Wide-area Overlay Networking to Manage Science DMZ Accelerated Flows

Prasad Calyam¹, Alex Berryman², Erik Saule³, Hari Subramoni², Paul Schopis², Gordon Springer¹, Umit Catalyurek², D. K. Panda²

¹University of Missouri-Columbia, ²The Ohio State University, ³University of North Carolina-Charlotte
{calyamp, gks}@missouri.edu, {berryman.15, subramoni.1, catalyurek.1, schopis.1, panda.2}@osu.edu, esaule@uncc.edu

Abstract—There is a new trend emerging across university campuses to deploy Science DMZs (demilitarized zones) to support science drivers that involve for e.g., data-intensive applications needing access to remote instrumentation or public cloud resources. Using advanced technologies such as “multi-domain” software-defined networking, zero-copy RDMA data transfers, active measurements and federated identity/access – accelerated flows are starting to be set up from Science DMZs over wider-area overlay networks, by-passing traditional campus firewalls. In this paper, we present a “campus Science DMZ reference architecture” for adaptively managing host-to-host accelerated flows of multiple researchers over wide-area overlay networks with shared under lay infrastructure components. We discuss our novel approaches in handling challenges of policy specification, security enforcement, and performance engineering within Science DMZs to support diverse accelerated flows on a scalable/extendible basis. Lastly, we present a multi-disciplinary case study of a bioinformatics science driver application in a double-ended campus Science DMZ tested. Our case study illustrates how our reference architecture can enable new “High-Throughput Computing services” that improve remote accessibility and peer-collaboration of data-intensive science users, and simplify related operations/management for campus network service providers.

I. INTRODUCTION

In recent years, cyberinfrastructures have been advancing rapidly to enable researchers to: (a) remotely access distributed computing resources and big data sets, and (b) effectively collaborate with remote peers, at a global-scale. Today’s cyberinfrastructures are driven by a growing convergence of data-intensive science needs (e.g., Bioinformatics for gene sequencing) that are served through increased end-to-end high-speed connectivity (i.e., availability of 40/100Gbps routers), and emerging computer/network virtualization management technologies. The examples of virtualization management technologies include: (i) software-defined networking (SDN) based on programmable OpenFlow switches [1], (ii) RDMA over Converged Ethernet (RoCE) implemented between zero-copy data transfer nodes [2], (iii) multi-domain network performance monitoring using perfSONAR active measurement points [3], and (iv) federated identity/access management using Shibboleth-based entitlements [4].

In order to configure such cyberinfrastructures at university research labs and federate them with remote sites, there is a new trend emerging across university campuses to deploy Science DMZ [5] (demilitarized zone) network designs.

This work was supported by the National Science Foundation under awards: ACI-1246001 and ACI-1245795, and Cisco Systems. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Cisco Systems.
Science DMZs alternately enable configuration of firewall friction-free overlay networks that can accelerate bandwidth-intensive or latency-sensitive flows even across wide-area networks (spanning multiple domains) by leveraging SDN techniques for centralized, rapid policy-directed traffic management. They allow for seamless integration with public clouds that support VLAN extensions such as AWS Direct Connect [8] or GENI Slices [9] to realize for e.g., loss-free wide-area network path provisioning with guaranteed end-to-end available bandwidth needed for technologies such as RoCE to accelerate data transfers of science driver applications. Obviously, there are significant investments needed for supplementing traditional campus infrastructures with parallel high-speed networks at 10+ Gbps speeds and for maintaining the various virtualization management technologies.

However, the greater challenges lie in developing and deploying solutions that can simplify operations while handling policy specifications of ‘on-demand’ researcher-specific overlays over shared infrastructure components at the campus department, campus border and backbone network levels. In addition, management of overlay network flows that are authorized to use Science DMZ resources (i.e., are allowed to by-pass traditional campus firewalls) also requires adequate security enforcement mechanisms to be in place to prevent abuse. Further, timely/accurate measurements across multi-domain overlay network paths are necessary for performance engineering of the acceleration between data transfer nodes, and for bottleneck troubleshooting.

