Journal of Physics: Conference Series

PAPER • OPEN ACCESS

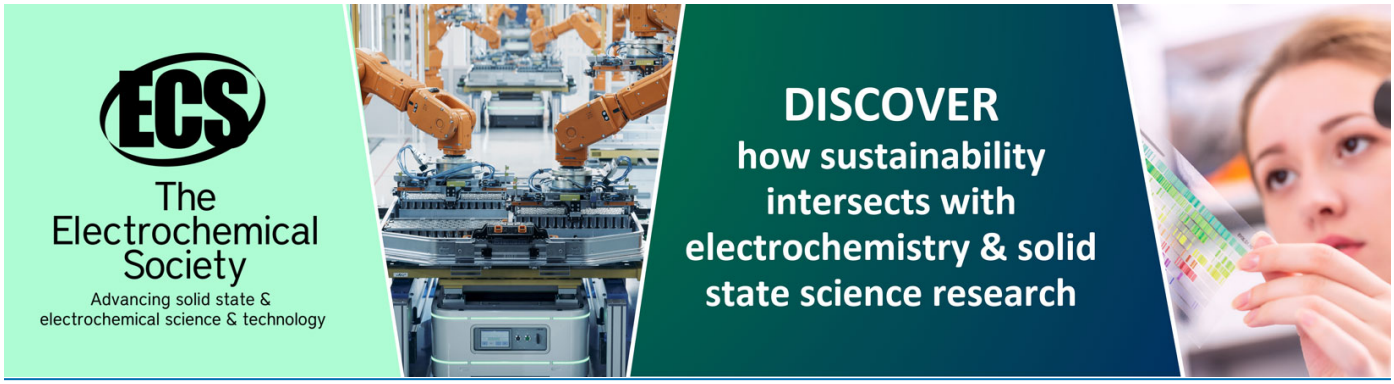
SDN-NGenIA, a software defined next generation integrated architecture for HEP and data intensive science

To cite this article: J Balcas *et al* 2017 *J. Phys.: Conf. Ser.* **898** 112001

View the article online for updates and enhancements.

You may also like

- Scheduling Data Flow between Data Centers Based on Software Defined Network
Siquan Hu, Zhao Huang and Zhiguo Shi
- Automatically reconfigurable optical data center network with dynamic bandwidth allocation
Xuwei Xue, Kristif Prifti, Bitao Pan et al.
- A Survey on Emerging Software-Defined Networking and Blockchain in Smart Health Care
Shivani Wadhwa, Himanshi Babbar and Shalli Rani



This content was downloaded from IP address 208.127.70.16 on 09/05/2024 at 05:43

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 112001 doi:10.1088/1742-6596/898/10/112001

SDN-NGenIA, a software defined next generation integrated architecture for HEP and data intensive science

J Balcas, T W Hendricks, D Kcira*, A Mughal, H Newman, M Spiropulu, J R Vlimant

California Institute of Technology, Pasadena, CA 91125, USA

E-mail: * dkcira@caltech.edu

Abstract. The SDN Next Generation Integrated Architecture (SDN-NGeNIA) project addresses some of the key challenges facing the present and next generations of science programs in HEP, astrophysics, and other fields, whose potential discoveries depend on their ability to distribute, process and analyze globally distributed Petascale to Exascale datasets. The SDN-NGenIA system under development by Caltech and partner HEP and network teams is focused on the coordinated use of network, computing and storage infrastructures, through a set of developments that build on the experience gained in recently completed and previous projects that use dynamic circuits with bandwidth guarantees to support major network flows, as demonstrated across LHC Open Network Environment [1] and in large scale demonstrations over the last three years, and recently integrated with PhEDEx and Asynchronous Stage Out data management applications of the CMS experiment at the Large Hadron Collider. In addition to the general program goals of supporting the network needs of the LHC and other science programs with similar needs, a recent focus is the use of the Leadership HPC facility at Argonne National Lab (ALCF) for data intensive applications.

1. Introduction

The SDN-NGeNIA project aims to enable the LHC and other leading programs in high energy physics and other global science domains funded by the Department of Energy (DOE) to operate with a new level of efficiency and control, through the development of a Next Generation Integrated Architecture (NGenIA) based on intelligent software defined network (SDN)-driven network systems coupled to high throughput applications. While the initial focus is on the challenging LHC use case, the systems and products being developed are general, and apply to many fields of data intensive science ranging from astrophysical sky surveys to bioinformatics and earth observation, as well as other organizations facing the challenges of extracting knowledge from distributed multi-Petabyte data stores.

