Fermilab Status Update
ICFA SCIC 2013 Report
Phil DeMar
January 26, 2013

Fermilab’s major network changes in 2012 were focused on planning and initial deployment of 100GE technology. Our efforts were largely directed at aligning the Laboratory’s WAN infrastructure with the rollout of ESnet5, ESnet’s 100GE network. In addition, the Laboratory upgraded its network/middleware R&D test bed facilities to provide 40GE and 100GE support as well. Upgrades in data center networking were relatively modest, with the emphasis being on sustaining stable operations for the CMS Tier-1 and Run-II networks. IPv6 was enabled for Laboratory DNS, e-mail, and central web services.

100GE WAN Capabilities

2012 marked a preparatory year for the Laboratory in terms of 100GE network technology deployment. The primary emphasis has been on upgrading the network perimeter (WAN) infrastructure to accommodate 100GE access to ESnet. With ESnet completing the deployment of its 100GE-based ESnet5 backbone in late 2012, the implementation of 100GE access for the Laboratory to that backbone is now proceeding. As with the Laboratory’s current offsite access infrastructure, a diverse, fully redundant metropolitan area network (MAN), called ChiExpress, is being deployed to serve both Fermilab and nearby Argonne National Laboratory. The initial implementation of the ChiExpress MAN will provide the Laboratory with one 100GE channel and three 10GE channels. The aggregate of 130Gb/s in total bandwidth represents a 60% increase in the Laboratory’s WAN bandwidth capacity over our current MAN infrastructure (8x10GE). Even more significant, the ChiExpress MAN is a WAN access infrastructure that provides a scalable and moderate cost path for WAN capacity upgrades in 100GE increments.

The 100GE channel currently being deployed will terminate at StarLight, and support the Laboratory’s off-site high impact science data movement. This traffic is largely carried across data circuits today, and traverses a separate network path from the Laboratory’s general internet traffic. The three 10GE ChiExpress channels will be used for that general internet traffic, providing full redundancy for it. The Laboratory has historically utilized a separate border router for its data circuit traffic. In 2012, a new, 100GE-capable border router (Nexus 7000) was procured and deployed to support the 100GE ChiExpress channel that ESnet is providing to the Laboratory. The configuration of the Laboratory’s off-site access infrastructure, once ChiExpress is fully deployed is shown in figure 1. Note that when completed, there will be two, geographically separate
ChiExpress hubs at the Laboratory in order to preserve full geographic diversity for the our WAN infrastructure and ensure highly resilient internet access. The transition to production use of ChiExpress by the Laboratory is expected in the 1st quarter of 2013.

![Diagram of ChiExpress network](image)

**Figure 1:** Fermilab Offsite Access via ChiExpress (deployment in progress)

A number of US CMS Tier-2 sites are planning to implement comparable 100GE network infrastructure, and several have inquired about the feasibility of achieving data movement at significantly higher rates than the 10Gb/s level. This will be an area of investigation in 2013 once the ChiExpress deployment is completed, and comparable 100GE capabilities exist at collaborating sites.

**Attachment to LHCONE**

In 2012, the Laboratory connected to the LHC Open Network Environment (LHCONE) in order to provide enhanced access to our Tier-1 data for Tier-2 sites, principally those located in Europe. An additional 10GE channel on the existing MAN infrastructure was brought into production use to accommodate the LHCONE traffic. Utilization levels routinely run at 2-4Gb/s, with peaks approaching 10Gb/s.

**Development of 100GE-Capable R&D Test Environment**

The Laboratory’s 100GE ChiExpress channel will serve a dual role. As previously mentioned, it will support the Laboratory’s off-site high impact science data movement, principally for CMS. The current bandwidth allocation for that type of data movement is 5x10GE, or 50Gb/s. With the LHC in shutdown for two years, this is considered to be more than adequate. The Laboratory intends to utilize the remaining bandwidth on the 100GE channel to support its network and middleware R&E activities. These activities
include connecting into the ESnet 100GE test bed network, as well as other R&D collaborations with wide-area data movement components. Data circuits will be employed for these R&D collaborations, in the same manner they are used for high impact science data movement. A minimum of 50Gb/s is assumed to be available for this R&D activity. Depending on production traffic needs, particularly with LHC in shutdown, significantly more than 50Gb/s may be made available for our network and middleware R&D efforts when those activities would benefit from additional bandwidth. Figure 2 depicts the dual use of the 100GE channel for production and R&D uses.

