



Education

# **Enterprise Class Mass Storage: Innovative Approaches to Disk-based Active Archive Storage**

William Mottram, Veridictus Associates Inc & The Wikibon Project

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the Author nor the Presenter is an attorney and nothing in this presentation is intended to be nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
- The information presented herein represents the Author's personal opinion and current understanding of the issues involved. The Author, the Presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

**NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.**

## ➤ **Enterprise Class Mass Storage: Innovative Approaches to Disk-based Active Archival Storage**

This session will appeal to Data Center Managers and those that are interested in how high capacity storage technologies are evolving. The presentation builds from a discussion of the problems created by massive data growth and the argument that traditional approaches to data storage will not meet the demands of large enterprise datastores. Demands such long term data storage but with near instant access times (sec), “*findability*,” robust data and device integrity, scalability and with the flexibility to able transparent service of unpredictable activity patterns.

The presentation will focus on the innovative technologies that are enabling these next generation solutions that are increasingly becoming identified as a separate class of storage.

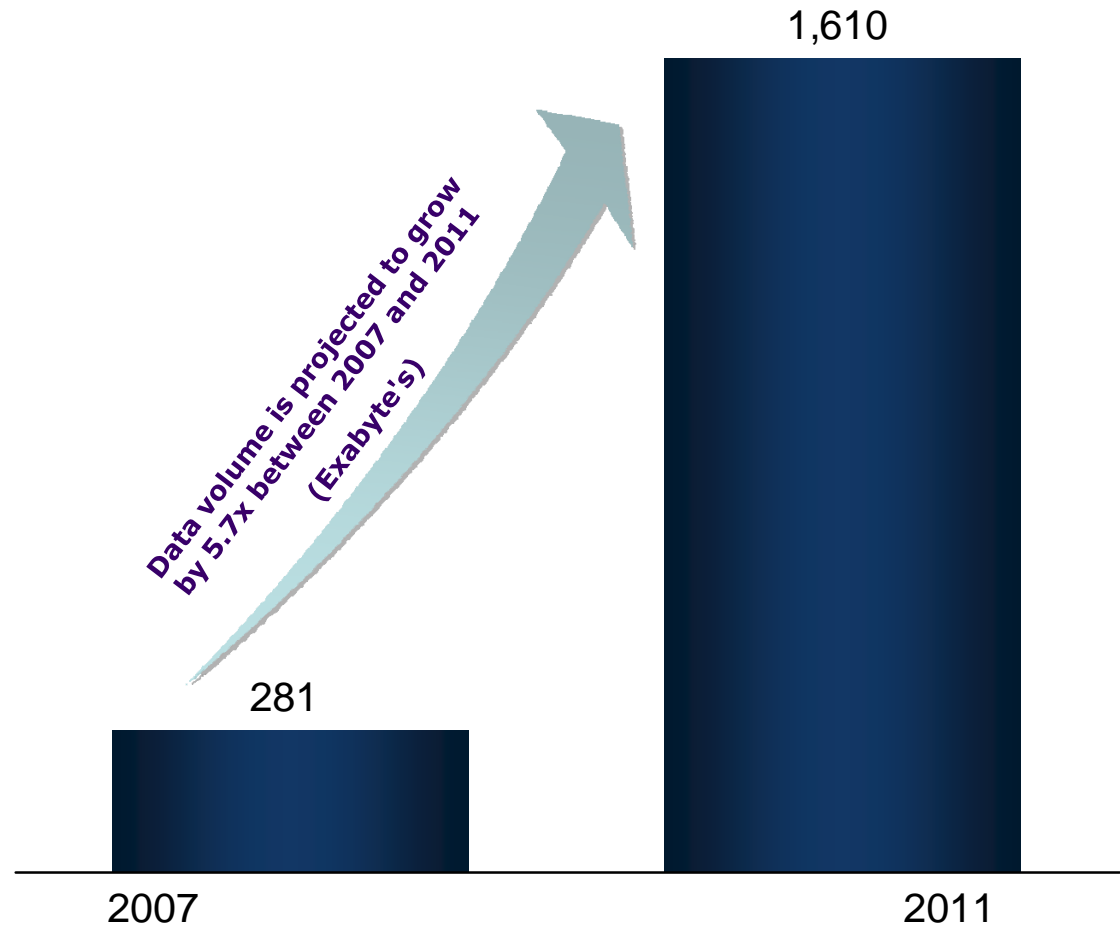
The presentation will conclude with a look to the future with a brief discussion on holographic storage and some interesting projects that are still in the academic domain.

- Understand the challenges presented by massive datastores and how these challenges influence storage strategies and purchase decisions.
- Understand how next generation storage architectures are designed to meet the challenges of today's active archive applications such as delivering acceptable QoS, i.e. access times, latency, bandwidth, scalability and solution longevity.
- Know what questions to ask when considering the purchase of these massively dense solutions.

