# Laboratoire International de Calcul Intensif et d'Informatique Ambiante

# Laboratório Internacional em Computação Intensiva e Informática Ambiente

## International Laboratory in HPC and Ambiant Computing

Jean-Marc.Vincent@imag.fr

Laboratoire d'Informatique de Grenoble,
Inria team MESCAL
University Grenoble-Alpes, France

HOSCAR 3$^{rd}$ Workshop Bordeaux, 2013, September 2-5

# The Licia a Brazilian/French collaboration

1. **A Long Term Collaboration**

2. Licia Infrastructure

3. Main Scientific Axes

4. Scientific Actions of Licia

# A Long Term Collaboration

More than 30 years of scientific collaborations between Porto Alegre and Grenoble scientific institutes

**At every levels :** invited professors, postdoc, PhD, Master (master/doctoral courses), under-grad students (Brafitec programs),

**Last 6 years :**
- more than 150 joint publications
- many joint PhD ( 30 in Parallel and Distributed Computing),
- many research missions (100+)

**International projects :**
- CAPES/COFECUB (10), CNPq/INRIA (2), CNPq/CNRS (1), Équipe Associée(2)
- STIC-AMSUD (3), European Projects (FP7)

**Teams involved :**
- MOAIS and MESCAL (Parallel and Distributed Computing (Inria))
- NanoSim (Multicore) - MAGMA (Artificial Intelligence),
- Steamer (Distributed Multimedia),
- Drakkar (Networking)
- Getalp (Computational Linguistic (TAL))
- other teams

# LICIA: a joint research laboratory

A new step in the collaboration :

A **mutual structure** between INF and LIG to help scientists of partners to **build more ambitious, new, research projects**.

- **Support punctual actions**
- **Deeper pre-existing collaborations**
  HPC, IA, Databases, Information Systems,. . .
- **Broader collaborations:**
  - Internally: new themes (embedded systems, Computer Graphics, software, . . . )
  -Externally: enhance partnerships in other contexts. Brazil/France (HOSCAR, GDRI), EU, Latin America. . .

Participants : members of LIG and UFRGS/INF + external collaborations

# The Licia a Brazilian/French collaboration

1. A Long Term Collaboration

2. **Licia Infrastructure**

3. Main Scientific Axes

4. Scientific Actions of Licia

# LICIA: Status

- The LICIA is a “**Laboratoire International Associé**” (LIA), of the CNRS. scientific tool, 4 years program.
- Includes members of the LIG and UFRGS/INF.
- Involves the French research institutes CNRS and Inria
- Grenoble Universities
- Involves the Brazilian agencies/university:
  - UFRGS, FAPERGS
  - CAPES, CNPq
- A board is in charge of the management of the Licia.

# LICIA: Organization

## Scientific Steering Committee (5+5 searchers)

- Bruno RAFFIN, Computer Graphics, Virtual Reality,
- Hervé GUIOL, Applied Probabilities,
- Jérôme GENSEL, Information Systems,
- José Palazzo M. DE OLIVEIRA, Information Processing,
- Laurent BESACIER, Automatic Processing of Natural Language,
- Luciana NEDEL, Computer Graphics,
- Luigi CARRO, Embedded Systems,
- Marcelo PIMENTA, Interaction Man/Machine,
- Rosa VICARI, Artificial Intelligence,
- Yves DEMAZEAU, Artifical Intelligence.

# LICIA: Organization (2)

## Coordination board

- Philippe O. A. Navaux and Yves Denneulin (directors)
- Nicolas Maillard and Jean-Marc Vincent (co-directors)

with Hervé Martin (LIG) and Flavio Wagner (UFRGS)

# The Licia a Brazilian/French collaboration

1. A Long Term Collaboration

2. Licia Infrastructure

3. **Main Scientific Axes**

4. Scientific Actions of Licia

# Scientific Scope

## Infra-structure and Middleware

- High-Performance Computing,
- Embedded Systems,
- Sensor Networks, . . .

## Algorithms, Software and Programming

- Models and Methods for HPC,
- Modeling and access to information,
- Software and theoretical computer science, . . .

## Interactions and Usage

- Visualization and Interactive Simulations,
- Interaction/Perception,
- Interactive Agents and Computer-Aided Learning, . . .

