Campus Network Design
Science DMZ

Dale Smith
Network Startup Resource Center
dsmith@nsrc.org

ESnet
ENERGY SCIENCES NETWORK

U.S. DEPARTMENT OF
ENERGY

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Making Networks Faster

- Lots of work has been done to try to understand how to make transfers of large data files go faster.
- ESnet, the USA Energy Sciences Network has done a lot of work on this issue
  - See http://fasterdata.es.net
- This talk will summarize some of those concepts

Science Needs Lots of Data

- Many science disciplines need large data sets for analysis
- Moving these large data sets over long distances is challenging

Moving 1 Terabyte File

- 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
- Compare these speeds to:
  - USB 2.0 portable disk: 20-30 hours

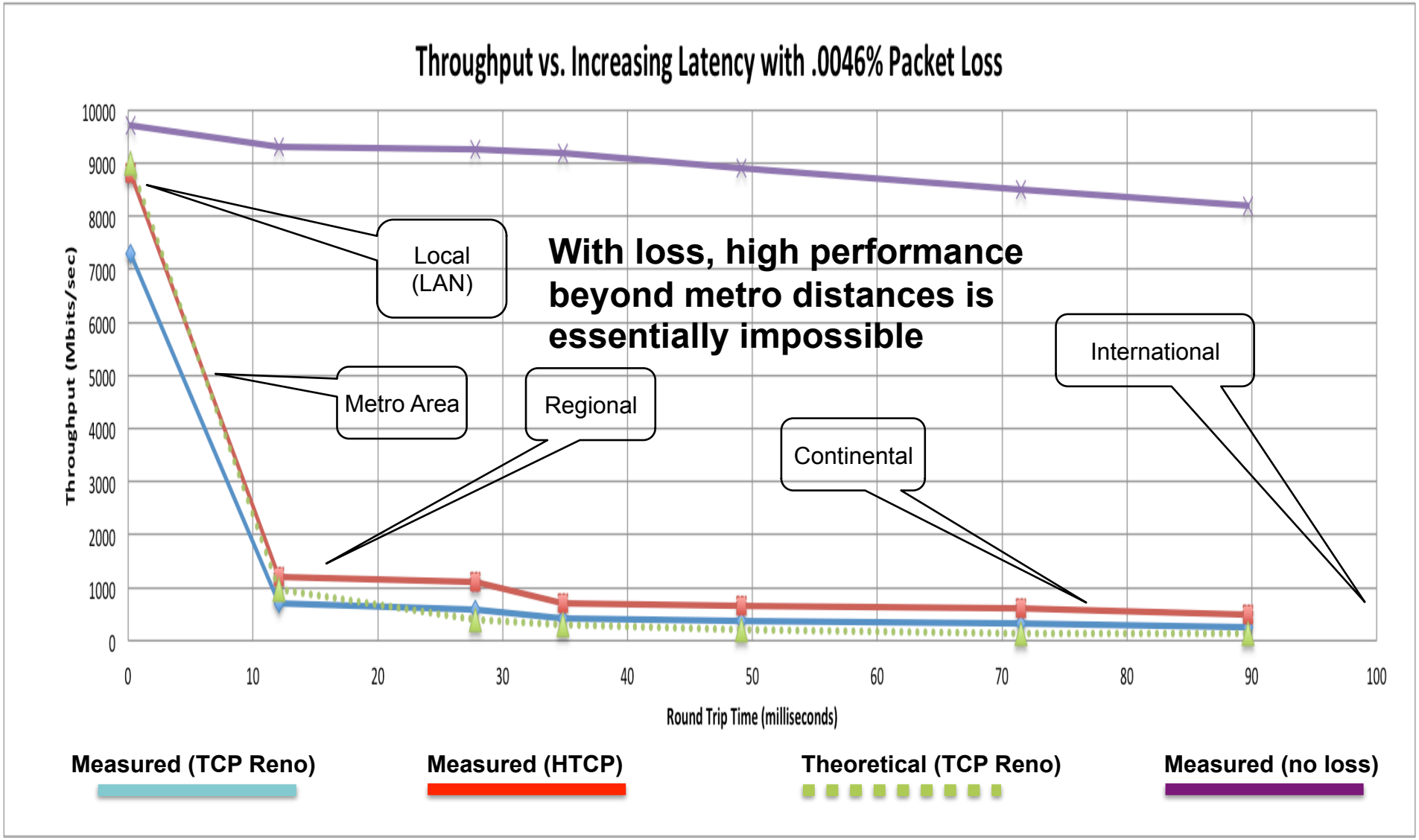
Science Use Case

- Alice & Bob are science collaborators
  - Experts in their field
  - Physically separated, on separate continents
  - Rely on networks, but are not IT experts
- Alice & Bob start a new project
  - Instrument @ one end generating large data sets
  - processing/analysis @ the other
  - How well is this going to work?

Use Case: Networks Look OK

- Pinging between Alice and Bob's systems show 1 packet lost in every 20,000 sent
- IT/Networking look at Internet use graphs and sees low usage (no congestion)
- However, data transfers are 1/10th what they expected and are taking 10 times longer than was predicted
- What has happened?

A small amount of packet loss makes a huge difference in TCP performance



Which leads us to the Science DMZ

- Causes of poor data transfer performance
  - Packet loss issues (even small amounts)
  - Un-tuned/under-powered hosts
- The Science DMZ is a design pattern to optimize for network performance
  - Not all implementations look the same, but share common features
  - Some choices don't make sense for everyone (Know Your Network)

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Traditional Network DMZ

- DMZ – "Demilitarized Zone"
  - Typically a network segment off of the border firewall that houses servers
  - This network segment is near the site perimeter and has a different security policy than the rest of the network
  - Commonly used architectural element for deploying WAN-facing services (e.g. email, DNS, web)

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Traditional Network DMZ

- Traffic on the DMZ does not traverse the campus 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
- Do the same thing for science – Science DMZ

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

