Performance Analysis and Optimisation of

METOFFICE UNIFIED MODEL ON A CRAY XC30

EASC 2015 Edinburgh

Karthee Sivalingam & NCAS - CMS, University of Reading

National Centre for Atmospheric Science

NATURAL ENVIRONMENT RESEARCH COUNCIL

Met Office

JWCRP

NERC

SCIENCE OF THE ENVIRONMENT

NCAS-CMS

The National centre for Atmospheric Science (NCAS) - Computational Modelling services (CMS)
NCAS is one of the Natural Environment Research Council's (NERC) research centres.

AIMS

- The science of climate change, including modelling and predictions
- Atmospheric composition, including air quality
- Weather, including hazardous weather
- Technologies for observing and modelling the atmosphere

HPC & CLIMATE

Top 500 list Nov 2014

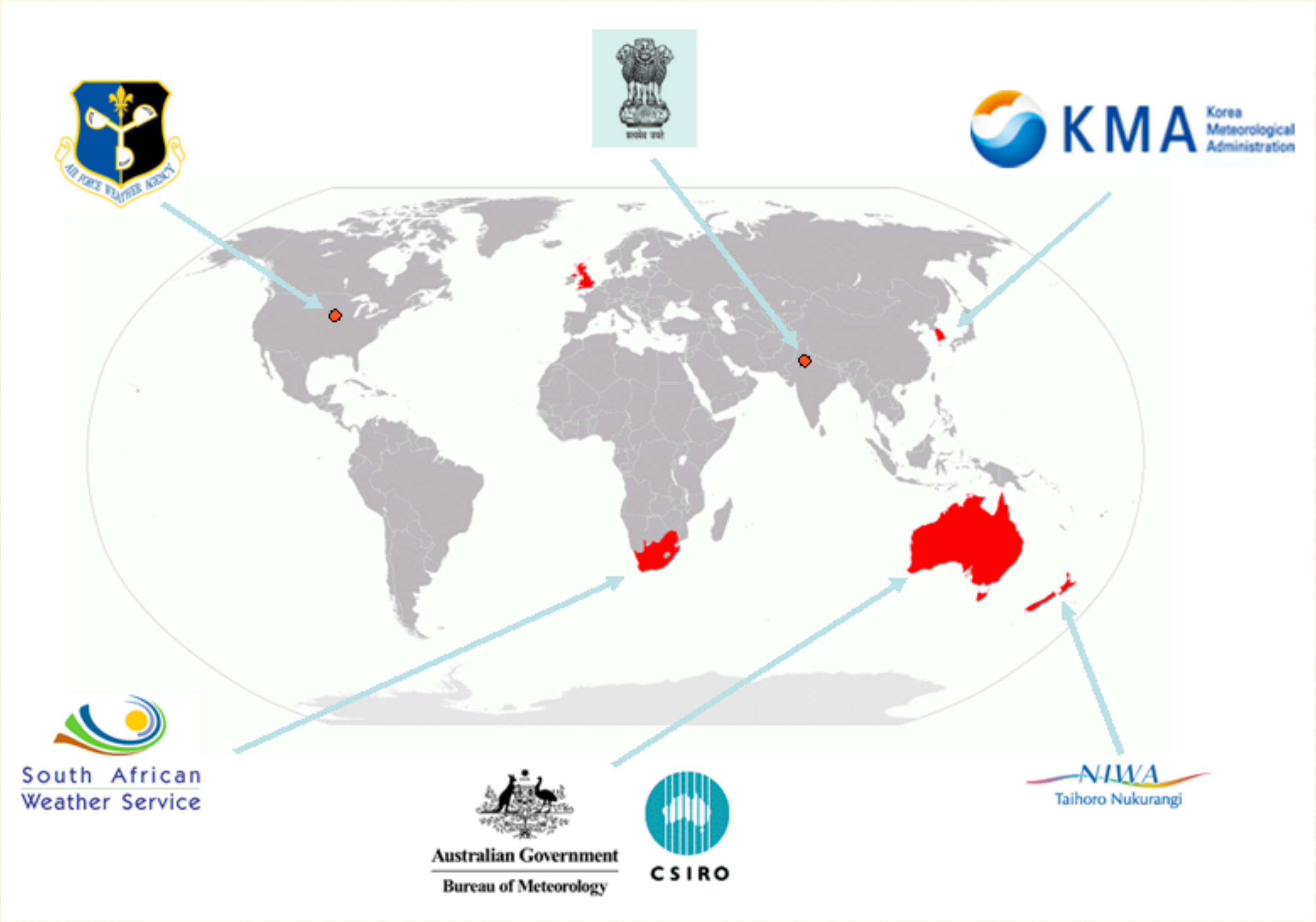
RANK	SITE	SYSTEM	CORES	RMAX (TFLOPS)	RPEAK (TFLOPS)
25	EPSRC/University of Edinburgh United Kingdom	**ARCHER** - Cray XC30, Intel Xeon E5 v2 12C 2.700GHz, Aries interconnect Cray Inc.	118,080	1,642.5	2,550.5
28	ECMWF United Kingdom	Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.	83,160	1,552.0	1,796.3
29	ECMWF United Kingdom	Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.	83,160	1,552.0	1,796.3
30	Science and Technology Facilities Council - Daresbury Laboratory United Kingdom	**Blue Joule** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	131,072	1,431.1	1,677.7
43	University of Edinburgh United Kingdom	**DiRAC** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	98,304	1,073.3	1,258.3
—	Met Office United Kingdom (BBC)	Cray XC40, Intel Haswell	480,000	16,000	16,000

OBJECTIVE

- Message passing , Threading and Vectorisation
- Message passing has scaling limitations
- Single node performance - Threading and Vectorisation
- Intel x86, IBM Power, Coprocessors, Accelerators, ARM 64 bit
- Need for a HPC library that abstracts the architecture

METOFFICE UNIFIED MODEL

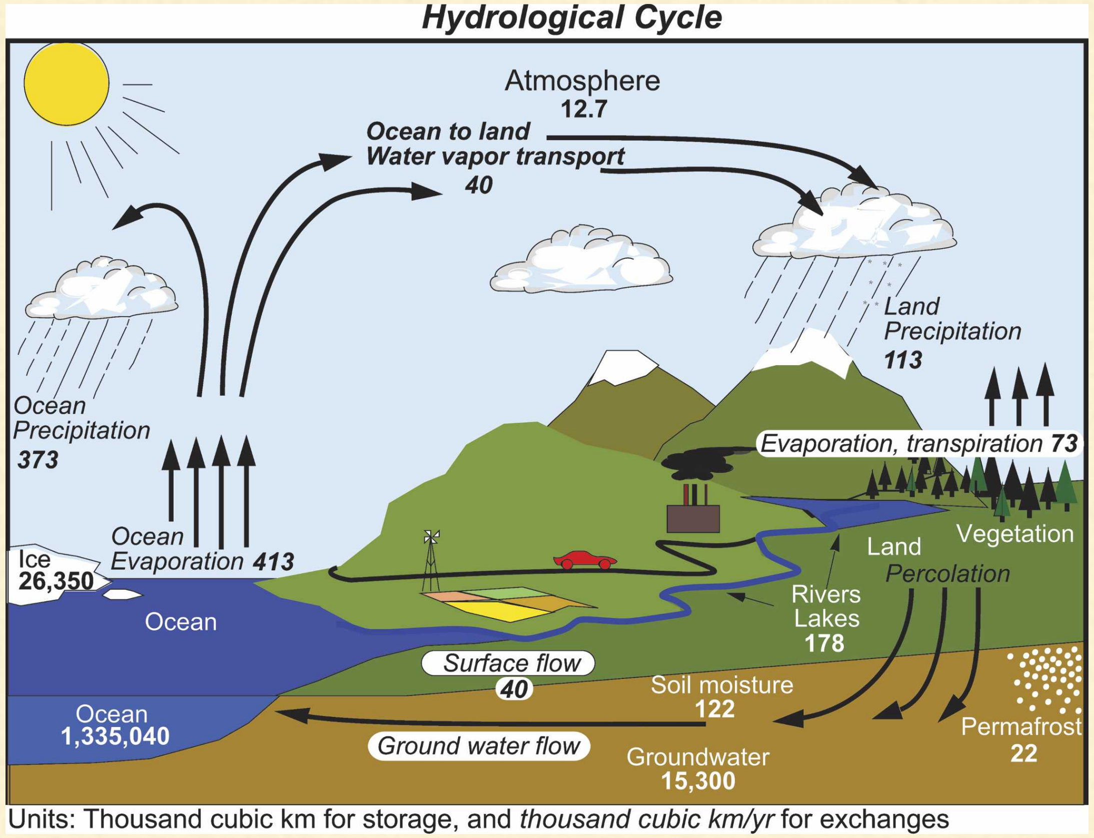
- The UM is a numerical modelling system, developed by the UK Met Office, and used for operational weather forecasting and climate prediction.
- It is licensed to the UK academic community for research.
- Joint Weather and Climate Research Programme (JWCRP), a strategic partnership between NERC and the Met Office for model development.
- It is used by forecast centres and climate agencies around the world

