Orchestrating Science DMZs for Big Data Acceleration: Challenges and Approaches

Saptarshi Debroy, Prasad Calyam, Matthew Dickinson
Department of Computer Science, University of Missouri-Columbia
Email: {debroysa, calyamp, dickinsonmg}@missouri.edu.

I. Introduction

A. What is Science Big Data?

In recent years, most scientific research in both academia and industry has become increasingly data-driven. According to market estimates, spending related to supporting scientific data-intensive research is expected to increase to $5.8 billion by 2018 [1]. Particularly for data-intensive scientific fields such as bioscience, or particle physics within academic environments, data storage/processing facilities, expert collaborators and specialized computing resources do not always reside within campus boundaries. With the growing trend of large collaborative partnerships involving researchers, expensive scientific instruments and high performance computing centers, experiments and simulations produce peta-bytes of data viz., Big Data, that is likely to be shared and analyzed by scientists in multi-disciplinary areas [2]. With the United States of America (USA) government initiating a multi-million dollar research agenda on Big Data topics including networking [3], funding agencies such as *National Science Foundation*, *Department of Energy*, and *Defense Advanced Research Projects Agency* are encouraging and supporting cross-campus Big Data research collaborations globally.

B. Networking for Science Big Data Movement

To meet the data movement and processing needs, there is a growing trend amongst researchers within Big Data fields to frequently access remote specialized resources and communicate with collaborators using high-speed overlay networks. These networks use shared underlaying components, but allow end-to-end circuit provisioning with bandwidth reservations [4]. Furthermore, in cases where researchers have sporadic/bursty resource demands on short-to-medium timescales, they are looking to federate local resources with 'on-demand' remote resources to form 'hybrid clouds', versus just relying on expensive over-provisioning of local resources [5]. Figure 1 demonstrates one such example where science Big Data from a Genomics lab requires to be moved to remote locations depending on the data generation, analysis, or sharing requirements.

Thus, to support science Big Data movement to external sites, there is a need for simple, yet scalable end-to-end network architectures and implementations that enable applications to use the wide-area networks most efficiently; and possibly control intermediate network resources to meet Quality of Service (QoS) demands [6]. Moreover, it is imperative to get around the 'frictions' in the enterprise edge-networks i.e., the bottlenecks introduced by traditional campus firewalls with complex rule-set processing and heavy manual intervention that degrade the flow performance of data-intensive applications [7]. Consequently, it is becoming evident that such researchers' use cases with large data movement demands need to be served by transforming system and network resource provisioning practices on campuses.

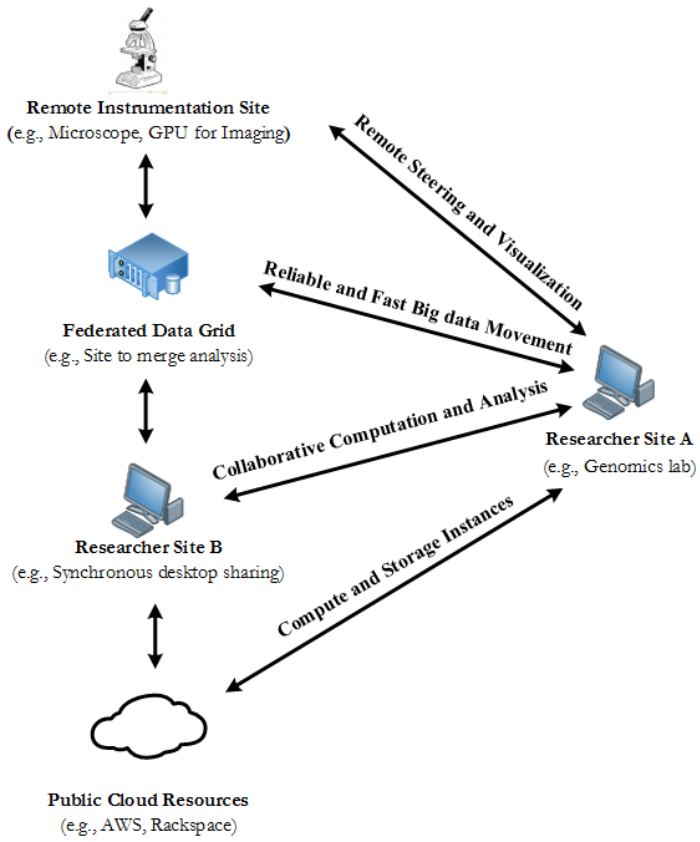


Fig. 1. Example showing need for science Big Data generation and data movement

C. De-Militarized Zones for Science Big Data

The obvious approach to support the special data movement demands of researchers is to build parallel cyberinfrastructures to the enterprise network infrastructures. These parallel infrastructures could allow by-passing of campus firewalls and support 'friction-free' data-intensive flow acceleration over wide-area network paths to remote sites at 1-10 Gbps speeds for seamless federation of local and remote resources [8] [9]. This practice is popularly referred to as building Science DMZs [10] (De-militarized Zones) with network designs that can provide high-speed (1 Gbps - upto 100 Gbps) programmable networks with dedicated network infrastructures for research traffic flows and allow use of high-throughput data transfer protocols. They do not necessarily use traditional TCP/IP protocols with congestion control on end-to-end reserved bandwidth paths, and have deep instrumentation and measurement to monitor performance of applications and infrastructure. The functionalities of Science DMZ as defined in [4] include:

- *A scalable, extensible network infrastructure free from packet loss that causes poor TCP performance;*
- *Appropriate usage policies so that high-performance applications are not hampered by unnecessary constraints;*
- *An effective "on-ramp" for local resources to access wide area network services; and*
- *Mechanisms for testing and measuring, thereby ensuring consistent performance.*

Following the above definition, the realization of a Science DMZ involves transformation of legacy campus infrastructure with increased end-to-end high-speed connectivity (i.e., availability of 10/40/100Gbps end-to-end paths) [11], [12], and emerging computer/network virtualization management technologies [13], [14] for "Big Data flow acceleration" over wide-area networks. The examples of virtualization management technologies include: (i) software-defined networking (SDN) [15], [16], [17] based on programmable OpenFlow switches [18], (ii) RDMA over Converged Ethernet (RoCE) implemented between zero-copy data transfer nodes [19], [20], (iii) multi-domain network performance monitoring using perfSONAR [21] active measurement points, and (iv) federated identity/access management (IAM) using Shibboleth-based entitlements [22].

Although Science DMZ infrastructures can be tuned to provide the desired flow acceleration and can be optimized for QoS factors relating to Big Data application "performance", the policy handling of research traffic can cause a major bottleneck at the campus edge-router. This can particularly impact the performance across applications, if multiple applications simultaneously access hybrid cloud resources and compete for the exclusive and limited Science DMZ resources. Experimental evidence in works such as [9] show considerable disparity between theoretical and achievable goodput of Big Data transfer between remote domains of a networked federation due to policy and other protocol issues. Therefore, there is a need to provide fine-grained dynamic control of Science DMZ network resources i.e., "personalization" leveraging awareness of research application flows, while also efficiently virtualizing the infrastructure for handling multiple diverse application traffic flows.

QoS-aware automated network convergence schemes have been proposed for purely cloud computing contexts [23], however there is a dearth of works that address the "personalization" of hybrid cloud computing architectures involving Science DMZs. More specifically, there is a need to explore the concepts related to application-driven overlay networking with novel cloud services such as 'Network-as-a-Service' to intelligently provision on-demand network resources for Big Data application performance acceleration using the Science DMZ approach. Early works such as our work on Application-Driven Overlay Network-as-a-Service (ADON) [24] seek to develop such cloud services by performing a direct binding of applications to infrastructure and providing fine-grained automated QoS control. The challenge is to solve the multi-tenancy network virtualization problems at campus-edge networks (e.g., through use of dynamic queue policy management), while making network programmability related issues a non-factor for data-intensive application users, who are typically not experts in networking.

