Network Evolution

Shawn McKee, Marian Babik

USATLAS Facilities Meeting

OSG All Hands Meeting March 19, 2018

Salt Lake City, UT

Introduction

- High Energy Physics (HEP) has significantly benefited from strong relationship with Research and Education (R&E) network providers
  - Thanks to LHCOPN/LHCONE community and NREN contributions, experiments enjoy almost “infinite” capacity at relatively low (or no-direct) cost
  - NRENs have been able to continually expand their capacities to overprovision the networks relative to the experiments needs and use
- Aim of this talk is to stimulate discussion on LHC Network Evolution
  - **Scope is mid to long-term**
  - Identify and discuss emerging trends and their potential impact
  - Propose areas of interests for LHC community and ways to follow up after the workshop

Current Network Status

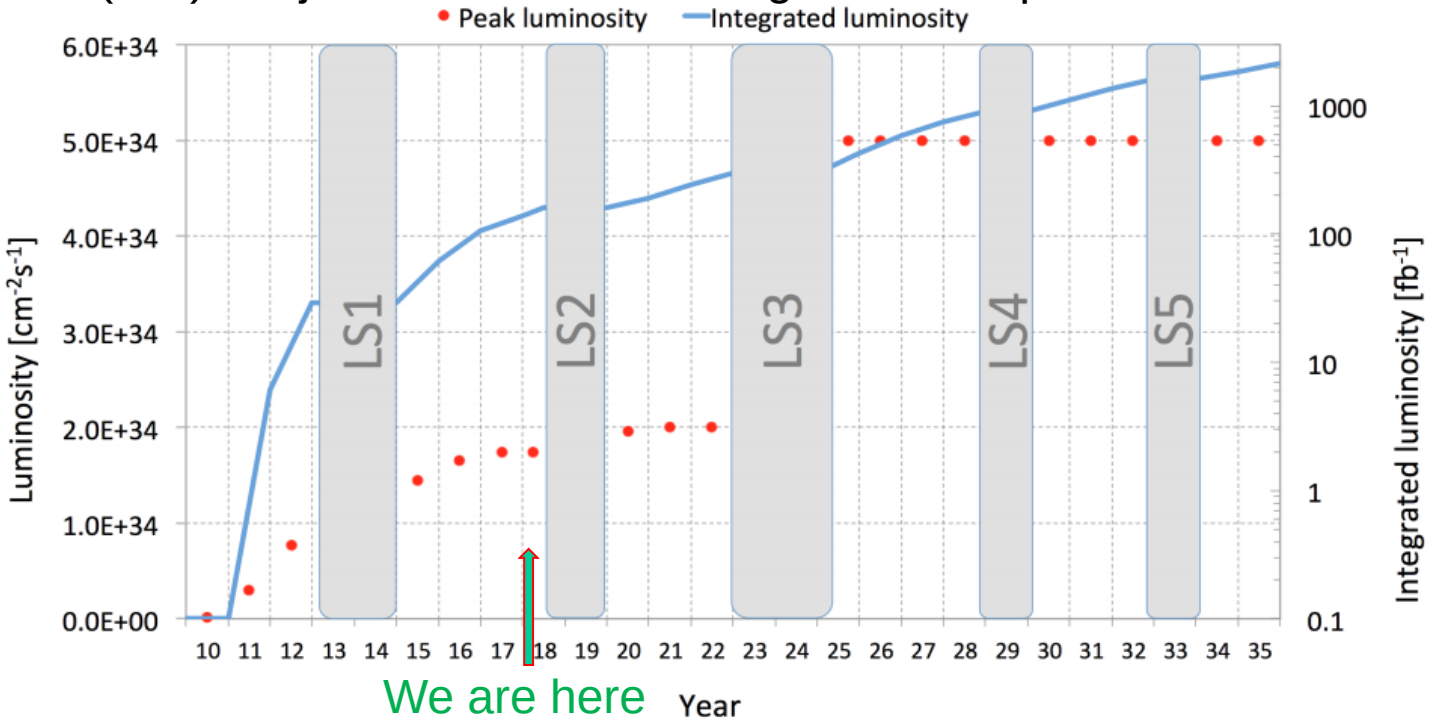
As noted this talk is on the mid-to-long term perspective but I did want to state a few things about our current status up front

- Our networks have performed very well for our community
- Most physicists are happy with the networking we have
- Some concerns exist around our ability to fully utilize **existing** networks
- Visibility is key to understanding, maintaining and fixing our networks

So there is **near-term** work to do regarding networking in optimizing, monitoring and fixing network problems, but we should also think longer term regarding how the situation may evolve and what that might mean for us.

