INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team reso

Optimized protocols and software for high performance networks

Grenoble - Rhône-Alpes

THEME NUM

Table of contents

1. Team

Head of project-team
Pascale Vicat-Blanc Primet [Research Director (DR2) Inria, HdR]

Administrative Assistant
Danielle Bianchetti [Secretary ENS – 20%]
Sylvie Boyer [Secretary (SAR) INRIA – 10%]

Staff member INRIA
Paulo Gonçalves [Research Associate (CR1) Inria]
Laurent Lefèvre [Research Associate (CR1) Inria]

Staff member Université Claude Bernard Lyon1 (UCB)
Jean-Patrick Gelas [Maître de conférences]
Olivier Glück [Maître de conférences]
Isabelle Guérin-Lassous [Professeur, HdR]

Project technical staff
Jean-Christophe Mignot [Research Engineer CNRS – 40%]
Aurélien Cedeyn [ENS Engineer – Grid5000]
Mathieu Imbert [Research Engineer INRIA – 40%]
Pierre Bozonnet [Expert Engineer INRIA –contrat Alcatel + CARRIOCAS]
Marcelo Pasin [Expert Engineer INRIA-FP6 EC-GIN]
Damien Ancelin [Expert Engineer INRIA -FP6 EC-GIN]
Olivier Mornard [Expert Engineer INRIA– ANR HIPCAL]

Postdoctoral position
Chen Cheng [Postdoc with IN2P3 CNRS]

Ph. D. students
Narjess Ayari [PhD student with France Telecom R-D - CIFRE - 2005/2008]
Romaric Guillier [PhD student, ENS - 2006/2009]
Ludovic Hablot [PhD student, MENRT - 2006/2009]
Patrick Loiseau [PhD student, ENS - 2006/2009]
Dino Lopez Pacheco [PhD student - Mexican Goverment Grant- 2004/2008]
Sébastien Soudan [PhD student, MENRT - 2006/2009]
Rémi Vanier [PhD student, INRIA - 2006/2009]
Dinil Mon Divakaran [PhD student, INRIA - 2007/20010]

Student internship
Anne-Cécile Orgerie [Internship ENS]
Walid El Dahabi [Internship Univ. Claude Bernard Lyon1]

Visiting scientists
Paul Roe [Queensland Univeristy of Technology, Brisbane, Australia from Sept. 2007 to Dec. 2007]

2. Overall Objectives

2.1. Project-team presentation overview

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 twenty members in average, including six 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" . In this direction, 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.

2.2. Context

Wavelengths multiplexing and wavelengths switching techniques on optical fibers allow core network infrastructures to rapidly improve their throughput, reliability and flexibility. Links of 40 gigabits per second will be soon available. New technologies like 10 Gigabit/s Ethernet or 10Gigabit/s Myrinet is also driving the increase of bandwidth in local area networks. These improvements have given the opportunity to create high performance distributed systems called "computational and data grids" that aggregate storage and computation resources into a virtual and integrated computing environment. Grid computing is a promising technology harnessing distributed resources into virtual organizations (or communities) for the future resource intensive scientific, business and domestic applications. On an other hand, the volumes of heterogeneous data that are produced by various distributed sources (sensor networks, sophisticated instruments and end users) and managed in distributed data centers are rapidly increasing. Complex computational models performed on supercomputers produce petabytes of data, which have to be accessed and analysed by various user groups. Moving such 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. One of the key challenge for large deployment of Grid technology is the provisionning of a secure, flexible, transparent and high performance transport infrastructure for data access and processing. Consequently, future high-speed optical networks are addressed not only to support the accelerating and dynamic growth of data traffic but also the new emerging network requirements such as fast and flexible provisioning, QoS levels, and fast recovery procedures of such data intensive computing applications. Enabling ultra high performance machine to machine communications lead then to new bandwidth sharing paradigms. Although grids theoretically offer solutions for resources aggregation, predictable and high performance for applications may be hard to obtain due to the improperness of bandwidth sharing paradigms (fairness, best effort, no QoS) , communication protocols and software. The fact that processors, memory, bus and disc speeds, involved into the protocol processing chain do no scale with network speeds is also an issue. In order to deliver this emerging traffic in a timely, efficient, and reliable manner over long distance networks, several issues such as quality of service, security, traffic metrology , traffic modeling and network resource scheduling have to be investigated.

2.3. Research area

To address some of these issues, 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 algorithms for efficient and customizable transport and QoS for heterogeneous traffic at very highspeed.

Last year, two new research topics complementing our second research axis have been introduced:

- Traffic metrology and statistical inference
- Overlay networks

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. This also concerns the study of the subtile coupling of computing and communication. Indeed protocol processing at very high speed requires the introduction of new approaches to distribute the load on different entities or to simplify the per-packet or per bit operations. Symetrically, computing on networked ressources requires adequate protocols for data exchanges. One of the main direction we follow is the exploration of the potential 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. This in some sense corresponds to offloading. 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 behavior, the design of adaptive algorithms dedicated to the expressed flow needs, definition and introduction in the network of end-to-end protocol-oriented mechanisms, making the interaction between packet processing and forwarding smooth and efficient.

The traffic metrology and statistical inference topic was recently brought into the RESO team activities (April 2006) and deals with the metrological aspects of grid and distributed system traffic. Intended as a diagnosis tool to serve our second research axis, we are led to first define pertinent metrics to get an instantaneous snapshot of the network, and to assess the corresponding quality of service. In a second step, we foresee to develop methodological approaches, along with the necessary tools, to fit the measures with realiable (in the widest sense) statistical models. Ultimately, based on these models, we plan to infer a short time forecast of the network capacities (instantaneous bandwidth, latency, losses,...), which will steer the current transport protocol to automatically adapt to the context.
This activity comes along with a necessary effort devoted to develop an experimental metrology platform, relying on the Grid5000 infrastructure.

The "overlay networks" topic center on resource virtualisation and resource sharing optimization problems that arise in overlay networks. We mainly focus on routing and load balancing problems. The goal is to provide distributed algorithms for these problems. We also explore the concept of virtualization of heterogeneous interconnected resources (network devices, computer, storage spaces, instruments...) in such context.

2.4. Application domains

RESO applies its research to the domains of high performance computing and to Grid communications. Grid computing is a promising technology that brings together large collection of geographically distributed resources (e.g., computing, storage, visualization, etc.) to build on demand very high performance computing environments for compute and data-intensive applications. These large scale cybernetic infrastructures gain increasing attention from a broad range of actors: from research communities to computer providers, large companies, and telecommunication operators (telcos). Whereas grids have been widely in use in the scientific community, they are now on the verge of moving into the commercial environment. American, european and japanese telcos plan to move forward grid computing. Different scenarios for telcos can be envisioned: telcos may (1) deploy grids internally, e.g. for rapid dynamic service provisioning to new customers; (2) link different sites via VPNs; (3) act as a service broker. Which senario will be developed remains an open question which we explore with our industrial partners OrangeLabs and Alcatel-Lucent but also our japanese collaborators.

Researches conducted these last years reveal that grid technology raise new challenges in terms of network optimisation as well as of protocol architecture and of transport paradigms. A broad deployment of the grid technology can modify and influence the design of the future Internet as other emerging communicating applications.

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. 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 is difficult to guarantee or predict. 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 simplify the programming and to optimize the execution of their communication parts while fully exploiting the capacities of the evolving networking infrastructure.

2.5. Methodology

The RESO approach relies on a methodology based on a three steps cycle: 1) a fine analysis of limitations encountered in existing protocols (TCP/IP), 2) the exploration of disruptive solutions, 3) the theoretical and experimental evaluation of these proposals. This research focuses an heavily ossified research object (TCP/IP protocols) and lies between a challenging emerging application domain on a specific network context. These factors induce a close interaction with both the application level and the underlying network level as well as a deep technical and scientifical knowledge of protocols and network equipments. The methodology is then based on a continous study of the high end and original requirements and on experimental evaluation of the functionalities and performance of emeging dedicated high speed infrastructures. RESO gathers expertise in advanced high performance local and cluster area networks protocols, in distributed systems and algorithmics, in protocol and protocol architecture design, in long distance networking, in time series analysis and in statistical inference. This background work provides the basis for innovative protocols and software design. Moreover, we implement and experiment our proposals on real, emulated local or wide area testbeds with real conditions and large scale applications.

2.6. 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 studies high speed network characteristics, grid application requirements, creates open source code, distributes it to the research community for evaluation and usage and help in shortening the wizard gap between network experts and novices. The long term goal is also to contribute to the evolution of protocols, standards and networking equipments, prompting the introduction of metrology as an intrinsic component of high-speed networks. An important effort is naturally dedicated to the dissemination of these new approaches.

2.7. Summary of the main contributions of the team in 2007

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

2.7.1. Direction 1: Optimized communication software and equipments

- Study and optimizations of Message Passing Interface implementations for Grid platforms : MPICH-Madeleine, GridMPI, OpenMPI. The experiments were conducted on the national GRID'5000 testbed.

- Design and development of a high performance autonomic network node adapted to industrial context (IAN2). Proposition of adapted autonomic services;
- Design of a fault tolerant and highly available architecture for clusterd network equipments;

2.7.2. Direction 2: end-to-end transport and service differentiation

- E2E Transport: Contribution to the design of a user oriented test suite towards a benchmark for new transport protocols evaluation and comparison;
- E2E Transport: Implementation of optimization algorithms for network resource sharing and delay constrainted flow scheduling in very high speed networks.
- E2E Transport: Design of end point time and rate limitation solutions for bandwidth allocation profile enforcement.
- E2E Transport: Study of differentiated chanels provisioning in Overlay Networks
- Metrology: Design and development of a fine grain traffic capture and traffic analysis system dedicated to high speed links.
- Metrology: Comparison of sampling methods for characterizing heavy tailed distributions in high speed networks traffic.
- Overlays: Network and system virtualisation for "on-demand" overlays creation. Evaluation of virtualisation cost of end to end communications.
- Overlays: Distributed algorithms for bandwidth sharing in mobile or very high speed environments.

2.7.3. Grid Network services and applications

- Design of a grid service for bulk data transfert scheduling in high performance Grid environments;
- Study of network requirements of business grid applications in the context of an ultra high speed network (Carriocas)
- Pursue the collaborations for the development of the GRID5000 international optical interconnections to Netherland (DAS3) and Japan (Naregi) in collaboration with RENATER;
- Traffic monitoring of Grid5000 at packet resolution to assess and enhance flow-level Grid Network performance;
- Design of a metrology infrastructure for performance and traffic monitoring in Grid5000.
- Design of the HIPerNet software for network-aware virtual custer management tool.
- Study and development of a network-centric P2P resource discovery system.
- Design of a lambda path reservation, scheduling and virtualisation system

3. Scientific Foundations

3.1. Optimized communication software and equipments

Participants: Narjess Ayari, Pierre Bozonnet, Jean-Patrick Gelas, Olivier Glück, Laurent Lefèvre, Pascale Vicat-Blanc Primet, Jean-Christophe Mignot, Ludovic Hablot, Sébastien Soudan.

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. In a cluster environment, high performance applications running on high-speed interconnects require both efficient communication between computing nodes and fast access to the storage system. In a grid environment, two key points in the communication layers need to be taken in consideration in order to execute efficiently high performance applications: the heterogeneity of high-speed interconnects composing the grid and the Wide Area Network used to achieve inter-site communications. We explore new mechanisms to improve the application performance when it executes on the grid. We study how a MPI application can benefit, during one execution, of several high-speed networks at the same time. In particular, it implies to find a way to communicate efficiently between these heterogenous interconnections. We also explore how to keep good performance execution when long-distance communications are necessary because the application is launched on multiple sites of the grid.

In this research axis, we explore the design of autonomic network equipments able to dynamically deploy adapted services. These equipments have been used in industrial context (TEMIC project, 3DDL collaboration). In order to support network functions in the embedded equipments, we propose a high performance autonomic network environment execution architecture ($\text{Tamanoir}_{\text{embedded}}$ software suite). High availability, fault tolerance and scalability issues of cluster-based network equipments have been and are currently explored.

3.2. End-to-end High performance transport

Participants: Pascale Vicat-Blanc Primet, Dino Lopez Pacheco, Laurent Lefèvre, Sebastien Soudan, Romaric Guillier, DInil Mon Divakaran.

