INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team reso

Optimized protocols and software for high performance networks

Rhône-Alpes

THEME NUM

Table of contents

1. Team

Head of project-team

Pascale Vicat-Blanc Primet [Maître de conférences Ecole Centrale de Lyon, HDR]

Administrative Assistant

Isabelle Antunes Pera [ENS]
Sylvie Boyer [INRIA]

Staff member INRIA

Laurent Lefèvre [Chargé de Recherches 1ère classe INRIA]

Staff member Université Claude Bernard Lyon1 (UCB)

CongDuc Pham [Maître de conférences, HDR]
Olivier Glück [Maître de conférences]

Project technical staff

Jean-Christophe Mignot [Permanent Engineer CNRS]
Fayçal Bouhafs [Temporary Engineer INRIA - CDD - projet RNTL e-Toile - up to 1/3/2004]
Fabien Chanussot [Temporary Engineer INRIA CDD - projet RNTL e-Toile]
François Echantillac [Temporary Engineer INRIA CDD -projet RNRT VTHD++]
Stéphane D'Alu [Temporary Engineer, co-Remap ENS CDD - projet Grid5000]

Postdoctoral position

Jingdi Zheng [INRIA Postdoc, 2004-2005]

Ph. D. students

Matthieu Goutelle [PhD student - 2003/2004 - MENRT]
Marc Herbert [PhD student - 2001/2004 - CIFRE SUN]
Eric Lemoine [PhD student - 2001/2004 - CIFRE SUN]
Julien Laganier [PhD student - 2002/2005 - CIFRE SUN]
Dino Lopez Pacheco [PhD student - 2004/2007 - Mexican Goverment Grant]
Brice Goglin [PhD student - 2002/2005 - BDI CNRS]
Antoine Vernois [co-Remap and IBCP - PhD student - 2002/2005 - MENRT ACI GRID]

Student internship

Jean-Paul Corvo [DEA DIF Student 1/3/04 - 15/7/04]
Dino-Martin Lopez-Pacheco [DEA DIF Student 1/3/04 - 15/7/04]
Pablo Neira Ayuso [INSA Student 1/3/04 - 15/7/04]
Sylvain Dattrino [Maitrise Informatique 1/12/2003-29/2/2004]
Grégoire Locqueneux [Maitrise Informatique 1/12/2003-29/2/2004]
Aweni Saroukou [Maitrise Informatique 1/12/2003-29/2/2004]
Cyril Otal [Ecole Centrale 1/4/2004-31/9/2004]

External collaborator

Loic Prylli [Myricom]

Long term visiting scientists

Lakhdar Derdouri [Visitor from University of Constantine, Algeria, 27/11/2004 - 26/12/2004]
Dieter Kranzlmuller [Visitor from University of Linz, Austria, 30/9/2004- 22/10/2004]

2. Overall Objectives

RESO is focusing on communication software, services and protocols in the context of high performance short and long distance networking and applying its results to the domain of Grids.

2.1.1. Project-team presentation overview

The INRIA RESO project has been created the 1st of December 2003. The RESO team belongs to the “Laboratoire de l’Informatique du Parallélisme” (LIP) - Unité Mixte de Recherche (UMR) CNRS-INRIA-ENS with Université Claude Bernard of Lyon. It consists of fifteen members in average, including four permanent researchers and teaching researchers. RESO is part of the " Numerical Systems " theme of the INRIA, part of the B subsection: Grids and high-performance computing. The research activities of the RESO project fits the first priority challenge of the INRIA’s strategic plan: "design and master the future network infrastructures and communication services platforms" .

2.1.2. Context

Wavelengths multiplexing and future wavelengths switching techniques on optical fibers allow core network infrastructures to rapidly improve their throughput and reliability. In a near future, links of many tens (some hundreds) of Gigabits per second will be made available. New technologies like 10 Gigabit/s Ethernet or 10Gigabit/s Infiniband will also drive the increase of bandwidth in local area networks. These improvements have given the opportunity to create high performance distributed systems called "computational grids" that aggregate storage and computation resources into a virtual and integrated computation intensive environment. Grid computing is a promising technology harnessing distributed resources into virtual organizations for the future resource intensive scientific and business applications. However, moving enormous quantities of data among grid elements and ensuring efficient message passing between communicating processes raise specific challenges on the communication protocols and their related mechanisms. Although grids theoretically offer solutions for resources aggregation, high performance for applications may be hard to obtain due to the improperness of communication protocols and software and to the fact that processors speeds, in charge of protocol processing do no scale with network speeds. In order to deliver grid traffic in a timely, efficient, and reliable manner over long distance networks, several issues such as quality of service, security, and network resource scheduling, have to be investigated.

2.1.3. Research area

Our work follows two major research axes :

- Optimized software architectures for efficient communications in end systems, cluster-based servers and programmable access equipments,
- Protocols and computations for efficient and customizable transport of heterogeneous streams.

The first research axis explores how communication subsystems in end systems, in cluster networks and programmable access equipments can be enhanced and optimized. Our researches focus on high performance software solutions for clusters, new active network solutions for IP networks and interconnection of IP networks, networks of clusters or networks of data storage. We search at optimizing both data movements and I/O management that are closely inter-dependant, by using the intelligence of network interface cards (NICs).

The second research axis explores the problem of efficient transfer of heterogeneous flows in a high performance and high speed long distance networking infrastructure. The scientific directions we follow concern the study of flexible solutions exploiting innovating networking services in routers and the addition of packet processing software components at the edge of the core network for controlling the flows. Problems to be solved are modeling and quantifying the influence of the different performance parameters on a transport connection and the end-to-end characterization of the network links with discontinuous network services, the design of adaptive algorithm dedicated to the expressed flow needs, definition and placement in the network of active behavior meaningful to the semantic of the end-to-end transport protocols or application, making the interaction between packet processing and forwarding smooth and efficient.

2.1.4. Application domains

RESO applies its research to the domains of high performance computing and to Grid communications. The geographical topology of the Grid depends on the distribution of the community members. Though there might be a strong relation between the entities building a virtual organization, a Grid still consists of resources owned by different, typically independent organizations. Heterogeneity of resources and policies is a fundamental result of this. Web Service Resource Framework (WS-RF), the coupling of the notion of resource and web service, has recently been introduced. It has added the performance issue to the re-usability, interoperability, and openness advantages of web services. Grid services involve operations and strategies from application layer down to network layer, with service agreements defined at application layer and middleware developed for the communication between layers. In a typical implementation scenario, the grid middleware provisions the resource, and passes the delivery criteria to the network services. The network, accordingly, follows up to enforce the appropriate data transfer. In a Grid, the network performance requirements are very high and may strongly influence the performance of the whole distributed system. The construction of grid networks over the optical transport layer tackles the problem of communication performance from the transport medium perspective. However, our vision is that Grid applications, due to the heterogeneity and large scale factors, will continue to use traditional IP packet protocols, at least in the end systems and will rely on a complex interconnection of heterogeneous networks. In such context end-to-end flow performance cannot be guaranteed or predicted. Thus, for achieving end-to-end QoS objectives, the remaining deficiencies of the network performance have to be masked by adaptation performed at the host level or somewhere in the datapath. RESO designs Grid network services and network middleware to avoid the applications to be network-aware, to simplify the programming and to optimize the execution of their communication parts while fully exploiting the capacities of the evolving network infrastructure.

2.1.5. Methodology

The RESO approach relies on the analysis of limitations encountered in existing systems or protocols and on the theoretical and experimental exploration of new approaches. This research framework between a challenging application domain and a specific network context, induces a close interaction with the application level and with the underlying network level. The methodology is based on a study of the high end and original requirements and on experimental evaluation of the functionalities and performance of high speed infrastructures. RESO gather expertise in advanced high performance local area networks protocols, in distributed systems and in long distance networking. This background work provides the context model for innovative and adequate protocols and software design and evaluation. Moreover, the propositions are implemented and experimented on real or emulated local or wide area testbeds with real conditions and large scale applications.

2.1.6. Goals

RESO aims at providing software solutions for high performance and flexible communications fully exploiting the very high speed networking infrastructure of computational and data grids. The goal of our research is to provide analysis of the limitations of the current communication software and protocols designed for standard networks and traditional usages, and to propose optimization and control mechanisms for the end-to-end performance and quality of service. RESO explores original and innovative end-to-end transport services and protocols that meet the needs of grid applications. These solutions must scale in increasing bandwidths, heterogeneity and number of flows.

RESO creates open source code, distributes it to the research community for evaluation and usage. The long term goal is also to contribute to the evolution of protocols and networking equipments and to the dissemination of new approaches.

2.1.7. Summary of the main contributions of the team in 2004

During this year, RESO team had main contributions in the following fields:

2.1.7.1. Axe 1: Optimized communication software and equipments

- Design and proposition of a new networking subsystem architecture built around a packet classifier executed in the Network Interface Controller (NIC). Development of the KNET software in collaboration with SUN;
- Study and design of efficient remote data access for clusters, that maximizes the underlying network utilization. Development of the ORFA (Optimized Remote File-system Access) software prototype on Myrinet networks. Integration of new feature in the new Myrinet programming interface, MX, in collaboration with Myricom;
- Development of a high performance active network architecture (Tamanoir) and associated tools (Echidna, Pangolin). Proposition of load balancing functions in cluster-based active routers;
- Validation of Tamanoir through internal and external projects (IBP, deployment of FPTP (LAAS, Toulouse), deployment of programmable nodes (3DDL project));

2.1.7.2. Axe 2: end-to-end transport and service differentiation

- Design, development and experimentation of an alternative solution for end-to-end service differentiation in TCP/IP environment, named **equivalent differentiated services (EDS)**. This proposition aims at enhancing performance differentiation at the IP level without requiring any control plane by offering to the packets a tradeoff between delay and loss in each forwarding hop. Experimentation within the DataTAG transoceanic network;
- Design, development and experimentation of a control service - QoSINUS - that manages the performance of a set of heterogeneous flows with adaptive packet marking. This control service attempts to optimize the utilization rate while honoring the requirement of individual flows. This service is implemented as a QoS API and an active service of the Tamanoir architecture;
- Proposition and exploration of a new approach of congestion control, based on back-pressure flow control for high transfer throughput in dedicated long fat networks;
- Analysis and experimentation of non-intrusive methods of throughput measurement and proposition of an original hop by hop method for link capacity estimation and path utilization rate evaluation. A tool implementing this method, **TraceRate** has been developed;
- Design and development of high performance active services (DYRAM, QoSINUS);
- Deployment of active and programmable solutions around VTHD backbone (RNRT VTHD++ project) and to support Grid applications (RNTL e-Toile).

