Next Generation Exascale Network Integrated Architecture for HEP and Global Science [Invited]

Harvey Newman (*), M. Spiropulu, J. Balcas, D. Kcira, I. Legrand,
A. Mughal, J. R. Vlimant, R. Voicu
High Energy Physics, California Institute of Technology
Pasadena, CA 91125

***Abstract*— NGENIA-ES is specifically designed to accomplish new levels of capability in support of global science collaborations through the development of a new class of intelligent, agile networked systems. Its path to success is built upon our broad based ongoing developments in multiple areas, strong ties among our high energy physics, computer and network science and engineering teams, and our close collaboration with key technology developers and providers deeply engaged in the National Strategic Computing Initiative (NSCI).**

Distributed exascale computing; Large Hadron Collider; Grid computing; Software defined networking;

I. Global Challenges in Data Intensive Research

We are entering a new era of exploration and discovery in many fields, from high energy physics (HEP) and astrophysics to climate science, genomics, seismology and biomedical research, each with its own complex workflow requiring massive computing, data handling and networks. The continued cycle of breakthroughs in each of these fields depends crucially on our ability to extract the wealth of knowledge, whether subtle patterns, small perturbations or rare events, buried in massive datasets whose scale and complexity continue to grow exponentially with time.

In spite of technology advances, the largest data- and network-intensive programs supported by the DOE and partner agencies, including the Upgraded High Luminosity LHC [1] program, the Large Synoptic Space Telescope (LSST) [2] and the Square Kilometer Array (SKA) [3] astrophysics surveys, photon-based sciences, the Joint Genome Institute applications, the Earth System Grid and any other data-intensive emerging areas of growth, face unprecedented challenges: in global data distribution, processing, access and analysis, in the coordinated use of massive but still limited computing, storage and network resources, and in the coordinated operation and collaboration within global scientific enterprises each encompassing hundreds to thousands of scientists.

The most data intensive program is currently the LHC high energy physics program, with more than 300 petabytes under management at 160 sites in the US and around the world, growing to an estimated volume of one Exabyte by the time the data taking run now underway completes in 2018. The data storage and computational needs are projected to grow by another two orders of magnitude by the HL LHC era in the middle of the next decade. This leads to a critical need to develop and deploy new data- and network-intensive operational modes making effective use of the US Leadership Computing Facilities (LCFs) starting now, while continuing to expand and refine the system driving data exchange with the LCFs as they progress through the pre-exascale and exascale stages and beyond, while taking full advantage of the state of the art in high performance storage and networks across several technology generations.

It is also to be expected that the SKA and the other programs cited, and the rise of societal developments such as genomics in support of precision medicine and sequencing of other species' genomes, may match or eclipse HEP's needs within the next decade. This makes the developments summarized in this paper, which are designed to accommodate multiple science workflows while making full use of the available network capacity and server throughput capabilities during each technology generation, all the more pressing.

II. Global Networked Systems for Next Generation Science

Our long term strategy is based on co-design of the methods that make best use of the network and computing and storage infrastructures, together with data structures and real-time adaptive algorithms. Rather than writing code for a distributed system assumed to be static and rigid,

Manuscript received June 10, 2016.

Harvey Newman is a professor of physics at the Department of Physics, Mathematics and Astronomy at the California Institute of Technology, Pasadena, CA 91125, USA. newman@hep.caltech.edu

the success of these programs will depend on the efficient interplay of software with an elastic and diverse set of resources – CPU, storage, and network. In finding an overall optimal solution, new modes of steering, use (and reuse) of data products produced and consumed at many locations, new modes of propagating information on data product availability and the cost of delivery versus re-computation in real time, and interactions among user groups, end sites and the network as a system will need to be developed.

The crux of the solution to this generational challenge lies in the remarkable synergy emerging between:

• Deeply programmable, agile software-defined network (SDN) infrastructures which are evolving towards multi-service multi-domain network "operating systems" interconnecting science teams across regional, national and global distances, and

• Worldwide distributed systems developed by the data intensive science programs, harnessing global workflow, scheduling and data management systems they have developed, which are enabled by distributed operations and security infrastructures riding on high capacity (but still-passive) networks.

As in many revolutions, the groups working at the intersection of their domain science and computational science and technology will have a crucial role.

A new overarching concept is one of "consistent operations", where the experiments' workflow management systems will be deeply network aware, reactive and proactive, responding to moment-to-moment feedback on actual versus estimated task progress, state changes of the networks and end systems, and a holistic view of workflows with diverse characteristics and requirements serving many fields. This will enable the science programs to develop a new more efficient operational paradigm based on software-driven bandwidth allocation, load balancing, flow moderation and topology reconfiguration on the fly where needed, leading to full use of the available network as well as computing and storage infrastructures while avoiding saturation and blocking of other network traffic.

