High Speed Scientific Data Transfers using Software Defined Networking

Harvey Newman
Caltech
+1 626 395 6656
newman@hep.caltech.edu

Azher Mughal
Caltech
+1 626 395 8758
azher@hep.caltech.edu

Dorian Kcira
Caltech
+1 626 395 2562
dkcira@caltech.edu

Iosif Legrand
Caltech
+1 626 395 6656
Iosif.Legrand@cern.ch

Ramiro Voicu
Caltech
+1 626 395 6656
Ramiro.Voicu@cern.ch

Julian Bunn
Caltech
+1 626 395 6656
Julian.Bunn@caltech.edu

* 1200 E California Blvd., Pasadena CA 91125

*Additional collaborators are listed at the end of this paper

ABSTRACT

The massive data volumes acquired, simulated, processed and analyzed by globally distributed scientific collaborations continue to grow exponentially. One leading example is the LHC program, now at the start of its second three year data taking cycle, searching for new particles and interactions in a previously inaccessible range of energies, which has experienced a 70% growth in peak data transfer rates over the last 12 months alone. Other major science programs such as LSST and SKA, and other disciplines ranging from earth observation to genomics, are expected to have similar or great needs than the LHC program within the next decade. The development of new methods for fast, efficient and reliable data transfers over national and global distances, and a new generation of intelligent, software-driven networks capable of supporting multiple science programs with diverse needs for high volume and/or real-time data delivery, are essential if these programs are to continue to progress, and meet their goals. In this paper we describe activities of the Caltech High Energy Physics team and collaborators, related to the use Software Defined Networking to help achieve fast and efficient data distribution and access. Results from Supercomputing 2014 are presented together with our work on the Advanced Network Services for the Experiments project, and a new project developing a Next Generation Integrated SDN Architecture, as well as our plans for Supercomputing 2015.

INDIS2015, November 15-20, 2015, Austin, TX, USA

ISBN 978-1-4503-4002-1/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2830318.2830320

Keywords

CMS, LHC, ANSE, SENSE, SDN-NGeNIA, PhEDEx, SC15, OpenFlow, SDN.

1. INTRODUCTION

Scientific innovation continues to exponentially increase the production of valuable research data. Exchange of this information typically involves the worldwide network infrastructure. To ensure success for large, high priority uses, such as those found in the Large Hadron Collider (LHC) community, the California Institute of Technology, Michigan, the University of Texas at Arlington and Vanderbilt University are collaborating on the ANSE [1] (Advanced Network Services for Experiments) Project. This project targets the integration of network awareness and strategic network resource use and management (in cooperation with the R&E network community) into the ATLAS and CMS data distribution and management software infrastructures.

The project's goal is to integrate network monitoring and network provisioning capabilities with the software stacks of the CMS and ATLAS experiments at the LHC. This is achieved by enabling more deterministic time to completion for a designated set of data transfers. Specifically, the project is developing a library of network-related functions that will provide the ATLAS PANDA and CMS PhEDEx systems access to network status, topology, metrics and management capabilities.

The year 2014 saw rapid evolution in Software Defined Networking (SDN). The ANSE team has adapted to and remained at the forefront of these developments. It has worked in the Floodlight (2013-14) and subsequently in the Open Daylight framework, developing intelligent path selection methods across complex networks supporting multiple data transfer requests. Each is carried out with multipath TCP and Caltech's Fast Data Transfer (FDT) application.

This work also builds on the results of the DOE/ASCR-funded OLiMPS (OpenFlow Link-level Multipath Switching) and a Cisco Research microgrant.

In the dynamic circuit domain, a complementary development to this is the migration of the dynamic circuits used by PhEDEx and

PanDA, two of the main data distribution and management systems of CMS and ATLAS respectively to NSI, when and where possible

In the area of network monitoring we have been working closely with the WLCG Network and Transfer Metrics Working Group (responsible for managing the acquisition of network and transfers metrics globally for LHC[1]) to ensure we have a reliable source of data for ANSE to process for use by ATLAS and CMS. A big component of this is the developing Open Science Grid (OSG) network datastore which is collecting and providing access to perfSONAR data measurements for all of OSG and WLCG. Additional sources of information will come from the global Xrootd testing framework enabled for both ATLAS and CMS and the "Sonar" (test data transfer) infrastructure used in ATLAS. ANSE is working on validating and summarizing this data to provide input to higher level services which need to make network path based decisions.

In this paper we describe the ambitious set of demonstrations at the Supercomputing 2014 (SC 14) conference, and how the use of these new strategic methods of network provisioning and workflow optimization will continue to be a strong theme of our collaboration in 2015. We describe the plans for several future activities in this domain. Recent specific work items include (1) adaptations that use widely used data transfer applications, such as FTS that manages transfers using gridftp, (2) site orchestration that provides stable QoS and rate limits at the site edge up to wire speed using Open vSwitch, (3) storage to storage dynamic circuits including flow matching to scattered sets of source and destination IP addresses, as is typical in LHC dataset transfers, (4) porting our multipathing Open Daylight controller to the Lithium release, (5) comparative experience with ONOS, in collaboration with AmLight (FIU) and on the ESnet testbed, and (6) development of the next generation SDN-integrated global system concept based on the above ongoing developments and series of tests.

2. Experience with the OpenDaylight Controller at Supercomputing 2014

During SC14 we tested the Hydrogen release of OpenDaylight, a community open-source software framework for controlling Software Defined Networks, particularly switchgear running the OpenFlow protocol.

2.1 Topology and Tests

The Supercomputing 2014 setup (Figure 1) demonstrated a system capable of efficiently moving LHC datasets among different LHC data centers, as well as among three end-points set up at the exhibition show floor. It featured several state-of-the art techniques and technologies:

- An OpenDaylight SDN controller providing smart load balancing of data flows, described in detail below.
- PhEDEx – the CMS data transfer management software enhanced through bandwidth reservation capability; the interface to the OSCARS reservation system has been implemented as part of the NSF funded ANSE project.
- PanDA – the scientific workload management system, originally developed in the ATLAS experiment.
- Multi-100Gbps WAN connectivity. 4x100G between the three show-floor booths: Caltech, iCAIR, MSU/UMICH, (1 Tbps possible) Plus a total of 4 100Gbps WAN uplinks providing connectivity to at least the following sites: Caltech, University of Victoria, University of Michigan, CERN, SPRACE.

