INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Project-Team reso*

# *Optimized protocols and software for high performance networks*

*Rhône-Alpes*

THEME 1B

# Table of contents

# 1. Team

**Head of project-team**

Pascale Vicat-Blanc Primet [Maître de conférences Ecole Centrale de Lyon détachement CR1 INRIA, HDR]

**Staff member INRIA**

Laurent Lefèvre [Chargé de Recherches INRIA]

**Staff member Université Claude Bernard Lyon1 (UCB)**

CongDuc Pham [Maître de conférences, HDR]
Olivier Gluck [Maître de conférences since 1/10/03]

**Staff member CNRS**

Loic Prylli [Chargé de Recherches CNRS up to 1/11/03]

**Project technical staff**

Jean-Christophe Mignot [Permanent Engineer CNRS]
Faycal Bouhafs [Temporary Engineer INRIA - CDD - projet RNTL e-Toile]
Fabien Chanussot [Temporary Engineer INRIA CDD - projet RNTL e-Toile]
Saad El Hadri [Temporary Engineer INRIA CDD - projet RNRT VTHD++, in RESO up to 15/11/2003]
Pierre Billiau [Temporary Engineer INRIA CDD -projet DataTAG - 16/2/2003 - 17/9/2003]
François Echantillac [Temporary Engineer INRIA CDD -projet DataTAG since 1/10/2003]

**Ph. D. students**

Benjamin Gaidioz [PhD student - 2000/2003 - MENRT]
Jean-Patrick Gelas [PhD student - 2000/2003 - MENRT]
Moufida Maimour [PhD student - 2000/2003 - algerian government]
Marc Herbert [PhD student - 2001/2004 - CIFRE SUN]
Eric Lemoine [PhD student - 2001/2004 - CIFRE SUN]
Julien Laganier [PhD student - 2002/2005 - CIFRE SUN]
Brice Goglin [PhD student - 2002/2005 - BdI CNRS]
Antoine Vernois [co-Remap and IBCP) (PhD student - 2002/2005 - MENRT ACI GRID]
Mathieu Goutelle [PhD student - 2003/2006 - MENRT]

**Student intern**

Pierpaolo Giacomin [INSA Student 1/3/03 - 15/7/03]

**Long term visiting scientists**

Alessandro Bassi [Visitor from LOCI Lab, Knoxville, USA up to 1/10/2003]

# 2. Overall Objectives

RESO is focusing on communication software, services and protocols in the context of High Performance Networking and applying its results to the domain of Grids.

### *2.1.1. Project-team presentation overview*

RESO has been an INRIA pre-project proposed in 1999 between INRIA and Université Claude Bernard of Lyon.

The team joined the “Laboratoire de l’Informatique du Parallélisme” (LIP) - Unité Mixte de Recherche (UMR) CNRS-INRIA-ENS with Université Claude Bernard of Lyonin January 2003.

RESO has presented a project proposition to the Rhône-Alpes Research Unit Project Committee in March 2003. The INRIA RESO project has been officially created the 1st of December 2003.

### *2.1.2. Scientific foundations*

The RESO approach relies on the analysis of limitations encountered in existing systems and on the theoretical and experimental exploration of new approaches. This research framework between a new specific application

context and challenging 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 performances of high speed infrastructures. RESO gather expertises 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.

### 2.1.3. Goals

RESO aims at providing software solutions for high performance and flexible communications in very high speed wired networks. Current communication software and protocols designed for standard networks and traditional usages expose strong limitations when applied to this context of high performance computing. The goal of our research is optimization and control of the end to end quality of service in high performance distributed systems called computational grids.

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.

### 2.1.4. Research area

The various research areas cover high performance communication software design and optimization for end systems or cluster, enhancement of network IP layer with service differentiation or active service processing, unicast and multicast transport protocols for long distance communication, network performance measurement and monitoring, real or emulated testbed design, deployment and performance evaluation.

### 2.1.5. Application domains

RESO applies its research to the domains of high performance computing and to Grid communications. In a Grid, the network performance requirements are very high and may strongly influence the performance of the whole distributed system. As Grid applications generally rely on a complex interconnection of heterogeneous IP domains, end-to-end flow performances cannot be guaranteed or predicted. Thus, for achieving end-to-end QoS objectives, the remaining deficiencies of the network performances have generally to be masked by adaptation performed at the host level. RESO designs services and software to avoid the applications to be network-aware and to simplify the programming and to optimize the execution of their communication parts while fully exploiting the capacities of the network infrastructure.

### 2.1.6. Main contributions

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

- Protocols and optimized software for High performance PC-cluster networks;
- End to End Service differentiation in IP networks;
- High Performance transport protocols;
- High Performance Active Networks and Services;
- Network Grid services and applications.

#### *2.1.6.1. Protocols and optimized software for High performance PC-cluster networks*

- 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 system;
- 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 protoype on Myrinet networks.

*2.1.6.2. End to End Service differentiation in IP networks*

- Proposition and design of an alternative solution for end to end service differentiation in TCP/IP environment, named **equivalent differentiated services(EDS)**. This proposition aims at enhancing performance differentiation at the IP level without requiring any control plane. Different end-to-end packet marking protocols have been designed and evaluated to prove the validity of the EDS incremental and soft evolution of the IP forwarding paradigm. EDS suite has been developed as a new queuing discipline and a modified SCTP module in LINUX kernel.
- Design of a control service that dynamically manages the Diffserv classes allocated to an access point [49]. This control service attempts to optimize the utilization rate while honoring the requirement of individual flows. This service is implemented as a QoS API and an active service of the Tamanoir architecture.
- Evaluation and analysis of the real behavior of TCP with IP Premium, Assured Forwarding and Less Than Best Effort DiffServ classes, in the context of the European GEANT backbone in collaboration with the DANTE consortium and in the VTHD national network. The aim was to measure the benefit that Grid flows can expect from such advanced network services and to verify the properties of the QoS services in a production environment.

*2.1.6.3. High Performance transport protocols*

Exploration of innovative approaches based on a better knowledge and light weight control of the path to solve the problem of high and controlled throughput in very high speed links with long latency. The contributions in this area are:

- Analysis of limitation of existing or proposed high performance transport protocol's design and implementation in very high performance environments;
- Proposition of a new approach of congestion control in this context, based on back-pressure flow control;
- Analysis of on-intrusive methods of throughput measurement and proposition of an original hop by hop method for link capacity estimation and path utilization rate evaluation. A tool implementing this method, **TraceRate** has been developed.

*2.1.6.4. High Performance Active Networks and Services*

We conducted several research investigations on the topic of programmable and active networks and services for Grid support :

- Development of a high performance active network architecture (Tamanoir) and associated tools (Echidna, Pangolin). Proposition of load balancing functions in cluster-based active routers.
- Validation of Tamanoir through internal and external projects (IBP, deployment of FPTP (LAAS, Toulouse), deployment of collaborative web caches (INSA, Lyon));
- Design and development of high performance active services (DYRAM, QoSINUS);
- Deployment of active and programmable solutions around VTHD backbone (RNRT VTHD++ project) and to support Grid applications (RNTL e-Toile).

*2.1.6.5. Grid Network services and applications*

We conducted several analyses on requirements and experiments on network services in the context of large scale grid projects.

- Important contribution to the design and development of the eToile, national Grid testbed.

- Contribution to the development and improvement of high-level end-to-end performance measurement network services for grid environment. In particular we define, develop and deploy a **Probes coordination protocol PCP** that aims to coordinate the concurrent measurements of this distributed system in a grid network.
- Design and development of a framework for **Network Cost Estimation Service, NCES** and evaluation of its pertinence and accuracy in a real Grid (DataGRID) for data replica optimization.

# 3. Scientific Foundations

## 3.1. End to End Service differentiation in IP networks

**Key words:** *DiffServ, Network Quality of Service, Alternative DiffServ, Packet Scheduling Algorithm, EDS.*

**Contributed by:** Benjamin Gaidioz, Mathieu Goutelle, Pierre Billiau, François Echantillac, Pascale Vicat-Blanc Primet.

*Glossary*

**Equivalent Differentiated Services** Alternative approach of IP service differentiation that aims, by a specific router scheduling and file management mechanisms, at differentiating packets forwarding with a trade-off between loss rate and latency

*This research on Service differentiation is conducted in the context of the National RNRT VTHD++ project, the national RNTL eToile project, the European DataGRID project and the European DataTAG project*
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. Different proposals for improving the IP stack, like the DiffServ architecture, have been proposed but still exhibits three types of limitations we are considering:

- the end to end performances that the DiffServ standardized services offer 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.

## 3.2. High performance transport protocols

**Contributed by:** Marc Herbert, Mathieu Goutelle, Pascale Vicat-Blanc Primet.

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

The conservative behavior of TCP with respect to congestion in IP networks (RFC 2581) is at the heart of the current performance issues faced by the high-performance networking community. Several theoretical

and experimental analyses have shown that the dynamic 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 performances are poor and disappointing. Proposed enhancements to TCP tackle this problem in different ways, while retaining backwards compatibility. Highspeed TCP [62] and Scalable TCP [69] increase the aggressiveness in high-throughput situations while staying fair to standard TCP flows in legacy contexts. **FAST** [57] leverages the queueing information provided by round-trip time variations, in order to efficiently control buffering in routers and manage IP congestion optimally. Since two year, these propositions are actively analyzed and experimented by the international community. 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, tools for measuring the end-to-end performance of a link 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

**Contributed by:** Pascale Vicat-Blanc Primet, Geneviève Romier, AbdelHamid Joumdane, Fabien Chanussot, Mathieu Goutelle, Franck Bonnassieux, Robert Harakaly, Jean-Christophe Mignot, Loic Prylli.

