SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

EDUCATION

The Storage Evolution: From Blocks, Files and Objects to Object Storage Systems

Christian Bandulet, Sun Microsystems

SNIA Legal Notice

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- 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.

Abstract

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

The Storage Evolution: From Blocks, Files and Objects to Object Storage Systems

This session will appeal to CIOs, CTOs, Consultants, System Architects and Technologists, and those that are seeking a fundamental understanding of the emerging object-based storage technologies. The audience will gain insight into the basic differences of block-, file- and object-based data access methods. The session will delve into the benefits of object storage and it's value and also outline how this technology might impact future directions of storage system architectures.

Topics

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

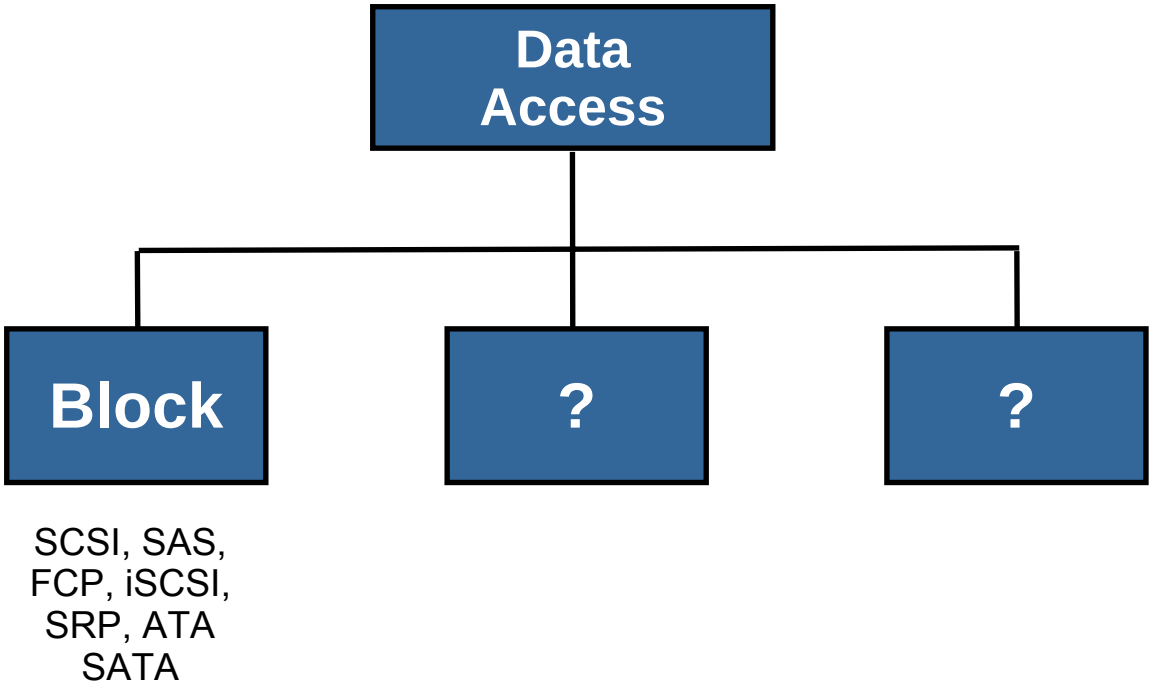
- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

Topics

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

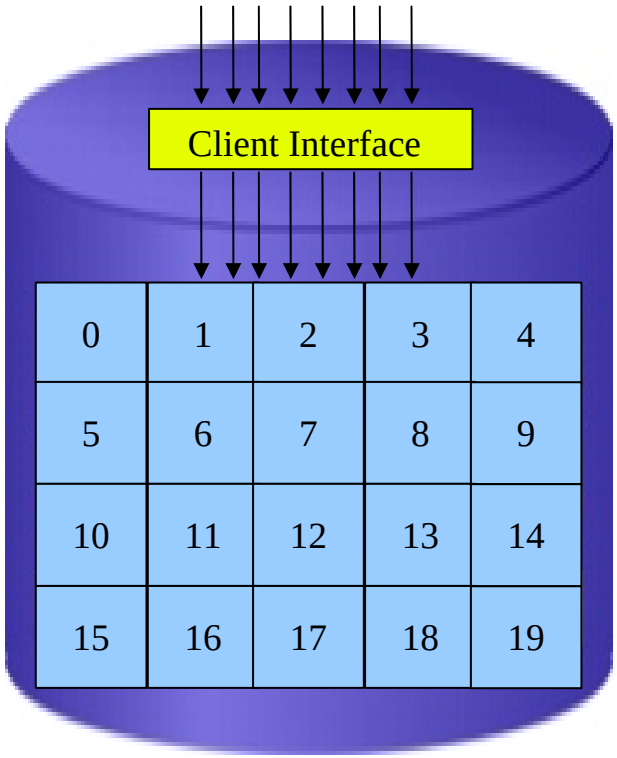
- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

The Data Access Taxonomy



The Block Paradigm

SCSI, SAS, FCP, SRP, iSCSI, ATA, SATA



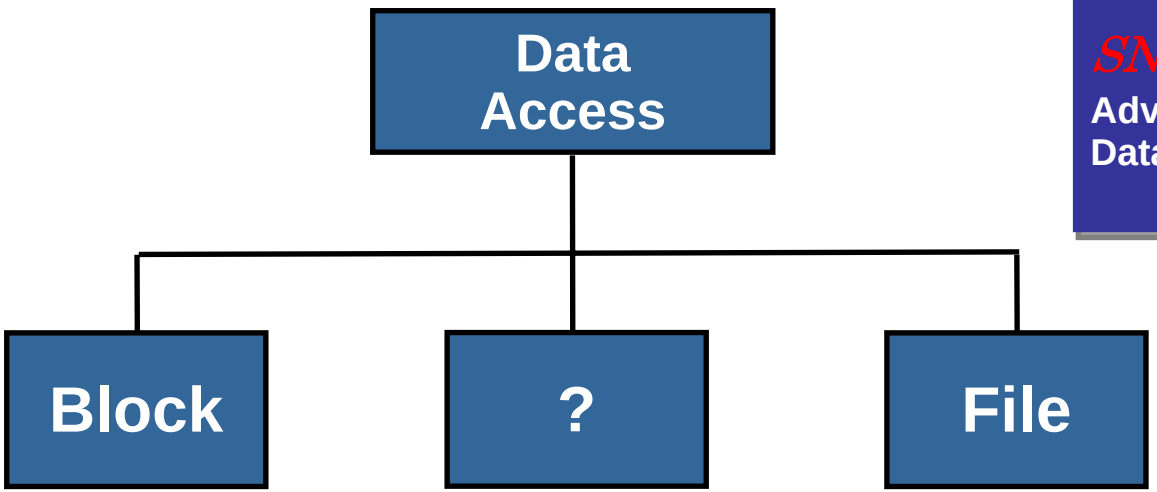
Physical Blocks:
e.g. 512 bytes

Topics

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

The Data Access Taxonomy



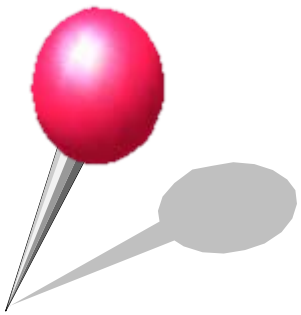
SCSI, SAS, FCP, iSCSI, SRP, ATA SATA

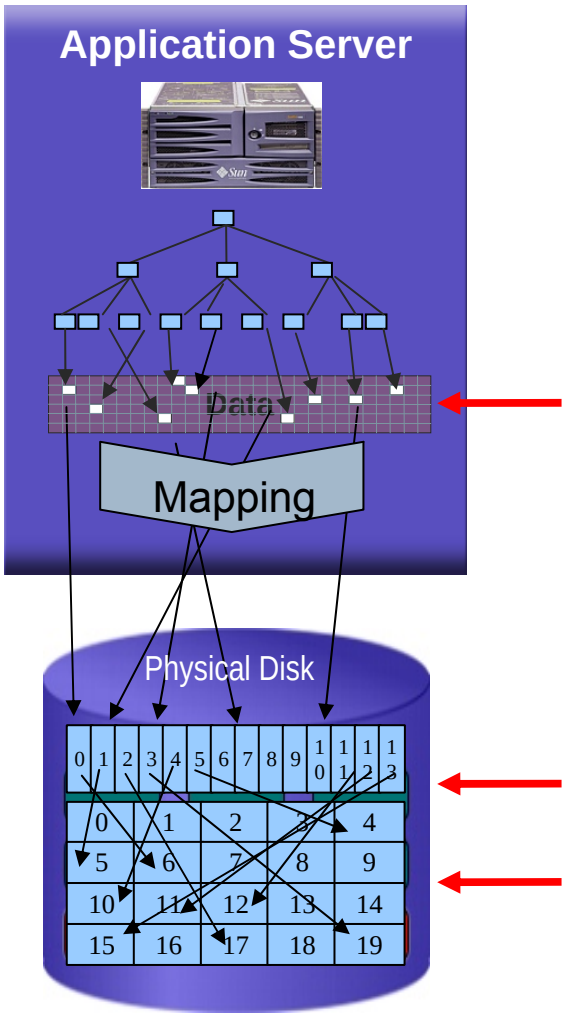
Local FS, Distributed FS: (NFS, CIFS, WAFS) Shared FS: (Cluster FS, SAN FS) Global, Distributed & Parallel FS

