Weathering the storm

Lessons learnt in managing a 24x7x265 HPC delivery platform.

Craig West
Data and Digital Group
Australian Bureau of Meteorology
Melbourne, Australia
Craig.West@bom.gov.au

***Abstract*— The Bureau of Meteorology is Australia's national weather agency. Its mandate covers weather forecasting, extreme weather events and operational advice to aviation, maritime, military and agriculture clients. In a country of significant weather extremes, checking the forecast on the BoM is a daily ritual for most Australians, "the BoM" provides one of the most widely used services in Australia.**

This paper discusses how we manage our Cray XC and CS infrastructure, Lustre and GPFS storage environments delivering true measured production workload availability above 99.86% per month, whilst balancing maintenance needs and minimising impacts from system outages. These, like the weather, are often outside our control. Improving our resilience is a major focus.

The supercomputer is a significant link in the Bureau's weather forecasting value chain. Our HPC umbrella covers mass data input, output formatting platforms and the critical schedulers that launch and monitor our model suites.

I. INTRODUCTION

The Bureau of Meteorology is Australia's national weather agency. In its 100 year history its mandate has grown to cover not just weather forecasting, but extreme weather events and operational advice to aviation, maritime, military and agriculture clients. We often use the comparison of the US Weather service, which covers an area just slightly bigger than us, with similar extremes of weather happening at any one time, but we have 1/10th of their population and, inevitably, resources and funding.

In a country of significant weather extremes, checking the weather forecast is a daily ritual for most Australians, "the BoM" provides one of the most widely used services in Australia.

Graphical weather forecasts are data visualisations of vast amounts of processed data produced from Numerical Weather Prediction (NWP) suites. We generate at least 4 Terabytes a day from 5 Gigabytes of radar and weather observations. The forecasts are generated from many thousands of scripts that make up each NWP suite run on our systems.

On average we run +100 suites on any given day and this increases substantially during Australia's extreme weather season November-April when we add additional suites as and when required for cyclone, fire, flood and air quality events.

Our total workflow throughput is around 60,000 individual tasks every 24 hours. This number is increasing, and so are the per task resources, as we increase the resolutions of existing models and add new models. All of these tasks form a dependency chain that delivers our weather forecasting output 24x7x365. The HPC service is a critical step in delivering on the Bureau's goals to provide timely and accurate weather advice.

The stated aim of the Bureau's forecasting services group is to have zero deaths from extreme weather events through timely and accurate weather advice.

II. SCHEDULERS

Just about every HPC site uses batch schedulers, but with 60,000+ jobs running every 24 hours delivering against a regular timeframe we probably use our schedulers differently to most sites. Our chosen batch scheduler is Altair's PBS Pro. We use it primarily to launch the jobs onto our HPC systems with support for queue priorities, and special node type requests. Additionally, the capability exists to support suspend/resume scheduling during times of heavy system usage and high priority workloads. PBS Pro is our interface between the workflow schedulers and the compute resources, giving the systems support teams the capability to steer the workloads and provide a level of resilience in a seemingly transparent form to the NWP applications.

ECMWF's SMS workflow scheduler has been used as the main workflow scheduler for over 20 years, it is aging and out of development so we have our own coding team helping to support it. However, in that time SMS has become immensely stable, if somewhat limited in its functionality. We're currently running SMS alongside a new deployment of UK Met Office Rose/Cylc workflow scheduler. Together these two workflow schedulers deliver jobs to PBS Pro for batch processing on our HPC facilities, and along with visualisations the workflow schedulers provide monitoring and diagnosis to the support teams as needed.

III. APPLICATIONS

There are two main categories of applications on our systems; utility applications/libraries and the numerical

prediction applications for the main forecast models that produce the data that are fed to downstream systems and eventually to our customers. The NWP applications fall into two main types. There are those that run on a regular schedule like the atmospheric and ocean models. And there are on-demand models like cyclone tracking and air quality event monitoring that run during the times they are needed.

The NWP suites themselves are rarely standalone. They require a large input data set from both Bureau data sources and from other weather monitoring sites. Additionally the suites themselves can have complex interactions between one another, which requires cross suite triggering. The applications assume that the computational and storage resources they require are always available. So we ensure we understand the capacity required for the applications to predict their utilisation, and in essence forecast the future needs. Some applications are considered more important than others, and for that reason as the system utilisation becomes closer to optimal limits the introduction of suspend and resume features will be utilised, along with other features available in both the batch and workflow schedulers.

