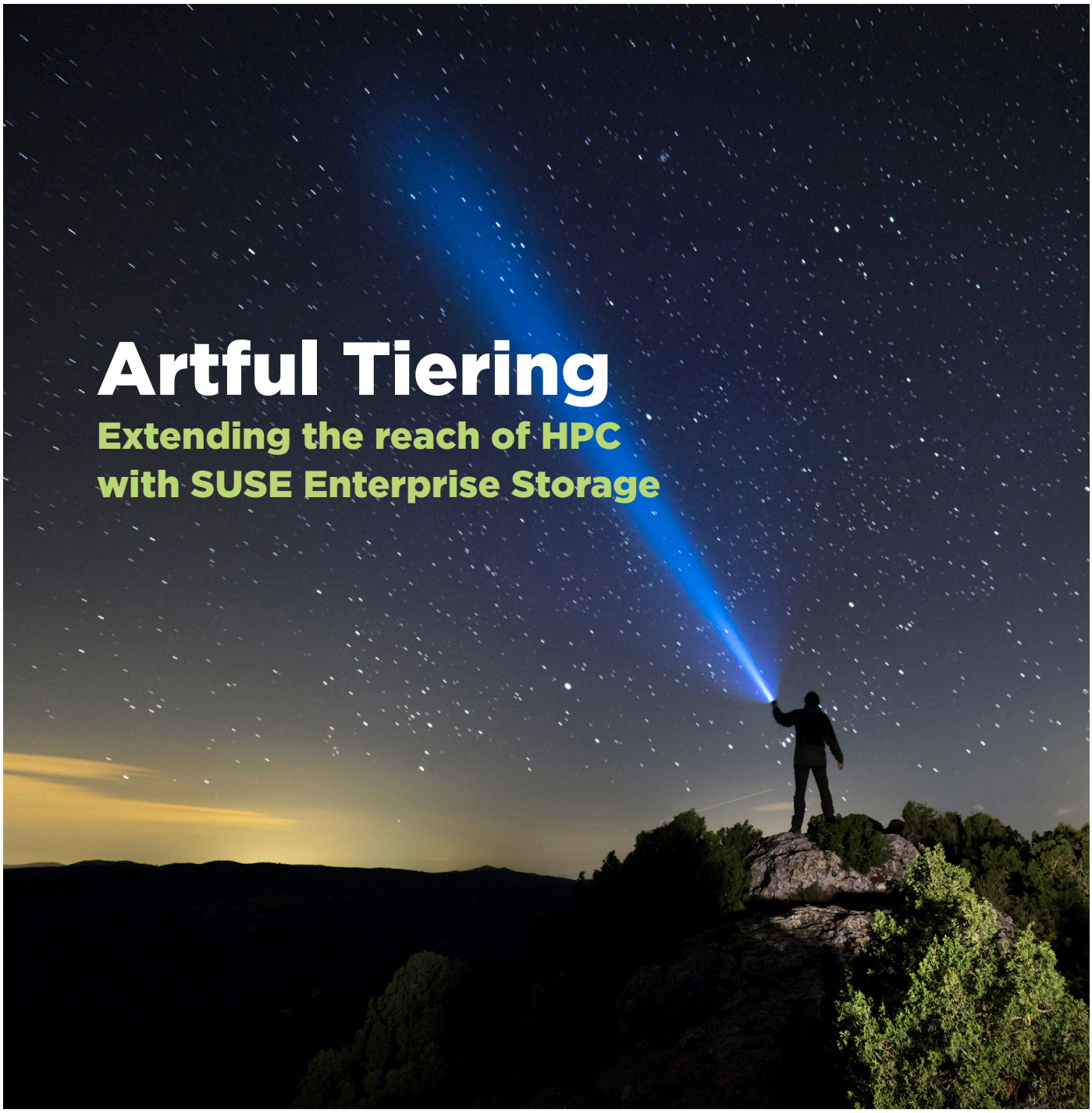
WHITE PAPER



Artful Tiering

Extending the reach of HPC with SUSE Enterprise Storage

suse.com

SUSE

Everything about the high-performance computing environment is big: hundreds or thousands of processors generating millions of CPU requests per second. Where does this data come from and where does it go? When you work in the HPC space, it doesn't take long to realize that data storage is an important component in achieving high performance at a massive scale.

