Scalable and Efficient Parallel and Distributed Simulation of Complex, Dynamic and Mobile Systems

Gabriele D'Angelo

joint work with

Luciano Bononi
Michele Bracuto
Lorenzo Donatiello

University of Bologna
Department of Computer Science

Workshop FIRB Perf 2005 – Torino

Presentation outline

- Simulation of large scale and complex models
- The ARTìS software architecture
- Simulation of massive, complex and dynamic systems
  - High Performance Computing (HPC)
- The CR-PADS logical structure and implementation
- Further optimizations for PADS:
  - Hyper-Threading / Multi-Core support
  - Data Marshalling
- Conclusions and future work

Simulation of large scale and complex models

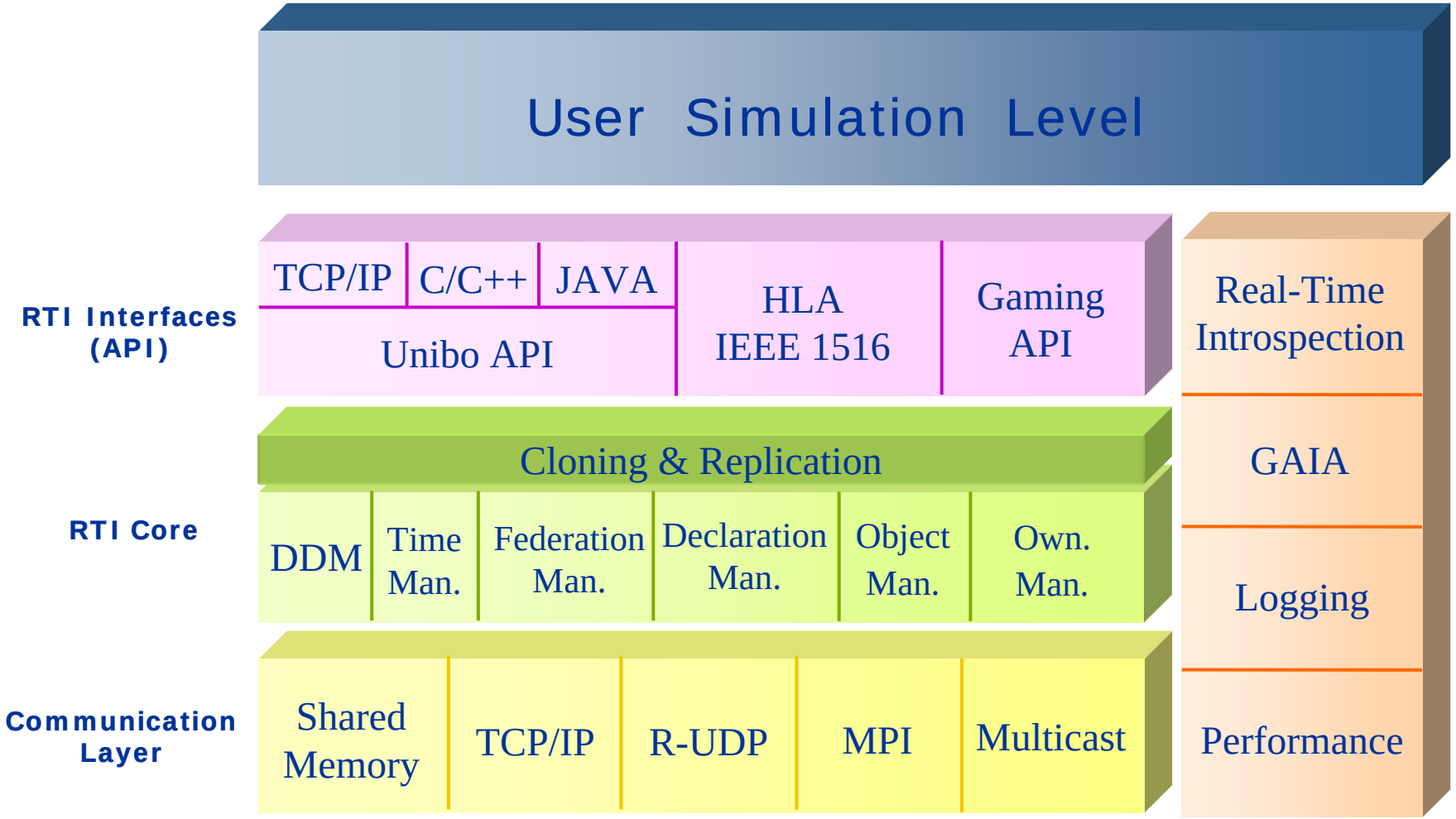
- Simulation models currently of interest may include a potentially huge number of simulated objects
- Large scale and complex simulation models may be unpractical to simulate on a single-processor execution unit: **huge memory requirements**, large amount of **time** required to complete the simulation runs
- The **memory** bottleneck reduction, **scalability** and **speed-up** can be achieved by using *parallel/distributed models and execution architectures*
- Goal: to **increase the simulation speed**, reduce the **Wall Clock Time (WCT)** required to complete the simulation runs

The ARTìS software architecture

Advanced RTI System (ARTìS), parallel and distributed simulation middleware:

- Model components' heterogeneity, distribution and reuse
- Adaptive evaluation of the communication bottlenecks and dynamic adaptation of the inter-process communication layer
- Generic Adaptive Interaction Architecture (GAIA): model components' migration mechanism to support load balancing and data distribution management (DDM) overhead reduction
- Support for High Performance Computing clusters (HPC): scalability evaluation of the middleware
- Concurrent Replication of Parallel and Distributed Simulation (CR-PADS) and cloning

ARTìS: logical architecture



Performance evaluation: Ad-Hoc wireless network model

Simulated model:

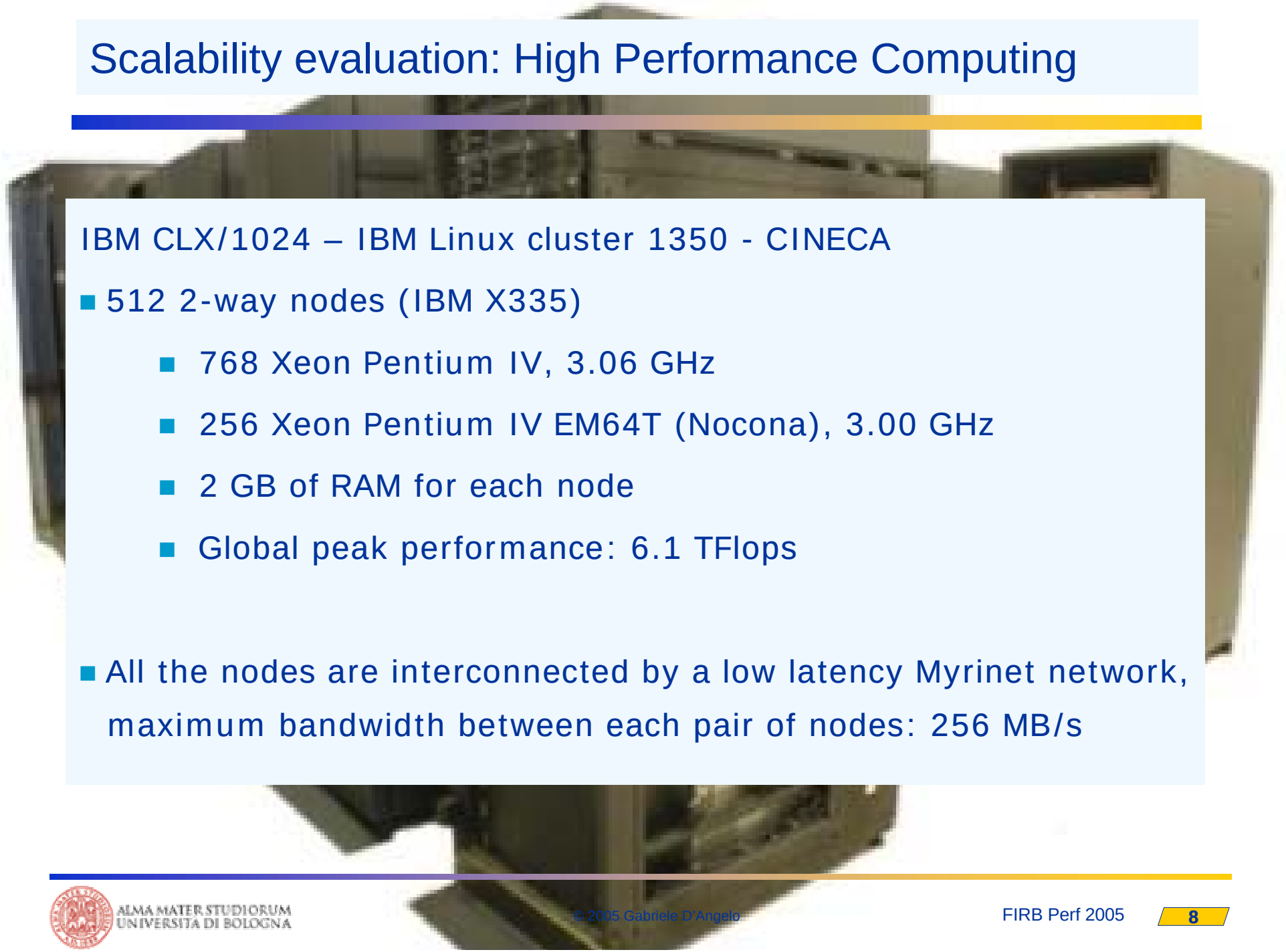
- A set of **Simulated Mobile Hosts** (SMHs)
- Mobility model:
  - Random Mobility Motion model (RMM)
  - uncorrelated SMHs' mobility
- Traffic model:
  - ping messages (CBR) by every SMH to all neighbors within the wireless communication range (250 m)
- Propagation model
  - open space (neighbor-SMHs within detection range)

Ad-Hoc network model characterization

Computation and communication issues:

- The **computation** required for each SMH per time-step is in the order of **O(#SMH^2)**: to determine the neighbor set
- The **communication** required among SMHs is in the order of **O(K*#SMH)** per time-step, with K defined as a constant value based on SMHs density (assumed as constant)

