HIGH PERFORMANCE INNOVATION

The Benefits of Tape and How to Maximize It

Nathan Schumann

Instrumental, Inc.

February 5, 2009

Agenda

- Introduction
- Storage Scalability
- Error Rates
- Power and Cooling Costs
- MAID
- Data De-duplication
- Tape Performance
- Final Thoughts

Introduction

Tape Sales

- For years many vendors claiming tape is dead
- Many performance obstacles for tape
  - Same is true for disk, but performance is far better understood
- Tape is often the forgotten child
  - Not because it's not critical to the data center
  - It's not understood
- Tape has many advantages over disk, including
  - Hardware compression
  - Hardware encryption

DR Requirements

- September 11th, 2001 changed everything we all know in terms of how DR is viewed
- Large sites now often
  - Backup disk to disk to tape (D2D2T) locally
  - Replicate to DR D2D
  - Backup the replication at DR to tape
- Tape Impact
  - Often more tape is used as multiple copies are used at both sites

Energy Costs

- What company does not have a green data center initiative?
- Power usage is a big issue and of course getting bigger
- Tape Impact
  - Tape is the most efficient storage in terms of power
  - Disk storage is a large part of the power usage profile for many organizations

Tape is Not Passé

- With disk drive density increasing, some see tape technology as passé
  - It is not and will never be - we'll review why
- Tape has its places in the tiered storage of every data center
  - From SMB to enterprise
- The argument that tape is not needed cannot stand up to critical analysis
- MAID cannot currently replace tape