Does faster storage mean a faster cluster? Maybe...but it all depends on the system requirements. Most HPC systems are designed for an overall budget, and storage is just one of several components. If you spend more than you need to on high-priced storage, you'll have less to spend on the other components -- the processors, switches, and high-speed network fabric -- you'll need to optimize the system.

Most experts believe the best way to maximize overall performance is to optimize storage design: provide low-latency, high-performance storage for the parts of the system that require it, and strive for cost efficiency in the parts of the system that don't need a high-priced performance boost.

This desire to maximize storage performance by optimizing storage efficiency has led to the rise of Ceph as a powerful storage option for the HPC space. Ceph is a fault-tolerant, self-healing, and self-managing software-defined storage solution that keeps the overall cost to a minimum and easily scales too very large data sets.

If your design benefits from splitting the storage by latency, or even if you would prefer to consolidate your storage resources within a single technology, Ceph can fit gracefully into your HPC solution. With an innovative subscription model and enterprise-grade technical support, SUSE Enterprise Storage provides best-in-class Ceph storage efficiency for many HPC workloads. This paper offers some ideas on how to integrate SUSE Enterprise Storage into your HPC storage environment.

Storage Savings with SUSE Enterprise Storage

SUSE Enterprise Storage is well known within the HPC space for its remarkably low cost of ownership. Several factors contribute to the minimal per-GB cost of SUSE Enterprise Storage, including:

- **Low acquisition cost** – SUSE Enterprise Storage runs on commodity hardware, without the need for a proprietary appliance or custom device. Ceph and other software components included with SUSE Enterprise Storage are open source.
- **Low administration cost** – SUSE Enterprise Storage is self-healing and self-managing. If a node goes down, the cluster re-balances and adapts automatically. Built-in fault tolerance protects the system from data loss, and a powerful set of management tools allows easy oversight for the cluster. A single Ceph admin can manage up to 4 Petabytes of data – up to 8 times more than the management capacity of a single admin in a block storage environment.
- **Low upgrade cost** – SUSE Enterprise Storage is easy to expand and upgrade. Add another server to the cluster, and the network rebalances with minimal reconfiguration.
- **Low subscription cost** – SUSE Enterprise Storage has an innovative per-server subscription model that eliminates capacity-based charges. Unlike other enterprise-grade Ceph solutions, you won't pay more money just because your storage needs expand.

The combination of these factors means that SUSE Enterprise Storage can deliver software-defined storage at 50% of the cost of comparable solutions while still maintaining the scalability and reliability of an enterprise solution.

Understanding Tiered Storage

Organizing data into storage tiers optimizes storage by keeping cost to a minimum yet offering elevated performance for mission-critical tasks. Many models for storage tiering exist, and the definitions of the storage tiers can vary depending on industry and application. Tier 1 is the name typically assigned to low-latency data that requires fast or frequent access and is thus an important candidate for optimization. All HPC workloads are different, so the details vary, but examples could include scratch space or application data that is likely to undergo frequent access or update.

The strict definition of Tier 2 data can also vary depending on the environment and workload, but you can think of Tier 2 as data that can reside at a higher latency without causing a bottleneck for the application. Examples could include data that is accessed infrequently or that is used with non-critical threads within an HPC application.

Some discussions refer to a lower, Tier 3 level of storage for near-line or infrequently accessed data, however, the distinction between Tier 2 and 3 is unnecessary with SUSE Enterprise Storage, which offers Tier 2 performance at a price point more often associated with Tier 3.

Ceph: Flexible by Design

SUSE Enterprise Storage is designed to fit easily into a tiered storage scenario. The underlying Ceph storage environment lets you organize your cluster into storage pools. A pool is served by several object storage daemons on different servers running within the cluster, but you can organize your cluster so that the data saved to the pool lands on a specific type of storage device (Figure 1). This design creates a seamless and convenient way to channel different types of data to different types of storage hardware – without sacrificing the flexibility and low cost of a clustered, software-defined storage environment.

