Theme 1: Hardware Software Techniques in Interconnect Design for Deep Learning Applications

Proposals are invited for research focused on scalable accelerator interconnects for deep learning applications. The rapid expansion of deep learning models has shifted the primary performance bottleneck to the interconnect fabric, necessitating fundamental research to address end-to-end latency, energy consumption, and network reliability at scale.

The research scope may include, but is not limited to, the following areas:

- **Interconnect Architecture and Topology:** Investigating novel physical topologies and the architectural innovations needed to bridge the gap between low-latency high bandwidth scale-up and highly scalable scale-out networks.
- **Network-on-Chip (NoC) Design:** Advancing on-chip communication for heterogeneous, multi-die accelerator systems to improve efficiency and reduce latency.
- **Collective Communication Algorithms:** Developing dynamic, topology-aware collective communication algorithms and the software support required to optimize data exchange in complex, asymmetric clusters.
- **Hardware and Software Fault Tolerance:** Designing and implementing intelligent fault tolerance mechanisms, including proactive, software-supported techniques, to ensure serving or training continuity in the event of hardware or network failures.

Theme 2: Runtime Memory Management for Distributed-Memory LLM Accelerators

Proposals are invited for research focused on scalable runtime memory management for large-scale LLMs serving on next-generation accelerators with distributed-memory architectures. Current software ecosystems, optimized for GPUs with a global shared memory model, are fundamentally mismatched with novel systems from vendors like Cerebras and Graphcore that employ a physically distributed memory model using local SRAM and on-chip mesh fabrics. This architectural divergence creates significant bottlenecks and renders existing serving optimizations like Paged Attention and Chunked Prefilling inefficient or inoperable. This problem is further exacerbated for sparse Mixture-of-Experts (MoE) models, where the distributed memory amplifies the All-to-All communication overhead, causing significant compute and memory bandwidth underutilization and idle time.

The proposed research should focus on fundamental runtime problems, including the efficient implementation of Paged Attention and Chunked Prefilling, by adapting these techniques for explicit data movement across distributed memory. Beyond improving traditional metrics like latency and throughput, research outcomes must also address critical system-level factors, including Host-to-Device interconnect performance, on-chip memory bandwidth utilization, and energy consumption.

Project Proposal Requirements

To ensure that proposals are thoroughly prepared and align with the goals of the research project collaboration program, the following requirements must be met. Proposals lacking detailed descriptions of the below items will not be considered for review.

- The proposal should be limited to 6 pages, excluding appendices.
- The proposal must demonstrate a comprehensive understanding of the challenges and propose an innovative approach that provides tangible, measurable improvements. This should include:
    - Quantifiable outcomes using Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs).
    - Expected deliverables, such as a novel architecture or algorithm, a software library extension, a detailed performance analysis report, and an academic publication.
- Clearly define the merit of the proposed innovation compared to competing approaches, including:
    - A detailed analysis of the current state-of-the-art and how the proposed innovation advances beyond it.
    - Quantitative projections for performance improvement, tied to representative values from authoritative publications, such as industry benchmarks or academic research.
    - An assessment of the prospective product's potential impact, supported by relevant justifications and a strategic plan for possible commercialization, including potential applications in industries that rely heavily on AI/ML workloads, target markets.

Budget and Funding

- The budget must be comprehensive and include overhead costs, with a clear breakdown of the calculation methodology.
- Please provide a reasonable budget amount considering the number of student researchers who will be fully dedicated to this project.

Compliance with Export Control Regulations

The Project Investigator must ensure compliance with all applicable export control regulations, including updated announcements from the Bureau of Industry and Security (BIS). This includes:

- Ensuring all personnel involved in the project are aware of and comply with relevant export control regulations.
- Implementing necessary measures to prevent unauthorized disclosure or transfer of controlled technology or items.
- Obtaining any required export licenses or authorizations prior to the transfer of controlled technology or items.

Failure to comply with export control regulations can result in severe consequences, including fines, penalties, and reputational damage. If you are unsure about any aspect of export control regulations, please seek guidance from relevant authorities or experts.

Samsung AGI Computing Lab University Program Overview
This program is open to world-leading universities and designed to create opportunities to explore breakthroughs and innovative research.

a. Timeline

Submission open	September 22, 2025
Application deadline	November 12, 2025, at 6pm PST
Interview	December 4, 5 2025
Final announcement of awardees (via email)	January, 2026

b. Eligibility for Funds
To be eligible for funds under the Samsung AGI Computing Lab's University Program, an applicant's university must accept the Research Agreement (RA) as part of the proposal-submission process. Key provisions of the RA specify project conditions including funding for the project, IP rights, and clarify other aspects of the research collaboration.

c. Evaluation Criteria
Samsung evaluates proposals in the following (but not limited to) criteria:

1) Innovativeness of research
2) Potential business and/or scientific/social impact
3) Feasibility of research with respect to planned time, objectives, intended results, and resources (subjected to availability)

Samsung will have sole discretion in the University Program Award Selection. No feedback will be provided to the applicant.

d. Confidential and Proprietary Information
Samsung does not wish to receive confidential or propriety information in the submitted proposals.

Samsung does not require, and does not desire to receive any information that may be deemed confidential by the University and its partners. Samsung will treat all information submitted in proposals as non-confidential and non-propriety.