Performance of the Met Office Unified Model on Intel Xeon clusters

16th ECMWF Workshop on HPC in Meteorology

27-31 October 2014, Reading, UK

Ilia Bermous

the Australian Bureau of Meteorology

Australian Government

Bureau of Meteorology

The Centre for Australian Weather and Climate Research

A partnership between CSIRO and the Bureau of Meteorology

CSIRO

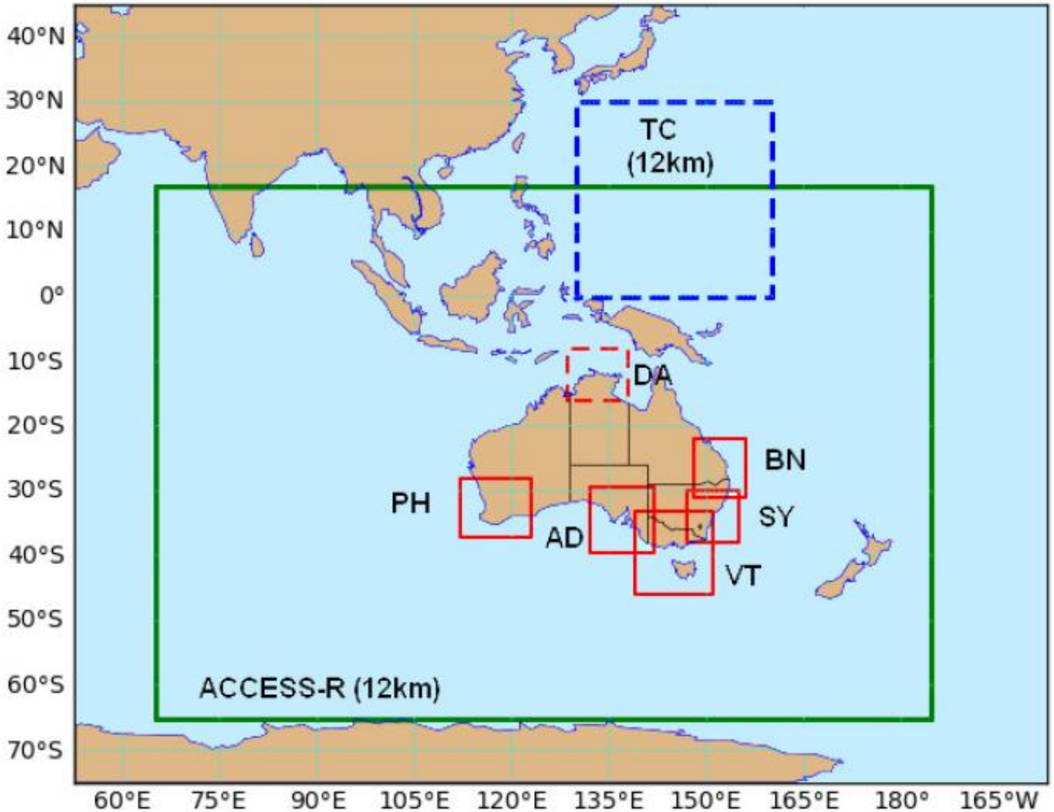
Presentation outline

- The Australian Bureau of Meteorology (BoM) forecast systems and their coming operational upgrade
- Description of Intel Xeon systems and software used
- Performance scaling analysis for limited area and global forecast models
- Practical benefits of the recommended method

Australian Government
Bureau of Meteorology

CSIRO

Forecast systems operationally run in BoM



ACCESS-G

ACCESS-R

ACCESS-C

ACCESS = the Australian Community Climate and Earth-System Simulator

G – global
R – regional
C – city

Forecast component of the systems is based on the UK Met Office Unified Model (UM)

APS1 → APS2 upgrade in early 2015

APS1 – Australian Parallel Suite 1

Australian Government
Bureau of Meteorology

CSIRO

APS2 upgrade for ACCESS-G

APS1: N320L70 => 40km & 70 levels

APS2: N512L70 => 25km & 70 levels

	APS1	APS2	factor
grid	640x481x70	1024x769x70	2.56
forecast length	10 days	10 days	1
time step	12 min	10 min	1.2

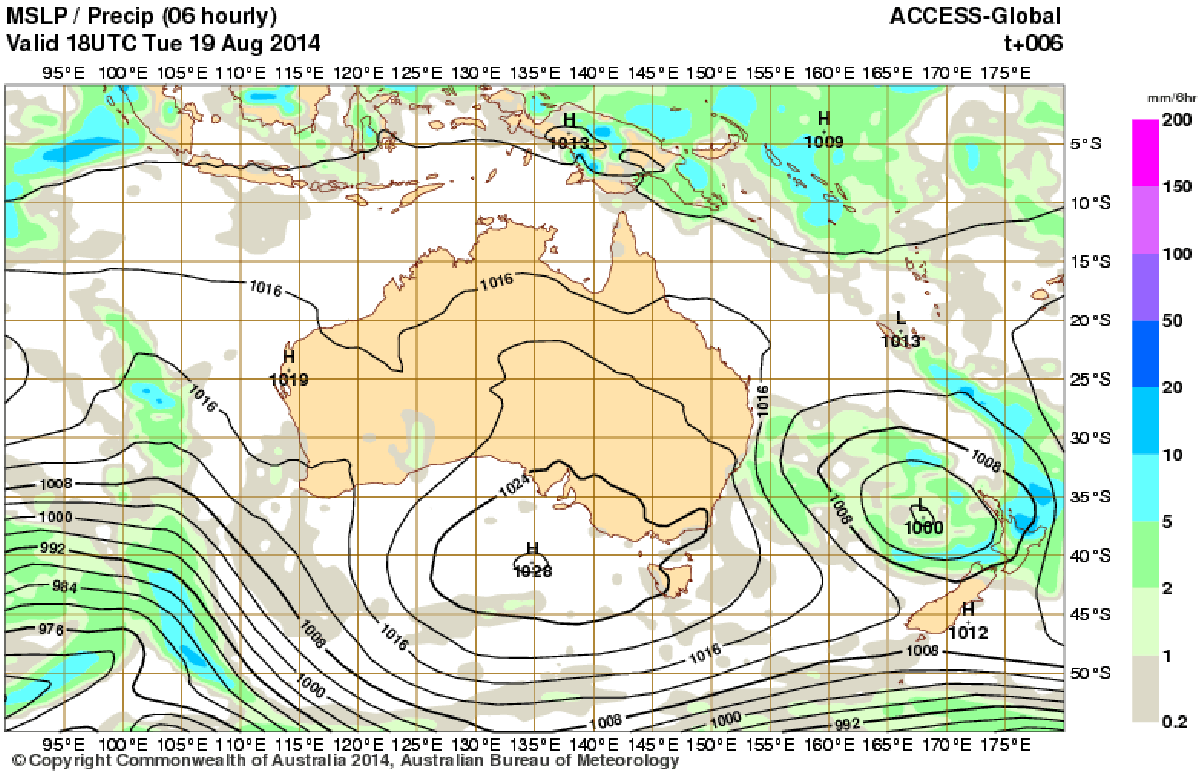
Total factor 3.07

Upgrade in the UM version: UM7.5 (APS1) → UM8.2 (APS2)
"New Dynamics" dynamical core

Australian Government
Bureau of Meteorology

CSIRO

APS2 upgrade for ACCESS-R



Regional or R12 model

resolution: 12 km 70 levels

grid: 1088x746x70

forecast: 75 hours

time step: 5 min

runs 4 times daily

APS1 → APS2: no changes in the resolution and the forecast length

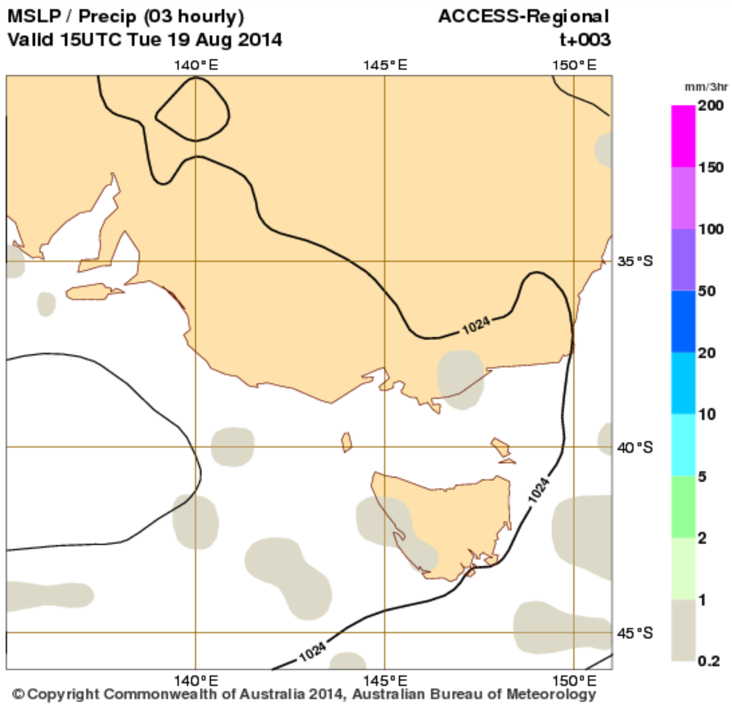
UM7.5 (APS1) → UM8.4 (APS2)

