

20.12.05

Deliverable DJ.3.3.1: GÉANT2 Bandwidth on Demand Framework and General Architecture



Deliverable DJ.3.3.1

Contractual Date:	30/06/05
Actual Date:	20/12/05
Contract Number:	511082
Instrument type:	Integrated Infrastructure Initiative (I3)
Activity:	JRA3
Work Item:	03
Nature of Deliverable:	R (Report)
Dissemination Level	PU (Public)
Lead Partner	GARR
Document Code	GN2-05-208v7

Authors: Maarten Büchli (DANTE), Mauro Campanella (GARR, Editor), Gábor Ivánszky (HUNGARNET), Radoslaw Krzywania (PSNC), Bram Peeters (SURFNET), Damir Regvar (CARNET), Victor Rejjs (HEANET), Laura Serrano (GARR), Afrodite Sevasti (GRNET), Kostas Stamos (GRNET), Chrysostomos Tziouvaras (GRNET), Dave Wilson (HEANET)

Abstract: The Bandwidth on Demand service, as defined in the context of the GN2 project, aims at providing a guaranteed capacity, connection oriented service between two end points. The deliverable elaborates on a technical framework and a draft architecture for the service to be deployed in a large multi-domain network with multiple transport technologies. This document outlines a BoD service architecture, which is sufficiently detailed to allow a modular first phase of implementation using manual or semi-automated procedures. The implementation will start from the Inter-Domain communication module and leverage existing components. Refinement and improvements will result from the implementation experience.

Table of Contents

0	Executive Summary	v
1	Introduction	1
2	Terminology and Assumptions	3
2.1	Service General Principles and Assumptions	4
2.1.1	End to End	4
2.1.2	Service Guarantees and Availability	4
2.1.3	Service User Interface	5
2.1.4	Authorisation, Authentication and Accounting	6
2.1.5	Multi-domain	6
2.1.6	Monitoring	6
2.1.7	BoD with Packet- or Circuit-based Technologies	7
2.1.8	Integration of Technologies	7
2.1.9	Choice between equivalent Transport Technologies	8
2.1.10	Layer 3 Routing and BoD	8
2.1.11	Time synchronization	9
2.2	Design Procedure	10
2.2.1	Identification and Definition of a Service	10
2.2.2	Identification of the Service's Main Building Modules	11
2.2.3	Definition of Functional Specification of Modules	11
2.2.4	Definition of the Communication Flow	11
2.2.5	Use Cases	11
3	Summary of Requirements	12
4	Definition of the Bandwidth on Demand Service	14
5	Architecture Definition	16
5.1	Architecture Overview	16
5.2	Abstract Representation of the Network	18
5.2.1	Abstract Representation: Objects	20

5.3	Interaction with other Services	22
5.3.1	Authentication and Authorization Infrastructure	22
5.3.2	Monitoring	23
5.3.3	Network Management Service	23
5.3.4	Premium IP	24
5.3.5	Database and Archival Service	28
5.4	Modules and Blocks, Functionalities	28
5.4.1	Inter-domain Manager	28
5.4.2	Domain Manager	31
5.4.3	Logic and Policies Module	33
5.4.4	Location Service Module	35
5.4.5	Technology Proxies Modules	36
5.4.6	Pathfinder Module	38
5.4.7	Information Storage System	42
5.5	Connection Diagrams and Data Flow	45
6	Conclusions	46
Appendix A	References	47
Appendix B	Acronyms	49
Appendix C	Terminology	52
Appendix D	Connection Diagrams and Data Flow	57

Table of Figures

Figure 5.1: Sample domain with some services enabled	17
Figure 5.2: Modules of the BoD system	18
Figure 5.3: A sample multidomain BoD service case	19
Figure 5.4: Abstract representation objects	20
Figure 5.5: Interaction between BoD system modules and external services (when available)	22
Figure 5.6: A PIP domain in a chain of BoD domains	26
Figure 5.7: Inter-domain Manager main blocks	29
Figure 5.8: Domain Manager Building blocks	32
Figure 5.9: The blocks in the Logic and Policies module	33
Figure 5.10: Technology Proxy blocks	37
Figure 5.11: Pathfinder blocks	41
Figure 5.12: Information Storage System	43
Figure D.1: Regular BoD service reservation process	58

0 Executive Summary

The Bandwidth on Demand service, as defined in the context of the GN2 project, aims at providing a guaranteed capacity, connection oriented service between two end points. The deliverable elaborates on a technical framework and a draft architecture for the service to be deployed in a large multi-domain network with multiple transport technologies. A domain will provide the BoD service between its demarcation points. In this context, each domain, as identified by its demarcation points, contributes to the end-to-end BoD service.

A key element of the BoD service provisioning architecture in each domain is therefore the Inter-Domain Manager, which is responsible for BoD service request reception, approval and instantiation. The inter-domain communication is the part of the system, which can rely less on known, proven and existing standards. It is the part that has to be fully developed within the JRA3 activity of the GN2 project.

When the end points are in different domains, the architecture requires the Inter-Domain Managers to cooperate on a peer-to-peer basis to create the requested end-to-end path. Such a distributed model allows a great level of independence between domains. Each domain can choose its own policies and technologies used to provide the BoD service within its borders. The peering model also allows scaling to a large number of domains and imposes loose or no synchronisation constraints upon the implementation.

The high degree of independence from the technologies used in the physical implementation of the network BoD service in each domain is achieved by adopting a service and resources' abstraction layer and by clearly defining the building modules and blocks of the system and their functions, so that their interfaces can be clearly implemented.

This document elaborates a BoD service architecture, which is sufficiently detailed to allow the first phase of implementation using manual or semi-automated procedures. It is expected that refinement and improvements will result from the feedback of the implementation experience.

During the development process of the architecture, it became clear that abstracting the network representation and the services definition would provide a simpler and more powerful environment. The effort would be beneficial not only to the bandwidth on demand service, but in general to the development of a multi-service network and to simplify and standardize inter-domain communication.

Although the BoD service architecture presented here is complete, and therefore complex, its modularity allows fast prototyping and a modular implementation, with inclusion of existing components through interfacing.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

In particular, the implementation will initially focus on the inter-domain module and provide reference implementation for the other modules or define interfaces towards existing systems with the required functionality, like intra-domain managers and network management systems.

1 Introduction

Traditionally, providing dedicated network capacity to users is a relatively slow process, in particular for high-speed connections. The development in network transmission technologies and the widespread use of optical fibres for data transmission allow new network usage scenarios, in which capacity may be provided to the end user in much shorter time.

This technical possibility arises at the same time with the appearance of specific requirements from the GRID community, as well as from research groups in Astronomy, Supercomputing and High-Energy Physics, to get access at dedicated high-speed connections at minimum notice, or even in real-time.

The GN2 Joint Research Activity 3 aims at the design, specification and implementation of a Bandwidth on Demand (BoD) service for the European research networks. The service should provide a connection oriented point-to-point service with guaranteed bandwidth through one or more domains. In other words, the service should be offered over the GEANT network and the European National Research and Education Networks (NRENs). Creating such a service over multiple domains represents a challenge, as most of the current specification and standards do not cover the inter-domain case. From previous GN1 and NREN experience [GN1-D.27] it is known that establishing inter-domain "light path" services requires a lot of interaction between the different involved networks. JRA3 aims to engineer this process, to streamline and finally automate it.

The previous work of GN2 JRA3 in work item 2 resulted in a user survey on user's BoD requirements [DJ3.2.1] and a state-of-the-art study of BoD-related technologies [DJ3.2.2]. The requirements for a BoD service are summarised in this document.

At the same time, relevant initiatives such as the UCLP [UCLP] (User-controlled LightPaths) platform of CANARIE and GLIF (the Global Lambda Integrated Facility) initiative have been analysed when conducting the aforementioned state-of-the-art study and have influenced the work presented here. As the Activity evolves, the issues of inter-working with such existing systems will be dealt with in more detail.

This document is the result of GN2 JRA3 work item 3. The aim is at specifying a generic framework and architecture for an inter-domain BoD service. Within the architecture the different functional modules, blocks and interfaces are identified and described. Future work of work item 3 will consist of detailing and extending this framework and architecture, in particular for inter-domain communication and system automation. The implementation will be initially based on manual or semi-automated procedure. The implementation of the functional modules will gradually be automated by means of software.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

In the process of defining the generic framework and architecture for an inter-domain BoD service, the need to broaden focus - in the early stages of GN2 JRA3 work - from the single domain, manual BoD service specification to the definition of the communication model and processes for inter-domain BoD provisioning clearly emerged. This approach is followed here and in the next phases of the GN2 JRA3 work. Solving the inter-domain BoD provisioning problem is the 'glue' that allows streamlining of the process and defining the information exchange necessary to set up inter-domain BoD service instances. This is the case regardless of the underlying individual per-domain implementations and applies also to manual per-domain BoD provisioning. Tackling first the inter-domain functionality of the BoD service is also more practical, since the inter-domain layer will be common among all domains and thus the corresponding specification and implementation work is and must be directly applicable to each domain involved from 'day 0'. This is considered a more pragmatic approach, it is allowed by the modularity of the architecture and allows the activity to start immediately the implementation phase.

As a result of this change of plans in the activity, the intra-domain BoD provisioning specification has been temporarily postponed and this document has been re-titled from 'Definition of Phase 1 BoD service (single domain, manual provisioning)' to 'GÉANT2 Bandwidth on Demand Framework and General Architecture'.

Sections 3 and 4 of the document discuss and define the framework upon which the architecture stands. After the terminology description, the key elements and basic assumptions for the service are presented. It includes a brief description of architectural design principles and a summary of users' requirements as emerging from the results of JRA3 work item 2.

The remaining sections define the architecture. The basic modules, their components and the service interaction with other services are described. Preliminary description of the interface parameters for each module is included.

The implementation of the BoD system modules will be carried by GN2 JRA3 work item 4, while work item 5 has the task to test the system. A continuous feedback process is expected between the different work items.

Although the BoD service architecture presented here is complete, and therefore complex, its modularity allows fast prototyping and a modular implementation, with inclusion of existing components through interfacing.

In particular the implementation will initially focus on the inter-domain module and provide reference implementation for the other modules or define interfaces towards existing system with the required functionalities, like intra-domain managers and network management systems.

Interfacing with other BoD architectures, as UCLP will be possible by developing proxy interfaces.

2 Terminology and Assumptions

The National Research and Education Networks are traditionally based on IP packet switching and NREN transport packet flows by statistical multiplexing and aggregation. The rapid developments in data transmission and computer processing speed are opening new possibilities for creating large distributed systems and operating them as a single entity.

For some of these applications, a connection-oriented, end-to-end (therefore multi-domain) “Bandwidth Allocation and Reservation Service” may well complement the classic Best Effort service. The term “connection-oriented” has been used in order to highlight the fact that the bandwidth should be reserved, not contended, exhibit deterministic performance and be logically separated from other traffic sharing the network. The service can be realized both through traditional Layer 2 circuit oriented transport technologies, like Ethernet or SDH and Layer 1 wavelengths on fibre (“lambdas”).

The creation of “circuits with assured capacity” using packet based technologies at Layer 3 is also possible, albeit less straightforward and requires more complex tasks in the network nodes. As an example, a virtual circuit, connection oriented technology like GMPLS can be used at Layer 3 to create Label Switched paths (LSP) which, to assure capacity, must be marked with QoS tags and handled appropriately by the network.

The proposed architecture focuses on Layer 1 and 2 circuits, but it does not overrule or exclude the use of Layer 3 technologies for the provisioning of BoD service, at least for some parts of the end-to-end path. This approach might be considered as a compromise to the required connection-oriented nature of the BoD service, but, taking into consideration the fact that a lot of the underlying infrastructure in the R&E community in Europe is still IP-enabled, the approach is considered pragmatic. Furthermore, this choice does not at all compromise or weaken the proposed framework in its efficiency to accommodate the use of Layer 1 & 2 technologies for BoD service provisioning.

Appendix A contains a definition of key terms used throughout this document. The definitions refer to the way in which these concepts are understood with respect to bandwidth on demand framework and architecture specification. A precise definition of terms like “capacity” or “bandwidth” is provided.

The following sub-sections report the summary on some general principles discussion. The general principles discussion highlights the key elements of a BoD service and explains the decision taken, when needed.

2.1 Service General Principles and Assumptions

The following section provides some general principles and service characteristics upon which the BoD service architecture will be built.

2.1.1 End to End

The “End to End” term refers to the end-to-end principle at the base of Internet, as detailed in the original article on “End-to-End arguments in system design [Saltzer84].

The BoD service will be defined initially as a service between demarcation points in a domain and may imply a certain amount of complexity in the network itself. BoD service enabled domains may create a daisy chain to offer a multidomain service.

The final goal of the service is to have widespread availability.

2.1.2 Service Guarantees and Availability

The service will provide to the user the equivalent of an end-to-end circuit with assured and fixed “capacity” [Appendix C, C.5]. The user is interested usually in assured “bandwidth” [Appendix C, C.4] which is the amount of user data that an application is capable of transferring per unit time. The bandwidth usable by the application is actually a function of the transport protocol used, the Layer 3 protocol, node performance, MTU and packet size and so on which are not under the network control.

It is then important to inform the user that the bandwidth seen by the application may vary when compared to the contracted capacity value and thus provide a clear definition of the service provided.

A simple estimate of the bandwidth which an application may be able to obtain, given a contracted capacity value can be computed as the maximum amount of bit/seconds available to the user (payload) for that capacity, when using end-to-end full size maximum transfer unit (MTU) packets, no IP options, TCP with only the timestamp option and Ethernet as datalink in the users’ host.

As the capacity resource is finite, the service might provide a negative answer to a user’s request. In addition, the user must not assume a 100% availability of the service, but a lower percentage due to unavoidable physical constraints, like Bit Error Rate, failures on unforeseen events. The need to formalise the user requirements and network assurances in a Service Level Agreement (SLA) as well as its content, will be evaluated during deployment of the service.

2.1.2.1 Service Granularity

The architecture will not specify a minimum or maximum amount of Bandwidth a user can request, as the limit is only related to technology. For each technology, the service granularity will be defined during implementation.

In the cases where both the BoD and Premium IP services are available in the same domain, both services might be used to procure the equivalent of an end-to-end path. As a rule of thumb, the BoD service may cover a range of provisioned capacities that the Premium IP (PIP) service is unable to provide due to the PIP service rule, which limits the aggregated PIP traffic to 10% of the capacity of each link. As the situation in Europe is uneven in term of capacity, the minimum BoD service capacity can be as low as 10 Mb/s. The upper limit in principle is defined only by the technology availability. Dynamic sharing of link capacity between services is not considered for the time being as it will add to the complexity of the service models.

When, on the same link, different services reserve capacity, the capacity value for each service varies between zero and a maximum, predefined value. The 10% rule of PIP service applies to the total capacity available to PiP and the rest of Layer 3 (IP) traffic and no longer to 10% of the physical total capacity. If the available capacity for PiP and Best effort decreases because of BoD reservation, the 10 % rule applies to the new value.

It is considered safer to plan in advance the maximum values of the capacity devoted to PiP and other services on the same link to simplify the allocation.

2.1.2.2 Backup, Resilience and Path Diversity

The BoD service will consider the feasibility to provide backup and resiliency for the end-to-end circuits. The task implies considering additional resources in order to provide guarantees to the user in case of failure of the primary circuit.

