Flash Memory SUMMIT

Deploying Flash in the Data Center

Or How this Flash

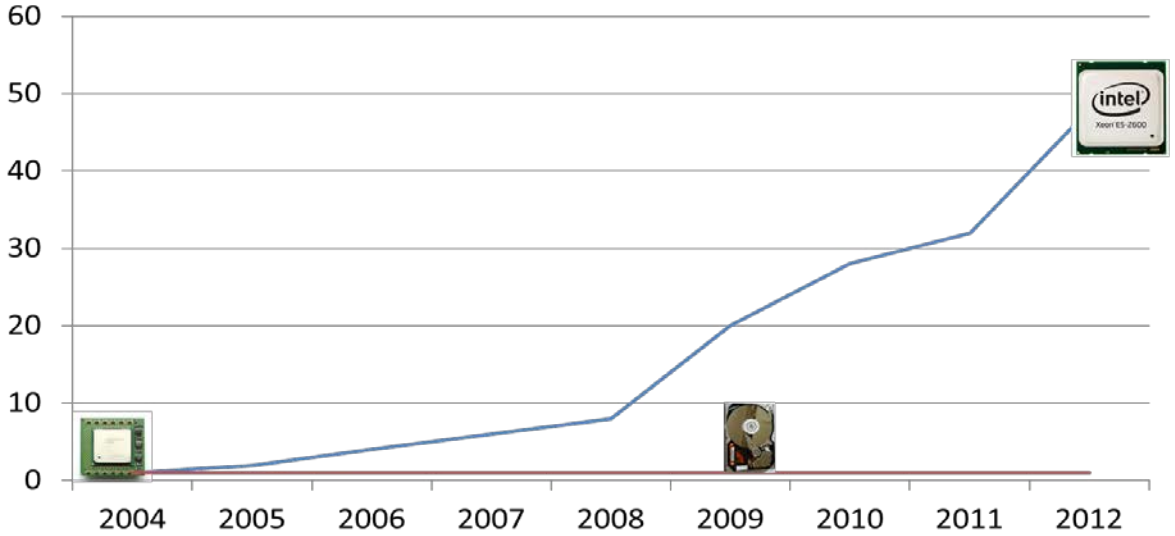
Makes Storage Like This Flash®

Today's Agenda

- The storage performance problem
- Flash to the rescue
  - A brief flash memory primer
  - Flash/SSD types and form factors
- The All Flash Arrays (AFA)
- Hybrid arrays
- Server side flash
- Converged architectures
- Choosing a solution

The IO Gap

- Processor speed doubles every 2-3 years
- Disks have been stuck at 15K RPM since 2000

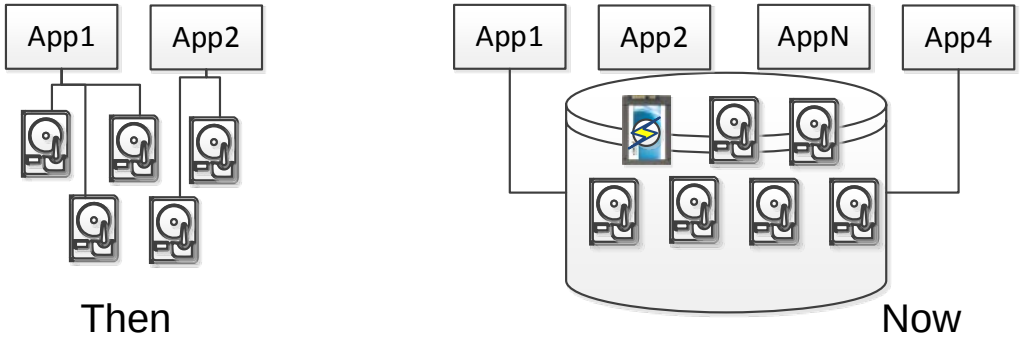


“The I/O Blender” Strains Storage

- Virtualization throws I/O into a blender… all I/O is now random I/O!
- 10 VMs doing sequential I/O to the same datastore=random I/O
- Disk drives are good at sequential, less good at random

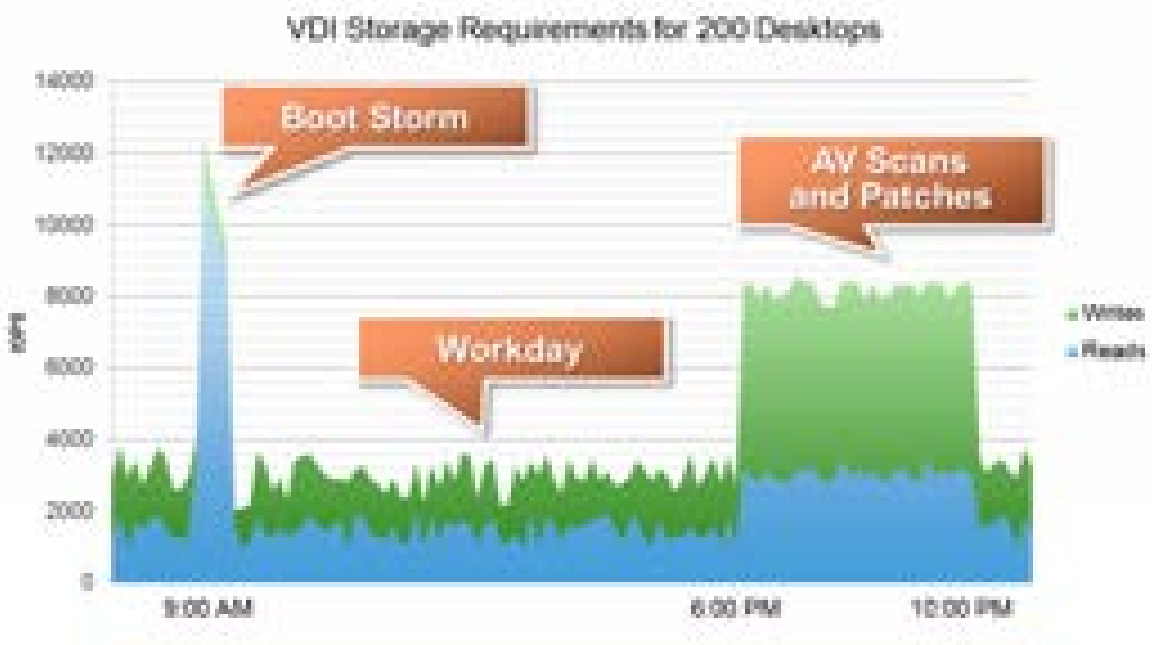
The Noisy Neighbor Moves In

- Dedicated spindles are like solid walls
  - Applications have limited effect on each other
    - Backups excepted – Same data
- Shared datastores provide no protection
  - 1 application demanding 10,000 IOPS will slow the others.

