GloMAP Mode on HECToR Phase2b (Cray XT6)

Mark Richardson
Numerical Algorithms Group

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 (EPCC based)
  - Cray Centre of Excellence
  - Portland Group (on-line forum)

What is HECToR?

- UK service for research
  - **H**igh-**E**nd **C**omputing **T**erafl**o**p **R**esource
- Currently phase2b of 3
  - 1855 nodes of 2x AMD "Magny-Cours" (24 cores)
- Theoretically 360TFlops
- LUSTRE storage
- Phase2a (Cray XT AMD Barcelona Quad Cores) still available soon to be connected by esFS (e-LUSTRE)

What is NAG's role in the HECToR service?

Core Support

- Helpdesk - applications queries
- Assistance with porting, tuning, migration (2a→2b →3), …

Training

- HECToR, Parallel programming, Optimisation, Software engineering, …

Distributed Support

- Dedicated 6-12 month projects to scale, renovate, restructure applications
- 23 projects completed so far, 25 in progress (>46 person-years)
- New call for proposals: deadline 6th December

HECToR

RESEARCH COUNCILS UK

What is GLOMAP?

- A computer program for simulating aerosol processes in the earth's atmosphere
  - TOMCAT an advection code 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.
    - Dr. Glenn Carver, University of Cambridge

HECToR

RESEARCH COUNCILS UK

Background

- MPI version in regular use on HECToR
- Open MP version for use on SMP machines
  - Typically 32 threads (with additional auto-parallel)
  - Developed elsewhere so cannot comment further
- Hybrid version was subject of DCSE
- Focus of the talk is placement of tasks on Cray XT6

Case description 1

- The T42 model is quite a low resolution model
  - 128 x 64 x 31 grid-boxes for whole atmosphere
- The incorporation of chemistry increases number of static arrays
  - causes compilation issues for large domains
  - minimum 4 nodes of phase1 XT4h using 8 cores
    - 3GB/core available
- The MPI version decomposes by two dimensions
  - longitude by latitude (but typically 4 patches across a lat)
  - retains full altitude on each “patch”
- Each hexahedral division is called a grid-box
- Typical use is 32 MPI tasks
  - each containing 32 x 8 x 31 grid-boxes