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 networks, 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. Considering the traditional feedback loop of window based transport protocols will not scale with higher rate level under loss or congesting traffic conditions, it seems judicious to start examining alternative radical solution for end to end transport as well as for congestion control. These solutions can be based on pair to pair approaches, buffer in line or flow scheduling, fully exploiting not only the rate dimension of data transfer but also space, time and cross-layer dimensions.

One of the direction recently investigated in Grids, is the capacity of dynamically establishing overprovisionned dedicated lambda path. The optical fiber communication will be the predominant mechanism for data transmission in the core. To address the anticipated terabit demands dynamically reconfigurable optical networks are envisionned. This vision will be realized with the deployment of configurable optical components, which are now becoming economically viable. To meet the terabit challenge, network designers will enhance

core functionality by migrating to, equipped with tunable transceivers, optical crossconnects (OXC), and optical add/drop multiplexers. At the opposite side of the spectrum, dedicated high bandwidth channels are critical in large scale applications to ensure timely task completion, which in turn necessitates a high-performance control plane capable of scheduling such channels in advance. The control-plane, traditionally in the hand of telco may migrate to the users. Optical Cross-Connects (OXCs) becomes more and more, cheap, simple and controllable, Prototyping and studying the interactions of components required to accomplish the tasks of user-specified bandwidth reservation, path computation and network signaling is of importance. This year RESO starts to integrate this new technological perspective to understand how this optical component interact with electronic component and how to configure, control and tune them with end computers in the context of our associated team with AIST (Japan) and the G-Lambda Project and in collaboration with Alcatel-Lucent in the context of the CARRIOCAS project.

Finally an other important issue is flow differenciation. Indeed, it is known that flows crossing IP networks are not equally sensitive to loss or delay variations because they do not have the same utility functions and the same final usage. 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. Since several years, RESO explores various complementary or fundamentally different solutions to carry end-to-end quality of service to grid applications to assure an efficient usage of the interconnected computing resources [53] by considering the flow abstraction within the network. This approach will be further explored within the new research axis we introduce in the framework of the INRIA-BellLabs laboratory where we will lead the "Semantic networking" thema.

3.3. Metrology and Statistical inference on grids' traffic

Participants: Pascale Vicat-Blanc Primet, Paulo Gonçalves, Isabelle Guérin-Lassous, Patrick Loiseau.

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.

That is the reason why a metrology activity including data processing, statistical inference, time series and stochastic processes analysis, deemed important to embed in the main research realm of RESO. Our goal is for these analyses to become in the near future a plain component not only in the study and in the development of infrastructures and computing networks, but also in real-time resources identification and management.

Grids specificities, such as the cooperating equipments number and heterogeneity, the number of independent processes, the treatments, bandwidth and stock capacities, turn indispensable to revisit the algorithms, as well as the control and operating mechanisms, in order to reach appropriate and optimal performances.

To validate a priori hypothesis that sustain already investigated approaches (e.g. overlay, virtualizing network resources, distributing network treatments, middleware programming), we foresee to resort to metrology and to the statistical analysis of the collected data. Indeed, we believe that automatic identification of static and dynamic properties of network resources is a prerequisite for developing adequate algorithms.

To drive us in this task, we will rely on the impressive amount of studies devoted to the internet traffic analysis, and on the established results that have been obtained in the last years [60], [64], [46], [47]. For instance, we are interested in verifying if the conjecture relating long range dependance (LRD) in traffic flows with heavy tailed distributions of files sizes [64], still holds with grid networks. Difficulty dwells in a reliable estimation of density functions from loosely sampled data, and in the even more tricky one of accurate estimation of

LRD parameters and tail exponents from incomplete data sets [3]. To tackle these issues, P. Gonçalves (former Mistis (ex IS2) project member) and P. Loiseau (PhD student, Ms in Physics) supply RESO with the necessary inter-disciplinary competence in signal processing and statistics.

Finally, the great investment that has been granted to Grid5000 (and to the interconnections Grid5000-NAREGI and Grid5000-DAS 3) will profitably be used providing us with a high-performance and quite novel experimental setup to confront the proposed theoretical models with real traffic measurements.

3.4. Grid Network services and applications

Participants: Pascale Vicat-Blanc Primet, Olivier Glück, Olivier Mornard, Sebastien Soudan, Marcelo Pasin, Pierre Bozonnet.

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 [54]. 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 OGF.

- 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...
- Advance reservations

4. Application Domains

4.1. Panorama

Keywords: *Autonomic 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 [62] [61], [63]. Loss probability, important and incompressible latencies, dynamic behavior of network paths question profoundly models and technic used in parallel and distributed computing [52]. 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 OGF, the European or National projects. These activities have been done in close collaboration with 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 pfldnet and Glif communities. This year RESO joined the CARRIOCAS project which studies and implements a very high bit rate (up to 40 Gb/s per wavelength) network interconnecting super computers, storage systems and high resolution visualization device to support data and computing intensive applications in industrial and scientific domains. Our activities cover networking intelligence for high performance distributed applications. Finally, the evolution of the Internet usage pushing the convergence of communication and computation at every level confirm RESO initial vision : the network should not be seen only as a black box providing pipes between edge machines, but as a vast cloud increasingly embedding the computational and storage resources to meet the requirement of emerging applications. These resources are generally located at important crossroads and access points throughout the network. During the last few years we have seen that the distinction between packet forwarding and application processing has become blurred. The network community now starts to worry not only about forwarding packets without regard to application semantics, but is increasingly trying to exploit new functionalities within the network to meet the requirement of the application. Reciproquely, distributed systems and applications have traditionally been designed to run ?on top of? the Internet, and to take the architecture of the Internet as given. Although the convergence of communication and computation at every level appears to be natural, it is still very difficult to efficiently explore the full range of possibilities it can bring. Most of the proposals exploiting this convergence break the initial design philosophy of the Internet protocol stack (end to end argument for example), or if implemented in the application layer present lot of performance, resilience and scalability issues. W e think that the Internet re-design raises the opportunity to better understand and assess higher-level system requirements, and use these as drivers of the lower layer architecture. In this process, mechanisms that are implemented today as part of applications may conceivably migrate into the network itself, and this is one of main driver of the future researches of RESO and of our strong involvment in the new INRIA-BellLabs "Semantic Networking" research axis.

- RESO is closely involved in the design and deployment of the Grid 5000 testbed, and responsible for the networking aspects. Grid5000 is a national initiative aiming at providing a huge experimental instrument to the grid software research community. RESO participate to the new INRIA development action ALADDIN construction. RESO is representing two scientific challenges related to the network
- We continue the investigation of limits of the existing communication services or protocols and evaluate more efficient approaches within the Grid5000 national experimental infrastructure based on the RENATER network and on its international extensions: Netherland (10Gb/s) and Japan (1Gb/s). Participating to the design, deployment and usage of such high performance experimental Network and Grid testbed allow us to gather a strong deep experience and unique expertize in high speed network and protocols exploration and tuning.
- RESO participate to the construction of an international community around Grid networks through the european EC-GIN project as well as with the OGF networking community.
- Through the ANR IGTMD project and is collaborating with the LCG and real physicists. A dedicated link deployed between IN2P3 (one of the largest computing center in France) and the FermiLab laboratory in Chicago, enable us to perform transport protocol experiments as well as traffic capture.
- RESO is bringing its expertize in Grids and Grid Networking to the CARRIOCAS project of the "pôle Ile de France System@tic.This collaboration enable us to explore the limits and the advantages of our previous results in the context of a 40Gb/s network.

5. Software

5.1. BDTS: Bulk Data Transfer Scheduling Service

Keywords: *Transfer jobs*, *bandwidth allocation profile*, *calendar*, *linear programming*, *scheduler*.

Participants: Marcelo Pasin, Sebastien Soudan, Dinil Mon Divakaran, Pascale Vicat-Blanc Primet.

The BDTS(Bulk Data Transfer Scheduling Service) is a modular software which is dedicated to the management of transfer jobs and of the bandwidth allocation along the rate and time axis.

5.2. NXE: Network eXperiment Engine

Keywords: *Network experiment*, *protocol evaluation*, *reproducible experiment*, *workflow*.

Participants: Pascale Vicat-Blanc Primet, Romaric Guillier.

NXE (for Network eXperiment Engine) is a tool developed to be able to execute any particular scenario over any given topology. A scenario is defined as a sequence of dates at which networking events such the start of a new bulk data transfer occur. This software automate the selection, deployement , configuration and activation on distributed resources of pieces of software required to execute a large scale and reproducible networking experiment. This software has been demonstrated during the SuperComputing'2007 event on the INRIA booth.

5.3. HSTTS: High Speed Transport protocols Test Suite

Keywords: *Performance evaluation*, *TCP*, *transport protocol comparison*.

Participants: Pascale Vicat-Blanc Primet, Romaric Guillier.

HSTTS (for High Speed Transport protocol Test Suite) is software imlementing a fixed set of data transfer scenarios. It is designed to help users evaluate the performance they ought to be able to get out of their networking infrastructure when they transfer data by using different types of transport protocols and services. This software has been presented during the SuperComputing'2007 event on the INRIA booth.

5.4. FLOC: Flow control

Keywords: *end-host based traffic shaping; rate enforcement*.

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan.

The FLOC (flow control) software is a low end-host rate and time enforcement mechanism. FLOC is a daemon in charge of ensuring a multi-interval bandwidth allocation profile assigned to a socket identified by a token and registred by user applications will be respected by the flow. FLOC changes GNU/Linux kernel's `qdisc` configuration according to current date and profile so that sockets can only send what they are allowed to.

5.5. DLPTsoft: Distributed Lexicographic Placement Table software

Participants: Pascale Vicat-Blanc Primet, Pierre Bozonnet.

DLPTsoft implements the DLPT distributed information system.

It combines good scalability properties with insertion and search capabilities adapted to resource discovery. The DLPT (Distributed Lexicographic Placement Table) stores services' or resources' references under the shape of (key, value) pairs. The DLPT supports exact match requests on a given key and partial search strings by providing automatic completion. It supports range queries. Multicriteria searches can be also achieved.

5.6. SNE (Stateful Network Equipment)

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

Participant: Laurent Lefèvre [contact].

Joint work with Pablo Neira Ayuso from University of Sevilla (spain).

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. SNE is an open source project, available on the web (CECILL Licence) at http://perso.ens-lyon.fr/laurent.lefevre/software/SNE.

5.7. Tamanoir^{embedded} (Active execution environment for embedded autonomic network equipments)

Keywords: *autonomic networking*, *programmable network equipments*.

Participants: Martine Chaudier, Jean-Patrick Gelas [contact], Laurent Lefèvre.

We designed an Execution Environment called $Tamanoir^{embedded}$ based on the Tamanoir software suite. The original Tamanoir version is a prototype software with features too complex for an industrial purpose (cluster-based approach, Linux modules, multi-level services...).

Due to some typical industrial constraints (e.g code maintenance), we reduced the code complexity and removed all unused classes and methods or actually useless for this project. It allows us to reduce the overall size of the software suite and make the maintenance and improvement of the code easier for service developers.

$Tamanoir^{embedded}$ is a dedicated software platform fully written in Java and suitable for heterogeneous services. Tamanoir provides various methods for dynamic service deployment. $Tamanoir^{embedded}$ also supports autonomic deployment and services updating through mobile equipments. Inside automatic maintenance projects, we deploy wireless based IAN^2 (Industrial Autonomic Network Node) nodes in remote industrial environments (no wire connections available) [58]. In order to download maintenance information, human agents can come near IAN^2 nodes to request informations. During this step, mobile equipments (PDA, Tablets, cellulars) are also used as mobile repositories to push new services and software inside autonomic nodes.

Tamanoir is an open source software suite, available on the web and protected by APP (Agence Francaise de Protection des Programmes).

5.8. XCP-i (Interoperable eXplicit Control Protocol)

Keywords: *XCP*, *high performance transport protocol*.

Participants: Dino Martin Lopez-Pacheco, Anne-Cécile Orgerie, Laurent Lefèvre.

