Low-Latency Communication and Acceleration in a liquid-cooled energy-efficient Prototype Rack

Manolis Katevenis (Coordinator) and the ExaNeSt Consortium

EuroHPC Summit Week – Exascale HPDA Workshop – 16 May 2019, Poznan

update of 30 August 2019

This project was funded by the European Union’s H2020-FET-HPC Programme under GA 671553
European Exascale System Interconnect & Storage

www.exanest.eu

in collaboration with:

ExaNoDe

ECOSCALE

EURO EXA
Funded by the European Union
Efficiency, Acceleration, Packaging, Network, Storage

• **Energy-efficiency:** ARM processors
  – even with poor FP performance (A53), yet – but preparing for EPI...

• **Reconfigurable Accelerators:** FPGA’s (with embedded A53 hard-macros)

• **Dense Packaging:** reduce volume & latency ⇒ liquid-cooling

• **Interconnection Network:** low latency, low cost, high thruput, resilient

• **Storage:** distributed, in-node NVMe, full systems software stack

  + Real, full **Applications** ported, optimized, evaluated
Dense Packaging 1: Quad-FPGA Daughter Board (QFDB)

- 4x Zynq FPGA = 16x ARM 64-bit A53 proc. cores + 10k DSP slices + 2.4 million logic elements
- 64 GBy DRAM 2133MHz, ECC
- 250 GBy SSD
- 10 off-board links x 10 Gbps

- 120x130 mm² board, 12 PCB layers, 1700 components, 46 power supplies, 16 power sensors
  - 24 high-speed serial x16 Gbps + 144 LVDS-pair on-board links
Dense Packaging 2: Liquid Cooling

- Electronics immersed in non-conducting liquid
- Currently: vertical blades, fully immersed
- Next Generation: horizontal, sprinkled

Rack-level water circulation
The HPC Testbed

- Currently: 8 Blades
  = 32 QFDBs = 128 FPGAs
  = 512 cores (64-bit A53)
  + 2 TBy DRAM + 8 TBy SSD
- Runs full systems software stack & HPC jobs mng’mnt
- Runs full, real Applications
- 4 more Blades to be added
The “ExaNet” Interconnection Network

• 3-Dimensional Torus, via on-chip (FPGA) routers
  – 10 Gbit/s (full-duplex) per external link, 16 Gb/s per QFDB-internal link
  – 70 ns one-way per chip-to-chip link; $17 \times 6.7 = 115$ ns on-chip per router hop
  – router cost (10 ports) = 22% of ZU9 programmable logic (60 kLUT’s, 0.5 MByte SRAM)

• Virtualized Network Interface, on-chip (FPGA)
  – 1024-channel, 8 protection domains, 64-bit virtual address Remote DMA Engine
  – virtualized packetizers to send, mbox queues to receive 16-Byte “atomic” messages
  – error checking, NACK / time-out, retransmissions, all in hardware
  – 490 ns one-way, one-hop, user-to-user software ping-pong latency (16 Bytes)
  – NI cost = 18 % of ZU9 (49 kLUT’s, 0.25 MByte SRAM) + 1 RT (“Real-Time”) core

• For Intra-Rack Network: simulation studies, Optical Switch chip fab
• avoid system call: multi-channel engine, virtual addresses, SMMU for translation
• avoid copies to/from non-cacheable pages or cache flushes: cache-coherent I/O
• avoid buffer registration: from/to anywhere in user memory w. SMMU transl’tn
• avoid kernel buffers for pinning: allow page-faults during DMA, then restart
• avoid copying: user/lib double-buf’rng; reduce latency: early match /eager send
Global + Local Parallel Storage, Virtualization

*Global Storage + per-job SSD/NVM on-demand temporary Parallel FS*

- **BeeGFS** parallel filesystem (open source), with replication extensions
- System integration with SLURM queue manager
  - provide local parallel storage to the jobs, thus benefiting from proximity
- Low-latency memory-mapped storage access path in Linux
- Virtual Machines (VM):
  - RDMA from within VM
  - transparent MPI remoting
  - acceleration for host-to-VM and VM-to-VM interactions, through RDMA mapping
Real, full Applications, ported and Optimized to ARM

• Material science: *LAMMPS, MiniMD*
• Climate forecasting: *REGCM*
• Engineering – Computational Fluid Dynamics: *openFoam, SailFish*
• Astrophysics – large-scale high-resolution simulations of cosmic formation and evolution: *Gadget, Pinocchio, ExaHiNbody*
• Neuroscience – brain simulation: *DPSNN*
• **Data Analytics:** *MonetDB*
MonetDB open-source RDBMS with BeeGFS locality

