# Research Statement

Jyotikrishna Dass | dass.jyotikrishna@gmail.com

## Objective

The problem of training machine learning models in a distributed setting is becoming increasingly important as data collected on edge devices grows at an annual rate of 33%, making up 22% of the total global datasphere. This issue represents a central aspect of the "Artificial Intelligence of Things (IoT)" phenomenon in many applications, such as connected autonomous vehicles, digital healthcare, smart grids, and edge-caching wireless networks. What is common to many AIoT problems is not only the scale of the data but also the complexity of the processes that generated the data. Most importantly, much of the data is generated and collected on edge devices, characterized by heterogeneous compute and memory capabilities, dynamic network connections, and growing concerns about data privacy. To effectively use such data and derive insights, it is often insufficient to merely scale up existing methods typically designed for cloud-based training using centralized benchmark datasets. One must also understand and incorporate decentralized data, device capabilities, and network constraints toward creating optimized machine learning models and efficient systems for edge computing.

**Research Vision.** Building upon this insight that training machine learning models in an AIoT setting requires a rethinking of off-the-shelf machine learning solutions, my research agenda (Figure 1) is centered around ***developing distributed machine learning approaches that bridge centralized learning and federated learning***, taking into account the prominent streaming characteristics of data, along with the widespread heterogeneity of devices and networks. The primary objective is to harness the strong convergence guarantees of centralized learning while processing the original data on local edge devices, similar to federated learning, to train a global model collectively. My research aims to realize distributed edge intelligence to align with the key objectives of AIoT, i.e., safeguarding data privacy, minimizing latency, conserving bandwidth, enhancing energy efficiency, constructing robust models, and integrating streaming data.

**Research Methodology.** I aim to ***formulate advanced optimization techniques*** characterizing the unique AIoT requirements on a distributed network, ***develop parallel algorithms*** for training and updating machine learning models with a focus on low-cost of memory, improved latency and privacy, and ***co-design energy-efficient computing systems*** for accelerating computations towards practical AIoT applications. These tools will facilitate the local processing of decentralized data streams at source devices while enabling distributed computations with minimal communication overheads to improve the machine learning systems' performance, efficiency, and resiliency.

