Flash and the Architecture of Storage Systems

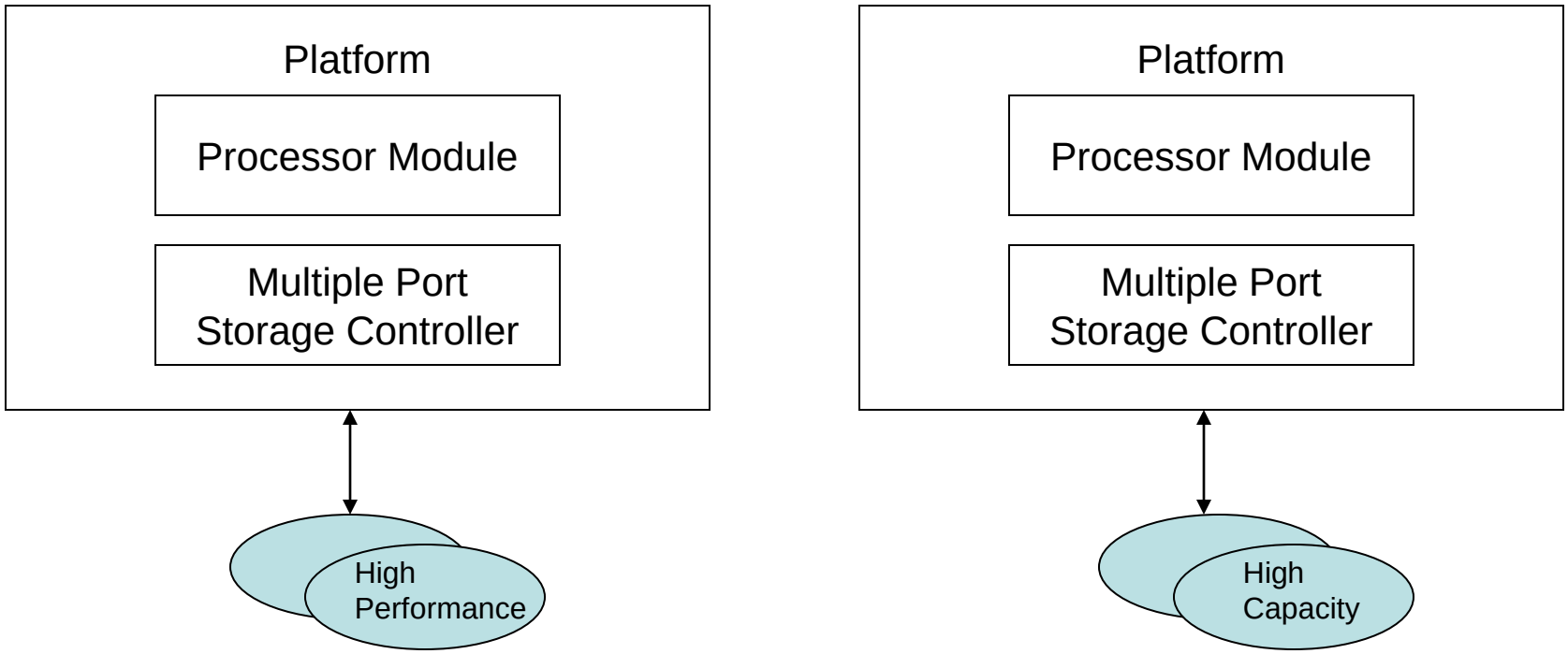
Moshe Selfin, Vice President Enterprise Solutions
Anobit

Storage Developer Conference
September 21st, 2010

Agenda

- SSDs in Midrange Storage Systems
- Typical High-End Storage Architecture
- SSD role in storage architecture
  - Tier 0
  - Machine Recommended Tiering
  - Auto Tiering
  - Caching (Read cache, write cache, combined caching)
- SSD additional contribution to storage management ideas
  - Meta-data storage
  - Static wear level
  - Hot/cold separation
- External/Internal Redundancies
- Conclusion

