07 Storage Reliability Technologies

www.huawei.com

HUAWEI

Foreword

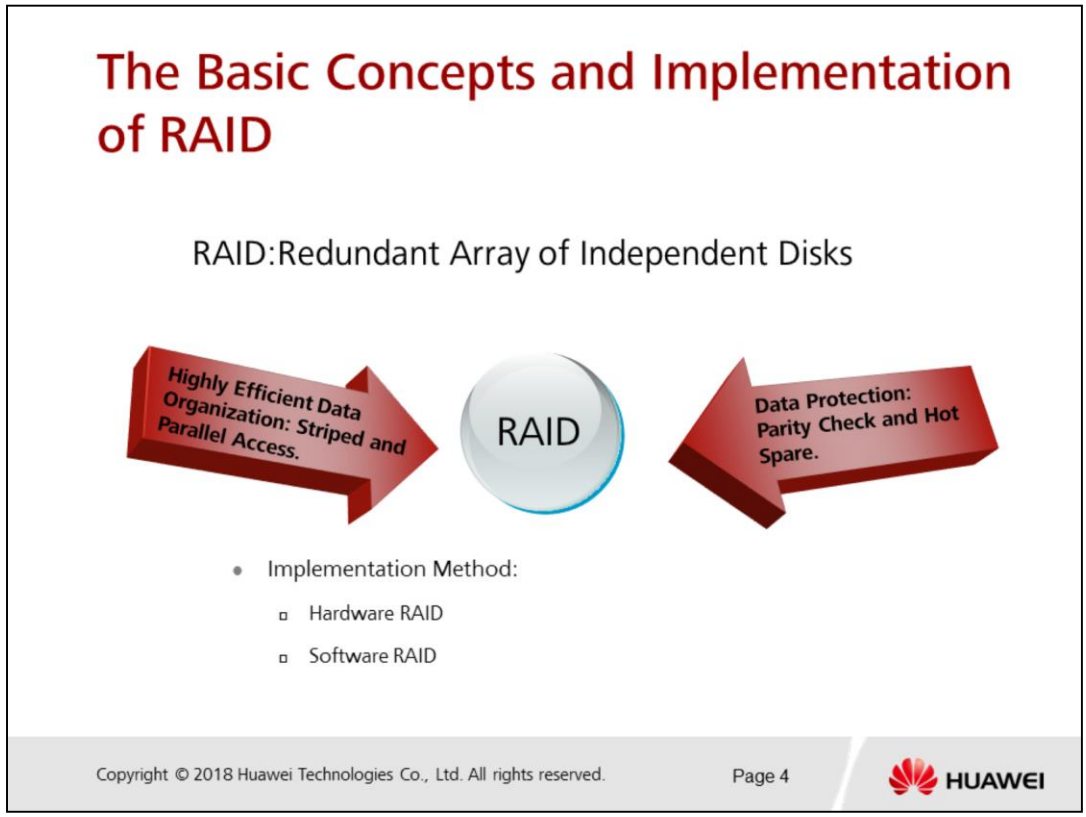
- This module mainly describes the traditional RAID technologies, RAID 2.0+ technologies, Multipathing technologies and disk reliability technologies.
- This module also describes the functions of these technologies towards the aspect of data protection.

Objectives

- Upon completion of this module, you will be able to:
  - Describe the concepts, principles and types of RAID technologies.
  - Describe the principles of RAID 2.0+ technologies.
  - Understand the Multipathing technologies of hosts.
  - Understand the technologies implemented for ensuring the reliability of hard disk.

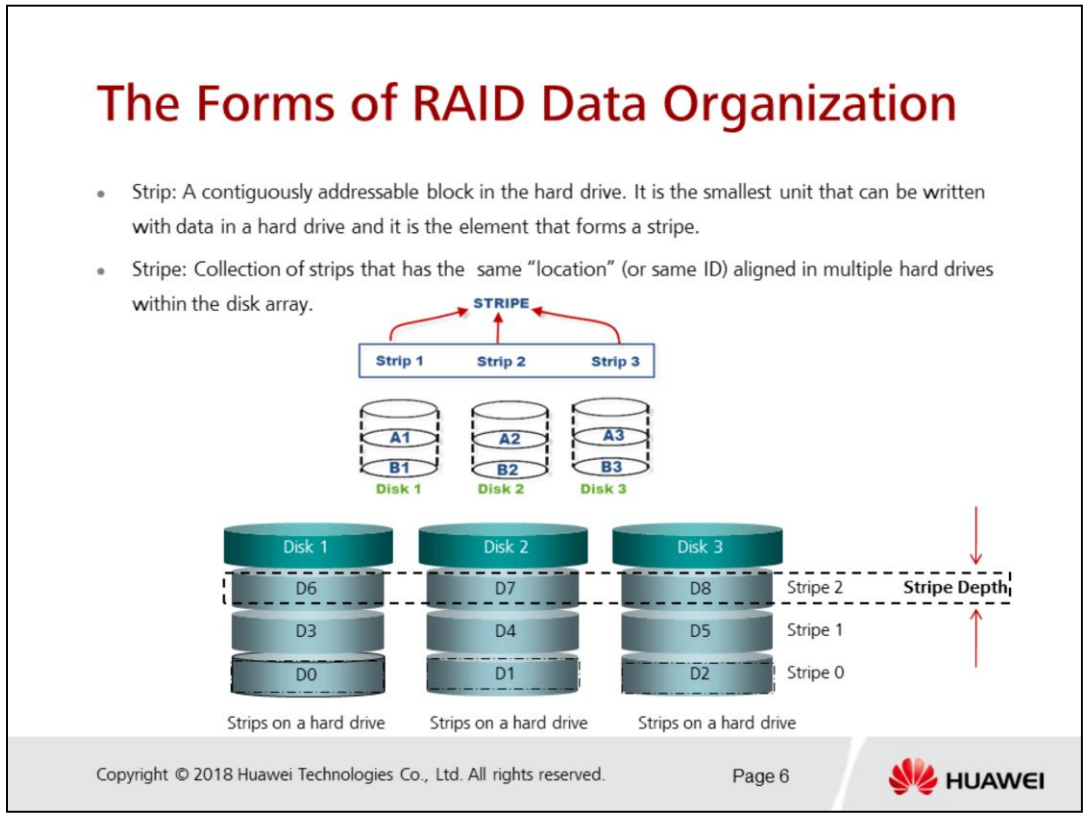
Contents

HUAWEI



- The initial purpose of RAID technology was to combine multiple hard disks with small capacities in order to obtain a larger storage space. The RAID technologies that we will be mentioning in this module is more related towards data protection. In other words, when the physical devices fails, RAID is able to prevent the loss of data.
- The main functions of RAID technologies:
  - By striping the data on the hard disk to achieve block access of data, it reduces the mechanical seek time of the drive and improves the data access speed.
  - By reading the data simultaneously from few drives in the array (Parallel Access), it reduces the total mechanical seek time of the drives and improves the data access speed.
  - By mirroring or parity check methods, it achieves redundant data protection in the drives.
- With the development of array technologies, it produced a lot of different types of RAID. However, only a few types of RAID are still in use actively. In this module, we will be discussing the most common types of RAID and learn the other related functions of RAID such as data protection. At the same time, we will learn on why choosing different types of RAID means that there will be a difference in the performance or cost.

- In the storage devices, there are two methods to implement the RAID feature, which are: Hardware RAID and Software RAID.
    - Hardware RAID uses dedicated RAID controllers, hard disk controllers or storage processors. RAID controllers have their own processors, I/O processing chips, and RAM in order to increase the resource utilization rate and data transmission speed. RAID controllers manages the data routing, data buffer and controls the data flow between the host and the RAID. Hardware RAID are commonly used in servers.
    - Software RAID has no dedicated processor or I/O processing chips, and is fully reliant on the host processor. Hence, low speed CPU is unable to meet the requirements of RAID implementation. Software RAID is commonly used in enterprise grade storage devices.



- Stripe Width:
  - Stripe width is the count of the number of member disk in a stripe. In simple words, stripe width refers to the number of parallel stripes that can be written to or read from simultaneously. This is equal to the number of disks in the array, so a four-disk striped array would have a stripe width of four.
- Stripe Depth:
  - Stripe depth refers to the storage capacity of a stripe in each individual disk in the array. For an example, the stripe depth a stripe size of 128KB across two hard disk in RAID 0 configuration will be 128KB/2=64KB.
  - In other forms of RAID that has parity involved, please note that you need to exclude the parity disk and only include the data disks before calculating the stripe depth. For example, a RAID 5 configuration with 3 hard disks with a stripe size of 128KB, will be 128KB/2=64KB. It is divided by 2 because the sum of storage capacity equivalent to 1 disk is used for parity (For RAID 5 with 3 disks configuration every 2 blocks of Data, there is 1 block of Parity Data, D0+D1+P0).

RAID Data Protection Methods

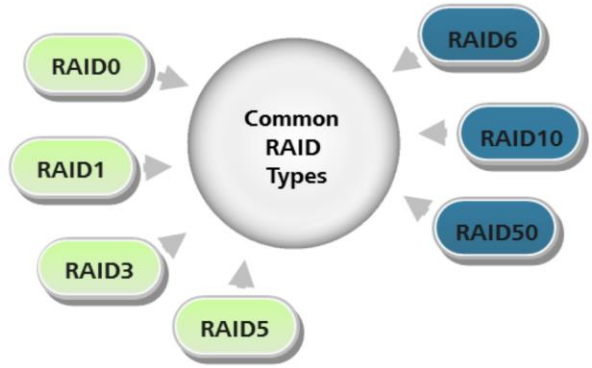
- Method 1: Save a copy of the data on another redundant hard drive.
- Method 2: Parity Check Calculation (XOR).
  - XOR computation is widely used in digital and computing science.
  - XOR computation--If both is equal then it is false, if both is different then it is true.
  - 0⊕0= 0; 0⊕1= 1; 1⊕0= 1; 1⊕1= 0;

XOR Parity Check Creates A Redundant Backup of Data.

- RAID technologies usually have 2 types of methods for data protection. One of the method is to store a backup of the data on another redundant hard drive(mirroring). The second method is by using XOR computation to calculate the parity of the data. The XOR computation will use the user data to calculate the additional data which is the parity that can be used to restore data in the event if one of the drives in the array fails. For RAID types that uses XOR computation for parity check, it means that there is a need for additional parity drives.
- A very simple Boolean operation is used at the binary level to create RAID parity. This operation is the Exclusive Disjunction operation also known as Exclusive OR (XOR). By using XOR, the raw binary data (1 or 0) is passed through an operation that results in a final binary result, which can be used for redundancy and error correction.
- The table for the operation of XOR on raw binary data is shown on the diagram above.

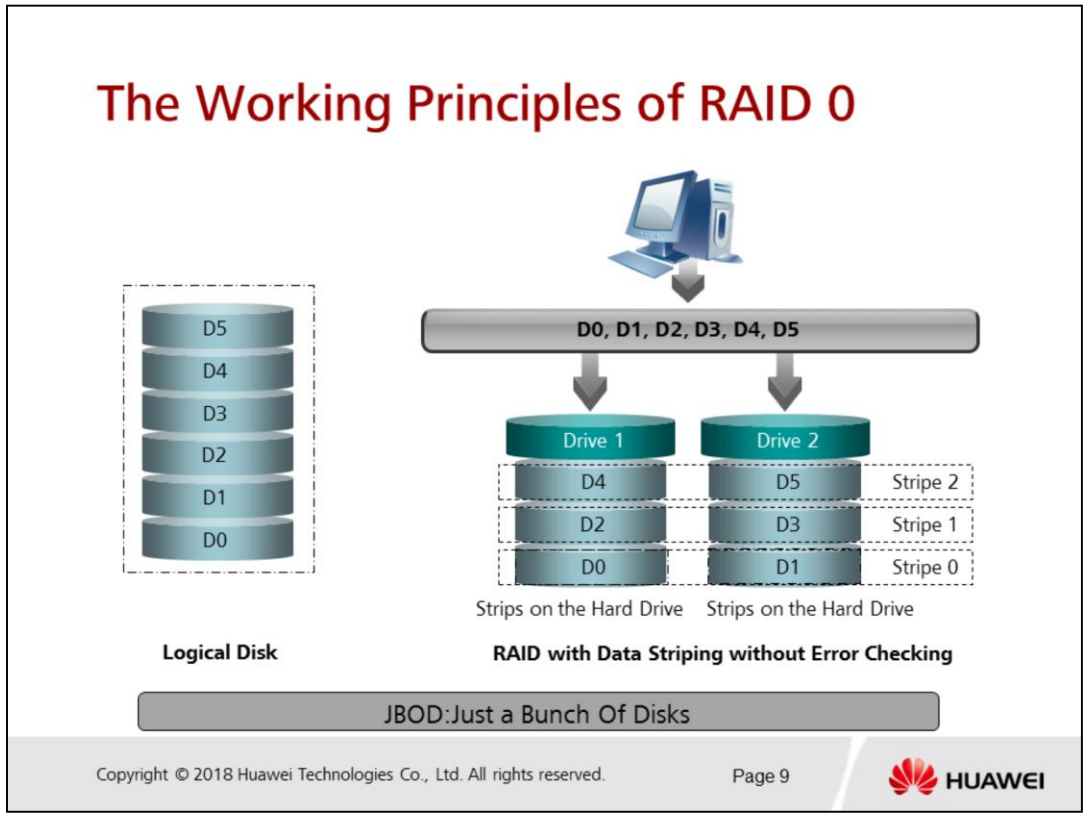
Common RAID Types and Classification Standards

- RAID technologies combines multiple independent physical disks into a logical disk using different methods, which increases the disks read/write performance and data security. The RAID types can be classified based on the methods used in the combination of the disks into a RAID array.



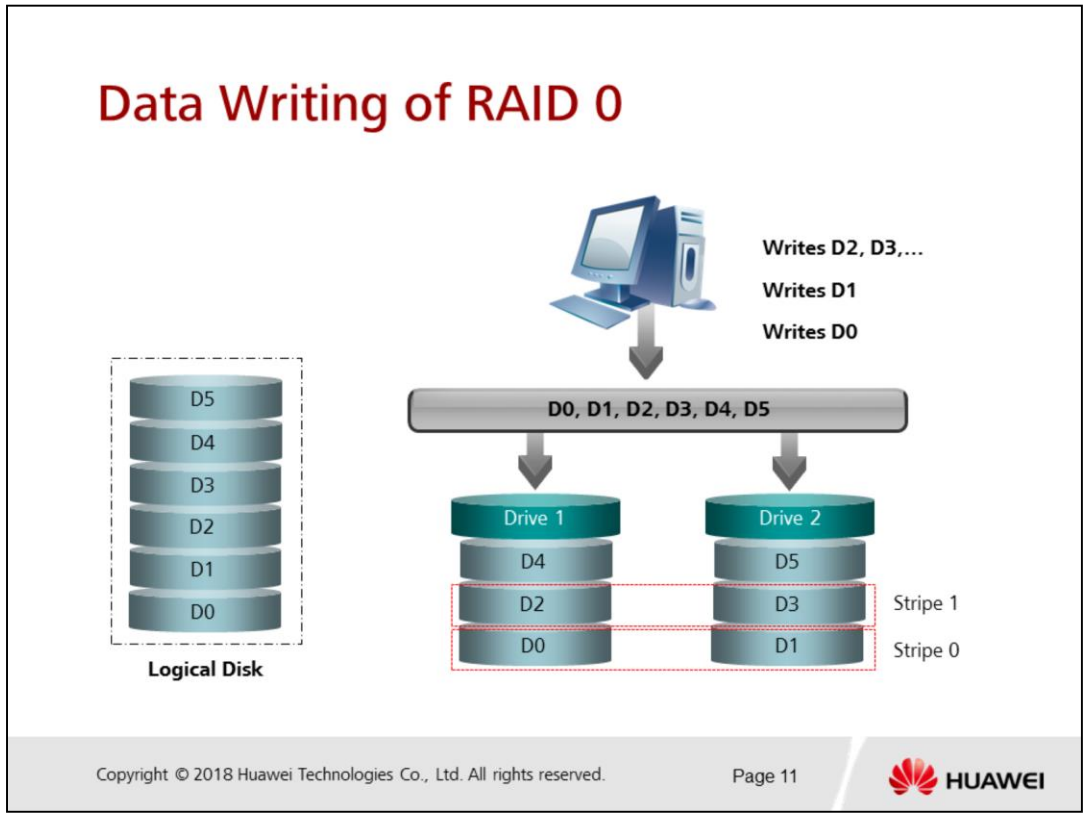
HUAWEI

- In summary, the advantages of RAID technologies can be shown in the following aspects:
  - Provides larger storage capacity by combining multiple hard disks into a logical array.
  - Data is cut up into data blocks, which allows parallel read/writes across multiple hard drives, and increases the data access speed in disks.
  - Provides the fault tolerance through mirroring or parity checks.

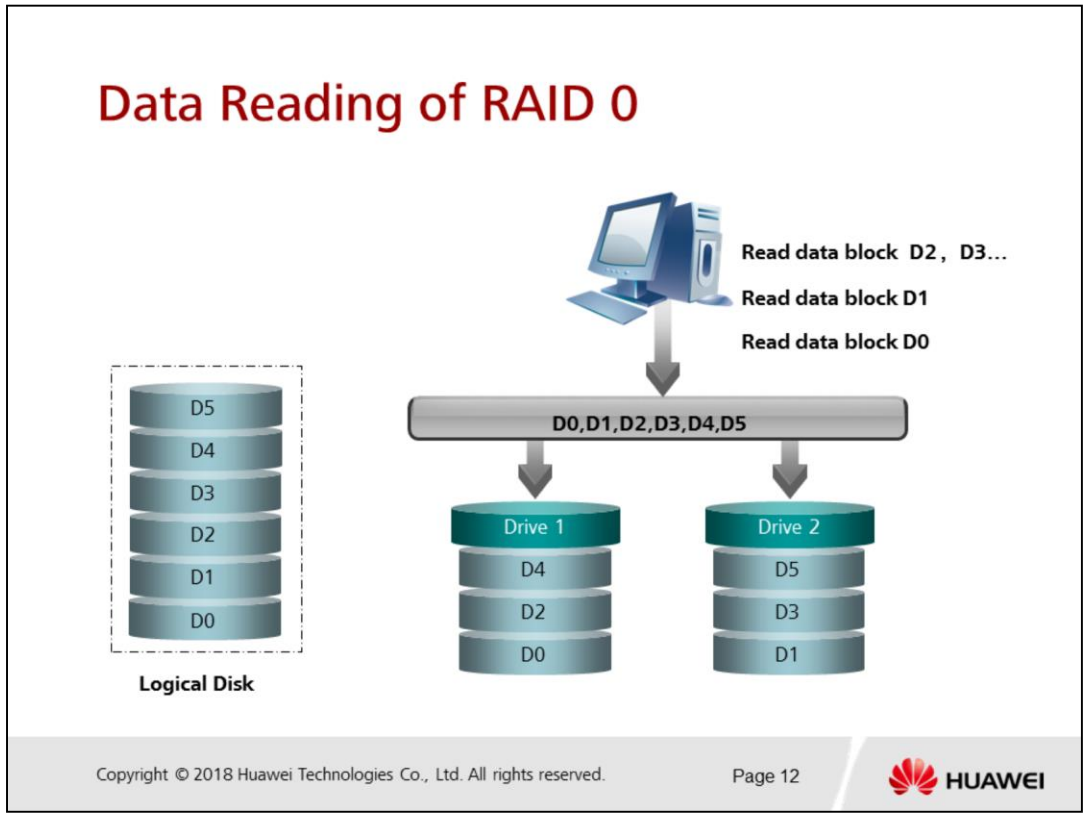


- Among all the RAID types, RAID 0 (also known as Striped RAID) has the highest storage performance. RAID 0 uses striping technologies to distribute the data across all the disks in the RAID group within the array.
- A RAID 0 implementation has a minimum of 2 disks. RAID 0 groups strips data of into data blocks of varying sizes ranging from 512 bytes to megabytes (usually it is the multiple of 512 bytes) and writes them into different drives. As shown on the two disk that forms RAID 0 at the diagram above: The first 2 data blocks are written into stripe 0, in which, the first data block is written in the stripe 0 of disk 1, and the second data block is written in parallel at the stripe 0 of disk 2. The next data block is written at stripe 1 of disk 1 and so on. Using this method, the workload of I/O is distributed and balanced across all the disks. Due to the higher speed of the data transmission bus compared to the disk read/write speeds, it can be considered that the drives within the RAID group can read/write simultaneously in parallel.

