Reconfigurable hardware for big data

Gustavo Alonso

Systems Group

Department of Computer Science

ETH Zurich, Switzerland

www.systems.ethz.ch

Systems Group

- 7 faculty
- ~40 PhD
- ~8 postdocs

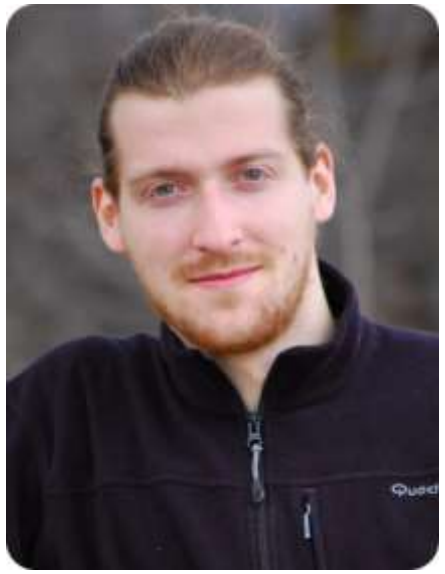
Researching all aspects of system architecture, sw and hw

The team behind the work:

David Sidler

Muhsen Owaida

Zsolt Istvan



Kaan Kara

Data processing today:

Appliances

Data Centers (Cloud)

What is a database engine?

- As complex or more complex than an operating system
- Full software stack including
  - Parsers, Compilers, Optimizers
  - Own resource management (memory, storage, network)
  - Plugins for application logic
  - Infrastructure for distribution, replication, notifications, recovery
  - Extract, Transform, and Load infrastructure
- Large legacy, backward compatibility, standards
- Hugely optimized

Databases are blindly fast at what they do well

Date Submitted	Company	System	Performance (tpmC)	Price/tpmC
11/25/14	SAP	Dell PowerEdge T620	112,890	.19 USD
03/26/13	ORACLE	SPARC T5-8 Server	8,552,523	.55 USD
02/22/13	IBM	IBM System x3650 M4	1,320,082	.51 USD
09/27/12	CISCO	Cisco UCS C240 M3 Rack Server	1,609,186	.47 USD
04/11/12	IBM	IBM Flex System x240	1,503,544	.53 USD
03/27/12	ORACLE	Sun Server X2-8	5,055,888	.89 USD
01/17/12	ORACLE	Sun Fire X4800 M2 Server	4,803,718	.98 USD

Databases = think big

ORACLE EXADATA

From Oracle documentation

Database Grid

- 8 Dual-processor x64 database servers

OR

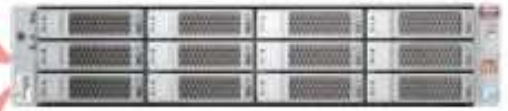
- 2 Eight-processor x64 database servers

InfiniBand Network

- Redundant 40Gb/s switches
- Unified server & storage network

Intelligent Storage Grid

- 14 High-performance low-cost storage servers



- 100 TB **High Performance** disk, or 336 TB **High Capacity** disk
- **5.3 TB PCI Flash**
- Data mirrored across storage servers

Database engine trends: Appliances

Oracle:

- T7, SQL in Hardware, RAPID

SAP:

- OLTP+OLAP on main memory
- Hana on SGI supercomputer

SAP Hana on SGI UV 300H
SGI documentation

The challenge of hardware acceleration

If it sounds too good to be true ..

Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU

Victor W Lee†, Changkyu Kim†, Jatin Chhugani†, Michael Deisher†,
Daehyun Kim†, Anthony D. Nguyen†, Nadathur Satish†, Mikhail Smelyanskiy†,
Srinivas Chennupaty⋆, Per Hammarlund⋆, Ronak Singhal⋆ and Pradeep Dubey†

victor.w.lee@intel.com

†Throughput Computing Lab,
Intel Corporation

⋆Intel Architecture Group,
Intel Corporation

ABSTRACT

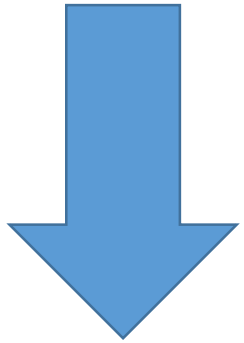
Recent advances in computing have led to an explosion in the amount of data being generated. Processing the ever-growing data in a timely manner has made throughput computing an important aspect for emerging applications. Our analysis of a set of important throughput computing kernels shows that there is an ample amount of parallelism in these kernels which makes them suitable for today's multi-core CPUs and GPUs. In the past few years there have been many studies claiming GPUs deliver substantial speedups (between 10X and 1000X) over multi-core CPUs on these kernels. To

The past decade has seen a huge increase in digital content as more documents are being created in digital form than ever before. Moreover, the web has become the medium of choice for storing and delivering information such as stock market data, personal records, and news. Soon, the amount of digital data will exceed exabytes (10^{18}) [31]. The massive amount of data makes storing, cataloging, processing, and retrieving information challenging. A new class of applications has emerged across different domains such as database, games, video, and finance that can process this huge amount of data to distill and deliver appropriate content to

Usual unspoken caveats in HW acceleration

- Where is the data to start with?
- Where does the data has to be at the end?
- What happens with irregular workloads?
- What happens with large intermediate states?
- What is the architecture?
- Is the design preventing the system from doing something else?
- Can the accelerator be multithreaded?
- Is the gain big enough to justify the additional complexity?
- Can the gains be characterized?