![Figure 2: Allocation of Fermilab 100GE Channel between Production and R&D](image)

The part of the 100GE ChiExpress channel allocated to network and middleware R&D represents only one component in the Laboratory’s network R&D test environment. In addition, the Laboratory has upgraded its local R&D network test bed infrastructure to a 100GE capability. A Nexus 7000 with one 100GE module and one 40GE module has been procured, and now serves as the core device for the network R&D test bed. 10GE-connected test systems have been migrated over to the Nexus, and two PCI Gen-3 systems equipped with 40GE NICs have been attached as well. The Nexus has a 100GE connection for wide-area access to the border router supporting the 100GE ChiExpress channel. The Catalyst 6509 that formerly served as the core network test bed device has been retained to support legacy 1GE-connected systems. Figure 3 shows the new 100GE-based local network test bed.
Figure 3: Fermilab 100GE-capable network test bed

It is worth noting that the local network test bed is a dual use facility, supporting network/middleware R&D activities, as well as providing an evaluation environment for new switch modules & software. With the Nexus 7000 emerging as the central platform within the Laboratory’s data centers, the latter function becomes increasingly important to stable operations. One example of this type of evaluation is 10GBase-T, which we expect to be deploying and evaluating in the test bed in the 1st quarter of 2013.

Data Center Networks:

In 2012, the emphasis for Laboratory data center networks was on stable operations, not infrastructure upgrades. With the Tevatron permanently shut down, and the LHC facing a lengthy shutdown of its own, upgrades were essentially limited to adding 10GE channels to ether-channeled links where traffic levels required them. New servers procured for CMS in 2012 were attached at 10GE (fiber). Evaluation of Fabric Path as a core data center interconnection technology was conducted in 2012, with deployment in early 2013 likely for the general data center network infrastructure.

Despite the lack of significant infrastructure upgrades in 2012, both the CMS Tier-1 and Run-2 data center networks handled increased traffic levels without instigating performance issues. Figure 4 depicts the CMS Tier-1 on a representative high-traffic day. Spread out over four separate computer rooms, The CMS Tier-1 LAN routinely fills most of 8x10GE ether-channeled links to 6509 access switches.
The aggregate traffic load sustained within the CMS Tier-1 LAN is probably more revealing. As shown in figure 5, those traffic levels reached ~170Gb/s on a frequent basis in 2012. In 2011, those peaks had been at the ~150Gb/s levels. The aggregated peak increase was due to improved system performance characteristics, not upgrades to the CMS Tier-1 network. The Tier-1’s network infrastructure is capable of handling even higher aggregate traffic loads.

Traffic graphs for the Run-II network tells a similar story. Like the CMS Tier-1, the Run-II network is spread across four separate computer rooms. Distribution switches within each computer room are interconnected at 8x10GE. In 2012, traffic levels on those links pushed towards 80Gb/s, as shown in Figure 6.
IPv6 Support:

In 2012, the Laboratory’s made significant progress in deployment of production IPv6 services. In conjunction with a US Government directive toward IPv6 support within Federal Agencies, the Laboratory successfully enabled IPv6 support for its DNS, e-mail, and central web services. Unlike most other National Labs and Federal Agencies, Fermilab’s IPv6 support extended beyond services deployed on the network perimeter, and into the core infrastructure of the campus network. This positions the Laboratory to be able to support IPv6 more widely in the future. In 2013-2014, Fermilab expects to be deploying IPv6 within select user networks as an initial step toward wide scale campus deployment. In addition, there is an active Laboratory effort to participate in a HEPix-based inter-site IPv6 test bed with an objective to evaluate IPv6 within the HEP application environment.