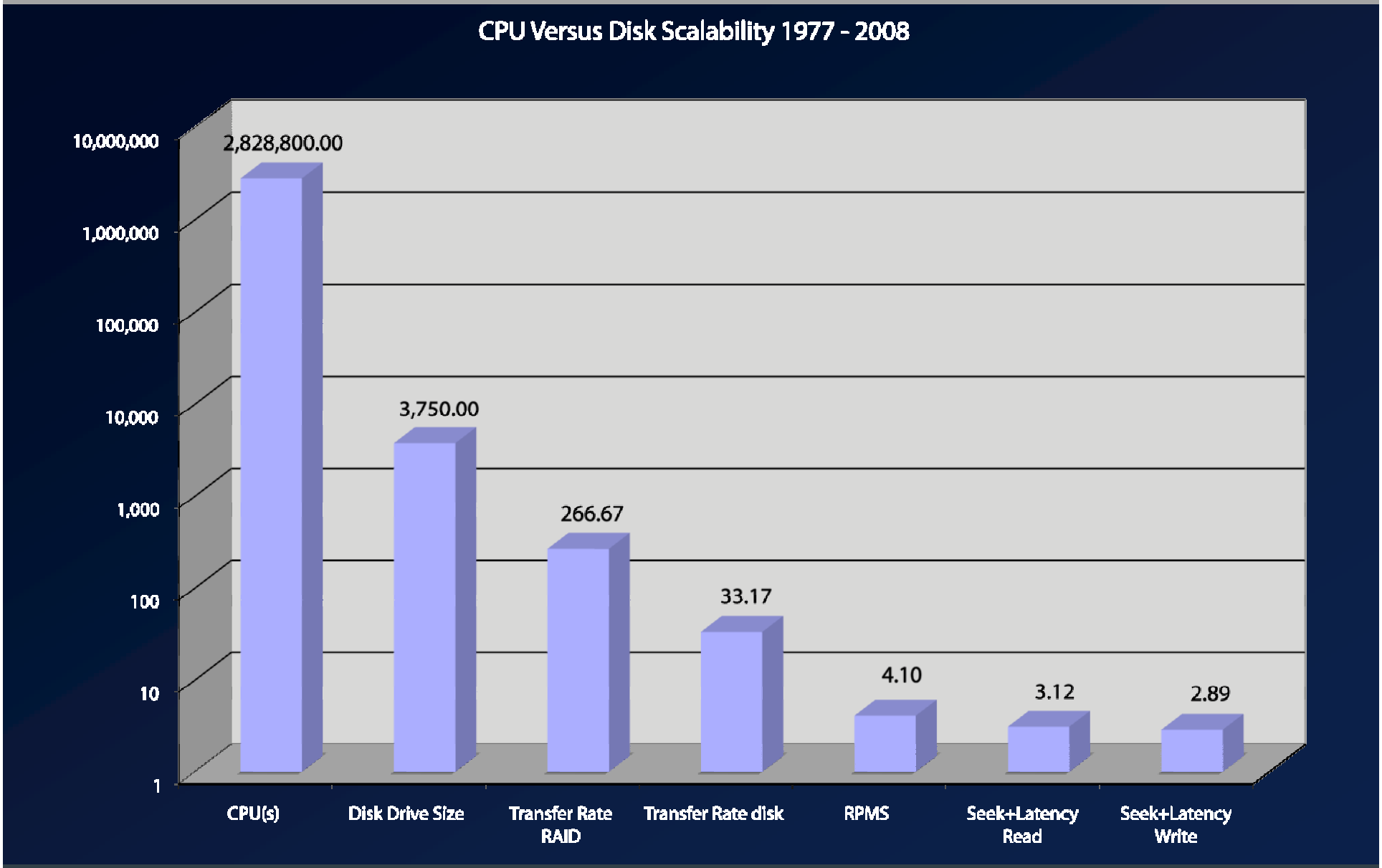
Instrumental

Storage Scalability

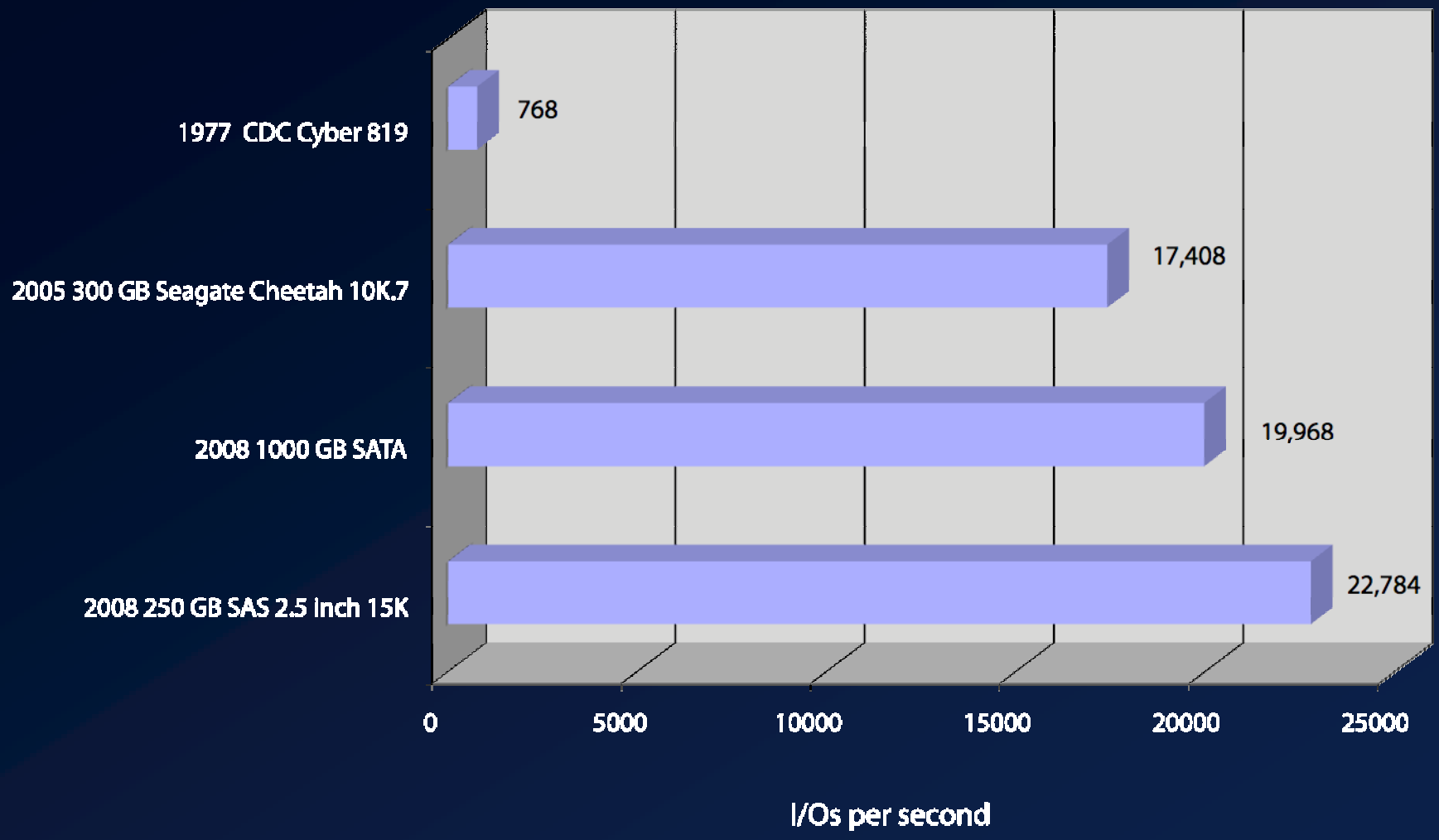
Limitations are Real

- Storage has not scaled well for decades
- Storage scaling limitations impact system, application design and hardware purchased
- Tape Impact
  - Storage scaling impacts tape sales
  - Tape latency (pick + load + position) is higher than disk by at least 14,400 times
    - This is why some companies are going to D2D backup
  - Tape performance and capacity is improving at a higher rate than disk