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 [63]. 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 performances. 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:

- Measuring and monitoring the end to end performances 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.
- Evaluating the advantages of differentiated services, like Premium or Less than Assured Services, offered by the network infrastructure for specific grid flows is of importance.
- Creating enhanced and programmable transport protocols to optimize heterogeneous data transfers within the grid may offer a scalable and flexible approach for performance control and optimization.

# 4. Application Domains

## 4.1. Panorama

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

RESO applies its research to the domains of high performance Cluster and Grid communications. Several actions have been conducted in the context of European or National projects. These activities have been done in close collaboration with the CNRS-UREC team, other INRIA and CNRS French teams involved in the eToile project, and other European teams involved in the DataGRID and DataTAG project.

- A study of the specific requirements of grid applications has been initiated. The characteristics and performances of several grid network infrastructure have been measured and analyzed [61],[23][24],
- We have participated to the design, development and deployment of an extensible Network Monitoring system that measures, gathers and publishes relevant monitoring information in the global information system of the Grid like MDS and R-GMA in the DataGRID testbed [22].
- We have deployed and evaluated the Network Cost Estimation Service and its associated functions in the DataGRID environment with the OGSA Reptor software in charge of replica access optimization [83].
- We have actively participate to the design and deployement of a Grid testbed based on a controlled private very high speed network: eToile. The innovative Network Services, Tamanoir environment, Dynamic Network Quality of Service Management and control (QoSINUS suite) Active Reliable Multicast (DyRAM) have been deployed and are used in this testbed of the RNTL French eToile Grid. The limits of the existing communication services and protocols are analyzed and more efficient approaches that aim to carry the gigabit performance to the grid user level and take into consideration the specific needs of grid flows are explored. Deploying such an high performance Grid testbed allows also to evaluate the benefit that grid middleware and applications can get from enhanced networking technologies.
- The Madeleine, multi-protocol communication library, has been adapted and integrated both in Globus and eToile middleware.

# 5. Software

## 5.1. QoSINUS suite

**Contributed by:** Fabien Chanussot (contact), Pascale Vicat-Blanc Primet.

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

QoSinus : QoSinus is an active QoS service that interfaces the application QoS specifications (SLS) with an adaptive packet marking at DiffServ domains frontiers. QoSinus is distributed in the RNTL eToile suite. All details of QoSinus suite are available at $http://www.ens-lyon.fr/LIP/RESO/QoSINUS$

## 5.2. EDS suite

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

**Key words:** *Proportional DiffServ, RED, adapted packet marking, SCTP*.

The EDS PHB (Equivalent DiffServ) and SCTP-based packet marking protocol SCTP-lm provide alternative DiffServ mechanisms (based on PDS and RED) and transport adaptive packet marking protocols developed as Linux modules. `EDS` is distributed inside EU DataTAG project. All details of EDS suite are available at $http://www.ens-lyon.fr/LIP/RESO/software/EDS$

## 5.3. MapCenter

**Contributed by:** Franck Bonnassieux (contact), Robert Harakaly, Pascale Vicat-Blanc Primet.

**Key words:** *Network monitoring, resource visualization, grid.*

MapCenter is an open source software for Grid Resource and Services visualization. MapCenter is distributed in the IST EDG suite and is currently monitoring many Grids (IST DataGRID, RNTL eToile , IST DataTAG, Atlas Grid, Grid Ireland, GRIDIS, LCG, CrossGRID, PlanetLab, Nanyang Campus grid...) All details of MapCenter suite are available at $http://www.ens-lyon.fr/LIP/RESO/software/MapCenter$

## 5.4. NetCost Estimation Service

**Contributed by:** Robert Harakaly (contact), Franck Bonnassieux, Pascale Vicat-Blanc Primet.

**Key words:** *Optimization, network performance estimation, end to end throughput, grid.*

Network Cost Estimation Service (NCES) is providing the grid schedulers with an aggregate estimate of the network performance in terms of achievable throughput of a dedicated end-to-end path. NCES is distributed in open source in the EDG suite. NCES is used by Reptor, a Replica Optimization OGSA service. All details of NetCost suite are available at $http://www.ens-lyon.fr/LIP/RESO/software/NetCost$

## 5.5. Probe Coordination Protocol

**Contributed by:** Franck Bonnassieux, Robert Harakaly (contact), Pascale Vicat-Blanc Primet.

**Key words:** *Performance measurement, clique protocol, distributed scheduling.*

The PCP probe coordination protocol is a tool for synchronizing and coordinating the end-to-end active performance measurements in a grid testbed. PCP is currentlyreplacing the cron measurement schedulers in the DataGRID network monitoring system. PCP is distributed in open source in the EDG suite. All details of the PCP protocol are available at http://www.ens-lyon.fr/LIP/RESO/software/PCP

## 5.6. TraceRate

**Contributed by:** Mathieu Goutelle (contact), Pascale Vicat-Blanc Primet.

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

`TraceRate` is a LINUX implementation of the hop by hop path rate estimation method. This tool is split into two modules. The first one is the measurement module, which sends many times a back-to-back packet pair and gather the dispersion measurements. The second module does the distribution analysis. The measures are done for each value of TTL between source and destination in order to investigate the whole path. By default, 500 packet pairs are sent for each loop with 1400 bytes. The tool is immunized from ICMP and UDP packets limitation, firewalls filtering. This tool is an adaptation of the well-known `traceroute` which sends TCP packets instead of ICMP packets.

## 5.7. Tamanoir

**Contributed by:** Jean-Patrick Gelas, Laurent Lefèvre (contact), Saad El Hadri.

**Key words:** *active and programmable networks, execution environment.*

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 in the RNTL eToile suite. It is urrently used by partners in RNTL eToile Project and in RNRT VTHD++ project. Tamanoir is also used by research teams for development of new network services : LAAS (Toulouse) for deployment of FPTP protocol, LISI (INSA, Lyon) for the design of active web caches and Univ. Vannes for deployment of internal monitoring systems. All details on Tamanoir are available at $http://www.ens-lyon.fr/LIP/RESO/Tamanoir$

## 5.8. Echidna

**Contributed by:** Saad El-Hadri, Laurent Lefèvre (contact).

**Key words:** *traffic generator, programmable networks* .

Echidna is a fully distributed active traffic generator. It allows the deployment of large scale active network tests. It can be adapted to any kind of execution environment. Opensource software, August 2003.

## 5.9. Pangolin

**Contributed by:** Saad El-Hadri, Laurent Lefèvre (contact).

**Key words:** *programmable network, Grid.*

Adaptation of visualization environment `MapCenter` from European DataGRID project. Pangolin provides management and visualization of large scale active network infrastructure on wide area networks. Based on active services deployment, it allows to manage active nodes, service repositories and active services.

## 5.10. ORFA (Optimized Remote File-system Access)

**Contributed by:** Brice Goglin (contact), Loic Prylli.

**Key words:** *SAN networks, filesystem.*

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

## 5.11. KNET

**Contributed by:** Éric Lemoine, Laurent Lefèvre, CongDuc Pham.

**Key words:** *Networking, sub-system, driver, embedded code, Linux, GM.*

The `KNET` parallel networking sub-system has been developed for `Linux` (as a separate module) and the `GM-1.5`'s driver and firmware (embedded code) have been adapted to `KNET`.

## 5.12. DyRAM

**Contributed by:** Faycal Bouhafs, Moufida Maimour, CongDuc Pham (contact).

**Key words:** *Reliable Multicast, programmable networks.*

`DyRAM` is a reliable multicast framework using lightweight services in routers to improve performances of multi-point communications. The implementation of `DyRAM` is being done within the RNTL e-Toile project. `DyRAM` consists of a library and of an API for developing applications with multicast support. For the moment, the main application target is file transfers, therefore an `ftp`-like program is also developed for the partners within the project.

## 5.13. sucvP

**Contributed by:** Julien Laganier (contact).

**Key words:** *verifiable identifiers, verifiable addresses, end-in-end, IPsec, decentralized security.*

The protocol sucvP has been implemented on FreeBSD, including the interfaces with the IPsec subsystem embedded within this operating system. The implementation uses the cryptographic functions of the OpenSSL library. INRIA holds intellectual property associated with the first version of this software, enforced by Agence Francaise de Protection des Programmes (APP). This software allows any IPv6 node which uses a cryptographic identifier as its IP address to prove to its interlocutors that it indeed " owns " its address. One can thus derive from this proof of ownership of IPv6 address a confidence allowing to secure the traffic exchanged by such nodes thanks to the use of IPsec in transport mode (Transport Mode Opportunistic IPsec).

This software provides the foundation of an architecture of security built on top of a cryptographic identifiers "infrastructure". With the aim of showing the applicability of this infrastructure at the network level (e.g., IP), this software was also adapted to provide the Tunnel Mode Opportunistic IPsec service. This software has also been adapted to support Host Identity Protocol, a protocol currently discussed at the IETF which provides an equivalent service.

# 6. New Results

## 6.1. Protocols and software for high performance PC-clusters networks

### *6.1.1. Optimized Remote File-system Access*

**Contributed by:** Brice Goglin, Loic Prylli.

**Key words:** .

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) [46] was developed on Myrinet networks to provide an efficient access to remote data. The user-level implementation showed that file transfers may saturate the physical link [47]. The need to cache metadata on the client's side leads to the idea of porting 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. We are currently preparing a collaboration with Myricom to work on a new interface that will fit both filesystem and usual communication that MPI applications use.