IV. Facility

Our facilities consist of a number of Cray XC and CS based systems for computation. These systems are paired with dedicated Lustre Sonexion and DDN/GPFS appliances for storage. The XC systems are currently XC40 based and include a single air-cooled rack XC40-AC which is our TDS/Exemplar, a three air-cooled rack XC40-AC which is our application development system, and a pair of XC40-LC halls with 6 liquid-cooled racks each for the production environments. Each liquid cooled rack is capable of holding three times more nodes than a single air-cooled rack.

The XC40-LC halls each have their own management systems, but jobs can be submitted to it seamlessly so it is considered a single system. These halls support production and pre-production applications. Lustre Sonexion 2000 storage systems accompany each set of the XC40 systems mentioned above in the following order; a single SSU Sonexion; a six SSU Sonexion; and two pairs (2 x 3 SSU and 2 x 6 SSU) Sonexions for the production environment.

The CS400 cluster comprises of a basic TDS/Exemplar system with a small DDN appliance, and a pair of CS400 clusters each of which have sixteen compute nodes, four GPU nodes, three application facing Service nodes, and a pair of management nodes. The nodes utilise dedicated NVMe drives for local application performance and GPFS caching. Accompanying the pair of CS400 systems is a pair of DDN/GPFS appliances with 10 disk shelves each.

Shared between the XC and CS systems are a group of nodes which provide high speed input and output connections to other systems with the Bureau. These nodes are batch scheduled like all others, but their primary purpose is to move the data around, including internally to given systems, between the XC and CS systems, and to systems external to our Cray environments.

The production XC40-LC halls share a PBS Pro batch scheduler, as do the production CS400 systems. This is to enable a seamless transition for applications when undertaking maintenance or recovering from system failures. All other systems have their own PBS Pro batch schedulers.

Additionally there is a small dedicated virtualisation cluster and an NFS storage cluster to support HPC production services. Non production systems have access to the Bureau enterprise virtualisation clusters and storage systems which are managed by a separate team.

The primary purpose for the HPC environment is to crunch the numbers needed to produce the weather forecasts. This includes short to long term forecasts, along with severe weather alerts and other on-demand applications such as tracking cyclones, volcanic eruptions and flood events.

V. Configuration Mangagement

Each of the four Cray XC systems arrived with a different system configuration, mostly minor differences. Ideally the two production capable halls should have a configuration that is as close to identical as possible, excluding items that need to be unique like hostnames. To do this meant we needed a way to compare our systems, so we created a Git repository to store all the configurations of our machines and we push our configurations to that system. This process has advantages such as providing us with a backup and historical information, along with being able to provide us with alerts when things change. Since doing this we have utilised the provided Git based differences to tweak our production systems to make them as similar as possible. One negative side to this as that things change constantly on the system (such as logs) we get a lot of false positive in the differences. This is because we are sending lots of information into the Git system. A refinement here has been the use of both white and black listing for files and directories. The Git repository also allows us to compare our production halls to our application development and TDS/Exemplar systems. This is a useful feature in that it confirms that the settings we applied during our testing of a setting change or patch are applied correctly on the production halls.

Alongside this our Git differences allow us to check the status of our Sonexion configuration settings and do some basic back up of the configuration settings. This has proved valuable a few times as the Sonexion updates tend to replace some files and settings with default options. We can quickly verify changes and reapply the changes we have done in the past.

The use of the Git diff system has provided many time saving features. One useful function is testing what has changed after a patch or reconfigure. For example, we have a specialised RSIP configuration, it is a Cray supported and documented one, but it is not part of the reconfiguration options for generating a new XC image. So after generating the image we need to modify RSIP settings, thankfully it doesn't need to be done often, but it would be an easy step to overlook. Running a Git difference update shows us the change quickly, and it could be easily fixed by just restoring the previous file, as we could with this method. Another option is to simply query the system to see what the differences are and apply any changes that are needed. This option is useful in the case that a file has been changed by the patch or some other process, as we can see how it differs rather

than just replace a file with the previous version. In the case where an option has changed we might need to maintain that specific changed option but alter other settings, so simply replacing the file could be dangerous.