D. Chapter Organization

This book chapter seeks to introduce concepts related to Science DMZs used for acceleration of Science Big Data flows over wide-area networks. The chapter will first discuss the nature of science Big Data applications, and then identify the limitations of traditional campus networking infrastructures. Following this, we present the technologies and transformations needed for infrastructures to allow dynamic orchestration of programmable network resources, as well as for enabling performance visibility and policy configuration in Science DMZs. Next, we present two examples of actual Science DMZ implementation use cases with one incremental Science DMZ setup, and another dual-ended Science DMZ federation. Lastly, we discuss the open problems and salient features for *personalization* of hybrid cloud computing architectures in an on-demand and federated manner. We remark that the contents of this chapter build upon the insights gathered

through the theoretical and experimental research on application-driven network infrastructure personalization at the Virtualization, Multimedia and Networking (VIMAN) Lab in University of Missouri-Columbia (MU).

II. Science Big Data Application Challenges

A. Nature of Science Big Data Applications

Humankind is generating data at an exponential rate; it is predicted that by 2020, over 40 zettabytes of data will be created, replicated, and consumed by the humankind [25]. It is a common misconception to characterize any data generated at a large-scale as Big Data. Formally, the four essential attributes of Big Data are: *Volume* i.e., size of the generated data, *Variety* i.e., different forms of the data, *Velocity* i.e., the speed of data generation, and finally *Veracity* i.e., uncertainty of data. Another perspective of Big Data from networking perspective is - any aggregate "data-in-motion" that forces us to look *beyond* traditional infrastructure technologies (e.g., desktop computing storage, IP networking) and analysis methods (e.g., correlation analysis or multi-variate analysis) that are state-of-the-art at a given point in time. From industry perspective, Big Data relates to the generation, analysis, and processing of user-related information to develop better and more profitable services in e.g., Facebook social networking, Google *Flu trends* prediction, United Parcel Service (UPS) route delivery optimization.

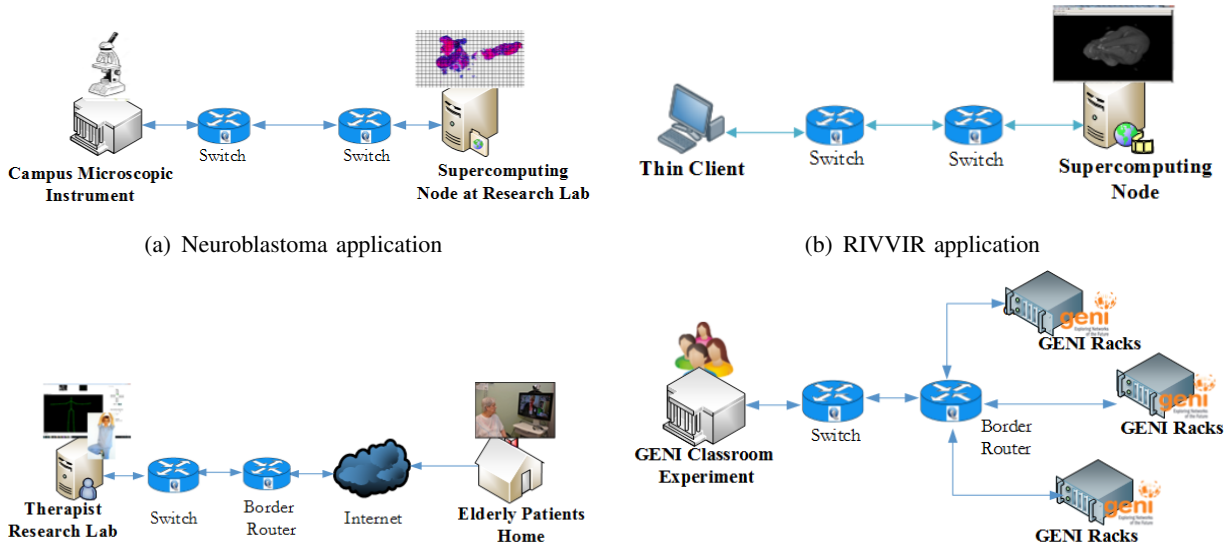
Although the industry has taken the lead in defining and tackling the challenges of handling Big Data, there are many similar and a few different definitions and challenges in important scientific disciplines such as biological sciences, geological sciences, astrophysics, and particle mechanics that have been dealing with Big Data related issues for a while. For example, genomics researchers use Big Data analysis techniques such as MapReduce and Hadoop [33] used in industry for web search. Their data transfer application flows involve several thousands of small files with periodic bursts rather than large single-file datasets. This leads to large amounts of small, random I/O traffic which makes it impossible for a typical campus access network to guarantee end-to-end expected performance. In the following, we discuss two exemplar cases of cutting-edge scientific research that is producing Big Data with unique characteristics at remote instrument sites with data movement scenarios that go much beyond simple file transfers:

1) High Energy Physics: High energy physics or particle mechanics is a scientific field which involves generation and processing of Big Data in its quest to find for e.g., the "God Particle" that has been widely publicized in the popular press recently. Europe's Organization for Nuclear and Particle Research (CERN) houses a *Large Hadron Collider* (LHC) [26], [27], the world's largest and highest-energy particle accelerator. The LHC experiments constitute about 150 million sensors delivering data at the rate of 40 million times per second. There are nearly 600 million collisions per second and after filtering and refraining from recording more than 99.999% of these streams, there are 100 collisions of interest per second. As a result, only working with less than 0.001% of the sensor stream data, the data flow from just four major LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication, which gets fed to university campuses and research labs across the world for access by researchers, educators and students.

2) Biological Sciences and Genomics: Biological Sciences have been one of the highest generators of large data sets for several years, specifically due to the overloads of omics information viz., genomes, transcriptomes, epigenomes and other omics data from cells, tissues and organisms. While the first human genome was a $3 billion dollar project requiring over a decade

to complete in 2002, scientists are now able to sequence and analyze an entire genome in a few hours for less than a thousand dollars. A fully sequenced human genome is in the range of 100 - 1,000 gigabyte of data, and a million customers' data can add up to an exabyte of data which needs to be widely accessed by university hospitals and clinical labs.

In addition to the consumption, analysis and sharing of such major instruments generated science Big Data at campus sites of universities and research labs, there are other cases that need on-demand or real-time data movement between a local site to advanced instrument sites or remote collaborator sites. Below we discuss the nature of four other data-intensive science application workflows being studied at MU's VIMAN Lab from diverse scientific fields that highlight the campus user's perspective in both research and education.



(a) Neuroblastoma application

(b) RIVVIR application

(c) ElderCare-as-a-Service application

(d) GENI classroom experiments application

Fig. 2. Science Big Data movement for different application use cases

3) Neuroblastoma Data Cutter Application: The Neuroblastoma application [9] workflow as shown in Figure 2(a) consists of a high-resolution microscopic instrument on a local campus site generating data-intensive images that need to be processed in real-time to identify and diagnose Neuroblastoma (a type of cancer) infected cells. The processing software and high performance resources required for processing these images are highly specialized and typically available remotely at sites with large GPU clusters. Hence, images (each on the order of several gigabytes) from the local campus need to be transferred in real-time to the remote sites for high-resolution analysis and interactive viewing of processed images. For use in medical settings, it is expected that such automated techniques for image processing should have response times on the order of 10 - 20 seconds for each user task in image exploration.

4) Remote Interactive Volume Visualization Application (RIVVIR): As shown in Figure 2(b), the RIVVIR application [28] at a local campus deals with real-time remote volume visualization of large 3D models (on the order of terabyte files) of small animal imaging generated by MRI scanners. This application needs to be accessed simultaneously by multiple researchers for remote

steering and visualization, and thus it is impractical to download such data sets for analysis. Thus, remote users need to rely on thin-clients that access the RIVVIR application over network paths that have high end-to-end available bandwidth, and low packet loss or jitter for optimal user Quality of Experience (QoE).

5) ElderCare-as-a-Service Application: As shown in Figure 2(c), an ElderCare-as-a-Service application [29] consists of an interactive videoconferencing based tele-health session between a therapist at a university hospital and a remotely residing elderly patient. One of the tele-health use cases for wellness purposes involves performing physiotherapy exercises through an interactive coaching interface that not only involves video but also 3D sensor data from Kinect devices at both ends. It has been shown that regular Internet paths are unsuitable for delivery adequate user QoE, and hence this application is being only deployed on-demand for use in homes with 1 Gbps connections (e.g., at homes with Google Fiber in Kansas City, USA). During the physiotherapy session, the QoE for both users is a critical factor especially when transferring skeletal images and depth information from Kinect sensors that are large in volume and velocity (e.g., every session data is on the order of several tends of gigabytes), and for administration of proper exercise forms and their assessment of the elders' gait trends.