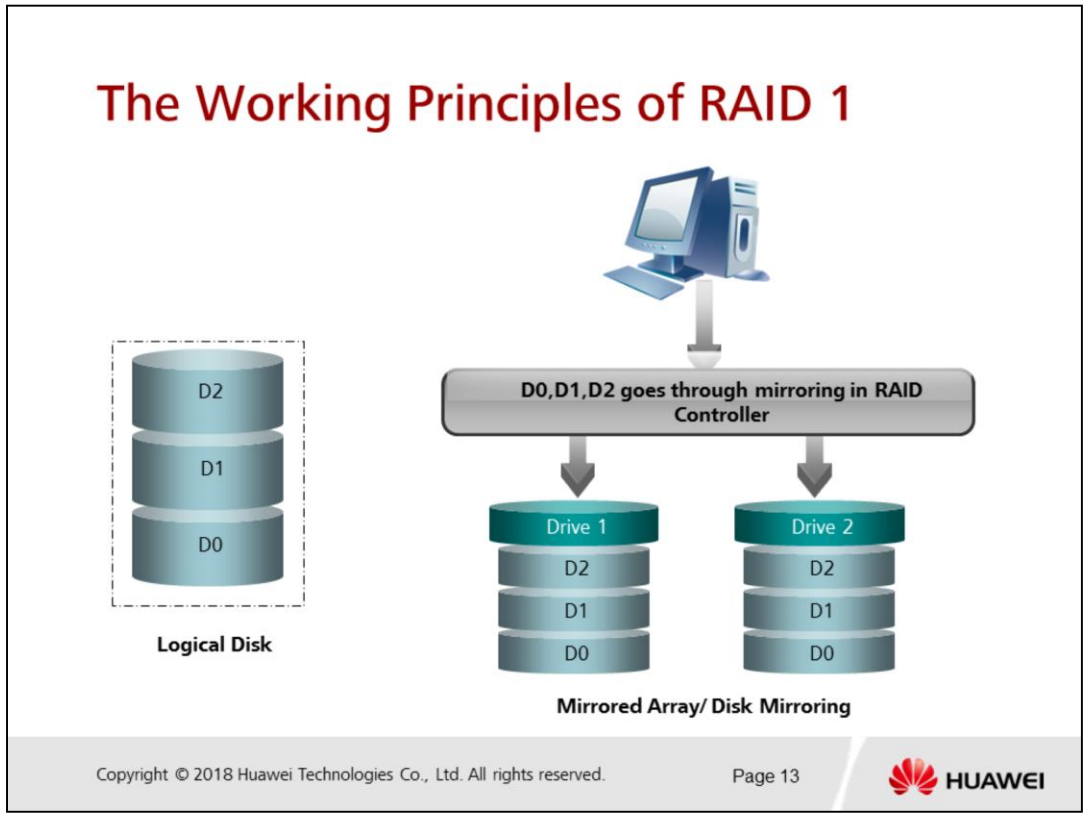
- The disks within a RAID 0 group must have the same capacity and same rotational speed. If a RAID 0 is created using 4 disks, in theory the read/write rate is 4 times compared to a single drive (in reality, there would be some system loss that brings down the actual speed). The storage capacity will be 4 times compared to a single drive. If the capacity and the rotational speed is different within the disk in a RAID 0 group, the usable storage capacity and speed will be 4 times of the smallest capacity and lowest performance drive respectively.
- RAID 0 has the unique feature of providing a single logical drive with large capacity and have very fast I/O at the same time. Before RAID 0 technologies were in use, another similar technology called JBOD exists. JBOD (Just a Bunch Of Disks), forms a logical drive by combining multiple hard disks. The main difference between JBOD and RAID 0 is that a JBOD data block cannot be written in parallel to different hard disks. In JBOD, only when the storage space of the first hard disk is fully used up, then the second hard disk will be used to store data. Hence, the total capacity of JBOD is the sum of all the storage capacity of the drives connected, but the performance of the system is only equal to 1 drive.



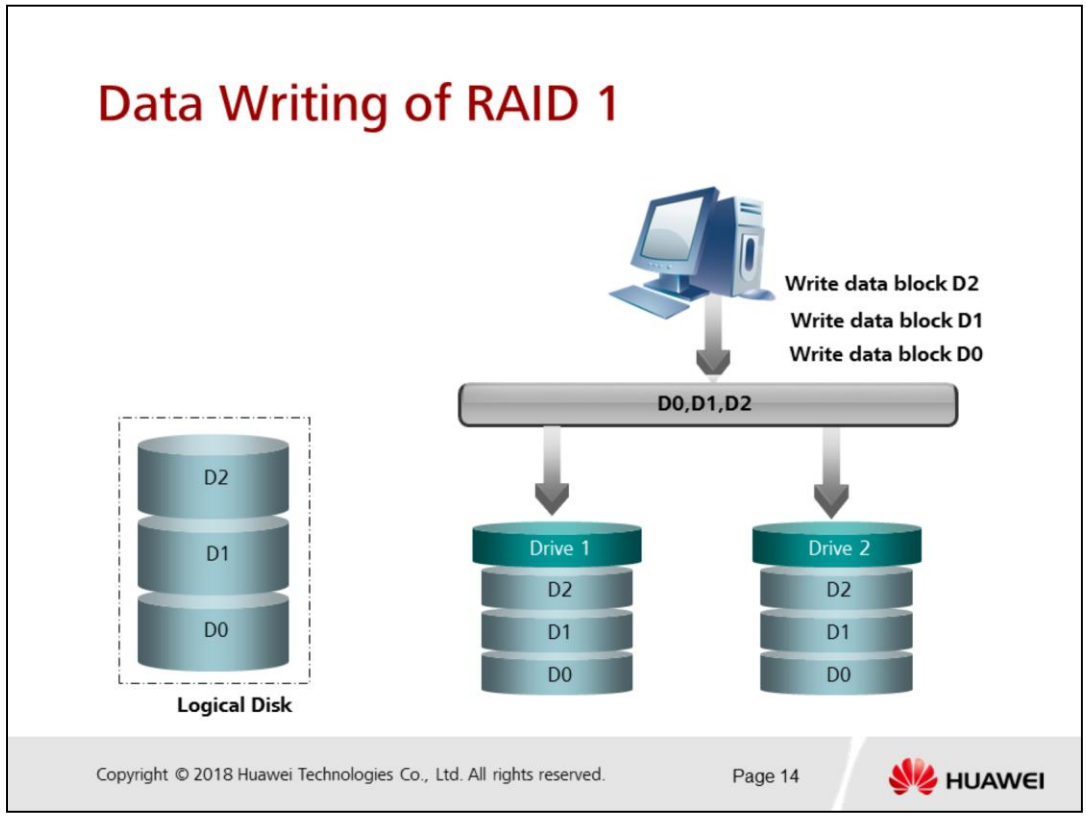
- RAID 0 uses data striping technologies to write data into the disk groups, it divides the data into data blocks, and evenly distribute them across all the hard disks within the RAID group. Only when the previous stripe is fully written by data, then the data is written to another stripe. As shown in the diagram above, data blocks D0, D1, D2, D3, D4, D5 is waiting to be written into a RAID 0 disk group. D0 is written into the first stripe (Stripe 0) into the first drive (Drive 1), and D2 is written to the first stripe in drive 2 and so on. All the remaining data is written to the disk group using the data striping method in a RAID 0 disk group.
- The data write performance of RAID 0 is directly proportional to the number of disks within the disk group. In simpler words, the more number of disk are in the RAID 0 disk group, the higher the data write performance.



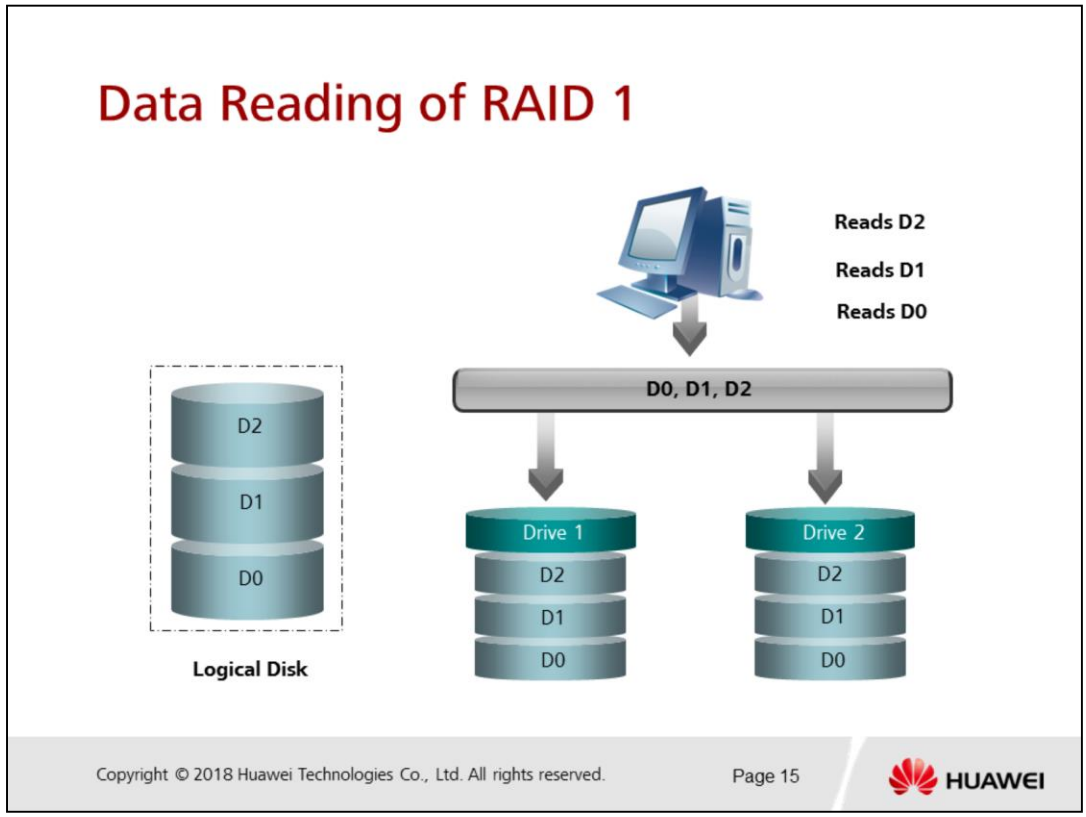
- When RAID 0 receives a data read request, it will search and retrieve the targeted data from all the disks within the RAID group. In the diagram above, we can see the whole data read process.
- Firstly, the array receives the read request for data blocks D0, D1, D2, D3, D4, D5. Subsequently, the array reads the data in parallel, which is reading D0 from drive 1 and D1 from drive 2, and so on for all the other data. When all the data blocks are read from RAID, they are integrated by the RAID controllers to form a complete data and then sent to the host. Since, RAID 0 is using data striping technologies where data is split across drives, the data blocks must be recombined to form the complete unstriped data before it is sent to the host.
- Similar to data writing, RAID 0 data read performance is directly proportional to the number of disks within the RAID group.



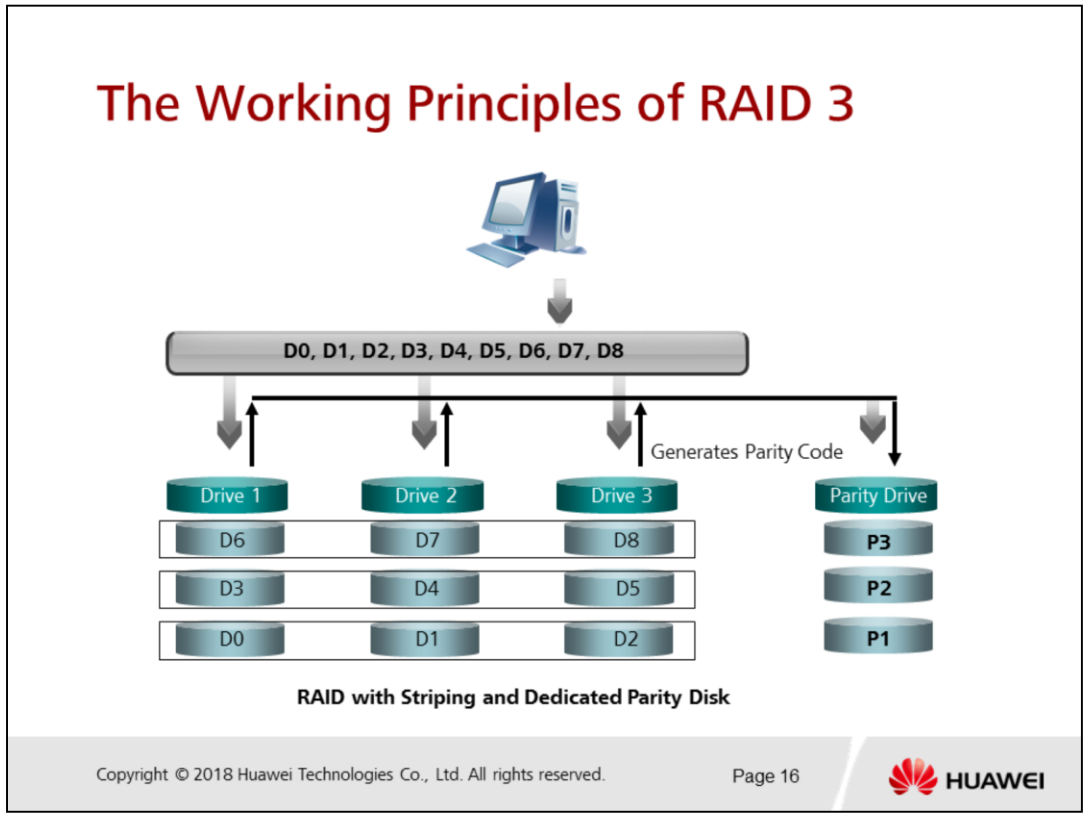
- RAID 1 (also known as mirrored disk array) is a type of RAID that is built on top of the principles of high safety. RAID 1 system uses 2 identical hard drives that is configured with mirroring. When data is written to one of the drives, the data copy will be stored simultaneously on the mirrored drive. When the source drive (physical drive) is faulty, the mirrored drive will takeover services to ensure service continuity. The mirrored drive serves as backup, and provides high data reliability.
- A RAID 1 disk group data storage capacity is only equal to the single disk capacity, the other disk stores the copies of the data, which means that for every 1 Gigabyte of data, it consumes 2 Gigabytes of storage space. Hence, the RAID 1 system that consist of 2 drives only has the space utilization rate of 50%.
- The two disks used for RAID 1 need to be identical in storage size. If the storage capacity of the drives are different, the usable storage capacity will be based on the smaller drive.



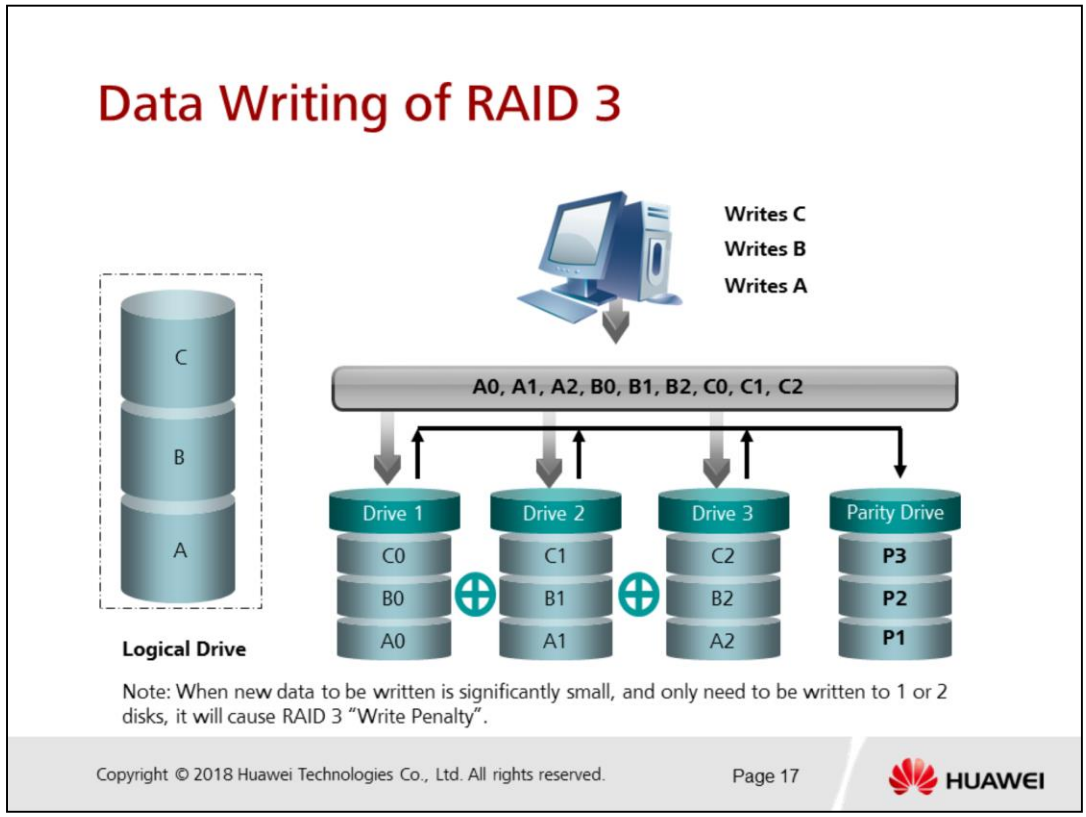
- RAID 0 uses data striping technologies to write different data blocks in parallel into multiple drives, but in contrast, RAID 1 writes the same data blocks to each disks within the RAID array. Data on all the member disks are the same. As shown on diagram above, data blocks D0, D1, D2 are waiting to be written to the drives. D0 and copy of D0 is written to both of the hard drives (Drive 1 and Drive 2) at the same time, and the other data uses the same method (mirroring) to be written to drives within the RAID 1 disk groups.
- Generally, a RAID 1 write performance is equal to the write performance of a single disk.



- When RAID 1 reads data, it will read from the data disk and mirrored disk simultaneously, which increases the reading performance. If one of the drives fails, the data can be read from the other drive.
- RAID 1 read performance is equal to the sum of the performance of both disk. But the write performance is only 50% since each data need to be written twice.

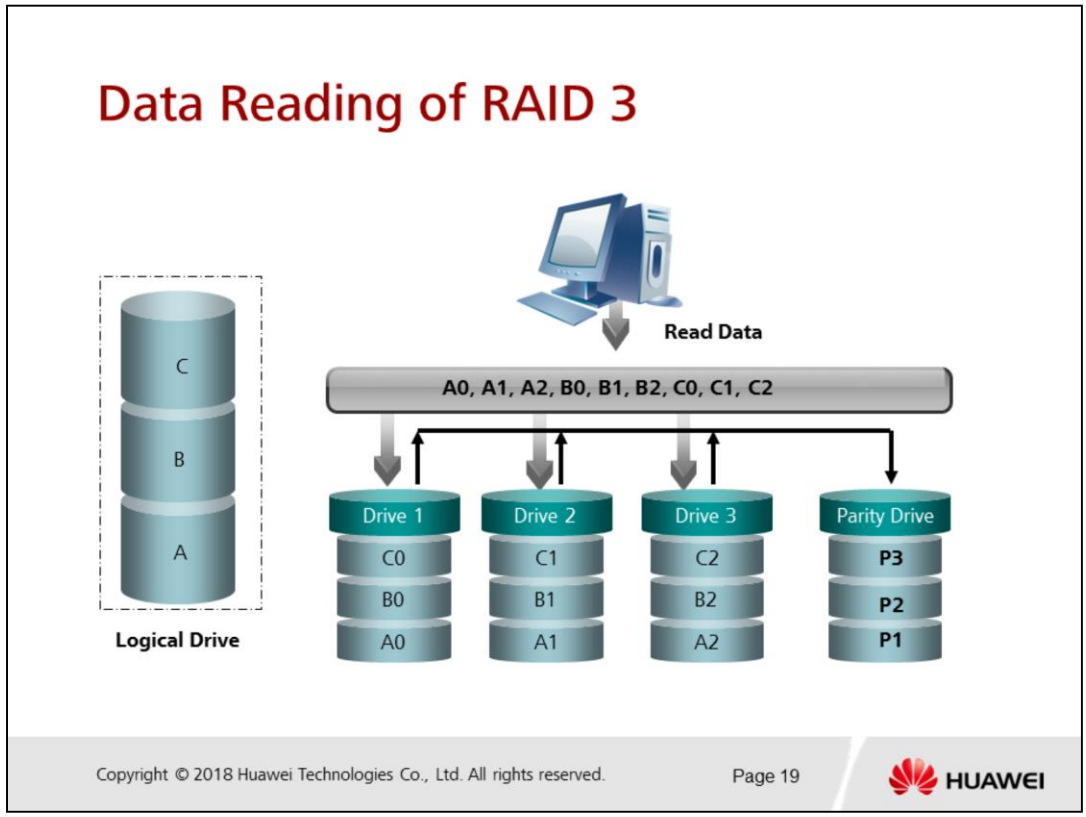


- RAID 3 is similar to RAID 0, but the difference lies with the fact that RAID 3 has a dedicated parity disk that stores parity information of the data. The dedicated parity disk stores the parity information of the data stripe. If any data error is detected or the drive is faulty, we can use the parity information to restore the data on the faulty drive. RAID 3 is suitable for data with high density or single user environments, that require long sequential read or write access to data such as uncompressed video editing or streaming media. RAID 3 assigns the data write operations to its data disk members of the RAID group. However, when there is new data to be written, no matter if it is written to any of the data disks, RAID 3 needs to recalculate and rewrite the parity information. Hence, when some applications requires large amount of data writing, the parity disk in RAID 3 will have high workload. Due to the fact that there is a wait time for the parity check, the RAID 3 read and write performance is affected to a certain level. Additionally, due to the high workload of the parity disk, it is commonly the disk that is most likely to fail within a RAID 3 system. This is why the parity disk is often considered as the bottleneck of RAID 3 systems.
- Assuming the number of disks in a RAID 3 system as N, the usable storage capacity will be N-1 (deducting the storage capacity of the 1 drive used as parity disk). Similar to the other RAID types, the member disks within the RAID group need to have the same storage capacity and rotational speed for optimum performance.

