GloMAP mode on XE6

HECToR DCSE Technical Meeting October2011

Mark Richardson, HECToR CSE Team,
Numerical Algorithms Group

Outline of this Case Study

- Enhance the existing MPI version of GLOMAP MODE with Open MP directives
- Enable the mixed-mode of parallel operation for better use of the multi-core systems that are being installed
- Allow higher resolution simulations

HECToR

RESEARCH COUNCILS UK

What is GLOMAP?

- A computer program for simulating aerosol processes in the earth's atmosphere
  - TOMCAT is an advection code and is the main program
    - Reads wind data and transports the chemistry around the atmosphere
    - Maintained and Supplied by Professor Martyn Chipperfield, University of Leeds
  - GLOMAP Mode, the aerosol process method
    - Replaces the built-in chemistry model of TOMCAT (the subroutine "CHIMIE")
    - Developed at University of Leeds by Dr. Graham Mann, NCAS
  - ASAD the chemical reaction solver (2004).
    - Dr. Glenn Carver, University of Cambridge

HECToR

RESEARCH COUNCILS UK

Acknowledgements

- NERC, NCAS
- Research Councils UK, HECToR Resource
- University of Leeds School of Earth and Environment
- HECToR DCSE programme
- NAG provide personnel
- Additional technical support
  - HECToR Service Helpdesk
  - Cray Centre of Excellence
  - Portland Group (on-line forum)

Background 1

- MPI version has been subject of two DCSEs to enhance its performance on HECToR
- DCSE 1 (6mm[1] Oct 2008-May 2009)
  - XT4 phase1a 2-core, phase1b 4-core
- DCSE 2 (4mm[1] Oct 2009-May 2010)
  - XT6 phase2a 24-core
- A solely Open MP version had been developed prior to the MPI version
  - Several years ago and completely separate from this project.

HECToR

[1] mm = man month

RESEARCH COUNCILS UK

Review objectives

- Enhance the existing MPI version of GLOMAP MODE with Open MP directives
- Enable the mixed-mode of parallel operation for better use of the multi-core systems that are being installed
- Allow higher resolution simulations

HECToR

RESEARCH COUNCILS UK

TABLE 1: Comparison of MPI and mixed-mode on XT4h

MPI tasks	16	32	64	128
XT4h GMM N4	4.227	2.174	1.426	1.085
XT4h GMH N1t1	2.696	1.498	0.826	0.665
XT4h GMH N1t2	1.692	0.979	0.58	0.574
XT4h GMH N1t4	1.337	0.735	0.489	0.473

HECToR

RESEARCH COUNCILS UK

Background 2

- Inauguration of XT6 with 24 cores per node at Edinburgh June 2010.
- A version of this was presented at CUG2010 with data from
  - XT6 phase2a 24core (i.e. No Gemini)
- As part of the CSE activity to monitor changes
  - XT6 phase2b provided e-LUSTRE
  - XE6 phase2c provided GEMINI

HECToR

RESEARCH COUNCILS UK

Case description 1

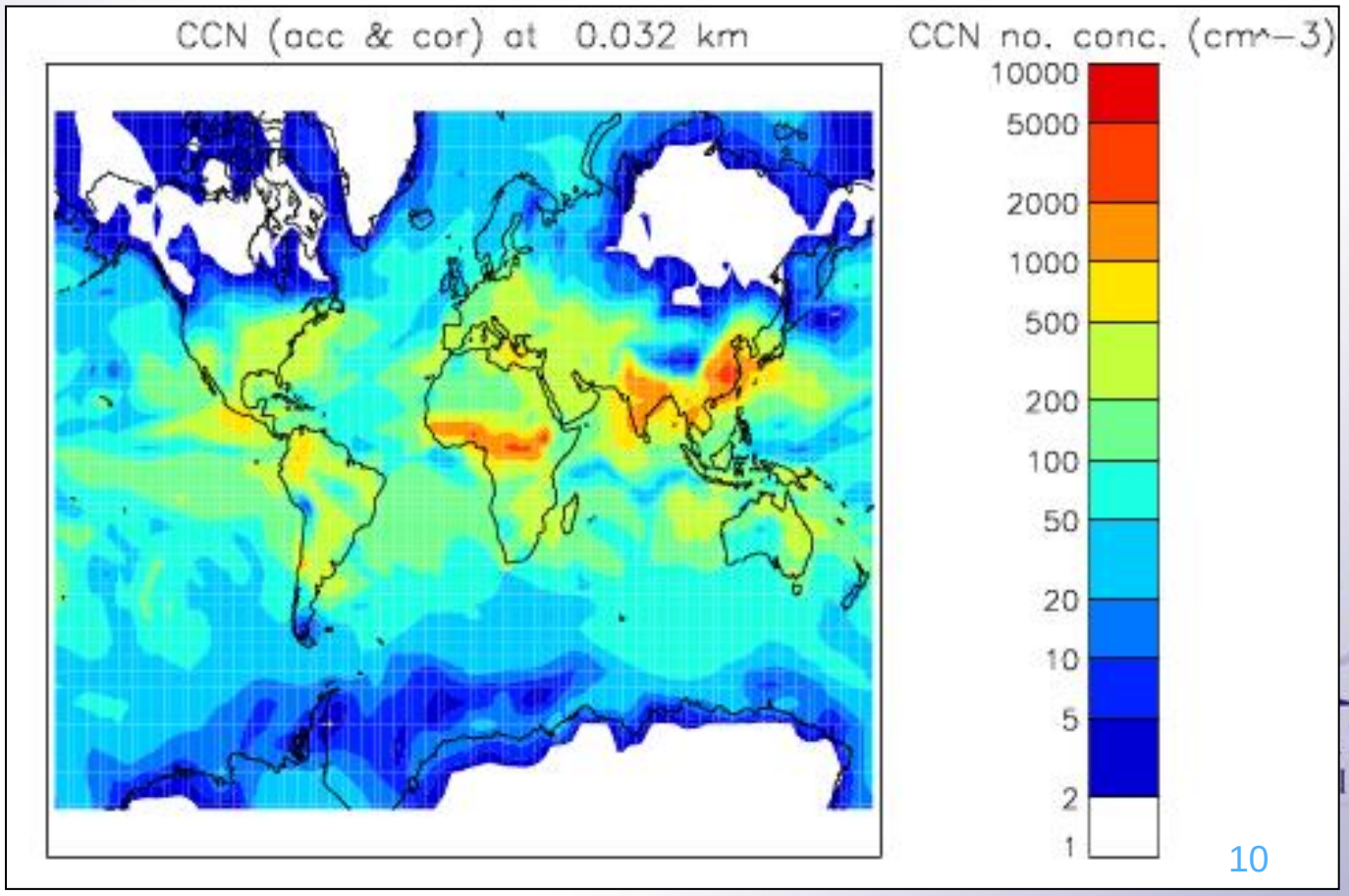
- The T42 model is a low resolution model
  - 128 x 64 x 31 i.e. 2.8^{o} per grid-box
- The chemistry requires significant memory
  - minimum 4 nodes of XT4h using 8 cores
  - i.e. hector phase 1, ~3GB per core
- The MPI topology is two dimensional (lon by lat)
  - but retains full altitude on each “patch”
- Typical use is 32 MPI tasks (each task has 32 x 8 x 31 grid-boxes)

