Australian Government
Bureau of Meteorology

Weathering the storm – Lessons learnt in managing a 24x7x265 HPC delivery platform

Craig West
HPC Systems Manager
Bureau of Meteorology

Australian Government
Bureau of Meteorology

Introduction

- We are Australia's national weather agency.
- Providing weather forecasting, extreme weather, space weather, flood/hydrology and climate advice
  - Aviation, maritime, military, agriculture, commercial groups
  - The public
  - The government
- Our teams
  - Scientific Computing Services
    - Systems Admins
    - Storage Admins
    - Applications Support
    - Optimisation Support
  - Model Build
  - National Operations Centre
  - IT Command Centre
  - Cray on-site staff

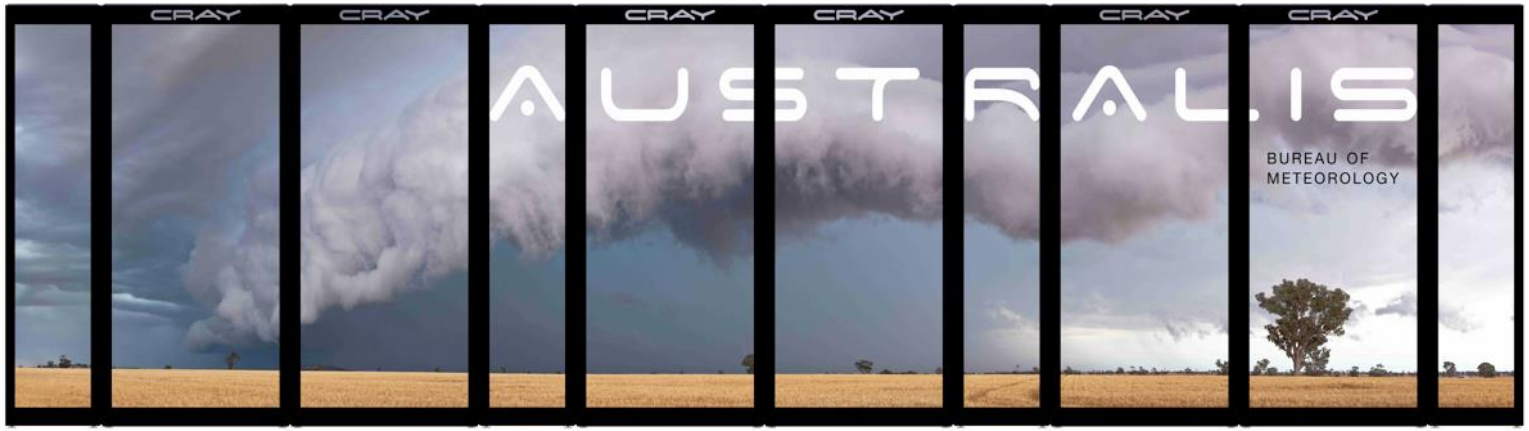
} HPC on-call support

} Model on-call support

Australian Government
Bureau of Meteorology

What we have - XC

- Compute environment – CLE 5.2UP04
  - Production – 2 x 6 cabinet XC40-LC (pair of halls)
  - Development – 3 cabinet XC40-AC
  - Test – 1 cabinet XC40-AC



- Storage systems
  - Production – 2 x 3 SSU Sonexion + 2 x 6 SSU Sonexion
  - Development – 6 SSU Sonexion
  - Test – 1 SSU Sonexion

Australian Government
Bureau of Meteorology

What we have - CS

- CS400 Compute x 2 halls – RHEL 7
  - 16 node cpu cluster
  - 4 gpu node
  - 3 service nodes
  - Local NVMe
- DDN Gridscaler 14K x 2
  - 10 enclosures
  - GPFS / Spectrum scale
- Also have a small test system with a few nodes and a DDN.

- New resource aimed at reducing need on MAMU nodes.
- Supports high I/O and integration with other platforms.

