November 18th 2013, SC13 Network Performance Tutorial

Jason Zurawski – Internet2/ESnet

Hardware & Network Placement Overview

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

The Basics

- Choosing hardware for a measurement node is not a complicated process
- Some basic guidelines:
  - Bare Metal (more on this later)
  - x86 or 64Bit Architecture
  - "Modern" limits for RAM, CPU Speed, Main Storage
    - E.g. it doesn't need to be brand new, but it should be no older than 8 years (e.g. we have evidence of old Pentium II desktop machines working, but not working well ☺)
  - Recycling is fine, unless you have money to burn on a new device (and who doesn't!)

J. Zurawski – zurawski@es.net

The Basics - Considerations

- The measurement device should fit neatly into your existing infrastructure.
  - E.g. if you have free rack space, a 1U server near the routing devices works
  - If you don't have rack space, perhaps something on the floor makes sense
- Homogenous device choice often means spares parts are available
  - E.g. if you expect failing hard drives and NICs
- Modern hardware is better about power management, resiliency to failure
- Accuracy is what we strive for in measurements:
  - We want accurate, stable clocks
  - We want capable NIC hardware/device drivers
  - We want to have a machine that can handle the moderate load of heavyweight network testing

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Time (is not) on Your Side

- The biggest requirement from the server's perspective, is being able to accurately represent time
- This is done in two ways
  - Computers have a local clock (features a battery backup).
    - New exception to this rule are machines similar to a Raspberry PI.
  - Computerized methods exist to keep the local clock honest
- If the clock is hosed, measurements are hosed

What is NTP?

- NTP is a protocol designed to synchronize the clocks of computers over a network to UTC
  - Servers will present the data in timezones as needed
  - Synchronize to Internet time servers or other sources, such as a radio or satellite receiver or telephone modem service.
- It can also be used as a server for dependent clients
- Attempts to keep time ***monotonically increasing*** while minimizing offset and skew
  - Sends signals to system clock to correct
  - 'skipping' may be large to start
- Provides accuracies typically ***less than a millisecond on LANs*** and ***up to a few milliseconds on WANs***
- Typical NTP configurations utilize multiple redundant servers and diverse network paths in order to achieve high accuracy and reliability.
  - Redundancy – enough choices to pick a 'good' clock
  - Diverse Paths – Minimize the effect of congestion on a common path

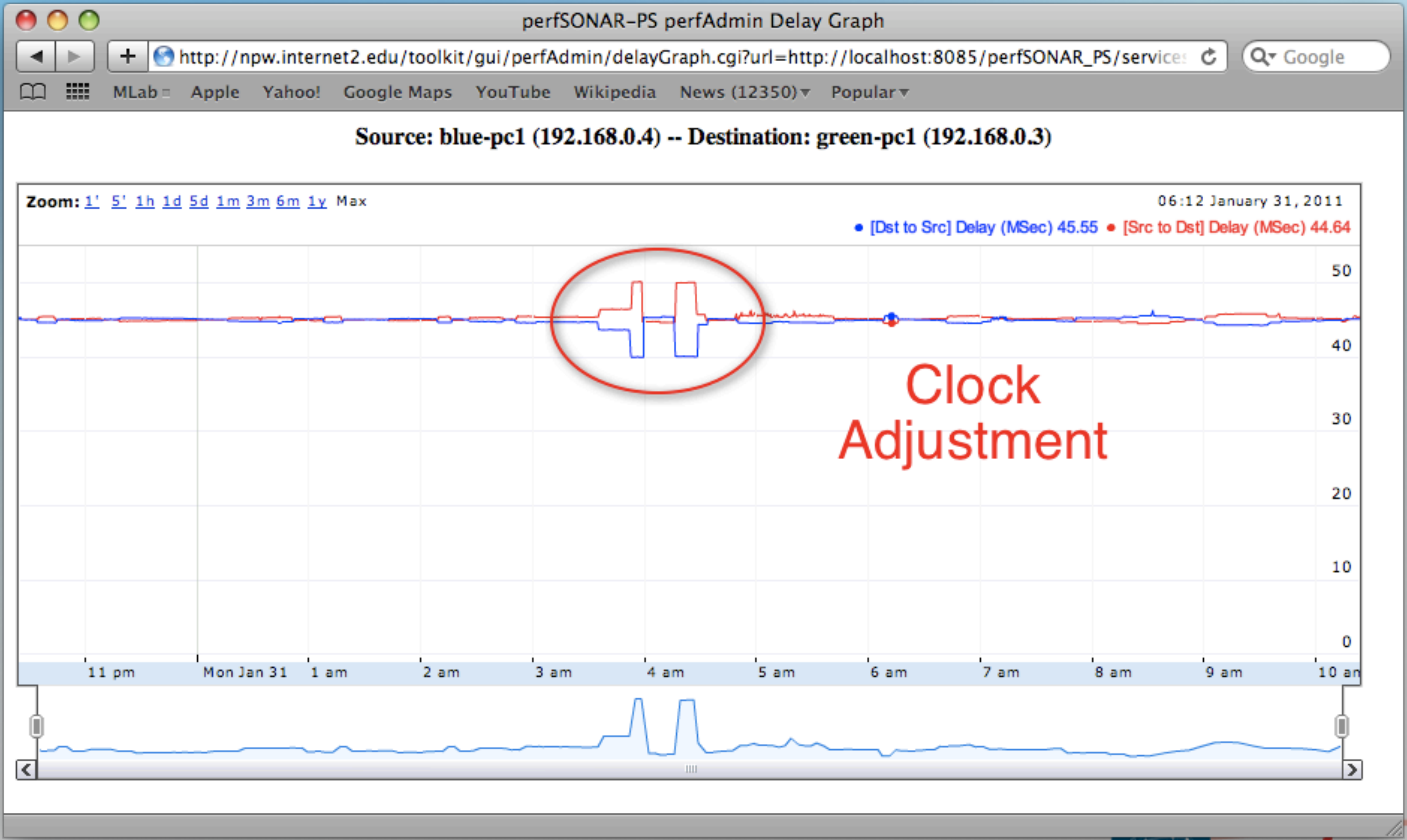
Utility for Measurement

- Scheduling requires coarse grain agreement on time (lets start/end together)
  - Agreement must be "global" in scope – UTC
  - Individual servers communicate with multiple other hosts
- Stability/Accuracy are important
  - Virtualization is still tricky…
- One-Way latency requirements
  - Jitter (requires stability of offset within sample)
  - Latency (requires accuracy)
- Sensible compromise
  - Well defined error representation

