draft-ggf-masum-grid-network-services-0
Category: Draft
GHPN

Masum Z. Hasan, Wayne Clark, Monique Morrow
Cisco Systems, Inc.

May 2004

Network Service Interfaces to Grid

Draft

Status of This Memo

This memo provides information to the Grid community. It currently does not define any standards or technical recommendations. Distribution is unlimited.

Copyright Notice

Abstract

This document provides an overview of potential network service interfaces to Grid. Further work or a recommendation may be pursued from the concepts described in this document.

TABLE OF CONTENTS

1 Introduction

The term Grid is widely used for distributed, parallel, and networked processing of (ISO/OSI) Layer 7 (L7) applications and services. A Grid network is an *overlay network* on top of underlying lower layer network consisting of network, data-link and physical (L3/2/1) layers. The functioning, configuration, and behavior of L3/2/1 networks are assumed to be independent of the overlay L7 Grid networks. That is the underlying L3/2/1 network remains as a cloud to the Grid. The Grid network elements (NE) are L7 devices, such as servers, workstations, supercomputers, and storage devices. The overlay L7 Grid network can be managed by a Grid middleware, such as *Globus*. In what follows, we use the term Grid to include L7 overlay Grid network, middleware, application, and L7 NEs. Even if the L3/2/1 network remains as a cloud to the Grid, it is desirable that Grid makes use of wide varieties of services (referred to as *network services*; contrast this with *Grid services*, which may or may not make use of network services directly) provided by the L3/2/1 network. The justifications for interfacing network services to the Grid are manifold, a non-exhaustive list of which is as follows:

- With interfaces to network services a Grid middleware or application will be able to perform network-aware Grid functions (scheduling, storage management, etc.).
- New types of Grids can be built. For example, a secure Grid where each Grid site is an MPLS VPN site.

In this document we discuss the need for network service interfaces to for Grid. We provide an overview of a number of potential network services and a few use cases of how Grid could make use of those services. We do not define the exact interfaces or the system/framework that can provide (implement) the interfaces.

2 Grid Resources and Functions

An overview of Grid resources and resource management functions is provided in this Section.

A Grid is an overlay network of L7 NEs on which Grid applications and services are run. A Grid may be managed by a Grid resource management system or middleware, such as Globus. Following is a list of resources that may be managed by a Grid middleware:

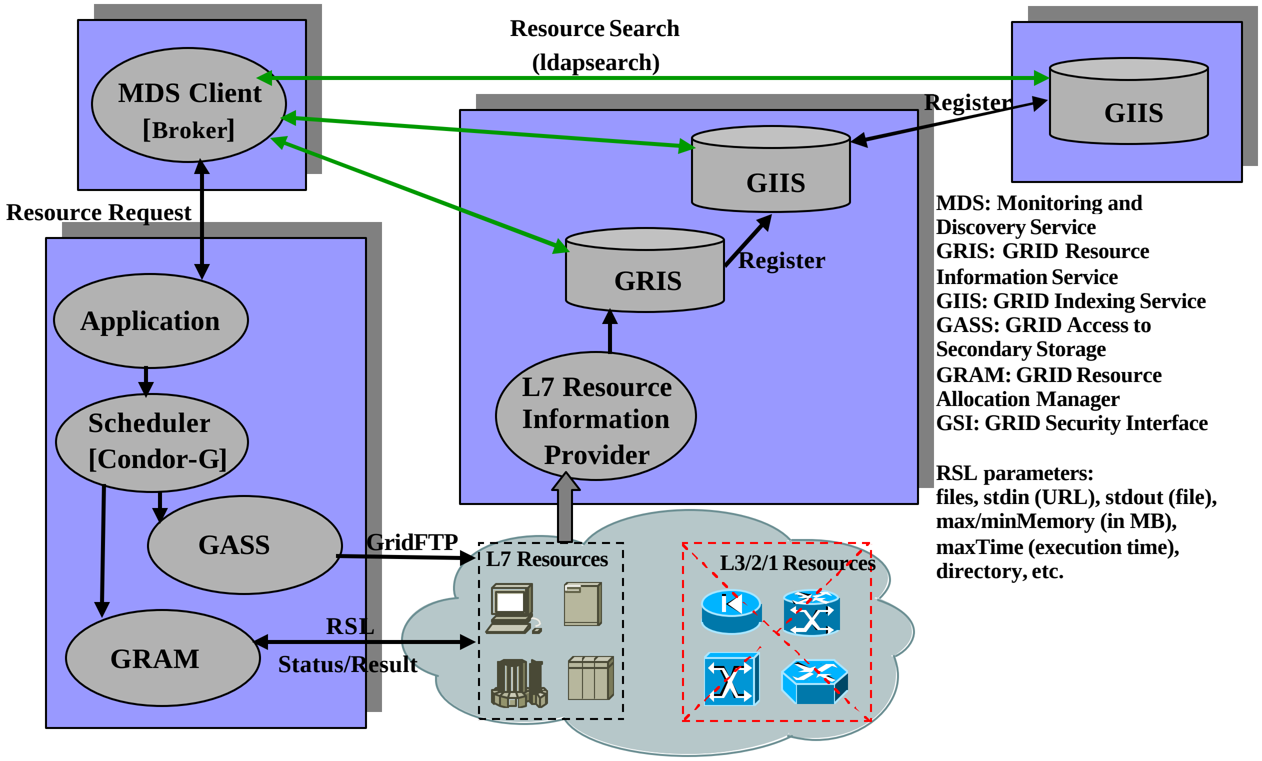
- L7 NEs: Workstations, Servers, Supercomputers
- Processes/Applications/Tasks (computational units)
- CPU, Memory, Files
- Storage
- Data-sets, Databases
- L7 services (web, database, e-commerce, gaming, etc.).

