Implementation and impact of an ultra-compact multi-FPGA board for large system prototyping

Plan

• Context
• Board implementation
• Prototype environment
• Board impact
• Conclusion
The Quad-FPGA DaughterBoard (QFDB) inception

- Two sister projects:
  - ExaNeSt (Towards ExaScale-level Network, Storage and density)
  - EcoScale (Towards unified remote Hardware acceleration)

- How to prototype new ideas for distributed systems?
  - Simulation
  - Real hardware

- Commercial hardware limits flexibility.
  - High-performance interconnect
  - Interconnected accelerators

- How to prototype these ideas at large scale?
  - Programmable Logic is needed to prototype ideas
  - Processors are necessary for control and applications
  - And environment is needed
A quick introduction to Xilinx Zynq Ultrascale+

- Xilinx Zynq Ultrascale+ is both MPSoC and an FPGA

- Processing System (PS):
  - 4x ARM-A53 64-bits in-order cores
  - 2x ARM-R5 cores
  - DDR4 controller (up to DDR4-PC2133)
  - ARM SMMUv2

- Programmable Logic (PL):
  - 16nm FinFET+
  - 2K5 DSP cores
  - 3MB SRAM
  - 16x 16Gb/s GTH transceivers

- Strong PS↔PL connection:
  - ~100ns read latency
  - PS→PL: 2x 128-bit @333MHz
  - PL→PS: 7x AXI4 128-bit @333MHz, coherence support

Zynq Ultrascale+ overview, www.xilinx.com
The QFDB architecture

- GTH transceivers up to 16Gb/s
- Centralized connectivity
- On-board SSD
- QSPI flash for boot
- 15 power sensors
Have a look!

Stack height: 25mm

120mm

130mm
A few words about bring-up

- Hardware is hard
  - High-density boards are harder
  - Need strong PCB&firmware&software understanding

- Stress boards early!

Programmable Logic Controlled Stress IP

Real-time Monitoring script

Memory borrowing
Prototype infrastructure

• Provide users convenient access
• Support various network configurations
• ManagerPC functionalities are implemented in VM for easier deployment
QFDB boot sequence

• Developed in-house yat tool
  • Apply patches automatically
  • Support different profiles
  • Git-friendly
  • Generate flash images

• Boot packages
  • Bitstream programming
  • Device tree
  • Kernel (if needed)

• Node-level configuration
  • Boot packages stored remotely
  • ID-based boot package selection

QFDB boot sequence

Typical boot

---

F. Chaix - Implementation and impact of an ultra-compact multi-FPGA board for large system prototyping – H2RC’19
Single FPGA acceleration capabilities

- Proof-of-concept Matrix Multiplication accelerator
- Designed using HLS flow
  - Single-precision, Tiled approach (128x128)
  - Loop unrolling \((k, j)\) by 4

\[
\begin{align*}
\text{for } i \text{ in } 0 \text{ to } n & \text{ do} \\
\text{for } j \text{ in } 0 \text{ to } n & \text{ do} \\
\text{for } k \text{ in } 0 \text{ to } n & \text{ do} \\
C[i][j] & := A[i][k] \times B[k][j]
\end{align*}
\]

- Adjusted to exploit PS\(\leftrightarrow\)PL bandwidth
- 275 FP32 Gflop/s @ 300MHz
- 17 FP32 Gflop/s/Watt (dynamic)

<table>
<thead>
<tr>
<th>Resource</th>
<th>LUTs</th>
<th>Flip-flops</th>
<th>DSPs</th>
<th>BRAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>153K</td>
<td>300K</td>
<td>2057</td>
<td>416</td>
</tr>
<tr>
<td>%</td>
<td>56%</td>
<td>55%</td>
<td>82%</td>
<td>46%</td>
</tr>
</tbody>
</table>

EcoScale prototype

- 8 QFDBs per ‘baseboard’
- 1U air-cooled rack
- Power supply and cooling
- High connectivity
Shared multi-FPGA reconfigurable acceleration

• Unified sharing of Programmable Logic and memory across nodes
  • Each core can access memory in any node
  • Each core can spawn accelerator in any node (and at any time)
  • Each accelerator can access memory in any node

• Take away many hassles
  • System appears as a large fragmented FPGA

I. Mavroidis et al., ECOSCALE: Reconfigurable computing and runtime system for future exascale systems, DATE 2016
ExaNeSt prototype

- 4x QFDBs plugged on ‘mezzanine’ boards
- Immersed into liquid-cooled blades
- Current: 48 QFDBs = 192 nodes
- Projected: 64 QFDBs = 256 nodes

Populated mezzanine

Liquid-cooled rack

(1) Power Supply Unit
(2) 48-port 1GbE switches
(6) 16-port 10GbE switches
(12) Blades
High-performance Interconnect research

• Leverage Low-latency coherent access to cores memory
• Minimize cores overhead
• Developed a full stack
  • ExaNet low-latency interconnect
  • 3D torus
  • Virtualized endpoint devices
  • R5 cores software for control
  • DMA/packetizer/mailbox libraries
  • MPI/PGAS implementation
• Other activities ongoing
  • Congestion control
  • Multipath

M. Ploumidis et al., **Software and Hardware co-design for low-power HPC platforms**, *ExaComm 2019*
Conclusions

• A Quad-FPGA DaughterBoard (QFDB) was developed and tested

• Two prototypes were built on that
  • ExaNeSt: Liquid-cooled, ExaNet interconnect, 192 → 256 nodes
  • EcoScale: Air-cooled, AXI interconnect, 64 nodes

• A software environment was built to accelerate research activities
  • Automated patching and profiles support
  • Boot packages to improve versatility

• The board has been used for various research avenues
  • High-performance interconnect
  • Shared Multi-FPGA accelerators
  • Stand-alone accelerators
THANK YOU!