- Introduction
  - ◆ Why the status quo is not sustainable.
- Framing the problem.
  - ◆ The multiple and demanding personalities of data
  - ◆ What is active archive?
- Today's Evolving solutions – nascent but available.
  - ◆ Architecture
  - ◆ Technology
  - ◆ Sustainability
- Look into the crystal ball.
  - ◆ Some happenings in academia
- Final thoughts.
  - ◆ Solving for tomorrows problems

# Why the Status Quo is Not Sustainable!

- Explosive Data Growth
  - Internet, e-business, digital media
  - Increasing OLTP
  - Accretive nature of data
- Storage Selection Factors
  - Data or use patterns are not homogeneous and drive different storage needs.
  - Worthless data has no value only cost and liability



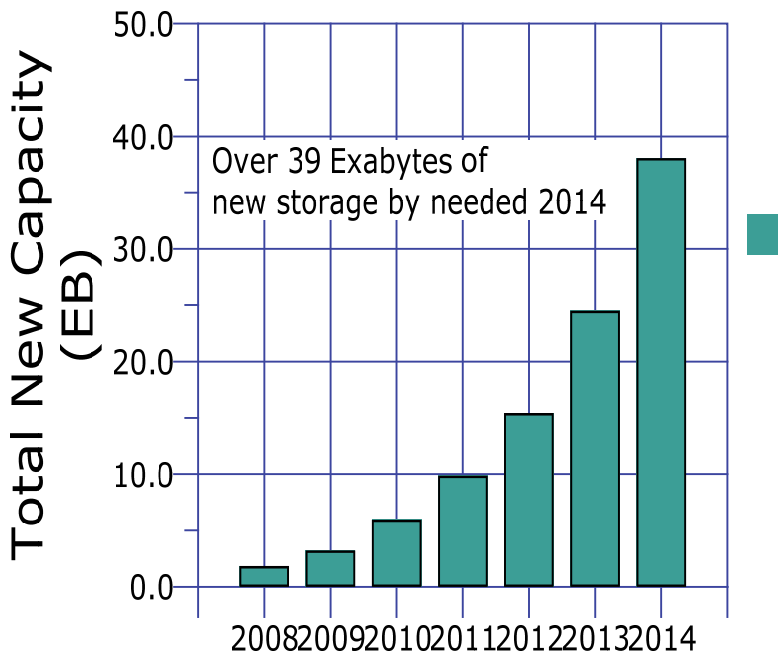
\* IDC: The Diverse and Exploding Digital Universe March 2008.

# Compliance and Corporate Governance Drive Accretive Growth.

|   |              |
|---|--------------|
| ➤ tax returns                           | Indefinitely |
| ➤ expense reports                       | 6 yrs        |
| ➤ bank statements                       | 3 yrs        |
| ➤ production correspondence             | 8 yrs        |
| ➤ employees exposed to toxic substances | 30 years     |
| ➤ medical records                       | Life + 2 yrs |
| ➤ monthly trial balances                | 6 yrs        |
| ➤ insurance claims post settlement      | 10 yrs       |
| ➤ Customer Information/BI               | Indefinitely |
| ➤ Product/Field Performance             | Indefinitely |

# Two Dynamic “Active Archive” Growth Sectors.

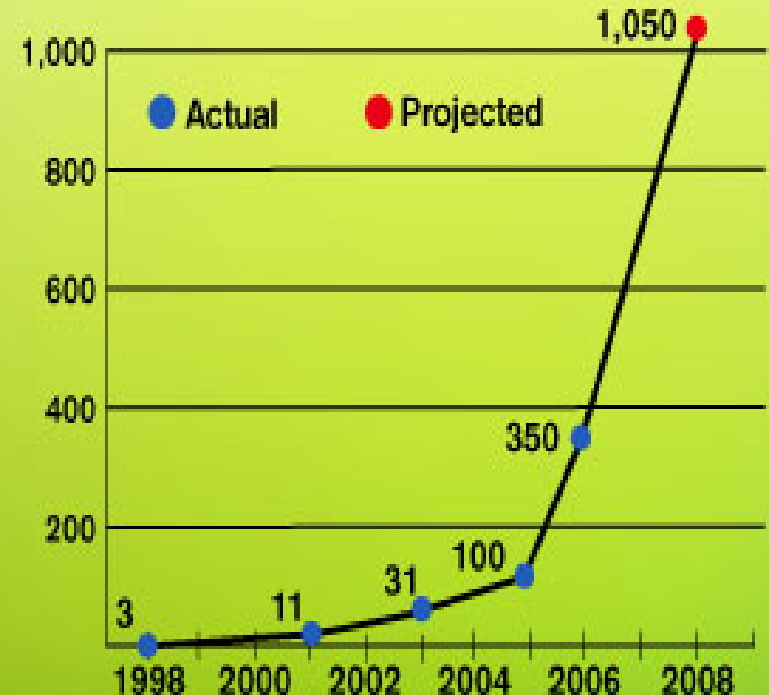
## Storage Capacity for Media and Entertainment