A central concept in this development program is a new paradigm *consistent network operations* among widely distributed computing and storage facilities. In these facilities, stable high throughput flows at set rates on load-balanced network paths, up to flexible high water marks. These marks are adjusted in real time to accommodate other network traffic. The large smooth flows are launched and managed by SDN services that act in concert with the experiments site-resident data distribution and management systems, to meet the expanding


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

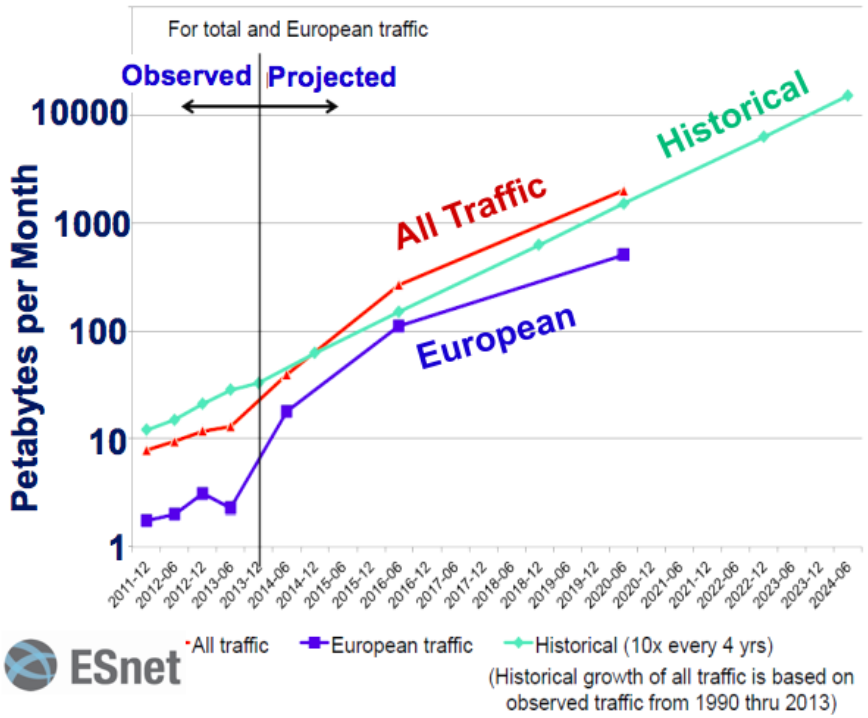


Figure 1. Observed and projected traffic over the ESnet national research and education network. A data volume of 43 PB per month was observed in April 2016. The rate of increase exceeds the historical trend of 10 times increase every 4 years.

needs of the science programs. A more detailed description of the concept and components of SDN-NG enIA is given in [2].

2. New challenges in exascale data and computing

The largest science datasets today from the Large Hadron Collider (LHC) are around 300 Petabytes. Exabyte datasets are on the horizon by the end of the LHC Run2 in 2018. During the High-Luminosity LHC (HL LHC) [3] era that starts in 2025 the size of the datasets will grow by an additional 100 times, reaching the range of 50-100 Exabytes. The rate of traffic growth over the national research and education networks (NRENs) that support the LHC and other science programs is given in figure 1. It shows the observed and projected traffic on the Energy Sciences Network (ESnet) [4] that connects the national labs in the united states and peers with European and other NRENs.

At the present stage of the LHC running we are already dependent on the high performance computer networks to be able to distribute the data to the data centers of the LHC Grid all over the world. With the increase of the volume of the data described above this dependency will become even larger. In addition, other fields of science will have growing needs (see also Section 3 below). There will be, therefore, a stiff competition for the use of large but limited network resources.

3. The future of big data

As mentioned already, in addition to the HL LHC, other fields of science will have big amounts of data, in fact they will probably eclipse the LHC. A table from [7] is shown in figure 2. It has

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Figure 2. Big Data: Astronomical or Genomical? A summary table of big data circa 2025 [5].

a summary of the big data circa 2025. For comparison, the HL LHC with have 2-10 EB data per year at the storage level, data acquisition at 10 TB/s, offline of the order of 0.1 TB/s and will require 0.065 to 0.2×10^{12} CPU hours for analysis.