"New Dynamics" dynamical core

Australian Government
Bureau of Meteorology

CSIRO

APS2 upgrade for ACCESS-C



City models for 6 domains:

Adelaide (AD), Brisbane (BN), Darwin (DN), Perth (PH), Sydney (SY), Victoria-Tasmania (VT)

APS1: 4 km & 70 levels

APS2: 1.5 km & 70 levels

UM7.5 (APS1) → UM8.2 (APS2)
"New Dynamics" dynamical core

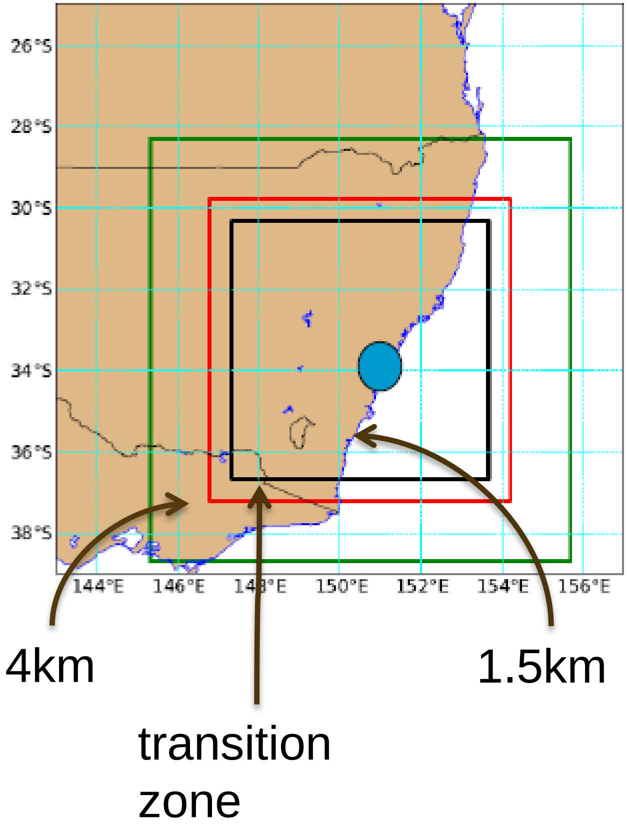
	APS1	APS2	factor
grid (VT)	334x362x70	890x964x70	7.10
forecast length	39 hours	39 hours	1
time step	100 sec	50 sec	2

Total factor 14.2

Australian Government
Bureau of Meteorology

CSIRO

Sydney UKV 1.5km forecast model

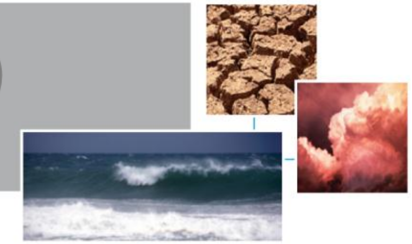


- **an experimental high-resolution system for Sydney domain, created in 2012**
- **uses the UKV modelling concept**
- **grid: E-W x N-S of 648 x 720 with 70 levels**

Australian Government
Bureau of Meteorology

CSIRO

UK Met Office Unified Model (UM) at BoM



- From Sep 2009 UM has been used operationally (research licence signed between BoM, CSIRO and the Met Office) as an NWP forecast component at BoM initially on NEC SX-6 then on Intel Xeon clusters

- UM was developed at the end of 1980's – beginning of 1990's on an MPP system using MPI as a single level of parallelism

- From the mid of 2000's the hybrid parallel programming paradigm was introduced in the UM7.0 and since then the OpenMP implementation has been consistently improving

Technical characteristics of 3 Intel Xeon clusters

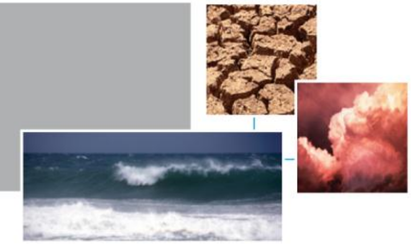
	Solar, BoM 49Tflops, 2009	Ngamai, BoM 138 Tflops, 2013	Raijin, NCI (ANU) 1.2 Pflops, 2013
processor	Intel Xeon X5570	Intel Xeon E5-2640	Intel Xeon E5-2670
nodes/cores	576/4608	576/6912	3592/57472
memory per node	24GB	64GB	32GB; 64GB; 128GB
node peak perf	85 GFLOPS	240 GFLOPS	332.8 GFLOPS
node max memory bandwidth	64 GB/s	85.3 GB/s	102.4 GB/s
Byte/Flop	0.753	0.355	0.308
infiniband interconnect	QDR	QDR	FDR
turbo boost	No	No	Yes
hyper-threading	No	No	No

NCI - the National Computational Infrastructure
ANU - the Australian National University in Canberra

Australian Government
Bureau of Meteorology

CSIRO

Software on the systems



➢Intel compiler

v12.1.8.273;   v14.0.1.106   with the following major compilation options

`-O3 -fp-model precise` (the latter option is for results reproducibility on a rerun)

`-xavx` (Intel Xeon E5, Ngamai & Raijin) for advanced vector extensions

`-g -traceback` (to get a failed subroutine call sequence, no impact on the performance)

➢MPI libraries

OpenMPI: v1.6.5; v1.7.4;   v1.8.2 is being tested now

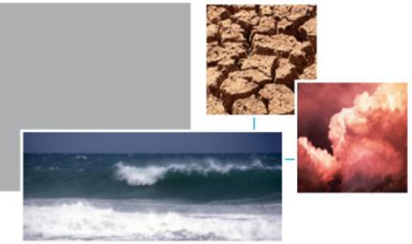
Intel MPI: v4.1.1.036 (for UM applications OpenMPI gives better performance)

➢ Lustre file striping on all systems

➢ 2 important aspects in the executables built on BoM and NCI systems

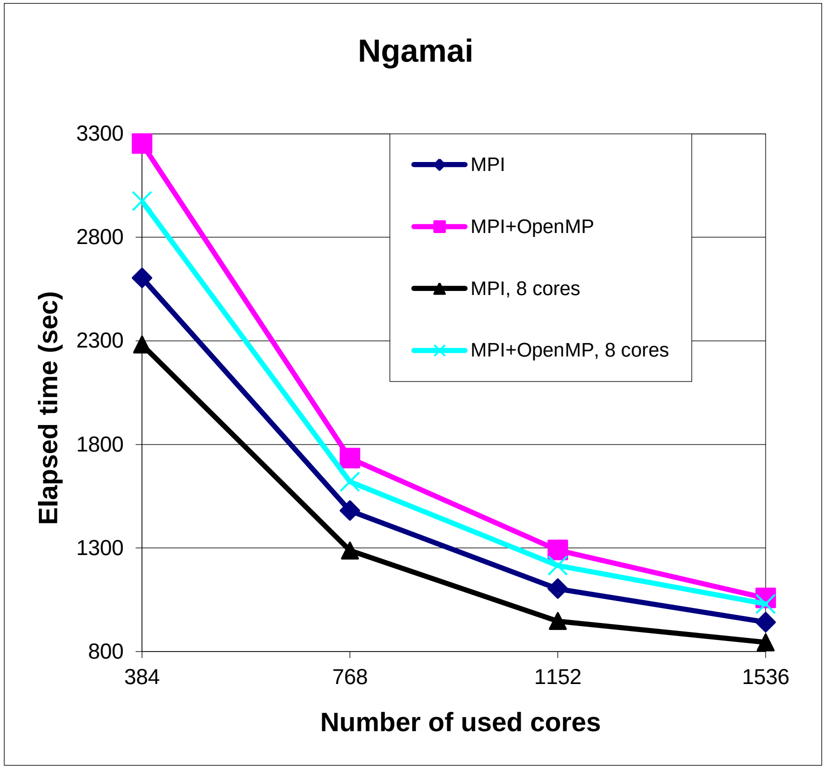
- **compatibility of executables across Ngamai & Raijin systems**
- **reproducibility of the numerical results on these systems**