### *6.1.2. Designing and evaluating the KNET system*

**Contributed by:** Éric Lemoine, Laurent Lefèvre, CongDuc Pham.

**Key words:** *Parallel networking sub-system, data locality, SMP machine, performance, robustness.*

We propose a new networking subsystem architecture built around a packet classifier executing in the Network Interface Controller (NIC). By classifying packets in the NIC, we believe that significant performance, scalability, and robustness gains can be achieved on shared-memory multiprocessor Internet servers. To show the feasibility and the benefits of the approach, we developed the KNET software prototype (consisting in extensions to the Linux kernel and modifications to the Myrinet NIC firmware and driver) and ran a series of experiments.

KNET's objectives are to parallelize packet processing in the operating system while maximizing data locality in the processor caches and eliminating the Receive Livelock effect that can severely affect the operating system's robustness. KNET uses per-processor network threads to achieve parallelism, packet demultiplexing in the Network Interface Controller to maximize connection data locality and ensure robustness[35]. KNET exhibits up to 35% improvement in throughput on a 4-way machine.

## 6.2. End to end service differentiation in IP networks

### *6.2.1. Equivalent Differentiated Services architecture*

**Contributed by:** Benjamin Gaidioz, Pierre Billiau, François Echantillac, Pascale Vicat-Blanc Primet.

**Key words:** *Network Quality of Service, DiffServ, PHB, RED, packet scheduling algorithm.*

In the light of the frustrating experience of deployment of existing IP QoS approaches, IntServ [56] and DiffServ [55], we have proposed a new differentiated service scheme called EDS : "Equivalent Differentiated Services".

*6.2.1.1. EDS design*

This proposition represents a radical departure from traditional "DiffServ" architectures which rely on bounded domain concept and pricing models. The EDS is merging the Alternative Best Effort ideas [67] and the Proportional DiffServ Principles [59]. The EDS scheme aims at providing a spectrum of "different but equivalent" network services that offer a trade-off between delay and loss rate to the end-to-end flows. EDS acts as a network layer protocol analogous to IP and, as TCP does, the end-to-end transport layer has to do some adaptation. As EDS offers a service differentiation based on packet marking, the corresponding transport layer has to adapt data transmission and packet marking accordingly. Considering that the Internet traffic is composed of real-time traffic, interactive traffic, WEB traffic and bulk file transfer traffic, different types of adaptive packet marking algorithms, integrated in a transport protocol stub can benefit differently from the network differentiate behaviors [12], [21]. The implementation in LINUX has been realized [44]. This software comprises different LINUX modules: a novel router mechanism merging an original RED-based active queue management algorithm and a proportional scheduling algorithm and a transport protocol for bulk data transfer that integrates an adaptive packet marking algorithm in the SCTP AIMD algorithm. This software has been functionally validated and is under performance evaluation within the European DataTAG project. The aim is to prove the EDS concept and to show that it can improve the transfer of a mix heterogeneous flows on long distance and heterogeneously provisioned links.

*6.2.1.2. Equivalent Differentiated Services transport layer*

On a plain best-effort network, there is of course no way to control neither the end-to-end delay nor the loss rate. Packets are forwarded in a delay and with a loss probability depending on the network load. The EDS system has been designed to reflect the IP design philosophy to the plane of performance differentiation. The same way TCP has been designed to provide reliability on top of the unreliable network layer IP, we have designed three transport protocols which provide specific soft quality of service properties to applications. The RT-TP over EDS protocol ensures *as best as possible* an end-to-end delay and a relative reliability to a real-time application [43]. The SM-TP over EDS protocol ensures *as best as possible* an end-to-end delay bound to reliable short message transport [45]. The LM-TP over EDS protocol ensures *as best as possible* an improved end-to-end delay to bulk data transport. The NS simulation and real tests in emulated testbed demonstrate that this architecture improves flow specific performance criteria in the context of a realistic mix of heterogeneous traffics.

*6.2.1.3. LM-TP over EDS implementation and evaluation*

All the adaptive packet marking algorithms have been implemented in the SCTP NS module. The LM-TP protocol is implemented also in an SCTP module in Linux. We choose SCTP for its modular implementations both in LINUX and NS comparing to the TCP's one. The implementation as a module in Linux facilitates the test. Extensive tests with the Nistnet emulator have shown that LM-TP over EDS is resistant to non-friendly UDP flows in short or long paths. LM-TP over EDS offers a smoother and slightly better throughput than SCTP over IP. Throughput obtaines with lkSCTP, SCTP-lm and TCP has been compared in different conditions (load, delay) on different router configuration (IP default, RED, EDS, EDSRED). When there is no delay the performance of SCTP-lm on EDSRED is comparable to that of TCP. And moreover all the protocols get the best performance on EDSRED. But when the delay is set to a great value using nistnet ( RTT=200ms), the protocols get the same performance on all routers. The behavior of this protocol over EDS will be evaluated on the DataTAG link.

### *6.2.2. Performance measurement of TCP over DiffServ in production networks*

Performance measurement of TCP over DiffServ in production networks

**Contributed by:** Mathieu Goutelle, Franck Bonnassieux, Fabien Chanussot, Pascale Vicat-Blanc Primet.

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

End to end performance and specifically TCP behavior in real Diffserv environment have not been well experimented in large scale. To verify the end to end properties offered by Diffserv in real production networks, we have conducted a set of tests in the GEANT European backbone and in the VTHD experimental network.

*6.2.2.1. Performance measurement of TCP over DiffServ in GEANT*

The aim of these experiments was to test the behavior of TCP flow in different classes of service offered by the GEANT European backbone [41]. TCP flows marked in IP Premium, BE and LBE were analysed when LBE class, BE class and even IP Premium class are congested. For these tests, we have two PCs in NL and UK GEANT POPs connected at 1Gb/s to the core routers. Two possible paths between the PCs where created by static routing. The network bottleneck is the same for both paths (between France and Netherland via Belgium) and is 2.5Gb/s. Other links are 10Gb/s. We were also able to generate background traffic with the SmartBits in UK and DE in order to artificially congest the bottleneck FR-BE-NL. The SmartBits can generate up to 2.5Gb/s of raw traffic (not TCP). The results we obtained in GEANT show that the DiffServ implementation respects the IETF specifications [42]: IP Premium is very well protected in throughput against BE / LBE, BE is very well protected against LBE, LBE keeps the 5% of bandwidth under BE congestion. These results confirm that applications that require stable throughput can reserve IP premium resources and that unfriendly or intrusive applications that are not rate sensitive can use all the available resources without disturbing classical traffic when using LBE class.

*6.2.2.2. Performance measurement of TCP over DiffServ in VTHD*

The results we obtained in VTHD allow the evaluation of the performance of the specific DiffServ implementation of VTHD. IP Premium offers good performance stability to TCP while Assured Forwarding, enhance performance but is not able to guarantee bounds. Several performance analysis with a distributed medical images processing software have been performed [80]

### *6.2.3. Dynamic DiffServ class management and end to end QoS control*

**Contributed by:** Fabien Chanussot, Pascale Vicat-Blanc Primet.

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

We propose a service that dynamically adaptes packet marking to best fit the requirements of individual grid flows and simultaneously to best allocate the shared differentiates resources [49].

*6.2.3.1. Adaptive packet marking strategies on classical DiffServ*

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

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

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

The second issue is addressed by a service architecture that combines flow aware and infrastructure aware components to map and dynamically adapt the QoS specification of the flows to the QoS facilities offered by

the network. IP premium is a finite and scare resource. To avoid to waste this resource, we propose algorithms that statically or dynamically adapt the packet marking according the real QoS of a TCP flow. The analysis of ACK permits to calculate periodically the amount of data transfered and to increase packet priority when required in order to meet some deadline requirement. The ultimate goal is to provide an Earliest Deadline First algorithm in an edge packet marking equipment, in order to serve the performance requirements of individual TCP flows. This algorithm has been implemented in an active service under the TAMANOIR environment.

## 6.3. High performance transport protocols

### *6.3.1. High performance transport*

**Contributed by:** Marc Herbert, Pascale Vicat-Blanc Primet.

**Key words:** *High speed transport, congestion control, flow control*.

The new congestion control solution we propose for high speed network is based on back-pressure flow control. Our Network of Queues proposal [66] suggests an outright departure from current TCP standards for some particular networks. The idea is to replace the current end-host-based TCP/IP congestion management by a network of flow-controlled links, according to a scheme known as “back-pressure”. The challenge is to control the flow queue by queue with the functionalities already present in the intermediate networking equipment. The idea is to activate the 802.3x flow control in very high speed Ethernet links to reduce the feedback loop and to efficiently prevent the congestion. It has been proved theoretically that the back-pressure approach is better than classical end system feedback control approach. The issue to solve is to prove its validity in the actual equipments and in operational IP networks and to solve cross layer issues. Our proposal argues that, in some specific networking contexts like those of grids, using back-pressure as an addition to existing TCP/IP/Ethernet networking hardware and software may offer a valuable tradeoff between performance gain and migration cost. In order to develop insights on how the current network hardware and software behave relative to flow control, we forced a 100 Mb/s bottleneck in local gigabit testbed. The result of the first, basic experiment is a sawtooth-shaped throughput curve. When several hosts compete for the same bottleneck, the cooperative **AIMD** algorithm of TCP gives an approximately fair share of the capacity to each flow. The first promising conclusions are that this approach is feasible in a IP/802.3x environment and offers a smoother reaction to congestion compared to TCP or HS-TCP and a rapid convergence to fairness.

