Acceleration for energy efficient, cost effective HPC

Lennart Johnsson
University of Houston
Royal Institute of Technology
Outline

• Energy Efficiency – Why?
  – The big picture
    • Environment
    • Economics
  – HPC
    • PDC experience
    • Exa-scale

• Where does the power go?
  – PUE
  – In the system
  – Reuse high density
Outline (cont’d)

• Energy efficiency – How?
  – Simple highly energy efficient CPUs
  – Slow low power CPUs
  – Standard CPUs
  – Acceleration (GPUs, FPGAs)

• Energy Efficiency – Implications
  – Programming
  – Applications
Energy efficiency – Why?
The Environment

BEIJING | Sun Feb 28, 2010 5:34am EST (Reuters) - China said on Sunday it will spell out greenhouse gas emissions goals and monitoring rules for regions and sectors in its next five-year plan, with monitoring to show it is serious about curbing emissions. The Chinese government said in November it would reduce the amount of carbon dioxide, the main greenhouse gas from human activity, emitted to make each unit of national income by 40 to 45 percent by 2020, compared with 2005 levels.

Bloomberg News (7/20) reported, "China, the world's biggest polluter, may spend about 5 trillion yuan ($738 billion) in the next decade developing cleaner sources of energy to reduce emissions from burning oil and coal, a government official said."
The Environment

1000 Years of CO₂ and Global Temperature Change

- Global Temperature Change (deg F)
- CO₂ Concentration (ppm)

Year: 1000, 1200, 1400, 1600, 1800, 2000
The Environment

Variations of the Earth’s surface temperature: year 1000 to year 2100
Climate Change is not reversible

• Climate Change is not like acid rain or ozone destruction where environment will quickly return to normal once source of pollution is removed.

• GHG emissions will stay in the atmosphere for thousands of years and continue to accumulate.

• Planet will continue to warm up even if we drastically reduce emissions.

Weaver et al., GRL (2007)

All we hope to achieve is to slow down the rapid rate of climate change.

http://www.slideshare.net/bstarn/ottawa-u-deploying-5g-networks
Urgency of Action

• “We’re uncertain about the magnitude of climate change, which is inevitable, because we’re talking about reaching levels of carbon dioxide in the atmosphere not seen in millions of years.
• You might think that this uncertainty weakens the case for action, but it actually strengthens it.
• This risk of catastrophe, rather than the details of cost-benefit calculations, makes the most powerful case for strong climate policy.
• Current projections of global warming in the absence of action are just too close to the kinds of numbers associated with doomsday scenarios. It would be irresponsible — it’s tempting to say criminally irresponsible — not to step back from what could all too easily turn out to be the edge of a cliff.”

Nobel Laureate
Paul Krugman


http://www.slideshare.net/bstarn/ottawa-u-deploying-5g-networks
The Global ICT Carbon Footprint is Roughly the Same as the Aviation Industry Today

But ICT Emissions are Growing at 6% Annually!

ICT represent 8% of global electricity consumption
Projected to grow to as much as 20% of all electrical consumption in the US

Future Broadband- Internet alone is expected to consume 5% of all electricity

http://www.slideshare.net/bstarn/ottawa-u-deploying-5g-networks
The Environment

Rapid retreat of glacier

Current Ice Extent
09/16/2007

Sea Ice edge
Sep. 16, 2007

September median ice edge
1979-2000

1932: Boulder Ice Cave, Glacier National Park

Total extent = 4.1 million sq km

1988: Boulder Ice Cave, Glacier National Park
Sea level has increased about 3 mm/yr between 1993 and 2005

- Projected increase from 1990-2100 is anywhere from 0.09 – 0.88 meters.

1/3rd due to melting glaciers
2/3rd due expansion from warming oceans

Source: Trenberth, NCAR 2005

Ocean Acidification

Lower $pH = \text{MORE ACID}$

Historical and Projected $pH$ and Dissolved $CO_2$

Feely, Sabine and Fabry, 2006

- Since 1850, ocean $pH$ has decreased by about 0.1 unit (30% increase in acidity). (Royal Society 2006)
- At present rate of $CO_2$ emission, acidity predicted to increase by 0.4 units $pH$ (3-fold increase in $H^+$ ions) by 2100.
- Carbonate ion concentrations decrease.

Source: http://alaskaconservationsolutions.com/acs/images/stories/docs/AkCS_current.ppt
Economy

Worldwide Server Installed Base, New Server Spending, and Power and Cooling Expense

The US: Cooling and power cost surpassed hardware cost ~2008/2009 in many locations

Power per rack has increased from a few kW to 30 – 80+ kW

Estimate for 2007 and spring 2008 purchases: 4 yr cooling cost ~1.5 times cluster cost

Source: IDC, 2006
Where does the power go?
How Does a Data Center Use Power?

- 70% of a “typical” data center’s power goes to Power & Cooling
- Percentage varies with data centers
- HP is working across the full spectrum to raise data center energy efficiency
- Figures to the left are “typical” of existing/legacy data centers

PUE = Power Usage Effectiveness
PUE = Total Facility Power
IT Equipment Power

Source: The Green Grid, 2007, “Guidelines for Energy-Efficient Datacenters” (www.thegreengrid.org). Notes: for PUE “lower is better”; DCiE = Data Center Infrastructure Efficiency, DCiE = 1/PUE, for DCiE “higher is better”

AOL, Manassas Technology Center, (MTC),
92,000 sf. of raised floor.

Google Data Center PUE

\[ PUE = \frac{E_{US1} + E_{US2} + E_{TX} + E_{HV}}{E_{US2} + E_{Net1} - E_{CRAC} - E_{UPS} - E_{LV}} \]

- **EUS1** Energy consumption for type 1 unit substations feeding the cooling plant, lighting, and some network equipment
- **EUS2** Energy consumption for type 2 unit substations feeding servers, network, storage, and CRACs
- **ETX** Medium and high voltage transformer losses
- **EHV** High voltage cable losses
- **ELV** Low voltage cable losses
- **ECRAC** CRAC energy consumption
- **EUPS** Energy loss at UPSes which feed servers, network, and storage equipment
- **ENet1** Network room energy fed from type 1 unit substitution

Power fundamentals

**Processor**
- Modern processors being designed today (for 2010) dissipate about 200 pJ/op total.
  - This is $\sim 200\text{W/TF 2010}$
- In 2018 we might be able to drop this to 10 pJ/op
  - $\sim 10\text{W/TF 2018}$
- This is then **16 MW** for a sustained HPL Exaflops
- This does not include memory, interconnect, I/O, power delivery, cooling or anything else

**Memory**
- Cannot afford separate DRAM in an Exa-ops machine!
- Propose a MIP machine with Aggressive voltage scaling on 8nm
- Might get to 40 KW/PF – **60 MW** for sustained Exa-ops

**Power fundamentals**

**Interconnect**
- For short distances: still Cu
- Off Board: Si photonics
- Need ~ 0.1 B/Flop

**Assume (a miracle)**
- 5 mW/Gbit/sec

**~ 50 MW** for the interconnect!

**I/O**
- Optics is the only choice:
- 10-20 PetaBytes/sec
- ~ a few MW (a swag)

**Power and Cooling**

Still 30% of the total power budget in 2018!
Total power requirement in **2018: 120—200 MW**!

