Julien Herrmann

Postdoctoral Researcher

Centre de Recherche Inria Bordeaux
200 Avenue de la Vieille Tour
33405 Talence, FRANCE

Mail: *julien.herrmann@inria.fr*
Web: http://people.bordeaux.inria.fr/julien.herrmann/

Professional Profile

I am a postdoctoral researcher in the Tadaam Team at Inria Bordeaux - Sud Ouest. My research interests are in algorithm design, discrete optimization, scheduling techniques for large-scale distributed platforms, data management in High Performance Computing systems,... My contributions include strong theoretical analysis, as well as efficient implementation for real-life implementations. I worked in different research field, including numerical algorithms for High Performance Computing, data analytics for biomedical informatics, backpropagation in deep neural networks,...

I have 28 co-authors from more than 10 different institutions, and I have more than 200 hours of experience in teaching responsibilities.

Positions

Oct. 2018 - Now	**PostDoc - Research assistant**, Inria Research Center of Bordeaux
Feb. 2016 - Sept. 2018	**PostDoc - Research assistant**, Georgia Institute of Technology
2012 - 2015	**PhD Student**, ENS Lyon - LIP, ROMA team
2009 - 2012	**Élève normalien**, ENS Lyon

Education

2012 - 2015 **PhD in Computer Science**, ENS Lyon

Title: *Memory-Aware Algorithms and Scheduling Techniques for Matrix Computations*

Defended on November 25th 2015 at ENS Lyon

Jury:

Examiners

Oliver Sinnen	Senior lecturer at Auckland University, New Zealand
Denis Trystram	Full Professor at Institut National Polytechnique de Grenoble, France

Thesis referees

Luc Giraud	Research Director INRIA Bordeaux
Pierre Manneback	Full Professor at Université de Mons, Belgium

PhD supervisors

Loris Marchal	CNRS researcher at ENS Lyon
Yves Robert	Full Professor at ENS Lyon

2010 - 2012 **Master's degree in Theoretical Computer Science**, ENS Lyon, Lyon, France
+5 month Internship with Marc Baboulin @Université Paris-Sud, France and University of Tennessee, USA
"Butterfly transformations to solve linear systems"
+5 month Internship with Yves Robert @Ecole Normale de Lyon, France and University of Tennessee, USA
"Mixing LU and QR factorisation to solve linear systems"

2009 - 2010 **Bachelor in Theoretical Computer Science**, ENS Lyon, Lyon, France
+2 month Internship with Charles-Edmont Bichot @Ecole Centrale de Lyon, France
"Genetic algorithms for graph partitioning problems"

Major publications

- Mathieu Faverge, Julien Herrmann, Julien Langou, Bradley Lowery, Yves Robert, and Jack Dongarra
  Mixing LU and QR factorization algorithms to design high-performance dense linear algebra solvers
  Journal of Parallel and Distributed Computing (JPDC)

- Julien Herrmann, Jonathan Kho, Bora Uçar, Kamer Kaya, and Ümit V. Çatalyürek
  Acyclic partitioning of large directed acyclic graphs
  Proceedings of Cluster, Cloud and Grid Computing (CCGRID)

- Guillaume Aupy, Julien Herrmann, Paul Hovland, and Yves Robert
  Optimal multistage algorithm for adjoint computation
  SIAM J. Scientific Computing (SISC)

- Marc Baboulin, Jack Dongarra, Julien Herrmann, and Stanimire Tomov
  Accelerating linear system solutions using randomization techniques
  ACM Transactions on Mathematical Software (TOMS)

Teaching responsibilities

Note that, in France, L1, L2, and L3 classes are undergrad levels, M1, and M2 are Master levels.

During the academic years 2010-2011 and 2011-2012, I was in charge of mathematics oral examinations in my former school, Les Lazaristes, Lyon. I was preparing the exercises, questioning and grading the students, one by one, two hours per week. They had a L1 or L2 level in mathematics.

During my PhD and postdoc, I was a teaching assistant in seven different computer science classes. I was in charge of the exercise sessions and/or the programming sessions. I was writing the exercise sheets, supervising the exercise sessions, and following the students' progress. I was also preparing, correcting, and grading the homework and the mid-term exams. I was in charge of the continuous monitoring of the students and I was responsible of half their final grade for the class. I was the only teaching assistant for most of the lessons. Some of them were taught in French, some in English depending on either there were foreign students in the class or not. Willing to diversify my teaching experience, I taught classes at my former school, the E.N.S Lyon, as well as at the University Lyon 1, where I was in charge of larger average class size. Lately, I have been a teaching assistant at the ENSEIRB-MATMECA engineering school, which allowed me to discover a new teaching environment.

