# Supercomputing Upgrade at the Australian Bureau of Meteorology

**Ilia Bermous**

**Bureau of Meteorology Research Centre**

**Melbourne, Australia**

Acknowledgements:

Bill Bourke, Mike Naughton, Peter Steinle

BMRC colleagues

Phil Tannenbaum & HPCCC colleagues

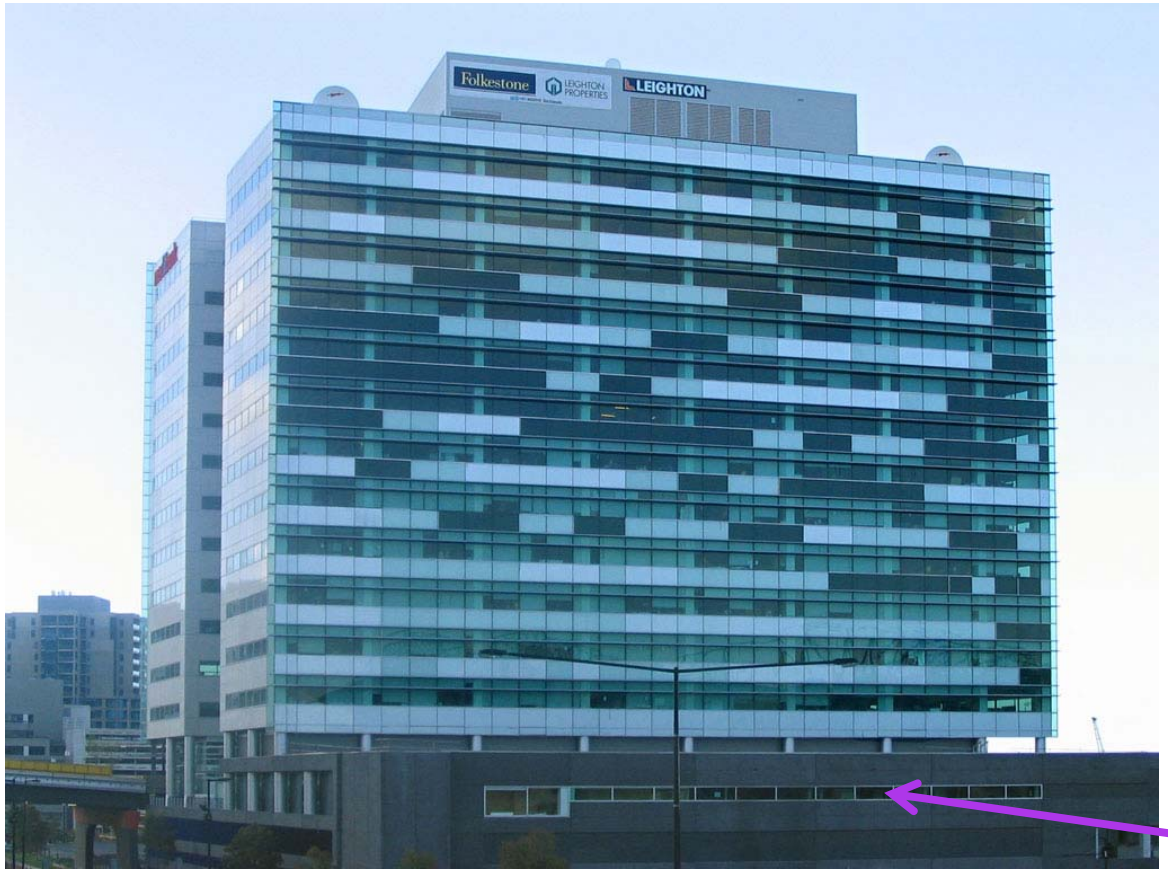
NEC/A applications support staff

*I.Bermous@bom.gov.au*

# Recent BoM/CSIRO HPCCC Systems History 1997-2007

| | | |
|---|---|---|
| NEC SX-4/16 | 1997 | |
| NEC SX-4/32 | 1998-2000 | 50:50 |
| NEC SX-4/32 & SX-5/16 | 2000-2001 | BoM/CSIRO |
| NEC SX-5/32M2 | 2001-2004 | Shared |
| NEC SX-6/144M18 | 2004 | BoM/CSIRO |
| NEC SX-6/224M28 | 4Q2004-2007 | Separate partitions<br>CSIRO 5 nodes |

# New Location at 700 Collins Street



# Phases of Current Contract with NEC

NEC
SX-6

TX7/i9510
1.3 GHz

| | April 2004 | Oct 2004 – 2007 |
|---|---|---|
| Nodes | 18(13 BoM) | 28(23 BoM) |
| CPUs | 144 | 224 |
| Peak Performance | 1,152 GFLOPS | 1,792 GFLOPS |
| Main Memory Unit | 1,152 GB | 1,792 GB |
| Disk Capacity | | |
| Total | 16 TB | 26 TB |
| GFS | 14 TB | 22 TB |

| | April 2004 | Oct 2004 – 2007 |
|---|---|---|
| Nodes | 2 | 2 |
| CPUs | 24 | 32 |
| Memory | 24 GB | 32 GB |

