Published Mar 20, 2018 • 4 min read

Project Denali to define flexible SSDs for cloud-scale applications

By Laura Caulfield, Senior Software Engineer, Azure Storage

SHARE

Thought leadership more

Last September, at SNIA's Storage Developer's Conference, I presented a prototype of the Project Denali SSD. Project Denali drives provide the flexibility needed to optimize for the workloads of a wide variety of cloud applications, the simplicity to keep pace with rapid innovations in NAND flash memory and application design, and the scale required for multitenant hardware that is so common in the cloud.

This month, I'm attending the Open Compute Project (OCP) U.S. Summit 2018 to begin formalization of the specification that will define the interface to Project Denali drives. Once in place, the specification will allow both hardware vendors and cloud providers to build and release their final products. The specification defines a new abstraction, which separates the roles of managing NAND, and managing data placement. The former will remain in the hardware – close to the NAND and in the product that reinvents itself with every new generation of NAND. The latter, once separated from the NAND management algorithms, will be allowed to follow its own schedule for innovation, and won't be prone to bugs introduced by product cycles that track solely with NAND generations.

The specification, with the refactored algorithms it requires, will give a renewed boost to innovation in both drive firmware and host software. Throughout the development of the Project Denali architecture, we've focused on the following four goals:

1. Flexible architecture for innovation agility: Workload-specific optimizations, FTL managed as cloud services component
2. Rapid enablement of new NAND generations: NAND follows Moore's Law; SSDs: hours to precondition, hundreds of workloads
3. Support a broad set of applications on massively shared devices: Azure (>600 services), Bing, Exchange, O365, others; up to hundreds of users per drive
4. Scale requires multi-vendor support & supply chain diversity: Azure operates in 38 regions globally, more than any other cloud provider

Addressing a growing divide

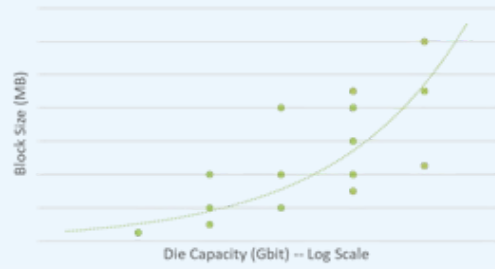
One of the primary drivers behind the Project Denali architecture is the mismatch of write and reclaim size between the hardware and the application. Cloud-scale applications tend to scale out the number of tenants on a machine, whereas the hardware scales up the size of its architectural parameters. As the number of cores in a server increases, a single machine can support more VMs. When storage servers increase their capacity, they typically increase the number of tenants using each as a back-end. While there are some notable exceptions, there is still a need for cloud hardware to provide enough flexibility to efficiently serve these multi-tenant designs.

SSDs' caching further increases this divide. To make the best use of the SSD's parallelism, its controller collects enough writes to fill one flash page in every plane, about 4MB in our current drives. If the data comes from multiple applications, the drive mixes their data in a way that is difficult to predict or control. This becomes a problem when the device performs garbage collection (GC). By this time, some applications may have freed their data, or updated it at a faster rate than other applications. For efficient GC, the ideal is to free or update the data on a single

Radian Memory Systems, LLC - Exhibit 2006
Page 2006 - 1
IPR2025-01289, Samsung Elecs. Co., Ltd. v. Radian Memory Sys., LLC

block. But when caching brings the effective block size to 1GB, it's very unlikely that the host can issue enough data to the drive for one tenant before servicing writes for the next tenant.

This caching design can also be viewed as an optimization for a single workload: sequential writes. It's not surprising for SSD makers to target this workload – optimizations for HDDs aim to make workloads sequential, and sequential writes on SSDs are consistently 4-5x faster than a workload of random writes (on a completely full drive with 7% OP). However, real-world cloud applications rarely consist of writes that are purely sequential on the whole drive. Much more common are workloads consisting of hundreds of sequential write streams per TB of storage.



The final nail in the coffin is in how flash block sizes change over time. Advances in the density of NAND flash typically increase the size of the block. For this reason, it will be even more important for the host to have access to place its data on the native block size in future designs.