CPU Versus Disk Scalability

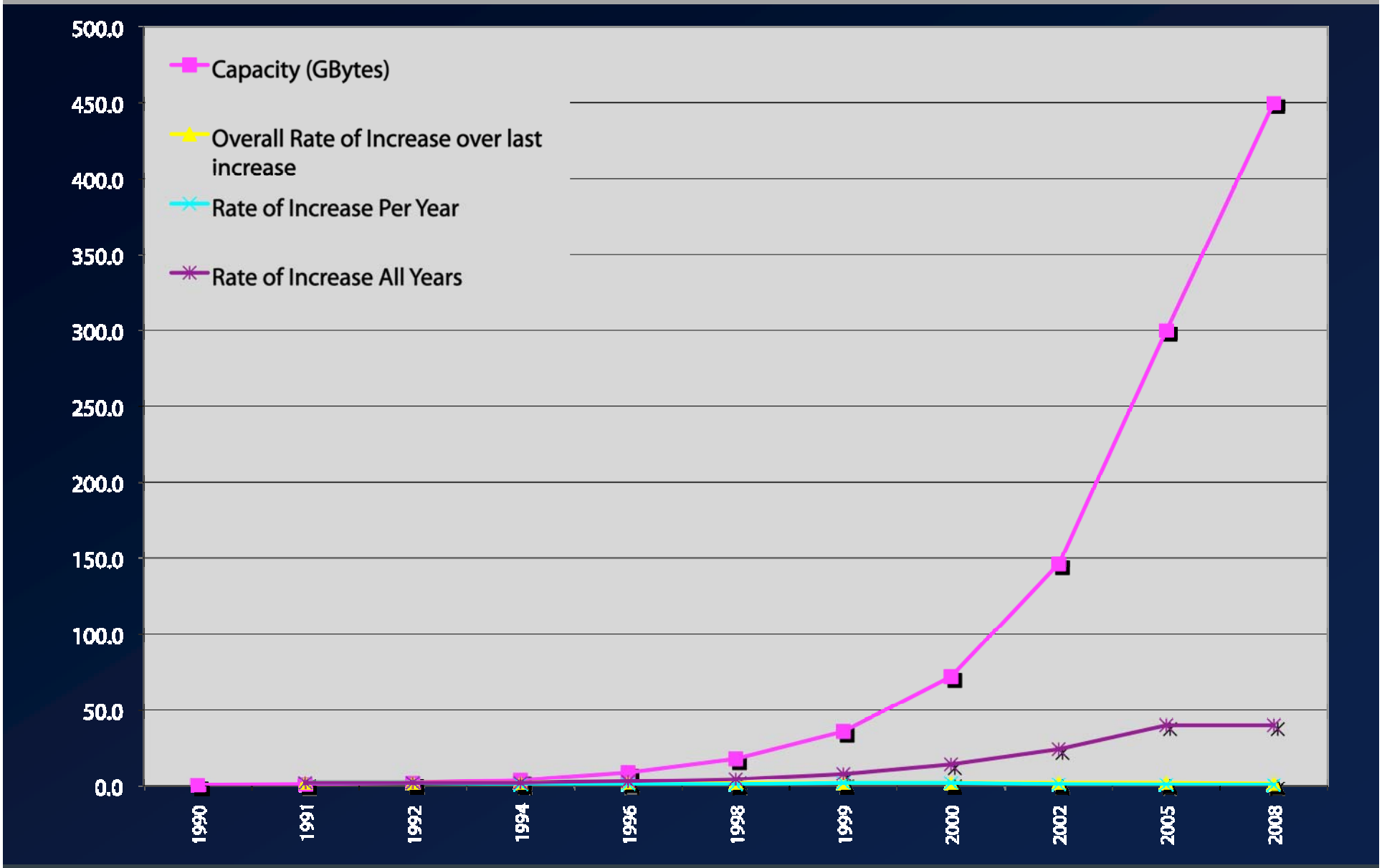


Disk 4 KByte I/Os Per Second

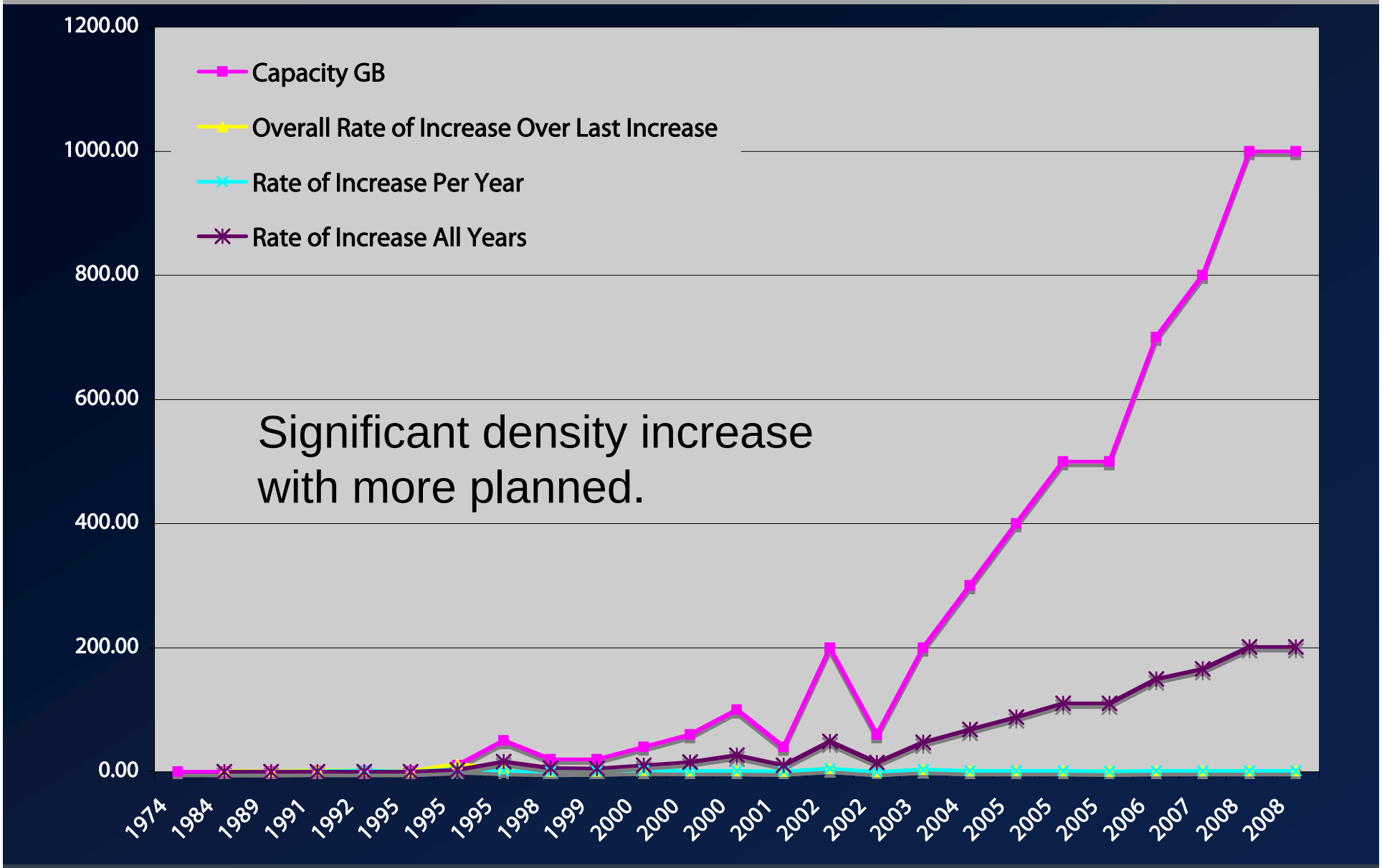


Only 29x difference in over 30 years

Seagate Disk History



Tape During the Same Period



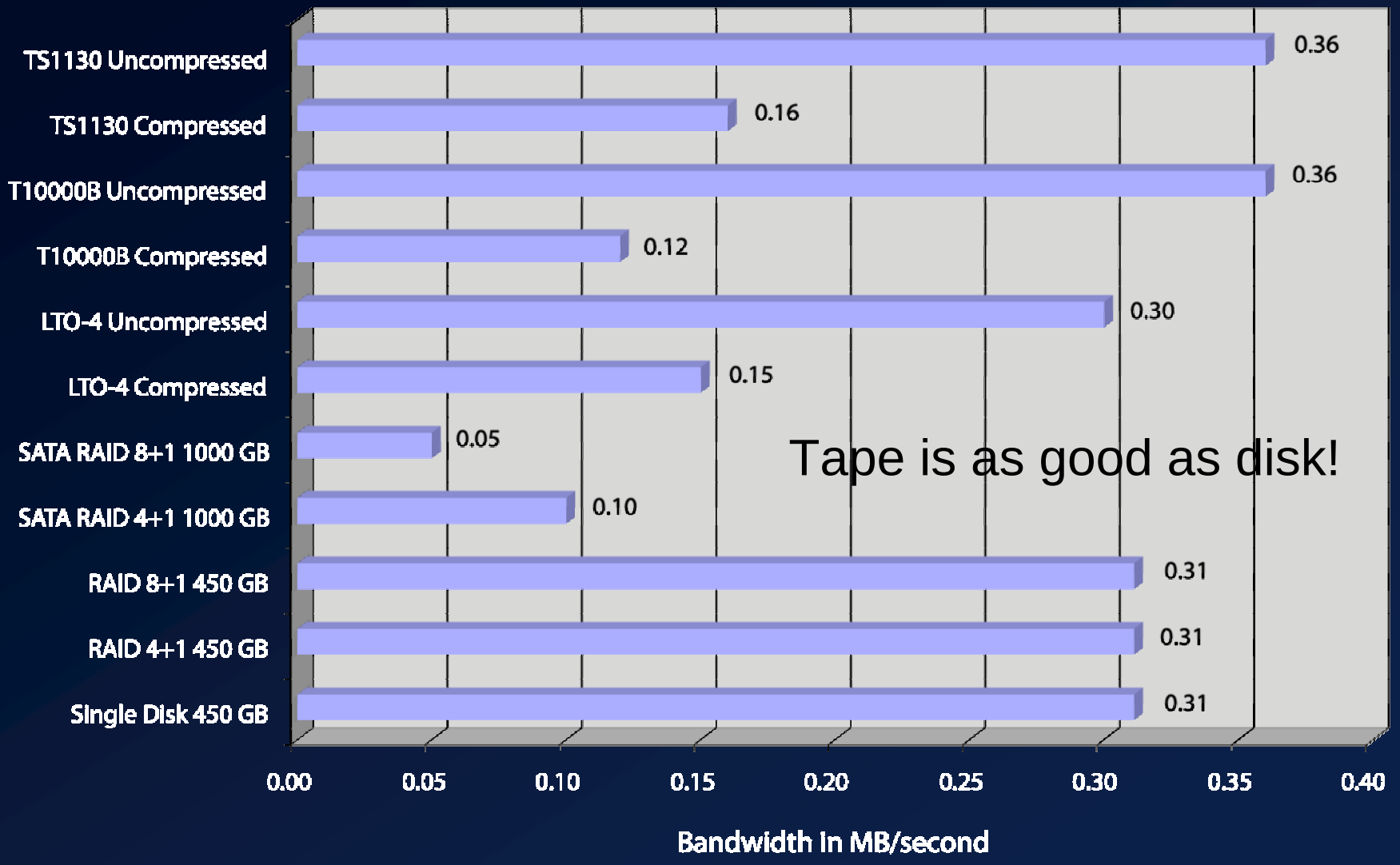
Performance Since 1990

Disk	Average Performance MB/second	Improvement Since 1990 x Times
FC/SAS	100	25
SATA	70	25

Tape	Compressed MB/second	Uncompressed MB/second	Compressed Improvement Since 1990 x Times	Uncompressed Improvement Since 1990 x Times
LTO-4	240	120	192	96
TS1130	360	160	288	128

Neither disk or tape performance is scaling, but tape is far better

Bandwidth Per GB of Capacity



Storage Scalability

- Bottom-line is neither tape or disk is scaling well in terms of bandwidth
- Tape is growing better than disk in terms of capacity and performance

Error Rates

Hard and Soft Errors

