The Science DMZ – Introduction & Architecture

Jason Zurawski - ESnet Engineering & Outreach

Operating Innovative Networks (OIN)

October 3rd & 4th, 2013
Introduction & Purpose

- The "Campus Cyberinfrastructure - Network Infrastructure and Engineering (CC-NIE)" program:
  - Invests in improvements and re-engineering at the campus level to support a range of data transfers supporting computational science and computer networks and systems research.
  - Supports Network Integration activities tied to achieving higher levels of performance, reliability and predictability for science applications and distributed research projects.

- The bolded items can be tricky: this series of talks will introduce some broad concepts that will help:
  - Capable network architectures
  - Federated End-to-End monitoring
  - Advanced data movement tools and procedures
  - We will not be digging too deep technically, but deep enough to give ‘hit the ground running’ experience. We encourage everyone to take discussions to the mailing list & forums:
    - [http://fasterdata.es.net/forums/](http://fasterdata.es.net/forums/)
    - [https://gab.es.net/mailman/listinfo/sciencedmz](https://gab.es.net/mailman/listinfo/sciencedmz)
What is there to worry about?

• **Genomics**
  - Sequencer data volume increasing 12x over the next 3 years
  - Sequencer cost decreasing by 10x over same time period

• **High Energy Physics**
  - LHC experiments produce & distribute petabytes of data/year
  - Peak data rates increase 3-5x over 5 years

• **Light Sources**
  - Many detectors on a Moore’s Law curve
  - Data volumes rendering previous operational models obsolete

• **Common Threads**
  - Increased capability, greater need for data mobility due to span/depth of collaboration space
  - Global is the new local. Research is no longer done within a domain. End to end involves many fiefdoms to cross – and yes this becomes your problem when your users are impacted
Overview

Part 1 (Today):

• What is ESnet?
• Science DMZ Introduction & Motivation
• Science DMZ Architecture

Part 2 (Today):

• PerfSONAR
• Science DMZ Security Best Practices

Part 3 (Today & Tomorrow):

• The Data Transfer Node
• Data Transfer Tools
• Conclusions & Discussion
What is ESnet?

- A high-performance network linking DOE Office of Science researchers to global collaborators and resources around the world, including:
  - Supercomputer centers
  - User Facilities
  - Multi-program labs
  - Universities
  - Connectivity to Internet and Cloud providers

- A national DOE user facility providing:
  - Tailored data mobility solutions for science
  - *Dedicated Science Engagement team to support researchers*
  - Collaboration services e.g. audio/video conferencing
The Office of Science supports:

- 27,000 Ph.D.s, graduate students, undergraduates, engineers, and technicians
- 26,000 users of open-access facilities
- 300 leading academic institutions
- 17 DOE laboratories
The Science Data Explosion

ESnet Accepted Traffic: Jan 1990 - May 2013 (Log Scale)

Projected volume for May 2014: 48.1 PB
Actual volume for May 2013: 11.0 PB

Expecting 100 Petabytes/month of data in 2015
• The capabilities required to support scientific data movement involve hardware and software developments at all levels:
  1. Optical signal transport
  2. Network routers and switches
  3. Data transport (TCP is still the norm)
  4. Network monitoring and testing
  5. Operating system evolution
  6. Data movement and management techniques and software
  7. Evolution of network architectures
  8. New network services

• Technology advances in these areas have resulted in today’s state-of-the-art that makes it possible for science to continue innovating

Topics we will explore over the next 2 days
Wide Area Network is Engineered for Elephants

Science data

ESnet Traffic (last 24 hours)

Total Traffic
OSCARS Traffic

perfSONAR
powered

Google
Netflix
hulu
YouTube

100 G

perfSONAR
powered

Copyright (c) 2012 AMS-IX B.V. Updated: 31-Jan-2012 06:15:35 +00:00
Time to Raise Your Network Expectations:
Time to Copy 1 Terabyte

On a...
- 10 Mbps network: 300 hrs (12.5 days)
- 100 Mbps network: 30 hrs
- 1 Gbps network: 3 hrs (are your disks fast enough?)
- 10 Gbps network: 20 minutes (need really fast disks and filesystems)

• These figures assume some headroom left for other users
  - Compare these speeds to:
    • USB 2.0 portable disk

  » 20-30 hours to load 1 Terabyte

See Also: http://what-if.xkcd.com/31/
## Bandwidth Requirements to move Y Bytes of data in Time X

