Grid High Performance Networking Research Group	Tiziana Ferrari INFN CNAF
GRID WORKING DRAFT	Gigi Karmous-Edwards MCNC Institute
Category: Informational Track	Mark J. Leese Daresbury Laboratory
http://forge.gridforum.org/projects/ghpn-wg/	Paul Mealor University College London
	Inder Monga Nortel Networks Labs
	Volker Sander Forschungszentrum Jülich

(complete list of authors under definition)

Status of this Memo

This memo provides information to the Grid community in the area of high performance networking. It does not define any standards or technical recommendations. Distribution is unlimited.

Comments: Comments should be sent to the GHPN mailing list (ghpn-wg@gridforum.org).

GRID NETWORK SERVICES USE CASES 1

1 INTRODUCTION 1

2 PATH-ORIENTED USE CASES 2

2.1 VISUALIZATION SESSION 2
2.2 REMOTE PARALLELIZED VISUALIZATION 5
2.3 HIGH THROUGHPUT FILE TRANSPORT WITH A DEADLINE 7
2.4 L2 VIRTUAL CONNECTIVITY 10
2.5 QUALITY OF SERVICE PATH FOR GRID APPLICATIONS AND MIDDLEWARE 14

3 KNOWLEDGE-BASED USE CASES 18

3.1 SERVICE OPTIMIZATION [MARK, PAUL, TIZIANA] 18
3.2 ADMINISTRATIVE SETUP OF SCHEDULES OF MEASUREMENTS 21

4 SECURITY CONSIDERATIONS 24

5 AUTHORS INFORMATION 24

6 INTELLECTUAL PROPERTY STATEMENT 24

7 FULL COPYRIGHT NOTICE 24

Grid Network Services Use Cases

1 Introduction

Network services are specialized in the handling of network-related or network-resident resources. A network service is further labeled as a *Grid network service* whenever the service has roles and/or interfaces that are deemed to be specific to a grid infrastructure.

This document contains a list of Grid network service use cases. We expect to expand the document with new use cases, if needed. This is a companion to the Draft-ggf-ghpn-netservices document, which is mainly to provide a set of functional requirements.

The purpose of this information document is: a) to provide a high-level but formal description of some well-understood Grid network services use cases; b) to facilitate the identification of network services critical to the Grid middleware and user applications; c) to help with the identification of the various relationships between Grid network services.

So far, we have identified the following two main use case areas: *Path-oriented* and *Knowledge-based*, as indicated in the following list. The former area includes use cases with special requirements in terms of traffic forwarding, while the latter includes use cases based on information about status and properties of the network.

1. **AREA 1: Path-oriented use cases**
   - Visualization session
   - Remote Parallelized Visualization
   - High Throughput File Transport with a Deadline
   - Quality of Service path for Grid applications and middleware
   - Layer 2 virtual connectivity

2. **AREA 2: Knowledge-based use cases**
   - Service optimization
   - Administrative setup of schedules of measurements

The list of use cases currently included in this document is not exhaustive and we expect to extend it with further contributions from the Grid community.

2 Path-oriented use cases

2.1 Visualization session

2.1.1 Use case summary

Visualization is one of the key methods used to represent data (raw or processed) and is used extensively by almost all fields of specialization for instance e-sciences, medicine, engineering and digital art. A visualization session may either use data-sets available either locally or remotely. Collaborative virtual-reality, distributed CAD, tele-immersion, distributed simulation analysis and haptic collaborations are examples of applications requiring a significant amount of Grid resources (network resources included). In this use-case we focus on requirements of compute and data-intensive visualization sessions.

2.1.2 Customers

Grid Resource Brokers catering to applications requiring visualization like collaborative virtual-reality, distributed CAD, tele-immersion, distributed simulation analysis and haptic collaborations.

2.1.3 Scenarios

An application requests a visualization session to be created between geographically distributed data sources and end users with visualization devices. There could be multiple end-users at geographically disparate locations looking at the same image or a different slice of the same image simultaneously. The application request may also include an interactive and/or collaborative component, for example a user can interactively choose to modify the image by choosing a different image-processing algorithm, zoom in-out or change the viewing angle, on the ongoing visualization session. The capabilities of the visualization devices at each location might be different in terms of display capabilities (resolution, size) and interactive capabilities (ability to modify, zoom etc).

A generic visualization session can be generically broken into the following components:

1. *Data*: The data for a visualization session may be accessed from a storage device, streamed from data access device (s) like a sensor, modality, microscope etc., or streamed from a computational algorithm.
2. *Computation*: The generated data is analyzed and interpreted to prepare it for a visualization session.
3. *Display*: The analyzed data is then rendered and rasterized before sending it to the display. Based on the display capabilities, different rendering and rasterization algorithms may need to be used.
4. *Interactive commands*: Interactive commands from the end-users may need new data from the sensors, or new computations to be performed before displaying the modified results.

