HPC – Algorithms and Applications

– Intro –

Michael Bader
Winter 2013/2014
Part I

Scientific Computing and Numerical Simulation
The Simulation Pipeline

- phenomenon, process etc.
- modelling
- mathematical model
- numerical treatment
- numerical algorithm
- parallel implementation
- simulation code
- visualization
- results to interpret
- embedding
- statement
- tool
Example – Millennium-XXL Project

- $N$-body simulation with $N = 3 \cdot 10^{11}$ “particles”
- compute gravitational forces and effects
  (every “particle” correspond to $\sim 10^9$ suns)
- simulation of the generation of galaxy clusters
  plausibility of the “cold dark matter” model

(Springel, Angulo, et al., 2010)
Simulation – HPC-Related Data:

- $N$-body simulation with $N = 3 \cdot 10^{11}$ “particles”
- 10 TB RAM required only to store particles positions and velocities (single precision)
- total memory requirement: 29 TB
- JuRoPa supercomputer (Jülich)
- simulation on 1536 nodes (each 2x QuadCore, thus 12288 cores)
- hybrid parallelisation: MPI plus OpenMP/Posix threads
- runtime: 9.3 days; 300 CPU years in total
Example – Gordon Bell Prize 2010

(Rahimian, . . . , Biros, 2010)

- direct simulation of blood flow
- particulate flow simulation (coupled problem)
- Stokes flow for blood plasma
- red blood cells as immersed, deformable particles
Example – Gordon Bell Prize 2010 (2)

Simulation – HPC-Related Data:

- up to 260 Mio blood cells, up to $9 \cdot 10^{10}$ unknowns
- fast multipole method to compute Stokes flow (octree-based; octree-level 4–24)
- scalability: 327 CPU-GPU nodes on Keeneland cluster, 200,000 AMD cores on Jaguar (ORNL)
- 0.7 Petaflops/s sustained performance on Jaguar
- extensive use of GEMM routine (matrix multiplication)
- runtime: $\approx 1$ minute per time step

Article for Supercomputing conference:
http://www.cc.gatech.edu/~gbiros/papers/sc10.pdf
“Faster, Bigger, More”

Why parallel high performance computing:

- **response time**: compute a problem in $\frac{1}{p}$ time
  - speed up engineering processes
  - real-time simulations (tsunami warning?)

- **Problem size**: compute a $p$-times bigger problem
  - Simulation of multi-scale phenomena
  - maximal problem size that “fits into the machine”

- **Throughput**: compute $p$ problems at once
  - case and parameter studies, statistical risk scenarios, etc.
  - massively distributed computing (SETI@home, e.g.)
Part II

HPC in the Literature – Past and Present Trends
Four Horizons for Enhancing the Performance of Parallel Simulations Based on Partial Differential Equations (David Keyes, 2000)

1. Expanded Number of Processors
   → in 2000: 1000 cores; in 2010: 200,000 cores

2. More Efficient Use of Faster Processors
   → PDF working-sets, cache efficiency

3. More Architecture-Friendly Algorithms
   → improve temporal/spatial locality

4. Algorithms Delivering More “Science per Flop”
   → adaptivity (in space and time), higher-order methods, fast solvers
The Seven Dwarfs of HPC

“dwarfs” = key algorithmic kernels in many scientific computing applications

P. Colella (LBNL), 2004:
1. dense linear algebra
2. sparse linear algebra
3. spectral methods
4. N-body methods
5. structured grids
6. unstructured grids
7. Monte Carlo
Computational Science Demands a New Paradigm

Computational simulation must meet three challenges to become a mature partner of theory and experiment (Post & Votta, 2005)

1. **performance challenge**
   → exponential growth of performance, massively parallel architectures

2. **programming challenge**
   → new (parallel) programming models

3. **prediction challenge**
   → careful verification and validation of codes; towards reproducible simulation experiments
Free Lunch is Over(*)
…actually already over for quite some time!