6) Classroom Lab Experiments: It is important to note that Big Data related educational activities with concurrent student access also are significant in terms of campus needs that manifest in new sets of challenges. As shown in Figure 2(d), we can consider an example of a class of 30 or more students conducting lab experiments at a university in a Cloud Computing course that requires access to large amount of resources across multiple data centers that host GENI Racks[1] [29]. As part of the lab exercises, several virtual machines need to be reserved and instantiated by students on remotely located GENI Racks. There can be a sudden bursts of application traffic flows at the campus-edge router whose volume, variety and velocity can be significantly high due to simultaneous services access for computing and analysis, especially the evening before the lab assignment submission deadline.

B. Traditional Campus Networking Issues

1) Competing with Enterprise Needs: The above described Big Data use cases constitute a diverse class of emerging applications that are stressing the traditional campus network environments that were originally designed to support enterprise traffic needs such as e-mail, web browsing and video streaming for distance learning. When appropriate campus cyberinfrastructure resources for Big Data applications do not exist, cutting-edge research in important scientific fields is constrained. Either the researchers do not take on studies with real-time data movement needs, or they resort to simplistic methods to move research data by exchanging hard-drives via 'snail mail' between local and remote sites. Obviously, such simplistic methods are unsustainable and have fundamental scalability issues [8], not to mention that they impede the progress of advanced research that is possible with better on-demand data movement cyberinfrastructure capabilities.

On the other hand, using the "general purpose" enterprise network (i.e., Layer-3/IP network) for data-intensive science application flows is often a highly sub-optimal alternative; and as

[1]GENI Racks are Future Internet infrastructure elements developed by academia in co-operation with industry partners such as HP, IBM, Dell and Cisco; they include APIs and hardware that enable discovery, reservation and teardown of distributed federated resources with advanced technologies such as SDN with OpenFlow, compute virtualization, and Federated-IAM.

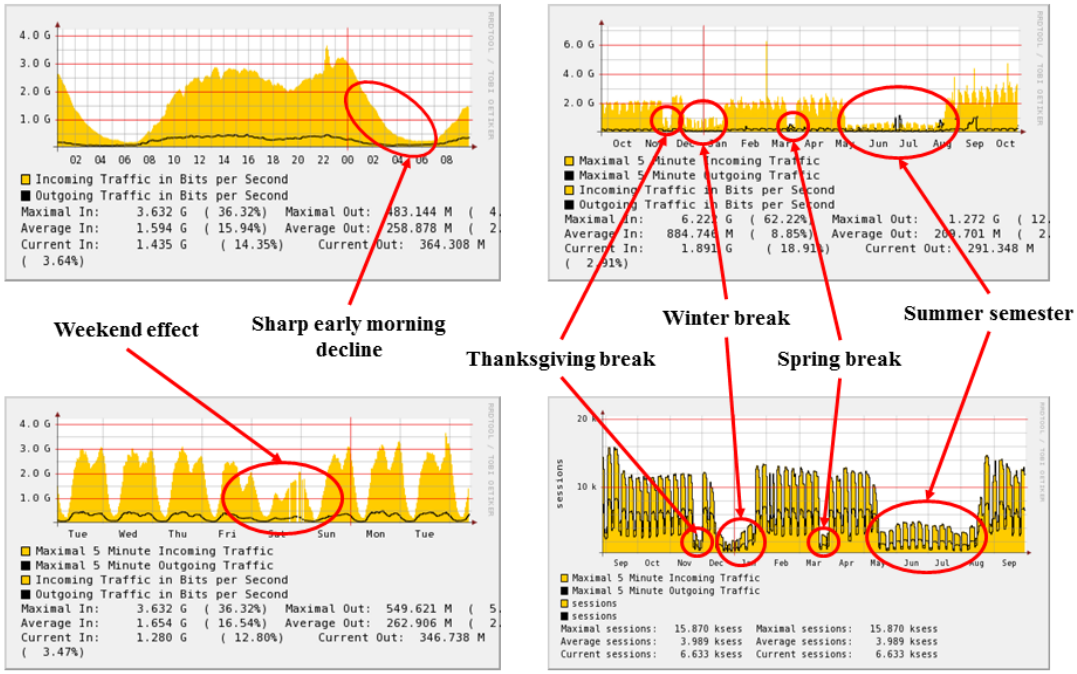


Fig. 3. Campus access network usage trend at MU

described earlier in previous section, they may not at all serve the purpose of some synchronous Big Data applications due to sharing of network bandwidth with enterprise cross-traffic. Figure 3 illustrates the periodic nature of the enterprise traffic with total bandwidth utilization and the session count of wireless access points at MU throughout the year. In Figure 3(a), we show the daily and weekly usage patterns with peak utilization during the day coinciding with most of the on-campus classes with a significant dip during the latter hours of night, and under utilization in the early weekends especially during Friday nights and Saturdays. Figure 3(b) shows seasonal characteristics with peak bandwidth utilization observed during Fall and Spring semesters. Intermediate breaks and Summer semester shows overwhelmingly low usage due to fewer students on campus. For wireless access points session counts shown in the bottom of Figure 3(b), the frequent student movements around the campus leads to a large number of association and authentication processes to wireless access points, and bandwidth availability varies at different times in a day, week or month time-scale. It is obvious that sharing such traditional campus networks with daily and seasonally fluctuating cross-traffic trends causes significant amount of 'friction' for science Big Data movement and can easily lead to performance bottlenecks.

To aggravate the above bottleneck situation, traditional campus networks are optimized for enterprise 'security' and partially sacrifice 'performance' to effectively defend against cyber-

attacks. The security optimization in traditional networks leads to campus firewall policies that block ports needed for various data-intensive collaboration tools (e.g., remote desktop access of a remote collaborator using RDP or VNC [31], GridFTP data movement utility [32]). Federal regulations such as HIPAA in the USA that deal with privacy issues of health-related data also increase the extent to which network access lists are tightly controlled and performance is compromised to favor higher security stances. The blocking of ports in traditional campus networks decreases the risk of malicious access of internal-network data/resources, however it severely limits the ability of researchers to influence campus security policies. Even if ad-hoc static firewall exceptions are applied, they are not scalable to meet special performance demands of multiple Big Data application related researchers. This is because of the 'friction' from hardware limitations of firewalls that arises when handling heavy network-traffic loads of researcher application flows under complex firewall rule-set constraints.

2) Hardware Limitations: In addition to the friction due to firewall hardware limitations, friction also manifests for data-intensive flows due to the use of traditional traffic engineering methods that have: (a) long provisioning cycles and distributed management when dealing with under or over subscribed links, and (b) inability to perform granular classification of flows to enforce researcher-specific policies for bandwidth provisioning. Frequently, the bulk data being transferred externally by researchers is sent on hardware that was purchased a number of years ago, or has been re-purposed for budgetary reasons. This results in situations where the computational complexity to handle researcher traffic due to newer application trends has increased, while the supporting network hardware capability has remained fairly static or even degraded. The overall result is that the workflows involving data processing and analysis pipelines are often 'slow' from the perspective of researchers due to large data transfer queues, to the point that scaling of research investigations is limited by several weeks or even months for purely networking limitations between sites.

In a shared campus environment, hosts generating differing network data-rates in their communications due to application characteristics or network interface card (NIC) capabilities of hosts can lead to resource misconfiguration issues in both the system and network levels and cause other kinds of performance issues [30]. For example, misconfigurations could occur due to internal buffers on switches becoming exhausted due to improper settings, or due to duplex mis-matches and lower rate negotiation frequently experienced with new servers with 1 Gbps NICs communicating with old servers with 100 Mbps; same is true when 10 Gbps NIC hosts communicate with 1 Gbps hosts. In a larger and complex campus environment with shared underlaying infrastructures for enterprise and research traffic, it is not always possible to predict whether a particular pathway has end-to-end port configurations for high network speeds, or if there will be consistent end-to-end data-rates.

