Suvinay Subramanian

Contact suvinay@csail.mit.edu, suvinay@google.com http://suvinay.com

Education

Massachusetts Institute of Technology (MIT) Cambridge, USA
S.M. (Jun'13), Ph.D. (Sep'18) in Electrical Engineering and Computer Science Sep'11—Sep'18
Advisors: Prof. Daniel Sanchez (Ph.D.) and Prof. Li-Shiuan Peh (S.M.)

Indian Institute of Technology (IIT) Madras Chennai, India
B.Tech in Electrical Engineering Aug'07—Jul'11

Awards and Distinctions

- ACM SIGMICRO Early Career Award. (2024)
- Honorable mention IEEE Micro "Top Picks from the Computer Architecture Conferences". (2017)
- IEEE Micro "Top Picks from the Computer Architecture Conferences". (2016)
- Finalist, Qualcomm Innovation Fellowship. (2014)
- Institute Blues Award, IIT Madras for all-round excellence. (2011)
- OP Jindal Engineering and Management Scholarship (OPJEMS). (2010)
- Honda Young Engineer and Scientist (YES) Award. (2009)
- Olympiad Achievements: (2007)
  - Gold medalist Indian National Chemistry Olympiad (INChO).
  - India top 1% in Indian National Physics Olympiad (INPhO).
  - Ranked 31st in India in National Science Olympiad (NSO).
- National Talent Search Scholar. 1st rank in Karnataka state. (2005—11)

Work Experience

Staff Software Engineer Google
Systems Infrastructure | ML Performance Team *Nov'18—Present*

- Architecture and codesign of high-performance custom AI processors and systems (TPUs), that power Google's AI capabilities including large embedding models (LEMs or recommendation models) and large language models (LLMs like Gemini).
- Core architect of the SparseCore, a novel dataflow processor for sparse, irregular AI workloads. Enables 2x improved performance for training production large embedding models (LEMs or recommendation models), that generate O(billions) of dollars of revenue (Ads, Search, YouTube).
- Performance engineering for multiple AI workloads, including high-value optimizations (e.g., multi-operator fusion FlatAttention) widely deployed for flagship LLMs (Gemini), and codesign for emerging deep learning paradigms such as mixture-of-experts (MoE) and sparsity.
- Developed hardware and system simulators in C++/Python, tuned assembly-level kernels, compiler optimizations, parallelization strategies and application-level performance tuning.
- Liaison and work with professors and graduate students in researching ideas, and identify promising techniques for TPU systems.

Research Assistant MIT
Advisor: Prof. Daniel Sanchez, EECS Department *Sep'14—Sep'18*

- Worked on Swarm a novel hardware-software codesigned architecture for pervasive parallelism.
- Designed new execution model and hardware multi-core architecture that employs aggressive speculation to scale difficult-to-parallelize ordered irregular applications near linearly to hundreds of cores.
- Developed Pin-based simulator for studying multi-core architectures up to 1000-cores, new benchmark suite of ordered irregular applications.

Teaching Assistant MIT
Instructors: Prof. Daniel Sanchez, Prof. Joel Emer, EECS Department *Feb'16—May'16*

- 6.823: Computer System Architecture, a graduate-level course spanning ISA and out-of-order instruction pipelines, to virtual memory, multicores, and memory systems. Developed new lab using Murphi for formal verification of cache-coherence protocol. Led 20 students in weekly recitations.

Research Assistant MIT
Advisors: Prof. Hari Balakrishnan, Prof. Mohammad Alizadeh, EECS Dept. *Sep'13—Dec'15*

- Designed a new abstraction and hardware primitive, Push-In-First-Out (PIFO) queue for programmable packet scheduling at line-rate in high-speed network routers. Synthesized prototype PIFO in 32 nm technology node.

Research Assistant MIT
Advisor: Prof. Li-Shiuan Peh, EECS Department *Sep'11—Dec'14*

- Member of team at MIT that designed and fabricated SCORPIO, a 36-core snoopy-coherent multi-core processor in 45 nm technology node.

- Designed and implemented ordered mesh network-on-chip (NoC) employing a novel distributed ordering scheme and supporting in-network snoopy coherence.
- Research and performance analysis on multiple network-on-chip (NoC) architectural ideas.

