Journal of the American Medical Informatics Association, 25(3), 2018, 267–274
doi: 10.1093/jamia/ocx104
Advance Access Publication Date: 6 October 2017
Research and Applications

Research and Applications

The medical science DMZ: a network design pattern for data-intensive medical science

Sean Peisert,[1,2,3] Eli Dart,[4] William Barnett,[5] Edward Balas,[6] James Cuff,[7] Robert L Grossman,[8] Ari Berman,[9] Anurag Shankar,[10] and Brian Tierney[4]

[1]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [2]Department of Computer Science, University of California Davis, Davis, CA, USA, [3]Corporation for Education Network Initiatives in California (CENIC), Berkeley, CA, USA, [4]ESnet, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [5]Indiana Clinical and Translational Sciences Institute and Regenstrief Institute, Indiana University, Indianapolis, IN, USA, [6]Global Research Network Operations Center, Indiana University, Bloomington, IN, USA, [7]Research Computing, Harvard University, Cambridge, MA, USA, [8]Center for Data Intensive Science, University of Chicago, Chicago, USA, [9]BioTeam, Middleton, MA, USA and [10]Pervasive Technology Institute, Indiana University, Bloomington, IN, USA

Corresponding Author: Sean Peisert, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA; Corporation for Education Network Initiatives in California (CENIC), Berkeley, CA 94710, USA; or Department of Computer Science, University of California, Davis, One Shields Ave., Davis, CA 95656, USA. E-mail: sppeisert@lbl.gov. Phone: (510) 486-4706

Received 28 April 2017; Revised 3 August 2017; Editorial Decision 24 August 2017; Accepted 30 August 2017

ABSTRACT

Objective: We describe a detailed solution for maintaining high-capacity, data-intensive network flows (eg, 10, 40, 100 Gbps+) in a scientific, medical context while still adhering to security and privacy laws and regulations.
Materials and Methods: High-end networking, packet-filter firewalls, network intrusion-detection systems.
Results: We describe a "Medical Science DMZ" concept as an option for secure, high-volume transport of large, sensitive datasets between research institutions over national research networks, and give 3 detailed descriptions of implemented Medical Science DMZs.
Discussion: The exponentially increasing amounts of "omics" data, high-quality imaging, and other rapidly growing clinical datasets have resulted in the rise of biomedical research "Big Data." The storage, analysis, and network resources required to process these data and integrate them into patient diagnoses and treatments have grown to scales that strain the capabilities of academic health centers. Some data are not generated locally and cannot be sustained locally, and shared data repositories such as those provided by the National Library of Medicine, the National Cancer Institute, and international partners such as the European Bioinformatics Institute are rapidly growing. The ability to store and compute using these data must therefore be addressed by a combination of local, national, and industry resources that exchange large datasets. Maintaining data-intensive flows that comply with the Health Insurance Portability and Accountability Act (HIPAA) and other regulations presents a new challenge for biomedical research. We describe a strategy that marries performance and security by borrowing from and redefining the concept of a Science DMZ, a framework that is used in physical sciences and engineering research to manage high-capacity data flows.
Conclusion: By implementing a Medical Science DMZ architecture, biomedical researchers can leverage the scale provided by high-performance computer and cloud storage facilities and national high-speed research networks while preserving privacy and meeting regulatory requirements.

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

Key words: computer communication networks, data-intensive science, high-performance computing, biomedical research, computer security, Health Insurance Portability and Accountability Act

INTRODUCTION

"Big Data"[1] now plays as significant a role in medical science as it does in other facets of modern life. However, storage, computation, and transfer needs to process the data are growing rapidly in medical schools, outstripping the capacity of on-premise IT resources. Two basic options are available to address this problem. The first is cloud computing using public "clouds" like Amazon and Google, "secure" clouds offered by telecommunication firms and other companies, or on-premise private clouds. Precision medicine will require participation in a national federation of interlinked data repositories and high-performance computing (HPC), cloud computing, and storage facilities that will serve biomedical researchers and ultimately care providers. Data generated by increasingly high-throughput and increasingly distributed sequencers and imaging facilities will need to be integrated with rapidly expanding national repositories of reference data such as The Cancer Genome Atlas. Any precision medicine effort will need to combine locally managed data, distributed reference data, and local and national computational services.

The National Institutes of Health are spearheading a "commons" initiative for data sharing, and have long provided reference data through the National Library of Medicine. The National Cancer Institute is exploring this option by funding a number of cloud "pilots" for cancer genomics.[2] National HPC facilities available at many academic institutions[3] are applying their capacity to biomedical research. These efforts are interconnected by high-capacity research networks such as Internet2, ESnet, and the Corporation for Education Network Initiatives in California. These networks are part of the so-called research and education (R&E) network ecosystem, which provides high-performance networks designed specifically for large-scale science and engineering data to interconnect research institutions globally. Such resources have traditionally been leveraged for applications at scale, such as high-energy physics research (eg, the Large Hadron Collider experiments, which use the Open Science Grid[4]), astronomy, climate modeling, and other "big science" initiatives that compute at the petaFLOPS scale. However, implementing large-scale computing and data storage for medical applications presents a number of challenges for academic medical centers, particularly security and regulatory compliance. Protecting patient privacy has not, however, traditionally been part of the equation in high-performance computing. Many organizations, such as the Coalition for Advanced Scientific Computing, are helping HPC centers meet Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH) requirements in response to this need.

