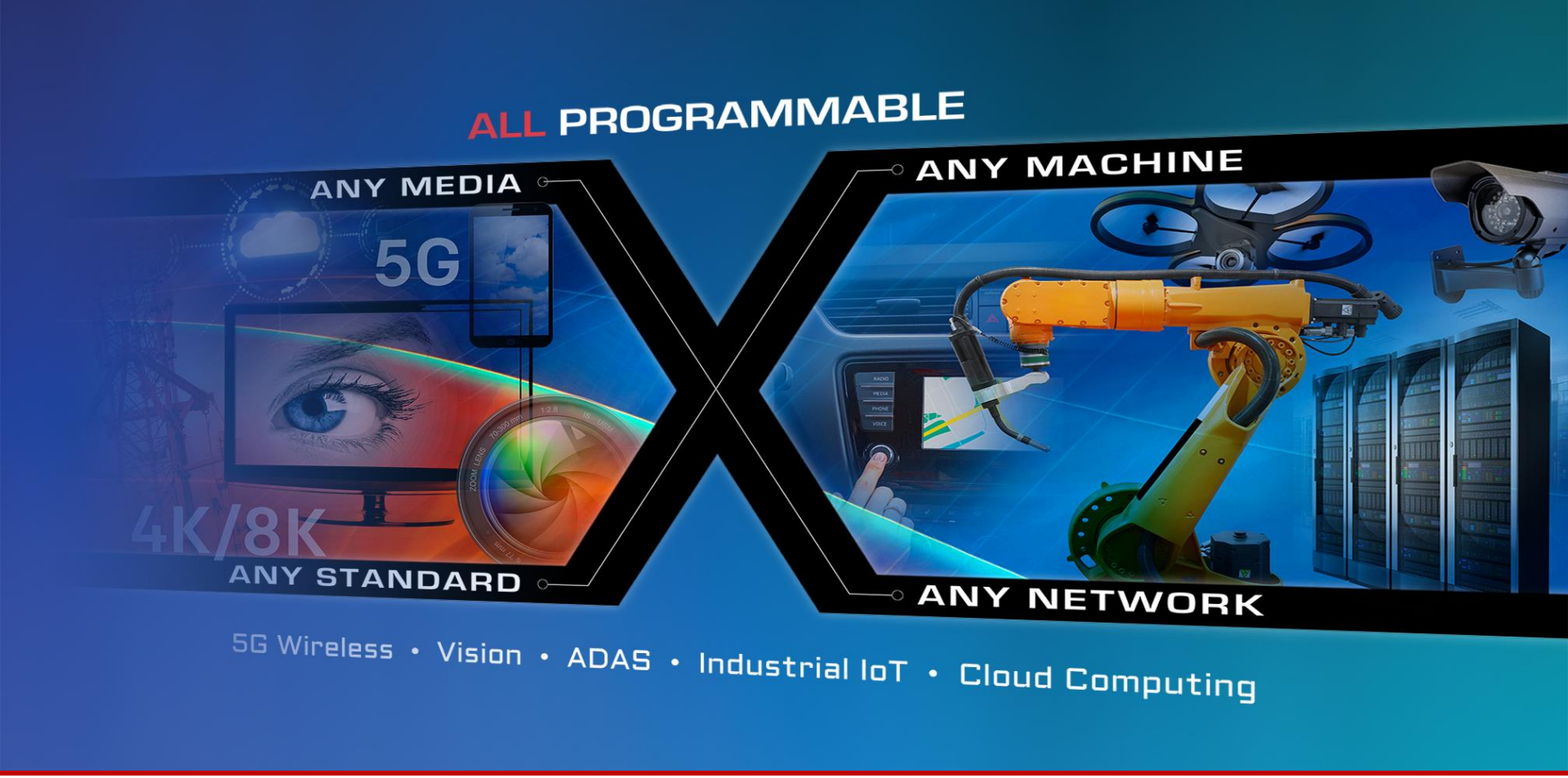


XILINX
ALL PROGRAMMABLE™

mle
missing link electronics

Heterogeneous Architectures for Implementation of High-Capacity Hyper-Converged Storage Devices

Endric Schubert (MLE),
Michaela Blott (Xilinx Research)

SDC, 2016

Content

Heterogeneous Architectures for Implementation of High-capacity Hyper-converged Storage Devices

- Who – Xilinx Research and Missing Link Electronics
- Why – High-capacity hyper-converged storage needs predictable scalability in performance, and programmability for flexibility
- What – A single-chip heterogeneous active storage solution for Terabit per second processing
- How – By combining modern FPGA design methodologies, including High-Level Synthesis, with IP cores for full acceleration of rich software

Xilinx Research and Missing Link Electronics

Xilinx – The All Programmable Company

XILINX
ALL PROGRAMMABLE™

XILINX - Founded 1984

- Headquarters
- Research and Development
- Sales and Support
- Manufacturing

$2.38B FY15 revenue

>55% market segment share

3,500+ employees worldwide

20,000 customers worldwide

3,500+ patents

60 industry firsts

Xilinx is Diversified Across Multiple Markets

Mars Rover

MRI scanners

3D televisions

Machine Learning

ADAS

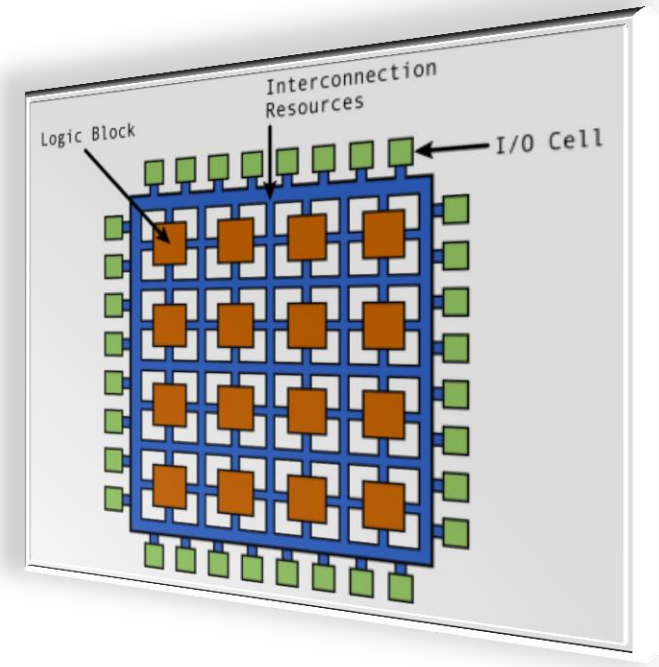
AR

What are FPGAs

- **A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by the customer or designer after manufacturing—hence " field-programmable"**

  (Source: Wikipedia)

- **In their simplest form FPGAs contain:**
  - Configurable Logic Blocks (AND, OR..)
  - Configurable interconnect
  - I/O Interfaces
- **Today:**
  - 3.4M
  - 6.3Tbps IO
  - + DSPs, ARM, 2.5D…

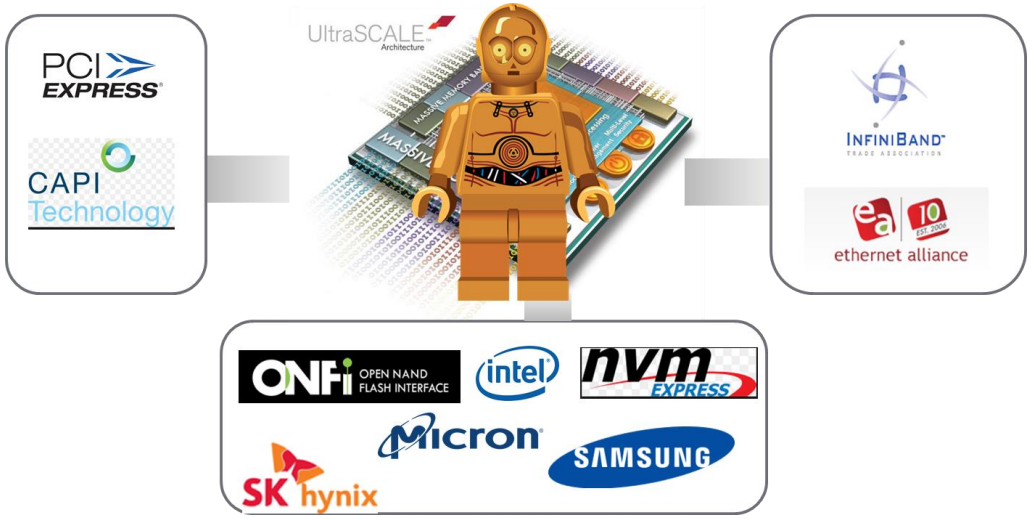


Custom-tailored hardware accelerator for your application, while providing programmability

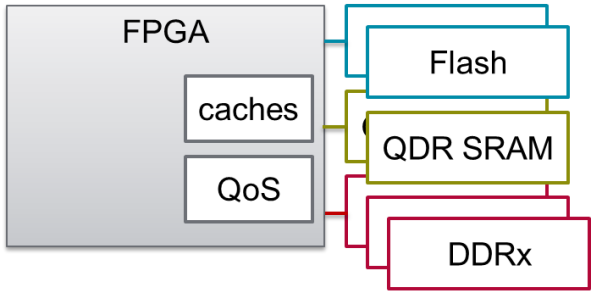
Customizable Interfaces & Memory Architectures

- **Flexibility to support many types of interfaces**

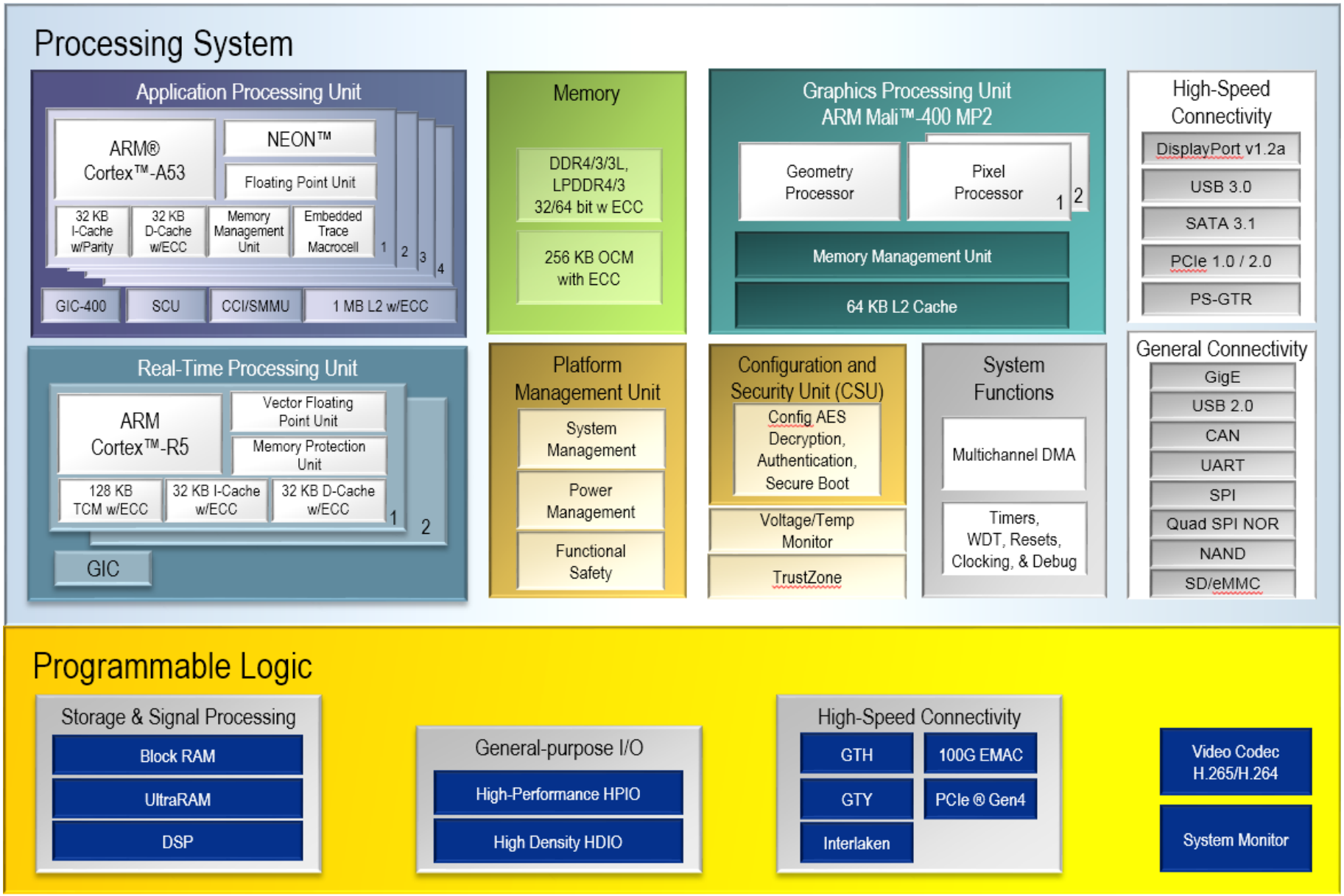
- **Single-chip solution**
  - BOM cost reduction
  - PCB footprint reduction
  - Power/energy reduction
  - Performance



- **Customized memory systems**
  - Hybrid memory
  - Application specific cache hierarchies



