Solid-state drive controller with embedded RAID functions

Jianjun Luo$^{1}$, Lingyan-Fan$^{1\text{a})}$, Chris Tsu$^{2}$, and Xuan Geng$^{3}$

$^{1}$ Micro-Electronics Research Institute, Hangzhou Dianzi University, Zhejiang, 310018, China
$^{2}$ Sage Microelectronics Corporation, 910 Campisi Way, Campbell, CA95008, USA
$^{3}$ Department of Electronic Engineering, College of Information Engineering of Shanghai Maritime University, Shanghai, 201306, China

$^{a})$ fanlingyan@hdu.edu.cn

Abstract: A novel solid state drive (SSD) controller integrated circuit structure is presented. This controller has up to five flash memory channels. These five channels, regarded as four independent virtual disks in a redundant array of independent disks (RAID) system, can be configured as RAID0 or RAID5 mode. This real silicon controller was silicon proven working well in a single SSD unit with intrinsic RAID (iRAID) functions configured as either RAID0 or RAID5 mode. The traditional concept that RAID has to be built up by multiple independent disks has been changed by this single SSD unit with intrinsic RAID embedded. It brings a new way to make big RAID storage systems at lower costs, higher performances, less power consumption, smaller in size and easier maintenance.

Keywords: NAND flash memory, solid state drive, RAID

Classification: Electron devices, circuits, and systems

References

1 Introduction

RAID has been widely applied in today’s storage systems. The term “RAID” was first defined as “Redundant array of inexpensive disks” [1] by David Patterson, Garth A. Gibson, and Randy Katz at the University of California, Berkeley in 1987 revealing the original expectation of low cost. However, a real RAID system was found “not inexpensive”. Therefore, the term changed from “inexpensive” to “independent”, defined as a technology that combines multiple disk drive components into a logical unit for the purposes of data redundancy and performance improvement. In technical terms, data is distributed across the drives in one of several ways (RAID levels) [2, 3], depending on the specific level of redundancy and performance required. A single drive has a high speed interface such as ATA (Advanced Technology Attachment), Serial ATA (SATA) [4], PCI (Peripheral Component Interconnect)/PCI Express (PCIE), etc.

Before the magnetic platter was replaced by flash memory as the dominating storage media, it was hard to make RAID functions embedded into a drive with mechanical parts. Flash memory with its 100% semiconductor compatible technologies make SSD very attractive in performance and reliability. A Solid State Disk (SSD) consists of a controller ASIC (Application Specific Integrated Circuit) and a group of flash memory chips. Usually data stream is striped among these flash memory chips by multiple flash channels.

If each flash memory channel is regarded as a virtual disk in a RAID system while RAID control logic is implemented inside a SSD controller, this controller chip can distribute data stream among these virtual disks following the exact same rule of RAID system. In this way, a SSD unit built up by such a controller is not only a single drive, but also a RAID micro-system providing the redundant capability with high reliability. S. Im et al. analyzed RAID0 and RAID5 micro-system with multi-channel flash memory chips [5], and here, a real RAID controller chip was implemented to prove the architecture effective.

2 The implementation

Fig. 1 is a SSD controller architecture with intrinsic RAID embedded. The supervisor CPU interprets SATA-II command sets and configures the RAID controller module. Each channel has a flash channel controller (FCC) with a small eight-bit RISC CPU core embedded to obtain intelligence channel configurations. FCC is taking care of flash management which includes wear-leveling algorithm, Error Correction Coding (ECC), bad block management, etc. Each channel with its corresponding buffer and flash memory chips can be regarded as a virtual disk in this intrinsic RAID (iRAID) micro-system. The channel number N can be an integer 1, 2, 3, 4 or more. Here, when it is realized as a real silicon chip, the channel number N is equal to 4. BCH algorithm [6] is applied to provide configurable ECC ability, which can correct up to 48 bit errors per 1K byte payload data.

The Controller can be configured under iRAID0 and iRAID5 modes:
1. RAID0 Mode
When configured in RAID0 mode, the RAID controller in Fig. 1 distributes data into virtual disks by striping operation. Currently most SSD controllers in use today can do this [7].