The Science DMZ in 2 Slides

- Consists of three key components:
  1. "Friction free" network path
     - Goal is **no** packet loss
     - Highly capable network devices (wire-speed, deep queues)
     - Security policy and enforcement specific to science workflows
     - Located at or near site perimeter if possible
     - Virtual circuit connectivity option

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

The Science DMZ in 2 Slides

2. Dedicated, high-performance Data Transfer Nodes (DTNs)
   - Hardware, operating system, libraries all optimized for transferring data quickly
   - Includes optimized data transfer tools such as Globus or GridFTP
3. Performance measurement/test node
   - perfSONAR

Details at http://fasterdata.es.net/science-dmz/

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

TCP – Ubiquitous and Fragile

- Virtually all communications between systems on the Internet is via TCP
- TCP – the fragile workhorse
  - TCP is (for very good reasons) timid – packet loss is interpreted as congestion
  - Packet loss in conjunction with latency is a performance killer
  - Like it or not, TCP is used for the vast majority of data transfers (including science data)

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

How do you make TCP go fast?

- High-performance wide area TCP flows must get loss-free service
  - Sufficient bandwidth to avoid congestion
  - No packet loss in network devices you can control
  - This is “Friction Free Networking”

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

1. Friction Free Network Path

- Goal is to Totally Eliminate Packet Loss
- Common sources
  - security devices (firewalls, NAT boxes, Intrusion Prevention System/IPS)
  - Routers or switches without enough buffering
    - Micro flows from a capable server can over run the buffers in some devices
    - This problem becomes more common as data rates increase and the capabilities of the hosts increase

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

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Equipment – Routers and Switches

- Requirements for Science DMZ gear are different than for campus LAN
  - No need to go for the long list of features
    - Support for the latest LAN integration magic with your Windows Active Directory environment is probably not super-important
  - A clean architecture is important
    - How fast can a single flow go?
    - Are there any components that go slower than interface wire speed?

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Please tell me – what should I buy?

