Australian Government
Bureau of Meteorology

BUREAU OF METEOROLOGY

Sustain Teraflops on a Production System
Challenges in Resilience and Consistency

Tim F. Pugh[1] and Joerg Henrichs[2]

Fourteenth Workshop on the Use of High Performance Computing in Meteorology

1 - 5 November 2010

[1] Australian Bureau of Meteorology

[2] Oracle Australia, Inc

The High Performance Computing and Collaboration Centre
A partnership between CSIRO and the Bureau of Meteorology
www.hpccc.gov.au

CSIRO

Outline of Topics

- New Programs in the Bureau
  - Water Act 2007 → Water Division in Bureau of Meteorology
  - National Plan for Environmental Information (NPEI)
  - Strategic Radar Enhancement Project (SREP)

- Notable changes in the last two years
  - Sun Constellation system declared ready for operational use
  - Unified Model and 4DVAR declared operational on Sun Constellation

- Production System
  - Production Challenges
  - Future Application Computing Estimates
  - Challenges for Consistency and Resilience

Water Resource Management

- Water Act 2007
  - $12.9 billion investment in a 10-year plan to secure long-term water supply
  - $450 million investment in Bureau of Meteorology for responsibility in compiling and delivering comprehensive water information
  - The new Water Division has four major areas of activity:
    - Water Data Management
    - Water Data Analysis and Reporting
    - Water Forecasting
    - Water IT Planning and Development
  - For more information, go to http://www.bom.gov.au/water

- 2008 Modernisation and Extension Funding program
  - $80 million program to help water data collection agencies upgrade and expand their streamflow, groundwater monitoring and water storage measurement networks

Australian Government
Bureau of Meteorology

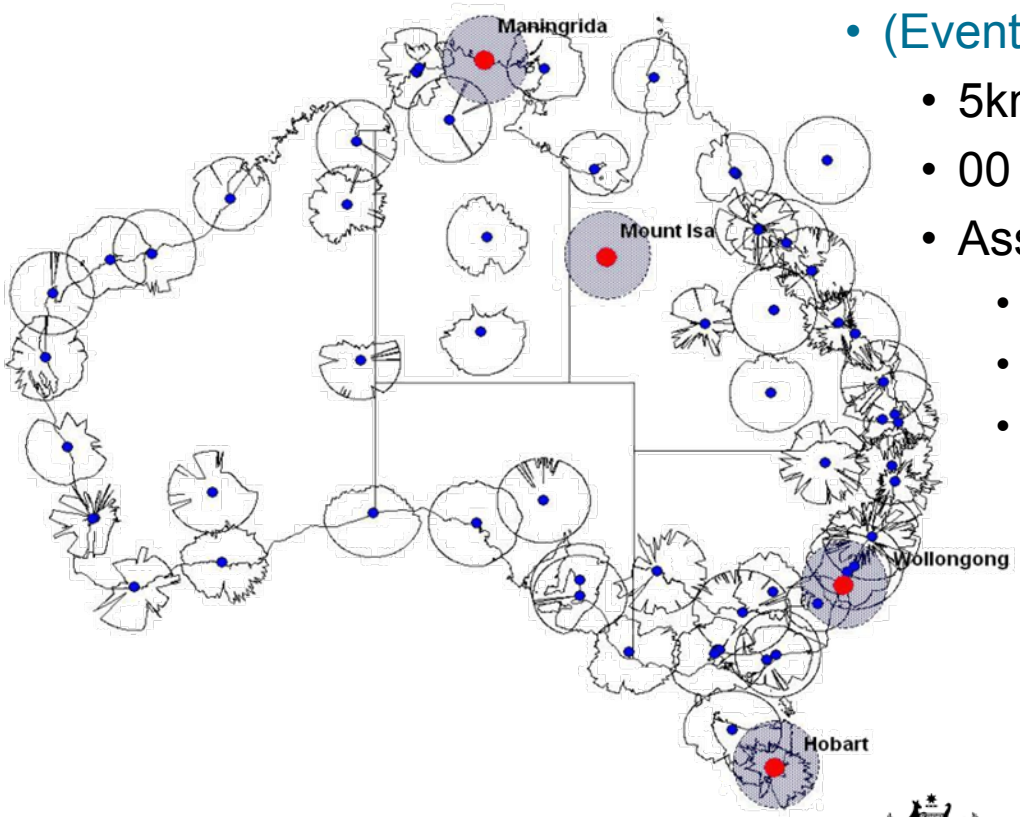
(NPEI) National Plan for Environmental Information

- 2010 Whole-of-Government Initiative for Environmental Information
  - The bureau will:
    - become the Australian Government authority for environmental information, building on their existing role for weather, climate and water information
  - In the first four years, the initiative will:
    - establish the Bureau of Meteorology as the Australian Government authority for environmental information
    - formalise arrangements to coordinate priorities and activities across government
    - review existing information resources, and environmental information activity
    - begin building priority national environmental datasets and the infrastructure to deliver them.

For additional information, go to http://environment.gov.au/npei

Australian Government
Bureau of Meteorology

SREP
Strategic Radar Enhancement Project



- CAWCR: *Improve the underlying science to assimilate radar data into the Bureau's NWP models*
- (Eventual) Upgrade for city based systems
  - 5km → ~2km
  - 00 & 12Z → 00, 03, 06, 09, 12, ….
  - Assimilation
    - In situ obs & satellite
    - Doppler winds
    - Precipitation

System and NWP Updates

March 2009 Contract signed with Sun Microsystems

March 2009 Exemplar system for software porting staff in San Diego

June 2009 Phase One: Initial system delivery

Sept. 2009 Data Centre upgrade completed

Oct. 2009 Phase Two: Full system build in progress

Nov. 2009 Phase Two: System access to software porting staff

….. resolution of a number of HW/SW issues …..