Local File Systems

One more level of indirection

- file/directory management(~**10%** of workload)
- block/sector management (~**90%** of workload)





File system structure (i.e. inodes)

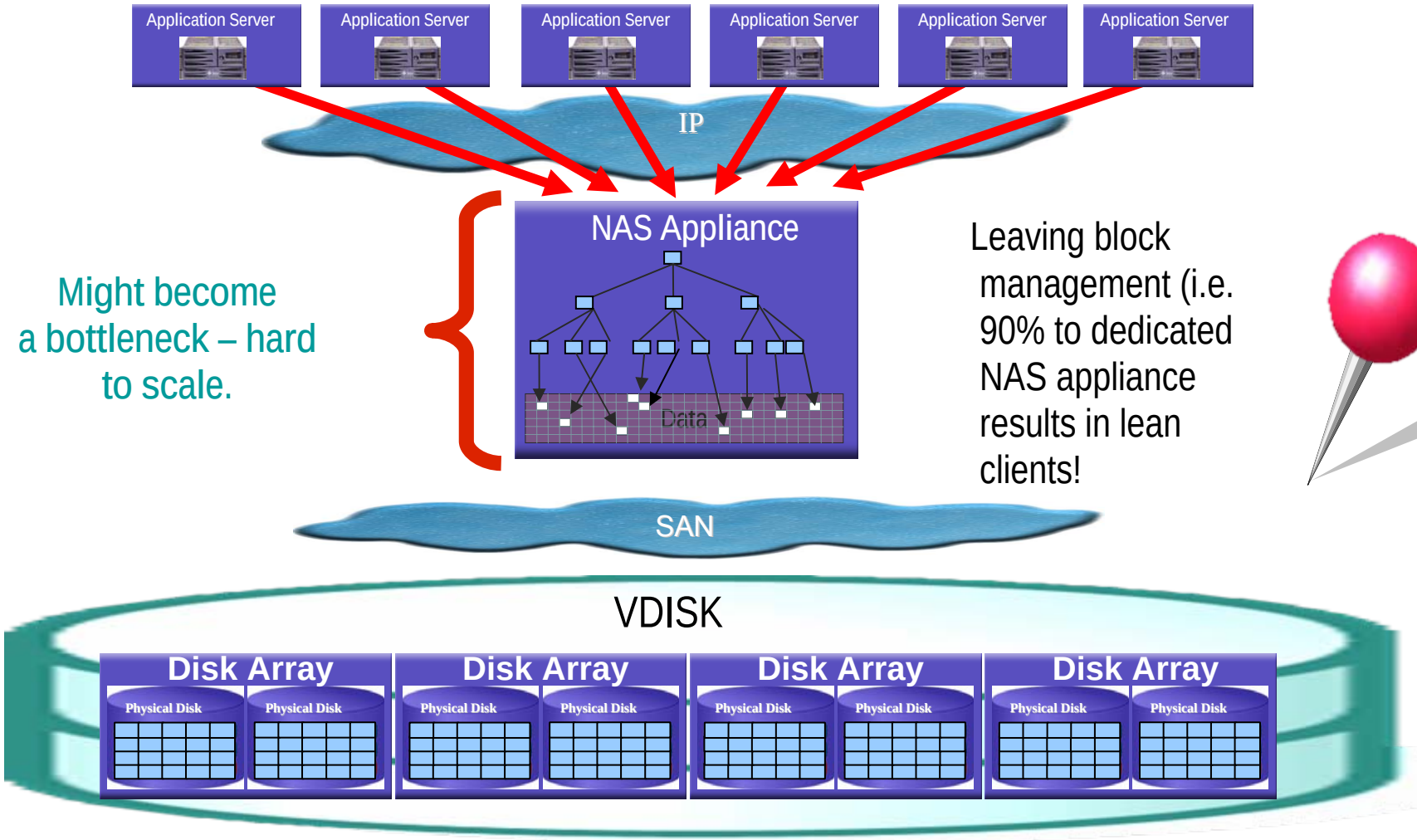
Logical Blocks on Disk

Physical Blocks on Disk

Distributed File Systems

e.g. NAS with NFS,CIFS Protocol

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION



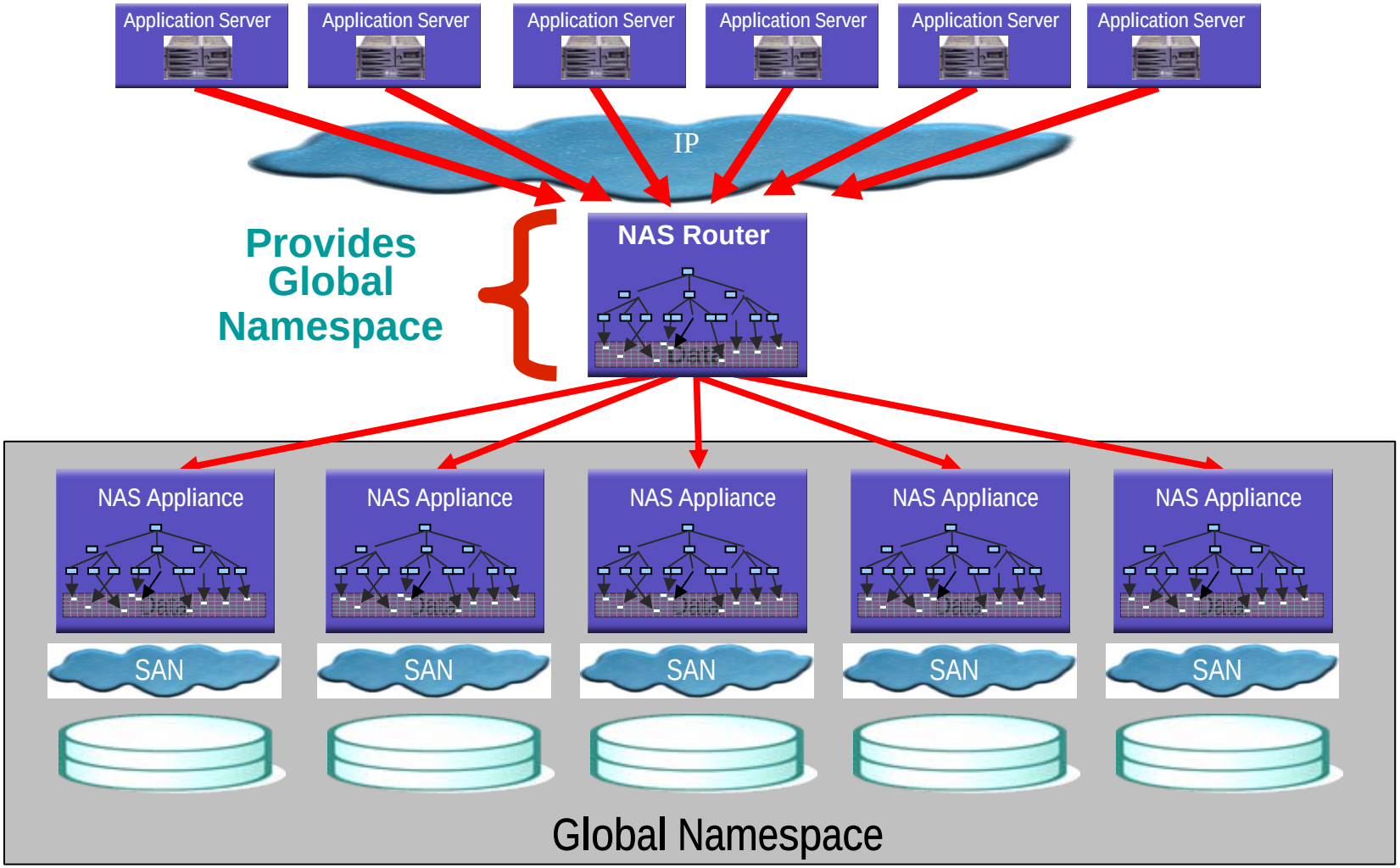
Might become a bottleneck – hard to scale.

