The Challenges and Opportunities of Micro-Servers in the HPC Ecosystem


Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
The Challenges and Opportunities of Micro-Servers in the HPC Ecosystem

Dimitrios S. Nikolopoulos

School of Electronics, Electrical Engineering and Computer Science
Queen’s University of Belfast

September 4, 2014
Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions
What we know

Technology alone can not bridge the gap$^1$

---

HPC and ARM

Single-core ARM

Figure 3: Single-core performance and energy efficiency results of micro-benchmarks with frequency sweep. Baseline is the Tegra 2 platform running at 1GHz.

Figure 4: Multi-core performance and energy efficiency results of micro-benchmarks with frequency sweep. Baseline is the Tegra 2 platform running on 1GHz.

3.2 Memory bandwidth
Figure 5 shows the memory bandwidth for each platform, measured using the STREAM benchmark [28]. Our results show a significant improvement in memory bandwidth, of about 4.5 times, between the Tegra platforms (ARM Cortex-A9) and the Samsung Exynos 5250 (ARM Cortex-A15). This appears to be mostly due to the better Cortex-A15 microarchitecture which also improves the number of outstanding memory requests [42], and due to an additional channel in memory controller. Compared with the peak memory bandwidth, the multicore results imply an efficiency of 62% (Tegra 2), 27% (Tegra 3), 52% (Exynos 520), and 57% (Intel Core i7-2760QM).

4. PARALLEL SCALABILITY
In this section we investigate the performance, energy efficiency and scalability of an ARM multicore cluster on production applications. Our single-node performance evaluation in Section 3.1 shows that the Tegra 2 is almost eight times slower than the Intel Core i7, both at their maximum operating frequencies, so the applications must be able to exploit a minimum of eight times as many parallel processors in order to achieve competitive time-to-solution. Fortunately, as described in the previous section, newer ARM SoCs are narrowing the gap. In this section, we also evaluate in detail the interconnection networks available on Tegra 2 and Exynos 5 platforms, in terms of latency and effective bandwidth.

---

HPC and ARM

...and the performance shortfall

Figure 3: Single-core performance and energy efficiency results of micro-benchmarks with frequency sweep. Baseline is the Tegra 2 platform running at 1GHz.

Figure 4: Multi-core performance and energy efficiency results of micro-benchmarks with frequency sweep. Baseline is the Tegra 2 platform running on 1GHz.

3.2 Memory bandwidth

Figure 5 shows the memory bandwidth for each platform, measured using the STREAM benchmark [28]. Our results show a significant improvement in memory bandwidth, of about 4.5 times, between the Tegra platforms (ARM Cortex-A9) and the Samsung Exynos 5250 (ARM Cortex-A15). This appears to be mostly due to the better Cortex-A15 microarchitecture which also improves the number of outstanding memory requests [42], and due to an additional channel in memory controller. Compared with the peak memory bandwidth, the multicore results imply an efficiency of 62% (Tegra 2), 27% (Tegra 3), 52% (Exynos 520), and 57% (Intel Core i7-2760QM).

4. PARALLEL SCALABILITY

In this section we investigate the performance, energy efficiency and scalability of an ARM multicore cluster on production applications. Our single-node performance evaluation in Section 3.1 shows that the Tegra 2 is almost eight times slower than the Intel Core i7, both at their maximum operating frequencies, so the applications must be able to exploit a minimum of eight times as many parallel processors in order to achieve competitive time-to-solution. Fortunately, as described in the previous section, newer ARM SoCs are narrowing the gap. In this section, we also evaluate in detail the interconnection networks available on Tegra 2 and Exynos 5 platforms, in terms of latency and effective bandwidth.