Research Intern Nvidia Research
Manager: Dr. Steve Keckler, Computer Architecture Research Group *May'14—Aug'14*

- Developed a memory system simulator and explored policies for heterogenous memory management in CPU-GPU systems.

Graduate Technical Intern Intel Labs
Manager: Dr. Mani Azimi, Platform Architecture Research Group *Jun'12—Aug'12*

- Studied scalability of on-chip interconnection networks to large-scale multicores (>100 core).
- Developed power and performance models, and explored topology alternatives for large-scale on-chip interconnection networks.

Undergraduate Researcher IIT Madras
Advisor: Prof. V. Kamakoti, CSE Department *Oct'10—Apr'11*

- Designed new algorithm to identify illegal states in a circuit for pseudo-functional testing of small-delay defects.

Undergraduate Research Intern UNB, Canada
Advisor: Prof. David Bremner, CS and Math Department *May'10—Jul'10*

- Developed a custom SAT solver that employed ideas from graph isomorphism to exploit structure and symmetry in the underlying problem of geometric realizability of convex polytopes.

Undergraduate Technical Intern Texas Instruments
Manager: Dr. Srivaths Ravi, DFT Lead, Texas Instruments India *May'09—Jul'09*

- Developed a framework and setup for gate level statistical power estimation on a 65 nm System-on-Chip (SoC). Automated large parts of the power estimation flow. Correlated power estimates from the framework with real silicon numbers.

Publications & Invited Talks

Effective Interplay between Sparsity and Quantization: From Theory to Practice ICLR 2025
S.B. Harma, A. Chakraborty, E. Kostenok, D. Mishin, D. Ha, B. Falsafi, M. Jaggi, M. Liu, Y. Oh, **S. Subramanian**, A. Yazdanbakhsh

The Journey Matters: Average Parameter Count over Pre-training Unifies Sparse and Dense Scaling Laws ICLR 2025
T. Jin, A.I. Humayun, U. Evci, **S. Subramanian**, A. Yazdanbakhsh, D. Alistarh, G.K. Dziugaite

Progressive Gradient Flow for Robust N:M Sparsity Training in Transformers CPAL 2025
A.R. Bambhaniya, A. Yazdanbakhsh, **S. Subramanian**, S.C. Kao, S. Agrawal, U. Evci, T. Krishna

Codesigning Computing Systems for Artificial Intelligence Talks 2023–24
MIT, Stanford University, Columbia University, Georgia Institute of Technology, Harvard University, New York University, KAIST, ISCA Keynote @ CogArch, University of South Carolina, UC Irvine, AMD, Cruise

JaxPruner: A Concise Library for Sparsity Research CPAL 2024
J.H. Lee, W. Park, N. Mitchell, J. Pilault, J. Obando-Ceron, H.B. Kim, N. Lee, E. Frantar, Y. Long, A. Yazdanbakhsh, S. Agrawal, **S. Subramanian**, X. Wang, S.C. Kao, X. Zhang, T. Gale, A. Bik, W. Han, M. Ferev, Z. Han, H.S. Kim, Y. Dauphin, G.K. Dziugaite, P.S. Castro, U. Evci

TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings ISCA 2023 (Ind.)
N.P. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, **S. Subramanian**, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, D. Patterson

FLAT: An Optimized Dataflow for Mitigating Attention Performance Bottlenecks ASPLOS 2023
S-C. Kao, **S. Subramanian**, G. Agrawal, T. Krishna

STEP: Learning N:M Structured Sparsity Masks from Scratch with Precondition ICML 2023
Y. Lu, S. Agrawal, **S. Subramanian**, O. Ryabakov, C. De-Sa, A. Yazdanbakhsh

Harmonizing Speculative and Non-Speculative Execution in Architectures for Ordered Parallelism MICRO 2018
M.C. Jeffrey, V.A. Ying, **S. Subramanian**, H.R. Lee, J. Emer, D. Sanchez

SAM: Optimizing Multithreaded Cores for Speculative Parallelism PACT 2017
M. Abeydeera, **S. Subramanian**, M.C. Jeffrey, J. Emer, D. Sanchez

Fractal: An Execution Model for Fine-Grain Nested Speculative Parallelism ISCA 2017
S. Subramanian, M.C. Jeffrey, M. Abeydeera, H.R. Lee, V.A. Ying, J. Emer, D. Sanchez