XCP (eXplicit Control Protocol) is a transport protocol that uses the assistance of specialized routers to very accurately determine the available bandwidth along the path from the source to the destination. We propose XCP-i[38] which is operable on an internetwork consisting of XCP routers and traditional IP routers without loosing the benefit of the XCP control laws

An ns-2 module simulating XCP-i has been developed and will be available on the web. Based on a Linux kernel, a software XCP-i router is currently under development.

6. New Results

6.1. Optimized communication software and equipements

6.1.1. Optimisation of MPI application executions on the grid

Keywords: *Grid*, *Grid5000*, *MPI*, *heterogeneity*, *high-speed interconnects*.

Participants: Ludovic Hablot, Olivier Glück, Jean-Christophe Mignot, Pascale Vicat-Blanc Primet.

The MPI standard is often used in parallel applications for communication needs. Most of them are designed for homogeneous clusters but MPI implementations for grids have to take into account heterogeneity and long distance network links in order to maintain a high performance level. These two constraints are not considered together in existing MPI implementations and raise the question of MPI efficiency in grids. Our goal is to significatively improve the performance execution of MPI applications on the grid.

We have done a state of the art, a performance evaluation, understanding and tuning of four recent MPI implementations for the Grid : MPICH-Madeleine, GridMPI, OpenMPI and MPICH2. The comparison is based on the executions of pingpong, NAS Parallel Benchmarks and a real application of geophysics. These experiments take place on the national GRID'5000 testbed. We show that a tuning of both TCP protocol and MPI implementation are necessary to obtain good performances on the grid. We study the impact on application time execution of a long-way latency between two groups of 8 MPI tasks for each NAS parallel benchmark. Our experiments and tunings presented in [24] lead to the conclusion that GridMPI performs better results than the others and that executing MPI applications on a grid can be beneficial if some specific parameters are well tuned.

Next year, we plan to study more precisely the impact of using the TCP protocol for WAN communications (inter-site communications in the grid) and its interactions with MPI applications. We think that the TCP protocol is not the better one for doing commnications between MPI tasks on the grid. In such a way, we want to propose some modifications in the use of TCP for running MPI applications in a grid platform. We also would like to propose a framework allowing a MPI application composed of several tasks to correctly match its tasks regarding the grid topology and the network state. The idea is to execute one time the application in order to know how the application communicates, then to watch which ressources are available on the grid and so, to propose an efficient placement of tasks on grid nodes.

6.1.2. High performance Autonomic Gateways for large scale distributed systems and Grids

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

Participants: Jean-Patrick Gelas, Laurent Lefèvre.

In the framework of a cooperative industrial maintenance and monitoring project (TEMIC project), in which we are involved with different academic and industrial partners, we design devices to be easily and efficiently deployable in an industrial context. Once the hardware deployed and used, it must also be easily removable at the end of the maintenance or monitoring contract. In this project, we deploy our devices in secured industrial departments, restricted areas, or in an out-of-the-way locations. These devices must act as auto-configurable and re-programmable network nodes. Thus, the equipments must be *autonomic* and must not require direct human intervention.

The design of an autonomic network equipment must take into account specific requirements of active equipments in terms of dynamic service deployment, auto-settings, self-configuration, monitoring but also in terms of hardware specification (limited resources, limited mechanical parts constraints, dimension constraints), reliability and fault tolerance.

We proposed an adaptation of a generic high performance active network environment (Tamanoir) in order to deploy on limited resources based network boxes and to increase reliability and scalability. The implementation process is based on a hardware solution provided by the Bearstech company. Through this approach we proposed the architecture of an **I**ndustrial **A**utonomic **N**etwork **N**ode (called IAN^2) able to be deployed in industrial platforms [50], [58]. We evaluated the capabilities of IAN^2 in terms of computing and networking resources and dynamic re-programmability.

6.1.3. High availability for clustered network equipments

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

Participants: Narjess Ayari, Laurent Lefèvre, Pascale Vicat-Blanc Primet.

A key component for improving the scalability and the availability of network services is to deploy them within a cluster of servers. The main objective of this work is to design a network traffic load balancing architecture which meets fine grained scheduling while efficiently spreading the offered network traffic among the available cluster resources.

- **A scalable architecture for balancing the offered network traffic**

  While a lot of researches have been conducted in the field of job and network load balancing, less interest has been granted to the impact of the granularity of the used mechanism on the reliable execution of the upper layer services. In fact, the currently used flow level network load balancing frameworks fail to achieve session awareness while efficiently spreading the offered network load among the available resources, typically, when the offered network session involves multiple and heterogeneous flows. Representative services range from familiar services like HTTP and FTP, to some recent services like multimedia streaming using RTSP/RTP/RTCP and Voice over IP using SIP. Our work aims to provide an architecture to efficiently balance the offered network sessions among the available processing resources within a cluster of servers.
- **A highly available architecture for balancing the offered network traffic**

  High availability allows service architectures to meet growing demands and to ensure uninterrupted service. In our work, we are interested in providing the continuous execution of the offered network sessions in case of failure of the legitimate entry point to the cluster as well as in case of the failure of the processing server inside the cluster. We noticed that current fault tolerant frameworks need to support consistent transport and application level failover mechanisms, and that transport layer protocols do not provide high availability capabilities. Indeed, TCP does not distinguish between a packet loss due to congestion, or a packet loss due to a server overload or due to a server/link failure. Thus, it reacts the same way to packet losses and to delays, by retransmitting the same segment to the same remote end point of the connection. Moreover, TCP tolerates short periods of disconnection not longer than a few RTTs. It disconnects the communicating hosts once specific timers expire. On the other hand, transport protocols rely on an explicit association between a service and its physical location for the wired Internet. Thus, when a host fails, the end-to-end flow terminates.

  In order to address this limitation, we proposed an active replication based system which enhances the reliability of the already established TCP flows. The proposed scheme is client transparent and does not incur any overhead to the end-to-end communication during failsafe periods, and performs well during failures. Parts of this work are protected by the Intellectual Property National Institute (INPI) patent disclosure N°FR0653546

  [17], [16], [41], [15], [42]

6.1.4. High availability for stateful network equipments

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

Participant: Laurent Lefèvre.

Joint work with Pablo Neira Ayuso from University of Sevilla (Spain).
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

6.2. E2E Transport and Service Differentiation

6.2.1. A study of large flow interactions in high-speed shared networks with Grid5000 and GtrcNET-10 instruments

Keywords: *bandwidth sharing*, *bulk data transfers*, *high speed transport protocol experimentation*.

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan, Ludovic Hablot, Romaric Guillier.

We consider the problem of huge data transfers and bandwidth sharing in the context of grid infrastructures where transfer delay bounds are required. In this work we investigate large flow interactions in a real very high-speed network and aim at contributing to high-speed TCP variants evaluation by providing precise measurements. We also explore the behaviour of protocols under different realistic congestion and long latency conditions in 10 Gbps experimental emulated environments. We show that using parallel streams with new TCP protocols like BIC is highly valuable in this context as it increases the multiplexing level. According to the modest RTT value of the grid testbed we use, the various TCP variants we evaluated present comparable results, and Reno still behaves quite well. When the latency increases, H-TCP and HS-TCP performs better than the others in these particular conditions.

6.2.2. TCP Variants and Transfer Time Predictability in Very High Speed Networks

Keywords: *bulk data transfers*, *high speed transport protocol experimentation*.

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan, Romaric Guillier.

In high performance distributed computing applications, data movements have demanding performance requirements such as reliable and predictable delivery. Predicting the throughput of large transfers is very difficult in paths that are heavily loaded with just a few big flows. In this work we explore how current high speed transport protocols behave and may improve transfer time predictability of gigabits of data among endpoints in a range of conditions. In a fully controlled long distance 10 Gbps network testbed, we compare several TCP variants behaviour in presence of diverse congestion level and reverse traffic situations. We show that these factors have a very strong impact on transfer time predictability of several transport protocols. We show that when bulk data transfers start simultaneously, transfer time efficiency and predictability are strongly affected. When the congestion level is high (> 1.2) both transfer time efficiency and predictability depend on the chosen protocol.The most important factor this study reveals is the reverse traffic impact. It strongly affects all protocols. We conclude that flow scheduling service controlling the starting time and the congestion level in forward and reverse path is mandatory in these low multiplexing environments

6.2.3. Towards a User-Oriented Benchmark for Transport Protocols Comparison in very High Speed Networks

Keywords: *High Speed networks*, *High Speed transport*, *Performance evaluation*, *Protocol Benchmark*, *TCP*.

Participants: Pascale Vicat-Blanc Primet, Romaric Guillier, Ludovic Hablot.

Standard TCP faces performance limitations in very high speed wide area networks, mainly due to a long end-to-end feedback loop and a conservative behaviour with respect to congestion. Many TCP variants have been proposed to overcome these limitations. However, TCP is a complex protocol with many user-configurable parameters and a range of different implementations. It is then important to define measurement methods so that the transport services and protocols can evolve guided by scientific principles and can be compared quantitatively. Users of these variants need performance parameters that describe protocol capabilities so that they can develop and tune their applications. The goal of this work to make some steps towards a user-oriented test suite and a benchmark, called HSTTS, for high speed transport protocols comparison. We first identified useful metrics. We then isolated infrastructure parameters and traffic factors which influence the protocol behaviour. This enabled us to define classes of representative applications and scenarios capturing and synthesising comprehensive and useful properties. We finally evaluate this proposal on the Grid'5000 experimental environment, and present it to the IRTF TRMG working group.

6.2.4. Evaluation of High Speed TCP variants and study of large flow interactions in high-speed shared networks

Keywords: *bulk data transfers*, *congestion control*, *high speed transport protocol*, *transfer delay predictability*, *transport protocol experimentation*.

Participants: Romaric Guillier, Ludovic Hablot, Sébastien Soudan, Pascale Vicat-Blanc.

We consider the problem of huge data transfers and congestion control in contexts where transfer delay bounds are required. We investigate high-speed TCP variants and contribute to their evaluation by providing accurate measurements. This work gives an insight on the behaviour of alternative protocols under different realistic congestion and long latency conditions in the 10 Gbps experimental environments provided by the Grid5000 testbed and by the GtrcNET10 latency emulation device. This work also gives experimental results on performance of a large number of parallel flows (up to 110 parallel streams) and on large flow interactions in a real very high-speed networks This work complements the general studies on transport protocol benchmarking which we explored within the international cpfldneet community [66].

6.2.5. Router assisted network transport protocol

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

Participants: Dino Martin Lopez-Pacheco, Laurent Lefèvre.

In heterogeneous networks, where many flows, non-regulated and/or with a high QoS level, share the resources, the available best-effort bandwidth varies over time. This changes can be represented by an aggregation of UDP ON-OFF sources what produces a step-based variation model. In this type of environments, we have tested the performance of many transport control protocols (TCP New Reno, High Speed TCP, TCP Westwood+ and XCP) using the ns2 simulator. In our studies, XCP showed always the best performance, with a high stability and fairness level. But in heterogeneous networks, the lost of packets is very common, so we have tested XCP in a network where the lost in the reverse path cause some ACK losses. In the new results, we have found that the ACK losses produce many problems in the connections, caused by a wrong calculus of the congestion window size, specifically when the available bandwidth decreases. That is because the success of XCP is based on the network state information, provided by the routers to the sender in the ACK packets. Since, the problem is generated by the wrong calculus of the congestion window size in the sender side, we proposed to compute this value in the receiver side. We have called this new approach XCP-r

We repeated the simulations set using XCP-r and we found that XCP-r shows always more stability and better fairness level.

6.2.6. XCP-i: a new interoperable XCP version for high speed heterogeneous networks

Keywords: *TCP*, *XCP*, *XCP-i*, *available bandwidth*, *congestion control*, *virtual XCP-i router*.

Participants: Dino Martin Lopez-Pacheco, Laurent Lefèvre.

XCP (eXplicit Control Protocol) is a transport protocol that uses the assistance of specialized routers to very accurately determine the available bandwidth along the path from the source to the destination. In this way, XCP efficiently controls the sender's congestion window size thus avoiding the traditional slow-start and congestion avoidance phase. However, XCP requires the collaboration of all the routers on the data path which is almost impossible to achieve in an incremental deployment scenario of XCP. It has been shown that XCP behaves badly, worse than TCP, in the presence of non-XCP routers thus limiting dramatically the benefit of having XCP running in some parts of the network. In this work, we address this problem and propose XCP-i which is operable on an internetwork consisting of XCP routers and traditional IP routers without loosing the benefit of the XCP control laws.