For the OpenDaylight tests we used Brocade MLXe16 switches at the Caltech, iCAIR and Vanderbilt booths, together with an Extreme switch collocated at the Caltech booth. The setup and interconnects are shown in Figure 2.

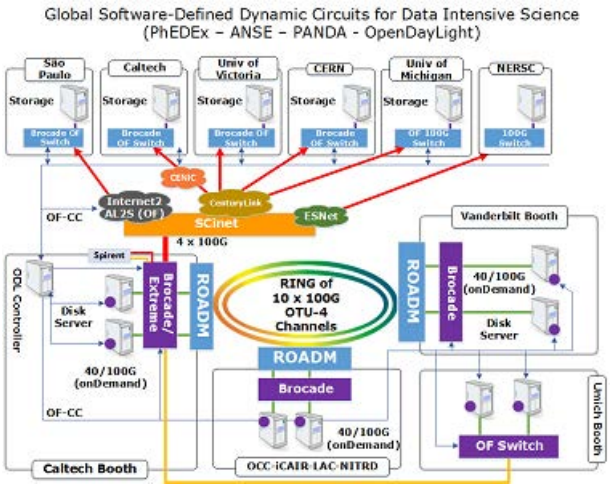


Figure 1: Global Software-Defined Dynamic Circuits for Data Intensive Science – Local and external connectivity.

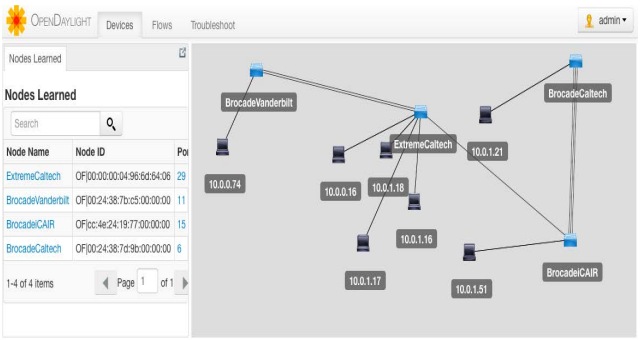


Figure 2: The OpenFlow switches and attached hosts used for the OpenDaylight tests at SC14.

The topology included a pair of 100Gbit links between the Vanderbilt Brocade and the Caltech Extreme, a single 100Gbit link between the Extreme and the iCAIR Brocade, and three 100 Gbit links between the iCAIR and Caltech Brocades. Each switch was configured to talk to the ODL controller, which was running in a dedicated Virtual Machine on a host at the Caltech booth.

The purpose of the OpenDaylight (ODL) tests was to evaluate the performance of plugins, developed at Caltech, that provide reactive intelligent routing of traffic between endpoints in the network. These plugins, Multipath and Multipath Northbound[2] offer control over several routing strategies, as well as RESTful endpoints for control and monitoring of the network. In particular, the reactive mode of the Multipath plugin will respond to emergent traffic flows by installing flow rules in the switches along the dynamically chosen path for the flow. The path selection algorithms that were tested at SC14 included:

[1] https://twiki.cern.ch/twiki/bin/view/LCG/NetworkTransferMetrics

[2] http://pcbunn.cacr.caltech.edu/multipath/northbound/doc/index.html and http://pcbunn.cacr.caltech.edu/multipath/doc/index.html

- Round Robin – each new traffic flow is assigned to the next possible path between source and destination
- Shortest Path – each flow is assigned to the shortest path (the least number of hops)
- Available Bandwidth – each flow is assigned to the path with the largest available bandwidth on the busiest hop
- Random Path – each flow is assigned to one of the available paths at random

These are but a few of the available selectors provided – see the documentation for more details.

To test the various selectors, the selector required was first configured using the Northbound API's REST interface, and then traffic was started between hosts using either Iperf or FDT, depending on the test. Each switch, on reception of a new frame from the source host, matches the frame against any existing flow entries in its switch tables. If the frame matches, the switch sends the packet on its way, as defined by the matching rule. If there is no match, the switch "punts" the packet to the ODL controller for inspection.

On reception of a punted packet/frame, the Multipath plugin decides whether to process the packet, depending on its protocol. For these tests ICMP and TCP packets in Ethernet were processed – other punted packets were ignored. For both types of processed packets, the source and destination host address were extracted, together with their incoming outgoing switch ports. For TCP traffic, the source and destination TCP port numbers were optionally extracted for creating flow rules. (In most tests, the source TCP port was used to distinguish the flows.)

Once the source and destination details for the punted packet were extracted, the Multipath plugin then used the current path selection algorithm to determine a path between the source and destination. Then, flow rules for each switch on the path were defined and accumulated in a list, before finally being sent to ODL's "FlowManager" module for writing to the switches themselves. Extensive error checking is used to ensure that only complete, correct paths are left active in the switchgear.

The flow entries themselves used idle timeouts of either 10 seconds or 60 seconds (the latter being used early on during initial testing).

2.2 Results and Issues

Several FDT flows were set up between Vanderbilt, iCAIR and Caltech following path selection with Multipath's available bandwidth algorithm. In Figure 3 the network traffic is shown filling the 40 Gbit/sec host adapters on a pair of nodes at iCAIR and Caltech. This traffic is the result of an FDT flow set up using the Multipath system.

