Investigating the impact of the Large Scale on distributed systems

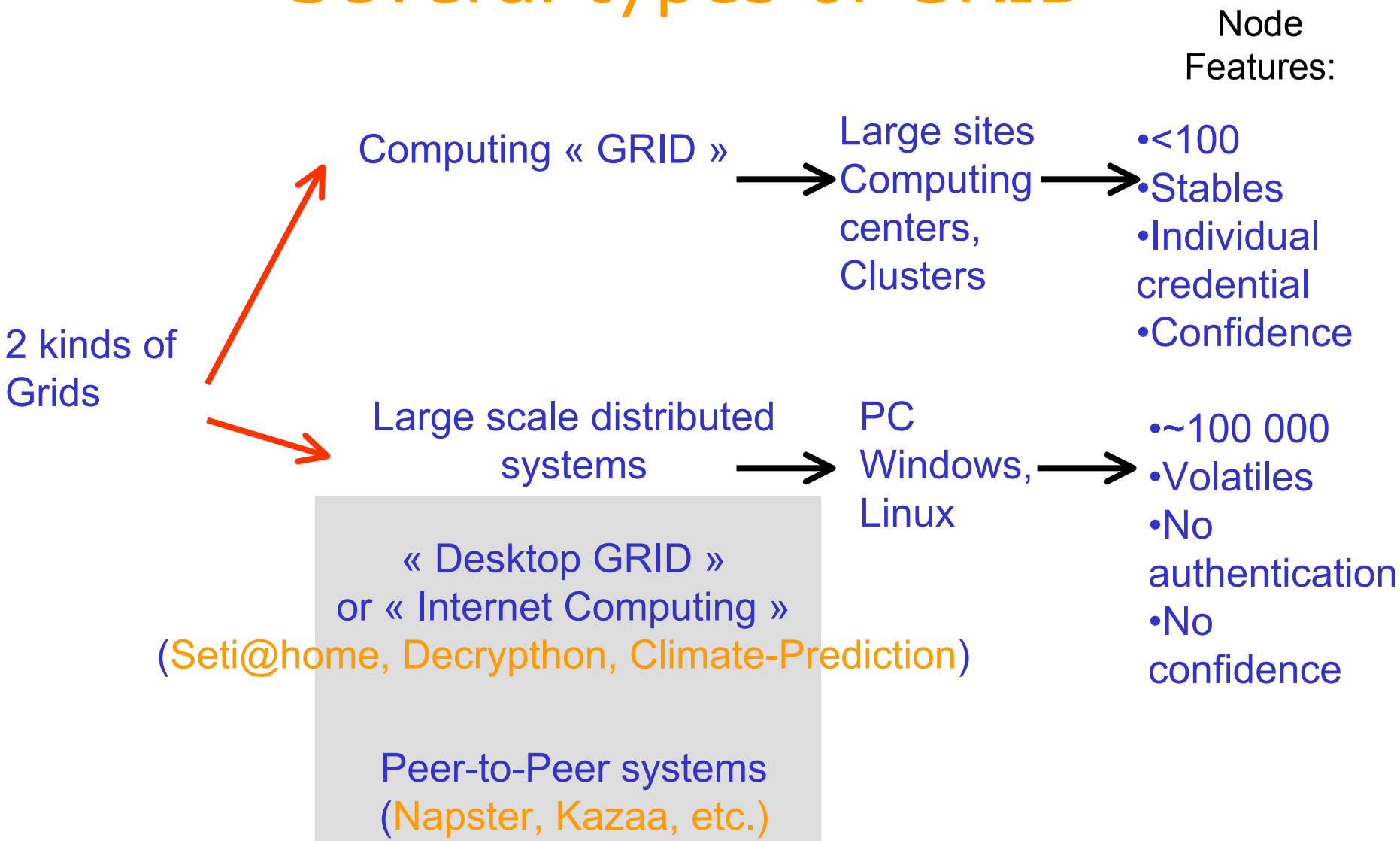
F. Cappello
INRIA

Grand-Large Project, INRIA/PCRI

LRI, Université Paris Sud
fci@lri.fr,
www.lri.fr/~fci

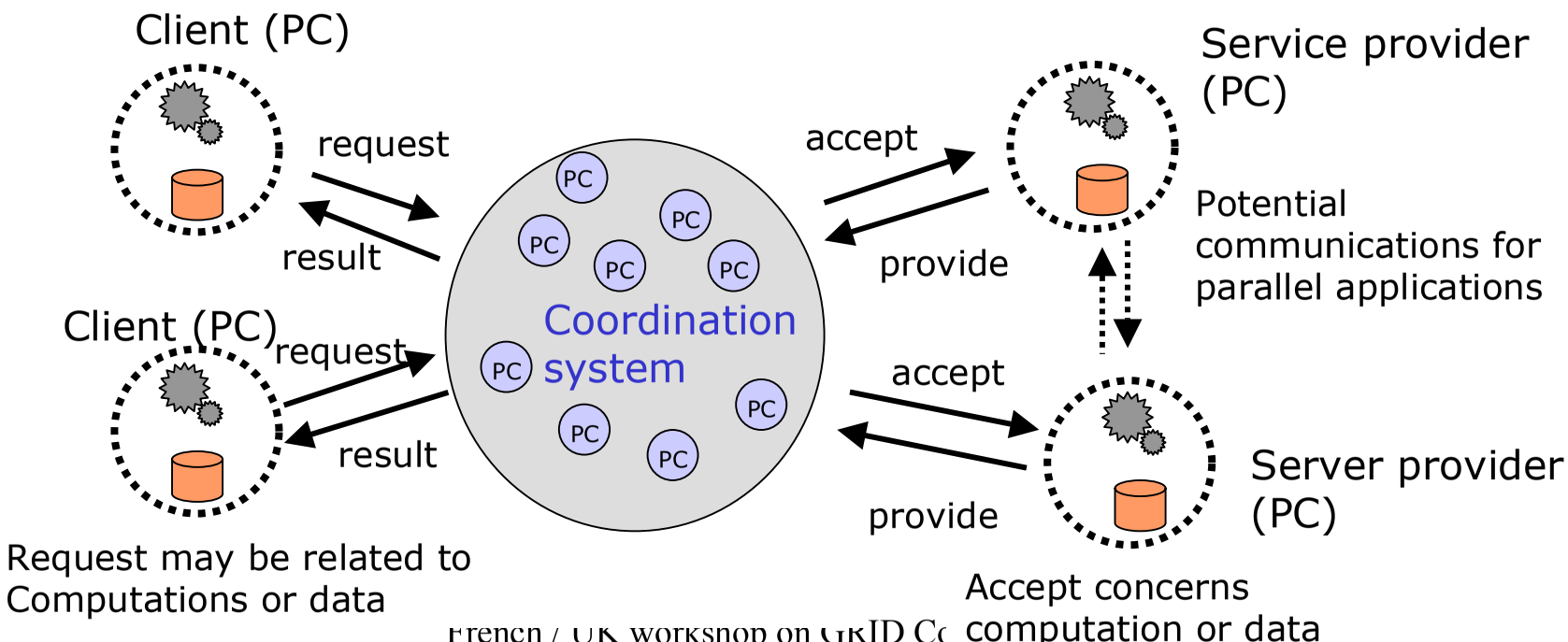
Action Concertée Incitative [ACI]
Globalisation des Ressources Informatiques
et des Données [GRID]

Several types of GRID



Fusion of Dgrid and P2P →General Purpose Large Scale Distributed systems

- Large computing infrastructures (~10 000 nodes or more)
- Geographically distributed / different administration domains
- With almost no control of the participating nodes
- Where any node to play different roles (client, server, system infrastructure)



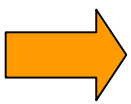
Distributed System Problematic renewal

A very simple problem statement but leading to a lot of research issues (classical OS):

Scheduling, Load Balancing, Security, Fairness, Coordination, Message passing, Data storage, Programming, Deployment, etc.

BUT « Large Scale » feature has severe implications:

- Node Volatility, Network failures, Asynchrony
- Lack of trust (very low control of participating nodes)
- No consistent global view of the system



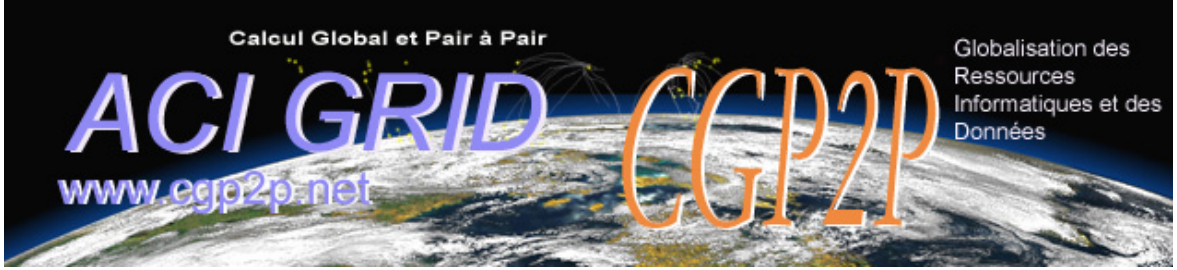
Conventional techniques/approaches may not fit

Ex: fault tolerance

- Classical fault tolerance (consensus impossible)
- Self-Stabilization (the system is always changing)

New approaches (intrinsically scalable/FT) are needed

- Autonomous decisions, Self-organization, etc.