Source: 2009 Digital Storage for Media and Entertainment Report, Coughlin Associates

## Biggest Just Gets Bigger

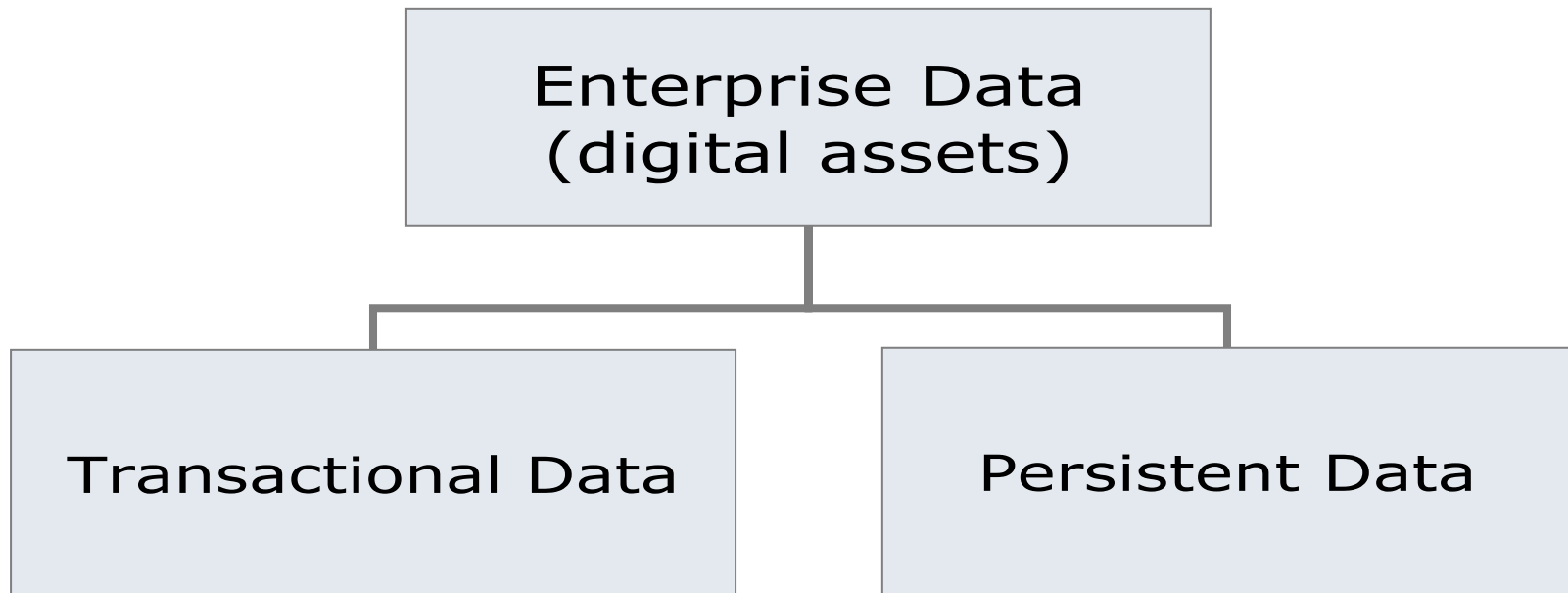
Size of largest data warehouse  
over the years (in terabytes)



Data: WinterCorp

# ➤ Framing the Problem

# Data is Not Homogeneous!



**Transactional Data** - the traditional view of data that has molded today's disk storage architecture.

**Persistent Data** - data that once created is rarely accessed or modified - the fastest growing segment of data storage

