PS Series Storage Array
Load balancers
Module Objectives

• Upon completion of this section the student will be able to:
  – Identify the different types of load balancers in the PS Series Arrays
  – Describe each of the different types of load balancers
  – Compare the different load balancers
  – Describe benefits and best practices for each of the load balancers
You need to understand servicing the customer workloads. As load increases in the SAN, latency will increase as SAN system resources (network, controller/RAM and Spindle to name a few) become taxed.

Ideally, latency is kept low, and application responsiveness is high.

When the number of workloads in the environment increase, the aggregate workload on the SAN becomes increasingly random.

Over time, workloads can change in profile, resulting in resource contention. Traditionally in the SAN world, our SRM tools could alert us of "hot spots" in the storage and SAN managers could reallocate resources to react to these new workload profiles.

Equallogic natively distributes the work across all available resources to reduce resource contention.

The methods by which we allocate and redistribute workloads will be covered in this session.
There are many methods by which the Equallogic system distribute work across the available resources in the system or “Group”
We call these engines “load balancers” and some are always available, while some are available only in multi-member pools.

Load balancers available to all pools:
- Network load balancer: The NLB is designed to distribute iSCSI connections across [appropriate] available network interfaces to load them proportionally.
- Equallogic MPIO: Enables Host interface to array port distribution and proportional loading

Multi-member Pool Load balancers:
- Capacity load balancer: The CLB maintains proportionate disk consumption across the pool to enable rapid changes in consumption while minimizing impacts on performance
- Free Space Balancing: Maintains free space distribution across the pool as consumption changes as space is no longer consumed
- Automatic Performance Load Balancer: APLB manages the overall latency of members in the pool by swapping hot pages for cold(er) pages within the pool.

Other load balancers:
- Hybrid array load balancer: The hybrid LB (RAID 6 Accelerated) manages the placement of high I/O data within the different media types within the array.
PS Series supports Tiered Storage, which enables you to define multiple tiers — or pools of storage in a single PS Series group (SAN).

PS Series SAN is comprised of a what is known as a Group. A Group is that component of the SAN that the Host will initially access to reach their volumes and it is also what is managed using either a Browser GUI interface or via a Command Line Interface.

A group is comprised of arrays, known as members. A group could be a single member group or it could be a multi-member group. To add additional storage all that needs to be done is to add a new array, member, to the group. This can be done while the applications are running with minimal to no impact to the application servers.

A member is added to the group and belongs to a pool. A pool is comprised of the storage that the members within that pool provide. Additional storage can be added to a pool by simply adding another member to the desired pool.

Pooling is logical segmentation of the SAN and provides a capability to perform tiered Storage. This provides administrators with greater control over how disk resources are allocated. While online, volumes can be allocated and moved between tiers of storage, providing high levels of service.

You can create multiple tiers of storage within a single, multi-member PS Series group by creating multiple storage pools. Pools allow you to control where data is located.

Pools - A pool consists of one or more group members (PS Series arrays). At any one time, a member can be assigned to only one pool — or storage tier. It is easy to assign a member to a pool and also to move a member between pools, with no impact on data availability, providing administrators with the flexibility needed to meet their business needs.

Volumes are also assigned to a pool and can be easily moved between pools, with no impact on availability. In addition, automatic data placement and load balancing occurs within a pool, based on the overall workload of the storage hardware resources within the pool.
In the PS Series Array environment there is no need for administrators to manually create RAID sets or map data onto disks or individual controllers. When a volume is created it will be striped across arrays in the pool. A pool is a logical container that contains arrays. The virtualization process of spreading the volume out over multiple disks makes it possible to increase a volume size by simply specifying a new size in the Group Manager GUI.

The storage firmware handles storage allocation and capacity balancing across the disks and arrays. When a new array is added to a pool, volumes within that pool will immediately distribute across to the new member regardless of RAID type or spindle speed.

