Single NAND Package Solutions for AI Applications

Jeff Yang

Storage research Dept. Director

Silicon Motion Technology Corp.

FMS the Future of Memory and Storage

SiliconMotion

AI's storage requirement

- Data ingestion, preparation and Training the model: load the data from storage.
- Checkpointing restore/reload: high BW sequential write/read.
- Inference: Sustained read bandwidth.
- Sudden huge sequential read(50GB/sec)/write(10GB/sec) bandwidth.
- Edge device may only needs good inference capability.
  - After adopting some proper data compression, it still needs 16GB/sec read bandwidth to provide minimum requirement from 5~10 token per second. (under the small language models)
- Inference capability with several tens GB devices may become popular everywhere.

Storage and Memory types for AI applications

Remote storage/memory

High ability,
share by different remote user,
Longer latency, but easier to extend

Through the networking

On the same substrate

Attached storage and memory.

High bandwidth Memory to Storage
Still need to reduce the cost.

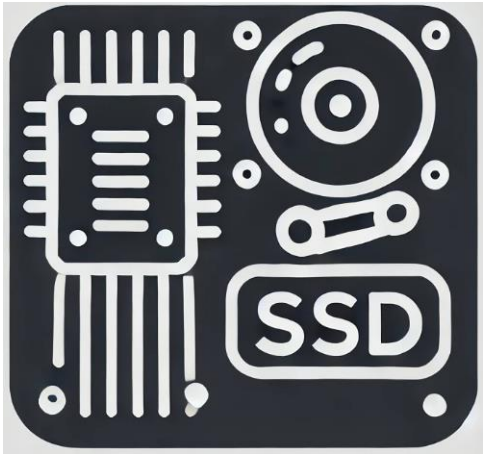
memory/storage access skill

Locality, LRU, compress, deduplications

Directly attached

NVDIMM,DIMM, PCIE hardware slot

Hardware extension limitation.
Acceptable latency for traditional computation, but not acceptable for AI.



In-memory Computation (ICM)

RRAM Matrix Vector Multiplications

It is possible to have a solution with Low cost, low energy.
Error bit on RRAM.

Present dynamic SLC/TLC/QLC management

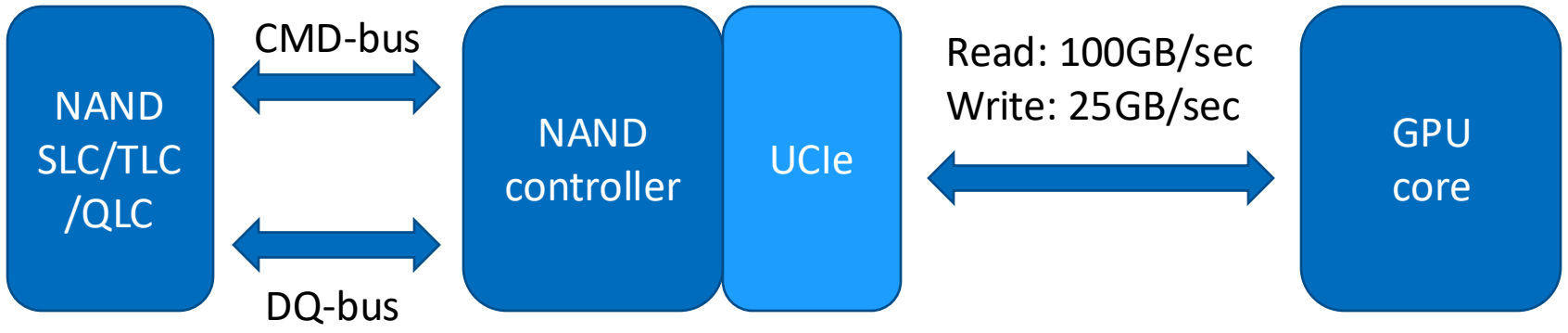
- Single channel NAND package from 4LUN to 8LUN or 16LUN in the future. (256GB per LUN)
- The plane number per LUN increases from 4 to 6 or 8 in the future.
- SLC(80usec for one page) and TLC(1msec for three pages) have one-pass program capability.
- QLC(7msec for four pages) is slower than SLC/TLC but provides huge capacity. It will be good to move the QLC data write into background.
- Extreme SLC capability: (Faster)
  - Write: 16LUN*8plane*16KB/80usec = 25.6GB/sec.
  - Read: 16LUN*8plane*16KB/20usec = 100GB/sec.
- Extreme TLC capability:
  - Write: 16LUN*8plane*16KB/330usec = 6GB/sec.
  - Read: 16LUN*8plane*16KB/40usec = 50GB/sec.

Current SLC is good enough but

1. SLC Cost expensive, TLC low write perf.
2. NAND IO-speed is too slow. (4800MTs)

From 1M IOPS to more than 10M: Enhancing NAND's Data Access Capability

- Reduce Memory usage and transfer to NAND flash for inference applications.
- According the JEDEC230G standard with DDR4800MTs I/O-speed, 1M IOPS become achievable in single NAND channel. But it is still too slow for AI applications.
- Enhance the NAND DQ-bus from 8-bit to 128-bit and transition PCIe to UCIe.
- Provide background operation to save the bus efficiency. Merge the SLC to TLC/QLC without consume DQ-bus resource.

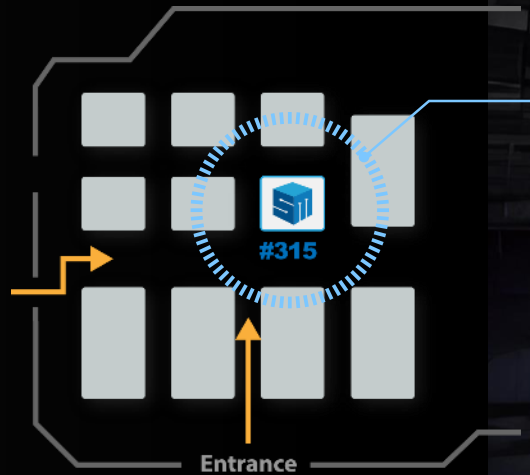


Summary

- Cost is one of the most important topics for a certain application become popular.
- Reduce the inference cost is critical.
- Most to the edge device cannot have a strong CPU or GPU.
- Put the inference capability into the flash controller is a best way.
- Improve the NAND’s IO bandwidth is the first item in our wish list.

Meet us at booth #315

Scan to learn more!



SiliconMotion

Driving AI Innovation in Flash Storage