Before We Begin

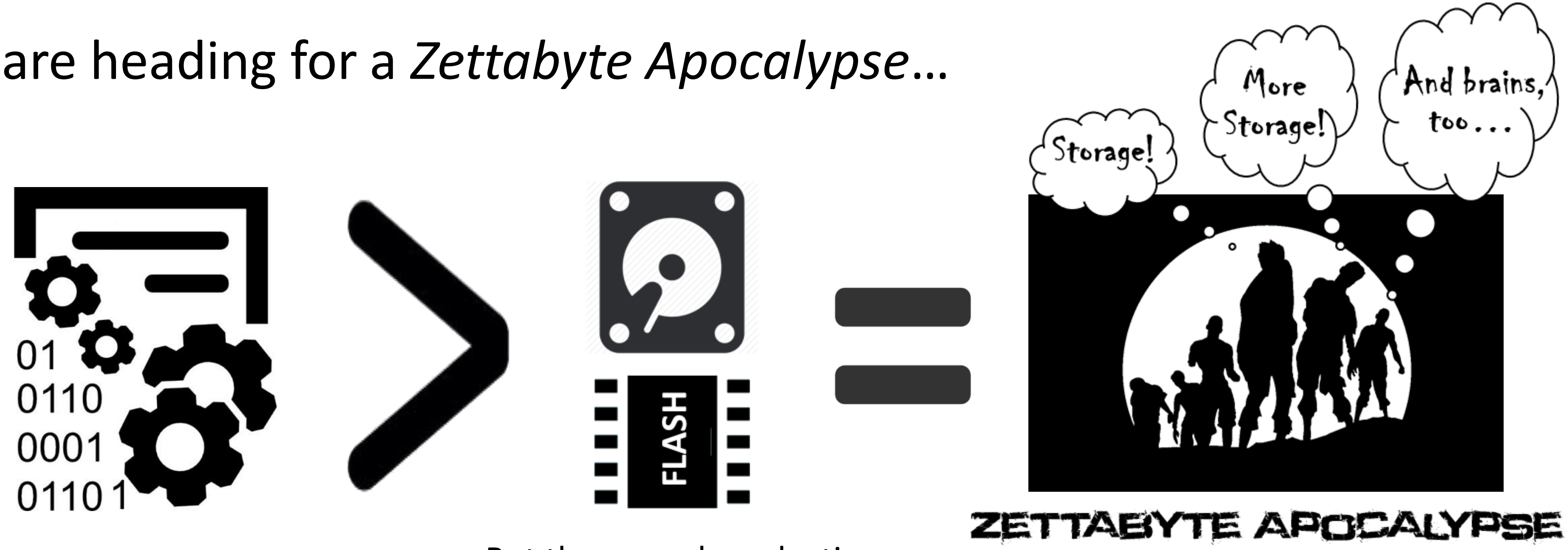
An Bit of Context for
Those Who Missed the Blockbuster
Film of the Summer

THE SECRET LIFE OF BITS

PRESENTED BY
JON TOIGO, CHAIRMAN
DATA MANAGEMENT INSTITUTE

Analysts agree…

- We are heading for a *Zettabyte Apocalypse*…



We will generate between 10 and 60 zettabytes of new data by 2020…

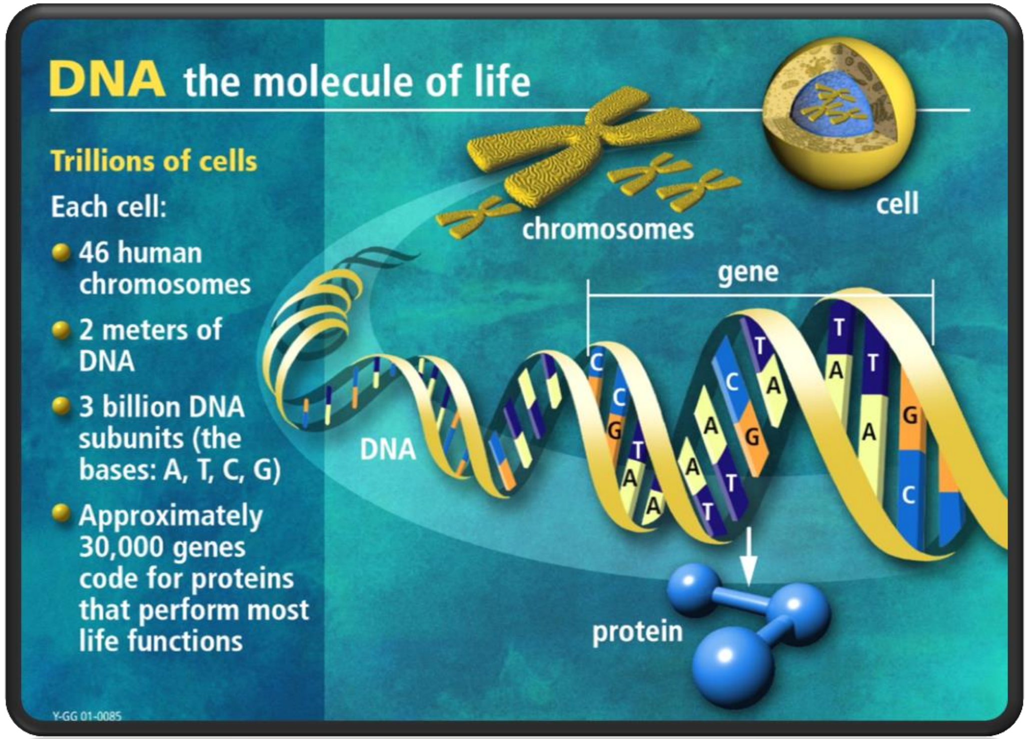
But the annual production of HDD and SSD tops out at less than 1.5 zettabytes of capacity…

What to do?

- Trillions of investment dollars would be needed to double capacity…still insufficient!
- Facebook likes optical disk for archival storage: good luck with that (still waiting for a TB of capacity on BluRay)
- First forays into exotic storage from Microsoft and others…*like DNA*

Yes. You heard correctly.

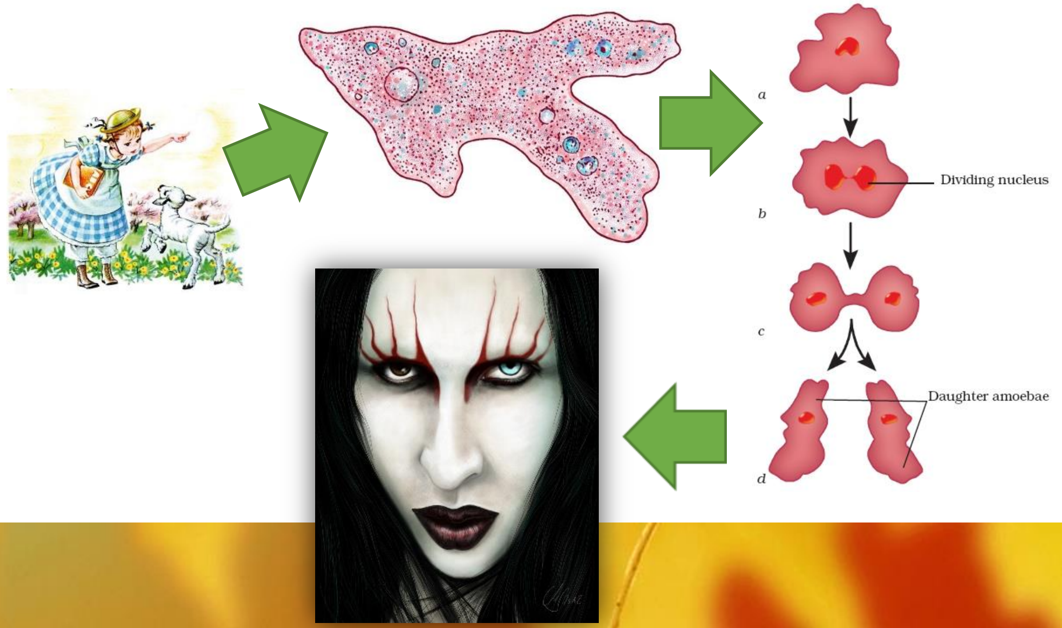
- *Dioxyribonucleic Acid.* The stuff of life.
  - Microsoft has actually invested in Twist Bioscience R&D to figure out how to store data in DNA
  - “DNA data storage could last up to 2,000 years without deterioration.”
  - “Furthermore, and perhaps more importantly for the exponential digital data deluge we are facing, *‘a single gram of DNA can store almost one trillion gigabytes (almost a zettabyte) of digital data’.”*



http://hexus.net/tech/news/storage/92486-microsoft-buys-synthetic-dna-digital-data-storage-research/