Midrange Storage System

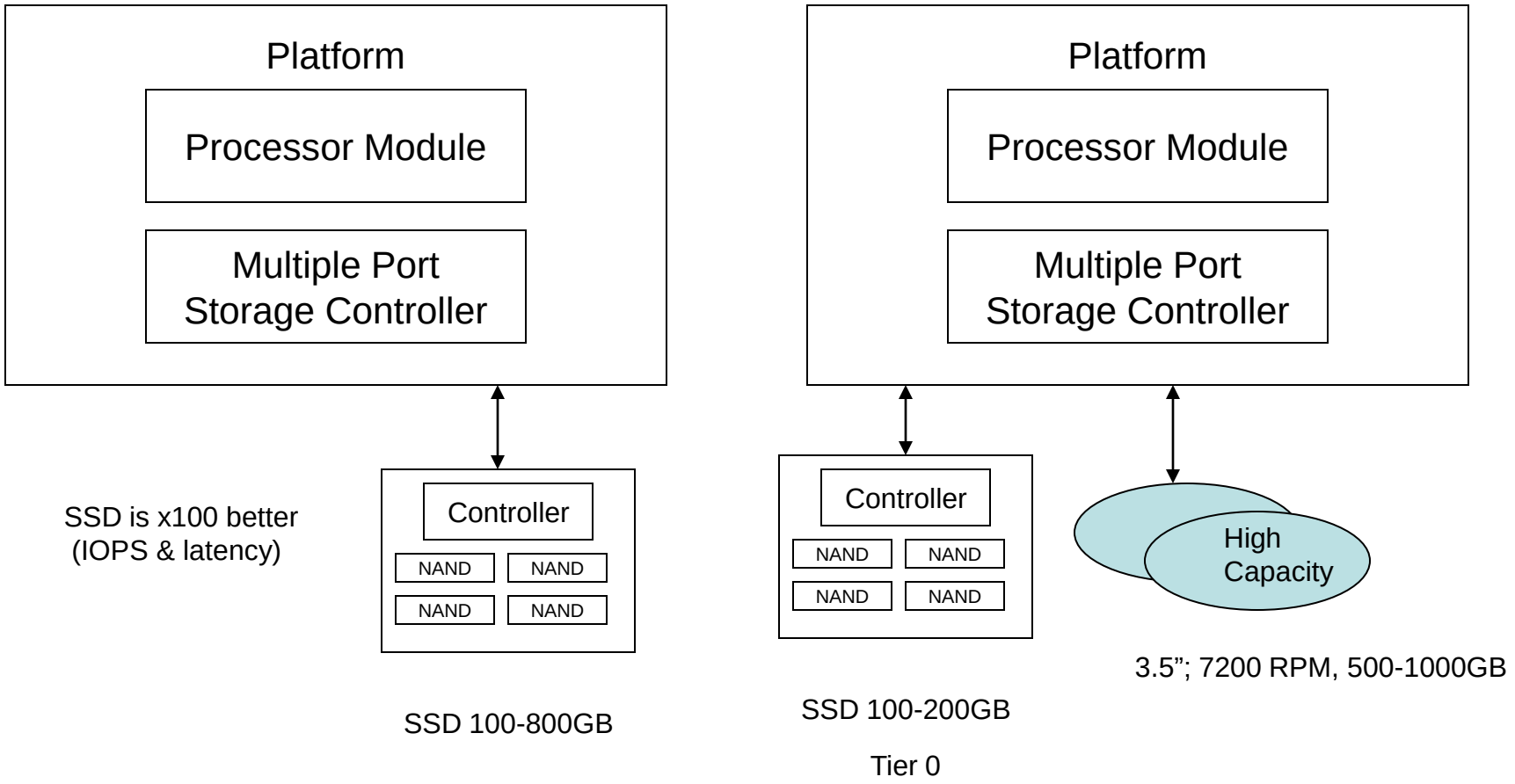


2.5”/3.5”; 15K RPM, 73-300GB

