MOAIS

Multi-Programming and Scheduling Design for Applications of Interactive Simulation

Jean-Louis Roch & al.

Moais Trasféré du Musée de l'Homme au
Musée du Quai Branly.
Origin: Rapa Nui [Easter Island], Cook Bay, West coast
Date : between the XIst and the XVth century
Height: 1,85 m.

« Tête sculptée en tuf volcanique. Recueillie en 1872 (…) sur le navire "La Flore(…). Ce buste a été détaché à l'époque du reste du corps devant l'impossibilité de transporter la pièce entière jusqu'au navire. »

http://moais.imag.fr

Visite AERES – 8-10 Février 2010

Staff

- **1/1/2005**: Creation of MOAIS team -- 1/1/2006: INRIA team-project / LIG
  27/3/2008: INRIA Evaluation

- Vincent Danjean [MdC, 9/05]
- Pierre-François Dutot [MdC, 9/06]
- Thierry Gautier [CR INRIA]
- Guillaume Huard [MdC]
- Grégory Mounié [MdC]
- Clément Pernet [MdC, 12/08]
- Bruno Raffin [CR INRIA, HDR]
- Jean-Louis Roch [MdC, Team leader]
- Denis Trystram [Prof]
- Frédéric Wagner [MdC, 9/06]

- 2 Postdocs: Ingo Assenmacher, *Veronika Sonigo*
- 12 PhD students, 2 engineers,
  - 17 PhDs defended (1/10/05->30/9/09) + 3 (L Schnorr, JD Lesage, T Roche)

- Christian Seguy [computing facilities]
- Ahlem Zammit-Boubaker, Annie-Claude Vial-Dallais [administration]

Parallelism everywhere

3

- Distributed, Heterogeneous

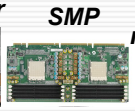
Global (P2P)

BOINC

Grids

Cluster

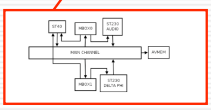
SMP



multi-core

GPU

MPSoC



MPI OpenMP

MapReduce [Google]

Cuda [NVidia]

TBB [Intel]

Cilk++ [CilkArts]

...SPIRIT

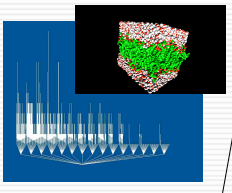
Fortress [Sun]

MOAIS objective

4

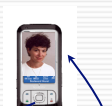
- ***To provide end-to-end parallel programming solutions for high-performance interactive computing with provable performances.***

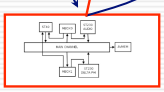
optimization computational steering, VR embedded



FVNano Interaction with molecules
[[LBT, LIFO, CEA/DIF]
QAP/Nugent on Grid'5000
[PRISM, GSCOP, DOLPHIN]

INRIA Grimage platform
[MOAIS, PERCEPTION, EVASION]