A non-exhaustive list of Grid resource management functions is as follows (functions can be closely interrelated):

- Resource allocation. For example, allocation of CPU, memory and data sets to computational units.

- Scheduling and distribution. While the resource allocation may involve a single NE, this function may involve multiple NEs, where resources (for example, tasks) are distributed and scheduled in an optimized way.
- Data, database, and storage management.
- Monitoring.
- Resource Discovery.
- Resource information searching.
- Security management.

Architecture of a Grid middleware (Globus) is shown in Figure 1.



MDS: Monitoring and Discovery Service
GRIS: GRID Resource Information Service
GIIS: GRID Indexing Service
GASS: GRID Access to Secondary Storage
GRAM: GRID Resource Allocation Manager
GSI: GRID Security Interface

RSL parameters:
files, stdin (URL), stdout (file), max/minMemory (in MB), maxTime (execution time), directory, etc.

Figure 1 Grid Middleware (Globus) Architecture

3 Overlay Grid Network

An overlay Grid network may be built on many different types of L3/2/1 networks. In this Section we provide a few examples of L3/2/1 networks over which Grid can be built. While the overlay Grid operates independently of the underlying network, as we have mentioned earlier, a Grid can benefit from interfacing with the services provided by the underlying network.

Figure 2 shows an overlay Grid network. The figure shows distributed Grid sites and copies of Grid middleware or its components running on those sites. Underneath the cloud there may be wide varieties of network types, a few examples of which are provided below.

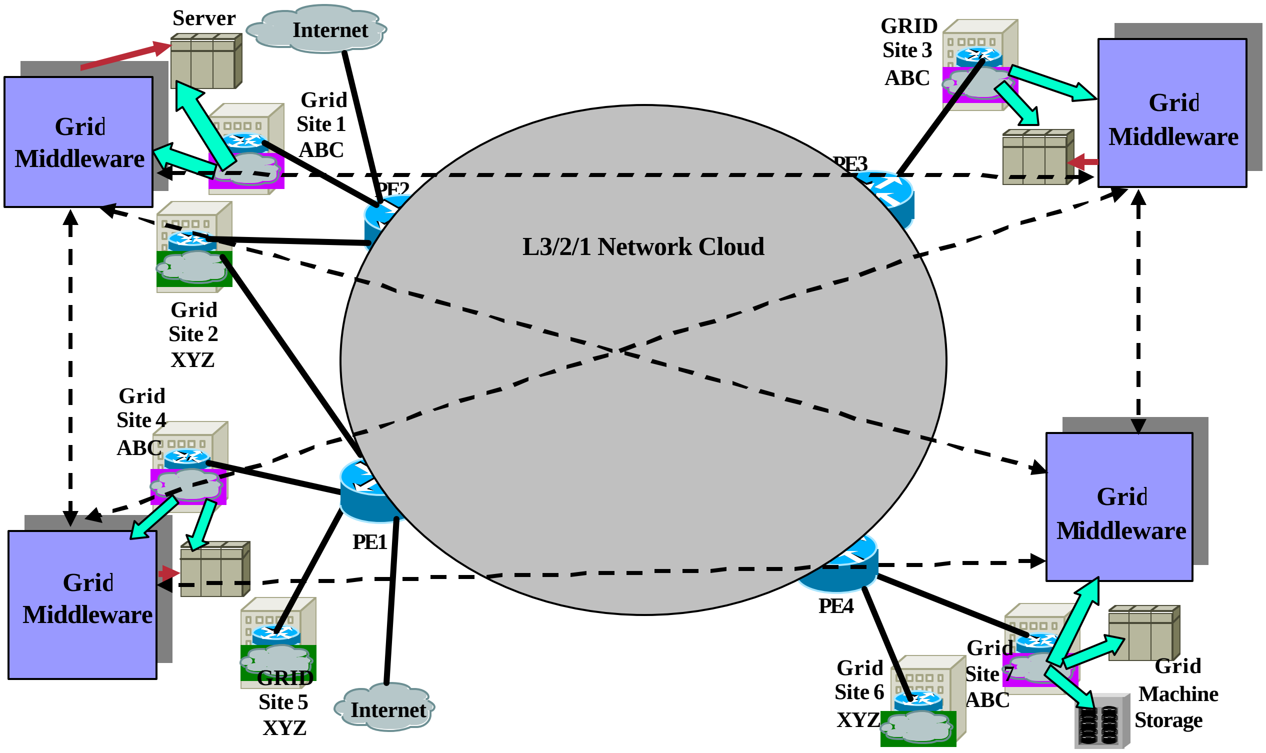


Figure 2: Overlay Grid Network

Figure 3 shows a Grid (or Grid MPLS VPN) over an MPLS network, which may be a private shared network either owned and operated by a service provider, or owned and operated by an enterprise owning the Grid. In the both the cases, the network can be a cloud with respect to the Grid.

Figure 4 shows a Grid (or Grid L2VPN) over an AToM (Any Transport over MPLS; for example, Ethernet over MPLS) network.

