Enzian – A platform for research in software-hardware co-design

Gustavo Alonso

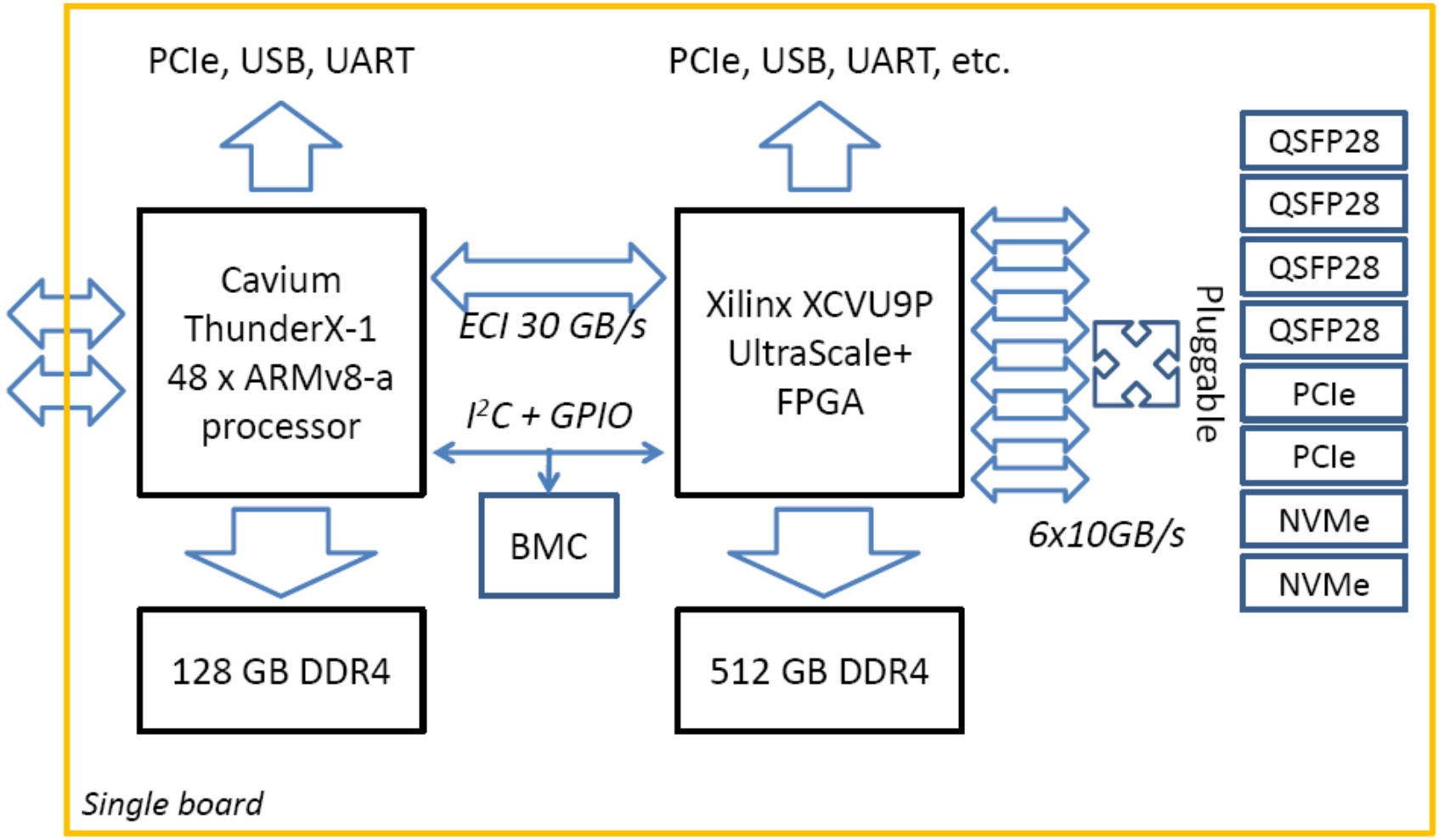
Systems Group

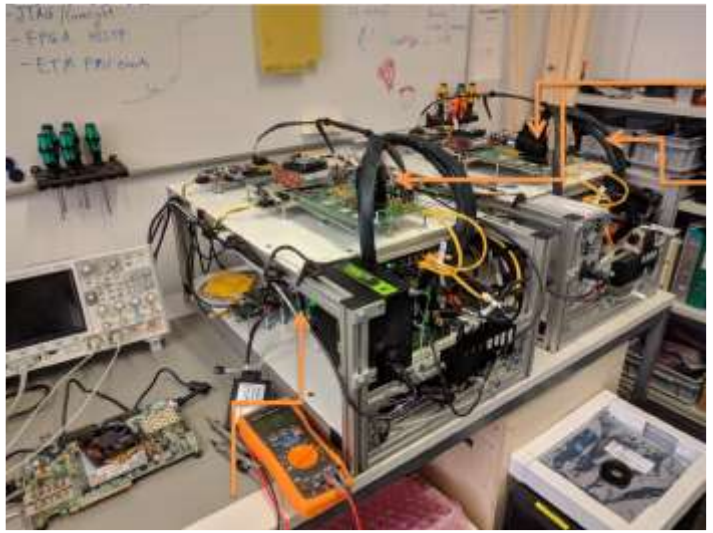
Department of Computer Science

ETH Zurich, Switzerland

Systems@ETH zürich

A platform for exploration: Enzian





Joint work with Timothy Roscoe (ETHZ)