While the systems to be developed should be targeted at many programs, taking diverse "process of science" paradigms into account, one fertile area for development (as well as progressive large scale field testing) is the LHC program, which is now in its second three year run, anticipated to yield a new round of groundbreaking discoveries, as well as a new level of "global data and network intensity". This is complemented by the very different but equally challenging real-time workflows in diverse fields, including bioinformatics, computational astrophysics, radio astronomy, and oceanic and atmospheric sciences.

III. Leadership Computing, Storage and Network (CSN) Ecosystems for Next Generation Data Intensive Science

To gauge the great opportunity in terms of CPU resources for the HEP program (using the CMS example at the LHC) one only has to recall that the CPU requirements are expected to grow by 65 to 200 times between now and the HL LHC [1], while the affordable CPU power obtainable within a fixed budget, including Moore's law and possible code improvements, is estimated to be an order of magnitude less.

The key issues to develop this vision include:

From the client site and science Virtual Organization side (using the HEP example):

• Recasting HEP's generation, reconstruction and simulation codes, case by case, to adapt to the emerging 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 LCFs 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 key issues for the LCF and other HPC facilities such as NERSC mirror several of the elements, and include, from the HPC facility side:

• 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 recasting step)

• Building dynamic and adaptive "just in time" systems that respond rapidly (on the required timescale) to offered demands as they occur; including client-side/server-side coordination for a consistent outcome

• Developing algorithms that effectively co-schedule CPU, memory, storage, IO port, local and wide area network resources; with the necessary coordination as above

• Developing an appropriate security infrastructure, and corresponding system architectures in hardware and software, that meet the security needs of the LCF. For the LCFs this means adopting a new mode of ongoing service to a major client in quasi-real time, in a way that can be adapted to meet the LCF's requirements.

• Applying machine learning coupled to game theory and system modeling, to optimizing 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 LCFs

into account as part of the workflow.

• Exploiting the intense ongoing developments of virtualization of computing systems and services in the research community and in industry: in the case of the LCFs, the recent developments of "site orchestration" of virtualized resources, and even newer concepts of secure ways to bridge the site edge, such as next generation Science DMZs [4] or similar edge-bridging methods are relevant.

IV. LCF-Edge Data Intensive System Operational Model

A promising direction centers on the use of a new class of LCF-Edge Data Intensive Systems. The use of secure systems at the site perimeter means that security (both human and AI) and countermeasures where needed can be focused on a limited number of subsystems and software entities (proxies), so that the manpower burden may be acceptable.

The operational concept is that HEP data be brought into the edge systems in petabyte-size chunks, far enough in advance so that the data is always waiting and ready when the corresponding jobs are scheduled to start. Multiple chunks for different stages of the overall workflow are foreseen, with each chunk identified to have a certain provenance and certain attributes (such as the ratio of CPU to I/O requirements) so that clusters of chunks are matched to an HPC subsystem configured to match the attributes while working with high efficiency of utilization. At a later stage, once can also foresee dynamic restructuring of the HPC resources, especially if they are virtualized in logical "sectors".

Considering the parameters in this problem yields interesting consequences. As of today, a 1 petabyte chunk would occupy a 100 Gbps link if used to 100% capacity for a full 24-hour day. Given the 300 petabytes currently stored by the LHC experiments and the fact that approximately 250 petabytes flowed over the networks in and out of the US in the past year, the 1 petabyte chunks each represent a relatively small "data transaction" compared to the whole task at hand, and so one would like to transport many chunks to and from the LCF. A typical configuration would thus preferably include several 100 Gbps links today, migrating to several 400 Gbps links within 3-5 years, and several 1 Tbps links by the startup of the High Luminosity LHC a decade from now, depending on the demand evolution and the cost evolution during this period.

As a result, the use and network requirements of such Data Intensive Leadership Facilities will no doubt present a significant challenge and equally well, an opportunity for the conception and development of the next generation of intelligent networked systems supporting data intensive science.

