

HEPiX Network Functions Virtualisation WG Update

Marian Babik, Shawn McKee
LHCOPN/LHCONE session Umea

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
- Other data intensive sciences will be coming online soon (SKA, LSST, etc.)
 - Besides Astronomy there are MANY science domains anticipating data scales beyond LHC
- Network provisioning will need to evolve
 - Focusing not only on network capacity, but also on other **network capabilities**
- It's important that we explore new technologies and evaluate how they could be useful to our future computing models
 - **While it's still unclear which technologies will become mainstream, it's already clear that software (software-defined) will play major role in networks in the mid-term**

Network Functions Virtualisation WG

Mandate: Identify use cases, survey existing approaches and evaluate whether and how Software Defined Networking (SDN) and Network Functions Virtualisation (NFV) should be deployed in HEP.

Team: 60 members including **R&Es** (GEANT, ESNNet, Internet2, AARNet, Canarie, SURFNet, GARR, JISC, RENATER, NORDUnet) and **sites** (ASGC, PIC, BNL, CNAF, CERN, KIAE, FIU, AGLT2, Caltech, DESY, IHEP, Nikhef)

Monthly **meetings** started last year (<https://indico.cern.ch/category/10031/>)

Mailing list: <https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepixonfv-wg>

Objectives/sub-tasks

- Work organised in **two phases**, phase I (exploratory):
 - Define use cases
 - Explore SDN/NFV approaches for compute, e.g. OpenStack/Kubernetes (mainly **intra-site** activities) - **Cloud Native Networking**
 - Explore SDN/NFV approaches for distributed storage/end-to-end transfers, e.g. data lakes (**inter-site** activity in collaboration with RENS/NRENS) - **Programmable Networks**
 - Evaluate existing approaches (ODL, Contrail, OVN/OVS, etc.), analyze readiness/gaps
 - Share experiences between the sites/RENS/NRENS
 - Tutorials/introductory material to help sites establish their testbeds;
Document deployment experiences, issues/gaps, production readiness
- Initial report/white paper - in progress - TBD today
 - If we agree that there should be phase II (mainly wrt cross-site SDN/NFV deployment) then:
 - Propose timetable and analyse resource needed to run cross-site experiments/testbeds
 - Implementation and configuration advice, organise scalability/performance testing

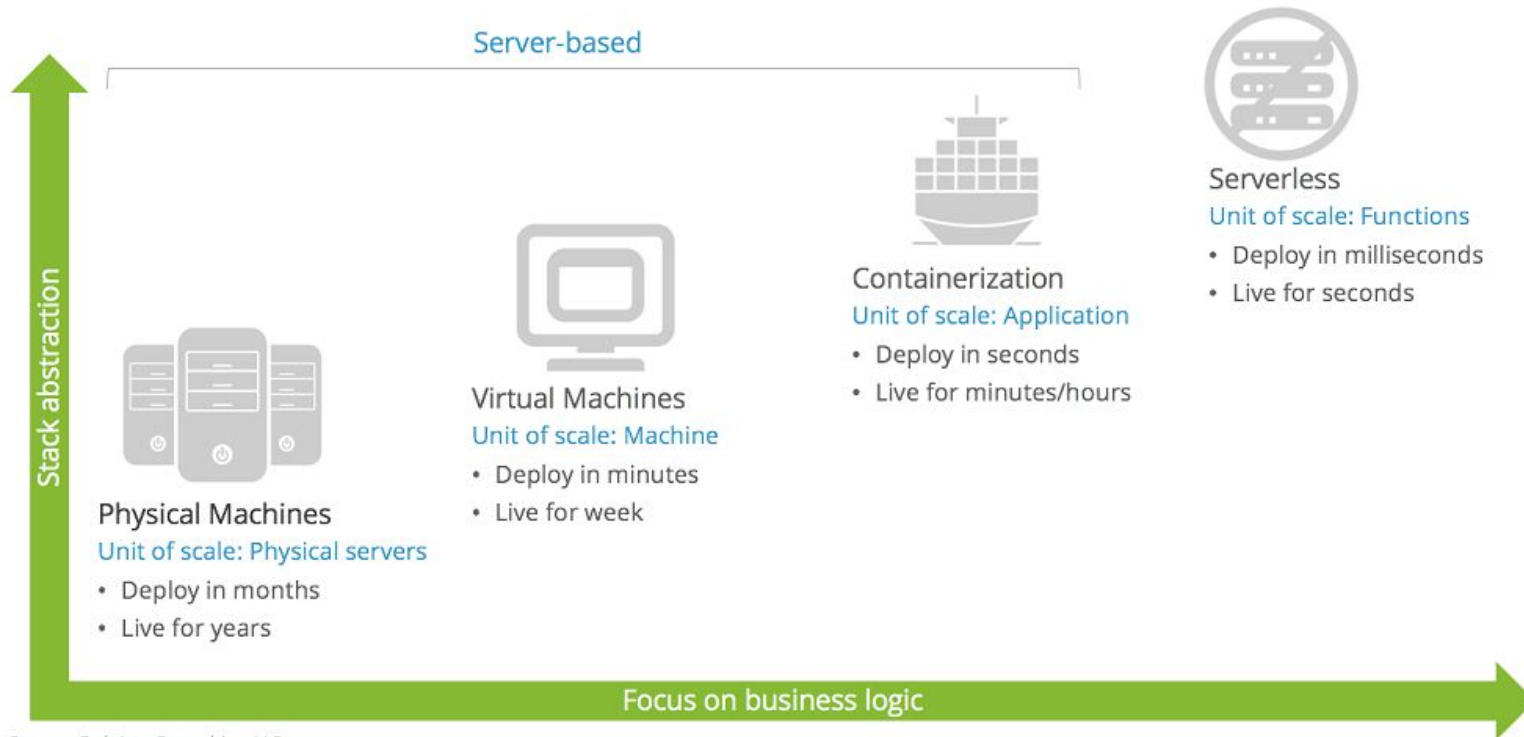
Past Meetings

- Cloud Native Networking oriented sessions
 - 25th April - SDN/NFV Demo Session
 - 23rd May - SDN/NFV Demo Session 2
 - 20th June - Tungsten Fabric Session (CERN/Nikhef)
- Programmable networks oriented sessions
 - 19th Sept - AMLight SDN project
 - 31st Oct - F2F @FNAL
 - Dec - BigData Express (Wenji Wu, FNAL)
 - Jan - SENSE & ESNet Evolution (Chin Guok, ESNet)
 - Feb - GEANT Evolution and Plans (Mian Usman, GEANT)
 - Mar/Apr - WP discussion
 - June - F2F @Umea - WP discussion
- Full meeting schedule + slides
 - <https://indico.cern.ch/category/10031/>
- HEPiX fall will be in Amsterdam 14-18th of October