The striping data size $D_{size}$ can be defined as a sector (512 byte). Considering the flash memory’s unique specifications in which flash memory has to be programmed by a minimum unit of page and the page size is normally 2K byte, and up to 8K byte for the most updated flash part numbers, $D_{size}$ is recommended to be defined as 4 K byte or 8 K byte or even larger. However, the larger the $D_{size}$ is, the more internal SRAM buffer should be applied to match it. Here, $D_{size} = 4$ K byte is applied for both RAID0 and RAID5 for evaluation.

2. RAID5 Mode
When configured in RAID5 mode, the RAID controller in Fig. 1 has $N+1$ virtual disks. The data stream from the host is striped to virtual disk 0, 1, ... and $N$ while the additional virtual disk (the parity disk) is accepting data result of Exclusive-OR (XOR) operation of all $N$ channels with data packet size of $D_{size}$. The content in the parity disk is calculated by

$$D(V_p) = D(V_{n+1}) = D(V_0) \oplus D(V_1) \oplus \ldots \oplus D(V_N)$$

(1)

Here, $D(V_i)$, $i = 0, 1, 2, 3, \ldots$ means Data packet distributed to virtual disk $i$. $V_p$ means the virtual disk for parity.

Fig. 1. SSD controller architecture with intrinsic RAID embedded
3. Flexibility with multi-modes
The integrated circuit in Fig. 1 here was designed with total five flash memory
buses supporting five virtual disks in maximum. It provided the flexibility of
being configured by firmware to work under RAID0 or RAID5 modes. The
controller can support the following configuration options:
(1) RAID0 mode: Data are striped to N channels and no redundant channel
for parity. N can be 2, 3, 4 or 5 here. Typically, four-channel mode can be
applied for 64 GB SSD with RAID0 function embedded if 16 GB per
TSOP package flash chip mounted on PCB (Printed Circuit Board) for
each channel or flash bus. And 80 GB SSD can be built up in five-channel
RAID0 mode.
(2) RAID5 mode: RAID5 SSD can be configured to N+1 channel. N can be 2,
3 or 4 for this controller. The typical application is to apply N = 4. This
means four channels for data striping while one additional channel for
parity. For example, if 16 GB per TSOP flash package applied, a 64 GB
RAID5 SSD is constructed with the 16 GB memory in channel 4 occupied
by parity.
It is possible to change a SSD working in RAID0 mode with five flash channels
to RAID5 mode. Here are two cases:
(1) If data are striped to four channels in RAID0 while keeping the channel 4
as a redundant (inactive) channel, the only thing firmware has to do is to
calculate the parity and store it into the redundant channel. Assume:
each channel has 16 GB flash memory; SSD in RAID0 mode has capacity
of 64 GB while 16 GB in channel 4 is not used for any data or parity. And
when changed to RAID5 mode, channel 4 has to be activated as the
parity channel. The RAID5 SSD has 64 GB capacity for data. When SSD
unit being changed from RAID0 to RAID5, no data lost.
(2) If data are striped to five channels in RAID0, all data stored in flash
memory chips have to be destroyed when re-configured to RAID5 mode.
This means RAID0 SSD has to be physically rebuilt and formatted. If
each channel density is 16 GB, SSD in RAID0 mode has capacity of
80 GB. And then changed to RAID5 mode, the SSD unit has visual
capacity 64 GB while 16 GB of channel 4 is used for parity.
It is easy to change a RAID5 mode SSD to work under RAID0 mode by
simply ignoring the parity channel and no data lost.

3 Results
The SSD controller with iRAID embedded had a real silicon die area as
3.1 mm * 3.1 mm under 0.11 um CMOS process (shrunk from 0.13 um proc-
cess), shown in Fig. 2. The performance results are shown in Table I. RAID0
had the full speed of all four channels and almost reached the maximum speed
of SATA-II throughput can support. Considering the overhead of SATA
protocol, RAID0 with five channels has speed up to 90% of SATA-II’s
theoretical speed (300 MB/s), and a little bit slower for RAID0 with four
channels. This means RAID0 speed can reach 80%–90% of SATA-II’s theo-
retical speed. One channel was scarified as the parity disk in iRAID5 mode, so the read/write speed has a discount about 75% in theory. In order to minimize such a discount, there were 32 KB FIFO inside the controller IC as the 1st level buffer. Furthermore, NCQ (Natural Command Que) was applied to make commands well-organized. Thus, the final measured read speed revealed RAID0 and RAID5 modes having 20 MB/s–30 MB/s discount.