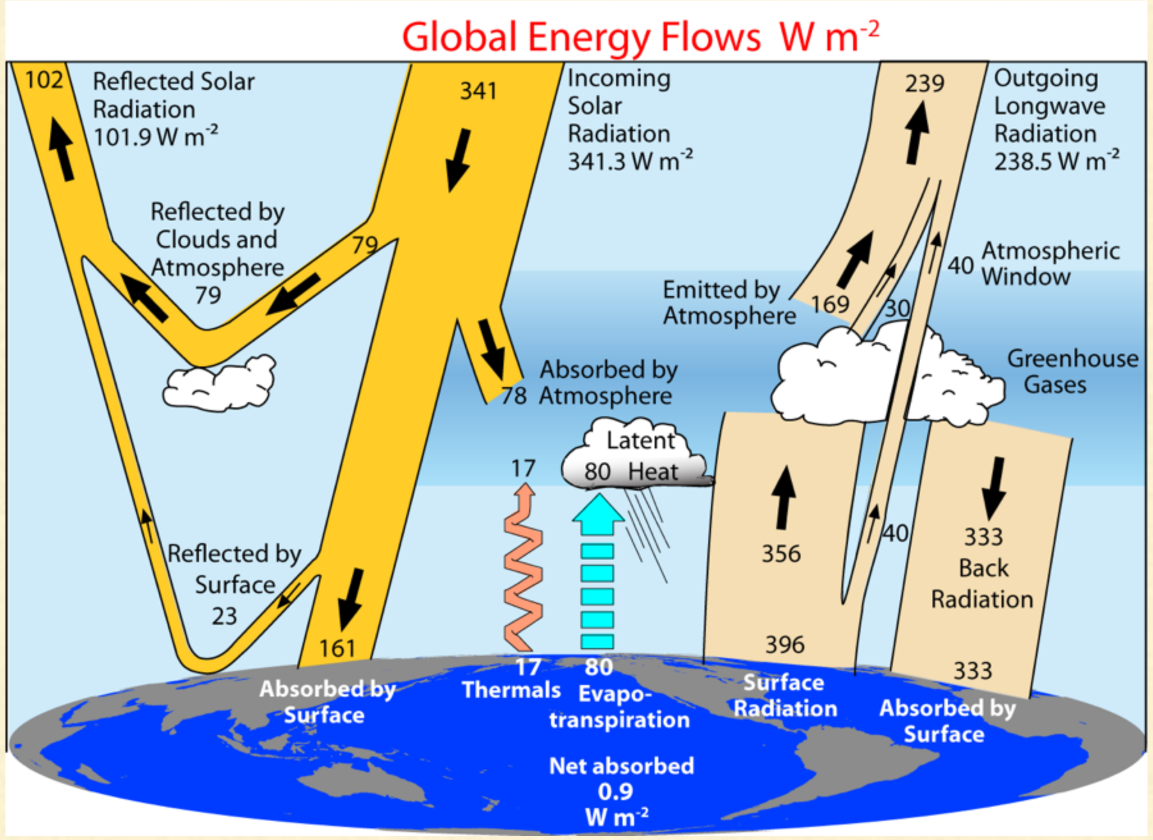


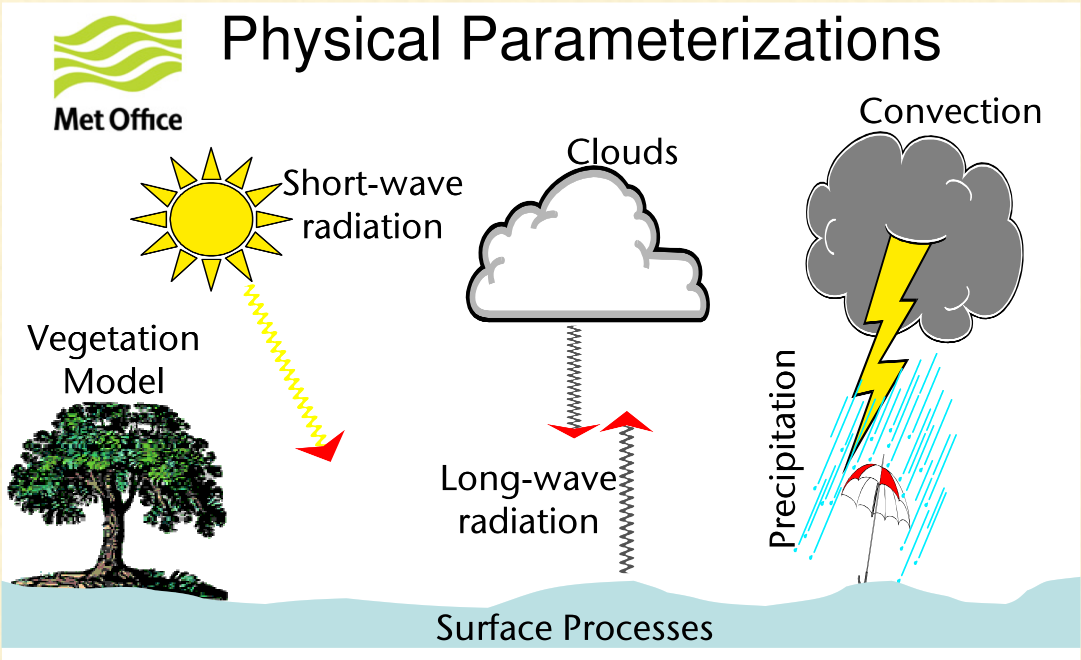
CLIMATE MODELLING

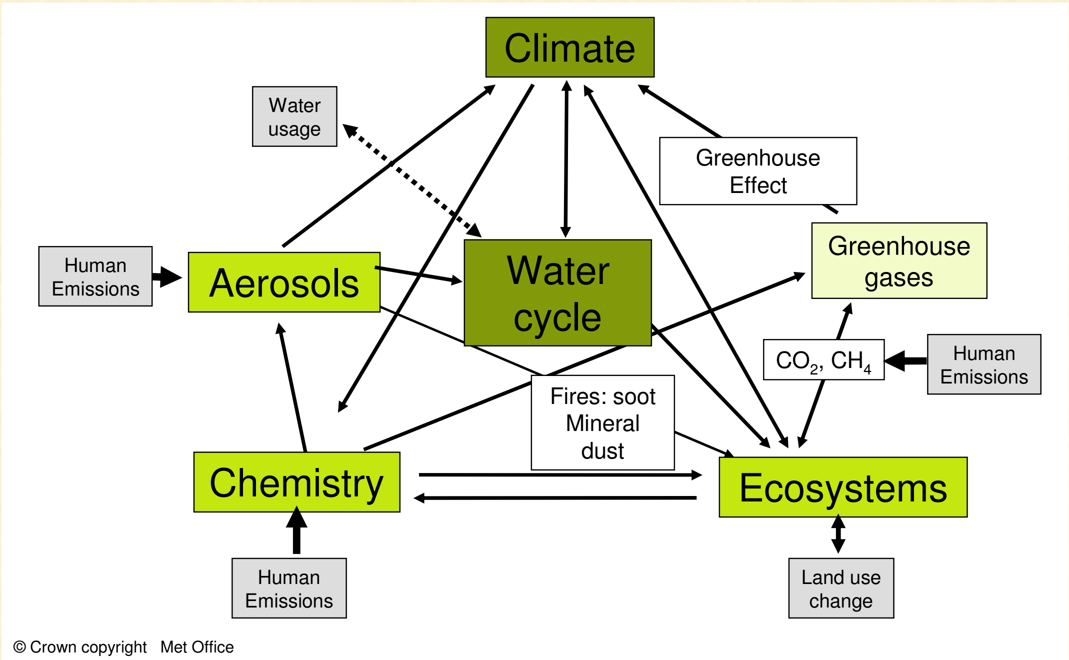
courtesy : Trenberth et al, 2007,2009

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL





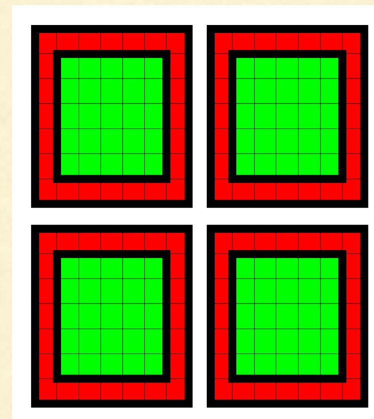


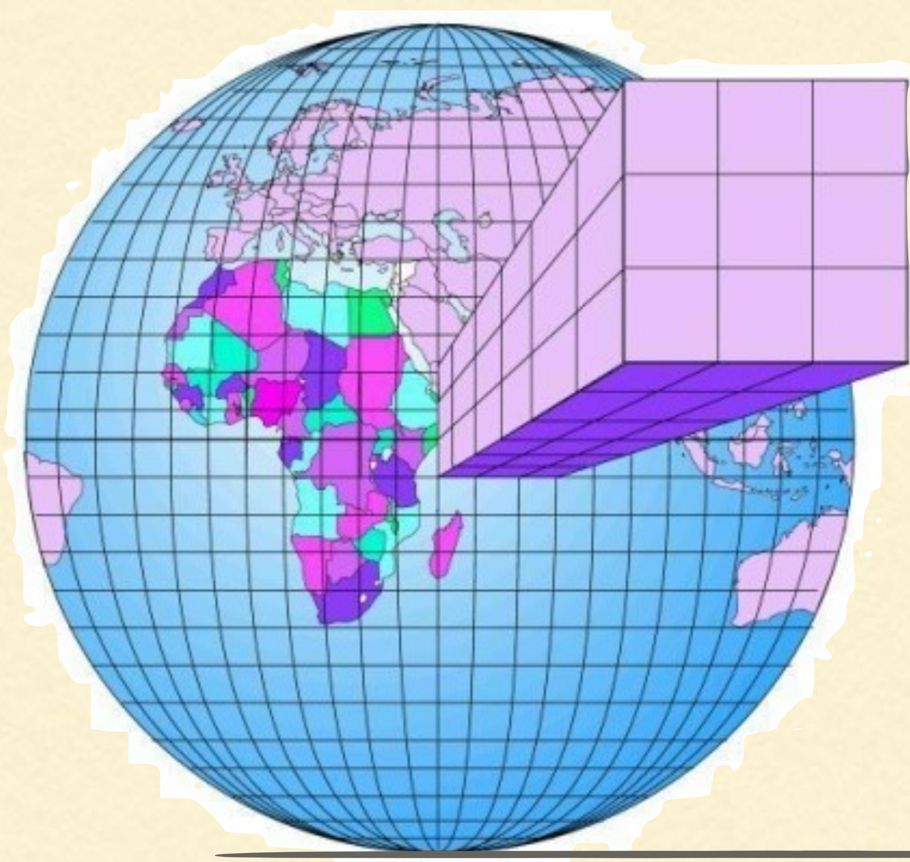


PARALLEL IMPLEMENTATION

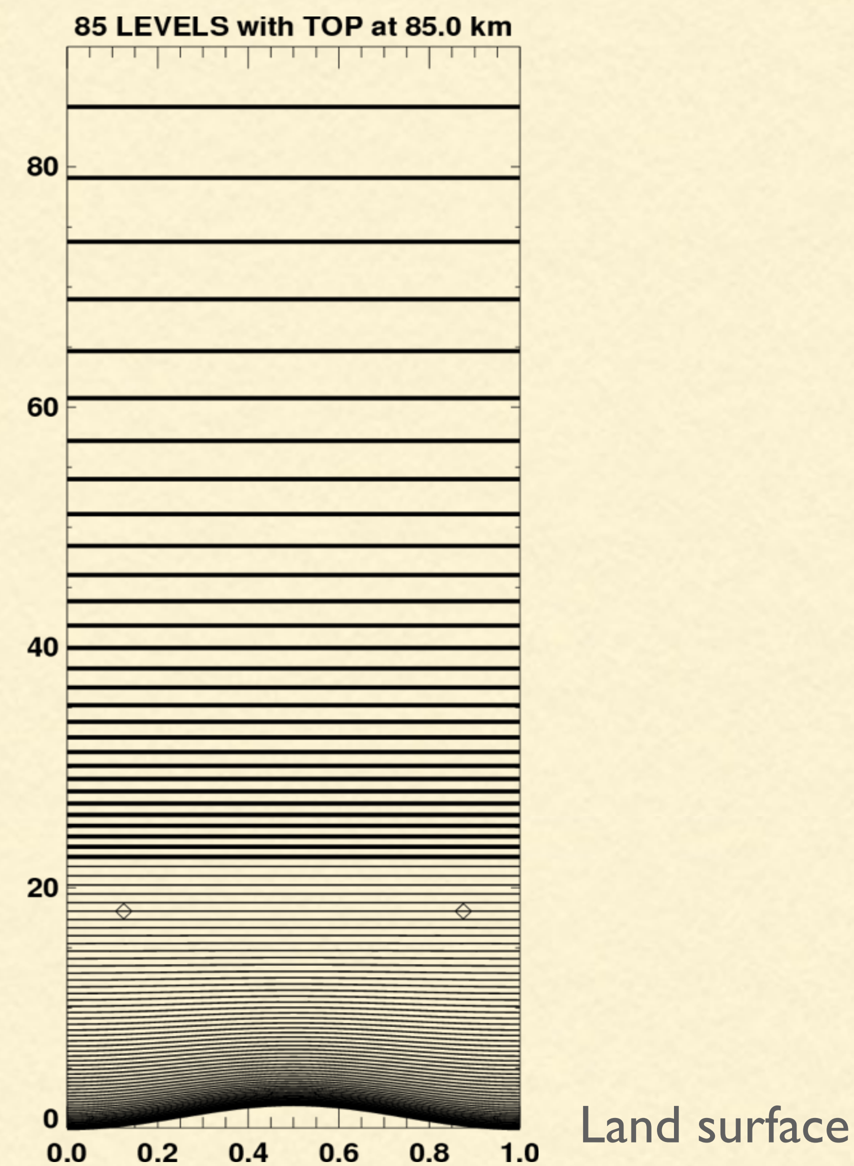
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Regular, Static, Lat-Long Decomposition
- Mixed mode MPI/OpenMP
- Asynchronous I/O servers
- Communications on demand for advection
- Multiple halo sizes (up to 8)





Horizontal resolution



Vertical resolution

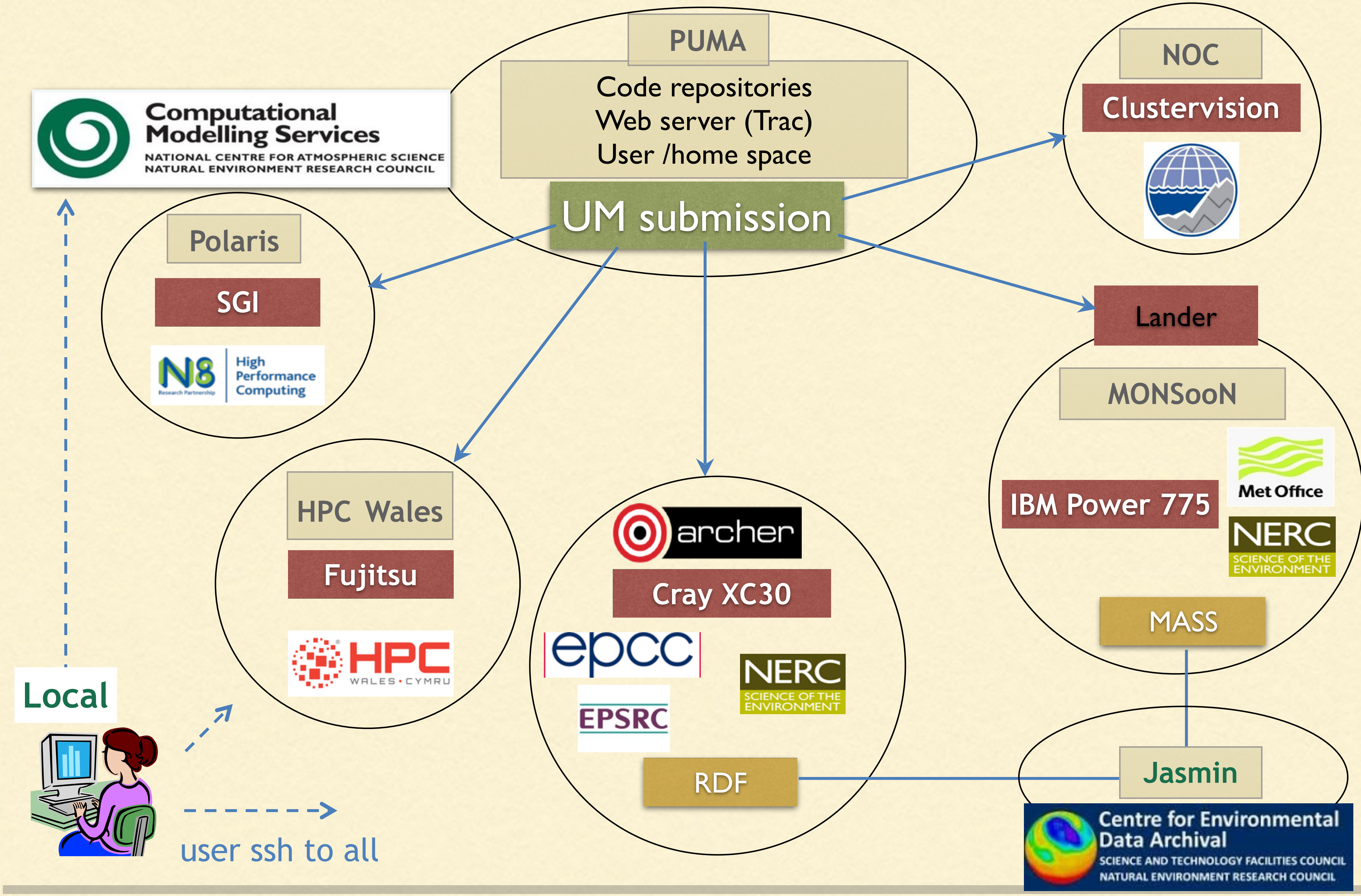
GLOBAL MODELS

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

N96	N144	N216	N320	N512	N768	N1024	N2048
(192 x 145)	(288 x 217)	(432 x 325)	(640 x 481)	(1024 x 769)	(1536 x 1152)	(2048 x 1536)	(4096 x 3073)
~135 km	~90 km	~60 km	~40 km	~25 km	~17 km	~12 km	~6 km

	NWP	Climate
Run length	10 day operational forecast, 15 day ensemble forecast	Months (seasonal) Years, decades, centuries+
Global resolution	Testing: N320 (40 km) with 15 min ts Operational: N768 (17 km) with 7.5 min ts	Low resolution: N96 (135 km) with 20 min ts High resolution: N512 (25 km) with 15 min ts
Dynamics	Non-bit reproducible	Bit-reproducible

NCAS SUPPORTED MACHINES



Climate modelling on Cray XC30

ARCHER - Cray XC30, Intel Xeon E5 v2 12C 2.700GHz,
Aries interconnect Cray Inc.
Cray Performance Analysis Tools

AIMS

- Performance scaling of the UM at different resolutions
- Performance analysis using Cray PAT tools
- Optimisations for MPI
- Cray Reveal for OPENMP

UM JOBS

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

JOB	COLUMN	ROWS	LAND	VERTICAL	TIME	RESOLUTION
N96	192	144	11271	85	20 min	135 km
N216	432	324	52614	85	15 min	60 km
N512	1024	768	280592	85	10 min	25 km

Number of columns and rows describes the grid of the global model in North-South and East-West (horizontal) direction respectively.
Land points refers to the number of simulated land points.
Vertical levels describes the vertical grid of atmosphere.
Time steps refers to the number of physics timesteps per simulated day.
Resolution refers to resolution of the global grid.

