novachips

Unleashing the Next Generation Flash Storage Solution For Data Centers

WHITEPAPER

Introduction

Because most people and machines are now connected through various smart devices, or the Internet-of-Things (IoT), an enormous amount of data is generated every day. Data has become critical for and analysis and processing. With more than a trillion sensors generating reams of data, storage demand will surge even more to store data between sensors (e.g., alarms and wearable devices, like health monitors) and big data applications.

In order to keep up with this volume, solid state drives (SSDs) have gained a lot of traction in the market. They come in a wide range of capacities, employing a variety of system interfaces (i.e., connection to the host) and form factors (physical dimensions) to meet data center applications. The most common SSDs, in the data center storage segment, use the traditional HDD 2.5" and 3.5" form factors with SATA and SAS interconnects and with up to 500 GB ~ 1TB capacities. These flash-based SSDs provide faster I/O performance than HDD and support larger storage capacities than earlier SSD technology. They consume less power and retain data when power is cut off abruptly, as in a power failure.

HDDs are not the best solution for applications that require low latency for read and write operations. This is where SSD works best. HDDs have performance and physical limitations that prevent them from keeping pace with the fastest data center applications and with increased server workloads. While basic servers can handle hundreds of thousands of IOPS, a traditional HDD can only deliver between 100 and 300 IOPS. Thus, there is a clear disconnect between storage and server throughput.

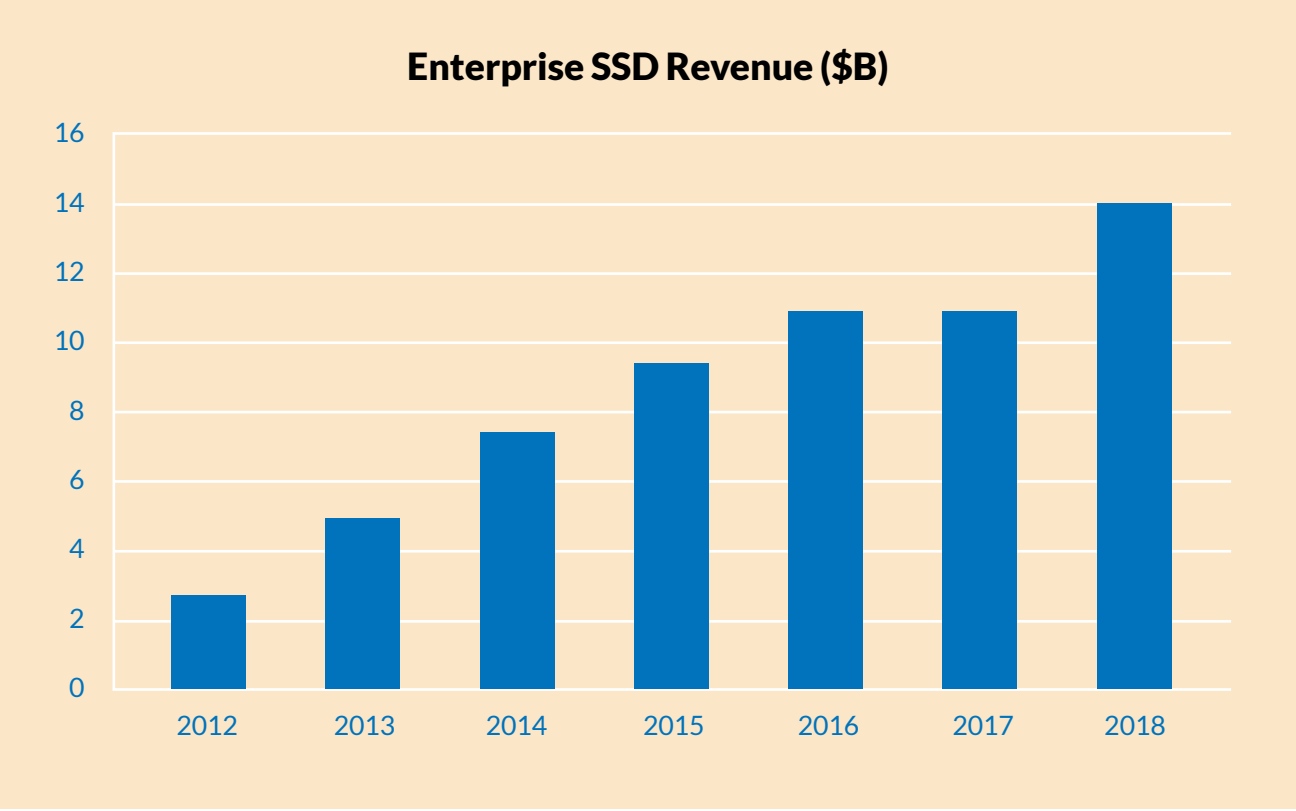
Every time that data is requested from a different location in HDD storage, the mechanical drive controller needs to physically move across the disk. This limits the speed by which a drive can read data. In contrast, SSDs have no moving parts. Because of their speed advantage, and falling prices, enterprise and cloud infrastructures are rapidly moving to SSDs for large-scale applications like virtual desktops, online transaction processing (OLTP), and web caches. PCI Express (PCIe) offers the lowest-latency, highest throughput host interface to take advantage of a direct connection to NAND flash. However, current generation SAS/SATA and FPGA-based PCIe SSDs have had architectural limitations that may result in performance limitations, shorter life space, and higher costs. In this paper, we look at what has been done to overcome those issues.

