Semi-RAID: A Reliable Energy-Aware RAID Data Layout for Sequential Data Access

Li Xiao   Tan Yu-An   Sun Zhizhuo
Beijing Institute of Technology

Outline

- Introduction to RAID architecture
- Storage requirement of video surveillance system
- Pros and Cons of traditional RAID architecture
- The idea of Semi-RAID (S-RAID)
- S-RAID 4 and S-RAID 5 data layout
- Improvement of S-RAID by grouping
- Power consumption and Performance of S-RAID

Introduction to RAID architecture

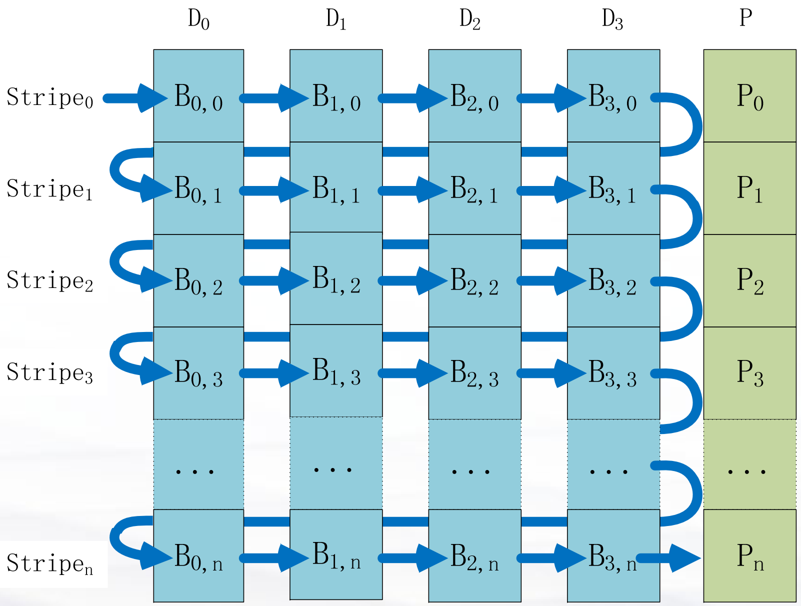
What is RAID?

- RAID combines multiple disk drive components into a logical unit.
- Data is distributed across the drives in one of several ways called "RAID levels".
- Advantages of RAID：
  - Increasing the reliability of data storage
  - Improving the read/write performance of data access

Introduction RAID architecture

RAID-4

- Improve performance by block-level striping
- Exploit XOR parity for fault tolerance
- Use one dedicated parity disk (bottleneck)



	D_0	D_1	D_2	D_3	P
Stripe_0	$B_{0,0}$	$B_{1,0}$	$B_{2,0}$	$B_{3,0}$	P_0
Stripe_1	$B_{0,1}$	$B_{1,1}$	$B_{2,1}$	$B_{3,1}$	P_1
Stripe_2	$B_{0,2}$	$B_{1,2}$	$B_{2,2}$	$B_{3,2}$	P_2
Stripe_3	$B_{0,3}$	$B_{1,3}$	$B_{2,3}$	$B_{3,3}$	P_3
	. . .	. . .	. . .	. . .	. . .
Stripe_n	$B_{0,n}$	$B_{1,n}$	$B_{2,n}$	$B_{3,n}$	P_n

$$P = \bigoplus_{i} D_i = D_0 \oplus D_1 \oplus \cdots \oplus D_{n-1}$$

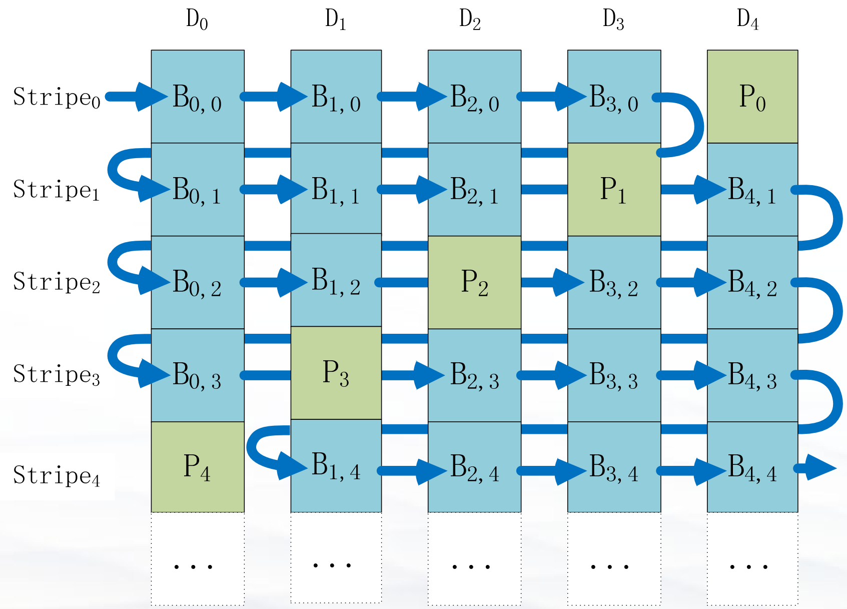
$$P_k = \bigoplus_{i=0}^{n-1} B_{i,k}$$

$$= B_{0,k} \oplus B_{1,k} \oplus \cdots \oplus B_{n-1,k}$$

Introduction to RAID architecture

RAID-5

- Improve performance by block-level striping
- Exploit XOR parity for fault tolerance
- Distribute parity blocks across data disk



	D_0	D_1	D_2	D_3	D_4
$Stripe_0$	$B_{0,0}$	$B_{1,0}$	$B_{2,0}$	$B_{3,0}$	P_0
$Stripe_1$	$B_{0,1}$	$B_{1,1}$	$B_{2,1}$	P_1	$B_{4,1}$
$Stripe_2$	$B_{0,2}$	$B_{1,2}$	P_2	$B_{3,2}$	$B_{4,2}$
$Stripe_3$	$B_{0,3}$	P_3	$B_{2,3}$	$B_{3,3}$	$B_{4,3}$
$Stripe_4$	P_4	$B_{1,4}$	$B_{2,4}$	$B_{3,4}$	$B_{4,4}$
	. . .	. . .	. . .	. . .	. . .