In order for precision medicine and other Big Data health care research strategies to be successful, there must be a national strategy for the secure transfer of patient data at scale. Many organizations are now working to meet HIPAA and HITECH requirements for their systems in response to this need. A de facto technical control in environments subject to regulations such as the HIPAA Security Rule[5] is to employ commercial firewalls. However, a significant tension exists between the standards that reference firewalls for sensitive data[6] and the performance and throughput requirements needed for data-intensive science. Specifically, a very small number of dropped packets, due to stateful and/or deep packet–inspecting firewalls, can lead to a severe degradation in network throughput.[7,8] The Science DMZ model is used in many scientific environments to solve performance problems for data-intensive science.

A Science DMZ is a portion of the network, built at or near the local network perimeter of an individual research institution, that is designed such that the equipment, configuration, and security policies are optimized for high-performance workflows and large datasets.[7,8] A Science DMZ is typically connected to an R&E network at high speed to allow the resources in the local Science DMZ to connect to other Science DMZs with the performance necessary to support large-scale data-intensive science. The basic Science DMZ model has been successfully implemented in numerous scenarios, including those involving astrophysics, photon science, high-energy physics, materials science, climate modeling, and genomics.[7,8] These efforts have been notably recognized by the National Science Foundation, which has awarded multiple rounds of funding to US academic institutions (as part of the Campus Cyberinfrastructure program[9]) to construct Science DMZ environments on their campuses to support research at scale.

The Science DMZ architecture also maintains the security of the data through a number of distinct techniques, but does not employ commercial firewalls due to their negative impact on performance. As a result, the Science DMZ model is not currently employed in environments subject to the HIPAA Security Rule[5] and HITECH requirements, due to the presumed technical controls based on de facto use of stateful and deep packet–inspecting commercial firewalls.[5] We believe that this problem has a solution, however.

We have taken a central of tenet of the Science DMZ[7,8] and reengineered it for restricted data as a Medical Science DMZ.[10] Science DMZs operate at scale using already-provisioned software and authentication stack as well as mature services at each site. Creating a high-capacity, secure, data-intensive enclave within each research institution and at major data repositories allows scientists across the country to securely move datasets at scale to the appropriate computational resources based on the trust relationships that govern each science collaboration. This provides the ability to compute on the data at scales previously reserved for much larger physical sciences and engineering problems, but at much lower cost and with much less effort than using commercial clouds.

While HIPAA defines and mandates certain safeguards, it allows latitude in addressing those safeguards. More importantly, it shifts the focus to risk-centric, as opposed to control-centric, practices. This approach to security is more nuanced and includes factors such as cost, likelihood of exploitation, impact, etc. To reflect this philosophy, we have defined a Medical Science DMZ as a method or approach that allows data flows at scale while simultaneously addressing the HIPAA Security Rule and related regulations governing biomedical data and appropriately managing risk. We emphasize use cases that involve scientists transferring and processing medical research data that have very different requirements than those of medical centers communicating with suppliers, service providers, and employees. Our network design pattern addresses Big Data and can be implemented using a combination of physical, administrative, and technical safeguards.

In this paper, we describe details of 3 implementations and how they balance the key aspects of a Medical Science DMZ of high-throughput and regulatory compliance.

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

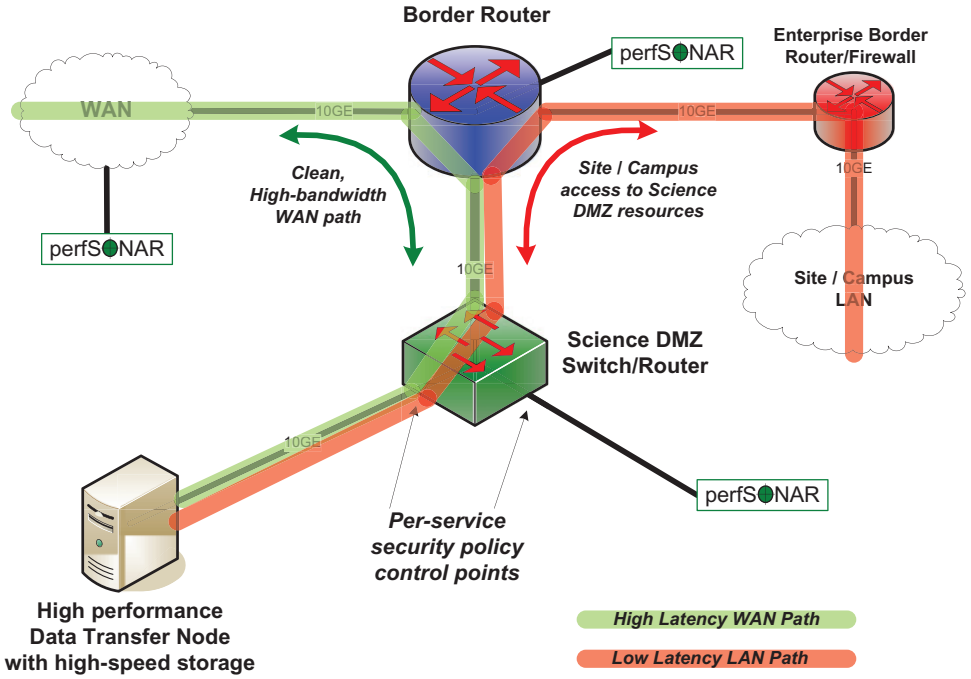

Figure 1. Classical Science DMZ.