This paper is targeted at SSD vendors, server and storage OEMs, and cloud computing architects who are looking for next-generation SSD technologies that will provide improved performance, reliability, and scalability. The Novachips NVS5700 series offers a next-generation flash storage processor solution. This is the world's first 28nm low power native PCIe controller. With industry-leading native PCIe SSD performance, the NVS5700 series solves the performance limitations of existing SATA SSDs. Furthermore, the NVS5700 series offers a highly scalable architecture to let customers to build large capacity SSDs at lower cost, with extraordinary performance to meet the needs of their target customers and market segments.

The Growth of Enterprise SSD Market

New enterprise products ranging from drives to caches to arrays have led to greater integration of SSDs into corporate storage systems. Market analysis says the current surge in growth is going to increase exponentially. Gartner says the SSD market will grow to $26.7 billion USD by the year 2018. Most growing portion will be with the enterprise SSD with PCIe interface. Gartner estimates total SSD shipments will to grow 50% this year and reach 236 million units by 2018. This is close to half the size of the current HDD market of 397 million units. As shown in Figure 1, the enterprise SSD market is expected to reach $13.9 billion in revenues by 2018, nearly six times that of 2012. And also, with over 100 million enterprise PCIe ports on the market, combined with rapidly growing offerings in term of port shapes and size, revenue forecasts for PCIe technology within the enterprise SSD market are increasingly optimistic.

Figure 1: The Growth of Enterprise SSD Market



(Source: Gartner October, 2014)

The Emergence of PCIe SSDs in Data Centers

Much of current storage infrastructure uses bridged PCIe SSD interfaces like serial-attached SCSI (SAS), SCSI-over-PCIe (SOP), PCIe based SATA, and FPGA-based PCIe SSDs. Global Industry Analysts, Inc. says, "These Bridged PCIe SSD solutions currently have inherent architectural limitations that make these solutions less cost effective in the longer run. In addition, these solutions have the ability to handle only medium density IOPS. Higher power consumption is another major drawback associated with these solutions." To overcome these obstacles, the NVM Express (NVMe) specification was developed to allow a native PCIe SSD to connect directly to the PCIe Root Complex with a simple, efficient command set written specifically for the current and next generation of non-volatile memories. Since the PCIe SSDs have a more direct path to the CPU, there is a significant reduction in latency versus devices connected through a host bus adapter (HBA). This provides an order of magnitude boost in performance to future-proof the investment in PCIe-based SSDs. NVMe provides:

- Ultra-low latency
- Very high throughput
- Low power consumption, resulting in a lower Total Cost of Ownership (TCO) and reduced carbon footprint
- Cost reduction and shorter time to market, through the use of a standardized interface
- Reliable performance across multiple cores, enabling quick access to critical data
- An optimized register interface and command set that reduces CPU utilization resulting in higher performance and lower power usage
- Scalable for current and future NVM performance requirements
- End-to-end data protection capabilities and support for standard security protocols, such as Trusted Computing Group
- Seamless integration into multiple operating systems with standard open driver interfaces

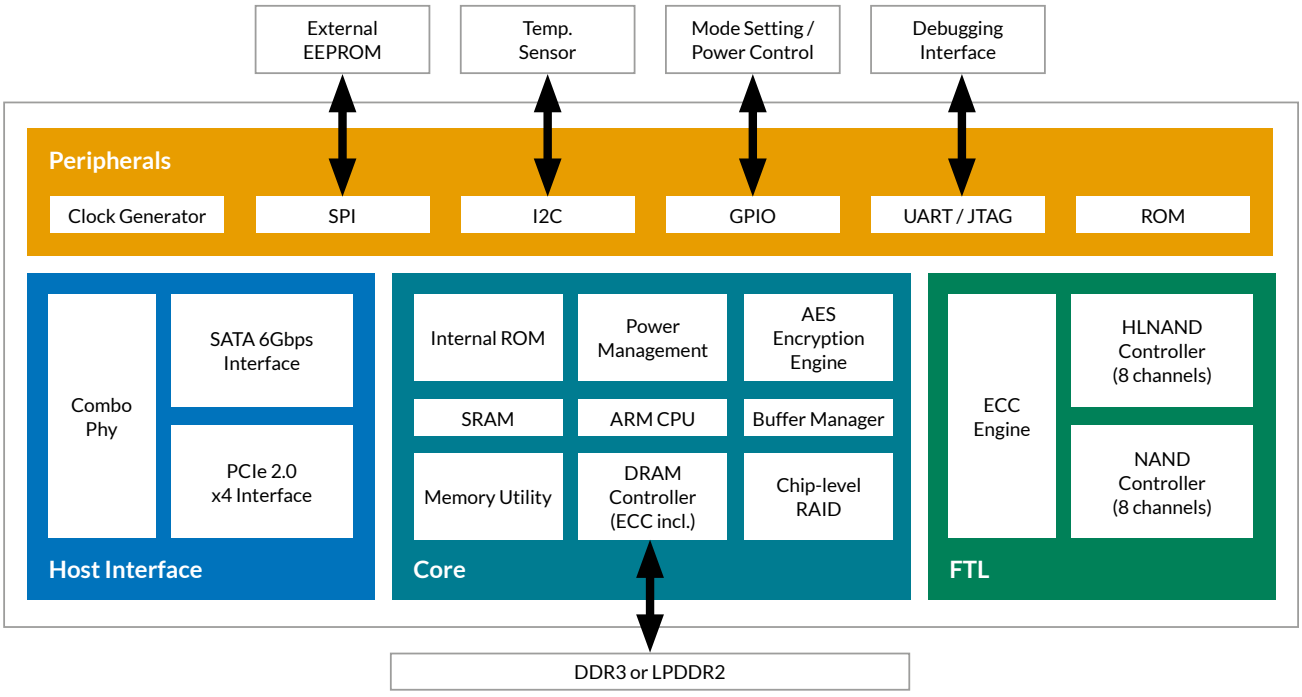
Figure 2: Novachips PCIe HLSSD (Express Series)

Novachips PCIe/SATA Combo Flash Storage Processor

The Novachips NVS5700 series controller can support higher storage densities than traditional SSD configurations. With the NVS5700 series controller, the Novachips SSD supports up to 16TB compared to the current 1TB or 2TB limit in enterprise SSD. This increase capacity 8x over current competitors' SSDs.

