Update on Network Requirements and Computing Model R&D for the HL-LHC Era

H. Newman, February 2020

Network Requirements and a New Network-Aware Computing Model for the HL-LHC Era

Summary: Recent estimates of network capacity requirements for the HL-LHC era indicate that these cannot be met through technology evolution and price/performance improvements alone within a constant budget. An in-depth consideration of HL-LHC Computing Model is thus needed, and an R&D program to formulate, design and prototyping of the new Model is recommended. This program could take advantage of current development projects that provide the capability to set up, allocate and end-to-end network paths with bandwidth guarantees, and to coordinate the use of network resources with computing and storage resources.

2020 Update in the Outlook for Network Requirements:

In January, at the 43rd LHCOPN/LHCONE meeting at CERN[1], the LHC experiments expressed the need for Terabit/sec links by the start of HL-LHC operations in 2027-28, preceded by the usual Computing and Storage (and Network) challenges starting during LHC Run3 (2021-4). This was reinforced by the requirements presented by the DOMA project[2] which *"foresees requiring 1 Tbps links by HL-LHC (ballpark) to support WLCG needs. This is for the network backbones and larger sites…"*

The quoted network capacity requirements are an order of magnitude greater than what is available now through the present national and transoceanic networks based on 100GE links. As discussed at the LHCONE meeting, in the GNA-G Leadership group meeting that followed, and in the HEPIX Techwatch technology tracking group, these requirements cannot be accommodated solely through the exploitation of technology evolution within a constant budget. As a result, the further development of managed end-to-end services for the LHC and other science programs and the associated plans presented in this note, could be of pivotal importance. Work in this direction should also be guided by DOMA statements that *"caching/latency hiding will be important. DOMA is exploring XCache as a mechanism, which provides latency hiding and support for diskless sites (with regional data lakes). Production of AODs (using RAW) will be a network driver, especially regionally. Effectively the 'site' is expanded to encompass a 'region'."*

- It was agreed in subsequent discussions that the HEPIX Technology Watch WG and/or the Global Network Advancement (GNA-G) leadership group that was formed in the fall of 2019[3], can help define how much of it can be satisfied through technology evolution by 2027, and by 2024 in the preparatory phase.
- The rest will involve a change in paradigm including the end-to-end services involving sites and networks, and orchestration, as is being developed in projects such as SENSE, SANDIE and NOTED (described below). Ongoing discussions should continue to conceptualize and define what the new class(es) of service required entail.
- An important part of this is the persistent testbed being deployed by SENSE in collaboration with AutoGOLE and other collaborating projects. This is proceeding starting with the current SENSE testbed sites, plus extensions to UCSD, CERN, Starlight in Chicago, and a few other sites in the US and overseas.

[1] https://indico.cern.ch/event/828520/. See E. Martelli, S. McKee LHCOPN-LHCONE Report to the Grid Deployment Board,

[2] See the DOMA project requirements presented at the LHCOPN/LHCONE meeting: https://indico.cern.ch/event/828520/contributions/3570904/attachments/1968554/3274036/LHCONE-DOMA-01-2020.pdf

[3] The GNA-G is an open volunteer group devoted to developing the blueprint to make using the Global R&E networks both simpler and more effective, operating under GNA-G. The primary mission of the Global Network Advancement Group (GNA-G) is to support global research and education using the technology, infrastructures and investments of its participants. See https://www.dropbox.com/s/qsh2vn00f6n247a/GNA-G%20Meeting%20slides%20-%20TechEX19%20v0.8.pptx?dl=0

- The testbed will first be put it in place starting with a number of key locations and then having it grow organically. Its operation would start with SENSE and AutoGOLE services in a shared setting with QoS (providing some bandwidth guarantees in allocating network resources), then moving later to a set of dedicated links (perhaps scheduled) in order to have some ongoing developments and tests "at scale".

