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

<table>
<thead>
<tr>
<th>System</th>
<th>Time Period</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEC SX-4/16</td>
<td>1997</td>
<td>50:50</td>
</tr>
<tr>
<td>NEC SX-4/32</td>
<td>1998-2000</td>
<td>BoM/CSIRO</td>
</tr>
<tr>
<td>NEC SX-5/32M2</td>
<td>2001-2004</td>
<td></td>
</tr>
<tr>
<td>NEC SX-6/144M18</td>
<td>2004</td>
<td>BoM/CSIRO</td>
</tr>
<tr>
<td>NEC SX-6/224M28</td>
<td>4Q2004-2007</td>
<td>Separate partitions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSIRO 5 nodes</td>
</tr>
</tbody>
</table>
New Location at 700 Collins Street
# Phases of Current Contract with NEC

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>18(13 BoM)</td>
<td>28(23 BoM)</td>
</tr>
<tr>
<td><strong>CPUs</strong></td>
<td>144</td>
<td>224</td>
</tr>
<tr>
<td><strong>Peak Performance</strong></td>
<td>1,152 GFLOPS</td>
<td>1,792 GFLOPS</td>
</tr>
<tr>
<td><strong>Main Memory Unit</strong></td>
<td>1,152 GB</td>
<td>1,792 GB</td>
</tr>
<tr>
<td><strong>Disk Capacity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16 TB</td>
<td>26 TB</td>
</tr>
<tr>
<td>GFS</td>
<td>14 TB</td>
<td>22 TB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>CPUs</strong></td>
<td>24</td>
<td>32</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>24 GB</td>
<td>32 GB</td>
</tr>
</tbody>
</table>
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
- T1239 (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

BMRC Atmospheric Model

seasonal forecasting (coupled)

prec. amip(1979-1988) jan

climatic and climatic change (coupled)

medium range forecasting

BAM

regional forecasting

GASP 24 hour forecast - mslp, 20040923 00Z

mesoscale forecasting
Performance Scaling to 16 SX-6 Nodes for T479L50
(Benchmark Configuration)

PERCENTAGE OF PEAK

TOTAL GFLOPS

Number of CPUs

Number of CPUs

GASP
LAPS
ASSIM
Principal Components of BoM HPC System

- NEC TX7
- NEC SX-6
- GFS / SAN
- Storage
- MARS TSM
- Storagetech silo
- SamFS
- Solaris
- HP-UX
- p690

11th Workshop on Use of HPC in Meteorology, 25-29 October 2004
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 \( \leq 64 \) 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

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

<table>
<thead>
<tr>
<th>Node comparison</th>
<th>SX-5</th>
<th>SX-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Peak Performance</td>
<td>128 GFLOPS</td>
<td>64 GFLOPS</td>
</tr>
<tr>
<td>Main Memory Unit</td>
<td>128 GB</td>
<td>64 GB</td>
</tr>
<tr>
<td>Memory Bandwidth per CPU</td>
<td>64 GB/sec</td>
<td>32 GB/sec</td>
</tr>
<tr>
<td>I/O Bandwidth</td>
<td>12.8 GB/sec</td>
<td>8 GB/sec</td>
</tr>
</tbody>
</table>
Performance tuning results for GenSI ASSIM

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

![Graph showing CPU time for SX-5, SX-6, and SX-6 improved with 1 CPU and 4 CPUs.]

- SX-5: +3.8% improvement
- SX-6: -14.5% improvement
- SX-6 improved: +44.2% improvement

+50.2% 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 0.375° Australian Region data assimilation and 3 day prediction system
  - TLAPS 0.375° Tropical Australasian Region 3 day prediction system
  - LAPS 0.125° Australian Region 2 day prediction system
  - TCLAPS 0.15° 3 day tropical cyclone prediction system
  - MESOLAPS 0.05° 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

![Bar chart showing CPU hours per month for SX-5 (April) and SX-6 (September). The chart indicates a significant increase in usage for SX-6.]
## Data Archiving

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

Storagetech silo
## Resolutions in the Future

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

( * ) 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