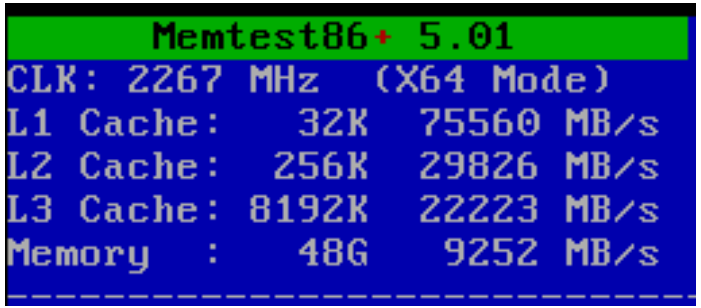


VDI Presents Unique Workloads

- Highly variable but coincident (boot/login in AM)
- Steady state 50+% write
- 40+% of projects fail due to storage performance



Data Access Performance



- L1 processor cache ~1ns
- L2 processor cache ~4ns
- Main memory ~100ns
- PCIe SSD read 16-60µs (16,000-60,00ns)
- SAS/SATA SSD read 50-200µs (50,000-200,000ns)
- Disk read 4-50ms (4-50,000,000ns)

Moral of the story: keep IOPS away from the disk

Traditional Performance Solutions

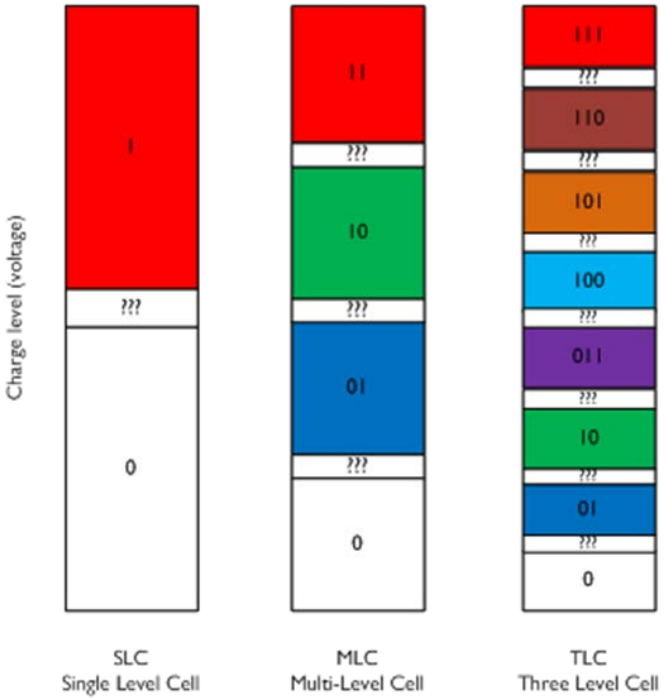
- Head per track disk drives, DRAM SSDs
  - Huge price premium limits use to the very few
- Wide Striping
  - A 15K RPM disk delivers 200 IOPS
  - For 10,000 IOPS spread load across 50 drives
    - Of course that's 15PB of capacity
  - Short stroking
    - Use just outside tracks to cut latency
- Wasting capacity wastes $ and OpEx (power, maint)

What Is Flash Memory?

- Solid State, Non-volatile memory
  - Stored charge device
  - Not as fast as DRAM but retains
- Read/Write blocks but must erase 256KB-1MB pages
  - Erase takes 2ms or more
  - Erase wears out cells
- Writes always slower than reads

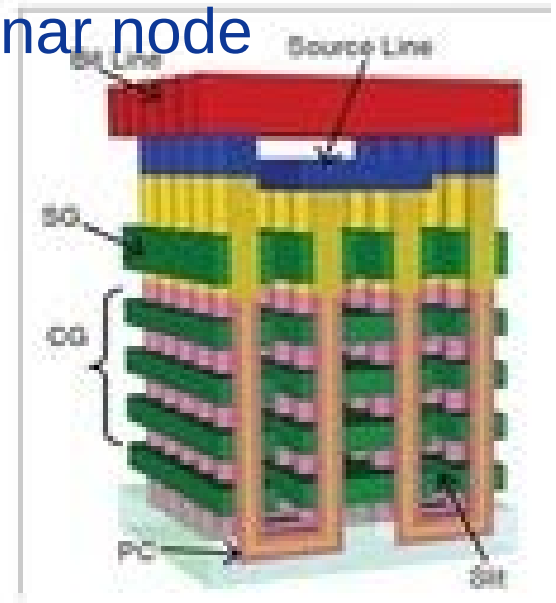
The Three, & ½, Types of Flash

- Single Level Cell (SLC) (1bit/cell)
  - Fastest
  - 50,000 program/erase cycle lifetime
- Multi Level Cell (MLC) (2 bits/cell)
  - Slower
  - 3,000 program/erase cycle lifetime
- eMLC or HET MLC (2 bits/cell)
  - Slightly slower writes
  - 12-20,000 cycles
- Triple Level Cell (TLC) (3 bits/cell)
  - 3D/LDPC boosts endurance to data center level



