HEPiX Network Functions Virtualisation WG Update

Marian Babik, Shawn McKee

HEPiX, 15 May 2018

Working Group Motivation

- Massive network automation is possible - in production and at large-scale
- Existing technologies offer novel ways on how to look at networks in DC
  - Bare metal is still important, but considerable amount of virtualization is moving up the stack
  - This means potential cost reduction and moving away from vendor lock-in
  - But will also mean considerable changes in how we design, operate and manage networks
- **OpenStack** and **Docker** are leading the way
  - Some major cloud providers are already running NFV in production and at large-scale
- Most of the T1 and big T2 sites will need to look into NFV
  - To enable multi-tenancy, improve bandwidth allocation, streamline integration with compute and storage
  - Sites have number of options to choose from therefore sharing experiences and knowledge would be clearly beneficial
  - Virtualisation of DC networking will facilitate integration with the existing SDN WAN projects - we should aim to work together with (N)RENs to understand requirements and draft roadmap for future production systems

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

Network Functions Virtualisation WG

Mandate: Identify use cases, survey existing approaches and evaluate whether and how SDN/NFV should be deployed in HEP.

Team: 48 members including R&Es (GEANT, ESNet, Internet2, AARNet, Canarie, SURFNet, GARR, JISC, RENATER, NORDUnet), sites (ASGC, PIC, BNL, CNAF, CERN, KIAE, FIU, AGLT2, Caltech, DESY, IHEP, NIKHEF) and also one company CORSA

Monthly **meetings** started in January (https://indico.cern.ch/category/10031/)

Mailing list: https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepix-nfv-wg

Objectives/sub-tasks

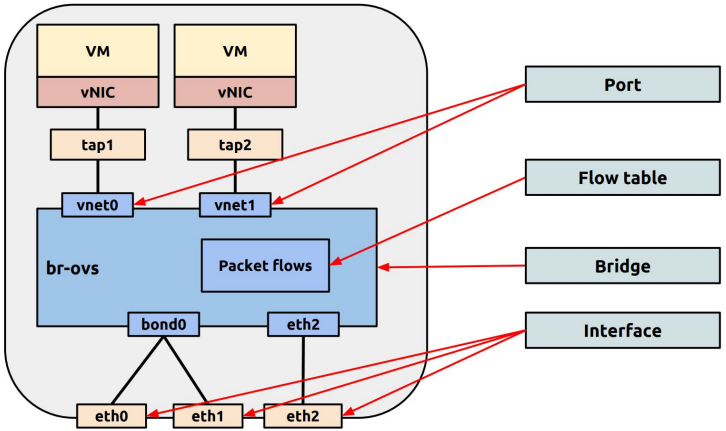
- Proposing **two phases**, phase I (exploratory):
  - Define use cases
    - Explore SDN/NFV approaches for compute, e.g. OpenStack/Kubernetes (mainly **intra-site** activities)
    - Explore SDN/NFV approaches for distributed storage/end-to-end transfers, e.g. data lakes (**inter-site** activity in collaboration with RENs/NRENs)
  - 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 by Q1 2019 - interim before 2019 Spring HEPiX
  - 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

SDN/NFV 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)

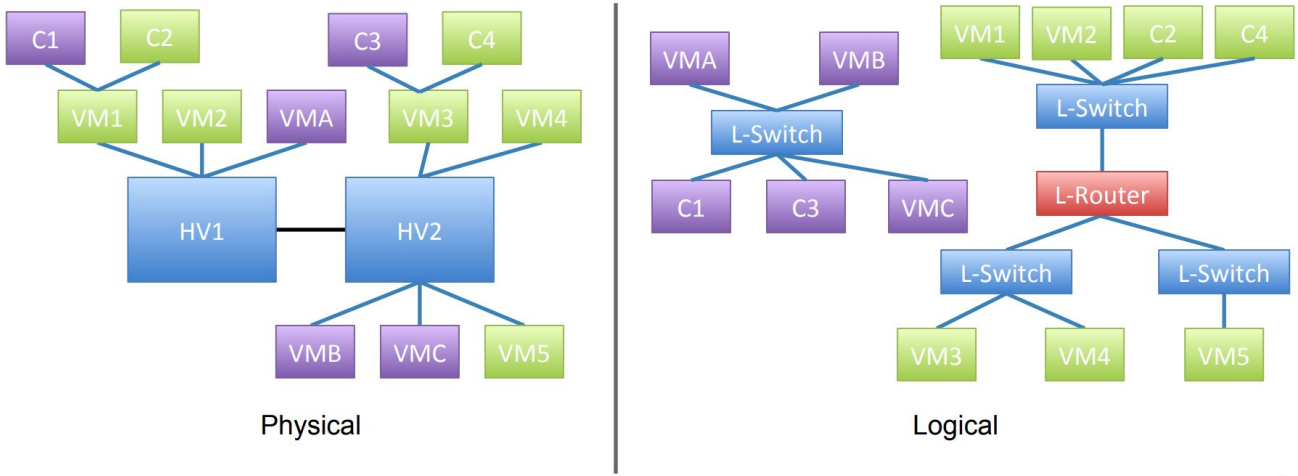
Enables massive network automation …

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, OVSDB, 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.