- We get this question a lot
  - Hard to answer what is right for you
- We don't recommend one vendor over another
- We have no idea what's right for your environment
  - Our goal is to describe our understanding of what works and why

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Some Stuff We Think Is Important

- Deep interface queues
  - Output queue or VOQ – doesn't matter
  - What TCP sees is what matters
  - No, this isn't buffer bloat
- Good counters
  - We like the ability to reliably count *every* packet associated with a particular flow, address pair, etc
    - Very helpful for debugging packet loss
    - Must not affect performance (just count it, don't punt it)
    - sflow support if possible
  - If the box is going to drop a packet, it should increment a counter somewhere indicating that it dropped the packet
    - Magic vendor permissions and hidden commands should not be necessary
    - Some boxes just lie – run away!
- Single-flow performance should be wire-speed

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

You are not alone

- Lots of community resources
  - Ask folks who have already done it
  - Ask the Science DMZ mailing list: sciencedmz@es.net
- Vendors can be very helpful – just ask the right questions
  - Request an eval box (or preferably two)
  - Ask for config examples to implement a particular feature
    - E.g. “Please give me the QoS config for the following:”
      - 1 queue for network control (highest priority) – 5% of interface buffer memory
      - 1 queue configured for tail-drop (lower priority) – 95% of interface buffer memory
      - With that config, how many milliseconds of buffer are in the tail-drop queue when measured at interface wire speed?

UNIVERSITY OF OREGON

NSRC Network Startup Resource Center

2. Dedicated Data Transfer Node

- As your network speeds increase, it becomes more and more critical to have hosts specifically designed to move data
  - Special hardware
    - Disk systems
    - Motherboards, processors, etc
  - Properly tuned operating system to be able to transfer data over long distances (with high latency)

DTN Hardware Considerations

- Motherboard and Chassis
  - 40Gbs requires PCI Express Gen3
    - Intel Sandy/Ivy Bridge for example
- Pay attention to PCI bus, not all the same
  - Some faster and wired with more “lanes” than others
- DTN should have lots of memory
  - 32Gb should be considered a minimum

DTN Disk Storage Considerations

- Whether local storage or Storage System doesn't matter. Only speed does.
- Disk speed is a changing game and anything we write here is out of date soon
- SSD storage is generally faster than rotating disks
- Raid controllers can help things go faster
  - Must do raid to go faster than 1Gbs

DTN Network Interfaces

- Almost any 1G network interface is fine
- When you get to 10G (and above) network interfaces, you need special hardware
  - Cheap is not always good
  - Look for
    - Interrupt coalescing
    - Support for MSI-X
    - Support for TCP offload engine
    - Support for zero-copy protocols such as RDMA

DTN Tuning

- Defaults aren't right for any operating system
- What needs to be tuned:
  - BIOS
  - Firmware
  - Device Drivers
  - Networking
  - File System
  - Application
- Can often double performance with tuning

And use the right transfer tool

- Sample Results: Berkeley, CA to Argonne, IL (near Chicago), RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
scp	140Mbs
HPN patched scp	1.2Gbs
ftp	1.4Gbs
GridFTP (4 streams)	6.6Gbs