# High Performance Computing

Some keywords

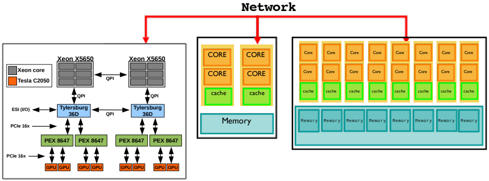
- **HPC and Low-power** *Jean-Francois Mehaut* Performance Analysis of HPC Applications on Low-Power Embedded Platforms
- **I/O and Large Scale Systems** *Rodrigo Kassik* Performance Analysis of Parallel File System: Application Characterization and Benchmarks
- **Runtime for multi-core** Bruno Raffin Efficient Multi-core Programming and In-Situ Result Analysis
- **Simulation and Visualization of Large Scale Execution** Lucas Schnorr and Arnaud Legrand
- **Performance Evaluation of Large-Scale Systems** Jean-Marc Vincent Paulo Fernandes (PUCRS)
- **Scheduling** Denis Trystram (also with USP)
- **...**

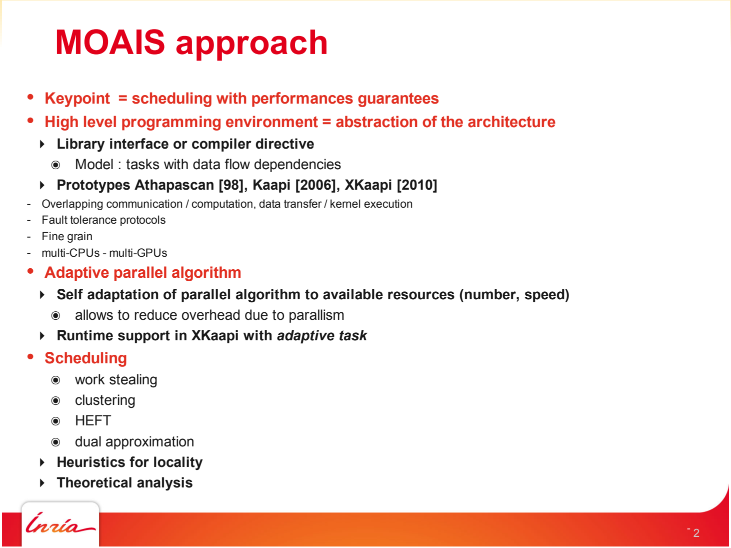
# Context: heterogeneous architecture

- **Complex architecture**
  - **Computing resources**
    - SIMD Units, CPU Core, GPU, Intel Xeon Phi,..., MPPA, FPGA, DSP ...
  - **Memory**
    - hierarchical memory

- registers, cache L1, L2, L3, local memory bank, global memory, remote memory
    - private / shared cache
  - **Interconnection networks**
    - between cores & memory = memory network (HyperTransport, QuickPath, PCI, ...)
    - between machines (Ethernet, InfiniBand,...)

➡ **High complexity**
  - **million of components**
  - **heterogeneous**
    - memory access
    - computing capability
  - **failure**



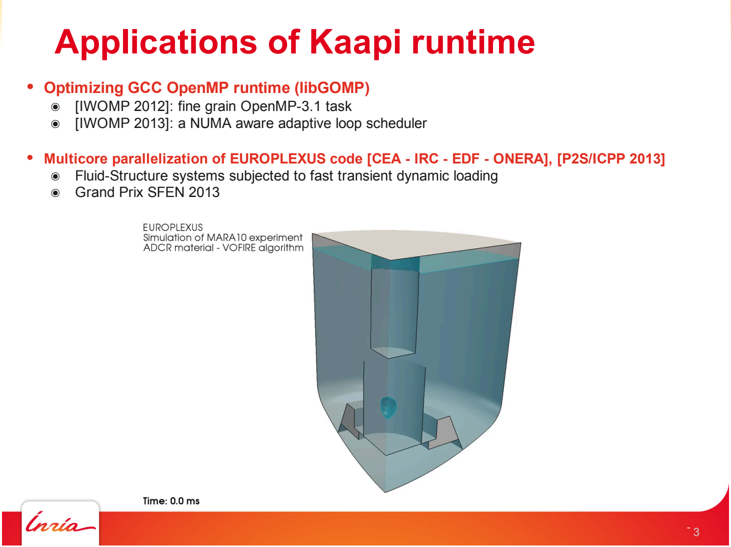


# MOAIS approach

- Keypoint = scheduling with performances guarantees
- High level programming environment = abstraction of the architecture
  - **Library interface or compiler directive**
    - Model : tasks with data flow dependencies
  - **Prototypes Athapascan [98], Kaapi [2006], XKaapi [2010]**

