Reliable Energy-Aware SSD based RAID-6 System

(FAST 2012, Feburary 14-17, San Jose, U.S.A)

Mehdi Pirahandeh, Deok-Hwan Kim*

Dept. of Electronic Engineering, Inha University, Incheon, South Korea

mehdi@iesl.inha.ac.kr, deokhwan@inha.ac.kr, * Corresponding Author

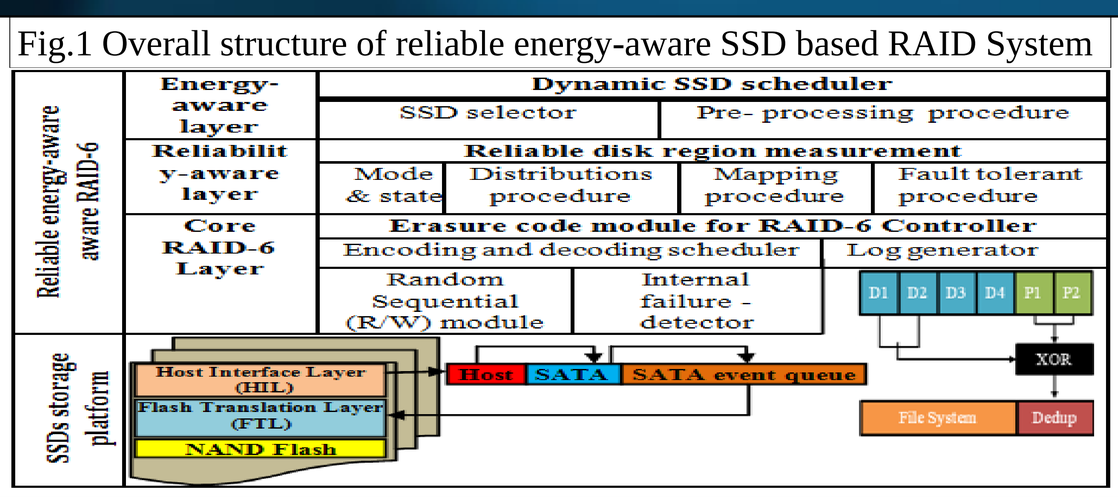
Motivation and goals

- **Challenges of SSD markets deal with**
- Breakthroughs in terms of energy consumption, reliability, performance of SSDs
- Achieve optimal energy consumption for HDD, SSD and large storage systems
  - ✓ Dynamic voltage measurement
  - ✓ Auto power management
  - ✓ Compiler directed energy optimization
- Performance and energy use of RAID systems with various type SSDs
- Reliability system and some criteria
  - ✓ Cost for replication based schemes are expensive
  - ✓ Erasure codes considering power management
  - ✓Need to measure repair transition rate
  - ✓ Average data loss during *N* iterations of IO operations on critically exposed sectors
  - ✓ Lower utilization level produces lower failure rates

- **Goals**
- Proposes an energy aware algorithm
- Model which enable SSDs to increase the performance and decrease the level of power consumption

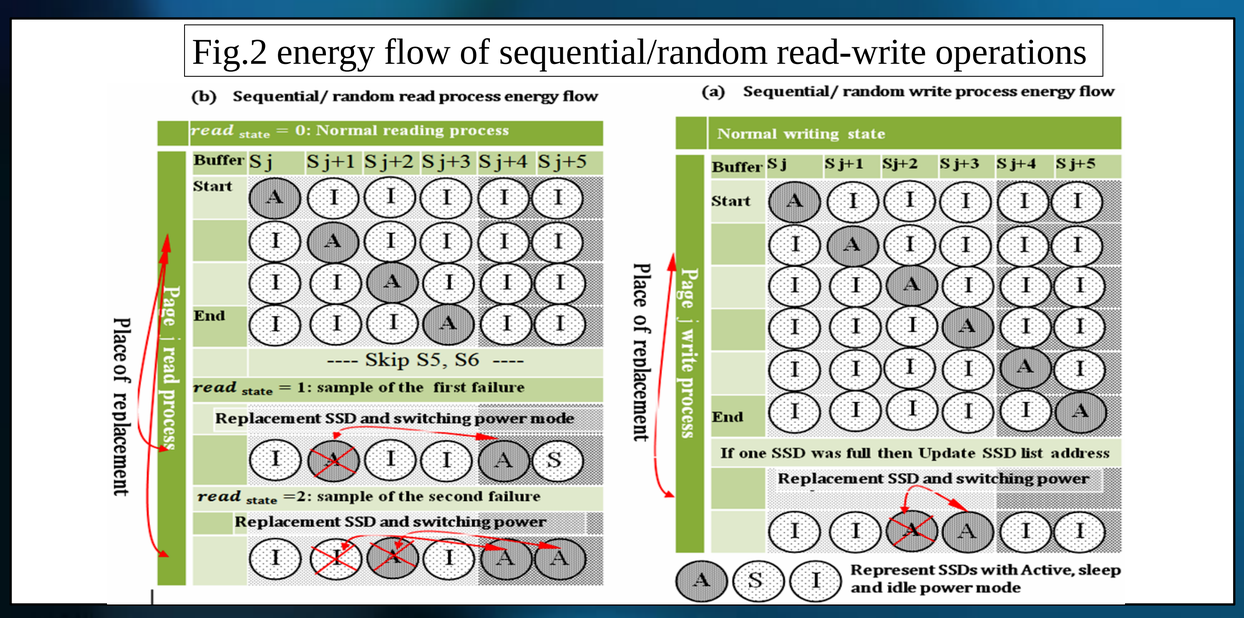
Proposed RAID System

Fig.1 Overall structure of reliable energy-aware SSD based RAID System



- Core layer in the RAID-6 controller
  - ✓ Encoding and decoding scheduler
  - ✓ Random and sequential read and write modules with an internal failure detector using erasure codes
  - ✓ Traces are updated using the log generator
- Reliability-aware layer
  - ✓ Procedures for prediction of SSD reliability
    - ➢ Use erasure codes generated from the core layer
- Energy-aware layer
  - ✓ Pre-processing procedure
    - ➢ Initialize the reliability measurement and utilization level
  - ✓ SSD selector and pre-processing procedure.
    - ➢ Set the status of selected SSD power mode to idle-sleep-active via imported traces
  - ✓ Dynamic SSD scheduler
    - ➢ Visualize the statistics of SSD energy consumption
    - ➢ Update the power mode into idle, sleep or active.
- Host interface layer, flash translation layer and NAND flash chips.