### *6.3.2. End to end throughput measurement*

**Contributed by:** Mathieu Goutelle, Pascale Vicat-Blanc Primet.

**Key words:** *Performance measurement, Packet Pair methods, hop by hop capacity discovery*.

To discover the characteristics of a path in terms of hop by hop capacity and utilization rate, we have proposed a new method and a tool. Our approach, using a hop-by-hop packet pair method and a fine analysis of the measurements, provides such information. The method consists in using the dispersion of a packet pair because it has many advantages compared with the *Variable Packet Size* method [79]. Cross-traffic taints the dispersion measurements with noise, which forces to elaborate complex analysis methodologies [60]. There are two kinds of errors : the first one is typically due to cross-traffic when packets are inserted between the two probes and hence the capacity is underestimated. The method we propose lies on an incremental discovery of the path characteristics. For this, we evaluate three parameters for each measurements distribution. **The maximal mode** is the easiest to determine. It corresponds to the interval with the maximum numbers of samples. **The previous mode** is the mode of the current distribution which has the same capacity value as the one estimated for the previous hop. **The new mode** is the mode with a capacity value strictly lower – a new mode implies that the capacity decreases – than the previous mode and which includes a sufficient number of samples (here 1% of the total number of measurements). For the first hop, the previous and new mode are the same. We evaluate the **noise area** too. This is defined as the little capacity values area which contains three or more side-by-side modes, *i.e.* not separated by an interval of at least a distribution step. This proposition has been validated in simulation, then implemented in Linux and validated experimentally. We have compared our method with

others to define its limits and the potential utilizations on the developed tool. We have shown that this method is relatively non-intrusive, robust, relatively accurate and reliable and keep these qualities under bad network conditions (high load, long path, *etc.*) [32][48], [65]. We have shown that our tool works up to 1 Gbit/s [48]. We have validated the Linux implementation and have demonstrated that it provides usable results in real life, without the participation of the receiving computer and path routers. Results show that TraceRate can also estimate path utilization rate. We are actually studying a new data analysis method that can rapidly extract such aggregate information. We will explore how such precise information can be used by a transport protocol to better control transmission rate and end to end transfer delay. The fact that our tool can give in a single and non-intrusive measure the capacity and the available bandwidth is very promising.

## 6.4. High performance active networks and services

### *6.4.1. Gigabit Active Network Execution Environment*

**Contributed by:** Jean-Patrick Gelas, Pierpaolo Giacomin, Saad El Hadri, Laurent Lefèvre.

**Key words:** *execution environments, programmable and active networks.*

We have proposed a new execution environment called `Tamanoir`, which focuses on performance problems of active and programmable network equipments and dynamic deployment of services[13]. Targeted equipments are deployed in access networks around high performance (Gbit/s) backbones. These networks must face heterogeneity problems in terms of equipments and bandwidth.

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

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

The high performance Tamanoir architecture has been implemented on a cluster-based infrastructure and supports active services inside the Linux kernel and on distributed resources [31][17][18].

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

Our environment has been used and deployed inside various applications context like :

- Active Web [33] to efficiently support on-the-fly protocol change for web sessions;
- Collaborative web caches [34] to deploy intelligent and lightweight caches inside the network;
- Active Logistical networks [25] to provide efficient storage functionalities inside the network for multimedia streams.

### 6.4.2. Active logistical networks

**Contributed by:** Alessandro Bassi, Jean-Patrick Gelas, Laurent Lefèvre.

**Key words:** *storage, active networks.*

Logistical networks provide efficient distributed storage solutions inside networks. The *Internet Backplane Protocol* (IBP) developed by LoCI laboratory (Univ. Tennessee Knoxville, USA) allows the sharing of storage resources through wide area networks. `IBP` is based on data blocs (disk, memory...) and proposes a complete data depot solution. `IBP` depots are distributed between sites and can be accessed remotely by data streams to deploy a global storage service

We have studied the integration and merging of logistical networks inside our active network solutions in order to allow active services to efficiently store data on the fly [25]. From this new proposed architecture, we have merged and developed an active logistical equipment based on Tamanoir execution environment and IBP.

### 6.4.3. Active network support for collaborative web caches

**Contributed by:** Laurent Lefèvre.

**Key words:** *web caches, collaboration, active networks.*

During the DEA internship of Sidali Guebli jointly supervised between J.M. Pierson (LIRIS, INSA Lyon) and L. Lefèvre, we studied the support of active networks to lightweight communications infrastructure for collaborative web caches. Some of the difficulty lies in the limited resources we want to deploy on the active nodes (in terms of CPU, memory and disk). But, we clearly benefit from active networks support by transparently deploying active caches through data path without modifying and re-configuring Web clients and servers. Collaborative web caches services have been developed in Java inside Tamanoir EE. These active cache services can be dynamically modified and communicates in point to point way through control communication channel between active nodes. Active dedicated services have been developed and deployed and validated on local experimental platform [34].

### 6.4.4. Load balancing in cluster-based active network equipments

Load balancing in cluster-based active network equipments
**Contributed by:** Pierpaolo Giacomin, Laurent Lefèvre.

**Key words:** *software router, cluster, load balancing.*

As programmable network equipments allow deployment of heterogeneous services, we propose new solutions to efficiently balance equipments based on clusters. We propose new load balancing policies added to the Linux Virtual Server Project (LVS).

### 6.4.5. New active services for reliable multicast communication

**Contributed by:** Moufida Maimour, CongDuc Pham.

**Key words:** *Reliable multicast, programmable networks, active services.*

Active services for reliable multicast proposed so far in the research community consisted in the cache of data packets and the feedback aggregation. Caching data is very costly to be implemented in routers therefore we investigated 4 new services to improve the performances of local recovery and heterogeneity support. These are:

- the dynamic replier election which consists of choosing a link/host as a replier one (the one which will send again the missing packet) to perform local recoveries;
- the early detection of packet losses and the emission of the corresponding NACKs;
- an accurate, on a per-hop basis, RTT (*Round-Trip-Time*) computation for congestion and rate adaptation purposes for interoperability with unicast TCP flows;
- the partitioning of the receivers to handle heterogeneity.

For the first two services, we have conducted analytical studies similar to those realized in 2001 [73] to model and evaluate their performances. The results have been published in [76] and [74],[14][15]. Regarding the two last services consisting in the RTTs aggregation and the receiver partitioning, preliminary results published in [37], [36] and [38] are very encouraging.

All the proposed services are lightweight active services that consume very few router's resources. However, combined with local recoveries, they are very beneficial to reduce the end-to-end latency and to provide the support of heterogeneity in a multicast session.

### *6.4.6. Congestion control in DyRAM*

**Contributed by:** Moufida Maimour, CongDuc Pham.

**Key words:** *Reliable multicast, programmable networks, congestion control.*

Congestion control in multicast is a difficult task because it is hard to get and take into account the status of the entire group of receivers and to satisfy all the receivers when they are heterogeneous (which is almost always the case).

The active service that estimates the RTTs from the receivers towards the source is an important component in the congestion control mechanism that we proposed. Active routers in the multicast tree estimate the RTT towards their parent node (another active router or the source) and aggregate these informations in order to propagate only one value towards the source. As in `RMANP` [54] or `NCA` [68], we benefit from the physical multicast tree to aggregate the RTT values, as opposed to `TRAM` [58] or `MTCP` [81] which use, and thus maintain, a logical tree.

The `AMCA` algorithm (*Active-based Multicast Congestion Avoidance Algorithm*) [37] that we proposed use this lightweight service to predict (and in most cases avoid) congestions by observing the RTT variation. The approach is similar in concepts to TCP Vegas but do not suffer from the path re-routing problem. AMCA is compatible with TCP.

### *6.4.7. The DyRAM active reliable multicast protocol*

**Contributed by:** Moufida Maimour, Congduc Pham.

**Key words:** *Reliable multicast, programmable networks, protocol.*

We have integrated in a protocol called `DyRAM` (*Dynamic Replier Active reliable Multicast*), the active services that we proposed (along with feedback aggregation and the subcast feature). The main objective of DyRAM is to avoid cache in routers and to provide low recovery latencies. DyRAM is therefore very different from `ARM` [71], `AER` [68] or `MAF` [82]. `DyRAM` and its performance are described in [77] and [19].

## 6.5. Grid Network services and applications

### *6.5.1. Network Cost Estimation Service for Grids*

**Contributed by:** Franck Bonnassieux, Mathieu Goutelle, Robert Harakaly, Pascale Vicat-Blanc Primet.

**Key words:** *network service, cost estimation, replica optimization.*

In a Data Grid, replicas are located at several different storage locations with a large range of possible current network throughput and latency. It is important to select replicas based on their minimal access latency. Although replica access optimization does not only depend on the network link and its capacity (the load and the latency of data servers have to be taken into account too), we have examined how an aggregate knowledge of the network behavior may have an important impact on the replica access optimization step.

In the last few years, dedicated Grid network monitoring systems have been developed and deployed within Grid environments [85], [84]. To provide the users with an abstract and homogeneous view of the complex set of interconnected resources, we designed and developed a performance measurement system that is characterized by simple and relevant metrics of a grid network cloud [28], [27]. However, to optimize the application performance, a Grid middleware component requires aggregate and simple estimations of *transfer costs* between defined end hosts. We examine how high level *Network Cost Estimation Functions* (NCEF) can

be computed and used in a Grid environment for network-based replica optimization [83]. In particular, we study how an estimation of the end to end transfer delay of a certain amount of data can be easily derived from the raw network performances measurements. A flexible and open Grid *Network Cost Estimation Service* (NCES) that permits Grid resource management services to use network monitoring data in a very simple fashion has been developed. We demonstrate that provided approximations are valuable in certain cases, are necessary to find a tradeoff between accuracy, efficiency and scalability, and to define an extensible set of *functions* within the framework of an open service.