Do not replace, enhance



Help the CPU to do what it does not do well

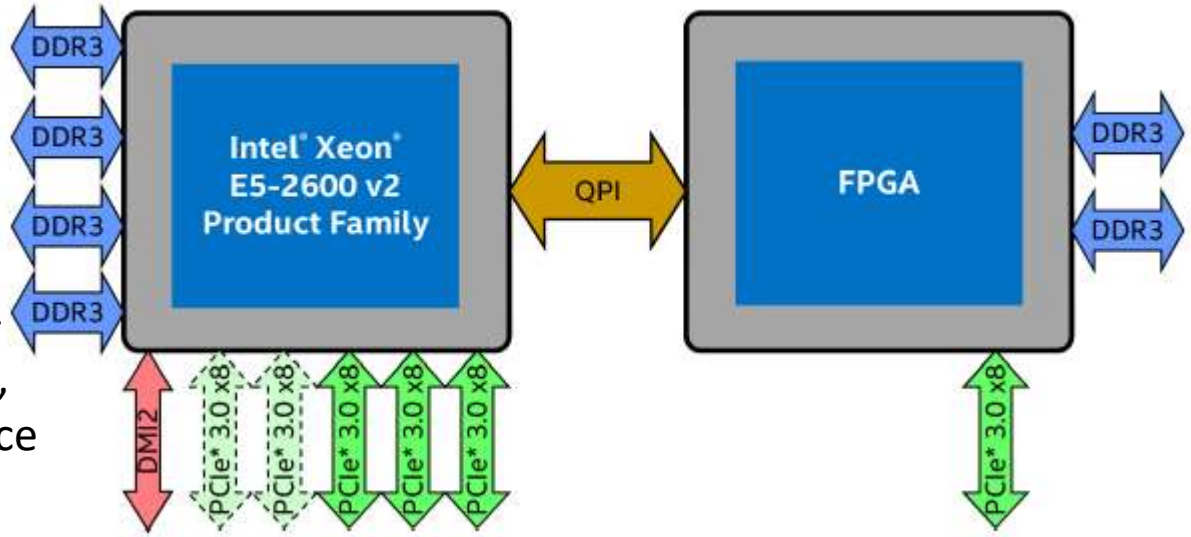
Text search in databases

Istvan et al, FCCM'16

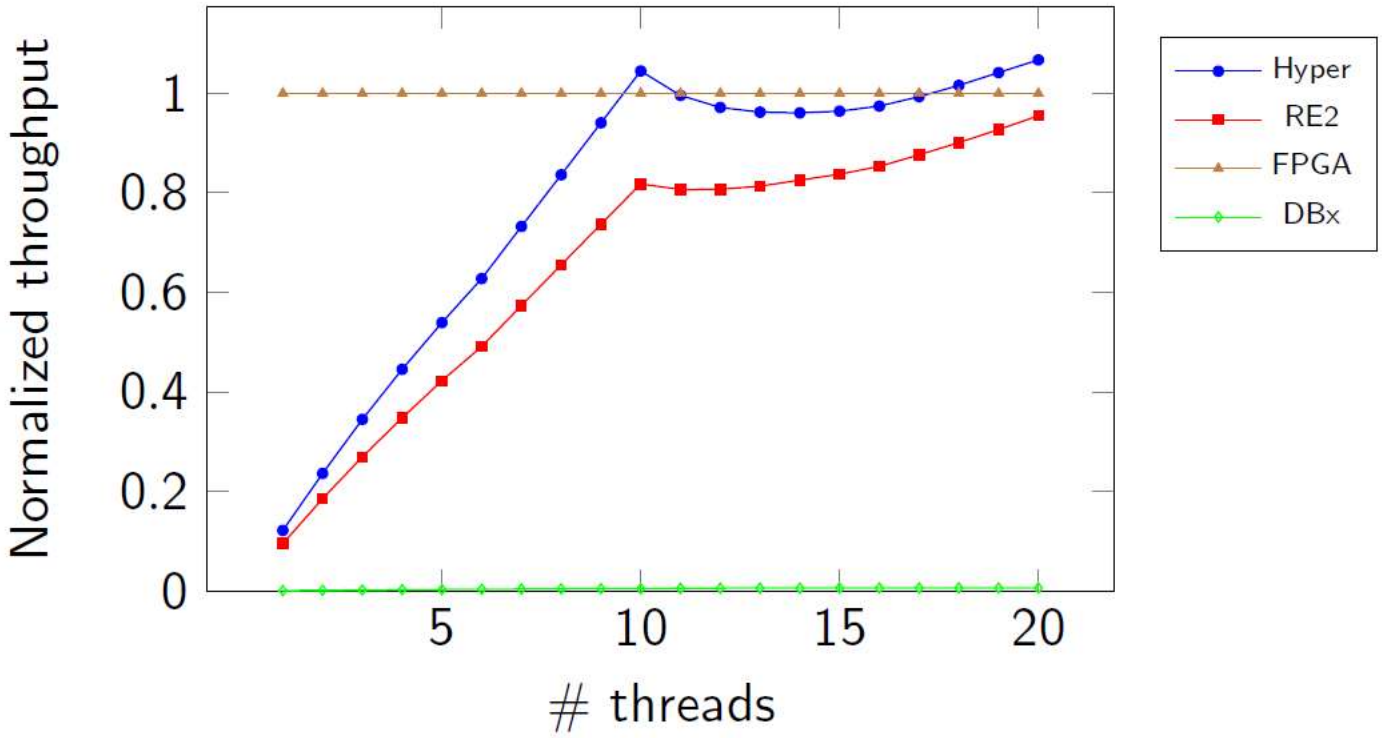
3 Queries with increasing complexity:

```
Q1:     SELECT count(*) FROM address_table
WHERE addr_string LIKE '%Strasse%';

Q2:     SELECT count(*) FROM address_table
WHERE addr_string LIKE '%Alan%Turing%Cheshire%';

Q3:     SELECT count(*) FROM address_table
WHERE REGEXP_LIKE(addr_string, '(Strasse|Str\.)
    .*(8[0-9]{4})');
```

INTEL HARP:
This is an experimental system provided by Intel any results presented are generated using pre-production hardware and software, and may not reflect the performance of production or future systems.