How we use it

- 24x7x365 Operations
  - Focus on keeping business critical workloads running
  - System design needs to be resilient
  - Software applications need to understand system design
  - Service Level Agreement with Cray
    - Maintain a production capability
- XC + CS = HPC service
  - Jobs run where they are most suited

Schedulers

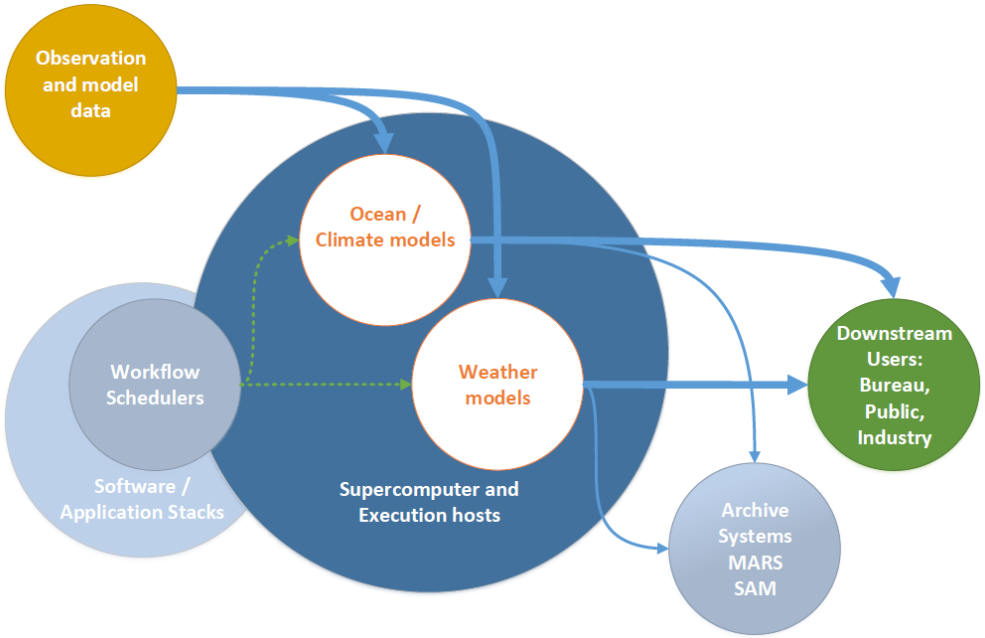
- PBS Pro batch scheduler
  - A batch scheduler per system
  - 2 production halls are considered one system
  - 60,000 jobs in production per day, and growing
- Rose/Cylc and SMS workflow schedulers
  - Regular and on-demand workflows
  - Submit the scheduled and on-demand applications to PBS Pro
  - Monitors health of running jobs

Australian Government
Bureau of Meteorology

Applications

- Built by a team of developers and tested through Dev and Pre-production
- Deployed by the HPC support team into production
- Provide capacity profiles
- Two primary types – Scheduled and On-demand

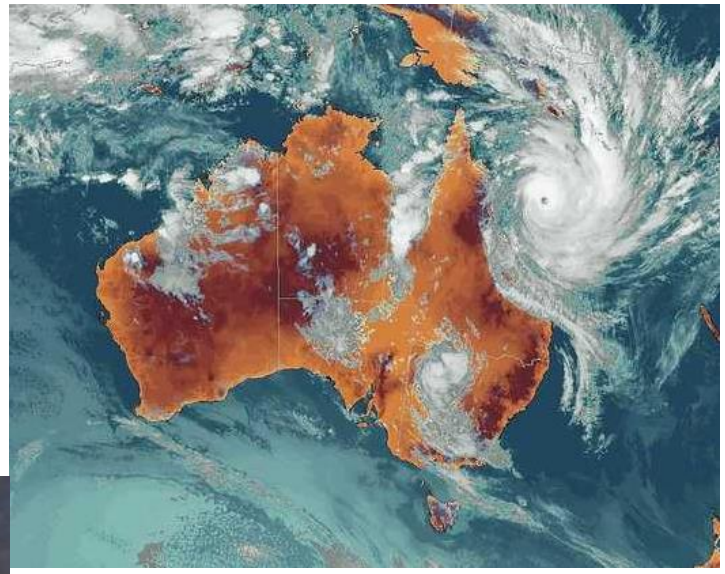
- Responsibility of applications to backup their own data for storage failovers
- Automated builds from GIT
- Artifactory binaries

