“Big Cache”, B+Tree Full Associative Very Large Scale DRAM-based Storage Cache

- Introduction of “Super Storage” using “Big Cache” technology -

Super SSD (G5) / Enterprise

Super SSD (G5) / Professional

Super RAID V

Super RAID TIER CACHE

Agenda

[1] Motivation of developing “Big Cache” technology
[2] Concept and basic idea of “Big Cache”
[3] Problem on the existing storage cache system
[4] Solution, B+Tree cache entry search
[5] Introduction of features and specifications
[6] Performance view and characteristics
[7] General comparison on HDD and SSD
[8] Effects and benefit for customers
[9] Conclusion (“Big Cache” adaptation products)

Motivation of developing “Big Cache”

- Background
  - Developed software for DRAM based solid state disk at first.
  - Product release as “Solid STOR” with maximum 64 GB from 2004.
  - Focused the market of database application and engineering use.
- Problem of DRAM based solid state disk
  - Expensive cost for adapting big data applications
  - Not easy, separating hot data of application to the limited capacity
  - Strong requirement to unlimited capacity of solid state disk

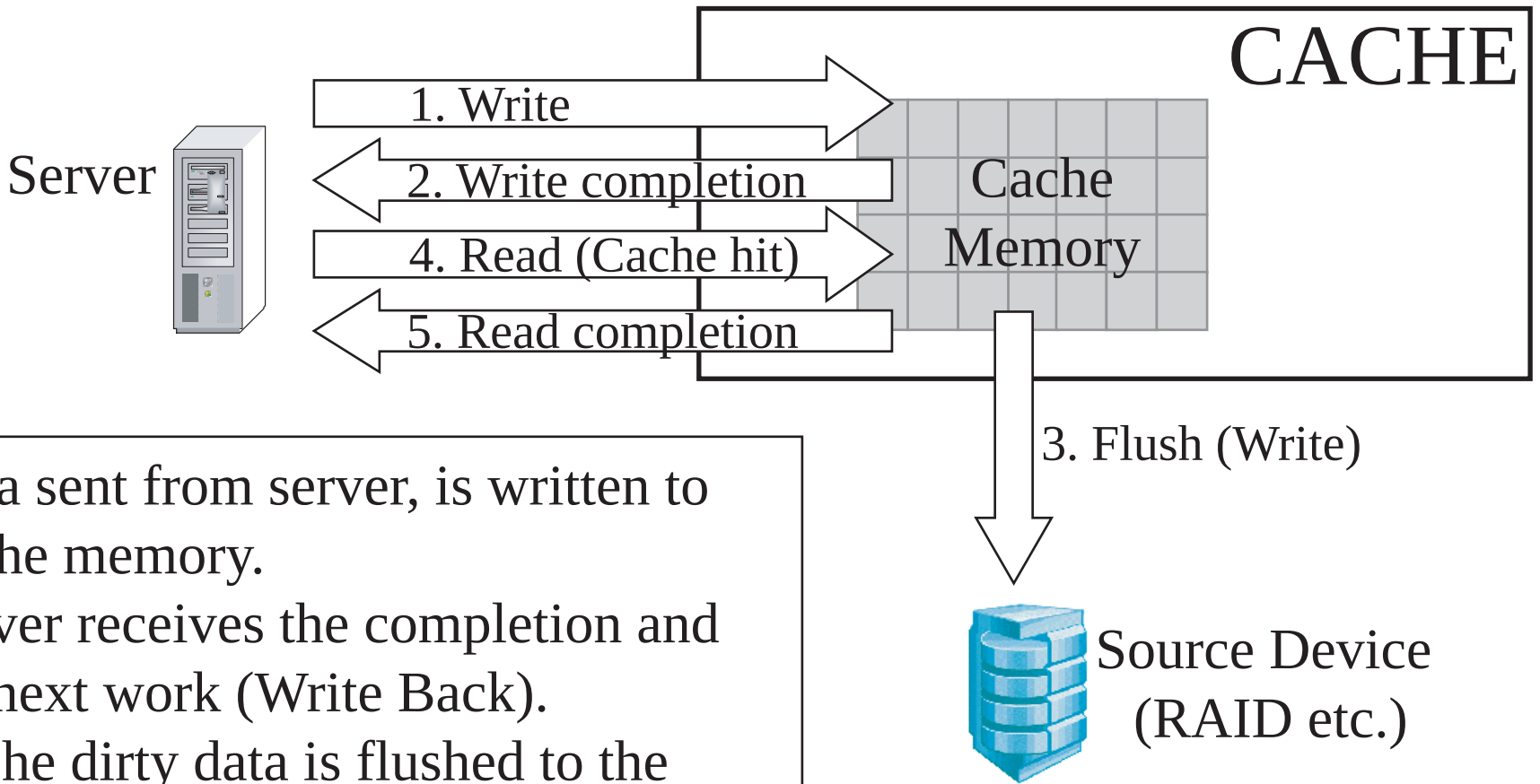
→ Developing “Big Cache” technology

→ Production as “Super Storage” series (2009)

→ Generation 5, production release as “Super Storege(G5)” (2014)

- 12G SAS Host I/F, 12G SAS RAID controller
- Intel Xeon Ivy Bridge v2 platform

Outline of data flow on storage cache



1. Data sent from server, is written to cache memory.
2. Server receives the completion and go next work (Write Back).
3. Cache dirty data is flushed to the source device (RAID etc) later.

Basic idea on “Big Cache”

- Application will refer only written data by own.
- Storage capacity is very large (Ex. several 10 / 100 TB), but accessed hot data size is limited (Ex. several 100 GB) in closing duration (Ex. one day).
- We need the cache capacity which application writes data frequently.
- If we can implement such very large scale storage cache, ...

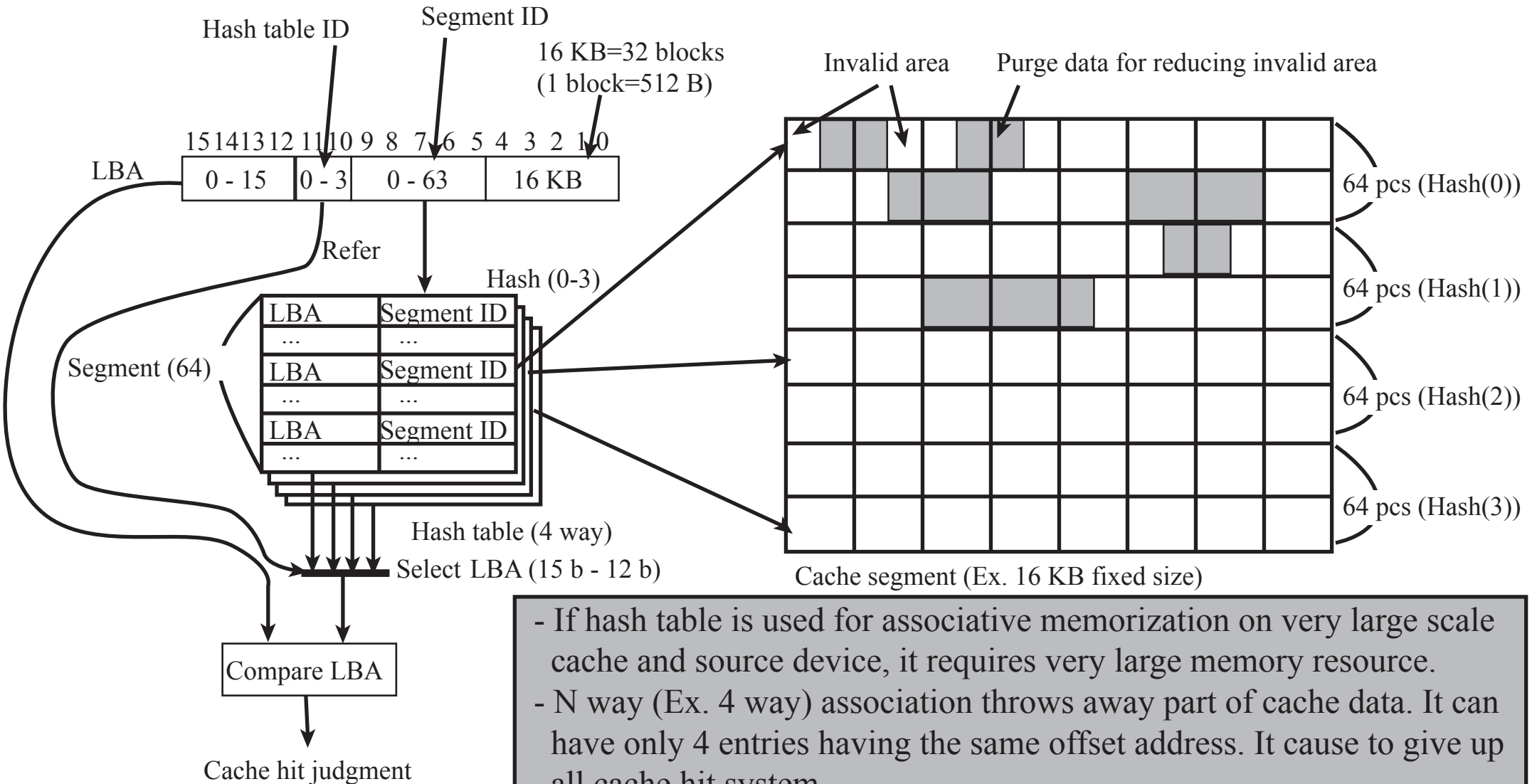
- We expect all cache hit on “Big Cache”.
- It behaves as very large DRAM based solid state disk and we can provide it to customers with effective price.
- We can find synergistic effect of DRAM cache and flash SSD also.

Basic concept and idea of “Super Storage”

Media	Features
DRAM	- Very high speed on R/W, random / sequential I/O. - Very wide bandwidth, very high speed and good stability , >40 GB/s throughput, >2 million IOPS. But it is very expensive as storage media.
Flash SSD	- High speed on sequential/random read access, 600 MB/s and 90K IOPS (read), SLC or enterprise-level MLC. - Weakness point on random write I/Os for a long time stress. - Not high performance on synchronous I/O access. Higher IOPS on many IO outstandings and asynchronous I/Os. Because Flash SSD improves performance with parallel I/Os processing by internal multiple channels. - Large capacity, 400 GB (SLC), 800 GB / 1.6 TB (MLC)
Magnetic Had Disk	- Middle-range speed on sequential access, 140 MB/s (outer side), 80 MB/s (inner side). - Very low IOPS on random I/O, 250 IOPS to 300 IOPS. - Very low price of bit cost. - Very large capacity, 900 GB (SAS), 4 TB (SATA).

