



Data Grid eXplorer: une plate-forme d'émulation de grilles

Franck Cappello, Pascale Primet, Olivier Richard,
Christophe Cérin, Pierre Sens



ACI
*Masses de
Données*





Plan de l'exposé

- Contexte des grilles
- Problématique de l'émulation
- L'approche GDX
- Les projets sur GDX

Contexte

2 types de
grands
systèmes
distribués

Les Grilles de calcul
ou « GRID »

→ Grands sites
de calcul
Clusters

Caractéristiques
des nœuds :

- <1000
- Stables
- Identification individuelle
- Confiance

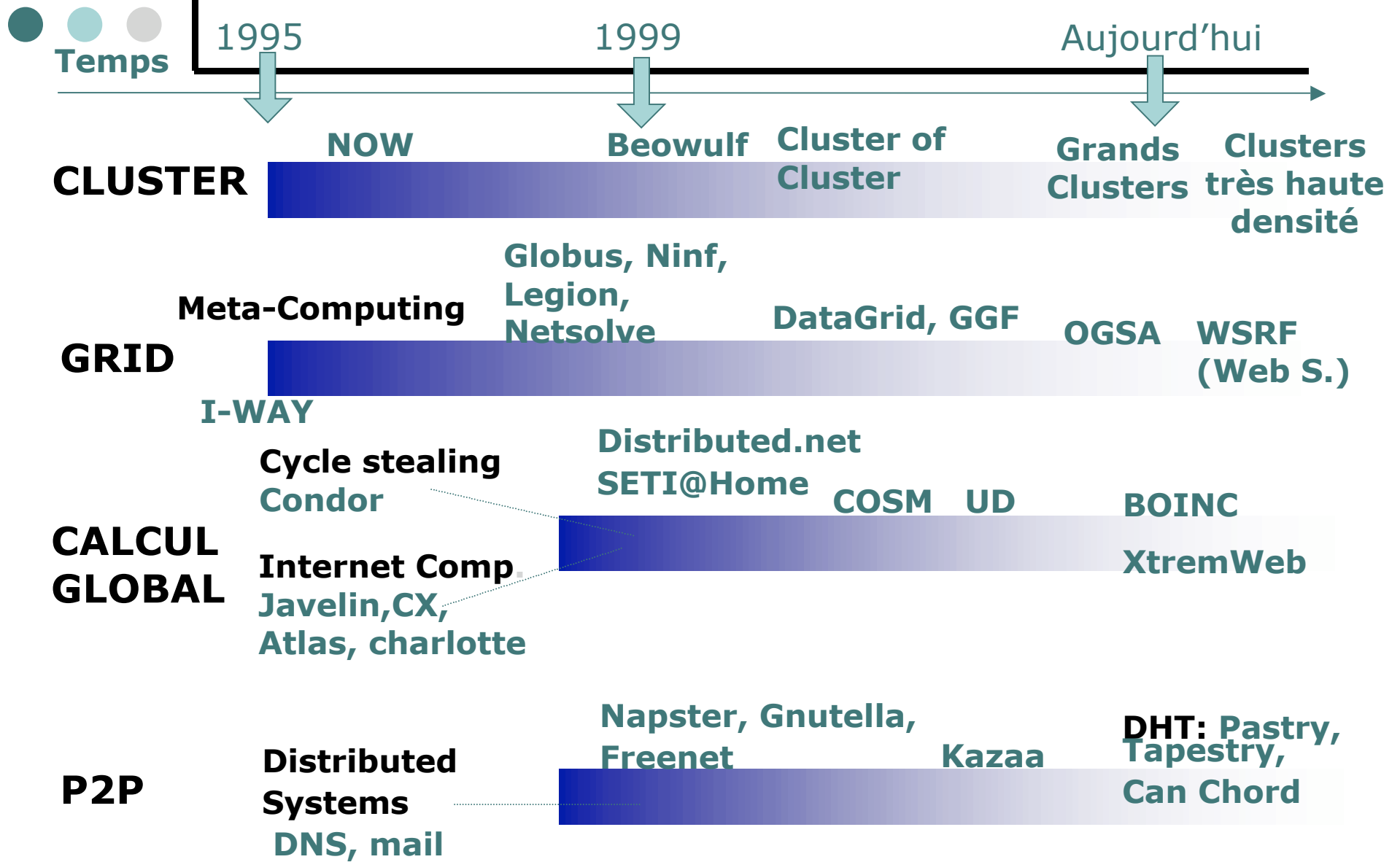
Les systèmes distribués
à grande échelle

→ Les systèmes de Calcul Global
→ Les systèmes Pair à Pair

→ PC

- ~100 000
- Volatiles
- Pas d'ident individuelle
- Pas de confiance

Perspective historique



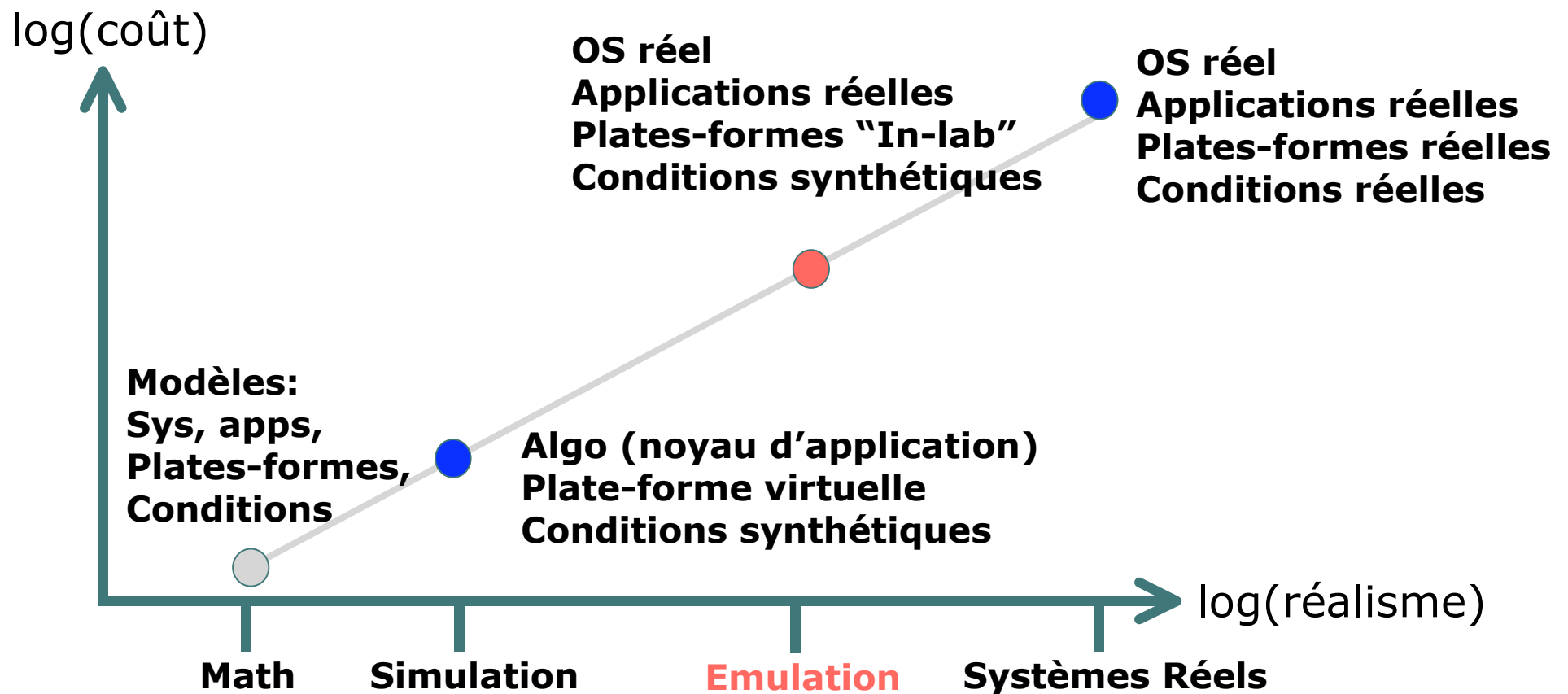
Défis des grands systèmes

Difficultés :

Passage à l'échelle => changement de problématique

- Délai de transmission variable
 - Grand nombre de ressources
 - Fautes
 - Dynamicité
 - Sécurité
 - Hétérogénéité
- } **Pas d'état global**
- Développement de nouveaux systèmes et algorithmes :
 - Prise en compte de la topologie – approche hiérarchique
 - Prise en compte de l'incertitude - algorithmes « indulgents »
 - Tolérances aux fautes
 - Importance d'expérimenter, dimensionner, calibrer, comparer les protocoles et les algorithmes

Outils de validation





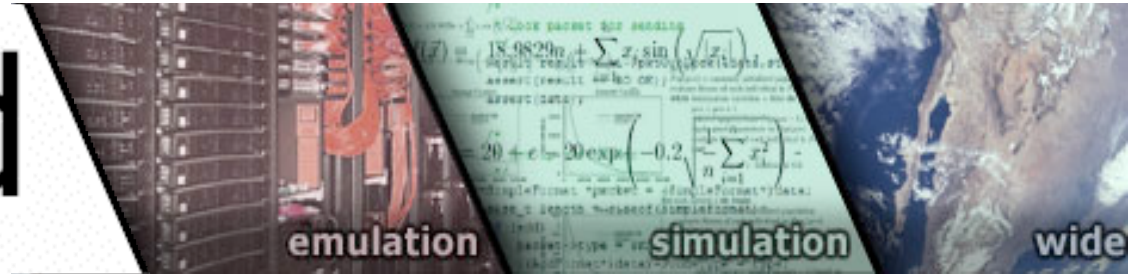
Systemes existants : Simulateurs

- **Simulateurs à événements discrets**
 - Dédiés à un type d'expériences
- **Pour les grilles :**
 - SimGrid / SimGrid2 (Univ. San Diego / ENS Lyon)
 - GridSim (Univ. Melbourne)
 - Dédiés aux études sur l'ordonnancement sur grilles
- Nombre de sites limités (~100)
- **Pour les systèmes pair-à-pair :**
 - SimPastry (Microsoft)
 - Large Scale Simulateur – LS3 (Univ. Paris 6 / INRIA)