Scalability evaluation: High Performance Computing



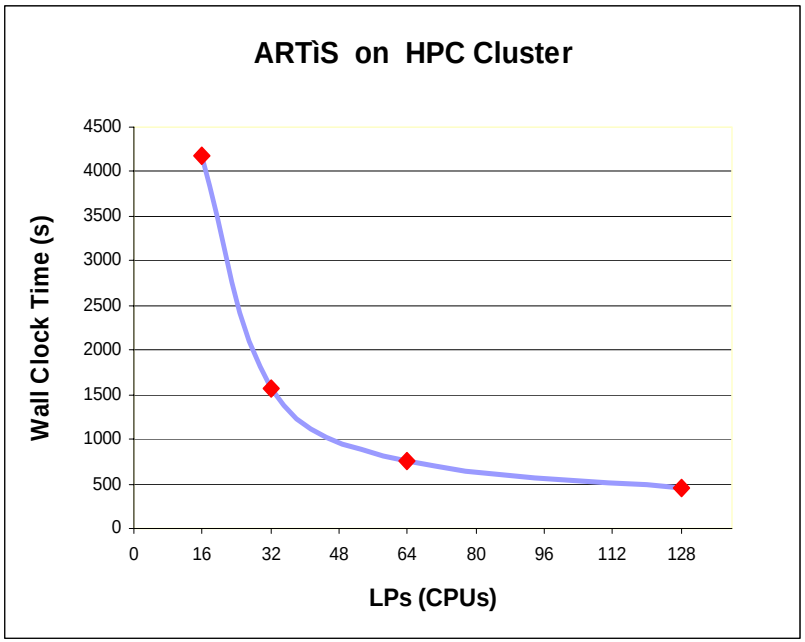
IBM CLX/1024 – IBM Linux cluster 1350 - CINECA

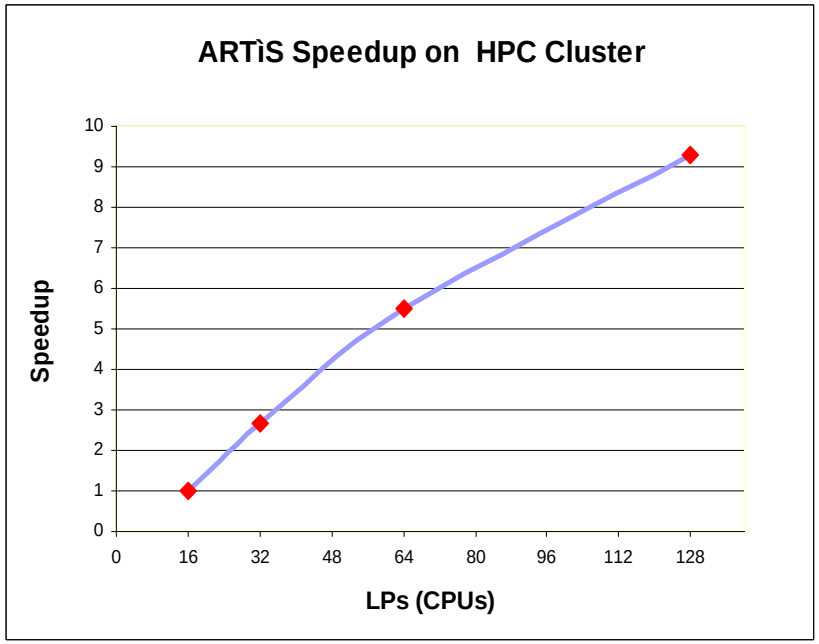
- 512 2-way nodes (IBM X335)
  - 768 Xeon Pentium IV, 3.06 GHz
  - 256 Xeon Pentium IV EM64T (Nocona), 3.00 GHz
  - 2 GB of RAM for each node
  - Global peak performance: 6.1 TFlops

- All the nodes are interconnected by a low latency Myrinet network, maximum bandwidth between each pair of nodes: 256 MB/s

ARTìS on High Performance Computing (HPC)

1 million of Simulated Mobile Hosts (*1 Timestep of simulated time*)