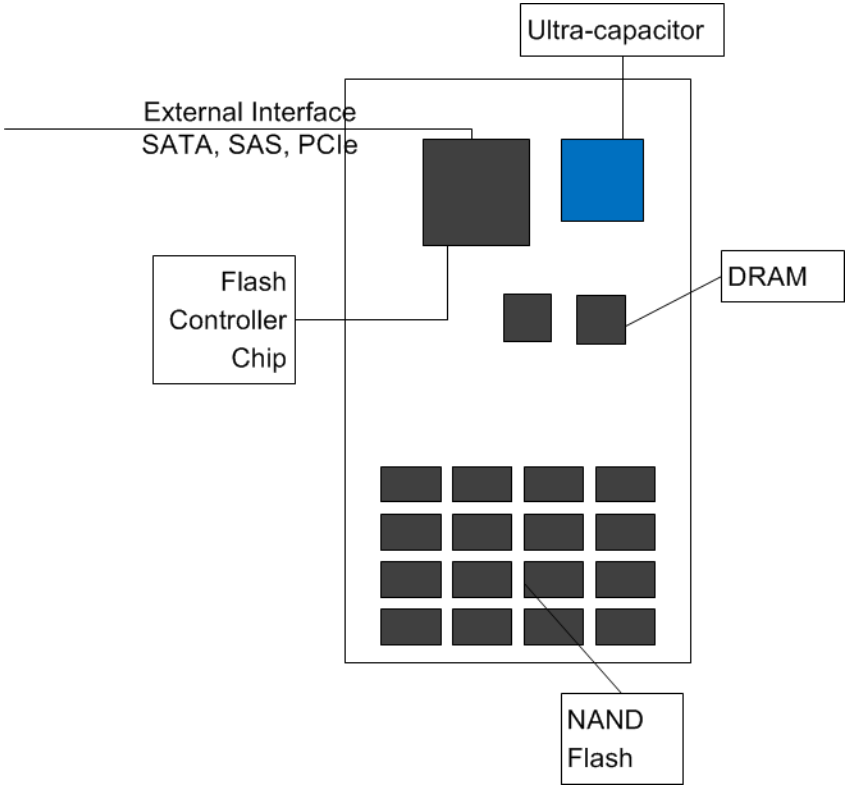
Flash's Future

- Smaller cells are denser, cheaper, crappier
  - Today's 1x nm cells (15-19nm) last planar node
- 3D is the future
  - Samsung now shipping 3D
  - Others sampling
    - SanDisk/Toshba 256GB/chip
- Other technologies post 2020
  - PCM, Memristors, Spin Torque,



Anatomy of an SSD

- Flash Controller
  - Provides external interface
    - SAS, SATA, PCIe
  - Wear leveling
  - Error correction
    - Shifting to LDPC
- DRAM
  - Write buffer & Metadata
- Ultra or other capacitor
  - Power failure DRAM dump
  - Enterprise SSDs only



Flash/SSD Form Factors

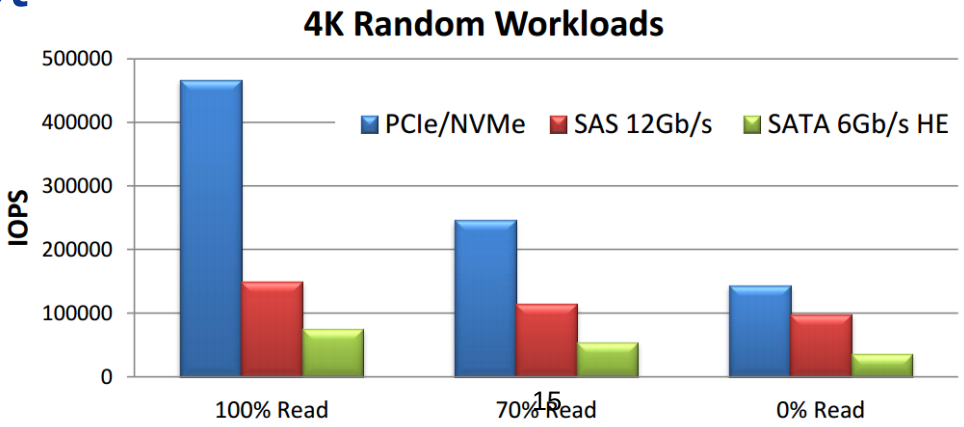
- SATA 2.5”
  - The standard for laptops, good for servers
- SAS 2.5”
  - Dual ports for dual controller arrays
- PCIe
  - Lower latency, higher bandwidth
  - Blades require special form factors
  - M.2 for small form factors like notebooks
  - U.2 for 2.5” hot swap

U.2/SFF-8639 PCIe for 2.5” SSDs

- Adds x4 PCIe 3.0 lanes to SAS/SATA connector
  - Dual ports to x2
- Appearing on new servers
  - Making PCIe/NVMe SSDs hot swappable
- Next step for storage arrays

Enter NVMe

- New logical interface for PCIe storage
  - Replaces ACHI
- More, deeper queues
- Simpler command set
- Lower latency
  - 300 vs 500+µs



SSDs use Flash but Flash≠SSD

- Fusion-IO cards
  - Atomic Writes
    - Send multiple writes (eg: parts of a database transaction)
  - Key-Value Store
  - FTL runs in host CPU

Diablo Puts Flash on the Memory Bus

- Memory Channel Flash
  (SanDisk UltraDIMM)
  - Block storage or direct memory
  - Write latency as low as 3µsec
  - Requires BIOS support
- Memory1
  - 400GB/DIMM
  - No BIOS/OS Support
  - Volatile

Selecting SSDs

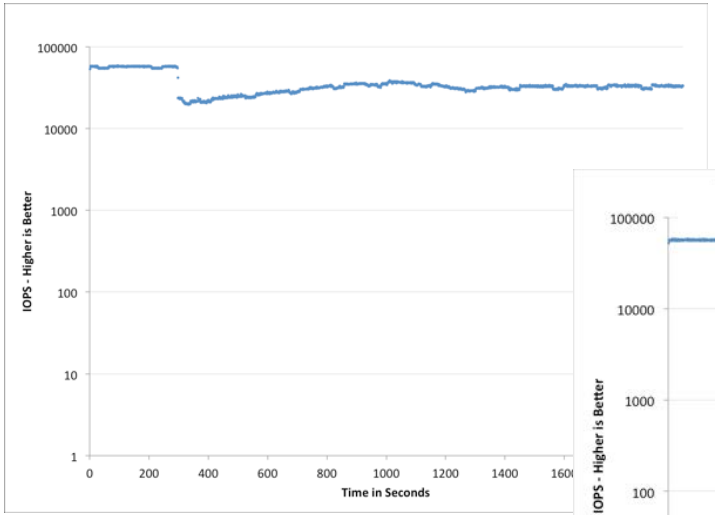
- Trust your OEM's qualification
  - They really do test
- Most applications won't need 100K IOPS
- Endurance ≠ reliability
  - SSDs more reliable than HDDs
    - 2 million hr MTBF
    - 10^17 BER vs 10^15 for near line HDD
  - Wear out is predictable
  - Consider treating SSDs as consumables
  - However don't use read optimized drive in write heavy environment