Acceptable clock use

- NTP should stabilize the clock over time
- Measurements (e.g. OWAMP) will reflect this change
  - Less ‘skipping’
  - No more ‘negative’ measurements
- NTP will remain in a steady state unless there are network/host problems
  - Selecting constantly between the best ‘peer’ clocks
  - Network routing causing delay between peers
  - Host temperature fluctuations, CPU variability

Acceptable clock use – OWAMP Data

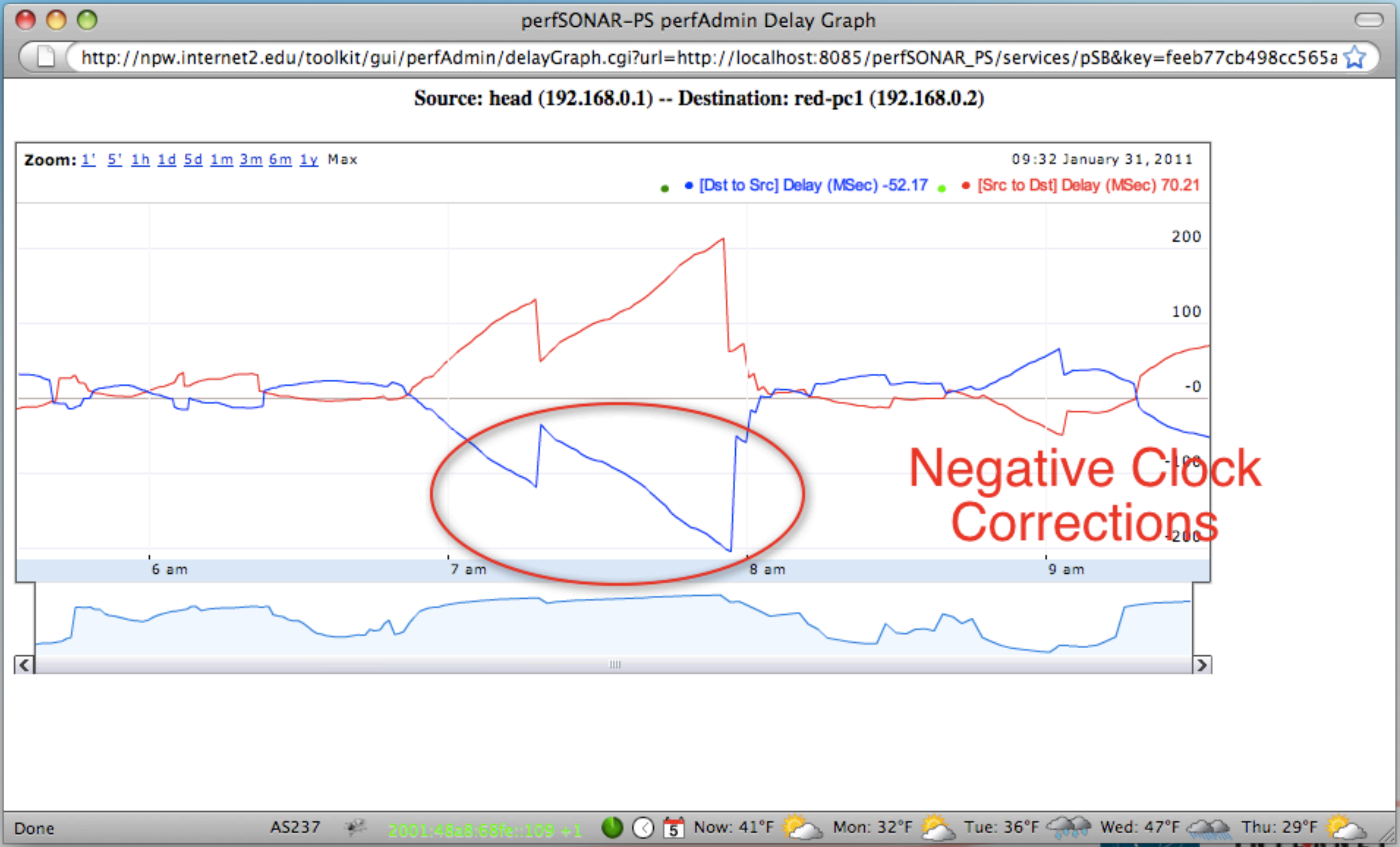

perfSONAR-PS perfAdmin Delay Graph
http://npw.internet2.edu/toolkit/gui/perfAdmin/delayGraph.cgi?url=http://localhost:8085/perfSONAR_PS/service
Google
MLab
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (12350)
Popular
Source: blue-pc1 (192.168.0.4) -- Destination: green-pc1 (192.168.0.3)
Zoom: 1' 5' 1h 1d 5d 1m 3m 6m 1y Max
06:12 January 31, 2011
● [Dst to Src] Delay (MSec) 45.55 ● [Src to Dst] Delay (MSec) 44.64
50
40
30
20
10
0
11 pm
Mon Jan 31
1 am
2 am
3 am
4 am
5 am
6 am
7 am
8 am
9 am
10 am

Clock Adjustment

Poor clock use

- NTP cannot stabilize the clock
  - CMOS battery failure
  - Poor selection of peers
  - Network congestion
  - Host invariability (temperature, CPU)
- Frequent skips in perceived time
- Measurement is unreliable (negative latencies)
- High Jitter