It is interesting to note that performance mis-match issues for data transfer rates are not just network related, and could also occur in systems that contain a large array of solid state drives (versus a system that has a handful of traditional spinning hard drives). Frequently, researchers are not fully aware of the capabilities (and limitations) of their hardware, and I/O speed limitations at storage systems could manifest as bottlenecks, even if end-to-end network bandwidth provisioning is performed as 'expected' at high-speeds to meet researcher requirements.

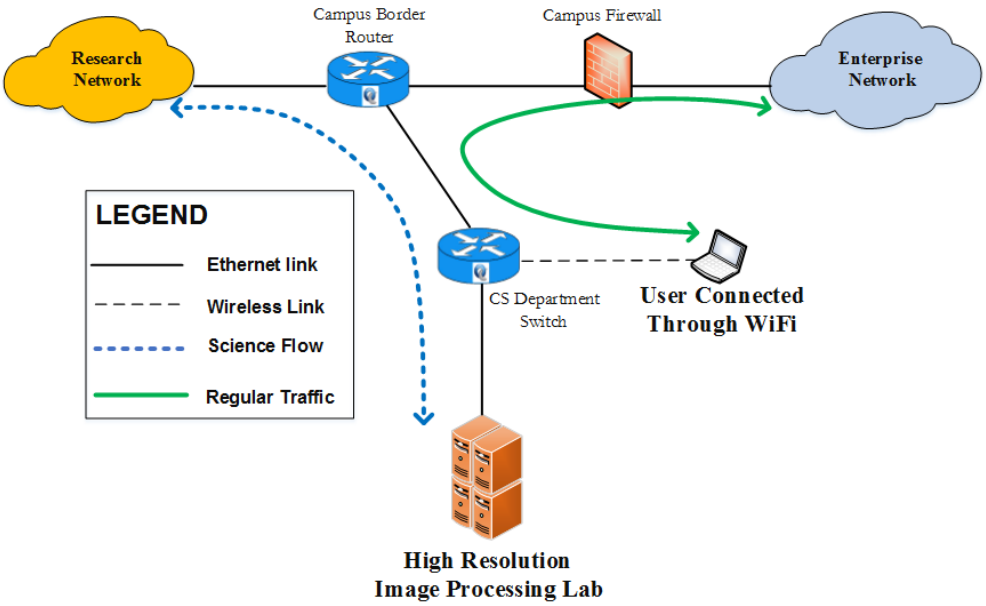


Fig. 4. Transit selection of Science flows and regular traffic within campus

III. Transformation of Campus Infrastructure for Science DMZs

A. An 'On-ramp' to Science DMZ Infrastructure

The inability of traditional campus infrastructures to cater to the real-time or on-demand science Big Data application needs is the primary motivation behind creating a "parallel infrastructure" involving Science DMZs with increased high-speed end-to-end connectivity and advanced technologies described previously in Section I-A. They provide modernized infrastructure and research-friendly firewall policies with minimal or no firewalls in a Science DMZ deployment. In addition, they can be customized per application needs for on-ramp of data-intensive science flows to fast wide-area network backbones (e.g., Internet2 in USA, GEANT in Europe, or APAN in Asia). The parallel infrastructure design thus features abilities such as dynamic identification and orchestration of Big Data application traffic to by-pass the campus' enterprise firewall and use devices that foster flow acceleration, when transit selection is made to leverage the Science DMZ networking infrastructure.

Figure 4 illustrates traffic flow 'transit selection' within a campus access network with Science DMZ capabilities. We can see how intelligence at the campus border and department-level switches enables bypassing of research data flows from campus firewall restricted paths onto research network paths. However, enterprise traffic such as web browsing or e-mails are routed through the same campus access network to the Internet through the firewall policed paths. The research network paths typically involve extended VLAN overlays between local and remote sites, and services such as AWS Direct Connect are used for high-speed layer-2 connections to public clouds. With such overlay paths, Big Data applications can use local/remote and the public cloud resources as if they all reside within the same internal network.

Moreover, research traffic can be isolated from other cross-traffic through loss-free, dedicated 'on-demand' bandwidth provisioning on a shared network underlay infrastructure. It is important

to note that the 'last-mile' problem of getting static or dynamic VLANs connected from the research lab facilities to the Science DMZ edge is one of the harder infrastructure setup issues. In case of Big Data application cases, having 100 Gigabit Ethernet (GE) and 40 - 100 Gbps network devices could be a key requirement. Given that network devices that support 40 - 100 Gbps speeds are expensive, building overlay networks requires significant investments from both the central campus and departmental units. Also, the backbone network providers at the regional (e.g., CENIC) and national-level (e.g., Internet2) need to create a wide footprint of their backbones to support multiple extended VLAN overlays simultaneously between campuses.

Further, the end-to-end infrastructure should ideally feature SDN with OpenFlow switches at strategic traffic aggregation points within the campus and backbone networks. SDN provides centralized control on dynamic science workflows over a distributed network architecture, and thus allows proactive/reactive provisioning and traffic steering of flows in a unified, vendor-independent manner [18]. It also enables fine-grained control of network traffic depending on the QoS requirements of the application workflows. In addition, OpenFlow enabled switches help in dynamic modification of security policies for large flows between trusted sites when helping them dynamically by-pass the campus firewall [16]. Figure 5 shows the infrastructural components of a Science DMZ network within a campus featuring SDN connectivity to different departments. Normal application traffic traverses paths with intermediate campus firewalls, and reaches remote collaborator sites or public cloud sites over enterprise IP network to access common web applications. However, data-intensive science application flows from research labs that are 'accelerated' within Science DMZs by-pass the firewall to the 10 - 100 GE backbones.

B. Handling Policy Specifications

Assuming the relevant infrastructure investments are in place, the next challenge relates to the Federated-IAM that requires specifying and handling fine-grained resource access policies in a multi-institution collaboration setting (i.e., at both the local and remote researcher/instrument campuses, and within the backbone networks) with minimal administrative overhead. Figure 6 illustrates a layered reference architecture for deploying Science DMZs on campuses that need to be securely accessed using policies that are implemented by the Federated-IAM framework. We assume a scenario where two researchers at remote campuses with different subject matter expertise collaborate on an image processing application that requires access to an instrument facility at one researcher's site, and a HPC facility at the other researcher's site.

In order to successfully realize the layered architecture functions in the context of multi-institutional policy specification/handling, there are several questions that need to addressed by the Federated-IAM implementation such as: (i) How can a researcher at the microscope facility be authenticated and authorized to reserve HPC resources at the collaborator researcher campus?; (ii) How can a OpenFlow controller at one campus be authorized to provision flows within a backbone network in an on-demand manner?; and even (iii) How do we restrict who can query the performance measurement data within the extended VLAN overlay network that supports many researchers over time?

Fortunately, standards-based identity management approaches based on Shibboleth entitlements [22] have evolved to accommodate permissions in above user-to-service authentication and authorization use cases. These approaches are being widely adopted in academia and industry enterprises. However, they require a central, as well as an independent 'Service Provider' that hosts an 'Entitlement service' amongst all of the campuses that federate their Science DMZ