Storage requirement of video surveillance system

- Large storage capacity
  - A disk array of 16 2TB-disk has 30TB available capacity
  - Assume 2Mb/s video code rate，30TB storage space is capable for video data of (Day · Channel)：
    - 24*3600s*2Mb/s=24*3600*0.25MB=21.6GB
    - 30TB/21.6GB=30000GB/21.6GB=1388 Days
  - With One camera installed, The disk array can keep 1388 days' video data.
  - With 32 cameras installed，The disk array can keep 1388/32≈43 days' video data.

Storage requirement of video surveillance system

- High reliability
  - ➢ Users of video surveillance system (airport, prison, etc.) need to meet strict regulations
  - ➢ video surveillance system runs 7X24 hours
  - ➢ The video fragment loss will cause extreme high risk ， so the intact of data must be guaranteed .
  - ➢ Performance of the video surveillance must be guaranteed in degraded mode and rebuild mode of RAID.

Storage requirement of video surveillance system

- Moderate performance
  - To support 32 cameras saving video data concurrently, the disk array should have a write bandwidth of
  - 32*2Mb/s=32*0.25MB/s=8MB/s
  - Indeed, 32 cameras only write 8MB data every second.
  - 100 Cameras: 100*0.25MB/s=25MB/s

Advantages of traditional RAID

- Large Storage Capacity requirement
  - ➢ It can be satisfied
  - ➢ Disk array supports at least 16 disks，scales well through Disk Expansion Enclosure。
- Data protection requirement
  - ➢ It can be satisfied
  - ➢ Use RAID-4/5, data can be rebuilt during disk failure
- Performance requirement
  - ➢ It can be satisfied，but don’t take full advantage of the performance of disk array
  - ➢ 32 cameras only need 8MB/s， 100 cameras only need 25MB/s

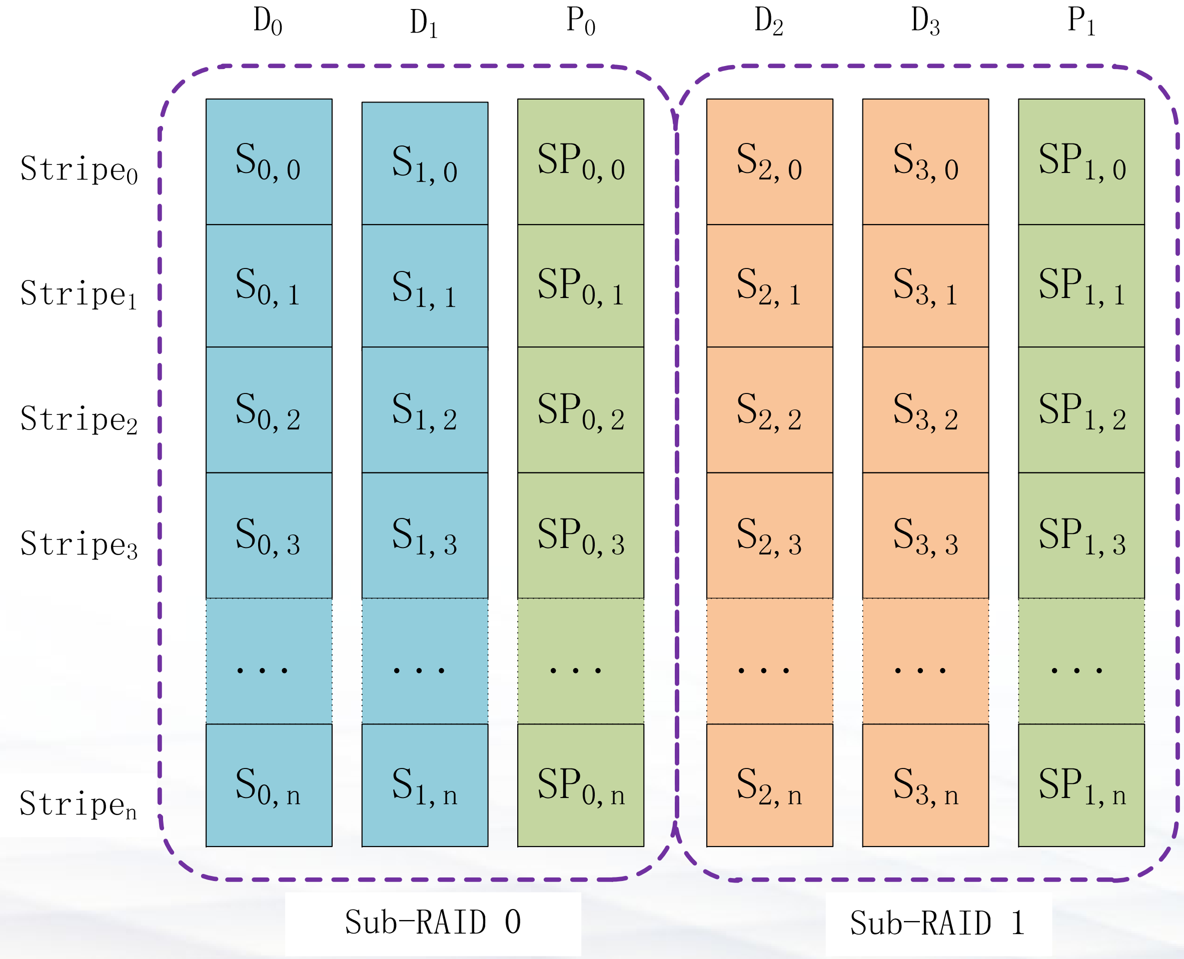
Disadvantages of traditional RAID

- High failure rate of individual disk
  - ➢ Disk lifetime depends on its working hours
  - ➢ All disks in RAID work 7X24 hours

  - ➢ Current solution: divides RAID into sub-RAID systems, idle sub-RAID can be put into sleep
    - ➢ Every sub-RAID needs a separate parity disk
    - ➢ Management of the sub-RAIDs is complicated.

