The Science DMZ: A Network Design Pattern for Data-Intensive Science

Eli Dart
Energy Sciences Network
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
eddart@lbl.gov

Lauren Rotman
Energy Sciences Network
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
lbrotman@lbl.gov

Brian Tierney
Energy Sciences Network
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
bltierney@lbl.gov

Mary Hester
Energy Sciences Network
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
mchester@lbl.gov

Jason Zurawski
Internet2
Office of the CTO
Washington DC, 20036
zurawski@internet2.edu

Abstract

The ever-increasing scale of scientific data has become a significant challenge for researchers that rely on networks to interact with remote computing systems and transfer results to collaborators worldwide. Despite the availability of high-capacity connections, scientists struggle with inadequate cyberinfrastructure that cripples data transfer performance, and impedes scientific progress. The *Science DMZ* paradigm comprises a proven set of network design patterns that collectively address these problems for scientists. We explain the Science DMZ model, including network architecture, system configuration, cybersecurity, and performance tools, that creates an optimized network environment for science. We describe use cases from universities, supercomputing centers and research laboratories, highlighting the effectiveness of the Science DMZ model in diverse operational settings. In all, the Science DMZ model is a solid platform that supports any science workflow, and flexibly accommodates emerging network technologies. As a result, the Science DMZ vastly improves collaboration, accelerating scientific discovery.

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

http://dx.doi.org/10.1145/2503210.2503245

Categories and Subject Descriptors

C.2.1 [**Computer–Communication Networks**]: Network Architecture and Design; C.2.3 [**Computer–Communication Networks**]: Network Operations—*network management, network monitoring*; C.2.5 [**Computer–Communication Networks**]: Local and Wide-Area Networks—*internet*

General Terms

Performance, Reliability, Design, Measurement

1. INTRODUCTION

A design pattern is a solution that can be applied to a general class of problems. This definition, originating in the field of architecture [1,2], has been adopted in computer science, where the idea has been used in software designs [6] and in our case network designs. The network design patterns we discuss are focused on high end-to-end network performance for data-intensive science applications. These patterns focus on optimizing the network interactions between wide area networks, campus networks, and computing systems.

The Science DMZ model, as a design pattern, can be adapted to solve performance problems on *any* existing network. Of these performance problems, packet loss has proven to be the most detrimental as it causes an observable and dramatic decrease in data throughput for most applications. Packet loss can be caused by many factors including: firewalls that cannot effectively process science traffic flows; routers and switches with inadequate burst capacity; dirty optics; and failing network and system components. In addition, another performance problem can be the misconfiguration of data transfer hosts, which is often a contributing factor in poor network performance.

Many of these problems are found on the local area networks, often categorized as "general-purpose" networks, that are not designed to support large science data flows. Today many scientists are relying on these network infrastructures to share, store, and analyze their data which is often geographically dispersed.

The Science DMZ provides a design pattern developed to specifically address these local area network issues and offers research institutions a framework to support data-intensive science. The Science DMZ model has been broadly deployed and has already become indispensable to the present and future of science workflows.

The Science DMZ provides:

- 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.

This paper will discuss the Science DMZ from its development to its role in future technologies. First, Section 2 will discuss the Science DMZ's original development in addressing the performance of TCP-based applications. Second, Section 3 enumerates the components of the Science DMZ model and how each component adds to the overall paradigm. Next, Sections 4 and 5 offer some sample illustrations of networks that vary in size and purpose. Following, Section 6 will discuss some examples of Science DMZ implementations from the R&E community. And lastly, Section 7 highlights some future technological advancements that will enhance the applicability of the Science DMZ design.

2. MOTIVATION

When developing the Science DMZ, several key principles provided the foundation to its design. First, these design patterns are optimized for science. This means the components of the system—including all the equipment, software and associated services—are configured specifically to support data-intensive science. Second, the model is designed to be scalable in its ability to serve institutions ranging from large experimental facilities to supercomputing sites to multi-disciplinary research universities to individual research groups or scientists. The model also scales to serve a growing number of users at those facilities with an increasing and varying amount of data over time. Lastly, the Science DMZ model was created with future innovation in mind by providing the flexibility to incorporate emerging network services. For instance, advances in virtual circuit services, 100 Gigabit Ethernet, and the emergence of software-defined networking present new and exciting opportunities to improve scientific productivity. In this section, we will mostly discuss the first principle since it is the driving mission for the Science DMZ model.

The first principle of the model is to optimize the network for science. To do this, there are two entities or areas of the network that should be considered: the wide area network and the local area networks. The wide area networks (or WANs) are often already optimized and can accommodate large data flows up to 100Gbps. However, the local area networks are still a choke point for these large data flows.

Local area networks are usually general-purpose networks that support multiple missions, the first of which is to support the organization's business operations including email, procurement systems, web browsing, and so forth. Second, these general networks must also be built with security that protects financial and personnel data. Meanwhile, these networks are also used for research as scientists depend on this infrastructure to share, store, and analyze data from many different sources. As scientists attempt to run their applications over these general-purpose networks, the result is often poor performance, and with the increase of data set complexity and size, scientists often wait hours, days, or weeks for their data to arrive.

Since many aspects of general-purpose networks are difficult or impossible to change in the ways necessary to improve their performance, the network architecture must be adapted to accommodate the needs of science applications without affecting mission critical business and security operations. Some of these aspects that are difficult to change might include the size of the memory buffers for individual interfaces; mixed traffic patterns between mail and web traffic that would include science data; and emphasis on availability vs. performance and what can be counted on over time for network availability.

The Science DMZ model has already been implemented at various institutions to upgrade these general-purpose, institutional networks. The National Science Foundation (NSF) recognized the Science DMZ as a proven operational best practice for university campuses supporting data-intensive science and specifically identified this model as eligible for funding through the Campus Cyberinfrastructure–Network Infrastructure and Engineering Program (CC-NIE).[1] This program was created in 2012 and has since been responsible for implementing approximately 20 Science DMZs at different locations—thereby serving the needs of the science community. Another NSF solicitation was released in 2013 and awards to fund a similar number of new Science DMZ's are expected.