Peter Kogge – DARPA Exascale study

• Last 30 years:
  – “Gigascale” computing first in a single vector processor
  – “Terascale” computing first via several thousand microprocessors
  – “Petascale” computing first via several hundred thousand cores

• Commercial technology: to date
  – Always shrunk prior “XXX” scale to smaller form factor
  – Shrink, with speedup, enabled next “XXX” scale

• Space/Embedded computing has lagged far behind
  – Environment forced implementation constraints
  – Power budget limited both clock rate & parallelism

• “Exascale” now on horizon
  – But beginning to suffer similar constraints as space
  – And technologies to tackle exa challenges very relevant

  Especially Energy/Power

http://www.ll.mit.edu/HPEC/agendas/proc09/Day1/S1_0955_Kogge_presentation.ppt
High Energy Efficiency and High Density of great importance for HPC

In particular for Exascale!!
Zero Emission Data Center

IBM Research – Cool with hot water!

Source: T. Brunschwiler, B. Smith, E. Ruetsche and B. Michel, IBM Research, Zurich
Zero Emission Data Center

Source: T. Brunschwiler, B. Smith, E. Ruetsche and B. Michel, IBM Research, Zurich
First Prototype at IBM Rüschildkon

- Reduce cooling energy by tailored water cooling system
  - Cooling the chip with “hot” water (up to 60 °C)
  - Free cooling: no energy-intensive chillers needed

- Reuse waste heat for remote heating
  - The prototype reuses 75% of the energy for remote heating
  - Obtain recyclable heat (60 °C) for remote heating.
  - Best in a cold climate with dense population

- Prototype
  - Similar Power of CPU and main board for air / liquid 60 °C cooled version
  - Large fan power reduction
  - Liquid pump much more efficient and can vary flow at the rack level

Source: T. Brunschwiler, B. Smith, E. Ruetsche and B. Michel, IBM Research, Zurich
Heat Capacity of this much air

Heat Capacity of this much water

## Cooling Limit rough estimates +/-

<table>
<thead>
<tr>
<th>Cooling Architecture</th>
<th>Limits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open data center</td>
<td>4 kW</td>
<td>Least efficient</td>
</tr>
<tr>
<td>Hot aisle/cold aisle</td>
<td>8-12 kW</td>
<td>Requires specific layouts for higher end</td>
</tr>
<tr>
<td>Cold aisle containment</td>
<td>30 kW</td>
<td>Best for legacy DCs</td>
</tr>
<tr>
<td>Hot aisle containment or chimney cabinets</td>
<td>30 kW</td>
<td>Best for new construction</td>
</tr>
<tr>
<td>Liquid cooled rack enclosure</td>
<td>35 kW</td>
<td>Ideal for legacy DCs with limited airflow</td>
</tr>
<tr>
<td>Liquid rear doors</td>
<td>15-30 kW</td>
<td>Can make rack room-neutral</td>
</tr>
<tr>
<td>In row coolers</td>
<td>20-35 kW</td>
<td>Space implications but rack independent</td>
</tr>
<tr>
<td>Close coupled coolers</td>
<td>10-25 kW</td>
<td>Overhead has space advantages</td>
</tr>
<tr>
<td>CPU liquid cooling</td>
<td>80 kW</td>
<td>Caution! must we cool the rest with air?</td>
</tr>
<tr>
<td>Integrated liquid cooling</td>
<td>100 kW</td>
<td>Benefits in Interconnect</td>
</tr>
</tbody>
</table>

Energy efficiency – How?

- Simple highly energy efficient CPUs
- Slow low power CPUs
- Standard CPUs
- Acceleration/Specialized (GPUs, FPGAs)
Processor sizes

How Small is Small

- Power5 (server)
  - 389mm\(^2\)
  - 120W@1900MHz
- Intel Core2 sc (laptop)
  - 130mm\(^2\)
  - 15W@1000MHz
- ARM Cortex A8 (toaster oven)
  - 5mm\(^2\)
  - 0.8W@800MHz
- Tensilica DP (cell phones)
  - 0.8mm\(^2\)
  - 0.09W@600MHz
- Tensilica Xtensa (Cisco Rtr)
  - 0.32mm\(^2\) for 3!
  - 0.05W@600MHz

Each core operates at 1/3 to 1/10th efficiency of largest chip, but you can pack 100x more cores onto a chip and consume 1/20 the power

# A comparison of some accelerators

<table>
<thead>
<tr>
<th></th>
<th>Cell BE</th>
<th>Nvidia G80</th>
<th>ClearSpeed CSX600</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>32-bit FP</strong></td>
<td>200+ GFLOPS</td>
<td>360+ GFLOPS</td>
<td>25+ GFLOPS</td>
</tr>
<tr>
<td><strong>64-bit FP</strong></td>
<td>20+ GFLOPS 100+GF</td>
<td>78 GF</td>
<td>25+ GFLOPS 96 GF</td>
</tr>
<tr>
<td><strong>Clock frequency</strong></td>
<td>3.2 GHz</td>
<td>575 MHz</td>
<td>210 MHz</td>
</tr>
<tr>
<td><strong>Transistors/chip</strong></td>
<td>~ 241M</td>
<td>~ 681M</td>
<td>~ 128M</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>~ 110 Watts</td>
<td>~ 145 W (for GeForce 8800 GTX board)</td>
<td>~ 10W 25W board</td>
</tr>
<tr>
<td><strong>nVidia GF8800GTX</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>nVidia Tesla S870 1U 1.3 TF</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clearspeed PCI-X board</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why this performance difference?

- Standard processors are optimized for a 40 year old model
  - Efficiency as defined in 1964
  - Heavy weight threads
  - Complex control
  - Big overhead per operation
  - Parallelism added as an afterthought
  - A large fraction of the silicon devoted to
    - Address translation
    - Instruction reordering
    - Register renaming
    - Cache hierarchy

FPU/SSE
Integer ALU
3Dnow

Itanium2 Madison
~4% devoted to ALU and FPU
L3 Cache
Green Flash Strawman System Design

Three different approaches examined (in 2008 technology)
Computation $0.015 \times 0.02 \times 100L$: 10 PFlops sustained, ~200 PFlops peak

- **AMD Opteron**: Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market
- **BlueGene**: Generic embedded processor core and customize system-on-chip (SoC) to improve power efficiency for scientific applications
- **Tensilica XTensa**: Customized embedded CPU w/SoC provides further power efficiency benefits but maintains programmability

<table>
<thead>
<tr>
<th>Processor</th>
<th>Clock</th>
<th>Peak/Core (Gflops)</th>
<th>Cores/Socket</th>
<th>Sockets</th>
<th>Cores</th>
<th>Power 2008</th>
<th>Cost 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD Opteron</td>
<td>2.8GHz</td>
<td>5.6</td>
<td>2</td>
<td>890K</td>
<td>1.7M</td>
<td>179 MW</td>
<td>$1B+</td>
</tr>
<tr>
<td>IBM BG/P</td>
<td>850MHz</td>
<td>3.4</td>
<td>4</td>
<td>740K</td>
<td>3.0M</td>
<td>20 MW</td>
<td>$1B+</td>
</tr>
<tr>
<td>Green Flash / Tensilica XTensa</td>
<td>650MHz</td>
<td>2.7</td>
<td>32</td>
<td>120K</td>
<td>4.0M</td>
<td>3 MW</td>
<td>$75M</td>
</tr>
</tbody>
</table>

PRACE Technology Prototypes
What is PRACE?

