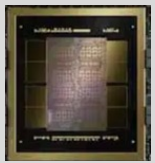
Prefill Node
(Compute Optimized)

Decode Node
(Memory Optimized)



GPU

Offload
Context

Load
Context

DRAM (10-25GB/s)

SSD (1-4GB/s)

ICMS

Send Context

(KV Cache)

Tokens/Context

LPU