Novachips Express-series HyperLink SSD (HLSSD) has a unique solid state drive designed to deliver massive storage at a lower cost and faster performance than other designs. Express does this by combining unique HyperLink NAND flash memory technology with comprehensive endurance management firmware and power loss data protection features. These improvements extend the lifetime and reliability of the SSD. The Novachips SSD family boosts data center performance, conserves power and cooling resources, and maximizes space efficiency. It overcomes the capacity bottlenecks associated with the conventional NAND flash I/O constraints with products backed by a company with proven experience in SSD design.

Figure 3: Novachips' Flash Storage Processor



Key Benefits of Novachips PCIe SSDs

Novachips' first generation NVMe controller core, the NovaExpress, implements a single PCI Function NVMe Controller that is compliant with the NVMe Specification Revision 1.1. It utilizes a PCIe Gen2 x4 interface core and bridges to an ARM based SOC subsystem interface where the non-volatile memory interface controller is embedded. Its descriptor based NVMe Interface operation is thoroughly visible to the embedded firmware and enables comprehensive management by the firmware while providing automatic processing of a set of key functions, such as queue management, and a selective set of command processing. As the first generation NVMe controllers from Novachips, it takes the most essential set of high-end features supported by the NVMe Interface and targets them for the entry-to-mid level enterprise applications. Plus it is built for growth: its flexible architecture makes the core scalable to easily integrate more features in future generation components.

Sustained performance

Current SSD design suffers performance degradation for several reasons related to the processor. The processor does not have the power to handle garbage collection and wear leveling at the same time. So those operations become blocked. This is an issue with high performance, high write requirements. Novachips PCIe SSDs solves this problem by removing these household tasks from the I/O path. This is accomplished with an advanced processor with sufficient processing power.

2-level Bad Block Management

No matter how smart the wear leveling algorithm is, an intrinsic limitation of NAND flash memories is the presence of bad blocks (BB), i.e., blocks that contain one or more locations whose reliability is not guaranteed. In addition to maintenance at the drive level, the SSD must also perform maintenance at the chip level. In every NAND cell, each page contains a few extra bytes of extra capacity that the SSD controller uses to store a parity bit. Error correction code uses the parity bit during read or write operations. When it detects a read failure, it can try to reconstruct the data from the parity bit. If that is not possible, the controller marks the block as bad and retrieves the data from another RAID location.

The Bad Block Management (BBM) module creates and maintains a map of bad blocks. Flash memory as it is manufactured at the factory always contains some bad blocks and memory cells. Factor testing creates this map during factory initialization of the memory device. The controller updates this information while the memory is operating.

Read Refresh by Bit Error Monitoring

As NAND flash floating gate scaling is approaching 10 nanometers, most NAND flash based suffer SSD significantly from low endurance as each NAND flash memory cell can tolerate only a limited number of program/erase (P/E) cycles (about 100,000). Bit error rate (BER) increases exponentially with P/E cycles, which causes further unrecoverable

data retention errors. Enterprise SSD's data retention must be guaranteed even under a power-off condition, so the SSD controller must ensure that some idling static data is refreshed regularly. The Novachips NVS5700 advanced flash storage processor performs an intelligent Background Data Refresh to detect bit errors more efficiently. This refresh is where the controller reads and recalculates the internal ECC on the complete SSD within certain days and dynamically optimizes page refresh based on the applied error correction level.

Enhanced NAND Reliability with Static & Dynamic Wear Leveling

Flash memory devices have a limited and predictable lifespan due to their physical properties. Write and erase operations stress the memory cell over time. Compounding this fact is the probability of an imbalance or a bias in cell activity. This creates an irregular distribution of these unstable cells. Wear leveling techniques rely on the concept of logical to physical translation—that is, each time the host application requires updates to the same (logical) sector, the memory controller dynamically maps the sector onto a different (physical) sector, keeping track of the mapping either in a specific table or with pointers. The out-of-date copy of the sector is tagged as both invalid and eligible to be erased. In this way, all the physical blocks are used evenly, thus extending the device's lifespan.

There are two techniques for wear leveling. Dynamic wear leveling writes to the lowest address block available with the lowest erase count. Static wear leveling marks a block as eligible for remapping as soon as its age (how many times it has been programmed) deviates from the average.

