ATEMPO
preserving data ecosystems

HPC Data Management Solutions

An Atempo White Paper

HPC Data Storage: Managing the Growth Curve

CONTENT

SUMMARY

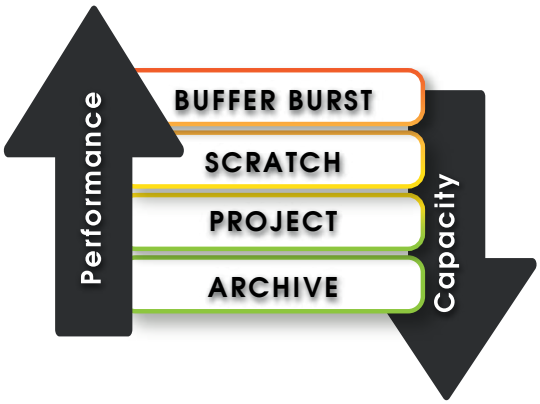
Data volumes stored in scientific research labs are growing at over 30% per year. Vastly increased compute performances are pushing storage loads to the limit. Moving data to middle and archive tier storages is a genuine technological and economical challenge.

This paper will look at the current HPC data lifecycle landscape illustrated by two specific customer use cases.

PERFORMANCE OR CAPACITY?
PERFORMANCE AND CAPACITY!

Formerly more distinct strands of HPC architectures, compute and storage components will continue to converge during this decade. On-prem and cloud data storages will need more speed and control to move data between storage tiers rapidly, securely and economically.

The levels of data stored on the high end of the IO performance scale remain relatively low. Burst and Scratch spaces where compute takes place occupies a smaller proportion of HPC storage infrastructures usually on very high-end SSD storage. Further down the storage tiers we have the computational ante-chamber (quota-controlled home directories) with a final archiving or store tier where many petabytes of data are kept often indefinitely on cheaper storage supports (cloud, tape…).



HPC-generated research data today has also been joined by other Big Data and AI applications requiring very high compute capacity. Moving this data to and from cloud storage can be long and costly. Hybrid offerings with on-prem or edge storage and the continued use of tape for archiving will play a growing role in HPC workflows enabling the intelligent movement of targeted data. For many HPC users it is not always possible to recompute and experiment because costs are too high. The initial results often need to be preserved and shared with peers for many years.

Different levels of service are available to HPC users within an organization. Universities for example can bill users on data retentions, data availability and data sharing. The complexity of moving data between different tiers, while ensuring optimal security and keeping costs to a minimum, requires HPC organizations to plan and refine their approach and workflows constantly.

This paper presents Atempo's Miria, the solution to orchestrate, move, protect, archive and migrate middle and storage-tier HPC data.

"Advances in HPC compute and storage workloads continue to drive data management challenges; as new storage technologies are leveraged comes the challenges of multiple storage silos, heterogeneous file systems and long-term storage requirements, this drives the need for reliable and efficient data orchestration tools."

Laurence Horrocks-Barlow, Technical Director, OCF Atempo, UK HPC Partner

STRENGTH IN NUMBERS: PARALLEL FILE SYSTEMS

Parallel file systems in general and Lustre-based file systems in particular are the mainstay of HPC file storage architectures (see www.top500.org). Lustre leverages end-to-end data throughput which can exceed 10 GB/sec and can also scale to many thousands of clients. This makes it ideal for HPC clusters and high-volume data environments.

Since 2010, the total volume of data stored in scientific research labs has increased from a few PBs to hundreds of PBs and continues to grow as fast as 30% per year, driven by a 1000-fold increase in system performance and a 100-fold increase in system memory. Data storage and management have become a key concern in effective HPC architecture choices.

To keep pace with data levels while managing budgets and offering a shared high computing service platform to multiple groups of users, research labs organize their HPC storage in tiers. Each tier has a dedicated purpose as well as distinct hardware, capacity and I/O characteristics. With the potential to rapidly generate many petabytes of data, users need to manage data generated at the Burst or Scratch tiers. The middle and storage tiers are where we need to move, manage and protect data.

- **Home Space:** In the middle, this storage tier is typically used by research teams to document processes, prepare code, write articles, manage their application source codes etc.