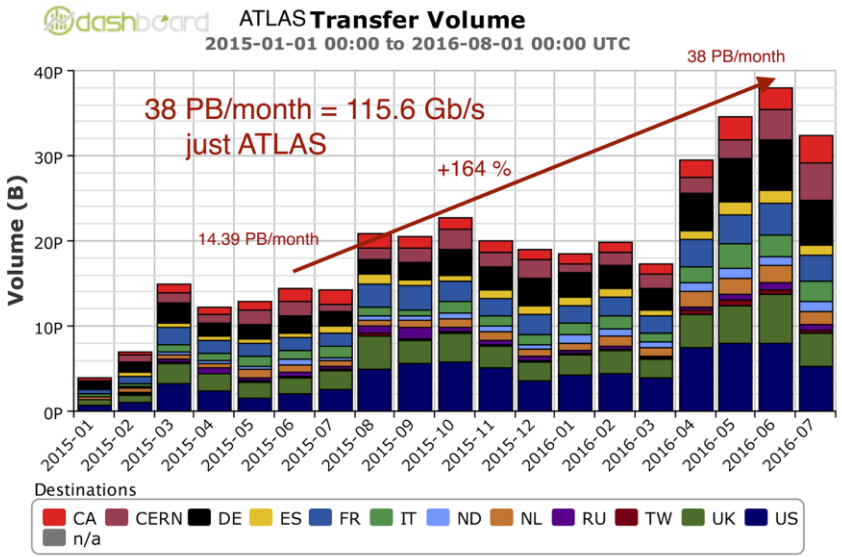
LHC schedule

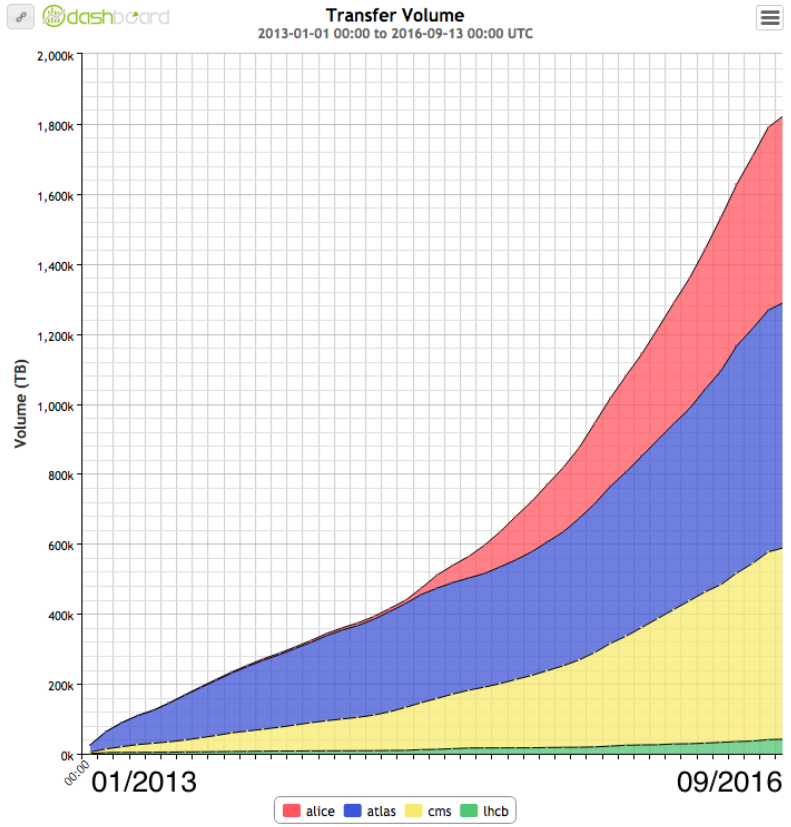
We will see significant pressure on network resources, which will likely accelerate in HL-LHC (x10). Major increases in funding are not expected and will likely remain flat.