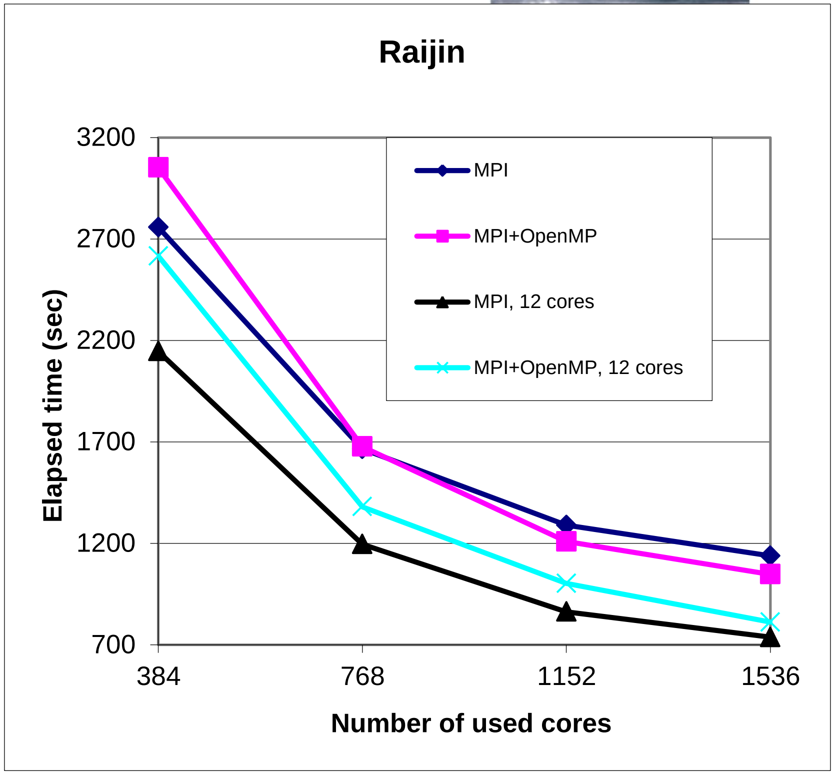
Some details on the application runs



- **UKV Sydney**
  - UM8.2 with Intel12.1.8.273 and OpenMPI1.6.5
  - 25 hour simulation with 50 sec time step
  - output size: 18 GB
  - the IO servers were not used due to relatively small I/O
- **APS2 ACCESS-G (N512L70)**
  - UM8.2 with Intel14.0.1.106 and OpenMPI1.7.4
  - 3 day simulation with 10 min time step => 432 time steps
  - output size: 137 GB
  - UM IO servers were used => this requires the usage of multithreading with at least 2 OpenMP threads

UKV performance scaling: pure MPI versus MPI / OpenMP hybrid



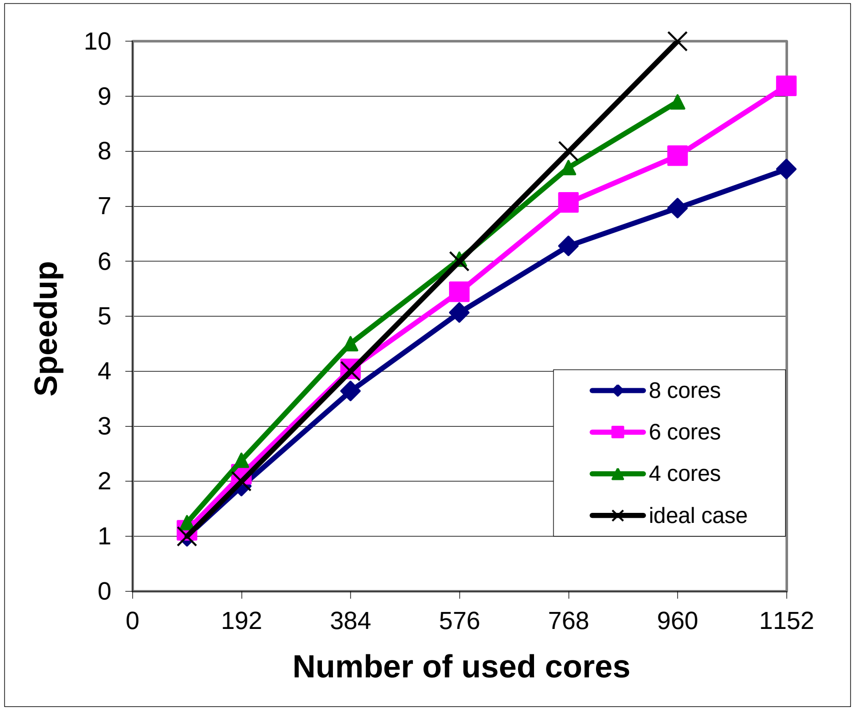


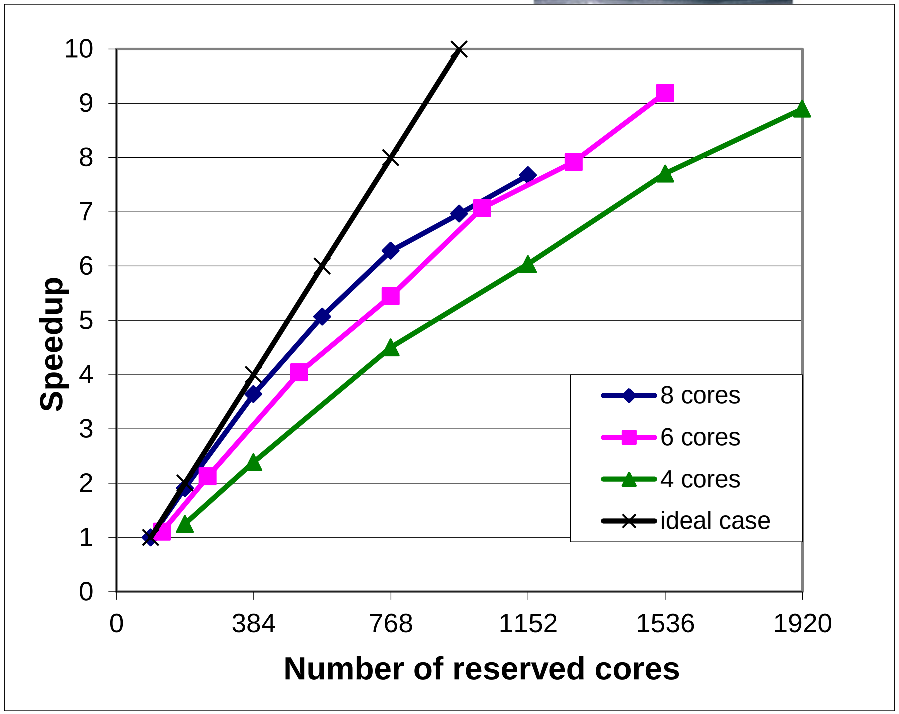
- **2 OpenMP threads were used in the hybrid parallelism case**
- **8 cores-per-node on Ngamai and 12 cores-per-node on Raijin represent partial node usage reserving full nodes but using 8 from 12 cores on Ngamai and 12 from 16 cores on Raijin**
- **"used cores" describes the actual cores used in a run**