Poor clock use – Skew in OWAMP Data



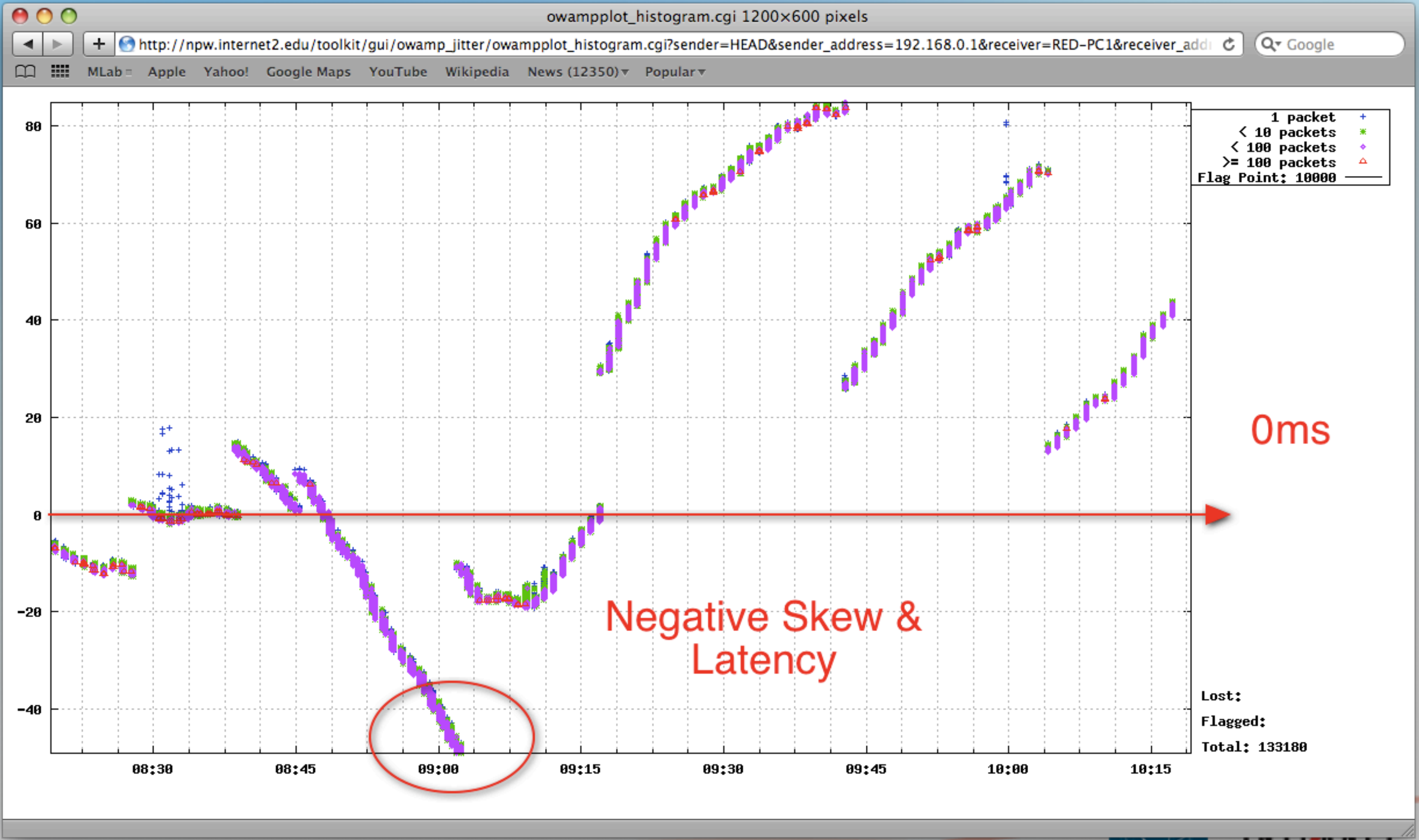
perfSONAR-PS perfAdmin Delay Graph

http://npw.internet2.edu/toolkit/gui/perfAdmin/delayGraph.cgi?url=http://localhost:8085/perfSONAR_PS/services/pSB&key=feeb77cb498cc565a

Source: head (192.168.0.1) -- Destination: red-pc1 (192.168.0.2)