# Overall Increase with SX-5 Comparison

- 11.5 times increase in peak CPU capacity
- 13 times increase in memory
- 14 times increase in disk capacity
- Similar disk speed increase
- Performance
  - \+ 5%-30% faster per processor for our major applications
  - – some performance degradation for multithreaded runs or multiple single CPU jobs within a node

The Domains of the Operational Suite of

# Numerical Weather Analysis and Prediction Systems


## GASP

Global Analysis and Prediction System

- $T_L$239 (smallest half wavelength resolved : 83 km)
- 29 levels
- Prediction to 8 days

## LAPS

Limited Area Prediction System
Australian Region

- 0.375° horizontal grid spacing
- 29 levels
- Prediction to 72 hours

## TLAPS

Tropical Limited Area Prediction System

- 0.375° horizontal grid spacing
- 29 levels
- Special features for tropical analysis
- Prediction to 48 hours

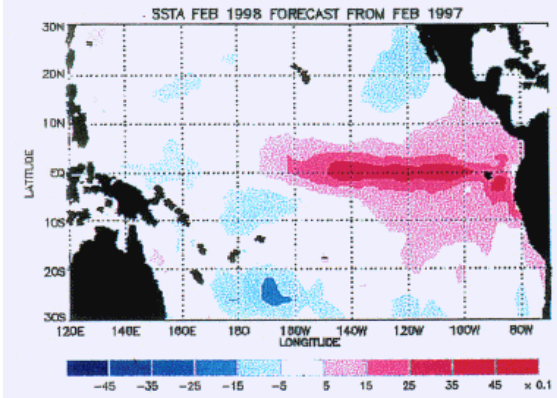
## MESO-LAPS

Smaller scale versions of LAPS

- 0.05°-0.125° horizontal grid spacing
- 29 levels
- Prediction to 36 hours

- Special version for tropical cyclone prediction

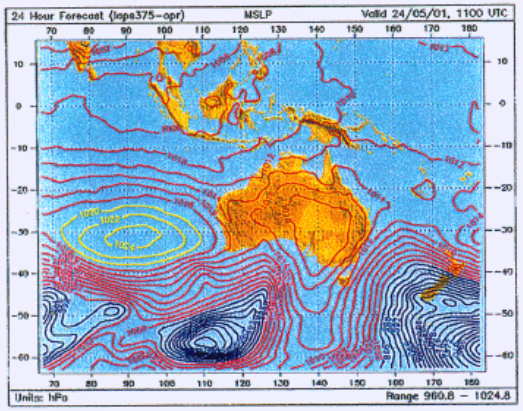
# Unified BMRC Atmospheric Model



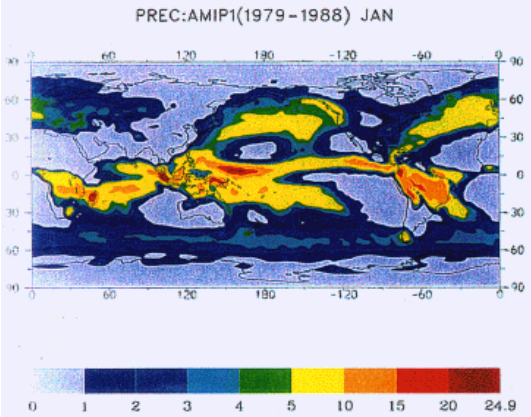

seasonal forecasting
(coupled)

**BMRC Atmospheric Model**

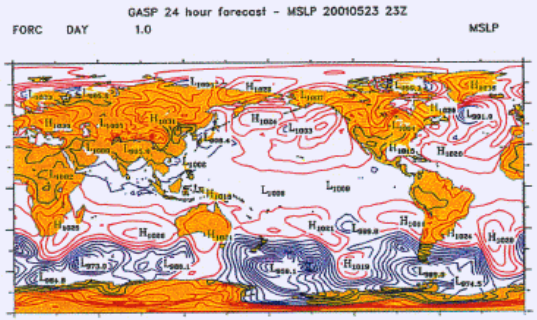
**BAM**



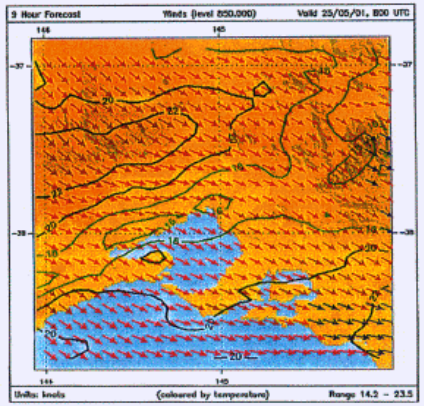

regional forecasting




climate and climate change
(coupled)