Australian Government
Bureau of Meteorology

CSIRO

UKV scalability on Solar using pure MPI



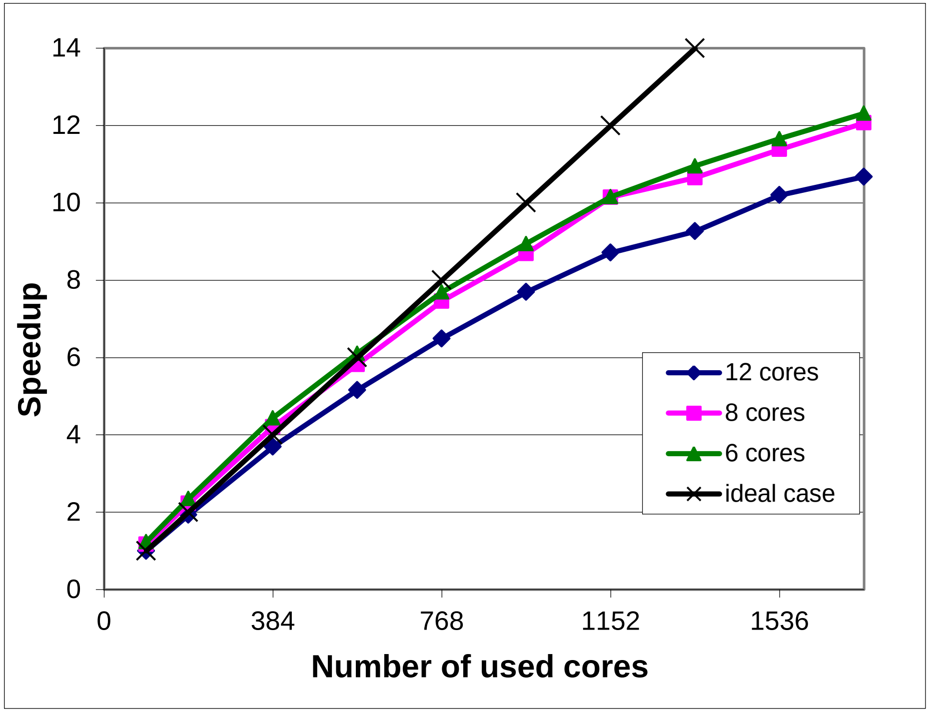


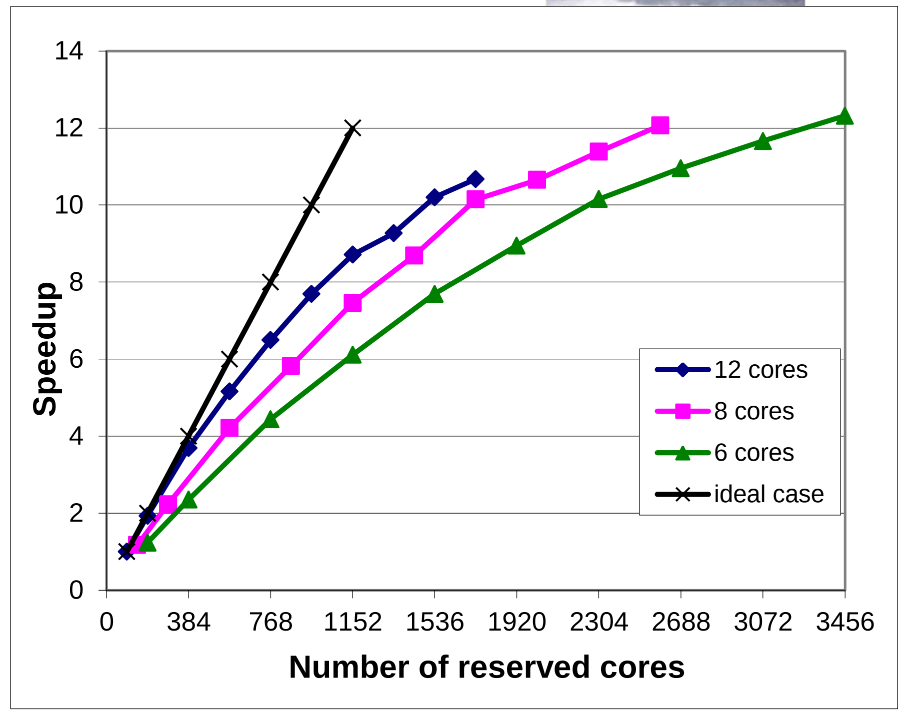
- **speedup was calculated in relation to the elapsed time of 11488sec obtained for a 96 core run on fully committed nodes**
- **usage of partial nodes improved run-times with 6 cores-per-node by 6.9-16.5%, with 4 cores-per-node a further reduction of 7.3-10.4% was achieved**
- **the most efficient system usage was on fully committed nodes up to 1152 core usage**
- **"reserved" cores describes the total number of the actually used and unused cores**

Australian Government
Bureau of Meteorology

CSIRO

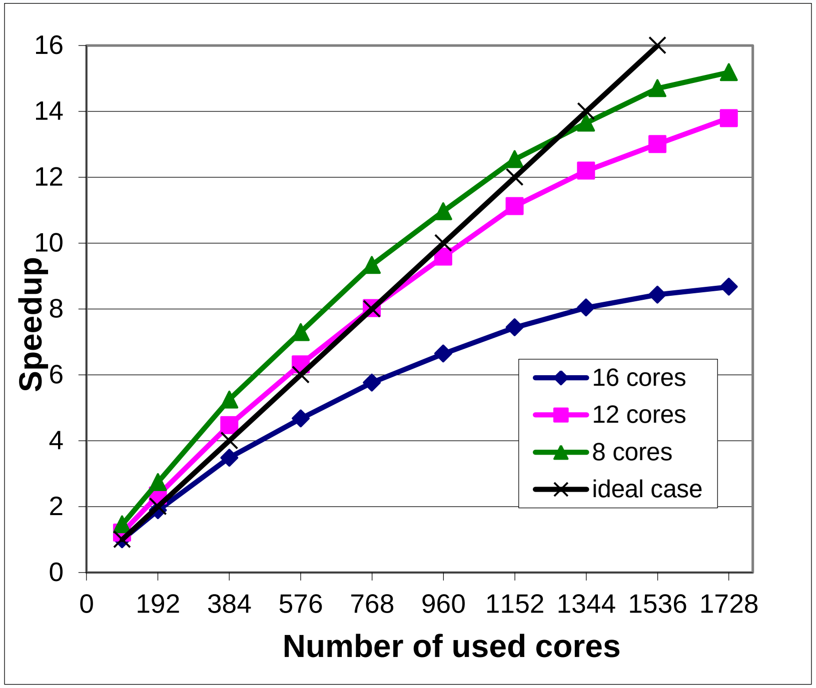
UKV scalability on Ngamai using pure MPI

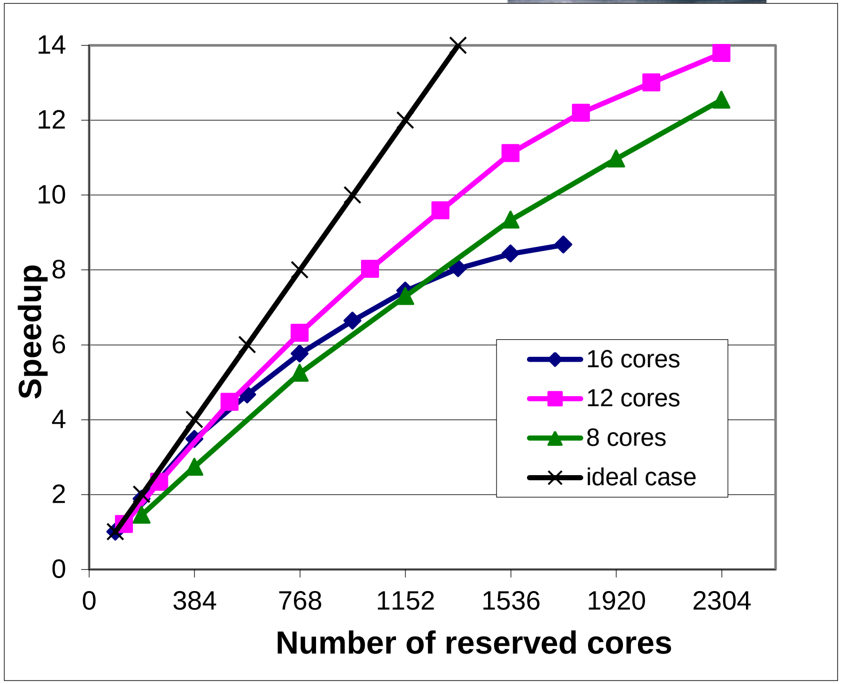




- **speedup was calculated in relation to the elapsed time of 9608sec obtained for a 96 core run on fully committed nodes**
- **usage of partial nodes improved run-times by 10.4-14.6% with 8 cores-per-node, less than 5.5% of performance improvement was achieved with half utilised nodes**
- **as on Solar the most efficient system usage was on fully committed nodes up to 1728 core usage**