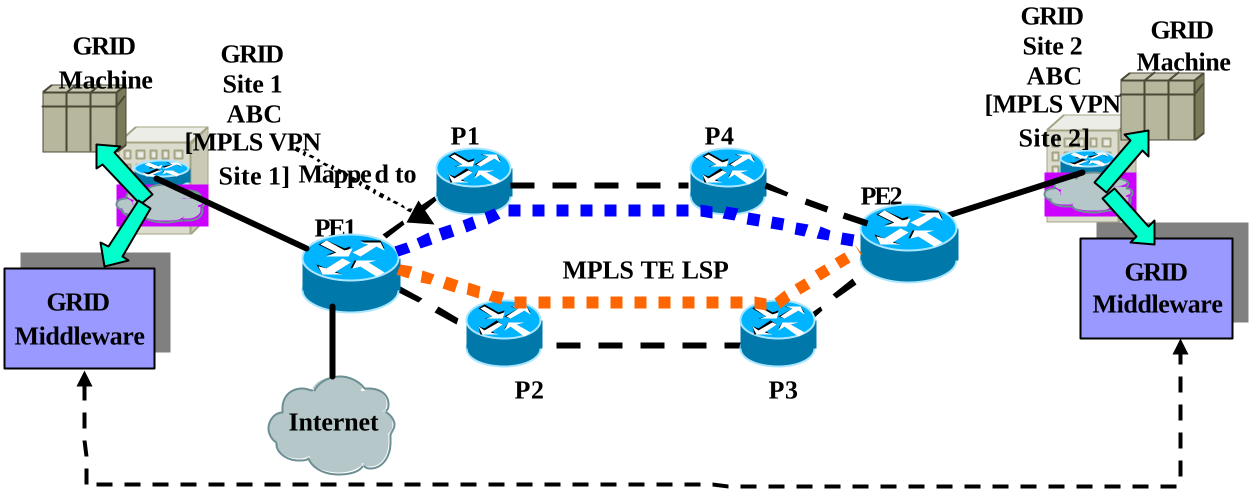


Figure 3: Grid over MPLS Network (Grid MPLS VPN)

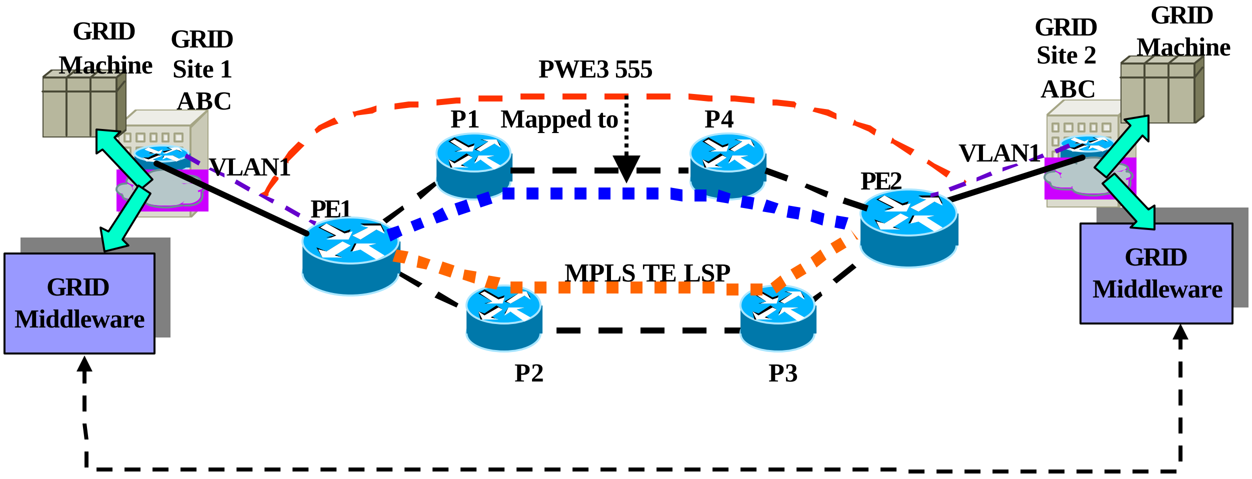


Figure 4: Grid over AToM (Any Transport over MPLS) (Grid L2VPN)

Figure 5 shows a Grid over an optical/transport/GMPLS/ASON[1] network. The upper portion shows a GMPLS network that is owned by the Grid network enterprise. Hence it is possible to provision site-to-site GMPLS TE LSP. The lower portion of the figure shows a network architecture where the interfaces between Grid network sites and the underlying network is O-UNI (OIF Optical-UNI). The underlying network may or may not be owned by the Grid enterprise.