XCP-i basically executes the next four steps to discover and compute a new feedback that reflects the state of the network where non-XCP routers are placed:

1. Discover where the non-XCP routers are in the data path.
2. Discover the upstream and downstream XCP-i routers of the non-XCP routers.
3. Estimate the available bandwidth where the non-XCP routers are placed.
4. Create a virtual XCP-i router that computes a new feedback using the estimated available bandwidth before.

The simulation results on a number of topologies that reflect the various scenario of incremental deployment on the Internet show that although XCP-i performances depend on available bandwidth estimation accuracy, XCP-i still outperforms TCP on high-speed links [38].

6.2.7. Flow scheduling in high speed networks

Keywords: *bandwidth reservation*, *flow scheduling*, *grid networks*, *multi-rate*, *transfer delay*.

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan, Dinil Mon Divakaran, Marcelo Pasin, Chen Cheng.

In this work, we consider the problem of bulk data transfers and bandwidth sharing in the context of grid infrastructures. Tight co-ordination of resource allocation among end points in grid networks often requires a service to transfer voluminous data sets from one site to another in a specified time interval. Given a set of such transfers, we studied the Bulk Data Transfer Scheduling (BDTS) problem and then developed the associated software, which provides grid users and grid application a service to specify their transfer request and ensure a transparent control of them.

The BDTS problem searches for the optimal bandwidth allocation profile for each transfer to minimize the overall network congestion. An important objective of scheduling, thus, is to minimize the (weighted) maximum required capacity in network along time axis. If request must be served with non-zero bandwidth in a continuous interval, the optimal scheduling is NP-complete even for a single link. In comparison, we were able to show that the multi-interval scheduling, which divides the active window of a task into multiple intervals and assigns bandwidth value independently in each of them, is both sufficient and necessary to attain the optimality in BDTS[20].

Specifically, we demonstrated that BDTS can be solved in polynomial time as a Maximum Concurrent Flow Problem[19].

The optimal solution attained is in the form of multi-interval scheduling with the number of intervals upper-bounded.

The concepts developed have proven to work with the implementation of two software services, one to manage and schedule transfer jobs and another to control the flow of the executing jobs (the last one is discussed in detail in [33].

BDTS was implemented using two original abstractions: profiles and calendars. Profiles represent a bandwidth function over the time, and is used to represent how much (time and space) every transfer uses of each link. It is also used to represent link capacity and to command flow control components to enforce the network allocation. The application that adapts to the bandwidth can have access to its profiles as well. Calendars on the other hand are the collection of allocations for a single link. They are composed of several profiles and keep track of all engagements of a link.

Experiments conducted over several representative topologies in Grid5000 demonstrated the significant advantage of optimal solutions presented.

6.2.8. End point flow time and rate control in very high speed networks

Keywords: *flow synchronisation*, *multi-rate*, *packet pacing*, *rate control*, *transfer delay*.

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan, Dinil Mon Divakaran.

In this work, we consider the problem of enforcing rate allocation made by BDTS scheduler in a packet network. As enforcement in network equipments is not available, we consider end-host enforcement mechanisms. We compared end-host based traffic shaping mechanisms combined with transport protocols to implement this flow scheduling architecture. The evaluation is carried out on a testbed in a range of latency conditions, which shows that, (1) TCP AIMD congestion control is neither efficient nor stable, especially when RTT is large; (2) a fine-grained traffic shaper is necessary to avoid temporal burst, esp. when router/switch 19 does not provide enough buffer; (3) a large enough buffer in sender is necessary to quickly change from low rate to high rate. We have designed and developed the FLOC software, which is present on each machine responsible to enforce the multi-interval bandwidth allocation profile. Experiments conducted in the Grid5000 testbed have shown the accuracy and responsiveness of the rate control mechanisms provided in the FLOC software. Future work will concentrate on larger experiments on the Grid5000 testbed, comparison with UDT-BLAST rate limitation implementation and will examine the scalability of the flow scheduling approach in real grid context with real applications. This work is partially supported by EC-GIN EU contract.

6.2.9. Flow scheduling and lambda-path reservation

Participants: Pascale Vicat-Blanc Primet, Sebastien Soudan, Romaric Guillier, Ludovic Hablot.

In this work we are studying how lambda-path reservation which enable to dynamically provision a networking links can be combined with a flow scheduler. This work is based on the AIST GNS-WSI2 service interface and the RESO BDTS service. GNS-WSI2 is intended to become a standard web services interface between Grid resource manager and operator-owned network resource manager for advance reservation of bandwidth. It is developped in the G-lambda project which is a joint project of KDDI R&D labs, NTT, NICT and AIST. BDTS is developed by INRIA RESO team. It receives requests of transfer specified by source, destination, volume, minimum start time and maximum end time. Then it finds network resources that allow to transfer this volume during the specified time window. BDTS acts as a resource manager for site network resources by keeping resource utilization information and as a resource coordinator for core network resources by doing reservation to core NRMs. In the first step we have interfaced GNS-WSI2 with BDTS to provision in advance provisioning of some network path. We plan to study further the time granularity issue and the flexibility offered by on demand or in advance lambda path provisionning. This collaborative topic between AIST GTRC team and INRIA RESO team in the context of the GridNet-FJ associated team as started in September 2007. We are also examining this problem in the framework of the CARRIOCAS project with Orange-Labs and Alcatel-Lucent.

6.2.10. Steady state load balancing

Keywords: *ditributed and dynamic sharing*, *fairness*.

Participants: Rémi Vannier, Isabelle Guérin Lassous.

Multiple applications that execute on an heterogeneous platform compete for CPU and network resources. We design a model and an algorithm to find a load balancing between loosely coupled applications (independant applications such as BOINC) that execute concurrently on such a platform. The algorithm being distributed among the nodes of the platform, it has nice properties such as scalability, and fault tolerance. Besides, the achieved solution optimizes the use of the platform while being fair between the applications, which means that every application has its share of computing time.

6.2.11. Time Fairness in Wireless Local Networks

Keywords: *MAC protocols*, *performance anomaly*, *time fairness*.

Participant: Isabelle Guérin Lassous.

In the widely used IEEE 802.11 standard, the so-called *performance anomaly* is a well known issue. Several works have tried to solve this problem by introducing mechanisms such as packet fragmentation, backoff adaptation, or packet aggregation during a fixed time interval. This year, we design and thoroughly analyze PAS, Performance Anomaly Solution, a dynamic and distributed approach solving the performance anomaly problem. PAS is based on packets' aggregation using a dynamic time interval, which depends on the wireless channel occupation time perceived by each node. Since each station senses the medium independently, this makes PAS a totally distributed solution. Even more, PAS may coexist with standard IEEE 802.11 nodes without any particular adaptation, yet being able to increase performances. Our solution differs from other proposition in the literature because of its dynamic and distributed nature, which makes it suitable in the context of multi-hop networks. Furthermore, it allows increasing fairness, reactivity, and in some cases efficiency.

6.2.12. Fairness and Efficiency in Ad Hoc networks

Keywords: *MAC protocols*, *efficiency*, *fairness*.

Participant: Isabelle Guérin Lassous.

The IEEE 802.11 MAC layer is known for its unfairness behavior in *ad hoc* networks. Introducing fairness in the 802.11 MAC protocol may lead to a global throughput decrease. It is still a real challenge to design a fair MAC protocol for ad hoc networks that is distributed, topology independent, that relies on no explicit information exchanges and that is efficient, *i.e.* that achieves a good aggregate throughput. The MadMac protocol deals with fairness and throughput by maximizing aggregate throughput when unfairness is solved. Fairness provided by MadMac is only based on information provided by the 802.11 MAC layer. MadMac has been tested in many configurations that are known to be unfair and compared with three protocols (IEEE 802.11 and two fair MAC protocols). In these configurations, MadMac provides a good aggregate throughput while solving the fairness issues.

6.2.13. Dynamic Bandwidth Sharing in Ad Hoc Networks

Keywords: *bandwidth sharing*, *differentiation*, *efficiency*.

Participant: Isabelle Guérin Lassous.

This year, we propose a new cross-layer protocol named DRBT (Dynamic Regulation of Best Effort Traffic) which supports QoS guarantees and provides a distributed regulation mechanism for best effort traffic in wireless ad hoc networks. By adapting the rate of best effort traffic at the MAC Layer, DRBT increases the acceptance's rate of QoS flows through the network. Our protocol also provides an accurate method to evaluate the available bandwidth in IEEE 802.11-based ad hoc networks which differentiates between real time applications and those which are less exigent in term of bandwidth more commonly called best effort traffic. Through simulation, we compare the performance of our proposal scheme with AODV.

6.2.14. Scheduling bulk data transfers in grid networks

Keywords: *bandwidth reservation*, *deadline*, *flexible start time*, *flow scheduling*, *grid networks*, *multi-rate*.

Participants: Chen Binbin, Sébastien Soudan, Pascale Vicat-Blanc Primet.

In this long term research area, we consider the problem of bulk data transfers and bandwidth sharing in the context of grid infrastructures and propose to explore a disruptive approach for congestion control in high pseed networks. Indeed, in grid computing which empowers high-performance computing in a large-scale distributed environment, network bandwidth, which makes the expensive computational and storage resources work in concert, plays an active role on carrying grid applications traffic. Due to specific traffic patterns and application scenarios, grid network resource management encounters new challenges. From the bandwidth sharing perspective, we look at network bandwidth shared among computing and storage elements and explore a session level network resource control approach.

- In our first investigations of this field, we introduced a specific network model: a hierarchical bipartite graph with two sets of bottlenecks called ingress and egress points and defined bulk data transfer job request. Referred to as short-lived, grid data requests with transmission window and volume are scheduled in the network. By manipulating the transmission window, the request accept rate and network resource utilization are to be optimized. The formulated optimization problem, considering this network model, is proven NP-complete. Associated with proposed heuristics, simulations are carried out to illustrate the pros and cons of each bandwidth sharing strategy and its application scenarios. A tuning factor, that allows for adapting performance objective, is introduced to adjust network infrastructure and workload
- We then continue this study of bandwidth reservation problem for bulk data transfers in grid networks. We generalize our grid networks model as a set of distributed sites interconnected by any network with potential bottlenecks, and transfer requests arrive online with specifed volumes and deadlines. Current reservation schemes such as RSVP are designed for requests with fixed transmission start time and single rate. In comparison, our definition of request in terms of volume and deadline allows more flexibility in the design of reservation schemes. We define the extended design space by formalizing three schemes families, namely, NOW (or immediate), Single Rate (SR) and Multirate (MR), with increasing generality, complexity and potential performance. Maximal packing (MaxPack) and minimize delay (MinDelay) are set as criteria to select candidate scheme from each family. The proposed reservation schemes is shown to achieve a much better performance then RSVP-type schemes, and can be implemented in both centralized and distributed architectures [51].
- The following study explored the same problem of bandwidth scheduling for transfers with specified volume, active time window (arrival time and deadline) and route, but consider both periodic and sporadic bulk data transfers. For periodic transfers, their request definitions are available off-line and network capacity is dimensioned to accept all of them. An important objective of scheduling, thus, is to minimize the (weighted) maximum required capacity in network along time axis. If request must be served with non-zero bandwidth in a continuous interval, the optimal scheduling is NP-complete even for a single link. In comparison, if the active window of request can be divided into multiple sub-intervals, each with different data rates (possibly zero), the optimal scheduling problem can be modeled as a multicommodity network flow problem which employs polynomial solution. Remained network capacity from periodic transfers is then used to serve sporadic transfers which arrive dynamically. The performance metric for sporadic transfers includes both accept probability and flow time, both of which can be potentially improved if bandwidth is scheduled flexibility
- Finally, we start examining how the advance reservation and off-line data transfers jobs scheduling will interfere with an unified control plane allowing the creation of bandwidth guaranteed tunnels across optical core network and Ethernet local network. We propose a model for such networks and study the problem of bandwidth sharing with bulk data transfers in this GMPLS context. Several allocation algorithms based on QoS routing works have been proposed and compared

6.2.15. Maximum likelihood estimate of heavy-tail exponents from sampled data

Keywords: *flow size*, *heavy-tail distributions*, *maximum likelihood estimation*.