March 2010 Oracle takes over Sun Microsystems

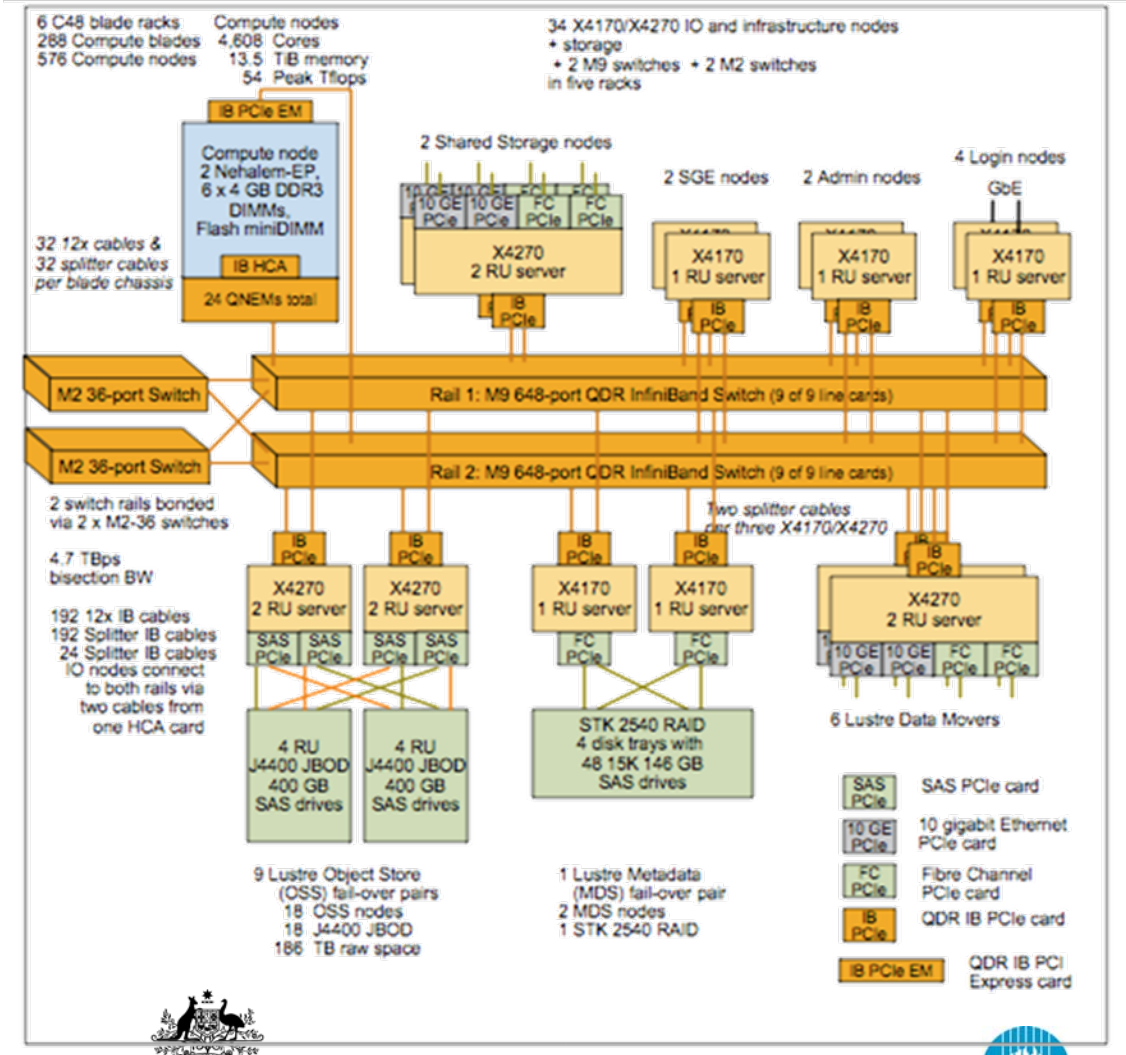
30 May 2010 Phase Two: Full system ready for production use

22 Jun 2010 Oracle and BoM declare system ready for operational use

20 Aug 2010 NEC SX-6 decommissioned

Australian Government
Bureau of Meteorology

Bureau's HPC system "Solar"
Sun Constellation System



Australian Government
Bureau of Meteorology

Operational requirements

- Solar is mission-critical system for Bureau
  - Full NWP suite run 4 times per day, 365 days per year
  - Natural disaster season (cyclones, bushfires) November – February inclusive ➔ system must be available to run simulation jobs on demand during this period
  - Failures have national significance
- Oracle responsible for system availability, stringent SLA
  - Max 3.6 hours system downtime on 30 day rolling schedule (file system, job scheduler, network)
  - 99.5% uptime for infrastructure nodes
  - 75% uptime for compute nodes

Current status

- Operating within SLA since acceptance
- Average 120,000 jobs per day, dominated by data mover jobs (94%)
- Roughly equal numbers of operational and research compute jobs, but research CPU is 80% versus operational 20%

Australian Government
Bureau of Meteorology

ANU/NCI HPC System "vayu"
Sun Constellation System

System Racks
- 16 x C48 blade racks
- 6 x 19" racks
- 34 sq m floor space
- 750kW power (HPL)

1,492 Compute nodes
- 11,936 Cores
- 140 Peak Tflops
- 37 TiB memory

13 Lustre Object Store pairs
- 26 x OSS nodes
- 52 x J4400 JBOD
- 834 TB usable storage
- 25 GBps BW

CSIRO share is ~24% of vayu

- IPCC AR5 climate runs
- Ocean reanalysis using 1/10 deg global ocean model

Australian Government
Bureau of Meteorology

2012 Petascale HPC Facility

- April 2010, the Australian government announced $50m contract to ANU/NCI to build a new Petascale HPC facility for climate change, earth system science and national water management research
    - To procure a new data centre
    - To procure a petascale computing system
    - Partners to provide operating budget of facility
    - Target date for system delivery is 2012