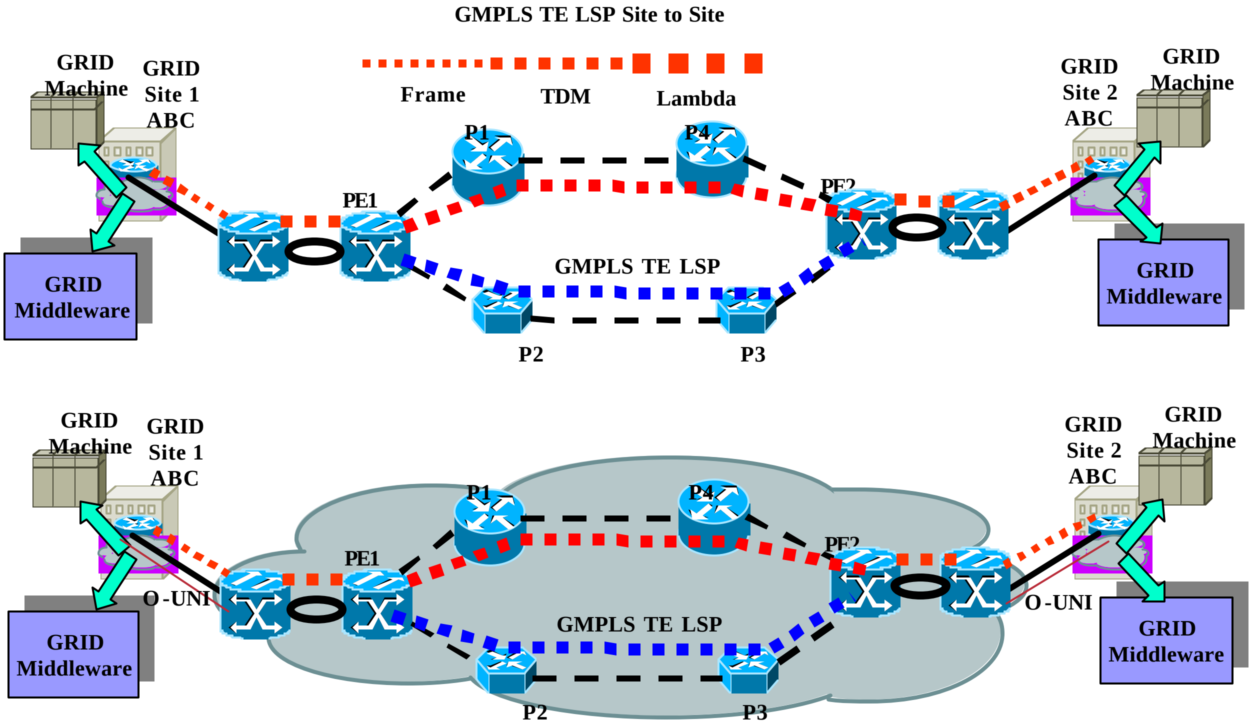


Figure 5: Grid over Optical/Transport/GMPLS/ASON Network

[1] ASON: ITU-T Automatic Switched Optical Network

Figure 6 shows possible details of distributed campus networks where Grid resources are housed. As shown in the figure a Grid may make use various network services provided in the campus (enterprise). For example, the figure shows that Grid traffic may be passed through a path with any combination of firewall, load-balancing, IDS, or SSL accelerator services. The campus Grid network (as an Autonomous System) may have multi-home (E-NNI) BGP connections with multiple WAN or Internet Service Providers. The campus network may be enabled with optimized BGP alternate path selection services (such as Cisco OER: Optimized Exit Routing). The campus Grid can interface with all these different types of services.

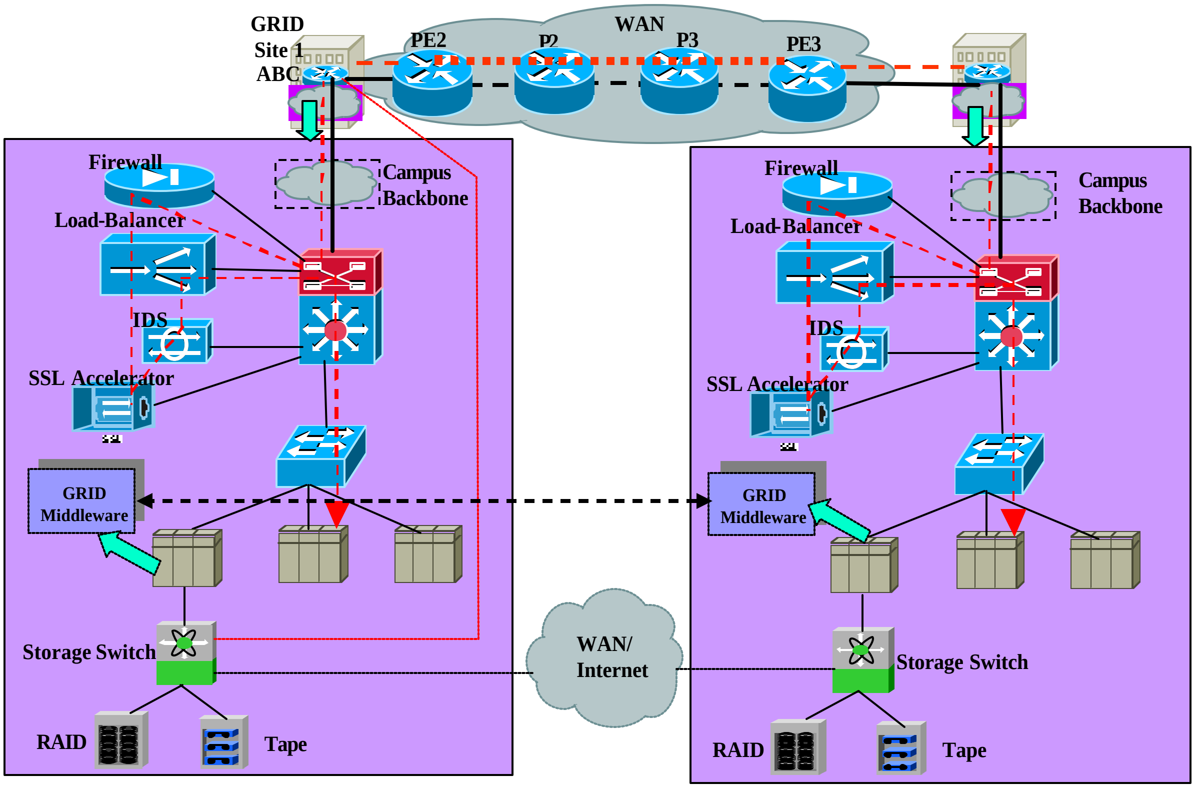


Figure 6: Grid Network over Campus Network and WAN

Figure 7 shows a Grid over public Internet (the cloud is Internet).