Providing an explicitly diverse path for the primary and the backup circuit is considered useful and should be possible. The user should have the possibility to specify the level and type, if any, of the backup required, possibly including backup at layer 3 as discussed in subsection 0 or physical path diversity.

2.1.2.3 Other Quality of Service Assurances

The service is targeted primarily at providing assured capacity and no packet loss. It will be possible at a later stage to consider requests with additional requirements on delay and IPDV or other QoS parameters.

2.1.3 Service User Interface

The requests will be accepted through a user interface, which can be a web interface or an application-programming interface. The time needed to process the requests will depend on the stage of BoD service implementation. In the first stage of the service there will be a manual configuration. In the last stage, it will be fully automatic.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

All accepted requests will correspond to reservations, each for a finite range of time. For the system management point of view, each request will have a start and an end time. Request for permanent allocation will be dealt with end time request renewal. Periodic reservation with a single user's request may be allowed according to the request characteristics.

2.1.4 Authorisation, Authentication and Accounting

Authorisation, Authentication and Accounting are considered essential for the service. The BoD service will rely on an external Authorisation and Authentication Infrastructure (AAI) to validate user requests and to also to provide security internal to the system.

The accounting functionality will be provided from the early stages of the service deployment. Its purpose is not only to check the resources' allocation, but also the service usage and performance. Part of the accounting system data will be internally collected by the BoD system itself while an external network monitoring system will provide the rest of the data.

2.1.5 Multi-domain

In case of an end-to-end path crossing various administrative domains, each with a different technology, the BoD service will be built by connecting the different BoD services in each domain to provide the end-to-end service.

In a multi-domain environment, the source domain has the responsibility to provide the service to the end user. The source domain has the role of starting the provisioning process and of providing a final response to the user within a maximum amount of time.

It is fundamental that each domain maintains the independency in defining its policies and in choosing its preferred technological implementation, whilst it must implement the standard inter-domain interface.

2.1.5.1 BoD supportive Domain

A domain along the requested BoD path may support the service through over-provisioning or similar techniques at Layer 3. In this case, the domain is not BoD capable, but may provide support to BoD requests from other domains.

2.1.6 Monitoring

Monitoring is the key to fulfil the user requirements for service performance [section 4] and needs to be provided on an end-to-end basis. To this goal the single domain, edge-to-edge, measurements need to be

concatenated to provide a full picture. The latter is a functionality that will be provided leveraging tools and services developed by JRA1.

For monitoring the BoD system itself, it is important that the basic system design provides hooks for getting important functioning information out of the system, as an example: log-files, real-time copies of messages between modules (to check protocol misbehaviour), time-stamping, etc.

This information can be, or is complementary to the accounting information of the system.

2.1.7 BoD with Packet- or Circuit-based Technologies

The BoD service can be realized at Layer 2 or 1 using dedicated circuits/channels/wavelengths or using Layer 3 technologies to transport Layer 2 frames. It can also interface directly with Layer 3 technologies for bandwidth provisioning, such as packet based QoS techniques, like Premium IP, and MPLS. The two approaches are in some way complementary and may coexist.

A BoD service based on Layer 2 circuits or Layer 1 wavelengths may be easier to deploy, as it is in practice equivalent to creating a network as the NRENs are used to do, but requires a widespread and abundant presence of unused physical resources in the network, which are not always available or too expensive to procure.

As many NREN networks are currently packet-based, Layer 3 packet-based QoS may provide a good alternative to the BoD service. Still, few NRENs offer QoS services at the IP Layer. Some rely on over provisioning, while some others are congested.

In any case, the various BoD technologies in the different layers will play a complementary role, which will change with time and will evolve according to the technology developments.

2.1.8 Integration of Technologies

The set of network technologies varies in each NREN domain, ranging from Layer 1 wavelengths to MPLS. Due to this reason, the creation of an end-to-end BoD path mandates the integration of the different technologies. As an example, a path can start using an Ethernet circuit, proceed through a SONET/SDH circuit and terminate on a domain, which uses MPLS.

The integration of technologies is therefore a key task for the deployment of the service. Integrating different technologies, in the scope of the BoD service, requires both a clear abstraction of each technology characteristic and some experiments to validate the integration procedure.

2.1.9 Choice between equivalent Transport Technologies

There can be cases in which the path finding process finds that the circuit between two endpoints can be realized using two different transport technologies. The choice can be done according to the set of parameters which fits better the user's request, as an example synchronous transmission, or which allows a better utilisation of network resources.

2.1.10 Layer 3 Routing and BoD

While BoD service can be fulfilled by Premium IP service at Layer 3, it has the limitation of following the Layer 3 routing paths. According to PIP specification, there will be no policy routing implementation, therefore features like route diversity or explicit routing requests will not be supported in the basic PIP service.

These features might be added using traffic engineering techniques enabled by the use of (G)MPLS coupled with QoS.

In cases where BoD will only use Layer 2 functionality, the circuit may not necessarily follow the IP routed path. The BoD system will not initially consider providing IP routing on top of the Layer 2 offered circuits.

Some users may wish to use BoD Layer 2 circuits to connect machines or subnets that are already connected to research networking by means of an NREN or educational institution's existing IP network. In this instance, the machines will already have IP addresses assigned by the administrators of their local network. Unless the requirement for independent (BoD) connectivity is known at the time of allocation, it is likely that these IP addresses are assigned specifically for the purpose of connecting via the local university's IP network. These users may still apply to their Local Internet Registry for globally unique IP space to be allocated for the purpose of the BoD circuit and surrounding equipment, or for them to agree on the use of RFC1918 space. In case the BoD circuits spans multiple administrative domains, coordination is needed and the most appropriate solution has to be defined on a case-by-case basis. In each of the above cases, the BoD circuit and infrastructure remains effectively a Virtual Private Network separate from the existing IP network.

IETF charter Zero Configuration Networking focused on this issue targeting to enable networking in the absence of configuration and administration. RFC 3927 describes how a host can automatically configure an interface with an IPv4 address within the 169.254/16 prefix that can be used for connecting with other devices on the same physical or logical local link. A recent successful demonstration of this technology was conducted within the framework of the iGrid 2005 event by the Advanced Internet Research Group of the University of Amsterdam [Dijkstra]. Testbed included three computers in Amsterdam and two computers in San Diego directly connected through a dedicated wavelength. Existing software implementations were used for:

- Automatic assignment of link-local IPv4 address, according to IETF RFC 3927.
- Hostname and address resolution, by using IETF's Multicast DNS [mDNS], which enables DNS-like operations on the local link without using any conventional DNS server.

- Service discovery, by using IETF's DNS Service Discovery [DNS-SD], which enables discovering a list of named instances of a particular service, using only standard DNS queries.

If a BoD user wishes to use their BoD circuit as a primary link, but also requires failover to the existing IP network, should the BoD circuit go down for some reason, then it is no longer the case that the BoD circuit is separate from the existing IP network and its routing infrastructure.

Successfully integrating new circuits at arbitrary locations into the existing IP infrastructure is out of the scope of the initial phase of the BoD architecture.

Nonetheless, it has to be dealt with soon, to allow the offering of a BoD service spanning both Layer 2 and Layer 3 and to ensure correct integration of the BoD packet service in the existing infrastructure.

Such integration requires the cooperation of multiple parties:

- The Local Internet Registry (LIR) must assign IP address space to be used by the networks connected by the BoD circuit
- The connected networks and their institutions' Computing Services departments must agree on a suitable routing protocol and implementation
- The institutions and their NRENs must agree on a suitable dynamic routing protocol and its implementation, which may include speaking BGP and allocating a globally unique AS number to the institution
- The NRENs and GEANT must agree on the exchange of the relevant IP address space and AS path using BGP.

Some users may attempt to "hack" failover by reusing existing IP addresses on the BoD circuit, with or without authorisation from their LIR. To implement this, it would require at least the use of routers at each end of the BoD circuit as opposed to using the BoD circuit as a LAN extension and some form of link failure detection, either by means of Ethernet Operation Administration and Management (OAM) (still in development by router vendors) or a dynamic routing protocol, running independently of the existing IP infrastructure.

As a conclusion, it is suggested to proactively analyse in the next step the interaction between BoD provisioning and IP routing. Rules, procedures and caveats should be defined for the user accessing BoD and configuring IP on top of it.

2.1.11 Time synchronization

The BoD system has to provide advance bandwidth reservations according at certain time scales and it is engineered as a distributed system over multiple domains. All the modules in a BoD System require then a

sufficiently good synchronisation and the same level of synchronization has to be active between the implementation in different domains.

It is foreseen that initially this synchronization requirements can be satisfied by the use of the Network Time Protocol (NTP), using as time sources hosts with Global Positioning System receivers. NTP can provide the time with a precision close to one millisecond at the European scale, which is considered good enough for the initial deployment of the service.

2.2 Design Procedure

A top down approach is followed in the framework and architecture design procedure. The main steps are:

- agreement on a BoD framework
- definition of the service and of its relationship with other services.
- identification of the service's modules and blocks inside a module
- definition of functional specification of a module and its blocks.
- definition of the communication flow between modules and blocks, including a functional interface specification.
- Validation through the analysis of sample use cases.

Some general principles have guided the design procedure for the Bandwidth on Demand service, in addition to the analysis of the requirements:

- scalability of the system
- the need to build an end to end service, even in the case of a multidomain environment
- interoperability to allow the possibility of different technical implementation
- existing standards and Open Software solutions and independence of commercial software

2.2.1 Identification and Definition of a Service

A network offers many services to its users and to itself, for internal purposes. Most of the network services exhibit a mutual dependency; the Bandwidth on Demand service needs, as an example, at least the Authorization and Authentication Infrastructure (AAI) service and the monitoring service. The instantiation of a

service as autonomous may be decided upon how many other services or components use it and on the possibility to build it independently from other services.

According to these criteria, the Bandwidth on Demand Service, Monitoring service, the AAI service, the Location Service, and the Network Management service are all identified as autonomous services. They all will benefit from the presence of the other services and use them when available.

2.2.2 Identification of the Service's Main Building Modules

Once a service has been identified and defined, this step identifies its main functional modules. It should be possible to implement a module as an autonomous entity. The modules cooperate to produce the desired service. The modules, then, should be as much as possible independent one from each other, but their number must be kept at a minimum to avoid excessive complexity and minimize inter-module communication needs.

2.2.3 Definition of Functional Specification of Modules

Once the main modules have been specified, the design procedure advances by analyzing each of them and specifying its main functions in term of blocks and their mutual dependence.

If the same function appears in more than one module, like as an example, the AA function, the same name will be used.

2.2.4 Definition of the Communication Flow

This step analyses the communication flow between modules and blocks. Mutual dependencies and use of external services, like AAI are reported. A functional interface specification in all the relevant communication channels must be drafted and flowcharts are produced.

The flow analysis is carried both for the data plane and for control plane.

2.2.5 Use Cases

The architecture will be validated through its application for few, but important use cases. The analysis must detail how the use cases requirements can be satisfied by the architecture and should look for and highlight problems as well as possible optimisations. These use cases will be used for testing the implementation and performance of the system.

3 Summary of Requirements

As a fundamental step in the definition of the BoD service, a questionnaire has been distributed to users to gather their requirements. Targeted users are currently running research projects that require high amount of bandwidth between different sites across Europe, such as research infrastructures and special networks sites (e.g. Starlight). Unfortunately, although more than a dozen of potential user groups were identified and asked for contribution, only six replied. Five answered to the majority of the questions in the BoD requirements' questionnaire and one provided only some basic information. Of those six, four are typical representatives of the GRID community, one is a successful example of implementation of a service with similar goals at those of the JRA3 BoD, and one (which provided very concise information) is a non-GRID representative.

A thorough analysis of the user groups' responses was carried out and the analysis is reported as part of the GN2 JRA3 deliverable DJ3.2.1 "BoD User and Application Survey" [DJ3.2.1]. The main results are highlighted in the following paragraphs.

The need for BoD-like services exists today for a number of different communities/projects/users groups both in Europe and worldwide. It is expected that the need for a BoD service will become even more apparent in the upcoming years, when such a service should be mature and widespread enough. A number of Grid-related projects are active at the time of writing and it is expected that they will soon be mature enough to use the advanced services of the underlying networking infrastructure.

In order to access and control the BoD service standardized interfaces have to be provided to users and applications for resource reservations and service monitoring of the end-to-end BoD provisioning path. The interfaces can be static/manual at the early stages of BoD service deployment and dynamic signalling will be the final goal. The interfaces should support more than simple reservation requests and responses. They should provide the hooks to applications and Grid middleware to constantly interact with the BoD service in order to retrieve request status, receive event notifications and query statistics of service usage.

The requirements for advance reservations and the duration of each single request have proven to be quite diverse making it difficult to create different classes based on time and frequency of BoD service use. The BoD system should be general enough to support a large and as granular as possible space of duration intervals and frequencies for the service requests.

Bandwidth requirements range from 10 Mbps to 10 Gbps. In terms of the foreseen traffic patterns, Grid applications requirements are quite diversified, ranging from few high bandwidth burst transmissions, to continuous traffic flows at low speeds, often with real-time constraints.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

Concerning the network architectures on which the multi-domain BoD service should be deployed, the analysis reveals requirements for a large variety of physical interfaces and technologies (Gigabit Ethernet, ATM, SONET, optical switches, MPLS-based overlay topologies etc.). This requirement makes the task of JRA3 quite demanding in terms of the variety of network technologies that have to be supported in order to succeed in providing end-to-end BoD services to multiple users. Moreover, the JRA3 BoD service is required to provide the mechanisms and the visualisation components for monitoring the various elements of the service such as the status of each end-to-end BoD service instance, the quality of the service provided, accurate estimates of the achieved bandwidth when compared to the agreed guarantees etc. As user-perceived quality seems to be one of the key factors for the success of a service, monitoring is also considered a key component of the BoD service, which must be available from its early stages.

Regarding the reliability for each BoD service instance, a number of priorities have been identified:

- no loss or reordering of frames
- restoration of a failing BoD service instance should be provided as an option
- possibility of confidentiality (encryption) of transferred data
- It is desirable to optimise the latency
- choice to configure the restoration time from short to longer intervals and control path restoration in case of failure from the user/application environment.

4 Definition of the Bandwidth on Demand Service

The Bandwidth on Demand service aims at providing a guaranteed capacity, connection oriented, point-to-point service between two end user points.

The BoD service has the following characteristics:

- Inter-domain: the end user points may be located in different domains.
- Capacity: the BoD service is not restricted by definition to a specific capacity value. The initial maximum capacity will be 10Gbps. This may change in the future when technologies develop further. The minimum amount of capacity that can be requested will depend on local domain policies and restrictions imposed by the technology used (e.g. SDH has a 155 Mbps granularity).
- Point to point: the BoD service provides Point-to-Point services. Point-to-Multipoint may be realised as a set of point-to-point services.
- Bi-directional: the service is a bi-directional service.
- Symmetric capacity: the provisioned amount of capacity in both directions between the end user points is equal. Asymmetry in capacity provisioning is technically viable only for some technologies and it makes much more complex the overall provisioning. For these reasons it will not be implemented.
- Symmetric paths: the BoD provides identical forward and return paths. This implies that the propagation delay in both directions is equal.
- Advance reservations: BoD services can be requested in advance. There will be a minimum time period required between the request and the actual provisioning of the service as a function of the level of automation in the process. A maximum time duration of the reservation is also foreseen. A time extension mechanism for a request ensures the possibility to have a longer service without interruption.

