3D Memory

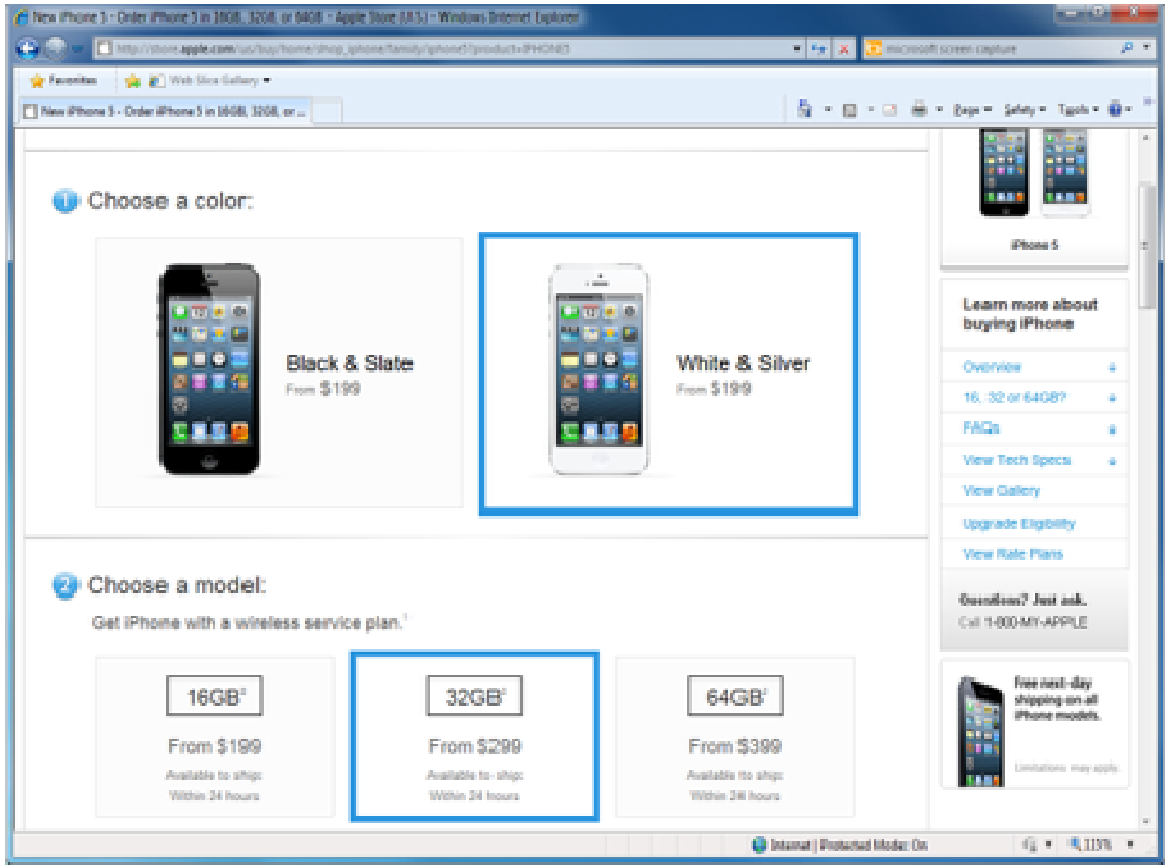
Trends and Obstacles

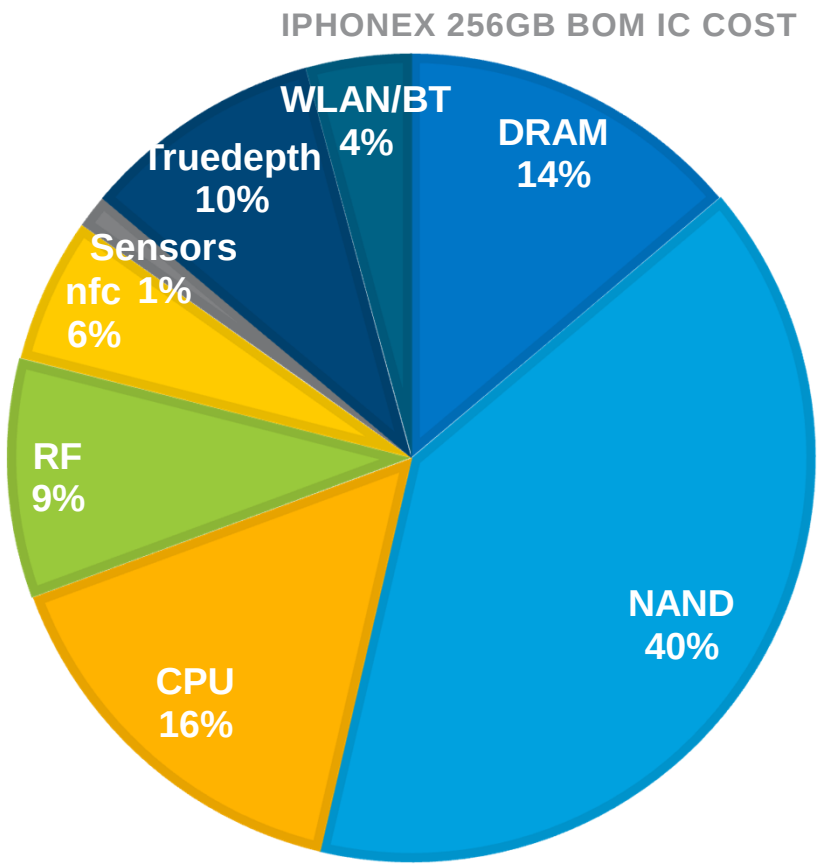
Anton Korzh: System Architect ACS Pathfinding

Outline

- History of 3D memory: what exists of today
- Why do people want to go 3D
- What problems could be seen?
- And what could be the solutions
- Conclusions

Memory is becoming more important then compute

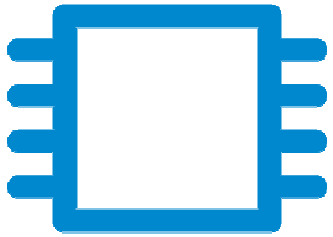




Source: IHS iSuppli Teardown Analysis, Oct 2017

Micron

Memory Trends



- Planar scaling is increasingly challenging
- DRAM expected to be ideal memory
  - Already exploit redundancy at fabrication time to maximize yield
  - Does the system need perfect memory
- And still DUMB – managed by external controller

- NAND exhibits some autonomy but hides behind HDD interface

