The Science DMZ and the CIO: Data Intensive Science and the Enterprise

Eli Dart, Network Engineer
ESnet Science Engagement
Lawrence Berkeley National Laboratory

ESnet/Internet2 Technology Exchange
Indianapolis, IN
October 29, 2014
Outline

• Science DMZ context
  – Where we are
  – Why would you build a Science DMZ?
• Success factors
  – What makes a Science DMZ successful?
• Enterprise traffic vs. science traffic
  – Differences
  – Security implications
The Science DMZ in 1 Slide

• “Friction free” network path
  – Highly capable network devices (wire-speed, deep queues)
  – Virtual circuit connectivity option
  – Security policy and enforcement specific to science workflows
  – Located at or near site perimeter

• Dedicated, high-performance Data Transfer Nodes (DTNs)
  – Hardware, operating system, config all optimized for data transfer
  – High-performance data transfer tools such as Globus

• Performance test and measurement – perfSONAR

• Science engagement
  – Map experiments onto cyberinfrastructure
  – Work with users to ensure they are successful

• Details at http://fasterdata.es.net/science-dmz/
Science DMZ Design Pattern (Abstract)

- **Border Router**
- **Enterprise Border Router/Firewall**
- **WAN**
  - Clean, High-bandwidth WAN path
- Science DMZ Switch/Router
  - Per-service security policy control points
  - Site / Campus access to Science DMZ resources
- Site / Campus LAN
- High performance Data Transfer Node with high-speed storage
- perfSONAR connections at various nodes.
Context: Science DMZ Adoption

- DOE National Laboratories
  - Both large and small sites
  - HPC centers, LHC sites, experimental facilities
- NSF CC-NIE and CC*IIE programs leverage Science DMZ
  - $40M and counting (third round awards coming soon, estimate additional $18M to $20M)
  - Significant investments across the US university complex, ~130 awards
  - Big shoutout to Kevin Thompson and the NSF – these programs are critically important
- National Institutes of Health
  - 100G network infrastructure refresh
- US Department of Agriculture
  - Agricultural Research Service is building a new science network based on the Science DMZ model
  - [https://www.fbo.gov/index?s=opportunity&mode=form&tab=core&id=a7f291f4216b5a24c1177a5684e1809b](https://www.fbo.gov/index?s=opportunity&mode=form&tab=core&id=a7f291f4216b5a24c1177a5684e1809b)
- Other US agencies looking at Science DMZ model
  - NASA
  - NOAA
- Australian Research Data Storage Infrastructure (RDSI)
  - Science DMZs at major sites, connected by a high speed network
Context: Community Capabilities

- Many Science DMZs directly support science applications
  - LHC (Run 2 is coming soon)
  - Experiment operation (Fusion, Light Sources, etc.)
  - Data transfer into/out of HPC facilities
- Many Science DMZs are SDN-ready
  - Openflow-capable gear
  - SDN research ongoing
- High-performance components
  - High-speed WAN connectivity
  - perfSONAR deployments
  - DTN deployments
- Metcalfe’s Law of Network Utility
  - Value proportional to the square of the number of DMZs? n log(n)?
  - Cyberinfrastructure value increases as we all upgrade
Strategic Impacts

• What does this mean?
  – We are in the midst of a significant cyberinfrastructure upgrade
  – Enterprise networks need not be unduly perturbed 😊

• Significantly enhanced capabilities compared to 3 years ago
  – Terabyte-scale data movement is much easier
  – Petabyte-scale data movement possible outside the LHC experiments
    • 3.1Gbps = 1PB/month
    • (Try doing that through your enterprise firewall!)
  – Widely-deployed tools are much better (e.g. Globus)

• Raised expectations for network infrastructures
  – Scientists should be able to do better than residential broadband
    • Many more sites can now achieve good performance
    • Incumbent on science networks to meet the challenge
      – Remember the TCP loss characteristics
      – Use perfSONAR
  – Science experiments assume this stuff works – we can now meet their needs
Why Build A Science DMZ?