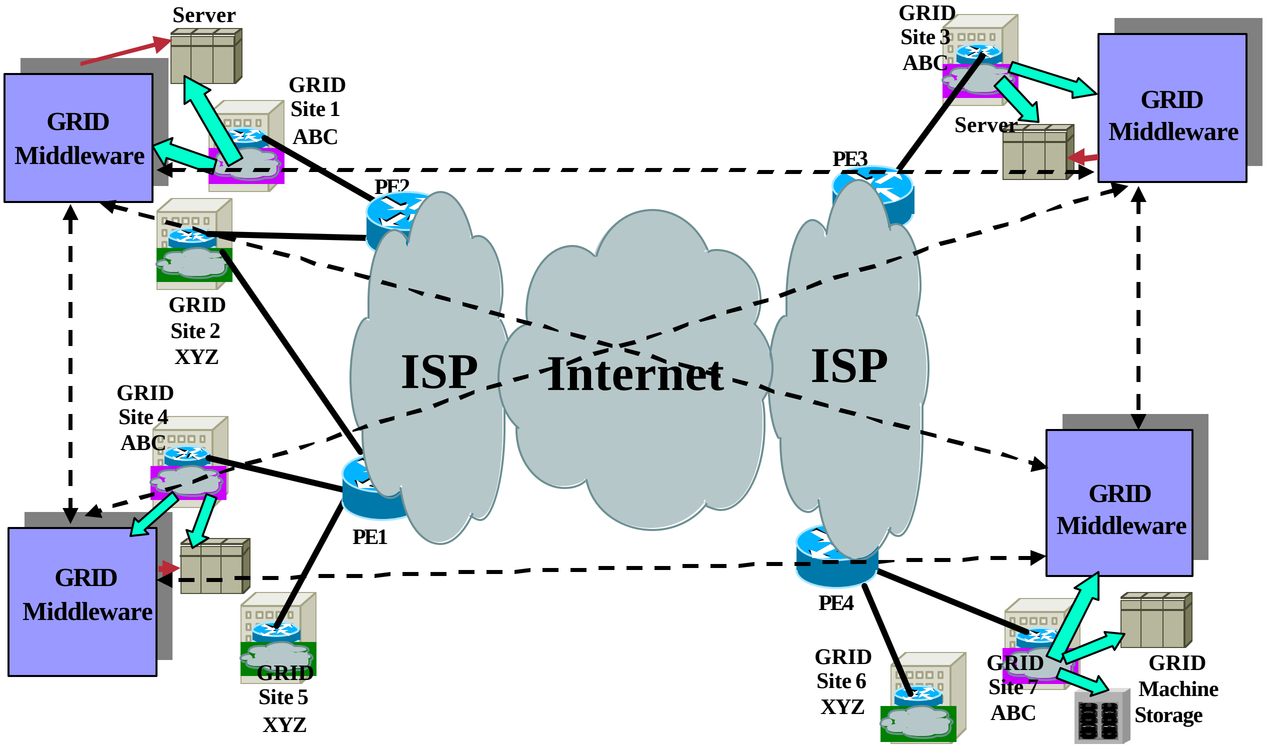


Figure 7: Grid over Public Internet

4 Network Services

Given the types of underlying networks, there are wide varieties of network services with which a Grid can interface. A non-exhaustive list is as follows:

- Bandwidth related services
- CoS/QoS related services
- L3/2/1 VPN/Security
  - L3 MPLS VPN
  - L3 IPSEC VPN
  - L2 VPN (Any Transport over MPLS or L2TPv3 based)
- G/MPLS Traffic Engineering (TE) based services
- Optical connection services
- Firewall services
- IDS services
- SSL Acceleration services
- Optimized BGP alternate path selection services (for multi-homed connections).

5 Use Cases

In this section we provide a few use cases of network services with which a Grid can interface.

5.1 Network aware Task Scheduling

Consider Figure 8, which shows an overlay Grid network belonging to a particular enterprise (for example, ABC; figure also shows an overlay Grid network for enterprise XYZ).

The Grid network is connected over a WAN. The WAN service may be provided over a *private shared network* (as opposed to *public shared network* like Internet as shown in Figure 7) owned by a service provider (SP) or carrier. The network can also be owned by a large enterprise (for example, ABC) or a consortium (a Grid consortium, for example).

Consider task scheduling and distribution (a Grid middleware function; see Figure 1). The task scheduler based on computing resource information (CPU cycle, memory, CPU speed, etc.) schedules tasks on Grid NEs. But consider the following:

- There is intensive communication between certain tasks.
- Some level of bandwidth and QoS is required on the communication paths.

With the availability of interfaces to L3/2/1 network services, the task scheduler could perform the following:

- Make a decision on task distribution (before actually distributing them), query the condition on the communication paths between Grid sites, and adjust distribution based on query result.
- Redistribute based on network condition after the tasks have been distributed.

- Request bandwidth and QoS constrained paths (pipes/tunnels), if supported, between relevant sites. For example, as shown in Figure 8, if the underlying (L3/2/1) network is [G]MPLS, then the Grid may request provisioning of MPLS TE LSP tunnels between relevant sites.

5.2 Network aware Grid data movement

Consider a Grid data mining application. For distributed or parallel processing of data, the data-mining agents are scheduled on multiple Grid computers distributed over the network. It may be desirable that the segment of data to be mined is moved as close to the mining agent computers as possible. This movement may not be necessary, if network condition or available bandwidth between NEs is adequate. A dedicated bandwidth may be reserved between a mining agent and its associated data repository (as shown in Figure 8). Based on network condition, Grid middleware (or a middleware component, such as GASS) can schedule data movement and, if necessary, request QoS or bandwidth constrained tunnels between relevant NEs (similar to the scheduler above).

