INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team reso

Optimized protocols and software for high performance networks

Rhône-Alpes

THEME NUM

Table of contents

1. Team **1**
2. Overall Objectives **1**
2.1. Overall Objectives 1
2.1.1. Project-team presentation overview 1
2.1.2. Context 2
2.1.3. Research area 2
2.1.4. Application domains 2
2.1.5. Methodology 3
2.1.6. Goals 3
2.1.7. Summary of the main contributions of the team in 2005 3
2.1.7.1. Direction 1: Optimized communication software and equipments 3
2.1.7.2. Direction 2: end-to-end transport and service differentiation 4
2.1.7.3. Grid Network services and applications 4
3. Scientific Foundations **4**
3.1. Optimized communication software and equipments 4
3.2. End-to-end high performance and deterministic transport and service differentiation 5
3.3. Grid Network services and applications 6
4. Application Domains **7**
4.1. Panorama 7
5. Software **8**
5.1. ORFA (Optimized Remote File-system Access) 8
5.2. eWAN (Emulator WAN) 8
5.3. SNE (Stateful Network Equipment) 8
5.4. NetLinkBench : Netlink Socket Benchmark Tool 8
5.5. Tamanoir 8
5.6. Tamanoir embedded : Active execution environment for embedded network equipments 9
6. New Results **9**
6.1. Optimized communication software and equipements 9
6.1.1. Optimized Remote File-system Access 9
6.1.2. Record and replay NIC-assisted mechanisms for MPI communications in clusters 10
6.1.3. Exploration of Network Processor technology for inline gigabit flow control 10
6.1.4. Context aware network services : supporting deployment of Java games on mobile platforms 10
6.1.5. Programmable network services for context aware adaptation 11
6.1.6. Design of an autonomic network node 12
6.1.7. High availability for clustered network equipments 12
6.1.8. Evaluation of TCP on variable bandwidth environments 13
6.1.9. Evaluating and designing estimation mechanisms for variable bandwidth environments 13
6.1.10. Router assisted network transport protocol 13
6.1.11. Integration of FEC codes into the DyRAM framework 14
6.2. Grid Network services and applications 14
6.2.1. High Performance Network Emulator 14
6.2.2. Integrating web services and programmable networks for improving flexibility of active Grids 15
6.2.3. Programmable network support of Grid middleware 15
6.2.4. Experimenting and deploying DyRAM on a grid infrastructure 15
6.2.5. Secure virtualization of the TCP/IP protocol suite: the HIP protocol 15

6.2.6. Fully distributed security in Grids 16
6.2.7. Overlay for Grids 16
6.2.8. Optimisation of network resource scheduling in grids 16
6.2.9. Traffic Isolation and Network Resource Sharing for Performance Control in Grids 16
7. Contracts and Grants with Industry 17
7.1. Myricom 17
7.2. France Telecom RD 17
7.3. INTEL 17
7.4. SUN Labs, Europe 17
7.5. 3DDL 17
7.6. Bearstech 18
8. Other Grants and Activities 18
8.1. Regional actions 18
8.1.1. Fédération Lyonnaise de Calcul Scientifique Haute Performance 18
8.2. National actions 18
8.2.1. RNRT Temic 18
8.2.2. ACI Grid GRIPPS 18
8.2.3. ACI Grandes Masses de Données GridExplorer 18
8.2.4. GRID5000 18
8.3. European actions 19
8.3.1. Programmes d'Actions Intégrées Amadeus with Linz Univ., Austria 19
8.4. International actions 19
8.4.1. Programme d'Actions Intiégrées "Fast" with Queensland University of Technology 19
8.4.2. NSF-INRIA with Aerospace Organization 20
8.4.3. AIST Grid Technology Research Center 20
8.5. Visitors 20
8.5.1. Collaboration with AIST GTRC, Japan 20
9. Dissemination 20
9.1. Conference organisation, editors for special issues 20
9.2. Graduate teaching 22
9.3. Miscelleneous teaching 22
9.4. Animation of the scientific community 23
9.5. Participation in boards of examiners and committees 24
9.6. Seminars, invited talks 24
10. Bibliography 25

1. Team

Head of project-team
Pascale Vicat-Blanc Primet [Directrice de Recherche INRIA]

Administrative Assistant
Isabelle Antunes Pera [ENS]
Sylvie Boyer [INRIA]

Staff member INRIA
Laurent Lefèvre [Chargé de Recherches 1ère classe INRIA]

Staff member Université Claude Bernard Lyon1 (UCB)
CongDuc Pham [Maître de conférences, HDR, until 31/9/05]
Olivier Glück [Maître de conférences]
Jean-Patrick Gelas [ATER (since 1/10/2005)]

Project technical staff
Martine Chaudier [Temporary Engineer INRIA CDD - project RNRT TEMIC (from 15/3/2005 to 11/12/2005)]
Jean-Patrick Gelas [Temporary Engineer INRIA CDD - project RNRT TEMIC (from 15/1/2005 to 31/9/2005)]
Jean-Christophe Mignot [Permanent Engineer CNRS]
Stéphane D'Alu [Temporary Engineer, co-Remap ENS CDD - projet Grid5000]

Postdoctoral position
Jingdi Zheng [INRIA Postdoc, 2004-2005, until 15/7/05]

Ph. D. students
Narjess Ayari [PhD student with France Telecom R-D - CIFRE - since 1/6/2005]
Julien Laganier [PhD student - 2002/2005 - CIFRE SUN]
Dino Lopez Pacheco [PhD student - 2004/2007 - Mexican Goverment Grant]
Brice Goglin [PhD student - 2002/2005 - BDI CNRS]
Antoine Vernois [co-Remap and IBCP - PhD student - 2002/2005 - MENRT ACI GRID]

Student internship
Cyril Otal [DEA DIF Student 1/3/05 - 15/7/05]
Sebastien Soudan [Ecole Centrale de Lyon Student 1/4/05 - 31/9/05]
Chien-Jon Soon [Student in Queensland University of Technology, Brisbane, Australia, 1/6/2005-1/10/2005 - co-advised by Laurent Lefevre]

External collaborator
Loic Prylli [Myricom]

Long term visiting scientists
Lakhdar Derdouri [Visitor from University of Constantine, Algeria, 29/11/2005 - 26/12/2005]

2. Overall Objectives

2.1. Overall Objectives

2.1.1. 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 twelve members in average, including four permanent researchers and teaching researchers. RESO is part of the " Numerical Systems " theme of the INRIA, part of the B subsection: Grids and high-performance computing. The research activities of the RESO project fits the first priority challenge of the INRIA's strategic plan: "design and master the future network infrastructures and communication services platforms" . In this context, RESO is focusing

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

2.1.2. Context

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

2.1.3. Research area

Our work follows two major research axes :

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

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

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

2.1.4. Application domains

RESO applies its research to the domains of high performance computing and to Grid communications. The geographical topology of the Grid depends on the distribution of the community members. Though there might be a strong relation between the entities building a virtual organization, a Grid still consists of resources owned by different, typically independent organizations. Heterogeneity of resources and policies is a fundamental result of this. Web Service Resource Framework (WS-RF), the coupling of the notion of resource and web service, has recently been introduced. It has added the performance issue to the re-usability, interoperability, and openness advantages of web services. Grid services involve operations and strategies from

application layer down to network layer, with service agreements defined at application layer and middleware developed for the communication between layers. In a typical implementation scenario, the grid middleware provisions the resource, and passes the delivery criteria to the network services. The network, accordingly, follows up to enforce the appropriate data transfer. In a Grid, the network performance requirements are very high and may strongly influence the performance of the whole distributed system. The construction of grid networks over the optical transport layer tackles the problem of communication performance from the transport medium perspective. However, our vision is that Grid applications, due to the heterogeneity and large scale factors, will continue to use traditional IP packet protocols, at least in the end systems and will rely on a complex interconnection of heterogeneous networks. In such context end-to-end flow performance 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 avoid the applications to be network-aware, to simplify the programming and to optimize the execution of their communication parts while fully exploiting the capacities of the evolving network infrastructure.

