Opportunities and Challenges of Using Solid State Drives in Large Scale Datacenters

Badriddine M. Khessib
Dileep Bhandarkar
Microsoft Corporation

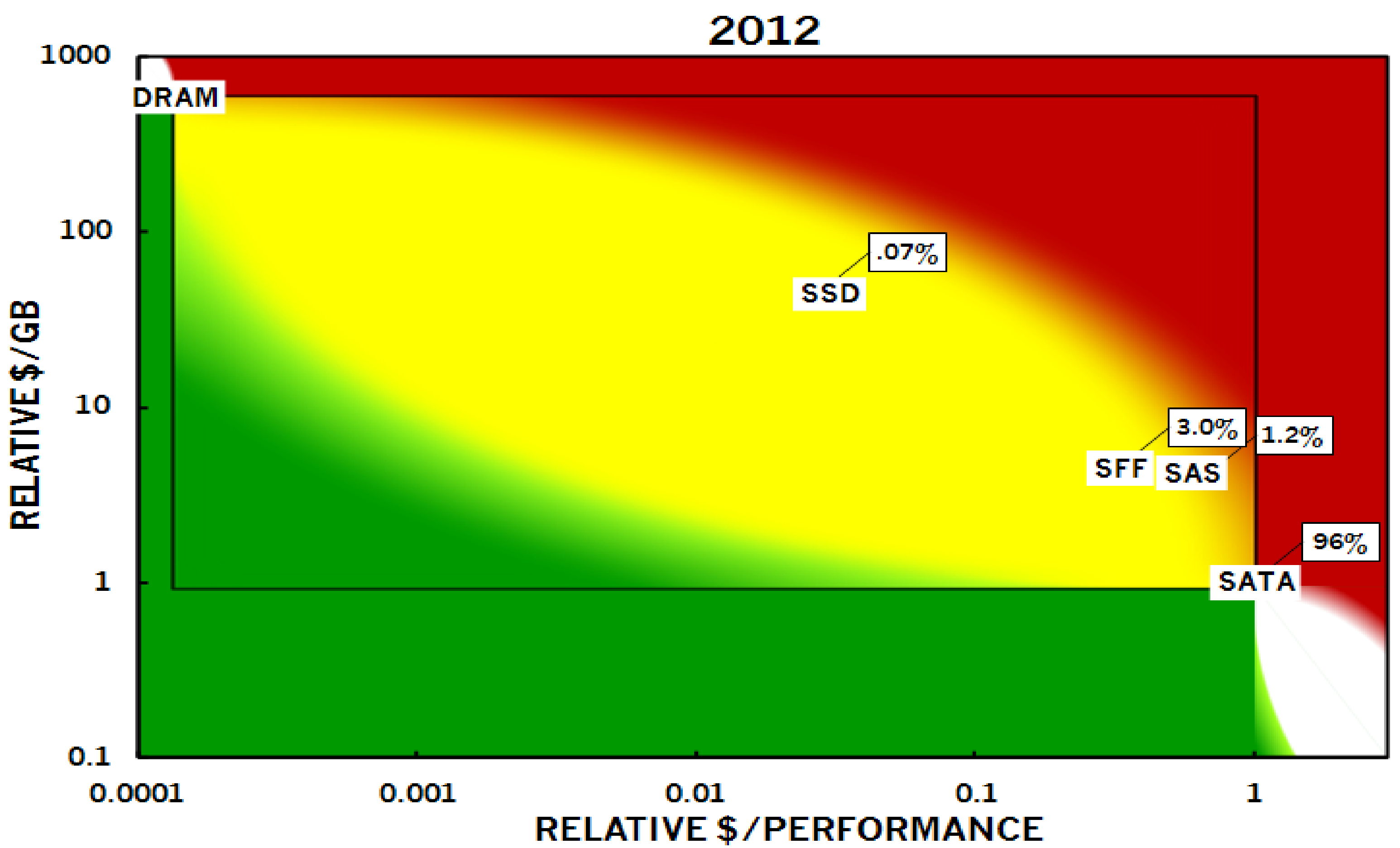
Overview

* SSD Usage Model and Application Classes
* Cost Model
* Application Specific Endurance Model
* Data Retention in the Data Center and end-of-life failure model
* Other SSD requirements

SSD by the numbers

Storage Type	Size (GB)	Price ($)	Perf	$/GB	$/Perf	Watts	W/GB
DRAM	4	143	1000000	35.75	0.000143	6	1.5
SSD (SLC)	120	1244	10000	10.37	0.1244	2	0.017
SSD (MLC)	160	480	10000	3.00	0.048	2	0.013
SAS(15K)	300	216	200	0.72	1.08	14	0.047
SAS(10K)	300	186	150	0.62	1.24	8	0.027
SAS(7.2K)	2000	293	100	0.15	2.93	5	0.003
SATA(7.2K)	2000	293	100	0.15	2.93	5	0.003

Storage Technologies by the $



*Hetzler, Steven R. The storage chasm: Implications for the future of HDD and solid state storage. *http://www.idema.org. [Online]* December 2008.

