# Factoring the Network into The GRID

Monique J.Morrow

Distinguished Consulting Engineer

Cisco

Cisco Team:
Masum Z. Hasan
Nino Vidovic
Wayne Clark
Horst Dumcke
Dragan Milosavljevic
Monique Morrow

# Agenda

1. Factoring the Network Into the GRiD

2. Video Rendering User Case

3. Multi-autonomous Domain Constructs and Challenges and Summary

# Network Factored Grid

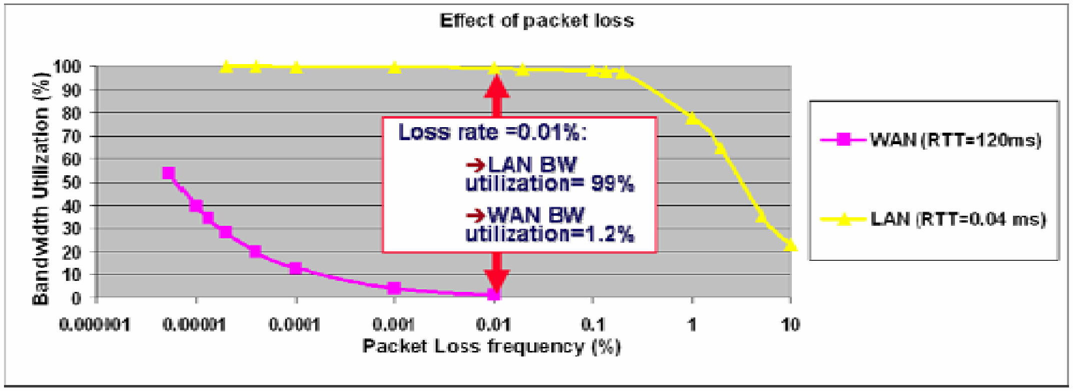
- Network
  - (ISO/OSI) network Layers 1, 2, 3, and 4 (L1/2/3/4)
- Issues we will encounter to factor network into Grid
  1. High speed data transfer
  2. Network abstraction
     A. Network Virtualization
  3. Network resources and services to factor into Grid
     A. Generally abstracted resources and services
  4. Dynamic provisioning of network resources for Grid
  5. Network based performance/QoS monitoring

# Advantages of Factoring Network into Grid

- Factoring network into Grid will facilitate
  - Better use of resource intensive applications
  - Better overall performance (performance and price ratio)
  - Network-aware Grid services
    - Network-aware global Grid job scheduling
  - BW, QoS and other network service provisioning (indirectly)
  - Fine-tuning of network parameters on-demand
- Fine-tuning of network parameters and network resource provisioning on-demand will facilitate HPC and other resource intensive applications to move to Grid

# High speed Data Transfer

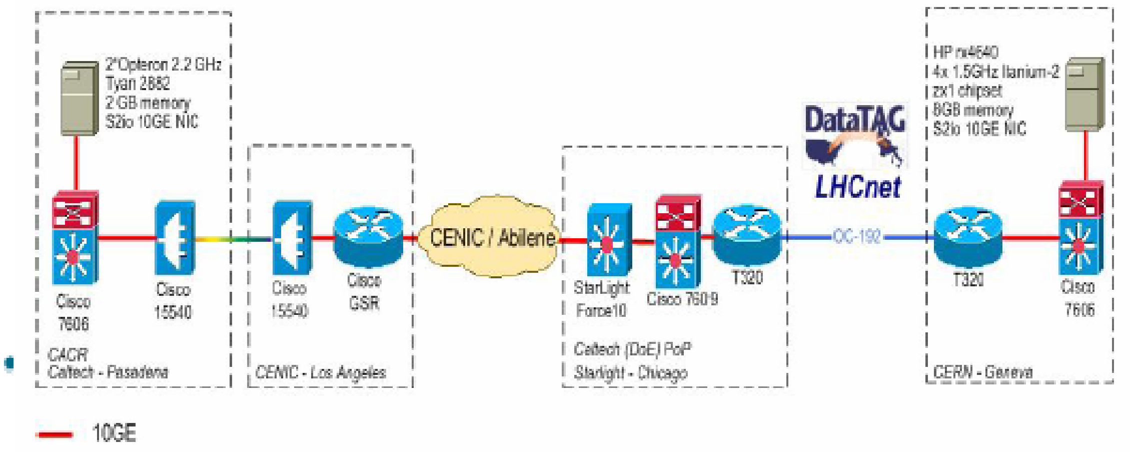
- Many Grid applications require high-speed data transfer over *high bandwidth-delay product (BDP)* links
- TCP or TCP based protocols are most widely used
- No matter what is the speed/capacity of a link, throughput may suffer
- TCP Throughput <= 1.2 * MSS / (rtt * sqrt(packet_loss))
- 180ms RTT (Geneva – LA), packet loss 0.1% (.001) → 1.2*1460*8/.18*sqrt(.001) → 2.5 Mbps throughput
- Frame size (Jumbo Ethernet frame) of 9000 bytes → 15 Mbps