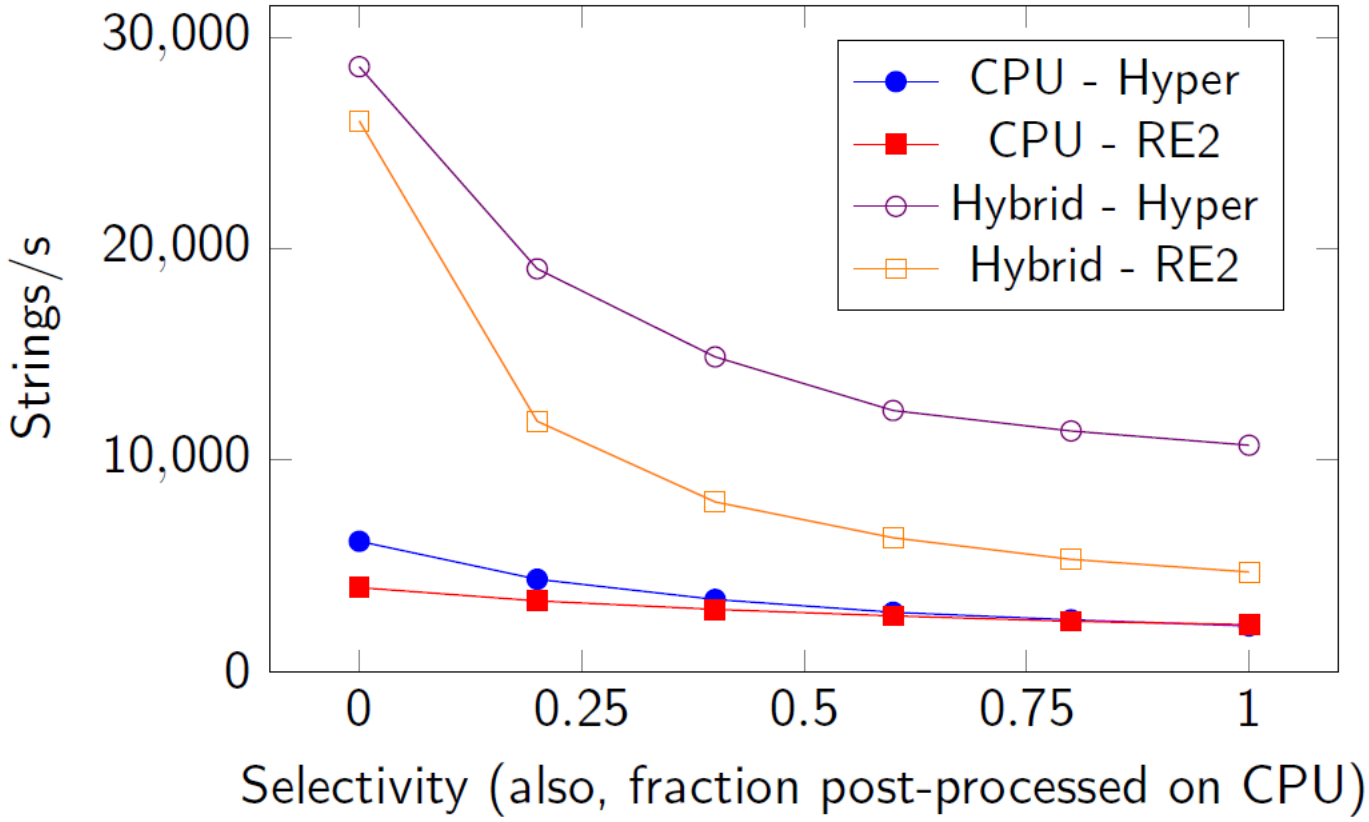


100% processing on FPGA



Q3: `WHERE REGEXP_LIKE(address_string, '(Strasse|Str\.).*(8[0-9]4)'`

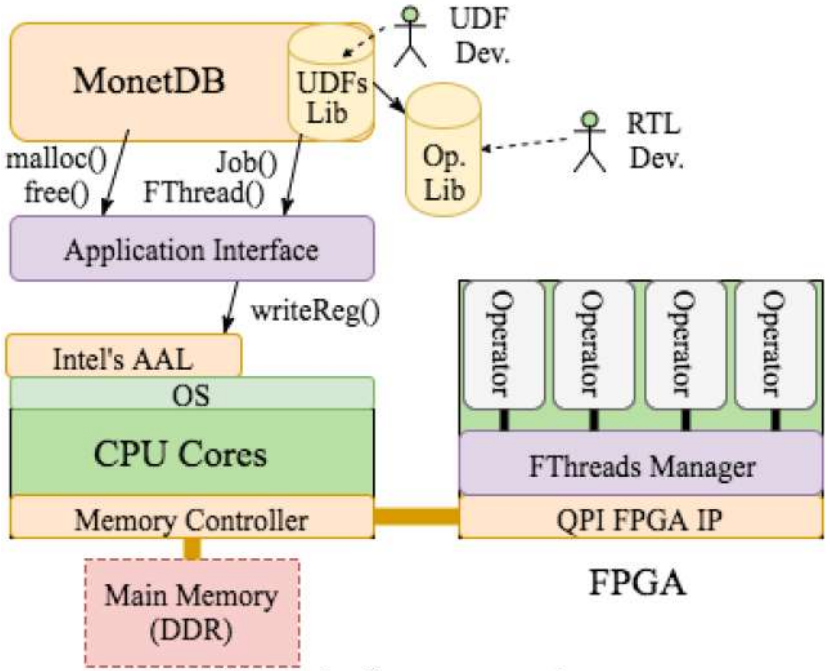
Hybrid Processing CPU/FPGA

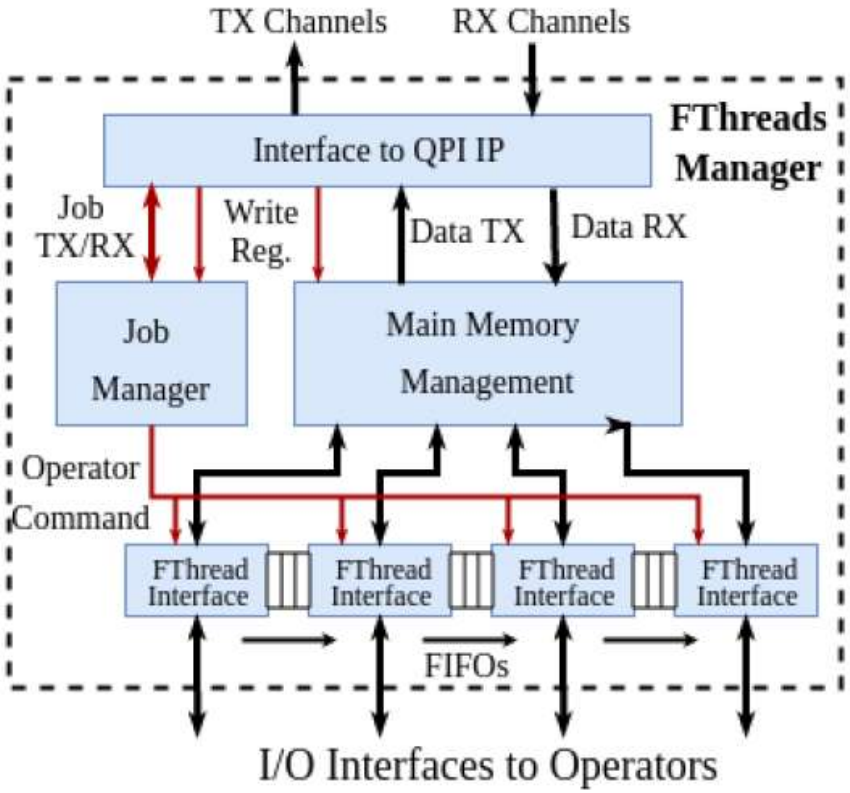


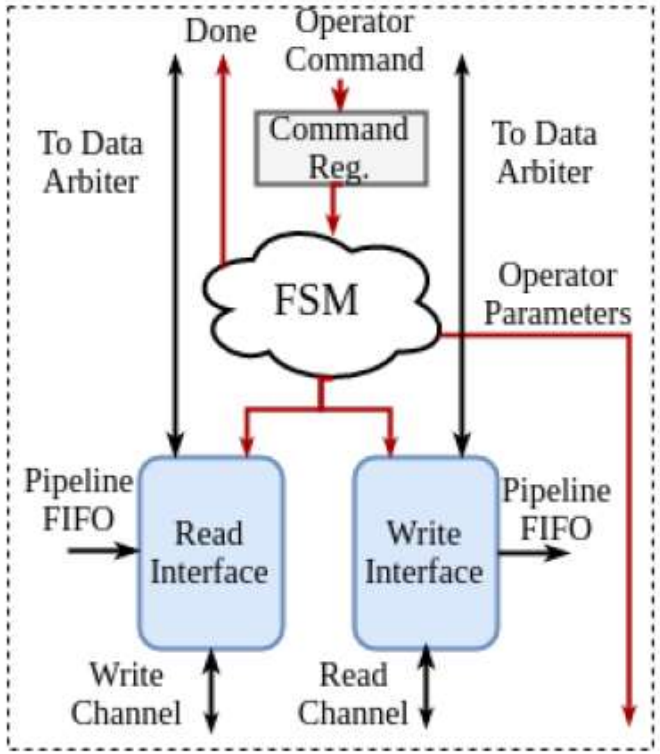
Regular expression: `'(Strasse|Str\.).*(8[0-9]4).*delivery'`