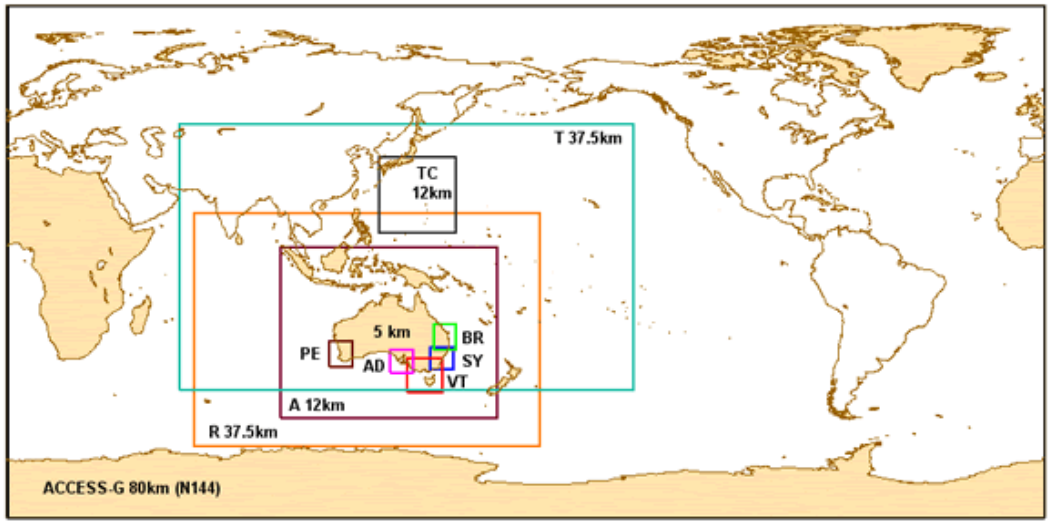
- http://nci.org.au/news-and-events/news/funding-agreement-signed-a-a-new-petascale-high-performance-com/
- National Computational Infrastructure (NCI) is located at the Australian National University (ANU), Canberra, Australia

Australian Government
Bureau of Meteorology

CSIRO

Australian Government
Bureau of Meteorology

ACCESS-NWP (APS0 timelines)



"APS0": ACCESS Parallel Suite 0

Met Office Unified Model vn6.4

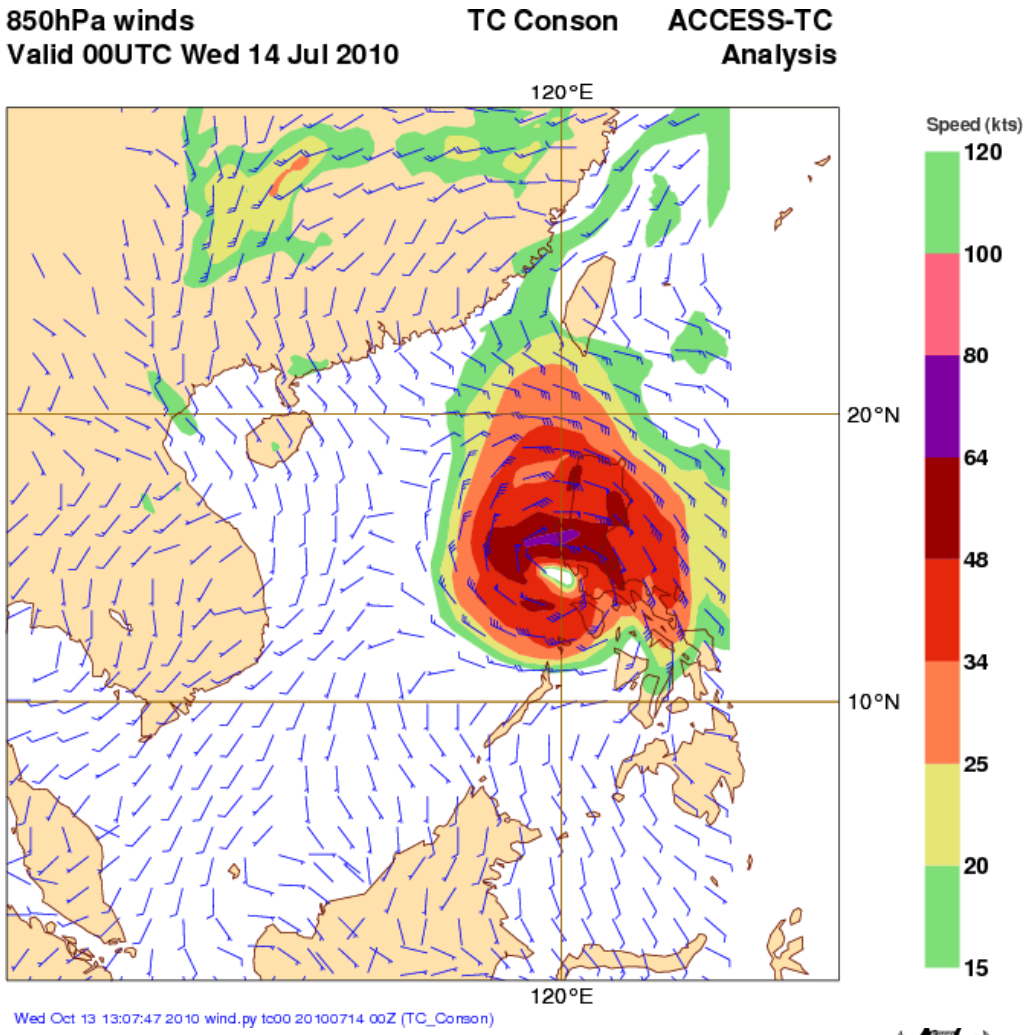
Replicates legacy domains

- G – Global (80km L50)
- R – Regional (37.5km L50)
- T – Tropical (37.5km L50)
- A – Australian Meso (12km L50)
- C – City (5kmL50)
- TC – Tropical Cyclone (12km L50)

