

Stéphanie Moreaud

Post Doctoral Researcher,
Innovative Computing Laboratory,
University of Tennessee

E-mail : smoreaud@eecs.utk.edu
Phone : (+1) 865-974-6321

ICL, University of Tennessee
Suite 413 Claxton
1122 Volunteer Blvd
Knoxville TN 37996-3450, USA

French nationality

Education

- 2007–2011** **PhD Thesis in Computer Science - Tasks and data movement for high performance communication on modern architectures.** INRIA, LaBRI, University of Bordeaux I, France. *Advisors : Raymond Namyst and Brice Goglin.*
- 2005–2007** **Master of Computer Science in distributed systems, networks and parallelism.** University of Bordeaux, France, Computer Science Department.
- 2002–2005** **Bachelor of Computer Science.** University of Bordeaux, France, Computer Science Department.

Research area

The increasing number of cores leads to a drastic complexification of hardware topologies, with multiple shared cache levels, memory nodes, I/O busses, etc. On hierarchical architectures, data exchange and communication efficiency between tasks depend on the physical location of tasks and data on the underlying topology. To achieve the performance level required by high-performance computing, tasks and data placement as well as communication strategies have to be carefully adapted according to hardware resources and software affinities. In this context, my work focusses on the evaluation of hardware topologies effects on communication performances and on the design of methods to adapt tasks and data placement or communication strategies according to these constraints.

Research Interests :

- High performance computing
- Multicore hierarchical architectures
- Hierarchies' effects on communication
- Message passing, MPI
- High speed networking in clusters.

Invited talks and research visits

- 2010** **Seminar,** “ *Adaptive MPI Multirail Tuning for Non-Uniform Input/Output Access* ”, at the French Alternative Energies and Atomic Energy Commission, Bruyères-le-Châtel.
- 2010** **PhD students seminar,** “ *Impact of multiprocessor architectures on cluster communication* ”, LaBRI, Bordeaux.
- 2009** **Research visit** at the Argonne National Laboratory, in the context of a collaboration with the Radix team to experiment our ideas in the MPICH2-Nemesis software stack (1 week).
- 2009** **National day of young researchers on multicore and multiprocessors,** “ *High performance communication between MPI processes on multicore architectures* ”, Paris.

Other professional experiences

- July–Aug. 2007** **Internship in the INRIA RUNTIME team**, “ *Study of non uniform network device access and threads scheduling* ”, INRIA Bordeaux Sud-Ouest, France.
- Jan.–June 2007** **Master thesis**, “ *Impact of multiprocessors architectures on communication in clusters : from NUMA effects to automatic placement* ”, LaBRI, University of Bordeaux, France, Computer Science Department.
- July–Aug. 2006** **Internship in the INRIA RUNTIME team**, “ *Spinlocks implementation in the MARCEL library* ”, INRIA Bordeaux Sud-Ouest, France.

Teaching experience

- 2010–2011** **Teaching Assistant** at *University of Bordeaux, Computer Science Department*.
- Introduction to Unix, 1st year at the Technology Institute.
- Network administration, 2nd year at the Technology Institute.
- Programming assistance, at the Technology Institute.
- Operating system, 1st year at the Technology Institute.
- System programming, 2nd year at the Ho Chi Minh Ville French University Center, Vietnam.
- 2009–2010** **Teaching Assistant** at *University of Bordeaux, Computer Science Department*.
- Operating system, 1st year at the Technology Institute.
- Advanced Network, 2nd year at the Technology Institute.
- 2008–2009** **Teaching Assistant** at *University of Bordeaux, Computer Science Department*.
- Basic Network, 1st year at the Technology Institute.
- Programmation of parallel architectures, 1st year of Master (postgraduate level).
- 2007–2008** **Teaching Assistant** at *University of Bordeaux, Computer Science Department*.
- Introduction to computer science, 1st year of Licence (undergraduate level).
- Programming environment, 2nd year of Licence (undergraduate level).