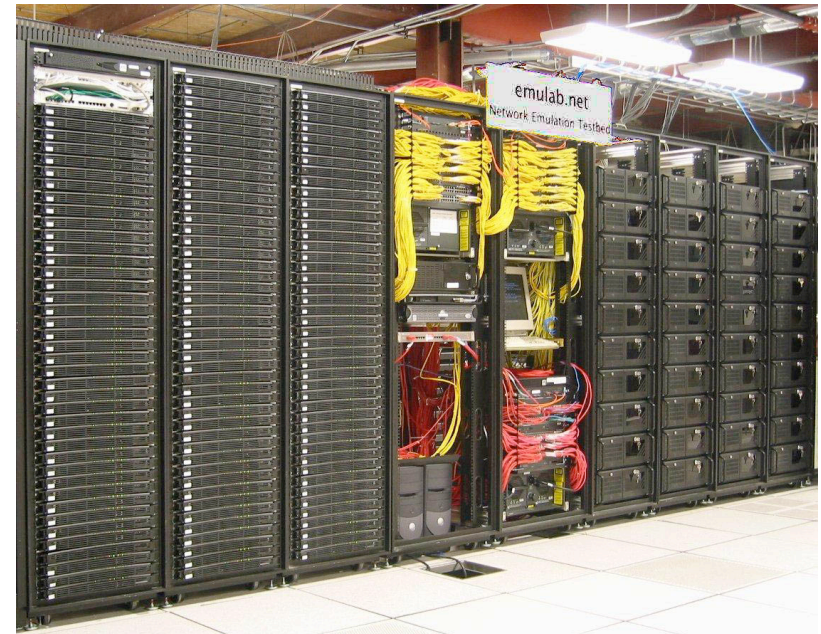


Systemes existants

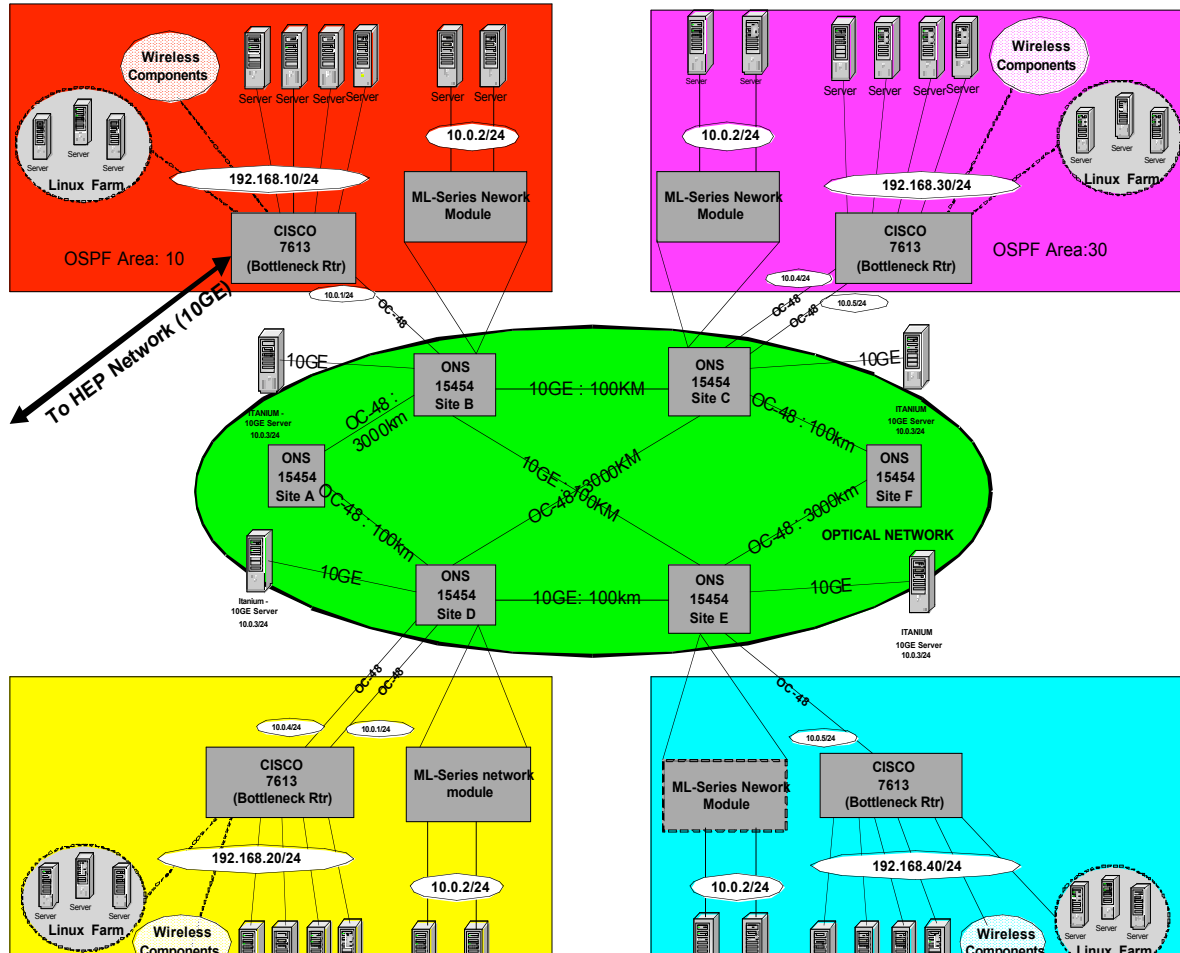
- Emulateurs :
 - Netbed - emulab (Univ. Utah)
 - Wan In Lab (Caltech)
- Systemes réels :
 - Planetlab (consortium):
 - Sites distribués sur Internet (Nov. 2005 : 631 noeuds sur 298 sites)
 - Conditions réseaux réelles
 - Non reproductibilité
 - Nombre de CPU faibles
 - Très fortes charges



- Description de la topologie => NS script
- Utilisation de DummyNet
- Outils de Mapping routeur_logique => machine physique
- Utilisation du simulateur NSE (ns emulation)
- Utilisation de noeuds extérieurs client (40 DSL)
- Environ 200 noeuds, partagés par les expériences



Wan In Lab



6 Cisco
ONS15454
4 routers
10s servers
Wireless devices
800km fiber
~100ms RTT

Systeme réel : PlanetLab



- Nombre de nœuds réduit (Novembre 2005 – 632 nœuds sur 298 sites)
- Expériences réelles : base de mesure pour calibrer les outils (simulateurs / émulateurs)
- Non reproductibilité

Data Grid eXplorer

Plate-forme expérimentale
mutualisée à l'échelle nationale

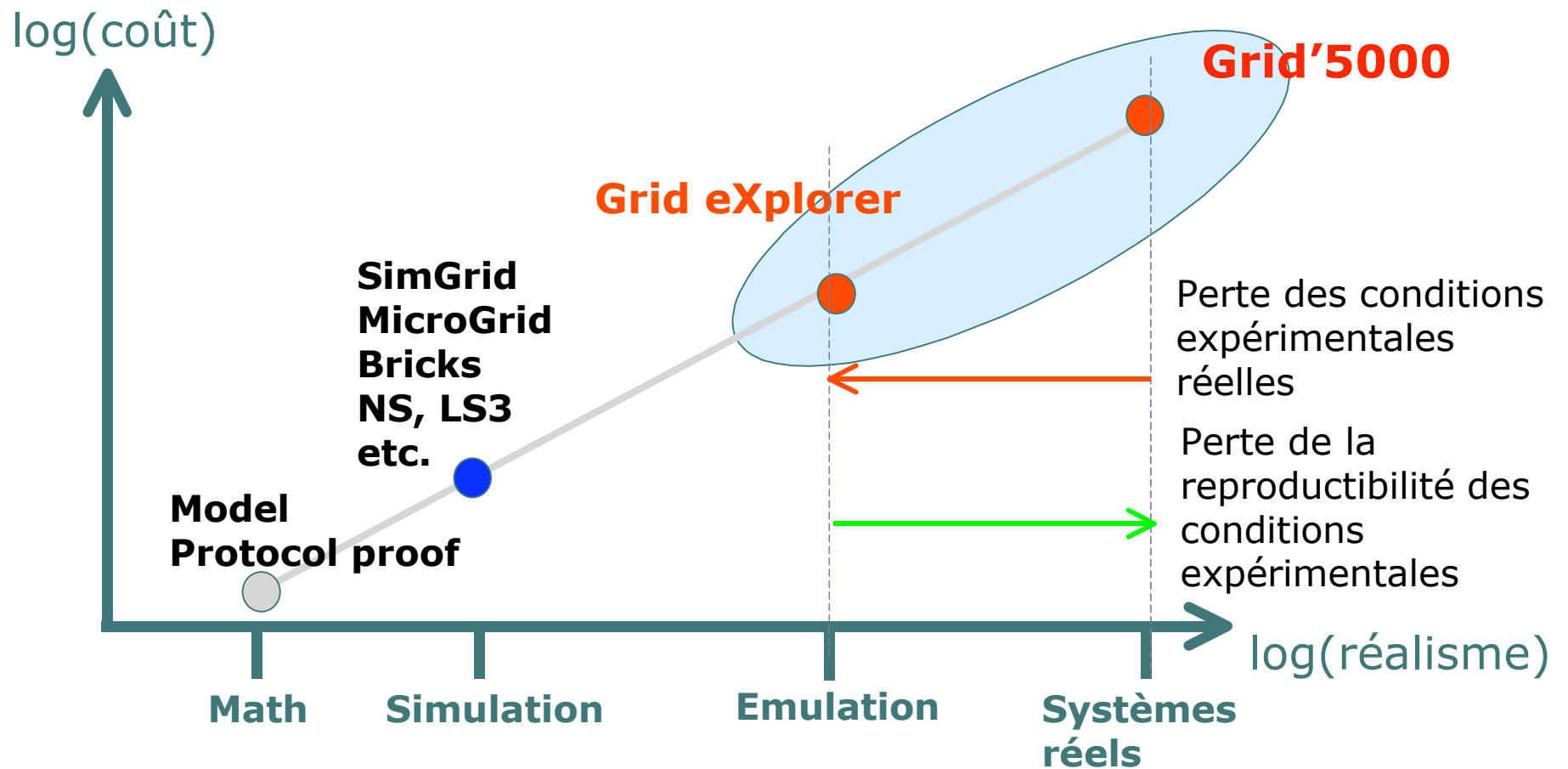
ACI Masse de Données
INRIA
CNRS
Région IdF



ACI
Masses
de
Données



Positionnement de Grid eXplorer





GDX : Les objectifs

- Cluster pour expérimentation Grid et P2P
 - Ce n'est pas un environnement de production
- Emuler les systèmes à large échelle
 - Virtualisation de nœuds :
 - Grid (~10 000 nœuds)
 - Pair-à-pair (de 100 000 à plusieurs millions de nœuds)
 - Emulation de longues distances :
 - Liaisons nationales (20ms) et transcontinentales (300ms)
- Observation :
 - Journalisation, Sondes pour observer points chauds
- Reproductibilité
 - Base de Fichiers et de Scripts de configuration
- Reconfigurations :
 - Systèmes de déploiement automatique d'applications et de noyaux



Emulation

- **Principe :**

- Un cluster dédié aux expérimentations à large échelle
- Exécution d'applications réelles dans un environnement contraint

- **Propriétés attendues :**

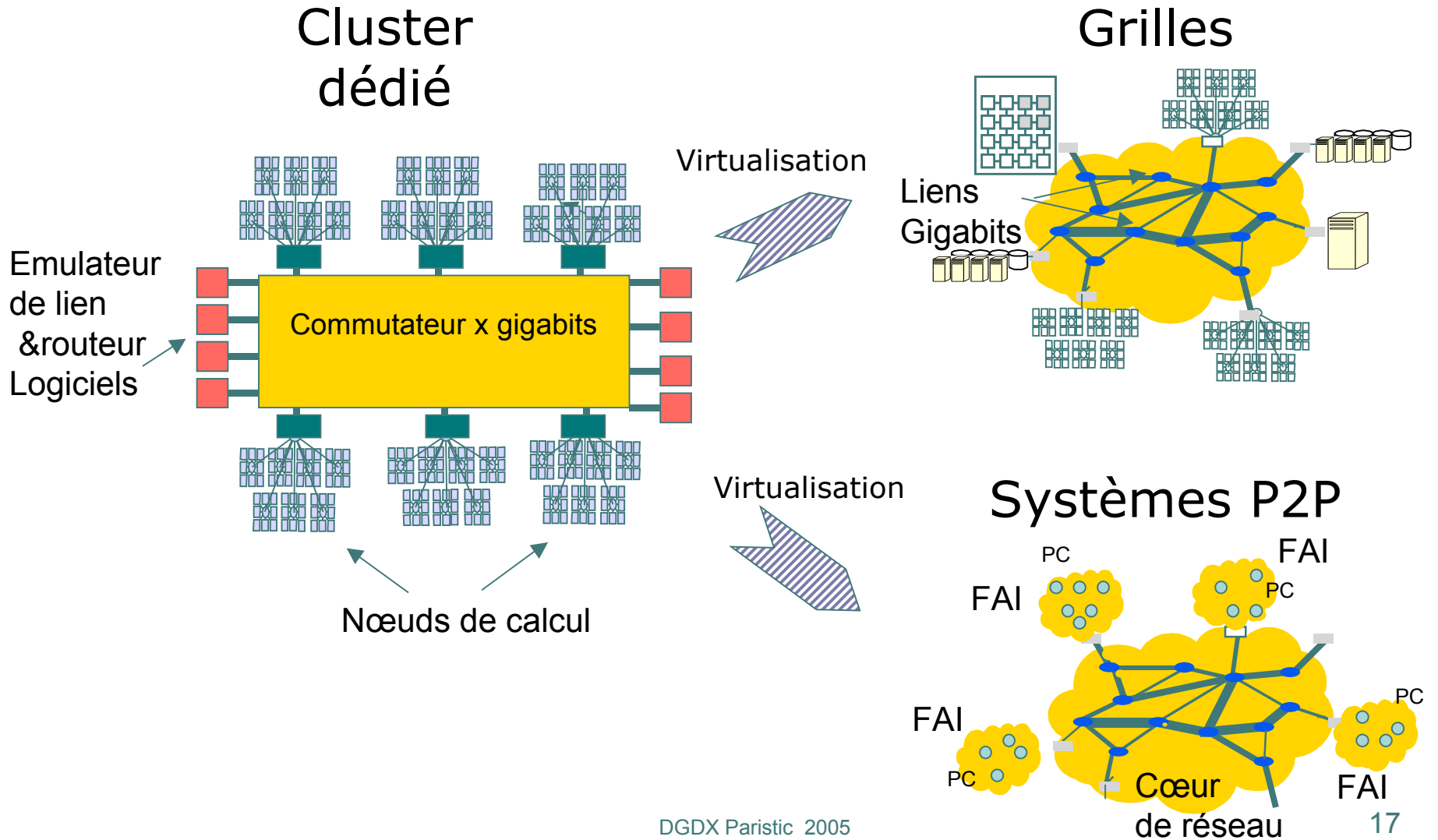
- Proche de la réalité (dépendant de la taille du cluster)
- **Reproductibilité** des mesures
- Observation fine des phénomènes
- Reconfigurabilité : possibilité de tester différentes configurations logicielles et matérielles