The production environment is the area that needs to be the most stable, but that doesn't mean the other areas are not important and protected. Like most IT teams we follow a change management process. Making changes into the production area takes time. For system changes the effort is spent working through our TDS/Exemplar system, followed by our development system, and then finally into each of the XC halls, one at a time. We build up the process and apply similar changes to the next environment. Software applications follow a similar process, they are developed and tested, moved into a pre-production stage where they may get some refinement and are given stability testing and finally deployed to production following a detailed approvals process. The HPC Support Team performs the deployment to the production environment, acting as gate keepers and ensuring the appropriate documentation and approvals are in place. The whole process is supported by those responsible for building the application. The source code for the applications are kept in software repositories and for the builds aimed at the production environment, the applications go through an automated process with artefacts stored in an Artifactory repository. It is from the Artifactory repository that the final deploy to the production environment takes place.

VI. FAILOVER RESILENCE

To provide the highest level of availability for the applications to operate we have devised a system whereby we can migrate the compute or storage workloads to specific parts of the system. This allows us to isolate components either because they are faulty or need to under maintenance. This has been achieved by ensuring that the system has two compute halls that are effectively identical, along with storage systems created in pairs. Our failover process is governed by a standard change management process that allows us to follow a set of defined steps to expedite the change.

The service level agreement covers having a compute hall available for production use, along with an associated storage target. So any single hall or storage system can be offline due to an issue or be undergoing maintenance without affecting the production level SLA. The affected hall still has an SLA associated, but it is at a lower level than the one representing production.

A. Compute halls

To achieve the resilience for the compute systems a single PBS Pro service exists for the production halls which is comprised of a failover pair of servers. This service schedules all jobs onto the two production XC halls and the associated CDL nodes. This includes all production, pre-production and system related PBS jobs. The majority of the jobs are submitted from either SMS or Rose/Cylc workflow scheduler.

Normally we run production on one hall, and pre-production and everything else on the other hall. A compute hall failover is a trivial matter of selecting which nodes are attached to which PBS queues. We have queues dedicated to production and pre-production, which are also given different priorities. This queue adjustment is done in about 2 minutes, during which time we suspend the launching of new jobs, but we don't stop already running ones. Depending on the reason behind the hall failover being triggered, we can suspend non-production queues and drain halls to ensure that capacity is available. However, in the case that a compute hall is non-responsive our action is to power off that hall to ensure that the workload ceases to exist before we restart the jobs from their previously known good states. This method is used as if the non-responsive system became responsive again it could try to continue processing the existing workloads, which may result in corruption. Upon recovery of the compute halls, the actual restart of jobs is performed by the suite schedulers, not PBS Pro.

If the halls are correctly functioning and there are applications running, then the failover can be done live in which the new jobs would launch on the opposite hall and the old jobs would be allowed to finish on their existing nodes. This is the process that is undertaken most often. It is used when doing things such as system patching, or testing compute failovers.

We also remove any unhealthy nodes from the queues for node level maintenance, and they are returned after a successful 24-hour period of triage testing. There is a queue in which we place any suspect nodes, or those returning from hardware fixes. It allows the Cray staff to ensure that the nodes are functioning correctly without them being interrupted by jobs. It also minimises the chances of production workloads failing due to a node being unhealthy. Additionally, we also undertake hardware maintenance on nodes not in the production hall.

B. Storage systems

The storage failover is a more complicated process than the compute failover. And, unlike the compute process, the storage failover requires that the NWP applications can detect which mode the storage targets are currently using. In our XC systems we effectively have two separate file storage targets, a large and a small one. Each target is a pair of Sonexion 2000 Lustre appliances, with one of the disks in each pair considered the primary and the other the backup target. Each pair can be placed into one of four modes called: Normal, Failover, Recovery and Isolation. The modes allow us to take a file system offline for maintenance, or isolate it if there are issues. If we consider each pair to have an A and a B disk, then the modes work as follows:

- Normal mode – primary writes are to disk A, with the backup to disk B. This is the day-to-day mode and is where we ideally want to be.
- Failover mode – primary writes are to disk B, whilst disk A is not available. No backups can be done in this mode.
- Recovery mode – primary writes are to disk B, with the backup going to disk A. This mode is used after Failover, and is to ensure that the data replication that should have taken place during Failover mode is able to be 'caught up'.

- Isolated mode – primary writes are to disk A, with disk B not available. This is the opposite of Failover.