2.1.7.3. Grid Network services and applications

- Contribution to the analysis of the Grid Networking issues within the GGF community;
- Contribution to the design and development of the GRID5000, national Grid testbed;
- Contribution to the design and development of the GridExplorer, national Grid emulator;
- Design and development of a software for configuring a cluster in a grid network cloud, eWAN;
- Contribution to the standardization of the HIP (Host Identity Protocol) allowing the Distributed security in Grids;
- Definition of a new architecture based on an overlay approach for offering a deterministic data transfer service to grid applications.

3. Scientific Foundations

3.1. Optimized communication software and equipments

Participants: Olivier Glück, Brice Goglin, Laurent Lefevre, Pascale Vicat-Blanc Primet.

The emergence of high performance parallel applications has raised the need of low latency and high bandwidth communications. Massively parallel supercomputers provided integrated communication hardware to exchange data between the memory of different nodes. They are now often replaced by clusters of workstations based on high-speed interconnects such as MYRINET or INFINIBAND which are more generic, more extensive, less expensive and where communications are processed by dedicated network interfaces.

A large amount of interesting work has been done to improve communications between cluster nodes at the application level through the use of the advanced features in the network interface card and *OS-bypass* techniques. Meanwhile, storage access needs to reach similar performance to read input data and store output data on a remote node without being the bottleneck. Parallel applications require both efficient communication between distant application tasks and fast access to remote storage. High performance distributed file systems have special requirements that have not really been considered when designing most underlying network access layers. While usual application communications should obviously occur at user-level, distributed file system were initially implemented in the kernel to supply transparent remote accesses. They were designed for traditional networks and caching was used to compensate the high latency.

In a cluster environment, two directions are studied to improve the performance of distributed file systems: distributing the workload across multiple servers or efficiently using the low latency and high bandwidth of the underlying high-speed network.

In this research axis we explore the de-localization of network functionalities in dedicated equipments (programmable NICs) or intermediate nodes (programmable network equipment).

We studied several techniques based on new functions in the network interface controllers to maximize the execution efficiency of the operating system's communication software. In particular, the main proposition of our work (KNET software suite) is to place a packet classifier in the network interface controller in order to smartly spread incoming network streams across the processors (or threads) of connected servers.

In order to support network functions in the network, we propose a high performance active network environment execution architecture (Tamanoir software suite). This architecture is based on various layers adapted on services and applications requirements : NICs for no state ultra lightweight services, kernel for few state lightweight services, user space for middle service and distributed resources for CPU/storage consuming services. We propose various adaptive solutions (load balancing, fault tolerance) to efficiently deal with heterogeneous services and applications.

3.2. End-to-end transport and services differentiation

Participants: Fabien Chanussot, Marc Herbert, Mathieu Goutelle, Dino Lopez Pachenko, François Echantillac, Cong-Duc Pham, Pascale Vicat-Blanc Primet.

In TCP/IP networks, the end-to-end principle aims at simplifying the network level while pushing all the complexity on the end host level. This principle has been proved to be very valuable in the context of the traditional low capacity Internet. In packet networking, congestion events are the natural counterpart of the flexibility to interconnect mismatched elements and freely multiplex flows. Managing congestion in packet networks is a very complex issue. This is especially true in IP networks where, at best, congestion information is very limited (e.g., **ECN**) or, at worst, non-existent, forcing the transmitter to infer it instead (e.g., based on losses or delay) in TCP.

The conservative behavior of TCP with respect to congestion in IP networks (RFC 2581) is at the heart of the current performance issues faced by the high-performance networking community. Several theoretical and experimental analysis have shown that the dynamics of the traditional feedback based approach is too low in very high speed networks that may lose packets. Consequently network resource utilization is not optimal and

the application performance is poor and disappointing. Many Grid-enabled computing applications wish to transfer large volumes of data over wide area networks and require high data rates in order to do so. However, Grid-enabled applications are rarely able to take full advantage of the high-capacity (2.5 Gbit/s, 10 Gbit/s and upwards) networks installed today. Recent data for Internet 2 show that 90% of the bulk TCP flows (defined as transfers of at least 10 Megabyte of data) use less than 5 Mbit/s, and that 99% use less than 20 Mbit/s out of the possible 622 Mbit/s provision. There are many reasons for such poor performance. Many of the problems are directly related to the end system, to the processor and bus speed, and to the NIC with its associated driver. TCP configuration (e.g., small buffer space or features such as SACK being improperly negotiated) will have a significant impact. TCP itself was designed first and foremost to be robust and when congestion is detected, TCP accommodates the problem but at the expense of reduced performance. There are also design problems with TCP itself. For example, for a standard TCP connection with 1500-byte packets and a 100 ms round-trip time, achieving a steady-state throughput of 10 Gbit/s would require an average congestion window of 83,333 segments, and a packet drop rate of at most one congestion event every 5,000,000,000 packet (or equivalently, at most one congestion event every 1 2/3 hours). HighSpeed TCP [44] and Scalable TCP [48] increase the aggressiveness in high-throughput situations while staying fair to standard TCP flows in legacy contexts. **FAST** [42] leverages the queueing information provided by round-trip time variations, in order to efficiently control buffering in routers and manage IP congestion optimally. These propositions are actively analyzed and experimented by the international community. RESO participates to the elaboration of a survey on protocols other than standard TCP in the framwork of the Data Transport research group of the Global Grid Forum [41]. RESO is organizing in 2005, the third edition of the leading international workshop in this domain (see http://www.ens-lyon.fr/LIP/RESO/pfldnet2005). Several issues have been already enlightened. Considering the traditional feedback loop will not scale with higher rate level under loss or congesting traffic conditions, it seems judicious to start examining alternative radical solutions.

On the other hand, flows crossing the IP networks are not equally sensitive to loss or delay variations. Since several years, research effort has been spent to solve the problem of the heterogeneous performance needs of the IP traffic. A class of solutions considers that the IP layer should provide more sophisticated services than the simple best-effort service to meet the application's quality of service requirements. Quality of service has been studied in IP networks in the context of multimedia applications [43]. Various complementary solutions have to be integrated to carry end-to-end quality of service to grid applications to assure an efficient usage of the interconnected computing resources [46]. Solution like Diffserv exhibits three types of limitations we are considering:

- the end-to-end performance that the DiffServ standardized services provide have not been largely studied in real networks;
- when experiment shows that end-to-end connection can benefit from advanced DiffServ QoS network functionalities, their usage by individual flows is not straightforward;
- the deployment of DiffServ architecture presents different scaling problems. Alternative approaches are proposed to solve this issue.

Finally, tools for measuring the end-to-end performance of a path between two hosts are very important for transport protocol and distributed application performance optimization. Bandwidth evaluation methods aim to provide a realistic view of the raw capacity but also of the dynamic behavior of the interconnection that may be very useful to evaluate the time for bulk data transfer. Existing methods differ according to the measurements strategies and the evaluated metric. These methods can be active or passive, intrusive or non-intrusive. Non-intrusive active approaches, based on packet train or on packet pair provide available bandwidth measurements and/or the total capacity measurements. None of the proposed tools, based on these methods, enable the evaluation of both metrics, while giving an overview of the link topology and characteristics.

3.3. Grid Network services and applications

Participants: Pascale Vicat-Blanc Primet, Jingdi Zheng, Olivier Glück, Julien Laganier, Fabien Chanussot, Mathieu Goutelle, Robert Harakaly, Jean-Christophe Mignot.

The purpose of Computational Grids is to aggregate a large collection of shared resources (computing, communication, storage, information) to build an efficient and very high performance computing environment for data-intensive or computing-intensive applications [47]. But generally, the underlying communication infrastructure of these large scale distributed environments is a complex interconnection of multi-IP domains with changing performance characteristics. Consequently *the Grid Network cloud* may exhibit extreme heterogeneity in performance and reliability that can considerably affect the global application performance. Performance and security are the major issues grids encountered from a technical point of view.

The performance problem of the grid network cloud can be studied from different but complementary view points. All these approaches are valuable and will fit the grid network services middleware framework under definition stage at GGF.

- Measuring and monitoring the end-to-end performance helps to characterize the links and the network behavior. Network cost functions and forecasts, based on such measurement information, allow the upper abstraction level to build optimization and adaptation algorithms.
- Optimally using network services provided by the network infrastructure for specific grid flows is of importance.
- Creating enhanced and programmable transport protocols adapted to heterogeneous data transfers within the grid may offer a scalable and flexible approach for performance control and optimization.
- Modeling, managing and controlling the grid network resource as a first class resource of the global environment: transfer scheduling, data movement balancing...

4. Application Domains

4.1. Panorama

Keywords: *Active Networks*, *Communication Software*, *End to End Transport*, *Grids*, *High Performance*, *Networks*, *Protocols*, *Quality of Service*, *Telecommunications*.

RESO applies its research to the domains of high performance Cluster and Grid communications. Existing GRID applications did already identify potential networking bottlenecks, either caused by conceptual or implementation specific problems, or missing service capabilities. We participated to the elaboration of the first GGF document on this subject [40] [49][50]. Loss probability, important and incompressible latencies, dynamic behavior of network paths question profoundly models and technic used in parallel and distributed computing [45]. The particular challenge arises from a heavily distributed infrastructure with an ambitious end-to-end service demand. Provisioning end-to-end services with known and knowable characteristics in a large scale networking infrastructure requires a consistent service in an environment that spans multiple administrative and technological domains. We argue that the first bottleneck is located at the interface between the local area network (LAN) and the wide area network (WAN).