Microserver concept

- Lightweight and scale-out oriented
  - 1U fits 24–48 cards
- Targeting datacenters, in particular web services
  - no FP, but latency-sensitive
- Shared fan and power supply
- Wide range of processor choices within low power envelopes
- Favoring commodity memory & interconnects (Ethernet vs. IB, LPDDR vs. DDR)
The NanoStreams proposition

Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions
The NanoStreams proposition

Gap in the server landscape

http://www.nanostreams.eu

D. Nikolopoulos (EEECS@QUB)
NanoStreams AoC block

- AoC host on Calxeda boards (A9 cores, 10 GigE)
  - Odroid boards explored as alternative: (A15 cores, GigE)
- AoC accelerator on Xilinx Zynq boards
NanoStreams software stack

Taming oversubscription and latency

- Space and time isolation of parallel components
- RDMA over raw Ethernet, user-level
- Soft real-time scheduling guarantees
- Locality exploitation both horizontally and vertically
Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions
Option pricing

- Datacenters co-located with trading venues
- No flexibility in moving the datacenter “where electricity is cheap”
- No flexibility in running the datacenter “when electricity is cheap”
- Not particularly compute- or data-intensive, low-latency workloads
  - Monte Carlo simulations, Black Scholes, Binomial Pricing
  - Instance runs in ms or µs
  - Heavily traded symbols trigger Koptions/session

\[
\begin{align*}
\text{Price} &= (-1)^p \left( SN((-1)^p d_1) - Pe^{-rT} N((-1)^p d_2) \right) \\
\text{Price} &= \frac{e^{-rT}}{N} \sum_{i=1}^{N} \max \left( 0, S - Pe^{(r - \frac{\sigma^2}{2}) T + \sigma \sqrt{T} x_i} \right) \\
u &= e^{\sigma \sqrt{T}} \quad \text{and} \quad d = \frac{1}{u}
\end{align*}
\]
Energy-efficiency metrics and measurement approaches

Real-time, latency-sensitive workloads

- **Joules**/option: Provider-side, sustained throughout trading day, reduction translates to less TCO
- **Time**/option: User-side, end-to-end latency.
- **QoS**: Calculating option before new price arrives; unknown deadline.

---

Replayed, real, trading day market feed with 617 option pricing instances on Facebook stock

### Table: Power profiles for standalone kernel kernels

<table>
<thead>
<tr>
<th>Kernel and Platform</th>
<th>N</th>
<th>PRE-VRM $\bar{P}(W)$</th>
<th>Time (s)</th>
<th>J/opt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MC Intel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5M</td>
<td>25.8</td>
<td>8.6</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2.0M</td>
<td>26.0</td>
<td>34.0</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td><strong>MC Viridis(1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5M</td>
<td>6.8</td>
<td>41.2</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>2.0M</td>
<td>7.4</td>
<td>163.7</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td><strong>MC Viridis(16)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5M</td>
<td>108.8</td>
<td>2.9</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>2.0M</td>
<td>118.4</td>
<td>10.1</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td><strong>BT Intel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000</td>
<td>24.5</td>
<td>8.6</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>24.9</td>
<td>32.8</td>
<td>1.86</td>
<td></td>
</tr>
<tr>
<td><strong>BT Viridis(1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000</td>
<td>5.0</td>
<td>42.0</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>5.2</td>
<td>132.0</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td><strong>BT Viridis(16)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000</td>
<td>88.0</td>
<td>2.8</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>97.6</td>
<td>8.0</td>
<td>1.27</td>
<td></td>
</tr>
</tbody>
</table>
Session-wide energy efficiency

Table: J/opt for execution of the standalone kernels using the PRE-PSU power measurement

<table>
<thead>
<tr>
<th>N</th>
<th>Intel</th>
<th>Viridis(1)</th>
<th>Viridis(16)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \bar{P} ) (W)</td>
<td>J/opt</td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td>109.1</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>0.5M</td>
<td>112</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>1.0M</td>
<td>114.1</td>
<td>6.29</td>
</tr>
<tr>
<td></td>
<td>2.0M</td>
<td>112</td>
<td>3.16</td>
</tr>
<tr>
<td>BT</td>
<td>4000</td>
<td>109.8</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>111.7</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>112.1</td>
<td>5.96</td>
</tr>
</tbody>
</table>
How QoS changes the overall picture