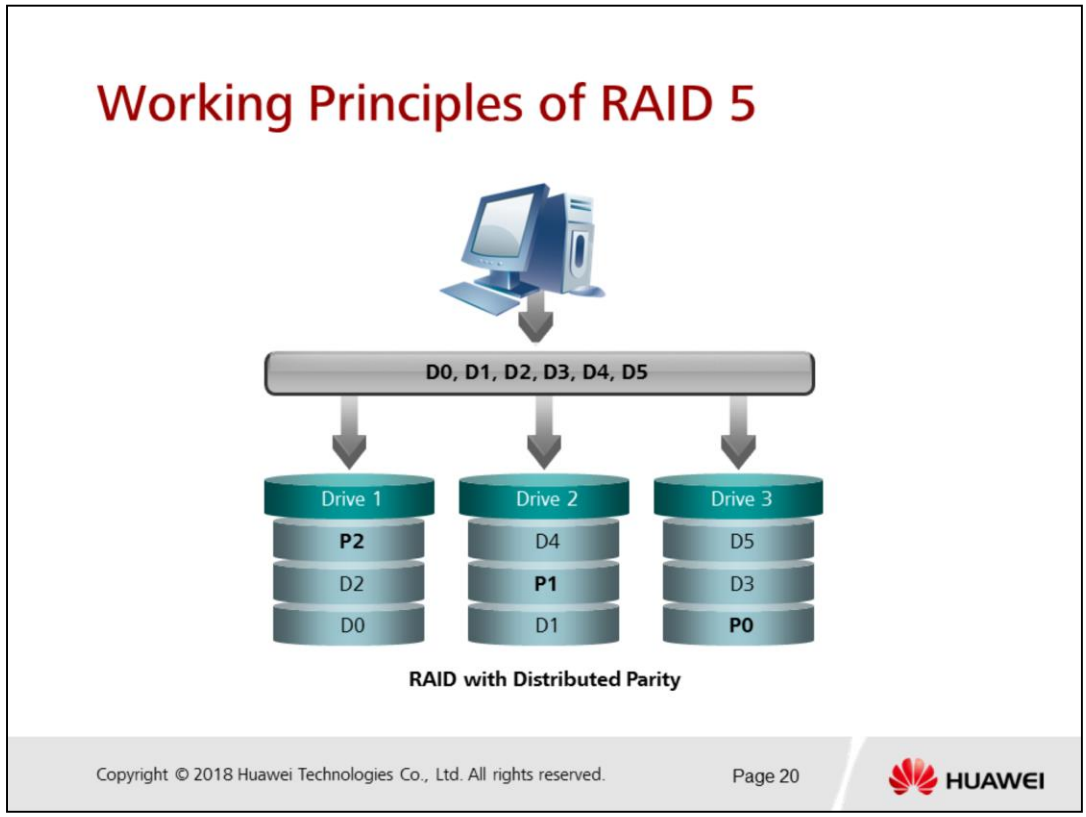


- RAID 3 implements single disk fault tolerance and parallel data transmission. In other words, RAID 3 uses data striping technologies to split up data into data blocks, runs XOR calculations to generate the parity information that is written into the dedicated parity drive. When a drive is faulty, the data is written to the drives that are not faulty and the parity check continues.
- RAID 3 performance is not fixed. In principle, RAID 3 uses N+1 data protection method. This means that when N number of drives containing user data that needs to be protected, you will need a extra 1 drive to store parity information. In this scenario, when a new data is written to the disks, parity information will be calculated and written to the parity disk.

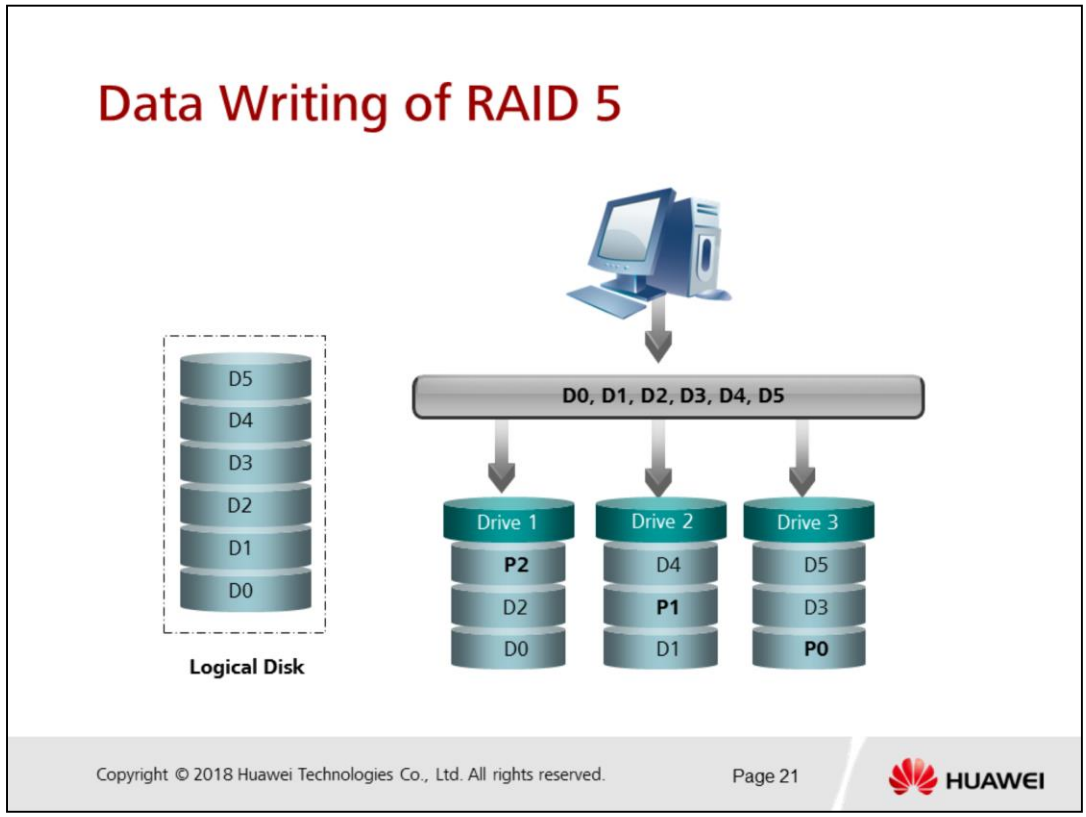
- In normal circumstances, all the member disks in the RAID group will cooperate in the data striping process, which means that N number of drives will be written in parallel. But when the amount of data to be written is very small, where it is only needed to be written to one or two of the drives, based on RAID 3 working principles, it still need to read all the disks to recalculate the parity value. This scenario where small amount of data is written but requires extra overhead on read and write operations compared to a data writing operation to a single drive, which does not increase the performance of the system is called as RAID 3 Write Penalty. For example, a single byte of data need to be written to a RAID 3 system, but the total overhead for the operation is 4 bytes assuming all operations only require 1 byte (Write the actual data of 1 byte in disk 1, read 1 byte of data on disk 2, read 1 byte of data in disk 3, calculate and write 1 byte of parity in the parity disk).
- RAID 3 write performance depends on the amount of data, the number of disk drives, and the time required to calculate and store the parity information. Assuming the number of disk drives in RAID 3 as N, when all of the member disk has the same rotational speed, in the scenario of full data striping without considering write penalty, the RAID 3 sequential IO write performance in theory will be slightly smaller than N-1 times the performance of a single drive (calculation of redundant check requires additional computing time).



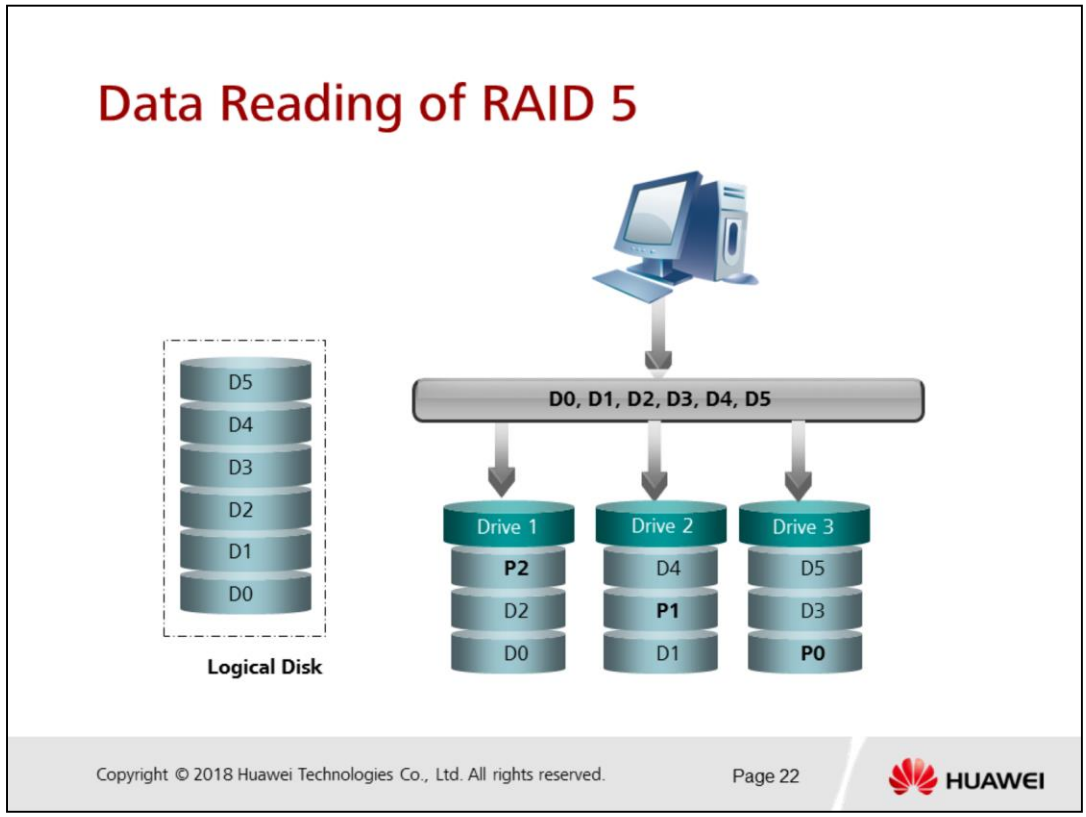
- In RAID 3, data are read in stripe method. Each drive in the RAID is controlled by the disk driver, so the data blocks in the same stripe can be read in parallel. Hence, each of the drives in the RAID 3 system is utilized during the data reading operation, which increases the red performance.
- RAID 3 uses parallel data read(write) mode.
- RAID 3 data read performance depends on the amount of data to be read and the number of disk drives within the RAID 3 array.



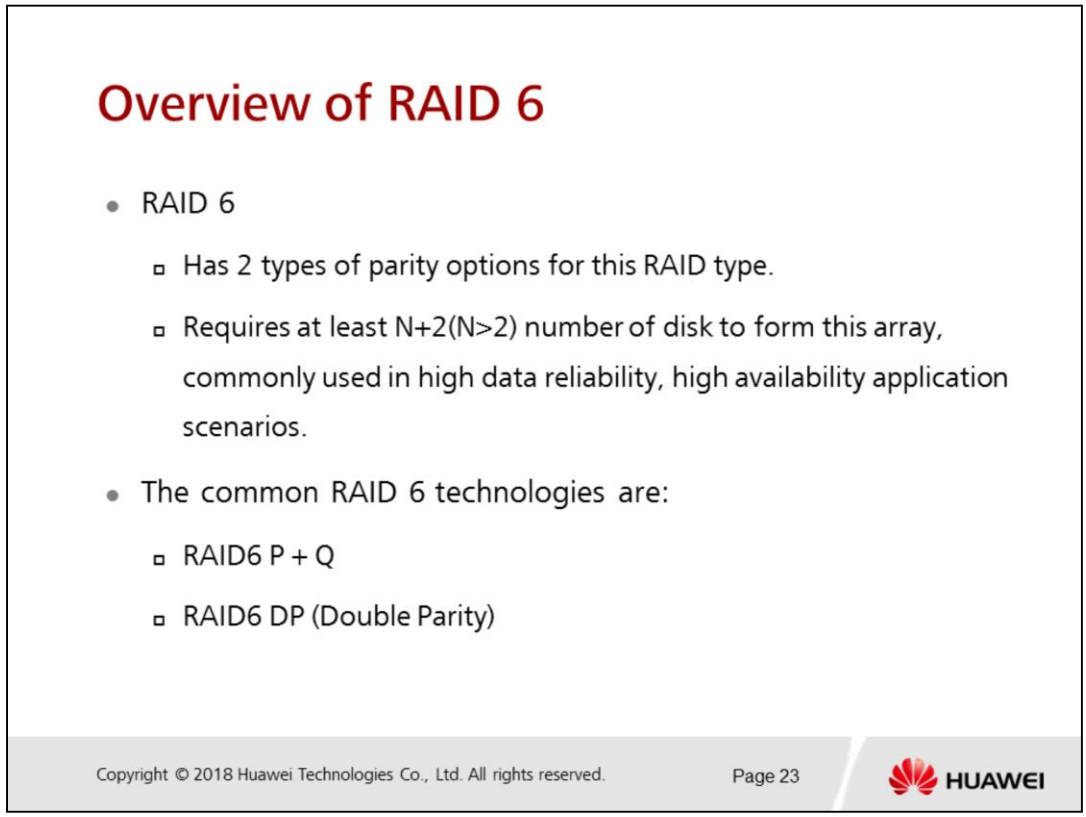
- RAID 5 is the improved version of RAID3, and it uses data striping, calculates and store parity information. RAID has a dedicated parity drive where all the parity information is read and written, this also causes the performance bottleneck of RAID 3 that we have mentioned previously. RAID 5 uses distributed parity, where each member disk in the RAID group will store data and parity information. So, RAID 5 do not have a bottleneck or a hot spot like RAID 3.
- Assuming the number of drives in a RAID 5 group is N, the usable data storage capacity is equal to N-1. Similar to the other RAID, all the member disks in the RAID 5 array need to have the same disk capacity and rotational speed.
- In the disk array of RAID 3 and RAID 5 types, if a disk fails, the RAID group will change from Online state to Failed state, until the data on the faulty disk is fully reconstructed. If another disks in the RAID failed at the same time, then all the data in the RAID array will be lost.



- In RAID 5, data is written in the form of stripes into the disk groups. Each member disk of the RAID 5 group stores data and parity information, when data is written in stripes, the parity information will be written in the corresponding disk. When RAID 5 is continuously writing data, different stripes uses different disks to store parity information. Hence, the parity information in RAID 5 is not saved in a fixed dedicated parity disk, but it is distributed and stored across the disks based on certain logics.
- When writing small amount of data, RAID 5 has write penalty similar to RAID 3.
- RAID 5 write performance depends on the amount of data written and the amount of disks in the RAID 5 array. Assuming the number of disks in a RAID 5 array is N, in the scenario where all the member disks has same rotational speed and without considering write penalty, the RAID 5 sequential IO write performance in theory will be slightly smaller than N-1 times the performance of a single drive (calculation of redundant check requires additional computing time).



- Data within a RAID 5 array is stored in the form of stripes on the disks. Only N-1 number of data disks is required to fully restore all the data. It has the fault tolerance of a single drive, meaning that one of the hard disk can fail without losing data, but if more than 1 fails at the same time then the data can not be reconstructed.
- The data read performance of RAID 5 depends on the amount of data and the number of disks within the RAID group.

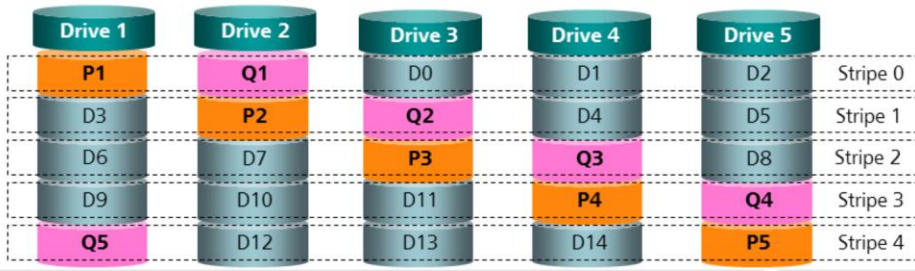


- The RAID group that we have mentioned earlier in this module only considered the scenario where only a single disk has failed (excluding RAID 0). Nowadays, the storage capacity has greatly increased, and the disk reconstruction has also increased. A RAID 5 system that is formed with a lot of large capacity drives may need a few days and not few hours to reconstruct the data of the failed drives. During reconstruction, the system is in the downgraded state, which means that in this situation, any additional drive failures will cause the array to fail and the data is lost. This is why some organization or enterprises requires a dual redundancy system. In other words: A RAID group should allow 2 drives to fail and at the same time all the data is still accessible. This type of dual redundancy data protection methods usually can be implemented in the following ways:
    - The first method is multi mirroring. Multi mirroring refers to the data stored in the main disk is stored in multiple copies to additional redundant hard drives. This method involves a very large expenditure and cost.
    - The second method is RAID 6 array. RAID 6 array provides data protection for up to 2 disk failures at the same time.

- RAID 6 is officially called distributed double parity RAID. In essence, it is just an improved version of RAID 5, which has the features of data striping and distributed parity. Currently, RAID 6 has dual parity, which means two things:
    - Firstly, when user data is written, the additional double parity calculation need to be run at the same time. Hence, among all the RAID types, RAID 6 can be considered as the “slowest”.
    - Secondly, the additional parity information requires the disk space of 2 disk drives. This is why RAID 6 is considered as N+2 type of RAID.
- For now, RAID 6 does not have a unified standard. Different vendors uses different methods to implement RAID 6. The following are the 2 most common implementation of RAID 6:
    - RAID P+Q : Implemented by Huawei, Hitachi Data Systems(HDS).
    - RAID DP : Implemented by NetApp.
- These 2 types of RAID 6 models uses different methods of data parity. But, both of them are able to protect data completeness and support data access during the scenario of 2 disk failures within the RAID group.