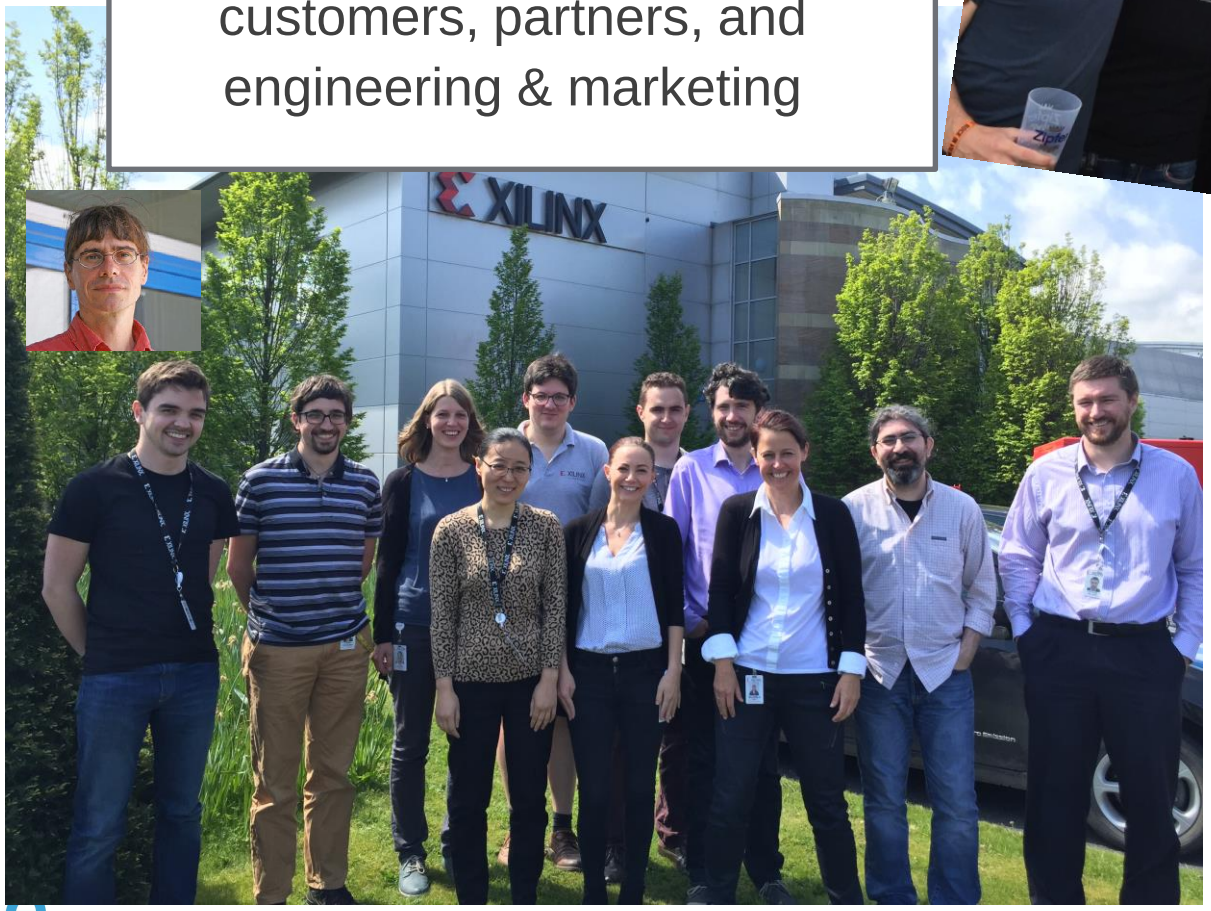
Heterogeneous Multicore with Programmable Logic



Xilinx Research - Ireland

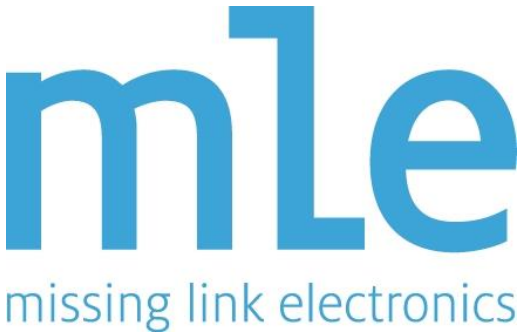
Applications & Architectures

Through application-driven technology development with customers, partners, and engineering & marketing



Missing Link Electronics

Xilinx Ecosystem Partner



Vision: The convergence of software and off-the-shelf programmable logic opens-up more economic system realizations with predictable scalability!

Mission: To de-risk the adoption of heterogeneous compute technology by providing pre-validated IP and expert design services.

Certified Xilinx Alliance Partner since 2011, Preferred Xilinx PetaLinux Design Service Partner since 2013.

Missing Link Electronics Products & Services

TCP/IP & UDP/IP Network Protocol Accelerators for FPGA (patent pending).

Patented Mixed Signal systems solutions with integrated Delta-Sigma converters in FPGA logic.

SATA Storage Extension for Xilinx Zynq All-Programmable Systems-on-Chip.

MLE markets and supports the Xilinx XPS USB 2.0 EHCI Host Controller IP core.

Tools for architecture analysis and optimization and RTL and C/C++ based FPGA design.

A team of FPGA and Linux engineers to support our customer's technology projects in the USA and Europe.

Motivation

Technology Forces in Storage

- **Software significantly impacts latency and energy efficiency in systems with nonvolatile memory**
- **However, software-defined flexibility is necessary to fully utilize novel storage technologies**
- **Hyper-capacity hyper-converged storage systems need more performance, but within cost and energy envelopes**

Software Considered Harmful

Nonvolatile memory (NVM) shifts the balance between hardware and software costs in storage systems, and thereby redefines software's role. In disk-based systems, the energy that the storage stack consumes running on a power-hungry CPU pales in comparison to a disk's energy requirements. As a result, it is possible to improve performance and save energy by adding software to a disk-based system.

But host-side software is slow and energy-hungry compared to NVM, and the more software the host executes to manage I/O requests, the slower those requests will be. This means that using existing storage stacks to manage NVM-based storage is a recipe for disappointment and inefficiency, and it will be difficult if not impossible to improve performance and efficiency by adding software to the system. Conversely, reducing interactions with software components and refactoring them to reduce their costs is an effective way to improve performance and efficiency.

Measurements of software and hardware costs in contemporary storage systems illustrate software's shifting role. In the off-the-shelf Linux storage stack, a single 512-byte I/O operation requires about 19 µs of processor time on a 2.27-GHz Nehalem processor. A single active Nahalem core consumes around 28 W, or 532 µJ, per I/O operation.

These software costs are roughly constant regardless of the underlying storage technology, but the relative cost of software changes completely. Figure A illustrates this shift. For disks, software accounts for just 0.27 percent of I/O operational latency, but for the ioDrive (a high-end flash-based solid-state drive) and Moneta (our prototype SSD for next-generation memory), it accounts for 22 percent and 70 percent, respectively. The shift is almost as dramatic for energy: 0.42 percent of the disk's I/O operational energy goes to software versus 73 percent and 95 percent for ioDrive and Moneta.

The shifting ratio of software to hardware costs has profound effects on how designers should approach crafting a storage system. As an example, consider the decision to add the logical volume manager (LVM) to a Linux storage stack to make expanding system capacity easier. Table A shows the comparison between a disk-based system, ioDrive, and the NVM-based Moneta SSD. Adding this layer increases software latency by 2 µs and energy consumption by 56 µJ per I/O operation. In the disk-based system, these increases are negligible, but they are much higher for the ioDrive, and highest of all for Moneta.

To minimize the harm that operating system software causes, system designers need to reengineer storage systems to minimize software's role. In some cases, this will require extensions or modifications to storage hardware, but often it means applying well-known design principles to refactor existing systems.

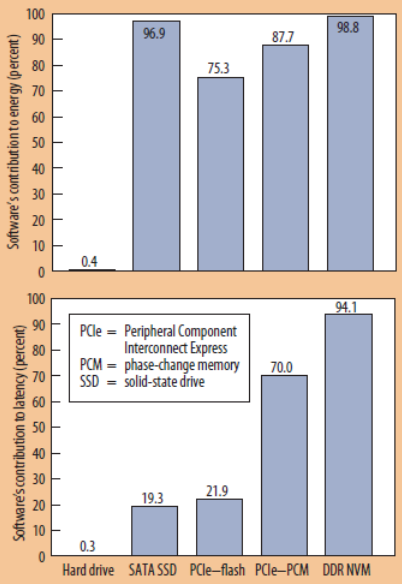


Figure A. Software's impact on latency and energy. Without significantly reengineering the storage software stack, software latency and energy per I/O operation will quickly dominate I/O costs.

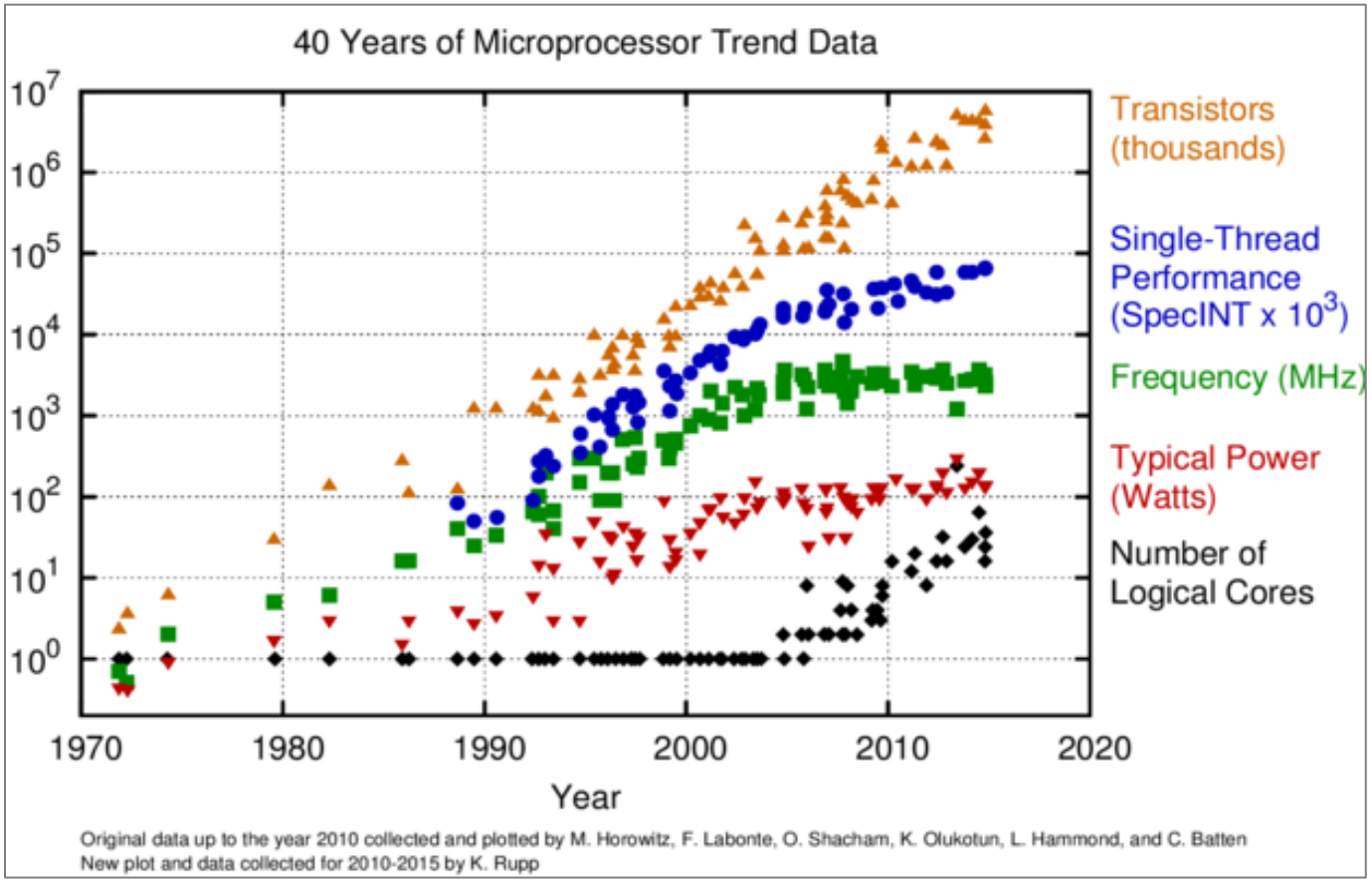
Table A. Impact of adding a logical volume manager (LVM) to three storage systems.

