The Medical Science DMZ

RECEIVED 2 December 2015
REVISED 17 January 2016
ACCEPTED 11 February 2016

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

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD UNIVERSITY PRESS

ABSTRACT

Objective We describe use cases and an institutional reference architecture for maintaining high-capacity, data-intensive network flows (e.g., 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 data sets between research institutions over national research networks.

Discussion The exponentially increasing amounts of "omics" data, the rapid increase of high-quality imaging, and other rapidly growing clinical data sets 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 data sets. Maintaining data-intensive flows that comply with HIPAA and other regulations presents a new challenge for biomedical research. Recognizing this, 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.

Keywords: Computer Communication Networks, Data Intensive Science, High Performance Computing, Biomedical Research, Computer Security, Health Insurance Portability and Accountability Act

INTRODUCTION

With a national commitment to precision medicine,[1] medical science is quickly moving into the realm of "big data."[2] Storage, computation, and transfer needs to process these data are growing rapidly in medical schools, outstripping the capacity of on-premise IT resources. 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 pursuing 3 cloud pilots for cancer genomics.[3] National HPC facilities are applying their capacity to biomedical research. These efforts are interconnected by high-capacity research networks such as the Internet2 and ESnet. 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. Academic computing resources connected to R&E networks have traditionally been leveraged for applications at scale such as high-energy physics research (e.g., 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. 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 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. The key control points for these data at each institution are firewalls that inspect network traffic, secure sensitive data, and mitigate risks. A de facto technical control in environments subject to regulations such as the Health Insurance Portability and Accountability Act (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 requirements needed for data-intensive science.

A "Science DMZ" is a portion of the network at the local network perimeter of an individual research institution that is designed such that the equipment, configuration, and security policies are optimized for high-

Correspondence to Sean Peisert, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; Department of Computer Science, University of California Davis, Davis, CA, USA; Corporation for Education Network Initiatives in California (CENIC), Berkeley, CA, USA; sppeisert@lbl.gov. For numbered affiliations see end of article.

performance workflows and large data sets.[5,6] 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 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 inadequate performance. The basic Science DMZ model[5,6] has been successfully implemented in numerous scenarios, including those involving astrophysics, photon science, high-energy physics, materials science, climate modeling, and genomics.

We have taken a central tenet of the Science DMZ[7,8] and re-engineered it for sensitive data as a "Medical Science DMZ." Science DMZs operate at scale using already-provisioned software and authentication stacks 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 data sets 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.

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. To reflect this philosophy, we define 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.

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."[9] NIST 800-41 defines multiple different types of firewalls, including:

- Packet-filtering firewalls that use attributes of the packet headers as the basis for their access control decisions.
- Stateful firewalls that track the connection state in the same way the end hosts do, and are able to detect protocol-level anomalies and other threats that a packet filter cannot.
- Application-layer firewalls that examine the contents of the packets and messages and grant or deny access based on inferred application state.

Commercial equipment providers have tended to only define stateful firewalls and application firewalls as "firewalls." Despite this, the standards body considered authoritative in matters of US government policy (NIST) considers a packet-filtering router a firewall.

From the perspective of NIST 800-41, a Science DMZ uses a non-stateful, packet-filter firewall implemented in the gateway, or a downstream router. In this model, a packet enters the firewall, its source and destination addresses are compared to a list of rules, and the firewall takes action (forward or discard) associated with the rule. Other compensating controls, such as Intrusion Detection Systems (IDSs) (e.g., the Bro system[10,11]), are often employed. A capable Science DMZ firewall can be configured to copy every packet it receives and send it to an IDS. The IDS analyzes the packets and can take action to block hostile traffic.

MEDICAL SCIENCE DMZ ARCHITECTURES

1. Classical science DMZ

In a Science DMZ, a network enclave is constructed using high-performance network routers at or near the institutional network perimeter. Because they are 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-throughput servers called Data Transfer Nodes (DTNs) are connected directly to these routers in the Science DMZ. The DTNs handle all data ingest/export tasks for the Science DMZ. The router to which the DTN is directly connected implements security controls for the DTN. The DTN typically also runs host-based firewalls or IDS packages, and the set of applications running on the DTN is strictly limited to system maintenance and data ingest/export tasks. The limited number of applications on the DTN is a critical point—it dramatically reduces the attack surface and makes the DTN a better fit for risk-based security controls.

So designed, a Science DMZ is inherently resistant to a wide variety of attacks. If data encryption is implemented, the data are not accessible to adversaries that might snoop on the communication between Science DMZs that share a trust relationship. The DMZ router controls which DTNs exchange data, limiting opportunities for data exfiltration. The IDS monitors for policy infractions and incoming hostile activity. 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 an IDS is employed, it can monitor DTN network traffic to ensure that the policies on the DTN are being followed. This defense in depth is 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 subenclaves within an institution, each with its own risk profile, security policies, compensating controls, and so forth. The following case studies describe the addition of capabilities to the classical Science DMZ, enhancing it for use in environments with protected data.

2. Three medical science DMZ implementations

Indiana University (IU), Harvard University, and the University of Chicago all implement Medical Science DMZs. Each one has implemented frameworks that allow the free flow of data where needed and address HIPAA using alternate controls that manage risk. To that end, each organization has implemented a specific "risk-managed" technical DMZ solution that encompasses the entire high-performance computing, storage, and network infrastructure.

IU has created a holistic environment called "SciPass"[12] that leverages a comprehensive, NIST-based risk management framework.[13,14] SciPass[12] is managed by the IU GlobalNOC.[15] The SciPass system contains 6 components: an OpenFlow Switch,[16] the SciPass controller, a cluster of IDS sensors, a PerfSONAR host,[17] a firewall, and a DTN. SciPass defines IDS policies to identify "good" flows. These policies contain a combination of time of the day and day of the week, and source and designation IP address, along with protocol and application layer data to determine if a flow should bypass an institutional firewall. Thus, users, network administrators, and security administrators are able to jointly define and enforce desired network behavior.