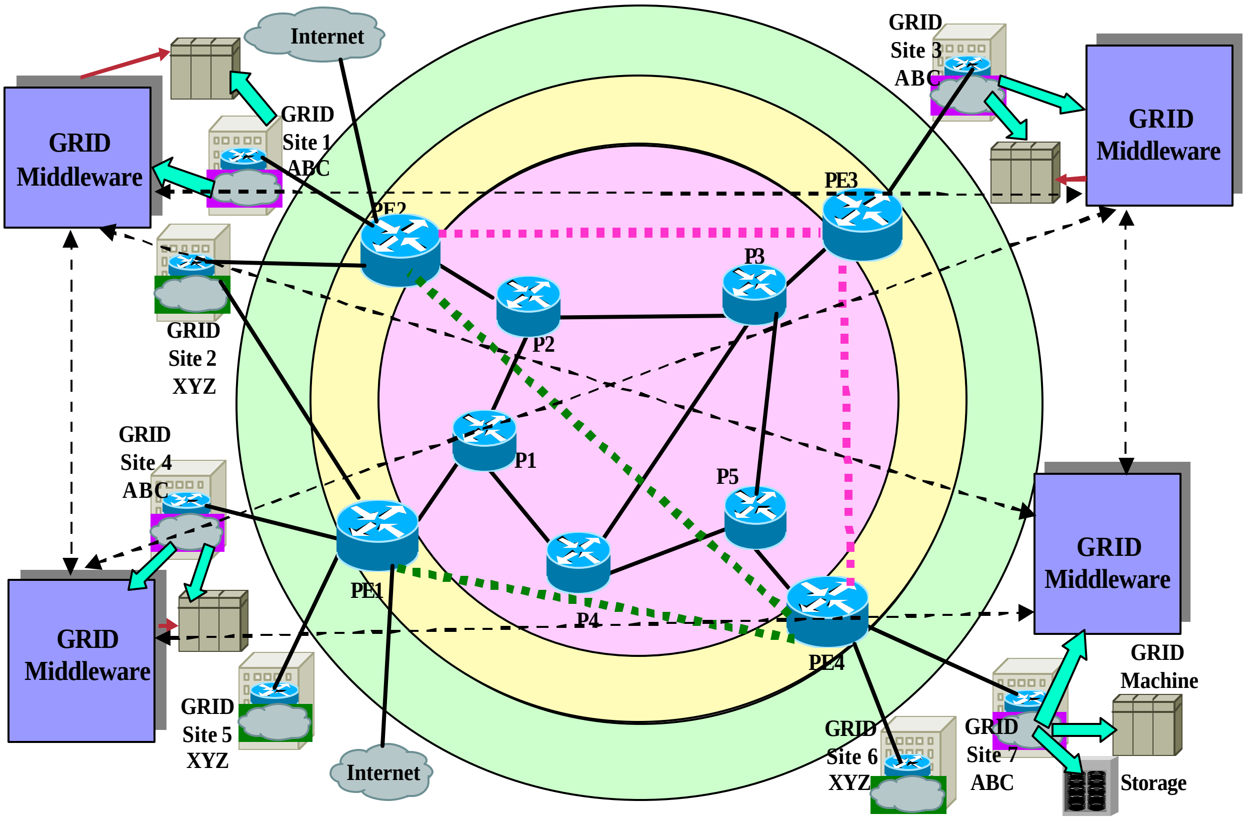


Figure 8: Network aware task scheduling and Data Movement

5.3 Grid VPN

Consider the example network in Figure 8. A portion (the circled ones; let us call the portion *core*) of this network may be owned and operated by a service provider or a Grid consortium. As shown in the figure there are multiple Grid "customers" (ABC and XYZ) on the core network. Each customer may want to build its own secure Grid VPN. For example, if the core is an MPLS network and MPLS VPN is supported on the provider edge (PE) devices, then the customers can build MPLS Grid VPN.

6 Network Service Interfaces

It is evident from the discussion above that there is a need for defining L3/2/1 network service interfaces with which a Grid can interface. There are four issues relevant to these interfaces that need to be considered:

1. Nature of the interfaces, that is, the requirements that these interfaces should adhere to. *It is expected that the interfaces provided are high-level, that hide details of the network and service configuration*. For example, a GRID could request the following:

    a. Create_path (source, destination, bandwidth, QoS), where the source and destination can be references to IP address, site, VPN, GRID service, etc., and QoS can be abstracted into Platinum, Gold, etc., that hides the detail of how QoS is provided. For example, Platinum CoS, depending on the QoS capabilities of the underlying network, can be transparently mapped to any of the following:
        i. DiffServ EF
        ii. Relevant IntServ QoS
        iii. Priority queue + DS-TE tunnel + FRR protection
        iv. Firewall + SSL Acceleration + IDS + Redundancy.

    b. Join_in_VPN (source, existing_VPN): a Grid site joins a VPN. The detail of VPN related configuration is hidden from Grid. For example, if the service is MPLS VPN, then configuration details, such as VRF (Virtual Routing Forwarding) Route Target or Distinguisher are hidden from the Grid using the interface.

    Following is a non-exhaustive list of requirements from which the interfaces should be defined:

    1. Details of the underlying networks must be hidden from the Grid.
    2. Network topology or model must be hidden from the Grid.
    3. Details of network configuration must be hidden from the Grid.
    4. Visibility of network status by a Grid must be limited.
    5. Types of underlying network can be hidden from the Grid. A non-exhaustive list of network types that may need to be considered is as follows:
        a. IP
        b. MPLS

        i. MPLS TE
        ii. MPLS DS-TE (DiffServ-aware TE)
        iii. GMPLS
        iv. AToM (Any Transport over MPLS)
    c. ATM
    d. TDM (Sonet/SDH)
    e. DWDM
    f. Ethernet
    g. LAN, WAN
    h. Private Shared Network
    i. Public Shared network (Internet).
6. Interfaces should be as simple as possible.
7. An interface should correspond to multiple operations in the network.
8. A single interface should be able to cover as many types of network as possible.

2. Major types of network service interfaces:
    a. Configuration related. The Create_path and Join_in_VPN are examples of such interfaces.
    b. Monitoring interfaces. Once the L3/2/1 network service related resources are configured or provisioned, they can be monitored. For example, the configured path can be monitored. A Grid middleware may monitor L7 Grid resources (mentioned in Section 2). For example, it may perform the following:
        i. TCP throughput between Grid NEs.
        ii. Disk-to-disk file copy throughput.
        iii. Grid NE to Grid NE delay, jitter, RTT.

