



SSD
Weather
Channel

Client SSD:  Sunny

Server SSD:  Cloudy 

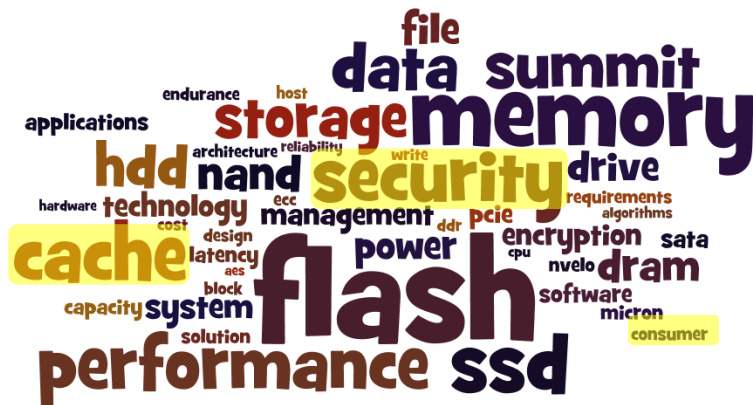
2012. Oct

S/W Development Team
Memory Division
SAMSUNG ELECTRONICS Co., LTD

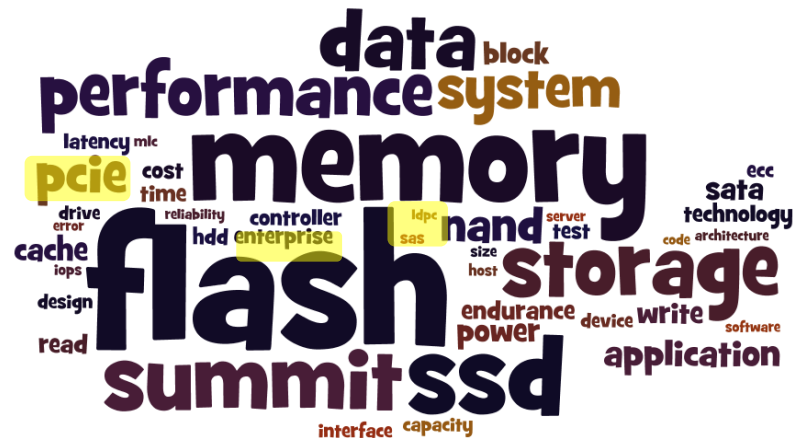


"Flash" the Server SSD Market

- 50 most frequent words in Flash Memory Summit 2010-2012



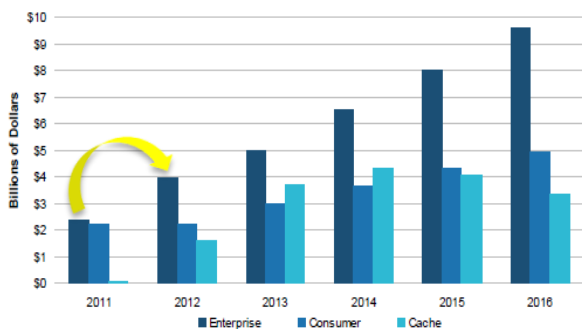
< Flash Memory Summit 2010 >



< Flash Memory Summit 2012 >

- Enterprise SSD surpassing client SSD in revenue by far

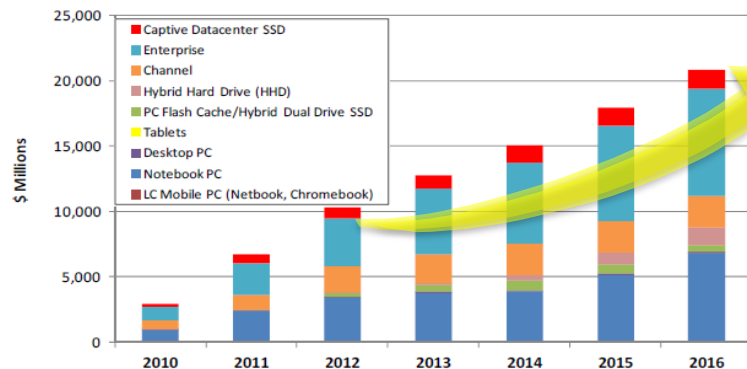
Figure 1: SSD Revenues by Segment, 2011-2016



