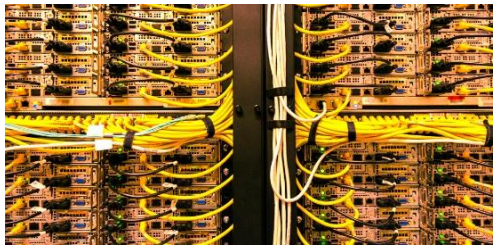
Designing SSDs for large scale cloud workloads

Kushagra Vaid
GM, Server Engineering, Microsoft

FLASH MEMORY SUMMIT, AUG 2014

Microsoft

5.8+ billion
worldwide queries each month.
bing



250+ million
active users.
SkyDrive

400+ million
active accounts.
Outlook.com

2.4+ million
emails per day.
Microsoft® Exchange Hosted Services

8.6+ trillion objects
in Windows Azure storage.
Windows Azure

48+ million
users in 41 markets.
XBOX LIVE

50+ million
active users.
Office 365

1 in 4
Enterprise customers.
Office

50+ billion
minutes of connections handled each day.
skype

200+ Cloud Services @ Microsoft

1+ billion customers. **20+ million** businesses. **90+** markets worldwide.

Cloud workloads are different!

Examples:

- Read-mostly, write-once per day
- Sequential write streams for object stores
- Synchronous replication for data durability ; No RAID
- Flash used as intermediate buffer pool between DRAM/HDD

SSD scale implications: Design

	Client	Enterprise	Cloud
Performance IOPS	Medium	High	Medium
Duty cycle	<100%	100%	100%
NAND endurance	0.5-3K P/E	30-40K	15-20K
Data retention	12 months	3	<= 1
Power loss protection	No	Yes	App-specific

SSD scale implications: Operations

	Client	Enterprise	Cloud
Data logging	Minimal	Differs across vendors	Rich and Consistent
Error handling	None	“Brick” or sometimes resume	Always Resume
Failure Recovery	None	Skilled technician	Automated, remote recovery
Security	None	Secure erase	Secure erase

Microsoft SSD reference design

- Built on proven commercial ASIC
- SATA interface
- eMLC NAND
- PFAIL functionality implemented
- Rich set of instrumentation/counters

Exploring NAND tradeoffs for Cloud

Abbr.	Parameter	Changes	Endurance	Performance	Power
C_R	Code Rate	Fixed → Variable	↑	Targeted Overheads	←→
$t_{ret.}$	Data Retention	Decrease	↑	←→	←→
ECC_{Type}	Code Style	Hard → Hard + Soft	↑	←→	←→
V_{TH}	Read Thresholds	Dynamic, Targeted Read Parameters	↑	←→	←→
t_{dwell}	Dwell Time	Increase	↑	↓	←→
$V_{prog.}$	Write Voltages	Decrease	↑	↓	↓

NAND tuning on Microsoft SSD

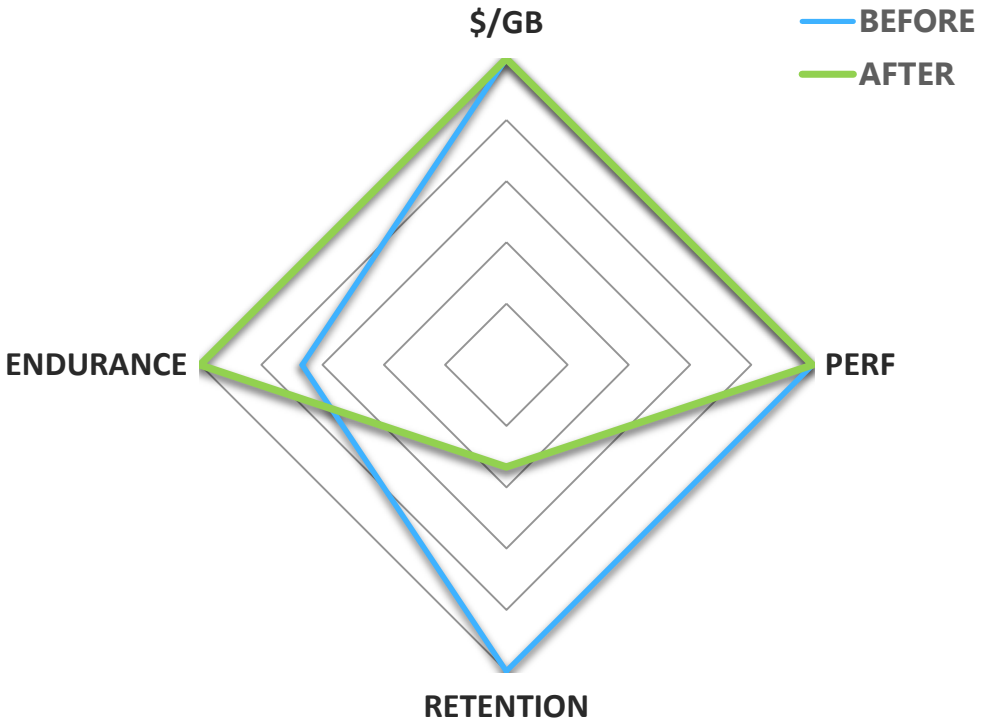
Abbr.	Parameter	Changes
C_R	Code Rate	24b → 29b
$t_{ret.}$	Data Retention	Decrease (3 → 1 mo)
ECC_{Type}	Code Style	Hard → Hard + Soft
V_{TH}	Read Thresholds	Dynamic, Targeted Read Parameters
t_{dwell}	Dwell Time	Increase
$V_{prog.}$	Write Voltages	Decrease