# Data Types Determine Storage Architectures?

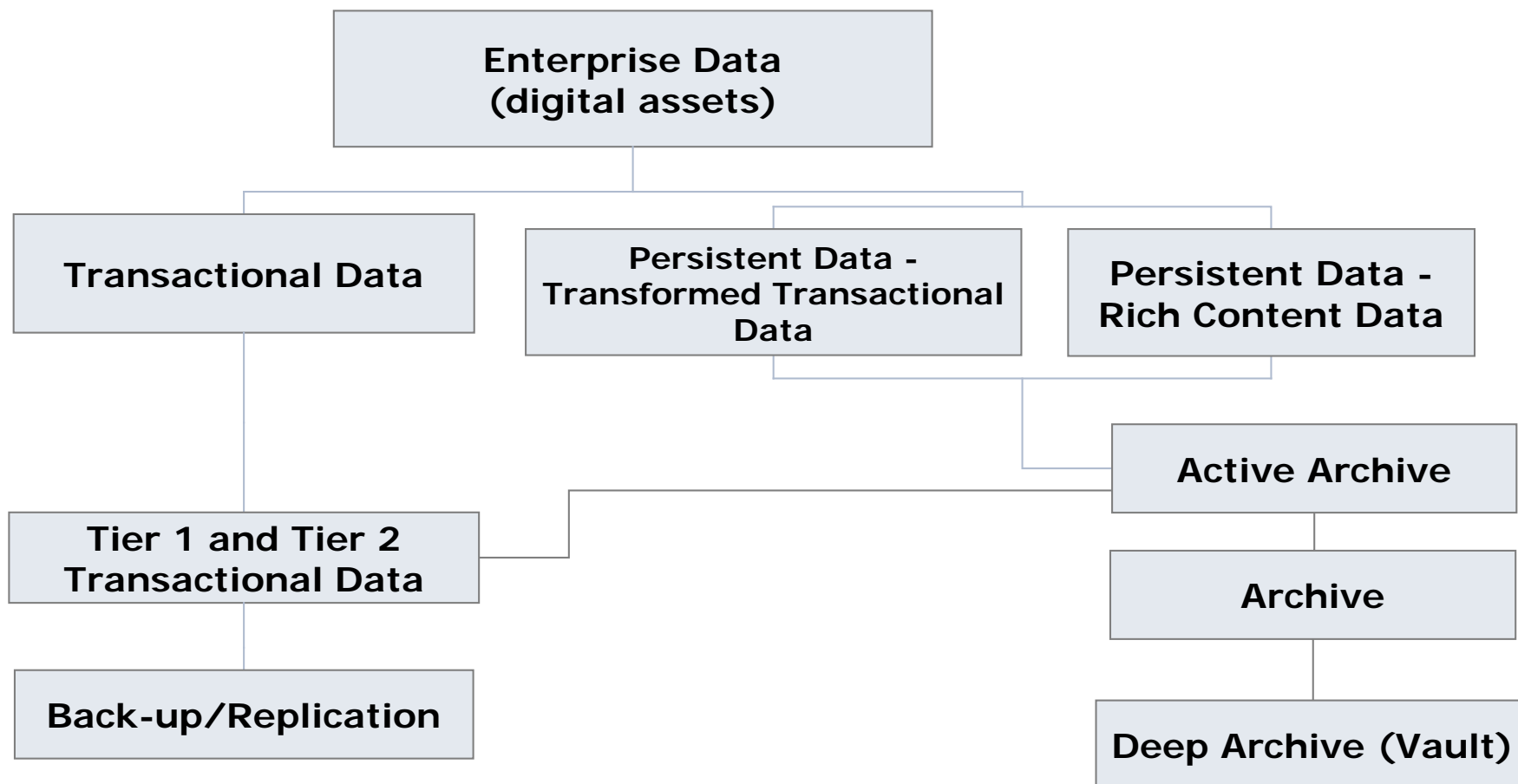
## Transactional Data *"Data Capture Storage"*

- Very low latency response, high IOP's data.
- random in nature
- Read-write-Modify Model
- Optimized to provide access to data at most if not all times.
- Exhibits temporal or spatial locality cacheable
- Optimized to small-grained data access
- Data retention tends to be short lived.
- Design principles that drive product cost upwards

## Persistent Data *"Data Retention Storage"*

- Latency tends not to be an issue. Once created data is rarely accessed or modified often immutable.
- tends to have lower IOPS requirements but bandwidth centric
- Low temporal access locality, caching tends to be a wasted expense
- long term retention - robust data care
- immutable, reference content
- Design principles that eliminate unnecessary sophistication eliminates cost.

# Data is Not Homogeneous!



- Archive is not back-up but a complementary process
  - ◆ Recovery vs. long term retention
  - ◆ If your back-up is more than 4 weeks old is it a back-up or archive?

## ➤ Back-up:

- ◆ A secondary copy of the data
- ◆ Primarily used for recovery actions - enabling predictable RTO and RPO
- ◆ Data retention measured in weeks or months
- ◆ Not a useful option for compliance/ediscovery

## ➤ Archive:

- ◆ Primary copy of the data
- ◆ Available for information retrieval
- ◆ Can significantly improve operational efficiencies
- ◆ Long term data retention
- ◆ Data is retained for corporate governance, business intelligence or regulatory compliance

- Draws on attributes of transactional and persistent storage
  - *Mass storage with performance*
- Transparent servicing of unpredictable workloads - Elastic architecture.
- Scalable and flexible
- Archived data must be easily “findable”
- Suitable for long term data retention
- Design principles that promote sustainability but eliminate unnecessary sophistication.

➤ **Nascent but available technologies and techniques.**

- Most solutions are developed to solve existing problems
- Rear view mirror perspectives tend to guide solution development.
- Solutions tend to be compromised to fit capabilities of existing technology
- Solutions based on standard components have a faster pace of evolution

# What Architecture is Right!

- **Scale-up (Monolithic) or Scale-out (Clustered)?**
  - ◆ Monolithic or Scale-Up - the traditional approach where storage sits behind one or two controllers/servers heads.
  - ◆ Clustered or Scale-Out - the flexibility to independently scale bandwidth, processing and capacity, non-disruptively and on the fly. Easiest architecture for the incremental introduction of new technologies.
  
