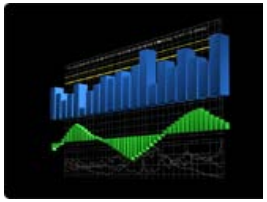
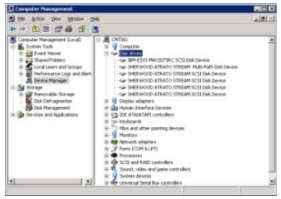
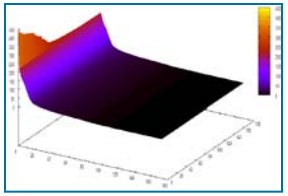
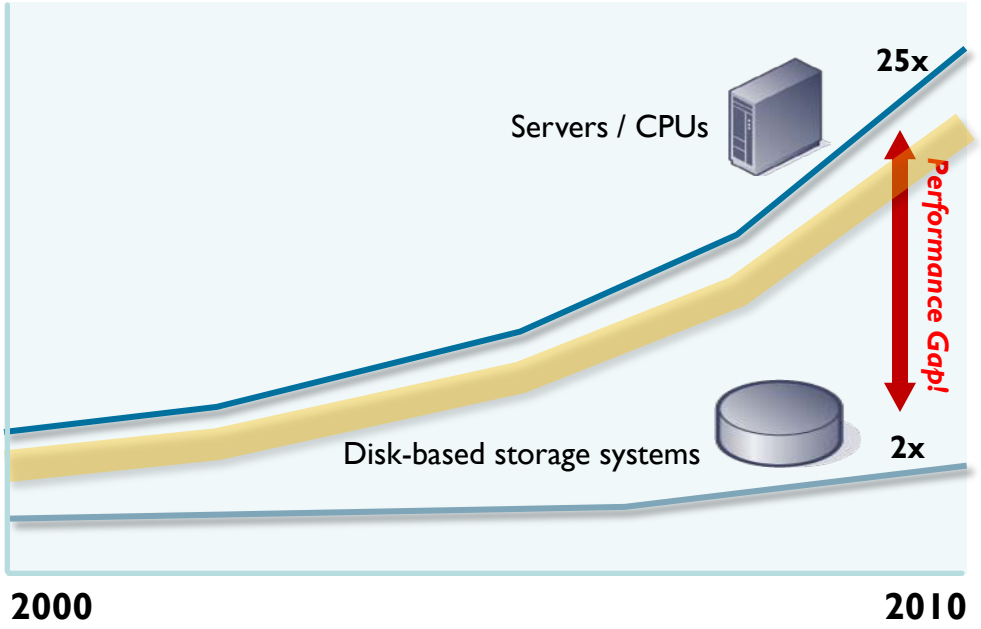
Storage Acceleration, Driven by Autonomic Software

Dr. Sam Siewert,
Formerly CTO Atrato, Inc.
Presently with Intel Architecture Group

The Storage Dilemma

Increasing Performance Gap between Servers and Storage



SSD Driven Performance

Tiered Storage and IO Page Cache
10+ Year Window… At Least

NVRAM Scaling for Storage Virtualization Emerging Quickly

- QPI Scaled DRAM to Terabytes Per Node
- Nand Flash SLC/MLC to 10's of Terabytes Today Per Node
  - PCI-e Nand Block NVRAM/Storage Devices
  - SAS/SATA Nand Solid-State Disk or SSD
- PCMS – Stackable PCM, Around the Corner…?
- Memristor, Racetrack, Nano-RAM, Further Out…?

Approx Cost	Device Type	Rand Latency	Scaling	Example
$100+/GB	QPI-DDR	640ns/sector (10ns/Word)	0.33TB/U (2TB/6U)	IBM 3690
$25+/GB	PCI-e SLC Nand	2 to 26 µsec per sector	0.8TB/U (320GB/Slot)	FusionIO, Micron
$5+/GB	SAS/SATA SLC SSD	75 to 200 µsec per sector	1.2TB/U 24x100GB In 2U	Intel, Pliant, STEC
$1+/GB	2.5"/3.5" HDD	3 to 100 milliseconds	30TB/U (60x2TB 3.5")	Atrato Inc.

The Best Thing That's Happened *To Storage Since…*

RAID, Storage Area Networking, Virtualization

…

Solid-State Tier and Primary Storage

- Tiering and Cache For Enterprise – Solid-State Thin Provisioning, As Needed, Avoid Concern about Cost and Data Protection

Drive Replacement in Consumer (Laptop, Netbook) Space First

Storage Has Suffered with Moore's Law for Capacity Alone (Not Access)

Storage Now Has Access Moore's Law Potential

- Long and Rich Roadmap of Devices Along with Access-Hungry (IOPs Sensitive) Applications
  - Cloud, Semantic-Web, Ontological Web, Business Intelligence, Analytics, OLTP, Meta-Data hosting, Desktop Virtualization

HSM Tiered Storage Architecture

Current Limitations

Sizing Storage Tiers?

