How Hardware Evolution is Driving Software Systems

Gustavo Alonso

Systems Group

Department of Computer Science

ETH Zurich, Switzerland

Systems@ETH zürich

DEBS'19 – Darmstadt, Germany

www.systems.ethz.ch

- Muhsen Owaida (senior researcher)
- Zeke Wang (senior researcher)
- Amit Kulkarni (senior researcher)

- David Sidler (PhD student)
- Kaan Kara (PhD student)
- Abishek Ramdas (PhD student)
- Fabio Maschi (PhD student)
- Dario Korolija (PhD student)
- Zhenhao He (PhD student)
- Monica Chiosa (PhD student)

Systems Group

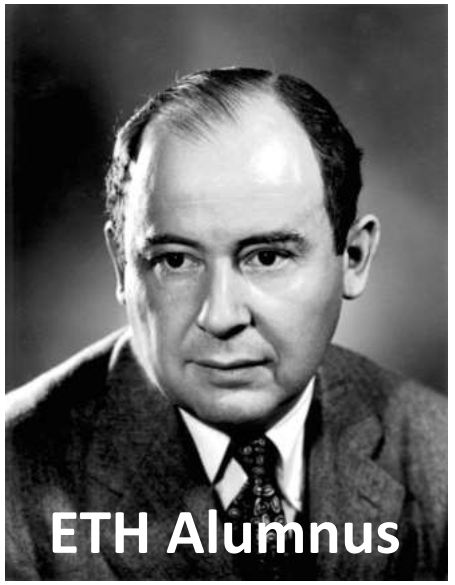
The usual starting point

- Moore’s Law
- Dennar scaling, physical limits
- Multicore
- GPU, TPU, FPGA
- Data centers and the cloud
- …

- Corollarium: Hardware is changing really fast looking for a way forward

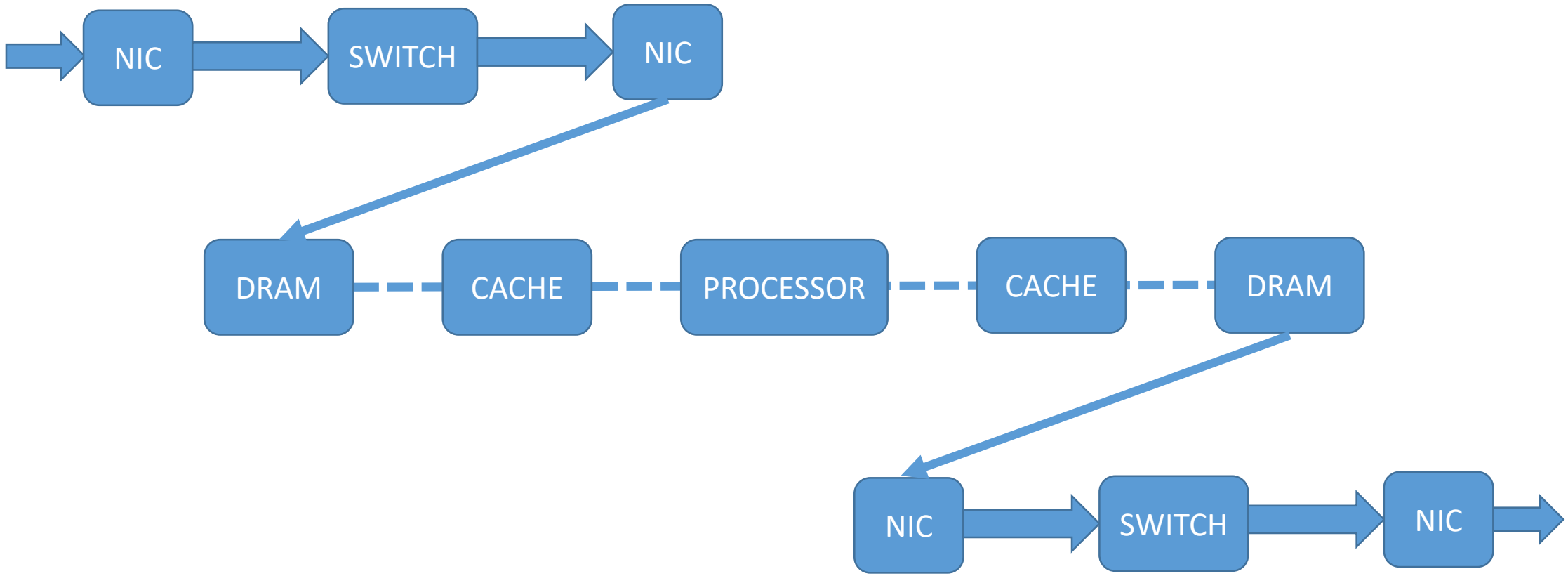
Exploring future data processing systems

- **Algorithms**: What can be done if we are not (or less) bound by the limitations of modern CPUs?
- **Architectures**: What can be done if we are not (or less) bound by the limitations of current Von Neumann and x86 style architectures?
- **Systems**: If we are no longer bound by CPU and architectural limitations, how would complete systems look like?



Abstraction

Reality

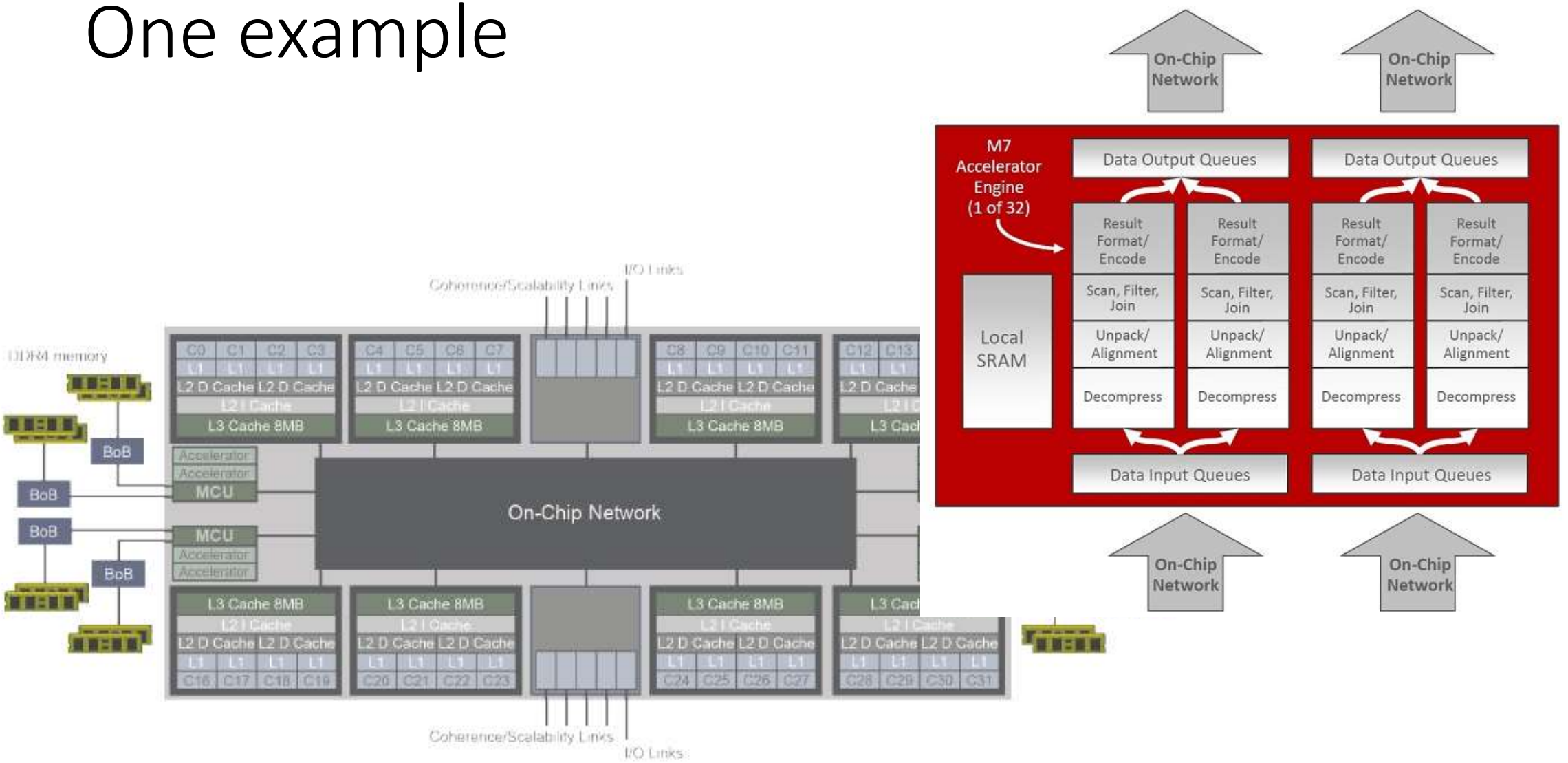

NIC
SWITCH
NIC
DRAM
CACHE
PROCESSOR
CACHE
DRAM
NIC
SWITCH
NIC

Efficient Data processing on new hardware

- Big Data implies there is a lot of data
  - If the data moves, you lose. Hence, ...
  - ... if the data moves, something useful better happen beyond moving the data

Every element in the system
(memory, bus, disk, cache, network card, network switch, ...)
should be a processing component