BACKGROUND

According to National Institute of Standards and Technology (NIST) publication 800-41, firewalls are "devices or programs that control the flow of network traffic between networks or hosts that employ differing security postures."[11] NIST 800-41 defines multiple different types of firewalls, including:

- Packet-filtering firewalls that use attributes of the packet headers, such as destination addresses, source addresses, and other options, as the basis for their access control decisions.
- Stateful firewalls that implement their own protocol state machines and track the connection state in the same way the end hosts do, and are thus able to detect protocol-level anomalies and other threats that a simple packet filter cannot.
- Application-layer firewalls that examine the contents of the packets and messages and grants or denies access based on inferred application state (eg, by detecting malicious web content destined for a web browser).

While NIST 800-41 has a rich and nuanced view of the breadth of firewall types and capabilities, the commercial marketplace only defines stateful firewalls and application firewalls as "firewalls." Whereas a packet-filtering router is not considered to be a firewall by many commercial providers, the standards body considered authoritative in matters of US government policy, NIST, does consider a packet-filtering router to be a firewall, albeit a simple one.

From the perspective of NIST 800-41, a Science DMZ uses a nonstateful packet-filter firewall that is implemented in the gateway or a downstream router (this is how NIST 800-41 defines the packet-filtering router that is a key component of a standard Science DMZ). The packet enters the firewall, and its source and destination addresses are compared to a list of rules. If it matches any of those rules, the action associated with the rule (forward or discard) takes place. In addition, other compensating controls are often employed in a Science DMZ, such as the use of an intrusion detection system (IDS) (eg, the Bro system[12,13]) or an intrusion prevention system (IPS) (eg, Snort[14]). A capable Science DMZ router (again, called a firewall by NIST 800-41) can usually be configured to copy every packet it receives and send that to an IDS. The IDS analyzes the packets and, based on the result of its analysis, can take action to block or otherwise interfere with any hostile traffic.

CLASSICAL SCIENCE DMZ

In a classical Science DMZ, shown in Figure 1, a network enclave is constructed using high-performance equipment (typically one or more switches/routers) at or near the institutional network perimeter. Because it is at the network perimeter, the resources in the Science DMZ have ready high-performance access to the global R&E network infrastructure and therefore have high-performance access to the resources in other Science DMZs, so long as security policies and trust relationships permit such access. High-performance servers, called data transfer nodes (DTNs), are connected directly to these high-performance routers in the Science DMZ. The DTNs handle all data ingest/export tasks, so the DTNs are the focus of security policy for the Science DMZ (other than the protection of the Science DMZ infrastructure itself). The security controls for the DTN are implemented by the router to which the DTN is directly connected. Additional layers of security are typically implemented as well: the DTN typically runs host-based firewalls or IDS packages; a network IDS such as Bro is often employed; and the set of applications running on the DTN is strictly limited to system maintenance and data ingest/export tasks. This limitation of applications on the DTN is a critical point – it dramatically reduces the network-visible attack surface, and it makes the DTN a better fit for risk-based security controls that can be implemented using high-performance technologies (and, in particular, without commercial stateful or application firewalls).

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

So designed, a Science DMZ is resistant to a wide variety of attacks. If the data-transfer tools deployed on the DTNs implement in-flight data encryption for all transfers, the data are not accessible to adversaries that might snoop on the communication between the Science DMZs that share a trust relationship. The stateless firewall (implemented by the Science DMZ switch or router) controls which DTNs exchange data, limiting the scope of data exfiltration. The IDS monitors both for policy infractions and for incoming hostile activity. The limitation of the application set on the DTNs limits the ways in which an attacker can compromise the system (and then only from an external host that is permitted by the stateless firewall capabilities in the Science DMZ). All of this can be done in a way that preserves the high-performance data-transfer capabilities necessary for effective collaboration in the era of Big Data.

If confidentiality is required (eg, encryption in flight or encryption at rest), these policies can be implemented and enforced on the DTN. In addition, if an IDS is employed, it can monitor the DTN traffic to ensure that the policies are in fact being followed. This defense in depth can serve as an important cross-check for DTN configuration changes, and is especially powerful in operational environments where the IDS policies are not routinely modified at the same time as the DTN configuration.

The flexibility of the Science DMZ model allows for multiple sub-enclaves within an institution, each with its own risk profile, security policies, compensating controls, etc. This segmentation of risk and the ability to apply sets of specific controls to sub-enclaves as required make the Science DMZ model applicable to a wide variety of network designs, threat models, and data protection requirements. The following case studies describe the addition of capabilities to the classical Science DMZ, enhancing it for use in environments with protected data.