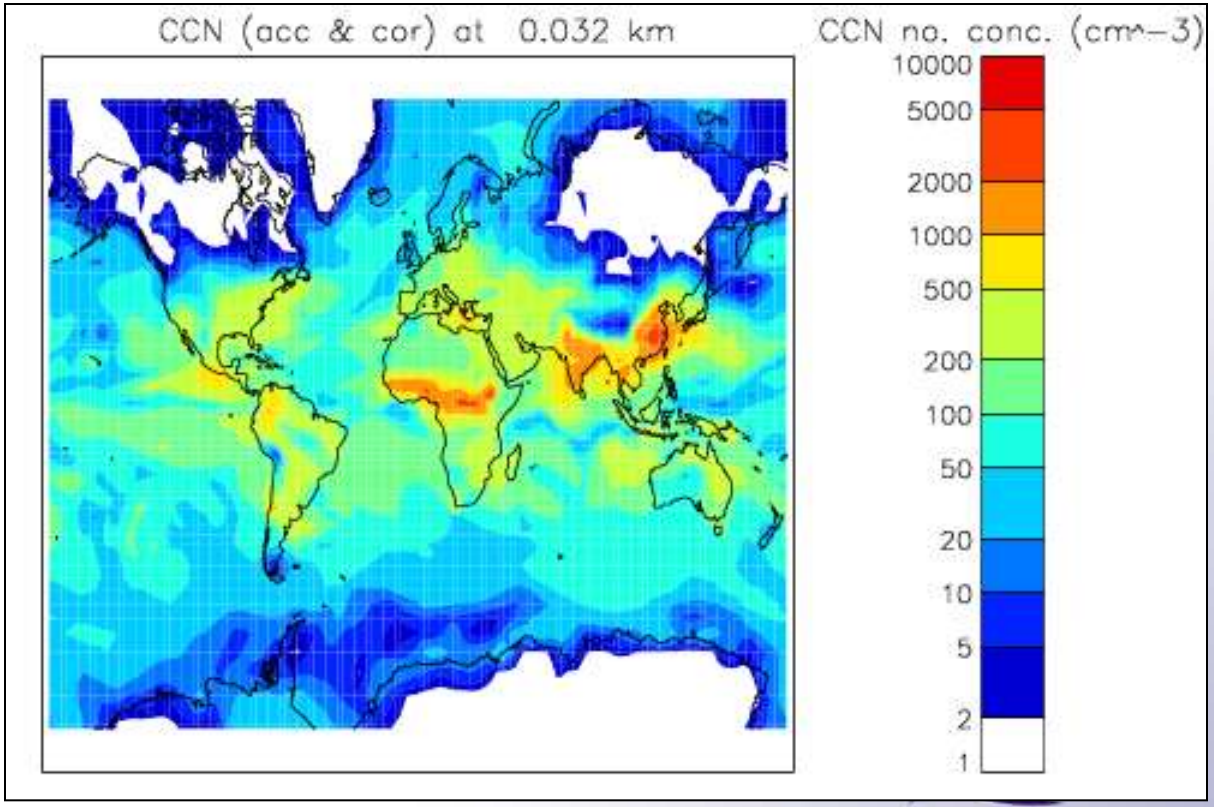
HECToR

RESEARCH COUNCILS UK

Case description 2

- Earth's atmosphere mapped into a 3-D Cartesian coordinate system
- T42 is 128x64x31 (grid-boxes)
- MPI 2-D topology creates uniform patches (e.g. 32 x 4 x 31)

- 197 scalars
- 3 days used for investigation
- (144 steps)
- Initial I/O stages omitted as they form ~30% of this simulation



HECToR

RESEARCH COUNCILS UK

Goals of the project

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

Table 1: phase2a Cray PAT results for fully populated nodes GLOMAP mode purely MPI version and three configurations

Phase 2a is one quad core per node, will use the fully packed pure MPI version for comparisons (i.e. production version)

	M32 % of whole sim	M32 % of GMM only	M64 % of whole sim	M64 % of GMM only	M128 % of whole sim	M128 % of GMM only
ADVX2	4.2	5.7	2.8	5.5	1.3	5.7
ADVY2	11	14.9	7.1	13.9	2.2	9.7
ADVZ2	4.9	6.6	3.2	6.3	1.4	6.2
CONSOM	5.4	7.3	3.6	7.1	1.1	4.8
CHIMIE	**40.9**	**55.3**	**27.4**	**53.7**	**12.4**	**54.6**
MAIN	7.7	10.4	7	13.7	4.4	19.4
TOTAL FOR GMM	74	100	51	100	22.7	100
MPI	13.3	-	28.3	-	47.4	-
MPI_SYNC	12.7	-	20.7	-	29.9	-

TOMCAT Analysis

- 4 subroutines in TOMCAT accounts for ~30% of the GMM
- ADVX2 an outermost loop over NIV
  - Upper limit 31 at the test case resolution
- ADVY2 an outer most loop over NIV
  - Upper limit 31 at the test case resolution
  - Has some extra MPI work for polar regions
- ADVZ2 an outermost loop over MYLAT
  - Upper limit (16,8,4,2) for (16,32,64,128) tasks
- CONSOM a second level loop over NTRA
  - Upper limit 36
  - The outer loop is over MYLAT

CHIMIE Analysis

- The CHIMIE subroutine accounts for ~55 %
  - has some MPI work and additional loops external to major loop
  - includes all 'UKCA' subroutines called by CHIMIE
- Contains a major loop over latitudes (MYLAT)
  - Upper limit (16,8,4,2) for (16,32,64,128) tasks
  - This is main Open MP directive
- Introduced
  - THREADPRIVATE common block for interfacing with ASAD
  - Large block of Open MP code, explicitly declaring:
    - private data passed into the GLOMAP sub-system
    - shared data
    - private common blocks

HECToR

RESEARCH COUNCILS UK

ASAD Analysis

- ASAD had already been converted for use with the earlier version of GLOMAP with Open MP
  - This is an ODE solver for chemical reactions
  - It is wholly within the CHIMIE subroutine
- Only common blocks have been treated to retain private data
- No “acceleration” from Open MP explicitly
  - The time spent within ASAD is per thread
  - Time for computation in the MPI task is reduced
- Possible future project on balancing chemistry calculations

Systems

- XT4h ("h" because of connected Cray X2)
  - Initially dual-core Opteron
    - 2 cores per node and ~3GB per core
  - Phase2a has a single quad-core "Barcelona"
    - 4 cores per node and ~2GB per core
- XT6 (phase2b)
  - Two "Magny-Cours" processors per node
    - 24 cores per node ; ~1.5GB per core (8GB per hex-core die)
    - Each hex-core die shares 6MB L3 cache (well, 5MB of it!)
  - Specific description is at:
    - http://www.hector.ac.uk/cse/documentation/Phase2b/