- Disk drive hard error rates (per Seagate)
  - SATA 1 sector per $10E^{-15}$
  - SAS 1 sector per $10E^{-16}$
  - FC 1 sector per $10E^{-16}$
- In 1996 the rate was 1 sector per $10E^{-14}$ for enterprise drives
  - Capacity was only 9 GB
- Tape Impact
  - Tape error rate has historically been 2 orders of magnitude better than disk

Enterprise Disk Reliability

	1996	2008	Comparison
Disk Mean Time to Failure (What do RAID vendors see?)	100K hours (<50K hrs?)	1.2M hours (~500K+ hrs)	~10x MORE reliable? (~10x)
Capacity per disk	9 GB	450 GB	~50x denser
Array Mean Time to Repair (Rebuild time @ 10%)	9 GB / 9.6 MB/sec x 10 = 9,375 seconds	450 GB / 99 MB/sec x 10 = 45,454 seconds	~ 3.2x LESS reliable

Problem 1: Time to repair disk (MTTR) is much worse!
Problem 2: This problem gets worse with SATA
Problem 3: This problem gets compounded with RAID

Data Reliability

- Not just hardware errors
- Silent data corruption can occur when an error occurs in both the packet and error check
- This results in
  - Undetected errors
  - Miscorrected errors
- Questions now raised
  - Is it in hardware or software?
  - Where is the error originating specifically?
  - What event caused the error?

Undetectable Bit Error Rate

Sustain Transfer Rate Per Second for a Year							
UDBER	0.5 GB/sec	1 GB/sec	10 GB/sec	100 GB/sec	1 TB/sec	10 TB/sec	100 TB/sec
1.E-21	0.0	0.0	0.0	0.0	0.3	2.7	27.1
1.E-20	0.0	0.0	0.0	0.3	2.7	27.1	270.9
1.E-19	0.0	0.0	0.3	2.7	27.1	270.9	2708.9
1.E-18	0.1	0.3	2.7	27.1	270.9	2708.9	27089.2
1.E-17	1.4	2.7	27.1	270.9	2708.9	27089.2	270892.2
1.E-16	13.5	27.1	270.9	2708.9	27089.2	270892.2	2708921.8
1.E-15	135.4	270.9	2708.9	27089.2	270892.2	2708921.8	27089217.7

This does not include errors as hardware degrades such as a failing drive and/or controller.