PERFORMANCE METRIC

$$M_{year} = \frac{1200}{T_{model}}$$

M_{year} - the number of model years simulated per day.

$$T_{model} = T_{wallclock} - T_{initial}$$

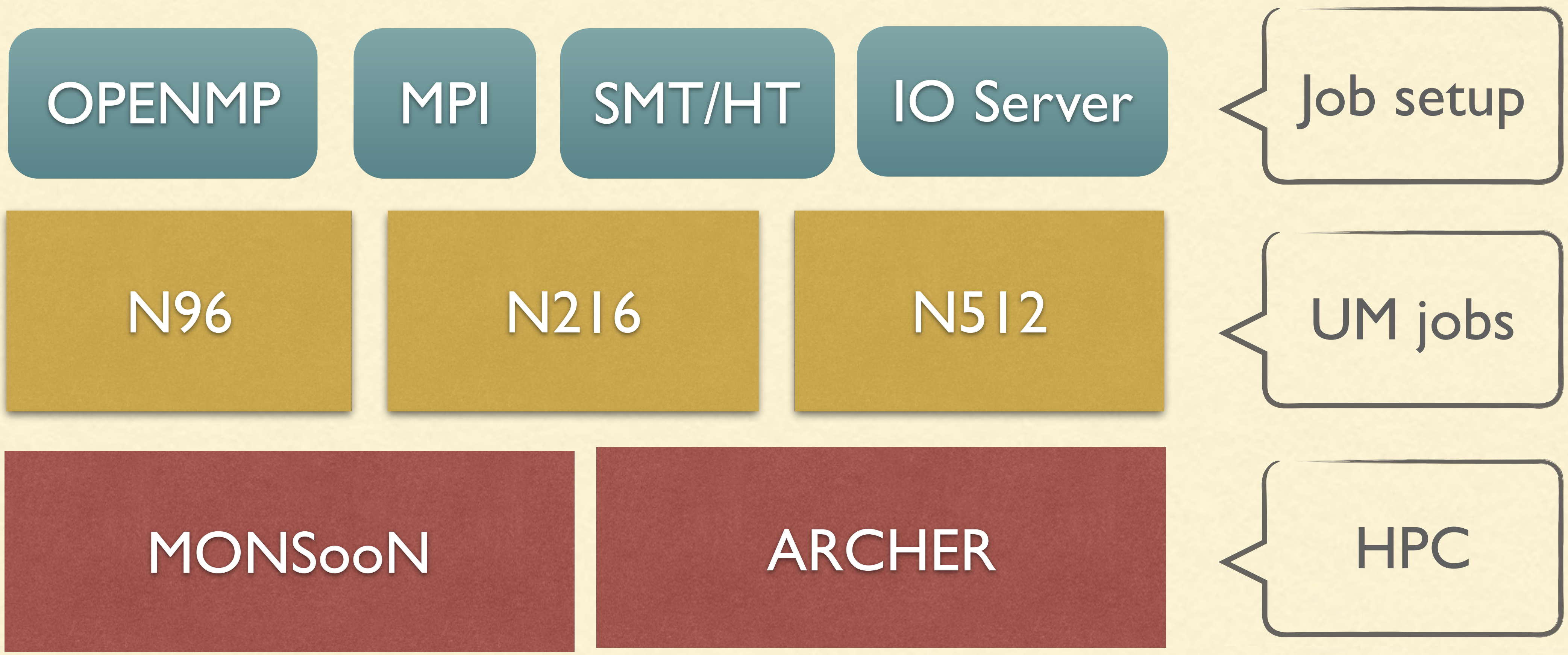
T_{model} - time to simulate 5 model days
$T_{wallclock}$ - total wallclock time
$T_{initial}$ - time to initialise

$$C = \frac{1}{M_{year}} \times n_{core} \times 24$$

C - Cost of simulating a model year in core-hours
n_{core} - number of physical cores

MODEL SETUP

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Hyper threading or Symmetric Multithreading	Hyper threading slows UM ; SMT achieved ~30% speedup
Bit reproducibility	-e m -s real64 -s integer64 -h O2 -hflex_mp=intolerant -h omp

PERFORMANCE SCALING

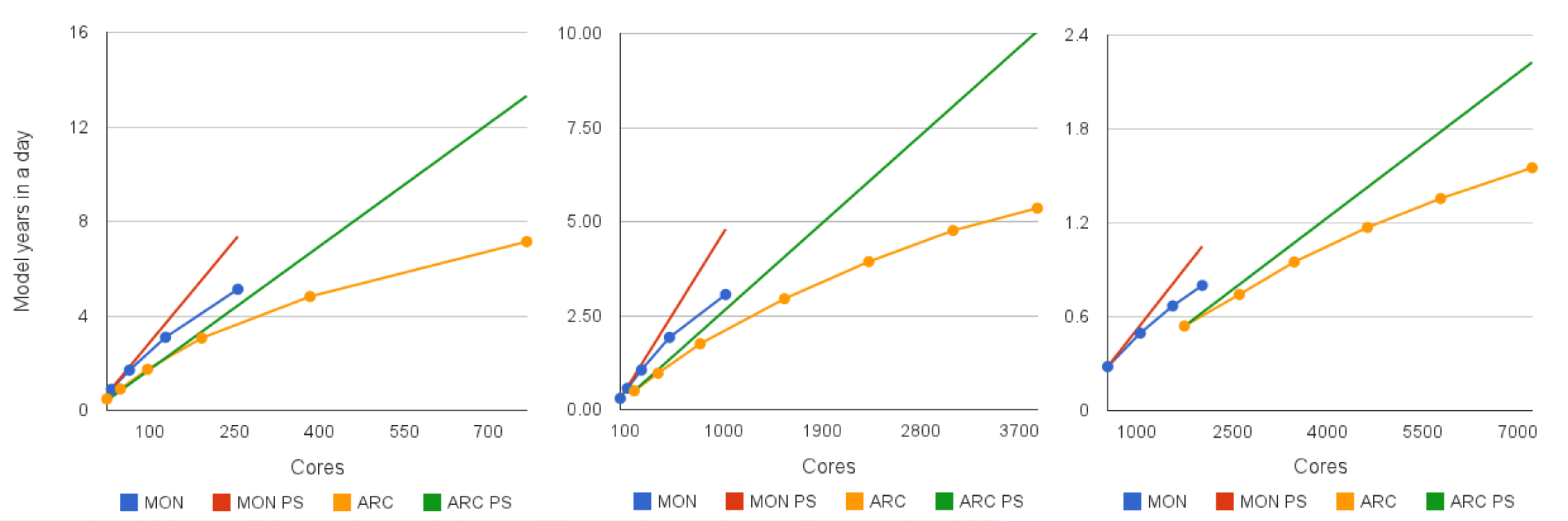
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

N96

12 MPI tasks x 2
OPENMP threads

N216

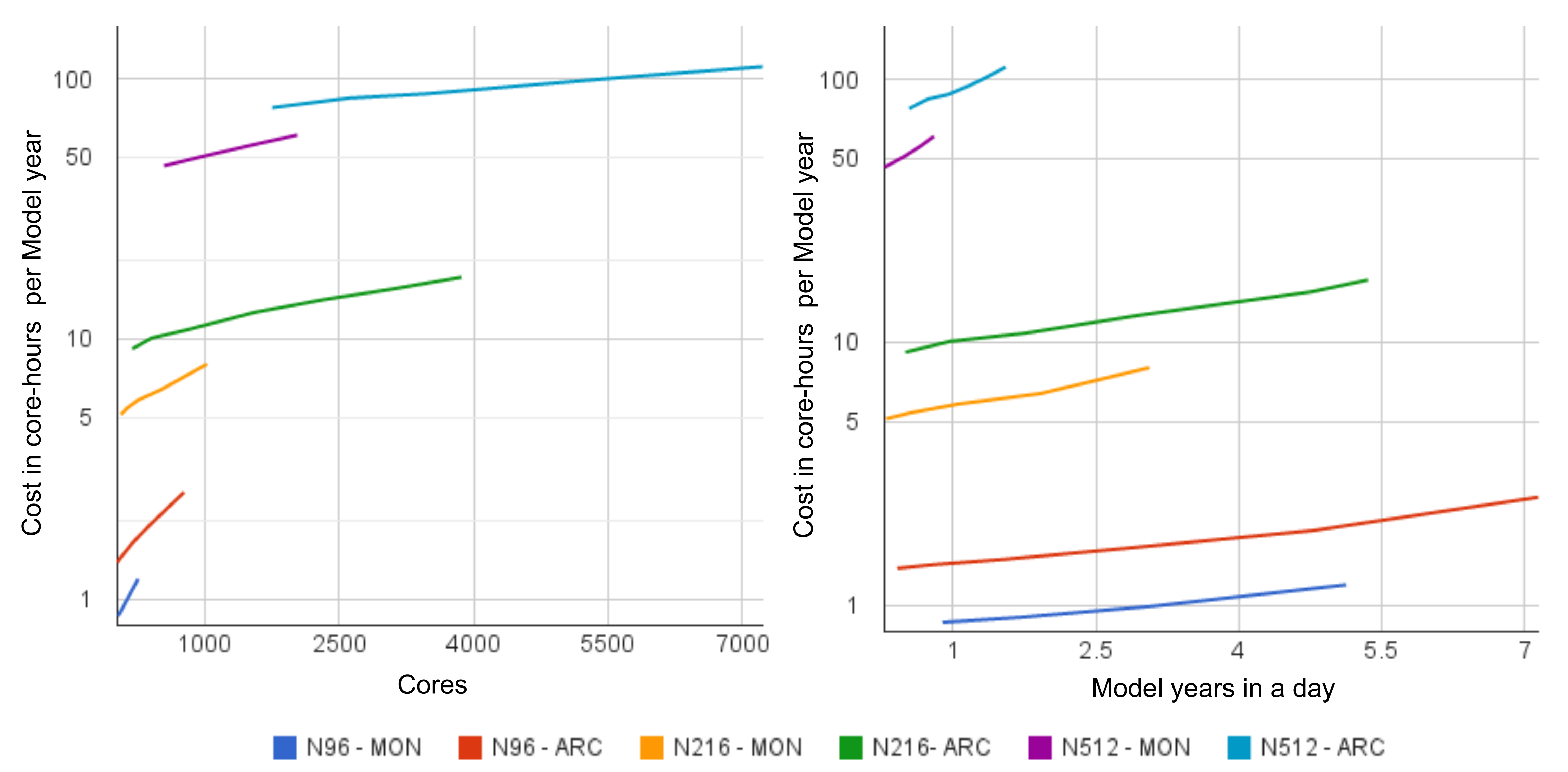
N512



Performance scaling of UM job on ARCHER(ARC) and MON- SooN(MON). Cores refers to the actual number of physical cores used and performance is measured as number of model years simulated in a day (Myear). MON PS and ARC PS refers to perfect scaling that can be expected on MONSooN and ARCHER respectively.