Source: IHS iSuppli | April 2012

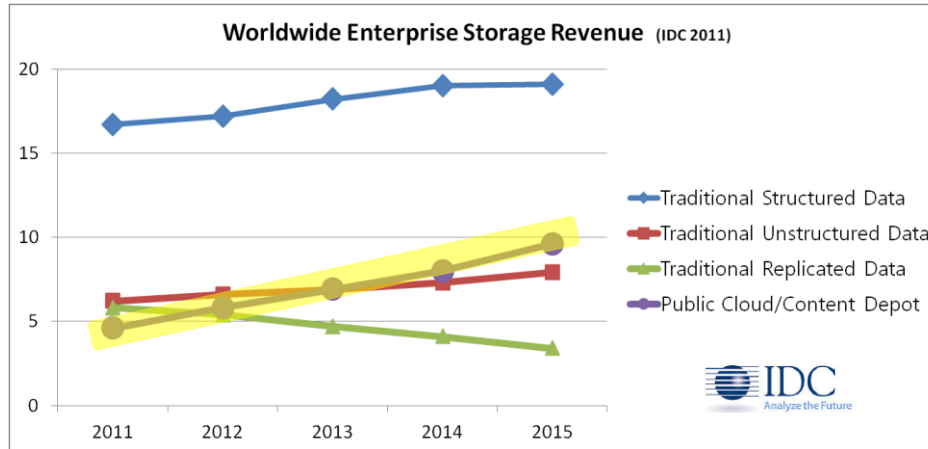


Computing SSD Revenues



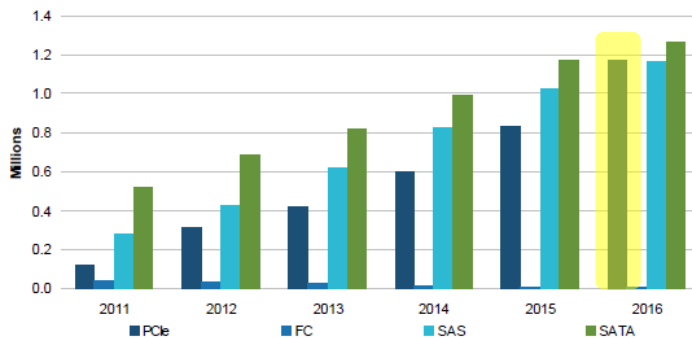
Market Trends of Enterprise Storage

- Public cloud/contents depot explodes by 58% (CAGR) in capacity

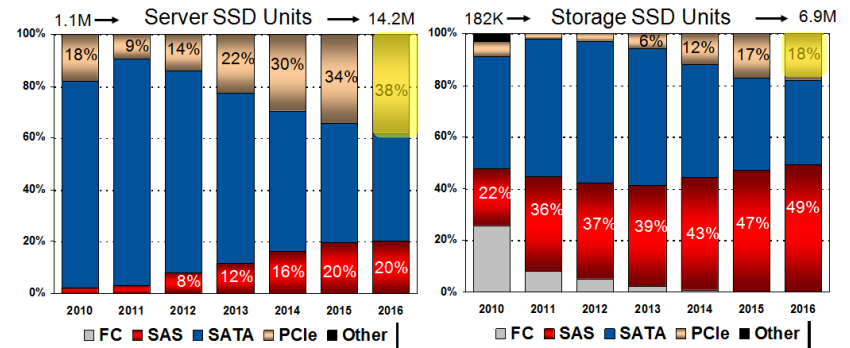


- PCI Express adoption grows up to 1/3 by 2016
 - SATA (server SSD) & SAS (storage SSD) still dominant in enterprise market

Figure 5: Enterprise SSDs by Interface, 2011-2016



Source: IHS iSuppli | April 2012

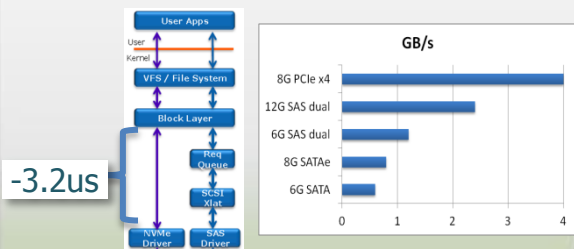


Graphs Show % of Enterprise SSD Shipments by Interface in Unit

Why PCI Express for Enterprise Storage?

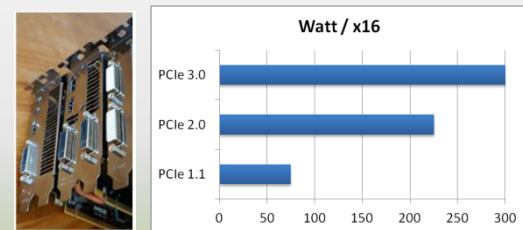
Low Latency + High Throughput

- I/O directly connected to CPU
- Gen3: 1GB/s/lane (scalable) w/ QD=64K

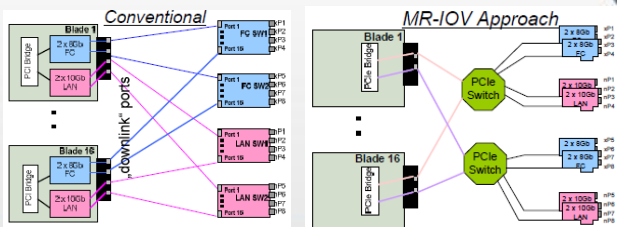


More Power Supply

- Gen3 x16: up to 300W (Graphics)
- 25W / SSD device + PCIe PM

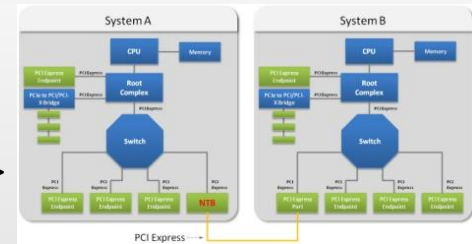


- VI O/H: Hypervisor → Adapter
- Devices shared by multiple hosts



IO Virtualization

- Non-Transparent-Bridge ports
- <\$200 adapter, <2us latency








< Clustering via NTB >

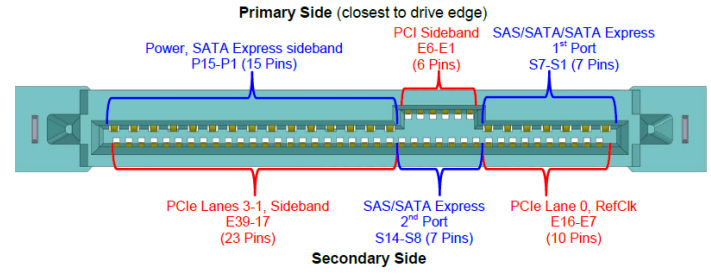
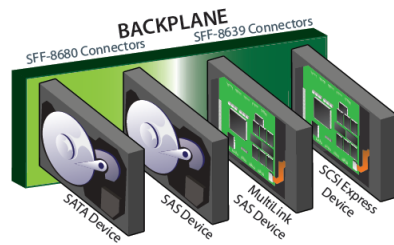
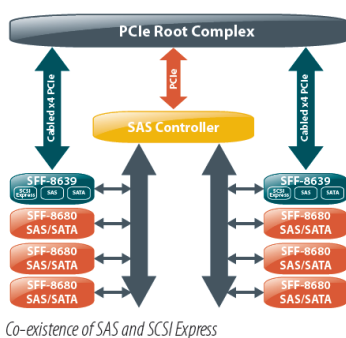
Cost Effective Clustering