Network R&D projects:

- The development of the HL-LHC Computing Model can leverage ongoing work in the following R&D projects (as examples):

  1. Work underway in the SENSE[4] (Software-defined network for End-to-end Networked Science at Exascale) project, building intelligent end-to-end deterministic and policy-guided network services coordinated with the end-site computing and storage resources, where Caltech has the primary responsibility for the “End-Site Resource Manager". This includes an ability to manage the network as a first-class schedulable resource akin to instruments, compute, and storage. A persistent SENSE testbed consisting of network and end-system resources has been deployed across DOE laboratories, university facilities, ESnet and other regional and national networks. A specific CMS use case includes integration of SENSE services with FTS and eventually Rucio.

     This approach provides the necessary tools and methods to coordinate the use of network resources through its “Network Resource Manager” and computing and storage resources (and job schedules) through its “Site Resource Manager”. The SENSE Orchestrator also can be used and is being further developed to handle multiple requests across many network domains. An example of the SENSE international testbed topology demonstrated at the Supercomputing 2019 (SC19) conference, including multiple Layer3 VPN (Virtual Private Network) overlays is shown in Figure 1.

     Starting in September 2019, during discussions and presentations at the Americas and Global Research Platform workshops and the Internet2 Tech Exchange in December, it has become increasingly clear that the services in the SENSE project could be further developed to serve as a mediator among the intelligent network software systems being developed in the various world regions including Europe (AutoGOLE), Latin America (AmLight), and Asia (Virtual Dedicated Networks). A living example of this was demonstrated at SC19[5] where interoperation of the SENSE and AutoGOLE network service frameworks, and integral control of the DTN systems and network systems by the SENSE Resource Managers developed by Caltech and ESnet were shown.

     This has led to plans for a persistent national and global R&D testbed as a venue for ongoing and future network developments in the context of the HL-LHC Computing Model. These developments are also planned to leverage NSF’s major investment in FABRIC[6], “a unique national research infrastructure to enable cutting-edge and exploratory research at-scale in networking, cybersecurity, distributed computing and storage systems, machine learning, and science applications”.

[4] SDN Enabled Networks for Science at the Exascale. See for example https://cs.lbl.gov/news-media/news/2019/sense-takes-software-defined-networking-to-the-next-level/ and https://sc19.supercomputing.org/app/uploads/2019/11/SC19-NRE-013.pdf

[5] SC19 Network Research Exhibition: “LHC Multi-Resource, Multi-Domain Orchestration via AutoGOLE and SENSE”, https://sc19.supercomputing.org/app/uploads/2019/11/SC19-NRE-020.pdf

[6] “FABRIC project launches with $20 Million NSF grant to test a reimagined Internet. https://fabric-testbed.net/news/fabric-award

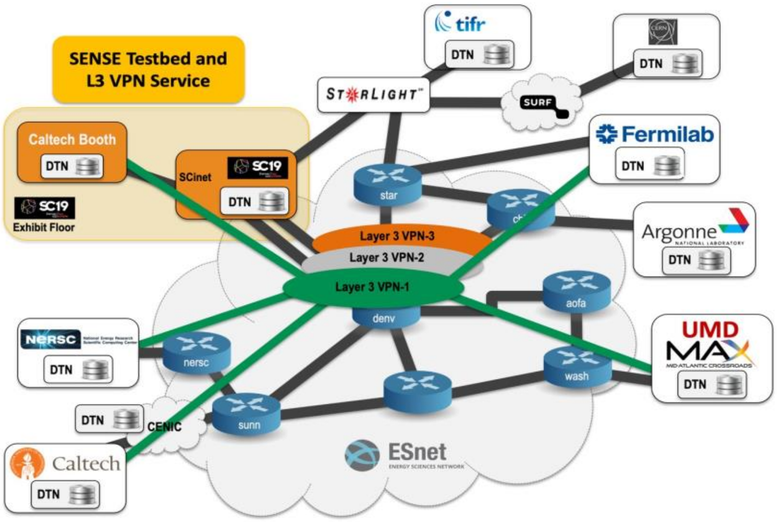

Figure 1 SENSE international testbed spanning multiple sites in the US, CERN, Amsterdam and India, as demonstrated at SC19.

The SENSE system also has been shown to be relatively easy to integrate with other network-oriented toolsets, including both AutoGOLE and Internet2's OESS SDN systems. Near-term development plans include an interface to CERN/IT's NOTED.

