Integrating Open Science Grid (OSG) with Software Defined Exchange (SDX) Projects Workshop Report

June 5-6, 2019

Dr. Julio Ibarra, Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU)
Dr. Heidi Morgan, Information Science Institute (ISI) Internet and Networked Systems group at the University of Southern California (USC).
Dr. Thomas DeFanti, California Institute for Telecommunications and Information Technology (Calit2) at University of California San Diego (UCSD)
Dr. Frank Wuerthwein, Open Science Grid (OSG) and San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD)

Table of Contents

1. Executive Summary

The report presents a summary of the activities at the workshop on the integration of the Open Science Grid with the Software Defined Exchange Points (SDX), which took place in June 2019 at the Calit2 at the University of California San Diego (UCSD). Representatives from StartLight SDX, PacificWave-SDX, and AtlanticWave-SDX presented current states of the SDXs along with the keynote presentation on OSG advancements in open science through distributed high-throughput computing. The main goals of the workshop were to understand what useful network services an SDX can provide domain scientists on OSG and what is required for the IRNC SDX projects to integrate with OSG.

2. Introduction

The Integrating Open Science Grid (OSG) with Software Defined Exchange (SDX) Projects Workshop took place on June 5-6, 2019 and was hosted by the California Institute for Telecommunications and Information Technology (Calit2) at the University of California San Diego (UCSD).

Organizers:
The California Institute for Telecommunications and Information Technology (Calit2, previously Cal(IT)2), also referred to as the Qualcomm Institute at its San Diego branch, is an academic research institution jointly run by the University of California San Diego (UCSD) and the University of California, Irvine (UCI).
CIARA is the Center for Internet Augmented Research and Assessment at Florida International University (FIU), created to serve as a catalyst for Internet innovation. CIARA enables faculty and student research by facilitating collaboration with the Cyberinfrastructure community.

3. Goals and Objectives of the OSG-SDX Workshop

The goal of this workshop was to explore the technical requirements for integrating the Open Science Grid (OSG) with the Software Defined Exchange (SDX) projects of the National Science Foundation (NSF) IRNC program.

OSG definition: The Open Science Grid (OSG) provides common service and support for resource providers and scientific institutions using a distributed fabric of high throughput computational services. The OSG does not own resources but provides software and services to users and resource providers alike to enable the opportunistic usage and sharing of resources. The OSG is funded through a diverse portfolio of awards from the National Science Foundation and the Department of Energy.

SDX definition: Software Defined Networking (SDN) offers direct control over packet-processing rules that match on multiple header fields and perform a variety of actions. A Software Defined IXP (SDX) enables application-specific peering (e.g., two networks peer only for streaming video traffic), has a programming abstraction for applications, applications isolation, scalability for rule-table size and computational overhead. SDX enables unlimited user's policies implementation for

participants who advertise full routing tables while achieving sub-second convergence in response to configuration changes and routing updates (Gupta et al., 2014).

4. Workshop Activities

The two-day OSG-SDX Workshop took place at Atkinson Hall at UCSD. More details about logistics of the meetings can be found at the Eventbrite registration page here: https://osg-sdx-workshop.eventbrite.com

Approximately 28 attendees participated (25 in person and three remotely). See Appendix B. The meeting gathered participants from 19 university organizations and research institutions from the USA and Latin America:

- California Institute of Technology (Caltech)
- Sao Paulo Research and Analysis Center (SPRACE) Brazil
- San Diego State University (SDSU)
- San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD)
- University of California San Diego
- Information Science Institute (ISI) USC
- University of California Merced (UC Merced)
- Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU)
- iCAIR/Northwestern University
- University of Utah
- International Center for Advanced Internet Research at Northwestern University
- University of California San Diego (UCSD)
- CENIC - Pacific Wave
- California Institute of Technology (Caltech)
- National Science Foundation (NSF)
- California Institute for Telecommunications and Information Technology (Calit2) at University of California San Diego (UCSD)
- Georgia Institute of Technology (Georgia Tech)
- University of North Carolina (UNC) Chapel Hill - RENCI
- Information Science Institute (ISI) University of Southern California (USC)

Calit2 offered a video Conference connection via ZOOM for the remote participants.
The workshop was comprised of two sessions:

- Day 1- OSG Challenges, landscape, and roadmap; and NSF IRNC Software Defined Exchange Panel
- Day 2 - Integration, Design, and Planning Jam Session

See Appendix A for agenda details.

4.1 OSG Challenges, landscape, and roadmap

The workshop began with welcoming remarks by Larry Smarr, Director of California Institute for Telecommunications and Information Technology (Calit2). Updates on the following NSF awarded project were presented:

- (2015-2020) Pacific Research Platform (PRP) connection on campus "big data freeways" to create a regional end-to-end science-driven "big data superhighway" system under the NSF CC*DNI Grant
- (2015-2020) NSF CI-New project Cognitive Hardware and Software ecosystem Community Infrastructure (CHASE-CI) has implemented 256 High Speed "Cloud" GPUs for 32 machine-learning faculty & their students at ten campuses to train AI algorithms on Big Data.
- (2018-2019) A National-scale pilot using CENIC & Internet2 to connect Quilt regional R&E Networks toward a National Research Platform (NRP)

Over 100 UCSD-designed FIONAs boxes have been deployed on the PRP to solve the disk-to-disk data transfer problem at nearly full speed on best-effort 10g, 40g, and 100g networks. PRP's Nautilus Hypercluster Kubernetes is used to orchestrate software containers on 15 campuses, containing 3300 CPUs Cores and 122 hosts (4PB). Currently, the OSG Data Federation (200K compute cores) built on 9 Data Caches to reduce network traffic and hide data access latencies. The IceCube Neutrino Observatory in Antarctica[1] research instrument is a use case with a significant increase in the usage of the OSG PRP network (190 GPUs and 1348 CPU-Cores) since the beginning of March 2019. Details about this presentation can be found here.

The workshop continued with a keynote presentation by Frank Wuerthwein, Executive Director of the Open Science Grid San Diego Supercomputer Center (SDSC) at the University of California San Diego (UCSD). OSG serves the needs of an undergrad to a big science lab. There are four distinct groups:

- Individual researchers and small groups on OSG-Connect
- The campus Research Support Organizations
- Multi-institutional Science Teams (e.g., XENON, GlueX, SPT, Simons)
- Big science projects (e.g., US-ATLAS, US-CMS, LIGO, IceCube)

OSG is the optimal partner for large distributed systems because of the distributed high throughput computing (dHTC) or latency tolerant computing that maximizes the effective capacity. The two challenges for successful dHTC are how to split big computing problems to a smaller one that can fit at an individual machine and how to minimize the user requirements for maximum effectiveness. In this way, researchers can curate, publish their software & data, and deliver them at runtime regardless of the location. Mindful parallelism is implied to efficiently use the dHTC resources, including CPU (200K 24hours x 30 days) and GPU (690K hours within 30 days). Over the last few years, the OSG data federation has exponentially grown. This requires another look at the researcher's connection points.

In the exaflop age, dHTC is unique because the scaling is only limited by the scaling behavior of the data distribution and the infrastructure software. Currently, Condor[2] is used to addressing the

[1] The IceCube Neutrino Observatory is the first detector of its kind, designed to observe the cosmos from deep within the South Pole ice: https://icecube.wisc.edu/

[2] HTCondor is a specialized workload management system for compute-intensive jobs: http://research.cs.wisc.edu/htcondor/description.html

science drivers that use OSG. OSG's goal is to grow effective capacity while growing the research community that can benefit from dHTC. For example, an exaflop hour will be scientifically useful and feasible.

OSG strategy for growing effective capacity is to minimize the effort it takes to integrate new resources with OSG. The new institutions that willing to join OSG should be able to integrate their resources simultaneously instead of requiring a complex software implementation (learn, operate, train) from their side. Currently, OSG is offering to operate services to the groups they serve in order to minimize the threshold to entry for everybody. A campus Research IT Organization should not have to learn anything "non-standard" in order to have their researchers benefit from OSG, or have their resources be available via OSG. A researcher should be able to have a single access point and decide how to proceed according to the importance of its project.