Multiple Standards for PCIe SSD

- Three major standards over PCIe: SATAe → NVMe → SCSIe

			
Primary Target	Enterprise Server SSD	Enterprise Storage SSD	Client/Hybrid SSD
Command Interface	NVMe	SCSI (SOP & PQI)	ATA (AHCI) / NVMe
Form Factor	2.5" / SIOM 	2.5" / Edge Card 	2.5" / 1.8" / mSATA
Key Drivers	Intel, Dell	HP	Intel
Standardization	NVMe Group	T10 & STA	SATA-IO
First SSD Products	2012	2013	2013
Revision	v1.0d	Under standardization	SATA rev3.2




- Express Bay (SFF-8639) supports multiple standard devices (~13)



< 12G SAS / 4x 6G SAS / 4x PCIe / 6G SATA / SATAe >

Characteristics of Cloud SSD

- Cloud SSD has distinctive requirements on Endurance/Cost

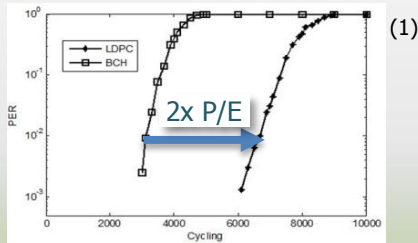
Category	Client SSD	Cloud SSD	Enterprise SSD
Price	Low	Low	High
Retention	>1 Year	<1 Month	>1 Year
P/E cycles	>3K (hard limit)	>10K (soft limit)	>30K
Power	Limited	Always On	Always On
Ta	<70°C	45~50°C	60~70°C
Capacity	IDEMA(2 ⁿ GB)	OverProvisioning	More OverProvisioning
Performance	Response Critical	Endurance Negotiable	Throughput Critical
Recovery	Meta-SPOR	Data-SPOR	Data-SPOR
Power backup	Ceramic cap (~us) 	Tantal cap (~ms) 	Super cap (~s) 

- To maximize endurance w/ low-cost solution → S/W & device engineering

H/W Ideas for Cloud SSD

Controller IP → Endurance ↑

- Enhanced ECC or LDPC
- Chip-level RAID



Component → Cost ↓

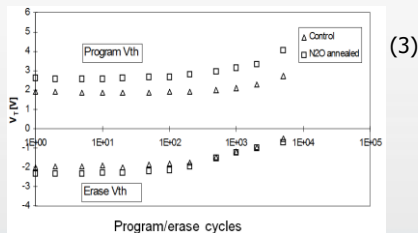
- MLC/TLC, not SLC/eMLC, even 3D?
- Tantal cap, not Supercap

	Gate First		Gate Last
	Toshiba/P-BICS	Hynix DC-SF	Samsung/TCAT
Type of 3D NAND			
Transistor	Gate all around; Salicided Poly Si gate	Gate all around; Salicided Poly Si gate	Gate all around; Damascene metal gate
Storage	Charge trap	Floating gate	Charge trap

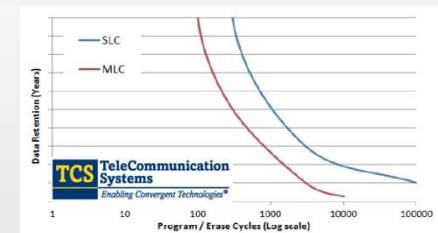
* Cloud SSD Priority

- Cost
- Endurance
- Performance

- Operating voltage control
- Lower temperature → more P/E



- Wear-level index
- Less retention → more P/E



Flash Operation → Endurance ↑

Flash Feature → Endurance ↑

(1) Xueqiang Wang, Flash Memories, ISBN: 978-953-307-272-2

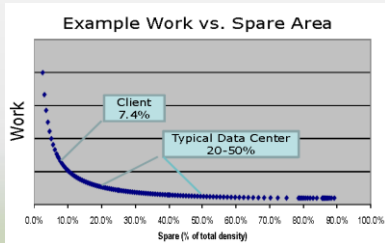
(2) Seung Suk Lee, Emerging Challenges in NAND Flash Technology, Hynix Semiconductor Inc., Flash Memory Summit 2011, Aug 2011

(3) Paolo Pavan, Flash Memory Cells – An Overview, Proceedings of the IEEE, Vol. 85, No. 8, Aug 1997

S/W Ideas for Cloud SSD

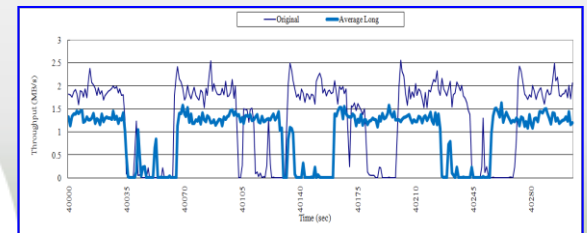
Capacity ↓ → Endurance ↑

- Over-provisioning: Client < 8%, Server > 28%
- SLC mode only: Lifetime multiplied



Performance ↓ → Endurance ↑

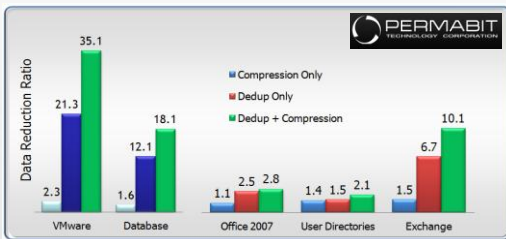
- Dynamic throttling by wear & temperature
- Recovery-period: > 3 days → 1/2 RawBER