Negative Clock Corrections

Poor clock use – Jitter in OWAMP Data



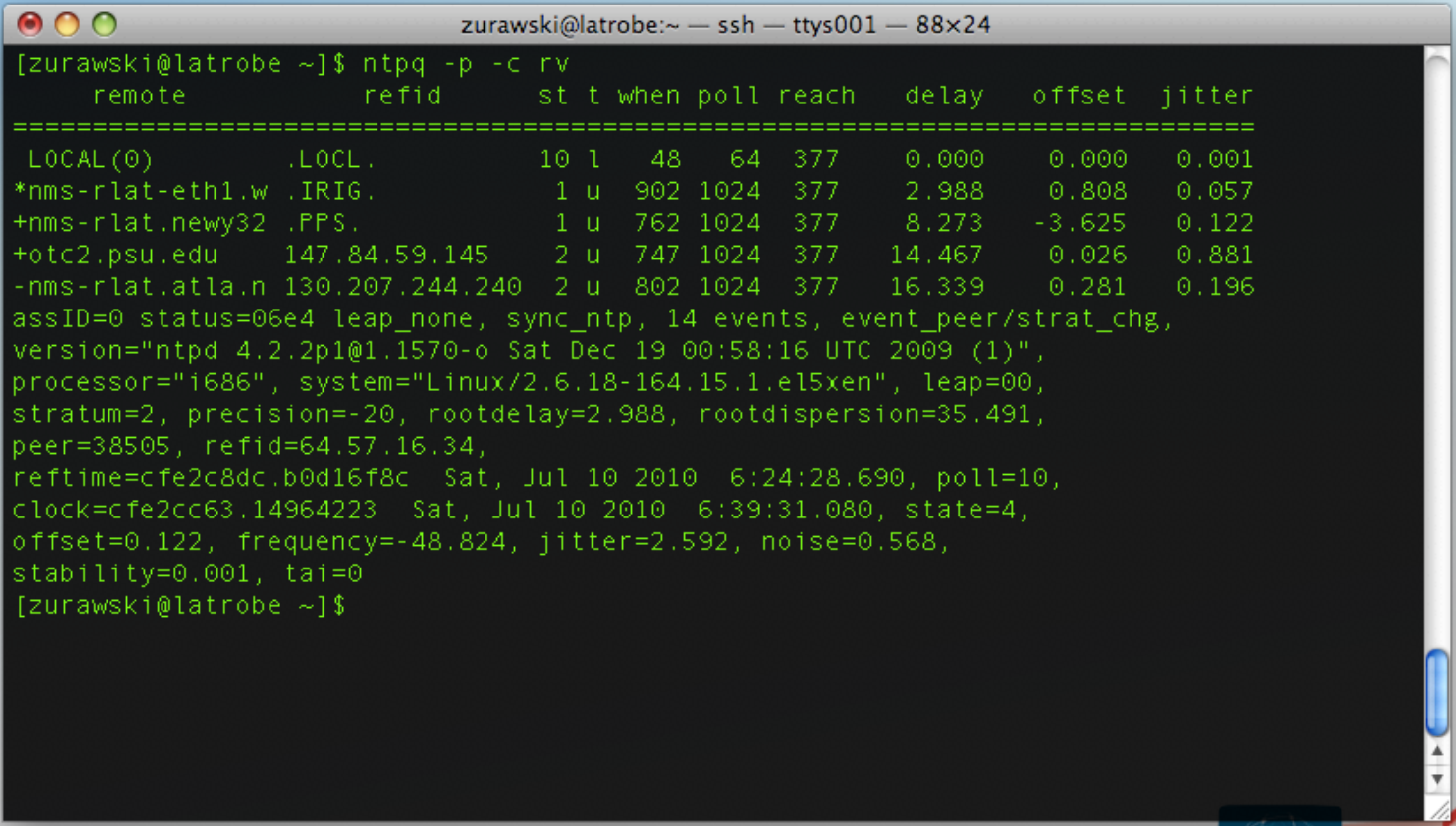
0ms

Negative Skew & Latency

J. Zurawski – zurawski@es.net

Verify NTP – Do this live if you Like

- **`ntpq –p –c rv`**



```
[zurawski@latrobe ~]$ ntpq -p -c rv
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
 LOCAL(0)        .LOCL.          10 l   48   64  377    0.000    0.000   0.001
*nms-rlat-eth1.w .IRIG.           1 u  902 1024  377    2.988    0.808   0.057
+nms-rlat.newy32 .PPS.            1 u  762 1024  377    8.273   -3.625   0.122
+otc2.psu.edu    147.84.59.145    2 u  747 1024  377   14.467    0.026   0.881
-nms-rlat.atla.n 130.207.244.240  2 u  802 1024  377   16.339    0.281   0.196
assID=0 status=06e4 leap_none, sync_ntp, 14 events, event_peer/strat_chg,
version="ntpd 4.2.2p1@1.1570-o Sat Dec 19 00:58:16 UTC 2009 (1)",
processor="i686", system="Linux/2.6.18-164.15.1.el5xen", leap=00,
stratum=2, precision=-20, rootdelay=2.988, rootdispersion=35.491,
peer=38505, refid=64.57.16.34,
reftime=cfe2c8dc.b0d16f8c  Sat, Jul 10 2010  6:24:28.690, poll=10,
clock=cfe2cc63.14964223  Sat, Jul 10 2010  6:39:31.080, state=4,
offset=0.122, frequency=-48.824, jitter=2.592, noise=0.568,
stability=0.001, tai=0
[zurawski@latrobe ~]$
```

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Use Cases

- When selecting hardware, consider there are two basic use cases in the measurement world:
  - Lightweight Testing, normally related to Latency (Ping, Traceroute, OWAMP, Passive Measurements)
  - Heavier Testing, normally related to Bandwidth (BWCTL, IPERF, NUTTCP, NPAD, NDT)
- Hardware requirements can be relaxed in the case of the former

Use Cases - Latency

- A 10G card isn't really need, 1G is recommended (100M would be ok as well, just be sure the driver is recent)
  - Be careful with TCP offload on some NICs, it can introduce OOP
- CPU load is minimal, single core single CPU is fine. Doesn't need to be a whole lot of MHz/GHz
  - Multi-core/processor systems can sometimes introduce jitter on their own if interpret processing is not handled efficiently
- RAM is also minimal, enough to support a modern Linux distro (1G should be sufficient)
- Main Memory is where you do need some power. OWAMP Regular testing data can build up over time. Several G a month depending on who you are testing against.
  - This can be cleaned out if you are space constrained
  - We recommend 200G to be safe.

Use Cases - Bandwidth

- 1G is a common use case, but if you can do 10G aim for this
  - Same caveat about drivers – there are some nasty kernel/driver interactions stories out there …
- CPU should be beefy, you do want a pretty good pentium/xeon on your side. Multi-cores/processors are not a requirement
- RAM should be consistent with the CPU, 2G+ is good
- The main memory requirements are not as great as the latency machine, 100G is more than enough.

perfSONAR powered

Agenda

- Hardware
    - The Basics
    - Time (is not) on Your Side
    - Use Cases
        - Latency
        - Bandwidth
    - Good Choices
    - Poor Choices
- Network Placement
    - Overview
    - Zones
    - Strategies
- Regular Testing Plan
    - Importance
    - Visualizations

Good Choices

- Modern Server Class Hardware
  - Internet2 uses Dell Power Edge 1950s (from 2005!) and these are still kicking
  - I have been testing some Dell R310s lately. Pretty cost effective (EDU pricing of around $1.5k if you add on a 10G card and some LR optics)
  - Supermicro makes a nice 1U/Half Size machine with an Atom processor. These are excellent for Latency testing (don't push it with the bandwidth though

Good Choices

- Desktop Towers
  - I don’t test these often, most are probably ok for temporary use cases.
  - “Energy Saving” models are a little suspect, these could reduce CPU power and effect the clock
- Laptops
  - I wouldn’t recommend this for longer term use, but for diagnostics they are mobile and effective

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Poor Choices

- Virtual Machines
  - Our largest concern is the clock
    - A VM gets its time updates from the Hyporvisor
    - The HV gets updates via the system (hopefully it is running NTP)
    - If the VM is also running NTP, it will attempt to keep the clock stable, but the 'backdoor' updates to the VM clock from the HV will skip time forward/backward – confusing NTP
    - Think about what happens if the VM is swapped out …
  - Situations where a VM is ok:
    - NDT/NPAD Beacon
    - 1G bandwidth testing
    - SNMP Collection, NAGIOS Operation
  - Situations where it is not:
    - OWAMP measurements
    - 10G Throughput

Poor Choices

- Mac Mini and similar micro-machines
  - Largest concern here is that the 1G NIC is on the motherboard, and competes for BUS resources.
    - This introduces jitter in latency measurements
    - Reduces throughput tests
  - Power management can be funky too
- Desktops/Laptops (for permanent placement)
  - Power management is a concern for aforementioned reasons
  - Onboard NICs are common here as well

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

perfSONAR Deployment Locations

- Critical to deploy such that you can test with useful semantics
- perfSONAR hosts allow parts of the path to be tested separately
  - Reduced visibility for devices between perfSONAR hosts
  - Rely on counters or other means where perfSONAR can't go
- Effective test methodology derived from protocol behavior
  - TCP suffers much more from packet loss as latency increases
  - TCP is more likely to cause loss as latency increases
  - Testing should leverage this in two ways
    - Design tests so that they are likely to fail if there is a problem
    - Mimic the behavior of production traffic as much as possible
- ***N.B.*** don't design your tests to succeed – ***it is not helpful***