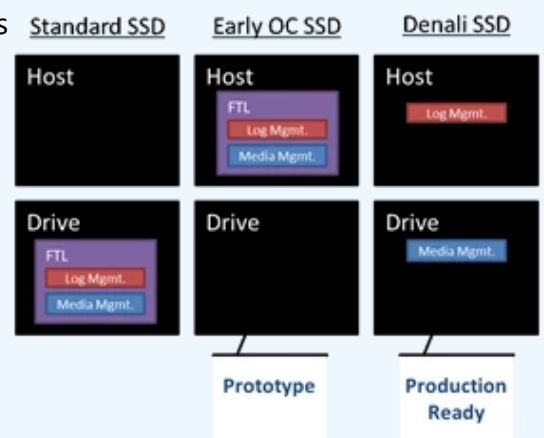
Flexible architecture for cloud hardware

To address the challenges described above, we took a closer look at the algorithms in SSD's flash translation layers, and in cloud applications and found the following two categories:

Log managers receive random writes and are responsible for translating these to one or more streams of sequential writes, and emitting these to the next layer down in the storage stack. To do this, log managers typically maintain an address map and must perform garbage collection.

Media managers are written for a specific generation of media, and implement algorithms such as ECC, read retry and data refresh in accordance with the physical properties of the media such as the patterns of where and when errors arise and data retention times.

These two components appear in many places. Standard SSDs mix both, and early open channel architectures require the host to implement both. Most systems have at least two or three log managers stacked on top of each other – a scenario SanDisk warned against in their paper, Don't Stack Your Log on My Log. In Project Denali, we propose a clear division between log and media managers, with placement of the media manager in the drive, and a single log manager encapsulated entirely in the host.



Successful prototype

In the first phase of our development, we collaborated with CNEX to build a prototype system. While the interface change opens up opportunities to optimize across many layers of the storage stack, we modified only two components: the firmware and the lowest level device driver in Azure's OS. This allowed for a quick evaluation of the ideas, provides infrastructure for legacy applications and sets up the system for future optimizations.

The pre-optimization results were better than we expected. The memory, write amplification and CPU overheads that are typically in the drive moved to the host (as expected), and the system's throughput and latency were slightly better than standard SSDs.

We look forward to finalizing the Denali specification in the months ahead through the Denali group and plan to make this specification available broadly later this year. Refactoring the flash translation layer will promote innovation across the storage stack, and I invite you to learn more about the new interface.

Related Posts

Radian Memory Systems, LLC - Exhibit 2006
Page 2006 - 2
IPR2025-01289, Samsung Elecs. Co., Ltd. v. Radian Memory Sys., LLC

Radian Memory Systems, LLC - Exhibit 2006
Page 2006 - 3
IPR2025-01289, Samsung Elecs. Co., Ltd. v. Radian Memory Sys., LLC

Explore
Azure AI Foundry

The future of AI starts here. Envision your next great AI app with the latest technologies. Get started with Azure.

Learn more about Azure AI Foundry

Connect with us on social

Explore Azure
- What is Azure?
- Get started with Azure
- Global infrastructure
- Datacenter regions
- Trust your cloud
- Azure Essentials
- Customer stories

Products and pricing
- Products
- Azure pricing
- Free Azure services
- Flexible purchase options
- FinOps on Azure
- Optimize your costs

Solutions and support
- Solutions
- Resources for accelerating growth
- Solution architectures
- Support
- Azure demo and live Q&A

Partners
- Azure Marketplace
- Find a partner
- Join ISV Success

Resources
- Documentation
- Blog
- Developer resources
- Students
- Events and Webinars
- Analyst reports, white papers, and e-books
- Videos

Cloud computing
- What is cloud computing?
- What is multicloud?
- What is machine learning?
- What is deep learning?
- What is AIaaS?
- What are LLMs?
- What are SLMs?
- What is RAG?

English (United States) Your Privacy Choices Consumer Health Privacy

Sitemap Contact Microsoft Privacy Terms of use Trademarks Safety & eco Recycling About our ads © Microsoft 2025