2.1.5. Methodology

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

2.1.6. Goals

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

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

2.1.7. Summary of the main contributions of the team in 2005

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

2.1.7.1. Direction 1: Optimized communication software and equipments

- Design and proposition of a new networking subsystem architecture built around a packet classifier executed in the Network Interface Controller (NIC). Development of the KNET software in collaboration with SUN;
- Study and design of efficient remote data access for clusters, that maximizes the underlying network utilization. Development of the ORFA (Optimized Remote File-system Access) software prototype on Myrinet networks. Integration of new features in the new Myrinet programming interface, MX, in collaboration with Myricom;

- Development of a high performance active network architecture (Tamanoir) and associated tools (Echidna, Pangolin). Proposition of load balancing functions in cluster-based active routers;
- Exploration of the Network Processor technology for designing high performance grid overlay gateways;
- Validation of Tamanoir through internal and external projects (IBP, deployment of FPTP (LAAS, Toulouse), deployment of programmable nodes (3DDL project));

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

- Contribution to the analysis of the limitation of TCP protocol for very long distance high speed networks within the GGF and pfldnet community;
- Optimization algorithms for network resource sharing and flow scheduling in grids.
- Evaluation and proposition around the XCP High Performance transport protocols;

2.1.7.3. Grid Network services and applications

- Contribution to the design and development of the GRID5000, national Grid testbed;
- Contribution to the design and development of the GridExplorer, national Grid emulator;
- Contribution to the standardization of the HIP (Host Identity Protocol) and to the definition of the HIPernet architecture for a fully distributed network security in Grids;
- Definition of a new architecture based on an overlay approach for offering a deterministic data transfer service to grid applications.
- Emulation instrument design and realization for high performance protocols and Grid software. We designed, developed and evaluate the eWAN software for configuring a cluster in a virtual high performance grid network cloud.

3. Scientific Foundations

3.1. Optimized communication software and equipments

Participants: Olivier Glück, Brice Goglin, Laurent Lefevre, Pascale Vicat-Blanc Primet, Cyril Otal, Sebastien 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. Parallel applications require both efficient communication between distant application tasks and fast access to remote storage. High performance distributed file systems have special requirements that have not really been considered when designing most underlying network access layers. While usual application communications should obviously occur at user-level, distributed file system were initially implemented in the kernel to supply transparent remote accesses. They were designed for traditional networks and caching was used to compensate the high latency. In a cluster environment, two directions are studied to improve the performance of distributed file systems: distributing the workload across multiple servers or efficiently using the low latency and high bandwidth of the underlying high-speed network. We studied this second direction with our ORFA (*Optimized Remote File-system Access* [51]) user-level

implementation and its kernel port, ORFS (*Optimized Remote File-System* [22]), as a distributed file system test platform to improve the usage of high-speed interconnects in the context of remote file access. High performance applications running on high-speed interconnects require both efficient communication between computing nodes and fast access to the storage system. Making the most out of these networks to access remote files requires a good interaction between their highly specific software interface and the special requirements of distributed file systems. In this research axis we explore the de-localization of network functionalities in dedicated equipments (programmable NICs) or intermediate nodes (programmable network equipment). We studied several techniques based on new functions in the network interface controllers to maximize the execution efficiency of the operating system's communication software. In particular, the main proposition of our work (KNET software suite) is to place a packet classifier in the network interface controller in order to smartly spread incoming network streams across the processors (or threads) of connected servers. In order to support network functions in the network, we propose a high performance active network environment execution architecture (Tamanoir software suite). This architecture is based on various layers adapted on services and applications requirements : NICs for no state ultra lightweight services, kernel for few state lightweight services, user space for middle service and distributed resources for CPU/storage consuming services. We propose various adaptive solutions (load balancing, fault tolerance) to efficiently deal with heterogeneous services and applications.

3.2. End-to-end high performance and deterministic transport and service differentiation

Participants: Pascale Vicat-Blanc Primet, Jingdi Zeng, Dino Lopez Pacheco, Cong-Duc Pham.

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

The conservative behavior of TCP with respect to congestion in IP networks (RFC 2581) is at the heart of the current performance issues faced by the high-performance networking community. Several theoretical and experimental analysis have shown that the dynamics of the traditional feedback based approach is too low in very high speed networks that may lose packets. Consequently network resource utilization is not optimal and the application performance is poor and disappointing. Many Grid-enabled computing applications wish to transfer large volumes of data over wide area networks and require high data rates in order to do so. However, Grid-enabled applications are rarely able to take full advantage of the high-capacity (2.5 Gbit/s, 10 Gbit/s and upwards) networks installed today. Recent data for Internet 2 show that 90% of the bulk TCP flows (defined as transfers of at least 10 Megabyte of data) use less than 5 Mbit/s, and that 99% use less than 20 Mbit/s out of the possible 622 Mbit/s provision. There are many reasons for such poor performance. Many of the problems are directly related to the end system, to the processor and bus speed, and to the NIC with its associated driver. TCP configuration (e.g., small buffer space or features such as SACK being improperly negotiated) will have a significant impact. TCP itself was designed first and foremost to be robust and when congestion is detected, TCP accommodates the problem but at the expense of reduced performance. There are also design problems with TCP itself. For example, for a standard TCP connection with 1500-byte packets and a 100 ms round-trip time, achieving a steady-state throughput of 10 Gbit/s would require an average congestion window of 83,333 segments, and a packet drop rate of at most one congestion event every 5,000,000,000 packet (or equivalently, at most one congestion event every 1 2/3 hours). HighSpeed TCP [47] and Scalable TCP [54] increase the aggressiveness in high-throughput situations while staying fair to standard TCP flows in legacy contexts. **FAST** [53] leverages the queueing information provided by round-trip time variations, in order to efficiently control buffering in routers and manage IP congestion optimally. These propositions are actively analyzed and

experimented by the international community. RESO participates to the elaboration of a survey on protocols other than standard TCP in the framwork of the Data Transport research group of the Global Grid Forum [41]. RESO is organizing in 2005, the third edition of the leading international workshop in this domain (see http://www.ens-lyon.fr/LIP/RESO/pfldnet2005). Several issues have been already enlightened. Considering the traditional feedback loop will not scale with higher rate level under loss or congesting traffic conditions, it seems judicious to start examining alternative radical solutions.