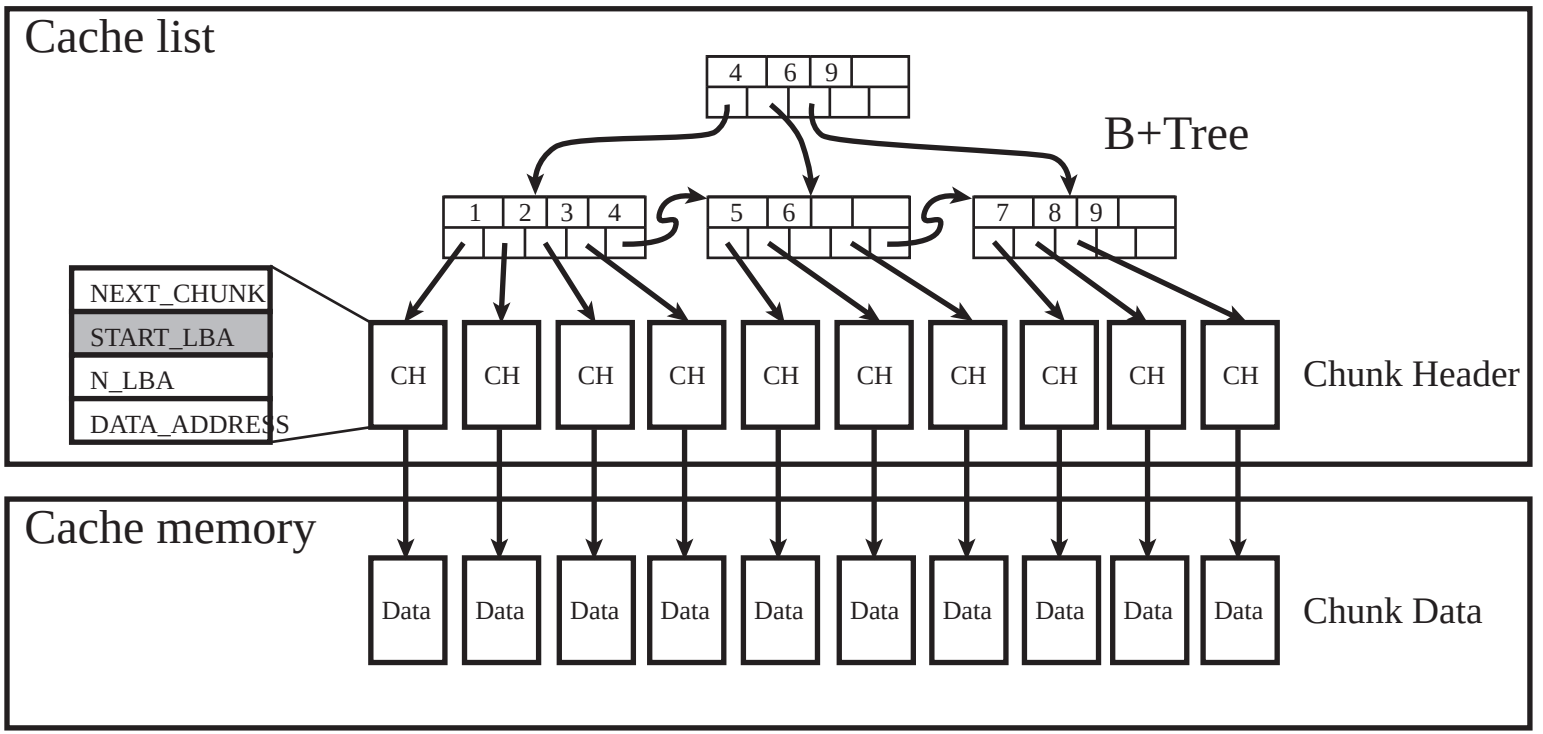
➡ Taking each strong point of each media that is, “Super Storage”

Problem on existing storage cache system



- If hash table is used for associative memorization on very large scale cache and source device, it requires very large memory resource.
- N way (Ex. 4 way) association throws away part of cache data. It can have only 4 entries having the same offset address. It cause to give up all cache hit system.
- Non-aligned data creates invalid area on cache memory.

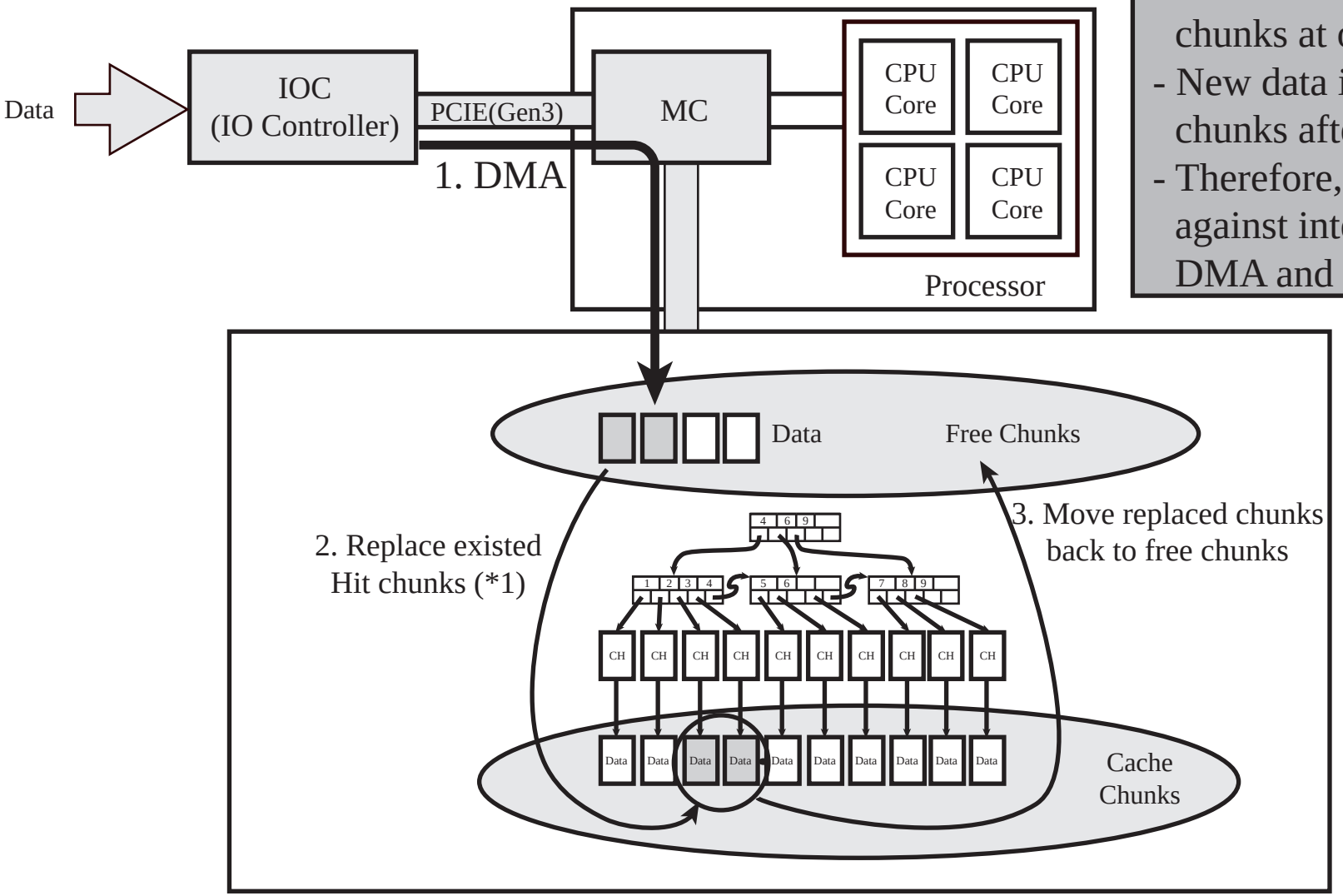
Cache entry search by B+Tree algorithm



- Full associative cache can realize all cache hit system.
- Start address information in chunk header can solve non-aligned address issue.
- All cache hit system and effective-use of memory resource for cache capacity.

Zero-copy data transfer (Write)