Inside a real database …

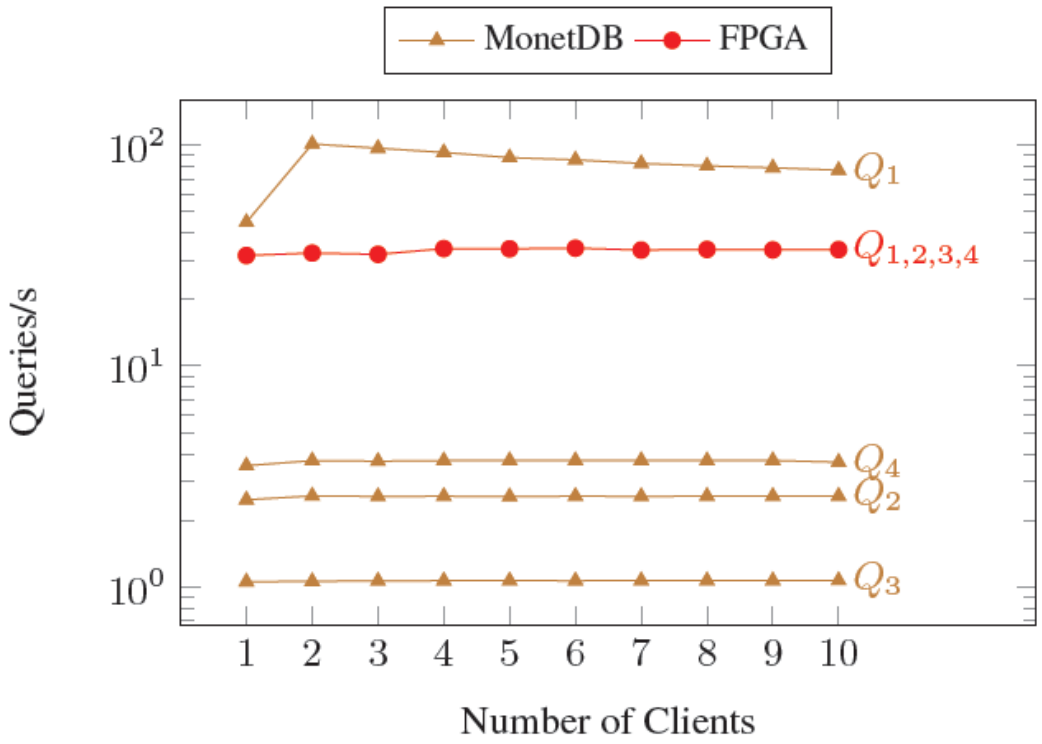
Sidler et al., SIGMOD'17



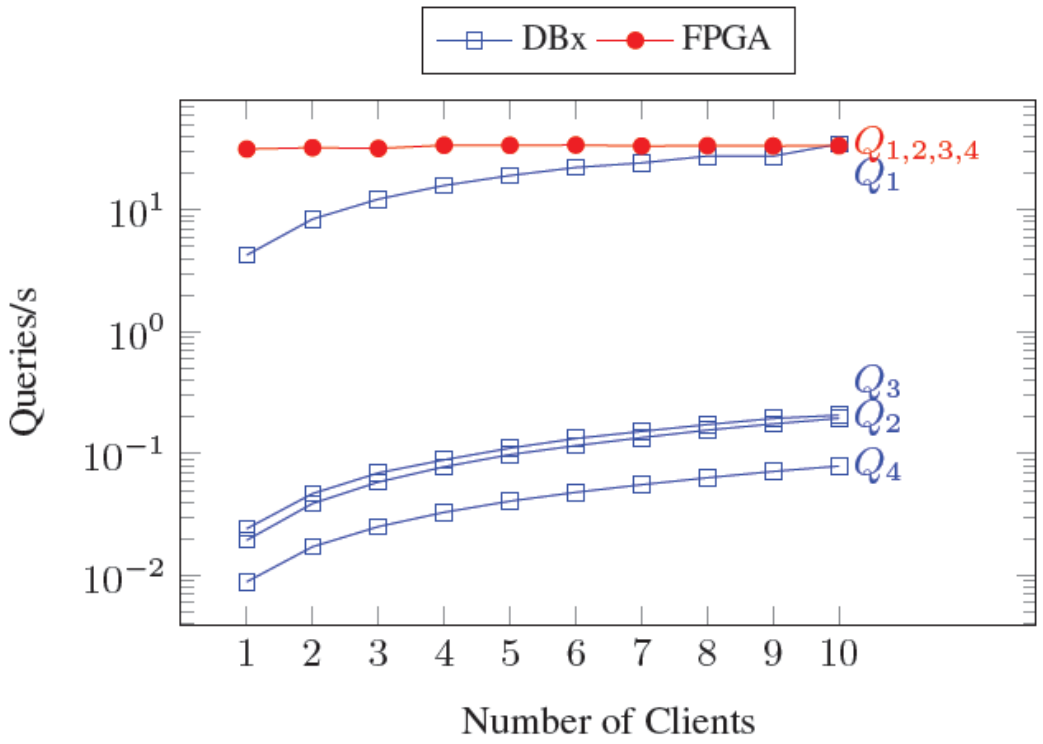




Accelerating real engines



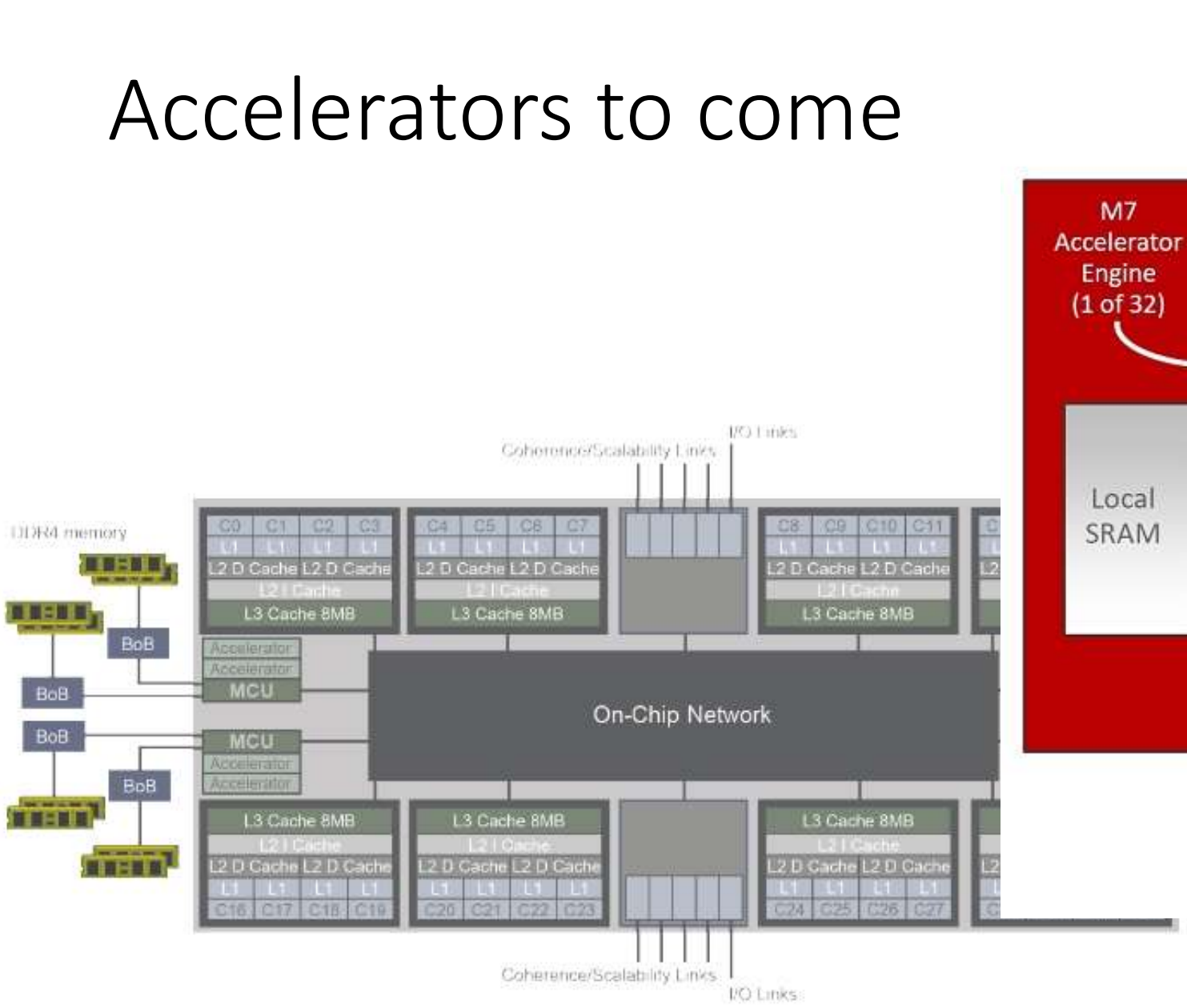
