Network Support for Clouds

Eli Dart, Network Engineer

ESnet Engineering Group

HPC and Cloud Computing Workshop

Berkeley, CA

June 20, 2011

Overview

Clouds from a network perspective

- Clouds imply networks
- Virtual Circuits – network services for science

Data Intensity

- Data Intensity as a driver for services
- Science DMZ
- Data Transfer Node (DTN)

Futures – The Advanced Networking Initiative

Clouds from a network perspective

A “cloud” looks similar to any other scientific instrument

- One or more data transfer hosts
- Need for significant data movement

Clouds imply networks

- Without networks there are no clouds
- Service interface to the network → flexibility
- Network services enable provisioning, virtualization, scheduling

Research and Education (R&E) networks provide high bandwidth as a given

Services increase the utility of networks, clouds, etc.

What Is “The Network” Anyway?

From the perspective of a user, “The Network” is not a bunch of routers, switches, fiber, and so on

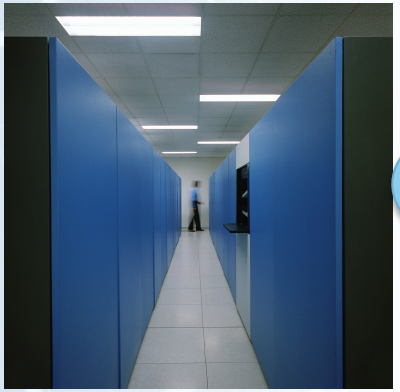
“The Network” is the thing that is broken when remote data transfers are hard (That’s cute and all, but what does this really say?)

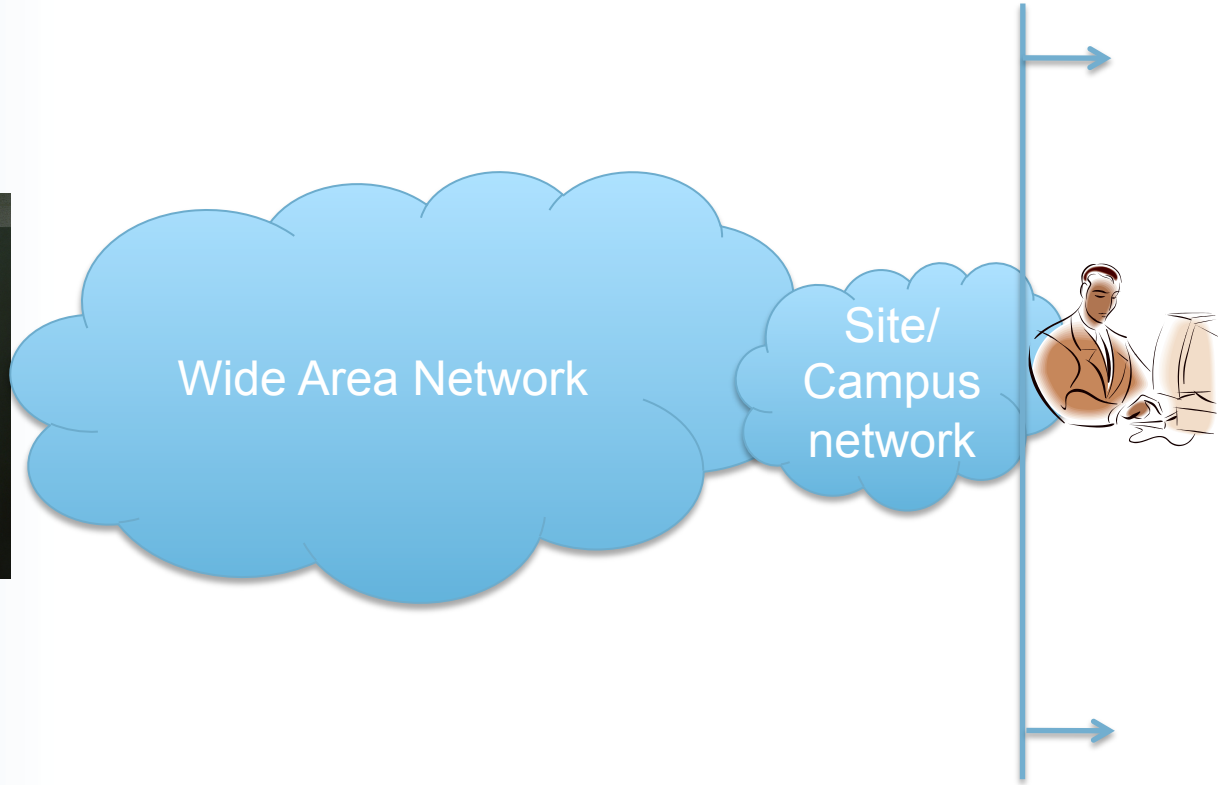
- “The Network” is the set of devices and applications involved in the use of a remote service (e.g. Cloud service or HPC Cloud resource)
- The primary user interface to “The Network” is typically a data transfer tool or other user-agent application that talks to a tool or application on a different host

Therefore, the utility of “The Network” or “The Cloud” depends on the existence and availability of well-configured end systems, and network services to support them

