Une introduction à la Problématique Grilles/P2P

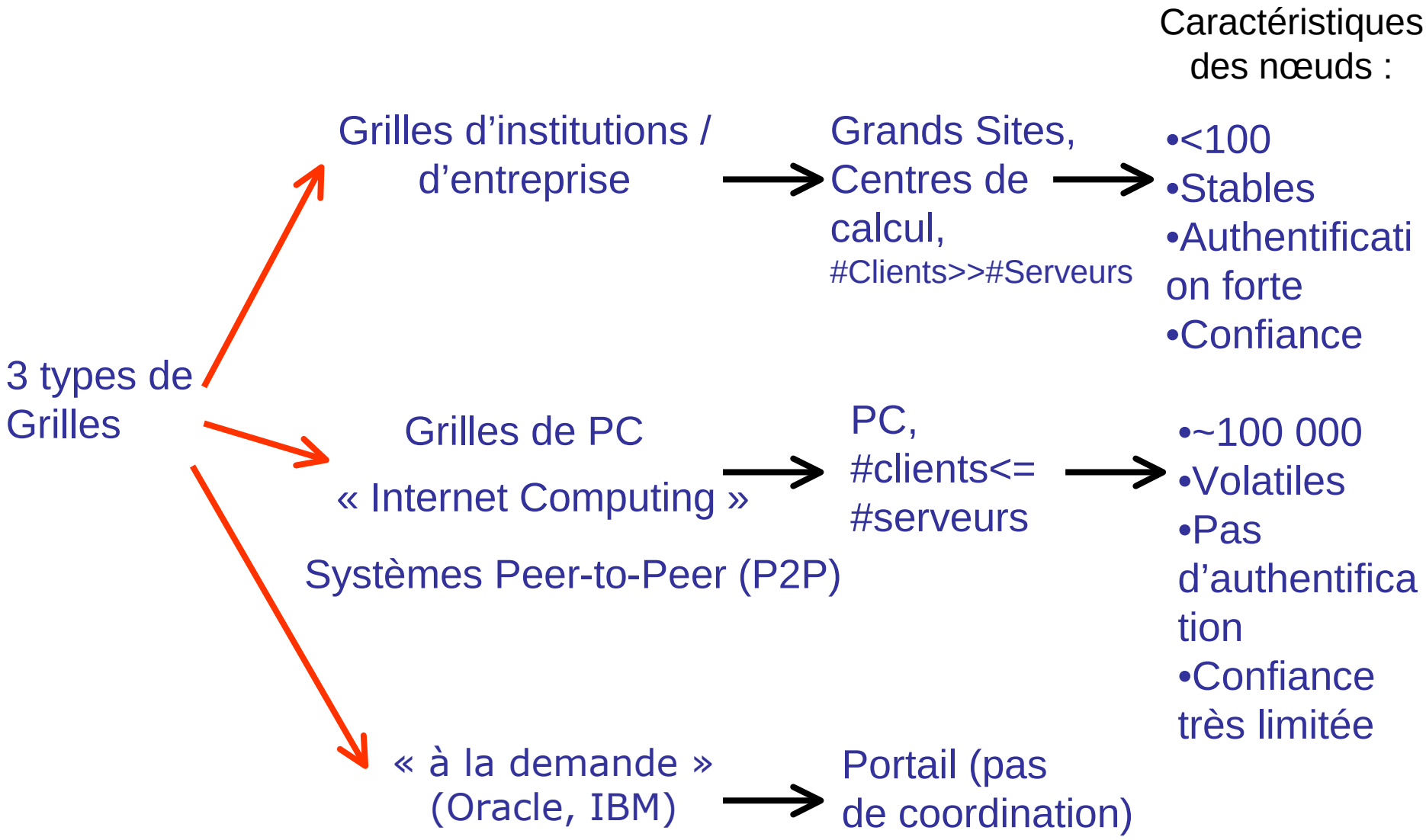
Franck Cappello
INRIA
LRI, Université Paris Sud
fci@lri.fr

Grille : analogie avec la distribution de l'électricité

Délivrer des capacités de calcul/stockage/communication
de façon transparente à l'utilisateur, quand et où il les demande

Plus pragmatique : construire des communautés dynamiques
et fournir à leurs membres l'accès aux ressources partagées

Différents types de Grilles



Grille d'entreprise != grille initiale

"Three Obstacles to Making Grid Computing Routine"

1) New approaches to problem solving
   - Data Grids, distributed computing, peer-to-peer, collaboration grids, …
2) Structuring and writing programs
   - Abstractions, tools

Programming Problem

3) Enabling resource sharing across distinct institutions
   - Resource discovery, access, reservation, allocation; authentication, authorization, policy; communication; fault detection and notification; …

System Problem

Les problèmes soulevés par les Grilles/P2P

Les Grilles et systèmes P2P sont des systèmes distribués !

Mais :

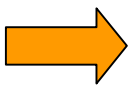
Leur échelle, leur dynamisme et la présence de très nombreux domaines d'administration renouvelle la problématique

Mécanismes :

- Déploiement,
- Communication,
- Stockage de données,
- Calcul,
- Coordination,
- Programmation.

Propriétés recherchées :

- Sécurité,
- Performance,
- Tolérance aux pannes,
- Extensibilité,
- Equilibrage de charge,
- Certification des données,
- Equité,



Puisque l'infrastructure matérielle est donnée, la recherche est essentiellement logicielle

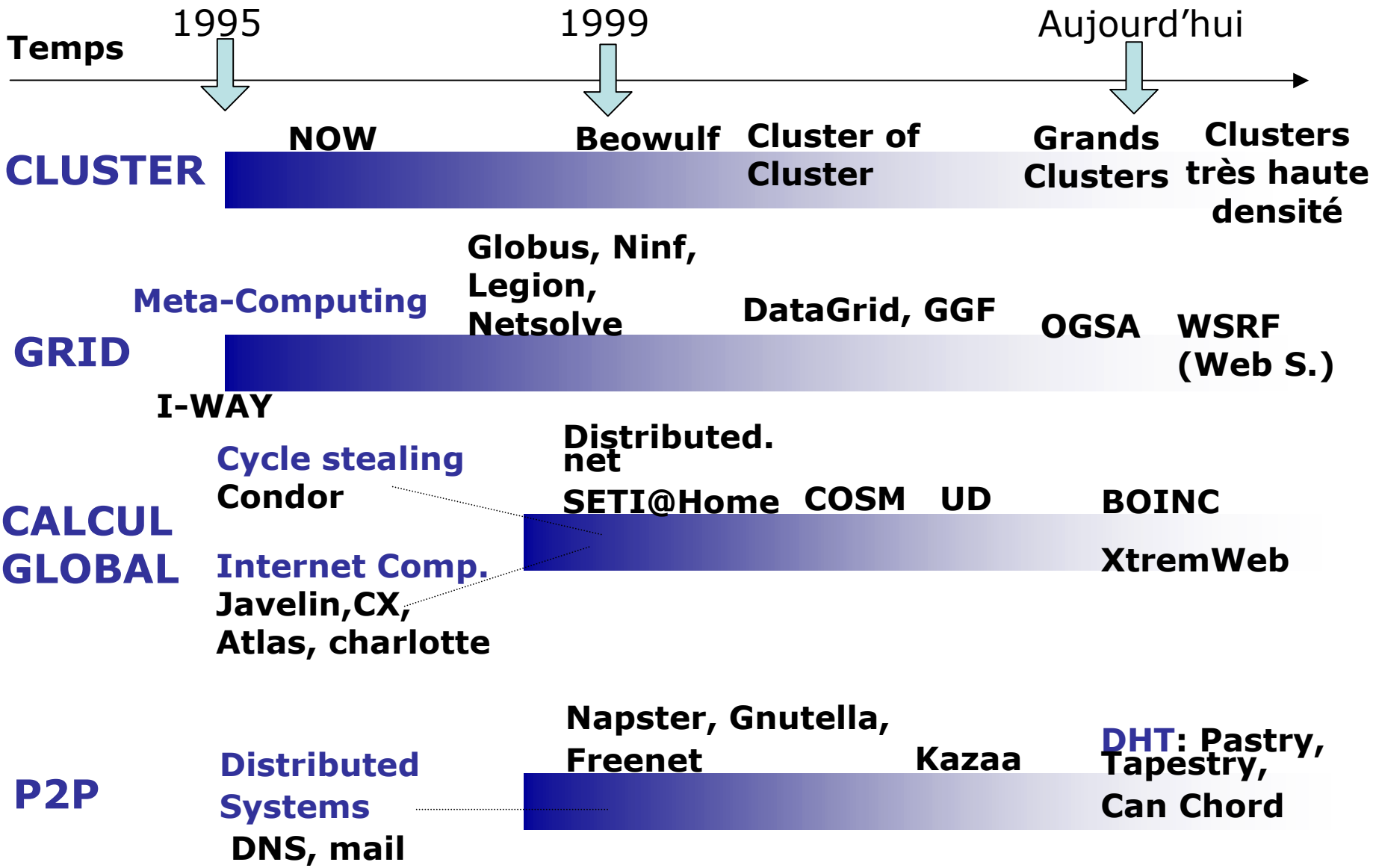
Données Massives et Grilles/P2P

Problèmes typiques liés aux données (projets ACI Masse de données) :