- Over-provisioning of costly SSDs
- Unable to predict/show performance gains
- No metrics to measure improvement

Need Scalable High Capacity/Density Arrays

- Not bandwidth matched to scale capacity
- Does not leverage HDD=Capacity, SSD=Access
- Designed for Disk-to-Disk Tiers and Migration

Inefficient Management

- Slow Activation Policy, Not adaptive to changing access patterns
- Requires IT time and resources
- Extra IO for Migration

IO Page Cache Architecture
Current Limitations

RAM is Too Small, Too Fast

- Highest $/GB
- Many Orders of Magnitude Faster than Disk

Battery Backing / UPS Required

- For Write-Back Ingest Cache
- Costly to Mirror

Hard to Scale

- TBs/Server at Best OTS

Why not Use Nand Flash?

Combine Cache and Tier Mgt?

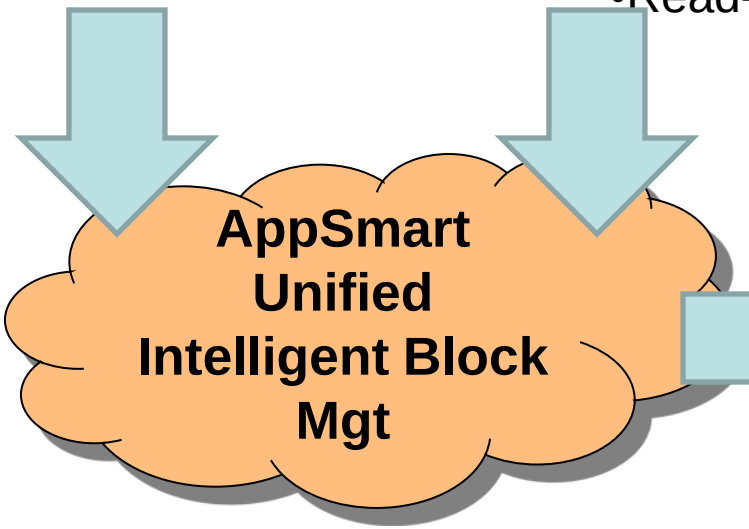
Tiering (HSM Data ***Migration***)

- Cost to Host Data
- How Active?, $/GB
- Tier Down to Save
- Policy Driven
- Slow Activation (Every Day)
- Extra IOs

v.

Cache (Data ***Replication***/Ingest)

- Access Cost (Latencny, $/IOP)
- Replication of Hot Data
- No Added Capacity
- Every IO (Page Replacement)
- Write-Back for Fast Ingest
- Read-Ahead for Sequential



Combined Tiering/Cache

- Fast Access, High Capacity
- Low $/GB
- Low $/IOP
- Scalable
- Thinly Provisioned SSD

New Data Center Storage
ApplicationSmart Self-Optimization

Storage Page Cache and HSM Limitations: Cache is limited in scale/scope, HSM is slowly activated

RAM Too Small
- GB, need TB
- Limited Spatial/Temporal Scaling

Traditional Storage Arrays

Page Cache & HSM

Imprecise, Not Adaptive
- RAM cost/GB high
- Cache hit rate efficiency with LRU/MRU low
- Tier Migration Adds IO

Bandwidth Mismatch w/ Cache & HDD

Missed Penalty, Too High
- Over-provisioning of cache
- Too much rack space for HDD
- HSM migration with SSD is slow and steals bandwidth

ApplicationSmart Provides Data Access Acceleration: Manages purpose-built set-cache Solid-State Tier

Identifies Cache-ability by Application and Computes Speed-up

Data Storage for Random Data Access

Profile Monitors Access Changes

Extracts Spatial and Temporal Patterns from Semi-Random Access Truly Random Requires Spindles and SSD