In this paper, we present a “campus Science DMZ reference architecture” and a case study for adaptively managing host-to-host accelerated flows of multiple researchers over wide-area overlay networks with shared underlay infrastructure components. To handle the above challenges of policy specification, security enforcement, and performance engineering within Science DMZs, we propose novel approaches that can support diverse accelerated flows on a scalable/extensible basis. More specifically, we define the functionality and workload of a “policy-directory/gatekeeper-proxy” maintained by a technician in a “Performance Engineer” role to securely enforce access policies specific to individual (authorized) researcher needs. The role of a “Performance Engineer” is vital for identity and access management of Science DMZ flows, and augments traditional System/Network Engineer roles on campuses. For serving the rapid-provisioning and adaptive management of application performance engineering within extended VLAN overlays, we describe requirements for a Science DMZ OpenFlow controller hosted within the gatekeeper-proxy that orchestrates multi-domain RoCE-based data transfer nodes, and perfSONAR-based measurement points.

Finally, we describe how we are implementing our proposed reference architecture in a multi-disciplinary case study of a bioinformatics science driver application that uses “Data-Cutter” [10] in a double-ended campus Science DMZ tested between University of Missouri-Columbia and The Ohio State University. The double-ended nature of the tested allows for a wide array of experimentation of ‘high-friction’, ‘reduced-friction’, and ‘friction-free’ configurations between host-to-host flows over wide-area networks. We believe our current development and tested experimentation efforts in the case study can foster new “High-Throughput Computing (HTC) services” on a scalable/extensible basis that can greatly simplify Science DMZ operations/management at campuses.

The remainder of the paper is organized as follows: Section II presents the campus science DMZ challenges and requirements for our reference architecture. Section III presents a bioinformatics science driver application case study within a double-ended campus Science DMZ testbed that implements our reference architecture. Section IV concludes the paper and presents future work.

II. CHALLENGES AND REQUIREMENTS

A. Infrastructure Investments

Fig. 2 presents our layered reference architecture illustration for deploying Science DMZs on campuses. 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 a microscope facility at one researcher’s site, and a High-performance Computing (HPC) facility at the other researcher’s site. At the networking layer, as described in Section I, both the researchers use their respective campus access networks, and connect over an extended VLAN overlay between their lab sites, and use the direct connect network to use public cloud resources. Thus, they can use each other’s resources and the public cloud resources as if they all reside within the same internal network, and their traffic can be isolated from other cross-traffic through loss-free, dedicated ‘on-demand’ bandwidth on the shared network overlay.

As expected in such scenarios, 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. Given that network devices that support 40 - 100 Gbps speeds are expensive at the time of this writing, building overlay networks requires significant investments from both the central campus as well as the departmental units. In addition, the backbone network providers at the regional and national level need to create a wide footprint of their backbones that can support these extended VLAN overlays between campuses. The end-to-end infrastructure should preferably feature OpenFlow-capable [1] switches at strategic traffic aggregation points, within the campus and backbone networks in order to proactively or reactively provision and steer traffic flows in a unified as well as vendor-independent manner. We can expect two possibilities, the first where there is no separate research network infrastructure within a campus. In such cases, the static or dynamic VLANs need to be built on top of the existing campus core network. In the second possibility where there is a separate research network infrastructure that is separate from the campus core, VLAN provisioning for Science DMZ flows can more ideally leverage the dedicated infrastructure.
B. Handling Policy Specifications

Assuming the relevant infrastructure investments are in place, the next challenge relates to the ‘federated identity and access management’ that requires specifying and handling fine-grained resource access policies in a multi-institution collaboration setting (i.e., at both the researcher campuses, and within the backbone networks) with minimal administrative overhead. There are several questions that need to be addressed in the context of multi-institutional policy specification and handling 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?; or even (iii) How do we restrict who can query the performance measurement data from the perfSONAR measurement archives within the extended VLAN overlay network that supports many researchers over time?