V. Key System Elements and Concepts

The key elements the NGENIA-ES architecture include:

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

VI. 6. Foundations and Progress

SuperComputing 2015: A Pilot with Terabit/sec Transactions for LCFs in the Pre-Exascale Era

During the November Supercomputing 2015 conference in Austin, an international team of Caltech, SPRACE Sao Paulo and the Univ. of Michigan, together with teams from FIU, Vanderbilt and support from vendors including Dell, Mangstor, Mellanox, QLogic, SGI and Spirent worked to demonstrate large data flow transfers across a highly intelligent SDN network. The networks for this work were supported by SCinet Network Research Exhibition (NRE), ESnet, Century Link, CENIC and Pacific Wave

The global picture of SC15 demonstrations involving various WAN paths is shown in Fig. 1. The SDN demonstrations revolved around an OpenFlow ring connecting seven different booths and the WAN connections to Caltech and other campuses in the Pacific Research Platform, Michigan, Starlight, CERN and Sao Paulo. Some of the WAN connections were built using NSI [6] dynamic circuits and stitched together to form end-to-end paths using a custom SDN application. All the remote switches were controlled by a single controller in the Caltech booth on the show floor. The results were presented at the ESnet INDIS [16] and SDN [17] workshops during the conference. A paper [18] was published in IEEE ACM.

While terabit/sec aggregate flows were already achieved during Supercomputing 2014, the SC15 demonstrations included the use of a large set (29) of the very first 100 Gbps network interfaces on servers, as needed to support the large data transactions foreseen with the Argonne Leadership Computing Facility (ALCF) in the pre-exascale era using Theta and subsequently Aurora in 2016-19, on the way to exascale operations with petabyte data transactions. The deployment at SC15 included several Dell and Supermicro servers each capable of stable bidirectional flows to and from a single port of greater than 100Gbps (illustrated in Fig. 2) and stable aggregate flows per server of > 300 Gbps using four network interfaces and Caltech's FDT. The overall throughput capability deployed

in one rack at SC15 was 1.5 Terabits/sec, matching (and exceeding) the local network connections.

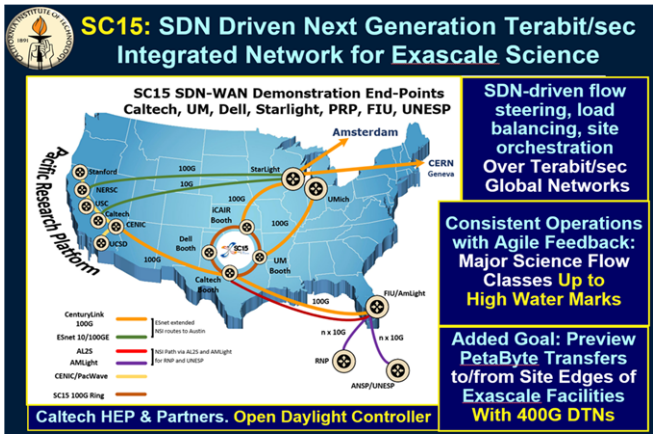

Fig. 1. Global picture of the Exascale NGENIA prototype demonstration at SC15.

The design of servers capable of 1 Terabit/sec each, appropriate for use with exascale-era computing systems and next (to next) generations of networks is currently underway with Orange Laboratories (Silicon Valley) and EchoStreams, to be demonstrated at SC16.

Network-Endsite Flow Rate Limits up to 100G Wire Speed with Open vSwitch (OVS)

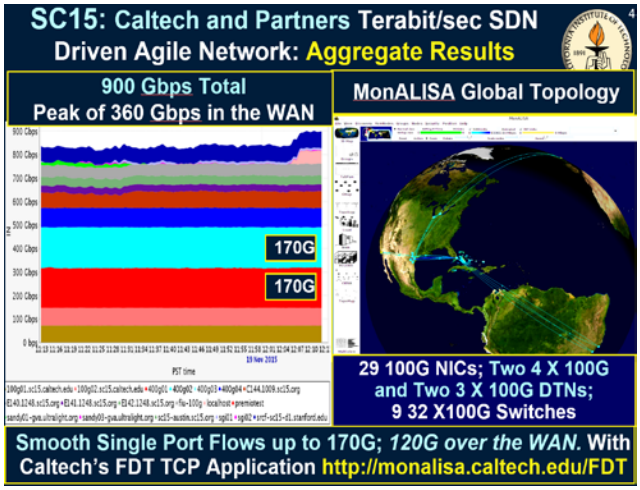

Fig. 2. Examples of record throughputs and network monitoring at the SC15 conference using Caltech's MonALISA monitoring and FDT. The results achieved are appropriate for supporting sub-petabyte transactions in the pre-exascale era, as part of the operational model now being developed with the ALCF.