Cloud Native Networking

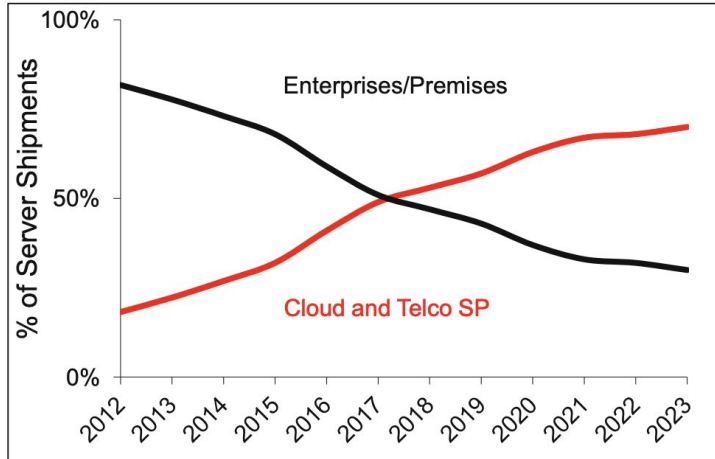
Cloud Computing Evolution

Figure 1: Road to Serverless Solutions
Paradigm shift in computing evolution



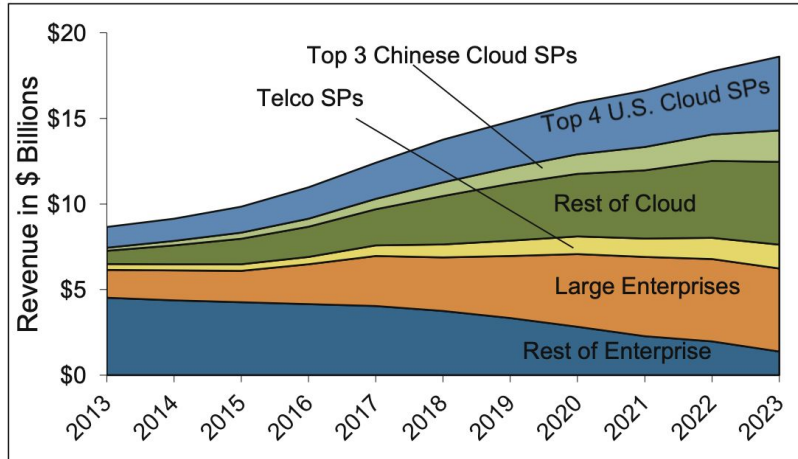
Arista's Cloud Networking Opportunity

Server Shipments



Source: Dell'Oro Group Server Research

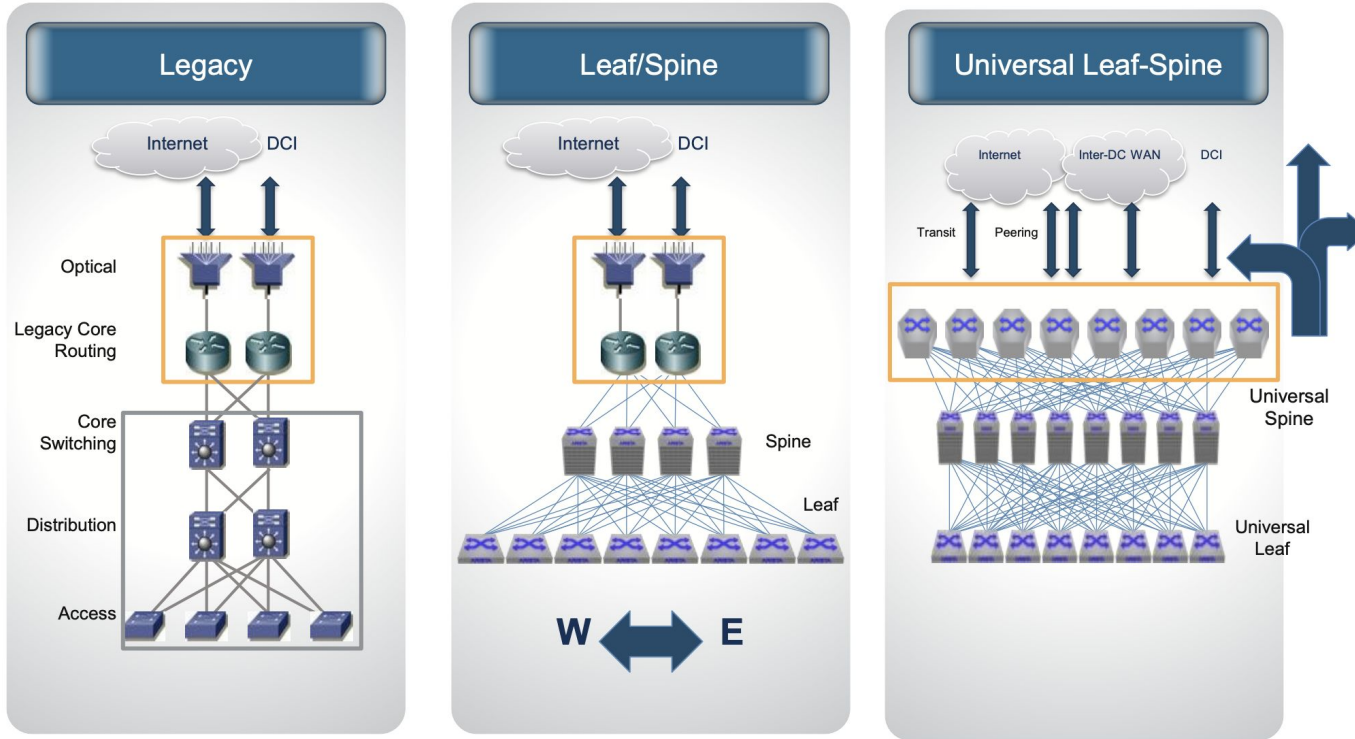
Data Center Ethernet Switch Revenue



Source: Dell'Oro Ethernet Switch Data Center 5 Year Forecast Report January 2019

- Enterprise workloads are migrating to public & hybrid clouds
- Traditional enterprise served <100,000 employees vs. Clouds @ hundreds of millions of users
- The emergence of cloud native apps & containers necessitates a new architecture