An end-to-end view on Cloud Computing

Distributed
Data Center

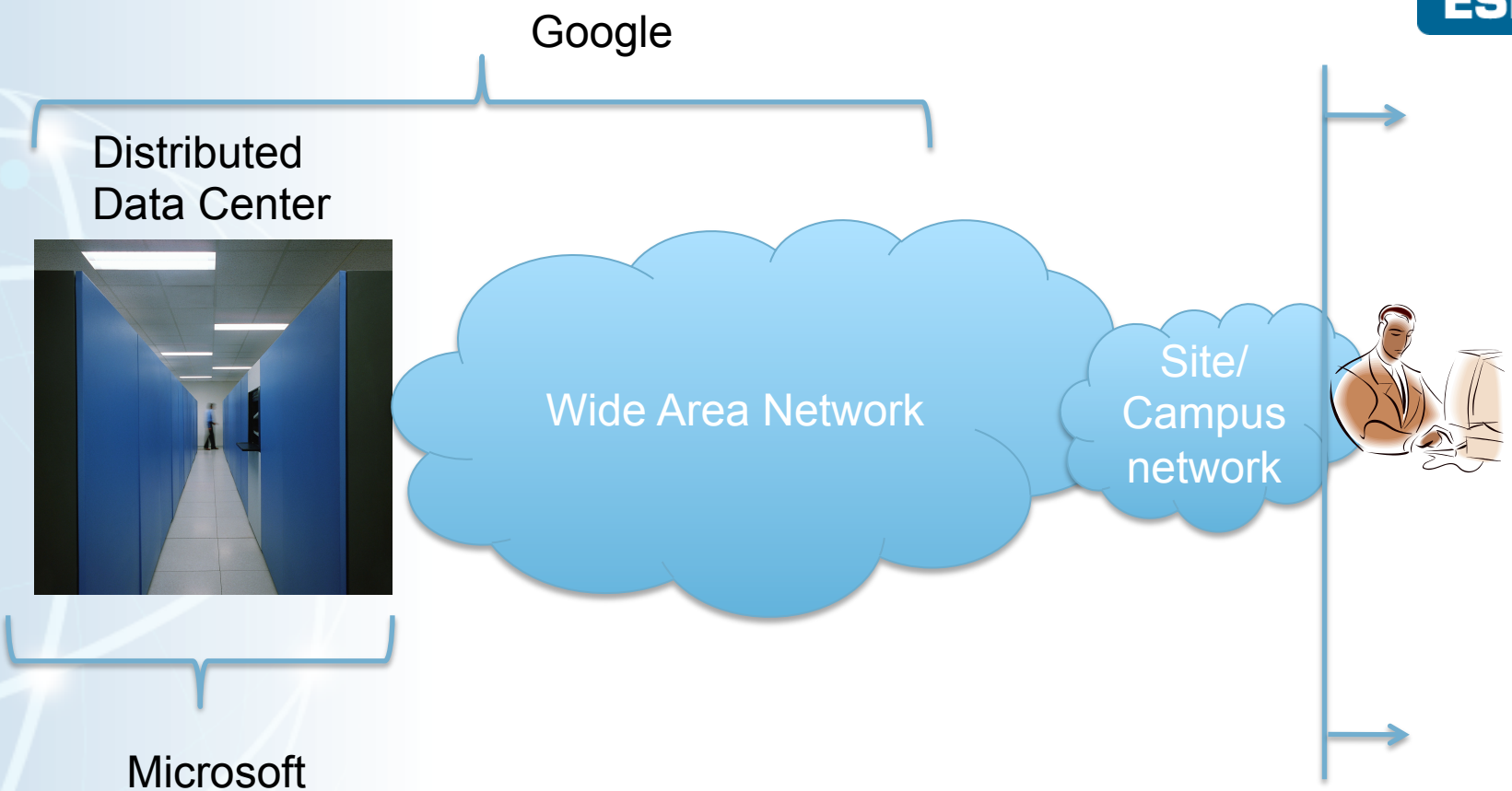




This is the cloud that
everyone thinks about!

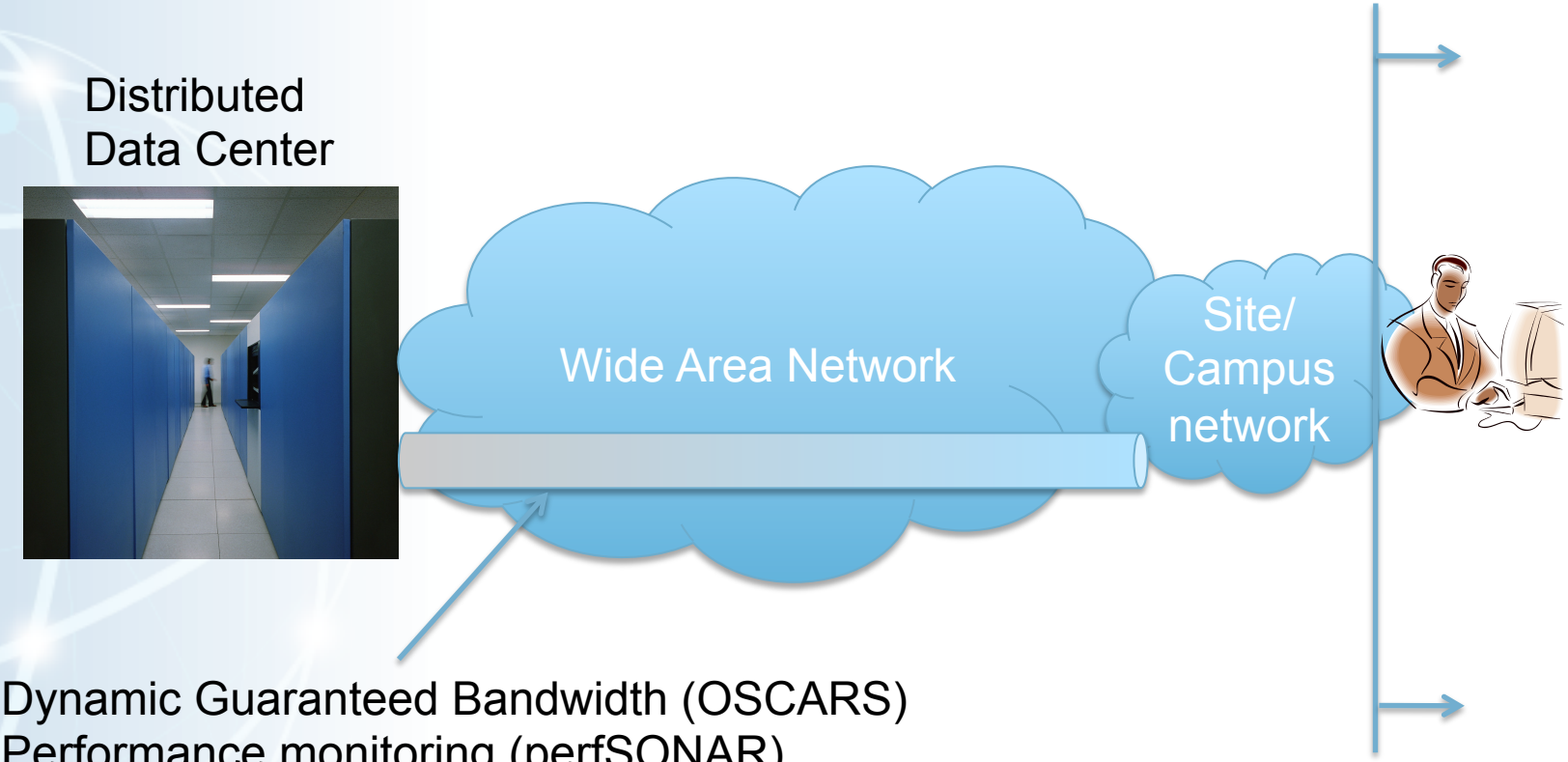
User experience =
Σ (Application +
Data center +
Campus + WAN)

User Experience Varies



User experience =
Σ (Application +
Data center +
Campus + WAN)