By default, traffic is forwarded from the OpenFlow switch through the institutional firewall, and copies of packets are sent to the array of IDS sensors. When policies determine that it is appropriate to route an individual flow around the firewall, a pair of OpenFlow rules are added to the switch so that packets associated with this flow are directly forwarded, bypassing the institutional firewall and the IDS array. These rules contain an idle timeout so that once the flow is complete, the rules are purged from the switch. The SciPass architecture allows exceptions for known high-performance data transfers. The flexibility of

the Science DMZ model allows for these enhancements, which significantly reduce the attack surface for the DTNs.

Harvard University's approach begins with an individual signing a Data Use Agreement. Dedicated systems inside a firewall, with dedicated virtual private networking, are controlled by the users' Organizational Unit. Individual machines are secured, logged, backed up, monitored, and sandboxed from the main shared cluster. Data sharing is enabled by leveraging encrypted virtual containers and networks. For data transfer, the data pass through a Secure Sockets Layer appliance onto the private secured system via a dedicated encrypted tunnel. The virtual private network (VPN), user id, and 2-factor authentication system enable access to the physical or virtual machine, with a subsequent login to finally access that system.

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 known as "Bionimbus" for the biomedical research community. Both of these support high-performance data transport through the Science DMZ, tightly integrated with the security services of the application.

Key architecture decisions include the notion that all network traffic from outside the application to the computing and storage infrastructure passes through one or more heavily monitored "head nodes." The storage and computing nodes are not connected directly to the Internet. Additionally, traffic containing sensitive or controlled access data, including traffic using high-performance data transport protocols through the Science DMZ, is encrypted.

SUMMARY

The national high-performance network infrastructure provides a scalable option for handling the biomedical data avalanche and the ensuing computational workflows. We have defined a Medical Science DMZ as a potential institutional approach to solving the security and regulatory issues introduced by HIPAA. The Medical Science DMZ is able to transfer data at high throughput by ensuring that endpoints are HIPAA-aligned, implement a large number of baseline NIST 800-53 controls,[13] and are supplemented by enterprise common NIST controls,[14] thus introducing alternate controls that lower or mitigate the risk of data exposure due to an absence of packet-filter firewalls. Finally, we have described several production implementations where this architecture is already being deployed to support research with sensitive data at scale.

CONTRIBUTORSHIP STATEMENT

S.P., W.B., and E.D. were the primary authors of this paper. J.C. and R.L.G. also contributed substantially to the paper and, in particular, contributed significantly to the "case study" portion of the paper. E.B., A.B., A.S., and B.T. all contributed intellectual value and text to the paper.

COMPETING INTERESTS

None.

ACKNOWLEDGEMENTS

This work was supported in part by the Director, 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.

IU 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., who provided the switch hardware support and technical input for the SciPass testbed.

REFERENCES

1. FACT SHEET: President Obama's Precision Medicine Initiative. https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative. Accessed November 25, 2015.
2. What is Big Data? http://bd2k.nih.gov/about_bd2k.html#bigdata. Accessed April 17, 2015.
3. NCI Cancer Genomics Cloud Pilots. https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots. Accessed November 25, 2015.
4. Open Science Grid. http://opensciencegrid.org. Accessed April 17, 2015.
5. U.S. Department of Health and Human Services (HHS). HIPAA Security Rule. http: //www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/. Accessed April 17, 2015.
6. National Institute of Standards and Technology (NIST). Risk Management Guide for Information Technology Systems — NIST Special Publication 800-30. http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/nist800-30.pdf, July 2002. Accessed November 25, 2015.
7. Dart E, Rotman L, Tierney B, *et al.* The Science DMZ: A Network Design Pattern for Data-Intensive Science. *Proceedings of the IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver CO, 2013.
8. Science DMZ Network Architecture. http://fasterdata.es.net/science-dmz/. Accessed April 17, 2015.
9. National Institute of Standards and Technology (NIST). 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 April 17, 2015.
10. Paxson V. Bro: a system for detecting network intruders in real-time. *Comput Networks.* 1999;31(23):2435–2463.
11. The Bro Network Security Monitor. https://www.bro.org. Accessed April 17, 2015.
12. SciPass. http://globalnoc.iu.edu/sdn/scipass.html. Accessed April 17, 2015.
13. National Institute of Standards and Technology (NIST). 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 April 17, 2015.
14. National Institute of Standards and Technology (NIST). 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. http://csrc.nist.gov/publications/nistpubs/800-66-Rev1/SP-800-66-Revision1.pdf. Accessed November 25, 2015.
15. GlobalNOC. https://globalnoc.iu.edu. Accessed April 17, 2015.
16. McKeown N, Anderson T, Balakrishnan H, *et al.* Openflow: enabling innovation in campus net- works. *ACM SIGCOMM Comput Commun Rev.* 2008;38(2):69–74.
17. 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.* Springer, LNCS 3826; 2005:241-254.

AUTHOR AFFILIATIONS

[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]Indiana Clinical and Translational Sciences Institute and Regenstrief Institute, Indiana University, Indianapolis, IN, USA

[5]ESnet, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[6]Research Computing, Harvard University, Cambridge, MA, USA
[7]Center for Data Intensive Science, University of Chicago, Chicago, USA
[8]Global Research Network Operations Center, Indiana University, Bloomington, IN, USA
[9]BioTeam, Middleton, MA, USA
[10]Pervasive Technology Institute, Indiana University, Bloomington, IN, USA