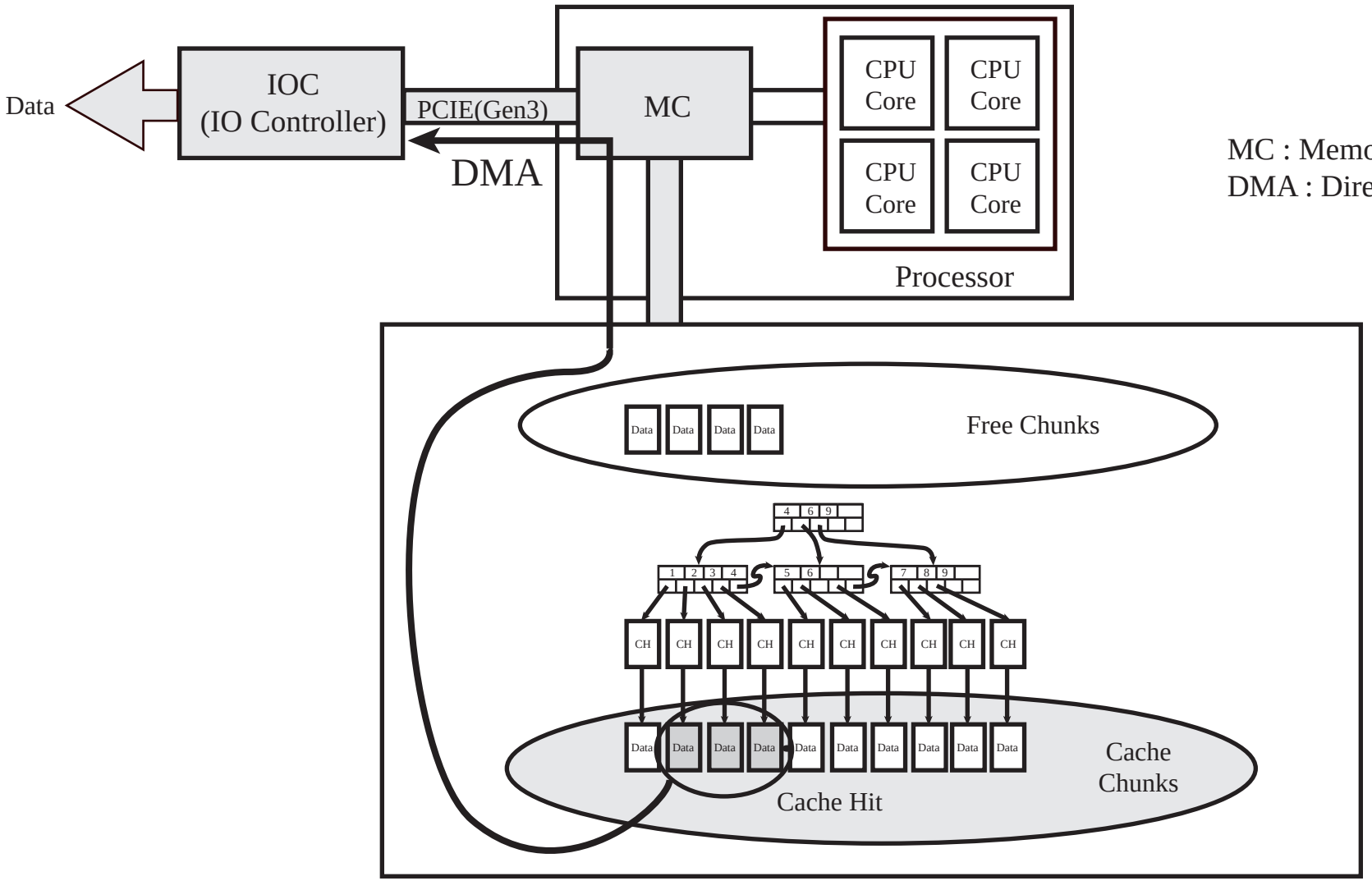
- Received data is written into free chunks at once.
- New data is replaced to cache chunks after DMA completion.
- Therefore, data clean is kept against intermediate fail of write DMA and abort command.



MC : Memory Controller
DMA : Direct Memory Access

Remarks (*1) At cache-miss case, new chunks are inserted as new cache entries.
If necessary, some old cache chunks are replaced by cache replacement algorithm (ex. LRU).

Zero-copy data transfer (Read)



MC : Memory Controller
DMA : Direct Memory Access

Effect of zero-copy data transfer way

○ Performance, up to limitation of hardware

- Throughput performance
  - > Up to memory bandwidth (>40GB/s) by DMA transfer to/from multiple IOCs.
- IOPS performance
  - > Per IOC performance x N (number) of IOCs.
  - > Up to CPU capability on MSI interrupt procedure by multiple cores.
  - > Load balancing “B+Tree Cache Entry Search” to multiple cores.

○ Effect of the latest CPU (Xeon Ivy Bridge v2) and the latest SAS/IOC

- More than 9 GB/s throughput per one IOC with PCIE (Gen3).
- Ultra wide memory bandwidth by multiple DDR3 (1,600MHz) channels.
- LSI 3004 (PCIE / Gen3 x 8), 12Gb x 8 / SAS one chip controller (IOC).
- 320 K IOPS@4KB per IOC x N (number) of IOCs=Total IOPS.

Statement of performance at cache hit case (1)

Performance dependent items	Factors	Details of factor and actual condition	Effects to product
Performance of hardware	IO Controller (IOC) processing capability	IOC has protocol engine and handles almost procedure of IOs. IO performance is limited by IOC's capability.	IO processing capability per one IOC is 250 K - 320K IOPS. For more IOPS performance, multiple IOCs improve total IOPS.
	Interrupt processing capability of CPU	Various interrupts happen and IO processing are processed in interrupt context and softIRQ context. Performance of IO processing never exceed CPU capability.	Interrupts from each IOC are dispatched to different CPU core (MSI-X and IRQ balance). Program codes for IO processing run at the same time at parallel on multiple cores. In current product, CPU interrupt capability is not bottle-neck. Maximum IOPS performance depends on IOC capability.
	Transfer bandwidth between PCIE and memory	Data is transfer by DMA between PCIE and memory. Throughput is limited by memory bandwidth between PCIE and memory.	Throughput of one LSI3004 (IOC, PCIE / Gen3) is about 9,400 MB/s@R&W. Increasing number of IOCs, improving throughput performance. We measured 34 GB/s with 4 IOCs.
Software overhead	Overhead of B+Tree search	Cache hit judgement and searching cache hit chunk list takes CPU overhead.	This procedure is processed in SoftIRQ raised by CPU after related interrupt context. This search process happened in each IOC, is processed in different CPU core at parallel.

Performance at cache hit case (2)

- Factor of performance at cache hit depends on IO processing capability of IOC.
- Performance of PCIE / Gen3 6G SAS IOC is 250 K IOPS and 5,500 MB/s (Throughput).
- Performance of PCIE / Gen3 12G SAS IOC is 320 K IOPS and 9,400 MB/s (Throughput).
- Multiple IOCs can improve system performance.
- Overhead of data access competition is existed and it disturbs linear improvement by multiple IOCs.
- Performance guideline according to number of multiple IOCs configuration.

Number of IOCs	IOPS@4KB, R	Throughput@512KB, R/W	IO-queue=1, 4KB-sync-write (For Journal write etc.)	Remarks
1 (6G/Gen3)	250 K IOPS	5,500 MB/s	8,000 IOPS	Measured
2 (6G/Gen3)	400 K IOPS	11,000 MB/s	↑	Measured
4 (6G/Gen3)	750 K IOPS	22,000 MB/s	↑	Measured
6 (6G/Gen2) *1	1,100 K IOPS	13,400 MB/s	↑	Measured
7 (6G/Gen2) *1	1,300 K IOPS	13,500 MB/s	↑	Measured
1 (12G/Gen3)	320 K IOPS	9,400 MB/s	27,000 IOPS	Measured
4 (12G/Gen3)	1,100 K IOPS	35,000 MB/s	↑	Measured
10 (12G/Gen3)	2,000 K IOPS	40,000 - 45,000 MB/s	↑	Presumed

*1 SuperSSD (G5) / Enterprise model is configured with PCIE / Gen2 IOCs.

Features of “Super Storage” (1)

Items	Features
Zero-copy based Large Scale Cache	- Full associative very large scale storage cache by B+Tree search. - Zero-copy data transfer, which decrease overhead and up to the limitation of hardware performance. - Effective use of memory resource for cache capacity by chunk header of cache indexing method. - Solution of non-aligned address issue by chunk header.
Object based Cache Management	- Assigning cache object to each source volume and multiple cache objects can run at parallel. - Setting different parameters to each cache object.
Stationed Cache	- Cache data never be purged before cache full condition which can improve cache hit ratio.
Background Dirty Data Flush	- Background flush with keeping high priority of front IOs. - Effective flush method of gathering dirty chunks having continuous address. - LBA (*1) sort flush for improving flush time.

*1) LBA : Logical Block Address
*2) Cache data management method is patented by CMS.

Features of “Super Storage” (2)

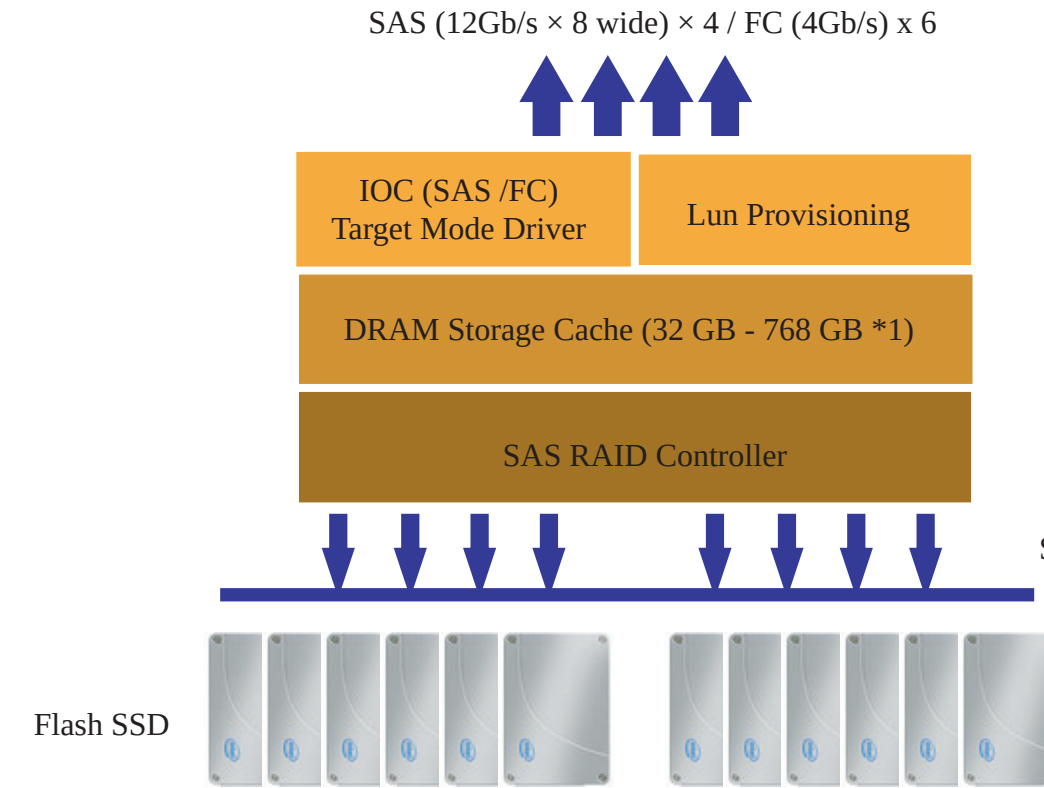
Items	Features
Cache Chunk Size	- Configuring cache chunk size according to application’s IO access pattern, in range of 4 KB, 8 KB, 16 KB, 32 KB and 64 KB. - Chunk threshold (0 - 99 %) which can select abandon (give-up) or move-in of smaller write data than chunk size.
Read Ahead	- Read Ahead means that BigCache read from source device with chunk size at read cache-miss case with small read size less than chunk size. - Configure Enable / Disable, to each cache object.
Read move-in disable Write move-in disable	- Read move-in Disable / Enable, which can be configured. - Write move-in Disable / Enable, which can be configured also.
Statistics	- Chunk counters showing number of free chunks, used chunks, dirty chunks, clean chunks and partial-data chunks. - Hit counters showing number and ration of cache hit, cache miss, give-up move-in chunks, partial data chunks. - IOPS / Throughput of host to cache and cache to source device. - IO access size distribution of host to cache and cache to source device. - IO access LBA distribution of host to cache and cache to source device.

