On-Demand Data Analytics and Storage for Extreme-Scale Simulations and Experiments

Franck Cappello[1,2], Katrin Heitmann[1], Gabrielle Allen[2], William Gropp[2], Salman Habib[1], Ed Seidel[2], Brandon George[4], Brett Bode[2], Tim Boerner[2], Maxine D. Brown[3], Michelle Butler[2], Randal L Butler[2], Kenton G. McHenry[2], Athol J Kemball[2], Rajkumar Kettimuthu[1], Ravi Madduri[1], Alex Parga[2], Roberto R. Sisneros[2], Corby B. Schmitz[1], Sean R Stevens[2], Matthew J Turk[2], Tom Uram[1], David Wheeler[2], Michael J. Wilde[1], Justin M. Wozniak[1].

[1]Argonne National Laboratory, [2]NCSA, [3]UIC, [4]DDN

This white paper presents a concept covering two elements of a plausible pathway towards the convergence of high performance computing and large-scale data analysis:

- Unifying vision for the computational and data-related processes as a foundation for convergence
- Common ecosystem where a data center and an HPC center are seamlessly coupled

Motivations

Data size and throughput is becoming one of the main limiting factors of extreme-scale simulations and experiments. With current system computational capabilities, extreme-scale scientific simulations and experiments can generate much more data than can be stored at a single site. The scientific community needs multi-level data access to perform complex and accurate analyses, while avoiding severe data reduction methods such as filtering and extrapolation. Also, often a single site cannot simultaneously satisfy both computing and data analytics requirements. Finally, extreme-scale simulations and experiments tend to push toward a model where a group runs simulations that are then analyzed by many other groups (the one-simulation, many-groups model), in contrast to one group running their own simulations and data analytics (the one-simulation, one-group model).

Use Case

The extreme-scale example considered here is taken from computational cosmology. The science requirements demand running very large simulations, such as N-body runs with trillions of particles. The resulting data products are scientifically very rich and of interest to many research groups. It is therefore very desirable that the data be made broadly available. However, as a fiducial example, a trillion-particle simulation with the HACC code generates 20 PB of raw data (40 TB per snapshot and 500 snapshots), which is is more than petascale systems such as Mira and Blue Waters can store for a single run in their file systems. An interesting point is that while one version of HACC is optimized for Mira and can scale to multi-millions of cores, Blue Waters offers exceptional data analytics capabilities with its thousands of GPUs. This suggests a combined infrastructure based on using Mira for the simulation, Blue Waters for a first-level data analysis, and a separate, possibly distributed, data center to store the distilled results. Users from other universities and labs would then pull the data from the data center and run further data analytics locally following their particular scientific interests.

On-Demand Data Analysis and Storage Concept

Other communities have analogous needs, and some have already put in place infrastructure to respond to them. One example is the high energy physics (HEP) community with the Large Hadron Collider and its dedicated data analytics infrastructure, where the data produced by the experiments is sent to remote distributed storage and the data analytics is performed on other servers. Another example is the genomics community, which has its own data banks. However, the associated infrastructure in these cases is dedicated (not shared by other large-scale experiments), specific (responds to specific data distribution needs), and permanent (resources and policies are in place for 24/7 analysis needs).

The concept we propose in this white paper is different in several ways. We consider the deployment of an *elastic virtual infrastructure* connecting several data production sites (comprising both simulation and experiment), data centers for storage, and data analysis centers. The infrastructure set-up is not permanent; it is dynamically instantiated on demand for transient needs (even if the data is stored on a long-term basis in the data centers). The same resources (data production side, data centers, and analytics centers) could be used or shared by other scientific communities instantiating infrastructure components, potentially aggregating other resources as well.

Our concept relies on technologies needed for the convergence of HPC and big data, such as virtualized resources, resource reservation, software deployment, user group management, policy management, etc. It combines some grid principles by aggregating geographically distributed resources owned by different institutions, as well as some cloud principles: Infrastructure as a Service, elasticity, and virtualization.

The SC16 Experiment

In order to demonstrate the feasibility of this concept, a team of about 15 researchers and staff from Argonne, UIUC, UIC, DDN, and including SCinet leaders, is preparing an experiment (Figure 1) which will be demonstrated at the 2016 International Conference for High Performance Computing, Networking, Storage and Analysis (SC16) in Salt Lake City. This experiment and demonstration covers several critical elements of on-demand data analytics and storage.

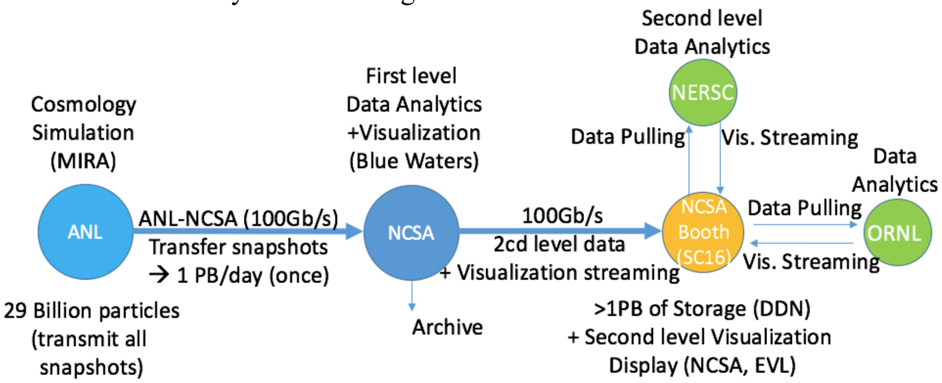

Figure 1: The SC16 on-demand data analytics and storage experiment

During the experiment a state-of-the-art cosmology simulation combining high spatial and temporal resolution in a large cosmological volume will be performed on Mira with 29 billion particles. All time snapshots (500) of the simulation (L1 data) will then be transmitted, as they are produced, to NCSA using Globus. Note that in previous simulations, users were able to analyze only ~100 snapshots because of infrastructure limitations. A first level of data analytics (L2 data) and visualization will be performed as snapshots arrive at NCSA using the GPU partition of Blue Waters. Blue Waters will also archive the L1 data on tape. The L2 data (1/2 the size of L1 data) will be sent immediately as produced to a DDN storage server (playing the role of a data center), hosted at the SC16 NCSA booth. At least two sites (NERSC and ORNL) will play the role of data analytic centers. They will pull and analyze the L2 data as it is stored on the storage server and perform specific and different data analytics using the PDACS framework. The goal is to demonstrate that L2 data can be used by several cosmology research groups. The data analysis results will be streamed in ultra-high-resolution using the SAGE2 software to the SC16 NCSA booth on the show floor. The whole experiment will be orchestrated by Swift.

This experiment will achieve three objectives never accomplished before: (1) run a state-of-the-art cosmology simulation and store or communicate all snapshots (currently only 1 snapshot of 5 or 10 is stored or communicated), (2) transmit 1 petabyte/day of raw data between two distant leading computing facilities (ANL and NCSA), and (3) perform multiple different extreme-scale data analyses from data stored on a remote data center.

Making the On-Demand Data Analytic and Storage Concept a Reality

The SC16 experiment will stress the simulation, communication, storage, and analysis infrastructure, and the orchestration and streaming software of the on-demand data analytics and storage concept. Several key elements of the concept will remain to be tested, including the software for resource virtualization, deployment, group management, and policy management. In particular, the software stack run by users is static in the SC16 experiment. In a more complete implementation, software will be deployed as the virtual infrastructure is instantiated.