    But since the Grid middleware may not be able to have access to L3/2/1 resources, *end-to-end, on-demand, or real-time* performance improvements may not possible (for example, by requesting an alternate bandwidth and QoS constrained path). Note that, end-to-end protocols, such TCP, RTP, RTCP, etc., monitor network conditions transparently, such as RTT, on behalf of applications using them, and adapts accordingly (for example, TCP timeout (re)calculation). But the connection is not rerouted to a new path based on monitoring results, a feature that can be provided to Grid via network interfaces.

3. The system or framework for providing the interfaces. The interfaces can be provided out of a service, network or element management system, or directly out of a network element. Let us call the system providing (implementing) the interfaces, L3/2/1 Resource Management System (L3/2/1 RMS). L3/2/1 RMS should be cleanly separated from the Grid (Grid middleware, whose function is to manage L7 Grid resources) as shown in Figure 9 and Figure 10. *Note that, the network service interfaces for Grid will have a higher level of abstraction (hiding details) than what is provided by a traditional Service, Network, or Element Management System*.

4. Discovery of network services and capabilities.

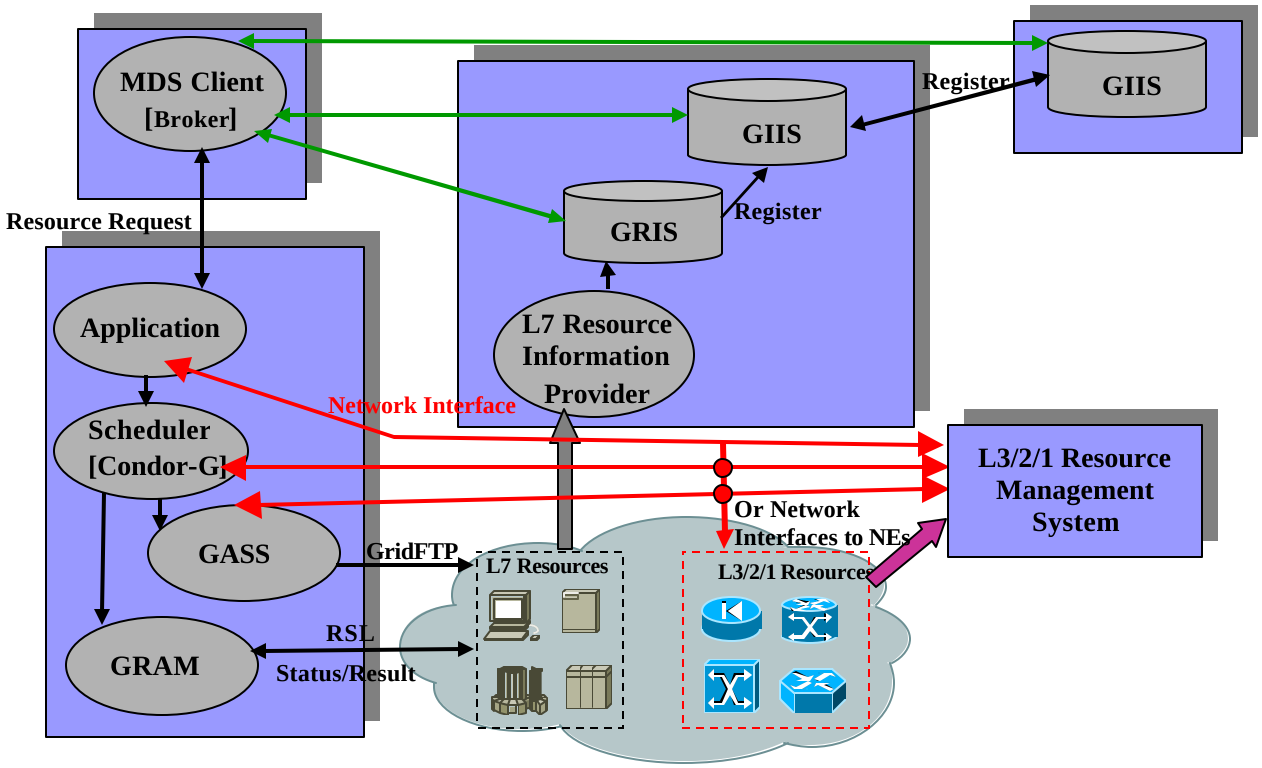


Figure 9 Network Service Interactions from Grid Middleware or Applications

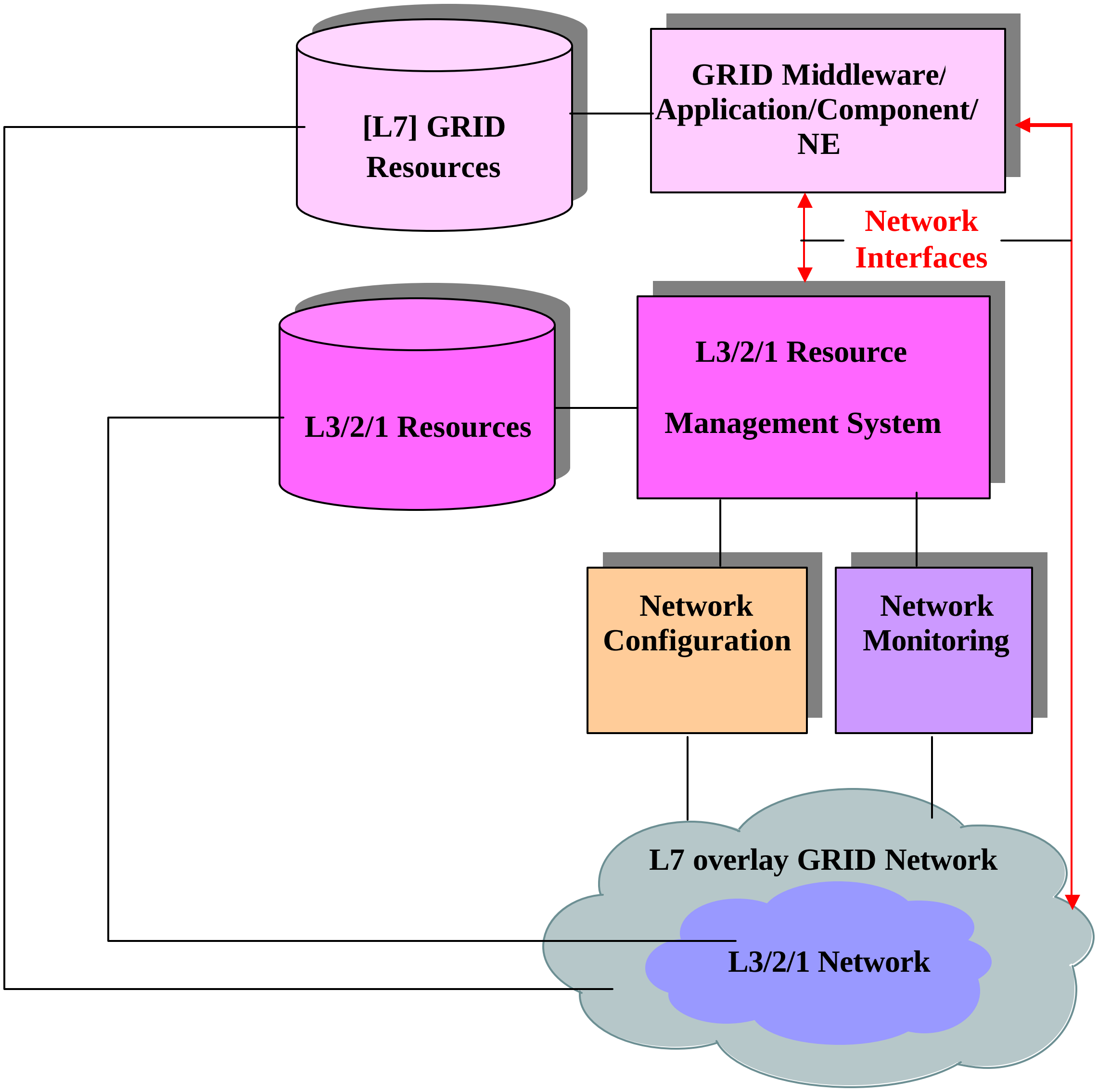


Figure 10 High-level Network Service Interface based Grid and L3/2/1 Resource Management Architecture

7 Discussion

7.1 Need for Managed Bandwidth and QoS

Why do we need to manage bandwidth and QoS?

In a public or private shared network we may need to manage bandwidth and QoS because of the following reasons:

- Networks are built around *onion model*. Consider Figure 8, as we move from the core to the edge, there is less and less available bandwidth. Hence, bandwidth and QoS on the edges, at least, need to be managed.
- Consider a core link.
    - Traffic may converge on it from all directions.
    - The link may be shared by many types of traffic.
    - There may be limited number of flows/sessions/applications sustained on the link at any point in time. As an example (just to make a point), assume that all the links of core are OC-12 (622Mbps) and 40% of the capacity of all links has been allocated for highest priority traffic (DifServ EF for VoIP traffic, for example). Assuming that each voice call takes about 30kbps (rate may vary from 20kbps to 64kbps depending on compression scheme used), in total about 8000 (600000000*.4/30000) simultaneous calls from all the edges can be admitted to a core link.