SSD Usage Models

* HDD Caching
  * Intermediate persistent cache between HDDs and memory
  * Caches hot HDD pages in SSD
  * Cost efficient, but ..
  * Require hardware/software support
* HDD replacement
  * Easy to implement but could be too expensive
  * We have cost model for this approach

The Cost Model (VLDB 2010 TPCTC workshop)

- HDD: IO is expensive
  - $Cost_{HDD}$= IOPS*$/$IOPS_{HDD}$ + $Power_{HDD}$ * $/Watt
- SSD: GB is expensive
  - $Cost_{SSD}$= GB*$/$GB_{SSD}$ + $Power_{SSD}$ * $/Watt

- For SSD to be viable:
  - $Cost_{HDD}$ > $Cost_{SSD}$
  - IOD * $/$IOPS_{HDD}$ + PD_{Δ} * $/Watt > $/$GB_{SSD}$

The Cost Model (cont.)

- $IOD * \$/IOPS_{HDD} + PD_{\Delta} * \$/Watt > \$/GB_{SSD}$
  - IOD: IOPS/GB, workload dependent
  - $\$/IOPS_{HDD}$: $1.24
  - PD_{Δ}: 0.01 Watt/GB
  - $/Watt: $10
  - $\$/GB_{SSD}$: $10.37 SLC & $3 MLC
- Solve for IOD:
  - IOD > 8.28 (SLC)
  - IOD > 2.34 (MLC)

SSD Usage and Application Classes

	HDD caching	HDD replacement
Commodity Systems	• Map/Reduce • File system • ECN	• Key/Value Store • Web Search
Reliable Systems	• Enterprise OLTP • Enterprise DSS	• ?

SSD is Consumable Storage

- Apps have to monitor State of SSD
  - SMART attributes
  - OS error events
  - App-level Page Checksums
- Costing Changes:
  - In enterprise a disk (HDD or SSD) is expected to last 3-4 and should be under warranty for that duration
  - For SSD Media is not covered with Warranty
  - Extra cost for the end user.

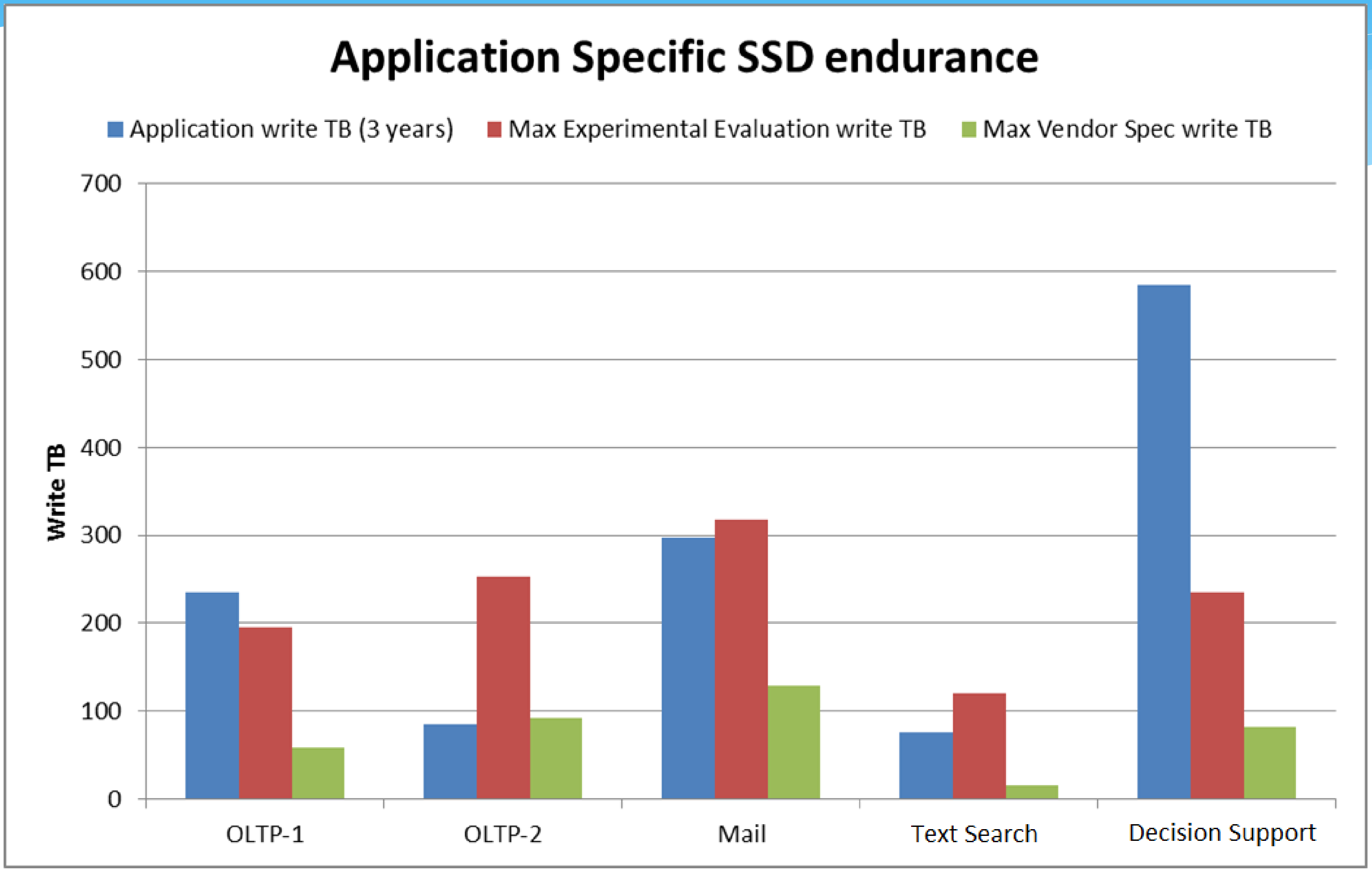
The Cost Model (Revisited)