On the other hand, flows crossing the IP networks are not equally sensitive to loss or delay variations. Since several years, research effort has been spent to solve the problem of the heterogeneous performance needs of the IP traffic. A class of solutions considers that the IP layer should provide more sophisticated services than the simple best-effort service to meet the application's quality of service requirements. Quality of service has been studied in IP networks in the context of multimedia applications [46]. Various complementary solutions have to be integrated to carry end-to-end quality of service to grid applications to assure an efficient usage of the interconnected computing resources [49]. Solution like Diffserv exhibits three types of limitations we are considering:

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

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

3.3. Grid Network services and applications

Participants: Pascale Vicat-Blanc Primet, Jingdi Zheng, Olivier Glück, Julien Laganier, Jean-Christophe Mignot.

The purpose of Computational Grids is to aggregate a large collection of shared resources (computing, communication, storage, information) to build an efficient and very high performance computing environment for data-intensive or computing-intensive applications [50]. But generally, the underlying communication infrastructure of these large scale distributed environments is a complex interconnection of multi-IP domains with changing performance characteristics. Consequently *the Grid Network cloud* may exhibit extreme heterogeneity in performance and reliability that can considerably affect the global application performance. Performance and security are the major issues grids encountered from a technical point of view.

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

- Measuring and monitoring the end-to-end performance helps to characterize the links and the network behavior. Network cost functions and forecasts, based on such measurement information, allow the upper abstraction level to build optimization and adaptation algorithms.

- Optimally using network services provided by the network infrastructure for specific grid flows is of importance.
- Creating enhanced and programmable transport protocols adapted to heterogeneous data transfers within the grid may offer a scalable and flexible approach for performance control and optimization.
- Modeling, managing and controlling the grid network resource as a first class resource of the global environment: transfer scheduling, data movement balancing...

4. Application Domains

4.1. Panorama

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

RESO applies its research to the domains of high performance Cluster and Grid communications. Existing GRID applications did already identify potential networking bottlenecks, either caused by conceptual or implementation specific problems, or missing service capabilities. We participated to the elaboration of the first GGF document on this subject [58] [57], [59]. Loss probability, important and incompressible latencies, dynamic behavior of network paths question profoundly models and technic used in parallel and distributed computing [48]. The particular challenge arises from a heavily distributed infrastructure with an ambitious end-to-end service demand. Provisioning end-to-end services with known and knowable characteristics in a large scale networking infrastructure requires a consistent service in an environment that spans multiple administrative and technological domains. We argue that the first bottleneck is located at the interface between the local area network (LAN) and the wide area network (WAN). RESO conducted several actions in the field of Grid High Performance Networking in the context of the GGF, the 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.

- 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. Participating to the design, deployment and usage of such high performance experimental Grid testbed allows us to evaluate and measure the benefit that grid middleware and applications can get from enhanced networking technologies. The experience and expertize we get from this work are a tremendous gain for our research on performance bottlenecks.
- Grid 5000 is a national initiative aiming at providing a huge experimental instrument to the grid software research community. Lyon, with RESO and GRAAL projects, is part of this initiative. RESO is closely involved in the design and deployment of the testbed, and responsible for the networking aspects.
- We participate to the definition of the Grid Explorer physical architecture and to the design of the configuring, tuning and monitoring software. Grid Explorer, the largest cluster of Grid 5000 platform will be a very large scale instrument for grid software evaluation.

5. Software

5.1. ORFA (Optimized Remote File-system Access)

Keywords: *SAN networks*, *filesystem*.

Participants: Brice Goglin (contact), Olivier Glück.

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

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

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

5.2. eWAN (Emulator WAN)

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

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

EWAN [60] is a software and hardware tool for configuring and programming a large PC cluster in a wide area network emulation instrument. High performance, fine parameter tuning and a great utilization flexibility are the main proposed features of this experimental tool. Details are available at http://www.ens-lyon.fr/LIP/RESO/Software/EWAN/

5.3. SNE (Stateful Network Equipment)

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

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

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

5.4. NetLinkBench : Netlink Socket Benchmark Tool

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

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

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

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

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

5.5. Tamanoir

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

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

Tamanoir is an open source software environment for high speed active networks. Available on the web and protected by APP (Agence Francaise de Protection des Programmes). TAMANOIR is distributed within the RNTL eToile suite. It is used by partners in RNTL eToile Project and in the collaboration with 3DDL company (for supporting deployment of Java based games on mobile platforms). All details on

Tamanoir are available at http://www.ens-lyon.fr/LIP/RESO/Tamanoir

5.6. Tamanoir embedded : Active execution environment for embedded 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). 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.

6. New Results

6.1. Optimized communication software and equipements

6.1.1. Optimized Remote File-system Access

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

Data storage in a cluster environment requires dedicated systems that are able to sustain high bandwidth needs and serve many concurrent clients. Several projects have already been proposed to address this issue. PVFS, GPFS or Lustre provide parallel file systems whose scalability is ensured by data stripping and workload sharing across several servers. We study the link between clients and these systems in order to maximize the underlying network utilization. Indeed cluster nodes are connected through a high bandwidth low latency network such as Myrinet, whose features lead us to the idea of using them for data storage. ORFA (*Optimized Remote File-system Access*) was developed on Myrinet networks to provide an efficient access to remote data. The fully transparent user-level client [52] allows any legacy application to saturate the physical link by accessing remote files. The need to cache metadata on the client's side leads to the idea of developing ORFS (*Optimized Remote File-System*), the port of ORFA into the Linux kernel. Besides, the use of ORFA-like techniques in parallel filesystems should enhance their performance to make the most out of the underlying network. This work also showed that the now well-known memory registration model that is used on asynchronous network interface such as Myrinet does not fit file system implementation needs. Maintaining a registration in the kernel to enable high-performance non-buffered remote file access has required to patch the Linux kernel so that an external module might be notified of address space modifications. Collisions between virtual addresses of different processes have been avoided by using a modified GM firmware in

the network interface card, making the registration cache as efficient in the kernel than in user-level in the ORFA implementation. Besides, buffered accesses have been improved by replacing the traditional memory registration with physical address based primitives which are much more suitable for such an environment. These make the ORFS implementation very efficient [51]. We have worked and we are still working with Myricom to integrate all this work in their new driver, MX (Myrinet Express), so that the interaction between it and file systems implementations will be much easier and efficient. Results have been published in [24] and in [22]. This whole work led Brice Goglin to pass his Ph.D. Thesis [12] in october 2005.

6.1.2. Record and replay NIC-assisted mechanisms for MPI communications in clusters

Keywords: *Communication system, Operating systems, programmable network cards.*

Participant: Laurent Lefèvre.