Unlike some storage products that run on proprietary appliances and other closed systems, SUSE Enterprise Storage allows nearly unlimited flexibility in how you shape and organize the cluster. In some cases, the scenario depicted in Figure 1 could lead to a single-source multi-tiered storage environment all managed within Ceph. In other environments, Ceph's storage pools can optimize and add granularity within the Tier 2 space.

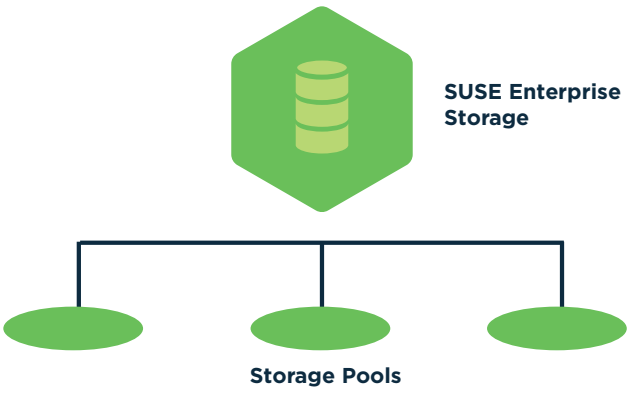


Figure 1: Ceph lets you organize the storage environment into storage pools.

Tiering for Better Performance

Figure 2 shows a single-source HPC storage environment, in which SUSE Enterprise Storage provides all storage services for the HPC cluster. Because SUSE Enterprise Storage offers the lowest per-GB storage cost for any equivalent storage alternative, this option is a strong contender for scenarios where it provides the necessary performance metrics. SUSE estimates that approximately 20% of total HPC use cases would benefit from a SUSE Enterprise Storage single-source storage solution. This type of solution works best on smaller clusters (with 250 or fewer compute nodes) and for use cases that can operate with some flexibility in latency requirements, such as analytics and AI workloads running on OpenStack. SUSE Enterprise Storage provides the concept of a caching tier, which can accelerate performance when used with faster media in front of the backing pool.

Recent performance improvements mean that even some larger HPC environments are starting to consider Ceph as a single-source, mulit-tiered storage solution. For instance, engineers at the massive CERN scientific research center in Switzerland use Ceph with a hyperconverged HPC infrastructure intended for analyzing particle accelerator data.

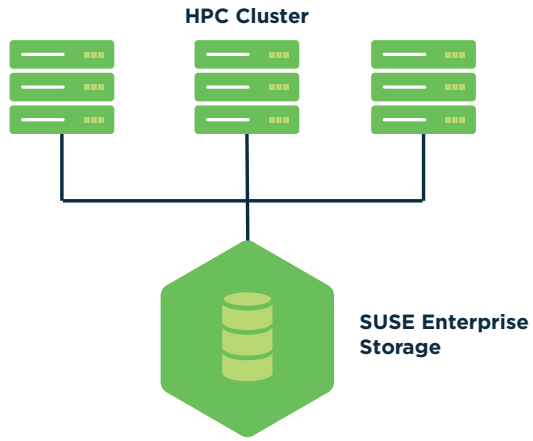


Figure 2: Single-source HPC storage with SUSE Enterprise Storage.

The scenario in Figure 3 is a cost-efficient approach for larger and more latency-sensitive clusters. A separate, Tier 1 storage component adds a specialized, low-latency layer, and SUSE Enterprise Storage provides for the rest of the storage needs. The Tier 1 component typically consists of a high-performance storage alternative such as NetApp, IBM Spectrum, or Lustre. In this case, SUSE Enterprise Storage handles long-term data, archival data, home directories, and other data receiving occasional access.

In some environments, both tiers depicted in Figure 3 interact directly with the HPC application. In other settings, the Tier 1 software interfaces with the HPC cluster, and SUSE Enterprise Storage acts as something more like an archive storage system for maintaining the complete data set. Before a job runs, data needed for the job is copied to the Tier 1 storage space. The program thus interacts only with Tier 1 storage, and, after execution, the results are copied back to SUSE Enterprise Storage. This scenario is best for data-intensive calculations, such as scientific studies, that do not require continuous access to the complete data set.

