Dell Storage PS Series Architecture: Load Balancers

Dell EMC Engineering
February 2017

A Dell EMC Technical White Paper

Revisions

Date	Description
August 2011	Initial release
February 2017	Updated to reflect new branding and formatting

Acknowledgements

This paper was produced by the following members of the Dell EMC™ Storage team:

Engineering: Chuck Armstrong

Table of contents

Executive summary

This document provides VMware® and Dell™ PS Series storage administrators with recommended best practices when configuring a virtual environment. It describes the function of PS Series load balancers, which are used to intelligently maintain system performance at optimum levels with PS Series storage. This document also covers connectivity and high availability, as well as some common performance-tuning best practices. In addition, there is a discussion on various methods to implement for data-drive connectivity for the virtual machines.

1 Introduction

In many organizations, significant effort is devoted to tuning resources: modifying network settings on switches, changing various parameters on servers, re-configuring QoS on the WAN, and continually altering the layout of data on the SAN. Estimates indicate that 80 percent of most IT budgets is dedicated to operations and maintenance, leaving only 20 percent for technical innovation to enhance the business. Dell PS Series storage solutions help reduce the operational cost associated with storage by automating management tasks and optimizing resource utilization. By leveraging PS Series solutions, organizations can maximize storage efficiency by ensuring that the right data is in the right place at the right cost.

2 PS Series overview

Dell PS Series storage is an award-winning, virtualized, scale-out SAN architecture that provides simplified management, ease of use, a comprehensive set of data protection and recovery tools, and great ROI for both large and small organizations. PS Series arrays are designed and built to be highly available and work together to provide virtualized SAN resources called a **Group**. The inclusive software model means never having to decipher a complex licensing scheme, and never having to purchase software upgrades to receive new features. The enhanced Automatic Performance Load Balancer (APLB) introduced in PS Series firmware version 5.1 is a good example of the outstanding value inherent in the PS Series licensing model: Customers who have valid support contracts are entitled to download the new firmware, and (once installed on all the members in the PS Series Group) they can immediately take advantage of its ability to better optimize their SAN resources — with nothing additional to purchase.

3 Why automatically balance SAN resources

The IT resources available to most organizations are limited. The realities of budgets, staffing, knowledge, and the business needs of the organization are often competing and conflicting, forcing IT managers to optimize the environment for the proper balance of these items. The usual result is that tuning for optimal performance is deprioritized in favor of more pressing items. Unfortunately, the demands that are placed on a SAN after installation are rarely the same as those that were anticipated at the time the SAN was installed. For example, applications may have been added or removed from the environment, and user loads may have increased or decreased based on the evolving needs of the business. Typically, due to data growth, the number of disks or other storage devices has increased, often dramatically.

For a classic non-virtualized storage architecture, there are typically only two options to optimize performance once these inevitable changes have occurred, and neither of them simple: Re-layout all of the data on the storage array manually, or replace the old array in its entirety and begin the process of performance tuning from scratch.

PS Series storage provides an alternative strategy for tuning that does not strain IT resources in either of those ways. Because of the virtualized scale-out architecture of the PS Series arrays, adding resources to meet the demands of the business is easy — administrators simply add an array to an existing PS Series Group and more space is immediately available. The system will automatically rebalance the load with no downtime and in a way that is completely transparent to the servers, applications, and users. Re-deploying or retiring older equipment is equally simple, just choose a member to remove, and it will automatically offload its data to other members in its pool and extract itself from the Group and be immediately ready for re-deployment or retirement. This, too, occurs seamlessly with no downtime.

4 PS Series load balancers in PS Series storage pools

When you initialize the first array and create a PS Series Group, a default **pool** is automatically established. An array added to the Group is referred to as a **member** of the Group. The first member added to the Group is initially placed into the default pool, and administrators subsequently deploy **volumes** from this pool. Within the pool, resources such as network bandwidth, disk capacity, and I/O are balanced automatically. Multiple pools can be created to isolate volumes and separate members. This may be done for a variety of reasons, including technical (placing specific application data on resources such as SSD) or business reasons (ensuring that legal department data is isolated from the data from other departments).

With more than one pool, administrators can initiate moving volumes or members between the pools seamlessly, with no downtime to the applications. Within a pool, PS Series arrays are designed to automate the placement of data to maximize the utilization of the resources that the customer has chosen for their SAN.

There are three load balancers that operate within a pool:

- The NLB (Network Load Balancer) manages the assignment of individual iSCSI connections to Ethernet ports on the pool members
- The CLB (Capacity Load Balancer) manages the utilization of the disk capacity in the pool
- The APLB (Automatic Performance Load Balancer) manages the distribution of high I/O data within the pool.

5 How the Network Load Balancer (NLB) works

Communications between application servers (iSCSI initiators) and volumes (iSCSI targets) are called **connections**. A PS Series Group will present all iSCSI targets through a single virtual address known as the **Group IP address**. This allows administrators to establish connections easily by configuring the iSCSI initiator with only the Group IP address instead of the IP addresses of all of the interfaces in the Group.

As the load increases or decreases on the various Ethernet ports, the NLB automatically distributes connections among the active Ethernet ports of the members using a feature of the iSCSI specification called **redirection**. Redirection defines how the iSCSI target instructs the iSCSI initiator to log out and close the connection to the IP address that it is currently using and immediately log in to another address and establish a new connection. Support for redirection is required for iSCSI initiators by the iSCSI specification. Redirection is utilized by the NLB within a PS Series Group to permit the application server to establish iSCSI connections as needed without first needing to be updated manually to know all of the possible IP addresses that the SAN is using. Leveraging redirection, the NLB ensures that all the network interfaces within the SAN are optimally used. The NLB and iSCSI connection redirection are also key functions used by the PS Series architecture to enable volumes and members to migrate seamlessly from one pool to another, and permit members to join or leave the Group as required with no interruption in service to the applications.