- HDD: IO is expensive
  - $Cost_{HDD} = IOPS * \$/IOPS_{HDD} + Power_{HDD} * \$/Watt$
- SSD: GB is expensive
  - $Cost_{SSD} = GB * EF * \$/GB_{SSD} + Power_{SSD} * \$/Watt$
- EF (Endurance Factor):
  - App 3-year Writes (GB)/SSD endurance
  - EF >= 1

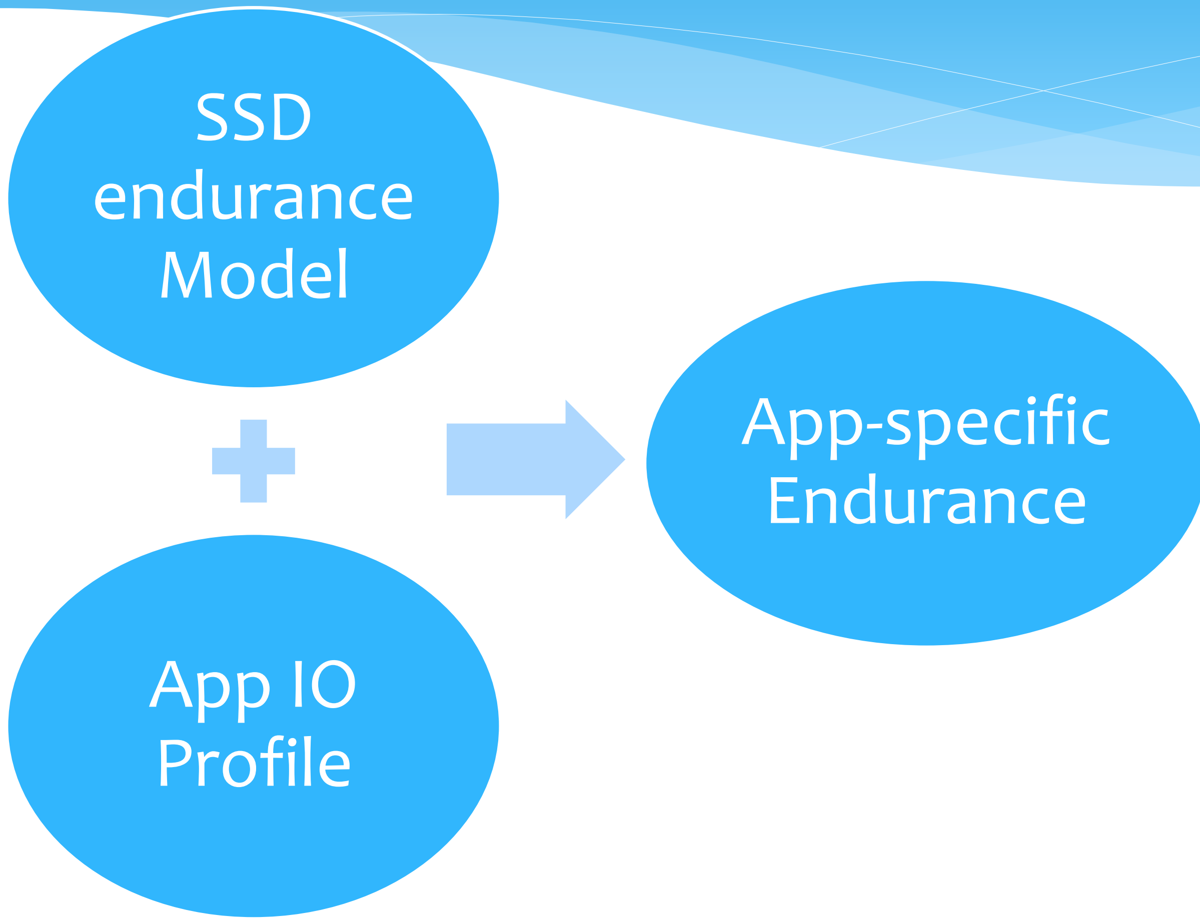
SSD Endurance

* No standard way of specifying endurance
  * Some provide a single number based on certain workload
  * Some provide sequential and random numbers
* All are inadequate
  * Ignore IO block sizes
  * Assume long retention period (1 year)