Fortunately, standards-based identity management approaches based on Shibboleth entitlements [4] have evolved to accommodate permissions in above user-to-service authentication/authorization use cases, and 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 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, and perfSONAR measurement points. Note that the user name and password needed for authentication of a researcher in the federation does not leave the researcher’s institutional boundary, and assertions from the Service Provider regarding researcher’s identity are handled exclusively by the campus Identity Provider. The Internet2 InCommon federation is an exemplar that supports the registration of such Service Providers [11] to facilitate multi-institutional collaborations.

C. Security Enforcement

In order to securely enforce the policy directories of the federation, and to allow institutional policy management of the Science DMZ flows, a ‘gatekeeper-proxy middleware’ as shown in Fig. 2 is required. The concept of the gatekeeper-proxy although is novel in the context of Science DMZs, it originates from the H.323 videoconferencing community where a similar middleware concept was used to allow H.323 sessions to by-pass firewalls when they used dynamic ports (port range: $>2^{10}$ and $<2^{16}$) to exchange voice and video traffic flows [12]. 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’, (b) Management toolkit that includes dashboards of flow performance status/issues and meta-data of flows for auditability through a ‘measurement engine’, and (c) Science Driver application (i.e., the image processing application in Fig. 2) as part of a HTC ‘service engine’ which runs the user-facing portal when a researcher requests access to external VLAN overlay resources.

To effectively maintain the gatekeeper-proxy to serve diverse researcher needs concurrently on the shared underlay infrastructure components, 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 “Performance Engineers” for related tasks.

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 co-ordinating with multi-domain entities at various levels (i.e., building-level, campus-level, backbone-level), and also has to cater to each researcher’s expectations of high-availability and peak-performance to remote sites – without disrupting core campus network traffic. 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. Further, the tools need to enable monitoring of the more important ‘control traffic’ patterns of researcher flows being authorized by the gatekeeper-proxy to use extended overlay resources, than actually monitoring ‘data traffic’, which is the primary focus in the core campus network protected by firewalls.

D. Performance Engineering

For serving the rapid-provisioning and adaptive management of the Science DMZ flows to consistently deliver peak performance and fault tolerance, the routing engine that runs the OpenFlow controller in the gatekeeper-proxy has to orchestrate end-to-end service engine comprising of data transfer node components, and the measurement engine comprising of measurement point appliance components.

1) Data Transfer Nodes: The main components of the service engine that produce the ‘acceleration’ for the researcher data flows within the extended VLAN overlay are the Data Transfer Nodes (DTNs). Particularly in the context of real-time applications such as the scenario in Fig. 2 between the microscope facility researcher and the HPC facility researcher, RoCE-based DTNs are more pertinent than traditional data transfer tools such as SCP, RSYNC, or GridFTP [7] that are frequently used over low throughput links. There is a growing consensus based on original works in [2] and [13], and recent works [14] and [15] that – RoCE-based DTNs are a strong successor for gaining higher acceleration in extended VLAN overlays for scientific data distribution, content replication and remote data backups.

RoCE has lower overhead owing to the zero-copy RDMA protocol characteristics than the TCP/UDP based file transfer protocols used in traditional tools. The zero-copy characteristic refers to the reduction in latencies when RoCE is used to place data between data centers (e.g., from the microscope computer’s memory to the remote HPC node’s memory in Fig. 2) with minimal demands on the memory bus bandwidth and CPU processing overhead on either sides. Scientific and
parallel computing domains that use the Message Passing Interface (MPI) as the underlying basis have been taking advantage of high performance interconnects such as iWARP and RoCE through implementations of MPI, such as MVAPICH and MVAPICH2 [2] that support these interconnects. The ability to provision loss-free, high throughput extended VLAN overlays with guaranteed end-to-end available bandwidth between remote data centers makes the currently possible use of RoCE feasible across Science DMZs.