Following SC15, the team developed the use of Open vSwitch to manage "orchestration" of operations among the end sites and the network, under software control. Open Virtual Switch (OVS), which can make end-hosts appear as a switch and thus an integral part of the network, is designed to enable network automation via programmatic interfaces, especially for the virtualized environments. It is available as part of the major standard Linux distributions, and supports standard, well-established protocols for internal management, security, monitoring, and automated control for traffic flows. Following developments using the latest OVS traffic control module versions by the team, OVS was shown to perform extremely well, with the ability to stably limit traffic rates at any level up to wire speed, with very low CPU overhead, as illustrated in Fig. 3.

In achieving smooth high rate flows, especially flows at the rates shown in the figure, it is important to note that transfers do not involve files per se, but rather data buffers sent and received by Caltech's FDT [8] application. FDT is able to decompose any structure of multiple files into buffers whose dimension is adapted to the I/O capability of the end-systems involved in a transfer, and send the buffers to the network at a rate compatible with the real-time capacity of the network path which is monitored (using MonALISA) in real-time, resulting in smooth "impedance matched" flows. The file structure is subsequently restored to its original form at the destination.

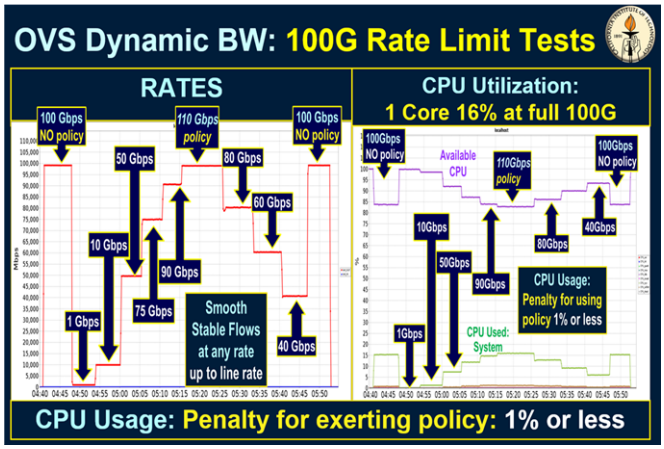

Fig. 3. Tests of Open vSwitch to dynamically control bandwidth between servers with 100G interfaces. The ability to smoothly control the flows using OVS and FDT at any level up to wire speed is shown on the left, and the very small added CPU load resulting from exerting a rate limit (1% or less) is shown on the right.

VII. Orchestration Among Sites with Multiple Host Groups, Paths and Policies

The design concept of NGENIA-ES's end-to-end orchestration of data flows involving multiple host groups at many sites, multiple diverse network paths among them, and diverse policies governing the path setup and prioritization of flows, is illustrated in Fig. 4. Diverse network paths are constructed to support sets of flows, each of which may be assigned bandwidth individually or in groups in response to (1) requests from applications (PhEDEx and ASO are shown as examples), (2) shell script commands, (3) other upstream SDN controllers.

Adjustment of the allocations can be triggered by (1) new requests, (2) real-time feedback based on the progress of transfers, (3) network state changes or error conditions, (4) proactive load-balancing operations, or (5) orchestration operations imposed by controllers or emerging network operating systems (the example of ESnet's SENOS [19] is

shown) that manage operations in the wide area network core.

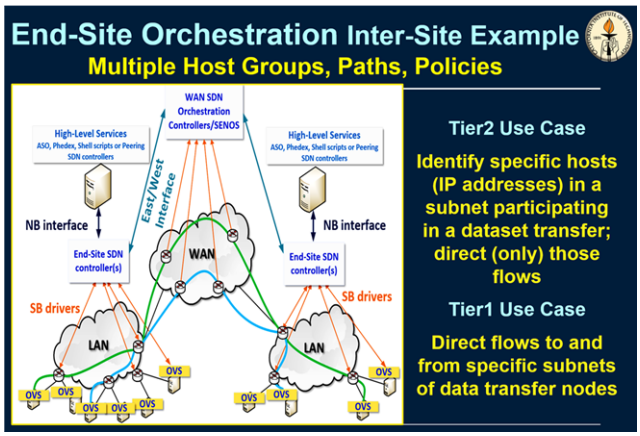

Fig. 4. Illustration of the construction of network paths and management of data flows among multiple clusters at multiple sites using a set of OVS instances, SDN switches and controllers.

Pervasive monitoring and tracking of operations supporting the orchestration functions will be provided by Caltech's MonALISA [20] monitoring system. As SDN frameworks such as OpenDaylight mature, it is expected that some of the monitoring functions may be migrated to standard SDN services.