Nondeterministic program behavior leads to different results in successive program executions, even if the same input data is provided. For this reason, re-executions of a program (as needed during cyclic debugging) are only possible, if certain precautions are taken. The most common solution is provided by record&replay mechanisms, where an initial record phase is used to extract characteristic behavioral data, which is afterwards used to control equivalent executions during subsequent replay phases. With the novel record&replay mechanism on Myrinet network interface cards (NIC), program perturbations during the record phase are avoided by performing the initial monitoring activities directly on the NIC. This approach ensures, that the CPU of the computing nodes is not affected by the monitoring activities, while subsequent re-executions can still be controlled with the data collected on the NICs. [26]

6.1.3. Exploration of Network Processor technology for inline gigabit flow control

Keywords: *Flow control, IXP2400, intelligent NICs, network processors.*

Participants: Sebastien Soudan, Pascale Vicat-Blanc Primet.

To control the networks links and flows at gigabit speed it is necessary to implement advanced mechanisms at very low level: in FPGA or network processors. For optimizing network resource utilisation and end to end performance without modifying actual TCP/IP protocols we propose to schedule and control the data movements (ie flows) inline. For this we implemented a dedicated controller able to test the conformance in time and rate of gigabit flows. We have programmed and evaluated the IXP 2400 Network processor provided by our collaboration with Intel and which have been integrated within the Grid 5000 testbed. Obtained performance are very impressive as we were able to attain 2.4Gb/s from the 2.5Gb/s specified by the provider, without any overhead on the central CPU. This exploration shows that network processor technology is very promising but development environments are very unmature. They need to be improved for making this solution a valuable solution for innovative network protocols or mechanisms prototyping and evaluation as well as for their industrialisation.

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

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

Participants: Aweni Saroukou, Laurent Lefèvre.

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

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

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

- to reduce development costs and the complexity for managing a version of a game for each mobile class. The active node will adapt the files on the fly;

- to reduce the usage of bandwidth and interactions between clients and games server;
- to efficiently support deployment of games without adding too much latency on real networks.

We design the architecture of an active transcoding service (ActiveWapS) deployed inside the Tamanoir Execution Environment. This service transforms on the fly, parts of the games (JAD files) in order to adapt them to target mobile phones. We also validate this approach on a local platform with emulated wireless network [29] [55].

6.1.5. Programmable network services for context aware adaptation

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

Participants: Martine Chaudier, Laurent Lefèvre.

Traditional industrial maintenance process (ie 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 [39]) in which we are currently involved with different academic and industrial partners, our team designed devices (including hardware and software) easily deployable in an industrial context, and also easily removable at the end of the maintenance contract. TEMIC proposes a hardware and software platform of collaborative remote maintenance : maintenance staff may work remotely and in collaboration with other experts. The TEMIC platform integrates various technologies:

- Networks: they may be wire (LAN, WAN) or wireless (GPRS, WiFi, Bluetooth). They present different characteristics (rate...).
- Terminals: PC, laptop, PDA, mobile phone. Resources, display capacity, communication protocol, are different.
- Multimedia applications: VOD, videoconference, file download. They run with specific protocols and video/audio formats.

This heterogeneity needs 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 Execution Environment in Java. Active services applying on multimedia streams crossing the network node may realise 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 scenarios[39], [36]. They were planned by the TEMIC project team to be used by a company through a maintenance contract on a restricted industrial area. The first is called "Gathering and Survey" and it deals with the transport of survey data coming from media sensors towards mobile terminals, and the second is "Analysis and Collaboration" and concerns the multimedia data exchange between mobile devices and a collaborative server.

At this time, three active services have been developed for this project. They are designed to adapt multimedia data on the fly. Active network services are deployed on active network nodes. They can be

dynamically loaded from a specific server (a service repository), or they can be deployed from next to next between the active network nodes. The location of the active network nodes (and so the place on the network where data will be adapted) is also an important point to take into account. When the adaptation service contributes to reduce the amount of data transmitted on the network (by degrading the encoding quality or by decreasing the frame size), it is judicious to set the active node nearest the data source (the server) in order to reduce the used bandwidth early enough, before data are sent on LAN and WAN. Due to resources characteristics of both server and mobile devices, multimedia data are stored on the server at the highest available quality and they will be adapted (possibly reduced) when transmitted to less resources devices. Our active services use the Java Media Framework (JMF), version 2.1.1 for the media adaptation. The JMF API enables the display, capture, encoding, decoding and streaming of multimedia data in Java technology-based applications. We conducted several experiments to evaluate the impacts of adaptation on the mobile client and on the network, and the performances of our industrial network node. For these tests we considered the transmission of video streams to different mobile devices and so the adaptation on the fly according to the capabilities of these devices. 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.

6.1.6. Design of an autonomic network node

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 [21], [42]. We evaluated the capabilities of IAN^2 in terms of computing and networking resources and dynamic re-programmability.

6.1.7. High availability for clustered network equipments

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

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

In operational networks, the availability of some critical elements like gateways, firewalls and proxies must be guaranteed. High availability allows service architectures to meet growing demands and to ensure uninterrupted service. These architectures deal with different flows of data, which can be classified in signaling and interactive non elastic data, and in non signaling elastic data. Typical elastic flows are file transfers such as those in email and world wide web services. Interactive flows carry streams of voice or video over virtual

circuits or sessions that are established and controlled on different networks via signaling protocols. Typically, today's VoIP service architectures need to meet the high availability requirements to compete with Public Switched Telephone Networks, which are reliable and offer short failover times. [20]

6.1.8. Evaluation of TCP on variable bandwidth environments

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

Participants: Dino Martin Lopez-Pacheco, Congduc Pham.

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

6.1.9. Evaluating and designing estimation mechanisms for variable bandwidth environments

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

Participants: Dino Martin Lopez-Pacheco, Congduc Pham.

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

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

6.1.10. Router assisted network transport protocol

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

Participants: Dino Martin Lopez-Pacheco, Laurent Lefèvre, Congduc Pham.

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 [30].

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

6.1.11. Integration of FEC codes into the DyRAM framework

Keywords: *FEC, Multicast, reliability*.

Participants: Sylvain Dattrino, Congduc Pham.

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

6.2. Grid Network services and applications

6.2.1. High Performance Network Emulator

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

Participants: Cyril Otal, Olivier Glück, Pascale Primet.

The Grid aims at expanding the cluster based parallel computing paradigm towards large scale distributed systems based on IP networks. EWAN [60] is a high performance network environment emulator. It takes place in the research effort on computer grids, aggregations of computer resources inter-connected by a wide area network. EWAN offers an emulation framework needed by experiments in this field, bringing a great flexibility, a high level of performance and a precise control. eWAN provides features to control key characteristics of grid or transport protocol evaluation scenarii. To achieve correct performance and enable test at gigabit speed with minimum noise and overhead, the different functional entities are deployed on separate, non shared and reserved machines and local networks. Compared to Emulab, the particularity of EWAN is to exploit within a limited time (one to few ours depending on the experiment needs) any cluster composed of several tens or more common PCs. The main functionnalities that have been identified are:

- link emulation with key characteristics control: like latency (from 1ms to 500ms) loss rate (with different distributions), capacities (from (10Mb/s to 10Gb/s)
- topology emulation (chain, star, ring, mesh, dumpbell, fish bone...)
- IP version (v4 and v6) and jumbo frame support
- traffic generation
- process application running
- traffic and performance monitoring and logging

The eWAN software can be divided into two main parts: an interface for creation of simple topologies and an engine for deploying every sort of topologies. We have evaluated several network emulation solutions and have configured a 12 nodes cluster to test EWAN software on this cluster. As emulation solutions, we have compared Nistnet, netem and GtrcNET. Nistnet is a well known software to emulate network link, netem is an equivalent recently included in the Linux kernel and GtrcNET is an hardware network emulator developed by the AIST. EWAN manages all the three solutions and allows the user to choose one of them. All stuff (source code, documentation, results, ...) about EWAN can be found at http://www.ens-lyon.fr/LIP/RESO/Software/EWAN/. A paper has been written in collaboration with Tomohiro Kudoh and Yuestu Kodama for the Pfldnet2006 conference.

