CUG 2008 HELSINKI • MAY 5–8, 2008
CROSSING THE BOUNDARIES

HECToR, the CoE and Large-Scale Application Performance on CLE

David Tanqueray,
Jason Beech-Brandt,
Kevin Roy*,
Martyn Foster,
Cray Centre of Excellence for HECToR

CRAY
THE SUPERCOMPUTER COMPANY

Topics

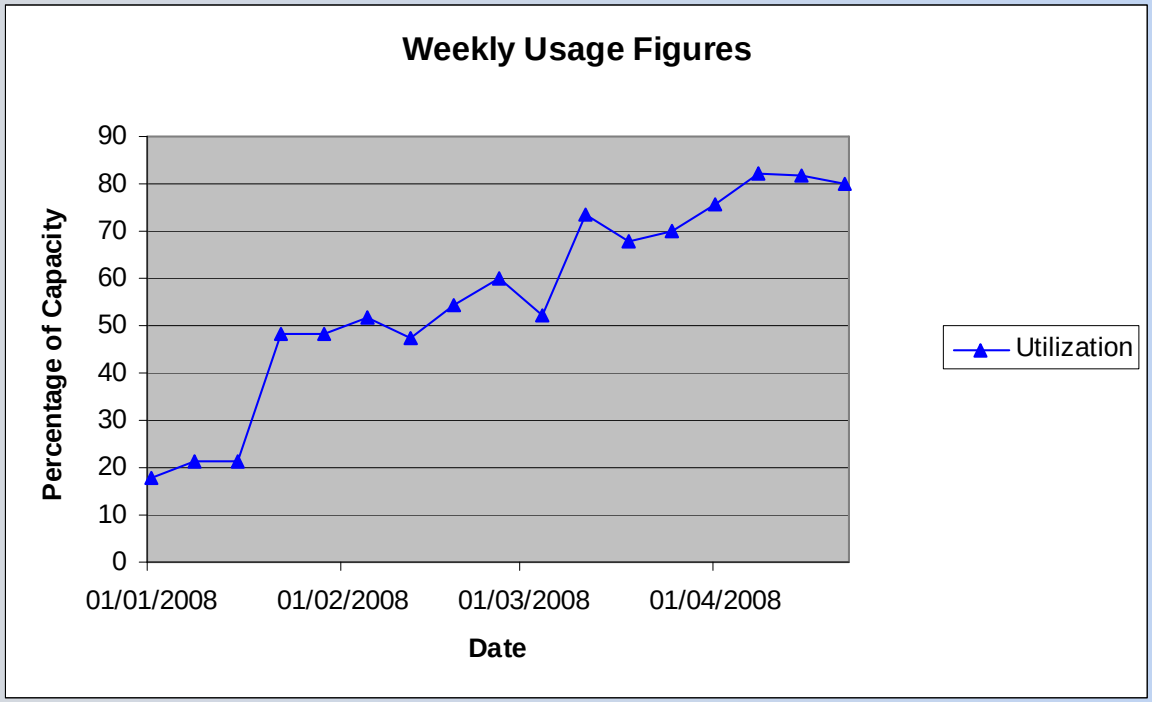
- HECToR
- The Centre of Excellence
- Activities
  - CASINO
  - SBLI
  - DLPOLY
  - HemeLB
  - Others
- The Future

HECToR

- High End Computing Terascale Resource (HECToR)
  - EPSRC, BBSRC, NERC – funding agencies
  - UoE HPCX Ltd – main contractor, administration, helpdesk and website
    - **Application Performance on the UK's New HECToR Service**, Fiona Reid, HPCX Consortium (HPCX) (3.15pm today)
  - NAG Ltd – CSE provider
  - University of Edinburgh - housing
  - Cray Inc – hardware (also some CSE support)
- Provides HPC facilities for UK academ
  - Using peer review process
  - Using directed calls
  - Covers wide spectrum of science
- Will contribute to resources to DEISA

HECToR

- HECToR is a 60 cabinet dual core XT4 system
  - Installed August 2007
- One of the first Cray Linux Environment (CLE) systems to go into user service
- A X2 upgrade is imminent
  - One cabinet 112 cores
- First hybrid system
  - soon
- It is well utilized



CSE Support on HECToR

- Partner with HECToR user community to assist in deriving maximum benefit from XT4/X2 etc.
  - Training, Documentation, case studies, FAQs
  - Assistance with porting, performance tuning and optimisation of user codes
- Teamwork
  - NAG HECToR CSE
    - Central Team: ~8 FTEs based in Oxford
    - Distributed Team: ~12 FTEs
      - seconded to particular users, research groups or consortia
      - Currently supporting NEMO, Castep, Casino. Others in the pipeline.
- The CoE complements this group

The CoE

- What is the CoE?
  - Cray Centre of Excellence for HECToR
- Work with all the partners and the user community
  - Look at upcoming software ready for integration into HECToR
  - Training
  - Application optimization
    - more focused on getting the best from the Cray Hardware.
  - Support CSE activities
  - Test future platforms for HECToR
  - A conduit to Cray Engineering

Casino Enhancements

- Time to read data set is excessive (ASCII file: 7.6GB up to 16.3GB)
  - Runs are in multiple batch jobs, each continuing from each other
  - Data file must be reloaded for each job before it gets going
  - Example: 7.6GB takes ~1200 seconds before useful work starts
- Solution:
  - First time through write out file in Binary (can be done on 1 node)
  - Subsequent runs detect binary file and use that
  - Results in size reduction too!