MEDICAL SCIENCE DMZ ARCHITECTURES

We reiterate that the focus of the Medical Science DMZ design is distinct from that of patient-centric medical center networks. The latter, for example, is well advised to employ techniques such as IPSs and virtual private networks (VPNs) that are not currently capable of scaling to the data volumes associated with data-intensive medical research. For example, the University of Chicago has 1500 users per day accessing data from the Genomic Data Commons at 10–20 Gbps (in aggregate, sometimes higher) over a Medical Science DMZ. All sensitive data flows are encrypted. The University of Chicago (for example) cannot handle this use case with VPNs. In contrast, a medical center uses VPNs so it can communicate with its suppliers, service providers, and employees, a starkly different use case.

As such, the goal of the Science DMZ, including the Medical Science DMZ, is to enable high-performance transfer of data at scale while maintaining adequate security. We note that safeguards such as IPSs can still be implemented institutionally to provide additional controls.

In this section, we describe 3 different approaches to building a Medical Science DMZ. Indiana University, Harvard University, and the University of Chicago all use a non-firewalled approach to HIPAA in their Medical Science DMZs. Each has implemented a framework that allows free flow of data where needed and addresses HIPAA using alternate, reasonable, and appropriate controls that manage the risk posed by the absence of stateful or application-layer firewalls. To that end, each organization has implemented a specific "risk-managed" DMZ solution that encompasses the entire high-performance computing, storage, and network infrastructure. In the following subsections, we describe these architectures in detail.

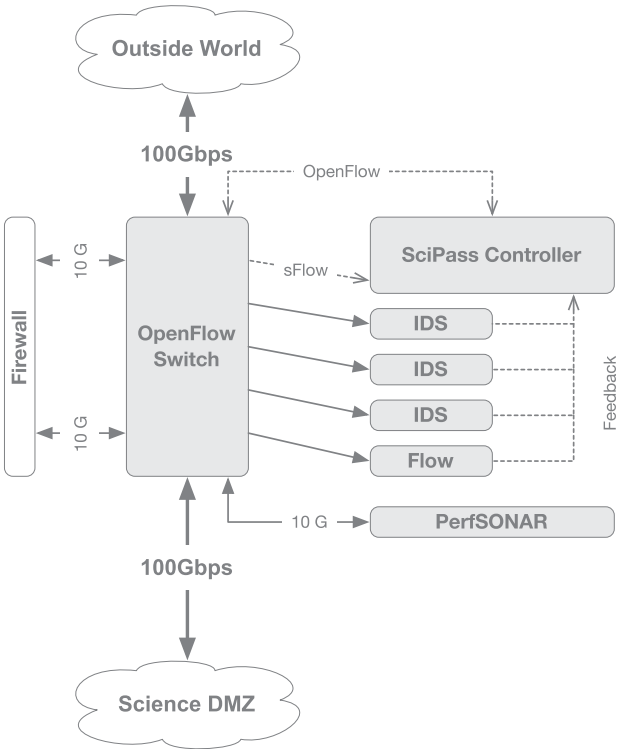

Figure 2. The SciPass system contains 6 components: an OpenFlow switch,[19] the SciPass controller, a cluster of IDS sensors, a PerfSONAR host,[20] a firewall, a network flow analysis system, and a DTN. The integration of these 6 components is designed to provide a secure, performant DMZ with a detailed history of all DMZ activity and a suite of tools to troubleshoot performance problems when they inevitably arise. The design also enhances network security with duty separation. In operation, the SciPass system can be administered by one set of individuals, with another set responsible for the individual DTNs. Both the DTNs and the SciPass switch would provide firewalling functions, with SciPass having enhanced temporal and address granularity.

The Indiana University SciPass Science DMZ

Indiana University's GlobalNOC[15] has a holistic, technical architecture under development called SciPass,[16] shown in Figure 2, that leverages a comprehensive NIST-based risk management framework[17,18] to support high-rate data flows that comply with regulations such as HIPAA.

Today, most IDS and flow sensors are unable to process 100 Gbps on a single server. To support aggregate traffic rates greater than single-sensor capacity, SciPass provides a load-balancing mechanism to allow for the use of an array of sensors, with traffic balanced in such a way as to ensure that packets from the same flow are examined by the same sensor. This is illustrated in Figure 3.

The SciPass design retains existing firewall infrastructure components as a technical control mechanism for all traffic in and out of the DMZ. Support for large data transfers, which have been shown to be a problem for most firewalls, is provided through the use of dynamic per-flow bypass of the firewall for known good data transfers. Bypassing is accomplished by reconfiguring the forwarding path on the OpenFlow Switch[16] to forward around the firewall and IDS after the flow is deemed acceptable via IDS examination, as

