Projets de recherche sur les grilles en France

Yves Denneulin
Laboratoire ID/IMAG
CNRS - INPG - INRIA - UJF

Journées IN2P3 - 18 septembre 2006

Bref historique

- La préhistoire : apparition du méta-computing
  - 93-96 : expérimentations sur PVM, MPI hétérogène
  - apparition de Nexus, base de ce qui deviendra Globus
- Les débuts : première version de Globus
  - lancement de Datagrid
  - démarrage des ACI GRID
  - projet RNTL E-toile

Les plate-formes

- Un souci : disposer d'outils pour expérimenter facilement des solutions logicielles innovantes
  - "clients" : chercheurs en informatique
  - pas d'objectifs directs de production
- Projets principaux
  - Grid Xplorer
  - Grid5000

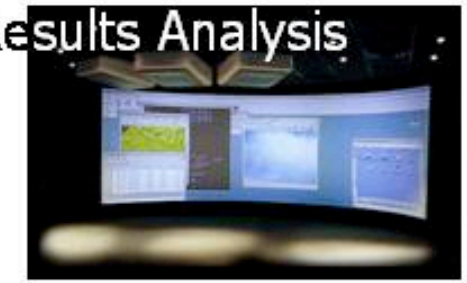
Grid eXplorer

Grid eXplorer
Analogy with physic instruments

A set of sensors
Inside real platform

A database of
Experimental
conditions

A tool set for
Observation
Results Analysis



Validation on
real platforms

F. Cappello, O. Richard, P. Sens ---oo Draft oo--- Contact us for experiment proposal

Grid eXplorer

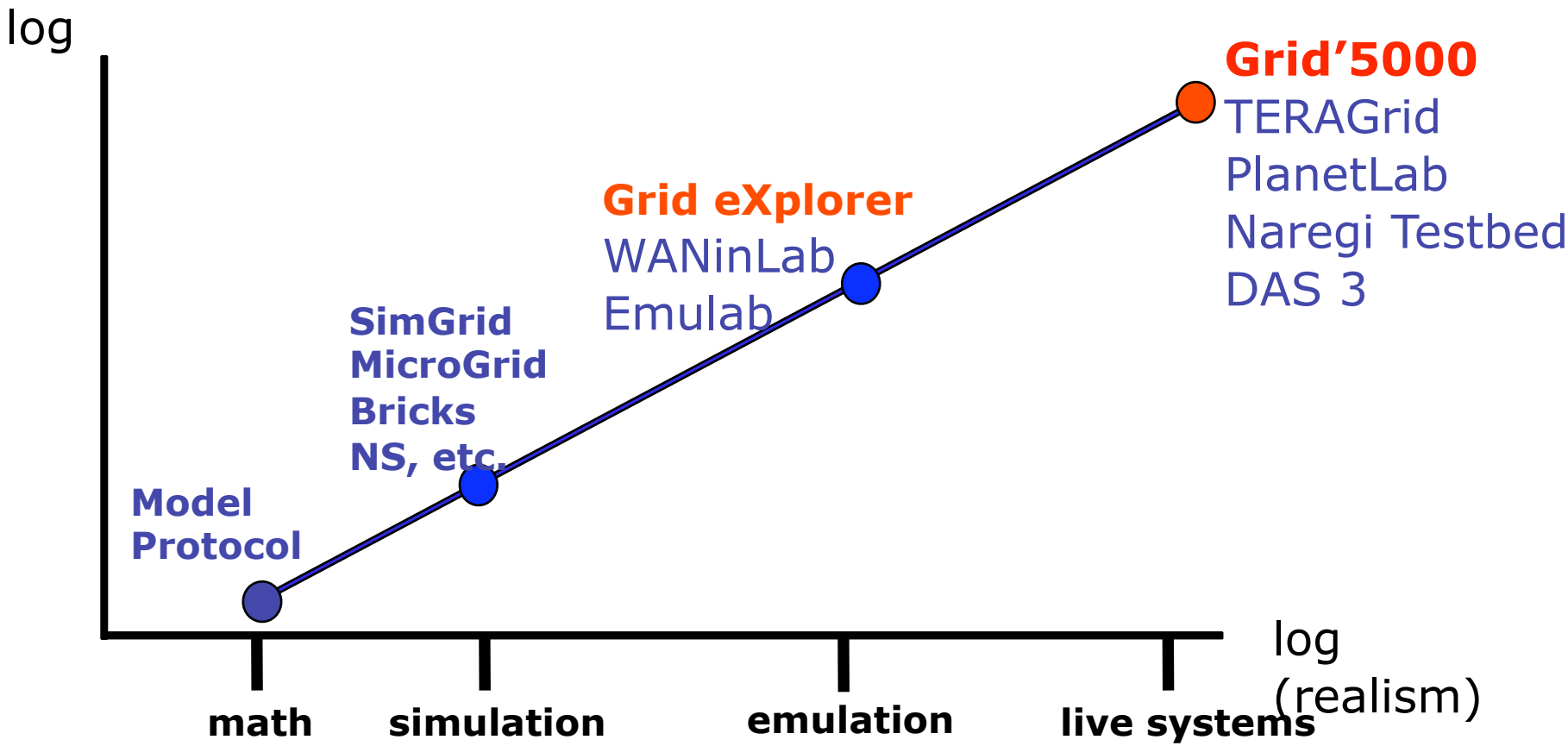
- Pour les chercheurs en grilles, P2P et réseaux
- Utilisation dans de nombreux domaines
  - émulation de systèmes grand échelle
  - injection de charge, ...
- Éléments
  - 600 CPUs, base de données de conditions expérimentales, simulateurs/émulateurs, aide à la visualisation