2.1.4 Involved resources

The Grid Resource Broker (GRB) has to acquire sensors, computational, storage and network resources depending on the distributed nature and complexity of the visualization session. Each of the components described above can be located remotely and/or require grid assistance to perform at an acceptable level.

2.1.5 Functional requirements

The data acquisition can be streamed from a high-throughput single sensor requiring transient storage and network resources or from many little sensors requiring only network resources. The data analysis portion can require computational grid resources that may or may not be local to the data acquisition site or data display site. The data display might require computation grid resources to render/rasterize the data which may/may not be local to the display screen. Interactive commands will typically be issued at or near the display terminal.

The functional requirements on some of the network services are listed below:

1. *Network Capability Discovery Service*: The geographically distributed nature of the visualization session will require the GRB to query the network capabilities like bandwidth, latency between the various data acquisition, data compute sites and data display sites.
2. *Network Resource Allocation Service*: The GRB might need to allocate the right Quality of Service (QoS) including bandwidth, latency, priority between the different visualization session locations. This quality of service reservation might depend on a pre-negotiated SLA for the visualization session
3. *Network SLA Monitoring Service*: This service might be used to monitor the ongoing network QoS for the visualization session and prompt the GRB in case the SLA negotiated is violated.
4. *Network Advanced Reservation Service*: This service may be used if the visualization/collaborative session is planned in advance and the requirements for the session are known. This works well especially when the data acquisition devices are one-of-a-kind and require prior reservation as well by the GRB
5. *Network Security Service*: It is possible for the visualization session to pass non-trusted network service providers. In this case, encryption, VPN, firewall or other network services might be requested by the GRB.
6. *Network AAA Service*: The GRB might need Authorization before allocating network resources and might need accounting records to provide to the application the amount of network resource used in a visualization session.

2.1.6 Security considerations

Signed and authorized requests from GRB will ensure no attacks or modifications to the network services requested.

2.1.7 Performance considerations

GRB's discovery of the computation and display capabilities as well as network capabilities between sites could modify the performance requirements of the computational aspect or the network aspect of visualization sessions. For example, a display with low resolution and low network bandwidth connection will require a

different rasterization algorithm to be run remotely and visualization data to be streamed to that display. A high-resolution display with a lot of processing capability might have the analyzed data streamed to it over a high-bandwidth network connection so the rendering of the data before display happens on the local compute cluster.

2.1.8 Use case situation analysis

The visualization use case has been discussed in research papers and presentations. There have been examples of such use-case implemented for certain science experiments.

2.1.9 References

Distance Visualization: Data Exploration on the Grid, Ian Foster et. al., IEEE Computer 1999.

Network Requirements of Ultra-High Resolution Visualization and Collaboration Environments--An Applications Perspective, Jason Leigh et. al., MCNC Workshop, April 23rd 2004

2.2 Remote Parallelized Visualization

2.2.1 Use case summary

Today's scientific community is relying more and more on visualization techniques for their scientific analysis and discovery. Increases in compute resources have led to larger data sets for analysis. Analytical tools such as remote visualization have found it necessary to find more effective mechanisms for rendering visualization of very large data sets. Parallelizing techniques have proved promising in three areas: i) server-side functions, ii) client side functions, iii) object rendering. Rendering and display have stringent bandwidth, latency and jitter requirements, especially when remoted.

2.2.2 Customers

The customers are researchers requiring visualization analysis of very large data sets on the order of terabytes to petabytes from remote locations.

2.2.3 Scenarios

The researcher initiates a parallel visualization session on remote compute resources via Grid service.

Multiple displays reside local to the researcher, together the display panels provide the rendering of a single visualization object. Prior to object rendering, the large data set was divided over multiple servers for more efficient computation and I/O. Each display is associated with a separate remote server for its server side numerically intensive computations.

Once the object is rendered on the displays the user can start their analysis. User input at the client (e.g., "change this isosurface level", "rotate display", "analyze variable X", "animate over time or space") generates control commands that are passed to the remote servers. The appropriate control commands are then sent to each remote server, triggering large flows of data/geometries across the network to the client side and then rendered to the display wall. Each remote server updates its associated client side display. Each update is latency and jitter sensitive.

As the user rotates the object via mouse movement, near-real-time object rendering occurs.

2.2.4 Involved resources

1. Grid network discovery service
2. Grid network monitoring service
3. Grid network connection service
4. Grid Resource manager service
5. Grid security service
6. Grid accounting service

2.2.5 Functional requirements (for Grid network service/services in our case)

. Functional requirements for the network:

- Grid network discovery service should exist and provide input network connection service.
- The user may first request a query on the remote destination address to determine if it is reachable.
- Prior to the user initiating a connection to the remote servers, the user may request availability of resources including the network resource.
- If advanced scheduling of portion of the resources are necessary, the user will make reservation for resources, otherwise will make requests on demand.
- The network resource request should contain bandwidth requirements, and other QoS parameter such as maximum tolerated jitter and latency.
- Due to the parallel nature of parallel visualization software, the network resources may be parallel end-to-end connections from each server to the associated displays. (This may be a very expensive/inefficient proposition to obtain). Each server update across the network should maintain the requested QoS parameters.
- Requested QoS parameters should be monitored to ensure compliance.
- Violations of QoS must be recorded, reported and attempted resolution.
- Each time the user maneuvers the visualization object , control messages are sent to the remote servers and in a near real-time fashion provide updates to the associated client side displays via a high speed network connection.
- Completion of the parallel visualization session should result in the release of resources.