Masse de Documents Pair à Pair, Grid Data Service, Pair à Pair, Data Grid eXplorer

- Gestion des données
  - Stockage, indexation, recherche, communication et partage de données (protocole P2P, algorithmique de Graphe)
- Tolérance aux pannes / extensibilité
  - Volatilité à grande échelle (persistance, intégrité, graphes dynamiques)
  - Asynchronisme de l'Internet (résultat d'impossibilité)
- Performance
  - Accès rapide (protocoles, réplication, cohérence), équilibrage de charge (modèle de charge/trafic),
  - Minimisation de la consommation de ressources liée au contrôle du système (sécurité, maintient des Metadata, etc.)

Perspective historique



L'échelle de ces systèmes distribués augmente significativement avec le temps

Data Grid eXplorer

Plate-forme expérimentale
mutualisée à l'échelle nationale
ACI Masse de Données

Franck Cappello
INRIA
fci@lri.fr

Avec tous les membres participants

Référent ACI MD : Luc Bougé

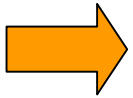
Plate-forme expérimentale Grilles/P2P
Motivations

Les sys. Grilles/P2P posent un très grand nombre questions:

Sécurité, Performance, Tolérance aux pannes, Extensibilité, Equilibrage de charge, Coordination, Passage de messages, Stockage, Programmation, Algorithmes, Protocoles de communication, Déploiement, etc.

Modèles théoriques et simulateurs ne permettent pas de capturer des conditions réalistes (codes réelles, OS, etc.)

Les plates-formes de production ont des difficultés importantes à reproduire les conditions expérimentales



Comment comparer ?

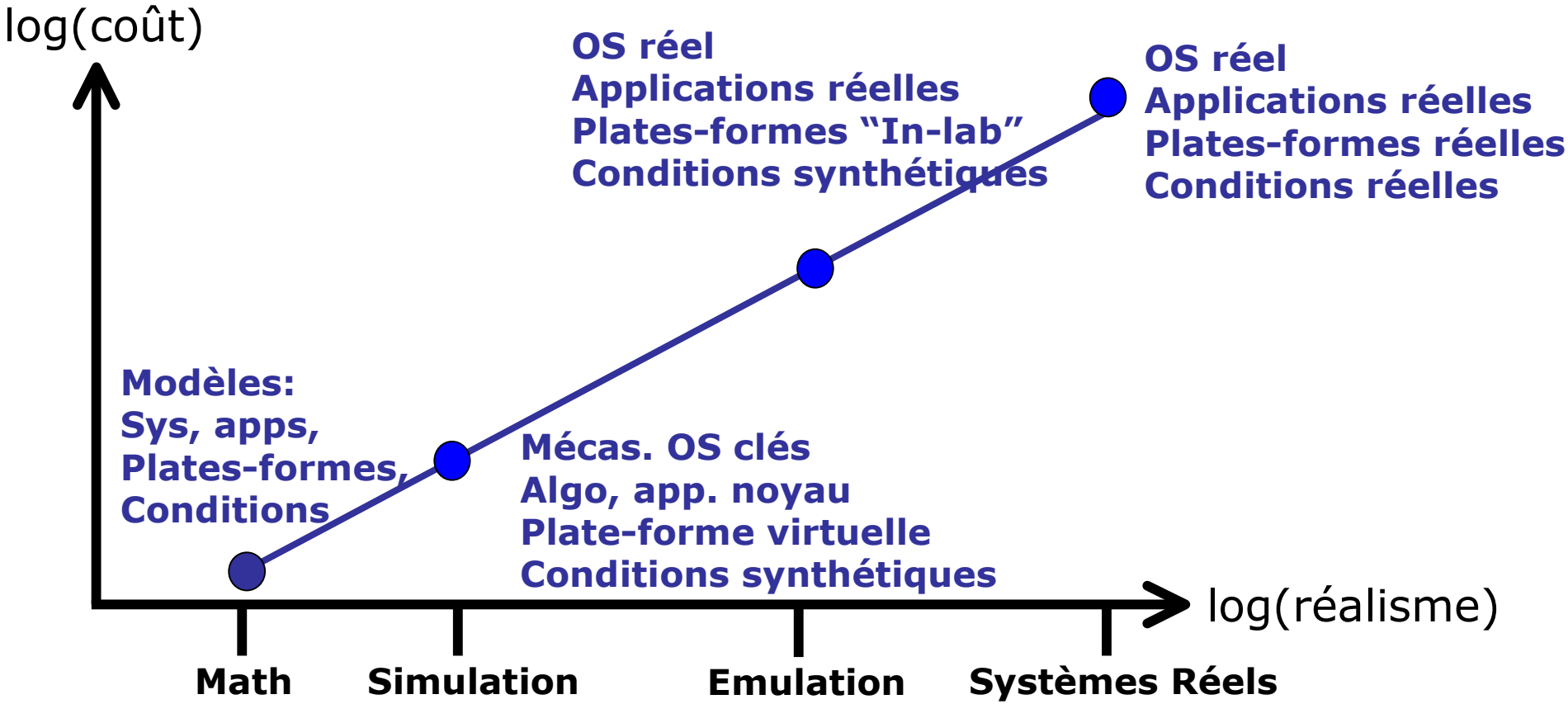
- Protocoles/techniques de tolérance aux pannes
- Mécanismes de sécurité
- Protocoles réseau
- etc.

A large échelle !!

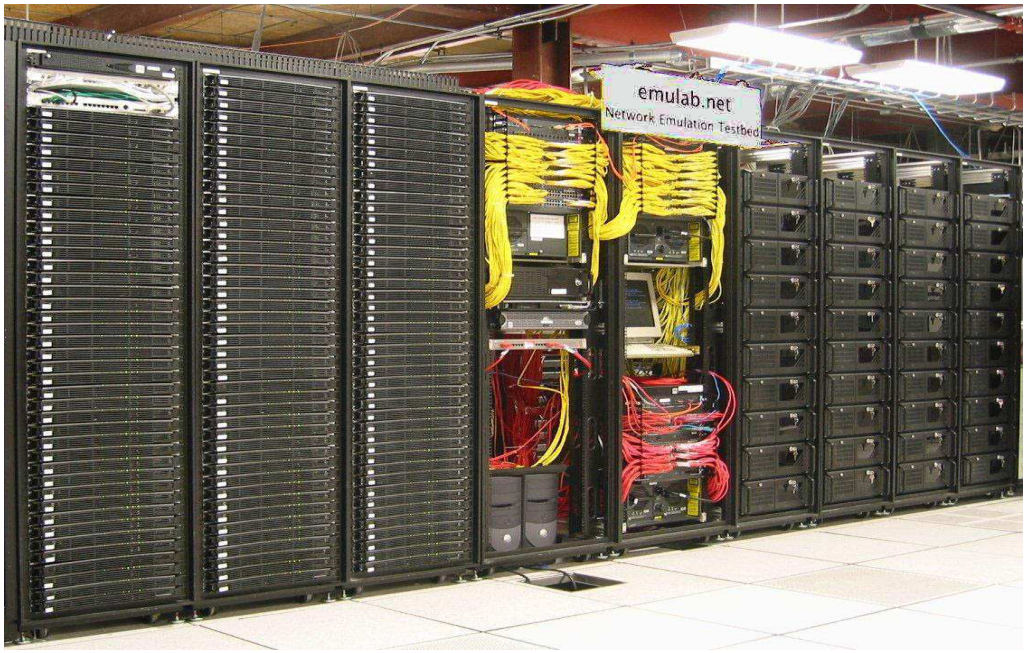
Outils pour l'étude des systèmes distribués

