IRNC 100G Supplements Kick Off Meeting Summary

Draft compiled by Jennifer M. Schopf, March 2014

1. Summary and Overview

On November 18, 2013, during the SC13 conference in Denver, a group of approximately 40 application scientists and network engineers met to discuss three recently announced supplements to IRNC awards, and what effort could be made to allow applications to make real use of 100G testbeds and eventually production 100 G networks going forward. The goals of the meeting were three-fold

- To identify common issues across the three awards
- To understand and leverage work already being undertaken to enable additional science, including by other communities
- To encourage collaboration among awards, as well as among network engineers and applications scientists.

The opening talk discussed the existing 100G testbed run by ESnet, which is national in scope, and the lessons learned by that project. This was followed by a series of presentations by application groups, primarily focusing on what they could do now, what they wanted to do going forward (and their need for more than 10G links), and what help they thought they needed. The third component was a panel of network engineers talking about the problems they had seen, and the services they hoped to enable. The meeting was highly interactive, and questions and discussions were strongly encouraged. A full agenda, including links to talks and a participant list, is available at http://tinyurl.com/nsf100g or http://internationalnetworking.iu.edu/news/events/science-trials_SC13_Workshop.html

The presentations and discussions included several major themes

1) The hardest part of using a 100G testbed wasn't the use of the testbed itself, but working with the links between the end user's site and the testbed as well as making sure end-to-end performance made met the needs of the applications.
2) It was unclear how many applications truly needed network capacity between 10 G and 100G currently.
3) There were language issues between application domain scientists and network engineers that could (and did) lead to very different understanding of issues.
4) Many domain scientists were interested in baseline functionality issues (reliable performance, basic predictions of behavior), while network engineers often wanted, or interpreted these requests to mean functions provided by higher level, often experimental, technology approaches (choosing paths on the fly, SDN approaches in general).
5) There was no single "user", and requirements between groups, or even within a single team, varies widely.
6) Cross-domain network debugging and measurement are very difficult on 100 G networks.
7) There was a need for better sharing of information in this space, and to address common problems.

Next steps were also discussed, including follow up meetings, sharing of community data, and advocating funding agencies to provide additional resources for resources to aid end use of networks.

2. Supplemental Overviews

Three supplemental awards were made to explore the use of 100G international testbeds to prior IRNC recipients. The International Networks group at Indiana University, led by Dr. Jennifer Schopf, received a supplement to the America Connects to Europe (ACE) award (ACI-0962973) to support end-user science, the use of the Lustre distributed file system, and better monitoring and prediction for 100G networks. Florida International University received a supplement to the Americas Lightpaths (AmLIght) award (ACI-0963053), led by Dr. Julio Ibarra, to deploy an unprecedented experimental 100G alien wave between the U.S. and Brazil. The StarLight consortium, led by Joe Mambretti, received a IRNC supplement (OCI-0962997) to create the Petascale Science Prototype Services Facility (PSPSF), an initiative to design and implement network architectures, services, and core capabilities in support of Big Data science over 100G trans-Atlantic paths.

3. The problem wasn't the 100G testbed

Brian Tierney gave an overview of the use of the ESnet 100G domestic testbed to start the meeting. With such a large capacity link, it has become apparent that the bottlenecks in data transfers have shifted and that getting 100G performance is increasingly about tuning the end-to-end system and the end points instead of backbone behavior. In a series of graphs, it was shown that end-host mismatches, incorrect buffering on one end or the other, and simple disk speeds could have significant impacts on the observed throughput. Tierney also cautioned that the old rules of thumb (such as simply using parallel streams) that have been developed for 10 G links were no longer sufficient, and often were detrimental to achieving higher performance because of unforeseen consequences on the end hosts.

The other problem Tierney emphasized was the scarcity of equipment that could drive 40G or higher flows in a production setting. All aspects of the end-to-end system were being improved, but equipment to drive large flows was expensive and rare. His group had had some success in acquiring loaned equipment, but this could have the additional burden of using other networking equipment that wasn't yet fully production ready and also needed debugging. He commented that every demo he had helped set up showing 40G or larger flows for SC had had issues somewhere in the path that had nothing to do with the 100G link in question.

In general, once the backbone was stable (which could take significant initial start up time), the link itself was not an issue. Getting a large flow of data to the backbone was the challenge.