(a) Comparison to MonetDB

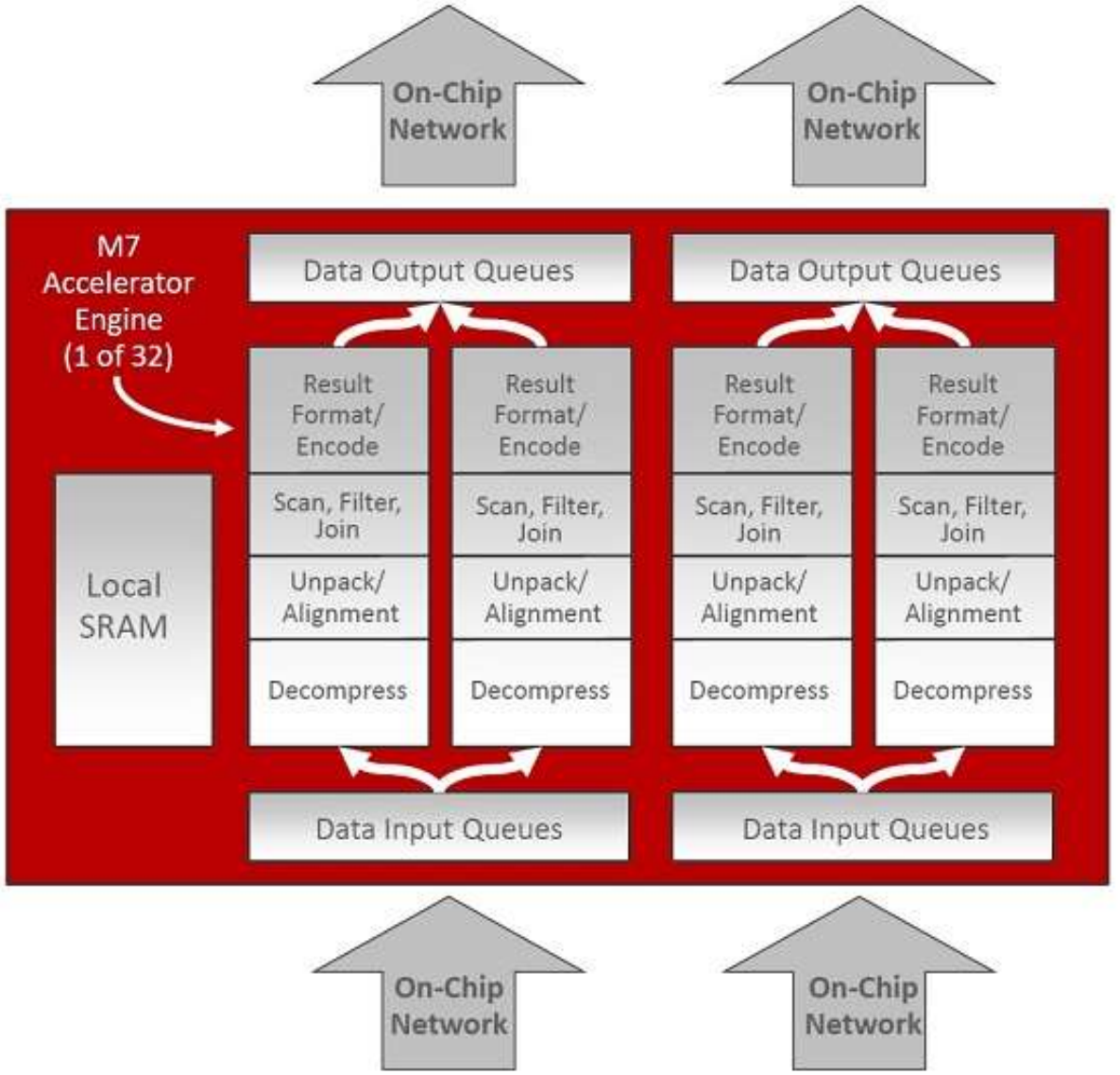


(b) Comparison to DBx

Throughput with increasing number of clients, number of records in table 2.5Mio.