LHC Traffic Growth

Experiments have been transferring exponentially increasing amount of data since startup. This trend is likely to continue as it's driven by increasing data volumes, more capable infrastructure and excellent networks.

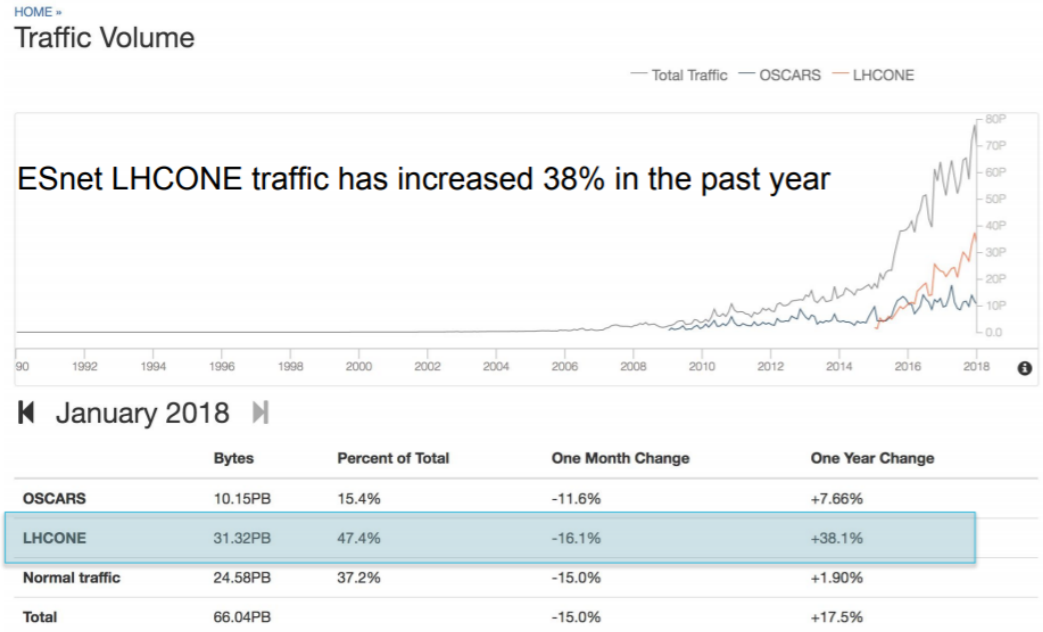




R&E Traffic Growth: ESnet View

ESnet Traffic Volumes

LHCONE handles more traffic than any other ESnet service



	Bytes	Percent of Total	One Month Change	One Year Change
OSCARS	10.15PB	15.4%	-11.6%	+7.66%
LHCONE	31.32PB	47.4%	-16.1%	+38.1%
Normal traffic	24.58PB	37.2%	-15.0%	+1.90%
Total	66.04PB		-15.0%	+17.5%

ESnet

LHCONE is now almost half of all ESnet Traffic and continues to drive annual increases.

In general, ESNet sees overall traffic grow at factor 10 every 4 years. Recent LHC traffic appears to match this trend.

GEANT reported LHCONE peaks of over 100Gbps with traffic increase of 38% in the last year.

This has caused stresses on the available network capacity due to the LHC performing better than expected, but the situation is unlikely to improve in the long-term.