PERFORMANCE SCALING

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

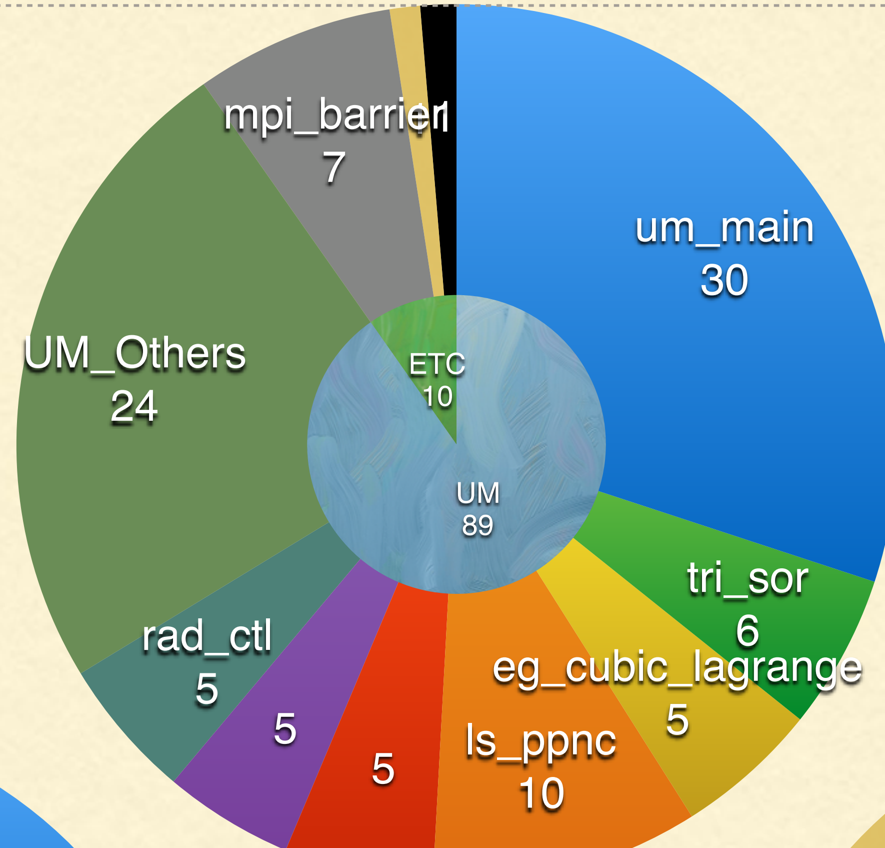


Cost of simulating a model year (C) of the UM job on ARCHER(ARC) and MONSooN(MON) compared to the number of physical cores - ncore (left) and Model years in a day Myear (right).

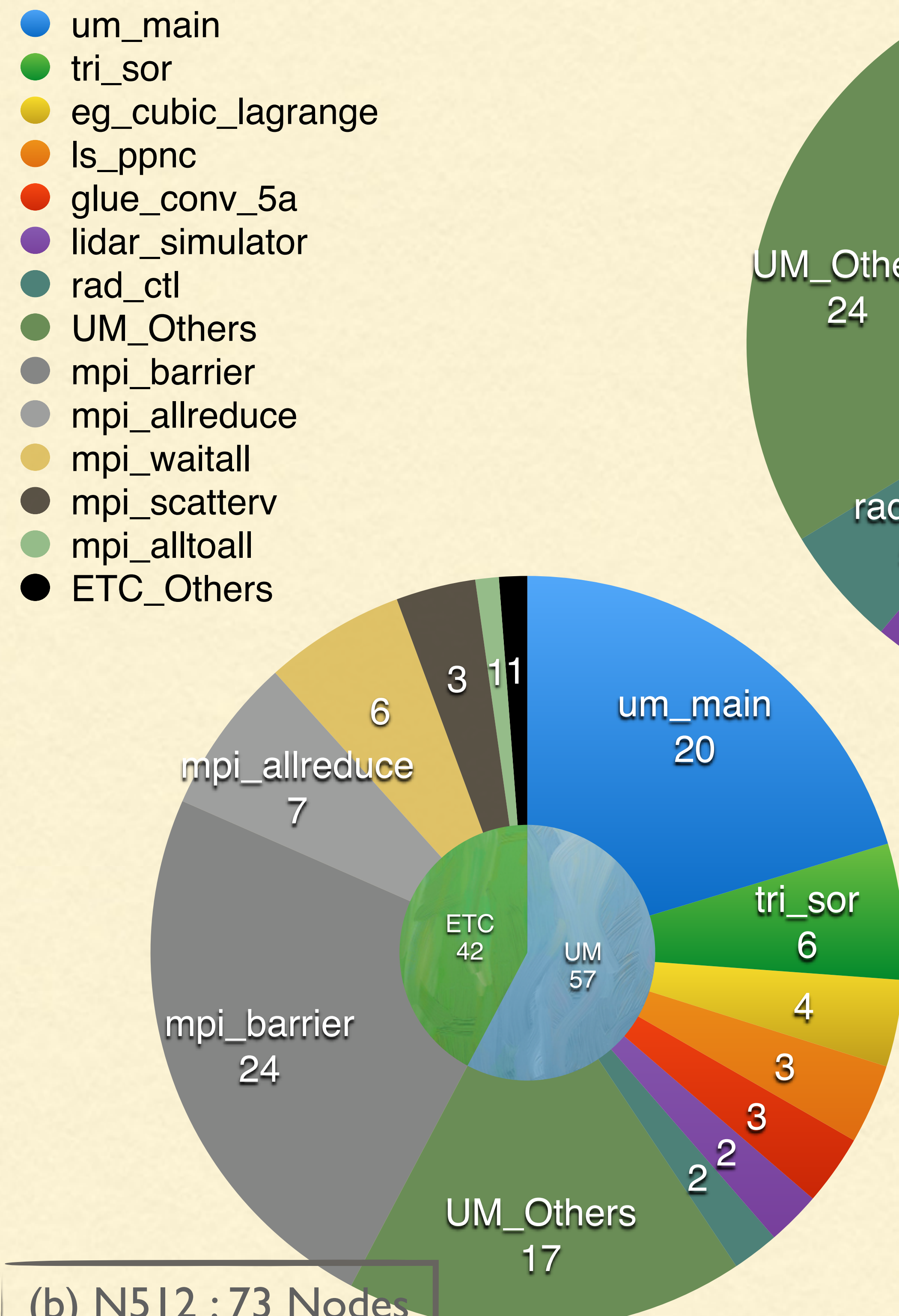
PERFORMANCE ANALYSIS

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

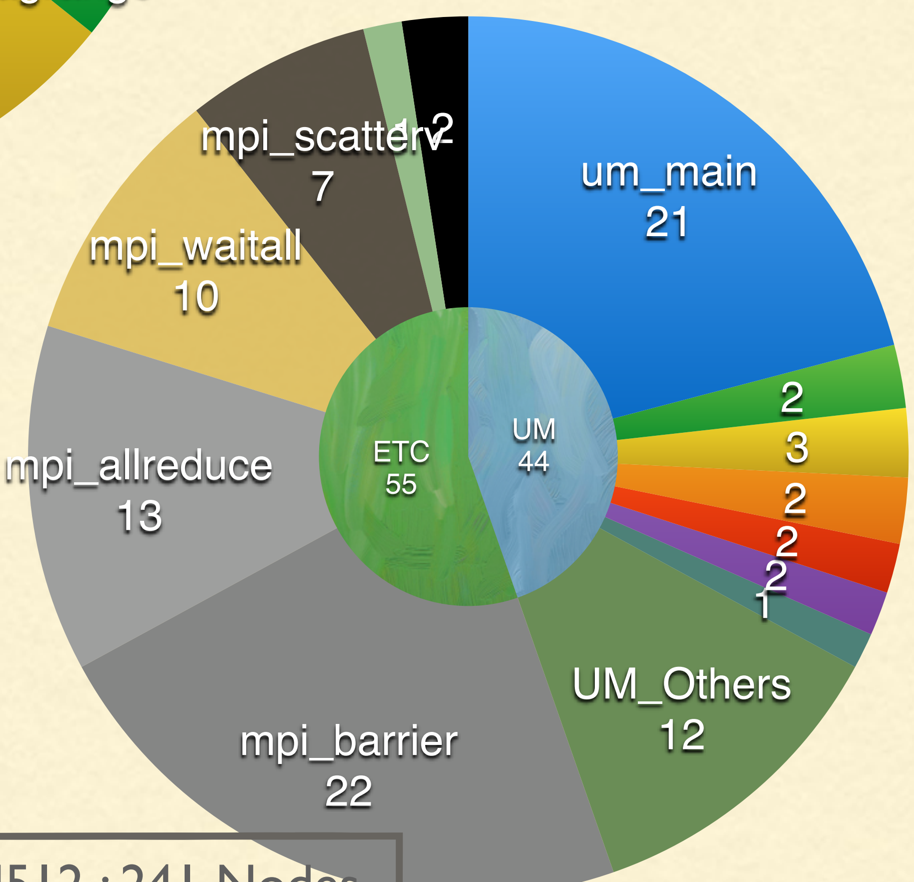
The inner pie chart shows the overall profile in which UM includes profile of all UM user code and ETC includes all other library calls.



(a) N96 : 1 Node



(b) N512 : 73 Nodes



(c) N512 : 241 Nodes

MPI RANK REORDER

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Nearest neighbour communications

MPI ranks - PE configuration 24 x 36

0	1	2	3	4	…	35
36	37	38	39	40	…	71
72	73	74	75	76	…	107
108	109	110	111	112	…	143
…	…	…	…	…	…	…
828	829	829	829	829	…	863

12 PEs per node

0	7	8
1	6	9
2	5	10
3	4	11

36	43	44
37	42	45
38	41	46
39	40	47

72	79	80
73	78	81
74	77	82
75	76	83

108	115	116
109	114	117
110	113	118
111	112	119

SMP Rank Order (Nodes 0,4,8,12)

MPI rank 37 (and 75) along with the ranks involved in nearest neighbour communications are highlighted. Rank order is based on using 12 MPI ranks per node on ARCHER.