Brief outline

- Who
- When
- Why
- What
- Wtf? (What’s that for?)

Who

Systems Group, ETH Zurich

- 4 (5) faculty
  - Databases
  - Operating Systems
  - Distributed Systems
  - Machine Learning
  - Networking
  - Data center and cloud
- 12 postdocs
- 32 PhD students

www.systems.ethz.ch

When

Now, but we have been at this for a while …

- … but now is more difficult to argue against it ☺
- At this stage, everybody shows the same graph
- And talks at length about
  - Moore’s law
  - Dennard’s scaling
  - Dark silicon
  - Feature sizes
  - Production cost
  - Trump and/or Brexit
  - …

A different take

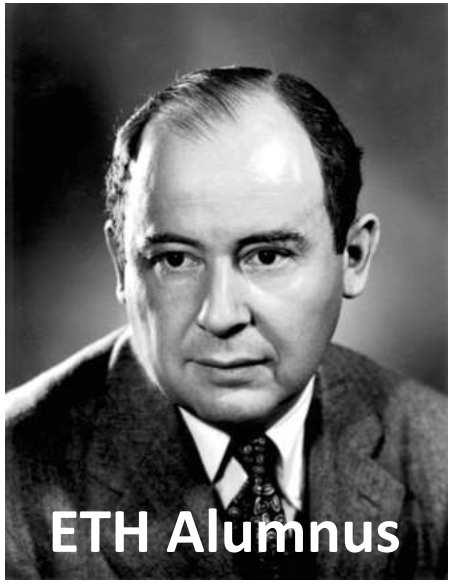
- General purpose is history
- Specialization is the name of the game
- Cloud computing is the game changer:
  - Turning IT into a service industry instead of a manufacturing gig
  - Hyperscalers dominating the market
  - Economies of scale
  - Not have to worry about client side deployment
  - Build what works best for the job (conventional computers ain’t it)

Systems@ETH zürich

Why

A research agenda

- **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?



Background

- Our focus in the last 10+ years has been to understand algorithm and system development on non von Neumann architectures

- Frequent item (ICDE'10, TKDE'11)
- Skyline (FCCM'13, TRETS'15)
- Histogram calculations (SIGMOD'14)
- Median (PVLDB'09)
- Sorting networks (VLDB Journal'12)
- Complex event processing (PVLDB'10)
- Decision Trees (FPL'17,FPL'18,VLDB'20)
- Stochastic Gradient Descent (FCCM'17, VLDB'19)
- Hashing (FPL'13, FPL'16)
- Joins, Group By (PVLDB'14)
- Data partitioning (SIGMOD'17)
- TCP/IP (FCCM'15, FPL'16, FPL'19)
- RoCE (ongoing)
- KVS (NSDI'16, PVLDB'17)
- String Processing (SIGMOD'17)
- K-Means (FPL'18)