The NLB should not be confused with Multi-Path I/O (MPIO), which is load-balancing that occurs on the application host. MPIO uses redundant physical interfaces to deliver high availability to shared storage. Using MPIO, servers can send multiple I/O streams to SAN volumes. Each of these paths uses an iSCSI connection managed by the NLB.

In addition to the standard functionality provided by MPIO, Dell EMC provides host tools to enhance the performance of MPIO and to automatically manage the connections for Windows (including Microsoft® Hyper-V®), VMware and Linux environments. For a further discussion on the various MPIO enhancements provided by Dell EMC, refer to the following Dell EMC technical reports:

Configuring and Deploying the Dell PS Series Multi-path I/O Device Specific Module with Microsoft Windows

Configuring and Installing the PS Series Multipathing Extension Module for VMware vSphere and PS Series SANs

6 How the Capacity Load Balancer (CLB) works

The CLB ensures that as volumes are created and deleted, and as members are added to and removed from a pool, the relative percent of capacity in use is maintained at a consistent level among the members in that pool. Keeping the members in the pool filled to the same percentage of their disk capacity helps to ensure that all of the resources in the pool are used equally, and helps avoid overloading one member compared to another. It can also help ensure that members have the necessary free space available to perform other tasks such as replication and internal maintenance properly.

When the CLB assigns a portion of a volume to an array, it is called a **slice**. The CLB will attempt to satisfy the capacity needs of each volume with a distribution policy that typically limits the number of slices per volume to three. More than three slices will only be created when the capacity requirements of a volume cannot be satisfied with three slices.

Most administrators choose the default **Automatic** RAID preference setting for the majority of their volumes. The CLB will normally choose the members to use without regard to RAID level unless the administrator selects a specific RAID preference type for the volume (for example, RAID 6).[1]

If an administrator chooses a specific RAID type and it is available in the pool, the CLB attempts to honor the preference request and place the volume on members with the requested RAID type. As long as all of the volumes that are requesting a particular RAID type can be accommodated on members of that RAID type they will be, even if this results in the members of the pool with the requested RAID type having higher capacity utilization than other members of the pool. If the request cannot be honored because there is insufficient capacity available (or no members) at the requested RAID type, volumes will be placed on other resources in the pool as if the RAID preference had been set to Automatic. Setting RAID preference for a volume in an existing environment may cause other volumes with their RAID preference set to Automatic to have their slices moved to members other than the ones that they resided on prior to the change.

When the CLB needs to re-distribute the data in the pool, it creates a rebalance plan (RBP). An RBP might be created in response to a change in the resources available in the pool (adding or removing a member), or a change in the way that the current resources are used (adding a volume; changing a snapshot or replica reserve; modifying delegated space for the replicas from another PS Series Group; or resulting from the growth of a thin provisioned resource). An RBP is influenced by any RAID preference settings for the volumes in the pool and will, when possible, honor RAID-preference settings for volumes as discussed previously. As resource usage is optimized, an RBP may temporarily create a capacity imbalance, but after the RBP is executed, the imbalance will be rectified.

Similar to an RBP, the CLB can also create free-space-trouble plans (FSTP). An FSTP is created when the CLB determines that a pool member has reached a critical point (90 GB) and there is free space available on

[1]In small pools, this will always be the case. In a larger pool with at least three members at RAID 10, for volumes whose RAID preference is set to Automatic with a historical profile of large amounts of random I/O (as measured in days to weeks), all of the volume slices for that volume will be placed on RAID 10 by the CLB if it is available in the pool. Since there is no prior history of I/O to reference when determining where to initially place a volume, in the absence of volume preferences, the CLB will determine where to place the slices for the new volume based on capacity alone in all pool sizes.

other members in the pool. An FSTP will cancel other RBPs. Once the low space issue that prompted the FSTP has been resolved, the CLB will create new RBPs if they are required.[2]

All data movement, regardless of whether caused by an RBP or FSTP, is handled in a transactional manner — data is only removed from the source of the transfer and internal metadata that tracks the location of the data is updated only after its receipt is confirmed by the target of the transfer.

[2]While this is not a commonly used feature of the PS Series, volumes may be explicitly bound to specific members by the volume bind CLI command. These bound volumes are exempt from movement when either an RBP or FSTP is created.

7 How the Automatic Performance Load Balancer (APLB) works

The APLB feature is designed to help alleviate the difficulties inherent in manually balancing the utilization of SAN performance resources. Operating on the resources in a pool, The APLB is capable of adjusting to dynamic workloads in real time and at a sub-volume level. It will provide both sub-volume-based tiering when presented with heterogeneous or tiered resources to work with, as well as hot-spot elimination when presented with homogeneous resources in the pool.

The APLB optimizes resources in a PS Series pool based on how the applications are actually using the SAN resources. When the slices have been assigned to members in the PS Series pool by the CLB and I/O begins, certain patterns of access may develop.[3] Due to the random nature of I/O, these access patterns are often unbalanced, which while normal, may place more demand on certain PS Series members than others. Often, the imbalance occurs within the same volume with portions of the volume exhibiting high I/O, while other portions of the volume exhibit low I/O. This imbalance can be detected and corrected by the APLB.

In a PS Series pool, all other PS Series products can adjust to this potential imbalance in latency. In the event that a workload causes a particular PS Series member to exhibit relatively high latencies compared to other members of the same pool, the APLB will be able to detect and correct this imbalance and by exchanging high I/O data from the PS Series member with high latency for low I/O data from a peer with low latency. This rebalancing results in better resource utilization and an overall improvement in the performance of all of the applications using the resource of the PS Series pool.[4] Figure 1 shows a basic example of this across two members.