Not really so far fetched…

- It only takes a teaspoon of genetic material to make a human…
- Early experiments have written “Mary had a little lamb” into the DNA of amoebas. It was retrievable when the organism reproduced…but degraded
- Scientists have decided that mutation was to blame and could be avoided by using more advanced lifeforms…

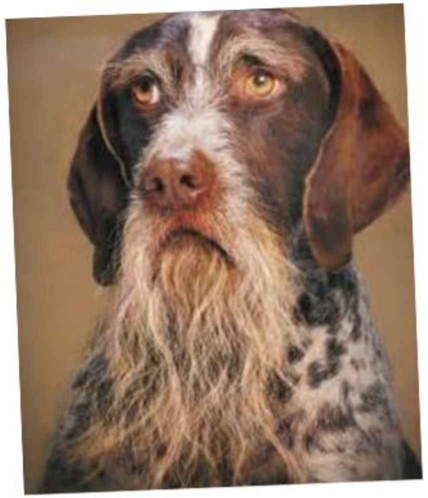


The mind boggles…

Portable Storage

Flash Storage

Heterogeneous Storage



Archival Storage

Redundant Portable Storage

Capacity Storage

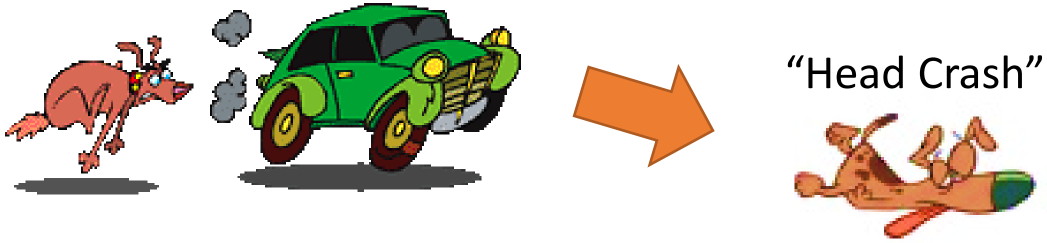
Secure Storage

Hyper-Converged Storage

Software-Defined Storage

But the strategy has its foibles…

- Data loss still possible, of course
- And the commercialization of the technology is still quite a few years off
- Perhaps, it would be better to focus on encoding our genetic material with more propensity for *common sense!*



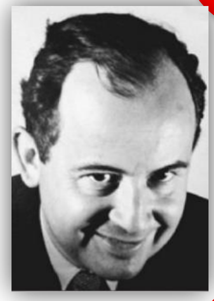
Looking back over 2016, it could be argued

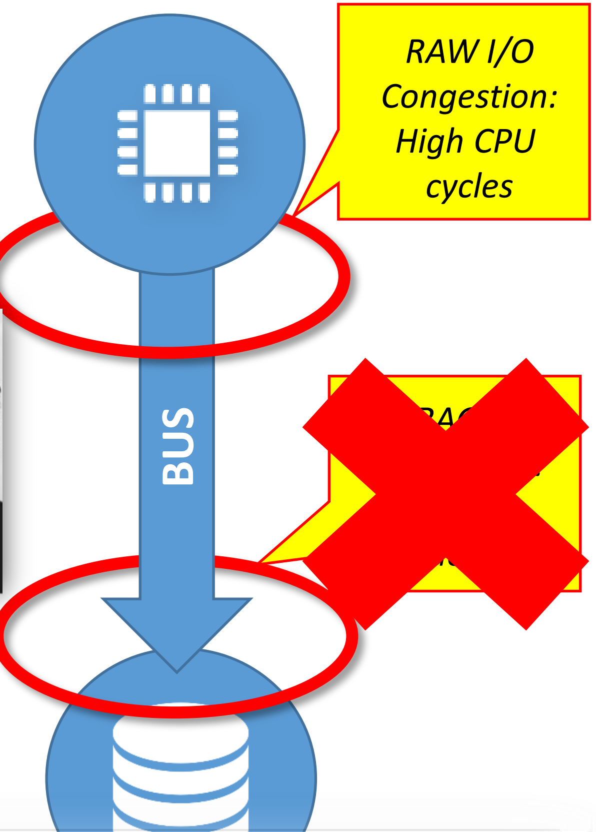
- That IQ's have dropped sharply in a lot of IT departments…
  - Lack of understanding of the I/O bus makes practitioners vulnerable to tech vendor woo
  - Technology tribalism leading to bad storage architecture decisions
  - Absence of even a basic knowledge of storage technology history and a tendency to regard cloud as a storage game changer
  - Tendency to favor tactical quick fix over strategic solutions
  - Triumph of marketecture over architecture: consumers flock to the "shiny new thing" (genetic programming to seek clean water)

Selling Maslow,
Not Multiprocessing

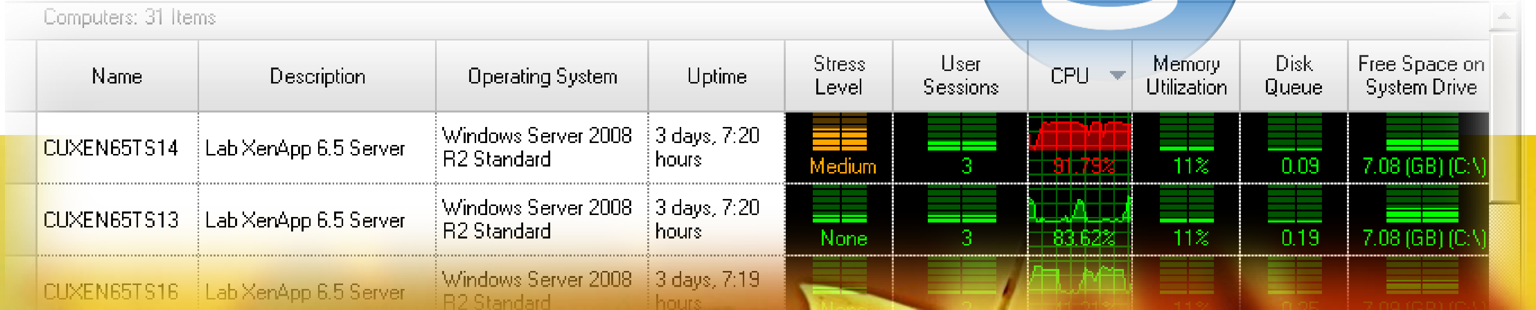
Let’s start with the I/O bus…

- Summarizing my report from last year in one slide…
  - Storage I/O is the end product of a chain of events that starts with the processing of RAW I/O by the CPU and its introduction onto the bus
  - Storage I/O congestion is signaled by storage queue depth; RAW I/O congestion is signaled by high CPU cycling rates
  - In most cases of slow app performance, there is no queue depth – hence, storage is not the chokepoint
  - But storage vendors (and hypervisor peddlers) still sell faster storage kit on the promise of faster performing applications





RAW I/O Congestion: High CPU cycles



Computers: 31 Items

Name	Description	Operating System	Uptime	Stress Level	User Sessions	CPU	Memory Utilization	Disk Queue	Free Space on System Drive
CUXEN65TS14	Lab XenApp 6.5 Server	Windows Server 2008 R2 Standard	3 days, 7:20 hours	Medium	3	91.79%	11%	0.09	7.08 (GB) (C:\)
CUXEN65TS13	Lab XenApp 6.5 Server	Windows Server 2008 R2 Standard	3 days, 7:20 hours	None	3	83.62%	11%	0.19	7.08 (GB) (C:\)
CUXEN65TS16	Lab XenApp 6.5 Server	Windows Server 2008 R2 Standard	3 days, 7:19 hours	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Further validated by benchmarks of Adaptive Parallel I/O from DataCore…

- Storage Performance Council™ SPC-1™ Benchmarking…
  - Previewed last year, published in December 2015: **459,290.87** SPC-1 IOPS™
  - Second benchmark of HA version, January: **1,510,090.52** SPC-1 IOPS™ (withdrawn)
  - Third benchmark of FC-attached storage, June: **5,120,098.98** SPC-1™ IOPS