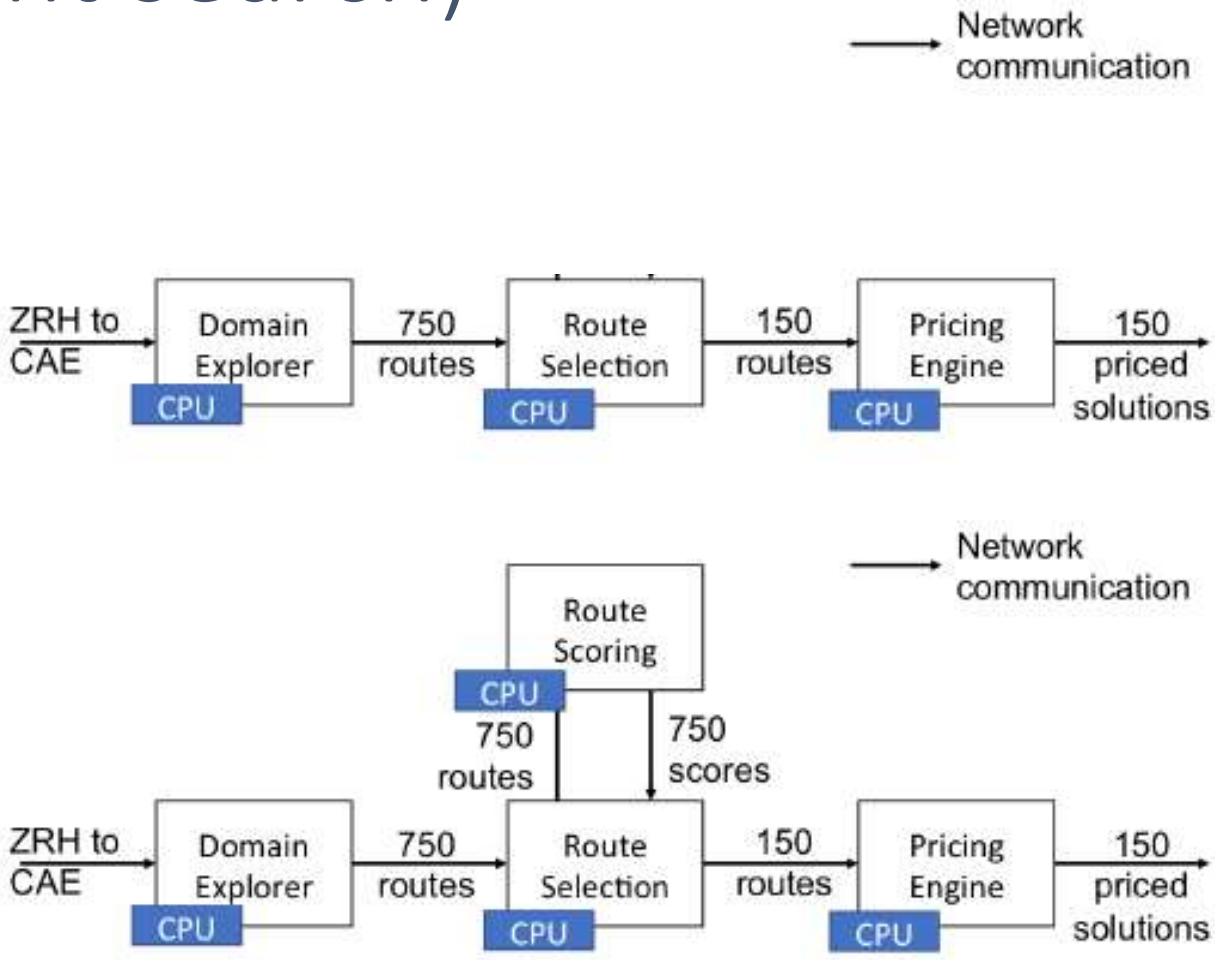
One example


On-Chip Network
On-Chip Network
M7 Accelerator Engine (1 of 32)
Data Output Queues
Data Output Queues
Result Format/ Encode
Result Format/ Encode
Result Format/ Encode
Result Format/ Encode
Scan, Filter, Join
Scan, Filter, Join
Scan, Filter, Join
Scan, Filter, Join
Unpack/ Alignment
Unpack/ Alignment
Unpack/ Alignment
Unpack/ Alignment
Decompress
Decompress
Decompress
Decompress
Local SRAM
Data Input Queues
Data Input Queues
On-Chip Network
On-Chip Network
Coherence/Scalability Links
I/O Links
DDR4 memory
BoB
BoB
BoB
BoB
Accelerator
Accelerator
MCU
MCU
Accelerator
Accelerator
On-Chip Network
L3 Cache 8MB
L2 D Cache
L2 I Cache
Coherence/Scalability Links
I/O Links

Events and streams in a real system

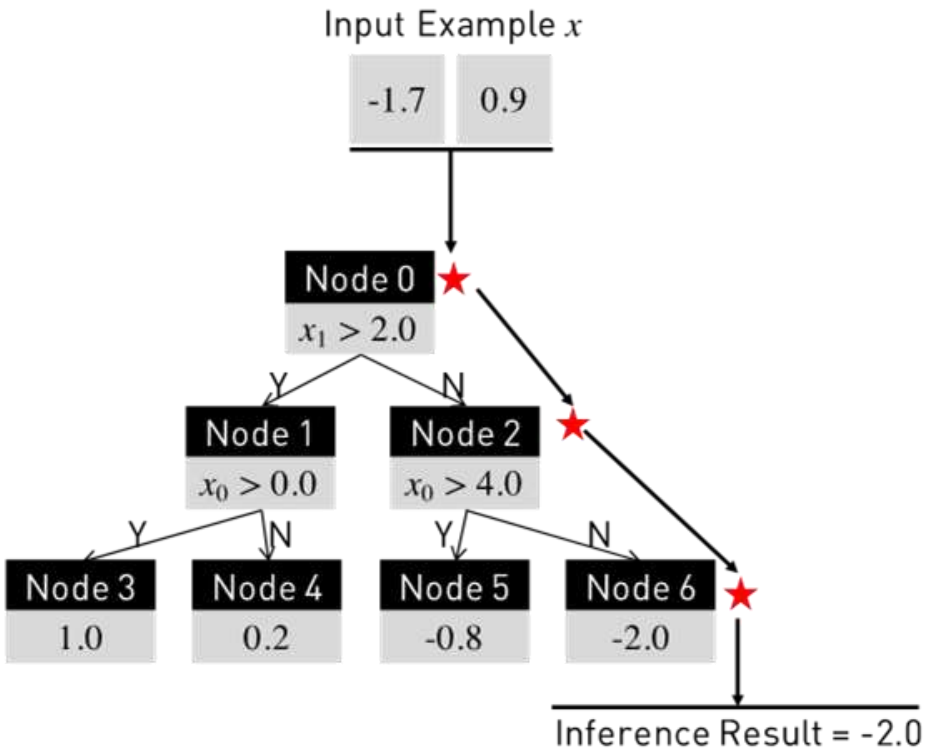
Amadeus use case (flight search)

- Complex systems involving events, rule engines, databases, and streams
- Typical recommender system trade-off:
  - latency vs throughput
  - Latency improved through reducing the amount of work at each stage and merging stages
  - Throughput improved by separating stages and parallelizing them across a cluster of machines
- Amount of data processed often restricted to meet requirements