4. Next generation integrated systems for exascale science

A major opportunity for helping solve the challenges of the exascale data and computing is the synergy among:

(i) Global operations data and workflow management systems developed by the High Energy Physics (HEP) programs, geared to work with increasingly diverse and elastic resources to respond to peak demands.
   - These are enabled by distributed operations and security infrastructures
   - They ride on high-capacity but presently mostly passive networks.

(ii) deeply programmable, agile *software defined networks* (SDN) emerging as multi-domain operating systems. Furthermore, new network paradigms are emerging that focus on the content, such as Content Delivery Networks and Named Data Networks,

(iii) Machine learning, modeling and simulation as well as game theoretical methods. A few steps are foreseen here: extraction of key variables, optimization, then real-time self-optimizing workflows.

A service diagram of NGenIA with the various components mentioned above as well as the Leadership and HPC Facilities (see Section 6), Cloud and other opportunistic resources is shown in figure 3.

5. SDN-NGenIA vision

SDN-NGenIA aims at building a distributed environment, where resources can be deployed flexibly to meet the demands. A natural path to this vision is *Software Defined Networking (SDN)*. SDN separates the functions that control the flow of traffic from the switching infrastructure that forwards the traffic through the use of open deeply programmable SDN controllers. This strategy has many benefits:

- It replaces stovepiped vendor hardware/software solutions with open platform-independent software services;
- It virtualizes services and networks, lowering cost and energy, with greater simplicity;
- It adds intelligent dynamics to system operations;
- It is already a major direction of network research and industry.

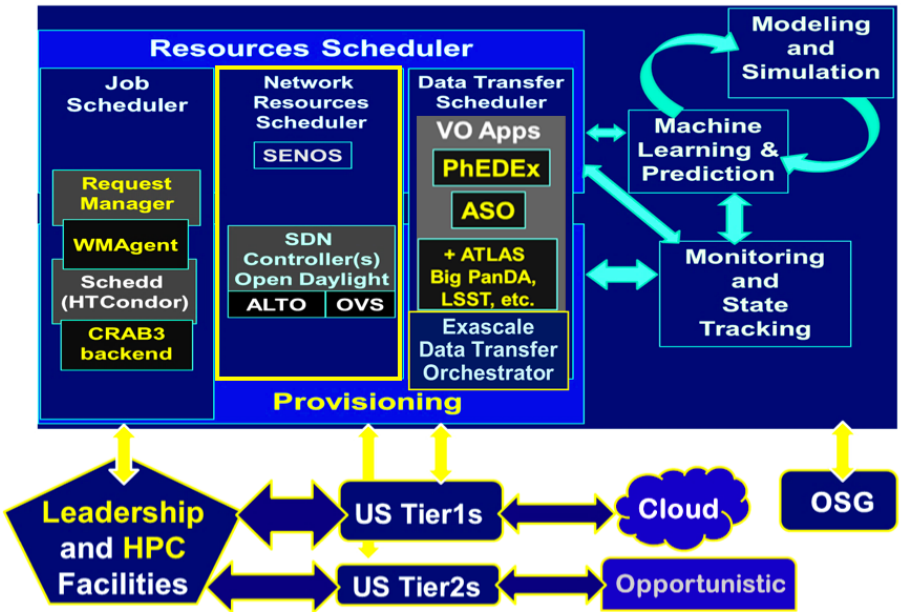


Figure 3. Components of NGenIA, including external resources such as LCF, Cloud, and other opportunistic.

The system envisioned here has built-in intelligence and will require excellent monitoring at all the levels.

6. Leadership computing facilities

Part of the NGenIA strategy is based on the need to develop and deploy new data- and network-intensive operational modes. Part of the solution is the use of the US Leadership Computing Facilities (LCFs). The Caltech team has already pilot projects with the Argonne Leadership Computing Facility (ALFC) [6] to test running workflows from the CMS experiment on the HPC clusters at Argonne.