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

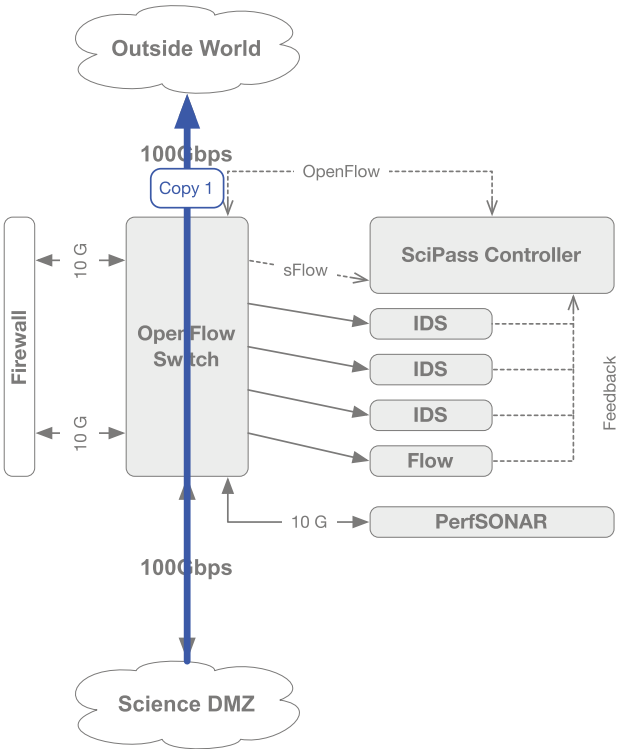

Figure 3. By default, traffic is forwarded through the institutional firewall via the OpenFlow switch. As this happens, copies of packets are sent to the array of IDS and flow analysis system sensors. SciPass uses a balancing mechanism that ensures that all packets for a given flow go to the same sensor for stream reassembly, and that flows are distributed as evenly as possible across the array of sensors. Using this approach allows for monitoring of individual 100 Gbps network connections using an array of 1 or 10 Gbps–capable sensors. SciPass supports the notion of sensor groups, where there can be a group of 3 different arrays of sensors that all need a copy of the same packet; this allows, for instance, the running of multiple types of IDS and flow analysis sensors. In testing and campus deployments, a combination of Bro, Snort, and Argus has been used. As copies of each packet are sent to an IDS, another copy is sent to a flow analysis system, which records summary records of all traffic in and out of the DMZ. These summaries are analogous to the level of detail contained in a typical residential phone bill and include the time of the transfer, the source and destination, the application used, and the volume data. They also contain information about packet loss or other performance problems observed in the flow. They do not, however, contain any of the data transmitted. These records are used for security and performance analysis purposes. On the performance side, the Argus utility has been used within the SciPass system to provide nonsampled flow accounting. For Transmission Control Protocol flows, it is able to passively detect performance impairments such as improperly tuned end hosts and packet loss. This makes it possible for DMZ operators to proactively detect performance issues without having to rely solely on end user reports and active testing of network performance. When performance problems are identified, the PerfSONAR utility is used to actively test the network path to help determine if the problems are network- or end system–based.

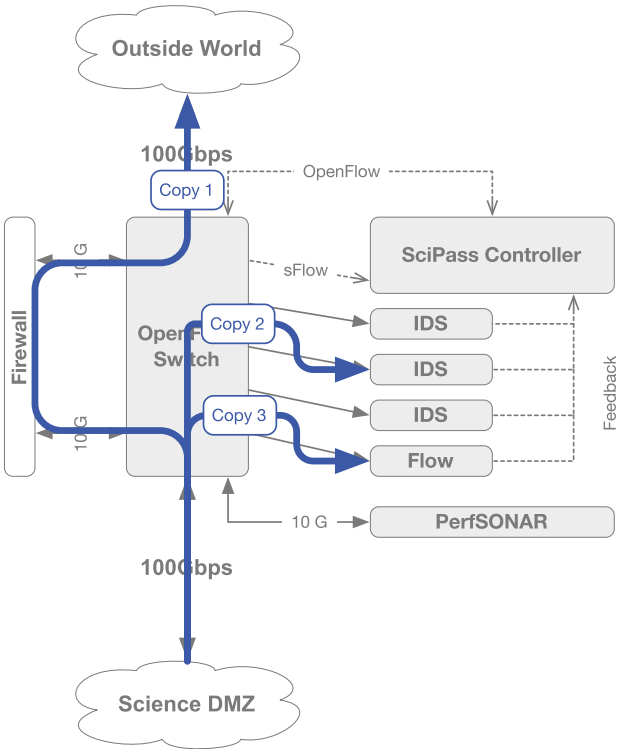

Figure 4. At the point that the SciPass system determines a flow is trustworthy, it will reprogram switch forwarding to bypass the firewall and IDS systems to reduce operational cost and improve network efficiency. For larger data transfers, this technique provides performance unconstrained by the firewall's limitations.

shown in Figure 4. This technique provides substantially improved end-to-end performance and reduces infrastructure costs by not having the IDS and firewall examine large volumes of traffic known to be uninteresting.

SciPass relies on IDS policies to identify "good" flows. These policies can contain a combination of time of day and day of the week, source and destination IP address, along with protocol and application-layer data to determine whether a flow should bypass an institutional firewall. For a user who uploads HIPAA data to the same facility across the country every Friday from a local DTN, SciPass could be configured to only bypass the firewall when transfers are made from a specific directory to a remote facility on Friday between 2 and 8 a.m. The policy allows users, network administrators, and security administrators to jointly define and enforce desired network behavior.

When the system determines that it is appropriate to route an individual flow around the firewall, a pair of higher-priority OpenFlow rules is added to the switch so that packets associated with the flow are directly forwarded, bypassing the institutional firewall and the IDS array. These rules contain an idle timeout to purge the rules once the flow completes. In essence, the system performs the same state tracking that a firewall does, and in appropriate situations optimizes the firewall out of the forwarding path. By doing so, operators reduce cost by using a lower-capacity firewall – for instance, one capable of supporting only 1–10Gbps transfers – while data transfers to or from the local DTN approach 40–100 Gbps.