Disadvantages of traditional RAID

Partitioning the storage system into RAID systems.



	D_0	D_1	P_0	D_2	D_3	P_1
Stripe$_0$	$S_{0,0}$	$S_{1,0}$	$SP_{0,0}$	$S_{2,0}$	$S_{3,0}$	$SP_{1,0}$
Stripe$_1$	$S_{0,1}$	$S_{1,1}$	$SP_{0,1}$	$S_{2,1}$	$S_{3,1}$	$SP_{1,1}$
Stripe$_2$	$S_{0,2}$	$S_{1,2}$	$SP_{0,2}$	$S_{2,2}$	$S_{3,2}$	$SP_{1,2}$
Stripe$_3$	$S_{0,3}$	$S_{1,3}$	$SP_{0,3}$	$S_{2,3}$	$S_{3,3}$	$SP_{1,3}$
	...	...	...	...	...	...
Stripe$_n$	$S_{0,n}$	$S_{1,n}$	$SP_{0,n}$	$S_{2,n}$	$S_{3,n}$	$SP_{1,n}$

Disadvantages of traditional RAID

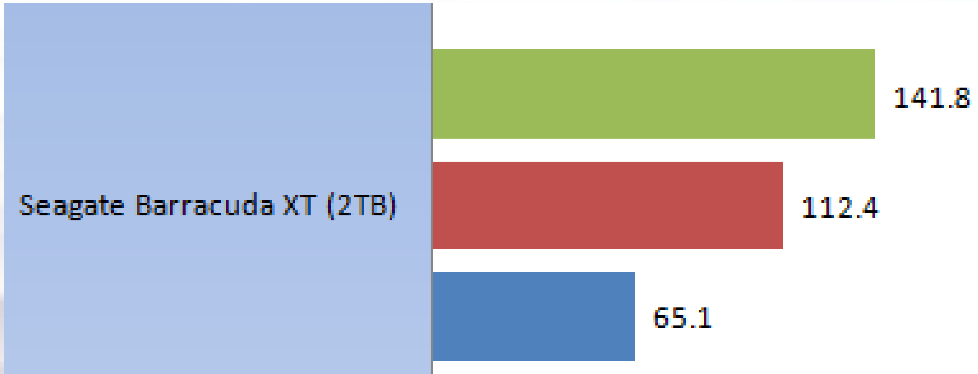
- High power consumption
  - Video surveillance has a moderate requirement on performance, it can be satisfied by one or a few disks' bandwidth.
    - **To meet the high performance requirement, a plurality of disks must work parallel**
  - All disks in RAID work 7X24 hours, consuming large amount of energy.
    - **The high performance provided by parallel working disks cannot be exploited by video surveillance application**
  - The heat generated by high-load disks needs extra cooling system

Observations

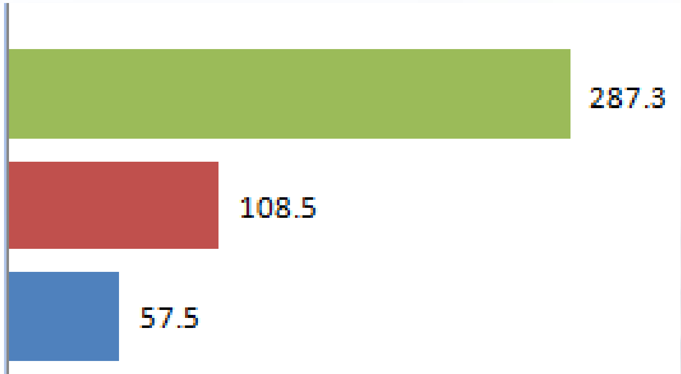
- Video surveillance system doesn’t need many disks work parallel
  - Rearranging the video stream data sequentially can save multi-channel video data into a single disk
  - Performance of a single disk can satisfy the requirement
    - The sequential write speed of SATA disk is around 100MB/s

Seagate 2TB SATA Test Results:

HDD Tune Pro(read)



HDD Tune Pro(write)



The ideas of Semi-RAID (S-RAID)

- Target:
  - In a RAID system，All disks are not working parallel, only a few disks needs to be in active mode.
  - Preserve the data protection function of tradition RAID

S-RAID is not DVR (DVR is equivalent to Non-RAID)

- DVR writes data to a single disk，and only move to the second disk when the first disk is full
- DVR has no data protection function

The ideas of Semi-RAID (S-RAID)

Ideas:

- S-RAID doesn’t need stripping
- Rearrange the data layout to make it suitable for video surveillance and other applications alike。

S-RAID 4

S-RAID 4 resembles RAID 4 in:

- Data is stored in data blocks
- Exploit XOR parity for fault tolerance
- Use one dedicated parity disk (bottleneck)