The optimum number of arrays that a volume will be spread across is 3, but this can vary depending on volume size and other factors. At initial volume creation, the firmware load balances based on proportional capacity. The administrator can set a RAID preference for a volume provided that they understand the likely volume I/O mix. Otherwise, the volume preference is set to the default value of “Automatic” which means that it will be spread over multiple arrays.

The system regularly free space balances. This is the normal process to ensure that there is an equal amount of free space on each member in a pool.

Administrators can guide the load balancer by making decisions on pool composition, which pool a volume is placed in and setting of RAID preferences within a pool to adjust the placement of data. These topics will be discussed in greater detail in a future module.
When the first PS Series array member is initialized and added to the group, it is added to the “default pool”.

If it is added using the CLI, a RAID type must also be applied to the array before any capacity is allocated to the group.

When a second member is added to the group, it can be added to the existing pool or to a new pool.

When the member is initialized, none of the available storage capacity of that member is allocated to the group/pool because a RAID level has not been applied to the new member.

Once the RAID has been specified, the new member’s storage capacity will be allocated to the pool it was added to.

A single pool can have multiple members in the pool. Each member can be a different RAID type as shown in the diagram.

Here you can have a PS6000X as a RAID 50 and a PS6000XV as a RAID 10 and a PS6000E as a RAID 5 all within the same pool. The volumes will move to the appropriate member based on I/O characteristics.

It is always recommended in a mixed spindle speed environment that you use pooling to separate the slower SATA arrays from the faster SAS arrays.

If you have multiple members with different RAID types within a single pool:

Volumes will move automatically between members based on the I/O characteristics of the volume. The I/O is based on the RAID type, if the volume preference is set to “Automatic”.

Volumes can be associated with a particular member, based on RAID type, by setting volume preference on the volume to the desired RAID type.
Tiered storage with multiple pools

Tiered storage with multiple pools allows:
- Separation of storage based on customer requirements
- Control of separate storage
- Appears like virtual SAN islands

Tiered storage with multiple pools prevents:
- Volumes from using available capacity in another pool. A volume is restricted to the pool it resides in.
- Automatic movement of volume using volume preferences only happens within a pool and not between pools

PS Storage arrays are added to the group and to a pool

When the array is added, it can be added to an existing or new pool

As of firmware version 3.2.x, pools can be merged. Prior to that, pool merging was manual.
Vacate is a process typically where the storage administrator can repurpose an array/member with no impact to the application servers and that are connected to volumes within the SAN.

The reason for a vacate is when the administrator reallocate the array to use it somewhere else in the organization, but they want to do it in a way that will not jeopardize the existing I/O to volumes within the SAN. When the vacate starts the array will confirm that there is enough storage in the pool it is leaving to support the volumes, if there is not then it will fail. If there is sufficient space to support the volumes then the vacate process will redistribute the volumes to the remaining members within the pool and once that is complete successfully the array will then be reset to factory defaults.
PS Series arrays provide dynamic network connection load balancing as well as the volume load balancing. As the connection workload changes, those connections on the more heavily loaded interfaces will be rebalanced to provide maximum performance to the volume data and insuring maximum network I/O. This is done across the arrays that have the connected volume striped across with no impact on applications and no user intervention.

The result of this rebalancing have connections connection being moved form one interface to a less busy interface on an array with the same pool.

It should be noted that pages do not move between members, If there is not a connection to a member here the data exists, then the member to member mesh will be used to retrieve that data, rather than moving the connection.
Why us

It is best practice to run MPIO on your servers
- Improves reliability of storage access
- Improves resilience to errors
- Improves performance

EqualLogic MPIO is easier to use than host/OS MPIO
- Automatically manage iSCSI sessions
  - Create iSCSI sessions based on the SAN configuration to ensure High Availability and maximize I/O performance
  - Sessions automatically raise and lower based on operating needs
- It comes with the product and does not cost extra.