There were some SDN and network issues faced during the tests:

1. The hardware configuration, and the switch ports assigned to OpenFlow control in the switches were dynamic during setup of the equipment, which made reliable testing challenging.
2. We observed some odd behavior of the firmware in the Extreme switch (e.g. punted packets matching existing flows, idle timeout values on flows not being respected) which we diagnosed with the help of Extreme engineers at SC14.
3. Patches were needed to the Brocade firmware to support some of the ODL switch commands, otherwise flow statistics were not available to ODL.
4. It was intended to include a remote Brocade (at the University of Victoria in BC) in the ODL topology, but problems with the layer 1 optical link between there and New Orleans prevented this.

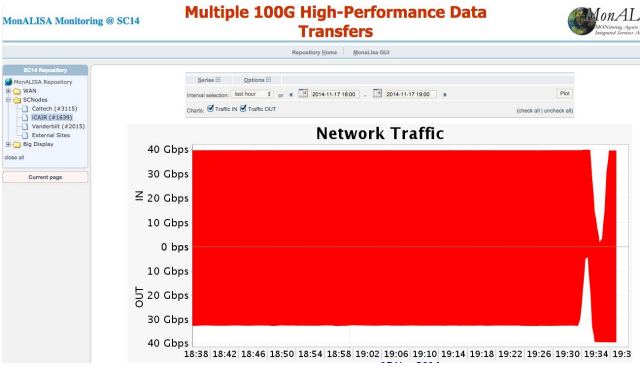


Figure 3: 40Gbit/sec traffic flow between hosts and the iCAIR and Caltech booths

3. Advanced Network Services for Experiments (ANSE)

The goal of the ANSE project (2012-15) is to improve the overall working efficiency of major science programs, starting with the CMS and ATLAS experiments, by integrating network-awareness and functions into their middleware stacks. For CMS, this has meant enabling control of virtual network circuits in PhEDEx, the CMS data-transfer management system, in order to achieve more deterministic transfer times. Members of the Caltech and Princeton teams in ANSE have enhanced PhEDEx, allowing it to create, use and destroy circuits according to its own needs.

One of the main components in PhEDEx is the ResourceManager, which was designed with reusability in mind. While PhEDEx can make direct calls to the ResourceManager, it can also be controlled by external applications via a REST interface. Additionally, the PhEDEx framework supports a plug-in system for the circuit backends, meaning it is not limited to using NSI circuits – if there is a need to support a different circuit provider, it can be easily added as a different plug-in to this system.

A proof-of-concept prototype has been built. It shows that PhEDEx can cleanly exploit layer 3 circuits, switching to take advantage of them when they exist, switching back to using the general purpose network when they go away, with no degradation in transfer quality as a result of the switchover. From this prototype, a refactored, production-ready version has been built. This is not PhEDEx-specific, nor even CMS-specific or HEP-specific, and can be used to bridge the gap between experiment middleware and a network that supports layer 2 layer 3 circuits. To fully exploit layer 2 circuits, more work must be done to fully extend them from storage-to-storage. A practical solution that will be developed and brought into production in 2016 has been identified, as described in the following sections.

3.1 Storage-to-Storage Circuits

Currently, there are no production-ready solutions which can set up a virtual circuit from storage to storage. NSI for example, only creates layer 2 circuits, which end at the site's border router, leaving the challenge of setting up the last mile to the sites themselves. In order to make efficient use of circuits, an automated solution in needed.

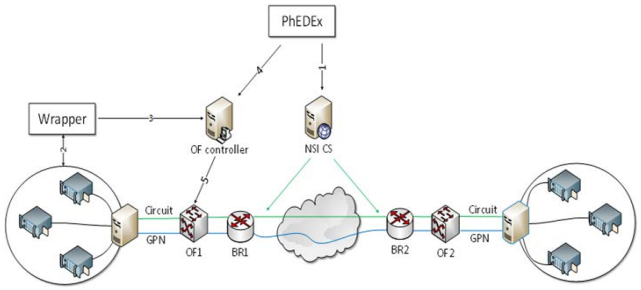


Figure 4: Detailed architecture view in which the ANSE framework is used to request an NSI circuit, with the last mile setup being handled by an OpenFlow solution

One the first of the challenges in automating this setup is knowing between which servers (or pairs of servers) a path needs to be created. This is not trivial since the middleware (PhEDEx) is too high level to provide such detailed information. PhEDEx only knows that two sites are connected, not the underlying network topology or configuration. Also, the source and destination servers involved in a given dataset transfer may be scattered within or among several subnets.

Moreover, when issuing a transfer, PhEDEx works with Storage URLs (SURLs), which only point to the hostname of the storage farm. The physical location of the files are given by Transfer URL (TURLs) and are chosen by the transfer stack itself from a set of available replicas. A wrapper has been proposed to deal with this issue. It will retrieve the IPs of the storage servers involved in a PhEDEx transfer and relay that information to an OpenFlow (OF) controller. The wrapper uses information about the current

PhEDEx transfers, which is provided directly by PhEDEx via the ResourceManager.

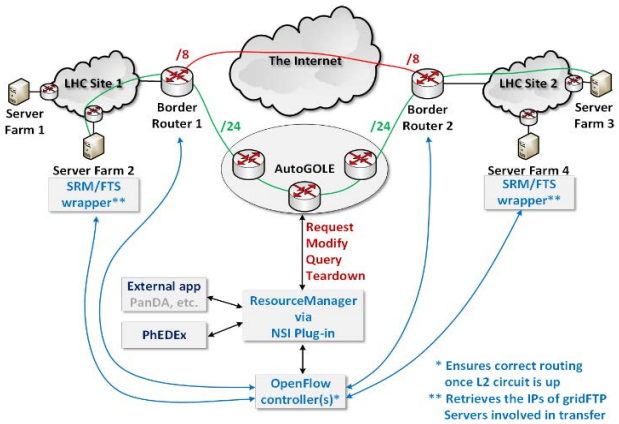


Figure 5: Global architecture of an SDN-NGenIA system in which the framework developed in ANSE requests NSI circuits with the last mile being handled by a generic layer 3 solution (OpenFlow based or BGP)

Figure 4 shows the interaction between the components:

1. PhEDEx requests a circuit (NSI or OSCARS [4]) via the ResourceManager
2. A wrapper to the transfer stack (FTS) locates all the storage servers involved in the transfer.
3. The wrapper informs the OF controller that a given list of servers will start transferring data.
4. Once the full circuit setup is complete PhEDEx informs the OF controller.
5. The OF controller installs flow rules on OF switches, routing only traffic from the set of IPs received from the wrapper, onto the new layer 2 path created via the NSI circuit.