Casino Enhancements

- Wants to use VERY LARGE wave function data sets
  - Having 2 copies (1 per core) on a node limits the problem size he can run
  - Array is read only (once loaded) so only 1 copy is really needed
- Solution:
  - OpenMP is an option, but the code already scales very well so engineering overhead in inserting enough OpenMP
  - Use a single SHARED array (between MPI tasks on node)
  - Can’t use Posix as /dev/shm is not user writeable
  - Use System V shared memory
    - BUT: System V shared memory uses int (32 bits) for size

Results

- Now can use larger wave functions sets (2x)
  - Will be 4x with quad core
  - 8x with XT5
- Binary option increases flexibility
- All with increased performance!!!

SBLI-Shock Boundary Layer Interaction (1/4)

- Finite difference code for turbulent boundary layers
- Higher-order central differencing, shock preserving advection scheme from the TVD family, entropy splitting of the Euler terms
- Used as an early access code on HECToR
  - Users were running on 4k cores within one hour
  - Allowed users to do simulations not possible on HPCx
  - Have enough data from early access time for a journal publication
  - Post-processing of this early-access data is ongoing
- Code scaled to over 12k cores on Jaguar at ORNL
- HPCx scaling stops at around 1200 processors
- Developers wanted to improve the single-CPU performance on HECToR

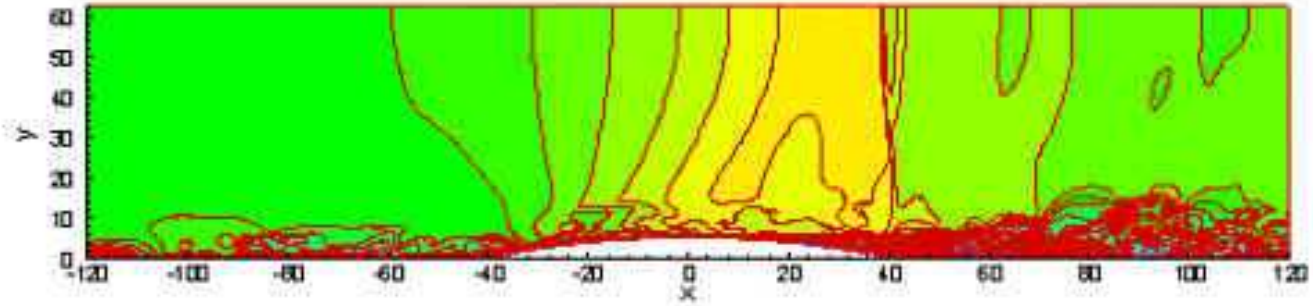


Figure illustrates instantaneous u-velocity contours of flow over a Delery bump

SBLI (2/4)

- Using the compiler feedback and CrayPAT it was possible make significant savings

- Showed that key parts were not vectorising
  - -Mneginfo tells us why certain optimizations are not being performed

    413, Loop not vectorized: data dependency

    real*8 temp(42) – this is used in place of individually declaring a large number of scalar temporaries
  - Doing this saved 20% in this routine, which itself is the most time consuming routine the code.
- Identified next region
  - Implemented appropriate cache blocking

Revised codes cache profile (3/4)

