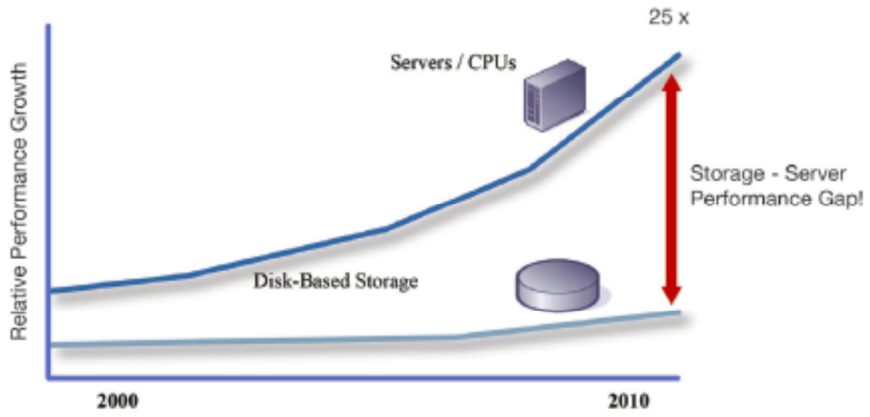
Increasing Performance Gap

Performance of CPU grows annually about 50%. Whereas, the storage performance grows only about 10%. The performance gap between CPU and Storage keep widening year-on-year. Applications such as transaction-intensive databases, business and financial analytics, data mining, data ware housing, production archive, VDI requires high storage performance to meet the business demands.

Figure 1. Increasing Performance Gap Between Servers and Storage



IT managers typically increase the number of spindles (disks) to increase the storage performance and try to deliver more IOPS to the application. However, this solution comes with a couple of problems. More disks means more money spent on the solution. It also creates lot of disk space left unutilized. More disks occupies expensive datacenter footprint and increases energy and cooling costs.

Once the disk usage increases, performance starts to drop. Again more spindles has to added to retain the required performance level.

Unfortunately, many performance-starved applications are often also capacity intensive.

To achieve a truly efficient data center, performance-starved applications must be addressed with a storage solution that meets both performance and capacity requirements.

Solid State Disk (SSD)

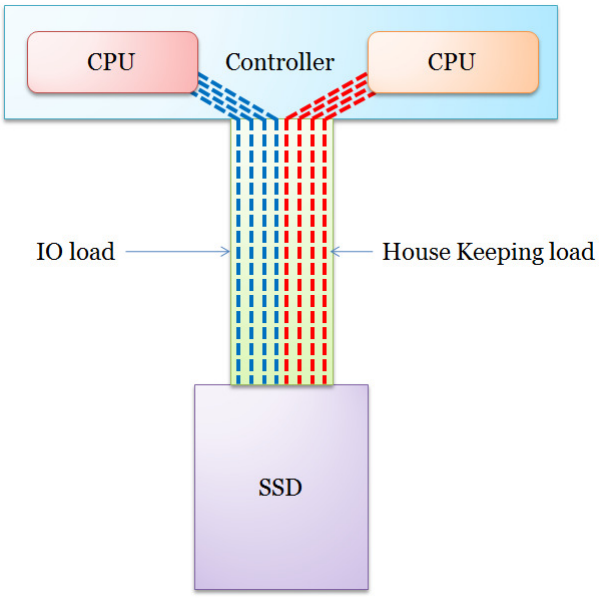
Solid State Disks (SSD) came to the rescue. It provided multi-times the performance of spindles based disks. Typically, a SSD delivers anywhere between 3000 to 4500 IOPS depending upon various workload conditions.

However the IOPS per SSD is not sustained throughout its life. Unlike spindle based disks, SSDs brings its own backend challenges. The following housekeeping activities happens in the flash based storage drives.