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

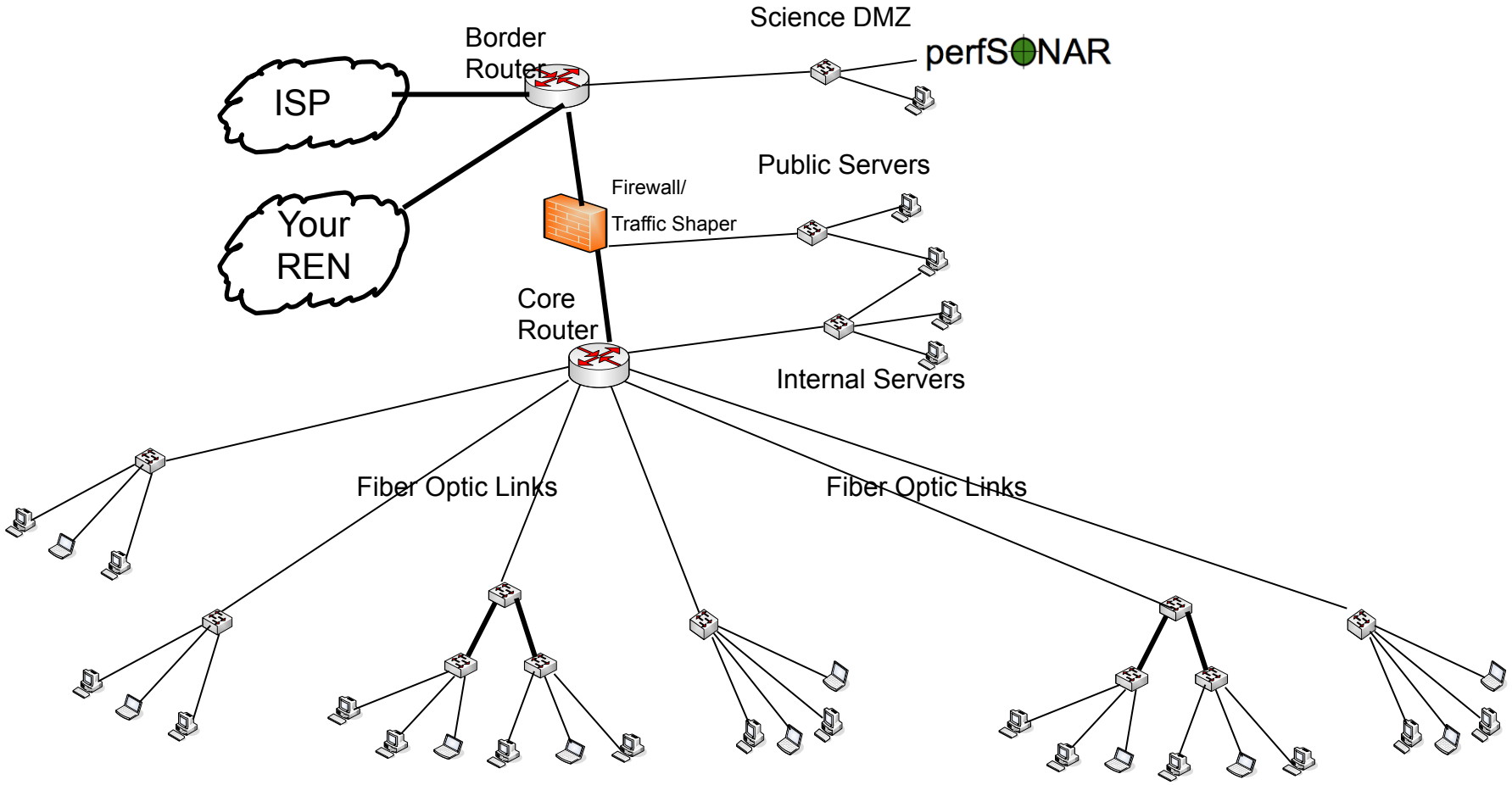
3. Performance/Measurement

- What is recommended for Science DMZ is to use perfSONAR
- Where should you put perfSONAR nodes?
  - Obviously, where the DTN is
  - But, what about other places?
  - Need perfSONAR in campus and in NREN
  - Being able to test to multiple locations and getting data from multiple places in your network is quite useful