Table: QoS metric and TCO in various setups

<table>
<thead>
<tr>
<th>MC 1M</th>
<th>QoS</th>
<th># Options priced</th>
<th>PRE-PSU $\bar{P}(W)$</th>
<th>TCO KWh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel(1)</td>
<td>13.2%</td>
<td>827593</td>
<td>112.0</td>
<td>0.73</td>
</tr>
<tr>
<td>Viridis(1)</td>
<td>2.6%</td>
<td>162873</td>
<td>135.5</td>
<td>0.88</td>
</tr>
<tr>
<td>Viridis(2)</td>
<td>5.2%</td>
<td>325048</td>
<td>141.9</td>
<td>0.92</td>
</tr>
<tr>
<td>Viridis(4)</td>
<td>10.4%</td>
<td>649402</td>
<td>158.0</td>
<td>1.03</td>
</tr>
<tr>
<td>Viridis(8)</td>
<td>20.8%</td>
<td>1305408</td>
<td>187.5</td>
<td>1.22</td>
</tr>
<tr>
<td>Viridis(16)</td>
<td>41.5%</td>
<td>2600416</td>
<td>244.6</td>
<td>1.59</td>
</tr>
<tr>
<td>*Intel(2)</td>
<td>26.4%</td>
<td>1655186</td>
<td>224.0</td>
<td>1.46</td>
</tr>
<tr>
<td>*Intel(3)</td>
<td>39.6%</td>
<td>2482779</td>
<td>336.0</td>
<td>2.18</td>
</tr>
</tbody>
</table>
Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions
Modeling the Energy of NVRAM

- NVRAM is a viable DRAM alternative with DRAM failing to scale beyond 22 nm.
- Various options: PCM, STT-RAM, RRAM.

\[ T(L) = \frac{N}{\phi} (CPI_0 + ML) \]  
\[ E_{\text{mem}} = E_{d,\text{mem}} NM + (P_{s,\text{mem}} S + P_{\text{cpu}}) T(L) \]
\[ \Delta E = \frac{N}{\phi} (\phi \Delta E_d M + CPI_0 \Delta P_s S + \Delta E_s M S + P_{\text{cpu}} M \Delta L) \]

---

### NVRAM versus DRAM

**Iso-energy-efficiency chart**

In the context of non-volatile memory technologies, the chart illustrates the comparison between different memory types, namely RRAM, STT-RAM, and PCM, against DRAM. The x-axis represents the memory size (in GB), while the y-axis indicates the number of memory accesses per 1k instructions. The chart shows the relative performance of each technology as a function of memory size, with DRAM generally performing better with smaller memories but facing diminishing returns as size increases. Conversely, NVM technologies like RRAM, STT-RAM, and PCM tend to have higher memory access numbers, especially at larger memory capacities, due to their inherent characteristics.
Workload characterization for column stores

![Diagram](image)

**Figure**: Object analysis tool

**Figure**: Workload Characterization of MonetDB.
Object placement in hybrid memories

< 20% of objects needed in DRAM

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>Supermicro Intel(R) Xeon(R) CPU E5-4650, 2.70GHz, 32 cores, 20 MB LLC</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>Dynamic Energy (64 bytes)</td>
</tr>
<tr>
<td>DRAM</td>
<td>61 (R), 61 (W)</td>
</tr>
<tr>
<td>PCM</td>
<td>268 (R), 732 (W)</td>
</tr>
</tbody>
</table>

Figure: AMAT versus AMAE

MonetDB

Moving Objects to DRAM

AMAT (Cycles)  AMAE (J) x 10^{-6}

- Q6
- Q9
- Q13
- Q21
Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions
Where do we go from here

- Micro-server concept is not a stranger to HPC
  - BG/P and BG/Q would be good examples of state-of-the-art micro-servers for datacenters

- What could make it a value proposition
  - Improved energy-efficiency in applications where performance requirements are easily met
  - Improved energy-efficiency in data-intensive applications
  - Scale-out and tight-sizing machine for workload, rather than over-provision

- What may not be a value proposition
  - HPC applications that do require absolute peak performance

- What is needed
  - Holistic approaches: whole system design for energy-efficiency (memories, interconnect), co-designed software stack
Credits

- EU FP7 Grant 610509, EPSRC Grants L000055/1, L004232/1

Charles Gillan, Giorgis Georgakoudis, George Tzenakis, Ahmad Hassan, Hans Vandierendonck, Bronis de Supinski