Characterize NAND for 1 month retention

Change firmware for increased ECC coverage (24b→29b)

What is the corresponding endurance improvement?

NAND parameter tuning results



50% endurance improvement with no change in product cost, performance or power

Exploring NAND parameter tradeoffs

Abbr.	Parameter	Changes	Endurance impact	Performance impact	Power impact
C_R	Code Rate	Fixed → Variable	↑	Targeted Overheads	↔
$t_{ret.}$	Data Retention	Decrease	↑	↔	↔
ECC_{Type}	Code Style	Hard → Hard + Soft	↑	↔	↔
V_{TH}	Read Thresholds	Dynamic, Targeted Read Parameters	↑	↔	↔
t_{dwell}	Dwell Time	Increase	↑	↓	↔
$V_{prog.}$	Write Voltages	Decrease	↑	↓	↓

Unlock the possibilities:

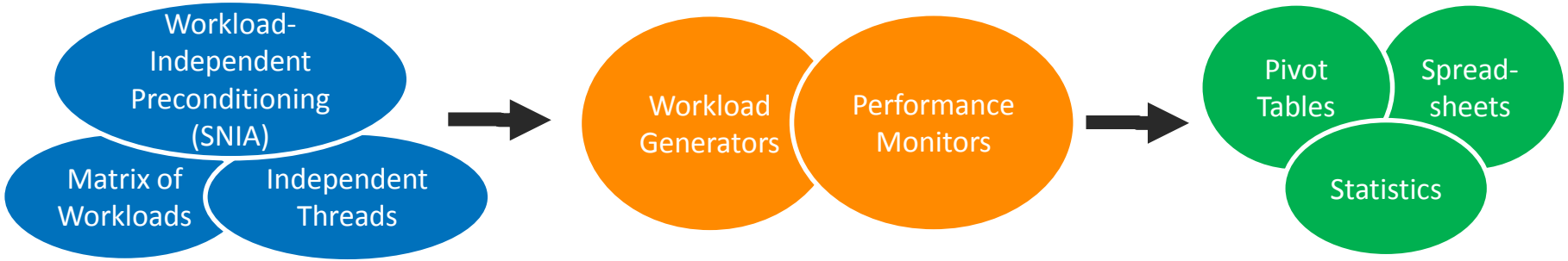
Research needed in industry and academia!

Goals for benchmarking SSDs at-scale

- Scalable: #SSDs x #Workloads x #Metrics
- Scriptable and Automated: Quickly construct and execute tests
- Simplified results data analysis and customized scoring
- Consistency in quantifying workload-specific endurance

Microsoft's SSD analysis tool - StorScore

Automated **== *Minimal Engineering Time*** ***Scripted*** **== *Quick & Easy to Modify***



Extensively used at Microsoft to drive SSD selection for Cloud workloads

StorScore benefits

Automatic pre-conditioning and push-button ease allow large scale parallel testing cutting down evaluation times

	BEFORE	AFTER
Test development effort	Tedious & Manual	Automated & Productive
Endurance qualification (1 TB example)	1 workload per SSD in 5 months	1 workload per SSD in 2 hours
Data analysis for scale tests	Unmanageable	Structured data with consumable reports

Contributing StorStore to the community

Available now for *free download*

http://aka.ms/storscore

For additional technical information, attend afternoon session

Presenters: Laura Caulfield and Mark Santaniello, Microsoft

Title: StorScore – Microsoft's System for SSD Qualification

Time: 2:40-3:55pm

Location: Ballroom G, Session 303-F (Testing Challenges)

Summary

Solution-level thinking - deeper understanding of cloud workloads requirements and datacenter environments

- Collaboration between End-user, Flash controller and NAND manufacturer

Workload-driven optimizations for APIs, FTL and NAND

- NVMe offers ideal canvas for driving next generation of innovations
- Non-linear improvements in endurance and device life

Microsoft's scalable test framework now available openly for standardized performance/endurance evaluation

THANK YOU

Microsoft