2.1 TCP Performance

The Transmission Control Protocol (TCP) [15] of the TCP/IP protocol suite is the primary transport protocol used for the reliable transfer of data between applications. TCP is used for email, web browsing, and similar applications. Most science applications are also built on TCP, so it is important that the networks are able to work with these applications (and TCP) to optimize the network for science.

TCP is robust in many respects—in particular it has sophisticated capabilities for providing reliable data delivery in the face of packet loss, network outages, and network congestion. However, the very mechanisms that make TCP so reliable also make it perform poorly when network conditions are not ideal. In particular, TCP interprets packet loss as network congestion, and reduces its sending rate when loss is detected. In practice, even a tiny amount of packet loss is enough to dramatically reduce TCP performance, and thus increase the overall data transfer time. When applied to large tasks, this can mean the difference between a scientist completing a transfer in days rather than hours or minutes. Therefore, networks that support data-intensive science must provide TCP-based applications with loss-free service if TCP-based applications are to perform well in the general case.

As an example of TCP's sensitivity, consider the follow-

[1] NSF's CC-NIE Program: `http://www.nsf.gov/pubs/2013/nsf13530/nsf13530.html`.

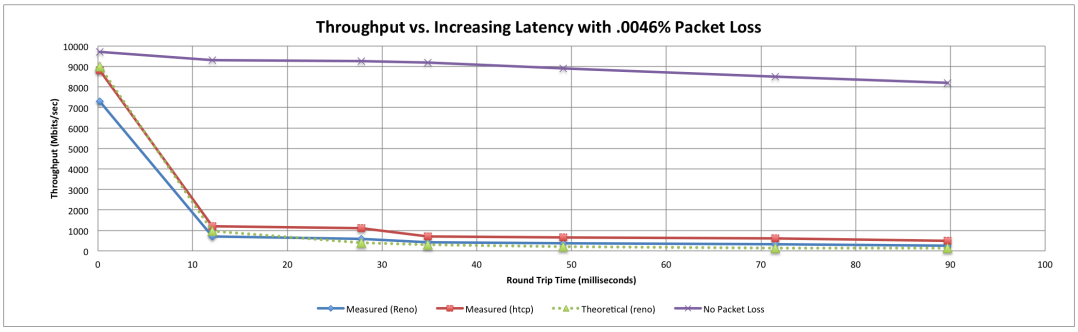

Figure 1: Graph shows the TCP throughput vs. round-trip time (latency) with packet loss between 10Gbps connected hosts, as predicted by the Mathis Equation. The topmost line (shown in purple) shows the throughput for TCP in a loss-free environment.

ing case. In 2012, Department of Energy's (DOE) Energy Sciences Network (ESnet) had a failing 10 Gbps router line card that was dropping 1 out of 22,000 packets, or 0.0046% of all traffic. Assuming the line card was working at peak efficiency, or 812,744 regular sized frames per second,[2] 37 packets were lost each second due to the loss rate. While this only resulted in an overall drop of throughput of 450 Kbps (on the device itself), it reduced the end-to-end TCP performance far more dramatically as demonstrated in Figure 1. This packet loss was not being reported by the router's internal error monitoring, and was only noticed using the *owamp* active packet loss monitoring tool, which is part of the perfSONAR Toolkit [3].

Because TCP interprets the loss as *network congestion*, it reacts by rapidly reducing the overall sending rate. The sending rate then slowly recovers due to the dynamic behavior of the control algorithms. Network performance can be negatively impacted at any point during the data transfer due to changing conditions in the network. This problem is exacerbated as the latency increases between communicating hosts. This is often the case when research collaborations sharing data are geographically distributed. In addition, feedback regarding the degraded performance takes longer to propagate between the communicating hosts.

The relationship between latency, data loss, and network capability was described by Mathis et al. as a mechanism to predict overall throughput [12]. The "Mathis Equation" states that maximum TCP throughput is at most:

$$\frac{\text{maximum segment size}}{\text{round-trip time}} \times \frac{1}{\sqrt{\text{packet loss rate}}}. \qquad (1)$$

Figure 1 shows the theoretical rate predicted by the Mathis Equation, along with the measured rate for both TCP-Reno and TCP-Hamilton across ESnet. These tests are between 10Gbps connected hosts configured to use 9KByte ("Jumbo Frame") Maximum Transmission Units (MTUs).

This example is indicative of the current operational reality in science networks. TCP is used for the vast majority of high-performance science applications. Since TCP is so sensitive to loss, a science network must provide TCP with a loss-free environment, end-to-end. This requirement, in turn, drives a set of design decisions that are key components of the Science DMZ model.

[2]Performance Metrics. `http://www.cisco.com/web/about/security/intelligence/network_performance_metrics.html`.

[3]perfSONAR Toolkit: `http://psps.perfsonar.net`

3. THE SCIENCE DMZ DESIGN PATTERN

The overall design pattern or paradigm of the Science DMZ is comprised of four sub-patterns. Each of these sub-patterns offers repeatable solutions for four different areas of concern: proper location (in network terms) of devices and connections; dedicated systems; performance measurement; and appropriate security policies. These four sub-patterns will be discussed in the following subsections.

3.1 Proper Location to Reduce Complexity

The physical location of the Science DMZ (or "location design pattern") is important to consider during the deployment process. The Science DMZ is typically deployed at or near the network perimeter of the institution. The reason for this is that it is important to involve as few network devices as reasonably possible in the data path between the experiment at a science facility, the Science DMZ, and the WAN.

Network communication between applications running on two hosts traverses, by definition, the hosts themselves and the entire network infrastructure between the hosts. Given the sensitivity of TCP to packet loss (as discussed in Section 2.1), it is important to ensure that all the components of the network path between the hosts are functioning properly and configured correctly. Wide area science networks are typically engineered to perform well for science applications, and in fact the Science DMZ model assumes that the wide area network is doing its job. However, the local network is often complex, and burdened with the compromises inherent in supporting multiple competing missions. The location design pattern accomplishes two things. The first is separation from the rest of the general network, and the second is reduced complexity.