- Protection: the system is capable of providing none or complete of protection to the service. A full protection can be realised by creating two completely separate paths from source to destination, including the physical layer.

The initial focus of BoD is on point-to-point Ethernet services. Ethernet can be either native or encapsulated in other technologies like Sonet/SDH. The end user points will be able to access the service through 1 Gbps Ethernet or 10 Gbps Ethernet ports. The BoD service will then transparently forward the Ethernet frames between the end user points.

SDH and packet-based technologies, like MPLS with QoS, may offer additional techniques for BoD service provisioning in the future. The user may then be able to suggest the desired technology using its UNI interface. The BoD User to Service interface can be based on the still evolving OIF or IETF UNI specifications, possibly with extensions, as an example for advance reservation.

5 Architecture Definition

The definition of an end-to-end BoD service across multiple domains is a non-trivial challenge, as each domain may use different technologies and apply different policies. The creation of the end-to-end path requires concatenating the various edge-to-edge paths in each domain through the stitching of different transport technologies, in order to provide the required service to the user.

The proposed architecture, detailed in this deliverable, leverages the layered model in data networking to facilitate the decoupling of the architecture in modules and blocks. In addition, the definition of an abstract network resource representation, in order to abstract the technology specific details in a standard format, can provide a non-ambiguous, simple and common language to all the system components. The creation of “technologies proxy” modules can then be considered the natural place for the translation from the abstract language definitions to each specific technology when needed. The network abstraction language will also allow a standard network data information interface exchange with other services.

For this reason the architecture overview is followed by a section on how to abstract the representation of the network and its applicability to the BoD model.

5.1 Architecture Overview

The BoD service architecture is engineered to be applicable to a set of independent collaborative domains. A key element is then the inter-domain communication. The architecture overview starts with a description of the possible interaction of the BoD service with other, existing services in the same domain, and proceeds to an overview of each module of the system.

A single administrative domain usually activates more than one service. A sample domain where BoD is enabled can be seen in **Figure 5.1**. In the same domain, other services, like the Network Management System (NMS), are usually available. Among those services the monitoring service and the Authorisation and Authentication Infrastructure (AAI) are of central importance to enable high level services like BoD and PIP.

The BoD system needs to integrate itself with the existing services and, at the same time, the BoD system should use them to avoid duplication of effort and to the overall domain structure and management.

The BoD architecture described here assumes therefore that in the same domain at least the AAI, the NMS and the monitoring services are available. In case those services are not available, the BoD system should implement their basic functionalities.

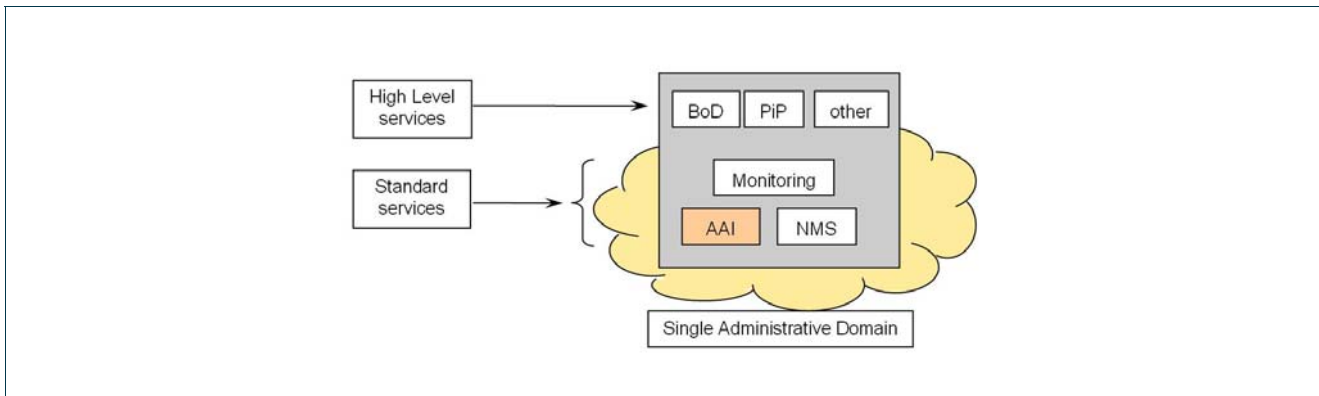


Figure 5.1: Sample domain with some services enabled

The core of the BoD system consists of the following main modules:

- Inter-domain Manager (IDM),
- Domain Manager (DM),
- Technology Proxies,
- Logic and Policies module,
- Pathfinder module.
- Information Storage System,
- Location Module

The IDM is responsible for processing each incoming BoD service request and for propagating the accepted request either locally to the DM or to another IDM in a neighbour domain. The service request is propagated along a chain of domains until the service endpoint is reached.

The main function of the DM is to setup BoD instances within the domain. In order to do that, it needs access to a detailed knowledge of the underlying topology. Based on the current resource utilization and possibly some other constraints, it communicates with the Pathfinder module to calculate the intra-domain path. Once this path is known it contacts the technology proxy to request the operative configuration of the BoD instance.

The Technology Proxies perform functions that translate the requests received in abstract language into vendor or equipment specific configurations. These modules are the components that deal with each specific technology in a BoD domain.

The Logic and Policies module contains all the rules and policies to be used by other modules when inspecting and elaborating a request.

The Pathfinder module contains the algorithms and the logic engines to search for a path that satisfies the request according to specific sets of rules and policies.

The Information Storage system and the Location Service module mainly offer support to the other modules. The Information Storage System is responsible for providing storage, archival and database functionalities to data explicitly relevant to the BoD system, while the Location module locates the address of all type of services and modules.

A summary picture of BoD system modules is shown in **Figure 5.2**. Each module corresponds to a different set of functions and internally has a different structure as described in following sub-sections.

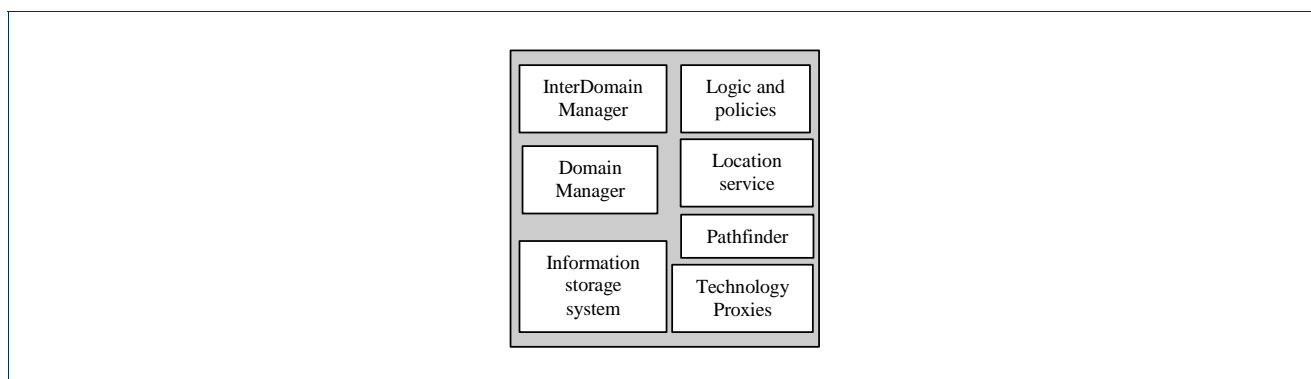


Figure 5.2: Modules of the BoD system

Error! Reference source not found. shows an example of a multidomain case in which IDMs collaborate to establish the end-to-end path. It shows how technologies may differ from one domain to the other.

5.2 Abstract Representation of the Network

An abstraction layer allows for the unambiguous definition of all the components of a seamless path between two BoD service provisioning end-points. The abstracted BoD path itself can be referenced as a single entity in all phases of BoD service provisioning, for monitoring, troubleshooting, re-routing etc.

One of the most important aspects of abstracting the entities involved in the BoD service is the ability to perform inter-domain path finding at the IDM layer, without looking into the details of domain-specific policies,

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

path properties and technologies. Using the abstract layer two IDMs can communicate in a technology neutral way.

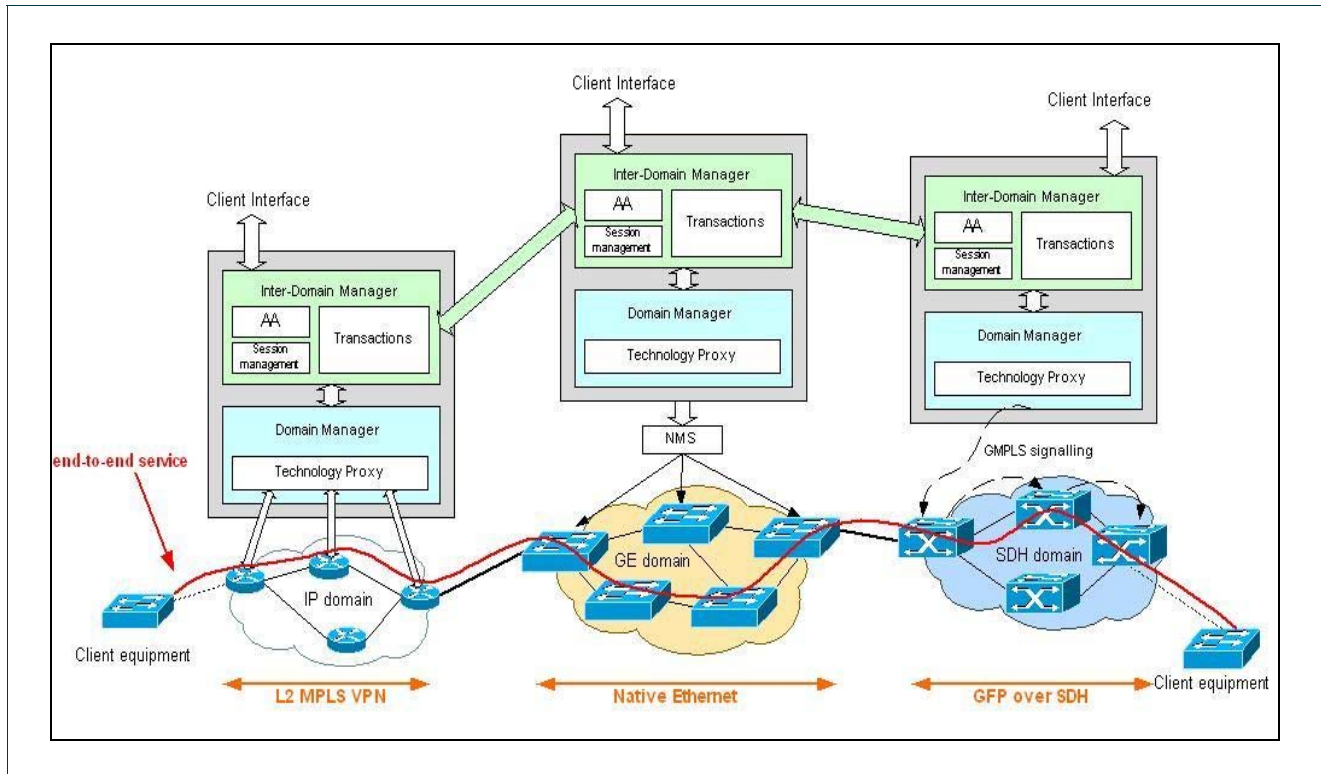


Figure 5.3: A sample multidomain BoD service case

The proposed abstraction layer is also useful for managing resources of the edge-to-edge path that are not entirely owned by a single entity, such as links interconnecting two domains. It also allows for a seamless treatment of resources within a single BoD domain where the BoD provisioning system can use more than one technology. More specifically, a methodology for the abstract representation of the available resources is an important tool for the successful deployment of a unified BoD topology database, which combines information from technology specific databases or network information systems that lie in the different technology proxies of the Domain Manager.

For the purposes of the abstract representation of network resources, it is useful to model physical and virtual resources as objects. Non-divisible, elementary or primitive objects may be combined in order to create complex objects. The aim of the BoD service in each provisioning case is to combine primitive and complex objects to create an edge-to-edge path.

The following sub-sections provide just a draft list of basic objects and their specification. The complete analysis definition of the abstract layer requires a significant effort. It will be carried on in the next phases of this work item and in collaboration with the implementation work item.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

5.2.1 Abstract Representation: Objects

5.2.1.1 Primitive Objects

This section provides the description of three proposed primitive building blocks of the BoD service abstract representation of network resources (see also **Figure 5.4**).

Node

A node represents a network element acting on one or more BoD circuit. It can, as an example, switch, groom or terminate a circuit. A node is physically associated with a network element (router, switch, OXC etc.), however it is possible that, due to some internal policies and domain-specific implementations of the DM, more than one “abstract” node maps to the same network element. In the latter case, policies can be defined and applied for the resources managed by each node. For example, access to a specific subset of interfaces of the network element belonging to the same node can be assigned to a specific group of users. A node has to be uniquely accessed and addressed by the DM within a certain domain

Port

A port is a physical and/or virtual interface on a node through which one or more dedicated capacity circuit enter or exit from a node. A port has to be uniquely accessed and addressed by the DM within a certain domain. Each port can be associated with a physical or virtual interface of the domain topology, so that a physical interface contains one or more ports.

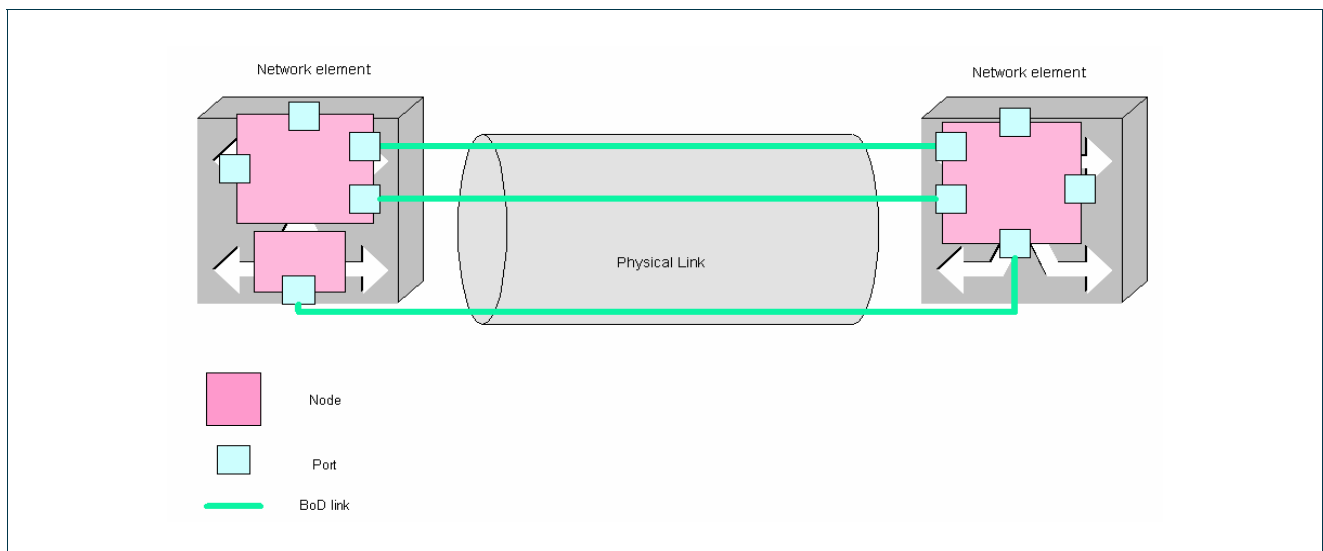


Figure 5.4: Abstract representation objects

Link

A link is a physical and/or virtual circuit between two ports that belong to two adjacent nodes of the abstract topology for a domain. Each link has to be associated with a physical circuit of the domain topology, so that one physical circuit contains one or more links.