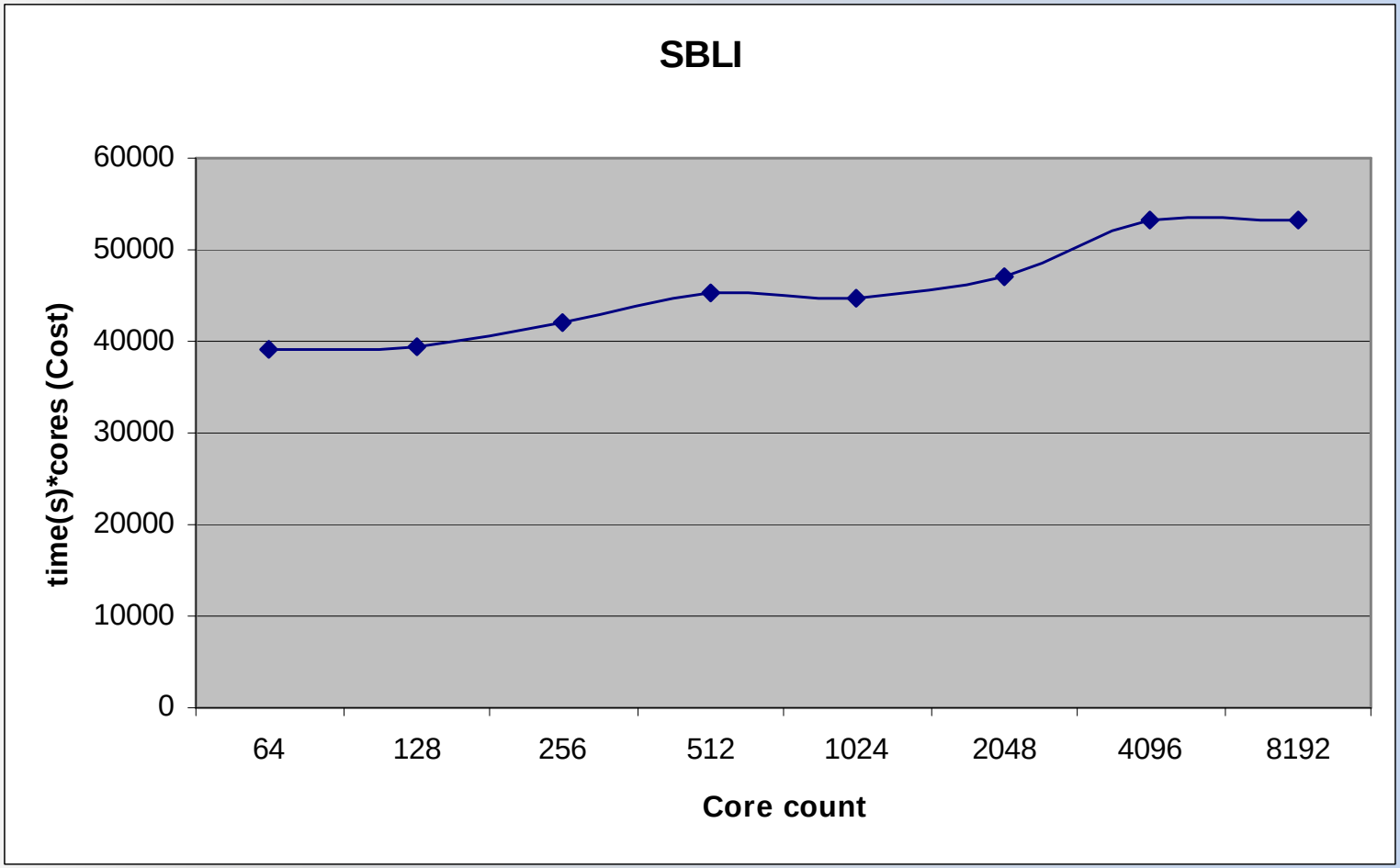
USER / deriv_d1eta_2_				orig
Time%		8.1%		12.4%
Time		22.654139		39.8
Imb.Time		3.048877		
Imb.Time%		12.1%		
Calls		2854		
PAPI_L1_DCA	910.346M/sec	14907115715	refs	
DATA_CACHE_REFILLS:SYSTEM	2.024M/sec	33136218	fills	
DATA_CACHE_REFILLS:L2_ALL	39.088M/sec	640067739	fills	
REQUESTS_TO_L2:DATA	63.320M/sec	1036880831	req	
Cycles	16.375 secs	42575593125	cycles	
User time (approx)	16.375 secs	42575593125	cycles	
Utilization rate		72.3%		
L1 Data cache misses	41.111M/sec	673203957	misses	
LD & ST per D1 miss		22.14	refs/miss	
D1 cache hit ratio		95.5%		89.8%
LD & ST per D2 miss		449.87	refs/miss	
D2 cache hit ratio		96.8%		90.2%
L2 cache hit ratio		95.1%		87.5%
Total cache hit ratio		99.8%		

+43%

Significantly better cache behaviour, and much less time is being spent doing these derivative calculations

SBLI (4/4)

- Also achieves better performance using PathScale



DLPoly v3 (1/3)

- Developed at STFC Daresbury Labs
  - Bill Smith, Ilian Todorov, Ian Bush
- Recently modified to use MPI-IO
- In the top 5 of HECToR applications
- CoE works with STFC Daresbury Labs to put changes into production release.

- Load balancing fix proposal in dlpoly4
  - Want to have CoE involvement
  - May get dCSE involvement
  - 2-4 Man years effort