End-hosts involved in each transfer will be automatically discovered using the SDN controller infrastructure shown as a distributed lookup service. As shown in the figure, the nature of the end hosts varies by use case. In the case of transfers to or from a Tier1 site such as Fermilab or a Leadership site such as the ALCF, transfers will involve data transfer nodes (DTNs) on specified subnets at the site edge. In the case of some Tier2 sites, the transfers may involve a set of hosts scattered throughout one or more subnets. In this case OVS is able to determine the source and destination IP addresses involved in a given transfer and direct only those flows to the path allocated for that transfer.

VIII. Next Generation SDN Developments: Consistent Operations with ALTO and OVS

We have recently begun work on developing a next generation "consistent operations" SDN paradigm, to support the large set of HEP and other data flows without impeding other network traffic (from other science programs or general purpose traffic), following the NGENIA-ES architecture described in the previous section.

In this approach two key components are involved, OVS to stably rate limit the flows at the edges and Application Layer Traffic Optimization (ALTO) [21] in OpenDaylight [13] for end to end optimal path creation, coupled to flow metering and high watermarks set in the network core. We have recently teamed up with Richard Yang's computer science group at Yale University [21] for the integration of ALTO modules with CMS PhEDEx and ASO applications. This methodology will query the replica tuples for a given set of datasets. Once the end nodes with datasets are identified then the MonALISA scheduler using ALTO modules will create multi-domain end to end transfer paths, assign each transfer to a given path, and allocate defined bandwidth levels to each transfer.

The allocations will subsequently be adjusted using OpenFlow flow-metering functions in response to requests from user applications or controllers, or by the requirements of the NGENIA-ES system itself, as described in the previous section. The flow-metering in the network core will be fed back to the OVS instances at the edges, and changes will be applied at a rate consistent with the smooth progress of end-to-end flows.

The near term concept for these developments, and trials are being implemented on a multi-domain testbed at Caltech, Starlight, Michigan and several sites in the Pacific Research Platform (PRP) and partners at CERN and Sao Paulo. In the near term, MonALISA services will be used to assist with the inter-controller (East-West) communications and interactions, until suitable East-West interactions are developed in future within the OpenDaylight framework itself. The high throughput transfers are accomplished with Caltech's FDT.

This will involve monitoring the occupancy of all parts of the infrastructure, steering flows so as to keep the occupancy of each sector of the infrastructure below some "high water mark" set in cooperation with the major network providers, and then reducing the rate of flows or the rate of handling service requests if and where needed.

The developments described above are being carried out on a testbed based at Caltech, with extensions to the other sites mentioned, involving 11 network switches with many 10G, 40G and 100G ports and 16 servers, several of which are capable of full 40G or 100G throughput. Most of the switches and servers shown have been obtained through the NSF DYNES [22-23], ANSE [24] and CHOPIN [25] projects, or through donations by the manufacturers.

An example of the automatically discovered real-time topology of the SDN testbed, using the Caltech OpenDaylight controller, showing the ports and interconnections and some of the flows installed using OpenFlow, is shown in Fig. 5.

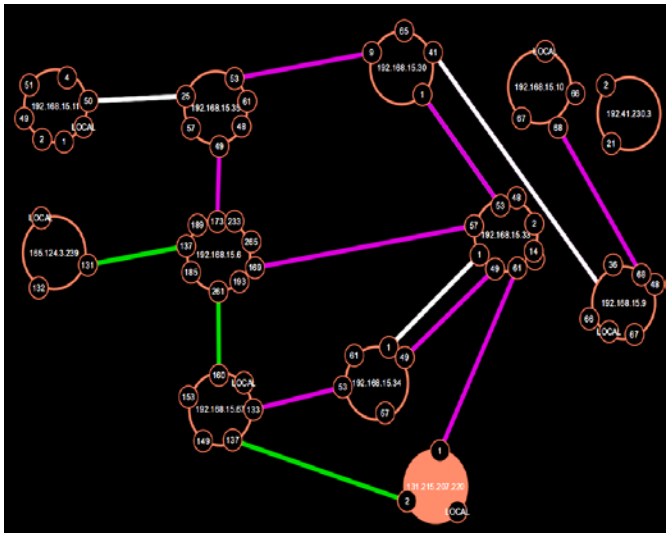

Fig. 5. An example of the auto-discovered topology of the Caltech SDN development testbed.

This work is expected to lead to new higher efficiency modes of stable operations across campuses and national and international network infrastructures, where major science programs have access to higher throughput, more predictable transfers, resulting in more efficient workflow. In this new paradigm, the major research and education network providers can be assured of the continuing compatibility of the services serving the science programs with other network operations, in spite of the programs' exponentially growing needs.

