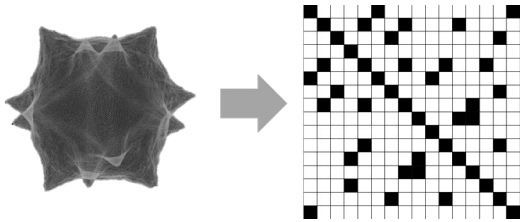


Sparsity, the Key Ingredient from HPC to Efficient LLMs
Co-Located with MICRO 2026, Athens, Greece

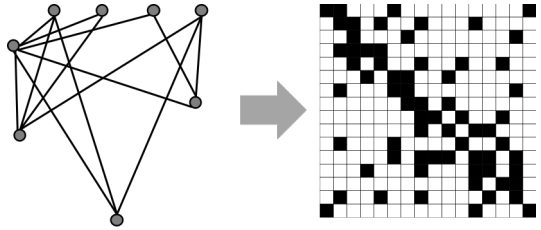
VOLANTIS SAMSUNG UNIVERSITY OF MARYLAND 18 56

Sparsity is Everywhere

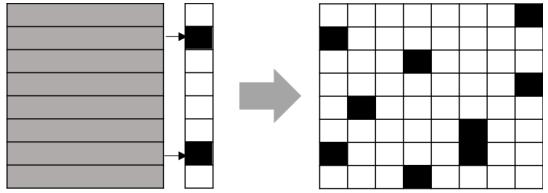
Scientific Computing



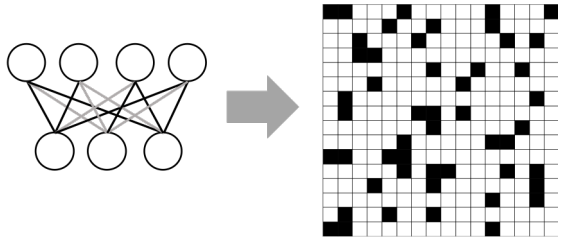
Graph Analytics



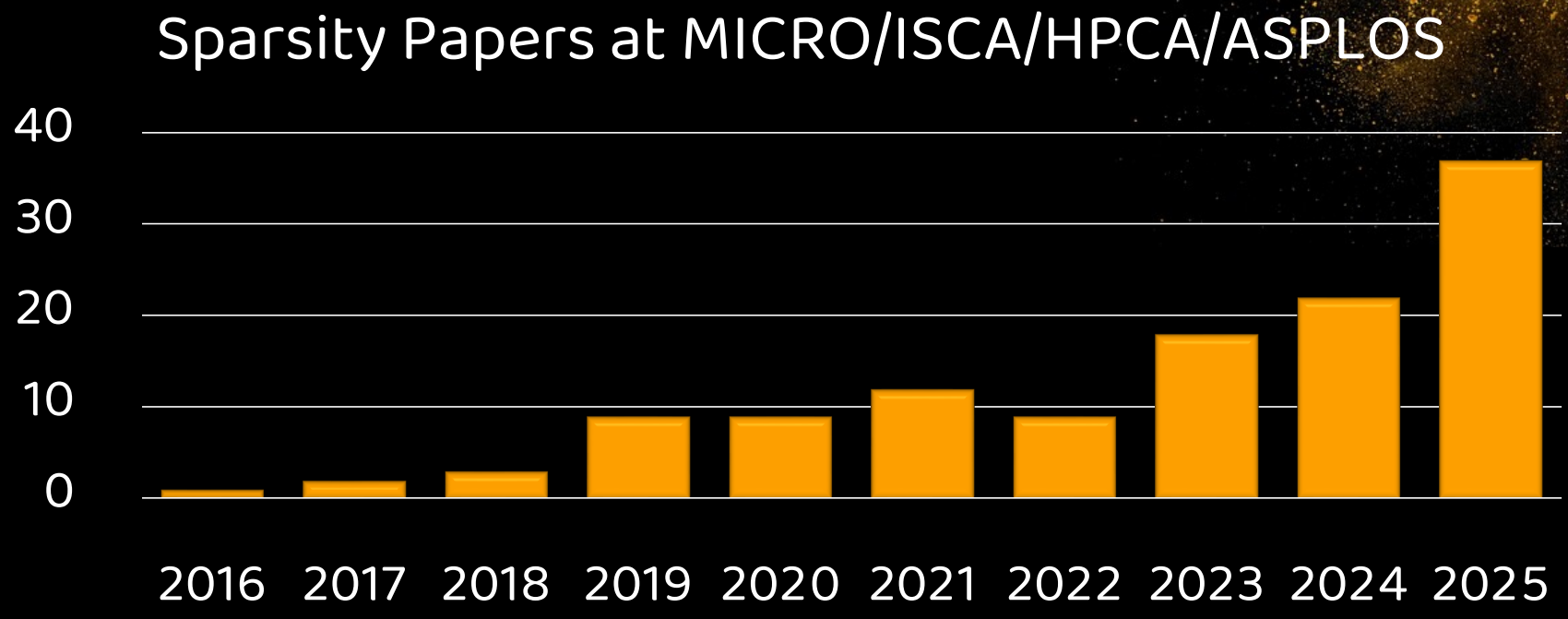
DLRMs



DNNs and LLMs



The number of papers on Sparsity is growing!