HECToR

RESEARCH COUNCILS UK

Case description 2

- Earth's atmosphere mapped into a 3-D Cartesian coordinate system
- T42 is 128x64x31 and MPI 2-D topology creates uniform patches (e.g. 32 x 4 x 31)
- 197 scalars
- A 3 days simulation used for investigation (144 steps)
- Initial I/O stages omitted as they form ~30% of the elapsed time for this simulation



HECToR

Analysis 1: Cray PAT sampling experiments

Accumulator category	32	64	128
MPI	21.2	39.5	56.6
User	61.1	46.6	33.5
ETC	17.7	13.9	9.9

Units represent percentage of run time

MPI function, selected calls	32	64	128
mpi_barrier	4.7	9.9	17.1
mpi_bcast	9.8	14.7	15.7
mpi_sendrecv	4.2	8.3	14.9

HECToR

RESEARCH COUNCILS UK

Analysis 2: High workload subroutines

Identified five subroutines that dominate the samples:
ADVX2, ADVY2, ADVZ2, CONSOM, CHIMIE

Subroutine	GMM n32	GMM n32	GMM n64	GMM n64	GMM n128	GMM n128
ADVX2	6.2	10.15	4.8	10.30	3.4	10.15
ADVY2	7.0	11.46	5.2	11.16	4.2	12.54
ADVZ2	5.0	8.18	3.7	7.94	2.5	7.46
CONSOM	6.4	10.47	3.8	8.15	2.5	7.46
CHIME+UKCA*	22.3	**36.50**	16.1	**34.55**	11.3	**33.73**
Other subroutines	14.2	23.24	13	**27.9**	9.6	**28.66**
	61.1	**100.00**	**46.6**	**100.00**	**33.5**	**100.00**

•Still omitting between 40-70% so do not expect great changes (Amdahl effect)

• *UKCA == all occur within CHIMIE OMP loop, eight counted here

HECToR

RESEARCH COUNCILS UK

Analysis 3

- Target outer loops, OMP directives applied to 5 loops
- Loop counts and decomposition
  - Vary with the number of chosen MPI tasks
  - Limitations to scaling through Open MP

Loop Location	M32	M64	M128
ADVX2	31	31	31
ADVY2	31	31	31
ADVZ2	8	4	2
CONSOM	36	36	36
CHIMIE	8	4	2

Open MP loop count upper bound

HECToR

RESEARCH COUNCILS UK

Results 1: case runs on Cray XE6

- The simulation can be placed on the nodes in several configurations
- Two scenarios: packed and uniform
  - Packed
    - In this mode the cores on each node are filled as per the job scheduler default
    - It is likely that the final node is under-populated
    - No opportunity for Open MP due to packing
    - Minimum number of nodes in use for the chosen MPI decomposition
  - Uniform
    - in this situation the options to "aprun" are used to limit the placement of the MPI tasks, for example:
      - 4 MPI tasks per hex-core
      - thus a uniform distribution of 16 tasks per node
    - Every node is "under-populated" w.r.t. MPI tasks