- **Questions to ask.**
  - ◆ Does it meet requirements?
    - > Access, Performance and Availability
  - ◆ Will it easily scale to meet future requirements?
    - > Capacity, Performance and Time
    - > Does the file system scale
  - ◆ Flexibility
    - > Can the solution be upgraded without disrupting data access?
  - ◆ Is the solution optimized to meet the requirements for long term data storage?
    - > Reliability, longevity, upgradability, sustainability and cost

# Scale-out Storage (Cluster)

- High density disk drive packaging driving space efficient storage
  - ◆ High TB/Sq ft ratio
- Tends to use large capacity disk drives (but not always)
- Cluster file system enabling transparent scalability and upgradability
- Highly scalable and flexible
  - ◆ Performance, capacity and time
  - ◆ Tunable
- Reliable:
  - ◆ Highly available and self healing
  - ◆ Immune (tolerant) to failure
- Affordable:
  - ◆ High volume hardware components to drive down cost

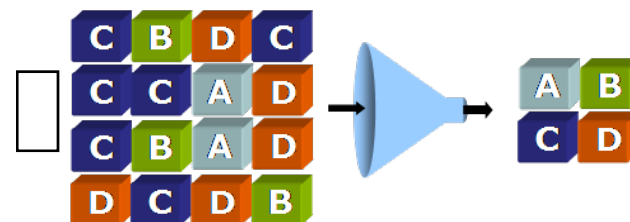
# When Not to Cluster.

- When traditional storage scales to meet your needs
- You do not need the flexibility
- Workload does not require the attributes of a cluster
- When all you need is a single, unified view of your files

# Storage Optimization – Data Footprint Reduction.

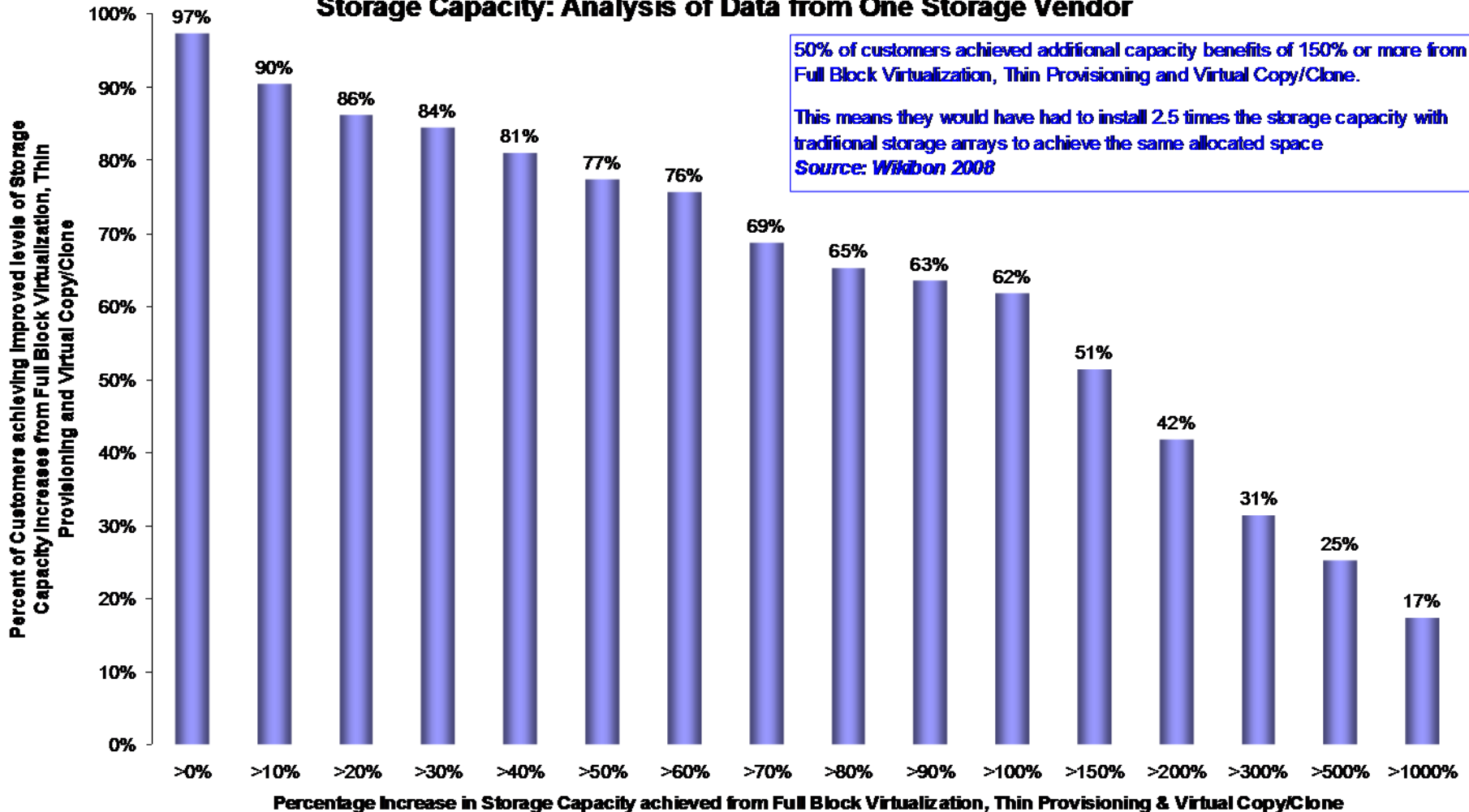
- Techniques that take advantage of the fact that data has statistical redundancy
- Compression
  - ◆ 2:1 reduction ratio is a good planning number but, it depends!
  - ◆ Primary vs. secondary
- Data-De-duplication
  - ◆ File and block level
  - ◆ Replaces duplicate data with pointers to a single shared copy
  - ◆ Data reduction – it depends!
- Thin Provisioning