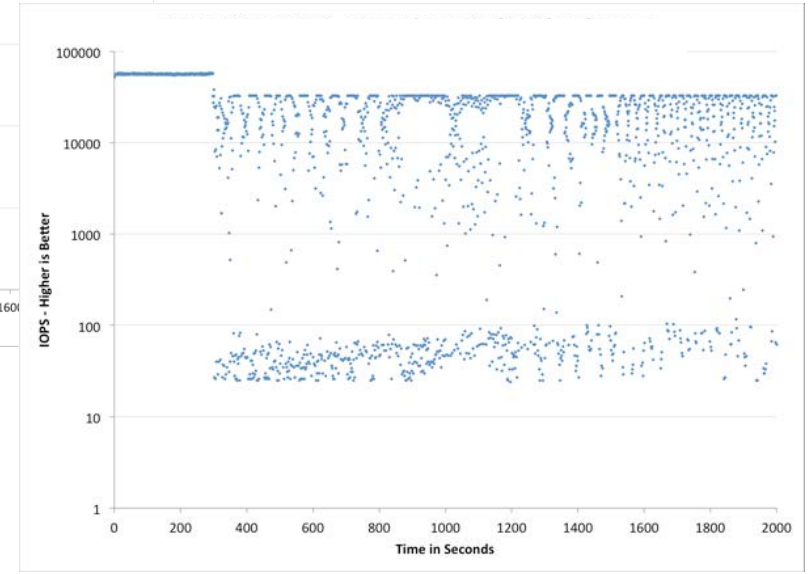
SanDisk's Enterprise SATA SSDs

Name	Sizes	IOPS r/w	Endurance	Application
Eco	240, 480, 960	80K/15K	1 DWPD, 3yr	Read intensive
Ascend	240, 480, 960	75K/14K	1 DWPD, 3yr	Read intensive
Ultra	200, 400, 800	75K/25K	3 DWPD, 5yr	General purpose
Extreme	100, 200, 400, 800	75K/25K	10 DWPD, 5yr	Write intensive

DWPD = Drive Writes Per Day

Consistent Performance is Key





Flash for Acceleration

- There are 31 flavors of flash usage
- What's best for you depends on your:
  - Application mix
  - IOPS demand
  - Tolerance of variable performance
  - Pocketbook
  - Organizational politics

Basic Deployment Models

- SSDs in server as disk
- All solid state array
- Hybrid arrays
  - Sub LUN tiering
  - Caching
- Storage Network Caching
- Server side caching
- Hyperconverged infrastructures

Flash in the Server

- Minimizes latency and maximizes bandwidth
  - No SAN latency/congestion
  - Dedicated controller
- But servers are unreliable
  - Data on server SSD is captive
  - Good where applications are resilient
    - Web 2.0
    - SQL Server Always On
- Software cross-server mirroring
  - But that adds latency to writes

All Flash Array Vendors Want You to Think of This

But Some Are This


Or Worse
This

What You Really Want

Rackmount SSDs

- Our drag racers
  - They go fast but that's all they do
- The first generation of solid state
- Not arrays because:
  - Single Controller
  - Limited to no data services
  - Thankfully dying out

The Hot Rods

- Legacy architectures with SSD replacing HDD
  - NetApp EF550
  - EMC VNX-F
  - Equallogic PS6110s
  - Many 2nd and 3rd tier vendor's AFAs

- Limited performance
  - 50-300,000 IOPS
- Full set of data management features
- Wrong architecture/data layout for flash

All Solid State Arrays

- Minimum dual controllers w/failover
- Even better scale-out
- Higher performance (1 megaIOP or better)
- Better scalability (100s of TB)
- Most have partial data management features
  - Snapshots, replication, thin provisioning, REST, Etc.
- May include data deduplication, compression
  - Lower cost w/minimal impact on performance

Legacy Vendors All Flash Array

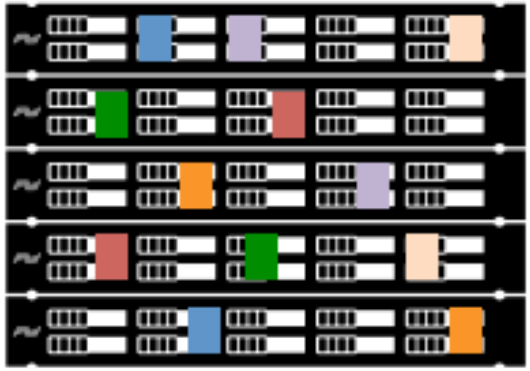
- 3Par and Compellent's data layout better for flash
  - Easier tiering, less write amplification
- Dell - Compellent
  - Mixed flash
    - SLC write cache/buffer, MLC main storage
  - Traditional dual controller
- HP 3Par Storeserv 7450
  - 220TB (Raw)
  - 2-4 controllers

Pure Storage Flasharray//m

- Dual x86 Controllers in canisters 3u
- 300 KIOPS
- SAS shelves, U.2 ports in head
- PCIe NVRAM shared
- 2.75-35TB raw capacity
- Always on compress ar
- FC, iSCSI or FCoE

SolidFire

- Scale out architecture
  - 5 node starter 174TB (raw) 375K IOPS
  - Scale to 100 nodes
- Always on dedupe, compression
- Content addressed SSDs
- Leading storage QoS
- Moving from cloud providers to enterprise
- iSCSI, FC via bridge nodes