Pour étudier les systèmes distribués à grande échelle :

1) Outils (modèles, simulateurs, émulateurs, plates-formes)
2) Interactions fortes entre les outils de recherche



- Comme « Modelnet »
  - 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 exterieurs client (40 DSL)

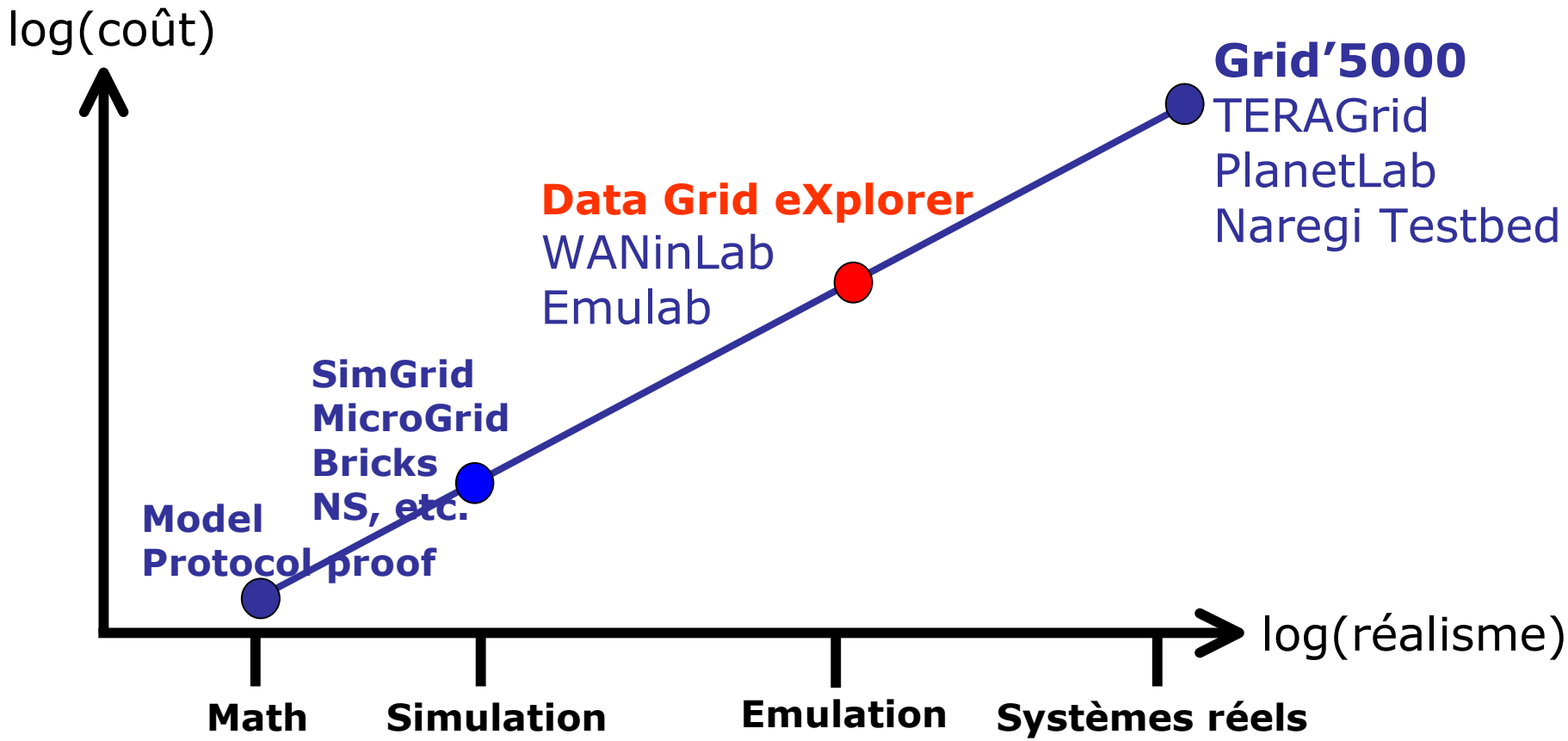


- 3 Sites

Outils pour l'étude des syst. Grilles/P2P

Au moment du montage du projet:
Il n'y a pas de plate-forme dédiée à l'étude des Grilles

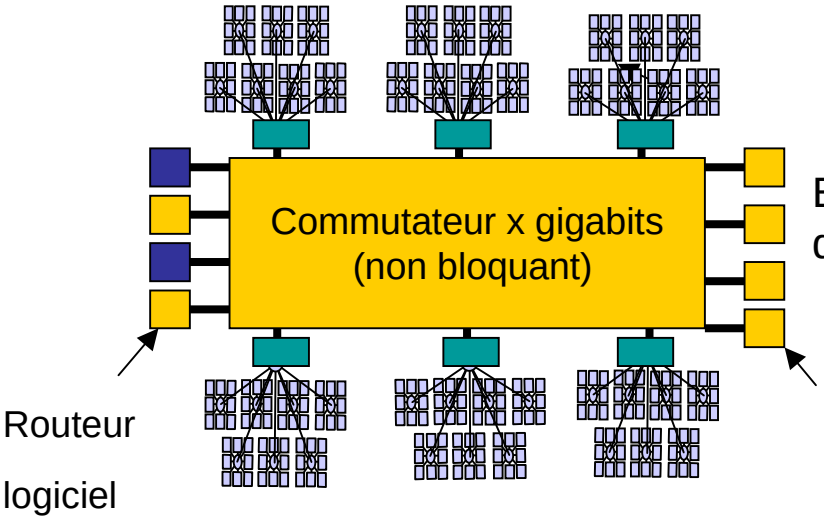
→ Data Grid eXplorer : un émulateur à grande échelle



Principe de l'émulation à large échelle

Systèmes à Grande échelle:

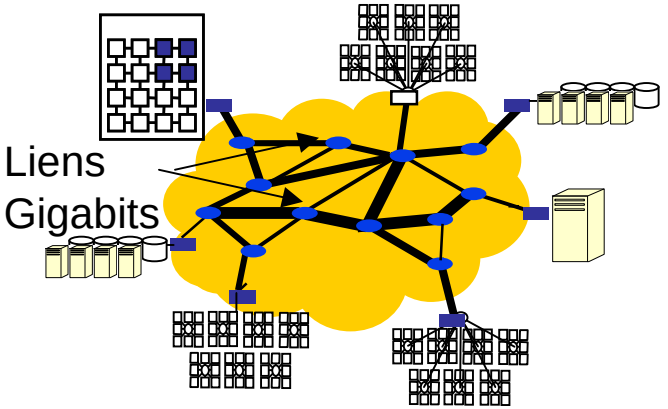
Cluster
Data Grid eXplorer



Virtualisation

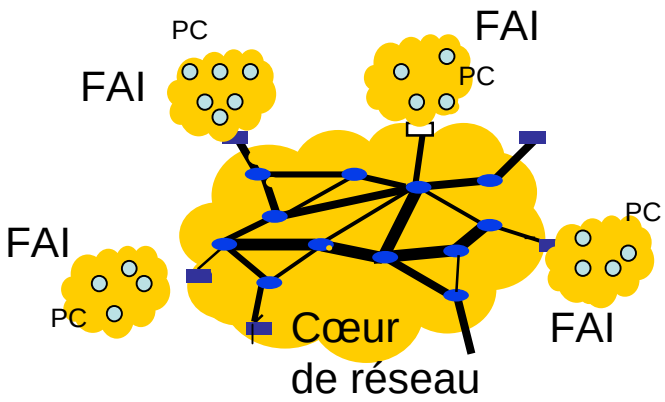
Emulateur
de lien

Grilles



Virtualisation

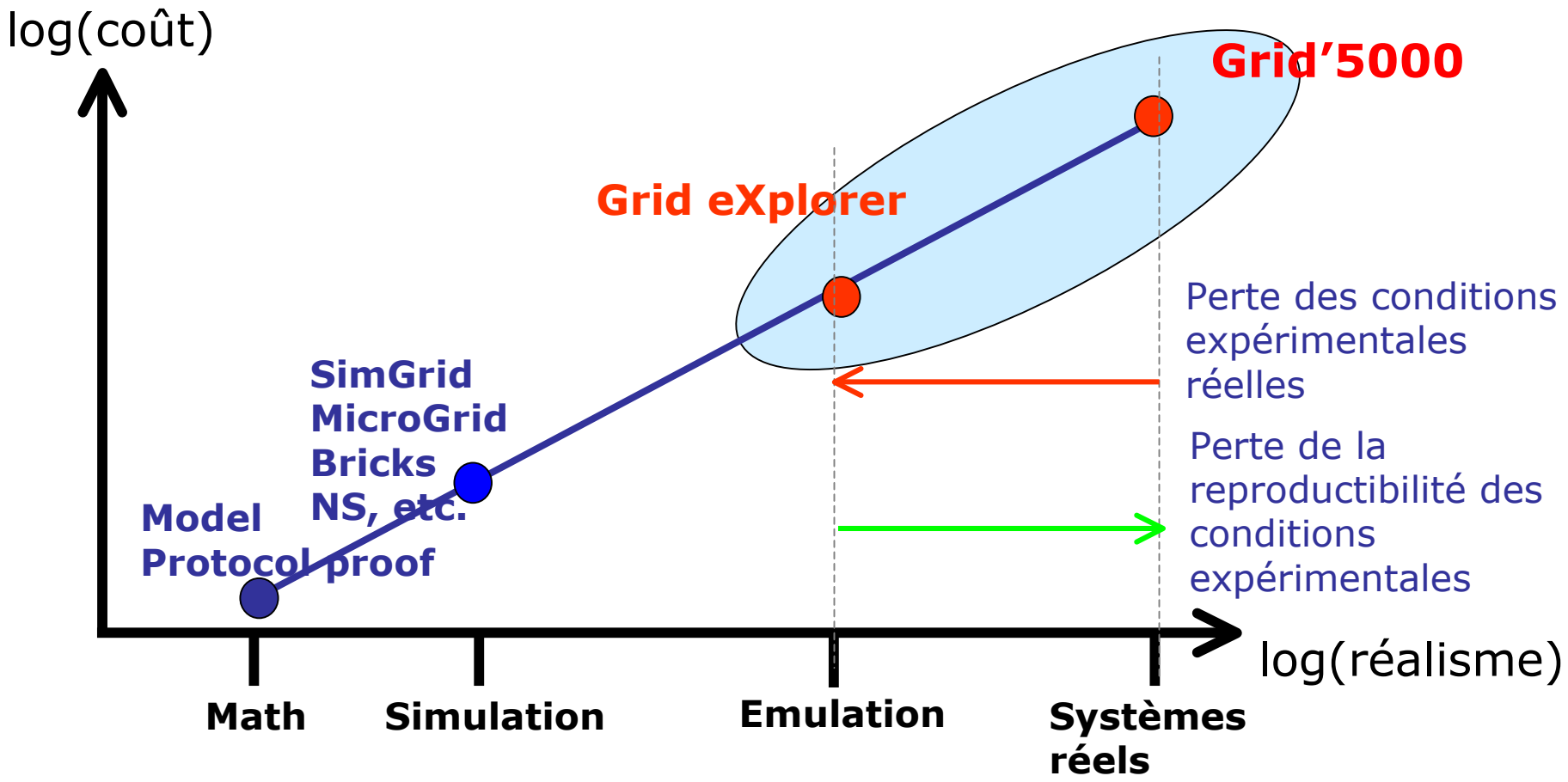
Systèmes P2P



Explorer les phénomènes à grande échelle liés au mouvement, stockage, calcul de données

Data Grid eXplorer + Grid'5000

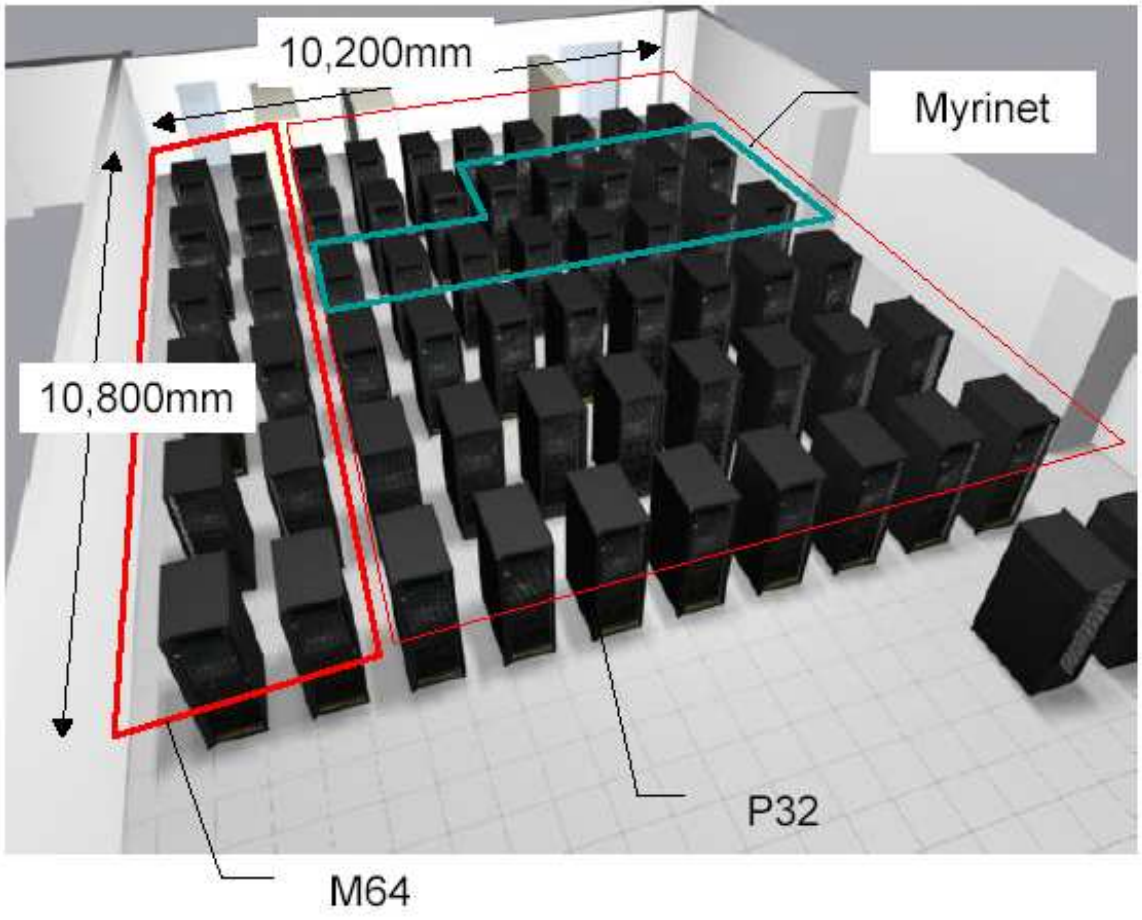
Combiner les deux plates-formes expérimentales nationales de Grille



Le montage du projet GdX a servi à comprendre les besoins des chercheurs en Grille et de fondation à la communauté Grid'5000