Successful but …

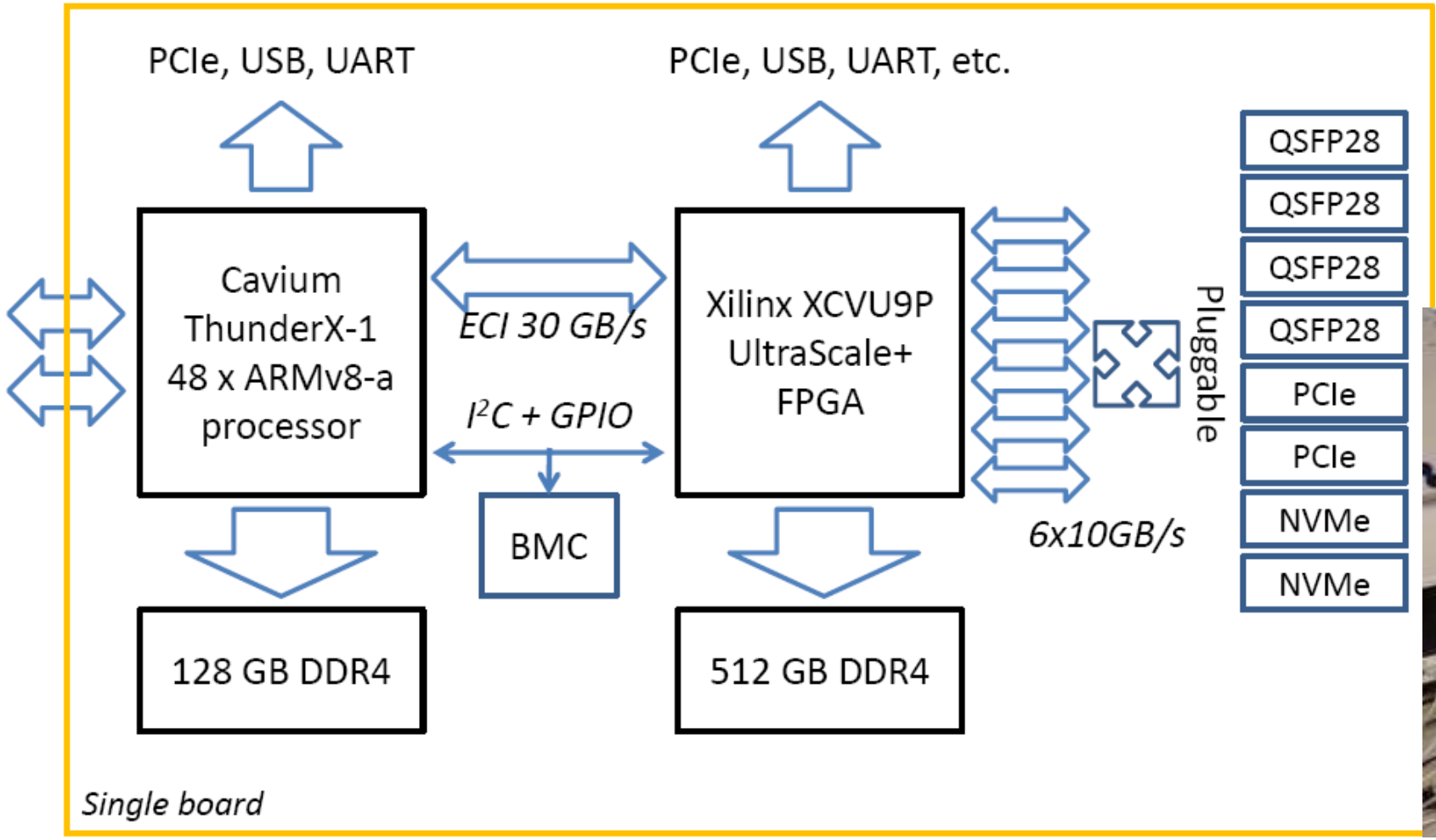
- We have a long list of open source systems and results already in use by industry, having built a number of systems before products appeared
  - TCP/IP stack (now at 100 Gbs)
  - RDMA stack (now at 100 Gbs)
  - Decision tree ensemble acceleration
  - String processing
  - SATA driver
  - …
- The lesson learned is that hardware always gets on the way

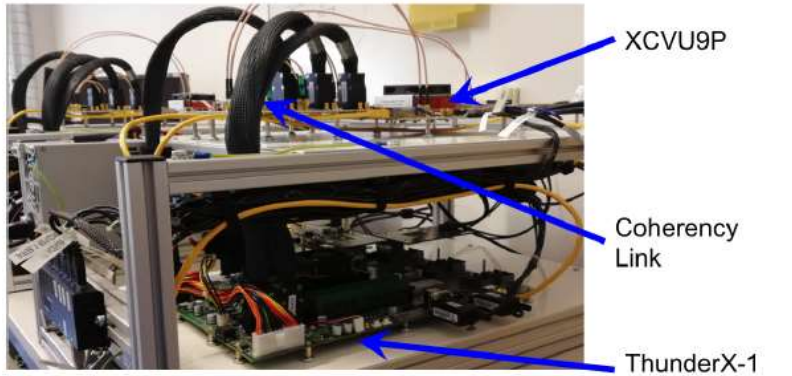
Research on hardware when hardware is the bottleneck