It can be argued whether bandwidth and QoS need to be managed in a Grid running on a dedicated private Giga/Tera/Petabit research network (such as TeraGRID/DTF, iVDGL). We may still need to manage bandwidth and QoS on such networks for the following reasons:

- The scale: the bandwidth requirements for applications that are run on these types of networks can be proportional to the bandwidth of the underlying network.
- As more bandwidth is available, more bandwidth consuming applications are introduced (this is true for any network).
- In a Grid community or consortium sharing a common infrastructure, each member will have its own demand into the network. Hence the bandwidth and QoS need to be managed.

7.2 End to End Signaling

Bandwidth and QoS resources can be provisioned in a number of ways:

- Via end-to-end signaling (between Grid middleware/applications), such as using (classical) flow-based RSVP[2].

- Via edge-to-edge signaling, such as RSVP-TE in MPLS-TE networks.
- Other means, such as configuration.
- A combination of all of the above.

Because of scale and other reasons flow-based RSVP may not be supported by all the network elements end-to-end. It is possible to support RSVP aggregation (per RFC 3175) at the edge (for example, from a Grid NE to a CE or PE router or switch), where the RSVP flows are aggregated on the edge, carried over the core, and de-multiplexed at the egress CE or PE.

7.3 Grid over a Public Shared Network

While a Grid middleware or application may be able to have access to L3/2/1 resources via the network service interfaces in a private shared and single-AS (autonomous system) network, it may not be able to have easy access to a network that is public shared and multi-AS (such as the Internet). Partial control, only at the edges of the network, may be possible. For example, if a Grid site is multi-homed (to single or multiple ISPs) and Cisco OER (BGP Optimized Exit Routing along multi-homed connections) like capability exists, then a Grid can request alternate exit path via network service interfaces.

8 Summary

It is expected that network service interfaces for Grid are defined following the guidelines set forth in Section 6.

[2] Note RSVP-TE is used for signaling MPLS TE tunnels. RSVP-TE is an aggregate version of classical RSVP.

Security Considerations

None

Author Information

Masum Z. Hasan
Wayne Clark
Monique Morrow
(masum, wclark, mmorrow)@cisco.com
Cisco Systems, Inc.

Glossary

Intellectual Property Statement

Full Copyright Notice

References