Energy efficient computing on Embedded and Mobile devices

Nikola Rajovic, Nikola Puzovic, Lluis Vilanova, Carlos Villavieja, Alex Ramirez
A brief look at the (outdated) Top500 list

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Center in Tianjin China</td>
<td>Tianhe-1A - NUDT TH MPP, X5670 2.93GHz 6C, NVIDIA GPU, FT-1000 8C NUDT</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Oak Ridge National Laboratory United States</td>
<td>Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.</td>
</tr>
<tr>
<td>4</td>
<td>National Supercomputing Centre in Shenzhen (NSCS) China</td>
<td>Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU Dawning</td>
</tr>
<tr>
<td>5</td>
<td>GSIC Center, Tokyo institute of Technology Japan</td>
<td>TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP</td>
</tr>
<tr>
<td>6</td>
<td>DOE/LLNL/LANL/SNL United States</td>
<td>Cielo - Cray XE6 8-core 2.4 GHz Cray Inc.</td>
</tr>
<tr>
<td>7</td>
<td>NASA/Ames Research Center/NAS United States</td>
<td>Pleades - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0GHz2 Xeon 5570/5670 2.93 Ghz, Infiniband SGI</td>
</tr>
<tr>
<td>8</td>
<td>DOE/Ohio州立大學/NERSC United States</td>
<td>Hopper - Cray XE6 12-core 2.1 GHz Cray Inc.</td>
</tr>
<tr>
<td>9</td>
<td>Commissariat a l'Energie Atomique (CEA) France</td>
<td>Tera-100 - Bull bulix super-node S6010/S6030 Bull SA</td>
</tr>
<tr>
<td>10</td>
<td>DOE/LLNL/LANL United States</td>
<td>Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8l 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM</td>
</tr>
</tbody>
</table>

- Most systems are built on general purpose multicore chips
  - Backwards compatibility
  - Programmer productivity
A brief look at the (soon to be outdated) Green500 list

- Most of the Top10 systems rely on accelerators for energy efficiency
  - ATI GPU
  - Nvidia GPU
  - IBM PowerXCell 8i
Some initial assertions

- You may disagree, but bear with me ...

- Power distribution
  - 5% Power Supply
  - 10% Cooling
    - Direct water
  - 10% Interconnect
    - Not always active
  - 10% Storage
    - Not always active
  - 32.5% Memory
  - **32.5% Processor**
Now, some arithmetic (and some assumptions)

- **Objective**: 1 EFLOPS on 20 MWatt
- **Blade-based multicore system design**
  - 100 blades / rack
  - 2 sockets / blade
  - 150 Watts / socket
- **CPU**
  - 8 ops/cycle @ 2GHz = 16 GFLOPS
- **1 EFLOPS / 16 GFLOPS**
  - 62.5 M cores
- **32% of 20 MWatt = 6.4 MWatt**
  - 6.4 MWatt / 62.5 M cores
  - 0.10 Watts / core
- **150 Watt / socket**
  - 1500 cores / socket
  - 24 TFLOPS / socket
Where are we today?

- **IBM BG/Q**
  - 8 ops/cycle @ 1.6 GHz
  - 16 cores / chip
    - 16K cores / rack
  - ~2.5 Watt / core

- **Fujitsu Ultrasparc VIIIfx**
  - 8 ops / cycle @ 2GHz
  - 8 cores / chip
  - 12 Watts / core

- **Nvidia Tesla C2050-2070**
  - 448 CUDA cores

- **ARM Cortex-A9**
  - 1 ops / cycle @ 800 MHz - 2 GHz
  - 0.25 - 1 Watt

- **ARM Cortex-A15**
  - 4 ops / cycle @ 1 - 2.5 GHz*
  - 0.35 Watt*

- All is there … but not together?

* Estimated from web sources, not an ARM Commitment
Can we build supercomputers from embedded technology?

- HPC used to be the edge of technology
  - First developed and deployed for HPC
  - Then used in servers and workstations
  - Then on PCs
  - Then on mobile and embedded devices

- Can we close the loop?
Energy-efficient prototype series @ BSC

- Start from COTS components
- Move on to integrated systems and custom HPC technology
Tegra2 prototype @ BSC

- Deploy the first large-scale ARM cluster prototype
  - Built entirely from COTS components

- Exploratory system to demonstrate
  - Capability to deploy HPC platform based on low-power components
  - Open-source system software stack

- Enable early software development and tuning on ARM platform
ARM Cortex-A9 multiprocessor

- Energy-efficient application processor
- Up to 4-core SMT
  - Full cache coherency
- VFP Double-Precision FP
  - 1 ops / cycle