Participants: Patrick Loiseau, Paulo Gonçalves, Pascale Vicat-Blanc Primet.

This work is partially in collaboration with the MISTIS team project.
We addressed the problem of estimating the flow size distribution corresponding to the traffic passing through an aggregated link. More precisely, we want to measure the tail exponent (hypothesizing that distributions are systematically heavy-tailed) from a sub-sampled measured series of packets stream. Considering a Pareto distribution as our theoretical a priori model, we formally derived the maximum likelihood estimate of the Pareto tail exponent α. Independently we proposed in a previous work a heuristic estimate of α (assuming the same Pareto a priori), which turned out to be conceptually very close to the maximum likelihood estimate, providing us with an intuitive interpretation of this latter. Based on simulated data, we performed a systematic

comparison of our estimator with different approaches proposed in the literature (e.g. stochastic counting, EM estimates, wavelet based estimates,...). Then, not only the proposed method significantly improves bias and variance estimates, but also, it still holds with small sampling rates (reasonably up to $1/100$), drastically contrasting with most rival estimators[26].

6.3. Grid Network services and applications

6.3.1. Development of a metrology platform on Grid5000

Keywords: *Gtrc-Net1*, *header extraction*, *metrology*, *monitoring*, *packet capture*.

Participants: Patrick Loiseau, Damien Ancelin, Aurélien Cedeyn, Paulo Gonçalves, Pascale Vicat-Blanc Primet.

We designed, implemented and deployed an experimental metrology platform able to perform a non-intrusive capture of grid traffic at packet-level granularity. Currently, we can probe a bi-directional link at 1Gb/s, with no data loss and with on-line extraction of packet headers.
Components of our experimental setup have the following characteristics:

- Gtrc-Net 1 (Gtrc-Net 10 in a mid-term future): This device allows to capture packet headers on a bi-directional link at 1 (at 10 Gbps respectively). This device is being developed within the framework of our GridNet-FJ associated team, with AIST GTRC Japan.
- Output of Gtrc-Net is treated by MAPI (Monitoring Application Programming Interface developed by the LOBSTER project) for extraction and possibly for visualization/monitoring of the probed link. We had to develop a specific driver for Gtrc-Net/MAPI interface. - The output format of MAPI is pcap readable by IPSUMDUMP. In addition, we resort to the toolbox "IP tools", independently developed by the OSCAR[1] ANR-project, and which allows for flow oriented treatments.

This work is supported by GridNet-FJ, EC-GIN EU contract and Grid5000/ALADDIN ADT.

6.3.2. SNMP-based Monitoring Agents and Heuristic Scheduling for large scale Grids

Keywords: *large scale Grids*, *monitoring*.

Participant: Laurent Lefèvre [contact].

Joint work with Edgar Magana (UPC, Cisco) and Joan Serrat (UPC, Barcelona, Spain).

6.3.3. Programmable network services for context aware adaptation

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

Participants: Laurent Lefèvre, Jean-Patrick Gelas.

Traditional industrial maintenance process (i.e. requiring regularly a human intervention on the exploitation area) are coming to their limits. Indeed, more and more industrial equipments are connected to communication networks. This allows us to consider optimised maintenance solutions. In addition to primary existing sensors (which only give some numeric values), we can now think about the use of multimedia sensors (video cameras, microphone, ...). Inside a cooperative industrial maintenance project (TEMIC project [65]) in which we are currently involved, our team designed equipments easily deployable in an industrial context, and also easily removable at the end of the maintenance contract.

[1] "Détection d'anomalies dans les réseaux de type overlay

The heterogeneity interms of networks, terminals and applications requires adaptive solutions for an efficient streams transmission on the platform networks. To respond to these various constraints, active services have to adapt and optimize the content of streams passing through the active network node. Multimedia data streams adaptation is performed dynamically in order to improve industrial maintenance solutions. The challenge is to provide an architecture running in a client/server environment, but involving no modification on the applications installed on the end-machines like web servers, video players,... For the Temic project, our team has worked on the design and adaptation of an industrial autonomic network node, which is derived from the Tamanoir environment. This Industrial Autonomic Network Node is designed to be deployed on limited resources based network boxes, and so to be integrated into industrial platforms. We developed and tested active adaptation network services, specially written for the Tamanoir^{embedded}. Active services applying on multimedia streams crossing the network node may realize data compression, format transcoding, frame resizing... This kind of adaptation contributes to the saving of network bandwidth (by decreasing the output data rate) and to the reduction of the resources used on the client terminal playing the multimedia data (by reducing the framerate and the frame size). The adaptation is thereby transparent for the applications.

We base our developments and experimentations on mainly two industrial maintenance scenariosTEMIC project team to be used by a company through a maintenance contract on a restricted industrial area.

At this time, three active services have been developed for this project. They are designed to adapt multimedia data on the fly.

Our experiments show that our solution is efficient in reducing the amount of data transmitted on the network, and so the bandwidth consumed by the application, and also in reducing the CPU and resources needed on the client machine to decode the streams. However, our experiments clearly show some limitations in the performances of our industrial network node. These low performances impact directly the display quality on the user's device. We have now to improve our hardware equipement to obtain better performances.

6.3.4. Inter-Planetary Grid Networking

Keywords: *Delay Tolerant Networking*, *Grid*, *autonomic networks*.

Participants: Jean-Patrick Gelas, Laurent Lefèvre.

The idea to extend the computer network protocols in order to tremendously extend the range of Internet through space was born and supported by the same persons who design TCP/IP 30 years ago, like Vint Cerf. Due to some constraints, transport protocols, among other (ex: routing, name space) must be radically changed to fit the requirement of this unusual environment, namely space! In the same time, the Delay Tolerant Networking (DTN) community works on networks which must deal with high latencies, frequent disconnections, no end-to-end path and power saving constraints. The new proposed protocols are designed to support high latencies and long disconnection (i.e. more than few milliseconds). They also should resist to planned or unplanned disconnection. We consider that the concept of Interplanetary Networks based on Disruption Tolerant Network solutions can be applied to Grid infrastructures.

Programmable and active networks allow specified classes of users to deploy dynamic network services adapted to data streams requirements. We have proposed the Active Grid Architecture (A-Grid) which focuses on active network adaptation for supporting Grid environments and applications. This Active Grid architecture proposes solutions to support multi-cluster infrastructures. This architecture is based on programmable network nodes distributed on network path used as gateways of clusters. In this architecture the network will take part in the Grid computing session by providing efficient and intelligent services dedicated to Grid data streams transport.

This tolerant design expects that applications remain efficient even if networks generate high latencies for communications. This approach focused on latency can be generalized to disrupted infrastructures. But, we want to propose global solutions as transparent as possible for users, applications and Grid middleware. Our approach allows us to modify only the system used as Programmable Network Gateway (PNG) located between clusters and the external network (i.e. Internet)

The proposed architecture of an Interplanetary Grid can also be applied to Grid infrastructures dealing with unreliable long distance network connections. We are currently implementing the model exposed in the previous section and we plan to emulate first experimentations and evaluations of this approach [55].

6.3.5. Integrating web services and programmable networks for improving flexibility of active Grids

Keywords: *Web services*, *programmable networks*.

Participants: Laurent Lefèvre, Pablo Pazos Rey.

Joint work with Chien-Jon Soon and Paul Roe from Queensland University of Technology, Brisbane, (Australia).
Active Grids [48], [49] are a form of grid infrastructure where the grid network is active and programmable. These grids directly support applications with value added services [59] such as data migration, compression, adaptation and monitoring. Services such as these are particularly important for eResearch applications which by their very nature are performance critical and data intensive.

We propose an architecture for improving the flexibility of Active Grids through web services. These enable Active Grid services to be easily and flexibly configured, monitored and deployed from practically any platform or application. The architecture is called WeSPNI ("Web Services based on Programmable Networks Infrastructure")

7. Contracts and Grants with Industry

7.1. Alcatel:Network-aware resource discovery

Keywords: *Network-aware resource discovery*, *P2P*.

Participants: Pascale Vicat-Blanc Primet, Pierre Bozonnet, Eddy Caron, Cedric Tedeshi, Dinil Mon Divakaran.

Our goal in this study was to explore the service & resource discovery problem in the context of new generation networks. We have studied centralized and distributed solutions and analysed how the networking constrains like inter-resource latency of throughput can be expressed and discovered. Then we have developed a P2P solution to evaluate the advantages and drawbacks of this approach in a real networking context. Basically our solution can be seen as a 2 layer infrastructure. The top layer (logical layer) is a distributed information system based on the DLPT model designed by the Graal EPI. It combines good scalability properties with insertion and search capabilities adapted to resource discovery and enable the integration of the network-awareness. The bottom layer (physical layer) is a P2P overlay that enables reliable communications between peers of the network. It has its own P2P routing protocol and allows peers to easily forward and handle queries and responses.

7.2. France Telecom R&D

Participants: Laurent Lefèvre, Pascale Primet.

In 2005, RESO has launched a collaboration with France Telecom R&D (Lannion) on "Network load balancing on layer 7 switching for high performance and high available Linux based platforms". A CIFRE grant has been accepted for supporting this collaboration. Ayari Narjess has begun her PhD on this topic in June 2005 [15], [16], [17], [42]. A patent has been deposed on "Multi-flow sessions management solutions" in 2007 [41]

8. Other Grants and Activities

8.1. National actions

8.1.1. GRID5000

Participants: Olivier Glück, Magi Sanchon, Sébastien Soudan, Romaric Guillier, Ludovic Hablot, Laurent Lefèvre, Pascale Vicat-Blanc Primet, Paulo Gonçalves, Patrick Loiseau, Jean-Christophe Mignot, Aurélien Cedeyn.

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 Grid5000 of Lyon comprises now around 300 processors interconnected with a network of 500Mb/s Ethernet bisection and a 2Gb/s Myrinet interconnection for 64 nodes.

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. Lyon site is nationaly recognized to gather the "networking expertize" with skilled researchers and engineers and dedicated networking equipments Metroflux, GNET10...). Working for the interconnection of the Grid5000 project at the international level, we are hosting the japanese Naregi project remote hosts and are accessing to dediated equipments within the Naregi testbed. We also participate to the ALADDIN ADT. Laurent Lefevre is responsible for the "défi 8". Aurélien Ceyden is member of the national technical committee of GRID'5000. Actual funding: 530K euros

8.1.2. ACI Grandes Masses de Données GridExplorer

Participants: Jean-Patrick Gelas, Olivier Glück, Laurent Lefèvre, Dino Lopez Pacheco, Pascale Vicat-Blanc Primet.

(2003-2006) : The aim of this project was 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 [67]

The second part of our contribution to this project was the evaluation of high speed transport protocol. This activity has started within the GridExplorer project and is now continuing within the ANR IGTMD project.

8.1.3. ANR IGTMD

Participants: Pascale Vicat-Blanc Primet, Chen Cheng, Romaric Guillier.

The aim of this project (2006-2008) is to design, develop and validate mechanisms that concretely make the interoperability of heterogeneous grids a reality. The project concentrates on the following topics: a) Bulk data transfers, b) replication and referring mechanisms, c) information system and job management interoperability, d) grid control and monitoring, e) usage of statistics and accounting data. A particular emphasis will be put on disk to disk bulk data transfers over very long distance with optimal performance. The key idea is to fully exploit the specificity of LCG applications (Computing Grid Project to find the grid middleware developed for the future Large Hydron Collider in CERN) and their real infrastructures to analyse and experiment new communication and replication models, alternative transport protocols emerging within the international scientific community. The participation to a standardization process for a generic grid transport service for bulk exchanges between heterogeneous grids will be a strong goal of the project. Despite the fact that the interoperability and the unification of a generic data transport in Grids are very often perceived as a necessity, they are in fact very little studied. The present project would allow France to get a leading position in this computing area that will be absolutely crucial to insure the Large Hydron Collider (LHC) data exploitation. The very strong experience of the partners in deployment and exploitation of international research and production computing instruments gives a promising perspective to this project and its ambitious

experimental approach. In this project, RESO is responsible for all research activities concerning high speed transport protocols and services. The key idea is to fully exploit the specificity of LCG applications and their real infrastructures to analyse and experiment new communication and replication models, alternative transport protocols emerging within the international scientific community. We are also exploiting the LCG traffic circulatingon the IGTMD link for packet capture and grid flow analysis

8.1.4. ANR DSLLAB

Participants: Laurent Lefèvre, Pascale Vicat-Blanc Primet, Jean-Patrick Gelas.