3.5”; 7200 RPM, 500-1000GB

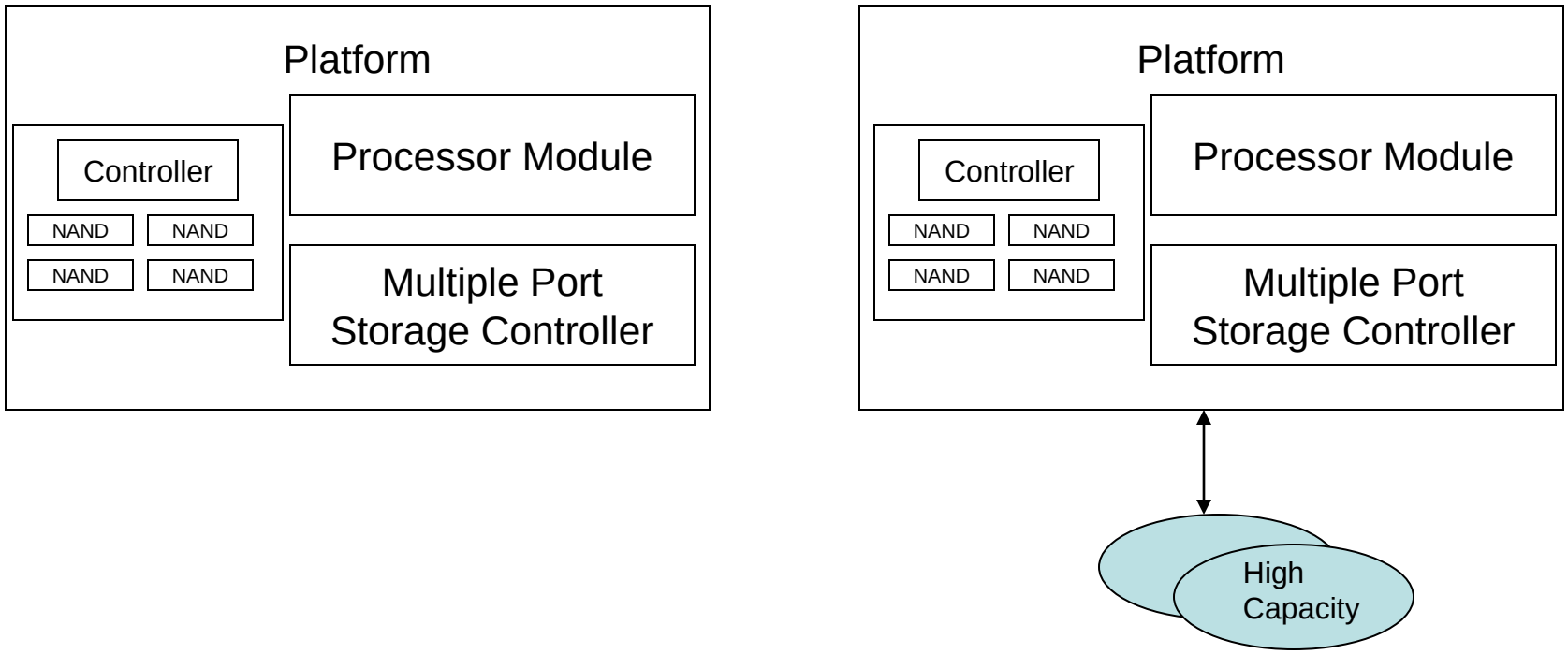
-

Midrange Storage System with SSD (1)



