HECToR

HIGH END COMPUTING TERASCALE RESOURCE

A Research Councils UK High End Computing Service

DCSE WS 2009: Improving Parallel Performance of GLOMAP Mode MPI

Mark Richardson

Numerical Algorithms Group Ltd

mark.richardson@nag.co.uk

HECToR

RESEARCH
COUNCILS UK

Personnel with input to the project

- NCAS
  - Prof. Carslaw
  - Dr. Graham Mann
- SEE
  - Prof. Martyn Chipperfield
  - Dr. Steven Pickering
- NAG Ltd CSE team
  - Mark Richardson
  - HECToR Support CSE team
- Cray CoE

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

UNIVERSITY OF LEEDS

nag

HECToR

RESEARCH
COUNCILS UK

Overview

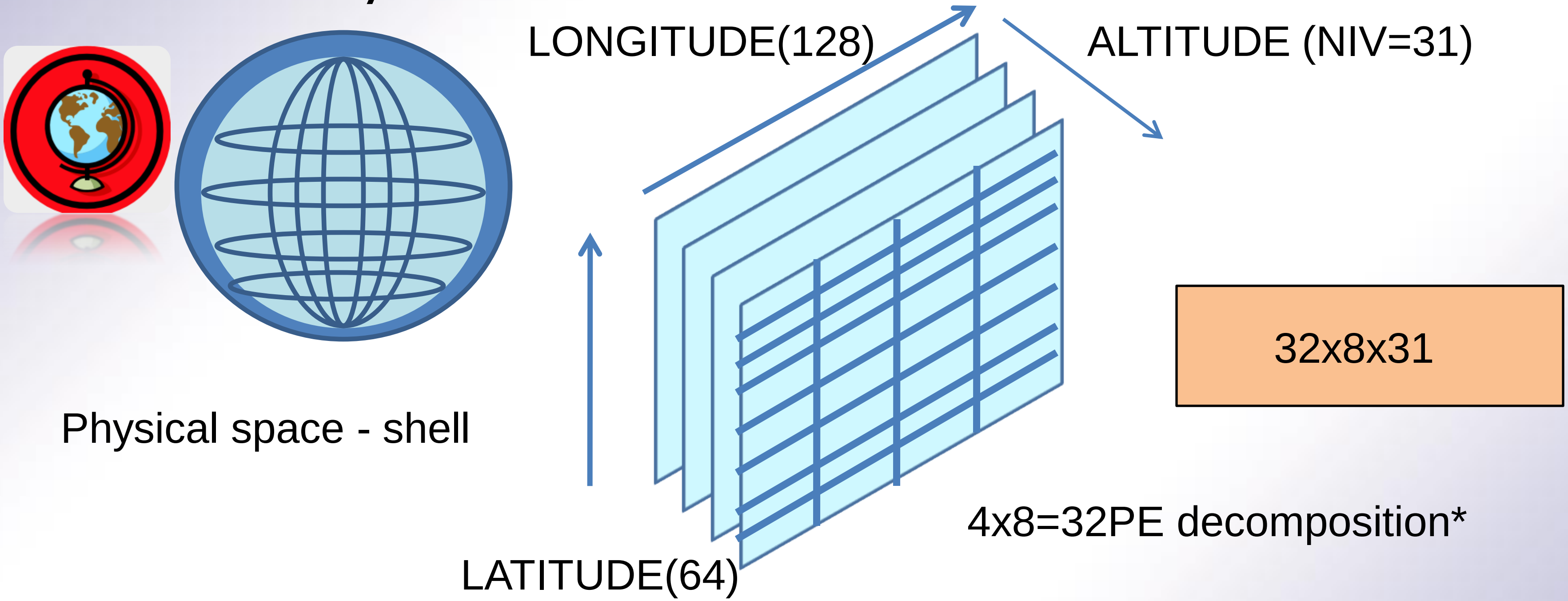
- Expect to give you an insight into some of the auxiliary effort needed to get the best use of HECToR
- Presented as a case study of GloMAP
  - Global Model of Aerosol Processes
- Follow three lines of investigation
  - Compiler options
  - Code structure
  - Parallel performance

The GloMAP simulation components

- TOMCAT advection code
  - Rectangular coordinate system for the numerical scheme
  - Mapping longitude, latitude and altitude
  - Resolution of this case T42 (128x64x31)
- GLOMAP chemistry University of Leeds
  - Per “gridbox” aerosol process model (>250 scalars)
  - Mode and Bin schemes (this project uses mode)
- ASAD from Cambridge
  - Numerical method for atmospheric chemical reactions
  - Time integration and user defined chemistry

Map physical space into computational space

- The GloMAP simulates decades of atmospheric chemistry



Recti-linear computational space

This project was originally 12 months.

- The project had been reduced to 6 months
- focus was on the shorter term goal of first 4 tasks
  - analyse GloMAP Mode MPI to provide a plan for enhancing its performance.
  - general code optimizations
  - MPI communication efficiency.
  - analyse the file handling and recommend a plan for parallel I/O to avoid the bottleneck of the MASTER-I/O model.