HECToR

RESEARCH COUNCILS UK

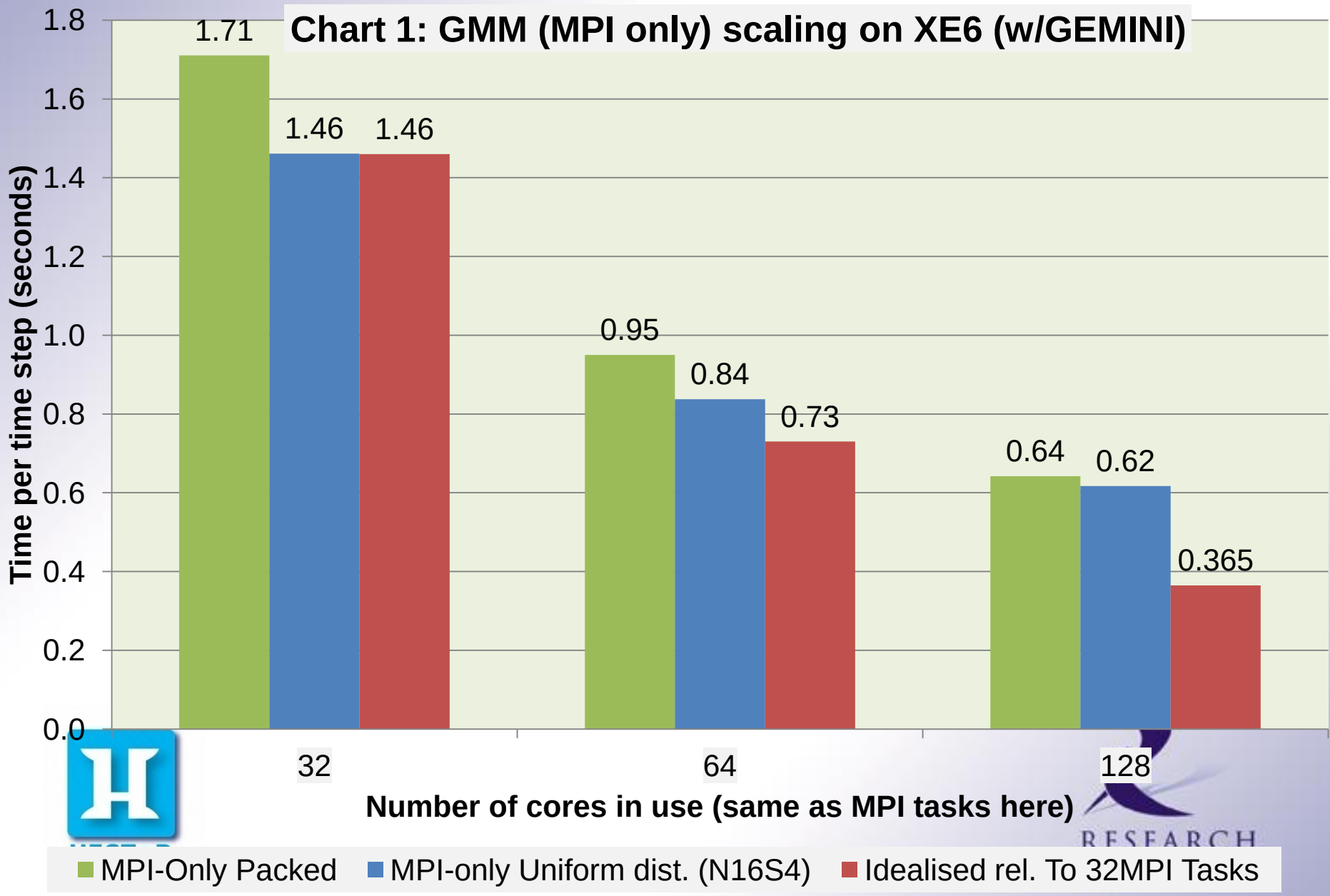


Chart 1: GMM (MPI only) scaling on XE6 (w/GEMINI)
Time per time step (seconds)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1.71
1.46
1.46
0.95
0.84
0.73
0.64
0.62
0.365
32
64
128
Number of cores in use (same as MPI tasks here)
MPI-Only Packed
MPI-only Uniform dist. (N16S4)
Idealised rel. To 32MPI Tasks