TABLE 2: Comparison of MPI and mixed-mode parallel on XT4h

Units are seconds per time step

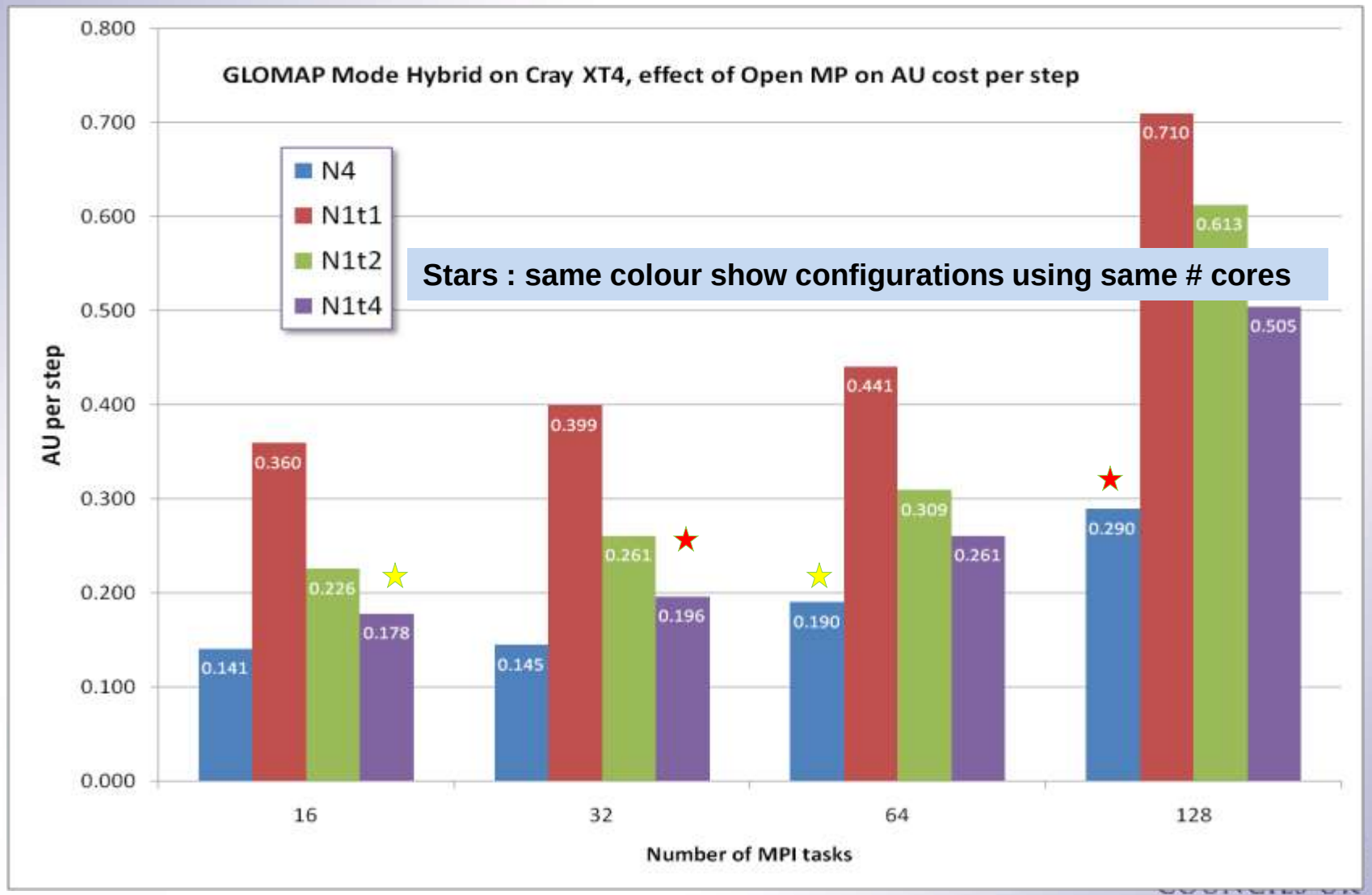
MPI tasks	16	32	64	128
XT4h GMM N4	4.227	2.174	1.426	1.085
XT4h GMH N1t1	2.696	1.498	0.826	0.665
XT4h GMH N1t2	1.692	0.979	0.58	0.574
XT4h GMH N1t4	1.337	0.735	0.489	0.473

Key point is: (32 MPI tasks)x(4 OMP threads) is better than 128 MPI tasks
i.e. a 3x speed-up over a 2x speed-up

HECToR

RESEARCH COUNCILS UK

Chart 1 : Effect of Open MP on cost per step on XT4h



Cray XT6: Naive placement is packed (N24)

- Will look at 32, 64 and 128 MPI task configurations
  - In pure MPI mode can look at "balanced" placements
- With hybrid version can look at spreading out
  - n32N4S1d4 n64N12S3d2 (i.e. 128 cores)
- Note the final node is not balanced
- Same amount of work per core for TOMCAT
- Extra MPI communications while diminishing data size
- Varying amount of work for chemistry per grid-box