HECToR

RESEARCH COUNCILS UK

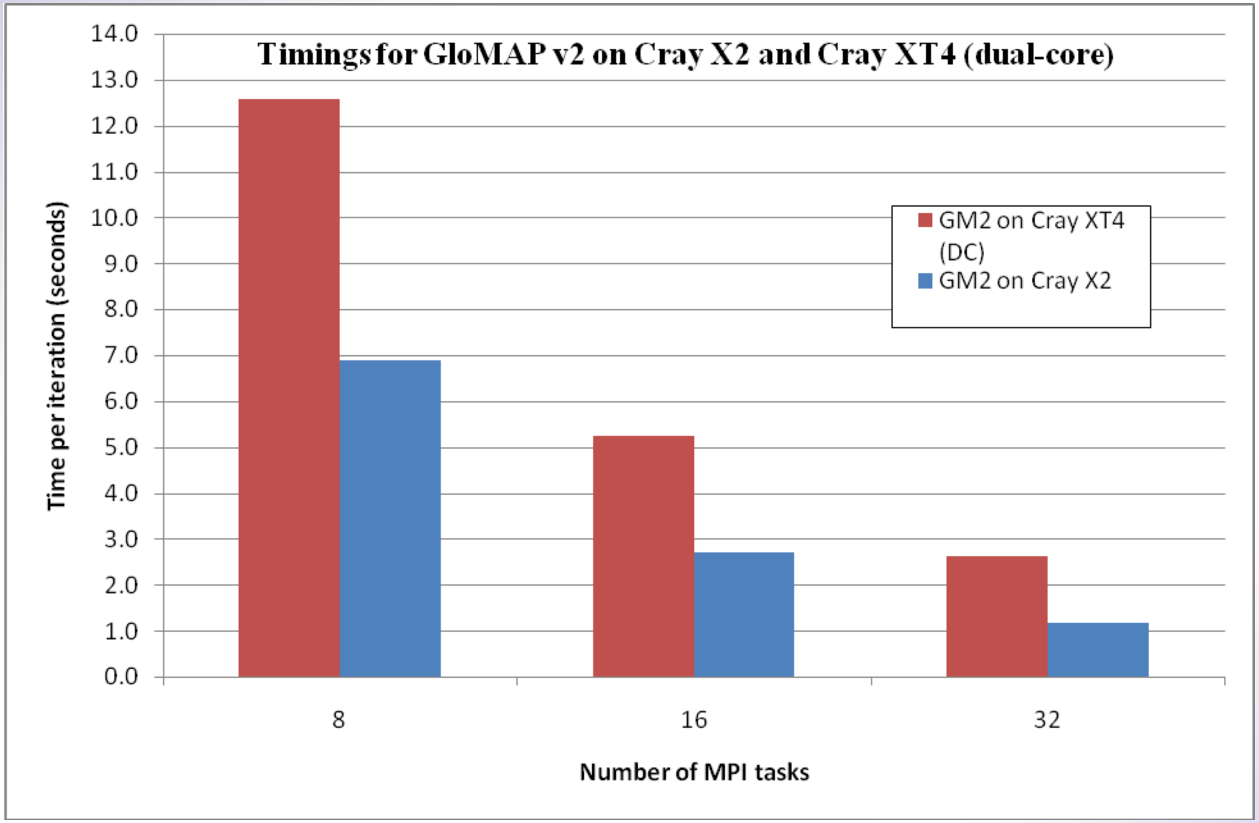
The GloMAP Working practice

- One large script with several sections
  - PBS directions
  - Shell commands, initialise variables
  - “here doc” TOMCAT updates (users work here)
  - “here doc” ASAD updates (users work here)
  - NUPDATE (serial process to create prog.f)
  - Compile glomap.exe (serial process)
  - Copy files (set up case directory)
  - APRUN (launch parallel program)
  - Post process (double to single)
- Strength is
  - that researchers have to make changes only to the “here doc” sections