<table>
<thead>
<tr>
<th></th>
<th>10PB</th>
<th>1PB</th>
<th>100TB</th>
<th>10TB</th>
<th>1TB</th>
<th>100GB</th>
<th>10GB</th>
<th>1GB</th>
<th>100MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bytes</td>
<td>25,020.0 Gbps</td>
<td>3,127.5 Gbps</td>
<td>1,042.5 Gbps</td>
<td>148.9 Gbps</td>
<td>34.7 Gbps</td>
<td>2,502.0 Gbps</td>
<td>312.7 Gbps</td>
<td>104.2 Gbps</td>
<td>14.9 Gbps</td>
</tr>
<tr>
<td></td>
<td>2,443.6 Gbps</td>
<td>305.4 Mbps</td>
<td>101.8 Mbps</td>
<td>145.4 Mbps</td>
<td>33.9 Mbps</td>
<td>244.3 Gbps</td>
<td>30.5 Gbps</td>
<td>10.2 Gbps</td>
<td>1.5 Gbps</td>
</tr>
<tr>
<td></td>
<td>24.4 Gbps</td>
<td>3.1 Gbps</td>
<td>1.0 Gbps</td>
<td>145.4 Mbps</td>
<td>33.9 Mbps</td>
<td>24.4 Gbps</td>
<td>3.1 Gbps</td>
<td>1.0 Gbps</td>
<td>145.4 Mbps</td>
</tr>
<tr>
<td></td>
<td>2.4 Gbps</td>
<td>30.54 Mbps</td>
<td>101.8 Mbps</td>
<td>145.4 Mbps</td>
<td>33.9 Mbps</td>
<td>2.4 Gbps</td>
<td>30.54 Mbps</td>
<td>101.8 Mbps</td>
<td>145.4 Mbps</td>
</tr>
<tr>
<td></td>
<td>238.6 Mbps</td>
<td>29.8 Mbps</td>
<td>9.9 Mbps</td>
<td>14.5 Mbps</td>
<td>3.4 Mbps</td>
<td>238.6 Mbps</td>
<td>29.8 Mbps</td>
<td>9.9 Mbps</td>
<td>14.5 Mbps</td>
</tr>
<tr>
<td></td>
<td>23.9 Mbps</td>
<td>3.0 Mbps</td>
<td>994.2 Kbps</td>
<td>142.0 Kbps</td>
<td>33.1 Kbps</td>
<td>23.9 Mbps</td>
<td>3.0 Mbps</td>
<td>994.2 Kbps</td>
<td>142.0 Kbps</td>
</tr>
<tr>
<td></td>
<td>2.4 Mbps</td>
<td>298.3 Kbps</td>
<td>99.4 Kbps</td>
<td>14.2 Kbps</td>
<td>3.3 Kbps</td>
<td>2.4 Mbps</td>
<td>298.3 Kbps</td>
<td>99.4 Kbps</td>
<td>14.2 Kbps</td>
</tr>
<tr>
<td></td>
<td>233.0 Kbps</td>
<td>29.1 Kbps</td>
<td>9.7 Kbps</td>
<td>1.4 Kbps</td>
<td>0.3 Kbps</td>
<td>233.0 Kbps</td>
<td>29.1 Kbps</td>
<td>9.7 Kbps</td>
<td>1.4 Kbps</td>
</tr>
</tbody>
</table>

This table available at [http://fasterdata.es.net](http://fasterdata.es.net)
Use Case = End to End Exchange

• Alice & Bob are collaborators
  o Experts in their field
  o Physically separated (common)
  o Rely on networks, but are not IT experts (common & expected)
  o They know their local IT staff. May also have an adversarial relationship with them (e.g. Alice and Bob are ‘troublemakers’ since they use the network, and expect it to work)

• Alice & Bob want to embark on a new project
  o Instrumentation @ one end, processing/analysis @ the other
  o Keep in mind they know about the science, not about the technology in the middle
  o Use infrastructure they are comfortable with, perhaps cobbled together by local support staff
Meet Alice & Bob - They are Science Cats

Stand back! Iz goin to do science!

MOAR DATA 4 U

INTERNETS

HAI, I CAN K THY HAZ DATA?

WAIT, I GET IT 4 U

© 2013 icanhascheezburger.com

© 2013 Susan's School Days

13 – ESnet Science Engagement (engage@es.net) - 10/2/13
Expectations & Realities

"In any large system, there is always something broken."

