Usability Investigations for High Energy Physics Analysis

Matt Crawford, Phil Demar, Dave Dykstra, Gabriele Garzoglio, Tanya Levshina, Ruth Pordes

October 1st 2010

Research Topic

Our applied research topic is to investigate and make recommendations for the integration and production use of the advanced network (100Gbps) fabric for the event simulation and analysis applications of physics experiments. Our methodology is to test, interpret and document the end-to-end usability and capabilities of the system.

The value proposition of investing in this applied research upfront is that the experiments can then take full advantage of, and will have already made the necessary adaptations for, 100 Gpbs as soon as it becomes available to them. Experiment application research-cases allow us to investigate the end-to-end system. For example to look in detail at the interactions and effects of the data manipulations in the multiple software layers, including effects of multiple layers of data caching, policies and I/O and the potential for disruptive "beating" across the layers.

Fermilab does applied research (development and support) to provide common services and systems for its scientific user communities across the Energy, Intensity and Cosmic Frontiers, including the Tevatron Run II, US CMS, Dark Energy Survey, neutrino and dark matter experiments. All collaborations are widely distributed with physicists in geographically dispersed groups needing access to datasets for analysis - where the dataset sizes and access rates vary significantly up to a Petabyte a week. A continuous, real-time and adaptive pipe of data from the acquisition system to nationally, and internationally, shared multi-stage analysis workspaces continues to be the nirvana.

We will study the data analysis patterns, error recovery scenarios and caching strategies generated and needed by the user application software. We will make parameter sweeps of the configurations of the user and OS driver-level interfaces and measure the robustness and performance of the end-to-end system. We will pay attention to integrity of the data, response to fault and error propagation, and to software optimization, comparing to the systems using the standard production networks.

Test plan

We will deploy experiment data servers, populate them with existing event and/or simulated data sets of, preferably, up to several 10s of TeraBytes. We will install the experiment services and applications. We will have scripts to simulate access by 10s of users. We will investigate the following end-to-end analysis scenarios:

1) CDF, D0 and MINOS application testing accessing data cached in various storage software back ends over the ANI testbed. We are interested in particular in the HDFS, dCache, and Lustre storage solutions. Initially we will deploy the storage software on the I/O nodes and applications on the host nodes of the testbed. We will transfer physics data on the storage servers and evaluate the efficiency and overall effectiveness of the new network towards regular data movement. Once the data is preinstalled, we will run I/O-bound applications to investigate the characteristics of the saturation points of each layer in the stack, now that network will not be the limiting factor in processing speed. The use case of event skimming is particularly relevant as applying simple event selection criteria does not involve any significant usage of CPU cycles, resulting in operations typically limited by the I/O capabilities of the system. Applications relying on ROOT for their I/O will be able to access data through the network in two access patterns: 1) by quickly skipping (seek) files to event header information and 2) by loading events into memory for non-trivial selection criteria. In particular the latter access pattern requires the transferring of all the event data into the client context in a non-CPU bound regime. We are in particular interested in evaluating the performance of the Lustre storage server in the configuration where only a few clients can already saturate the I/O bandwidth to storage. Typically, in fact, in such topology, the first layer to saturate is the network interface, rather than the storage software or the back-end storage hardware itself. A network connectivity of 100 Gbps is the ideal transport capability to probe the limits of the storage and application layers with the few-clients configuration.

2) CMS Analysis - running at least the data transfer portion of a sample event analysis accessing data over a Lustre distributed file system. This analysis will be split up into enough parallel pieces to take advantage of the high-speed network. The software and all data will be stored on Lustre servers on one side of the testbed and the computation will be on the Lustre client side. Normally the CMS software (~2.5K files are used totaling ~100MB) and the "conditions" data (~150MB) are shared for all the clients so they are either installed or cached on the client LAN, and the event data files (~2GB each) are copied ahead of time to the client LAN. In order to put a higher load on the network we plan to transfer all of it from the server to the client on demand. We will measure the performance in events read per unit of time, compare it to having everything on a local area network, and match that up with backbone network utilization plots. We will test both standard Lustre and the emerging Wide Area Lustre. We will repeat the tests with the HDFS file system, already used by US CMS Tier-2s.

3) CMS XROOTD Demonstrator, with the help of the Fermilab Tier-1 facility. Initial versions are already deployed at several smaller US CMS Tier-3 sites as well as in test mode at Fermilab. We will deploy and test the available versions on the testbed. There are other new application middleware services being proposed for the next phase of the CMS experiment. Once these have been ratified by the experiment (expected in 2011) we will be adding these to our regular testing program.

Testbed components used

1) End to end networking components. 2) Monitoring Hosts: for running the test analysis displays through our local monitoring and diagnostic systems. 3) Application Hosts: for running the experiment application and client software. 4) I/O test Hosts: to install the experiment data sets and storage implementations.

We will repeat all the tests with services based at Fermilab. Where feasible we will configure a network path that stretches to the ANI testbed components. To do this will require the addition of router equipment between Fermilab and ANL. Until the 100Gbps (Wide Area) Network endpoint is locally available we will aggregate multiple 10Gbps pipes to perform the tests as equivalently as possible.

Qualifications of the team

The technical work will be done by Dave Dykstra who is experienced in CMS applications, develops and supports the CMS-wide (and now ATLAS-wide) calibration information through a set of distributed Squid services (the Frontier software). We will hire a summer student to work with Dave. Other members of the team are leaders in the Computing Division's scientific facilities, storage and networking groups, as well as contributors to the US CMS distributed facilities.

Time and access requirements

We request 4 months of test time starting in, or soon after, May 2011. This will come at an appropriate time in existing development activities and Dave Dykstra's availability.

Reporting

We will provide regular updates in document form (modeled on the report at https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=3532). We will present the results at CMS, ESCC, ESNet, OSG and other technical meetings. We will publish the results in conferences such as ACAT, CHEP, HPDC, SC.