-

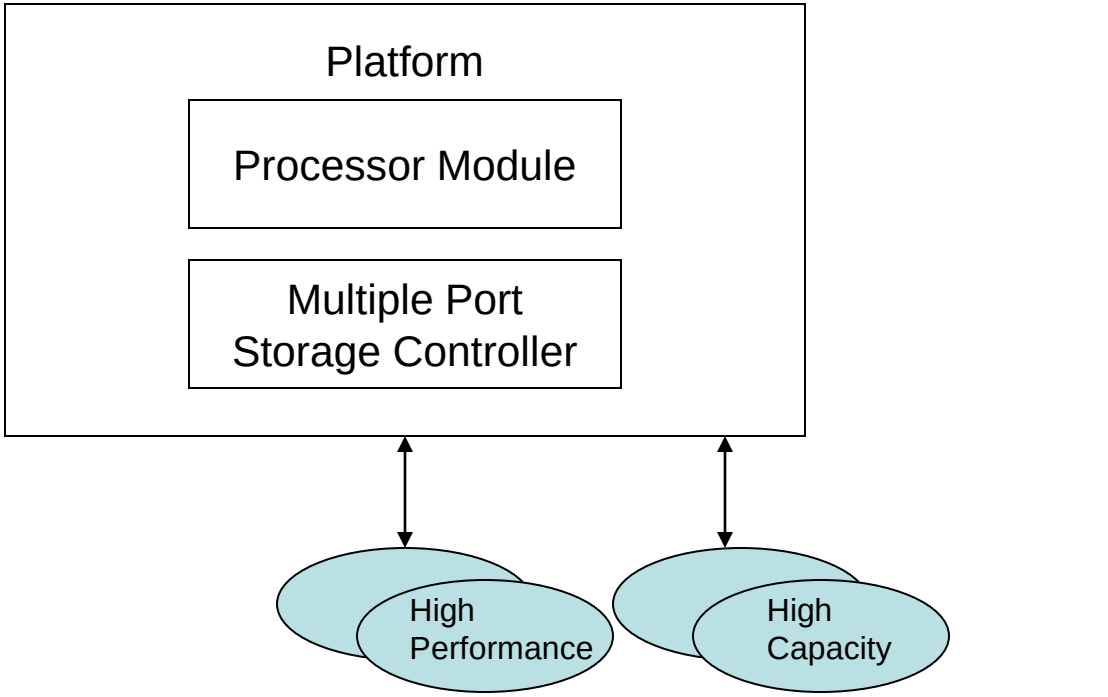
Midrange Storage System with SSD (2)