Leaving block management (i.e. 90% to dedicated NAS appliance results in lean clients!

NAS Aggregation/Virtualization

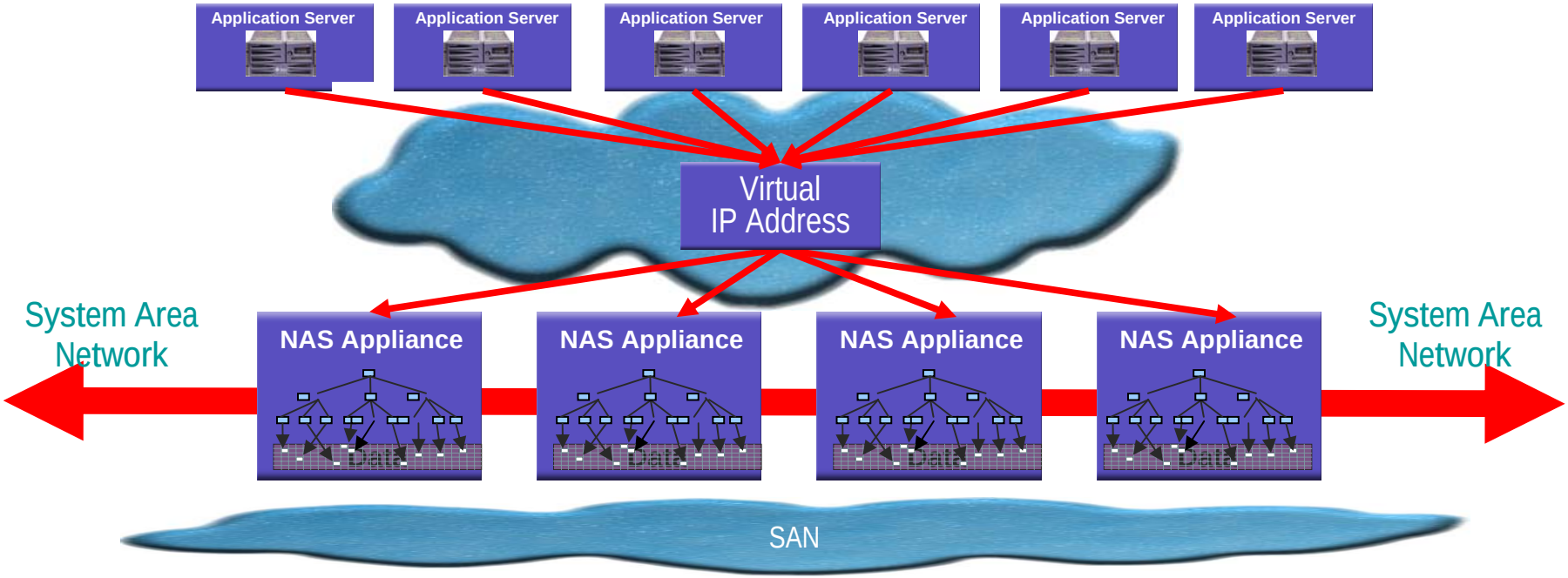
Global Namespace

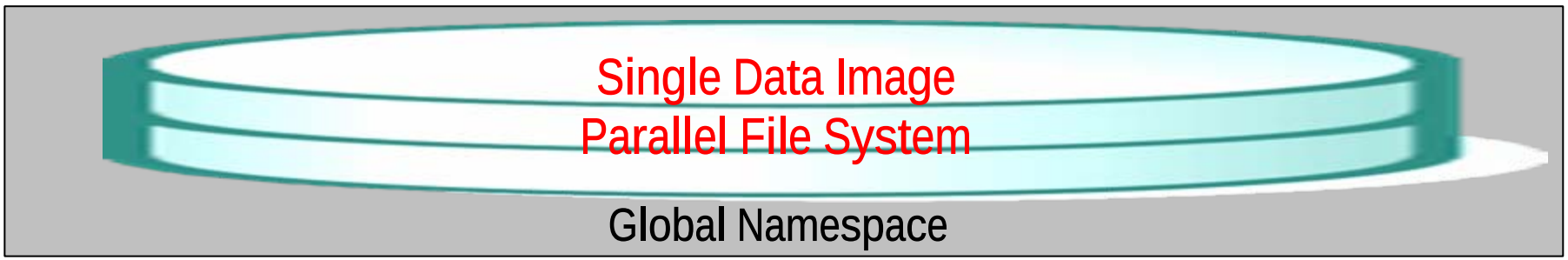
NAS Aggregation
Global Namespace



NAS Cluster

aka Tightly Coupled NAS



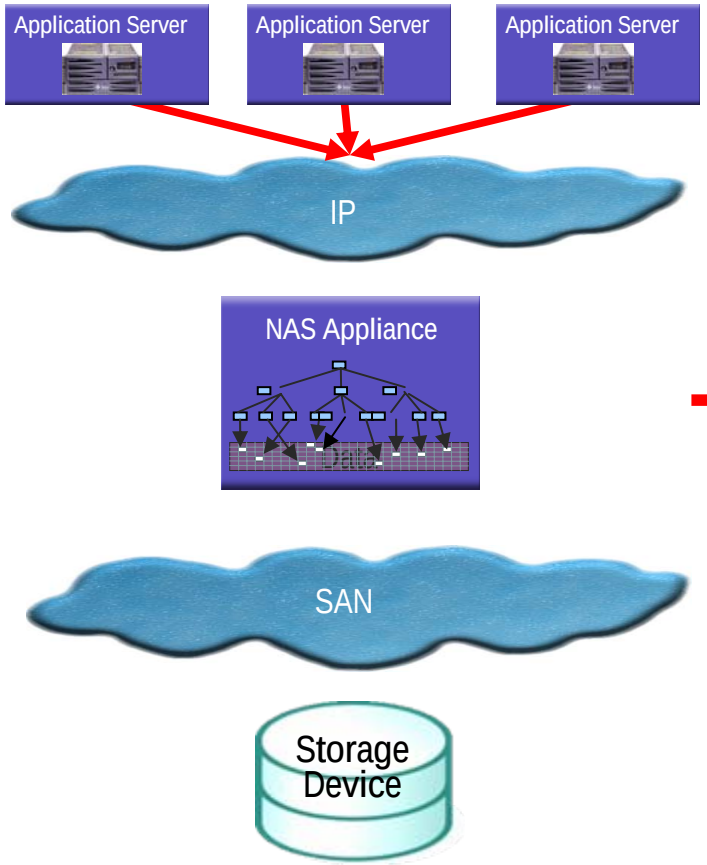


Scalable NAS