- September 2009 G, R, T went operational on NEC-SX6
- 29 June 2010 G, R, T, A went operational on Sun Constellation
- 12 August 2010 C went operational on Solar
- 17 August 2010 Last forecasts from legacy NWP - GASP, LAPS, TXLAPS, MesoLAPS,…
- Today TC is being ready for operations

Australian Government
Bureau of Meteorology

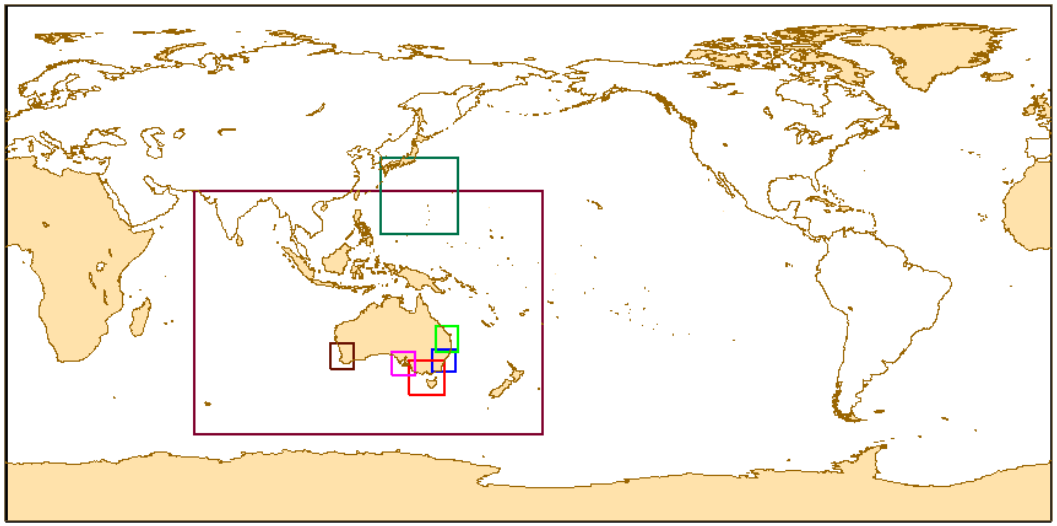
ACCESS-NWP (ACCESS-TC)



- ACCESS-TC
  - Assim/Forecast components in place, number of promising studies for NH and SH TCs
  - Post-processing/diagnostic tools being finalised
  - Addition of TC-bogus to other ACCESS systems
- Should be ready for Australian TC season

Australian Government
Bureau of Meteorology

ACCESS Parallel Suites



"APS1": ACCESS Parallel Suite 1

Met Office Unified Model vn7.5
Met Office 4D-VAR v26.1

- G – Global (40km L70)
- R – Regional (12km L70)
- C – City (4km L70)
- TC – Tropical Cyclone (12km L70)

NWP System	Domain	Type	Q1 2011	Q1 2012	Q1 2013
			APS1	**APS2**	**APS3**
ACCESS-G	Global	10 day FCST	**N320L70** (40 km) 640x481x70	**N512L70** (25 km) 1024x769x70	**20kmL90** (20 km) 1280x961x90
ACCESS-R	Australian Region	3 day FCST	**12kmL70** 1090x750x70	**12kmL70** 1090x750x70	**10kmL90** 1200x825x90
ACCESS-C	Cities	2 day FCST	**4kmL70** 200x200 to 300x300	**2kmL70** 400x400 to 600x600	**1.5kmL70** 533x533 to 800x800
ACCESS-TC	TC & Severe Wx	3 day FCST	**12kmL70** 300x300x70	**12kmL70** 300x300x70	**10kmL90** 330x330x90

Projected future APS on Sun Constellation for estimating computing requirements

Australian Government
Bureau of Meteorology

ACCESS Coupled Model for IPCC AR5

- Atmosphere: Unified Model Grid N96L38 (1.875 x 1.25)
- Ocean: MOM4p1 Tripolar Grid (1.0 x 1.0 x 46 levels)
- Sea Ice: CICE 4 Similar to Ocean Grid
- Coupler: OASIS 3.2.5
- Land surface/carbon cycle: "CABLE" (Kowalczyk et al. 2006) + CASA-CNP + LPJ dynamic vegetation

- Early version of HadGEM3, based on UM vn7.3
  - 144 cores = 96p Atmos.+ 40p Ocean + 6p CICE + 1p OASIS.
  - 3 hourly coupling
  - Typical run time is about 3.5 simulated years/day
  - Running at the ANU/NCI Sun Constellation system

Australian Government
Bureau of Meteorology

Operational Schedules

- Workflow Dependencies
  - Observations → Global → Regional → City Prediction Systems
  - Observation Products → Model Products → Forecasters → Public

- Drivers to Grow Capability Computing
  - To improve skill scores, need to improve grid resolution and physics
  - To complete a forecast cycle within a time window

- Why is runtime consistency important and what is acceptable?
  - Runtime consistency is needed to meet time windows and reduce risk.
    - Runtime inconsistency shows system arch./mgmt/application issues
  - 4 hours to … (Your Mileage May Vary)
    - assim obs into state of the atmosphere, make a prediction, output products
    - Then draw some maps, write some words i.e.“fine”, delivery the news

Australian Government
Bureau of Meteorology

APS0 Operational Schedule

Time (EST)	Time (UTC)	ACCESS-G	ACCESS-O3	ACCESS-R	ACCESS-A	VICTAS05	SYDNEY05
9:00 AM	23:00						
	23:15						
	23:30	AG18Z					
	23:45	256var 240um					
10:00 AM	0:00		AO3 18Z	AR18Z update			
	0:15		16var 240um	160var 80um			
	0:30				AA18Z update		
	0:45				(assim only)		
11:00 AM	1:00				240var 648um		
	1:15						
	1:30						
	1:45						
12:00 PM	2:00			AR00Z main			
	2:15			160var 80um			
	2:30			WW-3 96 cores	AA00Z main	VT00Z 250c	SY00Z 100c
	2:45				(assim+forc)		
1:00 PM	3:00				240var 648um		
	3:15						
	3:30						
	3:45				WW-3 96 cores		

