NILADRISH CHATTERJEE

Computer Architect

@ nilcsutah@gmail.com ☎ 801-554-1359 📍 Kirkland, WA 🔗 www.niladrish.org Google Scholar

OVERVIEW

Computer architect with more than 10yrs experience in industrial and academic research into high-performance, energy-efficient architectures. At NVIDIA research, I am in charge of inventing novel memory and system architectures that boost bandwidth and capacity for emerging AI and HPC workloads, while improving energy-per-access. My research has been influenced product roadmaps at NVIDIA and in partner memory vendors, and has led to patents and top-tier publications. Currently I lead an interdisciplinary effort into memory architectures to efficiently handle AI and Genomics workloads at scale. I have also been significantly involved in modeling and simulation efforts at NVIDIA and during my PhD - with my open source DRAM and memory-controller simulator, USIMM, having been used widely in academic and industrial memory system research. I regularly serve on the program and review committee of top-tier conferences and journals.

PUBLICATIONS

Memory Systems Architecture

- **HPCA-2018** Reducing Data Transfer Energy by Exploiting Similarity within a Data Transaction.
  D. Lee, M. O'Connor, and N. Chatterjee.
- **MICRO-2017** Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems.
  M. O'Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally.
- **HPCA-2017** Architecting an Energy-Efficient Memory System for GPUs.
  N. Chatterjee, M. O'Connor, D. Lee, D. R. Johnson, M. Rhu, S. W. Keckler, and W. J. Dally.
- **ISPASS-2016** Addressing Service Interruptions in Memory with Thread-to-Rank Assignment.
  M. Shevgoor, R. Balasubramonian, N. Chatterjee, and J. Kim.
- **SC-2014** Managing DRAM Latency Divergence in Irregular GPGPU Applications.
  N. Chatterjee, M. O'Connor, G. H. Loh, N. Jayasena, and R. Balasubramonian.
- **MICRO-2013** Quantifying the Relationship between the Power Delivery Network and Architectural Policies in 3D-Stacked Memory Devices.
  M. Shevgoor, J.-S. Kim, N. Chatterjee, R. Balasubramonian, A. Davis, and A. N. Udipi.
- **MICRO-2012** Leveraging Heterogeneity in DRAM Main Memories to Accelerate Critical Word Access.
  N. Chatterjee, M. Shevgoor, R. Balasubramonian, A. Davis, Z. Fang, R. Illikkal, and R. Iyer.
- **HPCA-2012** Staged-Reads: Mitigating the Impact of DRAM Writes on DRAM Reads.
  N. Chatterjee, R. Balasubramonian, N. Muralimanohar, A. Davis, and N. Jouppi.
- **ISCA-2010** Rethinking DRAM Design and Organization for Energy-Constrained Multi-cores.
  A. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. Jouppi.
- **ASPLOS-2010** Micro-pages: Increasing DRAM Efficiency with Locality-Aware Data Placement.

EXPERIENCE

NVIDIA

Sr. Research Scientist

📅 2013 – Ongoing 📍 Redmond, WA

- Researching fundamental advances in memory systems for HPC and AI platforms.
- Co-inventor of energy-efficient, high-performance DRAM substrate (FGDRAM) for post-HBM systems.
- Processing-near-memory for high-throughput systems.
- Performance modeling and architecture simulation.
- Currently working on scaling GPU memory systems for large-scale recommender systems.

HP Labs

Intern

📅 Fall 2012 📍 SLC, UT

- Disaggregated memory systems for high-capacity datacenters.

AMD Research

Intern

📅 Spring 2012 📍 Sunnyvale, CA

- Memory scheduling algorithms for multi-client SoC memory controllers.
- SIMT-aware memory controller to minimize latency-divergence while balancing sustained throughput.

EDUCATION

Ph.D. in Computer Engineering

University of Utah

📅 2008 – 2013

Designing Efficient Memory Schedulers For Future Systems

Advisor: Dr. Rajeev Balasubramonian

B.E. in Computer Science

Jadavpur University

📅 2003 – 2007

K. Sudan, N. Chatterjee, M. Awasthi, D. Nellans, R. Balasubramonian, and A. Davis.

- **MEMSYS-2016** CLARA: Circular Linked-List Refresh Architecture.
  A. Agrawal, M. O'Connor, E. Bolotin, N. Chatterjee, J. Emer, and S. W. Keckler.
- **MEMSYS-2015** Anatomy of GPU Memory System For Multi-Application Execution. A. Jog, O. Kayiran, E. Bolotin, T. Kesten, A. Pattnayak, N. Chatterjee, S. W. Keckler, M. T. Kandemir, and C. R. Das.