5.2.1.2 Complex Objects

Primitive objects can be combined to produce complex objects as described in the following sections.

Hop

A Hop is defined as the combination of a node, a port and a link.

Path

A path is defined as the abstract representation of a series of Hops

Composite Link

It is a complex link comprising of more than one adjacent BoD links. The initiating port of the leftmost BoD link and the terminating port of the rightmost BoD link are denoted as the initiating and terminating ports of the composite link. For the purposes of the abstract topology representation, a composite link is represented exactly as an ordinary BoD link and is treated as such by the path finding algorithms. The composite link notion is similar to the Forwarding Adjacencies in GMPLS.

5.2.1.3 Scenarios for Usage of the Abstract Topology

The definition of an abstract topology, based on the definitions and conventions stated above, allows for a number of different use-case scenarios:

- Within a certain domain, along certain physical links, only a part of the available capacity (e.g. an SDH VC) can be set aside for use by the BoD service. Thus, the BoD links forming the abstract topology are used to represent the available resources.
- For carrying out a requested reservation, the IDM needs to obtain a list of all possible paths between two ports of the domain, regardless of the technology used to implement each one of them.
- Due to a failure or changes in the reservation parameters of a BoD service-provisioning instance, the IDM decides to re-route (part of) a path. For those segments of the path where re-routing is not under the control of the IDM (but is controlled by e.g. an underlying NMS), composite links are defined in the abstract topology

- Due to a specific policy for the allocation of available resources for BoD to certain users/user groups/projects, a subset of the ports and BoD links in a domain’s abstract topology are declared as eligible for use in order to fulfil a certain reservation request.

5.3 Interaction with other Services

Many functions needed by the BoD system are already provided and used by other services. The BoD system will try to not duplicate or recreate such functions, but rather rely on existing services if available. The following sections detail the characteristics and the functions expected from these services.

Figure 5.5 shows the interaction between BoD components and the external services and system when available.

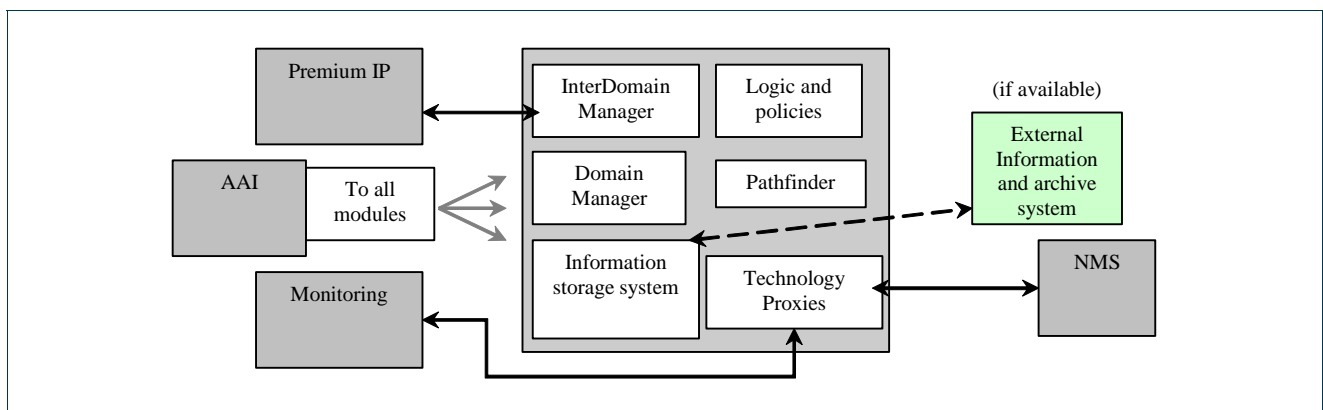


Figure 5.5: Interaction between BoD system modules and external services (when available)

5.3.1 Authentication and Authorization Infrastructure

The Authentication and Authorization Infrastructure (AAI) is a service dedicated to perform user and system request authentication and authorization, to enforce system security and to prevent unauthorized access and use of resources. A single domain requires at least one AAI service in order to process and instantiate a request.

The BoD service modules may interact with AAI multiple times during a single request execution.

After the initial authentication and authorisation check, the BoD system will apply additional, specific to BoD, rules and policies to the request.

In the case in which the AAI service is not available in a domain, a manual configuration of the AA policies between IDMs is required. As an example, a trust relationship can be agreed between two neighbouring domains. In this case, no authentication or authorization is performed on the request.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

The BoD system architecture described here assumes that an external AAI exists to authenticate and authorize access to the resources.

5.3.2 Monitoring

The BoD service needs to monitor the performance and health of the BoD software system itself and of the offered circuits. This is a key requirement to ensure that the service meets the user expectations as stated in DJ3.2.1 [DJ3.2.1].

Monitoring the internal health of the different modules and components of the system is also a key task in a distributed system and allows the identification and correction of performance bottlenecks. This task has to be performed by the BoD system itself.

Monitoring the end-to-end offered circuits, e.g. Ethernet or SDH/SONET circuits, implies a collaboration with the network monitoring system in each domain. As an example, to monitor SONET/SDH and optical circuits ITU-T [ITU-T] has defined various standards, mainly in the G (Transmission systems and media) and M (Maintenance) recommendation series (as an example [G.826] and [M.2101] and its addenda).

The monitoring data will be used for initial validation of the circuits' capacity in the set-up phase, during normal operation for performance analysis and conformance to applicable SLAs but also for debugging and accounting.

The exact monitoring procedures will be defined in collaboration with JRA1. Basic proposed circuit monitoring parameters are Availability and Bit Errors Rate (BER) of the circuit. Availability can be defined for a finite range of time, as an example, as the ratio between the minutes in which the circuit has been available and the total number of minutes in the chosen interval. In case of Sonet/SDH, the monitoring procedures are well defined in the ITU-T specification and metrics can be easily extracted from Sonet/SDH header counters. For other technologies, like Ethernet over DWDM systems, the monitoring procedures are not yet standardised, as well as the extensions needed to monitor the whole path. As the path may be composed by various technologies, the monitoring task is thus not trivial and requires additional research and development effort, which is already ongoing.

It is considered important that the monitoring data is stored in the same format as the one used by the JRA1 monitoring system to allow easy interchange and metric concatenation. This would also allow accessing the Layer 1 and 2 counters when investigating normal Layer 3 link performance issues.

5.3.3 Network Management Service

The Network Management System (NMS) is the fundamental task for network maintenance and operation. Its level of automation may vary from completely manual configuration procedures to complex automated systems.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

In the latter case, the NMS relies on Element Management Systems (EMS) that manage the specific individual network elements. These EMSs monitor the performance of the elements (devices), and provide an interface to their settings.

In the OSI model five different functions are defined for the NMS:

- Fault management: detecting and notifying faults, preferably also identifying the root cause and correcting the fault
- Configuration management: Inventory management and tracking the configuration of elements and systems
- Performance management: monitoring different performance parameters.
- Security management: managing the access to network devices, limiting access to authorised users.
- Accounting management: tracking the usage of network resources.

In practice, the NMS will use the interfaces and functions provided by the EMSs that have mostly local knowledge, and tie the different systems together to create a complete view of the network. This allows network-wide management of services. Communication between the NMS and the EMSs uses a wide variety of protocols, like CORBA, TL1, proprietary, SNMP or a CLI.

The NMS can be used to create end-to-end services on the network, monitor their state, manage faults and collect performance statistics. It may also allow the network operator to perform functions such as upgrades of software on the network elements, the management of user accounts to allow access to the network with different authority levels.

The BoD service will in several cases need to use the existing Network Management System to perform the actual BoD configurations upon a network. For this purpose an interaction interface is proposed between the technology proxies and the underlying NMS. The more complex the NMS, the simpler will the technology proxies task be. In the first stages of the BoD service activation, for domains in which an automated NMS is not available, the NMS operations can be performed by the Network Operation Centre operators (NOC) manually.

5.3.4 Premium IP

Both the Premium IP service [PIP] and the BoD service offer end-to-end assured capacity to the users. Premium IP is specifically targeted and working at the IP layer and relies on Quality of Service techniques for IP packet switching networks. Not all networks will have both services deployed, so just one or none may be present. Some networks will deploy PIP only and others will only choose to deploy BoD. In any case, the design and implementation of both services must allow for their co-existence.

A domain that implements BoD can be considered as PIP supportive, which means that a BoD domain should be able to serve and propagate a PIP request along a chain of PIP domains.

A domain that implements PIP can only provide a partial support to BoD, specifically at capacities below 1 Gigabit/s. At Gigabit capacities, PIP has not been defined and tested and it will not be able to replace the BoD system at Layer 2. It's quite important that the PIP service and the BoD service are capable of exchanging requests. For this to be possible, it is required that information flow between adjacent domains contains one or more fields to signal whether a request originates from an upstream domain implementing a different service. A BoD supportive or collaborative domain should be capable of forwarding the request parameters of the BoD service to the downstream domain in a transparent way.

Regarding the support of the BoD or PIP service across multiple domains with different capabilities, the following cases can be considered:

Case 1: A PIP domain in a chain of BoD domains (see **Figure 5.6**): The BoD request from domain 1 has to be translated to a request conforming to PIP specifications in order to be processed by domain 2. Domain 2 can simply relay the initial BoD request to domain 3 for further processing. In the figure, an indicative list of parameters for each service is provided.

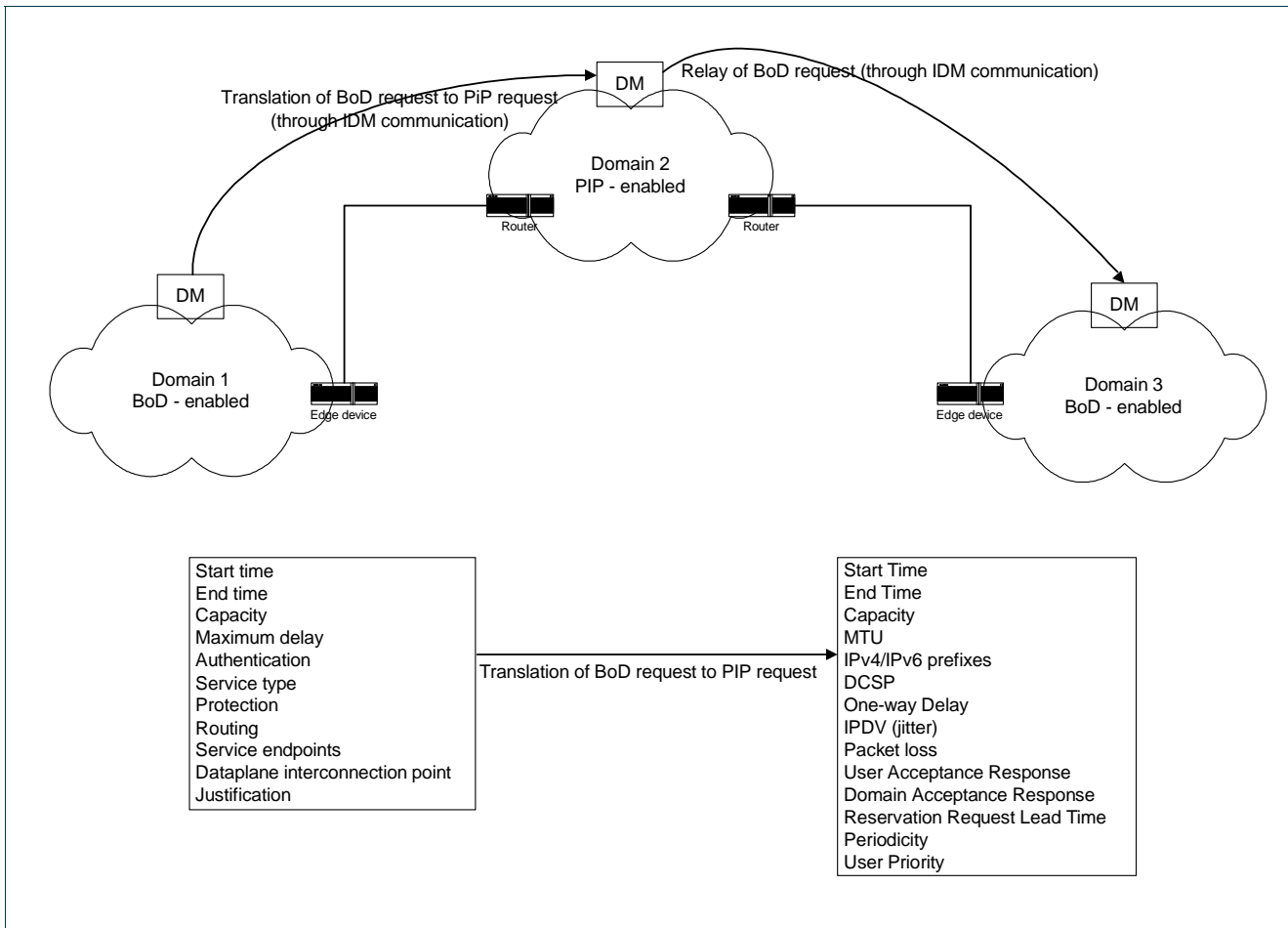


Figure 5.6: A PIP domain in a chain of BoD domains

Several parameters can be mapped directly between the two types of requests, while others are forwarded to the next domain, but not taken into account by the intermediate PIP domain. More specifically, translation of a BoD request to a PIP request can be done in the following way:

BoD parameter	PIP parameter	Comment
Start time	Start time	No change needed
End time	End time	No change needed
Capacity	Capacity	Possibly adaptation of capacity value is needed (as a function of technology used by BoD). An additional consideration is the granularity, which can be much coarser for circuit-based BoD service.
Maximum delay	One way delay	For the PIP service, delay consists of propagation delay, forwarding delay, queuing delay, transmission delay.
Authentication	Authentication	It is suggested that the AAI will be common to both services.

Service type		Relayed but not taken into account
Protection		Relayed but not taken into account
Routing		Relayed but not taken into account
Service endpoints		Relayed but not taken into account
Data plane interconnection point		Relayed but not taken into account
Justification		Relayed but not taken into account
MTU	MTU	Such a parameter will be included in the BoD interface definitions. It has to be taken into account that while a larger MTU poses no problem to PIP, an IP packet size that cannot be accommodated in the supported MTU will cause IP packets to be fragmented.
	IPv4/IPv6 prefixes	The values for the ingress/egress routers of the PIP domain
	DSCP	According to mapping in [PIP]
	IPDV (jitter)	Empty (optional parameter)
BER	Packet loss	As low as possible for PIP.
	User Acceptance Response Time	If such a parameter is not included in BoD request definition, then a set value must be mapped for this case
	Domain Acceptance Response Time	If such a parameter is not included in BoD request definition, then a set value must be mapped for this case
	Reservation Request Lead Time	If such a parameter is not included in BoD request definition, then a set value must be mapped for this case
	Periodicity	None
	User Priority	FCFS
Credit	Credit	May need mapping between the different credit unit of measure

Table 5.1: Translation of BoD requests to PiP request

Case 2: A BoD domain in a chain of PIP domains. Translating a PIP request to a BoD request in an intermediate BoD domain that lies in a PIP end-to-end path is a procedure similar to the one described for the previous case. Several parameters can be mapped directly between the two types of requests, while others are relayed to the next domain but not taken into account by the intermediate BoD domain.