There are several reasons to separate the high-performance science traffic from the rest of the network. The support of high-performance applications can involve the deployment of highly capable equipment that would be too expensive to use throughout the general-purpose network but that has necessary features such as high-performance filtering capabilities, sufficient buffering for burst capacity, and the ability to accurately account for packets that traverse the device. In some cases, the configuration of the network devices must be changed to support high-speed data flows—an example might be conflict between quality of service settings for the support of enterprise telephony and

the burst capacity necessary to support long-distance high-performance data flows. In addition, the location pattern makes the application of the appropriate security pattern significantly easier (see Section 3.4).

The location design pattern can also significantly reduce the complexity of the portion of the network used for science applications. Troubleshooting is time-consuming, and there is a large difference in operational cost and time-to-resolution between verifying the correct operation of a small number of routers and switches and tracing the science flow through a large number of network devices in the general-purpose network of a college campus. For this reason, the Science DMZ is typically located as close to the network perimeter as possible, i.e. close to or directly connected to the border router that connects the research institution's network to the wide area science network.

3.2 Dedicated Systems: The Data Transfer Node (DTN)

Systems used for wide area science data transfers perform far better if they are purpose-built for and dedicated to this function. These systems, which we call data transfer nodes (DTNs), are typically PC-based Linux servers constructed with high quality components and configured specifically for wide area data transfer. The DTN also has access to storage resources, whether it is a local high-speed disk subsystem, a connection to a local storage infrastructure, such as a storage area network (SAN), or the direct mount of a high-speed parallel file system such as Lustre[4] or GPFS.[5] The DTN runs the software tools used for high-speed data transfer to remote systems. Some typical software packages include GridFTP[6] [3] and its service-oriented front-end Globus Online[7] [**?**], discipline-specific tools such as XRootD,[8] and versions of default toolsets such as SSH/SCP with high-performance patches[9] applied.

DTNs are widely applicable in diverse science environments. For example, DTNs are deployed to support Beamline 8.3.2 at Berkeley Lab's Advanced Light Source,[10] and as a means of transferring data to and from a departmental cluster. On a larger scale, sets of DTNs are deployed at supercomputer centers (for example at the DOE's Argonne Leadership Computing Facility,[11] the National Energy Research Scientific Computing Center,[12] and Oak Ridge Leadership Computing Facility[13]) to facilitate high-performance transfer of data both within the centers and to remote sites. At even larger scales, large clusters of DTNs provide data service to the Large Hadron Collider (LHC)[14] collaborations. The Tier-1[15] centers deploy large numbers of DTNs to support thousands of scientists. These are systems dedicated to the task of data transfers so that they provide reliable, high-performance service to science applications.[16]

DTNs typically have high-speed network interfaces, but the key is to match the DTN to the capabilities of the wide area network infrastructure. For example, if the network connection from the site to the WAN is 1 Gigabit Ethernet, a 10 Gigabit Ethernet interface on the DTN may be counter-productive. The reason for this is that a high-performance DTN can overwhelm the slower wide area link causing packet loss.

The set of applications that run on a DTN is typically limited to parallel data transfer applications like GridFTP or FDT.[17] In particular, user-agent applications associated with general-purpose computing and business productivity (e.g., email clients, document editors, media players) are not installed. This is for two reasons. First, the dedication of the DTN to data transfer applications produces more consistent behavior and avoids engineering trade-offs that might be part of supporting a larger application set. Second, data transfer applications are relatively simple from a network security perspective, and this makes the appropriate security policy easier to apply (see Section 3.4).

Because the design and tuning of a DTN can be time-consuming for small research groups, ESnet has a DTN Tuning guide[18] and a Reference DTN Implementation guide.[19] The typical engineering trade-offs between cost, redundancy, performance, and so on. apply when deciding on what hardware to use for a DTN. In general, it is recommended that DTNs be procured and deployed such that they can be expanded to meet future storage requirements.

3.3 Performance Monitoring

Performance monitoring is critical to the discovery and elimination of so-called "soft failures" in the network. Soft failures are problems that do not cause a complete failure that prevents data from flowing (like a fiber cut), but causes poor performance. Examples of soft failures include packet loss due to failing components; dirty fiber optics; routers forwarding packets using the management CPU rather than the high-performance forwarding hardware; and inadequate hardware configuration. Soft failures often go undetected for many months or longer, since most network management and error reporting systems are optimized for reporting "hard failures", such as loss of a link or device. Also, many scientists do not know what level of performance to expect, and so they do not know when to alert knowledgeable staff about a potential problem.

A perfSONAR host [16] helps with fault diagnosis on the Science DMZ. It offers end-to-end testing with collaborating sites that have perfSONAR tools installed, which allows for multi-domain troubleshooting. perfSONAR is a network monitoring software suite designed to conduct both active and passive network measurements, convert these to a standard format, and then publish the data so it is publicly accessible. The perfSONAR host can run continuous checks

[4] Lustre. `http//www.lustre.org/`.
[5] GPFS. `http://www.ibm.com/systems/software/gpfs/`.
[6] GridFTP. `http://www.globus.org/datagrid/gridftp.html`.
[7] Globus Online. `https://www.globusonline.org/`
[8] XRootD. `http://xrootd.slac.stanford.edu/`.
[9] HPN-SSH. `http://www.psc.edu/networking/projects/hpn-ssh/`.
[10] LBNL ALS. `http://www-als.lbl.gov`.
[11] ALCF. `https://www.alcf.anl.gov`.
[12] NERSC. `http://www.nersc.gov`.
[13] OLCF. `http://www.olcf.ornl.gov/`.
[14] LHC. `http://lhc.web.cern.ch/lhc/`.
[15] US/LHC. `http://www.uslhc.us/The_US_and_the_LHC/Computing`.
[16] LHCOPN. `http://lhcopn.web.cern.ch/lhcopn/`.
[17] FTD. `http://monalisa.cern.ch/FDT/`
[18] DTN Tuning. `http://fasterdata.es.net/science-dmz/DTN/tuning/`
[19] Reference DTN. `http://fasterdata.es.net/science-dmz/data-transfer-node-reference-implementation/`.