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

Keywords: *Web services, programmable networks.*

Participants: Laurent Lefèvre, Chien-Jon Soon (Queensland University of Technology, Brisbane, Australia), Paul Roe (Queensland University of Technology, Brisbane, Australia).

Active Grids[13], [14] are a form of grid infrastructure where the grid network is active and programmable. These grids directly support applications with value added services [28] 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"). [56]

6.2.3. Programmable network support of Grid middleware

Keywords: *Globus, programmable networks.*

Participant: Laurent Lefèvre.

Efficiently and dynamically supporting Grid middleware with programmable and active network remains a challenging task. We explore some solutions to support the Globus Grid middleware with the Tamanoir active network environment [13], [14]. Inside the Globus XIO API (Globus 3.2), we develop some transform and transport drivers to propose network services alternatives for Grid applications.

6.2.4. Experimenting and deploying DyRAM on a grid infrastructure

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

Participants: Faycal Bouhafs, Congduc Pham.

Today's computational grids are using the standard IP routing functionality, that has basically remained unchanged for 2 decades, considering the network as a pure communication infrastructure. With the grid's distributed system point of view, one might consider to extend the *commodity Internet*'s basic functionalities. Higher value functionalities can thus be offered to computational grids. In this work, we report on our early experiences in building application-aware components for multicast and in defining an active grid architecture that would bring the usage of computational grid to a higher level than it is now (mainly batch submission of jobs). To illustrate the potential of this approach, we first present how such application-aware components could be built and then some experiments on deploying enhanced multicast communication services for the grid. Results published in [14] show that reliable multicast could deploy specific services based on the grid application needs.

6.2.5. Secure virtualization of the TCP/IP protocol suite: the HIP protocol

Participants: Julien Laganier, Pascale Vicat-Blanc Primet.

RESO is specifically following Host Identity Protocol (HIP) activities within the IETF because HIP has been identified has one of the major breakthrough technologies to enable secure virtualization of the TCP/IP protocol suite. Project RESO/Holonet seeks to network virtualization because of its applications to dynamic reconfiguration of the grid infrastructure. Three Working Group documents describing 'DNS extensions' (by Nikander & Laganier) and 'Rendezvous extensions' (by Laganier & Eggert) for HIP have been edited. The 'DNS extensions' allows a node to store HIP-related material in the DNS. This include its Rendezvous Server's IP address(es) or DNS names, as well as its Host Identity and its Host Identity Tag. The 'Rendezvous extensions' allows a HIP node to use another node, its Rendezvous Server (RVS), to maintain its reachability when changing its network attachment. A HIP node trying to communicate with such a HIP node would typically initiate communication towards its RVS, which will relay the initial packets of the HIP exchange

to its client. Then the two nodes can communicate without further assistance from the RVS. This allows fast moving node to maintain reachability even if there is too much update latency in the name-to-address lookup service.

6.2.6. Fully distributed security in Grids

Participants: Julien Laganier, Pascale Vicat-Blanc Primet.

Security in Grid environments appeals for fundamental primitives like the secure establishment of dynamic and isolated virtual trust domains. The security mechanisms currently used are generally based on a Public Key Infrastructure global to the grid environment, and a mix of global and local access control policies used to make an authorization decision. We think that such approaches do not scale well with the number of participating domains and entities. We propose a decentralized approach to the security in grid environment because we think it can better cope with its inherently distributed nature. The combination of network and operating system virtualization (Supernets) with the Host Identity Protocol (HIP) and Simple Public Key Infrastructure (SPKI) delegation/authorization certificates allows to create virtual trust domains onto multiple shared computer nodes connected by an untrusted network. Because our solution is fully decentralized, it can better adapt the vast diversity of trust relationships in the real world and has a better scalability with respect to the number of entities involved. Finally, the HIP basement of the architecture brings location independence to upper layers protocols experience, thus allowing to build lon-term communities despite entity movements[27].

6.2.7. Overlay for Grids

Keywords: *bulk data transfer*, *flow control*, *grid networking*, *overlay networks*.

Participants: Jingdi Zeng, Pascale Vicat-Blanc Primet.

As Grids allow users to share resources over long distance networks, critical, are the degrees by which data are effectively transported among these resources. A diversity of approaches, with enhanced transport protocols and different transport mediums, have been proposed for high throughput. Within these approaches, bridging grid applications and network services is critical for users to control data transfer and application performance. Adopting an overlay infrastructure, we are proposing a new network architecture for grid data transport. We identified three important issues of the infrastructure: the global network resource scheduling of a grid overlay network, the flow control mechanism of grid local networks, and, the edge-to-edge transfer performance guarantee of grid overlay routers. Further, grid data bursts are introduced to reliably improve transport control and data delivery [35].

6.2.8. Optimisation of network resource scheduling in grids

Keywords: *grid computing*, *heuristics*, *network resource*, *optimization*, *resource scheduling*.

Participants: Jingdi Zeng, Pascale Vicat-Blanc Primet.

While grid computing reaches further to geographically separated supercomputers, clusters, data warehouses, and disks, its previous loosely-coupled connections over the public network, that is, the Internet, has gained attention. What has been under pursuit, for these connections, is the end-to-end performance guarantee. From the perspective of resource sharing, we look at managing communication resources used by grid sites. A grid network model have been defined. Two resource request scenarios have been identied and studied. The corresponding solutions, after being proven as NP-complete, are obtained with heuristic algorithms. Simulation results show that the heuristics achieve fairly satisfying performance [31].

6.2.9. Traffic Isolation and Network Resource Sharing for Performance Control in Grids

Keywords: *diffserv*, *flow scheduling*, *grid computing*, *network resource sharing*.

Participants: Jingdi Zeng, Pascale Vicat-Blanc Primet.

Grid applications pose new demands on end-to-end data transfer performance control. Data-intensive grid applications rely on the underneath network to have distributed computational and storage resources work in concert. From computational grids to data grids, the focus of resource utilization is shifting from computing

power to network resources. We investigate network resource sharing in grids, especially data grids. We studie traffic characteristics and quality of service(QoS) mechanisms of grid applications. A hybrid approach, which combines classical QoS differentiation and advance resource reservation, is proposed to meet grid application performance requirements [40], [33].