Typical PDES + PADS problems

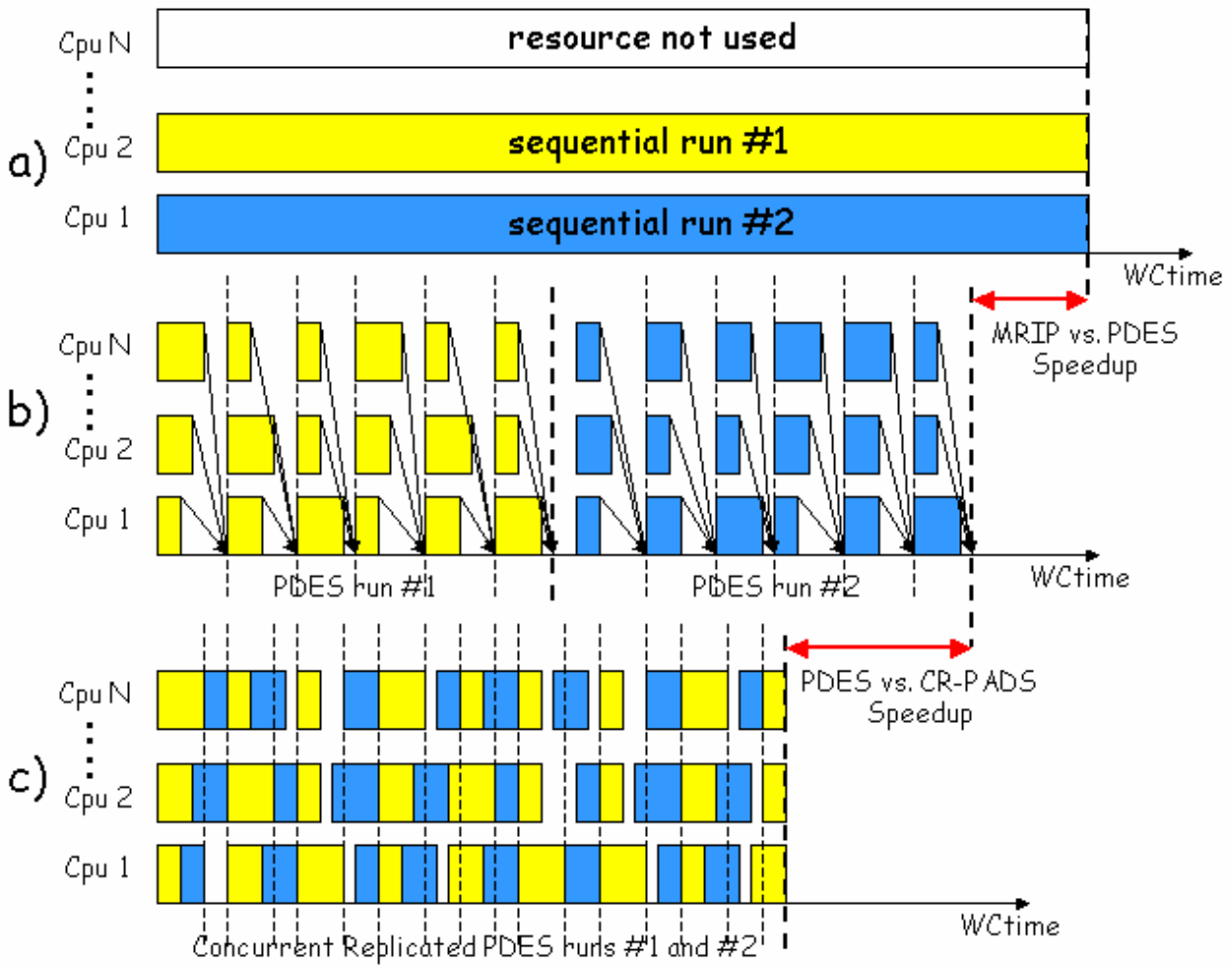
- A typical simulation-based investigation requires to collect many independent observations for a correct and significant statistical analysis of results
- In the PADS environment **frequent synchronizations** are required among the model components
- Each simulation component swings between **computation and communication phases**
- The whole set of processes advance with the **speed of the slowest**
- Some phases (usually before the synchronization barriers) could be communication intensive and may led to network congestions (further increasing the communication overhead)

Our idea is to obtain a more **fluent computation and communication** by concurrently merging the execution of more PADS replicas

Concurrent Replication of PADS (CR-PADS)

- The CR-PADS is a mechanism that **duplicates** the logical processes (LPs) of PADS runs starting from the **initialization phase** of every single run
- Every replica is based on the same model definition, realizes an **independent execution** based on local initial parameters, variable factors of the analysis and different random number generator seeds
- CR-PADS is absolutely **different** from **simulation cloning**
- The risk of this mechanism is to spend too much time in switching processes' execution, in the creation of communication bottlenecks and live-locks, resulting in trashing effects

The CR-PADS approach



Simulation test-beds

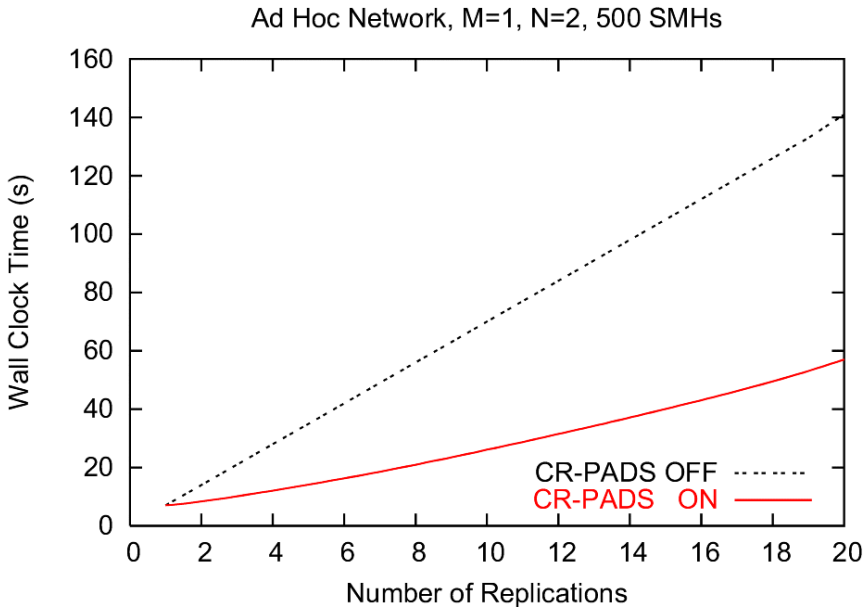
Two different environments:

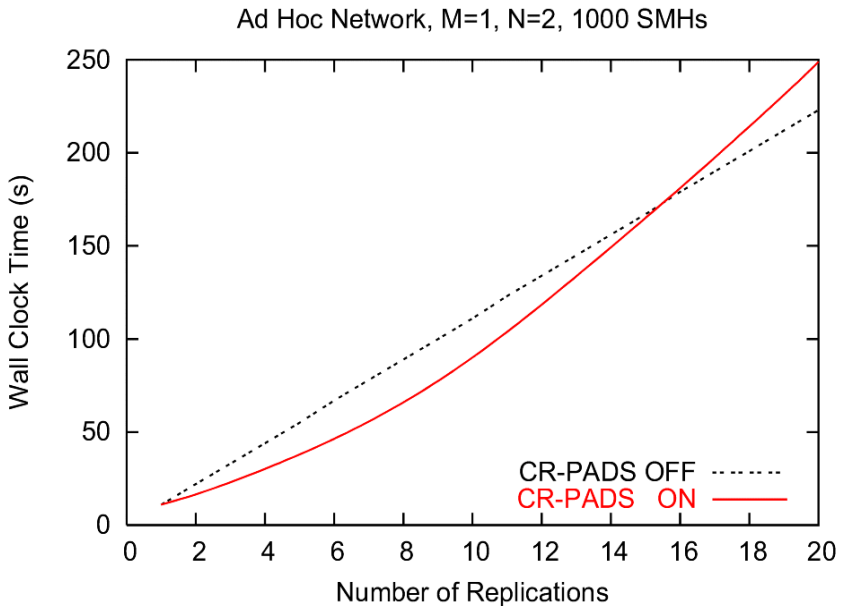
- **Parallel**: a single Physical Execution Unit (PEU)
- **Distributed**: a set of PEU, interconnected by a Fast Ethernet LAN (100 Mb/sec)
- Each PEU is an Intel Dual Xeon Pentium IV 2800 MHz, with 3 GB RAM, Debian GNU/Linux OS with kernel version 2.6.x
- Conservative time-stepped simulation: 300 time-steps

