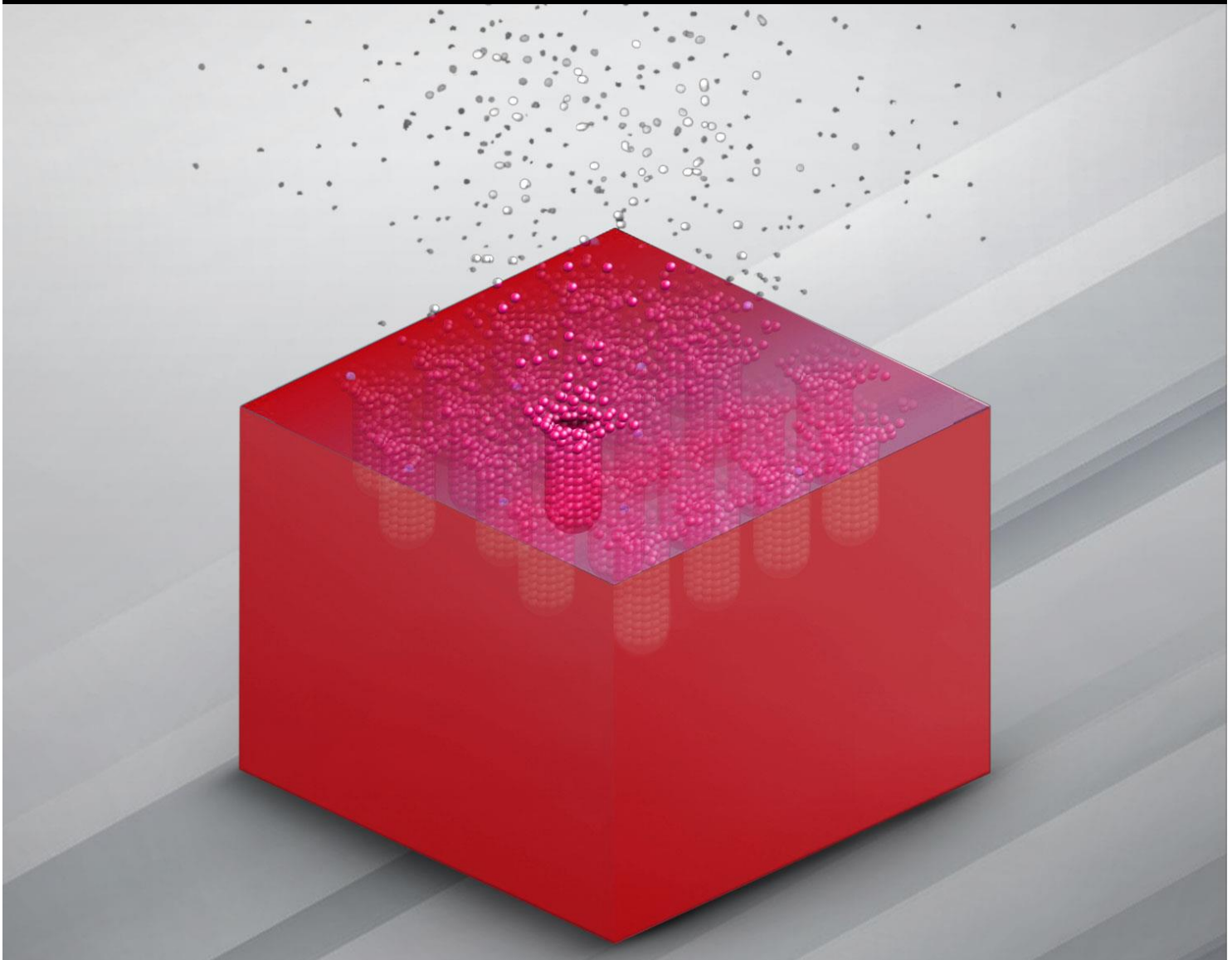
Counterpoint | Lam RESEARCH

Scaling to 1,000-Layer 3D NAND in the AI Era



July 2024

Executive Summary

3D NAND has become the mainstream architecture for NAND Flash memory as it has proliferated across key applications from mobile to cloud. The 3D NAND market remains dynamic as suppliers race to add more word-line layers (300+) this year to 1,000+ layers towards the end of the decade. The goal is to achieve denser but pristine 3D NAND architectures with impeccable performance that nicely aligns with the growing capabilities for compute and DRAM as we enter the AI era.