26 pers, 7 labs (started in 2001; end in July 2004)

Research topics and sub-projects:

Global architecture	(F. C. and O. R.)
User Interface, control language	(SPI, S. Petiton)
Security, sandboxing	(SPII, O. Richard)
Large scale Storage	(SPIII, Gil Utard)
Inter-node communications : MPICH-V	(SPIV, F. Cappello)
Scheduling -large scale, multi users-	(SPIV, C. G. and F.C.)
Theoretical proof of the protocols	(SPV, J. Beauquier)
GRID/P2P interoperability	(SPV, A. Cordier)
Validation on real applications	(G. Alléon, etc.)

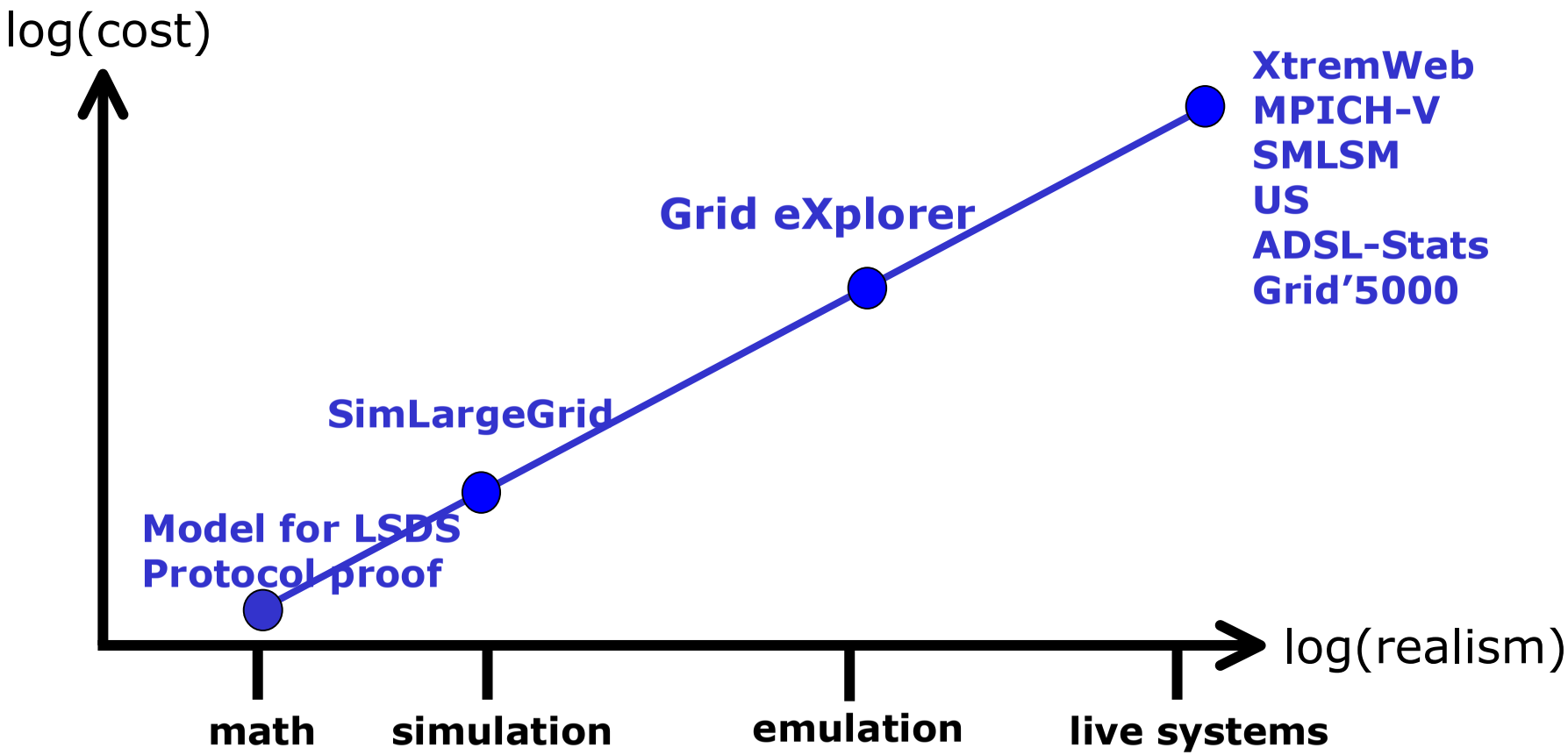
Action Concertée Incitative [ACI]
Globalisation des Ressources Informatiques et des Données [GRID]

Combining research tools

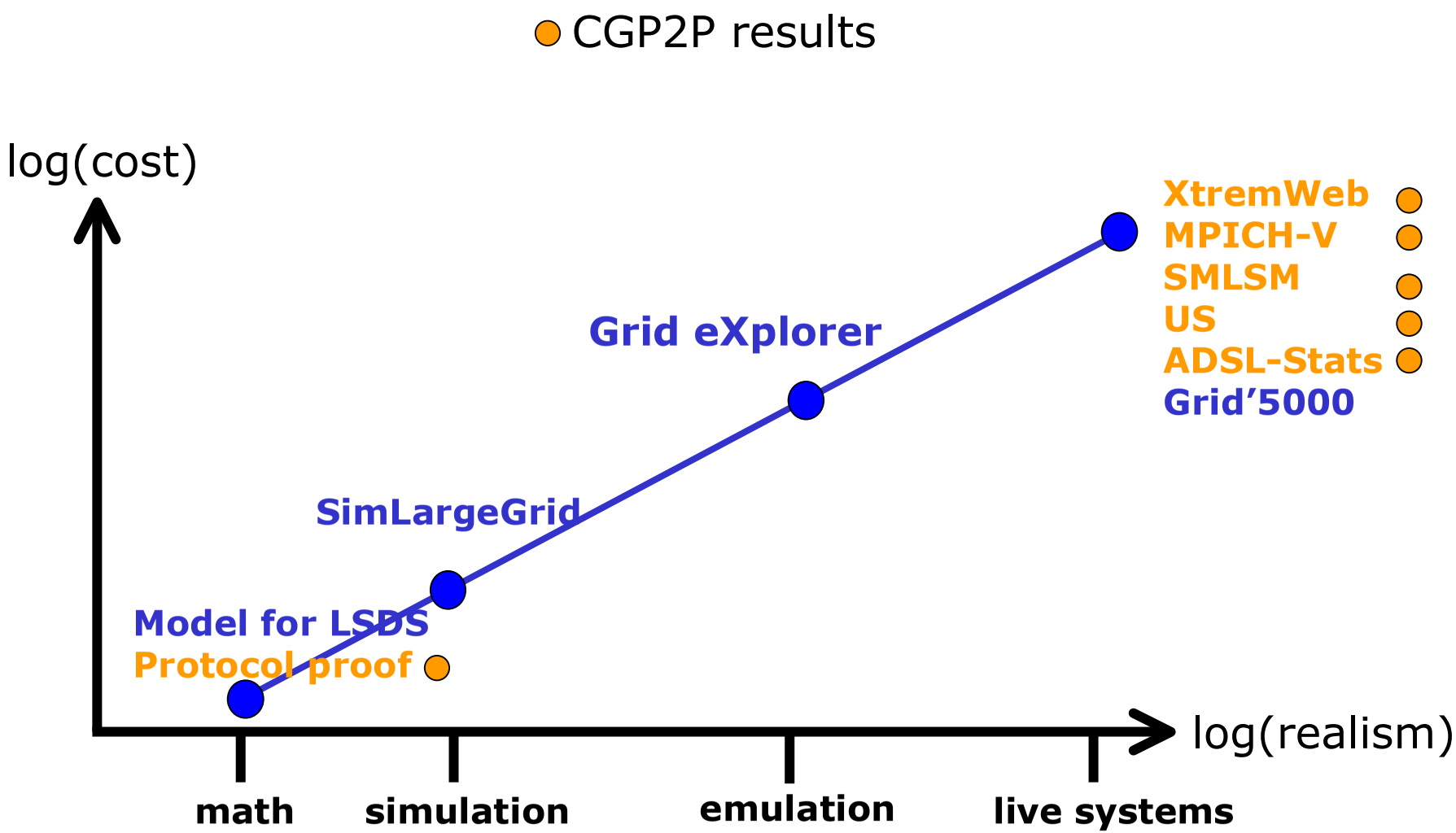
According to the current knowledge, we need:

1) New tools (model, simulators, emulators, experi. Platforms)
2) Strong interaction between research tools

Tools for Large Scale Distributed Systems



ACI Grid CGP2P Contribution

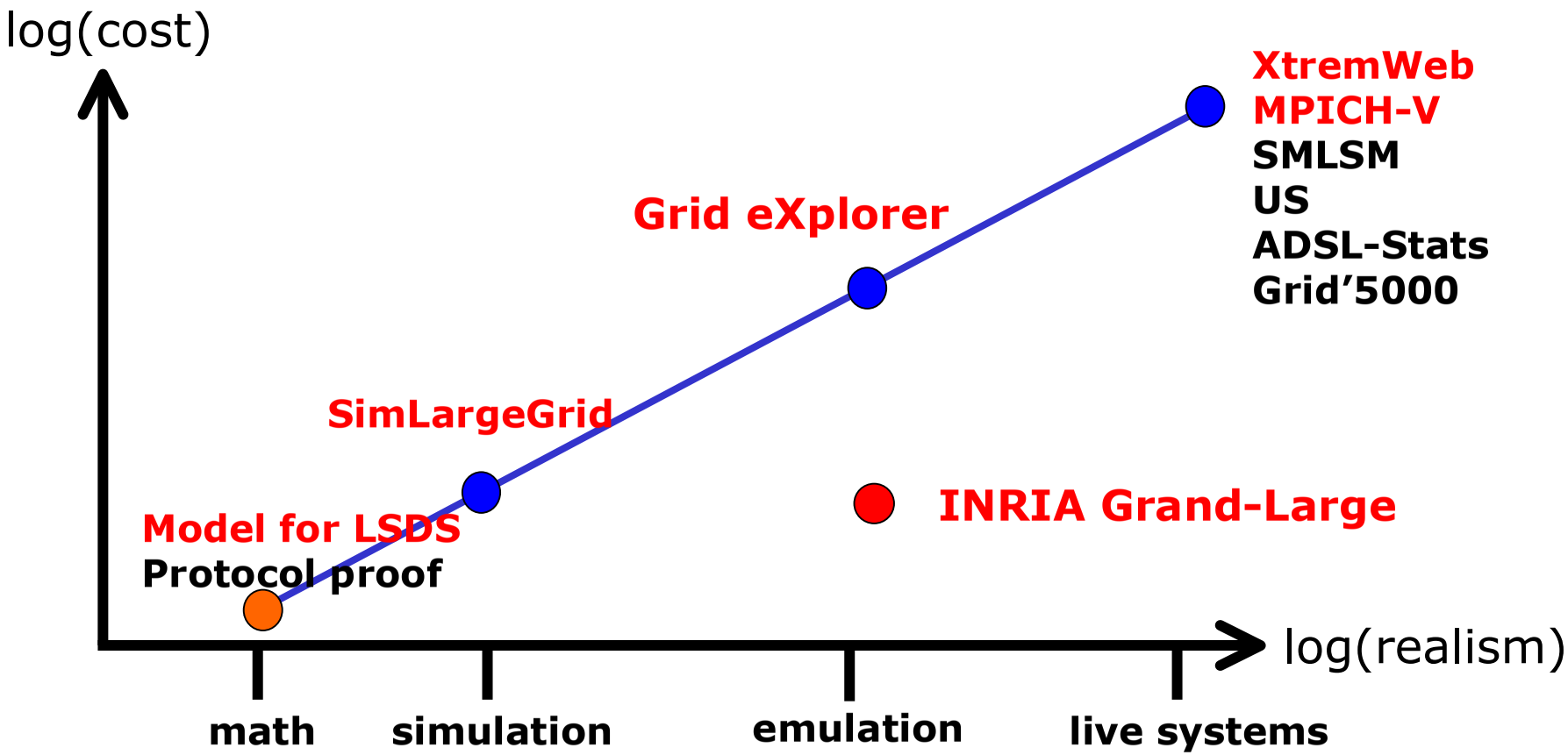


Combining research tools

According to the current knowledge, we need:

1) New tools (model, simulators, emulators, experi. Platforms)
2) Strong interaction between research tools

Tools for Large Scale Distributed Systems



Design of a theoretical model capturing LSDS characteristics

Network:

- ~10 k nodes or larger,
- Wide area network (Network failure, rare but to be considered)
- Standard protocols (TCP/IP)

Nodes:

- Volatile, Byzantine, crash may be permanent

TCP/IP + Very large scale + volatility

- higher levels protocols must be "connexionless" (<500 open connection with select)
  - If a connexion fails, what does it means?
    Either the target is down OR it cannot accept new connexions because all slots are full OR it does not see the incoming SYN message due to high network traffic
  - When a connexion is broken, what does it means?
    Etc.

Design of a theoretical model capturing LSDS characteristics

Current issues:

- LSDS systems seem to fall into the category of asynchronous systems! (consensus impossibility)

- Can fundamental mechanisms of LSDS systems be designed without requiring consensus?

→An interesting strategy would be to consider for each node an "horizon". Concensus would be guaranteed only inside this horizon.

These questions are not trivial!

Workshop:

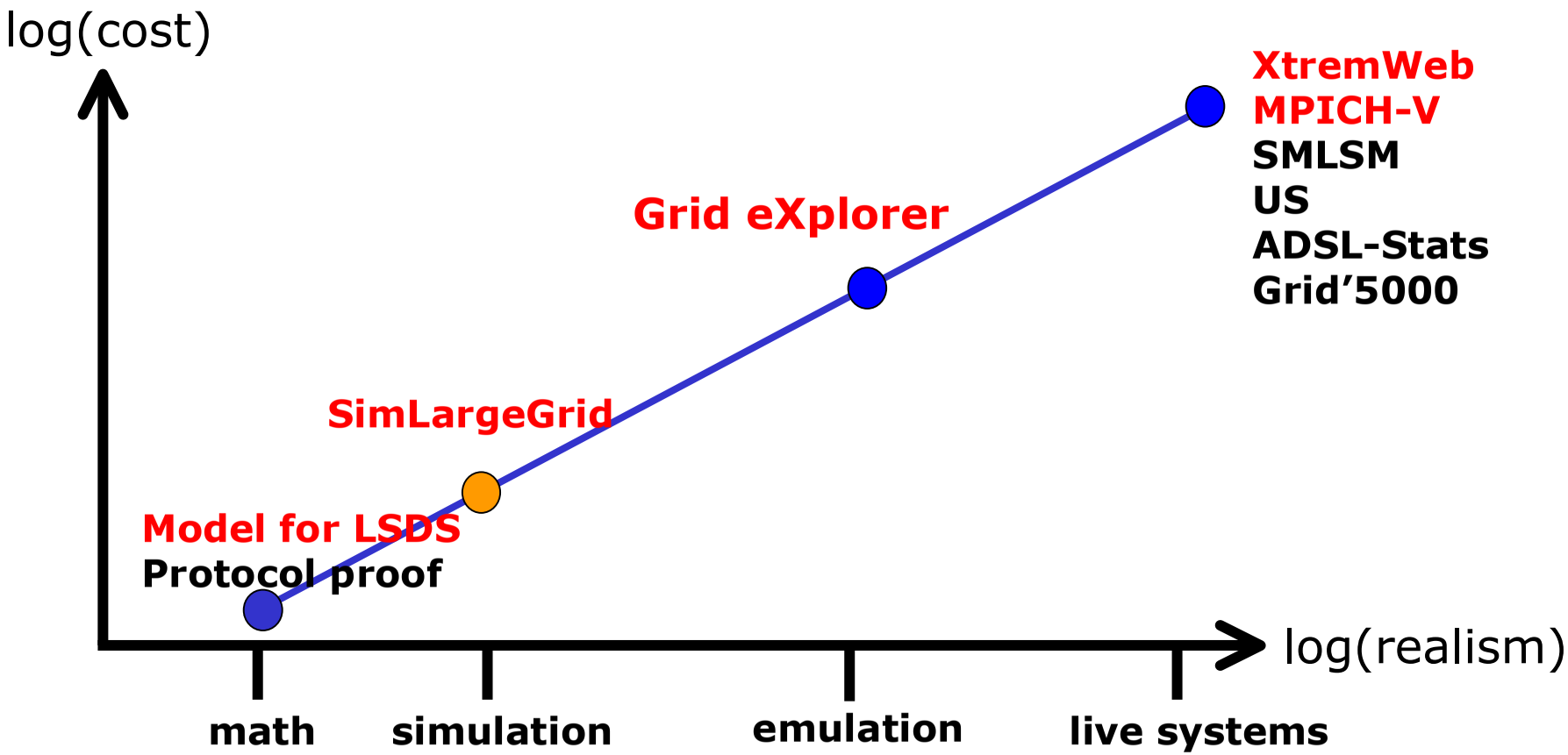
Hugues Fauconnier, Carole Delporte (Paris 7), Joffroy Beauquier, Franck Cappello, Colette Johnen, Sébastien Tixeuil, Thomas Herault (Paris 11)

Combining research tools

According to the current knowledge, we need:

1) New tools (model, simulators, emulators, experi. Platforms)
2) Strong interaction between research tools

Tools for Large Scale Distributed Systems



SimLargeGrid: Large Scale Nearest NeighborScheduling Simulator

Global coordination seems very difficult at large scale (Hierarchical solutions exist and may fit).