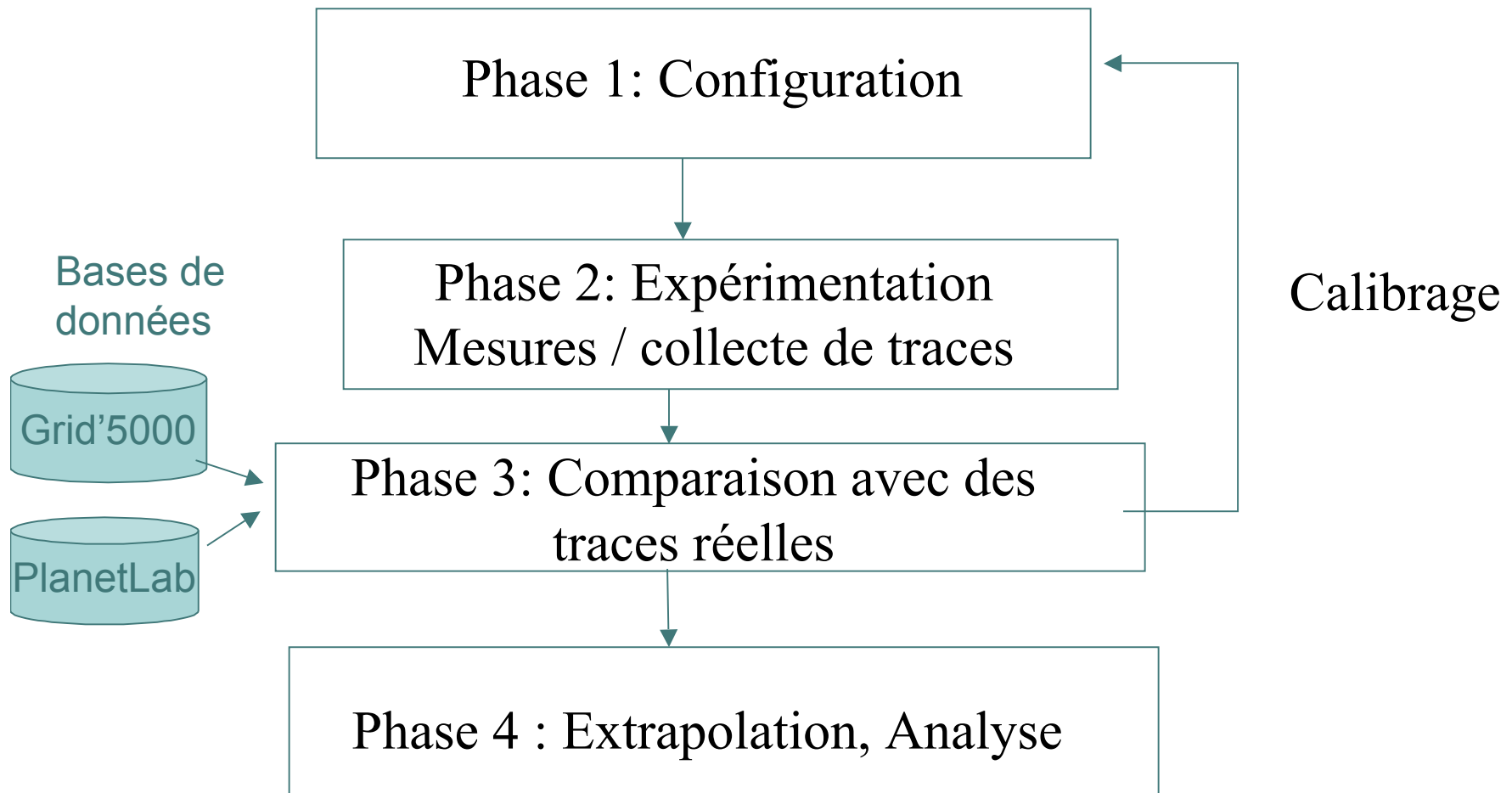
Emulation: Briques de bases

- **Routeurs logiciels et émulation de distance:**
 - Topologie virtuelle:
 - Réseau d'overlay
 - Emulation de distance:
 - Ajouts de latence, sauts (hops)
 - Emulation ou génération de charge
 - Taux d'erreurs & Variation de latence
 - Outils logiciels ou matériels:
 - DummyNet (FreeBSD), NISNet & netem(Linux), Modelnet (FreeBSD)
 - FPGA: GNET1 et GNET10, Network processors
- **Virtualisation de sites**
 - Virtualisation de nœuds logiques sur 1 nœud physique
 - Partage équitable de ressource (CPU, Mémoire)
 - Limiter les effets de bord
 - Outils :
 - Vserver, Xen, User Mode Linux

Principe



Principe : « boucle » d'émulation





Data Grid eXplorer : Ressources H

-Responsable Franck Cappello (LRI, Inria Grand-Large)
-Référent ACI MD Luc Bougé (ENS)

Les 4 thématiques transversales et leur responsable :

-Infrastructure (Matériel + système),	Olivier Richard (ID-IMAG)
-Emulation (Virtualisation),	Pierre Sens (LIP6, Inria REGAL)
-Réseau (infrastructure & emulation)	Pascale Primet (LIP, Inria RESO)
-Applications.	Christophe Cérin (LIPN)

•Comité Technique:

- Philippe Marty (philippe.marty@lri.fr) (CDD INRIA)
- Julien Leduc (leduc@imag.fr) (négociations INRIA)
- Jean-Claude Barbet (jean-claude.barbet@lri.fr) (bénévole - permanent LRI)
- Gilles Gallot (gilles.gallot@idris.fr) (bénévole - permanent IDRIS)



Les partenaires

- Institutionnels :
 - CEA, Ecole Centrale Paris
ID-IMAG, INRIA, IBCP
LAAS, LABRI, LARIA, LIFL
LIP, LIP6, LORIA, PRISM
- Industriels :
 - Alcatel Space, France-Télécom R&D, EADS
- Internationaux:
 - AIST (Japon), Argonne (USA)



GdX: Ressources financières

Financement Equipement:

ACI Masse de données:	750 K€ TTC
ACI Grid'5000 2004 :	155 K€ TTC
INRIA Rocquencourt:	150 K€ TTC
INRIA Futurs:	150 K€ TTC
Hébergement IDRIS	170 K€ TTC/an

SESAME Ile de France: 900 K€ TTC

Moyens humains:

36 mois (2x18 mois) ingénieur ACI Masse de données

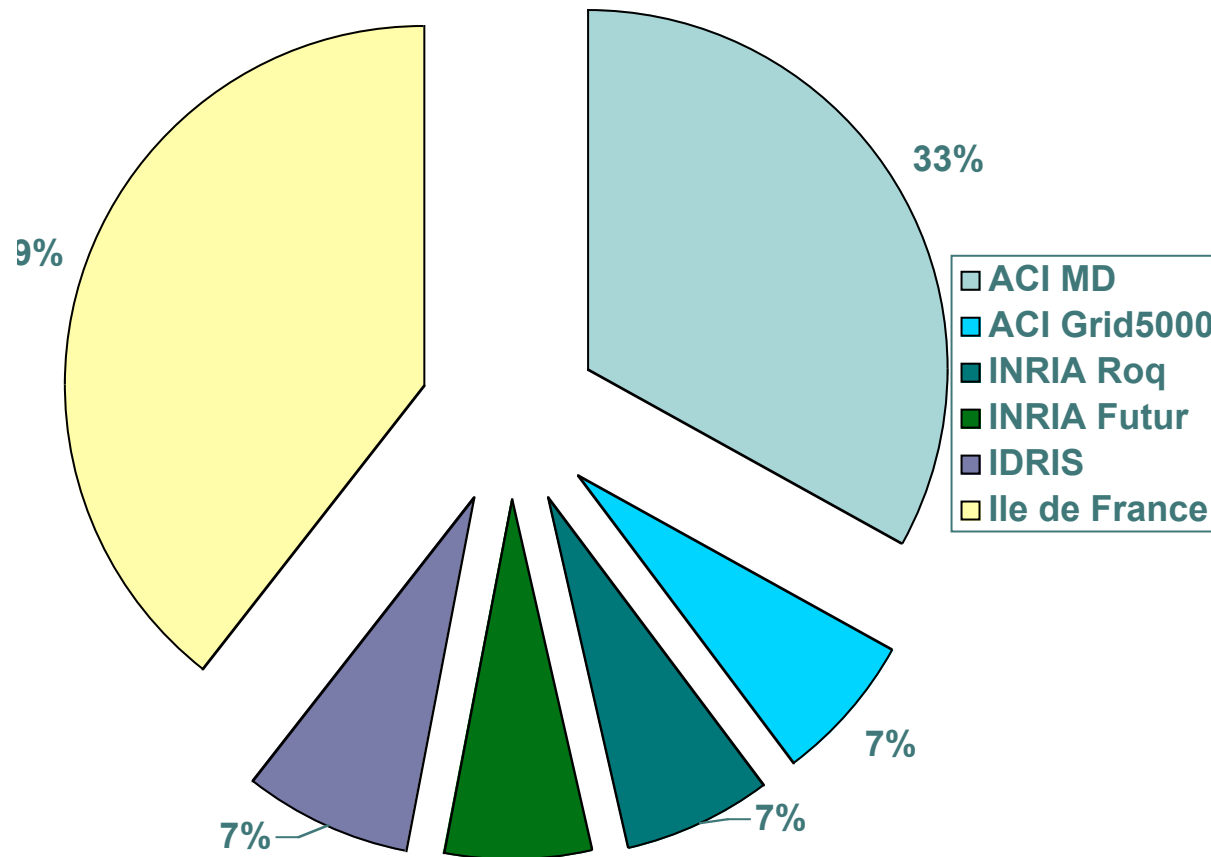
24 mois ingénieur associé INRIA

Ingénieurs IDRIS (difficile à quantifier)

Soutien ingénieurs LRI (difficile à quantifier)