Australian Government
Bureau of Meteorology

Global Forecast Cycle Estimated Compute Resources

NWP System	Domain	Type	Q3 2009	Q4 2010	Q2 2011	Q4 2011	Q1 2013
			APS0	APS1	APS2	APS2	APS3
			Intel Nehalem	Intel Nehalem	Intel Nehalem	Intel Sandy Bridge	Intel Sandy Bridge
ACCESS-G[1]	Global	10 day FCST 2 FCST / day	N144L50 (measured)	N320L70 (measured vn7.5)	N512L70 (measured vn7.6+)	N512L70 (estimated)	20kmL90 (estimated)
		grid pts	3,124,800	21,548,800	55,121,920	55,121,920	110,707,200
		timestep (min)	15.00	15.00	10.00	10.00	8.00
		cores	240	640	1280	782	1964
		% System (cores)	5.21%	13.89%	27.78%	8.49%	21.31%
		elapse time (min)	35	81	85	85	85
		Gflops (sustained)	132	352	704	704	1767
	Global	VAR 4 analysis/day	N108L50 (measured)	N144L70 (measured)	N216L70 (measured v26.1)	N216L70 (estimated)	50kmL90 (estimated)
		grid pts	1,760,400	4,374,720	9,828,000	9,828,000	17,740,800
		timestep (min)	20.00	15.00	15.00	15.00	12.66
		cores	256	504	756	462	988
		% System (cores)	5.56%	10.94%	16.41%	5.01%	10.72%
		elapse time (min)	13	13	32	32	32
		Gflops (sustained)	141	277	416	416	890

Australian Government
Bureau of Meteorology

Projected APS2 Operational Schedule

Time (EST)	Time (UTC)	ACCESS-G	~~ACCESS-O3~~	ACCESS-R	~~ACCESS-A~~	VICTAS05	SYDNEY05
9:00 AM	23:00						
	23:15						
	23:30	AG18Z					
	23:45	756var 1280um					
10:00 AM	0:00						
	0:15						
	0:30						
	0:45						
11:00 AM	1:00			AR18Z update			
	1:15			(assim only)			
	1:30			579var 1200um			
	1:45						
12:00 PM	2:00			AR00Z main			
	2:15			(assim+forc)			
	2:30			579var 1200um			
	2:45						
1:00 PM	3:00						
	3:15						
	3:30						
	3:45			WW-3 96 cores		VT00Z 1500c	SY00Z 600c

Consistency and Resilience

- System growth in processors, nodes, switches, and storage systems
  - Vector systems → Scalar systems → multicore systems →many-core
- Complexity in application interaction with system increasing…

- Resulting in …
  - More failures, more often due to increase exposure to equipment reliability
    - So fault tolerance and resilience are more significant
  - More issues due to system configuration inconsistency
    - Better management from firmware to O/S to services
  - More application elapse time inconsistency due to
    - Data transmission errors, symbol errors, retransmits
    - Operating system process scheduling and performance
    - Competing resources and priorities
    - Overcommitment of resources

Operating Systems
Unreliable runtime consistency

- Operating system tasks and performance
  - Operating system process scheduling
    - OS synchronization to avoid O/S jitter (Still have to wait)
  - Competing processes and priorities
    - Scheduling priority mistakes can cause starvation / lack of responsiveness
      - Application runtime inconsistency due to process priority mistake
        - Operational user job had equal or higher priority to Lustre process
  - Over commitment of resources and architecture issues
    - Processing loads out pace available resources (too much waiting)
    - NUMA latency, memory allocation and placement
  - Monitoring (e.g. ganglia, nagios)
    - Monitoring too frequently
      - tuned monitoring with N512L70 model