2.2.6 Security considerations

Prior to visualization initiation, all security requirement s must be met.

2.2.7 Performance considerations

Updates from the remote servers to client side displays require near-real-time updates.

The transfer of data updates across the network should be executed with

very low network jitter

very high network bandwidth for the transfer of large data sets

very low network latency

2.2.8 Use case situation analysis

2.2.9 References

2.3 High Throughput File Transport with a Deadline

2.3.1 Use Case Summary

A particular challenge that arises in Grid infrastructures is the coordinated use of multiple resources. Here, workflows with potentially complex interdependencies have to be mapped to a distributed environment. A grid network service that assures the local access of remote data at a particular time could be used in workflow management frameworks to synchronize the coordinated use of resources and thus to avoid unnecessary blocking times due to missing staging data. This leads to the use case of a high throughput file transport with a deadline.

2.3.2 Customers

Scientific computing relies on the availability of appropriate computational capabilities. Existing and emerging virtual organization will provide access to multiple high-performance computing facilities to serve science and engineering with the demanded computational capabilities. To ease the use of such an infrastructure, advances in resource management will allow end-users to specify a workflow that is handed to a community scheduler that takes care on resource selection and job submission. To build these future resource management functions, high-throughput file transport with a deadline gives the ability to explicitly consider the relocation of data in advanced scheduling algorithm. Consequently, there are two types of customers:

1. A resource management service such as a community scheduler [SNAP] that dynamically maps workflows to resources.
2. End-users that negotiate a particular time frame for the remote execution of their program. An example would be visualization and steering application that is served by a supercomputer application.

2.3.3 Scenarios

Large-scale supercomputing is expected to produce data at a similar rate than large-scale experiments. To post-process the computed results, high-throughput transfers are required to stage the data at the related computational resources. Similarly, high-end scientific computing also processes large amounts of input data that, from a performance perspective, should be accessible as fast as possible. Local parallel file systems are well suited for supporting the demanded I/O capabilities; however, the data has to be staged to the respective file system.

A community scheduler that controls multiple distributed computational resources has to select resources that serve an individual workflow. In modeling the transport of data as an individual service that finishes at a particular time, the scheduler can potentially create a service level agreement for the whole workflow that assures a particular end-time, even though the computation is scheduled on a resource where the processed data is not yet available.

2.3.4 Involved Resources

Data has to be staged from a source to a sink. The demanded service assures that a given amount of data has been transported to the sink at some time t. This involves adequate transport capabilities and the appropriate use by transport protocols.

2.3.5 Functional Requirements

- Access to a Guaranteed Rate Service that assures a requested bandwidth between two end-points available to the requester. Note that this service has to be available end-to-end.
- Ability to negotiate the amount of guaranteed bandwidth, the end-points, and the time interval it is assigned
- Availability of a high throughput transport protocol
- Effective use of the assigned bandwidth by the transport protocol and its application

2.3.6 Service Utilization

This service is intended to support the map of abstract workflows to Grid environments. The related service agreement is negotiated by the user – either an end-user or a high-level service such as a community scheduler. Service provisioning has either to be performed by the service provider, i.e. some management software that assures a timely provisioning according to the established agreement, or by the end-user that actively signals the service requests. Of course, in the latter case, the user has to provide appropriate policy information that refers to the existing agreement and that assures its right to use this agreement. Security Considerations

Access to the service has to be explicitly granted by a management system that implements the appropriate admission control. Appropriate AAA-mechanisms are required.

2.3.7 Performance Considerations

There are two types of performance considerations:

1. Performance of negotiating and claiming the service parameters

   Here, efficient factory mechanisms were required to implement a state full agreement negation. Similarly, appropriate authorization mechanisms have to be applied when the service is claimed, particularly because the network is composed of multiple administrative domains.

2. Performance of using the service.

   Here, for optimization, deadline file transports will likely rely on both: a Scavenger Service for getting a share of unused bandwidth and a Guaranteed Rate Service that assures a negotiated level of service. Of course, the challenge of effectively using the guaranteed rate remains. Pacing the low-level traffic by using traffic shaping mechanisms has been approved as an appropriate solution to

assure that transport protocols can effectively use the underlying transport capabilities.

2.3.8 Use Case Situation Analysis

The implementation of a High-Throughput file transfer with a deadline has been analyzed in the context of the General-purpose Architecture for Reservation and Allocation (GARA) [GARA], a former research thread of the Globus Project (now Globus Alliance), and in the context of the German government funded project Path-Allocation in Backbone Networks (PAB) [PAB]. Scientific papers have been published in [MCL] and [E2E].