More speculative approaches based on autonomous decisions, self organization are also good candidates.

Investigate this last idea with a concrete mechanism: Scheduler/Load balancer (SimGrid, Bricks, GriSim don't scale)

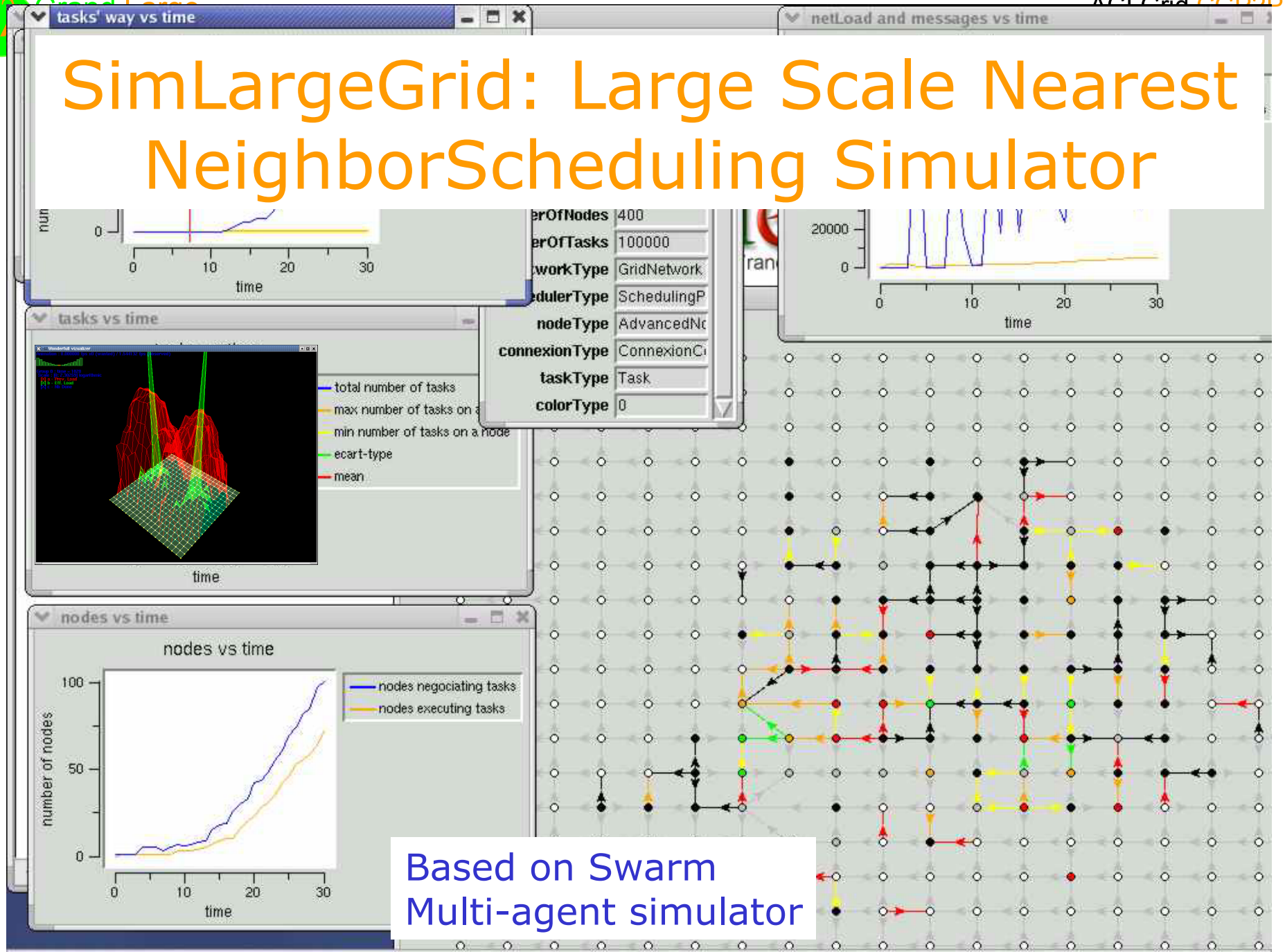
→Current status:

a simulation tool:

topology, volatility, asynchrony, latency/BW, heterogeneity

+ nearest neighbor scheduling algorithms
+ use the tool to compare them.

SimLargeGrid: Large Scale Nearest NeighborScheduling Simulator



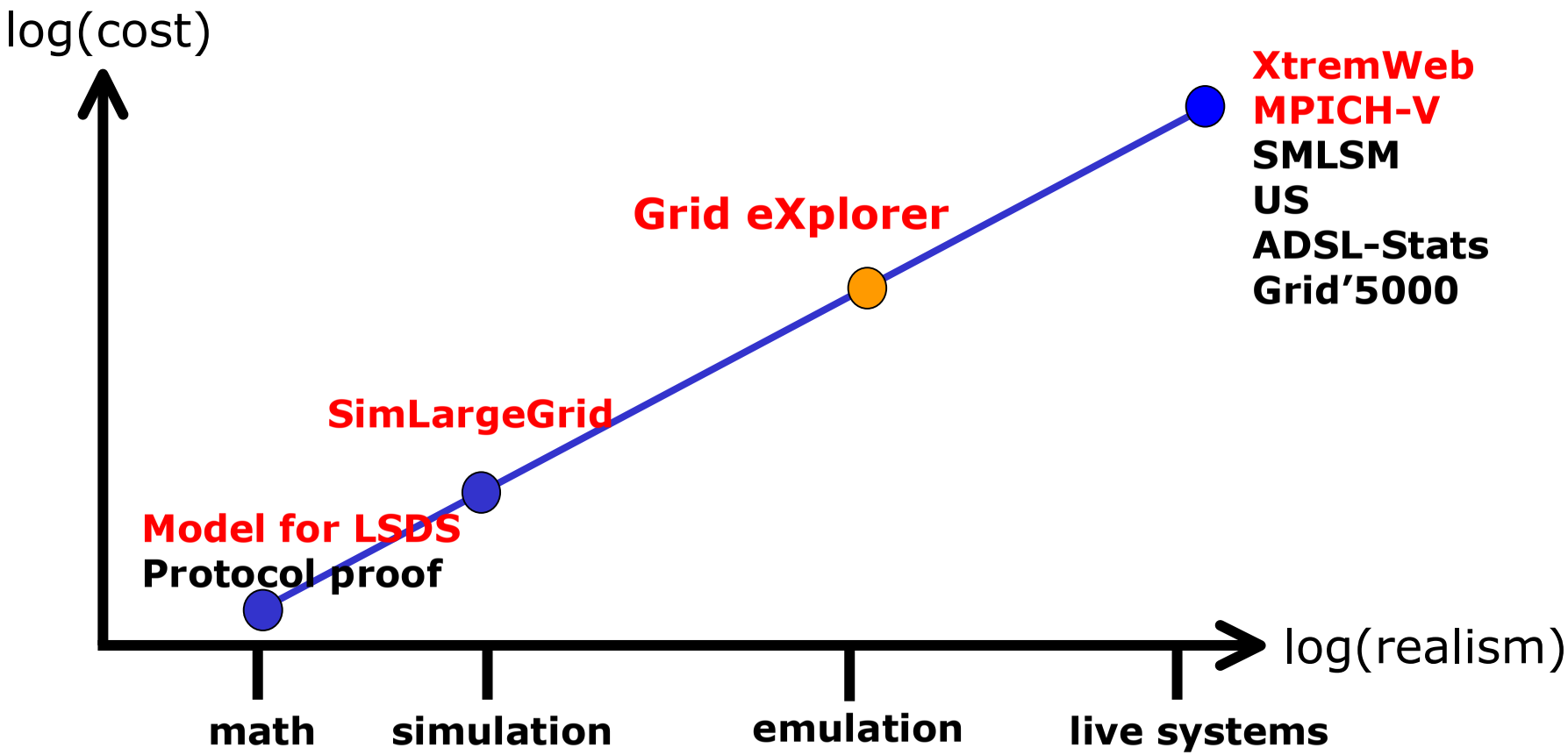
Based on Swarm
Multi-agent simulator

Combining research tools

According to the current knowledge, we need:

1) New tools (model, simulators, emulators, experi. Platforms)
2) Strong interaction between research tools

Tools for Large Scale Distributed Systems



Grid eXplorer

A "GRIDinLAB" instrument for CS researchers
Founded by the French ministry of research through the ACI "Data Mass" incentive + *INRIA*

For

- Grid/P2P researcher community
- Network researcher community

→Addressing specific issues of each domain
→Enabling research studies combining the 2 domains
→Ease and develop collaborations between the two communities.