RESO is partner of the DSLlab research project (2006-2008) which aims at building and using an experimental platform about distributed systems running on DSL Internet. The objective is twofold:

- to provide with accurate and customized measures of availability, activity and performances in order to characterize and tune the models of the ADSL resources;
- to provide with a validation and experimental tool for new protocols, services and simulators and emulators for these systems.

DSLlab consists of a set of low power, low noise computers spread over the ADSL. These computers are used simultaneously as active probes to capture the behavior traces, and as operational nodes to launch experiments. We expect from this experiment a better knowledge of the behavior of the ADSL and the design of accurate models for emulation and simulation of these systems which represents now a significant capability in terms of storage and computing power. The DSLLAB platform will be deployed in 2007.

In this project, RESO is reponsible for the definition, design and development of flow control algorithms and mechanisms, enabling distributed computing applications to fully exploit the DLS links.

8.1.5. ANR HIPCAL

Participants: Pascale Vicat-Blanc Primet, Jean-Patrick Gelas.

HIPerCAL studies a new paradigm (grid substrate) based on confined virtual cluster concept for resource control in grids. In particular, we propose to study and implement new approaches for bandwidth sharing and end to end network quality of service guarantees. The global infrastructure (computers, disks, networks) in partitioned in virtual infrastructures (aggregation of virtual machines coupled with virtual channels) dynamically composed. These virtual clusters are multiplexed in time and space, isolated and protected. The goal of this project is to explore an approach in a break with current services-oriented principles developed in grids to jointly enhance the application portability, the communications performance control and their security. The project aims at providing a grid substrate based on end to end bandwidth reservation, control overlay, network and system virtualization, cryptographic identification principles. The proposal will be validated and evaluated at different scales on the Grid5000 testbed with biomedical applications, demanding in security , performance and reliability. 10 to 1000 processors, links with 100Mb/s to 10Gb/s, few microseconds to 100ms will be involved in these experimentations. Comparison with Globus, Planetlab and Cluster on Demand approaches will be one of the specific goals of the experiments. We aim at demonstrating the functional transparency, enhanced predictability and efficiency for applications offered by the HIPerNET approach.

8.1.6. CARRIOCAS

Participants: Pascale Vicat-Blanc Primet, Pierre Bozonnet, Marcelo Pasin, Damien Ancelin.

Carriocas project studies and implements an ultra high bit rate (up to 40 Gbps per wavelength) network interconnecting super computers, storage servers and high resolution visualization devices to support data and computing intensive applications in industrial and scientific domains. The R&D activities cover high bit rate transmission systems, advanced networking intelligence, and high performance distributed applications. CARRIOCAS is a three year project started in October 2006 which aims to be an experimental step of the transition from local to external storage and computing systems. This transition is valuable to share the cost of powerful systems among several users, to provide scalable and resilient architecture through distributed resource and to enable virual collaborative working environments between different actors working on a same project. The following points will be especially investigated:

- Supporting the high bandwidth requirements through the migration of networks from 10 gbp/s to 40 Gbps/s per wavelength in a cost effective way.
- Building architectural, protocol and algorithmic solutions able to provide to the network the agility to dynamically adapt to the application needs with a high level of automation and optimizsation, while taking into account the administrative and business constraints.
- Developing and demonstrating on a network testbed distributed applications bringing perfomance enhancements for concrete scientific and industrial needs.
- Investigating the definition and the associated business models of high added value services integrating computing, vizulation, storage and network resources.

In this project, RESO in charge of the design and protoyping of the "Resource Scheduling Reconfiguration and Virtualisation - SRV" component.

8.2. European actions

8.2.1. AEOLUS

Participants: Isabelle Guérin-Lassous, Rémi Vanier.

AEOLUS (Algorithmic Principles for Building Efficient Overlay Computers) is an IP project that has been started since September, 1st, 2005. The university of Patras (Greece) is the prime contractor. The goal of this project is to investigate the principles and develop the algorithmic methods for building an overlay computer that enables an efficient and transparent access to the resources of an Internet-based global computer. In particular, the main objectives of this project are:

- To identify and study the important fundamental problems and investigate the corresponding algorithmic principles related to overlay computers running on global computers.
- To identify the important functionalities such an overlay computer should provide as tools to the programmer, and to develop, rigorously analyze and experimentally validate algorithmic methods that can make these functionalities efficient, scalable, fault-tolerant, and transparent to heterogeneity.
- To provide improved methods for communication and computing among wireless and possibly mobile nodes so that they can transparently become part of larger Internet-based overlay computers.
- To implement a set of functionalities, integrate them under a common software platform in order to provide the basic primitives of an overlay computer, as well as build sample services on this overlay computer, thus providing a proof-of-concept for our theoretical results.

8.2.2. EC-GIN

Participants: Pascale Vicat-Blanc Primet, Paulo Gonçalves, Patrick Loiseau, Sébastien Soudan, Romaric Guiller, Ludovic Hablot.

EC-GIN (Europe-China Grid InterNetworking) is an European STREP project started in November 1st 2006. The university of Innsbrück (Austria) is the prime contractor.
The Internet communication infrastructure (the TCP/IP protocol stack) is designed for broad use; as such, it does not take the specific characteristics of Grid applications into account. This one-size-fits-all approach works for a number of application domains, however, it is far from being optimal - general network mechanisms, while useful for the Grid, cannot be as efficient as customised solutions. While the Grid is slowly emerging, its network infrastructure is still in its infancy. Thus, based on a number of properties that make Grids unique from the network perspective, the project EC-GIN will develop tailored network technology in dedicated support of Grid applications. These technical solutions will be supplemented with a secure and incentive-based Grid Services network traffic management system, which will balance the conflicting performance demand and the economic use of resources in the network and within the Grid.

By collaboration between European and Chinese partners, EC-GIN parallels previous efforts for real-time multimedia transmission across the Internet: much like the Grid, these applications have special network requirements and show a special behaviour from the network perspective. However, while research into network support for multimedia applications has flourished, leading to a large number of standard protocols and mechanisms, the research community has neglected network support for Grid computing up to now. By filling this gap and appropriately exploiting / disseminating the project results, EC-GIN will, therefore, cause a "snowball effect" in the European and Chinese networking and Grid computing research communities.
Technically, EC-GIN will make the Grid work, operate, and communicate better. By appropriately utilising the underlying network, Grid resources in general will be used more efficiently and amplify the impact of Grid computing on the society and economy of Europe and China.

8.3. International actions

8.3.1. NEGST: JSPT-CNRS

Participants: Olivier Glück, Magi Sanchon, Sébastien Soudan, Romaric Guillier, Ludovic Hablot, Laurent Lefèvre, Pascale Vicat-Blanc Primet, Paulo Gonçalves, Patrick Loiseau, Jean-Christophe Mignot.

The objective of this project is to promote the collaborations of Japan and France on grid computing technology. In order to promote the collaborative researches, we consider that this project is organized for the following three parts:

1. Grid interoperability and applications
2. Grid Metrics
3. Instant Grid and virtualization of grid computing resources.

RESO mainly participates to the Grid Metrics topic.
Despite the development of strong technologies in all these domains, many issues are still open about the measurement methodology itself, the emulation or simulation of Grid platforms and the understanding of Grid software stack, application performance, and fault tolerance. The Grid Metrics topics, basically gathers all researches about applications, programming models, libraries, runtimes, operating systems and network evaluation, either in synthetic environment (emulators and simulators) or real environment (real network and Grids).

8.3.2. AIST Grid Technology Research Center: GridNet-FJ associated team

Participants: Pascale Vicat-Blanc Primet, Olivier Gluck, Ludovic Hablot, Sebastien Soudan, Romaric Guillier, Olivier Gluck, Paulo Goncalves, Patrick Loiseau.

Since 2007, RESO is puisuing its collaboration with AIST through the Gridnet-FJ associated team. We followed and even increased our working program on four parts: 1) High speed transport protocol over very high speed links, 2) Bandwidth allocation and control in Grids, 3) Optimisation of MPI communications in Grids, 4) Co-design of GtrcNET-packet capture functionality.

On point 1) with the high speed testbed for protocol evaluation we have deployed within Grid5000 and which integrates GtrcNET1 and GtrcNET10, we pursued our work on TCP variants comparison. We highlight the problem of congestion level which makes TCP behave very strangely (long TCP stops) and the problem of congesting reverse traffic. During our visit to AIST, we had long discussions on the TCP stop problem. This issue has been now solved and a patch to LINUX TCP stack has been posted; We also work together on the INRIA HSTTS (High Speed Transport Test Suite) and gathered very interesting and constructives remarks from AIST colleagues.

During the year 2007, we worked with RENATER, Grid5000, SINET and NAREGI people to establish an 1Gbps optical link between the NAREGI Grid, Japan and Grid'5000, France. This link has been provisionned during the summer. It provides a means for the INRIA RESO team and the japanese collaborators to study network behaviors over real large latencies (280 ms RTT in this case) and compare the results with the performance obtained in a emulated network.

The collaboration between AIST GTRC team and INRIA RESO team on the point 2) aims at studying how BDTS, a scheduled data transfer service could benefit this flexibility offered by advance provisioning of some network path and to develop a service which use the interface. This collaboration axis has been reinforced in the context of Sebastien Soudan's long stay in september 2007.

AIST GTRC is collaborating with Pr Ishikawa team at University of Tokyo on GridMPI implementation (point 3). We are studing together the GridMPI code to understand several performance issues we observed. We are now working to understand the impact of TCP slowstart on MPI communications. Intel microbenchmark experiments allowed us to compare the difference between collective operations in the MPI implementations. it has been aggreed that INRIA RESO will access to the GridMPI implementation to integrate its own optimisation directly within the source code.

During the stay of Yuestu Kodama and Tomohiro Kudoh at ENS, the INRIA RESO team and the AIST GTRC team have been working together to extend the functionnalities of the GtrcNET-1 by adding header capture at wire-speed (point 4) . These functionnalities have been implemented and a full system for flow analysis have been designed. The GtrcNET-1 box associated with the INRIA MetroFlux system is currently used by the Metrogrid project at the Grid'5000 Lyon site to identify flow patterns in the network traffic. We plan to design and develop this system for 10Gb/s speed and deploy 10 of such equipment within Grid5000/ALADDIN during 2008-2009 years.

8.3.3. Collaboration with University of Otago, New Zealand

Participant: Laurent Lefèvre.

Laurent Lefevre has been hosted as invited researcher in Otago University (Dunedin, New Zealand) from July to September 2007. He has worked in the team of Prof. Zhiyi Huang on advanced network solutions for media streaming [57], [56].

8.4. Visitors

8.4.1. Collaboration with Queensland University of Tehcnology, Australia

Participant: Laurent Lefevre.

RESO has hosted Professor Paul Roe for 3.5 months (September - December 2007) to work on network dynamic programmability of multimedia sensors.

8.4.2. Collaboration with AIST GTRC, Japan

Participants: Tomohiro Kudoh, Yuetsu Kodama, Pascale Vicat-Blanc.

RESO has hosted Dr Tomohiro Kudoh and Dr Yuetsu Kodama for 1 week as invited researchers to work on the design of a packet capture feature within GtrcNet1 equipment integrated in the Grid5000 cluster.

9. Dissemination

9.1. Conference organisation, editors for special issues

- Pascale Vicat-Blanc is
  - General co-Chair of the ACM International Conference on High Speed Networks for Grid Applications (GridNets2007) in Lyon.
  - General co-chair and Program co-chair of the international workshop on Protocols for Very Long Distance Networks
  - member of program committees : EUROPAR2007, GRIDNETS2007, CCGrid2008, CFIP2008, PFLDNET2008, BROADNET2008.

  - Guest editor of the special issue of Future Generation Computer System, the journal of the Grid on "High Speed Networks for Grid Applications", to be edited beginning of 2008.

- Paulo Gonçalves
  - was local co-chair of GridNets 2007;
  - was member of the technical program committee for 2007 IEEE Statistical Signal Processing (SSP) Workshop;
  - is member of the program committee of ASSESS workshop in conjunction with CCGrid 2008;
  - is co-editor of “Scaling, Fractals and Wavelets", John Whiley Ed.