Jon Postel

• Modern networks are large and complicated
• Many users will encounter unforeseen (and therefore challenging) situations:
  o Upgrading networks breaks them (loss of configuration, etc.)
  o Synergy between the new and the old
  o Statistical anomalies, e.g. that 7 year old interface will stop working eventually…

• Mitigating the risk can (and should) be done in a number of ways:
  o Analysis and alteration to architecture
  o Careful thought to security/data policies in target areas
  o Integration of software designed to exercise the network, and alert/visualize
  o Capable hosts and tools for scientific activities
Meet Alice & Bob; Sad Reality

15 – ESnet Science Engagement (engage@es.net) - 10/2/13
Which leads us to the Science DMZ…

- Significant commonality in issues encountered with science collaborations … and similar solution set
  - The causes of poor data transfer performance fit into a few categories with similar solutions
    - Un-tuned/under-powered hosts
    - Packet loss issues
    - Security devices
  - A successful model has emerged – the Science DMZ
    - This model successfully in use by CMS/ATLAS, ESG, NERSC, ORNL, ALS, and others

- The Science DMZ is a design pattern for network design.
  - Not all implementations look the same, but share common features
  - Some choices don’t make sense for everyone, caveat emptor
The Science DMZ in 1 Slide

Consists of **three key components**, all required:

“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 if possible

Dedicated, high-performance Data Transfer Nodes (DTNs)
- Hardware, operating system, libraries all optimized for transfer
- Includes optimized data transfer tools such as Globus Online and GridFTP

Performance measurement/test node
- perfSONAR

Details at [http://fasterdata.es.net/science-dmz/](http://fasterdata.es.net/science-dmz/)
Overview

Part 1 (Today):

• What is ESnet?
• Science DMZ Introduction & Motivation
• Science DMZ Architecture

Part 2 (Today):

• PerfSONAR
• Science DMZ Security Best Practices

Part 3 (Today & Tomorrow):

• The Data Transfer Node
• Data Transfer Tools
• Conclusions & Discussion
Science DMZ Background

The data mobility performance requirements for data intensive science are beyond what can typically be achieved using traditional methods

- Default host configurations (TCP, filesystems, NICs)
- Converged network architectures designed for commodity traffic
- Conventional security tools and policies
- Legacy data transfer tools (e.g. SCP)
- Wait-for-trouble-ticket operational models for network performance

The Science DMZ model describes a performance-based approach

- Dedicated infrastructure for wide-area data transfer
  - Well-configured data transfer hosts with modern tools
  - Capable network devices
  - High-performance data path which does not traverse commodity LAN
- Proactive operational models that enable performance
  - Well-deployed test and measurement tools (perfSONAR)
  - Periodic testing to locate issues instead of waiting for users to complain
- Security posture well-matched to high-performance science applications
Motivation

Science data increasing both in volume and in value
- Higher instrument performance
- Increased capacity for discovery
- Analyses previously not possible

Lots of promise, but only if scientists can actually work with the data
- Data has to get to analysis resources
- Results have to get to people
- People have to share results

Common pain point – data mobility
- Movement of data between instruments, facilities, analysis systems, and scientists is a gating factor for much of data intensive science
- Data mobility is not the only part of data intensive science – not even the most important part
- However, without data mobility data intensive science is hard

We need to move data – how can we do it consistently well?
Motivation (2)

Networks play a crucial role

- The very structure of modern science assumes science networks exist – high performance, feature rich, global scope
- Networks enable key aspects of data intensive science
  - Data mobility, automated workflows
  - Access to facilities, data, analysis resources

Messing with “the network” is unpleasant for most scientists

- Not their area of expertise
- Not where the value is (no papers come from messing with the network)
- Data intensive science is about the science, not about the network
- However, it’s a critical service – if the network breaks, everything stops

Therefore, infrastructure providers must cooperate to build consistent, reliable, high performance network services for data mobility

Here we describe a design pattern – the Science DMZ model – that works well in a variety of environments
Science DMZ Origins

ESnet has a lot of experience with different scientific communities at multiple data scales – e.g.
http://www.es.net/about/science-requirements/network-requirements-reviews/

N.B - If the above interests you, lets talk in the ‘community discussion’ tomorrow

Significant commonality in the issues encountered, and solution set

• The causes of poor data transfer performance fit into a few categories with similar solutions
  – Un-tuned/under-powered hosts and disks, packet loss issues, security devices

• A successful model has emerged – the Science DMZ
  – This model successfully in use by HEP (CMS/Atlas), Climate (ESG), several Supercomputer Centers, and others
Soft Network Failures

Soft failures are where basic connectivity functions, but high performance is not possible.