A large scale demonstration of this "consistent operations" paradigm is planned for SC16 this fall.

IX. Scheduling and Optimization: Max-Min Fair Resource Allocation Algorithms

The Yale team working with Caltech has begun to develop an iterative "Max-Min Fair Resource Allocation" (MRFA) scheduling solution to the resource allocation problem (RAP) whose aim is to minimize the maximal time to complete a transfer subject to a set of constraints, applied to the NGENIA-ES case. The Yale team has extensive research and experience with techniques for optimizing inter-datacenter transfers across a complex wide area network topology [26-30] and with practical solutions and extensions of the well-studied "Max-Min Fair Sharing" (MMF) problem [31] [32] [33], as well as frontline SDN developments in the OpenDaylight framework.

The constraints including the priority of each class of flows, expressed in terms of upper and lower limits on the allocated bandwidth between the source and destination for each transfer, and the capacity (or maximum sustainable aggregate throughput in practice) of each link in the network.

Additional objectives of the dataset transfer scheduler are to (1) fully utilize the network resources, i.e., the bandwidth, so that all dataset transfers can be completed in a timely manner; (2) allocate network resources fairly so that no transfer request suffers starvation, leading to an excessively long time to completion of the transfer, and (3) load balance among the sites and the network paths crossing a complex network topology so that no site and no network link is oversubscribed.

In the NGENIA-ES use cases, which are relatively complex, the algorithms for Max-Min fair sharing (MMF), the weighting among the factors used to define fairness, the metrics used to evaluate system performance, and the parameters that govern convergence and stability of the solutions in each time interval, all will require study and experimental test over the long term, as NGENIA-ES moves from demonstrations to pre-production deployments towards widespread production use. Once pilot solutions with stable first-round algorithms and performance metrics are developed at each stage, unsupervised learning to pick out the most influential variables followed by supervised learning to optimize system performance may be applied, to accelerate the development cycles.

While the initial use case will concern "consistent" high levels of network usage among representative Leadership (ALCF), HEP Tier1 (Fermilab) and HEP Tier2 (Caltech) sites along with other testbed and partner sites, the methodology including the algorithms, performance metrics and code base will be generalized and applied in the near term to real-world optimization of the workflows for the CMS [34] and ATLAS [35] experiments at the LHC. Longer term developments will be oriented towards optimization of operations across U.S. and international research and education networks, and among many sites including storage and computation resources in the scheduling, and other use cases with real-time delivery needs such as the Large Synoptic Space Telescope (LSST) [2] that will be based in Chile.

The basic idea of the MFRA algorithm is to iteratively maximize the volume of data that can be transferred subject to the constraints. The algorithm works in quantized time intervals such that it schedules network paths and data volumes to be transferred in each time slot. MFRA first marks all transfers as "unsaturated". Then it solves a linear programming model to find the common minimum transfer satisfaction rate (i.e., the ratio of transferred data volume in a time interval over the whole data volume of this request) that is satisfied by all transfer requests. With this common rate found, MFRA then randomly selects an unsaturated request in each iteration, increases its transfer rate as much as possible by finding residual paths available in the network, or by increasing the allocated bandwidth along an existing path, until it reaches its upper limit or can otherwise not be increased further, so it is "saturated".

At each iteration, newly saturated requests are removed from the subsequent process by fixing their corresponding rate value, and completed transfers are removed from further consideration. After all the data transfer rates are saturated in the given time slot, then a feasible set of data transfer volumes scheduled to be transferred in the slot across each link in the network can be derived. In this way, MFRA achieves a balance between fairness of data transfers and network utilization, with an iterative algorithm of relatively low computational complexity.

Note that since the allocations assigned to flows involve multi-segment network paths as well as bandwidth on each link, the problem can become difficult. However, MFRA can certainly handle the case where particular routing constraints are specified, e.g., where all routes are fixed ahead of time, or where each transfer request only uses one single path in each time slot. In this case, the problem can be well-handled by adding an additional set of linear constraints in the formulation of the MMF problem.

Once these developments are underway, we will also consider MMF extensions that scale well to large distributed network cases, such as the Upward Max-Min Fairness method which can be computed in a simple distributed manner, and the Iterative Exhaustive Waterfill (IEWF) algorithm developed by a Google research team [32] that extends the classical Waterfill method to the case of multiple available paths per transfer.

X. Engagement and Partnerships

The progress and achievements of the project pilots mentioned above have been made possible through our strong collaboration and partnership with colleagues and groups both at research institutions and private companies.