Example Measured Endurance



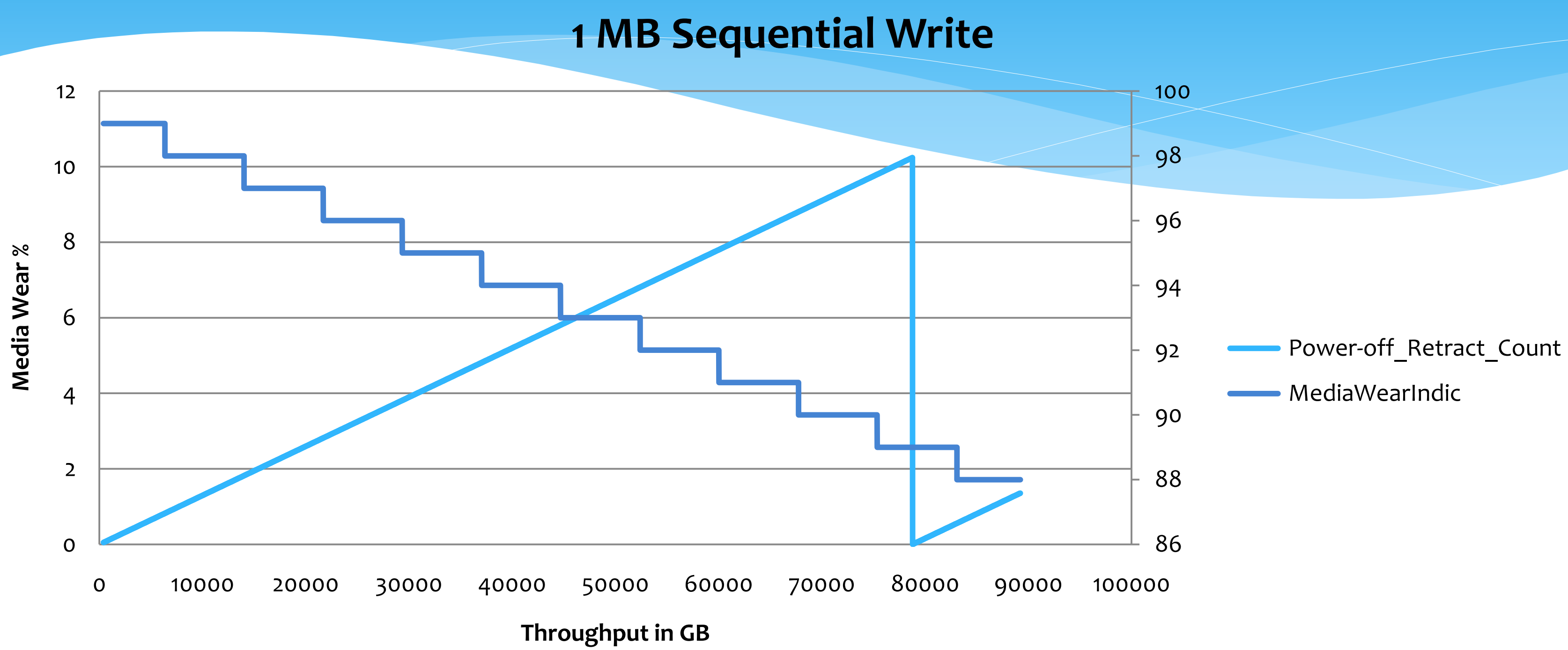
Application Centric Endurance Model



SSD Endurance Model

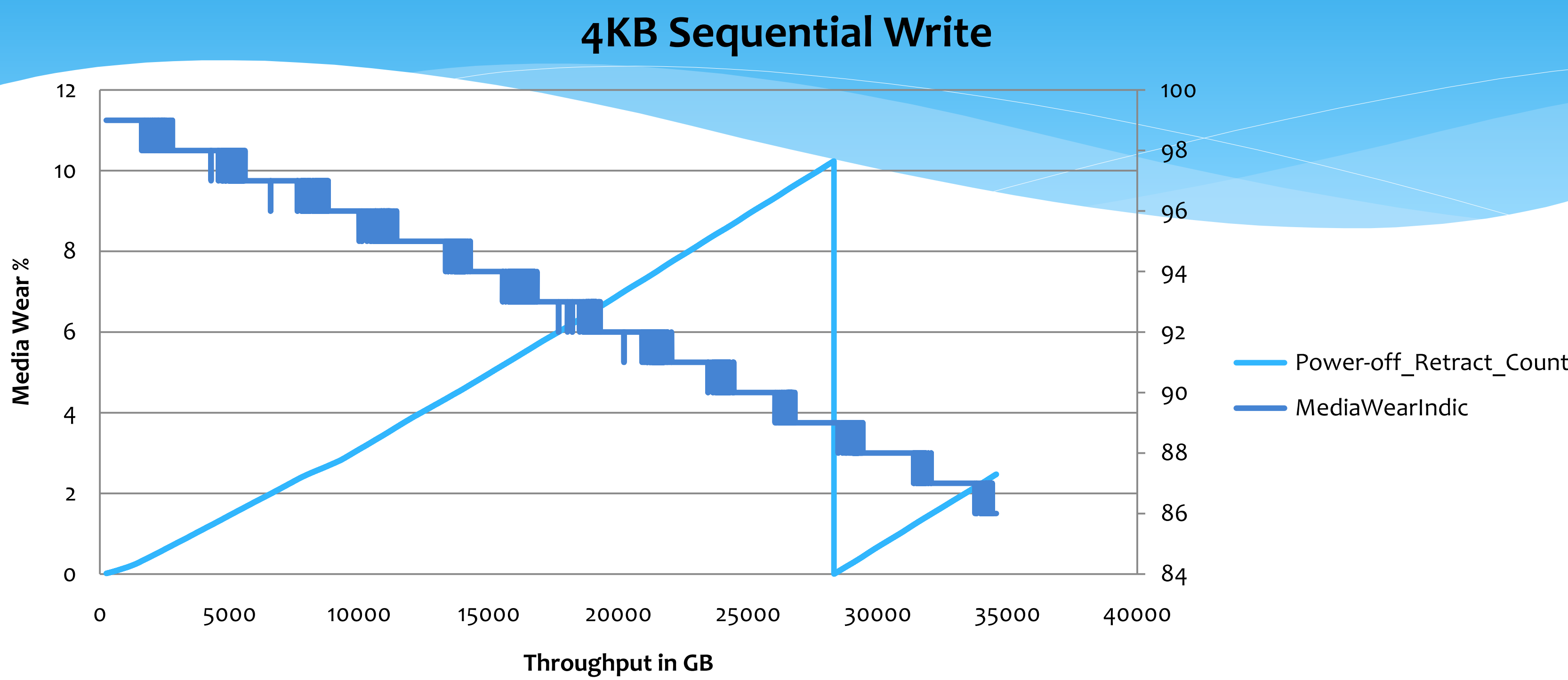
* Find random and sequential SSD endurance for most block sizes: 4KB, 8KB, 16KB, 32KB, 64KB, 128KB, … ,1MB
  * We collect SSD SMART attributes while running the above write workloads
  * The end result is figuring the write amplification model of the SSD.
* Disk used: 160GB MLC
  * 15TB (45TB overprovisioning) random endurance
  * 380 TB sequential endurance