During my PhD, I was a substitute teacher for the "Introduction to Computer Science" module in the Classe Passerelle at the E.N.S. Lyon. The goal of the Classe Passerelle is to welcome high school students from disadvantaged neighborhoods with a great academic level but not sufficient to get into college. They are offered an intensive transition year of study to get into a university. I was in charge of preparing from scratch 16 hours of class per year for two years on basic computer science concepts (high school level / first year of college).

The following table is a summary of my teaching responsibilities. In this table, OE means Oral Exam, L means Lecture, EC means Exercise Class, and PC means Programming Class.

Class name	Year	Number of students	Class Level	Number of hours			
				OE	L	EC	PC
ENSEIRB-MATMECA Bordeaux							
Probability and Statistics	2018-2019	30	L3	-	-	18	9
E.N.S. Lyon							
Parallel Algorithms	2013-2014 2014-2015	20	M1	-	-	24 24	-
Probability for CS	2013-2014 2014-2015	20	M1	-	-	24 24	-
Introduction to computer science	2013-2014 2014-2015	20	Classe Passerelle (L1)	-	16 16	-	-
Advance Algorithmic	2015-2016	25	L3	-	-	32	-
Université Lyon 1							
Unix Usage	2012-2013	42	L2	-	-	-	15
Introduction to programming	2015-2016	30	L1	-	-	-	25
Les Lazaristes Lyon							
Mathematics	2010-2011 2011-2012	1	L1 / L2	40 40	-	-	-
Subtotal				80	32	146	49
Total				297			

Collective responsabilities

- 2014-2016: Representative of the interns and PhD students at the laboratory council (ENS Lyon)
- 2013-2015: Volunteer in the Plaisir-Maths nonprofit organization, whose goal is to organize activities and serious games to teach Mathematics to young students (from Elementary School to High School). For two years, I helped creating games and hosting afternoon activities for classes going up to 40 students.
- 2014: Member of the MoMISSS local organizing commiteee
- 2013: Member of the ICPP local organizing commiteee
- 2012-present: I've been a frequent referee for the following journals: JPDC, Parallel Computing,...; and for the following conferences: Euro-Par, IPDPS, ICPP, HPCS,...

Supervision of research activities

- During my postdoc, I helped my advisor, Ümit V. Catalyürek, to supervise his PhD students. Jonathan Kho and Yusuf Ozkaya were working on centrality algorithms for large-scale graphs. They worked with me on a convex partitioning library for acyclic graphs and on a modular web application for the visualization and analysis of biomedical data. Abdurrahman Yasar worked on graph alignment problems.
- In summer 2015, I co-supervised the internship of Nicolas VIDAL (L3 at ENS Lyon), with Yves Robert. The research was about the estimation of the expectancy of the longest path in an acyclic task graph where tasks are subject to random failures. We proved that computing the exact expectancy of the critical path is a #P-complete problem, and we designed approximation algorithms, including an algorithm to compute the first order Taylor approximation of the expectancy.

List of publications

All my publications and the code developed are available on my website. The authors order is always the alphabetic order, except for [J5], [J6], [C1] , [C2], [P1], and [**?**].

Papers in international journals

[J1] Julien Herrmann, M. Yusuf Özkaya, Bora Uçar, Kamer Kaya, and Umit V. Catalyurek. Multilevel algorithms for acyclic partitioning of directed acyclic graphs. *SIAM Journal on Scientific Computing*, 41(4):A2117–A2145, 2019.

[J2] Henri Casanova, Julien Herrmann, and Yves Robert. Computing the expected makespan of task graphs in the presence of silent errors. *Parallel Computing*, 75:41–60, 2018.

[J3] Guillaume Aupy and Julien Herrmann. Periodicity in optimal hierarchical checkpointing schemes for adjoint computations. *Optimization Methods and Software*, 32(3):594–624, 2017.

[J4] Guillaume Aupy, Julien Herrmann, Paul Hovland, and Yves Robert. Optimal multistage algorithm for adjoint computation. *SIAM J. Scientific Computing*, 38(3), 2016.

[J5] Mathieu Faverge, Julien Herrmann, Julien Langou, Bradley Lowery, Yves Robert, and Jack Dongarra. Mixing LU and QR factorization algorithms to design high-performance dense linear algebra solvers. *Journal of Parallel and Distributed Computing (JPDC)*, 85C:32–46, 2015.

[J6] Julien Herrmann, George Bosilca, Thomas Hérault, Loris Marchal, Yves Robert, and Jack Dongarra. Assessing the cost of redistribution followed by a computational kernel: complexity and performance results. *Parallel Computing*, 2015.

[J7] Julien Herrmann, Loris Marchal, and Yves Robert. Memory-aware tree traversals with pre-assigned tasks. *Journal of Parallel and Distributed Computing (JPDC)*, 75:53–66, 2015.