SciPass is agnostic to the transfer protocols and local security measures employed on the DTN, and the model today places the responsibility for file and session encryption entirely on the DTN. For restricted data that require the use of file or session encryption, SciPass, and in particular the IDS, needs to be aware of what applications are permitted. However, SciPass can assist with the detection of privacy policy violations by adding Honeytokens or known bogus test records to the datasets and instructing the IDS to look for these patterns absent from conforming data transfers.

Future work at Indiana University will focus on how to effectively address HIPAA needs through the application and evolution of the SciPass architecture.

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

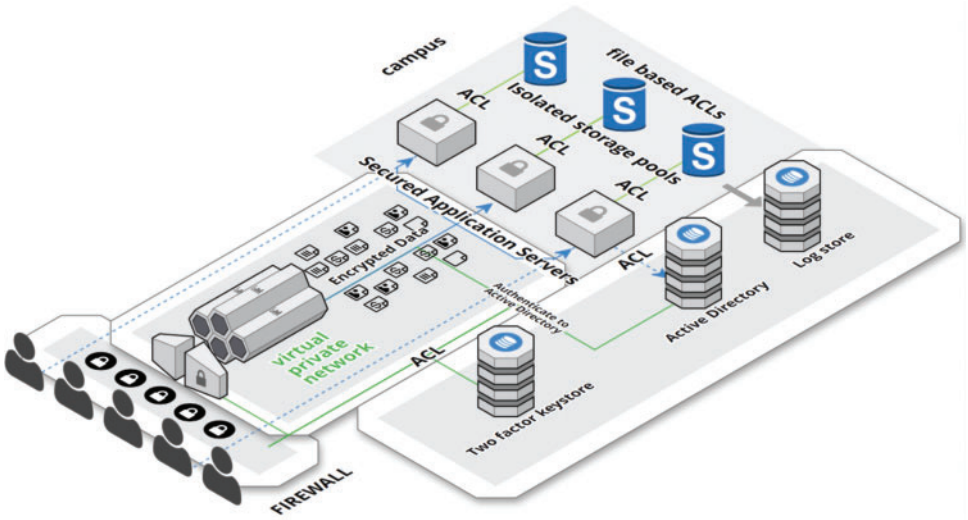

Figure 5. Diagram of the Medical Science DMZ architecture at Harvard University.

Harvard University

Running shared computing with open, national, and international network connections with collaborating researchers is clearly orthogonal to the methodologies required when managing and supporting restricted data. Supporting complex Data Use Agreements (DUAs) has also become increasingly more difficult, with serious federal and inordinately large financial penalties for any and all violations.

How does Harvard balance these 2 issues, maintaining open access to research faculty while controlling who has access? Given that it is the individual who physically signs the DUA, Harvard accordingly protects at the level of that individual and his or her sponsoring faculty. Dedicated systems are placed inside a firewall, with dedicated VPNs controlled by the user's defined Organizational Unit from the domain controller. Individual machines are secured, logged, backed up, monitored, and sandboxed from the main shared cluster. Only fully deidentified data based on the DUA are allowed within the shared environment. Thus, there is security, but there is no sharing (except for fully deidentified data). Of course, this does not scale.

Recently, virtualization of encrypted virtual devices has enabled some degree of scaling and "shared hosting." The virtual "container" and virtual network isolation removes the ability for an individual user to see any other user and/or system from inside of his or her system. Given that DUAs actively prohibit any and all sharing, the system needs to be designed accordingly, but once again, this unfortunately results in secure systems that do not scale.

Then there is the issue of data transfer. Because of the limitation to secure the endpoint via VPNs, the data pass through a Secure Sockets Layer appliance onto the private secured system via a dedicated encrypted tunnel. Harvard is effectively performing the digital equivalent of taking a virtual "armored" network cable out to its end user community, pulling "virtual wires" on an individual basis through the VPN.

Harvard bases its architecture on the "locked computer room" model, where, traditionally, health care records could only be accessed from stand-alone systems behind physical locked doors. The VPN, user ID, and 2-factor authentication enable access to the physical or virtual machine, with a subsequent login to finally access that system. Logs are pulled to a central facility that cannot be accessed, so chain of custody exists in the event of any issue. Backups are pulled to an offsite secured system.