Une nouvelle Génération d'outils

AIST Super Cluster (Appearance image) **17 M$**



P32: IBM eServer325
Opteron 2.0GHz, 6GB
2way x 1074 node
Myrinet 2000

M64: Intel Tiger 4
Madison 1.3GHz, 16GB
4way x 131 node
Myrinet 2000

F32: Linux Networx
Xeon 3.06GHz, 2GB
2way x 256+ node
GbE

2928 CPUs

AIST

Le projet ACI MD Data Grid eXplorer

13 Labs

IMAG, ID (UMR 5132), Laboratoire d'Informatique et Distribution, Université de Grenoble
LaRIA (UPRES EA 2083), Laboratoire de Recherche en Informatique d'Amiens, Université de Picardie Jules Verne
LRI (UMR 8623), Laboratoire de Recherche en Informatique, Université de Paris-sud
LAAS-CNRS (UPR 8001), Laboratoire d'Analyse et d'Architecture des Systèmes
LORIA (UMR 7503), Laboratoire lorrain de recherche en informatique et ses applications
LIP-ENS Lyon (URM 5668), Laboratoire de l'Informatique du Parallélisme
LIFL (ESA 8022), Laboratoire d'Informatique Fondamentale de Lille
INRIA Sophia Antipolis, UNSA, I3S-CNRS
LIP6 (UMR 7606), Laboratoire d'Informatique de Paris 6
LABRI (UMR 5800), Laboratoire Bordelais de Recherche en Informatique
IBCP (UMR5086), Institut de Biologie et Chimie des Protéines
CEA, Direction des Technologies de l'Information (Saclay)
IRISA, Institut de Recherche en Informatique et Systèmes Aléatoires

Alain Lecluse (IBCP),
Alexandre Genoud, (Projet OASIS, INRIA Sophia Antipolis)
Antoine Vernois, (IBCP)
Arnaud Contes, (Projet OASIS, INRIA Sophia Antipolis)
Aurélien Bouteiller, (LRI),
Bénedicte Legrand (LIP6)
Brice Goglin (doctorant), (INRIA LIP RESO),
Brigitte Rozoy (LRI)
Cécile Germain (LRI)
Christophe Blanchet, (IBCP)
Christophe Cérin, (Amiens, Laria)
Christophe Chassot, (LAAS-ENSICA),
Colette Johnen (LRI)
CongDuc Pham, (LIP)
Cyril Randriamaro, (LaRIA)
Denis Caromel, (Projet OASIS, INRIA Sophia Antipolis)
Eddy Caron, (LIP/ENS Lyon),
Emmanuel Jeannot, (Loria)
Eric Totel (Supélec Rennes)
Fabrice Huet, (Projet OASIS, INRIA Sophia Antipolis)
Faycal Bouhaf (DEA)(INRIA LIP RESO),
Franck Cappello, (LRI)
Françoise Baude, (Projet OASIS, INRIA Sophia Antipolis)
Frédéric Desprez, (LIP/INRIA Rhône-Alpes),
Frédéric Magniette, (LRI)
Gabriel Antoniu, (IRISA/INRIA Rennes),
George Bosilca, (LRI)
Georges Da Costa, (ID-IMAG),
Géraud Krawezik (LRI)
Gil Utard, (LaRIA)
Gilles Fedak, (LRI)
Grégory Mounié (ID-IMAG)
Guillaume Auriol, (LAAS-ENSICA),
Guillaume Mercier, (LaBRI),
Guy Bergère, (LIFL, GrandLarge INRIA Futur)
Haiwu He, (LIFL, GrandLarge INRIA Futur)
Isaac Scherson, (LIFL, GrandLarge, INRIA Futur)
Jens Gustedt (LORIA & INRIA Lorraine)
Joffroy Beauquier (LRI)
Johanne Cohen, (Loria)
Kavé Salamatian (LIP6),
Lamine Aouad (LIFL, GrandLarge, INRIA Futur)
Laurent Baduel, (Projet OASIS, INRIA Sophia Antipolis)
Laurent Dairaine, (LAAS)
Luc Bougé, (IRISA/ENS Cachan Antenne de Bretagne),
Luciana Arantes (LIP6),
Ludovic Mé, (Supélec Rennes)
Luis Angelo Estefanel, (ID-IMAG)
Marin Bertier (LIP6),
Mathieu Goutelle, (KIP)
Mathieu Jan, (IRISA)
Michel Diaz, (LAAS-ENSICA),
Michel Koskas (Amiens, Laria)
Nicolas Lacorne, (IBCP)
Nicolas Larrieu (LAAS-ENSICA),
Nicolas Viovy (CEA-DSM-LSCE)
Oleg Lodygensky, (LRI)
Olivier Richard (ID-IMAG),
Olivier Soyez, (LaRIA)
Pascal Berthou, (LAAS-ENSICA),
Pascale Primet (LIP),
Pascale Vicat-Blanc Primet, (INRIA LIP RESO),
Patrick Sénac, (LAAS-ENSICA),
Philippe d'Anfray (CEA-DTI/SISC),
Philippe Gauron, (LRI)
Philippe Owezarski (LAAS)
Pierre Fraigniaud, (LRI)
Pierre Lemarini, (LRI)
Pierre Sens (LIP6 / INRIA),
Pierre-André Wacrenier, (LaBRI),
Raymond Namyst, (LaBRI),
Samir Djilali, (LRI)
Sébastien Tixeuil (LRI)
Serge Petiton, (LIFL, GrandLarge INRIA Futur)
Stéphane Vialle (Supélec)
Tanguy Pérennou (LAAS)
Thierry Gayraud, (LAAS-ENSICA),
Thierry Priol, (IRISA)
Thomas Hérault, (LRI)
Timur Friedman (LIP6)
Vincent Danjean, (LaBRI),
Vincent Néri (LRI)

Data Grid eXplorer
Gouvernance

-Responsable	Franck Cappello (LRI)
-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)
-Réseau,	Pascale Primet (LIP, Inria RESO)
-Applications.	Christophe Cérin (Laria)

Data Grid eXplorer
Moyens

Obtenu:

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

Demandes:

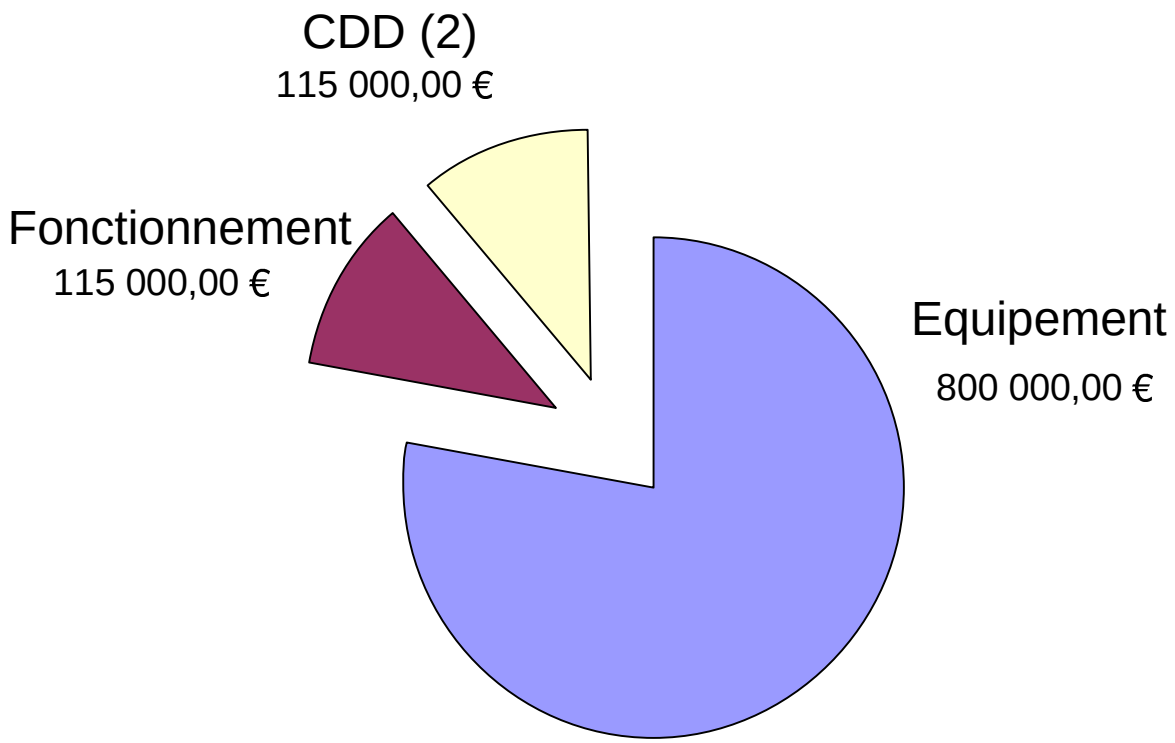
SESAME Ile de France:	900 K€ TTC
ASTRE 2005:	300 K€ TTC