5.3.5 Database and Archival Service

Within the BoD service, the database will serve two bookkeeping purposes. The first objective is to keep track of network resources, their availability and current utilization. This part of the BoD database will be referred to as 'Network Information System' (NIS). The path finding algorithms should rely on the NIS and its stored network representation. This NIS should cooperate with the NMS (if possible) or with a technology specific system through a technology proxy, which in some cases may complement or replace part of the database's functionality.

The second use of the system database is to store specific BoD system data, like reservation information, including queued, waiting for activation, in progress, and expired reservations. Interworking with the NIS is required in order to mark resources as used for current or future reservation requests.

Both objectives can be achieved with single database instance, but for scalability reasons two separated entities may be used.

The database system is also used to store all other types of information used by the BoD system, like policy and rules, monitoring data (if not stored in monitoring system) and accounting data.

If an external information archival system is available, it is important that the BoD internal information archival system and the external one cooperate, possibly through the NMS or a technology proxy. The cooperation should be defined to minimize the need for synchronisation and update between the two.

5.4 Modules and Blocks, Functionalities

The following sections provide a description of BoD functional modules and their proposed internal composition. The different modules will be described and analysed into their functional blocks.

5.4.1 Inter-domain Manager

The IDM represents the manager of the BoD system. It implements the following functionality:

- Receives and makes a high-level processing of BoD reservation requests from users or from other IDMs.
- Selects the next domain to contact to create the end-to-end path
- Participates in a commit process between all IDMs along the end-to-end path of a BoD reservation request
- Interacts with the AAI service, when available, to validate the identity of the BoD service requestor and his authorization privileges for requesting BoD services. In any case, it applies the appropriate rules and policies specific to the BoD service.

- Based on the authentication of each BoD service requestor, it implements a credit management system for the controlled allocation of bandwidth resources among the BoD users of the domain. Policies for credit charging for requests coming from other IDMs have to be defined.
- Operates the accounting system that keeps, processes and presents accounting data of the BoD service usage and availability per BoD session and in general within the domain. This functionality is particularly useful for assessing the successful deployment of the service as a whole and for signalling adjustments in resources set aside for the BoD service purposes in cases that these resources are under- or over-utilized
- Each IDM will implement its own policies and service management. Although it is suggested to have uniform rules between different administrative domains, the service must assume that the policies may be different

5.4.1.1 Inter-domain Manager Module building blocks

Figure 5.7 summarizes the building blocks and function present in the IDM.

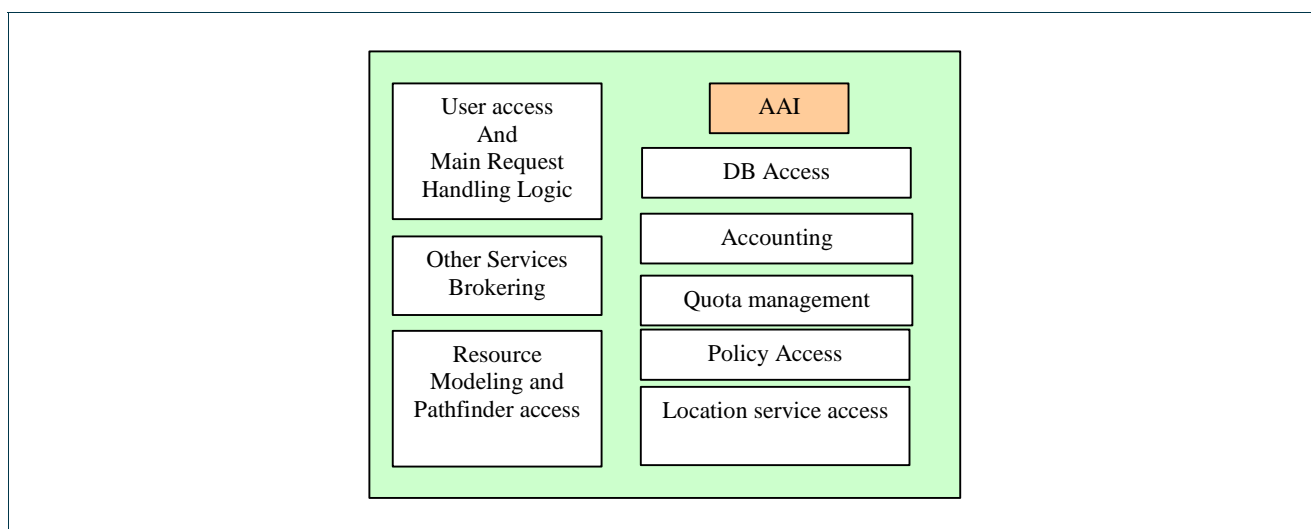


Figure 5.7: Inter-domain Manager main blocks

AAI

The AAI component of the IDM implements all functionalities for users and system components of the BoD service to be authenticated and authorized when using the service. The AAI service used will be an outcome of the GN2 JRA5 activity. The IDM should be able to exploit the AAI service to assess incoming BoD requests and all messages coming from upstream and downstream domains. The AAI component delivered by GN2 JRA5 is expected to be generic enough to apply to services other than BoD, however the IDM should deploy its own access block to it and build some custom functionality upon it.

Specific BoD authorisation will be dealt with in the user access block, through the logic and policies module. The separation is due to the need to clearly separate data managed by the external AAI system and internal authorisation rules.

In case the external AAI service is not available, all the functionalities have to be created inside the BoD system, possibly creating a specific module.

Quota Management

The Quota management will implement the functionality for managing a credit base algorithm for allocation of the BoD resources available within the domain among the authorized users. It will require communication to the logic and policies module and it will provide output for the Accounting component as well as for the DM. The latter concerns the cases when the availability of unused quota for a certain user needs to be confirmed before handing off the reservation request to the DM. When enough quota is not available for the requesting user, the request is rejected at the IDM level and a notification is sent as a response to the requestor.

Accounting

This block will perform a series of accounting activities for monitoring the whole system and the BoD resources use within the domain, such as the number of requests successfully served as well as those rejected, the average utilisation of BoD resources on the network links and for each request the effective use. It will interact both with the AAI and the Quota Management blocks for linking the service use with user identities and corresponding quotas.

It will also interact with the monitoring system to provide reports.

Resource Modelling and Pathfinder Access

When receiving a request, the IDM has to perform a series of topology checks and perform routing decisions. First, the IDM needs to discover if the starting point of the local user requested path is indeed internal to its domain. In case that the requestor is another IDM the point has to be at the correspondent edge. Afterwards it has to compute a path either completely inside its domain, or up to the domain border.

To perform these actions the pathfinder module will be called to provide a list of possible paths. The evaluation and choice of the preferred path amongst the list of paths received lies in the Logic and Policies module. Resource modelling is needed when evaluating alternatives for the BoD resource use.

In case a candidate path is received from another IDM, the IDM pathfinder can alter the candidate path for subsequent use. In a peer-to-peer mode, its task is to select the egress neighbour administrative domain.

Communication with other Services

This block contains, in a modular and extensible format, all the interfaces to communicate with other high level services like Premium IP. For each external service, the interface contains the rules to form a request, including how to convert resource representation from BoD abstract notation to the other service's specific notation when different.

User Access and Request-handling Logic

It contains the interfaces towards a generic requestor. The block can accept new interfaces definition and logic in a modular fashion. For example, it can start containing a web interface for human users and an application-programming interface, then evolve to accept request using XML. In the case the requestor is a BoD system with a different architecture (like UCLP), this block will use the appropriate proxy function to communicate.

Blocks to Access other BoD modules

The inter-module communication can be centralised in a single block or in separate blocks. The choice is more an implementation task and it will not be discussed here.

5.4.2 Domain Manager

The Domain Manager implements the following functionalities:

- Receipt of BoD reservation requests from the IDM. Each reservation request may or may not contain specific topological and technological requirements, like preferred path or technologies in addition to the original user requested parameters and additional IDM generated data (like a first indication of a path) and constraints.
- Preparation of a set of accounting data about the system itself and BoD service instances upon request from the IDM
- Processing of each BoD reservation request and replying on whether it can be fulfilled or not according to its knowledge of the network status through the technology proxies and/or the NMS.
- Use of the DM pathfinder module to compute the path through which a BoD reservation will be routed within the domain, based on the domain-specific technologies and/or policies
- Updating the BoD-specific network information and the general network status (possibly through the NMS and technology proxies) when an event occurs (as an example, the creation or end of a reservation or a fault)

5.4.2.1 Domain Manager Building Blocks

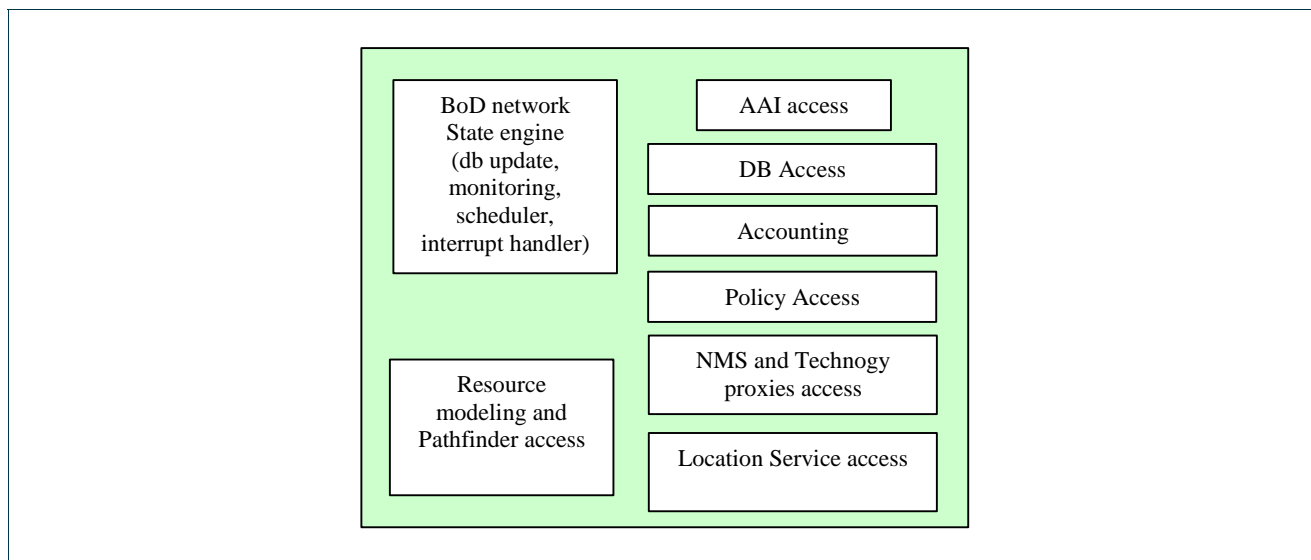


Figure 5.8: Domain Manager Building blocks

BoD Network State Engine

This block is responsible for maintaining an up-to-date status information of the BoD service for its domain. It is also responsible for handling asynchronous events (like failures), scheduling periodic tasks, starting and ending each reservation, monitoring and accounting.

This block acts also as the main I/O port for the module and keeps scheduling and queuing amongst requests.

Resource Modelling and Pathfinder Access

When receiving a request, the DM has to map a logical path to a physical path and to specific transport technology.

To perform these actions, the DM pathfinder module will be called to produce a list of paths. The logic used to evaluate the list of paths received lies in the Logic and Policies module, possibly including modelling of resource usage and comparing different solutions based on the abstract representation of elements.

NMS and Technology Proxies Access

The DM will not perform the network configuration by itself, but it will use existing services, like the NMS, or access the devices through specific technology proxies. In any case it is suggested that the interaction between the DM and the actual network entities flows through the technology proxies, as a way to provide an adaptation layer to the many different technological environments in each domain and to existing NMS systems.

AAI Access

The DM may need to authenticate and authorize the requests and command exchange with other modules. It should not need more complex AAI tasks.

Blocks to Access other Modules

The intermodule communication can be centralised in a single block or in separate blocks. The choice is more an implementation task and it will not be discussed here.

5.4.3 Logic and Policies Module

This section describes the architecture of the BoD logic and policies module. The module is responsible for providing decisions based on various set of policies and logic structures to apply them.

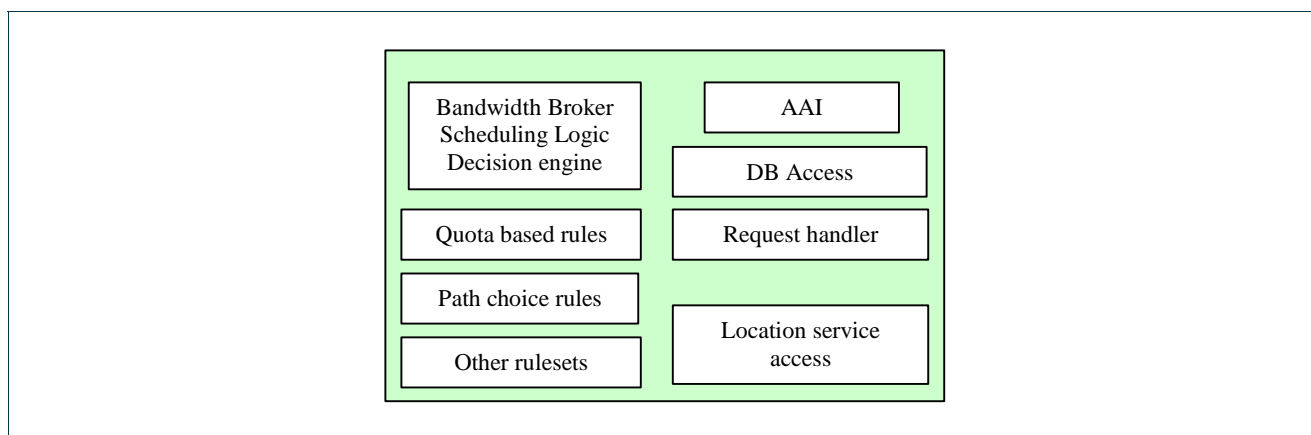


Figure 5.9: The blocks in the Logic and Policies module

The main part of the module is composed by the Bandwidth Broker (BB) block. The task of the BB is to manage and apply rules for usage of network resources for guaranteed capacity service to services requests.

The policies establish the rules for accepting a user request and starting a network reservation service. The logic and policies module uses as parameters the requestor's identity, the origin, destination and transit(s) BoD domains, in such a way to provide guarantees on capacity reservation, for example based on a given layer 2 circuit switching connection-oriented service. Its task is also to enforce the policies for the use of network resources by the introduction of credit-based mechanisms. BoD polices module would be queried at the ingress point to the BoD domain to validate the users' requests against current available policy rules and credit.

The policies module would therefore be used both by the IDM and in the DM module.. The relationship between the IDM and the DM and the Logic and Policies module is detailed further in this document.

The logic and policies module is based on following blocks parts:

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

- AAI – Authentication and Authorization for accessing BoD service;
- Credit based policies – shortly described as billing BoD user usage;
- Path choice policies – checking and approving network resource availability;
- Scheduling logic and decision engine – service reservation procedures and rules independent decision engine;
- Request handler – definition of BoD user request and procedure for handling user request;
- DB access – interaction with BoD database.

The module may also contain other sets of rules needed by the BoD system, for example for raising alarms or to create accounting reports.