Diagrams of task placement on phase2b

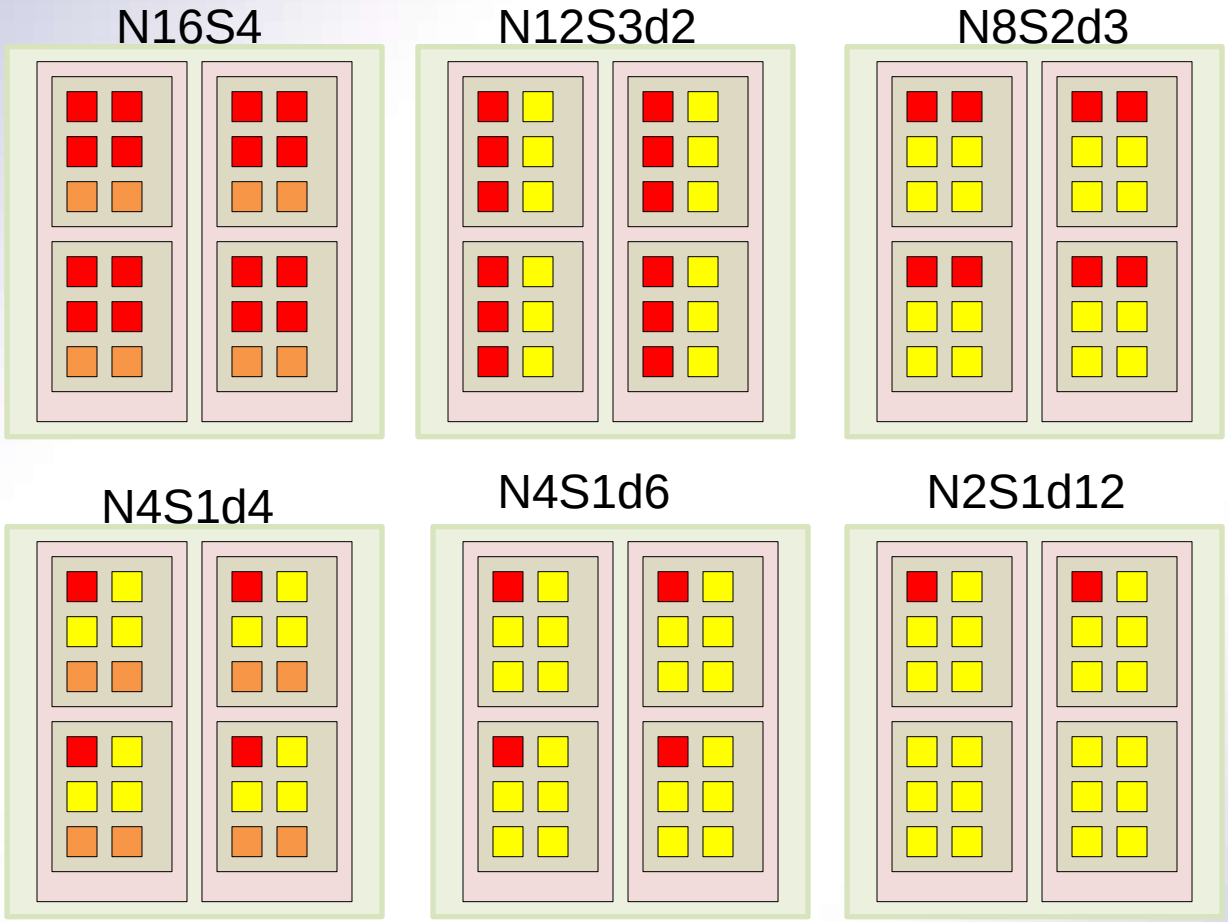


Table 3: Utilisation considerations

- Costs skewed by modulus 24
- Table shows number of nodes in use and utilisation
  - (cores in use)/(cores available)

	32 MPI	utilisation	64 MPI	utilisation	128 MPI	utilisation
N24	2, 0.67	32/48	3, 0.89	64/72	6, 0.89	128/144
N4 d6	8, 1.00	192	16, 1.00	384	32, 1.00	768
N4 d4	8, 0.67	128/192	16, 0.67	256/384	32, 0.67	512/768
N8 d3	4, 1.00	96	8, 1.00	192	16, 1.00	384
N12 d2	3, 0.89	64/72	6, 0.89	128/144	11, 0.97	256/264

Key point: can using fewer cores than are available improve the cost of simulation?