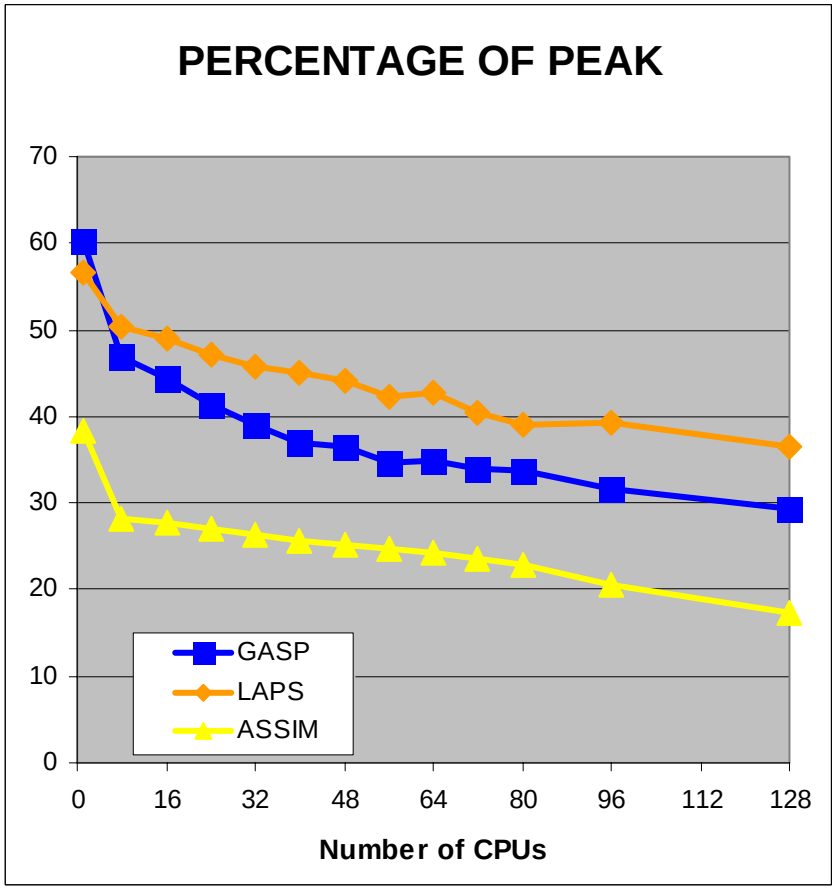

medium range forecasting

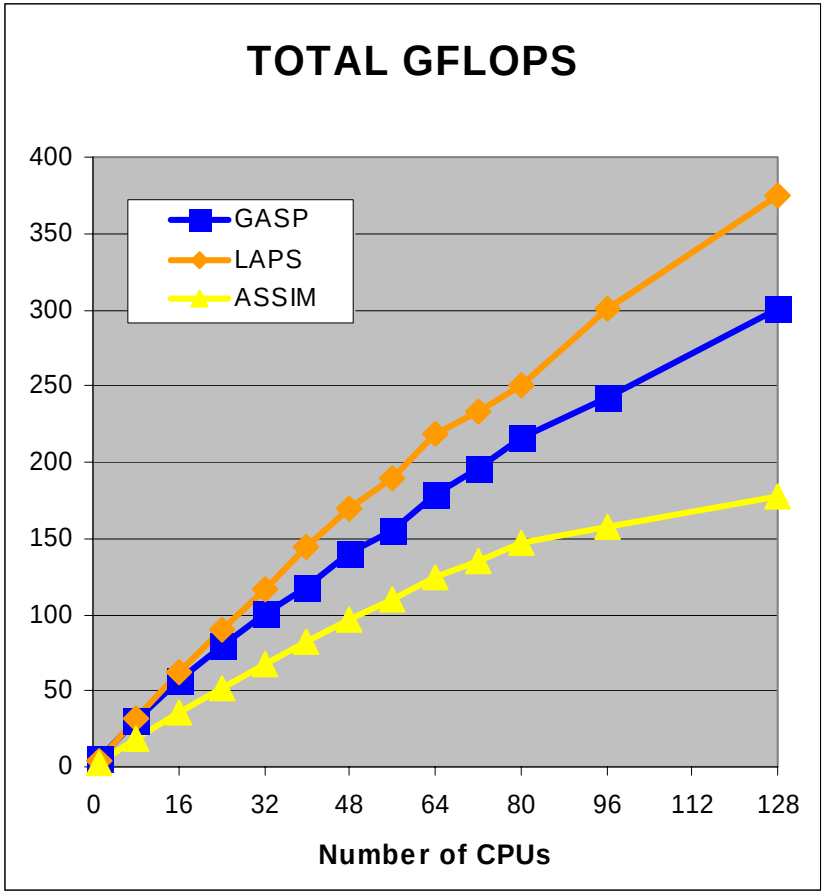



mesoscale forecasting

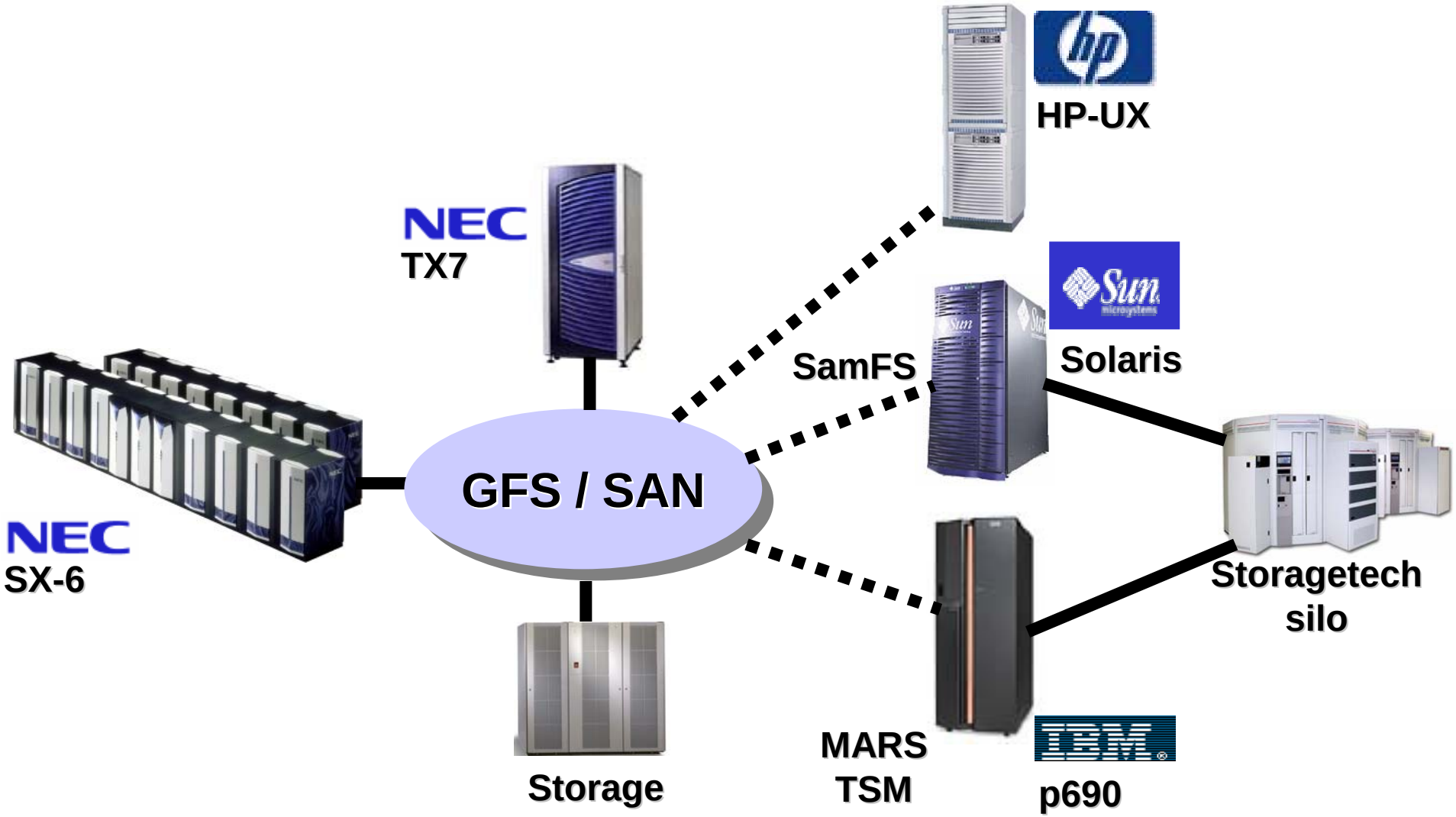
# Performance Scaling to 16 SX-6 Nodes for T479L50

(Benchmark Configuration)



# Principal Components of BoM HPC System



# GFS Overview

- User data is shared via NEC Global File System (GFS)
- GFS Server and Client on NEC TX7 IA-64 Linux front end server; GFS client on SX-6's
- XFS journaling file system on Linux
- SX-6 nodes access data via
  - NFS for **<=64** KB I/O requests
  - GFS for **> 64** KB I/O requests
- Availability of GFS clients for HP, IBM, Sun have also been promised in our contract

# Critical Features of TX7/GFS Design

- Need TX7's to perform to extremely high reliability thresholds to guarantee system availability for critical operations
- Need duplicated TX7; with fail over to avoid single point of failure

# Pros & Cons of GFS

- **Pros**
  - highly convenient file system visible from all SX-6 nodes and TX7
  - high performance data transfer (near local disc speed for large files)
  - fail-over capability between two TX7 servers
  - based on SGI's Open Source XFS for Linux

# Pros & Cons of GFS (cont.)

- **Cons**
  - Small block I/O uses NFS
  - No prioritisation of GFS I/O
  - No caching of GFS file systems compared with SX-6 file systems
  - GFS I/O bottleneck:
    - heavy I/O may cause significant performance degradation for applications using same file system, even on other nodes
    - better to use **local file system** (110GB per node) or MFF (6GB per node) for heavy I/O;
      can't migrate these jobs to other nodes if checkpointed