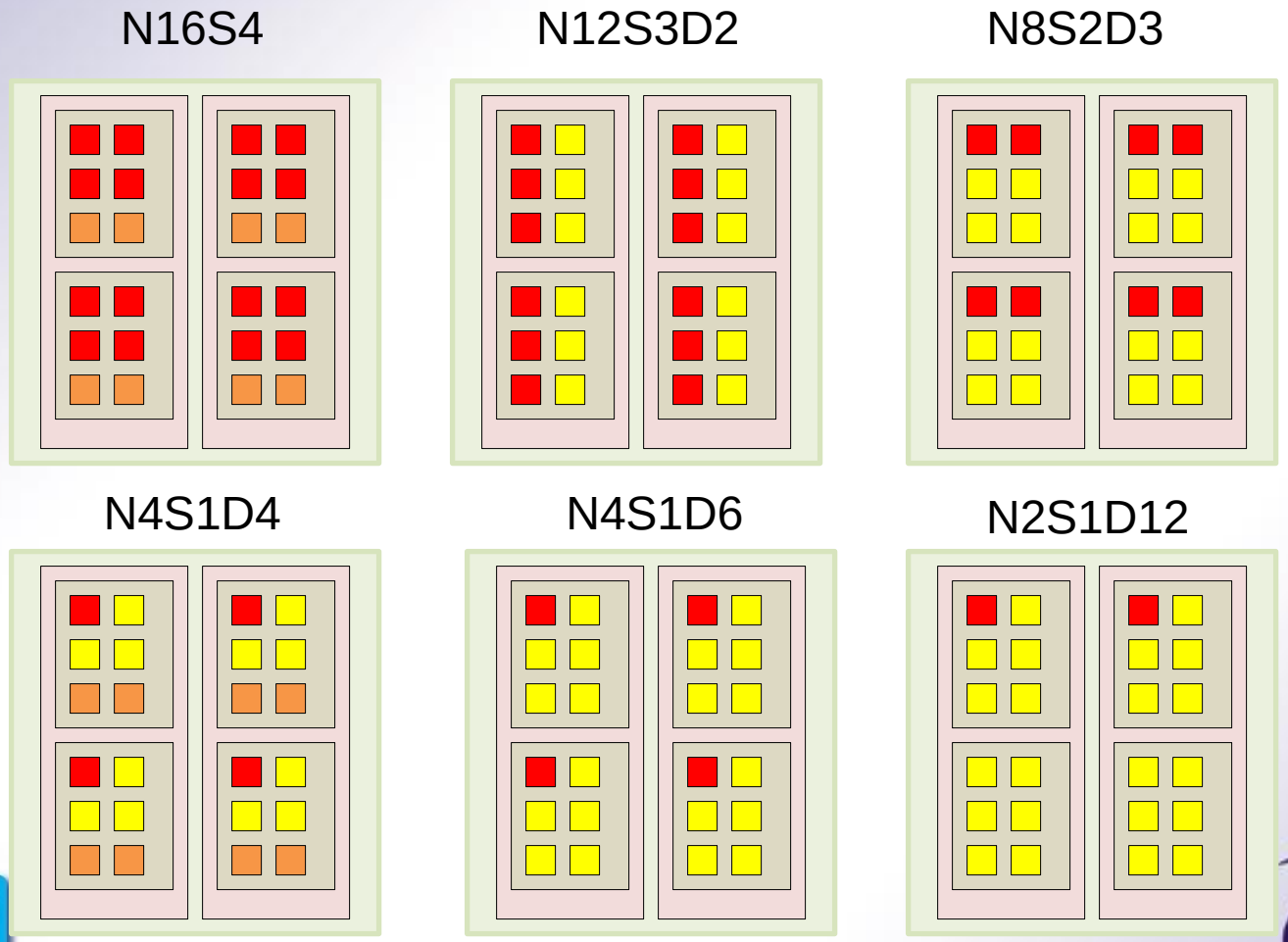
Mixed mode arrangements

- Consider how to place MPI tasks
  - Generally one would consider one MPI task per processor (Magny-Cours)
  - But probably better to consider on a per-hex-core basis
- How to activate OMP threads
  - Hex-core boundaries
- PBS job script
  - Reservation and claim
  - Per node

HECToR

RESEARCH COUNCILS UK

Diagrams of nodal arrangements



HECToR

RESEARCH COUNCILS UK

Resource allocation

```
#!/bin/bash --login
#PBS -N hyb_n64N2S1omp
# the PBS mpp values are to reserve enough resource
#PBS -l mppwidth=64
#PBS -l mppnppn=2
#PBS -l mppdepth=12
#PBS -l walltime=1:00:00
#PBS -v GMHPIO
# HECToR CSE nag account
#PBS -A z03
#
module list
echo "WORK is ${WORK} : GMHPIO is [ ${GMHPIO}  ]"
cd ${PBS_O_WORKDIR}
date > StartedJob.${PBS_JOBNAME}
# note the aprun command reflects exactly what you want
# the number MPI tasks is n; N is MPI per node; S is # tasks per NUMA Node
export OMP_NUM_THREADS=8
aprun  -n 64 -N 2 -d ${OMP_NUM_THREADS} ${GMHPIO}/src_64/xtgmh.exe > hyb_n64N2S1t${OMP_NUM_THREADS}
#
export OMP_NUM_THREADS=12
aprun  -n 64 -N 2 -d ${OMP_NUM_THREADS} ${GMHPIO}/src_64/xtgmh.exe > hyb_n64N2S1t${OMP_NUM_THREADS}
```

HECToR

RESEARCH COUNCILS UK

Results 2: charts

- Chart 2: 32 MPI tasks mixed mode
  - N16S4d1, N12S3d2, N8S2d3, N4S1d[4,6],N2d12
- Chart 3: 64 MPI tasks mixed mode
- Chart 4: 128 MPI tasks mixed mode
- Chart 5: Effect of mixed mode for several configurations