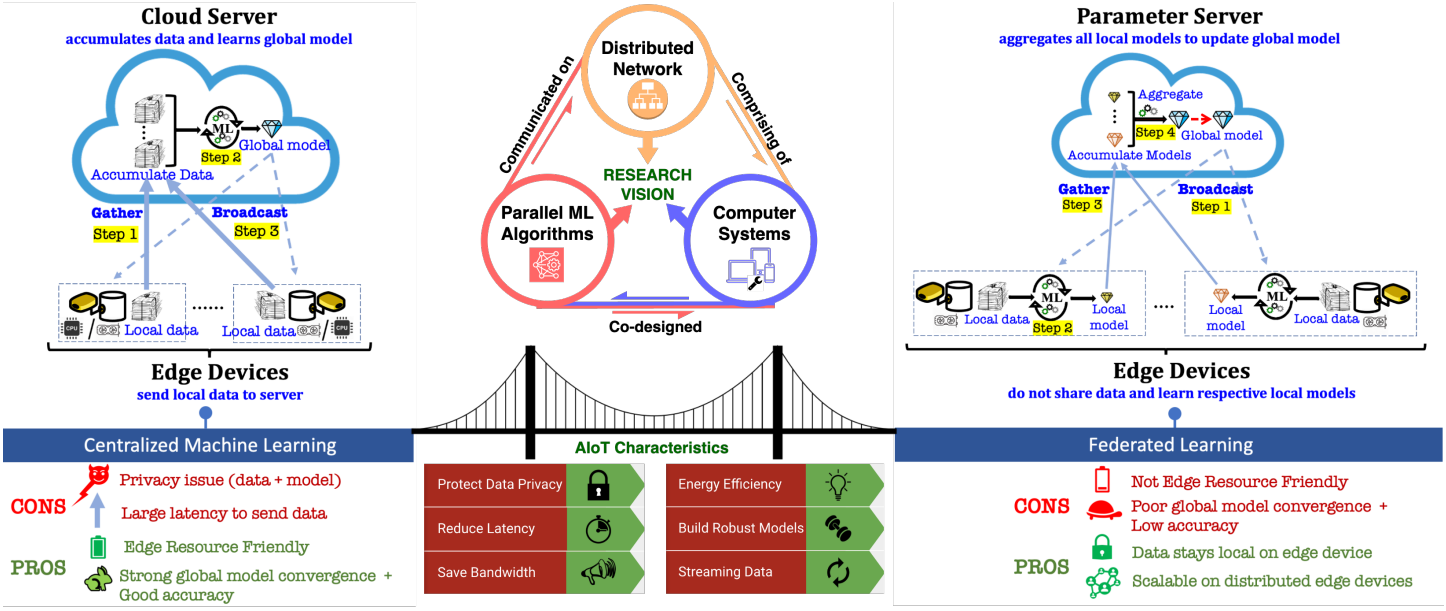



Figure 1: My research seeks to bridge centralized and federated learning through synergy among distributed networks, parallel ML algorithms, and computer systems for enabling distributed edge intelligence.

## Research Contributions

From the onset of my PhD journey, I was privileged to earn my advisor's support, which empowered me to explore my research interests encompassing machine learning and systems. This has involved independently formulating and spearheading various initiatives, fostering research collaborations for publications and grant writing, and adeptly juggling my teaching responsibilities in preparation for an academic career.

**Publications and Highlights.** My research has led to **10** publications (**7** as first-author) in some of the top-tier machine learning, parallel computing, and computer architecture conferences/journals such as ICML, ICDCS, IPDPS, TPDS, HPCA, Micro, TC, etc. I was nominated by Dept. of CSE (Texas A&M University) to compete at the Annual Computing@Southeastern Conference, where my Ph.D. dissertation received the **Best Poster Award** from a pool of top 40 Ph.D. candidates across 14 member universities. Moreover, during my postdoctoral tenure as a collaborator, I have substantially contributed to **3** successful research proposal grants from federal agencies, industry companies, and academic institutions. These include the NSF Core Programs, which received $1.2M in funding, the META Network for AI (selected amongst 6/38 proposals from 33 universities and institutions around the world) with a grant of $50K, and the Rice University Creative Ventures Fund which was awarded $20K. My involvement in these projects was comprehensive, spearheading the ideation process, team formation, proposal drafting, and budget planning stages.

My research pursuits have led to several projects that highlighted the richness of research questions and demonstrated holistic contributions spanning distributed networks, parallel ML algorithms, and computer systems.

(a) **Distributed Networks** - Towards formulating advanced optimization techniques for distributed networks, my work focused on minimizing synchronization time to optimize processor utilization and enhance latency. Additionally, it involved transforming dense, large-scale problems into memory-efficient (sparse) and separable problems conducive to parallel computations.

  (i) **Maximizing Processor Utilization.** In [7], I addressed *how to reduce idling due to synchronization across multiple nodes for solving parallel quadratic programming problems* (IPDPS'16). My work focused on reducing synchronization idling across multiple nodes by relaxing the synchronization frequency to maximize processor utilization. This involved developing a novel numerical algorithm, Lazily Synchronized Dual-Ascent (LSDA), which showed significant speedup and introduced a theory to determine the optimal synchronization period. LSDA is numerically stable and converges to the same result as the conventional tightly synchronized implementation.

  (ii) **Memory-Efficient Large-Scale Problem Solving.** In [4], I tackled the challenge, *how to efficiently formulate and solve a large-scale quadratic programming problem rather than solving the problem as a sequence of smaller sub-problems* (ICDCS'17). My work proposed a novel QR-decomposition framework (QRSVM) that leverages the low-rank structure of the kernel matrix to transform the dense matrix with quadratic memory complexity into one with a sparse and block-diagonal separable structure. This approach significantly reduced memory requirements and facilitated parallel computation. Furthermore, I derived an optimal step size for fast convergence of the dual ascent method for model training.

(b) **Parallel ML Algorithms** - Along this thrust, I developed a scalable distributed algorithm that minimizes communication costs for parallelizing model training. Moreover, my work has demonstrated the effectiveness of using memory-efficient data summaries and their parallel implementation as potential solutions for accelerating model learning.

