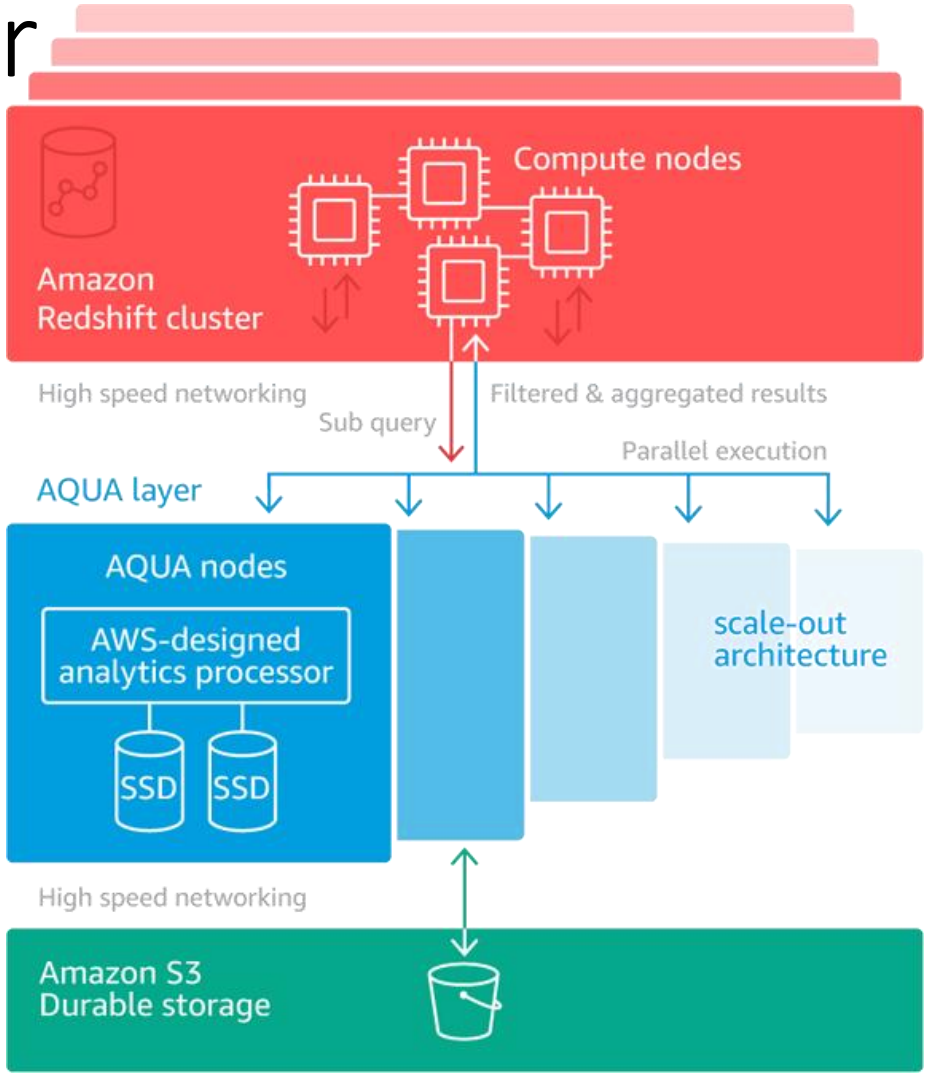
Azure Boost SoCs pair with each host and work in tandem to create a more secure hosting infrastructure.