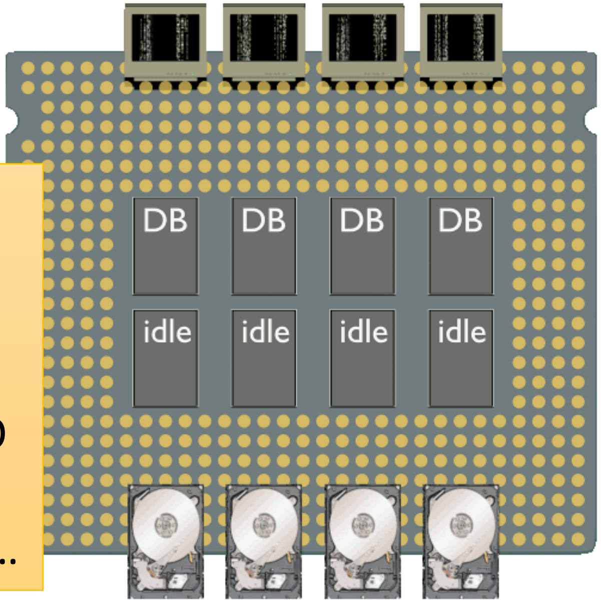
http://www.theregister.co.uk/2016/06/13/datacore_dominating_spc1_benchmark_on_priceperformance/

http://www.theregister.co.uk/2016/06/24/spc_says_up_yours_datacore/

http://www.theregister.co.uk/2016/06/15/datacore_drops_spc1_bombshell/

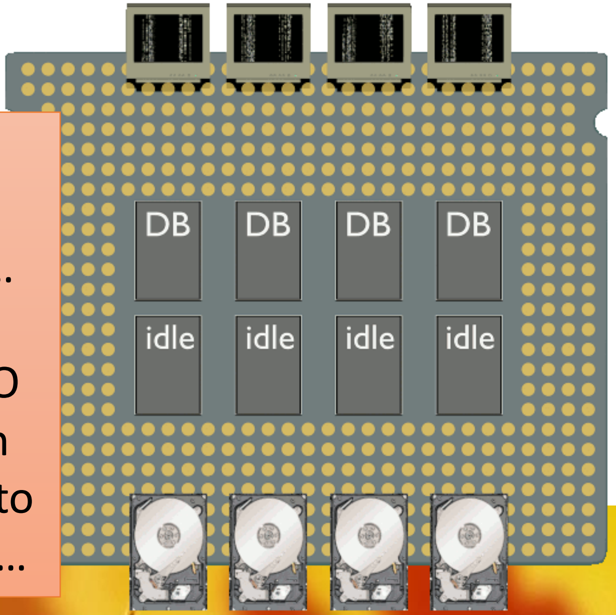
Without Adaptive Parallel I/O…

Database I/O processed sequentially…



With Adaptive Parallel I/O…

Database I/O processed in parallel, up to 300% faster…



Are we learning?

- DataCore results stick a hot poker in the eye of many tacit assumptions in industry marketing…
  - Slow performing workload usually RAW I/O bound: changing storage out for faster kit changes nothing
    - Goes for changing SATA SSD for NVMe flash
    - Goes for changing SAS disk for All Flash Arrays (AFA)
    - Goes for replacing SAN/NAS for internal/DAS storage
  - Converging/hyper-converging storage with servers means little to nothing from an app performance standpoint
  - Fibre Channel is not dead (the 5+ million IOPS result used less than half the bandwidth available on FC link connecting the storage kit to the server)

MAGICAL VIRTUAL SAN ELIXER
IMPROVED AGILITY
ENHANCED I/O
STORAGE LIKE IT WAS NEW!
Dr. VMWARE'S
SOFTWARE DEFINED STORAGE TONIC
FULLY COMPATIBLE WITH ALL ESX HYPERVISORS
HIGHLY AVAILABLE STORAGE INFRASTRUCTURE FOR YOUR DATA
3 NODE CLUSTER
SOOTHING SYRUP FOR NERVE AND BRAIN
HYPER-CONVERGENCE BITTERS
LEGACY STORAGE REMOVAL TONIC
CURE AND PREVENTIVE
EFFECTIVE ON SLOW APPLICATIONS SCALING AND CLUSTERING
SOFTWARE DEFINED
SNAKE OIL
PATENTED CURE FOR HYPER-CONVERGENT HYPOCONDRIA
IMPROVE PERFORMANCE
SIMPLIFY INFRASTRUCTURE
REDUCE COSTS

Of course, not everyone has heard of Parallel I/O or SPC-1 benchmarks…

- Hypervisor computing has gone tribal…
- Tattoos and colors soon to follow…

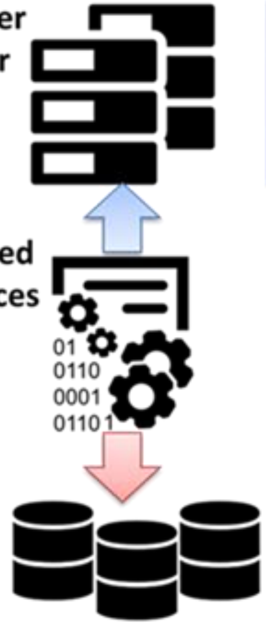
And each one has its own SDS stack…

- SDS = Software-Defined Storage

VALUE-ADD SOFTWARE (ON ARRAY CONTROLLER)

LEGACY STORAGE ARRAY

Services Controlled by Array–Based Software



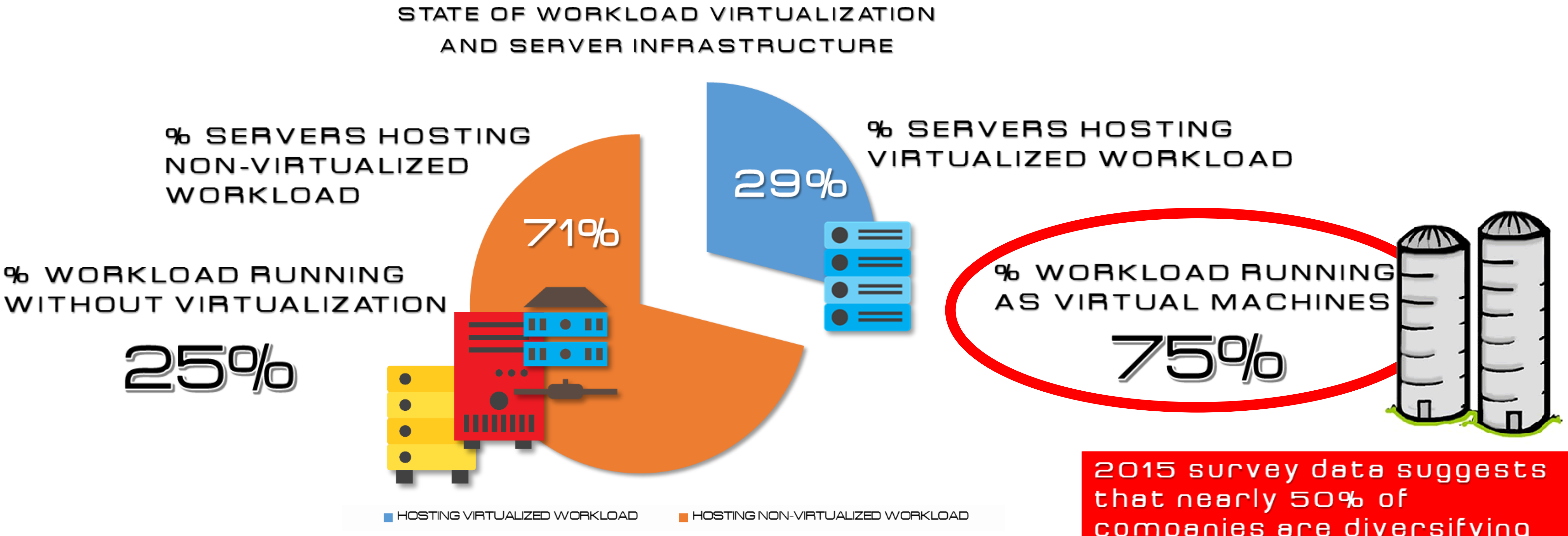
Services Controlled by Server –Based Software

SOFTWARE-DEFINED STORAGE STACK

SERVER

SDS abstracts value-add software services away from storage controller and instantiates services in a server-side software stack…

Leading to “mixed storage outcomes” in most environments…



2015 survey data suggests that nearly 50% of companies are diversifying their hypervisor choices…