Decision trees

Decision trees

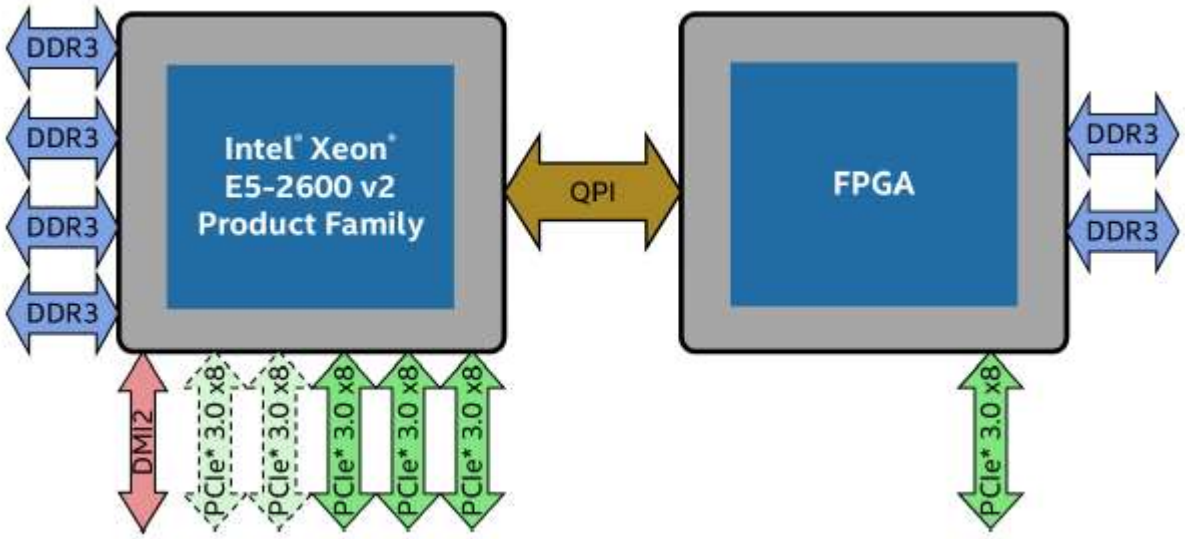


Decision tree ensembles

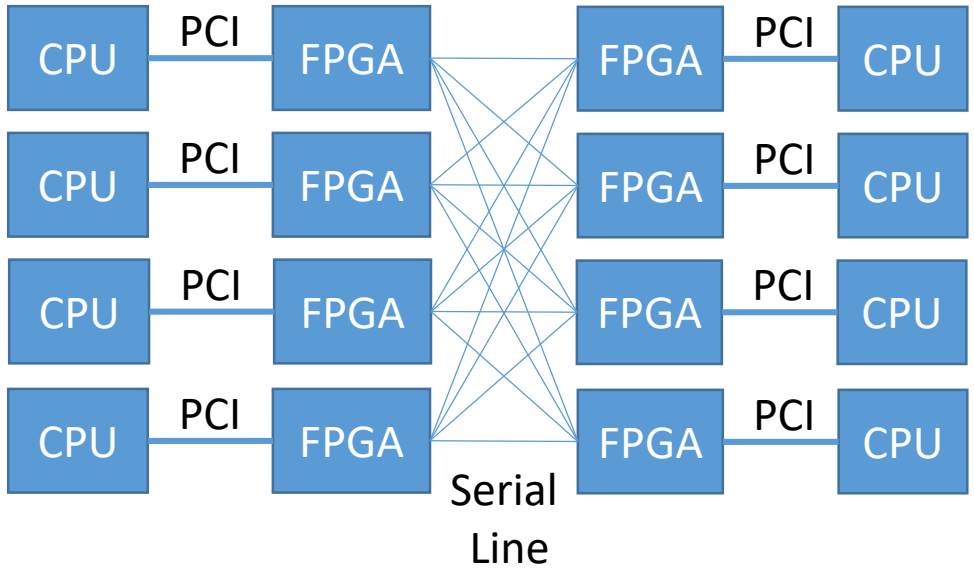
M. Owaida et al. FPL'17, FPL'18
Application Partitioning on FPGA Clusters: Inference over Decision Tree Ensembles