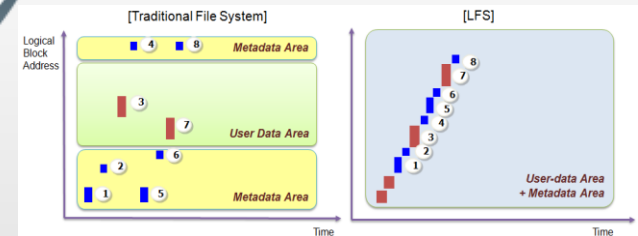
* Cloud SSD Priority

- Cost
- Endurance
- Performance

- De-duplication + Compression
- Hot/Cold Separation



- Log-structured filesystem
- Automatic storage tiering

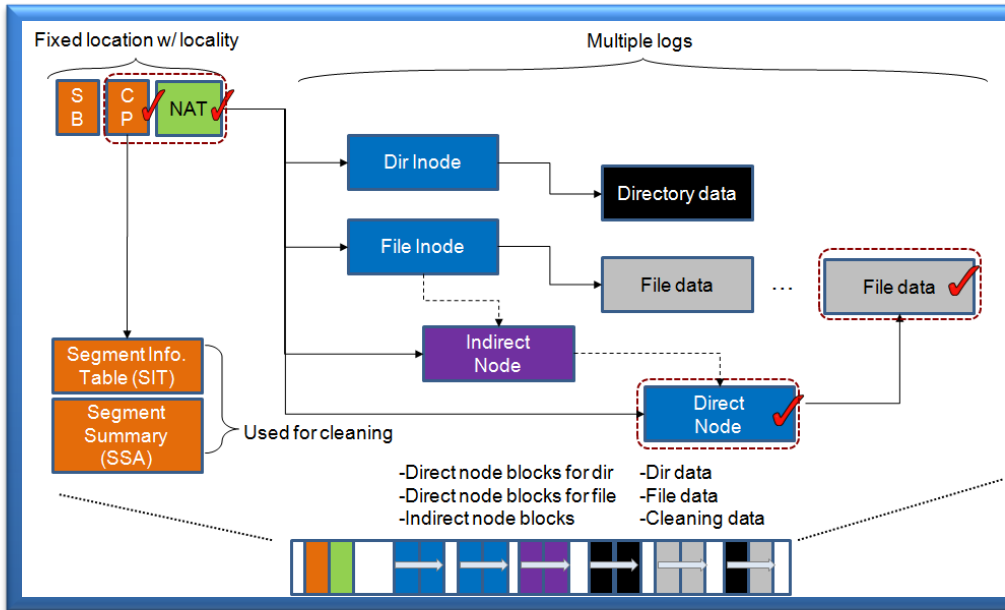


FW Algorithm → Endurance ↑

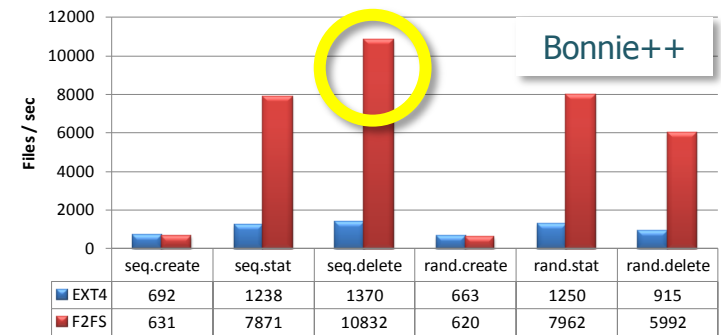
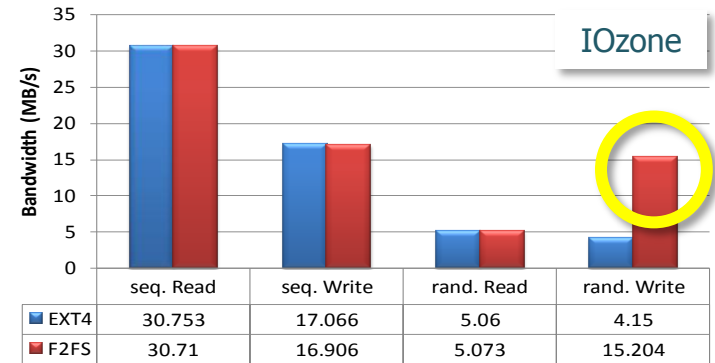
OS Algorithm → Endurance ↑


Flash-Friendly File System (F2FS)

- Samsung has released F2FS for flash storage to Linux open-source group
 - Wandering tree problem mitigated by NAT(Node Address Table)
 - Cleaning O/H reduced by background cleaning, hot/cold separation, adaptive logging
 - Can be configured by FTL-optimized parameters such as mapping unit
- ➔ Compared with ext4 in FS benchmarks, almost sequential-like random write performance



< F2FS Index Structure >



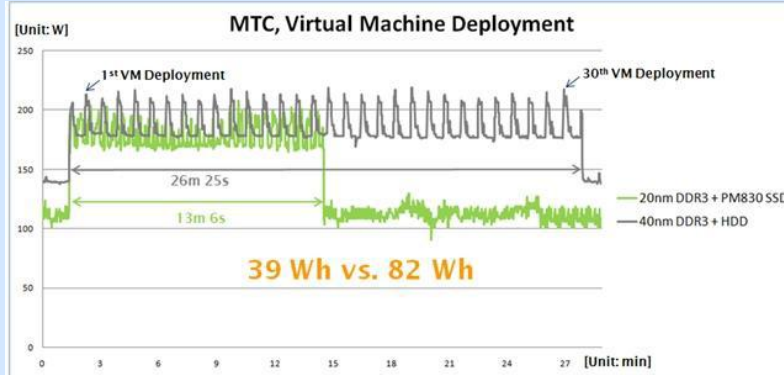
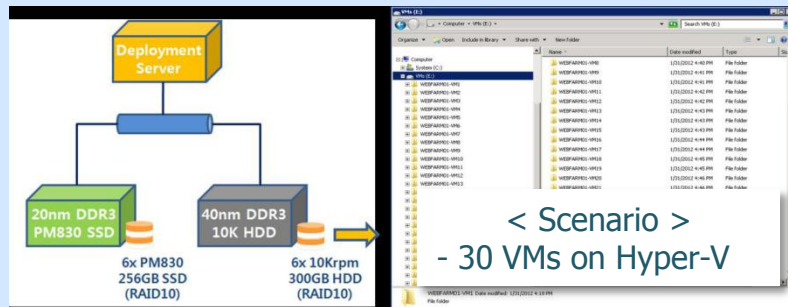

Greg Kroah-Hartman Oct 6, 2012 · Public
 Sweet, a new Linux file system from [Samsung](#) that is faster than existing ones when running on flash storage devices, submitted in a clean, easy-to-apply manner. This will be great for Android-based systems.

