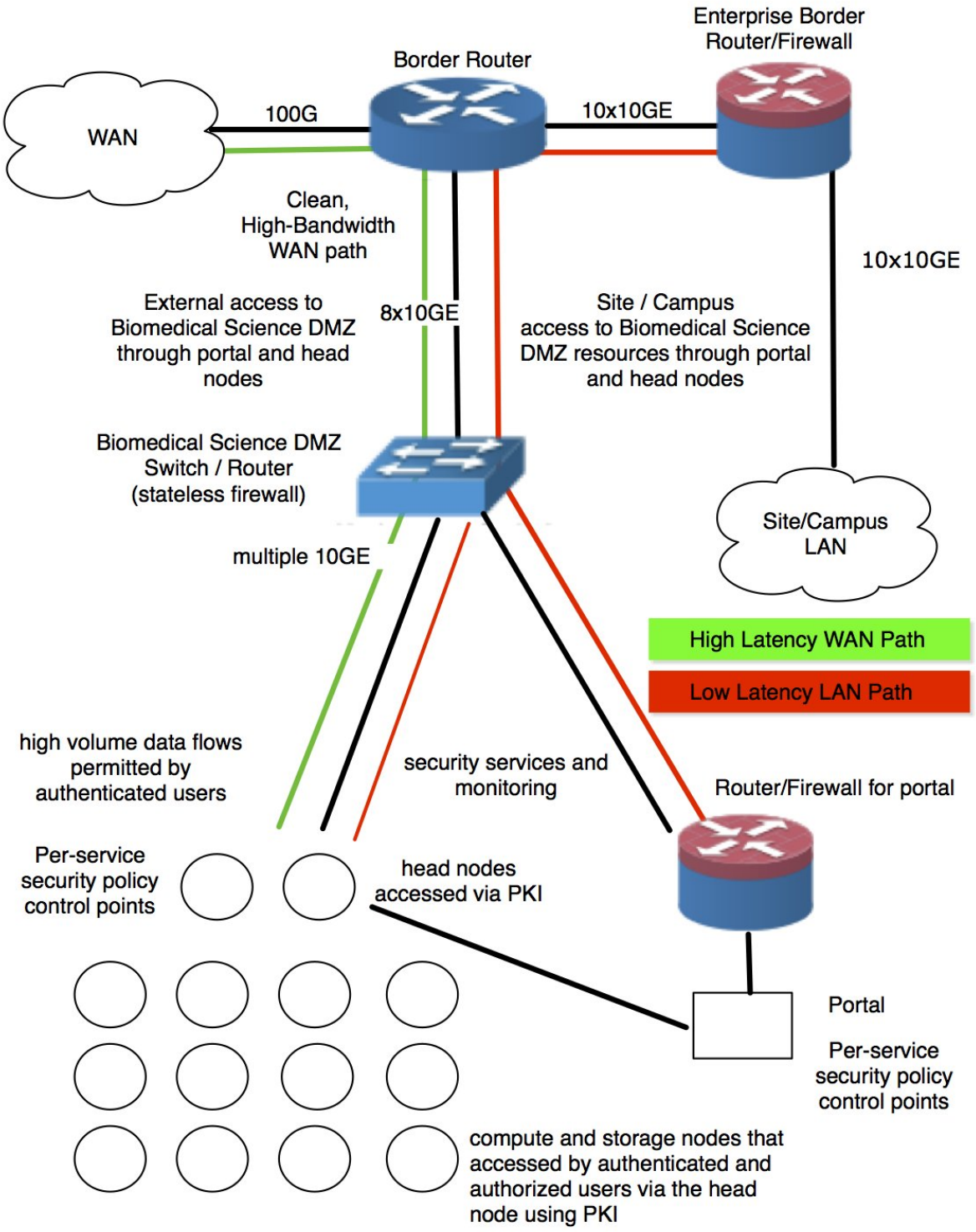
ESnet's science DMZ design could help transfer, protect medical research data

October 17 2017



Schematic drawing showing components of a Science DMZ. Credit: Sean Peisert, Berkeley Lab

Like other sciences, medical research is generating increasingly large datasets as doctors track health trends, the spread of diseases, genetic causes of illness and the like. Effectively using this data for efforts ranging from stopping the spread of deadly viruses to creating precision medicine treatments for individuals will be greatly accelerated by the secure sharing of the data, while also protecting individual privacy.