Storage system	Software latency increase (percent)	Energy consumption increase (percent)
Disk-based	0.03	0.04
ioDrive (flash-based SSD)	4.30	15.50
Moneta (NVM-based SSD)	10.70	18.70

Source: Steven Swanson and Adrian M. Caulfield, UCSD
IEEE Computer, August 2013

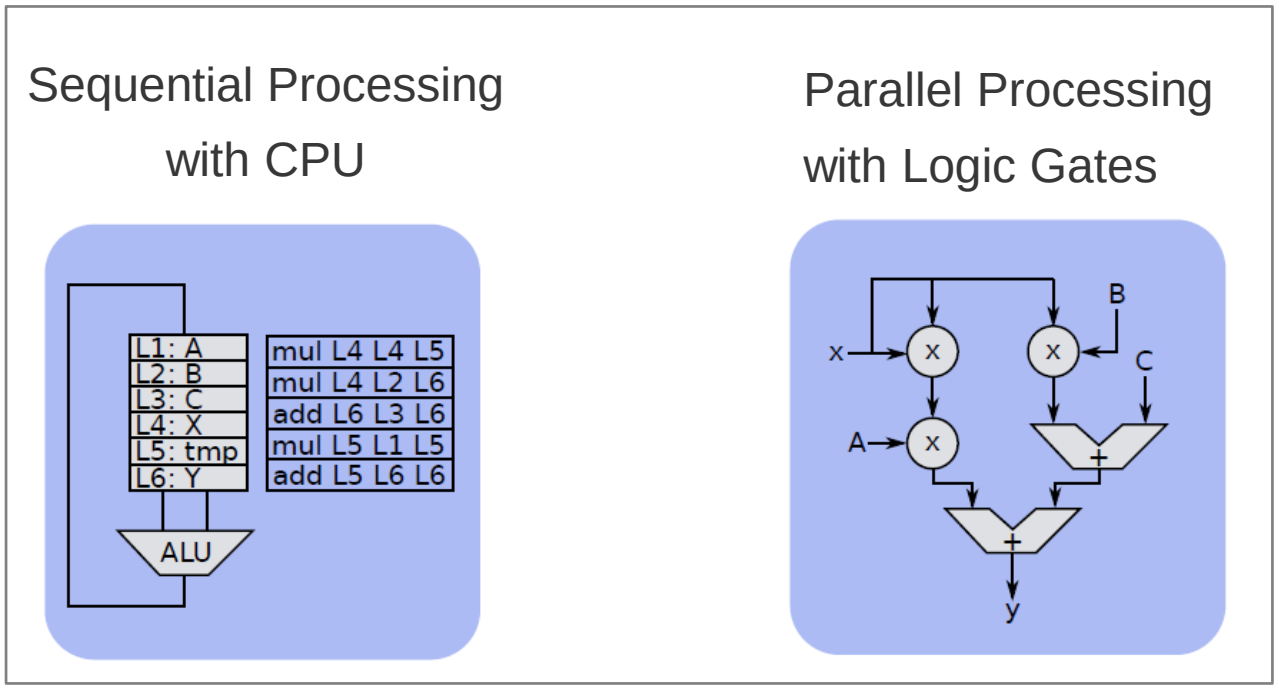
The Von Neumann Bottleneck [J. Backus, 1977]

- **CPU system performance scalability is limited**



New Compute Architectures are needed

Spatial vs Temporal Computing

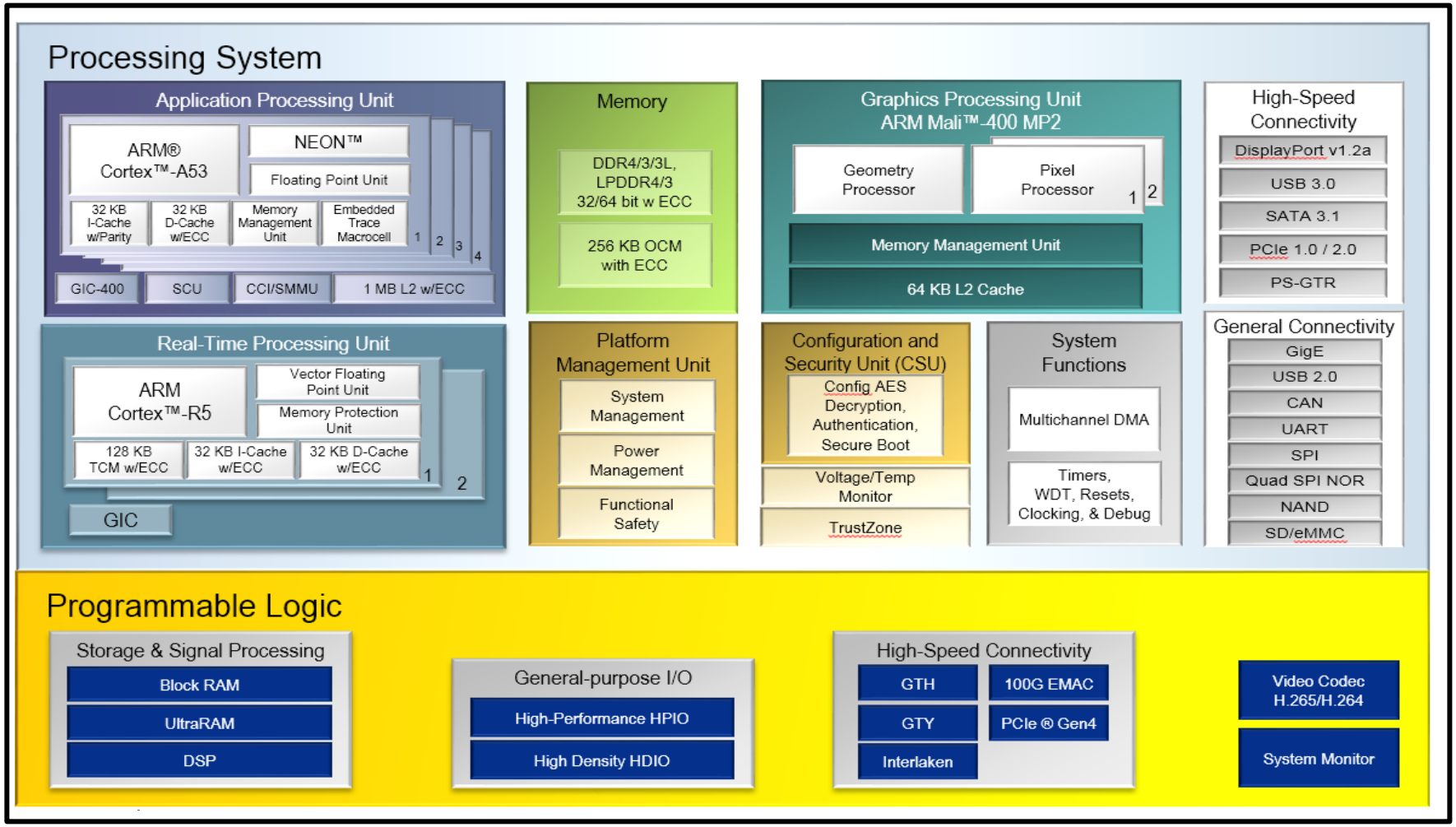


Source: Dr. Andre DeHon, Upenn: "Spatial vs. Temporal Computing"

- **CPU system performance scalability is limited**
- **Spatial computing offers further scaling opportunity**

New Compute Architectures are needed to take advantage of this

Architectural Choices for Storage Devices



Source: T.Noll, RWTH Aachen

Terabit Processing with Single-Chip Solutions

FPGA Comparison Table

	Kintex-7	Virtex-7	Kintex UltraScale	Kintex UltraScale+	Virtex UltraScale	Virtex UltraScale+
Logic Cells (K)	478	1,955	1,161	915	4,433	2,863
UltraRAM (Mb)	-	-	-	36.0	-	432.0
Block RAM (Mb)	34	68	76	34.5	132.9	94.5
DSP Slices	1,920	3,600	5,520	3,528	2,880	11,904
DSP Performance (symmetric FIR)	2,845 GMACs	5,335 GMACs	8,180 GMACs	6,287 GMACs	4,268 GMACs	21,213 GMACs
Transceiver Count	32	96	64	76		
Maximum Transceiver Speed (Gb/s)	12.5	28.05	16.3	32.75	30.5	32.75
Total Transceiver Bandwidth (full duplex) (Gb/s)	800	2,784	2,086	2,478	5,886	8,384
Memory Interface (DDR3)	1,866	1,866	2,133	2,133	2,133	2,133
Memory Interface (DDR4)	-	-	2,400	2,667	2,400	2,667
PCI Express®	x8 Gen2	x8 Gen3	x8 Gen3	x8 Gen 4 x16 Gen 3	x8 Gen3	x8 Gen 4 x16 Gen 3

800Gbps – 8.4Tbps

Source: http://www.xilinx.com/products/silicon-devices/fpga.html

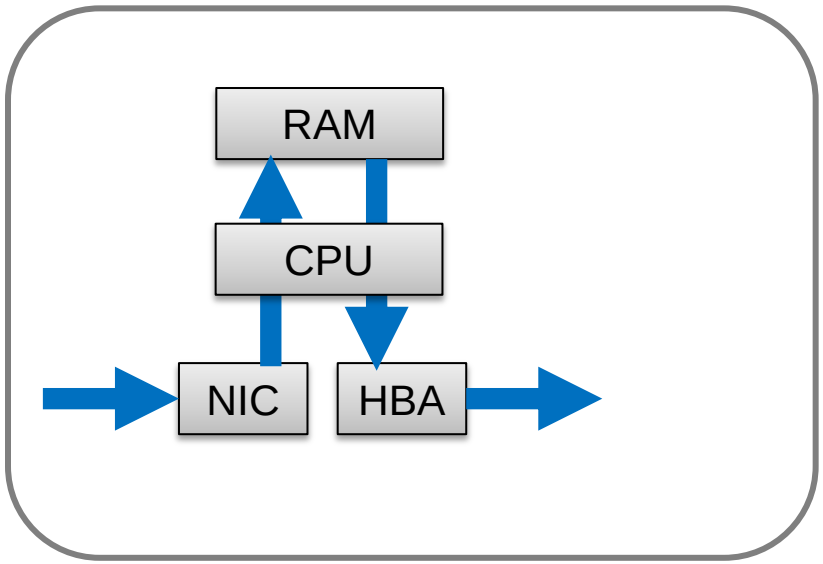
Hyper-Converge into one single device

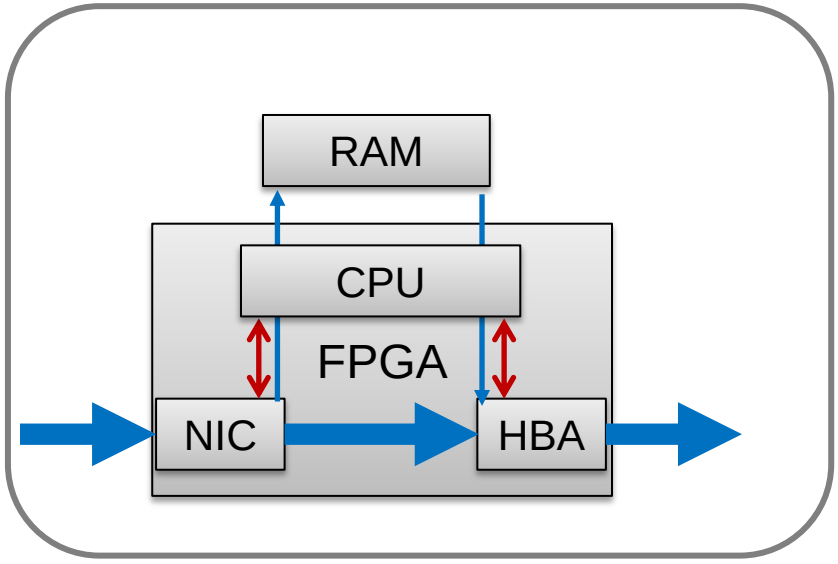
Tight coupling of compute + storage + networking

Current architecture limits maximum performance to total DMA bandwidth through many hops

Storage and compute directly integrated into the network

Dataflow processing for higher bandwidth via FPGA-based inline processing

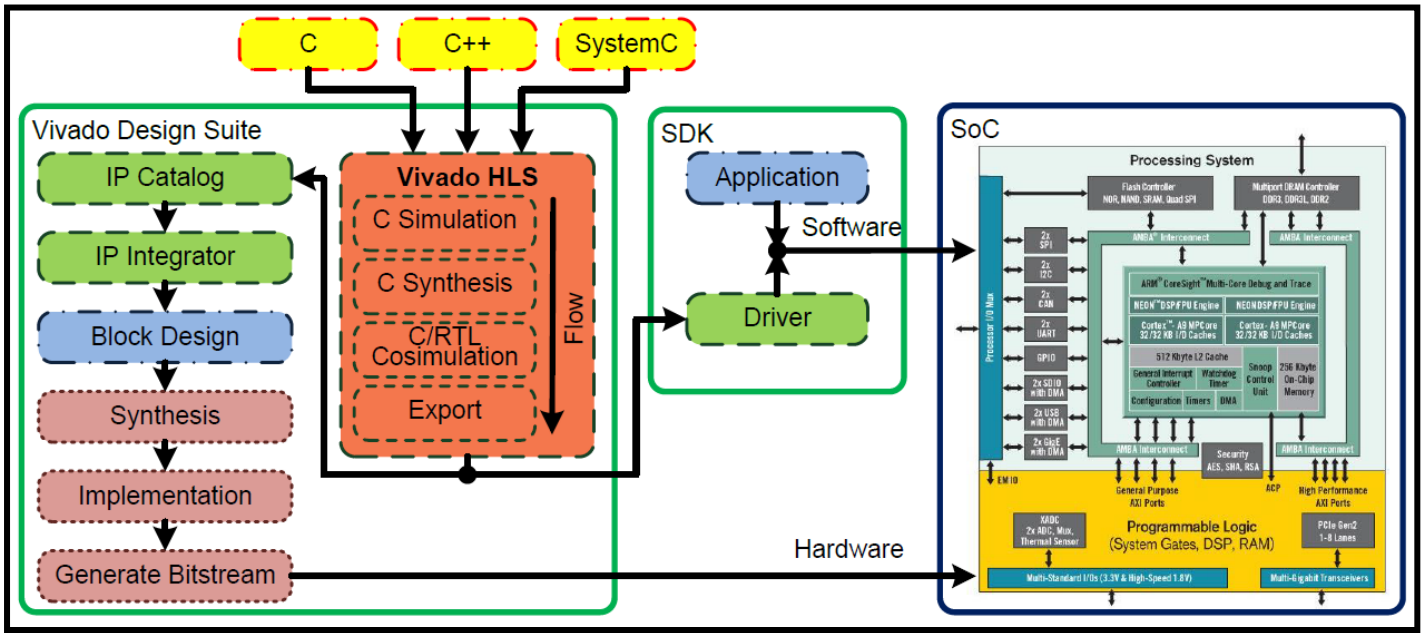




Source: Lim et al: Thin servers with smart pipes: designing {SoC} accelerators for memcached; ISCA 2013

