

Applied High-Performance GPU Computing

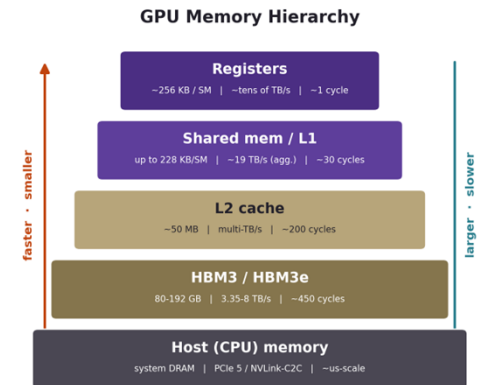
Hardware & Software Foundations

Autumn 2026 · UW ECE Certificate in GPU-Accelerated Computing & Visualization

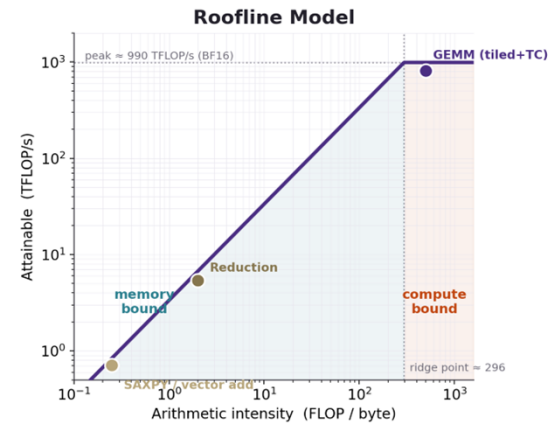
- **GPU architecture & the CUDA model** — streaming multiprocessors, warps & the SIMT model, thread-block clusters; the Volta → Hopper → Blackwell arc.
- **GPU Memory hierarchy & performance analysis** — coalesced access, shared-memory banking, the roofline model, and arithmetic intensity as the central design variable.
- **The canonical GPU patterns** — tiled GEMM, Tensor Cores (WGMMMA / TMA), reductions & online softmax, parallel scan, sort, and convolution.
- **Asynchronous execution** — CUDA streams & events, cooperative groups, cp.async pipelining, and an introduction to CUDA Graphs.
- **ML case studies** — FlashAttention v1/v2, the Mamba selective-scan, and 3D Gaussian Splatting rasterization.

Toolchain: CUDA C++ · Nsight Compute / Systems · Hopper & Blackwell GPUs (cloud-hosted, AWS)

CUDA C / C++ coding intensive



Bandwidth & latency across the on-chip memory hierarchy



Roofline: memory- vs. compute-bound kernels on an H100-class GPU