As a first step towards the next generation SDN-NGeNIA system described below in Section 4, we have proposed the architecture shown in Figure 5. The figure illustrates switching a preferred dataset transfer from the default path across R&E networks (in red) to a dedicated path with guaranteed bandwidth (in green), implemented by use of NSI dynamic circuits. Here two LHC sites are shown, each with storage farms attached. In the example shown, the ResourceManager is be used to request a dedicated path, shown going through the AutoGOLE [2] which is the current focus of point-to-point dynamic circuit development in LHCONE [3]. This is not a pre-requisite though – the implementation will cover the creation of circuits in any infrastructure supporting OSCARS/NSI [4].

The developments proposed in this project are essential to make the dynamic circuits services developed in ANSE and other projects fully functional at many sites, and to ensure that these services are well-adapted to the US CMS and US ATLAS data transport services and overall workflow. SDN-NGenIA also will support the concept of the next generation Science DMZ [5] as part of a modern research-oriented laboratory and campus infrastructure. This will ensure that the services deployed can be scaled out to serve hundreds of sites.

The SDN application involved the Fast Data Transfer Agent developed by Caltech [13], OESS [6] and flow space firewall [7] (FSF) software by Internet2. FSF provides a unique but contained environment through network slicing (a *function* of NFV) to any network operator. This sliced topology can be used to configure and connect another set of OpenFlow devices to create an overlay topology under a different administrative domain.

Such an application suite could be used to great benefit in large science projects such as LSST [8] where a single uncompressed image is typically 2.7 GB, and where up to a thousand images will be collected per night. Part of the challenge is that nightly pipelines are based on image subtraction and are designed to rapidly detect interesting transient events in the image stream and send out alerts to the community within 60 seconds from completing the image readout. Such large files can easily be detected using network flows and based on certain intents (e.g. they are needed within a few seconds!) and can be redirected over the high speed 100G routes once those routes are available.

Preparations are underway to extend the present 10G testbed during 2015 using the "OpenWave" multihop 100G path [9] from Miami to Sao Paulo, illustrated in Figure 6, which is being commissioned as of this writing. The network path being deployed is one of the critical design elements of LSST that is planned to be heavily used through 2031, connecting to the NCSA supercomputer and other

astrophysics centers through Internet2 and ESnet. The 100G Openwave testbed will be extended with paths across Internet2/OESS [10] and ESnet to LSST end sites including Fermilab and Argonne. The SDN-NGenIA controller will provide prototypical path selection services in the setup of dynamic circuits across the US, building on the work described in Section 6.1, adapted to the LSST use case.

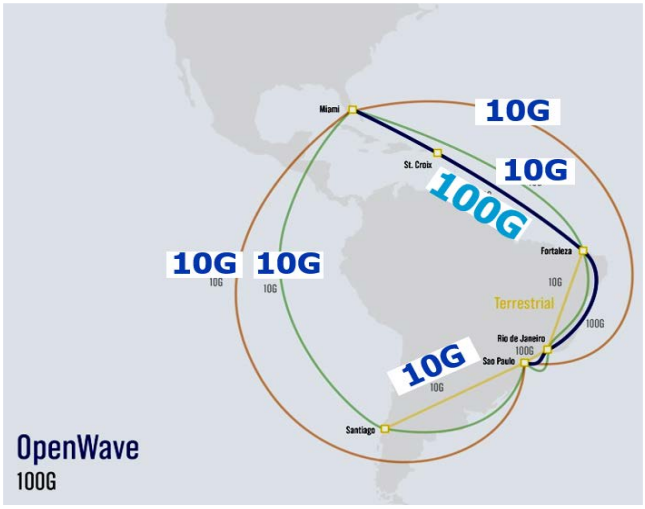


Figure 6: The "Openwave" 100G path being deployed by AmLight and the Brazilian networks, funded by NSF, that is now being commissioning on a Miami-St. Croix-Fortalezza-Rio-Sao Paulo backbone. Starting this summer, this is planned to be used by Caltech together with FIU, ANSP, RNP and GridUNESP

3.2 Intelligent Path Building for CRAB/ASO

The direct stage-out of users' data analysis job output files has been used in CMS since the first versions of the CRAB (CMS Remote Analysis Builder) system that helps manage users jobs and data. Analysis jobs are executed on worker nodes (WN) at the site where the input data is available, and the output files produced are then copied to a pre-defined destination site. Between 10 to 15% of analysis jobs have failed (until recently) when attempting to stage-out the files directly. This has required the design and implementation of Asynchronous Stage Out (ASO), a new component of the latest "CRAB3" analysis data management system architecture that first stages the output locally, then groups the transfers between a given source and destination site, and subsequently copies them to the destination site. CRAB3 stores the output to the local disk and informs the ASO component of a new transfer request by inserting a new entry in a CouchDB database set up for this purpose, and ASO then uses the File Transfer System (FTS) to execute the actual transfers.

As ASO reuses the design and components from PhEDEx, this gives ASO the flexibility to deploy dynamic circuits through the use of the PhEDEx REST API or its components. The Caltech group is currently working on integrating this functionality into ASO, in addition to the use of intelligent path building through the ANSE controller being developed in the Open Daylight Lithium release. This will allow CMS to extend the use of end-to-end dynamic circuits to help with the management of the many tens of thousands of analysis jobs per day in CMS. It will also improve the efficiency of use of computing and storage resources for data analysis.

