SCATE

IDENTITÉ / PERSONAL DETAILS

Genre / Gender : Homme / Man

Nom / Last Name : KUMAR

Prénom / First name : Suraj

Fonction / Function : Chargé.e de recherche / Research Fellow

Date de soutenance de thèse / Date of PhD defense : 04-2017

POSTE(S) ACTUEL(S) / CURRENT POSITION(S)

Organisme public français / French public entity

Code RNSR / R.N.S.R Code	Laboratoire / Laboratory	Code postal / Zip code	Localité / Locality
202224236C		69603	Villeurbanne

Organisme privé français / French private entity

Siret / Siret	Etablissement / Establishment	Direction service / Direction / department	Code postal / Zip code	Localité / Locality

Organisme étranger / Foreign entity

Etablissement / Establishment	Laboratoire / Laboratory	Localité / Locality	Pays / Country

AUTRES ACTIVITÉS / OTHER ACTIVITIES :

-- Program committee member for the International Conference for High Performance Computing, Networking, Storage, and Analysis 2024 (SC 2025)
-- Program committee member for the International Conference for High Performance Computing, Networking, Storage, and Analysis 2024 (SC 2024)
-- External reviewer for the 35th Symposium on Parallelism in Algorithms and Architectures (SPAA 2023)
-- Program committee member for the 51st International Conference on Parallel Processing (ICPP 2022)
-- External reviewer for the 48th International Conference on Parallel Processing (ICPP 2019).

-- Reviewer for the following international journals: IJPP since March 2018, JPDC since March 2020, CALC since March 2020, SIMAX since May 2020, TOMS since May 2020, SISC since May 2021, TPDS since May 2022.

RESPONSABILITÉS PÉDAGOGIQUES ET ADMINISTRATIVES / PEDAGOGICAL, TEACHING AND ADMINISTRATIVE RESPONSIBILITIES :

Fall 2025: "Resource-aware computations on CPUs and GPUs" for Master students at ENS Lyon
Fall 2024 and Fall 2023: "Data-aware algorithms for matrix and tensor computations" for Master students at ENS Lyon

POSTE(S) ANTÉRIEUR(S) / PREVIOUS POSITION(S) :

De / From	A / To	Nom de l'organisme / Name of the managing authority	Ville / City	Fonction / Function
2019	2022	Inria Paris	Paris, France	Autre / Other : Doctorant postdoctorant | PhD student, Post-doctoral fellow
2018	2019	Pacific Northwest National Laboratory	Richland, USA	Autre / Other : Doctorant postdoctorant | PhD student, Post-doctoral fellow
2017	2018	Ericsson Research	Bangalore, India	Ingénieur.e - chercheur.e (EPST) / Research Engineer (public institution)
2013	2017	Inria Bordeaux	Bordeaux, France	Autre / Other : Doctorant postdoctorant | PhD student, Post-doctoral fellow
2012	2013	IBM Research	New Delhi, India	Ingénieur.e - chercheur.e (EPST) / Research Engineer (public institution)
2010	2012	Indian Institute of Science	Bangalore, India	Autre / Other : Master Student

INTERRUPTION(S) DE CARRIÈRE / CAREER INTERRUPTION(S) :

FORMATION SUPÉRIEURE / HIGHER EDUCATION :

Date of the last academic (PhD) degree: 10-Nov-2017

PRIX, DISTINCTIONS, BOURSES / AWARDS, GRANTS :

-- Student travel awards to attend SC 2016 and IPDPS 2016.
-- Invited for Google PhD Student Summit on Compiler & Programming Technology, Munich, Germany 2014.
-- Recipient of the MHRD Scholarship, India (2010-2012) during my Masters.
-- All-India Rank 28 (top 0.03%) at the Gate Examination 2010, out of a total of about 107,000 candidates.

PRODUCTIONS SCIENTIFIQUES / SCIENTIFIC PRODUCTIONS :

N°	Publications / Publications	Quel est l'apport majeur de cette publication ? / What is the major contribution of this publication?