TCP was intentionally designed to hide all transmission errors from the user:

• “As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users.” (From RFC793, 1981)

Some soft failures only affect high bandwidth long RTT flows.

Hard failures are easy to detect & fix

• soft failures can lie hidden for years!

One network problem can often mask others
TCP Background

Networks provide connectivity between hosts – how do hosts see the network?

- From an application’s perspective, the interface to “the other end” is a socket or similar construct
- The vast majority of data transfer applications use TCP
- Communication is between applications – mostly over TCP

TCP – the fragile workhorse

- TCP is (for very good reasons) timid – packet loss is interpreted as congestion
- TCP has very limited ability to diagnose problems within the network (all it can do is measure packet loss and round trip time)
- Packet loss in conjunction with latency is a performance killer
- Like it or not, TCP is used for the vast majority of data transfer applications – we’re stuck with TCP
TCP Background (2)

It is far easier to architect the network to support TCP than it is to fix TCP

• People have been trying to fix TCP for years – limited success
• Here we are – packet loss is still the number one performance killer in long distance high performance environments

Pragmatically speaking, we must accommodate TCP

• Implications for equipment selection
  – Ability to provide loss-free IP service to TCP
  – Ability to accurately account for packets (aids loss localization)

• Implications for network architecture, deployment models
  – Infrastructure must be designed to allow easy troubleshooting
  – Test and measurement tools are critical – they have to be deployed
Common Soft Failures

Random Packet Loss
- Bad/dirty fibers or connectors – CRC error count is often related to this.
  - Note – ‘brand new’ jumpers need to be cleaned and sometimes scoped too …
- Low light levels due to amps/interfaces failing
- Duplex mismatch

Small Router/Switch Buffers
- Switches not able to handle the long packet trains prevalent in long RTT sessions and local cross traffic at the same time

Un-intentional Rate Limiting
- Processor-based switching on routers due to faults, ACL’s, or mis-configuration
A small amount of packet loss makes a huge difference in TCP performance

Throughput vs. increasing latency on a 10Gb/s link with 0.0046% packet loss

- On a 10 Gb/s LAN path the impact of low packet loss rates is minimal
- On a 10Gb/s WAN path the impact of low packet loss rates is enormous

**Implications:** error-free paths are essential for high-volume data transfers
How Do We Accommodate TCP?

High-performance wide area TCP flows must get loss-free service

- Sufficient bandwidth to avoid congestion
- Deep enough buffers in routers and switches to handle bursts
  - Especially true for long-distance flows due to packet behavior
  - No, this isn’t buffer bloat

Equally important – the infrastructure must be verifiable so that clean service can be provided

- Stuff breaks
  - Hardware, software, optics, bugs, …
  - How do we deal with it in a production environment?
- Must be able to prove a network device or path is functioning correctly
  - Regular active test should be run - perfSONAR
- Small footprint is a huge win
  - Fewer the number of devices = easier to locate the source of packet loss
Sample Soft Failures

Gradual failure of optical line card

Rebooted router with full route table

Normal performance

Degraded performance

Repair

Source: nersc-pt1.es.net (198.129.254.22) -- Destination: sunn-pt1.es.net

Gb/s

Source -> Destination in Gbps

Destination -> Source in Gbps

One month
Congestion on Link + Drifting Clock

Note that SNMP Poling Intervals could mask traffic bursts – OWAMP is active testing, and thus more indicative of true network behavior.
Expansion of ‘The Network’

Once the network is error-free, there is still the issue of efficiently moving data from the application running on a user system onto the network:

- Host TCP tuning
- Modern TCP stack
- Other issues (MTU, etc.)
- Data transfer tools and parallelism
- Other data transfer issues (firewalls, etc.)

“TCP tuning” commonly refers to the proper configuration of TCP windowing buffers for the path length:

- It is critical to use the optimal TCP send and receive socket buffer sizes for the path (RTT) you are using end-to-end
- Default TCP buffer sizes are typically much too small for today’s high speed networks
- Until recently, default TCP send/receive buffers were typically 64 KB
- Tuned buffer to fill CA to NY, 1 Gb/s path: 10 MB
  - 150X bigger than the default buffer size
Autotuning?