Accelerators in a data center

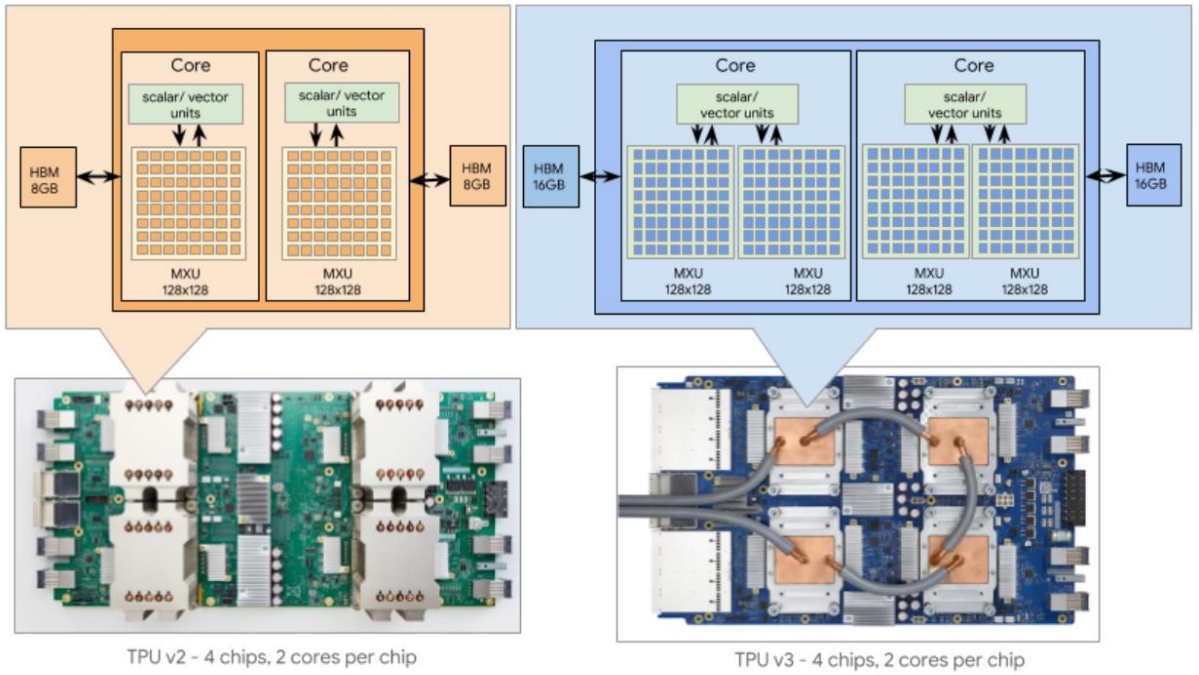
SELECT * FROM T WHERE id=3

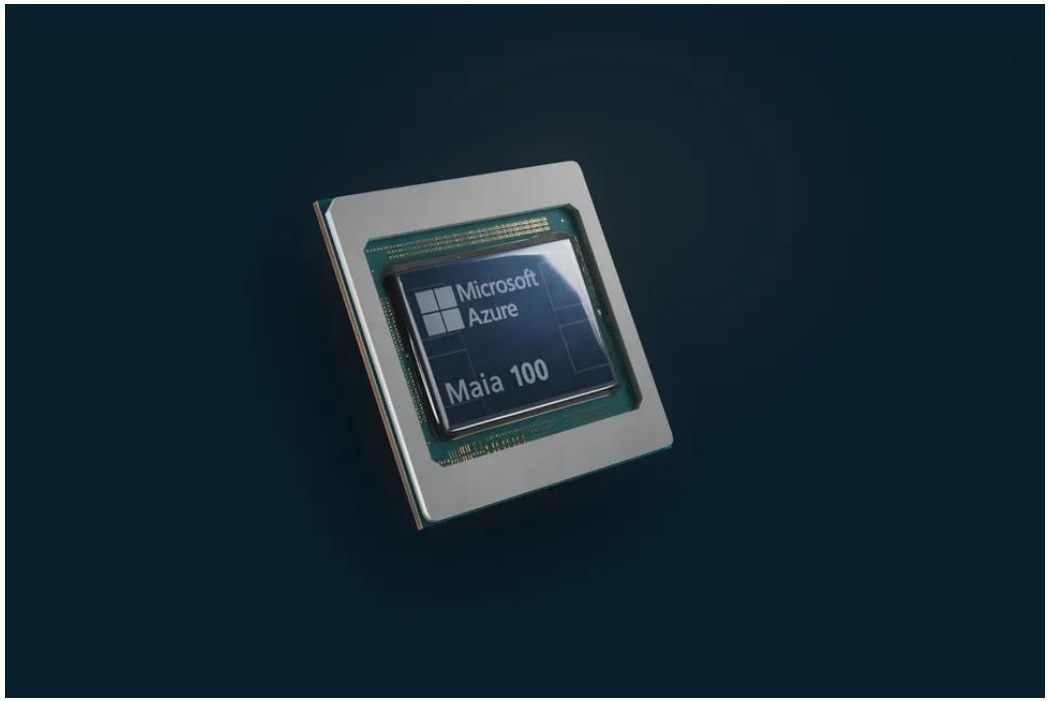
- A database will read the table from cloud storage
- Bring it all the way to the local memory, then to the CPU registers
- Just to throw away all tuples but 1
- Creates bottlenecks in storage, network, memory access, data buses, pollutes the caches, etc.