3.3 Site Orchestration

In-depth site orchestration seen from the site and science program point of view is one of the principal elements of SDN-NGenIA, a next generation SDN architecture being developed by Caltech and partners (see Section 4 below). The Caltech team's earlier work with dynamic circuits in the DYNES project [11] used a so-called "FDTAgent" [12] to couple the data transfer nodes (DTNs) at the end-sites running Caltech's FDT [13] as the high throughput data transfer application. The agent requests the circuit, waits for an answer, configures both end-hosts if the circuit provisioning succeeds, and modifies the local end-host routing including creating VLAN interfaces to use the new circuit. While this provided useful experience, the FDT transfer tool is not used widely yet. Apart from demos, the bandwidth allocated to the circuit was often not guaranteed because campuses did not agree to implement QoS to the end-nodes. The original DYNES and FDT agent work predated ANSE, our hands-on experience with SDN controllers, and the widespread use of Science DMZs.

In seeking standard methods of site orchestration, including solving the campus QoS problem in a generally applicable, transfer-protocol independent way, we decided to investigate Open vSwitch (OVS) [14. OVS, an open source platform-independent multilayer virtual switch, is designed to enable network automation via programmatic interfaces, especially for virtualized environments. It supports standard and well established protocols for management. One of the initial concerns was the performance of OVS on Linux compared to standard Linux bridging. However, starting with OVS version 2.x a series of improvements for the Linux implementation have resulted in OVS performing most of the packet flow processing directly in kernel space [15], instead of extra-routing into user space. The Caltech HEP Networking team conducted a series of OVS tests [16] which has shown that OVS is capable of stable traffic shaping at close to 10G wire speed over wide area networks, with very little CPU overhead. Based on the results of these tests it was concluded that OVS is a good end-point termination point, which enables a full, end-to-end, SDN-capable infrastructure.

From a deployment perspective we envision a seamless migration because the IP reachability will be the same for the end system whether it runs the OVS daemon or not. Another important aspect is that even if OVS can be controlled by means of OpenFlow it does not require a local SDN controller. OVS installation at a site can be done gradually without impacting IP connectivity, since only parts of the cluster need to run the OVS daemon. This enables us to envision a smooth deployment path where OVS can eventually run on all machines, including storage and compute nodes which are usually in the internal campus network, as well as DTNs (Data Transfer Nodes), which are usually located outside the campus firewall, at a number of leading early adopter sites.

There are several aspects of OVS that match the needs of the SDN-NGenIA architecture described in the next section:

- Monitoring via sFlow and/or NetFlow, as well as traffic mirroring (traffic taps)
- SDN-orchestrated configuration for data flows all the way to the end-hosts, which can be orchestrated from the local/campus SDN controller or brought down from a Regional/WAN controller
- QoS and traffic shaping right at the end-point of a data transfer, which is protocol agnostic and thus can be easily adapted for use in applications that use GridFTP, FDT, RDMA or another protocol

- Possible SDN-based configuration for the end-host, which should be controller-agnostic using the standard OpenFlow protocol
- Possibly replacing the standard (iptables) end-host firewall configuration with a more dynamic approach via the local controller. (Although OVS can run fine with local iptables firewall as well.)

An important OVS capability is traffic shaping right at the end-host. During our tests we observed that stable flows at any level, up to line rate if required, and can be turned on and off at will; also with TCP-based data transfers. This has a significant performance advantage in many cases, such as when the upstream switch has small buffers. Another application for the end-host traffic shaping is the possibility to dynamically impose a rate-limit for an aggregated flow (e.g. from multiple storage nodes, or even an entire rack, cluster, etc.) to protect the network infrastructure at the local site in case of big data flows. These aspects are perfectly matched to the SDN-NGenIA architecture, and pave the way to smoother and more predictable high throughput flows across a complex network topology serving multiple major science programs.

The OVS daemon on the end-host enables the campus SDN controller to use standard OpenFlow to configure the local network configuration at both layer 2 and layer 3. Some SDN controllers support other standard protocols to configure the end-host network (NETCONF, OVSDB) as well.

Figure 7 show a simplified generic view of a campus infrastructure. Since each local network may have different policies, and requirements/constraints at different campuses, the architecture leverages standard protocols and interfaces as much as possible without imposing any constraints on the network infrastructure (or policies) at the end-sites. The architecture envisions a standardized Northbound API, which is usually provided by all modern SDN controllers. The NB API will be used by external network services (or applications) to request and allocate network resources, or to request other functions (such as failover among subclusters) from the local SDN controller.

A useful feature of OVS is the possibility to interconnect different instances using standard overlay protocols (such as GRE or VXLAN) which may enable the dynamic connectivity between the local Science DMZ and local storage nodes and/or computing nodes elsewhere on campus. Together with the inter-site virtual circuit services already described, this will allow SDN-NGenIA to develop a rich set of site orchestration services as part of its prototypical architecture.

One challenging and important area of application of OVS is the LHC use case, where a large set of subsystems may have to be dynamically re-configured and tuned simultaneously to deal with changing conditions, such as shifting workloads and/or competing traffic, or errors. Site orchestration using OVS, and inter-site coordination through the Caltech controller and associated host and site agent-based services, and later an emerging network operating system, such as SENOS being developed by ESnet, will be needed to meet this challenge. This experience and the products of this work will subsequently inform, become interfaced, and potentially be integrated into the development of such operating systems in the latter part of the SDN-NGenIA project. The Caltech team has relevant expertise through its longstanding work on pre-SDN (pre-standard) autonomous systems with similar capabilities, which are currently deployed worldwide in the ALICE experiment [17] for example.

