Common Infrastructure for Data Intensive Science

Eli Dart, ESnet – dart@es.net

Data intensive science places new constraints on scientists and collaborations, even as it enables new discoveries. In many cases, science collaborations are faced with a sudden increase in the capabilities of their experimental apparatus – the instruments are capable of generating vastly larger volumes of data than current experimental and collaboration models can accommodate, with resolution, and complexity that is unprecedented in human history. This sudden increase in data volume requires that, in order to reap the benefits of the instruments that produce the data, scientists must become adept at a set of tasks that are often outside their areas of expertise. Data transfer, system maintenance, storage system architecture, and software automation are just a few of the tasks that can distract researchers from their focus – scientific discovery. While computers and networks (or their underlying technologies) are responsible in part for increase in data volume, computing and networking offer solutions to the mundane tasks associated with data intensive science.

ESnet conducts regular workshops to determine the network requirements of the experiments and science programs ESnet serves. This experience, along with the ongoing process of solving networking problems for science collaborations and building services to enable new modes of scientific discovery, has led us to an understanding of a core set of capabilities that underlie most successful data intensive science efforts.

These capabilities are:

- High-performance wide area networks to enable the transfer, distribution, sharing, and analysis of large data sets between instruments, facilities, laboratories, universities, and so forth
- Local networks that are architected so as to enable high-performance access to the wide area network by data transfer systems
- Well configured data transfer systems (i.e. computers with access to persistent storage) that run a common data mobility toolset
- Data mobility tools that are either common or interoperable, and can be automated by workflow tools – examples of this include the widely-deployed tools used by HEP, the Earth System Grid, Globus Online, etc.
- Workflow tools that permit the automation of data movement, placement, and analysis

There is one component that is not listed above that is also critical – that is a network test and measurement capability (such as the widely-deployed perfSONAR infrastructure). This was omitted from the above list because it is not in the data flow path from instrument to local storage across the network to remote storage and into an analysis system. However, test and measurement must not be omitted from production deployments as it is a critical component of a properly-functioning data-intensive infrastructure.

What, then, are the common components that can be addressed by a research and development program in collaborative technologies? Data mobility tools and workflow frameworks that are interoperable and easy to deploy by non-experts (i.e. by someone other than the people who wrote the code) are critical components of almost all data intensive science efforts. These tools, if designed and deployed correctly, can be a platform for data intensive science across facilities, institutions, experiments, and programs. If they are not deployed in an interoperable and scalable way, their functions will be replicated by each collaboration in turn out of necessity, since the functionality they provide is required in order to do data intensive science. Therefore, these technologies represent key points of leverage both for research and development and for deployment and sustaining maintenance. These are infrastructure, and will be relied upon by many science collaborations for years to come – they will be the foundations of future success.