CHART 2 : Mixed-mode XT6, 32 MPI tasks (May 2010)

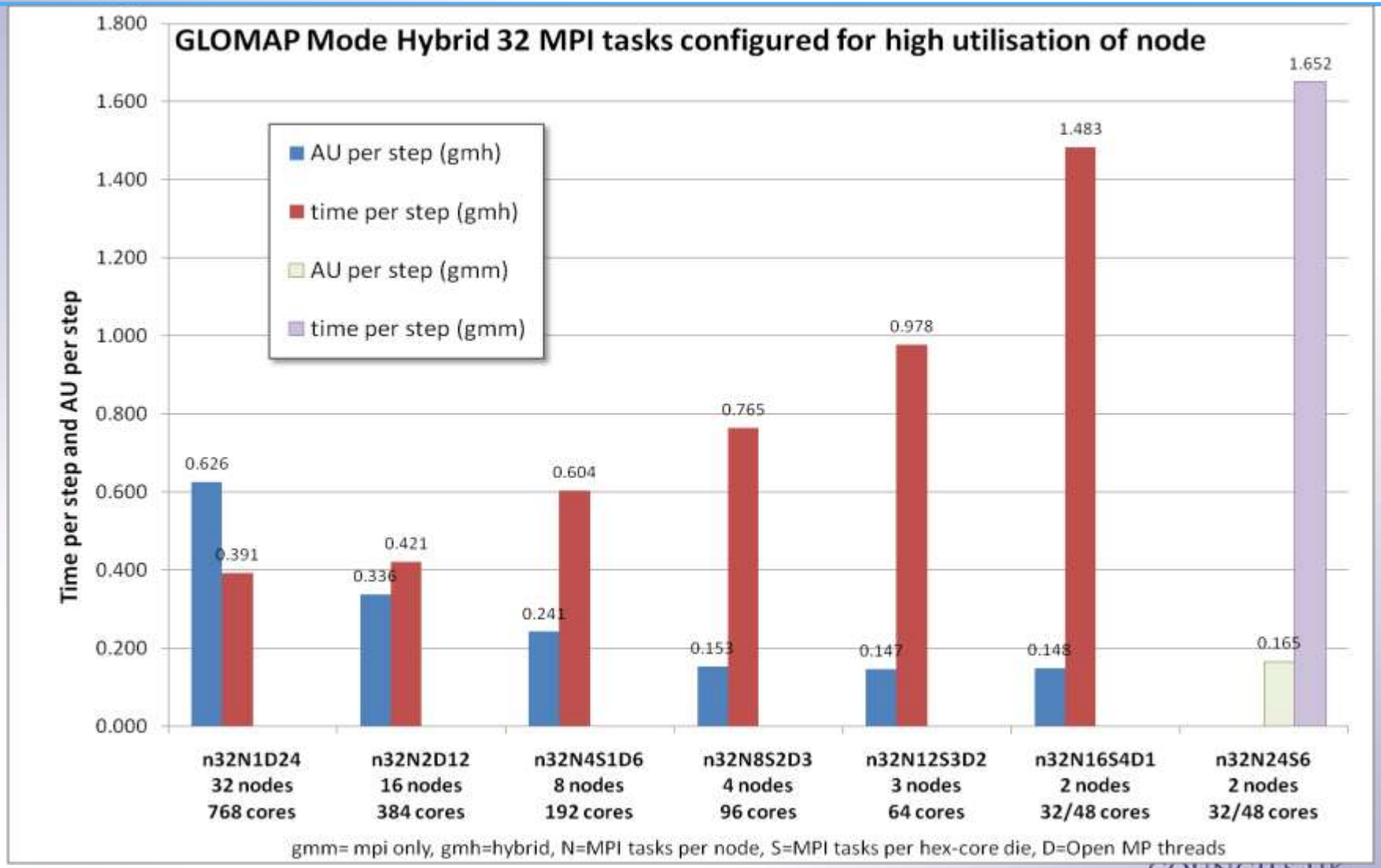


CHART 3 : mixed mode XT6, 64 MPI tasks (May 2010)