Grid experimental platform

Grid'5000 is large scale testbed dedicated to Grid experiments

→ Grid’5000 as a live system

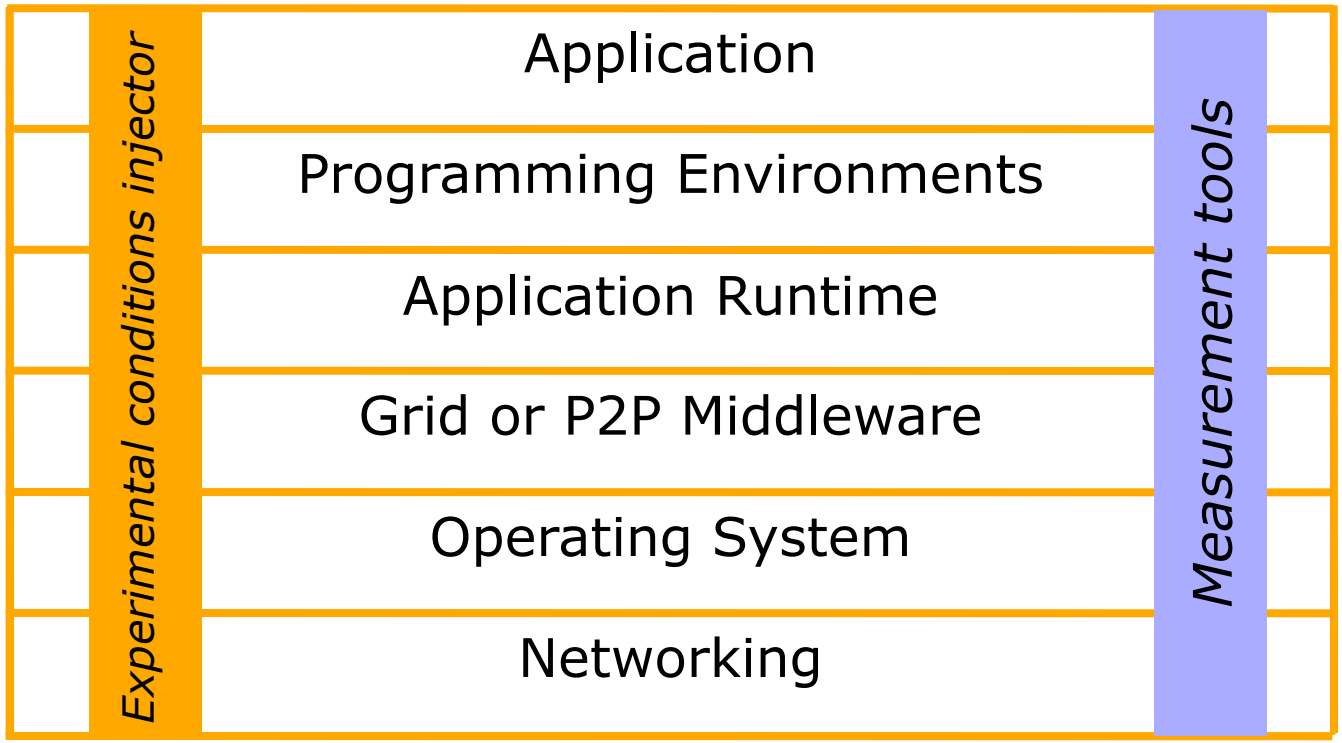
→ Grid eXplorer as a large scale emulator