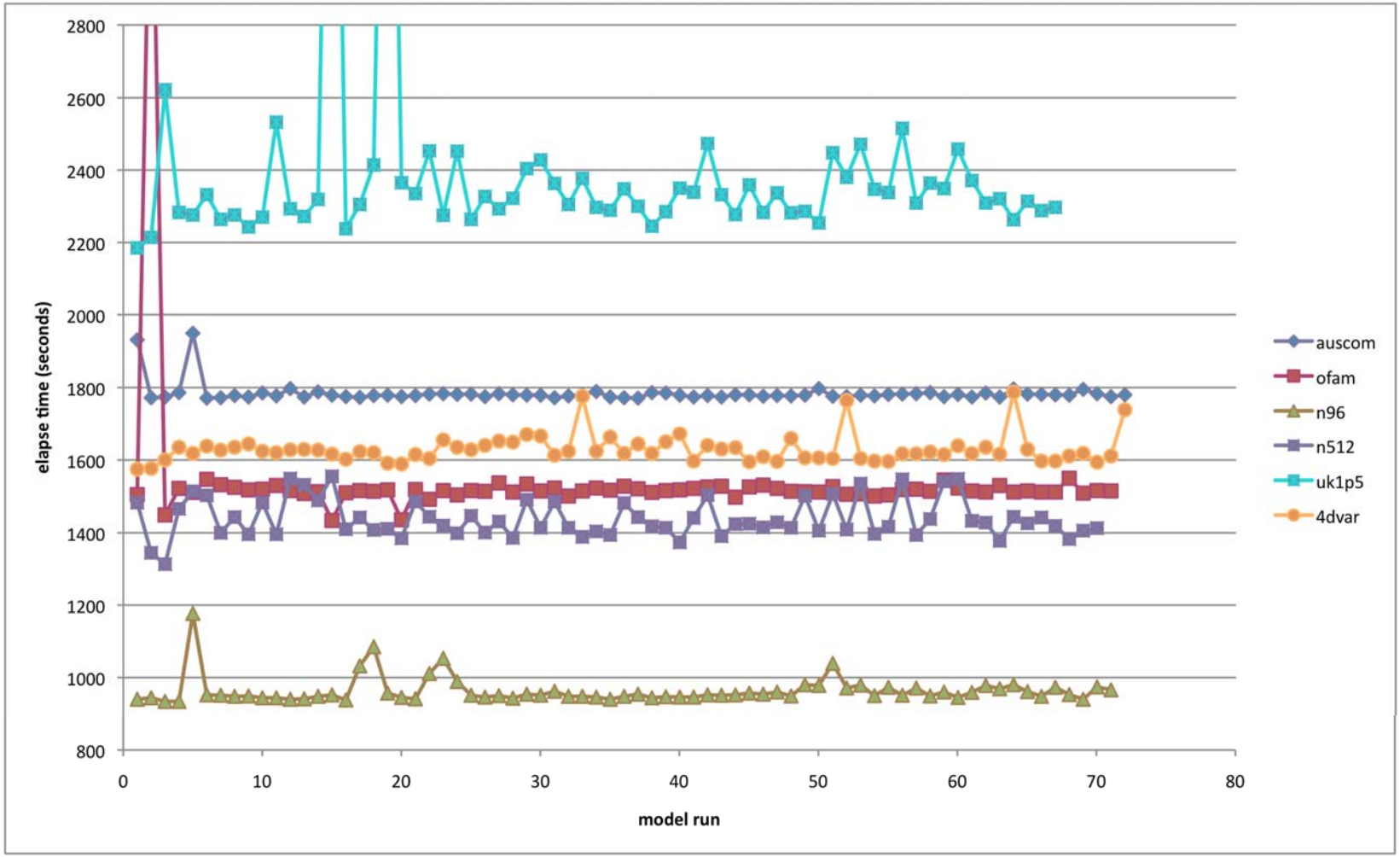
Node architecture does not offload operating system tasks
- Or is that Node Architecture does not offload computing tasks
- Dedicated processor(s) for Operating System tasks is desirable

Operating System Performance

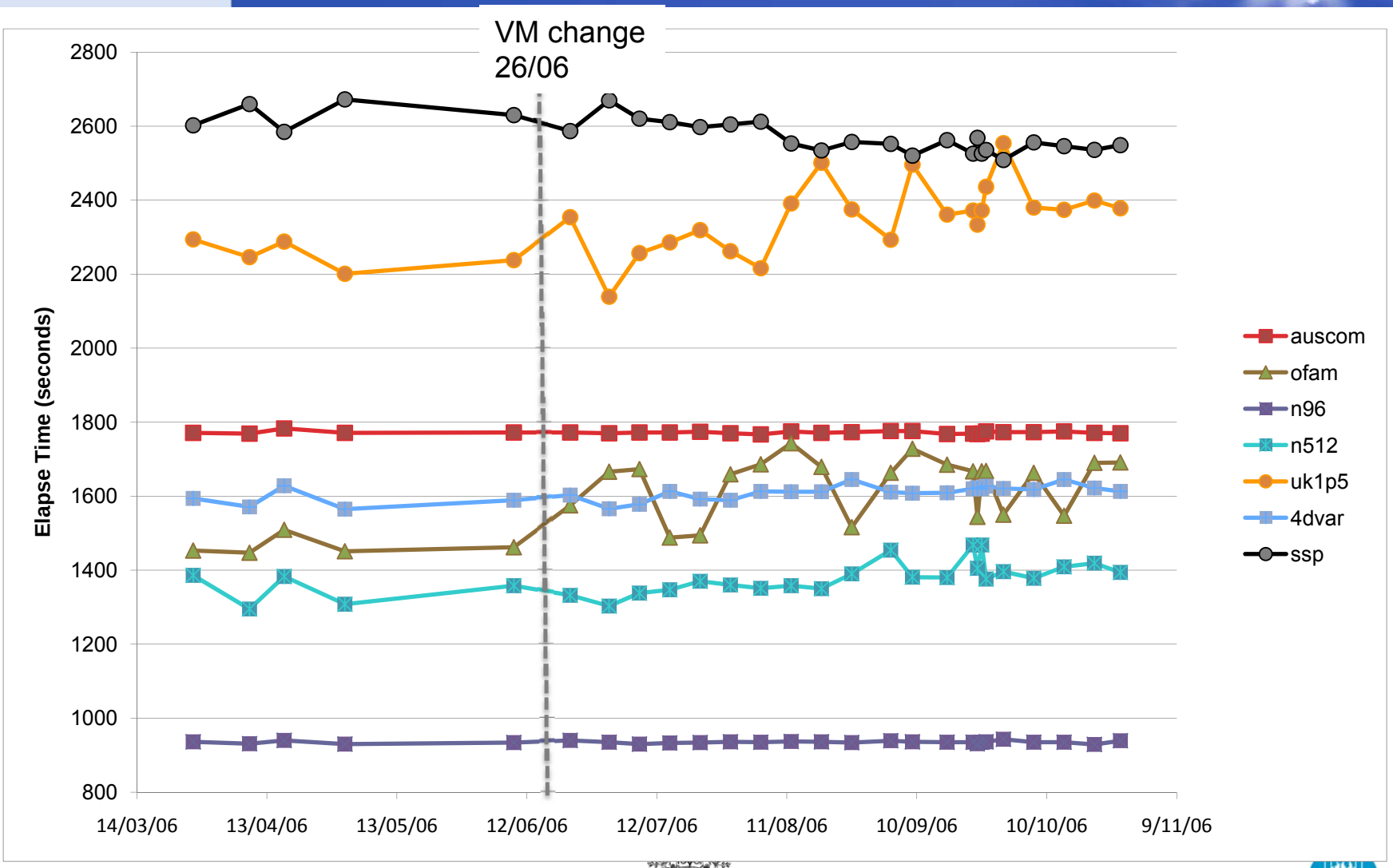
- Virtual memory performance for Linux kernel
  - Linux kernel 2.6.18-128.1.14.el5 for CentOS 5.3
  - Redhat Bugzilla ID is 457264
    - The workload was a CPU intensive application that periodically checkpoints to disk 1.5-2GB data. The 100% SYS was seen unless zone_reclaim_mode=0.
    - We were able to replicate the stall issue (every 3-5 attempts) while copying a 2GB file
    - Solution, turn off optimized zone reclaim code in kernel, zone_reclaim_mode=0
    - Still broken in Linux kernel 2.6.32 (ANU/NCI)
    - Workaround, 'cache dropping' script installed to help offset the impact of the zonal-reclaim-kernel bug
- Kernel driver issues
  - Datamover fault on 10Gig Ethernet, known RedHat problem
    - Workaround disable Intel hyperthreading on datamover nodes