Cloud Implementation w/ Samsung SSD

- Joint experiments w/ Microsoft Technology Center (Feb '12) ⁽¹⁾
 - (DDR3 20nm 8GB + PM830 SATA SSD) vs (DDR3 40nm 8GB + 10Krpm SAS HDD)

Case1 – VM Deployment

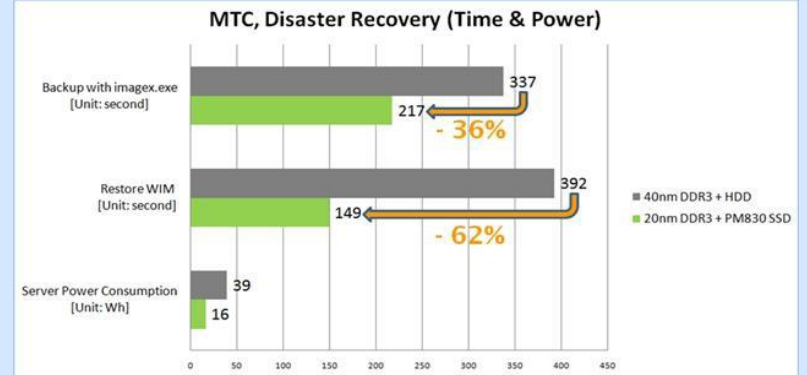
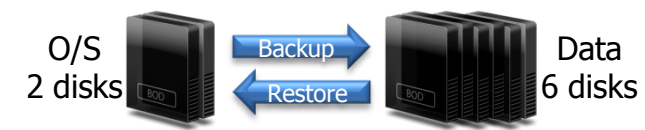
- 52% less power: 82Wh → 39Wh
- Completed 2 times faster: 26'25" → 13'6"



Case2 – Disaster Recovery

- 59% less power: 39Wh → 16Wh
- Stored 36% faster: 5'37" → 3'37"
- Recovered 62% faster: 6'32" → 2'29"

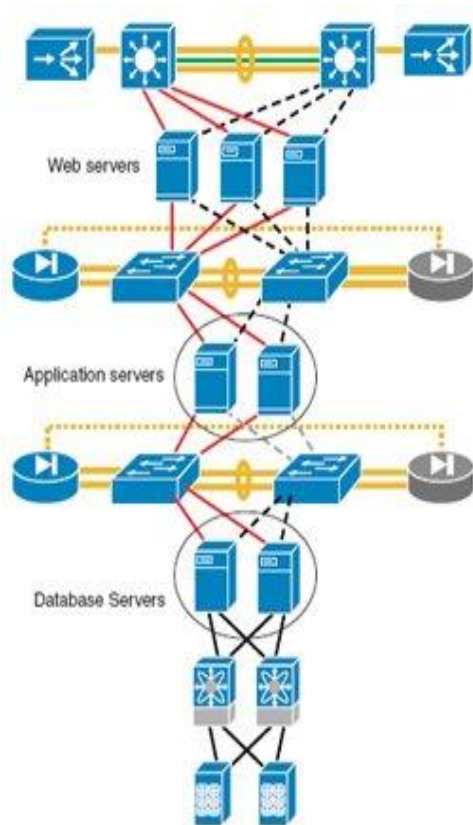
< Scenario >
- O/S image backup & restore w/ Microsoft ImageX



(1) Ramazan Can, Efficient Cloud Implementation, Microsoft Technology Center, Feb 2012

Tiered Data Center Architecture

- SSD solution: Front-end boot drives or back-end high-tier storage/cache
- Virtualization will increase SSD adoption even more in data center



< Source: Cisco Systems >

Category	CPU Load	DRAM Usage	Storage Load	Major Solution	Access Pattern	SSD Storage
Web Server	Low	Medium	Low	VM Web	Seq Write Ran Read	Boot MLC
Application Server	Medium	Medium	Low	VM App (WAS)	Seq Write Seq Read Ran Read	Boot MLC
Database Server	High	High	Medium	HPC Cache DBMS	Seq Write Ran Read	SATA eMLC (PCIe)
Storage Server	Low	Medium	High	Tiering	Ran Write Ran Read	SAS SLC/ eMLC

Flash Storage Category by Location

- Tier-1 storage is being replaced by hot-pluggable all-flash array
- Cache S/W becomes more important in tiered/virtualized storage systems