Features of “Super Storage” (3)

Items	Features
Cache Replacement Algorithm	- LRU (Least Recently Used). - LRU weakness point, which it causes all cache miss case when IO access range is larger than cache size with sweep IO access. - LIRS (Low Inter Reference Recency Set), which can cover its weakness point of LRU, LIRS improve the cache hit ratio to (cache size / IO access range) against such sweep IO access exceeding cache size. - LRU or LIRS, which can be selected in each cache object.
Flush method (2)	- Dirty and clean chunks merged flush, which dirty chunks and clean chunks having continuous address, are flushed at the same time to improve efficiency. - Dirty purge threshold parameter, default is 10 %.
Lun provisioning to multiple hosts	- Lun provisioning, which maps Lun to host with invisible, read only or read and write.

Super SSD (G5)

○ Flash SSD/RAID with very large scale DRAM storage cache and

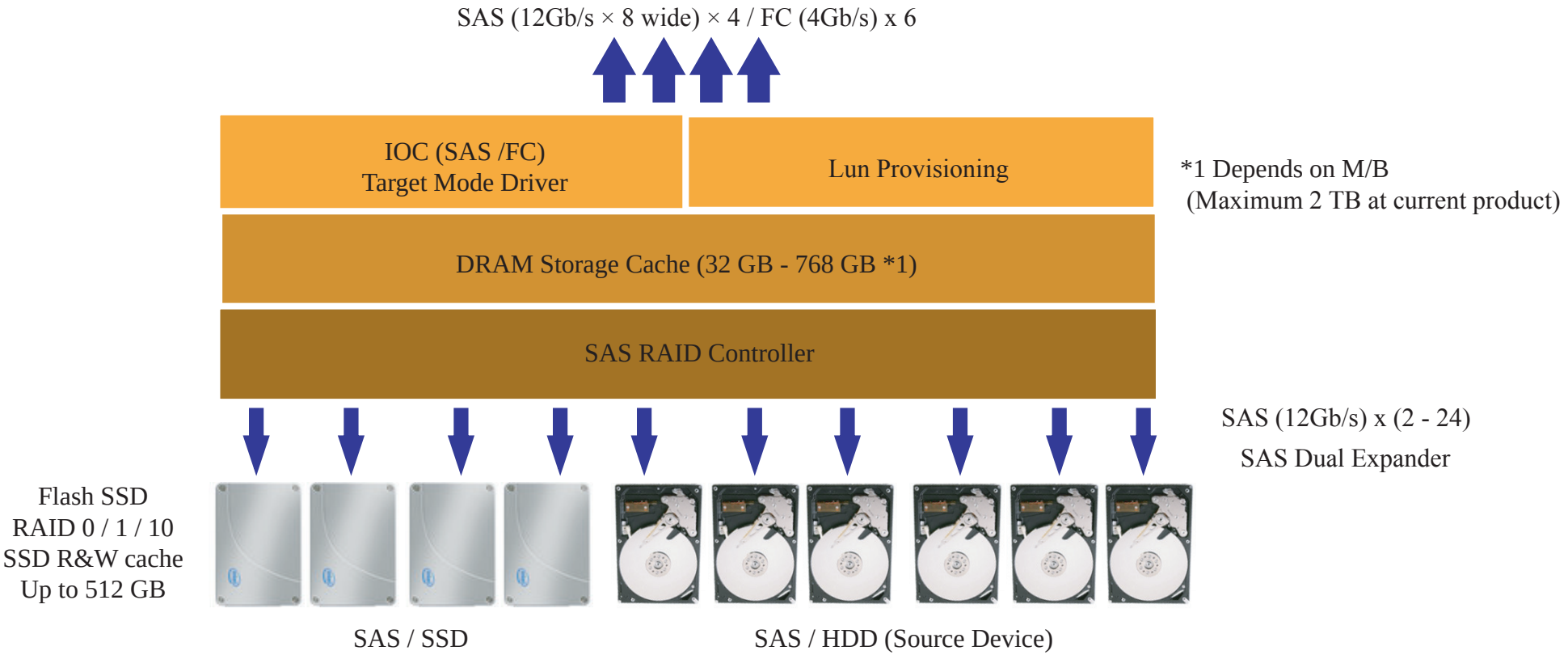


*1 Depends on M/B (Maximum 2 TB at current product)

- SSD RAID volume is 100 times, faster than HDD at read cache miss case.
- DRAM cache can cover the weakness point of Flash SSD at write IOs.
- DRAM cache hits improve more than 10 times faster than SSD volume at read and write.
- Much decreasing fault possibility by using Flash SSD without mechanical parts of HDD.

Super RAID TIER CACHE

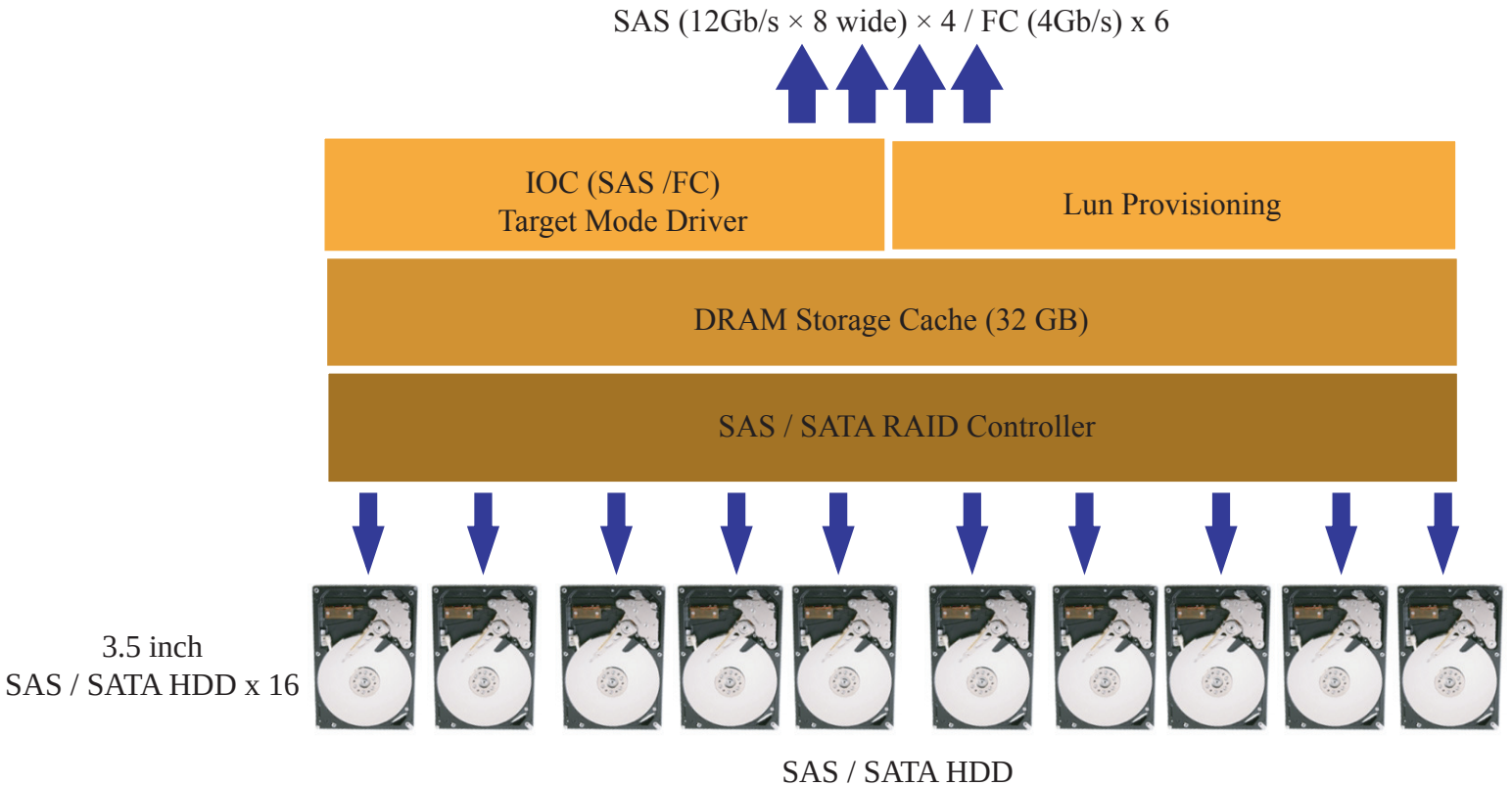
○ Large DRAM storage cache + Flash SSD cache + HDD/RAID volume



- Large DRAM storage cache and Flash SSD cache, two layered cache storage system.
- Combination of DRAM, Flash SSD and HDD, low cost solution for very high-speed and very large capacity which the source device is configured by SAS / HDD.

