

CUDA

A100 tensorcore

<https://zhuanlan.zhihu.com/p/620257581>

GPGPU v2.01.pdf

<https://zhuanlan.zhihu.com/p/166180054>

https://www.tinyedi.com/cuda_learning/#pipeline

<https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>

<https://zhuanlan.zhihu.com/p/486224812>

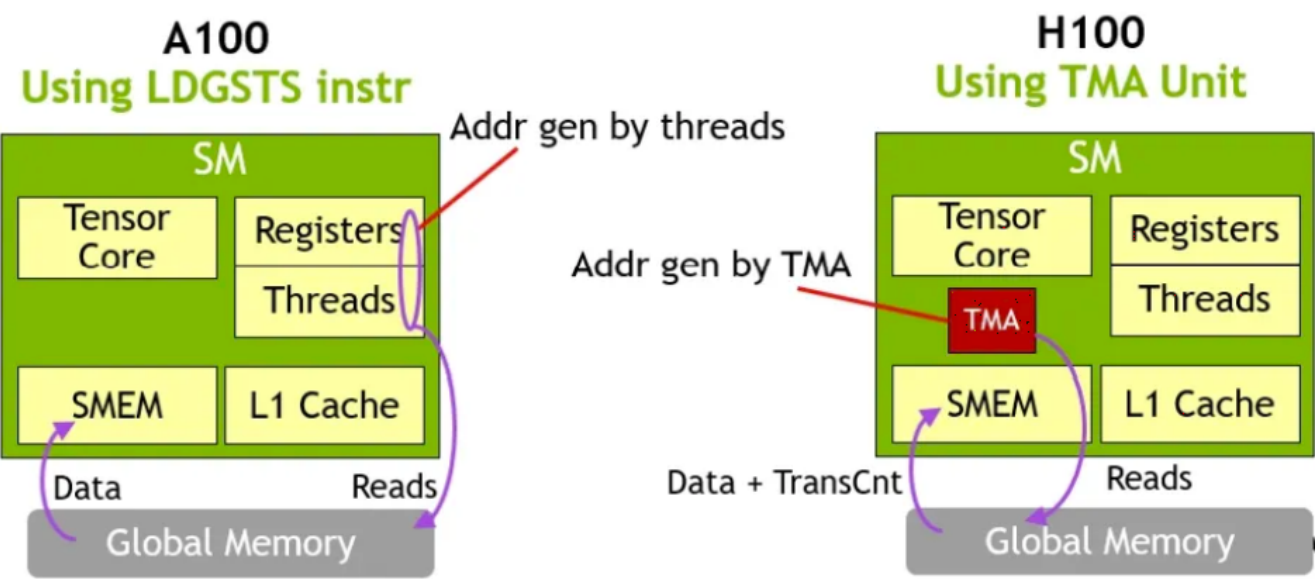
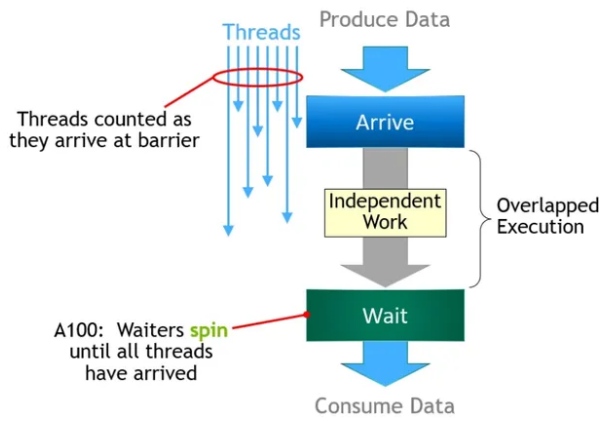
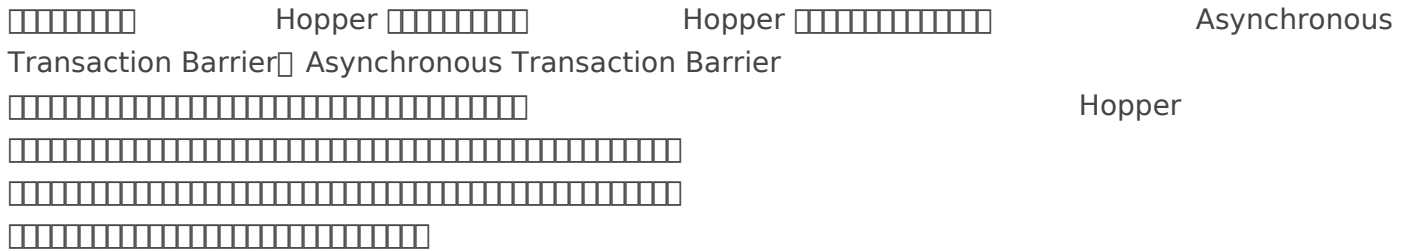
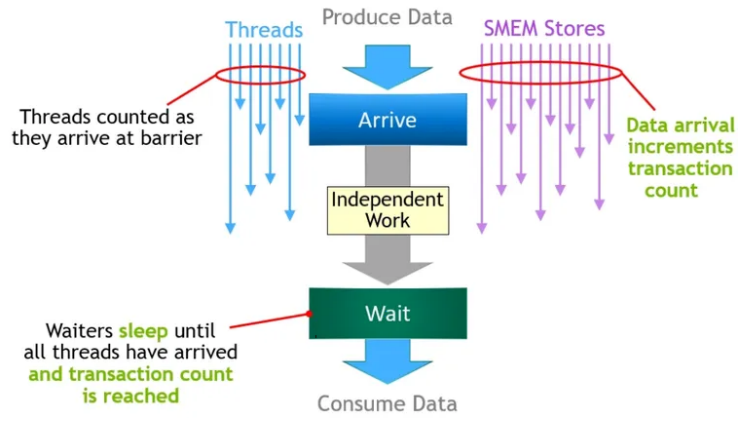