- Laurent Lefèvre is:
  - Workshop and Sponsor Chair of the GridNets2007 conference : First International Conference on Networks for Grid Applications, Lyon, France, October 17-19, 2007
  - program Co-Chair of ICPS 2006: International Conference on Pervasive Services, Lyon, France, June 26-30, 2006; He is co-editor in December 2007 of a Special issue from International Conference on Pervasive Services - Journal of System and Software [7]
  - co-organizer of the INRIA Booth during the Supercomputing conference (SC07) in Reno, USA, November 2007;
  - *Steering Committee* member of:
    * IEEE International Symposium on Cluster Computing and the Grid (CCGrid conference) series since 2004;
    * ICPS2007: IEEE International Conference on Pervasive Services, Istanbul, Turkey, July 2007;
    * IWAN2006: Eight International Workshop on Programmable and Active Networks, Paris, France, September 25-29, 2006 during Autonomic Networking 2006 conference;
  - member of the following Program Comittees: HotP2P’07 : Fourth International Workshop on Hot Topics in Peer-to-Peer Systems, DFMA’07 : The Third International Workshop on Distributed Frameworks for Multimedia Applications , INFOSCALE 2007 : The Second International Conference on Scalable Information Systems , ICA3PP-2007: The 7th International Conference on Algorithms and Architectures for Parallel Processing , HPDC 2007 : IEEE International Symposium on High Performance Distributed Computing, ISPDC 2007: 6th International Symposium on Parallel and Distributed Computing Conference, Symposium on Advances in Internet, Chinacom 2007, Grid2007 : The 8th IEEE/ACM International Conference on Grid Computing, Euro PVMMPI 2007 : 14th European PVM/MPI Conference, GADA’07 : Second International Conference on Grid computing, high-performAnce and Distributed Applications

- Isabelle Guérin Lassous is:
  - co-organizing the École d’Été ResCom 2007;

- Olivier Glück is:
  - a member of the following Program Committees : The 2007 International Conference on High Performance Computing and Communications (HPCC 2007), Houston, USA, 26-28 September 2007. The Rencontres Francophones du Paralléisme (RenPar’18), Fribourg, Suisse, 11-13 février 2008. The 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2008), May 19-22, 2008, Lyon, France.

– the local arrangement co-chair of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2008), May 19-22, 2008, Lyon, France.

9.2. Graduate teaching

- **started in 2007** P. Gonçalves
  Protocols and Stochastic Processes Analysis. Master Research (Ecole Normale Supérieure de Lyon, University Claude Bernard Lyon 1), lecture: 18h/year.
- **since 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.
- **since 2004** O.Glück
  Client/Server Model, Internet Applications, Network and System Administration.
  Master 2 SIR (University Claude Bernard Lyon 1), lecture 30h, others 30h.
- **since 2004** JP.Gelas
  High-speed networks, QoS and Multimedia ; Initiation to Java ; Local Area Networks .
  Master 2 SIR and CCI (University Claude Bernard Lyon 1), lecture 30h, others 40h.
- **since 2005** JP.Gelas
  Long distance networks ; Networks and Transport Protocols ; Routing ; Advanced Java and Web services Master 2 SIR (University Claude Bernard Lyon 1), lecture 45h, others 45h.

9.3. Miscelleneous teaching

- **since 2004**: O. Glück
  Computer Networks.
  Licence Informatique, (University Claude Bernard Lyon 1), lecture 30h, others 30h.
- L. Lefèvre
  High peformance Networks.
  Maitrise Informatique (Ho Chi Minh Ville University, Vietnam), 30h eq TD.
- P. Gonçalves
  Computer Science and Signal Processing. C. Shannon: from LP to the MP3 standard. Speaker at lecture "Applications of computer science to research and technological development" of École doctorale de Mathématiques et Informatique Fondamentale de Lyon. June 2007.
- L. Lefèvre
  is responsible of training periods for Research Master in ENS-Lyon
- **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.

9.4. Animation of the scientific community

- Pascale Vicat-Blanc

  - is member of the "Networks" expert committee of the CNRS.
  - participated to the "Telecom" expert committee of the ANR.
  - is within the Grid5000 project and ADT ALADDIN, member of the steering committee and co-leader of the Grid5000@Lyon site.
  - is leading the ANR CIS HIPCAL project.
  - is leading the INRIA team within the CARRIOCAS System@tic project.
  - is leading the INRIA team within the european EC-GIN project.
  - is leading the LIP team of the ANR (blanc) IGTMD project.
  - the INRIA scientific leader of a contract with ALU R&I,
  - is Scientific Advisor for the ETSI TC on Grid.
  - is leading the Technology and Business Councils on Grid Networks and Services of the ICST.

- Paulo Gonçalves was chairman of the “Time-Frequency / Time-Scale" session at GRETSI 2007 (Troyes, France).
- Isabelle Guérin Lassous is:
  - member of the CNRS TAROT action (Techniques Algorithmiques, Réseaux et d’Optimisation pour les Télécommunications);
  - the INRIA scientific leader of the european project AEOLUS (Algorithmic Principles for Building Efficient Overlay Computers);
  - the INRIA scientific leader of a contract with FT R&D, “Bandwidth problems in multihop wireless networks”;
  - member of the ARC INRIA Iramus (Radio Interface for Multihop Networks);

9.5. Participation in boards of examiners and committees

- Pascale Vicat-Blanc : president of the hearing committee of INRIA Rhône-Alpes;
- Isabelle Guérin Lassous is member of:
  - the specialists committee (section 27) of the ENS Lyon;
  - the hearing committee of INRIA Rhône-Alpes;
  - the SPECIF committee that allocates PhD awards;
  - four PhD examining boards: Nathalie Mitton (INSA de Lyon - Co-supervisor), Luigi Iannone (Paris 6), Dang Quan Nguyen (Paris 6 - reviewer) and Fanilo Harivelo (La Réunion - reviewer).
- 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 de l’UFR d’Informatique” of University Claude Bernard Lyon 1.
  - the “Conseil des Etudes et de la Vie Universitaire” of University Claude Bernard Lyon 1.
- Laurent Lefèvre
  - is member of the “commissions de spécialistes de 27ème section” of Ecole Normale Superieure (Lyon), University Antilles Guyane (Pointe à Pitre) and University Lumière (Lyon2);

  - has been reviewer of the PhD thesis of Sylvain Martin : "WASP - Lightweight Programmable Ephemeral State on Routers to Support End-to-End Applications", University of Liege, Belgium, October 2007

9.6. Seminars, invited talks

- Pascale Vicat-Blanc was:
  - Invited talk at INRIA-ALCATEL workshop meeting - Mars2007
  - Invited talk at Biarritz - CE CNRS - High Speed TCP - April2007
  - Invited talk at IM07 - Mai2007
  - Invited talk at STIC07 - November2007
  - Invited talk at the "" workshop organised by OrangeLabs : "". December 2007
- Laurent Lefèvre has been invited to give athe folowwing talks :
  - "Autonomic and programmable networks approach for supporting long latency (interplanetary) Grids", Laurent Lefèvre, Otago University, Seminar of New Zealand Distributed Information Systems group, New Zealand, August 28, 2007
  - "Towards new services and capabilities for next generation Grids", Laurent Lefèvre, Otago University, Seminar of Computer Systems group, New Zealand, August 2, 2007
  - "Next generation router-assisted transport protocols for high performance Grids : interoperability and fairness issues", Laurent Lefèvre, Ho Chi Minh Ville University, Seminary, Vietnam, May 2007

10. Bibliography

Major publications by the team in recent years

[1] F. BOUHAFS, J. GELAS, L. LEFÈVRE, M. MAIMOUR, C. PHAM, P. VICAT-BLANC PRIMET, B. TOURANCHEAU. *Designing and Evaluating An Active Grid Architecture*, in "The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications", vol. 21, n^{o} 2, February 2005, p. 315-330.

[2] B. GOGLIN, O. GLÜCK, P. VICAT-BLANC PRIMET. *An Efficient Network API for in-Kernel Applications in Clusters*, in "Proceedings of the IEEE International Conference on Cluster Computing, Boston, Massachussets", IEEE Computer Society Press, September 2005.

[3] P. GONÇALVES, R. RIEDI. *Diverging moments and parameter estimation*, in "Journal of American Statistical Association", vol. 100, n^{o} 472, December 2005, p. 1382–1393.

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

[5] D. LOPEZ PACHECO, C.-D. PHAM, L. LEFÈVRE. *XCP-i : eXplicit Control Protocol for heterogeneous inter-networking of high-speed networks*, in "Globecom 2006, San Francisco, California, USA", November 2006.

Year Publications

Books and Monographs

[6] O. AKAN, I. GUÉRIN LASSOUS (editors). *Fourth ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks (PE-WASUN)*, ACM, ACM, Chania, Crete Islands, Greece, October 2007.

[7] L. LEFÈVRE, J.-M. PIERSON. *Special issue from International Conference on Pervasive Services - Journal of System and Software*, vol. 80, n° 12, December 2007.

[8] J. TOUCH, K. KOBAYASHI, P. VICAT-BLANC PRIMET. *Special issue Hot topics in Transport Protocols for Very Long distance networks - International Journal of Computer Networks (COMNET)*, Elsevier, january 2007.

Articles in refereed journals and book chapters

[9] H. CARRÃO, P. GONÇALVES, M. CAETANO. *Contribution of multispectral and multitemporal information from MODIS images to land cover classification*, in "Elsevier, Remote Sensing of Environment", To appear, 2007.

[10] B. GOGLIN, O. GLÜCK, P. VICAT-BLANC PRIMET. *Interaction efficace entre les réseaux rapides et le stockage distribué dans les grappes de calcul*, in "Technique et Science Informatiques", 2007.

[11] E. PEREIRA DE SOUZA NETO, P. ABRY, P. LOISEAU, J.-C. CEJKA, M.-A. CUSTAUD, J. FRUTOSO, C. GHARIB, P. FLANDRIN. *Empirical mode decomposition to assess cardiovascular autonomic control in rats*, in "Fundamental & Clinical Pharmacology", vol. 21, n° 5, October 2007, p. 481–496.

[12] R. RAZAFINDRALAMBO, I. GUÉRIN LASSOUS, L. IANNONE, S. FDIDA. *Dynamic Packet Aggregation to Solve Performance Anomaly in 802.11 Wireless Networks*, in "Computer Networks", accepted, 2007.

[13] R. RAZAFINDRALAMBO, I. GUÉRIN LASSOUS. *Increasing Fairness and Efficiency using the MadMac Protocol in Ad Hoc Networks*, in "Ad Hoc Networks", accepted, 2007.

[14] G. RILLING, P. FLANDRIN, P. GONÇALVES, J. LILLY. *Bivariate Empirical mode decomposition*, in "IEEE, Signal Processing Letters", vol. 14, n° 12, 2007, p. 936–939.

Publications in Conferences and Workshops

[15] N. AYARI, D. BARBARON, L. LEFÈVRE, P. VICAT-BLANC PRIMET. *SARA: A Session Aware Infrastructure for High Performance Next Generation Cluster-based Servers*, in "ATNAC 2007 : Australasian Telecommunication Networks and Applications Conference, Christchurch, New Zealand", December 2007.

[16] N. AYARI, D. BARBARON, L. LEFÈVRE, P. VICAT-BLANC PRIMET. *Session Awareness issues for next-generation cluster-based network load balancing frameworks*, in "AICCSA07 : ACS/IEEE International Conference on Computer Systems and Applications, Amman, Jordan", May 2007, p. 180-186.

[17] N. AYARI, D. BARBARON, L. LEFÈVRE, P. VICAT-BLANC PRIMET. *T2CP-AR: A system for Transparent TCP Active Replication*, in "AINA-07 : The IEEE 21st International Conference on Advanced Information Networking and Applications, Niagara Falls, Canada", May 2007, p. 648-655.

[18] H. CARRÃO, P. GONÇALVES, M. CAETANO. *Land cover characterization through parametric modeling of intra-annual reflectance time series: a comparative study with MERIS data*, in "SPIE Europe Symposium on Remote Sensing, Firenze (Italy)", Sept. 2007.

