Networking: the view from HEP

Shawn McKee

1 Physics Department, University of Michigan, Ann Arbor, MI 48109-1040 USA
E-mail: smckee@umich.edu

Abstract. Networks have played a critical role in high-energy physics (HEP), enabling us to access and effectively utilize globally distributed resources to meet the needs of our physicists. National and global-scale collaborations that characterize HEP would not be feasible without ubiquitous capable networks. Because of their importance in enabling our grid computing infrastructure many physicists have taken leading roles in research and education (R&E) networking, participating in, and even convening, network related meetings and research programs with the broader networking community worldwide. This has led to HEP benefiting from excellent global networking capabilities for little to no direct cost. However, as other science domains ramp-up their need for similar networking it becomes less clear that this situation will continue unchanged. This paper will briefly discuss the history of networking in HEP, the current activities and challenges we are facing, and try to provide some understanding of where networking may be going in the next 5 to 10 years.

1. Introduction
This paper presents a HEP perspective on networking. While not presuming to speak for all of HEP, we do want to cover recent activities related to global HEP networking. In discussions about networking with physicists from around the world, we have gotten a broad range of feedback:

- Most physicists we have spoken with are very happy with the global R&E network infrastructure supporting HEP and its positive enabling role.
- Those involved in supporting our infrastructure have very basic concerns about our ability to effectively use the networks we have.
- Almost everyone would like the network to remain as transparent as possible while continuing to deliver excellent performance.

A quick summary of where we are now: HEP (and especially LHC) networking is global, foundational to our computing models and infrastructure and continuing an exponential increase in bandwidth use. Our networking is functioning well but facing some current and future challenges as we will cover in this paper. The HEP community has significantly benefited from the world-wide Research & Education (R&E) networking community. While our wide-area networking needs are significant and have historically been the "poster child" for globally distributed e-Science, this may be changing over the coming years.

2. A brief history of hep networking
There is a long history of work by physicists, led in large part by Harvey Newman with his modem connection to CERN from Caltech in 1985[1], to enable capable international networking. In the late 1990s the MONARC[2] team developed a model of how LHC experiments might construct a suitable infrastructure accounting for compute, storage and networking. The model assumed the network was
expensive, somewhat unreliable and not very performant and therefore relied on the Tier-0 through Tier-4 hierarchy of computing centers.

After the LHC turn-on the experiments found that the network was actually one of the most reliable and best performing components of our global infrastructure. Furthermore, that excellent wide-area networking (WAN) was generally being provided without direct cost to the experiments. Based upon their experience in Run-1 the LHC experiments evolved their computing models to take better advantage of the network. The two most visible changes were:

1) The hierarchical model was replaced by more egalitarian access to data and sites
2) Direct access to data across the WAN became part of the toolkits (AAA[3], FAX[4], etc)

3. Current hep network activities
A number of activities related to HEP networking are ongoing:

- The Open Science Grid (OSG) Networking Area which is gathering, storing and making available perfSONAR network metrics from both OSG and WLCG.
- The WLCG Network and Transfer Metrics Working Group which is responsible for ensuring all relevant network and transfer metrics are identified, collected and published.
- R&E backbone networks like ESnet, Internet2, GEANT, etc. that are providing the primary capacity for HEP traffic and working closely with HEP to plan for the future.
- Communities like LHCOPN/LHCONE, GLIF, perfSONAR that are exploring network technologies, measurement and management capabilities critical for HEP.
- The many institutions and communities worldwide supporting HEP networking.

One of the main challenges for HEP is to identify and integrate appropriate developments from any of the above work into its global infrastructure. The next two sections highlight areas relevant to this.

3.1. Osg and wlcg network efforts
Starting in 2012 the Open Science Grid added a networking area. OSG is now in its fifth year supporting WLCG/OSG networking and is focused on the following:

- Assisting its users and affiliates in identifying and fixing network bottlenecks
- Improving the ability to manage and use network topology and network metrics
- Developing an analytics platform based upon the Elasticsearch, Logstash and Kibana (ELK) software stack (see section 5)
- Supporting higher-level network services via network metric summarization
- Developing effective Alarming and Alerting for network problems

Meanwhile the WLCG has also been active in networking for HEP. In 2012 a perfSONAR deployment task force was created to get perfSONAR Toolkit instances deployed at all WLCG Tier-2 and above sites. The task-force concluded in 2014, having successfully deployed 250 instances at WLCG site all over the world[5]. Since 2014 WLCG has had the Network and Transfer Metrics working group[6] operating. Working groups are responsible for managing a particular operational area over time. In addition to tracking and tuning the network measurement infrastructure the working group has created a support unit to coordinate responses to potential network issues:

- Many issues are resolvable within the working group without the need to involve other support channels
- Real network issues can be identified and directed to the appropriate network support centers
- Many issues are resolved within hours mainly due to using perfSONAR information.