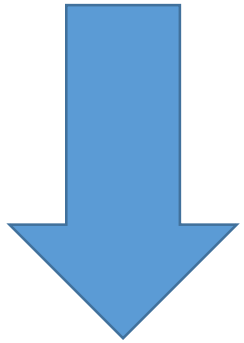
Accelerators to come



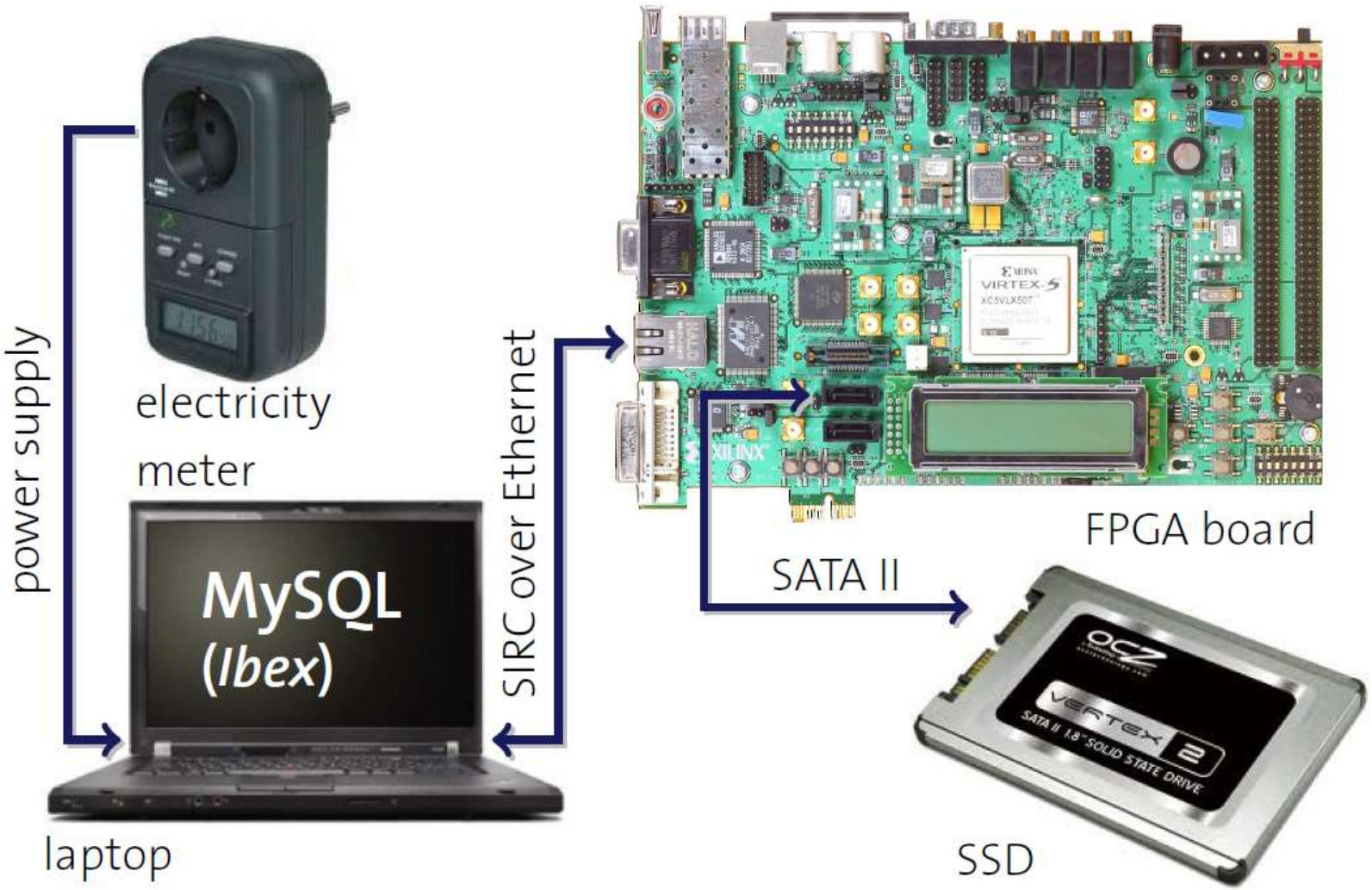


From Oracle M7 documentation

If the data moves, do it efficiently



Bumps in the wire(s)



(Woods, VLDB'14)

IBEX

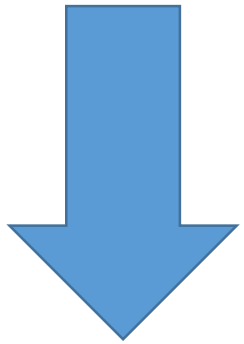
Sounds good?

The goal is to be able to do this at all levels:

- Smart storage
- On the network switch (SDN like)
- On the network card (smart NIC)
- On the PCI express bus
- On the memory bus (active memory)