Processor Unit

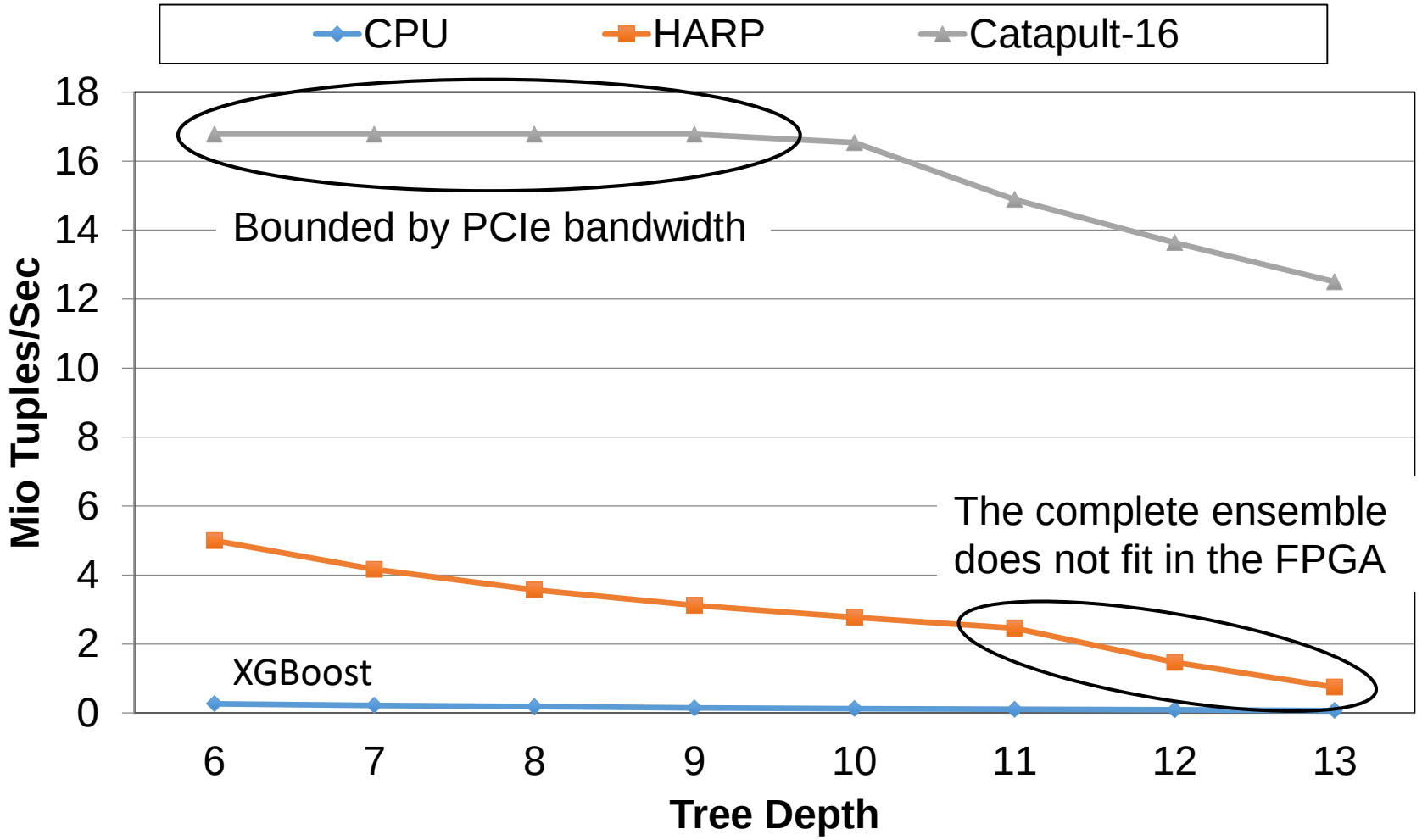
INTEL Xeon+FPGA v2



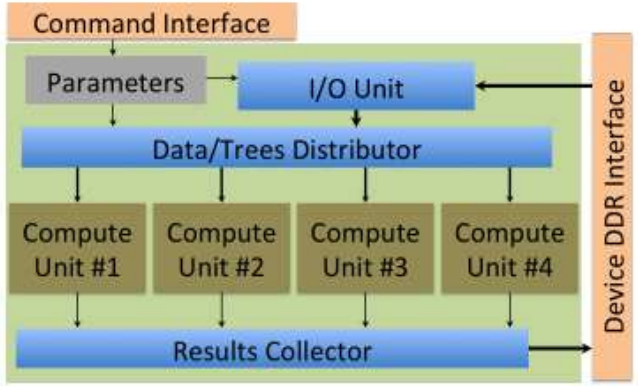
MICROSOFT CATAPULT v1

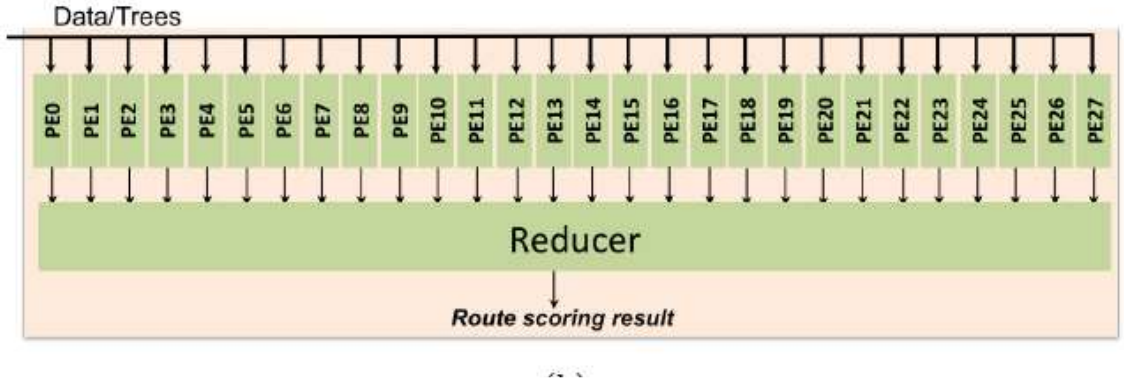


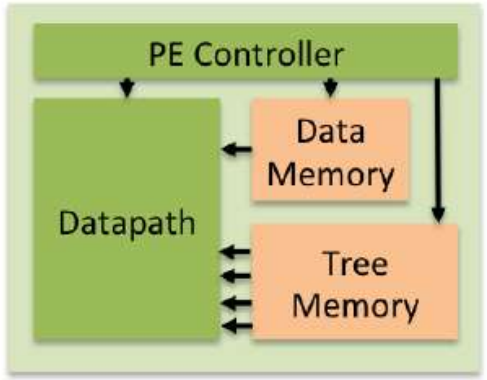
Making it work in practice

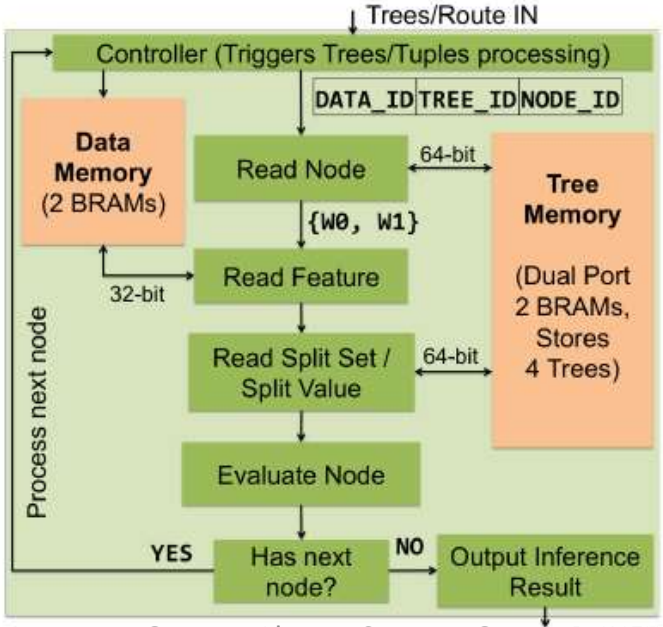


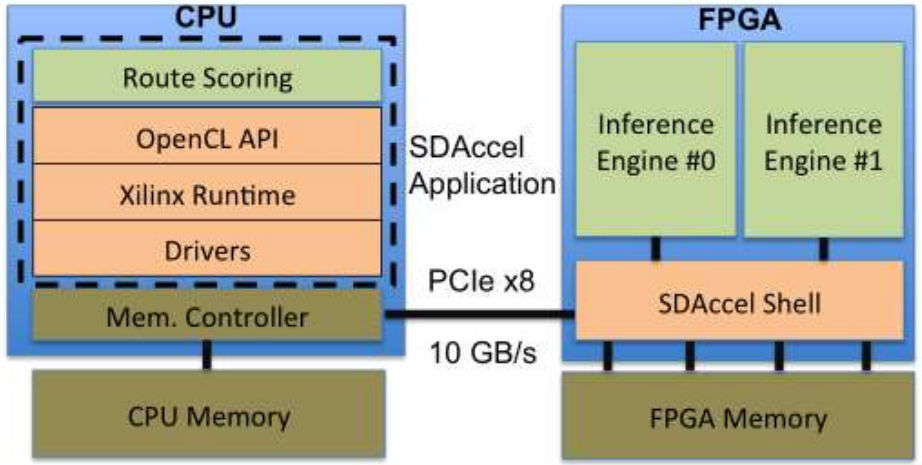
Parallelism on an FPGA (896 trees in one go)





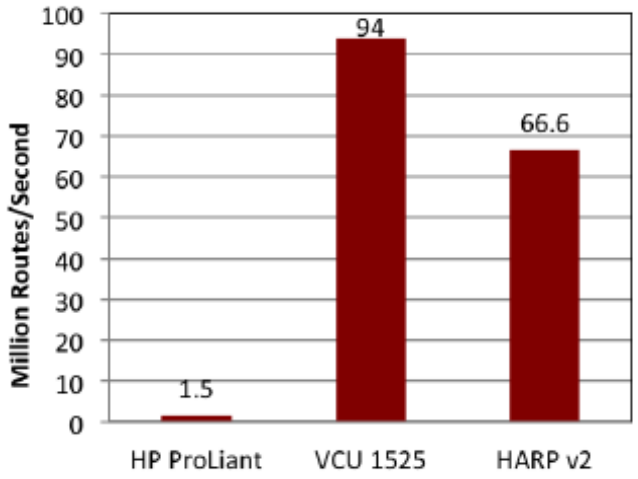


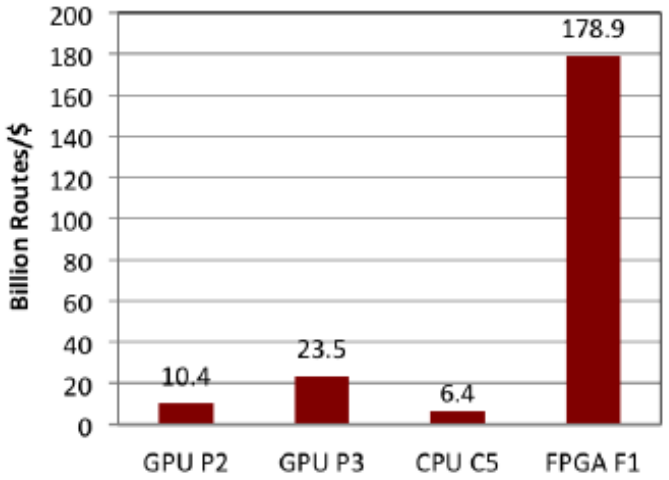


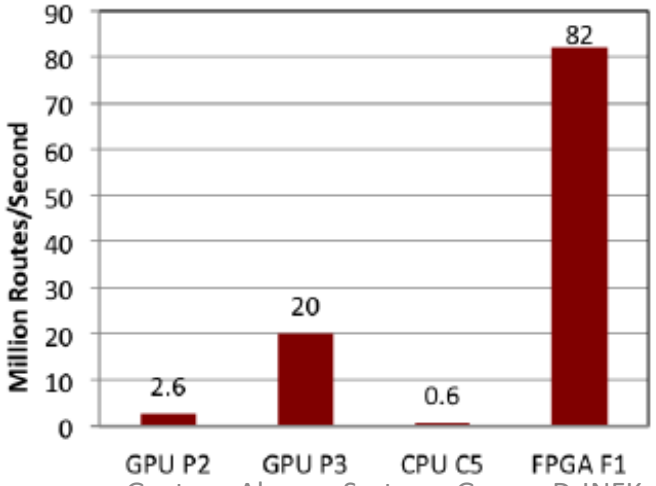


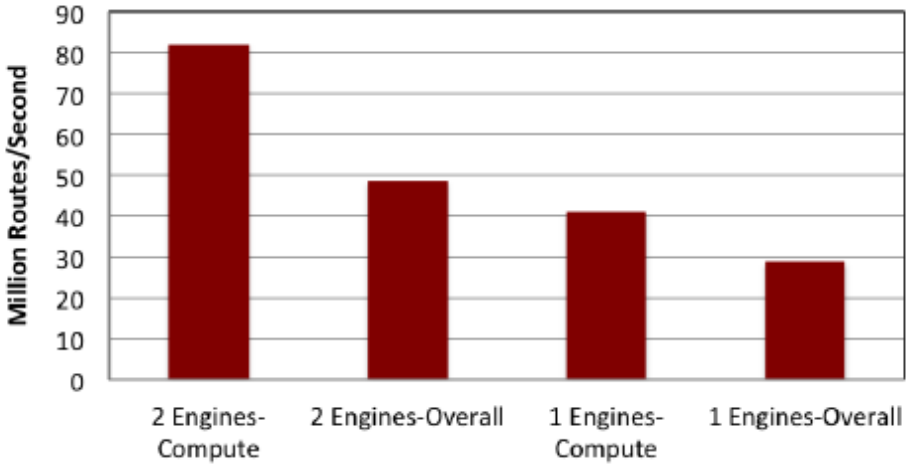
Flight Search on the cloud (Amazon F1)