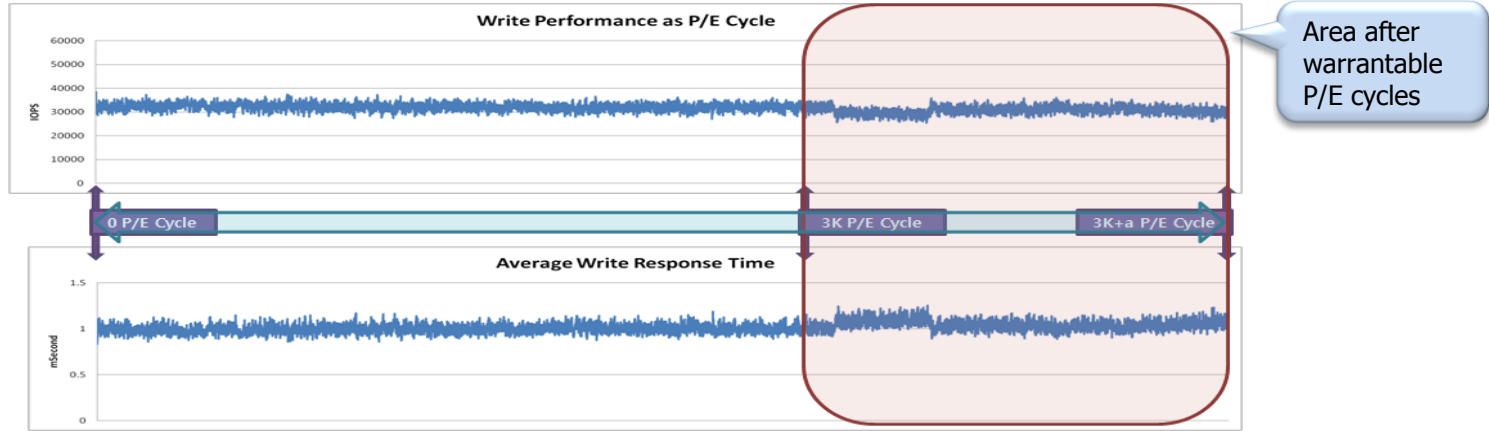


Category	Direct Attached	Host-Based Caching	Array-Based Caching	Array-Based Tiering	All SSD Array
Vendor -Product	<ul style="list-style-type: none"> •FusionIO -ioDrive •Virident -FlashMAX •LSI -WarpDrive 	<ul style="list-style-type: none"> •FusionIO -ioTurbine •Adaptec -MaxCache •Marvell -DragonFly 	<ul style="list-style-type: none"> •NetApp -FlashCache •EMC -FASTCache 	<ul style="list-style-type: none"> •HP -3PAR •EMC -Compellent 	<ul style="list-style-type: none"> •ViolinMemory -3000/6000 series •SkyEra -SkyHawk
Pros	•Best performance	•Low latency	•Good for hot data	<ul style="list-style-type: none"> •Automatic tiering •Capacity+availability 	<ul style="list-style-type: none"> •Best IOPS/\$ •Less space/power
Cons	<ul style="list-style-type: none"> •Worst cost •Limited capacity •No HA features 	<ul style="list-style-type: none"> •Data integrity •More irregular performance 	<ul style="list-style-type: none"> •Worse endurance •Irregular performance 	<ul style="list-style-type: none"> •Endurance issue •Performance O/H •Complex architecture 	<ul style="list-style-type: none"> •Low GB/\$ •Cloud scaling

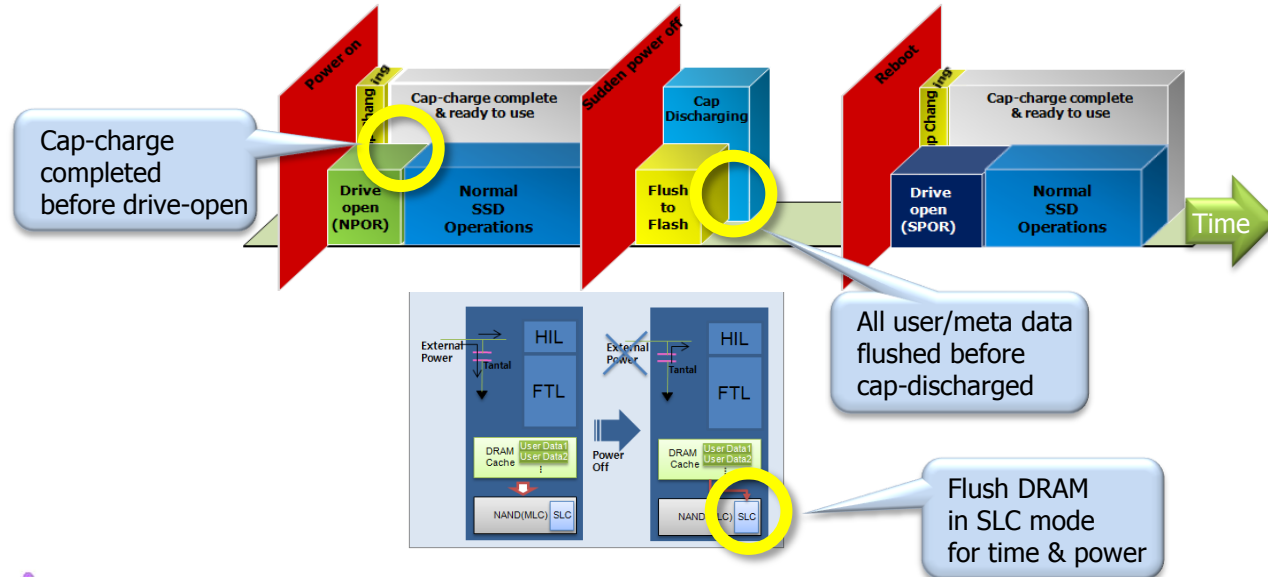
Samsung SSD for Data Center



- Consistent performance and response time is another key feature (QoS)