- **Archive Space:** storage used to preserve data and information for long-term purposes provides high volume capacities that can be organized to provide different levels of service, such as private archives or shared data archives. Depending on the research data and the domain of activity, this archive can be on-prem or cloud-based, or private or shared, or a mix.

Over and above the data which is generated on Lustre HPC architectures, other components also need protecting:

- Applications required for data computational needs,

- Lustre **MDS** and **MDT** components that are critical to managing and also recovering file system data (see the section below: "Reducing the impact of in-depth tree walking on Lustre").

DiRAC

PROJECT ARCHIVING

DiRAC - Distributed Research Utilizing Advanced Computing is the integrated supercomputing facility for theoretical modeling and HPC-based research in particle physics, astronomy and cosmology and nuclear physics. It is a key part of the infrastructure supporting the UK's Science and Technology Facilities Council (STFC) Frontier Science programme. Research scientists across the UK can apply for access to DiRAC's powerful computing facilities.

Four UK universities, Cambridge, Durham, Edinburgh & Leicester, are responsible for delivering DiRAC's HPC services. These universities provide core HPC facilities along with expertise to enable multiple research, support, knowledge transfer and industrial partnership projects.

Significant data considerations are at the heart of DiRAC's remit; petaflop compute and petabyte storage requirements are integral to DiRAC-supported projects.

DIRAC, SOME BACKGROUND

DiRAC's Memory Intensive facility in Durham has called on the services of Atempo, the Data Protection and Movement specialists, together with their UK partner, OCF, to implement a multi-petabyte archiving project for their Lustre and Spectrum Scale (GPFS) data.

We spoke to Dr Alastair Basden, Technical Lead for the DiRAC Memory Intensive service, in Durham University's Institute for Computational Cosmology (ICC). The DiRAC Memory Intensive service, the seventh increment in a series of HPC clusters at Durham, provides researchers with the computing power they need.

One of DiRAC@Durham's recent missions was to switch from Spectrum Scale (GPFS) storage to DDN's Lustre storage. The aim being to effectively enrich storage with a less expensive solution. The growing need for ever-higher memory intensive computing generates significant data volumes within their HPC environments.

Generated data storage by 2022 is projected to reach upwards of 20 petabytes with the roll out of DiRAC Phase 3. DiRAC's Data Management Plan includes an archival component for both the research database and finished peer-reviewed scientific research documents. It is the research data which requires archiving.

THE SOLUTION IN PRACTICE: ATEMPO MIRIA FOR ARCHIVING

The role of Miria is to archive research data and relieve higher cost disk storage by offloading DiRAC research data from primary storage to four LTO tape destinations. Current storage requirements stand at around 10 petabytes with an uplift to double that when DiRAC Phase 3 is rolled out. This phase will see a 10-fold increase in processing and a corresponding augmentation in data creation and storage requirements.

The Miria for Archiving User Interface allows users to perform their own rapid archiving tasks and to restore their data. Researchers can use the logical file tree format of their choice for each research project and preserve this tree structure when archiving their data. The Administration interface enables DiRAC to manage both archiving and backup of critical data.

The DiRAC technical teams at Durham have been impressed with the power of Miria (saturating an LTO8 drive with a single Miria Data Mover) with its rich feature set. "*Miria for Archiving is incredibly powerful and feature-rich and should meet our future needs*," remarks Alastair. "*Even though we've barely scratched the surface with the solution, we're very impressed so far.*"

"Miria for Archiving is incredibly powerful and feature-rich. We're very impressed so far. Archiving performance on Lustre file systems data flows is running at full tape speeds, which is perfect."

Dr Alastair Basden, Technical Lead for the DiRAC Memory Intensive Service, Durham University

Prior to Atempo, the incumbent archiving solution was slow and not as scalable as required. Atempo ran a Proof of Concept (POC) on Spectrum Scale GPFS first and then on Lustre. Atempo rapidly demonstrated that Miria could make short work of DiRAC's data archiving needs. A Miria Archiving Server and a dedicated Miria Data Mover directly access Lustre file systems and efficiently handle all data archive flows from source to destination.

The DiRAC technical teams at Durham are also using Miria to back up user home space along with archiving files selected by users. "*Every administrator and user action is handled through the HTTPS protocol which means it is*

very easy for us to set up an SSH tunnel and enable users to archive their files wherever they are in the world". All the physical data moving equipment along with the Miria server is installed and running and Alastair reports that: "*archiving performance on Lustre file systems data flows is running at full tape speeds, which is perfect*".