2) **Measurement Point Appliances**: The performance intelligence within the measurement engine to monitor Science DMZ flows and troubleshoot bottlenecks is provided by Measurement Point Appliances (MPAs) that perform multi-domain active measurements within the Science DMZ federation networks. Currently, perfSONAR [3] is the most widely-deployed (over 600 publicly registered MPAs) framework in measurement federations for performing multi-domain active measurements with tools such as Ping (for round trip delay measurements), Traceroute (for packet routes and network topology inference), OWAMP (for one-way delay measurements) and BWCTL (for TCP/UDP throughput measurements). One of the reasons for perfSONAR popularity is its ability to allow users to collect and share measurement archives from MPAs across administrative domains through interoperable web-service interfaces.

To avoid having barriers of wide-adoption (e.g., overheads in setting up federation and measurement level agreements), the current trends in implementation of perfSONAR MPAs are to have them in a default “totally open” mode to run tests on measurement points and view data within measurement archives. In this mode, the measurement points and data archives can be discovered by anonymous users and there is minimal regulation imposed by restricting maximum probing bandwidth utilization for active measurement tools such as Ping, Traceroute, OWAMP and BWCTL or traffic generators such as [19] that can be used to emulate actual Science DMZ application traffic. This totally open mode limits the security options for an enterprise and may result in undesired measurement tests and data shares within a domain’s measurement infrastructure resources.

To suit the establishment and enforcement of measurement level agreements within Science DMZ environments and related measurement federations of multiple collaborating institutions, we recently developed a “Resource Protection Service” viz., OnTimeSecure [16]. This service supplements the “totally open” mode through ‘user-to-service’ and ’service-to-service’ authentication/authorization through the Internet2 InCommon ‘Service Provider with Entitlement Service’ concept described in Section II-B. It features a hierarchical policy-engine and is built using RESTful APIs that are modular for extensibility, and are interoperable with perfSONAR standards based deployments for various measurement functions such as: measurement point discovery, test initiation and measurement data query. The policy-engine also interfaces with our metascheduler [17] [18] for prioritization of measurement requests and conflict-free scheduling within MPAs, while users concurrently attempt to utilize measurement resources.

3) **Orchestration Workflow**: Fig. 3 shows the overall gatekeeper-proxy orchestration workflow for: (i) policy-directed accelerated flow provisioning for installing the ‘control flows’ of the DTNs and the MPAs, and the (ii) performance-aware HTC Service delivery that involves monitoring the performance of the ‘data flows’ and adaptively configuring the service engine for consistent peak performance.

We can see that the gatekeeper-proxy orchestration workflow is initiated by a researcher authenticated in a Science DMZ federation portal, who is allowed to define the science driver application’s end-points, and the application’s monitoring objectives to sustain real-time session performance. The Performance Engineer receives the request, and assuming a HTC Service definition is in place (based on the federation’s Entitlement Service response), the corresponding policy-directed flow rules are generated using the gatekeeper-proxy management tools. We can envisage that the list of federation resources available to reserve an overlay at any moment can be discovered using perfSONAR Global Lookup Service, GENI Aggregate Manager API, AWS Console, and such other services. The rules are then sent to the OpenFlow controller to install the HTC flows for the Science driver application, and to install the concurrent measurement flows for monitoring the flow performance. If a HTC Service definition is not in place, we assume that the Performance Engineer is able to choose from a policy template library, and customize it to serve the specific researcher’s request. Next, the researcher is notified that the requested extended VLAN overlay resources are in place, and that the application session can be started.

Considering the same scenario in Fig. 2, let us assume a HTC Service that enables a researcher at the microscope facility to transfer large image files that require 10+ Gbps bandwidth links to meet the real-time needs for the analysis session. Given that the HTC flow of the researcher has a HTC Service definition in place that covers access to resources at both the researcher’s and collaborator’s Science DMZs, the image transfers can be expected to happen over the extended VLAN overlay between the DTNs within a few minutes of the original request by the researcher. However, in the case there are incorrect or non-existent HTC Service definitions, we can see that the researcher flows are treated as non-Science DMZ flows and are routed through the campus core networks with firewall friction laden paths that result in large response times. Further, if there are cases where multiple researchers (i.e., tenants) are contending for Science DMZ resources, the gatekeeper-proxy is responsible for: (a) ensuring isolation and load balancing between the tenant flows by leveraging SDN
capabilities, and (b) deferring immediate scheduling of new flows if real-time requirements of already provisioned flows will be impacted with any new flow admission.