EMC XtremIO

- Scale-out Fibre Channel
- X-Brick is 2 x86 servers w/SSDs
- Scales to 8 X-Bricks (but not online)
- Infiniband RDMA interconnect
- Shared memory requires UPS
- Full time dedupe, CAS
- 10-80TB raw

Violin Adds Services

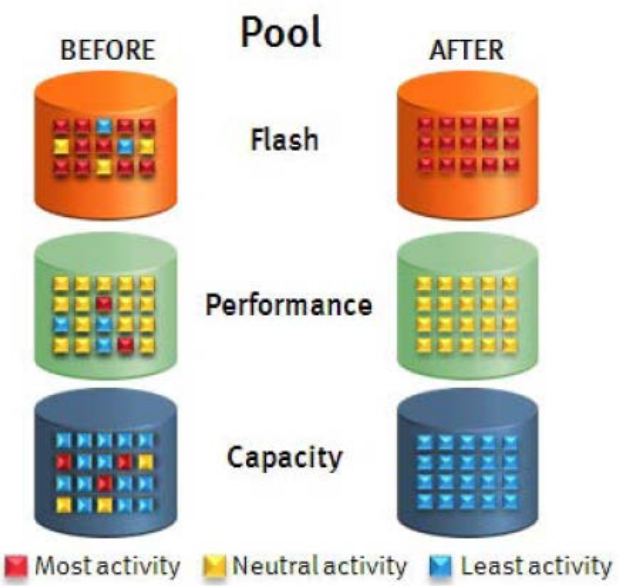
- Violin was market leader in hotrod era
- That's not enough
- Windows Flash Array - WSS on 6000 array
- Concerto 7000 storage routers ala Whiptail
  - Snapshots, replication Etc. via Falconstor
  - Scale to 280TB
- Unique flash modules
  - PCIe switched
  - Better consistency

Hybrid Arrays

- Combine flash and spinning disk in one system
  - Usually 7200RPM
- Legacy designs with SSDs added
- Next-Gen Hybrids
  - Tegile - Nimble
  - NexGen - Tintri
- High performance
  - 20,000 IOPS or more from 3-4u
  - 10% flash usually provides 2-4x performance boost
- Typically include deduplication, compression, virtualization features

Sub-LUN Tiering

- Moves “hot” data from slow to fast storage
- Only 1 copy of data
- Must collect access frequency metadata
- Usually on legacy arrays
- Ask about granularity, frequency
  - Up to 1GB, once a day
- Can give unpredictable performance



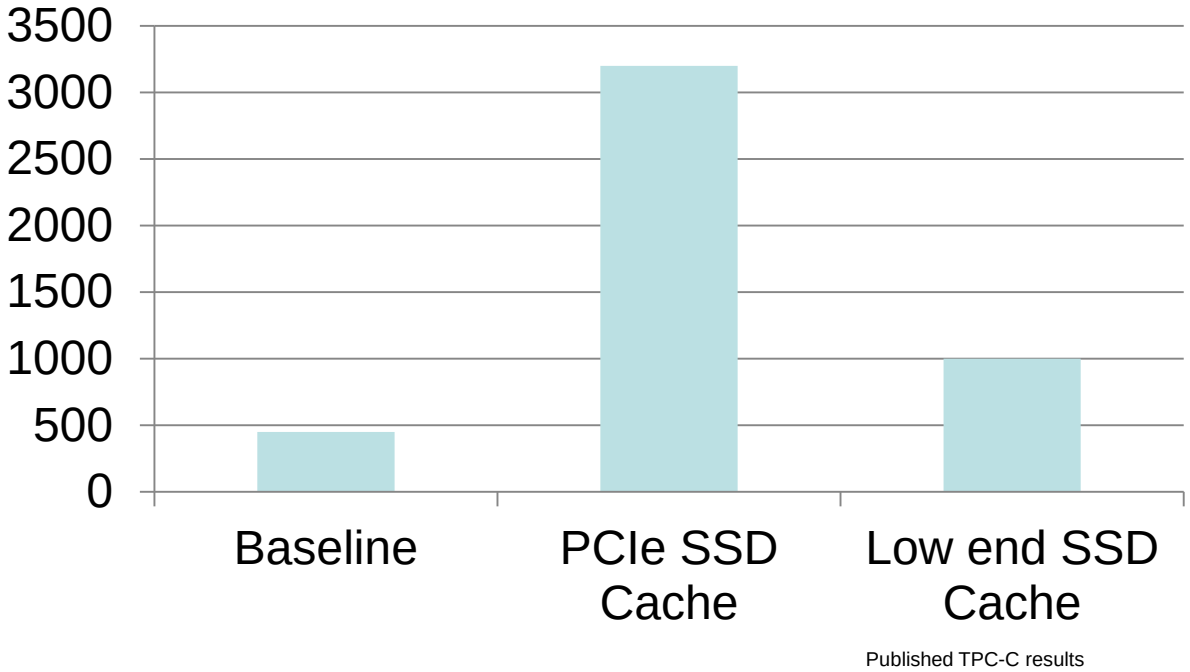
Flash Caching

- Data copied to flash on read and/or write
- Real time
- Write around
  - Reads cached
- Write-through cache
  - All writes to disk and flash synchronously
  - Acknowledgment from disk
- Write back cache
  - Write to flash, spool to disk asynchronously