Resident within the Novachips controller is the necessary intelligence to mitigate these challenges and optimize the life of the NAND flash. The Novachips controller wear leveling algorithm monitors and adjusts the rate of writes, erasures, and refreshes to eliminate biased activity to balance the rate of deterioration in the flash memory blocks. It then manages the physical data location with the best match between the I/O frequency on the data and the NAND flash block usage status.

2 Dimension Data Randomization

Not only can the Novachips NVS5700 SSD controller manage up to 128 NAND chips per channel, it can also support inline compression. Any data stream of "0" or "1" is compressed in real time and subject to an algorithm that remaps the data with a pointer. This technique can deliver up to a 94% savings in storage space. By enabling this space savings and preserving the reserved capacity for background tasks such as garbage collection and wear leveling, write performance is substantially increased.

Data Integrity

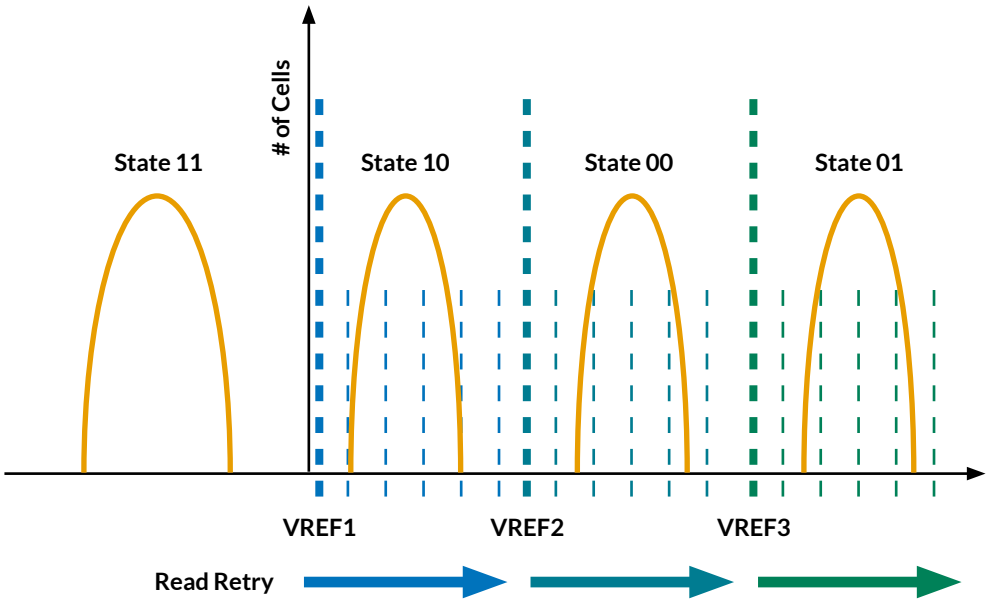
To verify data integrity, for every 2 KB of data written, 128 bits of ECC are appended. In addition, these 128 bits are used for data correction. This saves space over the MLC specification by 2½ times, which calls for 24 bits for each 1 KB of data.

Advanced Read Retry

As NAND memories cells get smaller, the potential for electrical fluctuations at the floating gate increases, especially with TLC and MLC flash memory. During the read operation of NAND flash, the memory cell's threshold voltage is iteratively compared to predefined read reference voltages. The upper and lower bound read reference voltages are identified, thus determining the stored cell value. However, when the threshold voltage distributions are distorted due to P/E cycling, charge loss over time, or program interference from the programming of neighboring cells, the threshold voltage distributions can shift. The distribution tails can enter the previously non-overlapping distribution margin regions, crossing the fixed read reference voltage levels.

To overcome such errors, a new mechanism called Advanced Read Retry (ARR) is implemented in the Novachips' NVS5700 series controller. Advanced Read Retry allows the read reference voltages to be dynamically adjusted to track changes in distributions and intelligently retries the failed reads with the newly adjusted reference levels such that read errors are decreased or even eliminated.