The cost of the scenario depicted in Figure 3 depends on the percentage of the total storage assigned to the Tier 1 configuration. Tools such as NetApp and IBM Spectrum use a capacity-based subscription model, so they can be significantly more expensive. Lustre does not use capacity-based subscriptions, but it can be difficult to configure and maintain, thus raising administration cost. Despite of the complications and extra cost associated with dedicated Tier 1 storage, the scene depicted in Figure 3 is perhaps the most common scenario for SUSE Enterprise Storage in HPC.

A variation on the alternative in Figure 3 is to allocate the high-performance Tier 1 storage to the scratch space used with HPC applications and use SUSE Enterprise Storage for everything else (Figure 4). The principal difference between Figures 3 and 4 is simply how you divide the subset of storage assigned to high-performance (and higher-cost) storage technologies. Figure 4 reduces the cost premium by reducing the data set assigned to the higher-cost storage option, but it might require a deeper knowledge of the use case to correctly analyze and optimize performance.

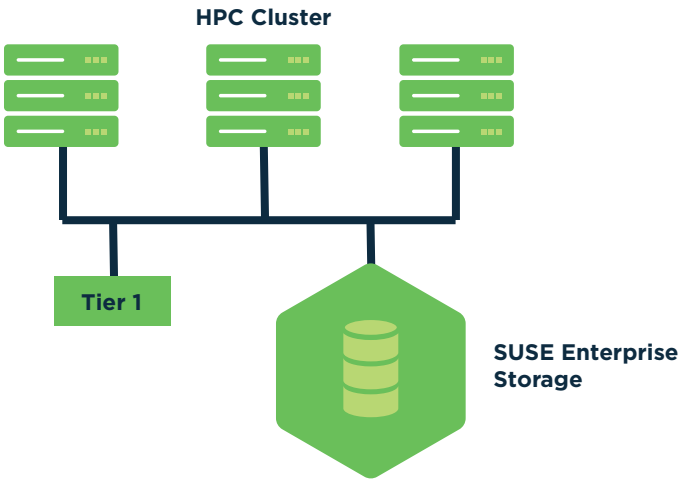


Figure 3: In latency-sensitive environments, it is often more cost effective to manage Tier 1 data through a separate storage technology.

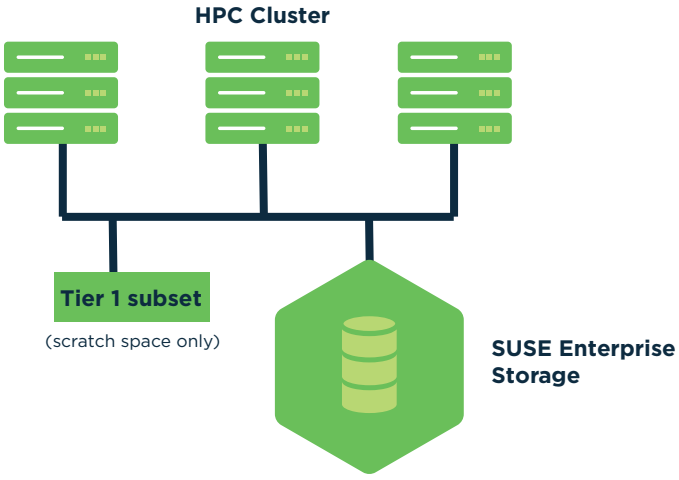


Figure 4: Some HPC systems just use Tier 1 storage for the scratch space and keep everything else in SUSE Enterprise Storage

Figure 5 shows another common scenario for SUSE Enterprise Storage in HPC. In this case, Tier 1 high-performance storage interacts directly with the HPC cluster, and some form of tiering software, such as HPE DMF or the Robin Hood open source policy engine, migrates the data from Tier 1 to Tier 2 if it is not accessed within a predefined period.

