Improving RAID Performance and Reliability with Non-volatile Write Journals

By Ravi Prakash, Senior Product Applications Manager, and Shivendra Singh, Applications Engineer Staff, Cypress Semiconductor Corp.

When RAID was first introduced, the term was defined as “Redundant Array of Inexpensive Disks”. It has now become a blanket term for data storage schemes with better fault tolerance and input/output (I/O) performance than a single disk drive or ‘Just a Bunch Of Disks’ (JBOD). RAID commonly represents a system consisting of many storage disks that are accessed by servers through high speed communication channels. The I/O performance of disk drives has always been the major bottleneck for any system requiring access to huge amounts of storage and there were several efforts being made at that time to increase the throughput of disks through mechanical means. Such efforts, however, led to cost increases of these high-density, high-speed disks, making them unaffordable for small businesses.

RAID was developed to encounter all the above shortcomings and has became a popularly accepted approach based purely on principles of redundancy and parallelism that allows the use of inexpensive storage media to achieve very high performance and safety from disk failures. The popularity of RAID was further boosted by the success of the Internet which generated huge demand for high performance data storage and management systems. Over time, the definition of RAID has changed to “Redundant Array of Independent Disks”, making the term more generic and applicable to higher cost disk drives like SAS, solid state drives, and others.

There are different implementations of RAID systems suitable for different applications. Most of these implementations were suggested in the original Berkeley RAID paper and some were invented based on industry needs. These different implementations are referred to as RAID ‘Levels’ which are differentiated based on parallelism, redundancy, and duplication. This article will show how to improve the performance, efficiency, and reliability of RAID systems by combining write journaling techniques with non-volatile memory technology.

**RAID Controller Architecture**

Figure 2 shows a generic RAID Storage system architecture. The data from the servers connected on the network is received by a switch which directs this data to different storage boxes. The processor in the storage box also interacts with multiple storage boxes through the switch to enable virtualization. The data received by the processor is sent to the RAID controller. The RAID controller may be implemented in software or in hardware as a RAID on Chip (RoC). A software implementation of RAID is typically done on the motherboard chipset.
Cache

The RAID controller has direct access to the cache memory that enables fast read/write access to the storage system. The cache is used to write the data in transition. The RAID system uses a cache to speed up the apparent I/O performance of the storage system so that the host processor is available for other tasks. In modern RAID systems, the size of these cache memories is very large and typically use high speeds SDRAMs.

Read Cache

A read cache memory is used to speed up the read process by predictive caching of expected reads. This cache memory helps reduce disk drive seek latency for read accesses by the processor.

Write Cache

The write cache is one of the most important components of a high performance RAID system. A RAID write may operate in two different modes:

- **Write Through Mode**: In write through mode, the data sent by the host is directly written to the disks by the RAID system and the cache is bypassed. The Host waits for the RAID controller to signal completion of data write before sending the next data. Though the write throughput in this mode is significantly lower, it ensures integrity of data as the host is only acknowledged when the data is committed to the disk.

- **Write Back Mode**: In write back mode, the RAID controller writes data from the host to the cache memory and acknowledges write completion to the host. The host is free to perform other tasks while the RAID controller transfers the data from the write cache to the disk drives. Although this approach significantly increases write performance, the actual transfer of data to the disks is transparent to the host. This raises a potential risk to data integrity in case of power failure after the host has acknowledged the write but before the data is committed to disk. There are several ways to improve data integrity while using a write back cache, with the most important method having a reliable power back up for the cache.
Write Journal

A write journal or write-in-progress record is used in ‘Cache Write Back’ mode to keep track of all ongoing write transactions. A write-back cache improves performance significantly but comes with an increased risk to the data integrity of the system. In write-back cache mode, the RAID controller does not wait for the data to be committed to disk before acknowledging the host about completion of the write. The data is transferred from the cache to disk internally and this movement of data is transparent to the host. After it receives acknowledgement from the RAID controller, the host assumes that the data sent has been securely stored to the physical disk drives. However, power failure, disk failure, or any other failure that inhibits the completion of data movement from the cache to disk drives may lead to system failure.

Consider a situation where the storage system acknowledges the host after writing data into the cache, but the power goes off before the system transfers the data from cache to the physical drives (as shown in Figure 3). The data stored is usually secured using a battery backup that retains the contents of the cache even during power failure. On the next power up, the storage system has to track back each transaction to find which blocks of data were not written onto physical drives. Therefore, even with battery backed cache, system recovery time may be significant as the system needs to be checked for the exact point of failure before recovery can begin.