OSG federation applies distributed control by enabling the recourse owners to determine the policy if use (connect or disconnect anytime), the type of recourse to be used (RAM, core per node, etc.), and matchmaking (locally, queue centrally, etc.) the requirement for the submitted tasks. In this way, if one recourse gets disconnected, the submitted task is sent back and executed at another location. OSG operates overlay system(s) as services for all of science. This is why the nature of distributed Research and Educational Networks can provide support for OSG via some interface. A researcher can submit data to the OSG Data Federation, which will still reside on his local institutional storage and be used via the OSG caches. Currently OSG Data Federation consists of six data origins:

- Fermi National Accelerator Laboratory (FNAL): HEP experiments
- University of Chicago: OSG community
- California Institute of Technology (Caltech): Public LIGO Data
- University of Nebraska–Lincoln (UNL): Private LIGO Data
- San Diego Supercomputer Center (SDSC): Simons Foundation
- National Center for Supercomputing Applications (NCSA) at the University of Illinois: DES & NASA Earth Science

Nine Data Caches are connected, and three are planned for the near future (Amsterdam PoP, Houston, and Sunnyvale). Caches are essential when it comes to guaranteeing performance across the nation. Currently, multiple caches can function as one. It is important to hide access latencies, reduce unnecessary network traffic from data reuse, and protect the data origins from overloads. There are minimum OSG requirements in place addressing the data origins.

Deployment. The operation model consists of a stack of hardware, OS, network performance, Data Federation services, and science applications. Following the current Pacific Research Platform (PRP), which runs the Kubernetes K8S cluster for OSG, hardware owners operate hardware, OS install and join K8S for container orchestration. Therefore, Kubernetes is the most suitable solution for OSG scaling out simultaneously by using containers on orchestration platform. The idea of Quality of Service (QoS) is that the Data Federation service can flow between the caches and origins (finite number of connections), and for elastic scale-out into the cloud.

In summary, the OSG's objective is to advance Open Science through distributed High Throughput Computing. To achieve this goal, OSG needs to maximize the integrated capacity with minimal invasion of the campus infrastructure. OSG is a natural partner for networking projects because it is distributed as the Software Defined Networks, federated by necessity, and depends on well-functioning networks.

4.2 NSF IRNC Software Defined Exchange (SDX) Panel

The NSF IRNC SDX Panel introduced the scopes, goals, and accomplishments of three SDX projects: StarLight SDX, PacificWave SDX, and the AtlanticWave SDX.

StarLight SDX

StarLight SDX has over nine national and international key participants. Updates on several projects were introduced.

In response to the needs of advanced communication services and technologies, the International Center for Advanced Internet Research (iCAIR)[3] has been established to provide a focal point for leading-edge Internet research, innovation, and early deployment. The iCAIR project goal is to accelerate leading-edge innovation and enhance global communications through advanced internet technologies, in partnership with the global community. iCAIR undertakes basic research in the areas of networking technology, which inform IRNC SDX development, as transition from legacy networks, extremely large capacity, specialized network services, architecture and technologies for data-intensive science, high degrees of communication services customization, highly programmable networks and network programming languages, network virtualization, tenant networks, highly distributed signaling processes, network operations automation, etc.

The IRNC StarLight SDX initiative is designing, implementing, and operating new services for global data-intensive sciences, based on emerging next-generation architecture and technologies, including virtualization, orchestration, segmentation (slicing), software defined resources, programmability, and customization. This project is transitioning network exchanges to open innovation platforms with over 36 supported science applications. For example, the Large Hadron Collider Optical Private Network (LHCOPN)[4] Tier1 Private Network connects over 15 international research communities ranging from 10Gbps, 20Gbps, 40Gbps, and 100Gbps. Six new science communities using the LHCOPN network: Belle II Experiment in particle physics, Pierre Auger Observatory (PAO) studying Ultra-High Energy Cosmos Rays, the PAO & LIGO, & Virgo collaboration on the Gravitational Wave origin, the NOvA experiment in neutrino physics, the XENON Dark Matter Project, and the DUNE/protoDUNE neutrino experiment. An example of High Energy Physics SDX Data Transfer Node (DTN) prototype service is the implementation of DTNs for the research community in Taiwan at the LHC network.

Additional support is provided for projects in Magnetic Fusion Energy[5] and the Argonne National Laboratory Advanced Photon Source[6]. Geoscience research, creating daily National Oceanic and Atmospheric Administration (NOAA)[7] research weather predictions, is also supported by the StarLight SDX network along with Data Commons & Data Sharing initiatives from the University of Chicago[8]. Another ongoing collaborative initiative is with the Square Kilometer Array (SKA)[9] by providing infrastructure and data sharing services in the domain of astronomy. Currently, StarLight operates two SDX, GENI[10] SDX testbeds (~25 network research testbed), and IRNC

[3] International Center for Advanced Internet Research (iCAIR): https://www.icair.org/about/index.html
[4] Large Hadron Collider Optical Private Network (LHCOPN): https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome
[5] Fusion Energy Sciences (FES) program https://www.energy.gov/science/fes/fusion-energy-sciences
[6] Advanced Photon Source (APS) at Argonne National Laboratory: https://www.aps.anl.gov/
[7] -Wave is NOAA's Enterprise network: https://noc.nwave.noaa.gov/
[8] Data commons provide an alternative by co-locating data, storage and computing resources: https://ctds.uchicago.edu/datacommons
[9] The Square Kilometre Array (SKA) project is an international effort to build the world's largest radio telescope, with over a square kilometre (one million square metres) of collecting area: https://www.skatelescope.org/
[10] GENI is an open US infrastructure for at-scale networking and distributed systems research and education: https://www.geni.net/

SDX. An example of a supported testbed is the Global P4 Experiment Network (G-P4EN) project and the Cross-Pacific SDN Testbed in collaboration with Korea[11]. A large scale reconfigurable experimental instrument for computer science projects called Chameleon[12] is also supported. A new project for StarLight is the open storage network support for the Sloan Digital Sky Survey (SDSS)[13] project, which shares data releases with global astronomical communities.
StarLight Supports DTN-as-a-Service and SCinet DTN on multiple Supercomputing conferences demonstrations. Other projects supported by Starlight are the Advanced Measurement Instrument and Services (AMIS)[14] project, the OSiRIS[15] project (a pilot project funded by the NSF to evaluate a software-defined storage infrastructure for primary Michigan research universities), and the Pacific Research Platform (PRP). In summary, StarLight SDX success includes close partnerships with science communities, deep understanding of requirements, successful translations of requirements to advanced services, architecture, and technologies The accomplishments consist of novelty and innovation in transitioning network exchanges to open innovation platforms with new SDX and SD-WAN architecture, services and technologies for global data-intensive science collaborations.

PacificWave SDX
The Corporation for Education Network Initiatives in California (CENIC) operates the California Research and Education Network (CalREN)[16]. CENIC operates an optical backbone network that extends to regional and international partners and service 20M Californians along with the University of California Campuses, California State Universities & Community Colleges, Private Universities (e.g., Caltech, Stanford, USC), KA-12, libraries, scientific organizations, and labs. PacificWave[17] is a joint project of CENIC, the Corporation for Education Network Initiatives in California, and the Pacific Northwest Gigapop (PNWGP)It is operated in collaboration with the University of Southern California and the University of Washington. A distributed, fully open, peering and exchange fabric with access points on a 100g west-coast backbone that spans Seattle, San Francisco, Sunnyvale, and Los Angeles to nearly all Pacific Rim R&E networks and Internet2. PacificWave is a geographically distributed peering facility with Open Exchange Points supporting both commercial and R&E peers and is partially supported through NSF grants. Currently serves over 30 countries across the Pacific connecting to the Western USA and enables science-driven high-capacity data-centric projects, such as the Pacific Research Platform (PRP). PacificWave enables researchers to move data between collaborator sites, supercomputer centers, and campus Science DMZs without performance degradation.
The PacificWave SDX supports dynamic circuit and services provisioning (AutoGOLE/NSI[18] and MEICAN[19]), and further looking to implement SDN for End-to-End Networking Exascale

[11] Korea Institute of Science and Technology Information: http://www.kreonet2.net/
[12] Chameleon testbed: https://www.chameleoncloud.org/about/chameleon/
[13] Sloan Digital Sky Survey: https://www.sdss.org/
[14] AMIS: Software-Defined and Privacy-Preserving Network Measurement Instrument: https://meetings.internet2.edu/2016-technology-exchange/detail/10004470/
[15] OSiRIS is a pilot project funded by the NSF to evaluate a software-defined storage infrastructure: http://www.osris.org/
[16] CENIC's California Research and Education Network (CalREN) is a multi-tiered, advanced network-services fabric serving the majority of research and education institutions: https://cenic.org/network/network-overview
[17] Pacific Wave is a wide-area advanced networking facility: http://pacificwave.net/
[18] CENIC and Pacific Northwest Gigapop (PNWGP) deployed Pacific Wave, a geographically distributed peering facility in 2004: https://cenic.org/blog/item/pacific-wave-and-cenic
[19] Management Environment of Inter-domain Circuits for Advanced Networks is a web application for the management of Dynamic Circuit Networks (DCNs): https://github.com/ufrgs-hyman/meican