New (Memory) Technologies Are Rare

TODAY'S AVAILABLE MEMORY TECHNOLOGIES EMERGED IN THE EARLY 70S

THE WAY MEMORY IS USED IN SYSTEMS, THE MEMORY HIERARCHY, HAS BEEN DEFINED BY THE EVOLUTION OF THESE MEMORIES OVER FOUR DECADES

The hierarchy and the hardware and software wrapped around it is as much defined by each memory technology's "limitations" as well as its "features"

Year of 1st Shipment	Memory Technology
1969	SRAM
1970	DRAM
1971	EPROM
1986	NOR Flash
1995	NAND Flash
1997	MLC Flash
2008	PCM

Electron Based

EPROM Derivatives

What is Next?

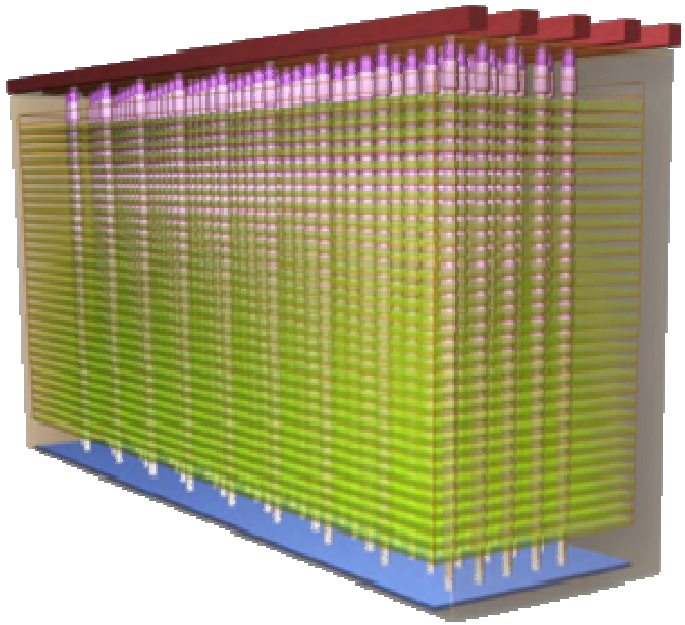
Memories that have shipped ≥ 1Gb densities

3D memory types

- 3D NAND
- 3DS TSV DRAM
- HMC
- HBM
- Emerging persistent memory
- Stacked SRAM on top of compute die