https://pages.awscloud.com/AQUA_Preview.html

Accelerating ML/AI



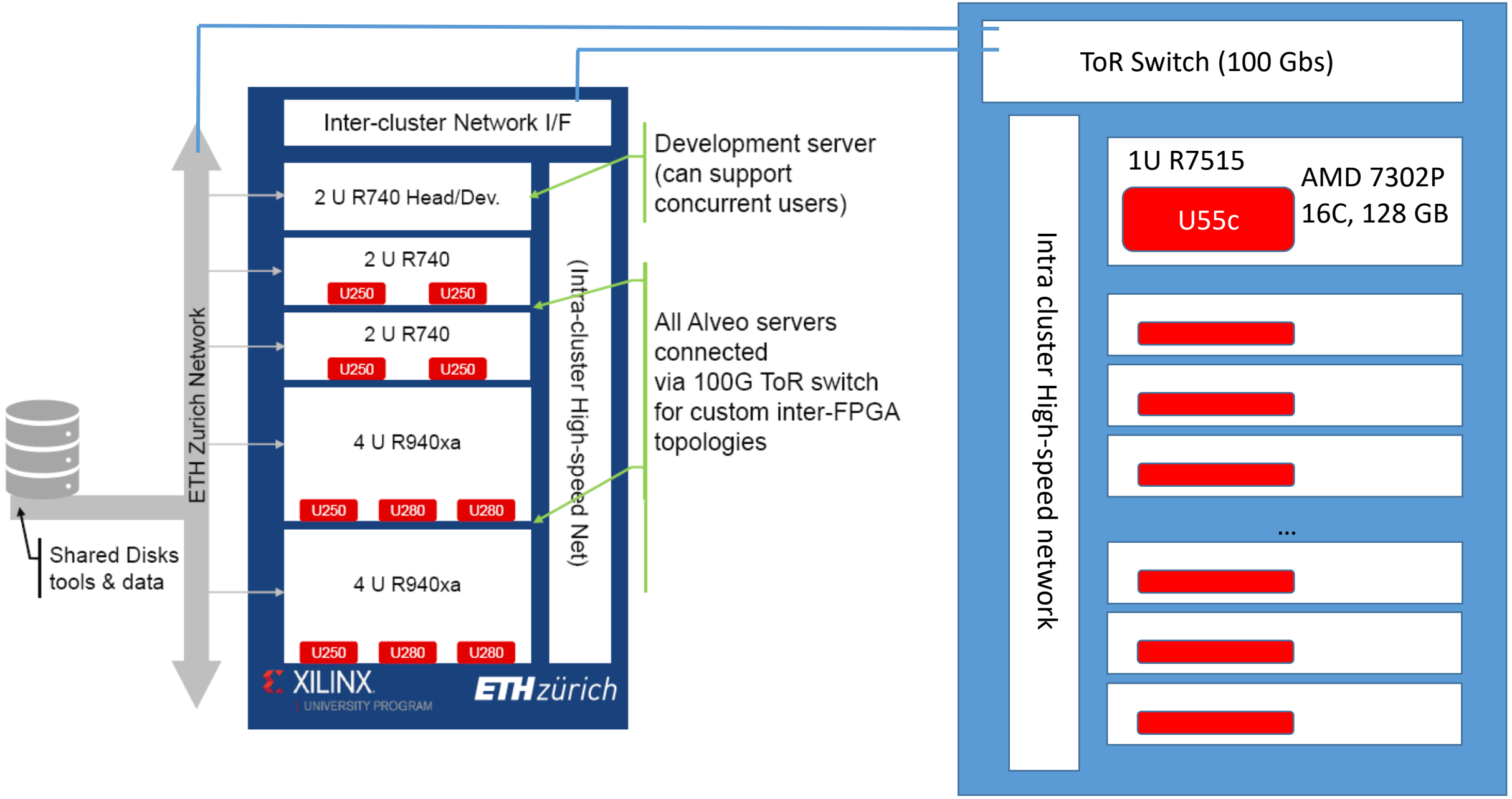


Infrastructure Building with new hardware

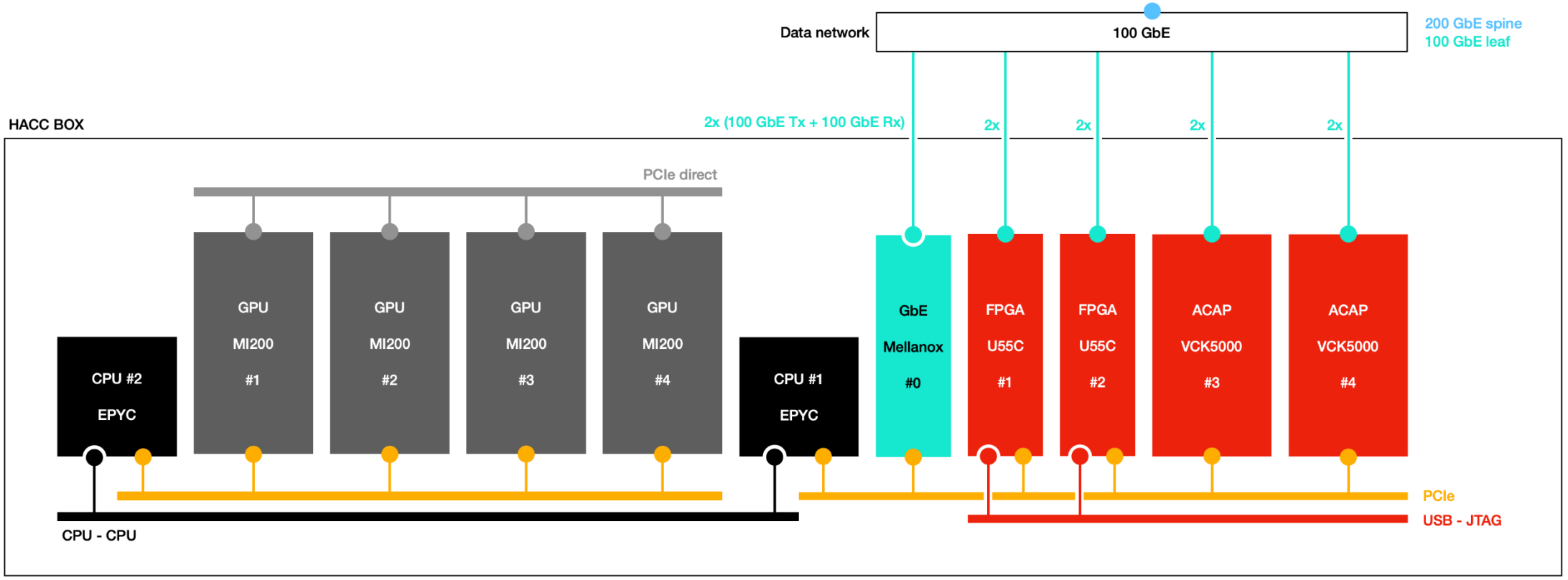
Infrastructure – HACC cluster

- The Heterogeneous Accelerated Compute Clusters (HACC) program is a unique initiative to support novel research in adaptive compute acceleration for data center settings and high-performance computing (HPC).

- ETH Zurich HACC
- https://systems.ethz.ch/research/data-processing-on-modern-hardware/hacc.html
- *https://github.com/fpgasystems/hacc*

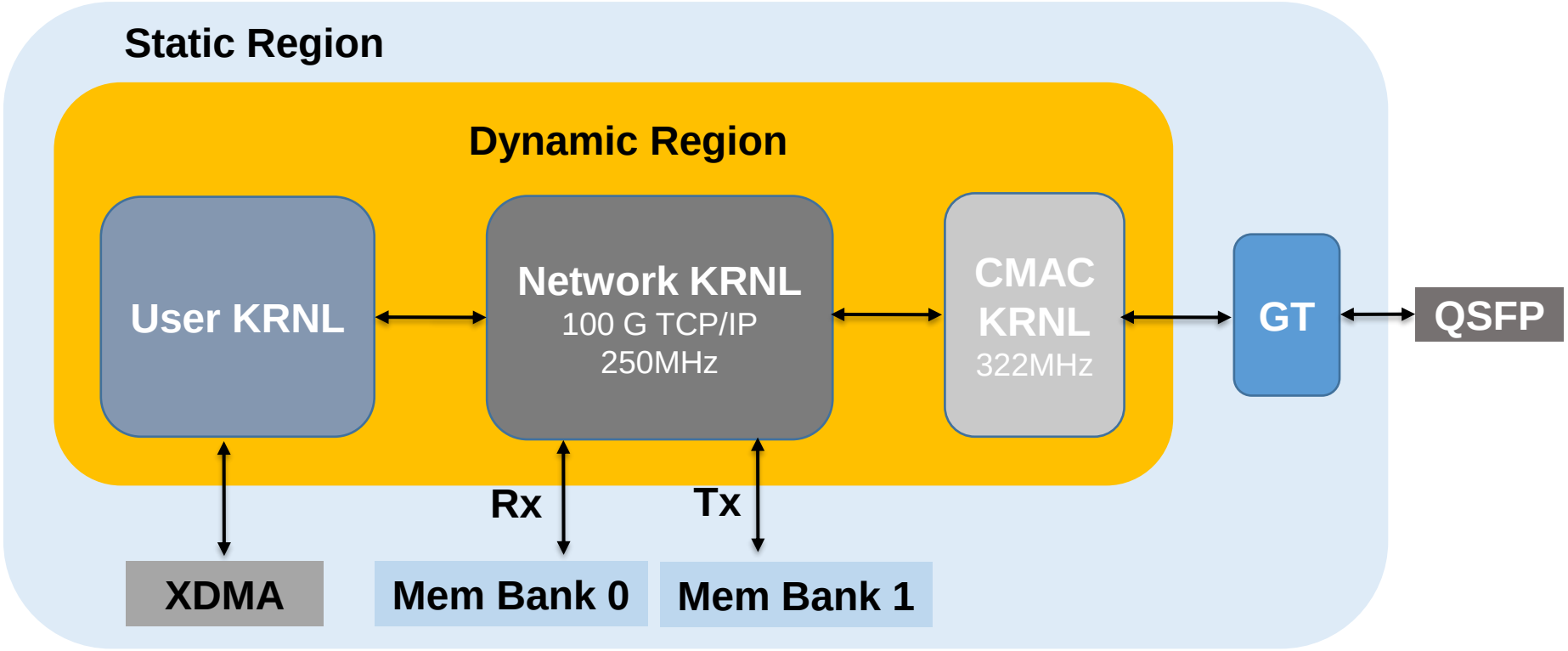

ToR Switch (100 Gbs)
Inter-cluster Network I/F
2 U R740 Head/Dev.
2 U R740
U250
U250
2 U R740
U250
U250
4 U R940xa
U250
U280
U280
4 U R940xa
U250
U280
U280
(Intra-cluster High-speed Net)
ETH Zurich Network
Shared Disks
tools & data
Development server
(can support
concurrent users)
All Alveo servers
connected
via 100G ToR switch
for custom inter-FPGA
topologies
XILINX
UNIVERSITY PROGRAM
ETH zürich
Intra cluster High-speed network
1U R7515
AMD 7302P
16C, 128 GB
U55c
...