Evolution of the Universal Cloud Network Architecture



Routing is Integrated into the Universal Spine and Leaf

Cloud Native (GCP example)

```
# create an app-vpc custom network
```

```
gcloud compute networks create app-vpc --subnet-mode=custom
```

```
# create a frontend subnet
```

```
gcloud compute networks subnets create frontend-subnet --network app-vpc  
--region us-west1 --range 192.168.0.0/24
```

```
# create a backend subnet
```

```
gcloud compute networks subnets create backend-subnet --network app-vpc  
--region us-west1 --range 192.168.1.0/24
```

```
# establishing peerings
```

```
gcloud compute networks peerings create test --network=frontend-subnet  
--peer-network=backend-subnet
```

```
# add firewall rules (per app/subnet/etc.)
```

Use Cases

Data centre networking offering standard cloud compute services

- **Native support for multi-stack**
 - Connecting and integrating multiple orchestration stacks like Kubernetes, Mesos/SMACK, OpenShift, OpenStack, VMware
- **Support for network equipment configuration and management**
- **Network support across-stack**
 - Networking and security across legacy, virtualized and containerized applications
- **Multi-tenancy - application level networking (VPNs on demand)**
- **Native support for creating federated services spanning DCs (e.g. fed k8s)**
 - Multi-cloud support - Data Centre Interconnect and Remote Compute
- **Security, monitoring and debugging**
 - Multistack and across-stack policy control, visibility and analytics
 - Networking and security across legacy, virtualized and containerized applications

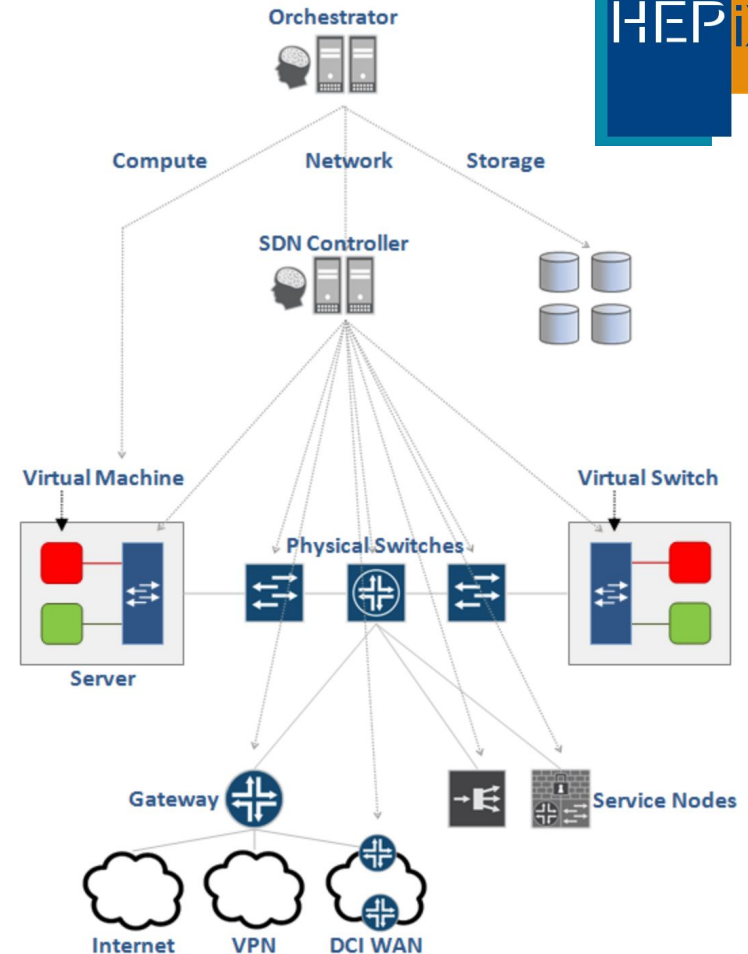
SDN/NFV DC Schema

Orchestrator (or VMI) manages critical aspects of DC

- Compute (VMs, containers, BM)
- Storage
- Network

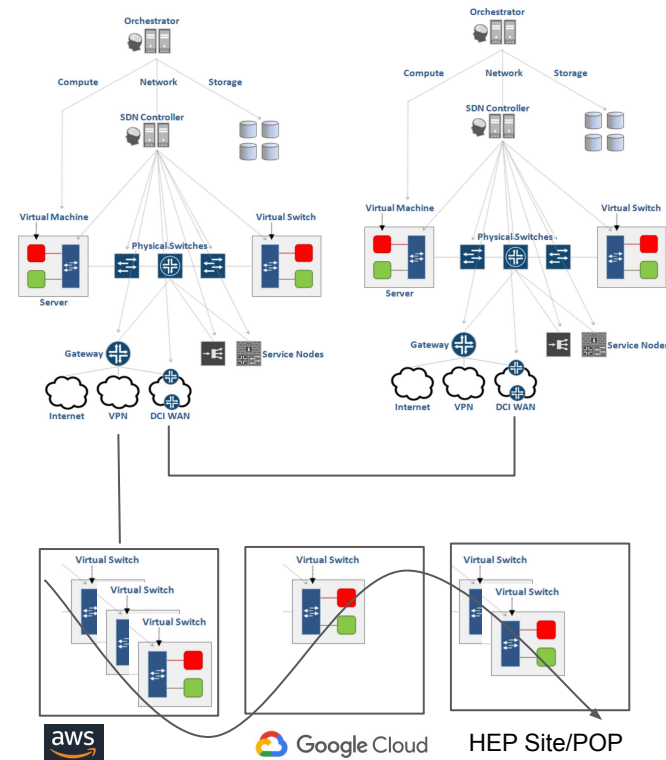
SDN Controller - manages network and network services - its core role is translate high-level (orchestrator) requests (e.g. create VM, connect tenant VMs to network, create network service such as load balancer, setup security policies for tenant, etc.) into concrete actions on physical and virtual switches, e.g. setup overlay networks, attach VMs to ports on virtual switch, route packets btw. virtual and physical networks, etc.

DC Network runs in a combination of virtual and physical switches, routers/gateways and tunneling protocols (VXLAN, MPLSoGRE, etc.)