Hardware Architectures for Deep Learning

- **ISPASS-2021** Learning Sparse Matrix Row Permutations for Efficient SpMM on GPU Architectures.
  A. Mehrabi, D. Lee, N. Chatterjee, D. Sorin, B. Lee, M. O'Connor.
- **ISPASS-2019** DeLTA: GPU Performance Model for Deep Learning Applications with In-Depth Memory System Traffic Analysis.
  S. Lym, D. Lee, M. O'Connor, N. Chatterjee, M. Erez.
- **HPCA-2018** Compressing DMA Engine: Leveraging Activation Sparsity for Training Deep Neural Networks.
  M. Rhu, M. O'Connor, N. Chatterjee, J. Pool, Y. Kwon, and S. W. Keckler.

Near-Memory Data Processing

- **SC-2019** Near-Memory Data Transformation for Efficient Sparse Matrix Multi-Vector Multiplication.
  D. Fujiki, N. Chatterjee, M. O'Connor, and D. Lee.
- **SC-2017** Toward Standardized Near-Data Processing with Unrestricted Data Placement for GPUs.
  G. Kim, N. Chatterjee, M. O'Connor, and K. Hsieh.
- **ISCA-2016** Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems.
  K. Hsieh, E. Ebrahimi, G. Kim, N. Chatterjee, M. O'Connor, N. Vijaykumar, O. Mutlu, and S. W. Keckler.

Modeling and Simulation

- **HPCA-2021** Need for Speed: Experiences Building a Trustworthy System-Level GPU Simulator.
  O. Villa, D. Lustig, Z. Yan, E. Bolotin, N. Chatterjee, N. Jiang, D. Nellans.
- **UUCS-TR-12-002** USIMM: the Utah Simulated Memory Module. A Simulation Infrastructure for the JWAC Memory Scheduling Competition.
  N. Chatterjee, R. Balasubramonian, M. Shevgoor, S. H. Pugsley, A. N. Udipi, A. Shaifei, M. Awasthi, K. Sudan, and Z. Chishti
- **SIGMETRICS-2018** What Your DRAM Power Models are Not Telling You: Lessons from a Detailed Experimental Study .
  S. Ghose, A. G. Yaglicki, R. Gupta, D. Lee, K. Kudrolli, W. X. Liu, H. Hassan, K. K. Chang, N. Chatterjee, A. Agrawal, M. O'Connor, and O. Mutlu .
- **SIGMETRICS-2017** Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms.
  K. K. Chang, A. G. Yaglikci, S. Ghose, A. Agrawal, N. Chatterjee, A. Kashyap, D. Lee, M. O'Connor, H. Hassan, and O. Mutlu.

ISSUED PATENTS

10,468,093 Systems and methods for dynamic random access memory (DRAM) sub-channels

9,910,605 Page migration in a hybrid memory device

9,846,550 Memory access methods and apparatus (divisional of 9,361,955)

9,535,831 Page migration in a 3D stacked hybrid memory

9,489,321 Scheduling memory accesses using an efficient row burst value

9,361,955 Memory Access Methods and Apparatus

HONORS

Best Paper Award
ISPASS 2016

Best Paper Nominee
HPCA 2018

Best Poster Awards
HiPC 2009. School of Computing Graduate Research Competition 2010 and 2013

School of Computing Fellowship
University of Utah teaching fellowship 2008. Awarded to outstanding incoming graduate students.

Media Coverage
HPCA 2017 work on GPU DRAM microarchitecture covered on The Next Platform

SERVICE

Conference Program Committee Member
HPCA 2019, HPCA 2018, IPDPS 2017, ICS 2016

External Review Committee Member
MICRO 2021/2019/2012, HPCA 2021/2015/2012, ISCA 2015/2014, HPG 2017

Journal Reviewer
ACM TACO

Co-organizer
3rd JILP Workshop on Computer Architecture Competitions (Memory Scheduling Championship). Held with ISCA-2012.