GRID Rank Order (Nodes 0,1,2,3)

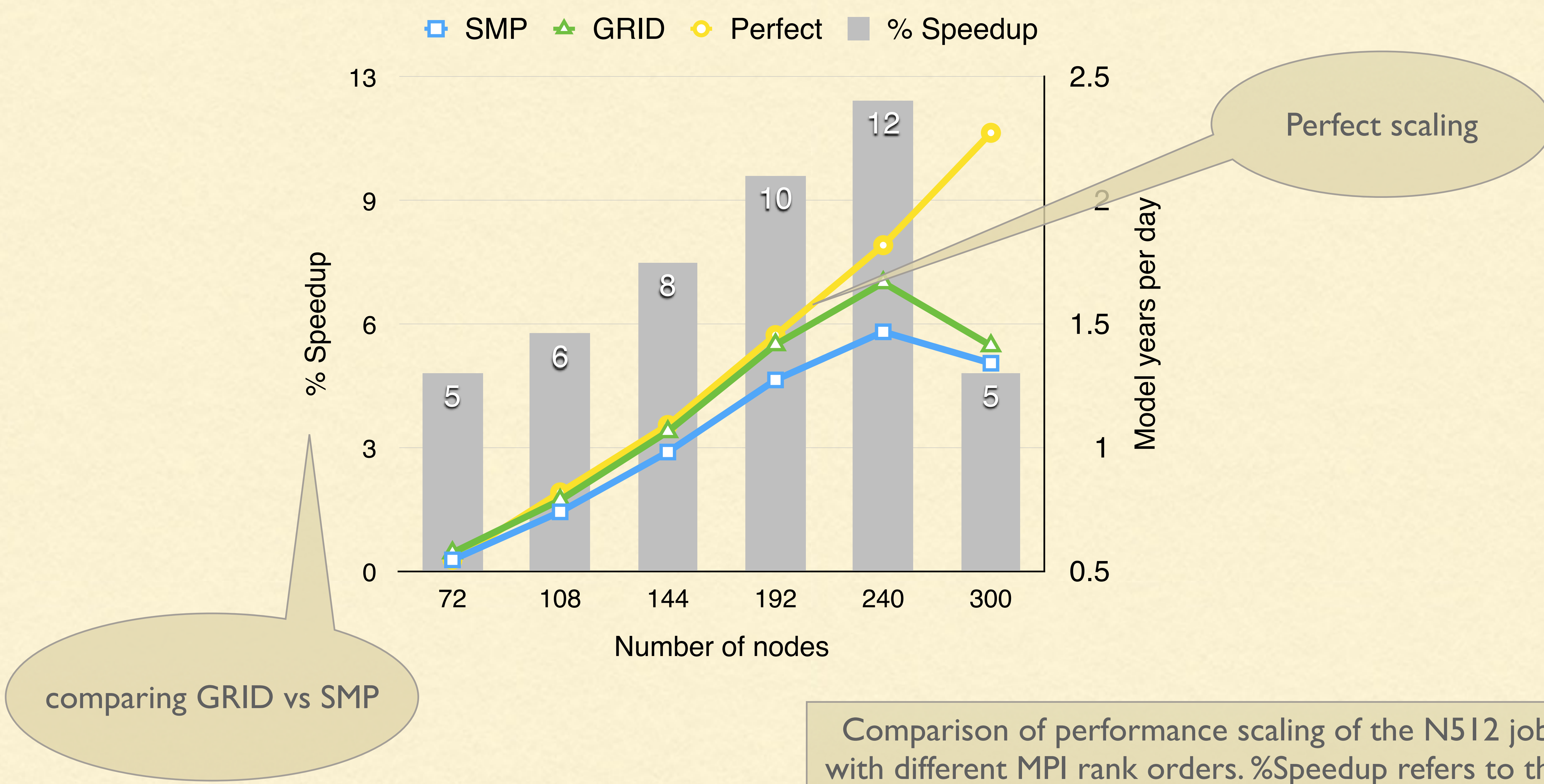
0	1	2
36	37	38
72	73	74
108	109	110

111	112	113
75	76	77
39	40	41
3	4	5

144	145	146
180	181	182
216	217	218
252	253	254

255	256	257
219	220	221
183	184	185
147	148	149

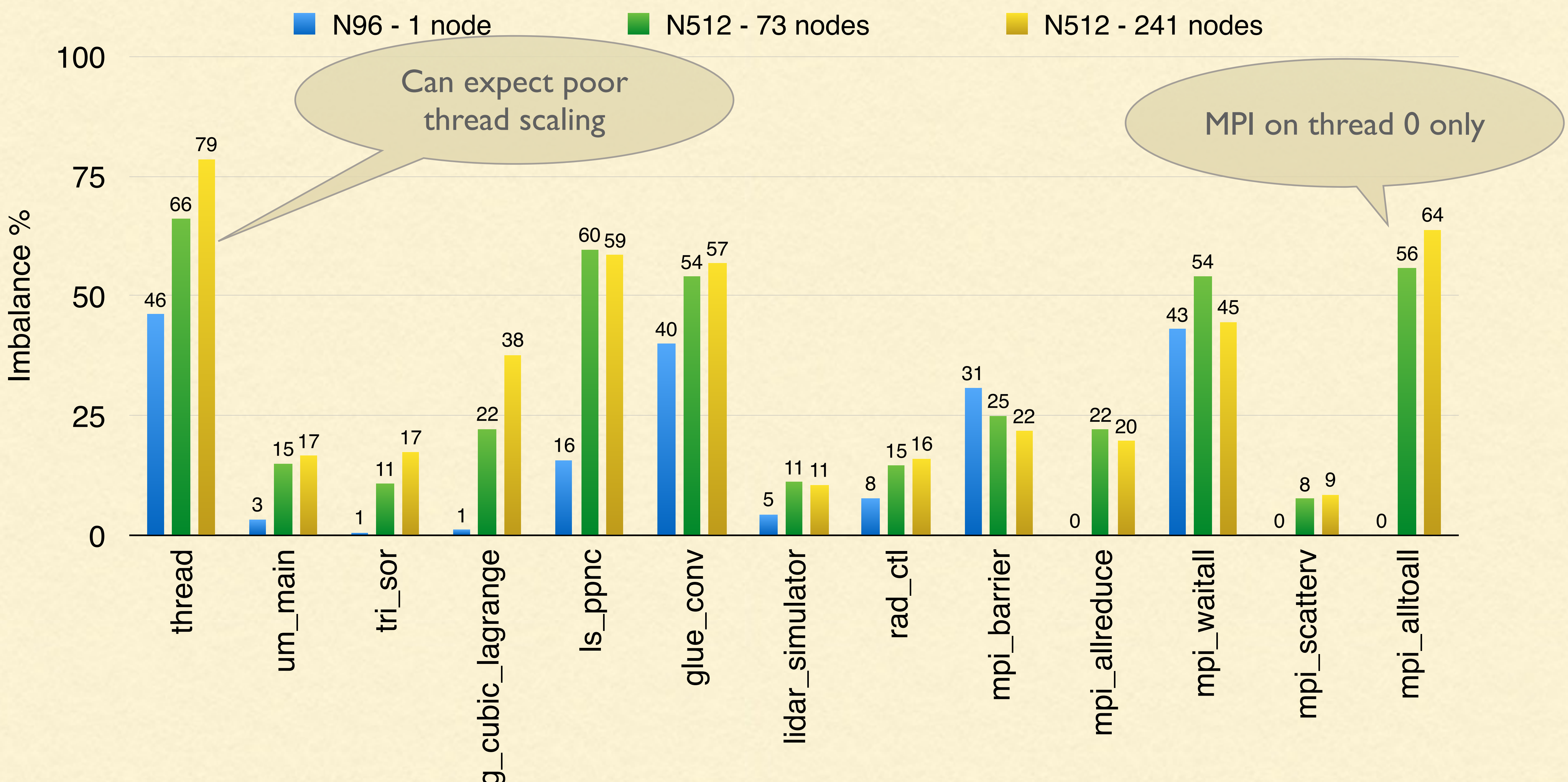
MPI RANK REORDER



Comparison of performance scaling of the N512 job with different MPI rank orders. %Speedup refers to the relative speedup achieved by using GRID rank order instead of SMP rank order.

