SEPTEMBER 2020

NUTANIX HCI PERFORMANCE EVOLUTION

Blockstore and SPDK

NUTANIX
YOUR ENTERPRISE CLOUD

Introduction

Blockstore in Nutanix AOS represents the latest improvements in a series of developments and growth to the core architecture of the first hyperconverged infrastructure operating system. It enables higher performance and scale and builds capabilities that will result in rapid adoption of new storage technologies, from SSD; NVMe; Intel Optane; to persistent and non-persistent memory architectures. This paper discusses the path to Blockstore, the underlying architectural components, and the scope of improvements it provides.

BACKGROUND

Nutanix is the first mover in Hyper Converged Infrastructure Software. Starting in 2009 with the introduction of a fully virtualized storage controller that runs on a hypervisor and eliminates the compute-network-monolithic storage hardware stack, Nutanix has evolved HCI from a solution for VDI environments to a solution capable of handling the highest performance and most demanding data intensive workloads. This evolution started with the technical decision to include fine-grained metadata that enables easy movement of data during failure scenarios and migration events. This provides resiliency and predictable scalability in performance and is part of the core Nutanix AOS architecture that keeps data local to the compute workload, rather than having to extend over a network hop for access providing data locality.

Building on those core tenets in AOS 5.11, Nutanix AOS introduced an Autonomous Extent Store(AES). This innovation extends the concept of data locality that only applied to application data before to include metadata as well. Instead of a single distributed key-value store, AES keeps the logical metadata in the distributed key-value store while physical metadata is stored locally with user data. This improves performance for sustained write workloads and set up AOS for the introduction of Blockstore.

Historically in servers or storage systems, storage devices have been orders of magnitude slower than CPU and RAM and operating system (OS) developers have handled this by the use of interrupts when applications are accessing storage. An interrupt from an application essentially submits a request to the OS kernel and then waits for the kernel to signal that the operation is complete, which adds an overhead in latency as the OS is responsible for a lot more than just storage. This latency overhead was insignificant compared to the significantly higher latency of the storage devices themselves. Further, when an application submits a storage request to the kernel, the storage data needs to be copied between the application memory space and the kernel memory space, which takes time and consumes valuable CPU resources.

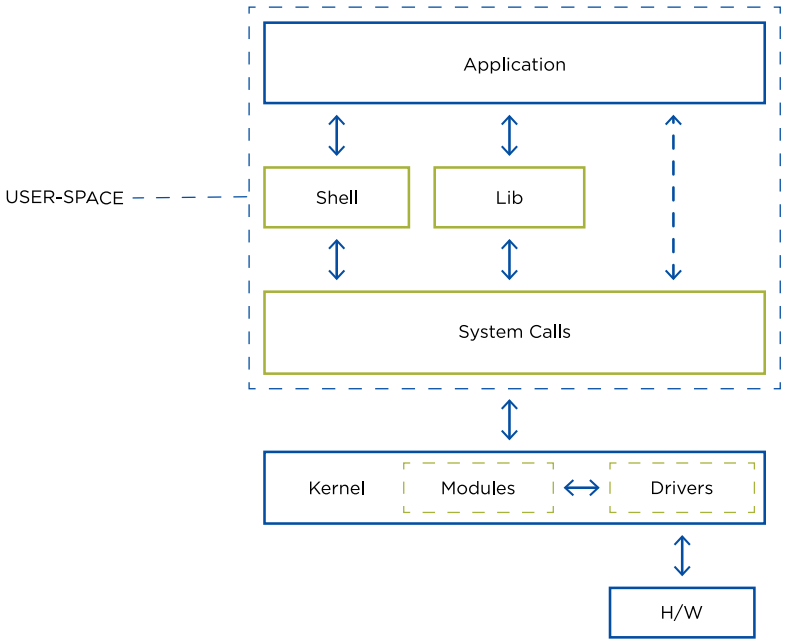

Figure 1: User and Kernel Space Interaction. Source: Nutanix Bible

Technological advances have now made physical storage much faster with the advent of SSDs which store data in circuits instead of on spinning magnetic disks. The NVMe protocol unlocks the full capabilities of SSDs by enabling them to connect directly to the PCIe bus, increasing the potential performance of not just a single drive but of the entire system. The performance capabilities of NVMe SSDs amplify the CPU cost associated with traditional storage access mechanisms and keeps new storage technologies from fully realizing their performance and scalability characteristics. The dominating factors contributing to latency in accessing Flash Based SSD and NVMe storage are in-system calls and kernel overhead, the legacy of spinning media and interrupt driven storage.