# GFS Usage Tips

## Efficiency

- Setting large enough buffer sizes for I/O (F_SETBUF)
- Specify optimal buffering for direct access I/O (F_HSDIR, should really be default)
- Large data transfer steps done by batch jobs executed on TX7

## Convenience

- **do_tx7** script to do a command in the same directory using rsh on a remote host;
  perform minor scalar operations such as rcp, gzip, etc on TX7 from SX-6 batch jobs
- Likewise use of **do_sx6** from TX7 to SX-6

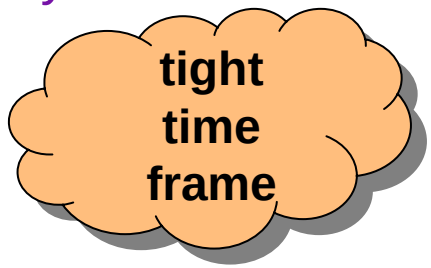
# Example of Local File System Impact

- Usage of local file system in operational GASP EPS (T119L19): 32 perturbed members, 4 batch jobs of 8 members executed in background with each member on 1 CPU, running in parallel on 4 nodes
  - **>40min** with I/O on GFS (time may vary a lot depending on GFS traffic to same device from other jobs)
  - **15-16min** with I/O on cached local file system ($LOCALDIR)
- Time for copying files from GFS to local file system and back is negligible in comparison with doing I/O from programs directly to GFS
- Further investigation is continuing with NEC cooperation, aiming to achieve near local file system performance with I/O direct to GFS

# Stages in SX-6 installation

- Sep 2003 – Jan 2004: porting applications on two nodes off-site SX-6 system
- Dec 2003 - Mar 2004: NEC installation and testing of the full system at 700 Collins Street

- 26 Mar 2004: user access
- 19 Apr – 19 May 2004: acceptance testing reliability trial
- 13 May 2004: BoM operations switched to SX-6
- 31 May 2004: the SX-5s switched off



# Porting experience from SX-5 to SX-6

- Binary compatible
- GFS worked smoothly
- Single node porting straightforward
- Not enough memory bandwidth

**performance degradation in multithreading and/or impact of other applications running on the same node at the same time**