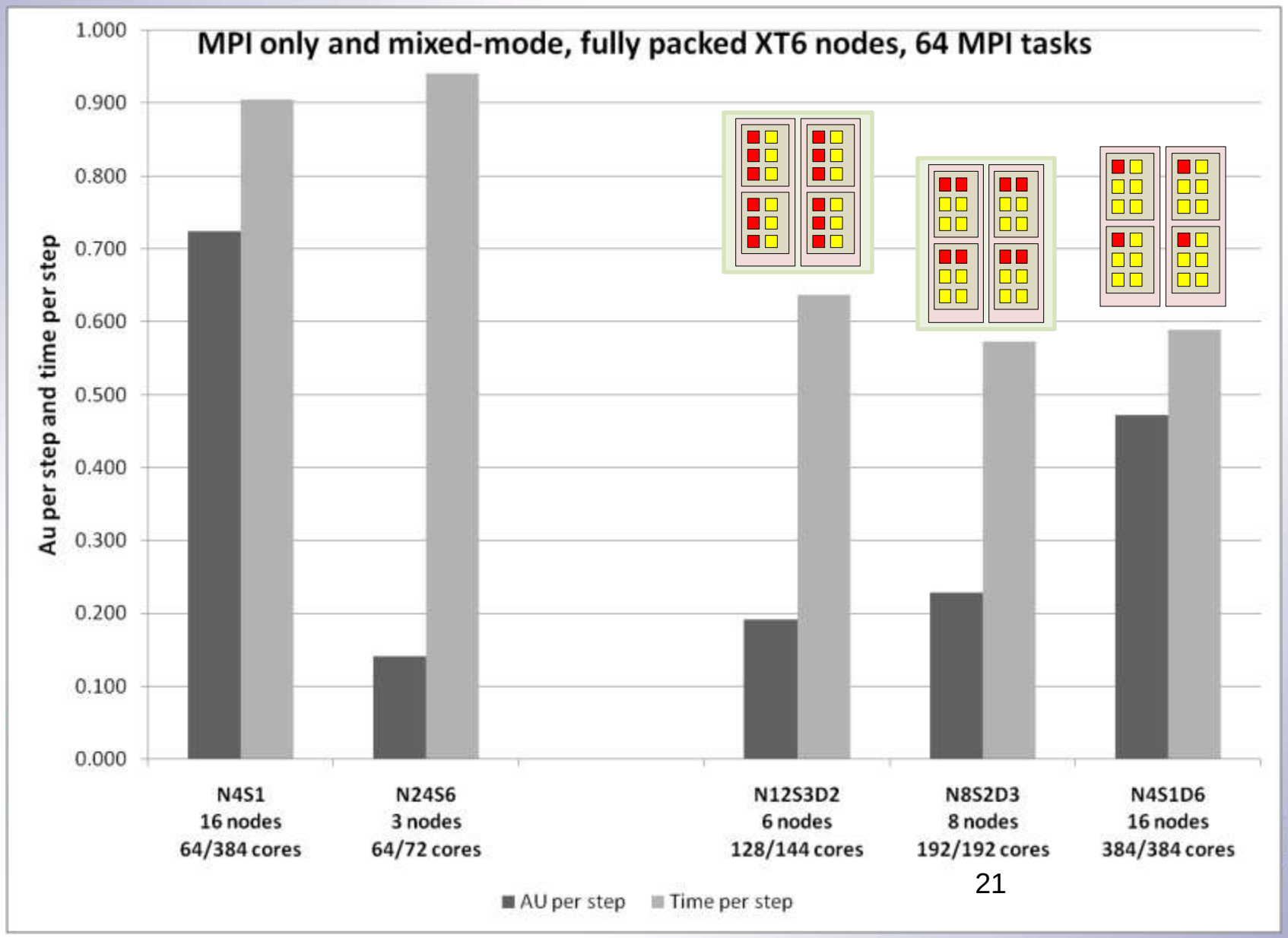
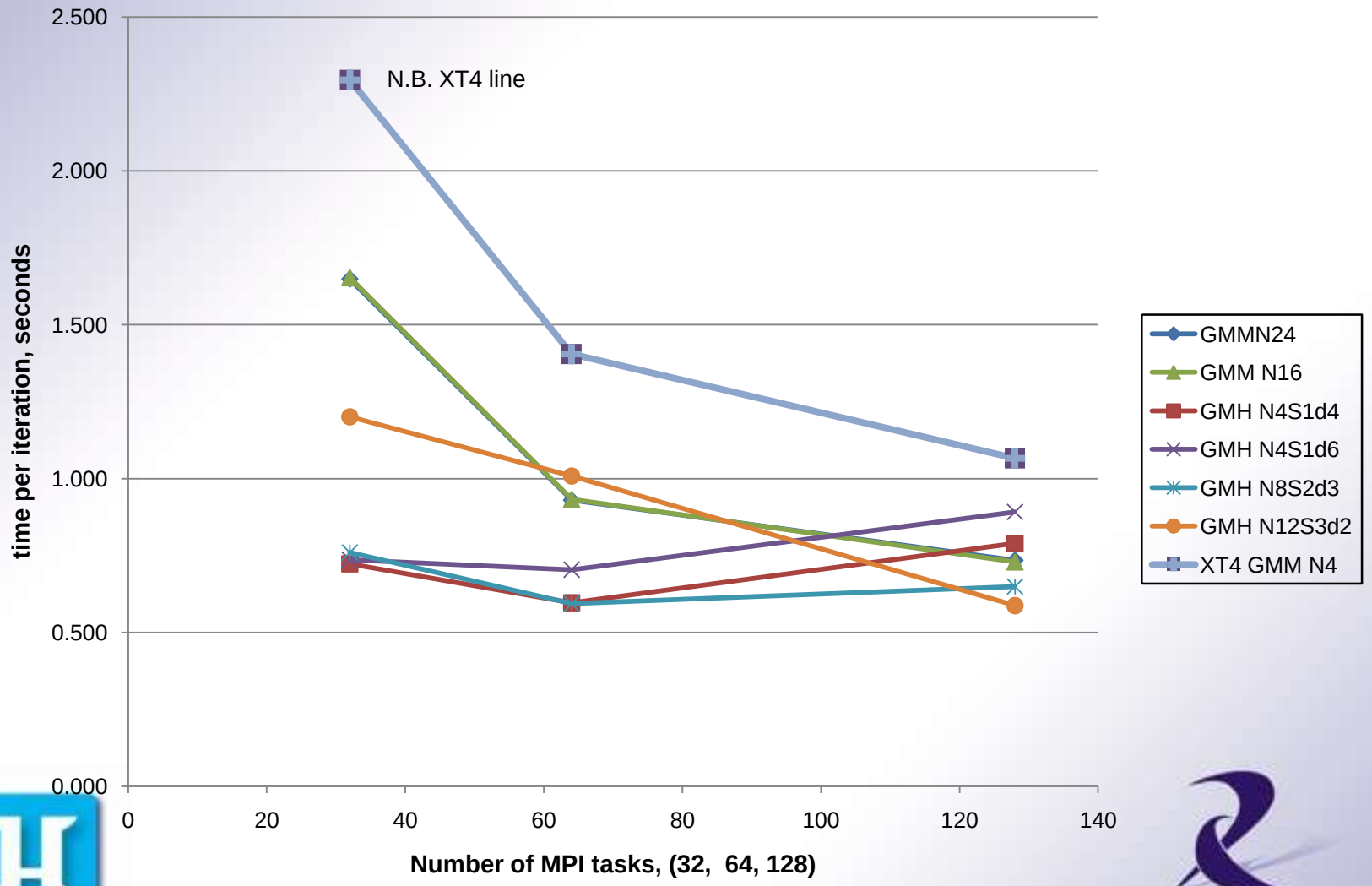