The major objective of monitoring a Science DMZ is to accurately gauge the performance of not only the network path, but also the end-to-end performance of the science driver applications to avoid cases of incorrect HTC Service definitions. When interpreting the performance of an application, it is essential to first obtain the baseline performance established by active measurement tools within an extended VLAN overlay under presumably ideal conditions. For e.g., if an end-to-end available bandwidth test shows 8 – 9 Gbps of throughput on a path, then passive monitoring of an application should be set to alert the Performance Engineer (e.g., using AWS CloudWatch alarm or perfSONAR alarm notification through static threshold setting of anomaly detection) if the achieved throughput of the actual application (as seen in data transfer application log files) falls below 8 – 9 Gbps. We remark that it is important to aggregate passive monitoring of network utilization classified by different scientific applications concurrently using the Science DMZ resources, as well as active measurements scheduled within MPAs that verify the potential of the network. If two applications are each using 4 Gbps of the available bandwidth on a 10 Gbps path, and an active throughput test reports less than 1 Gbps of available bandwidth, then no alarm should be raised since the link is still achieving nominal performance once the application’s passive performance metrics are factored in.

Thus, the passive monitoring can be configured to gauge the level of performance a particular scientific application is able to achieve at a given point in time. The accuracy of this relation between passive monitoring, and active measurements is critical and special care must be taken to ensure that the impact of one on the other is minimized and is accounted for in any interpretation of the results. In order to achieve the above performance awareness, the schedule and parameters of active measurement tests between MPAs must be cognizant of the presence and characteristics of current/expected scientific driver application flows. This type of awareness is feasible within the measurement engine of the gatekeeper-proxy since both the application flows and monitoring flows are provisioned and orchestrated centrally through the routing engine.

III. CASE STUDY

In this section, we describe how we are implementing the proposed reference architecture and workflow described in previous Section II in a multi-disciplinary case study. Our case study features a bioinformatics science driver application collaboration between two researchers in a double-ended, long-distance campus Science DMZ testbed that we are building between the University of Missouri-Columbia (MU), and The Ohio State University (OSU); the two sites are more than 500 miles apart geographically within the United States.

A. Double-ended Science DMZ Testbed Setup

Similar to the setup shown in Fig. 1, both the OSU and MU campuses are connected through an extended VLAN overlay that involves an Internet2 AL2S connection by way of local regional networks of OARnet in Ohio, and GPN/MoreNet in Missouri, respectively. Each Science DMZ has a matching DTN equipped with dual Intel 5E-2660, 128GB of memory, 300GB PCI-Express solid state drive, and dual Mellanox 10 Gbps network cards with RoCE support. Each Science DMZ has Narada Metrics MPAs that are based on the OnTimeSecure capabilities for continuous monitoring at 1 – 10 Gbps network speeds using perfSONAR-compliant web services.

A common Dell R610 node in the OSU Science DMZ is used to run the the GENI FOAM (Flowvisor OpenFlow Aggregate Manager) software [9] as well as a POX OpenFlow controller that controls both the OSU and MU Science DMZ OpenFlow switches. The Science DMZ at OSU contains a heterogeneous mix of OpenFlow switches from NEC and HP. Two HP 3800s are used to attach to the various nodes in the Science DMZ, and a single NEC PF5820 aggregates the two connections. The NEC switch is connected to OSU’s 100 Gbps Cisco Nexus router at 10 Gbps, but 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 from Brocade and HP; a Brocade VDX 8770 switch to attach various nodes in the Science DMZ, and a 100 Gbps Brocade MLX router at 10 Gbps interface speeds, with the ability to scale to 100 Gbps speeds.