LOAD IMBALANCE

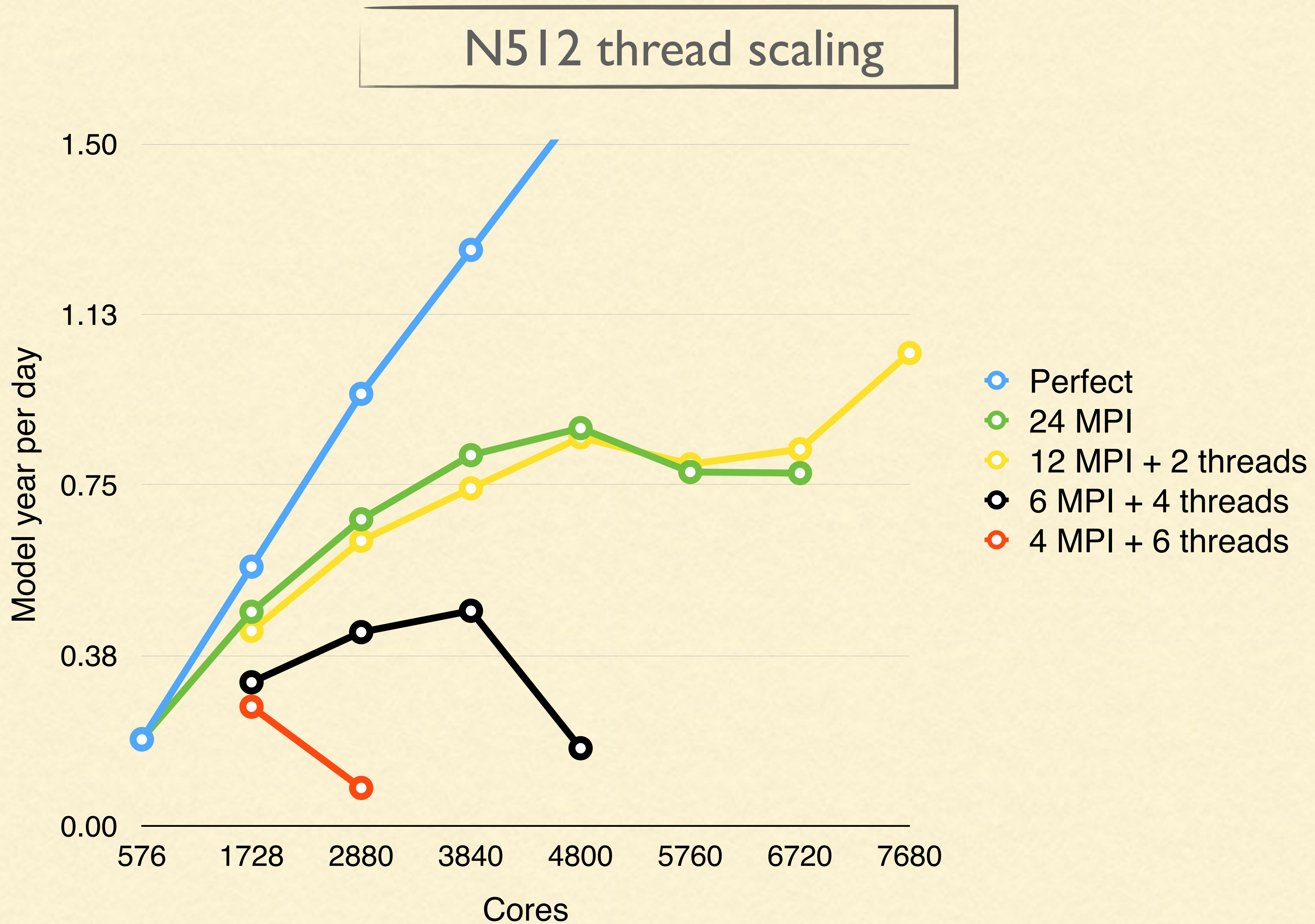
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Imbalance percentage of UM jobs : N96 job running on a single node; N512 job running on 73 nodes; N512 job running on 241 nodes; Thread imbalance percentage are relative to the set of threads used. Imbalance percentage of UM functions are relative to set of PEs.

PERFORMANCE SCALING

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



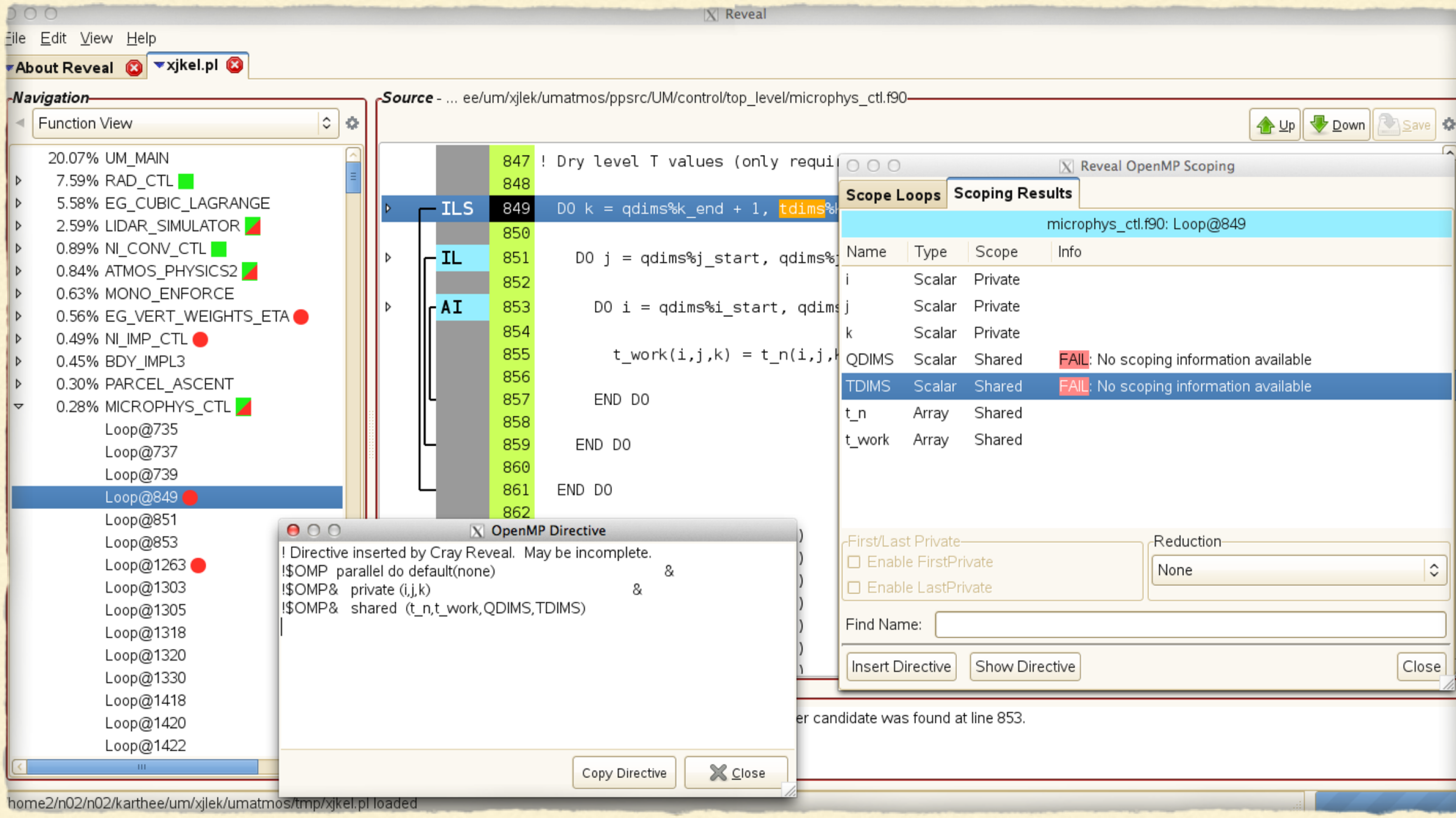
CRAY REVEAL

Cray Reveal - integrated performance analysis and code optimisation tool.

- provides loop analysis and scoping of serial loops.
- suggests OPENMP directive that can be inserted to a loop.
- can attach the performance data collected during execution to identify profile of loops.
- requires knowledge of OPENMP to resolve conflicts and issues.
- works only with Cray compiling environment.
- does not support tasks, barrier, critical or atomic regions.
- For more details - refer Cray documentation (not much)

CRAY REVEAL

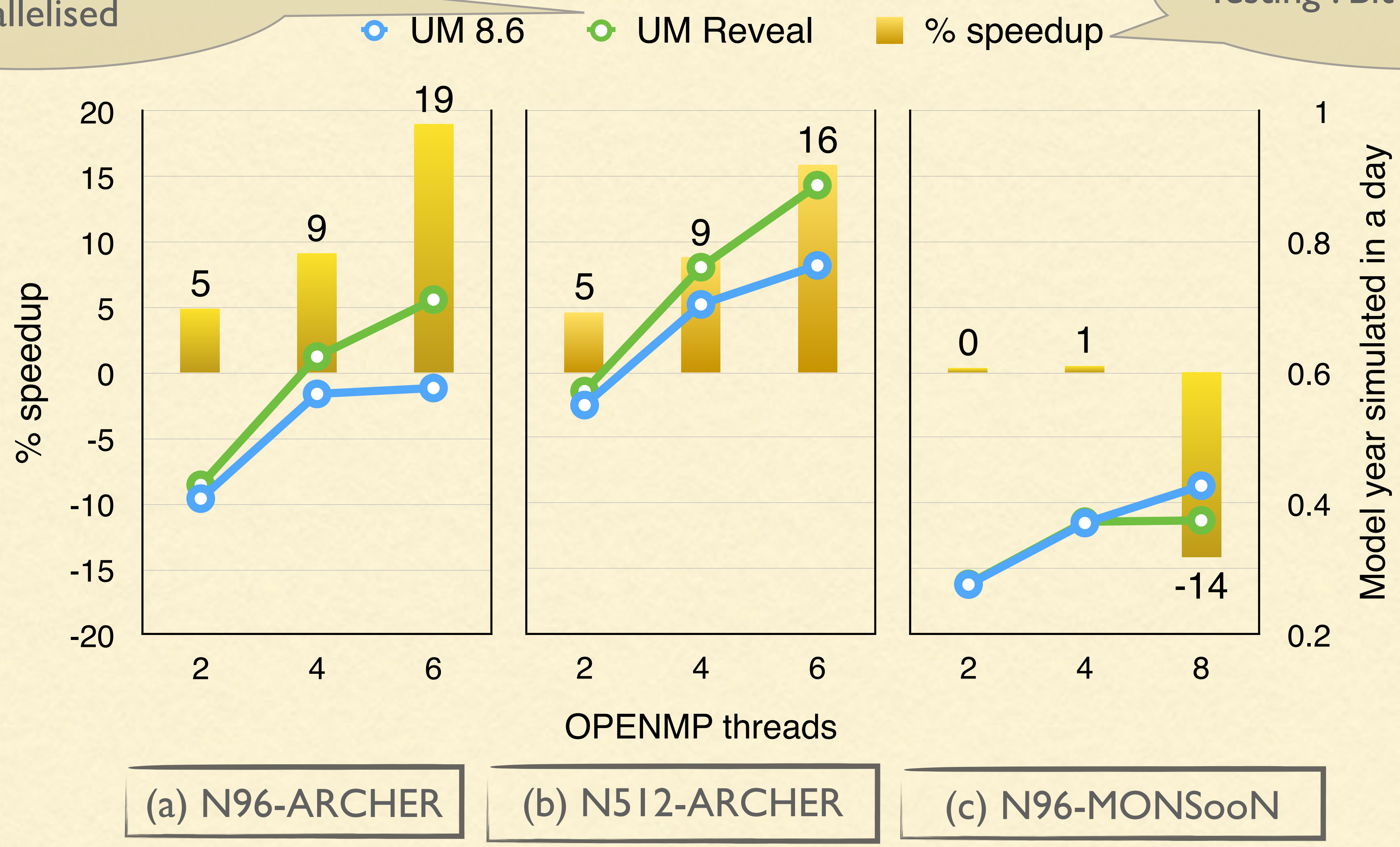
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