However, each of these suppliers are looking to increase the density of 3D NAND by stacking more layers in the same die and face significant challenges in scaling the architecture vertically, laterally and logically. These can be only addressed with advancements in wafer etching technologies and techniques from leading equipment manufacturers. This requires innovations across etching technologies and systems combined with novel chemistries. These advancements can enable deeper and faster etching, improve vertical scaling, minimize the profile deviation and boost density increase with repeatability.

This should help NAND suppliers march towards the 1,000-layer roadmap with better yields and lower costs. The icing on the cake would be if all of this is done in a highly sustainable manner.

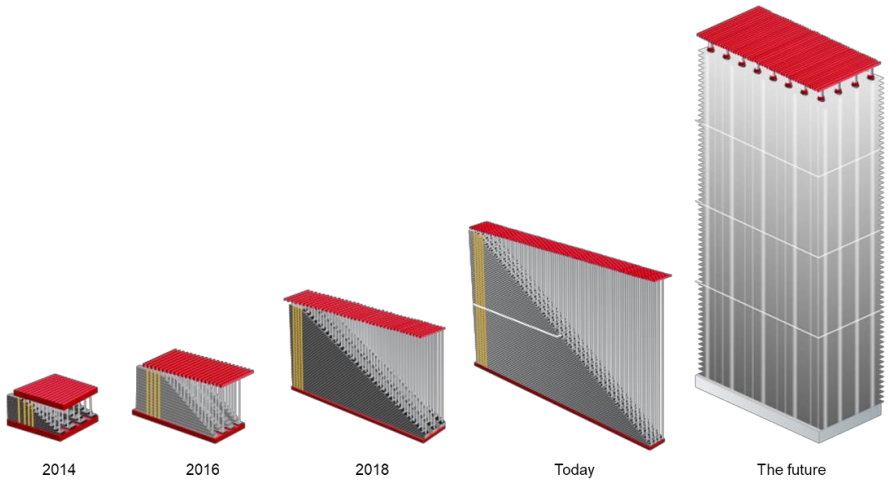
The 3D NAND Revolution

NAND FLASH technology has been the primary technology for low-cost and large-density data storage applications for several decades and has enabled the revolution of every end market from USB drives to mobile phones to servers. The non-volatile 2D or planar NAND FLASH memory has been the work horse of the solid-state storage industry, delivering increased performance, reliability and lower power consumption at lower cost per bit for the first couple of decades.

However, the advent of iPhones and Android smartphones has triggered an exponential need for storage in devices as well as cloud servers to store data generated from millions of applications across billions of devices. Over the years, user-generated content has grown from HD images to 4K videos shot via phone demanding more advanced and high-capacity storage solutions. High-quality streaming content including music and videos from the cloud or downloaded offline on device also require a lot of storage capacity.

This demand is being addressed by the industry's move to a 3D structure, where horizontal layers of memory cells are stacked and then connected using tiny vertical channels to increase the storage density. This approach allows the industry to keep pace with Moore's Law despite the challenging physics of making the memory cells smaller. With each subsequent generation from Single Layer Cell (SLC) to Quad-Level Cell (QLC) flash, the number of bits per cell kept on doubling the number of possible voltage states.

Evolution of 3D NAND



Sources: Lam Research, Counterpoint Research

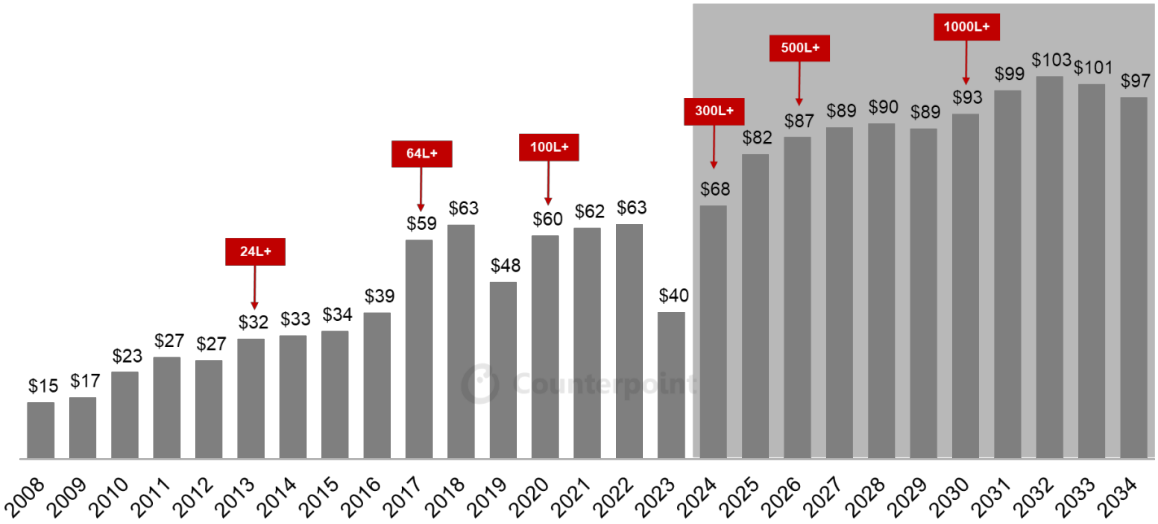
3D NAND flash memory has become the dominant non-volatile memory since the shipment of Samsung's second-generation 32-word-line layer in 2014, an inflection point that was reached soon after the introduction of first commercial 24-word-line