Moyens humains:

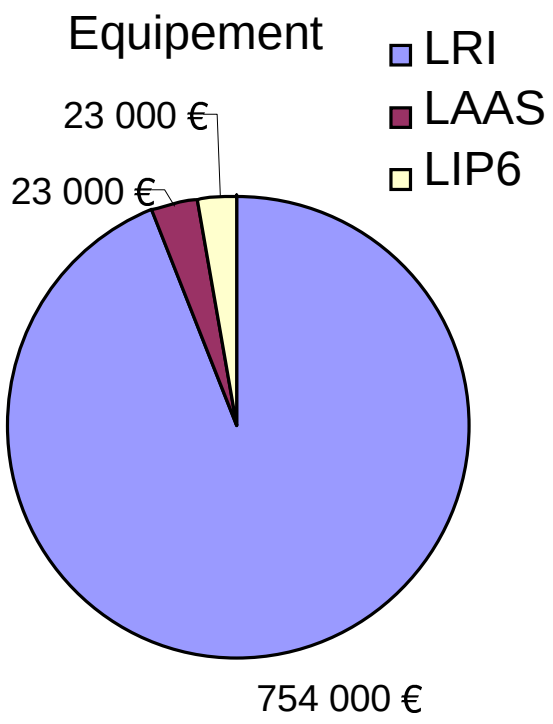
36 mois (2x18 mois) ingénieur ACI Masse de données
24 mois ingénieur associé INRIA
Ingénieurs IDRIS (difficile à quantifier)
Soutien ingénieurs LRI (difficile à quantifier)

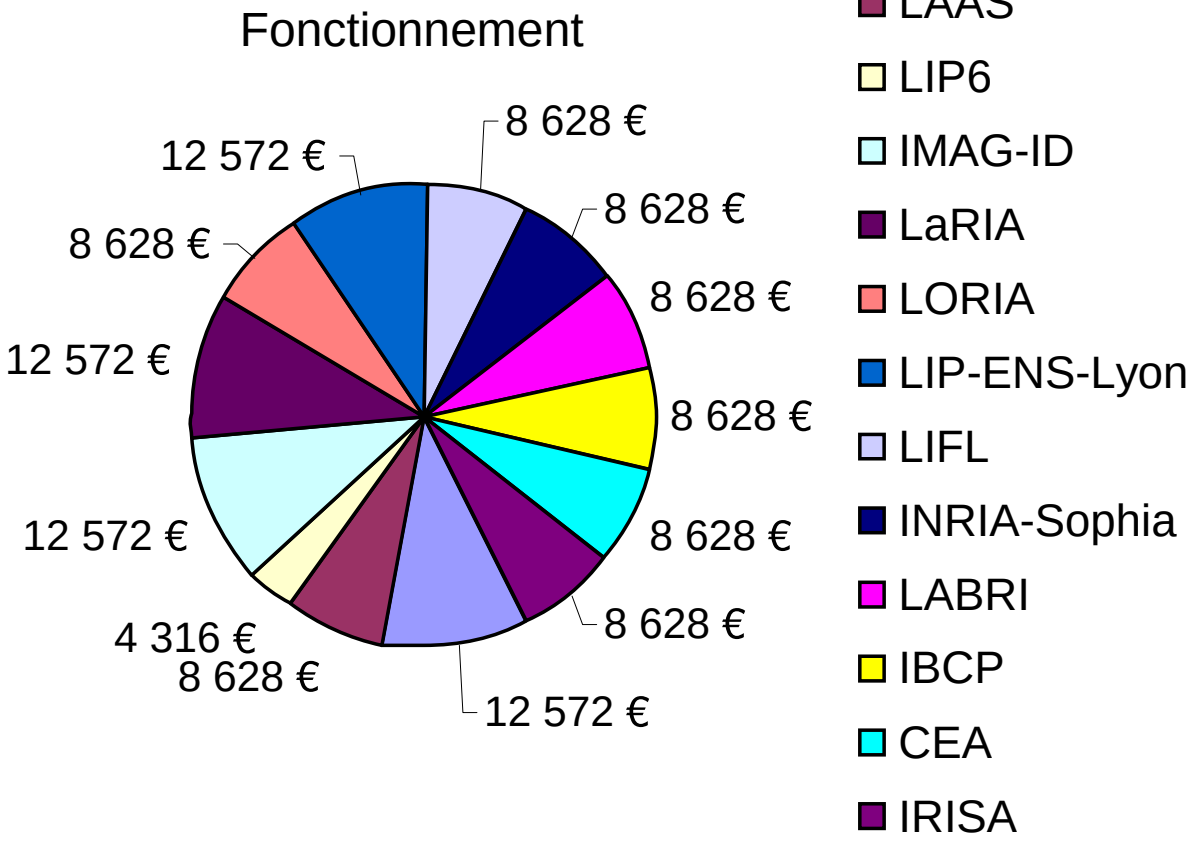
Data Grid Explorer
Budget ACI Masse de Données

Budget total 1030 K€ TTC



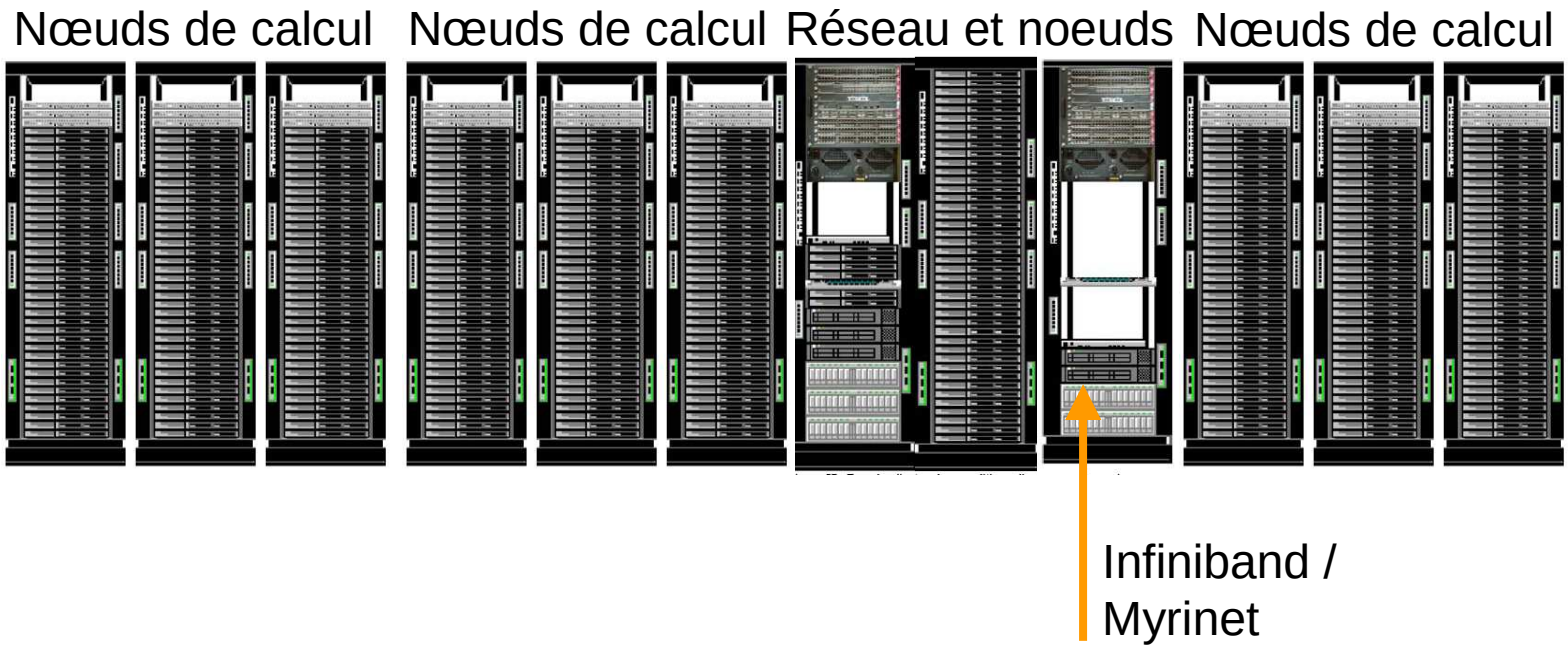
Data Grid Explorer
Budget ACI Masse de Données