Fig. 2. Snap-shot of the real silicon die

Table I. Performance measured by HD Tune Pro 5.00

<table>
<thead>
<tr>
<th>Test Items</th>
<th>RESULTS (Performance)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td>RAID5</td>
</tr>
<tr>
<td>Channel Number</td>
<td>3+1</td>
</tr>
<tr>
<td>Seq. read speed (MB/s)</td>
<td>224.8</td>
</tr>
<tr>
<td>Seq. write speed (MB/s)</td>
<td>211.3</td>
</tr>
<tr>
<td>Random read speed 4 KB (MB/s)</td>
<td>29.50</td>
</tr>
<tr>
<td>Random write speed 4 KB (MB/s)</td>
<td>16.26</td>
</tr>
</tbody>
</table>

In order to verify the recover ability, a flash memory daughter board was randomly plugged out, and replaced by a new memory board. Data in the corresponding channel was recovered and stored into the new daughter board in ten to twenty minutes depending on the system busy and idle time.

Table II shows the measured power consumption of the SSD units built up with this controller. Four SSD samples were tested. There were five TSOP package flash memory chips on each sample SSD board. In RAID5 with 4+1 channel mode and RAID0 with 5 striping channels, all five flash memory chips were activated. In RAID5 with 3+1 channel mode and RAID0 with 4 striping channels, flash channel 4 was inactive (deselected).

(1) Working Current: These four samples were configured to the corresponding RAID mode and keep them in read/write working status. There was no power consumption difference between RAID5 and RAID0 configuration if the same flash channel number was applied. When four channels working active, the power current was about 155 mA. When five channels active, there were about 50 mA current increased. The power consumption was dominated by the number of flash memory channels activated.
(2) Standby Current: All the four samples were configured to the corresponding RAID modes and keep them powered but no read/write activities. The power consumption results were in the range of 40 mA–50 mA, which dominated by SATA-II PHY bias, PLL current and the whole chip’s leakage current.

### 4 Conclusion

A single solid-state drive (SSD) has been built with a novel SSD controller and a group of flash memory chips. This SSD controller, designed with configurable RAID operation among the flash channels, was silicon proven working well with RAID0 and RAID5 functions. Although the prototype SSD controller with iRAID embedded was only an initial stage integrated circuit, the preliminary results of this architecture demonstrated that a RAID system would no longer be a bunch of independent drives, and possibly come with only a single disk. This will make those storage systems with RAID, such as cloud storage systems, lower in cost, better performance, consume less power, smaller in size and easier to maintain.

### Acknowledgments

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ12F01001, and the Scientific Research Fund of Zhejiang Provincial Education Department (No. Y200803237).

---

#### Table II. Power consumption

<table>
<thead>
<tr>
<th>Test Items</th>
<th>RESULTS (Performance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>RAID5</td>
</tr>
<tr>
<td>Channel Number</td>
<td>3+1</td>
</tr>
<tr>
<td>Standby Current (mA)</td>
<td></td>
</tr>
<tr>
<td>Sample 1</td>
<td>40.1</td>
</tr>
<tr>
<td>Sample 2</td>
<td>42.3</td>
</tr>
<tr>
<td>Sample 3</td>
<td>44.0</td>
</tr>
<tr>
<td>Sample 4</td>
<td>49.5</td>
</tr>
<tr>
<td>Working Current (mA)</td>
<td></td>
</tr>
<tr>
<td>Sample 1</td>
<td>150.4</td>
</tr>
<tr>
<td>Sample 2</td>
<td>161.6</td>
</tr>
<tr>
<td>Sample 3</td>
<td>155.6</td>
</tr>
<tr>
<td>Sample 4</td>
<td>154.1</td>
</tr>
</tbody>
</table>