• Data set scale
  – Detector output increasing
    • 1Hz → 10Hz → 100Hz → 1kHz ... → 1MHz
  – HPC scale increasing
    • Increased model resolution → increased data size
    • Increased HPC capability means additional problems can now be solved
  – Sequencers, Mass Spectrometers, ...

• Data placement
  – Move compute to the data?
  – Sure, if you can...otherwise you need to move it

• Who needs the raw data?
  – Anyone working on processing algorithms for raw data
  – Anyone aggregating/integrating data sets (absent perfect prior reduction)
  – Anyone doing data analysis for which a canned service does not exist

• Without a Science DMZ, this stuff is hard
  – Can you assume nobody at your institution will do this kind of work?
  – If this kind of work can’t be done, what does that mean in 5 years?
What Makes A Science DMZ Successful?

• A Science DMZ is successful when it’s useful
  – Contribution to science outcomes
  – Reduced cost for supporting science projects
  – Enable research that could not otherwise be done

• Several different parts to this
  – Networking organization must understand it
  – Systems organization must understand it
  – Security organization must understand it
  – Scientists/researchers must understand it

• Once everyone understands it and agrees, then it’s just implementation

• How do we bring this about?
Networking and Systems

• It’s pretty easy for networking folks to understand this stuff
  – It’s networking stuff, after all...
  – Sometimes a bit trickier to explain it to senior leadership
    • Roll up the technical detail
    • Strategic implications rather than bits and bytes

• Systems folks are generally on board as well
  – DTNs are straightforward
  – Most systems folks tend to like performance anyway
  – Systems people deal with users a lot – they like to be able to make users happy
Security

• Security folks can be harder
  – Firewall people in particular can be a challenge
  – We have seen some very “steadfast” firewall people
    • They can kill a project if they aren’t on board
    • Depending on the personalities involved, data may not be enough
    • In some cases, getting the security people on board means senior leadership giving them orders – try to avoid that if possible

• Remember – most of us work for science organizations
  – If science is the primary mission, then everybody works for the scientists
  – In a lot of cases security is reasonable – they just need to be included rather than dictated to
  – (Security people like performance too)
Science Engagement

• Scientists and researchers need to be able to use cyberinfrastructure
  – If I can’t use a tool, that tool doesn’t exist for me
  – There are already too many tools – we can’t expect folks to find the right ones at random
  – Scientists don’t have the cycles to be system integrators

• Science engagement bridges the gap
  – Understand what the scientists need to do with their data
  – Understand the capabilities of the cyberinfrastructure
  – Map the science onto the infrastructure

• Understanding the infrastructure is straightforward for us
  – We’re infrastructure people, right?
  – How do we understand the science?

• “Requirements and Relationships”
Requirements

• Requirements – what does the science project need?
  – Several different ways of getting to this
  – We can be told late, or we can go find out (I prefer to be proactive)
  – ESnet requirements process: http://www.es.net/requirements/

• Characterize science project from multiple angles
  – Instruments and facilities
    • “Hardware of science”
    • Detectors, telescopes, tokamaks, HPC facilities
    • Tells us about the data – where, how fast, how much, etc.
  – Process of science
    • How do scientists use the data for discovery?
    • Where does the data need to go? How is it analyzed? What time scale?
  – Assessments done in formal reviews
Requirements Review Structure

• Several key elements
  – Case studies provide a network-centric narrative of the science
  – Requirements synthesized from science instruments, facilities, and science process – *in collaboration with science programs*

• Process details are important
  – Four actors have the same conversation at the same time
    • ESnet, ESnet program management at DOE
    • Senior science program members, science program management at DOE
  – Open discussion about needs, issues, changes, best practice
    • All parties have the same conversation in the same room at the same time
    • Common understanding of program needs and the solutions ESnet undertakes to meet those needs
  – Review reports are vetted by ESnet and by both programs
Relationships

• Relationships – familiarity, understanding, trust
  – Work with science collaborations to understand their needs
  – Make their lives better
    • Fix problems
    • Give them better tools and workflows
  – Make sure you are accurate (expectations are important)
  – Check in with people periodically
    • I make it a practice to ask “is there anything we need to talk about?”
    • Often people won’t come to you first, but they will give you a chance to help if you check in
• Once you get a reputation for solving problems, it all gets easier
  – People come to you first
  – You get in early on the planning
  – People are more willing to push the envelope
Science Traffic – What Makes It Special?