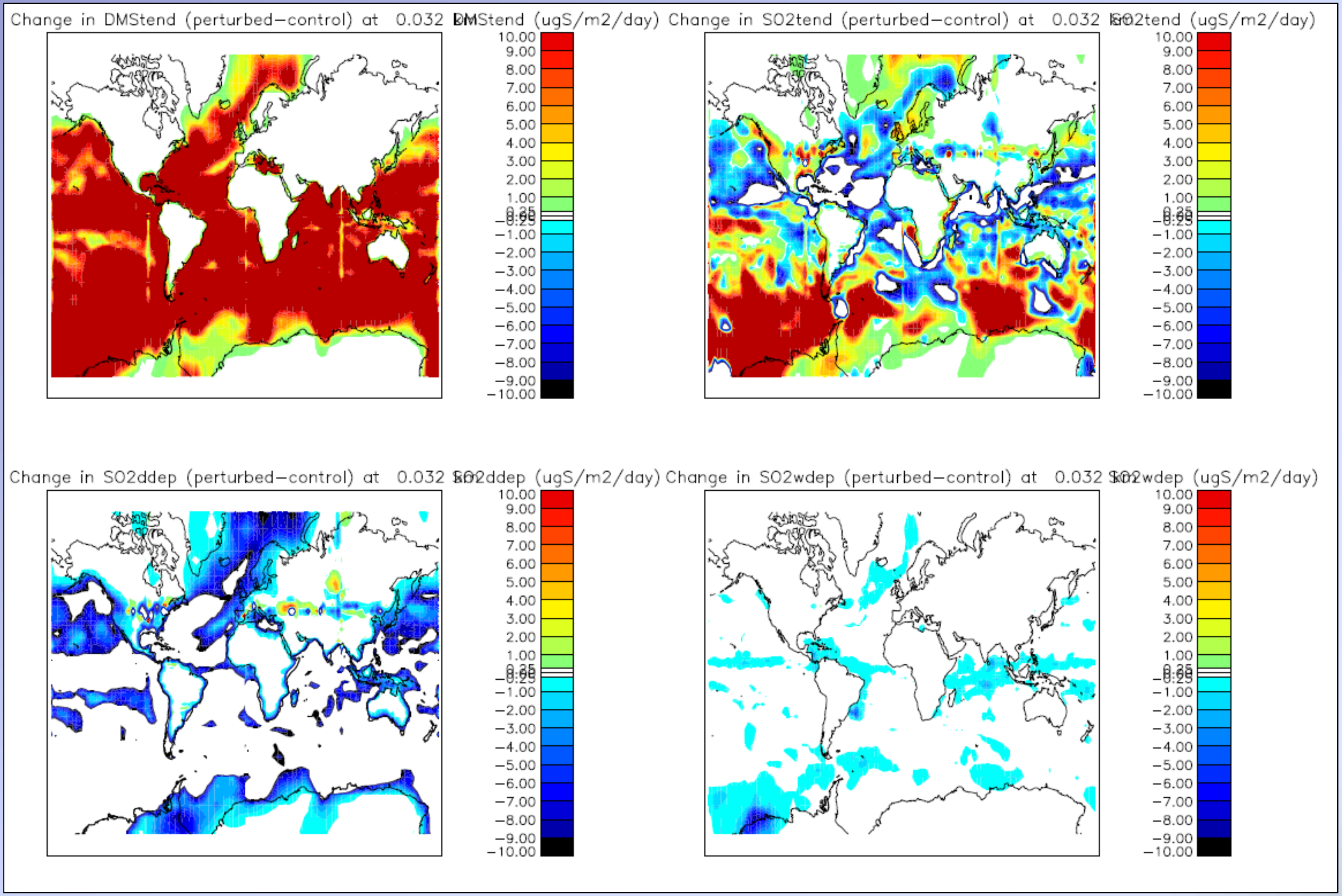
Porting to Cray X2

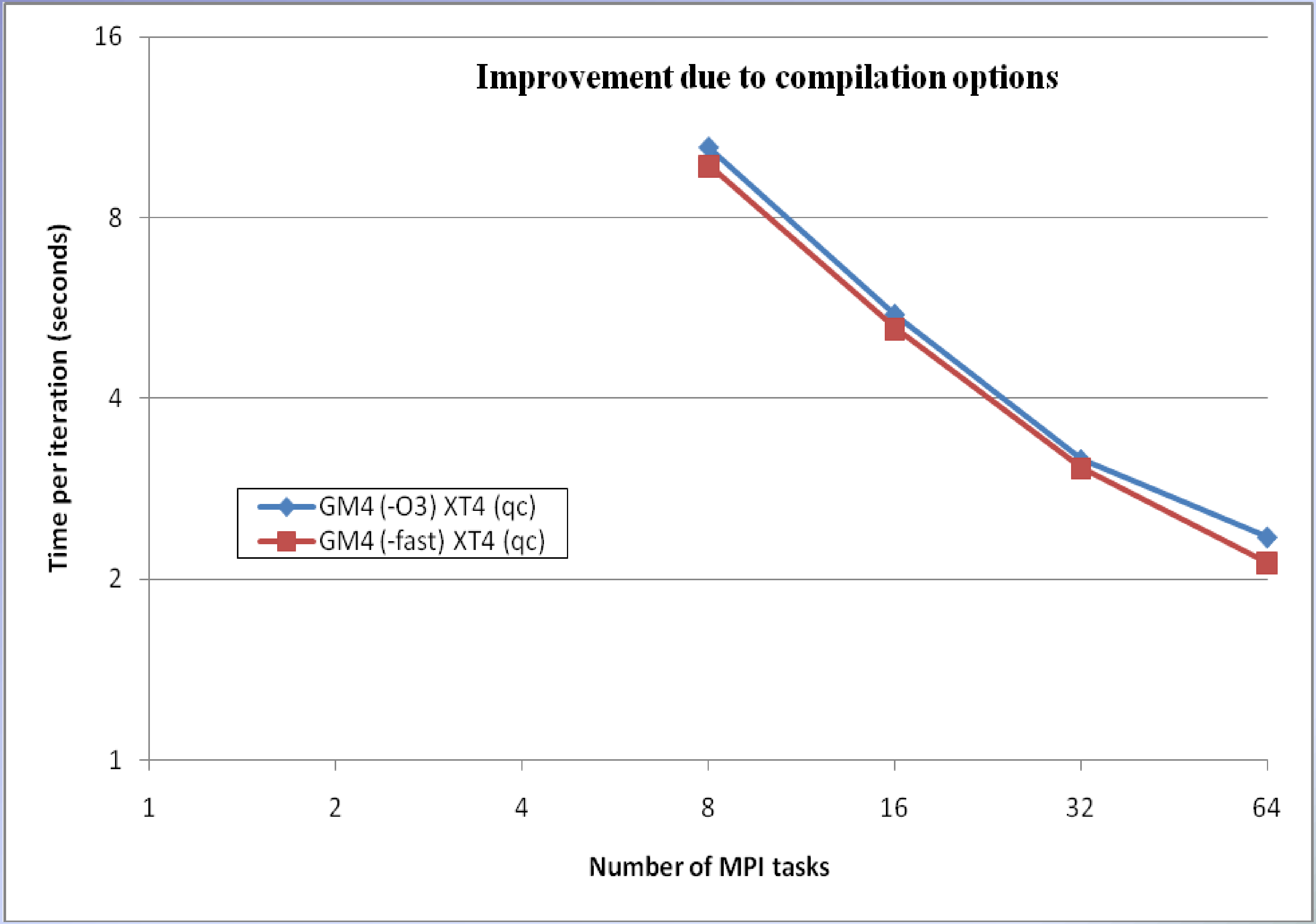
- Primarily used on HPCx with Open MP
- Code "already" vectorised
- Ported to HECToR XT4 using PG Fortran
- Some history of MPI implementation

Porting the code to Cray X2 vector system



Need to continually check quality of solution




Improvement due to compilation options
Time per iteration (seconds)
16
8
4
2
1
Number of MPI tasks
1
2
4
8
16
32
64
GM4 (-O3) XT4 (qc)
GM4 (-fast) XT4 (qc)

Analysing code structure

- Determine where the code is slow
- Use Cray PAT
- Read code (guided by CrayPAT and grep)
- Discussions with code owners (why?)

