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

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

SDC, 2016
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 compute 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
WHO?

Xilinx Research and Missing Link Electronics
Xilinx – The All Programmable Company

$2.38B FY15 revenue
>55% market segment share
3,500+ employees worldwide

20,000 customers worldwide
3,500+ patents
60 industry firsts
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“
  - Wikipedia

- In their simplest form FPGAs contain:
  - Configurable Logic Blocks
    - AND, OR, Invert & many other logic functions
  - Configurable interconnect
    - Enabling Logic Blocks to be connected together
  - I/O Interfaces

- Devices have up to millions of logic cells
  - Tens of millions of gates

With these elements an arbitrary logic design can be created
Customizable Interfaces & Memory Architectures

FPGA
- caches
- QoS

Flash
QDR SRAM
DDRx

Flexibility to interface to any other device and customize memory architectures
Heterogeneous Multicore with Programmable Logic

Processing System

Application Processing Unit
- ARM® Cortex™-A53
- NEON™ Floating Point Unit
- 32 KB I-Cache w/ECC
- 32 KB D-Cache w/ECC
- Memory Management Unit
- Embedded Trace Macrocell
- GIC-400
- SCU
- CCI/SMMU
- 1 MB L2 w/ECC

Real-Time Processing Unit
- ARM Cortex™-R5
- Vector Floating Point Unit
- Memory Protection Unit
- 128 KB TCM w/ECC
- 32 KB I-Cache w/ECC
- 32 KB D-Cache w/ECC
- GIC

Memory
- DDR4/3/3L, LPDDR4/3
- 32/64 bit w ECC
- 256 KB QCM with ECC
- 64 KB L2 Cache
- Memory Management Unit

Graphics Processing Unit
- ARM Mali™-400 MP2
- Geometry Processor
- Pixel Processor
- Configuration and Security Unit (CSU)
- System Management
- Power Management
- Functional Safety
- Voltage/Temp Monitor
- TrustZone
- Multichannel DMA
- Timers, WDT, Resets, Clocking, & Debug

Platform Management Unit
- Config AES Decryption, Authentication, Secure Boot

Programmable Logic

Storage & Signal Processing
- Block RAM
- UltraRAM
- DSP

General-purpose I/O
- High-Performance HPIO
- High Density HDIO

High-Speed Connectivity
- GTH
- GTY
- 100G EMAC
- PCIe @ Gen4
- Interlaken
- Video Codec H.265/H.264
- System Monitor
Xilinx is Diversified Across Multiple Markets

Aerospace and Defense
Industrial and Medical
Test, Measure and Emulation
Automotive and Transportation

Wireless Communications
Audio, Video Broadcast
Consumer
Wired Comms and Data Center

Mars Rover
MRI scanners
3D televisions
Machine Learning
ADAS
AR
Applications & Architectures

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

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.

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.
WHY?

High-capacity hyper-converged storage needs predictable scalability in performance, and programmability for flexibility
In systems with Nonvolatile Memory software now significantly impacts latency and energy efficiency.

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.

⇒ New compute architectures are needed!
The Von Neumann Bottleneck [J. Backus, 1977]

System performance does not scale with CPU anymore!
Data Processing in Hardware vs. Software

Andre DeHon, U Penn: “Spatial vs. Temporal Computing”

Sequential Processing with CPU
C, C++ Program

Parallel Processing with Logic Gates
VHDL, Verilog "Program"

Using both, spatial and temporal compute, allows to put the computational burden where it belongs!
Architectural Choices for Storage Devices

Log PERFORMANACE

Log FLEXIBILITY

General Purpose Processors

Digital Signal Processors

Application Specific Signal Processors

Field Programmable Devices

Application Specific ICs

Physically Optimized ICs

Source: T.Noll, RWTH Aachen

StrongARM110 0.4 MIPS/mW

TMS320C54x 3MIPS/mW

ICORE 20-35 MOPS/mW

10^3 ... 10^4

10^5 ... 10^6
Heterogeneous Processing Architectures for Higher Performance

Current architecture limits maximum performance to total DMA bandwidth.