Parallel environment: 500 and 1000 SMHs

M = # PEU

N = # LP

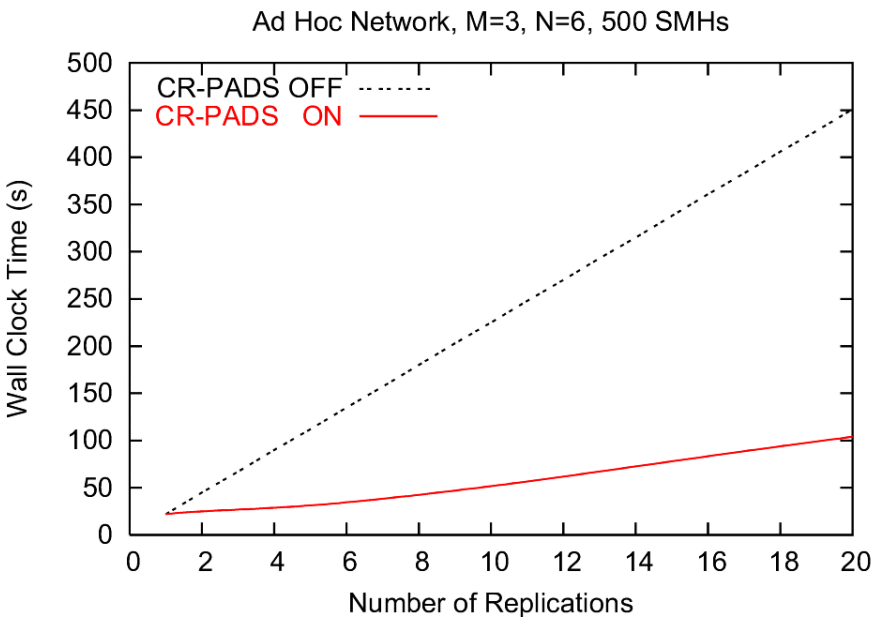


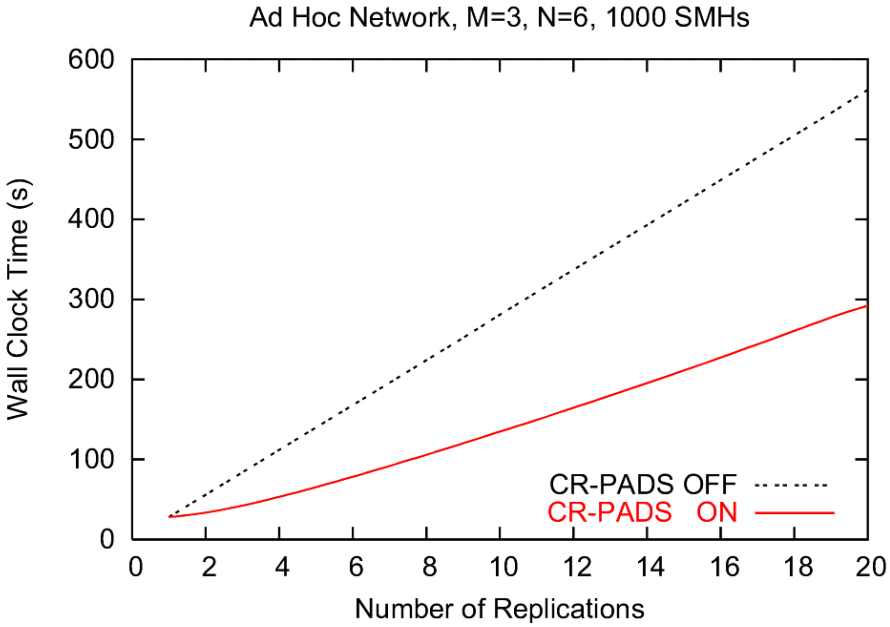


Distributed environment: 500 and 1000 SMHs

M = #PEU

N = #LP

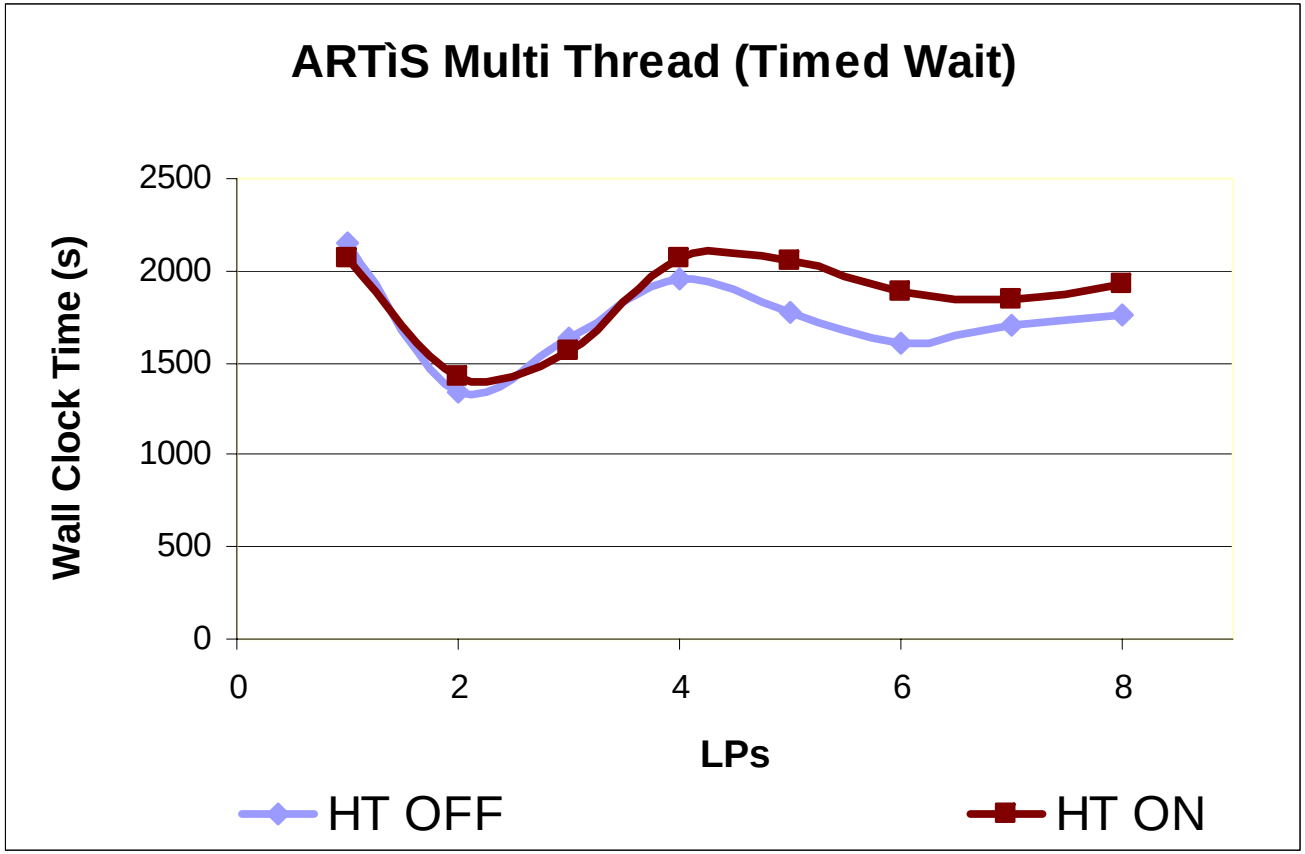




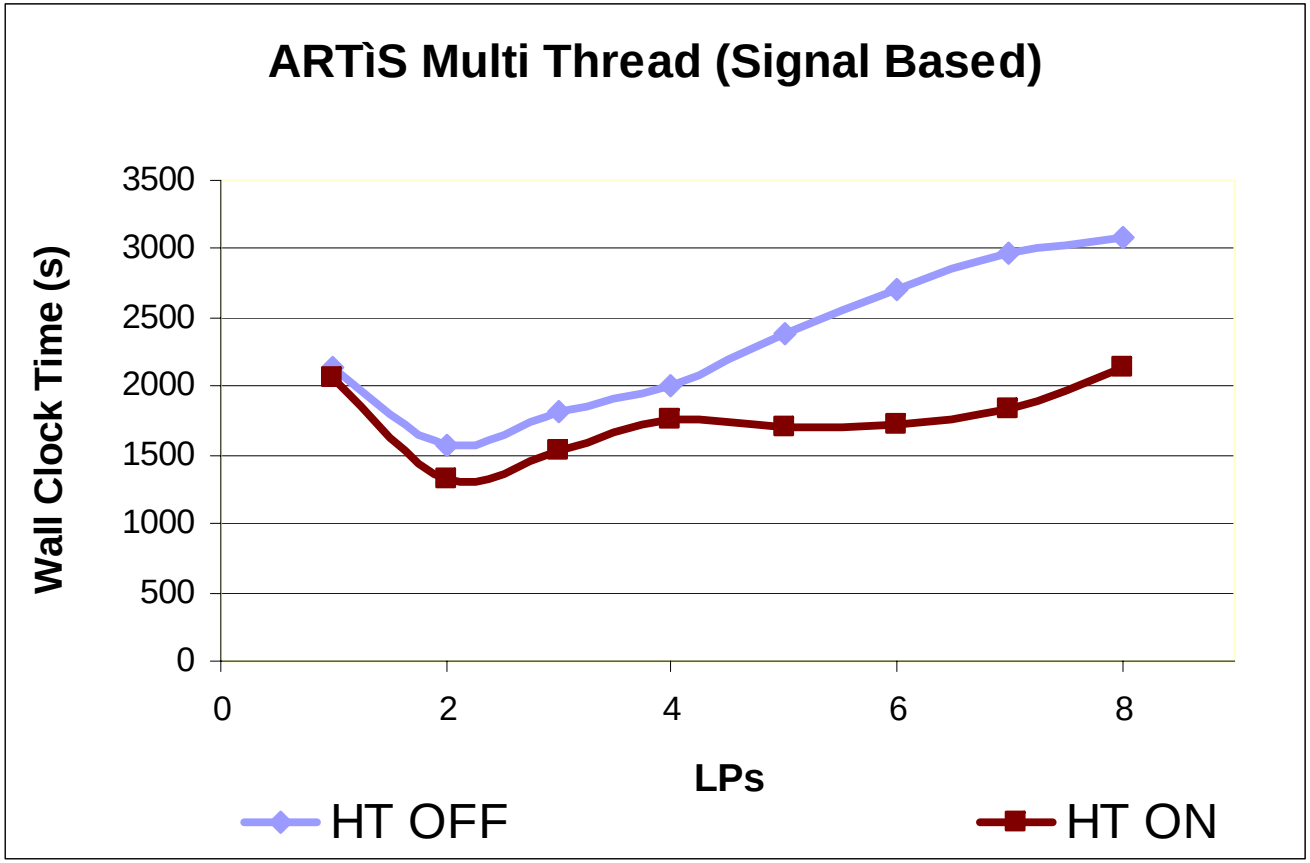
Optimizations: Hyper-Threading (HT) support

- The Hyper-Threading technology (HT) is a **new processor architecture** recently introduced by Intel
- The HT technology duplicates the high level portion of the architecture state on each logical processor, while logical processors share a subset of the physical processor execution resources
- HT technology makes a **single physical processor appearing as two logical processors** at the user level: one physical execution resource (CPU) is **shared** between two logical processors
- The influence of HT technology on PADS architectures and frameworks has not been investigated in detail

