

# Cyan Subhra Mishra

PhD Candidate, Computer Science and Engineering, The Pennsylvania State University  
cyan@psu.edu | cyanmishra92@gmail.com | linkedin.com/in/cyan-subhra-mishra | Personal Website

## Professional Summary

Hardware/software co-design specialist with extensive experience optimizing ML workloads across heterogeneous compute platforms. Expert in accelerator architecture, kernel optimization, and performance modeling for resource-constrained environments. Proven track record designing and implementing energy-efficient systems for large-scale ML deployment, with particular focus on model compression, microarchitectural innovations, and computational efficiency.

**Key words:** *Deep Learning at Edge, Continuous Learning Systems, Resource Aware ML, Intermittent Computing, ML-Systems, ML-Architecture, Power-aware design, Computational Storage*

## Education

<b>PhD, Computer Science and Engineering</b> Pennsylvania State University	<b>2018-2025 (Expected)</b> University Park, PA
Advisors: Dr. Mahmut Taylan Kandemir, Dr. Jack Sampson	GPA: 3.7/4.0
<b>B.Tech + M.Tech Dual Degree, Electronics and Communication Engineering</b> National Institute of Technology, Rourkela	<b>2011-2016</b> Odisha, India
Advisors: Dr. Sarat Kumar Patra, Tarjinder Singh (Intel)	CGPA: 8.39/10.00 (Honors)

## Technical Expertise

<b>Programming/ Compiler:</b>	Python, C++, LLVM, Apache TVM
<b>Hardware Architecture:</b>	Microarchitecture, FPGA, ReRAM crossbars, Systolic arrays, Low-power systems
<b>Accelerator Programming:</b>	CUDA toolkit, OpenCL, Low-precision computing, Parallel programming optimization
<b>ML Systems:</b>	TensorFlow, PyTorch, Quantization, Pruning, Kernel optimization
<b>Performance Engineering:</b>	Workload characterization, Power/Performance modeling, Microarchitecture simulation, profiling tools (Nsight, vTune, Pin tool, Valgrind)
<b>Hardware Design:</b>	SystemVerilog, Xilinx Vivado, Design Compiler, Hardware simulation

## Professional Experience

<b>Graduate Research Assistant</b>	<b>Microsystems Design Lab, Penn State (2018-Present)</b>
<ul style="list-style-type: none"><li>Led hardware/software co-design initiatives for ML systems at scale, optimizing performance, energy efficiency, and resource utilization across heterogeneous computing platforms.</li><li>Developed novel architectural solutions for efficient deployment of large language models on resource-constrained devices, achieving up to <b>22% accuracy improvement</b> with minimal computational overhead.</li><li>Engineered high-throughput computational storage architectures for data-intensive ML workloads, reducing data movement by <b>6.1×</b> while maintaining system-level performance.</li><li>Created comprehensive performance modeling frameworks for multi-dimensional optimization of ML deployments, balancing latency, accuracy, and energy constraints in both edge and cloud environments.</li></ul>	

<b>Research Intern</b>	<b>Bell Labs (Summer 2021)</b>
<ul style="list-style-type: none"><li>Developed optimization strategies for autonomous ML inference serving across heterogeneous hardware (GPUs, FPGAs), leveraging Apache TVM for cross-platform kernel optimization.</li><li>Implemented model compression techniques including quantization, pruning, and knowledge distillation to improve inference efficiency while maintaining accuracy targets.</li></ul>	

<b>Design Engineer</b>	<b>Intel (2016-2018)</b>
<ul style="list-style-type: none"><li>Led hardware/software co-design initiatives for ML accelerators, implementing systematic performance modeling methodologies for GPU and FPGA platforms.</li></ul>	

- Optimized ML kernels (convolution, softmax) for FPGA deployment, balancing computational efficiency with resource utilization through microarchitectural innovations.
- Conducted comprehensive timing analysis and workload characterization for large-scale ML deployment across heterogeneous computing environments.
- Developed simulation frameworks to validate accelerator designs, enabling rapid iteration and performance optimization prior to physical deployment.

## Research Intern

Intel (2015-2016)

- 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

Indian Institute of Technology, Bombay (Summer 2014)

- Designed algorithms and framework for hyperspectral image processing.
- Leveraged matlab, and atmospheric data to build atmospheric corrections for hyperspectral images.

## Research Intern

Center for AI and Reobotics, DRDO (Summer 2013)

- Designed novel algorithm and mathematical modeling for geonric azimuthal map projection.

