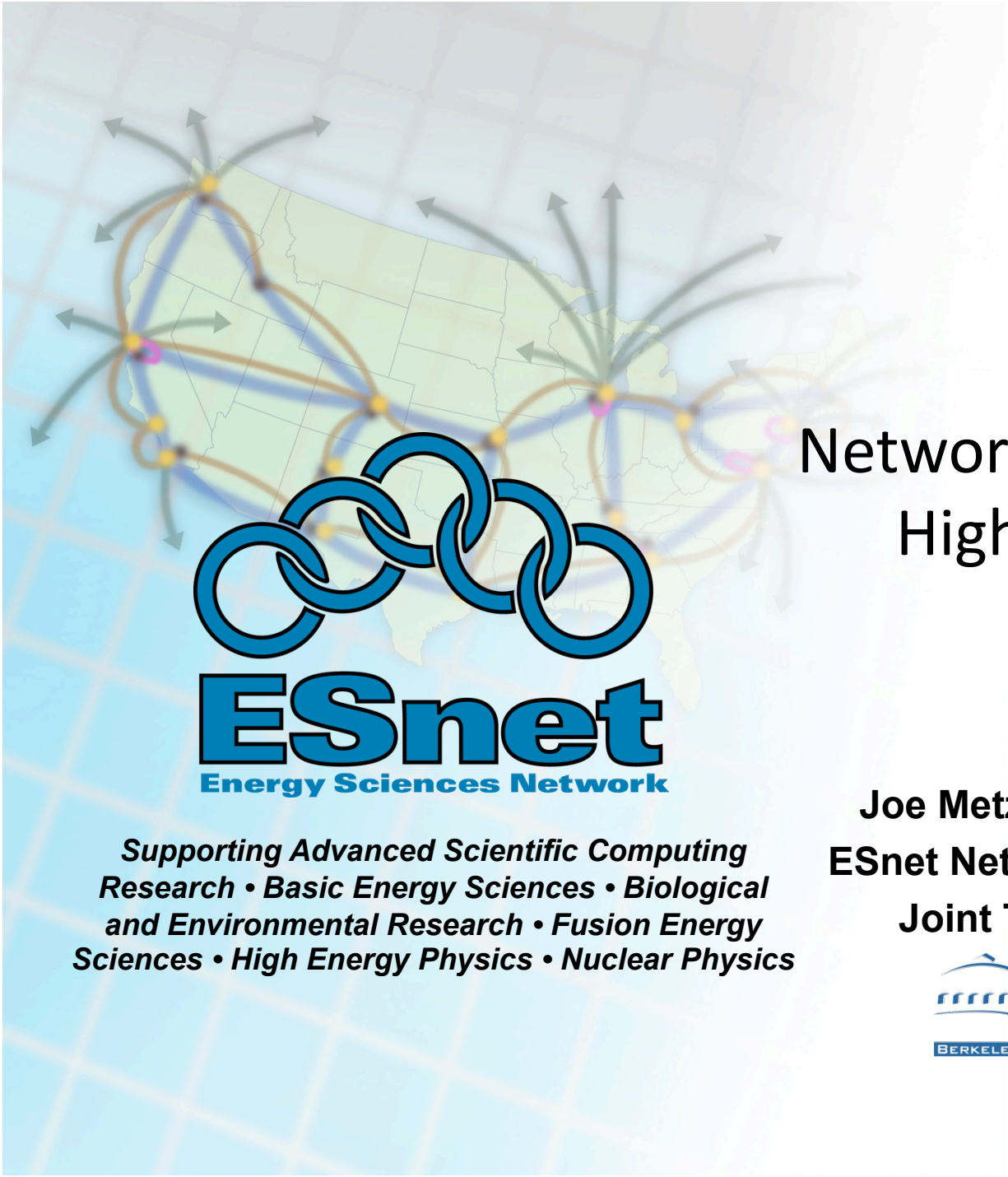


ESnet
Energy Sciences Network

Supporting Advanced Scientific Computing Research • Basic Energy Sciences • Biological and Environmental Research • Fusion Energy Sciences • High Energy Physics • Nuclear Physics

Network Architecture for High Performance

Joe Metzger, Network Engineer
ESnet Network Engineering Group
Joint Techs – July 28, 2009

Berkeley Lab

U.S. Department of Energy
Office of Science

Overview

- Architectural Considerations
- Network Troubleshooting
- Test and Measurement
- Host Tuning