• PRACE – Partnership for Advanced Computing in Europe
  A legal entity formed in April 2010 to provide a persistent pan-European Research Infrastructure for High-End Computing and associated services. Preceded by a 2+ years preparatory project. Member countries currently are
  
  • Austria
  • Bulgaria
  • Cyprus
  • Czech Republic
  • Finland
  • France
  • Germany
  • Greece
  • Ireland
  • Italy
  • Netherlands
  • Norway
  • Poland
  • Portugal
  • Serbia
  • Spain
  • Sweden
  • Switzerland
  • Turkey
  • United Kingdom
ESFRI – Estimated costs

- Unlike other European Research Infrastructures:
  - Tier-0 resources have to be renewed every 2-3 years
  - Construction cost 200 – 400 Mio. € every 2-3 years
  - Annual running cost 100 – 200 Mio. €
  - Additional effort needed for software

- A truly European challenge – also in terms of funding

- PRACE – The Partnership for Advanced Computing in Europe
  - An Initiative created to implement the ESFRI vision of a European HPC service

(ESFRI = European Strategy Forum for Research Infrastructures)
## PRACE Technology Prototypes

<table>
<thead>
<tr>
<th>Prototypes</th>
<th>Installation Site</th>
<th>Targeted Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>eQPACE</td>
<td>JSC, Germany</td>
<td>Interconnects, energy efficiency and density</td>
</tr>
<tr>
<td>RapidMind</td>
<td>BAdW-LRZ, Germany</td>
<td>Programming models for hybrid systems</td>
</tr>
<tr>
<td>LRZ-CINES 1</td>
<td>CINES, France, BAdW-LRZ, Germany</td>
<td>Intel Nehalem-EP, ClearSpeed and QDR Infiniband</td>
</tr>
<tr>
<td>LRZ-CINES 2</td>
<td>BAdW-LRZ, Germany</td>
<td>Intel Nehalem-EX, Numalink5, Intel Larrabee</td>
</tr>
<tr>
<td>Hybrid Technology</td>
<td>CEA, France</td>
<td>GPGPU, HMPP</td>
</tr>
<tr>
<td>Maxwell FPGA</td>
<td>EPCC, UK</td>
<td>FPGA, energy efficiency and programing</td>
</tr>
<tr>
<td>PGAS Compiler</td>
<td>CSCS, Switzerland</td>
<td>PGAS programming model</td>
</tr>
<tr>
<td>ClearSpeed</td>
<td>NCF, Netherlands</td>
<td>ClearSpeed</td>
</tr>
<tr>
<td>XC4-IO</td>
<td>CINECA, Italy</td>
<td>I/O and File System perf/, SSD for metadata,</td>
</tr>
<tr>
<td>Accelerator efficiency</td>
<td>PSNC, Poland, SFTC, UK</td>
<td>Power consumption, porting of applications</td>
</tr>
<tr>
<td>PGAS Programming</td>
<td>CSC, Finland</td>
<td>Performance of UPC and CAF</td>
</tr>
<tr>
<td>Parallel GPU</td>
<td>CSC, Finland</td>
<td>Parallelizing CUDA, porting CUDA to OpenCL</td>
</tr>
<tr>
<td>SNIC-KTH</td>
<td>KTH, Sweden</td>
<td>Energy efficient computing</td>
</tr>
</tbody>
</table>
eQPACE (extended QCD PARallel computing on Cell)

- Cell processor PowerXCell 8i
- eQPACE FPGA network processor (extension of QPACE)
- eQPACE board
- eQPACE with frontend at JSC
- **eQPACE**
  - nodes: IBM PowerXCell 8i processor with a custom FPGA-based network processor
  - 32 nodes per backplane with a peak DP performance of $32 \times 102.6 = 3.3$ TF
  - 8 backplanes in a rack, 26.3 TF
  - 4 racks

- **Communications Network**
  - 3D-torus with one dimension within the backplane arranged as $1 \times 4 \times 8$
  - nearest-neighbour communication between Local Stores of SPEs of adjacent nodes
  - custom low latency protocol using 10 GigE at the physical layer yielding 1 GB/s in each direction for each link

- **I/O Network**
  - The eQPACE network processor is connected to a Gbit-Ethernet transceiver with all ports in a rack connect to a standard Ethernet switches mounted in the rack
  - Tree network for evaluation of global conditions and synchronisation
  - Power consumption: Node 125W, Backplane with nodes 4kW. Rack 32kW plus switches
  - Closed node card housing connected to a liquid-cooled cold-plate
  - Front-end: An IBM e1350 cluster consisting of login and master nodes and six nodes for a parallel Lustre file system with two meta data and four object storage servers. The disk storage amounts to five enclosures, each with $14 \times 72$ GB disks.
• One *PowerPC Processor Element (PPE)*

• **8 Synergistic Processor Elements (SPE).**
  
  • Each of the SPEs runs a single thread executing two instructions/cycle performing up to 8 single-precision (SP) or 4 double precision (DP) floating point (FP) operations. The peak SP (DP) performance of all 8 SPEs on a single processor is 204.8 (102.4) GF at 3.2 GHz.

  • Each SPE has 256 KB on-chip Local Store (LS) accessed by DMA or by local load/store operations from/to 128 general-purpose 128-bit registers.

• On-Chip memory controller supporting 25.6 GB/s and a configurable I/O interface (Rambus FlexIO) supporting a coherent as well as a non-coherent protocol with a total bidirectional bandwidth of up to 25.6 GB/s.

• All units of the processor are connected to the coherent *Element Interconnect Bus (EIB)* by DMA controllers.

• In eQPACE, the I/O interface is used to interconnect the PowerXCell 8i processor with the network processor implemented on a Xilinx V5-LX110T FPGA

• In eQPACE, unlike other Cell-based parallel machines, data is moved via the EIB directly to or from the I/O interface enabling data to be moved directly from the LS of any SPE on one node to the LS of any SPE of one of the 6 neighboring nodes. Data do not have to be routed through main memory (to avoid the bottleneck of the memory interface) or the PPE.
LRZ – CINES 1

- 32 SGI XE dual socket blades
- 64 Intel Nehalem-EP 2.53 GHz CPUs (256 cores)
- 4 GB per core
- QDR Infiniband

Estimated peak performance: 2.59 TFlop/s

Accelerator
32 e710 ClearSpeed-Petapath cards
One per host blade
96 GF Double-Precision/card
4 GFlop/s / Watt

Estimated peak performance: 3 TFlop/s.

The total peak performance: 5.59 TFlop/s.
LRZ – CINES 2

- 48 SGI ICE dual socket blades with Intel Nehalem-EP 2.53 GHz 4 core CPUs and 24 GB memory
- One 4-Socket Nehalem-EX Intel “WhiteBox” (α-Prototype: "Sunrise Ridge")
- 16 “P1-Level” SGI “UltraViolet” prototype dual socket blades with 2.0 GHz Intel Nehalem-EX 8-core CPUs and 32 GB memory
- 4× DDR Infiniband switch and cables (44 IB ports plus 2 x 10GigE ports)
- SGI UV extension unit with
  - 1 Intel Larrabee accelerator
  - 4 ClearSpeed-Petapath Advance e710 accelerator boards
- Two racks with system administration infrastructure (head node)
- SLES10 SP2 operating system, SGI “Tempo” and SGI “ProPack” management software;
- Storage: the Lustre file systems of the BAdW-LRZ linux cluster (130 TB /scratch and 50 TB /project).