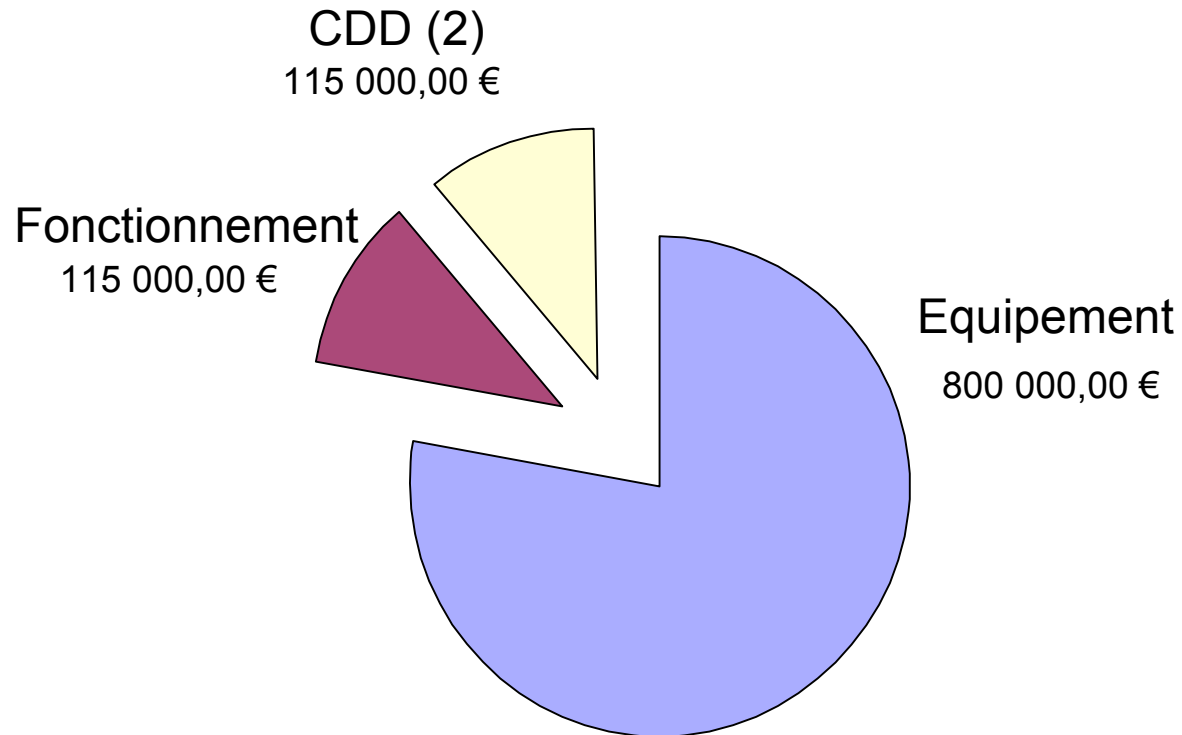
+ Connexion RENATER

Financement Equipement



Budget ACI Masse de Données

Budget total 1030 K€ TTC

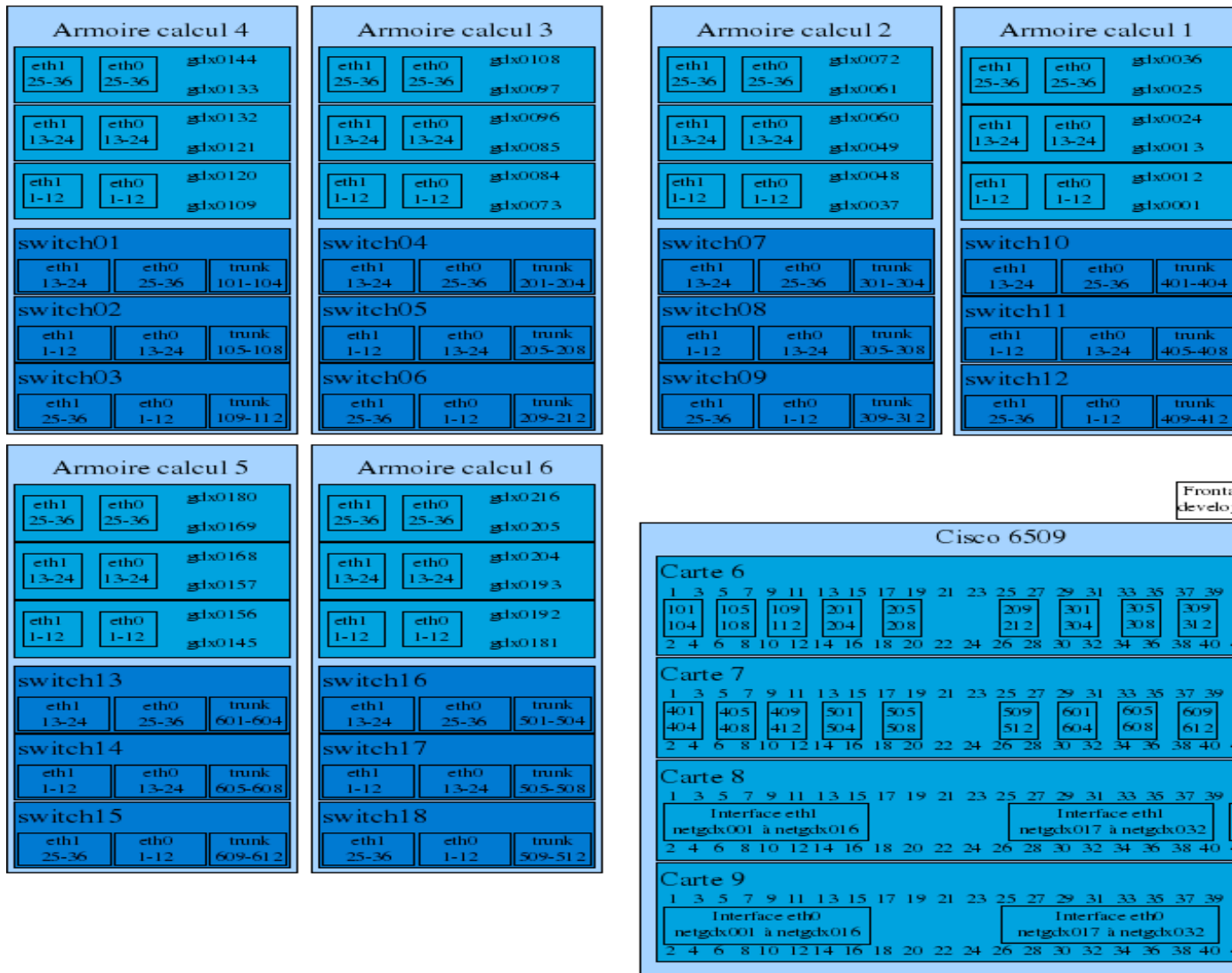




GdX: tranche 1 : fin 2004

- Noeuds « **de calcul** » :
 - 216 IBM eServer 325
 - 2 CPU AMD Opteron 64 bits : 2GHz Horloge / 1Mo Cache
 - 4 DIMM-DDR 512Mo
 - 2 interfaces Ethernet Gigabit
 - 1 HD 80 Go IDE
- Noeuds « **réseau** » :
 - 32 xSeries 335 (Xeon-32bits, 3GHz, 512Ko, 2Go, 2Gb, 40Go IDE)
- Noeuds « **de service** » :
 - 2 xSeries 346 (bi-Xeon-Nokona-64bits, 3GHz, 1Mo, 2Go, 2Gb, 1To Raid)
 - 2 eServeur 325 (bi-Opteron-64bits, 2GHz, 1Mo, 2Go, 4Gb, 80Go IDE)
- Réseau hiérarchique Gb Ethernet
 - Routeur Cisco 6905 (360 Gbps) 200 ports Gigabit Ethernet
 - Commutateurs Cisco 24 ports Gigabit Ethernet

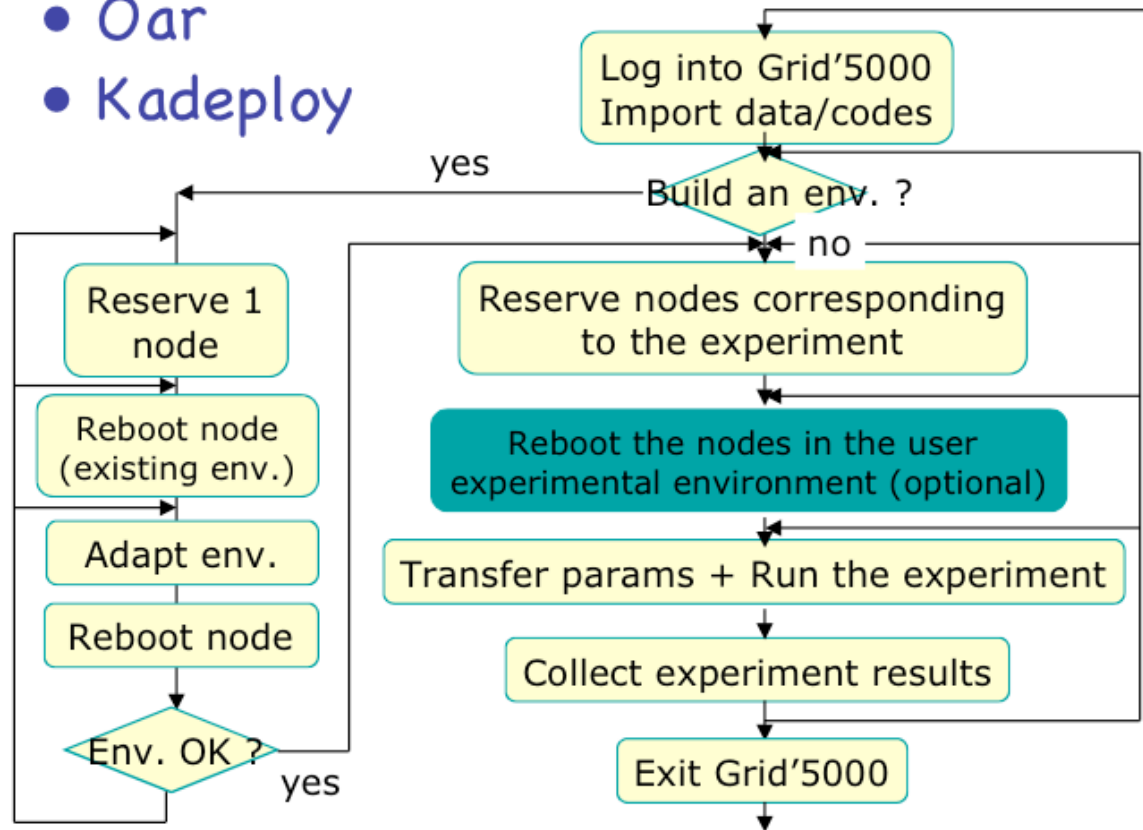
Câblage réseau



Les outils

Main tools developed:

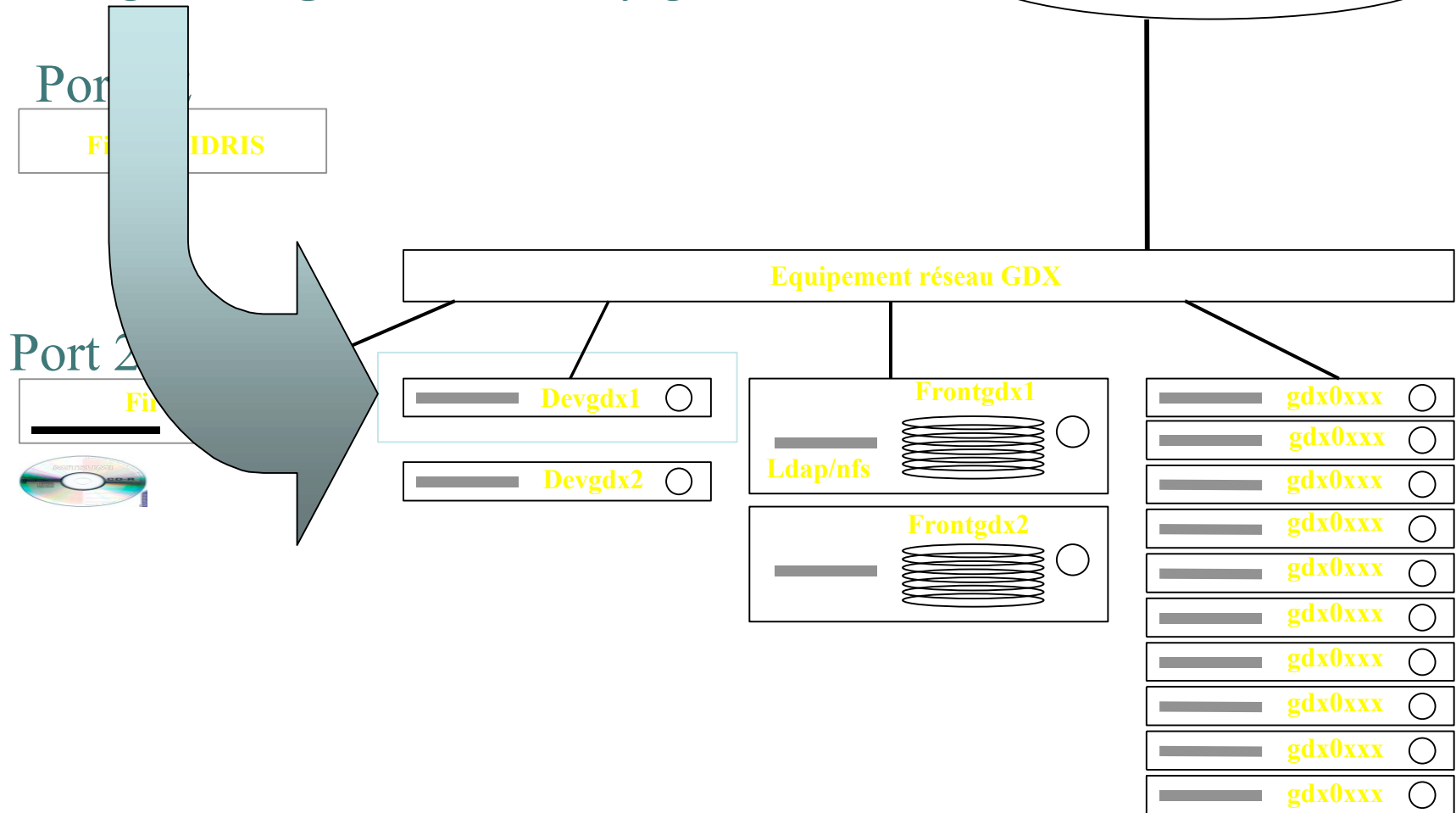
- Oar
- Kadeploy



Architecture de sécurité

« slogin bob@frontale.orsay.grid5000.fr »

VPN GRID5000



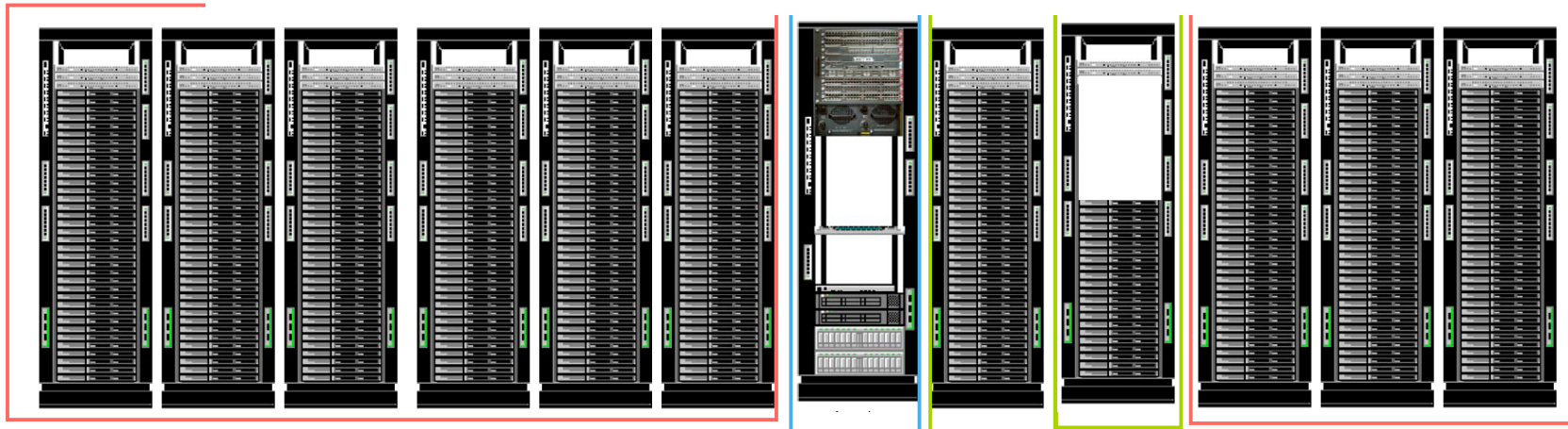
Architecture de sécurité

```
Terminal
File Edit View Terminal Go Help
aces*/users/archi/marty >ssh -X pmarty@frontale1.orsay.grid5000.fr
Password:
Last login: Wed Mar 16 21:33:27 2005 from netgdx032.orsay.grid5000.fr
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
! Bienvenue sur G5k-Orsay / GdX
!
! Vous etes sur frontale1 (devgdx001)
! => machine de reservation seulement
!
! Toutes les connexions sont journalisees
!
! infos => http://www.grid5000.fr/
!       => https://helpdesk.grid5000.fr/
!       => http://www.orsay.grid5000.fr/
! pb   => mailto:staff@orsay.grid5000.fr
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
=====
|
|          ATTENTION
|
| arret total de G5k-Orsay / GdX
| du 04 avril 2005 16h au 06 avril 08h
| (maintenance elec du batiment)
|
|=====
pmarty@devgdx001:~$
```

```
Terminal
File Edit View Terminal Go Help
aces*/users/archi/marty >ssh -X pmarty@frontale2.orsay.grid5000.fr
Password:
Last login: Wed Mar 16 19:41:42 2005 from netgdx032.orsay.grid5000.fr
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
! Bienvenue sur G5k-Orsay / GdX
!
! Vous etes sur frontale2 (devgdx002)
! => machine de reservation / compilation
!
! Toutes les connexions sont journalisees
!
! infos => http://www.grid5000.fr/
!       => https://helpdesk.grid5000.fr/
!       => http://www.orsay.grid5000.fr/ (pas encore en service)
! pb   => mailto:staff@orsay.grid5000.fr
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
=====
|
|          ATTENTION
|
| arret total de G5k-Orsay / GdX
| du 04 avril 2005 16h au 06 avril 08h
| (maintenance elec du batiment)
|
|=====
pmarty@devgdx002:~$ mozilla -remote &
```