[J8] Marc Baboulin, Jack Dongarra, Julien Herrmann, and Stanimire Tomov. Accelerating linear system solutions using randomization techniques. *ACM Transactions on Mathematical Software (TOMS)*, 39(2):8, 2013.

Papers in international conferences

[C1] M. Yusuf Özkaya, Anne Benoit, Bora Uçar, Julien Herrmann, and Umit V. Catalyurek. A scalable clustering-based task scheduler for homogeneous processors using DAG partitioning. In *IPDPS 2019 - 33rd IEEE International Parallel & Distributed Processing Symposium*, pages 1–11, Rio de Janeiro, Brazil, May 2019. IEEE.

[C2] Julien Herrmann, Jonathan Kho, Bora Uçar, Kamer Kaya, and Ümit V. Çatalyürek. Acyclic partitioning of large directed acyclic graphs. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017, Madrid, Spain, May 14-17, 2017*, pages 371–380, 2017.

[C3] Julien Herrmann, Zachary L. Witter, Nakul Patel, Jonathan Kho, Daniel A. Janies, and Ümit V. Çatalyürek. Visual analytics on the spread of pathogens. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, Seattle, WA, USA, October 2-5, 2016*, pages 519–520, 2016.

[C4] Henri Casanova, Julien Herrmann, and Yves Robert. Computing the expected makespan of task graphs in the presence of silent errors. In *45th International Conference on Parallel Processing Workshops, ICPP Workshops 2016, Philadelphia, PA, USA, August 16-19, 2016*, pages 141–150, 2016.

[C5] Emmanuel Agullo, Olivier Beaumont, Lionel Eyraud-Dubois, Julien Herrmann, Suraj Kumar, Loris Marchal, and Samuel Thibault. Bridging the gap between performance and bounds of cholesky factorization on heterogeneous platforms. In *Proc. of Heterogeneity in Computing Workshop (HCW)*, 2015.

[C6] Thomas Herault, Julien Herrmann, Loris Marchal, and Yves Robert. Determining the optimal redistribution for a given data partition. In *Proc. of IEEE International Symposium on Parallel and Distributed Computing (ISPDC)*, pages 95–102. IEEE, 2014.

[C7] Mathieu Faverge, Julien Herrmann, Julien Langou, Bradley Lowery, Yves Robert, and Jack Dongarra. Designing LU-QR hybrid solvers for performance and stability. In *Proc. of IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2014.

[C8] Julien Herrmann, Loris Marchal, and Yves Robert. Memory-aware list scheduling for hybrid platforms. In *Workshop on Advances in Parallel and Distributed Computational Models (APDCM)*, 2014.

[C9] Julien Herrmann, Loris Marchal, and Yves Robert. Model and complexity results for tree traversals on hybrid platforms. In *Proc. of Euro-Par Parallel Processing*, pages 647–658. Springer, 2013.

Posters

[P1] Julien Herrmann, Zachary L. Witter, Nakul Patel, Jonathan Kho, Daniel A. Janies, and Ümit V. Çatalyürek. PDG framework: Visual analytics on the spread of pathogens. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB), October 2016*, 2016.

Under review

[R1] Olivier Beaumont, Lionel Eyraud-Dubois, Julien Herrmann, Alexis Joly, and Alena Shilova. Optimal checkpointing strategy for general sequential models in pytorch. *Conference on Neural Information Processing Systems (NIPS).*

[R2] Olivier Beaumont, Julien Herrmann, Guillaume Pallez (Aupy), and Alena Shilova. Optimal memory-aware backpropagation of deep join networks. *Philosophical Transactions of the Royal Society.*

[R3] Guillaume Aupy and Julien Herrmann. H-revolve: A framework for adjoint computation on synchrone hierarchical platforms. *ACM Transactions on Mathematical Software (TOMS).*

Key contributions and Technology Development

Here is a summary of some of my major contributions:

- **Backpropagation in deep neural networks:** I focused on memory management problems that rise in the field of Automatic Differentiation or when computing the gradient descent in the training phase of deep neural networks. It consists in dynamically selecting which forward activations are saved, and then to automatically recompute missing activations from previous stored ones. I provided the first algorithm to compute the optimal checkpointing strategy for two storage levels [J4, J3] (it has been an open problem for 7 years in the field of Automatic Differentiation), and then, for any storage architecture [R3]. These theoretical results have been extended to more complex neural networks [R2] and I have implemented some of these checkpointing strategies in Pytorch to efficiently train deep neural networks when the GPU memory is not sufficient to store every forward activations [R1].