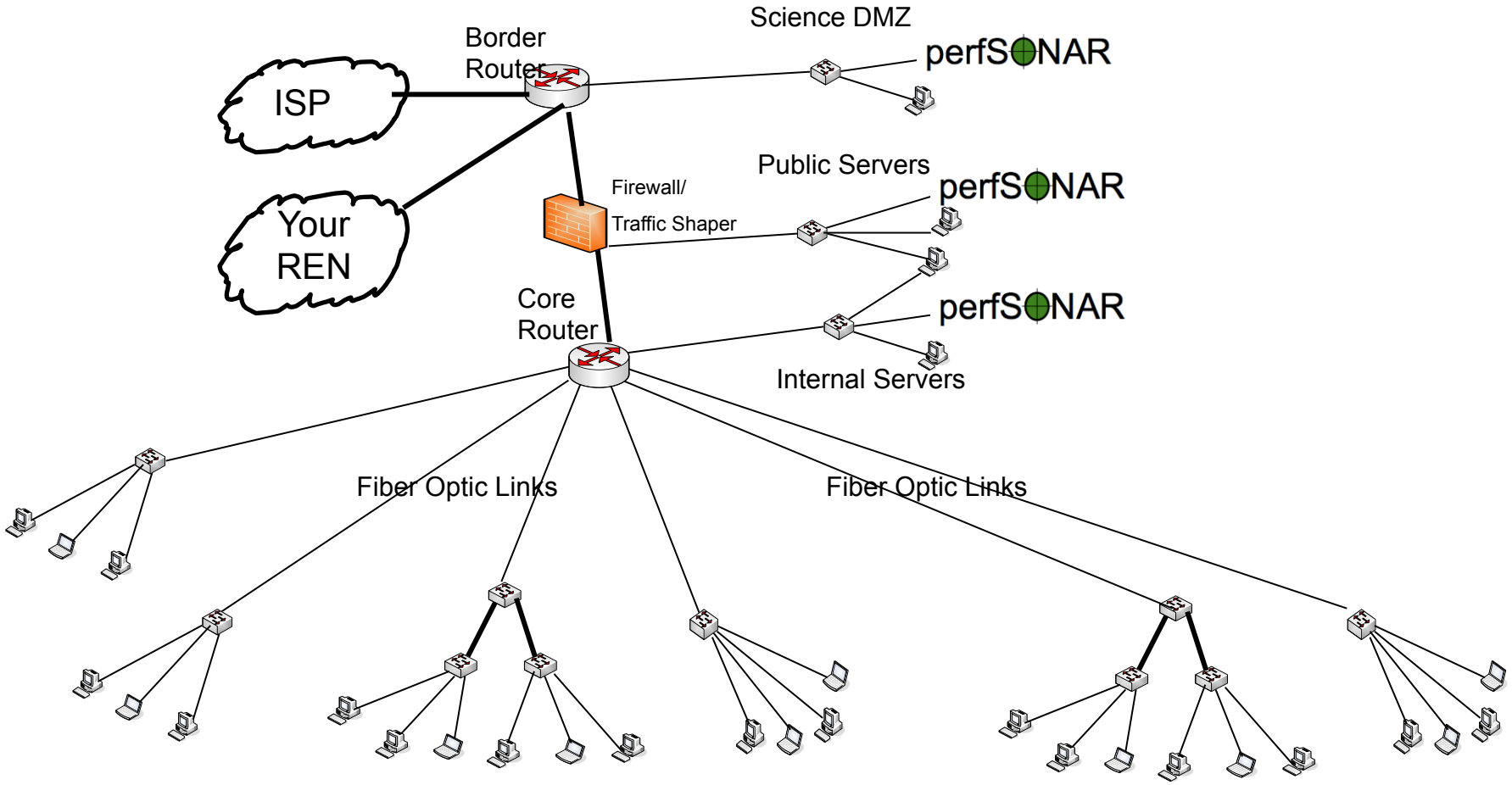
UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