RESO conducted several actions in the field of Grid High Performance Networking in the context of the GGF, the Europe or National projects. These activities have been done in close collaboration with the CNRS-UREC team, other INRIA and CNRS French teams (Grand Large, Apache, Graal) involved in the GRID5000 and the Grid Explorer projects and other European teams involved in the DataGRID and DataTAG project.

- We have participated to the design, development and deployment of an extensible Network Monitoring system that measures, gathers and publishes relevant monitoring information in the global information system of the Grid like MDS and R-GMA in the DataGRID testbed [17].

- We have actively participated to the design and deployment of a Grid testbed based on a controlled private very high speed network, VTHD: eToile. The innovative Network Services, Tamanoir environment, Dynamic Network Quality of Service Management and control (QoSINUS suite) Active Reliable Multicast (DyRAM) have been deployed and are used in this testbed of the RNTL French eToile Grid [30].
- The Madeleine, multi-protocol communication library, has been adapted and integrated both in Globus and eToile middleware.
- We continue the investigation of limits of the existing communication services or protocols and evaluate more efficient approaches within the DataTAG platform based on a transoceanic dedicated 10Gbit Ethernet link and Grid5000 national experimental infrastructure based on the RENATER network. Participating to the design, deployment and usage of such high performance experimental Grid testbed allows us to evaluate and measure the benefit that grid middleware and applications can get from enhanced networking technologies. The experience and expertize we get from this work are a tremendous gain for our research on performance bottlenecks.
- Grid 5000 is a national initiative aiming at providing a huge experimental instrument to the grid software research community. Lyon, with RESO and GRAAL projects, is part of this initiative. RESO is closely involved in the design and deployment of the testbed, and responsible for the networking aspects.
- We participate to the definition of the Grid Explorer physical architecture and to the design of the configuring, tuning and monitoring software. Grid Explorer, the largest cluster of Grid 5000 platform will be a very large scale instrument for grid software evaluation.

5. Software

5.1. ORFA (Optimized Remote File-system Access)

Keywords: *SAN networks*, *filesystem*.

Participants: Brice Goglin (contact), Olivier Glück.

ORFA is a user-level remote filesystem access protocol. It makes the most out of Myrinet networks through their GM or BIP interface for direct data transfer between user application buffers on the client's side and remote server file systems.

ORFS is the kernel port of ORFA. It runs on GM or MX interfaces over Myrinet networks. Both buffered and non-buffered accesses are implemented, with asynchronous or synchronous standard I/O primitives through the Linux kernel.

Details are available at http://perso.ens-lyon.fr/brice.goglin/work.php

5.2. eWAN (Emulator WAN)

Keywords: *eWAN*, *grid networking*, *network emulation*.

Participants: Cyril Otal, Olivier Glück (contact), Pascale Primet, François Echantillac.

EWAN [35] is a software and hardware tool for configuring and programming a large PC cluster in a wide area network emulation instrument. High performance, fine parameter tuning and a great utilization flexibility are the main proposed features of this experimental tool. Details are available at http://www.ens-lyon.fr/LIP/RESO/Software/EWAN/

5.3. SNE (Stateful Network Equipment)

Keywords: *High Availability*, *fault tolerance*.

Participants: Laurent Lefevre (contact), Pablo Neira Ayuso.

SNE is a complete library for designing a stateful network equipment (contains Linux kernel patch + user space daemon). The aim of the SNE library is to support issues related to the implementation of high available network elements, with specially focus on Linux systems and firewalls. The SNE library (Stateful Network Equipment) is an add-on to current High Availability (HA) protocols. This library is based on the replication of the connection tracking table system for designing stateful network equipments. Software is available at http://perso.ens-lyon.fr/laurent.lefevre/software/SNE

5.4. NetLinkBench : Netlink Socket Benchmark Tool

Keywords: *High Availability*, *fault tolerance*.

Participants: Laurent Lefèvre (contact), Pablo Neira Ayuso.

The Netlink sockets are an extension of the IP service which allow to exchange messages with the user space. Netlink sockets provide an efficient way to notify events and a smart interface from user space. They are implemented wrapped in socket syscalls operations, to be precise they are defined as a new socket type. They are proposed as an extension of the IP service.

To evaluate Netlink sockets in Linux, we propose the NetLinkBench tool which consists of two components, a kernel module and a user space tool. It allows to communicate broadcast messages between kernel and user space. By this way throughput and timestamping can be evaluated.

Software is available at http://perso.ens-lyon.fr/laurent.lefevre/software/netlinkbench

5.5. NOQ (Network of Queues)

Keywords: *High performance transport*, *congestion control*, *flow control*.

Participants: Marc Herbert, Pascale Vicat-Blanc Primet.

To tackle the problem of high performance transport over a dedicated large network, we investigate a hop-by-hop congestion control mechanism based on link level flow control (backpressure). We explain the dependence between tcp performance and burstiness at the Ethernet level, the stunning performance and fairness gains obtained by the activation of Ethernet 802.3x flow control. We propose a "Network of Queues" (noq) approach, where each network buffer protects itself against overflow, thus preventing global network congestion [25]. The software is available at http://marc.herbert.free.fr/noq/.

5.6. QoSINUS

Keywords: *DiffServ*, *Service Level Specification*, *active service*, *adapted packet marking*.

Participants: Fabien Chanussot (contact), Pascale Vicat-Blanc Primet.

QoSinus : QoSinus is an active QoS service that interfaces the application QoS specifications (SLS) with an adaptive packet marking at DiffServ domains frontiers. QoSINUS integrates a Earlist Deadline First algorithm for scheduling the flows. QoSinus is distributed in the RNTL eToile suite under a GPL licence. All details of QoSinus are available at http://www.ens-lyon.fr/LIP/RESO/Software/QoSINUS/index.html

5.7. EDS PHB and transport protocols

Keywords: *Proportional DiffServ*, *RED*, *SCTP*, *adapted packet marking*.

Participants: Benjamin Gaidioz, Mathieu Goutelle, François Echantillac (contact), Pierre Billiau, Pascale Vicat-Blanc Primet.

The EDS PHB (Equivalent DiffServ Per Hop Behavior) and SCTP-based packet marking protocol, SCTP-lm, provide alternative DiffServ mechanisms (based on Proportional Diffserv and RED) and transport adaptive packet marking protocols developed as Linux modules [16]. `SCTP-based protocols` have been developed and evaluated within the EU DataTAG project [18] . All details of EDS PHB and protocols are available at http://perso.ens-lyon.fr/francois.echantillac/EDS/Doc/EDS.html

5.8. MFTP (Multicast File Transfer Protocol)

Keywords: *Active networks*, *Multicast*, *Protocol*.

Participants: Fayçal Bouhafs, Congduc Pham.

MFTP is an implementation of the DyRAM protocol [15] that has been designed in our research group. MFTP is composed of a Java API that provides primitives for sending files across the Internet within a multicast group. It has mainly been developed to demonstrate the active reliable multicast features od DyRAM. The TAMANOIR execution environment is used for active services deployment. MFTP has been tested under the Linux operating system. The MFTP package can be found at http://www710.univ-lyon1.fr/~cpham/MULTICAST/index.html along with the documentation.

5.9. Java Wrappers for FEC library coding

Keywords: *FEC*, *Java*, *Multicast*.

Participants: Sylvain Dattrino, Congduc Pham.

We developed a Java Wrapper for the LDPC library written by V. Roca from the INRIA Planete project and for the Reed-Solomon library developed by Luigi Rizzo (Univ. Pisa, Italy). This wrapper allows the end-user to develop Java programs by using both the MFTP library to get active reliable multicast features and FEC codes to integrate a FEC mechanism. The Java wrapper package can be found at http://www710.univ-lyon1.fr/~cpham/MULTICAST/index.html

5.10. TraceRate

Keywords: *Network performance measurement*, *Packet Pair*, *Topology discovery*, *TraceRoute*, *capacity estimation*.

Participants: Mathieu Goutelle (contact), Pascale Vicat-Blanc Primet.

`TraceRate` is a LINUX implementation of the hop by hop path rate estimation method. This tool is split into two modules. The first one is the measurement module, which sends many times a back-to-back packet pair and gather the dispersion measurements. The second module does the distribution analysis. The measures are done for each value of TTL between source and destination in order to investigate the whole path. By default, 500 packet pairs are sent for each loop with 1400 bytes. The tool is immunized from ICMP and UDP packets limitation, firewalls filtering. This tool is an adaptation of the well-known `traceroute` which sends TCP packets instead of ICMP packets. `TraceRate` has been developed and evaluated within the EU DataTAG project[24] . All details of TraceRate are available at http://www.ens-lyon.fr/LIP/RESO/Software/TraceRate/index.html

5.11. Tamanoir

Keywords: *active and programmable networks*, *execution environment*.

Participants: Jean-Patrick Gelas, Laurent Lefèvre (contact).

Tamanoir is an open source software environment for high speed active networks. Available on the web and protected by APP (Agence Francaise de Protection des Programmes). TAMANOIR is distributed within the RNTL eToile suite. It is used by partners in RNTL eToile Project and in the collaboration with 3DDL company (for supporting deployment of Java based games on mobile platforms). All details on Tamanoir are available at http://www.ens-lyon.fr/LIP/RESO/Software/Tamanoir/index.html

6. New Results

6.1. Optimized communication software and equipements

6.1.1. Optimized Remote File-system Access

Participants: Brice Goglin, Olivier Glück, Pascale Vicat-Blanc Primet, Loic Prylli.

Data storage in a cluster environment requires dedicated systems that are able to sustain high bandwidth needs and serve many concurrent clients. Several projects have already been proposed to address this issue. PVFS, GPFS or Lustre provide parallel file systems whose scalability is ensured by data stripping and workload sharing across several servers.

