GPU-Accelerated Methods for Multi-GPU and Multi-Node

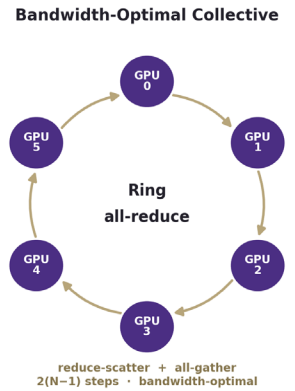
Scaling Machine Learning

Winter 2027 · UW ECE Certificate in GPU-Accelerated Computing & Visualization

- **Collective communication —** all-reduce, reduce-scatter, all-gather & all-to-all; bandwidth-optimal ring and latency-optimal tree algorithms; NCCL and in-network SHARP.
- **ML parallelism strategies —** data & sharded-data parallelism (ZeRO / FSDP2), plus tensor, pipeline, sequence, context, and expert parallelism.
- **Hybrid 5D parallelism & GPU-initiated comm —** composing DP × TP × PP × CP × EP; NVSHMEM / IBGDA, CUDA-aware MPI, and the torch.distributed stack.
- **Multi-GPU & multi-node hardware —** NVLink / NVSwitch, InfiniBand (NDR / XDR) & RoCE, GPUDirect RDMA;
- **ML Case studies —** Megatron-Core 5D training, DeepEP MoE all-to-all dispatch, and vLLM distributed inference.

Toolchain: PyTorch (FSDP2 / torch.distributed) · NCCL & NVSHMEM · CUDA-aware MPI · multi-node GPU clusters

PyTorch + CUDA C++ · distributed systems-intensive



Ring all-reduce: the collective that maps onto the fast intra-node fabric

Intra-node NVLink outpaces inter-node InfiniBand by ~20-40×