AWS Instance	Features	Cost
GPU P2.xlarge	1 NVidia K80	0.90 $/hour
GPU P3 2xlarge	1 NVidia V100	3.06 $/hour
CPU C5 2xlarge	8 vCPUs	0.34 $/hour
FPGA F1 2xlarge	1 Virtex UltraScale+	1.65 $/hour
On-premise		
HP ProLiant	56 CPU cores	11K $
Intel's HARP v2	1 Arria 10 FPGA	7.5K $
Xilinx VCU1525	1 Virtex UltraScale+	7.5K $









Thinking of the architecture

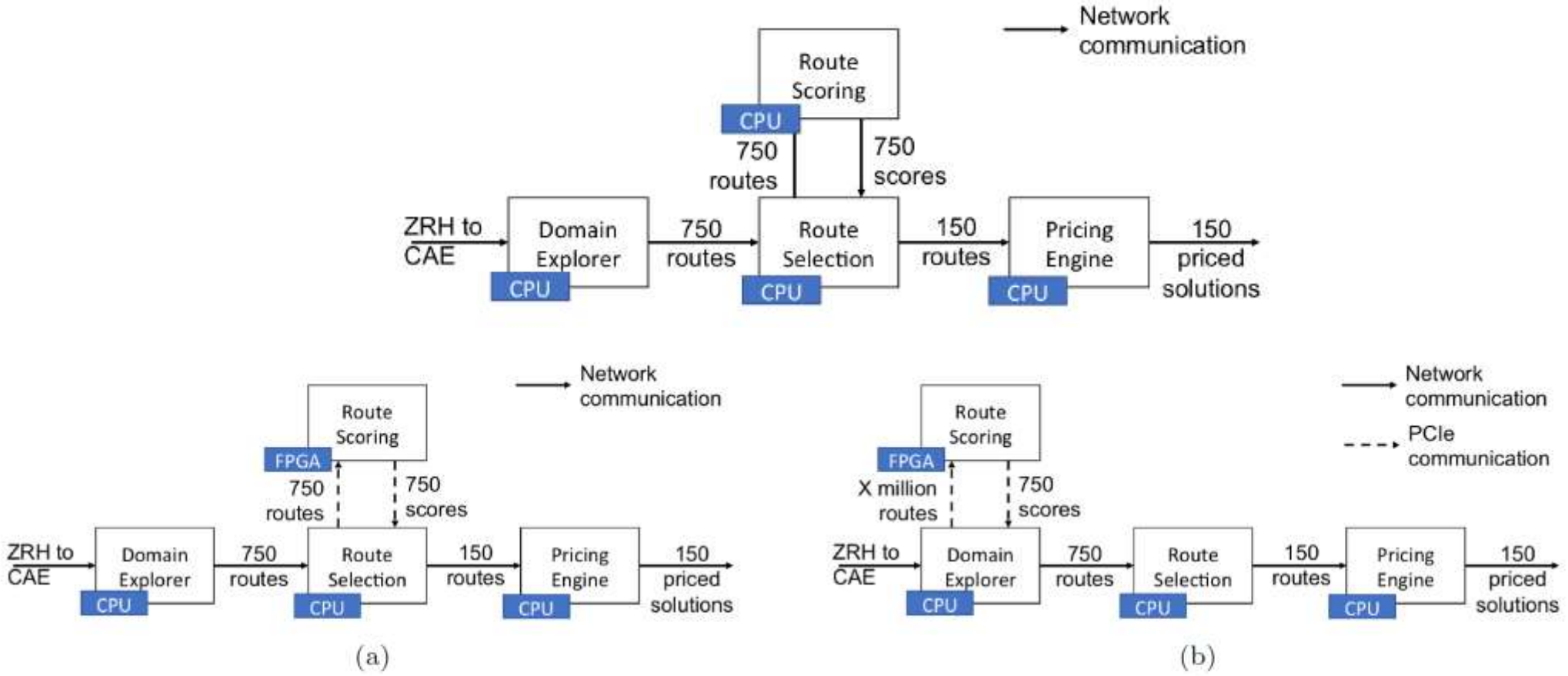


Figure 12: (a)Inserting a small FPGA card in each Route Selection server attached through PCIe.(b) Deploying the Route Scoring as part of the Domain Explorer by attaching an FPGA card to each Domain Explorer server.