GDXv2 : Evolution

2006 : 1040 processeurs (2GHz, 2Go)
64 nœuds routeurs (plusieurs interfaces)
Réseau haute-performance 10G (Myrinet/Ethernet/IB)



Nœuds de calcul (opteron bi-pro)

Frontal
Switch

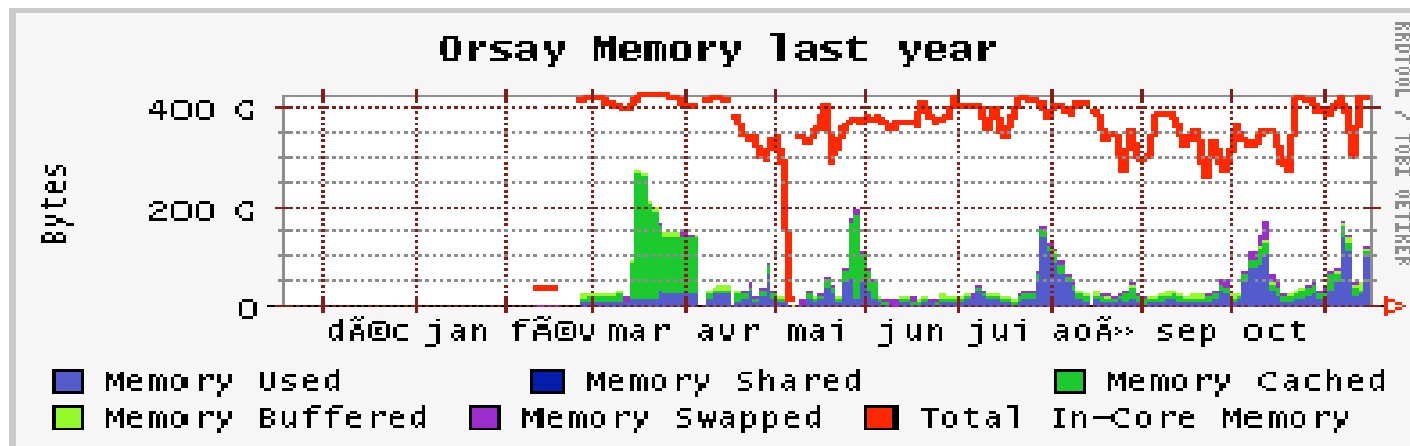
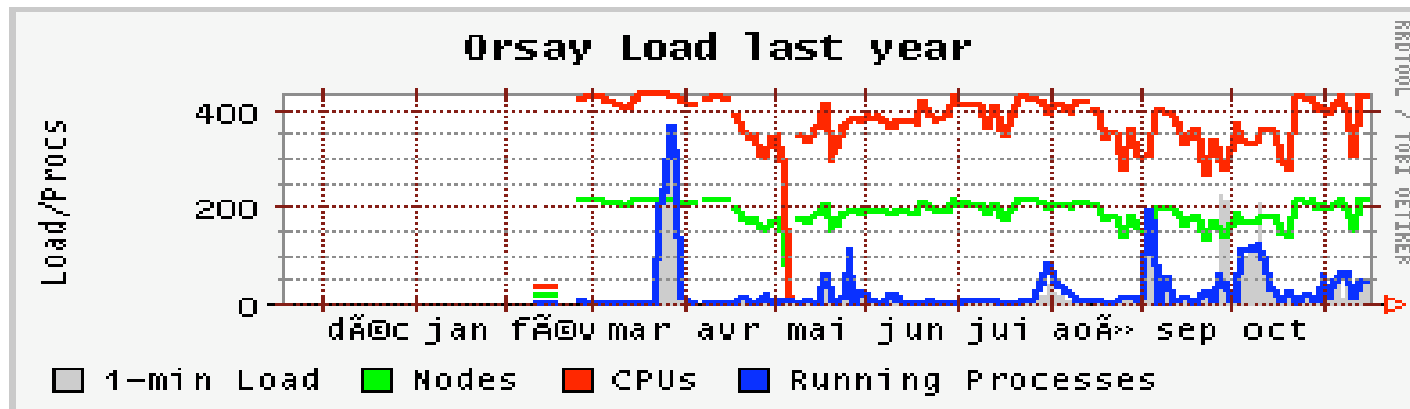
Routeurs
logiciels

Ministère (ACI Masses de Données)
INRIA Futurs / Rocquencourt

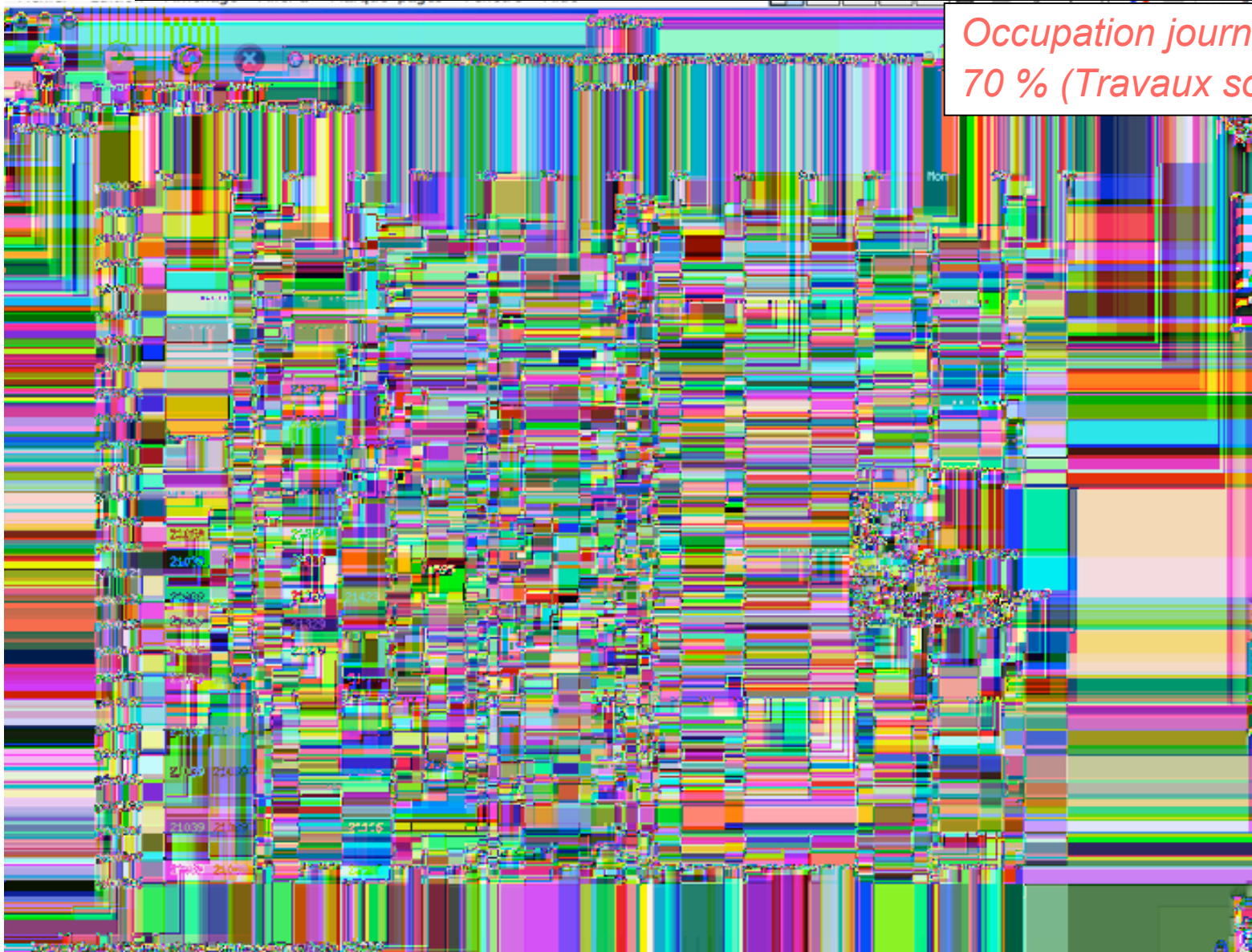
Région Ile de France
Ministère (ACI Grid'5000)

Utilisation GDX en 2005

Overview of Orsay

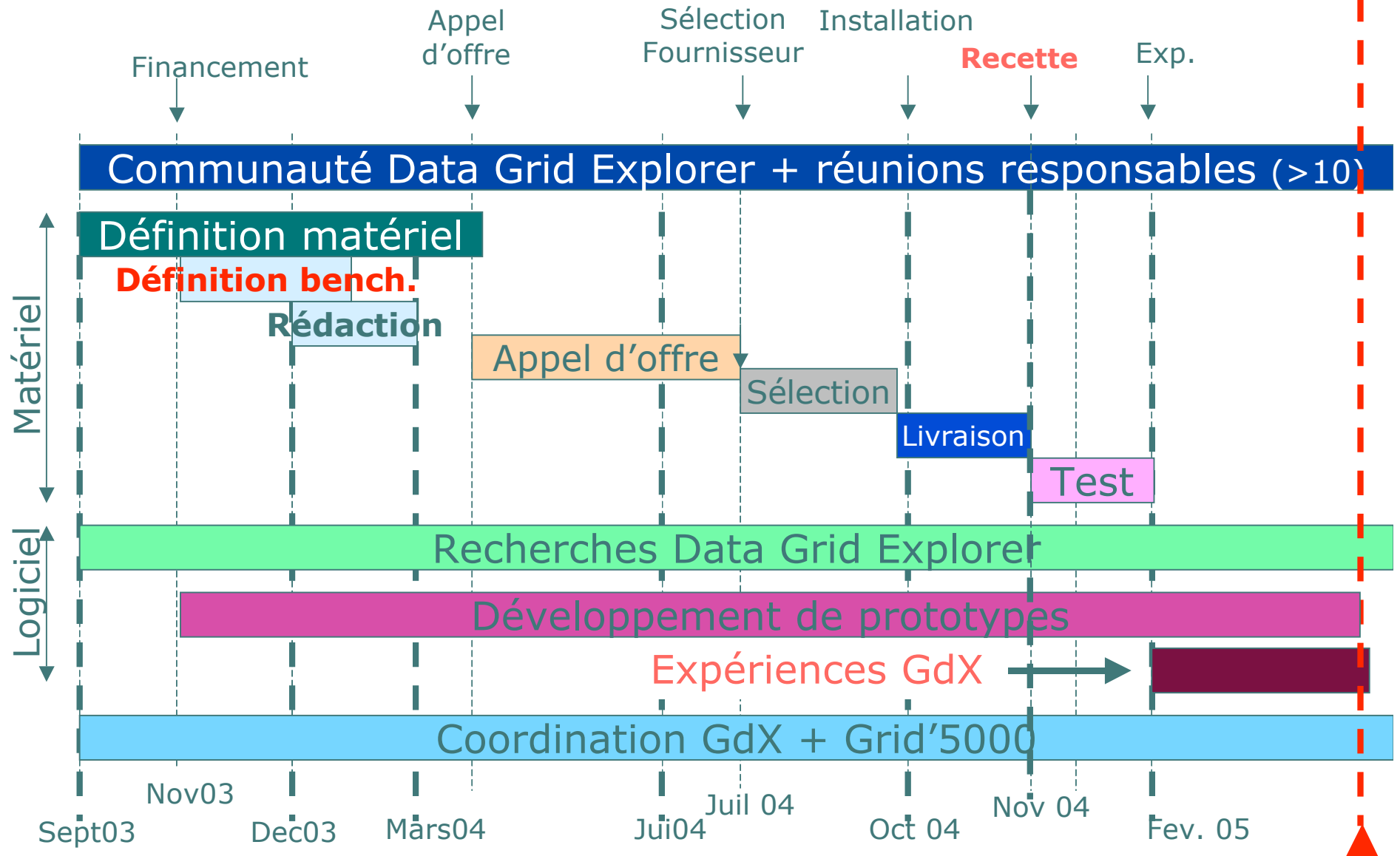


Utilisation de GDX (2)



GdX : Planning

Aujourd'hui





GDX : Projets scientifiques

1) **Construire l'instrument:**

- Cluster de 1K CPU
- Un réseau configurable (Ethernet, Myrinet/Infiniband)
- Un OS configurable (noyau, distribution, etc.)
- Un ensemble d'outils d'émulation
- Multi-utilisateurs

2) **Etudier l'impact de l'échelle sur les systèmes Grilles/P2P**

- Etudier des problèmes clés liés aux traitements données :
Extensibilité, Tolérance aux fautes, Ordonnancement, etc.
- Etudier des problèmes clés liés à la circulation des données:
Protocoles de transport haute performance, Partage de données, Stockage P2P, indexation répartie, etc.
- Applications
Simulation numérique, Bioinformatique, etc.