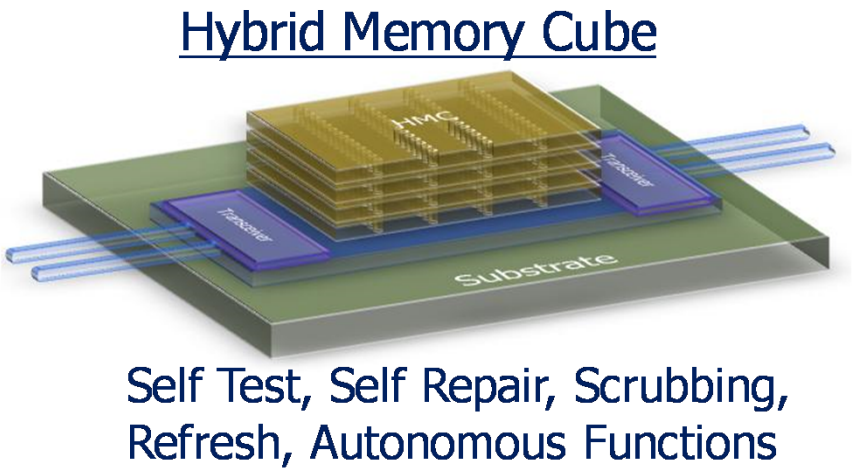
3D NAND

- Manufactured in layers not stacked
- Based on floating gate cell technology
- Up to 96 layers
- 1,2,3 or 4 bits per cell
- Lowest cost and highest density
- Limited endurance and need error correction
- 1Tb die produced