• Large scale data transfers are a hallmark of science traffic
  – Yes, scientists use web browsers, email, etc.
  – However, moving the data is the differentiator

• Enterprise traffic is typically composed of a large number of small flows
  – Web, email, document sharing, IP phones, ...
  – VPNs carrying all of the above

• We distinguish these in the following way:
  – Large-scale science traffic: Elephant flows
  – Enterprise traffic: Mouse flows
ESnet is not the Commercial Internet
Support For Science Traffic

• The Science DMZ is typically deployed to support science traffic
  – Typically large data transfers over long distances
  – In most cases, the data transfer applications use TCP

• The behavior of TCP is a legacy from the congestion collapse of the Internet in the 1980s
  – Loss is interpreted as congestion
  – TCP backs off to avoid congestion \(\rightarrow\) performance degrades
  – Performance hit related to the square of the packet loss rate

• Addressing this problem is a dominant engineering consideration for science networks
  – Lots of design effort
  – Lots of engineering time
  – Lots of troubleshooting effort
A small amount of packet loss makes a huge difference in TCP performance.

With loss, high performance beyond metro distances is essentially impossible.
Security Implications For Traffic Types

• We have two distinct traffic profiles
  – Commodity/enterprise traffic
    • Many, many mouse flows
    • High loss tolerance (they are low-bandwidth flows anyway)
  – Science traffic
    • Small number of elephant flows
    • Very sensitive to loss

• Traditional security approaches (i.e. enterprise firewalls) cause performance problems

• If we look at the security implications for science and commodity traffic, what do we see?
Commodity Traffic – Web Browsers

• What does a web browser do?
  – Download a text file from a web server (may be dynamically generated)
  – Render what we download
  – Fetch and render a bunch of other stuff based on links
  – When there is nothing left to fetch and render, the page is done

• What is all this stuff?
  – HTML (fine – it’s a web browser, after all)
  – Mobile code (sometimes useful, sometimes hostile)
  – Images to display
  – Rich media content (Flash and friends)

• Impossible to attribute content to people in practice
Science Traffic (DTN)

- A Data Transfer Node doesn’t run commodity applications
  - Or, at least, it shouldn’t
  - If people are running that goo on your DTN, shut it down
- What does a DTN do?
  - Negotiate a transfer with remote DTN
  - Open a few data connections
  - Push a few TB over those connections
  - Close the connections
- If data is being written, it’s being written by someone with an account (presumably you’ve already vetted that user if you gave them the account)
Commodity Traffic – Attack Surface

• There is a huge attack surface involved with commodity traffic
  – Mobile code execution
  – Media codecs
  – Image rendering libraries

• Go look at what some common web pages do
  – Say your users are into sports, or celebrity gossip, or news blogs
  – Go look at some popular sites (your users do at lunch time!)
    • You don’t have to get fancy – just turn on the javascript console
    • Look at all the places content comes from, what gets executed
      – Mobile code comes down as code, or maybe as text (but gets executed)
      – Images come down as images, or maybe as encoded text (but rendered as images)
      – Content comes from all over the place, depending on who bought what ad space
  – All of this comes over port 80 or port 443 – standard web stuff
Science Traffic – Attack Surface

• There is one listening service – the data transfer tool (assume Globus)
  – Port 2811: control
  – Ports 443 and 7512: oauth (ports depend on config)
  – Large data port range (1001 ports – 50,000 to 51,000)
    • Data ports are open during a transfer
    • Closed otherwise

• Nice clean behavior – just POSIX file operations (open, read, write, close)
  – No image rendering
  – No rich media
  – No document rendering
  – No mobile code
Attack Surface – Which is Greater?