Super RAID V

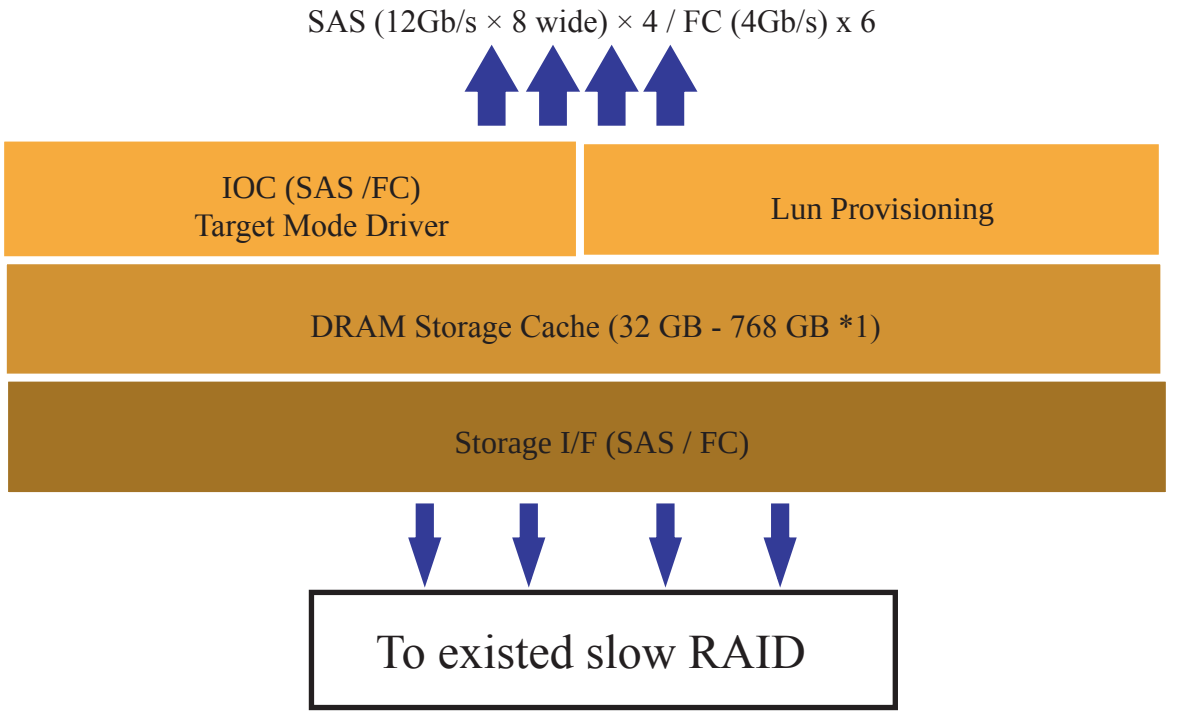
- Stable and high speed on write IOs



- Write buffer by DRAM storage cache.
- Smoothing throughput ripples on writing data to RAID 5 and RAID6.
- Capturing HD non-compressed video, video storage for non-linear video editor workstation.

Super CACHE

○ In-line cache storage cache deployed between host and the existed slow RAID



*1 Depends on M/B
(Maximum 2 TB at current product)

- Accelerating the existed slow RAID by in-line DRAM-based storage cache.
- Unnecessary adding driver and modifying software.
- Invisible from server operating system.

Comparison Flash SSD and HDD

[1] MTBF (Mean Time Before Failure)
- HDD is about 160 years, Flash SSD is about 220 years.
- When using 100 HDDs for one year, about 0.625 pcs of HDD is fault.
- When using 100 Flash SSDs for one year, about 0.455 pcs of SSD is fault.

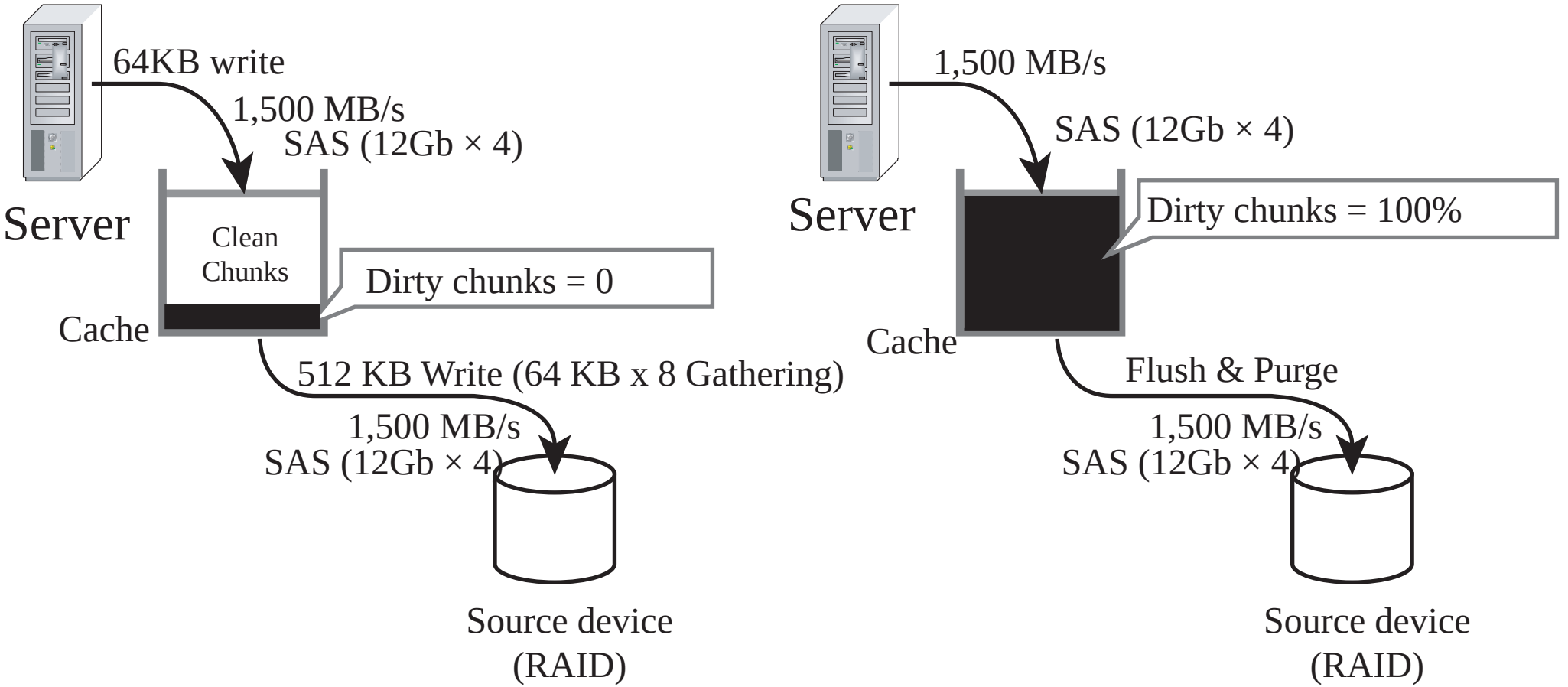
[2] Endurance and lifespan of Flash SSD
- Writing 400 GB data to Flash SSD (400 GB / SLC) every day,
  100,000 (Writable number of SLC) / (1 time / 1 day) = 100,000 days = 273 years
- Writing 250 MB/s (Maximum write performance) to Flash SSD (400 GB / SLC) continuously,
  400 GB×100,000 / 250 MB = 160,000,000 seconds = 1,852 days = about 5 years
- Writing 4 KB full-random IOs to Flash SSD (400 GB / SLC),
  About 16 years
  (In opened specification, 8,200 TB @ 100 GB (TBW) / (64MB/s) = 4.06 years)
- SSD Guard(TM) by LSI : Data migration from SSD with smart flag to hot spare SSD.

[3] Bit cost
  1,650 YEN@1GB (SSD) v.s 273 YEN@1 GB (HDD)

[4] Cost for 90,000 IOPS@Random-read
  90,000 IOPS(SSD) / 500 IOPS (HDD) = 180 drives (HDD)
  660,000 YEN (SSD) v.s 7,200,000 YEN (HDD)

*Remarks : Price is example, it is not actual value.

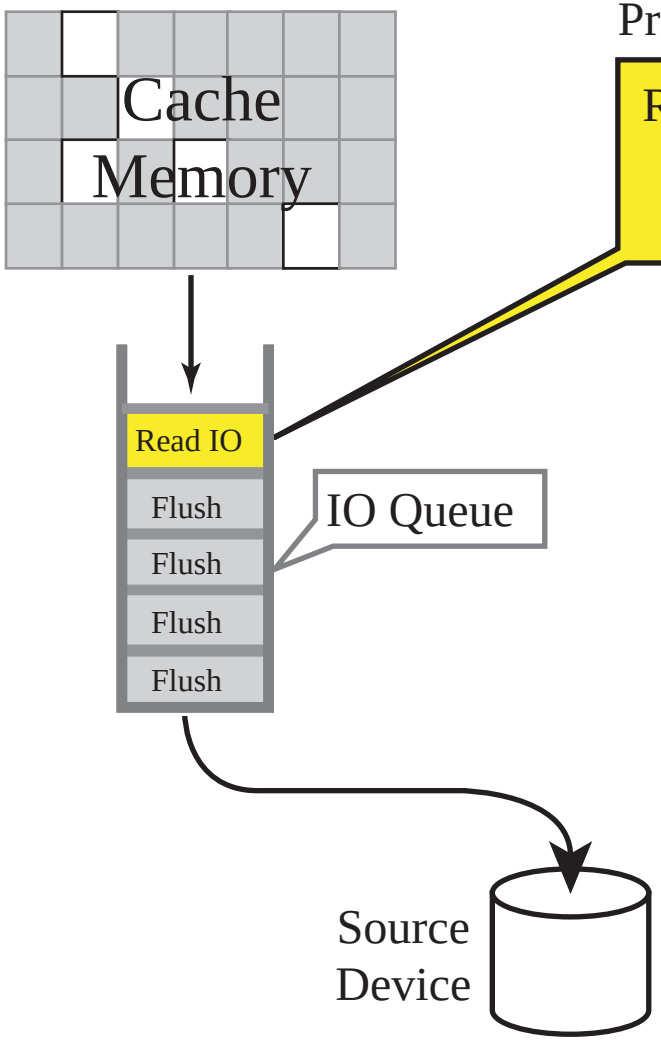
Behavior of flushing dirty chunks



Flush with gathering continuous address chunks.

No degrade at cache full condition.

Strategy on background flush



Problem

Read cache miss happen at running background flush
> IOs queues for dirty chunk flush, disturbs front IO.
> Large response time of read IO.

Strategy 1 : Delayed flush configuration

Stopping background flush at happening of cache miss read IO.