https://www.grid5000.fr

Grid'5000 principle:

A highly reconfigurable experimental platform

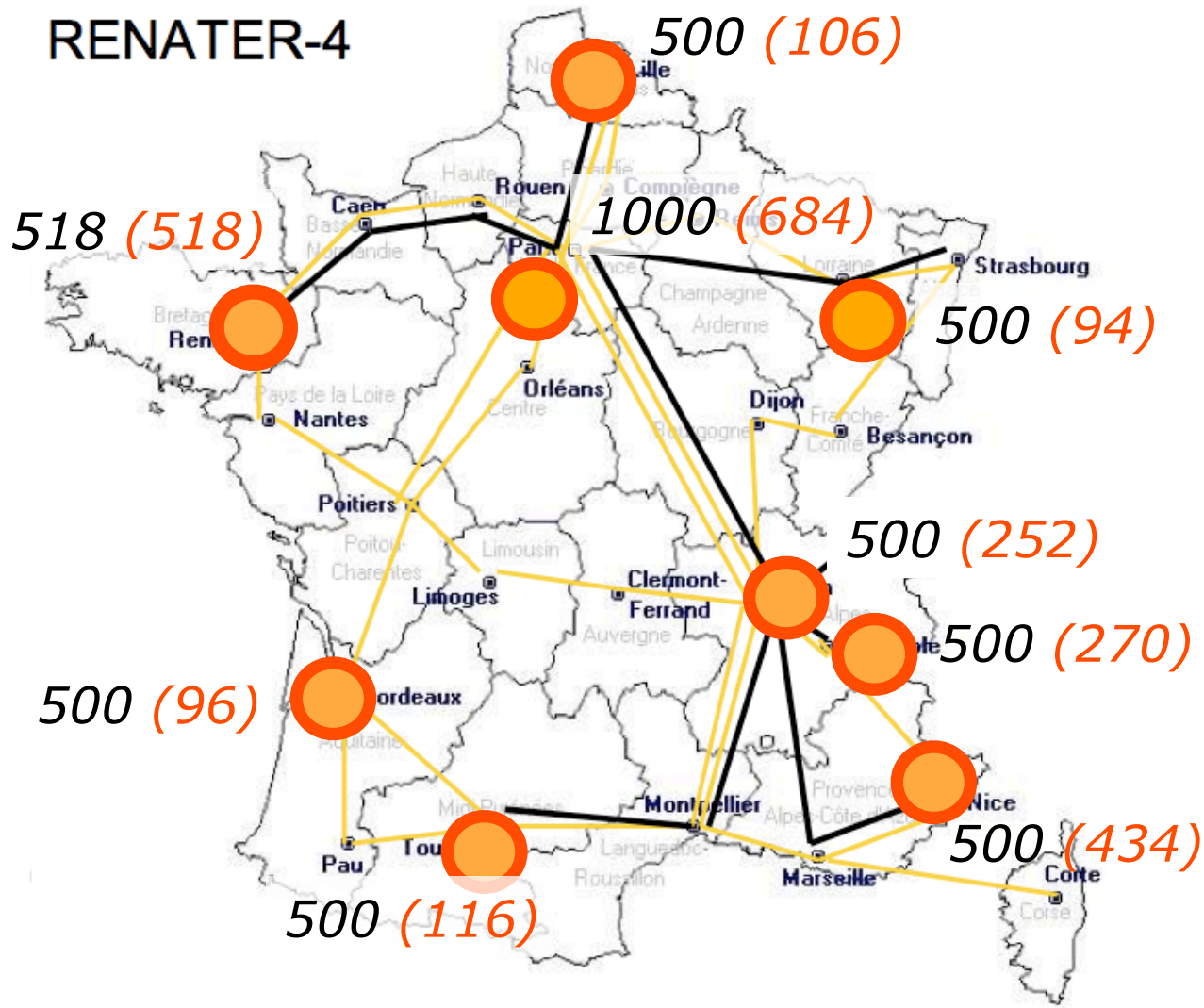


→ Let users create, deploy and run their software stack, including the software to test and their environment + measurement tools + experimental conditions injectors

https://www.grid5000.fr

Grid'5000 map

A large scale Instrument to study Grid issues



Dark fiber
Dedicated Lambda
Fully isolated traffic!

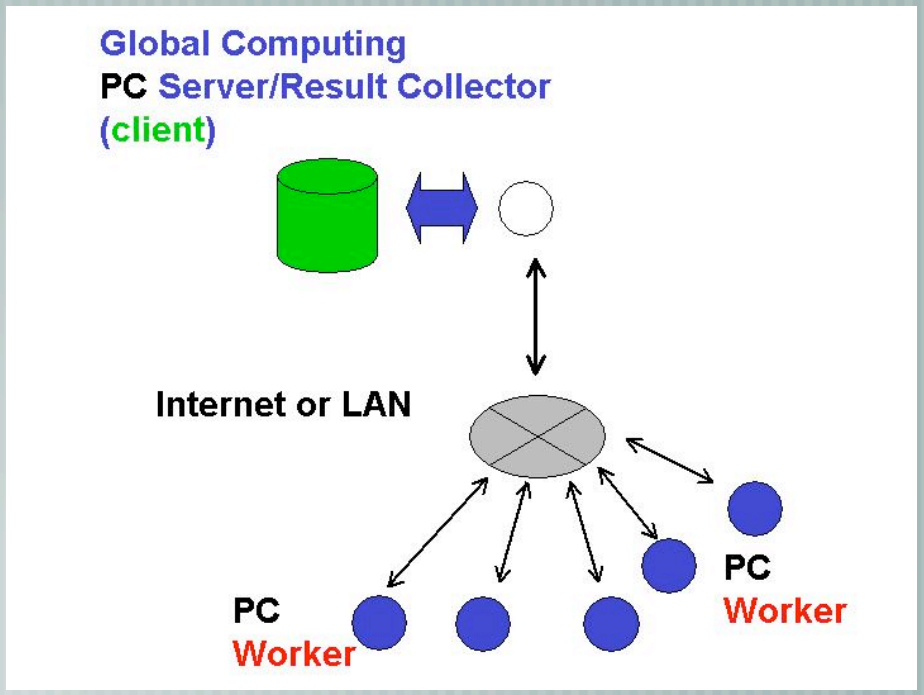
https://www.grid5000.fr

Projet Graal (Lyon)

- Algorithmique et ordonnancement pour plates-formes hétérogènes distribuées
- DIET : serveurs de calculs sur la grille, utilisation du RPC
- MUMPS : résolution linéaire de systèmes creux
- FAST : prédiction de performances sur la grille
- SimGrid : présenté plus tard
- Projet RESO : utilisation des réseaux très haut débit

XtremWeb

- LRI, Orsay
- software platform designed to serve as a substrate for Global Computing (Large Scale Distributed System) experiments
- Vol de cycles
- Différents travaux simultanées
- Tourne dans une JVM



Projet Padico

- Couplage de code d'applications
  - technologie Corba parallèle
  - implantation efficace de la norme Corba parallèle
- Paco++ : objets Corba parallèles
- GridCCM : composants Corba parallèles
- Adage : déploiement d'architectures CCM ou MPICH sur grilles