2.3.9 References

[SNAP] *SNAP: A Protocol for Negotiating Service Level Agreements and Coordinating Resource Management in Distributed Systems.* K. Czajkowski, I. Foster, C. Kesselman., V. Sander und S. Tuecke. Lecture Notes in Computer Science 2537, November 2002.

[GARA] *GARA: A Uniform Quality of Service Architecture*. A. Roy und V. Sander. "Grid Resource Management: State of the Arth and Future Trends", Edited by J. Nabrzyski, J. Schopf und J. Weglarz, Kluwer Academic Publisher, 2003 (ISBN 1-4020-7575-8).

[PAB] http://www.pab.rwth-aachen.de

[MCL] *Multi-Class-Applications for a Parallel Usage of a Guaranteed Rate and a Scavenger Service*. M. Fidler und V. Sander. In Proceedings of IEEE/ACM CCCGrid GAN 2003, May 2003.

[E2E] End-to-End Quality of Service for High-End Applications. I. Foster, M. Fidler, A. Roy, V. Sander und L. Winkler. Elsevier Computer Communications Journal, 2004. In press.

2.4

L2 virtual connectivity

2.4.1 Use case summary

Grid site managers and Virtual Organization (VO) administrators may be interested in the clustering of Grid nodes belonging to geographically dispersed Grid sites into a single Local Area Network (LAN), in order to enable a *closeness* relationship between nodes from different domains. Geographically distributed LANs are called Layer 2 VPNs.

The closeness to a Storage Element is particularly important during the resource matchmaking phase, when a Resource Broker needs to identify the best Computing Element for a given job in a range of different candidates. Computing Elements are *close* [GLUE] to the corresponding Storage Elements when they are member of the same local area network. The closeness can be reflected by

real (if the Computing and Storage Elements are physically connected to the same LAN) or virtual (they are part of the same L2 VPN).

The on-demand configuration of L2 VPNs allows the Resource Broker to select a Computing Element that is geographically remote, provided that the job input data can be retrieved from a Storage Element connected to the same L2 VPN, which implies that the input files do not need to be replicated close to the Computing Elements. This of course has an impact on the data replication policies that can be adopted for a given application, and it can contribute to limit the number of data replicas needed with a consequent reduction of the traffic generated on the links connecting large databases to the network and a simplification of the replica management.

2.4.2 Customers

There are two types of customers:

- the Grid site managers, who are responsible of defining for each local Computing Element the list of the Storage Elements that are *close* to it;
- the Virtual Organization (VO) administrators, who organize and supervise the membership of users and VO-specific resources to the VO.

2.4.3 Scenarios

The grouping of geographically dispersed resources, such as Storage Elements and Computing Elements, into the same virtual local area network can be used in the following scenarios.

1) Grid site managers may be interested in increasing the efficiency of the workload distribution and in limiting the amount of traffic induced by the replication of large data sets. This can be achieved by requesting that Computing Elements and Storage Elements that belong to different Grid sites, are part of the same (virtual) LAN, so that they can be considered *close* even if they are geographically distributed. In this way, thanks to the closeness relationship between computing and data resources, jobs can run on local computing facilities even when the input data are stored in a different Grid site. In order to do so, it is required that the resources belonging to the same virtual LAN are published in the local resource information directory, for example through the *close* attribute.

This can be possible if the resources from different sites are part of the same virtual LAN, in which case remote files can be accessed from other Grid sites through the technologies that are typically used in a LAN environment, such as NFS. A Grid node should be allowed to be part of different virtual LANs at the same time. If Computing and Storage Elements are member of the same L2 VPN, the files of interest to a job do not need to be replicated to a local storage system. This simplifies the data replication management, as it reduces the number of file replicas requested by the Grid. In addition to this, the traffic load produced by the periodic data replication process can be alleviated on the network links that provide connectivity between the Grid and the Storage Elements.

2) Small grid sites can be virtually extended to include remote resources such as Computing Elements and Storage Elements. In this way a larger set of Computing Elements can be selected to run a job, with a consequent increased efficiency in the distribution of the work load.

3) By dynamically clustering the distributed resources dedicated to a given VO, a "VO computing facility" can be simulated on top of the Grid. In this way the VO can make sure that the Computing Elements dedicated to it are always involved in the resource matchmaking phase for the job launched by its users, even when local copies of the input data are not available.

2.4.4 Involved resources

The list of the resources involved in this use case includes:

- Computing Elements;
- Storage Elements;
- Customer edge and transport networks;
- Network capacity and, in general, all the resources that need to be allocated in a network device in order to guarantee a given traffic forwarding behavior.

2.4.5 Functional requirements

The use case requirements are summarized in the following list.