Challenge of sampling experiments

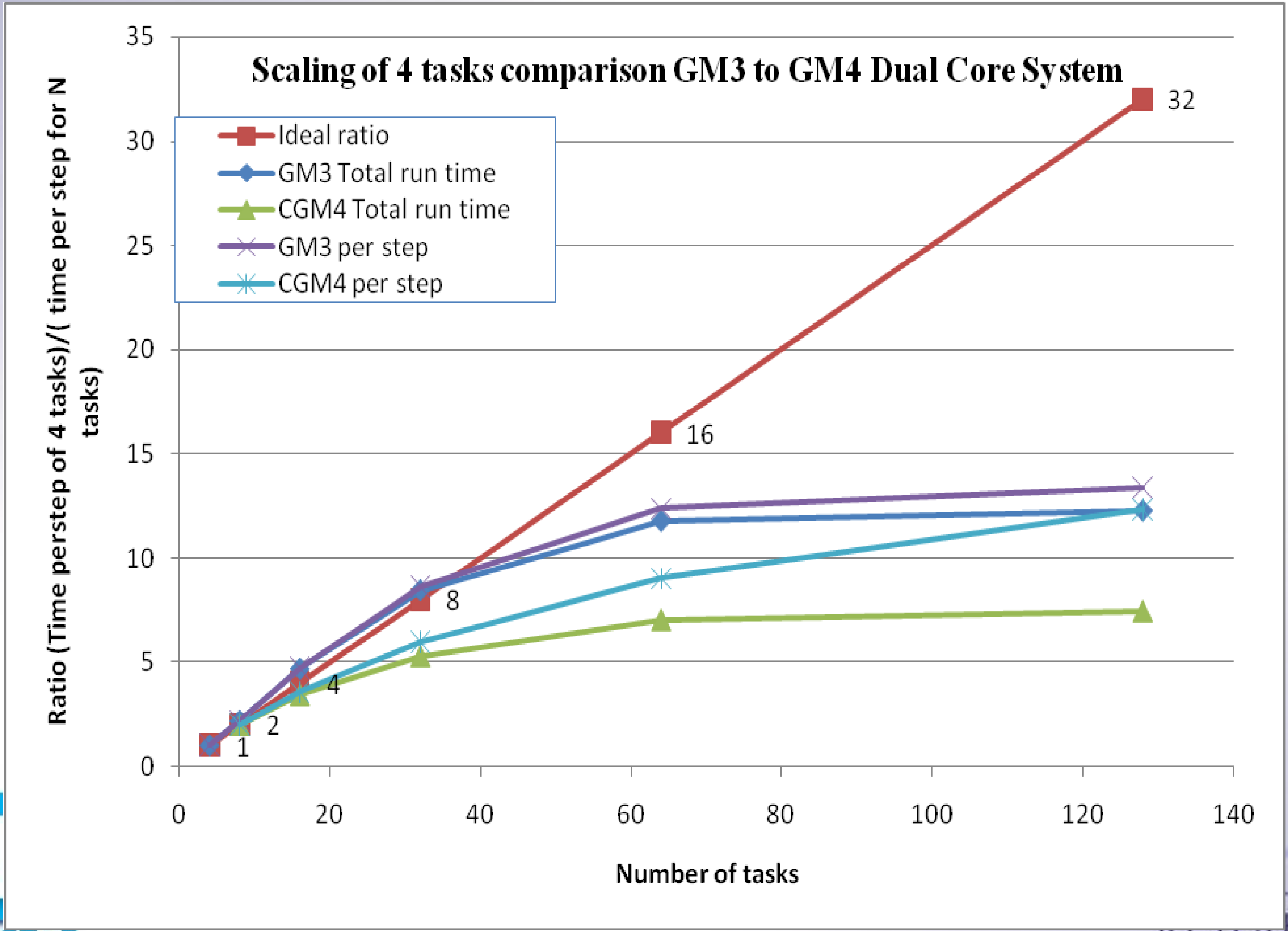
- How do you know you have not "quantised" the data?
  - Might be hitting a harmonic – so use trace to confirm
- Sampling for 8PEs gives higher resolution than for 64PEs (need to modify sample rate)
- Perhaps only useful for the rough guide

