Cyan Subhra Mishra

Graduate (Performance and Power) Engineer, Arm
email: cyanmishra92@gmail.com | Linkedin: cyan-subhra-mishra | Homepage

Summary

Specializing in hardware/software co-design for ML systems with expertise in accelerator architecture, performance modeling, and energy-efficient computing. Demonstrated experience optimizing neural network workloads across heterogeneous platforms (GPUs, FPGAs, ReRAM) with particular focus on kernel optimization, model compression techniques, and system-level performance for resource-constrained environments.
Key words: *Deep Learning at Edge, Continuous Learning Systems, Intermittent Computing, Non-volatile Processors, Energy Harvesting-Wireless Sensor Networks, Computational Storage*

Education

The Pennsylvania State University — University Park, PA
Ph.D. in Computer Science and Engineering — *Aug 2018 – Aug 2025 (Expected)*
Advisors: **Dr. Mahmut Taylan Kandemir, Dr. Jack Sampson** — Current GPA: 3.7/4.00

National Institute of Technology, Rourkela — Rourkela, Odisha, India
B.Tech + M.Tech Dual Degree in Electronics and Communication Engineering — *Aug 2011 – May 2016*
Advisors: **Dr. Sarat Kumar Patra, Tarjinder Singh** (Intel) — CGPA: 8.39/10.00 (***Honors***)

Experience

Graduate Engineer: SoC Performance Engineering — Jan 2026 – Present
Arm Inc. — *San Diego, CA*

- Performance and power modeling for next-generation System-on-Chip targeting emerging domains such as Edge AI, Internet of Things, and automotive, informing architectural trade-off decisions and design prioritization.
- Collaborating with cross-functional architecture, microarchitecture, and systems teams to quantify latency, throughput, energy, and area impacts of compute subsystems using analytical and empirical methods, guiding optimization strategies across hardware and software stacks.
- Developing and maintaining performance evaluation and automation frameworks using Python and C++ for large-scale workload profiling, simulation, data analysis, and regression benchmarking across workloads.
- Characterizing diverse workloads using performance profiling, hardware counters, and custom telemetry pipelines to identify bottlenecks and optimization opportunities.
- Communicating detailed performance insights to engineering stakeholders, elevating performance awareness and influencing product development roadmaps.
- Integrating performance data into continuous integration workflows, enabling automated performance regression detection and ensuring stability of architectural changes across development cycles.

Graduate Research Assistant — December 2018 – December 2025
Microsystems Design Lab, Penn State — *University Park, PA*

- Designing and implementing hardware/software co-design methodologies for ML systems, with focus on optimizing performance, energy efficiency, and resource utilization across heterogeneous computing platforms.
- Developing novel architectural solutions for efficient deployment of DNNs, LLMs, and RAGs on resource-constrained edge devices, achieving up to 22% higher accuracy with minimal computational overhead.
- Engineering high-throughput computational storage architectures for ML workloads that reduce data movement by 6.1× while improving system-level performance.
- Creating performance modeling frameworks for multi-dimensional optimization of large-scale ML deployments, balancing latency, accuracy, and energy constraints in both edge and cloud environments.

Research Intern — June 2021 – Aug 2021
Bell Labs — *Murray Hill, NJ*

- Developed optimization strategies for autonomous ML inference serving across heterogeneous platforms (GPUs, FPGAs), leveraging Apache TVM for cross-platform kernel optimization.
- Implemented model compression techniques including pruning, quantization, and knowledge distillation to improve inference efficiency while maintaining accuracy targets.

Design Engineer June 2016 – June 2018
Intel *Bengaluru, India*

- Led hardware/software co-design initiatives for ML accelerators, implementing systematic performance modeling methodologies for GPU and FPGA platforms.
- Optimized ML kernels (convolution, softmax) for FPGA deployment, balancing computational efficiency with resource utilization through microarchitectural innovations.
- Conducted comprehensive timing analysis and performance characterization for large-scale ML workloads across heterogeneous computing environments.
- Developed software/hardware simulation frameworks to validate accelerator designs, enabling rapid iteration and performance optimization prior to physical deployment.