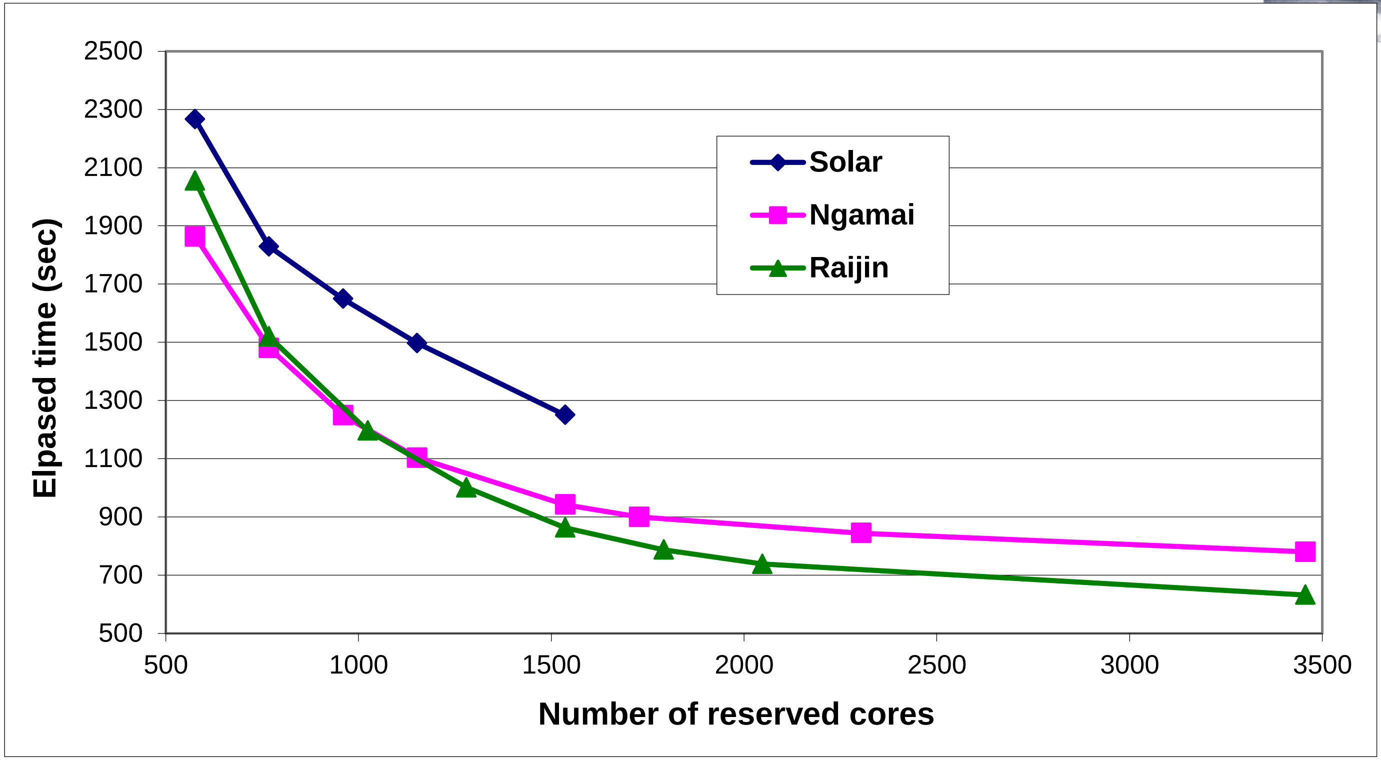
UKV scalability on Raijin using pure MPI





- **speedup was calculated in relation to the elapsed time of 9598sec obtained for a 96 core run on fully committed nodes**
- **usage of partial nodes with 12 cores-per-node significantly improved the model performance scaling, usage of half nodes gave over 10% of additional performance gain**
- **the shortest run-times were achieved on partially used nodes (12 cores-per-node) for over 576 cores reserved**

UKV performance comparison on 3 systems



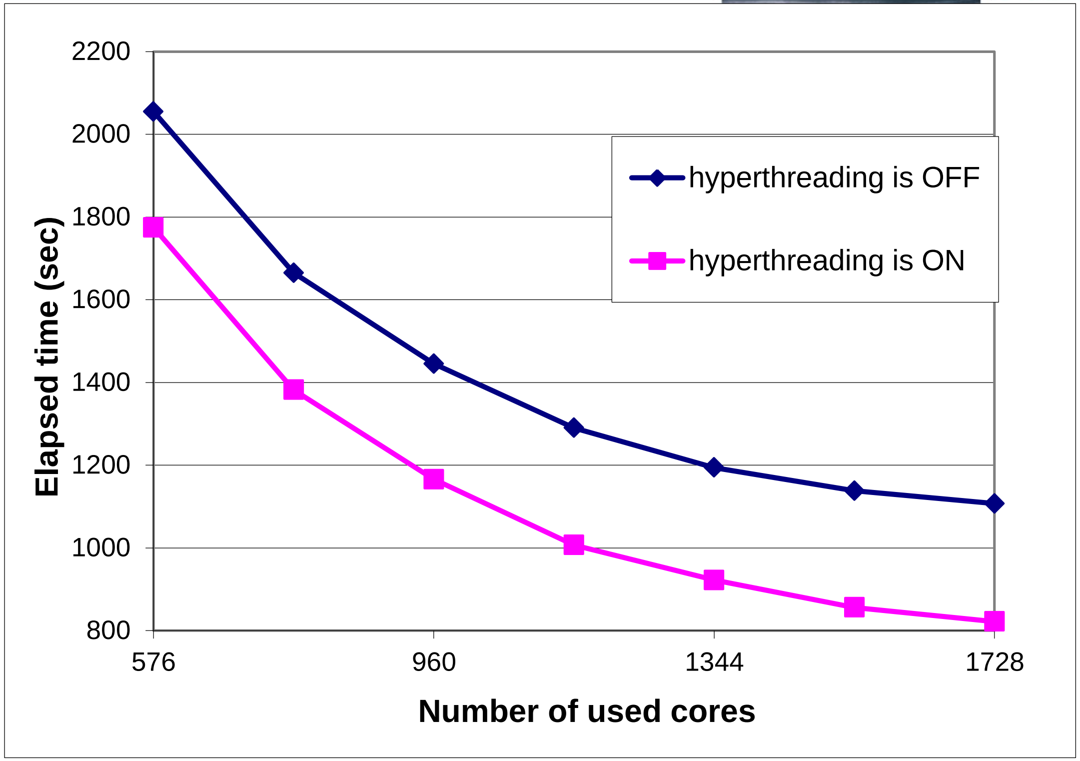
- **over 20% performance improvement on the latest systems**
- **Raijin vs Ngamai: being slower by 10% at 576 cores, the gap in the performance monotonically reduces to 0 at ~1000 cores and increases to 20% at 3456 cores due to a better performance on partial nodes and faster internode connect**

UKV performance with hyperthreading on Raijin

- **MPI executable**
- **fully committed node case**
- **additional PBS option**

  `-l other=hyperthread`

  for activating hyperthreads on the compute nodes

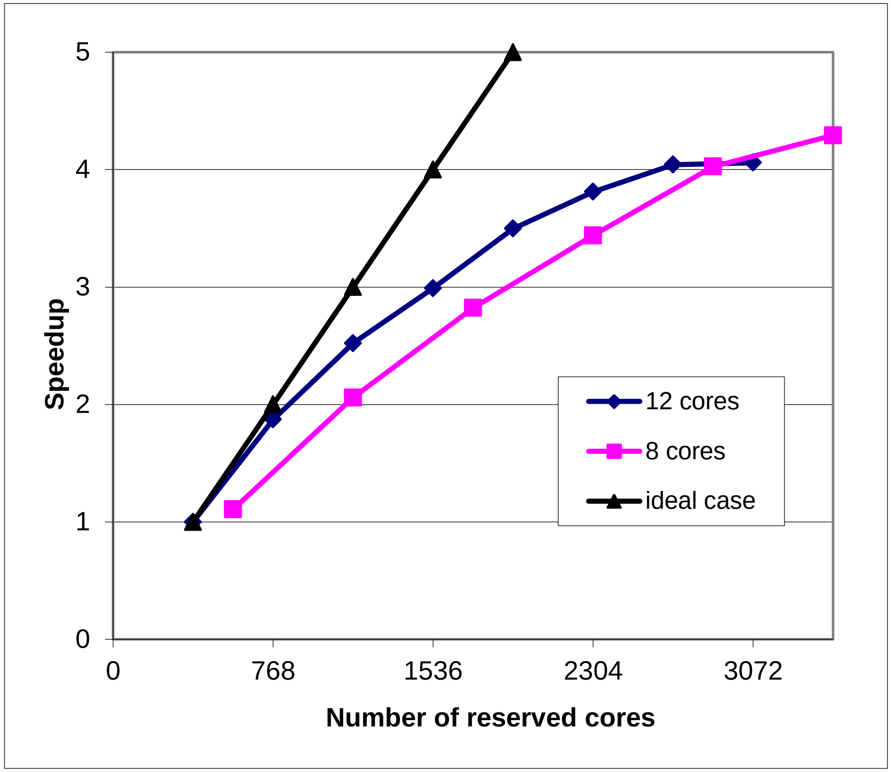


- **hyperthreading improves elapsed times by 13.6% - 25.7%, the improvement is monotonically increasing with an increase in the number of the used cores**
- **hyperthreading on partial nodes with 12 (out of 16) gives only a very modest improvement in the elapsed times of between 1% - 2.7%**