Loss rate =0.01%:
→LAN BW utilization= 99%
→WAN BW utilization=1.2%

# TCP Throughput Tuning

- Set TCP buffer size to bandwidth-delay-product
- High-speed TCP (RFC 3649)
- Parallel TCP stream
- GridFTP
- Jumbo Frame
- TCP Offload Engine (TOE)
- RDMS (Remote Direct Memory Access: bypass kernel for data copy)



- CERN-Geneva to CalTech transfer using Cisco devices
- All intermediate routers/switches on the path supported 9000 byte MTU
- TCP buffer was configured properly
- Memory-to-memory data transfer at 6,5 Gbps with a single TCP stream between CERN Geneva and LA (for CERN LHC data transfer)
- This is about factoring network into the Grid

# TCP Friendliness

- Aggressive TCP tuning, High-speed TCP may potentially starve traffic belonging to standard TCP
- Should be careful in deploying in commercial environments (mixing with standard stack)
- Tools that factor network into Grid can control this

# Network Abstraction – Domain and Layer Separation

- Domain and layer separation is predominant in the network world
  - ISO/OSI or Protocol layer separation
  - UNI (User to Network Interface) between network domains
    - Between client and server networks
    - Between a customer and a Network Service Provider (NSP)
    - Between an IP Data and a Sonet/SDH/Optical transport network
  - Separation of management and control between data and transport networks in NSP Environment
- If Enterprise IT (EIT) or NSP has to offer Grid services on existing infrastructure, similar domain and layer separation is needed in Grid environment