Bit error rates of most channels are $10E^{-12}$ and are corrected to $10E^{-17}$ for SATA, $10E^{-21}$ for SAS/FC.

Tape uses FC interface today, in the future potentially SAS interfaces, which are less susceptible to silent data corruption then SATA.

Therefore SATA is not a tape replacement unless parity is checked on read and corrections are made. Few vendors do this

Undetectable Bit Error Rate

- Tape Impact
  - Typically the weakest link is not the media, but the channel itself
  - Tape currently uses FC for the channel, potentially SAS in the future, which has roughly 4 orders of magnitude more ECC on the channel than SATA

Reduce Potential Errors

- A number of vendors provide products that monitor tape drives and tapes
  - Companies like Crossroads have products to address tape drive and cartridge errors
  - Monitoring errors and proactively removing tapes from the pool improves reliability
- Tapes have a lifespan just like disk drives
  - Tapes need monitoring similar to SMART monitoring of disks
  - Applications use raw SCSI commands to get tape drive status

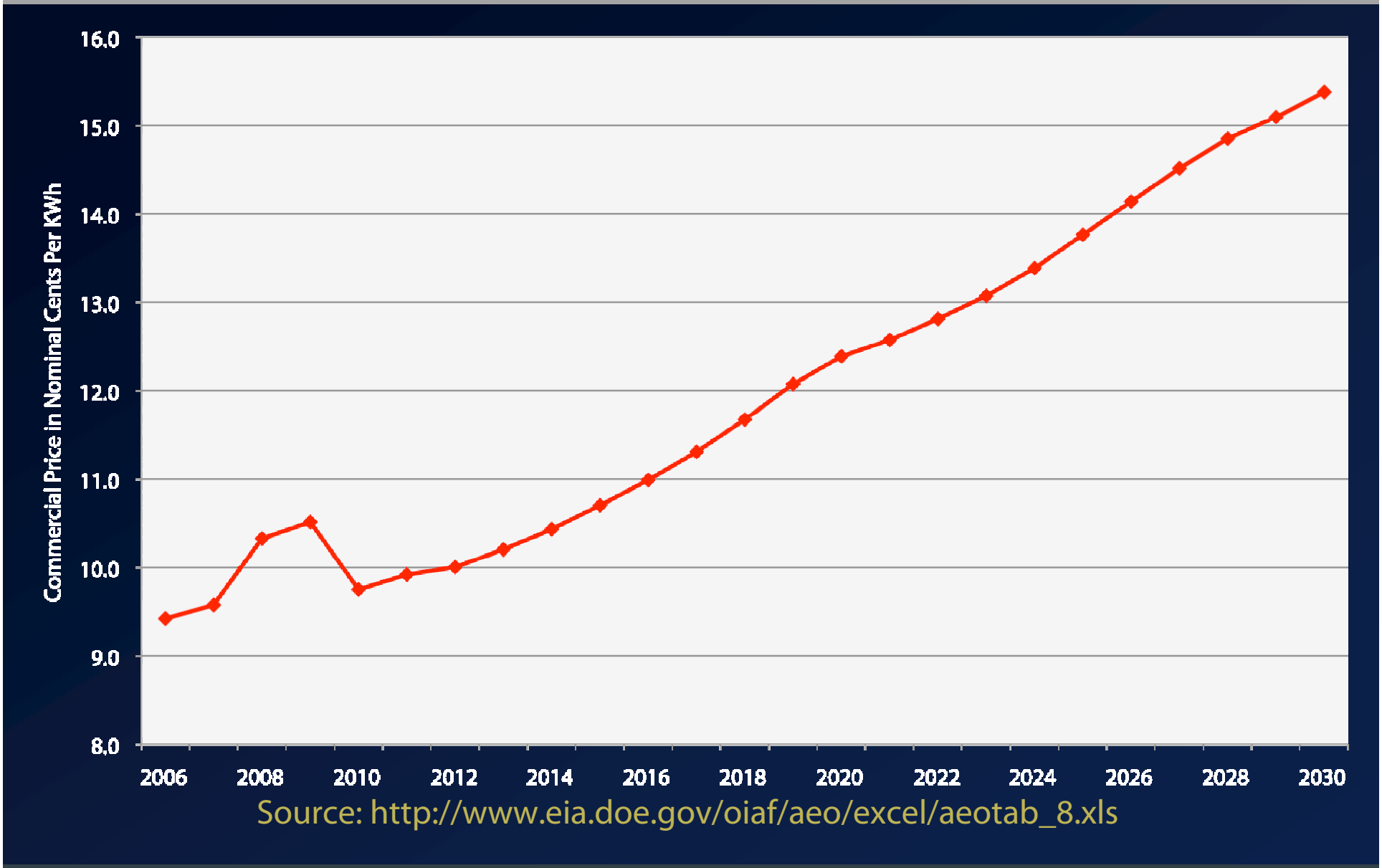
Instrumental

Power and Cooling Costs

Power Cost for Disk

- Power has become one of the biggest concerns for the data center
- Data centers are being built where power is located, not where businesses want them
- Power is such a problem that in Virginia AOL is paid by the power company to go on generated power in the summer sometimes
- Google moved to Oregon for power and cooling reasons