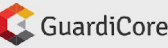








Multi-Cloud - DCI/Remote Compute

- SDN-based DC enables other interesting options
- **Data Center Interconnect (DCI)**
 - SDN services spanning multiple physical sites, each site with its own SDN deployment.
 - Agnostic to the Virtual Infrastructure Manager (Orchestrator) used.
- **Remote Compute**
 - Single SDN deployment extending its services to remote sites (POP/DC/Cloud). Ability to extend VPNs/VMs to another site without running a dedicated SDN cluster there.
- **Service chaining (NFV)**
 - Steering traffic between VPNs/VMs according to a policy, availability, etc.
- All the options are complementary and can be combined to create high-scale networking combining 100s or even 1000s of sites.



Cloud Native Networking Landscape

SDN solutions for data centres supporting **OpenStack, Kubernetes, OpenShift**. Range of approaches, both open-source (white) and commercial (grey) exist, tracked by the [Cloud Native Computing Foundation](#)

| | | | | | | | | |
|--|---|--|---|--|---|---|--|---|
|  <p>alcide</p> <p>Alcide Alcide Funding: \$12.3M</p> |  <p>Aporoto</p> <p>Aporoto Aporoto Funding: \$34.5M</p> |  <p>aviatrix</p> <p>Aviatrix Aviatrix Systems Funding: \$25M</p> |  <p>big switch</p> <p>Big Switch Networks Big Switch Networks Funding: \$120M</p> |  <p>cilium</p> <p>Cilium Isovalent ★ 3,836</p> |  <p>CNI</p> <p>Container Network Interface (CNI) Cloud Native Computing Foundation (CNCF) ★ 1,938</p> |  <p>Contiv</p> <p>Contiv Cisco ★ 93 M.Cap: \$232B</p> |  <p>CUMULUS</p> <p>Cumulus Cumulus Networks Funding: \$130M</p> |  <p>flannel</p> <p>Flannel Red Hat ★ 3,959 M.Cap: \$32.1B</p> |
|  <p>GuardiCore</p> <p>GuardiCore Centra GuardiCore Funding: \$48M</p> |  <p>LIGATO</p> <p>Ligato Cisco ★ 43 M.Cap: \$232B</p> |  <p>MULTUS</p> <p>Multus Intel ★ 399 M.Cap: \$240B</p> |  <p>vmware NSX</p> <p>NSX VMware M.Cap: \$73.6B</p> |  <p>nuagenetworks</p> <p>Nuage Networks From Nokia Nuage Networks Nuage Networks</p> |  <p>OCTARINE</p> <p>Oclarine Oclarine</p> |  <p>OvS</p> <p>Open vSwitch ★ 1,844</p> |  <p>PROJECT CALICO</p> <p>Project Calico Tigera ★ 951 Funding: \$53M</p> |  <p>tungstenfabric</p> <p>Tungsten Fabric Tungsten Fabric ★ 401</p> |

Our focus is on **Open vSwitch/Open Virtual Networking** and **Tungsten Fabric**

- Both open source, best match to our use cases – for comparison see [OVS talk by Y. Yang](#)

Existing Tungsten Fabric pilot deployments at CERN and Nikhef

Experiences with commercial deployments reported on regular basis by IHEP

Linux Foundation Networking

Additional projects that improve SDN/NFV performance, provide alternative SDN controllers, offer programmable off-loading capabilities, etc. are hosted by Linux Foundation



[Intel's Data Plane Development Kit](#) (DPDK) - accelerates packet processing workloads running on a wide variety of CPU architectures

[P4](#) - programming language for packet processing - suitable for describing everything from high- performance forwarding ASICs to software switches.

SmartNICs

- Now offered from multiple vendors - goal is to maximise capacity while providing full programmability for virtual switching and routing, tunnelling (VXLAN, MPLS), ACLs and security groups, etc.
- Three approaches are being followed:
 - FPGA based - good performance, but difficult to program, workload specific optimisation
 - ASIC based - best price/performance, easy to program but extensibility limited to pre-defined capabilities
 - SOC based - good price/performance, easily programmable, highest flexibility
- Datapath programmability ([tutorial](#))
 - Application level - OpenVSwitch, Tungsten vRouter, etc.
 - Packet movement infrastructure (part of data path) - BPF (Berkeley Packet Filter)/eBPF
 - Full description of data path - P4 language
- FPGA-based SmartNICs broadly deployed in Microsoft Azure
- Tungsten Fabric 5.1 release plans to [support](#) smartNICs
- Good overview provided in [ACM SIGARCH article](#)

Programmable Networks

Use Cases

Focus mainly on **end-to-end transfers** and **SD-WAN** like capabilities

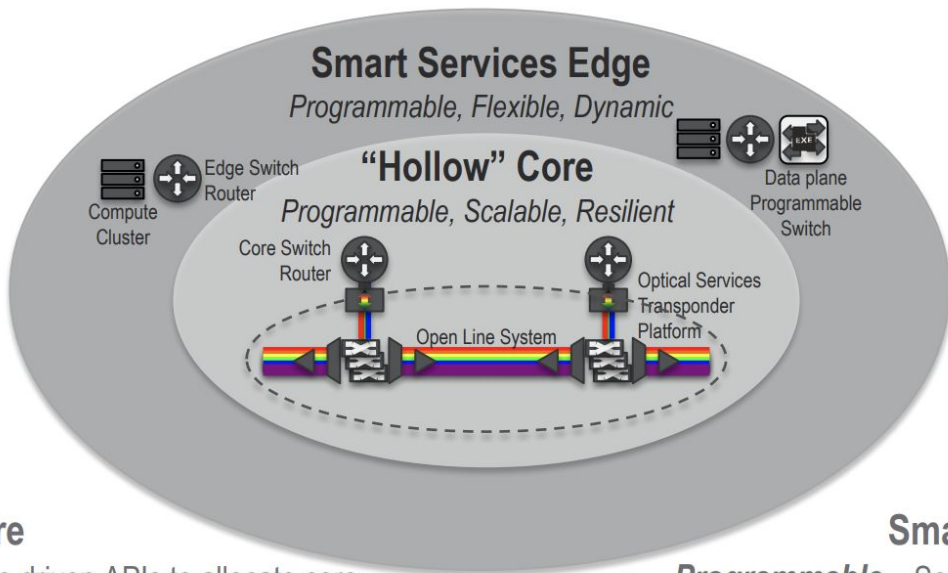
- **Increase throughput** by establishing on-demand circuits/flows
 - Additional capacity exists and can be provisioned by steering traffic via alternate paths
- **Provide QoS** for time critical transfers
 - Currently transfers are scheduled on first time/first basis and transfers are shared evenly - this doesn't always match the existing requirements (e.g. streaming data directly from exp).
- **Improve network to storage performance**
 - Currently there is often a mismatch between target storage and network performance
 - In addition, inefficiencies arise when existing data transfer tools run on DTNs (NUMA I/O, scheduling overheads, caches)
- **Effective use of HPC**
 - One of the major obstacle in fully utilising current HPCs are existing network limitations (mostly wrt. **capabilities** not capacity).
- **Capacity sharing**
 - Monitoring and managing network as a resource in a similar way we do compute and storage today is becoming likely in the future