ESnet's Open-source toolset for improving user experience

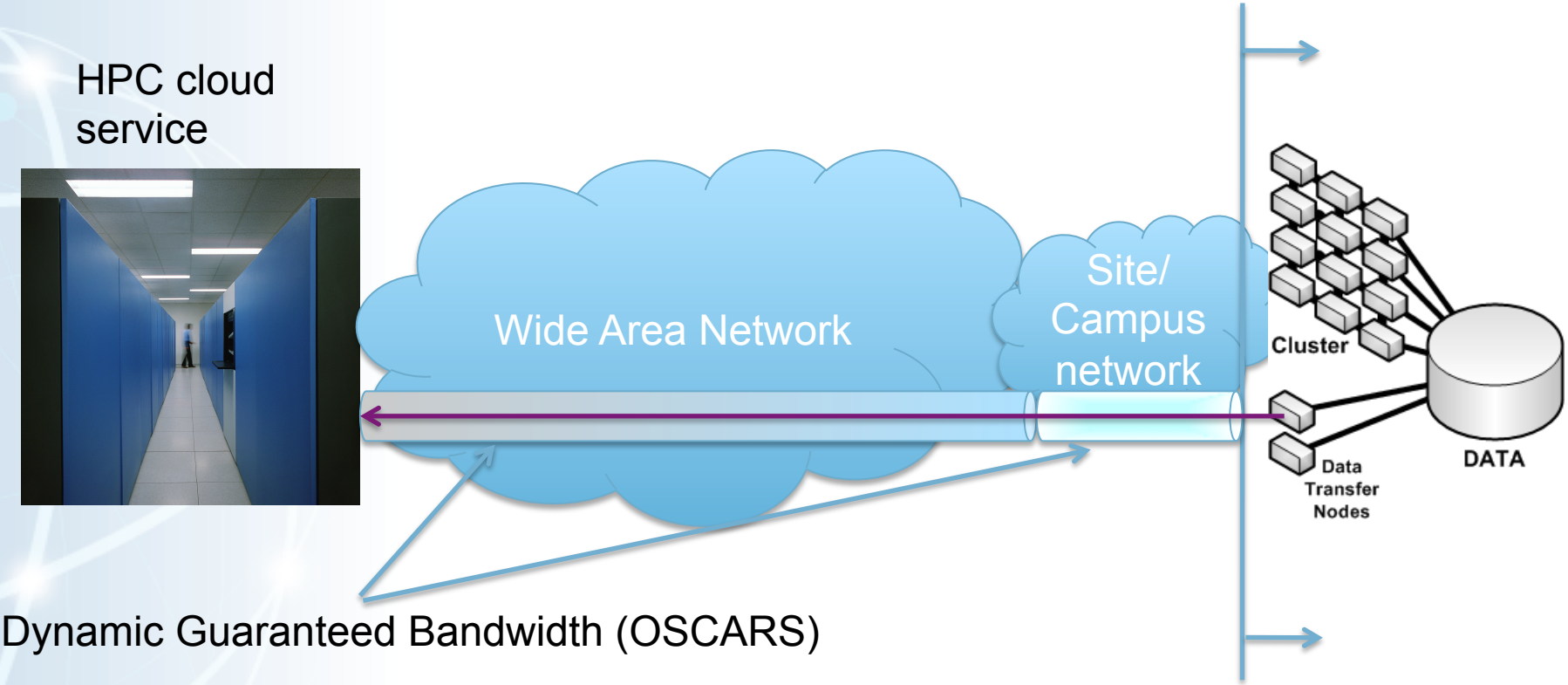


Dynamic Guaranteed Bandwidth (OSCARS)
Performance monitoring (perfSONAR)

Multi-domain capability (campus to DC)

User experience =
Σ (Application +
Data center +
Campus + WAN)

HPC use of cloud resources



User experience =
Σ (Application +
Data center +
Campus + WAN)

Virtual Circuits

Virtual circuits defined

- Benefits of circuits
- Current production use (e.g. LHC)

Multi-domain aspects

- VLAN tag as interdomain demarcation technology
- Enables per-domain implementations
- Controller provides topology, stitching, reservations, etc

OSCARS – ESnet production virtual circuit service

Virtual Circuits Defined

Dedicated paths for data

- A virtual circuit is a dedicated path or channel through a set of networks, the use of which is exclusive to the systems at the ends of the circuit
- Common example is an end to end VLAN
- Layer3 circuits exist as well – e.g. guaranteed bandwidth between two data transfer clusters, without changing config or addressing of the clusters

Multi-domain aspects

- Large-scale science is inherently multi-site, multi-domain
- Therefore, in order to be useful, virtual circuits must traverse multiple administrative domains in the general case

Benefits of Virtual Circuits

Bandwidth and service guarantees

- Virtual circuits can provide guaranteed bandwidth to applications on a scheduled basis
- Explicit paths for diversity (ensure no single failure disconnects two endpoints)

Service interface to the network

- A machine interface to virtual circuits can make middleware, schedulers, etc. network-aware
- Systems can make use of the network in ways that are more intelligent – network can be virtualized

Current Production Use of Circuits

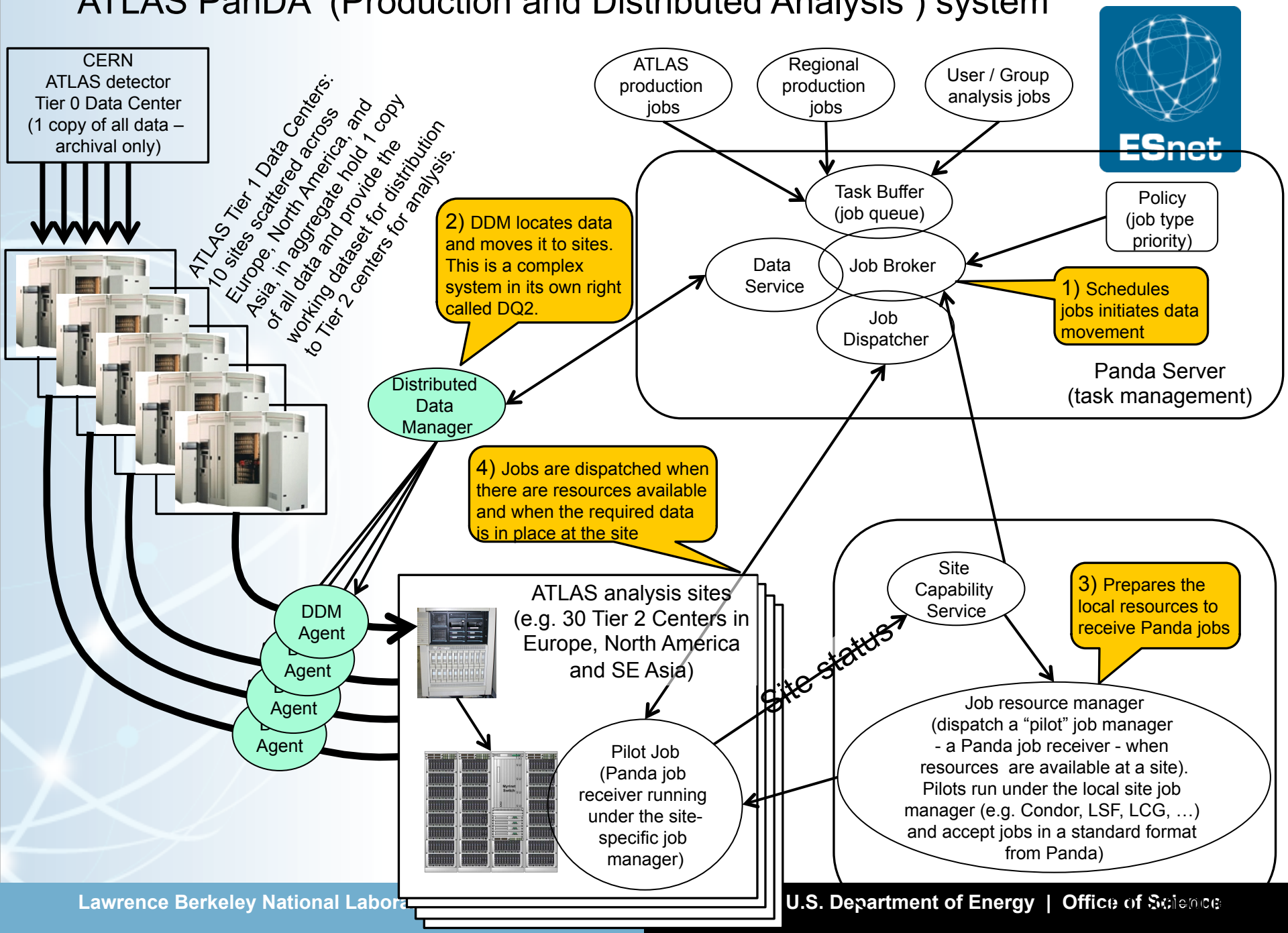
Largest single customer is the Large Hadron Collider