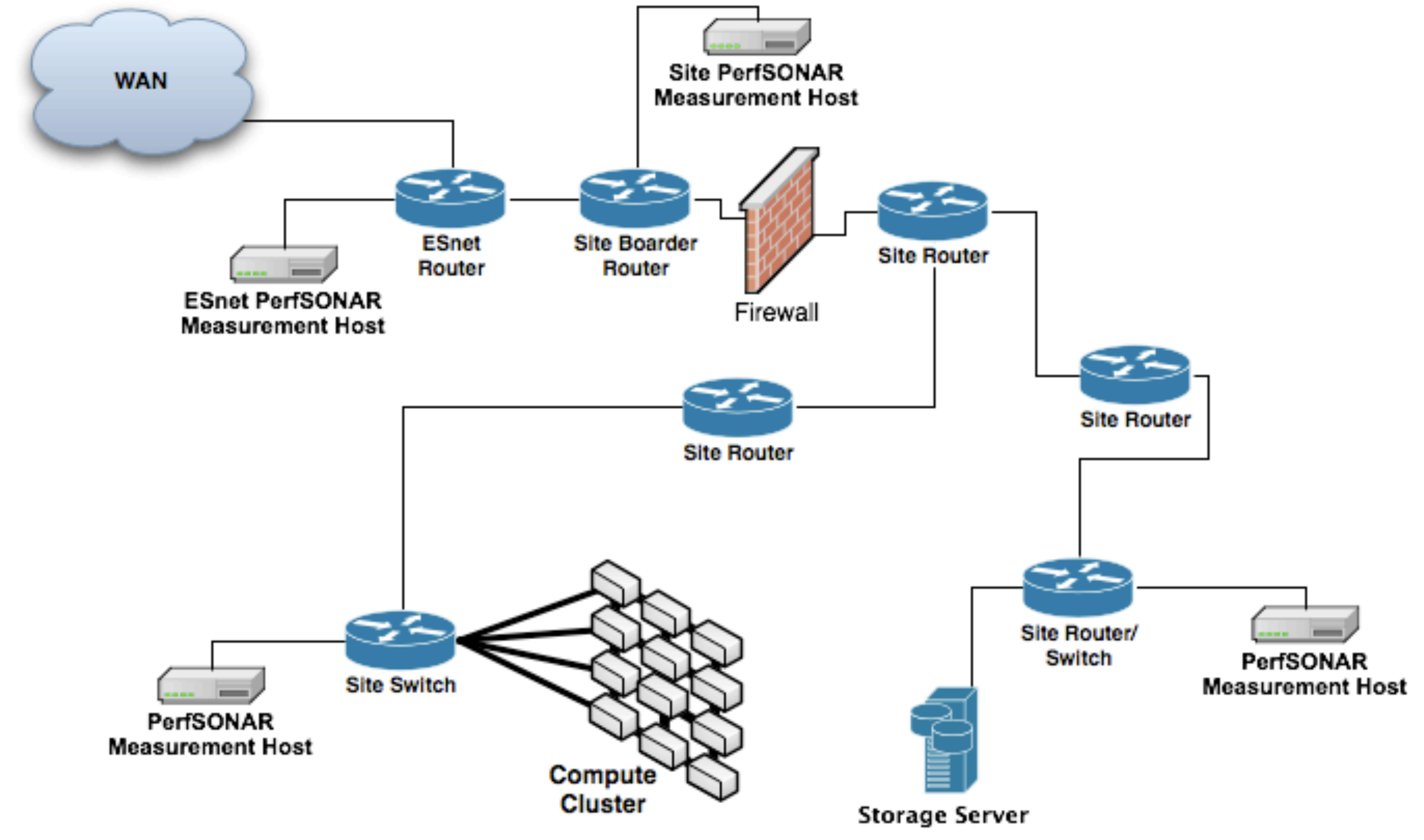
J. Zurawski – zurawski@es.net

Why is Placement Important

- Placement of a tester should depend on two things:
    - Where a tester will have the most positive of impacts for find/preventing problems
    - Where space/resources are available
- We want to find certain sets of problems:
    - Edge of your network to edge of your upstream provider
        - E.g. University to Regional Network
        - Regional Network to Backbone Network
    - Core of your network to Edge of your network and upstream providers
        - Campus core facility to demarcation point
        - Campus core to WAN Connectivity
    - Location of important devices to remote facilities and points in between
        - Data centers to consumers of said data (e.g. campus to campus)
        - Data centers to WAN Connectivity

J. Zurawski – zurawski@es.net

Sample Site Deployment



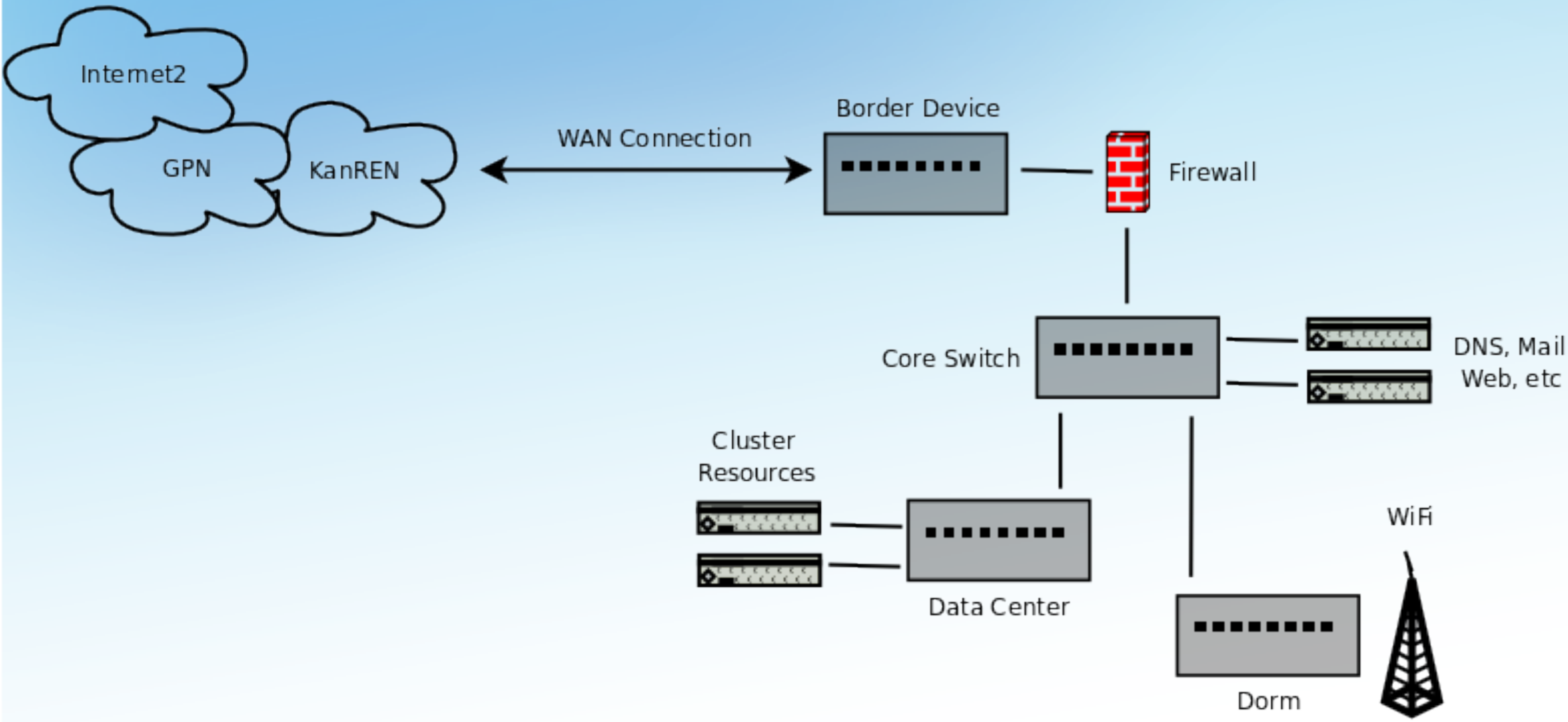
Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Constructing Zones