Separate control flow and dataflow for higher bandwidth via FPGA-based inline processing, FPGA-integrated NIC and HBA.
### FPGA Comparison Table

<table>
<thead>
<tr>
<th></th>
<th>Kintex-7</th>
<th>Virtex-7</th>
<th>Kintex UltraScale</th>
<th>Kintex UltraScale+</th>
<th>Virtex UltraScale</th>
<th>Virtex UltraScale+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic Cells (K)</td>
<td>478</td>
<td>1,955</td>
<td>1,161</td>
<td>915</td>
<td>4,433</td>
<td>2,863</td>
</tr>
<tr>
<td>UltraRAM (Mb)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>36.0</td>
<td>-</td>
<td>432.0</td>
</tr>
<tr>
<td>Block RAM (Mb)</td>
<td>34</td>
<td>68</td>
<td>76</td>
<td>34.5</td>
<td>132.9</td>
<td>94.5</td>
</tr>
<tr>
<td>DSP Slices</td>
<td>1,920</td>
<td>3,600</td>
<td>5,520</td>
<td>3,528</td>
<td>2,880</td>
<td>11,904</td>
</tr>
<tr>
<td>DSP Performance (symmetric FIR)</td>
<td>2,845 GMACs</td>
<td>5,335 GMACs</td>
<td>8,180 GMACs</td>
<td>6,287 GMACs</td>
<td>4,268 GMACs</td>
<td>21,213 GMACs</td>
</tr>
<tr>
<td>Transceiver Count</td>
<td>32</td>
<td>96</td>
<td>64</td>
<td>76</td>
<td>120</td>
<td>128</td>
</tr>
<tr>
<td>Maximum Transceiver Speed (Gb/s)</td>
<td>12.5</td>
<td>28.05</td>
<td>16.3</td>
<td>32.75</td>
<td>30.5</td>
<td>32.75</td>
</tr>
<tr>
<td>Total Transceiver Bandwidth (full duplex) (Gb/s)</td>
<td>800</td>
<td>2,784</td>
<td>2,086</td>
<td>2,478</td>
<td>5,886</td>
<td>8,384</td>
</tr>
<tr>
<td>Memory Interface (DDR3)</td>
<td>1,866</td>
<td>1,866</td>
<td>2,133</td>
<td>2,133</td>
<td>2,133</td>
<td>2,133</td>
</tr>
<tr>
<td>Memory Interface (DDR4)</td>
<td>-</td>
<td>-</td>
<td>2,400</td>
<td>2,667</td>
<td>2,400</td>
<td>2,667</td>
</tr>
<tr>
<td>PCI Express®</td>
<td>x8 Gen2</td>
<td>x8 Gen3</td>
<td>x8 Gen3</td>
<td>x8 Gen 4</td>
<td>x8 Gen3</td>
<td>x8 Gen 4</td>
</tr>
</tbody>
</table>

All-Programmable Design Flow Options

Proven block-based RTL Synthesis design flow combined with modern C/C++/SystemC High-Level-Synthesis
Design automation runs scheduling and resource binding to generate RTL code comprising data paths plus state machines for control flow.
Benefits of HLS-Based C/C++ FPGA Design

- Automated performance optimizations via parallelization at dataflow level
- Automatic interface synthesis and driver code generation for HW/SW connectivity
Idea: An Extensible Architecture for Storage Devices

Combination of multiple concepts:
1. Heterogeneous compute device as a single-chip solution
2. Direct network interface with full accelerator for protocols
3. Performance scaling with dataflow architectures
4. Scaling capacity and cost with a Hybrid Storage subsystem

Work-in-progress with a working Proof-of-Concept
An Extensible Architecture for Storage Devices

Data Node

Processing System with 64bit Quad-Core ARM A53

- Memory management
- Petalinux

FPGA fabric (PL)

- NIC/Bypass (routes traffic on basis of port)
- 10G if TCP/IP stack (limited session support)

KVS

Hybrid memory system

NVMe if

MIG

MLE IP

SD Services

Xilinx IP

DRAM channel

M.2 channel

M.2 channel

DRAM channel
WHAT?