HECToR

RESEARCH COUNCILS UK

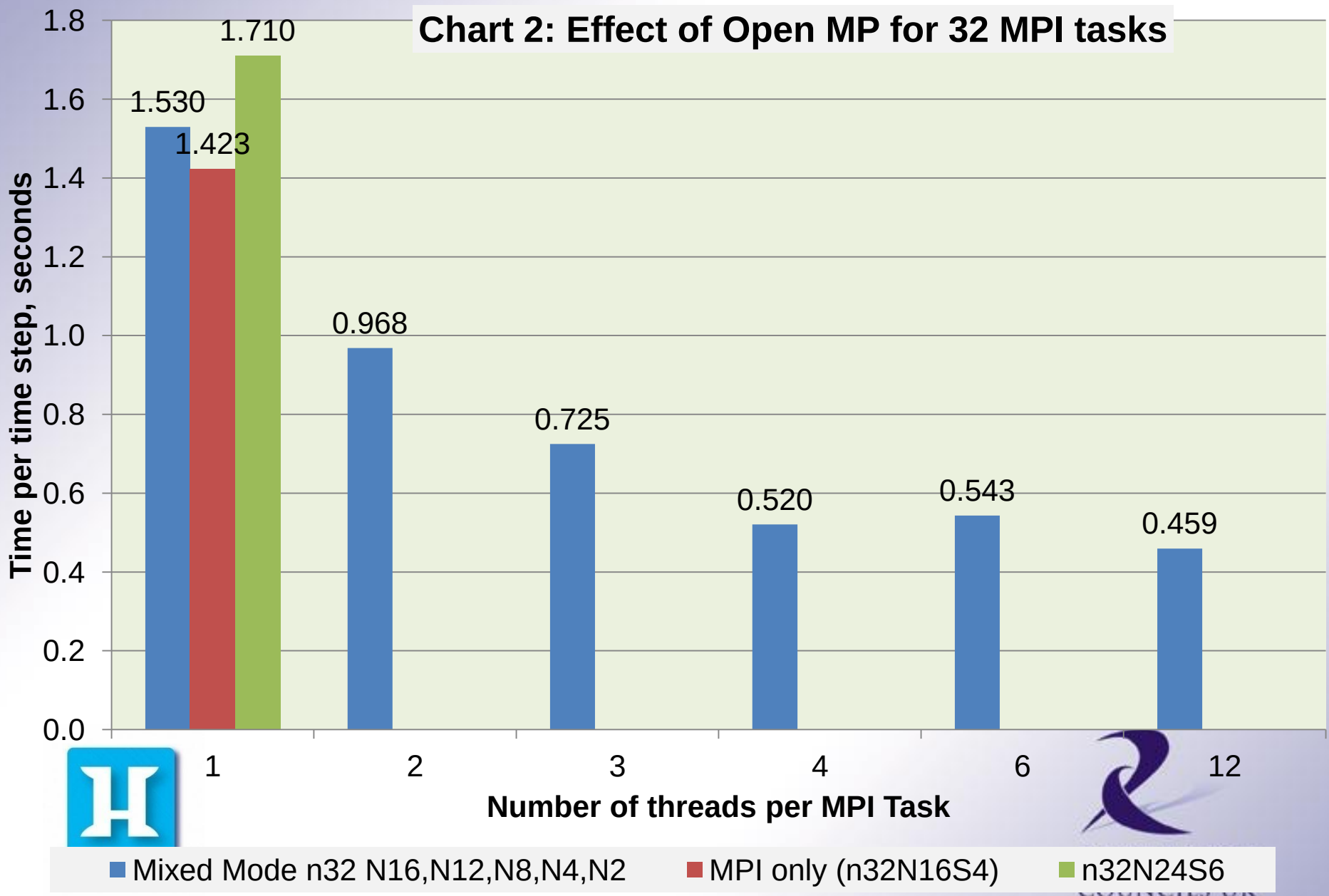


Chart 2: Effect of Open MP for 32 MPI tasks
Time per time step, seconds
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1.530
1.423
1.710
0.968
0.725
0.520
0.543
0.459
1
2
3
4
6
12
Number of threads per MPI Task
Mixed Mode n32 N16,N12,N8,N4,N2
MPI only (n32N16S4)
n32N24S6