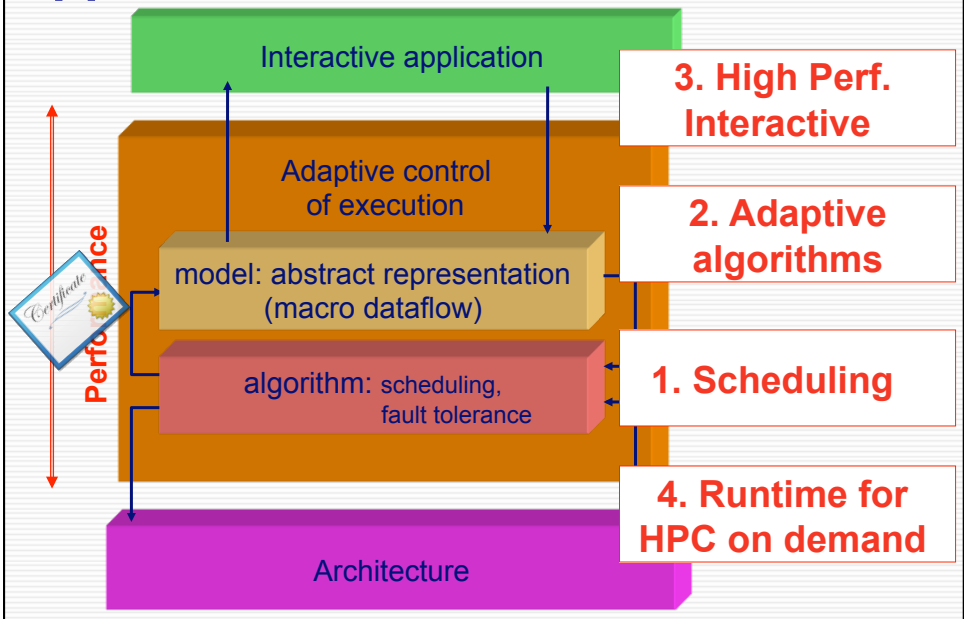


Streaming on MPSoCs
[ST]

Performance is multi-objective

Approach

Research Themes



Achievements for 2005-2009 and research directions

1. **Scheduling**
2. Adaptive algorithms
3. High-Performance Interactive Computation
4. Runtime for HPC on demand

1. Scheduling

- **Formalization of the related problems:**
  - Modeling of an adaptive application
  - Formalization and optimization w.r.t. multi-objective
  - Design of scalable scheduling algorithms

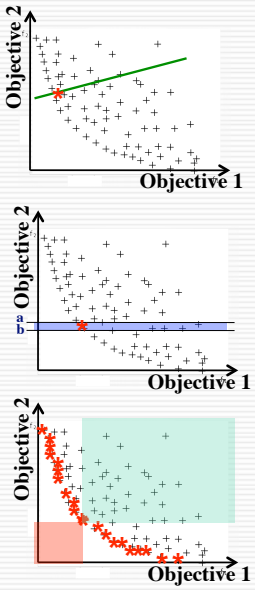
- **Approach**
  - Classical combinatorial optimization
    - complexity and bounds, approximation
  - Non standard methods
    - game theory, distributed analysis

Multi-objective approximation

$$\begin{bmatrix} Min\ \ Objective1 \\ Min\ \ Objective2 \end{bmatrix}$$

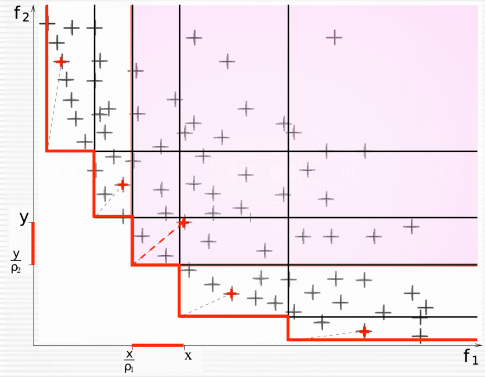
- **Reduction to a single objective:**
  - **To aggregate all objectives into one.**
    - Linear or convex combination to optimize.
  - **To constrain all objectives except one.**
    - E.g. minimize Makespan using Memory < S.

- **Computation of the Pareto set.**
  - **Pareto** : any solution non dominated by another one.



Approximation of the Pareto set

- Pareto set can be large and not convex,
- But can be approximated with a polynomial number of near-optimal solutions [Papadimitriou 2000]



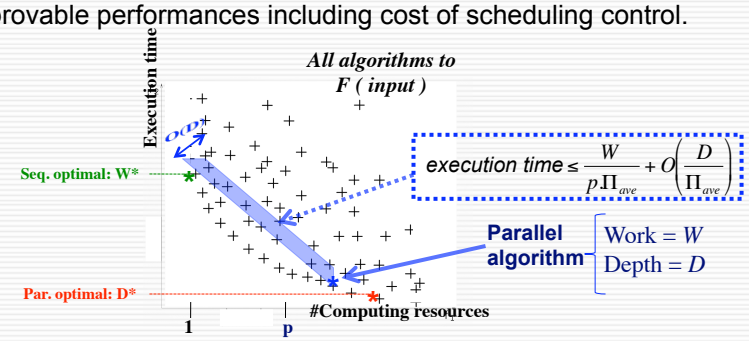
- Applications:
  - Makespan/Memory [IPDPS08]
  - Makespan/Reliability[JPDC09]

