3D NAND Technology Scaling helps accelerate AI growth

Jung Yoon, Ranjana Godse – IBM Supply Chain Engineering
Andrew Walls – IBM Flash Systems
Agenda

- 3D-NAND Scaling & AI
- Flash density trend
- NAND Layer Count scaling trend
- $/GB trend
- NAND technology Bandwidth comparison
- Flash Energy/Power trends
- Summary
Deep learning multilayered neural network require high throughput data ingestion, lower power consumption, small system footprints, and lower system cost. **AI acceleration to be realized by 3D-NAND scaling**
Characteristics of AI

• Deep Learning, Machine Learning and all forms of Analytics on Data requires ingestion of large amount of data at very high throughput – Real time AI require data pattern & relationship learning

• Workloads tend to be In-Memory database centric – need to ingest data at high rate & throughput
  • To have a large dataset it requires lots of memory and lots of servers.
  • DRAM is extremely expensive, challenges with bit cost reduction via scaling

• As data sets used for training ML/DL are growing over time – Flash with its low latency & high throughput is most optimal solution for AI storage
Attributes of AI

- Flash has very high read throughput density to assist workloads like deep learning.
- Real Time analytics require low latency access to lots of data.  
  - e.g., credit card fraud detection, image recognition
- Flash with increasing density can allow for lots of data to be accessed at fairly low latency.
- Flash enables High IOPs with low read latency, smaller footprint and lower power compared to HDD.
- Low Latency NAND and SCMs - fill performance and cost gap between Memory and Storage.
3D-NAND Scaling & AI

• 3D-NAND technology scaling provides high throughput density, lower power consumption, and smaller system footprints – critical to accelerating AI growth

• 3D-NAND enables multi-terabit TLC and QLC densities through cell layer count increases, innovations in circuit design, process technology, and stacked packaging

• The resulting cost reductions, throughput increases, and lower power consumption enable flash driven AI growth.

• Density, performance, and reliability tradeoffs must be considered for TLC, QLC, MLC, and low latency NAND. High throughput data ingestion and system reliability requirements for AI optimized workloads critical
Flash density growth will continue via 3D NAND layer count scaling 2018-2022+ - TLC, QLC, LL-NAND

3D NAND scaling on a 18-24 month cadence thru 2020 – 64L > 96L >120L >190L

Flash with increasing density can allow for lots of data – accelerating AI growth
Layer count trend for 3D NAND

Santa Clara, CA
August 2018

3D-NAND scaling expected to continue via 3D Layer count increase >200 layers & lithography shrink in 2022+ & beyond – no need for EUV for 8+ years in flash

Significant 3D-NAND Process, Architecture & Cell materials innovations needed for continued >200Layer scaling

Ref: 3D NAND process trends. TechInsights
Increase in number of layers make the high aspect ratio etch challenging-z-directional 3D-NAND cell process controls become critical with multiple etch processes.

Process innovations required in High AR Channel Etch, Stair case contacts, defect control, wafer yield.
• Flash Market Demand growth – driven by high density 3D NAND bit cost scaling with increased layer count
• Strong penetration of 3D TLC & QLC in Cloud Datacenter & Enterprise Storage in 2018-20
• 3D-NAND scaling enabled bit cost reduction combined with low latency, high throughput accelerating flash driven AI growth
• Drastic increase in Flash-SSD vs HDD in the past 5 years.
• SSD density growth driven by increase in number of layers and capacity for 3D TLC and QLC NAND
• Flash-SSD enables smaller system footprint over HDD– flash driven AI system acceleration (Petabytes of storage can be put in single rack-mount enclosure)
Density growth for 3D NAND

- Increase in 3D-NAND layer count drives significant increase in Si density/mm² & Package density/mm³
- 3D-NAND scaling density/mm³ increases driving significant storage system footprint efficiencies – optimal for AI Storage solutions

Max Flash memory density/mm³ in 16DP stacked Package

Max Flash memory density/mm² in 2D Si area

Santa Clara, CA
August 2018
• Bandwidth is the maximum amount of data that can travel through the channel.
• Complex workloads and volume of data is feed to Deep learning set.
• Thousands of threads processing this massive data have to reply on high data delivery rate from storage
• Flash has a very high read throughput density (over HDD) - optimal for deep learning workloads
With 3D-NAND generations, program and read Energy decreasing trend

Flash consumes less power vs HDD – significant difference in cost at large scale AI storage systems
NAND technology Energy trend

- Assuming 8% Read, 23% Program and 14% Erase workload
• AI growth will be accelerate by 3D-NAND technology scaling via high throughput density, lower power consumption, and smaller system footprints.

• 3D-NAND enables multi-terabit TLC and QLC densities through cell layer count increases, innovations in circuit design, process technology, and stacked packaging.

• The resulting cost reductions, throughput density increases, and lower power consumption expects to enable significant flash driven AI growth.