

LLM??AI????????

????

- 1. 1. MOE
- 2. 1. Attention FlashAttention
- 3. KVcache Deepseek Flash MLA
- 4.
- 5.
- 6. all to all
- 1. latency throughput
- 2.
- 7. MMU
- 8. Atomic reduction

??&??

- 1. / memory barrier
- 1. fence trigger leading tailing latency
- 2.
- 1.
- 2.
- 2. mailbox global sync

????

- 1. prefetch latency trigger latency
- 2.
- 3. launch DMA
- 4. 1D 2D DMA NoC launch
- 1. prefetch/preload
- 2. launch
- 5. “ ” launch join
- 1. “ ” launch /
- 2. launch launch
- 3.

??????

1.

--	--	--	--	--	--	--	--	--	--
1. MOE








--	--	--	--
2.

--	--	--	--

?????

1. Dequant GEMM
2. L1 L0 Tensor

?????

1.  dot  dot 
2.  dot  dot 
3.  linear