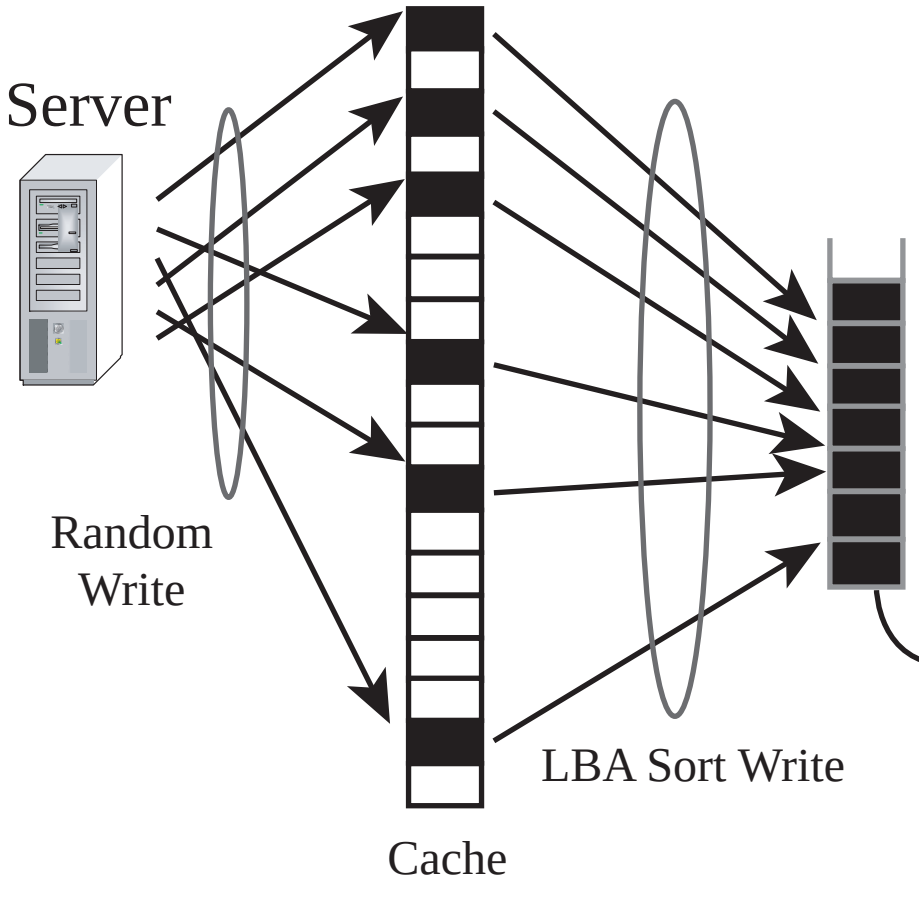
Background Flush policy :	DELAYED IO
Flush Delay Time :	1000 m seconds
Delay timer reset threshold :	2 IOs

Strategy 2 : Scheduled background flush

Backgroud flush at lower operating duration (Ex. At mid-night).

Flush start time :	02:00
Flush stop time :	05:00

LBA order flush



○ Flush performance (Example) *1 (IOPS)

	LBA order	LRU order
512 B	5,500 - 6,500	1,300 - 1,600
8 KB	2,200 - 2,700	1,300 - 1,500
64 KB	1,800 - 2,100	1,200 - 1,400

*1) Depends on the source device performance.

Improvement, 1.5 times to 4 times

Outline of SSD read cache (CacheCADE)

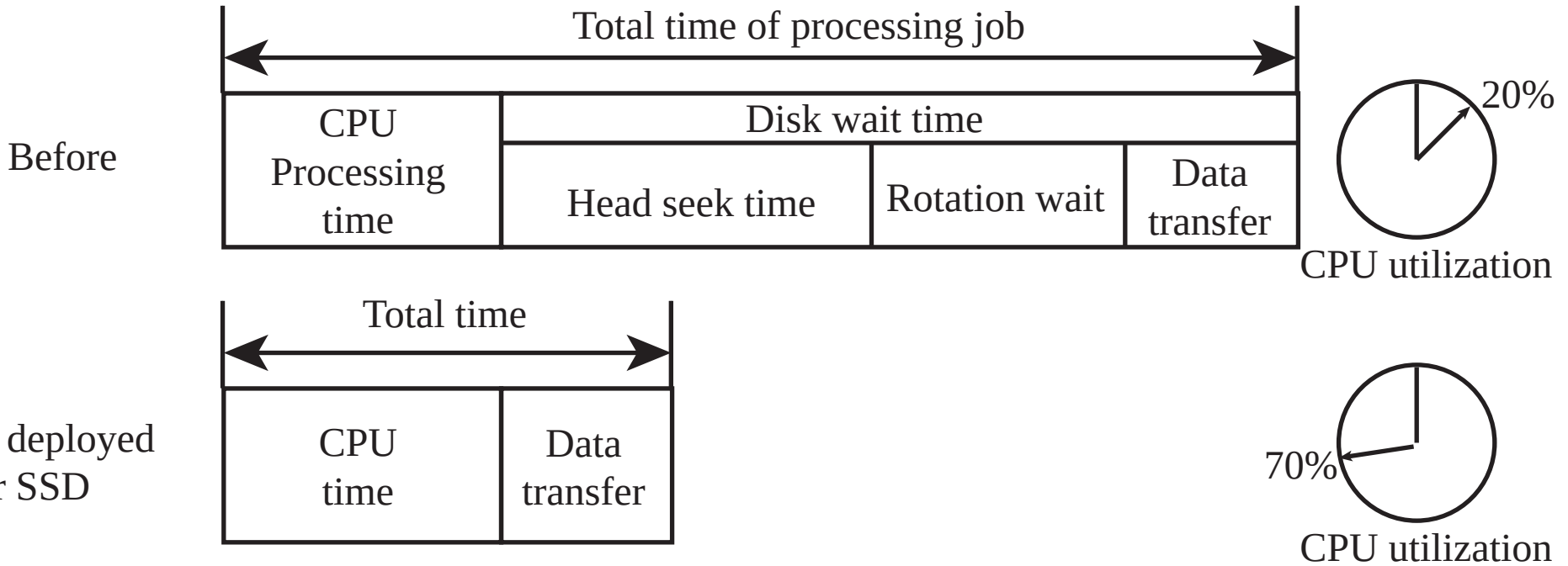
- Move-in read IO less than 60 KB to flash SSD.
- Configuring of multiple CacheCADE VD (Virtual Disk).
- Maximum 512 GB Flash SSD cache.
- Creating CacheCADE VD at accessing to the source device.
- Deleting CacheCADE VD at accessing to the source device.
- Load balancing to multiple flash SSDs.
- Reseting cache data at rebooting.

Outline of SSD R&W cache (CacheCADE Pro2.0)

- Not only read cache, Flash SSD read and write cache.
- RAID 0 / 1 / 10 configuration for CacheCADE VD.
- Maximum 512 GB Flash SSD cache.

*Remarks : CacheCADE is Trade Mark of LSI, which LSI developed software function running on MegaSAS RAID.

Effect example



- Reducing time of 10 hours heavy batch jobs to 3 hours completion, for example.
- Night batch job can be run at day time .
- Several 10 seconds response of online operation, which can be reduced to several seconds response.

Effect example (2)

○ Ecology

- Many servers and many HDDs to improve performance.
- After deploying Super SSD, reducing number of Servers and RAID systems and decreasing power consumption.

○ Cost viewpoint

- Very expensive cost for performance tuning ?
- Save cost for tuning by deploying Super SSD.

Super SSD (G5) / Enterprise

Itmes	Type	Specification
Interface	Host port	6Gb x 8 / wide SAS x 1 (Up to 6)
DRAM Cache capacity	-	32GB - 768GB
Flash SSD capacity	SAS / SLC /2.5 inch x n	800GB (400 GB × 2) - 9,600 GB (400 GB x 24)
IOPS	At cache hit case	1,500 K IOPS
Throughput	At cache hit case	15 GB/s
Volume share	LUN MASK/Provisioning	WWPN
SSD protection	RAID / SSD Guard (*1)	RAID 0, 1, 5, 6, 10, 50, 60 / Hot-spair
Configuration tool	Serial console / SSH	CUI
Monitor	-	LED, BMC, loging
Enclosure	-	19 inch EIA 2U rack mount
Management	-	SNMP / SNMP trap
Form factor	EIA 19’ 2U	-
Power (Consumption)	1,100W (350W)	Hot swappable and redundant

Remarks *1) LSI SSD Gurad (TM)
*2) Mentioned specifications will be changed.

Super SSD(G5) / Profesional

Itmes	Type	Specification
Interface	Host port	12Gb x 8 / wide SAS x 1 (Up to 4), FC(4G) option
DRAM Cache capacity	-	32GB - 768GB
Flash SSD capacity	SAS / MLC /2.5 inch x n	400GB (200 GB × 2) - 19,200 GB (800 GB x 24)
IOPS	At cache hit case	1,500 K IOPS
Throughput	At cache hit case	22 GB/s
Volume share	LUN MASK/Provisioning	WWPN
SSD protection	RAID / SSD Guard (*1)	RAID 0, 1, 5, 6, 10, 50, 60 / Hot-spair
Configuration tool	Serial console / SSH	CUI
Monitor	-	LED, BMC, loging
Enclosure	-	19 inch EIA 2U rack mount
Management	-	SNMP / SNMP trap
Form factor	EIA 19’ 2U	-
Power (Consumption)	1,200W (350W)	Hot swappable and redundant

Remarks *1) LSI SSD Gurad (TM)
*2) Mentioned specifications will be changed.