GM3 MPI sample experiment for 8PE (13s per iteration) and 64PE (2s per iteration)

```
GM3 (Cray XT4 Dual Core) PAT sample experiment 8PEs
 Samp % |    Samp |     Imb. |    Imb. |Group
        |         |     Samp | Samp % | Function
        |         |          |        |  PE='HIDE'
|----------------------------------------------
|  79.8% |  98686 |       -- |      -- |USER
||---------------------------------------------
||  27.3% |  33702 | 109.25 |    0.4% |chimie_
||   8.8% |  10857 | 174.38 |    1.8% |ukca_coagwithnucl_
||   6.0% |   7360 |  60.25 |    0.9% |advy2_
||   3.9% |   4795 | 238.50 |    5.4% |consom_
||   3.5% |   4364 |  29.88 |    0.8% |advz2_
||   3.2% |   3956 |  59.12 |    1.7% |advx2_
||   2.4% |   2945 |  90.12 |    3.4% |ukca_water_content_v
||   2.1% |   2586 | 169.75 |    7.0% |ukca_conden_
||   2.0% |   2448 |  13.50 |    0.6% |ukca_coag_coff_v_
||   1.8% |   2256 |  73.88 |    3.6% |ukca_solvecoagnucl_v
||   1.8% |   2171 |  79.12 |    4.0% |ukca_cond_coff_v_
||   1.6% |   2016 | 103.00 |    5.6% |ukca_volume_mode_
||   1.6% |   2003 |  50.38 |    2.8% |prls_
||   1.3% |   1583 | 110.62 |    7.5% |jac_
||   1.0% |   1274 |  63.75 |    5.4% |emptin2_
||   1.0% |   1188 |  34.00 |    3.2% |initer_
||=============================================
|  17.4% |  21498 |       -- |      -- |ETC
||---------------------------------------------
||   7.9% |   9808 | 309.00 |    3.5% |__c_mzero8
||   2.6% |   3212 |  83.75 |    2.9% |__c_mcopy8
||   1.1% |   1369 |  61.62 |    4.9% |__fmth_i_dexp
||=============================================
|   2.8% |   3466 |       -- |      -- |MPI
||---------------------------------------------
||   1.3% |   1587 | 584.38 |   30.8% |mpi_sendrecv_
||   1.0% |   1264 | 532.00 |   33.9% |mpi_recv_
|==============================================
```

```
GM3 (Cray XT4 Dual Core) PAT sample experiment 64PEs
 Samp % |   Samp |     Imb. |    Imb. |Group
        |        |     Samp | Samp % | Function
        |        |          |        |  PE='HIDE'
|----------------------------------------------
|  39.1% |  8647 |       -- |      -- |USER
||---------------------------------------------
||    5.3% |  1179 |  107.09 |    8.5% |advy2_
||    3.9% |   871 |   38.41 |    4.3% |chimie_
||    3.5% |   781 |   40.91 |    5.1% |ukca_coagwithnucl_
||    2.7% |   601 |   15.22 |    2.5% |advz2_
||    2.7% |   589 |   10.84 |    1.8% |consom_
||    2.3% |   512 |  270.48 |   35.1% |advx2_
||    1.6% |   348 |  112.12 |   24.8% |emptin2_
||    1.4% |   312 |   50.08 |   14.1% |ukca_water_content_v_
||    1.3% |   297 |  160.19 |   35.6% |fillin2_
||    1.3% |   279 |   66.77 |   19.6% |prls_
||    1.1% |   241 |   18.44 |    7.2% |ukca_coag_coff_v_
||    1.0% |   218 |  283.30 |   57.4% |spetru1_
||=============================================
|  32.2% |  7118 |       -- |      -- |MPI
||---------------------------------------------
||   15.8% |  3486 | 2032.61 |   37.4% |mpi_recv_
||   11.9% |  2638 | 2207.33 |   46.3% |mpi_sendrecv_
||    3.8% |   834 |  668.56 |   45.2% |mpi_ssend_
||=============================================
|  28.7% |  6352 |       -- |      -- |ETC
||---------------------------------------------
||    7.2% |  1595 |   90.95 |    5.5% |__c_mzero8
||    7.0% |  1548 |  421.45 |   21.7% |PtlEQPeek
||    1.9% |   429 |   54.33 |   11.4% |__c_mcopy8
||    1.8% |   395 |  139.33 |   26.5% |PtlEQGet
||    1.7% |   372 |  158.47 |   30.4% |PtlEQGet_internal
||    1.0% |   215 |   79.30 |   27.4% |ptl_hndl2nal
|=============================================
GM4 on 8 PEs (XT4 Dual Core) PAT sampling experiment report
 Samp % |  Samp |   Imb. |   Imb. |Group
        |       |   Samp | Samp % | Function
        |       |        |        |  PE='HIDE'
 100.0% | 73918 |     -- |     -- |Total
|---------------------------------------------
|  69.5% | 51363 |     -- |     -- |USER
||--------------------------------------------
||  10.9% |  8077 |  73.50 |   1.0% |advy2_
||  10.3% |  7580 | 202.38 |   3.0% |ukca_coagwithnucl_
||   6.6% |  4882 |  35.88 |   0.8% |consom_
||   6.4% |  4713 |  11.88 |   0.3% |advz2_
||   5.4% |  4012 |  66.00 |   1.8% |advx2_
||   2.9% |  2131 |  53.50 |   2.8% |ukca_water_content_v_
||   2.5% |  1811 |  97.25 |   5.8% |chimie_
||   2.4% |  1779 |  56.62 |   3.5% |ukca_conden_
||   2.2% |  1611 |  52.88 |   3.6% |ukca_calc_coag_kernel_
||   1.9% |  1418 |  30.38 |   2.4% |ukca_aero_step_
||   1.7% |  1273 |  21.00 |   1.9% |emptin2_
||   1.7% |  1254 | 219.25 |  17.0% |initer_
||   1.5% |  1143 |  17.75 |   1.7% |radabs_
||   1.3% |   939 |  43.00 |   5.0% |ukca_ddepaer_incl_sedi_
||   1.2% |   917 | 170.75 |  17.9% |fillin2_
||   1.2% |   875 |  67.75 |   8.2% |update_1dvars_by_cstep_
||============================================
|  26.5% | 19590 |     -- |     -- |ETC
||--------------------------------------------
||  11.1% |  8236 | 166.75 |   2.3% |__c_mzero8
||   3.6% |  2666 |  45.88 |   1.9% |__c_mcopy8
||   1.5% |  1093 | 421.00 |  31.8% |PtlEQPeek
||   1.3% |   937 |  44.50 |   5.2% |__fmth_i_dexp
||   1.0% |   729 |  41.38 |   6.1% |__fvdlog_long
||   1.0% |   715 |  61.25 |   9.0% |munmap
||============================================
|   4.0% |  2965 |     -- |     -- |MPI
||--------------------------------------------
||   1.7% |  1223 | 573.88 |  36.5% |mpi_recv_
||   1.6% |  1165 | 246.50 |  20.0% |mpi_sendrecv_
|=============================================
```

Change to code structure



HECToR

RESEARCH COUNCILS UK

Improve MPI communications

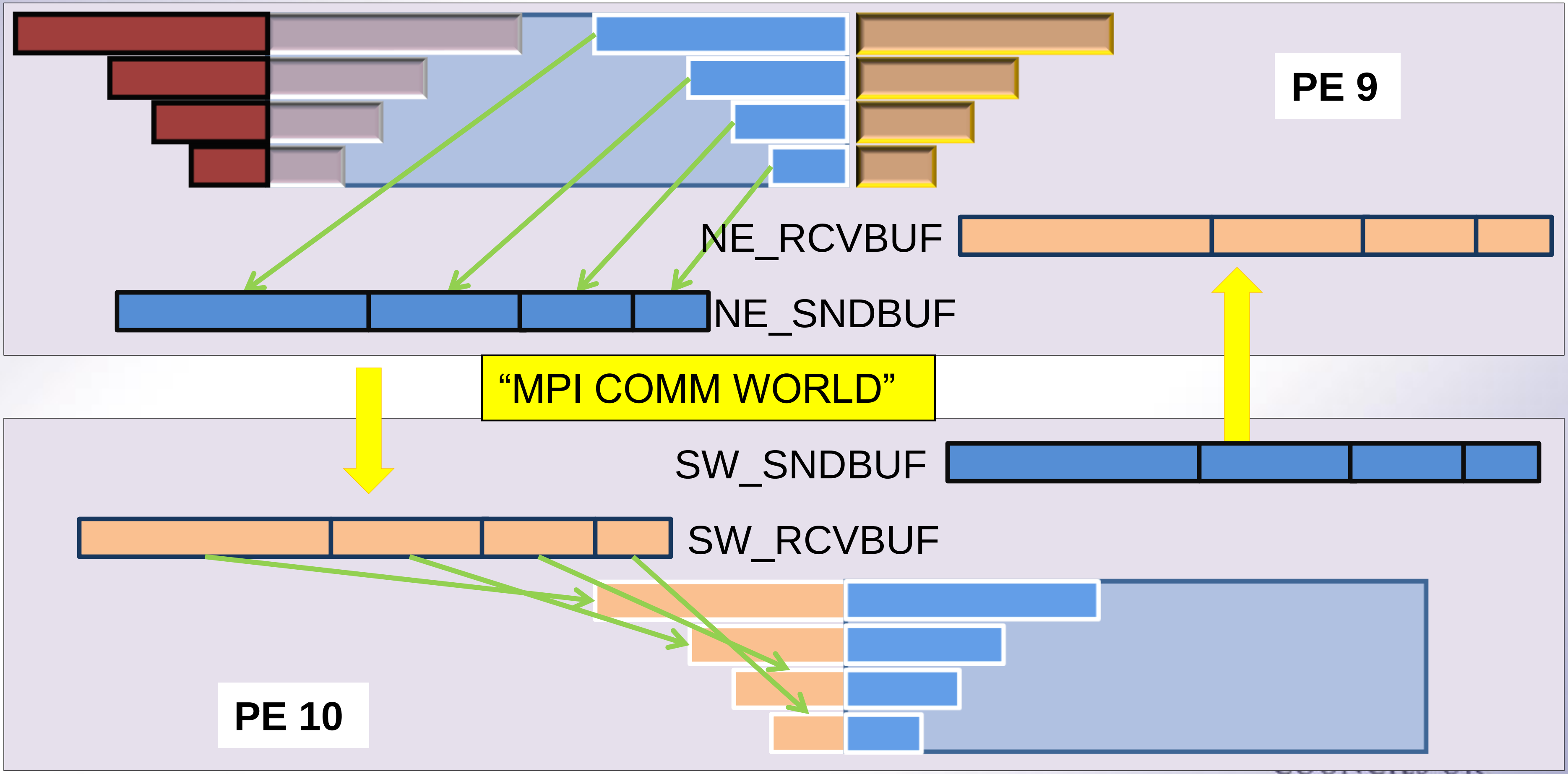
- Identify number of routines using send-receive pairs
  - 33 subroutines to visit – many different methods existed
- Read code, examine use of buffers
  - optimise filling buffers
- Observed a lot of MPI_BCASTS
  - Many associated with MASTER I/O requirement
- Too much global data
- Too much static memory