Overview (HACC heterogeneous boxes)



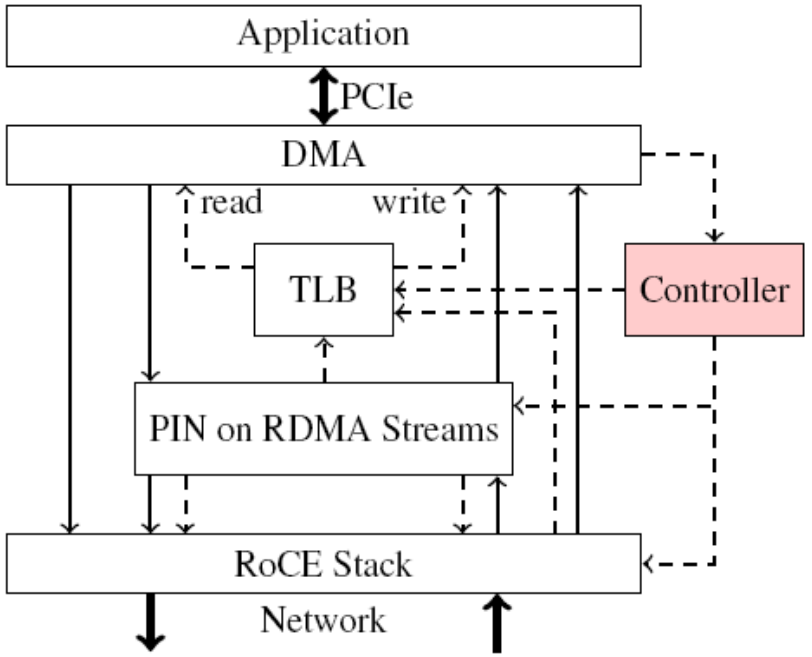
EasyNet & ACCL (100 GbE TCP/IP)

- CMAC: Ethernet subsystem, board specific
- Network: TCP/IP stack with streaming control and data interfaces
- User: Customized unit for application



He, Korolija, Alonso, "EasyNet: 100 Gbps Network for HLS", FPL'21

RDMA on FPGA



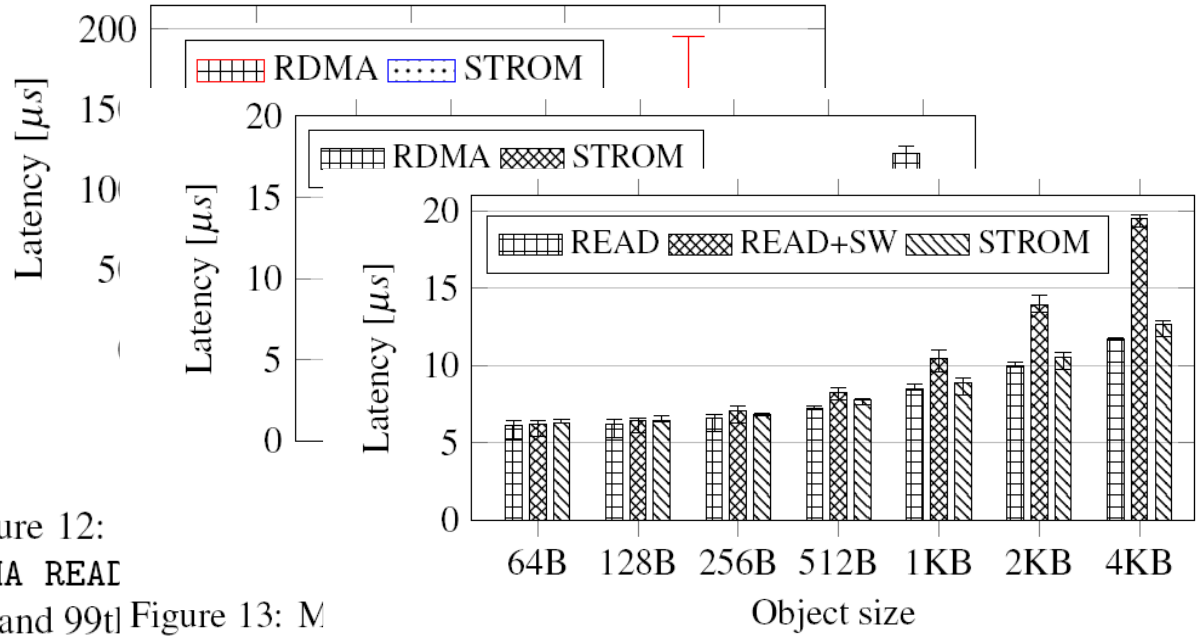
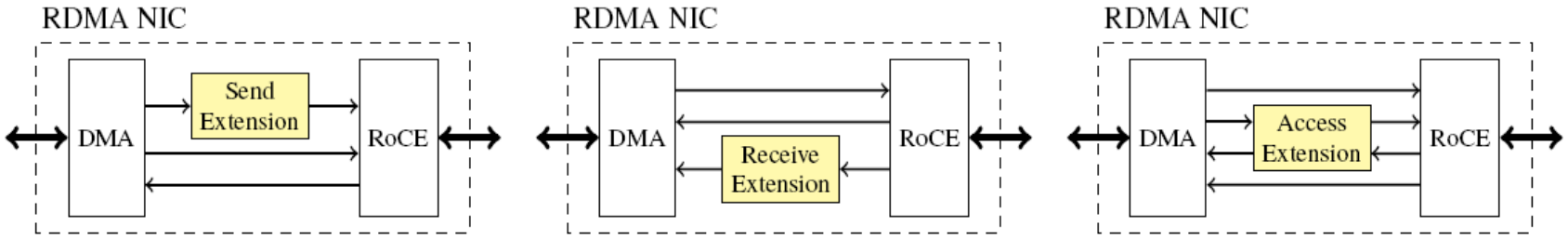


Figure 12:
RDMA READ
1st and 99t

Figure 13: M
READ vs a *tra*
Error bars ar

Figure 14: Median latency of reading a remote value without a consistency check, with a local CRC64 check in software, and with the CRC64 check offloaded to the *consistency kernel* on the remote NIC. Error bars indicate the 1st and 99th percentile.



