MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability

Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans

Problem: GPU Performance Scaling is Under Threat!

Solution: MCM-GPU
**Key Challenge:**
• On-package bandwidth is critical

**Revisit:**
1. Cache architecture
2. Thread-block scheduling
3. Data placement

**Results:**
• 5x on-package BW reduction
• 28.8% speedup over basic MCM-GPU
• Performance within 8% of ideal monolithic GPU

4:00 pm - Session 4B: GPUs