It performs better than host/OS MPIO
- Optimizes network performance and throughput
- Provides End to end network balancing (rather than array to switch balancing)
- Route I/O directly to Member which will be servicing it
- Reduces overall network traffic

EqualLogic Multipathing?
Connection Awareness to a PS Series SAN
- EHCM – EqualLogic Host Connection Module
  - Manages sessions to and from the iSCSI initiator
- Remote Setup Wizard – MPIO Settings Page
  - Allows management of DSM
    - Include / Exclude networks
    - Define load balancing policy
    - Configure session connections
- iSCSI Initiator MPIO Session Management Tab
  - Real-time view of MPIO session connections
- I/O Routing uses a two step process to choose the optimal path
  - Calculation is performed independently for each I/O
  - Least Queue Depth is preferred because it dynamically shifts more I/O to the paths with lower latency
- Network load balancing continues on array ports
The goal of Automatic Performance Load Balancer, APLB, is to detect an unbalance in the members latency within a pool and to move hot and cold pages between members to balance the member latency thus increasing performance.

Automatic Performance Load Balancer, APLB, is introduced within firmware version 5.1 version. The goal of APLB was to decrease member latency by moving data that is on a member within a pool that has high latency to another member within the same pool that has lower latency.

Data that resides on a member maybe be considered in one of three possible states:

- A Hot page, which is defined as a page that is currently being written to or that has recently been written to.
- A warm page, which is a page that has data on it but that data has not been access for some time.
- A cold page, is defined as a page that has not been written to.

When the volume is first created it is spread across the members of the pool where the pages are spread based on the volume and member capacity. As data begins to get accessed the latency of each member is analyzed at two minute intervals. If data is being accessed on a member that has a higher latency than another member within the pool then the pages associated with the data will then be moved to that member with a lower latency. As mentioned the member latencies are evaluated every two minutes.

Be aware:

- Page swapping will be suspended on a member when there is a RAID expansion/reconstruction occurring.
- Page swapping will also be suspended if all pool members have latencies of 100ms or greater.
- Page swapping is not helpful when there is high member latency but there is no hot or warm pages.
- This balancing can be disabled via the Group Management GUI.
- Hot and cold page swapping will not create additional volume slices as this violates the capacity balancer rules.

The Hot and cold page swapping algorithm is very similar to the Page swapping that occurs within the hybrid model array (PS6000 XVS & PS6010 XVS).
Member Eligibility
Not all pool members will be considered for hot/cold page swapping. Members undergoing RAID verification or RAID expansion will not be eligible to participate in hot/cold page swapping. However, hot/warm pages will be offloaded from degraded pool members.

Volume Eligibility
Pages from bound volumes will not be moved for hot/cold page swapping. This prevents page swapping from circumventing customer configuration.

Choosing a Member to Take on Hot Pages
Two member related parameters are considered when choosing the pool member best suited to consume hot pages. The first parameter is member latency and the second is a measure of the member’s utilization. The pool member with the best latency value and most headroom will be selected to take the hot pages. Note that pool members with latencies greater than 100 ms or headroom less than 5% will not be considered eligible to take hot pages.

Cold, Warm, and Hot Page Determination
Currently, cache is responsible for maintaining the statistics required to ascertain the hotness of a page. For each disk page, cache keeps a hotness counter. The hotness counter is not an I/O count. Instead, it is a representation of the frequency of page accesses within the last ten minutes.

Hybrid Array
When selecting hot pages for swapping on a hybrid array, pages will be chosen in a slightly different manner. For the hybrid, hot/warm pages residing on disk will be used for page swapping. If there are no hot/warm pages on disk, then hot/warm pages residing on SSD will be selected for page swapping.

Performance Thrashing
A pool member transitioning from offloading hot pages to being able to take on additional hot pages within a short period of time, is an indication that hot/cold page swapping is not helping pool member
latencies to converge. In this situation, page swapping will stop for 10 minutes and then resume with normal operation.
What/ When Page Balance

• What pages need to be moved?
  – This is determined by the number of reads and writes on a particular page in a given time set, using the current XVS hybrid model but expanded to the pool level
  – 10 pages are moved at a time

• The “cold” page is chosen in one of two ways:
  – A “cold” page will be chosen first if it is “unallocated” or not written to.
  – The second option is a “cold” page that has not been accessed in a long time