Research directions

- **Generic method for non-correlated objectives?**
  - Correlated: MinSum/Makespan : Min ΣC_i / Max C_i
  - Non correlated: Reliability/Makespan : $e^{-\lambda t}$ / Max C_i
- **Dealing with many objectives [diversity]**
  - For k objectives (eg Makespan for each user):
    constant approximation to the Pareto set [IPDPS09]
- **Towards a decentralized control, but keeping in mind a global objective**
  - Instead of global performance, local point of view (Game theory)
  - Consider classical method, but to provide distributed algorithms

List scheduling vs Work-stealing

- **List scheduling:** " Greedy": no idle processors while ready tasks.
  - Many variants to cope with multi-objectives (eg Makespan + Comm. Amount)
- **Work-stealing:** [... **Cilk90, Athapascan/Kaapi, ... CilkArts08,...** X10, TBB]
  - Restricted distributed implementation
  - But with provable performances including cost of scheduling control.



- **Need of a new analysis** to cope with multi-objective and various implementation features [mixed push/pull…]

Achievements for 2005-2009 and research directions

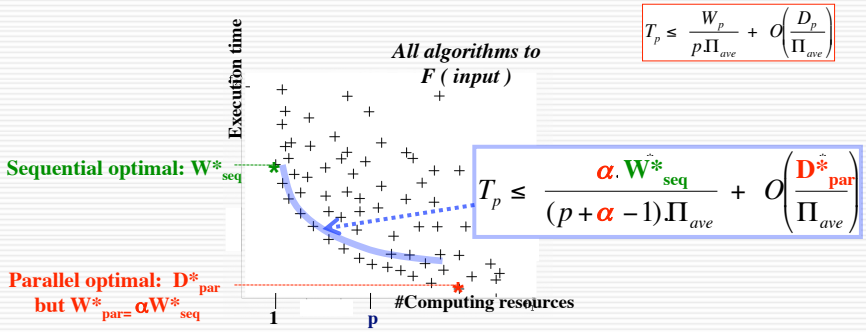
1. Scheduling
2. **Adaptive algorithms**
3. High-Performance Interactive Computation
4. Runtime for HPC on demand

2. Adaptive algorithms

***Objective**: To design and analyze algorithms that may **obliviously** adapt their execution under the control of the scheduling*

- **Heterogeneous resources,** variable speeds: *work-stealing based*

$$T_p \le \frac{W_p}{p\Pi_{ave}} + O\left(\frac{D_p}{\Pi_{ave}}\right)$$



- **Multi-objective problem**: work W_p increases when depth D_p decreases
  - **Adaptive recursive coupling of two algorithms Seq and Par** [PASCO'07,Europar'08]
  - Provable performances, both theory and practice *[**demo** Adaptive/Krash]*

Adaptive streaming

- Use case: HDTV on MPSoCs [ST Microelectronics]
  Minalogic SCEPTRE contract (ST , TIMA, MESCAL, ARENAIRE, COMPSYS, VERIMAG)

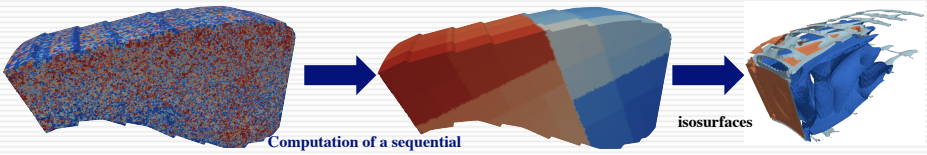
Architecture description
[SPIRIT / IP-XACT Simulator]

Application description
Sequential + potential parallelism

- Cache & processor oblivious stream computations [PDP08]
- Extensions to the programming of large MPSoC [Nano 2012 / ST, HiPeCoMP contract]

Mesh layouts for adaptive algos

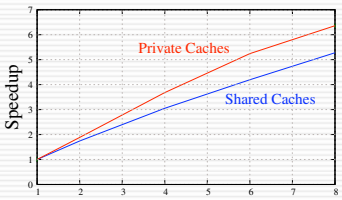
- **Intensive computations on mesh (CEA-DIF contract)**