A single-chip heterogeneous compute solution for Terabit-per-second processing
Networked Object Storage Node with MPSOC

Concept 1: Hardware Accelerated Network Stack

- **PS**
  - Fully hardware accelerated TCP/IP stack

- **FPGA fabric (PL)**
  - NIC/Bypass (routes traffic on basis of port)
  - 10G if TCP/IP stack (limited session support)

- **Data Path**

- **Petalinux**
Concept 1: Direct network interface with a full accelerator for network protocol processing

- 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 2: Dataflow architectures for performance scaling

- **Streaming architecture:** Flow-controlled series of processing stages which manipulate and pass through packets and their associated state.

  - **NIC/Bypass:** (routes traffic on basis of port)
  - **10G if TCP/IP stack**

 Dram channel

- **10Gbps demonstrated with a 64b data path @ 156MHz using 3% of FPGA resources**
- **80Gbps can be achieved by using a 512b @ 156MHz pipeline for example**

Concept 3: 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
Object distribution on the basis of size

<table>
<thead>
<tr>
<th>Value Size (B)</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>768</th>
<th>1K</th>
<th>4K</th>
<th>8K</th>
<th>32K</th>
<th>1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>0.55</td>
<td>0.075</td>
<td>0.275</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>Twitter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.85</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wiki</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.29</td>
<td>0.008</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Flickr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

- **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:*  
Dataflow architectures can accommodate high latency accesses without sacrificing throughput.

- In dataflow architectures: no limit to number of outstanding requests
- Flash can be serviced at maximum speed
Custom memory controllers

with out of order processing

Hybrid Memory Controller

Splitter

... 

Merger

DRAM Controller

SATA HBA

DRAM
Value Store

SSD
Value Store

© Copyright 2015 Xilinx
Concept 4: SD Services

Spatial computing of additional services at no performance cost until resource limitations are reached
By combining modern FPGA design methodologies, including High-Level Synthesis, with IP cores for full acceleration of rich software.
Results: Networked Object Storage Board with Xilinx Zynq Ultrascale+ MPSoC

50Gbps key value store with 2TB, 25W

- Dual DDR4 SODIMM 16GB x72 ECC DR 273 Gb/s @ 2133 Mb/s
- Dual M.2 2x SSD 512 GB
- Dual SFP+ 2x 10/25 Gbps
- 16nm MPSoC Quad A53 CPU Embedded FPGA
Results: Current Prototype Architecture

Data Node

PS

FPGA fabric (PL)

Memory management

Petalinux

NIC/Bypass (routes traffic on basis of port)

10G if TCP/IP stack (limited session support)

NVMe if

MIG

DRAM channel

M.2 channel

M.2 channel

DRAM channel

NVMe over IP

NetPerf

KVS
Experiments

PC connected to ZU9SN: Netperf, NVMe over IP

Spirent network tester connected to ZU9SN: Memcached @ 10Gbps

35 Watt under load
Results: Latency Analysis of Full Accelerator

- **Software (CPU + NIC)**
- **FPGA-based Full Accelerator**

![Latency histogram (100000 samples per transfer size)](image)

- Latency in ns: quantized (0-15 ns)
- Transfer size: 500 B, 1000 B

![Latency graph](image)

- Latency in ns: quantized (0-15 ns)
- Transfer size: 500 B, 1000 B
- PCIe delay

© Copyright 2015 Xilinx
Results: Throughput Analysis of Full Accelerator

Netperf V2.6 TCP-Stream Test Comparison

[root@x1rd157 ~]# netperf -L 192.168.1.101 -H 192.168.1.105 -t TCP_STREAM

MIGRATED TCP STREAM TEST from 192.168.1.101 () port 0 AF_INET to 192.168.1.105 () port 0 AF_INET