Australian Government
Bureau of Meteorology

CSIRO

N512L70 scalability on Ngamai



fully committed nodes
performance scaling with >1920 cores slow degrades and stops improving with the usage of 3072 cores

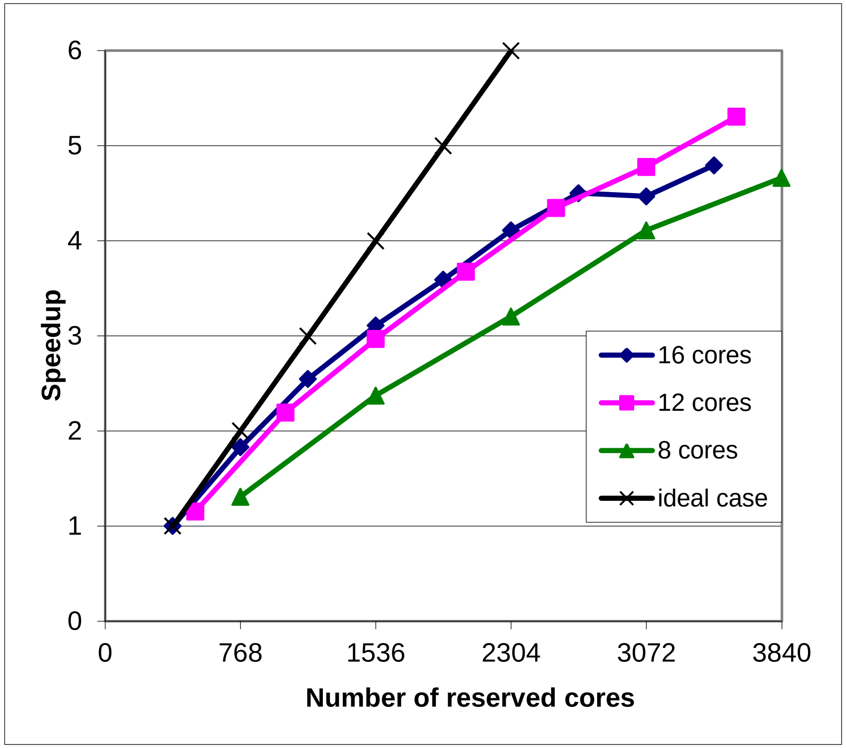
partial nodes (8 cores-per-node)
there is a relatively good performance scaling in the range of reserved cores up to 3456

from the efficiency point of view runs with 3072 reserved cores and higher should use partial nodes

speedup was calculated in relation to the elapsed time of 3068sec obtained for a 384 core run on fully committed nodes

6 cores-per-node: no improvement in the performance results in comparison with the 8 cores-per-node case (non-symmetry issue)

N512L70 scalability on Raijin

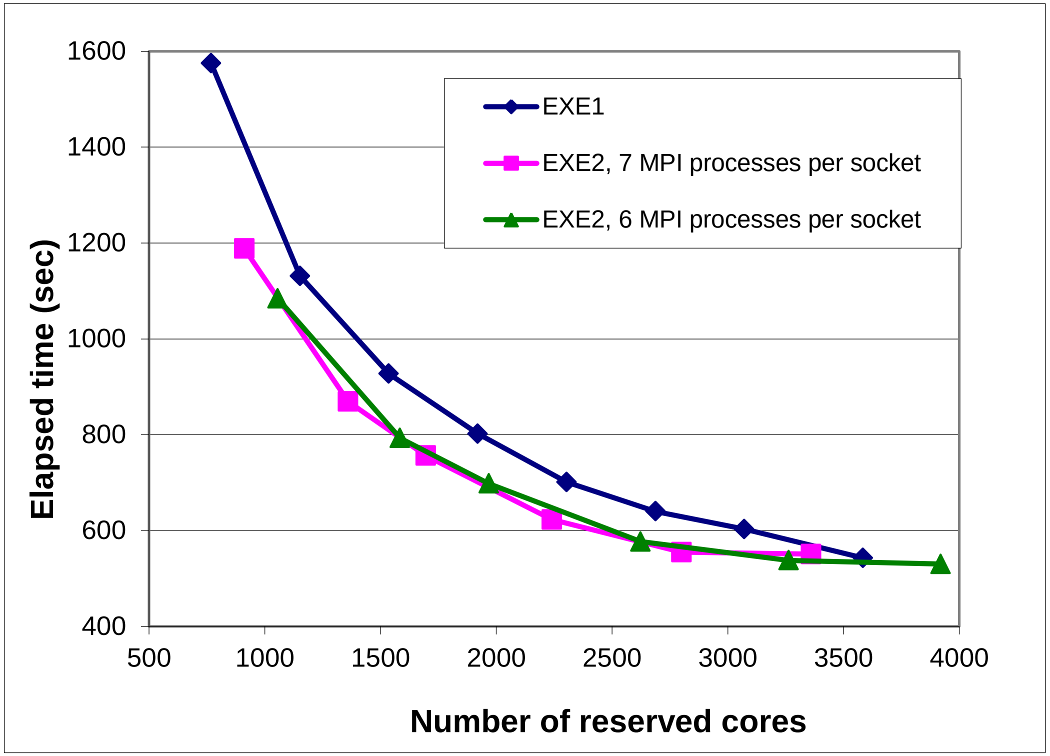


The most efficient system usage with 3072 core or higher is achieved on partially committed nodes with 12 cores-per-node

In the range of reserved cores between 384 and 2688 performance results on fully committed nodes and partial nodes with 12 cores-per-node are very close

speedup was calculated in relation to the elapsed time of 2881sec obtained for a 384 core run on fully committed nodes

N512L70 performance comparison for 2 executables on Raijin



EXE1 – an executable built with multithreading (the best performance results from the previous slide)

EXE2 – only IO server library was built with multithreading, the computational part of the sources is built without multithreading

- **a special approach in building UM executable allows some reduction of up to 10-12% from the run times obtained using a "standard" building procedure which diminishes with the usage of over 3000 cores**
- **from the efficiency point of view there is no preference between 2 curves obtained with the usage of EXE2**