Terascale

SSD

HDD

Petascale

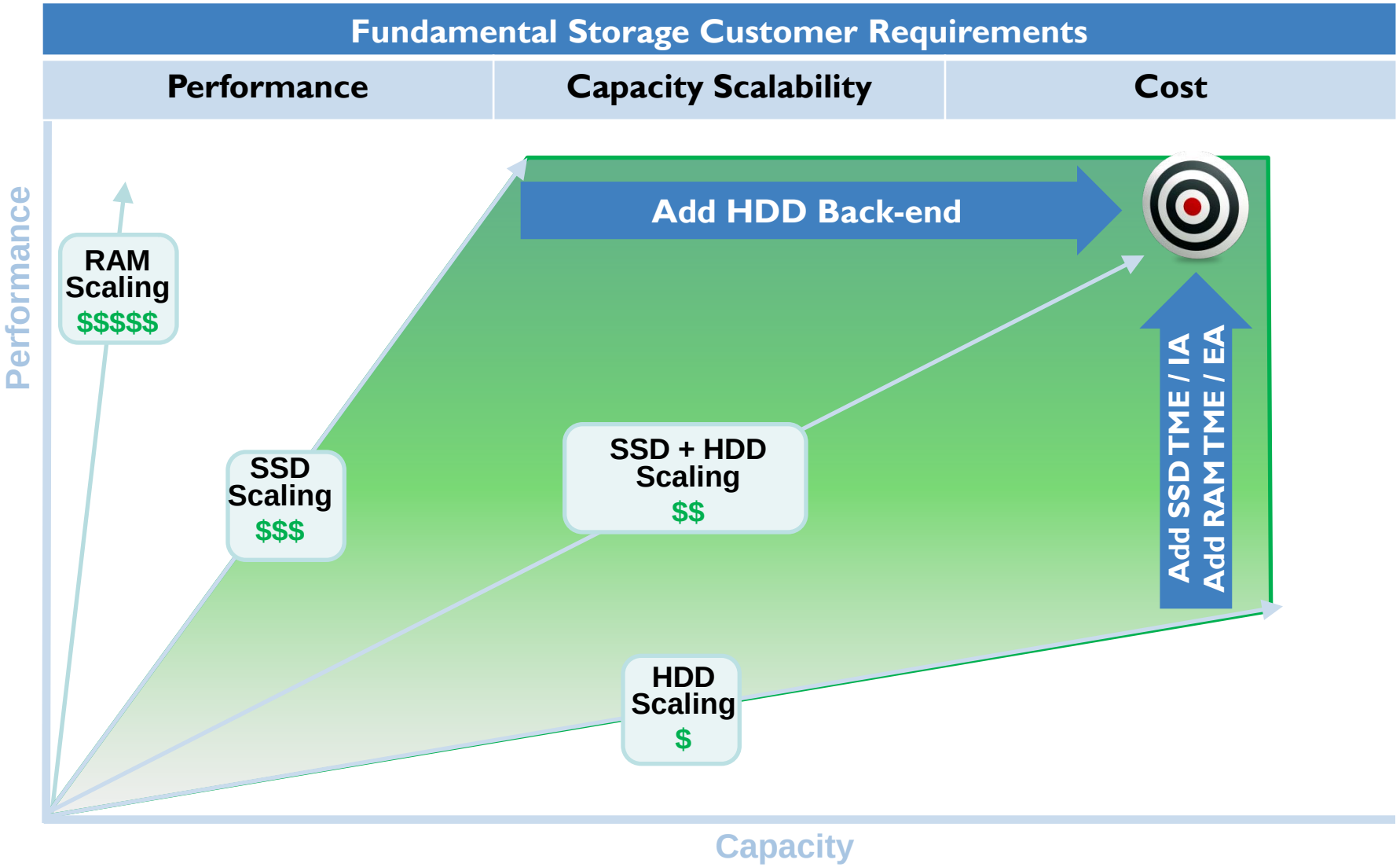
Tier 0 Replication (TME) , Log-Structured Write-Back (IA)