layer 128 Gbit chip in 2013. Depending on the NAND vendors (Samsung, Western Digital, Toshiba, Micron, SK Hynix and others), variations in the 3D NAND structure exist, and are known by different names, such as vertical-NAND FLASH (V-NAND FLASH) and bit cost scalable (BICS).

Over the last decade, the bit density has increased from 1 Gbit/mm^2 to an impressive 14.6 Gbit/mm^2 with a 232-tier layer solution now available. The continued advancement in compute for training and inferencing generative AI models in the cloud and edge coupled with growing high-quality user-generated content via smartphone cameras will drive significant need for faster, higher capacity storage for at least the next two decades. AI smartphones, AI PCs, AI servers in enterprises, autonomous vehicles and robotics, will be key end market drivers for 3D NAND growth.

Consequently, the overall NAND FLASH memory market is expected to more than double to reach **$93 billion by 2030,** up from $40 billion in 2023.

NAND Flash Revenues (In $ Billion) and 3D NAND Evolution Trends



Source: Counterpoint Research

This will be supported by the increasing number of layers being stacked to maintain the bit density scaling trendline and this is expected to continue in the years to come. We see the race continuing with the first 300+ layer 3D NAND likely to be shipped this year and jumping to possibly 1,000+ layer 3D NAND with shipments starting as early as 2030.

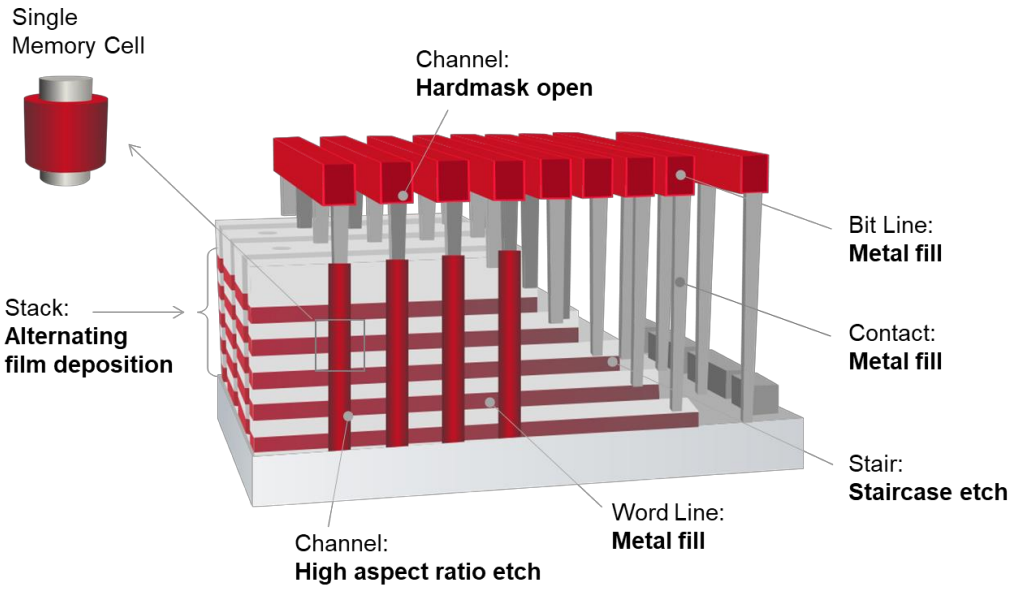
The Future Scaling of 3D NAND: Challenges

Bit growth and cost reduction dynamics have changed significantly when we compare 2D versus 3D NAND. 3D NAND has enabled higher bit growth rate but memory cost reduction per bit has been slowing due to the increased cost and complexities involved in scaling the growing number of layers.