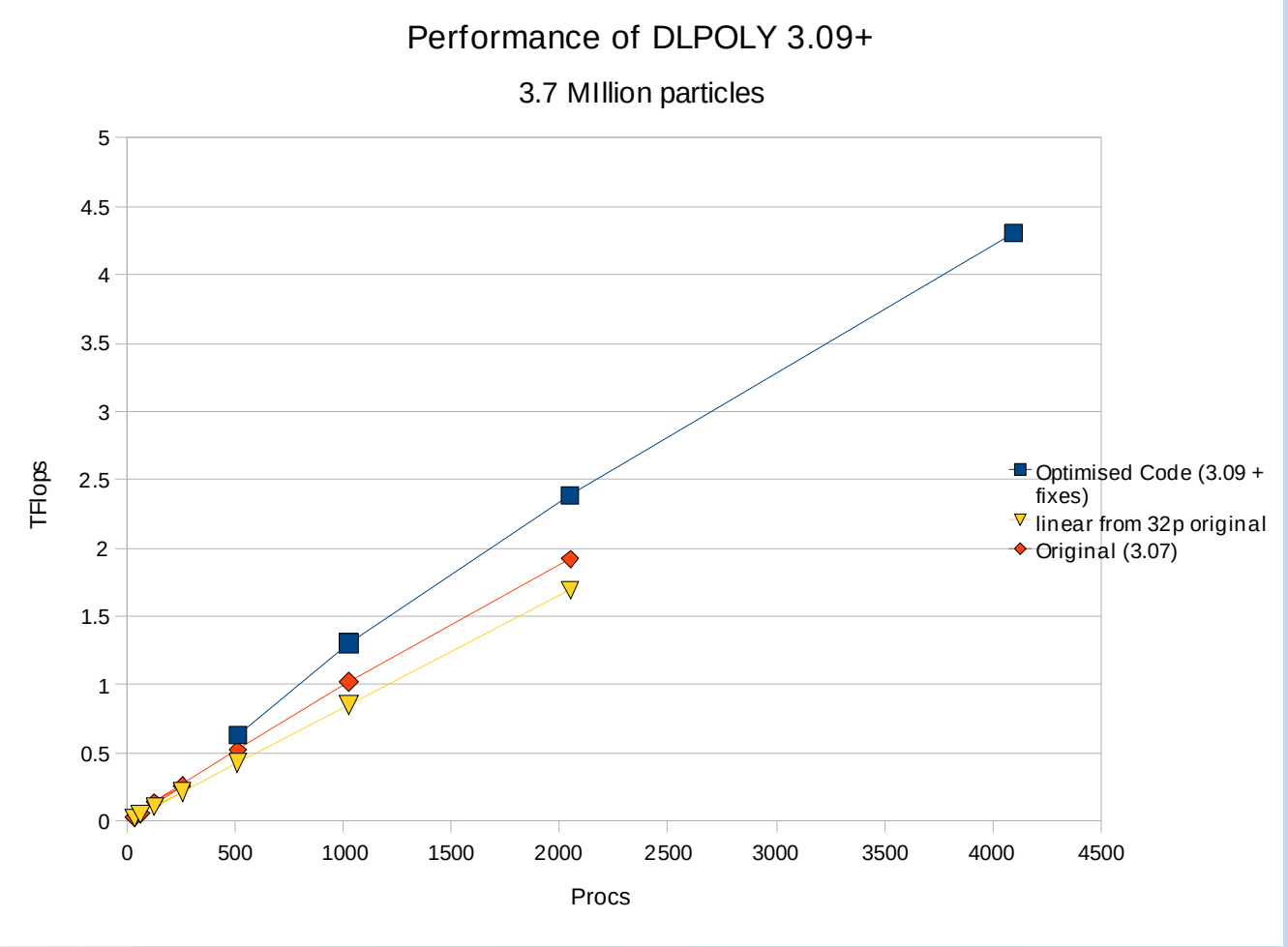
- Hector capability challenge (30 million AUs on HECToR)
  - Can we make egg shells without using chickens
  - ~7000 atoms in system
  - At 512p, only 136 atoms per core

  - Lots of IO
    - History every 500 cycles
    - Full dump every 1000 cycles
    - Formatted, sorted, ASCII

    - 20MB per write (approx 15s IO every 25s compute, **ouch**)