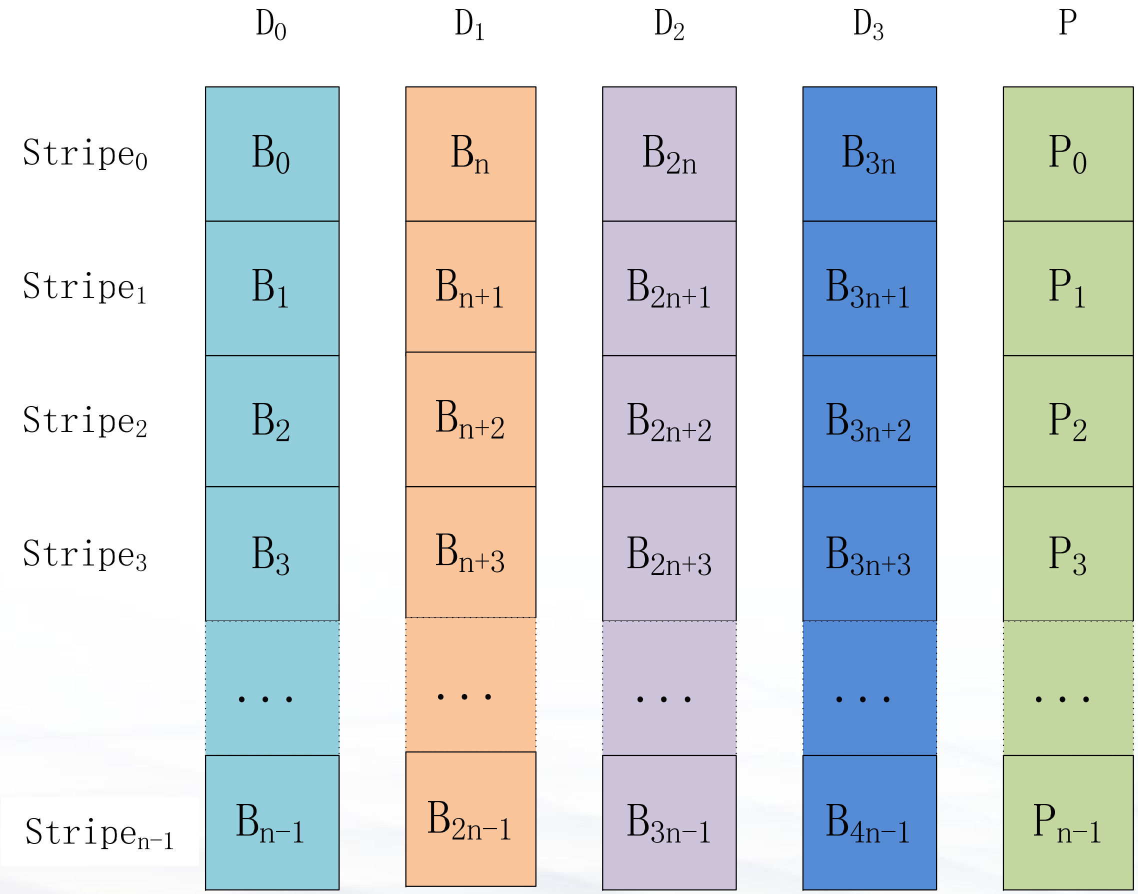
S-RAID 4 differs from RAID 4 in:

- Data Layout (like N-RAID)

  Disks cannot work in parallel to increase performance when reading / writing LBA adjacent data blocks

S-RAID 4 Data Layout

- Data Layout of Semi-RAID 4



	D_0	D_1	D_2	D_3	P
Stripe_0	B_0	B_n	B_{2n}	B_{3n}	P_0
Stripe_1	B_1	B_{n+1}	B_{2n+1}	B_{3n+1}	P_1
Stripe_2	B_2	B_{n+2}	B_{2n+2}	B_{3n+2}	P_2
Stripe_3	B_3	B_{n+3}	B_{2n+3}	B_{3n+3}	P_3
	. . .	. . .	. . .	. . .	. . .
Stripe_{n-1}	B_{n-1}	B_{2n-1}	B_{3n-1}	B_{4n-1}	P_{n-1}

Read and Write Operation in S-RAID

- Read operation is the same as RAID 4
  - ➢ Reading from standby disk needs to wake up the disk
- In the case of sequential write operation:
  - ➢ Use disk 1 first, then use disk 2, and so on…
  - ➢ Only one data disk and one parity disk are active at a time
  - ➢ All other data disks are in standby mode
  - ➢ While writing to the data disk, the parity should be recomputed at the same time (not like N-RAID)
- In the case of random write operation:
  - ➢ Other data disks may be woken up