R&E Plans

- R&E network providers have long been working closely with HEP
 - HEP has been representative of the future data intensive science domains
 - Often serving as testbed environment for early prototypes
- Surveying their plans for higher-level services and providing our feedback is critical for future evolution of HEP networking
- Different approaches are being followed - ranging from full SDN capable services (AMLight) up to a range of various low to higher-level edge services (ESNet6)
- Some important questions to address:
 - How do we tackle such range of network capabilities across R&Es ?
 - What services will be offered, how do we accommodate different functionality across R&Es ?
 - What interdependencies exist between Cloud Native Networking and R&E plans ? Can we effectively run DCIs over future networks ?
- **Our ability to use the programmable edge services will directly impact our ability to effectively use future networks.**

ESnet6 (“Hollow-Core”) Architecture Overview

Orchestration
and Automation



Monitoring and
Measurement



“Hollow” Core

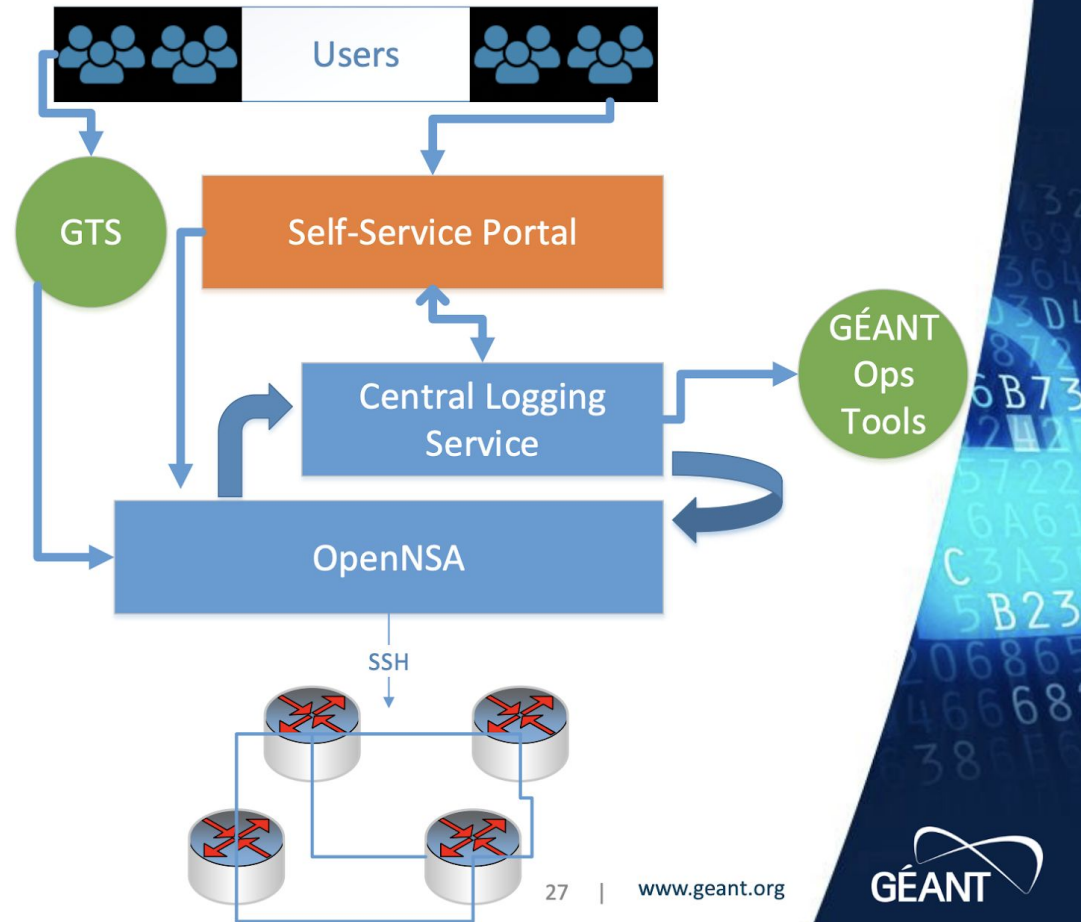
- **Programmable** – Software driven APIs to allocate core bandwidth as needed, and monitor status and performance.
- **Scalable** – Increased capacity scale and flexibility by leveraging latest technology (e.g. FlexGrid spectral partitioning, tunable wave modulation).
- **Resilient** – Protection and restoration functions using next generation Traffic Engineering (TE) protocols (e.g. Segment Routing (SR)).

Smart Services Edge

- **Programmable** – Software driven APIs to manage edge router/switch and retrieve telemetry information.
- **Flexible** - Data plane programmable switches (e.g. FPGA, NPU) in conjunction with compute resources to prototype new services (driven by Software Defined Networks (SDN))
- **Dynamic** – Dynamic instantiation of services using SDN paradigms (e.g. Network Function Virtualization (NFV), Virtual Network Functions (VNF), service chaining).

Connection Service

- Automated service delivery
- Self service portal
- Open interface (API) for use with 3rd party automated systems and orchestrators
- Automated monitoring and service inventory
- Multi-domain services

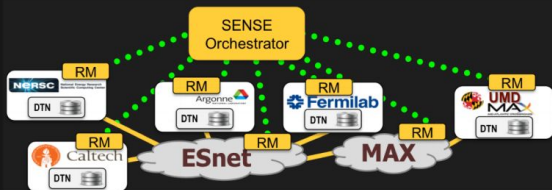


SENSE

SDN for End-to-end Networked Science at the Exascale (SENSE) - U.S. funded; ESN_{et}, FNAL, ANL, NERSC, Caltech and Univ. of Maryland

Application Workflow Agent

- Transforms the network into a first class resource for workflow planning and optimization
- Allows applications to “query and negotiate” with network
- The SENSE infrastructure is designed to develop these types of services in DevOps manner, and customize for individual application agents.
- Cannot do every computation possible, but can do any computation desired.