The double precision peak performance of the prototype is 6.5 TF.
LRZ – CINES 2

Intel MIC

32 cores, shared L2 cache with a local 256kB subset
L1: 32kB for instr.
32kB for data
NCF

• Accelerator units (AU)
  • Four e740 containing four Clearspeed CSX700 processors
  • 12 e780 units containing eight Clearspeed CSX700 processors
    (total 112 CSX700 processors)
  • Each CSX700 contains two 250 MHz 96 Multi-Threaded Array Processors,
    MTAPs, (arrays of SIMD processors) that can perform two DP ops/sec. Thus,
    a CSX700 processor has a peak performance of 96 GF DP (2 x 0.25 x 96 x 2).
    The power consumption is less than 25W.
• Hosts: Eight HP DL170h G6 dual socket nodes with Intel Nehalem-EP 2.67 GHz
  four core CPUs, 24 GB of DDR3 memory and dual disks. Each host connects to
  two AUs by PCI Express Gen. 2 16× (8 GB/s).
• Head node: HP DL160 G6.
• Interconnect between hosts: 4× DDR Infiniband
• Each e780 connects internally to its 8 CSX700 processors through a PLX 8648
  PCIe switch. The bandwidth to each individual processor is 1 GB/s.
• Each e740 connects internally to each CSX700 processor at 2 GB/s.

The peak DP performance of an e780 node is 768 GF and 384 GF for an e740 node.
### NCF

**Clearspeed-Petapath prototype**

<table>
<thead>
<tr>
<th>Rack 1</th>
<th>Rack 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GigE Switch</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e740</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e740</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e740</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e740</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>HP DL170</td>
<td>HP DL170</td>
</tr>
<tr>
<td>HP DL170</td>
<td>HP DL170</td>
</tr>
<tr>
<td><strong>Head Node (HP DL160)</strong></td>
<td><strong>KVM</strong></td>
</tr>
<tr>
<td>IB Switch</td>
<td>Monitor/Keyboard</td>
</tr>
<tr>
<td>HP DL170</td>
<td>HP DL170</td>
</tr>
<tr>
<td>HP DL170</td>
<td>HP DL170</td>
</tr>
<tr>
<td>Feynman e780</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e780</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e780</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e780</td>
<td>Feynman e780</td>
</tr>
<tr>
<td>Feynman e780</td>
<td>Feynman e780</td>
</tr>
</tbody>
</table>
• **Accelerator Units**
  • Eight Tesla S1070 unit
    each Tesla unit has two PCI-express 16x links, each
    connecting to two C1060 graphics boards (4 total)
  • Each C1060 has 30 1.3 GHz multiprocessors and 4 GB of
    memory.
  • Each multiprocessor has 8 single precision (SP) units and 1
double precision (DP) unit.

A C1060 have 240 SP and 30 DP units. A S1070 unit has 960
SP and 120 DP units with a peak performance of 2.496 TF SP
and 0.312 TF DP.

• **Hosts:** Four BULL R422 servers each with two Intel Harpertown
  CPUs and 16 GB of RAM memory.
  • Each host is connected to two Tesla S1070 units.
• **Host interconnect:** DDR Infiniband
• **Login node with four Nehalem-EP CPUs.**
Accelerator Units
- Eight nVIDIA Tesla S1070 servers
- Each S1070 has four C1060 1.3 GHz GPUs and 4 GB of RAM

Host:
- Eight Supermicro 825TQ-R700 LPB servers with motherboard X8Dai
- Each server has two Intel Nehalem-EP E5540 2.53 GHz (max TDP 80 W) CPUs and six 4GB DDR3 DIMMs
  SMT disabled

Network: Infiniband Voltaire HCA410-4EX, on-board dual gigabit Ethernet,

OS and software: Scientific Linux 5.3, gcc 4.1, icc 11.0, pgcc 9.0, CUDA 2.2.
Accelerator Units
• 32 Alpha Data Ltd cards using Xilinx Virtex-4 FPGAs
• 32 Nallatech Ltd cards using Loinx Virtex-4 FPGAs

Host: 32 node IBM HS21 BladeCenter system,
• each node has a single-core 2.8 GHz Intel CPU
• each node connects to two accelerator cards via a PCI-X bus

Interconnect: RocketIO connections on the FPGAs are used to create a nearest neighbour two-dimensional mesh interconnect between the FPGAs. This allows communication between FPGAs without using the (relatively) slow PCI- X bus.
• New 4-socket blade with 4 DIMMs per socket supporting PCI-Express Gen 2 x16
• Four 6-core 2.1 GHz 55W ADP AMD Istanbul CPUs, 32GB/node
• 10-blade in a 7U chassis with 36-port QDR IB switch, new efficient power supplies.
• 2TF/chassis, 12 TF/rack, 30 kW (6 x 4.8)
• 180 nodes, 4320 cores, full bisection QDR IB interconnect
SNIC/KTH

Network:
- QDR Infiniband
- 2-level Fat-Tree
- Leaf level 36-port switches built into chassis
- Five external 36-port switches
# Density - Examples

<table>
<thead>
<tr>
<th>Sockets/rack</th>
<th>Cores/rack</th>
<th>GF/core</th>
<th>TF/rack</th>
<th>kW/rack</th>
<th>TF/m²</th>
<th>kW/m²</th>
<th>TF/kW</th>
<th>Linpack TF/kW</th>
<th>Linpack Eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG/P</td>
<td>2048</td>
<td>4096</td>
<td>3.4</td>
<td>13.9</td>
<td>40</td>
<td>20.6</td>
<td>59</td>
<td>0.35</td>
<td>0.36 – 0.37</td>
</tr>
<tr>
<td>HP 2x blades (HTN, 80W)</td>
<td>256</td>
<td>1024</td>
<td>12 (3GHz)</td>
<td>12.3 (3GHz)</td>
<td>45</td>
<td>19.1</td>
<td>70</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>SGI Molecule</td>
<td>192</td>
<td>384</td>
<td>1.6</td>
<td>0.6</td>
<td>2</td>
<td>0.9</td>
<td>3</td>
<td>0.3</td>
<td>0.16</td>
</tr>
<tr>
<td>SiCortex</td>
<td>972</td>
<td>5832</td>
<td>1.4</td>
<td>8.2</td>
<td>22</td>
<td>2.2</td>
<td>8.5</td>
<td>0.37</td>
<td>0.22</td>
</tr>
<tr>
<td>SiCortex 2H09</td>
<td>972</td>
<td>11664</td>
<td>2.8</td>
<td>32.7</td>
<td>30</td>
<td>12.5</td>
<td>11.6</td>
<td>1.09</td>
<td>0.63</td>
</tr>
<tr>
<td>Supermicro PRACE prop</td>
<td>240</td>
<td>1440</td>
<td>9.6</td>
<td>13.8</td>
<td>32</td>
<td>19.2 – 21.5</td>
<td>44.4 – 49.8</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>Twin servers (quad core)</td>
<td>160</td>
<td>640</td>
<td>12 (3GHz)</td>
<td>7.7 (3GHz)</td>
<td>22</td>
<td>12</td>
<td>34.3</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>