Server Flash Caching Advantages

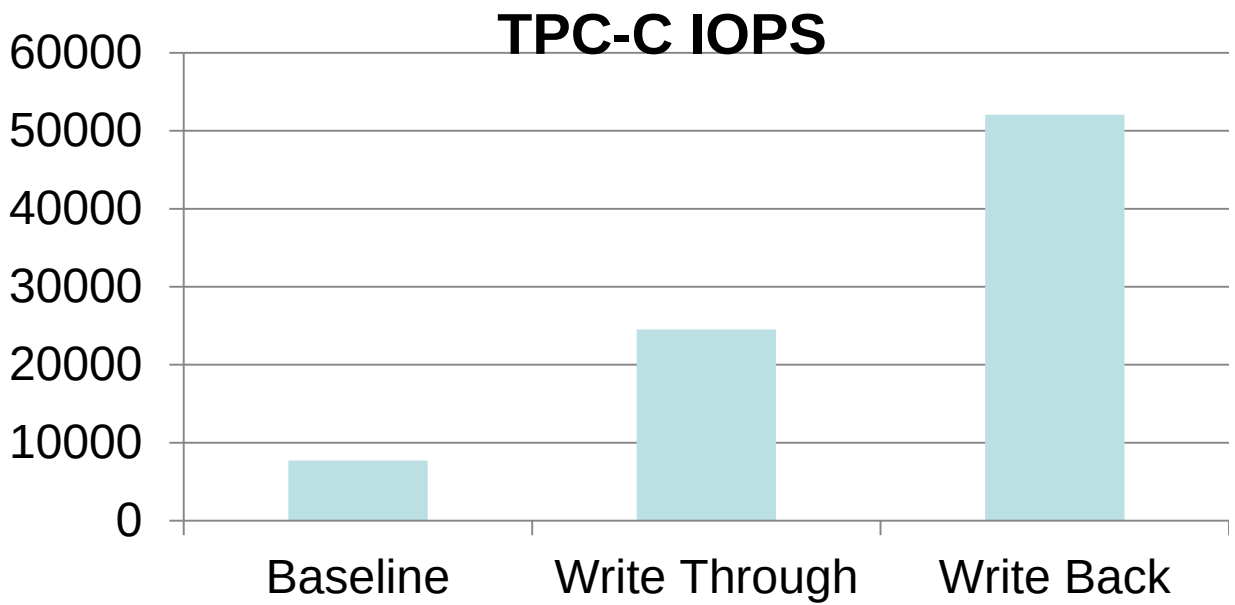
- Take advantage of lower latency
  - Especially w/PCIe flash card/SSD
- Data written to back end array
  - So not captive in failure scenario
- Works with any array
  - Or DAS for that matter
- Allows focused use of flash
  - Put your dollars just where needed
  - Match SSD performance to application
  - Politics: Server team not storage team solution

Caching Boosts Performance!



Published TPC-C results

Write Through and Write Back



- 100 GB cache
- Dataset 330GB grows to 450GB over 3 hour test

Server Side Caching Software

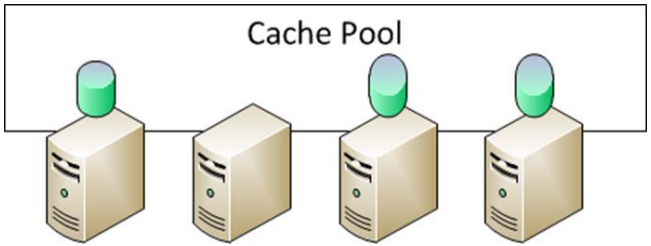
- Over 20 products introduced
- Some best for physical servers
  - Windows or Linux
- Others for hypervisors
  - Live migration/vMotion a problem
- Most provide write through cache
  - No unique data in server, only accelerates reads
- Duplicated, distributed cache for write back
- Applications cache too
  - SQL Server

Live Migration Issues

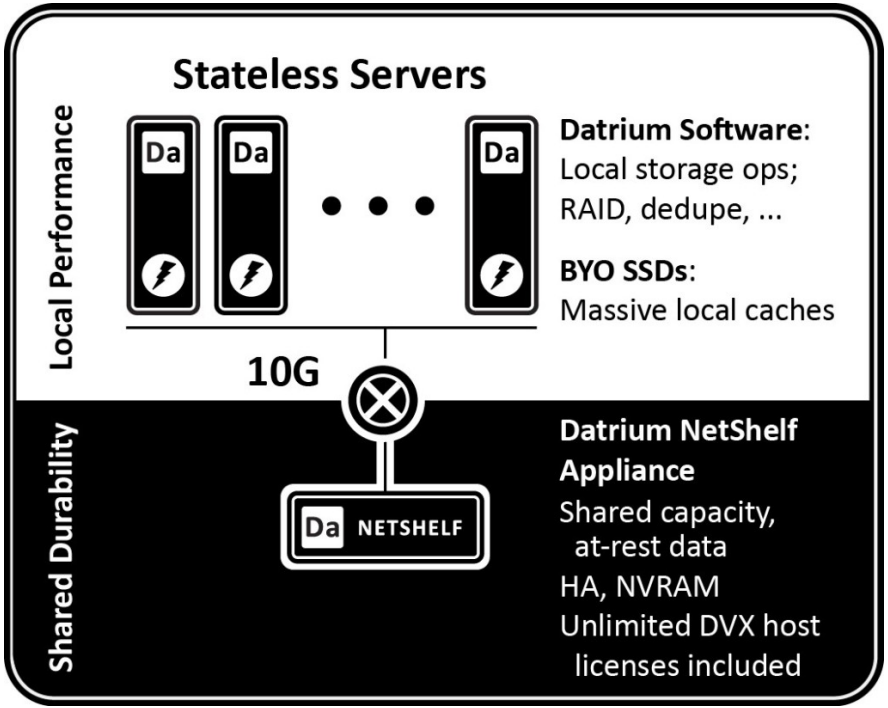
- Does cache allow migration
  - Through standard workflow
    - To allow automation like DRS?
- Is cache cold after migration?
- Cache coherency issues
- Guest cache
  - Cache LUN locks VM to server
    - Can automate but breaks workflow
- Hypervisor cache
  - Must prepare, warm cache at destination

Distributed Cache

- Duplicate cached writes across n servers
- Eliminates imprisoned data
- Allows cache for servers w/o SSD
- Solutions
  - PernixData
  - Dell Fluid Cache