Moore's Law for processors has been lagging the last few years but has held well for NAND so far. NAND scaling has slowed with respect to what has been the historical trend line of NAND and as huge investments are required different manufacturers are adopting techniques to stretch out their NAND roadmap.

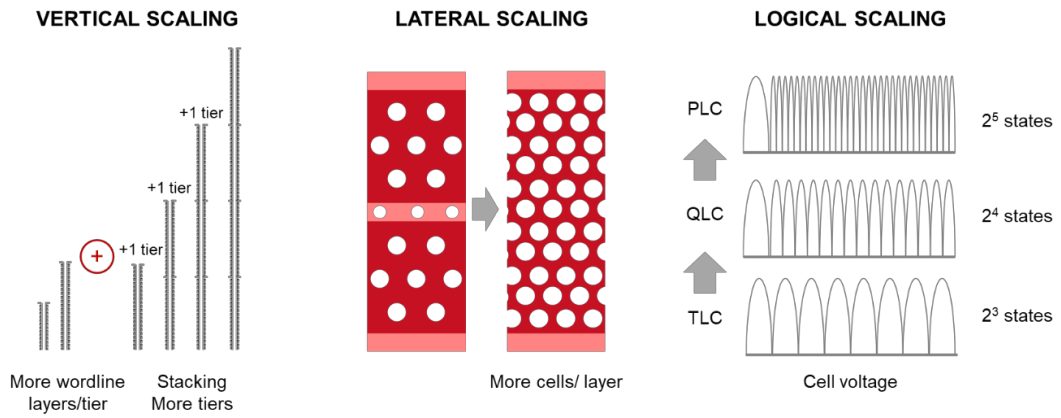
For 3D NAND scaling in the future, word-line (WL) stacking will continue to be a key driver along with XYZ dimension shrink of the cell to alleviate the cost and device challenges introduced by the WL stacking. However, the industry is anticipating NAND FLASH die density to reach 100 Gbit/mm^2 with 1,000-word-line (memory cell) layers going by the trends and improving NAND FLASH cell technology.

Key Process Steps in 3D NAND Memory Cell Formation



Sources: Lam Research, Counterpoint Research

Scaling across XYZ direction along with logic die design optimization, as represented in the figure below, to achieve a minimal cell footprint and die size will continue to be major drivers to alleviate the cost and device challenges.



Memory channel formation requires breakthrough innovation across multiple scaling vectors

Sources: Lam Research, Counterpoint Research

Challenges in Scaling 3D NAND:

3D NAND vertical stack scaling gives rise to challenges mostly on film deposition and etch, unlike devices scaling via feature size reduction. To pattern, isolate and connect vertically integrated 3D memory devices, difficult High Aspect Ratio (HAR) etches are required. The aspect ratio for a hole or trench in general is defined as the ratio of the depth to the width of the hole or trench. Critical processes in 3D NAND manufacturing includes alternate stack film deposition, high aspect ratio etching and word-line metallization. Finding the balance between bit density, read and write speeds, power, reliability and cost is crucial for applications. The process gets complicated as we add more layers to the structure and there are additional capital expenditures, as the number of layers rise, it becomes increasingly costly to add more storage capacity.

However, the process complexity and capital intensity of 3D NAND manufacturing adds to the difficulties fabs are facing in terms of process control, yield and cost per-bit.

NAND is a complex technology that presents some major manufacturing challenges such as high-aspect ratio (HAR) etch processes to enable tiny vertical channels, obtaining enough drive current between the memory cells, logic die design optimization and wafer warpage.

