

## **Input to Advanced Network and Middleware R&D Programs**

Jon Bakken, Phil Demar, Gabriele Garzoglio, Ian Fisk, Ruth Pordes, Panagiotis Spentzouris, Wenji Wu,  
Fermilab, February 2011

### **Scientific Program:**

The Fermilab HEP program supports data intensive experimental science, with currently more than 28 Petabytes of data on tape and 400 TBs data distribution per week with decade long analysis needs as well as an aggressive program of future developments for the accelerator, experimental, theoretical and astrophysics programs.

Two of Fermilab's scientific programs stand to gain the most value from and have the most challenging needs of 100 Gigabit and Terabit scale wide area networks. First is the US CMS global analysis of 10s of petabytes of data annually by more than 300 physicists in the U.S and more than a thousand worldwide; and second is the near- or real-time control and tuning of accelerators based on the results of large-scale multi-parameter accelerator modeling and simulation programs performed on the nations peta-, tera- and in the future exa- scale high performance computers.

CMS analysis of data acquired from the LHC is already involving an increasing number of the ~100 Tier-3 sites, the majority being in the US, where the science faculty together with their post-docs and students work locally as well as jointly with similar groups remotely on a common physics topic. These analysis ultimately depend on data distribution from the 7 CMS Tier-1 centers, of which the Fermilab Tier-1 is the largest and most performant. CMS depends on predictable and efficient bulk data movement between collaboration sites. Thus at Fermilab, the CMS analysis program imposes challenging requirements on the local area networking as well as the wide area network capabilities.

CMS is at the beginning of a major, and innovative, transition in its computing model, towards dynamic – rather than static – placement of the data sets (typically 30-100 Terabyte in size) needed for any particular analysis and from local storage (ie relying on the LAN) to wide area access (ie relying on the WAN) real time data access by the applications themselves. Clearly, both these transitions will make increased demands on the network capabilities and capacities – not only in terms of raw potential throughput but also for how the aggregate usage is managed and allocated and for the integrated and embedded nature of the dependence on the network resource for all computations.

Accelerator science is at a point of transformation in terms of being able to contemplate the use of peta- and tera-scale HPC facilities to generate results from the modeling programs with sufficient accuracy and fast enough turn-around to be able to influence real time steering of accelerators based on the results. As the application of particle accelerators in the service of medicine, industry, energy, the

environment, national security, and discovery science advances<sup>1</sup>, the benefit from such capabilities will increase. Such dynamic steering of operating accelerators requires data mining and interactive exploration of terabytes of simulated data sets in the control room itself – whereas the state of the art today is that only high level and synthesized results of the simulations are communicated back to the scientists and engineers.

Exascale computing will provide the opportunity for even larger simulations. It will enable full scale complete system ab-initio models of each accelerator and, potentially, real-time steering and control by making the results of the simulations rapidly available at the research facility itself.

### **Research and Development Needs:**

The challenge in the use of 100 Gigabit and Terabit networks is not only to ensure the raw end-to-end performance of the data transfer from source to sink, but also to provide network layer intelligence, operating system, middleware and applications software, that work together to ensure a manageable, usable, and robust system that effectively handles the diverse end-to-end network environments, the volatile network conditions, the many simultaneous and diverse users and applications, and the appropriate security and access controls.

#### *General R&D needs:*

- Network, host interface, OS, and storage layer software that correctly, efficiently, and at rate transfers data from advanced networks.
- Host interface, OS, storage layer software, and middleware that scale to 100 Gigabit and Terabit networks not only in terms of speed, but also in terms of the many number of parallel data streams.
- Information provision frameworks that provide information from all entities and interfaces in the system.
- Error detection and automated correction techniques – end-to-end - leading to enhanced fault tolerance and fast recovery capability.

#### *CMS specific R&D needs:*

- Network aware middleware that supports the globally distributed ensemble of sites to be treated and used as an integrated “local analysis system” and less of a set of independent resources.
  - Catalog Software that accurately keeps track of the whereabouts of data. The software must be distributed in nature and is scalable and extensible to many sites and many data
  - Fast-response query systems to quickly locate and answer data location queries.
- Adaptation of applications to respond to historical conditions in the planning of their data and job requests.

---

<sup>1</sup> <http://www.acceleratorsamerica.org/>

- Co-management of a complex set of resources to make more efficient use of tape, disk, network in a system of constraints.
- Methods and algorithms that analyse historical usage patterns, including trends and anomalies, to predict capabilities and capacities.
- Methods and algorithms that analyses network status and provide end-to-end network performance troubleshooting capability.
- Innovative I/O architectures and methods to make network I/O faster and more efficient.
- New task scheduling and assignment strategies and methods in middle ware to accelerate data processing.
- New task scheduling methods in OS to accelerate data processing.
- Approaches that most effectively integrate across the diverse network environments encountered in the end-to-end system.
- Understand how to scale "global task queues" more effectively and automatically - avoiding single point of failure, bottlenecks in optimizations etc.
- Wide-area data delivery and replica management.
- End to end scheduling and monitoring from source disk byte to sink disk byte
- Integration of the campus level (Tier-3) use, service and resource identity information with the global identity and access control environment.

*For near-line accelerator control based on modeling research and development needs include:*

- Guaranteed and reservable level of network service.
- Protected, auditable, and defensible end-to-end data delivery environments.
- Support for data set distribution and management from HPC (exascale computers) of 100 TB per run, with parameter optimizations achieved through multiple (100s-1000s) of runs.

*Operation and production system needs for applied research and development:*

Applied research and development in system design and engineering, software interfaces and capabilities, system and data center management and operation, are also critical needs in order to bring to fruition robust production systems that actually enable the innovations from research to benefit large-scale science. The needs include:

- Provision, analysis and communication (alarming) to the application layers of comprehensive, end-to-end, configurable, monitoring information (available on demand from every network box and host port, every interface, every piece of software). Summary of values outside nominal ranges, including error counters, IO rates, bonded IO rates; Statistics like the top resource "talkers".
- Understandable, synthesized, well crafted human interfaces for presentation and control of information –e.g. weathermap plots for all resources and I/O interfaces.
- Ability to combine information from diverse resource types – e.g. network information with information from data storage and management systems.

- Comprehensive, consistent, easy to deploy, use and interpret real-time testing tools and software.
- Software and algorithms in all services to ensure protection and throttling as needed to ensure the robustness of the system.