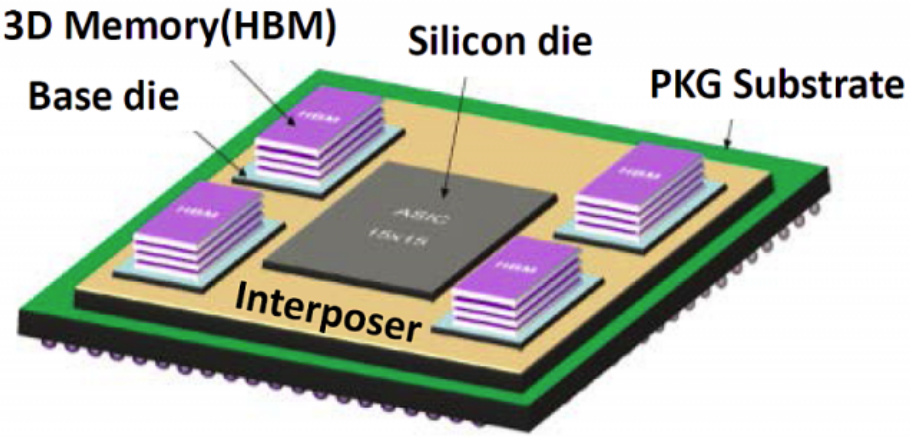
Hybrid Memory Cube

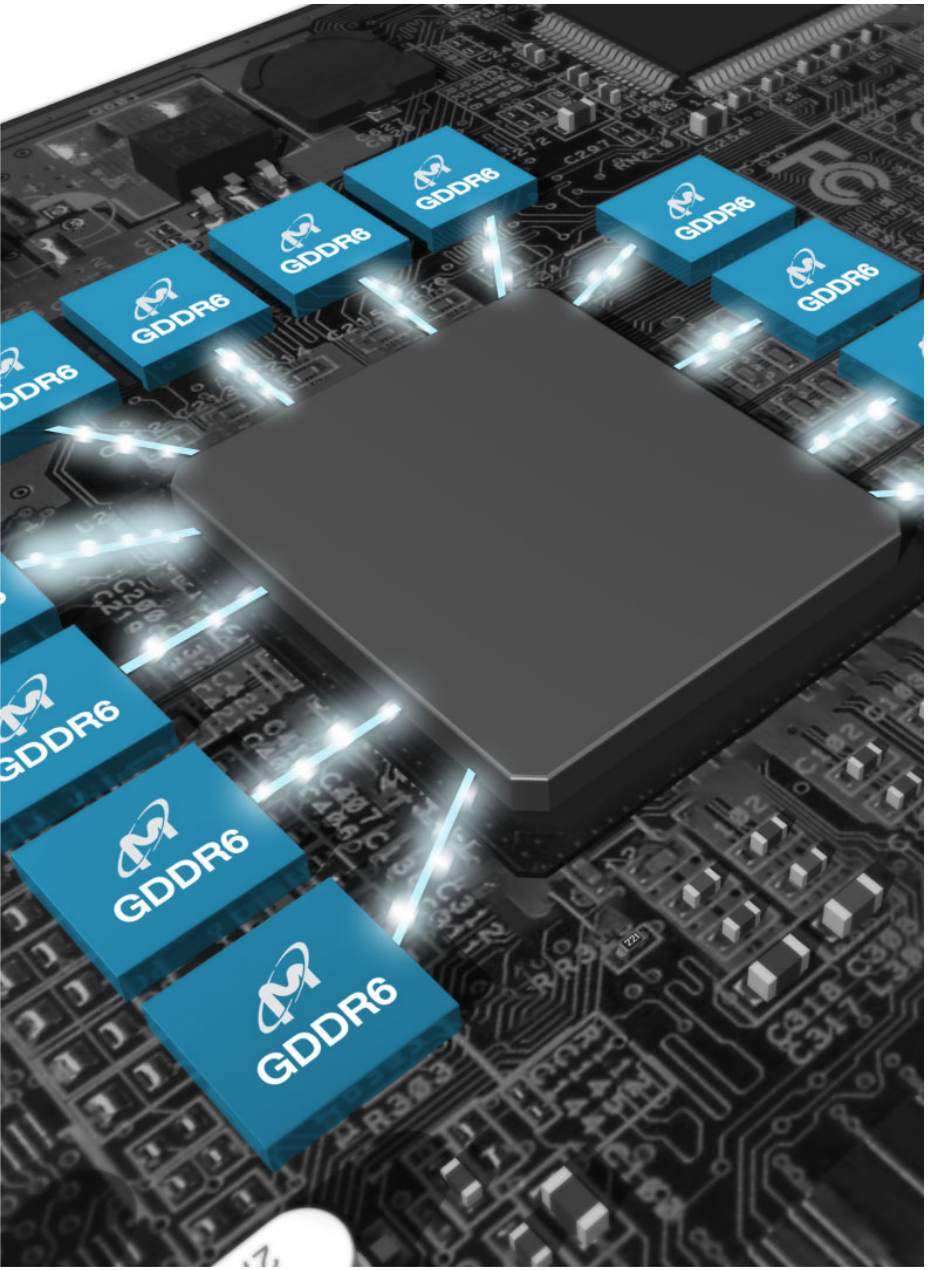
- First stacked DRAM introduced 2011
- Stacks DRAM dies on top of controller die
- Requires expensive TSVs
- More parallelism inside DRAM (vaults)



HBM

- JEDEC standard
- Same TSVs, but no logic layer (yet)
- Legacy interface scaled
- High bandwidth achieved not by increasing concurrency, but rather scaling number and frequency of IO pins





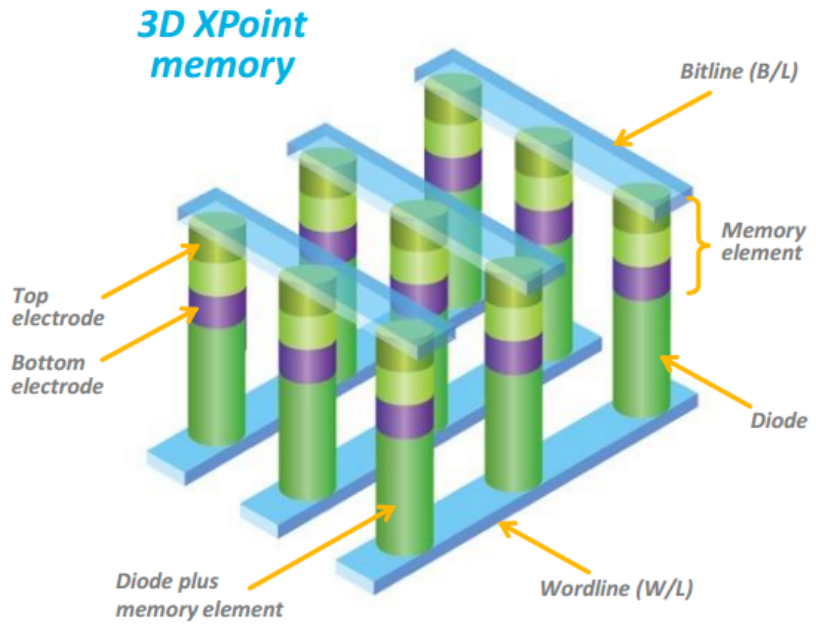
Lesson learned: HBM and GDDR5x

- GDDR5x was a huge success
- Solved a bandwidth problem with much smaller cost
- Hasn’t solved the density and still high energy
- Not everyone needs/wants stacked memory
- Adopted by industry as GDDR6
- GDDR6 is in mass production; demonstrated 20GB/s throughput

Micron

New emerging NVM memories

- Important new memory type
- Manufactured in 3D like NAND
- Much closer to DRAM in performance
- Non-volatile
- 1000x more endurance then NAND



Why do we want to stack memory?