We study the link between clients and these systems in order to maximize the underlying network utilization. Indeed cluster nodes are connected through a high bandwidth low latency network such as Myrinet, whose features lead us to the idea of using them for data storage. ORFA (*Optimized Remote File-system Access*) was developed on Myrinet networks to provide an efficient access to remote data. The fully transparent user-level client [23] allows any legacy application to saturate the physical link by accessing remote files. The need to cache metadata on the client's side leads to the idea of developing ORFS (*Optimized Remote File-System*), the port of ORFA into the Linux kernel. Besides, the use of ORFA-like techniques in parallel filesystems should enhance their performance to make the most out of the underlying network.

This work also showed that the now well-known memory registration model that is used on asynchronous network interface such as Myrinet does not fit file system implementation needs. Maintaining a registration in the kernel to enable high-performance non-buffered remote file access has required to patch the Linux kernel so that an external module might be notified of address space modifications. Collisions between virtual addresses of different processes have been avoided by using a modified GM firmware in the network interface card, making the registration cache as efficient in the kernel than in user-level in the ORFA implementation. Besides, buffered accesses have been improved by replacing the traditional memory registration with physical address based primitives which are much more suitable for such an environment. These make the ORFS implementation very efficient [22].

We are currently working with Myricom to integrate all this work in their new driver, MX (Myrinet Express), so that the interaction between it and file systems implementations will be much easier and efficient. Preliminary results have been published in [32].

6.1.2. NIC-assisted optimizations for multi-processor machines

Keywords: *Communication system*, *Network processors*, *Operating systems*.

Participants: Eric Lemoine, Laurent Lefèvre, Congduc Pham.

The high demand in network communication, notably with the success of the Internet, networks are in constant evolution. In particular, the throughput of networks has been increasing at an impressive pace these last years. However, progress in single-threaded microprocessor technologies is slowing down relatively to network technologies. Thus, it is more and more difficult for a single processor to keep up with the throughput of the network. With the multi-processor technologies becoming more and more commodity and with the emergence of multi-threaded processor technologies, a new opportunity for facing the high network throughput is opening: simultaneously using multiple execution threads to network processing. However, due to issues inherently related to multi-processing, such as cache misses and lock contentions, special care in the design and development of the communication software must be taken for efficiently executing this software. In this work, we present several techniques based on new functions in the network interface controllers to maximize the execution efficiency of the operating system's communication software. In particular, the main proposition of our work is to place a packet classifier in the network interface controller in order to smartly spread incoming network streams across the processors (or threads) of the machine. The KNET software prototype has been developed [27][33] and demonstrates the relevance of the proposed techniques. The experimental results are presented and analyzed in E. Lemoine PhD manuscript [11].

6.1.3. Lightweight software functionalities in programmable networks

Keywords: *execution environments*, *programmable and active networks*.

Participants: Aweni Saroukou, Jean-Paul Corvo, Laurent Lefèvre.

We have proposed a new execution environment called `Tamanoir`, which focuses on performance problems of active and programmable network equipments and dynamic deployment of services. Targeted equipments

are deployed in access networks around high performance (Gbit/s) backbones. These networks must face heterogeneity problems in terms of equipments and bandwidth.

The `Tamanoir` architecture is designed to be a high performance active router able to be deployed around high performance backbones. This approach concerns both a strategic deployment of active network functionalities around backbone in access layer networks and providing a high performance dedicated architecture.

Tamanoir Active Nodes (TAN) provide persistent active nodes supporting various active services applied to multiple data streams at the same time. Both main transport protocols (TCP/UDP) are supported by the TAN for carrying data. We rely on the user space level of the 4 layers of the Tamanoir architecture (Programmable NIC, Kernel space, User space and Distributed resources) in order to validate and to deploy our active collaborative cache services.

The high performance Tamanoir architecture has been implemented on a cluster-based infrastructure and supports active services inside the Linux kernel and on distributed resources .

Experimental tests have been made around high performance backbone (RNRT VTHD++ project), and for alternative support of Grid network infrastructure (RNTL e-Toile) [14][13][19].

6.1.4. Context aware network services : supporting deployment of Java games on mobile platforms

Keywords: *execution environments*, *programmable and active networks*.

Participants: Aweni Saroukou, Laurent Lefèvre.

Active Networks allow user or applications to inject customized programs into the network nodes. The creation of new services is an original way to think about development and deployment of customized modules to perform computation within the network. This can lead to massive improvement of network functionalities.

New mobile phone generations integrate more and more a Java Virtual Machine. This JVM allows providers to propose applications and games working on heterogeneous phones (without having to redo some specific development and to adapt them individually for specific features).

We propose to benefit from active and programmable networks by deploying active nodes on data path to efficiently adapt streams on the fly. This research follows three main goals :

- to reduce development costs and the complexity for managing a version of a game for each mobile class. The active node will adapt the files on the fly;
- to reduce the usage of bandwidth and interactions between clients and games server;
- to efficiently support deployment of games without adding too much latency on real networks.

We design the architecture of an active transcoding service (ActiveWapS) deployed inside the Tamanoir Execution Environment. This service transforms on the fly, parts of the games (JAD files) in order to adapt them to target mobile phones. We also validate this approach on a local platform with emulated wireless network.

6.1.5. High availability for stateful network equipments

Keywords: *fault tolerance*, *high availability*.

Participants: Pablo Neira Ayuso, Laurent Lefèvre.

In operational networks, the availability of some critical elements like gateways, firewalls and proxies must be guaranteed. Some important issues like the replication of these network elements, the reduce of unavailability time and the need of detecting failure of an element must be studied. We propose the SNE library (*Stateful Network Equipment*) which is an add-on to current High Availability (HA) protocols. This library is based on the replication of the connection tracking table system for designing stateful network equipments.

Proposing stateful network equipments on open source systems is a challenging task. We propose the basic blocks (SNE library) for building a stateful network equipment. This library can be combined with high-availability protocols (CARP, Linux HA...). We focus on Linux system in order to provide software solutions for designing high-available solutions for NAT, firewalls, proxies or gateways equipments...This library is based on components located in kernel and in user space of the network equipment. First micro-benchmark of communications mechanisms with Netlink sockets have shown the effectiveness of our approach [38].

6.2. High performance transport protocols and differentiated services

6.2.1. Dynamic end-to-end QoS control

Keywords: *DiffServ*, *Network Quality of Service*, *SLS*, *packet marking algorithm*.

Participants: Fabien Chanussot, Pascale Vicat-Blanc Primet.

A Grid oriented QoS API and a programmable QoS service [29][30][34] **QoSINUS** have been designed and developed within the context of the e-Toile project to introduce flexibility and dynamic in the management, the control and the achievement of end-to-end QoS in Grid context. Such an approach increases slightly the complexity at the Grid/WAN Network frontier points, but leaves the core network and the grid applications simple. This edge service aims at :

1. allowing heterogeneous Grid flows to specify individually and directly their QoS objectives,
2. mapping these objectives with the existing IP QoS services provided at the edge of the core inter-networks for improving the individual packet performance,
3. realizing a dynamic and appropriate adaptation according to the real state of the link, the QoS mechanisms configured and the experienced performance.

The first issue is addressed by an API that provides the user the ability to characterize the flow needs in terms of qualitative or quantitative end-to-end delay, end-to-end throughput, end-to-end loss rate or in terms of relative weight of these three main metrics. This API allows to define SLS (end-to-end service level specifications) in XML.

The second issue is addressed by a service architecture that combines flow aware and infrastructure aware components to map and dynamically adapt the QoS specification of the flows to the QoS facilities offered by the network. IP premium is a finite and scare resource. To avoid to waste this resource, we propose algorithms that statically or dynamically adapt the packet marking according to the real QoS of a TCP flow. The analysis of ACK allows to calculate periodically the amount of data transfered and to increase packet priority when required in order to meet some deadline requirement. The ultimate goal is to provide an Earliest Deadline First algorithm in an edge packet marking equipment, in order to serve the performance requirements of individual TCP flows. This algorithm has been implemented in an active service under the TAMANOIR environment. This service has been applied to a content based search in a medical imaging context, in collaboration with the Creatis laboratory [34].

6.2.2. Evaluation of TCP on variable bandwidth environments

Keywords: *TCP*, *congestion control*, *simulations*, *variable bandwidth*.

Participants: Dino Martin Lopez-Pacheco, Congduc Pham.

The assumption of constant bandwidth capacity may be not true anymore because many telco-operators and Internet providers (ISP) are beginning to deploy Quality of Service (QoS) features with reservation-like or priority-like mechanisms in their networks. Therefore, the available bandwidth for best-effort traffic can vary over time. This work studies the behavior and the performance issues of TCP and its new variants (HSTCP and XCP) in such environments. Both sine-based and step-based bandwidth variations models, which represents more closely dynamic bandwidth provisioning scenario, are used. The results highlight the problem of deterministic increase of the congestion windows which is not suitable for variable bandwidth environments and describe the different phases of TCP when facing bandwidth variations [28].

6.2.3. Evaluating and designing estimation mechanisms for variable bandwidth environments

Keywords: *TCP*, *congestion control*, *estimations*, *variable bandwidth*.

Participants: Dino Martin Lopez-Pacheco, Congduc Pham.

This work is a continuation of the previous study but with a focus on 3 transport protocols: TCP New Reno, TCP Westwood+, and XCP. TCP New Reno is the reference point for the comparison. TCP Westwood+ is an end-to-end approach that tries to detect the bandwidth variations by means of ACK filtering and monitoring. XCP, which is a router-assisted proposition, has additional functionalities that allows the source to get information about the available bandwidth for best-effort traffic along the path from the source to the receiver.

We also are working on improving our network model. We argue that the bandwidth variation can be represented by the aggregation of UDP on-off sources. As it turned out, the variation model produced by the UDP traffic is similar to the step variation model, but the first is more realistic because the routers' buffers are used by the cross-traffic as well. With this new network model, we found that TCP New Reno and TCP Westwood+ are not able to acquire the available bandwidth when it increases, even though the value of RTT is not very large. On the other hand, XCP is able to acquire the bandwidth available in almost all conditions, but it requires that all routers have XCP features. We are currently using the ns simulator to further investigate this research direction.