A two-edged sword…

- According to Gartner…

Utilization efficiency down by almost 10% since 2011

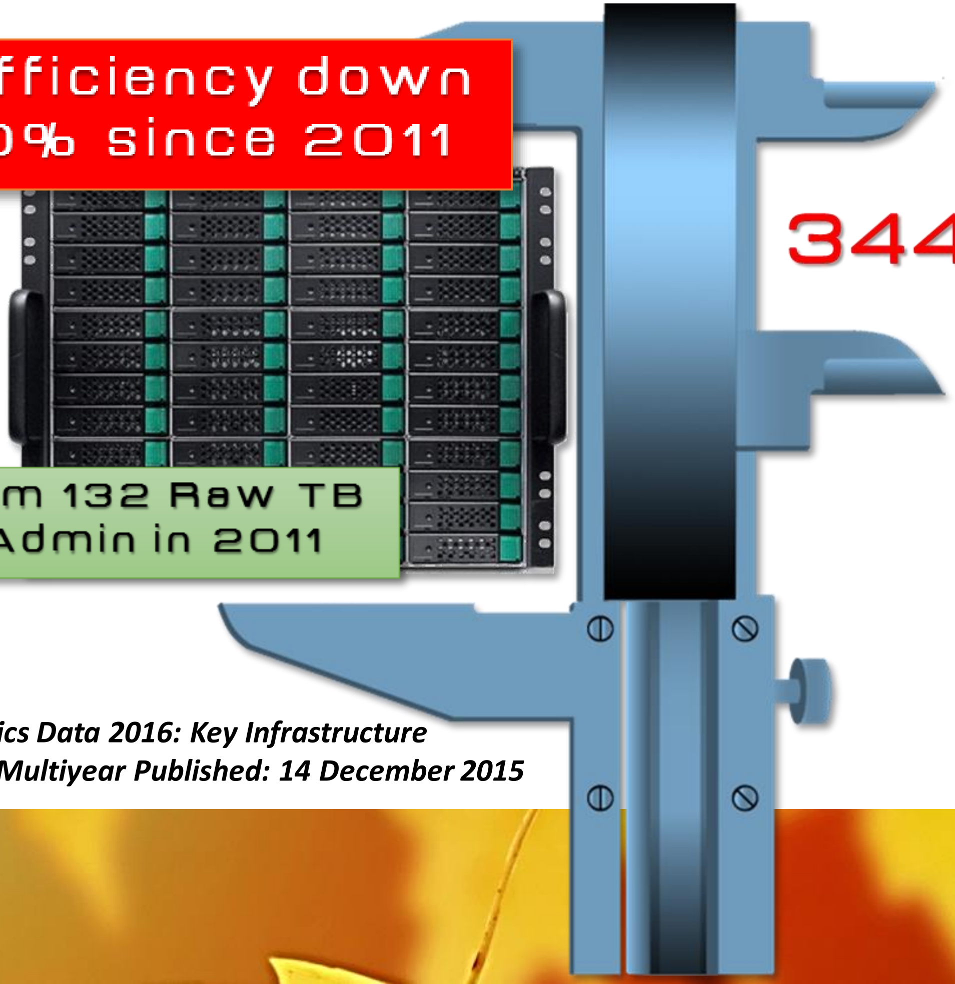
ANNUAL COST PER RAW TB

[NOT INCLUDING FACILITY COSTS]

$2009 (€1748)

Down by more than 50% since 2011

AVERAGE RAW TBs MANAGED PER STORAGE ADMINISTRATOR



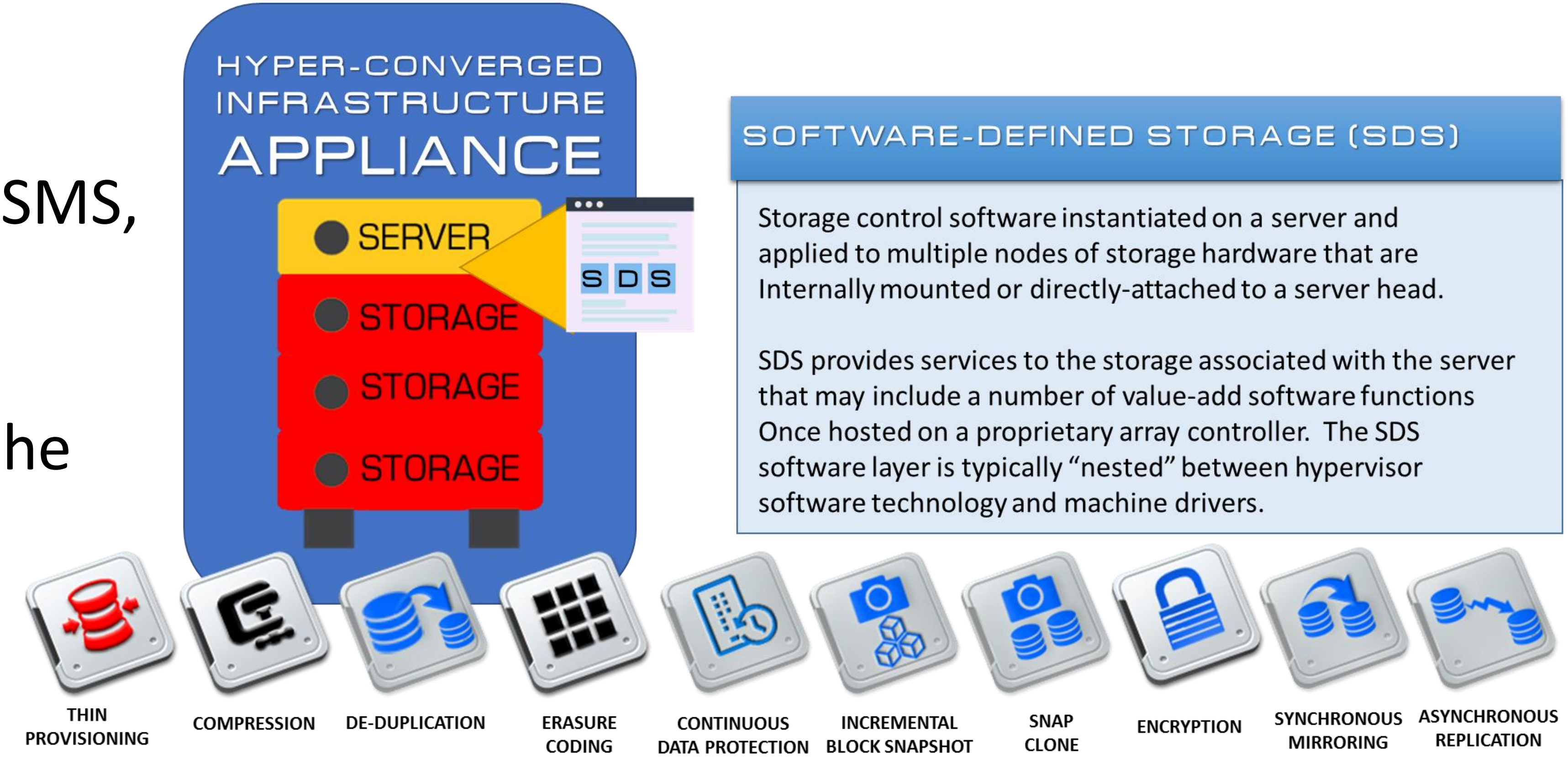
344

Up from 132 Raw TB per Admin in 2011

Source: Gartner, IT Key Metrics Data 2016: Key Infrastructure Measures: Storage Analysis: Multiyear Published: 14 December 2015

But SDS and Hyper-Converged Infrastructure are the *shiny new things* in storage…

- Well, sort of new…
- Actually, System Managed Storage (SMS, c. 1993) was SDS…
- HCI is simply the *appliantization* of the silo concept