GDX : Projets de recherches

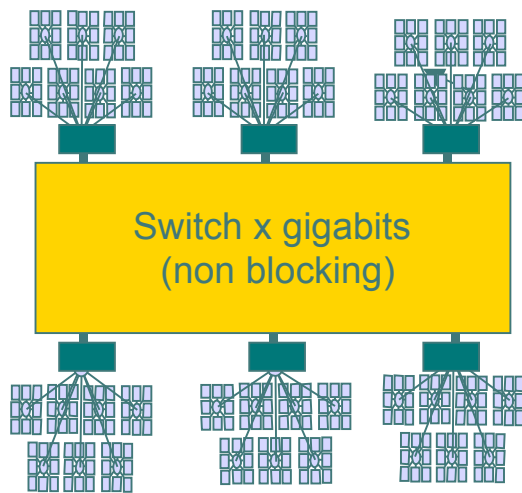
Experiences	Infrastructure	Emulation	Réseau	Application
I.1 Plate-forme	X	X	X	X
I.2 Virtual Grid		X	X	
I.3 Virt. Techniques	X		X	
I.4 Simul. Guidée par Emul.		X		
I.5 Emul. réseau	X	X	X	
I.6 Emul hétérogénéité		X		
I.7 Communication				X
I.8 Emul. Internet	X	X	X	
II.1 Engineering tech.		X	X	X
II.2 Objets Mobiles	X	X		
II.3 Tolérance aux pannes		X	X	
II.4 DHT		X		
II.5 Base de Données	X			X
II.6 Ordonnancement		X		X
II.7 Optimisation des comms.		X		
II.8 Partage de données		X		X
II.9 Uni et multicast		X	X	
II.10 Automate cellulaire		X		X
II.11 Bioinformatique				X
II.12 Stockage P2P			X	X
II.13 Adversaires		X	X	X
II.14 Sécurité		X	X	X
II.15 Internet nouvelle génér.	X	X	X	
II.16 Simulation numérique				X



Projets émulation (I)

- EWAN: émulation réseau très haut débit
- VGRID: émulation de processeurs
- H-GRID: émulation hétérogénéité
- Emulation trafic

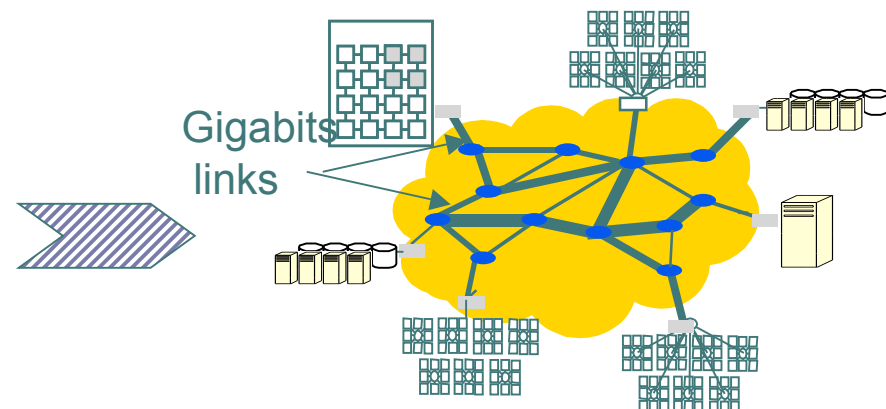
Projet eWAN – (LIP - RESO)



Etapes

1. Définition de la configuration
2. Génération des scripts et tables
3. Déploiement automatique
4. Lancement des expériences
5. Analyse des traces

Points durs: -> gigabit + longue latence
-> calibrage



Entrées

- Topologie, Délais
- QoS, limitation débit
- IPv6 ou v4

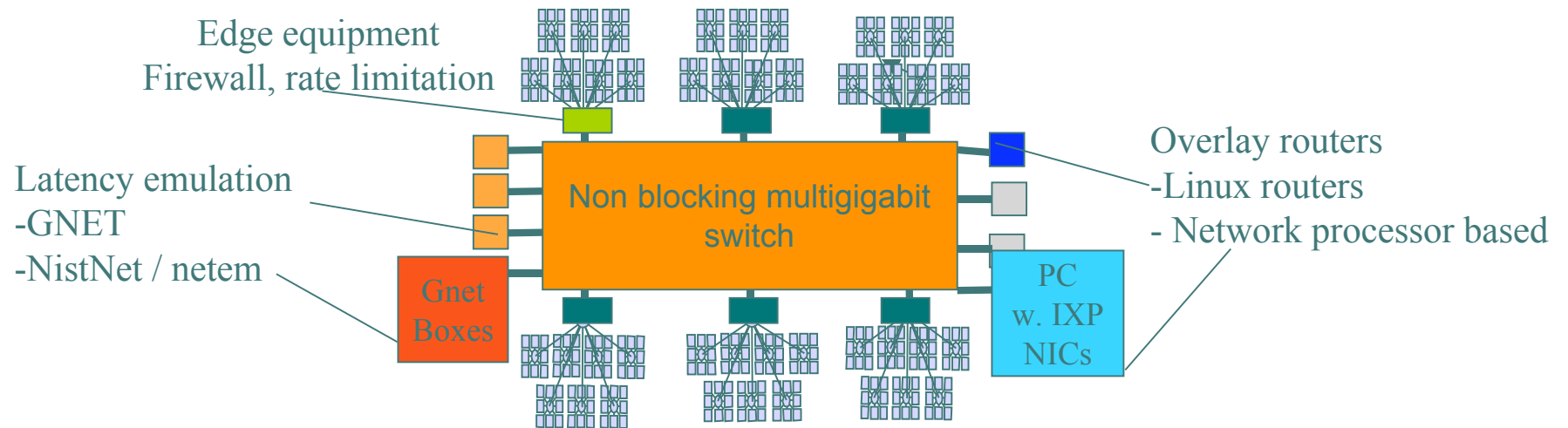
Expériences

- Sources: vrais flux
- Vraies applications ou squelettes.
- Trafic concurrent contrôlé.

Sorties

- Performance et comportement
- Influence des paramètres d'échelle
- Taux d'utilisation ressources

eWAN: infrastructure



Deployment

Scripts generation

Machine 0 IP : 140.77.12.61 emulate : Client c0
IP0: 192.168.4.2

Machine 1 IP : 140.77.12.62 emulate : la1
IP0: 192.168.3.2
IP1: 192.168.5.2

```
ip address flush label eth*;
ifconfig eth0 140.77.12.62 netmask 255.255.255.0;
route add default netmask 0.0.0.0 gw 140.77.12.1 dev eth0;
ifconfig eth0:0 mtu 1500 192.168.3.2;
ifconfig eth1 mtu 1500 192.168.5.2;
if [ -z "`cnistnet -Fd 2>/dev/stdout| grep command`" ];
then modprobe -r nistnet;
modprobe nistnet;
cnistnet -u;
cnistnet -a 0.0.0.0 0.0.0.0 --delay 5 --drop 5 > /dev/null;
else echo NIST Net not available;
fi;
route add -net 192.168.0.0 netmask 255.255.0.0 gw 192.168.3.1 dev eth0;
route add -net 192.168.6.0 gw 192.168.5.1 netmask 255.255.255.0 dev eth1;
```

Machine 4 IP : 140.77.12.65 emulate : Router rc0
IP0: 192.168.1.1
IP1: 192.168.2.1
IP2: 192.168.3.1

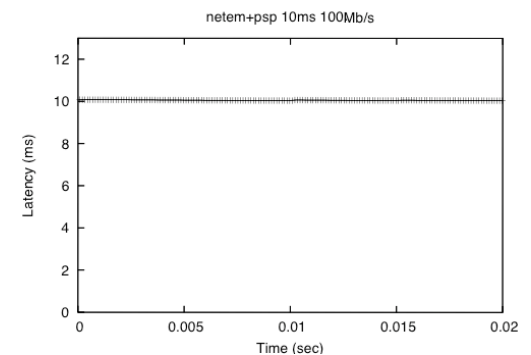
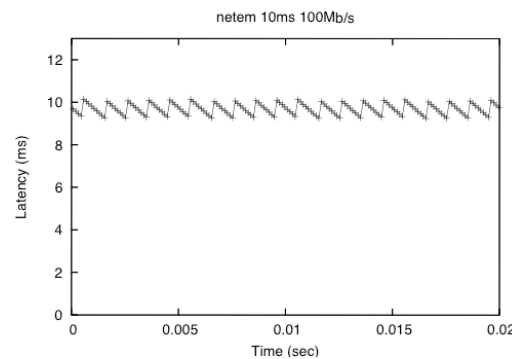
Machine 5 IP : 140.77.12.66 emulate : Access Point p0
IP0: 192.168.1.2
IP1: 192.168.4.1

```
ip address flush label eth*;
ifconfig eth0 140.77.12.66 netmask 255.255.255.0;
route add default netmask 0.0.0.0 gw 140.77.12.1 dev eth0;
ifconfig eth0:0 mtu 1500 192.168.1.2;
ifconfig eth1 mtu 1500 192.168.4.1;
tc qdisc replace dev eth1 root tbf rate 100mbit latency 1ms burst 15400000;
route add -net 192.168.0.0 netmask 255.255.0.0 gw 192.168.1.1 dev eth0;
```

Machine 6 IP : 140.77.12.67 emulate : Access Point p1
IP0: 192.168.5.1
IP1: 192.168.6.1

Precision & passage à l'échelle

- Emulation logicielle: netem: large déploiement
- Emulation matérielle : GNET(FPGA): haute précision
- eWAN utilise module netem linux pour deploiement large échelle + intègre émulateur matériel GNET1/GNET10
 - gestion des buffers pour l'émulation de délai rend le trafic très sporadique (high burstiness).
 - Couplage avec logiciel de pacing AIST PSP pour contrôler la sporadicité dans l'émulateur





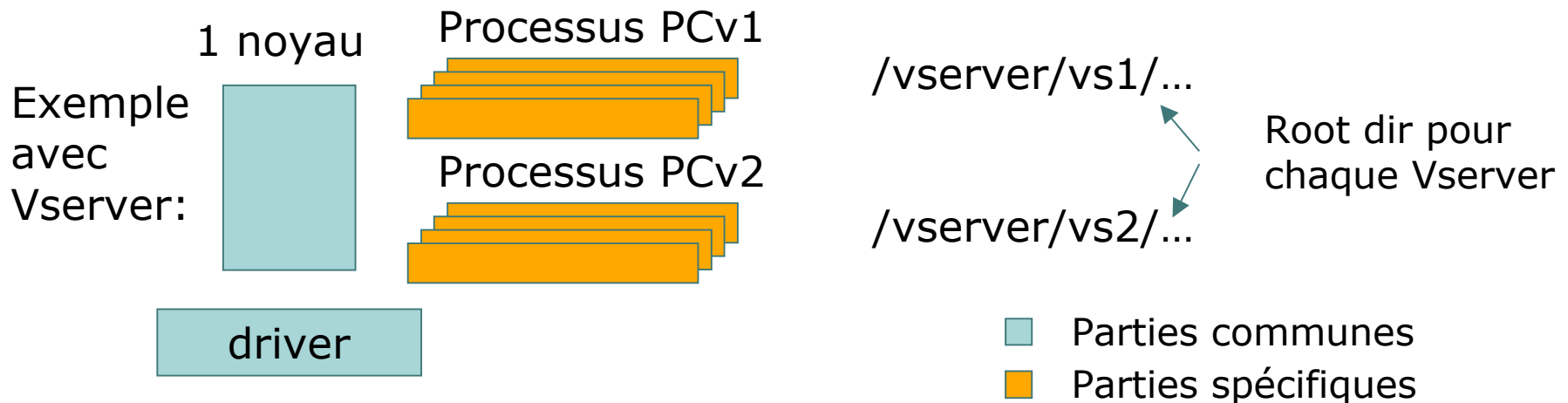
Emulation trafic Internet (LAAS)

- Conception (à l'aide du simulateur NS) d'une méthode de **reproduction des conditions réelles du trafic**
→ méthode et résultats publiés à ICC'2004, Paris
- Application de cette méthode à un émulateur
 - Mise en place de règles de configuration des émulateurs Dummynet
 - Développement d'une première version d'un logiciel de rejeu de traces de trafic réelles
- À venir:
 - Mise en œuvre sur GdX
 - intégration de cet outil dans un autre outil plus générique (DHS: développé au LAAS)
 - Recherches sur la génération de trafic réaliste à partir de la mesure et du calcul des premiers moments statistiques d'un trafic

Projet VGRID (LRI)

Emuler 100 PC virtuels sur 1 PC réel → 10 K PCv sur 100 CPUs (LRI),
100K PCv sur 1K CPUs (GdX), Non temps réel

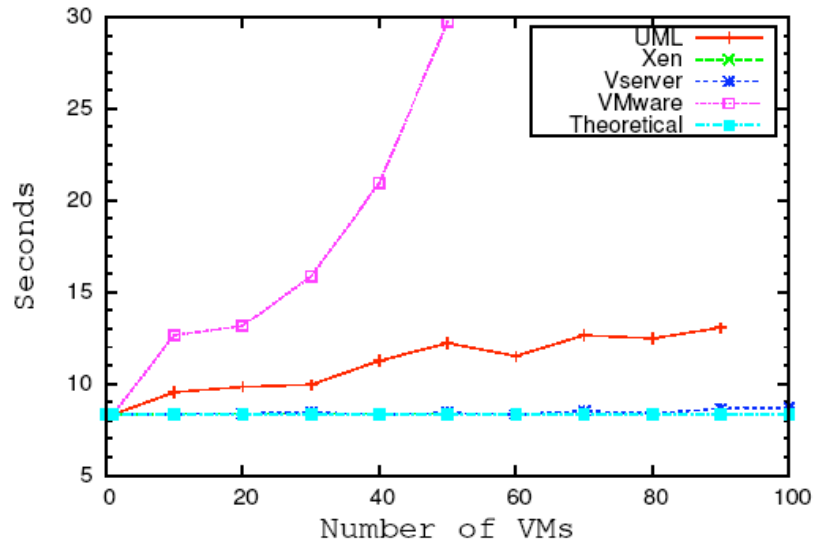
Etude de Vserver, Xen, Virtual PC, UML, VMware, Scheduler de noyaux,...