Design Flow Options

- **Semi-automated block-based RTL synthesis design flow combined with modern C/C++ High-Level-Synthesis for application specific functionality**



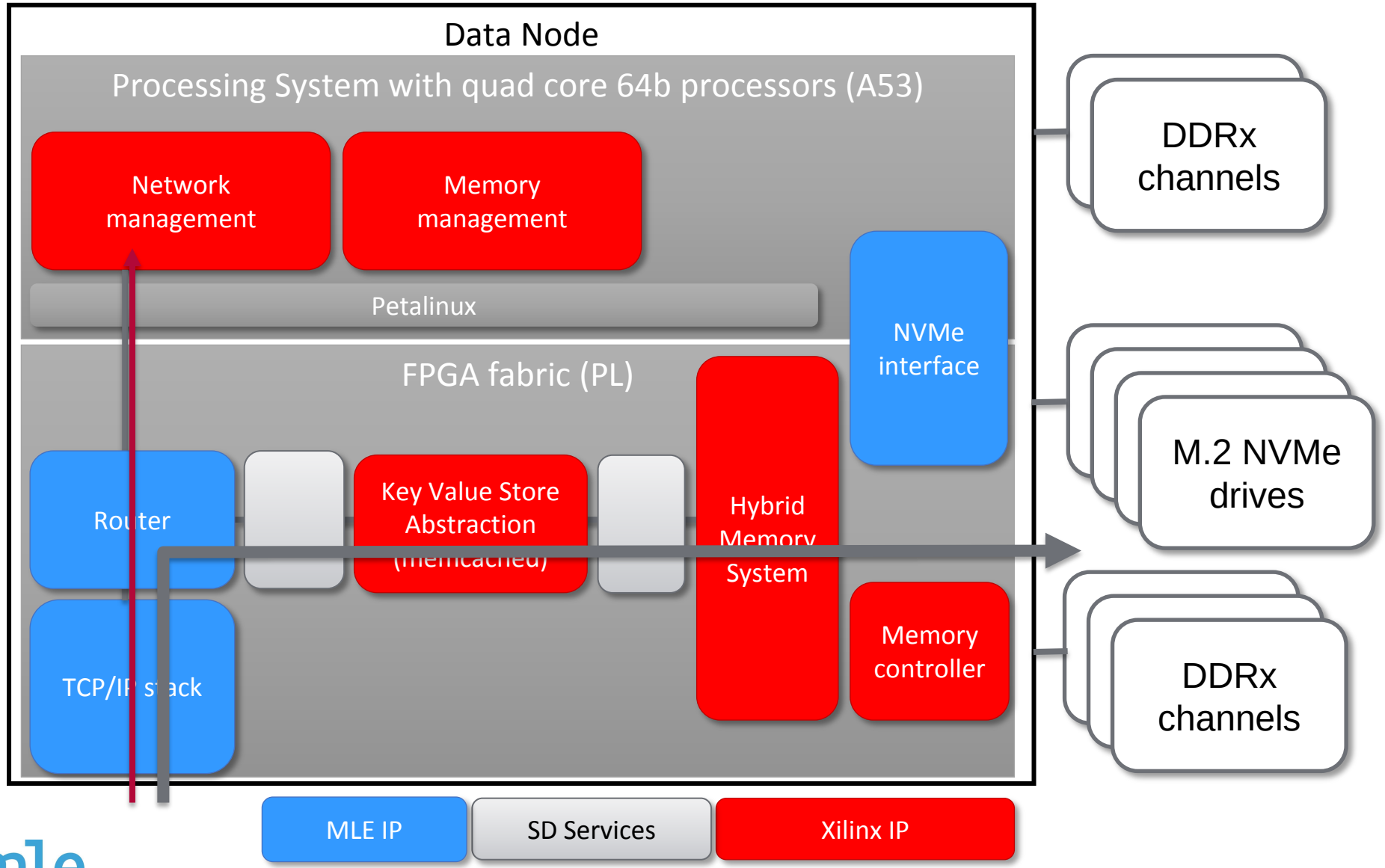
- **Fully integrated software environment are emerging to abstract data movement (SDSoC and SDAccel)**

Architectural Concepts

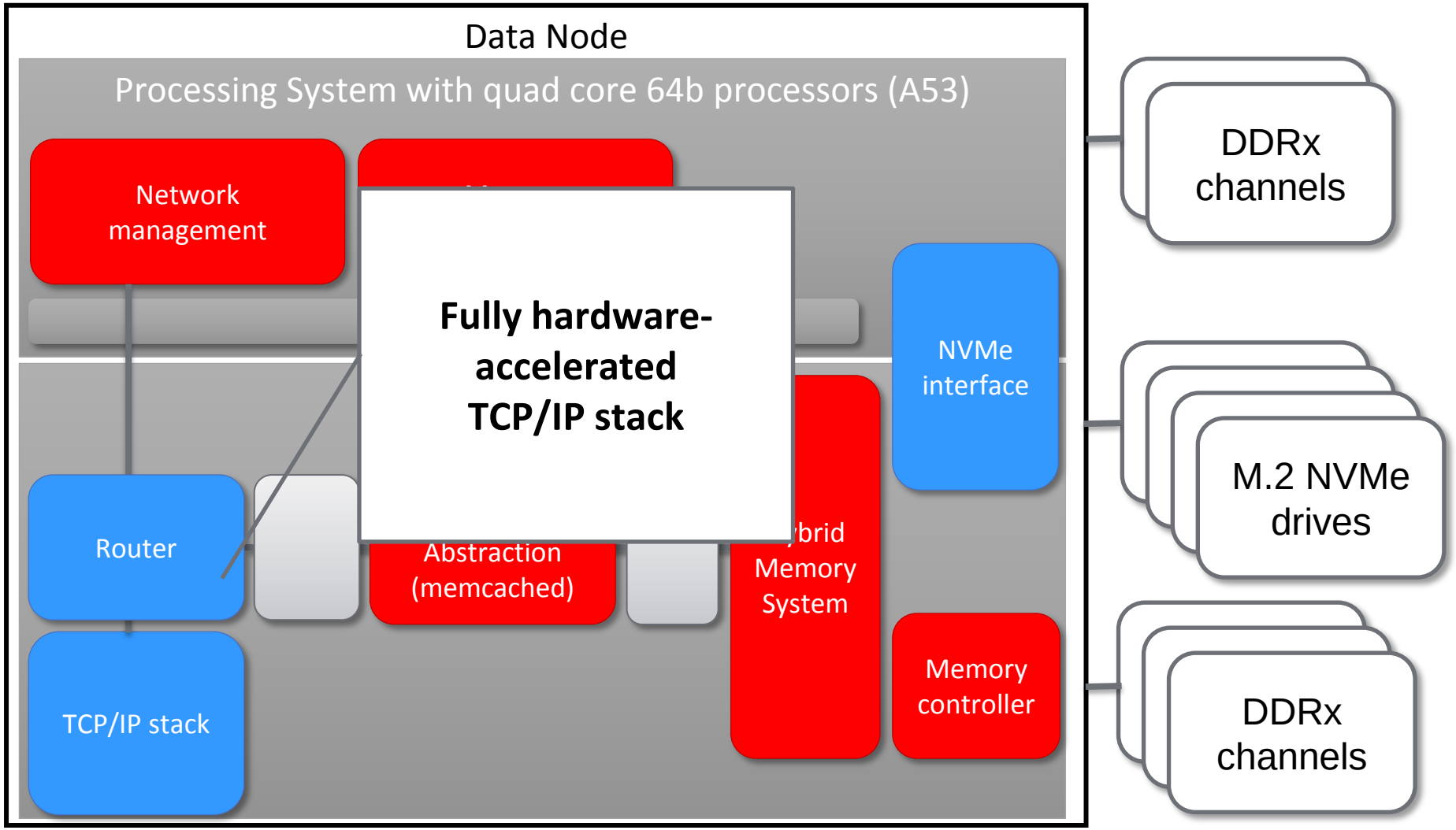
Key Concepts for an Extensible Architecture for Storage Devices

- **Heterogeneous compute device as a single-chip solution**
- **Direct network interface with full accelerator for protocols**
- **Performance scaling with dataflow architectures**
- **Scaling capacity and cost with a Hybrid Storage subsystem**
- **Software-defined services**

Concept 1: Single-Cip Solution for Storage



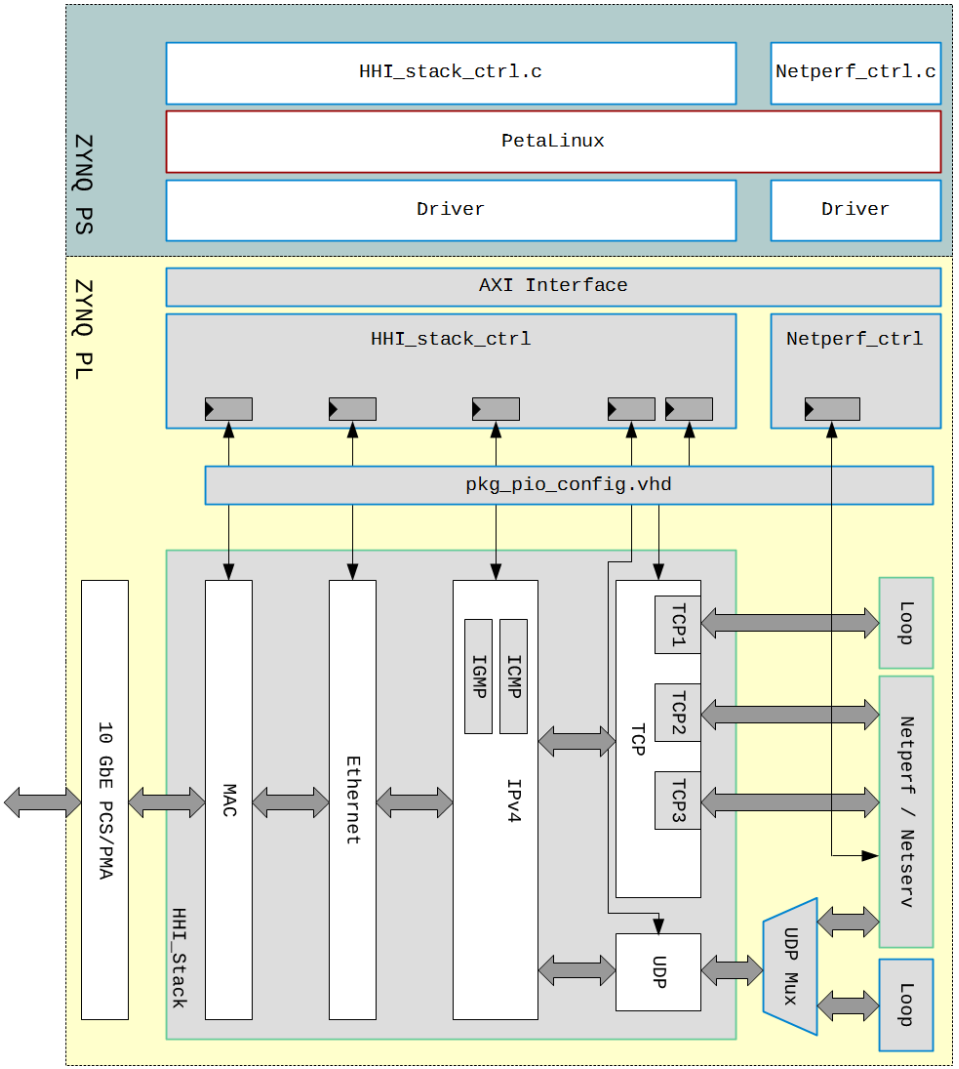
Concept 2: Hardware Accelerated Network Stack



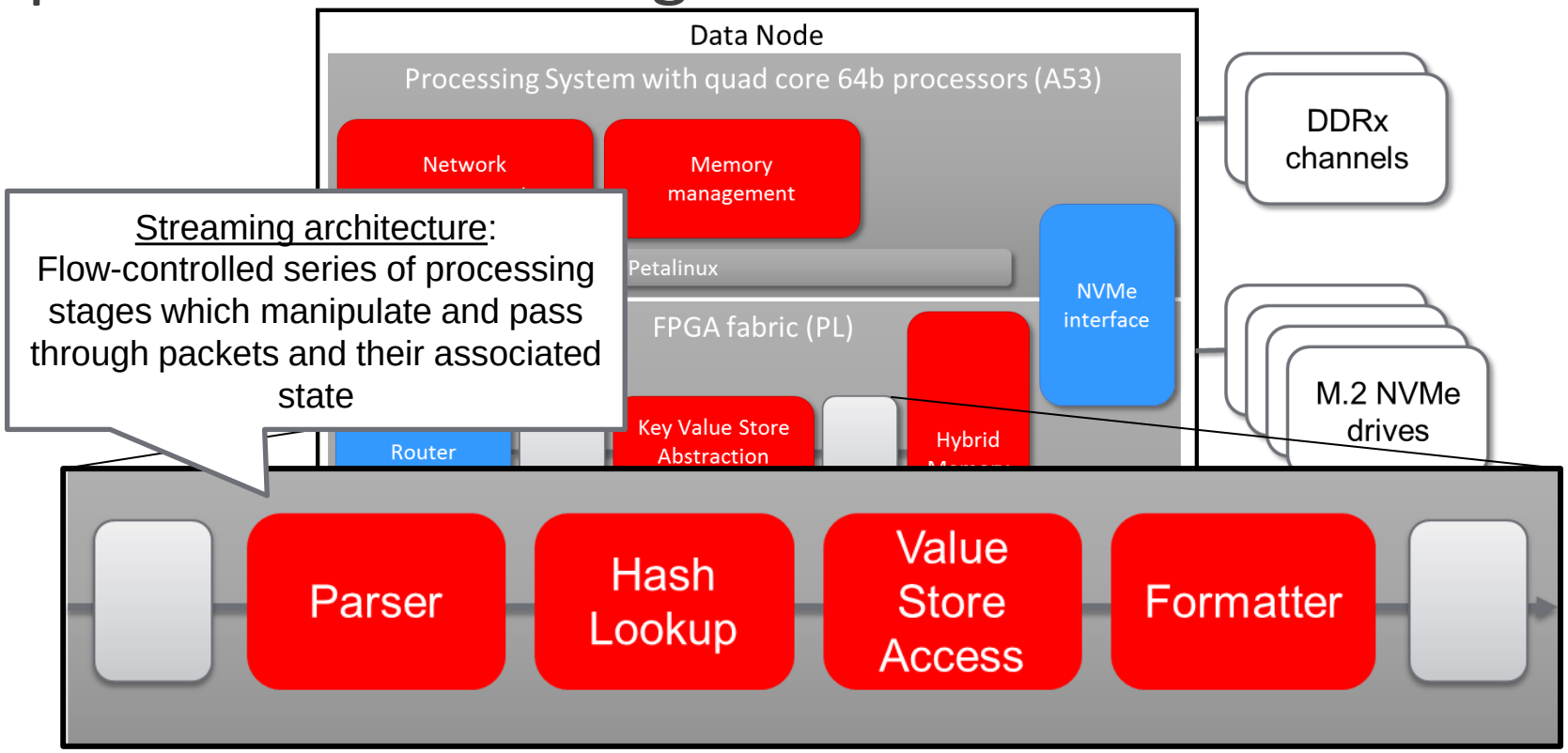
Concept 2: Hardware Accelerated Network Stack