In order to develop this vision, the key challenges from the client site and science Virtual Organization side (using the HEP example) are:

- Recasting HEP's generation, reconstruction and simulation codes, case by case, to adapt to the emerging High Performance Computing (HPC) architectures, addressing issues of memory, dataflow versus CPU etc.
- Identifying and matching the units of work in HEP's workflow to the specific HPC resources or sub-facilities well-adapted to the task (after the code recasting step)
- Building dynamic and adaptive just-in-time systems that respond rapidly (on the required timescale) to offered resources as they occur.
- Developing algorithms that effectively co-schedule CPU, memory, storage, IO port, local and wide area network resources.
- Developing an appropriate security infrastructure, and corresponding system architectures in hardware and software, that meet the security needs of the LCF.
- Applying machine learning to optimize the workflow of the HEP experiments, using self-organizing system methods which are well-adapted to such problems; while also taking the special parameters, conditions, and restrictions of LCF into account as part of the workflow.

- Exploiting the intense ongoing development of virtualized computing systems, networks and services in the research community and in industry: in the data center, campus and wide area network space aimed at coherent distributed system operations (including software defined networking, network function virtualization, and service chaining, along with emerging higher level concepts).

The challenges on the LHC and HPC facilities mirror those described above. In addition, another key issue for the LCFs is a new concept of secure ways to bridge the site edge, such as next generation Science Demilitarized Zones (DMZs) [8] or similar edge-bridging methods.

7. Open vSwitch for managing site interactions locally and over the WAN

Open vSwitch (OvS) [9] is a virtual switch licensed under the open source Apache 2.0 license, which is already part of the main Linux distributions. Furthermore, OvS understands OpenDaylight (ODL) [10], which is presently the most popular software platform for developing SDN applications and the one we are using in this project.

Part of the challenge for building an end-to-end SDN system is the configuration of data flows all the way to the end host. Through the use of OvS we will be able to orchestrate end-to-end configuration of data flows. They can be orchestrated from the local/campus SDN controller or brought down from the regional/WAN controller.

OvS provides quality of service (QoS) and traffic shaping right at the end-point of a data transfer. QoS via OvS is protocol agnostic: one can use TCP (GridFTP [11], FDT [12]) or UDP. The use of OvS helps to achieve better throughput by moderating and stabilizing data flows; e.g. in cases where the upstream switches have limited buffer memory. Under the hood, OvS uses the TC (Traffic control) part of iproute2 to configure and control the Linux kernel network scheduler. Monitoring is done with standard sFlow and/or NetFlow protocols.

8. Consistent network operations in OpenDaylight

The work described in this section is done in collaboration with the group of Prof. Richard Yang at Yale. The idea is to allow the tools of the CMS experiment at the LHC to interface with the OpenDaylight (ODL) controllers. The application layer traffic optimization (ALTO) [13] developed by Yang *et al* provides network information with the goal of modifying resource consumption patterns, while improving application performance.

The CMS tools will collaborate with an orchestrator (ExaO) developed by Caltech and Yale for the part related to network configuration and data transfers, while retaining the parts related to user requests, data locations and enforcement of policies. A schematic of the setup for testing ExaO is shown in figure 4.

Two additional ingredients of this solution are the network resident site abstraction (RSA) and control/path calculator (SPCE). They receive data requests and raw network state; compute on-demand, dynamic inter-domain network abstraction; and enforce application policies in networks. Furthermore, they present to the orchestrator only the necessary information about a site by hiding the details that would make optimal path calculations computationally expensive.

The consistency with the end hosts is reached using OvS as described in the previous section.

Summary

The NGENIA-ES program as envisioned will:

- Develop a synergy and convergence between data intensive science and exascale computing;
- Build a new class of intelligent, agile network systems;
- Generate novel, data-intensive workflows, accelerating the time to discovery of major science programs;