(SENSE)[20], Big Data Express[21], Network Function Virtualizations (NFV), Virtual Customer Premises Equipment (vCPE)[22], Infrastructure-as-a-service (IaaS), and containers from cloud providers, infrastructure and DTNs. Future collaborations also include federated access to SDX attached recourses as Inter-domain and inter cluster (CPU/GPU/TPU & Field Programmable Gate Arrays (FPGAs)[23] resources, storage, caching of data for high-energy physics and genomics, etc.). The PacificWave expansion, supported by NSF, enabled new instantiations and upgrades for existing connections to 100Gbps, SDN/SDX testbed deployment instrumentation measurement, monitoring, and analysis (perfSONAR and MaDDash), and collaboration with other IRNC awardees on SDX development. Current PacificWave SDX infrastructure implementation includes SDX middleware OpenFlow controllers (ONOS, Ryu), network testbed (NSI/OpenNSA), SDX OpenFlow switches at L2 (Seattle, Sunnyvale, Los Angeles), and other recourses as DTNs and PerfSONAR nodes.
Other initiatives between PRP and Network Startup Resource Center (NSRC) consists of organizing four workshops (90 participants from 60 organizations) to help network engineers to better understand and help researchers in their needs.
The goal of PRP is to support multi-institutional research groups working on large projects. Some of the science domains include particle physics (Compact Muon Solenoid -CMS experiment), biomedical genomics (San Diego Supercomputer Center -SDSC), earthquake engineering, telescope survey (Intermediate Palomar Transient Factory), visualization, virtual reality, and collaboration.

AtlanticWave SDX

The motivation behind the AtlanticWave SDX is to support the infrastructure for the science instruments.
Science instruments in South America are increasing in number (e.g., Atacama Large Millimeter Array (ALMA)[24], Dark Energy Camera (DECam)[25], Large Synoptic Survey Telescope (LSST)[26] with first light in 2022, Giant Magellan Telescope (GMT)[27] with first light 2024). For example, the Center for Scientific Computing in Sao Paulo/Brazil[28] process and analyze LHC Data, Open Science Grid (OSG) data, and data from other international instruments.
Currently, there is an increased data flows between the USA and Brazil research teams (Fermilab and Europe). For example, the received ~2,614 TB data flows from U.S. sources in 2017 and transmitted ~792 TB flows to U.S. destinations. There is a paradigm shift in south-north network traffic. The AmLight ExP project, sponsored by NSF, is addressing this demand by increasing AmLight network resilience and capacity. Additional links will be activated via the Pacific and the Atlantic Ocean for resiliency in case of natural disasters. Currently, there six links between the Boca Raton/USA and Fortaleza/Sao Paulo will be activated (four managed by RNP and two

[20] SDN for End-to-End Networking at Exascale (SENSE): https://www.es.net/assets/pubs_presos/SENSE-Thomas-20160217-on-Web.pdf
[21] BigData Express aims to provide schedulable, predictable, and high-performance data transfer service for DOE large-scale science computing facilities (LCF, NERSC, US-LHC computing facilities, etc.) and collaborators: https://bigdataexpress.fnal.gov/
[22] https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/solutions/vmware-vcpe-on-vmware-vcloud-nfv-brochure.pdf
[23] Field Programmable Gate Arrays (FPGAs) are semiconductor devices that are based around a matrix of configurable logic blocks (CLBs) connected via programmable interconnects: https://www.xilinx.com/products/silicon-devices/fpga/what-is-an-fpga.html
[24] Atacama Large Millimeter/submillimeter Array (ALMA) telescope: https://www.almaobservatory.org/en/home/
[25] Dark Energy Camera (DECam): http://www.ctio.noao.edu/noao/node/1033
[26] Large Synoptic Survey Telescope (LSST): https://www.lsst.org/
[27] Giant Magellan Telescope (GMTO): https://www.gmto.org/
[28] São Paulo Research and Analysis Center (SPRACE): https://sprace.org.br/

managed by FIU/ANSP/LSST). It will have a bandwidth capacity of 630Gbps. AMPATH[29] IXP is becoming a distributed exchange point by adding a PoP in Boca Raton, FL.
The goals of the AtlanticWave-SDX NSF project are to enable domain scientists to reserve network resources through a multi-domain SDX by simplifying the interface for domain scientists to request network resources and enabling science applications to react to network conditions in a more efficiently. Building a distributed SDX between the U.S. and S. America will support the dramatic increase in south-north science flows and will integrate the SDN infrastructures at AMPATH, SoX, SouthernLight, and AndesLight open exchange points. Partners and collaborators for the AtlanticWave SDX projects are Florida International University (FIU), University of Southern California Information Sciences Institute (USC-ISI), Georgia Institute of Technology (GT), Renaissance Computing Institute at UNC (RENCI), Academic Network of Sao Paulo (ANSP) [30], Association of Universities for Research in Astronomy (AURA), Rede Nacional de Ensino e Pesquisa (RNP, Brazil) [31], Red Universitaria Nacional (REUNA, Chile), Florida LambdaRail, and Internet2.
AtlanticWave SDX architecture consists of multiple layers: local controller (orchestrating the local network), AW-SDX Controller (Controls SoX, AMPATH, SouthernLight, and AndesLight), and Users (network operators and domain scientists) & science applications. AtlanticWave SDX strategy is to make the network as usable to the research community by enabling the users to request end-to-end science network services from the AW-SDX controller and consume end-to-end services composed by the AW-SDX controller.

AtlanticWave SDX characteristics are:

- AW-SDX provides a meet-me point where independent network administrative domains can interconnect to exchange compute, storage and network resources
- AW-SDX is designed as a multi-domain SDN framework, designed to provide end-to-end services
- AW-SDX provides a REST API to upper layers of the protocol stack
- AW-SDX automates the provisioning of dedicated end-to-end circuits by enabling the deployment of network policies between two or more SDN islands
- AW-SDX is designed to support intent-based networking; i.e., an application expresses a goal, then the SDN controller determines how to implement it
- AW-SDX is designed to support multiple southbound interfaces by which to communicate with local SDN controllers. For example, OpenFlow, P4, and NetConf
- AW-SDX exposes network services that domain scientists and science applications can consume through the application interface

4.3 OSG-SDX Integration architecture panel

The OSG-SDX Integration panel included presentations on architecture, components, interfaces, and other relevant information to reveal what resources are available (or not) for potential integration.

OSG Layers and Components

[29] International Academic Exchange Point at CIARA/FIU Miami: https://ampath.net/
[30] Academic Network of Sao Paulo Brazil: http://www.ansp.br/index.php/us/
[31] Research and Educational Network (RNP Brazil): https://www.rnp.br/en

OSG provides cyberinfrastructure (sites with compute, storage, and network recourses) and tools to Virtual Organization or VO (scientist) to collaborate by using shared cyberinfrastructure efficiently and easily.

Share resources, data, and software distribution using OSG

To convert users' scientific workflow request to OSG small executable tasks on shared resources, OSG is using High Throughput Computing (HTCondor), which understand Grid Security Infrastructure (GSI), sits at a site border and can interact with the local batch system(HTCondor, PBS, Slurm). A containerized version also exists for which a public IP and one open port are required. The access software is comprised of CernVM File System (CVMFL)[32] and Squid caching proxy (HTTP, HTTPS, FTP, and more)[33]. The shared access data OSG Solution is XCache/Stashcache[34], previously known as the XRootD proxy cache. XCache provides a caching service for data federations that serve one or more Vos. XrootD Based caching technology that also works together with CVMFS to provide data in Portable Operating System Interface (POSIX) using data locality. To accommodate the reusability of data by the scientific application, caching technology is designed on top of XrootD protocol and can use Squid proxies (many small files) or XCache (files >2GB and high throughput). The data still needs its source (origins). OSG uses XCache to distribute the data on demand based on locality after it is pulled by the cache from the source. The Xcache origin needs a public IP, 10Gbps NIC, and access to the storage.

Run science workflows on the OSG

HTCondor submit host is necessary for big collaborations as CMS, LIGO, IceCube, etc. For a single scientist, OSG Connect service is providing resources (CPU, memory, disk), immediate use, and a resource management system GlideinVMS which ties all sites together and creates an on-demand virtual heterogenous HTCondor pool.