- Overlapping communication / computation, data transfer / kernel execution
- Fault tolerance protocols
- Fine grain
- multi-CPUs - multi-GPUs

- Adaptive parallel algorithm
  - **Self adaptation of parallel algorithm to available resources (number, speed)**
    - allows to reduce overhead due to parallism
  - **Runtime support in XKaapi with *adaptive task***
- Scheduling
    - work stealing
    - clustering
    - HEFT
    - dual approximation
  - **Heuristics for locality**
  - **Theoretical analysis**



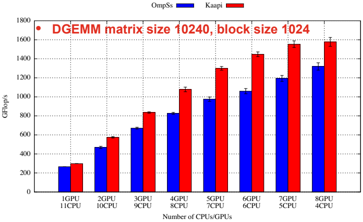
# Applications of Kaapi runtime

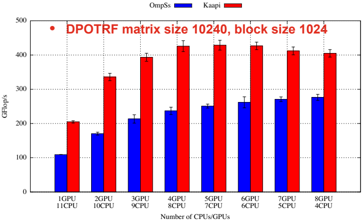
- **Optimizing GCC OpenMP runtime (libGOMP)**
  - [IWOMP 2012]: fine grain OpenMP-3.1 task
  - [IWOMP 2013]: a NUMA aware adaptive loop scheduler
- **Multicore parallelization of EUROPLEXUS code [CEA - IRC - EDF - ONERA], [P2S/ICPP 2013]**
  - Fluid-Structure systems subjected to fast transient dynamic loading
  - Grand Prix SFEN 2013

EUROPLEXUS
Simulation of MARA10 experiment
ADCR material - VOFIRE algorithm
Time: 0.0 ms
Inria
3

# Current results

- Heterogeneous architecture
  - **[OMPSS], [StarPU]**
  - **Interest in exploitation of high number of GPUs per node**
    - [SBAC-PAD 2012, Multiprog 2013, IPDPS 2013] on 8GPUs Fermi, 12 cores machine
- up to 2.43 TFlops DP on // DGEMM, 5.09TFlops SP
- up to 1.79 TFlops DP on // Cholesky factorization, 3.92TFlops SP
    - Low overhead, overlapping data transfer / kernel computation


- DGEMM matrix size 10240, block size 1024


- DPOTRF matrix size 10240, block size 1024

# Analysis of Large-Scale Executions

### Applications

NAS-DT Benchmark Analysis

- Traces obtained with the Simulated MPI (SMPI) tool

Non-cooperative Master Worker Applications

- Traces from the SimGrid toolkit

### Objective of the evaluation

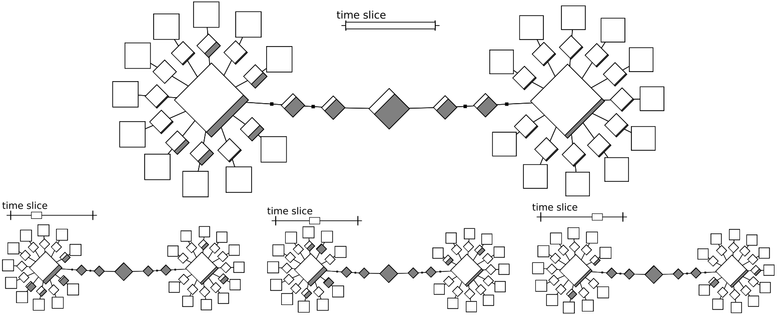
- Illustrate the potential of the approach
- Pin-point network contention
- Evaluate load balancing

### Trace metrics mapping

- Host are squares
- Network links are diamonds
- Filling is resource utilization

# NAS-DT Benchmark Analysis

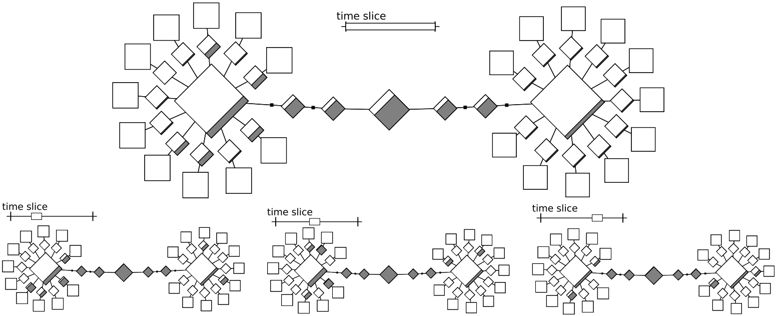
- Scenario configuration
  - NAS-DT Class A White Hole algorithm
  - Two clusters: Adonis (left) and Griffon (right)
  - Processes are allocated sequentially starting at Adonis
- Analysis
  - Central backbone is always the point of contention
  - Changing to smaller time-slices allows us to observe that