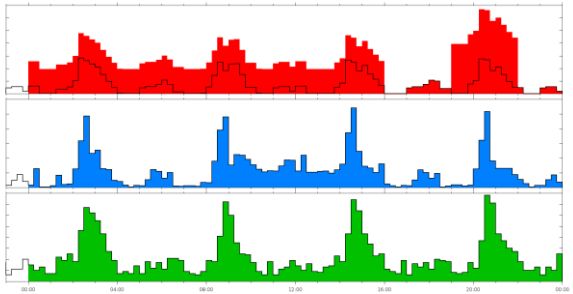


Australian Government
Bureau of Meteorology

Capacity Management

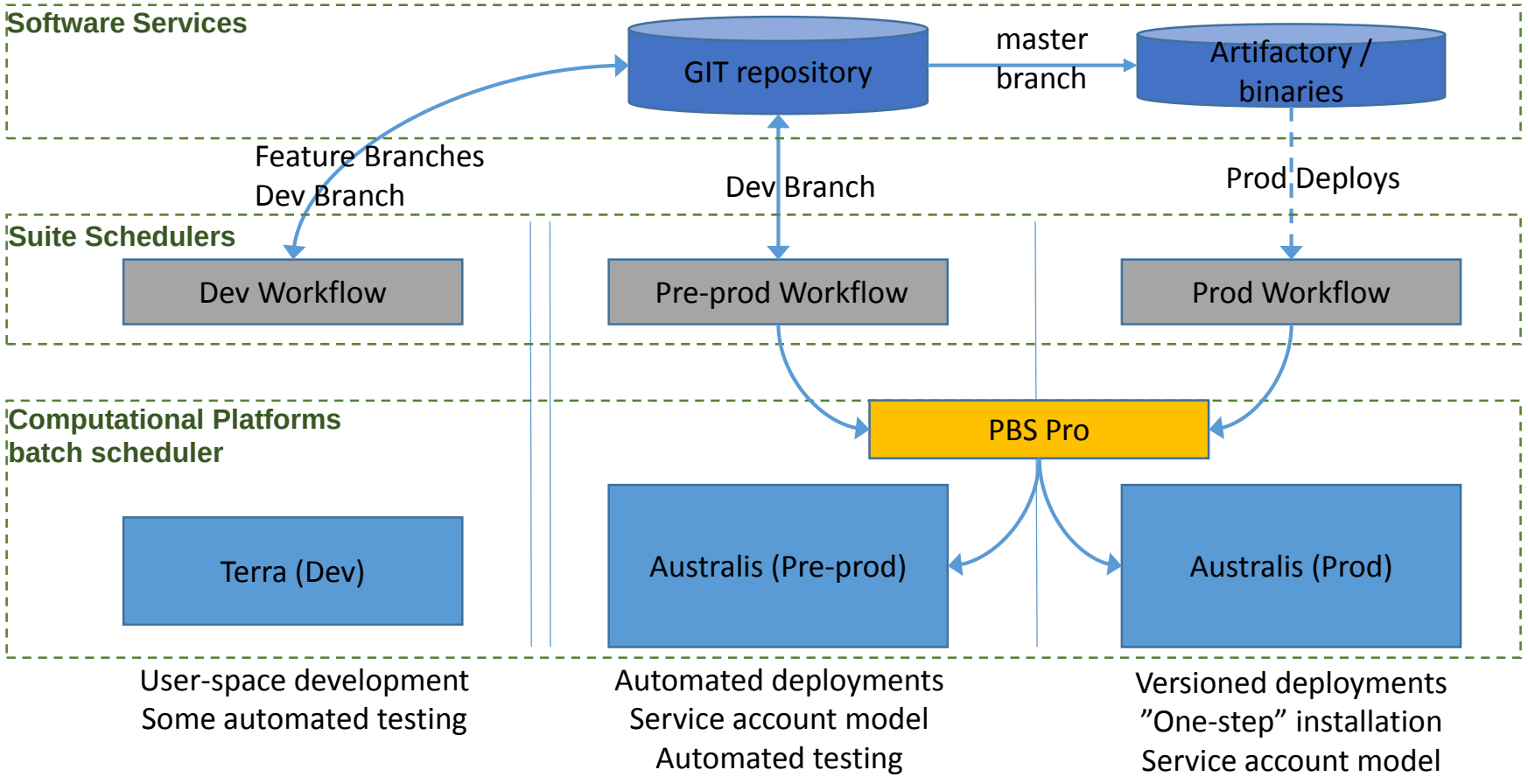
- Current system x16 time bigger than previous
- 5+ year plan for new and updated weather models
- Capacity profiles for each application
- Predict production utilisation
  - Ensure capacity available for production
  - Provide some resources for non-production
  - Seasonal variation
- Plan for growth





Australian Government
Bureau of Meteorology

Software lifecycle



Users

- People
  - Real users on our Development systems
  - Have access to view production environments (read only)

- Service accounts for applications
  - Automated accounts
  - Limited command set
  - Limited host access
  - Grouped by application area
  - Dev, Pre-prod, Prod