Throughput out to ~9000 km on a 10Gb/s network
32 MBy (autotuned) vs. 64 MBy (hand tuned) TCP window size

Effect of Window Size on Throughput vs. Increasing Latency

- 32 MBy window
- 64 MBy window

Round Trip Time (milliseconds) vs. Path Length, km

Throughput (Mbits/sec)

Path Length, km

No Packet Loss (hand tuned to 64MB)
No Packet Loss (autotuning to 32MB - the fasterdata.es.net...)
cumulative path length, km (approximate)
Linking it Together: Abstraction Helps & Hurts
Abstraction Helps & Hurts

MY INTERNETS

THEIR INTERNETS

OTHER INTERNETS

MOAR INTERNETS

??? INTERNETS

YOUR INTERNETS
Abstraction Helps & Hurts

The Seven Layers of OSI

- Physical Layer
- Data Link Layer
- Network Layer
- Transport Layer
- Session Layer
- Presentation Layer
- Application Layer

Physical Link

User

Transmit Data

Receive Data
Abstraction Helps & Hurts

Protocols, Etc.
Failure in the Layers

Redundancy Helps With Redundancy

Selected Connections Fail! But it's hard to fly.

IM DROPN UR PACKETS LOL

THPPTH!
Solution Space

• Basic idea:
  o Architectural changes
  o Solution for Monitoring/Emulation of User Behavior
  o Workflow Analysis/Adoption of New Tools

  o Architecture
    o Split out enterprise concerns from data intensive ones
    o Directed security policies, instead of blanket enforcement

• Monitoring:
  • Dedicated resources at different vantage points in the network
    o Running some standard and useful types of measurement
    o Integrated with tools that allow you to see/hear when a problem arises

  o Data Movement Solutions
    o Dedicated servers
    o High performance applications
Overview

Part 1 (Today):

• What is ESnet?
• Science DMZ Introduction & Motivation
• Science DMZ Architecture

Part 2 (Today):

• PerfSONAR
• Science DMZ Security Best Practices

Part 3 (Today & Tomorrow):

• The Data Transfer Node
• Data Transfer Tools
• Conclusions & Discussion
Traditional DMZ

DMZ – “Demilitarized Zone”

- Network segment near the site perimeter with different security policy
- Commonly used architectural element for deploying WAN-facing services (e.g. email, DNS, web)

Traffic for WAN-facing services does not traverse the LAN

- WAN flows are isolated from LAN traffic
- Infrastructure for WAN services is specifically configured for WAN

Separation of security policy improves both LAN and WAN

- No conflation of security policy between LAN hosts and WAN services
- DMZ hosts provide specific services
- LAN hosts must traverse the same ACLs as WAN hosts to access DMZ
The Data Transfer Trifecta: The “Science DMZ” Model

**Dedicated Systems for Data Transfer**
- High performance
- Configured for data transfer
- Proper tools

**Science DMZ**
- Dedicated location for DTN
- Proper security
- Easy to deploy - no need to redesign the whole network

**perfSONAR**
- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

Data Transfer Node

Science DMZ

perfSONAR

41 – ESnet Science Engagement (engage@es.net) - 10/2/13
Science DMZ Takes Many Forms

There are a lot of ways to combine these things – it all depends on what you need to do

• Small installation for a project or two
• Facility inside a larger institution
• Institutional capability serving multiple departments/divisions
• Science capability that consumes a majority of the infrastructure

Some of these are straightforward, others are less obvious

Key point of concentration: eliminate sources of packet loss / packet friction
The Data Transfer Trifecta: The “Science DMZ” Model

Dedicated Systems for Data Transfer
- High performance
- Configured for data transfer
- Proper tools

Science DMZ
- Dedicated location for DTN
- Proper security
- Easy to deploy - no need to redesign the whole network

Performance Testing & Measurement
- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

Data Transfer Node
- High performance
- Configured for data transfer
- Proper tools
Ad Hoc DTN Deployment

This is often what gets tried first

Data transfer node deployed where the owner has space

• This is often the easiest thing to do at the time
• Straightforward to turn on, hard to achieve performance