## Data Deduplication



# Block Storage Optimization.

## The Effects of Full Block Virtualization, Thin Provisioning & Virtual Copy/Clone on Storage Capacity: Analysis of Data from One Storage Vendor



- A must for long term “archival” data storage
  - ◆ Massive volumes of inactive data
  - ◆ Large disk pools
  - ◆ Automated maintenance eliminates embarrassing and expensive errors
    - › Fail in place vs. hot spare
  
- Vendors can avoid the obvious issues by thoughtful engineering
  - ◆ Packaging
    - › Density, Vibration and Heat
  - ◆ Predictive maintenance eliminates (minimizes) performance impacts
    - › S.M.A.R.T.
    - › Data integrity – managing media defects

## ➤ Fail-in-Place (Heal-in-Place)

- ◆ Sealed storage container
  - greater degrees of design freedom counter vibration and cooling issues
- ◆ Drive population includes operational as well as spare capacity
  - anticipated spare capacity to meet service the life expectations
- ◆ Eliminates hot spare tending
- ◆ Autonomic maintenance
  - Isolation of failed drives
  - Elimination of false positives
  - Rehabilitation or elimination process
  - Reduce duty cycle to effectively increase MTBF

## ➤ Recover usable capacity

- ◆ Partial capacity drives reintroduced into the storage pool

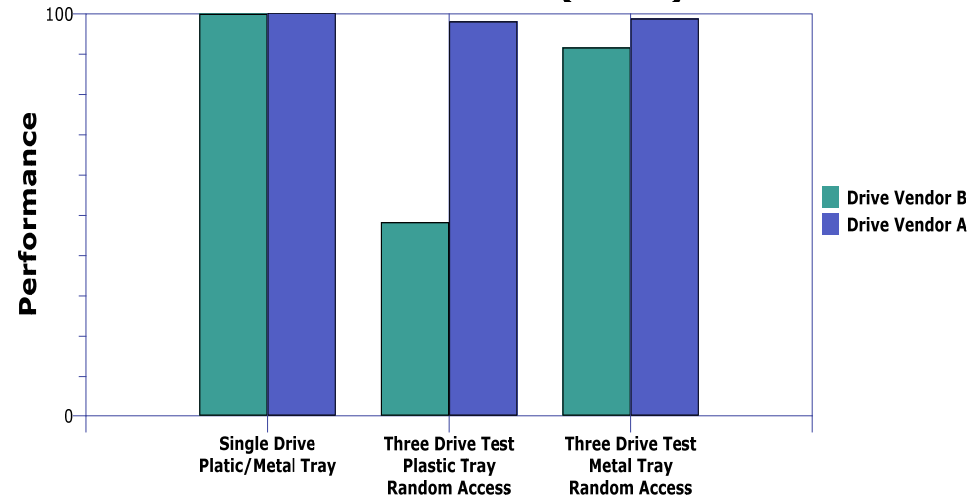
## ➤ Ongoing spare capacity monitoring

- ◆ Predicting life of storage container at full capacity

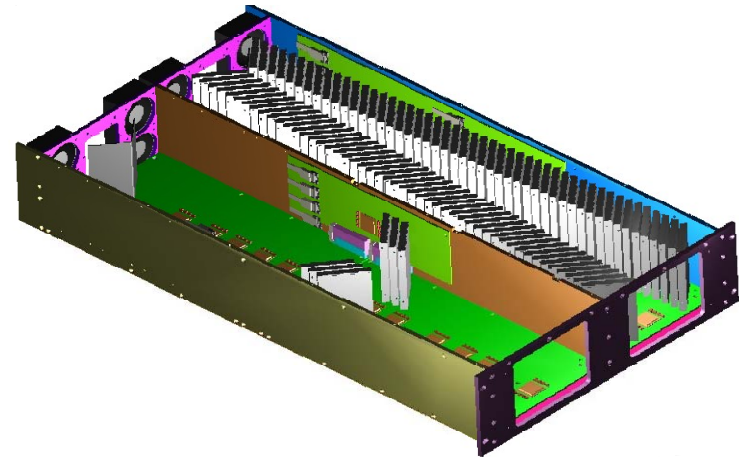
# Innovative Packaging Concepts

- Dense packaging has unique challenges
- Vibration
  - ◆ Rigidity & Placement
- Cooling
  - ◆ Airflow effectiveness