Architecture – Motivation

- ESnet has helped several different customers resolve network performance problems
  - Widely different communities
  - Patterns have emerged
- Architectural considerations for high performance science infrastructures
  - Science vs. Enterprise networks
  - Different requirements

Enterprise Network Requirements

- Business continuity
  - Risk management
    - Personally Identifiable Information (PII)
    - Financial information
    - Embarrassment due to security incidents
  - Relatively low bandwidth (100s of Mbps)
- Unsophisticated user base from a computer security perspective
  - Lots of desktop boxes
  - Laptops, visitors (hosts that visit other networks)
- Need network-level policy controls to mitigate risk (e.g. firewall protection)

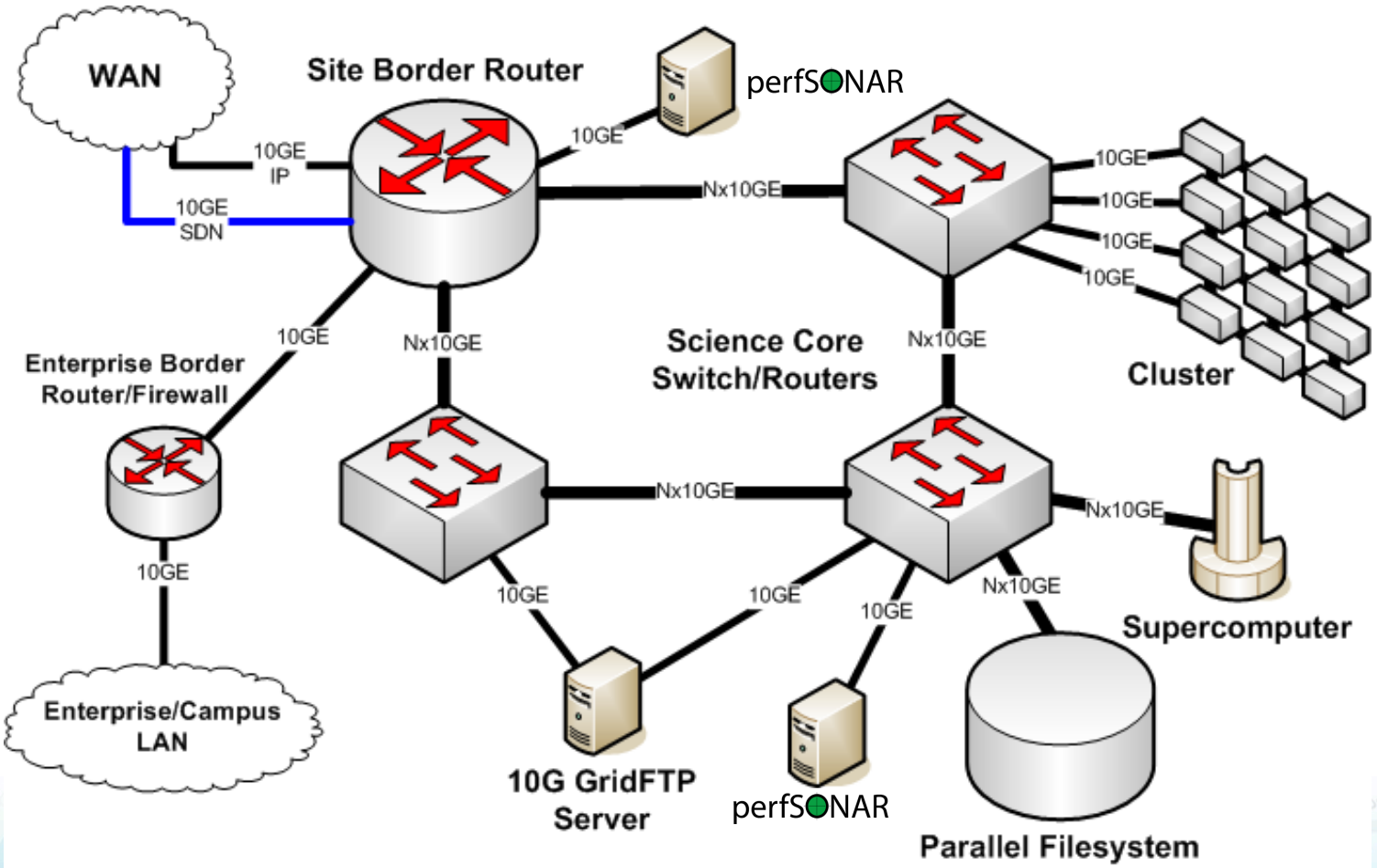
Science Networks are Different

- High bandwidth Requirement (10s of Gbps)
