NEXT-GENERATION SSD VERIFICATION PLATFORM WITH TERA-SCALE NAND CAPACITY AND STORAGE SIGNAL PROCESSING SUPPORT

NVRAMOS 2012-10-16

Seil Lee, Myunghyun Rhee, and Sungroh Yoon
Advanced Computing Laboratory, Seoul National University
Contents

Trends & Issues
- Trends of NVRAM
- Demand for Storage
- NAND Scaling and Challenges

Verification Platform for SSD Development
- SSD Core Technology and Controller
- Existing Verification Platforms
- Our Platforms

Implement of NAND Controller
- Specification
- NAND Status Monitor
- ONFI Interface

Experiments
- Maximum performance test
- FTL
- ECC
- NSM TEST
Trends & Issues

- Trends of NVRAM
- Demand for Storage
- NAND Scaling and Challenges
Trends of NVRAM

• Cost per GB Trend in 2012

• Market Trend in 2011

Source = DRAM Exchange/IDC/ASML

Advanced Computing Laboratory
http://acl.snu.ac.kr
Market of Storage Device

Storage Exploding

- 35M PB
- 4.6TB/person

NAND Flash Worldwide Usage

- Computing (SSD)
- Media Tablets
- Mobile Phone
- Consumer (Video, Music, Camera, USB)
- Other (including Auto, Gaming)

Compute CAGR growing at 250% YoY ('11-'15)

Source: Gartner December, 2011
"Forecast: Semiconductor Consumption by Electronic Equipment Type, Worldwide, 4Q11 Update"

eMMC Share of Total Flash Market

- 2009
- 2010
- 2011
- 2012
- 2013

Source: Micron Marketing, 2010
Embedded MultiMedia Card

- Not a card, but a chip
- Embedded storage solution with MMC interface, flash memory and controller
- Propelled by increased usage in smart mobile devices
• **Apple** acquired Israel-based **Anobit** for about $400 million (Dec, 2011)

• **SK Hynix** merged an SSD controller maker **LAMD** for ₩287 billion (Jun, 2012)
NAND Scaling and Challenges

- As NAND Technology node is scaled down
  - Capacity grows up
  - Number of electron per cell / signal Integrity decreases
  - Reliability (endurance) decreases

Source: "Advanced Flash Technology Status, Scaling Trends & Implications to Enterprise SSD Technology Enablement", Flash Memory Summit 2012
Verification Platform for SSD Development

- SSD Core Technology and Controller
- Existing Verification Platforms
- Our Platforms
SSD Core Components and Technology

Host Interface

- From Phy to Link/Trans. layer for High-speed interface
- Low-power / Reduced area

Controller

- Garbage Collection
- Wear-leveling
- Multi-channel
- Multi-core CPU
- Too heavy mapping info.
- Coding/Error Correction
- Better UBER
- Reduced area

NAND Flash

- From Phy to Link/Trans. layer for High-speed interface
- Low-power / Reduced area

DRAM

SSD (INDILINX Controller)

Multicore Architecture in Samsung SSD

Battery Runtime on Flash

Energy Consumption [Watts]

- SanDisk Solid State Drive
  32 GB, Flash, SATA/150
  Idle: 0.5, Load: 1.0
- MemoRight MR25.2-032S SSD
  32 GB, Flash, SATA/150
  Idle: 0.8, Load: 2.3
- Hitachi 7K200
  200 GB, 16 MB Cache, SATA/150
  Idle: 1.1, Load: 3.2
- Crucial SSD
  32 GB, Flash, SATA/150
  Idle: 1.6, Load: 2.9
- Mtron Flash SSD
  32 GB, Flash, SATA/150
  Idle: 2.0, Load: 2.9

HDD

ETC

- Low-power
- Sudden power loss protection

Advanced Computing Laboratory
http://acl.snu.ac.kr

10 / 43
Development of SSD Controller

- Tera-scale SSD controller should...
  - Be able to handle multi-channel (more than 10) NAND interface
  - Contain high speed host-interface
  - Carry out garbage collection very efficiently
  - Have high performance ECC module
Limitation of S/W Simulation

- SSD system contains various complicated components
  - CPU, SRAM, DRAM controller, Flash controller, ...
- The traditional S/W simulation is very useful but often slow to emulate the whole system
- FPGA-based simulation can help this out
# Existing Solutions