1 MB Sequential Write



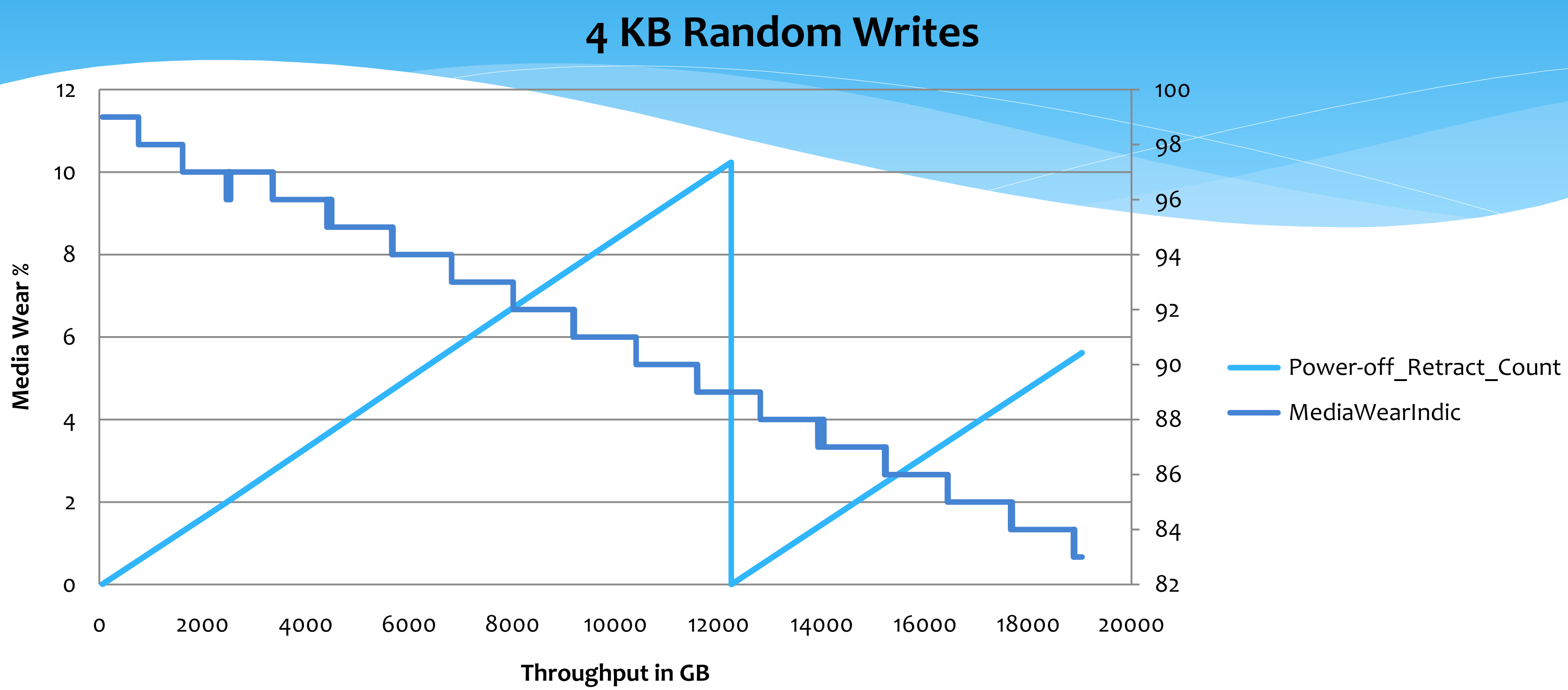
* Based on attribute 226: 785 TB
* Based on attribute 233: 743 TB
* OEM spec: ~385 TB.