HECToR

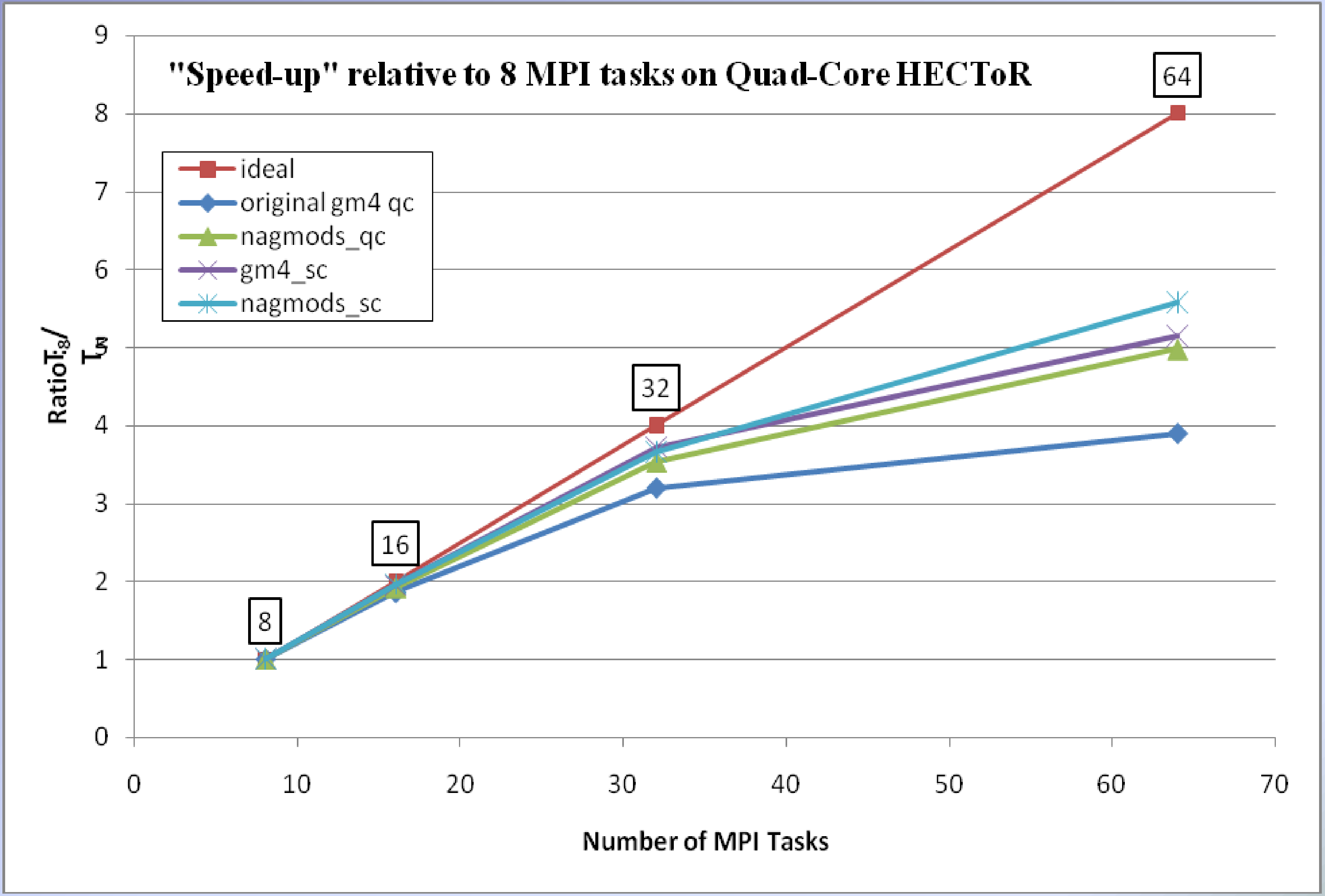
RESEARCH COUNCILS UK

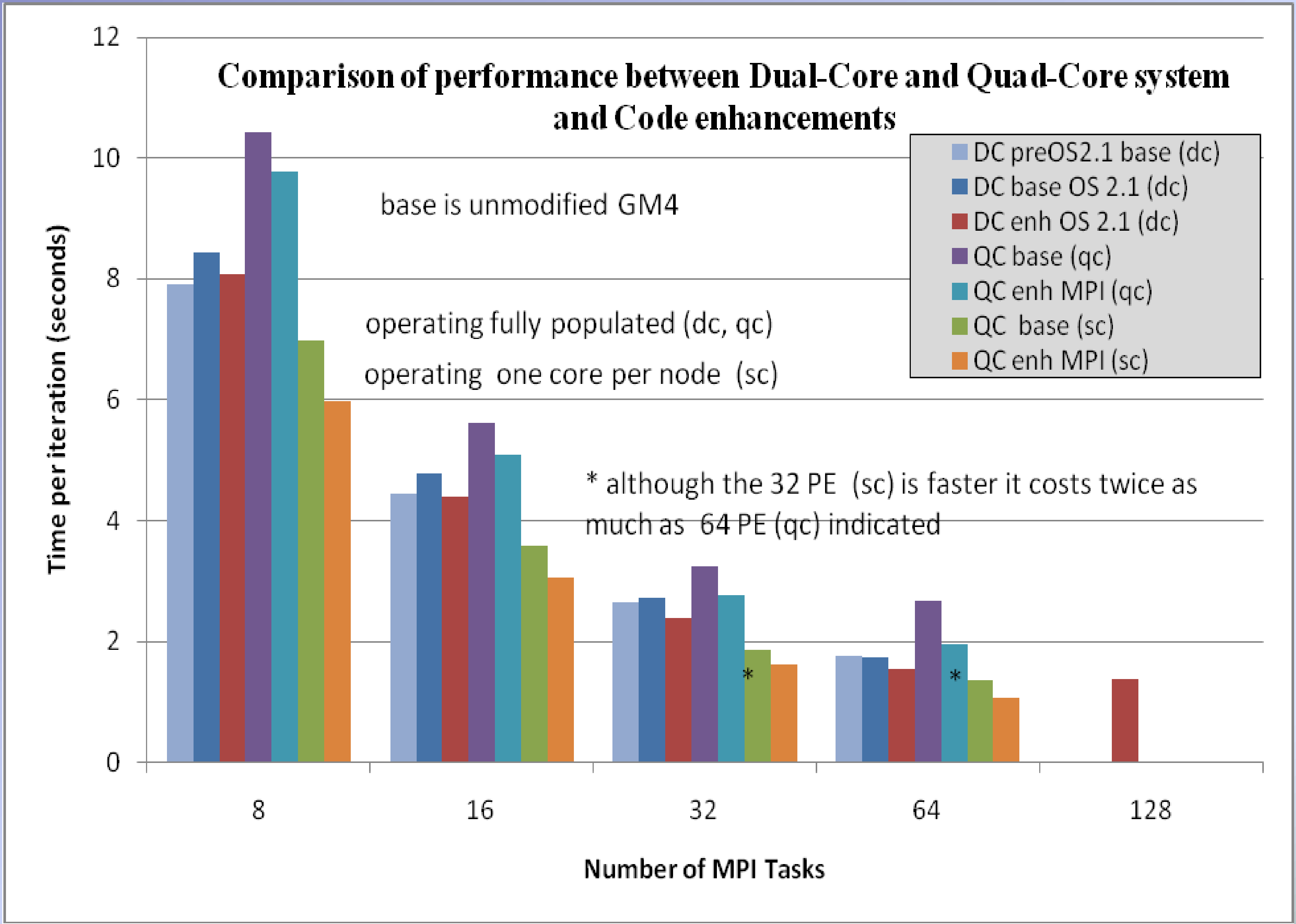
FILLIN2-LISTCOMM-EMPTIN2 process

- West and East halo and shadows (K=2)