5.4.3.1 AAI Access

The Authentication and Authorization Infrastructure is used to filter access to the module itself.

5.4.3.2 Credit Base Policies

The preferred method to provide resource sharing is the introduction of a service usage cost in the form of quotas. The cost is measured in credits and it is a method to quantify the rules and policies to be applied to each requestor. The cost in the BoD service has the role to prevent any misuse of the network resources provided in BoD compliant domains and to allow differential access to the BoD resources. The function of the credit-based block is to check the user quota (in BoD referred as credits). In accordance to the approach followed in the SA3 activity of GN2, credits will be charged in the form of BoD “volume” usage, i.e. the product of BoD capacity for the time it is requested. The credit base policies will be applied in co-ordination with the AAI service to account the service to a specific user and/or to a group.

5.4.3.3 Path Choice Policies

The path choice policies module provides rules to the pathfinder module, to perform path choice and define service parameters. The BoD service may provide an e2e backup path. The user may choose among network paths of equal cost, but with different characteristics e.g. delay, provided from the BoD service (see Pathfinder module for a more detailed description).

It also contains logic and policies to be used by IDM and DM to rank the paths in the list provided by the pathfinder module.

5.4.3.4 Scheduling Logic and Decision Engine

The Bandwidth broker architecture in BoD also relies on network time-stamps (which could also be implemented as a virtual reference time system). The time-stamp is important to maintain a coherent view of the BoD network both on the data and on the control plane. At the beginning it is assumed that the different components will use a simple tool like NTP to achieve the needed synchronization accuracy.

The ingress queue policy is supposed to be a First In First Out (FIFO) mechanism. Scheduling the service activation is the final step of the approval of a user request, and it includes inserting information about the network modifications and service parameters to the BoD database.

The decision engine is rule-set independent, it has the task to apply a set of rules to a set of input data and provide the output decision.

5.4.3.5 Request Handler

The Request handler module is, in simple terms, the service operational procedures, which describe how should the service handle each request and which modules should be contacted, within the BB module. The Request handler also acts as an I/O module between the user and the service.

5.4.3.6 DB Access

The service response time to a request is a function of the level of automatism achieved and of the reliability and efficiency in retrieval and management of network status information. In order to provide connectivity between different administrative domains, the BoD service will use a relational database, in which the BB has R/W access. The DB access module serves as an I/O module between the BB and other BoD service modules (pathfinder, AAI, monitoring).

5.4.4 Location Service Module

The BoD system is engineered as a distributed system in each domain and the service is provided by a collaborative effort of many domains. In such a distributed system there is a clear need for a discovery process of the different components inside the domain and the service access points between domains.

In case the Web Services technology is used to implement the intra-domain BoD system, each Web Service needs to discover the others. This functionality may initially be implemented through manual configurations, using for example a file containing URLs as the location of the different services. However, an automated approach can be taken as well by using existing web services discovery mechanisms such as UDDI [UDDI].

A discovery mechanism is also required to locate the IDM in an adjacent domain in order to exchange signalling messages along a chain of domains. Again, the location module can rely on static information, such

as a file or a database table that can be used to lookup the URL or the IP address of the other domains' IDM resource managers. In case the inter-domain discovery is dynamic there are several possibilities:

- UDDI, it may be possible to share UDDI registries between domains. It is not clear whether such an approach is scalable enough and how much coordination this requires from the two adjacent domains.
- Domain Name System (DNS) is an existing, proven and deployed mechanism to translate IP names to IP addresses and the other way around and to discover services for a certain domain (such as mail servers). As an example, appropriate "alias" resource records can be defined for each BoD Domain, or it may be possible to use SRV records [RFC 2782] to locate the BoD server in the other domain. Note that aliases provide just an IP address, while SRV records only provide an IP address and a port number while web services usually need a complete URL.
- Any other registry that a domain can update to advertise its services. The adjacent domain only needs to statically configure the location of the registry.

5.4.5 Technology Proxies Modules

The task of the technology proxies is to offer the BoD system a standard, bi-directional, access to different network technologies and Network Management Systems.

The BoD architecture foresees various proxy modules, each independent from the others, which can be modularly attached to the BoD system when needed. This solution is preferred to having just one proxy module, which contains all the specific technology blocks, as the latter is less scalable and does not allow physical decoupling of one proxy from the others.

In the BoD architecture, each technology proxy module creates an interface that abstracts a specific technology on a (technology) domain to the abstract format used by the BoD system. To achieve this, parameters that are used to specify the end-to-end BoD service and related areas, like monitoring and accounting, are mapped onto technology-specific parameters.

The proxy there is hence an abstract language translator. Each proxy contains a technology specific translator.

This means, as an example, that the parameters used in existing technology-specific UNIs - that allow the creation of connections on the domain - are mapped onto the parameters that are used to describe the BoD service in the abstract language. It could also mean that specific interfaces need to be written in a proxy that can drive the NMS to create a service.

The technology proxy also needs to provide an interface to the management of the service. It will need to translate technology-specific alarms and performance parameters to messages and values that are used by the BoD service. This allows managing the BoD service across domains in a generic way, without needing to deal with the specifics of the underlying technology. It could be that a technology does not have a direct way to

measure the parameters used to quantify the performance of the BoD service. In some cases the technology may not be able to provide a measurement such as packet losses, but it can provide a BER measurement instead. In that case the performance requirements of the end-to-end service need to be mapped onto the technology-specific measurements available by the technology proxy.

As an example of a technology proxy block, the case of a 740 Mb/s connection across a SONET/SDH network could be taken. In that case, the end-to-end connection request will - generically - be managed as a 740 Mb/s connection with specific end-to-end parameters, without referring to any technology-specific parameter (except perhaps the client-port protocol). On the domain that transports this service using SONET/SDH as the underlying technology the request could be interpreted by the technology proxy as a request for a VC4-5v circuit, and in case the network has an ASTN control plane the technology proxy could send that request through the standard UNI, and manage the service through there. This means the technology proxy translates the BoD request for a connection into a standard connection request for an SDH circuit, but will also provide all the handles such that it seamlessly fits into the end-to-end service.

It could be imagined that on a specific (administrative) domain more than one technology can provide a BoD service - each with its own technology proxy. In that case the proxies will provide identical north-bound interfaces to the other functions of the BoD service, and based on the capabilities and parameters of each technology a decision will need to be taken by the end-to-end BoD provisioning system as to what technology service is the most appropriate. Alternatively, the operator of the domain could integrate both technology proxies and provide a single interface to the BoD service.

Figure 5.10 details the blocks of each technology proxy. The key component is the Abstract representation to specific technology translator and I/O block, detailed in the following sub-section. The other blocks perform the same functions as in the other modules.

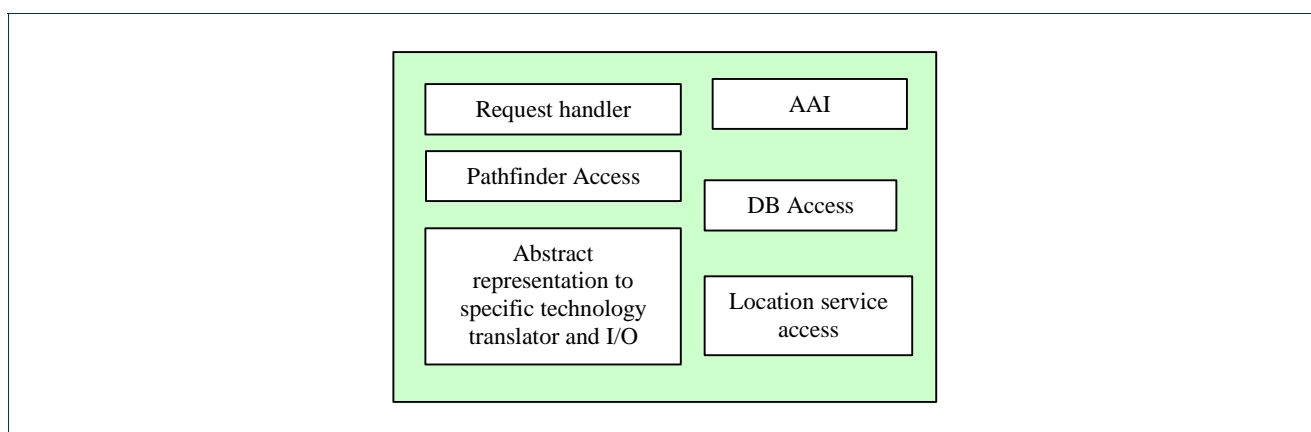


Figure 5.10: Technology Proxy blocks

5.4.5.1 *Abstract Representation to Specific Technology Translator and I/O*

The block supports the translation from the abstract network representation used by all the BoD modules to a single specific technology. The technology can range from a complex system, as an NMS, to Layer2 Ethernet switches or SDH cross-connects. Each proxy is responsible for just one technology.

The task of this block is then to map abstract language expressions to technology-specific or external system specific commands and expressions. It has to take into account the possibility that a translation may not be a simple one-to-one correspondence, as some characteristic may be absent or have a different meaning.

5.4.6 **Pathfinder Module**

The pathfinder (PF) module's task is to provide to the DM the choice of one (or a list of) paths between two service endpoints inside a single administrative domain. The pathfinder logic must be capable of finding solutions in the cases where different candidate inter-domain paths are imposing more than one candidate egress points from the domain.

5.4.6.1 *The Distributed Approach to Path-finding*

The different technologies and policies in each involved domain make a centralized computation of the precise end-to-end path for an inter-domain BoD provisioning request hard or even impossible. A centralized end-to-end path computation would for example require a detailed and up-to-date topology database of all participating administrative and technology domains. Such a database however is not easily created and is even more difficult to maintain, due to scaling issues in keeping the information up-to-date. These constraints impose the adoption of a distributed pathfinding model as opposed to a centralized one. A distributed pathfinding model will then be followed, initially focusing on the multi-domain path computation and the selection of egress and ingress points for each domain along the inter-domain path.

The path finding process acts differently at the technology domain level, at the policy domain level and at the inter-domain level. These differences suggest creating different blocks in the PF module. These PF blocks do not interact directly. The inter-domain PF, the Intra-domain PF and the technology domain PF interface respectively with the IDM, the DM, and, in some cases, with the NMS or the technology proxies.

5.4.6.2 *Constraint-based Path-finding*

The path(s) resulting by the PF blocks computations must meet the mandatory requirements defined by the service requestor, if any. These requirements constrain the set of possible solutions. The PF has to take into account both the attributes associated with resources (e.g. free capacity of a link) and the constraints in the request (e.g. required bandwidth). Solving the problem means that the PF finds a path that meets the required constraints. The following table provides a first summary of attributes of interest to the PF. The definitions are mainly based on [RFC 2702]. The abstract representation of the network will provide a complete definition of all the terms needed.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

	Attribute	Description
Link Resource	Available capacity for BoD	Amount of link capacity assigned for use to BoD
	Free Available capacity for BoD	Amount of BoD capacity currently not assigned
	Cost	Used for accounting, quotas and request admission purposes. The cumulative cost of all used resources to fulfil the request must be matched against the user's available credits
Request	QoS parameters	delay, ipdv, availability, BER
	Bandwidth or capacity	(see appendix A)
	Time parameters	Start and end time of the reservation, periodicity
	Resilience	Protection level requested for the BoD path
	Technology requested (if any)	Ethernet, SDH, other

Table 5.2: Basic PF attributes

It has to be noted that resource attributes can be grouped in resource attributes classes. Resource class attributes are administratively assigned parameters, which express some notion of "class" for resources. Resource class attributes can be viewed as "colours" assigned to resources such that resources with the same "colour" conceptually belong to the same class. Resource class attributes can be used to implement a variety of policies in a simpler form. The class is different from a composite object of the abstract representation.

In general, a resource can be assigned to more than one resource class attribute. For example, all OC-48 links in a given network may be assigned a specific resource class attribute. To a subset of OC-48 links may be assigned an additional resource class attributes in order to implement specific policies, or to abstract the network to a specific topology.

5.4.6.3 Backup and Resiliency

The possibility to provide a backup path in case of failure is considered an important feature of the BoD service. The PF must be able to compute paths with one of the following levels of protection:

- No backup

The BoD system doesn't provide any resiliency. The user has to take care about the restoration of a broken service, issuing a new request to the bandwidth broker. The BoD system has to be able to stop accounting for the usage of a broken service in case of a path failure.

- Automatic restoration

When a failure is detected, the BoD system automatically searches and builds a new pipe, if available, on the updated topology. In case a new pipe is not available, as a last fallback solution, the traffic may be redirected to use the Best Effort production service path, according to capacity availability.

- Full protection

Two totally distinct paths are built and available since the start time of service. Path separation has to be achieved creating physically disjoint paths in every hop. In this case, a domain must have more than one ingress and egress points. All resources used by the two paths are accounted to the user, irrespectively of whether a failure happened or not. The user can use both pipes since start time and, as an example, balance the traffic on them.

5.4.6.4 The PF's Input Parameters

The following table summarizes the key parameters to be used by the PF module.

Parameter	Comments
Service endpoint	Service endpoints identify the interfaces of the requested service. An endpoint can be specified as any supported Layer 1 or 2 service interface (e.g. an SDH timeslot, a dark fibre adapter on a patch-panel, a wavelength on a DWDM link, an Ethernet port, a 802.1q VLAN on an Ethernet port, and so forth)
Service start and end time	Advance reservation is a mandatory function of BoD,
Request and resource attributes and constraints	See Table 1 above.
Network topology	An abstract representation of the network devices and links falls under the PF's responsibility. The network topology must include resource attributes.

Table 5.3: Key PF parameters

5.4.6.5 Functionalities of Blocks in the Path Finder Module

The following subsections describe the blocks in the PF module, which contain PF specific functions. The access blocks take care of inter-module communication.

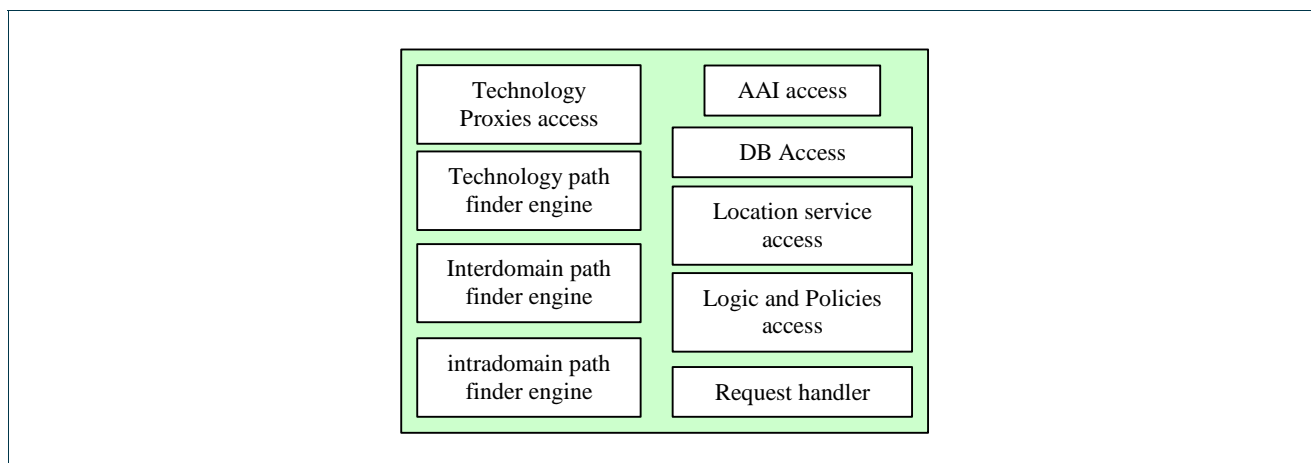


Figure 5.11: Pathfinder blocks

5.4.6.6 Inter-domain Path Finder Engine Block

The inter-domain pathfinder block is mainly used by the inter-domain manager (IDM). It is responsible for determining the list of domains along the e2e path across consecutive domains. This block is also responsible for the selection of the peering points between neighbouring domains, and more specifically for the selection of the ingress and egress interfaces through each of the involved domains . Two approaches are possible:

- Centralized path determination: the IDM of the source domain determines the complete path (as a list of domains) from the source domain to the destination domain. The initiating IDM uses the pathfinder block to contact in turn all domains relevant for the end-to-end path, according to their preference, and computes the best path (centralised model).
- Hop-by-hop path determination: the IDM of the source domain determines the preferred egress neighbour domain. The process continues on the second domain until the destination domain. All intermediate domain's PF determines just the next domain. At the end, a final end-to-end path is returned to the originating PF (distributed model).