  - Not just in connection speed, but in delivered performance of computational, visualization and storage resources
  - Different protocol/tool set
    - This isn't for desktop boxes
    - Special-purpose data movers
- Sensitive to perturbations caused by security devices
  - Numerous cases of firewalls causing problems
  - Often difficult to diagnose
  - Router filters can often provide equivalent security without the performance impact

Separate Enterprise and Science networks

- Build a science network for the science
- Attach the enterprise network to the science network
  - Put the Enterprise security perimeter at the edge of the enterprise network, not at the site border
  - Science resources are not then burdened by Enterprise firewall configuration

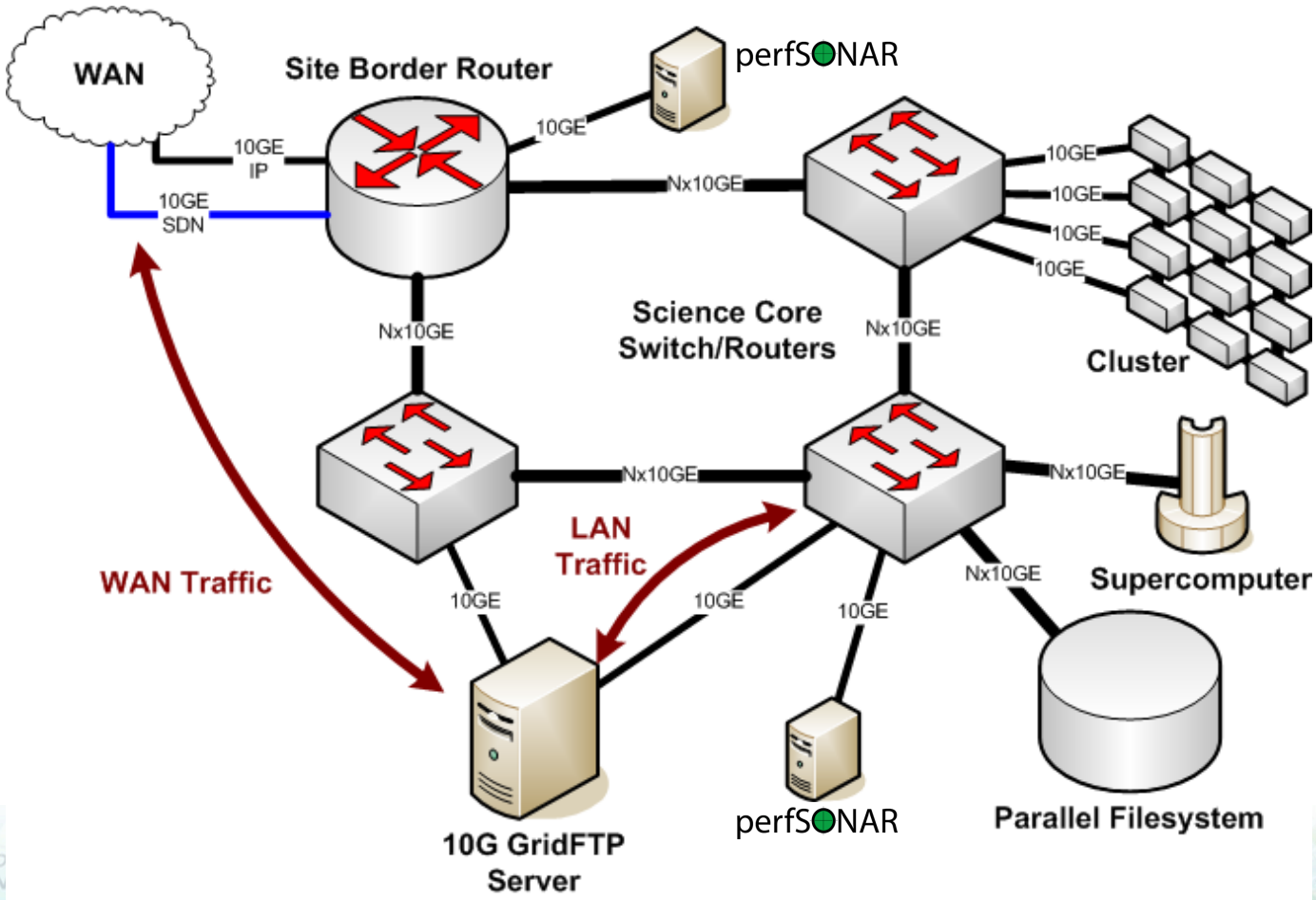
Separate Enterprise and Science Networks