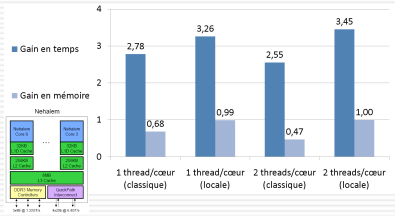
Computation of a sequential cache oblivious layout [IWSV08, IEEE Trans.VCG'10]

7 times less cache misses
5 times faster

Private caches (eg Opteron-875)
Parallel then sequential



Shared cache (eg Nehalem)
performances limited by cache size



Towards generic adaptation techniques cache/prcoessor ?

Research directions

- Heterogeneous and hierarchical architectures
  - Both processor and memory
- Impact of variable loads
- Adaptive resource allocation
  - Energy, time constraints
- Programming models for adaptive algorithms

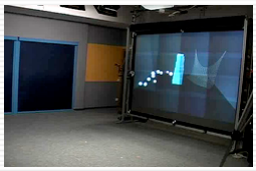
Achievements for 2005-2009 and research directions

1. Scheduling
2. Adaptive algorithms
3. **High-Performance Interactive Computation**
4. Runtime for HPC on demand

3. HPiC High Performance Interactive Computation

- *Motivation: parallelism to move one step towards the next generation of interactive applications*
  - Challenging platform : multi-cameras, multi-cpus, multi-GPUs, multi-display
  - **Grimage** platform associates
    - Parallelism [Moais]
    - computer vision [Perception]
    - realtime physics simulation [Evasion]

Adaptive realtime 3D modeling



[2004]

[2009]
-- Digitalis cluster (Grid 5000)
-- 20 cameras
-- 1 head-mounted display

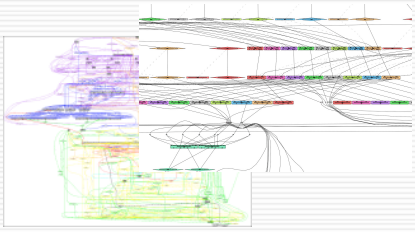
- **Positioning:** other platforms:
  - [Blue-C, ETH Zurich, 2005 …]
    [Tele-Immersion@UCBerkeley 2005]

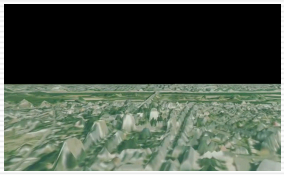
[http://grimage.inrialpes.fr]

3. HPiC : Interactivity

Middleware dedicated to interactive applications [Supercomputing08]

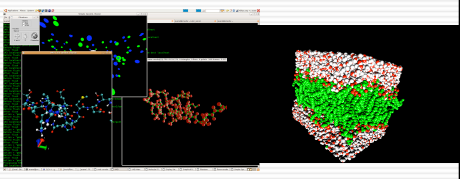
- Hierarchical components
- Data flow





Telepresence [ANR Dalia][VRST08]
3D reconstruction (Grenoble and Bordeaux)
Geographic information service (Orléans)

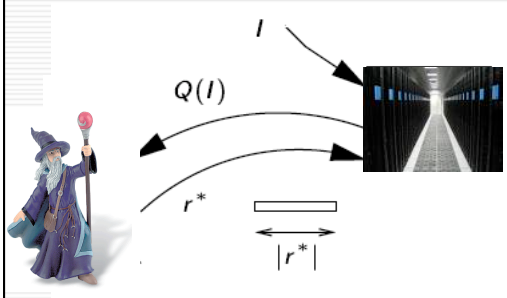
- Coarse grain mapping
- Online frequency calibration
  - multi-objective
  - Tuned sampling -> adaptive



Molecular dynamic
[ANR FVNano]

3. HPiC: Towards interactive computing

- Using interaction with **expert** to improve performances
  - ***"Human in the loop" => "Expert in the loop"***
- Theoretical approach: information (small) from an **Oracle** to improve complexity, performance ratio
  - Interactive complexity classes