The hop-by-hop method, based on a distributed collaborative protocol ensures the greater scalability and offers functionalities equivalent to inter-domain routing based on BGP. BGP with appropriate traffic engineering extensions may be used, as an example, as a robust protocol for the implementation of this functionality.

5.4.6.7 Intra-domain Path Finder Engine Block

This block is mainly used by the domain manager (DM). It is responsible for finding a detailed path inside a single administrative domain. Its operation is based on the abstract topology database containing the technology agnostic data of all BoD enabled network resources of an administrative domain. It is responsible for providing a list of alternative paths for each reservation request.

- from edge to edge (the ingress and egress interfaces provided by the inter-domain PF) in case of a transit domain
- to or from the domain edge to a destination internal to the domain.

It is also responsible for the determination of the path between the ingress and the egress provided by the IDM and of the connections between technology specific domains.

5.4.6.8 *Technology Path Finder Engine Block*

The block has the task to apply technology specific path finding algorithms. It's possible, that the network management system (NMS) already implements this function. If it does not, it must be implemented.

5.4.7 Information Storage System

The Information Storage System needs to provide the following functionalities:

- access to inter-domain network topology.
- access to detailed network topology at the local, intra-domain level,
- information about local network resources, their utilization and availability, as a function of current network state, scheduled and in-progress reservations,
- status of all reservations submitted to the service.
- access and storage of accounting and monitoring data
- support for storage and retrieval of any data relevant to the BoD system.

These functionalities may be provided directly by the module, or the module can act as a proxy towards external Storage and Information Systems, for example the DB managed by an NMS or the Monitoring system.

5.4.7.1 *Database Building Blocks*

To provide the reservation service with expected functionality, the database is decomposed in several building blocks. The interaction between the blocks is in scope of the database module internally, and access to the functionality is available through a Communication Interface.

Figure 5.12 presents the module construction.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

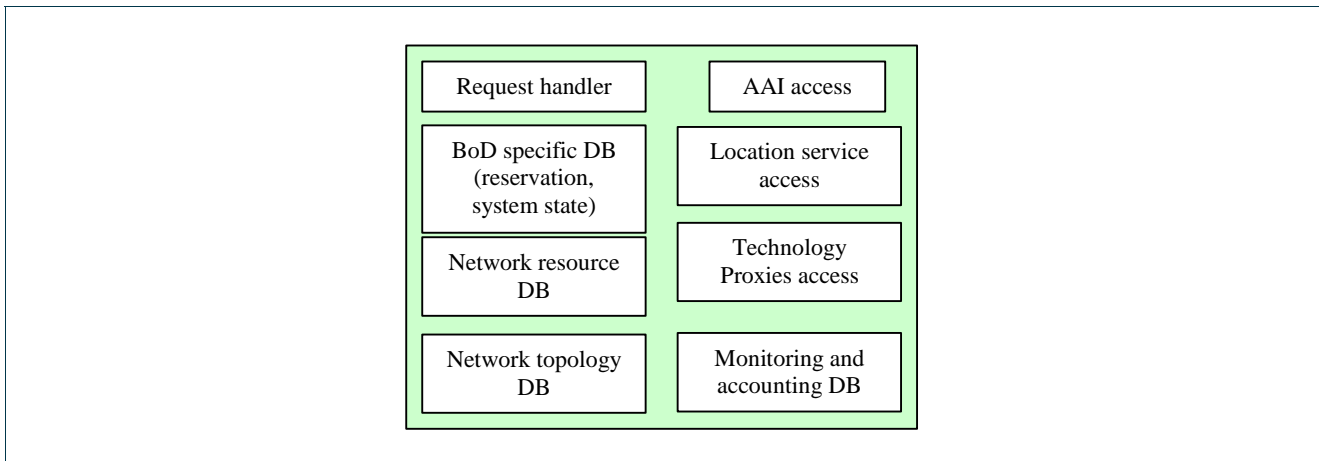


Figure 5.12: Information Storage System

5.4.7.2 Request Handler

The communication Interface of the request handler provides access to all database resources for the IDM, DM and Path Finder modules. The initial list of operations include requests to the database for:

- inter-domain network topology view – a view of network interconnections with neighbour domains, with only details about inter-domain links and no information about local domains topology,
- intra-domain network topology view – provides view of local network topology in the current domain, including network devices, technology specific parameters and detailed connection descriptions,
- resource availability and utilization on a specific link – provides detailed information about link status, utilization and free resources, that can be used for request realization,
- reservation status view and modification – provides an access to manipulate reservation requests including state modification, resource association, cancelling and scheduling options.
- Accounting and monitoring data – provides access to system and service usage statistic and to archival and retrieval function for monitoring and accounting data

5.4.7.3 Network Topology Database

The Network Topology Database stores two types of topology information:

- Inter-domain,
- Intra-domain.

The inter-domain topology data contains information about all neighbour domains. It may contain also cached data of the results of previous request results, which can provide hints on non-neighbouring remote domains and interconnection capabilities. The information about inter-domain connections contains at least the technology type and total capacity. This database will be used for the inter-domain path finding functionality.

The intra-domain topology data has to contain fully detailed information about the local domain topology, possibly in an abstract format. The purpose of this sub-module is to provide topology view for intra-domain path finding functionality, so all that information about the various network components (like equipment type, brand and capabilities, link technology type, link capacity should be available.

5.4.7.4 Network Resource Database

The Network Resource Database is an extra layer of information, which provides detailed status of current resource utilization. This information is also used by an intra-domain pathfinder to search for available reservation paths. The Network Resource Database needs to be updated upon every change of the network topology and configuration in order to represent a real-time, up-to-date network status. Cooperation between the Network Topology Database, the Network Resource Database and the NMS (when available) should be used to improve information quality and coherence and validate the network state.

5.4.7.5 BoD specific Database

This block will deal with information which is specific to the BoD system and which is not available or could not be available in external databases. In particular, the Reservation part of the Database objective has to keep track of reservations submitted to the system by local users, or from neighbour domains. The reservation status changes several times during request execution process, therefore updating entries in the database is extremely important for proper system behaviour. The request should be stored in such form to allow an easy retrieval of the following data:

- request owner – the user that has submitted the request to the system, either a human being or an application or middleware or a neighbour domain,
- request parameters – a group of general parameters associated with the request,
- reservation parameters – a group of parameters associated with specified reservation between two end points,
- reservation and request status – statuses of all reservations submitted within a specified request, and the status of the request itself, which indicates its execution progress,
- resources consumed by reservation – references to Network Topology Database and Network Resources Database must be provided instead of direct resources' descriptions, the cost of the reservation can also appear here.

5.5 Connection Diagrams and Data Flow

Appendix D reports an indicative simple test case of a reservation request and the flow of actions which need to happen in the BoD system. More cases need to be developed during the implementation phase.

6 Conclusions

This document elaborates the framework and the high level architecture of a BoD service, which has to be deployed between different administrative domains.

The architecture has been designed to scale to a large set of domains and to be capable of cooperation and integration with other services. The service implementation will initially be based on manual procedures. The level of automation will gradually increase up to near real-time provisioning. The architecture and its implementation will proceed using a feedback from the experience, starting from manual provisioning.

During the architecture engineering process, it has been noted that abstracting the network and the services in a formal representation, would provide a simpler, easier to create and more powerful service. A key task in the next step is thus to study the feasibility, devise and create a formal network abstraction and a set of formal services definition. The effort would be beneficial not only to the BoD service, but in general to create a multi-service network and to have a simple inter-domain communication. This document sketches the abstraction schema, which needs to be formalised, as an example using the XML language. To ensure the interoperability with other BoD services, the cooperation with other project is essential, possibly ending in a standardisation process. Initial discussions are already going on with the BRUW (Internet2), OSCARS (ESnet) and MUPBED (EC) projects.

The definition of the User-to-Network (UNI) and Network-to-Network (NNI) interfaces will follow the abstract definition as the next mandatory steps before service implementation. The specification will take into account the existing standards (e.g. OIF or IETF UNI) although an extension might be need, in particular to the NNI case. Once these interfaces have been defined they can initially be used in a manual provisioning process. The BoD service development will then automate the UNI and NNI interfaces and, in phased approach, the different modules and blocks that are required to provision BoD services internally within each domain. The final goal is a fully automated provisioning process.

All the steps that lead to this final goal and the experience gained will be reported on in future deliverables.

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

Appendix A References

- [DJ3.2.1] GEANT2 Deliverable DJ.3.2.1: BoD User and Application Survey.
- [DJ3.2.2] GEANT2 Deliverable DJ.3.2.2: Initial review of BoD related technologies.
- [Dijkstra] F. Dijkstra, J. van der Ham, C. de Laat "Using Zero Configuration Technology for IP addressing in Optical Networks", submitted for Future Generation Computer Systems, Feature topic iGrid 2005, October 2005.
- [DNS-SD] S. Cheshire, M. Krochmal, "DNS-Based Service Discovery", draft-cheshire-dnsext-dns-sd, Work in Progress, June 2005
- [G.821] ITU-T Rec. G.821, Error performance parameters and objectives for international, constant bit rate digital paths at or above the primary rate.
- [GN1-D.27] A. Patil, M. Enrico, M. Büchli, M. Karapandzic, "GN1, Deliverable 27: Resource Allocation and Reservation", 11 May 2005, <http://www.geant.net/upload/pdf/GEA-05-007v3.pdf>
- [M.2101] ITU-T Rec. M2101.1, Performance limits for bringing into service and maintenance of international SDH paths and multiplex sections.
- [mDNS] S. Cheshire, M. Krochmal, "Multicast DNS", draft-cheshire-dnsext-multicastdns, Work in Progress, June 2005.
- [PiP] Sevasti et. al, "GN-2 SA3 Deliverable D.S.3.9.1: Policy for allocation of Premium IP"
- M. Campanella, "SEQUIN D2.1 - Addendum 1 Implementation architecture specification for the Premium IP service", 2002
- A. Patil, T. Rodwell, GN2: Premium IP Provisioning System - Design Specification, December 2004
- M. Campanella, R. Sabatino, 'SEQUIN D4.2 QoS Implementation Plan', 31 May 2002

- [RFC 2119] Key words for use in RFCs to Indicate Requirement Levels. S. Bradner. March 1997.
- [RFC 2330] Framework for IP Performance Metrics. V. Paxson, G. Almes, J. Mahdavi, M. Mathis. May 1998.
- [RFC 2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M. and J. McManus, "Requirements for Traffic Engineering Over MPLS," RFC 2702, September 1999.
- [RFC 2782] A. Gulbrandsen, P. Vixie, L. Esibov "A DNS RR for specifying the location of services (DNS SRV)", RFC2782, February 2000
- [Saltzer84] J.H. Saltzer, D.P. Reed and D.D. Clark, "End-to-End arguments in system design", ACM Transactions in Computer Systems 2, 4 Nov, 1984, pp. 277-288, Scribe/FinalWord source: <http://web.mit.edu/Saltzer/www/publications/>
- [UCLP] <http://www.canarie.ca/canet4/uclp/> and its description in GN2 DJ3.2.2 "Initial review of BoD related technologies"
- [UDDI] <http://www.uddi.org>

Appendix B Acronyms

AAA	Authorization, Authentication and Accounting
AAI	Authorization and Authentication Infrastructure
ASTN	Automatic Switched Transport Network
BB	Bandwidth Broker
BGP	Border Gateway Protocol, the widely used inter-domain routing protocol specified by IETF
BoD	Bandwidth on Demand
CLI	Command Line Interface
CORBA	Common Object Remote Broker Architecture
DB	Database
DM	Domain Manager
DNS	Domain Name System
DSCP	Differentiated Services Code Point. - A Field In IP Header as defined In [RFC 2474]
e2e	End to End
EF	Expedited Forwarding (Differentiated Services Class, DSCP code 46)
EMS	Element Management Systems
GMPLS	Generalized Multi Protocol Label Switching

GPS	Global Positioning System – a satellite based surface system, which transmits very precise absolute time information (stratum 1)
IDM	Inter Domain Manager
IETF	Internet Engineering Task Force
IPDV	IP Packet Delay Variation
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
LSP	Label Switched Path
MTU	Maximum Transfer Unit
NMS	Network Management System
NOC	Network Operation Centre
NNI	Network to Network Interface
NTP	Network Time Protocol
NREN	National Research and Education Network
OIF	Optical Internetworking Forum
PF	Path Finder (module)
PIP	Premium IP Service
QoS	Quality of Service
RFC	Request for Comments
SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SLS	Service Level Specification
SNMP	Simple Network Management Protocol
SONET	Synchronous Optical Network

TL1	Transaction Language 1 – a widely used management protocol in telecommunications.
UCLP	User Controlled Light Paths
UDDI	Universal Description, Discovery and Integration protocol
UNI	User to Network Interface

Appendix C Terminology

The appendix contains the definitions of the key terms used in the document. The definitions follow current standards and best practice documents produced by IETF and ITU-T, in particular the definition in the RFC produced by the IETF IPPM working group [IPPM].

It's important to agree on clear definitions so that the technical description and the service level agreements presents no ambiguity. For example the bandwidth that the user requests is intuitively requested at the application layer and does not include the network overheads, which can constitute a substantial overhead and consume up to 10 % of "raw" capacity on a link.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALLNOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

C.1 Basic Glossary and Definitions

Term	Definition
Circuit	The path between two nodes of a network reserved for data communication. The term can refer to a physical (Layer 1 or 2) or a virtual (at any layer) path. Usually, a capacity is assigned to a circuit with or without guarantees, varying from the equivalent of a leased line to a standard Best Effort Virtual Circuit
Metric	The terms Metric and Characteristic are used interchangeably in this document. From IETF RFC2330 "Quantities related to the performance and reliability of the Internet that we'd like to know the value of. When such a quantity is carefully specified, we term the quantity a metric."
Node	The terms Host and Node are used interchangeably in this document.
Link	A single connection between two hosts.
Hop	A Link and a Node together form a hop
Path	A sequence of the form <h0, l1, h1,.....ln, hn> where n>=0 and 'hi'

	is a host or a node and 'li' is a link [RFC 2330]
Control Plane	The framework supporting the control & management of device operations in a network domain. It includes a signalling plane for handling calls, the synchronization & orchestration of events, the application of policies, the management of the topology management etc.
Data Plane	The components and functionality that support the core device operations (reception, processing and transmission of data) in a network element. It includes all the operations occurring real-time on a 'data path' (queuing/scheduling, data transmission, classification, data parsing, media access control, packet encapsulation/de-capsulation, packet fragmentation/reassembly)

Table C.1: Basic terms

C.2 Domain

The generic term “domain” is used with different meaning. In the text an adjective will be attached to the term to make its meaning explicit.