ARTìS and Hyper-Threading



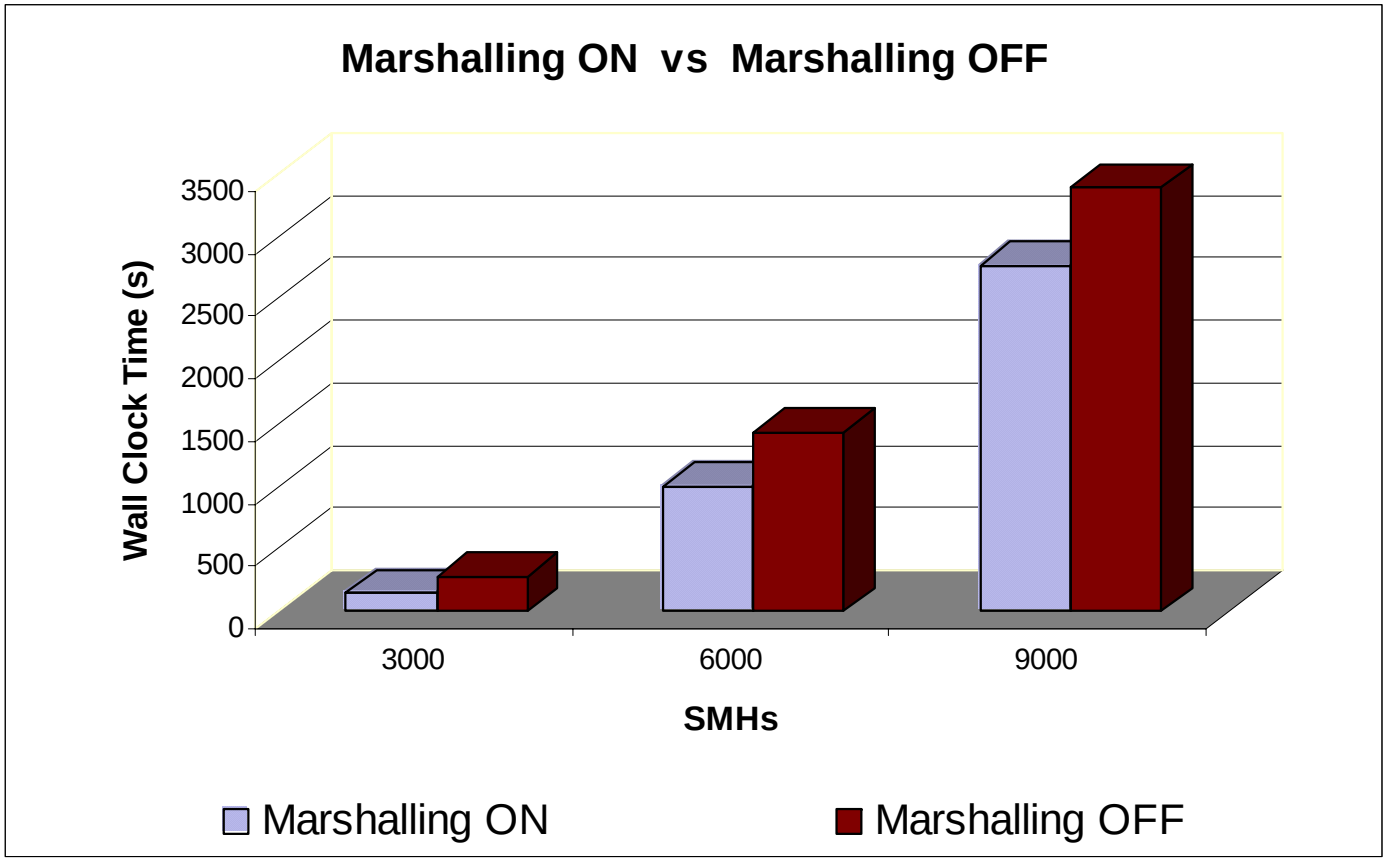
ARTìS and Hyper-Threading



Optimizations: communication marshalling

- The **communication efficiency** is one of the main factors determining the efficiency of a parallel or distributed simulation
- The data marshalling approach consists in the **concatenation** of more than one logical message in the same communication messages
- The data marshalling process is controlled by a timer: once every a maximum time limit the messages buffered on the LP are sent in a data marshalling packet (or frame)
- Inverse trade-off: degradation in the average communication **latency**
- The proposed optimization has been applied both to shared memory and TCP/IP communications

Communication Marshalling: distributed environment



Conclusion and Future work

- ARTÌS is a scalable, optimized parallel and distributed simulation middleware, used to simulate dynamic complex systems
- Careful and enhanced evaluation of the proposed optimizations
- Further improve the ARTÌS middleware
  - Data structures optimization (event-management)
  - SCTP support (IP based communication protocol)
  - IEEE 1516 full compatibility
  - New models (es. Detailed MAC protocols)
  - Migration based middleware for Internet games

References

1. L.Bononi, M.Bracuto, G.D'Angelo, L.Donatiello. **Analysis of High Performance Communication and Computation Solutions for Parallel and Distributed Simulation**. Proc. Of HPCC 2005
2. L.Bononi, M.Bracuto, G.D'Angelo, L.Donatiello. **Concurrent Replication of Parallel and Distributed Simulations**. Proc. of *ACM/IEEE/SCS PADS 2005*
3. L.Gardenghi, S.Pifferi, G.D'Angelo, L.Bononi. **Design and Simulation of a Migration-based Architecture for Massively Populated Internet Games**. Proc. of *IEEE NIME 2004*
4. L.Bononi, M.Bracuto, G.D'Angelo, L.Donatiello. **ARTIS: a Parallel and Distributed Simulation Middleware for Performance Evaluation**. Proc. of *ISCIS 2004*
5. L.Bononi, M.Bracuto, G.D'Angelo, L.Donatiello. **A New Middleware for Parallel and Distributed Simulation of Dynamically Interacting Systems**. Proc. of *IEEE DS-RT 2004*
6. L.Bononi, M.Bracuto, G.D'Angelo, L.Donatiello. **Performance Analysis of a Parallel and Distributed Framework for Large Scale Systems' Simulation**. Proc. of *ACM/IEEE MSWiM 2004*
7. L.Bononi, G.D'Angelo. **A Novel Approach for Distributed Simulation of Wireless Mobile Systems**. Proc. of *IFIP-TC6 PWC 2003*
8. L.Bononi, G.D'Angelo, L.Donatiello. **HLA-based Adaptive Distributed Simulation of Wireless Mobile Systems**. Proc. of *ACM/IEEE/SCS PADS 2003*

Scalable and Efficient Parallel and Distributed Simulation of Complex, Dynamic and Mobile Systems

Gabriele D'Angelo
<gda@cs.unibo.it>

joint work with

Luciano Bononi
Michele Bracuto
Lorenzo Donatiello

University of Bologna
Department of Computer Science

Workshop FIRB Perf 2005 – Torino