- LHCOPN (LHC Optical Private Network) provides a dedicated infrastructure for data transfer between CERN (Tier 0) and the LHC Tier 1 centers which distribute data to scientific sites
- This has performed well under multi-failure conditions
- Bandwidth guarantees, explicit paths provide robust service
- Service from Tier 1 centers to Tier 2 sites is guaranteed by circuits as well – each site is guaranteed a given level of service

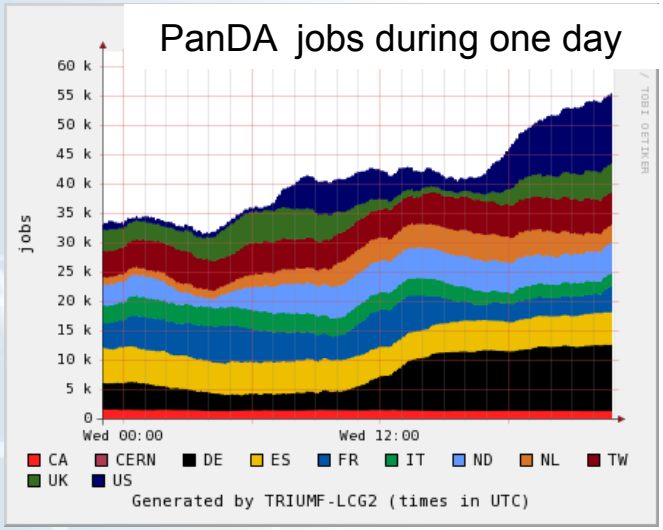
Other disciplines use circuits also

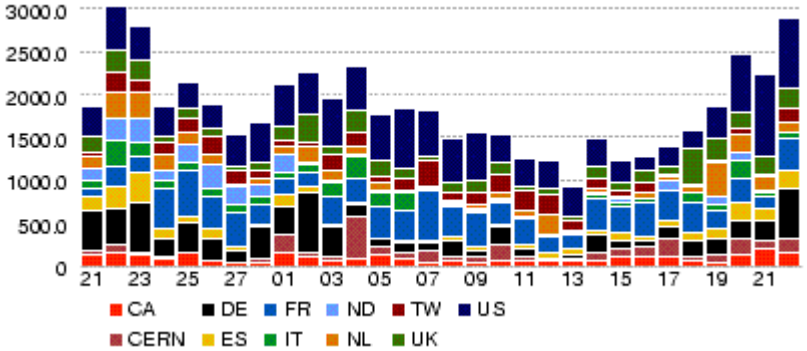
- Remote filesystem mounts between sites for data analysis
- Dedicated paths for network engineering purposes (e.g. route around expected outages before service failure occurs)

ATLAS PanDA (Production and Distributed Analysis) system

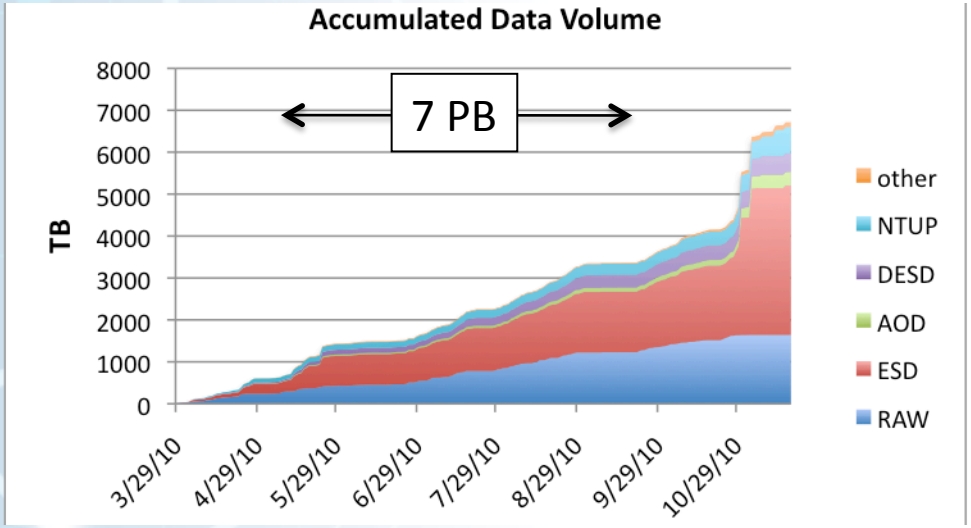


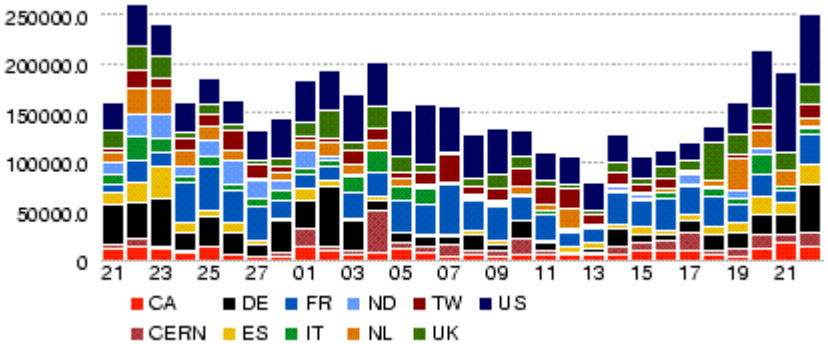
Scale of ATLAS Data Analysis





Tier 1 to Tier 2 throughput (MB/s) by day – up to 24 Gb/s – for all ATLAS Tier 1 sites





Data Transferred (GBytes) (up to 250 Tby/day)