• When is Hot & Cold page swapping put on hold?
  – The pool could possibly be overloaded
    ➢ All pool members have latencies 100ms or greater
  – A pool member quickly transitioning from being overloaded to being eligible to take on a load could indicate that some level of performance “thrashing” is going on in the pool.
    ➢ The page swapping will pause for 10 minutes and then resume and will repeat as necessary.
Observing Automatic Performance Load Balancing

Member A

Group performance results

Latency comes down

Member B

Avg. IOPs goes up

Queue depth reduced

Dell internal testing May 2021 - Group setup: a single pool containing 4 arrays, PS6500 I/O (fastest), PS6000 SX (second fastest) and PS6000E (slowest)
The members that are identified within the swap is defined by the member latency and the “Hardware Score”.
- The “Hardware Score” is based upon the hardware of the member, for example, SATA vs. SAS drives.

APBL is disabled via a checkbox in the advanced tab of the Group Configuration screen.
- Balancing is enabled by default

APBL does not interfere with the member free space
The hybrid array (PS6100xs) has Solid State Drives (SSD) and 10K SAS Hard Disk Drives (HDD) inside the chassis. When a volume is created on an XS system the volume will be spread across both the solid state drives and spinning disks. If there are multi-members within the pool, the volume will be spread across all members in the pool as it is currently done for non XS systems.

The XS system optimizes the performance of tiered application workloads by intelligently placing data across SSD and 10K RPM SAS drives based on when and how often the data is accessed. Data accessed more often is placed on the SSD and data accessed less often is placed on the spinning disks, this delivers excellent user responsiveness for tiered workloads.

Data is located on internal space called pages. Pages are classified inside the chassis by how often the data is accessed. The classifications are:
- **Hot page** - pages that contain data that has been access often
- **Warm page** - pages that were hot but are no longer hot
- **Cold page** - pages that have never been accessed

The advanced load balancing process operates on the I/O load to the actual pages. The process tracks the capacity in use by access frequency, and categorizes it into one of three classes, high I/O or Hot, medium I/O or Warm, and low I/O or Cold. Based on this categorization, the XS is then be able to evaluate the data residing on the two tiers of storage (SSD and 10K HD) within the array and determine if there is some part of the data on the HDD that would be better served by the SSD. If so, the data currently residing on the SSD tier is evaluated to determine if there is low I/O (cold) data that can be relocated to the HDD tier to make space available for the high I/O data. The XS will adjust the placement of the data accordingly. This allows in most cases for the more active data to be located in SSD and the less active to be on HDD. The action of moving pages form SSD to HDD will only occur within an array, not between arrays. So if there are multiple XS arrays in a pool and volume is spread across both arrays the data movement mentioned above is handled within each array of one another.

Another attribute of the XS system is that a portion of the SSD is allocated as accelerated cache for random writes. This is used in instances where blocks of data relating to a number of random writes can be directed to the accelerated cache and moved to SSD quickly, this is what we call "RAID 6 Accelerated".

---

The image explains the data movement process and categorization of pages in the XS system. The diagram illustrates the automatic and transparent load balancing across arrays, with classifications for hot, warm, and cold pages, and the process of moving data between SSD and HDD tiers based on access frequency.

---

PS Series Hardware Architecture

---

3-18
In PS groups there are many volume operations that consume and release free space dynamically

- Volumes grow/shrink – thin provisioning (map/unmap)
- Snapshots space grows/shrinks – create / delete, space grows as writes occur to volumes or other snaps
- Replication – recovery points, freeze of data for transmission

In multi-member pools, this can cause free space imbalances between members in the pool

- Data changes more quickly on faster members typically (e.g. consuming snapshot space faster)
  - **Free Space balancing** adjusts this in background shifting in-use and free pages between members

When capacity gets low, the member enters *free space trouble* state.

- In worst case scenarios, these imbalances can affect the dynamic operations
- When in free space trouble state, the load balancer works to more rapidly free space on member that is running low by swapping in use pages for free pages with other members.
- In pools with > 3 members it may also change the slice layout amongst members
Questions?