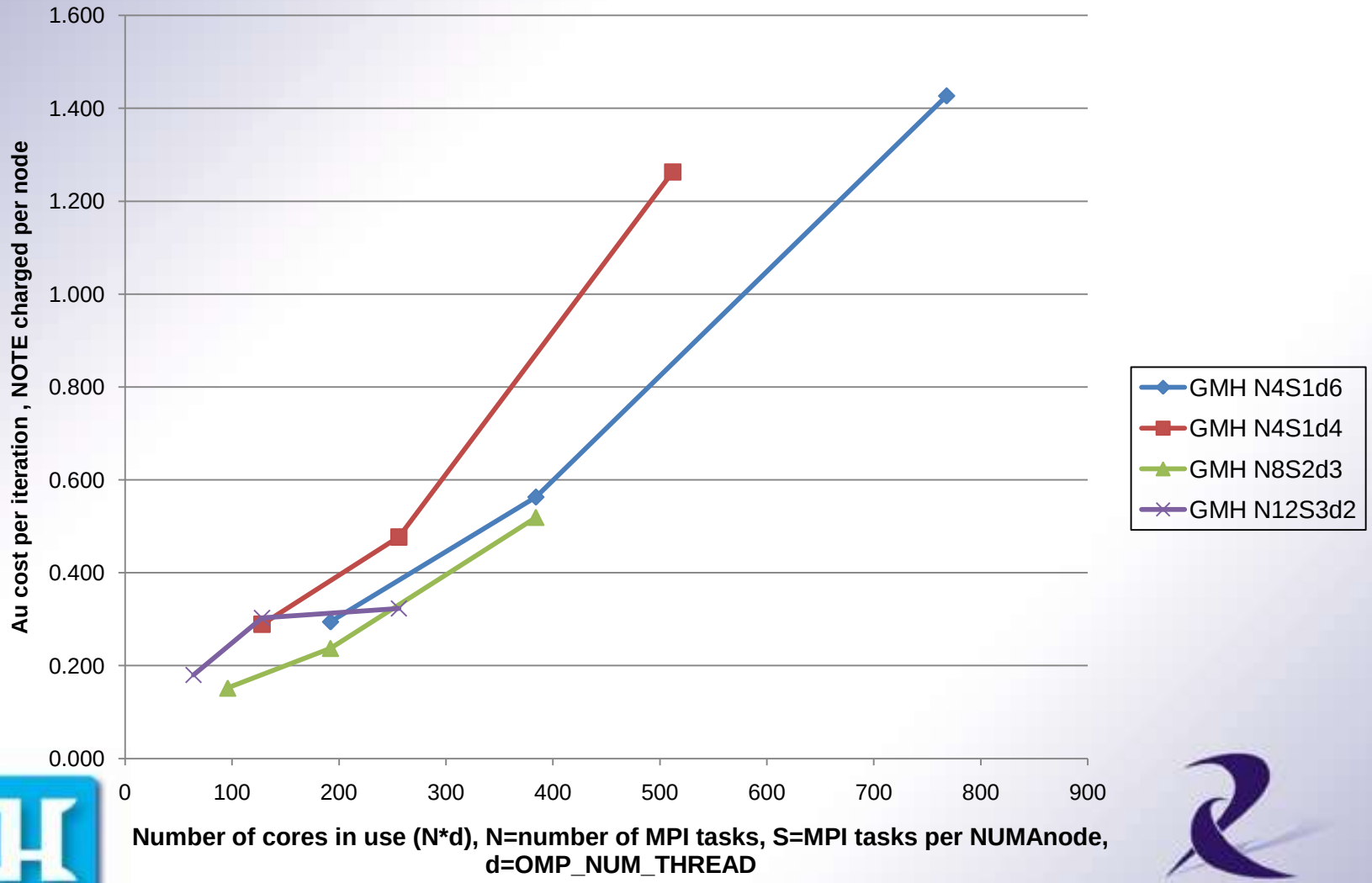