### 6.5.2. Active Grid

**Contributed by:** Laurent Lefèvre, Jean-Patrick Gelas.

**Key words:** *Grid, active networks.*

We have studied the benefits of programmable and active networks as an alternative solution for dynamic deployment of networks services adapted to Grid infrastructure. This proposition is called “Active Grid”. Preliminary works based on Tamanoir proposed collaborative usage of high performance execution environment with Grid middleware and applications [70][64] [29][30].

### 6.5.3. Distributed Security for applications and the grid

**Contributed by:** Julien Laganier, Alessandro Bassi.

**Key words:** *cryptographic Identifiers, distributed security, decentralized security, applications, grid.*

A new framework that will eventually allow to seamlessly secure any distributed application was described and partially implemented in [26] and [16]. This framework rests on the use of CBIDs by each entity combined with the use of SPKI authorization certificates, thus allowing a given CBID to delegate rights to another CBID. Amongst others things, this allows to secure the distributed and shared remote storage protocol Internet Backplane Protocol.

### 6.5.4. Security and Cryptographic Identifiers in the network layer

**Contributed by:** Julien Laganier, Laurent Lefèvre.

**Key words:** *Identifiers, distributed security, decentralized security, identification, localization.*

The implementation of an infrastructure of cryptographic identifiers in the network layer shows well the fundamental utility of such an infrastructure dice the network layer because it allows two nodes previously unknown from each other to communicate in a protected way on level IP This implies two other nodes acting as footbridges of IPsec safety, which will have to be discovered mutually and to exchange certificates of delegation proving that they are authorized by the two extremities of the data flow to act of the kind. Given the experimental and applied nature of this work, it could only profit from a use in real conditions, showing in this way its applicability and its practicality, it is judicious to standardize the stable components “as far as possible” of the aforesaid work, so that their use can spread in a way increased in the communities likely to use them. To this end, a work of standardization was started within the IETF around techniques of CGA/CBID, as described in [50], [52] and [51].

### 6.5.5. Madeleine

**Contributed by:** Jean-Christophe Mignot, Loic Prylli.

**Key words:** .

As part of the RNTL e-Toile project, the communication library, Madeleine [53], designed by the R. Namyst team at Bordeaux has been ported over Globus. This gives the opportunity to use the native Madeleine API as well as the MPI API. The port gives to the applications an API whose model is fully connected without having to establish the connections between the nodes nor having to know the underlying network (TCP, cluster, grid). Madeleine gives the possibility to use different bufferisation modes for transferring the data. More precisely as part of the RNTL e-Toile project, Madeleine transparency has been generalized to a multi-site deployment with the e-Toile authentification and security framework.

#### *6.5.6. Multicast in an active grid infrastructure*

**Contributed by:** Faycal Bouhafs, Moufida Maimour, Congduc Pham.

**Key words:** *Reliable multicast, programmable networks, computational grids.*

With a logical view closer to a distributed operating system than a pure communication infrastructure, one might consider extending for computational grids the basic functionalities found in the *commodity Internet*'s network infrastructure. Our work on multicast support for the grid is based on the possibility to easily (at least more easily than on the Internet) add processing elements (active routers) in the network infrastructure of a grid. These ideas and motivations are described in [78], [75] and [20] for multicast communications.

A first prototype of `DyRAM` has been developed by a master student, J. Mazuy, in 2002. The performance results have been published in [72]. This prototype has been improved and extended in the VTHD++ and E-Toile projects. We are implementing an `ftp`-like tool for data and code transfers on a computational grid. This work has lead to seminars and a demonstration at IPDPS 2003 (ACI GRID booth) and within the VTHD++ and E-Toile projects.

# 7. Contracts and Grants with Industry

## 7.1. SUN Labs, Europe

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

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

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.

## 7.2. Myricom

**Contributed by:** Brice Goglin, Loïc Prylli.

This long-term collaboration between our team and US based Myricom company is focused on the software Myrinet suite (GM) and works within the development of the ORFA (Optimized Remote File-system Access) software protoype.

## 7.3. EDF

**Contributed by:** Laurent Lefèvre.

RESO is involved in a GRECO project with EDF and IRISA (2000-2003). L. Lefèvre participates in the technical support of the PhD of G. Vallee (PARIS, IRISA). Supported by EDF.

## 7.4. 3DDL

**Contributed by:** Laurent Lefèvre.

**Key words:** *programmable networks, java* .

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

# 8. Other Grants and Activities

## 8.1. Regional actions

### *8.1.1. Region project*

**Contributed by:** Laurent Lefèvre, Cong-Duc Pham.

RESO is 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 Rhône-Alpes region (2001-2003).

## 8.2. National actions

### *8.2.1. ACI-Grid Jeune Equipe*

**Contributed by:** Laurent Lefèvre, Cong-Duc Pham, Pascale Primet.

RESO is investigating advanced research on network services for grid computing within an ACI GRID "Young Team" project (2002-2003).

### *8.2.2. RNTL eToile*

**Contributed by:** Faycal Bouhafs, Fabien Chanussot, Saad El-Hadri, Benjamin Gaidioz, Jean-Patrick Gelas, Laurent Lefèvre, Moufida Maimour, Congduc Pham, Geneviève Romier, Pascale Vicat-Blanc Primet.

The eToile project [39][23] is an experimental wide area grid testbed. The e-toile [1] has three complementary objectives:

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

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

### *8.2.3. RNRT VTHD++*

**Contributed by:** Faycal Bouhafs, Fabien Chanussot, Saad El-Hadri, Benjamin Gaidioz, Jean-Patrick Gelas, Laurent Lefèvre, Congduc Pham, Pascale Vicat-Blanc Primet.

(2002-2004) : RESO is responsible for Work Package 4 on "High performance active networks around VTHD backbone". Supported by RNRT, funding : 1 Engineer for 3 years.

### *8.2.4. ACI Grid GRIPPS*

**Contributed by:** 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 [40].

[1] e-toile is a RNTL project (réseau national de recherche en logiciel) funded by French Ministry of Research

### *8.2.5. ACI Grandes Masses de Données GridExplorer*

**Contributed by:** Olivier Gluck, Brice Goglin, Mathieu Goutelle, Marc Herbert, Julien Laganier, Laurent Lefèvre, Éric Lemoine, Congduc Pham, 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.

### *8.2.6. GRID5000*

**Contributed by:** Olivier Gluck, Brice Goglin, Mathieu Goutelle, Marc Herbert, Julien Laganier, Laurent Lefèvre, Éric Lemoine, Congduc Pham, Pascale Vicat-Blanc Primet.

(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*.

## 8.3. European actions

### *8.3.1. European DataGrid project*

**Contributed by:** Pascale Vicat-Blanc, Franck Bonnassieux, Robert Harakaly, Mathieu Goutelle.

**European DataGrid project** (2001-2003) : Research and Technological Development for an International Data Grid - contract IST-2000-25182 (CERN/CNRS/INFN/NIKHEF/PPARC/ESRIN). INRIA has been a sub-contractor of CNRS. This year we conducted experiments in the GEANT backbone for DiffServ evaluation. We finalized and deployed the NetCost services and the PCP protocol. Several experiments were conducted with partners of the University of Manchester (R. Hughes Jones team), the CERN and other work packages for validating these tools.

### *8.3.2. European DATATAG project*

**Contributed by:** Pascale Vicat-Blanc, Pierre Billiau, François Echantillac, Mathieu Goutelle, Marc Herbert.

IST-2001-32459 (CERN/INRIA/UvA/PPARC) Research and Technological Development for an International Grid Interconnection (2002-2003). RESO studies protocols of high performance data transport and quality of service provided by EDS on a long distance high performance backbone. Funding: 124K euros (18 month).

### *8.3.3. Programmes d'Actions Intégrées Amadeus with Linz Univ., Austria*

**Contributed by:** Laurent Lefèvre.

RESO is involved in a long term collaboration (1999-2000, 2001-2003) 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.

## 8.4. International actions

### *8.4.1. NSF-INRIA with Aerospace Organization*

**Contributed by:** Laurent Lefèvre.

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

## 8.5. Visitors

### *8.5.1. Collaboration with LOCI Lab., Tennessee, USA*

**Contributed by:** Alessandro Bassi, Jean-Patrick Gelas, Laurent Lefèvre.

We have a long term collaboration with LOCI lab (University of Tennessee, Knoxville, USA) on interactions between programmable and logistical networks. RESO has hosted A. Bassi as an invited researcher from 1/11/2001 to 1/11/2003. J.P. Gelas (PhD student in RESO) is going to spend one year in LOCI in 2004 as a postdocoral researcher.