### ARM Cortex-A9 Performance Power & Area

<table>
<thead>
<tr>
<th></th>
<th>Cortex-A9 Single Core Soft Macro Trial Implementation</th>
<th>Cortex-A9 Dual Core Hard Macro Implementations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>TSMC 65G</td>
<td>TSMC 40G</td>
</tr>
<tr>
<td>Optimization method</td>
<td>Performance Optimized</td>
<td>Performance Optimized</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power Optimized</td>
</tr>
<tr>
<td>Standard Cell Library</td>
<td>ARM SC12 + High Performance Kit</td>
<td>ARM SC12 + High Performance Kit</td>
</tr>
<tr>
<td>Performance (Total DMIPS)</td>
<td>2,075 DMIPS</td>
<td>10,000 DMIPS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,000 DMIPS</td>
</tr>
<tr>
<td>Frequency</td>
<td>830 MHz</td>
<td>2000 MHz (typical)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800 MHz (watts)</td>
</tr>
<tr>
<td>Energy Efficiency (DMIPS/mW)</td>
<td>5.2</td>
<td>5.26</td>
</tr>
<tr>
<td>Total power at target frequency</td>
<td>0.4 W</td>
<td>1.9 W</td>
</tr>
<tr>
<td>Silicon Area</td>
<td>1.5 mm² (excludes caches)</td>
<td>6.7 mm² (including L1 parity and all DFT/DFM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.6 mm² (including all DFT/DFM)</td>
</tr>
</tbody>
</table>
Nvidia Tegra2 SoC

- Dual-core Cortex-A9 @ 1GHz
  - VFP for DP (no NEON)
    - 2 GFLOPS (1 FMA / 2 cycles)
- Low-power Nvidia GPU
  - OpenGL only, CUDA not supported
- Several accelerators
  - Video encoder-decored
  - Audio processor
  - Image processor
- ARM7 core for power management
- 2 GFLOPS ~ 0.5 Watt
SECO Q7 Tegra2 + Carrier board

- **Q7 Module**
  - 1x Tegra2 SoC
  - 2x ARM Cortex-A9, 1 GHz
  - 1 GB DDR2 DRAM
  - 100 Mbit Ethernet
  - PCIe
    - 1 GbE
    - MXM connector for mobile GPU
  - 4" x 4"

- **Q7 carrier board**
  - 2 USB ports
  - 2 HDMI
    - 1 from Tegra
    - 1 from GPU
  - uSD slot
  - 8" x 5.6"

- **2 GFLOPS ~ 4 Watt**
1U multi-board container

- Standard 19" rack dimensions
  - 1.75" (1U) x 19" x 32" deep

- 8x Q7-MXM Carrier boards
  - 8x Tegra2 SoC
    - 16x ARM Cortex-A9
  - 8 GB DRAM

- 1 Power Supply Unit (PSU)
  - Daisy-chaining of boards
  - ~ 7 Watts PSU waste

- 16 GFLOPS ~ 40 Watts
Prototype rack

- Stack of 8 x 5U modules
  - 4 Compute nodes
  - 1 Ethernet switch

- Passive cooling
  - Passive heatsink on Q7

- Provide power consumption measurements
  - Per unit
    - Compute nodes
    - Ethernet switches
  - Per container
  - Per 5U

- 512 GFLOPS ~ 1.700 Watt
  - 300 MFLOPS / W
  - 60% efficiency ~ 180 MFLOPS / W
Manual assembly of board container
Manual assembly of containers in the rack + interconnect wiring
## System software stack

<table>
<thead>
<tr>
<th>Category</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open source system software stack</strong></td>
<td></td>
</tr>
<tr>
<td>Linux OS</td>
<td></td>
</tr>
<tr>
<td>GNU compiler</td>
<td></td>
</tr>
<tr>
<td>gcc 2.4.6</td>
<td></td>
</tr>
<tr>
<td>gfortran</td>
<td></td>
</tr>
<tr>
<td>Scientific libraries</td>
<td></td>
</tr>
<tr>
<td>ATLAS</td>
<td></td>
</tr>
<tr>
<td>FFTW</td>
<td></td>
</tr>
<tr>
<td>HDF5</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster management</strong></td>
<td></td>
</tr>
<tr>
<td>Runtime libraries</td>
<td></td>
</tr>
<tr>
<td>MPICH2</td>
<td></td>
</tr>
<tr>
<td>OpenMPI</td>
<td></td>
</tr>
<tr>
<td><strong>OmpSs toolchain</strong></td>
<td></td>
</tr>
<tr>
<td>Performance analysis tools</td>
<td></td>
</tr>
<tr>
<td>Paraver</td>
<td></td>
</tr>
<tr>
<td>Scalasca</td>
<td></td>
</tr>
<tr>
<td><strong>Runtime libraries</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Performance analysis tools</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Cluster Management</strong></td>
<td></td>
</tr>
<tr>
<td>slurm</td>
<td></td>
</tr>
<tr>
<td>GridEngine</td>
<td></td>
</tr>
<tr>
<td>Linux OS</td>
<td></td>
</tr>
</tbody>
</table>
Processor performance: Dhrystone

- Validate if Cortex-A9 achieves the ARM advertised Dhrystone performance
  - 2.500 DMIPS / GHz
- Compare to PowerPC 970MP (JS21, MareNostrum) and Core i7 (laptop)
  - ~ 2x slower than ppc970
  - ~ 9x slower than i7