BLOCKSTORE OVERVIEW

Nutanix AOS runs in a Linux based virtual machine on every node in a cluster, called a Controller Virtual Machine (CVM). To perform IO operations, Stargate, the main Nutanix process that manages data, needs to invoke system calls to the kernel filesystem and the block sub-system (units of application data).

AOS manages its own metadata. Stargate stores cluster-wide metadata in a distributed key-value store using a modified Cassandra database and local metadata in a high-performance key-value store utilizing RocksDB. Stargate does not need the additional full capabilities of Linux based filesystem for managing metadata and physical storage devices.

Nutanix Blockstore is a free space manager that controls space on physical devices at a block-level granularity, where blocks are the minimum units of allocation and deallocation. With Blockstore, the filesystem is removed from the kernel completely into application user space. This allows an application, in this case Stargate, to avoid the overhead of system calls and interrupts when accessing data from Flash Devices, and it also prevents memory copies needed for a kernel based file system. AOS has moved beyond the physical, spinning disk paradigm to a new one that is more appropriate for Flash media.

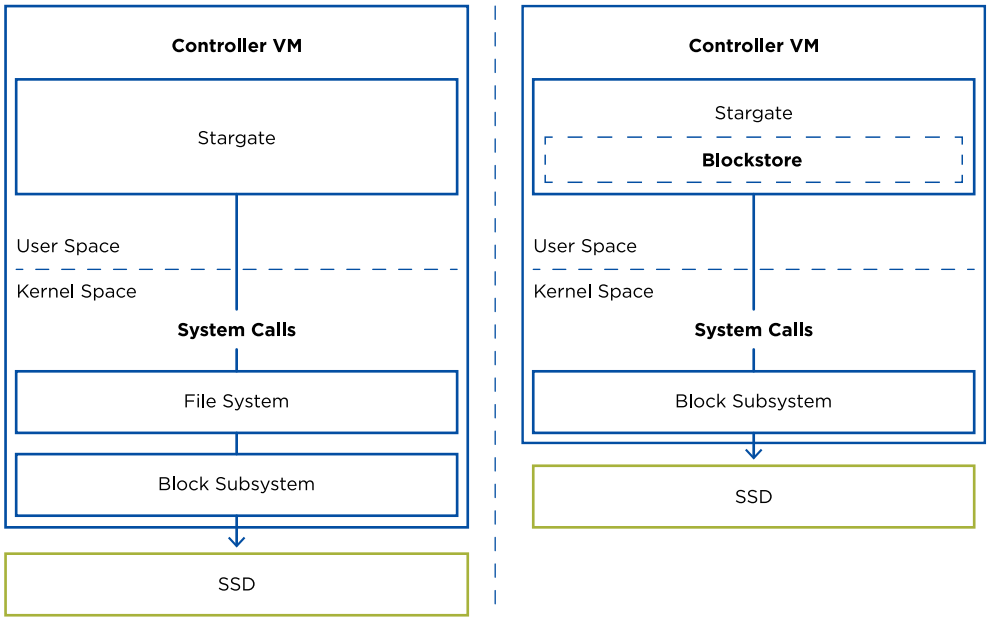

Figure 2: Controller VM Data path with Blockstore

BLOCKSTORE ARCHITECTURE

Blockstore's layered design delivers numerous benefits to Stargate. Blockstore is backed by abstracted linear address space called the backing store. The backing store can be a physical device like HDD, SDD, NVMe, Optane or a non-persistent linearly addressable device like memory or a virtual storage device such as vDisk. The design allows flexibility and the adoption of new technologies.