# 9. Dissemination

## 9.1. Conference organisation, editors for special issues

- Pascale Vicat-Blanc, as co-chair of the Global Grid Forum's Data Transport Research Group organized GGF6 DT-RG session in Tokyo (March 2003) and GGF7 DT-RG session in Seattle (June 2003).
- Pascale Vicat-Blanc is guest editor with Jean-Phillipe Martin-Flatin and Cees de Laat of a special issue of the international Future Generation Computer Systems (FGCS) Journal on "High Performance Protocols and Grid services". (to appear in summer 2004)
- Pascale Vicat-Blanc is member of program committees of CCGRID GAN2003, CCGRID GAN2004, Grid workshop in Supercomputing 2003, Pfldnet04. She has been reviewer for international journal and conferences : Communication Network Journal, Parallel letter, JPDC, Calculateurs Parallèles, TSI, IPDPS03, ICC04, Pfldnet04, CFIP03, JDIR03, INFOCOM2003.
- C. Pham is co-editor with B. Tourancheau of a special issue of FGCS on "Grid Infrastructures: Practice and Perspectives".
- Laurent Lefèvre is organizer and *program chairman* of workshops series "Distributed Shared MemOry on Clusters" DSM2003 (Tokyo) within IEEE International Symposium on Cluster Computing and the Grid (CCGrid).
- Laurent Lefèvre has been *Local chair* of Topic 9 on "Distributed algorithms" in Europar2003 conference, Klagenfurt, Austria, August 2003.
- Laurent Lefèvre, Pascale Vicat-Blanc and Craig Lee (AeroSpace Org.) have co-organized the Workshop "Grid and Advanced Networks" (GAN'03) in CCGrid 2003, Tokyo.
- Laurent Lefèvre is *Steering Committee* member of CCGrid conference.
- Laurent Lefèvre is member of following Program Comittee e ($i$) International journals : Parallel and Distributed Computing Practice (PDCP), Journal of Parallel and Distributed Computing (JPDC) 2003, FGCS Advanced Grid Techology 2003, ($ii$) French journals: Calculateurs Parallèles, TSI, ($iii$) International conferences: AGridM2003, Grid 2003, Europar 2003, AMS 2003, EuroPVMPI 2003, IWAN 2003, IEEE CCGrid 2003.

## 9.2. Graduate teaching

- **2003**: C. Pham
  High-speed Network and New Generation Internet.*Réseaux Haut-Débit et Internet Nouvelle Génération*.
  DEA DIF (University Claude Bernard Lyon 1, ENS-Lyon), lecture: 24h.
- **2002 & 2003**: C. Pham
  New Technologies for the Internet. *Les nouvelles technologies de l'Internet*.
  DEA DISIC (University Claude Bernard Lyon 1, INSA), lecture: 14h/year.

- **since 1998**: C. Pham
  Performance Evaluation and Simulation. *Evaluation de performances et simulation*.
  DESS IIR Réseaux (University Claude Bernard Lyon 1), lecture: 10h/year, experimental work: 40h/year.
- **since 1998**: C. Pham
  Wide Area Networks. *Réseaux grandes distances*.
  DESS CCI (University Claude Bernard Lyon 1), lecture: 10h/year.
- **2001 & 2002** : L. Lefèvre
  *Réseaux hautes performances*.
  DEA DIF (Université Claude Bernard Lyon 1, ENS-Lyon), lecture 24h.
- **depuis 2002**: L. Lefèvre
  *Réseaux hautes performances*.
  DESS IIR Réseaux (Université Claude Bernard Lyon 1, ENS-Lyon), lecture 10h.
- **2003**: O. Gluck
  *Internet et Outils Associés*.
  DESS IIR Réseaux (Université Claude Bernard Lyon 1, ENS-Lyon), lecture 10h.

## 9.3. Miscelleneous teaching

- **2002-2003**: P. Vicat-Blanc Primet
  *Computer Networks*.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year.
- **2002-2003**: P. Vicat-Blanc Primet
  *Multimedia Communications*.
  Engineer school (Ecole Centrale de Lyon), 20h lectures/year
- **2003**: P. Vicat-Blanc Primet
  *High Speed Networks and Quality of Service*.
  Maitrise IUP Réseaux (Université Claude Bernard Lyon1), 20h lectures/year.
- **since 2000**: C. Pham
  Communication Networks.
  *Les réseaux de communication*.
  MIM 2nd year (ENS-Lyon&University Claude Bernard Lyon 1), lecture: 30h/year.
- **since 1998**: C. Pham
  Communication Networks.
  *Réseaux de communications*.
  Maîtrise Informatique (Université Claude Bernard Lyon 1), lecture: 30h/year.
- **1998-2002** : C. Pham
  Wide Area Networks.
  *Réseaux grandes distances*.
  MIAG 3rd year (University Claude Bernard Lyon 1), lecture: 20h/year.
- **2002 & 2003** : L. Lefèvre
  *Réseaux, Internet et outils associés*.
  Maitrise Informatique (Université Antilles Guyane, Pointe à Pitre), 45h eq TD/an.

## 9.4. Animation of the scientific community

- Within the Global Grid Forum, standardization entity for grid middleware, Pascale Vicat-Blanc Primet 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 where C. Pham is co-author of the Grid Network requirements for the Grid.
- Pascale Vicat-Blanc Primet is responsible for the scientifical coordination of the RNTL e-Toile project. She has coordinated the intermediate project review, project seminars and public demonstrations : CNRS Paris January 2003, ENS Lyon June 2003, RNTL conference Grenoble October 2003.
- RESO members take part of the activities of the GDR "Architecture Réseaux et Parallélisme. We partipated to different actions like the Internet New Generation summer school (Porquerolles May 2003).
- We are member of the RTP "Communication Networks" of CNRS. We are also participating to the RTP CNRS "Grille" and more particularly in its *actions specifiques*: *enabling Grid5000* (2003-2004) and Grid programming methodology (2003-2004).

## 9.5. Participation in boards of examiners and committees

- Pascale Vicat-Blanc
  - participated to the board of examiners for recruitments of *Chargés de Recherche CR2* of the Rhône-Alpes INRIA research unit in 2003.
  - participated to the board of examiners for promoting Research engineer IR1 of BAP-E to the *Hors classe grade* IRHC in 2003.
  - has been member of the board of examiners of *DEA d'Informatique Fondamentale de Lyon*.
  - has been member of the board of examiners for recruitment of a system engineer for the Ecole Normale Supérieure de Lyon.
  - has been reviewer (rapporteur) and member of the PhD thesis jury of Pierre Lombard from IMAG (Grenoble) and Ernesto Exposito from LAAS (Toulouse).
- 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 member of the PhD thesis jury of Y. Calas from LIRMM, University of Montpellier, December 2003.

## 9.6. Seminars, invited talks

- Pascale Vicat-Blanc Primet has been invited to give a seminar on "Network issues in grids" to the Piloting meeting of the RTP CNRS Networks, in St Jean de Luz, January 2003.
- Pascale Vicat-Blanc Primet has been invited to give a seminar X-Aristote at the Ecole Polytechnique " Quality of Service in Grids ", Paris, May 2003.
- Pascale Vicat-Blanc Primet has been invited to give a seminar to the Club des Utilisateurs de l'Informatique du CEA (conférence CUIC2003) "Qualité de Service dans la Grille" , St Malo, june 2003.

- Pascale Vicat-Blanc Primet has been invited to give a seminar "High Performance Transport " to the Piloting meeting of the RTP CNRS Networks, in Marseille, September 2003.
- C. Pham did a tutorial "State-of-the-art in group communications: from protocols to applications" with V. Roca, ICT'2003, Papeete, Tahiti, February 23rd, 2003.
- Laurent Lefèvre has been invited in First International Workshop on Service-Oriented Grid and Utility Computing with a talk on "Achieving performances in active networks : a mandatory step to provide dynamic network services for Grid middleware and applications", GridBus workshop, Melbourne, Australia, june 2003.
- Laurent Lefèvre has presented the "INRIA activities on IPDPS Booth", in International Parallel and Distributed Processing Symposium, IPDPS 2003, Booth session, Nice, 24 april 2003.

# 10. Bibliography

## Reference Publications by the Team

[1] F. BONNASSIEUX, F. CHANUSSOT, R. HARAKALY, P. PRIMET. *Mapcenter: An Open Grid Status Visualization Tool.* in « Proceedings of the ISCA 15th International Conference on Parallel and Distributed Computing Systems », éditeurs W. SMARI, M. GUIZIANI., pages 173-178, September, 2002.

[2] F. BOUAHFS, B. GAIDIOZ, J. GELAS, L. LEFÈVRE, M. MAIMOUR, P. C., P. PRIMET, B. TOURANCHEAU. *Evaluating and Experimenting An Active Grid Architecture.* in « Future Generation Computer System », 2004, http://www.ens-lyon.fr/~cpham/Paper/, A paraître.

[3] B. GAIDIOZ, P. PRIMET. EDS*: A new scalable Service Differentiation Architecture for Internet.* in « Proceedings of International Symposium on Computer Communication (ISCC) », IEEE, pages 777-782, Taormina, Italy, July, 2002, downloads/gaidioz-iscc02.pdf.

[4] J.-P. GELAS, S. EL HADRI, L. LEFÈVRE. *Towards the Design of an High Performance Active Node.* in « Parallel Processing Letters », number 2, volume 13, jun, 2003.

[5] L. LEFÈVRE, C. PHAM, P. PRIMET, B. TOURANCHEAU, B. GAIDIOZ, J. GELAS, M. MAIMOUR. *Active Networking Support for the Grid.* in « IFIP-TC6 Third International Working Conference on Active Networks, IWAN 2001 », series Lecture Notes in Computer Science, volume 2207, éditeurs N. W. IAN W. MARSHALL., pages 16-33, October, 2001, ISBN: 3-540-42678-7.

[6] L. LEFÈVRE, J.-P. GELAS. *Programmable Networks and their Management.* Artech House Books, UK, mar, 2004, chapter Chapter 14 "High Performance Execution Environments", to appear.

[7] 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, June, 2003, http://www.ens-lyon.fr/~cpham/Paper/ISCC03.pdf.

[8] M. MAIMOUR, C. PHAM. *Dealing with Heterogeneity in a Fully Reliable Multicast Protocol.* in « Proceedings of IEEE International Conference On Networks (ICON 2003) », Sydney, Autralia, September, 2003, http://www.ens-lyon.fr/~cpham/Paper/ICON03.pdf.