  (i) **Scalable Distributed Algorithm.** In [5], I aimed at *designing scalable distributed algorithm to parallelize model training* (TPDS'18). I developed a communication-efficient implementation of the distributed machine learning framework, incorporating the following key improvements that reduced communication overheads during iterative model training: (i) skip communicating zero-rows in the memory-efficient representation of the local data, and (ii) communicate only a fixed small fraction of the dual variable during parallel dual ascent. The size of the truncated dual variable being communicated is independent of the number of workers and data samples. The resulting implementation demonstrated linear scalability across parallel workers, which is appealing for model training on distributed networks comprising many IoT devices.

(ii) **Speeding Up Model Learning.** In [2], I explored *how to speed up model learning under distributed settings* (ICML'21). My work leveraged data summaries to speed up model learning under distributed settings, advocating for classical Householder transformation as a strong candidate for sketching local data on each worker and accurately solving the family of Least Mean Squares problems by constructing a global data summary. I demonstrated it as a simpler, memory-efficient, and faster alternative that always existed compared to an award-winning baseline in [8]. My work highlighted the necessity of not hastily discarding traditional techniques, rethinking how certain comparisons are made, acknowledging widespread misconceptions, and reevaluating the most fitting algorithms for specific issues.

(c) **Computer Systems** - My contributions along this direction include co-designing software and hardware accelerators for energy-efficient distributed model training and improving the Vision Transformers' efficiency.

(i) **Energy-Efficient Model Training.** In [3], I focused on *optimizing the energy efficiency of distributed model training for edge devices* (TC 2020). My work implemented a first-of-its-kind system of multiple FPGAs as a distributed computing framework to fully parallelize, accelerate, and scale the distributed support vector machine training on decentralized data. Each FPGA unit has a pipelined model training IP logic core , with the distributed system faster and more energy-efficient than that of embedded edge processors and cloud processors.

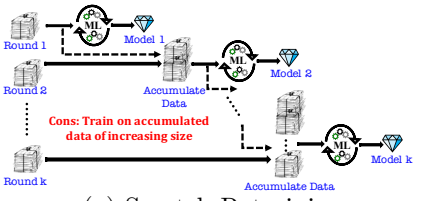
(ii) **Efficiency in Transformer Models.** My postdoctoral work [6] sought *how to overcome the quadratic cost in running Vision Transformers (ViTs) to improve its achievable efficiency* (HPCA'23). I proposed a novel algorithm-hardware co-designed framework, ViTALiTy, that decomposes Attention as a combination of low-rank and sparse components. At the algorithm level, I presented a first-order Taylor Attention as the low-rank component to linearize the cost of Attention blocks while demonstrating sparsity-based regularization for a few training epochs substantially boosts accuracy. In stark contrast to methods using dynamic sparsity computations for inference, the proposed framework implements a pipelined hardware-level accelerator that focuses solely on the fast execution of the relatively static linear Taylor attention component, thus eliminating computational overheads associated with dynamic sparsity and enhancing efficiency.

## Future Research Directions

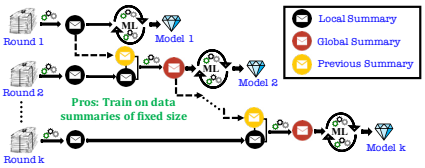
Significant strides have been made in the realms of Machine Learning (ML) and Systems to expedite computationally demanding ML training through parallelization. These techniques are typically designed for datacenter-scale deployment across high-performance computing servers for centralized learning, where training data is readily available or anticipated to be gathered from edge devices. Conversely, resource-limited edge devices are tasked with executing cost-effective sequential inference/prediction computations in real-time on unseen data. This has catalyzed advancements in the compression, pruning, and quantization of pre-trained ML models for edge deployment. At the hardware level, the resource heterogeneity and scarcity in typical edge devices have stimulated research into co-designing dedicated accelerators for energy-efficient inference computations. However, such workflows of training on the cloud and inference on the edge in a centralized learning setup are plagued by high network latency, privacy risks, and poor scalability in large connected networks of multiple heterogeneous edge devices. Federated learning (FL) is a recent ML trend that circumvents the need for data centralization on a cloud server by keeping data local at the source and pushing training computations to the edge. It enables a cluster of decentralized edge devices to collaboratively train a shared model. However, enabling efficient FL on the edge is challenging due to relatively poor model convergence than centralized learning [9] and the existence of lagging devices (i.e., stragglers or stale workers) owing to device heterogeneity, and network unreliability [1]. Hence, the FL setup might not be suitable for many online learning applications, such as news recommendations or interactive social networks that work on large amounts of data with a high temporality, i.e., data generally becomes obsolete very quickly. To unlock the immense potential of AIoT in transforming human life and significantly enabling distributed data analytics and smart systems for edge intelligence, I plan to undertake the following research endeavors to bridge centralized and federated learning setups.