4 KB Sequential Write



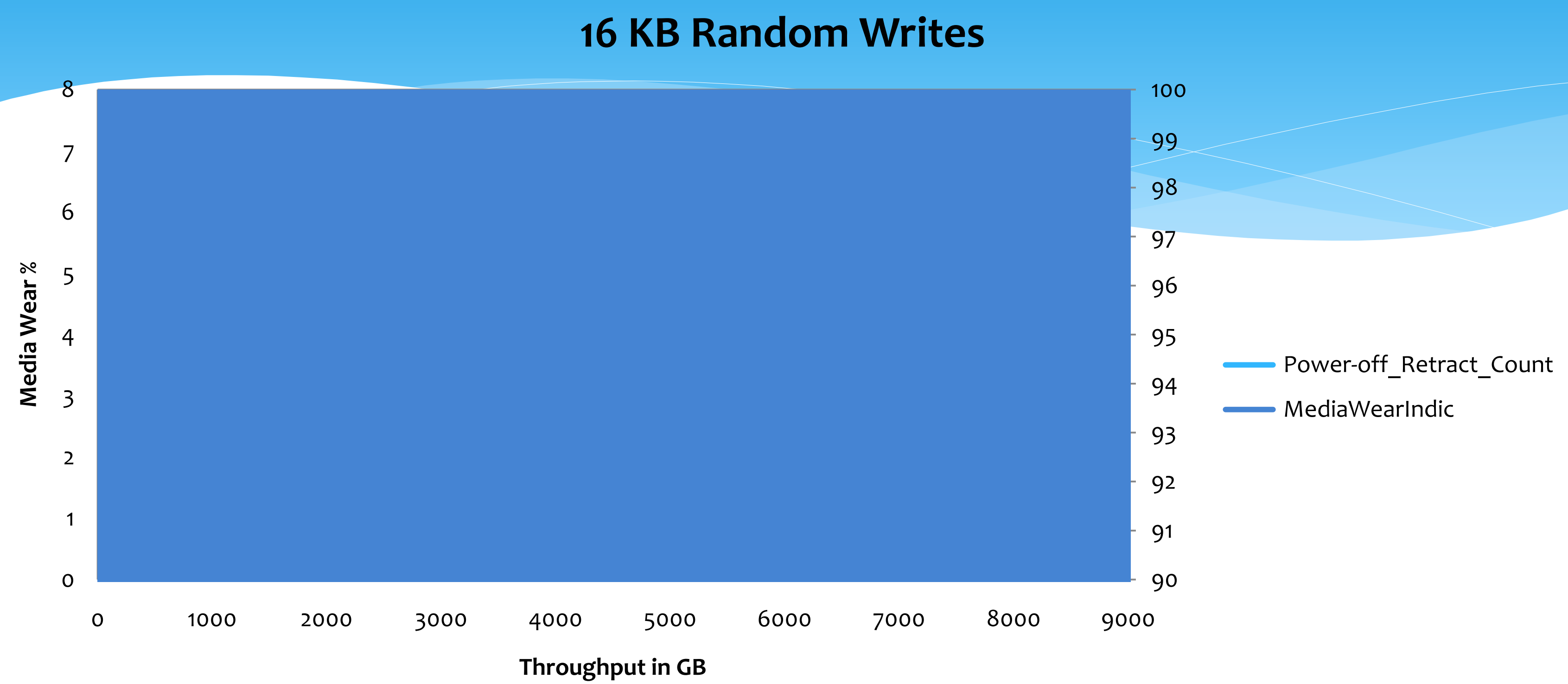
* Based on attribute 226: 277 TB
* Based on attribute 233: 246 TB
* No OEM spec.

4 KB Random Write



- Based on attribute 226: 122 TB
- Based on attribute 233: 112 TB
- OEM spec: 15TB