Some of our most important partners for this project are listed below.

Academic Partners: Pacific Research Platform (PRP) (includes UCSD, Stanford, UCLA, UC Berkeley and many other campuses throughout California), Fermi National Accelerator Laboratory, Brookhaven National Lab, Lawrence Berkeley National Lab, CERN, Florida International University, SPRACE and GridUNESP (Sao Paulo), University of Michigan, Florida International University, Vanderbilt University, KITT (Karlsruhe), CNAF (Bologna), KISTI (Korea), Yale University.

Research and Education Networks: ESNet, Internet2, CENIC (California), FLR (Florida), AmLight (Miami), SURFNet (Amsterdam), MiLR (Michigan), BCNET (British Columbia), Academic Network of São Paulo (ANSP), RNP (Brazil), KREONET (Korea).

Industry: Intel, Mellanox, Brocade Networks, Extreme Networks, EchoStreams, Padtec, Dell, Mangstor, QLogic, SGI, Spirent, Cisco, Juniper, Alcatel, Supermicro, Seagate, 2CRSI, Orange Labs, Arista, Inventec.

A key factor in the initial successes cited has been the support and engagement of the DOE Offices of Advanced Scientific Computing and High Energy Physics, and the NSF Directorate for Computer & Information Science and Engineering (CISE).

XI. Summary And Conclusions

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;

• 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 world's 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.

As reviewed in this paper, we made rapid progress, and launched our teams on a promising path towards these goals which, once accomplished, could have a generational impact on data intensive research and education.

Acknowledgment

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] L. Rossi and O. Brüning, "High Luminosity Large Hadron Collider A description for the European Strategy Preparatory Group," 1 Aug. 2012.
[2] "Large Synoptic Survey Telescope," http://www.lsst.org/lsst/.
[3] "SKA Project," https://www.skatelescope.org/project/.
[4] "Science DMZ," https://fasterdata.es.net/science-dmz.
[5] C. Guok, "A user driven dynamic circuit network implementation," Lawrence Berkeley National Laboratory; https://www.es.net/engineering-services/oscars/, 2009.
[6] G. Roberts et al., "Nsi connection service v2. 0," Open Grid Forum, GWD-RP, NSI-WG; https://www.ogf.org/ogf/doku.php/standards/standards, 2013.
[7] R. Egeland, T. Wildish and S. Metson, "Data transfer infrastructure for CMS data taking," PoS (2008): 033., 2008.
[8] Z. Maxa et al., "Powering physics data transfers with FDT," Journal of Physics: Conference Series 052014; fdt.cern.ch, 2011.
[9] "LHC Open Network Environment," http://lhcone.net.
[10] H. Riahi et al., "AsyncStageOut: Distributed user data management for CMS Analysis," Journal of Physics: Conference Series 2015: 062052., 2015
[11] "Super Computing 2015," sc15.supercomputing.org.
[12] "Project Floodlight," www.projectfloodlight.org/floodlight.
[13] "The OpenDaylight Platform," www.opendaylight.org.
[14] "Open vSwitch project," openvswitch.org.
[15] R. Voicu, "Traffic shaping using OVS," LHCOPN-LHCONE meeting, LBNL Berkeley; https://indico.cern.ch/event/376098/contribution/24/material/slides/1.pdf, June 2015.
[16] "INDIS, Innovating the Network for Data-Intensive Science," November 13, 2016.
[17] "Software Defined Networking (SDN) for Scientific Networking Workshop, Austin, Texas," November 2015.
[18] H. Newman et al., "High speed scientific data transfers using software defined networking," Proceedings of the Second Workshop on Innovating the Network for Data-Intensive Science 15 Nov. 2015, p. 2, 2015.
[19] I. Monga, "SENSE Kick-Off Meeting," SDN for the End-to-end Networked Science at the Exascale (SENSE) Organizational Meeting, February 16, 2016.
[20] I. Legrand, R. Voicu, C. Cirstoiu, C. Grigoras, L. Betev and A. Costan, "Monitoring and control of large systems with MonALISA," ACM Queue 7, 6;
[21] R. Yang et al., "Application-Layer Traffic Optimization (alto)," https://datatracker.ietf.org/wg/alto.
[22] "Development of Dynamic Network System (DYNES)," http://nsf.gov/awardsearch/showAward?AWD_ID=0958998.
[23] J. Zurawski, R. Ball, A. Barczyk, M. Binkley, J. Boote, E. Boyd, A. Brown, R. Brown, T. Lehman, S. McKee, B. Meekhof, A. Mughal, H. Newman, S. Rozsa, P. Sheldon, A. Tackett, R. Voicu, S. Wolff and X. Yang, "The DYNES Instrument: A Description and Overview," J. Phys.: Conf. Ser. 396 042065 doi:10.1088/1742-6596/396/4/042065, 2012.
[24] "ANSE (Advanced Network Services for Experiments)," NSF CC-NIE grant OCI-1246133: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1246133.
[25] "The Caltech High-performance Optical Integrated network, NSF CC-IIE grant OCI-1341024.," http://www.internet2.edu/presentations/ccnie201404/20140430-Newman-ProblemsEncountered-SolutionsDiscovered.pdf .