6.2.4. Network of queues

Keywords: *High speed transport*, *congestion control*, *flow control*.

Participants: Marc Herbert, Pascale Vicat-Blanc Primet.

The new congestion control solution we propose for high speed network is based on back-pressure flow control. Our Network of Queues proposal [25] suggests an outright departure from current TCP standards for some particular networks. The idea is to replace the current end-host-based TCP/IP congestion management by a network of flow-controlled links, according to a scheme known as "back-pressure". The challenge is to control the flow queue by queue with the functionalities already present in the intermediate networking equipment. The idea is to activate the 802.3x flow control in very high speed Ethernet links to reduce the feedback loop and to efficiently prevent the congestion. It has been proved theoretically that the back-pressure approach is better than classical end system feedback control approach. The issue to solve is to prove its validity in the actual equipments and in operational IP networks and to solve cross layer issues. Our proposal argues that, in some specific networking contexts like those of grids, using back-pressure as an addition to existing TCP/IP/Ethernet networking hardware and software may offer a valuable tradeoff between performance gain and migration cost. In order to develop insights on how the current network hardware and software behave relative to flow control, we forced a 100 Mb/s bottleneck in local gigabit testbed. The result of the first, basic experiment is a sawtooth-shaped throughput curve. When several hosts compete for the same bottleneck, the cooperative **AIMD** algorithm of TCP gives an approximately fair share of the capacity to

each flow. The first promising conclusions are that this approach is feasible in a IP/802.3x environment and offers a smoother reaction to congestion compared to TCP or HS-TCP and a rapid convergence to fairness. We participate also to the exploration of the TCP/IP stack implementation and the elaboration of a collective document within the DataTAG project.

6.2.5. Integration of FEC codes into the DyRAM framework

Keywords: *FEC*, *Multicast*, *reliability*.

Participants: Sylvain Dattrino, Congduc Pham.

FEC (Forward Error Correction) codes have been proposed for multicast communications to provide scalability mostly by reducing the amount of feedback traffic (RFC3453). In this work, we propose to integrate FEC codes into a NACK-based multicast protocol in order to support several file distribution constraints on a computational grid. For instance, interactive applications such as distributed simulations are better supported with a FEC approach which typically decreases the recovery latencies. We developed a Java wrapper that allows Java development of software using the C++ library (an LDPC library that implement large block codec encoder/decoder in C++ has been developed by Vincent Roca from the Planete INRIA project). This wrapper has been developed by S. Dattrino as part of a practical project during his internship at the RESO/LIP laboratory (Dec 2003-Mar 2004). This library has been developed and integrated into the DyRAM framework. Results show that the combination of FEC and NACKs is beneficial to the application.

6.3. Grid Network services and applications

6.3.1. High Performance Network Emulator

Keywords: *eWAN*, *grid networking*, *network emulation*.

Participants: Cyril Otal, Olivier Glück, Pascale Primet, François Echantillac.

The Grid aims at expanding the cluster based parallel computing paradigm towards large scale distributed systems based on IP networks.

EWAN [35] is a high performance network environment emulator. It takes place in the research effort on computer grids, aggregations of computer resources inter-connected by a wide area network. EWAN offers an emulation framework needed by experiments in this field, bringing a great flexibility, a high level of performance and a precise control. It can be divided into two main parts: an interface for creation of simple topologies and an engine for deploying every sort of topologies. We have evaluated several network emulation solutions and have configured a 12 nodes cluster to test EWAN software on this cluster. As emulation solutions, we have compared Nistnet, netem and GtrcNET. Nistnet is a well known software to emulate network link, netem is an equivalent recently included in the Linux kernel and GtrcNET is an hardware network emulator developed by the AIST. EWAN manages all the three solutions and allows the user to choose one of them. All stuff (source code, documentation, results, ...) about EWAN can be found at http://www.ens-lyon.fr/LIP/RESO/Software/EWAN/index.html. A demonstration and a poster have been shown at Super Computing 2004 [35].

6.3.2. Programmable network support of Grid middleware

Keywords: *Globus*, *active networks*.

Participants: Grégoire Locqueneux, Laurent Lefèvre.

Efficiently and dynamically supporting Grid middleware with programmable and active network remains a challenging task. We explore some solutions to support the Globus Grid middleware with the Tamanoir active network environment [19][12]. Inside the Globus XIO API (Globus 3.2), we develop some transform and transport drivers to propose network services alternatives for Grid applications. This software development has been conducted as part of a practical project during the internship G. Locqueneux in RESO team (Dec 2003-Mar 2004).

6.3.3. Experimenting and deploying DyRAM on a grid infrastructure

Keywords: *Multicast*, *active networks*, *grids*, *reliability*.

Participants: Fayçal Bouhafs, Congduc Pham.

Today's computational grids are using the standard IP routing functionality, that has basically remained unchanged for 2 decades, considering the network as a pure communication infrastructure. With the grid's distributed system point of view, one might consider to extend the *commodity Internet*'s basic functionalities. Higher value functionalities can thus be offered to computational grids. In this work, we report on our early experiences in building application-aware components for multicast and in defining an active grid architecture that would bring the usage of computational grid to a higher level than it is now (mainly batch submission of jobs). To illustrate the potential of this approach, we first present how such application-aware components could be built and then some experiments on deploying enhanced multicast communication services for the grid. Results that will be published in [12] show that reliable multicast could deploy specific services based on the grid application needs.

6.3.4. Distributed security for Grid

Participants: Julien Laganier, Pascale Vicat-Blanc Primet.

RESO is specifically following Host Identity Protocol (HIP) activities within the IETF because HIP has been identified has one of the major breakthrough technologies to enable secure virtualization of the TCP/IP protocol suite. Project RESO/Holonet seeks to network virtualization because of its applications to dynamic reconfiguration of the grid infrastructure. Three Working Group documents describing 'DNS extensions' (by Nikander & Laganier) and 'Rendezvous extensions' (by Laganier & Eggert) for HIP have been edited [36][37][39] . The 'DNS extensions' allows a node to store HIP-related material in the DNS. This include its Rendezvous Server's IP address(es) or DNS names, as well as its Host Identity and its Host Identity Tag. The 'Rendezvous extensions' allows a HIP node to use another node, its Rendezvous Server (RVS), to maintain its reachability when changing its network attachment. A HIP node trying to communicate with such a HIP node would typically initiate communication towards its RVS, which will relay the initial packets of the HIP exchange to its client. Then the two nodes can communicate without further assistance from the RVS. This allows fast moving node to maintain reachability even if there is too much update latency in the name-to-address lookup service.[21][20]

6.3.5. Overlay for Grids

Participants: Jingdi Zheng, Pascale Vicat-Blanc Primet.

As Grids allow users to share resources over long distance networks, critical, are the degrees by which data are effectively transported among these resources. A diversity of approaches, with enhanced transport protocols and different transport mediums, have been proposed for high throughput. Within these approaches, bridging grid applications and network services is critical for users to control data transfer and application performance. Adopting an overlay infrastructure, we are proposing a new network architecture for grid data transport. We identified three important issues of the infrastructure: the global network resource scheduling of a grid overlay network, the flow control mechanism of grid local networks, and, the edge-to-edge transfer performance guarantee of grid overlay routers. Further, grid data bursts are introduced to reliably improve transport control and data delivery.[31]

7. Contracts and Grants with Industry

7.1. Myricom

Participants: Brice Goglin, Loïc Prylli.

This long-term collaboration between our team and US based Myricom company is focused on their software Myrinet suites. The old driver (GM) was used as a experimentation platform for the ORFA

(*Optimized Remote File-system Access*) software prototype and its kernel port, ORFS (*Optimized Remote File-System*). This work is now being integrated in the new driver (MX) to make the interaction between file systems and Myrinet software layers much easier and efficient.

7.2. INTEL

Participants: Pascale Vicat-Blanc Primet, Olivier Gluck, Laurent Lefevre.

This collaboration aims at studying the potential of the network processor technology for building High performance (several Gigabits links) network emulators and dynamically programmable routers. The goal is to show that network processors improve performance and enhance capacities of Software network emulators and programmable routers based on Linux platforms. Network interface cards with network processors have been integrated within the GRID5000 testbed.

7.3. SUN Labs, Europe

Keywords: *Operating systems*, *SMP machines*, *Solaris*, *network protocols*, *networking sub-systems*, *security*.

Participants: Marc Herbert, Julien Laganier, Laurent Lefèvre, Eric Lemoine, Congduc Pham, Pascale Vicat-Blanc Primet.

RESO has established a long term collaboration with Sun Labs (3 CIFRE grants). This collaboration focuses on high performance transport protocols, optimizing protocols on high performance servers and distributed security. Within the networking sub-system optimization research theme, we have also developed tight collaborations with several research groups in SUN Microsystems, especially with the groups that develop new technologies for `Solaris`™and SUN's network interface cards. Within the Distributed Security field, we are collaborating with the Holonet project of Sun on the HIP studies.

7.4. 3DDL

Keywords: *java* , *programmable networks*.

Participant: Laurent Lefèvre.

Support to the innovation of a SME : 3DDL. Collaboration on the support of programmable network for the deployment of mobile applications on cellular. Funded by Région Rhone-Alpes with collaboration of LIRIS, INSA Lyon, 2003-2004.

7.5. Bearstech

Keywords: *embedded PC*, *network services* .

Participant: Laurent Lefèvre.

In 2004, RESO is launching a collaboration with this young company targeted on embedded computers and network equipments.

8. Other Grants and Activities

8.1. Regional actions

8.1.1. Fédération Lyonnaise de Calcul Scientifique Haute Performance

Participants: Laurent Lefèvre, Cong-Duc Pham.

RESO is a member of the "Fédération Lyonnaise de Calcul Scientifique Haute Performance", that is building a regional grid infrastructure with several high-performance clusters and parallel machines. Supported by the Rhone-Alpes region (2004-2005).