- Need more capacity and more bandwidth given scaling is slowing down
- Memory footprint reducing to increase compute density
- Bring memory closer to compute (energy)
- This all comes at a cost

Existing memory interfaces

- DRAM (RAS-CAS based)
  - DDR/GDDR/LPDDR
  - HBM
- NVM
  - SATA
  - NVMe
- GenZ, CCIX still not there
- Largely stuck in the simplest possible interface model from 70s
- Error-correction is centralized, preventing optimizations

Micron

What might we want from memory interface?

- Transaction based
  - More information about what accesses memory (not just addr and size)
- Very high concurrency
- Support for PIM and memory management
- Support for different types of memory (DRAM, NVM etc)
- Serialized
- Does that sound familiar? In HPC networks do support that already!

Micron

New interesting research direction : FGD

- Fine Grain DRAM papers by NVIDIA
- More fine grain channels instead of few fat ones
- Requires high concurrency from compute side

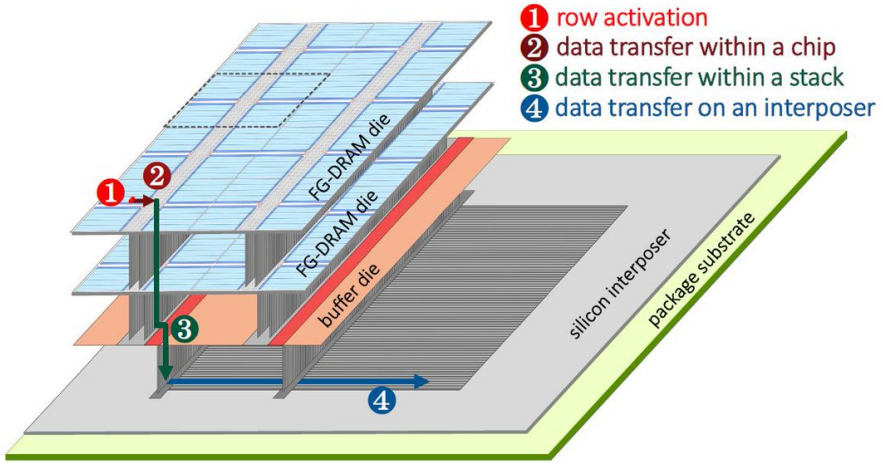
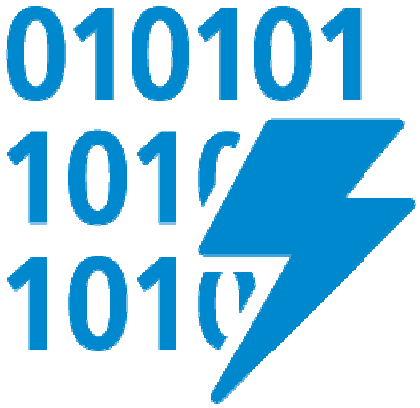


Figure 5: FGDRAM Die Stack Architecture

How else can we save energy?

- Most energy is spent on IO
- We can save lots of energy if do processing in memory
  - or in the controller die
- Not supported by DDR-like interfaces
  - Not only lack of commands
  - But controller wont let us open close rows on our own