Optimizations of S-RAID 4

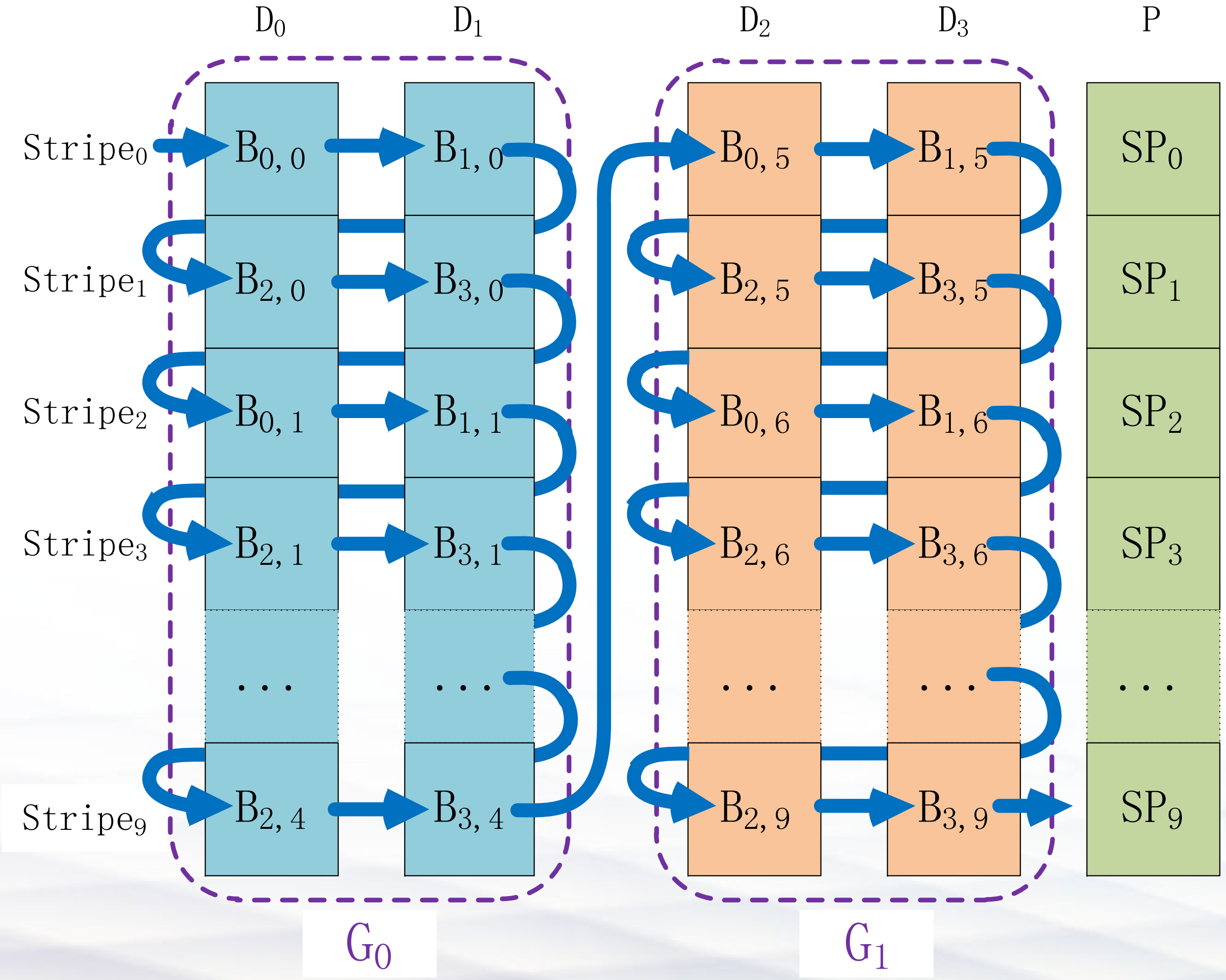
- One write request needs 4 I/O operation:
  - Read old data and old parity;
  - Write new data and new parity;
- Optimizations
  - Readahead (read old data and parity in large chunk)
  - Aggregation (aggregate new data into large chunk before write to disk);
  - Caching;
- Test results:
  - Single disk Seq Read/Write: 110MB/s / 107MB/s。
  - 16MB Cache, 1 disk S-RAID 4 Seq Write: 32MB/s。
  - 1GB Cache, 1 disk S-RAID 4 Seq Write: 52MB/s 。

Grouping S-RAID 4

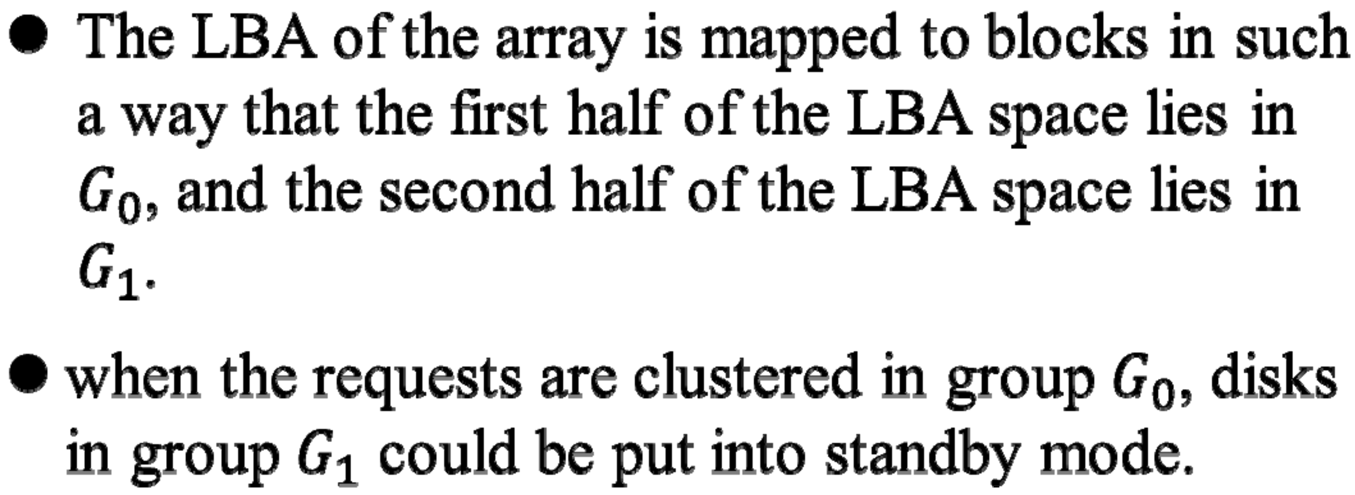
- Performance limitation
  - S-RAID4 has only one data disk work at one time
- Grouping Strategy
  - Allow more than one disk (group) working at the same time
  - use stripping in each group， write data blocks parallel
  - The more disks there are in active mode， the higher the performance will be. But power consumption increases and disk lifetime decreases accordingly.
    - When all disks are in active mode， the disk array is equivalent to a traditional RAID
  - The size of group is fixed， thus it must be planned in advance
  - Grouping can be used in both S-RAID 4 and S-RAID 5

S-RAID 4 Group Data Layout

- Group Data Layout of Semi-RAID 4



S-RAID 4 Group Data Layout



- The LBA of the array is mapped to blocks in such a way that the first half of the LBA space lies in G_0, and the second half of the LBA space lies in G_1.
- when the requests are clustered in group G_0, disks in group G_1 could be put into standby mode.

A group includes at least a whole data disk, therefore there is enough LBA space in one group for the sequential request to cluster in.