Data-Centric Execution of Speculative Parallel Programs .... MICRO 2016
M.C. Jeffrey, **S. Subramanian**, M. Abeydeera, J. Emer, D. Sanchez

Programmable Packet Scheduling .... SIGCOMM 2016
A. Sivaraman, **S. Subramanian**, A. Agrawal, S. Chole, S.T. Chuang, T. Edsall,
M. Alizadeh, S. Katti, N. McKeown, H. Balakrishnan

Unlocking Ordered Parallelism with the Swarm Architecture .... IEEE Micro 2016
M.C. Jeffrey, **S. Subramanian**, C. Yan, J. Emer, D. Sanchez
IEEE Micro's Top Picks from the Computer Architecture Conferences 2016

A Scalable Architecture for Ordered Parallelism .... MICRO 2015
M.C. Jeffrey, **S. Subramanian**, C. Yan, J. Emer, D. Sanchez

SCORPIO: A 36-core Research Chip Prototype Demonstrating Snoopy Coherence on a Scalable Mesh NoC with In-Network Ordering .... ISCA 2014
B.K. Daya, C.H.O. Chen, **S. Subramanian**, W.C. Kwon, S. Park, T. Krishna, J. Holt,
A. Chandrakasan, L.S. Peh

No Silver Bullet: Extending SDN to the Data Plane .... HotNets 2013
A. Sivaraman, K. Winstein, **S. Subramanian**, H. Balakrishnan

Single-Cycle Multihop Asynchronous Repeated Traversal: A SMART Future for Reconfigurable On-Chip Networks ... IEEE Computer 2013
T. Krishna, C.H.O. Chen, S. Park, W.C. Kwon, **S. Subramanian**, A. Chandrakasan, L.S. Peh

SMART: A Single-Cycle Reconfigurable NoC for SoC Applications .... DATE 2013
C.H.O. Chen, S. Park, T. Krishna, **S. Subramanian**, A. Chandrakasan, L.S. Peh

Patents

Programmable Accelerator for Data-Dependent, Irregular Operations .... US Patent 2023
R. Nagarajan, **S. Subramanian**, A.C. Jacob, C. Leary, T.J. Norrie, T.M. Vijayaraj, H. Hariharan

Sparse SIMD Cross-lane Processing Unit .... US Patent 2023
R. Nagarajan, **S. Subramanian**, A.C. Jacob

Streaming Transfers and Ordering Model .... US Patent 2023
R. Nagarajan, A.C. Jacob, **S. Subramanian**, H. Hariharan

Computer Skills

Languages	C, C++, Python, Perl, System Verilog, SQL.
Operating Systems	Linux, Mac OS X, Windows.
Tools	Intel Pin, MATLAB, SPICE, Cadence Virtuoso, Encounter, RTL Compiler, Synopsys DC, PrimePower, Tetramax, Nanosim.

Leadership, Service & Extra-curricular Activities

- Artifact Evaluation Co-chair: ASPLOS'21, ASPLOS'22, MICRO'23 (2021—23)
- Workshop Co-chair: Young Architect ASPLOS'23, MLArchSys ISCA'23,'24 (2023—25)
- Reviewer for: TC'18, TACO'19, CSUR'19, CAL'20, HPCA'20, MLSys'22, TCSI'22, ISCA'23, ASPLOS'23, MICRO'24, ISCA'25, ICLR'25 (2018—25)
- Co-host the Computer Architecture Podcast ($\sim 50K$ downloads) (2021—Present)
- Vice President and other operating roles, South India Fine Arts (SIFA), Bay Area (2021—2024)
- Cultural Chair, Indian Students Association, Sangam-MIT (2012—13)
- EECS Graduate Student Council (GSC) Representative, MIT (2011—12)
- Alumni Affairs Secretary, Saraswathi Hostel, IIT Madras (2009—10)
- Quality Management System Coordinator, Shaastra 2009, IIT Madras (2009)
- Web Operations Coordinator, Shaastra 2008, IIT Madras (2008)
- Trained in south Indian classical vocal, and percussion instrument, Mridangam (2000—Present)
- Member, MIT Ohms, acapella group. Released two albums, ICCA quarter-finalists. (2013—16)