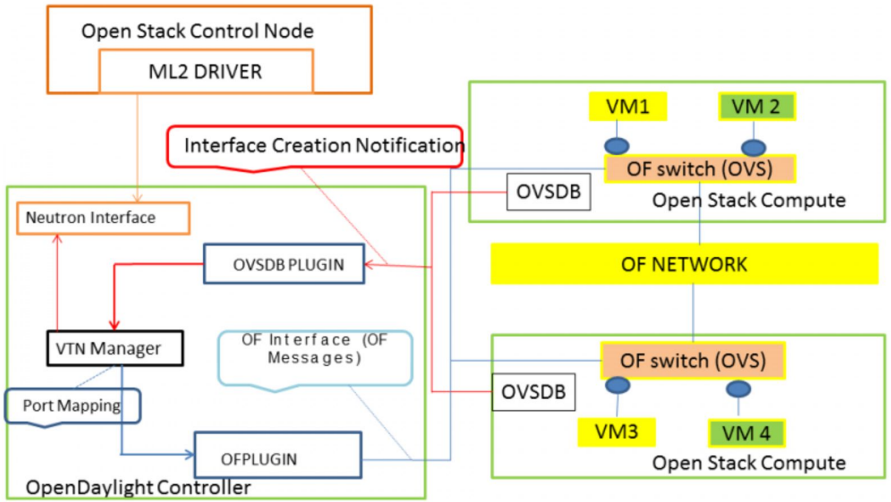
Controllers - Open Virtual Network (OVN)

- 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



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.

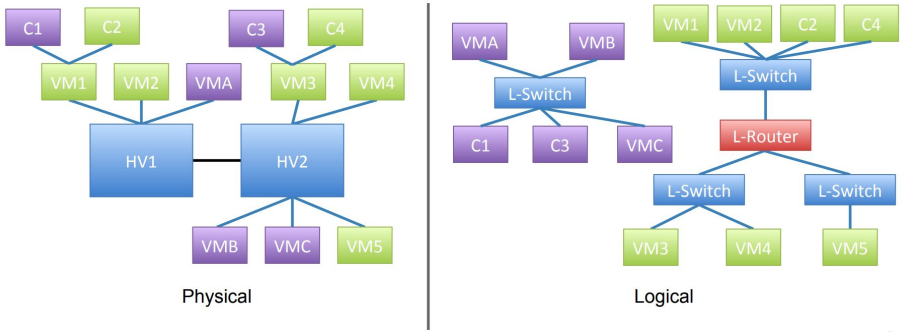
Both generic and vendor-specific plugins are available

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/)
  - 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.

SDN/NFV approaches for compute

- Agile service delivery on cloud infrastructures usually delivered via Network Functions Virtualisation (NFV) - with underlays such as **OpenStack**, **Kubernetes** and **Docker**
- Organise and effectively manage data-center (DC) networking using SDN/NFV, potential areas to look at:
  - On-demand network services (bandwidth scheduling, dynamic VPN), multi-tenancy support, allocation of network resources, bridging legacy and software-defined networks



- Explore existing technologies and tools and understand how they could be best deployed in HEP infrastructure DCs
- Significant interest from several sites to explore this area

SDN/NFV approaches for end-to-end transfers

- Network resource optimisation - dynamically optimising the network based on its load and state. Optimising the network using near real-time traffic, topology and equipment.
- Enable application workflows to drive network service provisioning
  - Dynamic interaction btw. network resident services and data distribution and management
  - Smart-middleware interfacing network paths with guaranteed network bandwidth to a set of high performance end-host DTNs
- Improving end-to-end transfer by optimising the network flow
  - Instrumenting storage systems with software switches that can be remotely controlled to smooth existing traffic
- Existing projects in **ATLAS** (SDN btw AGLT2/MWT2), **CMS** (SDN-NGenIA, **SENSE**), SDN aspects also in NSF-funded **SLATE** and **OSIRIS**
- We envision an operational model that coordinates use of network, CPU and storage resources, as it also seeks to improve workflow.

Plans

Introduction to SDN/NFV with hands on focused on OpenVSwitch (OVS)/Open Virtual Networking (OVN), VXLAN, integration/usage with OpenStack/Kubernetes was presented in the last meeting (https://indico.cern.ch/event/715631/)

We'd like to keep topical monthly meetings/discussions and organise at least one F2F during the year. The following topics are planned:

- **OpenContrail/Tungsten** (experiences from CERN and NIKHEF)
- **OpenDaylight/OpenVSwitch** (Caltech)
- **ATLAS SDN** experiments (AGLT2, KIT)

Meetings are announced to the NFV mailing list, and meeting notes will be posted there as well. Everyone interested is welcome to attend. We welcome any feedback and suggestions on how to improve. **Please join us!**

Questions ?