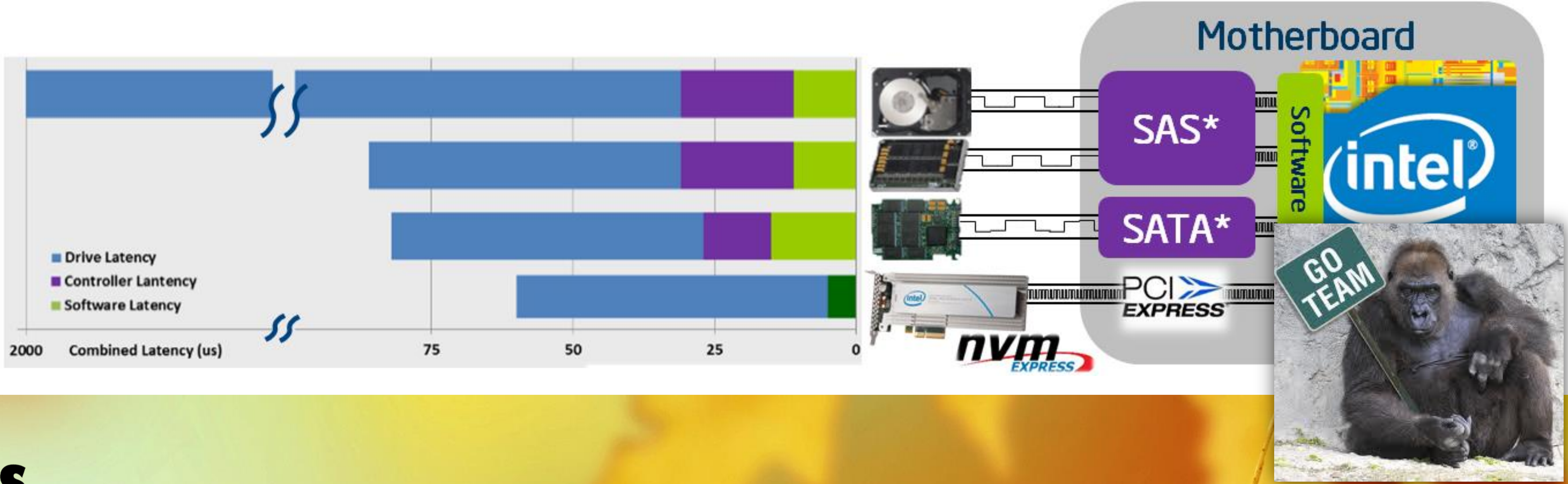
SOFTWARE-DEFINED STORAGE (SDS)

Storage control software instantiated on a server and applied to multiple nodes of storage hardware that are Internally mounted or directly-attached to a server head.

SDS provides services to the storage associated with the server that may include a number of value-add software functions Once hosted on a proprietary array controller. The SDS software layer is typically “nested” between hypervisor software technology and machine drivers.

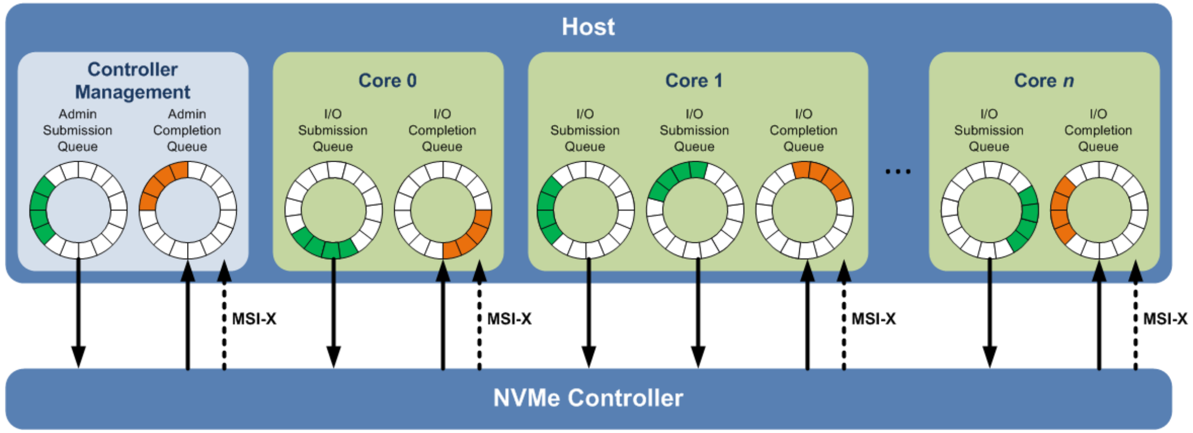
And “Tier 00” (NVMe Flash) is uber-cool…

- No, not SATA SSD (*that’s yesterday’s news*)
- NVMe Flash! (loud cheers, mom’s throw babies into the air, riotous applause…)



Fundamentals of NVMe

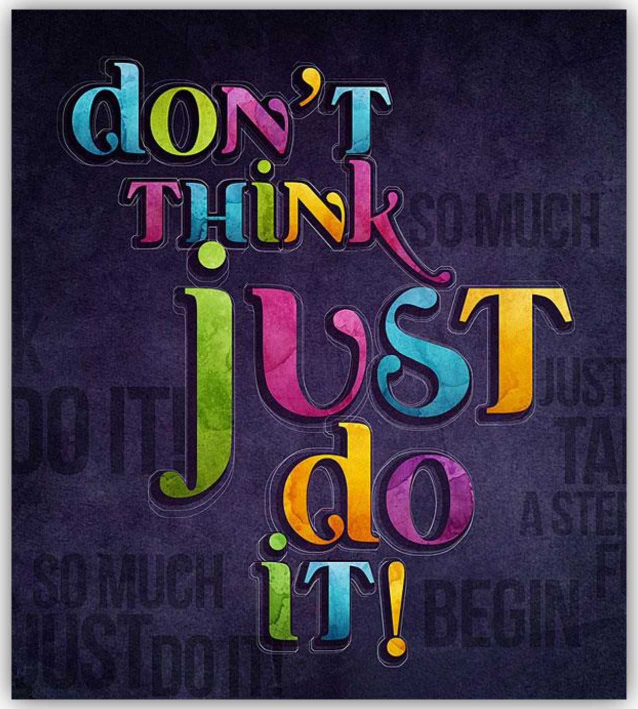
- Provide a PCIe interface for flash modules, eliminating the need to use the SATA controller and proprietary workarounds
- Attacks "bottlenecking" by expediting storage I/O processing
  - Multiple deep queues: SAS/SATA supported 256 commands/32 commands respectively in a single queue; NVMe supports 64K commands per queue and up to 64K queues
  - Eliminates I/O locking
  - Supports MSI-X and interrupt steering
- Uses half the number of CPU instructions to process an I/O request than SAS/SATA: higher IOPS per CPU cycle and lower I/O latency in host software stack



SOURCE: http://www.nvmexpress.org/wp-content/uploads/NVMe_Overview.pdf

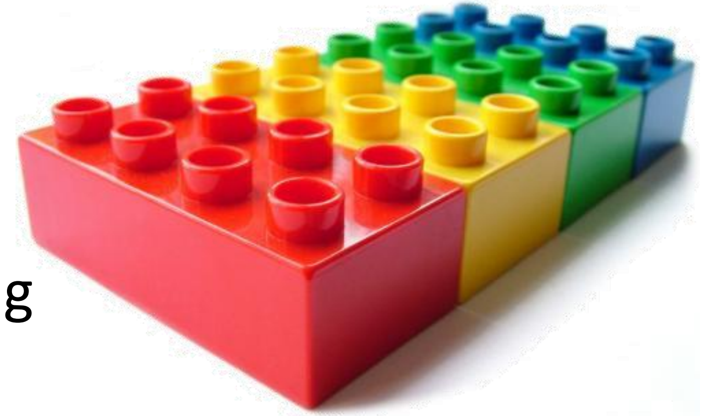
Truth be told: NVMe may be a solution in search of a problem…

- Acceleration of storage I/O being presented as enabler of faster virtual machines, in-memory databases…
  - Value to VM performance is questionable: cause of poor performing VMs rarely associated with STORAGE I/O, but with RAW I/O…which NVMe doesn’t address at all
  - Value to In-memory databases potentially greater, but only in IMDB can leverage NVMe technology
- NVMe does deliver a net overall reduction in latency over SATA-attached SSD, but flash is still not optimized for writes (compared to DRAM)
- Key value must be seen as an enabler of future architectures…



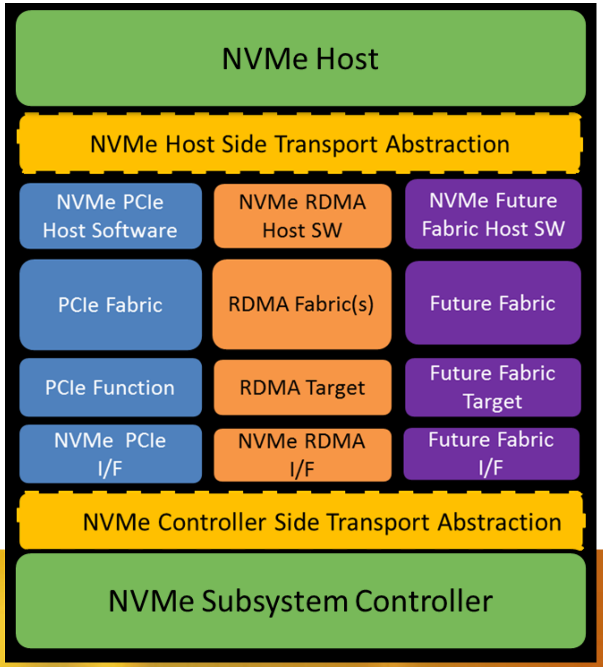
Possibly enabling a new data hosting strategy for in-memory databases?