<table>
<thead>
<tr>
<th></th>
<th>Case A</th>
<th>Case B</th>
<th>Case C</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPGA</td>
<td>Virtex-4</td>
<td>Virtex-5 x 2</td>
<td>Virtex-2</td>
</tr>
<tr>
<td>Capacity</td>
<td>4DIMM Bank 2GB/Bank</td>
<td>320GB</td>
<td>32GB</td>
</tr>
<tr>
<td>NAND</td>
<td>Legacy</td>
<td>Legacy</td>
<td>Legacy</td>
</tr>
<tr>
<td>Host Interface</td>
<td>PATA-to-SATA</td>
<td>N.A.</td>
<td>Ethernet</td>
</tr>
</tbody>
</table>
Existing Solutions

CASE A

Old FPGA and legacy NAND Gigabyte-scale

Source: "Development Platforms for Flash Memory Solid State Disks", Hongseok Kim at al., ISORC 2008

CASE B

Legacy NAND Gigabyte-scale Not a real host interface

Source: "FPGA-Based Solid-State Drive Prototyping Platform", Yu Cai at al., FCCM 2011

CASE C

Old FPGA and Legacy NAND Gigabyte-scale Ethernet Host interface

Source: "BlueSSD: An Open Platform for Cross-layer Experiments for NAND Flash-based SSDs", Sungjin Lee at al.
Running Short of FPGA Resources

- The More channel capacities, the more ECC modules
- The more reliable operation by ECC module requires the more of the FPGA’s resources
Proposed Platform

- Brand new FPGA : Xilinx Virtex-6
- Support Various NAND chips, not only legacy NAND but also ONFI NAND chips
- Extend NAND channel interface up to 14-Channel
- Various helpful functions for the development
  - Power measurement per channel interface

Source: "ONFi: Achieving Breakthrough NAND Performance", MEMCON08 2012
## Target Specification

<table>
<thead>
<tr>
<th></th>
<th>Goal</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>4,096</td>
<td>Gigabyte</td>
</tr>
<tr>
<td>Read Speed</td>
<td>1,000</td>
<td>MB/s</td>
</tr>
<tr>
<td>Write Speed</td>
<td>800</td>
<td>MB/s</td>
</tr>
<tr>
<td>IOPS</td>
<td>80</td>
<td>KIOPS</td>
</tr>
<tr>
<td>H/I Speed</td>
<td>6</td>
<td>Gbit/s</td>
</tr>
<tr>
<td>H/I Power Consumption</td>
<td>100</td>
<td>mW</td>
</tr>
<tr>
<td>SSP Throughput</td>
<td>1,200</td>
<td>Mbit/s</td>
</tr>
</tbody>
</table>

### Samsung SSD 840 Pro vs 830

<table>
<thead>
<tr>
<th></th>
<th>Samsung SSD 830 (256,512GB)</th>
<th>Samsung SSD 840 Pro (256,512GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential Read</td>
<td>520MB/s</td>
<td>540MB/s</td>
</tr>
<tr>
<td>Sequential Write</td>
<td>400MB/s</td>
<td>450MB/s</td>
</tr>
<tr>
<td>Random Read</td>
<td>80K IOPS</td>
<td>100K IOPS</td>
</tr>
<tr>
<td>Random Write</td>
<td>36K IOPS</td>
<td>78K IOPS</td>
</tr>
<tr>
<td>Active Power Use</td>
<td>0.24W</td>
<td>0.068W</td>
</tr>
<tr>
<td>Idle Power Use</td>
<td>0.14W</td>
<td>0.042W</td>
</tr>
</tbody>
</table>

[Advanced Computing Laboratory](http://acl.snu.ac.kr)
Platform v1

Real SSD

Verification Platform

Controller

FPGA

NAND Flash

Daughter Board
Platform v1 Details

FPGA Board
- Xilinx Virtex6 240LXT
- DDR3 SODIMM 512MB
- USB JTAG, UART
- PCIe x8

Daughter Board
- MAX up-to 4CH/ 6WAY
- Piggy-back (SODIMM)
- MICTOR connector for debugging
- Both legacy and ONFI NAND
Limitations of Platform v1

- Too few I/O pins, which limits capacity (30 signals per one channel)
- Use of Ready/Busy signals not possible (suboptimal I/O management)
Improvement Ideas

- Design a custom FPGA board
  - Selectively benchmark the reference board
- Maximize utilization of FPGA I/O pins for NAND interface
- Support up to 14-CH/8-Way
- Efficient use of Ready/Busy signals
Platform v2