Keynote Talk



Speeding-Up Communication in Distributed Sparse Applications

Prof. Josep Torrellas
The University of Illinois at Urbana-Champaign (UIUC)

Invited Speakers



Architecture Research for DL Sparsity in the Large Model Era: Challenges and Opportunities

Po-An Tsai
NVIDIA Research



The Sparse LLM Inference through Algorithm–Compiler Co-Design

Maryam Mehri Dehnavi
University of Toronto and NVIDIA

Call For Papers

- **NEW** Sparsity-aware networking, interconnects, and rack-scale deployment for large-model training and inference, including scale-up and scale-out fabrics.
- **NEW** Sparse retrieval, memory access, and context selection for long-context and agentic foundation models.
- **NEW** Runtime and scheduling frameworks for selective execution and sparse activation in multi-stage and multi-agent inference pipelines.
- Sparse inference and training techniques in LLMs and foundation models.
- MoE and dynamic routing models: performance, systems, and deployment.
- Quantization, pruning, and compression methods that induce or leverage sparsity.
- Sparsity in scientific and HPC applications.
- Architectural support for unstructured and structured sparsity.
- Compiler, runtime, and scheduling frameworks for sparse workloads.
- Benchmarks, metrics, and tools to evaluate sparse systems.
- Hardware/software co-design for sparsity-aware execution.
- Sparse acceleration in near-memory, neuromorphic, analog, edge, and low-power AI systems.
- Programming models and abstractions for sparse computing.
- Case studies of sparse systems deployed on a scale.

Papers

- All papers received 3+ reviews
- All double-blind reviews
- All papers are available on the website: spice.cs.umd.edu

- Submission Deadline was on September 11, 2026
- Author Notification sent on September 26, 2026 (before MICRO's early registration deadline)
- Submissions were on hotcrp

Schedule: Session 1

Time (EET)	Title	Speaker
1:00PM – 1:10PM	Opening Remarks	Bahar Asgari
1:10PM – 2:10PM	**Keynote Talk:** Speeding-Up Communication in Distributed Sparse Applications	Prof. Josep Torrellas
Session 1: Sparse Kernels | Session Chair: Ramyad Hadidi		
2:10PM – 2:40PM	**Invited Talk:** Architecture Research for DL Sparsity in the Large Model Era: Challenges and Opportunities	Dr. Po-An Tsai
2:40PM – 2:50PM	Profiler-Gated Fusion of V:N:M Sparse Tensor-Core Kernels for By-Construction Non-Regressive CNN Speedups	Peter Kogge
2:50PM – 3:00PM	SpeedSparse: Exploiting Dense General Matrix Multiplication for Practical Sparse Neural Networks	Shreya Alladi
3:00PM – 3:30PM	Coffee Break	

Schedule: Session 2

Session 2: Compression, Dataflow & Accelerators
Session Chair: Farzaneh Zokaee

3:30PM – 4:00PM	**Invited Talk:** Sparse LLM Inference through Algorithm–Compiler Co-Design	Prof. Maryam Mehri Dehnavi
4:00PM – 4:10PM	Connectivity, Not Coordinates: Accelerating SpMM and High-Order Tensor Contractions via Graph Abstractions	Amir Ahsaei
4:10PM – 4:20PM	Fractional-Exponent Floating-Point Numbers	Tianyun Zhang
4:20PM – 4:30PM	Cross-Layer Execution under Sparsity: Joint Exploration of Fusion Depth and Spatial Granularity	Hajar Falahati
4:30PM – 4:40PM	SQuaRE: Sparsity-Quantization Resource-Efficient Optimization	Nezam Rohbani
4:40PM – 4:50PM	LATTIX: A Sparse Processing-In-Memory Accelerator for Speculative LLM Inference	Seungkwan Kang
4:50PM – 5:00PM	Closing Remarks	

Logistics

☕ Coffee break is from 3:00pm to 3:30pm

🎁 Please feel free to grab a SPICE sticker!