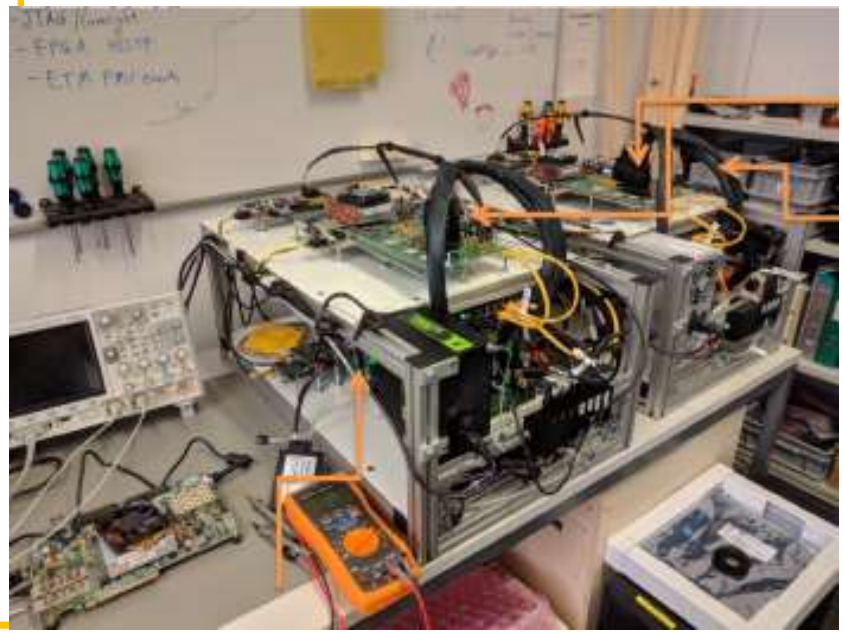
- We are very grateful for numerous donations of equipment and we have been privileged with very useful feedback from industry
  - Microsoft
  - Xilinx
  - Intel
  - Oracle
- But existing hardware just does not cut it
  - Designed by somebody else for something else
  - Close source, many black boxes
  - Woefully inadequate for data processing

What

A platform for exploration: Enzian

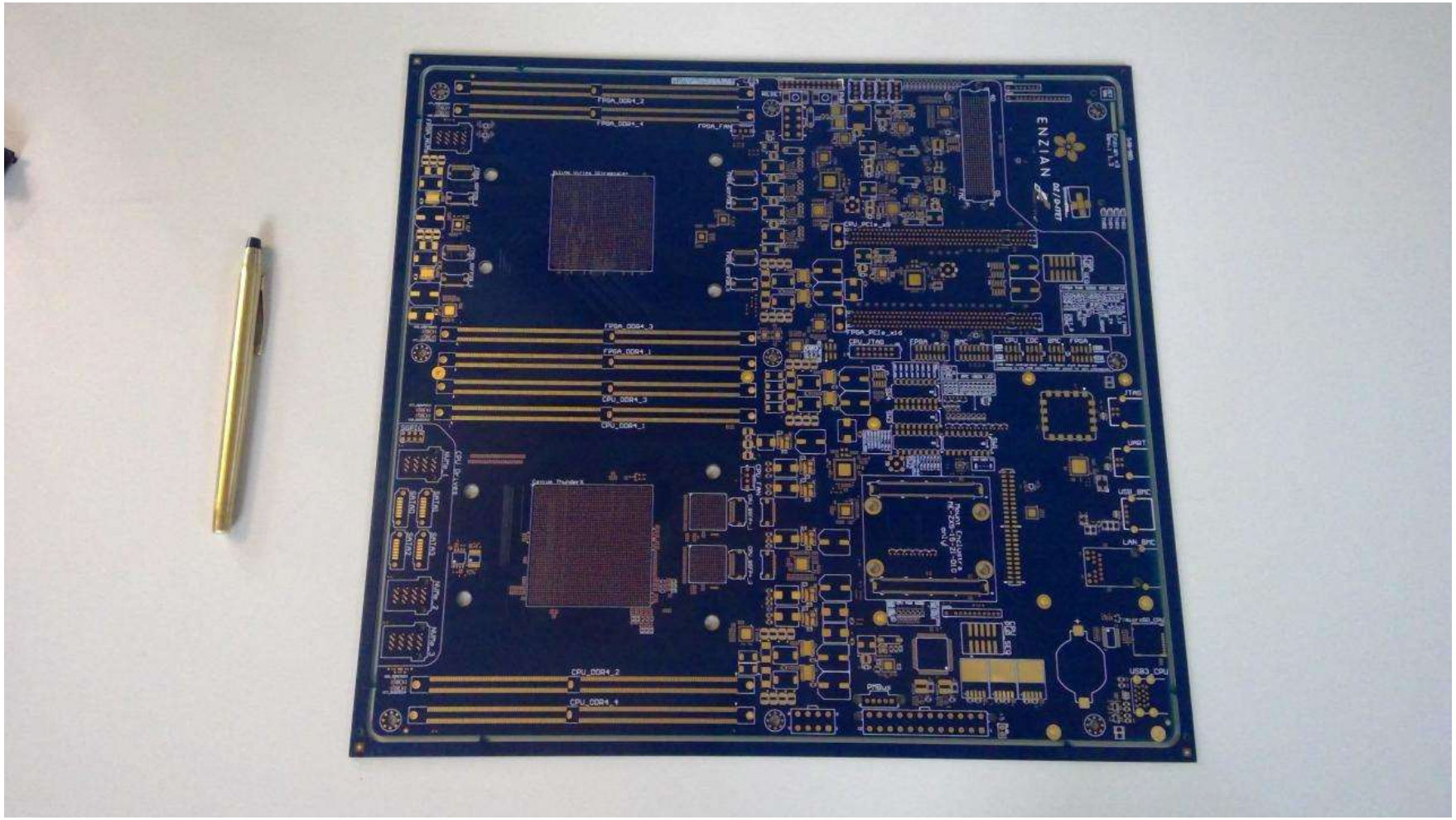






Joint work with Timothy Roscoe (ETHZ)