Many more possibilities

- Currently exploring how to replace a rule engine with an FPGA implementation capable of working on streams
  - Minimum connection time
  - >100.000 rules
  - Many attributes
  - Tight latency constraints

- Expect significant performance boost over existing engine (Drools)

Why FPGAs?

CPU

- Deterministic FA
- Sorting = classic algorithms
- Hashing (simple functions)
- Thread level parallelism

FPGA

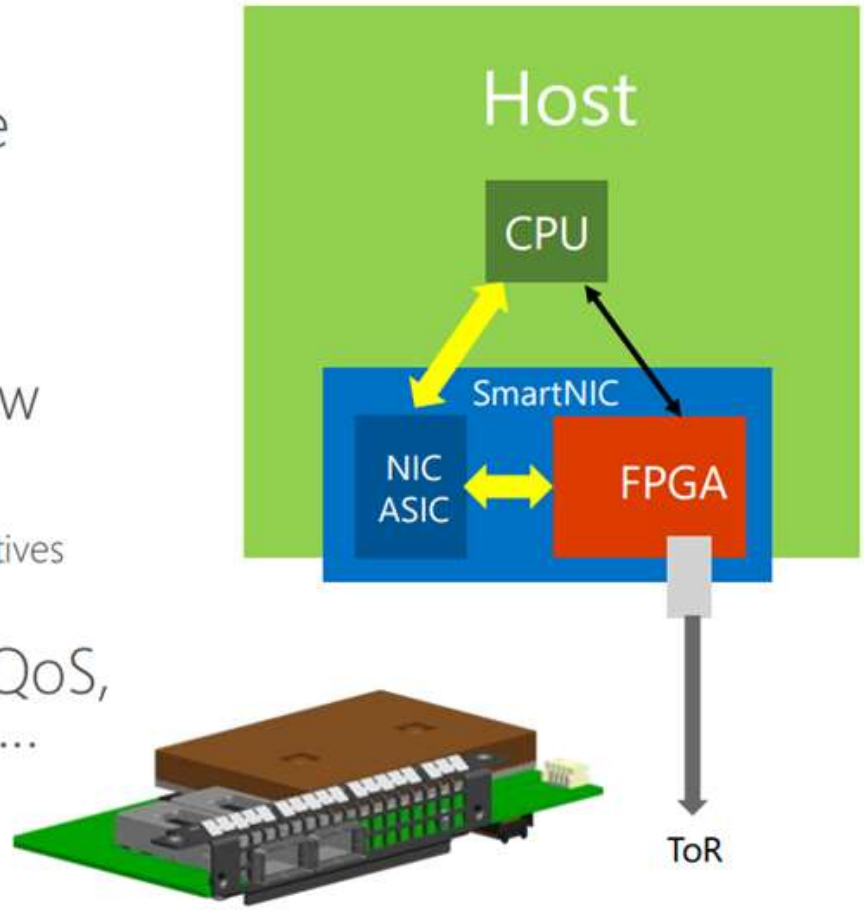
- Non deterministic FA
- Sorting network
- Robust hashing
- Deep pipelining

Architectures for future streaming engines

Rethink what processing means

Azure SmartNIC

- Use an FPGA for reconfigurable functions
  - FPGAs are already used in Bing (Catapult)
  - Roll out Hardware as we do software
- Programmed using Generic Flow Tables (GFT)
  - Language for programming SDN to hardware
  - Uses connections and structured actions as primitives
- SmartNIC can also do Crypto, QoS, storage acceleration, and more…



Microsoft

Local smart storage

IBEX

Smart Samsung SSD+FPGA

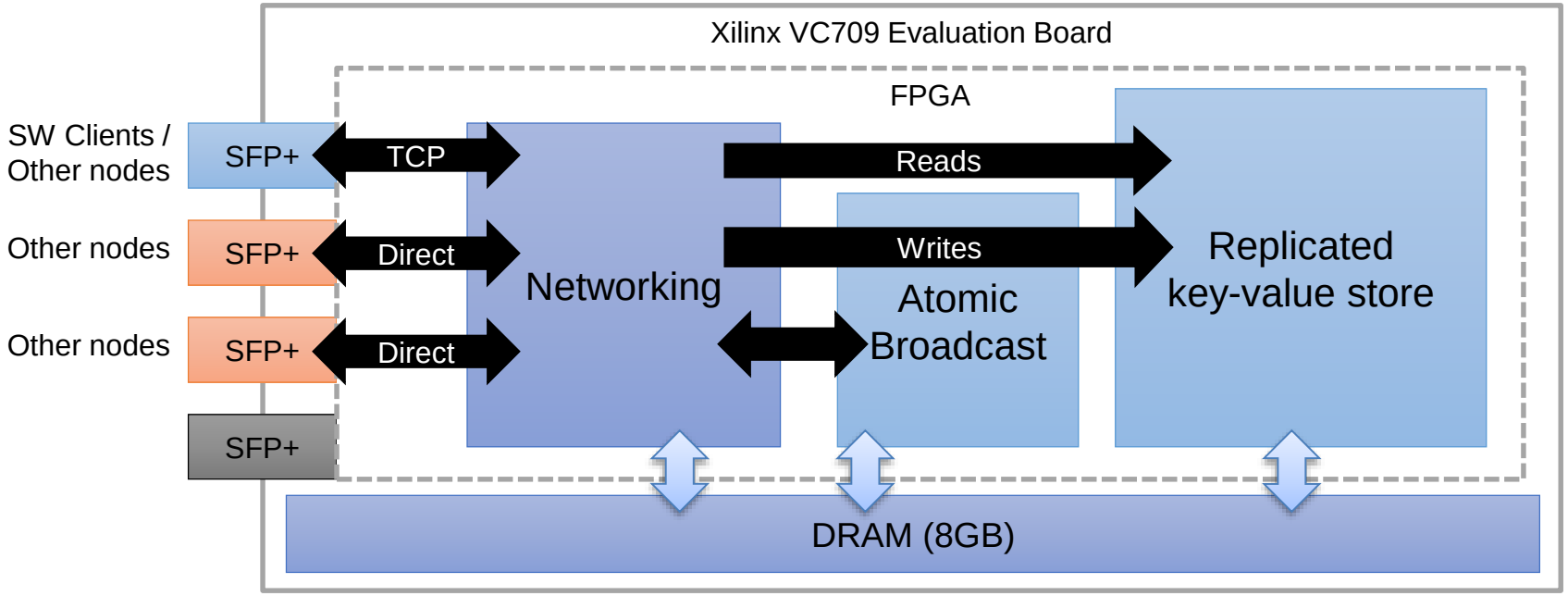
FPGA board

SSD

(Woods, PVLDB'10Woods, PVLDB'14; Woods, SIGMOD'13)