Fig.2 energy flow of sequential/random read-write operations



- ▪ Energy flow of sequential/random read-write operation using proposed model
  - ✓ Data pages are segmented into large chunks
  - ✓ Use power switching of SSDs after writing of each chunk is done.
    - ➢ After reading-writeing of current chunk in SSD ***Sj***, IO operation of next chunk is performed in SSD ***Sj*+1**
    - ➢ Measure SSD reliability and choose the parity SSD ***Sj*+4** and ***Sj*+5** with less utilization level.
  - ✓ Read operation needs to access four disks from ***Sj*** to ***Sj*+3** sequentially and skips two parity SSDs
    - ➢ When one disk fails, 1st parity SSD is "active" and 2nd parity SSD is in "sleep"
    - ➢When two disk failure occurs, both SSDs are in "active" mode
  - ✓ Write operation requires accessing six SSDs using power switching modes

Table 1: pseudo code of reliable energy-fault aware algorithm

Table 1: reliable energy-fault aware algorithm

```
1. Generate SSDs matrix
2. Initialize the estimated power modes and Set current mode as "idle"
3. Initialize utilization level and average data loss
4. Initialize Δ t_delay
5. Define Page with Chunk Matrix: P [p0, p1, p2, p3, p4, p5]
6. Define Page address : Pa[a0, a1, a2, a3, a4, a5]
7. if (current event == read)
8.   Initialize Pa[] list
9.   While ( Not end of Pa [] list) then
10.  If normal state then Set read state = 0
11.    Switch current SSD power mode from "idle" to "active"
12     read data in Pa[0]...Pa[3] and store them into current Page P[]
13.    Switch current SSD power mode from "active" to "idle"
14.  If first failure happen then read state =1, Increase t rebuild
15.    Set power mode of 1st parity as active and 2nd parity as sleep
16.    Set read state = 0
17.    Replace the failed SSD with 1st parity, Stop t rebuild
18.  If Second failure happen then Set read state =2,Increase t rebuild
19.    Set both 1st and 2nd parity SSDs power mode as active
20.    Set read state = 0
21.    Replace the failed SSD with 2nd parity, Stop t rebuild
22. if (current event == write )
24.    Create Pa[]
25.    While ( Not end of Pa[] list) then
26.      Switch the SSD power mode from "idle" to "active"
27       Write data in Pa [0]...Pa [5] into corresponding SSDs
28.      Switch the SSD power mode from "active" to "idle"
29.      if SSD corresponding to Pa[] is full, then update Pa[]
30.  Estimate total power consumption(E Total)
```

Preliminary Result

- ▪ Estimate power consumption manually for each SSD
  - ✓ Multiple power modes - waken(1.2w), active(2.4w), sleep(0.06w), idle(0.5w) and off
  - ✓ Average delay time between switching power mode smaller than that of traditional model
  - ✓ Dynamic SSD scheduler can activate each disk adaptively and minimize time of waken mode of SSDs.
  - ✓Average data loss is calculated based on the Markov model and utilization level
- SSD energy consumption for busy sever-like workload
  - ✓ The estimation period is 2 sec.
  - ✓ 60% energy saving for read operation and 48% for write operation compared to the general model
  - ✓ We can see that the proposed model significantly reduces the energy consumption by controlling the power switch of SSDs adaptively.

Table 2: SSD energy consumption for busy sever-like workload

	Power Modes	Idle	Sleep	Active	Waken	Total
Proposed Model	Read operation	0.5w*0.8s	0.06w*0.2s	2.4w*0.8s	1.2w*0.2s	2.572w
	Write operation	0.5w*0.4s	0.06w*0.2s	2.4w*1.2s	1.2w*0.2s	3.332w
General model	Read operation	0.5w*0.4s	-	4.8w*1.2s	1.2w*0.4s	6.44w
	Write operation	0.5w*0.4s	-	4.8w*1.2s	1.2w*0.4s	6.44w

Contribution

- Improved approach for periodic estimating the energy consumption of SSDs
- The reliability estimation considered to enhance the energy efficiency on the SSD based RAID-6 system
- A layered architecture for reliable energy-aware RAID-6 system
- Reduce energy consumption of parallel access of SSDs
  - ✓Dynamically switching of SSD power modes among active, idle, waken and sleep
  - ✓ Allowing one of parity SSDs in safe zone to be sleep mode
  - ✓ Segmenting data pages into large chunks
  - ✓ Using the power switching of SSDs after writing or reading of each chunk is done.
- Reduce energy consumption during active mode
  - ✓ Activating SSDs sequentially and minimizing delay time of power switching of SSDs
  - ✓ Avoiding repetitive accessing of same SSD by using large chunks
  - ✓ Increasing number of disks in idle mode

Acknowledgements

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0004114) and in part by the Ministry of Knowledge Economy(MKE) and Korea Institute for Advancement in Technology (KIAT) through the Workforce Development Program in Strategic Technology.