Research Intern Dec 2015 – June 2016
Intel *Bengaluru, India*

- Designed FPGA-based hardware accelerators for protein search algorithms, achieving significant speedups over CPU implementations.
- Leveraged OpenCL for rapid deployment and optimization of bioinformatics kernels (pairHMM, HMMer) on FPGA platforms, establishing performance benchmarks for production environments.

Research Intern May 2014 – July 2014
CSRE, IIT Bombay *Mumbai, India*

- Developed computational models for pre-processing hyperspectral satellite imagery, optimizing algorithms for data extraction and feature identification.

Research Intern May 2013 – July 2013
Centre for Artificial Intelligence and Robotics *Bengaluru, India*

- Developed a generic method for azimuthal map projection, implementing efficient algorithms for coordinate transformation.

Technical Skills

Hardware Architecture: GPU microarchitecture, FPGA acceleration, ReRAM crossbars, systolic arrays, energy-harvesting systems
Accelerator Programming: CUDA toolkit, OpenCL, Triton, parallel programming optimization
ML Systems: TensorFlow, PyTorch, Apache TVM, model compression, kernel optimization
Performance Engineering: Nvidia Nsight Compute/Systems, Intel vTune, workload characterization, power/performance modeling
Hardware Design: SystemVerilog, Xilinx Vivado, Design Compiler, microarchitecture simulation

Publications

Under Review

- [**R2025#3**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Soumya Prakash Mishra, Rui Zhang, Jack Sampson, Mahmut Taylan Kandemir, Chita R Das. **"Prophet: Neural Expert Prediction for Efficient Mixture-of-Experts Inference"** [link]

- [**R2025#2**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Jack Sampson, Mahmut Taylan Kandemir, Chita R Das. **"Hardware-Aware Neural Network Co-Design with Analog Activation for Energy-Efficient ReRAM Crossbars"** [link]

- [**R2025#1**] Deeksha Chaudhary, Rishabh Jain, **Cyan Subhra Mishra**, Mahmut Taylan Kandemir, Chita R Das. **"MaestroRAG: Orchestrated Pipeline Architecture for Efficient RAG on Edge Devices"**

arXiv Preprints

- [**arXiv-2024**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Jack Sampson, Mahmut Taylan Kandemir. **"Salient Store: Enabling Smart Storage for Continuous Learning Edge Servers"** arXiv preprint arXiv:2410.05435 (2024).

- [**arXiv-2024**] **Cyan Subhra Mishra**, Jack Sampson, Mahmut Taylan Kandemir, Vijaykrishnan Narayanan. **"Synergistic and Efficient Edge-Host Communication for Energy Harvesting Wireless Sensor Networks"** arXiv preprint arXiv:2408.14379 (2024).

- [**arXiv-2024**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Jack Sampson, Mahmut Taylan Kandemir, and Chita R. Das.**"Revisiting DNN Training for Intermittently Powered Energy Harvesting Micro Computers"** arXiv preprint arXiv:2408.13696 (2024)

- [**arXiv-2020**] **Cyan Subhra Mishra**, Jack Sampson, Mahmut Taylan Kandemir, Vijaykrishnan Narayanan. **"Seeker: Synergizing Mobile and Energy Harvesting Wearable Sensors for Human Activity Recognition."** arXiv preprint arXiv:2204.13106 (2022).

- [**arXiv'2020**] Jashwant Raj Gunasekaran , Prashanth Thinakaran, **Cyan Subhra Mishra**, Mahmut Taylan Kandemir, and Chita R. Das. **"Towards Designing a Self-Managed Machine Learning Inference Serving System inPublic Cloud."** arXiv preprint arXiv:2008.09491 (2020).

Conference Papers

- [**PACT-2025**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Mahmut Taylan Kandemir, Chita R Das. **"Salient Store: Enabling Smart Storage for Continuous Learning Edge Servers"** International Conference on Parallel Architectures and Compilation Techniques (PACT '25).