The scenario in Figure 5 often serves as the starting point for a hybrid storage solution, in which system designers mix and match storage technologies as needed to achieve minimum cost and optimum performance. For instance, the network could include multiple Tier 1 components (and even multiple Tier 2 components) operating together, each managing a different portion of the data. In this scenario, the versatile and cost-effective SUSE Enterprise Storage could serve a number of different roles, interacting with the cluster directly or operating in a near-line setting facilitated through tiering or backup software.

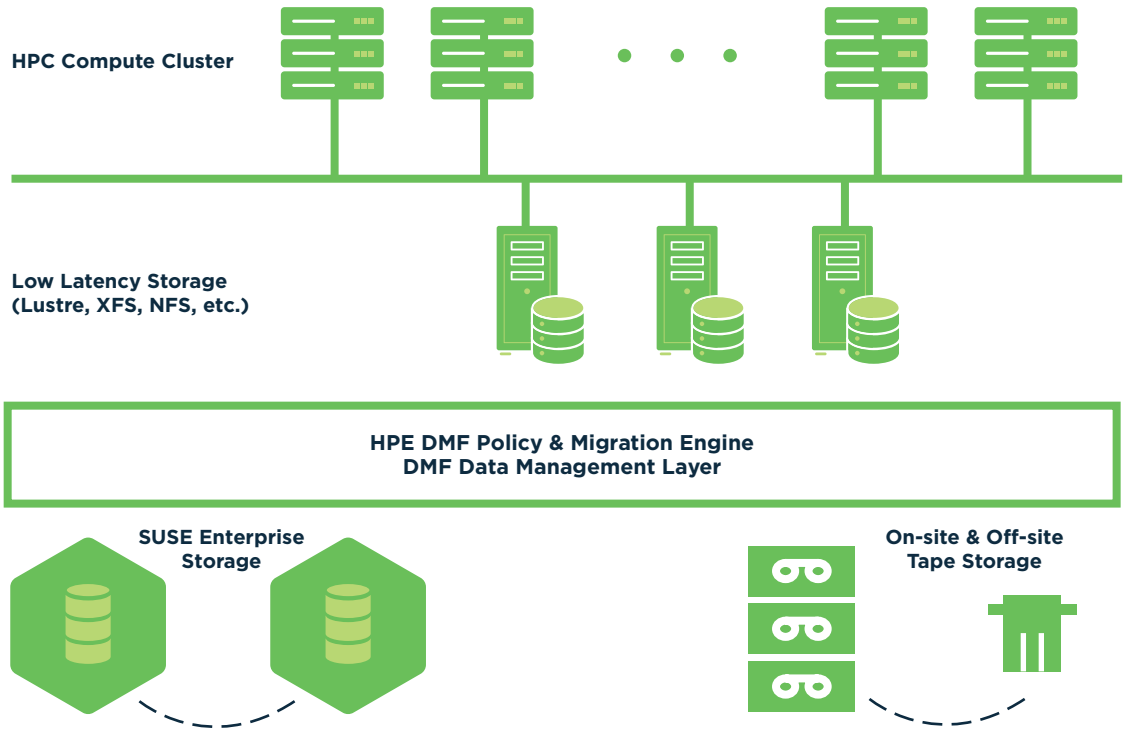


Figure 5: Use tiering software to manage the migration of data from Tier 1 to Tier 2 storage.

Conclusion

All HPC systems are different. Most high-performance systems are designed for a particular type of problem or problem-solving scenario, and the storage infrastructure is an important part of that design. This paper described how HPC engineers use tiering with SUSE Enterprise Storage to minimize storage cost and preserve more of the overall budget for other parts of the cluster design.

In each of the cases described in this paper, SUSE Enterprise Storage helps to optimize storage performance. If you think your HPC cluster is a candidate for a Ceph-based, single-source storage system, or if you would prefer to deploy SUSE Enterprise storage as a Tier 2 technology, the experts at SUSE will help you design your HPC system for maximum performance and minimal cost.

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

1800 Novell Place
Provo, UT 84606
United States

www.suse.com