Figure 4: Novachips' Advanced Read Retry



Hardware-based RAID

Novachips' Hardware-based RAID (HW-RAID) consists of block-level striping with distributed parity. (This parity information can be used to reconstruct data). In principle, the operation of HW-RAID inside the Novachips' NVS5700 series controller closely resembles conventional redundant array of independent disks (RAID) protection of hard disk drive (HDD) arrays.

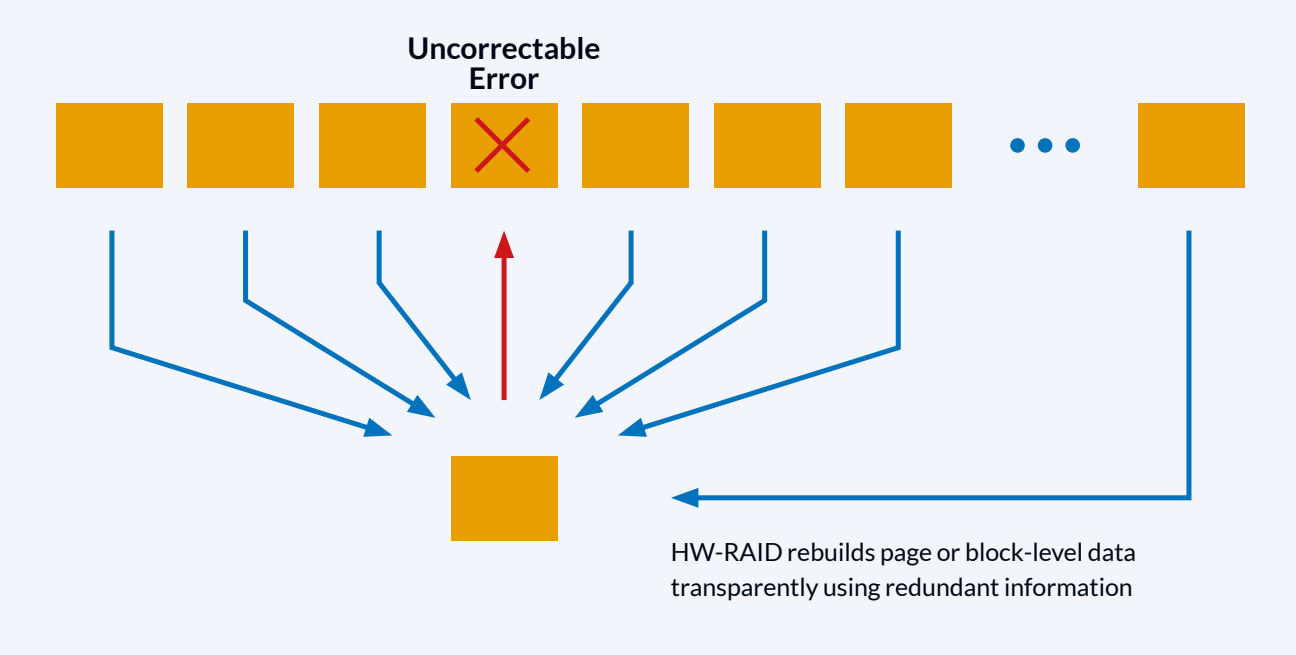
Data elements are used to calculate parity. Then the data and parity are stored as a single logical object. The specific HW-RAID implementation is a key element of Novachips' SSD design and is optimized for multiple factors such as NAND flash characteristics, intended usage model, and endurance requirements.

For example, with N+1 RAID; N pages of user data plus one page of parity is one stripe. Each element of the stripe is written to different blocks, planes, dies, and packages. This both protects user data from either cell failure of failure of the whole device.

To illustrate, each storage element could be thought of as a data block in a traditional disk array using RAID, the NAND dies could be disk drives and the multi-die-stacking package could be the drive shelf. Like a disk array, that's designed to lose an entire drive shelf without data loss, this kind of HW-RAID scheme allows an entire NAND flash package to fail, losing multiple NAND dies, without causing data loss.