- [**ICLR-2025**] **Cyan Subhra Mishra**, Deeksha Chaudhary, Jack Sampson, Mahmut Taylan Kandemir, and Chita R. Das.**"Revisiting DNN Training for Intermittently–Powered Energy–Harvesting Micro–Computers"**; [TO APPEAR]

- [**IPDPS-2025**] Wahid Uz Zaman, **Cyan Subhra Mishra**, Saleh AlSaleh, Abutalib Aghayev, and Mahmut Taylan Kandemir.**"CORD: Parallelizing Query Processing across Multiple Computational Storage Devices"**; [TO APPEAR]

- [**HPCA-2024**] **Cyan Subhra Mishra**, Jack Sampson, Mahmut Taylan Kandemir, Vijaykrishnan Narayanan and Chita R. Das.**"Uṣás: A Sustainable Continuous-Learning Framework for Edge Servers"**; In 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 891-907. IEEE, 2024.

- [**ICDCS-2022**] Ziyu Ying, Shulin Zhao, Haibo Zhang, **Cyan Subhra Mishra**, Sandeepa Bhuyan, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das.**"Exploiting Frame Similarity for Efficient Inference on Edge Devices"**; In 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS), pp. 1073-1084. IEEE, 2022.

- [**MICRO-2022**] Ziyu Ying, Shulin Zhao, Sandeepa Bhuyan, **Cyan Subhra Mishra**, Mahmut Kandemir, Chita R. Das. **"Pushing Point Cloud Compression to Edge"**; In 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO) (pp. 282-299). IEEE. In 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO) (pp. 282-299). IEEE.

- [**NSDI-2022**] Jashwant Raj Gunasekharan, **Cyan Subhra Mishra**, Prashanth Thinakaran, Bikash Sharma, Mahmut T Kandemir, Chita R. Das, **"Cocktail: A Multidimensional Optimization for Model Serving in Cloud"**, 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2022.

- [**NAS-2021**] Jashwant Raj Gunasekaran, **Cyan Subhra Mishra**, **"MLPP: Exploring Transfer Learning and Model Distillation for Predicting Application Performance"**, IEEE Network Architecture and Storage 2021 (NAS'21), 2021.

- [**SoCC-2021**] Vivek M. Bhas, Jashwant Raj Gunasekharan, Prashanth Thinakaran, **Cyan Subhra Mishra**, Mahmut T Kandemir, Chita R. Das, **"Kraken : Adaptive Container Provisioning for Deploying Dynamic DAGs in Serverless Platforms"**, ACM Symposium on Cloud Computing 2021 (SoCC'21), 2021.

- [**MICRO-2021**] Shulin Zhao, Haibo Zhang, **Cyan Subhra Mishra**, Sandeepa Bhuyan, Ziyu Ying, Mahmut T Kandemir, Chita R. Das, **"HoloAR: On-the-fly Optimization of 3D HolographicProcessing for Augmented Reality"**, 54th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2021.

- [**DATE-2021**] **Cyan Subhra Mishra**, Jack Sampson, Mahmut T Kandemir, Vijaykrishnan Narayanan, **"Origin: Enabling On-Device Intelligence for Human Activity Recognition Using Energy Harvesting Wireless Sensor Networks"**, Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2021. [**Best Paper Nominee**]

- [**WoSC-2020**] Jashwant Raj Gunasekaran, **Cyan Subhra Mishra**, Prashanth Thinakaran, Mahmut T Kandemir, Chita R Das, **"Implications Of Public Cloud Resource Heterogeneity for Inference Serving"**, Proceedings of the 6th International Workshop on Serverless Computing 2020.

- [**ISCA-2020**] Shulin Zhao, Haibo Zhang, Sandeepa Bhuyan, **Cyan Subhra Mishra**, Ziyu Ying, Mahmut T. Kandemir, Anand Sivasubramaniam, Chita R. Das, **"Déjà view: Spatio-temporal compute reuse for energy-efficient 360° VR video streaming"**, In Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture 2020 (**ISCA**) .