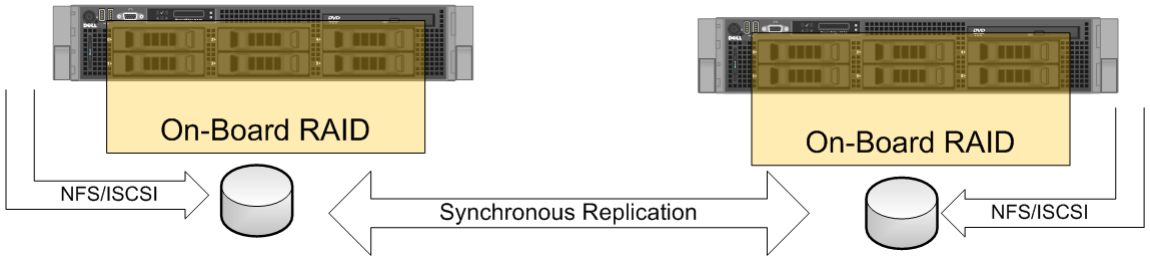
Datrium DiESL



- Host managed cache
- PCIe SSD in Host
  - Write through cache
- All flash NetShelf
  - Persistent layer
- NFS interface to vSphere
  - Per-VM data services
- Founders from Data Domain
  - Dedupe of course

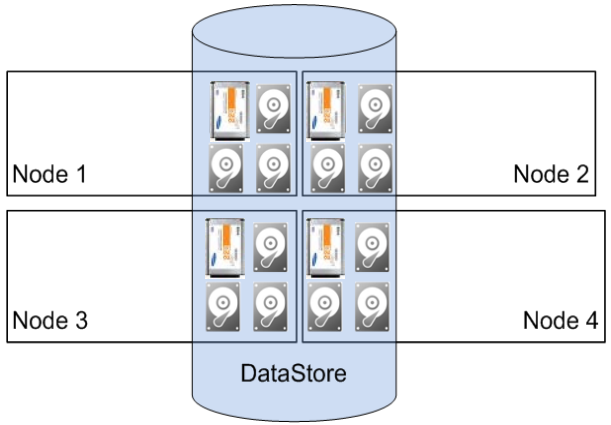
Virtual Storage Appliances

- Storage array software in a VM
- iSCSI or NFS back to host(s)
- Caching in software or RAID controller
- Players:
  - ~~VMware~~
  - StoreMagic
  - HP/Lefthand
  - Nexenta



Hyperconvirged Infrastructure (ServerSAN)

- Use server CPU and drive slots for storage
- Software pools SSD & HDD across multiple servers
- Data protection via n-way replication
- Can be sold as hardware or software
  - Software defined/driven



ServerSAN Products

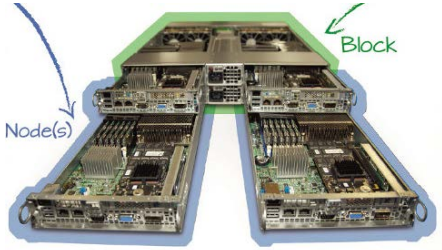
- VMware's VSAN
  - Scales from 4-32 nodes
  - 1 SSD, 1 HDD required per node
- Maxta Storage Platform
  - Data optimization (compress, dedupe)
  - Metadata based snapshots
- EMC ScaleIO
  - Scales to 100s of nodes
  - Hypervisor agnostic
- Atlantis Computing ILIO USX
  - Uses RAM and/or Flash for acceleration
  - Works with shared or local storage

ServerSAN Architecture Differentiators

- Data protection model
  - Per node RAID?
  - *N*-way replication
  - Network RAID?
- Flash usage:
  - Write through or write back cache
  - SubLUN tiering
- Prioritization/storage QoS
- Data locality
- Data reduction
- Snapshots and cloning

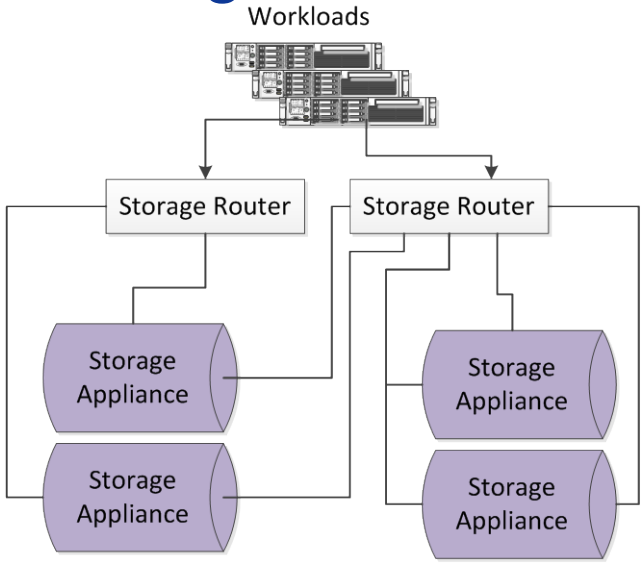
Hyper-converged Systems

- Nutanix
  - Derived from Google File System
  - 4 nodes/block
  - Multi-hypervisor
  - Storage for cluster only
- Simplivity
  - Dedupe and backup to the cloud
  - Storage available to other servers
  - 2u Servers
- No 20 other vendors incl. VMware's EVO:RAIL