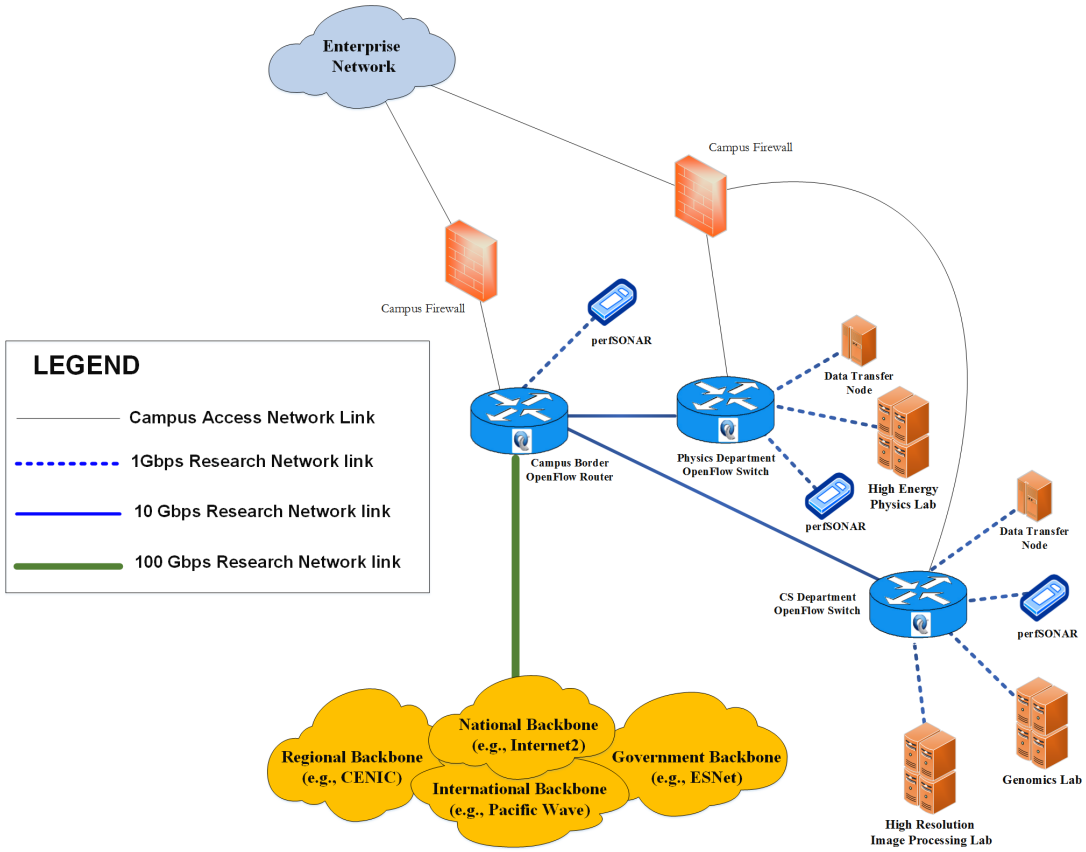


Fig. 5. A generic Science DMZ physical infrastructure diagram

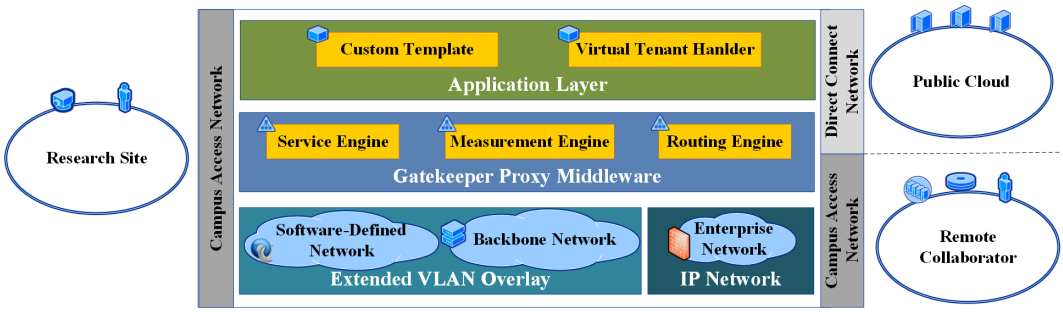


Fig. 6. Campus Science DMZ logical schematic showing architecture layers

infrastructures. Having a registered Service Provider in the Campus Science DMZ federation leads to a scalable and extensible approach, as it eliminates the need to have each campus have bilateral agreements with every other campus. It also allows for centrally managing entitlements based on mutual protection of privacy policies between institutions to authorize access to different infrastructure components such as inter-campus OpenFlow switches.

In order to securely maintain the policy directories of the federation, and to allow institutional policy management of the Science DMZ flows, a ‘gatekeeper-proxy middleware’ as shown in

Figure 6 is required. The gatekeeper-proxy is a critical component of the Science DMZ as it is responsible to integrate and orchestrate functionalities of a Science DMZ's: (a) OpenFlow controller through a 'routing engine' [40], (b) performance visibility through a 'measurement engine', and (c) 'service engine' which allows the functioning of the user-facing web portals that allow a researcher requests access to overlay network resources.

To effectively maintain the gatekeeper-proxy to serve diverse researcher needs concurrently on the shared underlay infrastructure, the role of a "Performance Engineer" technician within a campus Science DMZ is vital. We envisage this role to act as the primary 'keeper' and 'helpdesk' of the Science DMZ equipment, and the success of this role is in the technician's ability to augment traditional System/Network Engineer roles on campuses. In fact, large corporations that typically support data-intensive applications for their users (e.g., disaster data recovery and real-time analytics in financial sector, content delivery network management in consumer sector), have well-defined roles and responsibilities for a Performance Engineer.

Given that researcher data flows in Science DMZs are unique and dynamic, specialized technician skill sets and toolkits are needed. The Performance Engineer needs to effectively function as a liaison to researchers' unique computing and networking needs while coordinating with multi-domain entities at various levels (i.e., building-level, campus-level, backbone-level). He/she also has to cater to each researcher's expectations of high-availability and peak-performance to remote sites without disrupting core campus network traffic. For these purposes, the Performance Engineer can use 'custom templates' that allow repeatable deployment of Big Data application flows, and use virtualization technologies that allow realization of a 'virtual tenant handler' so that Big Data application flows are isolated from each other in terms of performance or security. Moreover, the tools of a Performance Engineer need to help serve the above onerous duties in conjunction with administering maintenance windows with advanced cyberinfrastructure technologies, and their change management processes.

C. Achieving Performance Visibility

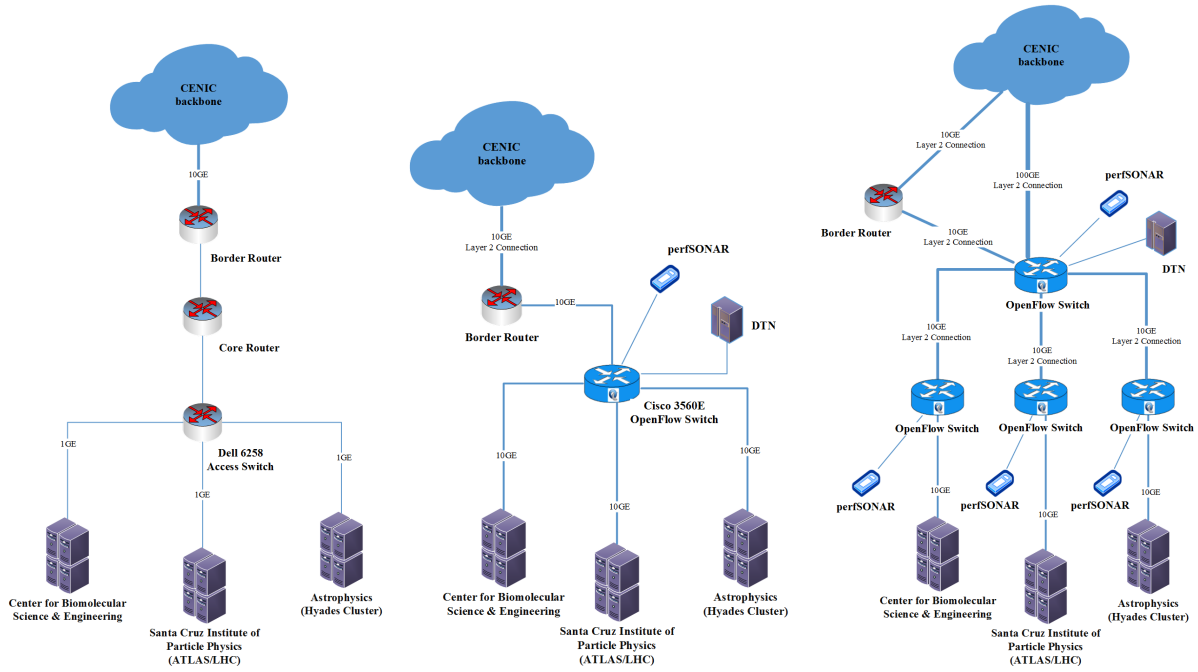
To ensure smooth operation of the fine-grained orchestration of science Big Data flows, Science DMZs require end-to-end network performance monitoring frameworks that can discover and eliminate the "soft failures" in the network. Soft failures cause poor performance unlike "hard failures" such as fiber cuts that prevent data from flowing. Particularly, active measurements using tools such as Ping (for round trip delay), Traceroute (for network topology inference), OWAMP (for one-way delay) and BWCTL (for TCP/UDP throughput) are essential in identifying soft failures such as packet loss due to failing components, mis-configurations such as duplex mismatches that affect data rates, or routers forwarding packets using the management CPU rather than using a high-performance forwarding hardware. These soft failures often go undetected as the legacy campus network management and error reporting systems are optimized for reporting hard failures, such as loss of a link or device.