Advantages of S-RAID 4

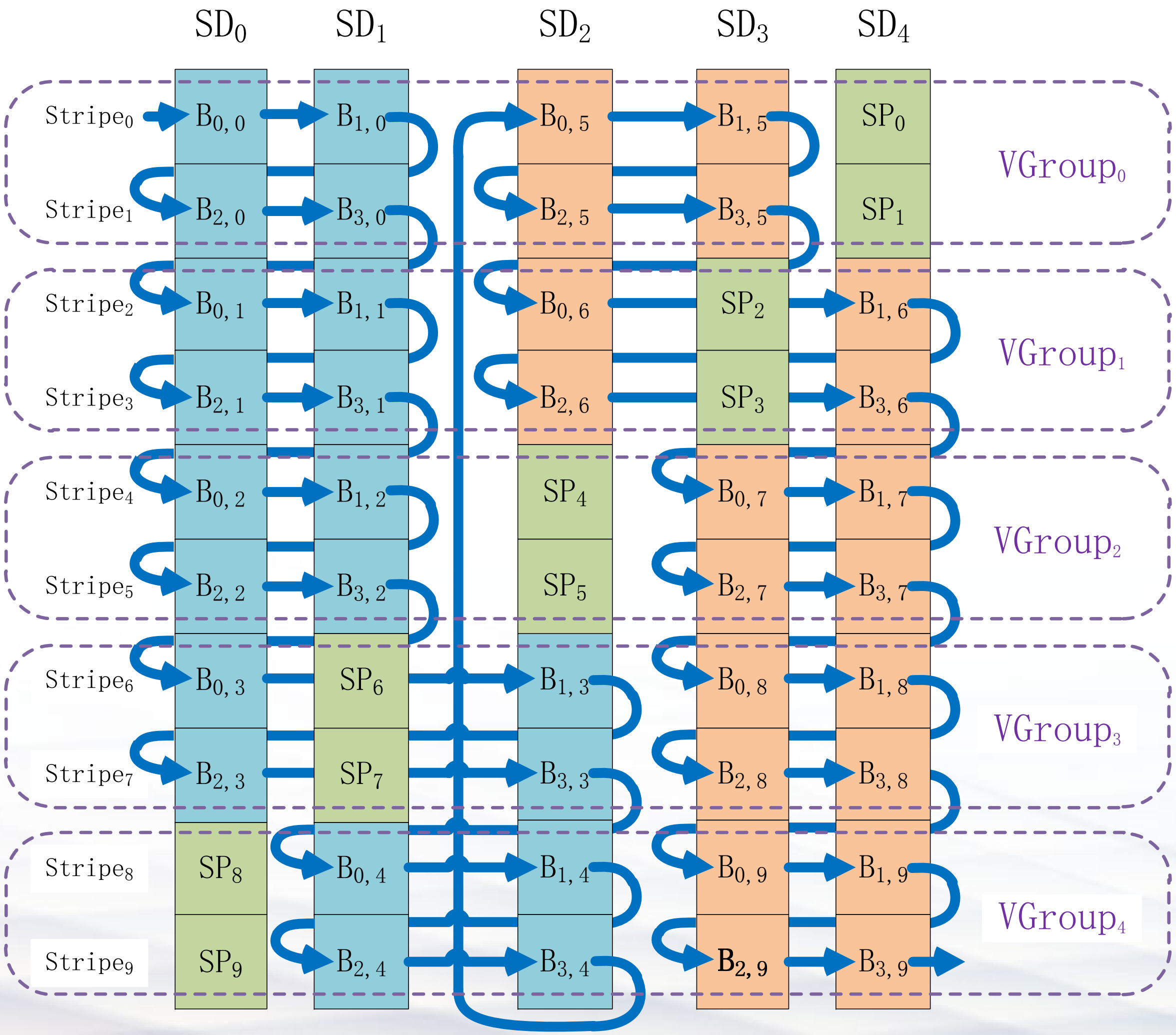
- Reduce the power consumption
  - Only part of the disks array are in active mode at the same time
- Enhance disk reliability
  - Working hours of individual data disk is much shorter than the working hours of the disk array
- Protect data from disk failure
  - S-RAID 4 data layout is like N-RAID data layout, but S-RAID provides data protection function of the traditional RAID

Limitation of S-RAID 4

- S-RAID 4 uses a fixed parity disk like the traditional RAID 4, hence the parity disk may also become a bottleneck.
- This not only affects the performance but also reduces reliability, because parity disk cannot be put into standby mode.
- To ease the bottleneck of parity disk, we introduce the S-RAID 5 data layout that uniformly distributes parity blocks among the disks.