Data Grid eXplorer
Matériel (ACI MD + ACI Grid'5000 + INRIA)

Configuration : 648 C CPUs + 48 N CPUs + 3 Frontaux



- Notre objectif est d'avoir cette configuration fonctionnelle pour début 2005 !
- Première tranche installée le 15 Octobre (2/3 du total)

De quoi avons-nous besoin pour l'émulation Grilles/P2P ?

1) Une Plate-forme pour exécuter les expériences (le cluster)

2) Des nœuds P2P/Grid contrôlables (protocoles, OS, middleware)

3) Un réseau « Contrôlable » et « Observable » entre les nœuds P2P/Grid → (émulation réseau)

4) Un ensemble d'outils logiciels pour, exécuter, surveiller et contrôler les expériences et collecter les résultats

Data Grid eXplorer

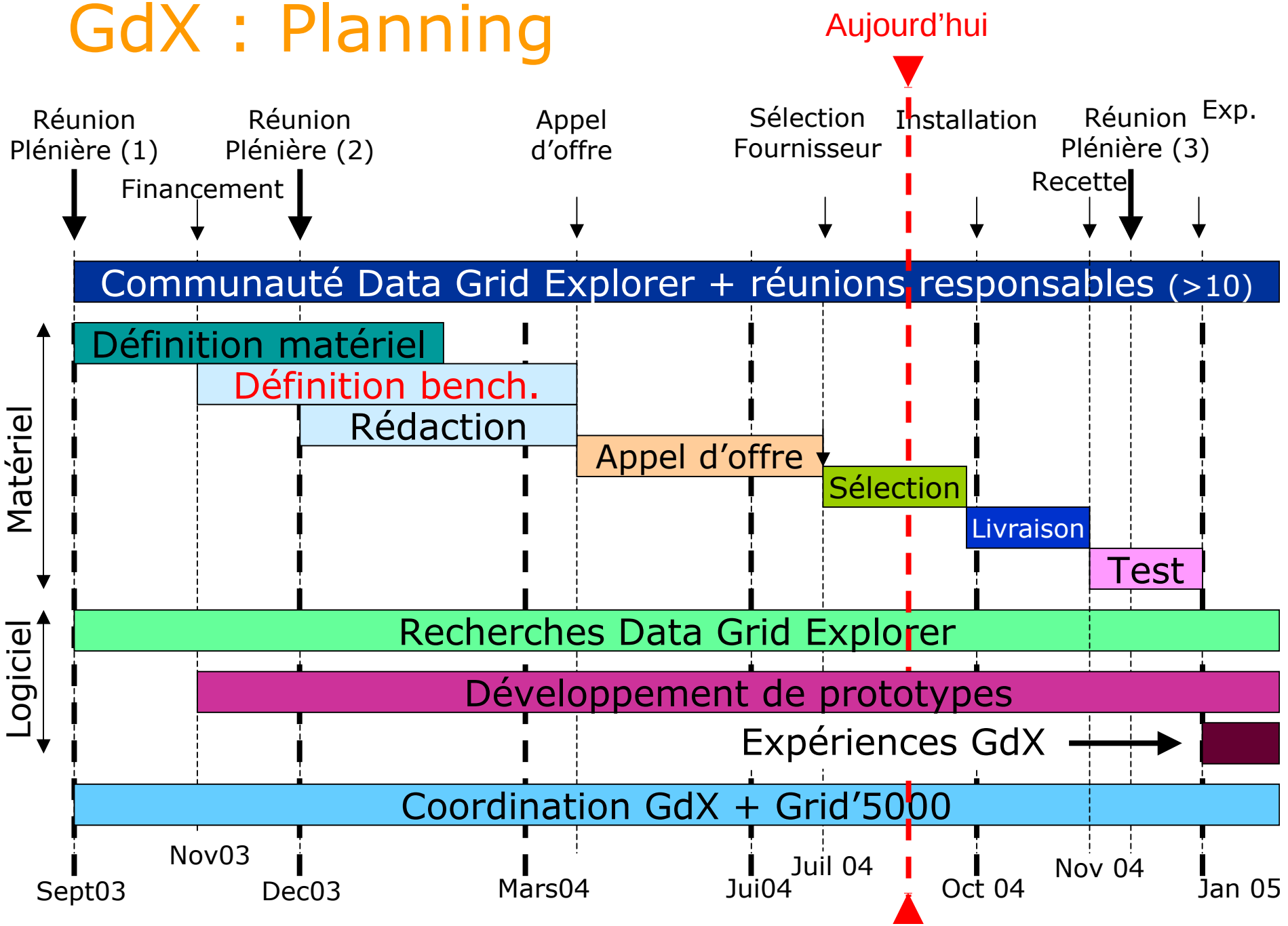
Travaux scientifiques

1) Construire l'instrument:
   - Cluster de 1K CPU (peut être limité à 600 en fonction du budget)
   - Un réseau configurable (Ethernet, Myrinet)
   - 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 traitement données :
     - Extensibilité, Tolérance aux pannes, 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.

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

GdX : Planning



Conclusion

- 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
  - Sera un plate-forme expérimentale pour les chercheurs en Grille/P2P
  - Un émulateur de système à grande échelle
  - Le matériel sera installé en Octobre 2004
  - Relation étroite avec le projet Grid'5000 (plusieurs chercheurs participent aux deux projets)

  - Présenté et Démo sur le stand Grid'5000 à SuperComputing 2004

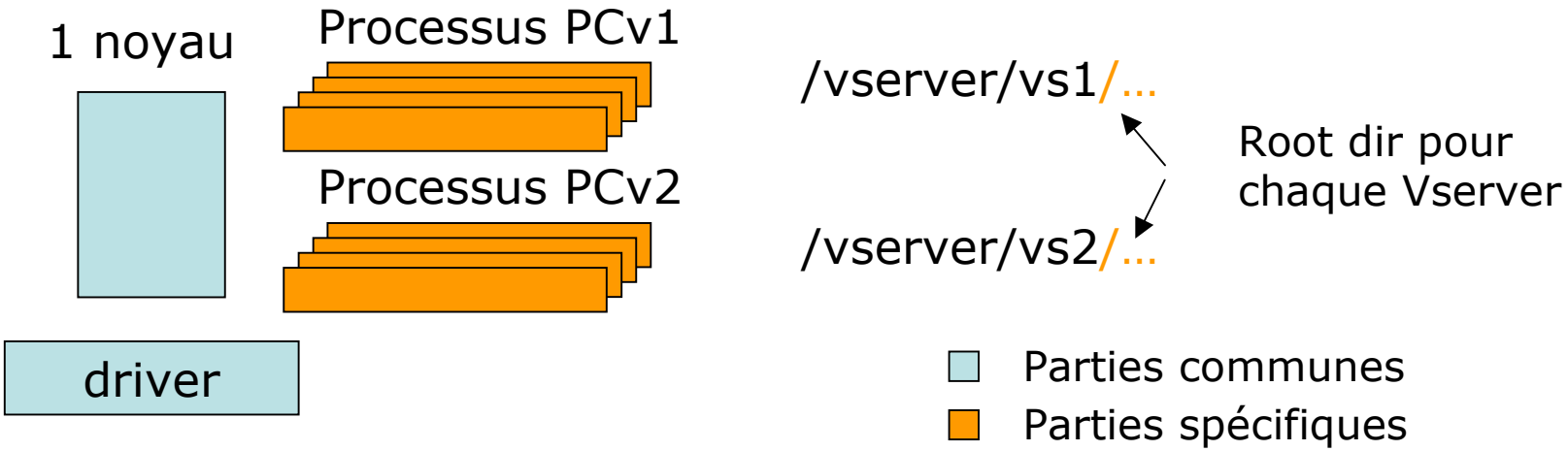

Questions ?