  – I am the main designer and developer of a Python library implementing optimal checkpointing strategies and heuristics for the backpropagation problem for any storage architecture.:
  https://gitlab.inria.fr/adjoint-computation/disk-revolve-public

  – This library was used to implement optimal checkpointing strategies in Pytorch allowing to perform the training phase of deep neural network despite restrictive memory constraints. Our fully automatic tool obtains almost 20% higher throughput on average compared to every existing alternatives in Pytorch [R1], and enables to use larger models, larger batch sizes or larger images while fitting into the memory limit of the training device.

  On these topics, I collaborated with automatic differentiation specialists (Paul Hovland, Argonne National Laboratory) and machine learning specialists (Alexis Joly, INRIA Sophia-Antipolis).

- **Convex partitioning of directed acyclic graphs:** I conceived the first multi-level convex partitionner for large directed acyclic graphs [C2, **?**]. The main difficulty of this work was to design and efficiently implement novel coarsening heuristics and refinement heuristics suited for large directed graphs. I conceived fast and efficient heuristics with strong theoretical certification for every steps of the multi-level convex partitionner. Convex partitions can be used to compute task-graph traversals that minimize the memory consumption or to design specific cluster-based list scheduling heuristics.

  – I am the main designer and developer of a convex partitioner for directed acyclic graphs (more than 14,000 lines of codes, in C++):
  http://tda.gatech.edu/software/dagP/index.html
  This library handles and partitions directed acyclic graphs with millions of nodes in matters of seconds. It has been later used to design memory-aware DAG schedulers [C1].

  This work has been conducted during my postdoc at Georgia Institute of Technology with Ümit V. Çatalyürek.

- **Visual analytics on the spread of pathogens:** During my postdoc at Georgia Institute of Technology, I have designed, developed, and implemented a modular, efficient framework of software tools to understand the spread of infectious disease through pathogen genetics and associated metadata for DTRA's Biosurveillance Ecosystem (BSVE). These metadata include host organisms, times and places of isolation, and clinical characteristics associated with a disease caused by pathogens. Our software aggregates this data and allows the user to analyze it with graphical structures. Our application, the Pathogen Dynamic Graph (PDG) organizes the interrelationships among the data elements in a highly interactive network.

  – I was in charge of designing analysis algorithms that work efficiently on the graph database (Neo4j) in which the metadata were stored [C3, P1].

  This work was done in colaboration with Zachary Witter and Daniel Janies from UNC Charlotte.

- **Performance and stability of linear solvers:** I focus on improving runtime systems for High Performance Computing applications on shared memory and distributed memory systems. I designed novel numerical algorithms to solve dense linear systems and efficient memory-aware scheduling techniques. I always attach great importance to implement these theoretical improvements in state-of-the-art runtime systems and to test them on real-life applications executed on supercomputers.

  - *Implementation of data redistribution techniques followed by a computational kernel:* On distributed memory clusters, performing data redistribution before a computational kernel can lead to better performance depending on the processes affinity of the application. I designed algorithms computing the best redistribution knowing the given initial distribution of the data and the parameters of the computational kernel [J6, C6]. The algorithms were coded in Python and I worked in close collaboration with the PaRSEC development team (a runtime system for distributed memory clusters developed at the University of Tennessee, Knoxville) to integrate the code into the PaRSEC runtime system (C++). They offer a significant improvement, especially for linear algebra applications, when the initial data distribution is not suited. The code is available in the public version of PaRSEC:
    http://icl.cs.utk.edu/parsec/software/index.html
  - *Implementation of dynamic scheduling techniques based on static analysis:* State-of-the-art runtime systems rely on dynamic scheduling and resource allocation mechanisms. I analyzed the performance of dynamic schedulers based on both actual executions and simulations, and I investigated how adding static rules based on an offline analysis of the problem to their decision process can indeed improve their performance [C5]. This work was implemented in StarPU, a runtime system for shared memory heterogeneous nodes developed at INRIA Bordeaux, and displayed performance improvement when scheduling dense linear applications. The code is available in the public version of StarPU:
    http://starpu.gforge.inria.fr/files/
  - *Implementation of an hybrid linear solver:* When solving a linear system, one can use an LU factorization (very efficiently parallelized) or a QR factorization (twice as costly but numerically stable). I designed a novel hybrid linear solver for distributed memory clusters that dynamically alternates LU with local pivoting and QR elimination steps [J5, C7]. I introduced several robustness criteria that forecast beforehand whenever the LU steps do compromise stability, and decide to perform a QR step if so. I implemented this hybrid solver in PaRSEC. It provides a continuous range of trade-offs between performance and stability that we can adjust with a parameter. The code is available in the public version of PaRSEC:
    http://icl.cs.utk.edu/parsec/software/index.html