Kubernetes Layers and Components

PRP CHASE-CI Nautilus Cluster[35] is a HyperCluster which is running Kubernetes container orchestration for Big Data applications. It consists of CILogon Federated Authentication[36] hook to provide access to resource scheduling (GPU, FPGA, TPU) and Cloud-Native storage (Rook, Ceph, EdgeFS). PerfSONAR, MaDDash, TCP_Info[37], ClusterWide, and Nagio[38] are used for network monitoring along with Prometheus and Grafana dashboards for visualization. Recently, ElastiFlow[39] EKS flow analysis dashboards are implemented. sFlow is considered for the near future.

The core to properly deploy Kubernetes is via multiple High-Availability (HA) Kubernetes masters, which gives a sense of control and ownership. The authentication via CILogon enable Federated Identity, allows Kubernetes to create a refresh token, and apply dynamic namespace which allows OSG to decide what gets to rune where. Geo-distributed storage pools are achieved by using Rook operator architecture for EdgeFS. Rook enables EdgeFS storage systems, which resolves multiple commits over multiple nodes simultaneously and run on Kubernetes using

[32] The CernVM File System provides a scalable, reliable and low-maintenance software distribution service: https://cernvm.cern.ch/portal/filesystem

[33] Squid is a caching proxy for the Web supporting HTTP, HTTPS, FTP, and more: http://www.squid-cache.org/

[34] XRootD is a highly-configurable data server used by sites in the OSG to support VO-specific storage needs: https://opensciencegrid.org/docs/data/xrootd/overview/

[35] Nautilus is a HyperCluster for running containerized Big Data Applications: http://pacificresearchplatform.org/nautilus/

[36] CILogon is an Integrated Identity and Access Management Platform for Science: https://www.cilogon.org/

[37] A mechanism in the Linux kernel for accessing information about a TCP socket: https://wiki.geant.org/display/public/EK/LinuxTcpInfo

[38] Nagio is IT Infrastructure Monitoring tool: https://www.nagios.org/

[39] Network flow Monitoring (Netflow, sFlow and IPFIX) with the Elastic Stack: https://github.com/robcowart/elastiflow

Kubernetes primitives. This allows low latency nodes to be tied together with high-performance local pools, remote cache, and meta-data pools. The EdgeFS ISGW (Inter-Segment GateWay) is allowing the generation of storage pools across clouds (e.g., metadata caches can be on Amazon, google cloud service on the bulk of the data, on-premise or off-premise pools in data centers, or distribute on PRP network). A sample hardware box configuration used for scheduling GPUs using Kubernetes consists of 24 CPU Cores, 32,000 GPU cores, 96 GB RAM, 2TB SSD, Dual 10Gbps ports (AMD and NVIDIA). Additional FIONA[40] boxes and Xilinx FPGA[41] are connected too. In order to run different requests on different cloud providers (Amazon, Google, CENIC, I2, GPN, Nysernet, LEARN, and Nautilus), the SDN concept can be applied. Stitching all networks (PRP, I2/The Quilt) together with Kubernetes can be done in multiple ways. Kubernetes v2 includes the Federation, which is consisting of two types of information, type and cluster configuration. Propagation, which is referring to mechanisms that distribute resources (template, placement, and override types) depending on the configuration, provides a concise representation of a resource, intended to appear in multiple clusters. They encode the minimum information required for propagation and are well-suited to serve as the glue between any given propagation mechanism and higher-order behaviors like policy-based placement and dynamic scheduling. These fundamental concepts provide building blocks that can be used by higher-level APIs (status, policy, and scheduling).

Project Calico[42] is an SDN overlay project currently used in the Nautilus cluster and recently added new hooks for external IP addresses and BGP wide-area servers. Cilium Multi-cluster service routing[43] is ope-source software for transparently providing and securing the network and API connectivity between application services deployed using Linux container management platforms like Kubernetes, Docker, and Mesos. Cilium is working with GridFTP, and MaDDash is hardwiring clusters and allows failover clusters, authentication, encryption. Graphana[44] dashboards are populated dynamically from Prometheus[45] service directly form Kubernetes. Next, perfSONAR deploys automatically as new nodes are joined to the cluster, MaDDash is configured automatically, and mesh config is handled by a golang[46] web-service. Nautilus hosts GitLab CI inside the cluster for automated deploying services easier, continuously update code, roll back/forward, and versioning. Future plans include a new Traceroute node-graph visualization tool to pull data from the last esmond result[47]. Currently, Nagios is used along with ElastiFlow sFlow for visualization. The latest Nautilus development includes Admiralty Multicluster scheduling[48] for Kubernetes, federation cluster to cluster using hooks to flow jobs in-between them. Multicluster-controller is a Go library for building Kubernetes controllers that need to watch resources in multiple clusters.

Kubernetes is the easy way for OSG to onboard SDX as this approach was already used over to deploy OSG over the Nautilus cluster.

[40] A low-cost, flash memory-based data server appliance to act as a Big Data hub: http://www.calit2.net/newsroom/release.php?id=2342
[41] FPGA-as-a-Service related project repositories: https://github.com/Xilinx/FPGA_as_a_Service
[42] Free and open source, Project Calico is designed to simplify, scale, and secure cloud networks: https://www.projectcalico.org/
[43] API-aware Networking and Security: https://cilium.io/blog/2019/02/12/cilium-14/
[44] An open platform for analytics and monitoring: https://grafana.com/
[45] Prometheus is open-source monitoring solution: https://prometheus.io/docs/visualization/grafana/
[46] Go is an open source programming language: https://golang.org/
[47] perfSONAR is a collection of software for performing and sharing end-to-end network measurements: https://docs.perfsonar.net/esmond_api_rest.html
[48] Admirality is a Multicluster-Controller, a library for building hybrid and multi-cloud Kubernetes operators: https://admiralty.io/blog/introducing-multicluster-controller/

AtlanticWave-SDX Layers and Components

AtlanticWave-SDX has three pillars: Software (architecture, implementation, and API), CI/CD, Testbed & experiments. Underneath the software needs to control the infrastructure via a networking mechanism to connect all pieces and provide an end-to-end solution. Five options addressing a multi-domain network environment.

Option 1: Control plane control using protocols like BGP, GMPLS.

Option 2: Using the NSI protocol.

Option 3: Using a hierarchical SDX controller. AtlanticWave-SDX infrastructure includes a multi-domain (SDN) network environment and DTN nodes at the edges. A distributed SDX consists of SDX controller and local controllers, and advance reservation. APIs can abstract operators and domain scientists. To apply such a network mechanism, AtlanticWave-SDX software can create a network slice using VLAN, P2P & P2MP circuits for provisioning and policies (emulated L1, L2, L3) for ingress and egress. The implementation could be done using Corsa switches, topology configuration files, SDX controller, and local controllers with open architecture. Currently, the AtlanticWave-SDX workflow includes BootStrapping (SDX operator, configuration manifest), participation policy APIs (participant1, participant 2, …participant n), policy breakdown at the central controller, and OpenFlow rules. AtlanticWave-SDX can extend the infrastructure to accommodate a transit service provider as OSG by implementing a network plumbing mechanism (L2, L3, reservation base, or best-effort) for networking and cluster domains. A possible option to extend the architecture is to customize the functions of the SDX central and local controller. To have an end-to-end system, coordination between an Access control policy implementation and an IP address is needed.

Option 4: When multiple clusters are connected to form a single Kubernetes with a WAN in-between, AtlanticWave SDX Core Network Insert (CNI) Plugin in-between the Kubernetes clusters can be used in the form of inter-cluster services.

Option 5: Controller of controllers or just an ansible-playbook. The APIs from SDX and k8s/Calico[49] are comprehensive. The orchestrator will need a complex resource reservation model, and the configuration automation playbook will have a query-based reservation workflow. The goal is to automate the end-to-end system as much as possible.

Continuous Integration & Continuous Delivery (CI/CD) can address a constant development of the software by using an open-source automation server as Jenkings[50] and hosting for software development as GitHub[51].

AtlanticWave-SDX testbed environment is deployed on the Breakable Experimental Network (BEN)[52] infrastructure at RENCI[53]. The testbed consist of experimental dark fiber facility, supports experimentation in metro-scale, four PoPs across Research Triangle Park in North Carolina, and several segments of dark fiber. The system requires an initial switch state to have created a Create Virtual Forwarding Contexts (VFC) (Virtual OpenFlow Switches), tunnels from virtual (logical) ports to the physical ports, and a set OpenFlow Controller (Local Controller). The AtlanticWave-SDX software is running in a docker container, and initiate start components, listen for Local Controller (LC) and User Interface (UI) connections.