Effect of communication enhancements





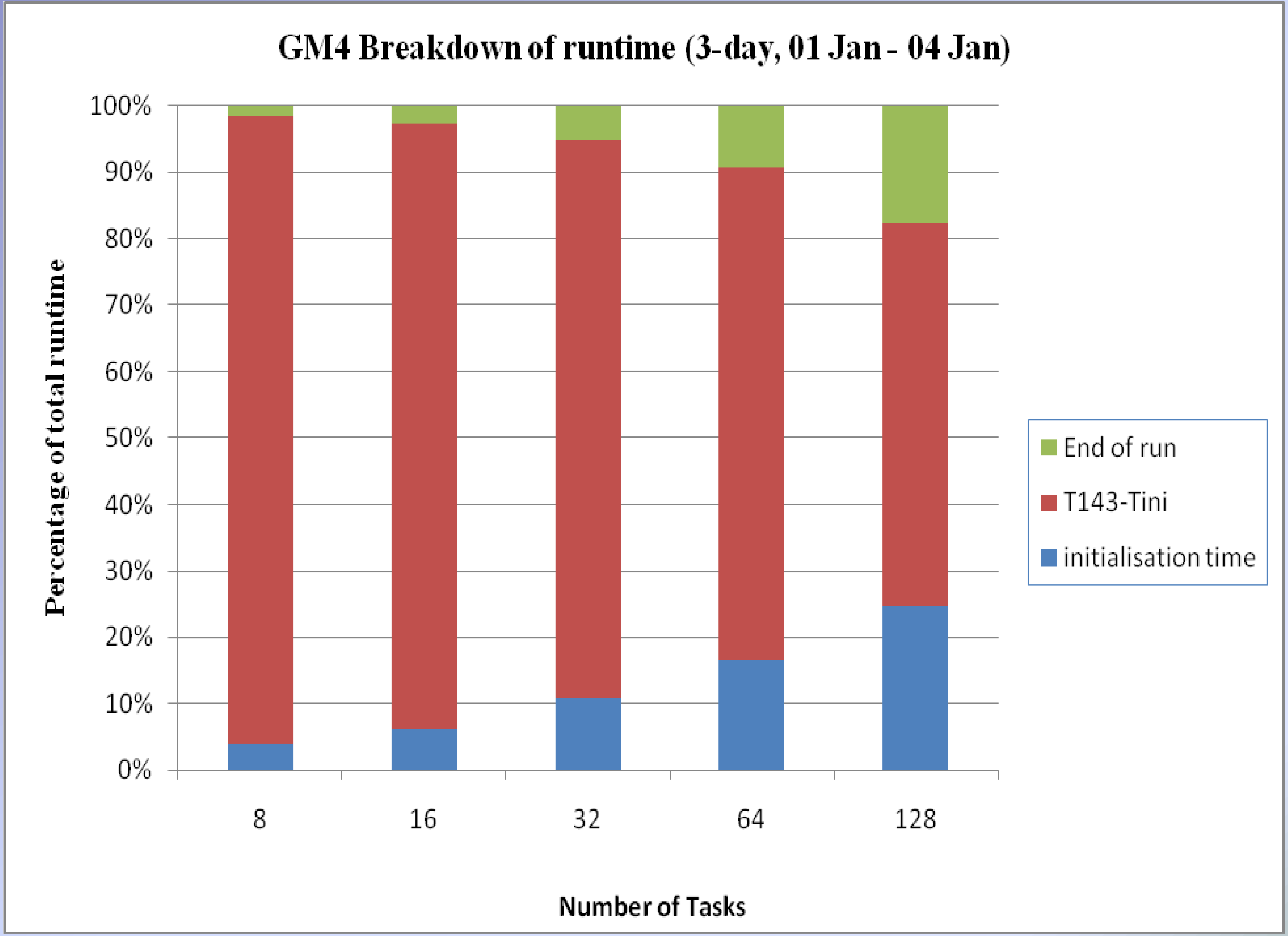
Comparison of performance between Dual-Core and Quad-Core system and Code enhancements