1. The client requests the configuration of a L2 VPN in order to cluster geographically dispersed nodes into the same (virtual) Local Area Network.
2. A given Grid node can belong to two or more L2 VLANs at the same time.
3. The client needs to be authenticated and before proceeding with the analysis of the request issued, which then needs to be authorized.
4. The type of L2 VPN requested is specified by a set of parameters including:
    - The IP addresses of the members of a given VLAN. The IP addresses correspond to Grid nodes such as Computing and Storage Elements.
    - The time interval during which the L2 VPN should be enabled. This allows the client to request the set-up of a L2 VPN in advance.
    - The type of traffic forwarding behavior to be associated to the requested VPN. In this way the Grid nodes that belong to a given Layer 2 VPN can experience a specifically tailored Quality of Service.

5. The client may want to request a change in the configuration of the L2 VPN dynamically. However, we expect the L2 VPN configuration to vary infrequently.
6. The privacy and security of data exchanges between the L2 VPN nodes should be guaranteed.
7. Data exchange performance across the L2 VPN should not be penalized by the presence of firewalls and the data transfer performance across the nodes of the L2 VPN should be guaranteed.

2.4.6 Service utilization

The support of the use case requires the interaction with a number of Grid general and network-specific services. The following list shows the sequence of operations generated by a client request.

1. The client (Grid site manager or VO administrator) requests in advance the set-up of a virtual LAN. A list of attributes needs to be specified, such as: the IP addresses of the remote nodes that will be member of the virtual LAN, the time interval, the traffic forwarding Quality of Service (QoS) selected, etc.
2. The client is authenticated and authorized.
3. The QoS level requested for the virtual LAN is negotiated.
4. If the L2 VLAN is successfully set-up, the Computing and Storage Elements connected to it are said to be virtually *close*. This change in the status of the resources has to be consequently reflected in the resource information directory.

2.4.7 Security considerations

Clients issuing a set-up request need to be authenticated and authorized.

Data exchanges across an on-demand L2 VPN require privacy and security

2.4.8 Performance considerations

The data exchange over the L2 VPN should not be affected by the performance limitations introduced by firewalls.

Considering the limited number of clients (VO Managers and Grid site administrators) that are involved in this use case, we expect the number of requests to be generated over time to be limited. Consequently, this use case does not causes particular problems in terms of scalability.

2.4.9 Use case situation analysis

This use case has been addressed in the framework of the IST project DataTAG. A prototype of the service based on the MPLS technology and the IP Premium and Less Than Best Effort services offered by GÉANT (the European research backbone network), was implemented [D2.5].

2.4.10 References

[GLUE] Andreozzi, S.; *GLUE Schema implementation for the LDAP model;* May 2003 (http://www.cnaf.infn.it/~sergio/publications/Glue4LDAP.pdf).

[D2.5] *Demonstration of Advance Reservation and Services*, DataTAG Deliverable 2.5, March 2003 (http://edms.cern.ch/file/431913/2/D2.5-1.5.pdf).

2.5 Quality of Service Path for Grid applications and middleware

2.5.1 Use case summary

The capability to transfer data across the Grid is of great importance given its inherent distributed nature. However, most of the transport networks today offer a single best-effort traffic forwarding behavior, which means that no performance can be guaranteed. Application and Grid middleware typically have different traffic forwarding requirements that can be satisfied by enabling traffic differentiation techniques in the network infrastructure. The traffic forwarding profile needs to be expressed by means of performance metrics such as: the achievable bandwidth [NM], one-way loss [OWPL, OWLP], one-way delay [OWD], delay variation [IPDV] etc. These terms are specified to the path provider during the negotiation phase. . Different categories of applications and middleware can be defined according to the different requirements they have, as shown in the following non- exhaustive list:

- *Applications handling audio/video/image content*: they require low packet loss and the minimization of one-way delay and instantaneous packet delay variation;
- *Short-lived, reliable data transactions:* they require data transfer reliability and the maximization of the number of completed transactions over time. Packet loss and delay minimization are important to reduce the number of retransmitted data units and to ensure a timely recovery in case of congestion;
- *Long-lived, intensive data transfers:* they require the maximization of throughput. The packet loss rate needs to be minimized especially at high speed, given the penalty introduced by some packet transport protocols such as TCP, which reduce the transmission rate every time a data unit is lost.

2.5.2 Customers

Grid user applications and middleware with specific traffic forwarding requirements.

2.5.3 Scenarios

Applications handling audio/video/image content

Videoconferencing, remote visualization, real-time remote analysis of images and tele-immersion are examples of applications performing remote processing of large databases of images and requiring remote visualization of the processed result. In addition, traditional videoconferencing applications producing audio and video traffic are used for computer-supported cooperative work. In this case, three critical parameters can affect performance: packet loss frequency (for good video and audio quality and image resolution), One-Way Delay (for timely delivery of images and voice) and Instantaneous Packet Delay Variation (IPDV) (for good audio quality).

Short-lived, reliable data transactions

Data-oriented applications requiring frequent access to small data segments as for remote file analysis and client/server transactions in GRID middleware, require data transfer reliability and are particularly affected by packet loss, which reduces the data rate when congestion control and avoidance algorithms are used at the transport protocol level. In

addition, packet loss reduces the number of completed transactions per time unit, a parameter that is more critical than throughput itself, given the relatively small amount of data exchanged. One-way delay minimization is also important for timely communication between servers and clients.