- kW/rack: estimated or nominal claimed, not measured peak
- TF/m²: HP 2x 220c and Twin blades assuming 0.6x1.07m² racks
- kW/m²: not including cooling and service areas
- Linpack TF/kW: BG/P from Top500 Nov 2008, SiCortex from company info, Twin server from SGI ICE from Top500 Nov 2008
- SGI Molecule: based on PRACE prototype offer (not concept presented at SC08)
## PRACE Prototype – Data movement

<table>
<thead>
<tr>
<th>Node (memory)</th>
<th>GF</th>
<th>GB</th>
<th>GB/core</th>
<th>GB/s</th>
<th>GF/s / GB/s</th>
<th>Link BW GB/s</th>
<th>TF/s / TB/s chassis</th>
<th>TF/s / TB/s C - C</th>
<th>TF/s / TB/s R - R</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG/P</td>
<td>13.6</td>
<td>2</td>
<td>0.5</td>
<td>13.6</td>
<td>1</td>
<td>0.425</td>
<td>32</td>
<td>43</td>
<td>51</td>
</tr>
<tr>
<td>HP 2x blades (HTN)</td>
<td>96</td>
<td>16</td>
<td>2</td>
<td>5.3</td>
<td>18</td>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGI Molecule</td>
<td>3.2</td>
<td>2</td>
<td>1</td>
<td>4.2</td>
<td>0.76</td>
<td>0.125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SiCortex</td>
<td>8.4</td>
<td>8</td>
<td>1.25</td>
<td>2.1</td>
<td>3.8</td>
<td>1.6</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SiCortex 2H09</td>
<td>33.6</td>
<td>12</td>
<td>1</td>
<td>4.3?</td>
<td>7.7?</td>
<td>3.5?</td>
<td></td>
<td></td>
<td>121</td>
</tr>
<tr>
<td>Supermicro PRACE</td>
<td>230.4</td>
<td>32</td>
<td>1.33</td>
<td>8*6.4</td>
<td>4.5</td>
<td>5</td>
<td>46</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Twin servers (quad core)</td>
<td>96</td>
<td>24</td>
<td>3</td>
<td>5.3?</td>
<td>18?</td>
<td>2.5</td>
<td>38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HP 2x 220c: 4 DIMM slots/node, table assumes 4GB DIMMs  
SiCortex: 2 DIMMs/socket, GB/s assume 533 MHz DDR2 (533/800 MHz DDR2 conflicting info from vendor)  
2H09: 3 DIMM slots/node, table assumes 4GB DDR3 DIMMs @1066MHz, no vendor response  
Supermicro: 4x4 DIMM slots/node  
Twin servers: 3 DIMM slots/node assumed (Supermicro), 4GB DIMMs
### SNIC/KTH

<table>
<thead>
<tr>
<th>Component</th>
<th>Power (W)</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-core 2.1 GHz CPUs (55W ADP)</td>
<td>2,880</td>
<td>56.8</td>
</tr>
<tr>
<td>Memory</td>
<td>800</td>
<td>15.8</td>
</tr>
<tr>
<td>PS</td>
<td>355</td>
<td>7.0</td>
</tr>
<tr>
<td>Fans</td>
<td>350</td>
<td>6.9</td>
</tr>
<tr>
<td>Motherboards</td>
<td>300</td>
<td>5.9</td>
</tr>
<tr>
<td>HT3 Links</td>
<td>120</td>
<td>2.4</td>
</tr>
<tr>
<td>IB HCAs</td>
<td>100</td>
<td>2.0</td>
</tr>
<tr>
<td>IB Switch</td>
<td>100</td>
<td>2.0</td>
</tr>
<tr>
<td>GigE Switch</td>
<td>40</td>
<td>0.8</td>
</tr>
<tr>
<td>CMM</td>
<td>20</td>
<td>0.4</td>
</tr>
<tr>
<td>Total</td>
<td>5,065</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Chassis design estimates**

- HPL observed: Max 4,647  Avg 4,625 W  Stream observed: 3,620 W
Memory study

- Elpida, Hynix, Micron, Samsung power consumption for DIMMs estimated using public tools and published chip specs
- Measurements carried out with Elpida, Hynix and Samsung DIMMs (on “old” motherboard and chipset, Istanbul 75W ADP CPUs)
Elpida and Samsung relative HPL performance

Elpida/Samsung performance on Supermicro 4-socket blade
Elpida and Samsung relative HPL performance/W

Elpida/Samsung HPL performance/W on Supermicro 4-socket blade
Memory selection

• Elpida vs Hynix on Phase-II motherboards
  – Hynix 97.6% power consumption of Elpida for HPL
  – Hynix 99.7% of Elpida HPL performance
  – Hynix 107.9% of Elpida Stream performance

Hynix selected
SNIC/KTH prototype

- 46U, 19” std depth (0.6mx1.07m) (42U for chassis + switches for fat tree: 3 racks total)
- 1440 cores, 5.7 sockets/U
- ~28 kW (HPL), 43.6 kW/m² (HPL) (BG/P 45.6 kW/m² est.)
- 12.1 TF TPP, 18.8 TF/m² (BG/P 20.6 TF/m²)
- Air cooled
Evaluation
Evaluation

- Performance
- Energy efficiency
- Density
- Programmability
- Portability
Evaluation

What do people do on high-end systems?

We (PRACE) decided to find out: Survey
PRACE Application Survey Results

How to compare/weight usage across different systems?

Linpack Equivalent Flops (LEFs) defined as

\[(\text{percentage of system utilization (time) relative to available time}) \times (\text{platform Linpack } R_{\text{max}})\]

was used

For simplicity applications with similar computational and data access patterns grouped

- Astronomy and cosmology
- Computational chemistry
- Computational engineering
- Computational fluid dynamics
- Condensed matter physics
- Earth and climate science
- Life science
- Particle physics
- Plasma physics
- Other
PRACE Application Survey Results

How to characterize codes?

The PRACE survey used “the 7 dwarves”, i.e. algorithm types that have similarity in computation and data movement; see Asanovic, et. al. 2006. The Landscape of Parallel Computing Research: A View from Berkeley. Technical Report No. UCB/EECS-2006-183. http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html

How to characterize codes using more than one dwarf, and used in more than one scientific category?

The application’s weighting was divided equally over each scientific area and “dwarf” used. For example, a code that was used in computational chemistry and condensed matter physics, and that used both spectral and map reduce methods, and had a weighting of 12 Tflop/s on a system, would contribute 3 Tflop/s to each of the four categories (scientific area x dwarf).
PRACE application code survey
100+ responses fall 2008/winter 2009

- 69 applications across 24 systems (PRACE partner systems with $R_{\text{max}}$ of at least 10TF)
- 47 application codes only on one system
- 16 applications on two systems
- 6 applications on three or more systems
  - 2 on three systems (Dalton ~4% of LEFs, Gromacs ~1% of LEFs)
  - 1 on four systems (NAMD ~4% of LEFs)
  - 2 on five systems (CPMD ~4% of LEFs, AVBP ~0% of LEFs)
  - 1 on nine systems (VASP ~15% of LEFs)
<table>
<thead>
<tr>
<th>Systems</th>
<th>%</th>
<th>$R_{\text{peak}}$</th>
<th>%</th>
<th>$R_{\text{max}}$</th>
<th>%</th>
<th>Cores</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEC</td>
<td>1</td>
<td>4.2</td>
<td>1.0</td>
<td>8,923</td>
<td>1.3</td>
<td>576</td>
<td>0.3</td>
</tr>
<tr>
<td>MPP</td>
<td>7</td>
<td>29.2</td>
<td>46.0</td>
<td>335,491</td>
<td>49.7</td>
<td>108,248</td>
<td>63.9</td>
</tr>
<tr>
<td>FNC</td>
<td>6</td>
<td>25.0</td>
<td>12.0</td>
<td>94,118</td>
<td>13.9</td>
<td>16,928</td>
<td>10.0</td>
</tr>
<tr>
<td>TNC</td>
<td>10</td>
<td>41.7</td>
<td>41.1</td>
<td>236,882</td>
<td>35.1</td>
<td>43,770</td>
<td>25.8</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>100.0</td>
<td>100.0</td>
<td>675,414</td>
<td>100.0</td>
<td>169,522</td>
<td>100.0</td>
</tr>
</tbody>
</table>

