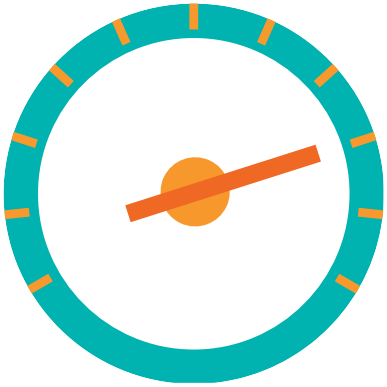
WHITE PAPER

Introducing BlueStore

Exploring an Exciting New Feature of SUSE Enterprise Storage 5

suse.com

SUSE®



With every major release, the Ceph storage system gets better, faster, and more capable. A new Ceph feature included with SUSE Enterprise Storage 5 provides inline compression and improves throughput by up to 200%.

The Ceph storage system is a highly efficient and cost-effective tool for safely storing large amounts of data. Ceph has been called “valet parking for data.” A storage client saves the data to the cluster and is then free to turn to other tasks, while, behind the scenes, the cluster snaps into action. Ceph uses a hash algorithm to distribute the data in a fault-tolerant manner to Object Storage Daemons (OSD) within the cluster.

The Ceph storage process requires several steps, but ultimately, the data must find its way to a disk or other permanent storage medium. Although Ceph was originally intended for direct-to-disk storage, recent versions have relied on a component called FileStore to save data to POSIX-compatible filesystems such as XFS, Btrfs, and Ext4.

Using a conventional Linux filesystem on the Ceph backend provided some benefits, however, it also had a cost. The principal issue with using a POSIX-compatible filesystem with Ceph is performance. XFS, Btrfs, and Ext4 were designed for versatility and general use, and they include unnecessary features that weigh down performance.

The most significant issue is that the storage process for a POSIX-compliant journaling filesystem requires that data and metadata be written twice for each write operation, once to the journal and once to the final data structure on the device.

Because the fault tolerance and isolation built directly into Ceph eliminate the need for this additional write operation, the double write does not provide value.

An innovative new component called BlueStore was introduced to address the performance issues associated with using a POSIX journaling filesystem on the Ceph backend. BlueStore, which was created specifically for Ceph, performs the functions necessary for its role within the Ceph cluster without the overhead of unnecessary features and guarantees.

BlueStore operates close to the hardware, thus eliminating the abstraction layers associated with POSIX filesystem operations. As you will learn later in this paper, the result is a significant improvement in storage performance. BlueStore also adds other improvements, including inline data compression and full data checksums.

BlueStore is one reason why SUSE Enterprise Storage 5 is faster, easier to manage, and easier to monitor than any previous version. Some experts predict that BlueStore will improve write speeds by a factor of two over previous versions, although the actual number could vary depending on your hardware and your Ceph configuration.

Inside BlueStore

BlueStore operates at a layer below the OSD, receiving data and writing it to a block storage device (Figure 1). BlueStore writes the data directly to the device, thus eliminating the overhead of a filesystem.

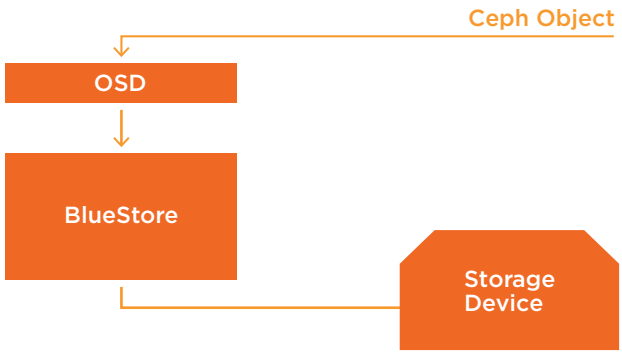


Figure 1.
BlueStore operates at a layer between the OSD and the storage device, receiving data from the OSD and saving it directly to disk.

Ceph makes extensive use of metadata for searching, managing, and reporting on data stored within the cluster. The metadata associated with a Ceph object is a collection of key-value pairs that locate and describe the data. Although the data and metadata are associated with the same object, they take different forms and are used in different ways. One of the innovations of BlueStore is that it uses a separate path for storing metadata. When the data is received and written to disk, the metadata associated with that data is written to a RocksDB structured database (Figure 2). RocksDB is an embeddable, persistent key-value store designed for fast storage. Keeping the metadata in a structured, easily accessible form optimizes search and statistical operations for the cluster. The RocksDB database uses the BlueRocksEnv wrapper to store data to the BlueFS filesystem. BlueFS is a very simple filesystem specifically created for use with BlueStore metadata.