Working Principles of RAID 6 P+Q

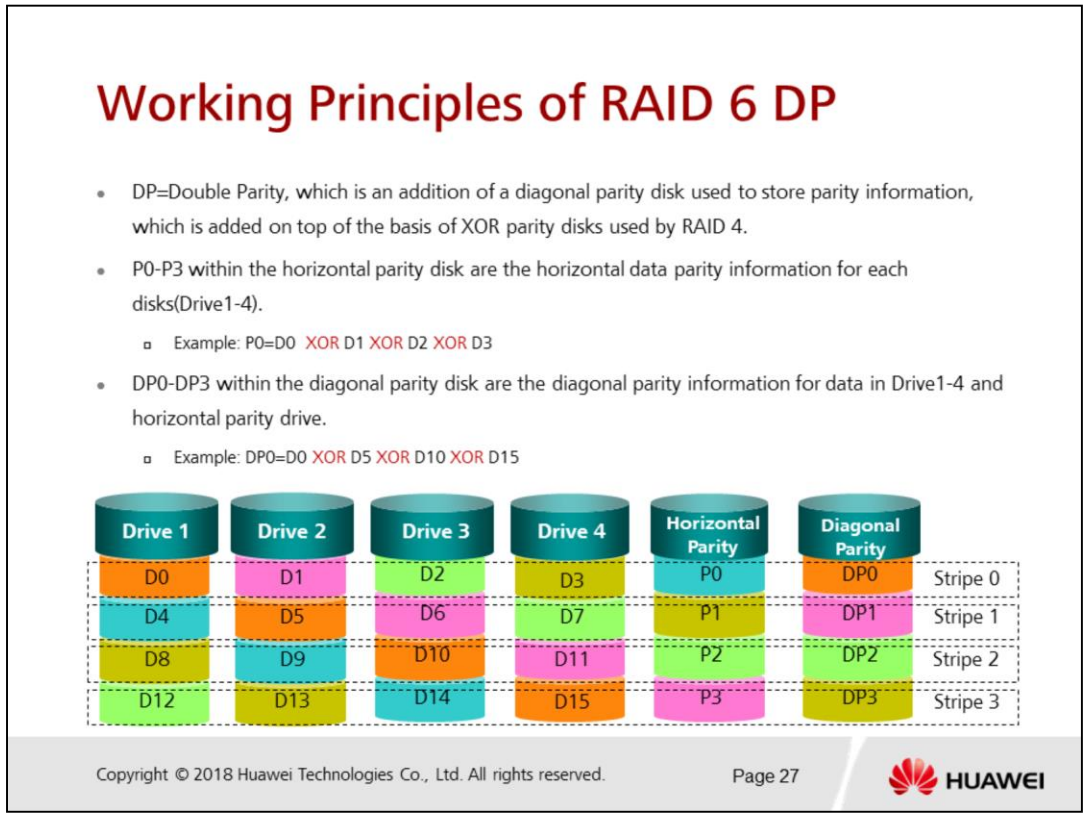
- RAID 6 P+Q calculates 2 different sets of parity data P and Q, when data is lost, the data can be recovered based on P and Q. The parity information of P and Q is calculated based on the formula below:
  - P = D0 ⊕ D1 ⊕ D2…
  - Q = (α ⊕ D0) ⊕ (β ⊕ D1) ⊕ (γ ⊕ D2)…



HUAWEI

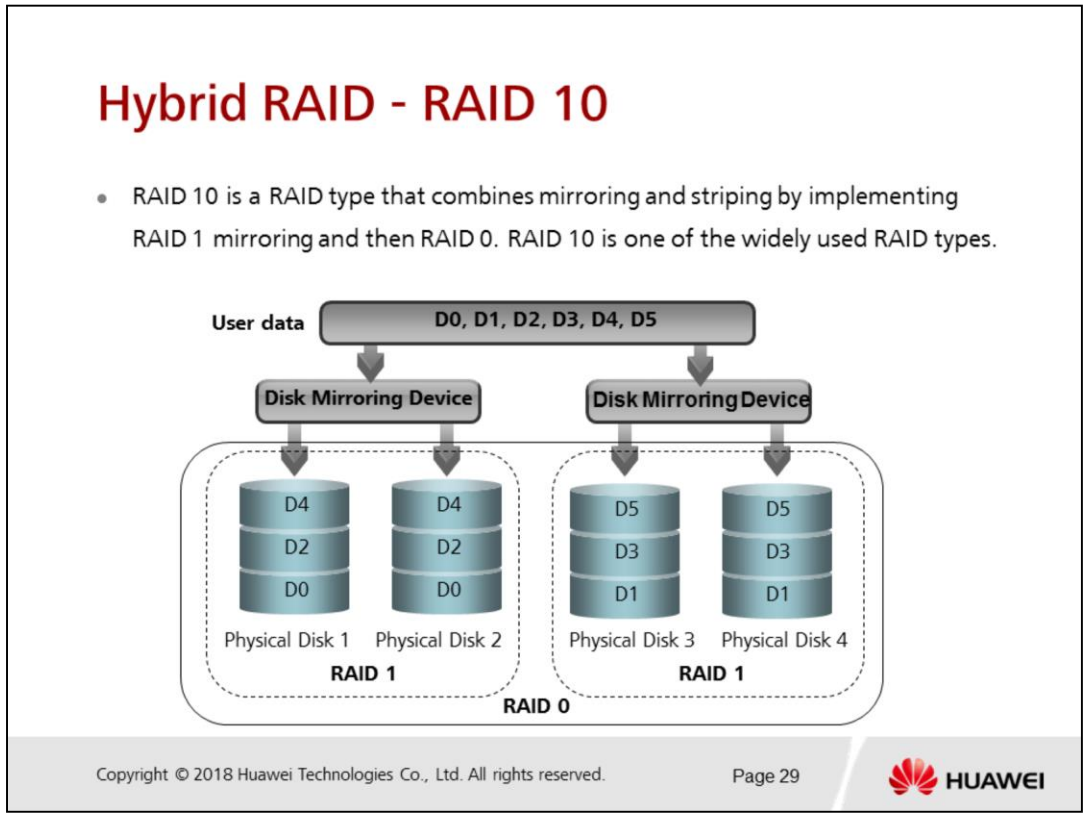
- When RAID 6 implements P+Q parity, P and Q are 2 independent set of parity value. They use different calculations, the user data and the parity information of the same stripe is distributed across the disks.
- P is calculated using simple XOR operation. Q is calculated by GF (Galois Field) conversion then with XOR operation. α (alpha) , β (beta) and γ (gamma)are constants used in the GF formula, and the resulting value will produce a "Reed Solomon Code". This calculation involves the GF conversion and XOR operation of the data in the same stripes within the data disks. Reed Solomon codes is widely used for error correction due to its capability to convert the raw data distributed across multiple disks into parity information, and it is able to restore the lost data by running calculations based on available parity information.
- As shown on the diagram above, P1 is derived by running XOR operation on the stripe 0 containing D0, D1 and D2. Similarly, P2 and P3 are derived using the same method on stripe 1 and stripe 2 respectively.

- Q1 is derived by running the XOR operation after the GF conversion of the stripe 0 containing D0, D1, and D2. Similarly, Q2 and Q3 are derived using the same method on stripe 1 and stripe 2 respectively.
- If a stripe is faulty within a disk, only P parity value is required to fully restore the data in the affected stripe by running the XOR operation on the P parity value and the data blocks in the other disks. If there is 2 disk failures in the same stripe, the are different ways of handling which differs with the scenario. In the scenario where the Q parity value is not in the faulty drives, data can be restored to the data disks then the parity value can be recalculated. In the scenario where Q parity value is in the faulty drive, then both methods need to be used on the two faulty drives, which means that P value is used to restore the Q value, then the Q value is used to restore the data in other disks.



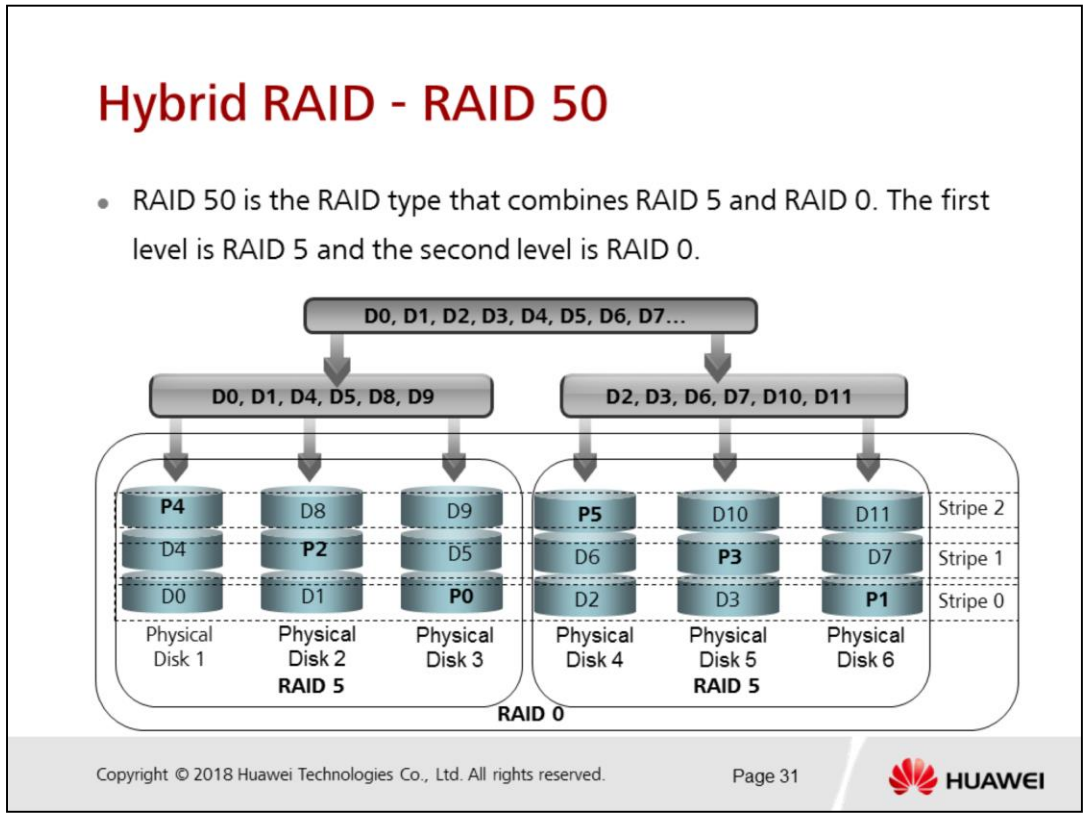
- Another type of RAID 6 implementation is RAID 6 DP(Double Parity). RAID 6 DP has two independent parity data blocks. The first parity information is the same as P parity of RAID 6 P+Q. The second parity information differs from RAID 6 P+Q, and uses diagonal XOR computation to calculate the parity information for diagonal data blocks across the disks in the array. Horizontal parity information is calculated using XOR computation on the data blocks in the same stripe. As shown on the diagram above, P0 is derived from XOR computation of the data blocks in stripe 0 which are D0, D1, D2 and D3. P1 is derived from XOR computation on data blocks D4, D5, D6, D7 on stripe 1 etc. Hence, we can conclude that P0 = D0 ⊕D1⊕ D2⊕D3， P1 = D4⊕D5⊕D6⊕D7 and so on for the other data on the remaining stripes in the disk array.
- The second parity data block is calculated using XOR computation on diagonal data blocks within the disk array. The process of choosing the data blocks for this computation is quite complex. DP0 is derived from XOR computation on data blocks: D0 on stripe 0 on drive 1, D5 on stripe 1 on drive 2, D10 on stripe 2 on drive 3 and D15 on stripe 3 on drive 4. Similarly, DP1 is derived from data blocks: D1 on stripe 0 on drive 2, D6 on stripe 1 on drive 3, D11 on stripe 2 on drive 4, and the horizontal parity data block P3 on stripe 3 on the horizontal parity disk. Hence, we can conclude that DP0 = D0⊕D5⊕D10⊕D15， DP1 = D1⊕D6⊕D11⊕P3, and so on for the other remaining stripes on the disk array.

- A RAID 6 array can tolerate 2 disk failures. As shown on the diagram above, if drive 1 and drive 2 fails, all the data in those two disk are lost, but since the other data blocks and parity information still exists on other drives, those lost data in drive 1 and drive 2 is still recoverable. Let us look into the process on how those data can be recovered. To recover D12, we use DP2 and diagonal parity computation (D12 = D2⊕D7⊕P2⊕DP2). To recover D13, we use P3 and horizontal parity computation (D13 = D12⊕D14⊕D15⊕P3). To recover D8, we use DP3 and diagonal parity computation (D8 = D3⊕P1⊕DP3⊕D13). To recover D9, we use P2 and horizontal parity computation (D9 = D8⊕D10⊕D11⊕P2). To recover D4, we use DP4 and diagonal parity computation and finally to recover D5 we use P1 and horizontal parity computation etc. All these operations are repetitive until all the data on the faulty drive are fully restored.
- A RAID 6 array performance no matter if it is DP or RAID P+Q, are comparatively slower to other RAID types. Hence, RAID 6 is suitable for 2 scenarios:
    - First scenario where data is really important, and need to be online and available over a maximum and longest period of time possible.
    - Second scenario where the disk capacity used are very huge (usually more than 2TB). Large capacity disk has longer disk reconstruction time, and two faulty disk will cause a period of inaccessible data for long hours. In RAID 6, it achieves a data protection level of able to tolerate a second disk failure even when it is during the disk reconstruction period of the first disk failure. Enterprises tends to use large capacity drives, and they prefer the vendors for storage arrays to be able to support and use a RAID group that has dual and redundant protection such as RAID 6.



- For most of the enterprise customers, RAID 0 is not actually a viable option for operation, and RAID 1 is limited in the sense of disk capacity utilization. RAID 10 combines RAID 1 and RAID 0, which provides the best solution especially for random data write, and since it does not have write penalty, there is a clear performance advantage.
- RAID 10 array always has a even disk count. Half of the disks are used for storing user data, and another half is used for mirror copies of the user data. The mirroring is executed based on data striping which means that each data is striped across the disks and mirrored.
- As shown on the diagram above, the physical disk 1 and 2 forms a RAID 1 and the other physical disk 3 and 4 forms another RAID 1. Both of these RAID 1 subgroups in turn forms a RAID 0 group.

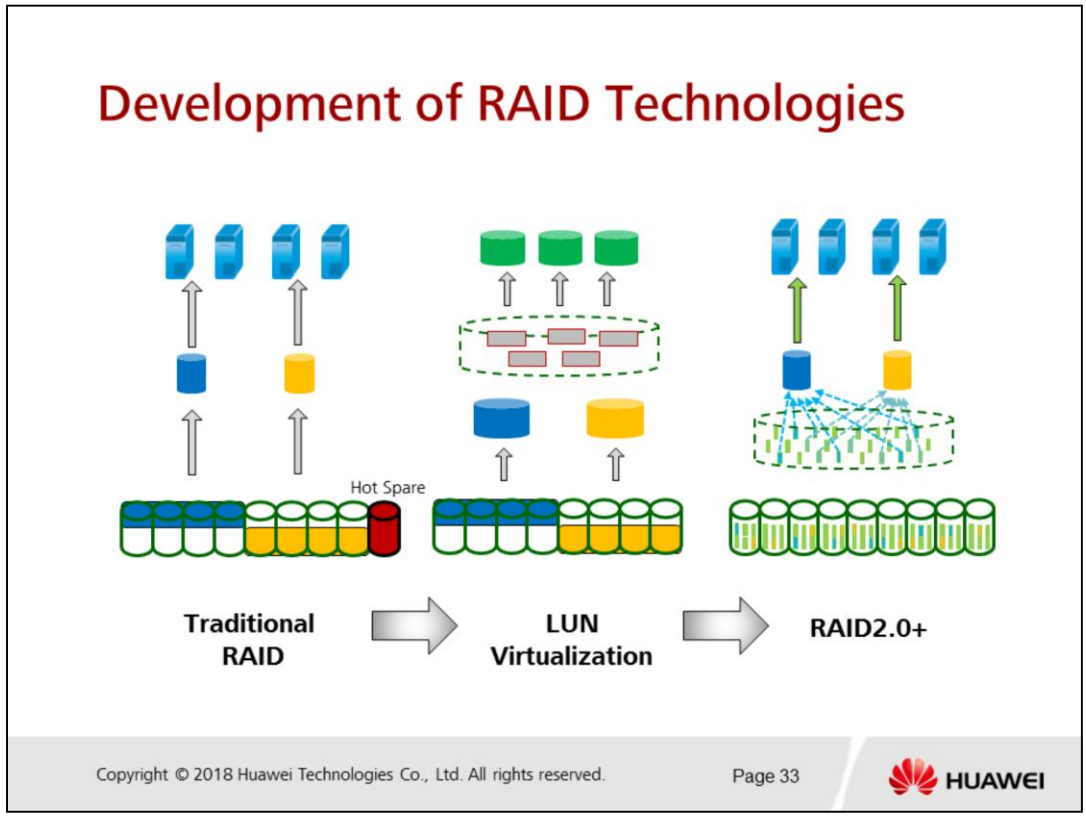
- When data is written in RAID 10, it uses parallel writing between the subgroups to write the data blocks, and uses mirroring to write the data blocks within the subgroup to the physical disks. As shown on the diagram above, when D0 is written into the RAID 10 array, it is written to physical disk 1 and it is mirrored to physical disk 2 as a copy.
- When disks on the different RAID 1 groups is faulty (e.g. Disk 2 and Disk 4), data access in RAID 10 is not affected. This is because the other 2 disks (Disk 1 and Disk 3) has the full copy of the data of the faulty disk 2 and 4. However, if both the disks in the same RAID 1 group is faulty at the same time (e.g. Disk 1 and Disk 2), data are no longer accessible.
- In theory, RAID 10 can tolerate half of the physical disk failure in the array, but in the worst case scenario, where both disks in the same subgroup fails, RAID 10 can also cause data loss. Generally, RAID 10 is used for single drive failure tolerant data protection.



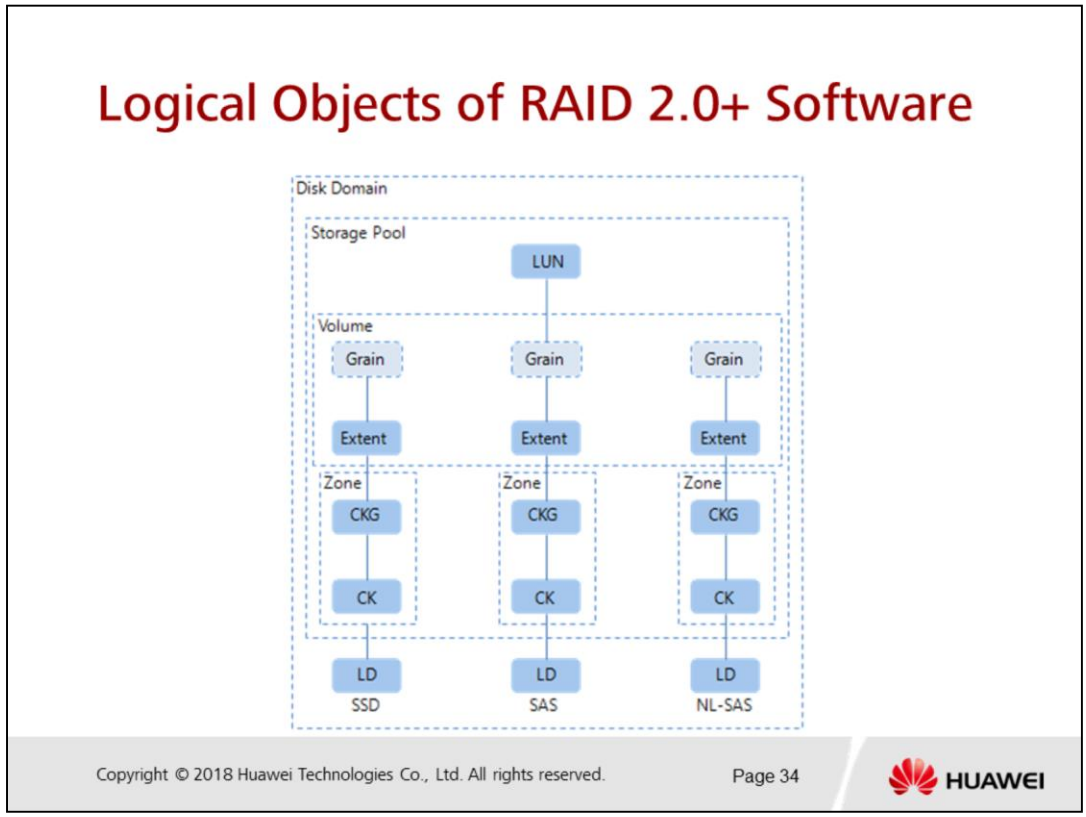
- RAID 50 is the combination of RAID 0 and RAID 5. Both subgroups are configured as RAID 5, and these 2 subgroups in turn forms a RAID 0. Each RAID 5 subgroup is independent from each other. RAID 50 requires at least 6 disks because a RAID 5 subgroup requires a minimum of 3 disks.
- As shown on the diagram above, physical disk 1, 2 and 3 forms a RAID 5 and physical disk 3, 4, and 5 forms another RAID 5 group. Both of theses RAID 5 subgroups forms another RAID 0.
- In RAID 50 array, it can tolerate multiple drive failures at the same time. However, in the event that 2 disks from the same RIAD 5 subgroup fails at the same time, the data on RAID 50 array will be lost.

