

How KIOXIA Is Powering the Future of AI with NAND Innovation

Maitry Dholakia | Vice President, Memory Business Unit

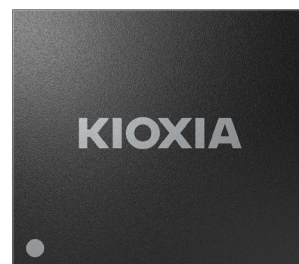
In my daily conversations with customers, AI has quickly moved from an emerging trend to a strategic priority. Our customers continue to push the boundaries of innovation and turn to KIOXIA to help power their vision. The questions we explore together reflect both the pace of change and the scale of opportunity in today's AI landscape. Below are some of the topics that come up most often – it's an exciting time as we explore the vast possibilities for our memory solutions.

Q: Why is NAND critical for AI?

A: To put it simply: no NAND means no scalable AI.

AI is compute-intensive - but above all, it's data-dependent. Training and running today's models requires massive datasets and constant iteration:

- Petabytes of training data
- Continuous dataset updates
- Checkpoints and model versioning
- Rapid access to inference data



All of this data needs a home. NAND flash in SSDs delivers high-density, cost-effective¹ storage capable of holding and streaming these enormous datasets.

In AI's early phase, demand focused on training large models - requiring heavy compute but relatively limited deployed storage. Today, AI has moved into high-speed inference at scale. Models are deployed widely across cloud, enterprise, and edge environments, driving sharp growth in storage requirements.

The current AI infrastructure buildout is driven directly by data growth. NAND flash enables this scalability by:

- Storing massive datasets
- Feeding GPUs and accelerators fast enough to avoid bottlenecks
- Making AI infrastructure economically viable
- Supporting both training and high-volume inference

Without high-density NAND flash, there's no practical way to store, access, and manage the data volumes AI demands.

Q: How is NAND better than traditional HDDs?

A: In AI environments, the advantages of NAND-based SSDs over HDDs are clearly seen in performance, power efficiency, and density.

SSDs deliver far lower latency and higher throughput, ensuring GPUs and accelerators are continuously fed with data, which is critical for AI workloads. They also consume less power per IOPS and per terabyte, helping data centers reduce energy use and cooling demands.

In terms of density, NAND enables significantly more capacity per rack unit. With QLC technology, even higher capacities can be delivered within the same footprint, improving scaling density per watt and per square foot.

As AI-driven data growth accelerates, QLC-based SSDs provide the performance, efficiency, and scalability needed - not only in hyperscale and enterprise data centers, but also across edge, client, and mobile applications.

Q: How, specifically, is KIOXIA addressing the demand for high density memory?

A: Having invented NAND flash memory more than 35 years ago, KIOXIA continues to advance the technology to meet evolving storage and performance requirements.

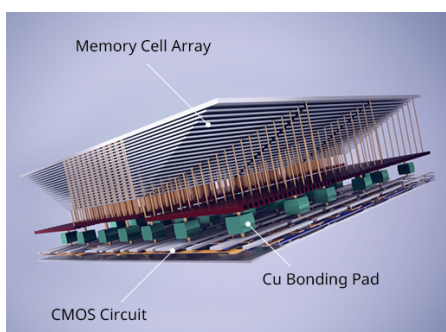
Our BiCS FLASH™ 3D flash memory architecture was developed to address the unceasing demand for improved density and performance. By stacking memory cells vertically, BiCS FLASH enables significantly higher capacities within a compact footprint.

In particular:

- **QLC-optimized solutions for data centers** improve power consumption, increase scaling density per watt, and deliver new levels of scalability for AI infrastructure.
- **UFS QLC solutions** are well suited for smartphones and tablets requiring high densities, and also support emerging categories such as PCs, networking, AR/VR, IoT, and AI-enabled devices.
- **UFS for mobile, PC, and automotive applications** enables AI functionality directly on devices, supporting fast, responsive user experiences.
- **XL-FLASH™ storage class memory (SCM)** is a low-latency NAND solution that fits between DRAM and traditional flash in the memory hierarchy. It offers higher capacity than DRAM with significantly lower latency than conventional NAND, making it well suited for hyperscale data centers, enterprise applications, and certain AI edge use cases that require fast, scalable storage.



Q: From a technology perspective what innovation/s did KIOXIA put into place to be ready to address this growth?



A: Anticipating the demand for higher density and scalable performance, we accelerated architectural innovation in our latest BiCS FLASH generations.

BiCS FLASH generation 8 introduced CMOS directly Bonded to Array (CBA), which separates the CMOS circuitry from the memory cell array and bonds them together. This enables more efficient lateral bit scaling, improving density, performance, and power efficiency while providing manufacturing flexibility.

In simple terms, CBA allows us to increase storage capacity without sacrificing speed or energy efficiency. This architecture positioned KIOXIA to meet the rapid growth in AI-driven storage demand with cost-effective, scalable solutions.

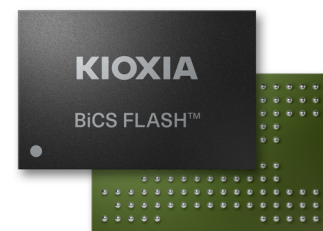
BiCS FLASH generation 8 is particularly well suited for customers whose storage requirements are being stressed by the explosion of data-intensive applications and AI capabilities.

Q: What does the future look like when it comes to AI and what is KIOXIA doing now to be ready to support it?

A: AI will continue expanding across cloud, enterprise, and edge environments. Models will grow more sophisticated, datasets will expand, and inference will move closer to end users and devices. All of this will increase demand for both higher capacity and higher performance storage.

KIOXIA is advancing a dual-axis development strategy to address these needs efficiently:

- **BiCS FLASH generation 9 products** leverage CBA technology to integrate proven memory cell technologies with the latest CMOS advancements. These products are designed to deliver high performance and strong power efficiency in low- to mid-level storage capacities. They will be integrated into enterprise SSDs, particularly those designed to maximize GPU efficiency in AI systems.
- **BiCS FLASH generation 10 products** expand the number of memory layers to address future demand for even larger-capacity, high-performance solutions - especially in hyperscale and data center environments.



Each generation builds on prior advancements, enabling optimized solutions for specific use cases while supporting efficient capital investment strategies.

Q: How is the expansion of AI use cases impacting memory needs?



A: AI at the edge is becoming increasingly important. Storing smaller AI models directly on devices - PCs, smartphones, tablets, or IoT systems - allows AI functionality to operate without a constant internet connection. This lowers latency, speeds up responses, protects user privacy, and reduces network load. As more AI processing happens locally, demand for high-density, power-efficient NAND in client and edge devices will grow.

The range of AI applications continues to expand, and all of them rely on memory.

From hyperscale data centers running large AI models, to enterprise systems enabling smart manufacturing and cybersecurity, to edge devices and AI-enabled PCs operating with low latency and greater privacy, every deployment depends on fast, scalable storage. As inference becomes more distributed and models move closer to the user, the need for high-density, power-efficient NAND only increases.

AI's future will be defined by data and NAND flash from KIOXIA will remain foundational to making that future possible.

Notes:

©2026 KIOXIA America, Inc. All rights reserved.

1: Cost effectiveness claim based on storing more bits per chip (3D stacking + multi-bit cells) and mass production efficiencies, which together lower the \$/GB of storage.

The views and opinions expressed in this blog are those of the author(s) and do not necessarily reflect those of KIOXIA America, Inc. All other company names, product names and service names may be trademarks of their respective companies.