If present, perfSONAR is at the border

• This is a good start
• Need a second one next to the DTN

Entire LAN path has to be sized for data flows

Entire LAN path is part of any troubleshooting exercise

This usually fails to provide the necessary performance.
Ad Hoc DTN Deployment

DTN traffic subject to firewall limitations

WAN

10G

Test and measurement not aligned with data resource placement

Site Border Router

Global security policy mixes rules for science and business traffic

Perimeter Firewall

10GE

Conflicting requirements result in performance compromises

Site / Campus LAN

High performance Data Transfer Node with high-speed storage

Building or Wiring Closet Switch/Router

DTN traffic subject to limitations of general-purpose networking equipment/config

Note: Site border router and perimeter firewall are often the same device
Multiple Ingress Data Flows, Common Egress

Hosts will typically send packets at the speed of their interface (1G, 10G, etc.)

- Instantaneous rate, not average rate
- If TCP has window available and data to send, host sends until there is either no data or no window

Hosts moving big data (e.g. DTNs) can send large bursts of back-to-back packets

- This is true even if the average rate as measured over seconds is slower (e.g. 4Gbps)
- On microsecond time scales, there is often congestion
- Router or switch must queue packets or drop them
Output Queue Drops – Common Locations

- Site Border Router
- Site Core Switch/Router
- Department uplink to site core constrained by budget or legacy equipment
- Inbound data path
- Common location of output queue drops for traffic inbound from the WAN
- Outbound data path
- Common locations of output queue drops for traffic outbound toward the WAN
- Wiring closet switch
- Workstations
- 32+ cluster nodes
- Cluster data transfer node
- Department cluster switch
- WAN

47 – ESnet Science Engagement (engage@es.net) - 10/2/13
Router and Switch Output Queues

Interface output queue allows the router or switch to avoid causing packet loss in cases of momentary congestion

In network devices, queue depth (or ‘buffer’) is often a function of cost

- Cheap, fixed-config LAN switches (especially in the 10G space) have inadequate buffering. Imagine a 10G ‘data center’ switch as the guilty party
- Cut-through or low-latency Ethernet switches typically have inadequate buffering (the whole point is to avoid queuing!)

Expensive, chassis-based devices are more likely to have deep enough queues

- Juniper MX and Alcatel-Lucent 7750 used in ESnet backbone
- Other vendors make such devices as well - details are important
- Thx to Jim: [http://people.ucsc.edu/~warner/buffer.html](http://people.ucsc.edu/~warner/buffer.html)

This expense is one driver for the Science DMZ architecture – only deploy the expensive features where necessary
Small-scale Science DMZ Deployment

Add-on to existing network infrastructure
- All that is required is a port on the border router
- Small footprint, pre-production commitment

Easy to experiment with components and technologies
- DTN prototyping
- perfSONAR testing

Limited scope makes security policy exceptions easy
- Only allow traffic from partners
- Add-on to production infrastructure – lower risk
A better approach: simple Science DMZ

- Border Router
  - Clean, High-bandwidth WAN path
  - Site / Campus access to Science DMZ resources

- Science DMZ Switch/Router
  - Per-service security policy control points

- High performance Data Transfer Node with high-speed storage

WAN

10G

Enterprise Border Router/Firewall

10GE

Site / Campus LAN

10GE

perfSONAR

10GE

WAN

perfSONAR

10G

10GE

Site / Campus LAN

10GE

Science DMZ Switch/Router

10GE

High performance Data Transfer Node with high-speed storage

10GE

perfSONAR

10GE

WAN

perfSONAR

10G

10GE

Enterprise Border Router/Firewall

10GE

Site / Campus LAN

10GE

Science DMZ Switch/Router

10GE

High performance Data Transfer Node with high-speed storage

10GE

perfSONAR

10GE

WAN

perfSONAR

10G

10GE

Enterprise Border Router/Firewall

10GE

Site / Campus LAN

10GE

Science DMZ Switch/Router

10GE

High performance Data Transfer Node with high-speed storage
Prototype Science DMZ Data Path

Border Router

Enterprise Border Router/Firewall

Science DMZ Switch/Router

Clean, High-bandwidth WAN path

Site / Campus access to Science DMZ resources

Per-service security policy control points

High performance Data Transfer Node with high-speed storage

High Latency WAN Path

Low Latency LAN Path

perfSONAR
Prototype With Virtual Circuits

Small virtual circuit prototype can be done in a small Science DMZ

- Perfect example is a Software Defined Networking (SDN) testbed
- Virtual circuit connection may or may not traverse border router