Autonomic Storage Tiering

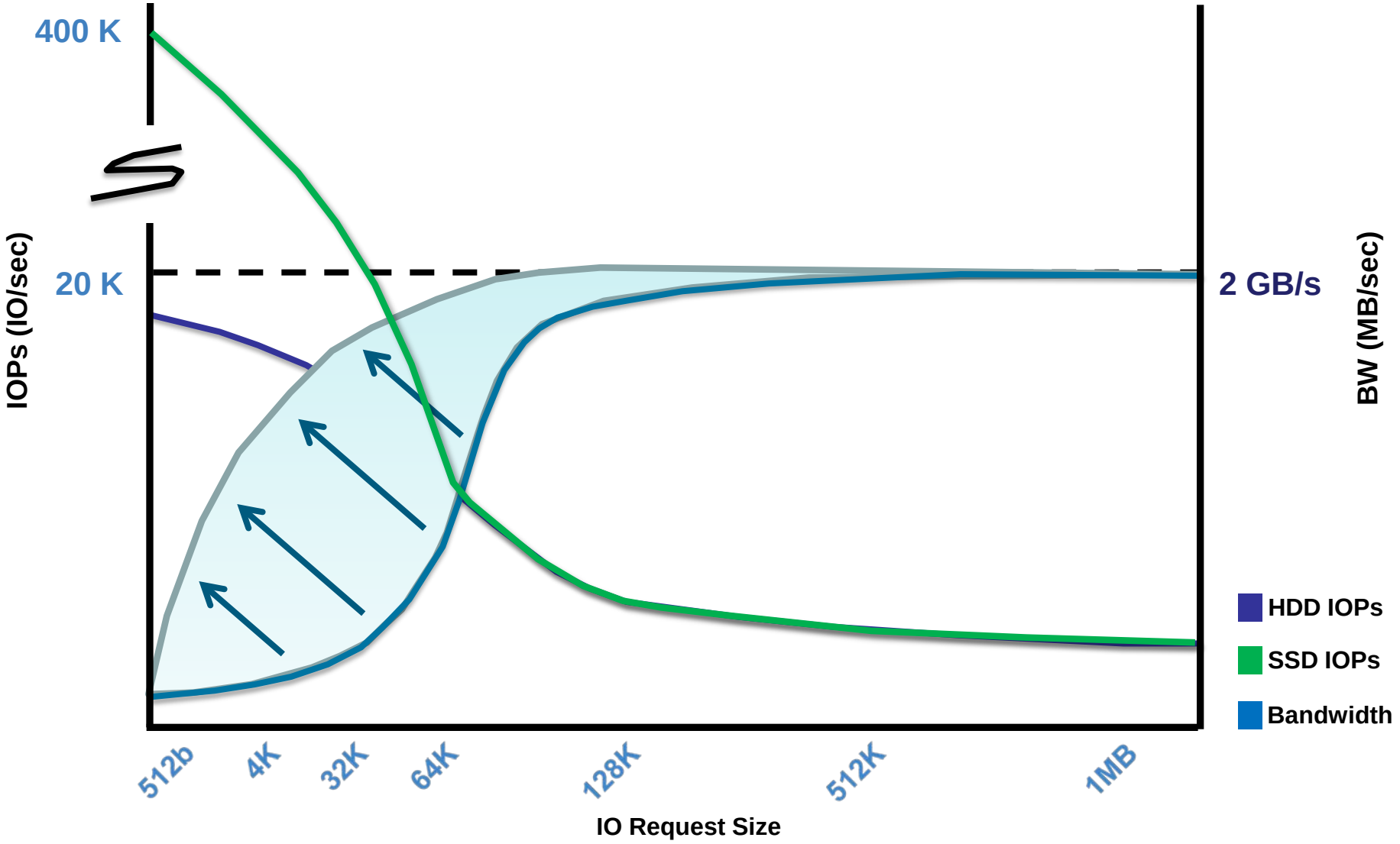
Hybrid VLUN spans SSDs and HDDs

Customer Benefit	Description
Sizes SSD requirements	• Analyzes and recommends amount of SSD *prior* to purchase • Only what is needed for applications, based on profile • No over-buying or over-provisioning
No added management	• Enables autonomic data tiering, no policies to set • Anticipates SSD needs based on data access patterns
Eliminates overhead	• Data is replicated but remains resident on HDDs • Avoids migration to and from HDD and SSD • No unnecessary IO, all tiering is opportunistic

The Bottom Line - Hybrid Storage Delivers Flexibility to Solve Problems



Performance Increase with SSDs



Multi-Tiered Management Software

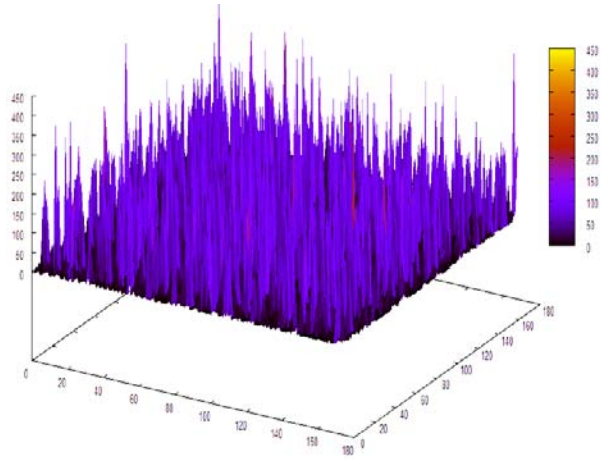
ApplicationSmart™

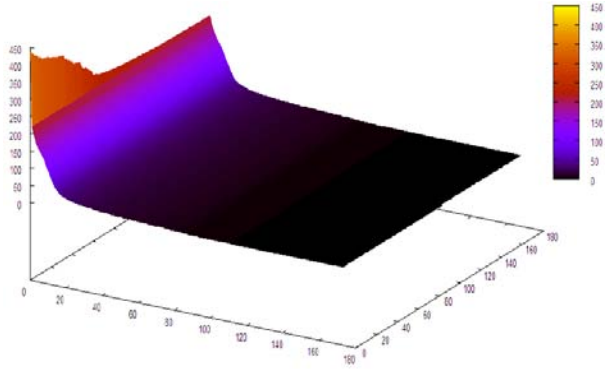
Access Profiler	• Adaptive histogram, highly compressed, scales to PB • Drives TME to accelerate IO for high access content
TME (Tiered Management Engine)	• Dynamic block replication with access pattern changes • Optimal FBR (or plug-in heuristic) set replacement • Mapped to LUNs or pools of LUNs
Ingest Accelerator	• Log-Structured Write-Back FIFO, Low Latency Completion • Tuned for RAID (aggregation & RAID Set IO reforming) • Check-point for Replay, Mirroring
Egress Accelerator	• Detector for sequential/random initiator streams • Read-ahead cache with auto enable/disable
SLM (SSD LUN Manager)	• Full AVS VLUN creation and management • SSD storage pool, data lifetime protection options

Data Access Profiler

Provides real-time application storage access patterns

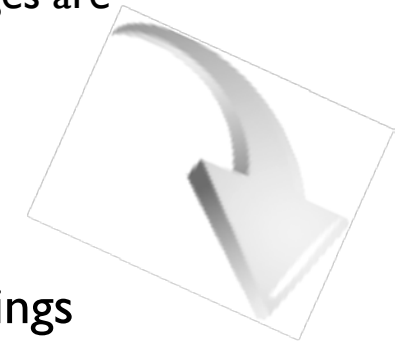
1000 Client VoD Workload





Histogram Analysis

- Identifies access hot-spots
- Notes when access changes are statistically significant
- Mapping integrates with virtualization engine