Première question scientifique : quelles métriques, quels benchmarks

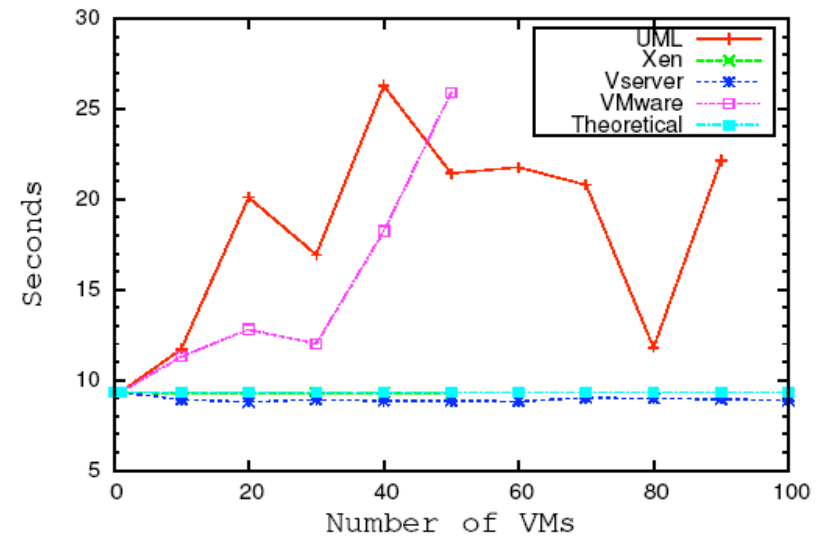
- Surcoût de l'émulation
- Equité entre les PCv (CPU, Mémoire Virtuelle, Disque, Réseau)
- Linéarité du ralentissement avec le nombre de PCv

Projet VGRID (LRI)



Overhead CPU / # VM

Overhead MEM / # VM



Projet H-GRID – (LORIA)

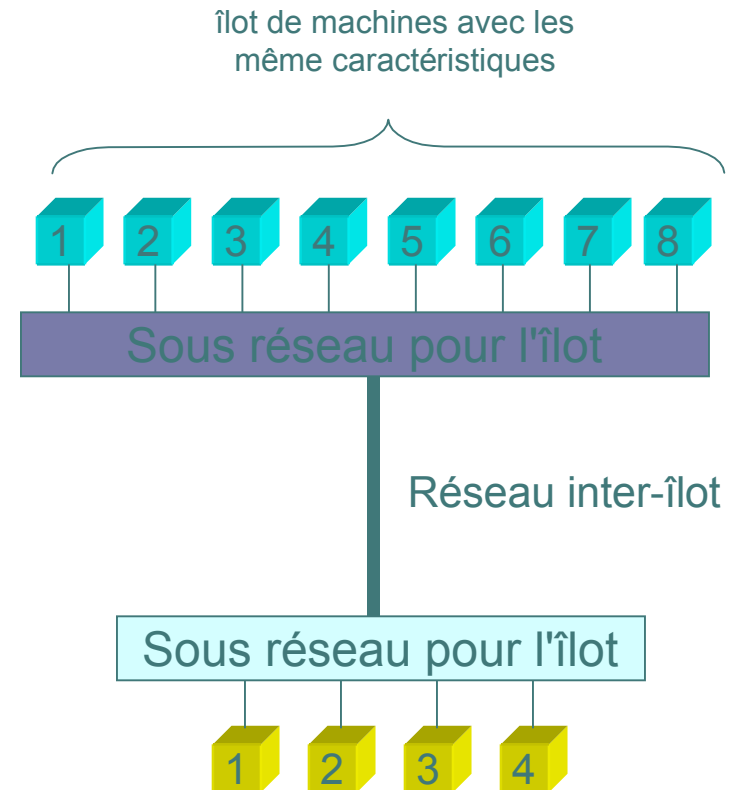
Objectif : rendre GdX hétérogène pour faire des expériences d'émulation

Moyen : dégrader les caractéristiques de la plate-forme :

- vitesse CPU,
- mémoire disponible ,
- latence réseau,
- bande passante réseau.

Mise en œuvre :

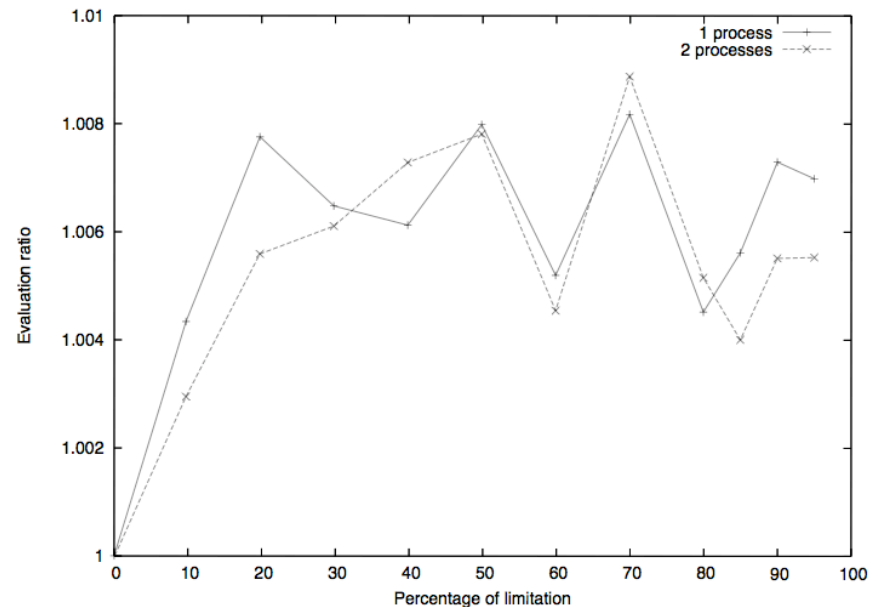
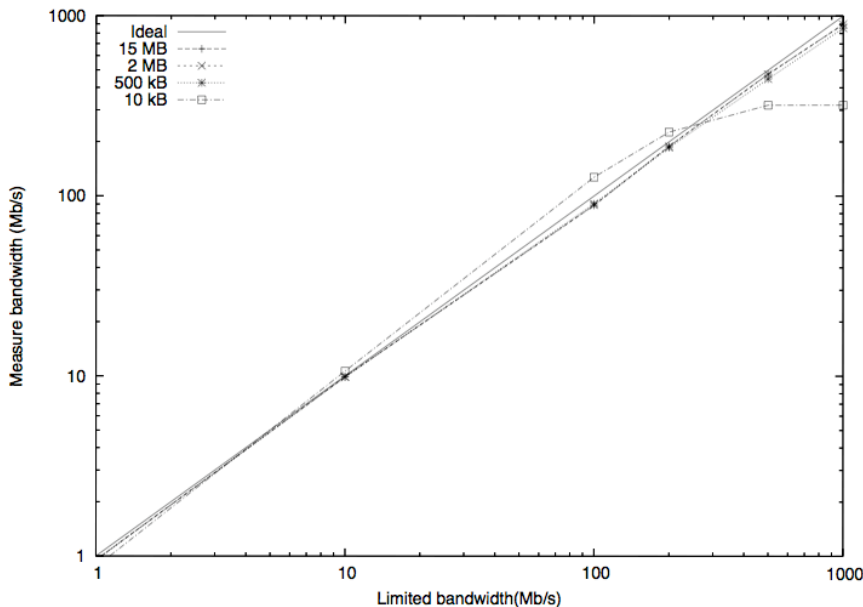
- configuration par **îlot** (union d'intervalle d'adresses IP),
- Définition des caractéristiques communes à chaque îlot,
- Définition des caractéristiques entre îlots.



```
ilot1 : [152.81.2.12-152.81.2.25]-[152.81.2.151-152.81.2.176]{  
SEED : 1 % -1 pour une graine aléatoire  
CPU : [800-1400] %chaque membre de l'îlot aura une vitesse CPU choisie uniformément entre 800 MHz et 1,4 GHz  
BPOUT : [1000;200] % Chaque membre de l'îlot aura une bande-passante en sortie choisie suivant une  
           % gaussienne de moyenne 1000 Ko/s et de variance 200 Ko/s  
...  
}  
!INTER : [ilot1;ilot2] [200-200] [300-300] [1;0] %caractéristiques réseau entre ilot1 et ilot2
```

Projet H-GRID – (LORIA)

Résultats de micro-benchmarks sur GdX



Tests de bande passante en fonction de la taille des données.
 Bande passante mesurée en fonction de la bande passante demandée.
 Pour 10 Ko le réseau n'est pas capable d'atteindre la bande passante crête (comme dans la réalité).

Tests de puissance CPU en fonction du pourcentage de dégradation demandé.
 Evaluation ratio : rapport entre temps mesuré et temps théorique (rapport=1 => émulation parfaite).
 Résultats : précision de l'émulation entre 1 et 0.4 %

Set latency	1	5	10	50	100	500	1000
RTT	2.12	10.05	20.12	100.06	200.2	1000.05	1999.75

Round trip time en fonction de la latence désirée (en ms)
 On voit que $RTT=2 \times \text{latence}$ car celle-ci est payée deux fois dans un aller-retour.



Etudes sur GDX

- GDS
- Grid Failure detectors
- Volatilités dans stockage P2P
- Automates cellulaires



II.3 Gestion des fautes (LIP6)

- Passage à l'échelle des détecteurs de fautes
 - Hiérarchie => gestion d'un grand nombre de fautes
 - Adaptation automatique des délais de surveillance => adaptation à la dynamique du réseau
 - Publications : [DSN 2003]
- Verrouillage tolérant les fautes sur Grille
 - Adaptation des algorithmes de verrouillage aux grilles
 - Algorithmes à jeton tolérant les fautes
 - Publications : [CCGrid 04], Rapport de recherche, JPDC
- Réalisations
 - Logiciels : Détecteurs de fautes, simulateur de système à large échelle, injecteur de fautes