Status



Plus accompanying software

- Lynx: our own FPGA OS with
  - Shared virtual address space host-accelerator
  - Multi-region dynamic reconfiguration
  - Networking stacks (TCP/IP, RoCE)
  - Memory management interfaces
- Cache coherent stack on FPGA
- Simulators and monitoring tools
- A database (doppioDB)
- Working on a data streaming platform

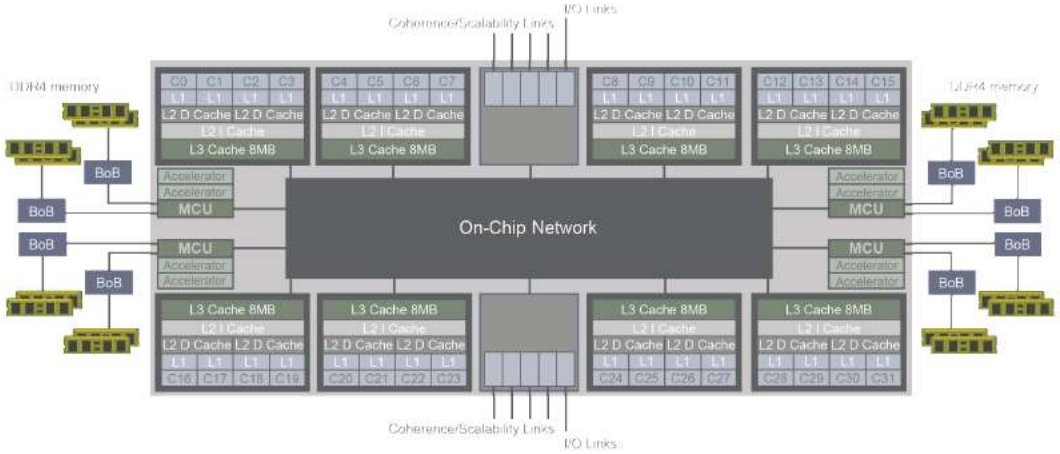
Goals for 2020

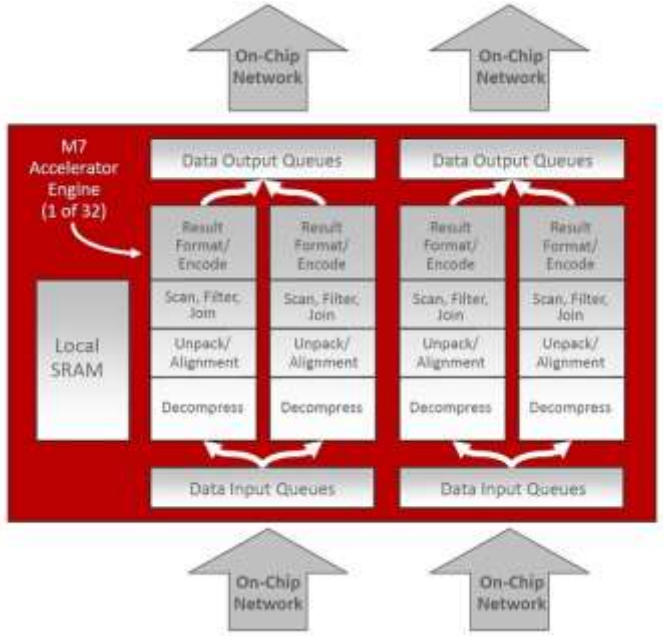
- Release everything as open source (including blue prints for the hardware)
- Cluster of 40 machines running at ETH (accessible to others)
- Several experimental systems (data processing, run time verification, operating systems, networking)
- Ability to deliver machines to those interested in purchasing them