**cause: memory contention**

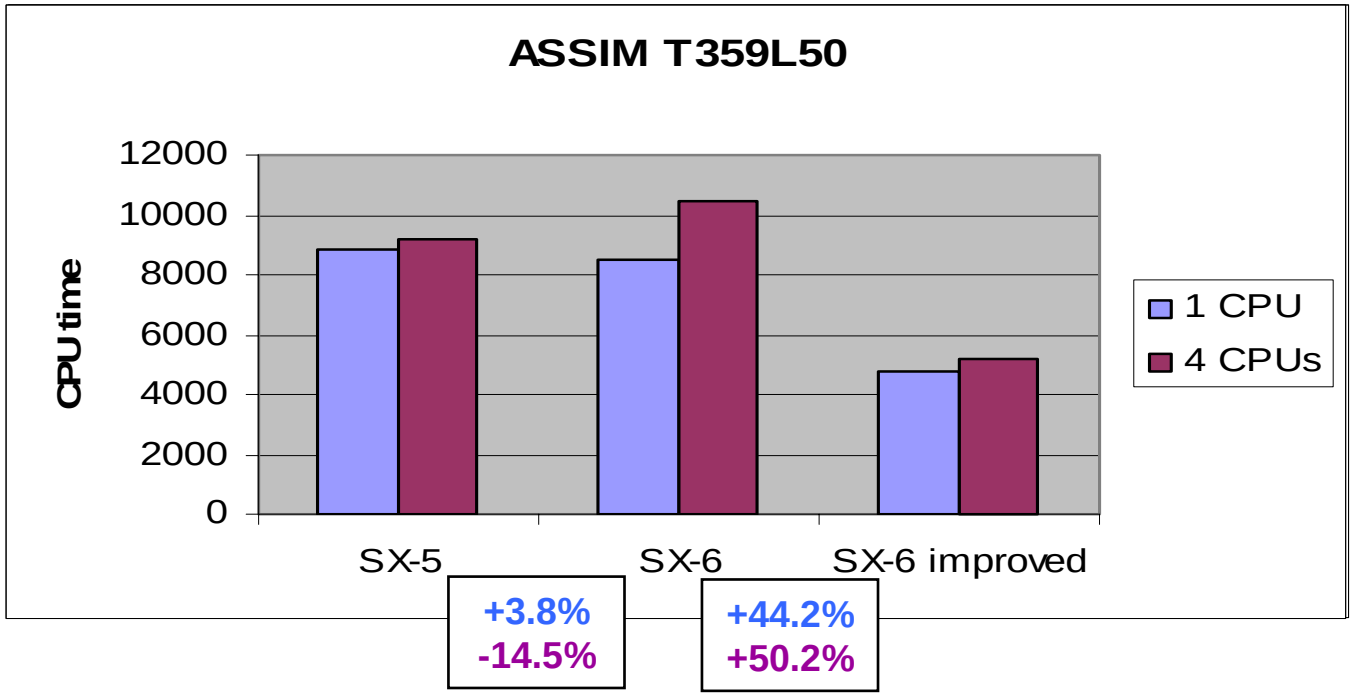
**Node comparison**

| | SX-5 | SX-6 |
|---|---|---|
| CPUs | 16 | 8 |
| Peak Performance | 128 GFLOPS | 64 GFLOPS |
| Main Memory Unit | 128 GB | 64 GB |
| Memory Bandwidth per CPU | 64 GB/sec | 32 GB/sec |
| I/O Bandwidth | 12.8 GB/sec | 8 GB/sec |

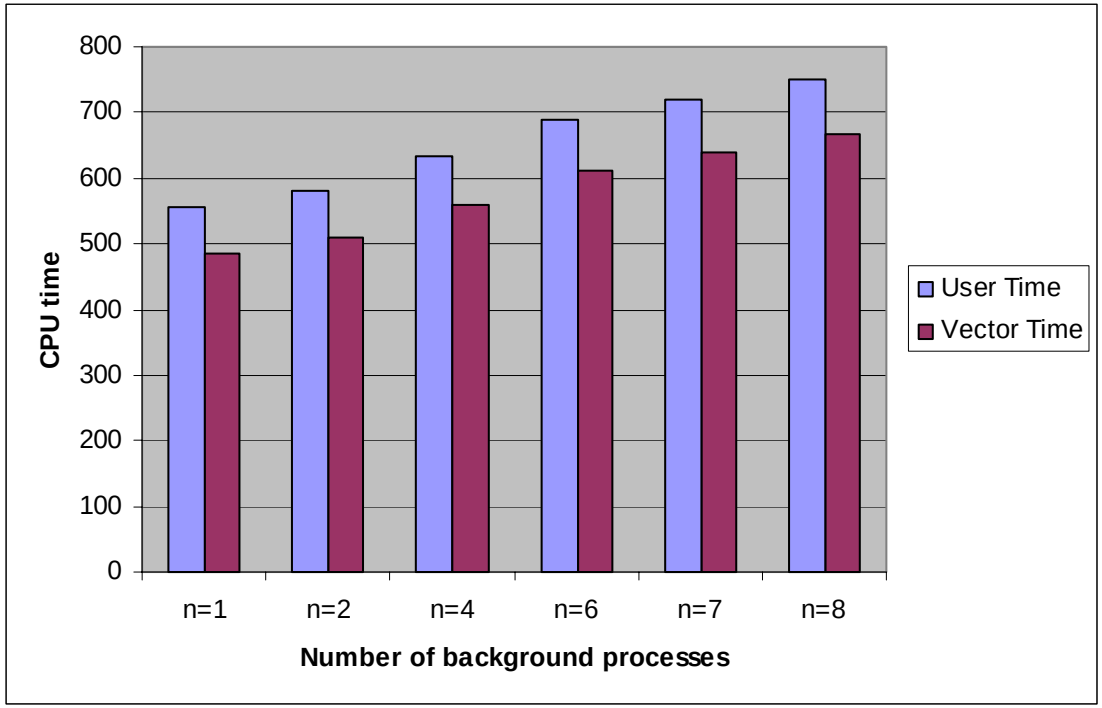
- Slow multithread performance in assimilation triggered performance tuning which delivered big improvement (see next slide)