A history of tickets and support incidents is tracked on the working group wiki[7]

3.2. Lhcopn and lhcone working groups
The LHCOPN[8] working group was established by CERN, the WLCG Tier-1 sites and the various HEP related research and education networks to define, deploy and operate the LHC Optical Private
Network interconnecting the Tier-1 and the Tier-0 at CERN before the startup of the LHC. It was very successful in delivering a dedicated, predictable and reliable network to distribute the data from the experiments at the LHC out to their world-wide collaborations.

The success of LHCOPN for the Tier-1s led to the creation of a similar network to support the Tier-2s and their interactions with the Tier-1s: The LHC Open Network Environment (LHCONE)[9]. The LHCOPN/LHCONE group now meets jointly 2-3 times per year to discuss policy, operations and future evolution necessary to support the LHC, and now the broader high-energy particle physics communities. This mostly volunteer effort composed of physicists, network engineers and network researchers has been very beneficial for LHC and continues to be one of the primary means of planning for and organizing networking for high-energy physics.

4. The importance of measuring our networks

Because of the central, foundational role networking plays for high-energy physics it is critical to continually measure and monitor our networks. End-to-end network issues are frequently difficult to spot and localize, and network problems are often multi-domain. Performance issues involving the network are complicated by the number of components involved end-to-end. Network problems can severely impact the WLCG experiment’s workflows and can take weeks, months and even years to get addressed.

There has been a consensus decision to standardize on specific tools and methods to allow our various groups to focus resources more effectively and better self-support. We have converged on perfSONAR which provides a number of standard metrics we can use to support our needs:

- Latency measurements provide one-way delays and packet loss metrics
  - Packet loss is almost always very bad for performance
- Bandwidth tests measure achievable throughput and track TCP retries (using Iperf3)
  - Provides a baseline to watch for changes; identify bottlenecks
- Traceroute/Tracepath track network topology
  - Measurements are only useful when we know the exact path they are taking through the network.
  - Tracepath additionally measures the Maximum Transmission Unit (MTU; the largest packet that can traverse the network without fragmentations) but is frequently, and inappropriately, blocked by firewalls.

4.1. Use of perfsonar

The use of perfSONAR for HEP started in 2006 with efforts in the US to use it to track the networks between various ATLAS sites. This proved very successful after helping to identify numerous, long-standing network issues at and between those ATLAS sites. Soon after there was interest from the LHCOPN community in having a similar capability. In 2011 all LHCOPN sites instrumented in the 4 months between two LHCOPN meetings. The increased visibility of the network and associated problems was an important tool for running and maintaining the LHCOPN. After this, the WLCG experiments agreed it would be beneficial to deploy perfSONAR and initiated the WLCG perfSONAR deployment task force mentioned above. By 2014, after deploying more than 250 instances the task force was decommissioned and the work handed over to the WLCG Network and Transfer Metrics working group.

5. Activities in network analytics

The volume and complexity of network related data being collected by OSG and the experiments is challenging to use, but holds the promise of providing much deeper insights into our networks and hard to identify network problems. To be most useful, the data requires cleaning, augmenting, transforming & correlating. Ilija Vukotic (Univ. of Chicago) has developed ELK/Jupyter stack for ATLAS Analytics and worked with Xinran Wang on anomaly detection and advanced alerting/notifications for network problems[10]. In addition they have also looked at detection of the
anomalies based on machine learning models with some success. There are additional efforts in network analytics for HEP by:

- Jerrod Dixon and Brian Bockelman (UNL) exploring network analytics for CMS
- Henryk Giemza (NCFB), Federico Stagni integrating perfSONAR metrics and associated analytics into DIRAC for LHCb
- Shawn McKee (Univ. of Michigan) working on real-time root cause analysis (PuNDIT) in collaboration with perfSONAR developers.
- Hendrik Boras and Marian Babik (CERN) working on developing models for a network cost-matrix to determine the relative performance of network paths.

These type of efforts will be critical for maximally benefitting from the large, complex set of network measurements we are acquiring.