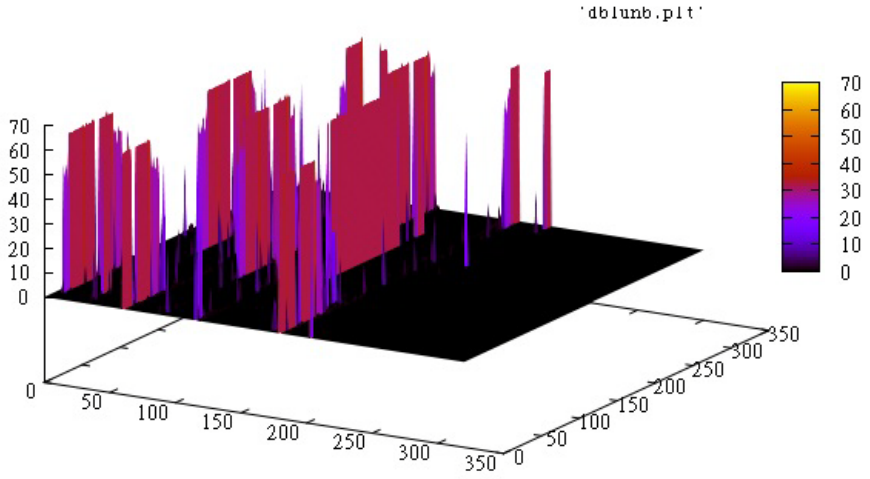


Histogram Groupings

- Drives TME IO acceleration
- Replicates blocks when statistically significant
- Provides continuous opportunistic updates
- Uses access visualization

Applications and Workloads are *Semi-Random*

Multi-Million OLTP Workload

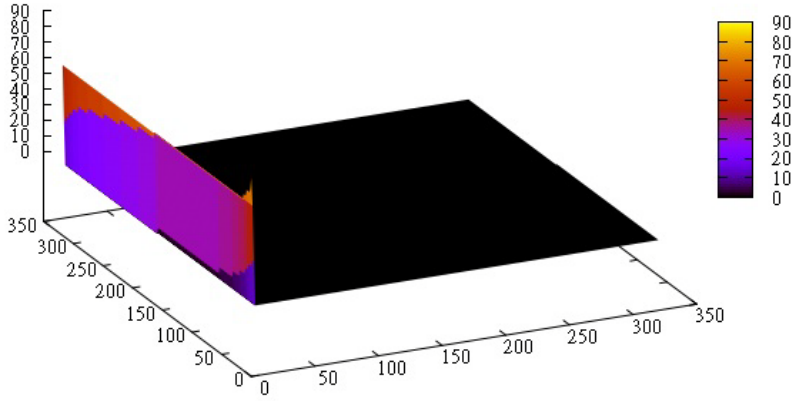


Total Active Data Small Subset

- Too Dynamic for HSM
- Flash/SSD Cache with FBR Dynamic and Large Enough

Profiling of More Applications

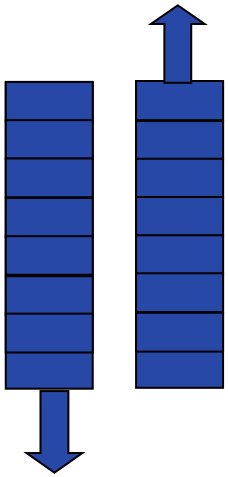
- Build a Knowledge Base by Application
- Being "Application Smart"
- Combine with Application Aware Storage Concepts



Handles Full Spectrum of Workloads

Sequential

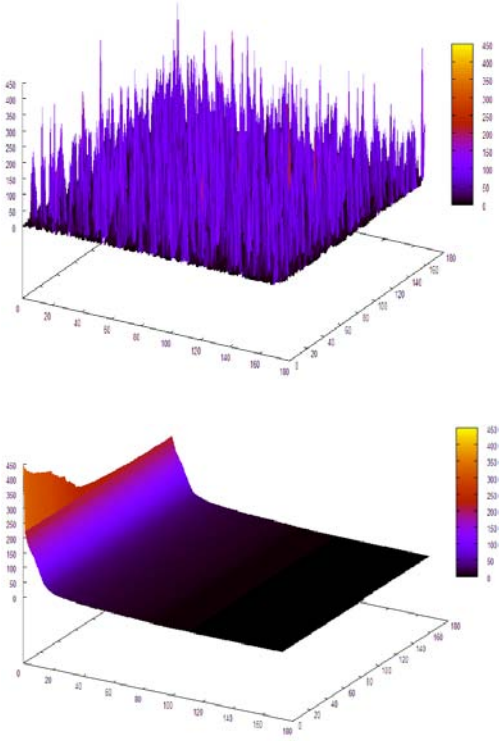
Egress IO read-ahead



Ingest IO reforming

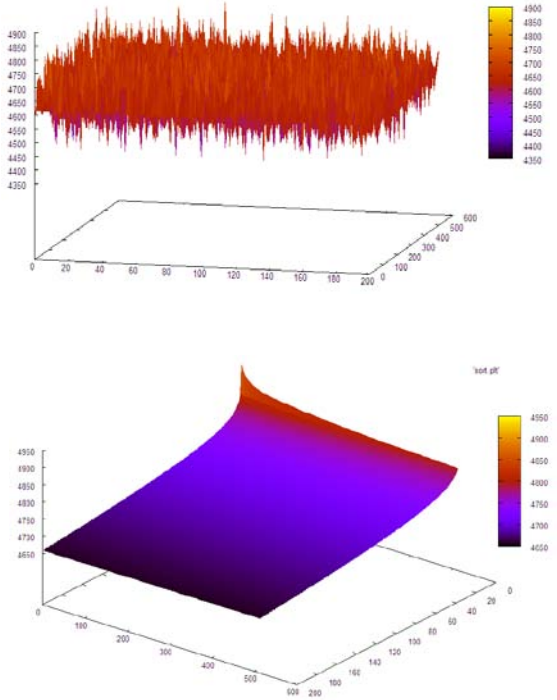
Fully Predictable
(Solid State FIFOs)