- Special use case: IMDB
  - In-memory databases have tended to be smallish read-intensive analytics databases: ideal for DRAM and flash, no queries to disk
  - SAP, et al, now seek to platform all databases – including on-line transaction processing (OLTP) DBs in-memory, a more challenging task…
- Could spawn a new class of HCI appliances
  - Leveraging dense DRAM and NVMe flash buffers for all data
  - Enabling a "Lego™-style" building block method for scaling to accommodate very large IMDB…



Perhaps leveraging next gen NVMe over Fabric architectures that are being discussed…

- “SCSI may be too slow in the future…”
- “Usage models will require extreme latency reduction…”
    - Protocol simplicity for automated I/O queue control and NVMe transport bridging
    - No translation to or from another protocol (SCSI)
    - Parallel NVMe mulitiple I/O queues exposed to host
    - Same architecture regardless of fabric type



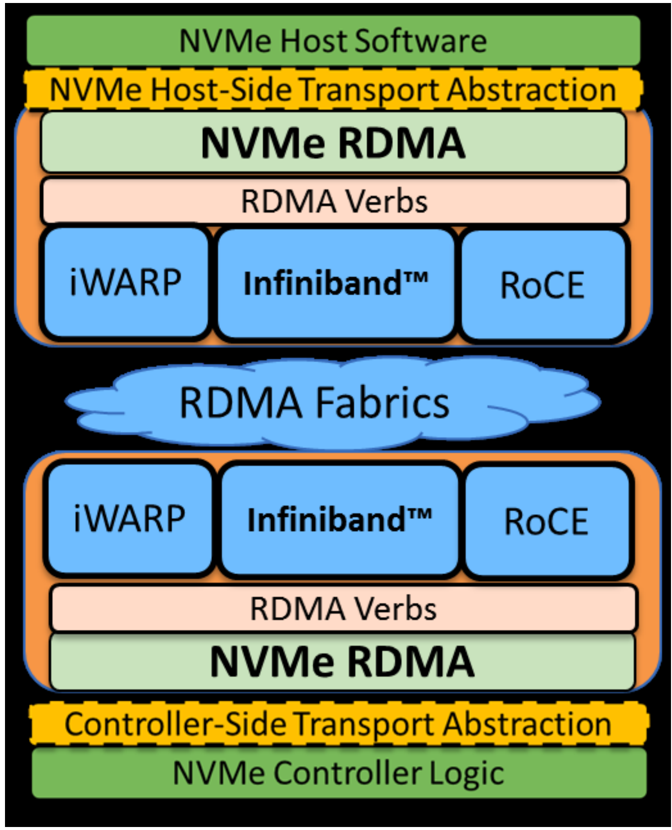
NVMe Host

NVMe Host Side Transport Abstraction

NVMe PCIe Host Software	NVMe RDMA Host SW	NVMe Future Fabric Host SW
PCIe Fabric	RDMA Fabric(s)	Future Fabric
PCIe Function	RDMA Target	Future Fabric Target
NVMe PCIe I/F	NVMe RDMA I/F	Future Fabric I/F

NVMe Controller Side Transport Abstraction

NVMe Subsystem Controller



NVMe Host Software

NVMe Host-Side Transport Abstraction

NVMe RDMA

RDMA Verbs

iWARP | Infiniband™ | RoCE

RDMA Fabrics

iWARP | Infiniband™ | RoCE

RDMA Verbs

NVMe RDMA

Controller-Side Transport Abstraction

NVMe Controller Logic

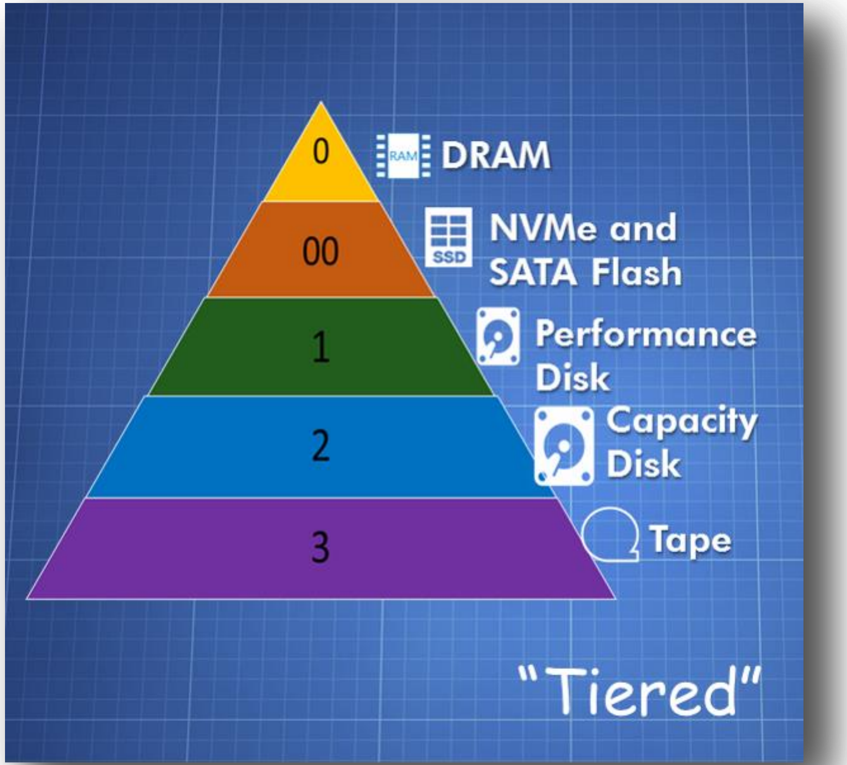
But, for now, the net effect…

- **Lower capacity allocation and utilization efficiency** at exactly the time when we need greater efficiency to reduce capacity demand (remember that *Zettabyte Apocalypse* thing…)
- **Higher storage costs** in the form of SDS node licenses and use of overpriced flash storage when unnecessary
- **Credence given to "flattened" and "friction-less" storage** infrastructure (huh?)

- ***Dogs and cats, sleeping together. Mass-hysteria!***

Flat storage infrastructure?