[49] Calico enables networking and network policy in Kubernetes clusters across the cloud: https://docs.projectcalico.org/v2.0/getting-started/kubernetes/

[50] Jenkins is a cross-platform for continuous integration and continuous delivery application: https://wiki.jenkins.io/display/JENKINS/Home

[51] GitHub provides hosting for software development version control using Git: https://github.com/

[52] BEN: Breakable Experimental Network: https://ben.renci.org/index.php?option=com_content&view=frontpage

[53] Renaissance Computing Institute (RENCI): https://renci.org/

The bootstrapping process set up the management process on the data plane, which includes clearing out existing rules, install management VLAN forwarding rules and set up an in-band management path for controllers. Ones the bootstrap process starts, LC clear the flows and pushes the initial learning flows for the control plane to the VFC, then negotiation between the LC and AtlanticWave-SDX controller starts. Scientists and network operators can reserve resources between two endpoints, servers, or DTNs via UI. AtlanticWave-SDX and LC controllers are aware of the topology by using a manifest file.
Another way of using the testbed is to create L2Tunnel[54] (P2P) connection with REST API and perfSONAR iperf3 tests for measurements.

StarLight-SDX Layers and Components
The selected SDX architectural attributes include control & network recourse APIs, Multi-domain integrated path controllers (with federation), controller signaling (including edge signaling), SDN/OF multi-layer traffic exchange services, multi-domain resource advertisement/discovery, topology exchange services, multiple highly customized services at all layers, granulated resource access (policy-based), foundation resource programmability, various types of gateways to open network environment, integration of OF & non-OF paths, and programmability for large scale large capacity streams.
The components available at StarLight are the network infrastructure of Internet2, Management Environment for Inter domain Circuits in Advanced Networks (MEICAN)[55] as an Orchestrator interdomain controller using NSI at L2 and L3, and an international SDN testbed. The starlight inter-domain SDN topology includes connections to PacificWave and Australia REN- AARNet[56]. The SDN-IP testbed with L3 SDX primarily operates from Asia by Taiwan Academic Network (TANet)[57] and National Applied Research Laboratories (NARLabs)[58] using ONOS[59] and OpenFlow. ExoGENI[60] testbed is another recourse available at StarLight with Open Resource Control Architecture (ORCA)[61] control framework as an orchestrator. StarLight also supports a distributed cloud InstaGENI for large-scale distributed research projects (34 sites). InstaGENI is a lightweight cluster with software-defined networking, Hardware-as-a-Service and Containers-as-a-Service, remote monitoring and management, and high-performance inter-site networking [62] with initial OpenFlow implementation and later transitioned to a virtual topology service. StarLight is accommodating varieties of frameworks and protocols corresponding to the project's requirements as GENI AM, ExoGENI, OpenFlow, NSI, ofNSI, and Open Science Data Cloud (OSDC)[63]. Starlight also participates in the ICAIR Global P4 experiment network (G-P4EN) project and Chameleon Cloud core network[64].

[54] Layer Two Tunneling Protocol (L2TP): https://tools.ietf.org/html/rfc2661
[55] Management Environment of Inter-Domain Circuits for Advanced Networks (MEICAN) developed by RNP Brazil: http://www.inf.ufrgs.br/~jwickboldt/?p=452
[56] Australia Research and Educational Network: https://www.aarnet.edu.au/
[57] TaiWan Advanced Research and Education Network (TWAREN): http://www.twaren.net/english/
[58] National Applied Research Laboratory: https://www.narlabs.org.tw/en
[59] ONOS is SDN controller platform: https://onosproject.org/
[60] ExoGENI is a GENI testbed: http://www.exogeni.net/
[61] ORCA is an IaaS software for managing meta-clouds. It is deployed in production on ExoGENI world-wide testbed https://github.com/RENCI-NRIG/orca5
[62] Nicholas Bastin, Andy Bavier, Jessica Blaine, Jim Chen, Narayan Krishnan, Joe Mambretti, Rick McGeer, Rob Ricci, and Nicki Watts. 2014. The InstaGENI initiative. Comput. Netw. 61, C (March 2014), 24-38. DOI=http://dx.doi.org/10.1016/j.bjp.2013.12.034
[63] Open Science Data Cloud provides the scientific community with resources for storing, sharing, and analyzing terabyte and petabyte-scale scientific datasets: https://www.opensciencedatacloud.org/
[64] A configurable experimental environment for large-scale cloud research: https://www.chameleoncloud.org/

The DTNs are a core resource to be used for the integration of the OSG in the SDX environment. As part of SCinet[65], StarLight participated in the development of a DTN ecosystem. Plans for demonstration at the Supercomputing conference 2019 includes six sides (USA, Australia, South Korea, Japan, and Singapore) and multiple DTNs deployed around the world. Starlight also has implemented the MREN research platform and Kubernetes cluster with PRP/TNRP.
Future initiatives include a new NSF Open Storage Network (ONS)[66] project, which will provide 1-2 Petabyte storage in the network.

PacificWave-SDX Layers and Component
Pacific Wave SDX stack includes SDX switches (L2), SDX middleware (OpenFlow, ONOS and Ryu Controllers), network testbeds (NSA/OpenNSA), testbed resources, science drivers' applications, and other resources as DTNs, perSONAR nodes, etc. There are three Pacific Wave exchange Switches collocated with three enhanced SDXchange switches at Seattle, Sunnyvale, and Los Angeles connected with 100G links. From those points, Pacific Wave SDX connects to GENI Mesoscale, ESnet, Internet2, and other participants. Traditional legacy peering between all Pacific Wave participants can remain on the Pacific Wave exchange switches. SDX services can be accessed via direct connection to the enhanced SDXchange switches or via the Pacific Wave exchange switches. A mixed connection can also take place between the participants.
Pacific Wave SDX/SDN testbed is migrating the bare-metal DTNs to containers with Kubernetes cluster and HA master nodes, federated access for the Nautilus cluster to the PRP recourses, additional GPU/TPU/FPGA resources at the exchange points, SENSE DTN-RM implementation, Big Data Express with mtdmFTP[67], etc.

Pacific Wave infrastructure-attached resources consist of 100G-connected DTNs / Kubernetes K8s Storage Nodes, 100G-connected perSONAR nodes, 10G-connected perSONAR nodes, 10G-connected dynamic perSONAR nodes / Kubernetes K8s master nodes, and 10G-connected x86 Hypervisor VM servers at each of the three locations.

Automated GLIF Open Lightpath Exchange (AutoGOLE)[68] fabric delivers dynamic network services among GOLEs and participating networks, and it is based on NSI connection service v2.0. The architectural standard is developed by the Open Grid Forum (OGF)[69]. It consists of a redundant aggregator backbone with a leaf ultimate provider agent (uPA) per network and 29 Network Service Agents (6 Aggregators, 23 uPA) advertising 30 networks. The autoGOLE uses Document Distribution Service (DDS) for NSA discovery[70] & document propagation between aggregators, monitoring/ troubleshooting & provisioning tools, Dashboard, MEICAN as an orchestrator, DDS Portal, etc. AutoGOLE has an advanced capability to enable experiments with new pathfinding and signaling algorithms and additional network modeling for optimizations. Pacific Wave has instantiated an autoGOLE (production L2 switch) with NSI uPAs (OpenNSA) to include SENSE network resource manager (RM) for interoperability. SENSE has already been instantiated at the Sunnyvale site and is planned for Los Angeles and Seattle sites.