---

[3]The response time required to service I/O requests to and from applications is referred to as **latency**. The I/O requests can be either **random** (access non-contiguous data) or **sequential**, where the data that is accessed in contiguous. Sequential data access is often exhibited by applications such as streaming media, and may be more tolerant of higher latency response from the storage as long as the stream of data is consistent, while random access is often less forgiving and requires lower latency when users are actively using an application in order to provide an acceptable user experience.

[4]Classifying data as high or low I/O is determined by the frequency of access to the data by the applications over a relatively brief interval.

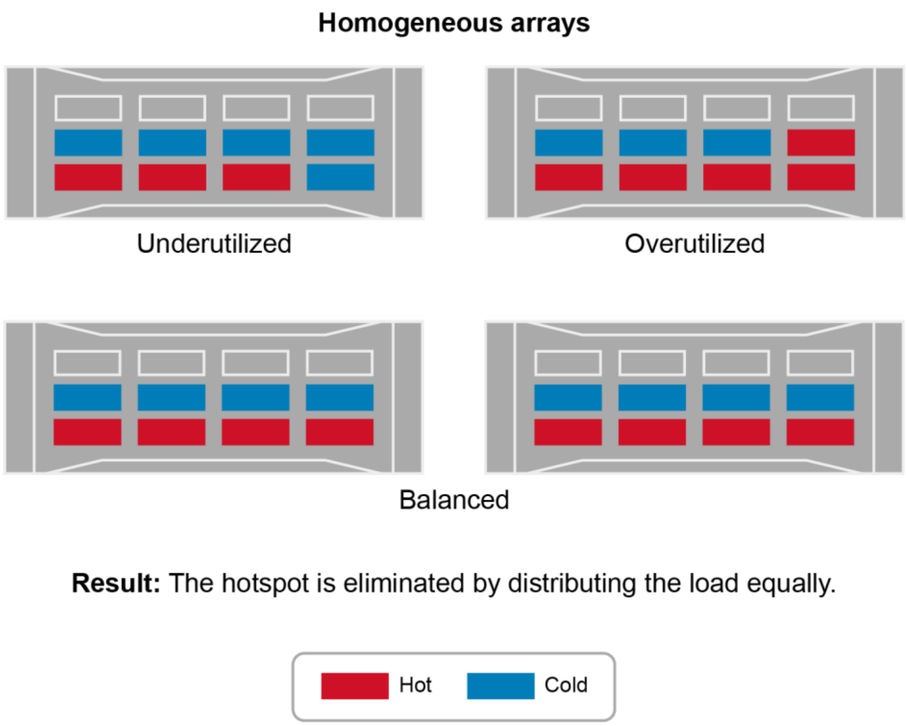

Figure 1 Example of APLB in a non-tiered environment

The APLB is surprisingly simple in its concept and execution, leveraging various aspects of the PS Series architecture to automatically balance the performance delivered to applications by the PS Series SAN. For example, the rebalance plans that the CLB uses to re-adjust the placement of data, are leveraged by the APLB as well. Instead of the typical one-way movement that the CLB usually performs, movement of data in the RBPs that the APLB creates is typically a two-way exchange between PS Series members to ensure that after a performance rebalance operation, the capacity balance is still maintained.[5]

As with all PS Series management tasks, the APLB runs with a lower priority than the processing of application I/O. Every few minutes, the APLB analyzes the range of latencies of member arrays in a PS Series pool, and determines if any of the members have a significantly higher latency (20 ms or greater) than the latency of the lowest latency members(s) in the pool. If it does, the APLB will attempt to identify workloads that could be rebalanced by moving high I/O data to less heavily loaded members (those with lower latency). If any are identified, an RBP will be created to exchange a portion of the high-I/O data from the member with high latency with an equivalent amount of low-I/O data with one of its peers supporting the workload that has been selected for rebalancing. The peer member chosen for the data exchange will be one of the other members in the pool already supporting a slice of the volume that has been selected to be rebalanced.

When the APLB is presented with more than one option for rebalancing (the volume selected for rebalancing has slices on two other members in a larger pool), and the latency of both options is similar, the APLB will use a second criteria to make the determination. This second criteria is the relative **busy-ness** of the arrays,

[5] In rare cases where a two-way exchange is not possible to achieve, the RBP created by the ALPB creates will still move some of the hot data to the less busy array. If this creates a low free space condition, then a FSB will be created by the CLB to address the situation.

which is a composite score of factors such as RAID type, disk speed, number of disks, as well as PS Series controller type and the current I/O load. The member with the lower relative busy-ness will be chosen for the data exchange.

The APLB works well in a variety of environments. For example, in PS Series pools with members displaying similar performance characteristics, the net effect is to eliminate hot spots in the pool. This is shown in Figure 1. In pools with members displaying dissimilar performance characteristics (for example, arrays with different drive types), the net result is tiering of the data such that the bulk of the active data will typically be serviced by the array(s) with the most I/O capability. This is shown in Figure 2.