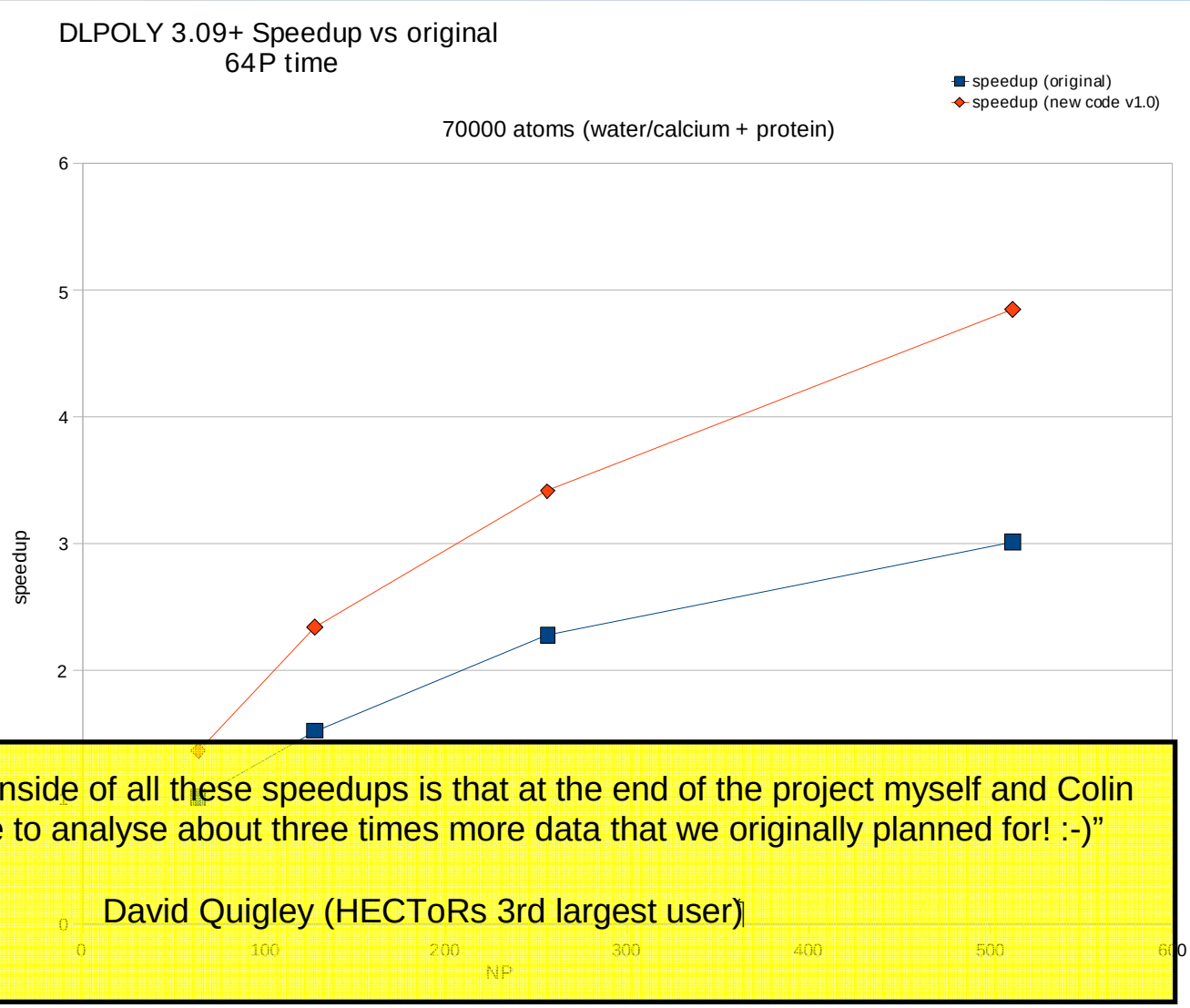
Science & Technology Facilities Council

DLPOLY Large Scalable systems (2/3)



Single Processor Improvements (3/3)

- Code change
  25%)
- Some propos
  not accepted
  - Maintainabili
    performance
  - Hopefully rev
    reintegrate th



"Of course the downside of all these speedups is that at the end of the project myself and Colin are going to have to analyse about three times more data that we originally planned for! :-)"

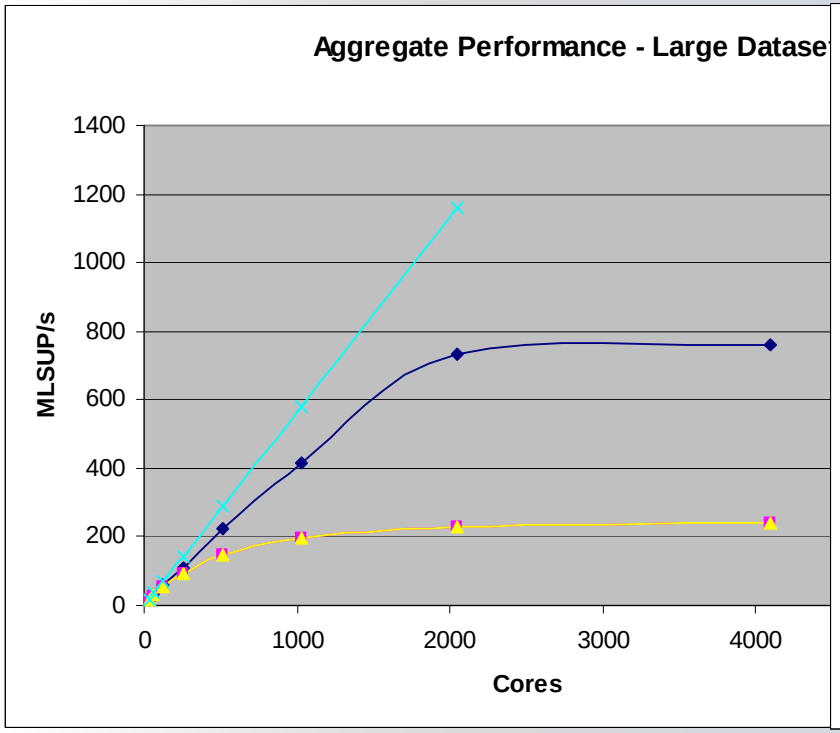
David Quigley (HECToRs 3rd largest user)

