Exploring Extreme Scalability in Scientific Applications

Mike Ashworth, Ilian Todorov
Computational Science & Engineering
STFC Daresbury Laboratory

Ian Bush
Numerical Algorithms Group Ltd

Mario Chavez
Institute of Engineering, UNAM, Mexico

mike.ashworth@stfc.ac.uk
http://www.cse.scitech.ac.uk/

ECMWF HPC Workshop
Outline

• Why explore extreme scalability?
• How are we doing this?
• What have we found so far?
• Where are we going next?
HPC Strategy Committee:

"... the UK should aim to achieve sustained Petascale performance as early as possible across a broad field of scientific applications, permitting the UK to remain internationally competitive in an increasingly broad set of high-end computing grand challenge problems."

... from A Strategic Framework for High-End Computing
1. Doubling performance every two years does not keep you competitive

2. Staying competitive is becoming more and more difficult
1. HECToR brings the total UK provision to over 100 Tflop/s

2. National provision is under pressure from University systems
1. The total computing power available to UK academia has doubled about every 15 months.

2. We can expect to be using Petascale resources in the UK in 2012 – about 4 years behind USA.
What will a Petascale system look like?

Current indicators:
- TOP500 #1 LANL Roadrunner 1.026 Pflop/s
  - 122,400 processors, attached Cell processors
- TOP500 #2 LLNL Blue Gene L 0.478 Pflop/s
  - 212,992 processors, dual-core nodes
- TOP500 #3 ANL Blue Gene P 0.450 Pflop/s
  - 163,840 processors, 2x dual-core nodes
- ORNL late-2008 upgrade to Cray XT4 ~1 Pflop/s
  - ~120,000 processors, quad-core nodes
- Japanese Petascale project
  - Smaller number of ~100 Gflop/s vector processors

Most likely solution for the UK is O(100,000) processors using multi-core components or attached processors
Challenges at the Petascale

Scientific:
- What new science can you do with 1000 Tflop/s?
- Larger problems, multi-scale, multi-disciplinary

Technical:
- How will existing codes scale to 10,000 or 100,000 processors?
  Scaling of time with processors, time with problem size, memory with problem size
- Data management, incl. pre- and post-processing
- Visualisation
- Fault tolerance
Daresbury Petascale project

Scaling analysis of current codes

Performance analysis on $O(10,000)$ procs

Forward-looking prediction to $O(100,000)$ procs

Optimisation of current algorithms

Development of new algorithms

Evaluation of alternative programming models
Machines
Cray XT4 *HECToR*
- DC 2.8 GHz Opteron 11328 cores

Cray XT3/XT4 *old-jaguar*
- DC 2.6 GHz Opteron ~12,000 XT4 cores

Cray XT3 palu CSCS
- DC 2.6 GHz Opteron 3328 cores

IBM p5-575 *HPCx*
- DC 1.7 GHz POWER5, HPS, 2560 cores

IBM BlueGene/L *jubl*
- DC 700 MHz PowerPC, 16384 cores
Applications
Applications

PDNS3D/SBLI
- Direct numerical simulation of turbulent flow

POLCOMS
- Coastal-ocean finite difference code

fd3d
- Earthquake simulation finite difference code

DL_POLY3
- Molecular dynamics code

CRYSTAL
- First principles periodic quantum chemistry code
What is a processor?

A processor by any other name ...

An applications view ...

A processor is what is has always been ...

- A short name for Central Processing Unit
- Something that runs a single instruction stream
- Something that runs an MPI task
- Something that runs a bunch of threads (OpenMP)
PDNS3D / SBLI
DNS results of near-wall turbulent flow
3D grid partitioning with halo cells

**Calculation cost:**
Scales as $n^3$

**Communication cost:**
Scales as $n^2$

**Strong scaling:**
Increasing $P$
Decreasing $n$
Comms will dominate
SBLI on jaguar

Turbulent channel flow benchmark

![Graph showing performance (Mgrid-points*iterations/sec) vs. number of processors. The graph plots the performance of T3 Cray XT4 jaguar against the number of processors, showing an increase in performance as the number of processors increases.]
**SBLI on Cray XT4**

**Turbulent channel flow benchmark**

Larger problems scale better

3rd Nov 2008  ECMWF HPC Workshop
% comms time from craypat

![Graph showing the percentage of communication time with different processor configurations.](image)
POLCOMS
POLCOMS is the finest resolution model to-date to simulate the circulation, temperature and salinity of the Northwest European continental Shelf important for understanding of the transport of nutrients, pollutants and dissolved carbon around shelf seas

We have worked with POL on coupling with ERSEM, WAM, CICE, data assimilation and optimisation for HPC platforms

Coupled Marine Ecosystem Model

- Physical Model
  - Wind Stress
  - Open Boundary
  - Heat Flux
  - Irradiation
  - Cloud Cover

- Pelagic Ecosystem Model
  - C, N, P, Si
  - River Inputs

- Benthic Model
  - Sediments
  - C, N, P, Si

- Pelagic Ecosystem Model
  - River Inputs
  - Wind Stress
  - Open Boundary
  - Heat Flux
  - Cloud Cover

- Benthic Model
  - Sediments
  - C, N, P, Si
Structured-grid finite difference code from POL
Sophisticated advection scheme to represent, fronts, eddies etc in the shelf seas
Halo-based partitioning
Complicated by land/sea issue

**Performance dependent on partitioning**

Known issue with communications imbalance – new version under test
Largest domain size limited by I/O through master
Efficient parallel I/O is essential for this code
fd3d
Seismic wave propagation
3D velocity-stress equations
Structured grid
Explicit scheme
• 2nd order accurate in time
• 4th order accurate in space
Regular grid partitioning
Halo exchange
fd3d output: synthetic seismograms
Locations of: a) the epicenter (red dot) of the 12 05 2008 Sichuan Ms 7.9; b) its rupture area and its kinematic slip; c) 9 seismographic stations sites (black dots) of the China Seismographic Network; d) the surficial projection of the 2400 x 1600 x 300 km3 volume used to discretize the region of interest; f) the geologic structure adopted for the volume
fd3d on Cray XT4 HECToR

- Craypat MPI: 10.4%
- 16.5%
- 14.5%
DL_POLY
Conventional routines (e.g. fftw) assume plane or column distributions. A global transpose of the data is required to complete the 3D FFT and additional costs are incurred re-organising the data from the natural block domain decomposition.