[65] SCinet: https://scinet.supercomputing.org/workshop/sites/default/files/SC18-ARCH-XNET-NRE-Workshop-Presentation_0.pdf
[66] Open Storage Network (OSN): https://www.openstoragenetwork.org/
[67] The Multicore-Aware Data Transfer Middleware (MDTM) Project: https://mdtm.fnal.gov/
[68] AutoGOLE fabric delivers dynamic layer 2 network services between Open Exchanges and networks, designed as a multi-domain system: https://github.com/jeroenh/AutoGOLE-Topologies
[69] Open Grid Forum (OGF) https://www.ogf.org/documents/GFD.212.pdf
[70] NSI Document Distribution Service: https://github.com/BandwidthOnDemand/nsi-dds

Pacific Wave dynamic circuit services (AutoGOLE / NSI) are currently available to participants traversing each of the Seattle, Sunnyvale, Los Angeles, and Tokyo GOLEs. Current OpenNSA capability can only manage a single backend device, and each of the locations has its own OpenNSA domain. Future Pacific Wave plans include consolidation under one domain and a single interface for the research community. The control plane includes peering with (NSI Aggregators) ESnet, NetherLight, StarLight, and SINET. The data plane includes peering with ESnet, StarLight, SINET, JGN-X, Caltech, and Calit2 UCSD (pending). Reservations and bandwidth requests can be made using the MAICAN interface. A 2015 demonstration for a circuit reservation using MEICAN was done in minutes instead of 100 emails iterations between the involved parties.
Another Pacific Wave ongoing initiative is "PRPv2 BGP Pilot: Route server for control plane," which aims to create a super-channel between the facilities that will enable selective route announcement of Science DMZ recourses (directly between SDX and NSA).
Future plans include placing a DTN at the Guam Research and Educational eXchange point. Such an expansion will connect to the University of Guam and other Pacific universities.

4.4 Jam Session on Integration, Design, and Planning

The jam session began with a presentation on the data use cases from OSG prepared by Frank Wuerthwein. This list below attempts to make a reasonably comprehensive list of data use cases in distributed environments like OSG, WLCG, etc.

1. Bulk transfers
    a) Rucio - FTS - third-party transfer between servers supporting GridFTP/HTTP/XRootD protocols.
    b) Globus online with its proprietary protocol
2. Not bulk transfer (e.g., transfer connected with processing)
    a) Input to a job:
        i) Pulling a full file into the worker node environment at runtime, prior to processing (i.e., HTCondor file transfer, or some other workflow governed mechanism that pulls a file from a server that is the origin for the file; or fully self-contained singularity containers with the applications in them).
        ii) Same as i) but pulling the full file from the closest cache. This may or may not trigger a cache miss, and thus the additional transfer of a file between cache and origin.
        iii) File open and random read access to file in cache.
    b) The output from a job:
        i) Pushing output files back to origin directly from the worker node. E.g., HTCondor file transfer or some other workflow engine. In that case, the workflow engines take care of the scheduling of all files it is responsible for to avoid overloads of receiving storage system. Job occupies a batch slot until this scheduled transfer is completed (i.e., leads to CPU inefficiency at the end of a job while the output is transferred).
        ii) Asynchronous Stage Out. The end of the job pushes output into a local storage system. File transfer is scheduled for later, possibly as part of a bulk transfer activity using FTS, or alike.
    c) Transfers initiated in response to cache misses (example XRootD):

i) Cache servers are optimized to serve data out. If disks are busy being read, then cache misses will get fetched but not written to disk. Data instead is buffered through RAM as it is served straight to the application that requested the data.

ii) XRootD protocol is a remote file access protocol that supports vectors of byte ranges (i.e., efficient random access to data inside files). The caching is at the partial file level, i.e. the byte ranges are stored on the cache, and future requests for overlapping byte ranges are dealt with correctly (i.e., misses are filled in, and hits are served from disk). In addition, the caches are configured to fill in holes as lazily over time. If the cache server is not heavily used, it will go through the holes in files and fetch them. This config is based on the premise that any file that was partially read is worth caching fully to prevent future cache misses.

The session continued with a discussion about the SDX & OSG objectives and further planning. Some of the main questions are identifying the science application drivers, the geographically involved parties, what would be easy to accomplish, and how it will map on the technology.

Application drivers:
The identified experiment drivers include LIGO[71], GlueX[72], CMS[73], OSG Data Federation in general, genomics (bioinformatics, food & health view), and FNAL TITE Subset Origins & Caches (e.g., Nova, Dune). Other science drives concerning the astronomy community include science observatory, the Dark Energy Survey (DES) (not a huge amount of data use yet on OSG), FNAL DES (4th largest OSG user 400 TB sitting in Chicago), and projects in South America (e.g., LSST, POLARBEAR[74]). The Very Energetic Radiation Imaging Telescope Array System (VERITAS)[75] science workflow could be used as a use case to investigate workflows specific to the astronomy and how data on disk falls into Origins if the researchers use it.

Data movement between Origins & Caches (e.g., output/input buffering to Brazil) could be possible via AtlanticWave SDX scheduling GUI. For example, CMS wants to bring data to Brazil, and GlueX would want to ship data from Sao Paulo to the north, to the US.

Origin definition: Origin is a local file system (or a server that instantiates a service that mounts the file system) and speaks HTTP or XRooD (or other protocol), data archive permanently, and serving data out to the entire OSG federation. It could range from TB to PB in size in the file system, which is mounted locally to a host that has bare metal OS Kubernetes installed. Origin Kubernetes joins the PRP Kubernetes Cluster, and then OSG spool the origin service as a pod into the Kubernetes.
From a software architecture perspective, the SDX can provide the connection between the pod and the file system at the network stack to use the SDX network capability without interfering with the OSG. At some location (e.g., Chicago origin) can be switched to Kubernetes to accommodate such an implementation. The SDX network capability would sit underneath the OSG pod of the

[71] The Laser Interferometer Gravitational-Wave Observatory (LIGO): https://www.ligo.caltech.edu/
[72] GlueX is a particle physics experiment: http://gluex.org/GlueX/Home.html
[73] The Compact Muon Solenoid (CMS) is a general-purpose detector at the Large Hadron Collider (LHC): https://home.cern/science/experiments/cms
[74] POLARBEAR is a Cosmic Microwave Background polarization experiment: http://bolo.berkeley.edu/polarbear/
[75] Very Energetic Radiation Imaging Telescope Array System (VERITAS): https://veritas.sao.arizona.edu/

origin service and connects to other pods used to deploy the caches and create a network overlay without interfering with the OSG.

Geographically involved parties:
In the case of South America, many instruments are located in the Atacama Desert, Chile, where the science array is procuring fiber connection to the rest of the world (e.g., ALMA). The connection in Sao Paulo (and maybe Rio) is Tier 2. The instrument currently being commissioned that needs to archive the data at NERSC[76]. Department of Energy (DOE) commits to archive the data at NERSC (in San Diego). PB scale data (operation and simulation) flows per year from Chile to NERSC and then to SDSC (used in the past for the Polarbear project).
National Institute of Health (NIH)[77] connect with 3-6 centers regarding Cryo Instruments. The OSG processing and archiving proposal are currently unfunded, except the Wisconsin OSG Cryo (Myron L) project are NSF funded. Now, the Gryo community facing data preservation challenges (newly produced from computations and old raw data) and the constant change of the used software.

Discussion about future plans of how the SDN programmability could be used to accommodate future compliance implications (e.g., HIPPA) that needs to be addressed in a protected environment for specific OSG users (e.g., genomics and NIH). Currently, OSG has decided not to be involved in because of the ever-changing legal environment. Funding for the NIH project does not require strict cost-effective usage of OSG but uses other recourses instead when necessary. The questions expressed by OSG is how authentication services can be implemented so the end-user can decide who has access to the data (e.g., Origins at Columbia University). Current tools cannot address this complex research and development challenge of how to create a closed environment where institutions can share access selectively. A key technology to be taken into consideration is the one used by the InCommon community for such a scenario. Other science domain that can benefit from such OSG implementation will be Intellectual Property (IP) projects dictating sharing guidelines and astronomy projects that have particular IP regulations. For this moment, OSG is not planning to use the SDX programmability as a solution.

SDX Engineers Response to OSG Science Drivers (Origins) is to start with a discussion between Brazil science community to connect to OSG via AWAVE SDX's (FW) at a temporary OSG Origin. The data produces in Sao Paulo can be stored locally and bulk transferred to JLAB via a schedule. In this way, the connection can be provisioned upfront. The cluster size at UNESP for CMS requirements is ~3,000 cores.