Miscellaneous

- 2011** **Organization of the Euro-Par conference** in Bordeaux, France. *Staff member*.
- 2010-2011** **Dissemination of scientific knowledge**, *presentation of research careers and university training at the Aquitec Forum and Students exhibition in Bordeaux*.
- 2010-2011** **Reviews** *Workshop A4MMC 2010, EuroMPI 2010 and CCGrid 2011*.
- 2009** **Participation at the ACACES summer school**, (*Advanced Computer Architecture and Compilation for Embedded Systems*), in Terrassa, Spain.
- 2008–2009** **Representative of PhD students** at the council of mathematics and computer science doctoral school.

Research activities

In the context of high-performance computing, the INRIA RUNTIME team ¹ is interested in the study and design of generic runtime systems for programming environments and applications in the intensive parallel computing field. These runtime systems have to enable an efficient use of the large, heterogeneous and hierarchical clusters. My work focusses on the problem of tasks and data placement on modern hierarchical multiprocessor architectures. The multiplication of cores and the new memory interconnects technology reintroduced *Non-Uniform Memory Access* architectures. Modern multicore architectures present complex topologies with affinities between processors sharing links or caches. On such structures, the distribution of tasks and data has a major impact on computing and communication performances.

NUMA effects on network communication and automatic placement in NEWMADELEINE

NUMA architectures present variable access times to the different memory banks, well-known to impact performances in the context of tasks scheduling. They are also responsible for different access times to the Input/Output busses depending on the physical location of processors. I evaluated the impact of these *Non-Uniform Input/Output Access* (NUIOA) effects on the network communication performance in clusters, during my Master thesis internship. This study reveals a large impact on communication performance and some asymmetric effects on throughput [9, 3].

To achieve reproductive and optimal performance, the communication tasks must be placed on processors close to the Network Interface Card (NIC). We thus proposed an automatic placement of tasks [3, 7] for the PM2² project communication library, NEWMADELEINE³. The implemented strategy collects topology information from the system, provides portable optimal performance and so relieves the users to do manual placement of tasks on the different architectures. This work was part of the NUMASIS and PARA projects of the french National Research Agency.

Adaptive MPI multirail tuning according to NUIOA effects in OPENMPI

To improve scalability, multicore architectures often benefit from several NICs simultaneously used by the communication library to offer a larger bandwidth. These *multirail* communication suffer from NUIOA effects. In the context of MPI applications, adapting the task placement to NUIOA constraints is complex. It requires to detect communication-intensive tasks, to give them a privileged access to the NICs wich might be difficult. Moreover such a placement can conflict with other binding policies chosen according to processes affinities. We thus looked at an orthogonal problem : to optimize the implementation of communication strategies within a predefined task placement. The multirail configuration study shows that MPI implementations should not blindly split large messages in halves but rather adapt the amount of data sent on each card according to the NUMA distance to the NICs and the communication context (point-to-point or collectives operations). We proposed to adapt the splitting ratio used in OPENMPI⁴ to determine the data amount sent on the different cards. Our strategy looks at the hardware topology thanks to the *hwloc*⁵ library and dynamically adjusts the ratio according to each task binding [4].

Including NUIOA constraints in MPI hierarchical collective operations

Modern MPI layers benefit from several collective operation algorithms wich result from a large research investment for the past decades. They are often based on a combination of multiple strategies depending on the underlying cluster topology, with local leader processes being used as intermediate. As local leaders are in charge of network communication, their intensive-communication schemes made them particularly sensitive to NUIOA effects. To counterbalance the NUIOA penalty occurring when leaders are bound far from the NICs, we also proposed to adapt MPI collective operations by electing these leaders according to the locality of processes with respect to NICs so to give them privileged network access [5].