base is unmodified GM4

operating fully populated (dc, qc)

operating one core per node (sc)

* although the 32 PE (sc) is faster it costs twice as much as 64 PE (qc) indicated


GM4 Breakdown of runtime (3-day, 01 Jan - 04 Jan)
Percentage of total runtime
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
8
16
32
64
128
Number of Tasks
End of run
T143-Tini
initialisation time

Results 1

Improvement due to changes in code structure, dual core system

Number of MPI Tasks	8	16	32	64
GM3 (DC -O3)	1952	872	451	276
GM4 (DC -fast)	1122	631	377	251
Improvement %	*42.48*	*27.62*	*16.26*	*9.08*

Time in seconds for simulation omitting first and final steps

Improvement due to changes in compiler optimization, quad core system

Number of MPI Tasks	8	16	32	64
GM4 (QC -O3)	1485	783	449	334
GM4 (QC -fast)	1387	742	434	302
Improvement %	*6.60*	*5.24*	*3.34*	*9.58*

Results 2

Improvement due to MPI enhancement				
Number of MPI Tasks	**8**	**16**	**32**	**64**
GM4 (-fast)	1387	742	434	302
GM4 (-fast) with MPI enhancement	1389	723	393	279
Improvement over GM4 baseline %	*-0.14*	*2.56*	*9.44*	*7.61*
Time in seconds for simulation omitting first and final steps				

Overall improvements (including previous optimisations)				
Number of MPI Tasks	**8**	**16**	**32**	**64**
GM4 (-O3)	1485	783	449	334
GM4 MPI enhancement	1389	723	393	279
Improvement over GM4 baseline %	*6.46*	*7.66*	*12.47*	*16.47*
Time in seconds for simulation omitting first and final steps				

HECToR

RESEARCH COUNCILS UK

Conclusions

- The code structure was revised to enhance cache usage
- Some coding errors were revealed by:
  - Cray X2 compiler
    - +subsequently NAG X86_64 compiler
  - Cray PAT
  - Code reading
  - Difference tool
- Improvement in buffer loading and unloading
  - Led to improvement in parallel performance

HECToR

RESEARCH COUNCILS UK

Recommendations

- Recommendations have been made for further improvement
  - MPI-IO will lead to;
    - Reducing BCASTS
    - Reduce memory; better use of cache
  - Re-use of buffers
    - will reduce memory requirement
- Some that have not been investigated
  - Pre-posting receives

Current work planned

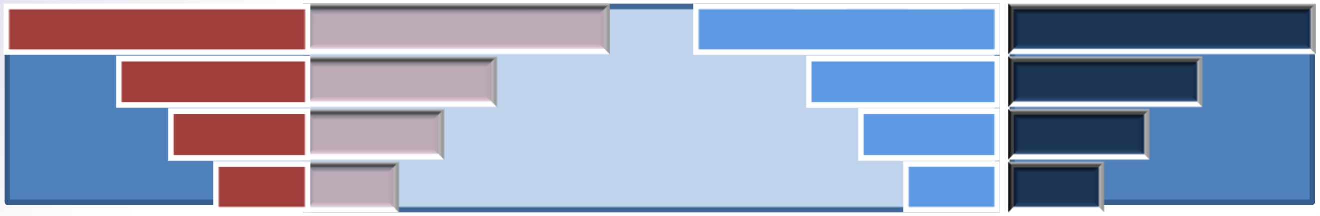
- Mixing MPI and Open MP
  - Can see that running single core per node gives advantages
  - Using Open MP will enhance that performance
    - E.g. If ¼ under-populate gives 2x speed-up and the inefficient SMP speed-up of 2.5 on 4 cores will result in a speed-up of 1.25 of the solely MPI version.

East West Communication pattern

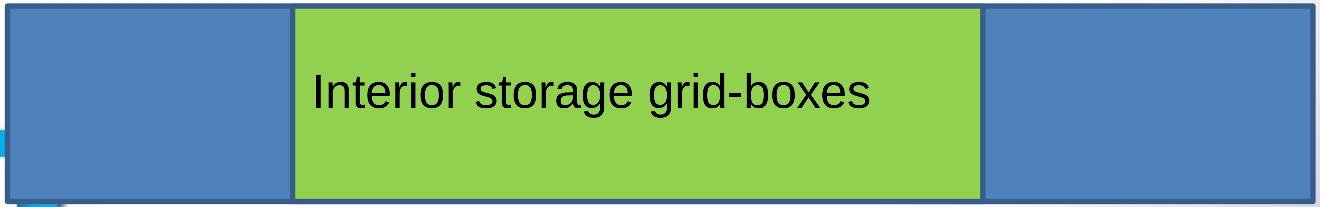
(if discussion requires it)

Halo data structure on one domain

- West and East halo and shadows



Required storage is excessive


Interior storage grid-boxes