Contents

HUAWEI

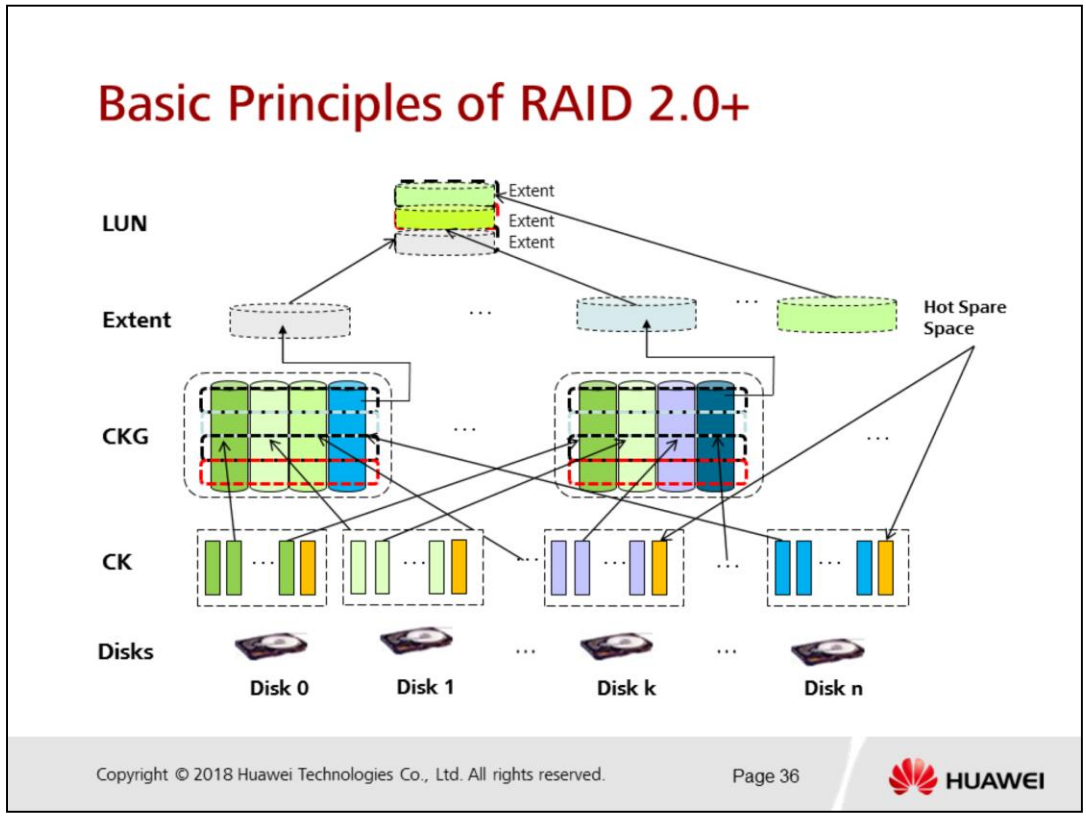


- In the earliest form of RAID technologies, it was just combination of couple smaller capacity inexpensive disks into a larger capacity logical disk for mainframe usage. Eventually, as the capacity of hard drives gets larger, the initial purpose of building RAID is longer towards creating a large capacity disks, but to utilize the RAID technologies to achieve data reliability and data security, and increase storage performance. As each individual disk capacity has become larger, and the storage capacity of digital disks in RAID are much larger, smaller logical unit number (LUN) are allocated from the RAID and used by servers. Following the development of disk technologies, the single disk capacity has reached few terabytes, and the time needed for disk reconstruction gets much longer which increases the risk of other disk failure during the disk reconstruction period that will cause data loss. In order to solve this issues, block virtualization technologies was introduced. RAID 2.0+ is a Huawei proprietary block virtualization technology, it distributes the physical and digital storage space into separate smaller blocks across multiple disks, and can utilize the data read and write performance of the storage system to its fullest extent. It is easier for expansion, allocation of storage space based on demand, distribution of hot and cold data, and data migration. It is the fundamental for all of the Huawei Smart storage software technologies. At the same time, since the hot spare space is also distributed across multiple disks, the data reconstruction can be done simultaneously across all disks, and it avoids the performance bottleneck of writing to a single hot spare disk, which in turn greatly reduces the data reconstruction time.

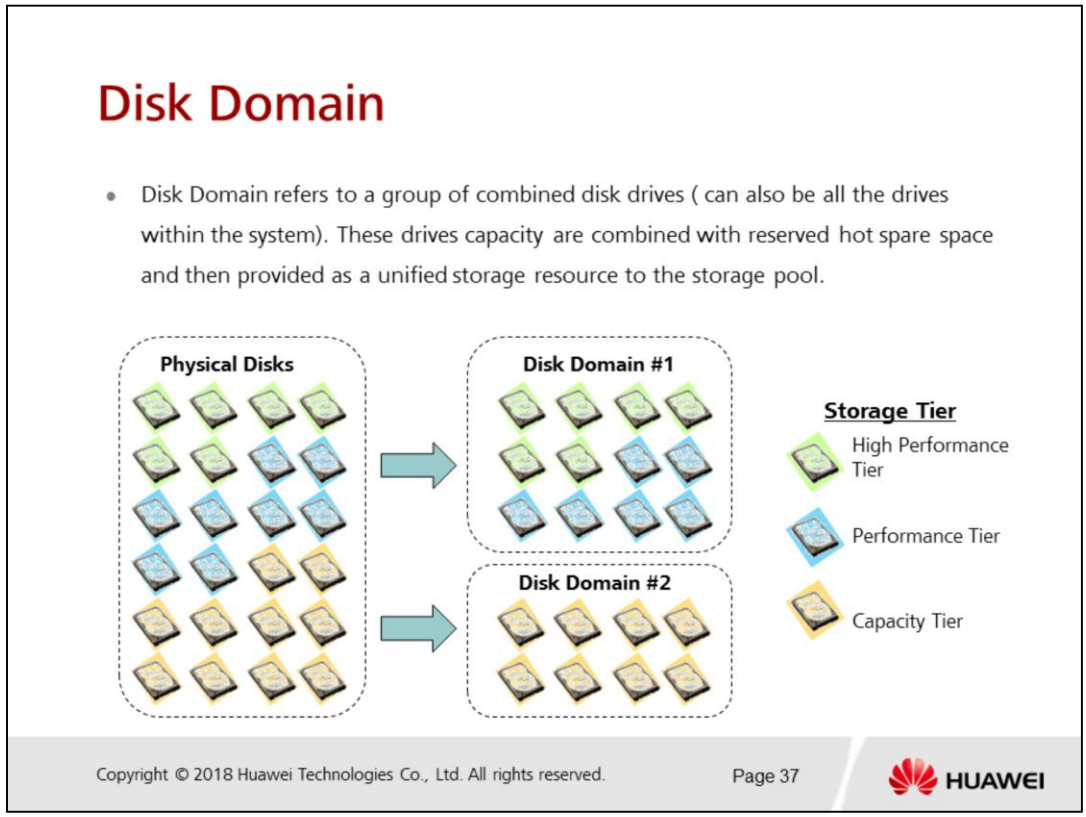


- Huawei RAID 2.0+ implements 2 layer virtualization management model that consist of low level disk management and high level resource management. Inside the system, the storage space of each disk is divided into smaller chunks, and these chunks are used as the fundamentals to build the RAID group, which allows data to be evenly distributed to all the disk in the storage pool. At the same time, chunks are used as the unit for resource management, which greatly increases the efficiency of resource management.
    - OceanStor storage system supports different types of disks (SSD, SAS, NL-SAS). In theory, SATA disks can also be used, but because of the low performance, it is less used in enterprise grade storage. These disks in the storage system forms a Disk Domain. In a disk domain, same type of disks can form a storage tier, and disks within the storage tier can be further allocated into different Disk Groups.
    - The disks in each storage tier is divided into Chunk(CK) of fixed size, in fact the CK size of SSD layer and SAS layer is 64MB and the CK size of NL-SAS layer is 256M by default.
    - OceanStor storage system uses random algorithm to form a Chunk Group (CKG) using the Chunk (CK) in each storage tier based on the user defined "RAID Policy". Users can set a RAID Policy for each of the storage tier within the Storage Pool.

    - OceanStor storage system will the cut up the Chunk Group (CKG) into smaller Extent. Extent is the smallest unit for data migration and it is the basic unit used to form a Thick LUN. The size of the Extent is by default 4MB but can be configured in the "Advanced" setting when creating a Storage Pool. For Thin LUN or File Systems, the Extent will be further divided into smaller unit called "Grain", and Grain is the unit used to create Thin LUNs and File Systems. (The concept of Thin LUN and Thick LUN will be described in the following modules).
- A group of Extent forms a Volume. Volume appears to the host as the LUN that it accesses (The LUN referred here is Thick LUN). When processing the user read/write request, or doing data migration, or application or release of storage space from storage system and LUN, all of these processes use Extent as a unit. For example: When a user creates a LUN, they can specify the capacity to be obtained from a certain storage tier, at this point the LUN will be created from the Extent available from the storage tier specified. When the user services are in operation, the storage system will migrate the frequently accessed and non frequently accessed data in between storage tiers based on the migration policies set by the user (This feature requires the purchase of the license for SmartTier). At this point of time, the data in the LUN in the unit of Extent will be distributed across the different storage tiers in the storage pool.
- When the user creates Thin LUN or File Systems, OceanStor storage system will further divide the Extent into smaller unit as Grain, and it serves as the basic unit for Thin LUN and File Systems. This further division of Extent into Grain allows the storage system to achieve much finer control in terms of the storage capacity.

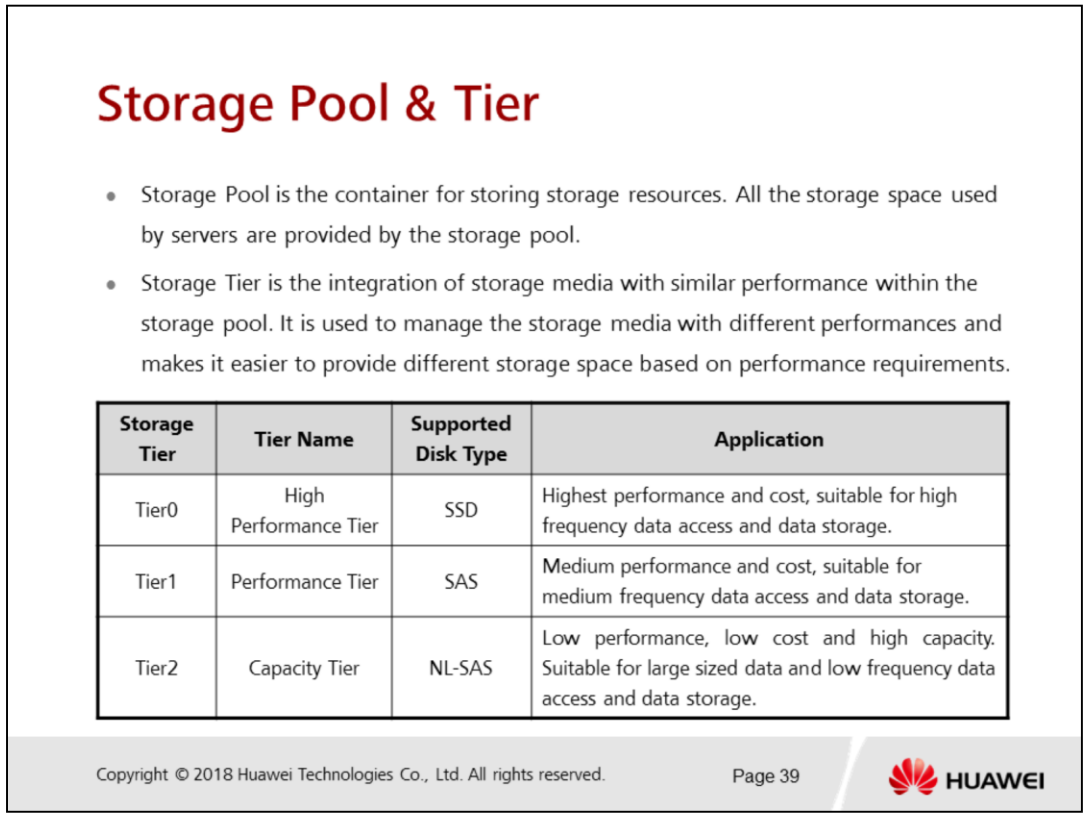


- In simpler terms, the overall principles of RAID 2.0+ can be represented as shown on the diagram above.



- In traditional RAID, the first step before providing the storage space to the host is to create a RAID group. But there are limitations and requirement for this step: The disk type within the RAID group must be the same, the storage capacity and the rotational speed of the disks must be identical, and the maximum number of disks are recommended to not exceed the count of 12 disks.
- Huawei RAID 2.0+ technology implements another way to provide storage space to hosts. In this method, the first step is no longer creating a RAID group but to create a Disk Domain. A disk domain refers to a group of disk drives. 1 disk drive can only belong to 1 disk domain. OceanStor storage system can create one or more disk domains.
- At a glance, the concept of disk domain and RAID groups seems similar, which are groups of disk drives, but in reality, there are big differences between the two. In the creation of RAID groups, the set of drives used to form the RAID group have already been configured to a fixed RAID type, and the drive type, drive capacity and drive rotational speed must be the same. However, in a disk domain, the number of drives can be huge such as hundreds of drives, and can have a maximum of 3 different drive types, where each type of the drive can be assigned to a storage tier.

- SSD (Solid State Drive) drive can be assigned to High Performance storage tier, SAS (Serial Attached SATA) drive can be assigned to Performance tier, and NL-SAS (Near Line SAS) drives can be assigned to Capacity tier. If the disk domain does not have the corresponding disk type for that storage tier, then that storage tier will not be available within the disk domain.
- The key differences between disk domain and RAID group lies in the fact that the main purpose of disk domain is for the isolation between different groups of hard drives and the total isolation in terms of failures, performance and storage resource. When creating a disk domain, the RAID type is not specified which means that the redundant data protection level has not been specified for that particular group of disks. In reality, RAID 2.0+ allows more flexible data protection configuration, and more finer grained control. The storage space formed from the combination of drive capacity within the disk domain are further divided into fine grained storage pools along with shared hot spare space shared between different storage tiers.
- The size of hot spare space is based on the configured hot spare policy(High, Low, None) set for the disk domain and auto configuration based on the number of disks within each storage tier in the disk domain. The hot spare space is no longer set as the specified drive designated by the storage administrator like in the traditional RAID.

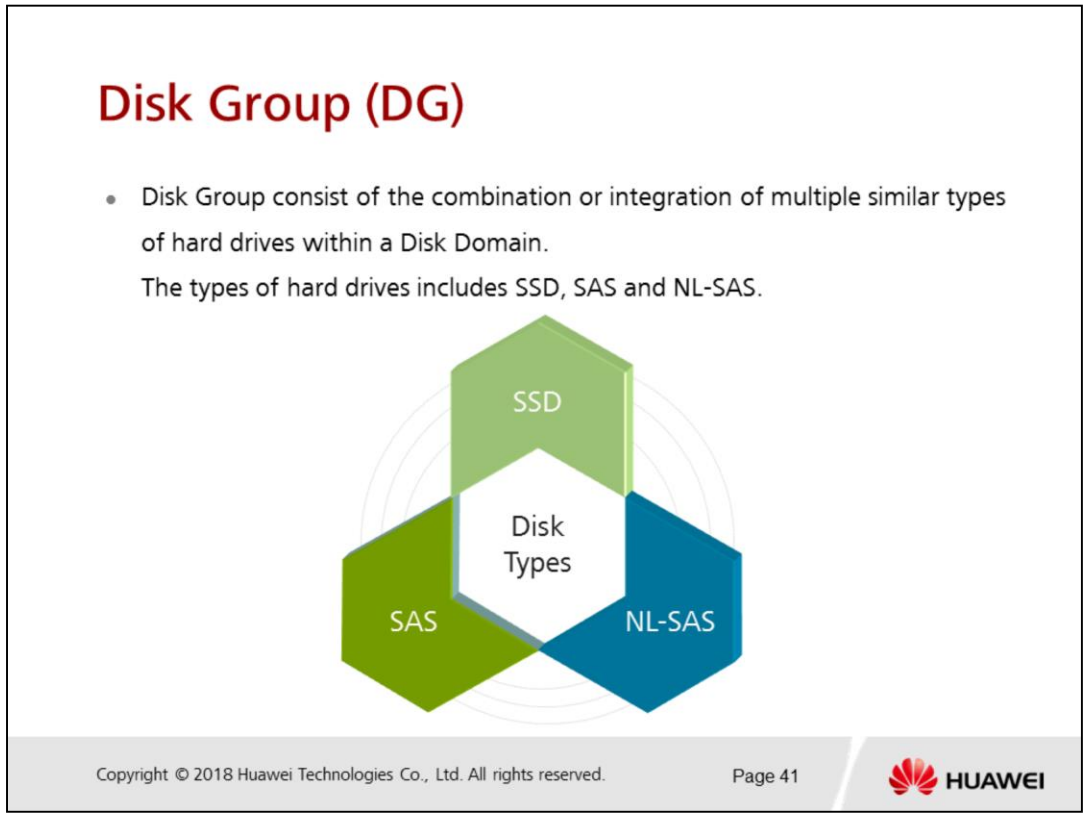


- A Storage Pool is created based on a disk domain, it can dynamically allocate the Chunk(CK) resources, form Chunk Group (CKG) based on the Raid Policy of each Storage Tier, to provide RAID protected storage resource to applications and users.
- A storage pool can be divided into different tiers, and the storage tiers and types of disk drives supported by the OceanStor storage system is shown on the table above.
- When you create a Storage Pool from a disk domain, you can specify the storage tier types and the “RAID Policy” and “Capacity” of each tier.
- The Capacity tier mostly consists of high capacity SATA or NL-SAS drives, and the recommended RAID policy is to use double parity RAID 6. (SATA drives are less commonly used nowadays, and part of the products no longer supports them in terms of specifications.)