Bulk data transfers

Data management operations causing replication of large database portions and jobs accessing very large data collections are examples of middleware software and applications moving a very large amount of packets across the network. The difference between this group and the previous consists in the different amount of data exchanged and the frequency of network transactions. Bulk transfers are likely to be rather infrequent but with a well-defined scheduling. In this case, throughput achieved for each data transfer is the critical parameter as it determines the data exchange completion time and the efficiency in network resource utilization. Throughput with reliable transfer protocols like TCP is critically influenced by the packet loss rate and loss pattern experienced during transmission. In fact, for every lost data unit, the output rate at the source is dynamically reduced in order to cope with congestion. For this reason, packet loss is highly undesirable, especially when running on high-speed infrastructures. Therefore, the packet loss pattern (isolated packet loss vs. packet loss bursts) determines the efficiency in resource utilization.

For this group, the guarantee of a minimum bandwidth is important to estimate the transaction finish time and to avoid network bottlenecks produced by multiple concurrent transactions for a given data source.

The required Quality of Service is expressed in terms of *guaranteed delivery* of a complete data file or in throughput predictability. The application or middleware may want to specify the ultimate delivery time, a target throughput or a stability level for data delivery. In this way, job schedulers and resource brokers are allowed to co-schedule data transfer/processing and to determine the best data sources from which information can be efficiently delivered to users and jobs.

2.5.4 Involved resources

- Computing and Storage Elements
- Hosts running Grid middleware with Quality of Service requirements
- Grid site networks and transport networks

2.5.5 Functional requirements

- The traffic forwarding behavior requested by a client needs to be quantitatively described through a list of metrics that are used as negotiation terms. The client needs to be provided with the list of attributes that can be used to specify the traffic forwarding profile.
- A client specifies to what traffic the requested traffic forwarding behavior needs to be applied. This is done by identifying the sources and destinations (through the IP addresses of the end-systems or the corresponding IP network addresses), the transport protocol and its port numbers, or a subset of this list, depending on the granularity of the traffic specification.

- Sources and destinations can belong to either the same administrative domain (intra-domain scenario) or to different domains (inter-domain scenario).
- The client should be given the possibility to request a traffic forwarding behavior in advance.
- The client should be given the possibility to negotiate the modification of the traffic forwarding behavior profile, if needed, and the time interval during which the behavior is requested.
- The clients need to be informed about the result of the negotiation process.
- The client should be given the possibility to issue short-term QoS requests.
- During the interval when the traffic forwarding behavior is guaranteed to the client, the client needs to be provided with feedback about the actual performance experienced by the traffic affected by the forwarding behavior requested.

2.5.6 Service utilization

- QoS requests need to be authenticated and authorized.
- The client establishes a negotiation session to request a given traffic forwarding behavior profile during a specified time interval.
- The client is provided with feedback with performance information of the traffic to which the requested traffic forwarding behavior applies.

2.5.7 Security considerations

Clients' requests are authenticated and authorized.

2.5.8 Performance considerations

The time needed to accomplish the allocation of a given traffic forwarding behavior needs to be sufficiently short to allow the client to issue short-term requests.

2.5.9 Use case situation analysis

Part of the functional specification defined in this use case have already been supported by a number of prototypes such as GARA [GARA], which is applicable in IP-based networks and relies on the Differentiated Services architecture [DS].

2.5.10 References

[NM] Lowekamp, B. et alt.; *A Hierarchy of Network Performance Characteristics for Grid Applications and Services*; the Network measurements Working Group, GGF, Work in progress.

[OWPL] Almes, G.; Kalidindi, S.; Zekauskas, M.; *A One-way Packet Loss Metric or IPPM;* RFC 2680, Sep 1999.

[OWLP] Koodli, R.; Ravikanth, R.; *One-way Loss Pattern Sample Metrics;* RFC 3357, Aug 2002.

[OWD] Almes, G. et alt.; *A One-way Delay Metric for IPPM,* RFC 2679.

[IPDV] Demichelis, C.; Chimento, P.; *IP Packet Delay Variation Metric for IPPM*, RFC 3393, Nov 2002.

[GARA] Roy, A.; Sander, V.; *GARA: A Uniform Quality of Service Architecture*; published in *Grid Resource Management: State of the Art and Future Trends*, Kluwer Academic Publishers, Fall 2003, pp. 377-394. Editors: Nabrzyski, J.; Schopf, J., M.; Weglarz, J.
[DS] Blake, S.; Black, D.; Carlson, M.; Davies, E.; Wang, Z.; Weiss, W.; *An Architecture for Differentiated Service*, Dec 1998.