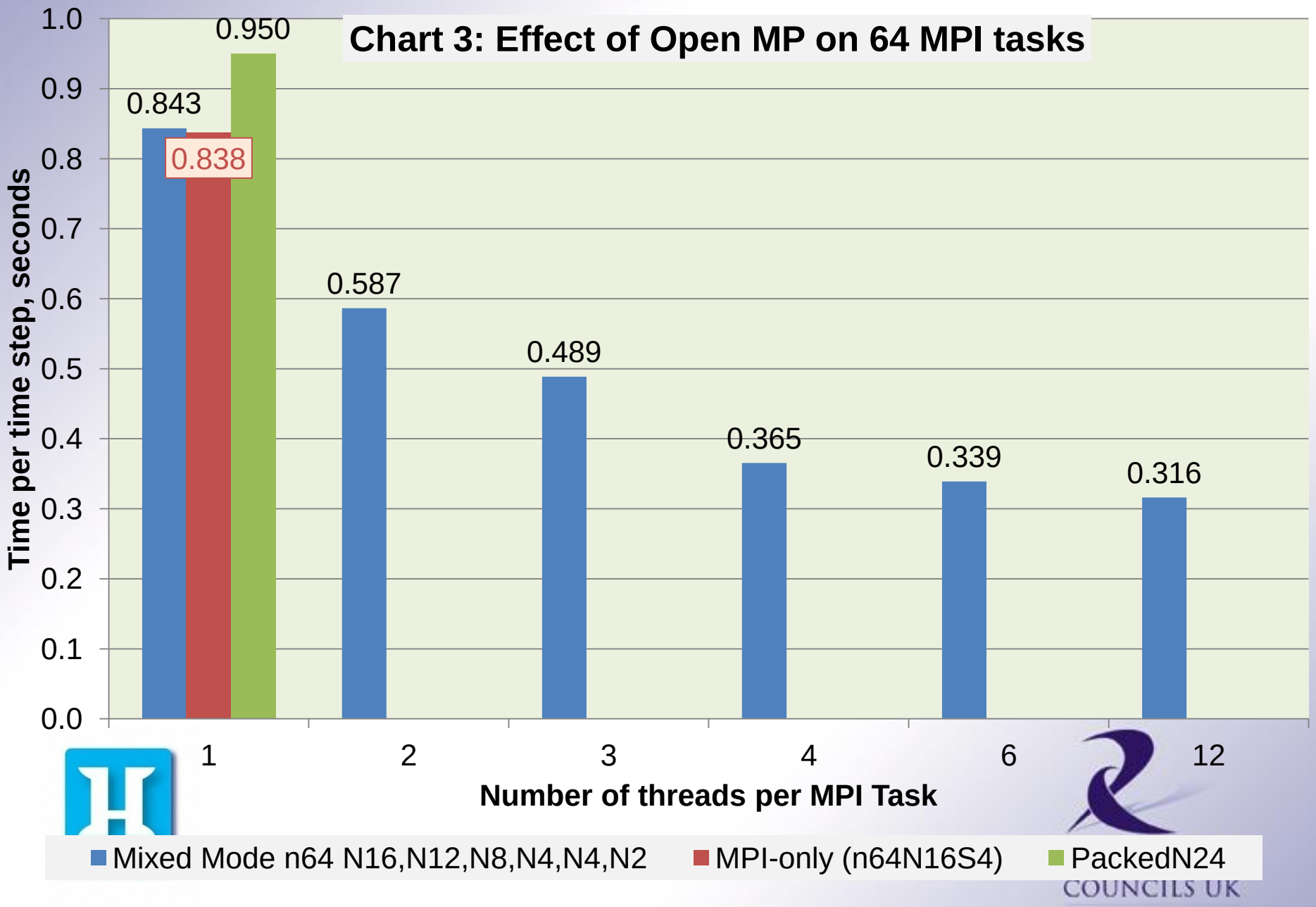


Chart 3: Effect of Open MP on 64 MPI tasks
Time per time step, seconds
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.843
0.838
0.950
0.587
0.489
0.365
0.339
0.316
1
2
3
4
6
12
Number of threads per MPI Task
Mixed Mode n64 N16,N12,N8,N4,N4,N2
MPI-only (n64N16S4)
PackedN24