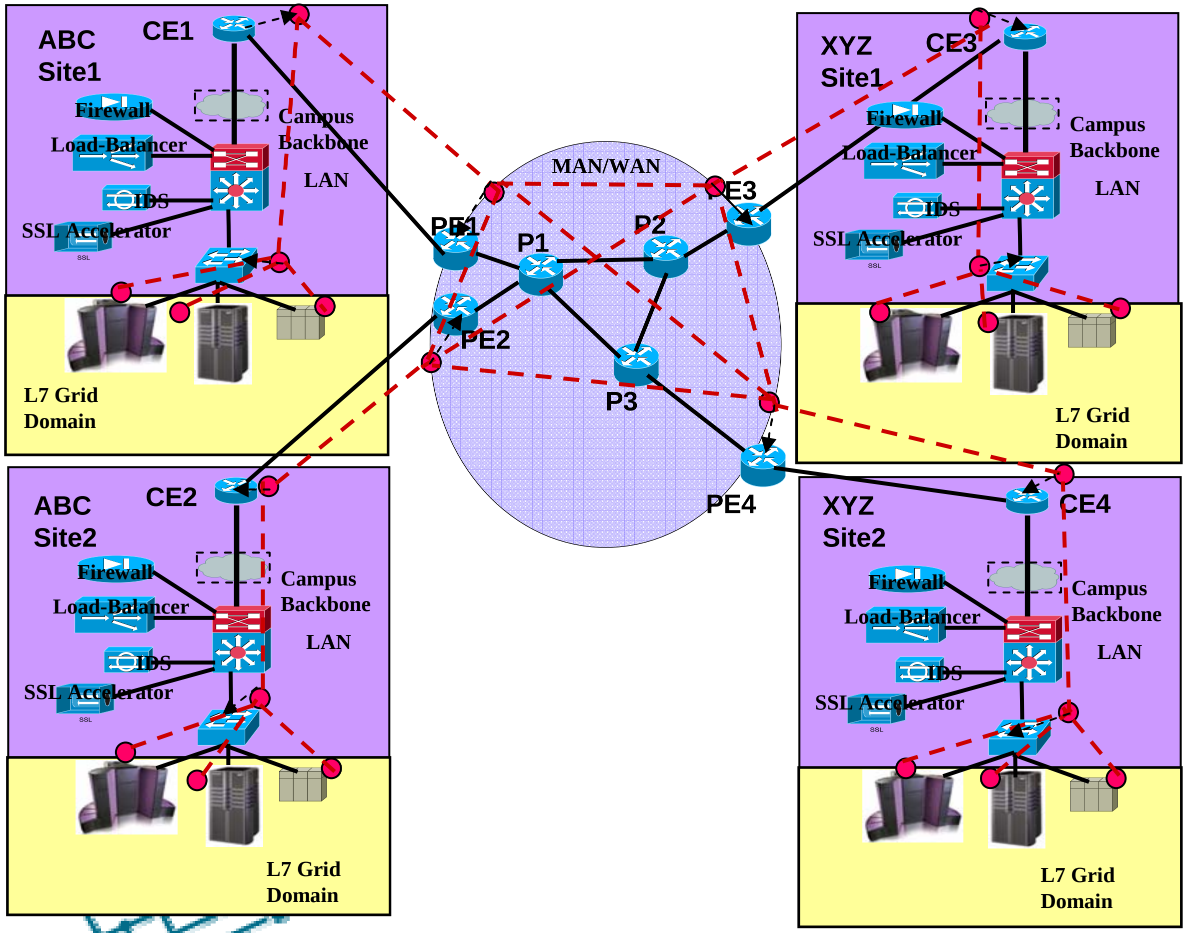
# Network Abstraction – Domain and Layer Separation – What can or cannot be done

- Consider the following questions: Should an L7 Grid system (such as a Grid application or a global Grid job scheduler)
  - Be allowed to configure full capabilities of a router, such as configure routing protocols?
  - Be allowed to get full access to network topology to make scheduling decisions ?
  - Be aware of wide varieties of network technologies (Ethernet, Optical, MPLS, etc.) on which Grid can be built?
  - Be aware of what QoS mechanism being used: 802.1p, IntServ, DiffServ, MPLS?
  - Be aware of various signaling protocols: classical RSVP, SIP, H.323?
- The general answer is no
- An NSP or Enterprise IT will not allow access to router configuration or full topology
- Separation can be ignored in research part of National Research Networks (NRN) or purely research networks, but it can not be, in commercial networks

ITU-T

# Network Abstraction – Domain and Layer Separation

L7 Grid Systems (L7GS): Grid Applications, Grid Clients, Grid Middleware Components



- An L7GS residing in the yellow region

  Can see the abstract topology shown red

  Cannot see the full physical (or logical) topology of NSP/EIT

  Cannot directly configure and provision the EIT or NSP network elements (NE)