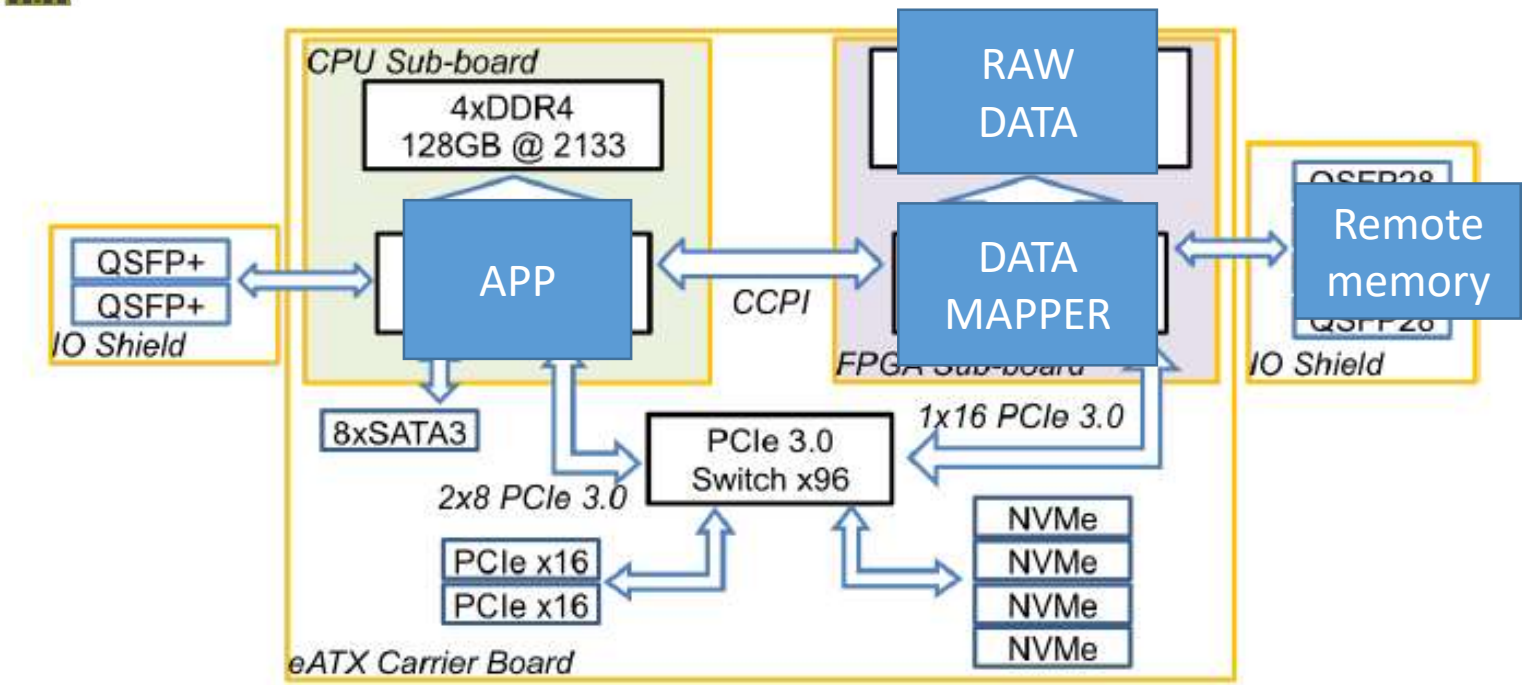
Systems@ETH zürich

Wtf? = What is that for?