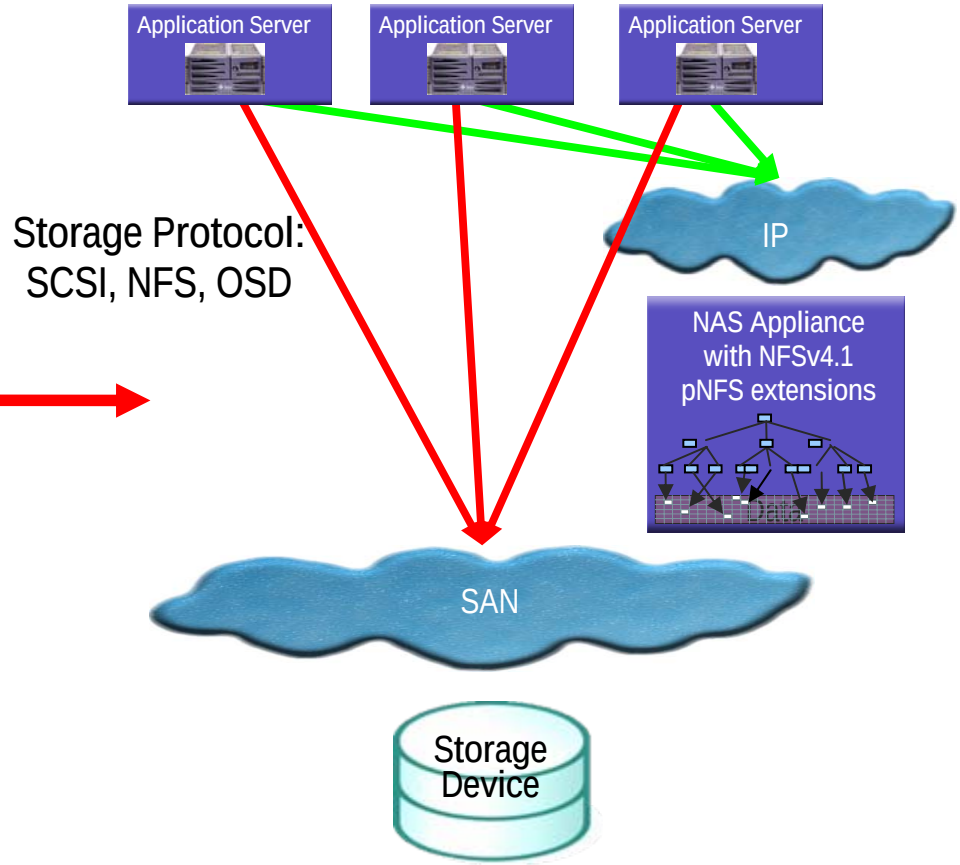
aka Loosely Coupled NAS
Global Namespace with NFSv4.1 and pNFS

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

In-Band NAS:



Out-of-Band NAS:



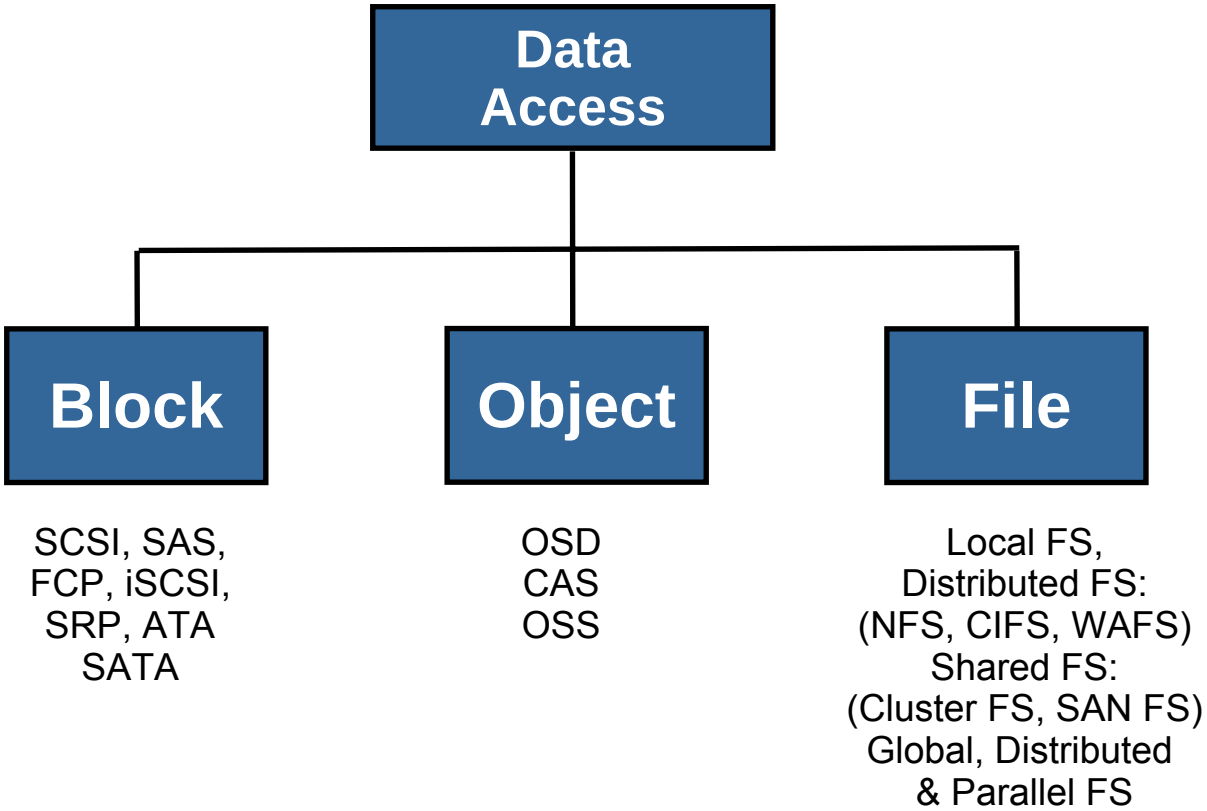
Scalable NAS
Loosely Coupled NAS Cluster

Topics

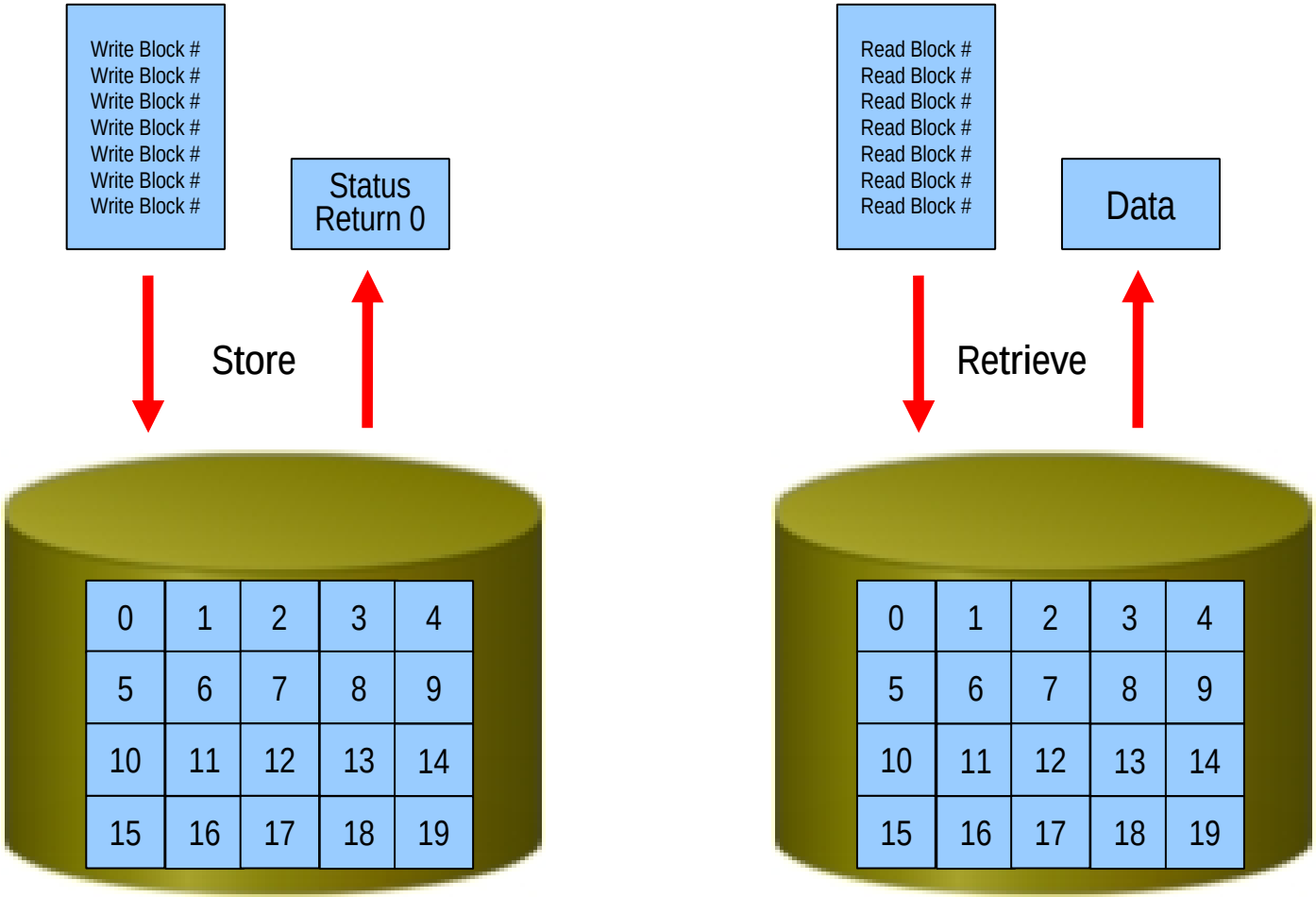
SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

