CS 758: Advanced Topics in Computer Architecture

Lecture #10: GPU Chiplets & Multi-Chip Modules (MCMs)

Professor Matthew D. Sinclair

Announcements

- Mid-Semester Evaluations
  - Who did receive the info?
  - Working with department on why many of you didn’t get link
  - Will post an announcement shortly
- Project Proposals due today
  - Will provide feedback soon
- Midterms released
  - Should be able to see on Gradescope with feedback
  - 2 weeks for regrades (until 11/4)

- (See Arunkumar slides)

Conclusion

- Moore’s Law petering out (even for GPUs – Dark Silicon paper)
  - Can’t scale a single GPU much/any further
  - Solution: multi core-like GPUs!
- Rely on data locality to avoid inter-GPM communication
  - Add third level of cache (L1.5) to store remote data
  - Redesign scheduler to avoid/exploit temporal locality across TBs
  - Redesign page placement to avoid/exploit temporal locality per GPM
- Lots of opportunities for interesting research!
  - Open questions: workload partitioning, imbalance, scheduling, coherence, consistency, synchronization