- The cells within a flash based storage has a limited lifespan. No cell should be repeatedly charged (write). In other words the cells within a flash drive should be charged (write) in a balanced manner.
- ECC data written on the cells has to be frequently read and recalculated to retain the integrity of the data.
- Similarly data on the cells has to rewritten once in a month to retain the data integrity.
- Logical to physical address conversion has to be done, to enable old data to be erased asynchronously.

Write Cliff

Storage controller and the backend disk IO path is utilized by both regular host IO and housekeeping activities. Housekeeping activities requires significant storage controller CPU cycles. Under write intensive workloads, housekeeping activities likely to hold the host IO. Such symptom is called “write-cliff”. During write-cliff, application performance suffers badly.



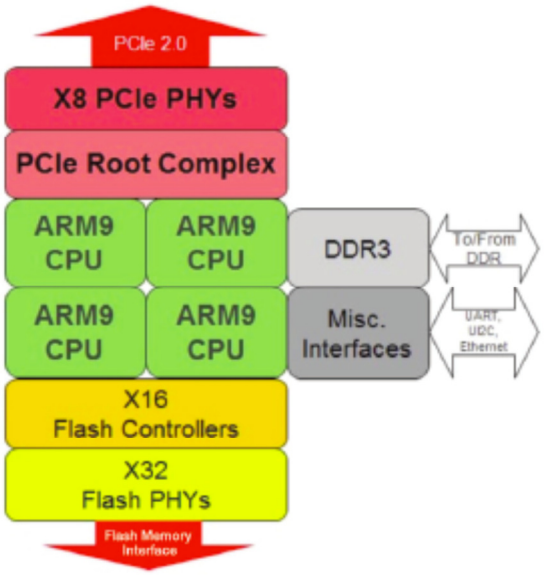
Hitachi Accelerated Flash (HAF)

HAF is not an SSD. It is unique flash technology, developed by Hitachi. HAF controllers are purpose-built to provide consistently high performance.

The frequency of writes determines the lifespan of flash. Hence, the number of times a flash memory cell is subject to a write the durability and hence reliability of the flash drive is impacted. Such frequent writing can't be changed. However, innovative controller designs can manage and optimize for this reality and extend the lifespan of flash drives.

Flash Controller ASIC

HAF uses the Hitachi custom-developed flash controller ASIC. This ASIC is a quad-core, 1GHz, 32-bit processor complex.



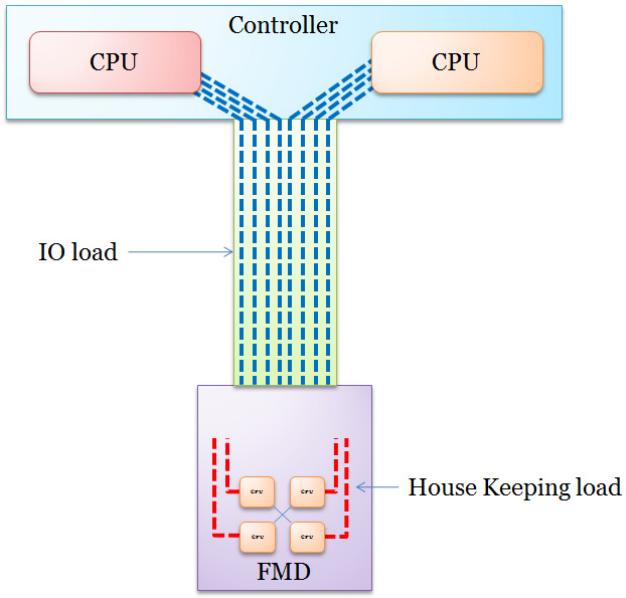
An area of contention within most flash controllers is the number of paths that access the flash storage. Most controllers offer 8, 12 or perhaps 16 paths. Hitachi chose to incorporate 32 parallel paths, combined with the power and flexibility of the multicore processor. This approach enables the parallel processing of multiple tasks. It also allows the removal of housekeeping tasks from the I/O path (wear leveling, ECC, and so forth), eliminating the potential of host IO blocking.