Please provide a listing of all available provisioning Endpoints

Endpoint Listing

What is the maximum bandwidth available for a P2P service between Caltech and Fermilab?

20 Gbps available for a P2P service between Caltech and Fermilab

Request 20 Gbps P2P service between Caltech and Fermilab. If 20 Gbps not available 10Gbps is ok.

15 Gbps P2P service between Caltech and Fermilab Instantiated

Service is not working, please check status

Failure on a network element, problem fixed

BigData Express

U.S DOE funded: FNAL, ESNNet, StarLight, KISTI, Univ. of Maryland, ORNL



Our Solution - BigData Express



- BigData Express: a schedulable, predictable, and high-performance data transfer service
 - ✓ – A peer-to-peer, scalable, and extensible data transfer model
 - A visually appealing, easy-to-use web portal
 - ✓ – A high-performance data transfer engine
 - A time-constraint-based scheduler
 - ✓ – On-demand provisioning of end-to-end network paths with guaranteed QoS
 - Robust and flexible error handling
 - CILogon-based security

Existing projects also in ATLAS (OVS btw AGLT2/MWT2/KIT), SDN aspects also in NSF-funded **SLATE**, **OSIRIS** and CERN's **NOTED** project

Outlook

WG plans (schedule & meeting topics)

- Topics/meetings to be scheduled
- Programmable networks oriented:
 - GEANT survey and JISC/JANET evolution (Tim Chown, JISC)
 - GEANT connect (higher level services)
 - KIT network evolution, DFN evolution (Bruno)
- Cloud native oriented:
 - Network design @USC, Arista CloudVision (Azher Mughal)
 - Tungsten deployment update (CERN/Nikhef)
 - P4 and smartNIC technologies (Mauro Campanella, GARR)

WP input

- Input from R&Es very important specifically related to plans for higher level services (APIs)
 - Challenges in bridging cloud native networks ?
 - Tungsten inter-DCI white-paper was circulated in the mailing list, mostly depends on MPLS - will this work ? If not, what are the alternatives ? Experiences/projects related to smartNIC/scaling out virtual networks ?
 - What R&D areas are being currently pursued (GEANT connect, ESNNet higher level services, etc.)
 - Transition from existing R&D projects to future platform offerings - what areas of existing R&D has support for going production (and at what time-scale)
 - NOTED, SENSE/BigData Express future

Working Group Plans

- Finalise whitepaper by HEPiX fall
 - Possibility organise SDN/NFV session there
- Need volunteers to collect input on different areas
 - Cloud Native Networking (Cumulus Linux, ONOS, CNI, etc.)
 - SmartNICs
 - Programmable Networks (R&E plans)
- Identify potential areas for further work
 - Performance studies (incl. smartNICs)
 - Programmable Networks Prototypes
 - DCI testing/evaluation
 - WLCG DOMA activities
 - etc.
- Always looking for feedback and additional volunteers/sites for help.

Summary

- Explored several existing SDN/NFV approaches and use cases
- In Cloud Native Networking focusing primarily on Tungsten Fabric, OVS, but also looking into commercial projects
 - CERN and Nikhef leading pilot projects for deploying Tungsten Fabric
- SENSE and BigData Express leading projects in programmable networks and data transfers
- Surveying R&E plans for higher-level services
- Introductory tutorial and material to help sites establish their testbed ready
- We plan to finalise white-paper process at the coming F2F (co-located with LHCOPN/LHCONE in Umea)
- Always looking for additional volunteers/sites for help

- Please see references for additional links

References

Whitepaper draft -

[https://docs.google.com/document/d/1w7XUPxE23DJXn--j-M3KvXlfXHUnYgsVU](https://docs.google.com/document/d/1w7XUPxE23DJXn--j-M3KvXlfXHUnYgsVU<hBpKFyjUQ/edit?usp=sharing)

WG meetings and notes: <https://indico.cern.ch/category/10031/>

SDN/NFV Tutorial: <https://indico.cern.ch/event/715631/>

Tungsten Fabric architectural overview:

<https://tungstenfabric.github.io/website/Tungsten-Fabric-Architecture.html>

OVN/OVS overview: <https://www.openvswitch.org/>

2018 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS) –

<http://conferences.computer.org/scw/2018/#!/toc/3>

Backup slides

Networking Challenges

- Capacity/share for data intensive sciences
 - No issues wrt available technology, however
 - What if N more HEP-scale science domains start competing for the same resources ?
- Remote data access proliferating in the current DDM design
 - Promoted as a way to solve challenges within experiment's DDM
 - Different patterns of network usage emerging
 - Moving from large streams to a mix of large and small frequent event streams
- Integration of Commercial Clouds
 - Impact on funding, usage policies, security, etc.
- Technology evolution
 - Software Defined Networking (SDN)/Network Functions Virtualisation (NFV)

Technology Impact

- Increased importance to oversee network capacities
 - Past and anticipated network usage by the experiments, including details on future workflows
- New technologies will make it easier to transfer vast amounts of data
 - HEP quite 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 (see later), we know that software will play a major role in the networks of the future
 - We have an opportunity here
- It's already clear that software will play major role in networks in the mid-term
- Important to understand how we can design, test and develop systems that could enter existing production workflows
 - **While at the same time changing something as fundamental as the network that all sites and experiments rely upon**
 - We need to engage sites, experiments and (N)REN(s) in this effort

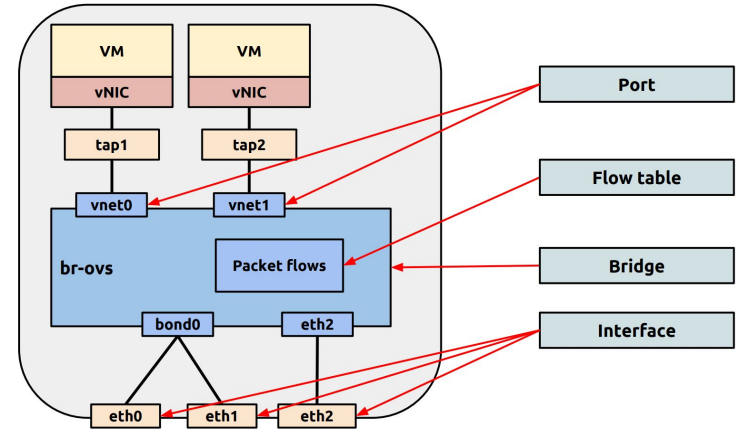
Software Defined Networks (SDN)

- Software Defined Networking (SDN) are 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

Software Switches

Open vSwitch (OVS) - open source multilayer virtual switch supporting standard interfaces and protocols:

- OpenFlow, STP 802.1d, RSTP,
- Advanced Control, Forwarding, Tunneling
- Primarily motivated to enable VM-to-VM networking, but grew to become the core component in most of the existing open source cloud networking solutions



Runs as any other standard Linux app - user-level controller with kernel-level datapath including HW off-loading (recent) and acceleration (Intel DPDK)

Enables massive network automation ...