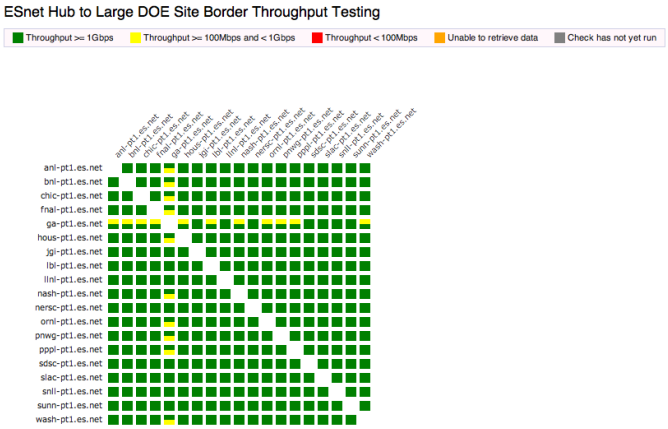

Figure 2: Regular perfSONAR monitoring of the ESnet infrastructure. The color scales denote the "degree" of throughput for the data path. Each square is halved to show the traffic rate in each direction between test hosts.

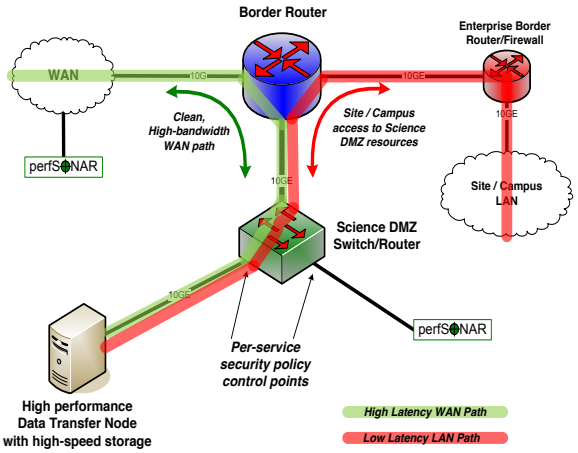

Figure 3: Example of the simple Science DMZ. Shows the data path through the border router and to the DTN (shown in green). The campus site access to the Science DMZ resources is shown in red.

for latency changes and packet loss using OWAMP,[20] as well as periodic "throughput" tests (a measure of available network bandwidth) using BWCTL.[21] If a problem arises that requires a network engineer to troubleshoot the routing and switching infrastructure, the tools necessary to work the problem are already deployed—they need not be installed before troubleshooting can begin.

By deploying a perfSONAR host as part of the Science DMZ architecture, regular active network testing can be used to alert network administers when packet loss rates increase, or throughput rates decrease. This is demonstrated by "dashboard" applications, as seen in Figure 2. Timely alerts and effective troubleshooting tools significantly reduce the time and effort required to isolate the problem and resolve it. This makes high performance the norm for science infrastructure, and provides significant productivity advantages for data-intensive science experiments.

3.4 Appropriate Security

Network and computer security are of critical importance for many organizations. Science infrastructures are no different than any other information infrastructure. They must be secured and defended. The National Institute for Standards and Technology (NIST) framework for security uses the CIA concepts—Confidentiality, Integrity, and Availability.[22] Data-intensive science adds another dimension—*performance*. If the science applications cannot achieve adequate performance, the science mission of the infrastructure has failed. Many of the tools in the traditional network security toolbox do not perform well enough for use in high-performance science environments. Rather than compromise security or compromise performance, the Science DMZ model addresses security using a multi-pronged approach.

The appropriate security pattern is heavily dependent on the location and the dedicated systems patterns. By deploying the Science DMZ in a separate location in the network topology, the traffic in the Science DMZ is separated from the traffic on the rest of the network (i.e., email, etc.), and security policy and tools can be applied specifically to the science-only traffic on the Science DMZ. The use of dedicated systems limits the application set deployed on the Science DMZ, and also reduces the attack surface.

A comprehensive network security capability uses many tools and technologies, including network and host intrusion detection systems, firewall appliances, flow analysis tools, host-based firewalls, router access control lists (ACLs), and other tools as needed. Appropriate security policies and enforcement mechanisms are designed based on the risk levels associated with high-performance science environments and built using components that scale to the data rates required without causing performance problems. Security for a data-intensive science environment can be tailored for the data transfer systems on the Science DMZ.

Science DMZ resources are designed to interact with external systems, and are isolated from (or have carefully managed access to) internal systems. This means the security policy for the Science DMZ can be tailored for this purpose. Users at the local site who access resources on their local Science DMZ through the lab or campus perimeter firewall will typically get reasonable performance, since the latency between the local users and the local Science DMZ is low (even if the firewall causes some loss), TCP can recover quickly.

4. SAMPLE DESIGNS

As a network design paradigm, the individual patterns of the Science DMZ can be combined in many different ways. The following examples of the overall Science DMZ model are presented as illustrations of the concepts using notional network diagrams of varying size and functionality.

4.1 Simple Science DMZ

A simple Science DMZ has several essential components. These include dedicated access to high-performance wide area networks, high-performance network equipment, DTNs, and monitoring infrastructure provided by perfSONAR.

[20] OWAMP. http://www.internet2.edu/performance/owamp/.
[21] BWCTL. http://www.internet2.edu/performance/bwctl/.
[22] FIPS-199. http://csrc.nist.gov/publications/PubsFIPS.html.

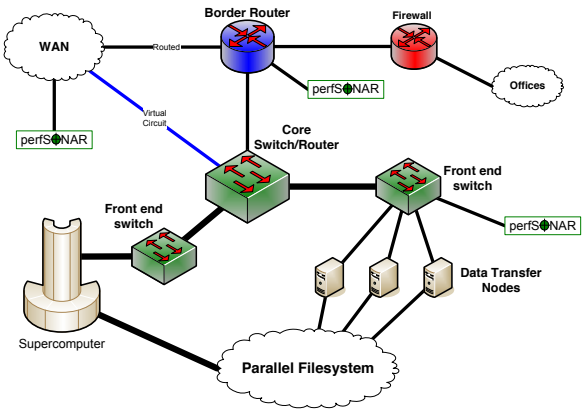

Figure 4: Example supercomputer center built as a Science DMZ.