WIFI Info:

MICRO 2026 Athens
Micro2026

📸 If you are a speaker, author or co-author of a paper, an organizer, or a program committee member of SPICE, **please come to the stage at the beginning of the break so that we take a group photo!**

Co-Chairs



Bahar Asgari
UMD



Ramyad Hadidi
Volantis Semiconductor

Organizer Committee



Farzane Zokaee
Samsung SRA



Olivia Hsu
Stanford/CMU



Antonino Tumeo
PNNL



Prashant Nair
UBC/d-Matrix



Ben Feinberg
Sandia NL

Program Committee & Student Volunteers

Program Committee:

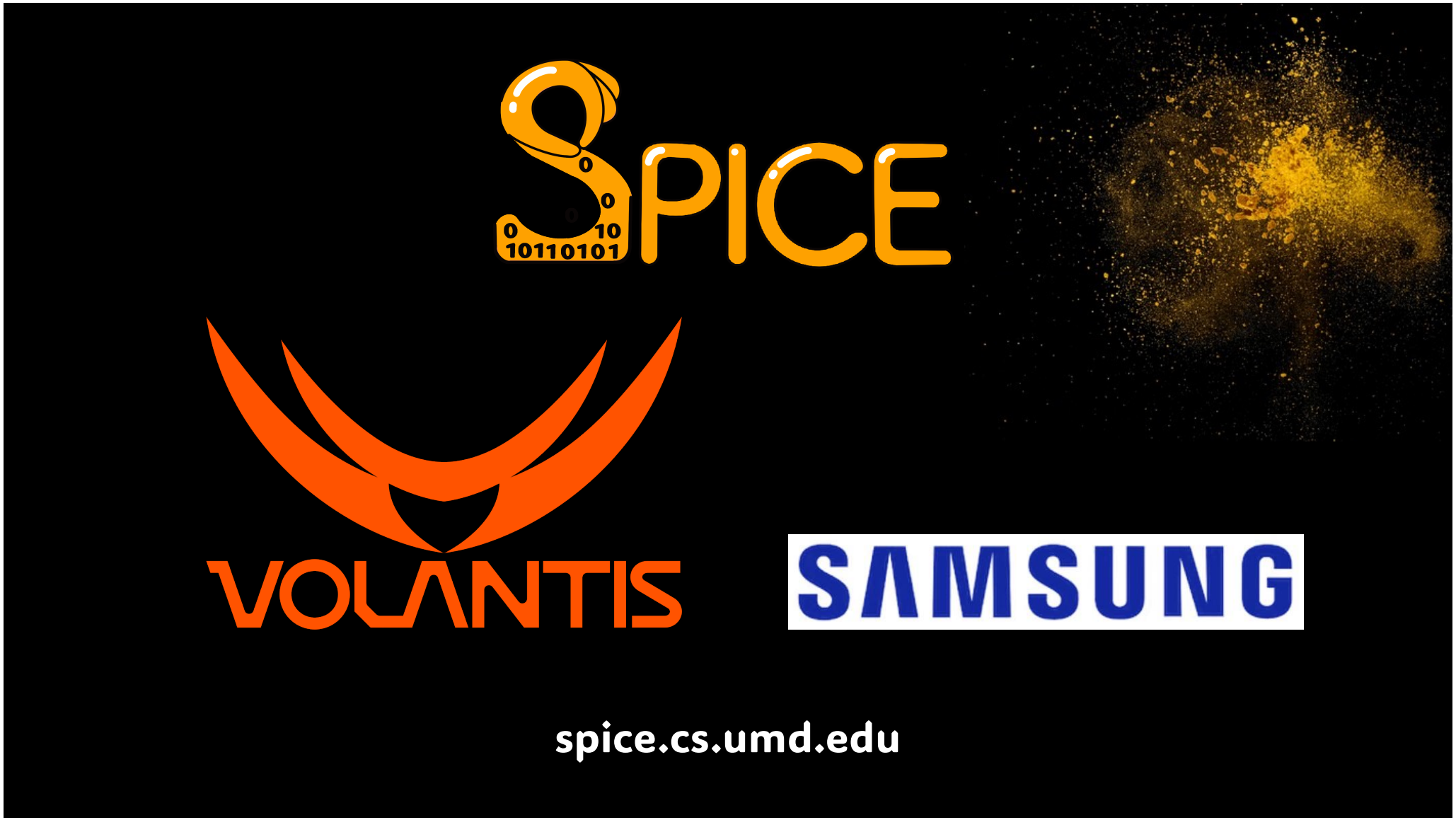
- **Anirudh Jain** | Tenstorrent
- **Anshuman Verma** | Microsoft
- **Christina Giannoula** | Max Planck Institute for Software Systems
- **Gerasimos Gerogiannis** | Google
- **Zishen Wan** | Columbia University

Submission Chair:

- **Johnson Umeike |** PhD Student, University of Maryland, College Park (UMD)

Publicity Chair:

- **Helia Hosseini |** PhD Student, University of Maryland, College Park (UMD)


SPICE
VOLANTIS
SAMSUNG
spice.cs.umd.edu

Keynote Talk



Speeding-Up Communication in Distributed Sparse Applications

Prof. Josep Torrellas
The University of Illinois at Urbana-Champaign (UIUC)