Building supercomputers from commodity embedded chips

Alex Ramirez
Barcelona Supercomputing Center
Technical Coordinator
The need for supercomputers

- Supercomputers have evolved into a basic research tool
  - Basic research
  - Industrial research
- Named by many as “the third pillar of science”
  - Theory
  - Practice
  - Computing
Supercomputer performance evolution

- 1000x performance improvement every 10-12 years
- 1 EFLOPS projected for 2019
# Power consumption of top supercomputers

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National University of Defense Technology</td>
<td>Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT</td>
<td>3,120,000</td>
<td>33,862.7</td>
<td>54,902.4</td>
<td>17,808</td>
</tr>
<tr>
<td>2</td>
<td>DOE/SC/Oak Ridge National Laboratory</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.</td>
<td>560,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LLNL</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
<tr>
<td>4</td>
<td>RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu</td>
<td>705,024</td>
<td>10,510.0</td>
<td>11,280.4</td>
<td>12,660</td>
</tr>
<tr>
<td>5</td>
<td>DOE/SC/Argonne National Laboratory</td>
<td>Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM</td>
<td>786,432</td>
<td>8,586.6</td>
<td>10,066.3</td>
<td>3,945</td>
</tr>
</tbody>
</table>

- Top supercomputers consume 4 to 18 MWatt of power
- Exponential performance progression at the cost of increased power consumption
Power defines performance

- **#1 Green500** expected to be <5 GFLOPS/Watt in 2014
- 1 EFLOPS would require >200MWatt
Providing power is not the problem

- Power defines performance, yes ...
- ...but it’s not that we can’t provide the power
Spending the power is not the problem

- Large Hadron Collider (LHC)
  - World’s largest high-energy particle collider
  - Built by CERN from 1998 to 2008
  - 7.5 billion ¤
    - The most expensive scientific instrument ever built

- 220-330 MWatts of power

- And nobody cares if it has no immediate impact into economy...or industrial use

- 15 PBytes of data / year
  - Will soon require a 1 EFLOPS supercomputer to process it
kWh cost for Medium size industries (ifice 2011)

- The power wall isn’t due to power availability
  - We could provide 500 MW to an HPC facility
- It’s the electricity bill at the end of the month!
  - 1 MWatt ~ 1 M EUR / year
The problem ...and the opportunity

- Europe represents ~40% of the HPC market
  - Yet, it does not have HPC technology of its own

- Nobody really knows how to build a sustainable EFLOOPS supercomputer
  - Power defines performance
  - Consensus about it being revolutionary, not evolutionary
  - Everyone starts from square zero, hence equal opportunities

- Europe is very strong in embedded computing
  - The most energy-efficient computing technology today
Mont-Blanc project goals

- To develop an European Exascale approach
- Leverage commodity and embedded power-efficient technology

Supported by EU FP7 with 16M € under two projects:

- Mont-Blanc: October 2011 – September 2014
  - 14.5 M € budget (8.1 M € EC contribution), 1095 Person-Month
  - 11.3 M € budget (8.0 M € EC contribution), 892 Person-Month
First, vector processors dominated HPC

- 1st Top500 list (June 1993) dominated by DLP architectures
  - Cray: vector, 41%
  - MasPar: SIMD, 11%
  - Convex/HP: vector, 5%
- Fujitsu *Wind Tunnel* is #1 1993-1996, with 170 GFLOPS
Then, commodity took over special purpose

- ASCI Red, Sandia
  - 1997, 1 TFLOPS
  - 9,298 cores @ 200 Mhz
  - Intel Pentium Pro
    - Upgraded to Pentium II Xeon, 1999, 3.1 TFLOPS

- ASCI White, LLNL
  - 2001, 7.3 TFLOPS
  - 8,192 proc. @ 375 Mhz,
  - IBM Power 3

Transition from Vector parallelism to Message-Passing Programming Models
The killer microprocessors

- Microprocessors killed the Vector supercomputers
  - They were not faster ...
  - ... but they were significantly cheaper and greener
- Need 10 microprocessors to achieve the performance of 1 Vector CPU
  - SIMD vs. MIMD programming paradigms
**The killer mobile processors™**

- Microprocessors killed the Vector supercomputers
  - They were not faster ...
  - ... but they were significantly cheaper and greener

- History may be about to repeat itself ...
  - Mobile processor are not faster ...
  - ...but they are significantly cheaper
**Mobile SoC vs Server processor**