- Possible explanation: bad process deployment of MPI

# NAS-DT Benchmark Analysis

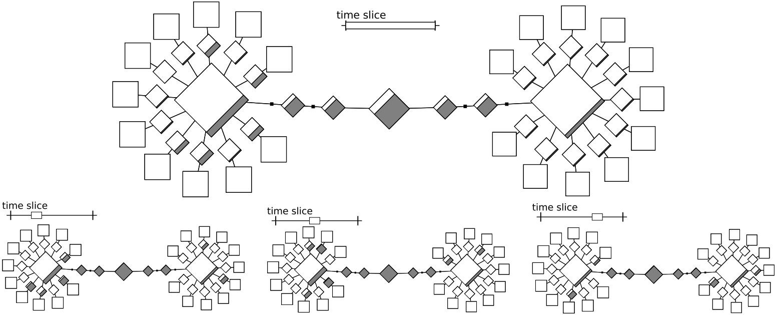
- Scenario configuration
  - NAS-DT Class A White Hole algorithm
  - Two clusters: Adonis (left) and Griffon (right)
  - Processes are allocated sequentially starting at Adonis
- Analysis
  - Central backbone is always the point of contention
  - Changing to smaller time-slices allows us to observe that



- Possible explanation: bad process deployment of MPI

# NAS-DT Benchmark Analysis

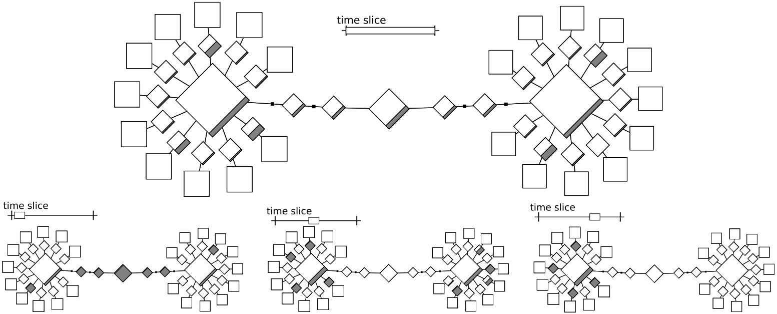
- Scenario configuration
  - NAS-DT Class A White Hole algorithm
  - Two clusters: Adonis (left) and Griffon (right)
  - Processes are allocated sequentially starting at Adonis
- Analysis
  - Central backbone is always the point of contention
  - Changing to smaller time-slices allows us to observe that



- Possible explanation: bad process deployment of MPI

# NAS-DT Benchmark Analysis (Best Deployment)

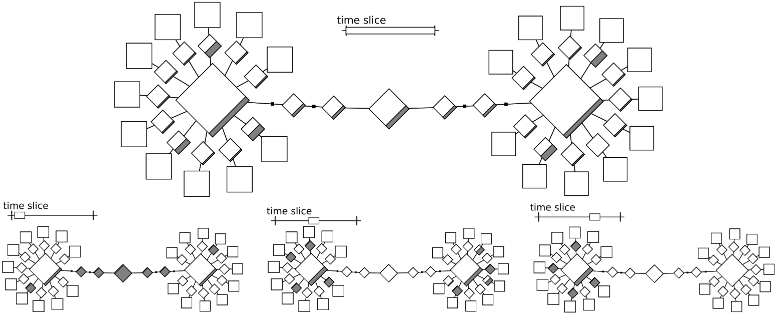
- New execution with a better deployment
  - Exploit locality, avoid backbone link
- Analysis
  - Central backbone is no longer a contention point
  - Smaller time-slices shows that it is only at the beginning