Exemples de travaux: I.2 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,…

Exemple avec Vserver:

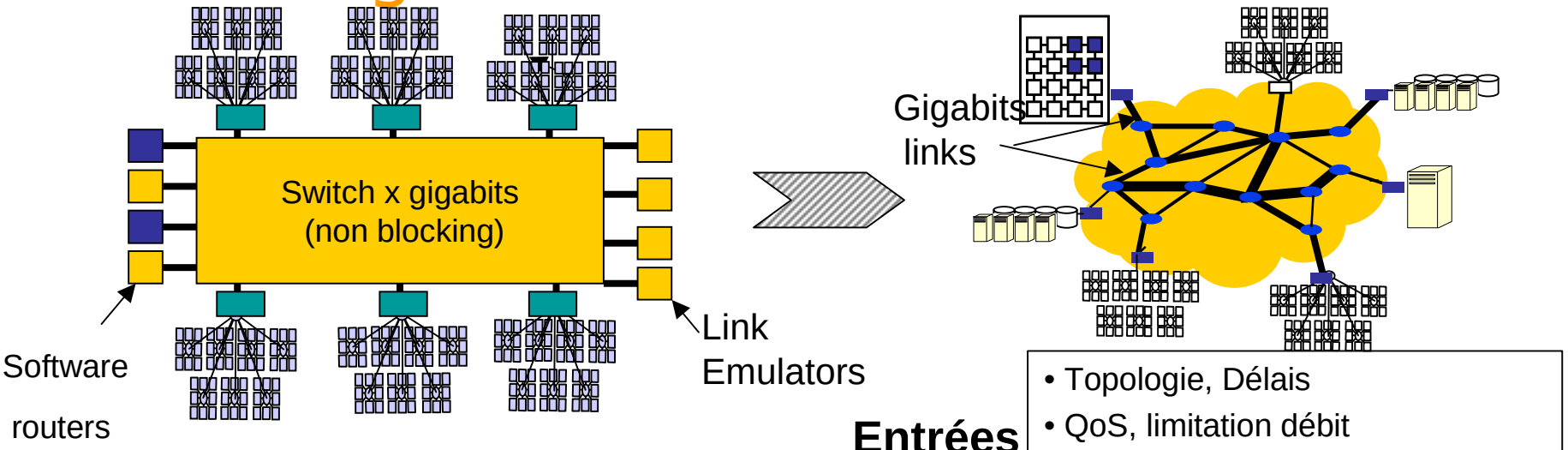


Première question scientifique : quelles métriques, quels benchmarks

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

Nouveau Thésard à partir de 2004

Exemple de travaux : I.5 eWAN: Nuage réseau haut débit émulé



Entrées
- Topologie, Délais
- QoS, limitation débit
- IPv6 ou v4, jumbo frames

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

Sorties
- Performance et comportement E2E
- Influence des paramètres d'échelle
- Taux d'utilisation ressources

Etapes

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

Point dur: Le multigigabit longue latence

Exemples de travaux:

I.6 Emulation de l'hétérogénéité (LORIA)

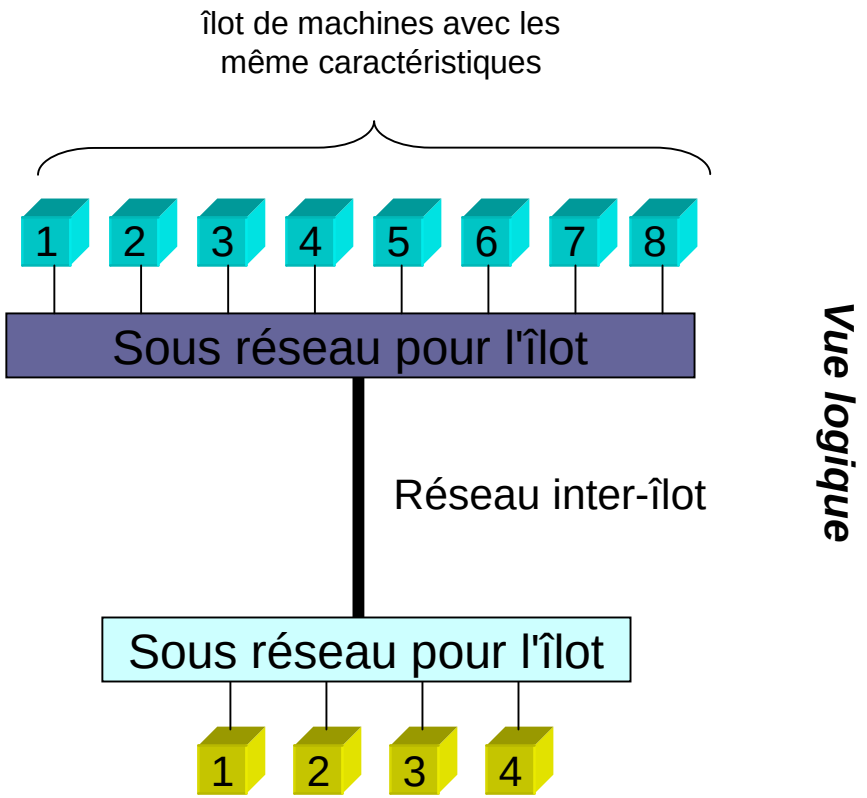
Objectif : rendre GdX hétérogène pour faire des expériences d'émulation

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'adresse 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
                    % gaussiène de moyenne 1000 Ko/s et de variance 200 Ko/s
…
}
!INTER : [ilot1;ilot2] [200-200] [300-300] [1;0] -1 %caractéristiques réseau entre ilot1 et ilot2
```

Exemples de travaux: I.8 Emulation Internet (LAAS)

- Conception (à l'aide du simulateur NS) d'une méthode de reproduction des condition 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

Exemples de travaux : II.3 Gestion des fautes dans les configurations de type Grid (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 dynamicité 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, soumissions à JPDC, soumission prévue à IPDPS
- Réalisations
  - Matériel : plate-forme d'emulation : 20 nœuds de calcul, 2 nœuds d'emulation
  - Logiciels : Detecteurs de fautes, simulateur de système à large échelle, injecteur de fautes

[Gri]-P-{PS}

Grid Protein Pattern Scanning

The Grid Protein Pattern Scanning-GriPPS project aims to adapt a bioinformatic algorithm of protein pattern scanning to the grid infrastructure. The behavior of the algorithm and the data wil be studied on several experimental grids.The tested middleware are those from the projects DataGrid (Eur-FP5), e-Toile (Fr-RNTL) and GASP (Fr ACI GRID).

Protein pattern scanning: *PattInProt*

–Hypothesis on function of unknown proteins (and also genes)
–Allowing biologically mismatch to enhance the sensitivity
–Clustering proteins into family
–Sequence annotation and biological crosslinks
-PattInProt is integrated into our current software (MPSA) and web portal (NPS@

GRID computing context

–Adapt PattInProt bioinformatic algorithm and data to the grid in order to foresee their behavior on a grid plateform
–Identifying bioinformatic specific constraints on the grid
–Test on several middleware and model of grid: DataGrid, e-Toile, DIET.

GRID benefit

–Complexer analyses on larger data set, lower threshold
–Distributing sequence databanks
–Integrity, privacy and security of data and method software
–Recommendations on gridification of similar bioinformatic algorithmes

Institut de Biologie et Chimie des Protéines

IN2P3 Centre de Calcul

Contact: http://gripps.ibcp.fr
Christophe.Blanchet @ ibcp.fr

Exemples de travaux : II.15 Internet Nouvelle Génération (LAAS)

- 3 expérimentations prêtes
  - QoS multi-domaines 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