Electricity Price Estimates



Source: http://www.eia.doe.gov/oiaf/aeo/excel/aeotab_8.xls

Cost Per Petabyte

- The cost per Petabyte is
  - LTO-4 native $357,048.69
  - LTO-4 with compression $178,524.32
  - Sun 6540 $3,456,744.79
- Almost 10x more without compression
  - The 6540 has about 1.8 GB/sec of bandwidth while 20 tape drives native is about 2.4 GB/sec
  - Software is not included in prices and is not cheap
  - http://www.enterprisestorageforum.com/outsourcing/features/article.php/3722171

Power Cost for Disk

Amount of Storage	Drive Count	Watts/Drive	Total KWatts (Drives and Trays)	Cost $0.10/KW Hour	Yearly Cost of Disks and Trays
4.6 PB	5355	13	195.25	$19.52	$171,030.24

- Disks always use power if they are spinning
- Power for tape drives in use and robots is comparatively small
  - Even when disks are spun down, the interface to the hardware is powered on

Power Cost for Disk

- Tape Impact
  - Tape uses virtually no power in comparison to disk
  - Power consumption from Quantum LTO-4
    - Idle (no cartridge): 6.4 Watts
    - Standby (with cartridge): 9.5 Watts
    - Typical: 28.8 Watts
    - Max: 30.1 Watts
  - Equivalent to 2.3 drives of power

Cooling Cost for Disk

- Number of BTUs required for cooling varies with the disk drives used and capacity
  - 3.5 inch requires more power
  - (16) 1 TB Seagate SATA drives = 169 Watts
  - (16) 450 GB Seagate SAS drives = 277 Watts
  - (16) 2.5 inch 146 GB Seagate SAS drives = 121 Watts
  - The best power density is 1 TB drives at 13 Watts per drive
    - Not enterprise level drives
    - Not fast
    - Not reliable

Cooling Cost for Disk

- Cooling costs about 1.45 the amount of power
  - So the $171,030.24 is really $247,993.85
  - The cost will go up since these numbers were generated with $0.10 KW hour
  - Power usage per GB will drop about 30% with 2.5 inch drives
- Tape Impact
  - Tape require virtually zero cooling
  - Again, operational power consumption is very low

Massive Array of Idle Disks (MAID)

MAID Facts

- MAID schemes power drives on and off based on needs
  - Most MAID devices limit the number of disks that can be powered on at any given time
  - MAID devices are configured as RAID from 3+1 to 8+1 depending on vendor
- With random recalls, some requests might have to wait based on the usage of the MAID device
  - Some MAID vendors allow only 25% of the system to be active at any give time
  - That INCLUDES RAID rebuild

MAID Facts

- MAID generally uses SATA drives which have known reliability issues
  - This increases the chance that the MAID device will be rebuilding rather than servicing I/Os
- MAID does not support hardware compression
  - Compression is done in software
  - Consumes CPU cycles slowing overall performance

MAID Facts

- Tape Impact
  - Hardware compression and encryption is always preferred to software
  - The channels for both SATA and SAS/FC are rated to $10E^{-12}$, SAS/FC are corrected to $10E^{-21}$ versus $10E^{-17}$ for SATA
  - Power and cooling are still cost considerations for MAID
  - Host bandwidth is the limiter to the number of tape drives that can be used, not the hardware itself

Instrumental

Data De-duplication

Data De-duplication

- Data de-duplication breaks files into pieces and compares this hash against existing files
- Similar to standard compression, but occurs across many files rather than one
- Concerns
  - Good data on disk, bad read, what is the outcome?
  - Good data in memory, but bad write. How much data is corrupted?
  - Is it possible to find bad data to correct the rest?