• A web browser typically uses just two ports – 80 and 443
  – Most firewall people count that as two
  – Web is a normal service, critical for business functions
• A Globus DTN uses over 1000 ports
  – Many naïve security people count that as “too many”
  – Weird service, they don’t understand it, and “too many ports”
• A naïve security person will view the DTN as more dangerous because of the high port count
  – This is not rational
• If you look at attack surfaces, the web browser is far more dangerous
  – Web browsers render and execute whatever the ‘net hands them
  – Port count has little to do with an application’s attack surface
To Each Its Own

• The preceding is another example of why science traffic and commodity traffic should be separated

• Enterprise traffic is enterprise traffic, and requires enterprise engineering solutions
  – Sufficient aggregate bandwidth
  – Inexpensive hardware
  – Firewalls
  – Proxies
  – Virus scanners

• Science traffic is science traffic, and requires science engineering solutions
  – Highly capable gear
  – Loss-free IP layer for TCP performance
  – High per-flow bandwidth, and tools that can use it
  – High visibility (perfSONAR)
  – Specific security policy tailored to science applications
Science DMZ – Moving Forward

• This is about the science
• Build it well, make sure they can use it

• ESnet vision:
  – Scientific progress is *completely unconstrained* by the physical location of instruments, people, computational resources, or data
  – Collaborations at every scale, in every domain, will have the *information and tools* they need to achieve maximum benefit from global networks
Thanks!

Eli Dart dart@es.net
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory

http://fasterdata.es.net/
http://my.es.net/
http://www.es.net/
Extra Slides
Science DMZ Security

• Goal – disentangle security policy and enforcement for science flows from security for business systems

• Rationale
  – Science data traffic is simple from a security perspective
  – Narrow application set on Science DMZ
    • Data transfer, data streaming packages
    • No printers, document readers, web browsers, building control systems, financial databases, staff desktops, etc.
  – Security controls that are typically implemented to protect business resources often cause performance problems

• Separation allows each to be optimized
Performance Is A Core Requirement

• Core information security principles
  – Confidentiality, Integrity, Availability (CIA)
  – Often, CIA and risk mitigation result in poor performance

• In data-intensive science, performance is an additional core mission requirement: CIA → PICA
  – CIA principles are important, but if performance is compromised the science mission fails
  – Not about “how much” security you have, but how the security is implemented
  – Need a way to appropriately secure systems without performance compromises
Placement Outside the Firewall

• The Science DMZ resources are placed outside the enterprise firewall for performance reasons
  – The meaning of this is specific – Science DMZ traffic does not traverse the firewall data plane
  – Packet filtering is fine – just don’t do it with a firewall

• Lots of heartburn over this, especially from the perspective of a conventional firewall manager
  – Lots of organizational policy directives mandating firewalls
  – Firewalls are designed to protect converged enterprise networks
  – Why would you put critical assets outside the firewall???

• The answer is that firewalls are typically a poor fit for high-performance science applications
Firewall Internals

• Typical firewalls are composed of a set of processors which inspect traffic in parallel
  – Traffic distributed among processors such that all traffic for a particular connection goes to the same processor
  – Simplifies state management
  – Parallelization scales deep analysis

• Excellent fit for enterprise traffic profile
  – High connection count, low per-connection data rate
  – Complex protocols with embedded threats

• Each processor is a fraction of firewall link speed
  – Significant limitation for data-intensive science applications
  – Overload causes packet loss – performance crashes
Thought Experiment

• We’re going to do a thought experiment
• Consider a network between three buildings – A, B, and C
• This is supposedly a 10Gbps network end to end (look at the links on the buildings)
• Building A houses the border router – not much goes on there except the external connectivity
• Lots of work happens in building B – so much that the processing is done with multiple processors to spread the load in an affordable way, and results are aggregated after
• Building C is where we branch out to other buildings
• Every link between buildings is 10Gbps – this is a 10Gbps network, right???
Notional 10G Network Between Buildings
Clearly Not A 10Gbps Network

• If you look at the inside of Building B, it is obvious from a network engineering perspective that this is not a 10Gbps network
  – Clearly the maximum per-flow data rate is 1Gbps, not 10Gbps
  – However, if you convert the buildings into network elements while keeping their internals intact, you get routers and firewalls
  – What firewall did the organization buy? What’s inside it?
  – Those little 1G “switches” are firewall processors

• This parallel firewall architecture has been in use for years
  – Slower processors are cheaper
  – Typically fine for a commodity traffic load
  – Therefore, this design is cost competitive and common
Notional 10G Network Between Devices
Notional Network Logical Diagram
Firewall Performance Example

• Observed performance, via perfSONAR, through a firewall:

  Almost 20 times slower through the firewall

• Observed performance, via perfSONAR, bypassing firewall:

  Huge improvement without the firewall
What’s Inside Your Firewall?