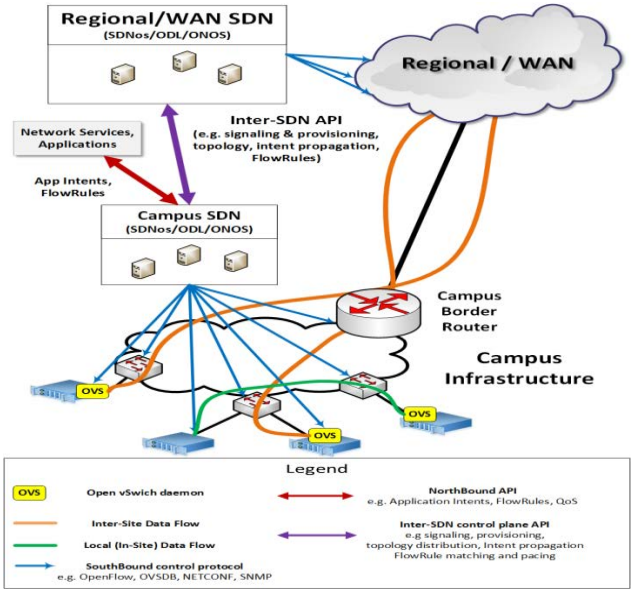
4. A Next Generation Integrated SDN Architecture (SDN-NGenIA) for HEP and Global Scale Science

Some of the largest data- and network-intensive programs in operation or planned today are the LHC and its follow-on High Luminosity LHC program, the LSST and SKA astrophysics surveys, the Joint Genome Institute and many other data-intensive emerging areas of growth[3]. These programs face unprecedented challenges in global exascale data distribution, processing, access and analysis, and in the coordinated use of massive but still limited

Figure 7: Schematic view of end-to-end SDN control, seen from one campus and its neighboring regional network.

CPU, storage and network resources. In response to these challenges and opportunities for science, Caltech, together with ESnet, Fermilab, Starlight/iCAIR, and other key laboratory, university and industry partners are designing and developing the first stages of the **SDN N**ext **Ge**neration **I**ntegrated **A**rchitecture (SDN-NGenIA) for HEP and global scale science.

The overarching goal is to maximize the discovery potential of scientific collaborations through the development of revolutionary open source products and methods in the SDN, virtualization and global system operation and optimization space. This will be accomplished by exploiting and contributing to the remarkable synergy emerging between: Deeply programmable software-defined agile and adaptive network infrastructures that are emerging as multi-service multi-domain network "operating systems" interconnecting next generation Science DMZs, and the systems developed by the data intensive science programs harnessing global workflow, scheduling and data management systems. While the initial focus will be on the challenging LHC use case, the products developed will be general, and apply to many fields of data intensive science. These will be informed by the LSST and bioinformatics/genomics use cases, which will be explored during the latter part of the planned program.



[3] See for example: http://www.es.net/assets/pubs_presos/BER-Net-Req-Review-2012-Final-Report.pdf

We plan to construct autonomous, intelligent site-resident services that dynamically interact with network-resident services, and with the science programs' principal data distribution and management tools, to request or command network resources in support of high throughput petascale to exascale workflows, using:

(1) smart middleware to interface to SDN-orchestrated data flows over network paths with guaranteed bandwidth all the way to a set of high performance end-host data transfer nodes (DTNs),
(2) protocol agnostic SDN-based QoS and traffic shaping services at the site egress that will provide stable, predictable data transfer rates, and auto-configuration of data transfer nodes, and
(3) host and site agent systems coupled to machine learning and global system modeling.

Specific work items include: (1) deep site orchestration among virtualized clusters, storage subsystems and subnets to successfully co-schedule CPU, storage and network resources; (2) science-program designed site architectures, operational modes, and policy and resource usage priorities, adjudicated across multiple network domains and among multiple virtual organizations; (3) seamlessly extending end-to-end operation across both extra-site and intra-site boundaries through the use of next generation Science DMZs; (4) novel methods of file system integration that enable granular control of extreme scale long distance transfers through flow matching of scattered source-destination address pairs to multi-domain dynamic circuits; (5) funneling massive sets of streams to DTNs at the site edge hosting petascale buffer pools configured for flows of 100 Gbps and up, exploiting state of the art data transfers where possible; (6) adaptive scheduling based on pervasive end-to-end monitoring, including DTN or compute-node resident agents providing comprehensive end-system profiling; and (7) unsupervised and supervised machine learning and modeling methods to drive the optimization of end-to-end workflow involving terabyte to multi-petabyte datasets.

The services to be developed are intended to interface seamlessly to the adaptive network operating system being developed in companion projects by ESnet, Fermilab, Caltech and others, to ensure that authorized applications achieve full throughput. The accumulated knowledge also is expected to inform the design of the following generations of distributed petabit/sec systems, including continental scale instruments such as SKA, and real-time leadership computing systems of the next decade harnessing zettabyte datasets.

5. Supercomputing 2015

The Caltech and Michigan HEP groups have been actively involved in the design and optimization of high speed Data Transfer Nodes (DTN). In the past the team has successfully transferred data across a pair of systems with peaks of above 90Gbps using multiple 40GE NICs. This year at SC15 given several hardware level advancements in both network and storage, we intend to demonstrate two variants of next generation DTN nodes capable of moving data at 100Gbps from storage to storage. These DTN nodes are based either on the general purpose SSD drives using off the shelve I/O controllers or specialized storage using NVMe PCIe storage cards. DTN nodes will be equipped with 100GE NICs.

Design of a 4 X 100G DTN well matched to the transfer of petabyte data sets, to and from the edge of the labs where next-generation HPC leadership systems will be installed by 2018-19, is now underway.

5.1 Innovation and Demonstration Plans

The group believes that optimization of large scientific data flows and providing dedicated WAN network paths over large research networks can be successfully achieved through an agent based intelligent software system having the visibility across the whole network. Such software when coupled with software defined networking (SDN) and the OpenFlow protocol to control the underlying switching fabric, can provide the best paths to data flows among any set of servers, while taking into account the state and policies in a multidomain network, as well as other large flows in progress.