Open vSwitch Features

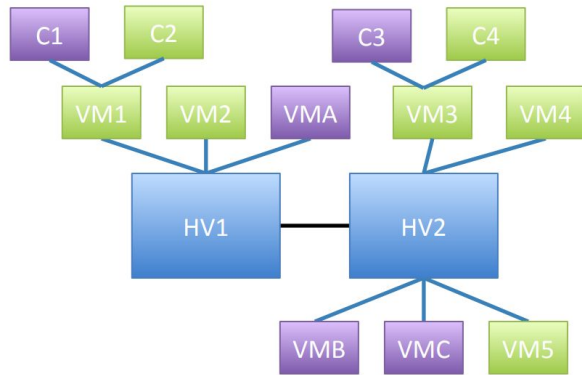
- Visibility into inter-VM communication via NetFlow, sFlow(R), IPFIX, SPAN, RSPAN, and GRE-tunneled mirrors
- LACP (IEEE 802.1AX-2008)
- Standard 802.1Q VLAN model with trunking
- Multicast snooping
- IETF Auto-Attach SPBM and rudimentary required LLDP support
- BFD and 802.1ag link monitoring
- STP (IEEE 802.1D-1998) and RSTP (IEEE 802.1D-2004)
- Fine-grained QoS control
- Support for HFSC qdisc
- Per VM interface traffic policing
- NIC bonding with source-MAC load balancing, active backup, and L4 hashing
- OpenFlow protocol support (including many extensions for virtualization)
- IPv6 support
- Multiple tunneling protocols (GRE, VXLAN, STT, and Geneve, with IPsec support)
- Remote configuration protocol with C and Python bindings
- Kernel and user-space forwarding engine options
- Multi-table forwarding pipeline with flow-caching engine
- Forwarding layer abstraction to ease porting to new software and hardware platforms

Controllers - Open DayLight

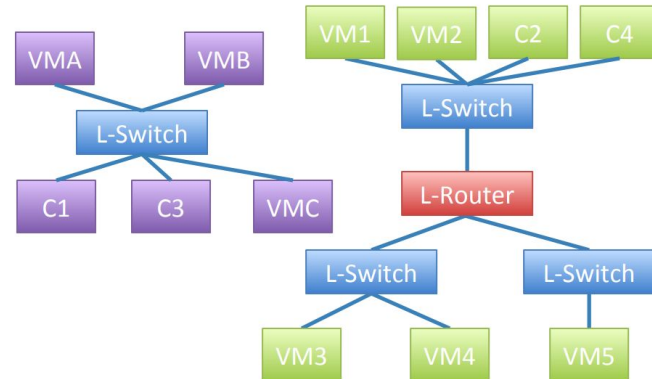
- Modular open platform for customizing and automating networks of any size and scale. Core use cases include:
 - **Cloud and NFV** - service delivery on cloud infrastructure in either the enterprise or service provider environment
 - **Network Resource Optimisation** - Dynamically optimizing the network based on load and state; support for variety of southbound protocols (OpenFlow, OVSDDB, NETCONF, BGP-LS)
 - Automated Service Delivery - Providing on-demand services that may be controlled by the end user or the service provider, e.g. on-demand bandwidth scheduling, dynamic VPN
 - Visibility and Control - Centralized administration of the network and/or multiple controllers.
- Core component in number of open networking frameworks
 - ONAP, OPNFV, OpenStack, etc.
- Integrated or embedded in more than 50 vendor solutions and apps
- ODL is just one of [many](#) controllers that are available:
 - OpenContrail, ONOS, MidoNet, Ryu, etc.

Controllers - Open Virtual Network (OVN)

- Open source logical networking for OVS
- Provides L2/L3 networking
 - Logical Switches; L2/L3/L4 ACLs
 - Logical Routers, Security Groups
 - Multiple Tunnel overlays (Geneve, VXLAN)
 - Top-of-rack-based & software-based physical-to-logical gateways



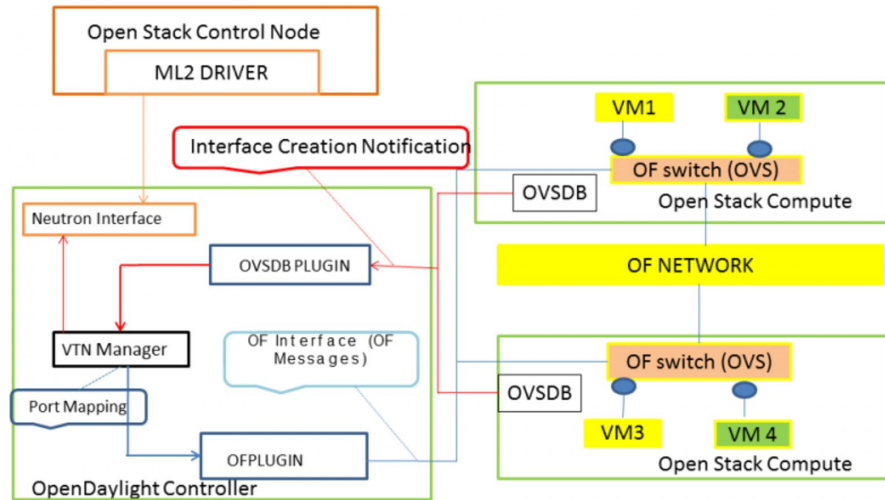
Physical



Logical

Cloud Compute - OpenStack Networking

- Cloud stresses networks like never before
 - Massive scale, Multi-tenancy/high density, VM mobility
- OpenStack Neutron offers a plugin technology to enable different (SDN) networking approaches - brings all previously mentioned techs together



ML2 driver is what makes controllers pluggable, so you can easily replace Neutron controller with OpenDaylight, OVN, etc.