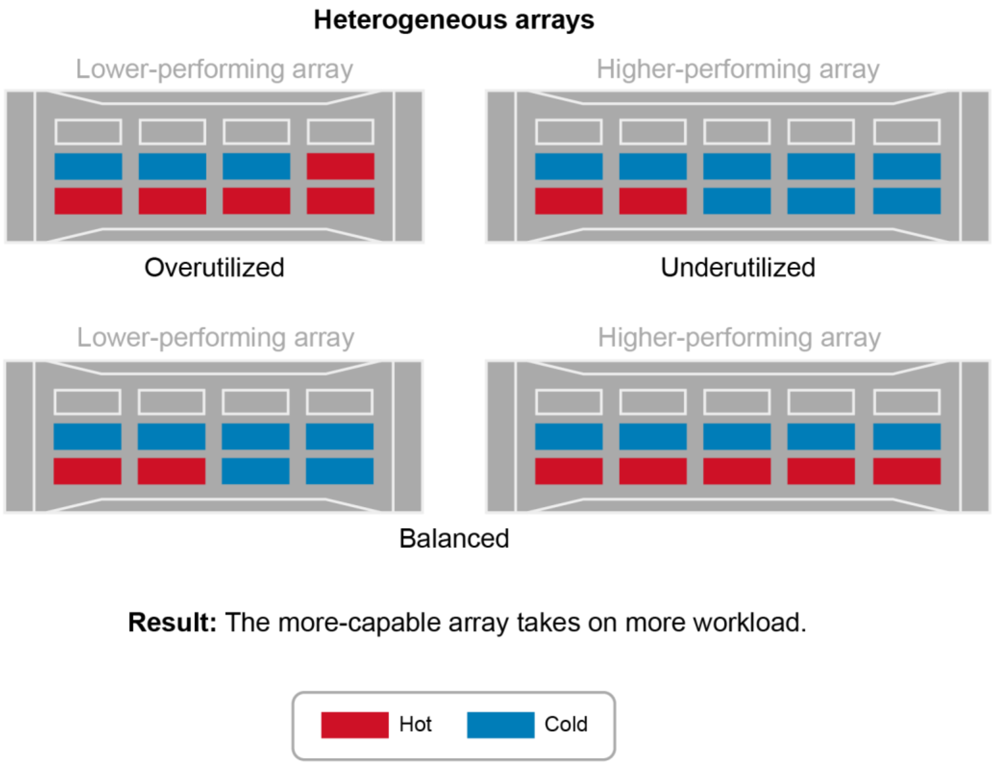

Figure 2 Example of APLB in a tiered environment

The data used to determine what portion of the workload is high I/O is based on recent activity, (on the order of minutes) so the APLB is able to adapt to a change in an application I/O pattern quickly. The APLB is also dynamic, constantly evaluating the environment and making small adjustments as required. When an application has reduced its demand for resources, the APLB does not continue to optimize the formerly active data. The advantages of the APLB approach are fourfold:

Seamless support of 24/7 business activities: By adjusting incrementally, there are no large batch movements of data. Instead, the APLB spreads the overhead of rebalancing into small operations through the day instead of in one large activity.

Ability to adjust to cyclical or one-time workload changes: By evaluating a relatively recent window of activity, the APLB detects the temporary nature of certain increases in I/O load (such as end-of -month financial activity), and they do not continue to influence the balancing of data after they are no longer relevant.

Reduction of worst-case-scenario purchasing: By working continually, the APLB can detect and act on cyclical business processes, such as increased end-of-month activity by the finance group enabling the resources of the SAN to be leveraged in near-real-time. This may enable IT management to purchase fewer resources since each application can better leverage the storage when it needs it most.

Future-proofed load-balancing: By using latency as the primary criteria, the APLB does not need to explicitly evaluate any other details of the storage, such as disk type (SAS or SATA), spindle speed, number of disks, or PS Series controller type. This makes the APLB a very simple and robust mechanism that does not need to be re-trained when new hardware configurations are introduced to the PS Series product line. This also ensures that when unplanned events occur that may influence the ability of certain arrays to serve their workload (for example, a RAID rebuild or bad NIC), the system automatically compensates.

8 Observing latencies and high I/O data

Various metrics central to the APLB behavior (such as array latencies and breakdown of data classified as high, medium, and low I/O) are displayed using SAN HeadQuarters. These metrics are part of the I/O subsection of each Group and can be readily observed for Groups, pool and the individual members within a pool. Observing this information can be helpful in understanding the operation of the APLB. For more information on SAN HQ, please consult the product documentation or Dell EMC document, *Analyzing PS Series Storage with SAN HeadQuarters*.

9 Tiering with the APLB

When provided with tiered resources in a pool, such as arrays with different spindle speeds or set to different RAID types, the APLB is able to use them to tier the workload. This is not limited to any particular RAID type, interface I/O type, spindle speed, number of disks, or PS Series controller generation, since all of these factors are abstracted by the use of latency, which is the primary factor when deciding when to rebalance the workload. The ability to tier gives the customer greater flexibility in selecting products that provide the correct combination of performance and capacity for their environment, since any of the factors mentioned previously could be relevant to creating differences in latency between PS Series members in a pool. For example, combining large-capacity PS66x0 class arrays with lower capacity PS62x0 arrays using disks that provide higher I/O to get better total system ROI may be the appropriate design for some customers. Others might choose to combine members with 10K SAS and members with SSD to meet their application workload requirements. Many other configurations are possible, these are simply examples. For one example of tiering, see Figure 2. For additional examples, see appendix A.

The ability of the APLB to tier also means that PS Series products are capable of placing the inactive data on lower-cost media, allowing more efficient use of resources to provide higher I/O to customer applications at a lower overall cost. This includes snapshot data, which may end up on lower cost media compared to the active data that they are protecting. Similarly, applications like OLTP databases or those that leverage PS Series thin clones or similar technologies, (such as VDI) that concentrate the majority of the read activity into a small portion of the total capacity can take advantage of the APLB. In these cases, when provided with heterogeneous resources (those with different random I/O capabilities), the APLB optimizes these workloads by migrating the frequently used data to members with greater random I/O capabilities, such as SSDs, while utilizing capacity-optimized members with lower I/O capabilities to host the bulk of the content.

When setting up a tiered environment, the capacity load balancer typically chooses a maximum of three pool members to distribute the slices for a given volume; if the ultimate goal of a particular SAN design is to provide an automatically tiered storage environment, then the pool should be made up of heterogeneous storage arrays and be limited to a maximum of three members for the most predictable results. Creativity in design, such as using a combination of tiered resources at the same RAID level and RAID preference settings, can provide predictable behavior with larger pools containing tiered resources. This is generally not a consideration when tiering is not a primary goal, in which case homogeneous resources would be deployed in the pool, up to the limits of the pool.