As with any Science DMZ deployment, this can be expanded as need grows

In this particular diagram, Science DMZ hosts can use either the routed or the circuit connection
Virtual Circuit Prototype Deployment

Border Router

Clean, High-bandwidth path to/from WAN
Dedicated path for virtual circuit traffic
Per-service security policy control points

High performance Data Transfer Node with high-speed storage

WAN

10G Routed

10G Virtual Circuit

Science DMZ Switch/Router

Site / Campus access to Science DMZ resources

10GE

Enterprise Border Router/Firewall

Site / Campus LAN

perfSONAR

perfSONAR

perfSONAR

53 – ESnet Science Engagement (engage@es.net) - 10/2/13
Research Project Requirements

Science DMZ model used to support research

Some research projects are networking research projects

• The network is both the environment and the subject of research
• Science DMZ is a good fit for several reasons
  – Isolate research from production when research is in the unstable phase
  – Separation of administrative control

Some research projects need high-performance end to end networking, but are not network research

• HEP/LHC, Astronomy, “Big Data,” etc.
• The Science DMZ is production cyberinfrastructure

Ideally, both network research and production data-intensive science could coexist
Science DMZ With Separate Research Area

Border Router

Enterprise Border Router/Firewall

Production WAN path

Site / Campus access to Science DMZ resources

Site / Campus LAN

Research WAN path

Production WAN path

Production Science DMZ Switch/Router

Research Science DMZ Switch/Router

Per-service security policy control points

Science DMZ Connections

Production DTN

Research DTN powered

PerfSONAR

PerfSONAR

PerfSONAR

55 – ESnet Science Engagement (engage@es.net) - 10/2/13

Lawrence Berkeley National Laboratory
Science DMZ – Flexible Design Pattern

The Science DMZ design pattern is highly adaptable to research Deploying a research Science DMZ is straightforward

• The basic elements are the same
  – Capable infrastructure designed for the task
  – Test and measurement to verify correct operation
  – Security policy well-matched to the environment, application set is strictly limited to reduce risk

• Connect the research DMZ to other resources as appropriate

The same ideas apply to supporting an SDN effort

• Test/research areas for development
• Transition to production as technology matures and need dictates
• One possible trajectory follows…
Science DMZ – Separate SDN Connection

SDN Science DMZ Switch/Router

Per-service security policy control points

High performance routed path

Site / Campus access to Science DMZ resources

WAN

Border Router

Enterprise Border Router/Firewall

Production Science DMZ Switch/Router

Science DMZ Connections

Research DTN

58 – ESnet Science Engagement (engage@es.net) - 10/2/13
Science DMZ – Production SDN Connection

- WAN
- Border Router
- Enterprise Border Router/Firewall
- Site / Campus LAN
- Science DMZ Switch/Router
- Production SDN Switch/Router
- Research SDN Switch/Router
- Production DTN
- Research DTN

Key Points:
- High performance routed path
- Site / Campus access to Science DMZ resources
- Per-service security policy control points
- perfSONAR connections

Diagram shows the integration of SDN (Software-Defined Networking) for high performance routed paths and security control points in a network infrastructure.
Science DMZ – SDN Campus Border

Border Router

Enterprise Border Router/Firewall

WAN

Site / Campus LAN

Production DTN

Research DTN

Science DMZ

Switch/Router

Production SDN

Science DMZ

Switch/Router

Research Science DMZ

Switch/Router

High performance multi-service path

Site / Campus access to Science DMZ resources

Per-service security policy control points

Science DMZ Connections

60 – ESnet Science Engagement (engage@es.net) - 10/2/13

60 – ESnet Science Engagement (engage@es.net) - 10/2/13
Support For Multiple Projects

Science DMZ architecture allows multiple projects to put DTNs in place

• Modular architecture
• Centralized location for data servers

This may or may not work well depending on institutional politics

• Issues such as physical security can make this a non-starter
• On the other hand, some shops already have service models in place

On balance, this can provide a cost savings – it depends

• Central support for data servers vs. carrying data flows
• How far do the data flows have to go?
Multiple Projects

Border Router

Enterprise Border Router/Firewall

WAN

Clean, High-bandwidth WAN path

Site / Campus access to Science DMZ resources

Science DMZ Switch/Router

Project A DTN

Project B DTN

Project C DTN

Site / Campus LAN

Per-project security policy control points

perfSONAR

perfSONAR powered

62 – ESnet Science Engagement (engage@es.net) - 10/2/13
Supercomputer Center Deployment

High-performance networking is assumed in this environment

- Data flows between systems, between systems and storage, wide area, etc.
- Global filesystem often ties resources together
  - Portions of this may not run over Ethernet (e.g. IB)
  - Implications for Data Transfer Nodes

“Science DMZ” may not look like a discrete entity here

- By the time you get through interconnecting all the resources, you end up with most of the network in the Science DMZ
- This is as it should be – the point is appropriate deployment of tools, configuration, policy control, etc.