7. Contracts and Grants with Industry

7.1. Myricom

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

This long-term collaboration between our team and US based Myricom company is focused on their software Myrinet suites. The old driver (GM) was used as a experimentation platform for the ORFA (*Optimized Remote File-system Access*) software prototype and its kernel port, ORFS (*Optimized Remote File-System*). This work is now being integrated in the new driver (MX) to make the interaction between file systems and Myrinet software layers much easier and efficient. Brice Goglin is in a Postdoc position at Myricom since november 2005.

7.2. France Telecom RD

Participant: Laurent Lefèvre.

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. [20]

7.3. INTEL

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

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

7.4. SUN Labs, Europe

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

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

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

7.5. 3DDL

Keywords: *java*, *programmable networks*.

Participant: Laurent Lefèvre.

RESO has established a long term collaboration with 3DDL SME. This collaboration concerns the design and deployment of software components inside the network in order to support the deployment of mobile applications on heterogeneous terminals[29] [55]. Funded by Région Rhone-Alpes with collaboration of LIRIS, INSA Lyon.

7.6. Bearstech

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

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

Since 2004, RESO is launching a collaboration with this young company targeted on embedded computers and network equipments. This collaboration has allowed an improved design of "Tamanoir embedded" software suite.

8. Other Grants and Activities

8.1. Regional actions

8.1.1. Fédération Lyonnaise de Calcul Scientifique Haute Performance

Participants: Laurent Lefèvre, Cong-Duc Pham.

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

8.2. National actions

8.2.1. RNRT Temic

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

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

8.2.2. ACI Grid GRIPPS

Participants: Antoine Vernois, Pascale Vicat-Blanc.

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

8.2.3. ACI Grandes Masses de Données GridExplorer

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

(2003-2006) : The aim of this project is to create a large scale grid and network emulator. RESO is involved in the design of the platform and is interested in designing a high performance transport protocol test methodology in this environment. EWAN [60], our high performance network emulator, is one of the main RESO contributions to this project. Pascale Vicat-Blanc is responsible of the network theme. RESO has participated to the definition of the architecture and technical choices of cluster hardware.

8.2.4. GRID5000

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

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

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

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

With the help of funding of INRIA Rhone-Alpes, the plateform has been upgraded with 150 processors and a 10Gb/s core lan. The Grid5000 of Lyon comprises now around 300 processors interconnected with a network of 250Mb/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. She defines with Renater the new dark fiber core infrastructure that will enable to interconnect each sites with 10G/s access links. She is also working for the interconnection of the Grid5000 project and the japanese Naregi project. Olivier Glück, Stéphane D'Alu and Jean-Christophe Mignot are members of the national technical committee of GRID'5000. Actual funding: 530K euros

8.3. European actions

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

Participant: Laurent Lefèvre.

RESO is involved in a long term collaboration (1999-2000, 2001-2003, 2004) with University of Linz, Austria (Prof. J. Volkert team) on the field of "Deporting services on Network Programmable cards". Supported by French Ministry of Foreign affairs. During 2004, RESO has hosted Dieter Kranzlmuller for a 3 weeks period. Even if this Amadeus project ended in 2004, the collaboration with Univ. Linz is still active in 2005 [26].

8.4. International actions

8.4.1. Programme d'Actions Intiégrées "Fast" with Queensland University of Technology

Participants: Laurent Lefèvre, Paul Roe (Queensland University of Technology, Brisbane, Australia, Australia).

This project focuses on the design of Web Services based on Programmable Networks Infrastructure (WeSPNI). This collaboration between RESO team and Programmi ng Language and System group (PLAS) in Queensland University of Technology (Brisbane, Australia) aims to bring together researchers able to design next gener ation of overlay networks. We observe a real convergence between Grid infrastructure and Web Service solutions. Based on the Open Grid Service Infrastructure , Grid researchers have proposed the WSRF (Web Service Resource Framework) where Web Services naturally fit in Grid requirements. This collaboration exploits this convergence by providing network solutions adapted to Grid requirements. This Fast project is supported by French Ministry of Foreign affairs (2005-20 06). [56]

8.4.2. NSF-INRIA with Aerospace Organization

Participant: Laurent Lefèvre.

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

8.4.3. AIST Grid Technology Research Center

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

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

8.5. Visitors

8.5.1. Collaboration with AIST GTRC, Japan

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

RESO has hosted Dr Tomohiro Kudoh and N. Yuetsu for 1 week as invited researchers from 5/12/05 to 9/12/05 to work on experimentation with GtrcNet1 equipment integrated in the Grid5000 cluster.

9. Dissemination

9.1. Conference organisation, editors for special issues

- Pascale Vicat-Blanc is Workshop chair of the IEEE International Conference on High Performance Distributed Computing (HPDC2006) in Paris.
- Pascale Vicat-Blanc is co-chairing the International Workshop on Grid networks (GridNets2006) of the IEEE Broadnet Conference in San Jose (California- USA).
- Pascale Vicat-Blanc was Co-chair and organizer of the International Workshop on Protocols for Long Distance Networks (pfldnet2005 - feb 2005) in Lyon (ENS). She is member of the steering committee of the Pfldnet Conference.

- Pascale Vicat-Blanc, as a co-chair of the Global Grid Forum's Data Transport Research Group organized DT-RG session in Berlin (March 2004) and gives a talk to the GHPN session in Hawai (June 2004).
- Pascale Vicat-Blanc is guest editor with Jean-Phillipe Martin-Flatin of a special issue of the International Future Generation Computer Systems (FGCS) Journal on "High Performance Protocols and Grid services", April 2005.
- Pascale Vicat-Blanc is member of program committees : VECPAR2006, GRIDNETS2006, PFLDNET2006, GRID2005 workshop of the SC 05, GRIDNETS2005, IEEE CCGRID GAN2005, PFLDNET2006. She has been reviewer for international journal and conferences : Communication Network Journal, Parallel letter, JPDC, Calculateurs Parallèles, TSI.
- Laurent Lefèvre and Pascale Vicat-Blanc have co-organized the Workshop "Grid and Advanced Networks" (GAN'05) in CCGrid 2005, Cardiff, UK.
- Laurent Lefèvre was Program chair and organizer of the "IWAN2005 : Seventh Annual International Working Conference on Active and Programmable Networks", Nice, Sophia Antipolis, November 2005
- Laurent Lefèvre was Vice-Chair of topic 6 on "Grid and Cluster Computing: Models, Middleware and Architecture" on Euro-par 2005 Conference, Lisbon, Portugal, 30th August - 2nd September 2005
- Laurent Lefèvre is organizer and *program chairman* of workshops series "Distributed Shared MemOry on Clusters" DSM2005 (Cardiff) within IEEE International Symposium on Cluster Computing and the Grid (CCGrid).
- Laurent Lefèvre is *Steering Committee* member of CCGrid conference.
- Laurent Lefèvre is member of the following Program Comittee (i) International journals : Parallel and Distributed Computing Practice (PDCP), Journal of Parallel and Distributed Computing (JPDC), FGCS Advanced Grid Techology, (ii) International conferences: DFMA05, e-Science05, Grid2005, ICA3PP-2005, ICCS2005, EuroPVMPI 2005, IWAN 2005, APGAC05, ICPP-05
- C. Pham is guest editor with B. Tourancheau of a special issue of the International Future Generation Computer Systems (FGCS) Journal on "Grid Infrastructures: Practice and Perspectives", Vol. 21(2), February 2005 [18].
- C. Pham is co-editor with G. Leduc of a special issue of Annals of Telecoms on "Transport protocols for Next Generation Networks.