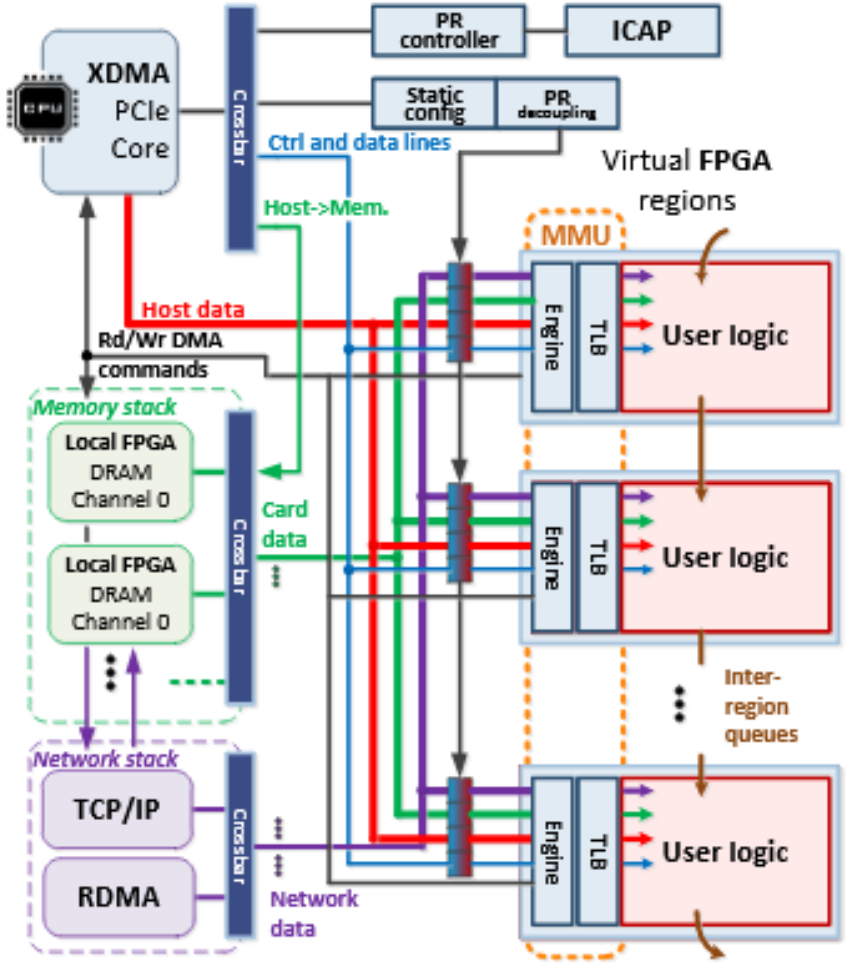
(a) Send Extension (b) Receive Extension (c) Access Extension

Sidler et al. Strom, EuroSys 2020

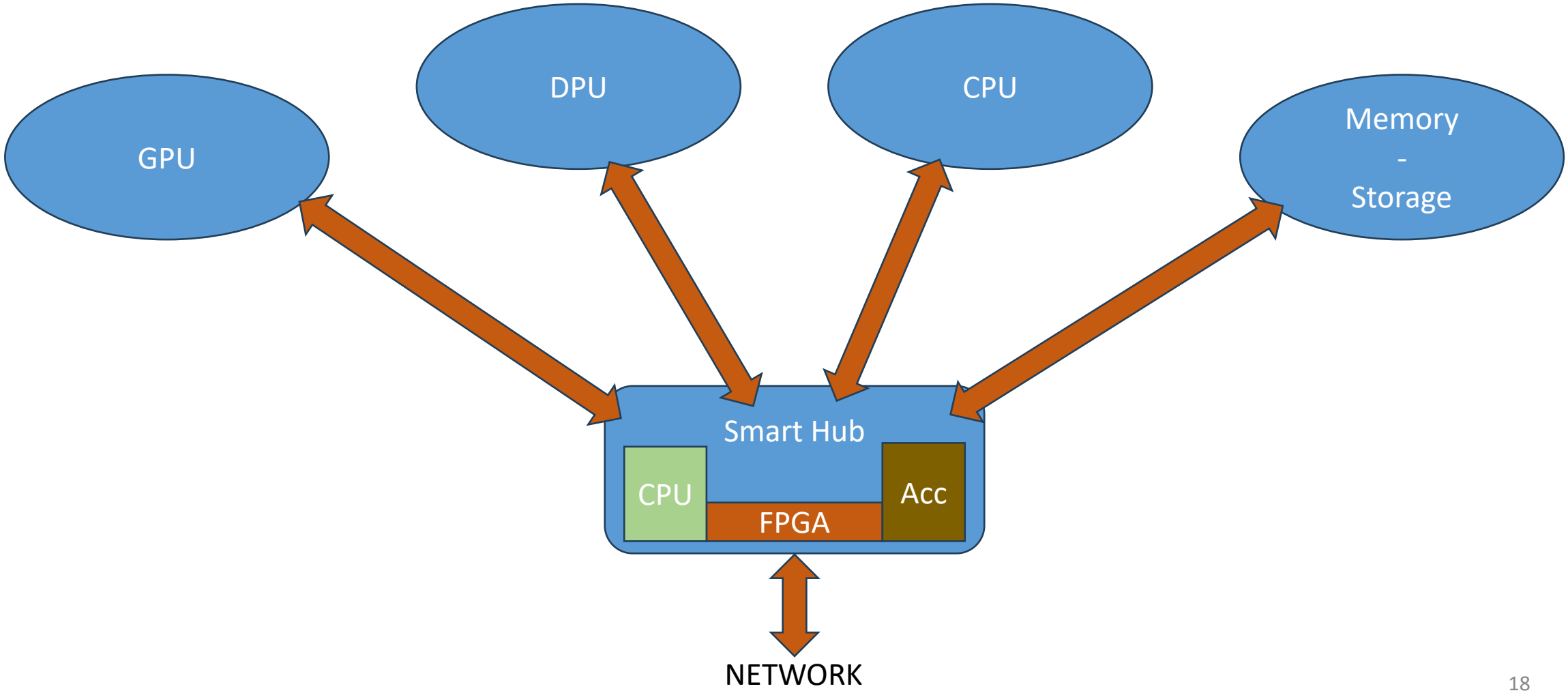
Coyote: a better FPGA shell

- Multiple user regions (6 to 10)
- RDMA/TCP network stack
- Unified memory space host-FPGA
- Virtual memory
- Multi-user memory management on FPGA

Do OS abstractions make sense on FPGAs?, Dario Korolija, Timothy Roscoe, and Gustavo Alonso, OSDI 2020