How we operate our systems

- Configuration management
  - Systems Admin – two distinct groups
  - Cray – hardware and OS support
  - Bureau staff – OS support, integration, applications

- Operations
  - Staff work normal business hours
  - ITCC triggered callouts to on-call staff
  - Maintenance during business hours

How we manage the systems

- System monitoring and configuration
- Daily reporting with a summary per system
  - Cray provided functions where available (some custom)
- System monitoring
  - Gathers real-time settings
  - Compares the configuration
  - Does this daily with a summary per system
  - Comparison can be trigged between systems

- Change Process steps through the systems to minimize risk to production
  - Work through Test, Development, and Production

Git tracking

- Out of the factory our systems were different
  - For the production halls we needed them matched
  - Our Dev and Test systems also needed to be similar
- GIT differences
  - Generated as a part of our backup systems
  - Enables daily comparisons to any given system
  - Can compare differences between systems
  - White/blacklisting of files and directories
  - Tracks changes in an environment where two parties support the systems
  - SMW and CIMS
  - Sonexion
- Example: RSIP configuration

Backups

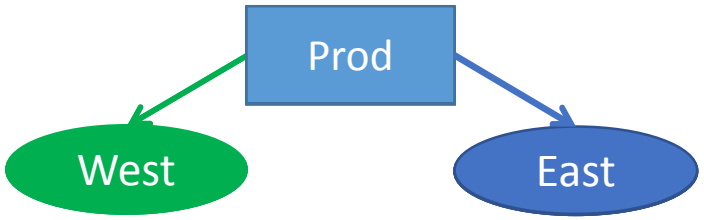
- Follow Cray recommended procedures, but with a tweak
  - Blue/Green/Red images
- Extensive backups done during patching cycles
- Customised process for off system backups
  - SMW & CIMS initiate their own and their sibling backups
  - Files pushed to external NFS target
  - NFS target snapshots and tape

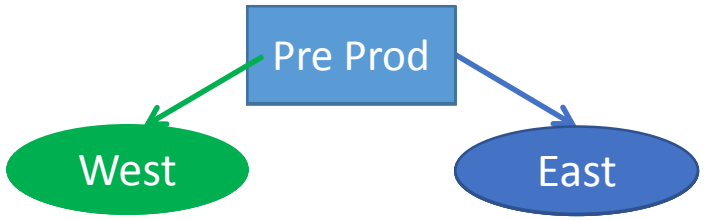
- Lustre data is not backed up.
  - Applications replicate their necessary data
  - Data is generally short lived

Australian Government
Bureau of Meteorology

Failovers - Compute

- Applications are not aware of which mode system is operating in
- Failover controlled by PBS Pro
- Change which Nodes assigned to each Queue
- Can suspend queues to drain halls, can also suspend workflows

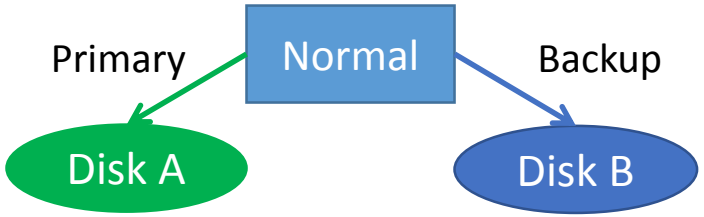


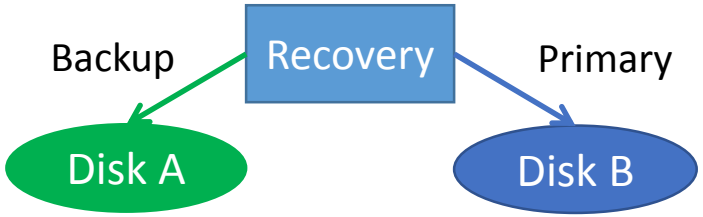


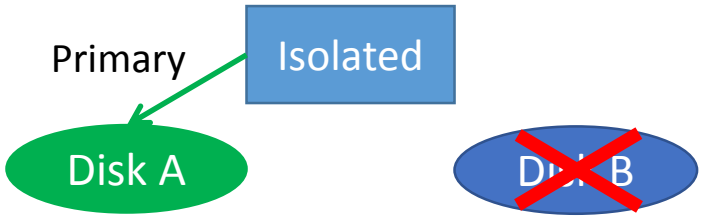
Australian Government
Bureau of Meteorology

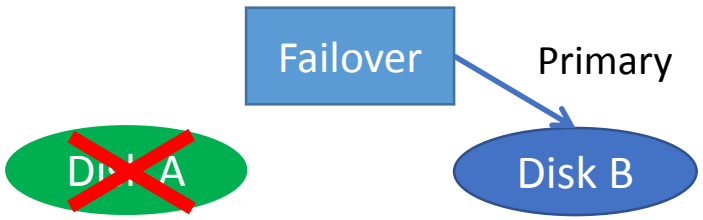
Failovers - Storage

- Storage targets are configured in pairs, small and big pairs
- Writes are done to a primary, and the application copies only necessary information to a backup target.
- NWP Suites know about storage modes and automatically adapt as needed.