B. Bioinformatics Science Driver Application

The double-ended Science DMZ deployment between OSU and MU has garnered support and attention from a number of researchers interested in multiple areas of collaboration between the two campuses. For the purpose of our case study, we now describe an exemplar Bioinformatics science driver application that uses “Data Cutter” [10] to highlight our reference architecture implementation work with two research labs, one at a Microscope lab facility at MU interested in data-intensive medical image processing, and the other at a HPC lab facility at OSU interested in HPC code optimization and RoCE, similar to the setup in Figs. 2 and 3.

1) Real-Time Medical Image Processing Need: The HTC Service we are designing as part of the case study aims to utilize the double-ended Science DMZ testbed to lower the time-to-result when performing real-time medical image processing to diagnose ‘Neuroblastoma’, a kind of cancer that affects immature or developing cells. This type of cancer mostly develops in children of less than 5 years old and more than half the cases affect children of less than 2 years old. The appropriate treatment for this cancer is decided based on a histopathological evaluation of tissue samples. That is to say, a doctor takes an image of a tissue sample and manually counts the number of diseased cells at different magnification levels. This process is extremely long and error prone, and thus automated techniques have become necessary for the analysis process. Recently computer-aided automatic diagnostic methods have been developed that have shown the proposed methods to be pleasingly parallel and can be executed using GPUs to obtain accurate results. However, at the resolution needed to perform the repeated analysis of a single slide (an image of 100K x 100K pixels), a single modern GPU (NVIDIA C2050) still needs about 3 minutes.
Considering the example shown in Fig. 4 for a set of analyzed images at 3 different magnification levels, automatic techniques for such analysis in DataCutter applications can only be approved for medical use if a decision can be obtained in a short time, such as ≈10 – 20 seconds. This could be achieved by a small GPU cluster system that is expensive to procure and maintain. Moreover, requiring every lab to own such a Neuroblastoma analysis system would be inefficient and cost-prohibitive. Creating a central GPU cluster that is remotely accessible via Science DMZ based friction-free paths could be seamlessly used by multiple remote researchers for rapid and accurate analysis process pertaining to Neuroblastoma or other such diseases. Such an use would help in alleviating the overall cost of analysis, and can foster the collaboration of GPU processing algorithm experts with the image analysis users. For this particular science driver application, a sustained 10 Gbps+ connection between the microscope facility at MU and the Owens HPC cluster site at OSU is found to be required to be able to satisfy the real-time analysis constraints.

For the case study experiments, we are testing image streaming configurations from the disk of a DTN located at MU. The Data Cutter application with a DTN is setup to be running on the Owens HPC cluster in order to access the images at MU DTN through the double-ended Science DMZ network paths and to feed image tiles to multiple GPU-enabled nodes for real-time analysis. We remark that the DataCutter is a component-based middleware that relies on the filter stream programming model. Therefore, it is very easy to make a DataCutter application cross the frontier of a cluster. On a typical Internet connection, it is enough to instantiate two instances of the DataCutter application on each side of the Science DMZ and add routing filters that abstract the network communicating over TCP. Multiple TCP streams can be added to improve performance up to a point after which the performance improvement saturates.

In the double-ended Science DMZ testbed experiments, we are also augmenting DataCutter to use the RoCE-capable DTNs as described in the next section. Fig. 5 shows the RoCE integration with image processing application components used to capture and analyze the Neuroblastoma images. We are evaluating testbed setup configurations that can help us study the impact of the number of nodes on the computation time of the real-time analysis, as well as how many network streams are necessary to saturate the extended VLAN overlay links for peak performance.