Meetings: https://indico.cern.ch/category/10031/ (live notes are attached)

Mailing list to join: https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepix-nfv-wg

F2F meeting is planned to be co-located with LHCOPN/LHCONE workshop, which will take place at FNAL in October 2018

Backup slides

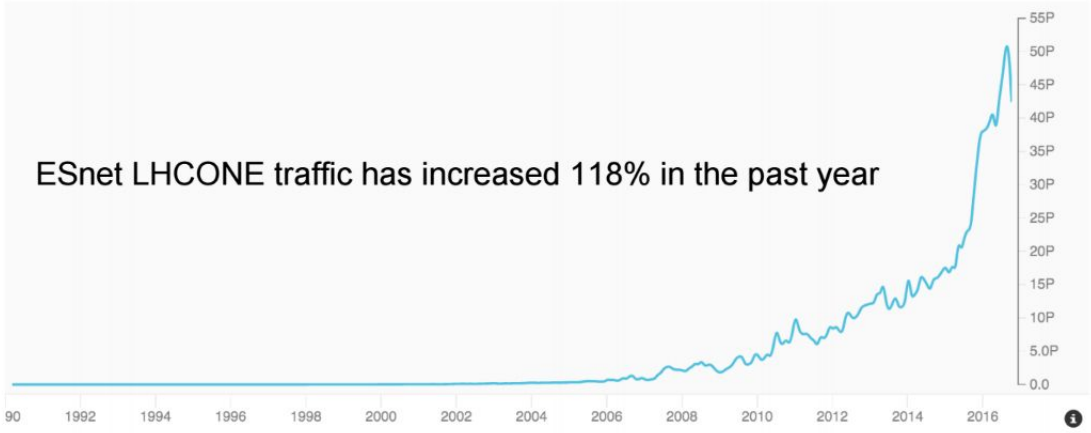
Open vSwitch Features

- 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

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

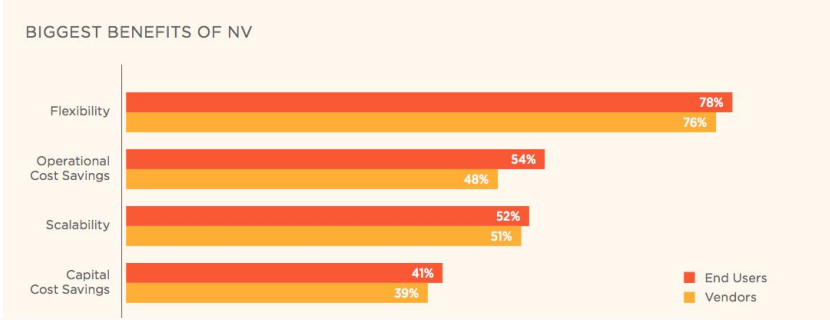
- TCP more stable in CC7, throughput ramp ups much quicker
  - Detailed report available from Brian Tierney/ESNet
- Fair Queueing Scheduler (FQ) available from kernel 3.11+
  - Even more stable, works better with small buffers
  - Pacing and shaping of traffic reliably to 32Gbps
- Best single flow tests 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

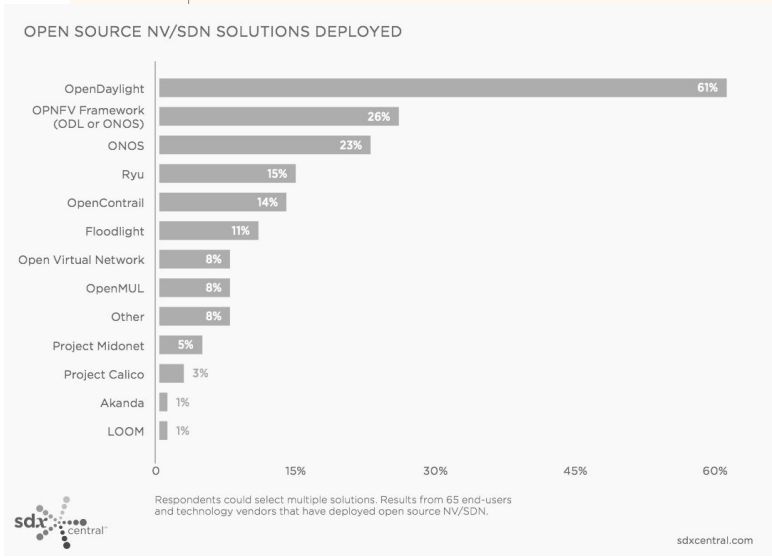
R&E Networking

- 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: Software Defined Networks (SDN)

- SDN is a set of technologies offering solutions for many of the future challenges
  - Current links can handle ~ 6x more traffic if we could avoid peaks and be more efficient
  - SDN driven by commercial efforts
- 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
  - Will experiments have effort to engage in the existing SDN testbeds to determine what impact it will have on their data management and operations ?





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