Australian Government
Bureau of Meteorology

CSIRO

Arithmetic performance and memory bandwidth trends

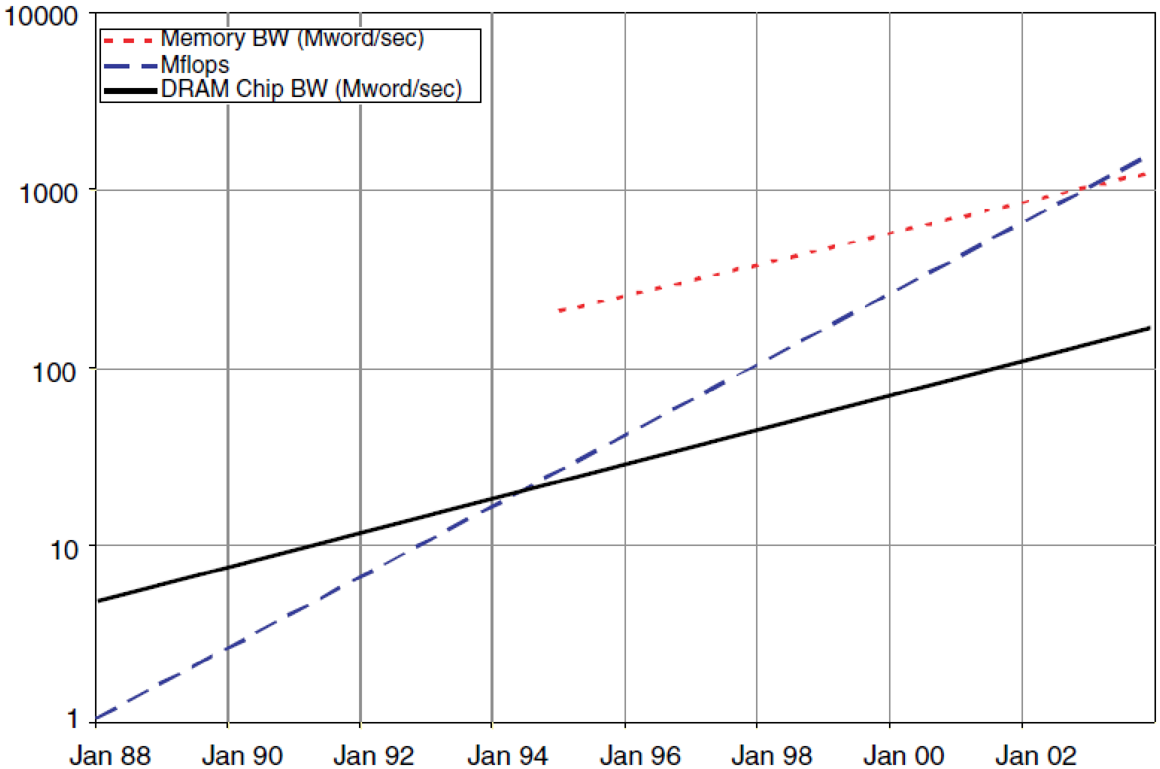


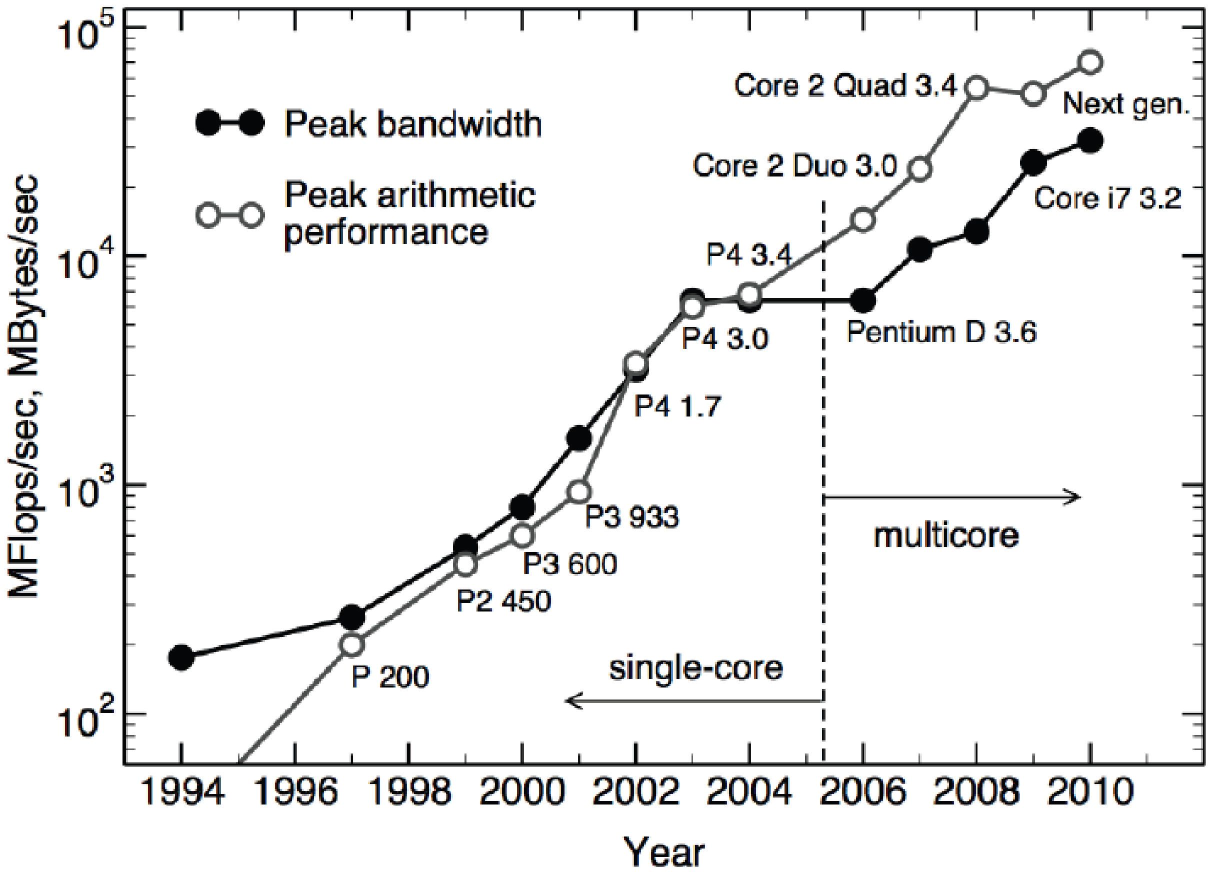
FIGURE 5.3 Arithmetic performance (Mflops), memory bandwidth, and DRAM chip bandwidth per calendar year.

"the gap between processor and memory performance continues to grow"

The authors concluded that the memory bandwidth speed on the current HPC may effect performance for memory bound applications

Graham, S.L., Snir, M., and Patterson, C.A., 2005: Getting Up To Speed: The Future Of Supercomputing. (http://research.microsoft.com/en-us/um/people/blampson/72-cstb-supercomputing/72-cstb-supercomputing.pdf)

Arithmetic performance and memory bandwidth trends (cont #2)



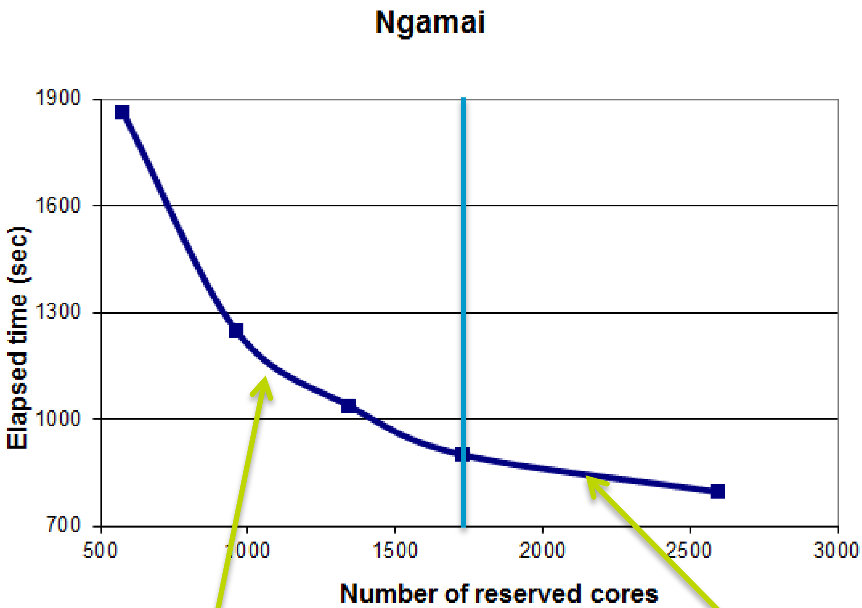
the gap is growing especially fast for multicore processors

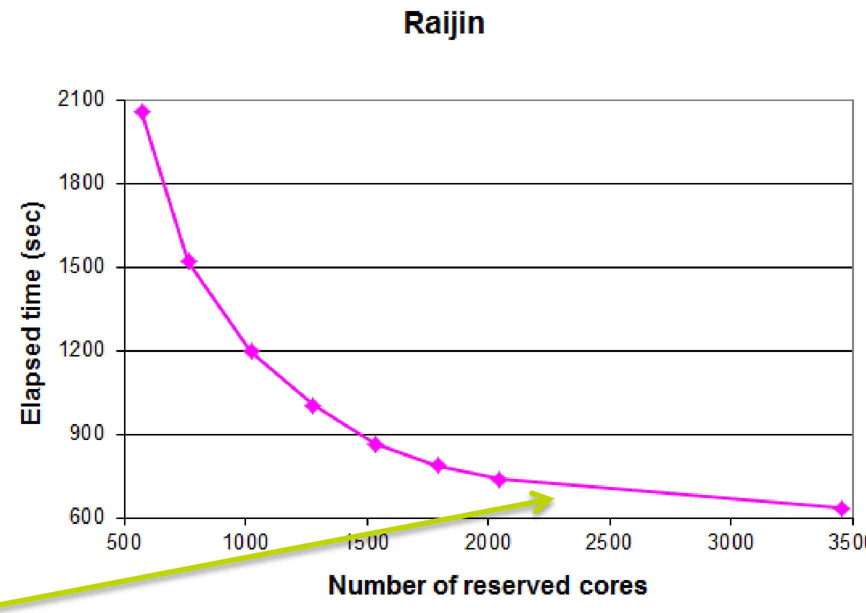
Wellein, G., Hager, G., Kreutzer, M.: Programming Techniques for Supercomputers: Modern Processors, 2012 (https://wiki.engr.illinois.edu/download/attachments/217842128/performance-bandwidth.pdf?version=1&modificationDate=1359049396000)

Australian Government
Bureau of Meteorology

CSIRO

Benefits in partial node usage (case #1): limit on scaling due to application constrains