Archiving requirements include creating file data copies on two tape locations. The DiRAC service at Durham is using both LTO7m and LTO8 tapes. Atempo fully supports this technology mix.

LOOKING AHEAD

Miria is not just an archiving tool. The Miria for Backup component also protects user pool data. During the POC Atempo also demonstrated FastScan capabilities for GPFS. FastScan optimizes how new and modified files are recognized and stored to avoid lengthy file system rescanning. A key future integration for DiRAC will be to leverage Miria for its advanced backup capabilities with Atempo's FastScan capacity for Lustre.

The strength of Atempo's R&D teams lies in their in-depth industry knowledge of file systems, handling data attributes and striping options for example. DiRAC are keen to work alongside the OCF and Atempo teams for this integration and provide access to DiRAC's valuable beta testing on an HPC-scale file and storage environment. This is important for institutes such as DiRAC – the ability to inform stakeholders that they are on the forefront of new and innovative solutions in many domains including data movement and data orchestration.

The Atempo and DiRAC teams will continue to work in a spirit of cooperation and mutualized re-sources to build on this initial success. ■

REDUCING THE IMPACT OF IN-DEPTH TREE WALKING ON LUSTRE

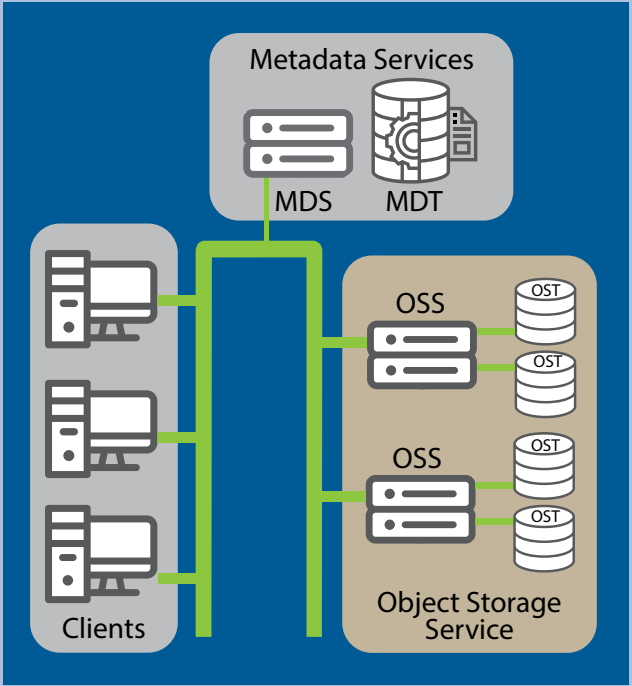
One of the big challenges of performing data management (backup, archiving, data migration) on a Lustre cluster consists in collecting the list of new and changed files. Many NAS platforms collect lists of new and changed files. These change lists can be leveraged by backup and archive solutions to speed up data movement.

When such service is not available, as is the case on Lustre, traditional data movement solutions go back to basics: they "walk" the filesystem ("tree walk") in order to identify changes. This procedure is not appropriate in a Lustre filesystem as is it a very lengthy process which will often exceed most backup windows. Backup or storage admins are then faced with making difficult choices for protecting or moving only a part of their data in the time slot available.

WHY IS THIS A COMPLEX PROCESS?

At one level, tree walking is a simple operation that consists of browsing the contents of a data structure, in this case the file system of a storage to collect the list of files and their attributes in each folder (inode). The sum of the list of files obtained is equal to the list of files in the storage.

The simple approach which recursively walks the tree hierarchy until reaching the leaf node, the actual files to be copied, is not as obvious as it may seem. Lustre is a distributed system made of multiple nodes that also distributes or splits data and metadata on several different components.



Components of a Lustre filesystem

Using parallelized processes does makes sense since multiple directories are explored

simultaneously. However, the process still has its challenges. The depth of the filesystem and the branching factors are important. In addition to having a complex tree structure, Lustre is a distributed system made of multiple nodes that stores metadata on different components. When a file is reached, its selection in the list of new and changed files is dependent on its associated meta data being collected from another component.

The above process creates a list with redundant information that requires a further consolidation step. In short, it's complex. And long.