One interesting problem in large data flow management systems (e.g. PhEDEx, Panda, etc.) is the possibility to dynamically adjust the bandwidth of different types or classes of data flows, based on scheduling parameters. This can be achieved to some degree using traffic shaping and QoS policies on the network devices. In one of the demos the group will demonstrate the possibility of dynamic bandwidth control of data flows right on the end-host using an SDN controller to configure the end system. The end-host systems will run Open vSwitch (OVs), which will be controlled via standard OpenFlow protocol. During the exercise, a single SDN instance will be used to dynamically configure an aggregated traffic shaping among multiple end systems.

The group will also demonstrate the use of SDN controllers to manage dataset transfers among end hosts. The controllers will write OpenFlow rules that direct specific transfers along specific paths through the network, according to a desired strategy. The paths are calculated (proactively or reactively) by software algorithms running in the controllers. The group will demonstrate SDN codes that implement several strategies, including those that select a random path, the shortest path, the least loaded path, the path with the fewest flows, as well as the maximum bandwidth path. The demonstration will show how each data set transfer can be associated with a specific path selection strategy. These demonstrated capabilities will help scientists to manage and prioritize timely access to large experimental data sets over the WAN.

In a separate demonstration, the group plans to present a high performance DTN system designed to achieve an aggregate network traffic flow of 400Gbps among a pair of systems connected through a multiport 100GE switch.

5.2 Broad HPC and Science Relevance

ESnet's bandwidth utilization projection for the next few years clearly puts a requirement on the end sites to actively plan and deploy high speed DTN server nodes together with SDN enabled paths. The proposed solutions directly address the HPC and HTC (High Performance Computing and High Throughput Computing) requirements and are applicable to various science projects in general.

As noted above there are many areas of science that are beginning to generate large amounts of data that need to be shared and accessed across multiple institutions. The infrastructure we will be demonstrating foreshadows what may soon be typical in many different domains of science and can serve as a working roadmap for others to explore. Certainly research universities will want to understand what is possible and what will be required to ensure that they can effectively support their researchers, as they collaborate in data-intensive programs that span many campuses and many countries.

5.3 ONOS

ONOS is an open source SDN networking operating system for Service Provider networks architected for high performance, scale and availability. ON.Lab announced the availability of ONOS in December 2014. The main partners include AT&T, China Unicom, Ciena, Cisco, Ericsson, Fujitsu, Huawei, Intel, the U.S. National

Science Foundation, NEC, NTT Communications and SK Telecom.

Some of the key characteristics of ONOS that are attractive for the LHC use case as well as the LSST, bioinformatics and weather now-casting applications that require deadline or near-real-time scheduling are: (1) Coherent messaging, flow tables, state management and leadership election among controllers; all of which are designed to survive an outage of the leader (master) instance and subsequent failover for high availability, (2) Support for some of the path selection algorithms implemented in our controller implemented in ODL: shortest path, least-loaded, etc. with hooks to add policy-based elements to the path construction procedure, (3) Designed to work with Open VSwitch across complex network topologies, for protocol independent control and impedance matching end-to-end of large flows, (4) Automatic load balancing through traffic engineering to allow higher than traditional levels of utilization and availability, (5) Graph based topology tracking and presentation, which makes traffic engineering and load balancing simpler, and (7) The ability to create mutually transparent network slices or islands where each "community" sees only the hosts relevant to it, and where each community has effectively its own network. Use of different protocols, parameter settings, policies, etc. is therefore possible.

The SDN-IP Peering application developed by the ONOS project team and leveraging the open source Quagga BGP suite (see Figure 8) enables peering among the SDN-based Internet2 network with traditional IP-based networks and other SDN-based networks.

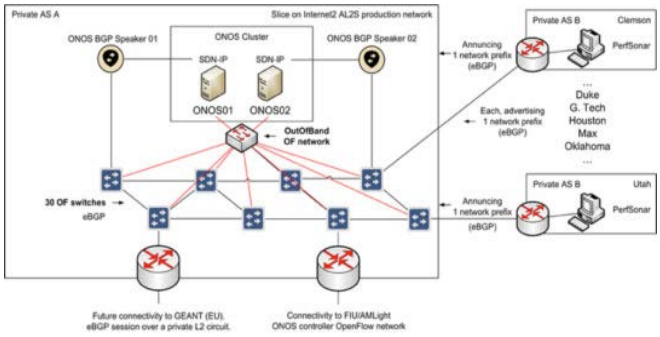


Figure 8: SDN Layer 2 topology map and IP peering using ONOS.

6. Conclusions

Fast and efficient data distribution and access, as required by distributed scientific instrumentation such as the LHC experiments' computing infrastructures, rely on the smooth interplay of many components. On top of the raw network capacity, the network architecture, switching equipment features and performance, end-system I/O architecture, the transfer applications and data management system software need to be tuned, and to some extent co-designed and co-developed, for frictionless operation. This has become increasingly challenging as the volumes and transfer rates required by the target science programs continue to grow exponentially.

The emergence of SDN, including Open Daylight, ONOS, and the controllers and end-to-end operating systems being developed by the network and science research communities provides a promising path towards an intelligent, adaptive ensemble of national and global networks that can meet these otherwise daunting needs.

The Supercomputing 2014 demonstration by the Caltech team and many partners, has brought together these major components, including state-of-the art commercially available hardware, and software and system components, SDN controllers and the first network operating system components, developed in several NSF and DOE funded projects.

We have shown what is achievable with state-of-the-art components and applications including full Terabit/s data movement between nodes both at the exhibition floor as well as at several LHC computing sites reachable over 100G WAN infrastructures, coupled to intelligent path construction and flow distributions over multiple paths. We have also advanced the state of the art in the management and optimization of flows with SDN methods at layers 1, 2 and 3 and advanced transfer protocols, integrated with some of the mainstream global data management applications of the LHC experiments.

The SC15 demonstrations will further advance the state of the art, moving these methods closer to production readiness for the LHC and other use cases, exploiting the latest SDN Open Daylight and ONOS releases and related methods (such as OVS), and spreading the knowledge to a growing circle of SDN developers in the HEP community. The scale will also advance as 100GE network interface cards for servers have appeared in recent months.