Speedup of your software can only come from parallelism:

- clock speed of CPU has stalled
- instruction-level parallelism per core has stalled
- number of cores is growing
- size of vector units is growing

Performance Development in Supercomputing

(source: www.top500.org)
## Top 500 (www.top500.org) – June 2013

<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, China</td>
<td><strong>Tianhe-2 (MilkyWay-2)</strong> - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT</td>
<td>3120000</td>
<td>33862.7</td>
<td>54902.4</td>
<td>17808</td>
</tr>
<tr>
<td>2</td>
<td>DOE/SC/Oak Ridge National Laboratory, United States</td>
<td><strong>Titan</strong> - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.</td>
<td>560640</td>
<td>17590.0</td>
<td>27112.5</td>
<td>8209</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LLNL, United States</td>
<td><strong>Sequoia</strong> - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM</td>
<td>1572864</td>
<td>17173.2</td>
<td>20132.7</td>
<td>7890</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>705024</td>
<td>10510.0</td>
<td>11280.4</td>
<td>12660</td>
</tr>
<tr>
<td>5</td>
<td>DOE/SC/Argonne National Laboratory, United States</td>
<td><strong>Mira</strong> - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM</td>
<td>786432</td>
<td>8586.6</td>
<td>10066.3</td>
<td>3945</td>
</tr>
<tr>
<td>6</td>
<td>Texas Advanced Computing Center/Univ. of Texas, United States</td>
<td><strong>Stampede</strong> - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, InfiniBand FDR, Intel Xeon Phi SE10P Dell</td>
<td>462462</td>
<td>5168.1</td>
<td>8520.1</td>
<td>4510</td>
</tr>
</tbody>
</table>
Top 500 Spotlights – Tianhe-2 and K Computer

**Tianhe-2/MilkyWay-2 → Intel Xeon Phi (NUDT)**
- 3.1 mio cores(!) – Intel Ivy Bridge and Xeon Phi
- Linpack benchmark: 33.8 PFlop/s
- ≈ 17 MW power(!!)
- Knights Corner / **Intel Xeon Phi** / Intel MIC as accelerator
- 61 cores, roughly 1.1–1.3 GHz

**Titan → Cray XK7, NVIDIA K20x (ORNL)**
- 18,688 compute nodes; 300,000 Opteron cores
- 18,688 **NVIDIA Tesla K20 GPUs**
- Linpack benchmark: 17.6 PFlop/s
- ≈ 8.2 MW power
Top 500 Spotlights – Sequoia and K Computer

**Sequoia** → IBM BlueGene/Q (LLNL)
- 98,304 compute nodes; 1.6 mio cores
- Linpack benchmark: 17.1 PFlop/s
- ≈ 8 MW power

**K Computer** → SPARC64 (RIKEN, Kobe)
- 88,128 processors; 705,024 cores
- Linpack benchmark: 10.51 PFlop/s
- ≈ 12 MW power
- SPARC64 VIIIfx 2.0GHz (8-core CPU)
Performance Development in Supercomputing

(source: www.top500.org)
International Exascale Software Project Roadmap

Towards an Exa-Flop/s Platform in 2018 (www.exascale.org):

1. **technology trends**
   - concurrency, reliability, power consumption, . . .
   - blueprint of an exascale system: 10-billion-way concurrency, 100 million to 1 billion cores, 10-to-100-way concurrency per core, hundreds of cores per die, . . .

2. **science trends**
   - climate, high-energy physics, nuclear physics, fusion energy sciences, materials science and chemistry, . . .