During any given run of an application the majority of reads and writes of data are done to the primary target. One of the final steps in the NWP suite run is to ensure that any data required for the next suite run is replicated to the backup target. It is the responsibility of the application to do this as the HPC support staff don't know what data is important nor can we copy it at the appropriate time. We also want to minimise the data that is replicated to save space, as some of the data is only used for temporary runs, or is able to be regenerated or recovered from elsewhere.

During the above mentioned modes there are some restrictions on having applications running on a given storage target. The basic rule is that if the primary target is changing, then no applications can be running whilst the storage mode for that target is altered. If only the backup target is changing, then this is considered safe to do whilst applications are running. We also need to spend a certain amount of time in the Normal or Recovery modes before switching. This gives the applications time to perform backups as most of them would run at least once during any 24-hour window. It is possible to trigger applications to run their backup commands independently from any scheduled or on-demand requirements. However, that is a human initiated process. If we were to consider disk A in a given target needed to offline, then the process we follow ensures we had been in Normal mode for at least 24 hours. Next we drain the jobs that need to use that storage target, change to Failover mode and resume the jobs. Now we can work on disk A with applications running only on disk B. Once disk A is returned we can change to Recovery mode without interrupting jobs. We stay in that mode for 24 hours and stop the affected jobs again, switch to Normal mode, and resume.

Unmounting a faulty Lustre disk from inside an XC platform is not always easy or successful. In some cases we may need to reboot a compute hall and not allow a faulty Lustre disk to remount before resuming operations. However, by changing where the targets point means we should be able to isolate applications from attempting to access a faulty storage target. This can allow us to delay a hall reboot, or fault find the storage issue whilst keep production operating.

VII. Patching the systems

Downtime to undertake patching of systems is a constant issue many sites need to deal with. The Cray XC environment is not well suited to sequential patching, and although the Cray CS product is a little better adapted it still has fundamental parts of the system that just can't be updated easily whilst the machines are operating. As such, our method of operations considers methods by which patching can be done without affecting production. As mentioned above we can migrate the production workloads between the compute halls, and even so with the storage system, albeit with minor pauses to workloads.

Therefore, our fundamental method of patching an XC system is to disable workloads on the hall about to undergo maintenance, ensure that the jobs have drained, or kill them off if necessary, and then undertake the maintenance. The hall under maintenance would already have been the non-production hall. To re-activate the hall after maintenance, we run non-production jobs for 24+ hours (ideally more than 36 hours) before we migrate the production workload back onto that hall. The production workload gets at least further 18+ hours before we then take the second hall offline to undertake the same maintenance as the first. A key goal here is that we are attempting to patch both halls within the same week as each other, and we are ensuring that the system is stable before moving production workloads onto it again.

We have had issues during patching, but the majority of those affect the pre-production environment, and this gives us a limited amount of time to resolve the issue. An example issue was the changing of a user environment variable post a patch. The environment variable that changed caused only four applications to have faults, but as it was detected early after the patch was applied we were able to update the applications before progressing too far. In this case we delayed the production hall failover by a few more hours. This sort of fix is called a "fix forward" in which we proactively fix the issue rather than rolling the patch back where we would then need to schedule a further outage to reapply the patch.

When patching any system there is an uncertainty attached as to if and what issues might be introduced. To minimise the potential for an issue in production, we first patch our TDS/Exemplar, followed by the application development system, and then finally the two production halls (one at a time as mentioned above). The other mitigation methods are to ensure we always have good system backups by following, where possible and practical, the Cray guidelines for system backups. We use the three image process with Blue/Green/Red images.

Our method may differ from other sites, so a quick explanation of our method is provided. Our Blue is our Normal image, if we are not running a Blue image then that tells the system admins that something is out of norm. The Green image is a cold backup of a good Blue image, it was taken with the XC cabinets powered down, and normally just prior to any patching of the Blue image. This is done during the same outage window as any patching, and is also generated straight after a clean reboot of the Blue image. The Red image is created as a live hot-backup of the Blue image. This Red image is taken the night before a patch and again about two days after the patch is considered successful, and any other time we decide necessary. The whole process is designed to allow us to roll back the complete image if the patch has issues, which is to switch back to the Green image. Or if there is a minor problem with the Blue image we can recover using the Red image.

VIII. Testing

Having plans and procedures for certain failures is only of limited use if systems are not tested regularly to see the procedures actually work. One method of doing this testing is used during the patching cycles, where either the compute hall patching triggers a hall failover or file system patching results in at least one of the two pairs of storage systems being unavailable. This isn't the only time testing is done. We also do testing in the pre-production environment so that

applications can be tested to verify that their storage target failover succeeds.

