12-14 February 2018 | Rydges Hotel Queenstown
eResearch NZ
2018

Coping with "Big Data" in Research Computing

David Fellinger
iRODS Consortium
davef@renci.org

ABSTRACT

Initially much of the work accomplished in high performance computing (HPC) sites revolved around simulation or visualization. Compute clusters could allow researchers to develop environments which were difficult if not impossible to physically duplicate. For the first time Schrodinger's equation could be solved in multiple dimensions giving new insights into complex quantum mechanical systems. Researchers could simulate and understand nuclear reactions at a level that allowed the complete elimination of physical nuclear testing. The entire process of radioactive decay could be simulated at a scale and over time periods that could not have been physically duplicated.

HPC also had a significant impact in design methodology. Aeronautical work that could have only been accomplished in large wind tunnels could now be completed in efficient compute clusters. The dynamics of an entire aircraft could be simulated including situations that would be impossible or damaging to physically execute. It can be argued that parallel computing has had an impact on the scientific method where multiple hypotheses can be checked and re-proposed while a simulation is being run vastly shortening the time to discovery.

In the beginning of the 21st century we saw the strong shift of collected data migrate from analogue to digital storage. Film, audio tape, and video tape disappeared in a very short period and forms of digital collaboration launched a literal data explosion. Data collection methods also evolved enabled by inexpensive network bandwidth and storage. Experiments like the Large Hadron Collider and large scale radio telescope arrays produced huge data sets. The study of climate change drove a worldwide growth of sensors of various types to record real time data. Finally, the "internet of things" produced huge data sets ranging from automotive diagnostics to a proliferation of automated video and other surveillance data. The term "big data" was coined relating to growing collections having the properties of Volume, Velocity, Variety, Variability, and finally Veracity implying verification of provenance. The analysis of these data collections launched a new era in high performance computing primarily devoted to data reduction and analysis. This new paradigm had a major impact on cluster requirements. Simulation environments involved loading a parallel process then collecting output data on a "scratch" file system. At times, checkpoints were also written during a long process to the file system so that a failure of a system component or a code crash would not necessitate a complete start from the beginning of the process. In the data analysis scenario, the bandwidth requirements of the input/output (I/O) infrastructure more than doubled. Large amounts of data had to be loaded onto machine cache or the scratch file system before the analysis could

start. Requirements of metadata tracking became critical because some analytical processes required comparisons of data which had variations of time, spatial, or other characteristics. Finally, the resource requirements grew because data was handled both onto and off of the compute cluster in a largely manual process. Parallel file systems like Lustre were developed which helped with the bandwidth problem so that the scratch file system could effectively act as a storage cluster which was able to increase in throughput based upon the number of parallel servers in the system. System schedulers like the Slurm Workload Manager enabled efficient machine usage by logically apportioning cluster components for specific operations. While efficient file system and scheduling technology improved the process to I/O flops ratio, moving large data sets to and from the cluster was still largely a manual task.

Automating the workflow

The Integrated Rule-Oriented Data System (iRODS) was launched about 10 years ago as an open source project dedicated to enabling the establishment and maintenance of research collections. A consortium was formed 4 years ago to continue this development with community support and a dedicated group of developers. This middleware is currently in use in large data repositories like EUDAT in the European Union and the Wellcome Trust Sanger Institute in the UK. Through the use of iRODS, system administrators can completely maintain all phases of data collection, metadata cataloging, and storage management through the application of mission specific rules applied to the entire collection. Metadata can be collected upon ingestion to build a catalog and access controls can be applied based upon security requirements all under iRODS management. Data integrity can be checked and a chain of ownership and custody can be maintained and audited. Recently, the iRODS Consortium started a program of building the tools necessary to apply data management in cluster computing. The thought process was straightforward, "If iRODS can be utilized by research teams for large scale data management why not consider the compute cluster effectively a team that can both ingest and publish at very high rates?"

The first HPC file system to be closely integrated with iRODS is Lustre which is an open source product maintained by OpenSFS in the US. This file system is used in many HPC centers worldwide and is noted for its scalability. Lustre is a distributed file system with the metadata co-located in a MetaData Server (MDS). This server maintains knowledge of the file names and related macro data placement locations but does not have the burden of placing the data on individual storage devices. Object storage servers (OSSs) have the burden of data placement but many of them can be used in parallel so the overall data placement and recovery is very efficient. The iRODS interface consists of a service node which operates effectively in parallel with the MDS updating an internal database to be consistent with a changelog produced by the MDS. It is important to note that iRODS monitors the data infrastructure and placement but is not in the data path which is critical to compute cluster performance. The iRODS database termed the iCAT contains the same knowledge as the MDS but can contain additional descriptive metadata that is critical to researchers who must search the computer cluster output results.

An HPC job scheduler may also be linked to iRODS so that complete workflow automation can be maintained based upon rules written for the collection and the process. As an example, oil discovery seismic study data may be available for a specific region and the data reduction process has been scheduled. Data can be moved by iRODS to the parallel scratch file system at the appropriate time for processing and then the resultant files can be moved to another file system for distribution. Distribution to users or subscribers is automated based on iRODS rules which are used to maintain data usage policies. Metadata can be extracted from headers while the files are being moved and the iCAT can be updated with the critical file content information and the data placement data. The scratch storage system can then be purged and reloaded with the next job in the queue. Researchers and system administrators can be notified as each step of the process is completed and all of this can be done without administration intervention. Finally, iRODS can also replicate either data, metadata, or both to other sites for collaborative study or disaster recovery purposes

The end result is a system dedicated to minimizing decision latency at every step of ingestion, processing, and distribution to end users based entirely on a flexible rule set that can be updated as requirements change. Each step of this process can be audited to meet the most stringent requirements of data management. The key to any research program is to provide data to researchers in a timely manner. The iRODS HPC tools have been developed to realize that goal by relieving administrators from burdensome data management tasks.

ABOUT THE AUTHOR

Dave Fellinger is a Data Management Technologist with the iRODS Consortium. He has over three decades of engineering and research experience including film systems, video processing devices, ASIC design and development, GaAs semiconductor manufacture, RAID and storage systems, and file systems. As Chief Scientist of DataDirect Networks, Inc. he focused on building an intellectual property portfolio and representing the company to conferences with a storage focus worldwide.

In his role with the iRODS Consortium Dave is working with users in research sites to assure that the functions and features of iRODS enable fully automated data management through data ingestion, security, maintenance, and distribution. He serves on the External Advisory Board of the DataNet Federation Consortium and was a member of the founding board of the iRODS Consortium.

He attended Carnegie-Mellon University and hold patents in diverse areas of technology.