Office networks can look like an afterthought, but they aren’t

- Deployed with appropriate security controls
- Office infrastructure need not be sized for science traffic
Supercomputer Center Data Path

- WAN
- Border Router
- Firewall
- Core Switch/Router
- Front switch
- Front end switch
- Supercomputer
- Data Transfer Nodes
- Parallel Filesystem

Lines represent:
- High Latency WAN Path
- Low Latency LAN Path
- High Latency VC Path

(perfSONAR powered)

65 – ESnet Science Engagement (engage@es.net) - 10/2/13
Major Data Site Deployment

In some cases, large scale data service is the major driver

- Huge volumes of data – ingest, export
- Individual DTNs don’t exist here – data transfer clusters

Single-pipe deployments don’t work

- Everything is parallel
  - Networks (Nx10G LAGs, soon to be Nx100G)
  - Hosts – data transfer clusters, no individual DTNs
  - WAN connections – multiple entry, redundant equipment
- Choke points (e.g. firewalls) cause problems

66 – ESnet Science Engagement (engage@es.net) - 10/2/13
Data Site – Architecture

WAN

Provider Edge Routers

Border Routers

HA Firewalls

Site/Campus LAN

Virtual Circuit

Virtual Circuit

perfSONAR

Data Transfer Cluster

Data Service Switch Plane

perfSONAR

perfSONAR

perfSONAR

perfSONAR

perfSONAR

perfSONAR
Distributed Science DMZ

Fiber-rich environment enables distributed Science DMZ
- No need to accommodate all equipment in one location
- Allows the deployment of institutional science service

WAN services arrive at the site in the normal way

Dark fiber distributes connectivity to Science DMZ services throughout the site
- Departments with their own networking groups can manage their own local Science DMZ infrastructure
- Facilities or buildings can be served without building up the business network to support those flows

Security is potentially more complex
- Remote infrastructure must be monitored
- Several technical remedies exist (arpwatch, no DHCP, separate address space, etc)
- Solutions depend on relationships with security groups
Distributed Science DMZ – Dark Fiber

- **WAN**
- **Border Router**
  - Clean, High-bandwidth WAN path
  - Site/Campus access to Science DMZ resources
- **Enterprise Border Router/Firewall**
- **Science DMZ Switch/Router**
  - Dark Fiber
  - Per-project security policy control points
- **Project A DTN (remote)**
- **Project B DTN (remote)**
- **Project C DTN (remote)**

69 – ESnet Science Engagement (engage@es.net) - 10/2/13
Multiple Science DMZs – Dark Fiber

Diagram showing a network setup with a Border Router connecting to Enterprise Border Router/Firewall, Site/Campus LAN, and various science DMZ switch/routers connected to Project A DTN (building A), Facility B DTN (building B), Cluster DTN (building C), and perfSONAR nodes.

Legend:
- WAN
- perfSONAR

Network Connections:
- 10G
- 10GE

Security Policy:
- Per-project security policy

Contact:
- engage@es.net
- 10/2/13
Common Threads

Two common threads exist in all these examples

Accommodation of TCP
- Wide area portion of data transfers traverses purpose-built path
- High performance devices that don’t drop packets

Ability to test and verify
- When problems arise (and they always will), they can be solved if the infrastructure is built correctly
- Small device count makes it easier to find issues
- Multiple test and measurement hosts provide multiple views of the data path
  - perfSONAR nodes at the site and in the WAN
  - perfSONAR nodes at the remote site
Summary So Far

There is no single “correct” way architect a Science DMZ

• these are a few “design patterns”

It depends on things like:

• site requirements
• existing resources
• availability of dark fiber
• budget

The main point is to reduce the opportunities for packet loss
The Science DMZ – Introduction & Architecture

Questions?

Jason Zurawski - zurawski@es.net
ESnet Science Engagement – engage@es.net

http://fasterdata.es.net