Separate LAN and WAN Traffic

- Use multiple host interfaces to eliminate problems caused by high-speed LAN flows saturating interfaces
  - WAN-facing interface used for WAN flows (e.g. remote users and resources)
  - LAN-facing interface used for LAN flows (e.g. access to local parallel filesystem)
  - LAN flows must not saturate WAN-facing interface!
- Dedicated resources for WAN transfers eliminate local system dependencies (e.g. WAN transfers to supercomputer interactive nodes)

Internal / External Traffic Separation



LAN Troubleshooting

- High performance between LAN hosts does not mean your network is clean
  - TCP recovers very quickly at LAN latencies
  - This hides packet loss
  - The same loss with WAN latencies causes crippling performance problems
- Router and Switch configuration is key
- Test and measurement hosts can help locate loss in the LAN, but only by testing to WAN destinations
- Testing helps locate soft failures

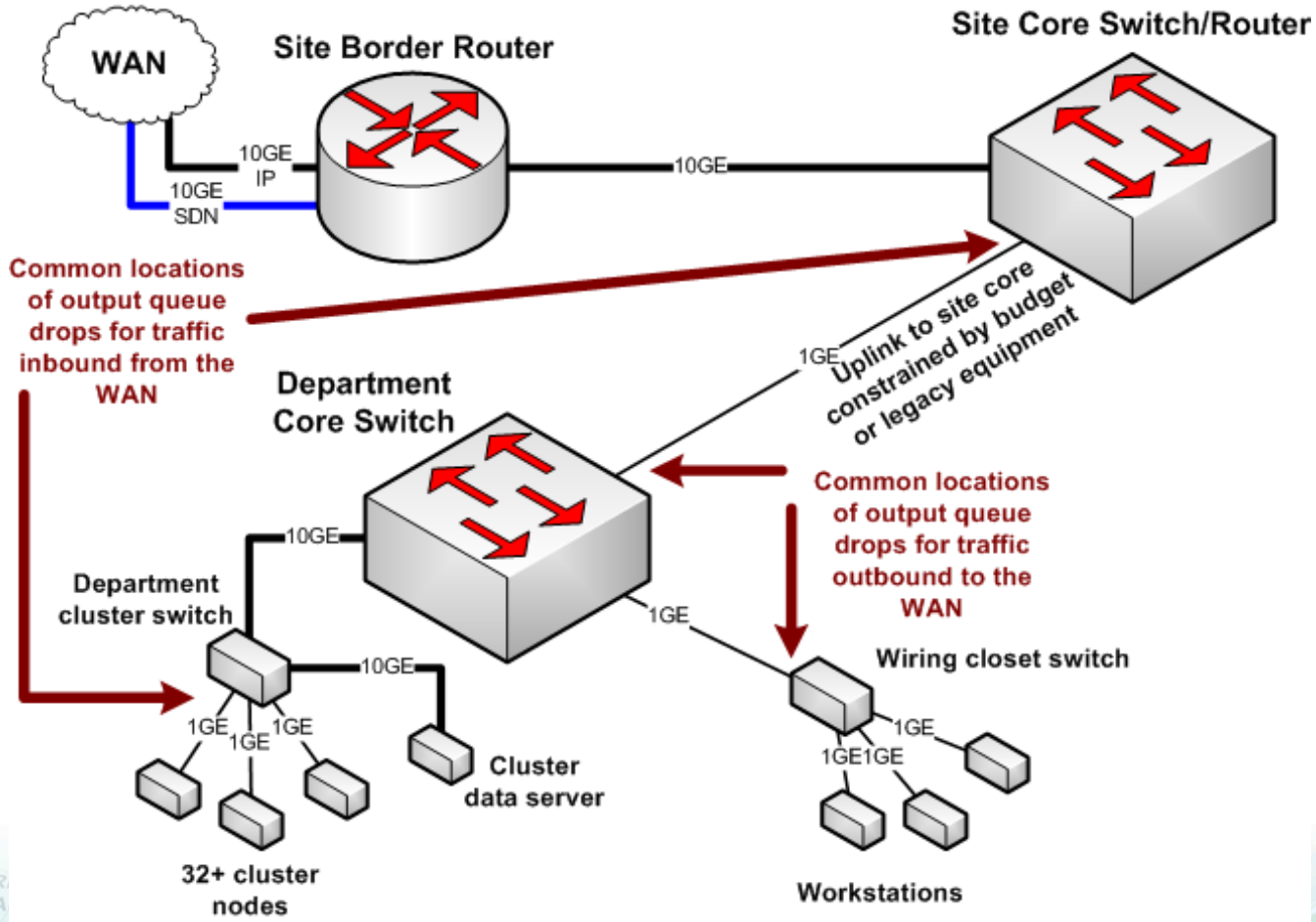
Soft Failures

- "Soft Failures" are network problems that don't result in total loss of connectivity
  - The network (or a particular router or link) is up, but does not perform well
  - Packet loss often goes unnoticed until you try to use the WAN for high throughput
- Soft failure examples
  - Process switching ("punting")
  - Dirty fiber
  - Failing optics
  - Misconfigured buffers/queues
  - Routing table overflow in Cisco devices (causes punting)

Router and Switch Configuration

- Buffer/queue management
  - TCP traffic is bursty
  - Coincident bursts destined for a common egress port cause momentary oversubscription of the output queue
  - Momentary oversubscription of output queues causes packet loss
  - Default configuration of many devices (e.g. Cisco routers running IOS) is inadequate
- Cisco commands
  - ‘sho int sum’ – check for output queue drops
  - ‘hold-queue 4096 out’ – change from default 40-packet output queue depth
- See http://fasterdata.es.net/cisco.html

Output Queue Oversubscription



Performance Limitations

- You get what you pay for
  - Cheap switches cause performance problems
    - Small buffers
    - Inadequate diagnostics
  - Cheap disks don't perform
    - Theoretically, most SATA disks have 3GB/sec transfer rate
    - In reality, they are more than 10x slower than that