3 Knowledge-based use cases

3.1 Service optimization [Mark, Paul, Tiziana]

3.1.1 Use case summary

Network performance data can provide very useful information to the Grid middleware and user applications involved in network transactions. Performance metrics can be used to estimate the "cost" of transmission between two given nodes, where the cost model varies depending on the Grid application and middleware requirements. The cost model is based on a set of network performance metrics and provides a summarized high-level view of a networked session. The cost can be used to identify the *best* destination node (clients, servers etc) from a set of candidates.

Customers

The network cost can be useful in a number of scenarios. Resource brokers and data replication managers are two examples of possible customers who could use this information to optimize their networked sessions:

3.1.2 Scenarios

Grid job scheduling

The Grid Job Scheduling Service selects a computing node from a list – often geographically distributed – candidates. This decision can be taken according to selection rules that take into consideration the requirements of the job (such as the software environment available on a given Computing Element, the amount of free disk space, available CPU, etc), and network performance, which characterizes the quality of data transmission on the path connecting a Computing Element to its input files sources. In order to do so, the "network cost" of a given path needs to be estimated on the basis of historic or estimated future network performance.

1) The user submits a job description to a Resource Broker using a user interface. The job description contains:
   a) The logical names of the data required for the job;
   b) Possibly the Storage Element on which output files from the job should be stored;
   c) Other information regarding requirements on how long the job must last, processor and software requirements and so on.
2) The Resource Broker finds all the Compute Elements that match the job's requirements.
3) The Resource Broker finds the location of all the replicas of the logical data files required by the job.
4) For each Compute Element, the Resource Broker finds the total cost to make the replicas available to the job.
5) The Resource Broker also finds the expected cost of storing the output of the job at a Storage Element for each Compute Element, if that has been specified in the job description.
6) The Resource Broker then chooses the Compute Element with the lowest total cost for running the job.

3.1.2.1 Input/output file management

Not only can the job execution be optimized by taking in consideration network performance information, but also the management of input/output files during a job execution can be more efficient when network costs are considered, as a job input/output process can produce considerable data traffic across the Grid. Storage Elements can be selected so that the amount of traffic to be exchanged from/to a Computing Element is minimized and/or the nodes with a suitable network connectivity performance are given higher priority.

1) A job requires a data file which has a logical name, and replicas might be stored at any number of Storage Elements.
2) Location and retrieval of that file is handled by specific Replica Management middleware.
   a) The Replica Management middleware finds the locations of all replicas of the required file.
   b) The total cost of making those replicas available to the job are then calculated.
   c) The replica with the lowest cost is then selected and transferred to a location where it is available to the job.

3.1.2.2 Data replication

Estimated network costs can be used to improve the efficiency of data management among different Storage Elements (SEs), e.g., to select the best replica of a given file (if there are copies in different SEs), to identify the most appropriate SEs when a given amount of data has to be replicated, and to manage input/output data fragments in a single SE. For example, in the last use case, it may happen that input/output data of a given job is fragmented and distributed among a number of SEs. If the fragments need to be gathered into a single SE, then the most appropriate SE has to be identified. The cost model can be based on principles such as the minimization of the amount of data exchanged between SEs, the identification of the SE with the lowest packet loss probability or with the maximum available bandwidth.

3.1.2.3 Adaptive remote file access

In some job execution scenarios, an application may decide what file/files it needs to access only at run time. In this case the information about the identity of the input files accessed is missing and it cannot be used by a Resource Broker during the matchmaking phase to statically allocate suitable Computing Elements to the application. For this reason, it becomes important to provide the application itself with that the possibility to dynamically adapt the source of its input file at run time. The optimization can be based on the dynamic adjustment of the Storage Element set that the application is using as the file access pattern changes, by taking into account the network performance experienced on the paths connecting the Computing Element to the Storage Elements in use.

3.1.3 Involved resources

- Storage Elements
- Computing Elements

- Network connections used to transfer replicas

3.1.4 Functional requirements

A cost must be calculable for every Grid resource (Compute and Storage Elements), including brand new resources. Changes in the performance of networks associated with resources should quickly be made available to brokering and optimization middleware so that resources are optimally used.

3.1.5 Service utilization

3.1.6 Security considerations

Information about the state of the networks and components of the Grid might be useful in targeting malicious attacks.

Incorrect information associated with resources might result in poor decisions by brokering middleware. Injecting bad information into the system might constitute an effective denial-of-service attack as certain resources are swamped, while others remain unused.

3.1.7 Performance considerations

The use of network cost information in the workload management scenario requires scalability and good responsiveness from the server, as the number of jobs handled by resource brokers can be considerable. For each job matchmaking may generate multiple cost estimation requests.

3.1.8 Use case situation analysis

A prototype of a service addressing the first scenario described in this document has been implemented in the framework of the EU project DataGrid [D7-4]. The following documents provide information on the prototype features and implementation [FG, NCES].

3.1.9 References

[D7-4] *Final Report on Network Infrastructure and Services, DataGrid Deliverable 7-4*, Jan 2004, https://edms.cern.ch/file/414132/2.1/DataGrid-07-D7-4-0206-2.0.pdf.