Information (guesses) to improve approximation schemes

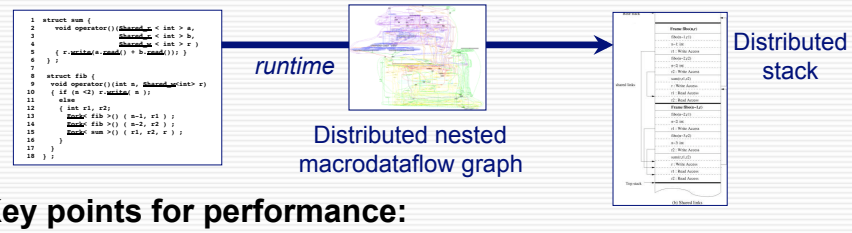
algorithm	approx. ratio	$\lvert r^* \rvert$	complexity
MA^{G_1}	$(k-g)$	$glog(m)$	$O(m^g * kn)$
$MA^{G_1}_{\beta}$	$(k-g)(1-\beta)$	$log(k)+glog(m)$	$O(k * m^g * kn)$
MA^{G_2}	$\frac{k}{g+1}$	$g(log(k)+log(m))$	$O((km)^g * kn)$

Discrete Resource Sharing Scheduling Problem [IPDPS09]

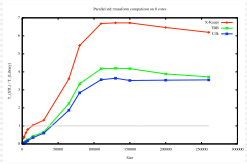
4. Runtime for HPC "on demand"

Objective: **scheduling control** at runtime based on
efficient distributed macro dataflow representation.

➢ **Kaapi middleware**



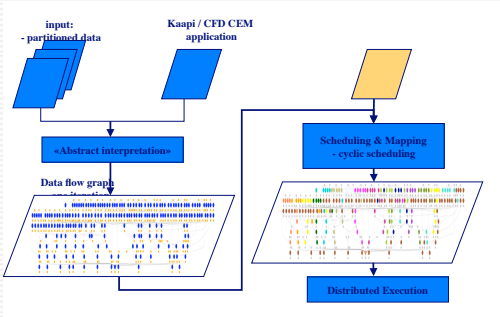
- **Key points for performance:**
  - Local serialization ("*work-first principle*") [J. CLSS'07, ICCS07]
  - **Scheduling** : Pre-partitioning + "work-stealing" (Lock-free / wait-free)
  - Fault-tolerance protocols, from scheduling properties
    - *coordinated protocol* [ICTTA'06, TSI07, MCO08] + *TIC protocol* [EIT05, TDSC09]

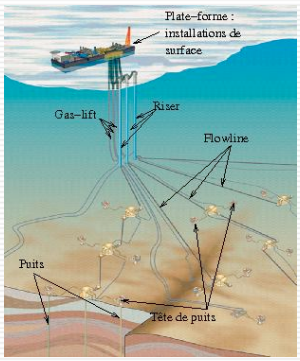


- **Positioning:**
  - Work stealing: Intel TBB, Cilk++, X10
  - Grid / global platforms:
    tolerate failures and falsification: Satin (FT)

Kaapi applications

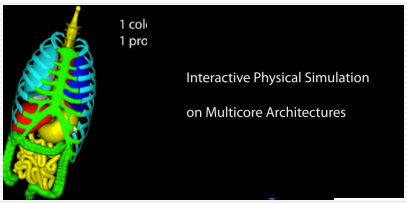
- *Process engineering - CAPE-OPEN standard [IFP]*
- Combinatorial Optimization (QAP, Q3AP: PRiSM/Gilco) [ANR CHOC]
- Numerical Kernel [ANR DISCOGRID]
- Chemical combustion [NUMED]
- Computer algebra […LinBox…]