CRAY REVEAL

2389 serial loops parallelised

Testing : Bit reproducibility

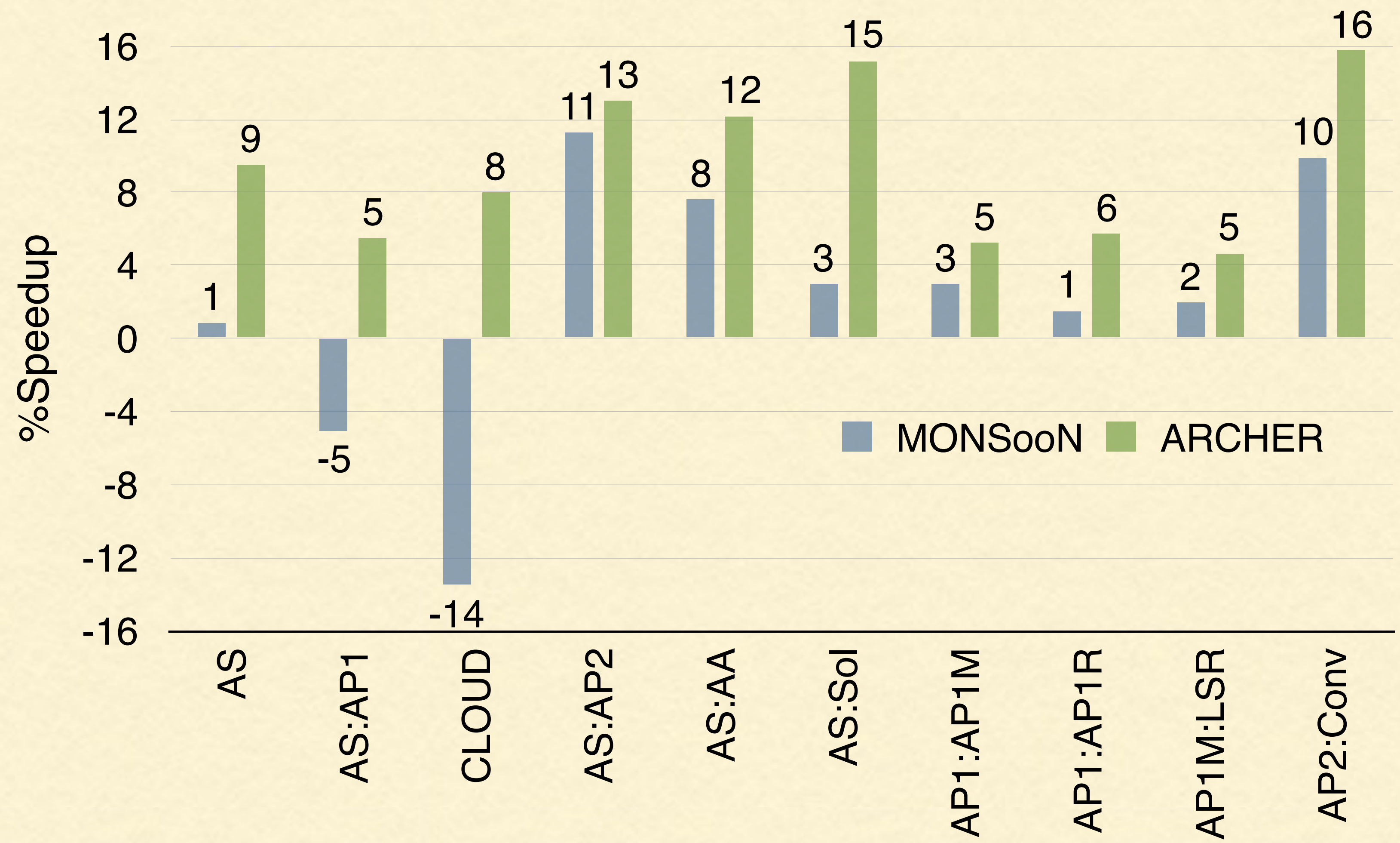


Performance scaling of the UM jobs using increasing number of OpenMP threads on ARCHER/MONSooN. Wallclock time refers to the time taken to complete 2 model days. UM8.6 refers to the original UM code and UMReveal to the code with new OpenMP directives. %Speedup is measured as a relative performance improvement achieved by adding new OpenMP directives.

OPENMP PERFORMANCE

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

AS	Atm Step
AS:AP1	Atmos Physcis 1
CLOUD	Cloud Simulator
AS:AP2	Atmos Physcis 2
AS:AA	S-L Advect
AS:Sol	Solver
AP1:AP1M	Microphys
AP1:AP1R	Radiation
AP1M:LSR	LS Rain
AP2:Conv	Convection



%Speedup of UM functions on MONSooN and ARCHER. %Speedup is measured as the relative improvement achieved by adding OpenMP regions. Performance of UM is measured using 6 MPI and 4 OpenMP threads per node. In X-axis labels, the calltree of the function is specified using ':' delimiter. For example in AS:AP1, 'Atm Step' calls function 'Atmos Physics 1'.

VECTORISATION

- ARCHER compute nodes contain two 2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processors
- supports AVX Instruction set extensions - 256-bit vector SIMD extension
- AVX floating point arithmetic 8*x* faster compared to scalar
- AVX2 and AVX512 also available

VECTORISATION IN UM

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

	LOOPS	%
TOTAL LOOPS	70000	100.00
VECTORIZED	9769	13.96
FUSED	6418	9.17
REPLACED WITH LIBRARY CALLS	2207	3.15
PARTIALLY VECTORIZED	2089	2.98
NOT VECTORIZED	49522	70.75

VECTORISATION IN UM

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

NOT VECTORIZED BECAUSE A BETTER CANDIDATE WAS FOUND	22382	31.97
NOT VECTORIZED BECAUSE OF A POTENTIAL REASSOCIATION ISSUE	10418	14.88
NOT VECTORIZED BECAUSE A RECURRENCE	5762	8.23
NOT VECTORIZED BECAUSE IT CONTAINS A CALL TO SUBROUTINE/ FUNCTION/IRREGULAR EXPRESSION	4876	6.97
NOT VECTORIZED BECAUSE IT DOES NOT MAP WELL ONTO THE TARGET ARCHITECTURE	3482	4.97
WAS NOT VECTORIZED BECAUSE THE TARGET ARRAY (X1) WOULD REQUIRE RANK EXPANSION	371	0.53
NOT VECTORIZED BECAUSE IT CONTAINS A REFERENCE TO A NON-VECTOR INTRINSIC	365	0.52
NOT VECTORIZED BECAUSE THE ITERATION SPACE IS TOO IRREGULAR	408	0.58
NOT VECTORIZED BECAUSE OF UNKNOWN REASON	1458	2.08

VECTORISATION IN UM

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

<u>*not vectorized because of a potential reassociation issue*</u>

If -h flex_mp=intolerant is specified on the command line, then loops are rejected as vector candidates if they contain floating point or complex operations which can potentially cause subtle result differences due to optimization variances between the main vector loop body and any left-over remainder iterations.

```
DO k=1,qdims%k_end
  DO j=1,rows
    DO i=1,row_length
      qrain_inc_step(i,j,k)   = qrain_star(i,j,k)   - qrain_n(i,j,k)
      qrain_conv(i,j,k)   = qrain_n(i,j,k)                          &
                 + fraction_step*qrain_inc_step(i,j,k)
    END DO ! i
  END DO ! j
END DO ! k
```

N	22382	31.97
N	10418	14.88
N	5762	8.23
N FU	4876	6.97
N A	3482	4.97
V RI	371	0.53
N VI	365	0.52
N	408	0.58
N	1458	2.08

VECTORISATION IN UM

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

__not vectorized because of recurrence__

Scalar code was generated for the loop because it contains a linear recurrence. The following loop would cause this message to be issued:

```
      DO I = 1,100
            A(I) = A(I-1)
      ENDDO

   DO j = first_row, last_row
       DO I = row_start_pt, row_end_pt

           iloc = LBC_address(i,j)

           rho_lbc(iloc,k) = p_zero/(R*temp)                    &
                       *exner_lbc(iloc,k)**((1.0-kappa)/kappa)

           exner_lbc(iloc,k+1) = exner_lbc(iloc,k)

       END DO
   END DO
```

N	22382	31.97
N	10418	14.88
N	5762	8.23
N FU	4876	6.97
N A	3482	4.97
V RI	371	0.53
N VI	365	0.52
N	408	0.58
N	1458	2.08

VECTORISATION IN UM

not vectorized because it does not map well to the target architecture

The loop contains too many operations that have no clean vector equivalent in hardware for the targeted architecture, and has been left to run as a purely scalar loop. Although the loop is vectorizable from a dependency and idiom standpoint, it would be unprofitable to emulate vector execution using scalar operations.

On other architectures with different hardware capabilities, this loop may be cleanly vectorized.

```
DO k = 1, dim_k_out
  DO j = 1, dim_j_out
    DO i = 1, dim_i_out

      i_out(i,j,k) = i_out(i,j,k) - datastart(1) + 1

    END DO  ! i = 1, dim_i_out * dim_j_out * dim_k_out
  END DO
END DO
```

22382	31.97
10418	14.88
5762	8.23
4876	6.97
3482	4.97
371	0.53
365	0.52
408	0.58
1458	2.08

SUMMARY

On Cray XC30 , UM performance is characterised by

- 2 OPENMP threads has a load imbalance of 46% that increases as the UM is scaled to higher resolution
- Newly added OPENMP directives result in 5 to 19% speedup
- ~57% of the loops cannot be vectorised by the compiler. Can be improved by not enforcing bit reproducibility
- Message passing does not scale well above 2880 MPI ranks as it consumes more than 50% of the total wallclock time.
- Using a GRID rank reorder results in 5 to 12% speedup.

GUNGHO FOR FUTURE

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

TERASCALE

PETASCALE

THANK YOU

EXASCALE