Adapting the strategy selection for intranode communication in MPICH2-Nemesis according to processors affinity

In collaboration with the Radix team of the Argonne National Laboratory (ANL), the RUNTIME team ported the MPICH2-Nemesis⁶ software stack on the PM2 platform. Results of this collaboration, new

¹<http://runtime.bordeaux.inria.fr/Runtime/>

²Available on INRIAGforge : <http://gforge.inria.fr/projects/pm2/>

³Documentation about PM2 components available on <http://runtime.bordeaux.inria.fr/Runtime/software.html>

⁴<http://www.open-mpi.org/>

⁵<http://www.open-mpi.org/projects/hwloc/>

⁶<http://www.mcs.anl.gov/research/projects/mpich2/>

shared memory data transfer strategies such as the use of the kernel module KNEM⁷ have been developed and integrated to MPICH2-Nemesis. We studied the intranode MPI communication performances obtained by the different methods available (double buffering data transfer, direct copy thanks to KNEM, Intel I/OAT DMA,...) on several experimentation platforms. Their analyze showed various results linked to the algorithms complexity, the communication context (message size, point-to-point or collective communication, etc.) and the hardware topology characteristics (memory interconnect technology, cache hierarchy, etc). To obtain optimal performance, we proposed to adapt the switching threshold between the available approaches according to the communication properties and the underlying topology characteristics, detected thanks to *hwloc* [2, 8, 1, 6].

Publications

International conferences

- [1] François Broquedis, Jérôme Clet-Ortega, Stéphanie Moreaud, Nathalie Furmento, Brice Goglin, Guillaume Mercier, Samuel Thibault, and Raymond Namyst. *hwloc: a Generic Framework for Managing Hardware Affinities in HPC Applications*. In *18th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP2010)*, Pisa, Italia, February 2010. IEEE Computer Society Press.
- [2] Darius Buntinas, Brice Goglin, Dave Goodell, Guillaume Mercier, and Stephanie Moreaud. *Cache-Efficient, Intranode Large-Message MPI Communication with MPICH2-Nemesis*. In *38th International Conference on Parallel Processing (ICPP-2009)*, Vienna, Austria, September 2009. IEEE Computer Society Press.
- [3] Stéphanie Moreaud and Brice Goglin. *Impact of NUMA Effects on High-Speed Networking with Multi-Opteron Machines*. In *19th IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS 2007)*, Cambridge, Massachussets, November 2007.
- [4] Stéphanie Moreaud, Brice Goglin, and Raymond Namyst. *Adaptive MPI Multirail Tuning for Non-Uniform Input/Output Access*. In *17th EuroMPI*, Stuttgart, Germany, September 2010.

International workshops

- [5] Brice Goglin and Stéphanie Moreaud. *Dodging Non-Uniform I/O Access in Hierarchical Collective Operations for Multicore Clusters*. In *CASS 2011: The 1st Workshop on Communication Architecture for Scalable Systems, held in conjunction with IPDPS 2011*, Anchorage, AK, May 2011. IEEE Computer Society Press. To appear.
- [6] Stéphanie Moreaud, Brice Goglin, David Goodell, and Raymond Namyst. *Optimizing MPI Communication within large Multicore nodes with Kernel assistance*. In *CAC 2010: The 10th Workshop on Communication Architecture for Clusters, held in conjunction with IPDPS 2010*, Atlanta, GA, April 2010. IEEE Computer Society Press.

National conferences

- [7] Stéphanie Moreaud. *Impacts des effets NUMA sur les communications haute performance dans les grappes de calcul*. In *18ème Rencontres Francophones du Parallélisme (RenPar08)*, Fribourg, Switzerland, February 2008.
- [8] Stéphanie Moreaud. *Adaptation des communications MPI intra-nœud aux architectures multicœurs modernes*. In *19ème Rencontres Francophones du Parallélisme (RenPar09)*, Toulouse, France, September 2009.

Thesis

- [9] Stéphanie Moreaud. *Impact des architectures multiprocesseurs sur les communications dans les grappes de calcul : de l'exploration des effets NUMA au placement automatique*. Master thesis, University of Bordeaux, June 2007.
- [10] Stéphanie Moreaud. *Mouvement de données et placement des tâches pour les communications haute performance sur machines hiérarchiques*. Phd thesis, University of Bordeaux, 2011.

⁷<http://runtime.bordeaux.inria.fr/knem/>