10 Using the APLB with hybrid arrays

The PS Series hybrid arrays (such as the PS4210XS, PS6210XS, and PS6610ES) use their own internal implementation of the APLB to compensate for unbalanced or tiered workloads. These arrays shift the active parts of the workload in the volume slices that they are responsible for to SSD media for extremely low-latency I/O, while leaving the bulk of the workload on HDD media for normal latency I/O. This lowers the overall latency of the hybrid products for the workloads that they are supporting.[6]

Since the CLB treats hybrid arrays like any other member in the PS Series Group, capacity will be assigned to them in relation to their overall capacity contribution to the pool. The APLB is able to work in conjunction with the hybrid arrays, permitting other, non-hybrid arrays to migrate active data to the hybrids in exchange for inactive data. In the case where more active data is placed on a hybrid array than the SSD tier is able to hold, the APLB is able to migrate active data from the HDD tier to another member such as a second hybrid or standard array that may have more resources available to better service the I/O requests. An example of a pool containing both a hybrid array and a non-hybrid array is shown in Figure 3. An example of two hybrid arrays in the same pool is shown in appendix A.

Additional details on APLB can be found in the document, *EqualLogic PS Series Architecture: Hybrid Array Load Balancer*.

[6] The hybrid arrays use only access frequency, not latency, when deciding what data to optimize internally. This, and the lower overhead of moving data internally, permit the hybrid arrays to optimize their workloads internally at a higher rate than the inter-array optimization of the APLB within the pool.

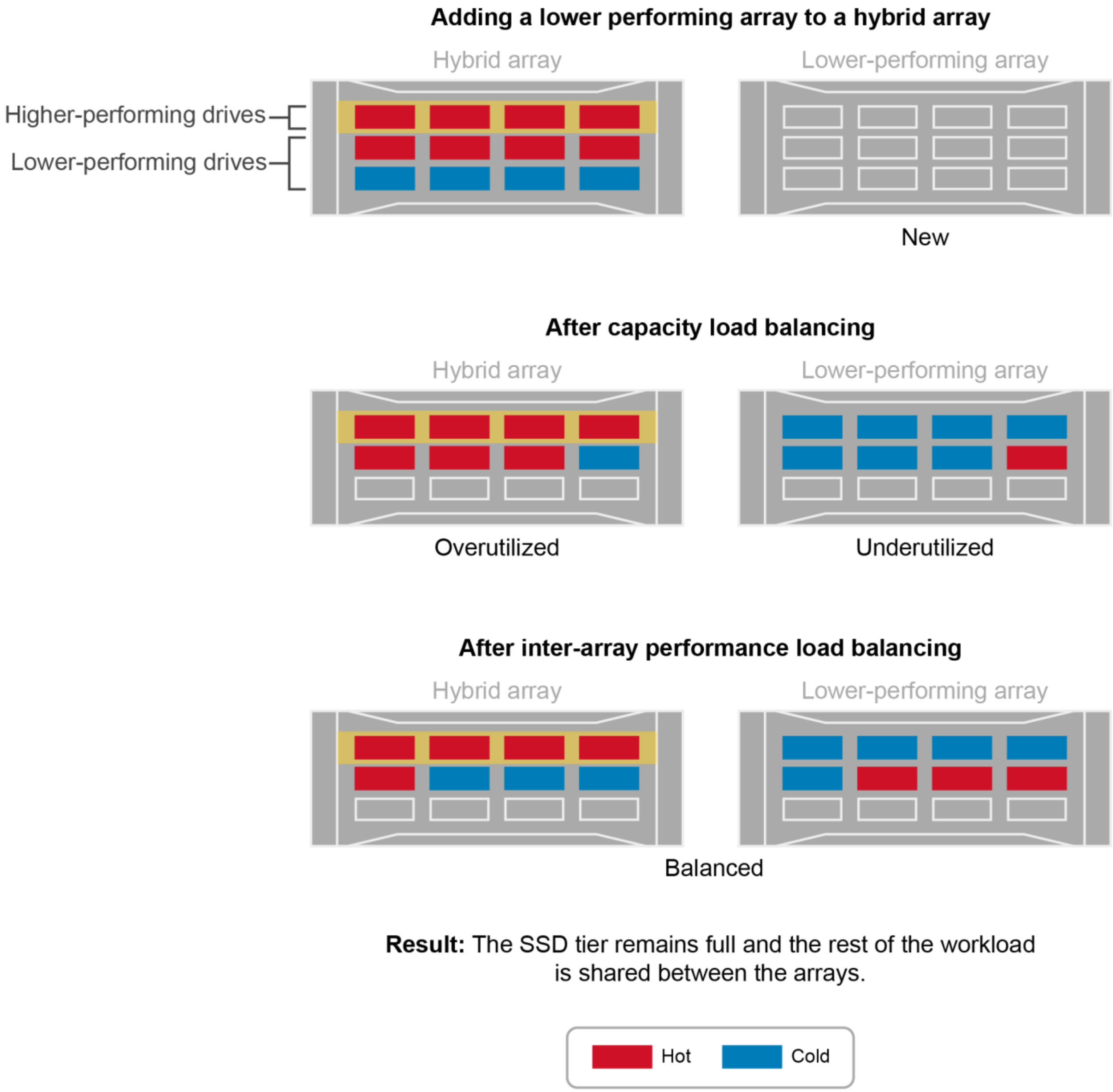

Figure 3 Example of APLB in a tiered environment

11 Scenarios in which APLB action is limited

There are several cases when the APLB will not take action. First, if all of the members of the storage pool are exhibiting low latencies, then the resources do not need to be balanced. In this instance, the workload is receiving an appropriate level of performance (low latency); it is possible that in a low-resource utilization scenario such as this, a member with lower performance characteristics could host more active data than a member with higher performance characteristics. In this case, the APLB merely monitors the latency of the pool members, but since there is nothing in need of optimization, it takes no further action. This is shown in Figure 4.