Slide from Michael O'Connor, LHCONE operations update

R&E Networking - provisioning

- R&E network providers have long been working closely with HEP community
  - HEP has been representative of the future data intensive science domains
  - Often serving as testbed environment for early prototypes
- Other data intensive sciences will be coming online
  - SKA (Square Kilometer Array) plans to operate at data volumes 200x current LHC scale
  - Besides Astronomy there are MANY science domains anticipating data scales beyond LHC, cf. ESRFI 2016 roadmap
- Network provisioning will need to evolve
  - LHCOPN as a dedicated/private network might become shared with other experiments
    - Already the case for Belle II, other experiments might be coming
    - Understanding how we monitor/share existing capacities will be needed
  - LHCONE has its own challenges - will we see AstroONE, BioONE ?
    - In its present status - it's already extensively complex
- **Monitoring and managing network as a resource in a similar way we do compute and storage today is becoming likely in the future**

Improving Our Use of the Network

- TCP more stable in CC7, throughput ramp ups much quicker
  - Detailed report available from Brian Tierney/ESNet
- Fair Queueing Scheduler (FQ) available from kernel 3.11+
  - Even more stable, works better with small buffers
  - Pacing and shaping of traffic reliably to 32Gbps
- Best single flow tests show TCP LAN at 79Gbps, WAN (RTT 92ms) at 49Gbps
  - IPv6 slightly faster on the WAN, slightly slower on the LAN
- **In summary: new enhancements make tuning easier in general**
  - But some previous "tricks" no longer apply
- New TCP congestion algorithm (TCP BBR) from Google
  - Google reports factor 2-4 performance improvement on path with 1% loss (100ms RTT)
  - Experimental evaluation by KIT is less conclusive
  - However this is a work in progress, BBR version 2 is in the works (not yet open source)
  - It will likely become serious contender and we'll need to plan its evaluation/deployment
- We can also explore TCP alternatives (UDP+control, other protocols for non-shared paths)

Software Defined Networks (SDN)

- Software Defined Networking (SDN) a set of new technologies enabling the following use cases:
  - **Automated service delivery** - providing on-demand network services (bandwidth scheduling, dynamic VPN)
  - **Clouds/NFV** - agile service delivery on cloud infrastructures usually delivered via Network Functions Virtualisation (NFV) - underlays are usually Cloud Compute Technologies, i.e. OpenStack/Kubernetes/Docker
  - **Network Resource Optimisation (NRO)** - dynamically optimising the network based on its load and state. Optimising the network using near real-time traffic, topology and equipment. This is the core area for improving end-to-end transfers and provide potential backend technology for DataLakes
  - **Visibility and Control** - improve our insights into existing network and provide ways for smarter monitoring and control
- Many different point-to-point efforts and successes reported within LHCOPN/LHCONE
  - **Primary challenge is getting end-to-end!**
- While it's still unclear which technologies will become mainstream, it's already clear that software will play major role in networks in the mid-term
  - Massive network automation is possible - in production and at large-scale
- HEPiX SDN/NFV Working Group was formed to bring together sites, experiments, (N)RENs and engage them in testing, deploying and evaluating network virtualization technologies

Cloud Networking

- Commercial cloud providers already operate big networks at global scale with significantly higher capacities that are available in R&E
- Cloud computing is also becoming an important topic and eventually we'll need to find ways how to effectively bridge commercial and R&E networks
  - ATLAS/Rucio project with Google is one the examples going in this direction
- Will commercial WAN become available in a similar manner we are now buying cloud compute and storage services ?
  - The underlying cost will be decisive
  - Transit within major cloud providers such as Amazon/Google currently not possible and likely challenging in the future, limited by regional business model - but great opportunity for NRENs

Data Lakes

- Simone and Rob have already talked on this, but a few comments relevant to networking:
- Data Lakes will rely on the networks to provide both in-lake and out-of-lake transport
  - Decoupling storage and data management is an opportunity for us to re-think how we currently manage and operate our networks
  - Finding ways how to integrate network monitoring/feedback with storage and provide near-real time information to the controllers will be important to get an efficient system
- Streaming/Caching is another area which is currently unaware of the network
  - Non-managed caches directly in the network hubs could provide significant benefits
  - This is already provided by commercial CDNs and is a potential opportunity for (N)RENs
  - It could be coupled with other services CDNs provide such as security (DoS protection, etc)
- Unless we find ways how to better interact with the networks, network operational cost of data lakes can be significant
  - Debugging ASGC to NDGF transfers where storage is located in Slovenia gives a hint at the challenges we'll likely face