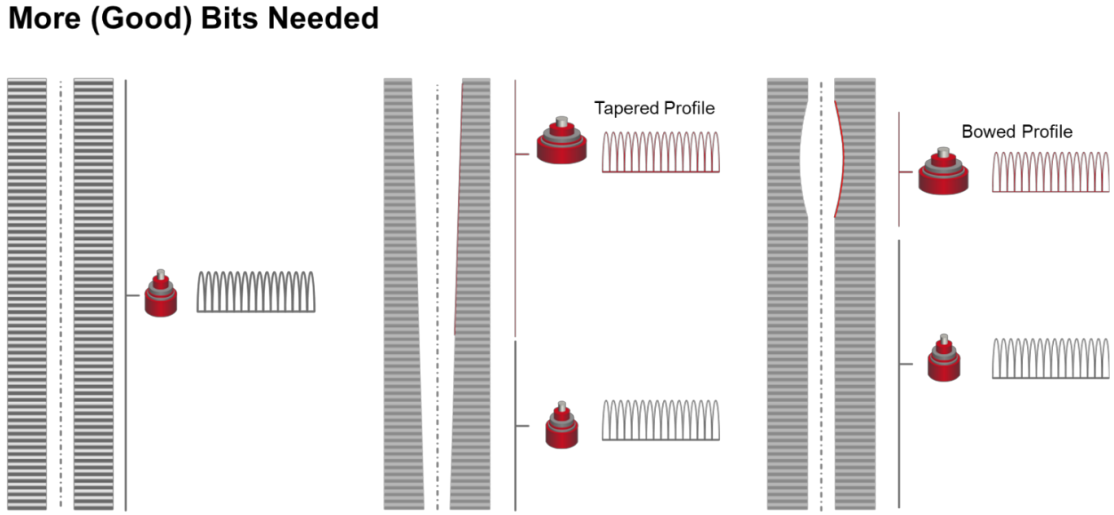
Challenge 1: Reduced Etch Rate

The traditional HAR process slows down the etch rate with aspect ratio increase as the number of layers increases. This is caused by the attenuation of ion and neutral fluxes as a function of aspect ratio reducing the etching rate. This undesired effect is called aspect ratio-dependent etching (ARDE).

The slowdown in etch rate as the etch proceeds deeper results in higher cost and further limits the depth the layer can be etched which could be a barrier for the progress of 3D NAND scaling. As smaller features etch slower and with the feature sizes industry is anticipating, it is of critical importance to enhance the transport of species to reach the bottom of the feature enabling faster etching rate.

Challenge 2: Variabilities in Profiles

Another challenge is maintaining vertical profile from top to bottom. Formation of cylindrical holes inside the films which house the storage devices called high-aspect ratio (HAR) ONON (oxide and nitride) channel hole etch with high uniformity and repeatability when employed in high volume production is a challenge, as it involves forming trillions of perfect channel holes from top to bottom. Further, as hundreds of layers of ONON films are patterned to about 100 nm hole size, the etch aspect ratio (depth/width) is pushed above 100:1 for 1,000-layer 3D NAND with multiple tiers.



A perfect channel shape, formed by etching, is needed for all forms of scaling

Source: Lam Research

Etching at low temperature is known to enhance etch rates and increase surface diffusion and physisorption of neutrals thereby enhancing the etch rate, on the contrary high deposition rates of passivating species may cause clogging at the top of the structures at low temperatures. **Polymer control** becomes a critical challenge in high aspect ratio dielectric etching where polymers on the top of the feature shrink the cross

section of the opening and thereby reduce the ion flux to the bottom of the feature, slowing the etch rate causing profile bowing.

HAR etching plays the most significant role in patterning channel holes, insights on ion scattering and mask interactions through feature-scale modelling is another crucial area of focus to improve the channel hole profile which deteriorates due to profile bowing and tapering with increased depth/aspect ratio. Understanding the evolution of profile (top CD, bow CD, and taper) continuously with time and how these non-ideal hole shapes are formed in a HAR etch is important for continued scaling of 3D NAND. Simulations provide valuable insight which helps in alleviating some of the hardware and process development challenges.

The Future Scaling of 3D NAND: Solutions

Without major innovations in tools such as deposition and etch, this evolution will struggle to improve the cost efficiency of the NAND FLASH storage products. Further, to enable scaling roadmap, the increase the number of processing steps and time will drive an increased investment in highly advanced deposition and etch tools and could cause the storage roadmap to slow down a bit.

To meet growing data demands driven by AI, memory vendors are finding smarter ways to pack more bits into tighter spaces, without sacrificing cost or too much performance. The industry is looking at innovations in equipment and manufacturing techniques to better address the above challenges. Some of the leading equipment manufacturers for 3D NAND manufacturing such as Lam Research have been working on co-optimizing advanced deposition and etch technologies with the memory vendors. For example, Lam's proprietary cryogenic technology, better etch chemistries and polymer management solutions helps to mitigate 3D NAND scaling challenges:

- Cryogenic Technology: Cryogenic etching is a process, initially developed in the 1980s, and is re-emerging as a method of dry etching. The etch process is carried out at a low on-wafer temperature, typically below 0°C, with the temperature control unit operating at even lower temperatures. Cryogenic etching helps increase adsorption of reactive species while limiting the lateral etch rate. Leveraging low-temperature benefits and different plasma chemistries to deliver increased high aspect ratio etch capability enhances the etching rate. Adoption of novel chemistries during low temperature process to improve circularity and sidewall roughness of the etched holes. This technology also reduces the overall environmental impact of the etching process.

