Vignesh Balaji

Webpage: https://bvignesh.github.io/ | Email: vbalaji@nvidia.com

Research Interests

Architectural and Software Optimizations for Sparse, Irregular Workloads (e.g. graph analytics)

Education

PHD | CARNEGIE MELLON UNIVERSITY | 2015-2021

- Thesis: *Input, Representation, and Access Pattern Guided Cache Locality Optimizations for Graph Analytics*
- Advisor: Brandon Lucia
- Major: Electrical and Computer Engineering (Computer Architecture)
- GPA: 3.86/4
- My thesis research focused on architectural and software optimizations for sparse, irregular memory access workloads (particularly, graph analytics). A fundamental tenet of my research was to leverage the unique properties of graph analytics workloads and input graphs to design locality and scalability optimizations for efficient graph processing on multi-core processors.

B.E. | BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE (BITS) PILANI | 2011-2015

- Thesis: *Design of a Resource Tracker for a Runtime Reconfigurable Coprocessor*
- Advisor: S.K. Nandy (IISc, Bangalore)
- Major: Electronics and Instrumentation
- GPA: 9.34/10 (Department Rank: 1)
- For my undergraduate thesis, I designed the simulation infrastructure to model a resource availability tracker used for scheduling kernels on a dynamically reconfigurable polymorphic coprocessor (REDEFINE).

Professional Experience

RESEARCH SCIENTIST | NVIDIA RESEARCH | JUL 2021 - PRESENT

- I am a member of the Architecture Research Group (ARG) at NVIDIA Research where I am broadly working on optimizing sparse irregular workloads for GPUs

SUMMER INTERN | INTEL LABS | MAY 2019 – AUG 2019

- Explored optimizations for streaming sparse tensor factorization by leveraging temporal characteristics of real-world input tensors

SUMMER INTERN | NVIDIA RESEARCH | MAY 2018 – AUG 2018

- Developed analytical models for an accelerator targeting graph processing and sparse linear algebra. The analytical models were used to explore the trade-off space for on-chip buffer management on the accelerator

SUMMER INTERN | IBM SRDC | MAY 2014 – AUG 2014

- Explored different organizations for Tunnel Field Effect Transistors (TFETs) to produce similar output responses as a CMOS transistors.

SUMMER INTERN | IGCAR KALPAKKAM | MAY 2013 – AUG 2013

- Designed a SoC-based system to detect the health of an electrochemical hydrogen sensor deployed in a Fast Breeder Test Reactor.

Publications

COMMUNITY-BASED MATRIX REORDERING FOR SPARSE LINEAR ALGEBRA OPTIMIZATION [ISPASS 2023]

- Authors: *Vignesh Balaji,* Neal Crago, Aamer Jaleel, Steve Keckler
- Paper in IEEE International Symposium on Performance Analysis of Systems and Software 2023

IMPROVING LOCALITY OF IRREGULAR UPDATES WITH HARDWARE ASSISTED PROPAGATION BLOCKING [HPCA 2022]

- Authors: *Vignesh Balaji,* Brandon Lucia
- Paper in International Symposium on High Performance Computer Architecture 2022
- ***(Nominated for Best Paper Award)***

P-OPT: PRACTICAL OPTIMAL CACHE REPLACEMENT FOR GRAPH ANALYTICS [HPCA 2021]

- Authors: *Vignesh Balaji,* Neal Crago, Aamer Jaleel, Brandon Lucia
- Paper in International Symposium on High Performance Computer Architecture 2021
- ***(Nominated for Best Paper Award)***

OPTIMIZING GRAPH PROCESSING AND PREPROCESSING WITH HARDWARE ASSISTED PROPAGATION BLOCKING [ARXIV 2020]

- Authors: *Vignesh Balaji*, Brandon Lucia

PEACENIK: ARCHITECTURE SUPPORT FOR NOT FAILING UNDER FAIL-STOP MEMORY CONSISTENCY [ASPLOS 2020]

- Authors: Rui Zhang, Swarnendu Biswas, *Vignesh Balaji,* Michael D. Bond, Brandon Lucia
- Paper in International Symposium on Architectural Support for Programming Languages and Operating Systems 2020

COMBINING DATA DUPLICATION AND GRAPH REORDERING TO ACCELERATE PARALLEL GRAPH PROCESSING [HPDC 2019]

- Authors: *Vignesh Balaji*, Brandon Lucia
- Paper in International Symposium on High-Performance Parallel and Distributed Computing 2019

WHEN IS GRAPH REORDERING AN OPTIMIZATION? [IISWC 2018]

- Authors: *Vignesh Balaji*, Brandon Lucia
- Paper in IEEE International Symposium on Workload Characterization 2018
- ***(Won the Best Paper Award)***

FLEXIBLE SUPPORT FOR FAST PARALLEL COMMUTATIVE UPDATES [ARXIV 2017]

- Authors: *Vignesh Balaji*, Dhruva Tirumala, Brandon Lucia

AN ARCHITECTURE AND PROGRAMMING MODEL FOR ACCELERATING PARALLEL COMMUTATIVE COMPUTATIONS VIA PRIVATIZATION [PPOPP 2017]

- Authors: *Vignesh Balaji*, Dhruva Tirumala, Brandon Lucia
- Poster presented in Principles and Practice of Parallel Programming 2017

INTERMITTENT COMPUTING: CHALLENGES AND OPPORTUNITIES [SNAPL 2017]

- Authors: Brandon Lucia, *Vignesh Balaji*, Alexei Colin, Kiwan Maeng, Emily Ruppel
- Paper in Summit on Advances in Programming Languages 2017

OVERCOMING THE DATA-FLOW LIMIT ON PARALLELISM WITH STRUCTURAL APPROXIMATION [WAX 2016]

- Authors: *Vignesh Balaji*, Brandon Lucia
- Paper presented in Workshop on Approximate Computing (WAX) 2016, co-located with ASPLOS 2016

Honors

- **Best Paper Nominee,** HPCA 2022
- **Best Paper Nominee,** HPCA 2021
- **Best Paper Award,** IISWC 2018
- **Deans Fellowship,** Carnegie Mellon University 2015
- **Merit Scholarship,** BITS Pilani 2013-2014

Service

- **Program Committee Member:** MICRO 2023
- **External Review Committee Member:** HPCA 2023, MICRO 2021
- **Reviewer:** IEEE transaction on Computers (Special Issue on Domain-Specific Architectures for Emerging Applications) 2019
- **Shadow Program Committee Member:** ASPLOS 2018

Relevant Coursework (at CMU)

- **Computer Architecture (18-740)** Fall 2015
- **Energy Aware Computing (18-743)** Fall 2015
- **Machine Learning (10-701)** Spring 2016
- **Optimizing Compilers for Modern Architectures (10-701)** Spring 2016
- **Advanced and Distributed Operating Systems (15-712)** Fall 2016
- **Networks in the Real World (18-755)** Fall 2016
- **Parallel Computer Architecture (18-742)** Spring 2017

Skills

- **Languages:** (CUDA) C++, C, Python, x86 assembly
- **Tools/Simulators:** Pin, Sniper, Gem5, perf, LIKWID, nsight, nvprof, PAPI, Intel VTune