- 128bit datapaths for Rx and TX
- Scales to 40 GigE (@250 MHz)
- No CPU needed – although embedded CPUs can be utilized for administrative or Layer 7 processing
- Extensible via HDL or via C/C++ using High-Level Synthesis
- Technology from:

Fraunhofer
Heinrich Hertz Institute



Concept 3: Dataflow architectures for performance scaling

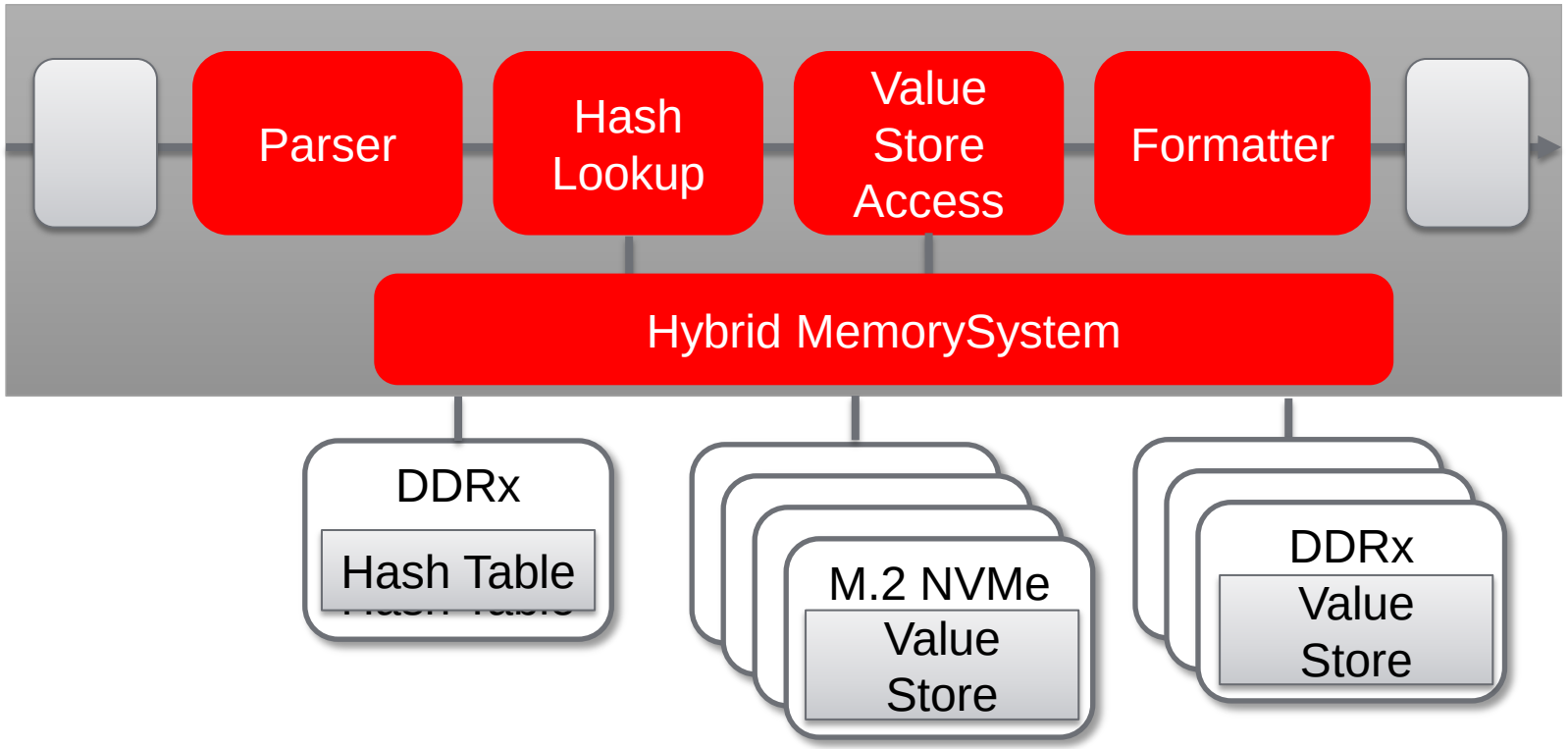


- **Now: 10 Gbps demonstrated with a 64b data path @ 156MHz using 20% of FPGA**
- **Next: 100 Gbps can be achieved by using a 512b @ 200MHz pipeline for example**

Source: Blott et al: Achieving 10Gbps line-rate key-value stores with FPGAs; HotCloud 2013

Concept 4: Scaling Capacity

- **SSDs combined with DDRx channels can be used to build high capacity & high performance object stores**
- **Concepts and early prototype to scale to 40TB & 80Gbps key value stores**



Source: HotStorage 2015, Scaling out to a Single-Node 80Gbps Memcached Server with 40Terabytes of Memory

Concept 4:Object distribution on the basis of size

Value Size (B)	128	256	512	768	1K	4K	8K	32K	1M
Facebook	0.55	0.075	0.275	0	0	0	0	0	0.1
Twitter	0	0	0	0.1	0.85	0.05	0	0	0
Wiki	0	0	0.2	0.1	0.4	0.29	0.008	0.001	0.001
Flickr	0	0	0	0	0	0.9	0.05	0.03	0.02

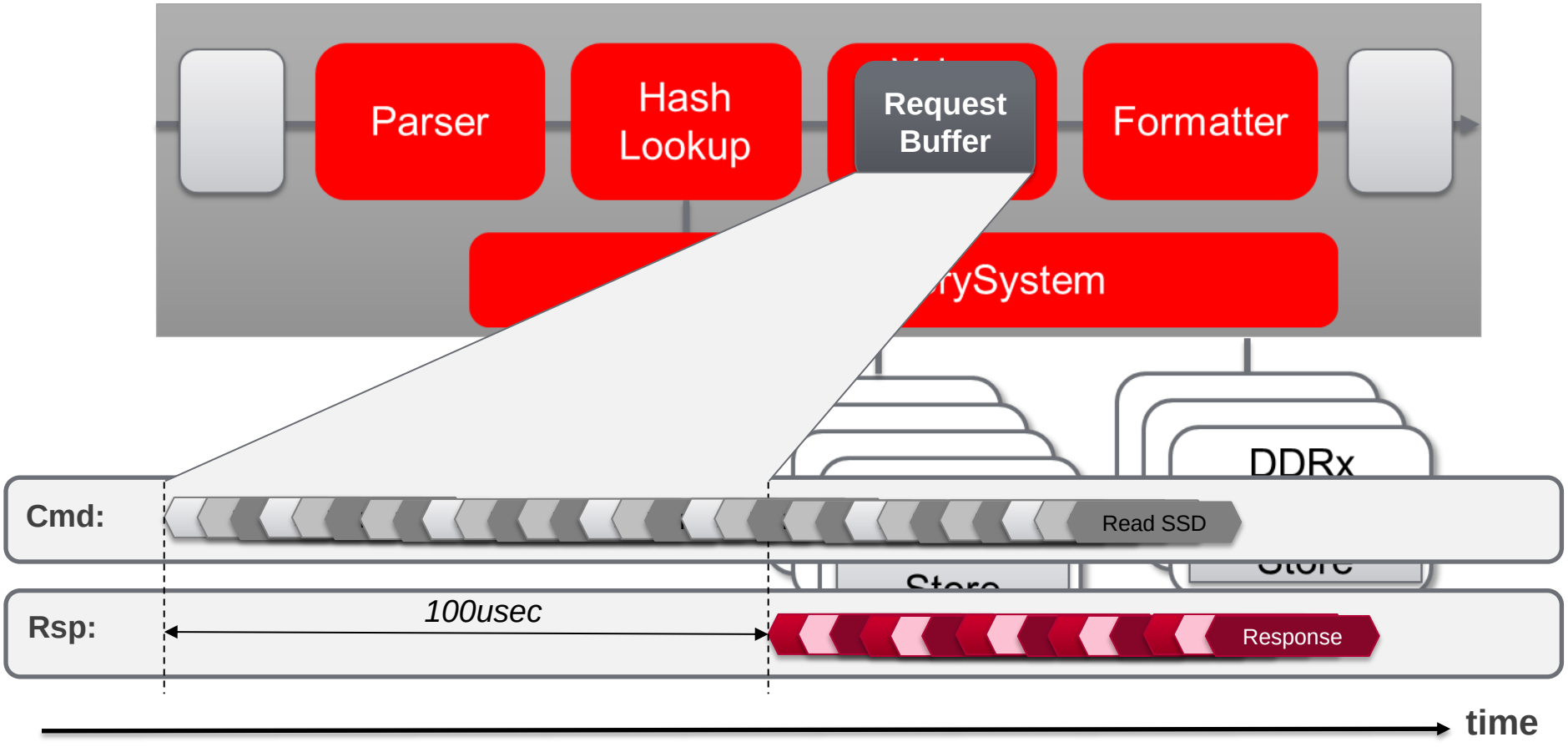
Stored in DRAM

Stored in Flash

- **Advantages:**
  - Larger objects require larger storage
  - Larger granular access to flash suits page-size access granularity of flash
- **Concerns:**
  - Large access latency on flash
  - Variations in access bandwidth and latency between DRAM and flash

Source: [3] Atikoglu et al: Workload analysis of a large-scale key-value store; SIGMETRICS 2012

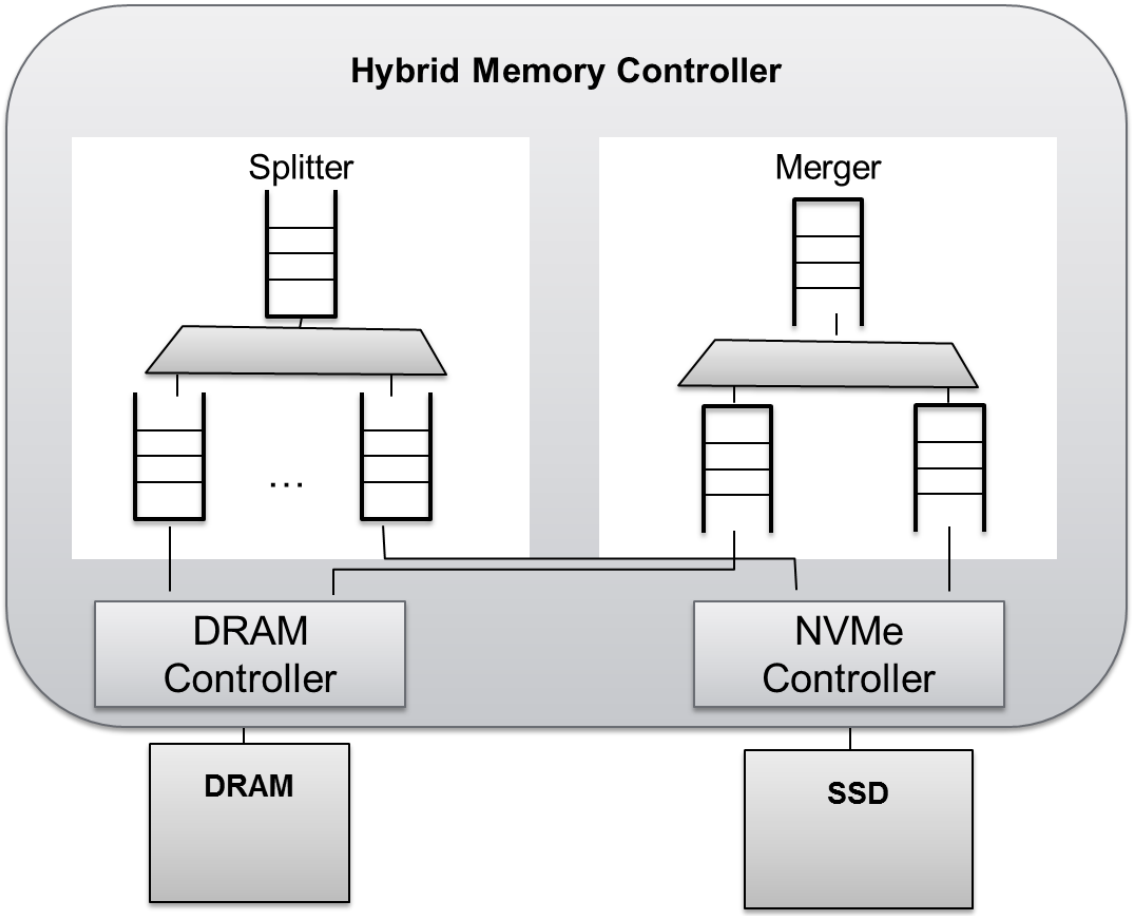
Concept 4: Handling High Latency Accesses without Sacrificing Throughput