To avoid such recovery delays, many storage systems log each transaction of data from the host to the cache and from cache to the disk in a nonvolatile circular buffer. During the next power-up, this log can be retrieved to identify the state of the storage system before failure and substantially speed recovery. This method is called write-in-progress record or ‘Write Journaling’.

Figure 3 RAID Controller – Data Flow
Choice of Memory Device for Write Journaling

Choosing the right nonvolatile memory type for write journaling is important for performance and reliability of a RAID system. While a RAID system can use battery-backed DDR cache for data caching, it can’t for write journal because of:

- **Efficiency**: In practice, large blocks of data are being written to the DDR cache continuously at very high speed. However, only small blocks of data are being written to the journal memory. Using a part of DDR cache as a write journal requires the system to stall to complete writing journaling data. This, in effect, leads to inefficient usage of cache bandwidth and adversely impacts regular RAID performance.

- **Redundancy**: Placing the journal memory in the write cache creates a single point of failure, for if the write cache is compromised, both cache data and the write journal will be lost.

- **Durability**: Cache memories are typically battery-backed SDRAMs. Usually, these have a limited power down lifetime of 72 hours. A dedicated nonvolatile memory for write journaling ensures that even if the cache data is lost, the system can identify the status of operations before the power failure occurred.

- **Availability**: Independent access to the write journal allows the system to accept new data to be written from the host even when it is reading the write journal. This helps to identify completed and pending writes to the disk while at the same time increasing system availability, a factor of prime importance to any storage system.

**NVRAM Parameters**

Choosing an appropriate non-volatile memory for write journaling is critical to system performance and reliability. Engineers should consider the following critical parameters:

- **Fast Read/Write Access**: The NVRAM should be able to handle fast read and write accesses. A slow journal can potentially slow down the whole storage system.

- **High Reliability**: Data reliability of journal entries is important if the system is to meet RAID atomicity, consistency, and more importantly, durability goals. This memory space is also used to store some configuration details and failure logs with time stamping. Therefore, high reliability and data integrity is essential for RAID application.

- **High Endurance**: Data needs to be written in a circular buffer fashion at a very high speed. Given the volume of data passing through RAIDs, the journal memory must have a very high endurance (i.e., support a large number of write/erase cycles).

**nvSRAM for Write Journaling**

The standard asynchronous SRAM interface on an nvSRAM enables high-speed read and writes access (up to 20 ns) while allowing the nvSRAM to be written or read effectively indefinitely. This is possible because the nonvolatile elements of nvSRAM provide reliable data storage without the need of a battery to protect data during power down. In addition, the endurance of nvSRAM is counted in terms of number of times power is lost for it is only when the system power goes down that a store operation to NV elements takes place. Specifically, data backup to nonvolatile memory is done automatically on power failure using stored energy on a small capacitor connected to the VCAP pin of the nvSRAM chip.
- **Fast Access**: nvSRAM is the fastest nonvolatile RAM in the industry with 20 ns read and writes access time. This enhances system performance by reducing access time.

- **Reliability**: SONOS technology provides nvSRAM with unmatched reliability. Data retention of 20 years at industrial temperature range is the highest among any nonvolatile RAM technologies available today.

- **Unlimited Endurance**: The SRAM interface of nvSRAM allows the write journal to be written infinitely without the worry of running out of endurance or wear leveling. This takes away all the effort required for implementing wear leveling or monitoring endurance cycles in software.

- **Environmental Friendly**: Unlike a battery-backed SRAM, nvSRAM is a green technology with no battery involved. A small capacitor is used to store power and utilized for AutoStore operation during power fail conditions.

- **Real Time Clock**: Many RAID systems time stamp write journal data. This typically requires an additional RTC device and thus increasing the total BOM cost and board area. With an integrated RTC function, nvSRAM frees up resources on RAID system such as communication channels and reduces software development.

Write journaling using nvSRAM is an effective and reliable way to implement a non-volatile cache for RAID systems. With high-speed read and write operations, infinite write endurance, battery-less operation, scalability, and reliable SONOS technology, nvSRAM-based designs provide a robust foundation for non-volatile cache in RAID systems.