VEC = Vector systems
MPP = Massively Parallel Processors (BlueGene and Cray XT)
FNC = Fat Node Cluster ("big" SMP nodes)
TNC = Thin Node Cluster

PRACE EU Deliverable D6.1
Language use by PRACE surveyed application codes

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran90</td>
<td>50</td>
</tr>
<tr>
<td>C90</td>
<td>22</td>
</tr>
<tr>
<td>Fortran77</td>
<td>15</td>
</tr>
<tr>
<td>C++</td>
<td>10</td>
</tr>
<tr>
<td>C99</td>
<td>7</td>
</tr>
<tr>
<td>Python</td>
<td>3</td>
</tr>
<tr>
<td>Perl</td>
<td>2</td>
</tr>
<tr>
<td>Mathematica</td>
<td>1</td>
</tr>
</tbody>
</table>

About 50% use more than one base language

16 out of the 69 application codes combine Fortran with C or C++
Application parallelization techniques

- 1 is sequential (BLAST)
- 1 code uses OpenMP only (Gaussian)
- 67 codes use MPI
  - 45 codes use MPI only, one having an MPI-2 version
  - 6 codes have one MPI version, one OpenMP version
  - 3 codes have one MPI version, one SHMEM version
  - 10 codes have hybrid MPI/OpenMP versions
  - 2 codes have hybrid MPI/SHMEM versions
  - 1 code has hybrid MPI/Posix threads
PRACE survey aggregated LEFs by scientific area

- Particle Physics 23.5
- Computational Chemistry 22.1
- Condensed Matter Physics 14.2
- CFD 8.6
- Earth & Climate 7.8
- Astronomy & Cosmology 5.8
- Life Sciences 5.3
- Other 5.8
- Computational Engineering 3.7
- Plasma Physics 3.3

PRACE EU Deliverable D6.1
PRACE survey applications by Dwarfs based on LEFs

- Map reduce methods: 45.1%
- Spectral methods: 18.4%
- Dense linear algebra: 14.4%
- Structured grids: 9.0%
- Particle methods: 7.2%
- Unstructured grids: 2.4%
- Sparse linear algebra: 3.4%

PRACE EU Deliverable D6.1
### PRACE survey applications characteristics in TF LEFs

<table>
<thead>
<tr>
<th>Area/Dwarf</th>
<th>Dense linear algebra</th>
<th>Spectral methods</th>
<th>Structured grids</th>
<th>Sparse linear algebra</th>
<th>Particle methods</th>
<th>Unstructured grids</th>
<th>Map reduce methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy and Cosmology</td>
<td></td>
<td>0.62</td>
<td>4.91</td>
<td>3.59</td>
<td>5.98</td>
<td>2.99</td>
<td>0</td>
</tr>
<tr>
<td>Computational Chemistry</td>
<td>15.35</td>
<td>26.09</td>
<td>1.80</td>
<td>3.45</td>
<td>7.49</td>
<td>0.53</td>
<td>12.98</td>
</tr>
<tr>
<td>Computational Engineering</td>
<td></td>
<td>0</td>
<td>0.53</td>
<td>0.53</td>
<td>0</td>
<td>0.53</td>
<td>2.8</td>
</tr>
<tr>
<td>CFD</td>
<td></td>
<td>1.70</td>
<td>7.37</td>
<td>3.05</td>
<td>0.32</td>
<td>3.00</td>
<td>0</td>
</tr>
<tr>
<td>Condensed Matter Physics</td>
<td>9.10</td>
<td>15.07</td>
<td>1.62</td>
<td>0.73</td>
<td>1.76</td>
<td>0.28</td>
<td>5.70</td>
</tr>
<tr>
<td>Earth and Climate Science</td>
<td></td>
<td>2.03</td>
<td>5.83</td>
<td>1.33</td>
<td>0</td>
<td>0.26</td>
<td>0</td>
</tr>
<tr>
<td>Life Science</td>
<td></td>
<td>4.72</td>
<td>0.94</td>
<td>0.13</td>
<td>0.94</td>
<td>0.28</td>
<td>3.46</td>
</tr>
<tr>
<td>Particle Physics</td>
<td>12.50</td>
<td>0</td>
<td>4.59</td>
<td>0.92</td>
<td>0.10</td>
<td>0</td>
<td>89.27</td>
</tr>
<tr>
<td>Plasma Physics</td>
<td></td>
<td>0</td>
<td>1.33</td>
<td>1.33</td>
<td>3.55</td>
<td>0.42</td>
<td>0.63</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
PRACE Benchmark Suite

- QCD synthetic benchmark (particle physics)
- VASP (comp. chemistry, condensed matter)
- NAMD (comp. chemistry, life sciences)
- CPMD (comp. chemistry, condensed matter)
- Code_Saturne (comp. fluid dynamics)
- Gadget (astronomy, cosmology)
- Torb (plasma physics)
- Echam5 (atmospheric science)
- Nemo (earth and climate sciences)
PRACE Benchmarks, possible extensions

- AVBP (CFD)
- CP2k (comp. chemistry)
- Gromacs (comp. chemistry)
- Helium (Helium atom modeling)
- SMMP (life sciences)
- Tripoli4 (comp. engineering)
- PEPC (plasma physics)
- Ramses (astronomy and cosmology)
- Cactus (astronomy and cosmology)
- N3D (CFD)
Technology prototypes - Simplified assessment

• EuroBench covering some of the 7 dwarves
  – Dense matrix multiplication (mod2am)
  – Sparse matrix-vector multiplication (mod2as)
  – FFT (mod2f)
  – Random number generation (mod2h)
• HPL
• Some Stream results
  – Copy (a(i) = b(i))
  – Scale (a(i) = q*b(i))
  – Add (a(i) = b(i) + c(i))
  – Triad (a(i) = b(i) + q*c(i))
• Some GUPS
• Some applications on some prototypes
  – Host institution selected application of particular interest
Reference platform

Of the systems on the November 2009 Top500 list 79% were based on Intel EM64T processor technology.