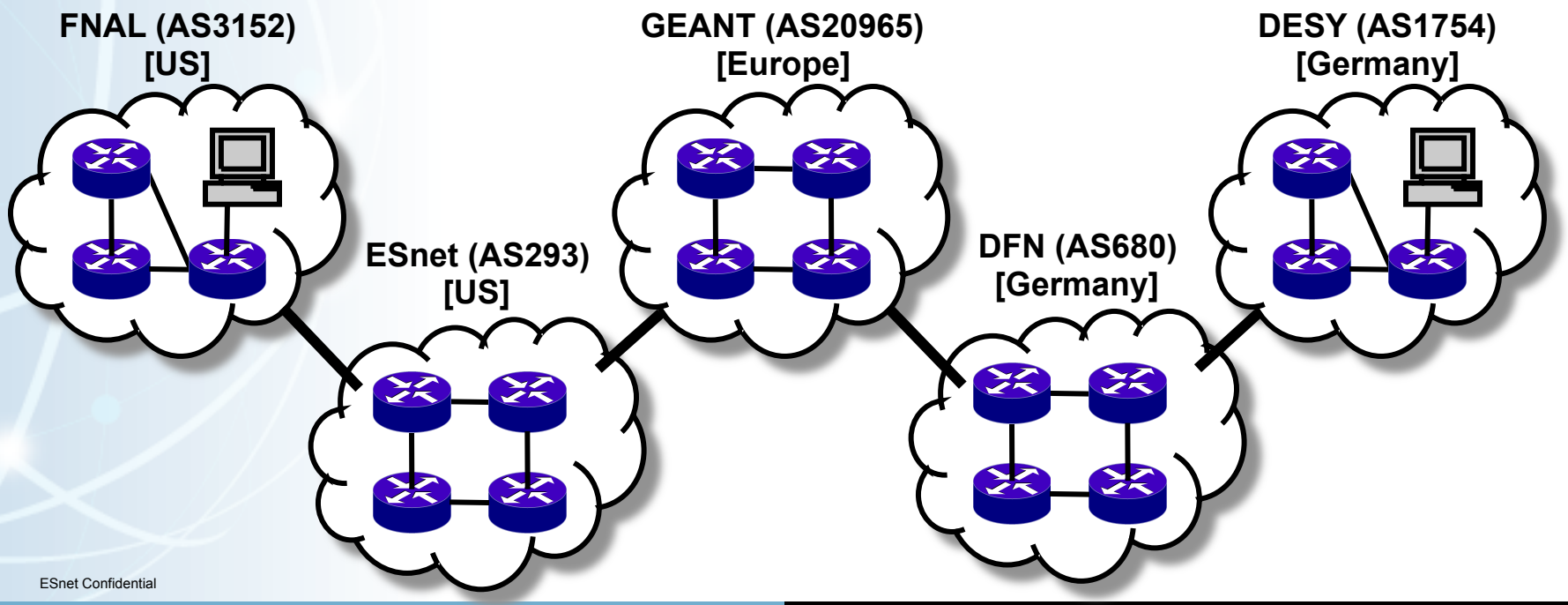
It is this scale of data movement and analysis jobs, going on 24 hr/day, 9+ months/yr, that networks must support in order to enable this sort of large-scale science

Environment of Science is Inherently Multi-Domain

End points will be at independent institutions – campuses or research institutes - that are served by ESnet, Internet2, GÉANT, and their regional networks

Complex inter-domain issues – typical circuit will involve five or more domains

For example, a connection between FNAL and DESY involves five domains, traverses four countries, and crosses seven time zones



Multi-domain Aspects

Each network may have its own circuit implementation

- MPLS
- VLANs
- SONET time slots

There must be a standard method of interconnection between domains

- Per-domain implementations must be able to exchange traffic
- The standard technology is a layer2 VLAN tag
- VLAN tag is service demarcation point at provider edge – this means that per-network implementations need not be directly compatible
- Requires topology exchange to “stitch” end to end circuits (this is done by InterDomain Controllers – IDCs)

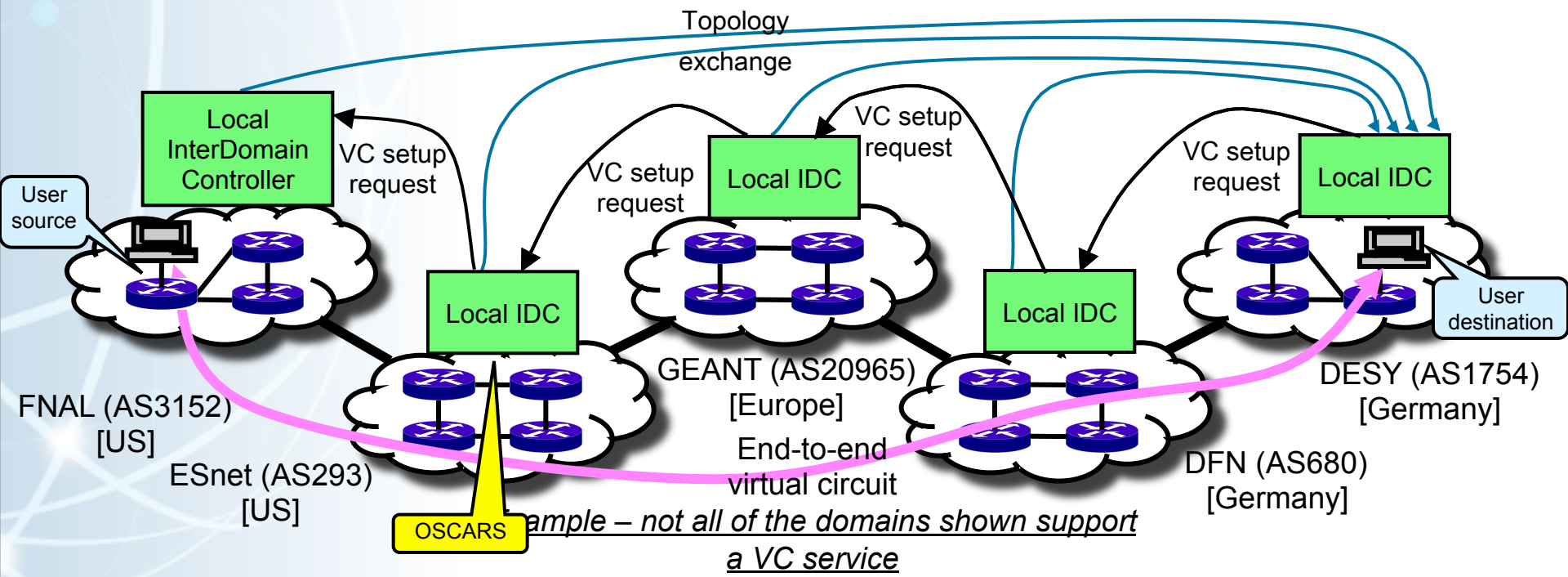
Interdomain Circuits via Federated IDCs

Inter-domain interoperability is crucial to serving science and is provided by an effective international R&E collaboration

In order to set up end-to-end circuits across multiple domains:

1. The domains exchange topology information containing at least potential VC ingress and egress points
2. VC setup request (via IDC protocol) is initiated at one end of the circuit and passed from domain to domain as the VC segments are authorized and reserved

A "work in progress," but the capability has been demonstrated



ESnet OSCARS – a service-oriented virtual circuit service

Guaranteed, reservable bandwidth with resiliency

- User specified bandwidth and time slot
- Explicit backup paths can be requested
- Paths may be either layer 3 (IP) or layer 2 (Ethernet) transport

Requested and managed in a Web Services framework

Traffic isolation allows for high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport

Secure connections

- The circuits are “secure” to the edges of the network (the site boundary) because they are managed by the control plane of the network which is highly secure and isolated from general traffic
- If the sites trust the circuit service model of all of the involved networks (which, in practice, is the same as that of ESnet) then the circuits do not have to transit the site firewall

Traffic engineering (for ESnet operations)

- Enables the engineering of explicit paths to meet specific requirements, e.g. bypass congested links; using higher bandwidth, lower latency paths; etc.