• Added **Scale-Out** features to MonetDB
  – The MonetDB Relational DBMS had so far only focused on vertical **Scale-Up**
  – Other Distributed RDBMS’s do not scale-up well
  – Other Open-Source RDBMS’s are not as strong as MonetDB in Analytics

⇒ first step into combining the best of both worlds, Scale-Out & Scale-Up

• Preliminary experimental results
  – Analytics benchmark based on 26 years of US domestic flights data, ~150 million records
  – Speed-up of up to ~2x
• System Energy to solution ~6x to 10x less on ExaNeSt Prototype versus Intel Linux Cluster with Infiniband ConnectX®-3 Pro Dual QSFP+ 54Gbps
• Problem size doubles every time system size doubles

<table>
<thead>
<tr>
<th>Number of FPGA's = Number of Intel Sockets, 4 cores each</th>
<th>HPCG (High Performance Conjugate Gradient)</th>
<th>HPL (High Performance Linpack)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ExaNeSt [kJ]</td>
<td>Intel Cluster [kJ]</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>449</td>
</tr>
<tr>
<td>8</td>
<td>83</td>
<td>686</td>
</tr>
<tr>
<td>16</td>
<td>219</td>
<td>1264</td>
</tr>
<tr>
<td>32</td>
<td>440</td>
<td>2864</td>
</tr>
</tbody>
</table>
Evaluation, 2 of 6 – Energy to Solution (b) Astrophysics

- **GADGET** Astrophysics Application (2,097,152 dark matter particles)
- Compute & memory & communication intensive
- ARM cores (no accelerator), 1 QFDB = 16 cores
- Compared to 2016 Intel cluster, 40 cores
- **Time** to solution: ARM is **3x slower** than Intel
- **Energy** to solution: ARM is **6x better** than Intel
Evaluation, 3 of 6 – Energy-Delay-Product – ARM+FPGA

- **ExaHiNBody** (pure N Body, *Compute intensive*, 2,097,152 particles)
  - Hybrid MPI + OpenMP + OpenCL ⇒ with/without FPGA Accelerator
- Compared to 2016 Intel cluster, 40 cores, with GPUs:
  - Nvidia GTX1080, a gaming GPU, or
  - Nvidia V100, probably the most powerful current GPU
- **Time** to solution:
  - ARM-only is *12x slower* than 2016 Intel cluster;
  - with FPGA is *2x faster* than Intel, *6x faster* than GTX1080 Nvidia
- **Energy-Delay-Product**:
  - ARM-only is *1.3x worse* than 2016 Intel cluster;
  - with FPGA: *600x better* than Intel, *10x* than GTX1080, *2x* than V100
Evaluation, 4 of 6 – Top Green 44 with FPGA Acceleration

- DGEMM benchmark (similar to Linpack)
- Compute bound
- ExaNeSt with FPGA Acceleration
- Double Precision

⇒ ExaNeSt is at least in the top 20% of June’19 Green 500
Evaluation, 5 of 6: Reconfigurable Accelerators (with EcoScale)

• **Oil Reservoir Simulation** (Rachford-Rice equation)
  – applied to selected grid points of the conceptual grid
  – QFDB: **8.5 GFLOPS/Watt, 200 GFLOPS** (300 MHz)
  – Quad-core i5-6685R: **25 GFLOPS**

• **SGEMM** (single precision FP matrix multiply)
  – QFDB: **17 GFLOPS/Watt, 1100 GFLOPS** (at 300 MHz, 82% DSP utilization)
  – Intel Xeon Platinum 8180 (28 cores): **3493 GFLOPS**
  – Nvidia P100 GPU (56 Stream Multiprocessors x 64 cores): **6828 GFLOPS**
Evaluation, 6 of 6: Reconfigurable Accelerators (with EuroEXA)

• **Smart City** (real-time video proc., Lucas-Kanade alg.)
  – preliminary: one ZU9 FPGA, yet
  – One FPGA: **36 ms/frame, 8 Watt**
  – Quad-core Xeon E3-1241: **5900 ms/frame, 10 Watt**
  – NVidia GTX 960 (16 SM): **43 ms/frame, 75 Watt**

• **Space-CNN** (Convolutional Neural Network weight compression for space data classification)
  – QFDB: **265 GFLOPS**, (at 250 MHz)
  – NVidia Quadro K2200: **123 GFLOPS**