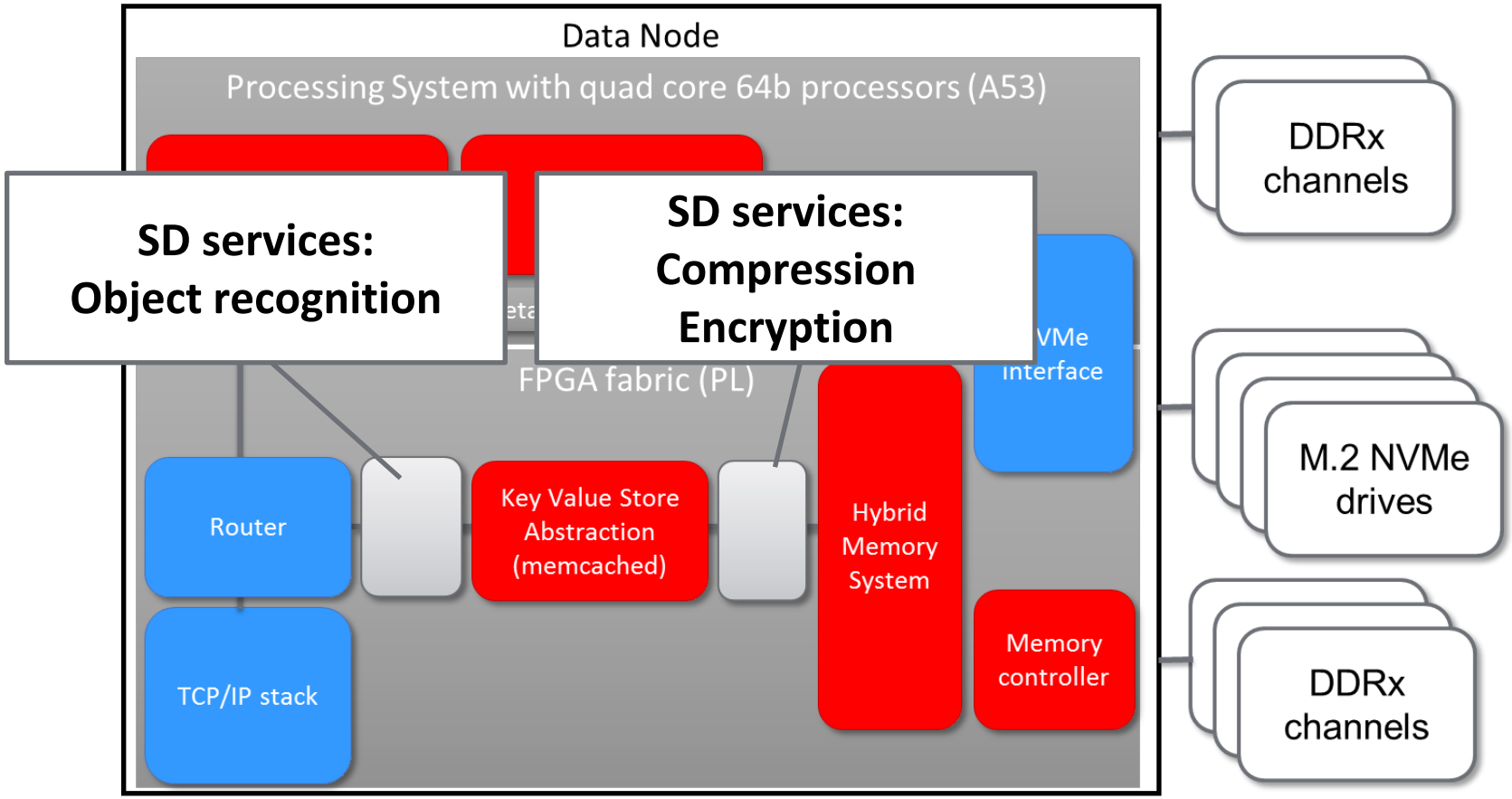
- **Dataflow architectures: no limit to number of outstanding requests**
- **Flash can be serviced at maximum speed**

Concept 4: Custom memory controllers

with out of order processing



Concept 5: SD Services

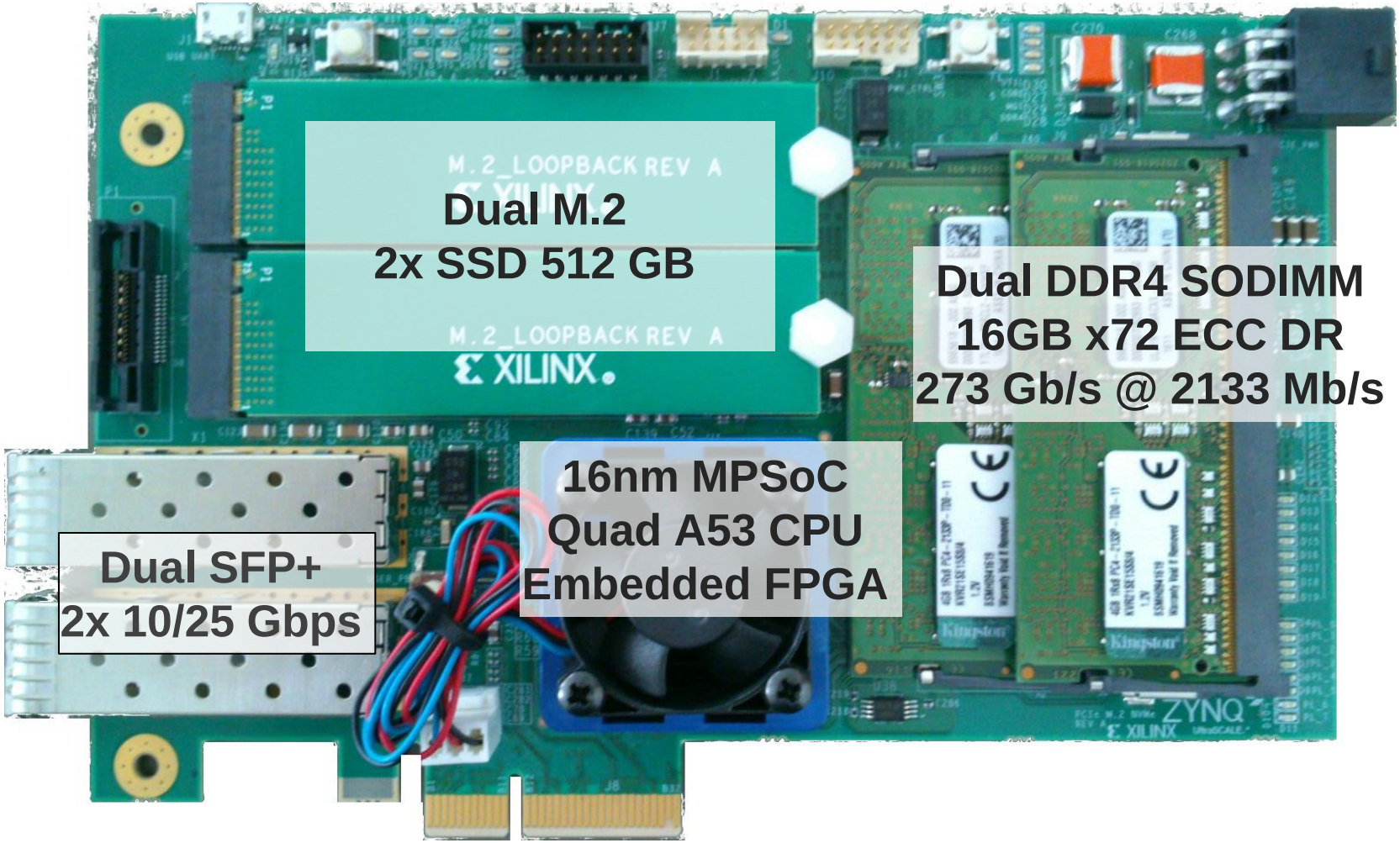


- **Spatial computing of additional services at no performance cost until resource limitations are reached**

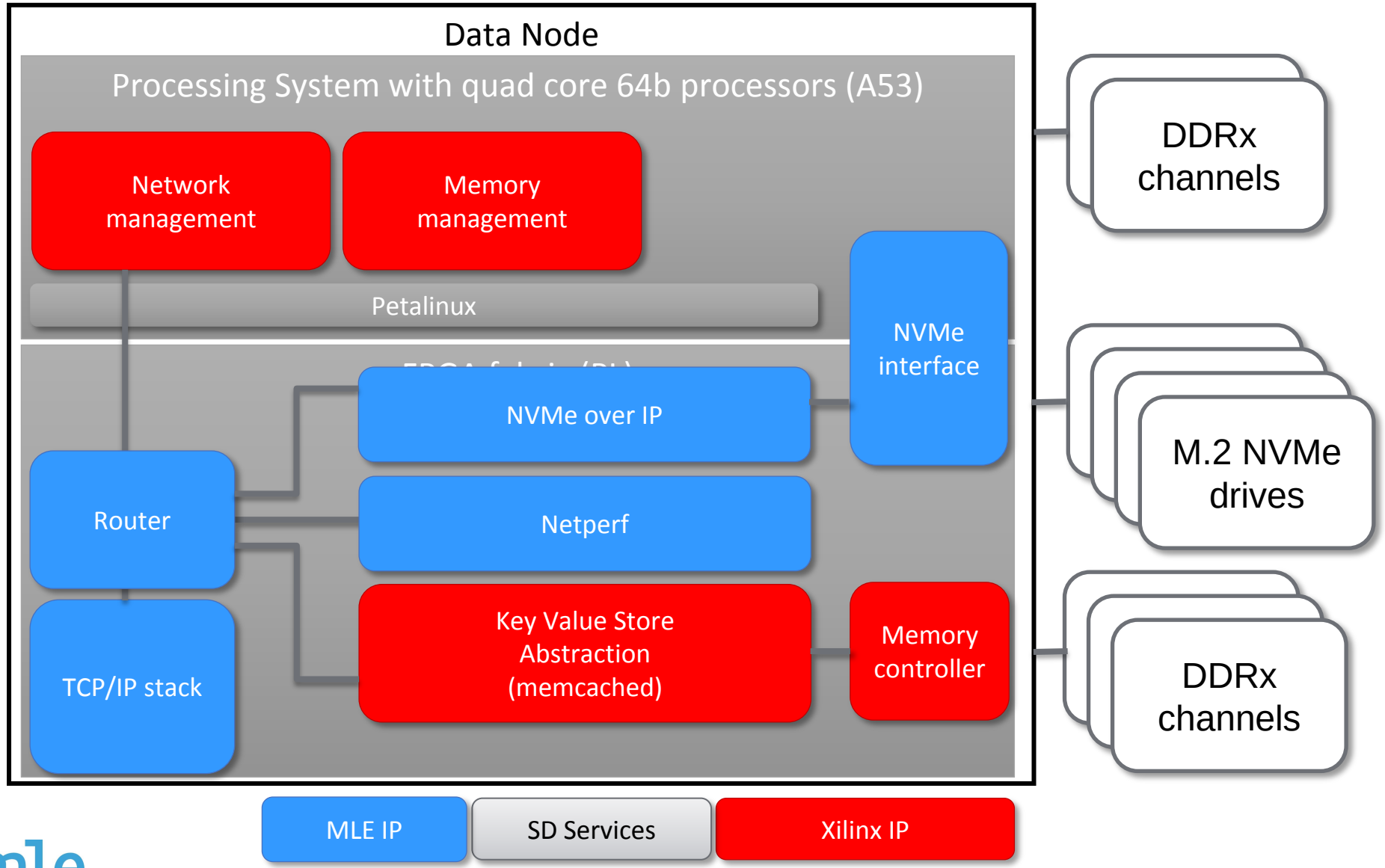
Results

Results: Networked Object Storage Board with Xilinx Zynq Ultrascale+ MPSoC

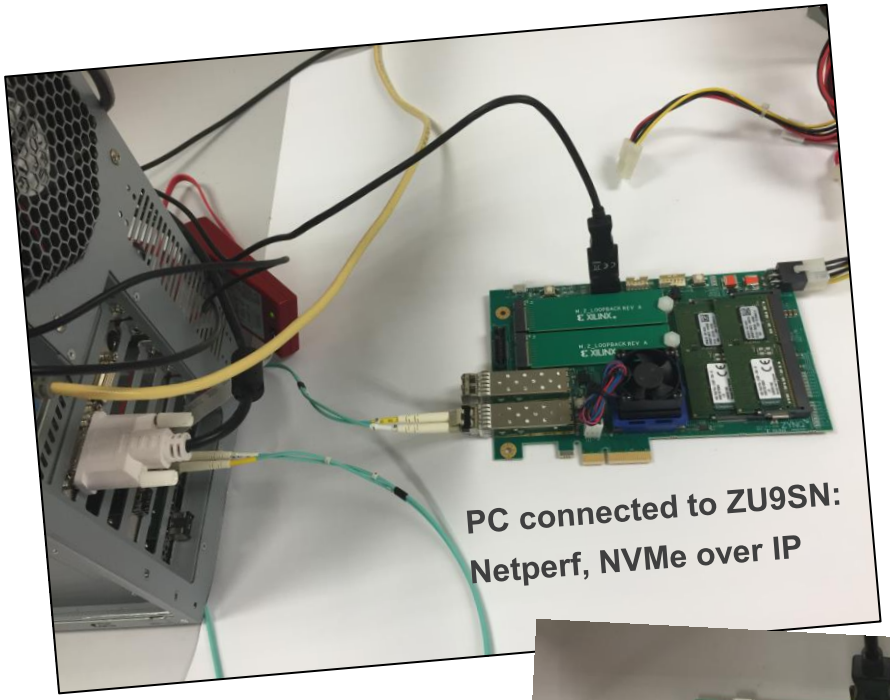
50Gbps key value store with 2TB, on a 35W board

