EE 7722
GPU Microarchitecture

Where/When
116 Tureaud Hall MWF 11:30–12:20 **Spring 2024**
`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.

```
@"EF_CUDA_SM89 EF_CUDA_VIRTUAL_SM(EF_CUDA_SM89)"
// tc.cu:228   const int tid = blockIdx.x * blockDim.x + t
S2R R17, SR_CTAID.X ;
IMAD.MOV.U32 R3, RZ, RZ, c[0x3][0xd0] ;
S2R R2, SR_TID.X ;
IMAD R17, R17, c[0x0][0x0], R2 ;
// tcu.cu:274 load_matrix_sync( tile_word, hp + i_word_0 +
// tcu.cu:278 mma_sync( tile_hq[i_wd], tile_qkv, tile_word
MOVM.16.MT88 R20, R20 ;
MOVM.16.MT88 R21, R21 ;
HMMA.16816.F32.BF16 R16, R12, R22, R16 ;
LEA R22, P1, R28, R34, 0x2 ;
LEA.HI.X R23, R28, R35, R29, 0x2, P1 ;
HMMA.16816.F32.BF16 R8, R12, R20, R8 ;
LD.E R12, [R24.64+0x138000] ;
```

GPU Assembler for Transformer Layer Demo

Prerequisites
By Topic: Computer architecture and digital logic.
Also, students must also have a familiarity machine language and C++ programming.

Topics
Overview—Accelerators (*e.g.*, NVIDIA GPU, Google TPU, Tesla DOJO) and their role in typical systems. Characteristics of workloads. 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. Scientific, graphic, and machine-learning workloads. Reduction techniques. Sorting.

Tensor Processing—Machine Learning Dataflows. Specialized Instruction Sets. Large-Systolic-Array Accelerators (*e.g.*, Google TPU4). Wafer-Scale Message-Driven MAC Grid (*e.g.*, Cerebras CS-2). Exploiting weight regularity and sparsity.

Research Directions—Workload studies. Acceleration of transformer and other ultra-high-compute workloads. Scratchpad v. cache. Many threads v. prefetch. Mixed CPU/GPU chips. 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.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`.