Data De-duplication

- Unless vendors provide checksum for both data and hash, there is a risk of data corruption
  - Data Domain and a few others do this
- If placed on less reliable storage what is the risk of a silent data corruption
- Data de-duplication may be better suited for email, rather than enterprise critical data
- Tape Impact
  - Compression occurs on one file at a time reducing the risk of widespread corruption

Instrumental

Tape Performance

General Tape Performance

- Stream data to tape
  - Tape drives perform best when streaming data using large blocks
  - Starving the drive will reduce performance due to start/stop of the drive
  - Some drives can slow down in response to incoming data, but not all
- Block sizes
  - Enterprise tape drives use anything from 256 KByte to 2 MByte block sizes

General Tape Performance

- Block sizes, continued
  - Most backup and HSM software are aware of the correct block sizes, but not always
  - Trust, but verify that the application is using the correct block size
  - Too small of a block size, the system will coalesce application I/O requests to form one large request
  - End up spending time forming requests rather than performing I/O
  - This leads to devices being busy when not actually doing valuable work

General Tape Performance

- Application tape buffers
  - Many applications provide a tunable for the number of tape buffers to use
  - Increasing this circular buffer to multiple MB or GB can help applications queue data more efficiently
  - Having an efficient queue keeps tape drives streaming and performing well

General Tape Performance

- Maximizing tape loads
  - Keep on the look out for tapes that are loaded, but little or no data being written/read from the drive
  - Load/thread/rewind/unload takes a lot of time and effects the overall performance of the tape subsystem
  - Use the tunables in your application to wait for enough data to accumulate before writing

General Tape Performance

- Backup/Archive Parallelization
  - If your system has multiple tape drives, use them
  - Spread the workload as much as possible over multiple drives to achieve higher overall performance
  - http://publib.boulder.ibm.com/infocenter/iseries/v5r3/index.jsp?topic=/rzalw/rzalwtape.htm

Network Performance

- Sites complain about the time to perform backups, but is this the fault of the network?
  - 1 Gbps Ethernet peaks at 100 MB/sec
  - LTO-4 without compression is faster leading to starvation of the drive and overall slowdown
- Tapes try to operate at fastest speed possible and will slow down to match the incoming data
  - If a drive is rated (in MB/sec) at 120, 90, 60 or 30 and data rate is 59 MB/sec, the drive will likely operate at 30 MB/sec
  - Some vendors offer variable speed drives

Network Performance

- Network performance and design impact tape performance
  - This is not the fault of the tape drive or media

Network Performance

- High latency networks over WANs are another problem, similar to network bandwidth
- Again, the drive will operate at the slowest data rate to attempt to keep data streaming
- Bottom-line
  - Network performance is just as important for performance as other components
  - Think of the systems ability to ingest data over the network before blaming the tape drives

Server Performance

- Servers have two performance bottlenecks
  - Memory bandwidth (memory to PCI bus)
  - PCI bus bandwidth
- Some servers currently limit memory bandwidth to less than 10 GB/sec
  - To read or write to tape you must also be reading and writing to a file system
  - This means that the total bandwidth is doubled
  - 2 GB/sec to tape means a minimum of 4 GB/sec of memory bandwidth

Server Performance

- Historically memory bandwidth has been one of the bottlenecks that impact tape performance
  - Also affects other components in the I/O path
- Small servers can limit tape performance with a slow PCI bus
  - Each PCIe 1.1 bus supports 2.5 GB/sec of I/O
  - Each slot has a lane count and each lane is 250 MB/sec
  - Slots have either 1, 2, 4, 8 or 16 lanes
  - PCIe 2.0 doubles performance for buses and lanes

Server Performance

- Small servers may not have enough PCI bandwidth to support bandwidth requirements
- Bottom-line
  - Server sizing is critical to a well performing backup or archive system
  - Must have enough PCI slots for the job at hand
  - Remember memory bandwidth is double the rate when using tape with a file system
  - Be realistic when looking at the theoretical performance of a server

File System Performance

- Direct I/O is important for tape performance
  - Direct I/O bypasses kernel paging and writes/reads data directly from application buffer to storage
  - Having to hop from tape to kernel pages and finally to the file system, vice versa for writes
  - This can DRAMATICALLY reduce performance
- File system block sizes affect how data is written to tape
  - File system block sizes that are smaller than the tape drive will result in extra time coalescing buffers to form large requests

File System Performance

- File system block sizes affect how data is written to tape
  - Read-ahead on the RAID only works if the files are sequentially allocated
- Bottom-line
  - A poorly performing file system will negatively impact tape performance
  - Tape drives are simple and easy to understand
  - Blame shifts to tape, but what's further up the I/O data path that could be affecting performance?

Final Thoughts

The Future of Tape

- Tape density is growing and has been growing at a higher rate than disk density
  - The trends continue to show this
- Tape is green
  - Power managed disk storage is not enterprise quality yet and has severe limitations
- Cost in terms of device usage and cooling
  - The cost to power and cool a tape device is negligible compared to disk
  - Power is much more important now and will continue to be in the future

Success of Tape Depends On

- Tape is the first choice for shipment
  - Shipping disk drives is scary and expensive
- Tape should be used differently than disk
  - Inherent in the technology
- Good architecture and tuning
  - Without either tape looks bad and disk better

Thank You