# Performance tuning results for GenSI ASSIM

**Slow multithreading performance in assimilation code triggered tuning which delivered big improvement**



# Memory Contention in GASP EPS



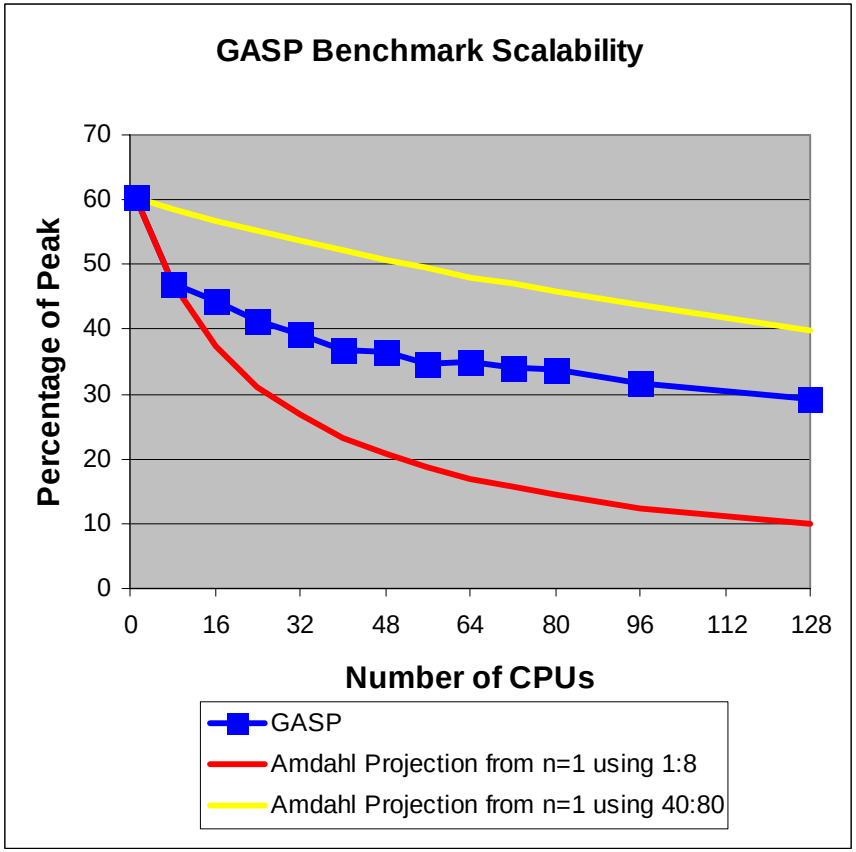

**1-8 parallel 10 day forecasts each running on 1 CPU on the same dedicated node**

- **CPU time per forecast increases by 195sec (~35%) overall**
- **CPU time increases almost linearly by ~26sec per parallel forecast**
- **CPU time increase is wholly due to vector time increase**

*CPU times are reproduced with forecasts run*

- *in batch parallel jobs*
- *via MPI wrapper*

# Intra-node versus Inter-node Scalability




- Intra-node scalability is ~96% (i.e. 1 CPU to 8 CPUs)
- Inter-node scalability is ~99.6% (i.e. 5 nodes to 10 nodes)
- Difference is mainly due to memory bandwidth/memory contention effects

# SX-6 usage

- **Operational runs**
  - jobs are submitted via SMS (ECMWF Supervisor Monitor Scheduler) scheduler in batch from operational HP server
  - high priority

- **Research**
  - batch jobs are submitted from TX7 or HP-UX
  - users not bound to specific nodes
  - very limited interactive access to the SX-6 nodes (only for debugging)
  - one node dedicated for performance testing
  - tight limits on TX7 usage
    - cross compilation is done on HP-UX and not on TX7
    - not more than 2 window sessions open by a user
    - no processing apart from GFS, NQSII, development scripts & jobs

# Scheduling

- ERSII and NQSII
- Gang scheduling
- Different queues for research, operational jobs
  - operational jobs spread across nodes to avoid contention between high priority tasks
  - single node and multi-node queues for research
- Max of 7 CPUs per process can be used on single node (to avoid contention with system accounting, etc daemons)
- Multi-node scheduling still in very early stages (not operational; tuning probably still required)