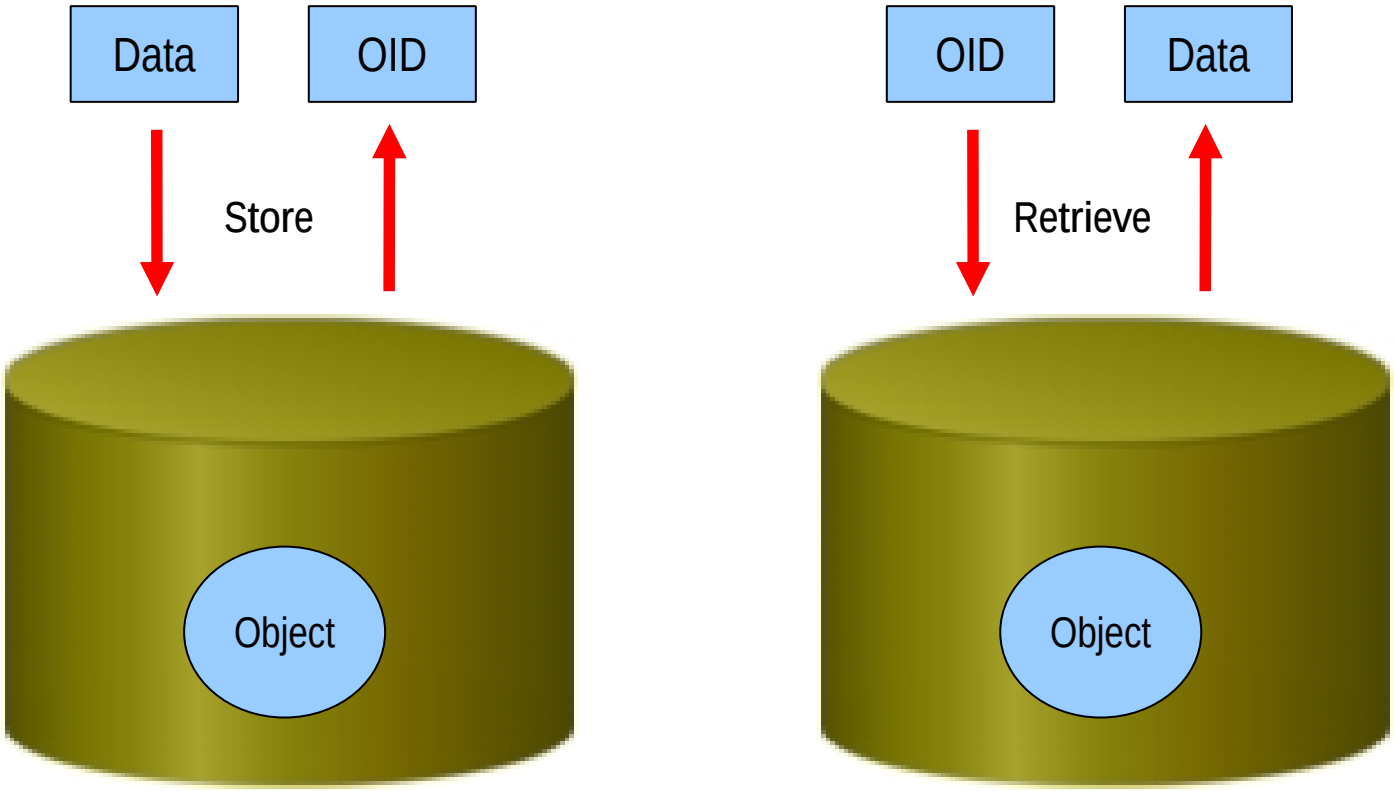
The Data Access Taxonomy



The Old Block Paradigm

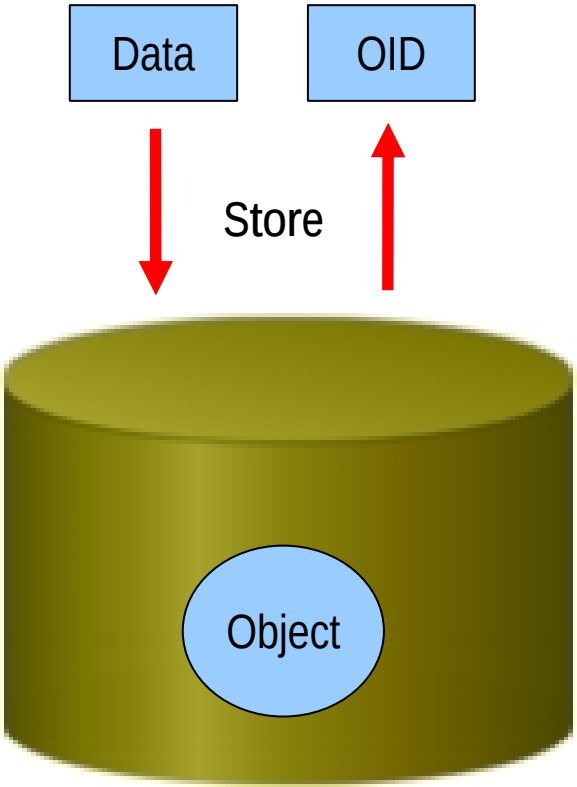


The New Object Paradigm



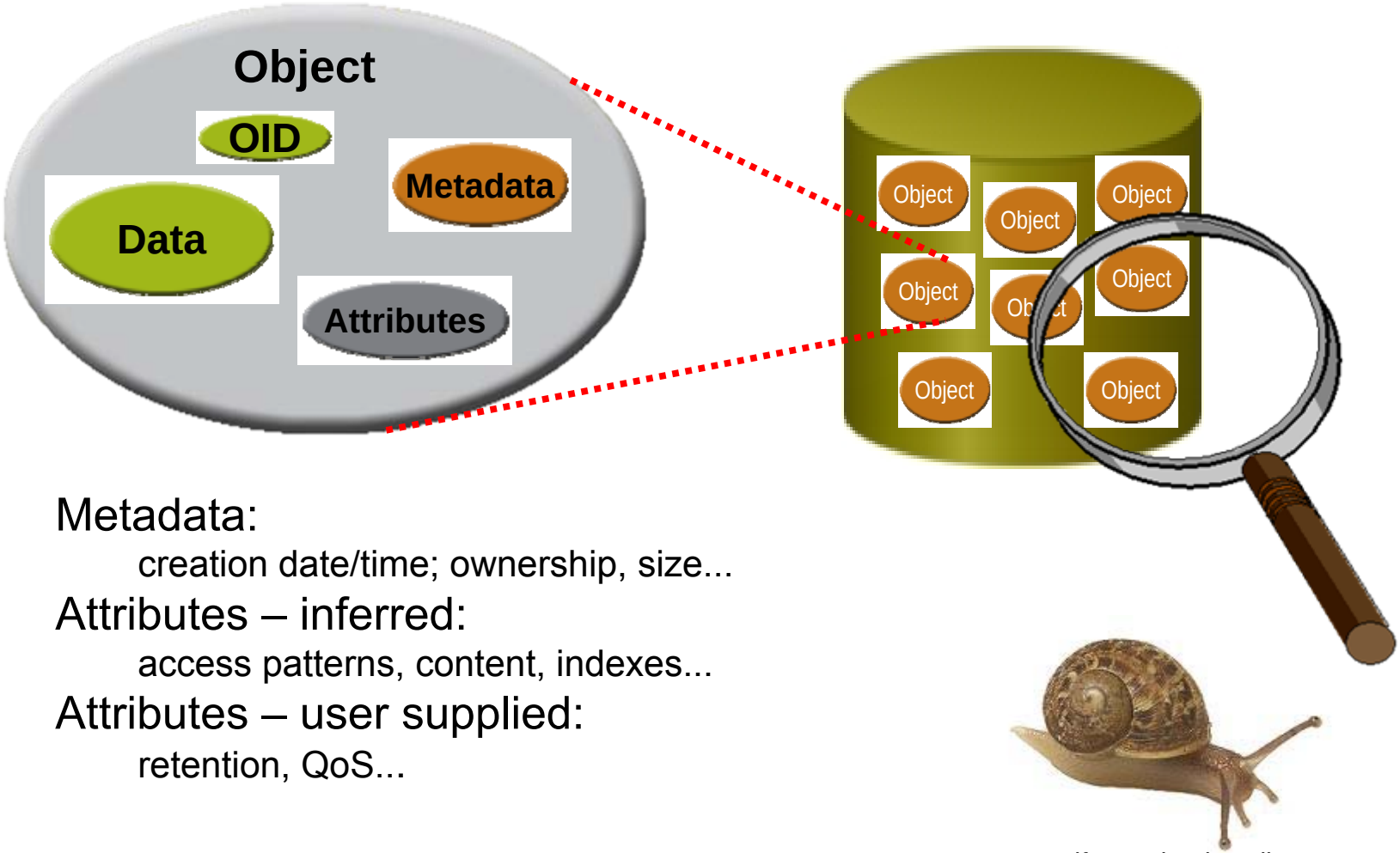
The New Object Paradigm
(cont'd)