Administrative, single policy domain: e.g. GÉANT2. This is the default case

Service domain: a domain, which provides the BoD service. One or more subsets of the administrative/policy domain, providing a single BoD service using multiple technologies (e.g. layer-2 Ethernet framed, with Ethernet Over MPLS and VLANs or Sonet/SDH channels)

Single technology domain: a set of nodes and links participating to the BoD service, featuring a specific type of transport technology and operating systems/brand/type.

C.3 Layer

The framework will adopt a layered model of a network. Although the ISO/OSI model and the generalized Internet model with 5 layers provide a good specification of layer boundaries, their interface and offered services, some technologies, like SDH do not fit precisely in the models. A summary of layers definition follows in accordance with the standard Internet definitions:

- Layer 5 is the application layer.
- Layer 4 is the transport layer. It offers transport service at the application layers. Common transport protocols are TCP and UDP.
- Layer 3 corresponds to the network domain. It takes care of routing packets and circuits.

- Layer 2 corresponds to the data link. For the BoD framework both SDH and Ethernet are considered part of this layer.
- Layer 1 corresponds to the physical media of a network, like wavelengths and bit stream encoding on wire.

Cabling is not considered part of a specific layer although its characteristics must be part of the complete specification of a specific junction.

C.4 Bandwidth

The maximum data transfer rate of information (bits/second) at the application layer that can be transmitted along an end-to-end path. For the BoD service is defined for the user as the maximum amount of bits/second that an application sees when using IP packets as large as the maximum MTU that the end-to-end path allows, using no IP options, TCP with only the timestamp option and Ethernet as data link technology in the users' host.

C.5 Capacity

Capacity is defined at layer 2 and layer 3.

At Layer 3 it is defined for a link as the maximum amount of bit/second of IP traffic, including the IP header that a link can carry when there is no other traffic, using the maximum MTU.

At Layer 2 it is defined as the maximum amount of bit/second of L2 payload for that link, which means L2 headers are excluded.

In case of an end-to-end circuit it is defined both at Layer 2 and 3 as the capacity of the link with the minimum value of the capacity, which corresponds at the bottleneck of the circuit.

C.5.1 Bandwidth and Capacity Sub-categories

Some additional terminology will be used in the document related to capacity and bandwidth:

- Available Bandwidth: at Layer 4 it is the amount of bandwidth in the circuit not currently used by any application
- Available Capacity: at Layer 2 and 3 it is the amount of capacity currently not used
- Used Bandwidth: The bandwidth the applications are currently using on the circuit.

- Used Capacity: The amount of capacity currently used on the circuit.
- Assured bandwidth: the amount of bandwidth assigned to the user with specific guarantees, similar to ATM constant bit rate service or Frame Relay circuit specification using Peak Information rate, Committed Information Rate and Burst Size.

At layer 2, the capacity is a constant and it is usually computed in term of bits for each second. The definition is unique and corresponds to the offered value. When computing these values at Layer 3, the averaging time interval is a parameter that must be agreed and reported to use correctly the definition.

As an example, a site may use a constant 100 Mbps during the day (between 8:00 AM and 8:00 PM) at Layer 3 and nothing in the rest of the day. If the capacity offered were computed as a daily average, this would imply that 50 Mbps are the average value. In presenting traffic statistics, 5 minutes (300 seconds) are a common averaging period, but it may be too long to correctly display bursts. The monitoring system and the SLA must then take into account the “averaging” effect for a proper use of the data.

C.5.2 Overhead

User (application) data is sent usually in the form of a byte stream through the network using encapsulation. Each layer wraps the data received from the layer above with layer-specific information using a header and in some cases a trailer. This leads to an increase in the overall amount of data to be transferred. For example the use of IP and TCP adds a minimum of 40 bytes of overhead to each segment. The maximum amount of information, which can be sent in a single layer 2 frame is a function of the technology, used and is referred as the maximum transfer unit size (MTU). Data link layer technologies in turn add their specific header and trailer.

The maximum amount of user information, which can be transferred per unit time in case of no errors or losses, varies significantly depending on the network setup including TCP stacks at the end-nodes. Accurate information on the technologies used at every layer in each hop and in the end nodes is necessary to compute the payload size available to the user byte stream and hence, the maximum bandwidth available to the application for a single flow on a single link.

The computation of actual overhead on a long multi-hop path is possible only when complete information on each hop is available. The bandwidth available as payload at the data link layer on a single link should then be seen as an upper bound.

Capacity varies from one layer to another as a function of the payload. As an example, on an Ethernet link, in addition to the TCP and IP overhead, 26 bytes containing Ethernet specific information will have to be added. A minimum of 66 bytes is needed to transmit 1500 user bytes, which is an overhead of 4.2 percent. The overhead on a link using PPP over SONET is different.

C.5.3 On Demand

In the context of the BoD service, it indicates that a request is initiated by a user to obtain guaranteed capacity in the near or distant future. Advance reservation is thus considered within scope of the service. Exact time boundaries and limits will be defined by the implementation technology.

It is possible that a request obtains a negative response. In any case the Service should ensure that in a bounded amount of time the user receives a response.

C.5.4 Technology

As a general definition it refers to the manner of accomplishing a task using technical processes, methods, or knowledge. In the context of this document it refers explicitly to a technique, which is able to provide a BoD service as defined here. Examples are Ethernet over MPLS plus QoS, SDH, Ethernet, Generic Framing Encapsulation.

Appendix D Connection Diagrams and Data Flow

This section analyses a simple case in which a user submits a reservation. More complex cases will be developed during implementation.

To make a reservation, the user is required to submit a request to the BoD reservation service. The BoD system will perform several actions prior to establishing reservation paths and controlling them. The following sequence chart shows the step-by-step request processing by system components.

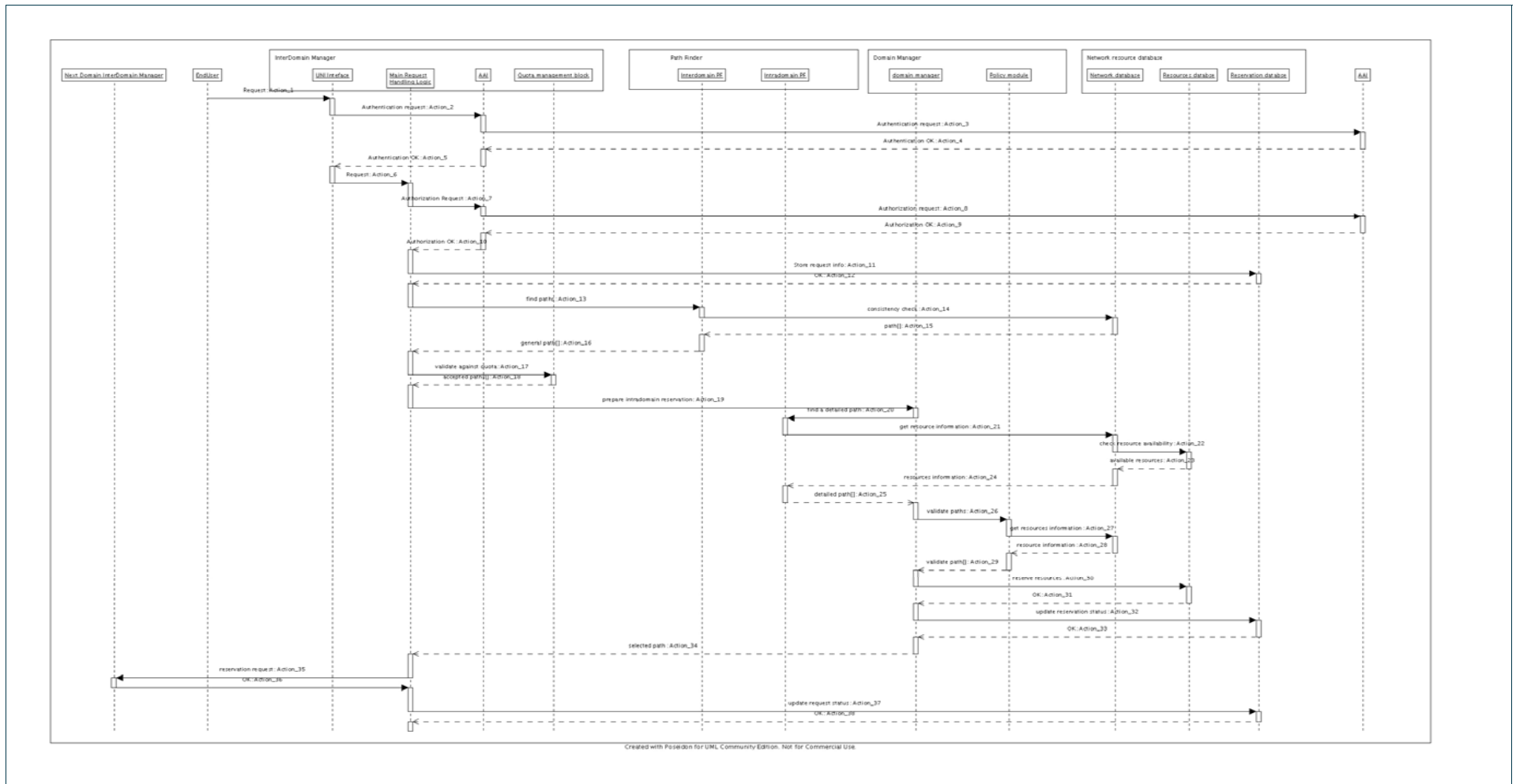


Figure D.1: Regular BoD service reservation process

Project:	GN2
Deliverable Number:	DJ.3.3.1
Date of Issue:	20/12/05
EC Contract No.:	511082
Document Code:	GN2-05-208v7

Use case UC1: Regular Reservation Process

Primary actor: User

Characters and Interests:

User: requests to perform a reservation between two end points placed in two different domains, waits for a response from the system about the success or failure of his request and –in the former case- the request scheduling data.

Domain Service Administrator: wants to have a view of system status and changes, needs to have user credits updated each time a reservation is performed

AAI: needs to receive AA requests in correct format.

Pre-conditions: User is granted access to his home domain's BoD service. The BoD system at the home domain and in the required domains along the end-to-end path are up and running. The start point of the reservation belongs to the user's home domain.

Post-conditions (in case of reservation success): The reservation is scheduled; resources are reserved in the system and cannot be used for other reservations at the same time. The user can check his request status and must receive information about the successful reservation. The user credit is decreased.

Main success scenario (Basic Flow):

1. The user sends a service request (for example through a service web portal) to the IDM in its home domain.
2. The I/O block of the IDM receives the request and forwards authentication data to AAI (IDM) block.
3. The AAI (IDM) block forwards authentication request to external AAI service.
4. The external AAI service replies with a successful user authentication to the AAI (IDM) block.
5. The AAI (IDM) block confirms user authentication to I/O (IDM) block.
6. User request is forwarded to Main Request Handling Logic (IDM) block.
7. Main Request Handling Logic (IDM) asks for authorization the AAI (IDM) block.
8. AAI (IDM) block forwards the authorization request to external AAI service.
9. External AAI service confirms access to service and responds to AAI (IDM) block.
10. AAI (IDM) block confirms access to services and responds to Main Request Handling Logic (IDM) block.

11. Main Request Handling Logic (IDM) block sends request information to Reservation Database (Network Resource Database).
12. Reservation Database (Network Resource Database) confirms successfully saved data to Main Request Handling Logic (IDM).
13. Main Request Handling Logic (IDM) requests for possible inter-domain reservation paths from Inter-domain Path Finder (Path Finder).
14. Inter-domain Path Finder (Path Finder) queries Network Database (Network Resource Database) for possible inter-domain reservation paths.
15. Network Database (Network Resource Database) replies with a number of possible reservation paths to Inter-domain Path Finder (Path Finder).
16. Inter-domain Path Finder (Path Finder) forwards the selected paths to Main Request Handling Logic (IDM).
17. Main Request Handling Logic (IDM) chooses the most suitable inter-domain path, and requests Quota Management Block (IDM) to check User request against User quota.
18. Quota Management Block (IDM) updates User quota, and responds to Main Request Handling Logic (IDM).
19. Main Request Handling Logic (IDM) requests the Domain Manager (DM) for intra-domain path reservation between domain ingress and egress points.
20. Domain Manager (DM) requests for possible intra-domain paths from Intra-domain Pathfinder (Path Finder).
21. Intra-domain Pathfinder (Path Finder) queries Network Database (Network Resource Database) for possible intra-domain reservation paths.
22. Network Database (Network Resource Database) queries the Resource Database (Network Resource Database) for available resources on the selected group of intra-domain reservation paths.
23. Resource Database (Network Resource Database) replies to Network Database (Network Resources Database) with resource information on selected group of intra-domain reservation paths
24. Network Database (Network Resources Database) responds to Intra-domain Path Finder (Path Finder) with the data necessary to produce the reservation paths that can be used to fulfil the reservation request.

25. Intra-domain Path Finder (Path Finder) forwards the collection of reservation paths that can be used to fulfil reservation request to Domain Manager (DM).
26. Domain Manager (DM) requests Policy Module (DM) to validate collection of reservation paths against domain and general policy.
27. Policy Module (DM) queries the Network Database (Network Resources Database) for resource information, regarding resources used for collection of reservation paths.
28. Network Database (Network Resources Database) responds with queried information to Policy Module (DM).
29. Policy Module (DM) decreases the number of possible reservation paths regarding the domain and general policy. The collection of approved paths is forwarded to Domain Manager (DM).
30. Domain Manager (DM) updates Resources Database (Network Resources Database) with schedule of resource usage (regarding previous path finding and policy operations).
31. Resources Database (Network Resources Database) confirms success of operation to Domain Manager (DM).
32. Domain Manager (DM) updates status of the user request in Reservation Database (Network Resources Database).
33. Reservation Database (Network resources Database) confirms success of operation to Domain Manager (DM).
34. Domain Manager (DM) reports success of intra-domain reservation and forwards chosen reservation path to Main Request Handling Logic (IDM).
35. Main Request Handling Logic (IDM) forwards the User request (updated with required local domain reservation data) to neighbour domain reservation system (neighbour IDM), which was chosen as the next hop along the e2e path.
36. Neighbour IDM performs same operations as local IDM (step 1 to 34, except that the IDM instead of the user is now the request owner) and reports success to local Main Request Handling Logic (IDM) block.
37. The Main Request Handling Logic (IDM) block updates a reservation status in Reservation Database (Network Resources Database).
38. The Reservation Database (Network Resources Database) confirms success of operation to Main Request Handling Logic (IDM) block.
39. From this point on, User request about reservation status reports a reservation success to User.

If the flow at any time fails, then:

1. Request processing is stopped. The Main Request Handling Logic (IDM) block needs to get the Reservation Status from the Reservation Database (Network Resources Database) to check if the current request was part of a request for multiple reservations.
2. The Reservation Database (Network Resources Database) returns reservation data to the Main Request Handling Logic (IDM) block.
3. The Main Request Handling Logic (IDM) block cancels each successful reservation of the current batch of requests updating the Resource Database and the Reservation Database (Network Resources Database).
4. When the user requests for the reservation status the system will report the reservation failure.