Currently, perfSONAR [21] is the most widely-deployed framework with over 1200 publicly registered measurement points worldwide for performing multi-domain active measurements. It is being used to create'measurement federations' for collection and sharing of end-to-end performance measurements across multiple geographically separated Science DMZs forming a research consortium [36]. Collected measurements can be queried amongst federation members through interoperable web-service interfaces to mainly analyze network paths to ensure packet loss free paths and identify end-to-end bottlenecks. They can also help in diagnosing performance

bottlenecks using anomaly detection [37], determining the optimal network path [38], or in network weather forecasting [39].

D. Science DMZ implementation use cases

Below we discuss two ideologically dissimilar Science DMZ implementation use cases. First, we present a 3-stage transformation of a campus science infrastructure for handling data-intensive application flows. Next, we shed light on a double-ended Science DMZ implementation that connects two geographically distant campus Science DMZs for Big Data collaboration between the two campuses.



(a) Past Science application flow path (b) Present Science DMZ architecture (c) Future Science DMZ upgrade plan

Fig. 7. An exemplar incremental campus Science DMZ implementation

1) An Incremental Science DMZ Implementation: In Figure 7, we show the stages of the University of California-Santa Cruz (UCSC) campus research network evolution to support data-intensive science applications [35]. Figure 7(a) shows the UCSC campus research network before Science DMZ implementation with a 10 Gbps campus distribution core catering the three main Big Data flow generators, e.g., Santa Cruz Institute of Particle Physics (SCIPP), a 6-Rack HYADES cluster, and Center for Biomolecular Science & Engineering. A traditional (i.e., non-OpenFlow) Dell 6258 access switch was responsible to route research data to campus border router through core routers and ultimately to the regional CENIC (Corporation for Education Network Initiatives in California) backbone. However, buffer size limitations of intermediate switches created bottlenecks in both research and enterprise networks, particularly dedicated 10GE links to research facilities could not support science data transfer rates beyond 1 Gbps. In 2013, UCSC implemented a quick-fix solution to the problem as shown in Figure 7(b), which

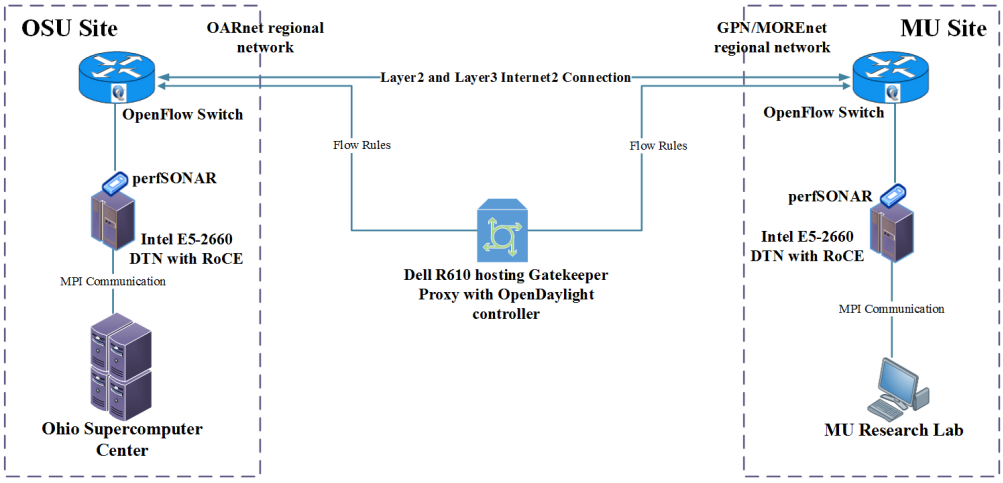


Fig. 8. An exemplar double-ended campus Science DMZ implementation

involved a Cisco 3560E OpenFlow switch connected with perfSONAR nodes and multi-core DTNs. The Science DMZ switch currently at this time of writing, has direct links to all Big Data applications and is connected to the border router with 10GE on both ends. In future, UCSC has plans to install dedicated Science DMZ switches connected through 10GE links with individual data intensive services, and a master switch connected to CENIC backbone through 100GE link, as shown in Figure 7(c). This 3-stage transformation is a common trend of campus network evolution that has been seen in many other cases, and is increasing the footprint of software-defined network elements (OpenFlow-compatible) in support of Big Data applications across campus environments.

2) A Double-ended Science DMZ Transformation Example: Figure 8 shows how both The Ohio State University (OSU) and MU campuses have transformed their Science DMZ infrastructures for inter-campus research collaborations. They are both connected through an extended VLAN overlay that involves an Internet2 Advanced Layer 2 Service (AL2S) connection by way of local regional networks of OARnet in Ohio, and GPN/MoreNet in Missouri, respectively. Each Science DMZ has a matching DTNs equipped with dual Intel E5-2660, 128 GB of memory, 300 GB PCI-Express solid state drive, and dual Mellanox 10 Gbps network cards with RoCE support. Each Science DMZ has perfSONAR measurement points for continuous monitoring at 1 – 10 Gbps network speeds. A common Dell R610 node in the OSU Science DMZ is used to run an OpenFlow controller that controls both the OSU and MU Science DMZ OpenFlow switches. Two HP 3800s are used to attach to the various nodes in the Science DMZ, and a single NEC PF5820 aggregates the two connections at OSU. A NEC switch is connected to OSU's 100 Gbps Cisco Nexus router at 10 Gbps, and has the ability to scale to 40 Gbps as the Science DMZ grows to support future researchers and applications. At MU, the Science DMZ features OpenFlow switches include a Brocade VDX 8770 switch to attach various nodes in the Science DMZ, and a 100 Gbps Brocade MLXE router at 10 Gbps interface speeds, with the ability to scale up to 100 Gbps speeds. This double-ended Science DMZ deployment between OSU and MU has garnered support and fostered new collaborations between a number of researchers on the two campuses, and is being viewed as model infrastructure for 'team science' projects.

IV. Network-as-a-Service within Science DMZs

If multiple applications accessing hybrid cloud resources compete for the exclusive and limited Science DMZ resources, the policy handling of research traffic can cause a major bottleneck at the campus edge router and impact the performance across applications. Thus, there is a need to provide dynamic Quality of Service (QoS) control of Science DMZ network resources versus setting a static rate limit affecting all applications. The dynamic control should have awareness of research application flows with urgent or other high-priority computing needs, while also efficiently virtualizing the infrastructure for handling multiple diverse application traffic flows [34]. The virtualization obviously should not affect the QoS of any of the provisioned applications, and also advanced services should be easy-to-use for data-intensive application users, who should not be worrying about configuring underlying infrastructure resources.

Consequently, there is a need to provide fine-grained dynamic control of Science DMZ network resources i.e., "personalization" leveraging awareness of research application flows, while also efficiently virtualizing the infrastructure for handling multiple diverse application traffic flows. More specifically, there is a need to explore the concepts related to application-driven overlay networking with novel cloud services such as 'Network-as-a-Service' to intelligently provision on-demand network resources for Big Data application performance acceleration using the Science DMZ approach. Early works such as our work on Application-Driven Overlay Network-as-a-Service (ADON) [24] seek to develop such cloud services by performing a direct binding of applications to infrastructure and providing fine-grained automated QoS control. The challenge is to solve the multi-tenancy network virtualization problems at campus-edge networks (e.g., through use of dynamic queue policy management), while making network programmability related issues a non-factor for data-intensive application users, who are typically not experts in networking. The salient features of ADON are as follows:

- ADON intelligently provisions on-demand network resources by performing a direct binding of applications to infrastructure with fine-grained automated QoS control in a Science DMZ.
- In ADON, 'network personalization' is performed using a concept of "custom templates" to catalog and handle unique profiles of application workflows.
- Using the custom templates and VTH concepts, ADON manages the hybrid cloud requirements of multiple applications in a scalable and extensible manner.
- ADON ensures predictable application performance delivery by scheduling transit selection (choosing between Internet or extended VLAN overlays) and traffic engineering (e.g., rate limit queue mapping based on application-driven requirements) at the campus-edge.