- WRITE 26,763 Bytes
- QoS = High
- Description = “X-Ray”
- Retention = 50 years
- Access Key = *&^%#
- Data Payload.....



- Object Storage Responsibilities:
- Space Management
- Access Control (Identity Mgmt)
- QoS Management
- Cache, Backup
- Policy Migration, Retention

Self-Contained Objects



Metadata:

creation date/time; ownership, size...

Attributes – inferred:

access patterns, content, indexes...

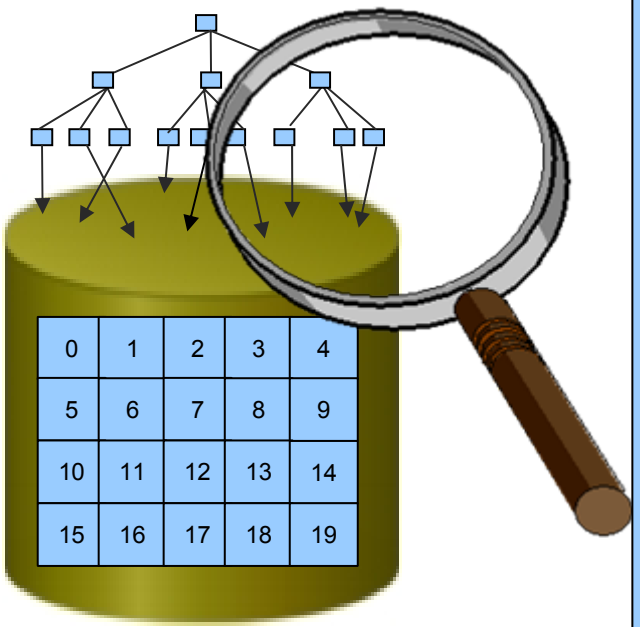
Attributes – user supplied:

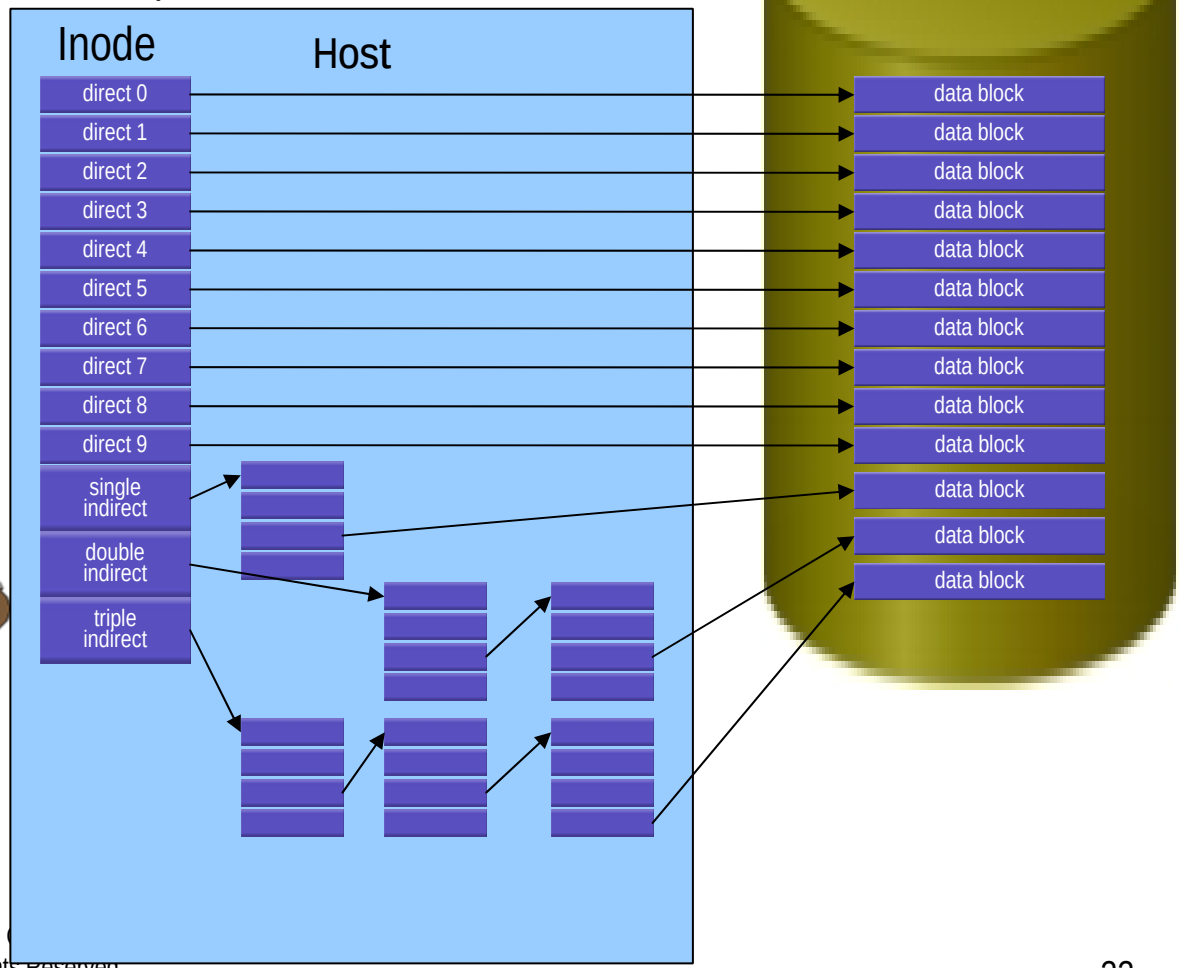
retention, QoS...