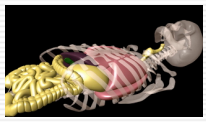




Scheduling: partitioning + work-stealing

Multi CPUs/GPUs Kaapi

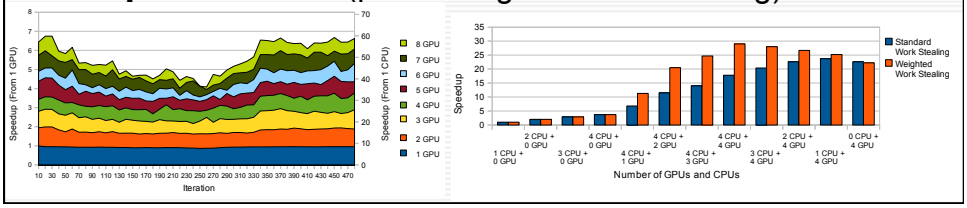




SOFA

Alcove / Evasion
Asclepios

- **Parallelization implicit to SOFA user** [Demo]
- **Mixed scheduling in Kaapi**
  - Initial distribution: graph partitioning CPU&GPU
  - Online work stealing CPUs & GPUs
- **+ Improvements** (partitioning and work stealing)

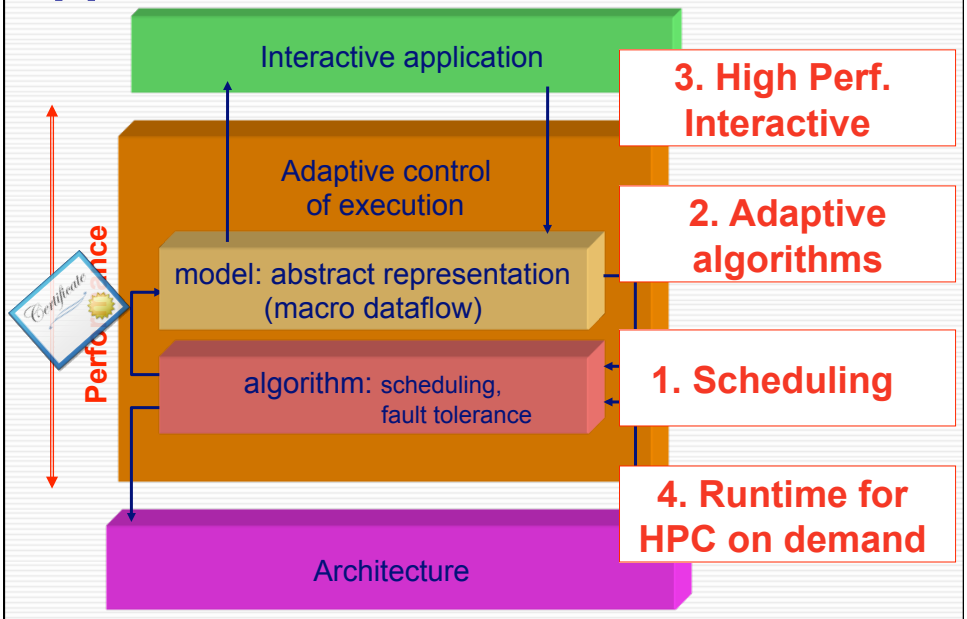


HPC on demand: research directions

- **Secure computing on exascale platforms**
- **Efficient parallel execution with tiny size grain**
  - ➔ Speedup on thousands of cores at fixed problem size
  - Adaptive algorithms [Prototype Xkaapi]
- **Hierarchical memory**
  - work-stealing provable performances?
- **“Oblivious tuning” of parallel applications**
  - ➔ Monitoring and retro-control
  - ➔ Deployment and schrinkening