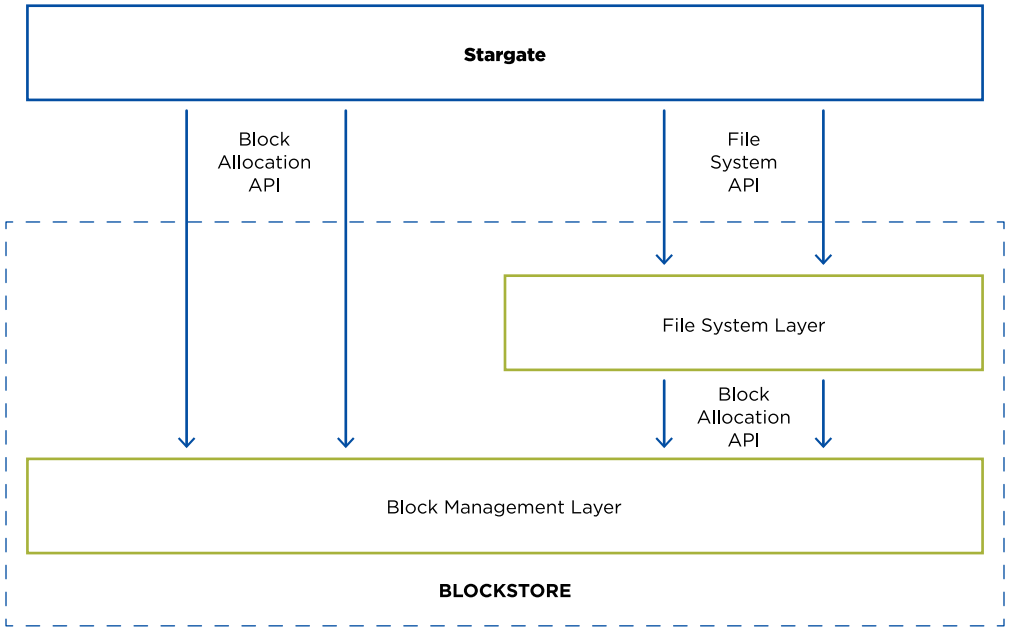

Figure 3: Blockstore Architecture

Blockstore has two different addressable layers:

1. The Block Management Layer

The block management layer is responsible for allocation and deallocation of blocks. The minimum block size that the block store supports is provided as an argument to the block store format operation. Stargate data paths that do not need filesystem capabilities now talk directly to block management layer.

2. The File System Layer

The file system layer sits on top of the block management layer and provides a POSIX-like file system API to stargate. This layer is intended for use by those parts of stargate that need filesystem semantics decoupling from parts that do not need it optimizing performance.

SPDK OVERVIEW

In addition to the benefits of Blockstore, which removes system calls for lookups, it also gives the ability to leverage libraries and APIs which allow access to storage devices directly from user space. One such library is Storage Performance Development Kit (SPDK) developed by Intel. SPDK allows a user space process to access an NVMe device directly. Nutanix implemented SPDK to build on the benefits of having Blockstore. SPDK architecture further improves performance with three key optimizations:

- Avoid interrupts by using polling that helps in reducing variance of tail latencies for applications.
- Avoid kernel locks as it is lockless, so performance scales linearly with the numbers of NVMe devices.
- Avoid system calls as the filesystem can now directly access the storage device from user space making the IO data path efficient.