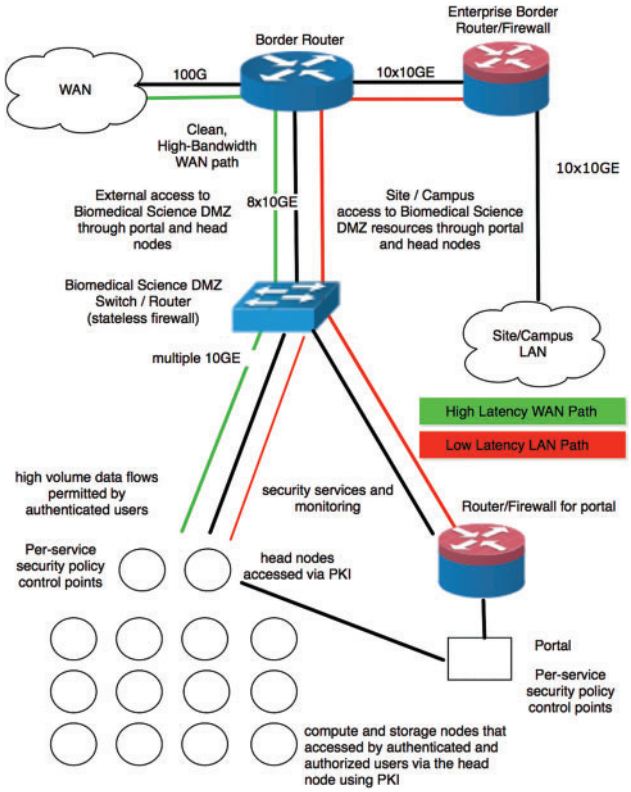

Figure 6. Diagram of the Bionimbus Science DMZ architecture at the University of Chicago. As shown, with this architecture, the University of Chicago does not use a commercial firewall between the storage and compute nodes and Science DMZ per se, but instead has a number of "compensating" controls and procedures to provide the required security. The traffic between the commodity and research networks and the applications portals is standard web traffic, and these connections can use firewalls.

Harvard medical computing leadership now works with the provost's Office for Research, in that any DUAs or requests for datasets are automatically passed to the assistant dean for research computing, and a conversation is then started with a documented process

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

Table 1. Sample risk matrix for high-speed transfers

Risk	How mitigated
Since there is no firewall, an attacker will discover open ports.	Only 3 ports are open. The IDS will detect port scans and generate alerts.
The attacker launches a denial of service (DoS) attack.	The IDS will detect and stop a DoS attack.
The attacker steals/guesses a user password and gains access to the user's data.	The system has a very small number of users. They have been trained to detect phishing, etc., and encrypt data prior to storage. Long passphrases are mandatory. The passphrase strength is very high.
The attacker exploits unpatched software accessible through open ports.	System is patched within 6 hours of a critical vulnerability.
The attacker gains system entry and engages in suspicious activity.	Logs are sent to a central log host and monitored in real time using a security information event management system. Alerts are generated when suspicious activity signatures are detected.
A successful attack is detected.	Mature incident response and reporting procedures are in place. System is immediately isolated.

To understand how a risk-based DMZ works for HIPAA-aligned data-intensive flows, consider data transfers between a supercomputer and the high-performance data storage system in a Medical Science DMZ. Neither system uses a firewall; a user can transfer data at a high transfer rate between the two. Both the supercomputer and the high-performance storage system have been HIPAA aligned, as described earlier. They implement a large number of baseline NIST 800-53 controls,[17] which, when supplemented by enterprise common NIST controls, act as alternate controls that lower or mitigate the risk of data exposure due to the absence of firewalls. Table 1 shows a sampling of how risk is analyzed. Risks found are balanced against the need to transfer large volumes of data to accomplish research goals, history of attacks, cost, etc. As mentioned in "Classical Science DMZ" section, we stress that while DTNs can be seen as an exposed single point of failure, the Medical Science DMZ architecture focuses risk on them by design, thus emphasizing the key parts of the architecture that require hardening and monitoring.

with the principal faculty member to design the specific system with sufficient security, speed, capacity, and capability.

After 3 years under this operating model, Harvard has observed that the research computing organization has become significantly more efficient. However, the DUA that comes directly from the data provider still becomes the bottleneck; each is like a snowflake. They are unique and special, and although they look similar from a distance, close up, no two are quite alike. Virtualization has thus far been the toolkit for research computing to be able to exploit shared physical infrastructure for velocity and agility, but remaining as restricted and controlled "containers" for each project to allow for segregation and separation. Configuration management has also enabled "templates," to ensure that the systems have appropriate logging, backup, software stack, and access controls; in light of recent Secure Sockets Layer flaws, this has been invaluable to be able to rapidly and automatically patch systems at risk. The configuration management database and automation are critical to maintaining control of the "one-off" project-based systems.

Self-provisioning of secure private tunnels with encrypted underlying storage and isolation of the containerized system is clearly where we need to be. Effectively, traditional university systems are behind the curve from a pure technology perspective, but clearly very much aligned in terms of policy, access control, and documentation. Science DMZ flexibility is the telecommunications equivalent of the "last mile" of security for restricted data sets.

Bionimbus protected data cloud and data commons architecture

Over the past 4 years, the University of Chicago, in collaboration with the not-for-profit Open Cloud Consortium, has developed and operated cloud-based computing infrastructure and data commons for the biomedical research community with a common technical architecture but different exposed services. Both of these are private and are housed at one of the University of Chicago data centers, and both support high-performance data transport through the Science DMZ that are tightly integrated with the security services of the application. Both also employ the NIST risk management framework and implement a large number of NIST 800-53 controls.

Key architecture decisions include the following:

- All network traffic from outside the application to the computing and storage infrastructure pass through one or more "head nodes," which are heavily monitored. In this sense, the storage nodes and computing nodes are not "connected directly to the Internet," which is a requirement of many of our applications.
- Iptables host-based firewall rules are used to restrict access to the system.
- Only authenticated and authorized users may access computing and storage services and controlled access and sensitive data associated with the applications. This access is only through public and private encryption keys, with paths that are routed through the head node. No password-based access is used in the application's nodes, except for accessing the application portal.
- All remote access to the system is monitored, and all access to the data and to the applications services is also monitored. The corresponding log files are reviewed regularly for irregularities and potential breaches.
- The University of Chicago regularly scans and run a variety of information security services across the entire infrastructure and records the results of these scans.
- High-performance utilities for moving data through the Science DMZ to connected organizations can only be initiated by authenticated and authorized users and are fully logged and regularly reviewed.
- Traffic containing sensitive or restricted data, including traffic using high-performance data transport protocols through the Science DMZ, is encrypted.