[26] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," Synthesis Lectures on Communication Networks 3, no. 1, pp. 1-211, 2010.
[27] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," IEEE Journal on Selected Areas in Communications, Volume 24, Issue. 8, pp. 1439-1451, 2006.
[28] E. L. Lawler, J. K. Lenstra, A. Kan, H. Rinnooy and D. B. Shmoys, "Sequencing and scheduling: Algorithms and complexity," Handbooks in Operations Research and Management Science 4, pp. 445-522, 1993.
[29] R. K. Ahuja, T. L. Magnanti and J. B. Orlin, "Network Flows: Theory, Algorithms, and Applications," ISBN-13: 978-0136175490.
[30] D. Bertsekas and R. Gallager, "Data Networks," Prentice-Hall, Englewood Cliffs, 2001.
[31] X. Lu, F. Kong, X. Liu, J. Yin and Q. Xiang, "Bulk savings for bulk transfers: Minimizing energy cost of inter-data-center traffic," Technical Report, 2015.
[32] E. Danna, A. Hassidim, H. Kaplan, A. Kumar, Y. Mansour, D. Raz and M. Segalov, "Upward max min fairness," IEEE INFOCOM Proceedings IEEE, pp. 837-845., 2012.
[33] M. Allalouf and Y. Shavitt, "Centralized and distributed algorithms for routing and weighted max-min fair bandwidth allocation," Networking, IEEE/ACM Transactions on 16, no. 5, pp. 1015-1024, 2008.
[34] CMS Experiment, cms.web.cern.ch
[35] ATLAS Experiment, atlas.web.cern.ch

Harvey B. Newman (newman@hep.caltech.edu), BS'68, PhD'74 in physics (MIT), became a member of IEEE in 2002 and is a member of the Communications and Computer Societies as well as a Fellow of the American Physical Society. He was co-spokesman of the Mark J collaboration that discovered the Gluon (the carrier of the strong force) in 1979 and is a member of the CMS Collaboration that discovered the Higgs Boson in 2012. He was US Collaboration Board Chair during 1998-2008 and is currently Chair of the US LHC Users Association. He has been Professor of Physics at the California Institute of Technology in Pasadena since 1982, and has chaired Caltech's exchange and study abroad programs since 2001.

He originated the use of international networks in high energy physics in 1982, led the MONARC project that defined the worldwide grid-based Computing Model of the LHC experiments in 1998-2000, and has chaired the ICFA Standing Committee on Inter-Regional Connectivity since 2002. Newman has had a leading role in originating, developing and operating state of the art international networks and collaborative systems serving the high energy and nuclear physics communities for the last 35 years. He served on the IETF and the Technical Advisory Group that led to the NSFNet in 1985-6. He originated the worldwide LHC Computing Model in 1996, and the LHC Open Network Environment (LHCONE) in 2010. He led US-CERN network operations and development as head of the US LHCNet project between 1995 and 2015. He currently represents the physics community on the Internet2 Network Policy and Operations Advisory Group and the Open Daylight Advisory Board.

Prof. Newman is developing the next generation of software-defined global networks together with ESnet, Internet2, CENIC, Starlight, SURFnet, and many other leading network partners, as well as Fermilab and the Pacific Research Platform. Since 2015 he and his Caltech team and partners have been developing the architecture and methodology for the use of exascale computing facilities for high energy physics and other data intensive science areas. As Chair of the ICFA Standing Committee on Inter-Regional Connectivity since 2002, he has worked to foster greater equality among scientists through the development and deployment of modern network and computing grid infrastructures in many countries including Brazil, Mexico, Pakistan, India, Romania, Slovakia and China. He was awarded Doctor Honoris Causa degrees by the Politechnica University in Bucharest, Romania, and the Pavel Jozef Safarik University in Kosice, Slovakia in 2007, and received the "José Bonifacio" medal of the State University of Rio de Janeiro in Brazil in 2009.