The team's SC15 demonstrations will also include first uses of Named Data Networking (NDN) for HEP data analysis including object collection extraction, building on experience in the NDN project and the climate science community. These aspects are covered in more detail in a companion paper focusing on NDN that has been accepted for presentation at the NDM 15 workshop.

7. SC14, ANSE and SC15 Team Members

An Academic Network at Sao Paulo (ANSP): Jorge Marcos, Luis Lopez
California Institute of Technology (Caltech): Artur Barczyk, Azher Mughal, Dorian Kcira, Harvey Newman, Iosif Legrand, Julian Bunn, Michael Bredel, Ramiro Voicu, Samir Cury, Vlad Lapadatescu, Maria Spiroulu, Jean-Roch Vlimant, Wayne Hendricks
CANARIE: Thomas Tam
CERN: Edoardo Martelli, David Foster
Colorado State University: Christos Papadopoulis, Susmit Shannigrahi
DE-KIT (Karlsruhe): Bruno Hoeft
Echostreams: Andy Lee, Gene Lee
ESnet: Brian Tierney, Eric Pouyoul, Inder Monga, Chin Guok, Greg Bell, Bill Johnston
Imperial College London: David Colling, Duncan Rand, Simon Fayer
Internet2: Eric Boyd
Fermilab: Phil DeMar, Wenji Wu, Panagiotis Spentzouris
Florida International University/AmLight: Julio Ibarra, Heidi Alvarez, Jeronimo Bezerra
Michigan State University: Andrew Keen
Northeastern University: Edmund Yeh, Ran Liu
Northwestern University: James Chen, Joe Mambretti
Padtec: Sergio Timoteo, Jorge Salamao, Chip Cox
Princeton University: Tony Wildish
Rede Nacional de Ensino e Pesquisa (RNP): Alex Moura, Gustavo Dias, Leandro Ciuffo, Michael Stanton
SurfNET: Gerben von Malenstein
Universidade Estadual de Campinas (UNICAMP): Christian Esteve Rothenberg, Dalton Soares Arantes, Darli Melo, Felipe Rudge Barbosa
Universidade Estadual Paulista (UNESP): Sergio Novaes, Rogerio Iope, Beraldo Leal, Eduardo Bach, Marcio Costa

Universidade Federal do ABC: Gustavo Pavani
University of Michigan (UMICH): Jorge Batista, Robert Ball, Roy Hockett, Shawn Mckee
University of Texas at Arlington (UTA): Kaushik De
University of Victoria (UVIC): Ian Gable, Randal Sobie
Vanderbilt University: Alan Tackett, Andrew Melo, Paul Sheldon

ACKNOWLEDGMENTS

We thank the agencies for the following grants, under which much of this work was carried out:

- ANSE – NSF award# 1246133
- CISCO – Award # 2014-128271
- Tier2 – NSF award # 1120138
- OLIMPS - DOE award # DE-SC0007346 (through 7/14/14)
- US LHCNet - DOE # DE-AC02-07CH11359 (through 5/31/15)

Our thanks also go to our commercial partners who made the Supercomputing demonstration possible by donating state-of-the art equipment: Brocade, Dell, Echostreams, Extreme Networks, Intel, Padtec and Spirent.

We also thank the SCinet network team for their strong support over the past and upcoming editions of Supercomputing exhibition, culminating in Terabit/sec infrastructure

REFERENCES

[1] NSF CC-NIE Integration ANSE (Advanced Network Services for Experiments) http://www.nsf.gov/awardsearch/showAward?AWD_ID=1246133
[2] AutoGOLE: http://www.glif.is/meetings/2015/spring/vanmalenstein-autogole.pdf
[3] LHCONE - LHC Open Network Environment: http://lhcone.net
[4] OSCARS - On-Demand Secure Circuits and Advance Reservation System: http://es.net/engineering-services/oscars/
[5] ESnet Science DMZ - https://fasterdata.es.net/science-dmz/
[6] OESS: Open Exchange Software Suite: http://globalnoc.iu.edu/sdn/oess.html
[7] FSFW: FlowSpace Firewall: http://globalnoc.iu.edu/sdn/fsfw.html
[8] Large Synoptic Survey Telescope: http://www.lsst.org/lsst/
[9] Openwave 100G Project. http://news.fiu.edu/?attachment_id=77505
[10] Internet2/OESS http://www.internet2.edu/products-services/advanced-networking/oess/
[11] Development of Dynamic Network System (DYNES): http://nsf.gov/awardsearch/showAward?AWD_ID=0958998
[12] J. Zurawski, R. Ball, A. Barczyk, M. Binkley, J. Boote, E. Boyd, A. Brown, R. Brown, T. Lehman, S. McKee, B. Meekhof, A. Mughal, H. Newman, S. Rozsa, P. Sheldon, A. Tackett, R. Voicu, S. Wolff and X. Yang, "The DYNES Instrument: A Description and Overview", 2012 J. Phys.: Conf. Ser. 396 042065 doi:10.1088/1742-6596/396/4/042065
[13] Fast Data Transfer: http://fdt.cern.ch
[14] Open vSwitch project: http://openvswitch.org/
[15] Accelerating Open vSwitch to "Ludicrous Speed": http://networkheresy.com/2014/11/13/accelerating-open-vswitch-to-ludicrous-speed/
[16] R. Voicu, "Traffic shaping using OVS": https://indico.cern.ch/event/376098/contribution/24/material/slides/1.pdf, June 2015, LHCOPN-LHCONE meeting, LBNL Berkeley
[17] Legrand, I., Voicu, R., Cirstoiu, C., Grigoras, C., Betev, L., and Costan, A., "Monitoring and control of large systems with MonALISA", ACM Queue 7, 6 (2009), 40–49, Online at: http://queue.acm.org/detail.cfm?id=1577839