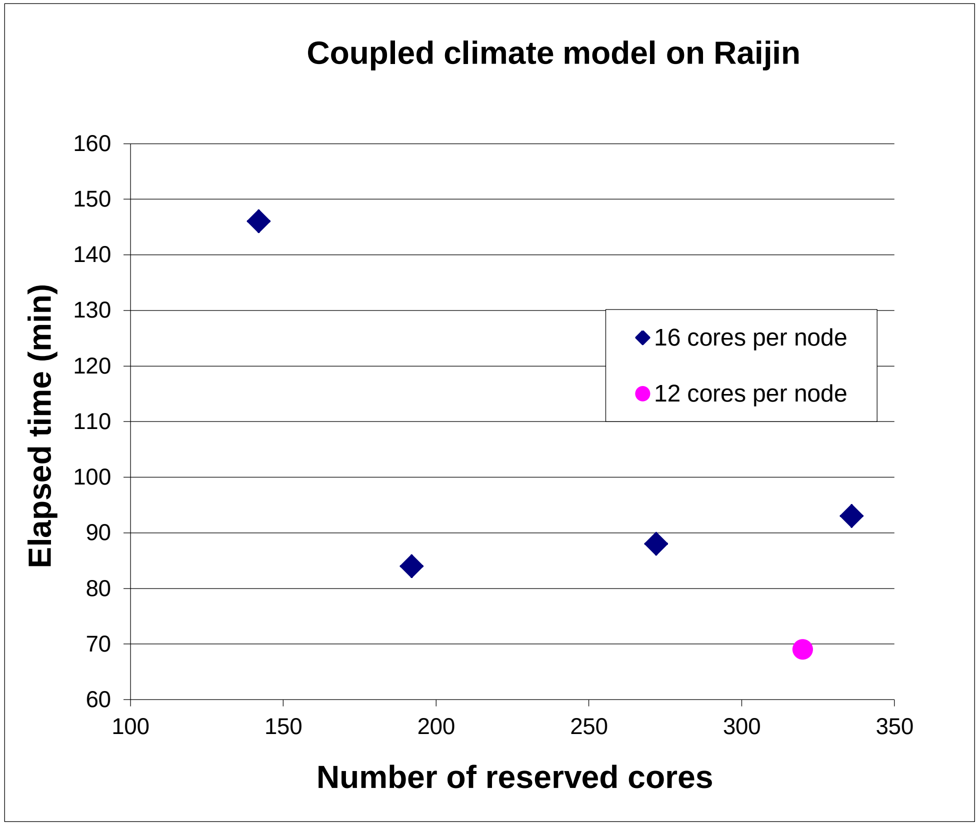




fully committed nodes **partially committed nodes**

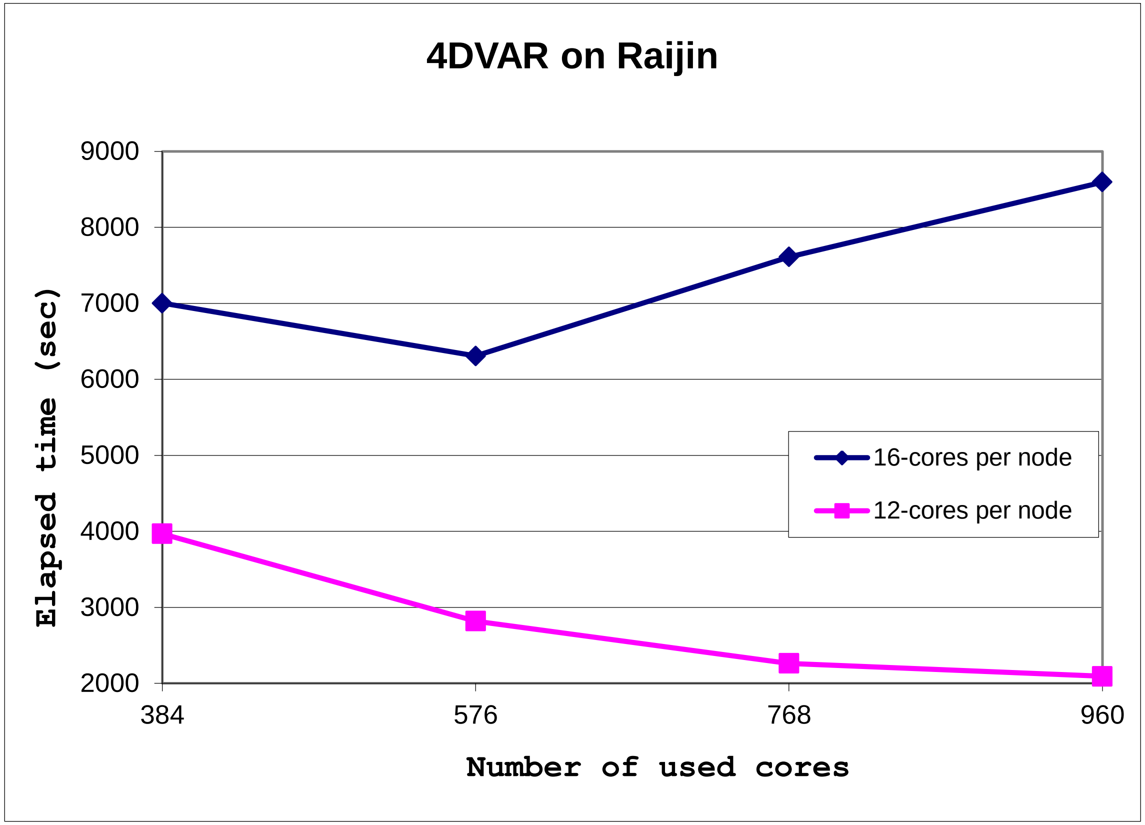
- **due to the UKV model constraints 1728 core case represents almost the maximum decomposition size which can be used for the application to achieve the shortest elapsed time**
- **Ngamai: usage of 8 cores (from 12 available) on each node allows to reduce this time by 11.5% on 2592 cores, our expectations are that this time could be improved further by 5% -10% if turbo boost would be available on the system**
- **Raijin: still using partial nodes a 20% improvement in the elapsed time is achieved using 3456 cores**

Benefits (case #2): performance flattened



Climate coupled model with the atmosphere modelled by a low resolution of N96L38

Benefits (case #3): quick improvement



The Met Office four-dimensional variational analysis system:

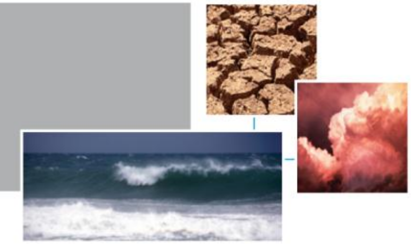
significant performance improvement for an anti-scaling performance problem with the usage of fully committed nodes

Practically this approach has been successfully used in reduction of run-times of the operational APS1 ACCESS-R system to meet the operational deadlines

Australian Government
Bureau of Meteorology

CSIRO

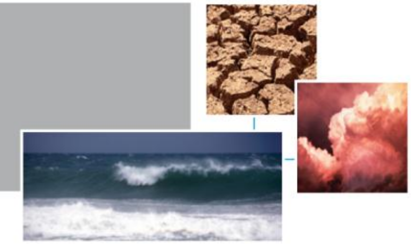
Conclusions



- **Examples of the UM applications running on Intel Xeon Sandy Bridge clusters show that in some cases for memory bandwidth intense applications the most efficient usage of the modern HPC can be achieved on partially used nodes**
  - **climate models usually run at a relatively low resolution could be sped up using this approach**

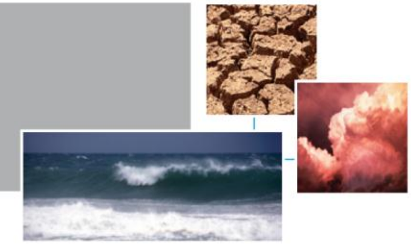
- **Trend in the HPC development shows that**
  - **Byte/Flop ratio is consistently decreasing especially on multicore processors => memory bandwidth for some applications is becoming a major bottleneck**
  - **systems have much more relatively cheap cores**
  - **usage of partial nodes and availability of turbo boost has a couple of advantages**
    - **cores run at a higher peak speed**
    - **memory bandwidth per core improves**

Conclusions (cont #2)



- Models usually do not scale up to the available number of cores or constrains in the model implementation do not allow to use more cores
  - MPI communication eventually kills scaling (only a very few threads can be used with UM, the Met Office 4DVAR runs with pure MPI only)
- All above mentioned aspects show that on the modern HPC systems usage of partial nodes can be an option for efficient usage of the systems as well as in getting the shortest run times for memory bandwidth intense applications
- Aspects of efficiency and speed in running applications on HPC
  - efficiency using limited available resources and meeting time deadline requirements for production runs of operational applications
  - speed when the human factor is involved

Thank you



Dr Ilia Bermous

Senior Information Technology Officer

email: i.bermous@bom.gov.au

Thanks to support personal at ANU/NCI and Oracle and the UM developers at the UK Met Office

Australian Government

Bureau of Meteorology

CSIRO