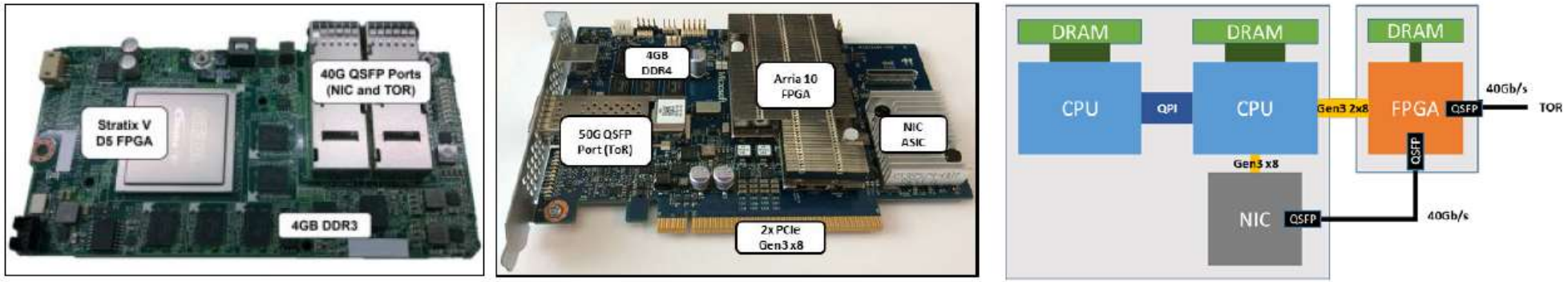
Boldly go where Oracle has gone before





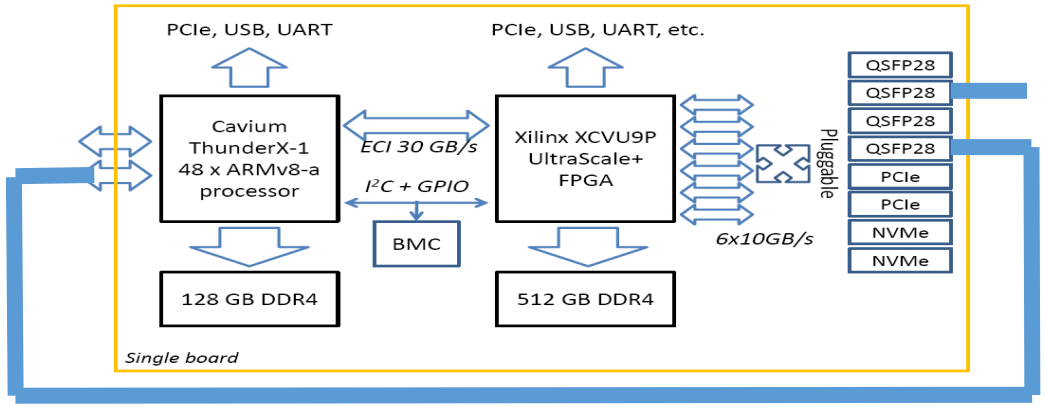


Boldly go where Microsoft has gone before

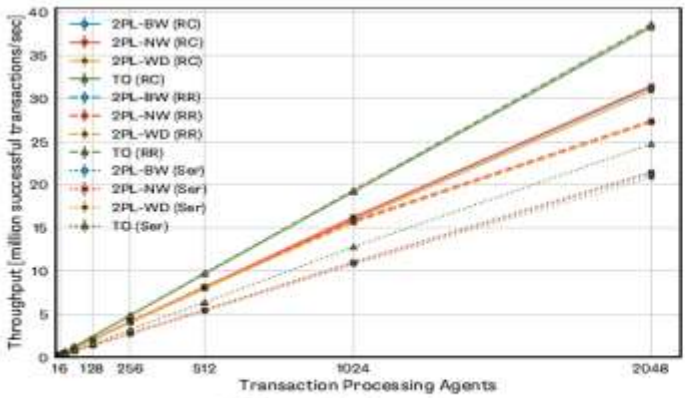


(a) Azure SmartNIC Gen1, 40GbE w/ external NIC (b) Azure SmartNIC Gen2, 50GbE w/ on-board NIC (c) *bump-in-the-wire* architecture

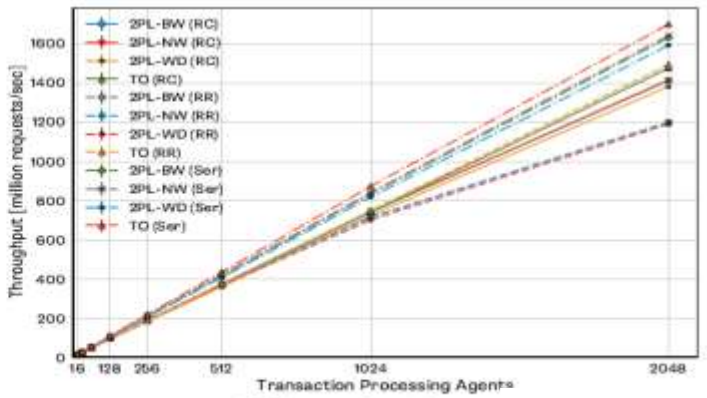
Figure 2: Azure SmartNIC boards with Bump-in-the-Wire Architecture



Transactions on a supercomputer (RDMA)



