实验九：综合实验

郑海刚

HITSZ 实验与创新实践教育中心
Education Center of Experiments and Innovations, HITSZ

H200 GPU算力

图：4块H100

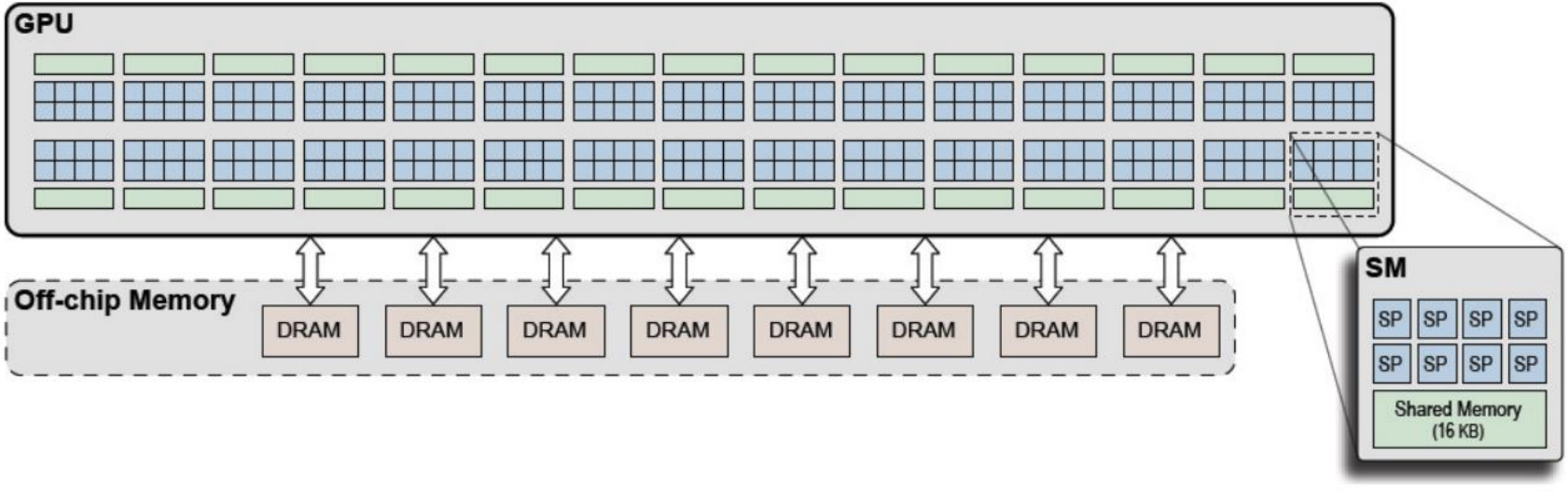
NVIDIA H200 Tensor Core GPU

	H200 SXM[1]	H200 NVL[1]
FP64	34 TFLOPS	34 TFLOPS
FP64 Tensor Core	67 TFLOPS	67 TFLOPS
FP32	67 TFLOPS	67 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS	989 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP16 Tensor Core[2]	1,979 TFLOPS	1,979 TFLOPS
FP8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
INT8 Tensor Core[2]	3,958 TFLOPS	3,958 TFLOPS
GPU Memory	141GB	141GB
GPU Memory Bandwidth	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Confidential Computing	Supported	Supported
Max Thermal Design Power (TDP)	Up to 700W (configurable)	Up to 600W (configurable)

A100 GPU算力

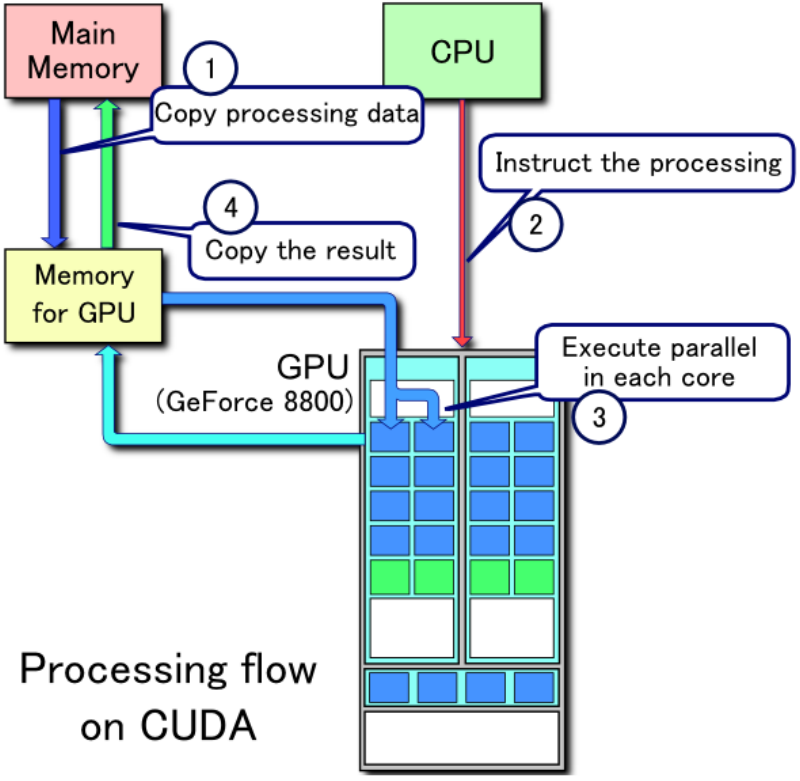
	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935 GB/s	2,039 GB/s
Max Thermal Design Power (TDP)	300W	400W ***

GPU实现DGEMM


GPU
Off-chip Memory
DRAM
DRAM
DRAM
DRAM
DRAM
DRAM
DRAM
DRAM
SM
SP
SP
SP
SP
SP
SP
SP
SP
Shared Memory
(16 KB)

CUDA

- Nvidia GPU的并行编程框架



课程评分安排

- 基于框架的DGEMM实现70分：lab1-lab6
  - naive矩阵乘、openblas矩阵乘、多线程矩阵乘、openMP矩阵乘
- DGEMM的MPI实现：10分
- HPL性能测试：5分
- 应用优化：上限10分
- 答辩（可选）：5分

答辩

- 最后一次课答辩，限5分钟：
  - 分享应用优化的实现
  - 包括但不限问题分析、方案设计、实现、结果分析
- git提交截止时间：最后一次课结束一周内