Figure 9 shows how through the ADON, data-intensive applications can co-exist on top of a shared wide-area physical infrastructure topology, with each application demanding local/remote network or compute resources with unique end-to-end QoS requirements. We can notice how multiple science Big Data applications such as Neuroblastoma, ECaaS and Classroom Lab with different QoS requirements are orchestrated through overlay networking without compromising the overall end-to-end performance. Such an approach of application-driven orchestration of Science DMZs seeks to lay the foundation for solving even harder issues that may transform the way science Big Data research is carried out through collaborations across communities.

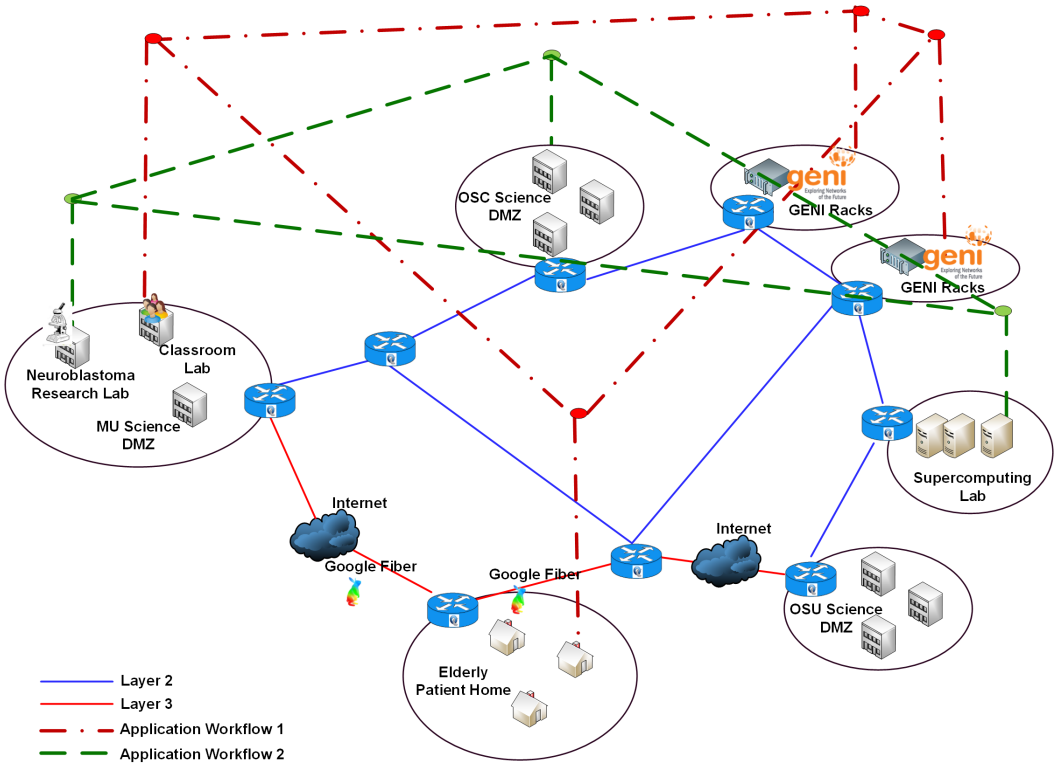


Fig. 9. Multi-tenant application workflows on a shared wide-area physical infrastructure

V. Concluding Remarks

A. What have we learnt?

To summarize, the exponential growth of science Big Data traffic and ensuing accelerated movement requirements are revolutionizing the way campus networks are being designed and operated. The volume and velocity of science Big Data has necessitated a complete paradigm shift from legacy campus network architecture principles and policies, especially in terms of resource provisioning, QoS management, and coupling performance visibility and control. To meet the need to make cross-campus Big Data movement friction free, today's campus networks are making provisions for on-demand proprietary access of high-speed physical resources by Big Data applications through various network virtualization approaches. Such exclusive access in order to isolate Big Data traffic from enterprise traffic is an intellectual evolution from the traditional ways campus access networks were designed a decade ago where applications used to share resources using best-efforts networks.

Moreover, in legacy campus access networks, the service providers used to jointly manage QoS for campus related business and research traffic. However, with adoption of Science DMZ infrastructures by research campuses to accelerate Big Data movement, the QoS initiative is shifting from being service provider governed - to - Big Data researcher and application steered. On-demand application-driven orchestration using intelligent resource allocation and network virtualization with application-imposed QoS guarantees has become the widely-accepted future direction of campus access network design for effective Big Data handling.

Finally, making performance monitoring an integral part of science Big Data networking to better steer data-intensive science flows within-and-outside the campus network, is providing robustness and fault-tolerance capabilities to the campus network in its goal to handle Big Data. Strategically located measurement instrumentation, and application-driven performance metrics have facilitated better visibility and control on Big Data transfer performance between campuses, and are enabling the identification and proactive avoidance of performance bottleneck scenarios and network soft-spots.

B. The road ahead and open problems

Transformation of legacy campus infrastructures with adoption of Science DMZs has facilitated "Big Data Highways" that use cutting-edge technologies, such as software-defined networking, end-to-end performance monitoring, and network-virtualization. Typically, these efforts so far have been mostly incremental where network engineers and designers upgrade the campus access network with faster devices, gigabit fiber cables and intelligent flow control to cater specific data-intensive applications' needs. However, a truly 'Research-defined Network' (RDN), built to control and troubleshoot all the networking aspects of data-intensive science applications is still to be realized by network researchers and service providers. Such RDNs need to support the full life-cycle of campus Big Data, from creation to computation and consumption.

Through RDNs, Big Data researchers will be able to dictate policies, security features, and QoS guarantees specific to their applications. Making data-intensive research a driving force behind campus Big Data network design will enable the network designers to better address open issues, such as: (a) assurance of satisfactory user QoE when simultaneously scheduling multiple science Big Data applications, (b) standardizing performance engineering techniques and protocols for easy use and wide-adoption, and (c) selectively replicating time sensitive or mission critical data across multiple platforms for reliability purposes and prudently selecting replication sites to avoid end-to-end performance bottlenecks.

Building such RDNs is a first step to federate different 'Big Data Highways' to create a 'Big Data Interstate system' where different campus Big Data network infrastructures seamlessly come together. Creating such federations should be aimed towards faster sharing of research data, enhancing cross campus research collaboration, and quicker troubleshooting of network performance bottlenecks. Although early efforts led by Clemson University [36] are taking shape in creating such multi-campus Science DMZ federations, there exists a number of open challenges in realizing such collaborations.

The open challenges can be summarized as follows: (a) co-ordination of federated resources with adherence to policies of multiple-domains, (b) enforcing federated and transparent access control mechanisms over local autonomy to facilitate broader sharing, (c) building secured middlegrounds for performance visibility and network control across Science DMZ domains, and (d) creating social platforms or extending existing platforms for scientific collaborations, such as Science Gateway [43] or HUBzero [44] where Big Data researchers, network designers, and policymakers belonging to the same society can mingle, share data and expertise, collaborate, and create new policies and rules for the federation. Solving these open issues are fundamental in the future explorations that will lead to a *Internet for Big Data* in our society.

VI. SUMMARY

To summarize, the key take-aways from this book chapter are:

- The unique characteristics and data movement requirements of science Big Data pose novel networking challenges.
- These challenges motivate the need for creation of Science DMZs that are parallel infrastructures to enterprise infrastructures to accelerate performance of science Big Data flows.
- Application-driven orchestration of science Big Data flows is essential within the Science DMZ to obtain expected performance and avoid performance bottlenecks.