Our SLAs with Cray also require that our systems continue to perform at levels similar to those that they were purchased at. To do this we run weekly Sustained System Performance (SSP) tests which test compute, memory, network and storage. Effectively these provide a number by which we can evaluate that our system has not lost performance. These tests are actually representative of the applications we run and not just off the shelf benchmarks. We also use the SSPs at the conclusion of a patching cycle as a verification that the system is both capable of running jobs and an early validation that the applications are performing. We have avoided an issue in the past where our Lustre targets were not performing optimally after a patch due to running a disk performance test. Steps were then taken to quickly locate and fix the issue without any impact on production jobs.

IX. Backups

The Lustre and DDN storage systems are not backed up, they contain petabytes of data, and much of it is short lived. Any data that an application needs kept is sent to external systems. As mentioned earlier data that is needed for current and future model runs is replicated by the application itself. The NFS home areas are backed up with both file system snapshots and to tape. This is where the main configuration and binaries for the models are located. It is also where application output logs are stored during runs, before they are offloaded elsewhere to ensure they are kept for some time for the applications developers.

For the HPC systems the Cray backup procedures are followed. However these procedures don't cover all the backups that we would like to have, so there are additional backup methods we use and small tweaks to the way we manage backups. For the XC system backups we use the Blue/Green/Red images in a specific method which is detailed below in the patching description.

Our biggest backup customisation is for the SMW / CIMS systems and for the nodes that they manage. These backups are sent to an NFS target which has its own snapshots and provides a backup to tape. The customisation also allows us to keep the configurations in an off-system manner which in turn, as mentioned above, allows us to actually do some extra work with that data around monitoring the configuration and changes.

X. Users

Our data and products have a critical aspect so we greatly restrict user interaction on the production system. Suite developers and their support staff are segregated to our pre-production and development realms to develop and deliver fixes and patches when required.

Managing the quality of the applications deployed into the production realm is a constant process of improvement and worthy of a separate paper. Our recent lessons have educated us to the need for more robust acceptance testing, however the nature of the suites deployed, their complexity and with a development chain that stretches across years and multiple organisations, surprises and headaches in the form of overnight call-outs, unexplained system slowdowns and on occasion the imminent threat of a complete stop, are frequent operational considerations for us. However the schedule cannot stop, so we work in collaboration with the NWP suite development and operations support teams to fix and patch as soon as is practical.

The operations team described are supported by just shy of 60 people. There are more than a dozen of us that directly manage the HPC platform; the system, storage and scheduler administration, shared libraries and low level applications. We also have a number of onsite Cray engineers. The NWP suites are developed and supported 10-20 staff with the number varying as the need for application deployment changes. Monitoring and escalation is managed by a rostered team of 15 in our 24 hour IT Command Centre.

The Supercomputer Computer Programme is at the confluence of the Bureau's Observation, Business, Science and Technology streams. It directly manages stakeholder requirements from these groups and indirectly from multiple external business clients, the federal government, and global responsibilities as part of the Unified Model consortium.

The HPC support team spends time both supporting the day-to-day operations as well as building the next generation targets for both hardware and software systems.

XI. Conclusion

Keeping any modern HPC facility operating is a full time job for a large number of support personnel. When you add supporting operations 24x7x365 with allowed system downtimes expected to be less than one hour per month it gets very hard. Add in the complexity of the NWP applications and the uncertainty of the actual weather means our jobs are constantly evolving and require not only rapid responses but environments that have resilience built in. Our record thus far has shown we are managing this task, with nearly two years of production operations on our XC platforms our consistency in our uptime statistics are exemplary. The choices we've made in the build and configuration have been informed by the experience we've gained in decades of production output. We continue to adjust our environments to suit and add functionality that allows our resilience to increase, along with providing platforms that are suitable for the NWP product generation to occur.

This paper has given an overview to some of the methods used by the Australian Bureau of Meteorology to operate its Cray based HPC systems.

For further information please contact the Craig West, who is a member of the Bureau’s Data & Digital Group, Scientific Computing Services Team.

XII. Acknowledgment

Thanks to Tim Pugh, Andrew Khaw, Richard Oxbrow, Tim Connors and Mark Dean from the Bureau's SCP Supercomputer Programme and Scientific Computing Services SCS Teams for valuable input and review.