4. Are the applications there?

There were 10 presentations discussing potential domain science use of a 100G link, and 2/3 of the participants in the meeting classified themselves as domain scientists, as opposed to network engineers. However, one of the underlying themes was the open question of whether or not a 100G link was truly needed by today's science, and if not, when might it be required.

There was a fair amount of discussion about how applications had changed over the last 10 years. Rion Dooley commented he'd "never seen a web page drive 100G", which was underscored by other scientists talking about how they don't use command line any more, and how much work is done through portals and other middleware approaches. Data sizes are growing, and flows are increasing, but end-to-end behavior was changing as well.

Overall there was agreement that some elephant flows would indeed need 100G. Included in those were applications that might have deadlines (weather or disease spread modeling), and others that involved instruments producing so much data it needed to be processed and removed from the instrument in real time (astronomy, HEP).

It was noted that many new instrument and data centers were being built expecting 100G connections. John Cobb pointed out (from his XSEDE experience) that this might involve re-architecting the applications involved to not be fully dependent on local data as well. Several domain scientists noted that as data sizes and use in applications changed, so did how they interacted with the data and the kinds of questions they asked. The over all feeling was that 100G links had the potential to enable one of these changes – if end-to-end performance could be achieved.

5. Engineers often don't speak domain science. And vice versa.

Even at this meeting, where the network engineers all had a long history of working closely with application end users, there were interesting mismatches of language during the meeting. This was witnessed both in misunderstandings (as discussed below when application scientists might ask for something simple, but network engineers would interpret the request as far more complicated), but also terminology. Discussions of check sums, data planes, and other topics were cut short by the moderator when the room was asked "Who knows what is meant by X", and fewer than 10% of the domain scientist gave a positive response.

Similarly, many domain scientists don't know information that the network engineers assumed they would. For example, none of the application people present knew what kind of packet loss could be tolerated by their applications, and several didn't know what packet loss was at all. Most didn't know what network performance they could or should expect, which made knowing that they weren't getting the performance they needed even harder. It was pointed out that packet loss might not be a good metric to use in this space, but the difference in basic assumptions was evident.

6. Functionality vs Shiny

Part of the communication gulf between the network engineers and the application scientists could be seen in what they were interested in discussing. In general, the domain scientists were more interested in baseline functionality issues (reliable performance, basic predictions of behavior), while network engineers often wanted, or interpreted these requests to mean functions required by higher level, often experimental, technology approaches (choosing paths on the fly, SDN).

A great example of this was the domain scientists wanted to know (in their minds very simply) why things weren't working, or even, IF things weren't working on the network.

However, the network engineers argued persuasively that this was not a cut and dried statement, and gave details about the difficulties in testing and monitoring that showed how complicated the underlying system could be. It was agreed that basics, such as where bottlenecks happened, needed to be described, but not how to achieve this information.

Part of the confusion over these issues was rooted in a possible difference in approach. It appeared in some cases that the network engineers wanted a high level, technical solution, whereas the application scientists often needed more basic data. For example, there was a long discussion and interest in providing a network speed map, similar to a Google map, that could show expected data transfer times, perhaps even by time of day. Most of the network engineers assumed one goal of this map would be to be able to change the route that the network used on the fly to enable a better path for the data, an increasingly popular load balancing approach used in SDN networks. However, when asked to clarify, the application scientists uniformly agreed they didn't want to change paths, they just wanted stable, reliable predictions of transfer times.

7. There is no single user

A point that came up several times was that domain scientists cannot be easily categorized into a single "user". As much as there might be shared approaches across some of the groups, even within an application team there were bound to be different points of view. Requirements between groups, or even within single group, could vary widely. For example, while often the discussion was about file transfer times, it was agreed that this was the wrong metric for some applications. It was also pointed out that different applications would have different needs at different stages of development, or by different members of the team with different end goals.

8. There is no single engineer

Debugging cross-domain networking problems, especially in the context of using a 100G trans-oceanic link, is an ongoing problem. When more than one NOC was in play, or the lines of communications for assistance in debugging a networking problem were unclear, many application end users found themselves without a clear path for assistance. While in theory, every organization should have a responsible party responsible for offering a given service; in practice following the lines of ownerships was difficult to impossible.