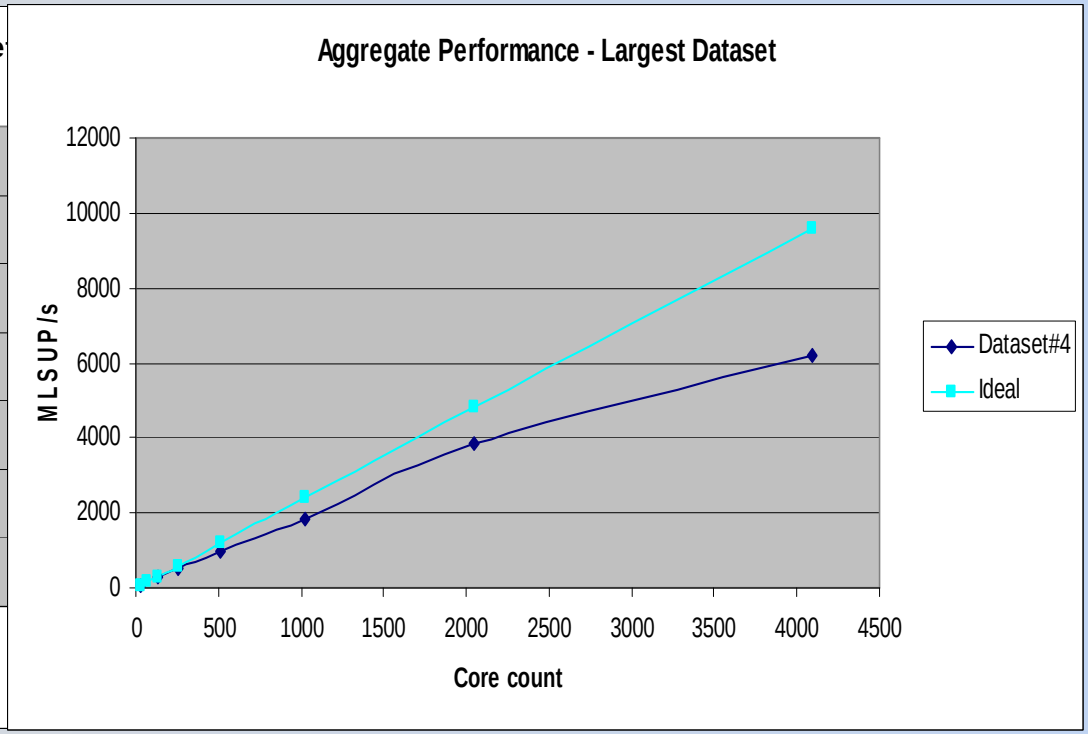
HemeLB (1/3)

- The HemeLB code is a parallel implementation of the Lattice-Boltzmann method for simulation of blood flow in cerebro-vascular systems.
- The code is designed to run on distributed and single multiprocessor machines using implementations of the well known MPI standard and is highly scalable, in order to be used on massively parallel computers and computational Grids.

HemeLB (2/3)

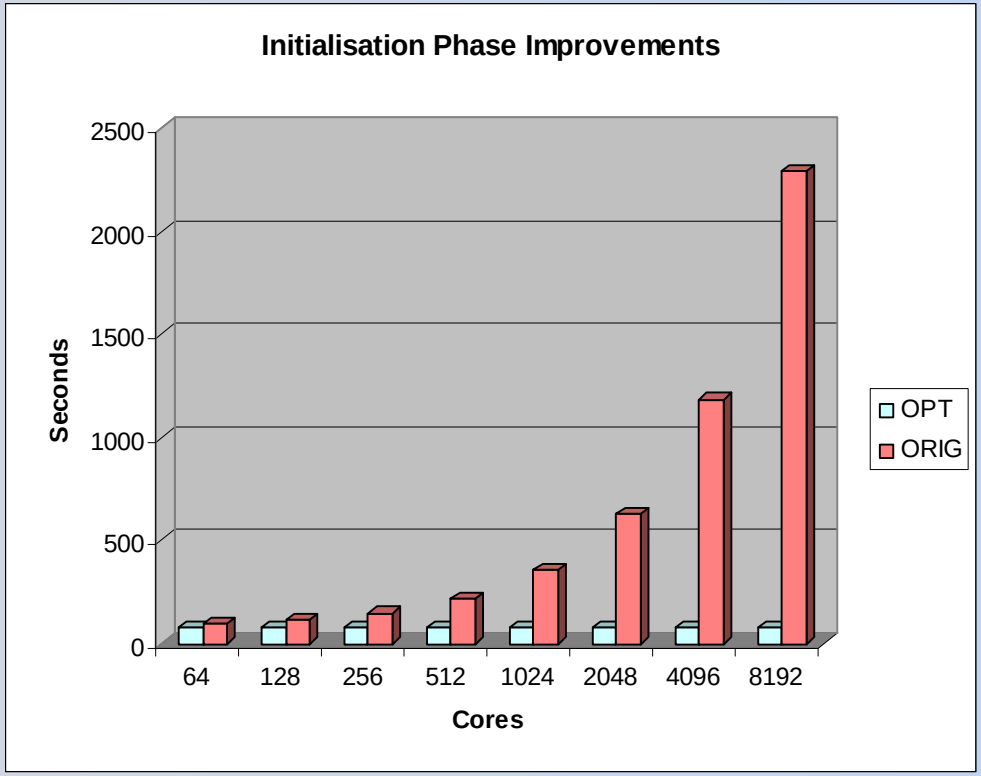
- UK application code.
- Has not been used at scale before!
  - The large dataset runs out of work at 2048 cores.
  - The other modes of use scale less well but this is expected as it involves serialization steps.