**Incremental Learning with Streaming Data Summarization.** This objective seeks to address *how can we efficiently and accurately update the model using summaries of streaming data on a single device?*

In the face of the burgeoning volume of streaming data generated across a distributed network of heterogeneous devices, the necessity for efficient data summarization to perform incremental updates for the global model has become increasingly paramount. This is especially critical in mission-critical applications that demand low-latency and real-time inference. I am interested in exploring an innovative approach to incrementally train ML models with streaming data (depicted in Figure 2), eliminating the need to transfer the original data from edge devices to the cloud while circumventing the need to deploy local models on heterogeneous devices. Specifically, I envision a restructured decentralized incremental ML framework for edge devices, focusing on the construction and communication of local data summaries to efficiently update the global model across successive streaming rounds.




(a) Scratch Retraining

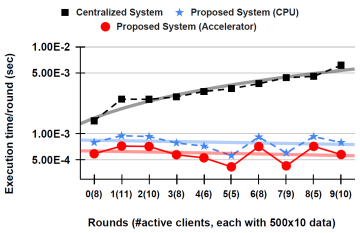



(b) Planned Incremental Learning

Figure 2: Streaming workflow.

**Co-Designing Energy-Efficient and Heterogeneity-Aware Decentralized Incremental Learning System.** My research seeks to address *how can we extend and scale the above solution to a decentralized incremental learning system comprising heterogeneous edge devices?*

The challenge lies in achieving incremental learning while ensuring real-time inference in resource-constrained edge devices. To address this, I am interested in co-designing solutions for an energy-efficient accelerator to efficiently summarize streaming data into a memory-efficient representation, thereby enabling the overlap of on-device inference computations and incremental learning. A preliminary version of the proposed accelerator simulated on an FPGA achieves a speedup of $\mathbf{2.3}\times - \mathbf{3.6}\times$ for data matrices with streaming batch sizes, $n = \{500, 5000\} \times 10$ compared to an Intel i7 processor. The planned incremental system with this preliminary accelerator as a potential edge client outperforms a simulated edge-based CPU network as observed in Fig. 3 for solving ridge regression model. To implement the aforementioned decentralized incremental learning framework for various machine and deep learning models, my research will develop a heterogeneity-aware system comprising multiple accelerators and edge-based commercial devices as potential clients, effectively aggregating and communicating data summaries under device and network heterogeneity. This endeavor represents a significant stride toward the future of decentralized learning, promising to revolutionize the way we handle and process data in an increasingly connected world.



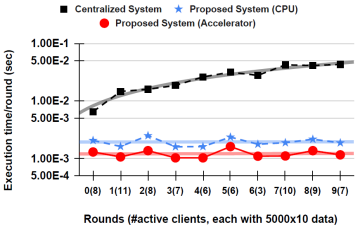




Figure 3: Scaling studies.

**Enabling Vision Transformers and Language Models for Edge Deployment.** The advent of Large Language Models (LLMs) and Vision Transformers (ViTs) has opened up new avenues for advancements in natural language processing and computer vision tasks towards generative AI applications. However, there is still much to uncover about the potential of these models, particularly in terms of their architectural innovations, context length improvements, efficiency, and more. Moreover, the application of these models in real-world scenarios, such as deployment on edge devices, presents its own set of challenges and opportunities.