perfSONAR Placement - Campus



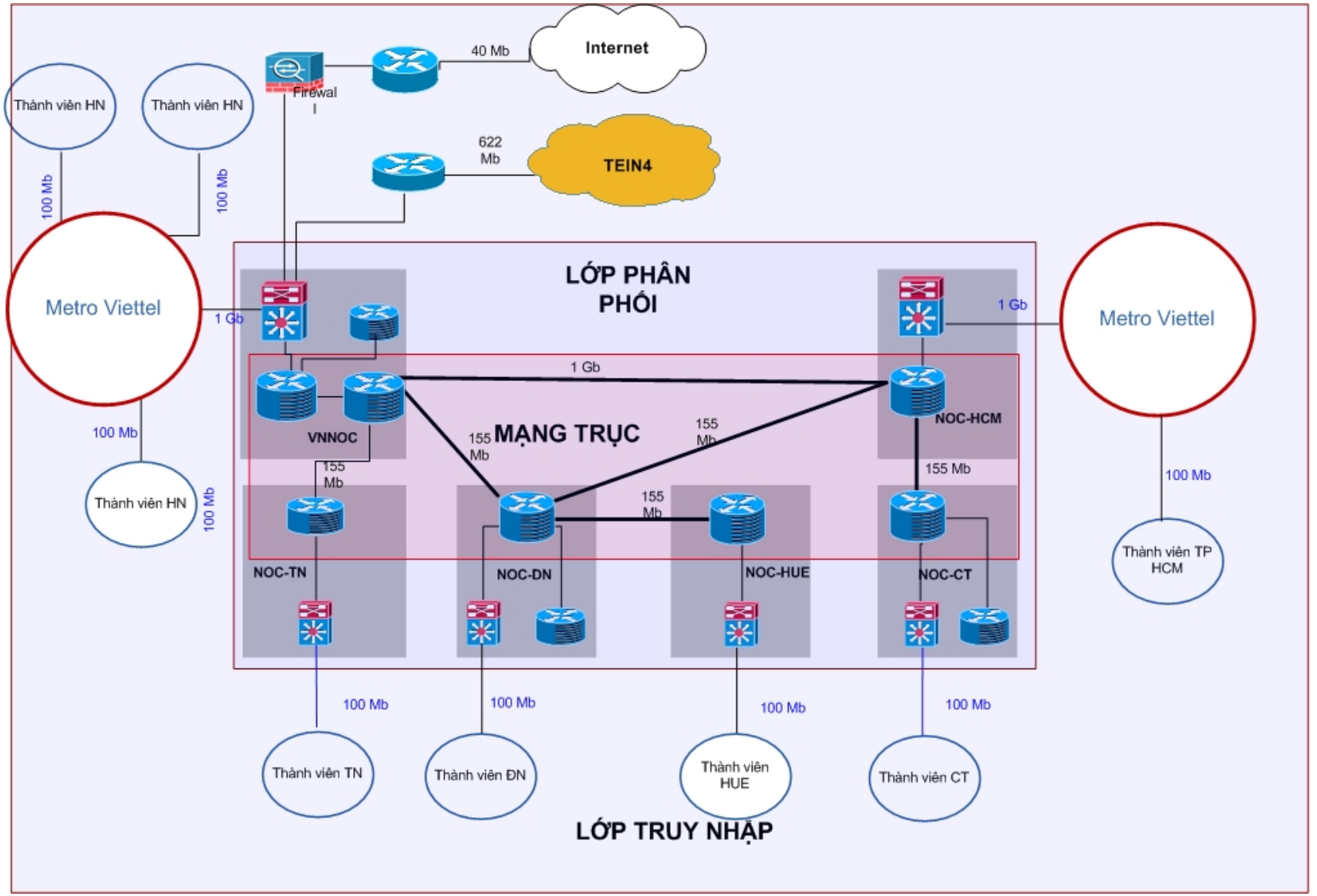
perfSONAR Placement - Campus



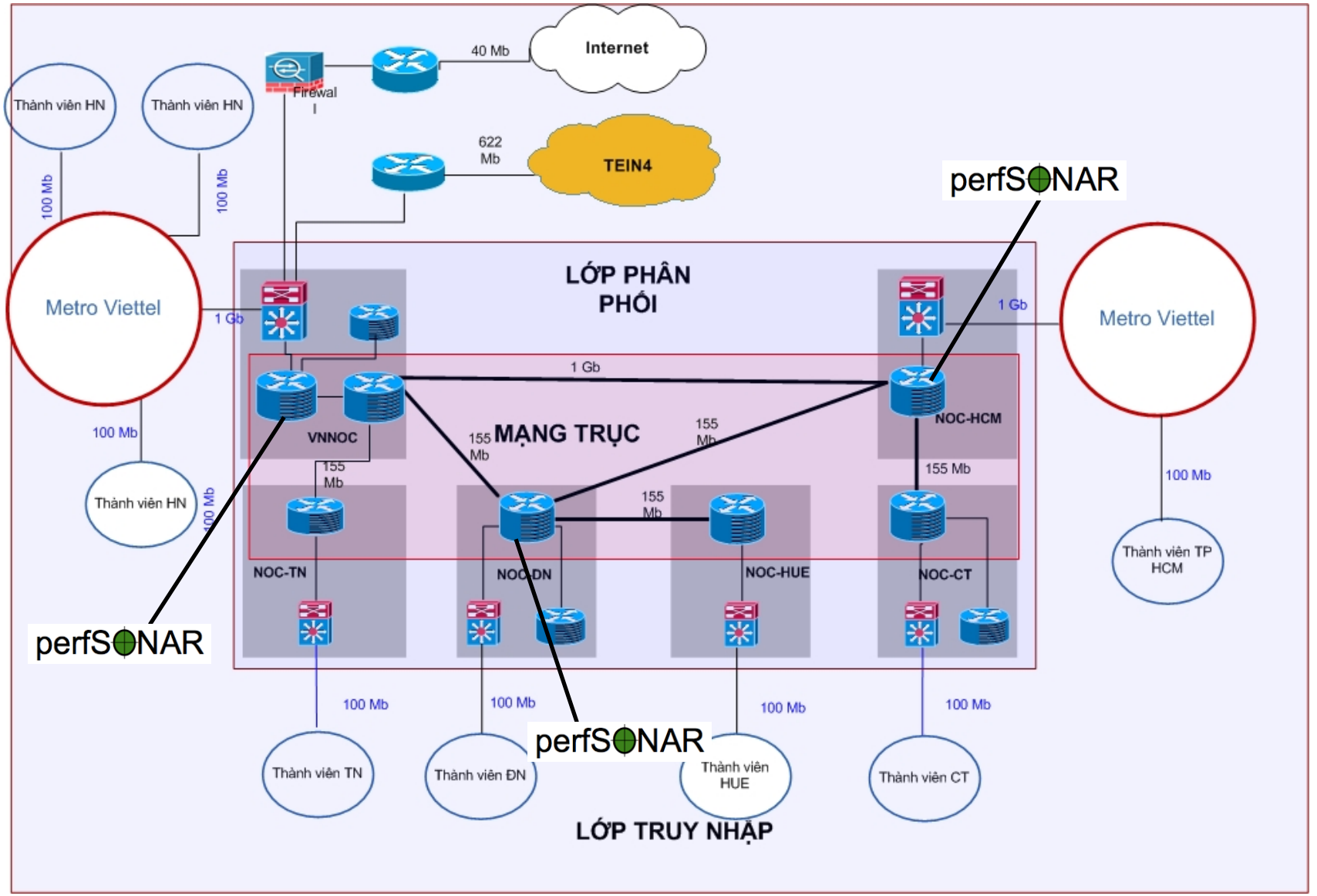
perfSONAR Placement - NREN

- NREN designs vary widely
- NREN should consider placing a perfSONAR node in every place that the NREN has a backbone or customer aggregation router