• Vendor: “but wait – we don’t do this anymore!”
  – It is true that vendors are working toward line-rate 10G firewalls, and some may even have them now
  – 10GE has been deployed in science environments for over 10 years
  – Firewall internals have only recently started to catch up with the 10G world
  – 100GE is being deployed now, 40Gbps host interfaces are available now
  – Firewalls are behind again

• In general, IT shops want to get 5+ years out of a firewall purchase
  – This often means that the firewall is years behind the technology curve
  – Whatever you deploy now, that’s the hardware feature set you get
  – When a new science project tries to deploy data-intensive resources, they get whatever feature set was purchased several years ago
The Firewall State Table

• Many firewalls use a state table to improve performance
  – State table lookup is fast
  – No need to process entire ruleset for every packet
  – Also allows session tracking (e.g. TCP sequence numbers)

• State table built dynamically
  – Incoming packets are matched against the state table
  – If no state table entry, go to the ruleset
  – If permitted by ruleset, create state table entry
  – Remove state table entry after observing connection teardown

• Semantically similar to punt-and-switch model of traffic forwarding used on many older routers
State Table Issues

- If the state table is not pruned, it will overflow
  - Not all connections close cleanly
    - I shut my laptop and go to a meeting
    - Software crashes happen
  - Some attacks try to fill state tables

- Solution: put a timer on state table entries
  - When a packet matches the state table entry, update the timer
  - If the timer expires, delete the state table entry

- What if I just pause for a few minutes?
  - This turns out to be a problem – state table timers are typically in the 5-15 minute range, while host keepalive timers are 2 hours
  - If a connection pauses (e.g. control channel waits for a large transfer), the firewall will delete the state table entry from under it – the control connection now hangs
  - We have seen this in production environments
Firewall Capabilities and Science Traffic

- Firewalls have a lot of sophistication in an enterprise setting
  - Application layer protocol analysis (HTTP, POP, MSRPC, etc.)
  - Built-in VPN servers
  - User awareness
- Data-intensive science flows typically don’t match this profile
  - Common case – data on filesystem A needs to be on filesystem Z
    - Data transfer tool verifies credentials over an encrypted channel
    - Then open a socket or set of sockets, and send data until done (1TB, 10TB, 100TB, …)
  - One workflow can use 10% to 50% or more of a 10G network link
- Do we have to use a firewall?
Firewalls As Access Lists

• When you ask a firewall administrator to allow data transfers through the firewall, what do they ask for?
  – IP address of your host
  – IP address of the remote host
  – Port range
  – *That looks like an ACL to me!*

• No special config for advanced protocol analysis – just address/port

• Router ACLs are better than firewalls at address/port filtering
  – ACL capabilities are typically built into the router
  – Router ACLs typically do not drop traffic permitted by policy
Security Without Firewalls

• Data intensive science traffic interacts poorly with firewalls
• Does this mean we ignore security? **NO!**
  – We **must** protect our systems
  – We just need to find a way to do security that does not prevent us from getting the science done

• **Key point** – *security policies and mechanisms that protect the Science DMZ should be implemented so that they do not compromise performance*

• Traffic permitted by policy should not experience performance impact as a result of the application of policy
If Not Firewalls, Then What?

• Remember – the goal is to protect systems in a way that allows the science mission to succeed

• I like something I heard at NERSC – paraphrasing: “Security controls should enhance the utility of science infrastructure.”

