ESnet Response: National Big Data R&D Initiative RFI

From: Inder Monga, CTO, and Eli Dart, Senior Network Engineer, Energy Sciences Network (ESnet), Lawrence Berkeley National Lab, Berkeley, CA
imonga@es.net and dart@es.net

Introduction

The Energy Sciences Network (ESnet)[1], a division of Lawrence Berkeley National Laboratory[2], is the Department of Energy's high-performance networking facility, engineered and optimized for big-data science. ESnet interconnects the entire national laboratory system, including its supercomputer centers and user facilities. This enables tens of thousands of scientists, at 40 DOE sites, to transfer Big Data, access remote resources, and collaborate productively via access to over 100 research and commercial networks worldwide. The richness of ESnet's global connectivity is motivated by the fact that approximately 80% of its traffic originates or terminates outside the national laboratory complex. This pattern in turn reflects the collaborative, increasingly international, nature of scientific research

Experience

ESnet was founded in 1986, soon after the creation of the global Internet. ESnet's mission has been to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools. Since 1990, ESnet's traffic has increased by a factor of 10 every 48 months as seen in Figure 1, roughly double the growth rate of the commercial Internet[3].

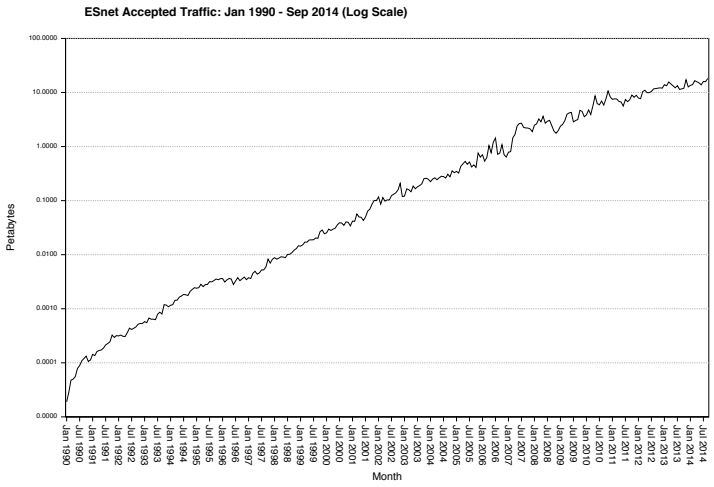

Figure 1: ESnet accepted traffic since 1990 (PB/month)

The structure of modern science now presumes the availability of reliable, high-bandwidth, feature-rich networks for interconnecting instruments and collaborators globally. As science is now largely data-intensive, ESnet focuses its efforts on the effective support of data-intensive science. ESnet has developed its architecture, capabilities and services primarily to serve scientific users and use cases, and maintains its understanding of the needs of its constituents through a formal requirements-gathering process[4].

[1] ESnet web site, http://www.es.net
[2] Lawrence Berkeley National Laboratory, http://www.lbl.gov
[3] IEEE 802.3™ Industry Connections Ethernet Bandwidth Assessment Ad Hoc Report, http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf
[4] ESnet Requirements Review Program, http://www.es.net/requirements/

In addition, ESnet has developed a wealth of real-world experience in enabling data-intensive science. This experience is captured in the fasterdata knowledge base (fasterdata.es.net) which also includes the Science DMZ architecture, created by ESnet. Science DMZs have been widely deployed in the US, with the NSF creating programs (CC-NIE and CC-IIE) to fund the adoption of this architecture. These programs have laid the foundation for a revolution in data-intensive science in the US, and many countries around the world are following suit. ESnet also develops open-source tools for science networks, including perfSONAR, which has seen over 1200 deployments globally, and OSCARS, which received an R&D 100 award in 2013.

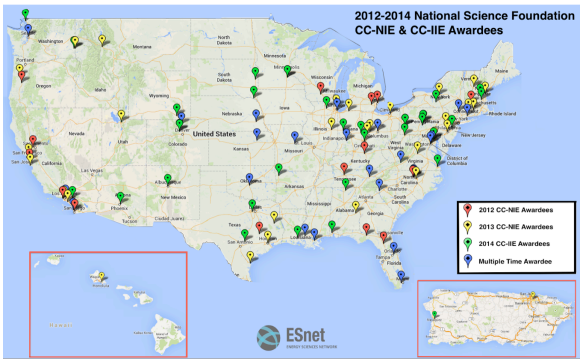

Figure 2: Locations of CC-NIE and CC-IIE awardees

Data movement, networking and its importance in big-data initiatives

As we look at the science constituency worldwide, ***unconstrained data mobility*** is extremely important and many times largely ignored by 'Big Data' initiatives. Large data sets are often moved to facilities where they can be analyzed; moved again for secondary analysis; and eventually shared with collaborators. This pattern applies broadly, and is a hallmark of data-intensive science.

Computing has become essential in virtually every area of science - from extreme-scale simulations to high-throughput analysis of instrument data, the use of computing in science is ubiquitous. Networking is also critical to data-intensive science – it provides for remote access, data movement, workflow integration, data integration, and collaboration. The effective use of networking technologies works as a lever, significantly increasing the value and productivity of interconnected scientific facilities, supercomputer centers, and collaborations.

Why this contribution should be accepted

We welcome an environment where scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data. To achieve this, we believe, future Big Data initiatives must:

- Produce friction-free network architectures, free of physical or logical flaws, which do not impact the ability of protocols to function in a correct and efficient manner.
- Support full-featured applications that have access to a deeper understanding of network construction, behavior, and expectations.
- Foster the deployment and operation of intelligent networks that can function with autonomy when needed, or through direct control by higher-level entities to deliver higher classes of service.
- Support high-performance data migration, data placement, and data access in ways that are productive for users and allow the development of a new class of data-driven scientific methods