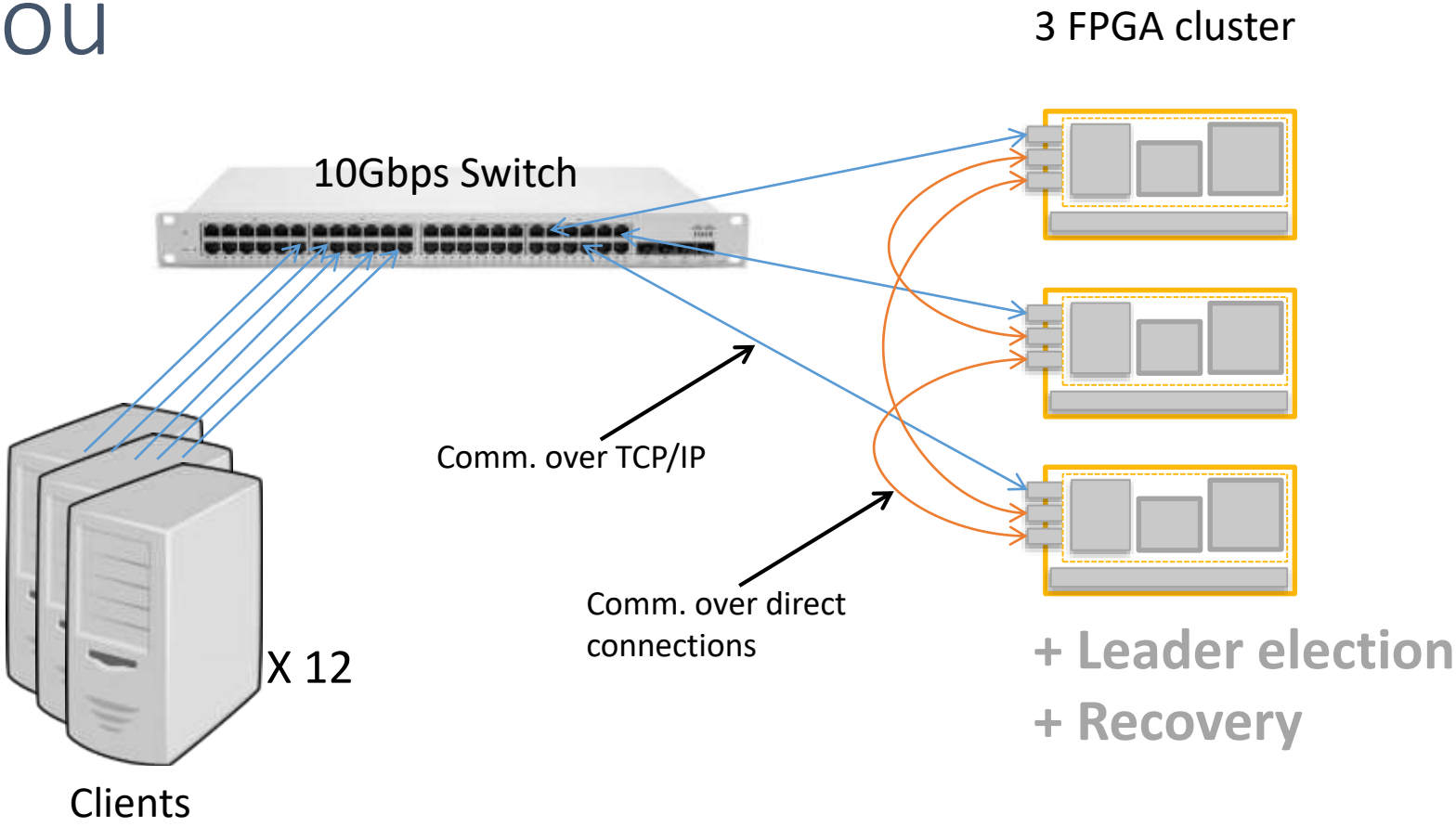
Remote Smart Storage/Memory

Caribou



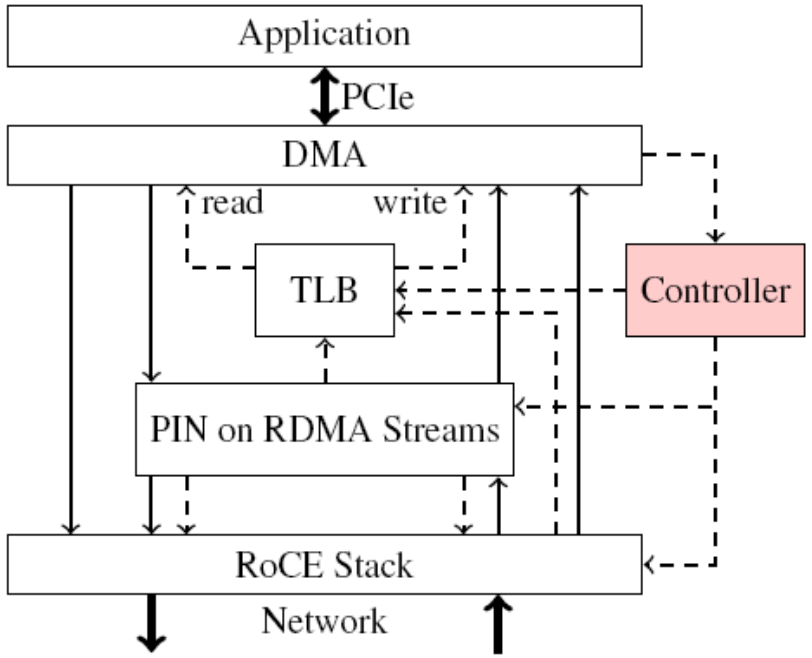
(Istvan et al, NSDI'16; Sidler, FPL'16, Istvan, PVLDB'17)

Caribou



- Drop-in replacement for memcached with Zookeeper's replication
- Standard tools for benchmarking (libmemcached)
  - Simulating 100s of clients

RDMA on FPGA



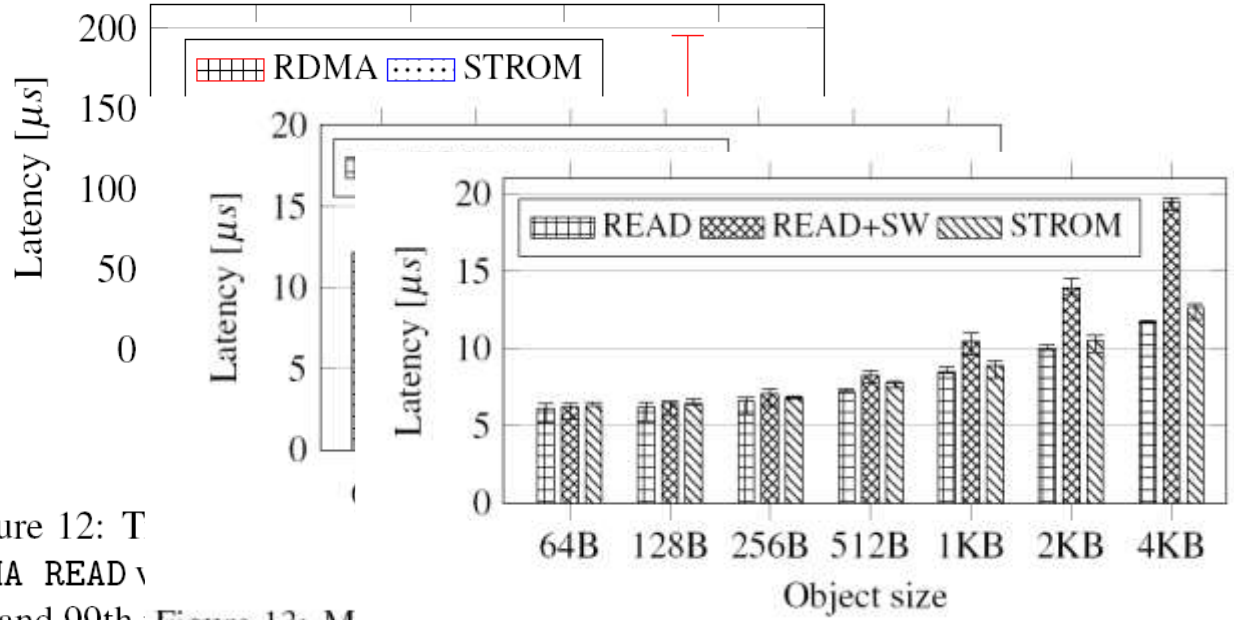
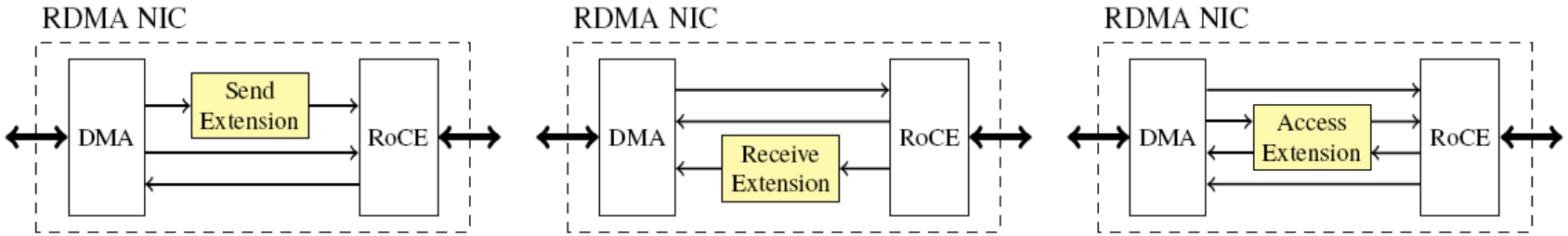


Figure 12: T
RDMA READ
1st and 99th

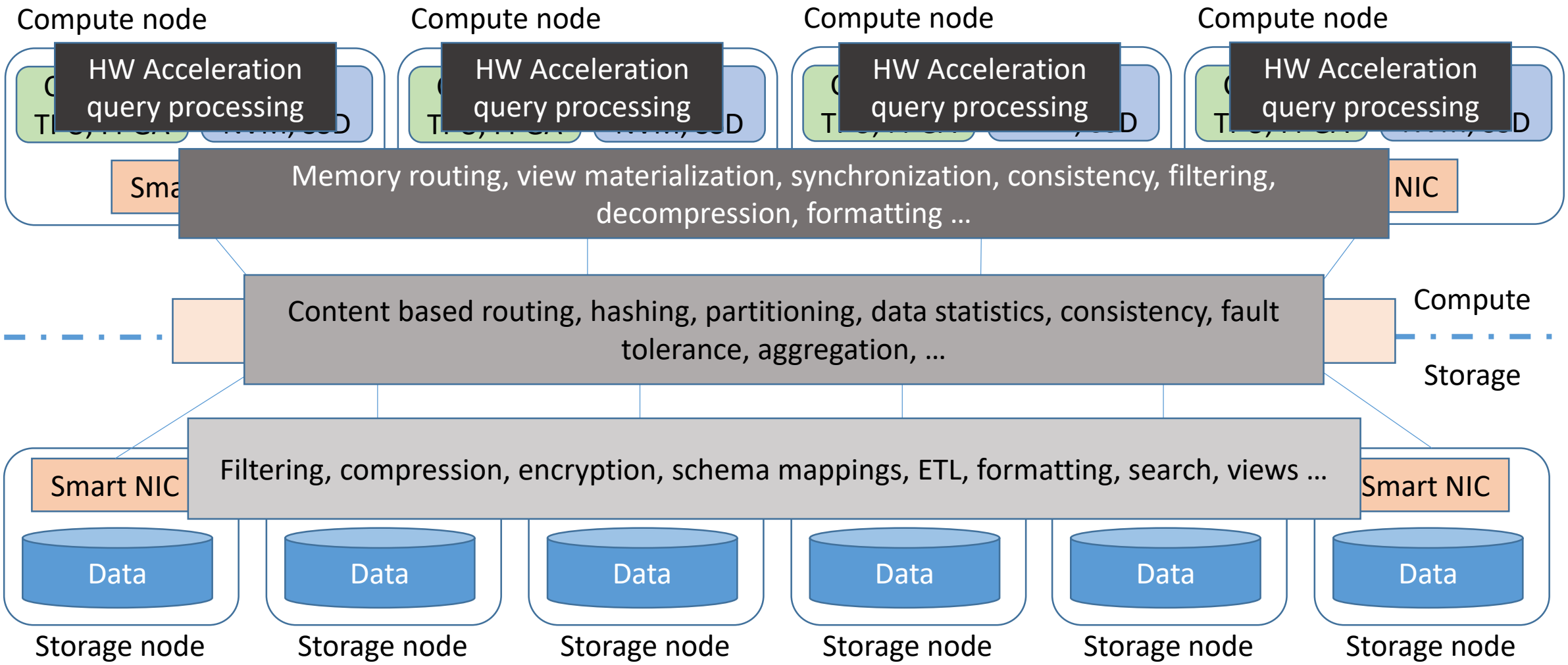
Figure 13: M
READ vs a *tra*
Error bars are