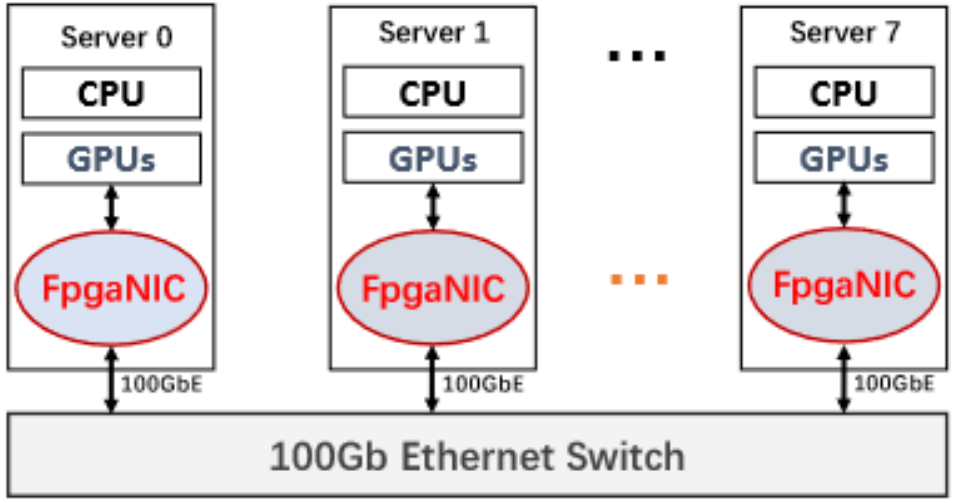


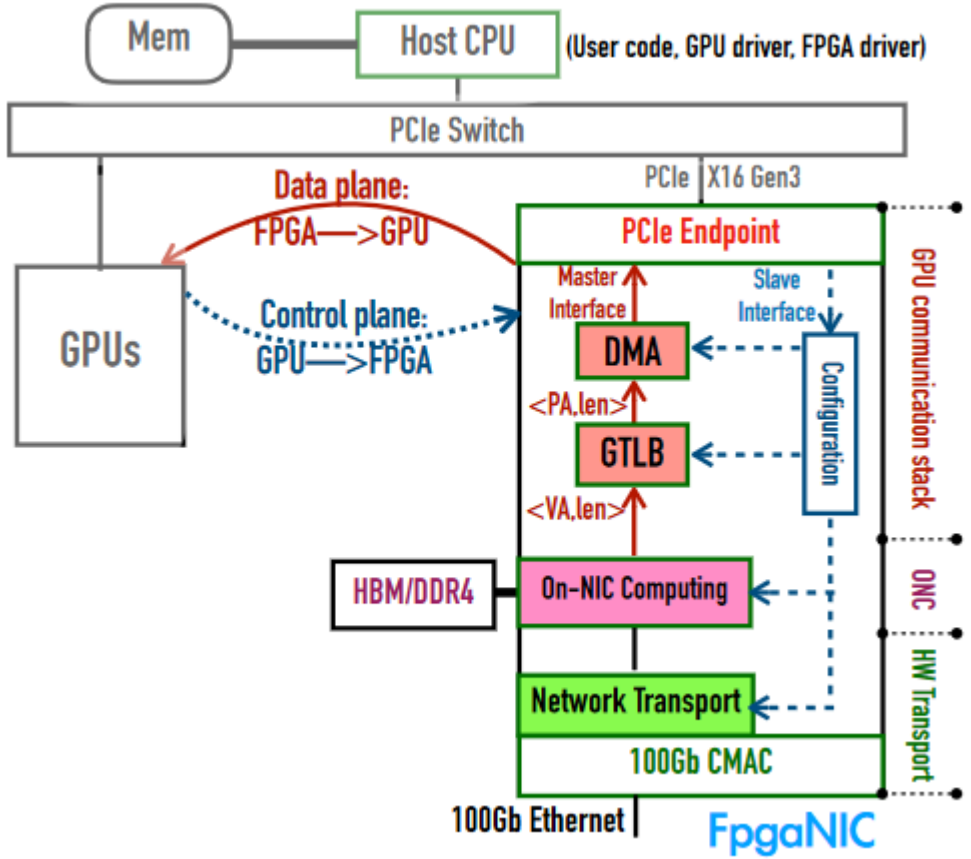
New Focus: SLASH (jointly with AMD Dublin)



FPGA GPU-FPGA

FpgaNIC: An FPGA-based Versatile 100Gb SmartNIC for GPUs
Wang et al. USENIX ATC 2022





Use Cases

Example

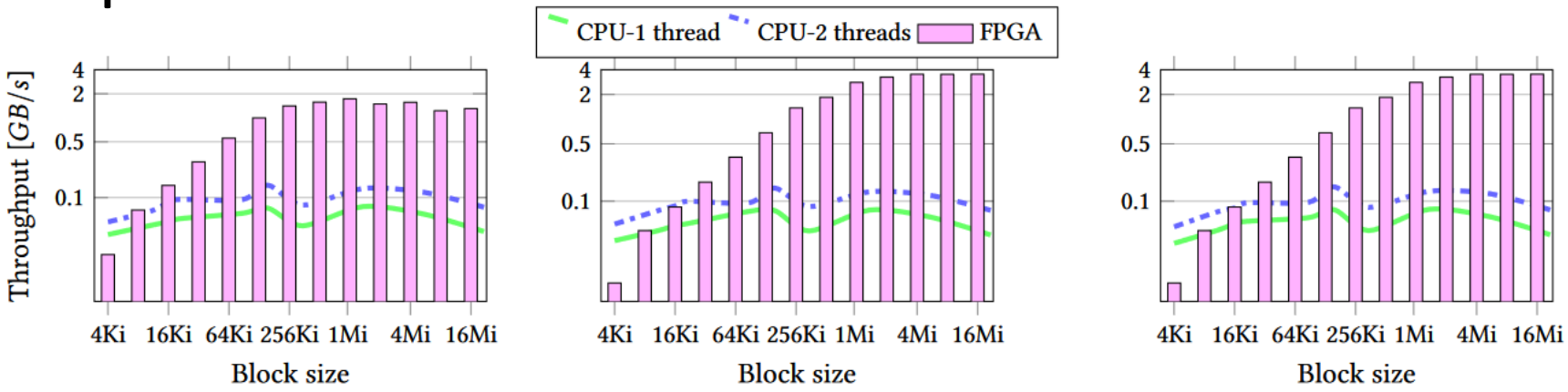


Figure 12: Full pipeline - with 1 and 2 threads on CPU vs. FPGA design. Note the logarithmic scale of the y axis.