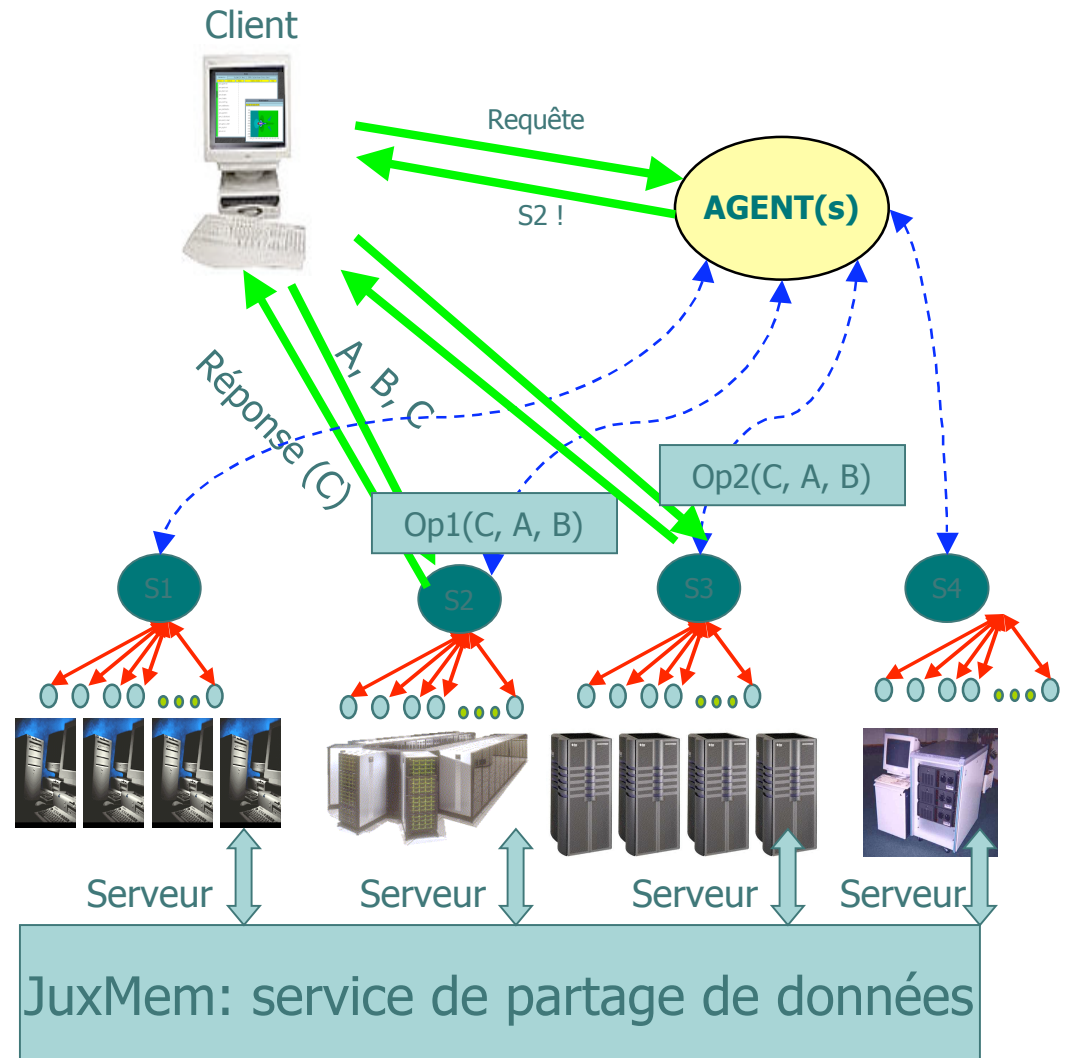


II.3 Volatilité dans stockage P2P (LIP6)

- Objectifs : Etude de l'impact du taux de volatilité dans les DHT (Distributed Hash Table)
- Cadre : le système de fichiers Pastis
- Experiences :
 - déploiement d'un reseau virtuel de 1000 nœuds sur 150 processeurs
 - Utilisation des routeurs logiciels Modelnets
 - Injection de différents modèles de volatilité

II.8 Partage de données (ACI MD GDS)

- Projet ACI MD GDS, partenaire du projet GdX
 - PARIS, GRAAL, REGAL
- Service de partage de données
 - Persistance
 - Localisation transparente
 - Cohérence des données
 - Architecture dynamique
 - Tolérance aux fautes
- Implementation
 - Multi-protocoles (réplication, cohérence)
 - Mise en œuvre sur JXTA 2.0 (Sun Microsystems)



II.8 JuxMem sur GdX

- Buts : tester le passage à l'échelle de JuxMem à travers l'utilisation d'une application DIET (GridTLSE)
 - Evaluation de l'algorithme d'allocation mémoire de JuxMem
 - Tests d'extensibilité des protocoles de cohérence tolérants aux fautes
 - Performances ? Taux de volatilité supporté ?
- Utilisation prévue de GdX
 - Nombre de nœuds visés : 1 000
 - Nombre de pairs visés : 100 000 (100 par machine)
 - Première évaluation poussée du passage à l'échelle de JXTA
 - Emulation de la plate-forme Grid'5000 sur GdX

II.10 Cellular Automata on Grid

A programming model for R&D of fine grained cellular systems:

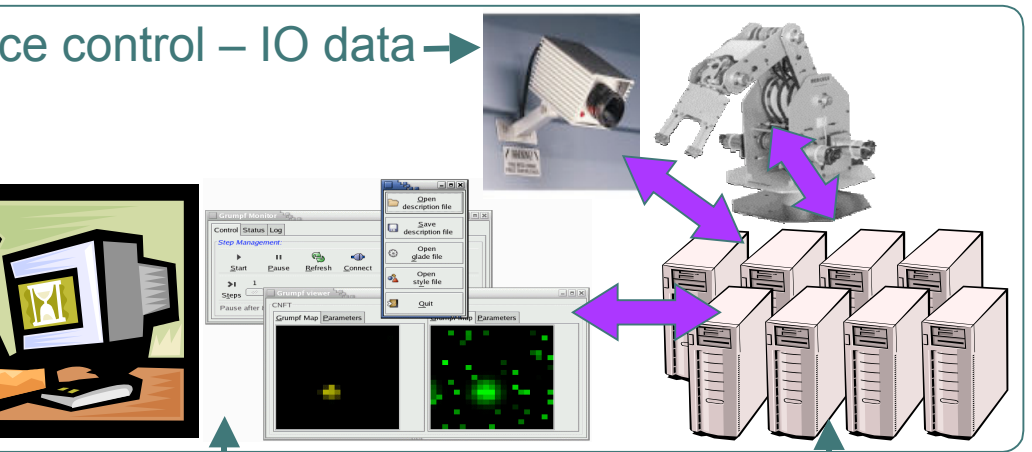
- CPU consuming
- frequent prototype update

_ Minimize ($T_{dev} + \sum T_{exec}$)

Main features:

- Cellular intensive computing steps: considers distributed memory & optimized communications.
- Sequential interactive slow steps: requires virtual shared memory.

2nd prototype under development



Client machine: **comfortable**
fine grained development env.

Distributed server:
coarse grained arch.

- large cluster
- cluster of clusters
- cluster of clusters across a WAN

Track efficient architectures



Nov 2005
July 2006

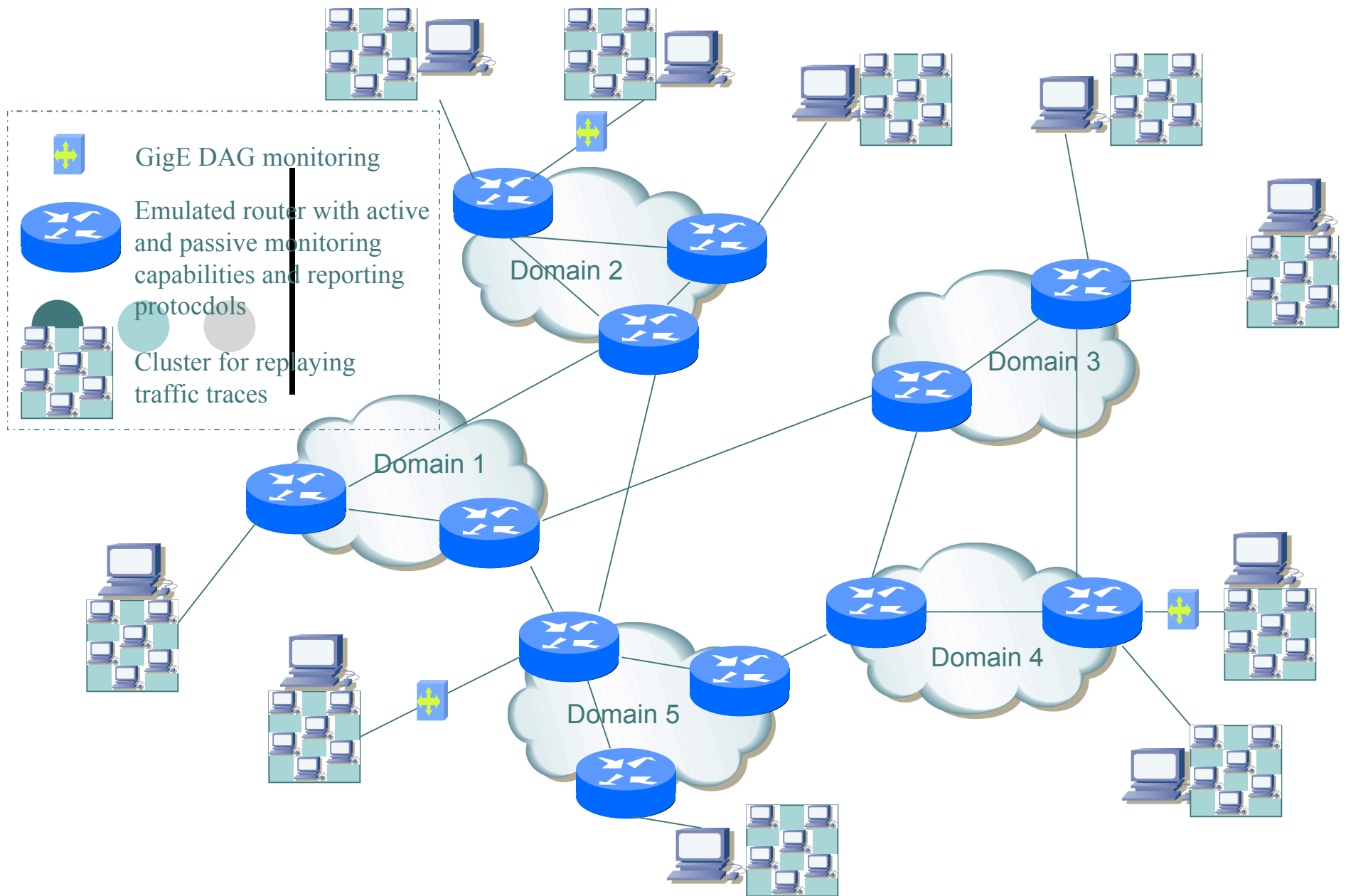
Collaboration Supélec – Loria – Potenza University:
(S. Vialle, J. Gustedt, A. De Vivo)





II.15 Internet Nouvelle Génération (LAAS)

- 3 expérimentations prêtes
 - QoS multi-domaine dans l'Internet avec DiffServ
 - Gestion des réseaux de l'Internet à partir de mesures: approche MBA (Measurement Based Architecture) / MSP (Measurement Signaling Protocol) → Expérience sur le contrôle de congestion avec MBCC (Measurement Based Congestion Control)
 - Coordination des activités dans des applications collaboratives distribuées (Middleware au dessus de Corba)
- Expérimentations menées pour l'instant sur une plate-forme de 10 machines en local → En attente de GdX



Large scale multi-domains Internet experiments for validating measurement based Network protocols and architectures

<http://research.microsoft.com/barc/SortBenchmark/>

	Daytona	Indy
Penny	(new 2005) 15 GB (163 M records) PostmanSort, doc pdf 979 sec on a \$951 Wintel 2 SATA Robert Ramey	(2002) 40 GB (433 M records) SheenSort.pdf 1541 seconds on a 614\$ Linux/AMD system Lei Yang, Hui Huang Zheng Wan, Tao Song Tsinghua University, Beijing, China
Minute	(2004) 32 GB (340 million records) Nsort pdf, word, htm Windows, 32 Itanium2, 2,350 SAN disks Chris Nyberg, Charles Koester Ordinal Technology	116GB (125 M records) 58.7 seconds Linux, 80 Itanium2, 2,520 SAN disks Jim Wyllie
TeraByte	(2004) 33 minutes Nsort pdf, word, htm Windows, 32 Itanium2, 2,350 SAN disks Chris Nyberg, Charles Koester Ordinal Technology	(new 2005) 435 seconds (7.25 minutes) SCS pdf Linux, 80 Itanium2, 2,520 SAN disks Jim Wyllie, IBM Almaden Research



**Need a Single
Image File
System
(Kerrighed,
GPFS...)**

```
ccerin@gdx0001:~/TRI$ mpirun -np 192 a.out size=1160000000 verbose=0 ifName=/tmp/titi ofName=/tmp/toto
We Generate data first, and we needed 170.063589 secs...then to store them on disk!
We start sorting NOW!
0.782573 + 10.891525 + 66.524312 + 32.727524 + 35.484297 = 146.583187 secs to sort 116000000000 bytes
on 192 procs
PID=0 has sorted: 5458892 records of 100 bytes
PID=16 has sorted: 5807629 records of 100 bytes
PID=2 has sorted: 6832457 records of 100 bytes
PID=8 has sorted: 4927854 records of 100 bytes
PID=1 has sorted: 5824897 records of 100 bytes
```

**Partitionning +
Communication time**



Conclusions

- Les recherches en Grille/P2P nécessitent des plates-formes à grande échelle
 - Pour étudier les questions liées au mouvement, stockage et calcul sur les **données** (les protocoles, systèmes, intergiciels, langages et modèles de programmation et les applications)
 - Avec des conditions expérimentales reproductibles
- Data Grid eXplorer
 - Une plate-forme expérimentale pour les chercheurs en Grille/P2P
 - Un émulateur de système à grande échelle
 - Relation étroite avec le projet Grid'5000 (plusieurs chercheurs participent aux deux projets)
- Beaucoup d'expériences en cours
 - Ecriture d'articles en cours



Questions ?



Calendrier des réunions

Réunions plénières

Réunion thématiques (réseau, émulation,
application, architecture)

Réunions téléphoniques entre les 5 coordinateurs

Périodicité plus forte au début de projets (pour fixer
les grands choix d'architectures)