self-contained snail

Block Access - Inodes

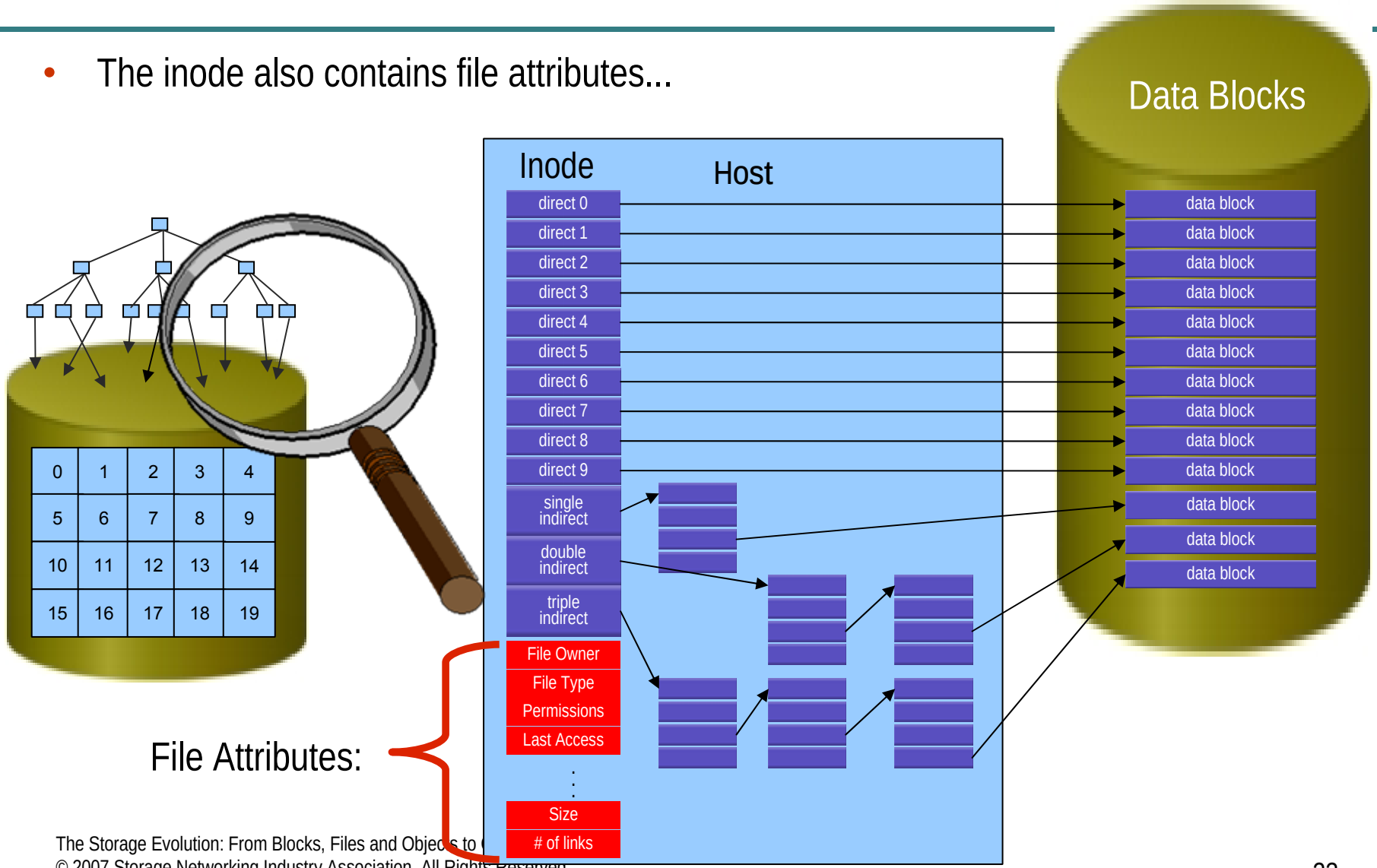
- The inode contains a few block numbers to ensure efficient access to small files. Access to larger files is provided via indirect blocks that contain block numbers



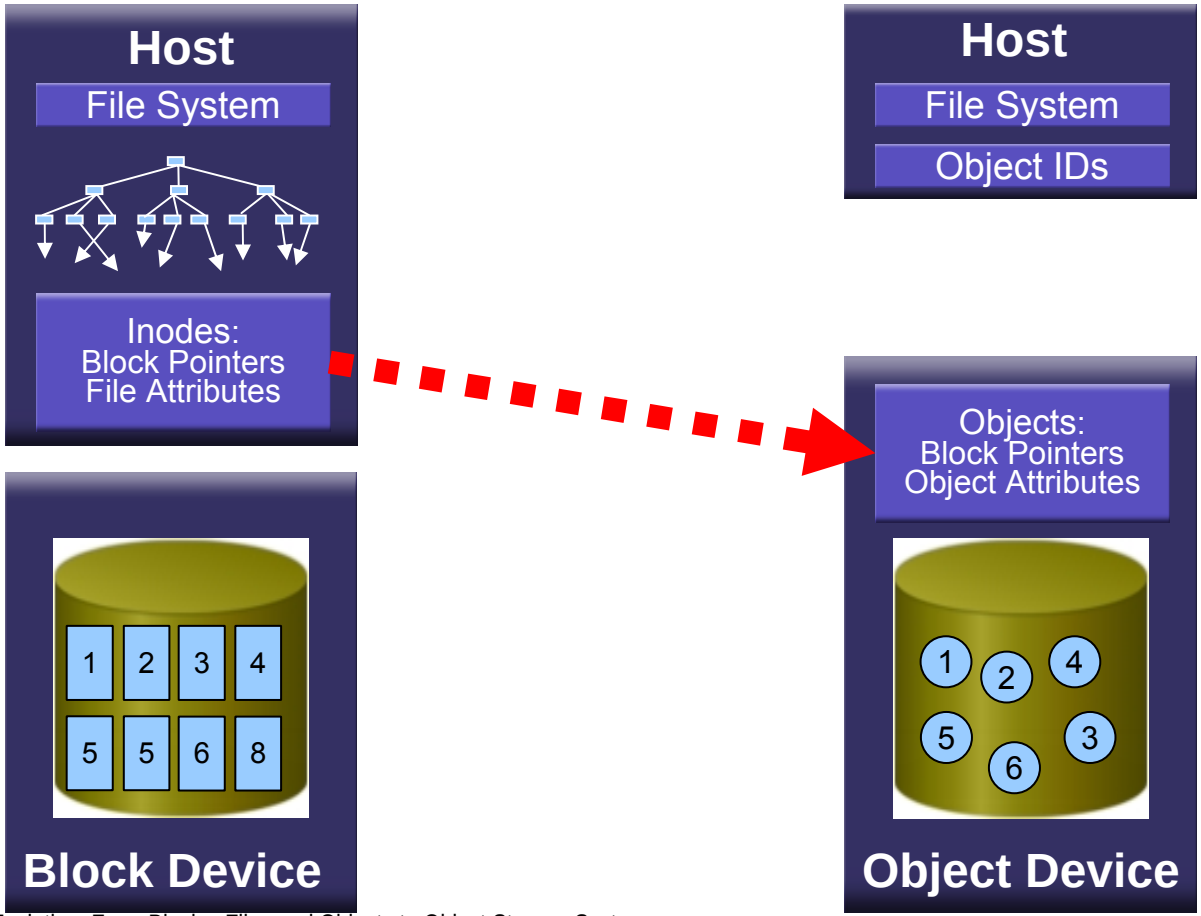


Block Access – Inodes (cont'd)

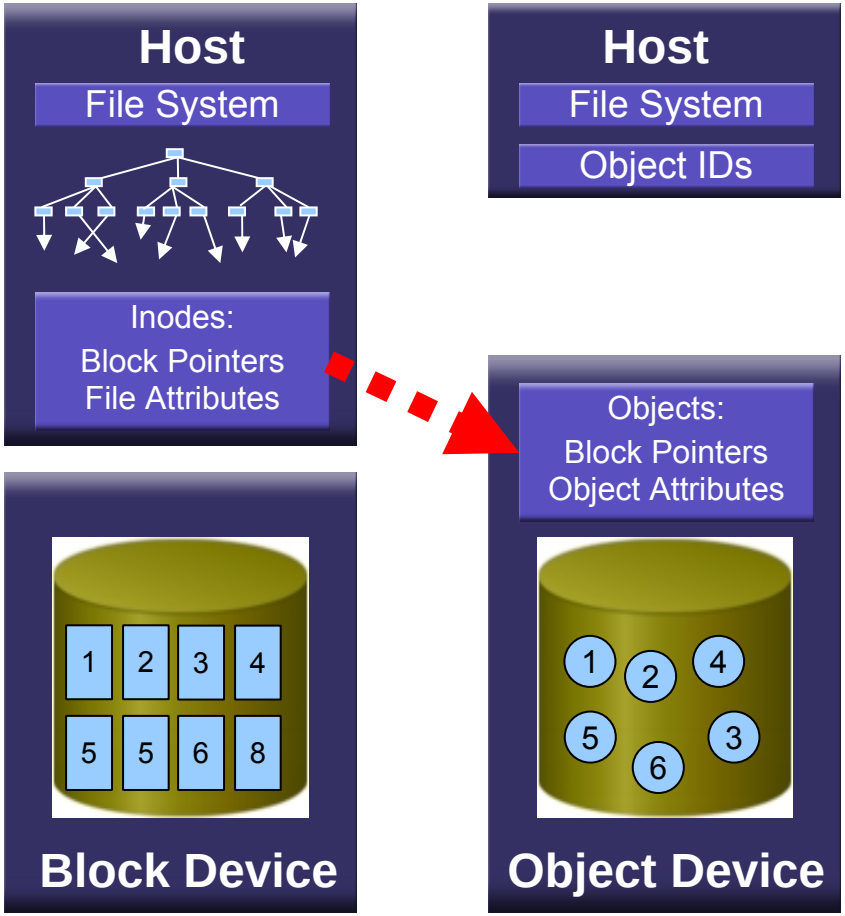
- The inode also contains file attributes...