Hot-Spots



Semi-Predictable
(Scalable Hybrid Flash/Disk)

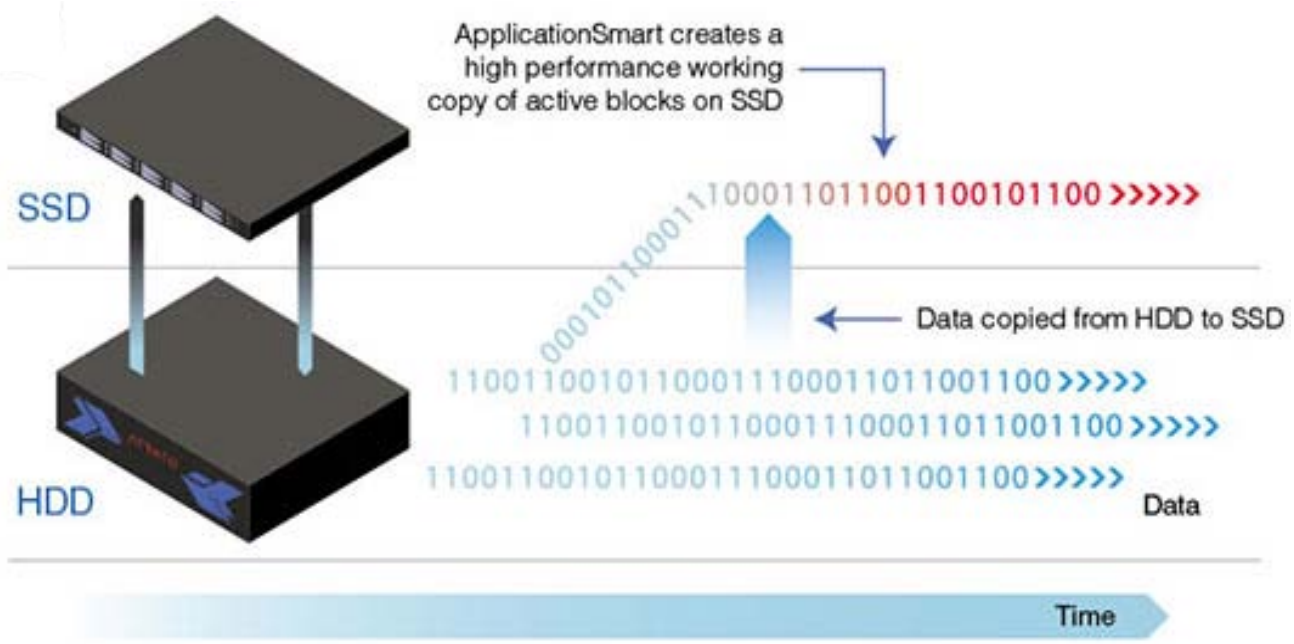
Random



Non-Cacheable
(Solved by Spindle Density or Random Access Solid-State Storage)

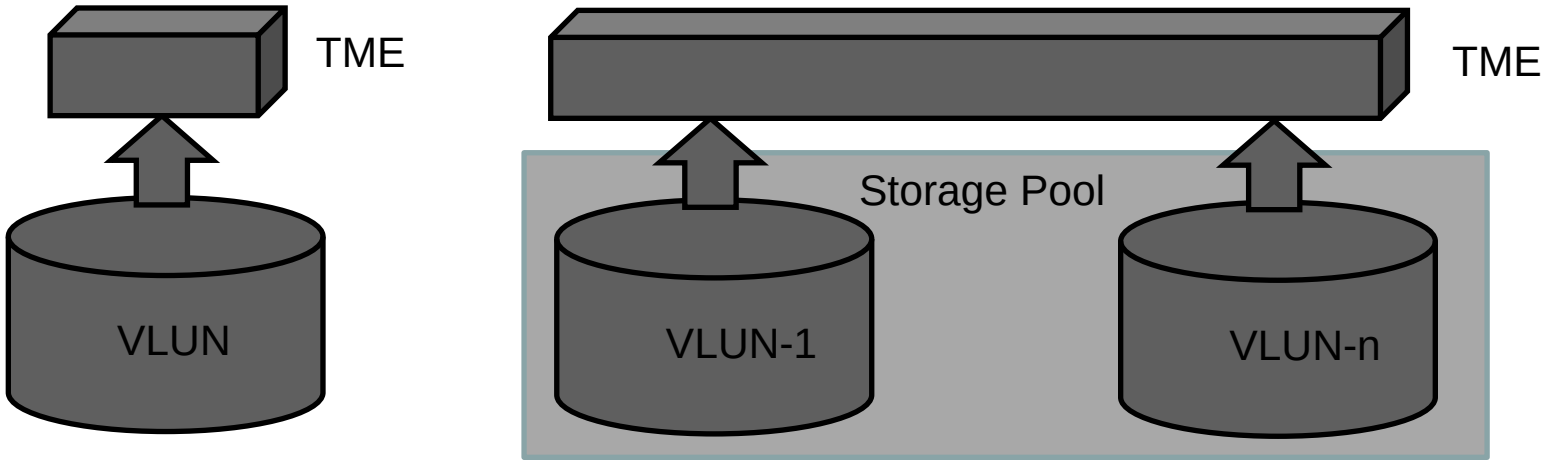
Tiered Management Engine (TME)