Tech Trends: Containers

- Recently there has been a strong interest in the container-based systems such as Docker
  - They offer a way to deploy and run distributed applications
  - Containers are lightweight - many of them can run on a single VM or physical host with shared OS
  - Greater portability since application is written to container interface not OS
- Obviously networking is a major limitation to containerization
  - Network virtualization, network programmability and separation between data and control plane are essential
  - Tools such as Flocker or Rancher can be used to create virtual overlay networks to connect containers across hosts and over larger networks (data centers, WAN)
- Containers have great potential to become disruptive in accelerating **SDN** and **merging LAN and WAN**
  - But clearly campus SDNs and WAN SDNs will evolve at different pace

Network Operations

- Deployment of perfSONARs at all WLCG sites made it possible for us to see and debug end-to-end network problems
  - OSG is gathering global perfSONAR data and making it available to WLCG and others
- A group focusing on helping sites and experiments with network issues using perfSONAR was formed - WLCG Network Throughput
  - Reports of non-performing links are actually quite common (almost on a weekly basis)
  - Most of the end-to-end issues are due to faulty switches or mis-configurations at sites
  - Some cases also due to link saturation (recently in LHCOPN) or issues at NRENs
- Recent cases have shown that MTU (+LB) is a significant issue
  - This is not going to change and might even get worse once overlays (VXLAN) start to be used more broadly
  - New WG was started within LHCOPN/LHCONE to understand the issues and propose an MTU policy for LHCONE
- It is increasingly important to focus on site-based network operations

Summary

- Increased importance to oversee network capacities
  - Past and anticipated network usage by the experiments, including details on future workflows
  - Sites vs NREN capacities
- New technologies will make it easier to transfer vast amounts of data
  - And HEP likely no longer the only domain that will need high throughput
- Sharing the future capacity will require greater interaction with networks
  - While unclear on what technologies will become mainstream, we know that software will play a major role in the networks of the future
  - We have an opportunity here with a little planning
- Containers might become the “accelerator” for adoption of SDNs on campus
  - With impact on skills and effort required to manage local networks

Questions or Comments?

Additional Slides from 2017 WLCG Discussion

WLCG Network Discussion

We have had three presentations: IPv6, LHCOPN/LHCONE and Network Evolution and we want to have an open discussion about network planning for WLCG in a moment

Prior to this workshop we asked for responses to a number of WLCG related network questions and the summary follows

Many thanks to all those who replied!

WLCG Net-survey: General summary

We received **17** responses, some encompassing many sites.

Sites(**18**): CERN, NDGF-Tier-1, AU-Melbourne, CA-TRIUMF / SFU, ES-PIC Tier-1, UK-RHUL Tier-2, UK-Cambridge Tier-2, UK-RAL Tier-1, UK-Imperial College Tier-2, KISTI Tier-1, INFN Tier-1, NorthEast Tier-2, FR-IN2P3 Tier-1, US-Nebraska Tier-2, US-Purdue Tier-2, US-Caltech Tier-2, FI-HIP Tier-2, FNAL Tier-1

Experiments: ATLAS, ALICE, LHCb

There were also some extra views and comments sent along

WLCG Net-survey: What fraction of your LAN/WAN capacity is used?

LAN	WAN
80	80%
100	100%
	100%
<50	100% LHCOPN, 50% LHCONE
Low	High; traffic bursty & traffic hits 100% when flowing but sometimes no traffic
5	2%
	75%
50	50%, but with weekly periods at 100%

LAN	WAN
Very low	10%
Very low	20% annual, but periods at 60%; usage increasing
10-70%	10-100% (bursty)
25%	50%
20-30%	10-20%
50-80%	20%
10-30%	15-40%
100%	80%

Broad range of usage among respondents

WLCG Net-survey: What are the obstacles in upgrading your network?

None/Just done	4
Money	8
Traffic	6
Guidance	4

Upgrade obstacles
Summary: Guidance appreciated; sites upgrade as they see the need based on traffic, but money is an obstacle---especially if Tier managers don't control the network.

WLCG Net-survey: What obstacles do you face in enabling IPv6?