8.2. National actions

8.2.1. RNTL eToile

Participants: Fayçal Bouhafs, Fabien Chanussot, Laurent Lefèvre, Congduc Pham, Geneviève Romier, Pascale Vicat-Blanc Primet.

The eToile project is an experimental wide area grid testbed. The e-toile project (e-toile is a RNTL project (réseau national de recherche en logiciel) funded by French Ministry of Research) had three complementary objectives:

- to build an experimental high performance grid platform that scales to France;
- to develop original Grid services to fully exploit the services and capacities offered by a very high performance network. The e-toile middleware integrates the most recent and relevant works of the French computer science laboratories (INRIA, CNRS) focused on enhanced communication services;
- to evaluate the deployment cost of chosen computing intensive and data-intensive applications and to estimate the performance gain they may obtain over the grid.

This national scale platform was the first initiative of this scale in France. Pascale Vicat-Blanc was coordinating the scientific efforts of this national project. In particular, she coordinated the project review, the project workshop and official demonstrations. RESO conducts also specific researches on Grid High performance networking and on active network services for middleware and Grid applications flows. RESO participated in the adaptation of the Madeleine software to the Globus and eToile Grid middleware. A strong collaboration within the VTHD++ project permits to test and tune the VTHD Network services like DiffServ.

8.2.2. RNRT VTHD++

Participants: Fayçal Bouhafs, Fabien Chanussot, Laurent Lefèvre, Congduc Pham, Pascale Vicat-Blanc Primet.

(2002-2004) : Laurent Lefèvre is responsible of Work Package 4 on "High performance active networks around VTHD backbone". Supported by RNRT, funding : 1 Engineer for 3 years. Software and delivrables provided during this project are available at http://www.ens-lyon.fr/LIP/RESO/Projects

8.2.3. RNRT Temic

Participant: Laurent Lefèvre.

(2003-2006) The RNRT Temic project is focused on providing solutions for collaborative management solutions of large and complex industrial process. In this project, RESO provides dynamic and adaptative networking solutions for efficiently supporting heterogeneous data streams and equipments. Experiments and platforms based on active and programmable network technology will be designed. Funding : 1 Engineer for 2 years

8.2.4. ACI Grid GRIPPS

Participants: Antoine Vernois, Pascale Vicat-Blanc.

(2003-2004) : RESO studies the problem of quality of service and end-to-end performance for genomic applications. A data intensive use case is developed and evaluated in the context of the eToile testbed.

8.2.5. ACI Grandes Masses de Données GridExplorer

Participants: François Echantillac, Olivier Glück, Cyril Otal, Pascale Vicat-Blanc Primet.

(2003-2006) : The aim of this project is to create a large scale grid and network emulator. RESO is involved in the design of the platform and is interested in designing a high performance transport protocol test methodology in this environment. EWAN [35], our high performance network emulator, is one of the

main RESO contributions to this project. Pascale Vicat-Blanc is responsible of the network theme. RESO has participated to the definition of the architecture and technical choices of cluster hardware.

8.2.6. GRID5000

Participants: Olivier Glück, Stéphane D'Alu, Brice Goglin, Marc Herbert, Julien Laganier, Laurent Lefèvre, Pascale Vicat-Blanc Primet, Jean-Christophe Mignot.

(2003-2005) : RESO is participating in the design of the *Ecole Normale Supérieure* site belonging to the experimental Grid platform GRID5000. We are particularly interested in building and collaborating in this national initiative for research and development of our innovative communication, transport and network services. We are also focusing on long distance networking issues of this national project within the CNRS AS *enabling Grid5000*.

ENS Lyon is involved in the GRID'5000 project, which aims at building an experimental Grid platform gathering eight sites geographically distributed in France. ENS Lyon hardware contribution is done for now by two distinct set of computers. The first unit, which is mainly intended for network emulation, is composed of 13 single processor SunFire V60x equipped with 2 Gb of memory and 3 Gigabit NICs each. The second unit consists of 61 2Ghz biprocessor Opteron IBM e325 (56 nodes, 1 gateway, 2 servers and 2 frontend), they are equipped with 2 Gb of Memory and 80Gb of disk each, 2 Gigabit NICs including one dedicated to administration, furthermore each server is also equipped with 584 Gb of storage provided by scsi disks. Network interconnection is realized using Ethernet Gigabit Foundry FES X448 switches, and FastEthernet switches for the management network; it is also expected, in the near future to have a Myrinet interconnection.

The operational status of Lyon's part of Grid'5000, is as follow. For the hardware, Foundry switches and IBM computers have been upgraded to the latest firmware. For the security point of view a firewall has been configured to run on the gateway to provide an isolation from the computers which are not part of Grid'5000, and a set of proxy to render basic services (DNS, NTP) have been activated. Finally for the users point of view, they have a set of computers whose clocks are synchronized (by NTP), where they have a uniform login/account handled by an LDAP server (previously done using NIS), and a common home directory delivered by NFS, the frontend provide them with the necessary compilation tools for the AMD64 architecture (optimized PathScale compiler is available), the currently available distribution on the node are Debian or Gentoo, which run using the native 64bits mode.

RESO has been strongly involved during this year in the design of the national prototype platform of GRID'5000 and in the choices of network components and architecture. Pascale Vicat-Blanc Primet is member of the national committee (comité de pilotage) of GRID'5000, co-responsible of the Lyon site with Frederic Desprez, and coordinates networks aspects with Renater and RMU, Lyon's metropolitan network. Olivier Glück, Stéphane D'Alu and Jean-Christophe Mignot are members of the national technical committee of GRID'5000. Actual funding: 350K euros

8.3. European actions

8.3.1. European DATATAG project

Participants: Pascale Vicat-Blanc, François Echantillac, Mathieu Goutelle, Marc Herbert.

IST-2001-32459 (CERN/INRIA/UvA/PPARC) Research and Technological Development for an International Grid Interconnection (2002-2004). RESO studies protocols of high performance data transport and quality of service provided by EDS on a long distance high performance backbone. A method and tool for hop by hop capacity estimation, TraceRate, has been designed, developed and experimented. Funding: 124K euros

8.3.2. Programmes d'Actions Intégrées Amadeus with Linz Univ., Austria

Participant: Laurent Lefèvre.

RESO is involved in a long term collaboration (1999-2000, 2001-2003, 2004) with University of Linz, Austria (Prof. J. Volkert team) on the field of "Deporting services on Network Programmable cards". Supported

by French Ministry of Foreign affairs. During 2004, RESO has hosted Dieter Kranzlmuller for a 3 weeks period [26].

8.4. International actions

8.4.1. NSF-INRIA with Aerospace Organization

Participant: Laurent Lefèvre.

A NSF-INRIA project is running with Aerospace Organization-USA (C. Lee team) on support of programmable networks for Grid middleware and overlays. (2004-2006).

8.4.2. AIST Grid Technology Research Center

Participants: Pascale Vicat-Blanc Primet, François Echantillac, Olivier Gluck.

After the first France-Japan Grid workshop in Paris (Mars 2004), INRIA RESO team and AIST GTRC group decided to work together. Both team focus their activities in High Performance GridNetworking area. AIST GTRC networking group is studying approaches that activate and use some intelligence in a dedicated equipment in the path, named GtrcNet1 . Two GtrcNet1 equipments have been installed within the GRID5000 node in Lyon. Both our teams adopt the same type of solutions based on IP technology, and exploiting some "intelligence" within the network (i.e., in network interface cards or in programmable equipments located in edge networks) to tackle the same kind of problems: high end-to-end throughput, performance control and measurement. A Memorandum of Understanding has been signed between INRIA and AIST GTRC in july 2004. A “Programme d’Actions Integrees” SAKURA project has recently been accepted for the 2005-2007 period.

8.5. Visitors

8.5.1. Collaboration with Univ Linz, Austria

Participants: Dieter Kranzlmuller, Laurent Lefèvre.

We have a long term collaboration with GUP, Univ. Linz on interactions between programmable network cards and tools for monitoring clusters and distributed applications. RESO has hosted D. Kranzlmuller for 3 weeks as an invited researcher from 1/11/2004 to 21/11/2004.

8.5.2. Collaboration with AIST GTRC, Japan

Participants: Tomohiro Kudoh, Yuetsu, Pascale Vicat-Blanc, François Echantillac.

RESO has hosted Dr Tomohiro Kudoh and N. Yuetsu for 1 week as invited researchers from 1/10/2004 to 8/10/2004 to work on the configuration of GtrcNet1 equipment and its integration within the eWAN software.

9. Dissemination

9.1. Conference organisation, editors for special issues

- Pascale Vicat-Blanc, as a co-chair of the Global Grid Forum’s Data Transport Research Group organized DT-RG session in Berlin (March 2004) and gives a talk to the GHPN session in Hawai (June 2004).
- Pascale Vicat-Blanc is guest editor with Jean-Phillipe Martin-Flatin of a special issue of the international Future Generation Computer Systems (FGCS) Journal on "High Performance Protocols and Grid services" (to appear in december 2004)

- Pascale Vicat-Blanc is member of program committees : IEEE HPDC2004, GRIDNETS2004, GNEW2004, IEEE CCGRID GAN2004, Pfldnet04. She has been reviewer for international journal and conferences : Communication Network Journal, Parallel letter, JPDC, Calculateurs Parallèles, TSI, IEEE ICC04, Pfldnet04, CFIP04, IEEE INFOCOM04, IEEE Supercomputing04.
- C. Pham is co-editor with B. Tourancheau of a special issue of FGCS on "Grid Infrastructures: Practice and Perspectives".
- Laurent Lefèvre and Pascale Vicat-Blanc have co-organized the Workshop "Grid and Advanced Networks" (GAN'04) in CCGrid 2004, Chicago.
- Laurent Lefèvre has co-organized "CCGSC 2004" : 6th Cluster and Computational Grid" Chateau Faverges de la Tour, France, Sept. 2004, co-organization with J. Dongarra and B. Tourancheau.
- Laurent Lefèvre is organizer and *program chairman* of workshops series "Distributed Shared MemOry on Clusters" DSM2004 (Chicago) within IEEE International Symposium on Cluster Computing and the Grid (CCGrid).
- Laurent Lefèvre is *Steering Committee* member of CCGrid conference.
- Laurent Lefèvre is member of the following Program Comittee (i) International journals : Parallel and Distributed Computing Practice (PDCP), Journal of Parallel and Distributed Computing (JPDC), FGCS Advanced Grid Techology, (ii) International conferences: AGridM2004, Grid 2004, LCR2004, MediaNet'2004, EuroPVMPI 2004, IWAN 2004, IEEE CCGrid 2004, HPCS2004, PPGaMS'04, APGAC04