- Uses output from Access Profiler to drive TME
- Dynamic block replication (Read/Write-Thru Load & FBR)
- As patterns change, new blocks are replicated
- Overwrites less active data

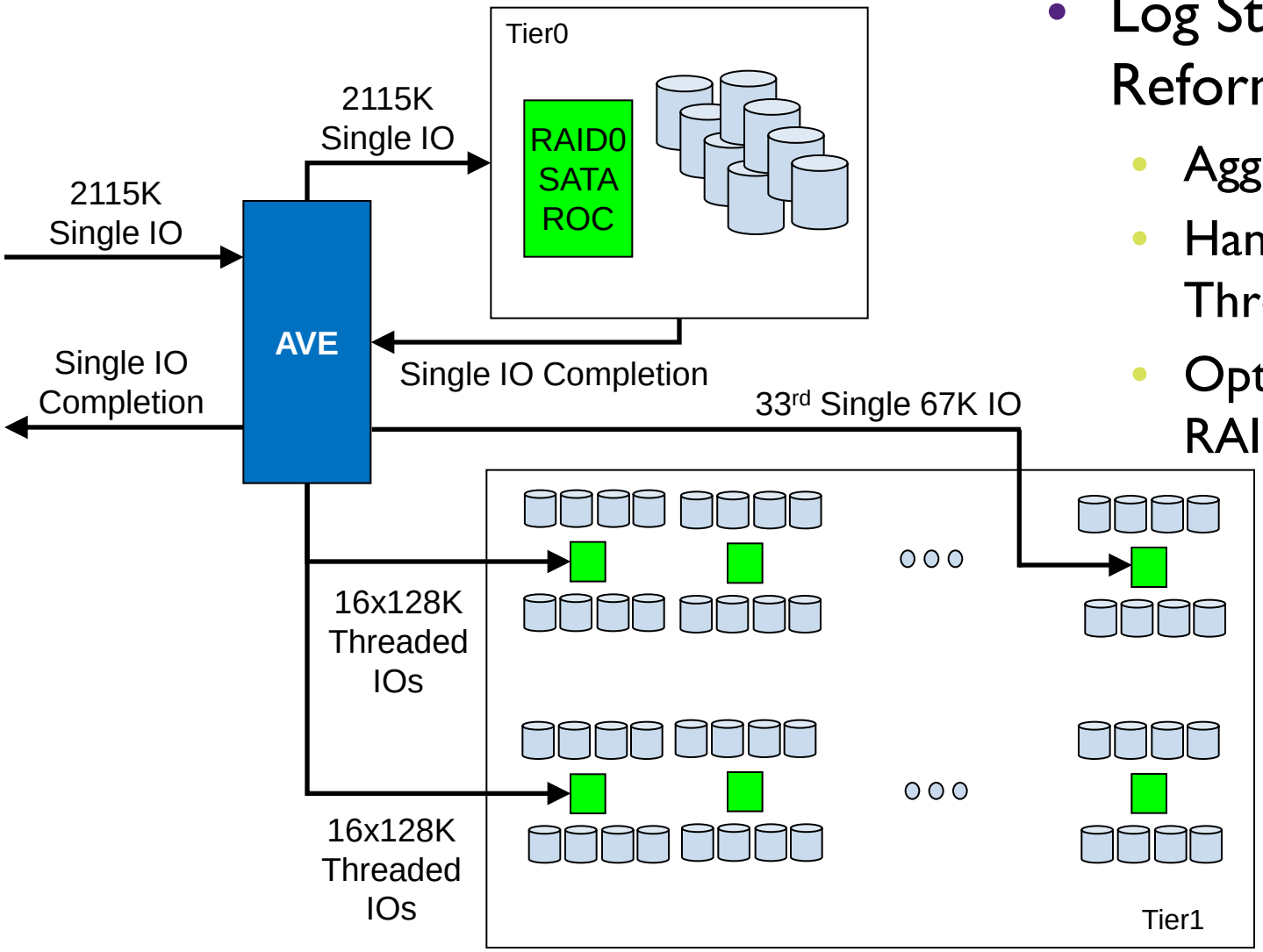


TME Spatial Locality Features

- Per LUN SSD Mapping
  - Initiator-Target-LUN Scope
  - Exclusive Provisioning of SSD to a Hybrid LUN (Firewalled)
- Per Storage-Pool SSD Mapping
  - Group of LUNs Scope
  - Allows for Competition for Set/Page Loading Between Multiple Initiator-Target-LUN Contexts