The Cryogenic technology addresses the above challenges arising due to the traditional HAR process:

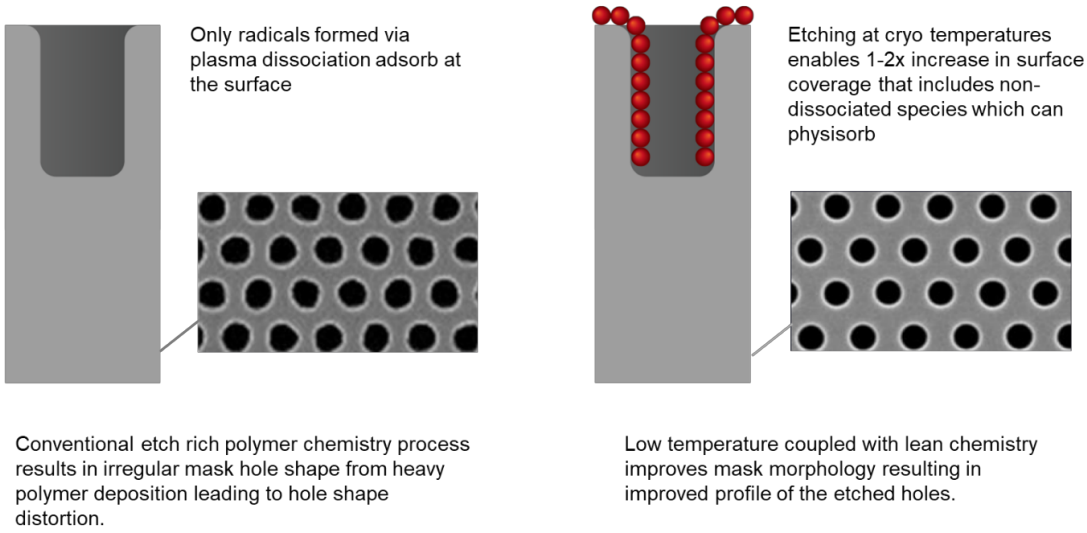
Solution 1: Enhancing Etch Rate

To mitigate ARDE, the transport of neutrals and ions must be enhanced, and this can be done either by increasing the ion energy and/or adjusting the plasma chemistry. Leveraging low-temperature benefits and different plasma chemistries through Lam's proprietary cryogenic etching technology helps in addressing the challenges of reduced etch rate with depth. Lam's pulsed power plasma technology utilizes increasing peak power at very short bursts drives ions much deeper with higher efficiency alleviating the challenges as the stack gets taller. Further, leveraging innovation in chemistry used for cryogenic etching is necessary to increase the etch rates for enabling taller structures.

Solution 2: Perfect Vertical Profiles - Reducing Variabilities

Adoption of leaner chemistry during low temperature process helps mitigate the clogging issue. In addition, adopting leaner chemistry during low temperature etching also improves mask morphology resulting in improved circularity and sidewall roughness of the etched holes. Polymer deposition on the sidewalls of the mask is the main root cause of sidewall roughness. This mechanism is suppressed for lean, low-temperature processes as the mask morphology and polymer deposition on the sidewalls of the mask are more consistent from hole to hole.

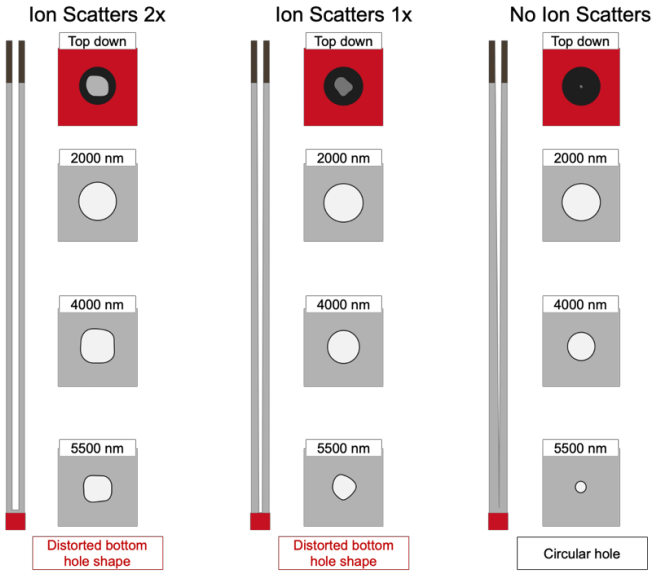
High Efficiency Cryogenic SiO_2 / SiN Etching