- [**HPCA-2020**] Keni Qiu, Nicholas Jao; Mengying Zhao, **Cyan Subhra Mishra**, Gulsum Gudukbay, Sethu Jose, Jack Sampson, Mahmut T Kandemir, Vijaykrishnan Narayanan, **"ResiRCA: A resilient energy harvesting ReRAM-based accelerator for intelligent embedded processors"**, In Proceedings of IEEE International Symposium on High Performance Computer Architecture 2020 (**HPCA**).

Journal Papers

- [**ASTM-SSMS-2022-0031**] Abhishek Hanchate, Parth Sanjaybhai Dave, Ankur Verma, Akash Tiwari, **Cyan Subhra Mishra**, Soundar R. T. Kumara, Anil Srivastava, Hui Yang, Vijaykrishnan Narayanan, John Morgan Sampson, Mahmut Taylan Kandemir, Kye-Hwan Lee, Tanna Marie Pugh, Amy Jorden, Gautam Natarajan, Dinakar Sagapuram, and Satish T. S. Bukkapatnam **"A Graphical Representation of Sensor Mapping for Machine Tool Fault Monitoring and Prognostics for Smart Manufacturing. "**

- [**IEEE-ESL-2022**] Tianyi Shen, **Cyan Subhra Mishra**, Jack Sampson, Mahmut Taylan Kandemir, Vijaykrishnan Narayanan.**"An Efficient Edge-Cloud Partitioning of Random Forests for Distributed Sensor Networks."**

- [**Defence Science Journal-2015**]; Narayan Panigrahi, **Cyan Subhra Mishra**, **"A Generic Method for Azimuthal Map Projection"**, Defence Science Journal 65 (5).

Academic Projects

Preliminary Works

- **NExUME:** Dynamically adjusts dropout and quantization in response to intermittent energy, letting DNNs adapt to fluctuating power. Uses an intermittency-aware optimizer (*DynFit*) and scheduling (*DynInfer*) to achieve up to 22% higher accuracy with minimal extra compute. **Key Concepts:** Energy Harvesting, Intermittent Computing, Dynamic Dropout/Quantization.

- **Seeker:** Offloads only compressed "coresets" from ultra-low-power sensors to the host, cutting communication by up to 8.9×. Sensors run partial DNNs with store-and-execute on non-volatile memory; the host completes inference on the coreset. **Key Concepts:** Coreset Compression, EH-WSN, Partial Inference, Edge-Host Synergy.

- **SaLT:** Uses computational storage for neural compression, encryption, and redundancy in continuous-learning edge servers. FPGA-accelerated storage drives offload data-intensive tasks, saving data movement and boosting throughput by 6×. **Key Concepts:** Computational Storage, Neural Compression, FPGA Acceleration, Edge Servers.

- **Usás:** Runs continuous learning on solar-powered edge servers with no batteries. A morphable accelerator reshapes itself to available power, while a teacher-student approach helps train robust models. Saves hundreds of kWh/year compared to standard continuous-learning setups. **Key Concepts:** Battery-Free, Morphable Accelerator, Continuous Learning, Solar-Powered.

- **Origin:** Schedules and ensembles multiple energy-harvesting sensors for human activity recognition. Each node runs a compact DNN, coordinated by an activity-aware scheduler and ensemble aggregator. Improves robustness and achieves $\sim$ 2.5%–5% higher accuracy under tight power constraints. **Key Concepts:** Human Activity Recognition, Ensemble Learning, EH-WSN, Scheduling.

Other Works

- **Cloud Computing:** Analyzes resource heterogeneity in public clouds for inference services, focusing on cost efficiency and performance. Proposes a scheduling approach to map inference requests to diverse VM instance types. Designing a resource management framework for dynamic DAG-based serverless platforms.

Minimizes container over-provisioning in function-chains by combining proactive and reactive scaling, plus a weighting policy for short-latency workflows. Designing ensembling-based model-serving framework in the public cloud, balancing cost, accuracy, and latency. Employs dynamic model subset selection, weighted majority voting, and transient VM usage to reduce deployment costs and maintain SLO.