6. Challenges for hep networking

There are a number of significant challenges the HEP faces in networking right now that will be discussed in the following sections.

6.1. Network operations

Network operations for HEP refers to deploying, maintaining and monitoring all the services and components in use to measure and monitor our networks. The initial deployment of perfSONARs at all WLCG sites made it possible for us to see and much more easily debug end-to-end network problems. These perfSONAR instances are now tracked and configured by the Network and Transfer Metrics working group discussed above. This group is also helping sites and experiments with network issues using perfSONAR. Reports of non-performing links are actually quite common (almost on a weekly basis) and most of the end-to-end issues are due to faulty switches or mis-configurations at sites. Some problems are due to link saturation (recently in LHCOPN) or issues at NRENs. Recent network analytics of LHCOPN/LHCONE perfSONAR data also pointed out a very interesting fact about our networks: we observe significant packet loss on almost 5% of all LHCONE links. Specifically, we monitor all source-destination network paths for loss by sending a 10Hz stream of packets (used to measure one-way delay) on each path and tracking the losses every minute. Paths are considered bad if they average more than 2% loss during a 3 hour period. It is therefore becoming increasingly important to focus on site-based network operations and identifying associated problems.

6.2. Network diversity

Another network challenge facing HEP is the range of capacities and funding models across our global set of sites. Some sites do not explicitly pay for WAN but others may need to pay for:

- Their Wide-Area Network (WAN) connections
- Any “special” services, such as virtual circuits, non-standard peerings, traffic shaping, etc.
- Bandwidth use beyond some threshold
- Support

Also Tier-1 and Tier-2 sites have a wide range of bandwidth connectivity to the WAN from 1 - 200 Gbps. This diversity in capability and cost leads to differences in perspectives about HEP networking planning and goals. Sites with excellent networking and small costs want to see the network emphasized (to reduce other costs or improve capability). Conversely, sites with lower capacity or “expensive” networking want to have the infrastructure configurable to conserve their networking use. It can be challenging to get consensus in how much we should emphasize use of the network.

6.3. Making the most of our networks

One thing we have learned in our efforts is that much of our WLCG infrastructure is NOT tuned to take the best advantage of the networks we currently have. There is a wide range of mis-configurations, non-optimal tunings and incorrect application and hardware settings that lead to
inefficient use of our networks. As mentioned above, we have a wealth of data now available and ready for analysis to identify bottlenecks and poor performance.

As we identify bottlenecks and poor performance we need to take the next step and work to improve our end-host’s ability to effectively utilize the network we have. This “low-hanging fruit” has significant advantages. It does not require SDN, new hardware or new networks but can make a huge difference in network throughput for our sites. It may be worthwhile to organize a workshop to share best practices, tools and tuning information to help remedy this situation.

7. Planning for the future
As a community HEP needs to think about what it wants to do regarding networking and at what timescales. Within the networking research community there is a vision of a long-term technology evolution producing “Smart Nets”. What characteristics will such networks have and how much work will HEP need to do to take best advantage of them? What things should we worry about in the near-term? The mid-term? The long-term? Before addressing questions about the future use and characteristics of the network, we want to cover some areas that provide a basis for thinking about those answers.

7.1. Lhc data growth
The LHC experiments have been transferring exponentially increasing amounts of data since their startup. This trend is likely to continue as it is driven by increasing data volumes, more capable infrastructure and excellent networks. Figure 1 demonstrates the exponential increase.

![Figure 1 Shown: the exponential increase in transfer volumes for the 4 main LHC experiments](image)

7.2. Evolution of the LHC Program
We will see significant pressure on network resources, which will likely accelerate in high-luminosity (HL) LHC by a factor of 10. Funding is not expected to increase and will likely remain flat. Shown in Figure 2 is the LHC schedule and the HL-LHC should be online in 2025.

7.3. Improving end-host networking
New operating systems and associated end-host improvements in hardware are making it easier to get higher performance on our wide-area networks. The TCP network stack is more stable and better configured out-of-the-box in recent operating systems like Community CentOS 7 (CC7) which enables
the network throughput to ramp up much more quickly[11]. There has been a significant number of other enhancements and results recently:

- Fair Queueing Scheduler (FQ) available from kernel 3.11+
  - Even more stable, works better with small buffers[12]
- Best single flow tests show TCP LAN at 79Gbps, WAN (RTT 92ms) at 49Gbps[11]
- IPv6 slightly faster on the WAN, slightly slower on the LAN[11]
- New TCP congestion algorithm (TCP BBR) from Google
  - Google reports 2-4x performance improvement on path with 1% loss (100ms RTT)[13]
  - Early testing from ESNet less conclusive, there is also question how tolerant BBR will be with other congestion algorithms on the same link.[11]

The trend is that those working on operating system development are now aware of the wide-area network and will be continuing to improve end-system ability to fully utilize the WAN. This should be a positive development for HEP but will have implications in WAN network use.