2) Flow Performance Acceleration in HTC Service: Referring to Fig. 5, we are working on enhancing our Advanced Data Transfer Service (ADTS) designed in [2] to take advantage of all the capabilities that high performance interconnects offer. Our current approach is to re-use the best communication run-time designs that have already been developed for high performance scientific computation and adapt them to be used for long-distance data transfers with DTNs across the double-ended Science DMZ testbed between OSU and MU. To this end, we are implementing the basic put and get functionalities of an FTP client/server program on top of our ADTS interface. We are also extending the basic ADTS design proposed in [2] to have end-to-end RoCE support between the MU and OSU DTNs. We are also working to integrate an ADTS software variant with support for RoCE in the GridFTP file transfer package that is widely-used in HPC user communities. As indicated in Fig. 5, the ADTS software is capable of dynamically detecting the kind of networking protocol supported by the underlying hardware and adapting itself to use either the Verbs API as well as RDMA enabled technologies in user space, or the TCP sockets and non-RDMA technologies in kernel space. Note the fact in Fig. 5 that the communication between the Imaging Microscope as well as the Image Processing Cluster with the DTNs is based on MPI protocols within their respective LAN infrastructures.

We plan to evaluate the performance of our FTP design in WAN scenario experiments considering a wide array of experimentations of ‘high-friction’ (i.e., firewalls on both sides, absence of SDN-capable paths), ‘reduced friction’ (i.e., firewalls on one side, partially SDN-capable network aggregation points) and ‘friction-free’ (i.e., double-ended Science DMZ overlay with fully SDN-capable strategic network aggregation points) using RAM disks, high speed Solid State Drive (SSD) devices as well as traditional Hard Disk Drives (HDDs) as source and sink for data. We expect our ADTS design to perform better than FTP-UDP and GridFTP in low-delay, high-bandwidth LAN settings.
From our earlier studies using the Obsidian WAN simulators [13], we know that for low-delay scenarios, the best performance is obtained when we use smaller staging buffers. Buffering packets helps improve file transfer performance only when I/O performance is worse than the network performance or when there is high delay factors in the intermediate network path. Thus, we are working on performing tests between the OSU and MU DTNs by using different staging buffer size configurations. In addition, we plan to use the SDN capabilities end-to-end within Internet2 AL2S and within regional and last-mile campus networks to evaluate the performance at different delay scenarios and route topologies between OSU and MU.

However, we were able to readily perform tests of client-perceived transfer times of both put and get operations between OSU and MU DTNs for a fixed buffer size of 4 MB and with different file size configurations over the Internet2 AL2S connection. Figure 6 shows the improved performance observed in our tests with ADTS due to memory-to-memory transfers over the GridFTP UDT (reliable UDP-based application-level Data Transfer protocol) configuration, which shows relatively lower performance due to effects of disk I/O. It is interesting to note that the higher improvements in performance with ADTS are more significant in cases of smaller file sizes. This has favorable effects for real-time processing applications, which typically decompose large images into smaller file sizes of the different image regions for parallel processing with multiple cores. Since the processed files to be transferred back for rendering can also have small file sizes, the overall communication time for the client-side image analysis is greatly improved with the ADTS enhancements in the DTNs. If the file sizes are larger, ADTS still outperforms UDT but the improvement is not as significant as seen with the lower file size ranges of 16 – 32 MB.

IV. CONCLUSION AND FUTURE WORK

In this paper, we described a campus Science DMZ reference architecture to show how multi-domain accelerated flows can be setup to enable inter-institutional collaborations of data-intensive science driver applications. Our reference architecture descriptions addressed the handling of policy specification, security enforcement and performance engineering within Science DMZs to support diverse accelerated flows on a scalable/extendible basis. We defined the role of a Performance Engineer who is the 'keeper' and 'help-desk' of the Science DMZ equipment that features RoCE-based DTN, OpenFlow-capable switches, perfSONAR-compliant MPAs, as well as Shibboleth-based entitlements as part of a federation that has a unified ‘gatekeeper-proxy middleware’.

We also presented an exemplar case study that illustrates our reference architecture implementation in a multi-disciplinary, real-time application based on bioinformatics science driver for Neuroblastoma image analysis using shared GPU cluster resources across a double-ended Science DMZ testbed. Our gatekeeper proxy implementation and testbed experimentation efforts are targeted towards fostering new HTC services that can simplify Science DMZ operations/management at campuses, and more importantly – improve the remote accessibility and peer-collaboration of data-intensive science users across multiple institutional boundaries with Science DMZ accelerated flows management over extended VLAN overlays.

REFERENCES