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

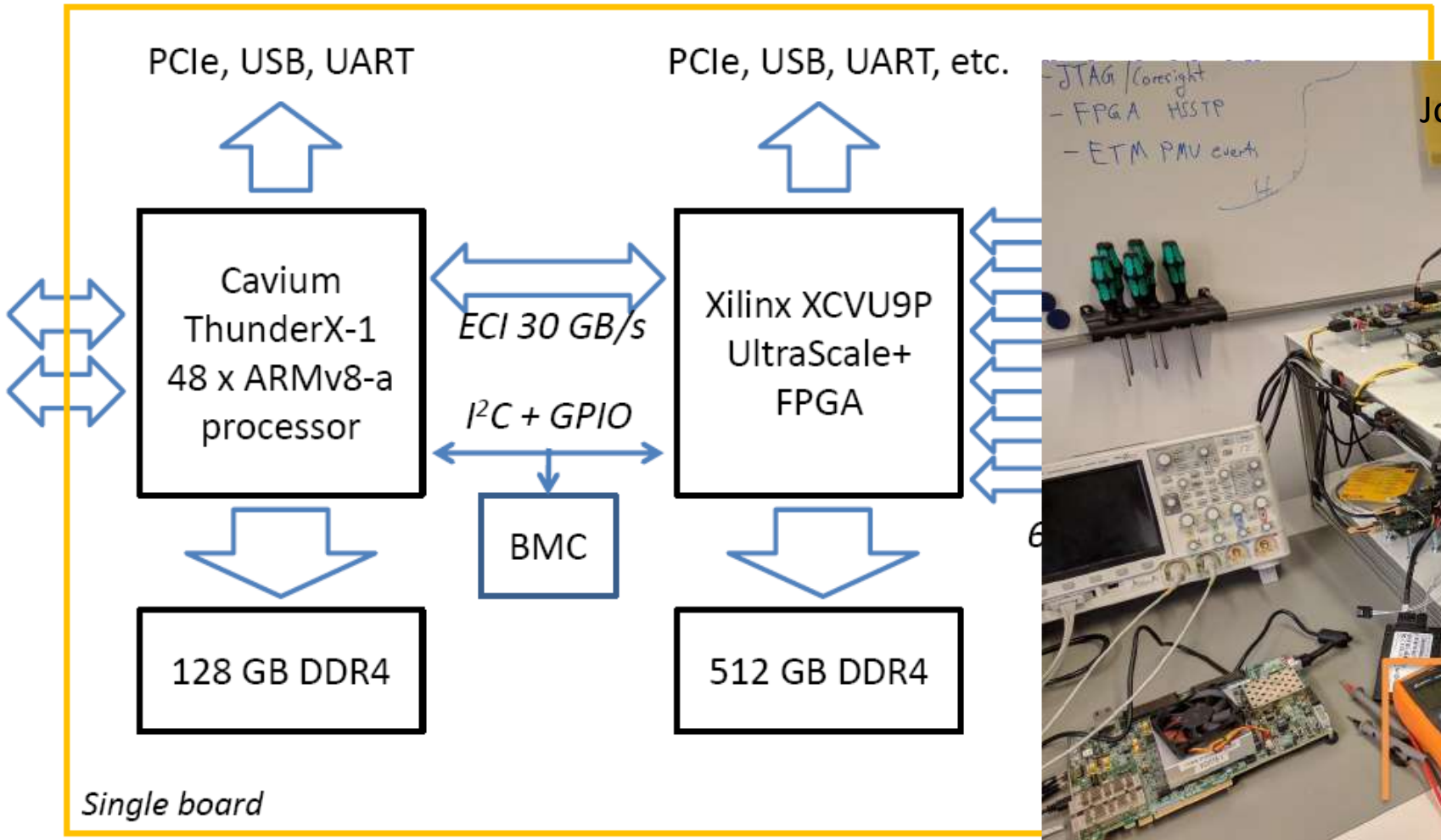
A vision for future data processing

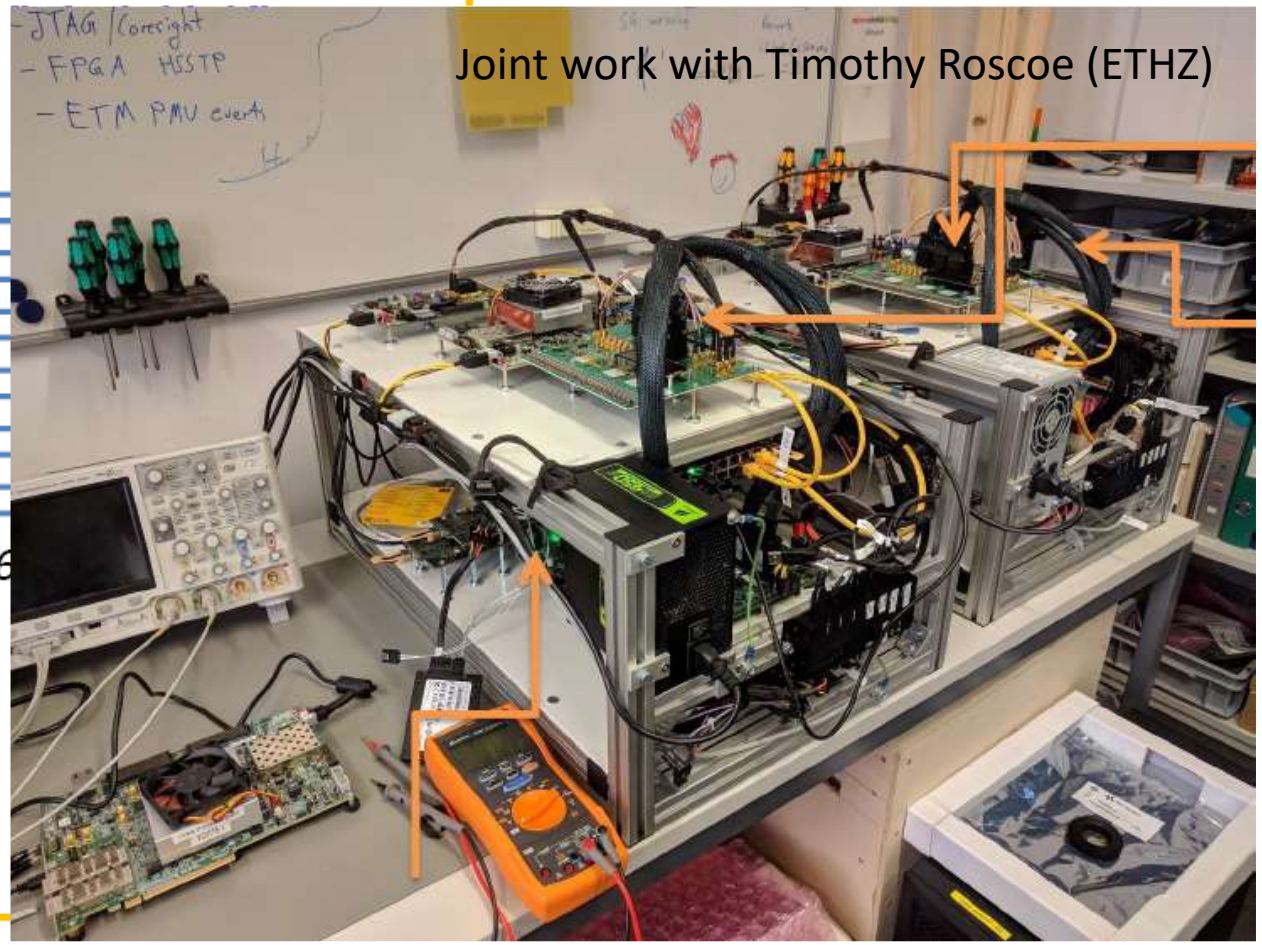


Systems@ETH zürich

Rethinking computing nodes

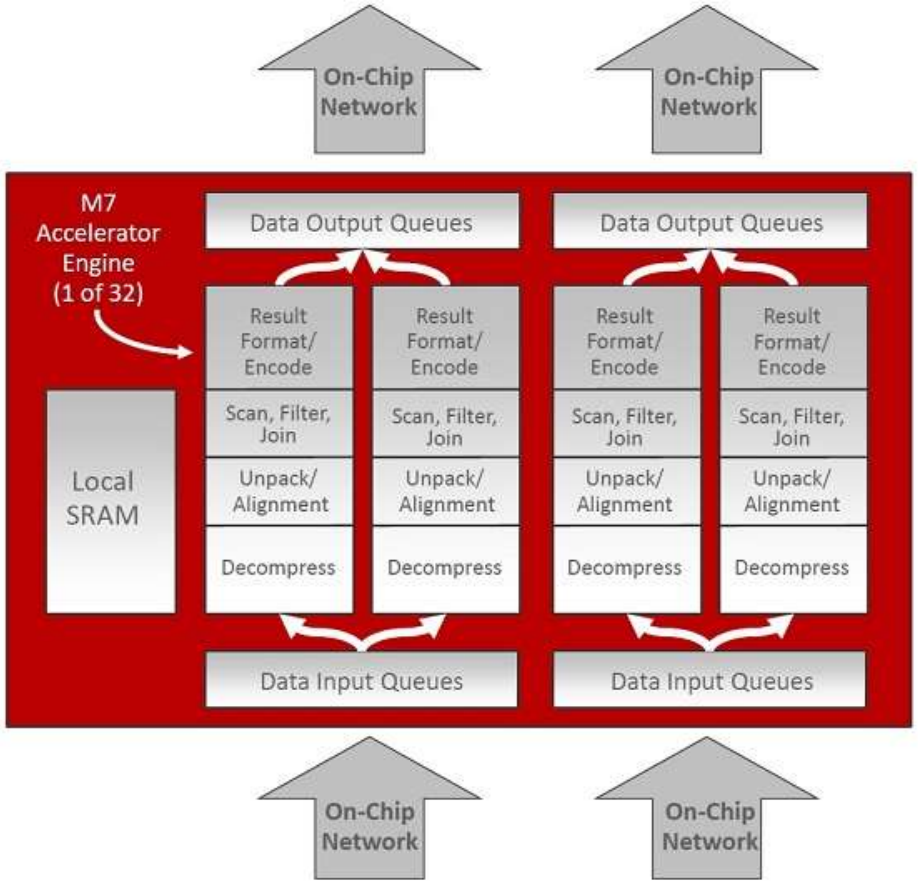
Enzian

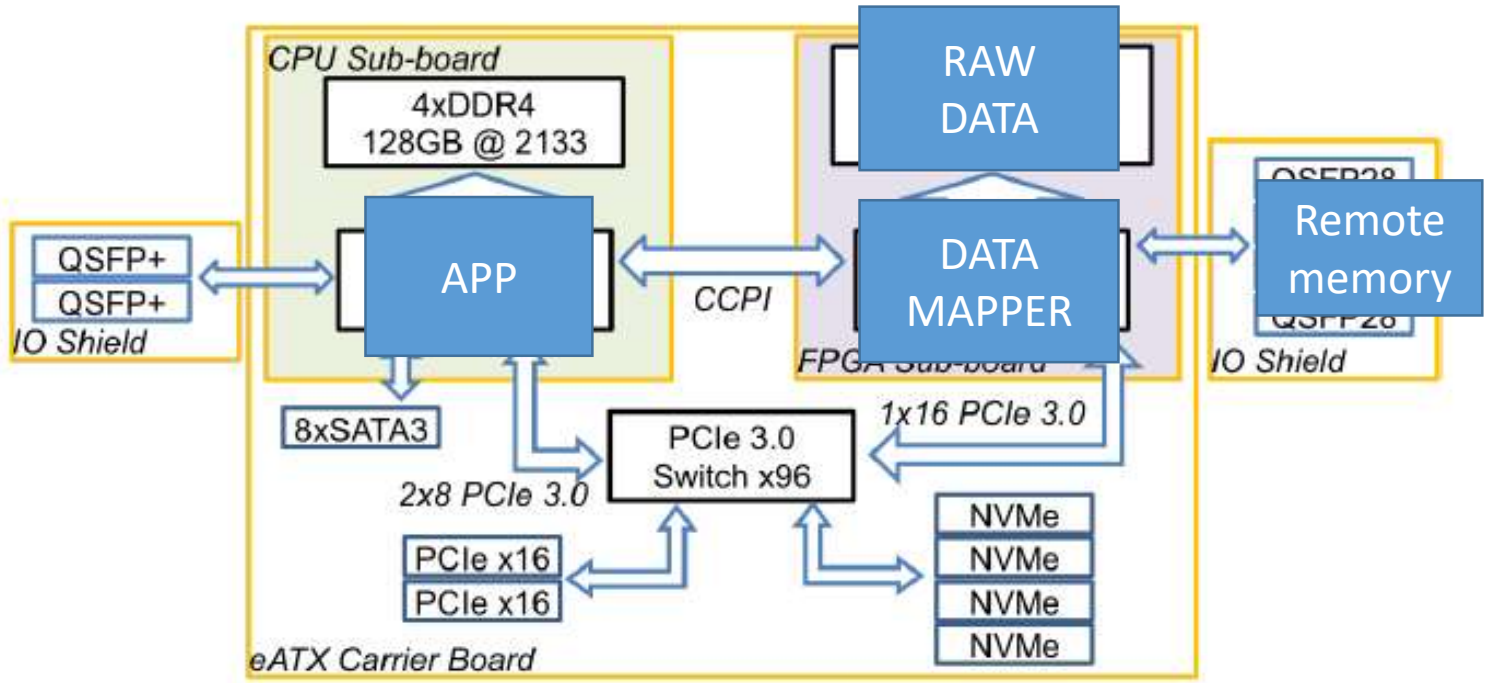




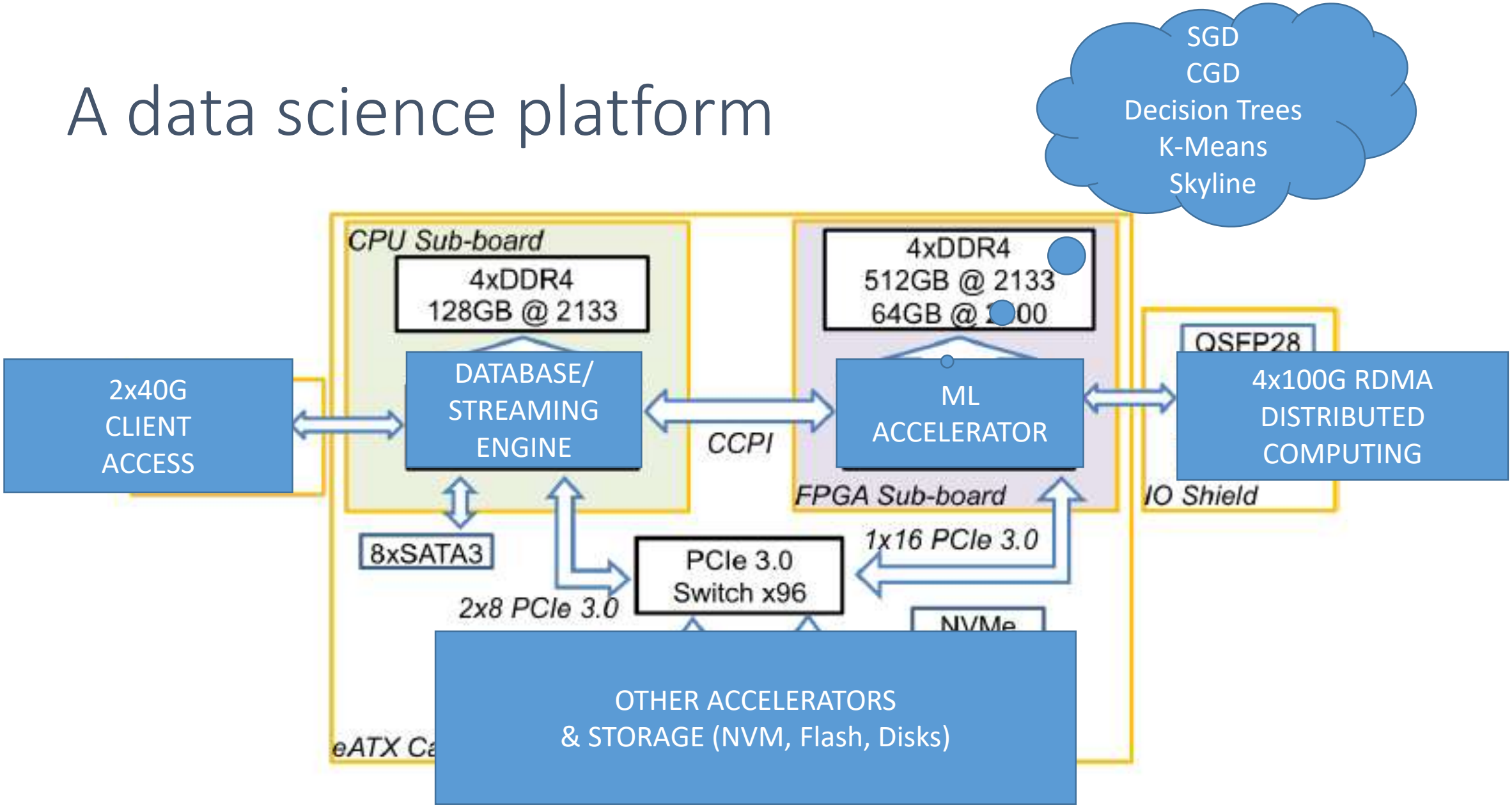
Joint work with Timothy Roscoe (ETHZ)

Near-memory processing as streams





A data science platform



Conclusions

- Hardware is opening a wealth of design opportunities
- Software needs to become more flexible and versatile, CPU only designs will not have the necessary performance
- New hardware trends can help to develop the new event and streaming systems