<table>
<thead>
<tr>
<th>Platforms</th>
<th>Energy (J)</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tegra 2</td>
<td>110.6</td>
<td>1.0</td>
</tr>
<tr>
<td>Core i7</td>
<td>116.8</td>
<td>1.056</td>
</tr>
</tbody>
</table>
Processor performance: SPEC CPU 2006

- Compare Cortex-A9 @ 1 GHz CPU performance with 3 platforms
  - ppc970 @ 2.3 GHz ~ 2-3x slower (= if we factor in freq.)
  - Core2 @ 2.5 GHz ~ 5-6x slower
  - Core i7 @ 2.8 GHz ~ 6-10x slower (2-4x slower if we factor freq.)

- Is it more power efficient?
Tegra2 not always more power-efficient than Core i7
- i7 efficiency is better for benchmarks where it outperforms A9 by 10x
Standard HPL, using ATLAS library
- ATLAS microkernels also achieve 1 GFLOPS peak performance

1.15 GFLOPS ~ 57% efficiency vs. peak performance

~200 MFLOPS / Watt
- In line with original predictions
Cluster performance: Linpack

- 24 nodes
  - 3 x 8 boards (48 GFLOPS peak)
  - 1 GbE switch
- 27.25 GFLOPS on 272 Watts
  - 57% efficiency vs. peak
  - 100 GFLOPS / Watt
- Small problem size (N)
  - 280 MB / node
- Power dominated by GbE switch
  - 40 W when idle, 100-150 W active

- 32 nodes
  - 4 x 8 boards (64 GFLOPS peak)
  - 1 GbE switch
- ... runs don’t complete due to boards overheating
  - Boards too close together
  - No space for airflow
Lessons learned so far

- Memory + interconnect dominates power consumption
  - Need a balanced system design
- Tuning scientific libraries takes time + effort
  - Compiling ATLAS on ARM Cortex-A9 took 1 month
- Linux on ARM needs tuning for HPC
  - CFS scheduler
  - softfp vs. hardfp
- DIY assembly of prototypes is harder than expected
  - 2 Person-Month just to press screws
- Even low-power devices need cooling
  - It’s the density that matters
## Validate the use of energy efficient CPU + compute accelerators
- ARM multicore processors
- Mobile Nvidia GPU accelerators

## Perform scalability tests to high number of compute nodes
- Higher core count required when using low-power components
- Evaluate impact of limited memory and bandwidth on low-end solutions

## Enable early application and runtime system development on ARM + GPU

### Tegra3 + GeForce 520MX:
- 4x Corext-A9 @ 1.5 GHz
- 48 CUDA cores @ 900 MHz
- 148 GFLOPS ~ 18 Watts
- ~ 8 GFLOPS / W

### Rack:
- 32x Board container
- 256x Q7 carrier boards
  - 1024x ARM Corext-A9 Cores
  - 256x GT520MX GPU
- 8x 48-port 1GbE LBA switches
- 38 TFLOPS ~ 5 Kwatt
  - 7.5 GFLOPS / W
  - 50% efficiency
  - 3.7 GFLOPS / W
What comes next?

European Exascale approach using embedded power-efficient technology

1. To deploy a **prototype HPC system** based on **currently available** energy-efficient embedded technology
2. To design a next-generation HPC system and **new embedded technologies targeting HPC systems** that would overcome most of the limitations encountered in the prototype system
3. To port and optimise a small number of **representative exascale applications** capable of exploiting this new generation of HPC systems

http://www.montblanc-project.eu
Integrate energy-efficient building blocks

- Integrated system design built from mobile / embedded components

- ARM multicore processors
  - Nvidia Tegra / Denver, Calxeda, Marvell Armada, ST-Ericsson Nova A9600, TI OMAP 5, ...

- Low-power memories

- Mobile accelerators
  - Mobile GPU
    - Nvidia GT 500M, ...
  - Embedded GPU
    - Nvidia Tegra, ARM Mali T604

- Low power 10 GbE switches
  - Gnodal GS 256

- Tier-0 system integration experience
  - BullX systems in the Top10
Hybrid MPI + OmpSs programming model

- Hide complexity from programmer
- Runtime system maps task graph to architecture
- Automatically performs optimizations
  - Many-core + accelerator exploitation
  - Asynchronous communication
    - Overlap communication + computation
  - Asynchronous data transfers
    - Overlap data transfer + computation
  - Strong scaling
    - Sustain performance with lower memory size per core
  - Locality management
    - Optimize data movement
Trade off bandwidth for power in the interconnect

- Hybrid MPI + SMPSs Linpack on 512 processors
- 1/5th the interconnect bandwidth, only 10% performance impact
- Rely on slower, but more efficient network?
Energy-efficient prototype series @ BSC

- A very exciting roadmap ahead
- Lots of challenges, both hardware and software!