- Reduction of execution time of approximately 20%

# NAS-DT Benchmark Analysis (Best Deployment)

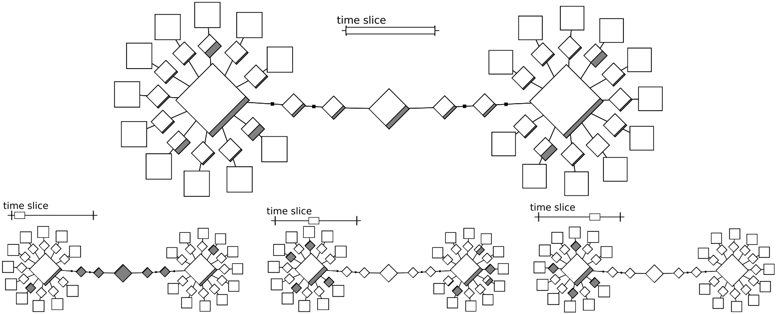
- New execution with a better deployment
  - Exploit locality, avoid backbone link
- Analysis
  - Central backbone is no longer a contention point
  - Smaller time-slices shows that it is only at the beginning



- Reduction of execution time of approximately 20%

# NAS-DT Benchmark Analysis (Best Deployment)

- New execution with a better deployment
  - Exploit locality, avoid backbone link
- Analysis
  - Central backbone is no longer a contention point
  - Smaller time-slices shows that it is only at the beginning



- Reduction of execution time of approximately 20%

# Non-Cooperative Master Worker Applications

Scenario configuration

- Two master-worker applications competing for resources
- Based on a realistic model of the Grid5000 platform
  → 2170 computing hosts + network links and routers
- Each application uses a bandwidth-centric optimal strategy

Tasks layout for each application

- First application: tasks are CPU bound → cheap to transfer, hard to calculate
- Second application : tasks are network bound → tasks are a little harder to transfer

Mapping trace metrics

- Each application has its own resource utilization variable
- First application (light gray) Second application (dark gray)

Three Expected Behaviors

- First application should achieve better resource utilization
- Form of locality from the second application
- Applications may interfere on computing resources

# Non-Cooperative Master Worker Applications

Scenario configuration

- Two master-worker applications competing for resources
- Based on a realistic model of the Grid5000 platform
  → 2170 computing hosts + network links and routers
- Each application uses a bandwidth-centric optimal strategy

Tasks layout for each application

- First application: tasks are CPU bound → cheap to transfer, hard to calculate
- Second application : tasks are network bound → tasks are a little harder to transfer

Mapping trace metrics

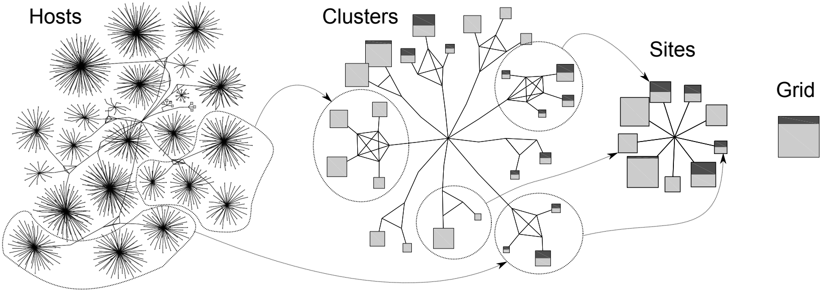
- Each application has its own resource utilization variable
- First application (light gray) Second application (dark gray)

Three Expected Behaviors

- First application should achieve better resource utilization
- Form of locality from the second application
- Applications may interfere on computing resources

# Non-Cooperative Applications – Spatial

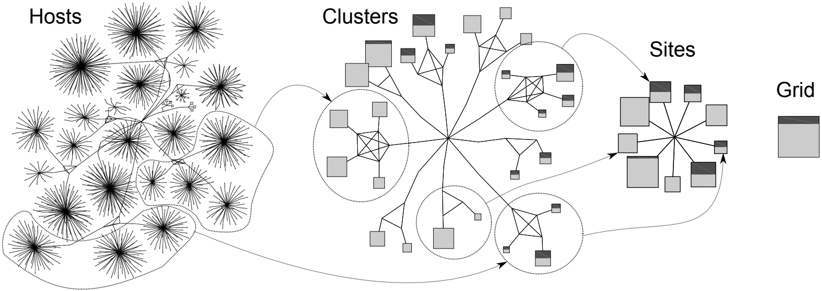
- Fixed temporal aggregation, changing spatial aggregation
- From left to right
  Hosts (no aggregation), Clusters, Sites, Grid (full aggregation)