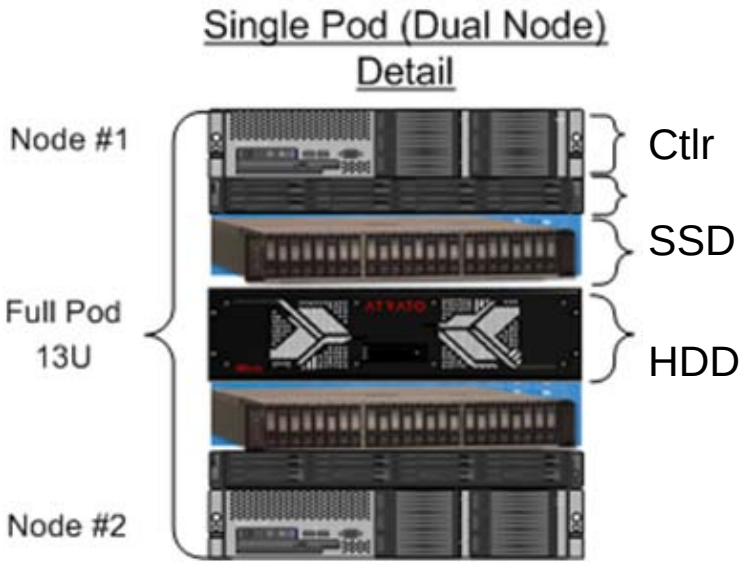
Ingest Acceleration (IA), Write-Back



- Log Structured IO Reforming
  - Aggregation
  - Handle Single Threaded IO
  - Optimizes for RAID Backend

Hybrid Node Design - Scaling

- **240TB, 8GB/sec, 1.8M Rand Read IOPs**
- **Basic Pod is 2 Nodes:**
  - **80TB**
  - ***300K IOPs/Node, 1.4GB/sec Random read/Node***
- **3 Pods (6 nodes):**
  - **12 x 10G iSCSI**
  - ***1.8M IOPs, 8+ GB/sec Total***



Similar to AppSmart?

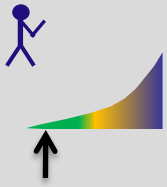
Numerous Intelligent Block Mgt, Page Cache and HSM Tiering Solutions for Virtualized Storage (To Name a Few)

- Open Source - Flashcache, Memcached
- Proprietary - IBM EasyTier, EMC FAST

Distinguishing Features To Look For

- ***Efficiency*** - Page Replicaiton/Replacement Algorithms and Hit-Rates (LRU/MRU, FBR, MFU/LFU, Combined, Advanced Heuristics)
- ***Intelligence*** - Profiling Algorithms (How Adaptive?, Overhead?)
- ***Write and Read Cache*** - Log-Structured Aggregation and IO Reforming
- ***Speed & Overhead*** - Read-Cache Look-Up, Meta-data Scaling, O(1)?
- ***Scaling*** - Tier-0 and Main-Store Scaling (Terascale/Petascale?)
- ***Data Protection*** (Mirroring on Ingest? Backed by RAID?, Dual-Controller?)
- ***Solid-State Device Support***, Thin/Hot-Plug Provisioning?

Performance Tiering Checklist

Performance Optimization		**Key Benchmarks:** • *Integrates high velocity storage tiers (Tier 0, Tier 1)* • *Block level movement for increased granularity* • *Works across multiple application workloads*
Dynamic Management		**Key Benchmarks:** • *Understands access patterns and changes* • *Moves data in real-time* • *Supports multiple VLUN configuration options*
Cost Efficiency		**Key Benchmarks:** • *Integrates efficient storage architectures (Tier, 0, Tier1)* • *Recommends SSDs only when needed* • *Fully autonomic, minimizes human intervention*

For a Deeper Dive…

"Storage-class memory: The next storage system technology", R.F. Freitas, W.W. Wilcke, IBM Journal of Research and Development, Vol. 52, No. 4/5, July/September, 2008.

"File classification in self-* storage systems", M. Mesnier, E. Thereska, G.R. Ganger, D. Ellard, Margo Seltzer, Proceedings of the First International Conference on Autonomic Computing, May, 2004.

"Application Awareness Makes Storage More Useful", Neal Leavitt, IEEE Computer Society, July 2008.

S. Siewert, D. Nelson, "*Solid State Drives in Storage and Embedded Applications*", Intel Technical Journal, July 2009.

"*Systems and Methods for Block-Level Management of Tiered Storage*", US Patent Application # 12/364,271, February, 2009.

HP, R. Stanley Williams, http://www.hp.com/hpinfo/newsroom/press/2010/100408xa.html, http://spectrum.ieee.org/semiconductors/devices/memristor-inside, http://spectrum.ieee.org/semiconductors/design/the-mysterious-memristor

IBM, Dr. Stuart Parkin,Racetrack Memory - http://www.youtube.com/watch?v=q5jRHZWQ0sc

Questions?