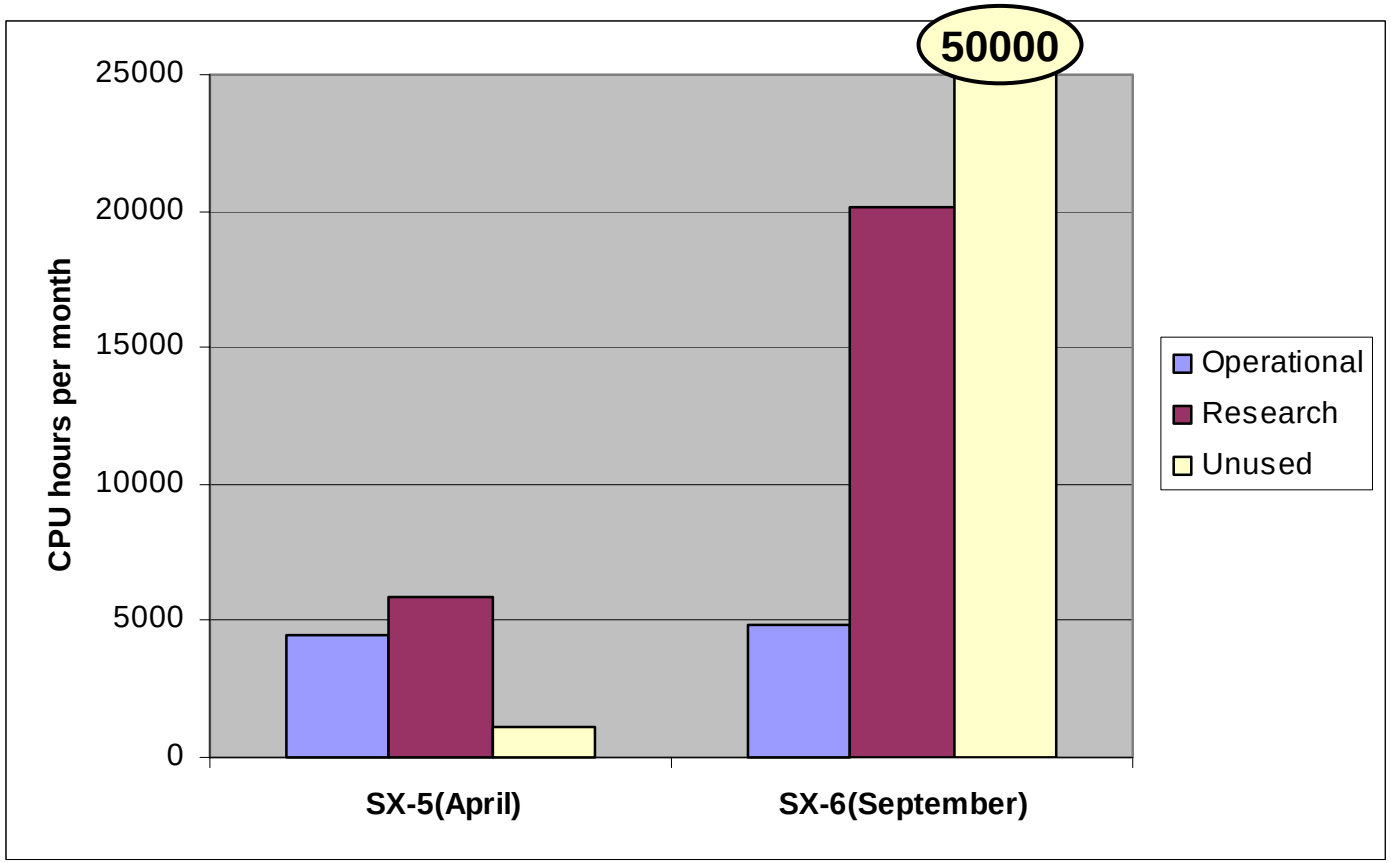
# Current Operational SX-6 Use

- **Short Range (runs twice a day, L29 for all models)**
  - LAPS 0.375º Australian Region data assimilation and 3 day prediction system
  - TLAPS 0.375º Tropical Australasian Region 3 day prediction system
  - LAPS 0.125º Australian Region 2 day prediction system
  - TCLAPS 0.15º 3 day tropical cyclone prediction system
  - MESOLAPS 0.05º mesoscale 2 day prediction system for Victoria-Tasmania, Sydney domains and South East Queensland

- **Medium Range (runs twice a day)**
  - GASP (T239L29, 75km) 2x6hours assimilation analysis and 10 days forecast
    - GASP (T239L33, scatterometer data) 2x6hours assimilation analysis and 10 days forecast run routinely in parallel, will become operational very soon
  - GASP EPS (33 members, T119L19, 150km, 10 days forecast)

- **Seasonal Forecasting**
  - POAMA (Predictive Ocean Atmosphere Model for Australia, T47L17, 350km) coupled model 8-9months seasonal forecast; run daily to produce time lagged ensemble

# BoM HPC System Usage: SX-5 vs SX-6



# Data Archiving

| System | Platform | Current archiving rate | Current archive volume | Transfer rate |
|---|---|---|---|---|
| MARS1OP/TSM<br>MARS1RE/TSM | IBM p690 4CPUs<br>IBM p690 4CPUs | 5 Gb/day<br>(0.15 Tb/mth) | 6Tb | 70Mb/sec |
| SamFS | Sun SF 4800<br>8 CPUs | 300-350Gb/day<br>(9-10 Tb/mth) | 185Tb | 50Mb/sec |

**Storagetech silo**

# Resolutions in the Future

| | Current Resolution | Future 2005-2007 ( * ) | Upgrade Resolution Factor | Upgrade CPU Time Factor |
|---|---|---|---|---|
| GASP | T239L29, 75km | T359L50, 50km | 1.5/1.7 | ~7 |
| GASP EPS | 33 members T119L19, 150km | 50 members T159L29, 112km | 1.6/1.3/1.5 | ~6-7 |
| LAPS | 0.375º, 29 levels | 0.25º, 50 levels | 1.5/1.7 | ~8 |
| LAPS EPS | 24 members 0.5º, 29 levels | 50 members 0.5º, 50 levels | 2.1/1.7 | ~4 |
| TCLAPS | 0.15º, 29 levels | 0.10º, 50 levels | 1.8/1.7 | ~9 |
| MESOLAPS | 0.05º, 29 levels | 0.05º, 50 levels | 1.7 | ~2 |
| POAMA Seasonal Forecasting | T47L17, 375km | T63L50, 280km | 1.3/3 | ~6 |

( * ) subject to demonstrating positive impact

# Conclusions

- NEC SX-6 is in place, delivering major HPC increase to BoM
- Transfer from SX-5 to SX-6 successfully made without major problems
- Operational jobs mostly faster than on SX-5
- TX7/SX6/GFS provides seamless environment for research and development
- System reliability is very high so far (no major problems over 6 months)

## Challenges

- System utilisation will need to rise as resolutions increase
- Job scheduling will need refinement as demand increases
- Some performance bottlenecks remain, which will need to be improved to realise future upgrade goals