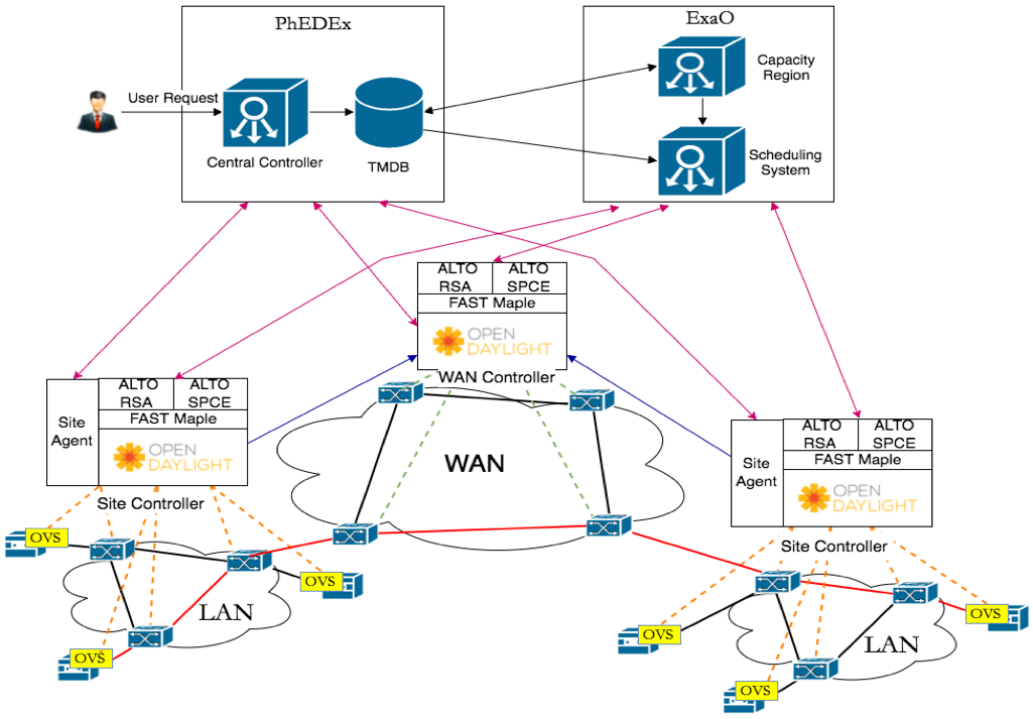


Figure 4. Consistent operations between tools of the CMS experiment and the SDN components as well as between SDN controllers and end-hosts.

- Work together with Leadership Computing Facilities to create Computing, Storage and Network (CSN) ecosystems for next-generation data intensive science;
- Develop new modes of network operations that promise to redefine the state of the art in high throughput while remaining compatible with the tide of smaller flows exchanged over the worlds research and education networks;
- Create new high throughput workflow and global system control and optimization methodologies, coupled to novel proactive, reactive and predictive Software Defined Network system designs; and
- Use data-driven methods both for optimizing the workflow of the science experiments and for scheduling and optimization of the network resources.

Acknowledgments
The work presented in this paper was supported through the following projects from the U.S. Department of Energy:

- OLiMPS, DOE/ASCR, DOE award # DE-SC0007346,
- DOE/ASCR SDN NGenIA, project id 000219898,
- SENSE, FNAL PO # 626507 under DOE award # DE-AC02-07CH11359,

and from the National Science Foundation:

- ANSE, NSF award # 1246133,
- CHOPIN, NSF award # 1341024,
- US CMS Tier2, NSF award # 1120138.

References

[1] LHC Open Network Environment, lhcone.web.cern.ch
[2] Newman H, Spiropulu M, Balcas J, Kcira D, Legrand I, Mughal A, Vlimant J, and Voicu R, "Next-Generation Exascale Network Integrated ArchiFaltotecture for Global Science [Invited]," J. Opt. Commun. Netw. 9, A162-A169 (2017).
[3] Rossi L and Brüning O, "High Luminosity Large Hadron Collider A description for the European Strategy Preparatory Group," 1 Aug. 2012.
[4] ESnet, Energy Sciences Network, www.es.net
[5] Stephens, Zachary D., et al. "Big data: astronomical or genomical?." PLoS Biol 13.7 (2015): e1002195.
[6] Argonne Leadership Computing Facility, ALCF, www.alcf.anl.gov
[7] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. "Big Data: Astronomical or Genomical?" PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195 (2015)
[8] "Science DMZ," https://fasterdata.es.net/science-dmz
[9] Open vSwitch (OvS), openvswitch.org
[10] OpenDaylight (ODL), www.opendaylight.org
[11] Allcock, William, et al. "GridFTP: Protocol extensions to FTP for the Grid." Global Grid ForumGFD-RP 20 (2003): 1-21.
[12] Maxa Z, Ahmed B, Kcira D, Legrand I, Mughal A, Thomas M and Voicu R, "Powering physics data transfers with FDT," Journal of Physics: Conference Series 052014; fdt.cern.ch, 2011.
[13] Yang R, "Application-Layer Traffic Optimization (alto)," https://datatracker.ietf.org/wg/alto.