Statistics:

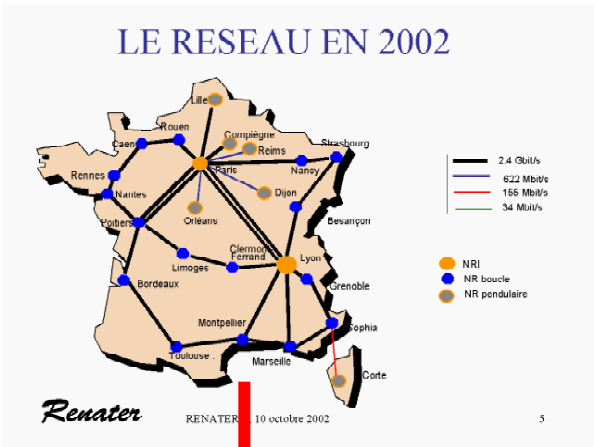
13 Laboratories
80 researchers
24 Research Experiments
>1M€ (not counting salaries)
Installed at IDRIS (Orsay)

Grid eXplorer: the big picture

- 13 Laboratories
- 80 researchers

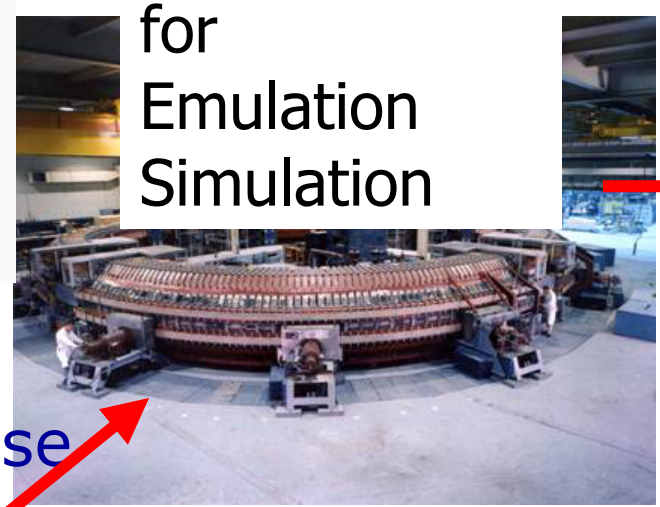
Close to Emulab and WaniLab

A set of sensors

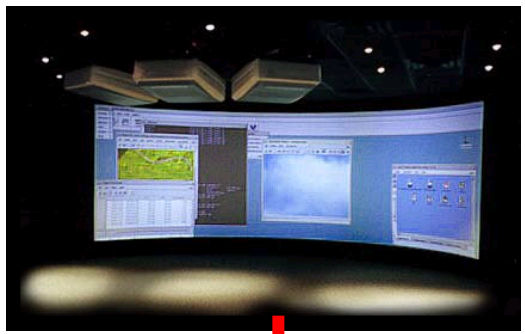


Emulator Core

Hardware + Soft for Emulation Simulation

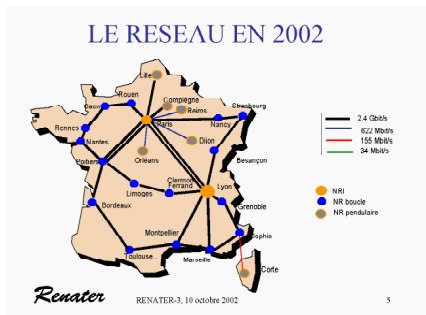


A set of tools for analysis



An experimental Conditions data base

Validation on Real life testbed



Grid eXplorer (GdX) current status:

- First stage: Building the Instrument
  - First GdX meeting was on September 16, 2003.
  - Hardware design meeting planned for October 15.
  - Hardware selection meeting on November 8

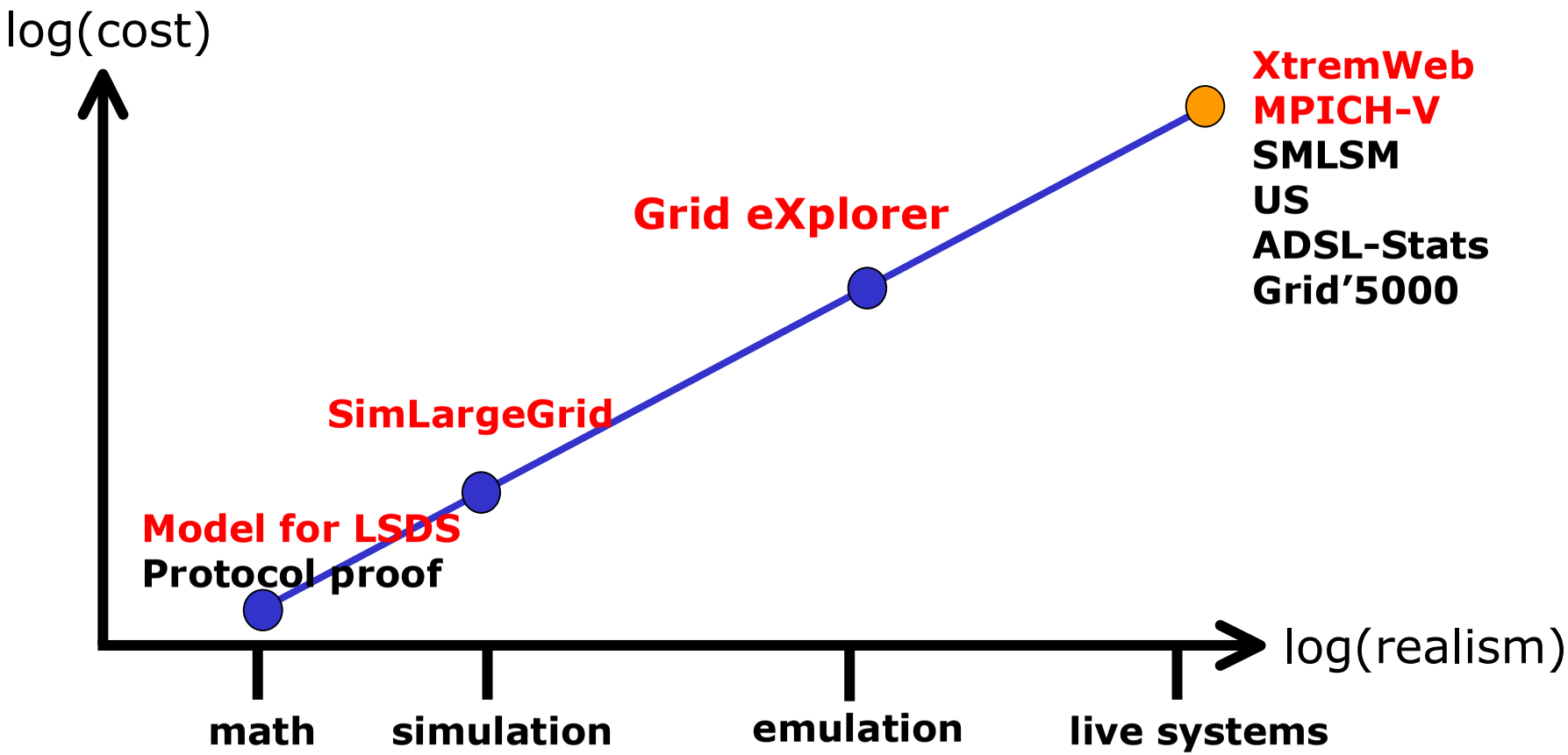
  - Choosing the nodes (single or dual?)
  - Choosing the CPU (Intel IA 32, IA64, Athlon 64, etc.)
  - Choosing the experimental Network (Myrinet, Ethernet, Infiniband, etc.)
  - Choosing the general experiment production architecture (parallel OS architecture, user access, batch scheduler, result repositoty)
  - Choosing the experimental database harware
  - Etc.