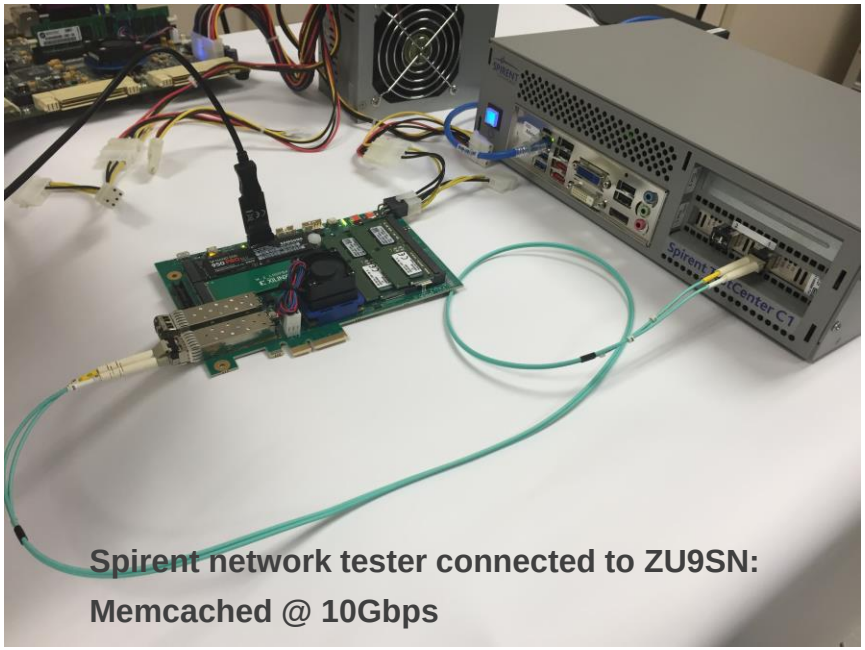


Results: Current Prototype Architecture



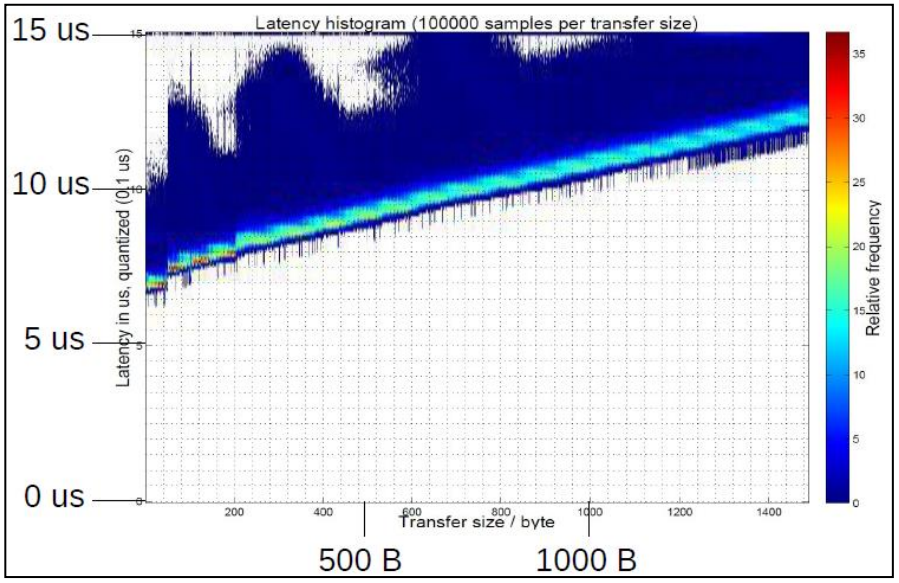
Experiments



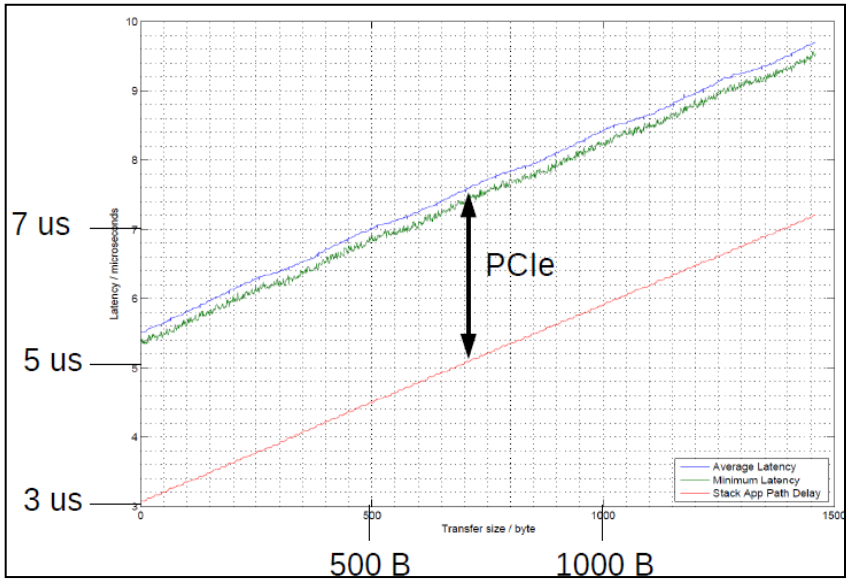


Results: Latency Analysis of Full Accelerator

Software (CPU + NIC)