Sources: Lam Research, Counterpoint Research

Lam's latest feature scale modelling coupled with reactor-scale modelling helps identify the potential origins of feature hole distortions such as bowing, striation and twisting and provides a recipe to manufacturers to minimize the effect of aspect ratio dependence. Capturing hole shape distortion and twisting through feature scale models is a powerful tool for understanding the etch profile and processes. The feature-scale model data suggest that apart from mask shape and morphology, ion scattering within the hole causes hole distortions, however in the absence of hard mask evolution which is facilitated using low temperature and leaner chemistry as highlighted earlier, profile distortion can be minimized as shown below.

The feature profile model with a non-evolving hard mask suggests ion scattering within the hole, causing distortions. However, the profile distortion can be minimized in the absence of hard mask evolution, which is enabled using low temperature and leaner chemistry.

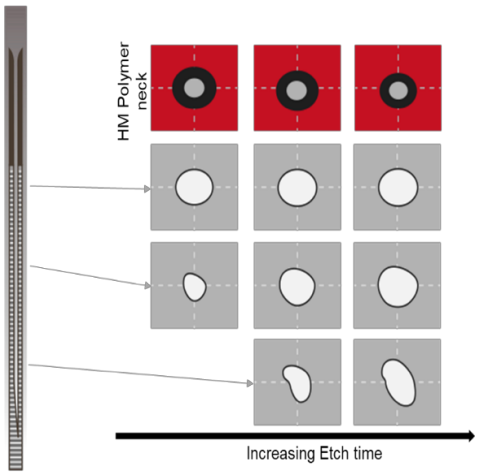


Sources: Lam Research, Counterpoint Research

To further understand the effects of hard mask evolution and the ARDE for the low-temperature etch in addition to ion scattering, a HAR ONON etch feature-scale model was developed. The simulation results show controlling hole shape distortion early in the process can mitigate hole distortion transfer deeper in the etch as shown below.

Feature profile simulator results modeling both hard mask (HM) evolution and the channel hole etch shows the time evolution of a distorted hole can become more circular with time at the same depth.

Leveraging the data obtained from feature scale model along with low temperature coupled with lean chemistry offers an additional window for improvement to NAND manufacturers unlike conventional etch process.



Sources: Lam Research, Counterpoint Research

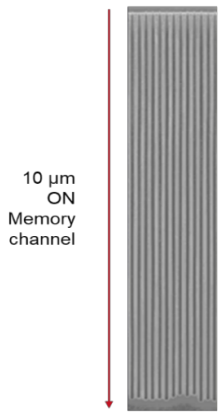
Leveraging low temperature and leaner chemistry restricts channel holes to exhibit minimum non-ideal hole behaviour such as twisting, roughness, and distortion at very HARs and this approach can be leveraged to enable more tiers for next-generation nodes.

The Future Scaling of 3D NAND: Enabling Roadmap for AI Era

Lam has developed this technology over the years and is a market leader in 3D NAND dielectric etching. With the introduction of cryogenic etching technology in 2019, Lam's installed base has grown to more than 7,500 chambers, with nearly 1,000 of these in production with cryogenic etch technology. This strong installed base gives Lam an experience of etching more than five million wafers using the cryogenic etch technology.

The latest innovation Lam Cryo™ 3.0 leverages the benefits of utilizing scalable high power and ion energy confined plasma reactors, unique pulsed plasma technology and temperatures as low as -60°C to efficiently etch with tight profile control. It enables deeper etching without compromising the feature shape. The repeatability of this highly precise, predictable etching process is a big differentiator for Lam as it helps NAND providers to achieve consistent output during production.

Lam Cryo™ 3.0 Performance



Parameter	Cryo™ 3.0	NAND scaling vector improved
Depth	10µm	**Vertical** *1.6x vs. current production*
Etch rate vs. non-cryo etch	2.5x faster	
Max CD	108 nm	**Lateral** *>10% cell density increase*
Bottom CD	99 nm	
ΔCD (Max-Bottom)	9 nm	**Logical** *>2x improvement in profile deviation*
Profile deviation ΔCD/Depth	<0.1%	

Source: Lam Research

- Predictably and repeatedly etch uniform memory channels as deep as 10 microns with less than 0.1% deviation in the channel's critical dimension from the top to the bottom.
- Enables etches 2.5x faster than conventional etch.
- The combination of speed, precision and profile control helps to enable high volume and high-quality yields.
- Delivers substantial carbon footprint reduction as compared to conventional etch processes including 40% reduction in energy consumption per wafer and a 90% reduction in emissions. *

*(Source: Lam Research. Up to a 90% reduction in Kg CO_2 per wafer. Estimated emissions reduction calculated using IPPC (Intergovernmental Panel on Climate Change) guidelines for greenhouse gas inventories. The estimated reduction has not been independently verified.)