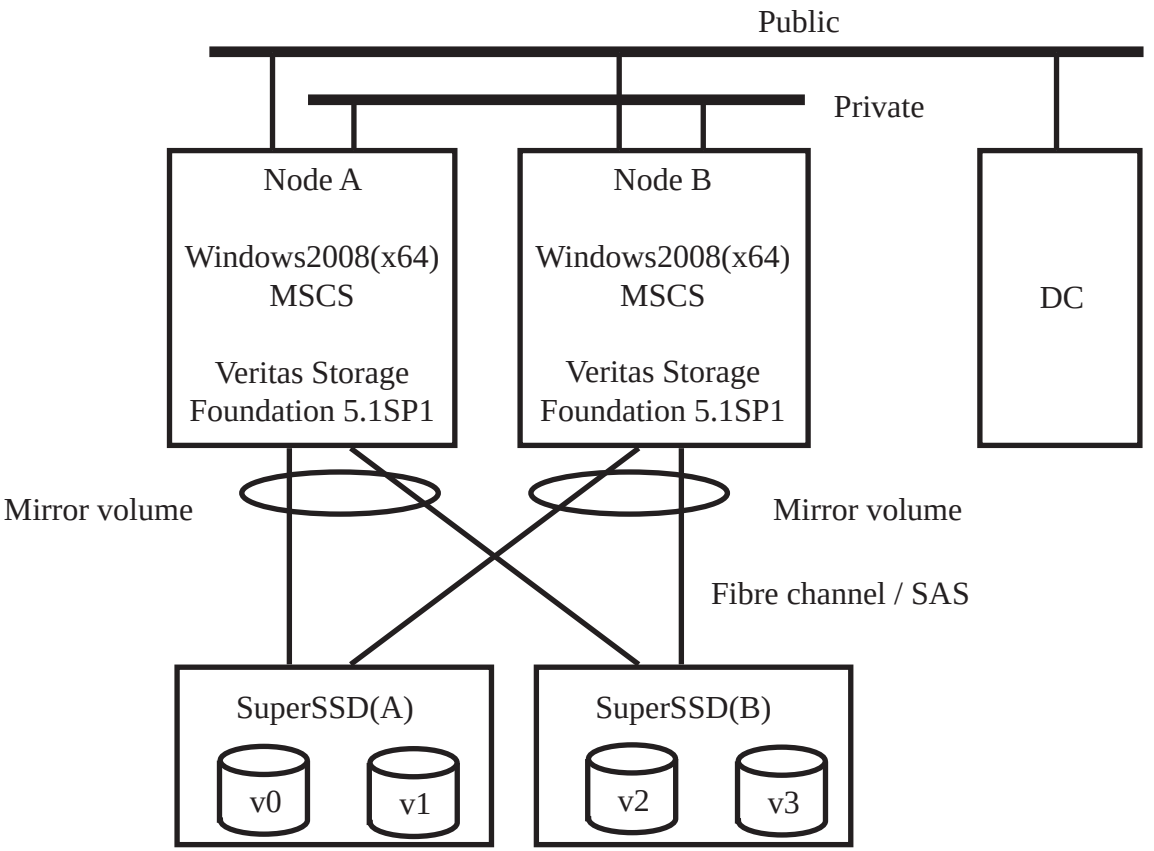
Super RAID V

Itmes	Type	Specification
Interface	Host port	12Gb x 8 / wide SAS x 1 (Up to 4), FC(4G) option
DRAM Cache capacity	-	32 GB
HDD capacity	SATA / 3.5 inch x 16	16 TB / 32 TB
IOPS	At cache hit case	1,500 K IOPS
Throughput	At cache hit case	22 GB/s
Volume share	LUN MASK/Provisioning	WWPN
HDD protection	RAID	RAID 0, 1, 5, 6, 10, 50, 60 / Hot-spair
Configuration tool	Serial console / SSH	CUI
Monitor	-	LED, BMC, loging
Enclosure	-	19 inch EIA 3U rack mount
Management	-	SNMP / SNMP trap
Form factor	EIA 19’ 3U	-
Power (Consumption)	1,200W (500W)	Hot swappable and redundant

Remarks *1) Mentioned specifications will be changed.

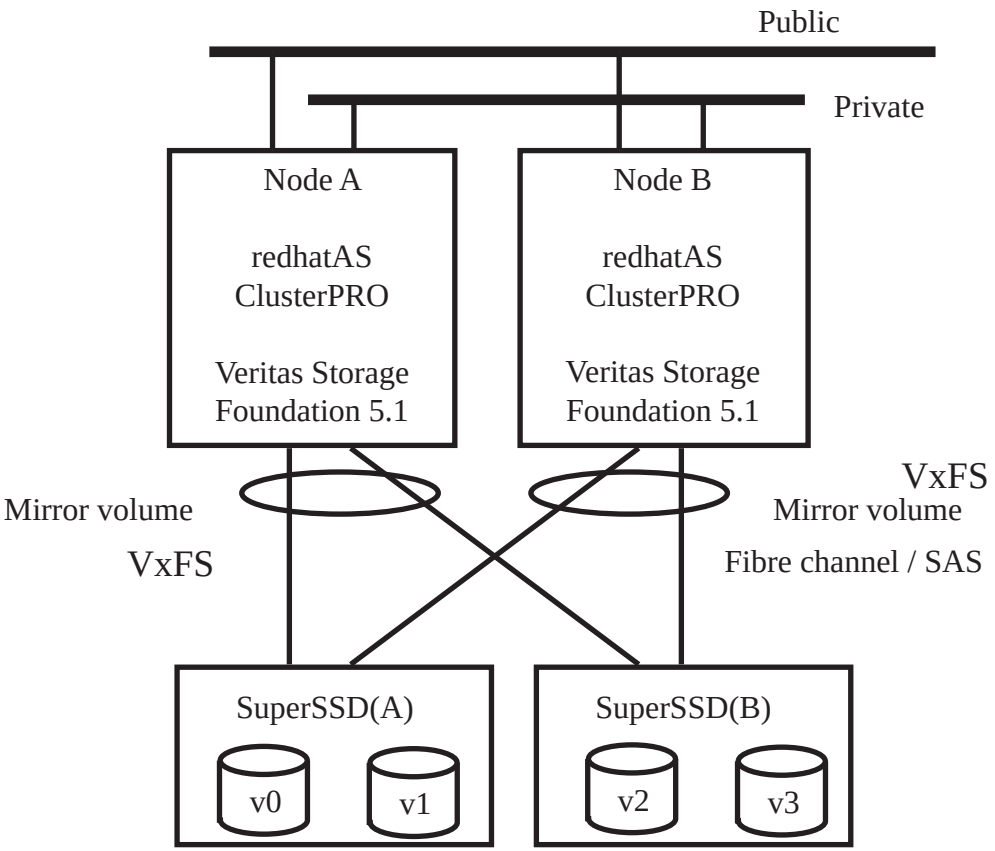
Super SSD solution (1)

- Windows2008/MSCS + Veritas Storage Foundation for Windows, high available system -



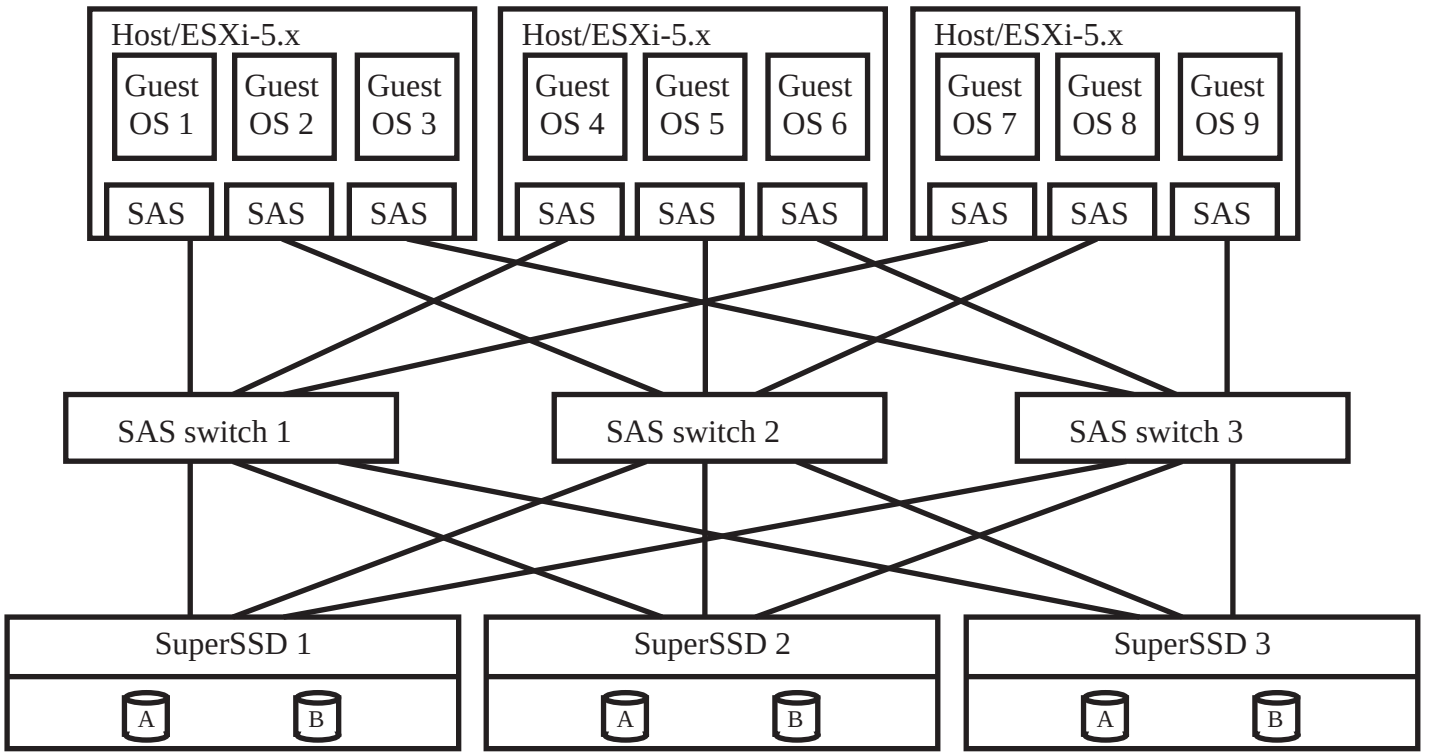
Super SSD solution (2)