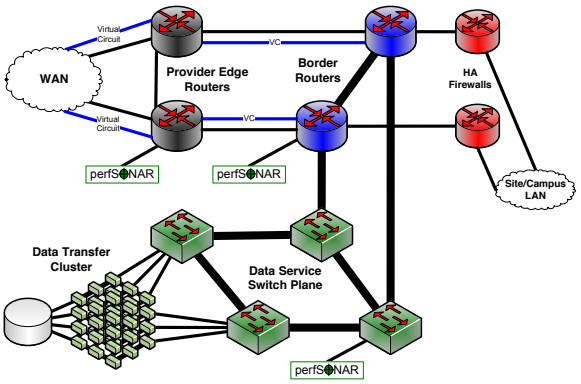

Figure 5: Example of an extreme data cluster. The wide area data path covers the entire network front-end, similar to the supercomputer center model.

These components are organized in an abstract diagram with data paths in Figure 3.

The DTN is connected directly to a high-performance Science DMZ switch or router, which is attached to the border router. By attaching the Science DMZ to the border router, it is much easier to guarantee a packet loss free path to the DTN, and to create virtual circuits that extend all the way to the end host. The DTN's job is to efficiently and effectively move science data between the local environment and remote sites and facilities. The security policy enforcement for the DTN is done using access control lists (ACLs) on the Science DMZ switch or router, not on a separate firewall. The ability to create a virtual circuit all the way to the host also provides an additional layer of security. This design is suitable for the deployment of DTNs that serve individual research projects or to support one particular science application. An example use case of the simple Science DMZ is discussed in Sections 6.1 and 6.2.

4.2 Supercomputer Center Network

The notional diagram shown in Figure 4 illustrates a simplified supercomputer center network. While this may not look much like the simple Science DMZ diagram in Figure 3, the same principles are used in its design.

Many supercomputer centers already use the Science DMZ model. Their networks are built to handle high-rate data flows without packet loss, and designed to allow easy troubleshooting and fault isolation. Test and measurement systems are integrated into the infrastructure from the beginning, so that problems can be located and resolved quickly, regardless of whether the local infrastructure is at fault. Note also that access to the parallel filesystem by wide area data transfers is via data transfer nodes that are dedicated to wide area data transfer tasks. When data sets are transferred to the DTN and written to the parallel filesystem, the data sets are immediately available on the supercomputer resources without the need for double-copying the data. Furthermore, all the advantages of a DTN—i.e., dedicated hosts, proper tools, and correct configuration—are preserved. This is also an advantage in that the login nodes for a supercomputer need not have their configurations modified to support wide area data transfers to the supercomputer itself. Data arrives from outside the center via the DTNs and is written to the central filesystem. The supercomputer login nodes do not need to replicate the DTN functionality in order to facilitate data ingestion. A use case is described in Section 6.4.

4.3 Big Data Site

For sites that handle very large data volumes (e.g., for large-scale experiments such as the LHC), individual data transfer nodes are not enough. These sites deploy data transfer "clusters", and these groups of machines serve data from multi-petabyte data storage systems. Still, the principles of the Science DMZ apply. Dedicated systems are still used for data transfer, and the path to the wide area is clean, simple, and easy to troubleshoot. Test and measurement systems are integrated in multiple locations to enable fault isolation. This network is similar to the supercomputer center example in that the wide area data path covers the entire network front-end, as shown in Figure 5.

This network has redundant connections to the research network backbone, each of which is capable of both routed IP and virtual circuit services. The enterprise portion of the network takes advantage of the high-capacity redundant infrastructure deployed for serving science data, and deploys redundant firewalls to ensure uptime. However, the science data flows do not traverse these devices. Appropriate security controls for the data service are implemented in the routing and switching plane. This is done both to keep the firewalls from causing performance problems and because the extremely high data rates are typically beyond the capacity of firewall hardware. More discussion about the LHC high-volume data infrastructure can be found in Johnston et al.'s paper presented at the 2013 TERENA Networking Conference [9].

5. NETWORK COMPONENTS

When choosing the network components for a Science DMZ, it is important to carefully select networking hardware that can efficiently handle the high bandwidth requirements. The most important factor is to deploy routers and switches that have enough queue buffer space to handle "fan-in" issues, and are properly configured to use this buffer space, as the default settings are often not optimized for bulk data transfers. (The fan-in issue is described in detail at the end of this section.) One should also look for devices that

have flexible, high-performance ACL (Access Control List) support so that the router or switch can provide adequate filtering to eliminate the need for firewall appliances. Note that some care must be taken in reading the documentation supplied by vendors. For example, Juniper Network's high-performance router ACLs are actually called "firewall filters" in the documentation and the device configuration. In general, it is important to ask vendors for specifics about packet filtering, interface queues, and other capabilities.

As discussed, two very common causes of packet loss are firewalls and aggregation devices with inadequate buffering. In order to understand these problems, it is important to remember that a TCP-based flow rarely runs at its overall "average" speed. When observed closely, it is apparent that most high-speed TCP flows are composed of bursts and pauses. These bursts are often very close to the maximum data rate for the sending host's interface. This is important, because it means that a 200 Mbps TCP flow between hosts with Gigabit Ethernet interfaces is actually composed of short bursts at or close to 1Gbps with pauses in between.

Firewalls are often built with an internal architecture that aggregates a set of lower-speed processors to achieve an aggregate throughput that is equal to the speed of the network interfaces of the firewall. This architecture works well when the traffic traversing the firewall is composed of a large number of low-speed flows (e.g., a typical business network traffic profile). However, this causes a problem when a host with a network interface that is faster than the firewall's internal processors emerges. Since the firewall must buffer the traffic bursts sent by the data transfer host until it can process all the packets in the burst, input buffer size is critical. Firewalls often have small input buffers because that is typically adequate for the traffic profile of a business network. If the firewall's input buffers are too small to hold the bursts from the science data transfer host, the user will suffer severe performance problems caused by packet loss.