Combining research tools

According to the current knowledge, we need:

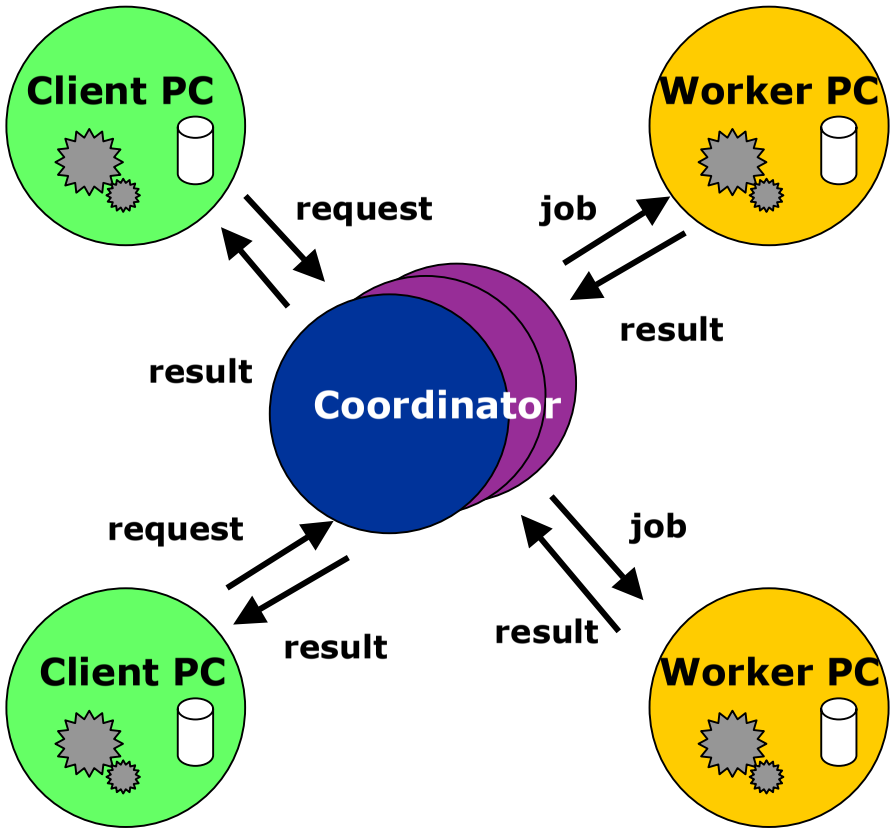
1) New tools (model, simulators, emulators, experi. Platforms)
2) Strong interaction between research tools

Tools for Large Scale Distributed Systems



XtremWeb

Middleware for Desktop Grid Computing



For research on DGrid:

- Scalability
- Fault tolerant
- Programming models
- GridRPC
- Security (Sandbox)
- Scheduling
- DGrid Services
- Deployment (Firewall/ NAT/Proxy bypass)

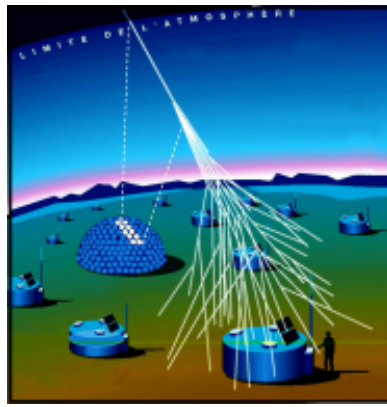
Main international users:

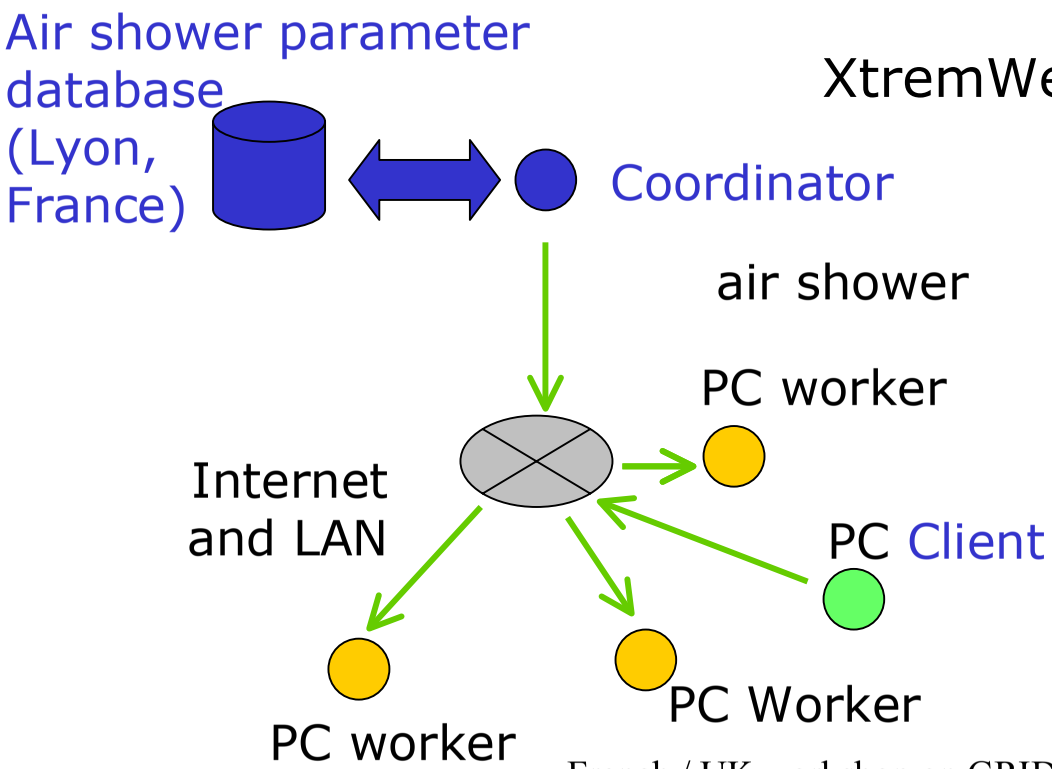
- UCSD (Chien, Casanova)
- U. Tsukuba (Sato)
- U. Geneva (Abdenader)

Production Example: XtremWeb-Auger

Understanding the origin of very high cosmic rays:

- Aires: Air Showers Extended Simulation
  - Sequential, Monte Carlo. Time for a run: 5 to 10 hours