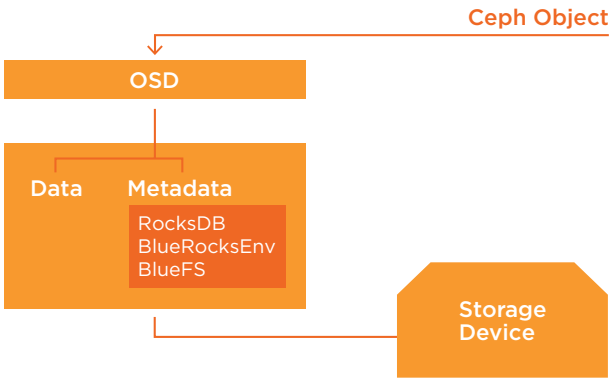


Figure 2.
Inside BlueStore: Metadata lands in a RocksDB structured database and is then saved to the Ceph-optimized BlueFS filesystem.

Separating the metadata from the object data leads to interesting scenarios for multi-disk configurations. For instance, you could use a large and inexpensive conventional disk for the object data and store the RocksDB metadata on a faster but smaller SSD device. Or, you might want to store the main RocksDB database on the big device and reserve the small SSD system for the RocksDB write ahead log (WAL). The SUSE technical support staff can help you determine the optimum storage device configuration for your network.

Benefits

Tests reveal that adopting BlueStore can result in an improvement of up to 200% throughput for write operations on the same hardware, and mixed read-write throughput can improve 70-100% depending on IO size.

BlueStore provides other advantages that will benefit many Ceph configurations. One interesting new feature is support for inline compression. One of the reasons for using Ceph is to maximize the efficiency of the storage infrastructure. Integrating data compression brings additional efficiency by increasing the total amount of data that can be stored within the cluster. Tunable parameters set the max and min compaction settings to help you find the balance point between latency, disk usage, and read amplification. A built-in garbage collection feature helps to minimize wasted space.

Another important benefit of BlueStore is its support for partial writes for erasure-coded pools. Erasure coding is a fault-tolerance technique that often provides a more efficient use of storage space than conventional RAID-based methods. Past versions of Ceph only supported erasure coding for full-object writes and appends, which limited the use of erasure coding to object interfaces like the RGW interface, unless the cluster employed a cache tier. With BlueStore, you can now use erasure coding directly with RBD/iSCSI block storage and the CephFS filesystem without cache tiering. A cache tier is still recommended in some scenarios for performance reasons.

Next Steps

SUSE recommends BlueStore for most Ceph use cases.

BlueStore is utilized automatically with SUSE Enterprise Storage 5. If you are setting up a new Ceph cluster with SUSE Enterprise Store 5, BlueStore is enabled by default, so you won't have to take any additional steps to install and configure BlueStore.

The BlueStore system operates at a low level within the Ceph cluster, and there are few reasons to interact with it directly. The one case when you might need to work directly with BlueStore is when you are migrating from an earlier, FileStore-based Ceph cluster to a BlueStore-based cluster. SUSE Enterprise Storage 5 comes with a migration tool that will help you manage the process of moving to BlueStore. The migration tool evacuates the old disk and reprovisions it for BlueStore, providing continuous operation with no downtime or even reduced redundancy. See the *SUSE 5 Enterprise Storage Administration and Deployment Guide* for more on the SUSE Enterprise Storage 5 migration tool.

Conclusion

The advanced Ceph storage system is evolving rapidly, and each release brings new features and capabilities. The BlueStore backend storage component included with SUSE Enterprise Storage 5 brings a vast improvement to storage performance by delivering direct-to-disk block storage for Ceph OSDs.

If you're setting up a new Ceph cluster with SUSE Enterprise Storage 5, BlueStore is enabled by default. If you're upgrading a cluster from a previous version of Ceph, the migration tool included with SUSE Enterprise Storage 5 will manage the process, allowing seamless migration with no downtime.

For more information, contact your local SUSE Solutions Provider, visit us online or call SUSE at:

1-800-796-3700 (U.S. and Canada)
1-801-861-4500 (Worldwide)

SUSE
Maxfeldstrasse 5
90409 Nuremberg
Germany

www.suse.com