The OSCARS Service

OSCARS is a virtual circuit service that

- Provides bandwidth guarantees at specified times
- Is capable of communicating with similar services in other network domains to set up end-to-end circuits
- Clearly separates the network device configuration module ("Path Setup") that interacts with the network hardware
  - ESnet uses MPLS transport
  - OSCARS is also implemented in other networks with very different Path Setup modules to configure other sorts of transport (e.g. SONET time slots)

OSCARS is widely deployed

- Foundation for many virtual circuit deployments
- http://www.es.net/services/virtual-circuits-oscars/

Data Intensive Science

Many science disciplines rely on data analysis

- Materials science
- Biomedicine
- Genomics
- High Energy Physics

Data transfer and data sharing are critical to scientific collaborations – in fact, scientific productivity is often determined by the ability to transfer/stream/share data

Data intensity increases are making default configurations obsolete

- Default configurations assume a LAN environment
- Default configurations assume user-agent applications

Dedicated resources allow configuration to match the task

Traditional DMZ (from network security)

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

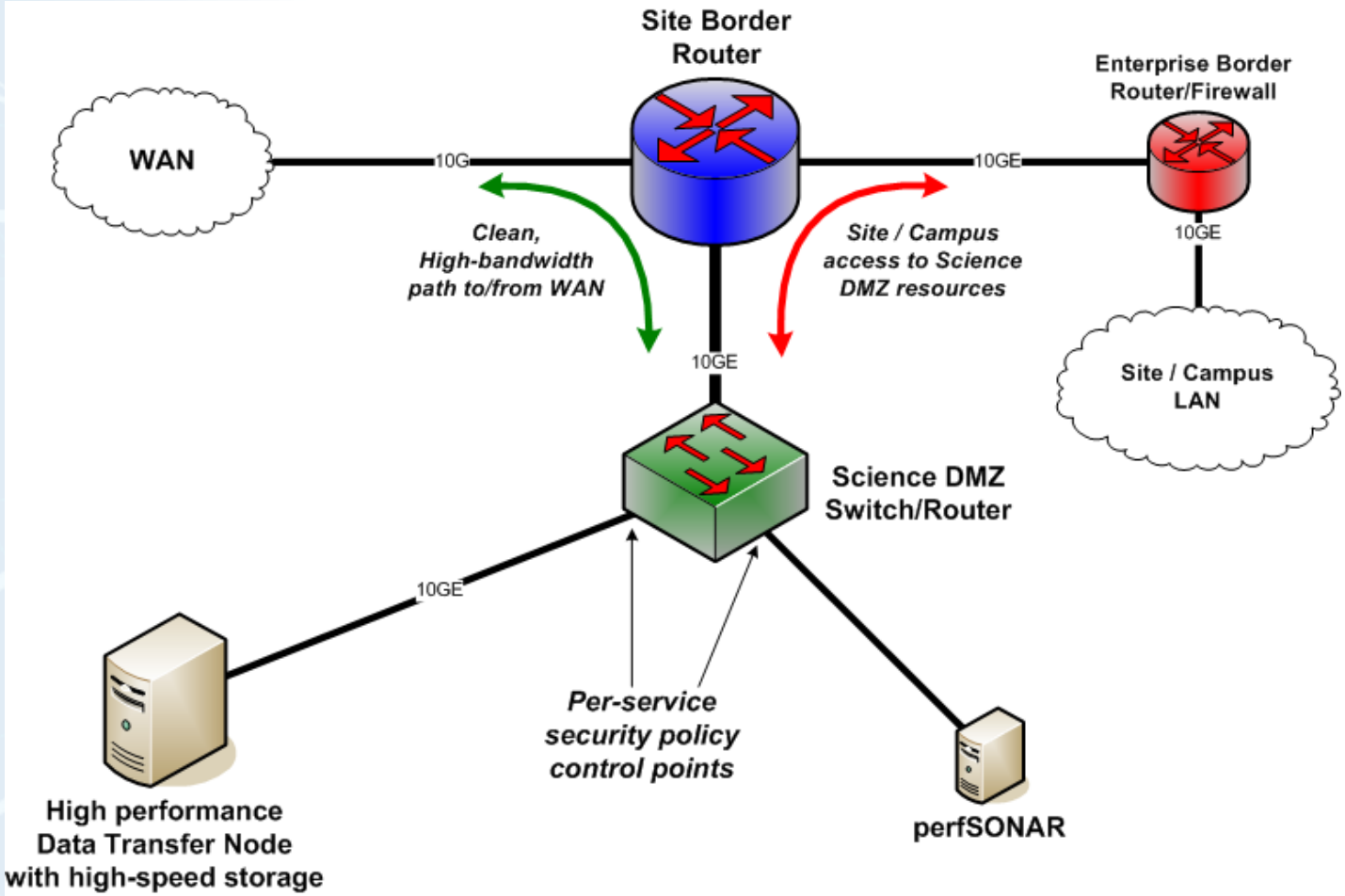
The Science DMZ

Science DMZ – a well-configured location for high-performance WAN-facing science services

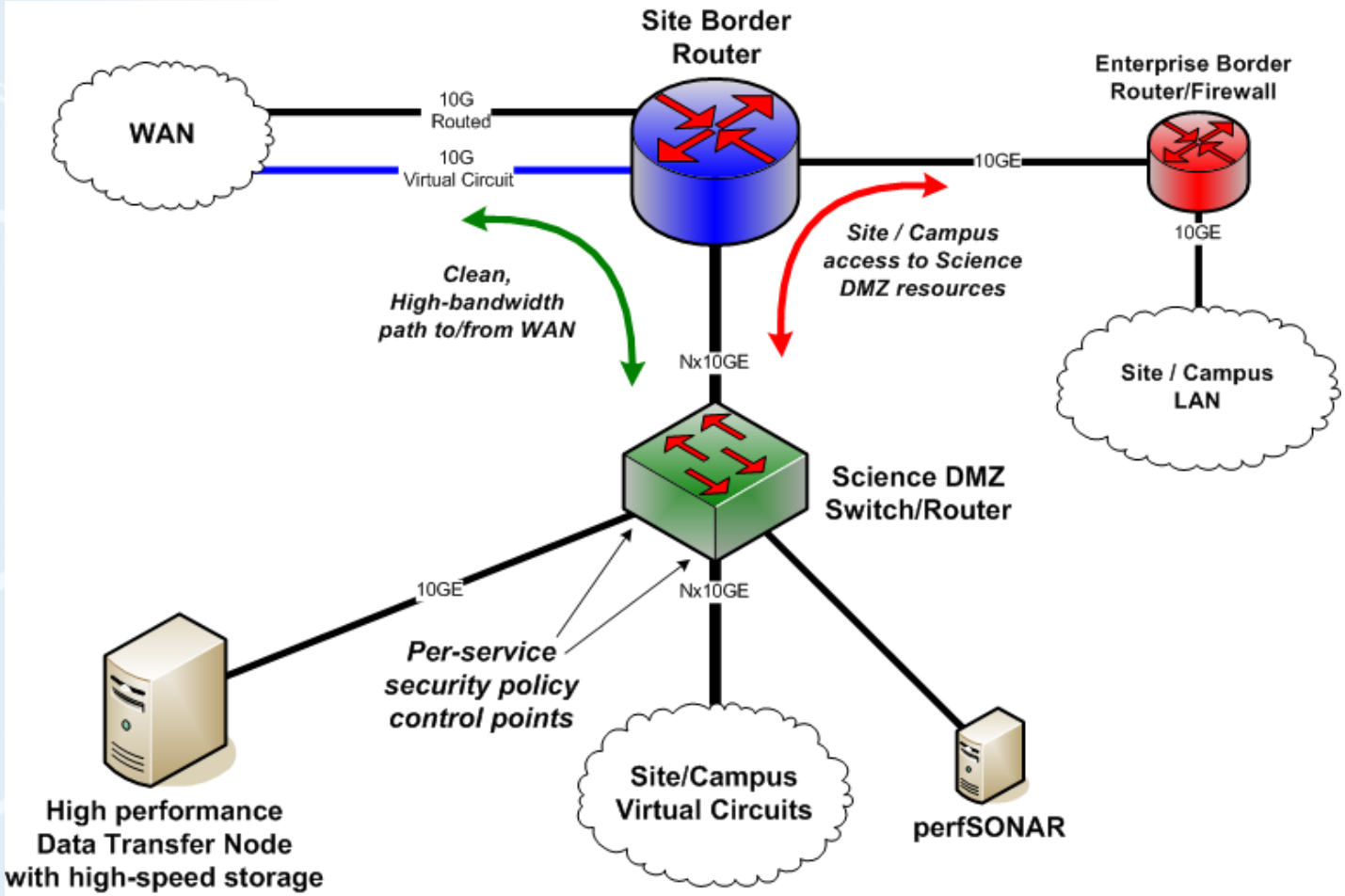
- Located at or near site perimeter on dedicated infrastructure
- Dedicated, high-performance data movers
- Highly capable network devices (wire-speed, deep queues)
- Virtual circuit infrastructure
- perfSONAR for test and measurement

