ElectronicDesign

3D LAYER FAQ FROM BILL WONG

HOW 3D FLASH MEMORY STACKS UP

3D flash memory stacks a lot of transistors in its multilayer implementation. This technology has driven significant increases in capacity - even as designers have had to adjust to meet the challenges that 3D stacks entail.

What attributes should developers consider with 3D flash memory?

Designers need to keep the performance requirements and characteristics of the application in mind, along with other features like die density, package density and write/erase endurance parameters. Interface speed needs to be considered as well, along with the type of I/O interface and supply voltage for the system.

3D flash generally requires better error correcting code (ECC) support than 2D flash memory. Supporting the new NAND features being defined at JEDEC, such as read and write interface training, will also be a consideration.

What factors contribute to the success of 3D flash memory?

A few factors come to mind when it comes to the success of 3D flash memory.

First is the accelerating data creation that is also driving an increased need for high density storage solutions. With triple level cell (TLC) die densities up to 512 Gbits and quad level cell (QLC) as high as 1.33 Tbits/die, 3D flash is the solution to this industry challenge. Stacking up to sixteen dies in a package enables a TLC 1 Tbyte package solution — and with QLC, an even higher package density of 2.66 Tbytes is achieved.

Second is the flash value proposition – flash has been a disruptive technology, and we all use it on a daily basis. Flash enables the mobility of content on our smart phones, makes notebooks thinner and lighter, gives us the ability to stream on the internet – and much more. Now, 3D flash is extending this value proposition to new applications. 3D flash in the hyperscale memory storage hierarchy enables a high density, low latency, cost effective solution.

The evolution of flash memory to a 3D structure, from a 2D planar structure, was inevitable, as two-dimensional scaling had reached its limits. As 2D memory cell transistors got smaller, the amount of charge they stored steadily decreased. This reduced the ability to retain data over time and temperature, and results in lower write/erase cycles. We see this effect in DRAM, as the need to increase refreshing has grown. 3D flash has enabled the continued increase in bits stored per die, which allows higher density per package and lower cost per bit.

How many layers can we have and what are the technical challenges to increasing the number of layers?

Over 200 layers is now commonly discussed - and some vendors are suggesting as many as 800 layers are possible. A higher layer count comes with technical challenges in addition to the challenges of increasing die size and performance.
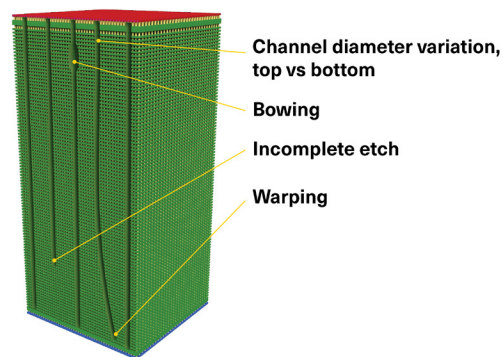
The biggest technical challenges to adding more layers show up in manufacturing process control. Keeping hole uniformity from the top layer all the way down to the bottom is more difficult with more layers. Hole alignment and depth all come into play.

Channel diameter variation, top vs bottom

Bowing

Incomplete etch

Warping

3D Flash Memory: Higher layer count increases manufacturing challenge

From a manufacturing standpoint, a higher layer count will, in turn, increase the fabrication time and yield loss due to an increased number of operations. Physical stress on the die is difficult to predict at this time, but since the thermal coefficient of expansion will not be the same between the Si substrate and the memory layers above it, there will be thermal stress (reminiscent of how a bimetallic strip curves with temperature).

How important are the number of layers?

The actual number of layers does not define technology leadership. It is more important to meet market and customer performance requirements.

More layers tend to correlate with higher densities, but the cost per bit does not necessarily scale with the layer count. With 2D flash, smaller lithography always translated to a lower cost per bit due to a smaller memory cell. With 3D flash, the layer count is not lithography and it adds incremental cost to the storage device.

What criteria should be used to evaluate 3D flash memory solutions?

There are five key attributes that should be used to evaluate 3D flash memory solutions. These include interface speed, overall performance, system reliability, power requirements and, of course, cost.

KIOXIA

Higher performance interfaces that will enable PCI Express and the corresponding NVMe protocol create higher bus speeds. Sequential read/write performance improvements are being made with each new generation. New features like improved random read performance are also in the mix. From a system reliability perspective, number of write/erase cycles, data retention and ECC capability requirements are important.

IO voltage has migrated from 3 V to 1.8 V to 1.2 V. This helps with higher interface speeds as well as improved power efficiency.

Continued innovation using different design approaches will help improve cost structure. Innovations like increased layer count, improved memory hole density and increasing number of bits per cell have helped make 3D flash a competitive memory solution.

How does CAPEX come into play and what difference does it make for those already invested in 3D technology?

CAPEX is significant because it is very expensive to develop 3D flash memory. It is typically about three to five times more expensive compared to the usual floating gate architectures due to the new equipment required, including etch and deposition machines. Production throughput time is also longer. As suppliers increase the number of layers in 3D generations, ongoing CAPEX investments are required for the tooling and equipment.

Supporting market performance expectations while also optimizing CAPEX investments is the key to a successful 3D flash memory project. Industry suppliers must consider the internal tradeoffs of squeezing a bit more from current investments versus making additional CAPEX investments to meet changing market performance expectations. This includes how the layer count comes into play and explains why suppliers' 3D generations with similar performance may have different numbers of layers.
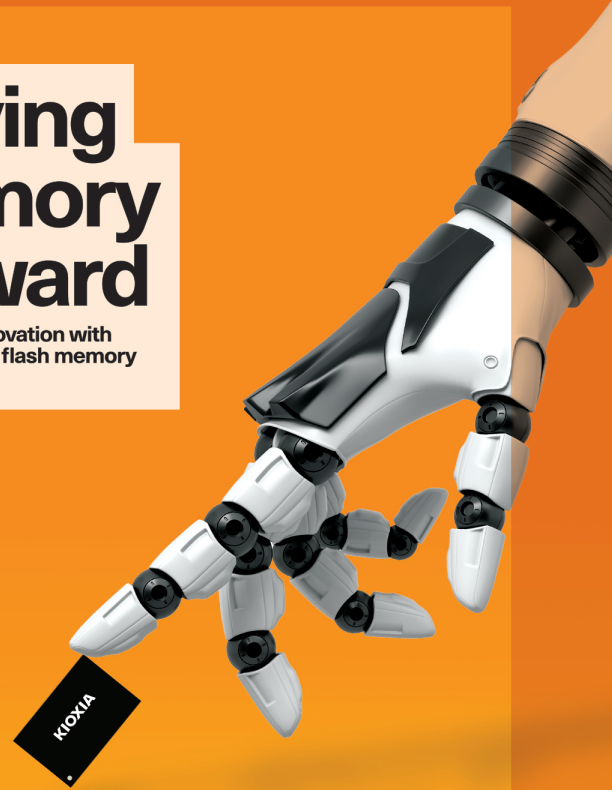
Sponsored by

KIOXIA