**Performance**
- 5.2 GFLOPS
- 153 GFLOPS
- 32.3 GFLOPS

**Cost**
- 21$^2$
- 1500$^3$
- 21$ (?)$

---

1. 6.8 GFLOPS from CPU + 25.5 GFLOPS from embedded GPU
2. Leaked Tegra3 price from the Nexus 7 Bill of Materials
3. Non-discounted List Price for the 8-core Intel E5 Sandy Bridge
Tibidabo: The first ARM HPC multicore cluster

**Q7 Tegra 2**
- 2 x Cortex-A9 @ 1GHz
- 2 GFLOPS
- 5 Watts (?)
- 0.4 GFLOPS / W

**Q7 carrier board**
- 2 x Cortex-A9
- 2 GFLOPS
- 1 GbE + 100 MbE
- 7 Watts
- 0.3 GFLOPS / W

**1U Rackable blade**
- 8 nodes
- 16 GFLOPS
- 65 Watts
- 0.25 GFLOPS / W

**2 Racks**
- 32 blade containers
- 256 nodes
- 512 cores
- 9x 48-port 1GbE switch
- 512 GFLOPS
- 3.4 Kwatt
- 0.15 GFLOPS / W

- Proof of concept
  - It is possible to deploy a cluster of smartphone processors
- Enable software stack development
HPC System software stack on ARM

- Open source system software stack
  - Ubuntu Linux OS
  - GNU compilers
    - gcc, g++, gfortran
  - Scientific libraries
    - ATLAS, FFTW, HDF5,...
  - Slurm cluster management

- Runtime libraries
  - MPICH2, OpenMPI
  - **OmpSs toolchain**
  - Performance analysis tools
    - Paraver, Scalasca
  - Allinea DDT 3.1 debugger
    - Ported to ARM
Limitations of current mobile processors for HPC

- 32-bit memory controller
  - Even if ARM Cortex-A15 offers 40-bit address space
- No ECC protection in memory
  - Limited scalability, errors will appear beyond a certain number of nodes
- No standard server I/O interfaces
  - Do NOT provide native Ethernet or PCI Express
  - Provide USB 3.0 and SATA (required for tablets)
- No network protocol off-load engine
  - TCP/IP, OpenMX, USB protocol stacks run on the CPU
- Thermal package not designed for sustained full-power operation

- All these are implementation decisions, not unsolvable problems
  - Only need a business case to justify the cost of including the new features …such as the HPC and server markets
Get ready for the change, before it happens …

- Mobile processors have qualities that make them interesting for HPC
  - FP64 capability
  - Performance increasing rapidly + energy efficient
  - Embedded GPU accelerator
  - Large market, many providers, competition, low cost

- Current limitations are due to target market conditions
  - Not real technical challenges

- A whole set of ARM server chips is coming
  - Solving most of the limitations identified
Samsung Exynos 5 Dual Superphone SoC

- 32nm HKMG
- Dual-core ARM Cortex-A15 @ 1.7 GHz
- Quad-core ARM Mali T604
  - OpenCL 1.1
- Dual-channel DDR3
- USB 3.0 to 1 GbE bridge

- All in a low-power mobile socket
Exynos 5 Dual vs. Intel i7

<table>
<thead>
<tr>
<th></th>
<th>Performance</th>
<th>Energy per iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quad Cortex-A9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dual Cortex-A15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quad Mali-T604</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quad i7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mali-T604 results could still improve with better algorithms, runtime and compilers to come...
Mont-Blanc SoM

- CPU + GPU + DRAM + storage + network ... all in a compute card just 8.5x5.6 cm

- Exynos 5 Dual: 2x ARM Cortex-A15
- USB 3.0 to 1 GbE
- 4 GB of DDR3-1600
- uSD slot, up to 64 GB
Mont-Blanc server blade

- 15 node-cluster in a standard Bull B505 enclosure

15 compute cards
30 x ARM Cortex-A15
15 x ARM Mali-T604 GPU
120 GB DDR3

Cluster management
1 GbE crossbar switch
2 x 10 GbE links
Mont-Blanc server chassis

- 9 blades in a standard 7U BullX chassis
- Shared cooling, PSU, chassis management
Exynos 5 Octa (5450)

- Quad-core ARM Cortex-A15, for performance
- Quad-core ARM Cortex-A7, for energy efficiency
- Six-core ARM Mali-T628, for OpenCL accelerator
  - 50% more GPU cores than Exynos 5 Dual
  - 25% higher compute performance per core
- Higher CPU and DDR frequencies
Conclusions

• Computing at scale is key to industry and research
  • “To outcompute is to outcompete”

• Supercomputers have hit the power wall
  • Energy efficiency determines the performance to deploy

• Europe has a strong position in the most energy efficient computing technology … embedded computing

• Leverage on it to build a new class of sustainable computer

montblanc-project.eu
MontBlancEU
@MontBlanc_EU