Additional complications could rise from the interactions between groups servicing a WAN and someone working on a campus network. It was agreed that the end-to-end performance was what mattered, and so campus network engineers needed to be part of the discussion of performance issues, but doing this in practice remains challenging. Also, measurement devices for 100 G networks, which can be required for debugging, are extremely costly.

9. Need for better community data

There was a discussion of the need for additional community resources such as the information through the ESnet Faster Data Knowledge Base (http://fasterdata.es.net/). In general, it was acknowledged that the community as a whole could do a better job at sharing basic data and best practices for use of large and long-haul networks. In general, it was agreed upon that sharing this information would be useful, but the exact way to do so was left undetermined.

10. Next steps

Next steps were also discussed, including follow up meetings, sharing of community data, and advocating funding agencies to provide additional resources for resources to aid end use of networks.

Several next steps for this work were proposed, including:

1) Additional follow up meetings between the three project PIs
2) Ongoing conversations between end user groups and network engineers supporting the proposed testbeds
3) Extensions to the fasterdata.es.net website to include additional best practices
4) Additional meetings at related conferences to talk about the use of 100G networks and large flows in practice.

Appendix 1: Agenda

http://internationalnetworking.iu.edu/news/events/science-trials_SC13_Workshop.html

100 Gbps transatlantic science trials workshop at SC13

When: Monday, November 18th
Time: 1:00pm-5:00pm
Place: Grand Hyatt Denver
1750 Welton Street
Denver, CO 80202-3999 US

1pm: Opening statement and introductions, Dr. Jennifer Schopf, IU

- NSF and 100G Support, Kevin Thompson, NSF
- Overview of the StarLight supplement, Joe Mambretti, NWU
- Overview of the CIARA supplement, Julio Ibarra, FIU (press release here)
- Overview of the ACE Supplement, Jennifer Schopf, IU

1:30: Experiences with the ESNet 100G Testbed, Brian Tierney, ESNet
2:00: Applications and their challenges (Structured as: Experiences to date on prototyping 100 Gbps services, migrating to a wider range of services, additional challenges to be addressed.)

- Lustre use over 100G, Abhinav Thota, IU
- PanSTARRS, Steve Smith, U Hawaii
- Science Gateways: Rion Dooley
- Distributed data storage: Paul Sheldon
- LHCONE: Artur Barczyk, CERN
- Applications focused on data intensive science and distributed research environment, Joe Mambretti, NWU
- Sage at 100G, Maxine Brown, UIC
- Science dmz and 100g – Ohio to Brazil: Marcio Faerman, OSU
- Open Science Data Cloud: Heidi Alvarez, FIU
- Monitoring 100G Networks, Martin Swany, IU

3:15: BREAK
3:30 Special Issues Related to 100 Gbps Networking

- Dale Finkleson, I2, Eli Dart, ESnet, and Jeronimo Aguiar, AMPath
- Panel to address known issues already encountered and anticipated by 100 G applications
- Anticipated new network approaches required for effective use of 100 G networks

4:00 Concluding remarks and discussion

Appendix 2: Attendees

Jeronimo Agular, RNP
Heidi Alvarez, FIU
Artur Barczyk , Caltech/USCHCNet
Maxine, Brown, UIC
Jim Hao Chen. NWU
John Cobb, ORNL
Eli Dart, ESnet
Rion Dooley, TACC
Marcio Faerman, OSC
Dale, Finkelson, I2
Arvind Gopu, IU
Robert Grossman, UC
Robert Henschel, IU
John Hicks, IU
Richard Hughes-Jones, Dante
Julio Ibarra, FIU
Rogerio Iope, Sao Paulo State University
Jason Leigh, Hawaii
Joe Mambretti, NWU
Terry Moore, UT
Edward Moynihan, I2
Matthias Müller. Achen
Robert Quick, IU/OSG
Luc Renambot, UIC
Jennifer Schopf, IU
Paul Sheldon, Vanderbilt
Steve Smith, U Hawaii
Jerry Sobieski, NORDUnet
Martin Swany, IU
Alan Tackett, Vanderbilt
Kevin Thompson, NSF
Abhinav Thota, IU
Brian Tierney, ESNet
Artur Varczyk, CERN
Alan Verlo, UIC
Rob Vietzke, I2
Thomas William. Dresden
Jim Williams, I2/IU
Wolfgang Wuensch, Dresden