Midwest RandNLA Conference 2026

Breakout 2: Hardware-Aware RandNLA

Participants: Daniel Bielich, Shambavi Suryanarayanan, Rob Webber, Chris Geoga, Longfei Gao, D. Adrian Maldonado, Tunan Wang

Facilitator: Sachin Garg

Current hardware trend

Modern accelerator architectures and HPC systems are increasingly constrained by memory movement and communication rather than actual arithmetic.

- Data movement across memory hierarchies frequently dominates runtime.
- Large multi-GPU systems are often limited by synchronization and interconnect costs.
- Modern accelerators increasingly prioritize low-precision tensor-core computation (e.g., FP8/FP4).

RandNLA for modern hardware

Sketching compresses matrices using one or a few passes over the data.

- Restricting computation to sketched data leads to reduction in data movement, for example in Sketch-and-solve and Block-Lanczos.
- Sketch-and-precondition approaches can improve conditioning and reduce iterative solver costs.
- Adoption of RandNLA can be challenging since randomized methods introduce variability, while many scientific applications and users require reproducible behavior.

Hardware aware algorithm design

There is a need for randomized algorithms built around GEMMs and other GPU-friendly operations.

- Structured sketches such as randomized Hadamard transforms may be less well-suited for GPUs than dense GEMM-based projections.
- GPU-oriented computational models may enable improved parallel complexity for iterative randomized methods such as linear system solvers.
- Randomized algorithms for large multi-GPU systems should be robust to communication delays.

Hardware aware algorithm design

Recent accelerator architectures increasingly emphasize low-precision tensor-core computation (e.g., FP8/FP4).

- There is a growing need for randomized algorithms that are robust to low-precision arithmetic.
- Iterative refinement provides a promising paradigm where a low-precision approximate solution is progressively refined using higher-precision correction steps.

Hardware aware algorithm design

- Continued support for high-precision accelerator hardware remains important for scientific computing and numerically sensitive RandNLA workloads.
- Unified memory and near-memory computing architectures could support RandNLA workloads on datasets exceeding individual accelerator memory capacity.