- Meaning *no tiers*…



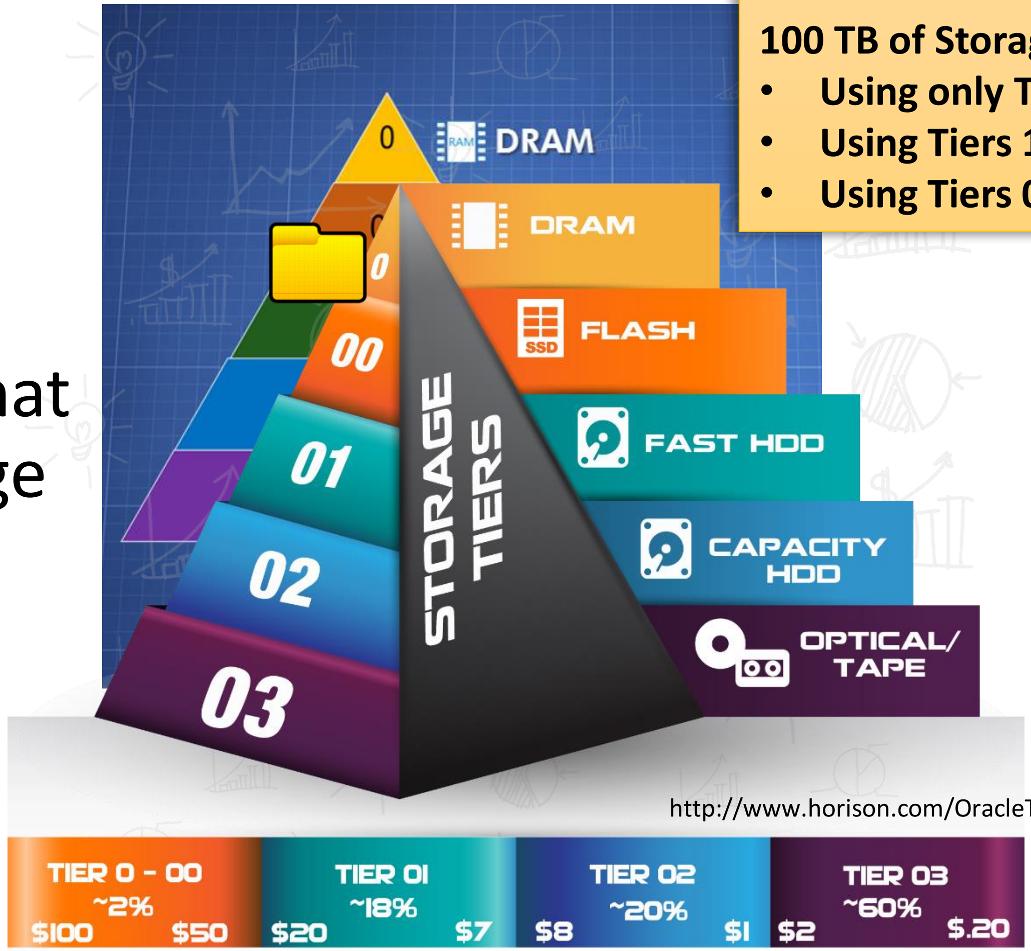
Why would you do such a thing?

- Tiering is intended to reduce storage costs by constantly migrating older data to less expensive forms of storage that are better suited to data usage characteristics…
- So *“day before yesterday”* !

Back of Envelope Math

100 TB of Storage

- **Using only Tiers 1 and 2: $765,000**
- **Using Tiers 1-3: $359,250**
- **Using Tiers 0-3: $482,250**



http://www.horison.com/OracleTieredStorageTakesCenterStage.pdf

Answer: To Eliminate All Friction!

- “Data movement equals friction.”
- “Friction equals latency.”
- “Latency is bad.”
- “Therefore, friction (and data movement) are bad.”

- *Oh, and tiering is hard work.*

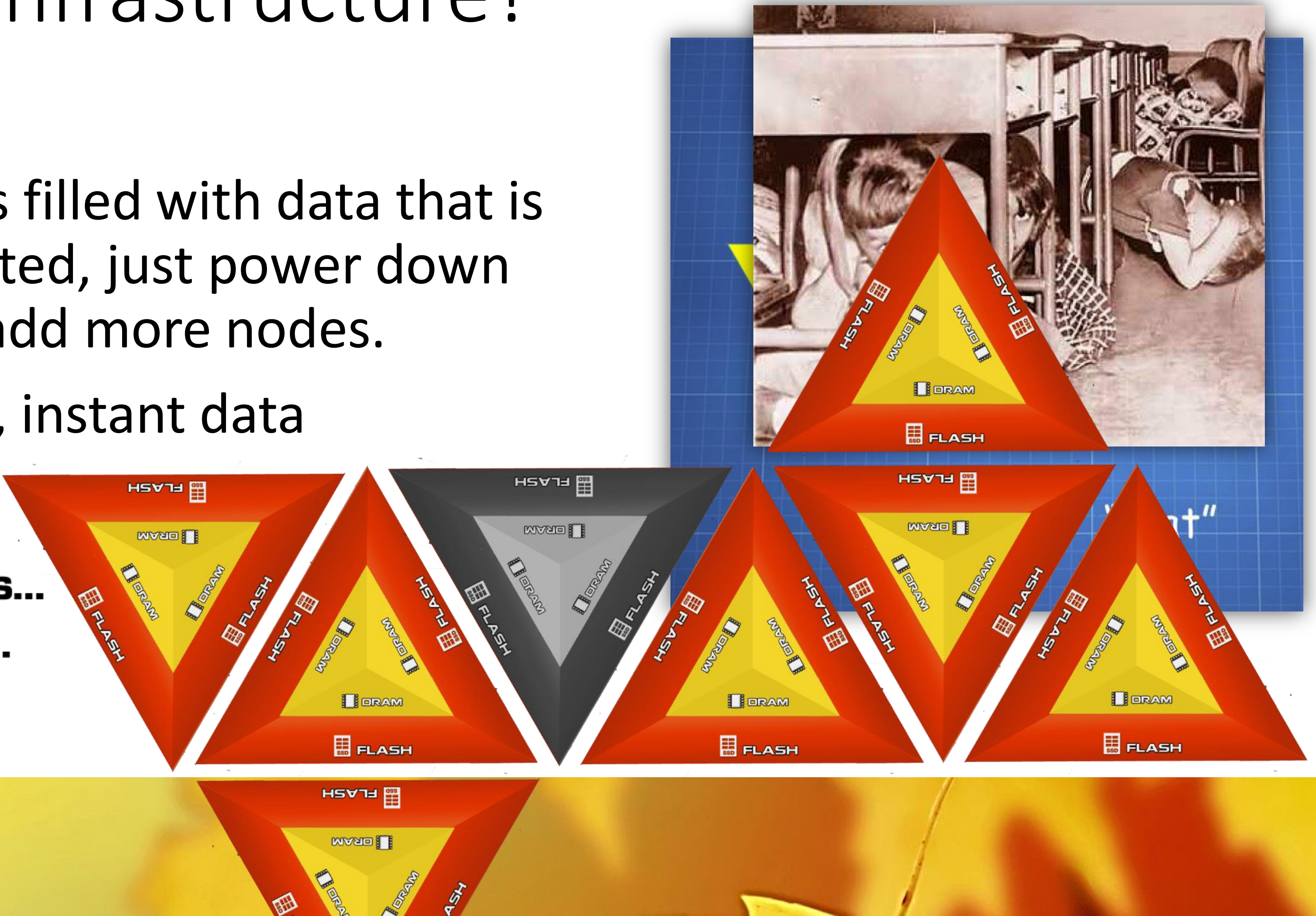
Okay, slick, so how do we preserve data in a “friction-less” infrastructure?

- “Shelter in place.”
- When storage node is filled with data that is not accessed or updated, just power down the drive. Then just add more nodes.
- *Voila!* Instant archive, instant data preservation…

ALL FLASH STORAGE NODES…

NODE FULL: POWER DOWN…

ADD ANOTHER NODE…



Apparently, no consideration given to…

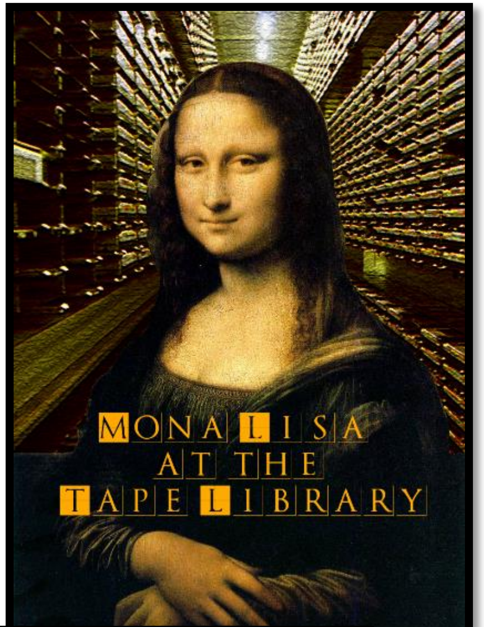
- Failure rates of disk AND SSD when powered down after continuous use…because there is little in the way of published reports
- The possibility of facility or milieu level disasters that could consume both active and powered down nodes
- The cost of infinite nodal replacement (a conceit left over from high performance computing cluster experiments…)

In the final analysis, tape remains a key technology….