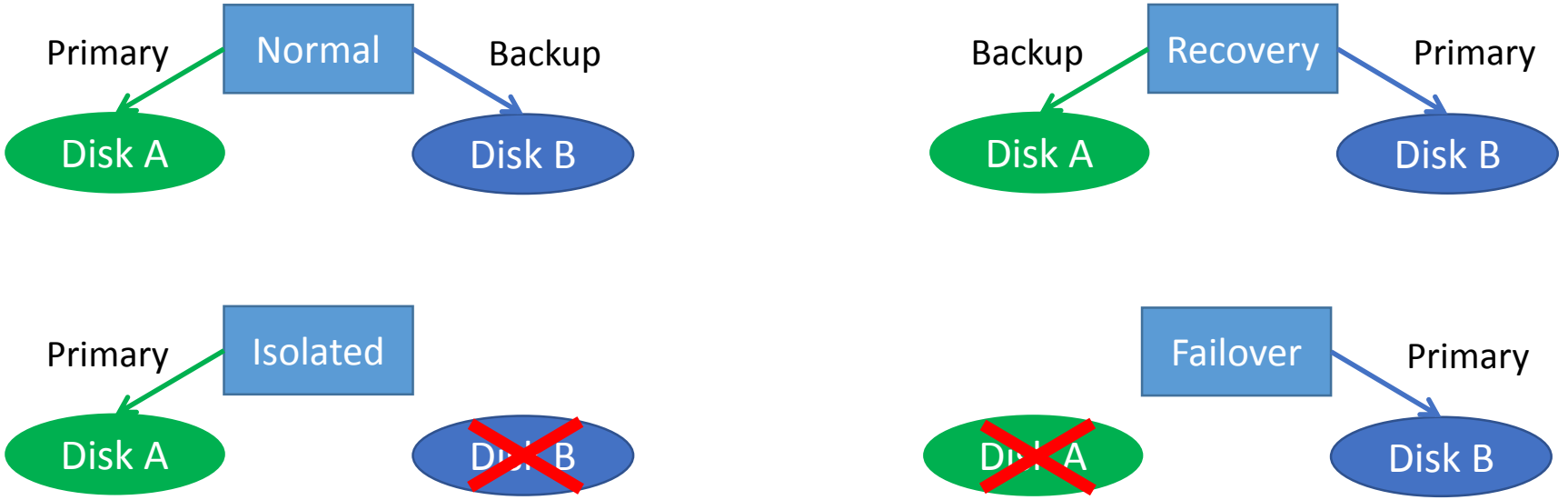






Australian Government
Bureau of Meteorology

Failovers – Storage example



System and Storage Patching

- Take advantage of having a *similar* TDS
- Patch the Development system next *(can be skipped if urgent)*
- Patch only a non-production hall (hall is offline during patch)
- Test the patched system is stable as pre-prod
  - Get applications team to run many different models
- Move production to patched hall
- Following day patch the now non-production hall
- Verify the halls are again as similar as possible

- Storage patching is similar
  - Isolate affected target and unmount
  - Patch
  - Mount and test
  - De-isolate target

Testing

- Carried out often – not regularly – depends on weather!
- Compute failovers testing during hall patching
- Storage failovers tested during Sonexion upgrades
- Sustained System Performance (SSP)
  - Runs weekly, reported monthly
  - Run as part of patching
  - Has caught issues
- Applications in pre-prod are test cases for production.
- Applications test storage failovers

Conclusion

- We have a good track record for keeping operations working 24x7x365
  - Almost two years of operations
- We track what is happening to the system
  - During changes
  - Between systems
- Our Systems and Applications work together
- Resilience is in the system design
  - Two halls increase our production availability

- Forecasting the weather is a complex process
  - So is running the machines that support our forecasts

Australian Government
Bureau of Meteorology

Thank you

Contact details:
Craig.West@bom.gov.au