Next level up: Campus support	4
Dual stack storage services	3
Time/people	2
Technical knowledge/Support from the community	1

IPv6 obstacles

WLCG Net-survey: Short-term question responses(1)

Short term (the rest of Run2): can we cope with the extra load due to the better than expected LHC performance? Can we be IPv6 ready? Can we monitor, diagnose and manage our networks sufficiently for HEP needs?

- Network will be OK; lack of storage is likely the major problem
- Networks underlie our global infrastructure and ideally ATLAS can continue to (semi-transparently) benefit from them. I think that in the near-term ATLAS needs to ensure we have the right amount of effort in place to monitor, test and debug our networks.
- In the short-term (1-2 years), we want alerting on network issues, the ability to understand our network paths and associated bottlenecks and the ability to make more intelligent use of the network based upon what we know from our monitoring.
- For FNAL in the short term yes, we have not yet seen our resources saturated at the WAN level, we have at the LAN level, however workflow changes by the VO's are making this much less an issue.

WLCG Net-survey: Short-term question responses(2)

Short term (the rest of Run2): Can we be IPv6 ready?

- For IPV6 we are readying this now, and expect we'll be able to meet WLCG timelines. FNAL though does not need IPV6 capability on our own, if anything our internal address space is shrinking due to increased core counts of machines, etc.
- IPv6 is currently being tested by LHCb, we are almost ready from the LHCb side to be ipv6 compatible. There are a few agents / services (notably the sandbox service) that need to be moved to dual-stack machines. And once that is done, we should be completely compatible with ipv6.

Can we monitor, diagnose and manage our networks sufficiently for HEP needs?

- I don't think we've run into cases where any network issues were a "mystery", so the right things appear to be monitored. You can always have fancier monitoring :)

WLCG Net-survey: Mid-term question responses

Medium term (LS2/Run 3): Upgrading the networks for Run 3; moving towards complete availability of data over IPv6. How do we integrate Software Defined Networking capabilities that exist in our networks to support our needs?

- We will cope

Questions:

Do we need to provide some level of effort thinking about SDN for WLCG?

Are industry changes coming in networking useful/relevant for WLCG?

WLCG Net-survey: Long-term question responses

Long term (HL-LHC): Do we influence the international network environment or live in an environment built for more demanding customers? To frame this another way, will HEP continue to enjoy abundant bandwidth and prioritised attention or will we need to adapt to an environment filled with other network intensive science domains competing for the network?

- Whatever; networking will be OK

Questions:

Do we foresee network budgets (sites, R&E backbone, NRENs) scaling with the number of LHC-scale science users of the network?

Do we need to worry about competitively sharing our networks on the timescale of 7-10 years?

WLCG Net-survey: What LAN/WAN bandwidths for Run 3?

The LAN and WAN throughput rates should be proportional to the resources evolution, i.e. to the number of cores, local storage capacity and data transfers (T0->T1s) for RAW data replication as a particular case. For the latter (as specified in the Run3 upgrade TDR) we need bandwidth to T1s sufficient to export 1/3 of data of standard PbPB run (10 PB out of 30) in reasonable time (1 month) which translates into 40Gbit/s to T1s from CERN

In general, we don't have a network throughput jump foreseen for RUN3 beyond the aforementioned standard site capacity and RAW data replication requirements.

- 40Gb/s to CERN (LHCOPN)
- Multi 100Gb/s
- 100Gb/s for Tier-2.
- Need to use networks more intelligently, not just throw bandwidth at the problem

WLCG Net-survey: Do you want Tier2 sites to separate clearly into ones with disk and ones without?

No clear answers.

Entirely diskless not feasible: “without disk” must mean with disk only for local cache purposes, i.e. without custodial responsibility for data.

But Tier-2s don’t have custodial responsibility and cached data can be sent on WAN just as easily as to the LAN provided there is the relevant bandwidth.

Managing storage at Tier-2s should be made easier.

Open Discussion

We should ask for comments, suggestions and questions

One item that seems to have come to the surface: storage policy, plans and corresponding estimates are going to be very important in network planning.