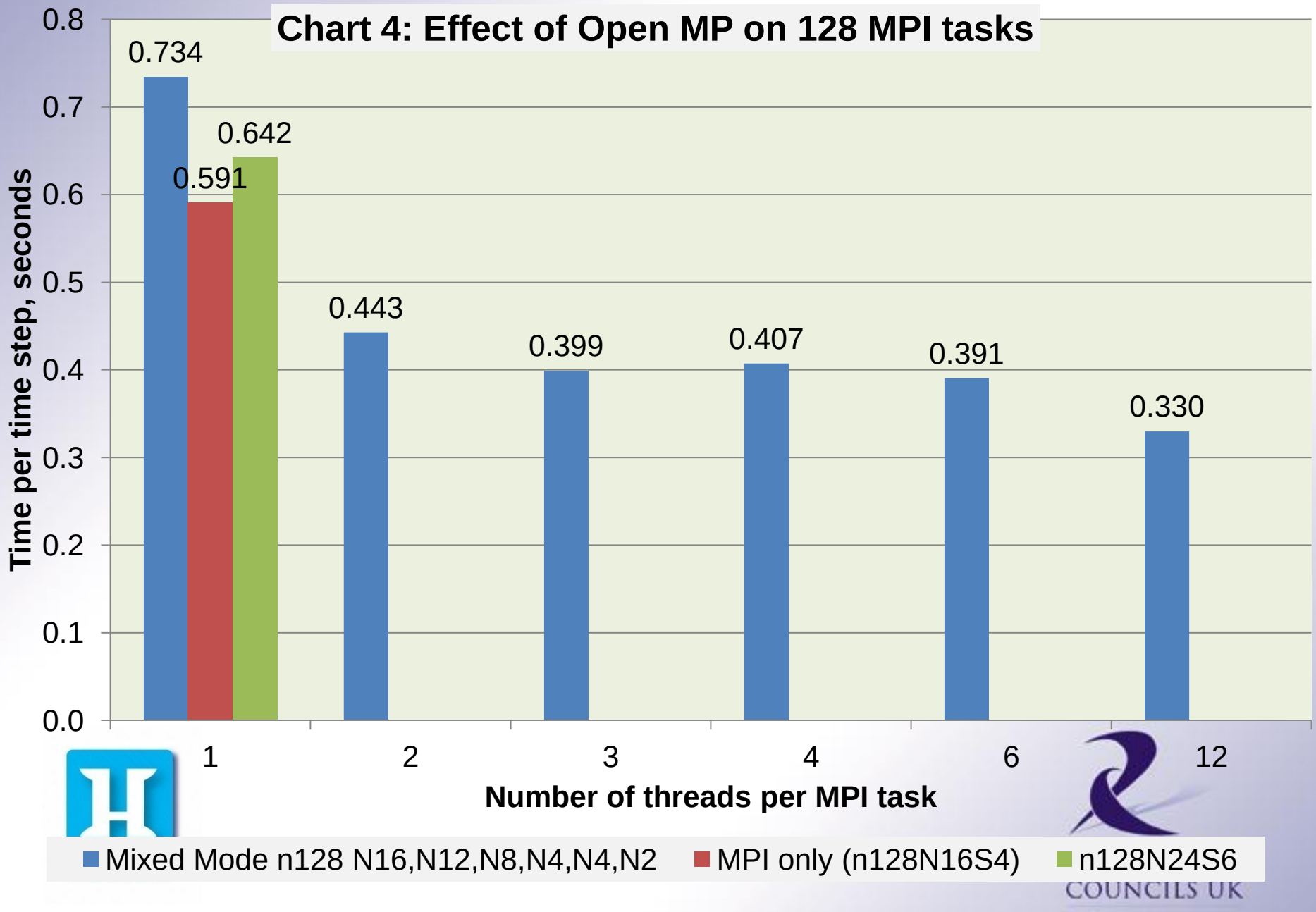
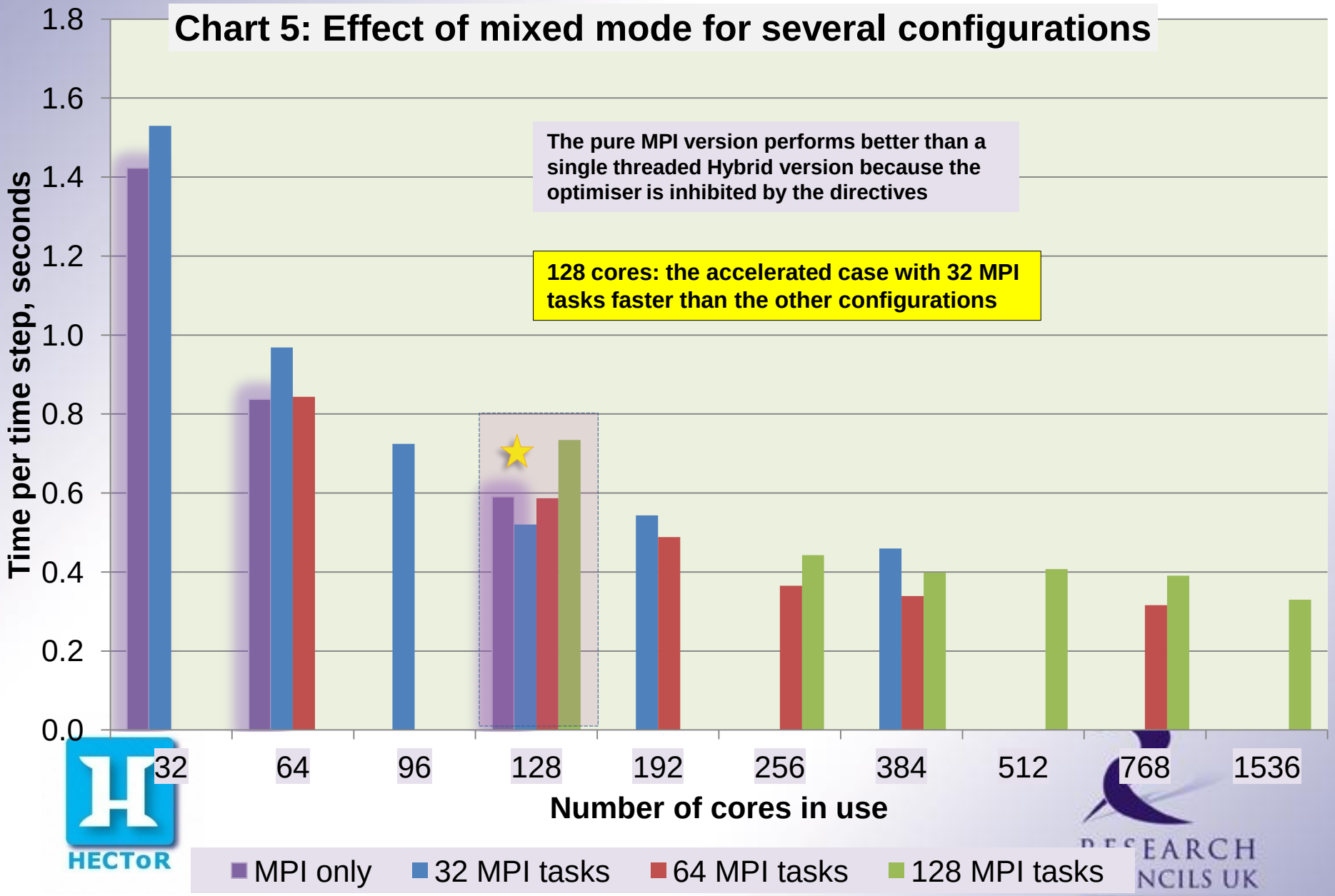


Chart 4: Effect of Open MP on 128 MPI tasks
Time per time step, seconds
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.734
0.591
0.642
0.443
0.399
0.407
0.391
0.330
1
2
3
4
6
12
Number of threads per MPI task
Mixed Mode n128 N16,N12,N8,N4,N4,N2
MPI only (n128N16S4)
n128N24S6