Key benefits of Hitachi Flash

Write Cliff Elimination

ASIC within the Hitachi Flash handles the housekeeping activities and storage controllers are freed-up. It improves controller performance and eliminates “write cliff”. This approach is possible thanks to use of a quad-core processor that has more than sufficient processing power.

Following picture shows housekeeping activities handled locally by Hitachi Flash ASICs.



Wear Leveling

Flash has a limited lifespan. Each write or erase cycle stresses the cell causing a deterioration of the cell longevity over time. If the cells are not written or erased in a balanced fashion, the flash drive can have an irregular distribution of unstable cells.

Hitachi Flash has patented intelligent algorithm to minimize such imbalance and write/erase cells in a uniform pattern. In turn it optimizes the life space of flash memory cells. The Hitachi Flash ASIC monitors the rate of writes, erasures and refreshes and balances the rate of deterioration in the flash memory cells.

Data Integrity

To preserve the integrity of data writes, 42 bits of ECC are added for every 1KB of data written. This action translates to each FMD ECC being able to correct up to 42 bits per 1.4KB, which exceeds the standard SSD spec of 24 bits per 1KB of data. It ensures that even if a bit error is discovered it can be easily recovered.

Adaptive Data Refresh

To catch bit errors quickly, the FMD will perform a high frequency "adaptive" data refresh. During “adaptive” refresh the controller reads and recalculates the internal ECC on the complete FMD every 2 days and dynamically optimizes page refresh based on applied error correction.

Note: To secure the integrity of previously written data, all data is rewritten at least every 30 days. This practice is key to not only extending flash cell longevity but also improving the overall sustained performance.

Periodic Data Diagnosis or Recovery

If the bit errors exceed the ECC correction capability of the Hitachi Flash, then the data is read out by the read retry function. This function adjusts the parameters of the flash memory and reads the data. The area is then refreshed, meaning that the data is read and copied to a different area before the data becomes unreadable.

High-Speed Formatting

The formatting is done autonomously in the Hitachi Flash. This highly efficient process is completed in approximately 60 minutes, regardless of the number of drives to format. When compared to the 280 minutes that a similarly configured array of SSDs (22.4TB) would take, it is apparent that systems employing the Hitachi Flash have a reduced install time.

Block Write Avoidance

Hitachi flash ASIC can manage 128 flash memory chips and it also supports inline compression. Any data stream of "0" or "1" is compressed in real time with an algorithm that remaps the data with a pointer. This technique can deliver up to a 94% savings in storage space. By enabling these space savings and preserving the reserved capacity for background tasks such as garbage collection and wear leveling, substantial improvement in the sustained write performance is realized.

This feature not only drastically reduces format times, but also, by eliminating unnecessary write/erase cycles it effectively extends the effective life of the flash memory.

Overprovisioning

Overprovisioning is the practice of including flash memory above the advertised capacity. This overprovisioning increases the write endurance of the specified flash memory capacity and its overall performance. Hitachi Flash enjoys 25% overprovisioning.

SUMMARY OF HITACHI FLASH BENEFITS

- ECC has been extended to correct 48 bits per 1.4kB. This correction enhances the ability to monitor the degradation of pages and avoids any premature page rewrites.
- The controller reads and recalculates the internal ECC on the complete Hitachi Flash every 2 days and dynamically optimizes page refresh based on applied error correction.
- Logical-physical address conversion, to enable old data to be erased asynchronously, minimizes housekeeping tasks.
- Buffered write area, to reduce formatting for small writes, efficiently manages formatted page availability.
- Data is refreshed at least every 30 days to avoid retention time degradation.
- Stream of “0”and “1” compaction reduces unnecessary writes by up to 94%.
- Wear leveling is done locally across the pages in a flash module and globally across all modules in a pool of flash modules. This approach distributes wear and extends the life of a flash module.
- 25% of the flash capacity is overprovisioned.