4. Semi-RAID Data Layout

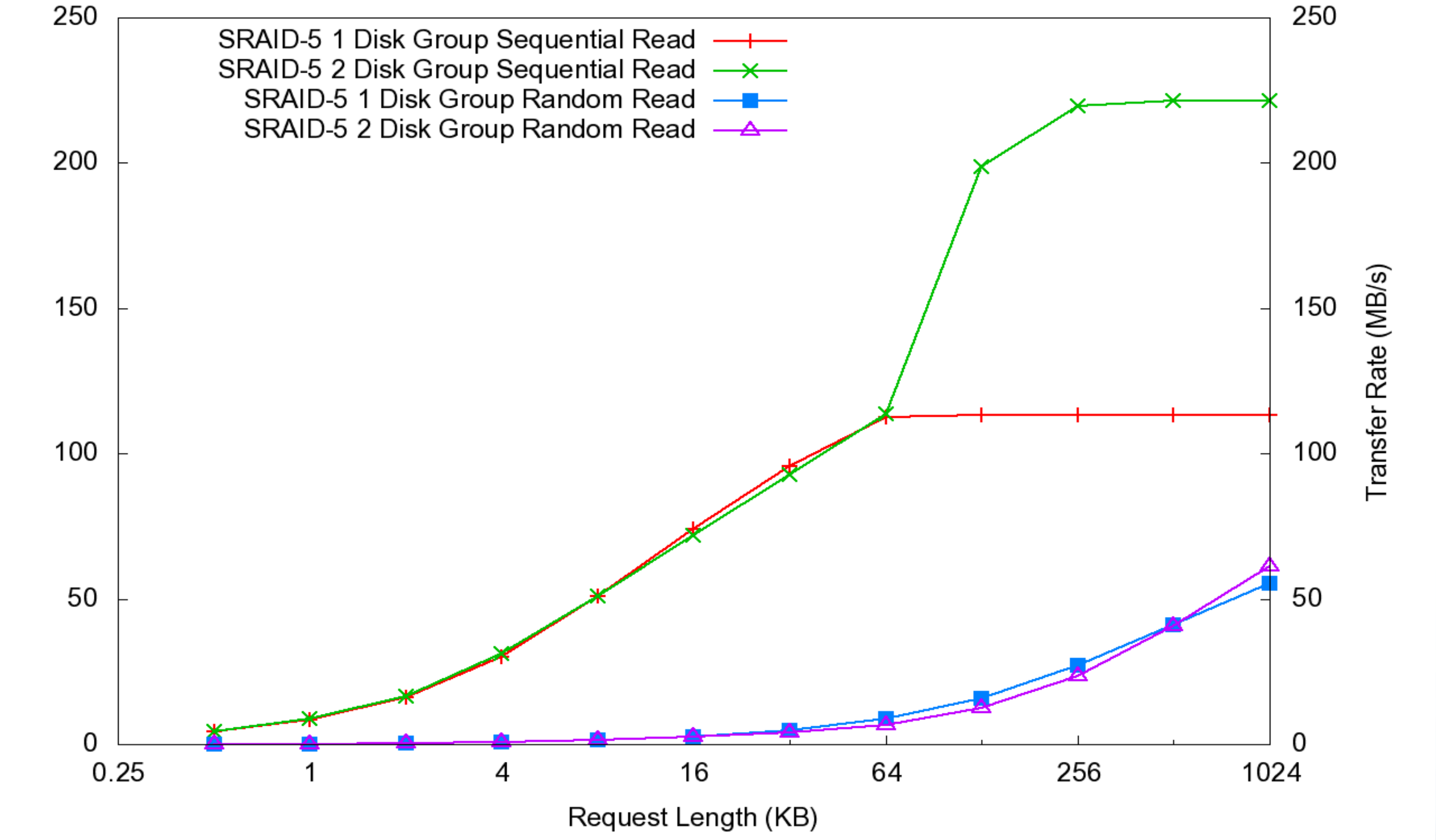
Data Layout of Semi-RAID 5



	SD_0	SD_1	SD_2	SD_3	SD_4	
$Stripe_0$	$B_{0,0}$	$B_{1,0}$	$B_{0,5}$	$B_{1,5}$	SP_0	$VGroup_0$
$Stripe_1$	$B_{2,0}$	$B_{3,0}$	$B_{2,5}$	$B_{3,5}$	SP_1	
$Stripe_2$	$B_{0,1}$	$B_{1,1}$	$B_{0,6}$	SP_2	$B_{1,6}$	$VGroup_1$
$Stripe_3$	$B_{2,1}$	$B_{3,1}$	$B_{2,6}$	SP_3	$B_{3,6}$	
$Stripe_4$	$B_{0,2}$	$B_{1,2}$	SP_4	$B_{0,7}$	$B_{1,7}$	$VGroup_2$
$Stripe_5$	$B_{2,2}$	$B_{3,2}$	SP_5	$B_{2,7}$	$B_{3,7}$	
$Stripe_6$	$B_{0,3}$	SP_6	$B_{1,3}$	$B_{0,8}$	$B_{1,8}$	$VGroup_3$
$Stripe_7$	$B_{2,3}$	SP_7	$B_{3,3}$	$B_{2,8}$	$B_{3,8}$	
$Stripe_8$	SP_8	$B_{0,4}$	$B_{1,4}$	$B_{0,9}$	$B_{1,9}$	$VGroup_4$
$Stripe_9$	SP_9	$B_{2,4}$	$B_{3,4}$	$B_{2,9}$	$B_{3,9}$	