9.2. Graduate teaching

- **2004**: P. Vicat-Blanc Primet
  Advanced protocols for high speed networks. *Réseaux avancés et leurs protocoles*.
  Master Research (Ecole Normale Supérieure de Lyon, University Claude Bernard Lyon 1), lecture: 28h/year.
- **2004**: C. Pham
  New Technologies for the Internet. *Les nouvelles technologies de l'Internet*.
  DEA DISIC (University Claude Bernard Lyon 1, INSA), lecture: 8h/year.
- **since 1998**: C. Pham
  Performance Evaluation and Simulation. *Evaluation de performance et simulation*.
  Master 2 SIR, formely DESS IIR Réseaux, and (University Claude Bernard Lyon 1), lecture: 10h/year, experimental work: 40h/year.
- **since 1998**: P. Vicat-Blanc Primet
  Wide Area Networks. *Réseaux grandes distances*.
  master CCI, formely DESS CCI, (University Claude Bernard Lyon 1), lecture: 20h/year.
- **since 2004**: C. Pham
  High-Speed Networks and QoS. *Réseaux haut-débit et QoS*.
  Master 2 SIR, formely DESS IIR Réseaux, (University Claude Bernard Lyon 1), lecture: 20h/year.
- **since 2003**: O. Glück
  Internet and programming on the Web.
  DESS IIR Réseaux (Université Claude Bernard Lyon 1), lecture 10h.
- **since 2004**: O. Glück
  Client/Server Model, Internet Applications, Network and System Administration.
  Master 2 SIR, formely DESS IIR Réseaux, (University Claude Bernard Lyon 1), lecture 30h, others 30h.

9.3. Miscelleneous teaching

- **since 1998**: C. Pham
  Communication Networks.
  Master 1 Informatique, formely Maîtrise d'Informatique, (Université Claude Bernard Lyon 1), lecture: 30h/year.
- **since 1991**: P. Vicat-Blanc Primet
  Computer Networks.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year.
- **since 2002**: P. Vicat-Blanc Primet
  Multimedia Communications.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year
- **since 2003**: P. Vicat-Blanc Primet
  High Speed Networks and Quality of Service.
  Maitrise IUP Réseaux (Université Claude Bernard Lyon1), 20h lectures/year.
- **since 2002** : L. Lefèvre
  Réseaux, Internet et outils associés.
  Maitrise Informatique (Université Antilles Guyane, Pointe à Pitre), 40h eq TD/an.
- **since 2003**: O. Glück
  LAN and WAN Networks.
  Licence IUP Réseaux (Université Claude Bernard Lyon 1), lecture 30h, others 30h.
- **since 2003**: O. Glück
  Computer Networks and Applications.
  Licence IUP Réseaux (Université Claude Bernard Lyon 1), lecture 30h, others 30h.
- **2004**: O. Glück
  Computer Networks.
  Licence Informatique, (University Claude Bernard Lyon 1), lecture 30h, others 30h.
- **2004**: O. Glück
  Programming on the Web.
  Master 1 SIR, formely Maîtrise d'Informatique, (University Claude Bernard Lyon 1), lecture 15h, others 15h.

9.4. Animation of the scientific community

Pascale Vicat-Blanc

- Within the Global Grid Forum, standardization entity for grid middleware, is co-chair of the Data-Transport Research Group. RESO is also active in the Network Monitoring Working Group as in the Grid High Performance Networking.
- Within the Grid5000 project, is a member of the steering committee (piloting).
- organized a national day on "Networks in Grid5000 project" in CNRS, february 2004.
- is a member of the steering committee of the RTP1 (Réseau) du CNRS.
- is a member of the INRIA delegation in Japan for the France-Japon workshop on Grid technology and participates to the setup of collaborations with the NAREGI project, the AIST Gtrc, the Tokyo Institute of Technology (Titech) and the Osaka University.

9.5. Participation in boards of examiners and committees

- Pascale Vicat-Blanc
  - participated to the board of examiners for recruitments of *Chargés de Recherche CR2* of the Rhône-Alpes INRIA research unit in 2004.
  - has been member of the board of examiners of *DEA d'Informatique Fondamentale de Lyon*.
  - has been reviewer (rapporteur) and member of the PhD thesis jury of David Fuin from Université de Besançon and Florestan de Belleville du Laboratoire TeSA (INT Toulouse) in December 2004.
- Laurent Lefèvre is member of the "commissions de spécialistes de 27ème section" of University Jean Monnet, Saint-Etienne and University Antilles Guyane, Pointe à Pitre.
- Congduc Pham has been reviewer of the PhD thesis jury of R. Beuran from University of St-Etienne and University of Bucarest, July 2004.
- Olivier Glück is a member of
  - the "commissions de spécialistes 27ème section" of University Claude Bernard Lyon 1 and University Pierre et Marie Curie Paris 6.
  - the "conseil d'UFR Informatique" of University Claude Bernard Lyon 1.

9.6. Seminars, invited talks

- Pascale Vicat-Blanc has been invited to give a tutorial at the Internet Nouvelle Generation summer school in Obernay (june 2004), a tutorial in the GridUse summer school in Metz (june2004), a seminar at the GridExplorer emulation meeting in November 2004, to give a talk at the CCGSC conference (Lyon September 2004).
- C. Pham has been invited to give a talk "New Internet and Networking Technologies and their Application to Computational Science" at the COSCI 2004 conference, University Bach Khoa, Ho Chi Minh City, Vietnam, 3-5 March 2004, a seminar "Advanced Networking: New Trends in Internet Technologies" at University Bach Khoa, Ho Chi Minh City, Vietnam, November 2004, a tutorial "New Internet and Networking Technologies for Grids and High-Performance Computing", HiPC 2004, Bangalore, India December 22nd, 2004.

10. Bibliography

Major publications by the team in recent years

[1] B. BOUAHFS, J. GELAS, L. LEFÈVRE, M. MAIMOUR, P. C., P. PRIMET, B. TOURANCHEAU. *Designing and Evaluating An Active Grid Architecture*, in "Future Generation Computer System", To appear in February 2005, vol. 21, n° 2, 2004, http://bat710.univ-lyon1.fr/~cpham/.

[2] J.-P. GELAS, S. EL HADRI, L. LEFÈVRE. *Towards the Design of an High Performance Active Node*, in "Parallel Processing Letters", vol. 13, n° 2, jun 2003.

[3] M. GOUTELLE, P. PRIMET. *Study of a non-intrusive method for measuring the end-to-end capacity and useful bandwidth of a path*, in "Proceedings of the 2004 International Conference on Communications, Paris, France", IEEE Communication Society, June 2004.

[4] L. LEFÈVRE, J.-P. GELAS. *Programmable Networks for IP Service Deployment*, A. GALIS, S. DENAZIS, C. BROU, C. KLEIN (editors)., chap. Chapter 14 on "High Performance Execution Environments", Artech House Books, UK, may 2004, p. 291-321.

[5] M. MAIMOUR, C. PHAM. *AMCA: an Active-based Multicast Congestion Avoidance Algorithm*, in "Proceedings of the 8th IEEE Symposium on Computers and Communications (ISCC 2003), Antalya, Turkey", Best paper award, June 2003.

[6] M. MAIMOUR, C. PHAM. *DyRAM: an Active Reliable Multicast framework for Data Distribution*, in "Journal of Cluster Computing", vol. 7, n° 2, 2004, p. 163-176, http://bat710.univ-lyon1.fr/~cpham/Paper/ccj04.pdf.

[7] J.-P. MARTIN-FLATIN, P. VICAT-BLANC PRIMET. ELSEVIER (editor). *special issue on High Performance Networking and Grid Services. The DataTAG project*, December 2004.

[8] G. MONTENEGRO, B. GAIDIOZ, P. PRIMET, B. TOURANCHEAU. *Equivalent Differentiated Services for AODVng*, in "ACM SIGMOBILE Mobile Computing and Communications Review", vol. 6, n° 3, July 2002, p. 110-111.

[9] P. PRIMET, B. GAIDIOZ, M. GOUTELLE. *Approches alternatives pour la différenciation de services IP*, in "TSI: Techniques et Sciences Informatiques, special issue Nouveaux Protocoles pour l'Internet", October 2004, p. 651-674.

[10] P. VICAT-BLANC PRIMET, F. BONNASSIEUX, R. HARAKALY. *Network monitoring in the European DataGRID project*, in "International Journal of High Performance Computing Applications", vol. 18, n° 3, January 2004, p. 293-304.

Doctoral dissertations and Habilitation theses

[11] E. LEMOINE. *Nouvelles fonctions dans les interfaces de communication pour l'augmentation des performances réseau des machines multi-processeur*, Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, jul 2004.

Articles in referred journals and book chapters

[12] F. BOUHAFS, J. GELAS, L. LEFÈVRE, M. MAIMOUR, C. PHAM, P. PRIMET/VICAT-BLANC, B. TOURANCHEAU. *Evaluating and Experimenting An Active Grid Architecture*, in "Future Generation Computer System", To appear in February 2005, vol. 21, nº 2, 2004, http://bat710.univ-lyon1.fr/~cpham/Paper/.

[13] J.-P. GELAS, L. LEFÈVRE. *Flexibilité et performance dans les routeurs actifs logiciels pour un support efficace des services déployés sur des réseaux gigabits*, in "Annales des Télécoms", jun 2004, p. 645-685.