- Analysis
  - Host level is too cumbersome
  - Cluster, resource usage favors the CPU-bound application
  - Site, enables to quantify perfectly how much it is
  - Grid, gives the overview for this particular time slice

# Non-Cooperative Applications – Spatial

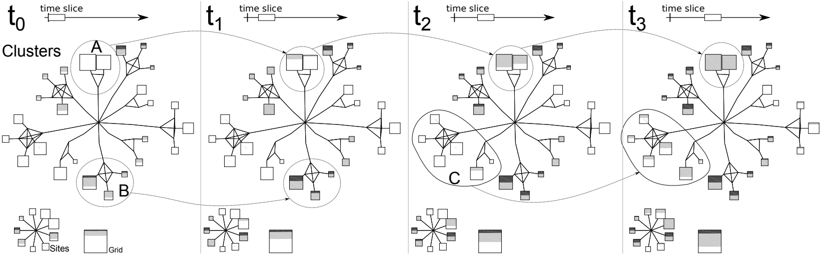
- Fixed temporal aggregation, changing spatial aggregation
- From left to right
  Hosts (no aggregation), Clusters, Sites, Grid (full aggregation)



- Analysis
  - Host level is too cumbersome
  - Cluster, resource usage favors the CPU-bound application
  - Site, enables to quantify perfectly how much it is
  - Grid, gives the overview for this particular time slice

# Non-Cooperative Applications – Temporal

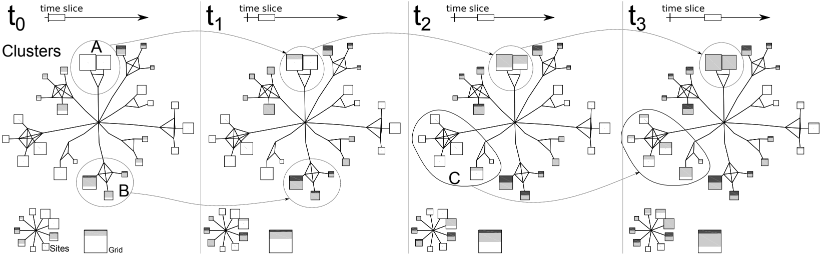
- Fixed spatial aggregation to Cluster level
- From left to right, time slice sliding forward
  - These are actually screenshots from an animation



- Analysis
  - Resource usage is not uniform in the platform for the CPU-bound app
    - $t_3$ shows cluster A full of tasks, but not cluster B
  - Site B is full of tasks at $t_2$
    while Site C has to wait until time slice $t_3$
- Possible explanation: single master process, big platform

# Non-Cooperative Applications – Temporal

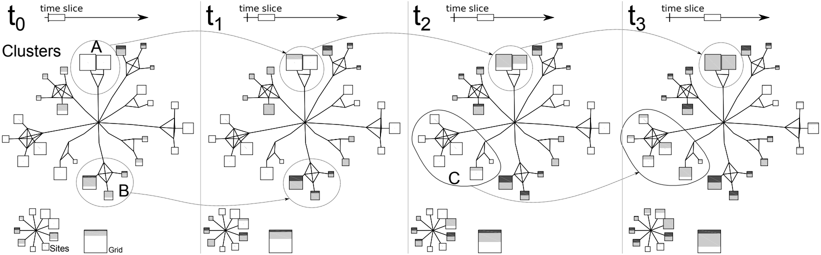
- Fixed spatial aggregation to Cluster level
- From left to right, time slice sliding forward
  - These are actually screenshots from an animation



- Analysis
  - Resource usage is not uniform in the platform for the CPU-bound app
    - $t_3$ shows cluster A full of tasks, but not cluster B
  - Site B is full of tasks at $t_2$
    while Site C has to wait until time slice $t_3$
- Possible explanation: single master process, big platform

# Non-Cooperative Applications – Temporal

- Fixed spatial aggregation to Cluster level
- From left to right, time slice sliding forward
  - These are actually screenshots from an animation



- Analysis
  - Resource usage is not uniform in the platform for the CPU-bound app
    - $t_3$ shows cluster A full of tasks, but not cluster B
  - Site B is full of tasks at $t_2$
    while Site C has to wait until time slice $t_3$
- Possible explanation: single master process, big platform