The SDX programmability can offer to the OSG Origin stack possible bandwidth usage of a campus network, streaming, connection to IXPs, and multi-domain multi-path routing as a service. Precise characteristics of the OSG Origin (host functions and location) can define the correct hooks to establish an overlay network. Kubernetes deployment mechanism can be called by the application layer exploiting the SDX programmability. Another project, managed by Cees de Laat (the University of Amsterdam in Netherland), is involving namespaces implementation within the Kubernetes and introducing caches at the IXPs. The proximity of caches could also improve the performance of OSG. In general, OSG does not want to orchestrate the network below the caches,

[76] The National Energy Research Scientific Computing Center (NERSC): https://www.nersc.gov/
[77] National Institutes of Health (NIH): https://www.nih.gov/

add additional nodes simultaneously, OSG node to see each other, and use more of the Kubernetes functionality.
Some of those requests can be accommodated and automated by SDX by having an inventory of the switches that OSG uses, creating a pre-configuration of the network pieces, and assigning IPs. Another option could be to create a virtual network slice (except LHC) for individual Origins. The. limiting factor today is that the speed the cache could be filled up depends on the network. Therefore, network performance should be monitored closely. Different organizations and universities own origins, and they are using different hardware, which can create an interoperability issue. Internet2 uses a heterogeneous model for disks with the same federation but not comparable between I2 PoPs.
The SDX controller needs to have a mechanism to implement the current OSG topology, which only grows and when a new site (Origin or cache) is initiated (via Kubernetes) to add the site automatically. It takes 30 seconds to a min to bring a new OSG cache with Kubernetes, and it already has the network layers, IPs, and protocols pre-configured.
SDX can add value to the OSG network because it can connect a variety of networks at the IXP, and some of them go into the campuses where the Origins exist. I can reduce the labor necessary to interconnect sites between NREN, Regional network, and campus Science DMZ networks. A pilot test with Sao Paulo, New Zealand, and Europe (e.g., Netherlands) could be planned for the future. The number of caches in Europe is growing faster than in the USA. Data Origin is in the USA and computing in Europe. OSG participants prefer to have one Origin and using the closest available cache. SDX will not be a possible solution for data moving or replication. This problem could be addressed via a content delivery approach similar to Netflix and Akamai.
Another issue to be addressed is to deploy the right size of the cache when multiple requests (e.g., a huge number of CPUs at the site as Cardiff) hits the cache at once. For OSG it is easier to operate a single cache per Origin. When a result of computing has a large volume is also presents a problem of where to store that result. Most OSG participants manage a smaller output and prefer to use an inexpensive hardware solution (e.g., adding a FIONA box without changing anything else). Currently, the output problem could only be resolved if the organization has its own storage disk at the location that the output is produced.

The strategy for SDX implementations is:

- The main goal is to populate the OSG cache as quickly as possible
- Deploy the SDX Local Controller (LC) at each cache site
- Use path optimization to do the cache updates
- Router/ Switch in front of cache site would be SDX LC
- Use the cache on campus and Science DMZ
- SDX LC would be a pod in the Kubernetes host – resource footprint very negligible
- Deploy SDX Pod into Kubernetes on the same physical hosts as OSG deploys Origins and Caches (e.g., SDX Cache, SDX Origin) creating a parallel control plane
- OSG could benefit by understanding the TCP/IP (network performance) visibility on a per-flow basis
- Get a baseline now and compare to experiment before and afterward
- Measure performance of the flows and the impact on OSG customer
- Measure performance with Elastic Flow(Service name), sFlow stat, TCP metrics, Measurement Lab, Nagios, T Stack
- Apply data analytics to understand improvements if any

To answer the question of what can go into a Kubernetes Pod: Anything you can do on a host you can do in a Kubernetes pod. A Pod is a VM without the OS and it has one IP address. A Pod is an object that has containers (thin walls around a process) inside.

Locations & Customers
The table below shows the OSG caches and Origins location.

Table 1 OSG caches and Origins locations

Caches		**Origins**
U Chicago (bare metal)	UNL	* SDSC: Simons
Kansas City	Georgia Tech	* NCSA: DES
UCSD	Cardiff	**Caltech: LIGO - public
Manhattan	KISTI	*U Chicago: OSG General
Chicago PoP	Amsterdam PoP	FNAL: Dune, DES, Minerva (easiest)
Amsterdam	Houston PoP	***UNL: LIGO - private, maybe UNL General
Syracuse	Sunnyvale PoP	*Legend: *Easy, ** Harder,*** Very Hard*

The action items for implementing of SDX with OSG are:

- Decide on initial target application driver (e.g., FNAL, LIGO, OSG general)
- Decide on the data source (discuss with Phil Demar)
- OSG General (discuss with Rob Gardner)
- Possible use case with NCSA (DES, LIGO triggers public data distribution) (discuss with Don Petravick)
- Decide on test node to be used for the emerging services
- Research a way for SDX to help OSG feed its caches
- Deploy SDX Pod, SDX Origins, SDX Caches
- Deploy a control plane network at caches and Origins (10G to 100G)
- SDX's to offer virtualized (slices) network services
- Nautilus access to try test the solutions (discussion with John Graham)
- I2/StarLight can provide a node to be used to replicate LIGO Origin at Chicago (discuss with Edgar Fajardo)
- OSG Kubernetes pod can be deployed for testing and shut down when resources are repurposed (discuss with Dima Mishin)
- Translate workshop output into actionable items and regroup
- Follow up meeting at GRP Sept 17-18, 2019 at Calit2
- Other action items

How do we define success?
To define the successful implementation of OSG with SDX, the following points can be used:

- Provide more predictability for access to caches and Origins
- Provide measurement
- Get familiar with the mutual technologies and how they interplay

- Define a location where OSG-SDX can have an implementation model
- Simulate traffic where data can fill up one of the OSG caches
- Monitor network behavior with a use case
- Locate bottlenecks on the network and measure disk write speed
- Generate monthly traffic report integrated with the OSG
- Create a baseline for OSG traffic
- Discuss SDX bandwidth reservation (AtlanticWave SDX inter-domain only)
- Discuss SENSE implementation between domains
- Discuss the use of OSG Orchestrator for inter-domain bottlenecks for a real time report
- AtlanticWave SDX can provide a time series or Grafana page
- Define SDX topology locations with IP address to feed to SENSE

Appendix A. Program for the OSG-SDX Workshop

Wednesday, June 5, 2019

08:30 Registration (coffee kiosk is conveniently located outside the building - no breakfast is served)

- Room 5302, Atkinson Hall, UC San Diego

09:00 Welcome Address (15 minutes presentation, 5 minutes Q&A) - Larry Smarr | Download presentation
09:20 Context and Goals of the workshop (15 minutes presentation, 5 minutes Q&A) - Julio Ibarra
09:40 OSG Keynote (40 minutes presentation, 10 minutes Q&A) - Frank Wuerthwein | Download presentation

- *What are the challenges OSG has in its landscape and foresees in its roadmap?*

10:30 Refreshment break (30 minutes)- (*coffee kiosk is conveniently located outside the building*)

11:00 NSF IRNC Software Defined Exchange Panel (Introduction, 10 minutes) - Julio Ibarra (Moderator)

- *Introduction about each of the IRNC SDX projects: Scope, Goals, Accomplishments, etc.*

11:10 StarLight-SDX (20 minutes presentation) - Jim Hao Chen | Download presentation
11:30 PacifiWave-SDX (20 minutes presentation) - John Hess | Download presentation
11:50 AtlanticWave-SDX (20 minutes presentation) - Julio Ibarra | Download presentation
12:10 Q&A session (30 minutes)

12:30 Lunch (1 hour) -(*wear comfortable shoes we will walk to the food court*)

13:30 Integration Panel (15 minutes presentation, 5 minutes Q&A for each presenter) Tom DeFanti (moderator)

- *Architecture, Components, Interfaces and other relevant information to reveal what resources are available, and also not available, for potential integration*

13:40 OSG Layers and components - Edgar Fajardo | Download presentation
14:00 Kubernetes Layers and components - John Graham | Download presentation
14:20 AtlanticWave-SDX layers and components - Yufeng Xin | Download presentation
14:40 StarLight-SDX layers and components - Jim Hao Chen | Download presentation

15:00 Refreshment break (30 minutes)- (*coffee kiosk is conveniently located outside the building*)

15:30 PacificWave-SDX layers and components - John Hess | Download presentation
15:50 Integration panel Q&A and wrap up (20 minutes)
16:10 Roadmap to the integration with OSG: Plenary

- *Describe and represent what OSG and SDXs want to accomplish. The aim of this session is to provide input to Thursday's design and integration session.*

17:10 Plans for day 2 of the workshop Julio Ibarra, Tom DeFanti, Frank Wuerthwein

18:00 Dinner (on your own). Please see below some suggested restaurants. **A full list can be found here.**

- Bella Vista (tasty nice outdoor seating, across Torrey Pines); Home Plate (has great hamburgers), by RIMAC; OceanView (pizzas); The Bistro (sushi); James' Place Prime (Seafood Sushi by the theater complex is really good); Rock Bottom (if you like noisy sports bar places).