For this reason PRACE technology prototype results were compared to results obtained on a standard dual-socket Intel Nehalem-EP 2.53 GHz processor platform (E5540), whenever possible. This CPU has theoretical peak performance of 10.12 GF per core and 80.96 GF per dual socket server.
Reference Platform single core

mod2am: Dense matrix–matrix multiplication

- C + MKL
- Fortran

Max efficiency
96+ %

PRACE EU Deliverable D8.3.2
Reference platform single core

mod2as: Sparse matrix–vector multiplication

Max efficiency < 14%

C+MKL
Fortran

Matrix order
Mflop/s
0 500 1000 1500
100 1000 10000

mod2as (sparse matrix-vector multiplication)
Reference platform single core

mod2f: Fast Fourier Transform

- C + MKL
- Fortran

Max efficiency < 28%

PRACE EU Deliverable D8.3.2
Reference platform single core

mod2h: Random Number generation

C only

Mega-ops/s

10^5 10^6 10^7

Sequence length

mod2h

3600 3650 3700 3750 3800

PRACE EU Deliverable D8.3.2
Reference platform dual socket

The total power consumption of one reference node running 8 threads of mod2am is 303 W. Hence the power efficiency is 251 MFlop/s per Watt.

Max efficiency 93+ %
Reference platform dual socket

mod2as: Sparse matrix-vector multiplication

Cache friendly size

Max efficiency < 5%
(mostly < 3%)
Reference platform dual socket

No multi-threaded MKL version available
Only MPI version tried
No MPI, thread or OpenMP version exist. Therefore, the accumulated performance of 8 concurrent serial mod2h instances have been used as for reference at the node level.
Numeric issues - GPU

GPUs are not fully IEEE-754 compliant.

Some single-precision examples taken from nVidia programming manual

- Addition and multiplication are combined into a single multiply-add instruction (FMAD), which truncates the intermediate result of the multiplication
- For addition and multiplication, only round-to-nearest-even and round-towards-zero are supported via static rounding modes; directed rounding towards +/- infinity is not supported
- Division is implemented via the reciprocal in a non-standard-compliant way
- Square root is implemented via the reciprocal square root in a non-standard-compliant way

The GPU hardware used for assessment of the impact in the case of Gram-Schmidt (GS) orthogonalisation was an nVIDIA Tesla S1070 with four C1060 GPU cards each with 4 GB of 512 bits GDDR3 memory.
Assessment of non IEEE-754 compliance for GS orthogonalisation

The implementation of the dense orthogonalisation process on classical CPUs follows the standard algorithm and uses ATLAS subroutines. For GPUs the implementation uses mainly CUBLAS with special attention to memory operations between host and device memory: allocation and transfers.

For the sparse case, sparse matrix vector multiply from nVIDIA is used for the GPU implementation together with CUBLAS.

CGS is known to be very sensitive to machine precision.
The precision of the GS process for one subspace size is calculated using the greatest absolute dot product of every dot product of the orthogonal basis: \[
\max |(v_i, v_j)|, \text{ with } i \neq j \text{ and } v_i, v_j \in V \text{ basis. }
\]
A value of 0 corresponds to no errors, and a value of 1 indicates a completely wrong result. The accumulator of the dot products uses the same precision as the rest of the calculations: if the precision is SP, then the accumulator is SP. Same applies for DP.
Comparison of CPU and GPU results for Andrews sparse matrices (Univ of Florida collection) using CGSr (Classical Gram-Schmidt with re-orthogonalisation).

The GPU results are less regular due to a different execution order of the floating-point operations as well as very small numbers in the V basis that tend to be truncated during computations on the GPU.

PRACE EU Deliverable D8.3.2
EuroBen results – mod2am

Performance relative to reference platform of mod2am

Observation: only the tuned ClearSpeed accelerator with a 42% host assist is about 50% faster than the reference platform.
EuroBen results – mod2as

Performance relative to reference platform of mod2as

No accelerator is able to attain a decent fraction of the performance of the reference platform (1,392 MFlop/s).

Contributing causes:
- irregular memory references
  low computation to memory reference ratio (2/3)
- memory bandwidth limitations
- poor support for inner-products for all accelerators evaluated
EuroBen results – mod2f

Performance relative to reference platform of mod2f

As for mod2am only the Clearspeed accelerator achieved a significant improvement over the reference platform.
EuroBen results – mod2h

Only the Clearspeed card was able to implement it. For the ClearSpeed card, a library routine with a Mersenne Twister algorithm equivalent to the one in the original source code was used. The Clearspeed library version was in some cases about 1.5 times faster than the reference implementation (at 3,700 MoP/s).
HPL result for
48 node cluster of reference nodes

Peak performance: 3,886 TF
Measured performance: 3.538 TF
Efficiency: 91%
The memory-filling ratio was 2,58 out of 3 GB/s per core, or 86%.
Power efficiency: 230 MF/W
HPL data: N=360,360  NB=168  P=16  Q=24

Measurements carried out on the LRZ – CINES 2 system

PRACE EU Deliverable D8.3.1
HPL on 32 node cluster of reference nodes with Clearspeed accelerators

The accelerators add 1.4 TF for the given HPL parameters when N=361,152. This is 43.75 GF added per CSX e710 card with a peak of 96GF. Efficiency: 45.6%

The final LINPACK score is 3.5 TF
Efficiency: 61.8%

Without acceleration the peak achieved HPL performance is 2.24TF
Efficiency: 86.5%
HPL on 32 node cluster of reference nodes with Clearspeed accelerators

Only six of the eight cores are used by the MKL library (configured by the OMP_NUM_THREADS environment variable) because one core is required to manage the CSX710 processor and MKL operated more efficiently with an even number of cores.

To offload DGEMM computations matrix dimensions must for M and N be a multiple of 192 and for K a multiple of 288. For dimensions that are not a multiple of these values, the DGEMM is split into parts where the largest part is performed on the CSX710 and the DGEMM “edges” are performed on the host. For the benchmarking “ideal” dimensions were chosen.
HPL on 32 node cluster of reference nodes with Clearspeed accelerators

LRZ – CINES 1

DGEMM performance (K=1152)

- Host assist: (42%) [OMP_NUM_THREADS=6]
- Host assist (100%) [OMP_NUM_THREADS=6]
- Host assist (100%) [OMP_NUM_THREADS=8]
- Host Assit (0%)
- Host Assist (42%) [CS_BLAS_HOST_M_THRESHOLD=1152]
HPL on 32 node cluster of reference nodes with Clearspeed accelerators

Power efficiency

Performance increase: 56%
Power increase: 10%
Power efficiency increase: 42%
Power efficiency: $230 \times 1.42 = 326 \text{ MF/W}$
HPL on an eight note cluster of the reference platform with Tesla S1070

Single core plus single GPU
Peak: $10.12 \times 10^7 + 78 = 88.12$ GF
Measured: 59.5 GF
Efficiency: 67.5%

Power efficiency:
Reference platform alone: 240 MF/W
Reference platform with Tesla S1070: 270 MF/W

Eight node cluster with eight S1070
Peak: $3.144 \times 10^12$ TF
$(8 \times 8 \times 4 \times 2.53 + 8 \times 4 \times 78)$
Measured: 1.650 TF
Efficiency: 52.5%

Measurements by STFC
HPL on the SNIC/KTH prototype

Still under investigation due to IB firmware issues
So far best observed

DGEMM node efficiency: 93.6%
HPL node efficiency: 82.7%
HPL chassis efficiency: 79%
HPL power efficiency: 344 MF/W (preliminary)
HPL on eQPACE

Efficiency: 79.9% (44.5TF of 55.7TF)

Power efficiency: 773 MF/W
HPL Summary

Efficiency

Reference platform: 91%
Reference platform + Clearspeed: 61.8%
Reference platform + GPU: 52.5%
SNIC/KTH prototype: 79% (preliminary)
eQPACE (Cell): 79.9%
HPL Summary

Power Efficiency

<table>
<thead>
<tr>
<th>Description</th>
<th>MF/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference platform</td>
<td>240</td>
</tr>
<tr>
<td>Reference platform + Clearspeed</td>
<td>326</td>
</tr>
<tr>
<td>Reference platform + GPU</td>
<td>270</td>
</tr>
<tr>
<td>SNIC/KTH prototype</td>
<td>344 (prelim.)</td>
</tr>
<tr>
<td>eQPACE (Cell)</td>
<td>773</td>
</tr>
</tbody>
</table>
An application result – SNIC/KTH

Gromacs scaling on 24-core AMD blade PRACE prototype
331,776-atom system, reaction-field, 2fs steplength

AMD Istanbul blades compared to BlueGene/P:
~7x higher Gromacs performance per core
~1.8x higher Power consumption per core
3.9x higher power efficiency for MD!
Programming

- PGI compiler with accelerator features
- HMPP (from CAPS)
- RapidMind (bought by Intel)
- ......