[9] G. MONTENEGRO, B. GAIDIOZ, P. PRIMET, B. TOURANCHEAU. *Equivalent Differentiated Services for AODVng.* in « ACM SIGMOBILE Mobile Computing and Communications Review », number 3, volume 6, July, 2002, pages 110-111.

[10] P. PRIMET, B. GAIDIOZ, M. GOUTELLE. *Approches alternatives pour la différenciation de services IP.* in « TSI: Techniques et Sciences Informatiques, special issue Réseaux et Protocoles », january, 2004, to appear.

[11] L. PRYLLI, B. TOURANCHEAU. *BIP: a new protocol designed for high performance networking on Myrinet.* in « Workshop PC-NOW, IPPS/SPDP98 ».

## PhD and "habilitation" theses

[12] B. GAIDIOZ. *Traitements différenciés et marquage adaptatif de paquets pour l'amélioration du transport des flux hétérogènes dans l'Internet.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, dec, 2003.

[13] J.-P. GELAS. *Vers la conception d'une architecture de réseaux actifs apte à supporter les débits des réseaux gigabits.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, dec, 2003.

[14] M. MAIMOUR. *Design, analysis and validation of router-assisted reliable multicast protocols in wide area networks.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, dec, 2003.

[15] C.-D. PHAM. *Simulations parallèles sur grappes de machines, Multicast fiable actif, Optimisations de systèmes de communications : quelques contributions pour la résistance au facteur d'échelle.* Habilitation à diriger les recherches, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, dec, 2003.

## Articles and Book Chapters

[16] A. BASSI, M. BECK, J. LAGANIER, G. PAOLLINI. *Enhancing Grid Capabilities:IBP over IPv6.* in « Future Generation Computer Systems special issue on Advanced Grid Technologies », March, 2004, to appear.

[17] J.-P. GELAS, S. EL HADRI, L. LEFÈVRE. *Towards the Design of an High Performance Active Node.* in « Parallel Processing Letters », number 2, volume 13, jun, 2003.

[18] L. LEFÈVRE, J.-P. GELAS. *Programmable Networks and their Management.* Artech House Books, UK, mar, 2004, chapter 14 of "High Performance Execution Environments", to appear.

[19] M. MAIMOUR, C. PHAM. *Dynamic Replier Active Reliable Multicast (DyRAM).* in « Journal of Cluster Computing », 2004, http://www.ens-lyon.fr/~cpham/Paper/, to appear.

[20] M. MAIMOUR, C. PHAM. *Experimenting Active Reliable Multicast on Application-Aware Grids.* in « Journal of Grid Computing », 2004, http://www.ens-lyon.fr/~cpham/Paper/, to appear.

[21] P. PRIMET, B. GAIDIOZ, M. GOUTELLE. *Approches alternatives pour la différenciation de services IP.* in « TSI: Techniques et Sciences Informatiques, special issue Réseaux et Protocoles », january, 2004, to appear.

[22] P. VICAT-BLANC PRIMET, F. BONNASSIEUX, R. HARAKALY. *Network monitoring in the DataGRID project.* in « International Journal of High Performance Computer Applications », January, 2004, to appear.

[23] P. VICAT-BLANC PRIMET, P. D'ANFRAY. *Les grilles haute-performance et le projet Etoile.* in « Matapli », number 71, May, 2003.

[24] P. VICAT-BLANC/PRIMET. *High Performance Grid Networking in the DataGrid Project.* in « special issue Future Generation Computer Systems », January, 2003.

## Conference and Workshop Publications, etc.

[25] A. BASSI, J.-P. GELAS, L. LEFÈVRE. *A Sustainable Framework for Multimedia Data Streaming.* in « International workshop on active networks (IWAN2003) », Kyoto, Japan, dec, 2003.

[26] A. BASSI, J. LAGANIER. *Towards an IPv6-based Security Framework for Distributed Shared Storage.* March, 2003, http://www.inria.fr/rrrt/rr-4817.html, In IFIP CMS'03. Also published as research report INRIA #4817 and LIP #2003-19.

[27] F. BONNASSIEUX, R. HARAKALY, P. PRIMET. *Automatic services discovery, monitoring and visualization of grid environments: the MapCenter approach.* in « Across Grid workshop », February, 2003.

[28] F. BONNASSIEUX, R. HARAKALY, P. PRIMET. *MapCenter : un modèle ouvert pour la découverte, la supervision et la visualisation des environnements distribués à large échelle.* in « Conference JRES », November, 2003.

[29] A. GALIS, J.-P. GELAS, L. LEFÈVRE, K. YANG. *Active Network Approach to Grid Management & Services.* in « Workshop on Innovative Solutions for Grid Computing - ICCS 2003 Conference », pages 1103-1113, Melbourne, Australia, jun, 2003, LNCS 2658, ISBN 3-540-40195-4.

[30] A. GALIS, L. LEFÈVRE. *Programmable and Active Networks : a network infrastructure for next generation GRIDs.* in « Parco20003, Parallel Computing 2003 conference - Mini Symposia on Grid Computing », Dresden University of Technology, Germany, sep, 2003.

[31] J.-P. GELAS, S. EL HADRI, L. LEFÈVRE. *Tamanoir: a software active node supporting gigabit networks.* in « ANTA 2003 : The second International Workshop on Active Networks Technologies and Applications », pages 159-168, Osaka, Japan, may, 2003.

[32] M. GOUTELLE, P. PRIMET. *Study of a non-intrusive method for measuring the hop-by-hop capacity of a path.* in « Best 2003, Bandwidth Estimation Workshop », CAIDA DOE, San Diego (CA) BEst 2003, December, 2003.

[33] L. LEFÈVRE, J.-P. GELAS. *Active Web : active networking support for web transport.* in « ANTA 2003 : The second International Workshop on Active Networks Technologies and Applications », pages 147-156, Osaka, Japan, may, 2003.

[34] L. LEFÈVRE, J.-M. PIERSON, S. GUEBLI. *Collaborative web caching with active networks.* in « International workshop on active networks (IWAN2003) », Kyoto, Japan, dec, 2003.

[35] E. LEMOINE, C. PHAM, L. LEFÈVRE. *Packet Classification in the NIC for Improved SMP-based Internet Servers.* in « Proceedings of IEEE 3rd International Conference on Networking (ICN'04) », Guadeloupe, French Caribbean, March, 2004, http://www.ens-lyon.fr/~cpham/Paper/ICN04.pdf, to appear.

[36] M. MAIMOUR, C. PHAM. *A RTT-based Partitioning Algorithm for a Multi-rate Reliable Multicast Protocol.* in « Proceedings of the IEEE High-Speed Network and Multimedia Communications Conference (HSNMC 2003) », Estoril, Portugal, July, 2003, http://www.ens-lyon.fr/~cpham/Paper/HSMNC03.pdf.

[37] 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, June, 2003, http://www.ens-lyon.fr/~cpham/Paper/ISCC03.pdf.

[38] M. MAIMOUR, C. PHAM. *Dealing with Heterogeneity in a Fully Reliable Multicast Protocol.* in « Proceedings of IEEE International Conference On Networks (ICON 2003) », Sydney, Autralia, September, 2003, http://www.ens-lyon.fr/~cpham/Paper/ICON03.pdf.

[39] P. V.-B. PRIMET, F. CHANUSSOT, C. BLANCHET, N. LACORNE, P. D'ANFRAY.. *E-Toile: High performance Grid Middleware.* in « IEEE International Cluster Conference. Grid Demo session », December, 2003.

[40] A. VERNOIS, P. VICAT-BLANC, F. DESPREZ, F. HERNANDEZ, C. BLANCHET. *GriPPS : Grid Protein Pattern Scanning.* in « in proceedings of the International HealthGrid 2003 : 1st Conference Workshop of HealthGrid cluster », Elsevier, January, 2003.

## Technical Reports

[41] T. FERRARI, P. PRIMET, R. HUGUES-JONES, M. GOUTELLE, P. C., ET AL.. *Network Monitoring Architecture.* rapport de contrat, EU DATAGRID IST-2000-25182 report Deliverable D7.3, October, 2003.

[42] P. P. FRANCK BONNASSIEUX. *Network Services - final report.* Rapport de Recherche, European DataGrid project, December, 2003, http://.

[43] B. GAIDIOZ, P. PRIMET. *End-to-end delay constrained protocol over the EDS service differentiation.* Technical Report, number RR-5030, INRIA, December, 2003, http://www.inria.fr/rrrt/rr-5030.html.

[44] B. GAIDIOZ, P. PRIMET. *Implementation of proportional loss rate differentiation in EDS |using Proportional Loss Rate and RED.* Technical Report, number RR-5029, INRIA, December, 2003, http://www.inria.fr/rrrt/rr-5029.html.

[45] B. GAIDIOZ, P. PRIMET. *Reliable and interactive protocol for short messages over the EDS service differentiation.* Technical Report, number RR-5031, INRIA, December, 2003, http://www.inria.fr/rrrt/rr-5031.html.

[46] B. GOGLIN, L. PRYLLI. *Design and Implementation of ORFA.* Technical Report, number TR2003-01, LIP, ENS Lyon, Lyon, France, September, 2003, http://www.ens-lyon.fr/LIP/Pub/Rapports/TR/TR2003/TR2003-01.ps.gz.

[47] B. GOGLIN, L. PRYLLI. *Performance Analysis of Remote File System Access over High Bandwidth Local Network.* Research Report, number RR2003-22, LIP, ENS Lyon, Lyon, France, April, 2003, ftp://ftp.ens-lyon.fr/pub/LIP/Rapports/RR/RR2003/RR2003-22.ps.gz, Also available as Research Report RR-4795, INRIA Rhône-Alpes.