- Firewalls are meant to stop traffic
  - Often have inadequate buffering
  - Often slower than wire-speed
  - Even if it says "10G Firewall" on it, the device probably has per-flow capacity of much less than 10G

Network Test and Measurement

- Test and measurement tools
  - PerfSONAR measurement hosts
  - Multiple measurement hosts are very helpful
    - One host at site border
    - One host near Science resources
  - Monitor data transfer performance (NetLogger)
- Regular throughput testing is key
  - Changes in performance are visible when they occur
  - Timing of performance changes helps identify causes
- See http://fasterdata.es.net/ps_howto.html

Network Performance – Host Tuning

- TCP autotuning should be turned on by default
  - Typically as good or better vs. hand tuning
  - Eliminates per-destination tuning overhead
- Modern Congestion Recovery is key
  - Cubic, HTCP perform well
  - Except when there are kernel bugs – see http://fasterdata.es.net/TCP-tuning/linux.html
- If you don't do the host tuning, the network can't help you
  - Linux distribution maintainers keep defaults low
  - Low default parameters mean hosts can't effectively use a high-performance network without manual intervention
- See http://fasterdata.es.net/

Changing Host Defaults is Simple

- Changing a few parameters makes a huge difference, often 4x to 10x improvement
- Just adding the following to /etc/sysctl.conf will usually help a lot:
  - (Linux systems with kernel version 2.6.x)
  - net.core.rmem_max = 4194304
  - net.core.wmem_max = 4194304
  - net.ipv4.tcp_rmem = 4096 87380 4194304
  - net.ipv4.tcp_wmem = 4096 65536 4194304
- Necessary but not necessarily sufficient
- http://fasterdata.es.net/tuning.html

Data Transfer – Use Dedicated Resources

- Any to any high-performance data transfers are hard
  - Most resources are configured for high-speed LAN performance (e.g. no stack tuning)
  - Some operating systems do not have TCP autotuning, or have other constraints that make TCP stack tuning difficult (e.g. AIX HPSS movers)
- WAN traffic has different requirements
  - LAN traffic recovers quickly from loss – latency is low enough that LAN loss is often invisible
  - WAN flows collapse with even minor packet loss
  - WAN transfer resources should be placed as close to the site border as is feasible

Questions?

- Thanks!