• There are multiple ways to solve this – some are technical, and some are organizational/sociological

• I’m not going to lie to you – this is harder than just putting up a firewall and closing your eyes
Other Technical Capabilities

• Intrusion Detection Systems (IDS)
  – Bro is high-performance and battle-tested
    • Bro protects several high-performance national assets
    • Bro can be scaled with clustering: [http://www.bro-ids.org/documentation/cluster.html](http://www.bro-ids.org/documentation/cluster.html)
  – Other IDS solutions are available also

• Netflow and IPFIX can provide intelligence, but not filtering

• Openflow and SDN
  – Using Openflow to control access to a network-based service seems pretty obvious
  – This could significantly reduce the attack surface for any authenticated network service
  – This would only work if the Openflow device had a robust data plane
Other Technical Capabilities (2)

• Aggressive access lists
  – More useful with project-specific DTNs
  – If the purpose of the DTN is to exchange data with a small set of remote collaborators, the ACL is pretty easy to write
  – Large-scale data distribution servers are hard to handle this way (but then, the firewall ruleset for such a service would be pretty open too)

• Limitation of the application set
  – One of the reasons to limit the application set in the Science DMZ is to make it easier to protect
  – Keep desktop applications off the DTN (and watch for them anyway using logging, netflow, etc – take violations seriously)
  – This requires collaboration between people – networking, security, systems, and scientists
Collaboration Within The Organization

• All stakeholders should collaborate on Science DMZ design, policy, and enforcement

• The security people have to be on board
  – Remember: security people already have political cover – it’s called the firewall
  – If a host gets compromised, the security officer can say they did their due diligence because there was a firewall in place
  – If the deployment of a Science DMZ is going to jeopardize the job of the security officer, expect pushback

• The Science DMZ is a strategic asset, and should be understood by the strategic thinkers in the organization
  – Changes in security models
  – Changes in operational models
  – Enhanced ability to compete for funding
  – Increased institutional capability – greater science output
Commodity vs. Science Traffic

- Stark difference in behavior
- Commodity traffic
  - When there are eyeballs, there is traffic
  - No eyeballs, no traffic
  - Web, email, etc.
  - Many, many, many mouse flows
- Science traffic
  - When there is data to move, there is traffic
  - Science facilities run 24x7
  - Small number of elephant flows
  - Individual workflows are sometimes visible in aggregate statistics
Commodity Traffic: Peering Interface
Science Traffic: Peering Interface
Security Footprint of a Globus Transfer

TCP ports 50000-51000

TCP ports 443, 2811, 7512

TCP ports 443, 2811, 7512

Logical data path

Physical data path

Logical control path

Physical control path

<table>
<thead>
<tr>
<th>Src Address</th>
<th>Src Port</th>
<th>Dst Address</th>
<th>Dst Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab1 DTN</td>
<td>TCP 50000-51000</td>
<td>Lab2 DTN</td>
<td>TCP 50000-51000</td>
</tr>
<tr>
<td>Lab1 DTN</td>
<td>TCP 443, 2811, 7512</td>
<td>Globus Cloud</td>
<td>TCP unprivileged</td>
</tr>
<tr>
<td>Lab2 DTN</td>
<td>TCP 50000-51000</td>
<td>Lab1 DTN</td>
<td>TCP 50000-51000</td>
</tr>
<tr>
<td>Globus Cloud</td>
<td>TCP unprivileged</td>
<td>Lab1 DTN</td>
<td>TCP 443, 2811, 7512</td>
</tr>
</tbody>
</table>
Security Footprint of a Globus DTN

**Diagram:**
- Local DTN
- Science DMZ
- Remote DTNs
- World
- DTN security filters
- Orchestration
- TCP ports 443, 2811, 7512

**Table:**

<table>
<thead>
<tr>
<th>Src Address</th>
<th>Src Port</th>
<th>Dst Address</th>
<th>Dst Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local DTN</td>
<td>TCP 50000-51000</td>
<td>World</td>
<td>TCP 50000-51000</td>
</tr>
<tr>
<td>Local DTN</td>
<td>TCP 443, 2811, 7512</td>
<td>Globus Cloud</td>
<td>TCP unprivileged</td>
</tr>
<tr>
<td>World</td>
<td>TCP 50000-51000</td>
<td>Local DTN</td>
<td>TCP 50000-51000</td>
</tr>
<tr>
<td>Globus Cloud</td>
<td>TCP unprivileged</td>
<td>Local DTN</td>
<td>TCP 443, 2811, 7512</td>
</tr>
</tbody>
</table>