Figure 19. Asynchronous Memory Copy with TMA on H100 vs LDGSTS on A100

Asynchronous Barrier (from A100)



Async Transaction Barrier (New on H100)



Predicate [] [] @P0, @!P3 [] predication [] PTX [] [] guard
 predicate [] [] predicate register [] P0, P6, PT [] []
 [] [] 4bit [] []
 [] predicate [] 3bit [] [] $2^3=8$ predicate register P0~P7 [] [] P7=PT
 [] True [] 1bit [] [] @PT [] [] @!
 PT [] []
 Predicate [] [] [] conditional branch []
 [] [] predicate [] [] warp [] [] []
 [] [] divergence [] [] branch [] latency [] [] instruction cache []
 [] [] [] branch [] [] divergence [] [] []
 mask [] [] thread [] active [] [] [] mask [] PTX [] [] warp
 vote [] [] load [] [] %lanemask_* [] [] [] warp [] [] mask [] []

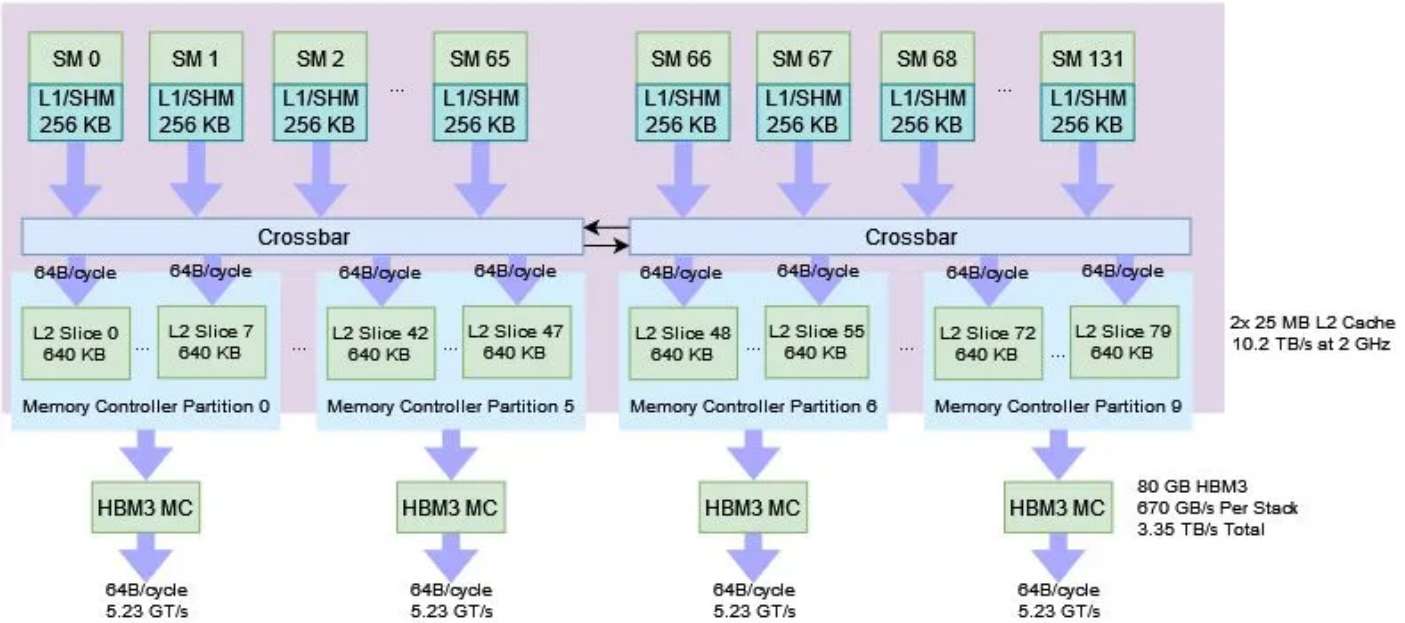
Cache control [] [] [] cache line [] []
 [] [] [] load [] [] [] cacheline [] [] cache [] []
 [] [] [] cache [] [] [] load [] latency [] [] [] PTX [] [] []
 [] [] Turing [] [] [] TLB [] latency [] [] [] latency [] []
 [] [] [] [] [] [] [] [] [] []

Control code [] [] reuse [] read barrier [] write barrier [] wait barrier [] yield hint [] stall count [] []
 [] [] reuse [] [] [] [] GPR [] bank conflict [] [] []
 [] [] [] GPR [] [] [] [] barrier [] [] [] thread [] [] []
 [] [] [] [] [] [] [] [] yield [] stall count [] [] [] warp [] [] []

memory [] [] [] [] memory bound [] [] []
 [] [] [] [] memory bound [] [] [] [] []
 [] [] [] [] [] [] [] [] [] [] latency [] [] [] [] []

stall□□□□ cycle□□□□□□□□□□

Nvidia Hopper (H100 SXM5)



CUDA

- Adjust kernel launch configuration to maximize device utilization.
- Minimize redundant accesses to global memory whenever possible.
- Avoid long sequences of diverged execution by threads within the same warp.

TMA

TMA SM 1D
5D
add/min/max /

memcpy_async API, group, barrier, pipeline.

- group: cooperative_groups::wait(group), group.sync(), cooperative_groups::wait, group.sync().
- barrier: barrier.arrive_and_wait(), barrier count, async_memcpy.
- pipeline: queue