Real SSD

Controller

NAND Flash

Verification Platform

FPGA

Daughter Board

Advanced Computing Laboratory

http://acl.snu.ac.kr
Platform v2 Details

**FPGA Board**
- Our own design
- Xilinx Virtex6 240LXT
- DDR3 SODIMM 512MB
- MAX up-to 14CH / 8WAY

**Daughter Board**
- 2 Daughter board
- Each one for 8CH - One of 8CH, another with 6CH (2CH not used)
- High speed/reliable SAMTEC connector interface
- Piggy-back (SODIMM)
- MICTOR connector for debug
Implement of NAND Controller

- Specification
- NAND Status Monitor
- ONFI Interface
The first stage of NAND Controller
- Target: Platform v1
- 4-channel 4-way
- Legacy NAND flash
- No R/B pins

Tera-scale NAND Controller
- Target: Platform v2
- 14-channel 8-way
- ONFI NAND flash
- 1 R/B pins per channel
Baseline Architecture

- **Specifications**
  - CPU: 150MHz Microblaze softcore
  - RAM: 512MB SDRAM
  - BUS: PLB
NAND Controller

• NAND Flash:
  – 8GB Samsung Legacy NAND
    • 33.3MHz interface clock
  – 16GB Micron ONFI 2.2 NAND
    • 100MHz interface clock
• Shared Buffer (for multi channel)
  – 8 – 32 pages buffer
  – Buffer sharing policy : 4-way associativity
• ECC : 28-bit BCH
  – Parity 392 bits
• Capacity
  – 4 channel 4 way : 128GB (v1)
  – 14 channel 8 way : 1792GB (v2)
NAND Controller

- **Channel structure**
  - Command / Address / Status Register
  - 16KB Read/Write Buffer (v1)
  - No R/B signal (v)
    - 1 bit R/B per 1 channel (v2)
  - FPGA usage
    - Microblaze system : 6%
    - NAND Controller : 2% per channel
Register Map

- **Register**
  - NAND interface에 System의 커맨드를 전달

- **Address Register**
  - Using Physical Address: FTL에서 직접 입력
  - Table size overhead 없음
  - Physical address를 통해 page 정확한 상태 정보 FTL에 전달

- **Command Register**
  - Chip selection, Command data, NAND interface 상태 정보 저장

- **Status Register**
  - 모든 chip의 Ready/Busy 및 커맨드 성공/실패 상태 확인
Data Flow

• NAND Controller Data Flow

Host Interface Queue
- In DRAM
- priority Queue

Data Copy
- DRAM <-> Read/Write buffer

MPMC

ECC

Shared Read/Write buffer
- In BRAM
- 8-32 page size (64~256KB)
- To support multi channel

CPU (DMA)

Interleaving Scheduler
- Command & Address Register
- Register decoder
- Control NAND interface

Interleaving scheduler

NAND Interface

NAND

NAND

PLB
FPGA usage

FPGA usage of NAND Controller

FPGA usage limit (60%)

Reserved space
FPGA Usage

NAND Interface only
Microblaze system only
1ch 4way
4ch 4way
14ch 8way (estimate)

System
Interface
Buffer
ECC

1ch-4way
4ch-4way
14ch-8way

Ratio
3%
12%
3%
1.6%
35%
11%
2.3%
19%
57%
24%
81%
51%
11%
NAND Status Monitor

**Motivation**
- Insufficient user I/O pins
- Pin counts without R/B

- 644 pins → 420 pins

**NAND status monitor (NSM)**
- Use only 1 R/B or even without R/B
- Update NAND status in near real time

**Idea**
- Throw read status command automatically – polling strategy
- Save past operation time in table
Real-time NAND Status Monitor

- **R/B Signal per channel:** 1 R/B signal per 1 channel (Platform v2)
- **Channel interleaving**
  - Initial read status point: 1 Bank 상황에서 측정한 operation time OR Typical Time
  - 2개 이상의 Bank가 Busy인 경우는 R/B signal이 없을 때와 동일하게 동작

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully depend on R/B</td>
<td>Ideal case</td>
<td>많은 수의 IO pin 필요</td>
</tr>
<tr>
<td>NSM with No R/B (using timer)</td>
<td>IO pin 필요 없음</td>
<td>여러 개의 bank가 동시에 busy일 경우 정확한 완료 타이밍 알기 어려움</td>
</tr>
<tr>
<td>NSM with 1 R/B</td>
<td>1개의 R/B로 정확한 Initial Value 설정 가능</td>
<td>Pin count가 ch 당 1개 늘어 남</td>
</tr>
<tr>
<td>Time multiplexing R/B pins</td>
<td>채널당핀 수: log(ways) NAND가 느린 특성 이용</td>
<td>MUX를 FPGA 외부에 장착 보드 제작 후 수정 어려움</td>
</tr>
</tbody>
</table>
Real-time NAND Status Monitor