My work on ViTALiTy [6] has laid the groundwork for enhancing the efficiency of ViTs for edge. I am keen on delving deeper into the innovations in linear Taylor Attention and exploring the potential of low-rank and sparse decompositions of softmax attention for efficient LLMs. The objective is to design a unified framework that caters to both Language and Vision Transformer models, as illustrated in Figure 4. Moreover, my work on NetDistiller [10] introduces a framework that boosts the task accuracy of Tiny Neural Networks (TNNs) for edge deployment, without incurring any inference overhead. Leveraging my experience, I am enthusiastic about parallelizing Transformer models and TNNs across a distributed network. My exploration aims to democratize the remarkable capabilities of such models, making them accessible for the edge.

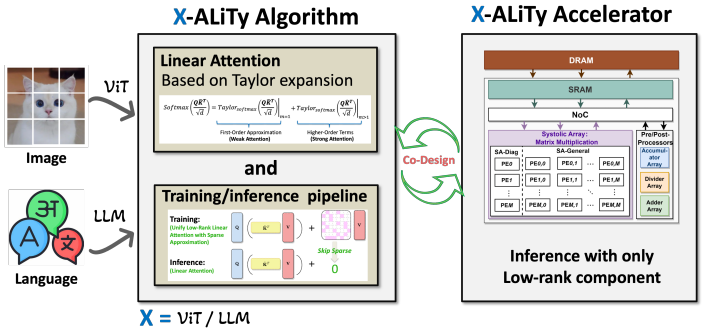



Figure 4: Unified Linear Taylor Attention Framework.

## References


[1] Sagar Dhakal, Saurav Prakash, Yair Yona, Shilpa Talwar, and Nageen Himayat. "Coded federated learning". In: *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE. 2019, pp. 1–6.

[2] **Jyotikrishna Dass** and Rabi Mahapatra. "Householder Sketch for Accurate and Accelerated Least-Mean-Squares Solvers". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 2467–2477. URL: https://proceedings.mlr.press/v139/dass21a.html.

[3] **Jyotikrishna Dass**, Yashwardhan Narawane, Rabi N. Mahapatra, and Vivek Sarin. "Distributed Training of Support Vector Machine on a Multiple-FPGA System". In: *IEEE Transactions on Computers* 69.7 (2020), pp. 1015–1026. DOI: 10.1109/TC.2020.2993552.

[4] **Jyotikrishna Dass**, V.N.S. Prithvi Sakuru, Vivek Sarin, and Rabi N. Mahapatra. "Distributed QR Decomposition Framework for Training Support Vector Machines". In: *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 2017, pp. 753–763. DOI: 10.1109/ICDCS.2017.222.

[5] **Jyotikrishna Dass**, Vivek Sarin, and Rabi N. Mahapatra. "Fast and Communication-Efficient Algorithm for Distributed Support Vector Machine Training". In: *IEEE Transactions on Parallel and Distributed Systems* 30.5 (2019), pp. 1065–1076. DOI: 10.1109/TPDS.2018.2879950.

[6] **Jyotikrishna Dass**, Shi Wu Shang, Li Huihong, Chaojian, Zhifan Ye, Zhongfeng Wang, and Yingyan Lin. "ViTALiTy: Unifying Low-rank and Sparse Approximation for Vision Transformer Acceleration with a Linear Taylor Attention". In: *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 2023, pp. 415–428. DOI: 10.1109/HPCA56546.2023.10071081.

[7] Kooktae Lee, Raktim Bhattacharya, **Jyotikrishna Dass**, VNS Prithvi Sakuru, and Rabi N Mahapatra. "A relaxed synchronization approach for solving parallel quadratic programming problems with guaranteed convergence". In: *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE. 2016, pp. 182–191.

[8] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. "Fast and accurate least-mean-squares solvers". In: *Advances in Neural Information Processing Systems*. 2019, pp. 8305–8316.

[9] Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. "Is local SGD better than minibatch SGD?" In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10334–10343.

[10] Shunyao Zhang, Yonggan Fu, Shang Wu, **Jyotikrishna Dass**, Haoran You, and Yingyan Lin. "NetDistiller: Empowering Tiny Deep Learning via In Situ Distillation". In: *IEEE Micro* 43.6 (2023), pp. 84–92. DOI: 10.1109/MM.2023.3324261.