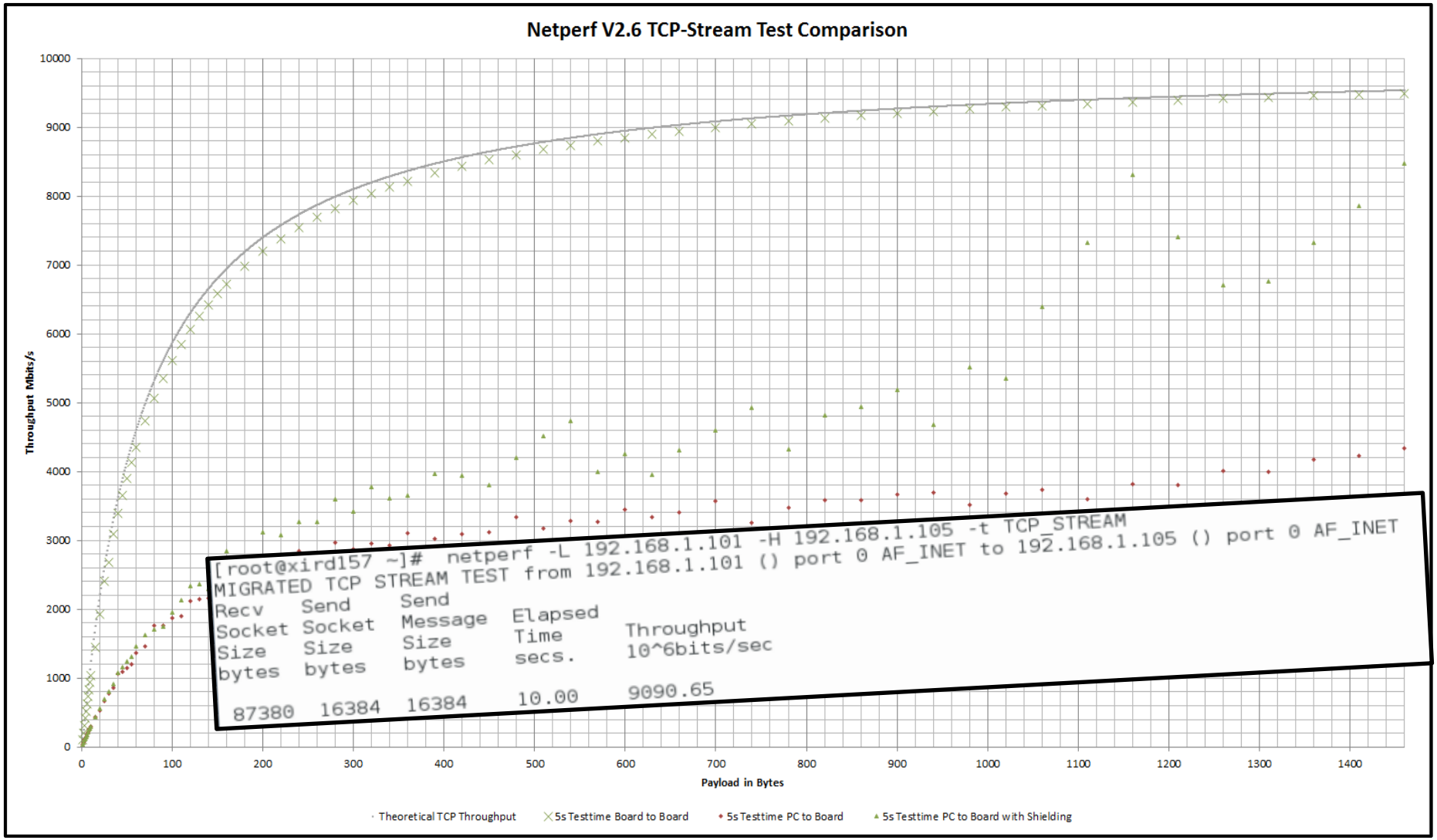


FPGA-based Full Accelerator

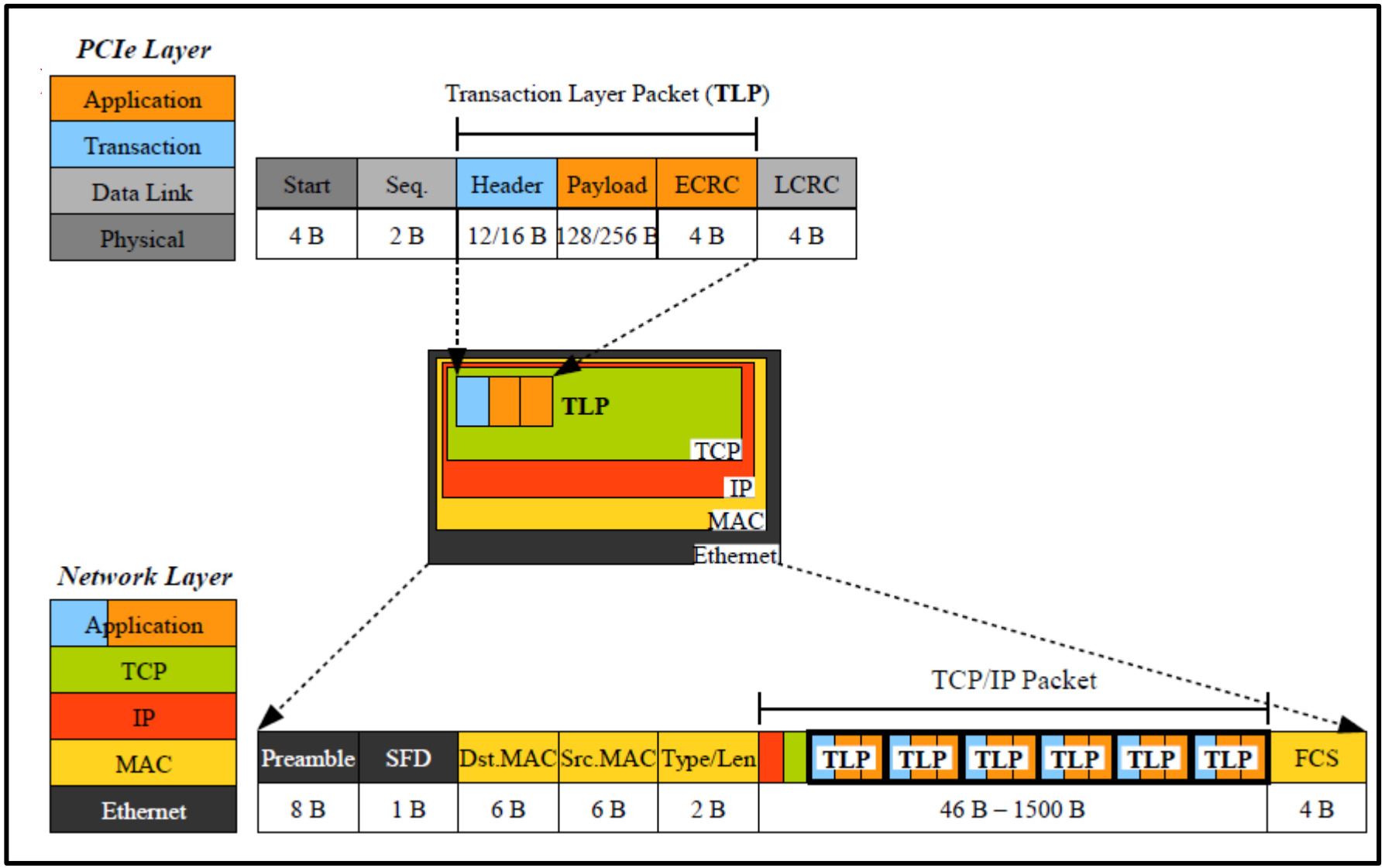


- **Lower and predictable latency with very little jitter**

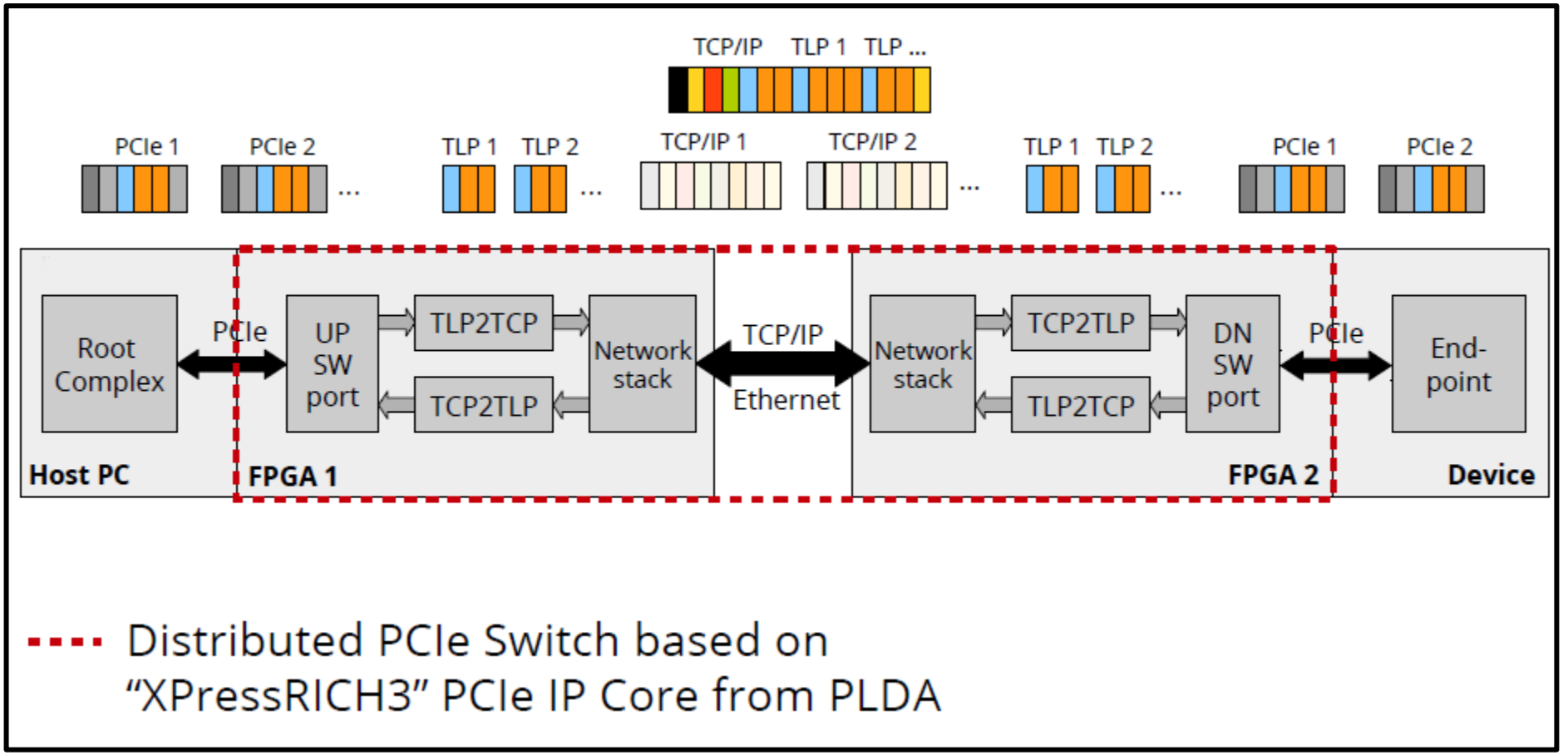
Results: Throughput Analysis of Full Accelerator



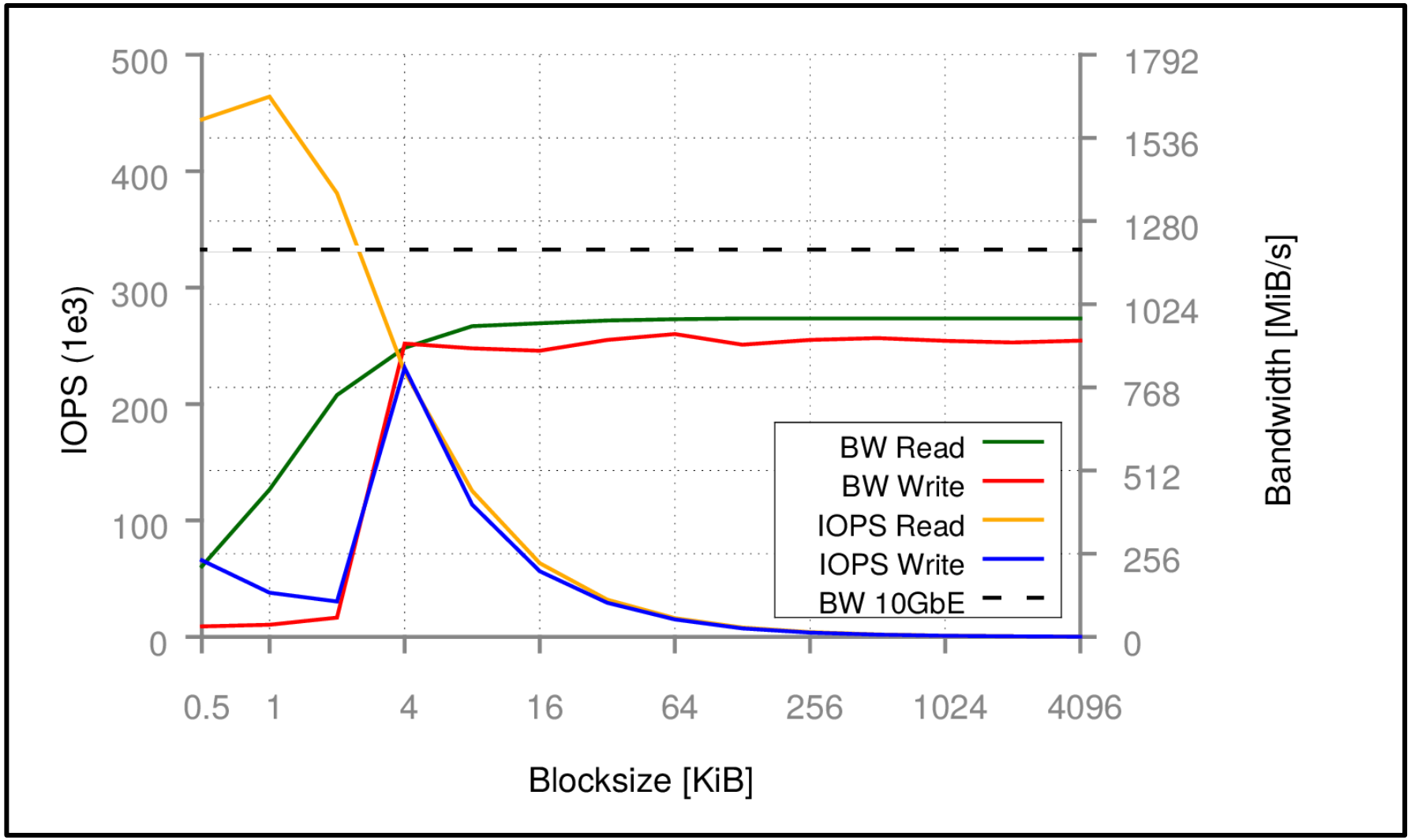
Results: Feasibility of “NVMe-over-IP”



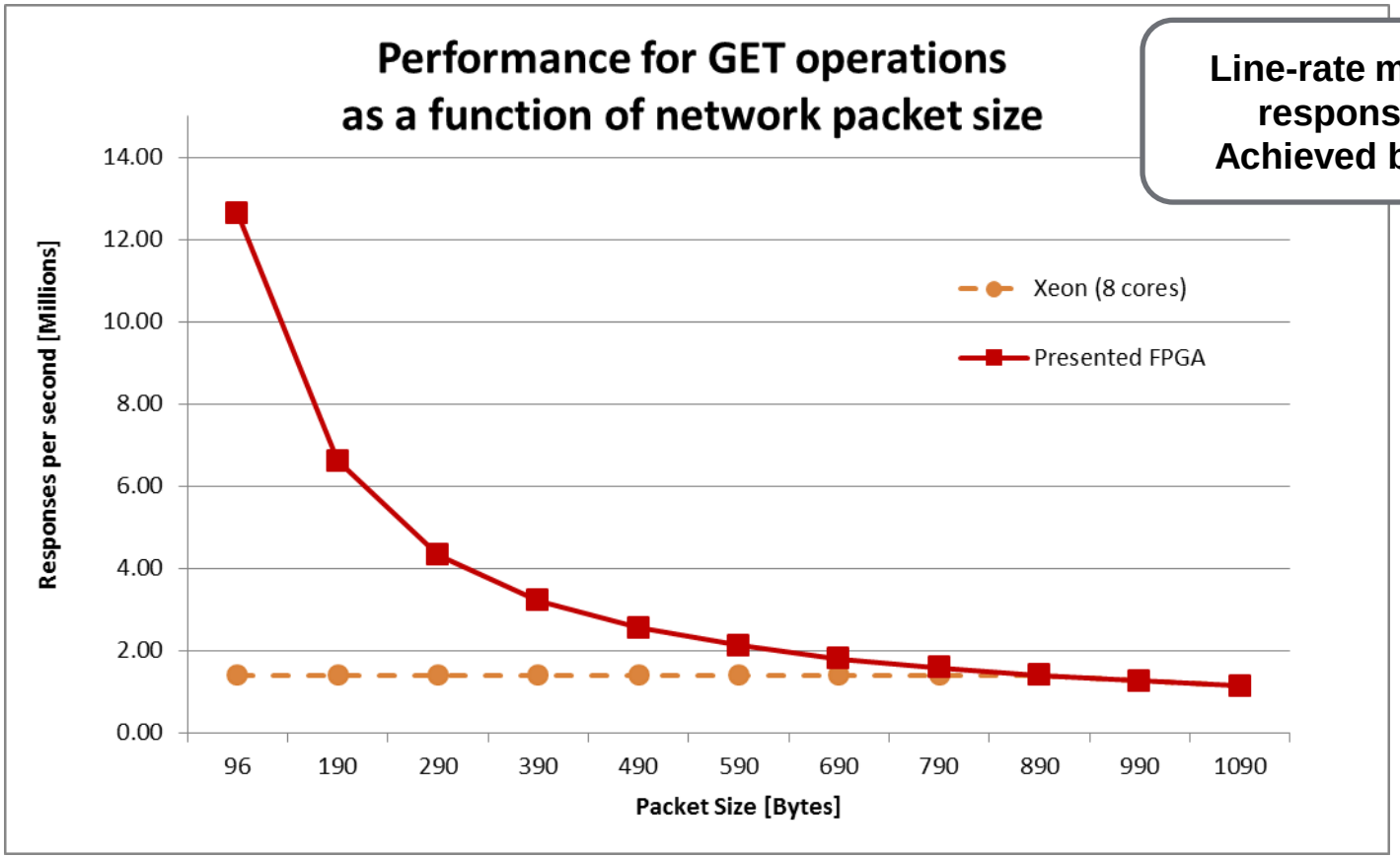
Results: Feasibility of non-legacy NVMe-over-IP



Results: Feasibility of non-legacy NVMe-over-IP



Results: Dataflow Architecture for Acceleration of Key-Value-Stores (KVS)



Line-rate maximum response rate Achieved by FPGA

FPGA is network bound, supports currently 52MRPS

Results: Comparison with best published results

X86 Platforms	GB	KRPS	Watt
Dual x86 (Mica) [12] – research	64	76,900	478
Dual x86 (FlashStore) [6] - research	80	57	84

X86 are limited extracting performance out of flash
- **Support either high performance or high capacity**

FPGA Platforms	GB	KRPS	Watt
Prototype (10Gbps)	512	13’000	35
Estimation (50Gbps, 2 M.2)	8’192	65’000	TBD
Estimation (100Gbps)	40’000	130’000	TBD

Dataflow architecture on MPSoC can support high performance and high capacity at lower power

Source:
HotStorage 2015, Scaling out to a Single-Node 80Gbps Memcached Server with 40Terabytes of Memory
[6] Debnath et al: Flashstore: High throughput persistent key-value store; PVLDB 2010
[12] Lim et al: Mica: A holistic approach to fast in memory key-value storage; NSDI 2014

Conclusion & Outlook

Conclusion

- **Trend towards unconventional architectures**
  - A diversification of increasingly heterogeneous devices and systems
  - Convergence of networking, compute and storage within single nodes
- **Key concepts for implementation of hyper-converged storage nodes**
  - Heterogeneous compute device as a single-chip solution
  - Direct network interface with a full hardware TCP/IP stack
  - Data flow architecture to accelerate all data processing
  - NVMe for multi-terabyte storage capacity
  - Hybrid memory system (DRAM & flash) for high capacity and high performance
- **Results:**
  - First prototype board build for 50Gbps with 2TB key value store
  - Proof of concept demonstrates:
    10Gbps TCP/IP stack, 13MRPS, 10Gbps key value store, 35Watt

Outlook

- **Exploration of first software defined services**
- **Joint evaluation with potential customers & universities, MLE and Xilinx to measure system-level benefits**