Hardware Acceleration of Compression and Encryption in SAP HANA

Monica Chiosa*
ETH Zurich
monica.chiosa@inf.ethz.ch

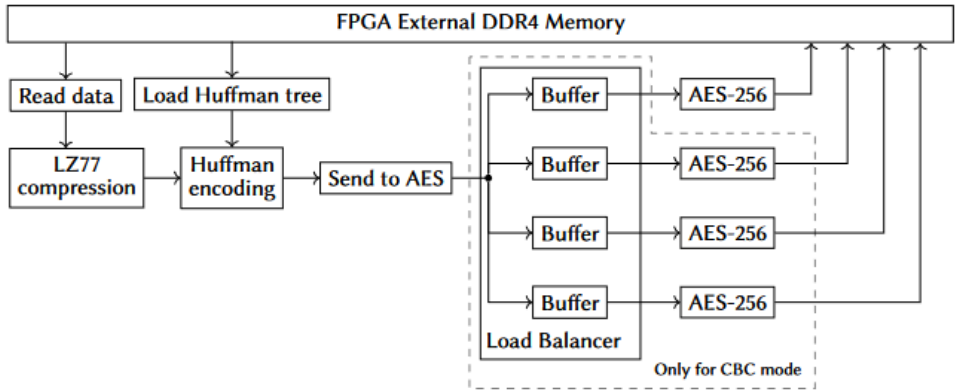
Fabio Maschi*
ETH Zurich
fabio.maschi@inf.ethz.ch

Ingo Müller
ETH Zurich
ingo.mueller@inf.ethz.ch

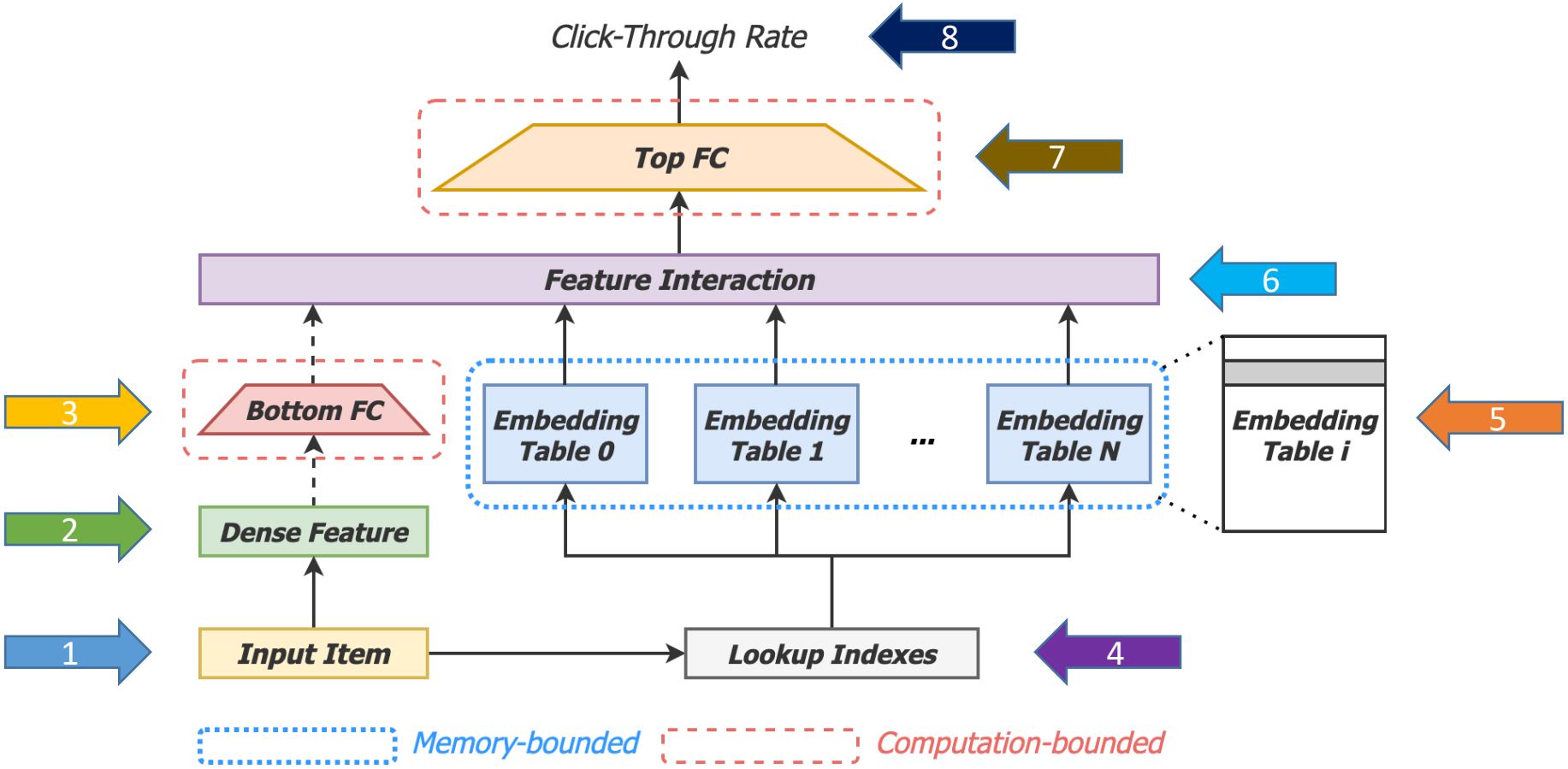
Gustavo Alonso
ETH Zurich
alonso@inf.ethz.ch