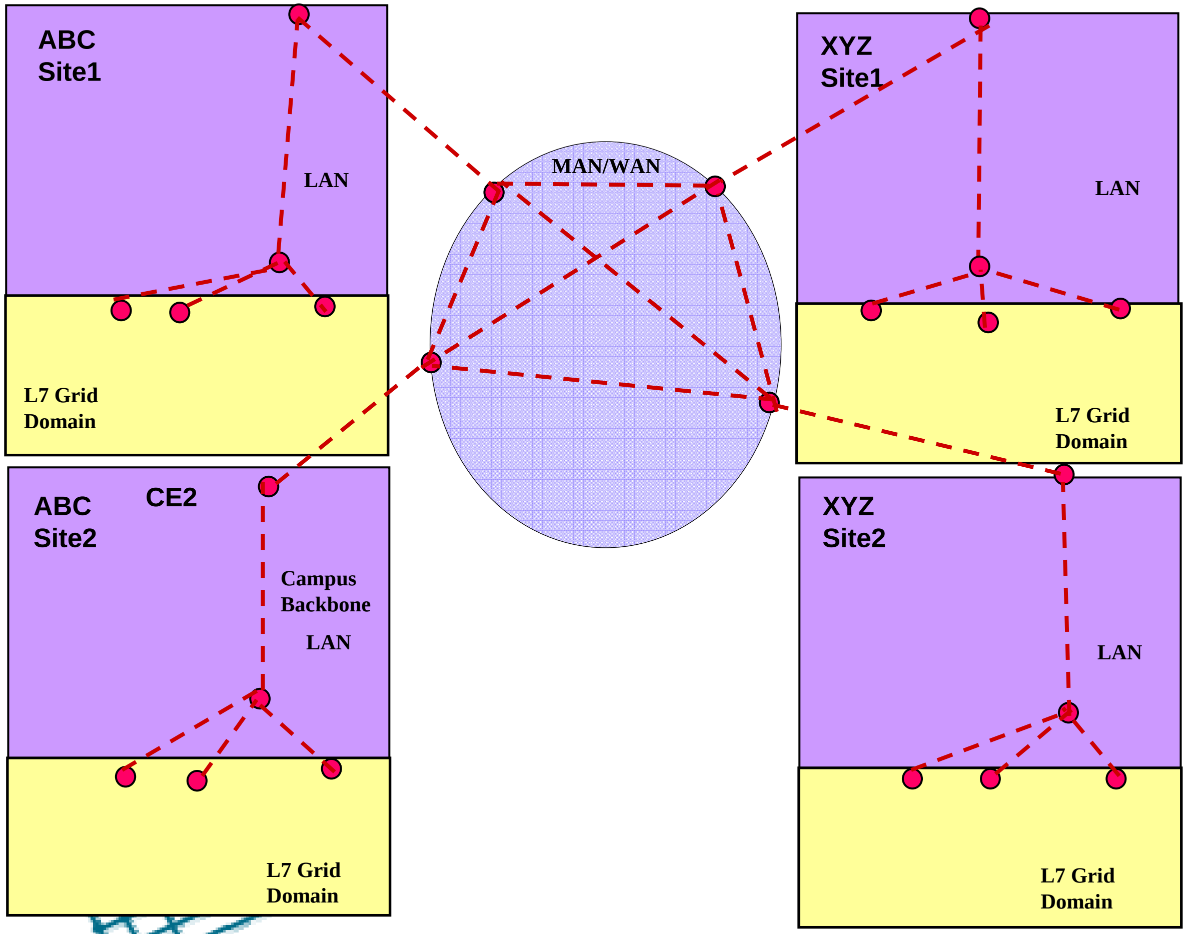
  with the knowledge of the abstract topology, L7GS can perform network-aware Grid functions

OpenGridForum

ITU-T

# Network Abstraction – Domain and Layer Separation

L7 Grid Systems (L7GS): Grid Applications, Grid Clients, Grid Middleware Components



- This abstract topology is what an L7GS may be allowed to see
- Requires modeling of abstract resources
- Function serves to an L7GS
  - Such abstracted topology
  - And relevant network service interfaces
    - Provisioning
    - Monitoring
    - Network parameter tuning