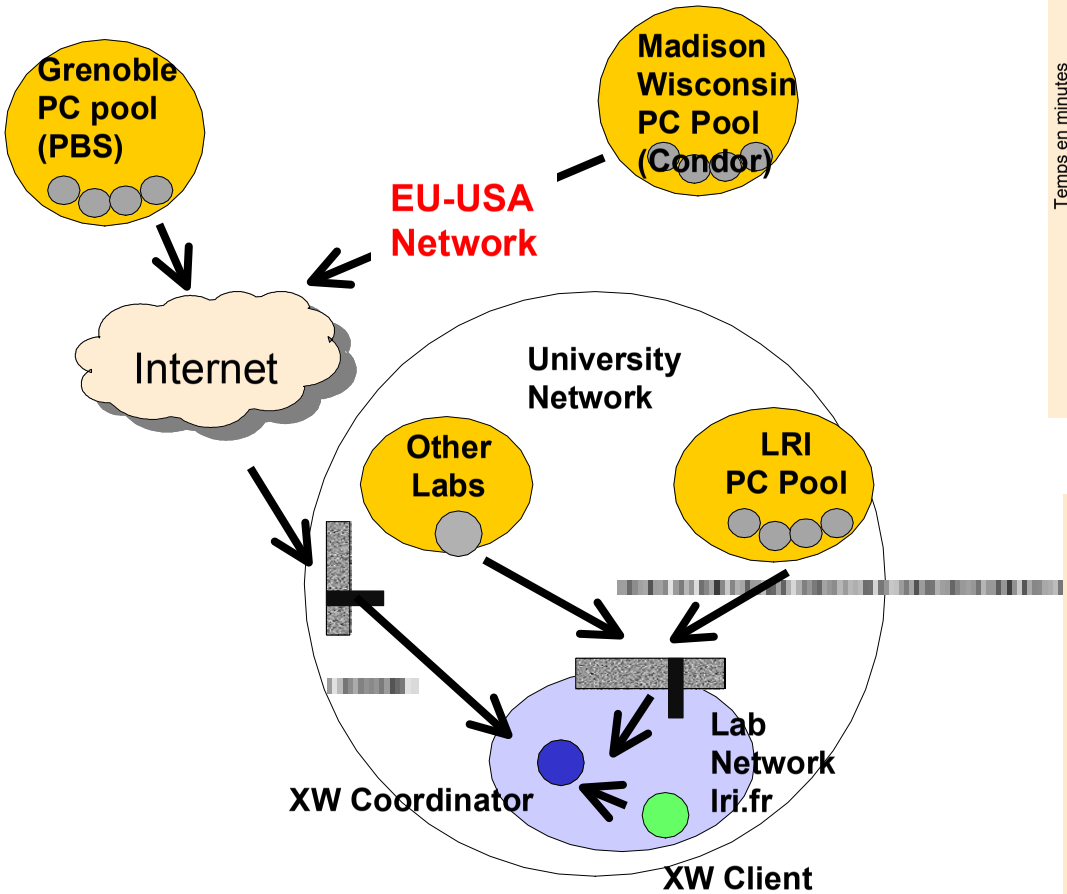




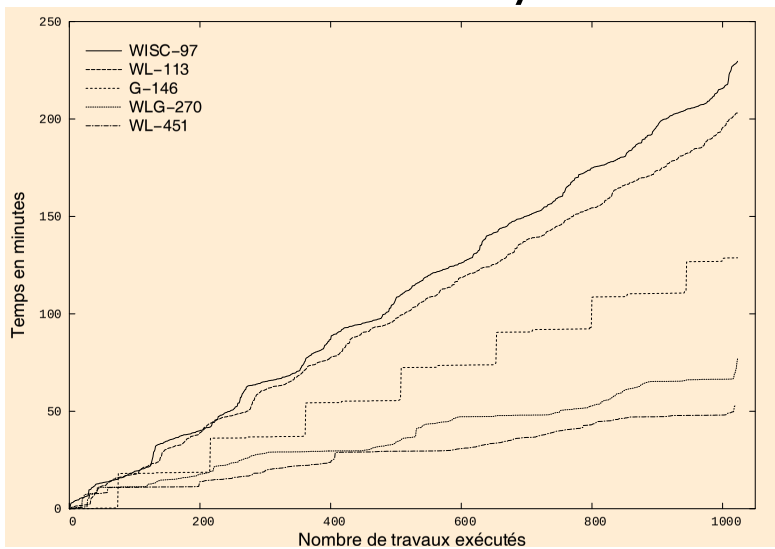
- Tasks are submited form
  - params. Database
  - users
- Estimated PC number: ~ 5000
- Production should start by the fall December
- Result certification by replication

XtremWeb-Testbed

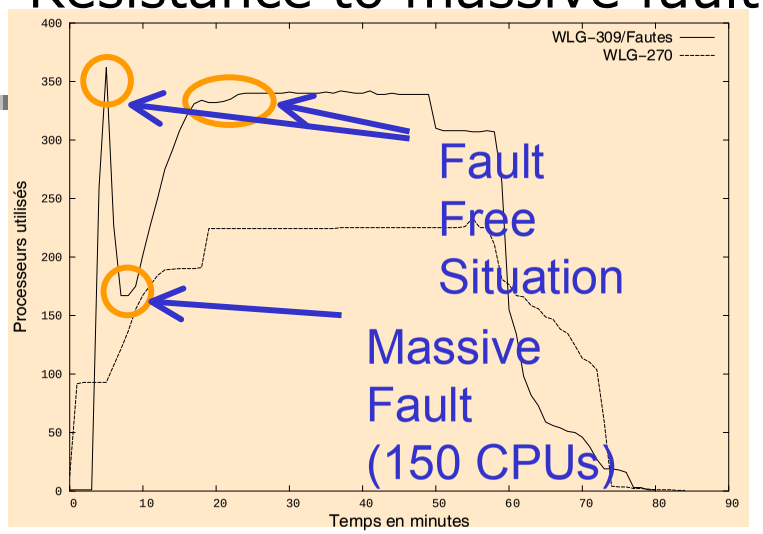
About 1K CPUs



Scalability



Resistance to massive fault



MPICH-V

MPI Implementation for Volatile resources

Toward an automatic/scalable fault tolerant MPI for Clusters & Grids

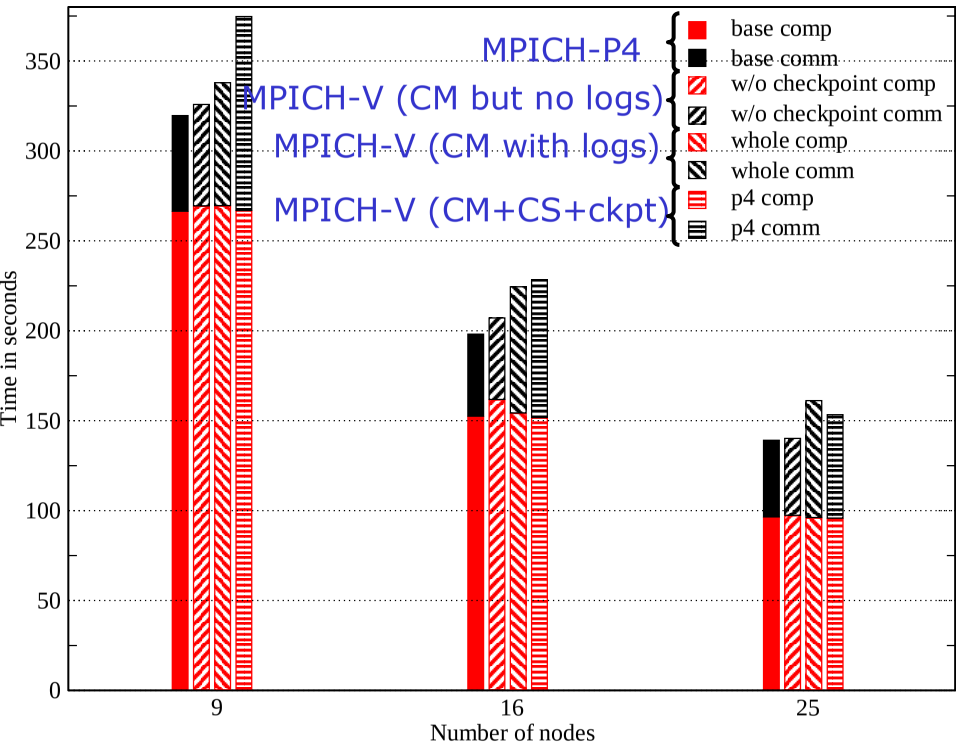
MPICH-V is a research effort

- with theoretical studies,
- experimental evaluations,
- pragmatic implementations,

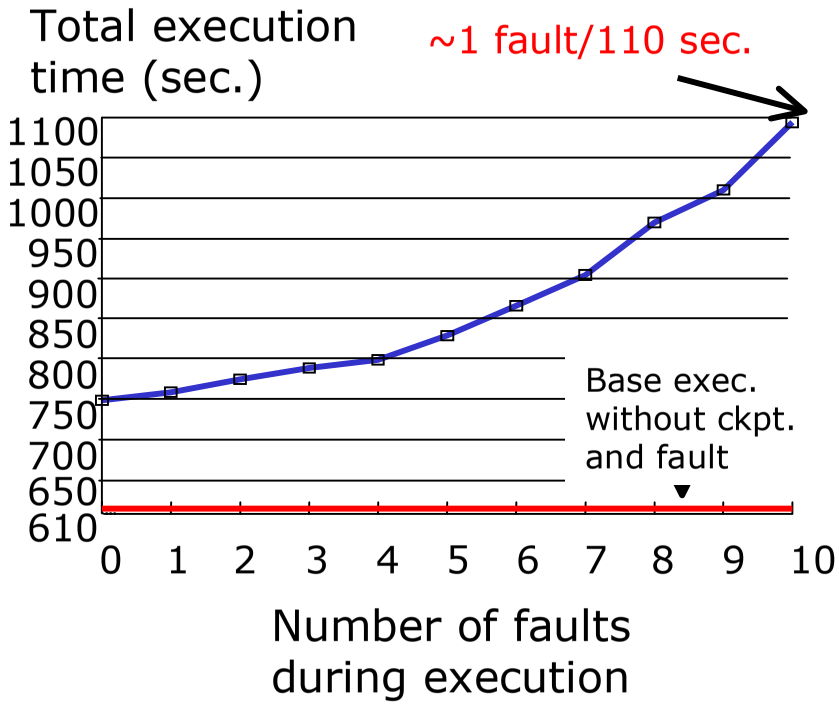
aiming to provide a MPI implementation based on MPICH, featuring multiple fault tolerant protocols (3 currently), for Desktop Grids, Large Clusters and Grids

Main Results

MPICH-V vs. MPICH-P4 BT.A



Execution time with faults (Fault injection) BT.A.9



Performance similar to MPICH-P4
Resistance to a very frequent faults

Other work and Conclusion

Many of the CGP2P Participants are also involved in:

- Grid'5000
- CoreGrid (NoE) proposals for FP6

Summary:

We are involved in different projects related to large scale distributed systems:

- From Theoretical Studies to Actual Grid Deployments
- About Fault Tolerance and Performance
- Middleware Design and Implementation: XtremWeb, MPICH-V
- Large Scale Experimental Platforms: Grid eXplorer, Grid'5000

Contact: fci@lri.fr

Links

ACI Grid CGP2P:	www.lri.fr/~fci/CGP2P
XtremWeb:	www.XtremWeb.net
MPICH-V:	www.lri.fr/~gk/MPICH-V
Grid eXplorer:	www.lri.fr/~fci/GdX
eGrid'5000:	www.lri.fr/~fci/AS1

Grid eXplorer
4 Research Topics

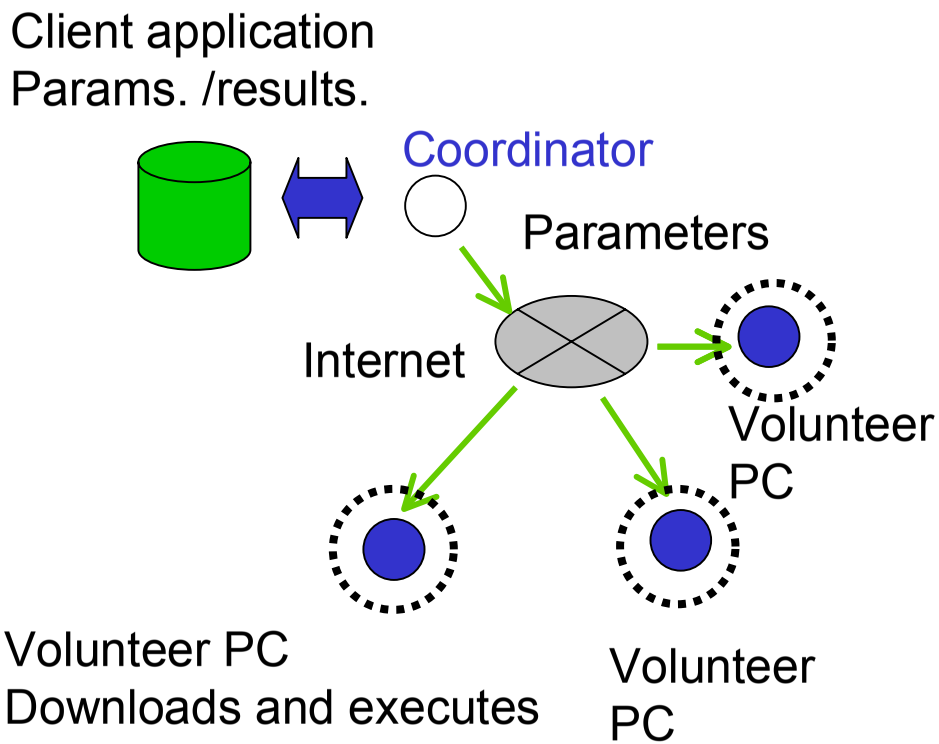
The 4 research topics and their leaders:

-Infrastructure (Hardware + system),	Olivier Richard (ID-IMAG)
-Emulation,	Pierre Sens (LIP6)
-Network,	Pascale Primet (LIP, Inria RESO)
-Applications.	Christophe Cérin (Laria)

Experiences	Infrastructure	Emulation	Network	Application
I.1 Platform	X	X	X	X
I.2 Virtual Grid		X	X	
I.3 Virt. Techniques	X		X	
I.4 Emul driven Simul		X		
I.5 Network.	X	X	X	
I.6 Heterogeneity emul		X		
I.7 Communication				X
I.8 Internet Emul.	X	X	X	
II.1 Engineering tech.		X	X	X
II.2 Mobile objects	X	X		
II.3 Fault tolerance		X	X	
II.4 DHT.		X		
II.5 Data base	X			X
II.6 Scheduling		X		X
II.7 Comm. Optimizat.		X		
II.8 Data sharing		X		X
II.9 Uni and multicast		X	X	
II.10 Cellul. automaton		X		X
II.11 Bioinformatique				X
II.12 P2P storage			X	X
II.13 Reliability		X	X	X
II.14 Security		X	X	X
II.15 NG. Internet	X	X	X	
II.16 Grid coupled sys.				X

Desktop Grids

A central coordinator schedules tasks on volunteer computers,
Master worker paradigm,
Cycle stealing

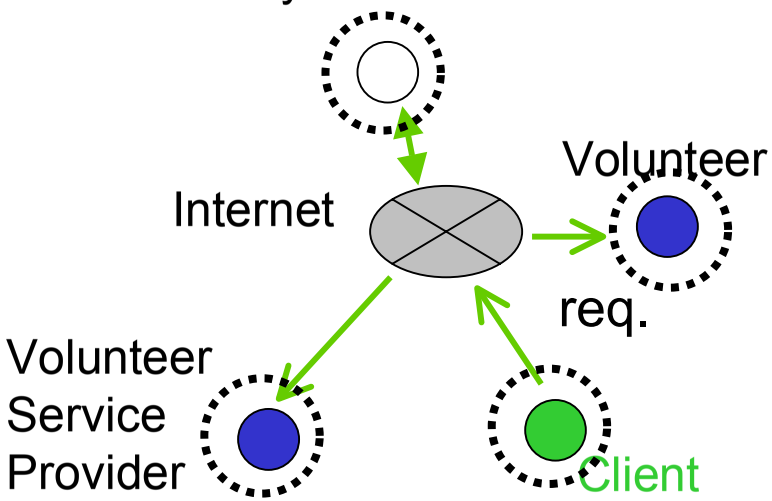


- Dedicated Applications
  - SETI@Home, distributed.net,
  - Décrypthon (France)
- Production applications
  - Folding@home, Genome@home,
  - Xpulsar@home,Folderol,
  - Exodus, Peer review,
- Research Platforms
  - Javelin, Bayanihan, JET,
  - Charlotte (based on Java),
- Commercial Platforms
  - Entropia, Parabon,
  - United Devices, Platform (AC)

Peer to Peer systems (P2P)

All system resources
-may play the roles of client and server,
-may communicate directly
Distributed and self-organizing infrastructure

Volunteer PC participating to the resource discovery/coordination

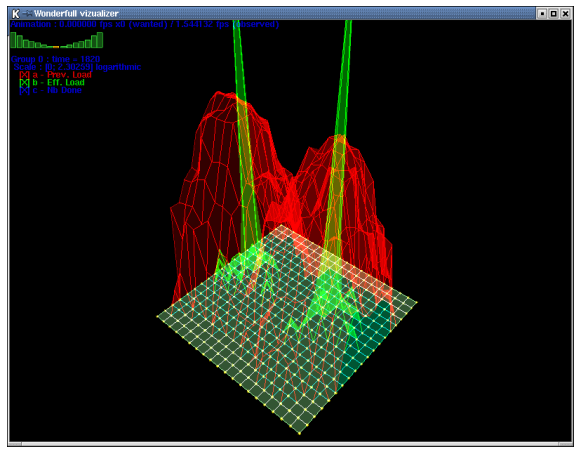


- User Applications
  - Instant Messaging
  - Managing and Sharing Information
  - Collaboration
  - Distributed storage
- Middleware
  - Napster, Gnutella, Freenet,
  - KaZaA, Music-city,
  - Jabber, Groove,
- Research Projects
  - Globe (Tann.), Cx (Javalin), Farsite,
  - OceanStore (USA),
  - Pastry, Tapestry/Plaxton, CAN, Chord,
- Other projects
  - Cosm, Wos, peer2peer.org,
  - JXTA (sun), PtPTL (intel),

Nearest Neighbor Scheduling with a 3D visualization tool

10K tasks on 900 nodes in mesh

- Negotiation (*red movie*)
- Distribution (*blue movie*)
- Execution (*green movie*)



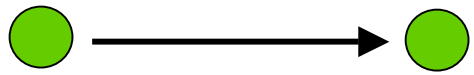
- Observation results:
  - —Symmetry for the negotiation phase
  - —Asymmetry for Distribution and Execution phases.
  - —Waves

Several hours to get 1 movie → parallel simulation is required!

Objectives and constraints

Goal: execute existing or new MPI Apps

Programmer's view unchanged:

PC client MPI_send()  **PC client MPI_recv()**

Problems: 1) volatile nodes (any number at any time)
2) non named receptions (→ should be replayed in the same order as the one of the previous failed exec.)

Objective summary:

1) Automatic fault tolerance
2) Transparency for the programmer & user
3) Tolerate n faults (n being the #MPI processes)
4) Scalable Infrastructure/protocols
5) Avoid global synchronizations (ckpt/restart)
6) Theoretical verification of protocols