7.4. Wan vs lan capacity and cost

Historically WAN capacity has not always had a stable relationship compared to data-centers LAN capacity. In recent history WAN technologies grew rapidly and for a while outpaced LAN or even local computing bus capacities. Today 100Gbps WAN links are the typical high-performance network backbone link speeds, but LANs are also in the same range. The list price for 100Gbit dual port card is ~ $1000[14], but significant discounts can be found (as low as $400), while a list price for 16 port 100Gbit switch is $9000. Because of this, it is relatively easy to over-subscribe WAN links in terms of the relatively inexpensive cost of local hardware at many sites.

Will WAN be able to keep up? Likely yes, but it should be noted that HEP did benefit from the fact that 100Gbit WAN was deployed on time for Run2. Equivalent upgrades from the current 100Gbit to 400Gbit or 1000Gbit might not be in place at the same relative time for Runs 3 and 4, respectively. For example, even though 800 Gbps waves will likely be available by the end of 2020, they will have a significantly larger relative cost, since waves at this bandwidth are foreseen to be deployed at proportionally shorter distances (thus more repeaters, and therefore more dollars, are needed).

7.5. Software defined networks

Software Defined Networks (SDN) are a set of technologies offering potential solutions for many of the current and future challenges in networking. Currently, networks are typically sized based upon peak-use and upgraded to the next capacity step when average use rises about ~30%. Network providers are paying for the relatively rare peak-use and paying for a significant amount of unused
capacity. In a recent presentation, ESnet’s Joe Metzger showed that current links might handle ~3-6x more traffic if we could avoid peaks and be more efficient in use of the available bandwidth[15]. SDN capabilities are being explored in many different point-to-point efforts within the LHCOPN/LHCONE communities. The challenge remains getting this capability end-to-end across HEP’s infrastructure. Also, while it is still unclear which technologies will become mainstream, it is already clear that software will play a major role in networks in the mid-term (commercially driven). Will experiments have the manpower to engage in the existing SDN testbeds to determine what impact it will have on their data management and operations?

7.6. Research & education networking

Research and Education network providers have been working in mutually beneficial ways with HEP for a long time because:

- HEP (especially LHC) has been representative of future data intensive science domains
- HEP serves as a testbed environment for early prototyping of evolving capabilities

HEP is no longer the only field requiring high throughput global networking. SKA (Square Kilometer Array) plans to operate at data volumes **200x current LHC scales**. Besides Astronomy there are many science domains anticipating data scales beyond LHC, cf. ESRFI 2016 roadmap[16]. So the obvious question is: “*What does n more HEP-scale science domains competing for the same network resources imply?*” Will HEP continue to enjoy “unlimited” bandwidth and prioritized attention or will it need to compete for the networks with other data intensive science domains? Will there be AstroONE or BioONE soon? Perhaps more importantly, will they bring additional network funding?

8. Summary and conclusion

HEP networking has been a reliable, high-performing infrastructure component for HEP’s globally distributed resources. Even so there remains a significant amount of work to do in finding and localizing problems (new or existing) and fixing bottlenecks.

New technologies and the ongoing work in HEP networking will continue to make it easier to increase data transfer rates between sites. It is likely that HEP will continue to exponentially increase its use of the network for the foreseeable future and, considering only HEP, this seems sustainable from the perspective of the R&E networks.

We still must track network capacities and technology changes that significantly modify (up or down) HEPs use of the network. What sites and NRENs (National R&E Networks) do regarding capacity upgrades can have a large impact on the use of the network. Additional HEP computing model evolution and its timing relative to changes in site and NREN upgrades need to be watched.

Perhaps the most important consideration is that HEP will likely no longer are the only domain using global R&E networking in the 5+ year time-frame. If true it means that sharing the future available capacity successfully will require greater interaction with networks. While it is unclear what technologies will become mainstream, we know that software will play a major role in the networks of the future and we need to be ready to use it and integrate it within our architectures.