Given all the problems with firewalls, one might ask what value they provide. If the application set is limited to data transfer applications running on a DTN, the answer is that firewalls provide very little value. When a firewall administrator is collecting the information necessary to allow a data transfer application such as GridFTP to traverse the firewall, the firewall administrator does not configure the firewall to use a specialized protocol analyzer that provides deep inspection of the application's traffic. The firewall administrator asks for the IP addresses of the communicating hosts, and the TCP ports that will be used by the hosts to communicate. Armed with that information, the firewall administrator configures the firewall to permit the traffic. Filtering based on IP address and TCP port number can be done on the Science DMZ switch or router with ACLs. When done with ACLs on a modern switch or router, the traffic does not need to traverse a firewall at all. This is a key point: by running a limited set of applications on the Science DMZ DTNs, the application profile is such that the Science DMZ can typically be defended well without incurring the performance penalties of a firewall. This is especially true if the ACLs are used in combination with intrusion detection systems or other advanced security tools. However, an intrusion detection system should be used even if a firewall is present.

Aggregation ("fan-in") problems are related to the firewall problem in that they too result from the combination of the burstiness of TCP traffic and small buffers on network devices. However, the fan-in problem arises when multiple traffic flows entering a switch or router from different ingress interfaces are destined for a common egress interface. If the speed of the sum of the bursts arriving at the switch is greater than the speed of the device's egress interface, the device must buffer the extra traffic or drop it. If the device does not have sufficient buffer space, it must drop some of the traffic, causing TCP performance problems. This situation is particularly common in inexpensive LAN switches. Since high-speed packet memory is expensive, cheap switches often do not have enough buffer space to handle anything except LAN traffic. Note that the fan-in problem is not unique to coincident bursts. If a burst from a single flow arrives at a rate greater than the rate available on the egress interface due to existing non-bursty traffic flows, the same problem exists.

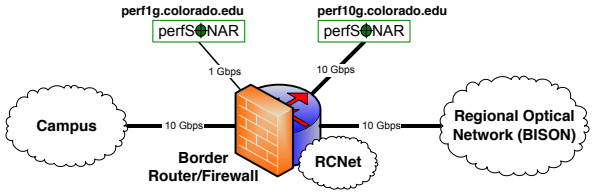

Figure 6: University of Colorado campus network, showing RCNet connected at the perimeter as a Science DMZ.

6. USE CASES

In this section we give some examples of how elements of the Science DMZ model have been put into practice. While a full implementation of recommendations is always encouraged, many factors influence what can and cannot be installed at a given location due to existing architectural limitations and policy. These use cases highlight the positive outcomes of the design methodology, and show that the Science DMZ model is able to please both administrative and scientific constituencies.

6.1 University of Colorado, Boulder

The University of Colorado, Boulder campus was an early adopter of Science DMZ technologies. Their core network features an immediate split into a protected campus infrastructure (beyond a firewall), as well as a research network (RCNet) that delivers unprotected functionality directly to campus consumers. Figure 6 shows the basic breakdown of this network, along with the placement of measurement tools provided by perfSONAR.

The physics department, a participant in the Compact Muon Solenoid (CMS)[23] experiment affiliated with the LHC project, is a heavy user of campus network resources. It is common to have multiple streams of traffic approaching an aggregate of 5 Gbps affiliated with this research group. As demand for resources increased, the physics group connected additional computation and storage to their local network. Figure 7 shows these additional 1 Gbps connections as they entered into the portion of the RCNet on campus.

Despite the initial care in the design of the network, overall performance began to suffer during heavy use times on

[23]CMS. `http://cms.web.cern.ch`.

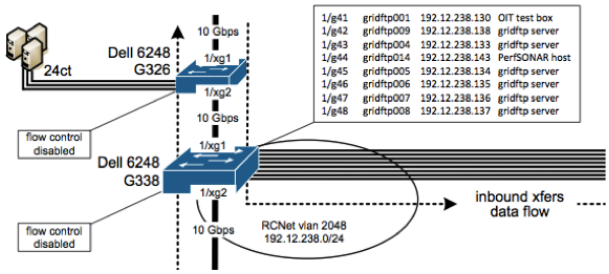

Figure 7: University of Colorado Network showing physics group connectivity.

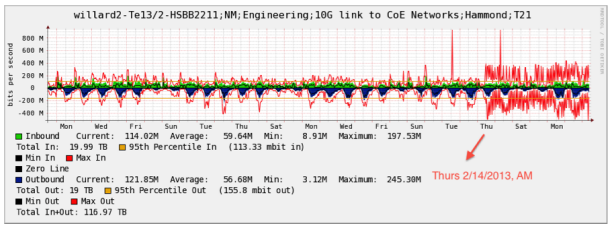

Figure 8: Penn State College of Engineering network utilization, collected passively from SNMP data

the campus. Passive and active perfSONAR monitoring alerted that there was low throughput to downstream facilities, as well as the presence of dropped packets on several network devices. Further investigation was able to correlate the dropped packets to three main factors:

- Increased number of connected hosts,
- Increased network demand per host,
- Lack of tunable memory on certain network devices in the path.

Replacement hardware was installed to alleviate this bottleneck in the network, but the problem remained upon initial observation. After additional investigation by the vendor and performance engineers, it was revealed that the unique operating environment (e.g., high "fan-out" that featured multiple 1Gbps connections feeding a single 10Gbps connection) was contributing to the problem. Under high load, the switch changed from cut-through mode to store-and-forward mode, and the cut-through switch was unable to provide loss-free service in store-and-forward mode.

After a fix was implemented by the vendor and additional changes to the architecture were implemented, performance returned to near line rate for each member of the physics computation cluster.