REFERENCES

[1] "Network modernization is key to leveraging big data", http://www.federaltimes.com/.

[2] H. Yin, Y. Jiang, C. Lin, Y. Luo, Y. Liu, "Big Data: Transforming the Design Philosophy of Future Internet", *IEEE Network Magazine*, Vol.28, 2014.

[3] "Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments", *Office of Science and Technology Policy, The White House*, 2012.

[4] E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, "The Science DMZ: A Network Design Pattern for Data-Intensive Science", *Proc. of IEEE/ACM Supercomputing*, 2013.

[5] A. Das, C. Lumezanu, Y. Zhang, V. Singh, G. Jiang, C. Yu, "Transparent and flexible network management for big data processing in the cloud", *Proc. of USENIX HotCloud*, 2013.

[6] X. Yi, F. Liu, J. Liu, H. Jin, "Building a Network Highway for Big Data: Architecture and Challenges", *IEEE Network Magazine*, Vol.28, 2014.

[7] L. Borovick, R. L. Villars, "The Critical Role of the Network in Big Data Applications", *Cisco White paper*, 2012.

[8] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, F. Lau, "Moving Big Data to The Cloud: An Online Cost-Minimizing Approach", *IEEE Journal on Selected Areas in Communications*, Vol.31, 2013.

[9] P. Calyam, A. Berryman, E. Saule, H. Subramoni, P. Schopis, G. Springer, U. Catalyurek, D. K. Panda, "Wide-area Overlay Networking to Manage Accelerated Science DMZ Flows", *Proc. of IEEE ICNC*, 2014.

[10] "Science DMZ Network Design Model", http://fasterdata.es.net/science-dmz.

[11] A. Rajendran, P. Mhashilkar, Hyunwoo Kim; D. Dykstra, G. Garzoglio, I. Raicu, "Optimizing Large Data Transfers over 100Gbps Wide Area Networks", *Proc. of IEEE/ACM CCGrid*, May 2013.

[12] E. Kissel, M. Swany, B. Tierney, E. Pouyoul, "Efficient wide area data transfer protocols for 100 Gbps networks and beyond", *Proc. of NDM*, 2013.

[13] H. Luo, H. Zhang, M. Zukerman, C. Qiao, "An incrementally deployable network architecture to support both data-centric and host-centric services", *IEEE Network Magazine*, 2014.

[14] Y. Ren, T. Li, D. Yu, S. Jin, T. Robertazzi, B. Tierney, E. Pouyoul, "Protocols for Wide-area Data-intensive Applications: Design and Performance Issues", *Proc. of IEEE/ACM Supercomputing*, 2012.

[15] Y. Cui, S. Xiao, C. Liao, I. Stojmenovic, M. Li, "Data Centers as Software Defined Networks: Traffic Redundancy Elimination with Wireless Cards at Routers", *IEEE Journal on Selected Areas in Communications*, Vol. 31, 2013.

[16] I. Monga, E. Pouyoul, C. Guok, "Software-Defined Networking for Big Data Science", *Proc. of IEEE/ACM Supercomputing*, 2012.

[17] M. V. Neves, C. A. De Rose, K. Katrinis, H. Franke, "Pythia: Faster Big Data in Motion through Predictive Software-Defined Network Optimization at Runtime", *Proc. of IEEE IPDPS*, 2014

[18] N. McKeown, T. Anderson, H. Balakrishnan, et. al., "OpenFlow: Enabling Innovation in Campus Networks", *ACM SIGCOMM Computer Communication Review*, Vol. 38, No. 2, 2008.

[19] P. Lai, H. Subramoni, S. Narravula, A. Mamidala, and D. K. Panda, "Designing Efficient FTP Mechanisms for High Performance Data-Transfer over InfiniBand", *Proc. of ICPP*, 2009.

[20] E. Kissel, M. Swany, "Evaluating High Performance Data Transfer with RDMA-based Protocols in Wide-Area Networks", *Proc. of IEEE HPCC*, 2012.

[21] A. Hanemann, J. Boote, E. Boyd, et. al., "perfSONAR: A Service Oriented Architecture for Multi-Domain Network Monitoring", *Proc. of ICSOC*, 2005. (http://www.perfsonar.net)

[22] R. Morgan, S. Cantor, et. al., "Federated Security: The Shibboleth Approach", *EDUCAUSE Quarterly*, Vol. 27, No. 4, 2004.

[23] W. Kim, P. Sharma, J. Lee, S. Banerjee, J. Tourrilhes, S. Lee, P. Yalagandula, "Automated and Scalable QoS Control for Network Convergence", *Proc. of INM/WREN*, 2010.

[24] S. Seetharam, P. Calyam, T. Beyene, "ADON: Application-Driven Overlay Network-as-a-Service for Data-Intensive Science", *Proc. of IEEE CloudNet*, 2014.

[25] "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", *IDC: Analyze the Future*, 2012.
[26] "LHC Guide, English version. A collection of facts and figures about the Large Hadron Collider (LHC) in the form of questions and answers", *CERN-Brochure-2008-001-Eng. LHC Guide*, Vol. 20, 2013.
[27] G. Brumfiel, "High-Energy Physics: Down the Petabyte Highway", *Nature 469*, Vol. 19, 2011.
[28] P. Calyam, A. Berryman, A Lai, M. Honigford, "VMLab: Infrastructure to Support Desktop Virtualization Experiments for Research and Education", *VMware Technical Journal*, 2012.
[29] P. Calyam, S. Seetharam, R. Antequera, "GENI Laboratory Exercises Development for a Cloud Computing Course", *Proc. of GENI Research and Educational Experiment Workshop*, 2014.
[30] M.A. Sharkh, M. Jammal, A. Shami, A. Ouda, "Resource Allocation in a Network-based Cloud Computing Environment: Design Challenges", *IEEE Communications Magazine*, Vol. 51, 2013.
[31] P. Calyam, A. Kalash, N. Ludban, et. al., "Experiences from Cyberinfrastructure Development for Multi-user Remote Instrumentation", *Proc. of IEEE e-Science*, 2008.
[32] W. Allcock, "GridFTP: Protocol Extensions to FTP for the Grid", *Global Grid Forum GFD-R-P.020*, 2003.
[33] Hadoop Tutorial - http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
[34] R. Jain, S. Paul, "Network virtualization and software defined networking for cloud computing: a survey", *IEEE Communications Magazine*, Vol.51, 2013.
[35] L. Smarr, "UC-Wide Cyberinfrastructure for Data-Intensive Research", *Invited Talk @ UC IT Leadership Council in Oakland, California*, 2014.
[36] J. B. Bottum, R. Marinshaw, H. Neeman, J. Pepin, J. B. von Oehsen, "The Condo-of-Condos", *Proc. of XSEDE*, 2013.
[37] P. Calyam, J. Pu, W. Mandrawa, A. Krishnamurthy, "OnTimeDetect: Dynamic Network Anomaly Notification in perfSONAR Deployments", *Proc. of IEEE/ACM MASCOTS*, 2010.
[38] S. Tao, K. Xu, A. Estepa, et. al., "Improving VoIP Quality through Path Switching", *Proc. of IEEE INFOCOM*, 2005.
[39] B. Gaidioz, R. Wolski, B. Tourancheau, "Synchronizing Network Probes to avoid Measurement Intrusiveness with the Network Weather Service", *Proc. of IEEE HPDC*, 2000.
[40] W. Jia, L. Wang, "A Unified Unicast and Multicast Routing and Forwarding Algorithm for Software-Defined Datacenter Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 31, 2013.
[41] H. Subramoni, P. Lai, R. Kettimuthu, D. K. Panda, "High Performance Data Transfer in Grid Environment using GridFTP over InfiniBand", *Proc. of IEEE CCGrid*, 2010.
[42] B. Tierney, E. Kissel, M. Swany, E. Pouyoul, "Efficient Data Transfer Protocols for Big Data", *Proc. of IEEE e-Science*, 2012.
[43] N. Wilkins-Diehr, "Science Gateways: Common Community Interfaces to Grid Resources", *Concurrency and Computation: Practice and Experience*, Vol. 19, No. 6, 2007.
[44] HUBzero Platform for Scientific Collaboration - https://hubzero.org