[48] M. GOUTELLE, P. PRIMET. *Study of a non-intrusive and accurate method for measuring the end-to-end useful bandwidth in a high rate/latency product link.* Rapport de Recherche, number RR-4959, INRIA Rhône-Alpes, October, 2003, http://www.inria.fr/rrrt/rr-4959.html.

[49] P. PRIMET, F. CHANUSSOT. *Spécification du service actif QoSINUS.* Technical report, number RR-0287, INRIA, 2003.

## Miscellaneous

[50] J. LAGANIER, G. MONTENEGRO. *Using IKE with Cryptographically Generated Addresses.* Internet draft `draft-laganier-ike-ipv6-cga-01.txt`, June, 2003, http://www.ietf.org/internet-drafts/draft-laganier-ike-ipv6-cga-01.txt, Work in progress, expired in December 2003.

[51] G. MONTENEGRO, J. LAGANIER, C. CASTELLUCIA. *Securing IPv6 Neighbor Discovery.* Internet draft `draft-montenegro-send-cga-rr-01.txt`, March, 2003, http://www.ietf.org/internet-drafts/draft-montenegro-send-cga-rr-01.txt, Work in progress, expired in August 2003.

[52] E. NORDMARK, S. CHAKRABARTI, J. LAGANIER. *Source Address Selection API for IPv6.* Internet draft `draft-chakrabarti-addrselect-api-02.txt`, October, 2003, http://www.ietf.org/internet-drafts/draft-chakrabarti-ipv6-addrselect-api-02.txt, Work in progress, expired in April 2004.

## Bibliography in notes

[53] O. AUMAGE, L. BOUGÉ, J.-F. MÉHAUT, R. NAMYST. *Madeleine II: A Portable and Efficient Communication Library for High-Performance Cluster Computing.* in « Parallel Computing », number 4, volume 28, April, 2002, pages 607–626.

[54] A. AZCORRA, M. CALDERÓN, M. SEDANO, J. I. MORENO. *Multicast Congestion Control for Active Network Services.* in « European Transactions in Telecommunications », number 3, volume 10, May/June, 1999.

[55] S. BLAKE, D. BLACK, M. CARLSON, E. DAVIES, Z. WANG, W. WEISS. *An architecture for differentiated services.* in « RFC 2475 », December, 1998.

[56] R. BRADEN, D. CLARK, S. SHENKER. *Integrated services in the internet architecture: an overview.* in « RFC 1633 », June, 1994.

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

[58] D. M. CHIU, M. KADANSKY, J. PROVINO. *A Congestion Control Algorithm for Tree-based Reliable Multicast Protocols.* in « Infocom 2002 », 2002.

[59] C. DOVROLIS, P. RAMANATHAN. *A case for relative differentiated services and the proportional differentiation model.* in « IEEE Networks », number 5, volume 13, September, 1999, pages 26–34, http://citeseer.nj.nec.com/article/dovrolis99case.html.

[60] C. DOVROLIS, P. RAMANATHAN, D. MOORE. *What Do Packet Dispersion Techniques Measure?.* in « Proceedings of INFOCOM'01 », pages 905-914, 2001, citeseer.nj.nec.com/479183.html.

[61] T. FERRARI, R. HUGHES-JONES, P. VICAT-BLANC PRIMET. *Collaborative investigations between DataGrid WP7 and DANTE.* Technical Report, European DataGrid Technical Document, May, 2002, http://ccwp7.in2p3.fr, In english.

[62] S. FLOYD. *HighSpeed TCP for Large Congestion Windows.* in « Internet draft, draft-floyd-tcp-highspeed-01.txt, work in progress, 2002 », pages work in progress, 2002, citeseer.nj.nec.com/479183.html.

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

[64] J.-P. GELAS, L. LEFÈVRE. *Towards the design of an Active Grid.* in « Computational Science - ICCS 2002 », volume 2230, éditeurs L. N. IN COMPUTER SCIENCE., pages 578-587, Amsterdam, The Netherlands, apr, 2002, http://www.ens-lyon.fr/~llefevre/Papers/LG02.ps.gz, ISBN 3-540-43593-X.

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

[66] M. HERBERT, P. V.-B. PRIMET. *Network of Queues.* in « submitted to International Protocol for Long Distance Conference. », 2004.

[67] P. HURLEY, J.-Y. LE BOUDEC, P. THIRAN, M. KARA. *ABE: Providing a Low-Delay Service within Best Effort.* in « IEEE Networks », number 5, volume 15, May, 2001, pages 60–69, http://citeseer.nj.nec.com/hurley01abe.html.

[68] S. KASERA, S. BHATTACHARYA. *Scalabe Fair Reliable Multicast Using Active Services.* in « IEEE Network Magazine's Special Issue on Multicast », 2000.

[69] T. KELLY. *Scalable TCP: Improving Performance in Highspeed Wide Area Networks.* in « Protocol for Long Distance Networks Conference », number Pfldnet-1, February, 2003.

[70] L. LEFÈVRE, C. PHAM, P. PRIMET, B. TOURANCHEAU, B. GAIDIOZ, J. GELAS, M. MAIMOUR. *Active Networking Support for the Grid.* in « IFIP-TC6 Third International Working Conference on Active Networks, IWAN 2001 », series Lecture Notes in Computer Science, volume 2207, éditeurs N. W. IAN W. MARSHALL., pages 16-33, October, 2001, http://www.ens-lyon.fr/~cpham/Paper/IWAN01.ps.gz, ISBN: 3-540-42678-7.

[71] L. LEHMAN, S. GARLAND, D. TENNEHOUSE. *Active Reliable Multicast.* in « Proc. of the IEEE INFOCOM, San Francisco, CA », March, 1998.

[72] M. MAIMOUR, J. MAZUY, C. PHAM. *The Cost of Active Services in Active Reliable Multicast.* in « Proceedings of the 4th IEEE Annual International Workshop on Active Middleware Services (AMS 2002) », pages 67-72, Edinburg, UK, July, 2002, http://www.ens-lyon.fr/~cpham/Paper/AMS02.ps.gz.

[73] M. MAIMOUR, C. PHAM. *A Throughput Analysis of Reliable Multicast Protocols in an Active Networking Environment.* in « Proceedings of the Sixth IEEE Symposium on Computers and Communications (ISCC 2001) », pages 151-158, Hammamet, Tunisia, July, 2001, http://www.ens-lyon.fr/~cpham/Paper/ISCC01.ps.gz.

[74] M. MAIMOUR, C. PHAM. *A Loss Detection Service for Active Reliable Multicast Protocols.* in « Proceedings of the International Network Conference (INC'2002) », Plymouth, UK, July, 2002, http://www.ens-lyon.fr/~cpham/Paper/INC02.pdf.

[75] M. MAIMOUR, C. PHAM. *An Active Reliable Multicast Framework for the Grids.* in « Proceedings of the International Conference on Computational Science (ICCS 2002) », series Lecture Notes in Computer Science, volume 2330, pages 588-597, April, 2002, http://www.ens-lyon.fr/~cpham/Paper/ICCS02.ps.gz.

[76] M. MAIMOUR, C. PHAM. *An Analysis of a Router-based Loss Detection Service for Active Reliable Multicast Protocols.* in « Proceedings of the 11th IEEE International Conference on Networks (ICON 2002) », pages 49-56, Singapour, August, 2002, http://www.ens-lyon.fr/~cpham/Paper/ICON02.ps.gz.

[77] M. MAIMOUR, C. PHAM. *Dynamic Replier Active Reliable Multicast (DyRAM).* in « Proceedings of the 7th IEEE Symposium on Computers and Communications (ISCC 2002) », pages 275-282, Taormina, Sicily, July, 2002, http://www.ens-lyon.fr/~cpham/Paper/ISCC02.ps.gz.

[78] M. MAIMOUR, C. PHAM. *Towards an application-aware communication framework for computational grids.* in « Proceedings of the Asian Computing Science Conference (ASIAN 2002) », pages 140-152, Hanoi, Vietnam, December, 2002, http://www.ens-lyon.fr/~cpham/Paper/ASIAN02.ps.gz.

[79] R. S. PRASAD, C. DOVROLIS, B. A. MAH. *The effect of layer-2 store-and-forward devices.* in « Proceedings of INFOCOM '03 », San Fransisco, CA, April, 2003.

[80] P. V.-B. PRIMET, J. MONTAGNAT, F. CHANUSSOT, M. GOUTELLE. *Network Quality of Service in Grid environments: the QoSinus approach.* in « submitted to IEEE International IPDPS Conference. », 2004.

[81] I. RHEE, N. BALLAGURU, G. N. ROUSKAS. *MTCP : Scalable TCP-like Congestion Control for Reliable Multicast.* in « Infocom 1999 », 1999.

[82] P. SPATHIS, K. L. THAI. *MAF: un protocole de multicast fiable.* in « CFIP 2002 », January, 2002.

[83] K. STOCKINGER, H. STOCKINGER, R. HARAKALY, F. BONNASSIEUX, P. PRIMET. *Optimisation of File Replication in a Data Grid Using Network Cost Function.* in « submitted to Journal of Grid Computing in September 2003 », 2003.

[84] A. WAHEED, W. SMITH, J. GEORGE, J. YAN. *An Infrastructure for Monitoring and management in Computational Grids.* in « IEEE International Symposium on High-Performance Distributed Computing », number HPDC-10, 2001.

[85] R. WOLSKI, N. SPRING, J. HAYES. *The Network Weather Service: A Distributed Resource Performance Forecasting Service for Metacomputing.* in « Future Generation Computing Systems », 1999, www.cs.ucsd.edu/groups/hpcl/apples/hetpubs.html.