Experiences with Mobile Processors for Energy Efficient HPC

Alex Ramirez, Nikola Rajovic, Alejandro Rico, James Vipond, Nikola Puzovic

Paul Carpenter
Grenoble, 20 March 2013
Outline

Killer mobile processors
- Past: From vector CPUs to commodity components
- Now: Current trends for mobile CPUs

Our experiences
- Tegra 2, Tegra 3, + discrete GPU
- Gap in performance starting to close

The future
- Prototype roadmap at BSC
“Killer microprocessors”

Microprocessors killed the Vector supercomputers

– They were not faster ...
– ... but they were significantly **cheaper** and **greener**

10 microprocessors approx. 1 Vector CPU

– SIMD vs. MIMD programming paradigms
The next step in the commodity chain

- ASCI Red (1997): microprocessors entered HPC
- 2010s: “killer mobiles”?

- Total cores in Jun'12 Top500
  13.5 M cores
- Tablets sold in Q4 2011
  27 M tablets
- Smartphones sold Q4 2011
  > 100 M phones
Background: Mont-Blanc project

To develop a European Exascale approach
- Based on embedded power-efficient technology

Objectives
- Develop first prototype system, limited by available technology
- Design a next-generation system, to overcome the limitations
- Develop a set of Exascale applications targeting the new system
But are we there yet?

History may be about to repeat itself…

But, are mobile SoCs suitable for HPC yet?

Early performance and energy-efficiency results

- Today’s mobile platforms
  - Lower GFLOPS per core, higher GFLOPS/W?
- HPC-specific micro-kernels
  - Stress different architectural features
  - Wide range of HPC domains
  - Reduce porting effort to new architectures
Evaluated Platforms

**Tegra 2**

- **40nm**
- Dual-core ARM Cortex-A9 (1GHz)
- 32KB L1 + 1MB shared L2
- 1GB DDR2
- 100 MbE NIC (on USB) and 1GbE NIC (on PCIe)
- No GPU

**Tegra 2 developer board**

**Tegra 3**

- **40nm**
- Quad-core ARM Cortex-A9 (1.3GHz)
- 32KB L1 + 1MB shared L2
- 2GB DDR3
- No GPU

**CARMA kit**

**Quadro 1000M**

- **40nm**
- NVIDIA Quadro 1000M GPU (CUDA)
- Single 1GbE NIC (USB)
## Micro-benchmarks

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Operation (vecop)</td>
<td>Common operation in regular codes</td>
</tr>
<tr>
<td>Dense Matrix-Matrix Multiplication (dmmm)</td>
<td>Common operation: measures data reuse and compute performance</td>
</tr>
<tr>
<td>3D stencil (3dstc)</td>
<td>Strided memory accesses (7-point 3D stencil)</td>
</tr>
<tr>
<td>2D Convolution (2dcon)</td>
<td>Spatial locality</td>
</tr>
<tr>
<td>Fast Fourier Transform (fft)</td>
<td>Peak floating-point, variable-stride accesses</td>
</tr>
<tr>
<td>Reduction (red)</td>
<td>Varying levels of parallelism (Scalar sum)</td>
</tr>
<tr>
<td>Histogram (hist)</td>
<td>Histogram with local privatisation, requires reduction stage</td>
</tr>
<tr>
<td>Merge Sort (msort)</td>
<td>Barrier operations</td>
</tr>
<tr>
<td>N-Body (nbody)</td>
<td>Irregular memory accesses</td>
</tr>
<tr>
<td>Atomic Monte-Carlo Dynamics (amcd)</td>
<td>Embarrassingly parallel: peak compute performance</td>
</tr>
<tr>
<td>Sparse Vector-Matrix Multiplication (spwm)</td>
<td>Load imbalance</td>
</tr>
</tbody>
</table>
Single-core results

**Single precision**

**Double precision**

<table>
<thead>
<tr>
<th></th>
<th>Tegra2</th>
<th>Tegra3</th>
<th>Quadro 1000M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speedup</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multi-core results

Single precision

<table>
<thead>
<tr>
<th>Speedup</th>
<th>Single precision</th>
<th>Double precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tegra2 : 2th</td>
<td>Tegra3 : 4th</td>
<td>Quadro 1000M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tegra2 : 2th</td>
</tr>
</tbody>
</table>

Double precision

<table>
<thead>
<tr>
<th>Speedup</th>
<th>Single precision</th>
<th>Double precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tegra2 : 2th</td>
<td>Tegra3 : 4th</td>
<td>Quadro 1000M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tegra2 : 2th</td>
</tr>
</tbody>
</table>
Summary of results

- Micro-benchmarks show reducing energy-to-solution
  - Tegra 2 and Tegra 3 both 40nm, same micro-architecture
  - 67% reduction in energy-to-solution through multi-core density

- Gap in performance is starting to close
  - Previous work: Tegra 2 vs Intel Core i7, 1.2x less energy, 9x slower

- Integrated CUDA GPU will improve HPC applicability
  - Mont-Blanc prototype: Samsung Exynos 5 Dual (Mali GPU)
  - Shared cache coherent memory
  - No power wasted on PCIe bus and GDDR5 memory
  - Higher energy efficiency, lower cost

- Future mobile SoCs will be increasingly competitive
  - in terms of performance while increasing energy efficiency
  - Samsung Exynos 5 dual Arndale board
    - > 2x speedup on double precision over Tegra 3 (micro-architecture)
  - 64-bit ARM Cortex-A50 series
  - Increasing multi-core density
Prototypes are critical to accelerate software development
  - System software stack + applications
Next steps

**Killer mobile processors**
- Not there right now, but getting close
- Will see supercomputer with mobile SoCs soon
- Question is whether it will become mainstream

**Mont-Blanc architecture is shaping up**
- ARM multicore + integrated OpenCL accelerator
- Ethernet NIC
- High density packaging

**Stay tuned!**

www.montblanc-project.eu
MontBlancEU
@MontBlanc_EQ