Australian Government
Bureau of Meteorology

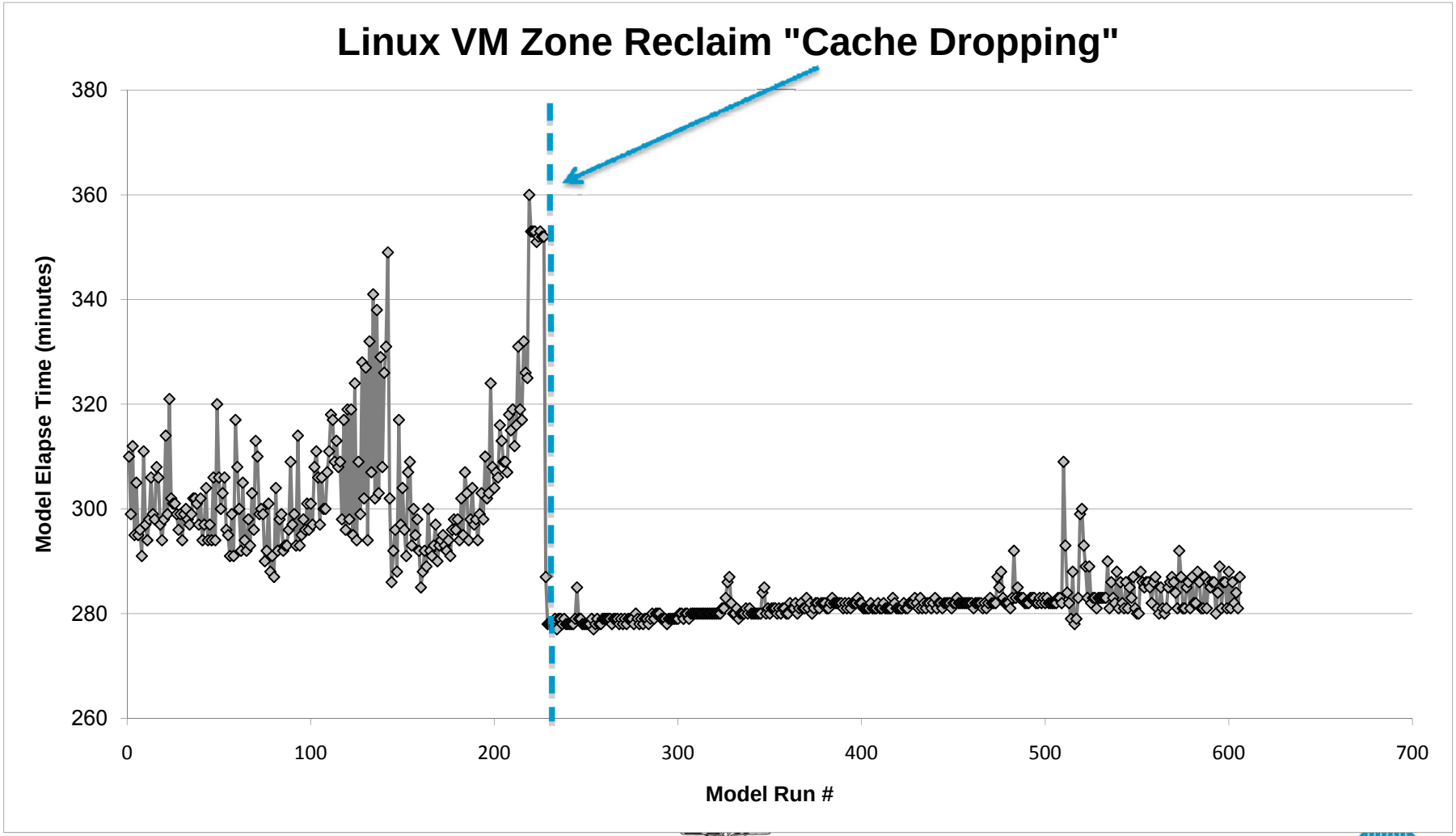
Optimized Linux VM Page Reclaim Mode



Measure of SSP



Workaround, “VM cache dropping”



Application Dependencies

- Unreliable runtime consistency
  - Application task and threads
    - Overcommitment of tasks per node
      - Threaded libraries (e.g math library)
    - MPI communications
      - Network retransmits and latency
      - Cable issues
      - PCIe board issues
    - System locks for system resources (e.g. file system)
      - Slow release of locks cause starvation issues if not careful
    - I/O performance
      - Local FS: SSD fragmentation issue over time (3-4 weeks, out of core solver)
      - Remote FS: Parallel file system concurrency and contention
    - Application programming issues
      - Characteristics of application programming on an architecture cause inconsistencies
        - MPI message sizes and frequency
        - Application memory placement on node

File Systems

- Unreliable runtime consistency
  - File systems
    - Concurrency → overcommitment, contention for same storage, meta-data, etc.
    - Complexity of high performance storage I/O (parallel file systems)
    - Meta-data server latency during high I/O transactions

- Lustre on the Bureau's system
  - We're in the process of moving from Lustre 1.8.1+patches to 1.8.4
  - Many bug fixes and improved stability.