1	Communication Lower Bounds and Optimal Algorithms for Multiple Tensor-Times-Matrix Computation, SIAM Journal on Matrix Analysis and Applications, Volume 45, 2024 (available at https://inria.hal.science/hal-03950359)	The main contributions of this article are the development of a new communication lower bound and a new algorithm for Multi-TTM computation. This is a key computation in algorithms for computing the Tucker tensor decomposition, which is frequently used in multidimensional data analysis. We proved that with correct choices of processor grid dimensions, the communication cost of our algorithm attains the lower bounds and is therefore communication optimal. I feel this work is impactful from both theoretical and practical points of view. The concepts presented in the article are quite generic and would help one to obtain communication optimal parallel algorithms for different (tensor) computations.
2	Brief Announcement: Tight Memory-Independent Parallel Matrix Multiplication Communication Lower Bounds, ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2022), Jul 2022, Philadelphia, PA, USA. (Extended version is available at https://arxiv.org/abs/2205.13407)	The main contribution of this work is establishing memory-independent communication lower bounds with tight constants for parallel matrix multiplication. Our constants improve on previous work in each of three cases that depend on the relative sizes of the matrix aspect ratios and the number of processors. The ideas introduced in the paper can be easily extended to arbitrary loop nesting and would help one to obtain tight communication lower bounds for computations with uneven dimensions.
3	Parallel Tensor Train through Hierarchical Decomposition (Available at https://hal.inria.fr/hal-03081555)	Most existing approaches try to parallelize certain steps of the sequential algorithms. To the best of our knowledge, we are the first one to modify the Tensor Train algorithm such that the depth of the computation tree is optimal. This approach also reduces the amount of communications and number of messages along the critical path. Our idea of expressing the computation tree as a balanced tree is simple, however it was difficult to prove theoretical guarantees and to decide what should we transmit on both sides of the tree.
4	Performance Models for Data Transfers: A Case Study with Computational Chemistry Kernels, International Conference on Parallel Processing (ICPP 2019), Aug 2019, Kyoto, Japan.	Data transfers are the main bottleneck for many applications on modern HPC systems. The paper addresses this topic and presents strategies to decide the order of data transfers such that the overlap between computations and data transfers is maximum.
5	Approximation Proofs of a Fast and Efficient List Scheduling Algorithm for Task-Based Runtime Systems on Multicores and GPUs IEEE International Parallel & Distributed Processing Symposium (IPDPS 2017), May 2017, Orlando, Florida, USA.	This paper provides a theoretical insight on the performance of a resource centric list scheduling algorithm, HeteroPrio, by proving approximation bounds compared to the optimal schedule in the case of independent tasks on two types of unrelated resources. Our results establish that spoliation (a technique that enables resources to restart uncompleted tasks on another resource) allows to prove bounded approximation ratios for a list scheduling algorithm on two unrelated resources, which is known to be impossible otherwise. We also establish that almost all our bounds are tight.

N°	Jeux de données, codes sources, logiciels, data paper, etc. / Data set, software, source code, data paper, etc.	Description succincte / Brief description
1	TAMM: tensor algebra for many body methods (I worked on this during my postdoc at Pacific Northwest National Laboratory, USA.)	This library is part of the NWChemEx Exascale Computing Project. The NWChemEx project redesigns the functional capabilities of NWChem library for the pre-exascale and exascale systems. TAMM is the compiler of NWChemEx.

N°	Jeux de données, codes sources, logiciels, data paper, etc. / Data set, software, source code, data paper, etc.	Description succincte / Brief description
2	Performance optimizations of TTI RTM on GPU based hybrid architectures (I worked on this during my stay at IBM Research, India.)	TTI RTM algorithm is widely used in seismic imaging. It has huge computational cost which makes it challenging for large scale exploration. We developed and implemented GPU-based parallel kernels for this algorithm for Statoil company.

VALORISATION / VALORISATION :