- For handling storage capacity demand
- For ensuring data preservation and protection
- For rationalizing storage infrastructure expense
- Regardless of what some of the “cool cats” might say…

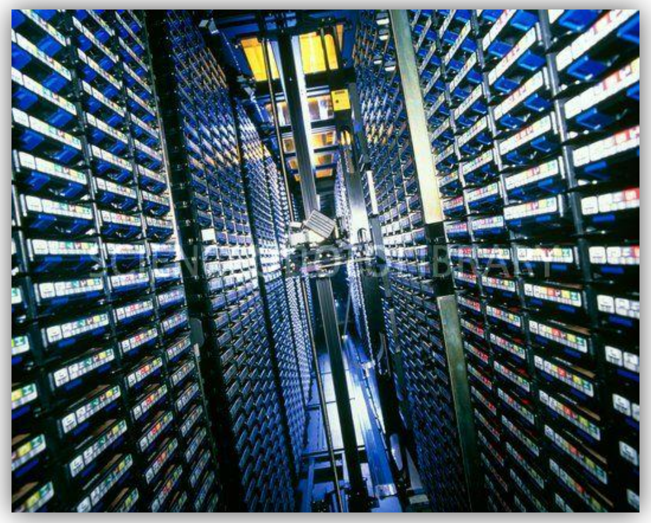
The Tape Renaissance has arrived…

- 75% of world’s data is on tape
- As an archive medium…
  - Capacity improvements outstripping all other kinds of storage: 220TB with BaFe media in LTO cartridge demonstrated in labs
  - Ideal for storing less frequently accessed and modified data
  - Retrieve speed adequate for cloud-based archive and a great modality for “cloud seeding”
  - And using tape is getting much simpler thanks to
    - Linear Tape File System (LTFS)
    - Media Lifecycle Management Automation



“Active” Archive is a key role for tape…

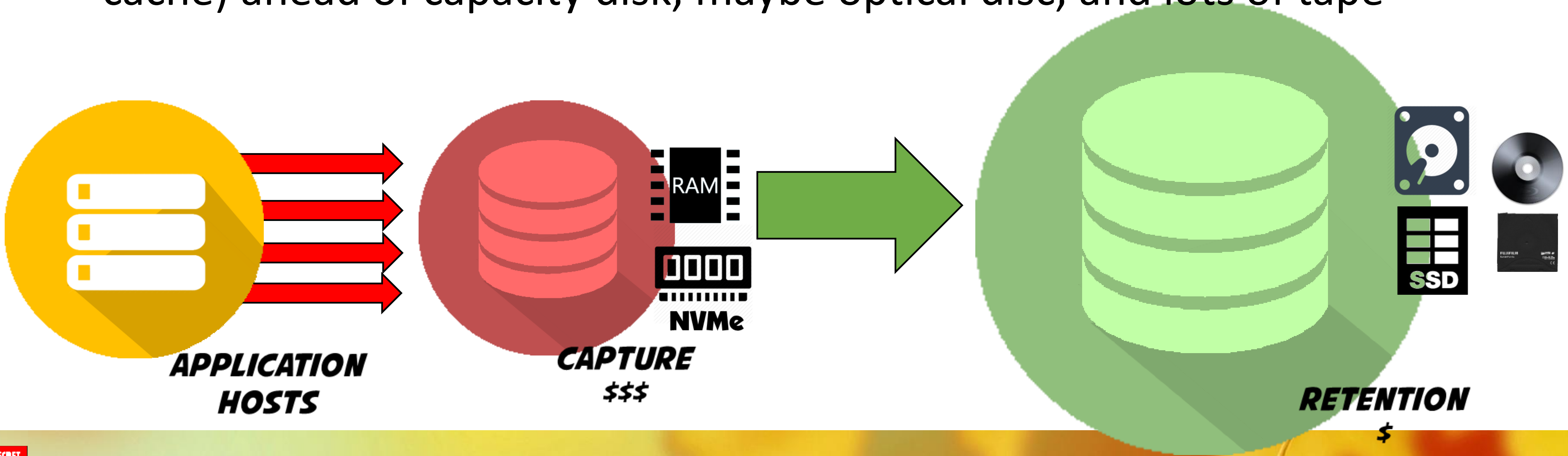
- Leveraging the durability of the media
  - 30 year durability rating
  - Lifespan equal to ~364 full file passes (writing enough data to fully fill the tape, which usually requires between 44 and 136 end-to-end passes)
  - Lifespan can be doubled by writing half of the media capacity
- LTO uses an automatic verify-after-write technology to validate that data has been written (superior to backup software processes that validate after write, increasing the number of end-to-end passes and reducing tape life)
- Plus, LTFS enables the writing of files and objects to tape in their native file/object system construct, eliminating the need for complex archive or backup software in many cases (software that would be needed to read data back from tape)



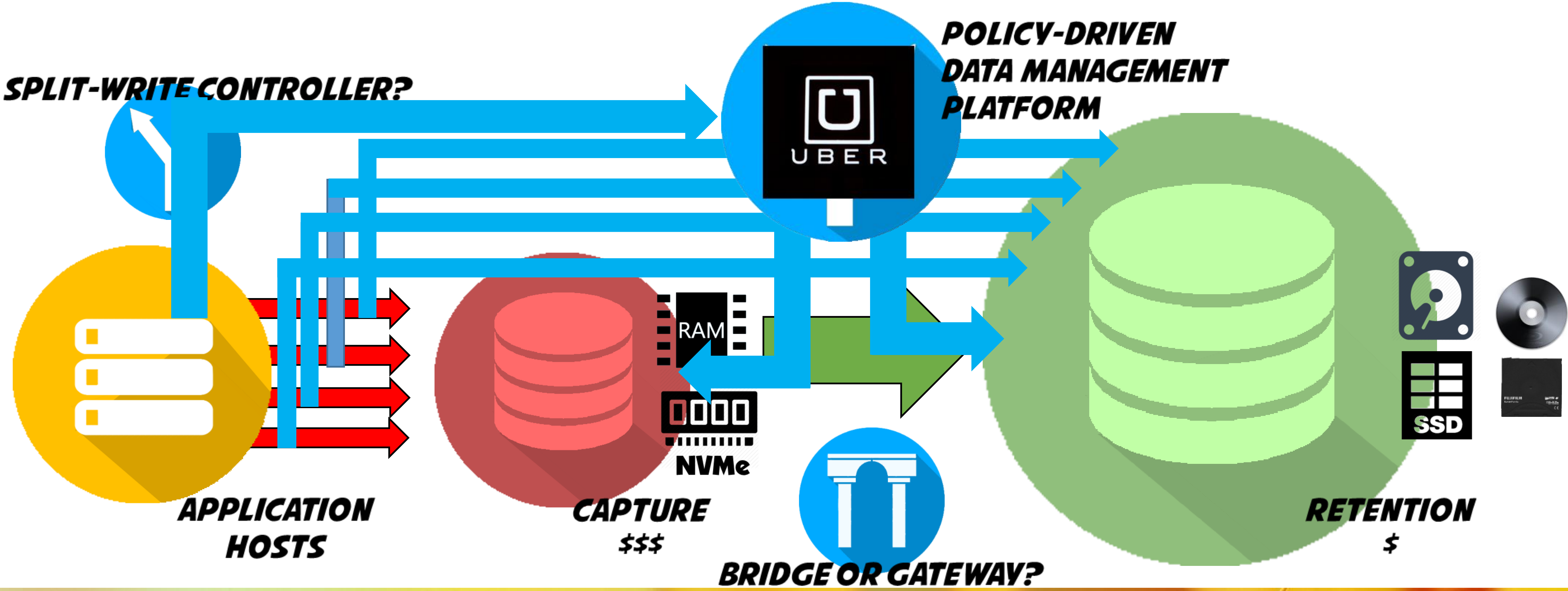
“It was much nicer before people started storing all their personal information in the cloud.”

IMHO: Future storage will require, at a minimum, two storage tiers…

- “Capture storage” – most likely DRAM with a flash chaser
- “Retention storage” – a mix of flash, performance disk (as a buffer or cache) ahead of capacity disk, maybe optical disc, and lots of tape



Question: *How to tier in a cloud-based world?*



Hasty conclusion…

- If I am invited back to next year’s Summit, we should talk about *Cognitive Data Management*
    - Intelligent classification of data
    - Policy-based and automated movement of data across tiers to optimize
        - Accessibility
        - Availability
        - Protection
        - Preservation
        - Privacy
        - Cost
- And, yes, tape will play a critical role

My two *centavos*…

- Questions? Thanks!
- Please keep in touch
  - Email: jtoigo@toigopartners.com
  - Linkedin: Jon Toigo
  - Twitter: @JonToigo
  - Blog: DrunkenData.com
  - Websites:
    - IT-SENSE.org
    - Data Management Institute
    - Toigo Partners International

IT-$ENSE

Data Management Institute

Toigo Partners International