100G Networking and Science Application Integration – Experiences, Challenges, Future Plans

Tom Lehman
Director of Research
Mid-Atlantic Crossroads (MAX)

Real Use of 100G Networks: The Good, the Bad, and What to Expect When You Go Beyond a Demo

Internet2 Global Summit, April 10, 2014, Denver, Colorado
Context for this Discussion

• MAX is a regional network in the Washington D.C. metropolitan area.
  – DWDM infrastructure across 11 PoPs
  – 100G services at Layer 3, Layer 2, and Layer 1
  – Fujitsu 9500 (Layer 1), Juniper MX960 (Layer 2 and Layer 3), Brocade MLXe (Layer 2)
  – “ScienceDMZ” on UMD Campus/MAX Edge
  – Advanced Services/Research (AWS, GENI, OSCARS, DCN, OESS)

• Connecting to Internet2 at 2 x 100G
• Peering with ESnet at 100G thru exchange fabric
• Many connected organizations and associated researchers
Working with multiple researchers and science domains to enable use of the 100G regional and national infrastructure.

This includes multiple projects from Astronomy, Physics, Bioinformatics, Climate Science, and others.

When moving beyond demos, the reality of this is that we cannot just focus on the networking aspects.

We have to take a more holistic approach which encompasses networks, storage systems, compute resources, science applications/workflows.

These must be addressed from an “end-to-end” perspective.

Where this “end-to-end” paradigm no longer applies to a single network flow, but instead is related to the application/workflow performance which may include resources at multiple locations and of different types – “Application/Workflow End-to-End”
Thanks to our Research Project Sponsors

- This work is enable by support from National Science Foundation (NSF) and Department of Energy (DOE) sponsored Research Projects
- High Performance Computing with Data and Networking Acceleration (HPCDNA) (NSF CC-NIE)
- Software Defined Network eXchange: End-to-End, Dynamic Science DMZ (SDNx) (NSF CC-NIE)
- 100G Connectivity for Data-Intensive Computing at JHU (NSF STCI)
- Resource Aware Intelligent Network Services (RAINS) (DOE)
100G networks have the promise to greatly enhance the researchers experience

So what are the Challenges?
Three broad categories identified:

• Networking Issues
• End System(s) Issues
• Application/Workflow End-to-End Issues
Networking Issues

• Link Speed Mismatches/Transitions/Bursts
  – Data flows from a 10G sender to a 1G receiver is a well known issue and can cause performance issues (due to burst profiles and resulting link and host system buffer overruns)
  – With new 100G infrastructures we see many more opportunities for mismatches due to possible transitions to/from 1G to 10G to 40G to 100G along the end-to-end path
  – This is similar to the over subscription of uplinks due to bursty behavior of local flows from clusters or individual hosts.
  – Network element buffers need to be tuned/sized properly
Networking Issues

• Network Element tuning
  – Network elements do offer feature sets to help with the above. There are often multiple tuning points available within a single chassis at line card and switch fabric levels. It is often necessary to become expert at tuning intermediate network elements, especially where there are transitions from 100G to 40G or 10G links.

• Link Aggregation (LAG)
  – LAG configurations are a convenient method to manage network topologies and data flows.
  – However it adds to the complexity of the end-to-end dataplane and can complicate troubleshooting
Networking Issues

• Network Path Troubleshooting
  – Complex end-to-end paths which cross multiple domains still really difficult to debug
  – We need a per flow automated debugging infrastructure
  – This is especially true when we use Layer 2 paths across multiple domains
  – Much of what the humans do now can be automated
End System(s) Issues

• The host, storage, compute systems are often located deep within the campus infrastructure
• Researchers often do not have high performance and tuned systems in their labs
• Combination of science DMZs like approaches and lab system build and deployment help are the corrective methods
• Our Science DMZ approach often includes moving some or all of the application to the Science DMZ location (via high performance virtual machine co-location)
• The NSF CC-NIE projects are really helping out here
End System(s) Issues

• Sometimes we have been able to extend the 100G networking infrastructure all the way to the campus End System(s).

• These are generally larger, well funded systems.

• We are also interested in helping the smaller projects and individual researchers to better utilized the 100G network infrastructure.

• We think this will contribute greatly to increased usage of the network infrastructure.

• There are a lot of these smaller research projects, which could benefit from a more “automatic” high performance environment.
Application/Workflow End-to-End Issues

• Researchers have often learned to accept the standard “bad performance” they are used to getting and have adjusted their research activities and goals accordingly.

• Just letting them know that there is a 100G network available will not change this. So we have to work with them directly to solve all the local end system(s) and networking issues.

• However, once we solve all the local system access and networking issues, we still may not get good end-to-end workflow performance due to the remote resource side. We typically have a close working relationship with the local researchers, but the remote resources with which they want to interact is another story.
Application/Workflow End-to-End Issues

• This requires engaging the remote side directly, or getting creative in how the remote side interaction is handled in the workflow.

• Solutions may involve deploying systems like GridFTP if remote system supports or custom scripts/modifications to realize multiple streams in a manner compatible with the remote system and unique to the specific data set.
Application/Workflow End-to-End Environment
So out of these three categories, which one is the most difficult and time consuming?

- Networking Issues
- End System(s) Issues
- Application/Workflow End-to-End Issues

Eventually wan and regional networking issues will be our bottleneck again, but first need to standardize solutions in the other areas.
How do we expect the multiple 100G network capacity will be utilized?

• We have some researchers who can saturate a 100Gbps link by themselves with demo traffic.
• We have some researchers who will likely be able to routinely fill half or maybe more of a 100G link with science research traffic.
• We have a lot of researchers we think we can move to 1~6 Gbps typical transfer rate who are now living with rates of 1-100s Mbps.
• This last group is more time consuming to work with because they do not have the networking expertise on their projects as compared to the larger projects.

• We believe by taking care of an initial 10 or so in this category, we will find ways to standardize the process/technology which will allow others to achieve similar results without as much help needed on our part.

• The result could easily be saturation of our 100 Gbps links during peak usage times.
Future Plans

• With availability of 100G networking and the other issues describe earlier, we will revisit some technologies previously explored
  – High Performance File Systems over the WAN
  – Various protocols (RoCE, iWARP, UDP based)
  – Layer 2 paths in combination with the above

• Intelligent workflow coordination between compute, storage, and network usage and associated application execution to maximize workflow performance

• Services embedded in the network such as storage and compute