6.2 The Pennsylvania State University & Virginia Tech Transportation Institute

The Pennsylvania State University's College of Engineering (CoE) collaborates with many partners on jointly funded activities. The Virginia Tech Transportation Institute (VTTI), housed at Virginia Polytechnic Institute and State University, is one such partner. VTTI chooses to collocate computing and storage resources at Penn State, whose network security and management is implemented by local staff. However, due to policy requirements for collocated equipment, a security mechanism in the form of a firewall was required to protect both the campus and VTTI equipment. Shortly after collocation, VTTI users noticed that performance for hosts connected by 1Gbps local connections were limited to around 50Mbps overall; this observation was true in either direction of data flow.

Using perfSONAR, network engineers discovered that the size of the TCP window was not growing beyond the default value of 64KB, despite the fact that hosts involved in data transfer and measurement testing were configured to use auto-tuning—a mechanism that would allow this value to grow as time, capacity, and demand dictated. To find the correct window size needed to achieve network speeds close to 1Gbps, the sites were measured at 10 ms away in terms of round-trip latency, which yielded a window size of:

$$\frac{1000\text{Mb/s}}{8\text{B}/b} * 10\text{ms} * \frac{1\text{s}}{1000\text{ms}} = 1.25\text{MB}. \qquad (2)$$

This theoretical value was 20 times less than the required size. Further investigation into the behavior of the network revealed that there was no packet loss observed along the path, and other perfSONAR test servers on campus showed performance to VTTI that exceeded 900Mbps. From some continued performance monitoring, the investigation began to center on the performance of the CoE firewall.

A review of the firewall configuration revealed that a setting on the firewall, *TCP flow sequence checking*, modifies the TCP header field that specifies window size (e.g., a clear violation of tcp_window_scaling, set forth in RFC 1323 [**?**]). Disabling this firewall setting increased inbound performance by nearly 5 times, and outbound performance by close to 12 times the original observations. Figure 8 is a capture of overall network utilization to CoE, and shows an immediate increase in performance after the change to the firewall setting.

Because CoE and VTTI were able to utilize the Science DMZ resources, like perfSONAR, engineers were able to locate and resolve the major network performance problem. Figure 8 also shows that numerous users, not just VTTI, were impacted by this abnormality. The alteration in behavior allowed TCP to reach higher levels of throughput, and allowed flows to complete in a shorter time than with a limited window.

6.3 The National Oceanic and Atmospheric Administration

The National Oceanic and Atmospheric Administration (NOAA) in Boulder houses the Earth System Research Lab, which supports a "reforecasting" project. The initiative involves running several decades of historical weather forecasts with the same current version of NOAA's Global Ensemble Forecast System (GEFS). Among the advantages associated with a long reforecast data set, model forecast errors can be diagnosed from the past forecasts and corrected, thereby dramatically increasing the forecast skill, especially in forecasts of relatively rare events and longer-lead forecast.

In 2010, the NOAA team received an allocation of 14.5 million processor hours at NERSC to perform this work. In all, the 1984–2012 historical GEFS dataset totaled over 800 TB, stored on the NERSC HPSS archival system. Of the 800TB at NERSC, the NOAA team sought to bring about 170TB back to NOAA Boulder for further processing and to make it more readily available to other researchers. When

the NOAA team tried to use an FTP server located behind NOAA's firewall for the transfers, they discovered that data trickled in at about 1–2MB/s.

Working with ESnet and NERSC, the NOAA team leveraged the Science DMZ design pattern to set up a new dedicated transfer node enabled with Globus Online to create a data path unencumbered by legacy firewalls. Immediately the team saw a throughput increase of nearly 200 times. The team was able to transfer 273 files with a total size of 239.5GB to the NOAA DTN in just over 10 minutes—approximately 395MB/s.

6.4 National Energy Research Scientific Computing Center

In 2009, both NERSC and OLCF installed DTNs to enable researchers who use their computing resources to move large data sets between each facility's mass storage systems. As a result, WAN transfers between NERSC and OLCF increased by at least a factor of 20 for many collaborations. As an example, a computational scientist in the OLCF Scientific Computing Group who was researching the fundamental nuclear properties of carbon-14, in collaboration with scientists from Lawrence Livermore National Laboratory (LLNL) and Iowa State University, had previously waited more than an entire workday for a single 33 GB input file to transfer—just one of the 20 files of similar size that needed to be moved between the sites. With the improved infrastructure, those researchers were immediately able to improve their transfer rate to 200 MB/sec enabling them to move all 40 TB of data between NERSC and OLCF in less than three days.

Since 2009, several science collaborations including those in astrophysics, climate, photon science, genomics and others have benefitted from the Science DMZ architecture at NERSC. Most recently, it has enabled high-speed multi-terabyte transfers between SLAC Linear Accelerator National Lab's Linac Coherent Light Source and NERSC to support protein crystallography experiments as well as transfers between Beamline 8.3.2 at Berkeley Lab's Advanced Light Source and NERSC in support of X-ray tomography experiments.

7. FUTURE TECHNOLOGIES

In addition to solving today's network performance problems, the Science DMZ model also makes it easier to experiment and integrate with tomorrow's technologies. Technologies such as dynamic "virtual circuits", software-defined networking (SDN), and 40/100Gbps ethernet can be deployed in the Science DMZ, eliminating the need to deploy these technologies deep inside campus infrastructure.

7.1 Virtual Circuits

Virtual circuit services, such as the ESnet-developed On-demand Secure Circuits and Reservation System, or OSCARS platform [7, 13], can be used to connect wide area layer-2 circuits directly to DTNs, allowing the DTNs to receive the benefits of the bandwidth reservation, quality of service guarantees, and traffic engineering capabilities. The campus or lab "inter-domain" controller (IDC)[24] can provision the local switch and initiate multi-domain wide area virtual circuit connectivity to provide guaranteed bandwidth between DTN's at multiple institutions. An example of this configuration is the NSF-funded Development of Dynamic Network System (DYNES) [**?**] project that is supporting a deployment of approximately 60 university campuses and regional networks across the US. Virtual circuits also enable the use of new data transfer protocols such as RDMA (remote direct memory access) over Converge Ethernet (RoCE) [4] on the Science DMZ DTNs. RoCE has been demonstrated to work well over a wide area network, but only on a guaranteed bandwidth virtual circuit with minimal competing traffic [11]. Kissel et al. show that RoCE can achieve the same performance as TCP (39.5Gbps for a single flow on a 40GE host), but with 50 times less CPU utilization.