Recv  Send  Send  Socket  Socket  Message  Elapsed  Throughput
Size  Size  Size  bytes  bytes  bytes  time  10^6 bits/sec
87380 16384 16384 10.00  9090.65

Payload in Bytes

Throughput in Mbits/s

• Theoretical TCP Throughput
• 5s Testtime Board to Board
• 5s Testtime PC to Board
• 5s Testtime PC to Board with Shielding
Results: Feasibility of “NVMe-over-IP”
Results: Feasibility of non-legacy NVMe-over-IP

Distributed PCIe Switch based on “XPressRICH3” PCIe IP Core from PLDA
Results: Feasibility of non-legacy NVMe-over-IP
Results: Dataflow Architecture for Acceleration of Key-Value-Stores (KVS)

Performance for GET operations as a function of network packet size

Demonstrator:
- Up to 36x in performance/power demonstrated
- Supports 52MSearches/second + 52MUpdates/second
- Plus 10-100x in latency
- Scalability to higher rates possible

Line-rate maximum response rate Achieved by FPGA
## Comparison with best published results

<table>
<thead>
<tr>
<th>Platforms</th>
<th>GB</th>
<th>KRPS</th>
<th>Watt</th>
<th>KRPS/Watt</th>
<th>GB/Watt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual x86 (Mica) [12]</td>
<td>64</td>
<td>76,900</td>
<td>478</td>
<td>161</td>
<td>0.1</td>
</tr>
<tr>
<td>Dual x86 (FlashStore) [6]</td>
<td>80</td>
<td>57</td>
<td>84</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>FAWN (SSD) [2]</td>
<td>32</td>
<td>35</td>
<td>15</td>
<td>2.3</td>
<td>2.1</td>
</tr>
<tr>
<td>FPGA (theoretical)</td>
<td>40,000</td>
<td>104,000</td>
<td>435</td>
<td>239</td>
<td>92</td>
</tr>
<tr>
<td>FPGA (facebook)</td>
<td>20,254</td>
<td>32,657</td>
<td>343</td>
<td>95.2</td>
<td>59</td>
</tr>
<tr>
<td>FPGA (prototype)</td>
<td>272</td>
<td>1,340</td>
<td>27</td>
<td>49</td>
<td>10</td>
</tr>
</tbody>
</table>

**Source:**

- HotStorage 2015, Scaling out to a Single-Node 80Gbps Memcached Server with 40 Terabytes of Memory
- Debnath et al: Flashstore: High throughput persistent key-value store; PVLDB 2010
- 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, 14MRPS, 10Gbps key value store, < 50Watt
Finalizing prototype
Exploration of first software defined services
Joint evaluation with potential customers & universities, MLE and Xilinx to measure system-level benefits
Applications require
- Increasing compute (machine learning, data analytics)
- Increasing storage capacity (photos, videos)
- Lower power (OPEX = 2x CAPEX)

Architectural innovation is required
- to provide further performance and storage scaling while reducing power

Integration of compute, memory and network within individual nodes
- Hyper-converged storage
Value of Heterogeneous Multicore Processing

- Maximize Performance via Application Specific Processing
  - Function mapping to optimal processing element

- Reduce Power via System Integration
  - Function optimization to processing element
  - Reduced interconnect power

- Reduce BOM Cost via System Integration
  - Fewer components
  - Reduced board space

- Increased Safety via Functional Safety Components
  - Critical component redundancy
  - Increased integration

- IP, Peripheral & Processing Customization via Programmable Logic
Concept: SD Services

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

PS

- Memory management
- Petalinux

FPGA fabric (PL)

- NIC/Bypass (routes traffic on basis of port)
- 10G if TCP/IP stack (limited session support)
- KVS
- Hybrid memory system
- MIG
- NVMe if

Connections:
- DRAM channel
- M.2 channel
- M.2 channel
- DRAM channel

Other Components:
- MLE IP
- SD Services
- Xilinx IP

© Copyright 2015 Xilinx