- Networks are large and complex, but can be broken into a couple of common components:
  - Main Distribution Frame (MDF) where the WAN connectivity will land.
  - Intermediate Distribution Frames (IDF) in other buildings (major components on a LAN)
  - The Network "core" which may be data center that houses key components (Mail, DNS, HTTP, Telephony)
  - Population centers (Dorms, Offices, Labs, Data Centers)

The Network



J. Zurawski – zurawski@es.net

Constructing Zones - Demarcation

- MDFs where the WAN connectivity Lands
  - May have provider hardware there too – position either between, or on your end of the demarcation
  - Some providers (e.g. ESnet) locate a server on their end too.
- If you have rack space/power, try to allocate a server directly connected to the border device
  - ***DO NOT*** put it behind the firewall/shaper(!!)
  - To get the best indication of "network" performance, we need to remove network devices from the equation.
- If security is really a concern, consider using a LiveCD and calling it an 'appliance'. This works at US National Labs ☺

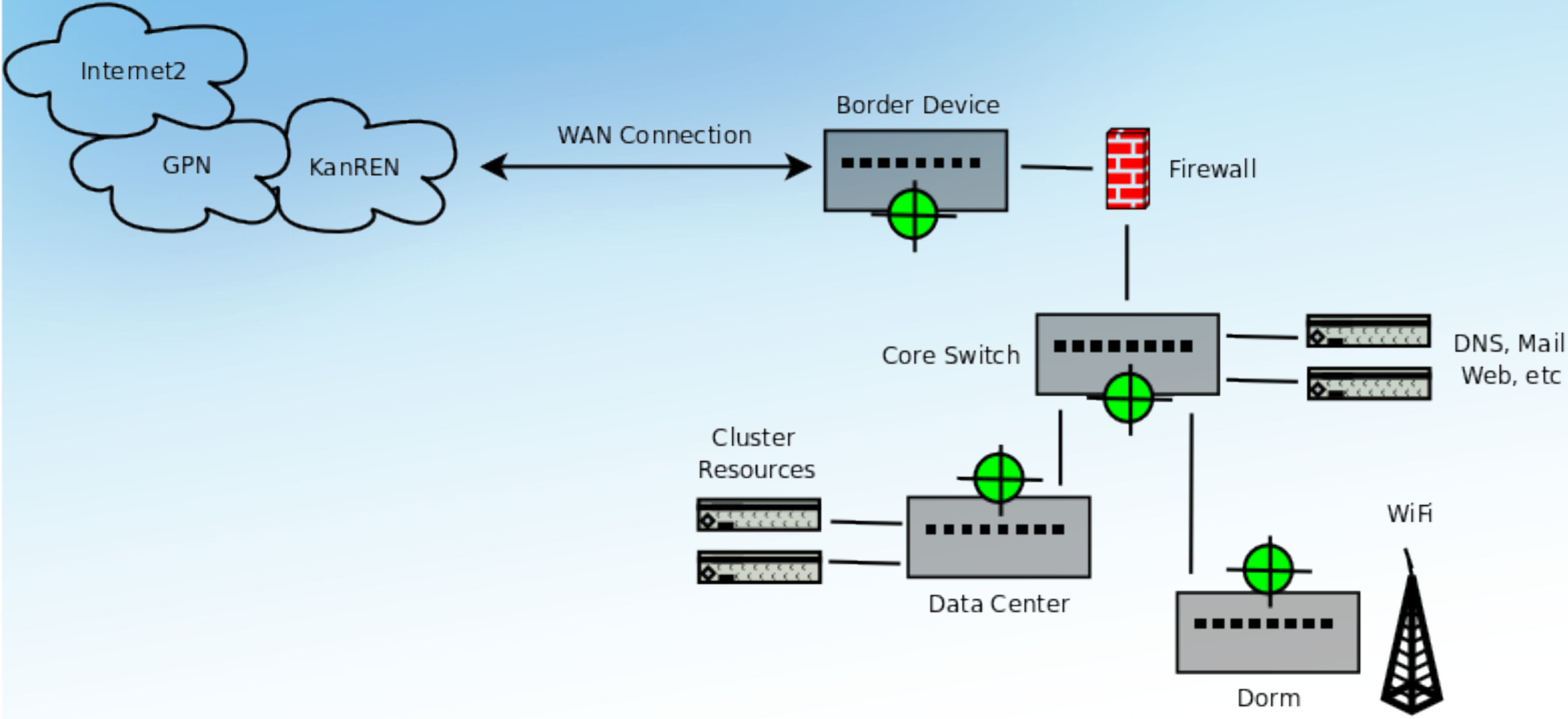
Constructing Zones - Core

- Most likely essential services are in a temperature controlled environment somewhere on campus.
- Rack a device here if power/space allow
  - Ideally it should be on the main Switch/Router for this facility.
  - This CAN be behind firewalls/shapers (but remember this when you are looking at results)
    - Packet loss/jitter/low bandwidth may be caused by the firewall shaper.
    - Comparing performance from behind the firewall to a remote location (KanREN) vs from in front of it is a fun exercise.
- Having the device here will allow testing from all over campus to the most central part of the network

Constructing Zones – Busy Areas

- Busy areas:
  - High Performance Computing Centers (HPCCs)
  - Large Labs
  - Offices/Dorms with a large population of untrusted users
  - Aggregation Points (e.g. ingress of several connections before using a larger access link)
- Space/Power rules apply (e.g. the access switch in a dorm may be in a cramped closet).
- If you are a mostly WiFi shop, put it near the controller.
- Portable devices may be a good choice here, for diagnostic use instead of regular testing

Placement Strategy



Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Testing Strategies

- Once the machines are placed, think about use patterns:
  - Internal performance assurance around the campus
    - For the users
    - For the operations team
  - Keeping the upstream providers honest (verifying SLAs)
  - Testing end to end, verifying that multi-domain uses (video, data movement) will work

Testing Strategies - Internal

- Edge Machine
  - Use this as a beacon, allow the Core and Population Center machines to test toward this
- Core Machine
  - Test to the Edge and all Population Centers
- Population Centers
  - Test to the Edge and Core – testing to other population centers is not required.