- ClusterPRO + Veritas Storage Foundation, high available system -



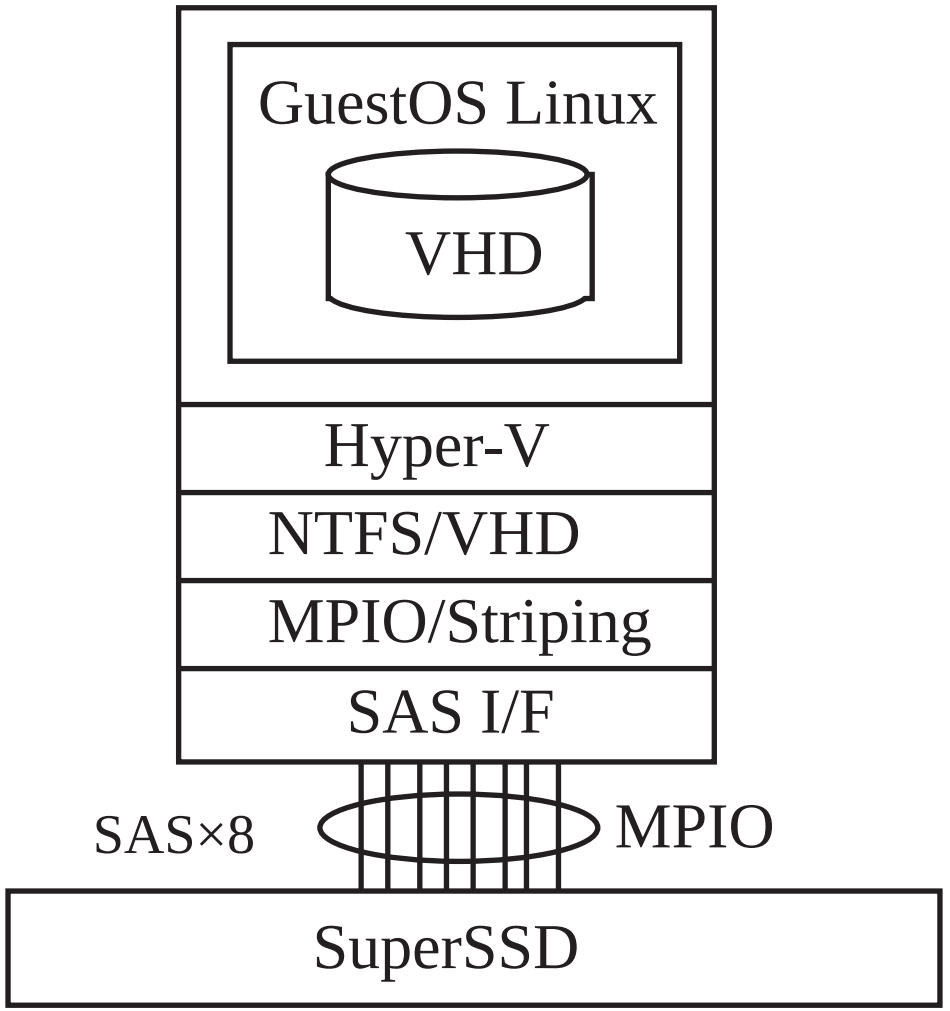
Super SSD solution (3)

- VMware, Cloud system -



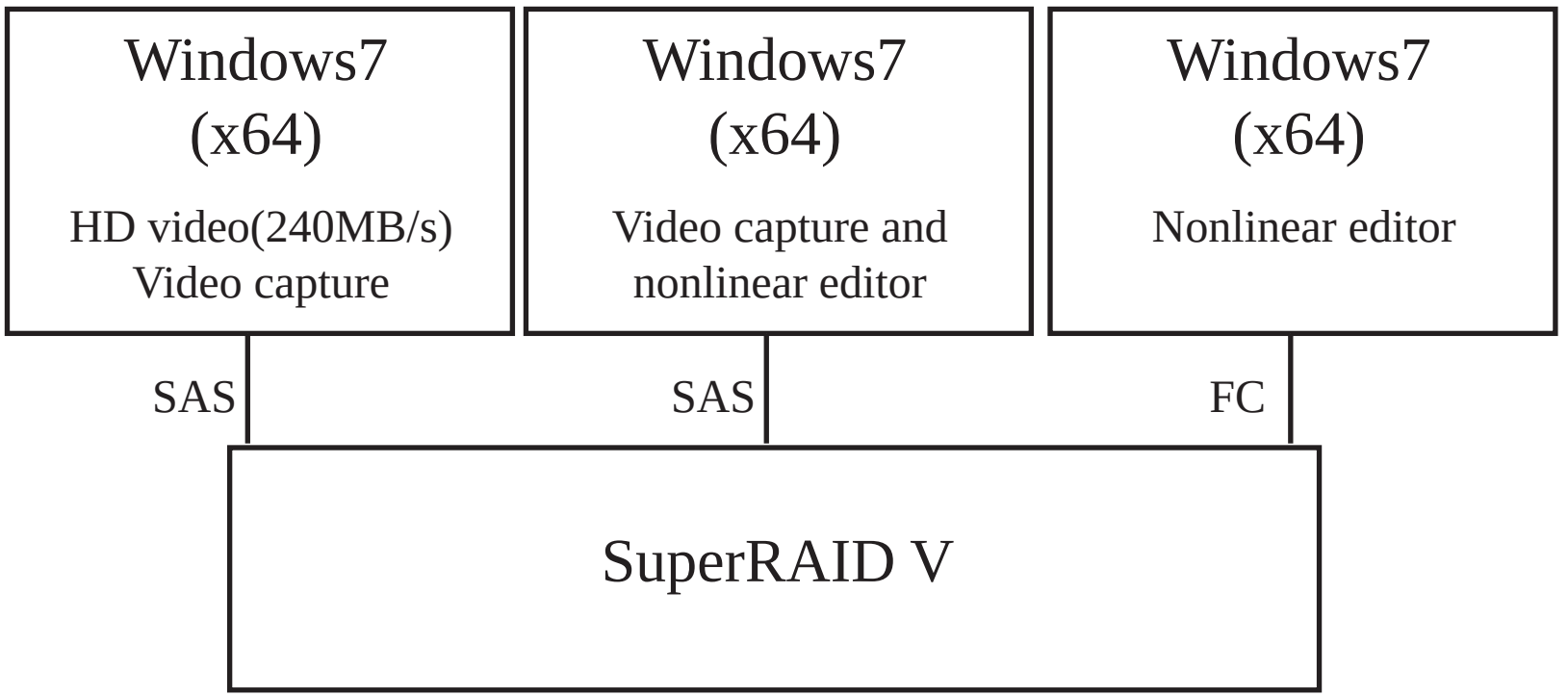
Super SSD solution (4)

- Windows2008R2 / Windows2012R2 Hyper-V -



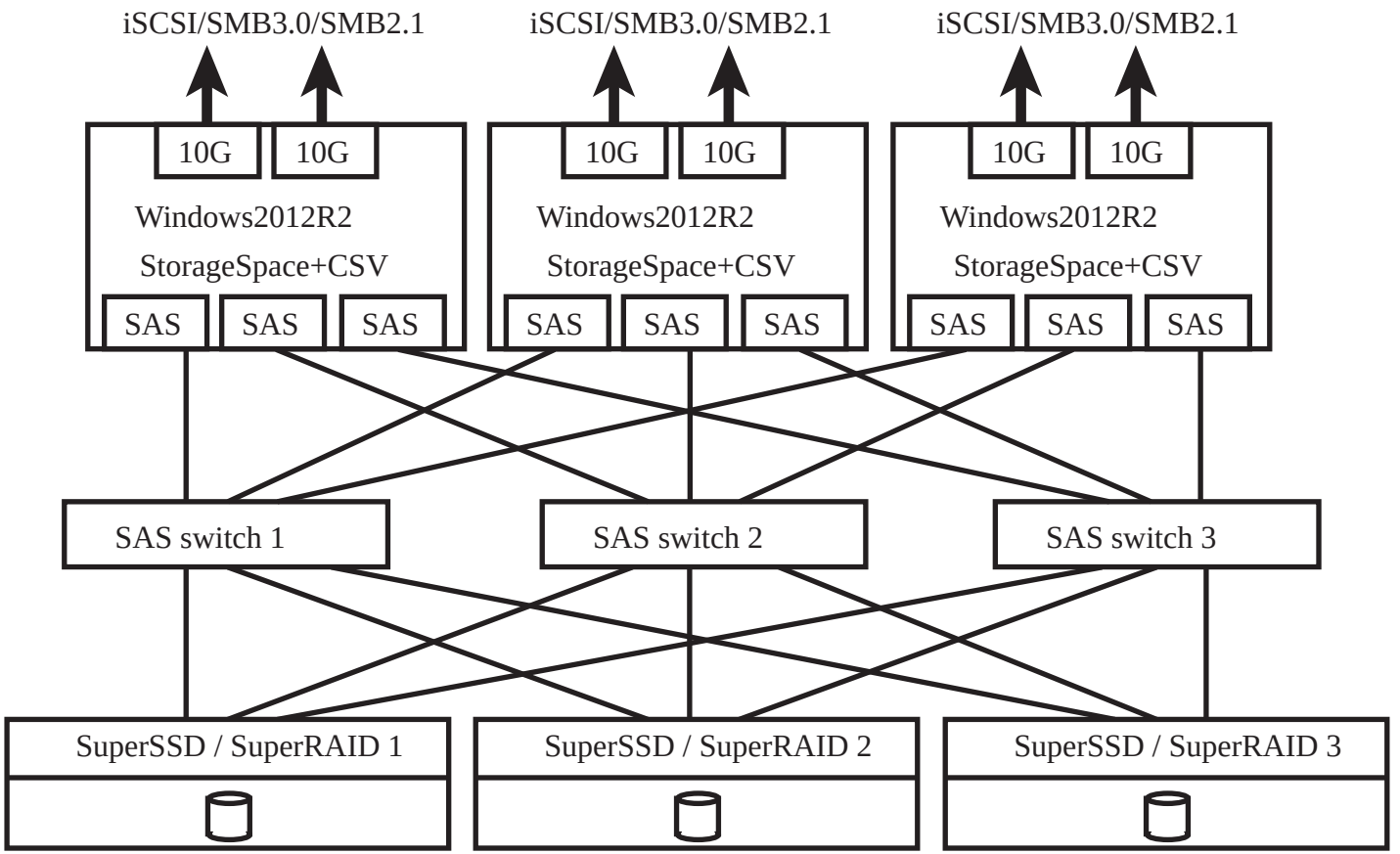
Super RAID V solution

- High speed shared file system (Super RAID V + MetaSAN)



Super Storage solution

- Windows2012R2 scale out file server -



Core Micro Systems, Inc.

URL : http://www.cmsinc.co.jp/ Mail : sales@cmsinc.co.jp

Wacore Kaname-cho Bldg. 9F
11-2, Nakamaru-cho, Itabashi-Ku, Tokyo 173-0026
TEL : +81-5558-5410 (IP Phone)
TEL : +81-3-5917-6451
FAX : +81-3-5917-6452