SUMMARY

The national high-performance network and storage infrastructure provides an attractive, scalable alternative to the risk and cost of implementing clouds to handle the biomedical data avalanche and

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018

the computational workflows they entail. We have defined a Medical Science DMZ as a potential institutional approach to solving the security and regulatory issues introduced by HIPAA, and described several production implementations where this architecture is already being deployed to support research with sensitive data at scale.

The Medical Science DMZ is able to transfer data at high throughput by ensuring that endpoints are HIPAA aligned and implement a risk management framework such as NIST, thus introducing alternate controls that lower or mitigate the risk of data exposure due to the absence of packet-filter firewalls. Finally, we described several production implementations where this architecture is already being deployed to support research with restricted data at scale.

FUNDING

This work was supported in part by the director of the Office of Science, Office of Advanced Scientific Computing Research of the US Department of Energy, under contract number DE-AC02-05CH11231. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of any of the employers or sponsors of this work.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

SP, ED, and WB were the primary authors of this paper. EB, JC, and RLG also contributed substantially to the paper and, in particular, contributed significantly to the "case study" portion of the paper. AB, AS, and BT all contributed intellectual value and text to the paper.

ACKNOWLEDGMENTS

Indiana University thanks ESnet for hosting a set of DTN test points and readily accessible performance tuning guides. These resources were very helpful in SciPass evaluations. Thanks also to Brocade Communication Systems Inc., which provided the switch hardware support and technical input for the SciPass testbed.

REFERENCES

1. What Is Big Data? https://datascience.nih.gov/bd2k/about/what. Accessed September 20, 2017.
2. *NCI Cloud Resources*. https://cbiit.cancer.gov/ncip/cloudresources. Accessed September 20, 2017.
3. LeDuc RD, Vaughn M, Fonner JM, *et al*. Leveraging the national cyberinfrastructure for biomedical research. *J Am Med Inform Assoc*. 2014;21:195–99.
4. *Open Science Grid*. http://opensciencegrid.org. Accessed April 17, 2015.
5. US Department of Health and Human Services. *HIPAA Security Rule*. www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/. Accessed April 17, 2015.
6. National Institute of Standards and Technology. *Guide for Conducting Risk Assessments — NIST Special Publication 800-30 Rev. 1*. https://csrc.nist.gov/publications/detail/sp/800-30/rev-1/final. September 2012. Accessed September 20, 2017.
7. *Science DMZ Network Architecture*. http://fasterdata.es.net/science-dmz/. Accessed April 17, 2015.
8. Dart E, Rotman L, Tierney B, *et al*. The Science DMZ: A Network Design Pattern for Data-Intensive Science. *Proc IEEE/ACM Annual SuperComputing Conference (SC13)*. Denver; 2013.
9. National Science Foundation. *Campus Cyberinfrastructure*. www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748. Accessed March 23, 2017.
10. Peisert S, Barnett WK, Dart E, *et al*. The Medical Science DMZ. *J Am Med Inform Assoc*. 2016;23(6):1199–1201.
11. National Institute of Standards and Technology. *Guidelines on Firewalls and Firewall Policy – NIST Special Publication 800-41, revision 1*. http://csrc.nist.gov/publications/nistpubs/800-41-Rev1/sp800-41-rev1.pdf. September 2009. Accessed September 20, 2017.
12. Paxson V. Bro: a system for detecting network intruders in real-time. *Comput Networks*. 1999;31(23):2435–63.
13. *The Bro Network Security Monitor*. www.bro.org. Accessed April 17, 2015.
14. *Snort – Network Intrusion Detection & Prevention System*. www.snort.org. Accessed July 25, 2017.
15. *GlobalNOC*. https://globalnoc.iu.edu. Accessed April 17, 2015.
16. *SciPass*. http://globalnoc.iu.edu/sdn/scipass.html. Accessed April 17, 2015.
17. National Institute of Standards and Technology. *Security and Privacy Controls for Federal Information Systems and Organizations – NIST Special Publication 800-66, revision 1. An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule*, October 2008.
18. National Institute of Standards and Technology. *Security and Privacy Controls for Federal Information Systems and Organizations – NIST Special Publication 800-53, revision 4*. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf, April 2013. Accessed September 20, 2017.
19. McKeown N, Anderson T, Balakrishnan H, *et al*. Openflow: enabling innovation in campus networks. *ACM SIGCOMM Comput Commun Rev*. 2008;38(2):69–74.
20. Hanemann A, Boote JW, Boyd EL, *et al*. PerfSONAR: A service oriented architecture for multi-domain network monitoring. In *Proceedings of the Third International Conference on Service Oriented Computing*. 2005:241–54.

Downloaded from https://academic.oup.com/jamia/article-abstract/25/3/267/4367749
by Ruth Lilly Medical Library user
on 29 August 2018