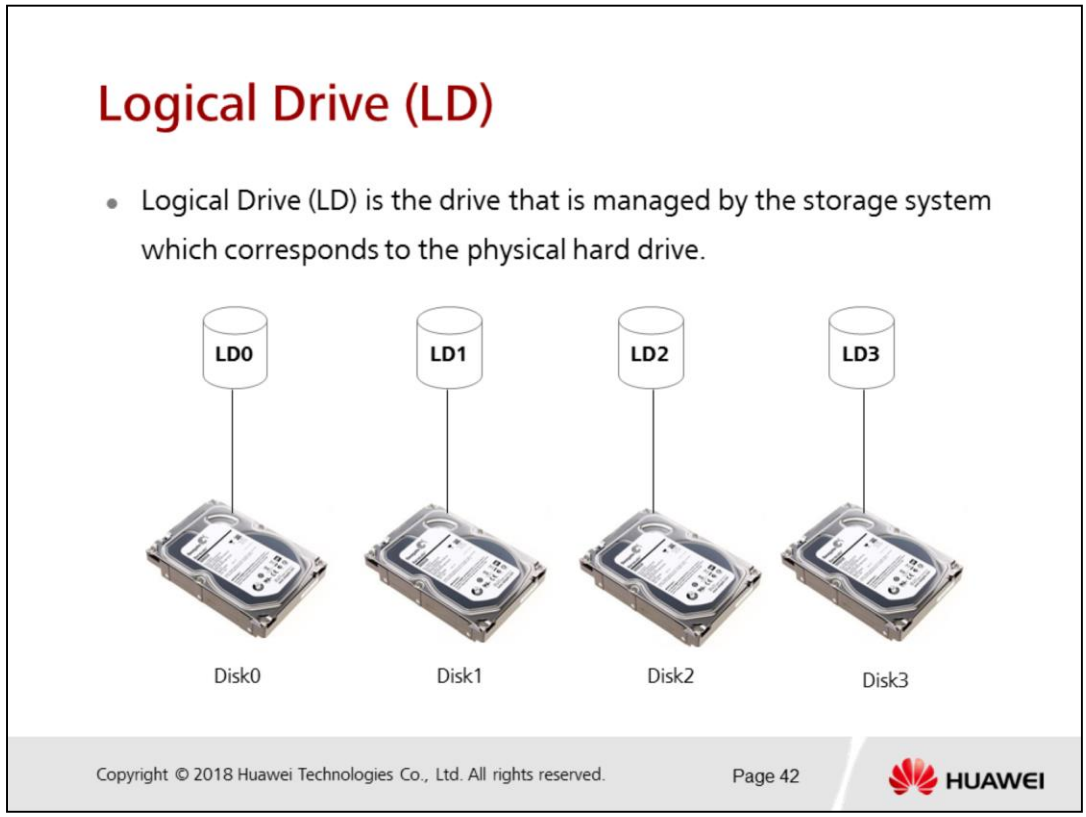
Storage RAID Type & RAID Policy

RAID Type	RAID Policy
RAID1	1D+1D，1D+1D+1D+1D
RAID10	System automatically chooses 2D +2D or 4D+4D
RAID3	2D+1P，4D+1P，8D+1P
RAID5	2D+1P，4D+1P，8D+1P
RAID50	(2D+1P)*2，(4D+1P)*2，(8D+1P)*2
RAID6	2D+2P，4D+2P，8D+2P，16D+2P

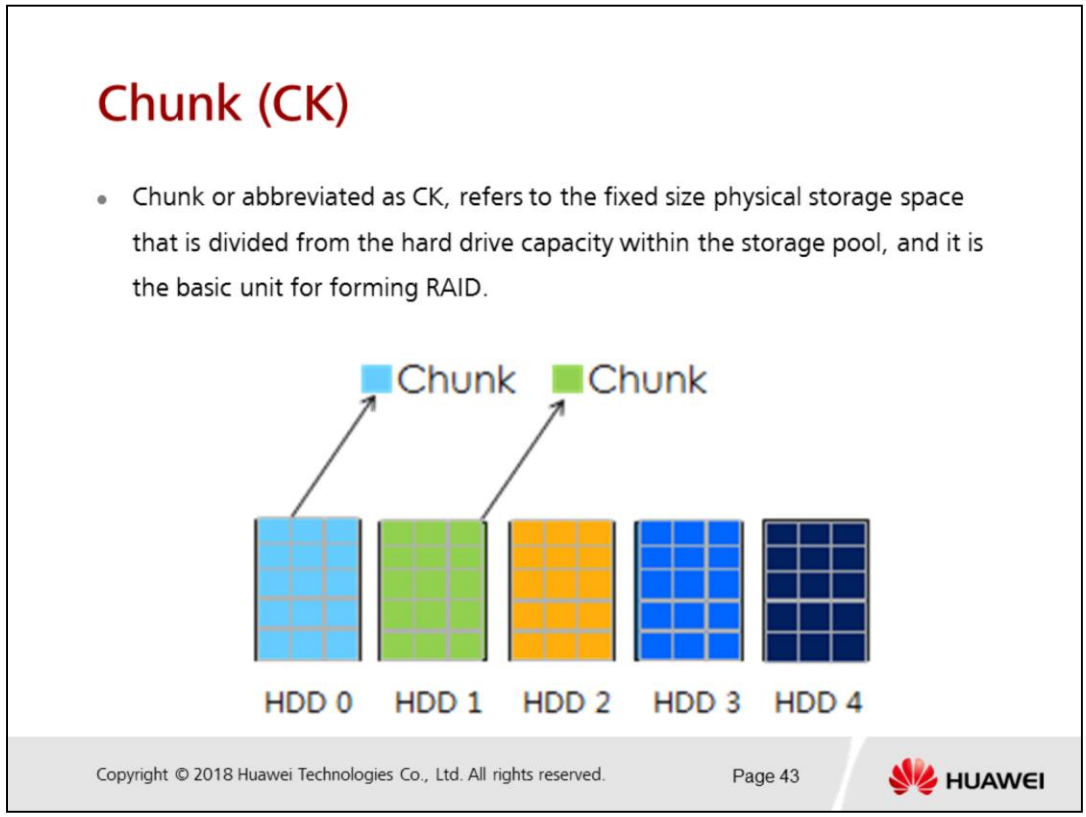
- OceanStor storage system supports RAID 1, RAID 10, RAID 3, RAID 5, RAID 50, and RAID 6. The supported RAID policy and configurations are shown as the table above.



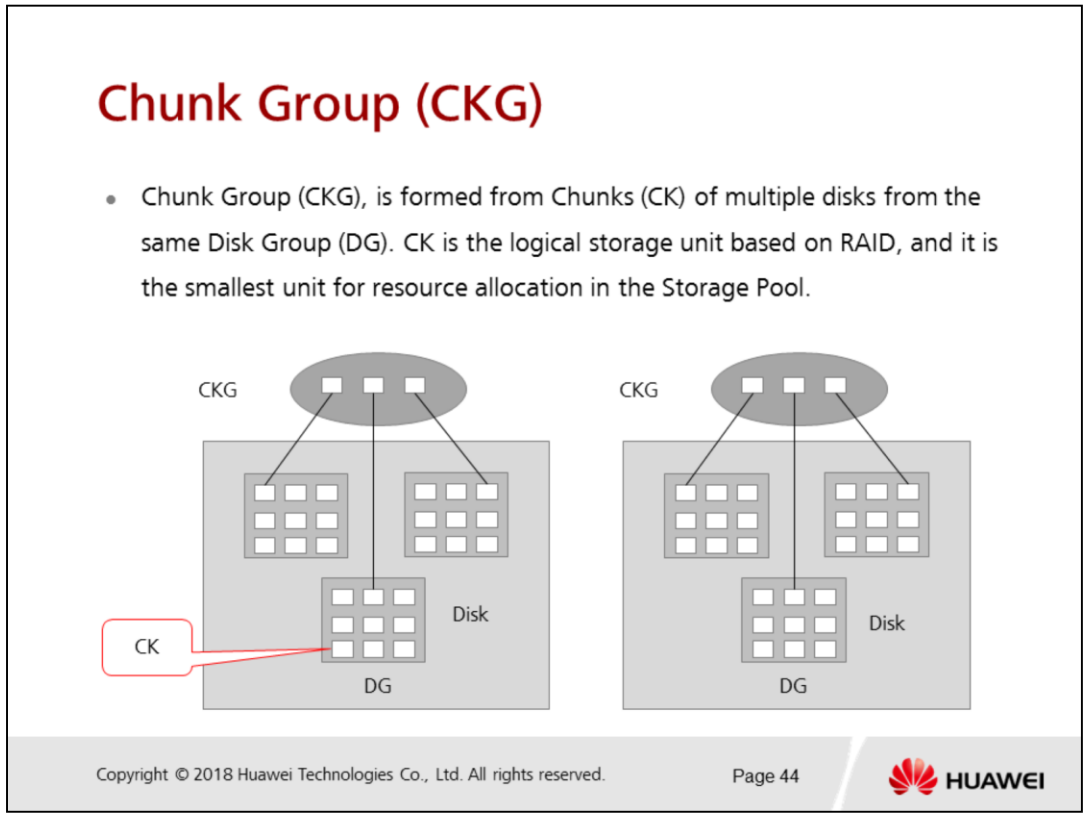
- OceanStor storage system will automatically divide the disks within the disk domain into one or more Disk Groups (DG) based on the number of disks within each disk type.
- 1 Disk Group (DG) can only contain 1 type of hard disk drives.
- Anyone of the many Chunks (CK) in a CKG (Chunk Group) originates from the different hard drives within the same Disk group.
- Disk Group (DG) belongs as the internal system object, and its main purpose is for fault isolation, it is automatically configured by the OceanStor storage system and not represented for external visualization for users.



- A logical drive is a drive space that is logically created on top of a physical hard disk drive. A logical drive is a separate partition with its own parameters and functions, and it operates independently. A logical drive can also be called a logical drive partition or logical disk partition.
- A physical drive refers to the hardware unit within a computer, laptop, or server. It is a data storage device that can store and retrieve digital information using one or more platters coated with a thin magnetic layer. Physical drives are usually attached to a computer or laptop. In some cases, they can be external, such as a USB pen drive or an external memory. A physical drive can be partitioned into two or more logical disks which can operate independently of each other.



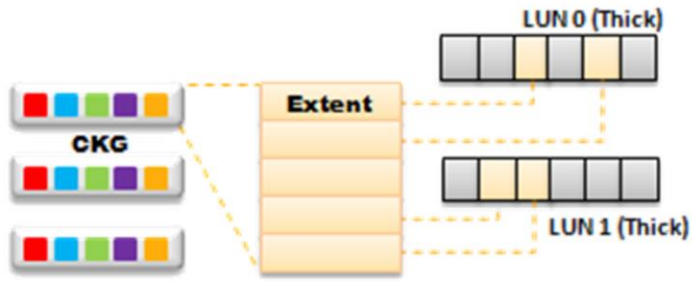
- The Chunk size is the set of consecutive physical storage space of fixed size divided by the storage system. The Chunk size cannot be changed.
- In simpler terms, a Chunk is the unit for dividing the disk space in a storage system. A chunk is 64 MB in size. It is the fundamental unit in the Huawei RAID2.0+ technology.



- The CK in the CKG are obtained from the hard disks within the same DG. CKG has RAID properties (RAID properties are actually configured on the Storage Tier). Both CK and CKG belong to internal system objects, and are automatically configured by the storage system which are not represented visually to external application and users.
- In a storage pool, a specified number of chunks make up a chunk group based on a specified RAID group.
- Chunks in a chunk group must come from disks of the same type.

Extent

- Extent is the logical storage space of fixed size that is formed using the CKG as the fundamental unit. The size of extent can be changed, and it is the smallest unit for hot data analysis and data migration (data migration granularity).

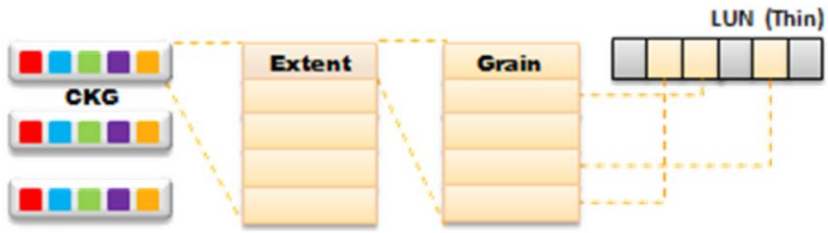


HUAWEI

- An Extent belongs to a Volume or a LUN, the size of the Extent can be configured when creating the storage pool. Once it is configured, the size of Extent can no longer be changed. The size of Extent in different storage pool can be different, but the size of the extent in the same storage pool is unified and identical.
- An extent is fixed-sized logical space by which space in a storage pool is divided. Extent is the smallest unit for applying for and freeing up space and migrating data. An extent is usually 512 KB to 64 MB in size. The default size is 4 MB.

Grain

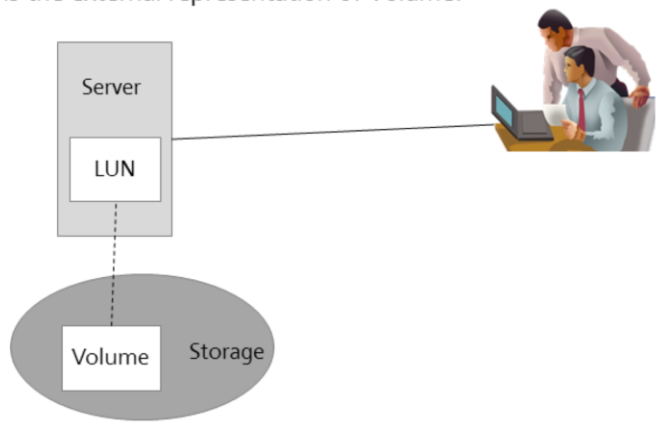
- In Thin LUN mode, the Extent is further divided into finer and smaller blocks. These blocks are called Grain. Thin LUN uses Grain for space allocation. The LBA within the Grain is continuous.



- Thin LUN is mapped to the LUN using Grain as the unit. This means that multiple Grains forms a LUN which is used to store user data. Grain is only used to form Thin LUNs and there is no such object used to form the Thick LUN. Thick LUN is formed using Extent instead of Grain.
- When a user creates a thin LUN, OceanStor storage system divide the extents into grains and map the grains to the thin LUN. In thin LUN mode, extents are divided into 64 KB grains. In this way, fine-grained management of storage capacity is implemented.
- Logical block addressing (LBA) is a common scheme used for specifying the location of blocks of data stored on computer storage devices, generally secondary storage systems such as hard disk drives. LBA is a particularly simple linear addressing scheme where blocks are located by an integer index, with the first block being LBA 0, the second LBA 1, and so on.

Volume & LUN

- Volume is the internal management object within a storage system.
- LUN is the storage unit that can be directly mapped to host for data read/write. It is the external representation of Volume.

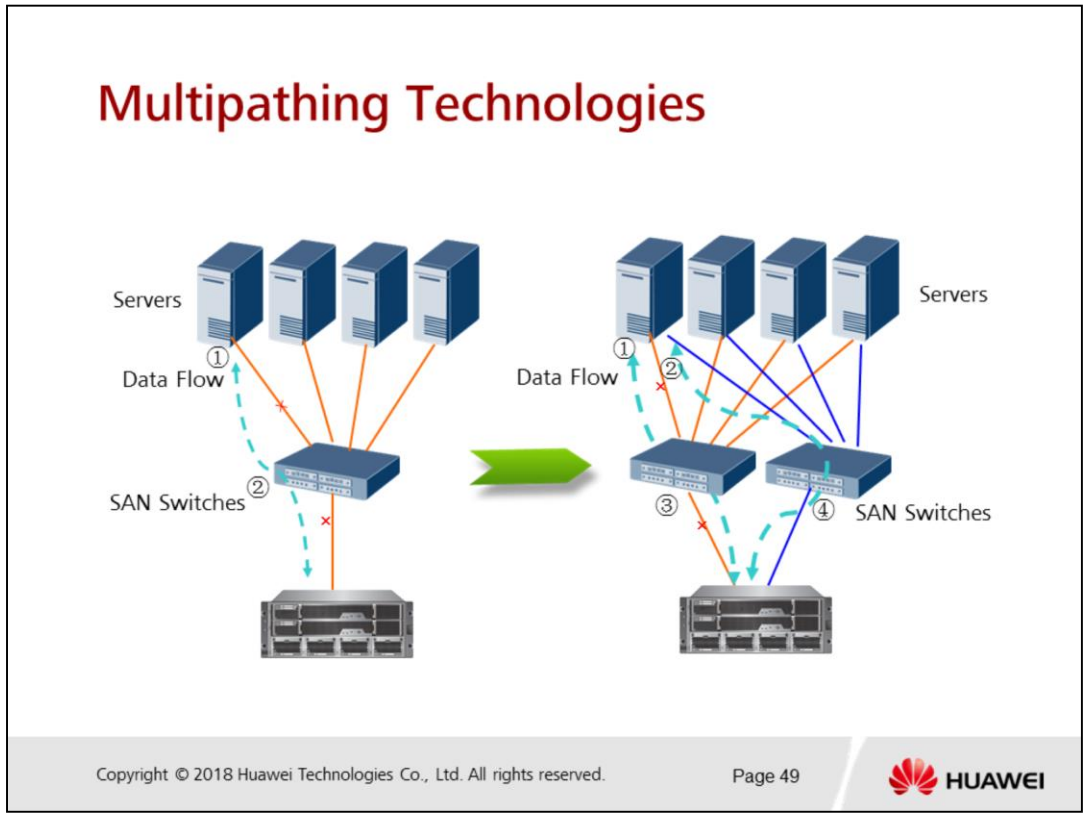


HUAWEI

- A Volume object is used to organize all the logical storage units such as Extent and Grain in a LUN. We can dynamically apply or release the amount of Extents to increase or decrease the actual storage space of a Volume.
- A LUN is made up with multiple extents. A LUN can contain multiple extents that are from different storage media, so more disks participate in data read and write, which ensures fast response, high efficiency, and improved performance.

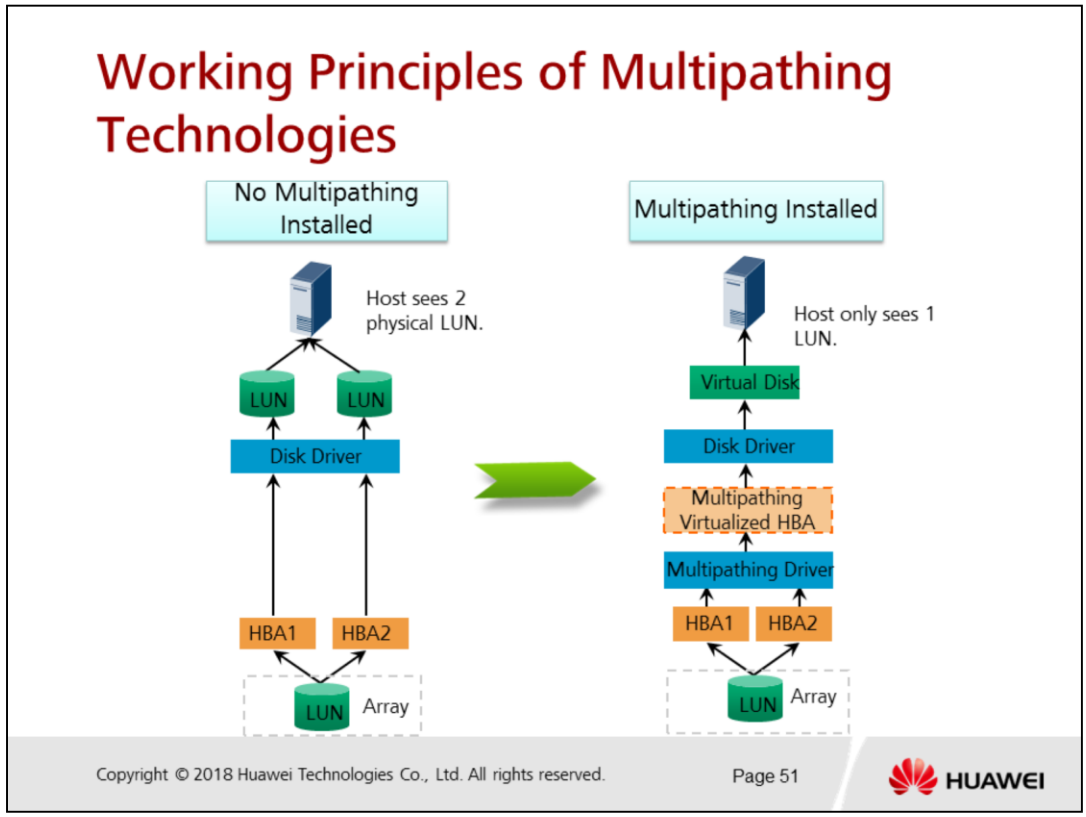
Contents

HUAWEI

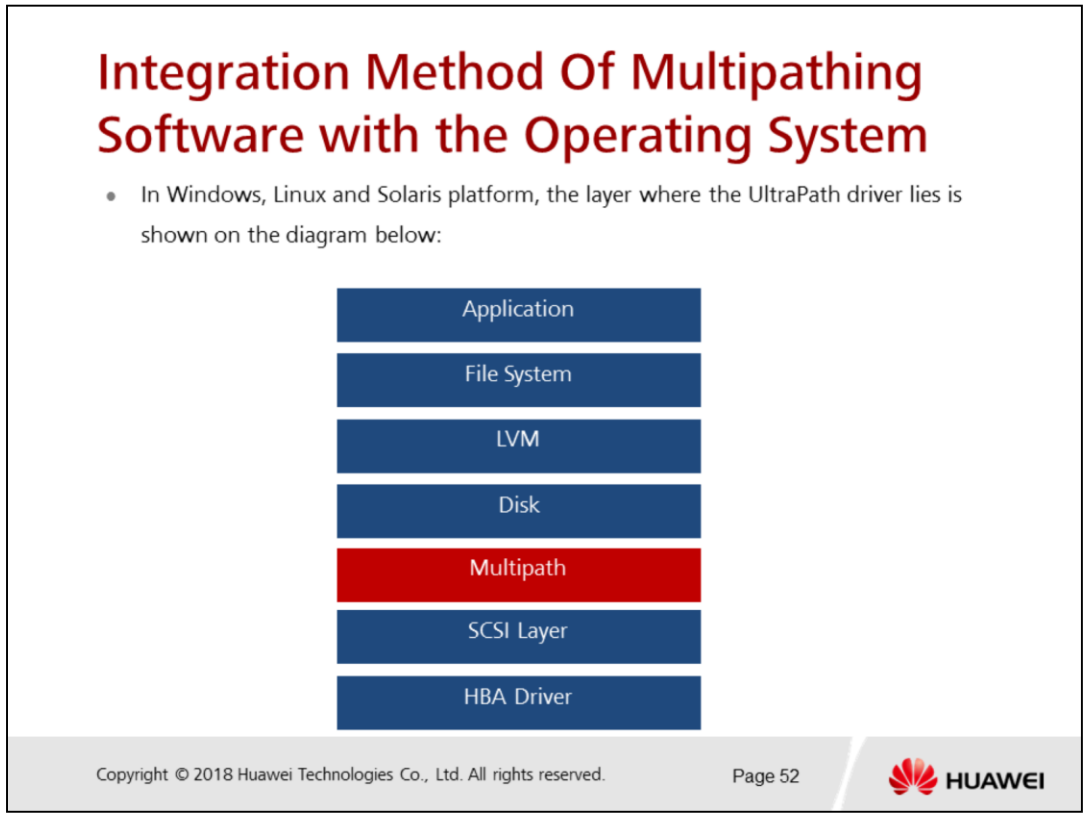


- The common path for user data to reach the disk array: Host → SAN Network → Storage System Controller → Disks in Storage System.
- Multipathing technologies refers to implementation of multiple connections and paths between the host and the storage array, which allows data to be transferred in more than 1 path between the host and storage array across multiple switches to avoid single point of failures. As shown on the right diagram above, there are 2 possible paths between the host and the storage array which are (1, 3) & (2, 4), and there are 2 independent switches along these paths. In this mode, when path (1, 3) is down, data flow from the host will be redirected by the Multipathing software to choose the path (2, 4) to reach the storage array. At the same time, if one of the switch is faulty, it will redirect the data flow towards the other path with the working network switch. When the 1st path has been recovered, the IO flow will automatically switch back to the normal route. The whole path switching and recovery process is transparent towards the host applications, which fully avoids the IO interruption between the host and storage array due to path failures.