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

Importance of Regular Testing

- You can't wait for users to report problems and then fix them (soft failures can go unreported for years!)
  - Things just break sometimes
    - Failing optics
    - Somebody messed around in a patch panel and kinked a fiber
    - Hardware goes bad
- Problems that get fixed have a way of coming back
  - System defaults come back after hardware/software upgrades
  - New employees may not know why the previous employee set things up a certain way and back out fixes
- Important to continually collect, archive, and alert on active throughput test results

Develop a Plan

- What are you going to measure?
  - Achievable bandwidth
    - 2-3 regional destinations
    - 4-8 important collaborators
    - 4-10 times per day to each destination
    - 20 second tests within a region, longer across oceans and continents
  - Loss/Availability/Latency
    - OWAMP: ~10 collaborators over diverse paths
    - PingER: use to monitor paths to collaborators who don't support owamp
  - Interface Utilization & Errors
- What are you going to do with the results?
  - NAGIOS Alerts
  - Reports to user community
  - Post to Website

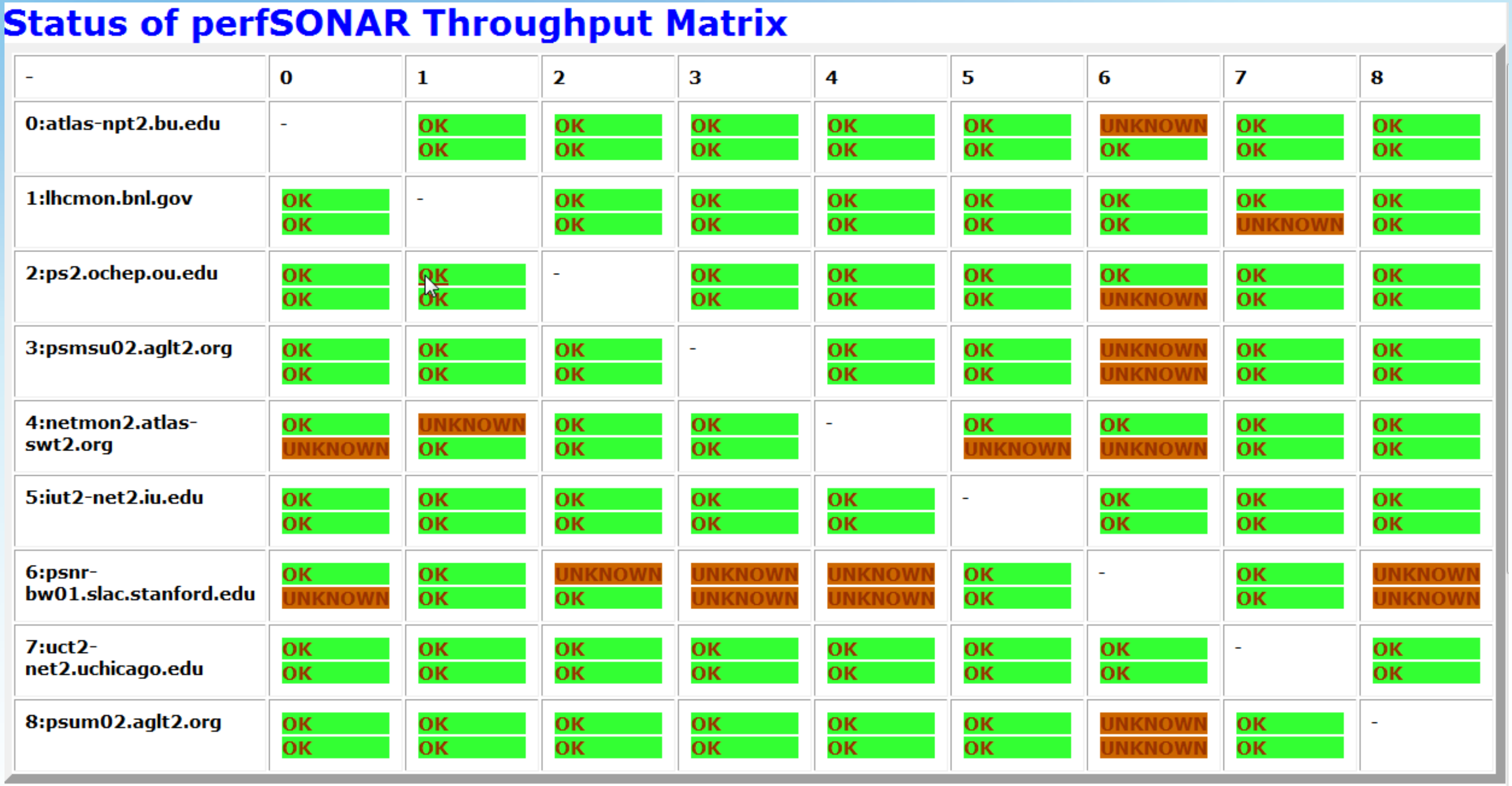
perfSONAR powered

J. Zurawski – zurawski@es.net

Agenda

- Hardware
  - The Basics
  - Time (is not) on Your Side
  - Use Cases
    - Latency
    - Bandwidth
  - Good Choices
  - Poor Choices
- Network Placement
  - Overview
  - Zones
  - Strategies
- Regular Testing Plan
  - Importance
  - Visualizations