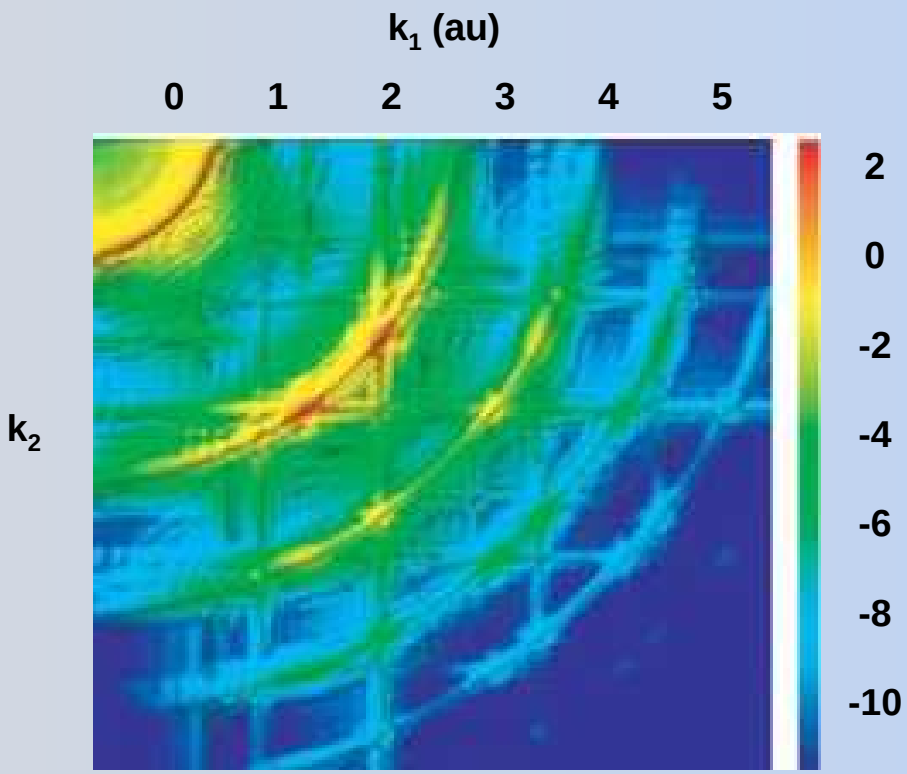
HemeLB (3/3)

- Startup phase was prohibitive to benchmarking and optimization at large processor counts
- IO Optimization performed to stop growth in time
  - Not so useful below 64 processors
  - Cost amortized in medium sized runs
  - This section have never really been examined before



HELIUM

- Solves time-dependent Schrodinger equation in full dimensionality
- Used to model interaction between an intense linearly polarized laser light and the Helium atom
- Highly optimized for HPCx
  - Six months were spent re-engineering the code specifically for this platform
- Largest problem on HPCx – 1200 processors
  - 50% of time is spent on communication
- Initial simulations on HECToR – 2048 processors
  - 5% of time is spent on communication
- The code authors are now preparing to do simulations at the quantum wavelength, which are not possible on HPCx