Many high-bandwidth science sites have moved to this architecture already as a matter of necessity – this model is adopted because it works

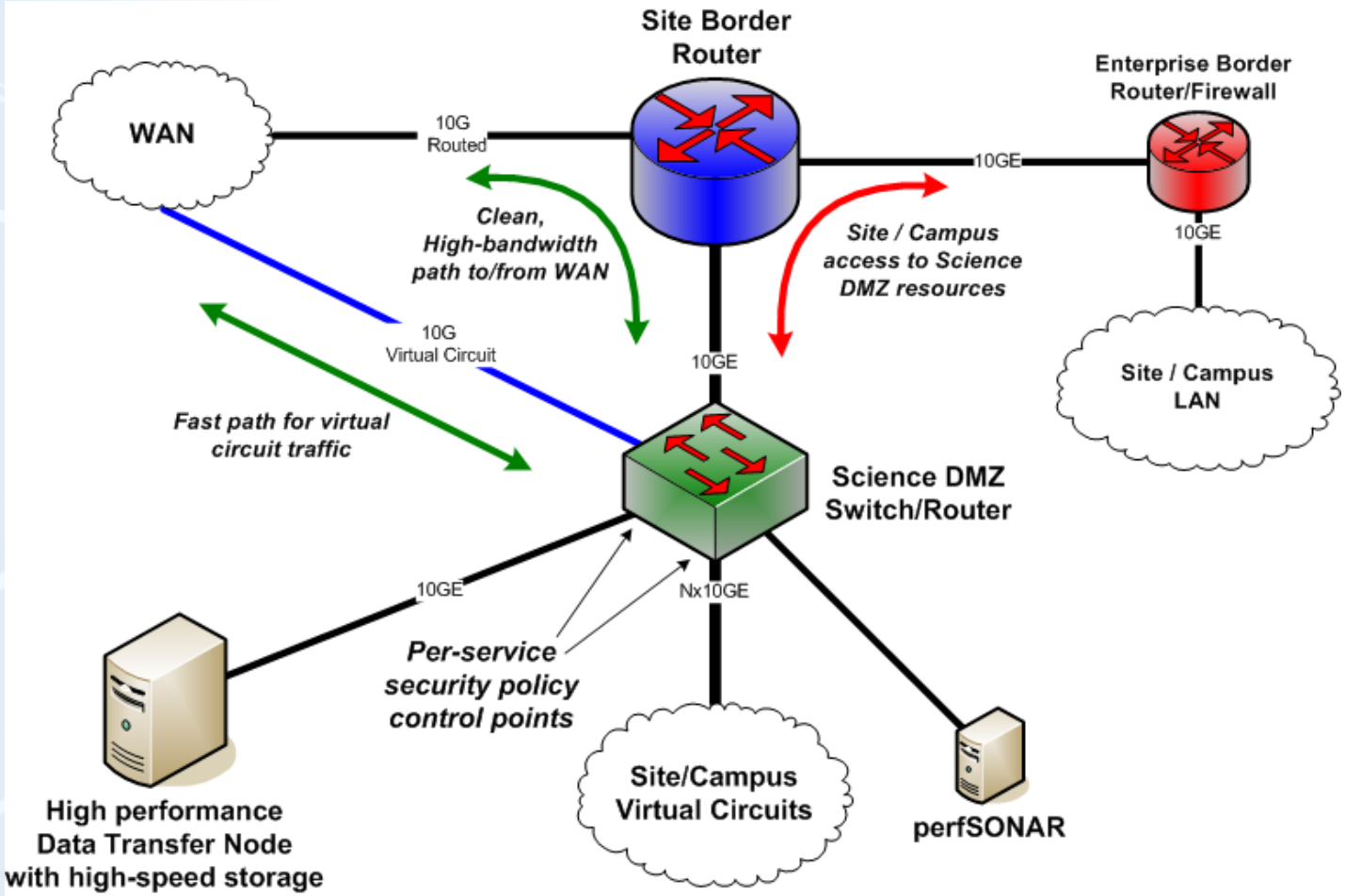
Science DMZ – Conceptual Diagram



Science DMZ - Advanced



Science DMZ – Separate Circuit Connection



Science DMZ Features and Components

Direct connection to site perimeter

- LAN devices eliminated from high-bandwidth data path
  - LAN infrastructure need not be sized for science flows (reduced hardware costs)
  - LAN infrastructure need not be configured to support science flows (e.g. deep output queues for support of science flows can conflict with VOIP requirements)
  - LAN infrastructure need not implement features necessary for virtual circuits (reduced costs, reduced complexity)
- Security policy for science data movers is not conflated with policy for business systems, wireless devices, printers, VOIP, etc

Science DMZ Features and Components

Dedicated infrastructure for science applications

- LAN devices are not part of the troubleshooting mix
- Dedicated devices are easier to configure properly and maintain
- Data Transfer Node for high-performance data movement

Test and measurement deployed on same network infrastructure as science resources

Note well – this is *not* a research project! It is based on hard-won experience

Note also that networks without a monolithic firewall need not add a monolithic firewall just because it's in the picture – it is still very beneficial to move large-scale data transfers close to the site perimeter even if there is no firewall

The Data Transfer Node (DTN)

Dedicated, high-performance host for long-distance data transfer

High performance disk, for example:

- High-speed local RAID
- Fibrechannel attachment to SAN, if available
- Lustre or GPFS filesystem mount (e.g. when deployed at supercomputer center)

High-speed network connection (1G or 10G)

- Connected to Science DMZ
- Separate security policy from business traffic

Multiple sites and facilities are deploying DTNs (Supercomputer centers, labs, experiments, etc)

Significant performance gains from DTN deployment

Science DMZ Security

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

Rationale

- Science flows are relatively 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, staff desktops, etc.
- Security controls that are typically implemented to protect business resources often cause performance problems
- Sizing security infrastructure on business networks for large science flows is expensive

Advanced Networking Initiative

$62.4M in ARRA funding

Two components:

- 100G Prototype Network
  - 4 sites (Supercomputing centers, exchange point in NY)
  - Stretch goal of including other DOE Office of Science sites
- Network Research Testbed
  - Dark fiber IRU
  - Infrastructure for researchers
  - Advisory committee

ANI Testbed Overview

Consists of 3 Phases

- Phase 1: “Tabletop testbed” at LBL (June 2010 to Summer 2011)
- Phase 2: Move to Long Island MAN when dark fiber is available (Summer 2011)
- Phase 3: Extend to WAN when 100Gbps available (late 2011)

Capabilities

- Ability to support end-to-end networking, middleware and application experiments, including interoperability testing of multi-vendor 100Gbps network components
- Researchers get “root” access to all devices
- Use Virtual Machine technology to support custom environments
- Detailed monitoring capabilities

Sample Projects

Examples of the types of projects the current testbed will support include the following:

- Path computation algorithms that incorporate information about hybrid layer 1, 2 and 3 paths, and support 'cut-through' routing
- New transport protocols for high speed networks
- Protection and recovery algorithms
- Automatic classification of large bulk data flows
- New routing protocols
- New network management techniques
- Novel packet processing algorithms
- High-throughput middleware and applications research

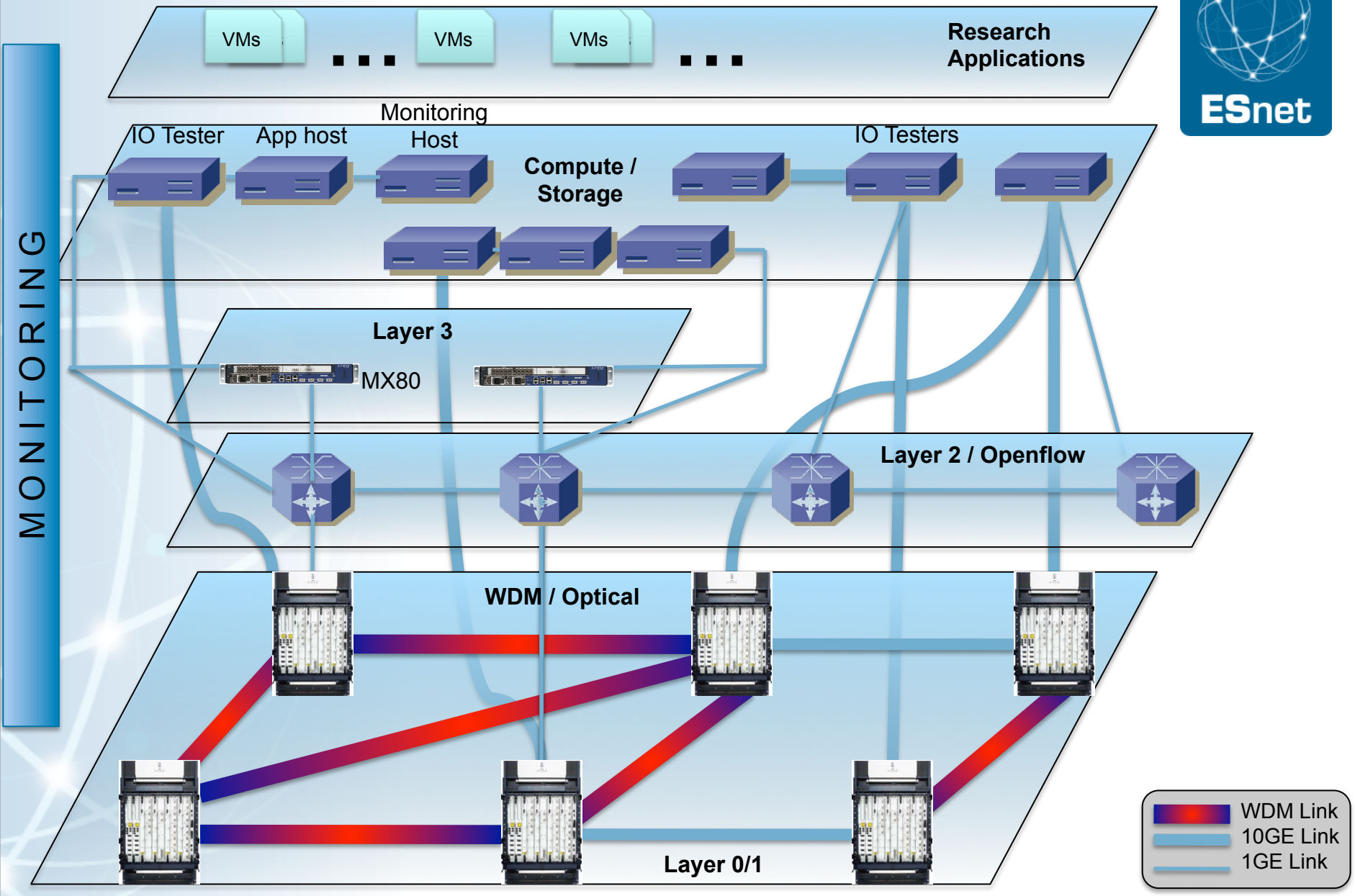
Network Testbed Components

Table Network Testbed consists of:

- DWDM devices (Layer 0-1)
- Layer 2 switches supporting Openflow
- Layer 3 Routers
- Test and measurement hosts
  - Virtual Machine based test environment
  - 4 or 6 x 10G test hosts initially
    - Eventually 40G and 100G from Acadia 100G NIC project

- This configuration will evolve over time

ANI Testbed: A Layered View



Testbed Access

Proposal process to gain access described at:

https://sites.google.com/a/lbl.gov/ani-testbed/

Testbed is available to anyone:

- DOE researchers
- Other government agencies
- Industry

Must submit a short proposal to the testbed review committee

- Committee is made up of members from the R&E community and industry

Plan is to accept roughly five proposals every 6 month review cycle

- Last round of proposals was due April 1, 2011
- Next round of proposals due Oct 1, 2011

Proposals are reviewed by a review committee with DOE, NSF, University, Industry, and International R&E community members

Network Performance Knowledge Base

http://fasterdata.es.net/

Host tuning information:

- http://fasterdata.es.net/fasterdata/host-tuning/

Data transfer tools (including SCP/SFTP issues):

- http://fasterdata.es.net/fasterdata/data-transfer-tools/

Data Transfer Node, including sample hardware config:

- http://fasterdata.es.net/fasterdata/data-transfer-node/

Questions?

Thanks!