Statement of Purpose
of Ao Shen (PhD applicant for Fall—2025)

I am determined to redefine the future of AI-driven systems. My journey began when I joined a supercomputing team for the ASC 23 (Asia Supercomputing) competition, optimizing CPU performance for molecular dynamics simulations. This early exposure to high-performance computing sparked my passion for system optimization. At Purdue University, I deepened my expertise through various research projects, but it was during my internship at the Shanghai Qizhi Institute (Founded by Turing Award winner Professor Andrew Yao) that my ambition reached new heights. There, I tackled cutting-edge challenges, from optimizing neural architecture search to reducing latency in large language model inference systems. These experiences solidified my fascination with the intersection of machine learning and system architecture and ignited my desire to push the boundaries of what's possible in computing.

While AI tools like ChatGPT are transforming industries, the cost of model training and inference remains a significant challenge, especially as hardware improvements slow in the post-Moore's Law era. My goal is to explore new architectures that bridge the gap between rapidly evolving algorithms and underutilized hardware potential. I am committed to pursuing a Ph.D. to develop efficient, AI-optimized systems that can accelerate the broader adoption of AI technologies.

As highlighted in my favorite article, *Pathways: A next-generation AI architecture* by Jeff Dean [1], the future of AI lies in building systems that can handle multiple tasks efficiently, learn new tasks quickly, and generalize across diverse domains. This vision deeply resonates with me and inspires my interest in three key areas of research. First, emerging AI workloads, such as large-scale machine learning models, demand specialized system optimizations to extract maximum performance and scalability. Second, the rise of edge computing and embodied intelligence calls for breaking traditional hardware-software boundaries to achieve real-time, efficient computations in resource-constrained environments. Finally, high-performance computing (HPC) for AI-driven scientific discovery excites me, as it holds the potential to unlock groundbreaking insights across disciplines like molecular dynamics and climate modeling. These three areas—AI-specific system optimizations, edge and embodied system design, and HPC for AI in science—are where I aim to focus my Ph.D. research, motivated by the aspiration to craft the systems that will power the next generation of intelligent technologies.

My broader vision is to bridge real-world industrial challenges with research-driven solutions, fostering a cycle of innovation between academia and industry. By focusing on both cutting-edge research and its practical applications, I aim to contribute to AI systems that are not only efficient but also impactful in real-world settings.

Systems for Emerging AI Workloads

The cost of training and inference for AI workloads has become a critical challenge, even for industry leaders like OpenAI, who grapple with balancing operational expenses. This issue is further exacerbated by the rapid evolution of AI models, which typically undergo significant changes every 5–6 years. During my internship at the Shanghai Qizhi Institute, I worked on optimizing inference systems for large language models (LLMs), which deepened my interest in building scalable and reliable systems to better support AI workloads and optimize hardware utilization. By refining preemption-based scheduling strategies, we achieved significant latency reductions, with key metrics like the p99 Time To First Token (TTFT) and Total Blocking Time (TBT) showing speedups of up to $11.2\times$. This project aligns with my broader interest in developing systems that efficiently handle the growing demands of AI workloads.

As the first author of the resulting paper, after conducting the majority of experiments and finalizing my thesis, I gained a foundational understanding of the experimental process and methodologies. I also recognized the importance of adhering to strict standards throughout the research. Additionally, I learned how to structure an academic paper to ensure logical flow, clarity, and precision in scientific writing.

Moreover, I am particularly interested in systems that leverage concepts such as **locality**. In transformer models, both static model parameters and dynamic KV cache allocations demand substantial resources. Techniques like quantization and distillation are vital for reducing computational load, making them essential for future system designs.

This brings up new questions: Which memories can be discarded, and which should be retrieved from secondary storage? Traditionally, this has been a systems challenge, but I see potential for AI to determine which information is truly necessary, beyond heuristic algorithms. In my future research, I aim to combine insights from AI architecture and systems theory to build scalable, efficient infrastructures that can meet the demands of emerging AI workloads.

Edge Computing and Embodied Intelligence

My research on edge-side technologies, including Neural Architecture Search (NAS), aligns with my broader interest in edge computing and embodied intelligence. My earliest project at the Shanghai Qizhi Institute is about NAS, though ultimately unsuccessful, was instrumental in shaping my research approach. The project aimed to expand the search space for operator design in edge-side applications, but the large search space generated too many candidates, making training and testing inefficient. My limited experience with computing and data management at the time further hampered progress, and I often ended up presenting incomplete data. From this experience, I learned the vital importance of clearly defining core assumptions, staying focused on the main objectives, and maintaining transparency throughout the research process — lessons that have greatly benefited my subsequent work. At the Institute, edge devices, including embodied AI, are a central focus of research. I've come to realize that future robots will likely operate as clusters of agents, with frameworks and systems like SWARM robotics playing a pivotal role, which will inevitably raise many system-level challenges. I am eager to further explore the field of robotics, leveraging my connections at the Institute to pursue new collaborations, particularly in research related to edge devices and embodied AI.

High-Performance Computing (HPC) for AI in Science

When I first entered university, I had the fortunate opportunity to join the school's supercomputing team. After rigorous efforts and passing several rounds of selection, I successfully won the First Prize as a group member in the ASC 23, where I optimized CPU performance for molecular dynamics simulations. Through techniques like OpenMP, vectorization, and multi-GPU communication using MPI, we achieved a 25% performance improvement, with further optimizations resulting in a total 50% increase. The computational optimization skills I developed through this competition have directly informed my current interest in HPC for AI4Science.

I see high-performance systems as critical tools to accelerate breakthroughs in fields such as climate modeling and molecular biology, where AI-driven scientific discovery holds immense potential. While AI has the power to revolutionize these areas, its computational demands often lead to performance bottlenecks. Throughout this process, I became fascinated by the workflow of profiling various metrics, identifying bottlenecks, proposing solutions, and ultimately achieving optimizations, thereby empowering other natural sciences. My work optimizing simulations highlighted the importance of addressing these challenges. By enhancing AI performance through HPC, we can unlock faster, more scalable solutions for scientific research.

References

[1] Jeff Dean. Pathways: A next-generation ai architecture. https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/, 2021. Accessed: 2024-11-08.