- Lustre OST incorrectly reported file system full
  - Configuration change to grant leak
  - OSTs successively reset, reclaiming leaked grant
  - Fixed in Lustre 1.8.4

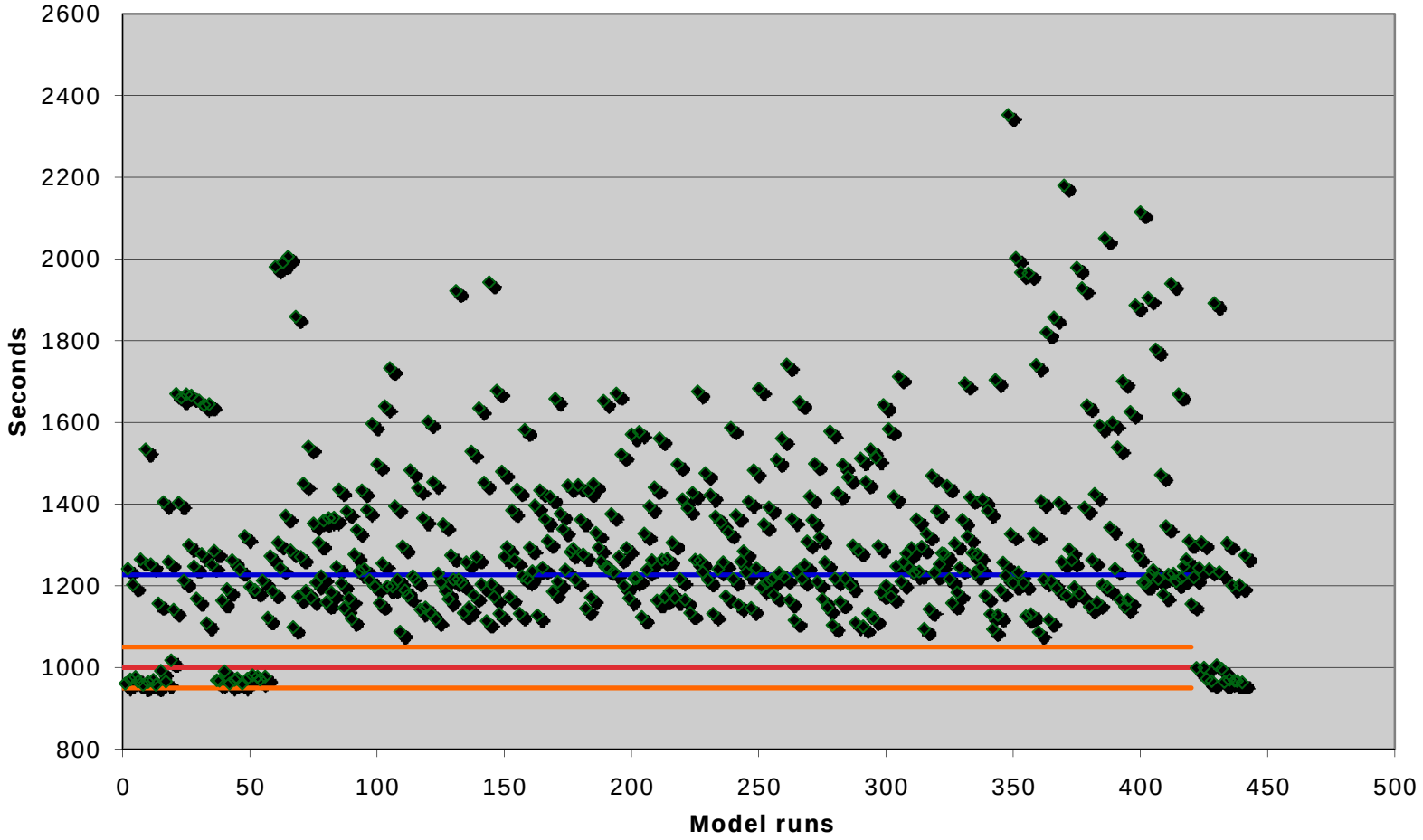
Network Interconnects

- Network interconnects
  - More data transmission errors, symbol errors, retransmits
    - Switches – monitoring of switches for error, retransmit statistics
    - Cables
      - Copper – poor signal integrity over distances
      - Fibre optic – fragile, handle with care

  - Network topology
    - Fat Tree or Torus characteristics
    - Fault tolerant and resilient, but elapse time is still affected
  - QDR Infiniband PCIe board
    - PCIe board network timing problem, firmware update fixed problem
    - BIOS upgrade fixing programming fault with PEM repeater

Australian Government
Bureau of Meteorology

N96L38 72-hour runtime variability test
single-rail IB with MPI over 4 nodes
poor runtime variability due to faulty IB cables



Future Challenges

- Resilience does not mean consistency
  - The application may continue to run, but it may not finish on time.
  - In the end, the schedule may kill the job for exceeding elapse time limits.

- With greater working parts, we'll experience more inconsistencies
  - Further understanding of interactions is important, ..and testing.
    - Will testing of future systems may mean longer acceptance periods?
  - Consistency problems are largely overlooked
    - Our weekly SSP tests show some results of inconsistency, but not to the degree of intense, repeat runs of a stress test.
    - We'll watch our operational model runs more closely to monitor system behavior.
  - Jitter comes in many forms, but we use one word to account for it.
    - Application runtime inconsistency is due to jitter, but what does that mean?

Australian Government
Bureau of Meteorology

The High Performance Computing and Collaboration Centre
A partnership between CSIRO and the Bureau of Meteorology

www.hpccc.gov.au

CSIRO

Thank you

Thanks to our partners in HPC:

ANU/NCI, CSIRO, Intel, Oracle, and the UK Met Office

Tim F. Pugh
Australian Bureau of Meteorology
Phone: +61 3 9669 4345
Email: tpugh@bom.gov.au