We believe HEP needs to properly prepare for an uncertain future. My recommendations covering the short, medium and long-term are:

- **Short-term (1-2 years):** Focus on network monitoring, debugging and analytics. Find and fix network problems, improving our ability to utilize the networks we have.
- **Medium-term (3-7 years):** Plan for and evaluate the use of SDN for our infrastructures. Work on integration of those aspects deemed beneficial. Estimate the impact of other data-intensive science domains on our R&E networks and collaborate with them on their ramp-up to our scale.
- **Long-term (8-12 years):** Plan for and deal with the R&E network environment: sharing, orchestration, automation and the implementation of smart networks. Ensure our software
can interact with smart network capabilities and agilely respond to dynamically changing infrastructure capacities and problems.

HEP has been very successful in benefitting from and capitalizing on networking and with proper planning and effort it should continue to do so for the foreseeable future.

**Acknowledgements**

The author would like to acknowledge the contributions of the perfSONAR project, namely staff from ESnet, GEANT, Indiana University, Internet2, the University of Michigan and all the contributors shown at [http://www.perfsonar.net/about/who-is-involved/](http://www.perfsonar.net/about/who-is-involved/) as well as the efforts of all those in OSG, and WLCG involved in networking for HEP. Of course it is also very important to acknowledge the significant support HEP has received from the global research and education networking community!

Lastly the author gratefully acknowledges the support of the Department of Energy (grant DE-SC0007859) and the National Science Foundation (grant PHY-1148698) which supported the work in this paper.

**References**

[1] Cook G April 2009 Building the World's Largest Optical Network to Enable Global Science Collaboration. In: The COOK Report on Internet Protocol Technology, Economics and Policy, pp 1-30

[2] Legrand I C and Newman H B 2000 The Monarch Tools for Simulating Large Network-Distributed processing Systems. In: Winter Simulation Conference, ed R R B J. A. Joines, K. Kang, and P. A. Fishwick

[3] Bloom K, Boccali T, Bockelman B, Bradley D, Dasu S, Dost J, Fanzago F, Sfiligoi I, Tadel A M, Tadel M, Vuosalo C, Würthwein F, Yagil A and Zvada M 2015 Any Data, Any Time, Anywhere: Global Data Access for Science. In: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), pp 85-91

[4] Robert G, Simone C, Guenter D, Johannes E, Andrew H, Friedrich G H, Jan I, Federica L, Ilija V, Wei Y and the Atlas C 2014 Data federation strategies for ATLAS using XRootD Journal of Physics: Conference Series 513 042049

[5] Campana S, Brown A, Bonacorsi D, Capone V, Girolamo D D, Casani A F, Flix J, Forti A, Gable I, Gutsche O, Hesnaux A, Liu S, Munoz F L, Magini N, Mohammed K, Rand D, Reale M, Roiser S, Zielinski M and Zurawski J 2014 Deployment of a WLCG network monitoring infrastructure based on the perfASONAR-PS technology Journal of Physics: Conference Series 513 062008

[6] Babik M and McKee S 2014 The WLCG Network and Transfer Metrics Working Group.

[7] Babik M and McKee S 2017 WLCG network support incidents.

[8] Martelli E 2015 The LHC Optical Private Network (LHCOPN).

[9] Martelli E and Stancu S 2015 LHCOPN and LHCONE: Status and Future Evolution J.Phys.Conf.Ser. 664 052025

[10] Vukotic I and Gardner R 2017 Big Data Analytics Tools as Applied to ATLAS Event Data. In: 22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP2016). (Journal of Physics Conference Series)

[11] Tierney B 2016 Recent Linux TCP Updates, and how to tune your 100G host

[12] Tierney B 2017 Packet Pacing using the FQ (Fair Queuing) scheduler. (Fasterdata: ESnet)

[13] Cardwell N, Cheng Y, Gunn C S, Yeganesh S H and Jacobson V 2016 BBR: Congestion-Based Congestion Control Queue 14 20-53

[14] 2017 Dell Accessories - (Network Adapter. Dell) Dell accessories web page showing the cost of a dual-ported Mellanox 100G network interface card ($829.99)

[15] Metzger J 2017 ESnet Network Upgrade Plans: ESnet6. In: HEPiX Conferences: HEPiX Presentation on ESnet's network upgrade plans

[16] 2016 STRATEGY REPORT ON RESEARCH INFRASTRUCTURES.