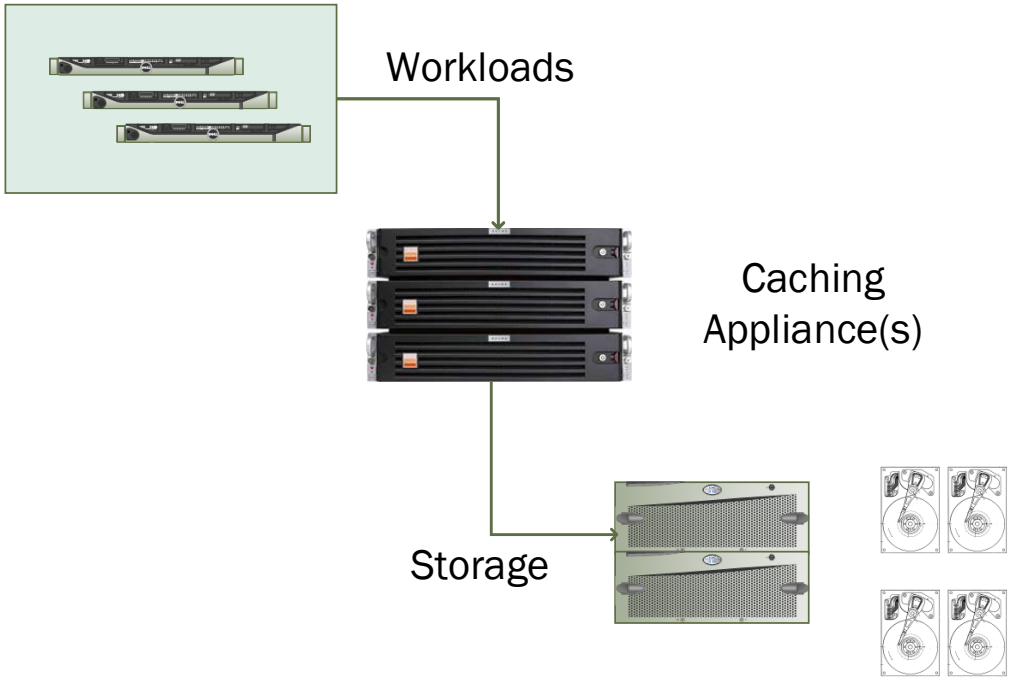
Questionable Idea 1: Smart Shelves

- How rack mount SSDs become arrays
- Adds data services/scale out in storage router
- Players:
  - Cisco Whiptail
  - Violin
  - XIO



Questionable Idea #2: Storage Network Caching

- "bump in the wire" cache
- Several vendors offered Fibre Channel versions
  - All discontinued
- NAS versions
  - Flopped in general market
  - Work as cloud gateways
    - Avere
  - HPC
    - DDN

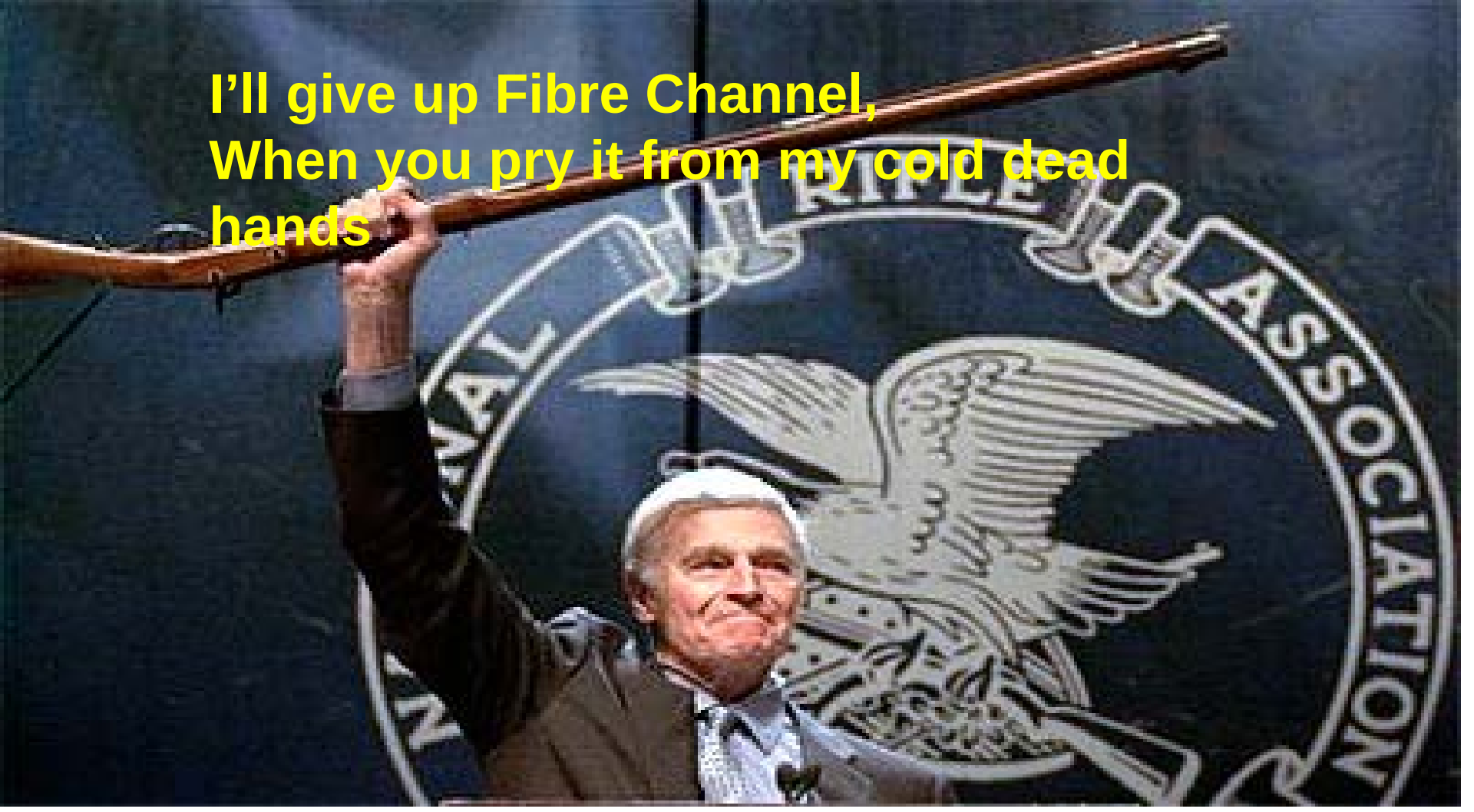


All Flash Array?

- If you need:
  - More than 75,000 IOPS
  - For one or more high ROI applications
- Expect to pay $4-7 GB
- Even with dedupe
- Think about data services
  - Snapshots, replication, Etc.

Hybrids

- Hybrids fit most users
  - High performance to flash, capacity from disk
  - All automatic
- Look for flash-first architectures
  - Usually but not always from newer vendors
- Ask about granularity and frequency for tiering
- Again data services
  - Snaps on HDD
  - Per-VM services


I’ll give up Fibre Channel,
When you pry it from my cold dead hands

Server Side Caching

- Decouples performance from capacity
- Strategic use
  - Pernix data write back cache w/low cost array
- Tactical solution
  - Offload existing array
  - Boost performance with minimal Opex

Questions and Contact

- Contact info:
  - Hmarks@deepstorage.net
  - @DeepStoragenet on Twitter