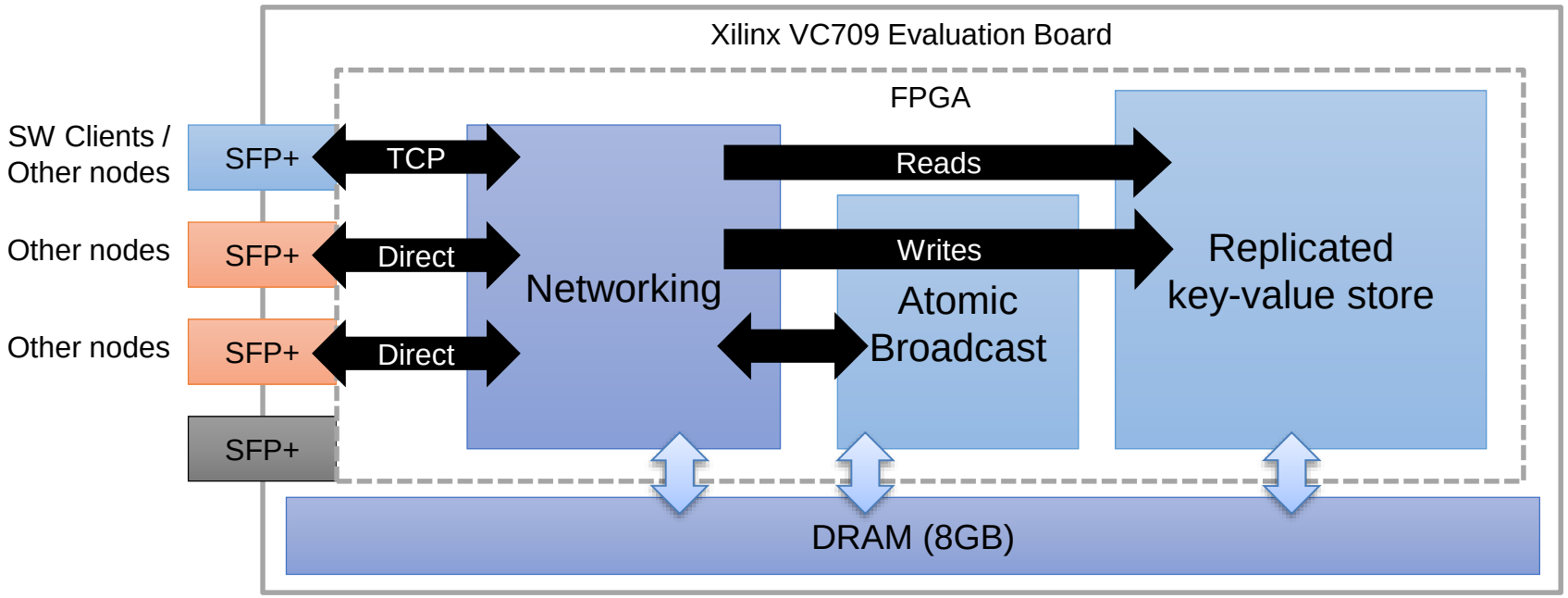
Every element in the system
(a node, a computer rack, a cluster)
will be a processing component

Disaggregated data center

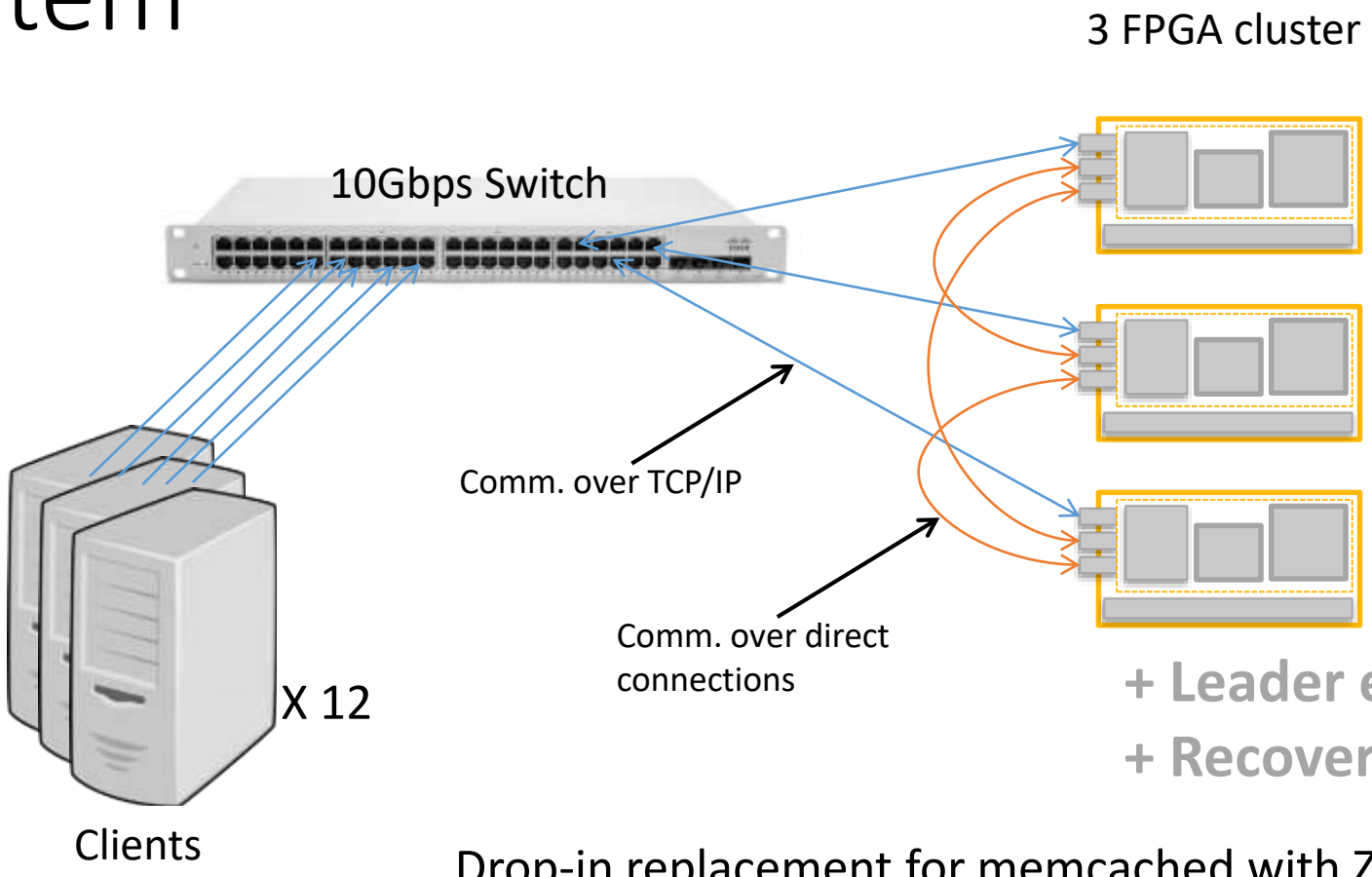


Near Data Computation

Consensus in a Box (Istvan et al, NSDI'16)