Norman May
SAP SE
norman.may@sap.com

VLDB 2022

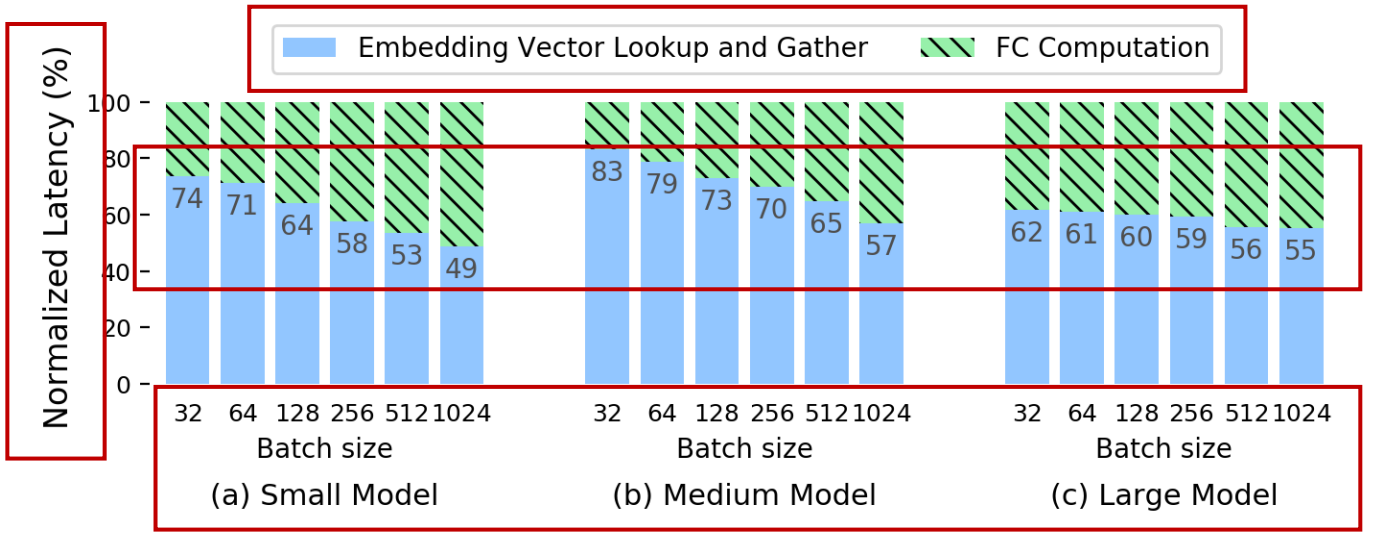


Deep recommendation models involve intensive embedding table lookups



Wenqi Jiang, Zhenhao He, Shuai Zhang, Thomas B. Preußer, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, Gustavo Alonso: MicroRec: Efficient Recommendation Inference by Hardware and Data Structure Solutions. MLSys 2021

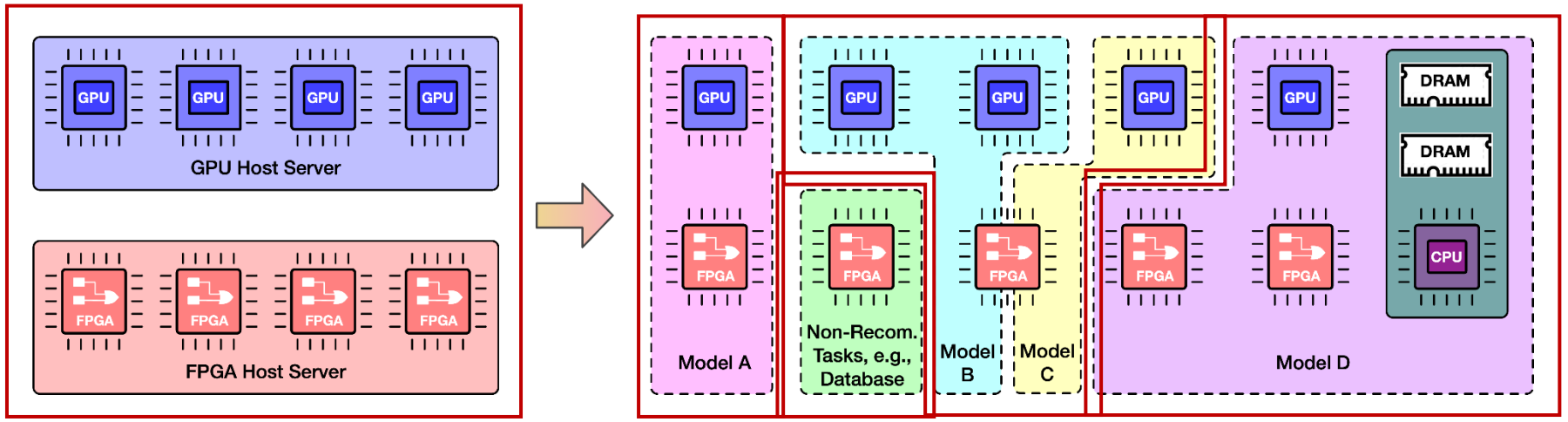
Workload profiling on Alibaba's real models



- Embedding lookup comprises more than half of the inference

FleetRec: bridging CPUs, GPUs and FPGAs by network in the cloud

- Using existing server



Interconnect through network

Wenqi Jiang, Zhenhao He, Shuai Zhang, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, Gustavo Alonso: FleetRec: Large-Scale Recommendation Inference on Hybrid GPU-FPGA Clusters. KDD 2021

Using distributed clusters

- We have used the HACC cluster to implement such a distributed system over 10 FPGAs, enabling us to explore the options for accelerating every step of the process

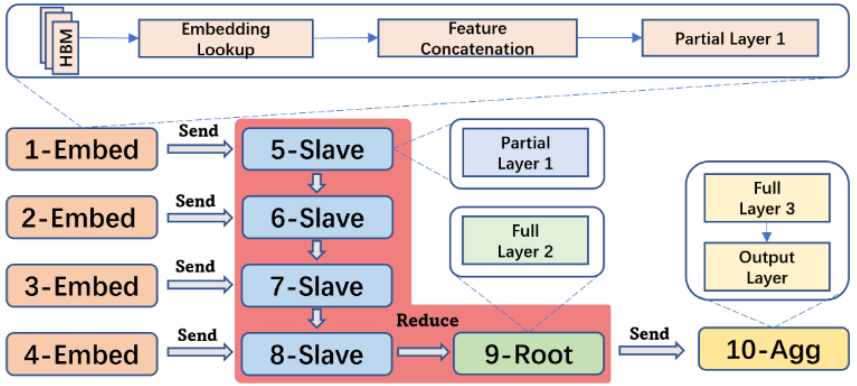


Figure 16: Conceptual design of partitioned DLRM, with $FC1$ decomposed and $FC2$, $FC3$ pipelined across nodes.

ACCL+: an FPGA-Based Collective Engine for Distributed Applications

Zhenhao He, Dario Korolija, Yu Zhu, and Benjamin Ramhorst, *Systems Group, ETH Zurich;* Tristan Laan, *University of Amsterdam;* Lucian Petrica and Michaela Blott, *AMD Research;* Gustavo Alonso, *Systems Group, ETH Zurich*

https://www.usenix.org/conference/osdi24/presentation/he

Conclusions

- Hardware acceleration and specialization is here to stay
- The bottleneck and inefficiencies caused by data movement will only become worse over time because eof the workloads and the growing use of the cloud
- Use one to solve the other
  - In-network data processing
  - Reconfigurable implementations (FPGA) for flexibility
  - Enabling new architectures
- This is not just about hardware:
  - The software needs to evolve to match the new systems