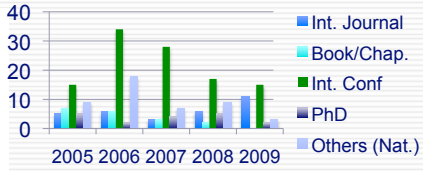
Approach

Research Themes

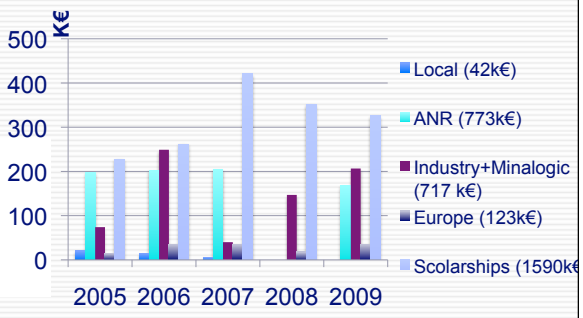


Publications and contracts

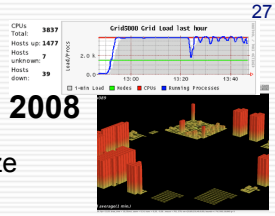
- **Publications**
  - 222 in 4 years , 45 rank #1
  - 19 Int. Journal (SIAM J.Comp, IEEE TC, TPDS, TDSC, EJOR, FCS, JPDC …)
  - 109 Int. Conf (SPAA, IPDPS, CCGrid VR, VIS, Europar, ICCS, Siggraph…)



- **Contracts**
  - Industry partners: STM, IFP, CEA, Bull, C-S, DCN,
  - 2 ARC, 7 ANRs +1
  - 2 pole MINALOGIC +1
  - 2 Europe



Highlights



- **1st prize Plugtest Nov. 2007 and Nov. 2008**

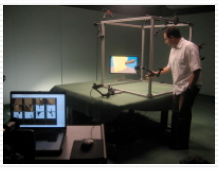
KAAPI

Dec. 2006: special Jury prize
Nov. 2007: 1st prize
- Nqueens(23) in 2107s with 3654 cores

Nov. 2008: Pricing: 1st prize Grid5K + Intrigger (Japan)

- **SIGGRAPH Aug. 2007/2009 Emerging Technologies**

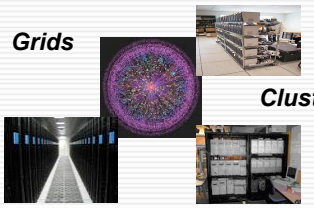
FlowVR
High Performance Virtual Reality



~4000 visitors

- **Valorization**: start-up 4D View Solutions (Sep. 2007)
  - co-founded by former PhD C. Menier [joined MOAIS / PERCEPTION]
  - transfer: parallel 3D modeling

Softwares



Grids

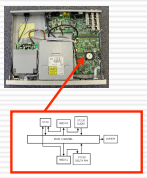
Cluster

SMP

multi-core

GPU

MPSoC



TakTuk

FlowVR
High Performance Virtual Reality

KAAPI

X-Kaapi

+ Dmake, Krash, Triva…

Multi-objective **Adaptive** **Performance**

29

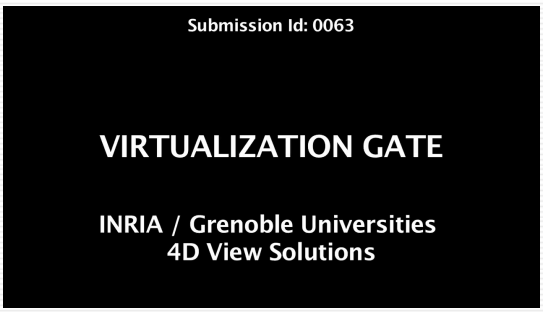
Perspectives summary 2010-2014

- **Manycore / exascale programming: dealing with uncertainties**
  - Heterogeneous computing
    - GPUs, MPSoCs, manycore
  - Large scale autonomous computing
    - faults : Fail-stop and malicious
  - Oblivious algorithms with provable performances
    - Game theory

30

Summary

"To provide parallel programming schemes, interfaces and tools for high performance interactive computing that enable to achieve provable performances on distributed parallel architectures, from multi-processors system-on-chip to lightweight grids and global computing platforms."

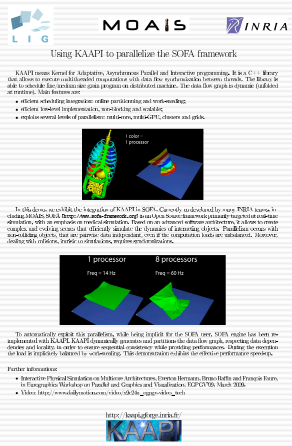


SIGGRAPH'09 [MOAIS - PERCEPTION - EVASION]

Moais Exhibitions

Grimage / FlowVR

SOFA / Kaapi



Adaptive Xkaapi/ Krash