## Impact of Cabinet Design and RV features on Performance (IOPS)



Source: Vendor Technical Whitepaper



# Access Density and Unpredictable (WEB 2.0) Workloads

- Access Density = IOPS/Storage Density (IOP/GB)
  - ◆ a measure of drive or subsystem response time.
- Management of unpredictable workloads
  - ◆ Rapid ingest and delivery of data
  - ◆ peak query volumes
  - ◆ variable query complexity
- Data Access Challenge
  - ◆ *“the capacity of a disk drive has increased 6000 times since 1964 the raw performance, seek, latency and transfer rate has only increased by a factor of 8”. Fred Moore. CTR, April 2000*

- Two classes of disk drive are evolving – capacity and performance.
- However, capacity growth is exceeding performance improvement encouraging “bad” practices.
  - ◆ Short stroking
- Active Archive need capacity and performance.
  - ◆ SSD evolving as the high performance tier.
  - ◆ Imbedded policy driven QoS
  - ◆ Tiering
  - ◆ Increase actuator count
    - > 2½" technology

- Drive Capacity
  - ◆ 2.5" FF now at 500GB (1TB by 2010)
- Volumetric Efficiency
  - ◆ 3.7x to 5.8x space utilization advantage
- Volumetric Density
  - ◆ 1.88 to 2.9 volumetric storage advantage
- Energy Efficiency
  - ◆ It depends – “consumer” vs. “enterprise” class
- Double IOPS

- Uses Memory technology with an online storage (disk) device image. (DRAM, NAND)
- Extremely fast, microsecond access times
- Announced capacities exceeding 1TB
- Expensive, but pricing dropping
- Claim –
  - ◆ SSD 30% to 40% more energy efficient
  - ◆ Up to 30x I/O performance (IOPS)
  - ◆ More efficient than short stroking

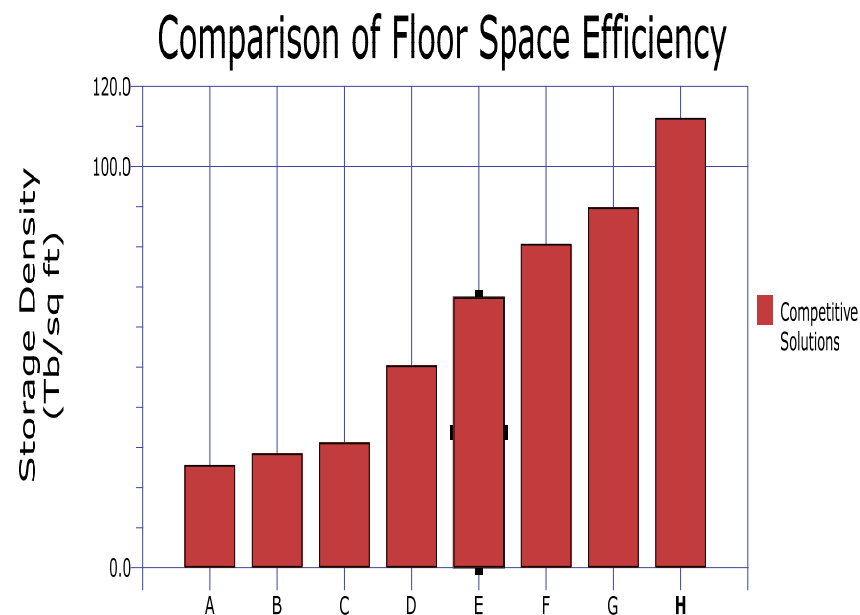
## ➤ High Storage Density

- ◆ Simplify management
- ◆ Save floor space

## ➤ Energy Efficiency

- ◆ Engineering for efficiency
  - Power supplies; variable speed fans, cache battery options
- ◆ Power Management
  - Policy managed
- ◆ Cooling
  - “rule of thumb” - 1KW of power consumed requires an additional 1KW of cooling.
  - Engineering concepts that drive efficiency
- ◆ Drive spin down, drive power down and MAID

➤ *“Certainly, there are environmental reasons for going green, but a green focus also can result in significant savings”- eWEEK.com*