Further, employing AI, Lam's feature-scale models also capture the physics of the ion species, transport effect as well as the surface-level chemistry dependences in feature profile evolution and utilizes the vast amount of information in improving the solution every generation. Basically, these trained AI models are built on years of telemetry data, customer feedback and experience help simulate perfect recipes in different scenarios to optimize the etching performance within a pristine environment for achieving deeper etch performance. Another benefit of the softer component is the ability for customers to seamless upgrade their existing infrastructure.

In addition to addressing the above benefits, Lam is also working closely in finding solutions for other approaches such as Word-Line Pitch Scaling, which is essential to enable more bits per device, Advanced Word-Line Metallization to enable thinner connections and faster devices and Wafer Stress Management, to control wafer bow enabling taller devices.

As the importance and capabilities of advanced compute (e.g. GPU, NPU) and DRAM (e.g. HBM) has increased, advancement of 3D NAND also remains critical to drive leading generative AI performance on the edge or the cloud. For example, the models which are stored in 3D NAND are regenerated and loaded into HBM DRAM, however, it is key to maintain those efficient and faster rewrite speeds between the two memory solutions. This is straightaway a function of the quality of 3D NAND based on the channel mobility which is dependent on the HAR channel etch profile. As a result, Lam Cryo™ 3.0 and the novel chemistries it enables are critical to have an all-round AI performance between compute, DRAM and 3D NAND. This will be a big differentiator for 3D NAND suppliers looking to succeed in the AI era.

Key Takeaways

- The demand for 3D NAND FLASH is likely to have significant growth in the next ten years and the industry is anticipating going up to 1,000-word-line (memory cell) layers would enable NAND FLASH die density to reach 100 Gbit/mm^2.

- Scaling 3D NAND beyond 400 layers will introduce significant challenges such as slow etch rate and variabilities in vertical profiles, which inhibit vertical scaling as more layers are added.

- This warrants consistent innovations in etch and deposition technologies from the equipment vendors and meaningful investments from NAND vendors to adopt those advanced technologies and techniques.

- 3D NAND manufacturing equipment vendors such as Lam Research are leading in this space, innovating with proprietary techniques that leverage cryogenic dielectric etch technology to enable NAND providers' roadmaps to scale to 1,000+ layers 3D NAND.

- Lam Research utilizes the most advanced hardware and process technology innovations, bringing significant improvements in HAR plasma etch systems.

- The third generation of Lam's cryogenic technology, advanced etch chemistries and polymer management, combined with other etch innovations, offer better control of channel hole shape from top to bottom. This capability is critical in very high aspect ratio etches and is the key to scaling the 3D NAND vertically.

- Thus, Lam's innovative solutions in etch technologies enable scaling across all the three vectors – lateral, vertical and logical – which can translate to significant benefits to the chipmakers.

- Further, this technology also reduces the overall environmental impact of the etching process.

Glossary

AI - Artificial Intelligence

ARDE - Aspect Ratio Dependent Etch

BICS - Bit Cost Scalable

CD - Critical Dimension

DRAM - Dynamic Random Access Memory

Gbit/mm^2 - Gigabit per millimetre square

GPU - Graphics Processing Unit

HAR - High-aspect Ratio

HBM - High Bandwidth Memory

NPU - Neural Processing Unit

ONON - Oxide Nitride stack

ON - Oxide Nitride

QLC - Qual-level Cell

SLC - Single-level Cell

V-NAND - Vertical-NAND

WL - Word-line

Authors, Copyright, User Agreement and Other General Information

Lead Authors

Dr Ashwath Rao
Senior Analyst
ashwath.rao@counterpointresearch.com

Neil Shah
Research, Vice President
neil@counterpointresearch.com

This research paper has been written in collaboration with Lam Research

COUNTERPOINT TECHNOLOGY MARKET RESEARCH
USA | UK | India | China | Taiwan | South Korea | Japan
info@counterpointresearch.com