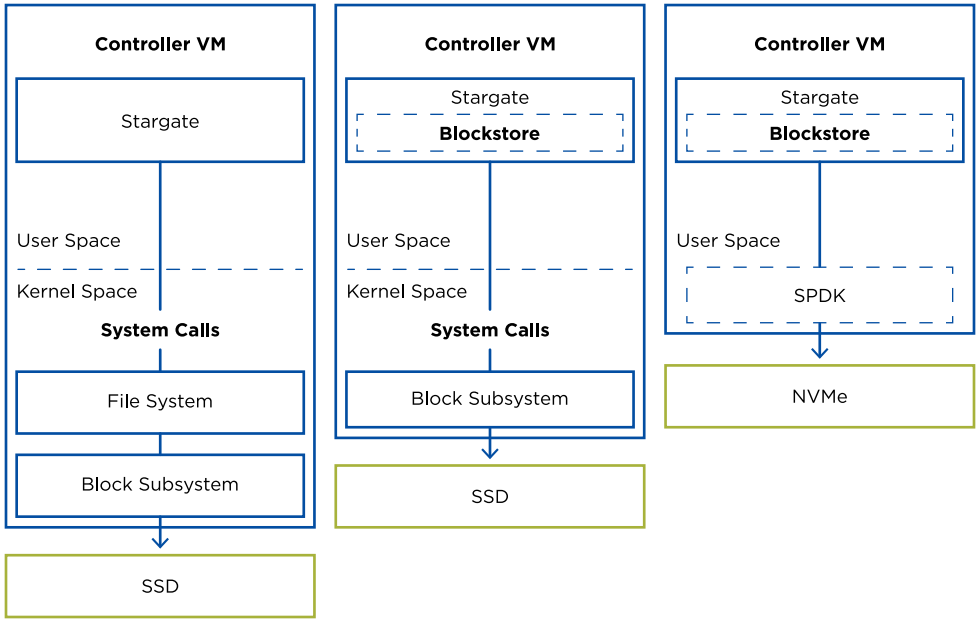

Figure 4: Controller VM Data path with Blockstore and SPDK

Blockstore makes it efficient to do storage and block allocation in user space and SPDK allows Nutanix AOS to access devices directly from user space, fully utilizing the capabilities and performance of NVMe based storage devices.

PERFORMANCE IMPROVEMENTS

Nutanix AOS was architected with performance at scale in mind. Fine-grained metadata helps in effective tiering and movement of data on the node and across cluster. Data locality ensures applications and user VMs are served data locally from within the node without a network hop. With AES, in addition to data locality, it also now provides metadata locality which helps in improving sustained write performance.

Blockstore builds on these key architectural tenets by removing the OS kernel bottleneck, enabling efficient storage operations and high performance, and SPDK maximizes the capabilities of NVMe and Optane devices. ESG has published Nutanix Architecture and Performance Optimization whitepaper which describes the performance benefits in detail and the impact on application performance.

These are the performance takeaways of using Blockstore and SPDK:

- Blockstore and SPDK improved and optimized performance for AOS by an average of ~20%-25% for read intensive workloads on the same hardware.
- Latencies and response times improved with Blockstore and SPDK which are crucial for Tier-1 workloads like databases and medical records applications.
- Blockstore and SPDK also resulted in lower and consistent tail latencies in medical records workloads vs pre-Blockstore, which meant there was less variance in performance on the same hardware.
- A cluster with Blockstore and SPDK with NVMe based devices gave ~60% better IOPS and ~45% lower latency compared to an All-Flash cluster without Blockstore and SPDK running SATA based SSDs.

CONCLUSION

Starting with data locality and fine-grained metadata, continuing with AES, Nutanix AOS delivers further architectural advancements in software with Blockstore. With Blockstore, Nutanix AOS is uniquely positioned to leverage the benefits of libraries like SPDK and storage hardware technologies like NVMe and Optane.

References

- The Nutanix Bible
- Nutanix BlockStore Performance Paper, Enterprise Strategy Group
- Storage Performance Development Kit

Nutanix makes infrastructure invisible, elevating IT to focus on the applications and services that power their business. The Nutanix Enterprise Cloud OS leverages web-scale engineering and consumer-grade design to natively converge compute, virtualization, and storage into a resilient, software-defined solution with rich machine intelligence. The result is predictable performance, cloud-like infrastructure consumption, robust security, and seamless application mobility for a broad range of enterprise applications. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

NUTANIX™
Your Enterprise Cloud Platform

info@nutanix.com | www.nutanix.com | @nutanix

LIST OF TABLES

LIST OF FIGURES