16 KB Random Write



- Based on attribute 226: 120 TB
- Based on attribute 233: 93 TB
- No OEM spec

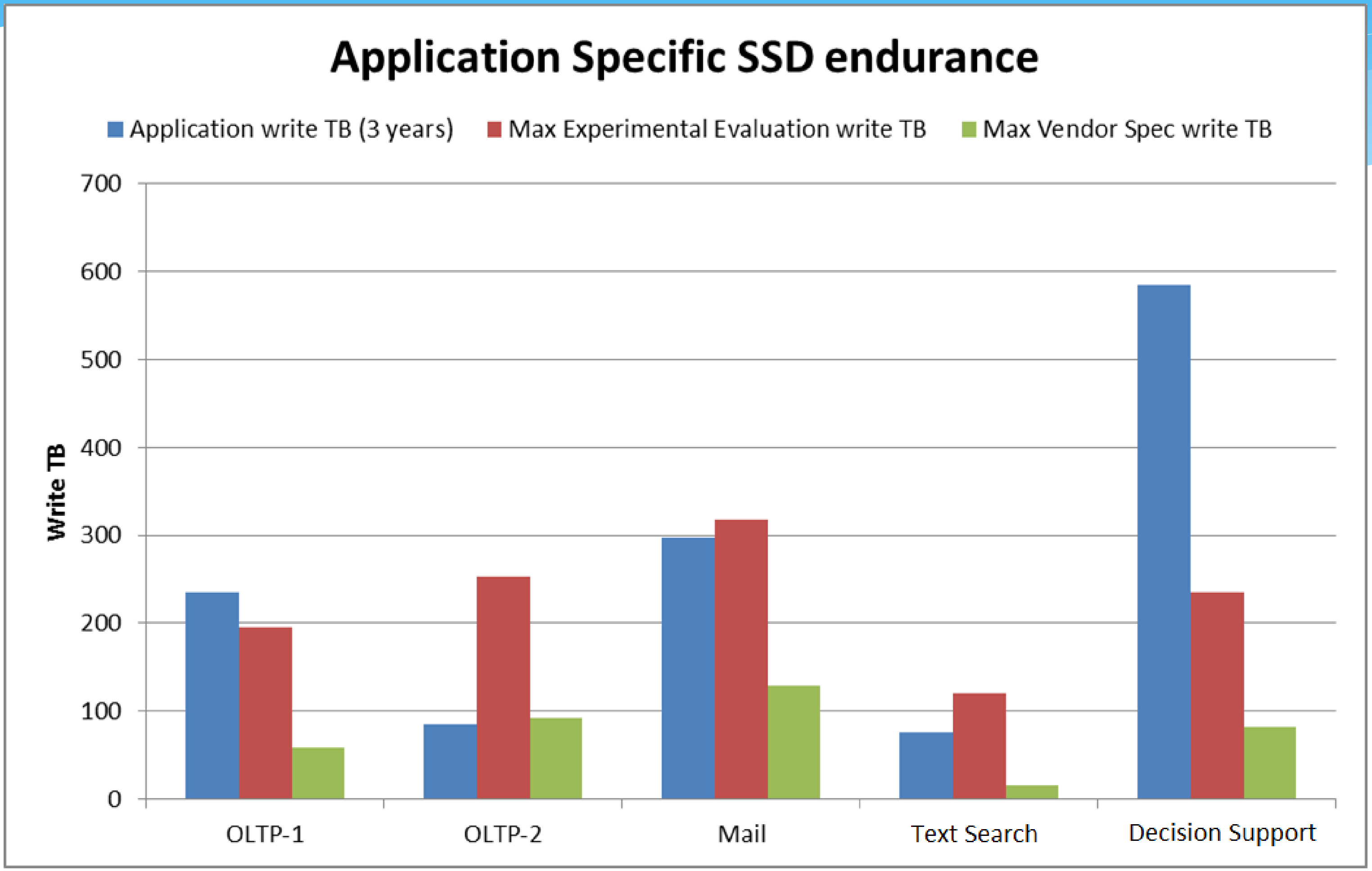
Workload IO Characterization

* We use Windows ETW infrastructure to collect Disk IO traces
* Wrote tools to process those traces and extract:
  * Read/Write distribution by block size
  * Randomness
  * IO density

IO Characterization of an App

Request Size	Total	% Total	Reads	% Read	Writes	% Writes
6	35860203	100%	35860203	100%	0	0.0%
256	84559	0.0%	0	0.0%	84559	97.3%
4	1749	0.0%	39	0.0%	1710	2.0%
32	205	0.0%	0	0.0%	205	0.2%
28	133	0.0%	0	0.0%	133	0.2%
8	103	0.0%	0	0.0%	103	0.1%
24	82	0.0%	0	0.0%	82	0.1%
12	70	0.0%	0	0.0%	70	0.1%
16	65	0.0%	0	0.0%	65	0.1%

Example Measured Endurance



Data Retention in the Data Center

* Very minimal data retention requirements
* Days not weeks
  * Data is replicated across servers and across data centers
  * If a server is down for few hours, rebuild server
* Servers always on
* We want to use SSD post 100% Media wear

End of Life Failure Model

* Can we push SSD beyond per-spec 100% media wear?
  * Answer is yes, based on our reduced data retention
  * We already collect all SMART attributes/OS events
  * Need the right SMART counters to predict "end" of life
    * Correctable ECC errors
    * Free blocks/retired blocks.
* But …
  * Certain SSDs will disable writes at 100% media wear
  * Others do not throttle writes but provide no mechanism of detecting true end-of-life

Other Disk Requirements

* SMART counters:
  * Must:
    * % media wear
    * Host writes (GB)
  * Like:
    * Free blocks/retired blocks
    * FTL writes
    * ECC corrections
* No endurance or end of life write throttling
  * Need for guaranteed SLA
* Secure Erase

Conclusion

* Endurance specification are ineffective and useless for Cloud apps:
  * Proposed a new app-specific endurance model
* Data retention requirement in the cloud is not strong (few hours – days max)
  * We need to go beyond 100% media wear to fully ustilize the disk
    * Need some visibility into the health of the disk (ECC corrections, for example)
* No throttling at any stage: we need a predictable performance to maintain our SLA
* Rich set of SMART counters will help us monitor and manage SSDs
  * Standardizing counters will simplify our software