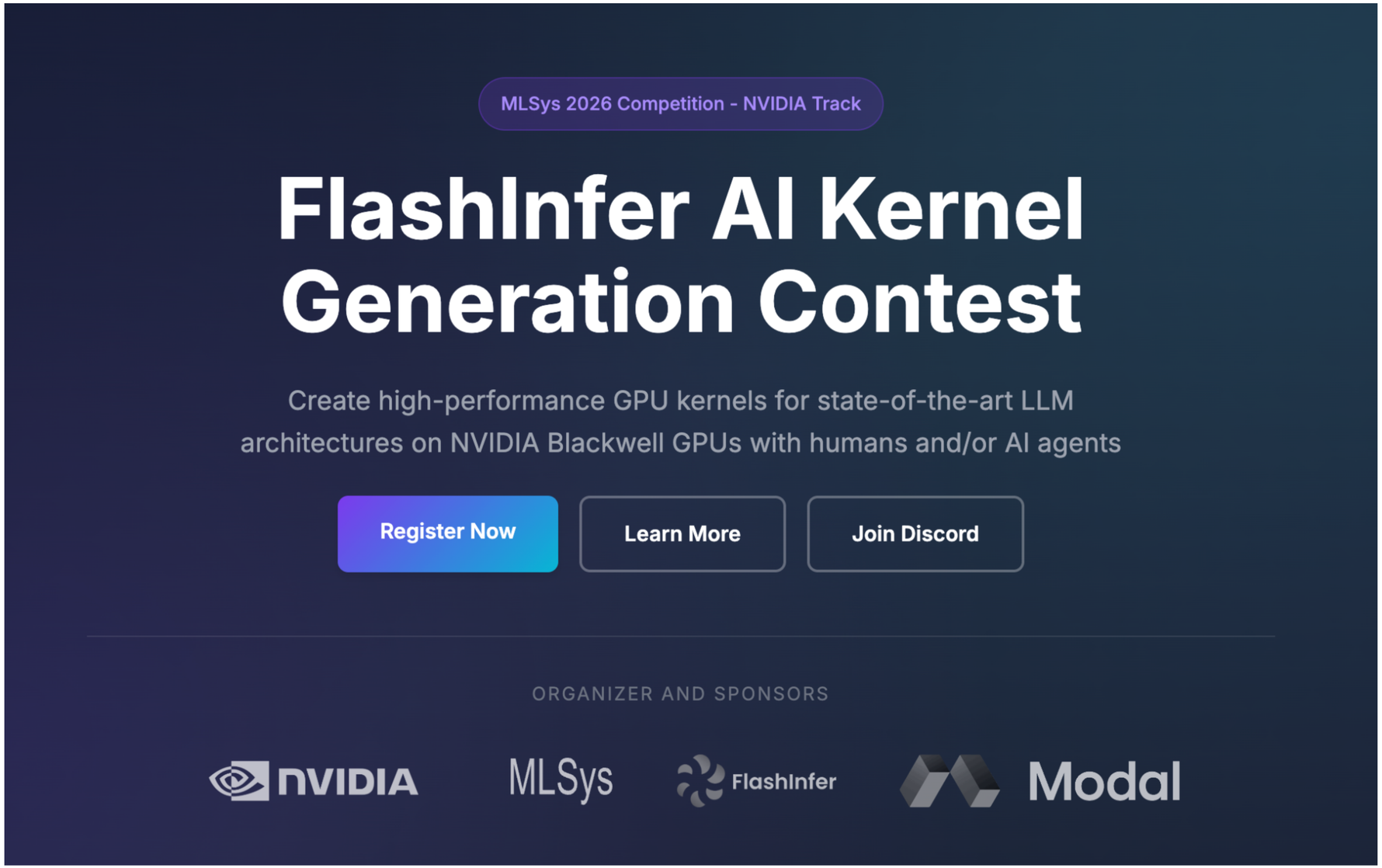
NVIDIA at MLSys 2026

Papers with NVIDIA authors

- **FlashInfer-Bench: Building the Virtuous Cycle for AI-driven LLM Systems**
- **Event Tensor: A Unified Abstraction for Compiling Dynamic Megakernel**
- **BLASST: Dynamic BLocked Attention Sparsity via Softmax Thresholding**
- **TiDAR: Think in Diffusion, Talk in Autoregression**
- **FlashAttention-4: Algorithm and Kernel Pipelining Co-Design for Asymmetric Hardware Scaling**



Checkout the competition session on Friday