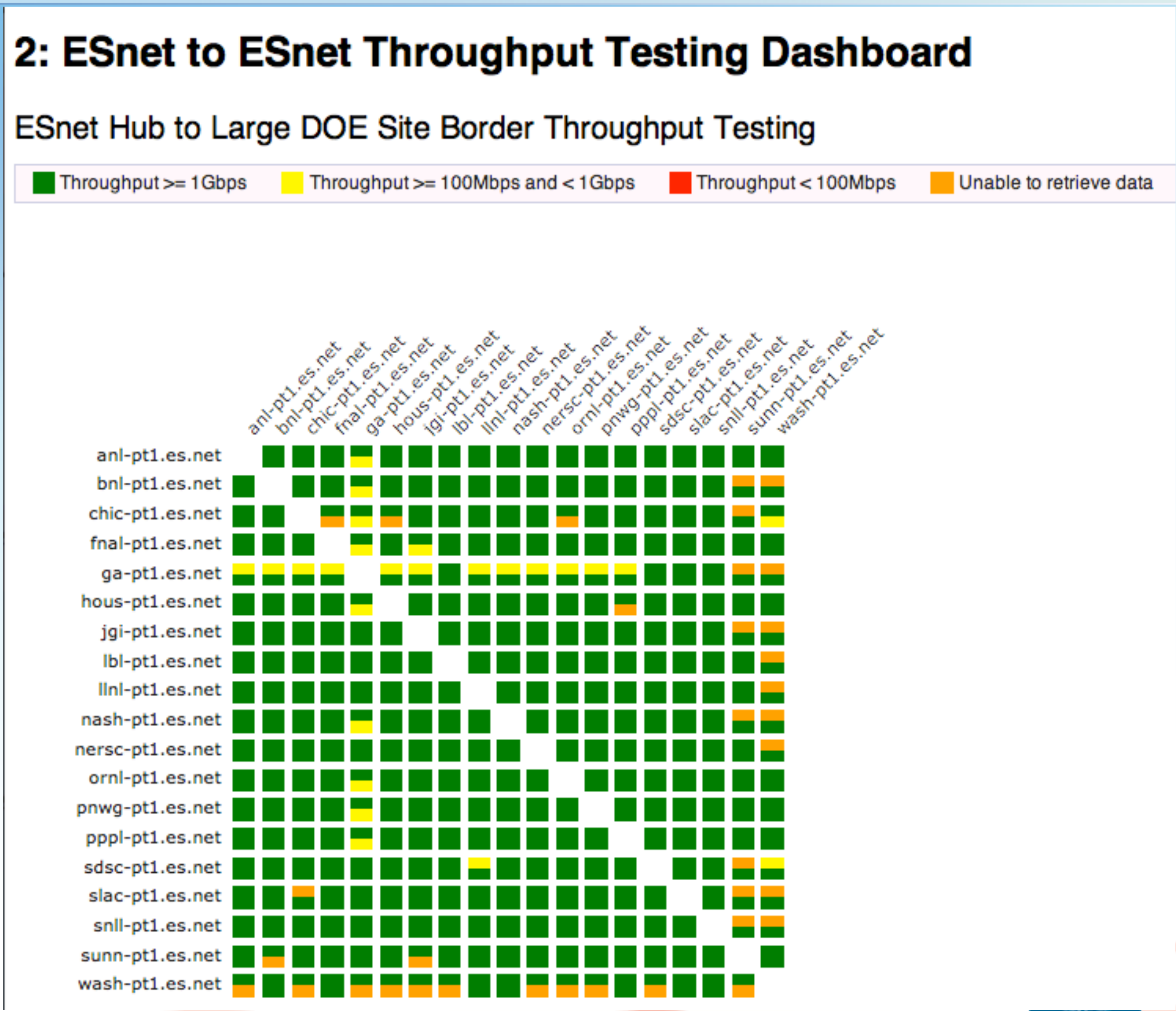
Dashboards – What the Serious Use



Status of perfSONAR Throughput Matrix

-	0	1	2	3	4	5	6	7	8
0:atlas-npt2.bu.edu	-	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN OK	OK OK	OK OK
1:lhcmon.bnl.gov	OK OK	-	OK OK	OK OK	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK
2:ps2.ochep.ou.edu	OK OK	OK OK	-	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK	OK OK
3:psmsu02.aglt2.org	OK OK	OK OK	OK OK	-	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	OK OK
4:netmon2.atlas-swt2.org	OK UNKNOWN	UNKNOWN OK	OK OK	OK OK	-	OK UNKNOWN	OK UNKNOWN	OK OK	OK OK
5:iut2-net2.iu.edu	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK	OK OK	OK OK
6:psnr-bw01.slac.stanford.edu	OK UNKNOWN	OK OK	UNKNOWN OK	UNKNOWN UNKNOWN	UNKNOWN UNKNOWN	OK OK	-	OK OK	UNKNOWN UNKNOWN
7:uct2-net2.uchicago.edu	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK
8:psum02.aglt2.org	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	-

Dashboards – What the Serious Use



2: ESnet to ESnet Throughput Testing Dashboard

ESnet Hub to Large DOE Site Border Throughput Testing

🟩 Throughput >= 1Gbps 🟨 Throughput >= 100Mbps and < 1Gbps 🟥 Throughput < 100Mbps 🟧 Unable to retrieve data

	anl-pt1.es.net	bnl-pt1.es.net	chic-pt1.es.net	fnal-pt1.es.net	ga-pt1.es.net	hous-pt1.es.net	jgi-pt1.es.net	lbl-pt1.es.net	llnl-pt1.es.net	nash-pt1.es.net	nersc-pt1.es.net	ornl-pt1.es.net	pnwg-pt1.es.net	pppl-pt1.es.net	sdsc-pt1.es.net	slac-pt1.es.net	snll-pt1.es.net	sunn-pt1.es.net	wash-pt1.es.net
anl-pt1.es.net		🟩	🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
bnl-pt1.es.net	🟩		🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩	🟧 🟩
chic-pt1.es.net	🟩	🟩		🟩 🟧	🟩 🟨	🟩 🟧	🟩	🟩	🟩	🟩	🟩	🟩 🟧	🟩	🟩	🟩	🟩	🟩	🟧 🟩	🟩 🟨
fnal-pt1.es.net	🟩	🟩	🟩		🟩 🟨	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
ga-pt1.es.net	🟨 🟩	🟨 🟩	🟨 🟩	🟨 🟩		🟨 🟩	🟨 🟩	🟩	🟨 🟩	🟨 🟩	🟨 🟩	🟨 🟩	🟨 🟩	🟨 🟩	🟩	🟩	🟩	🟧 🟩	🟧 🟩
hous-pt1.es.net	🟩	🟩	🟩	🟩	🟩 🟨		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩 🟧	🟩	🟩	🟩	🟩	🟩
jgi-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩	🟧 🟩
lbl-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩
llnl-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩
nash-pt1.es.net	🟩	🟩	🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩	🟧 🟩
nersc-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 🟩
ornl-pt1.es.net	🟩	🟩	🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩
pnwg-pt1.es.net	🟩	🟩	🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩
pppl-pt1.es.net	🟩	🟩	🟩	🟩	🟩 🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩
sdsc-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨 🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟧 🟩	🟨 🟩
slac-pt1.es.net	🟩	🟩	🟧 🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟧 🟩	🟧 🟩
snll-pt1.es.net	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟧 🟩	🟧 🟩
sunn-pt1.es.net	🟩	🟩 🟧	🟩	🟩	🟩	🟩	🟩 🟧	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩
wash-pt1.es.net	🟩 🟧	🟩	🟩 🟧	🟩	🟩 🟧	🟩 🟧	🟩 🟧	🟩 🟧	🟩	🟩	🟩 🟧	🟩 🟧	🟩 🟧	🟩	🟩 🟧	🟩	🟩	🟩 🟧	

Hardware & Network Placement Overview

November 18th 2013, SC13 Network Performance Tutorial

Jason Zurawski – Internet2/ESnet

For more information, visit http://www.internet2.edu/workshops/npw

https://submissions.supercomputing.org/?page=SessionEval&id=sess118