[24]IDC, `http://www.controlplane.net/`

7.2 100-Gigabit Ethernet

100 Gigabit Ethernet (GE) technology is being deployed by research networks around the world, to support data-intensive science. The NSF CC-NIE program is increasing the rate of 100GE deployment at US campuses with solicitations offered in 2012 and 2013. While 100GE promises the ability to support next-generation instruments and facilities, and to conduct scientific analysis of distributed data sets at unprecedented scale, 100GE technology poses significant challenges for the general-purpose networks at research institutions. Once a site is connected to a 100GE backbone, it would be very costly to distribute this new increased bandwidth across internal campus infrastructure. With the Science DMZ model, all hosts needing the increased bandwidth are near the border router, making it much easier to benefit from the 100GE connection.

7.3 Software-Defined Networking

Testing and deploying software defined networking, particularly the use of OpenFlow as a platform [**?**], is a timely example of how the Science DMZ model could be used for exploring and hardening new technologies.

Software-defined networking concepts and production uses of OpenFlow are still in their early stages of adoption by the community. Many innovative approaches are still being investigated to develop best practices for the deployment and integration of these services in production environments. ESnet and its collaborators at Indiana University have demonstrated an OpenFlow-based Science DMZ architecture that interoperates with a virtual circuit service like OSCARS. It is easy to set up an OSCARS virtual circuit across the WAN, but plumbing the circuit all the way to the end host must be done by hand. OpenFlow simplifies this process.

Another promising use of OpenFlow is as a mechanism to dynamically modify the security policy for large flows between trusted sites. Multiple groups have demonstrated the use of OpenFlow to dynamically bypass the firewall (e.g., Kissel et al.'s research on SDN with XSP [10]). Further, one could also use OpenFlow along with an intrusion detection system (IDS) to send the connection setup traffic to the IDS for analysis, and then once the connection is verified allow the flow to bypass the firewall and the IDS.

8. CONCLUSION

The Science DMZ model has its roots in operational practices developed over years of experience, and incorporates aspects of network architecture, network security, performance tuning, system design, and application selection. The Sci-

ence DMZ, as a design pattern, has already been successfully deployed at multiple sites across the US, and many through NSF funding. The Science DMZ model and its contributing technologies are well-tested and have been effectively used at supercomputer centers, national laboratories, and universities as well as in large-scale scientific experiments.

The Science DMZ model provides a conceptual framework for the deployment of networks and network-enabled tools and systems for the effective support of data-intensive science. With many science collaborations moving to large-scale or distributed experiments, the purpose of sharing best practices is becoming more important. This paper shares our work in developing the Science DMZ for the larger science community.

9. ACKNOWLEDGMENTS

The authors would like to thank NOAA, NERSC, the Pennsylvania State University, and the University of Colorado, Boulder, for their contributions to this work.

The authors wish to acknowledge the vision of the National Science Foundation for its support of the CC-NIE program.

10. DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

11. REFERENCES

[1] C. Alexander. *The Timeless Way of Building*. Oxford University Press, New York, 1979.

[2] C. Alexander, S. Ishikawa, and M. Silverstein. *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, New York, August 1977.

[3] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster. The Globus Striped GridFTP Framework and Server. In *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, SC '05, page 54, Washington, DC, USA, 2005. IEEE Computer Society.

[4] I. T. Association. InfiniBand. Architecture Specification Release 1.2.1 Annex A16: RoCE, 2010.

[5] E. Dart. Optimizing data management at the advanced light source with a science dmz. `http://www.globusworld.org/program`, 2012.

[6] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[7] C. Guok, D. Robertson, M. Thompson, J. Lee, B. Tierney, and W. Johnston. Intra and Interdomain Circuit Provisioning Using the OSCARS Reservation System. In *Third International Conference on Broadband Communications Networks and Systems, IEEE/ICST*, October 2006.

[8] V. Jacobson, R. Braden, and D. Borman. TCP Extensions for High Performance, 1992.

[9] W. E. Johnston, E. Dart, M. Ernst, and B. Tierney. Enabling high throughput in widely distributed data management and analysis systems: Lessons from the LHC. In *TERENA Networking Conference (TNC) 2013*, June 2013.

[10] E. Kissel, G. Fernandes, M. Jaffee, M. Swany, and M. Zhang. Driving software defined networks with xsp. In *Workshop on Software Defined Networks (SDN'12), International Conference on Communications (ICC)*. IEEE, June 2012.

[11] E. Kissel, B. Tierney, M. Swany, and E. Pouyoul. Efficient Data Transfer Protocols for Big Data. In *Proceedings of the 8th International Conference on eScience*. IEEE, July 2012.

[12] M. Mathis, J. Semke, J. Mahdavi, and T. Ott. The macroscopic behavior of the tcp congestion avoidance algorithm. *SIGCOMM Comput. Commun. Rev.*, 27(3):67–82, July 1997.

[13] I. Monga, C. Guok, W. E. Johnston, and B. Tierney. Hybrid Networks: Lessons Learned and Future Challenges Based on ESnet4 Experience. In *IEEE Communications Magazine*, May 2011.

[14] perfSONAR Consortium. perfSONAR Technical Overview. `http://www.perfsonar.net/overview.html`, June 2009.

[15] J. Postel. Transmission Control Protocol. Request for Comments (Standard) 793, Internet Engineering Task Force, September 1981.

[16] B. Tierney, J. Boote, E. Boyd, A. Brown, M. Grigoriev, J. Metzger, M. Swany, M. Zekauskas, and J. Zurawski. perfSONAR: Instantiating a Global Network Measurement Framework. In *SOSP Workshop on Real Overlays and Distributed Systems (ROADS '09)*, Big Sky, Montana, USA, Oct. 2009. ACM.