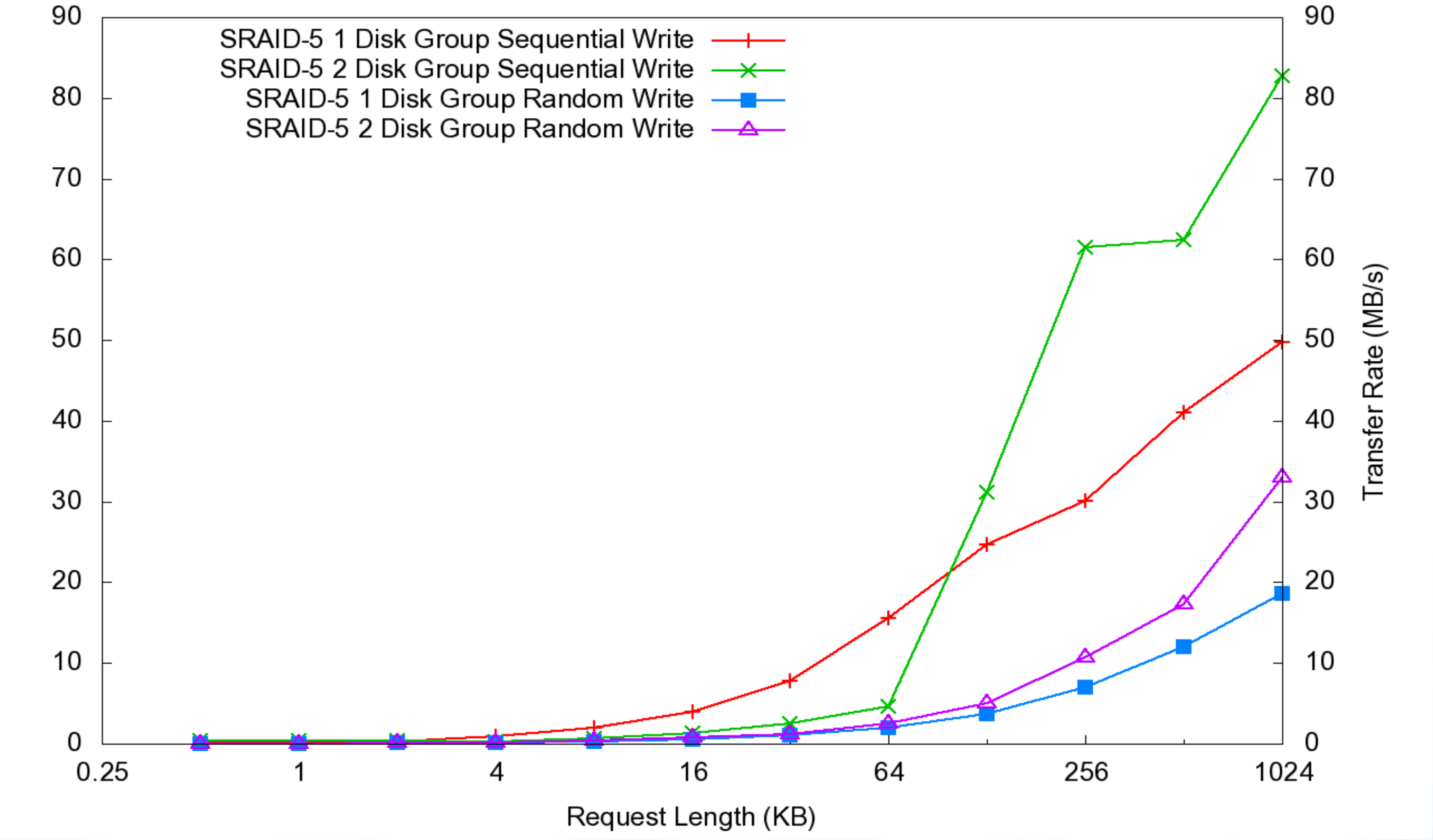
Performance of S-RAID 5

- Read Performance of S-RAID 5



Performance of S-RAID 5

- Write Performance of S-RAID 5



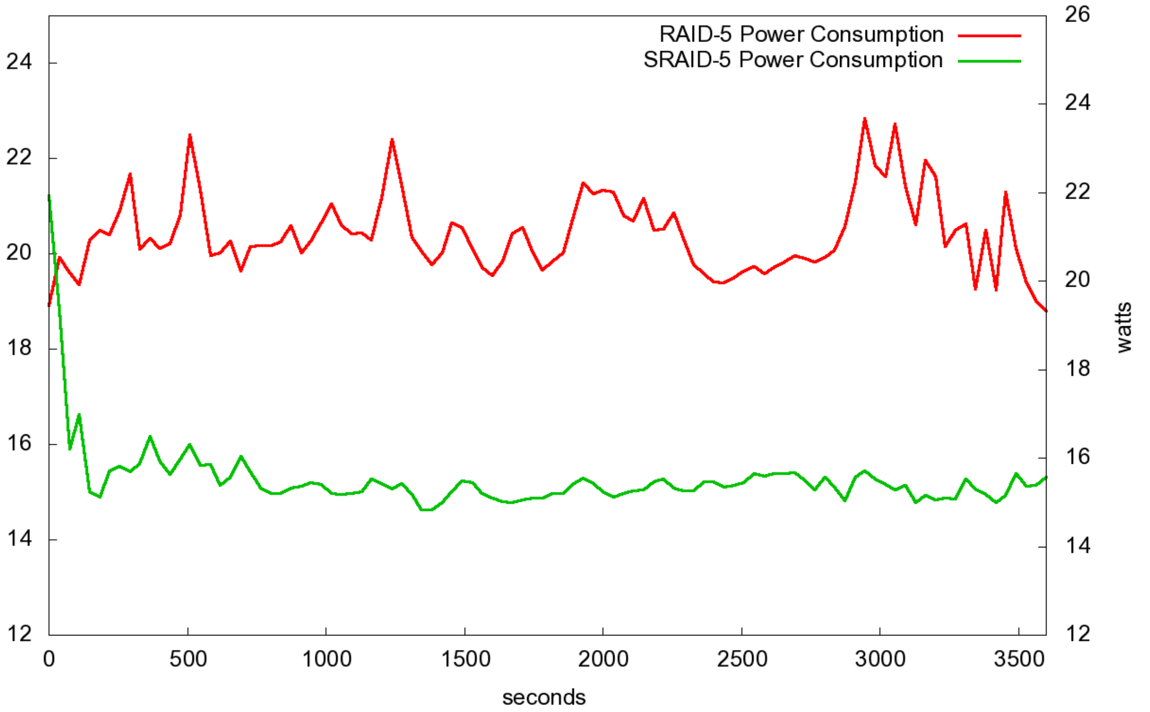
Power Consumption of S-RAID 5

To evaluate the power saving effect of S-RAID in actual situation, we test the power consumption of a video surveillance system with 32 digital cameras.

We run the experiment for a time period of 1 hour and measure the power consumption of each disk in the S-RAID 5 every second.

Power Consumption of S-RAID 5

Experiment Results:



The S-RAID 5 includes 5 Seagate ST3500418AS 500G 7200RPM Disks, and is divided into 2 groups of 2 disks. The number of vertical group is set to 5, the same as the number of the disks.

Conclusion

- S-RAID is an alternative RAID data layout optimized for sequential data access, S-RAID provides extra reliability and high energy efficiency.
- The trade-off is that, the performance drops in S-RAID especially for write request. So, S-RAID is only suitable for applications like video surveillance, CDP, VTL, etc.
- S-RAID addresses performance issue by adjusting the group size.

Conclusion

- Applicable scenarios: (Sequential data access)
  - Video surveillance
  - CDP (Continuous Data Protection)
  - VTL (Virtual Tape Library)
- Inapplicable scenarios:
  - Database (exhibits random data access pattern)
  - Video-on-demand, File sever, etc.(ask for high performance)

Further Work

- Set and manage dynamic group size in S-RAID, Therefore the same S-RAID can adapt to the variations of data transfer rate of the application.
- Design fine-grained schedule algorithm for disk spin-down and spin-up. instead of waiting for the idle disk for a constant length of time
- Exploit log-structure file system to obtain sequential write workloads.

Thank you!