- **Point Cloud:** Targets point cloud compression for real-time and energy-constrained systems. Proposes a two-stage pipeline for geometry (via parallel octree) and attribute compression, achieving up to 35× speedup and major energy savings on edge.
- **AR/VR:** Explores spatio-temporal compute reuse in 360° video streaming on mobile VR headsets. Skips redundant projection computations, reducing total power and meeting real-time constraints. Focuses on holographic rendering for AR headsets, using approximate or foveated-based strategies to reduce compute overhead. Demonstrates 2.7× speedup and 73% energy savings compared to naive holographic rendering.
- **Intermittent Computing:** Presents a reconfigurable ReRAM crossbar accelerator for DNN inference on intermittently-powered IoT devices. Dynamically adjusts partial activation under fluctuating harvested power, maintaining high energy efficiency and throughput.
- **Edge:** Proposes efficient edge–cloud partitioning for Random Forest inference. Introduces threshold-based offloading and training to preserve data privacy and meet latency objectives across distributed sensor deployments. Exploits frame-level and region-level similarity for video analytics on low-power edge devices. Skips redundant DNN computations in object detection, leading to speedups and energy improvements.

Talks

- Cocktail: A Multidimensional Optimization for Model Serving in Cloud: NSDI'22, Renton, Washington, USA, April'22.
- MLPP: Exploring Transfer Learning and Model Distillation for Predicting Application Performance: NAS'21, Riverside, California, October'21.
- PowerPrep: A power management proposal for user-facing datacenter workloads: NAS'21, Riverside, California, October'21 [**Best Paper Award**]
- Origin: Enabling On-Device Intelligence for Human Activity Recognition Using Energy Harvesting Wireless Sensor Networks: DATE'21, Online.
- Future of Machine Learning in Computer Science; Invited Talk: IIIT Vadodara, India, November 2017.

Teaching Experience

Instructor — Penn State
CMPEN 431: Introduction to Computer Architecture — *Spring 2024, Spring 2022*

Guest Lecture — Penn State
CSE 530: Computer Architecture – Graduate Level — *Fall 2023, Fall 2024*

Guest Lecture — Penn State
CMPEN 431: Introduction to Computer Architecture — *Fall 2024, Spring 2023*

Guest Lecturer — Penn State
CMPSC 497: Architecture for Deep Leaning — *Fall 2019*

Teaching Assistant — Penn State
CMPEN 270 – Introduction to Digital Design — *Fall 2018*

Teaching Assistant — NIT Rourkela
EC 270 – Basic Electronics Laboratory — *Fall 2015*

Course/Hobby Projects

- Branch Prediction Simulator
  Python-based simulation tool to evaluate the performance of different branch prediction algorithms.
- Few-shot Facial Recognition
  Using few-shot learning techniques and efficient memory organizations for edge devices.
- Distributed File System Management
  Distributed file system for both client and server side using RPC calls.
- Synchronization using Path Expressions
  Solving the bounded-buffer and readers-writers synchronization problem using path expressions.
- Compression of DNNs
  Exploring different DNN compression techniques like sparsity, pruning, quantization, coresets, SVD and knowledge distillation for making larger DNNs more suitable for low power low compute devices.
- AcuLive Classifier
  A naïve classifier to tell recorded voice against real voice for digital personal assistants

Honors AND Awards

2021 **Student Travel Award for NAS'21**
2021 **Best Paper Nomination of DATE'21**

Services and Memberships

Submission Chair: IISWC: IEEE International Symposium on Workload Characterization, 2025
Reviewer: TPDS: Transactions on Parallel and Distributed Systems, 2024, 2023, 2019
Reviewer: TC: Transactions on Computers, 2024, 2023, 2021, 2022
Student Member: ACM, IEEE

Relevant Coursework

Graduate: Large Scale Machine Learning, Computer Architecture, Operating System, Compilers.
Online Courses: Deep Learning, Introduction to Parallel Programming, Quantum Computation.