# The Licia a Brazilian/French collaboration

1. A Long Term Collaboration
2. Licia Infrastructure
3. Main Scientific Axes
4. **Scientific Actions of Licia**

# Summary of Activities in 2012

- **Visits :** Br → Fr 8, Fr → Br 11
- **PhD :** 13 (some in sandwich + double diploma)
- **New collaborations :**
  - Computer Graphics,
  - Embedded Systems
  - Computational Linguistic
- **New projects :**
  - European Projects : Projet FP7 Strep ICT Call EU Brazil OC4ES (Open Cloud for Earth Science) Partenaires EU: ISMB (Italie), INGV (Italie), Deltares (Pays Bas), CNRS (Institut des Grilles et LIG), LMU (Allemagne), UFRGS (Porto Alegre), INPE (Sao Jose do Campos), FURB (Blumenau), CPTEC (Cachoeira Paulista), UFSC (Florianopolis), UFSM (Santa Maria), UNIVAP (Sao Jose do Campos), CENEPAD (Cachoeira Paulista)
  - Joint Laboratory on Petascale Computing (JLPC) UJF-LIG/Nanosim, UFRGS supported by INRIA, UIUC, NCSA and Argone National Laboratory.
  - CAPES-COFECUB : Técnicas para Modelagem e Solução de Alta-Performance para Redes de Autômatos Estocásticos PRiSM/UVSQ Jean-Michel Fourneau et LIG Jean-Marc Vincent, PUC RS Paulo Fernandes et UFRGS-INF Nicolas Maillard
  - CAPES-COFECUB : Starship: scalable tools and algorithms for resilent, scalable hybrid interactive processing. LIG Thierry Gautier UFRGS-INF Lucas Schnorr

# LICIA Workshops

## Organization : every year

- Working groups
- New trends
- Keynote Talks
- PhD student session (master students presentations)
- Steering Committee and Board meeting

## Previous Workshops

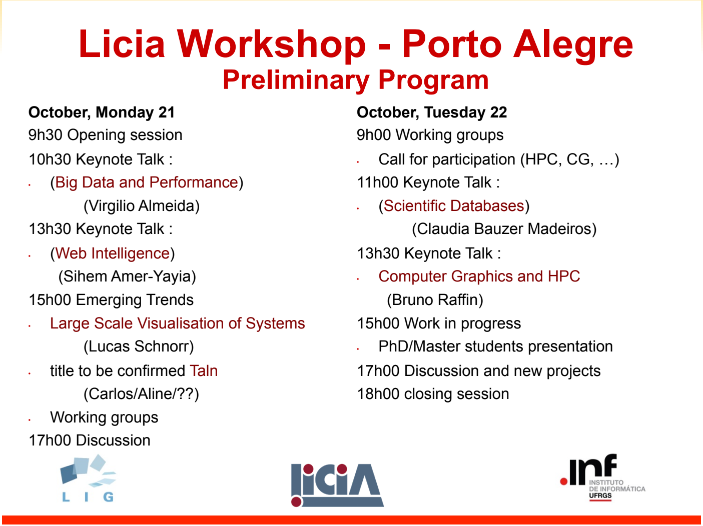
- 2011, October 29-30 Porto Alegre, participants : 11 + 30
- 2012, September 5-7 Grenoble, participants : 18 + 41

## Laboratoire International en Calcul Intensif et Informatique Ambiante

3$^{rd}$ Workshop LICIA
October 21-22, 2013
Porto Alegre

## Laboratório Internacional em Computação Intensiva e Informática Ambiente

.Inf INSTITUTO DE INFORMÁTICA UFRGS



# Licia Workshop - Porto Alegre

## Preliminary Program

**October, Monday 21**

9h30 Opening session

10h30 Keynote Talk :

- (Big Data and Performance)

  (Virgilio Almeida)

13h30 Keynote Talk :

- (Web Intelligence)

  (Sihem Amer-Yayia)

15h00 Emerging Trends

- Large Scale Visualisation of Systems

  (Lucas Schnorr)

- title to be confirmed Taln

  (Carlos/Aline/??)

- Working groups

17h00 Discussion

**October, Tuesday 22**

9h00 Working groups

- Call for participation (HPC, CG, …)

11h00 Keynote Talk :

- (Scientific Databases)

  (Claudia Bauzer Madeiros)

13h30 Keynote Talk :

- Computer Graphics and HPC

  (Bruno Raffin)

15h00 Work in progress

- PhD/Master students presentation

17h00 Discussion and new projects

18h00 closing session