3.5”; 7200 RPM, 500-1000GB

-

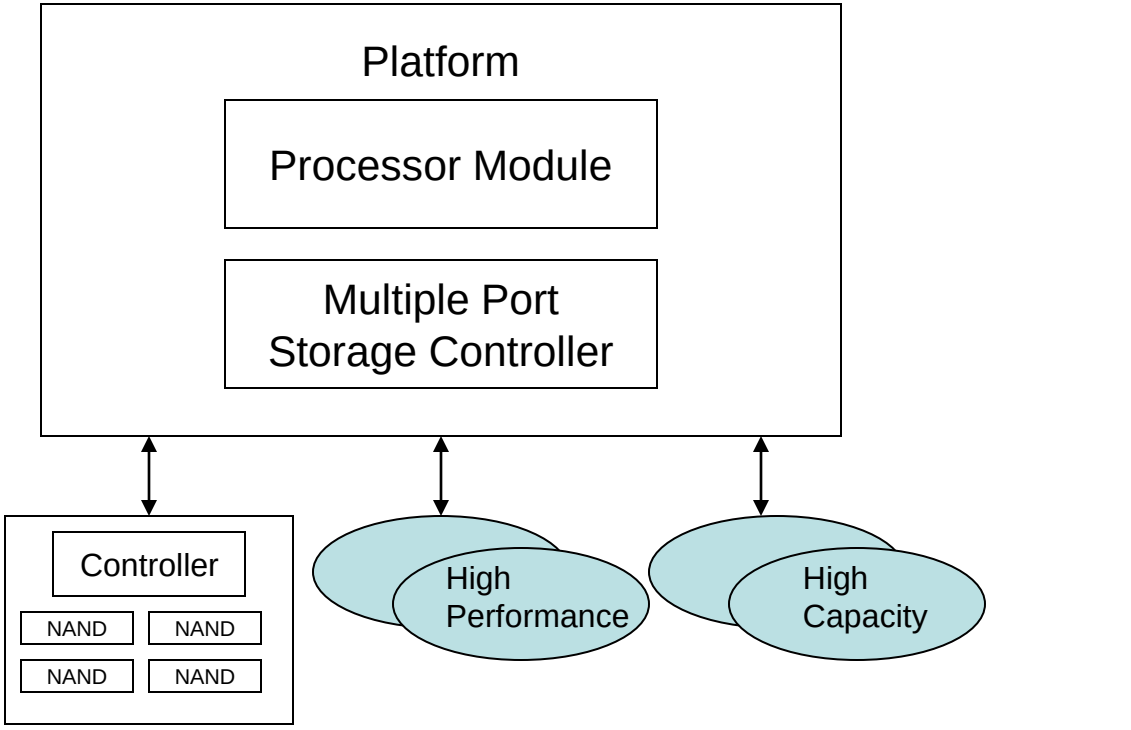
Typical High-End Storage System



-

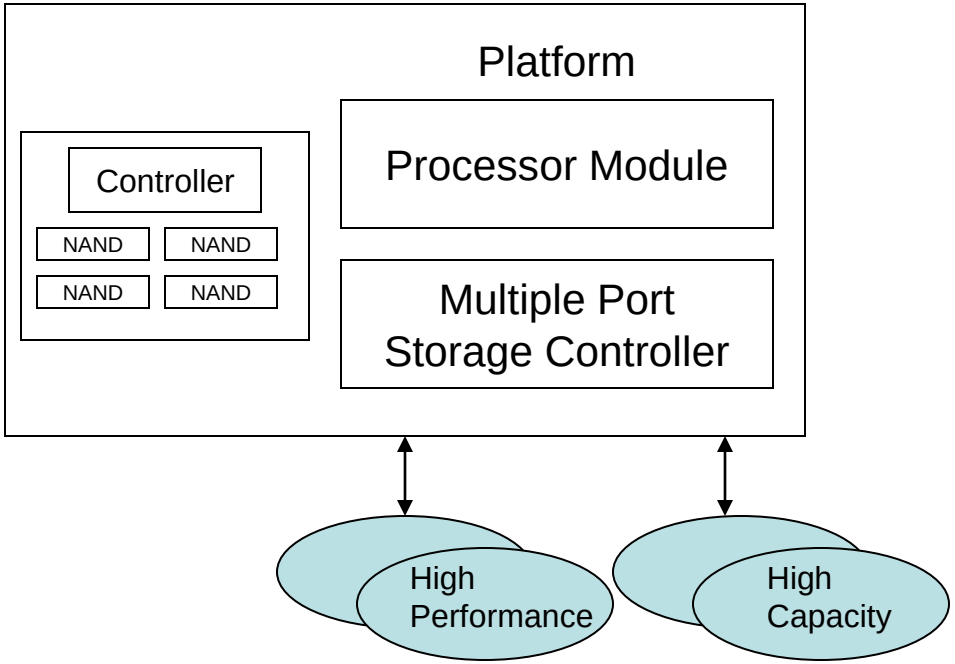
Typical High-End Storage System with SSD (1)

SSD is x100 better
(IOPS & latency)



-

Typical High-End Storage System with SSD (2)



15K RPM, 73-300GB

Tier 1 (On Line)

3.5”; 7200 RPM, 500-1000GB

Tier 1 (Near Line)

-

SSD Role as Cache

- General Caching Policy
  - Store most frequent used LBA/block in the fast memory
- Separating Caches
  - Read Cache
    - Store in SSD most frequent read LBA/block
    - In many cases such cache might need only 1-2 media cycles per day
    - Low cost memories can be used as reliability is not an issue (copy of the data is stored in the main storage)
    - No need for RAID → can be used in DAS
  - Write cache
    - Actually it is persistence storage for some limited time (few days)
    - Need to support high write IOPS and high endurance (10 cycles per day) → High end SSD should be used.
    - Must work with RAID to support "no single point of failure"