- Storage system redundancy solutions involves the field of Multipathing. From the host side and SAN network perspective, by integrating the Huawei UltraPath Multipathing software or other Multipathing software, it ensures that there is no single point of failures between the data flow paths. From the storage controller's perspective, using fully redundant hardware and hot plugging technologies allows us to achieve dual controller active-active redundant protection. From the disks and disk enclosure's perspective, using dual port technologies and disk Multipathing technologies allows us to achieve redundant disk protection. All these redundancy technologies working together allows high availability and high business continuity operations for enterprises and businesses.

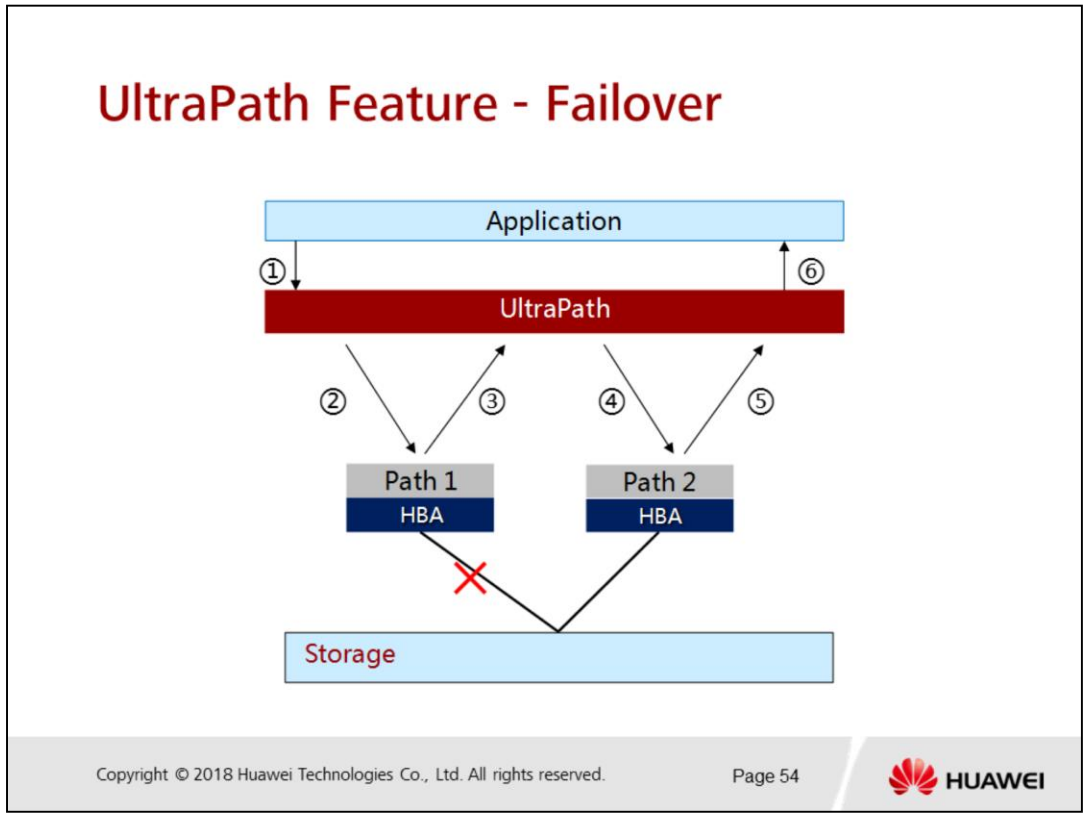


- Functions of Multipathing software:
  - Avoids logical error in operating systems caused by multiple paths directing to the same LUN.
  - Increases links and connection reliability, and avoids system failure due to single path or link failures.



- Generally, Multipathing software works underneath the disk drivers to provide virtualized disk to upper level application access. But it can also work above the disk drivers and below the LVM, like most mainstream Multipathing software in the Linux platform.
- Why is there a need to mask the LUNs when using Multipathing software?
  - In a redundant storage network, an application server with no Multipathing software detects a LUN on each path. Therefore, a LUN mapped through multiple paths is mistaken for two or more different LUNs. Redundant LUNs exists because each path reports a LUN directly to the application server. Due to the identification errors of the application server, different applications on the application server repeatedly write different data to the same location of the LUN, resulting in data corruption.
  - To resolve this problem, the application server must identify which is the real and available LUN. As UltraPath is able to acquire configuration information of the storage system, it clearly knows which LUN has been mapped to the application server.
  - UltraPath installed on the application server masks redundant LUNs on the operating system driver layer to provide the application server with only one available LUN, the virtual LUN. In this case, the application server only needs to deliver data read and write operations to UltraPath that masks the redundant LUNs, and properly writes data into LUNs without damaging other data.

- The are 2 main differences in the aspect of Multipathing software and OS integration:
    - Advantages of masking the native disks and create a virtual SCSI disk:
        - Security: Users cannot use the native redundant disks, which avoids scenarios like data corruption due to improper operation.
        - Transparency: It is transparent to upper level applications. There are no changes to the disks after deploying the Multipathing software, hence there is no need to change the configuration of upper level applications.
        - Compatibility: It has better compatibility as certain applications only recognizes standard SCSI disks.
    - Disadvantages of masking the original disks and create a virtual SCSI disk:
        - Non OS native Multipathing software requires OS driver stack to support the event blocking mechanism in order to mask the native disk, or it needs to use the hook method to replace the system functions. There is a risk of coexistence with third party drivers which means that certain drivers may affect the functionality of the Multipathing driver. If the masking functionality of the Multipathing driver is affected, data corruption due to wrong identification of LUN or disk as mentioned earlier can occur.
- Advantages of not masking the native disk, but to create additional virtual disk:
    - Better compatibility with third party drivers.
- Disadvantages of not masking the native disk, but to create additional virtual disk:
    - Users can use the native disks, which may cause data corruption due to improper operation.
    - Need to reconfigure the applications on the upper layer.
    - Compatibility issues exists.



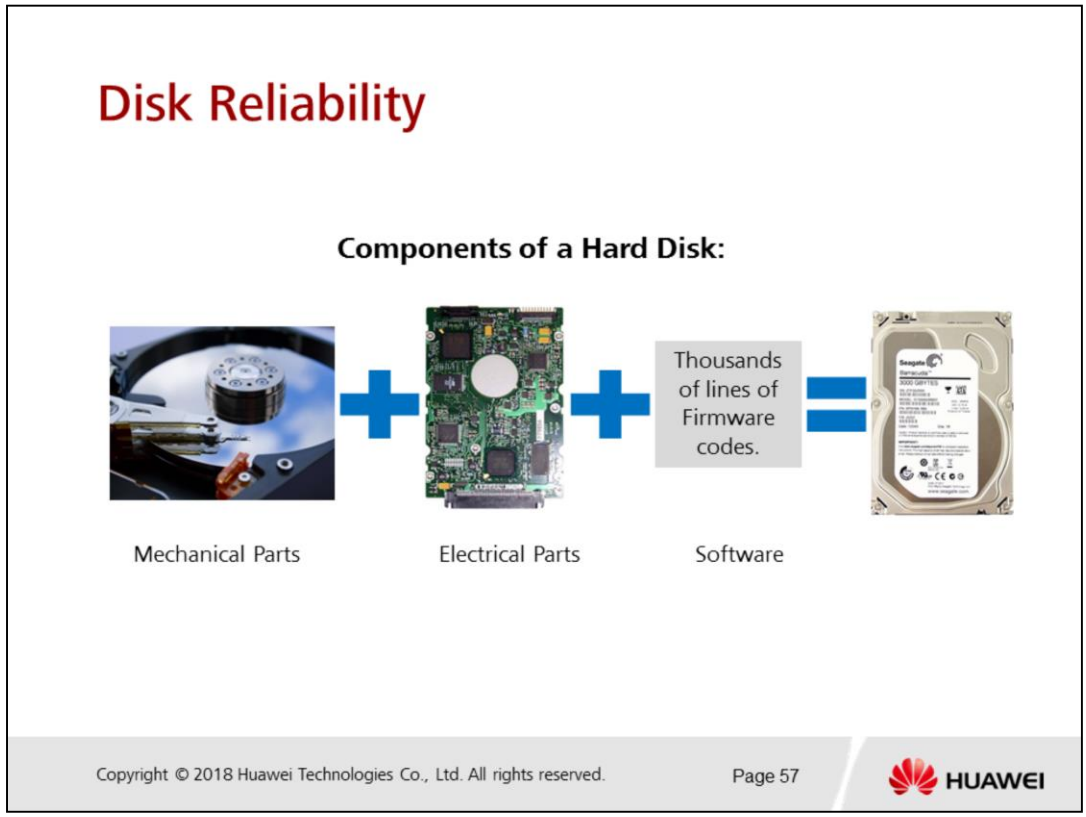
- When the same storage controller has multiple paths connecting to the host, and when the LUN belong to that particular storage controller, if a main path has failures, it will preferably choose the path on the other ports on the same storage controller. If no such path exists, then it shall choose the path that is connected to the other storage controller to send or receive I/O from host.
- UltraPath will automatically transfer the I/O using other paths in the event of a path failure. The process for this failover is shown on the diagram above:
    1. Application sends I/O towards the virtual disk generated by the UltraPath Multipathing software.
    2. UltraPath forwards the I/O onto Path 1 (SCSI device).
    3. Faulty path causes the I/O on the Path 1 fails.
    4. UltraPath retransmit the I/O on the other alternative Path 2.
    5. Path 2 I/O process is successful.
    6. UltraPath sends the I/O process success message to the upper layer application.

UltraPath Feature - Failback

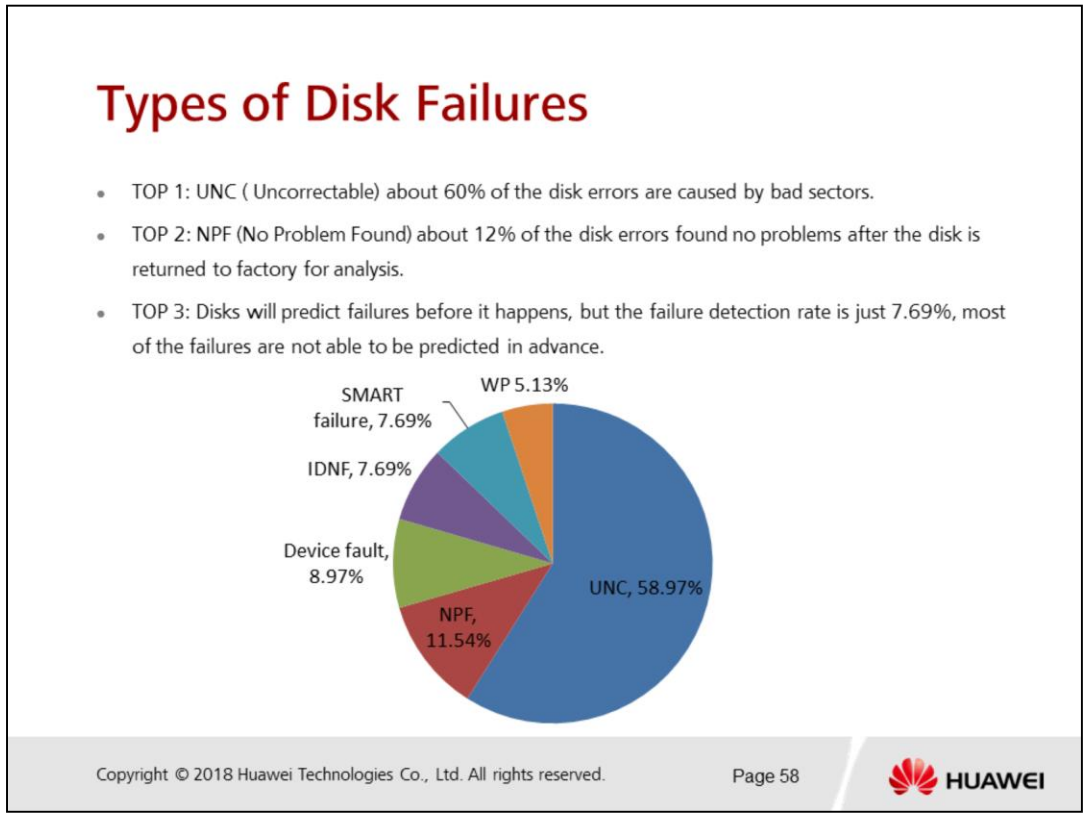
- UltraPath will retransmit the I/O back to the normal route after detecting that the normal path has recovered from failure.
- There are two methods of path recovery:
  - For systems that supports hot plugging technologies (such as Windows), the broken path between the host and storage will cause the SCSI device to be removed, correspondingly when the link has recovered it will trigger the automatic connection of the SCSI device. In this scenario, UltraPath will detect the path recovery instantly.
  - For systems that doesn't support hot plugging technologies (such as AIX, earlier version of Linux), UltraPath detects path recovery in periodic interval cycle. This means that UltraPath will check the paths in the configured specific interval of time to see if the path is recovered or not.

Contents

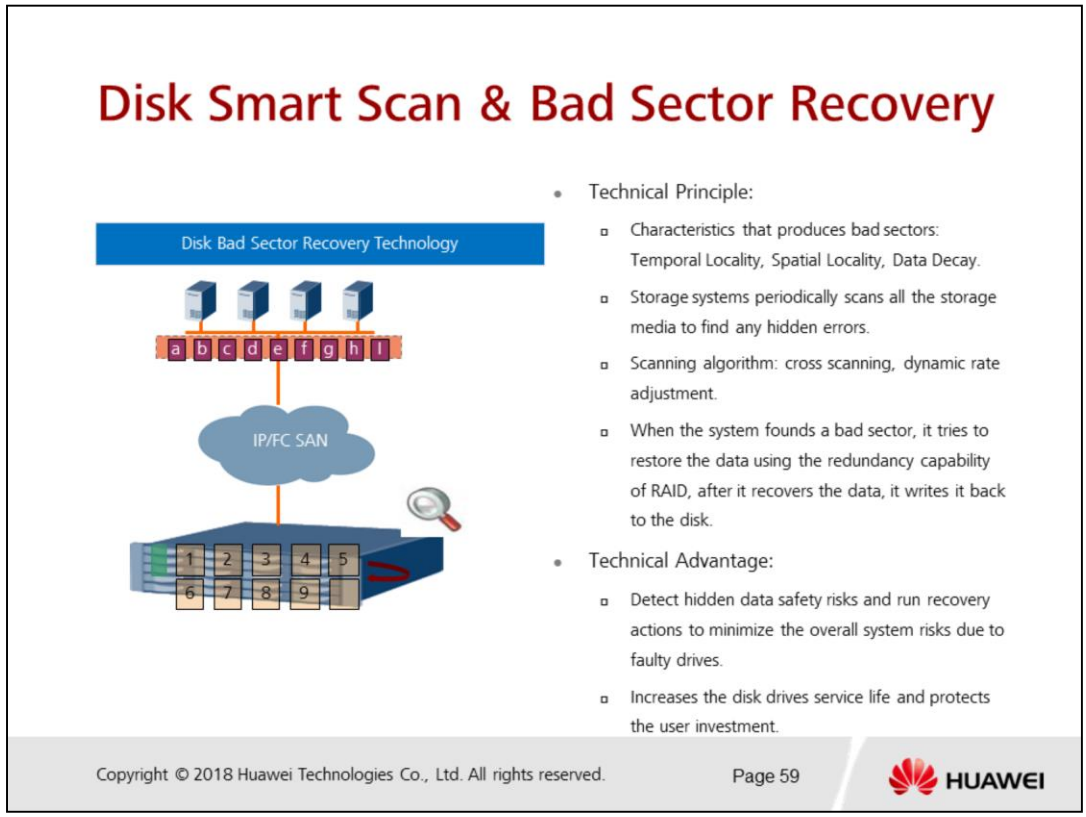
HUAWEI



- Electrical parts and software's function is to move the mechanical parts to complete the data access, retrieval and data storing.
- Hard disks are the component that store the user data in the storage systems, it is a piece of hardware that is frequently accessed for data reading and data writing. Core applications and important business data are stored in these devices which is crucial for it to be reliable when the data within it is needed.
- As disk capacity grows, more and more data is stored on a single drive. This makes it important to ensure that the hard disks used are reliable because a faulty drive may mean hours of data reconstruction or data loss.
- As the hard disks are made from mechanical, electrical and software parts, it means that there are many ways and factors that can cause a drive failure.
- In the following segment, we will look through the different types of disk failures, how to recover from those failures and the types of technologies used to increase the disk reliability to ensure higher data protection.

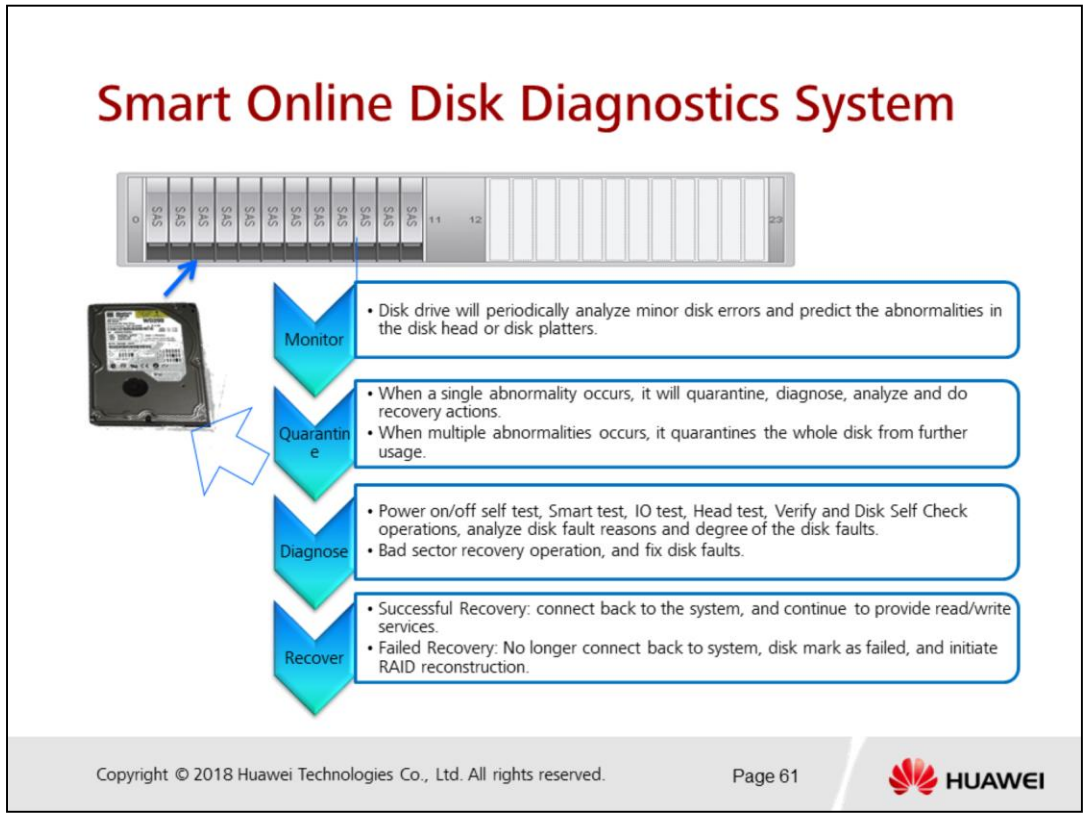


- UNC: Uncorrectable error which is an ECC error in the data field could not be corrected (a media error or read instability).
- IDNF: ID Not Found error where the required cylinder, head, and sector could not be found, or an ECC error occurred in the ID field.
- WP: Write Protected error where data could not be written to the disk.
- Factors that cause bad sectors:
  - Logical bad sectors: The read/write disk head has stains, or dust particles within the hard drive, external disturbances causing instable movement of the head, accidental interruption when data writing(e.g. power issues), vibration.
  - Physical bad sectors: Flaws on the disk platter, metal particles within the hard drive, external shock causing head crash etc.
  - Degeneration of disk head: Electrostatic discharge (ESD) or electrostatic overstress (EOS), material issues of the disk head, bad working environment for the disks accelerates the degeneration (high temperature, instable voltage) etc.
- EOS is a term used to describe the thermal damage that may occur when an IC is subjected to a current or voltage that is beyond the datasheet specification limits of the device. An EOS event can degrade the IC or cause permanent functional failure. EOS is a much slower phenomenon than ESD, but the associated energy is very high.
- Electrostatic discharge is the sudden flow of electricity between two electrically charged objects caused by contact.