## Key Hardware/Software Co-Design Projects

- **Hardware-Aware Neural Network Co-Design:** Developed analog activation techniques for ReRAM crossbars, demonstrating energy efficiency improvements while maintaining accuracy through custom hardware-software optimizations.
- **Morphable Accelerator Architecture (Usás):** Designed a reconfigurable compute engine that dynamically adapts to available power constraints through systolic array reshaping and workload prioritization, saving hundreds of kWh/year compared to traditional approaches.
- **Computational Storage for ML Workloads (SaLT):** Engineered FPGA-accelerated storage architecture achieving 6× higher throughput for data-intensive ML operations through near-storage processing and data path optimization.
- **Dynamic Neural Network Training (NExUME):** Implemented precision/energy income/hardware-aware training techniques that maintain accuracy under intermittent power conditions, using adaptive quantization and dropout strategies that respond to energy availability.
- **Distributed Inference Framework (Seeker):** Developed a cross-platform inference system that efficiently partitions DNN execution across resource-constrained devices and more powerful hosts, reducing communication overhead by 8.9×.

## Selected Publications

- **Mishra CS**, Chaudhary D, Sampson J, Kandemir MT, Das CR. "Hardware-Aware Neural Network Co-Design with Analog Activation for Energy-Efficient ReRAM Crossbars." Under review, 2025.
- **Mishra CS**, Chaudhary D, Sampson J, Kandemir MT, Das CR. "NExUME: Adaptive Training and Inference for DNNs under Intermittent Power Environments." *International Conference on Learning Representations (ICLR)*, 2025.
- **Mishra CS**, Sampson J, Kandemir MT, Narayanan V, Das CR. "Usás: A Sustainable Continuous-Learning Framework for Edge Servers." *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024.
- Chaudhary D, Jain R, **Mishra CS**, Kandemir MT, Das CR. "MaestroRAG: Orchestrated Pipeline Architecture for Efficient RAG on Edge Devices." *Under review*, 2025.
- **Mishra CS**, Sampson J, Kandemir MT, Narayanan V. "Synergistic and Efficient Edge-Host Communication for Energy Harvesting Wireless Sensor Networks." *arXiv preprint*, 2024.
- Gunasekaran J, **Mishra CS**, Thinakaran P, Sharma B, Kandemir MT, Das CR. "Cocktail: A Multidimensional Optimization for Model Serving in Cloud." *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2022.
- **Mishra CS**, Sampson J, Kandemir MT, Narayanan V. "Origin: Enabling On-Device Intelligence for Human Activity Recognition Using Energy Harvesting Wireless Sensor Networks." *Design, Automation & Test in Europe Conference (DATE)*, 2021. [Best Paper Nominee]

- Qiu K, Jao N, Zhao M, **Mishra CS**, Gudukbay G, Jose S, Sampson J, Kandemir MT, Narayanan V. "ResiRCA: A resilient energy harvesting ReRAM-based accelerator for intelligent embedded processors." *IEEE Symposium on High Performance Computer Architecture (HPCA)*, 2020.

## Teaching Experience & Service

---

<b>Instructor</b>	Introduction to Computer Architecture (CMPEN 431)	Spring 2022, 2024
<b>Guest Lecturer</b>	Computer Architecture - Graduate Level (CSE 530)	Fall 2023, 2024
<b>Guest Lecturer</b>	Hardware Architecture for Deep Learning (CMPSC 497)	Fall 2019

**Professional Service:** Submission Chair for IEEE IISWC 2025; Reviewer for IEEE TPDS (2019, 2023, 2024, 2025) and IEEE TC (2021-2025); Student Member of ACM and IEEE

## Honors & Awards

---

- Best Paper Nomination, Design Automation and Test in Europe (DATE), 2021
- Student Travel Award, IEEE Network Architecture and Storage (NAS), 2021

## Complete List of Publications

---

### Under Review

- [R2025#1] **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#2] Deeksha Chaudhary, Rishabh Jain, **Cyan Subhra Mishra**, Mahmut Taylan Kandemir, Chita R Das. "**Mae-stroRAG: 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 in Public Cloud.**" arXiv preprint arXiv:2008.09491 (2020).

### Conference Papers



- [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. "**Uşás: 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 Gunasekharan, **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 Holographic Processing 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 Gunasekharan, **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).

## Talks

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

## 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 overprovisioning 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  $35\times$  speedup and major energy savings on edge.
- **AR/VR:** Explores spatio-temporal compute reuse in  $360^\circ$  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  $2.7\times$  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.