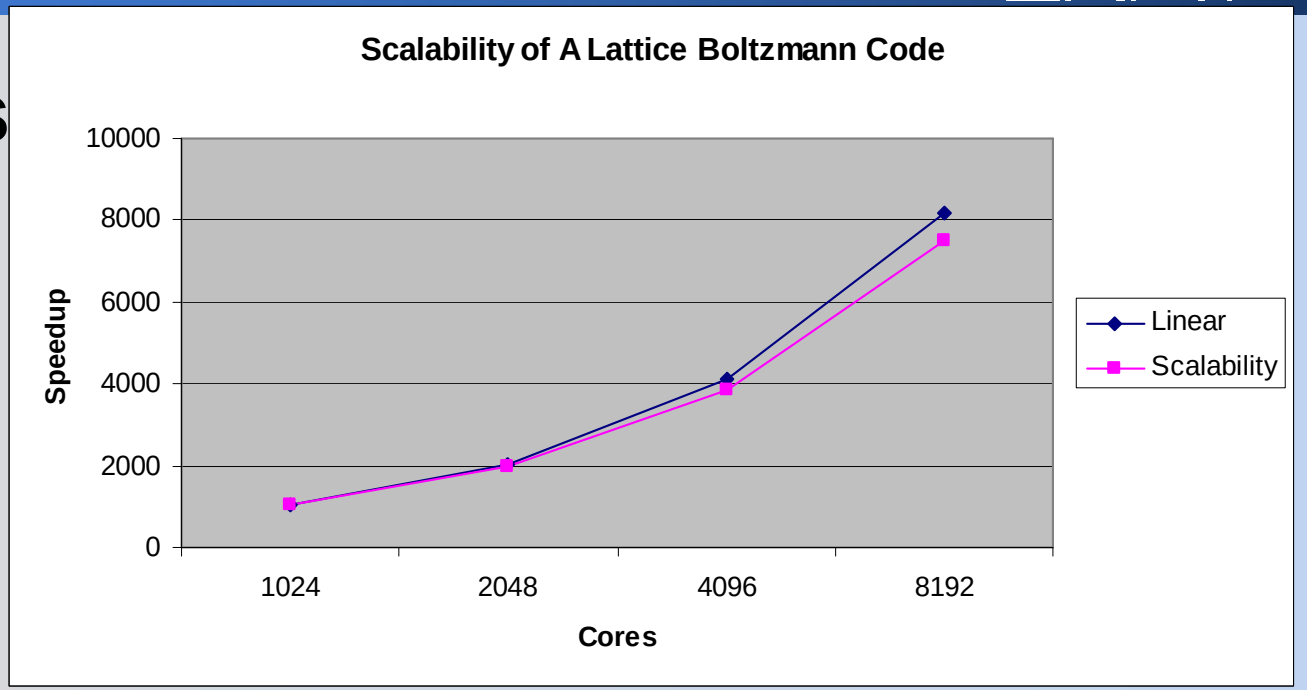


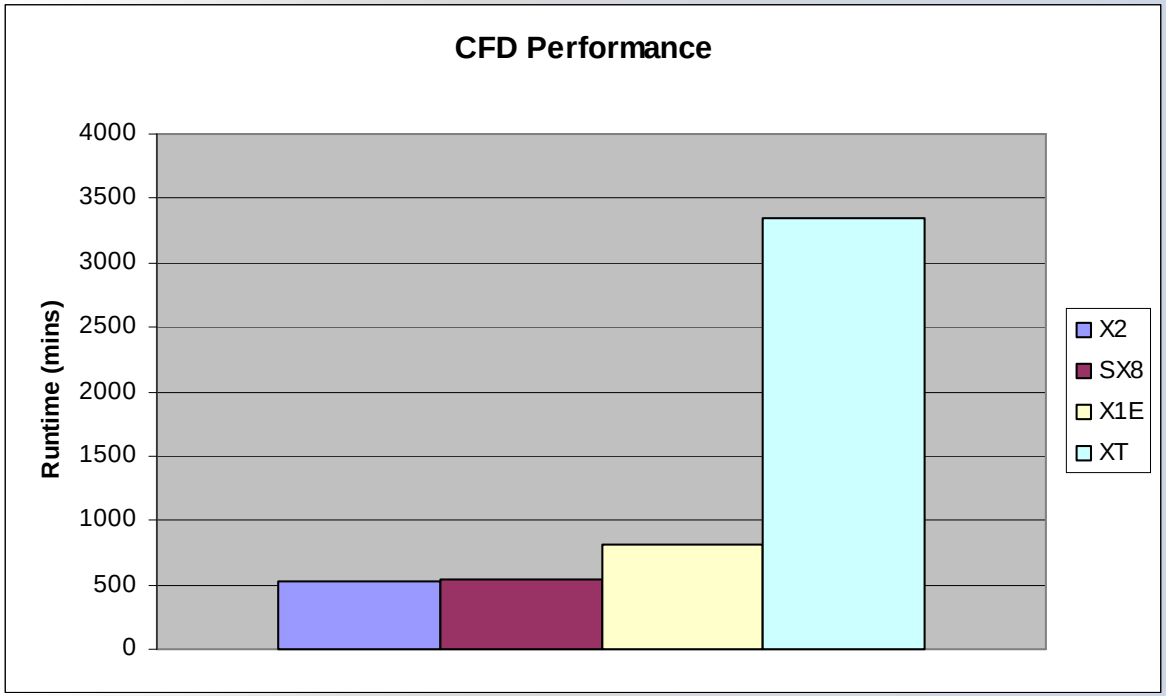
Quantum-mechanical state of helium prepared by a short intense laser pulse.

"I haven't seen anything this nice since the Cray T3D/E."

Jonathan Parker (Owner of one of HECToR's highest scaling codes)

Other Results





Conclusions and Future Activities

- Vector Workshop
- Concerted Vector code effort
- Continued XT effort – see job mix moving up the curve
- Mid next year begin preparing for quad-core