Unsurprisingly, this process impacts the Lustre cluster performance and results in very lengthy operations sometimes lasting 3 to 4 weeks, just for the tree walking part! However, this is a necessary step to any smooth data movement progress. So, what solutions are there for a research institute or a media major storing 500 million files or more?

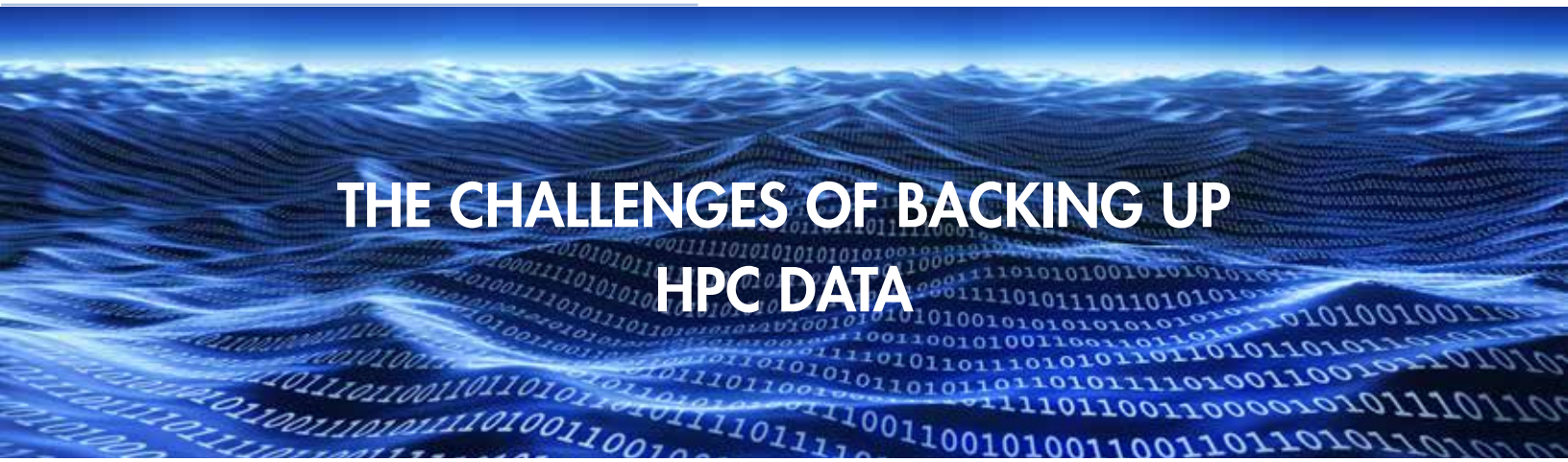
AVOIDING THE IMPACT OF TREE WALKING ON A LUSTRE CLUSTER

The idea developed by the Atempo team for Luster is simple: it consists in developing the core mechanism for collecting changes and offering the service to Miria Data Management components.

Miria solution is by design separating operations into two distinct streams: the change detection capability, and the data movement component – as this architecture delivers higher performances and more scalability.

The change detection mechanism (code name FastScan) is integrated with Lustre core components and nodes. The FastScan collects new, changed and deleted files at node level, enriched with meta data. The information is automatically consolidated in a database and available as a service call to the Miria server. For instance, when a backup or an archiving job is launched, the Miria Server gets the list of new and changed files from the FastScan DB, then it automatically split the list in multiple sub-tasks that are sent to several Miria Data Movers that are in charge of a part of the list. Data Movers adjusts the number of streams launched simultaneously in order to find the right balance to movement of files while preserving operational access to the file tree.

This architecture cuts the long wait induced by the filesystem tree walk and delivers an early start to actual data movement. Coupled with the high performance and scalability of Miria, it makes for a very efficient data management solution for distributed filesystems and petabytes of data.



THE CHALLENGES OF BACKING UP HPC DATA

MANAGING STORAGE COSTS AND NEW AND MODIFIED DATA

In some very high data volume environments –including HPC- backup has become increasingly complex or even impossible.

Above a certain volume threshold (around 100 TBs) or number of files (> 1m), backup windows become ever-tighter and full backups can take too long to complete. This is certainly the case for traditional NDMP backups.