Both generic and vendor-specific [plugins](#) are available

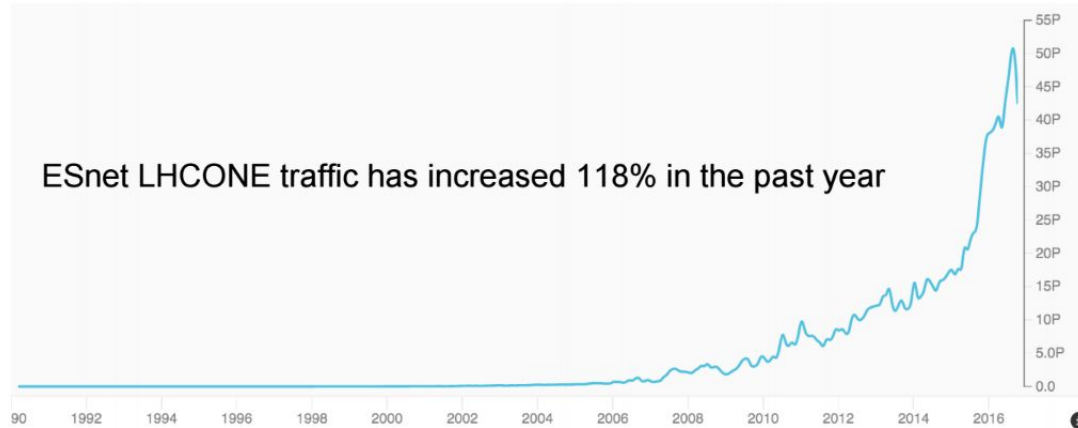
Cumulus Linux

- **Alternative to OVS** - uses separate apps/kernel functions to program different functionality such as STP/RSTP (mstpd), VXLAN (ifupdown2), VLAN (native linux bridge) etc.
- It does contain OVS to enable integration with controllers:
 - VMware NSX, Midokura Midonet, etc.
- Unlike OVS, Cumulus Linux is not an app, but a distribution, which is certified to run on bare metal switches
 - The list of supported HW is at [\(https://cumulusnetworks.com/products/hardware-compatibility-list/\)](https://cumulusnetworks.com/products/hardware-compatibility-list/)
 - Mainly Broadcom Tomahawk, Trident2/+, Helix4 and Mellanox Spectrum ASICs
- Otherwise runs like standard Linux, which means compute and network “speak the same language”
 - E.g. automation with Ansible, Puppet, Chef, etc.

R&E Traffic Growth Last Year

ESnet Traffic Volumes

LHCONE represents more than 32% of ESnet accepted traffic



◀ August 2016 ▶

| | Bytes | Percent of Total | One Month Change | One Year Change |
|----------------|----------|------------------|------------------|-----------------|
| OSCARS | 11.22 PB | 26.5% | -8.93% | +19.7% |
| LHCONE | 13.7 PB | 32.3% | -25.6% | +118% |
| Normal traffic | 17.46 PB | 41.2% | -15.1% | +29.7% |
| Total | 42.38 PB | | -17.4% | +45.5% |

[Slide from Michael O'Connor, LHCONE operations update](#)

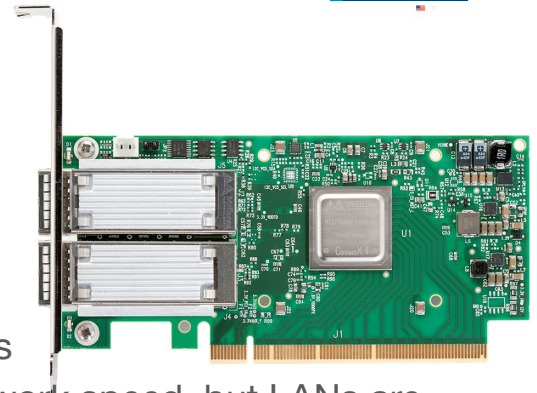
In general, ESNet sees overall traffic grow at [factor 10 every 4 years](#). Recent LHC traffic appears to match this trend.

[GEANT](#) reported LHCONE peaks of over 100Gbps with traffic increase of 65% 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.**

WAN vs LAN capacity

- Historically WAN capacity has not always had a stable relationship compared to data-centre
 - 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 speed, but LANs are also getting in the same range
 - List price for 100Gbit dual port card is ~ \$1000, but significant discounts can be found (as low as \$400), list price for 16 port 100Gbit switch is \$9000
- Today it is easy to over-subscribe WAN links
 - in terms of \$ of local hardware at many sites
- Will WAN be able to keep up ? **Likely yes**, however:
 - We did benefit from the fact that 100Gbit was deployed on time for Run2, might not be the case for Run3 and 4
 - By 2020 800 Gbps waves likely available, but at significant cost since those can be only deployed at proportionally shorter distances
- Planning of the capacities and upgrades (NREN vs sites) will be needed



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)
 - Early testing from ESNet less conclusive and questions need answering

R&E Networking

- 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
- Big data analytics requiring high throughput no longer limited to HEP
 - 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](#)
- **What if N more HEP-scale science domains start competing for the same network resources ?**
 - Will HEP continue to enjoy “unlimited” bandwidth and prioritised attention or will we need to compete for the networks with other data intensive science domains ?
 - Will there be **AstroONE**, **BioONE**, etc., soon ?

Tech Trends: SD-WAN

- Large Network as a Service providers include several well established CSPs such as Amazon, Rackspace, AT&T, Telefonica, etc.
- Recently more niche NaaS providers have appeared offering SD-WAN solutions
 - Aryaka, Cloudgenix, Pertino, VeloCloud, etc.
 - Their offering is currently limited and not suitable for high throughput, but evolving fast
- SD-WAN market is estimated to grow to \$6 billion in 2020 (sdxcentral)
- Will low cost WAN become available in a similar manner we are now buying cloud compute and storage services ?
 - Unlikely, our networks are shared, not easy to separate just LHC traffic
 - Transit within major cloud providers such as Amazon currently not possible and unlikely in the future, limited by regional business model - but great [opportunity for NRENs](#)

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 network analytics of LHCOPN/LHCONE perfSONAR data also point out some very interesting facts:
 - Packet loss greater than 2% for a period of 3 hours on almost 5% of all LHCONE links
- Network telemetry (real-time network link usage) likely to become available in the mid-term (but likely not from all NRENs at the same time)
- It is increasingly important to focus on site-based network operations