Novachips' HW-RAID feature is constructed from redundant information stored in multiple locations on the SSD's NAND flash devices. This allows it to rebuild page or block level data transparently, to a known good point in time, as illustrated in Figure 5. If a failure occurs, the Novachips' NVS5700 series controller automatically detects it and rebuilds the data. During this HW-RAID rebuilding process, the SSD's performance is reduced temporarily, but it recovers after the rebuild process completes.

Figure 5: Novachips' Hardware-based RAID (HW-RAID)



End-to-End Data Path Protection (DPP)

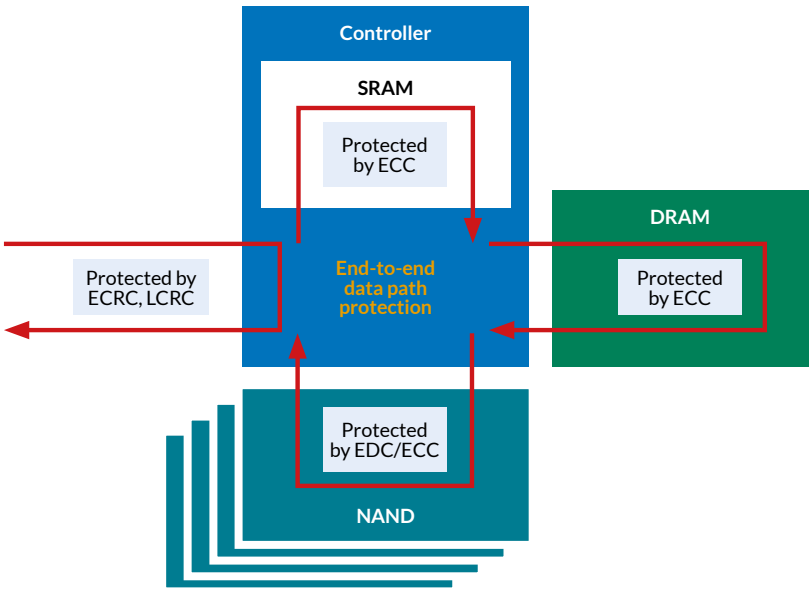
In order to ensure reliable write operations, data need to be verified as it moves from the host to the SSD interface and NAND flash media. This includes metadata such as Logical Block Address (LBA), which it the data block's index.

With each new generation, the technology node shrinks, the complexity of managing smaller NAND flash cell geometries increases and write/erase endurance decreases. Accordingly, End-to-End Data Path Protection (DPP) feature has become the recommended solution by NAND flash manufacturers to manage and enhance reliability of Flash-based solid state storage devices.

If uncorrectable bit errors occur and are not detected by the SSD's ECC algorithm, invalid data can be returned to the user application. If this is a mission-critical application, it could damage the business's reputation—imagine an incorrect stock trade—if this is not fixed and corrected immediately.

In order to protect mission critical user data, even in unexpected situations, Novachips NVS5700 series controller's End-to-End Data Path Protection (DPP) implements end-to-end cyclical redundancy check (ECRC) and link cyclical redundancy check (LCRC) parity codes. This is the first step of data transmission from the host. It stores this parity code together with data in internal memory. At the same time, all of the NAND flash memory data is fully protected by 8-bit error-detection-code (EDC) and Novachips' proprietary BCH error-correction-code (ECC). In addition, internal SRAM buffer memory is fully protected by the ECC algorithm inside the Novachips' NVS5700 series controller cores. This sophisticated and intelligent end-to-end data path protection eliminates the possibility of false error correction, and unexpected silent errors, plus guarantees that mission critical data is well protected.

Figure 6: Novachips' End-to-End Data Path Protection (DPP)



Conclusion

Novachips PCIe based SSD is the most advanced and cost-effective storage for data centers requiring the performance of SSD. Novachips has overcome the challenges of how to provide enterprise applications with long-endurance, reliable, high-capacity SSD storage. Our goal was to deliver this solution in a cost-effective manner for our customers, by driving down the required power consumption and utilizing competitively priced high-density MLC technology. All of this promises to push Novachips to frontrunner position in this fast growing market.

PN 14HL151

novachips.com