An alternative FFT algorithm has been designed to reduce communication costs.

- the 3D FFT is done as a series of 1D FFTs, each involving communications only between blocks in a given column
- The data distribution matches that used for the rest of the DL_POLY energy routines
- More data is transferred, but in far fewer messages
- Rather than all-to-all, the communications are column-wise only (see sparse comms structure, left)
BlueGene/L times

14.6 million particle Gd$_2$Zr$_2$O$_7$ system

- MD total
- Ewald - k space
- Link
- Other
- Van der Waals
- Ewald - Real Space

Number of Processors vs. Seconds / Evaluation
Cray XT4 & BGL performance

![Graph showing Cray XT4 hector and IBM BlueGene/L jubl performance](image)
Scaling analysis BGL

![Graph showing speed-up vs number of processors for different computational tasks]
Scaling analysis XT4

- Van der Waals
- Ewald - Real space
- Link
- Other
- Ewald - k space
- MD total
- Ideal

Number of Processors

Speed-up

0 2048 4096 6144 8192

3rd Nov 2008

ECMWF HPC Workshop
Excellent scaling with \(>\sim 1000\) particles per processor
Scalability limited by long-range forces
Can use force-shifted Coulomb electrostatics
Fast multipole electrostatics for even larger systems

I/O is a major bottleneck
Efficient parallel I/O is essential for this code
Plus tools to handle & visualize large output datasets

“The Need for Parallel I/O in Classical Molecular Dynamics”, Ilian Todorov, CUG 2008
CRYSTAL
Crystal

Electronic structure and related properties of periodic systems

All electron, local Gaussian basis set, DFT and Hartree-Fock

Under continuous development since 1974

Distributed to over 500 sites world wide

Developed jointly by Daresbury and the University of Turin
Crambin Results – Electrostatic Potential

Charge density isosurface coloured according to potential
Useful to determine possible chemically active groups
SCF cycle scaling

1737 Atoms, 23268 Basis functions

Number of Processors

Performance (arbitrary)

Cray XT4 HECToR
IBM p5-575 HPCx
ideal
SCF breakdown

The graph illustrates the percentage execution time across different numbers of processors for various computational tasks. The tasks include HPCx Integrals, HECToR Integrals, HPCx Diag, and HECToR Diag. The x-axis represents the number of processors, while the y-axis shows the percentage execution time. The graph shows a trend where the percentage execution time decreases as the number of processors increases, indicating improved efficiency with more processors.
CRYSTAL

SCF cycle dominated by two parts
Integral evaluation for the Kohn-Sham matrix
  - Time scales linearly
  - Difficult to distribute so poor scaling in memory
Dense linear algebra (diagonalization)
  - Standard libraries (e.g. ScaLaPack D&C)
  - Communications-heavy so poor scaling

Starts with integral evaluation dominating
For larger systems and larger number of processors the diagonalization dominates
Will need to look at diagonalization-less methods

We have looked at five codes up to 16384 procs

- Mainly to 8192 on Cray XT4, also BlueGene/L and /P

Most codes scale well to $O(10,000)$ procs:

- Need large problem sizes
- Need efficient parallel I/O (in progress)
- Need diagonalization-less methods for quantum chemistry

Prospects look good to exploit higher numbers

- Scaling isn’t everything, need to look also at efficiencies – especially for quad-core, multi-core and beyond
- Fortran+MPI works just fine (so far!)
Several speakers concluded that:

- The MPI send-receive model may hit limitations at very high processor numbers
- Hybrid programming e.g. MPI/OpenMP may help, only one MPI task per multi-core node, esp. for collectives, also saves memory
- Single-sided messaging may be needed and the PGAS languages (e.g. Co-Array Fortran, UPC) may be a good high-level interface

“Migrating a Scientific Application from MPI to Co-Arrays”, Ashby & Reid, CUG 2008
Conclusions

Petascale computing will soon be available in the UK

Largely achieved by massive increases in the number of processors

Systems will be based on multi-core nodes

We need to look now at scalability and other issues on O(10,000-100,000) processors

We may need to look at alternatives/additions to the existing programming model (serial language + MPI)
New Opportunities

Computational Science is evolving very rapidly

Hardware is moving rapidly towards the Petascale
- Extreme scalability is required to 100k processors at beyond
- Clusters of multi-core SMP nodes

Scientific demands are also changing
- Multi-scale
- Multi-disciplinary

We need to deliver on the evolving aspirations of the community across a broad spectrum of scientific and engineering disciplines

3rd Nov 2008
ECMWF HPC Workshop
The Hartree Centre will be a new kind of Computational Sciences institute for the UK that will:

– stimulate a step change in modeling capabilities for strategic science themes – Grand challenge projects
– multi-disciplinary, multi-scale, effective and efficient simulation
– have at its heart the collaborative development, support and exploitation of scientific applications software – this is the key to real scientific and economic impact and will be Hartree’s essential driver.
If you have been ... ... thank you for listening

Mike Ashworth

http://www.cse.scitech.ac.uk/