SSD Role in Storage System

- HDD replacement
  - Targets Midrange storage
  - Replace a dozen of performance rack HDDs with just a few SSDs
- Tier 0
  - SSD offer 100 times IOPS over tier 1 HDD
  - IT manager decides which applications will use the SSD
  - SSD should be operated under RAID in order to assure "no single point of failure"
  - Serviceability – SSDs should be in a serviceable form factor (e.g. Hot Swap)
- Machine Recommended Tiering
  - A machine runs statistics on the storage system
  - The machine recommends the IT manger which LBA should reside in the SSD External/Internal Redundancies
  - SSD should be operated under RAID in order to assure "no single point of failure"
- Auto Tiering
  - A machine runs statistics on the storage system
  - The machine decides automatically which LBA should reside in the SSD
  - SSD should be operated under RAID in order to assure "no single point of failure"
- Caching
  - Refer to previous slide
- Intermediate storage
  - SSD stores intermediate data (for example Business Intelligence Data). Doesn't have to be RAIDed
- Combinations
  - Part caching and part fixed tiering etc.

SSD Additional Contribution - Metadata

- Metadata
  - Tiering/caching management requires significant amount of metadata
  - The metadata can be:
    - Attached with the data itself
    - Centralized
    - Combined
- SSD Role
  - Use "large sector" (520/528 bytes) to store attached metadata
  - Provide fast-performance/low latency partition to store centralized metadata

SSD Additional Contribution - Hot/Cold

- Tiering/Caching management
  - Hot/cold – most of the decisions related to what should be cached/tiered are related to how frequent is the access to some LBA range
- SSD role
  - SSD has internal Hot/Cold mechanisms. For example static wear-level detects "cold" area and replaces those block with "highly used" blocks
  - SSD can report "LBA temperature" to the host in order to make the decisions related to caching/tiering easier

External/Internal Redundancies

- External SSD redundancy
  - Enterprise Reliability – One of the most important issues in enterprise applications is data reliability. As a result "no single point of failure" policy is used in storage systems
  - RAID (Redundant Array of Inexpensive Disks) – In order to avoid single point of failure, the disks are grouped and using simple ECC techniques someone can recover the whole data of failed disk
- Internal SSD redundancy
  - Description
    - Number of NAND dice – SSD naturally include large number of NAND dice (32-256)
    - SSD reliability will be significant low if the basic NAND die will not be excellent
    - Alternate solution – have spare dices
  - Spare die management
    - Automatic internal recovery – SSD manages internally ECC mechanism that enables it to recover the data when a die fails
    - Automatic external recovery – SSD has spare area, when a die fails, it uses the external RAID to recover the data and store it in the spare area

External/Internal Recovery

Feature	Internal Recovery[1]	External Recovery	Comments
Write Performance	Good	Better	Additional dice are used for over provision
Endurance	Good	Better	1. No need to write parity data (less data to program) 2. Additional dice are used for over provision (Lower write amplification)
SSD complexity	High	Fair	
Storage system complexity	Low	Fair[2]	

1. Drive using internal recovery must have sufficient reliability that overall probability for double fault ("Same LBA" on the RAID group will fail) is very very low.
2. No need to change the system design as upon each read failure the storage system will correct it. However it is recommended to enable the SSD to request for failed LBA at the moment it detects failure.

Conclusions

- **Tiering/Caching**
  SSD internal mechanism can provide additional value to the storage system

- **SSD Internal Redundancy**
  External recovery mechanism, due to it's superior performance, is highly recommended

About Anobit

- Founded in 2006
- Based in Israel
- Subsidiaries in the US and Korea
- 130 Employees

Thank You