In a paper published Friday, Oct. 6 by the *Journal of the American Medical Informatics Association*, a group of researchers led by Sean Peisert of the Department of Energy's (DOE) Lawrence Berkeley National Laboratory (Berkeley Lab) wrote that the Science DMZ architecture developed for moving large data sets quick and securely could be adapted to meet the needs of the medical research community.

The Science DMZ traces its name to an element of network security architecture. Typically, located at the network perimeter, a DMZ has its own security policy because of its dedicated purpose - exchanging data with the outside world.Exponentially increasing amounts of data from genomics, high quality imaging and other clinical data sets could provide valuable resources for preventing and treating medical conditions. But unlike most scientific data, medical information is subject to strict privacy protections under the Health Insurance Portability and Accountability Act (HIPAA) so any sharing of data must ensure that these protections are met.

"You can't just take the medical data from one site and drop it straight in to another site because of the policy constraints on that data," said Eli Dart, a network engineer at the Department of Energy's Energy Sciences Network (ESnet) who is a co-author of the paper. "But as members of a society, our health could benefit if the medical science community can become more productive in terms of accessing relevant data."

For example, an authenticated user could query a very large data base

stored at multiple sites to learn more about an emerging medical issue, such as the appearance of a new virus, said Peisert, who works in Berkeley Lab's Computational Research Division. In this way, teams of widely dispersed experts could collaborate in real-time to address the problem.

According to the authors of the paper, the storage, analysis and network resources needed to handle the data and integrate it into patient diagnoses and treatments have grown so much that they strain the capabilities of academic health centers. At the same time, shared data repositories like those at the National Library of Medicine, the National Cancer Institute and international partners such as the European Bioinformatics Institute are rapidly growing.

"But by implementing a Medical Science DMZ architecture, we believe 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," Peisert said. "Access would of course need to be properly authenticated, but unlocking the world's medical information could yield enormous benefits."

The authors 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 and appropriately managing risk." Their network design pattern addresses Big Data and can be implemented using a combination of physical, administrative and technical safeguards.

The paper was written as the National Institutes of Health (NIH) are spearheading a "Commons Initiative" for sharing data; the NIH have long provided reference data through the National Library of Medicine. The National Cancer Institute funded a number of pilot projects to use

cloud computing for cancer genomics in 2016, and the initiative has since continued and expanded beyond the pilot phase.s. Many universities with high-performance computing facilities available are increasingly applying their capacity to biomedical research.

The Science DMZ network architecture, which is used by more than 100 research institutions across the country, provides speed and security for moving large data sets. Dart led the development of the Science DMZ concept, formalized it in 2010, and has been helping organizations deploy it ever since.

A Science DMZ is specifically dedicated to external-facing high-performance science services and is separate from an organization's production network, which allows bulk science data transfers to be secured without inheriting the performance limitations of the infrastructure used to defend enterprise applications.

Data transfers using Science DMZs are straightforward from a network security perspective: the data transfer nodes (specially tuned servers) exchange security credentials to authenticate the transfer and then open several connections to move the specified data. One the job is completed, the connections close down. In the case of moving medical data, the information is encrypted both while it is being stored and while it's moving across the network.

"There's no magic," Dart said. "The security is easy to manage in that the sites are known entities and nothing moves without proper security credentials."

In fact, Dart said, such transfers pose less of a security problem than surfing the web on a personal computer connected to an open network. When someone browses a web site, the user's computer downloads content from many different locations as specified by the web page,

including ads that are sold and resold by firms around the world and may contain malware or other security threats. A data transfer between Science DMZs is a comparatively simple operation that doesn't involve image rendering or media players (which are common attack surfaces), and only transfers data from approved endpoints.

In their paper, the authors present the details of three implementations and describe how they balance the key aspects of a Medical Science DMZ of high-throughput and regulatory compliance. Indiana University, Harvard University, and the University of Chicago all use a non-firewalled approach to moving HIPAA-protected data in their Medical Science DMZs. Each site has implemented frameworks that allow free flow of data where needed and address HIPAA using alternate, reasonable and appropriate controls that manage risk.

In each case the data transfers are encrypted, and can only be initiated by authenticated and authorized users. The interactive network traffic needed to initiate such transfers still passes through one or more systems that are heavily protected and monitored. Although firewalls are not removed entirely from the system, they are used intelligently and overall system security is maintained while still permitting the transfer of sensitive data, such as large biomedical datasets.

"We wrote this paper as a starting point," Peisert said, "and hope that it will allow a lot of great things to happen."

Provided by Lawrence Berkeley National Laboratory