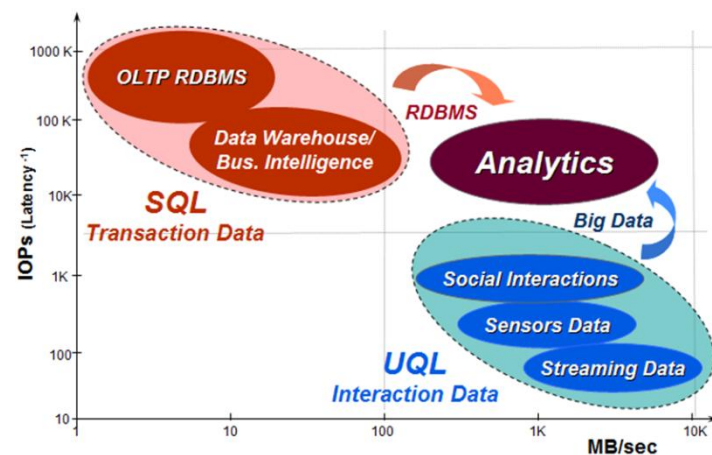
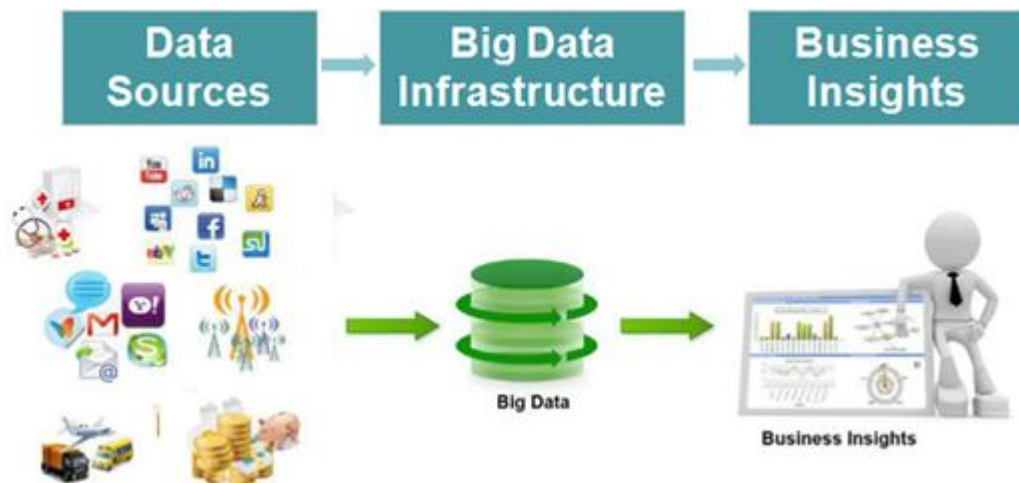


- All transferred data is protected by new F/W algorithm even at power loss



Big Data Infrastructure Galore

- Huge & complex data sets, impossible to process on traditional DBMS
- Big data analytics will need real-time distributed storage systems



Information is at the center of New Wave of opportunity

44x

as much Data and Content Over Coming Decade

2020
35 zettabytes

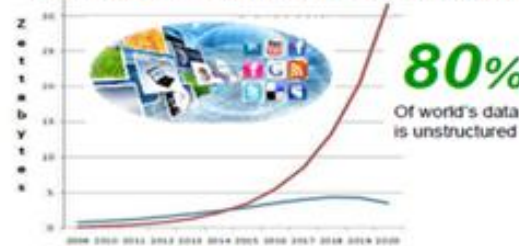
2009
800,000 petabytes

Velocity
Variety
Volume

Majority of data growth is being driven by unstructured data and billions of large objects



80% of world's data is unstructured driven by rise in Mobility devices, collaboration machine generated data.