# Network Resources and Services

- Not all resources or interfaces may be factored in
  - Network Elements themselves or their links
  - Protocol resources and their configuration interfaces
- Resources and interfaces that may be factored in are *network service related*
  - Bandwidth
  - QoS
  - VPN
  - Firewall
  - …
- Resources will be abstracted; For example:
  - Exact nature of QoS mechanism, such as DiffServ, IntServ, 802.1p, or MPLS will be hidden from L7GS
  - Exact nature of bandwidth "tunnel", such as MPLS TE LSP tunnel or Sonet/SDH Circuit or DWDM Lambda LightPath, will be hidden from L7GS

# On-demand / Dynamic Network Provisioning

- Network administrators or operators (Service Providers) generally limit or prioritize resource usage
  - Example: Policing, Queuing (based on packet marking)
- Demand for more bandwidth, priority may not be serviced immediately even though resources available, because of *static configuration*
- Demand for new features, such as new QoS priority for a new application cannot be serviced immediately
- Applications, of course, can use dynamic signaling protocols like classical RSVP, but the signaling solution has limitations
  - Wide varieties of Grid applications need to be modified to support signaling
  - Applications may need to support more than one signaling mechanisms in a heterogeneous environment like Grid
  - Signaling, such as classical RSVP may not be supported end-to-end (E2E)
- But applications somehow should be able to "signal" (request) bandwidth, QoS and other network services (such as VPN, firewall related) on-demand

# On-demand / Dynamic Network Provisioning contd ...

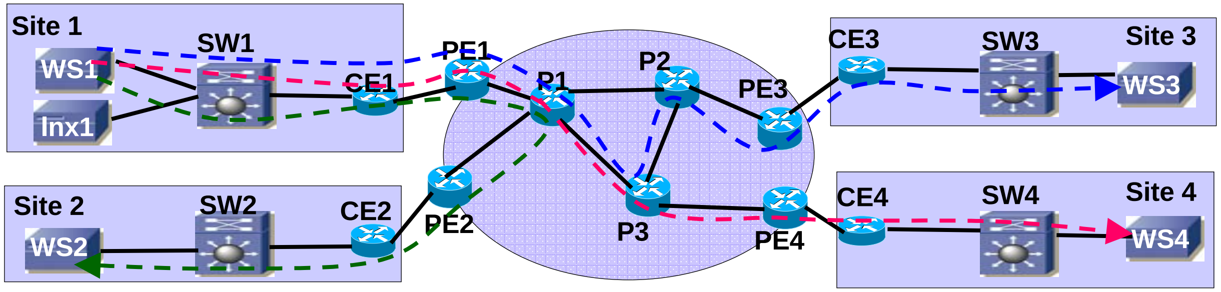
- On-demand or dynamic network provisioning does not necessarily mean
  - Immediate provisioning (rather future resource scheduling)
  - Full provisioning
    - Network resources are provisioned where and when necessary
    - May make use of existing configurations
  - One single signaling
    - Resources may be provisioned E2E using combination of, for example, classical RSVP, RSVP aggregation, CLI, XML, and other interfaces

# Grid Across Wider Network Domain

- In a Grid environment there is strong need to facilitate
  - Resource sharing across wider network domain
    - Not just single Cluster or LAN, but also Enterprise campus network, multi-site Data Center, MAN and WAN
  - Deployment of applications over wider network domain, which Grid users generally shy away from for lack of "control" over wider network
  - Execution of High-Performance Computing (HPC) applications across wider network domain (WND)
    - Latency in WND may far exceed fast interconnects (IPC fabric) used in (single-rack/room) dedicated clusters or supercomputers
    - Many HPC applications with high resource demand can be deployed in WND
  - Facilitate *cost-effective* Grid resource sharing and deployment
    - With "control" over network the domain of resources to be shared is widened
      - For example, resources (servers, clusters) in an Enterprise distributed globally across multiple locations (SJ, RTP, Europe, India) can be pooled into a Grid
      - Not necessarily need supercomputer or high-end clusters