[14] L. LEFÈVRE, J.-P. GELAS. *Programmable Networks for IP Service Deployment*, A. GALIS, S. DENAZIS, C. BROU, C. KLEIN (editors)., chap. Chapter 14 on "High Performance Execution Environments", Artech House Books, UK, may 2004, p. 291-321.

[15] M. MAIMOUR, C. PHAM. *DyRAM: an Active Reliable Multicast framework for Data Distribution*, in "Journal of Cluster Computing", vol. 7, nº 2, 2004, p. 163-176, http://bat710.univ-lyon1.fr/~cpham/Paper/ccj04.pdf.

[16] P. PRIMET, B. GAIDIOZ, M. GOUTELLE. *Approches alternatives pour la différenciation de services IP*, in "TSI: Techniques et Sciences Informatiques, special issue Nouveaux Protocoles pour l'Internet", October 2004, p. 651-674.

[17] P. VICAT-BLANC PRIMET, F. BONNASSIEUX, R. HARAKALY. *Network monitoring in the European DataGRID project*, in "International Journal of High Performance Computing Applications", vol. 18, nº 3, January 2004, p. 293-304.

[18] P. VICAT-BLANC PRIMET. *Experiment with the Equivalent Differentiated Services model in Grid context*, in "Journal of Future Generation Computer Systems (FGCS)", available on-line at www.sciencedirect.com, vol. 1269 Elsevier Science Press, December 2004.

Publications in Conferences and Workshops

[19] A. BASSI, M. BECK, F. CHANUSSOT, J.-P. GELAS, R. HARAKALY, L. LEFÈVRE, T. MOORE, J. PLANK, P. PRIMET. *Active and Logistical Networking for Grid Computing: the e-Toile Architecture*, in "First International Workshop on Active and Programmable Grids Architectures and Components - APGAC'04, Krakow, Poland", jun 2004, p. 202-209.

[20] C. CASTELLUCIA, G. MONTENEGRO, J. LAGANIER, C. NEUMANN. *IPv6 Opportunistic Encryption*, in "Proc. of 9th European Symposium on Research in Computer Security (ESORICS 2004)", vol. 3193/2004, Lecture Notes in Computer Science (LNCS), Springer-Verlag, September 2004, p. 309–321.

[21] L. EGGERT, J. LAGANIER, M. LIEBSCH, M. STIEMERLING. *HIP Resolution and Rendezvous Mechanisms*, in "Proc. of 1st Workshop on HIP and Related Architectures", November 2004.

[22] B. GOGLIN, L. PRYLLI, O. GLÜCK. *Optimizations of Client's side communications in a Distributed File System within a Myrinet Cluster*, in "Proceedings of the IEEE Workshop on High-Speed Local Networks (HSLN), held in conjunction with the 29th IEEE LCN Conference, Tampa, Florida", IEEE Computer Society Press, November 2004, p. 726-733.

[23] B. GOGLIN, L. PRYLLI. *Transparent Remote File Access through a Shared Library Client*, in "Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04), Las Vegas, Nevada", vol. 3, CSREA Press, June 2004, p. 1131-1137.

[24] M. GOUTELLE, P. PRIMET. *Study of a non-intrusive method for measuring the end-to-end capacity and useful bandwidth of a path*, in "Proceedings of the 2004 International Conference on Communications, Paris, France", IEEE Communication Society, June 2004.

[25] M. HERBERT, P. PRIMET. *A Case for Queue-to-Queue, Back-Pressure-Based Congestion Control for Grid Networks*, in "Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04)", P. H. R. ARABNIA (editor)., CSREA Press, June 2004.

[26] L. LEFÈVRE, D. KRANZLMULLER, M. MAURER. *Incremental Monitoring on Programmable Network Interface Cards*, in "The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications - PDPTA'04, Las Vegas, USA", jun 2004.

[27] E. LEMOINE, C. PHAM, L. LEFÈVRE. *Packet Classification in the NIC for Improved SMP-based Internet Servers*, in "Proceedings of IEEE 3rd International Conference on Networking (ICN'04), Guadeloupe, French Caribbean", March 2004, http://bat710.univ-lyon1.fr/~cpham/Paper/ICN04.pdf.

[28] D. M. LOPEZ-PACHECO, C. PHAM. *Performance of TCP (New Reno, Westwood), HSTCP and XCP in high-speed, highly variable-bandwidth environments*, in "Proceedings of IEEE 3rd International Conference on Network Protocols (ICNP'04), Berlin, Germany", Student Poster Session, October 2004, http://bat710.univ-lyon1.fr/~cpham/Paper/ICNP04.pdf.

[29] P. V.-B. PRIMET, J. MONTAGNAT, F. CHANUSSOT, M. GOUTELLE. *Network Quality of Service in Grid environments: the QoSinus approach*, in "IEEE Cluster Conference", poster paper, September 2004.

[30] P. VICAT-BLANC PRIMET. *Dynamic control and flexible management of quality of service in Grids: the Qosinus approach*, in "in proceeding of the 1st IEEE Broadnet Conference, GridNets workshop, San José", October 2004.

[31] J. ZENG, P. VICAT-BLANC PRIMET. *An Overlay Infrastructure for Bulk Data Transfers in Grids*, in "in Proceedings of the Third International Workshop on Protocol For Very Long Distance Fat Networks, Pfldnet2005", to appear, vol. Lyon, 2004.

Internal Reports

[32] B. GOGLIN, O. GLÜCK, P. V.-B. PRIMET. *Accès optimisés aux fichiers distants dans les grappes disposant d'un réseau rapide*, Also available as Research Report RR-5458, INRIA Rhône-Alpes, Research Report, n° RR2004-56, LIP, ENS Lyon, Lyon, France, December 2004, http://www.inria.fr/rrrt/rr-5458.html.

[33] E. LEMOINE, C.-D. PHAM, L. LEFÈVRE. *A new mechanism for Transmission Notification on SMP*, Technical report, n° RT-0295, INRIA, may 2004, http://www.inria.fr/rrrt/rt-0295.html.

[34] P. V.-B. PRIMET, J. MONTAGNAT, F. CHANUSSOT, M. GOUTELLE. *Flexible Management and Dynamic Control of Network Quality of Service in Grid environments : the QoSinus approach*, Research Report, n°

RR-5083, INRIA, Lyon, France, January 2004, http://www.inria.fr/rrrt/rr-5083.html.

[35] P. VICAT-BLANC, O. GLÜCK, C. OTAL, F. ECHANTILLAC. *Emulation d'un nuage réseau de grilles de calcul: eWAN*, Research Report, nº RR2004-59, LIP, ENS Lyon, Lyon, France, December 2004, http://www.ens-lyon.fr/LIP/Pub/Rapports/RR/RR2004/RR2004-59.pdf.

Miscellaneous

[36] J. LAGANIER, L. EGGERT. *Host Identity Protocol (HIP) Rendezvous Extensions*, Work in progress, expired in April 2005, October 2004, IETF HIP WG Internet Draft draft-ietf-hip-rvs-00.txt.

[37] J. LAGANIER, L. EGGERT. *Host Identity Protocol (HIP) Resolution and Rendezvous Problem Description*, Work in progress, expired in April 2005, October 2004, IETF Internet Draft draft-eggert-hiprg-rr-00.txt.

[38] L. LEFÈVRE, P. NEIRA AYUSO. *Fault tolerant library for designing stateful network equipments*, Poster INRIA Booth, Supercomputing 2004, Pittsburgh, USA, nov 2004.

[39] P. NIKANDER, J. LAGANIER. *Host Identity Protocol (HIP) Domain Name System (DNS) Extensions*, Work in progress, expired in April 2005, October 2004, IETF HIP WG Internet Draft draft-ietf-hip-dns-00.txt.

[40] V. SANDER, J. CROWCROFT, F. TRAVOSTINO, P. VICAT-BLANC PRIMET, C. PHAM, AL.. *Networking issues of GRID Infrastructures*, http://forge.gridforum.org/projects/ghpn-rg/document/draft-ggf-ghpn-netissues-0/en/1/draft-ggfghpn-netissues-1.pdf, GGF Informational Document.

[41] M. WELTZ, M. GOUTELLE, E. HE, P. VICAT-BLANC PRIMET, E. AL.. *Survey of Protocols other than TCP*, Work in progress, December 2004, GGF Informational Document.

Bibliography in notes

[42] D. X. W. CHENG JIN, S. H. LOW. *FAST TCP: motivation, architecture, algorithms, performance*, in "IEEE Infocom", March 2004.

[43] V. FIRIOUS, J. L. BOUDEC, D. TOWSLEY, Z.-L. ZHANG. *Theories and Models for Internet Quality of Service*, in "IEEE", May 2002.

[44] S. FLOYD. *HighSpeed TCP for Large Congestion Windows*, in "Internet draft, draft-floyd-tcp-highspeed-01.txt, work in progress, 2002", 2002, work in progress.

[45] S. FLOYD, V. JACOBSON. *Link-sharing and Resource Management Models for Packet Networks*, in "IEEE/ACM Transaction on Networking", 4, vol. 3, August 1995.

[46] I. FOSTER, M. FIDLER, A. ROY, V. SANDER, L. WINKLER. *End to end Quality of Service for High End applications*, in "Computer Communications, special Issue on Network Support for Grid Computing", 2002.

[47] I. FOSTER, C. KESSELMAN. *The Grid : Blueprint for a new Computing Infrastructure*, in "Morgan Kaufmann Publishers Inc.", 1998.

[48] T. KELLY. *Scalable TCP: Improving Performance in Highspeed Wide Area Networks*, in "Protocol for Long Distance Networks Conference", n° Pfldnet-1, February 2003.

[49] V. SANDER. *Networking issues of GRID Infrastructures*, in "GRID Working Draft of the GRID High-Performance Networking Research Group, Global GRID Forum", 2003.

[50] D. SIMEONIDOU. *Optical Network Infrastructure for Grid*, in "Grid Working Draft of the Grid High-Performance Networking Research Group, Global GRID Forum", 2003.