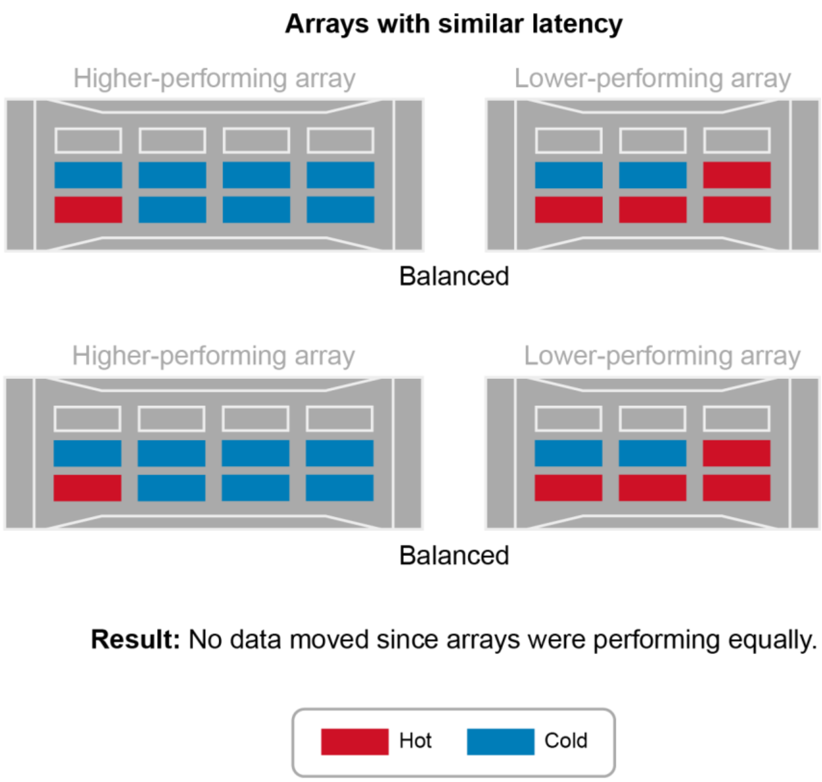

Figure 4 Fair workload distribution

Second, if several members are all supporting a workload and all are experiencing high but similar latency, then no amount of data exchange will improve the situation where more is being asked of the resources than they can provide. The APLB will not attempt to optimize arrays when there is no configuration that would provide improved performance. SAN HQ can be used to observe the situation, and if not a result of network or server misconfiguration, the likely resolution is to add additional members to the pool so there are more resources to handle the workload. For more information on SAN HQ, consult the document, *Analyzing PS Series Storage with SAN HeadQuarters*.

The third case in which the APLB will not take action is if a member is changing latency very rapidly, swinging from high latency to low latency and then back again. The APLB will detect this thrashing and will not attempt to alter the distribution of active data since the application is not behaving in a predictable manner and is not able to be optimized.

Fourth, if a customer has chosen to override the automatic placement of volumes by the capacity balancer by using the CLI bind command, all balancing of any sort is disabled for that volume. Note that this does not disable the APLB for other application workloads that may be using the same member that the volume is bound to.

A case where the APLB behavior should be observed more carefully is when volume RAID preference is utilized. If the pool contains only one member of the selected RAID type, the effect of the RAID preference option is equivalent to the CLI bind command, and the APLB will not be able to optimize the workload. On the other hand, if there are multiple members in the pool providing the selected RAID level, then the CLB will be able to create multiple volume slices. In this case, then the APLB will work as expected.

12 APLB and backup operations

Backup operations place a large burden on storage resources but should not be a problem for the APLB. Backup operations largely result in sequential reads, and only access a given portion of the data for a brief portion of the function. Thus, since there is no repeating pattern of access, the APLB will not take action, which is desirable since there is no optimization that may be achieved.

13 Disabling the APLB

While not recommended, the APLB may be disabled by unchecking the **Enable Performance Balancing in Pools** checkbox in the GUI. Note that this is a Group-wide setting and will affect all pools in the PS Series Group.

14 Summary

The various load balancers that work in a PS Series pool provide a very flexible, dynamic framework that is able to quickly adapt the PS Series storage to the shifting requirements of the customer environment. Changing I/O activity is easily balanced between the available network interfaces on an iSCSI-connection-basis by the NLB. Working in conjunction, the CLB and the APLB are able to help actively manage the data on the members using whatever resources the customer has chosen to provide. Overall, the load balancers permit the PS Series SAN to adapt in a fluid manner to maintain optimal resource utilization.

A Additional examples of APLB

The following are some additional examples of the data distribution before and after APLB has performed its load balancing function.