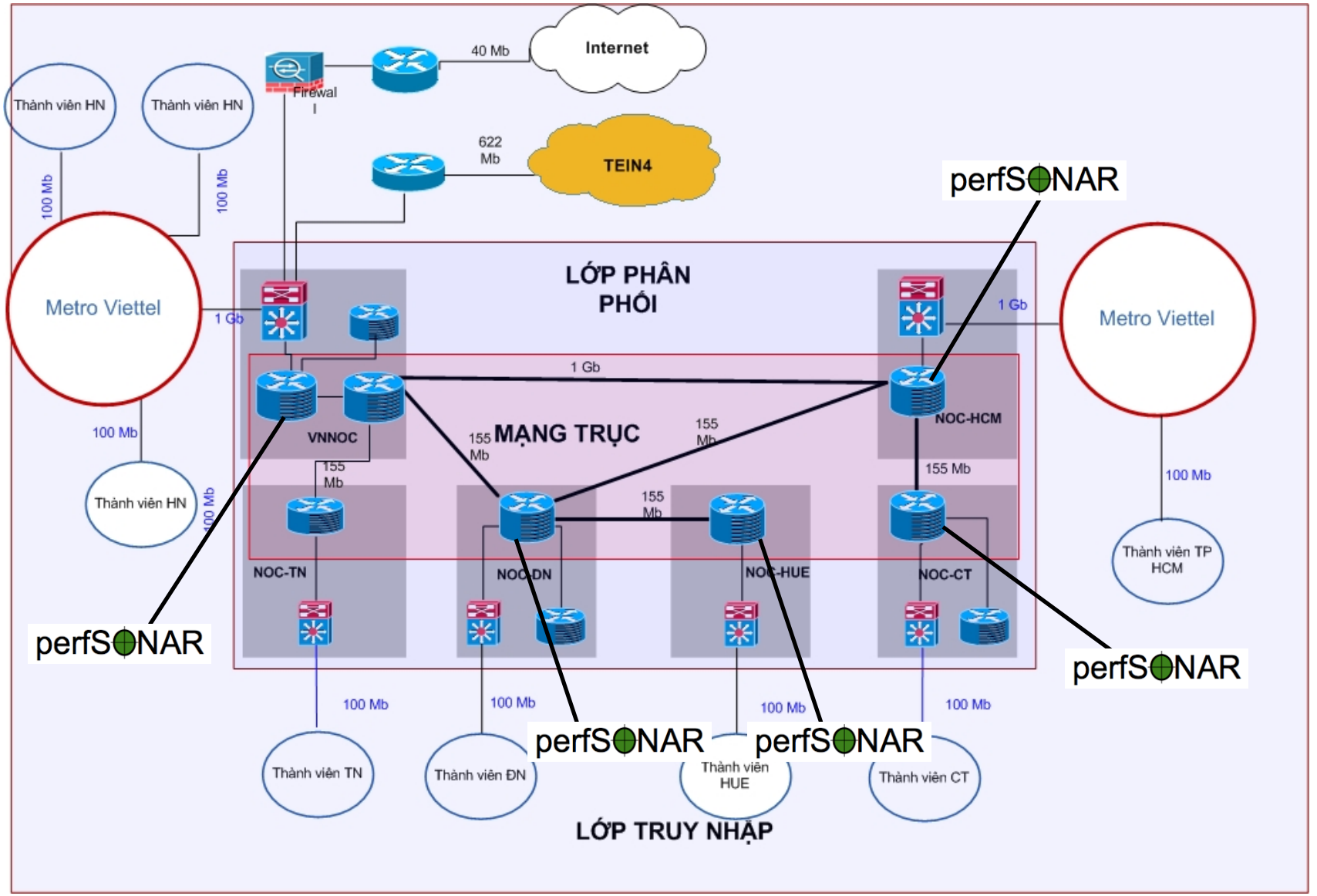
perfSONAR Placement - NREN



perfSONAR Placement - NREN



perfSONAR Placement - NREN



But what about Security?

- Just because there is no firewall doesn't mean you can't do security
  - Firewalls have security policies that say "allow this", "deny that"
    - That looks a lot like an access control list (ACL) on a router
    - You can duplicate most firewall policies using ACLs on routers
- You can do security without firewalls!

Questions?

UNIVERSITY OF OREGON

NSRC
Network Startup Resource Center

Symbols to use for diagrams