2. CERN/IT's NOTED project[7] that was presented at the LHCOPN-LHCONE meeting at CERN in January. NOTED's goal is to publish network aware information on on-going massive data transfers, that can be used to provide additional capacity by orchestrating the network behavior (e.g. more effective use of existing network paths; load balancing). The advantage of starting with NOTED is that its Transfer Broker, as illustrated in Figure 2, can already interpret Rucio and FTS queues and translate them into network aware information with the help of the WLCG's database[8]. While it is still in the prototyping stage, NOTED has already demonstrated the full chain with transfers between CERN and the Tier1s in Germany (DE-KIT) and the Netherlands (NLT1).

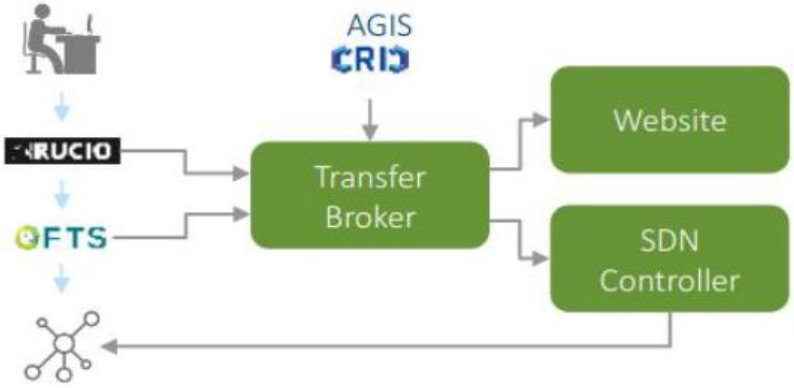

Figure 2 The NOTED architecture including a transfer broker and bandwidth allocation services that can interpret Rucio and FTS queues and translate them into network-aware information.

3. The SANDIE (SDN Assisted NDN for Data Intensive Experiments) project[9] is developing a new system for data access and analysis, based on a novel yet well-founded Named Data Networking (NDN) architecture, supported by advanced (SDN) services (as developed in the SENSE and other

[7] C. Busse-Grawitz, https://indico.cern.ch/event/828520/contributions/3570905/attachments/1968456/3273836/presentation_noted.pdf

[8] CRIC: a unified information system for WLCG and beyond, https://doi.org/10.1051/epjconf/201921403003

[9] https://github.com/cmscaltech/sandie-ndn, https://sc19.supercomputing.org/app/uploads/2019/11/SC19-NRE-035.pdf

projects by the group). SANDIE builds on the use of NDN protocols and services, integrated with the SDN methods and subsystems already developed and under continued development by Caltech and Northeastern and other collaborating groups.

One of the key components of SANDIE is an NDN-based XrootD plugin that is designed to replace the present XrootD implementation and redirectors, together with new data forwarding and caching methods, that are designed to improve the efficiency of CMS' data access, distribution and processing using CMSSW. The plugin architecture is shown in Figure 3.

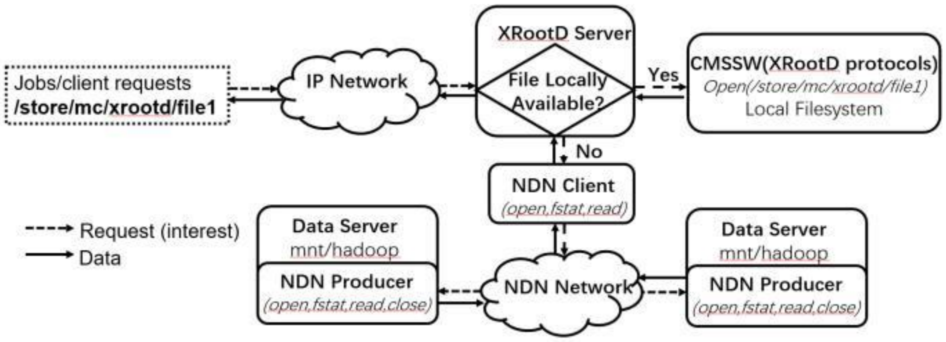

Figure 3 The NDN-based XRootD plugin architecture, developed in the SANDIE project