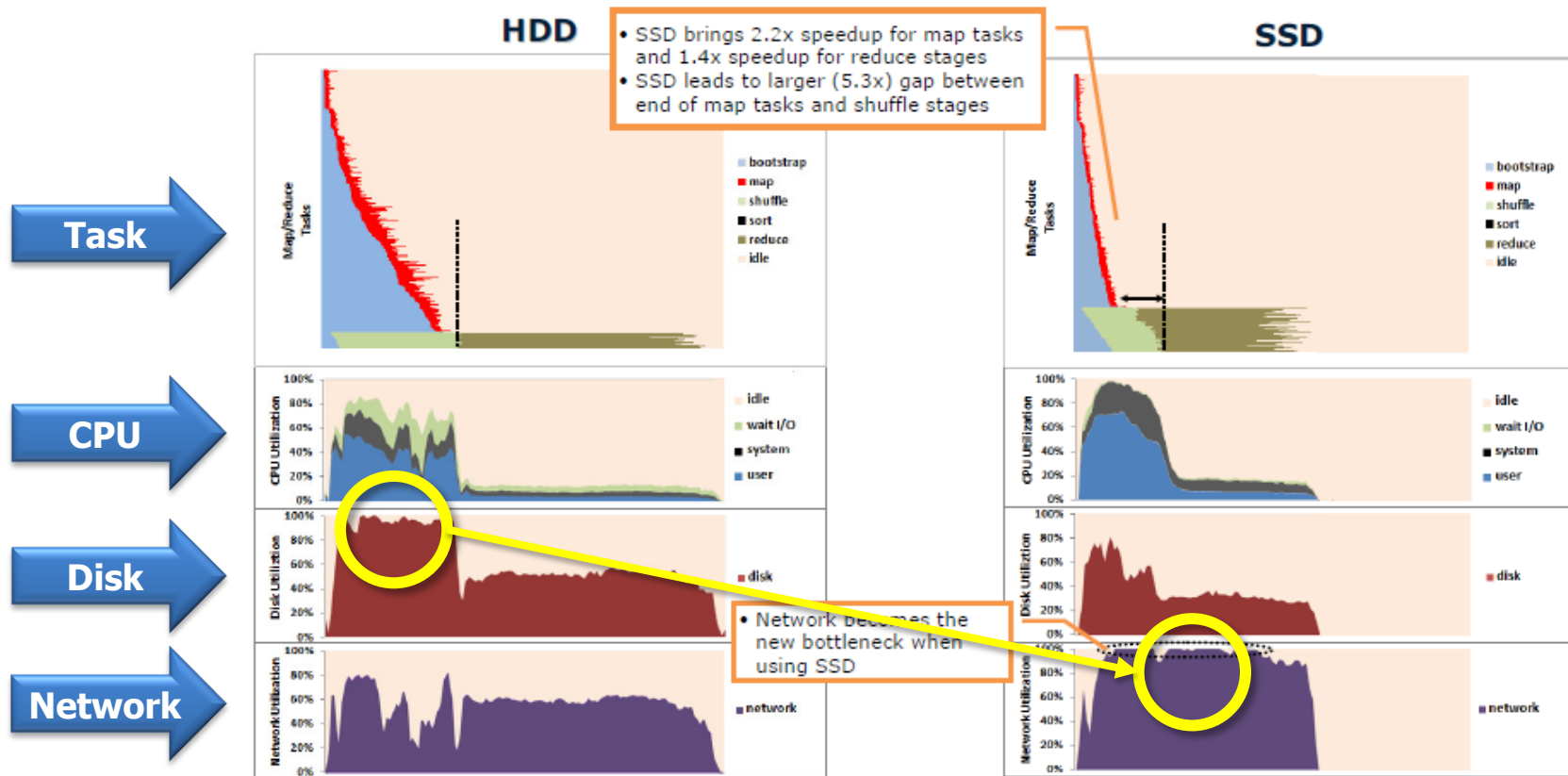
IMEX
RESEARCH.COM

SAMSUNG

Bottlenecks in Big Data Processing

- Balanced I/O subsystem (H/W + S/W) is critical in big data processing
- With SSD deployment, network infra & S/W stack should evolve as well

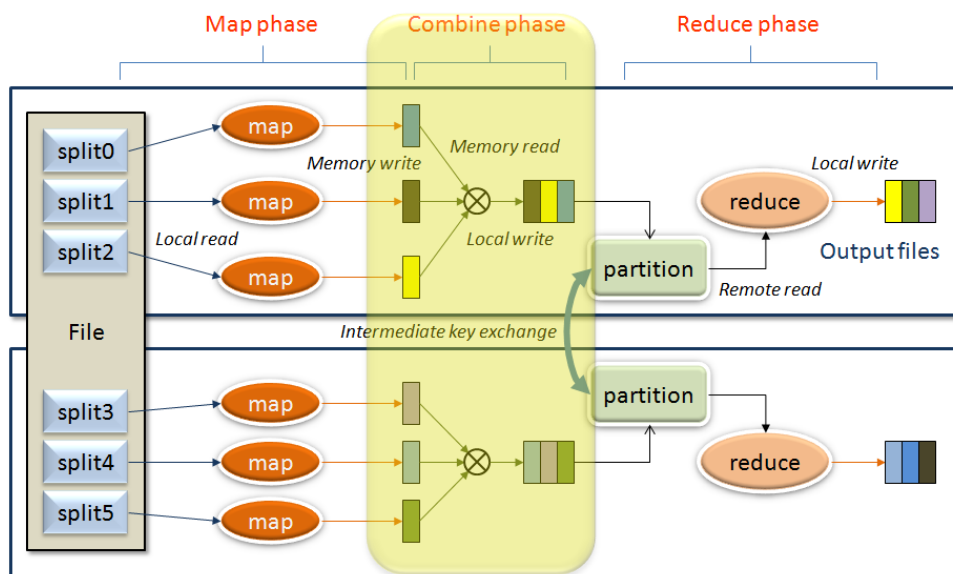
HDD vs. SSD for Hadoop Sort



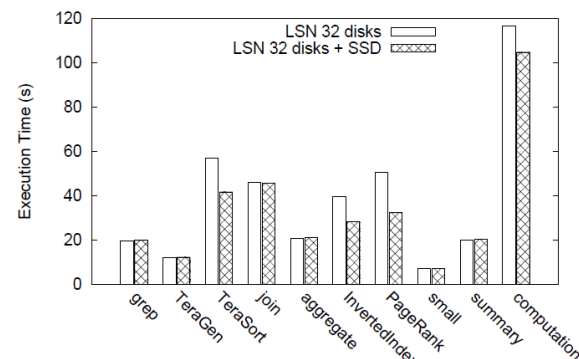
* SOURCE : Jinquan Dai (Intel), "Performance, Utilization and Power Characterization of Hadoop Clusters using HiBench", Hadoop in China 2010

SSD for Hadoop System

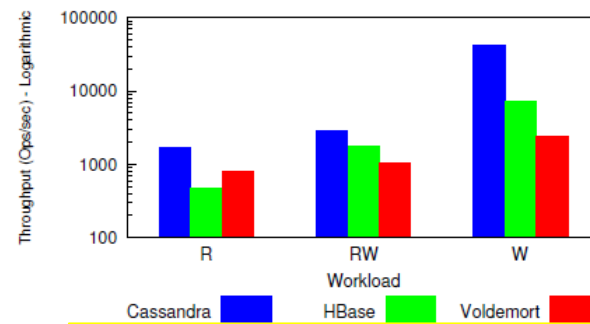
- Hadoop performance issues still being improved (ex) Hadoop-2.0, CDH, Cassandra
 - No failover mechanism, low parallelism, imbalanced namenode, FIFO scheduling, etc.
- Shuffle/merge phase generates intensive random writes → SSD preferable (1)
- Optimization in other Hadoop layers can give more chances to SSD (2)
 - (ex) Cassandra – no locking, log-structured, highly parallel, compression



< Hadoop Map/Reduce >



< (1) LSN performance w/ Hadoop+SSD >



< (2) Cassandra DB Throughput >

(1) Guanying Wang, Evaluating MapReduce System Performance, Ph.D thesis, Virginia Tech.

(2) Tilmann Rabl, Solving Big Data Challenges for Enterprise Application Performance Management, Proceedings of VLDB Endowment, Vol.5, No.12, Aug 2012



- Big data / Cloud computing is opening paradigm shift in Flash Storage, *but system-level optimization leaves a lot to be desired.*
- Enterprise storage I/F is converging on PCIe, *but storage industry is still much based on customized tiered architecture.*
- Cloud storage has distinctive requirements: Cost > Endurance > Latency, *but cost-endurance trade-off delays adoption of all-flash storage.*

Align with your imagination



Thank you