Le laboratoire ID (Grenoble)

- Travaux sur les grilles depuis 1998
  - supports d'exécution hétérogènes,
  - gestion et mouvement de données,
  - ordonnancement,
  - évaluation de performances,
  - déploiement et administration de grandes configurations.
- Approches théorique ET pratique

Un besoin commun au chercheurs en Grid Computing

Des questions récurrentes dans notre communauté :

- Quel est le meilleur algorithme d'ordonnancement pour une application donnée sur un type de plate-forme donné ?
- Quelle est la meilleure stratégie de réplication/distribution/gestion du cache pour un type de service donné ?
- Comment mon architecture passe-t-elle à l'échelle et quelle est sa résistance à différents types de pannes ?.
- Comment contrôler l'accès aux différentes ressources pour satisfaire les utilisateurs ?
- ...

On a besoin de faire des expériences pour guider des recherches ou valider des techniques.
En informatique, comme en biologie ou en physique, les expériences grandeur nature sont indispensables mais la simulation est un outil puissant permettant d'obtenir des résultats pertinents dans de nombreuses situations.

Simulation = reproductibilité, contrôle, rapidité.

SimGrid

Un projet vieux de 5 ans originellement développé pour la communauté ordonnancement.

Application L'application est modélisée en terme de processus communiquants pouvant exécuter des taches de calcul.

Ressources Une ressource (calcul, réseau, disque) est définie par un taux de service (pouvant varier dans le temps pour simuler une charge externe), une latence d'accès, un scénario de pannes.

Tâches Les tâches peuvent dépendre de plusieurs ressources (transfert de données sur un ensemble de liens, calcul utilisant un disque et un CHU,...).

Dans la plupart des autres projets, le partage des ressource "émerge" de simulation à évènement discret de bas niveau $\rightsquigarrow$ relativement lent.
SimGrid utilise un mélange de modèles fluides et de simulation à évènement discrets, ce qui permet d'obtenir des résultats réalistes et très rapides.

http://simgrid.gforge.inria.fr/

Softwares for Large Computing and Deep Experiments

It manages ressources of clusters as a **traditional batch scheduler** (Torque/LSF/SGE).

Its design is based on **high level tools** :
- **relational database** engine MySql,
- **scripting langage** Perl
- **scalable exploiting tools** Taktuk (parallel launcher).

It is **flexible** enough to be suitable for production clusters and research experiments.

Some features:

Advance Reservation, Modable Jobs, associate resource* (as licence or vlan), isolation by cpuset*, checkpointing support*, besteffort jobs, job dependencies*....

**in version 2*

Software for Large Computing and Deep Experiments

CiGri

It manages the **execution of large sets of parametric jobs** (>100K) on *ligthweight* grids by submitting individual jobs to each batch scheduler.

Associated to OAR, it allows to **exploit unused nodes (besteffort jobs)**.

Users can easily **monitor** and **control** their set of jobs through a web portal.

System provides mechanisms to **identify** job **error causes**, to **isolate faulty components** and to **resubmit job** in a safer context.

Some new features:

