Can You Manage at Petabyte Scale?

© 2012 Storage Networking Industry Association. All Rights Reserved.

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.

- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the 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, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

- The information presented herein represents the author's personal opinion and current understanding of the relevant 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.
Abstract

Storage Management for Data Center Managers

- Stored data growth within the data center now averages 50-80% on a compounded annual growth rate (CAGR) basis. However, it is anticipated that storage administrators will experience accelerating growth rates over the next three to five years. Acceleration will be driven by a number of factors we can identify now including mobile devices, analytics applications, and a growing list of data sources that are external to the data center. The cost of acquiring, managing and maintaining data center storage is already the largest line item in the IT hardware budget. But, with an accelerating demand for added capacity, IT management must now develop new and more sustainable practices and processes to adequately manage data growth. This presentation will review the storage-based technologies and best practices for managing at petabyte scale.
Agenda

Storage Technologies for 2012 and Beyond

- Can you manage storage at Petabyte scale?
- The impact of Pervasive Computing
- Scale-out or Scale Up
- Data Protection at Petabyte Scale
- Cloud storage = Data Dumpster?

Storage for Virtualized IT

- What changes for storage when you go from physical to virtual servers?
- Storage considerations for Virtual Desktop
Storage Technologies for 2012 and Beyond
Are Petabytes in your Future?

- Growth in capacity demand
  - 50-100% annual growth – compounding
  - 99% of all documents are in digital form now
  - ~50 billion emails per day
  - Unstructured data (data not in databases) is largest and fastest growing type of data

- Demand changes with information captured from Pervasive Computing moving into IT – the Big Data Problem
Are Petabytes in your Future?

50 - 80% will end up in traditional Data Centers

Data generated from traditional Data Centers

Non-Traditional Data Growth

3-5x CAGR

Traditional and Non-Traditional Data Center Growth
Outlook 3 -5 years
Managing at Petabyte Scale

- Do you need another/newer data center?
- Is managing your environment so complex that you have to tear it up and start over?
- Will you migrate the old to the new and sow the seeds that recreate the same old problems?
- Is storage the largest line-item in the hardware budget?
Managing at Petabyte Scale

- Can you manage a capacity growth rate of TBs per week? Per day?
- Are your current data protection processes sustainable?
- Will you add complexity as you add capacity?
- Will petabyte-scale storage break the budget?
Scale-out or Scale Up?

What to evaluate in scale-out storage

- Attributes of the global namespace file system
- The scale-out architecture
- Thin provisioning
- Data reduction
- SSD as “Tier Zero” or cache
- Inside and outside the box tiering
- Snapshots and clones
- The management app

What to evaluate in scale-up storage

- Virtualization
- Thin provisioning
- Data reduction
- SSD as “Tier Zero” or cache
- Inside and outside the box tiering
- Snapshots and clones
- The management app
Are your existing data protection processes sustainable at Petabyte scale?

- Time to get the backups done – growing capacity
- Amount of space consumed for backups - generations
- Cost of technology transitions – capital and migration
- Cost of backup software and administration
## Business Continuance
Continuation of full operations. Requires DATA, platforms, applications, and people for seamless failover and continuance.

## Disaster Recovery
Resumption of full operations after a period of recovery time. Requires DATA, platforms, applications, and people. Recovery operations may be involved.

## Data Protection
Data protection from accidental, inadvertent, or malicious destruction or corruption. Restoration to a known point (RPO) within a given time (RTO).
90% Drop in access rate within 60 days
Is RAID Sustainable at Petabyte Scale?

The RAID problem

- Large capacity disk drives take a long time to rebuild
  - Potentially days
  - Probability of a second error occurring is high – loss of data
Selectable data protection levels
- Example: 12 of 16 drives must be present – tolerate loss of 4
- Example: 12 of 16 sites must respond with date – tolerate 4 sites not responding

Potential benefits
- Eliminates RAID scale issues
- Potential for data protection and business continuance with same architecture
- Failures tolerated over time – no replacements expected (with warning and exception conditions)
- No service planned
  - Reduces warranty cost for vendor
  - Reduces service interruption / impact for IT
- Protection set calculated for lifespan of storage system
Is the Storage Cloud in your Future?

Data Ownership Models:

1. All data owned and in Corp. Data Center
2. All data owned but split between traditional data center and cloud
3. Blend of owned and “rented” data in both the Corp. data center and cloud

Non-Traditional Data Growth
3-5x CAGR

Traditional and Non-Traditional Data Center Growth
Outlook 3-5 years

(c) Evaluator Group, Inc
Storage in the Cloud

Information typically stored in the cloud
- Data when SaaS is used
- Archive information
- Backup data / Replication for DR
- Content distribution

Technologies used
- Applications that write data directly to cloud
- Storage-based migration software
- On-ramps – devices that stage data and move to the cloud over time
- Storage systems that make copies to cloud locations
Storage for Virtualized IT
Virtual Machines & Storage

Servers with
- More processors and more cores
- Multiple applications on same physical hardware

Create storage problems
- Not equivalent number of storage capabilities
  - Paths
  - Cache
  - Disk drives
- Creation of bottlenecks
- Need a storage system and network that can handle virtualized environment
High-value storage system features for server virtualization

- SSD
- Array-based snapshots and clones
- Thin provisioning
- In-the-box tiering / caching
- Array management app
- Array features specific to vendor – VAAI, VASA support, etc.
Virtual Desktop Infrastructure (VDI) and Storage

Same problems as multiple virtual machines

Except – multiplied

- More users
- Likelihood of “boot storms”
- Simultaneous application startup
- Protection and recovery – frequency and user-level control
- Provisioning

Duplicate Data: OS and App > 90% same

All data – Thru Server

Storage – each App and OS contending for resources
Sample Results From VDI Storage Benchmark Runs

- I/O ranges from 5 to 20 IOPS per desktop
- Block sizes include 512, 1K, 2K, 4K, 8K, 16K, 32K, 64K, 128K, 512K, 1M and 2M
- VDI does most I/O from 4K – 2 MB
- Read / Write mix varies from 30 / 70 (r/w) to 60 / 40 depending on apps
- Overall Read / Write mix is about 50 – 50
  - Lighter workloads are skewed towards more writes
  - 4K blocks, are the most common for writes, 32K for reads
Storage for VDI

Snapshots
- Writeable and space efficient – ROW/COW
- Large number supported
- Cascadable – snap of snap

Thin Provisioning
- Allocate capacity on demand only
- Reclaim deleted space

Wide striping
- Maximize number of simultaneous disk operations

Caching
- Read and write data
- Managed cache and pinning

Tiering – SSDs and disks
High performance
Automated provisioning
Summary

- Prepare to manage storage at Petabyte scale
- Virtualization is now driving storage decisions
- Vendors are differentiating on the basis of advanced functionality expressed in software
The SNIA Education Committee would like to thank the following individuals for their contributions to this Tutorial.

Authorship History

John Webster
Updates:
  John Webster – 8/26/2012

Additional Contributors

Randy Kerns

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org