Micron

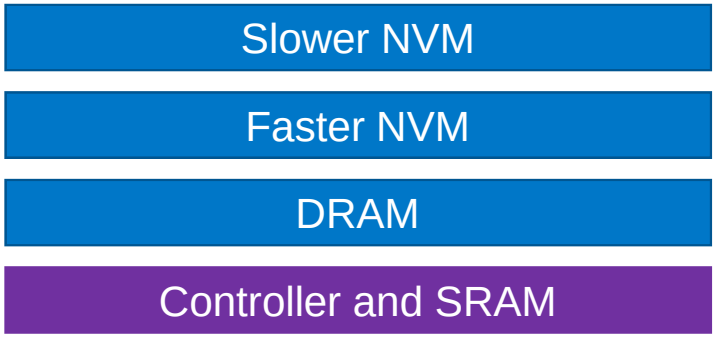
In-memory intelligence by Micron

Test case	IMI	Comparison	Comparison hardware
Gaussian Blur: stencil operation using separable kernel in CompuBench benchmark scaling from 1 × 1 to 15 × 15.	9.9 Gpixel/s	147 Mpixels/s	Samsung Exynos 7
	6 W	5.49 W	
Alpha Blend: 24 bits per pixel planar images.	600 Gbytes/s	202 Gbytes/s	Nvidia Titan X
	7 W	200 W	
CSV Parsing: four fields per record with one field converted to a non-negative integer.	2 Gbytes/s	28 Mbytes/s	Intel XEON E5
	7 W	90 W	
SHA1 (Bitcoin mining): with a large number of independent hashes.	2.2 ms/9M hashes	800 ms/8M hashes	Intel XEON E5
	7 W	90 W	
Image Fusion: registered images using HIS method. 2,048 × 1,080 32-bpp LRMI data with 4,096 × 2,160 16-bpp LiDAR data.	16 ms/frame	68 ms/frame	Intel XEON E5
	7 W	90 W	

Micron

Another appealing idea multi-tiered memory

- Every SSD is already multi-tiered
  - DRAM, SLC, TLC flash
- Hard to hide anything behind RAS-CAS interface
  - Deterministic timing
- We can do many interesting things
  - Memory side cache



Micron

Might need changes not in interface only

- Todays OS/VMM treats DRAM as perfect memory
  - Constant in size and perfectly reliable, equally accessible
  - Manages DRAM without letting hardware know

- NVM devices on contrary are supported as block device and FS

OpenChannel SSDs and Project Denali

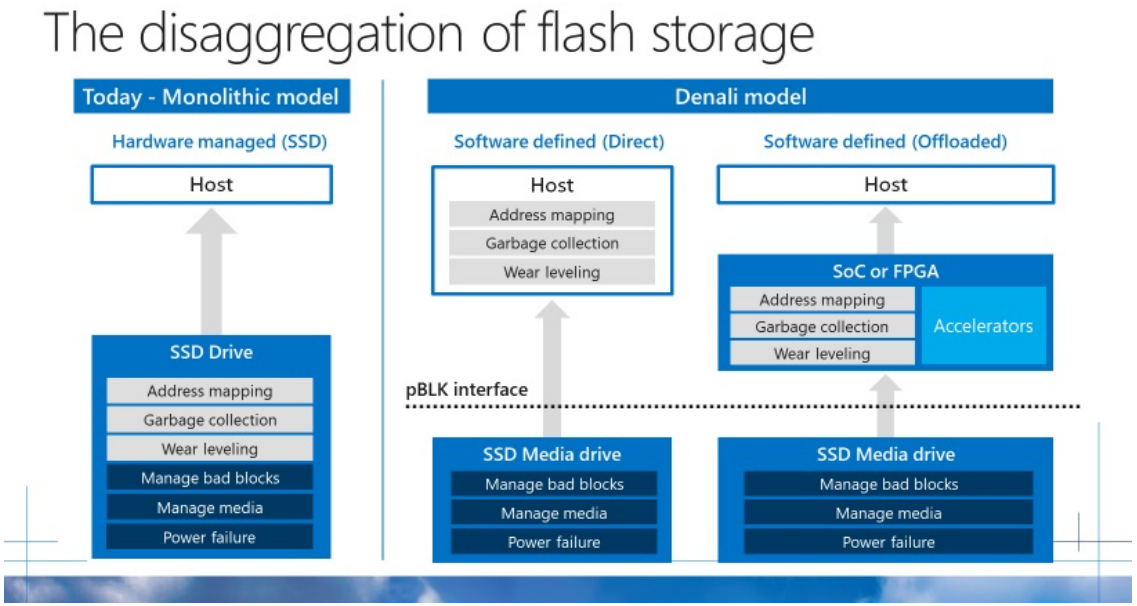
Datacenter folks not happy with flash interfaces also

Open Channel SSD exists for Linux

Microsoft started project Denali

Enables more control and efficiency on the software side and more innovations on memory side

Why nobody is doing the same for DRAM



Conclusions

- New interfaces are required to support memory innovations
- Mixing different memory types together is an appealing solution
- Some intelligence MUST be present in memory subsystem
- We can offer more intelligence, once supported by industry


Micron