PGI Accelerator version

Dense matrix-matrix multiplication (mod2am)
PGI-compiler on Nehalem versus NVIDIA C1060

Sparse matrix-vector multiplication (mod2as)
PGI-compiler on Nehalem versus NVIDIA C1060

mod2am on a single core Nehalem vs. nVIDIA C1060

mod2as on a single core Nehalem vs. nVIDIA C1060

PRACE EU Deliverable D8.3.2
Nehalem-EP single socket mod2am performance vs nVIDIA C1060 using HMPP

Nehalem-EP single socket mod2as performance vs nVIDIA C1060 using HMPP

PRACE EU Deliverable D8.3.2
The modification of an existing code to HMPP is lightweight to get a first non optimized running version.

Code modifications are needed if the code architecture does not take into account architectural features, such as vectorisation, as well as clean module isolation. Furthermore, some constructions (such as reductions) are very difficult to parallelize and won’t achieve decent performance on a graphic card.

Optimizing a code for a hybrid machine requires in-depth knowledge of the hardware. (NOT specific to HMPP; has been noticed in the CUDA porting, too)

Astute directives for code generation (such as loop reordering, loop fusion, etc.) are a great help to boost performance. HMPP does implement some.

With a bit of effort, performance offered by HMPP programs can be equal or better than those offered by a vendor’s library. This is an exciting result. It should encourage GPU experts to explicit common programming constructs.

More investigations are needed to fully grasp the advantage of using HMPP. In the coming months the usage of the same HMPP coding on different platforms (nVIDIA CUDA and ATI graphic cards) will be compared.
RapidMind

**Dense matrix-matrix multiplication --- GPU results (1 C1060)**

- sp, RM gpu-opt
- dp, MKL (8 N-EP cores)
- dp, CUDA
- sp, RM simple
- dp, RM gpu-opt.
- dp, RM simple

**Dense matrix-matrix multiplication --- CELL results (sp)**

- Cell SDK (16 SPUs)
- RM cell-opt. (16 SPUs)
- Cell SDK (8 SPUs)
- RM cell-opt. (8 SPUs)
- MKL (8 N-EP cores)
- RM simple (8 SPUs)

**Sparse matrix-vector multiplication (mod2as, dp)**

- MKL (8 N-EP cores)
- CUDA (C1060 GPU)
- RapidMind (8 N-EP cores)
- RapidMind (C1060 GPU)
- RapidMind (Cell, 8 SPUs)

**mod2as**

mod2am on Nehalem-EP vs nVidia C1060

mod2am on Nehalem-EP vs (CELL)

PRACE EU Deliverable D8.3.2
The diagram illustrates the performance of Fast Fourier Transformation (mod2f, sp) using different technologies: MKL (1 N-EP core), CUDA (C1060 GPU), RapidMind (C1060 GPU), RapidMind (8 N-EP cores), and RapidMind (Cell, 8 SPUs) across various data lengths. The performance is measured in Gflop/s, with the x-axis representing the length of the data and the y-axis showing the Gflop/s. The graph shows that CUDA (C1060 GPU) outperforms the other technologies, especially at higher data lengths.
RapidMind observations

Pure code portability (taking existing code and running it on a different backend) usually succeeds (except for some minor bugs found in the RM “Cell” backend)

However, performance is not at all portable across backends and performance prediction is very hard.
Programming productivity
Programmer Productivity

#Lines of Code (reported)

- A few languages stick-out positively:
  - Chapel (~1/16 LoC)
  - RapidMind (~1/4 LoC)

- Others require many lines of source code:
  - Cn (~2x)
  - CUDA (~3x, as soon as one cannot make use of a library call)

- No clear picture can be seen for CellSs
  - Few for the FFT, medium for MxM, many for SpMV

- All other languages are in the same range
  - MPI+OpenMP often being the shortest
  - UPC being one of the largest
Programmer productivity

Development time

- Most ports done within 5 days
- Very short (less than one day) development time for Chapel, MPI+OpenMP and CAPS hmpp
- Relatively long (5-10d) development time for CUDA and related ports (OpenCL, MPI+CUDA)
- Development time for UPC was unnaturally high due to many problems with the immature compiler
- CellSSs ports consumed much time, but programmers tried to get optimal performance, first version was always running within a few days
Programmer productivity

Developer diary example (2/2)

Developer Diary MPI+OpenMP sparse matrix-vector multiplication

Performance in Mflops vs Development time in minutes

- D1 = 30000 (3.33%)
- D2 = 30000 (13.33%)
- D3 = 40000 (4.24%)

PRACE Workshop “New Languages & Future Technology Prototypes” (March 1/2, 2010 at LRZ, Germany)
Summary Processors

• For the investigated computational kernels the present accelerators only incidentally will show performance benefits as compared to an 8-core, 2-socket Nehalem-EP node at 2.53 GHz.

• The comparisons were made with one accelerator per CPU socket, except for eQPACE. With additional accelerator cards per CPU socket the situation could be more advantageous.

• The performance of the accelerators are highly dominated by the host-accelerator bandwidth. Presently none of the accelerators are directly connected to the processor fabric, i.e., HyperTransport for AMD processors and Quickpath for Intel processors.

• The functionality of accelerators will be enhanced, for instance by adding support for reduction operations.

• Despite the lack of stellar results for most accelerators, with the exception of Cell, they will likely play a significant role in the compute nodes of future machines.
Summary – Energy efficiency

• Acceleration can significantly enhance the energy efficiency as demonstrated by the PRACE eQPACE prototype, ranked number 1 on the November 09 and June 2010 Green500 lists.

• But, the SNIC-KTH prototype illustrates that designs based on commodity hardware can achieve an energy efficiency that exceeds common accelerator designs without adding programming complexity.

• Present state of the art servers consumes several tens of Watt even when in idle state. Likewise, the power consumed by many of today’s accelerators is significant in idle state. Mechanisms to reduce the idle power are a major prerequisite to enhance the energy efficiency of accelerated systems.
Summary - Memory

Evolution and forecast of DDR memories

Memory performance is improving, but its fraction of the energy budget is rising.

Improving memory design in regards to energy consumption is a must for future large-scale systems. Support must be introduced for different power modes, with techniques such as e.g. control of refresh rates, throttling bandwidth or powering down banks. Software/OS support will be required to efficiently manage these mechanisms.
Summary programming

Programming heterogeneous systems is still difficult and tools need significant improvement.
Thank YOU!