# Use case:
# Video Rendering on Grid

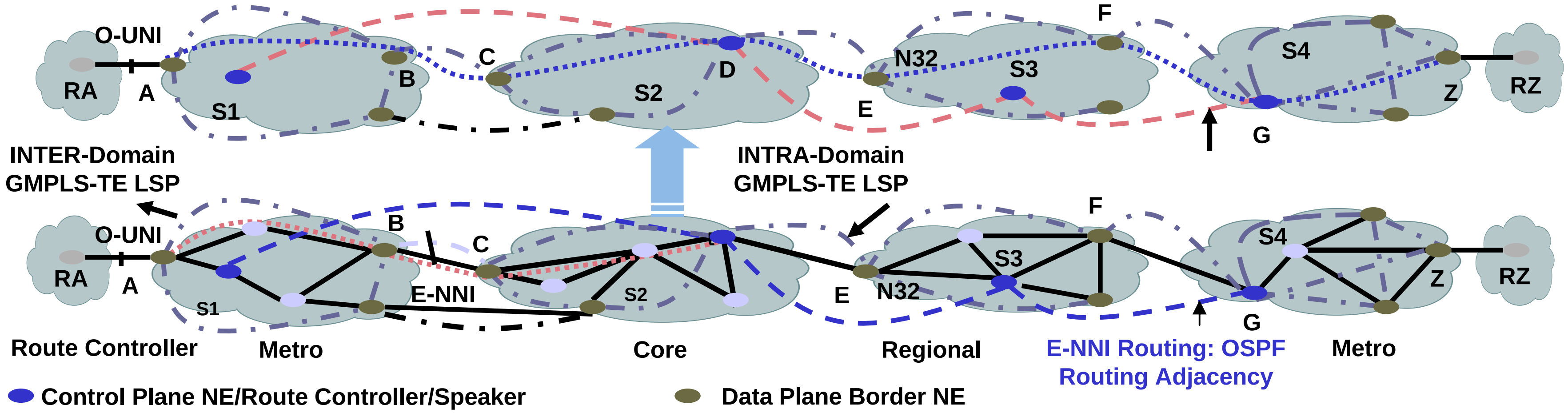
- Video rendering and encoding after post-production video editing
  - This is a resource intensive application from both compute and network perspective:
    - Compute resources
      - 1 min video may take 10 min to render
        - For example, Pixar: Monster Inc. cartoon: 150 days on 2000 node clusters with one GFLOP each
    - Network resources
      - Initial loading of application on large number of processors
      - During execution data may be carried frequently back and forth between processors and between processors and storage
- In the testbed Video data rendered on 4 (grid) machines in parallel
  - Function that creates 3 CosLinks from source machine (WS1) to 3 other machines (WS2-WS4) to transfer data
  - Video data transferred over provisioned CoSLinks, then rendered and merged back
  - One CosLink then created from machine containing rendered video to a client machine and video streamed



# Factoring Network in GriD Multi-Autonomous System Network

- E2E provisioning and QoS is specifically required in Grid environments as multiple organizations may participate in a Grid for resource sharing
- Should be able to provide E2E Multi-AS provisioning support in [G]MPLS networks
  - Inter-AS/CsC TE LSP, MPLS VPN supported in MPLS devices
- Multi-AS QoS is a challenge
- Standard forum support, such as IPSphere, IETF needed

Example of E2E Provisioning in Multiregion/Provider, Multitechnology, Multilayer (IP, Transport/Optical) Networks


O-UNI
RA
A
S1
B
C
S2
D
E
N32
S3
F
S4
Z
RZ
G
INTER-Domain GMPLS-TE LSP
INTRA-Domain GMPLS-TE LSP
O-UNI
RA
A
S1
B
C
E-NNI
S2
E
N32
S3
F
S4
Z
RZ
G
Route Controller
Metro
Core
Regional
E-NNI Routing: OSPF Routing Adjacency
Metro
● Control Plane NE/Route Controller/Speaker
● Data Plane Border NE