HPC typically generates petascale volumes of data which needs to be managed at each storage tier. Long-term storage is usually to tape or cloud but this is typically for archiving purposes only. Storages areas such as the Projects or Home Spaces areas need more day-to-day protection in order to provide flexible and granular restore capabilities to their users as well as full disaster recovery coverage.

Storage solutions frequently propose using **snapshot** and/or **replication** as an alternative to backup. The former does not completely guarantee Disaster Recovery for large data sets, and the latter is very sensitive to cyberattacks which tend to replicate as fast as your data.

Having many backup versions with suitable retention periods consisting in full backups once a week and incremental on a daily basis, adds flexibility to data recovery.

With backup in data intensive environments such as HPC, there are two principal challenges: the first is managing and budgeting storage space. The second is the ability to detect and to protect new, modified and deleted files rapidly and securely.

FASTSCAN FOR LUSTRE

Lustre is one of those storages that does not come with built-in snapshotting capability, and collecting the list of new, changed and deleted files by doing a filesystem scan is not a realistic option (see also the section above: "Reducing the impact of in-depth tree walking on Lustre") .

The FastScan capability of Miria is collecting new, changed, and deleted files at node level, enriched with meta data. The information is automatically consolidated in a database and made available to a Miria server which then orchestrates data movement.

Many HPC compute results cannot be recreated or are very costly to run more than once. Backing up is a cost effective way of preserving essential data over time. Miria means Lustre backups to the storage destination of your choice are now possible.

FOCUS ON HPC DATA MANAGEMENT WORKFLOWS

BACKUP & ARCHIVE OF HPC STORAGES

Petabyte-scale volumes and billions of files should not make the backup and recovery of file data sets more complex! Here are the two top reasons why you should consider using Miria for Backup or Archive:

- If you have millions or even billions of files on a given storage (Dell/EMC Isilon, StorNext, Qumulo, Lustre, ...) and you need a flexible flexible restore solution for both daily incidents and Disaster Recovery.
- If using Lustre or a GPFS shared file system and you are unable or barely able to run a full storage backup.

MIGRATION OF LARGE-SCALE FILE SETS BETWEEN PLATFORMS

When considering the migration of massive file-based data sets (>100 TB to many PB) between different storage manufacturers or storage types, here are 3 typical situations for using Miria for Migration and getting your file storage migration project under control.

- migrate several hundred terabytes or even petabytes of file data from a large scale-out NAS to another storage in a different location,
- migrate your massive file sets from one storage to another type of storage (a different vendor or version for example),
- achieve fast data transfer for a large number of unstructured files in a limited period of time.

DATA MOVEMENT BETWEEN DIFFERENT STORAGES

Synchronizing full storages between different manufacturers (Isilon, GPFS, Lustre and more) is a frequent request in many different research centers. Here are 3 reasons to choose Atempo Miria to synchronize or move file-sets between petabyte-scale storages. You need to:

- keep a large number of files to keep in sync with big daily changes. On many platforms, Miria leverages Atempo's FastScan capability to quickly identify the list of changed & new files on the storage. No need to wait for days to start moving files.
- move files between different vendor storages. Data is collected on source storages and converted automatically to the right format on target storage. Storage list includes any NAS or file storage (CIFS/NFS), parallel and distributed filesystems such as Lustre or GPFS and many other including object storages and Cloud.
- Sync on-demand for one or more file storages.

Miria data synchronization solution offers four levels of sync, ranging from one-way replication with or without replication of deletions, selective one-way replication of a file subset, and full bi-directional replications.

About Atempo

Atempo is a leading independent European-based software vendor with an established global presence providing solutions to protect, store, move and recover all mission-critical data sets for thousands of companies worldwide. With over 25 years' experience in data protection, Atempo offers a complete range of proven solutions for physical and virtual servers' backup, workstations and migration between different storages of very large data volumes. Atempo's three flagship solutions, Lina, Miria and Tina are labeled "As used by French Armed Forces" and "France Cybersecurity".

Selected to join the initial selection of the "French Tech 120", a government program designed to nurture 25 unicorns by 2025, Atempo is headquartered in Paris and is present in Europe, the US and Asia with a partner network in excess of 100 partners, integrators and managed service providers.

Want to talk about your project?

Contact an Atempo expert

ATEMPO
preserving data ecosystems

For more information: www.atempo.com.