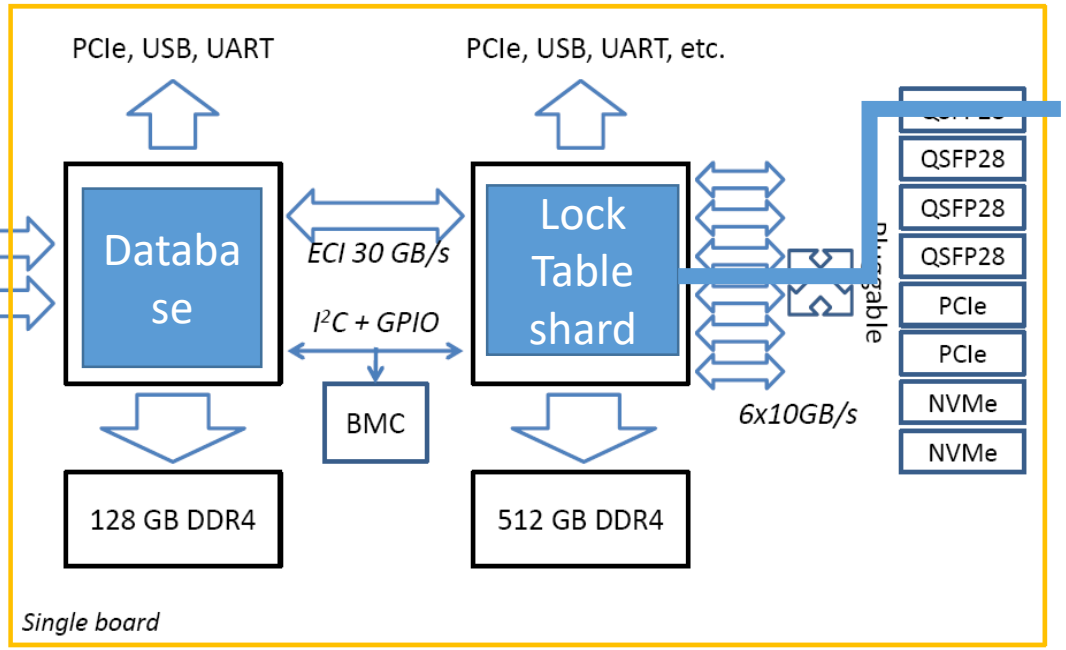
(a) Transaction throughput



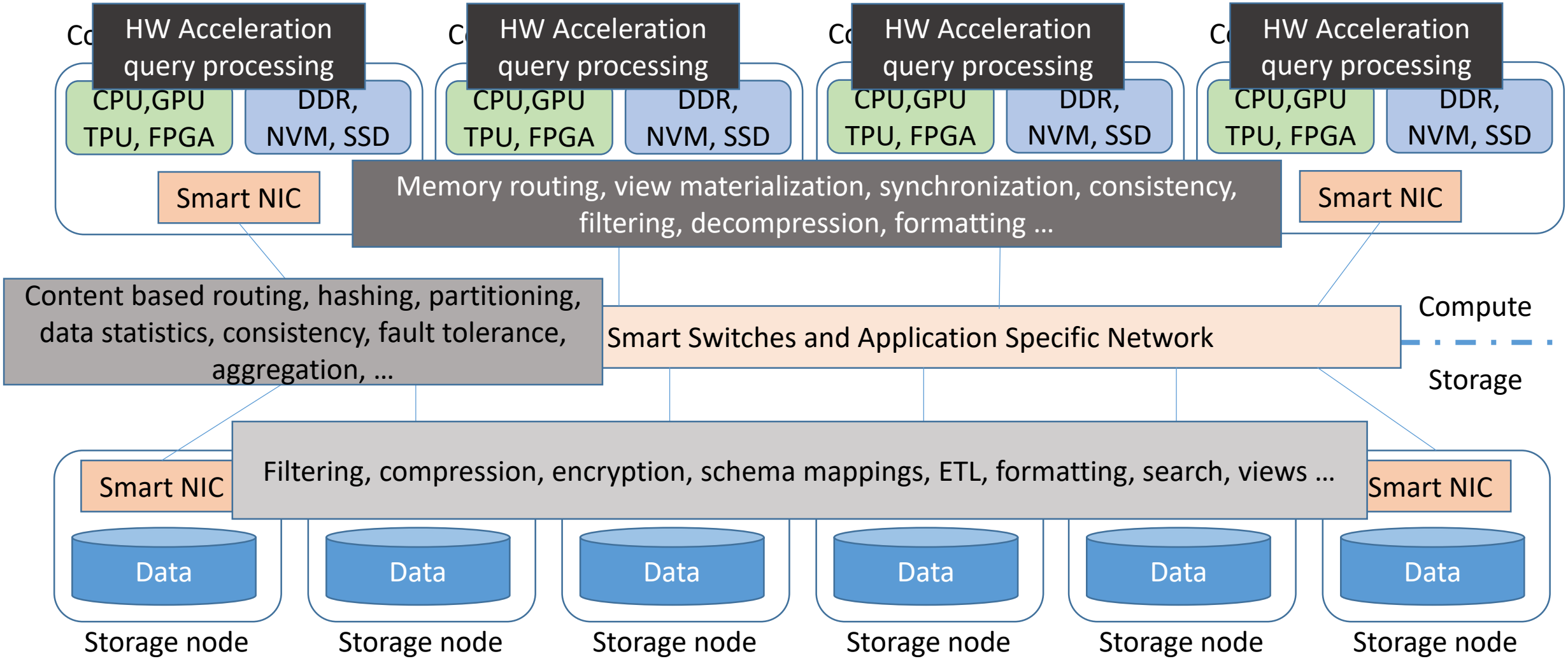
(b) Request throughput

Figure 3: Throughput of TPC-C with 2048 warehouses

Barthels et al. PVLDB 2019



A vision for future databases



For the data processing fans

- A database with shared memory architecture (over RDMA) with near data processing and network protocol customization
- A smart memory controller and buffer manager that can be dynamically programmed to pre-fetch, ETL, index, filter, sort, shape data from remote sources (memory, fabric attached storage, NVMs ...)
- A relational/streaming engine with line rate, in-network stream processing
- Auto tuning based on real time monitoring of all CPU parameters
- Extensibility to new data types and operations

Enzian is an open project

- Open source (hw and sw)
- Open to collaboration and contributions
- Aiming a creating a community
- Boosting relevant research in academia
- Creating a vehicle for exploring data processing on modern hardware

- Join us!