Chart 5: Effect of mixed mode for several configurations



Cost effective nature of hybrid runs

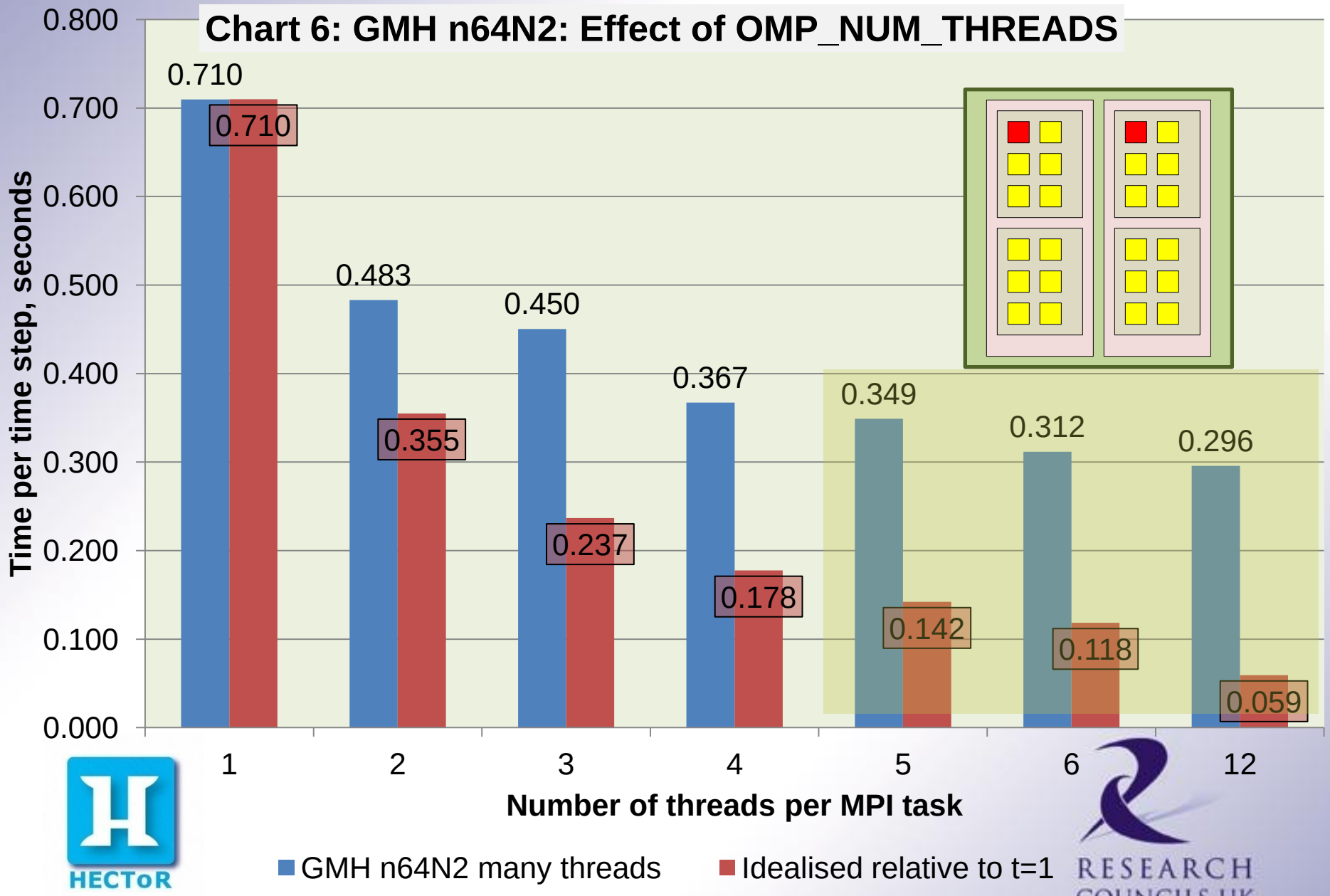
- comparing the cases in the chart 5, consider the number of nodes in use
  - n128N16d1 is 8 nodes : costs 960AU/hr
  - n32N4S1d4 is 8 nodes : costs 960AU/hr
  - n64N12d2 is 5.25 nodes : costs 720AU/hr
- Potentially better to use ¾ nodes and matching speed than the 10% speed improvement

HECToR

Based on 120 AU per hour for a standard node
Partial use of a node is charged at full node rate

RESEARCH COUNCILS UK

Chart 6: GMH n64N2: Effect of OMP_NUM_THREADS



HECToR

RESEARCH COUNCILS UK

Review objectives

- Enhance the existing MPI version of GLOMAP MODE with Open MP directives
  - Implemented on five loops
- Enable the mixed-mode of parallel operation for better use of the multi-core systems that are being installed
  - Examined different placement strategies
  - Uniform gives advantage over “packed”
- Allow higher resolution simulations
  - Not been able to determine (no case supplied yet)

HECToR

RESEARCH COUNCILS UK

Summary

- A complex code has been analysed
  - Changes made to modify the parallel operation
  - Extend its utilisation of the new hardware
- MPI-only would be limiting:
  - fully populated nodes return good value for charges
  - high task count communication profile impedes scaling
- Adding Open MP
  - allows the numbers of MPI task to be kept low
  - matching the performance of MPI-only

HECToR

RESEARCH COUNCILS UK

Conclusions

- Previous work had shown that under-population could provide a reduction in runtime
  - (XT4 Quad-core Compute Nodes, 2GB/c)
- Same is true for the XE6
  - Approx. 10% improvement over “packed”
  - Likely due to more space for intra-node comms that use shmem channels
- The choice of how to add the Open MP constructs is crucial to delivering expectations

HECToR

RESEARCH COUNCILS UK