EE 7722
GPU Microarchitecture

Where/When
1126 Patrick F. Taylor Hall 11:30–12:20 **Spring 2027**
`https://www.ece.lsu.edu/gp/`

Who
David M. Koppelman, Room 3316R P.F. Taylor Hall
(225) 578-5482, `koppel@ece.lsu.edu`
Office Hours: Monday–Friday: 14:00–15:00.

```
.target sm_120
LDG.E R165, desc[UR14][R208.64+0x10250] ;
LDG.E R2, desc[UR16][R208.64+0x18340] ;
LDG.E R3, desc[UR18][R208.64+0x18350] ;
// mm.cu:1054  for (ssize_t i_k=0; i_k<A_ncols; i_k+=tk )
IADD.64 R202, R202, 0x20 ;
ISETP.GT.U64.AND P0, PT, R202, 0x7ff, PT ;
IADD.64 R206, R206, 0x40 ;
IADD.64 R204, R204, 0x40 ;
HMMA.16816.F32 R24, R64, R210, R24 ;
HMMA.16816.F32 R28, R64, R212, R28 ;
HMMA.16816.F32 R108, R64, R198, R108 ;
HMMA.16816.F32 R4, R60, R210.reuse, R4 ;
HMMA.16816.F32 R16, R56, R210.reuse, R16 ;
HMMA.16816.F32 R136, R52, R210, R136 ;
```

GPU Assembler for Tensor Core Demo

Prerequisites
By Topic: Computer architecture, digital logic, programming in a C-like language, exposure to assembly language programming.

Topics
Overview—

Accelerators (*e.g.*, NVIDIA GPUs, Google Trillium TPUs, Meta MTIa, Tesla DOJO) and their role in typical systems.

Characteristics of workloads, including transformer attention. Survey of GPU APIs and systems, such as CUDA, OpenCL, OpenMP, OpenACC. A brief history.

Parallelism Related Concepts—Threads, Speedup. Latency, Throughput. Performance Limiters.

CPU Execution Features—Dynamic scheduling, caches, branch prediction.

Execution Architecture—Thread organizations. Hardware organization, functional units, systolic arrays. Scheduling, instruction execution, latency hiding.

Storage Hierarchy, Synchronization—Address spaces. Scratchpad stores and cache hierarchy. Basic data access strategies. Synchronization, atomic operations, message-driven execution. Basic reduction strategies.

Instruction Sets—Memory access, address spaces, locking, reduction. Predication, reconvergence.

Basic Algorithms and Techniques—Matrix multiplication, tensor operations. Machine-learning, scientific, and graphic workloads. Reduction techniques. Sorting.

Tensor Processing—Machine Learning Dataflows. Specialized Instruction Sets. Large-Systolic-Array Accelerators (*e.g.*, Google Trillium TPU). Wafer-Scale Message-Driven MAC Grid (*e.g.*, Cerebras CS-3). Exploiting weight regularity and sparsity.

Research Directions—Workload studies. Acceleration of transformer attention and other ultra-high-compute workloads. Processing near/in/with Memory Many threads v. prefetch. Low-energy execution.

Text
Papers and other references.

Grading
35% Midterm Exam • 35% Final Exam • 30% Homework and Projects
Midterm exam grade may be replaced with final exam grade if it is higher. Late assignment penalty: 10% per day late deducted. Missed-midterm-exam policy: at instructor's discretion either a makeup exam, use final exam grade for midterm grade (*i.e.*, 70% final exam weight), or use of zero for midterm grade. Daily attendance: optional, however students are responsible for all material, instructions, and notices presented in class.

✣ A.I.

Students may use AI, including LLMs such as Claude and ChatGPT, to assist in learning how to solve homework assignments. This includes asking about course material, clarifying the questions in an assignment, developing approaches to solve the problems, and helping with compiler or execution errors in solution code. After using AI and traditional resources (references, knowledgeable friends, etc.) to learn how to solve an assignment students are expected to then complete the assignment (maybe after several tries) without these resources. **Test questions will be written for students who learned how to solve the homework assignments unassisted.**

A.D.A.

Louisiana State University is committed to providing reasonable accommodations for all persons with disabilities. This syllabus is available in alternate formats upon request. Any student with a documented disability needing academic adjustments is requested to speak with Disability Services and the instructor as early in the semester as possible. All discussions will remain confidential. Please contact Disability Services in 115 Johnston Hall, 225-578-5919 or at `http://www.lsu.edu/disability`.