9.2. Graduate teaching

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

9.3. Miscelleneous teaching

- **since 2003**: O. Glück
  LAN and WAN Networks.
  Licence IUP Réseaux (Université Claude Bernard Lyon 1), lecture 30h, others 30h.
- **since 2003**: O. Glück
  Computer Networks and Applications.
  Licence IUP Réseaux (Université Claude Bernard Lyon 1), lecture 30h, others 30h.
- **2004**: O. Glück
  Computer Networks.
  Licence Informatique, (University Claude Bernard Lyon 1), lecture 30h, others 30h.
- **2004**: O. Glück
  Programming on the Web.
  Master 1 Informatique, formely Maîtrise d'Informatique, (University Claude Bernard Lyon 1), lecture 15h, others 15h.
- **since 2002** : L. Lefèvre
  Réseaux, Internet et outils associés.
  Maitrise Informatique (Université Antilles Guyane, Pointe à Pitre), 40h eq TD/an.

- **since 1998**: C. Pham
  Communication Networks.
  Master 1 Informatique, formely Maîtrise d'Informatique, (Université Claude Bernard Lyon 1), lecture: 30h/year.
- **2005**: C. Pham
  Performance Evaluation and Simulation. *Evaluation de performance et simulation*.
  Master 1 Informatique, formely Maîtrise d'Informatique (University Claude Bernard Lyon 1), lecture: 10h/year, lab studies: 20h/year.
- **since 1991**: P. Vicat-Blanc Primet
  Computer Networks.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year.
- **since 2002**: P. Vicat-Blanc Primet
  Multimedia Communications.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year
- **since 2003**: P. Vicat-Blanc Primet
  High Speed Networks and Quality of Service.
  Maitrise IUP Réseaux (Université Claude Bernard Lyon1), 20h lectures/year.

9.4. Animation of the scientific community

Pascale Vicat-Blanc

- member of the "Networks" expert committee of the CNRS.
- member of the INRIA delegation in Japan for the France-Japon workshop on Grid technology and participates to the setup of collaborations with the NAREGI project, the AIST Gtrc, the Tokyo Institute of Technology (Titech) and the Osaka University in december 2004.
- Within the Global Grid Forum, standardization entity for grid middleware, is co-chair of the Data-Transport Research Group. RESO is also active in the Network Monitoring Working Group as in the Grid High Performance Networking.
- Within the Grid5000 project, member of the steering committee.
- Within the DataGrid explorer project (ACi Masses de Donn»es), member of the steering committee.
- organized a national day on "Network and Grid Emulation" in ENS, june 2005.

9.5. Participation in boards of examiners and committees

- Olivier Glück is a member of
  - the “commissions de spécialistes 27ème section” of University Claude Bernard Lyon 1 and University Pierre et Marie Curie Paris 6.
  - the “conseil d’UFR Informatique” of University Claude Bernard Lyon 1.
- Laurent Lefèvre is member of the “commissions de spécialistes de 27ème section” of University Jean Monnet, Saint-Etienne and University Antilles Guyane, Pointe à Pitre.
- Congduc Pham
  - has been reviewer of the PhD thesis jury of R. Beuran from University of St-Etienne and University of Bucarest, July 2004.
  - has been reviewer of the PhD thesis jury of Le Khac Nhien An from INPG, March 2005.
- Pascale Vicat-Blanc
  - participated to the board of examiners for recruitments of *Chargés de Recherche CR2* of the Rhône-Alpes INRIA research unit in 2004 and 2005.
  - has been member of the board of examiners of *DEA d’Informatique Fondamentale de Lyon*.

9.6. Seminars, invited talks

- Laurent Leferve has been invited to give a talk on "Programmable networks" at the Queensland University of TEchnology, Brisbane, Australia, June 2005.
- P. Vicat-Blanc has been invited to give a talk on "GridNetworking " at the University of Osaka (JP) in july 2005.
- P. Vicat-Blanc has been invited to organise a session on Grid Networking and give a talk at the InterNetworking conference at ENST (july 2005) within the context of the IETF meeting in Paris.
- P. Vicat-Blanc has been invited to give a talk at the AIST booth during SuperComputing at Seattle in november 2005.
- P. Vicat-Blanc has been invited to give a talk on "Security in Grids" at the Security@INRIA seminar in december 2005.
- C. Pham gave a tutorial "Multicast technology: past, present, future" at IEEE DFMA 2005.

10. Bibliography

Major publications by the team in recent years

[1] B. BOUAHFS, J. GELAS, L. LEFÈVRE, M. MAIMOUR, C. PHAM, P. VICAT-BLANC PRIMET, B. TOURANCHEAU. *Designing and Evaluating An Active Grid Architecture*, in "Future Generation Computer System", To appear in February 2005, vol. 21, n° 2, 2004, http://bat710.univ-lyon1.fr/~cpham/Paper/FGCS03.pdf.

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

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

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

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

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

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

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

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

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

Books and Monographs

[11] *Proceedings of the 3rd International Workshop on Protocols for Veru Long Distance networks*, Ecole Normale Superieure de Lyon - INRIA, February 2005.

Doctoral dissertations and Habilitation theses

[12] B. GOGLIN. *Réseaux rapides et stockage distribué dans les grappes de calculateurs : propositions pour une interaction efficace*, 194 pages, Ph. D. Thesis, École normale supérieure de Lyon, 46, allée d'Italie, 69364 Lyon cedex 07, France, October 2005.

Articles in refereed journals and book chapters

[13] 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° 1, January 2005, p. 199-208.

[14] 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° 2, February 2005, p. 315-330.

[15] J.-P. MARTIN-FLATIN, P. VICAT-BLANC PRIMET. *Editorial of the special issue "High Performance Networking and Services in Grids: the DataTAG project*, in "International Journal of Future Generation Computer System, FGCS", vol. 21, n° Issue 4, April 2005, p. 439-623.

[16] J.-P. MARTIN-FLATIN, P. VICAT-BLANC PRIMET. *Special issue "High Performance Networking and Services in Grids: the DataTAG project*, in "International Journal of Future Generation Computer System, FGCS", vol. 21, n° Issue 4, April 2005, p. 439-442.

[17] C. PHAM, M. MAIMOUR. *Le contrôle de congestion dans les communications multicast*, A. BENSLIMANE (editor). , TRAITE IC2, Multicast Multimédia sur l'Internet, chap. 7, Hermes-Lavoiser, March 2005.

[18] C. PHAM, B. TOURANCHEAU. *Grid Infrastructures: practice and perspectives*, in "Future Generation Computer System", Editorial, vol. 21, n° 2, 2005, p. 247-248, http://bat710.univ-lyon1.fr/~cpham/Paper/.