3. **X-stack** (software stack for exascale)
   - energy, resiliency, heterogeneity, I/O and memory

4. **Polito-economic trends**
   - exascale systems run by government labs, used by CSE scientists
Exascale Roadmap

“Aggressively Designed Strawman Architecture”

<table>
<thead>
<tr>
<th>Level</th>
<th>What</th>
<th>Perf.</th>
<th>Power</th>
<th>RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPU</td>
<td>FPU, regs,. instr.-memory</td>
<td>1.5 GF</td>
<td>30 mW</td>
<td></td>
</tr>
<tr>
<td>Core</td>
<td>4 FPUs, L1</td>
<td>6 GF</td>
<td>141 mW</td>
<td></td>
</tr>
<tr>
<td>Proc. Chip</td>
<td>742 cores, L2/L3, Intercon.</td>
<td>4.5 TF</td>
<td>214 W</td>
<td></td>
</tr>
<tr>
<td>Node</td>
<td>Proc. chip, DRAM</td>
<td>4.5 TF</td>
<td>230 W</td>
<td>16 GB</td>
</tr>
<tr>
<td>Group</td>
<td>12 proc. chips, routers</td>
<td>54 TF</td>
<td>3.5 KW</td>
<td>192 GB</td>
</tr>
<tr>
<td>rack</td>
<td>32 groups</td>
<td>1.7 PF</td>
<td>116 KW</td>
<td>6.1 TB</td>
</tr>
<tr>
<td>System</td>
<td>583 racks</td>
<td>1 EF</td>
<td>67.7 MW</td>
<td>3.6 PB</td>
</tr>
</tbody>
</table>

approx. 285,000 cores per rack; 166 mio cores in total

Source: ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems
Exascale Roadmap – Should You Bother?

Your department’s compute cluster in 5 years?

- a **Petaflop System**!
- “one rack of the Exaflop system”
  → using the same/similar hardware
- extrapolated example machine:
  - peak performance: 1.7 PFlop/s
  - 6 TB RAM, 60 GB cache memory
  - “total concurrency”: $1.1 \cdot 10^6$
  - number of cores: 280,000
  - number of chips: 384

Source: ExaScale Software Study: Software Challenges in Extreme Scale Systems
Your Department’s PetaFlop/s Cluster in 5 Years?

**Tianhe-1A** (Tianjin, China; Top500 # 10)
- 14,336 Xeon X5670 CPUs
- **7,168 Nvidia Tesla M2050 GPUs**
- Linpack benchmark: ≈ 2.6 PFlop/s
- ≈ 4 MW power

**Stampede** (Intel, Top500 # 6)
- 102,400 cores (incl. Xeon Phi: MIC/“many integrated cores”)
- Linpack benchmark: ≈ 5 PFlop/s
- Knights Corner / **Intel Xeon Phi** / Intel MIC as accelerator
- 61 cores, roughly 1.1–1.3 GHz
- wider vector FP units: 64 bytes (i.e., 16 floats, 8 doubles)
- ≈ 4.5 MW power
Top 500 (www.top500.org) – # Cores

Cores per Socket System Share

- 8 Cores: 54.8%
- 6 Cores: 20.6%
- 4 Cores: 10.8%
- 16 Cores: 3.7%
- 12 Cores: 1.9%
- 2 Cores: 0.4%
- 1 Core: 0.3%
Top 500 (www.top500.org) – # Cores

Cores per Socket Performance Share

- 8 cores: 37.2%
- 6 cores: 17.3%
- 4 cores: 31.5%
- 2 cores: 9.4%
- Other: 10.8%

TOP10 June 2013
1. Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P (/system/177999) NUDT
Part III

Organisation
Lecture Topics – The Seven Dwarfs of HPC

Algorithms, parallelisation, HPC-aspects for problems related to:

1. dense linear algebra
2. sparse linear algebra
3. spectral methods
4. N-body methods
5. structured grids
6. unstructured grids
7. Monte Carlo
Examples on GPUs using CUDA:

- dense linear algebra
- structured grids
- sparse linear algebra
- (N-body methods)

Tutors and time/place:

- Oliver Meister
- roughly bi-weekly tutorials (90min)
  (see website for exact schedule)
- small “dwarf-related” projects
  (1–2 tutorial lessons per project)
Exams, ECTS, Modules

Exam:
- written or oral exam depending on number of participants
- include exercises (as part of the exam)

ECTS, Modules:
- 4 ECTS (2+1 lectures per week)
- Informatik/Computer Science: Master catalogue
- CSE: Application area
- others?