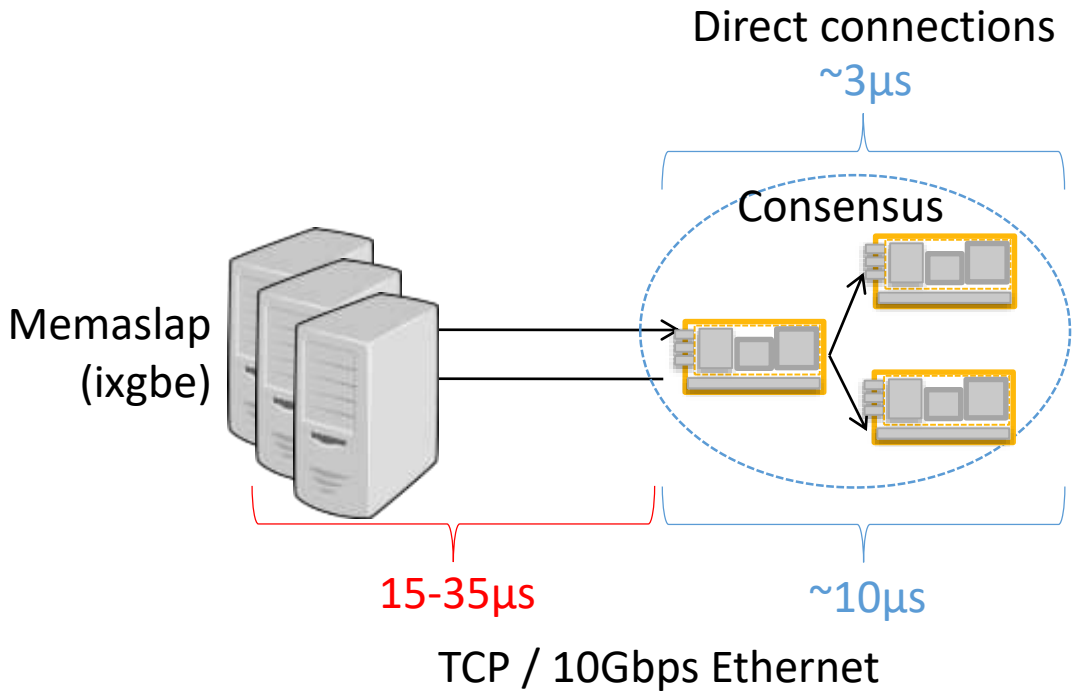
The system



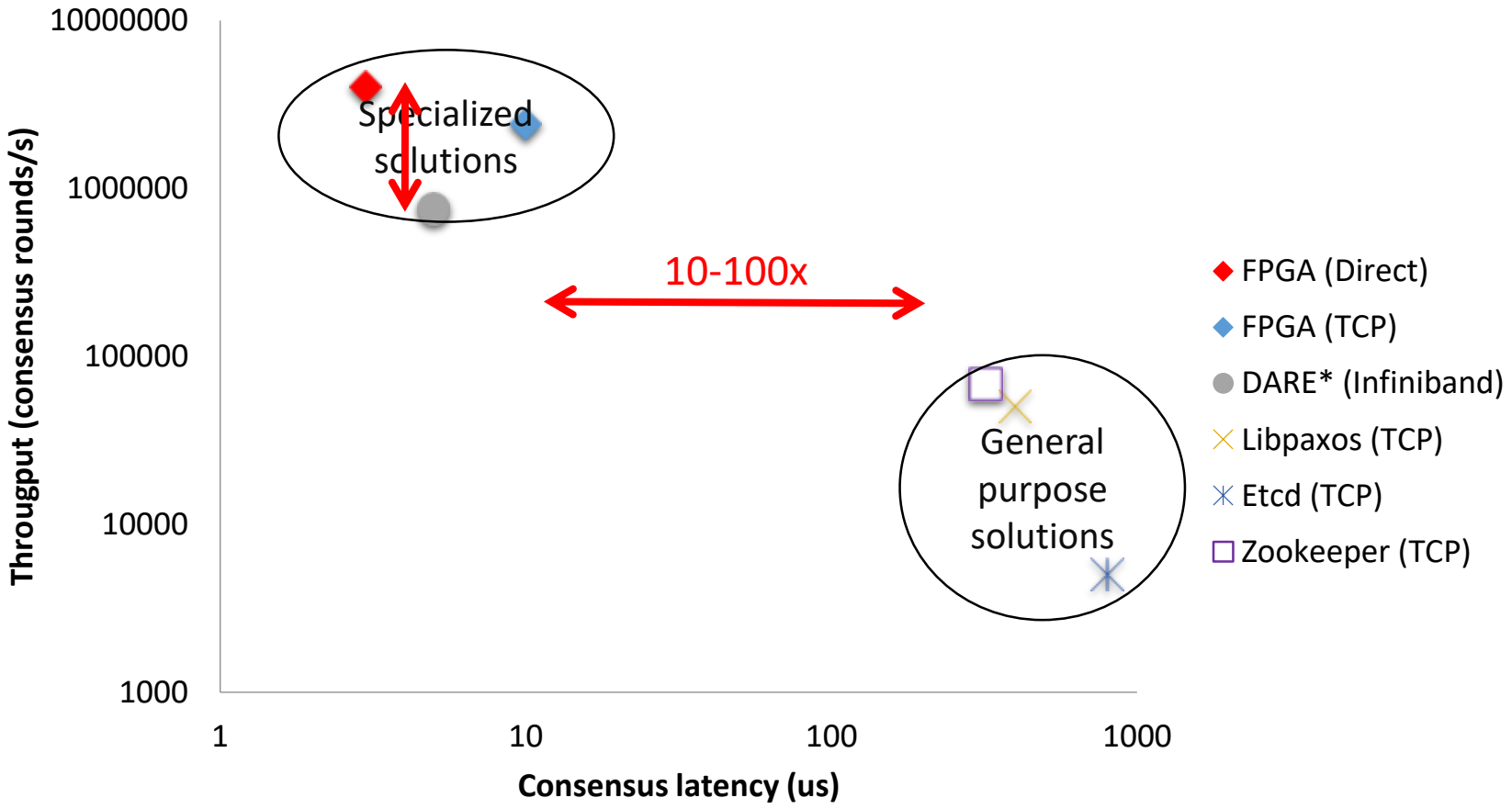
Drop-in replacement for memcached with Zookeeper's replication

- Standard tools for benchmarking (libmemcached)
  - Simulating 100s of clients

Latency of puts in a KVS



The benefit of specialization…



[1] Dragojevic et al. FaRM: Fast Remote Memory. In NSDI'14.
[2] Poke et al. DARE: High-Performance State Machine Replication on RDMA Networks. In HPDC'15.
*=We extrapolated from the 5 node setup for a 3 node setup.

This is the end …

There is a killer application (data science/big data)

There is a very fast evolution of the infrastructure for data processing (appliances, data centers)

Conventional processors and architectures are not good enough

FPGAs great tools to:

Explore parallelism

Explore new architectures

Explore Software Defined X/Y/Z

Prototype accelerators