No R/B
Only depend on timer

Indilinx Barefoot
- 2 chip마다 R/B
- 2 chip 묶어 사용

Full R/B pin per
every single chip

NAND Status Management
- 목표:
  채널 당 하나의 R/B으로 full R/B와 근접한 효과
- Platform v2에 적용

Advanced Computing Laboratory
http://acl.snu.ac.kr
Core Features

- ONFI Interface
  - ONFI 2.2 Interface 구현 - using DQS(Data Strobe)
  - ONFI 2.2 최대 interface 속도 지원(100 MHz, Legacy는 33.3MHz)
  - DDR protocol 적용 (Interface 속도 2배 향상)

- Backward compatibility
  - Legacy NAND 와 ONFI NAND 동시 사용
  - Legacy – Asynchronous mode
  - Synchronous mode (ONFI 2.2)

- No R/B pins
  - NAND Status Management 모듈 사용

- Implementation
  - Verilog RTL
  - Implementation & debugging in Platform v1
## Legacy vs ONFi 2.2

<table>
<thead>
<tr>
<th>Model</th>
<th>K9LCG08U1M</th>
<th>MT29F128G08CECABH1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Legacy</td>
<td>ONFi2.2 (Synchronous mode)</td>
</tr>
<tr>
<td>Test capacity</td>
<td>16GB</td>
<td>32GB</td>
</tr>
<tr>
<td>사용 BD</td>
<td>Original Platform 1</td>
<td>Revised Platform 1 for ONFI</td>
</tr>
<tr>
<td>Interface Speed</td>
<td>33.3 MB/s</td>
<td>200 MB/s</td>
</tr>
<tr>
<td>Write Speed</td>
<td>3.47 MB/s</td>
<td>6.4 MB/s</td>
</tr>
<tr>
<td>Read Speed</td>
<td>25.9 MB/s</td>
<td>73.4 MB/s</td>
</tr>
</tbody>
</table>

- Legacy & ONFi NAND Interface 속도 최적화

![Graph showing Legacy vs ONFi performance](chart.png)
Experiments

- Maximum performance test
- FTL
- ECC
- NSM TEST
• Verifying Maximum Capacity
  – Capacity: 14 channel 8way -> 1.792 GB
  – FPGA elements 사용량 60% 이하 조건
  – 1.792 GB 전체 읽기/쓰기/지우기 정상 동작 확인

• Maximum throughput test
  – Throughput: 순차/랜덤 읽기/쓰기 속도
  – Bank(8GB)의 개수: 14 x 16 = 224
  – 하나의 Bank 최대 쓰기 속도: 6.4MB/s
  – ASIC 컨트롤러 쓰기 속도 예상값: 1433.6MB/s
  – FPGA 예상 최대 쓰기 속도 예상값: 약 150MB/s
• FTL
  – Page mapping
  – Block mapping
  – BAST
  – LAST
  – User custom FTL
  – 속도, 수명, 오버헤드 비교

![Bar chart comparing Page mapping, BAST, and LAST]

**Page mapping 대비 FTL 성능 비교(예상도)**

- 쓰기 속도
- 읽기 속도
- 수명
- 오버헤드
Architecture exploration

- **Architecture exploration**
  - # of channels: 1, 2, 4, 8, 14
  - # of ways: 1, 2, 4, 8
  - Shared buffer
    - Various size & sharing policy

- **Measuring**
  - Channel, way 변화에 따른 throughput
  - Power consumption
  - channel-way trade off
  - ECC 크기 변화에 따른 error 변화

SSD의 대부분의 Factor에 대해 측정 가능
Other experiments

- **ECC**
  - Basic option: 28bit BCH (parity 392 bits)
  - NAND Spare area 제약으로 (440 bits) 더 큰 BCH 적용은 힘듬
  - Other options
    - Small size BCH
    - LDPC
    - Architecting ECC module(s)
  - 결과
    - ECC 사용에 따른 오버헤드 (속도, 게이트 사용량) 비교
    - Error 비율과 오버헤드에 따른 trade off 분석

- **NAND Status Monitor**
  - No R/B vs 1bit R/B per 1 channel 성능 비교