[19] P. VICAT-BLANC PRIMET, F. ECHANTILLAC, M. GOUTELLE. *Experiments of the equivalent differentiated service model in grids*, in "in International Journal Future Generation Computer Systems FGCS, special issue on "High Performance Networking and Services in Grids", vol. 21, n° Issue 4, April 2005, p. 512-524.

Publications in Conferences and Workshops

[20] N. AYARI, D. BARBARON, L. LEFÈVRE, P. VICAT-BLANC PRIMET. *A Survey on High Availability Mechanisms for IP Services*, in "HAPCW2005 : High Availability and Performance Computing Workshop, Santa Fe, New Mexico, USA", October 2005.

[21] 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.

[22] 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.

[23] 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.

[24] B. GOGLIN, O. GLÜCK, P. VICAT-BLANC PRIMET, J.-C. MIGNOT. *Accès optimisés aux fichiers distants dans les grappes disposant d'un réseau rapide*, in "Actes de RenPar'16, CFSE'4, SympAAA'2005, Le Croisic, Presqu'île de Guérande, France", April 2005.

[25] B. GOGLIN, O. GLÜCK, P. VICAT-BLANC PRIMET, J.-C. MIGNOT. *Accès optimisés aux fichiers distants dans les grappes disposant d'un réseau rapide*, in "Actes de RenPar'16, CFSE'4, SympAAA'2005, Le Croisic, Presqu'île de Guérande, France", April 2005.

[26] D. KRANZLMULLER, L. LEFÈVRE. *A Record and Replay mechanism on programmable network card*, in "The IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN 2005), Innsbruck, Austria", February 2005.

[27] J. LAGANIER, P. VICAT-BLANC PRIMET. *HIPernet: fully distributed security for grid environments*, in proceedings of Grid 2005 - 6th IEEE/ACM International Workshop on Grid Computing, Seattle, Washington, USA, November 2005.

[28] 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.

[29] L. LEFÈVRE, A. SAROUKOU. *Active network support for deployment of Java-based games on mobile platforms*, in "The First International Conference on Distributed Frameworks for Multimedia Applications (DFMA'2005), Besancon, France", IEEE Computer Society, February 2005, p. 88-95.

[30] D. M. LOPEZ-PACHECO, C. PHAM. *Robust Transport Protocol for Dynamic High-Speed Networks: enhancing the XCP approach*, in "Proceedings of IEEE MICC-ICON 2005, Kuala Lumpar, Malaysia", November 2005, p. 404-409, http://www.univ-pau.fr/~cpham/Paper/icon05.pdf.

[31] L. MARCHAL, P. VICAT-BLANC PRIMET, Y. ROBERT, J. ZENG. *Optimizing Network Resource Sharing in Grids*, IEEE GLOBECOM'05, USA, November 2005.

[32] P. VICAT-BLANC PRIMET, O. GLÜCK, C. OTAL, F. ECHANTILLAC. *Emulation d'un nuage réseau de grilles de calcul:* EWAN, in "Soumis à Colloque Francophone sur l'Ingénierie des Protocoles, Bordeaux, France", avril 2005.

[33] P. VICAT-BLANC PRIMET, J. ZENG. *Traffic Isolation and Network Resource Sharing for Performance Control in Grids*, In procs of the IEEE joint Int. Conf. on Autonomic and Autonomous Systems (ICAS'05) and Int. Conf. on Networking and Services (ICNS'05), T ahiti, French Polynesia, October 2005.

[34] P. VICAT-BLANC PRIMET, J. ZENG. *An Overlay Infrastructure for Bulk Data Transfer in Grids*, in "Proceedings of 3rd International Workshop on Protocols for Fast Long-Distance Networks (PFLDnet'05), Lyon, France", February 2005.

[35] J. ZENG, P. VICAT-BLANC PRIMET. *An Overlay Infrastructure for Bulk Data Transfers in Grids*, in "in Proceedings of the Third International Workshop on Protocol For Very Long Distance Fat Networks, Pfldnet2005, Lyon", February 2005.

Internal Reports

[36] M. CHAUDIER, L. LEFÈVRE. *Description des réseaux actifs déployés*, Delivrable D2.3 Projet RNRT Temic, Technical report, May 2005.

[37] B. GOGLIN, O. GLÜCK, P. VICAT-BLANC PRIMET. *An Efficient Network API for in-Kernel Applications in Clusters*, Also available as Research Report RR-5561, INRIA Rhône-Alpes, Research Report, n° RR2005-18, LIP, ENS Lyon, Lyon, France, April 2005, http://www.inria.fr/rrrt/rr-5561.html.

[38] L. MARCHAL, Y. ROBERT, P. VICAT-BLANC PRIMET, J. ZENG. *Optimizing Network Resource Sharing in Grids*, Technical report, n° RR-5523/RR-2005-10, INRIA/ENS-LIP, March 2005, http://www.inria.fr/rrrt/rr-5523.html.

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

[40] P. VICAT-BLANC PRIMET, J. ZENG. *Traffic Isolation and Network Resource Sharing for Performance Control in Grids*, Technical report, INRIA/ENS-LIP, March 2005.

[41] M. WELTZ, E. HE, P. VICAT-BLANC PRIMET. *Survey of Protocols other than TCP*, Submitted as GFD document, Technical report, Global Grid Forum, April 2005.

Miscellaneous

[42] M. CHAUDIER, J.-P. GELAS, L. LEFÈVRE. *IAN2 : Industrial Autonomic Network Node*, Poster INRIA Booth, Supercomputing 2005, Seattle, USA, November 2005.

[43] J. LAGANIER, P. VICAT-BLANC PRIMET. *HIPernet: fully distributed security for grid environments*, submitted to GRID2005, USA, April 2005.

[44] J.-P. MARTIN-FLATIN, P. VICAT-BLANC PRIMET. *Special issue "High Performance Networking and Services in Grids: the DataTAG project*, vol. 21, n° Issue 4, Elsevier, April 2005.

[45] P. VICAT-BLANC PRIMET, J. ZENG. *Traffic Isolation and Network Resource Sharing for Performance Control in Grids*, submitted to ACNS'05, USA, April 2005.

Bibliography in notes

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

[47] S. FLOYD. *HighSpeed TCP for Large Congestion Windows*, in "Internet draft, work in progress", 2002, work in progress, http://www.icir.org/floyd/talks/floyd-tsvwg-Jul02.pdf.

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

[49] 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.

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

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

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

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

[54] T. KELLY. *Scalable TCP: Improving Performance in Highspeed Wide Area Networks*, in "Protocol for Long Distance Networks Conference", nº Pfldnet-1, February 2003.

[55] L. LEFÈVRE, J.-M. PIERSON. *Just in time Entertainment deployment on mobile platforms*, in "ICIW'06 : International Conference on Internet and Web Applications and Services, Guadeloupe, French Caribbean", February 2006.

[56] L. LEFÈVRE, P. ROE. *Improving the flexibility of Active Grids through Web Services*, in "4th Australasian Symposium on Grid Computing and e-Research, Hobart, Australia", January 2006.

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

[58] 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/.

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

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