[FG] Ferrari, T.; Giacomini, F.; *Network Monitoring for GRID Performance Optimization*, Computer Communications Journal, pre-print version, Mar 2003 http://www.cnaf.infn.it/~ferrari/papers/myarticles/comp-comm2002.ps.

[NCES] The Network Cost Estimation Service, http://ccwp7.in2p3.fr/nces/

3.2 Administrative setup of schedules of measurements

3.2.1 Use case summary

Administrators require regularly scheduled and ad-hoc measurements for a variety of reasons.

3.2.2 Customers

- Administrators setting up measurements for monitoring the state of the network, and to provide data for use when diagnosing possible problems later.
- Administrators might also wish to manually set up measurements to aid middleware in optimising the functions of the Grid.
- Middleware services might also set up measurements in response to changes in configuration or usage.

3.2.3 Scenarios

Administrator setting up a single ad-hoc measurement

1) An administrator wishes to make a single measurement between two end-points, either of which might be outside his immediate control.
   a) The administrator knows the metric and certain parameters for the measurement, plus the names of the two endpoints. These parameters might include application-level and protocol-stack-level settings, as well as more exotic settings, such as requiring a particular type of service.
2) The administrator must be able to retrieve a measurement of the metric with the parameters and endpoints he specified after the measurement is completed.

Administrator setting up a temporary schedule of measurements

1) An administrator wishes to set up a series of regular measurements for a short period of time, in order to monitor an expected change, or for troubleshooting purposes.
   a) The administrator knows the metric and certain parameters for the measurement. These parameters might include application-level and protocol-stack-level settings, as well as more exotic settings, such as requiring a particular type of service.
   b) He also knows the approximate frequency at which measurements should be made.
   c) Finally, he knows that no more measurements should be made after some cut-off time, as they might be intrusive.
2) The user will watch the results of measurements as they are made.
3) After some time, the user decides that more measurements must be made past the original cut-off time.
4) After further time, the user decides that he has all the information he needs, and so stops the measurements altogether.

Administrator setting up a permanent schedule of measurements

1) An administrator wishes to set up a permanent schedule of regular measurements between two end-points. The results of these measurements might be used for a number of reasons:

a) to inject new information into the grid information systems for use by optimisation services;
b) to allow changes in the state of the networks to be flagged quickly, and so provide early warning of failures or other problems;
c) to ensure that service-level agreements are kept to.

2) The administrator must instantiate regular measurements between two nodes either of which may be outside his immediate control.
   a) The administrator knows the metric and certain parameters for the measurement. These parameters might include application-level and protocol-stack-level settings, as well as more exotic settings, such as requiring a particular type of service.
   b) He also knows the approximate frequency at which measurements should be made.
   c) Finally, he knows that no more measurements should be made after some cut-off time, as they might be intrusive

Setting up a permanent schedule of measurements for use by other middleware

1) An administrator wishes to set up a permanent schedule of regular measurements to characterise the connection between two resources on the Grid. The results of these measurements might be used by other middleware for optimisation purposes.
2) The administrator knows the names of the resources (that is, the name of a computing element and a storage element, for example). The names of the monitoring points themselves are unknown, and either of them might be outside the direct control of the administrator.
3) The administrator wants to ensure that no duplication of effort occurs: that is, if measurements are already being made which are applicable to the resources, these measurements are not duplicated for the new resources.
4) The measurement schedule should effectively run for the lifetime of the Grid, unless the resources change, or different types or volumes of information are required. Therefore:
   a) the administrator may wish to change any of the settings of the schedule; or
   b) the administrator may wish to cancel the measurements altogether.

3.2.4 Involved resources

- The *endpoints* of the measurements must not be loaded (processor or otherwise) such that measurements are perturbed.
- The networks in between the endpoints are usually fairly passively involved in the situation – they can be thought of as passive conduits of data. However, in certain cases (when measuring SLAs for example), the networks may be affected by other network services, such as bandwidth allocation services.

3.2.5 Functional requirements (for Grid network service/services in our case)

-

3.2.6 Service utilization

Advance reservation of resources is being dealt with by Working Groups of the GGF. Any relevant results of this work must be utilised.

3.2.7 Security considerations

- Making intrusive measurements between dedicated and privileged machines on a network can result in poor network performance for other users. The potential for a service to be used as a platform for denial-of-service attacks is very great.
- This mechanism could be used as a method for injecting bad information into the grid information system, if, for example, a measurement could be engineered to appear unfavourable. If the results of users' measurements are used by other Grid components, appropriate safeguards must be in place to ensure that those measurements cannot adversely affect the operation of the Grid.

3.2.8 Performance considerations

3.2.9 Use case situation analysis

3.2.10 References

I think I may be able to find some references from either E2E piPES or network troubleshooting work done by Richard HJ, Mark and so on.

4 Security Considerations

TBD

5 Authors Information

TBD

6 Intellectual Property Statement

The GGF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the GGF Secretariat. The GGF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the GGF Executive Director (see contacts information at GGF website).

7 Full Copyright Notice