[19] B. B. CHEN, P. VICAT-BLANC PRIMET. *Supporting bulk data transfers of high-end applications with guaranteed completion time*, in "IEEE ICC2007 International Conference on Computer Communication", IEEE, 2007.

[20] B. CHEN, P. VICAT-BLANC PRIMET. *Scheduling bulk data transfers in grid networks*, in "IEEE CCGRID 2007", IEEE, 2007.

[21] P. CONÇALVES, P. ABRY, G. RILLING, P. FLANDRIN. *Fractal dimension estimation: empirical mode decomposition versus wavelets*, in "IEEE Int. Conf. on Acoust. Speech and Sig. Proc., Honolulu, Hawaii (US)", April 2007.

[22] R. GUILLIER, L. HABLOT, Y. KODAMA, T. KUDOH, F. OKAZAKI, R. TAKANO, P. VICAT-BLANC PRIMET, S. SOUDAN. *A study of large flow interactions in high-speed shared networks with Grid5000 and GtrcNET-10 instruments*, in "PFLDnet 2007", Feb. 2007, http://wil.cs.caltech.edu/pfldnet2007/paper/Grid5000.pdf.

[23] R. GUILLIER, S. SOUDAN, P. VICAT-BLANC PRIMET. *TCP variants and transfer time predictability in very high speed networks*, in "Infocom 2007 High Speed Networks Workshop", May 2007.

[24] L. HABLOT, O. GLÜCK, J.-C. MIGNOT, S. GENAUD, P. VICAT-BLANC PRIMET. *Comparison and tuning of MPI implementation in a grid context*, in "In Proceedings of 2007 IEEE International Conference on Cluster Computing (CLUSTER)", September 2007, p. 458-463.

[25] S. KHALFALLAH, C. SARR, I. GUÉRIN LASSOUS. *Dynamic bandwidth management for multihop wireless ad hoc networks*, in "VTC-Spring, Dublin, Ireland", April 2007.

[26] P. LOISEAU, P. GONÇALVES, P. VICAT-BLANC PRIMET. *A comparative study of different heavy tail index estimators of the flow size from sampled data*, in "MetroGrid Workshop, GridNets, New York, USA", ACM Press, October 2007.

[27] D. M. LOPEZ PACHECO, L. LEFÈVRE, C.-D. PHAM. *Fairness issues when transferring large volume of data on high speed networks with router-assisted transport protocols*, in "High Speed Networks Workshop 2007, in conjunction with IEEE INFOCOM 2007, Anchorage, Alaska, USA", May 2007.

[28] E. MAGAÑA, L. LEFÈVRE, J. SERRAT. *Autonomic Management Architecture for Flexible Grid Services Deployment Based on Policies*, in "Architecture of Computing Systems - ARCS 2007, ETH, Zurich, Switzerland", vol. 4415, Springer Berlin / Heidelberg, March 2007, p. 157-170.

[29] E. MAGAÑA, L. LEFÈVRE, M. HASAN, J. SERRAT. *SNMP-based Monitoring Agents and Heuristic Scheduling for large scale Grids*, in "Grid computing, high-performAnce and Distributed Applications (GADA'07), Vilamoura, Algarve, Portugal", November 2007.

[30] G. RILLING, P. FLANDRIN, P. GONÇALVES. *Une extension bivariée pour la Décomposition Modale Empirique: Application à des bruits blancs complexes*, in "Proceedings of the 21th Colloquium GRETSI, Troyes (France)", September 2007.

[31] C. SARR, C. CHAUDET, G. CHELIUS, I. GUÉRIN LASSOUS. *Amélioration de la précision pour l'estimation de la bande passante résiduelle dans les réseaux ad hoc basés sur IEEE 802.11*, in "8es Journées Doctorales Informatique et Réseau (JDIR), Marne-la-Vallée, France", January 2007.

[32] S. SOUDAN, R. GUILLIER, L. HABLOT, Y. KODAMA, T. KUDOH, F. OKAZAKI, R. TAKANO, P. VICAT-BLANC PRIMET. *Investigation of Ethernet switches behavior in presence of contending flows at very high-speed*, in "PFLDnet 2007", Feb. 2007, http://wil.cs.caltech.edu/pfldnet2007/paper/EthernetSwitches.pdf.

[33] S. SOUDAN, R. GUILLIER, P. VICAT-BLANC PRIMET. *End-host based mechanisms for implementing Flow Scheduling in GridNetworks*, in "GridNets 2007", Oct. 2007.

Internal Reports

[34] A. CEDEYN, J.-P. GELAS, O. MORNARD, P. VICAT-BLANC PRIMET. *Document d'aide au deploiement d'IPv6 sur Grid5000*, "Also avalaible : LIP report TR2007-01", Technical Report, n^{o} 0346, INRIA, October 2007, https://hal.inria.fr/inria-00184555.

[35] R. GUILLIER, L. HABLOT, P. VICAT-BLANC PRIMET. *Towards a User-Oriented Benchmark for Transport Protocols Comparison in very High Speed Networks*, Also available as LIP Research Report RR2007-35, Research Report, n^{o} 6244, INRIA, 07 2007, https://hal.inria.fr/inria-00161254.

[36] L. HABLOT, O. GLÜCK, J.-C. MIGNOT, S. GENAUD, P. VICAT-BLANC PRIMET. *Comparison and tuning of MPI implementations in a grid context*, Research Report, n^{o} 6200, INRIA, 05 2007, https://hal.inria.fr/inria-00149411.

[37] D. M. LOPEZ PACHECO, L. LEFÈVRE, C. PHAM. *Fairness Issues When Transferring Large Volumes of Data on High Speed Networks With Router-Assisted Transport Protocols*, Also available as LIP Research Report RR2007-46, Research Report, n^{o} 6386, INRIA, December 2007, https://hal.inria.fr/inria-00195675.

[38] D. M. LOPEZ PACHECO, L. LEFÈVRE, C. PHAM. *XCP-i : eXplicit Control Protocol pour l'interconnexion de réseaux haut-débit hétérogènes*, Also available as LIP Research Report RR2007-47, Research Report, n^{o} 6385, INRIA, December 2007, https://hal.inria.fr/inria-00195634.

[39] S. SOUDAN, R. GUILLIER, P. VICAT-BLANC PRIMET. *End-host based mechanisms for implementing Flow Scheduling in GridNetworks*, Research Report, n^{o} 6205, INRIA, 05 2007, https://hal.inria.fr/inria-00150334.

[40] P. VICAT-BLANC PRIMET, J.-P. GELAS, O. MORNARD, D. MON DIVAKARAN, P. BOZONNET, M. JAN, V. ROCA, L. GIRAUD. *State of the Art of OS and Network virtualization solutions for Grids*, "Delivrable #1 : HIPCAL ANR-06-CIS-005", Technical report, INRIA, September 2007.

Miscellaneous

[41] N. AYARI, D. BARBARON, L. LEFÈVRE. *Procédés de gestion de sessions multi-flux. France Telecom R&D Patent*, June 2007.

[42] N. AYARI, D. BARBARON, L. LEFÈVRE, P. VICAT-BLANC PRIMET. *Implementation of an Active Replication based Framework for Highly Available Services*, NetFilter Workshop 2007, Karlsruhe, Germany, September 2007.

[43] J.-P. GELAS, L. LEFÈVRE. *MoonGrid: Bring Processing Power to the Moon*, ISU Annual International Symposium: Why the Moon? , Strasbourg, France, February 2007.

[44] J.-P. GELAS, L. LEFÈVRE, E. ROHMER. *Network support for long distance telerobotic platform* , Poster INRIA Booth in collaboration with Tohoku University (Japan), Supercomputing 2007, Reno, USA, November 2007.

[45] R. GUILLIER, P. VICAT-BLANC PRIMET. *TCP variants and transfer time predictability in very high speed networks*, poster, Ecole d'été RESCOM 2007, session doctorant, June 2007.

References in notes

[46] P. ABRY, R. BARANIUK, P. FLANDRIN, R. RIEDI, D. VEITCH. *Multiscale nature of network traffic*, in "IEEE Signal Processing Magazine", vol. 19, 2002, p. 28–46.

[47] P. ABRY, P. FLANDRIN, D. VEITCH. *Internet : comment réguler le trafic ?*, in "La Recherche", n^{O} 384, March 2005, p. 50–53.

[48] A. BASSI, M. BECK, F. CHANUSSOT, J.-P. GELAS, R. HARAKALY, L. LEFÈVRE, T. MOORE, J. PLANK, P. VICAT-BLANC PRIMET. *Active and Logistical Networking for Grid Computing: the e-Toile Architecture*, in "The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications", Elsevier B.V (ed),ISSN 0167-739X, vol. 21, n^{O} 1, January 2005, p. 199-208.

[49] F. BOUHAFS, J. GELAS, L. LEFÈVRE, M. MAIMOUR, C. PHAM, P. VICAT-BLANC PRIMET, B. TOURANCHEAU. *Designing and Evaluating An Active Grid Architecture*, in "The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications", vol. 21, n^{O} 2, February 2005, p. 315-330.

[50] M. CHAUDIER, J.-P. GELAS, L. LEFÈVRE. *Towards the design of an autonomic network node*, in "IWAN2005 : Seventh Annual International Working Conference on Active and Programmable Networks, Nice, France", November 2005.

[51] B. B. CHEN, P. VICAT-BLANC PRIMET. *Supporting bulk data transfers of high-end applications with guaranteed completion time*, Submitted to the IEEE ICC2007 International conference on computer communication, 2007.

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

[53] 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.

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

[55] J.-P. GELAS, L. LEFÈVRE. *MoonGrid: Bring Processing Power to the Moon*, ISU Annual International Symposium: "Why the Moon?", Strasbourg, France, February 2007.

[56] S. HASAN, L. LEFÈVRE, Z. HUANG, P. WERSTEIN. *Cross Layer Protocol Support for Live Streaming Media*, in "AINA-08 : The IEEE 22nd International Conference on Advanced Information Networking and Applications, Okinawa, Japan", March 2008.

[57] S. HASAN, L. LEFÈVRE, Z. HUANG, P. WERSTEIN. *Supporting Large Scale eResearch Infrastructures with Adapted Live Streaming Capabilities*, in "6th Australasian Symposium on Grid Computing and e-Research, Wollongong, Australia", January 2008.

[58] L. LEFÈVRE, J.-P. GELAS. *IAN2 : Industrial Autonomic Network Node architecture for supporting personnalized network services in industrial context*, in "Submitted to The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications", 2007.

[59] L. LEFÈVRE. *Heavy and lightweight dynamic network services : challenges and experiments for designing intelligent solutions in evolvable next generation networks*, in "Workshop on Autonomic Communication for Evolvable Next Generation Networks - The 7th International Symposium on Autonomous Decentralized Systems, Chengdu, Jiuzhaigou, China", ISBN : 0-7803-8963-8, IEEE Society, April 2005, p. 738-743.

[60] K. PARK, W. WILLINGER. *Self similar traffic analysis and performance evaluation*, Wiley, 2000.

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

[62] V. SANDER, F. TRAVOSTINO, J. CROWCROFT, P. VICAT-BLANC PRIMET, C. PHAM. *Networking Issues of Grid Infrastructures*, Technical report, october 2004, http://forge.gridforum.org/projects/ghpn-rg/.

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

[64] M. TAQQU, W. WILLINGER, R. SHERMAN. *Proof of a fundamental result in self similar traffic modeling*, in "Computer Communication Review", vol. 27, 1997, p. 5–23.

[65] H. TOBIET. *Panorama des réseaux utilisés et services à valeur ajoutée TEMIC*, Delivrable D2.1 Projet RNRT Temic, Technical report, February 2005.

[66] J. TOUCH, K. KOBAYASHI, P. VICAT-BLANC PRIMET. *Special issue "Hot topics in Transport Protocols for Very Long distance networks"*, in "International Journal of Computer Networks (COMNET)", january 2007.

[67] P. VICAT-BLANC PRIMET, O. GLÜCK, C. OTAL, F. ECHANTILLAC. *Emulation d'un nuage réseau de grilles de calcul: eWAN*, Research Report, n^{o} RR2004-59, LIP, ENS Lyon, Lyon, France, December 2004, http://www.ens-lyon.fr/LIP/Pub/Rapports/RR/RR2004/RR2004-59.pdf.