transparent environment setting (rsync & kadeploy), checkpointing support

http://cigri.imag.fr

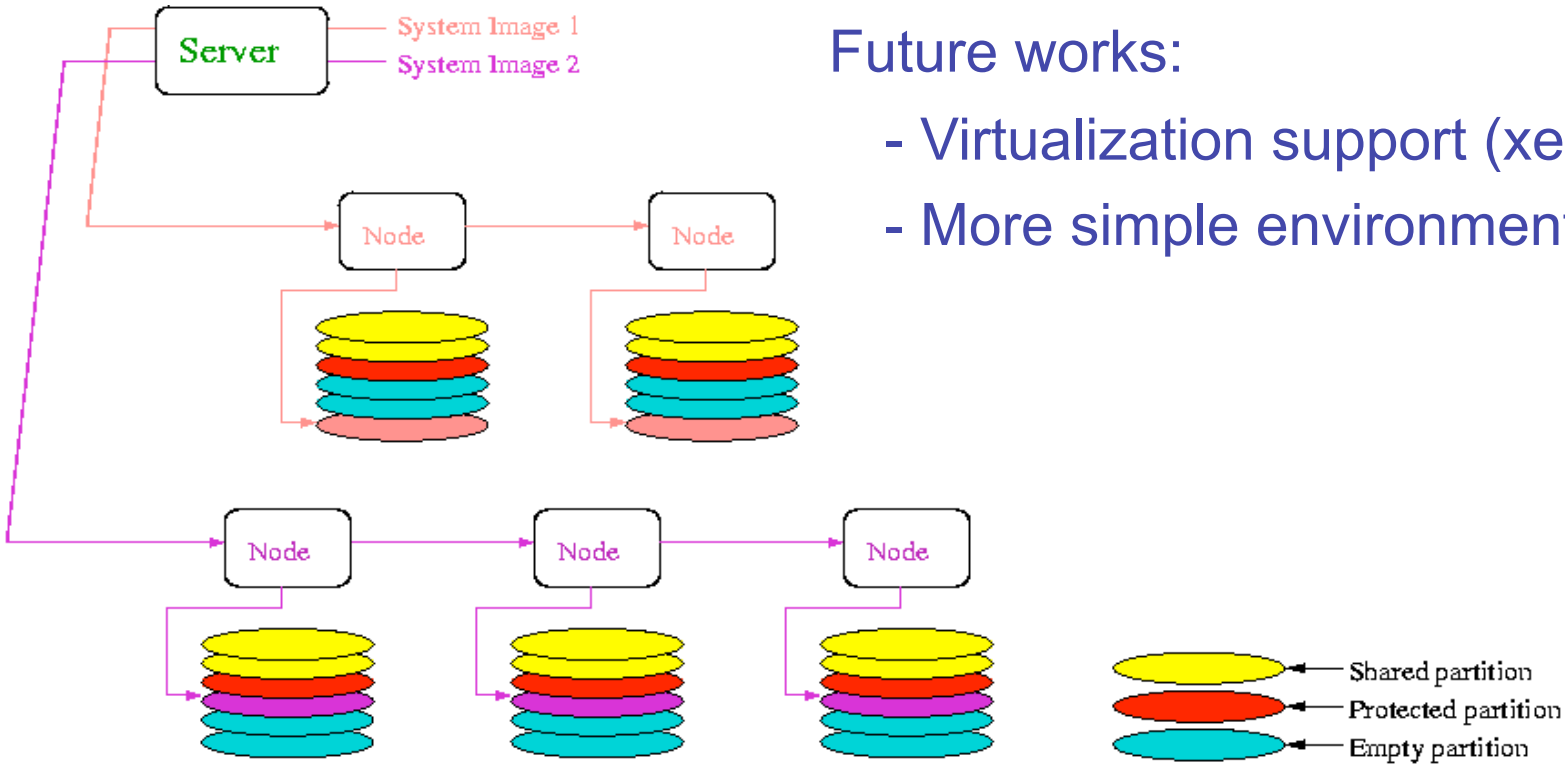
Softwares for Large Computing and Deep Experiments

KADEPLOY

is an OS installer which provides :

- an **efficient pipelined cloning system** and
- **post-installation** tools.

Performance example : 5 min for 100 nodes



Future works:

- Virtualization support (xen,vserver)
- More simple environment management

http://www-id.imag.fr/Logiciels/kadeploy/

Gestion de données

- Stockage sur grappes (NFSP)
- Transfert entre grappes (Gxfer - projet E-toile)
- Ordonnancement d'I/O (aIOLi)
- Indexation de données non structurées (ACI MD Gedeon)
  - "à la SRB" mais plus flexible
  - pas de centralisation, cache de méta-données et de requêtes
  - distribution des fichiers de méta-données

Gestion d'erreurs

- Architecture de grande taille => apparition d'erreurs
- Comment les caractériser ?
- Comment les classifier ?
- Quels traitements leur appliquer ?
- Comment les corriger ?
  - calcul redondant