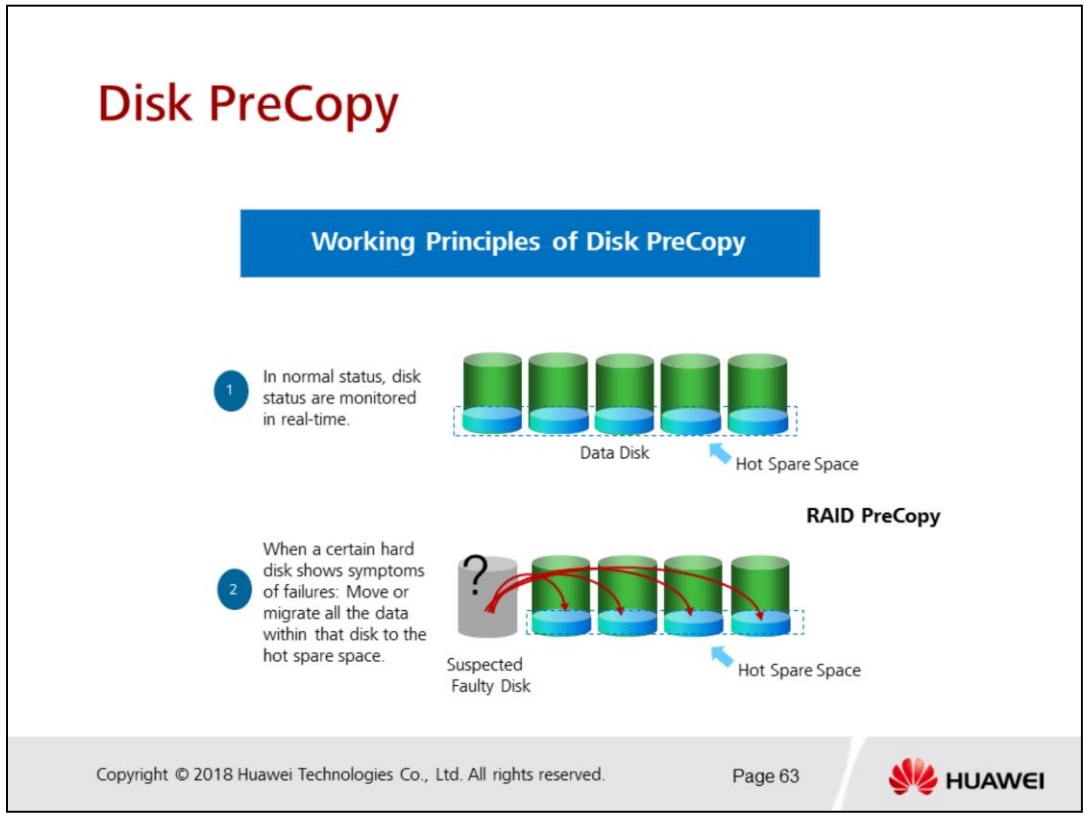
- Bad sectors are often caused during write operations or scratches, and these bad sectors are not found instantly and are only found during data reading process.
- Hence, the background scanning for bad sectors are very important, if not when we read the data, there might be a chance that both data on the RAID stripe is damaged which will cause data loss because it can no longer be recovered. Example: In a RAID 5 with 3 disks scenario, both 1 block of data and 1 block of parity data is corrupted at the same time, which makes the data recovery not possible to recover the lost data.
- As the disk capacities are getting larger and larger, simple sector by sector scanning will take 1 month or even a few months to complete 1 cycle of scanning in large storage systems. This in reality loses the effect of finding the bad sectors quickly in time to recover before further damage is done.
- Huawei has discovered some patterns in the generation of bad sectors during research:
    - If one sector is bad or damaged, the other nearby sectors are likely to be damaged.
    - If a bad sector is found at a certain point of time, at that same point of time there is a high possibility that other sectors might be damaged.
- Based on those pattern found, Huawei designed a scanning algorithm: It will periodically leap through and examine certain sectors within the disk, once it found a bad sector, it will check the nearby sectors and increase the speed of scanning, once it found that the bad sectors is reducing then the speed of scanning is lowered.

- Now, with this algorithm, the accuracy of the background scanning and its time efficiency have increased few times which allows bad sectors to be identified and recovered much more efficiently.
- Locality of reference, also known as the principle of locality, is a term for the phenomenon in which the same values, or related storage locations, are frequently accessed, depending on the memory access pattern. There are two basic types of reference locality – temporal and spatial locality. Temporal locality refers to the reuse of specific data, and/or resources, within a relatively small time duration. Spatial locality refers to the use of data elements within relatively close storage locations. Sequential locality, a special case of spatial locality, occurs when data elements are arranged and accessed linearly, such as, traversing the elements in a one-dimensional array. Locality is merely one type of predictable behavior that occurs in computer systems. Due to the frequent access of these data, the chances of producing bad sectors increases in those sectors of the disks.

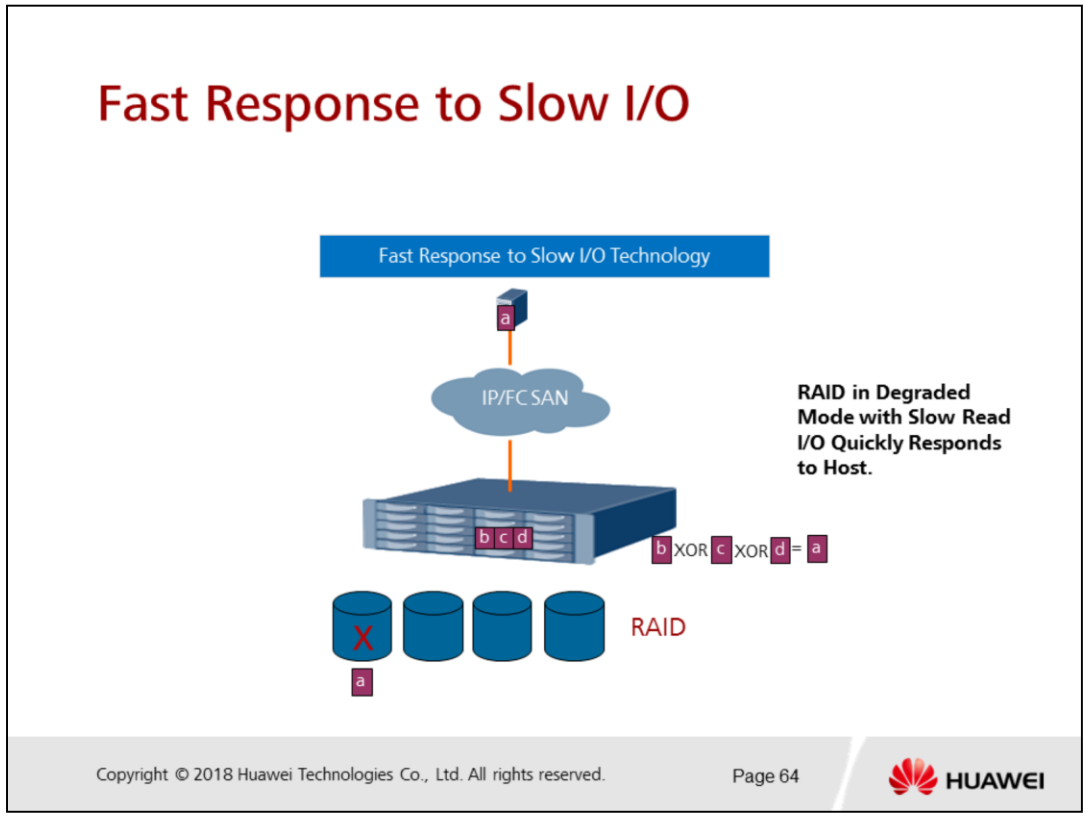


- Advantage: Solves the NPF(No Problem Found) issue (due to abnormality on disk firmware, short external vibration shock or other factors causing disk failures), minimizes the rate of RAID failure, and maximizes the business continuity and data reliability to the greatest extent.
- Based on analysis, annual SATA drives failures amounts to 2.5%, 70% of it due to repairable bad sectors. Periodic scanning on the disks in a 30 days cycle, means that bad sectors on the disks can be found and repaired every 30 days. Hence, we can calculate the annual disk failure rate after implementing the periodic disk scanning to be $\lambda=2.5\%\times0.3+2.5\%\times0.7\times30/365=0.89\%$. The RAID failure rate can also be lowered by 1 level.
- The are two methods to detect the bad sectors on the disks
  - Automatic Analysis of Read/Write Failures: There are many reasons for read/write failures, such as spoilt disk head, disk interface failure, connection cable failure, unplugged hard disks etc. When a disk read/write failure occurs, the storage system will analyze whether the cause of failure is due to bad sectors based on checking the current state of the storage system, current state of hard disk, I/O failure information.

- Active Scanning of Storage Media: Storage systems supports background scanning technologies for the storage media, which uses the idle period of the disk to check itself and find out the bad sectors in time to prevent further accumulated errors. During the scan, storage system has dropped the traditional method of sequentially reading all the sectors within the disk, and directly use the built in media scanning feature in the disk, which avoid the bandwidth consumption at the backplane due to disk scanning. This minimizes the impact on system performance. When the disk that is undergoing scanning receives a read/write request, the scanning operation will stop automatically, and priority is given to process read/write operations. After the read/write process is complete, the scanning operation can resume from the point where it previously stop scanning.



- Disk drives has built in mechanism to predict failures, but those mechanisms are not accurate to fully rely on.
- Huawei has design its own failure prediction mechanism based on the fundamentals of the built in failure detection mechanism of disk drives, which increased the accuracy of the failure detection by a few times.
- Technical Principle:
  - The storage system runs health monitoring on the disks, when it finds that the amount disk errors has went over the preconfigured threshold, it will immediately mark that disk as faulty.
  - RAID group uses the copy method to move the data in the affected disk to the hot spare space, and send an alarm to the admin personnel at the same time to remind them to change the suspected faulty disks.
- Technical Advantage: Greatly lowers the rate of RAID reconstruction, and increase system reliability.



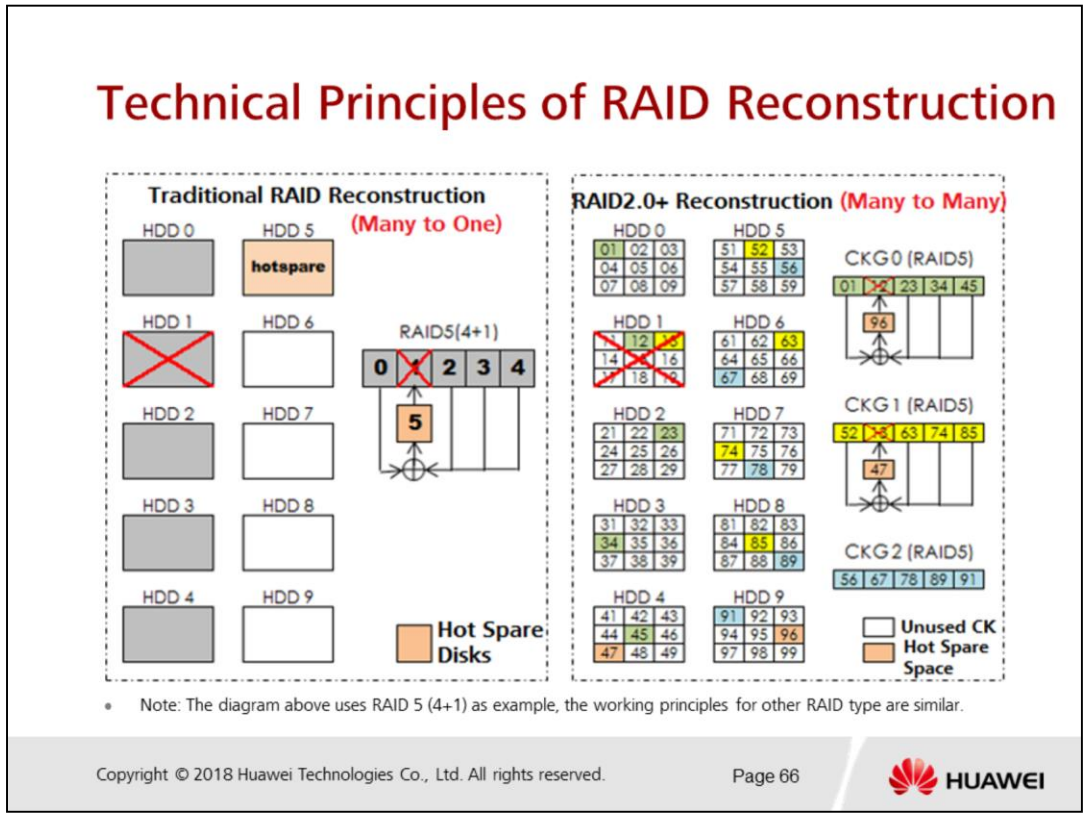
- Disk drives have mechanical parts, so some slow I/O at times is quite normal, and normal application may not even noticed the difference. But for customer core applications, it is quite sensitive towards access delay, the slow I/O may even affect the application or even cause interruption in service.
- Huawei storage fully utilizes the RAID technologies in this kind of scenario, when a certain disk I/O is quite slow, it quickly gives up on waiting for that disk and uses the data in other disks and parity information to quickly generate the data that is required by the core application. This effectively ensures the business continuity of the core applications.
- Technical Principle:
    - Disks causes service to hang or disrupted due to the long I/O response time (seconds level) because of physical bad sectors, head problems, vibration shock, multiple retries etc.
    - Storage system monitor the disk I/O in real-time, if it is slower than the preconfigured threshold, it will use RAID degrade mode to quickly generate the data using parity information and respond to host.
- Technical Advantage: Quickly respond to business services, truly support the core enterprise applications that are very sensitive towards time delay.

Overview of RAID Reconstruction Technologies

Traditional RAID	RAID2.0+
Need to manually configure the independent global or local hot spare disks.	Distributed hot spare space, no need for manual configuration.
Many to one reconstruction, where the reconstructed data are written in sequence to a single hot spare disk.	Many to many reconstruction, where reconstructed data in written in parallel to many disk drives.
Hot spot exists and longer reconstruction time.	Load balancing and short reconstruction time.

HUAWEI

- OceanStor storage system has implemented multiple fault tolerant designs towards the disks, which includes reliable protection features such as online disk diagnostics, DHA (Disk Health Analyzer) for disk diagnostics and failure prediction, background bad sector scanning. RAID 2.0+ technology will automatically reserve a certain amount of hot spare space in the disk domain based on the hot spare policies configured, and the users no longer have to manually configure them independently.
- When the system detects a region on the disk that in unrepairable or the whole disk has failed, the system will automatically begin data reconstruction, which will quickly reconstruct the affected data to the hot spare space of other hard disks to achieve fast self healing of the storage system.

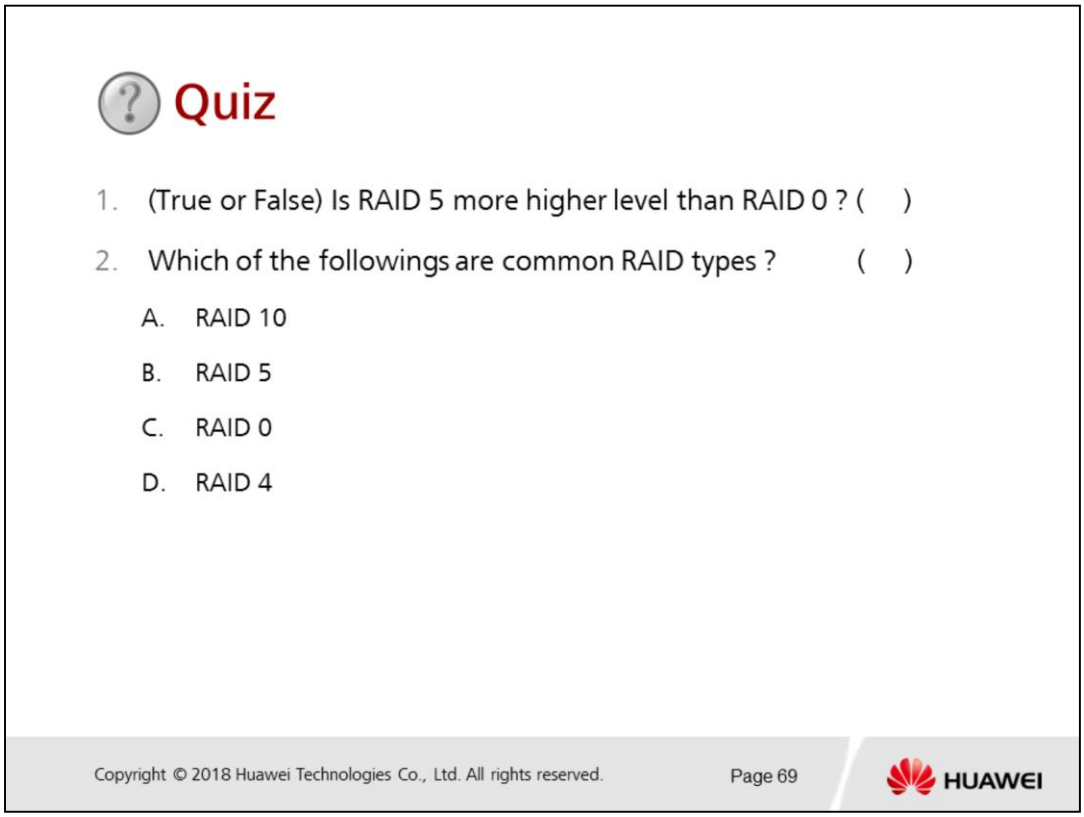


- The growth of disk capacity makes Traditional RAID to face a serious problem: 10 years ago, reconstructing a disk only takes minutes, but now a single disk reconstruction could take few hours or dozens of hours. Longer disk reconstruction makes the enterprise storage to stay in a Degraded RAID mode without any fault tolerance for long hours, this poses a huge risk for data loss. Real life scenarios such as data loss due the pressure of data reconstruction process and business operations during the reconstruction process are quite common.
- RAID2.0+ technology based on low level block virtualization solves the problem of the bottleneck target disk (hot spare disk) in traditional RAID reconstruction, making the write bandwidth of the reconstructed data flow no longer the bottleneck in data reconstruction speeds. This effectively increased the reconstruction speed greatly and lowered the rate of dual disk failures, and at the same time increased the overall system reliability.

- The diagram above shows the comparison of RAID reconstruction principles of Traditional RAID and RAID2.0+:
    - On the left part of the diagram, which shows the traditional RAID, consists of HDD0~HDD4 which forms a RAID 5 with HDD5 as the hot spare disk. When HDD1 has failed, other disks which are HDD0, HDD2, HDD4 and HDD4 reconstruct the data using XOR operation into the hot spare disk HDD5.
    - On the right part of the diagram, which shows the RAID2.0+ technology, when the HDD1 fails, the data in HDD1 is reconstructed in the form of CK, and only the allocated and used CK are reconstructed (CK12 and CK13 as shown in diagram). All the hard disks within the storage pool are involved within the reconstruction process, and the reconstructed data is distributed across multiple disk drives (HDD4 and HDD9 as shown in diagram).

Summary

- This module introduced:
  - Concepts, Working Principles and Types of RAID Technologies.
  - Technical Principles of RAID2.0+ Technologies.
  - Functions of Host Multipathing Technologies.
  - Disk Reliability Technologies.



- Answers:
  - F.
  - ABC.

Thank You

www.huawei.com