Thursday, June 6, 2019

08:30 Registration
09:00 Integration, Design, and Planning Jam Session - Moderator: Julio Ibarra

- *The goal of this session is to design how to integrate the SDXs with Kubernetes and OSG. Detailed diagrams and descriptions about how to accomplish the integration should be an outcome.*

Appendix B. List of Participants

In Person:

1. **Balcas, Justas** - Software Engineer California Institute of Technology (Caltech) - (juztas at gmail.com)
2. **Bezerra, Jeronimo** - IT Assistant Director Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU) - (jbezerra at fiu.edu)
3. **Breen, Joe** - Senior IT Architect University of Utah - (joe.breen at utah.edu)
4. **Cevik, Mert** - Sr.Linux Network Administrator University of North Carolina (UNC) Chapel Hill - RENCI - (mcevik at renci.org)
5. **Chen, Jim** - Associate Director iCAIR/Northwestern University - (jim-chen at northwestern.edu)
6. **Chergarova,** Vasilka - Research Coordinator Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU) - (vchergar at fiu.edu)
7. **Davila, Diego** - Computational Data Science Research Specialist San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD) - (didavila at ucsd.edu)
8. **DeFanti, Thomas** - Research Scientist University of California San Diego (UCSD) - (tdefanti at ucsd.edu)
9. **Donovan, Sean** - Research Scientist Georgia Institute of Technology (Georgia Tech) - (sdonovan at gatech.edu)

10. **Fajardo, Edgar** - OSG Developer San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD) - (emfajard at ucsd.edu)
11. **Hess, John** - Network Engineer CENIC - Pacific Wave - (jhess at cenic.org)
12. **Hutton, Thomas** - Network Architect San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD) - (hutton at ucsd.edu)
13. **Ibarra, Julio** - Assitant VP Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU) - (julio at fiu.edu)
14. **Leal, Beraldo** - Systems Engineer Sao Paulo Research and Analysis Center (SPRACE) Brazil - (beraldo.leal at cern.ch)
15. **Mishin, Dima** - Applications developer University of California San Diego - (dmishin at ucsd.edu)
16. **Morgan, Heidi** - Senior Computer Scientist Information Science Institute (ISI) USC - (hlmorgan at isi.edu)
17. **Moya, Andress** - Computing Research assistant at Caltech Tier2 California Institute of Technology (Caltech) - (amoya at caltech.edu)
18. **Paolini, Christopher** - Assistant Professor of Electrical Engineering San Diego State University (SDSU) - (paolini at engineering.sdsu.edu)
19. **Polizzi, Joel** - Visualization Engineering Technician University of California San Diego (UCSD) - (jpolizzi at eng.ucsd.edu)
20. **Ravi, Srivatsan** - Research assistant professor Information Science Institute (ISI) University of Southern California (USC) - (sravi at isi.edu)
21. **Smarr, Larry** - Director California Institute for Telecommunications and Information Technology (Calit2) at University of California San Diego (UCSD) - (lsmarr at ucsd.edu)
22. **Thompson, Kevin** - PD National Science Foundation (NSF) - (kthompso at nsf.gov)
23. Weekley, Jeffrey - Director of CI & RC University of California Merced - (jdweekley at ucmerced.edu)
24. **Wuerthwein, Frank** - Executive Director of the Open Science Grid San Diego Supercomputer Center (SDSC) at University of California San Diego (UCSD) - (fkw888 at gmail.com)
25. **Xin, Yufeng** - Sr. Scientist University of North Carolina (UNC) Chapel Hill - RENCI - (yxin at renci.org)

Remote:

1. **Graham, John** - System Integration Engineer University of California San Diego (UCSD) - (jjgraham at eng.ucsd.edu)
2. **Sanders, Matt** - Research Scientist Georgia Institute of Technology (Georgia Tech) - (msanders at gatech.edu)
3. **Stealey, Michael** - Distributed Systems Software Engineer University of North Carolina (UNC) Chapel Hill - RENCI - (stealey at unc.edu)

Appendix C. Acronyms

ALMA	Atacama Large Millimeter Array	NOAA	National Oceanic and Atmospheric Administration
AMIS	Advanced Measurement Instrument and Services	NRP	National Research Platform
ANSP	Academic Network of Sao Paulo	NSF	National Science Foundation
AURA	Association of Universities for Research in Astronomy	NSRC	Network Startup Resource Center
BEN	Breakable Experimental Network	OGF	Open Grid Forum
Calit2	California Institute for Telecommunications and Information Technology	OGF	Open Grid Forum
CALREN	California Research and Education Network	ONOS	SDN Controller platform
Caltech	California Institute of Technology	ONS	Open Storage Network
CENIC	Corporation for Education Network Initiatives in California	ORCA	Open Resource Control Architecture
CHASE-CI	Cognitive Hardware and Software ecosystem Community Infrastructure	OSDC	Open Science Data Cloud
CI/CD	Continuous Integration & Continuous Delivery	OSG	Open Science Grid
CIARA	Center for Internet Augmented Research and Assessment	OSiRIS	NSF pilot project NSF to evaluate a SDN CI
CMS	Compact Muon Solenoid	PAO	Pierre Auger Observatory
CNI	Core Network Insert	PNWGP	Pacific Northwest Gigapop
CPU	Central Processing Unit	POLARBEAR	Cosmic Microwave Background polarization experiment
CVMFL	CernVM File System	PoP	Point of Presense
DDS	Document Distribution Service	POSIX	Portable Operating System Interface
DECam	Dark Energy Camera	PRP	Pacific Research Platform
dHTC	Distributed high throughput computing	RENCI	Renaissance Computing Institute at UNC
DTN	Data Transfer Node	RM	Resource Manager
DUNE/protoDUNE	Neutrino experiment	RNP	Rede Nacional de Ensino e Pesquisa REN Brazil
EdgeFS ISGW	Inter-Segment GateWay	RSO	Research Support Organizations
FES	Fusion Energy Sciences	SD-WAN	Software Defined Wide Area Network
FIU	Florida International University	SDSC	San Diego Supercomputer Center
FNAL	Fermi National Accelerator Laboratory	SDSS	Sloan Digital Sky Survey
FPGA	Field-Programmable Gate Array	SDSU	San Diego State University
FPGAs	Field Programmable Gate Arrays	SDX	Software Defined Exchange
GIT	Georgia Institute of Technology	SENSE	SDN for End-to-End Networking Exascale
GlueX, SPT, Simons	Multi-institutional Science Teams	SKA	Square Kilometer Array
GMT	Giant Magellan Telescope	SoX	South America eXchange point
GPU	Graphic Porcessing Unit	SPRACE	Sao Paulo Research and Analysis Center
GRP	Global Research Platform	TANet	Taiwan Academic Network

GSI	Grid Security Infrastructure	TPU	Tensor Processing Unit
GT	Georgia Institute of Technology	UC	University of Chicago
HA	High-Availability	UC Merced	University of California Merced
HEP	High Energy Physics	UCI	University of California, Irvine
HIPPA	Health Insurance Portability and Accountability Act of 1996	UCSD	University of California San Diego
HTCondor	High Throughput Computing	UI	User Interface
IaaS	Infrastructure-as-a-service	UNC	University of North Carolina
iCAIR	International Center for Advanced Internet Research	UNL	University of Nebraska–Lincoln
ISI	Information Science Institute	uPA	Ultimate provider agent
LC	Local Controller	US-ATLAS	Big science projects like CMS, LIGO, IceCube
LHCOPN	Large Hadron Collider Optical Private Network	USC	University of Southern California
LIGO	Laser Interferometer Gravitational-Wave Observatory	UU	University of Utah
LSST	Large Synoptic Survey Telescope	vCPE	Virtual Customer Premises Equipment
MEICAN	Management Environment for Inter domain Circuits in Advanced Networks	VERITAS	Very Energetic Radiation Imaging Telescope Array System
NARLabs	National Applied Research Laboratories	VFC	Virtual Forwarding Contexts
NCSA	National Center for Supercomputing Applications	WLCG	Worldwide LHC Computing Grid
NERSC	National Energy Research Scientific Computing Center	XENON	Dark Matter Project
NFV	Network Function Virtualizations	XrootD	Protocol and can use Squid proxies
NIH	National Institute of Health		

References

Gupta, A., Vanbever, L., Shahbaz, M., Donovan, S. P., Schlinker, B., Feamster, N., . . . Katz-Bassett, E. (2014). SDX: a software defined internet exchange. *SIGCOMM Comput. Commun. Rev., 44*(4), 551-562. doi:10.1145/2740070.2626300