Chart 4: Effect of Open MP on Scaling (Oct 2010)



Key point: no clear advantage for “many” threads

Chart 5: Effect of Open MP on cost of simulation



HECToR

RESEARCH COUNCILS UK

Open MP acceleration: for this case

- This is reasonable on the XT4h
- This is non-ideal on XT6
  - Although should not expect to scale directly with number of threads
    - Amdahl effect
  - Loop limits are not well balanced for more than 8/4/2 threads
    - Depending on MPI decomposition (32/64/128)
  - MPI version works better than on XT4
- There is X% of code where Open MP implemented
  - X varies with MPI : (e.g. On XT4 90, 86, 80)
  - Only applied to the significant workload loops
  - Much larger effort would be required to fix every subroutine (330)

HECToR

RESEARCH COUNCILS UK

Summary and Conclusions

- GLOMAP mode has been converted to hybrid and tested on XT6
- The performance of the pure MPI production code on phase2b has improved ~30% over phase2a timings
- Activating the Open MP directives allows the code to use more cores per MPI task
  - Reduced number of MPI tasks keeps the code in the "good" scaling region (64 MPI tasks)
- Filling nodes in a "balanced" manner has little effect on performance
  - Except when considering the AU usage
- Open MP is limited
  - can only be as effective as the loop count
  - Plenty more to go for in this code
  - Larger case will be possible

HECToR

RESEARCH COUNCILS UK

Postscript:
Table 4: Amdahl effect, what should I expect?

Percentage of original run	M32 % of whole sim	Max loop count	Scaled by number of Open MP threads							
			1 thread	**2**	**4**	**6**	**8**	**12**	**24**	1024
ADVX2	4.2	31	4.20	2.10	1.05	0.70	0.53	0.35	0.18	0.14
ADVY2	11	31	11.00	5.50	2.75	1.83	1.38	0.92	0.46	0.35
ADVZ2	4.9	8	4.90	2.45	1.23	0.82	0.61	0.61	0.61	0.61
CONSOM	5.4	36	5.40	2.70	1.35	0.90	0.68	0.45	0.23	0.17
CHIMIE	40.9	8	40.90	20.45	10.23	6.82	5.11	5.11	5.11	5.11
MAIN	7.7	1	7.70	7.70	7.70	7.70	7.70	7.70	7.70	7.70
TOTAL FOR GMM	**74**		**74.10**	**40.90**	**24.30**	**18.77**	**16.00**	**15.14**	**14.28**	**14.08**
MPI	13.3		13.3	13.3	13.3	13.3	13.3	13.3	13.3	13.3
MPI_SYNC	12.7		12.7	12.7	12.7	12.7	12.7	12.7	12.7	12.7
Total for program	**100**		**100.10**	**66.90**	**50.30**	**44.77**	**42.00**	**41.14**	**40.28**	**40.08**
Speed up			***1***	***1.495***	***1.99***	***2.23***	***2.38***	***2.43***	***2.48***	***2.49***

Not as much as these idealised figures!

HECToR

RESEARCH COUNCILS UK