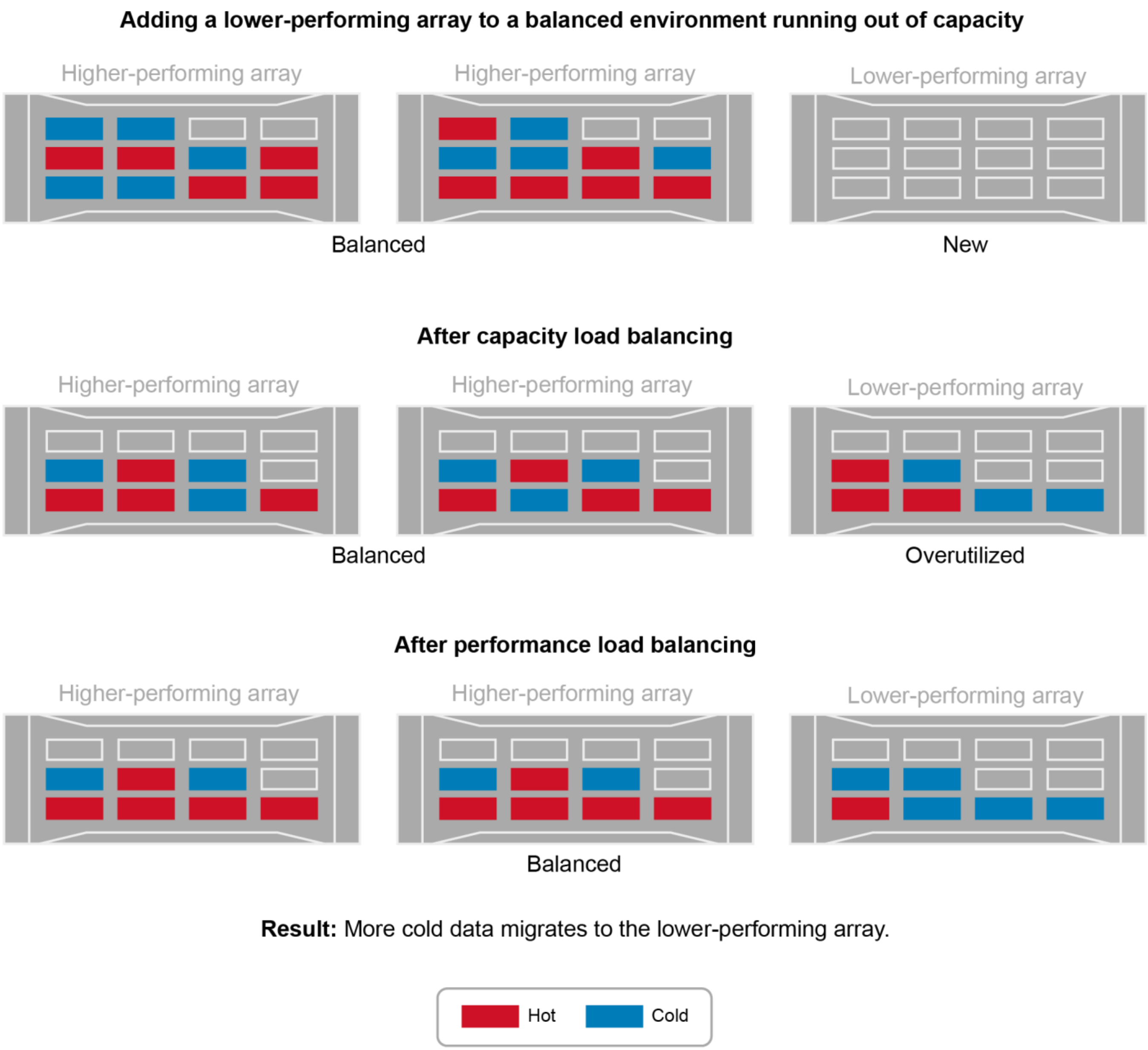

Figure 5 Rebalancing after adding capacity-optimized resources to a pool

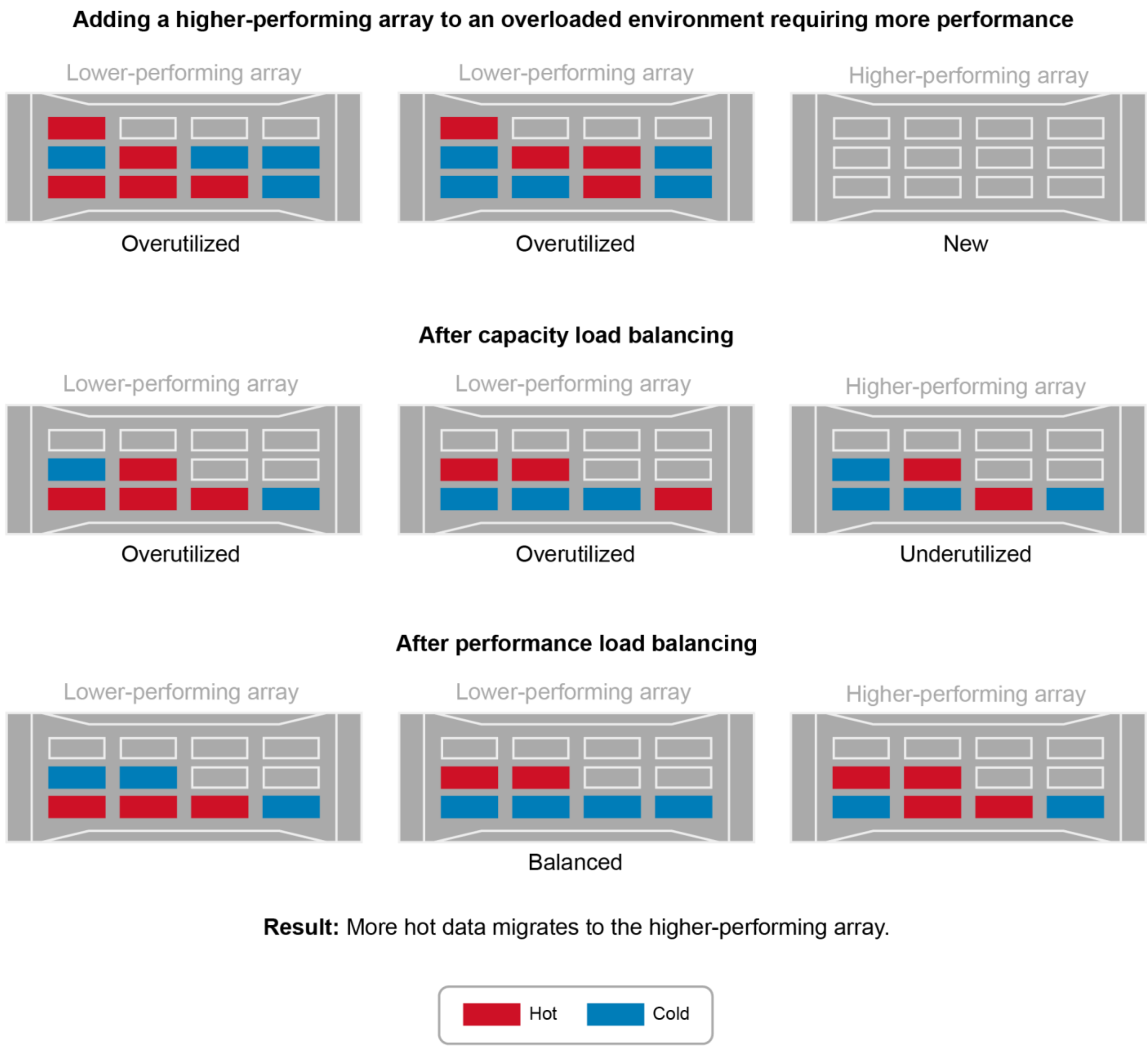

Figure 6 Rebalancing after adding performance-optimized resources to a pool

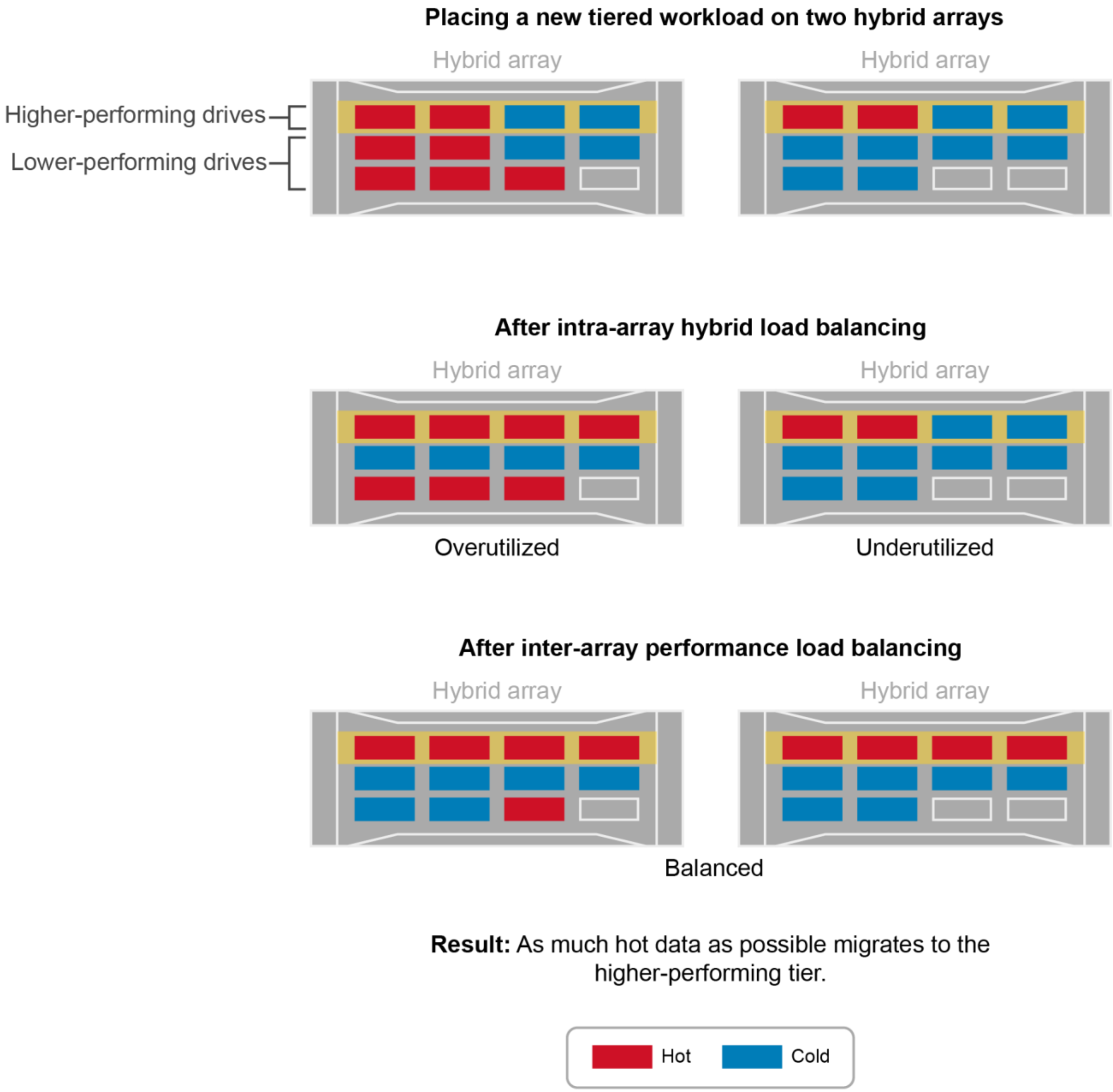

Figure 7 Rebalancing with multiple hybrid arrays

B Technical support and resources

Dell.com/Support is focused on meeting customer needs with proven services and support.

Dell TechCenter is an online technical community where IT professionals have access to numerous resources for Dell software, hardware and services.

Storage Solutions Technical Documents on Dell TechCenter provide expertise that helps to ensure customer success on Dell EMC storage platforms.

B.1 Related documentation

Articles, demos, technical documentation and more details about the PS Series Storage product family are available at PS Series Technical Documents.

For additional information, see the following referenced documentation:

- *Dell Storage Compatibility Matrix*
- *Dell PS Series Arrays: Advanced Storage Features in VMware vSphere*
- *Configuring iSCSI Connectivity with VMware vSphere 6 and Dell PS Series Storage*
- *Configuring and Deploying the Dell PS Series Multi-path I/O Device Specific Module with Microsoft Windows*
- *Configuring and Installing the PS Series Multipathing Extension Module for VMware vSphere*