# Drive spin down, drive power down and MAID

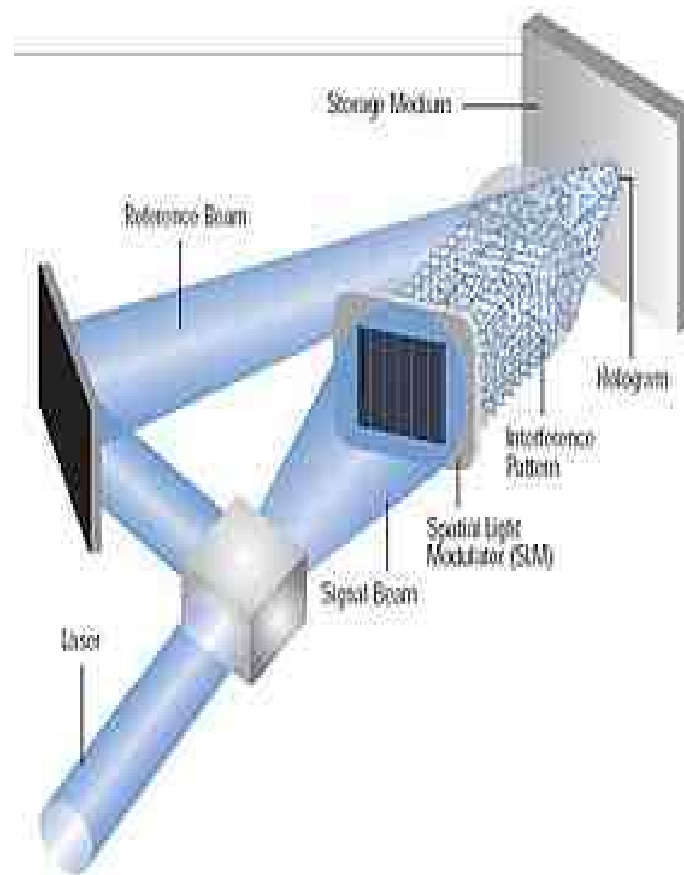
## ➤ Native MAID

- ◆ A storage system that powers down drives completely, individually or in groups when not required.
- ◆ Original definition (Univ. of Colorado) specifies that no more than 50% of the maximum installable drives could spin concurrently. Power provisioning advantage (CapEx)
- ◆ Very dense packaging
- ◆ Energy savings claim of up to 85% (OpEx)

## ➤ Soft MAID

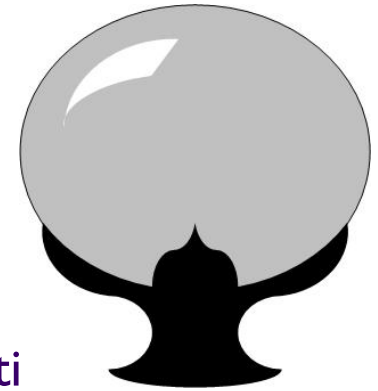
- ◆ Basic implementation is to simply park the actuator
- ◆ Drive rotational speed reduced or stopped but remain “hot”
- ◆ Reduce energy consumption from 15% to 60% (OpEx)
- ◆ However no data center power provisioning advantage (CapEx)
- ◆ Slower speed “econo” drive

- A wise man once told me that holographic recording is a future technology that always will be!
- Technology is now shipping
- Performance
  - ◆ Now 300GB - 20MB/s
  - ◆ 2011 800GB - 80MB/s
  - ◆ 2013 1600GB - 120MB/s
- 50 year archival life
- Encased media



# A Brief Look to the Future

- Pergamum:
  - ◆ long term evolvable storage built from intelligent network attached bricks (Tome) hosting disk, low processor and NVRAM
- Deep Store:
  - ◆ building more efficient archival storage using de-duplication advantage of intra-file and inter-file redundancy to increase storage density
- OceanStore:
  - ◆ An internet-scale persistent data store designed to scale to “billions” of users; incremental scalable, secure sharing and long-term durability
- MEMMS- (MicroElectroMechanical Systems)
  - ◆ Non-volatile storage blending magnetic recording material with thousands of recording heads
  - ◆ Photolithographic process producing a storage density up to 10GB/cm<sup>2</sup>, >1ms access times and bandwidths <50Mb/s



- Avoid rear view mirror perspectives; yesterdays solutions are not going to solve tomorrows problems
- Active archive storage is mass storage with performance
- Architecture must be transparently scalable in performance, management, capacity and time.
- Architecture should be self monitoring and self healing (device and data)
- Architecture must be sustainable
- Avoid compromising you archive solution to fit the capabilities of traditional technology.

- Please send any questions or comments on this presentation to SNIA: [trackstorage@snia.org](mailto:trackstorage@snia.org)

**Many thanks to the following individuals  
for their contributions to this tutorial.**

**- SNIA Education Committee**

**Bill Mottram  
David Vellante  
Rob Peglar**