Inodes vs. Objects



Object Autonomy

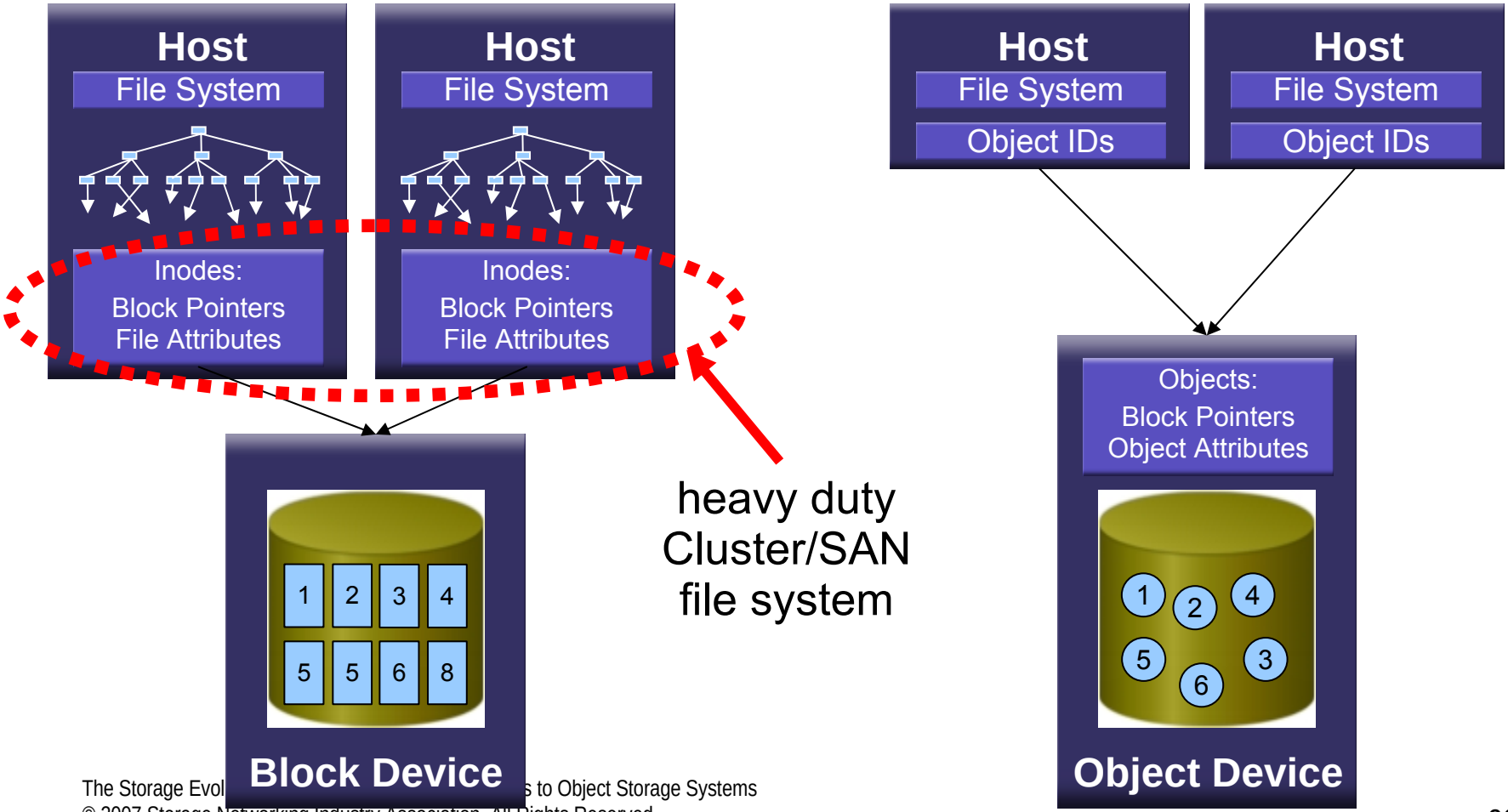


Storage becomes autonomous:

- capacity planning
- load balancing
- backup
- QoS, SLAs
- understand data/object grouping
- aggressive pre-fetching
- thin provisioning
- search
- compression/de-duplication/encryption
- strong security
- compliance/retention/secure delete
- availability/replication
- audit

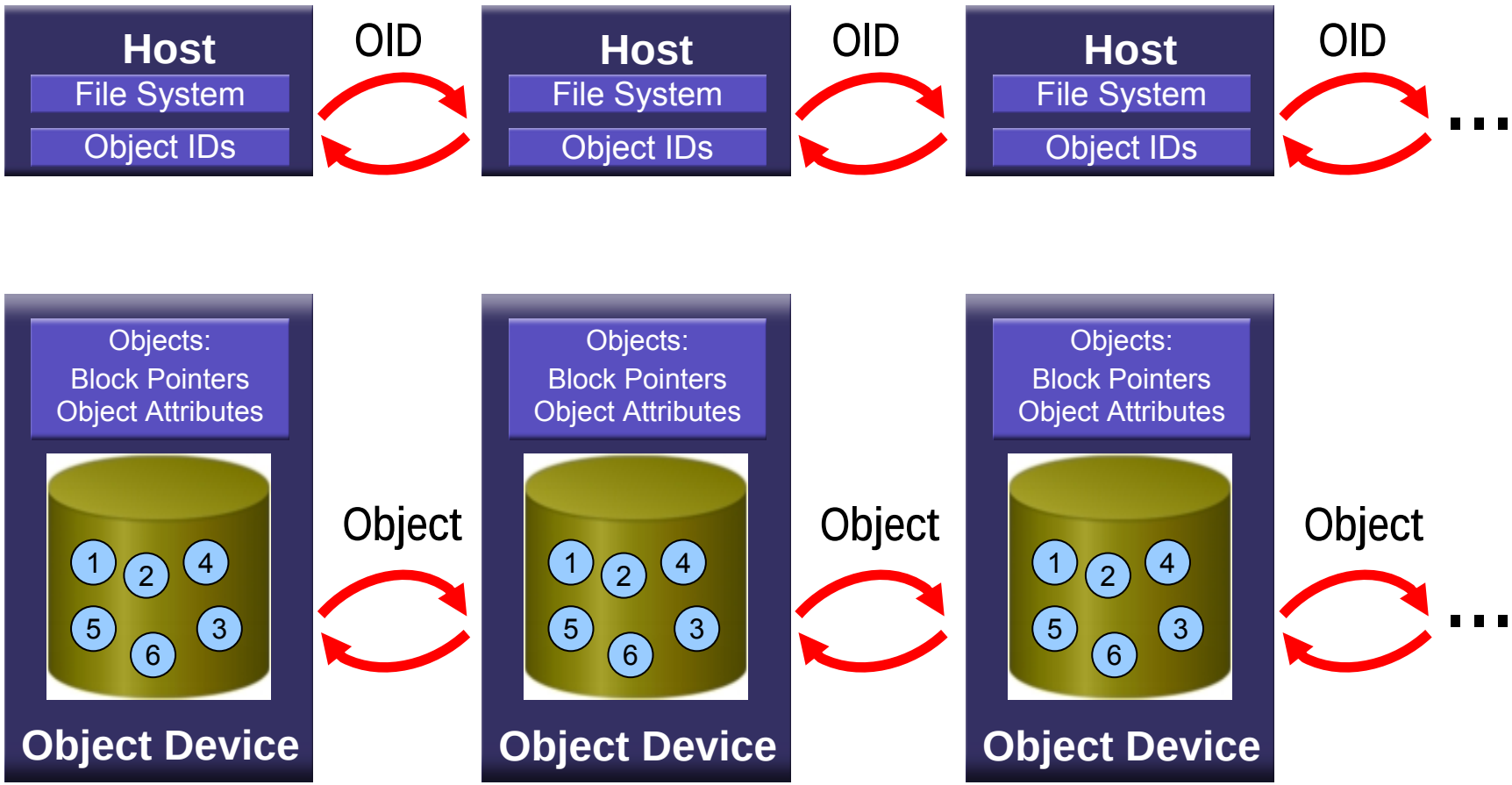
⋮

Data Sharing
Homogeneous/Heterogeneous

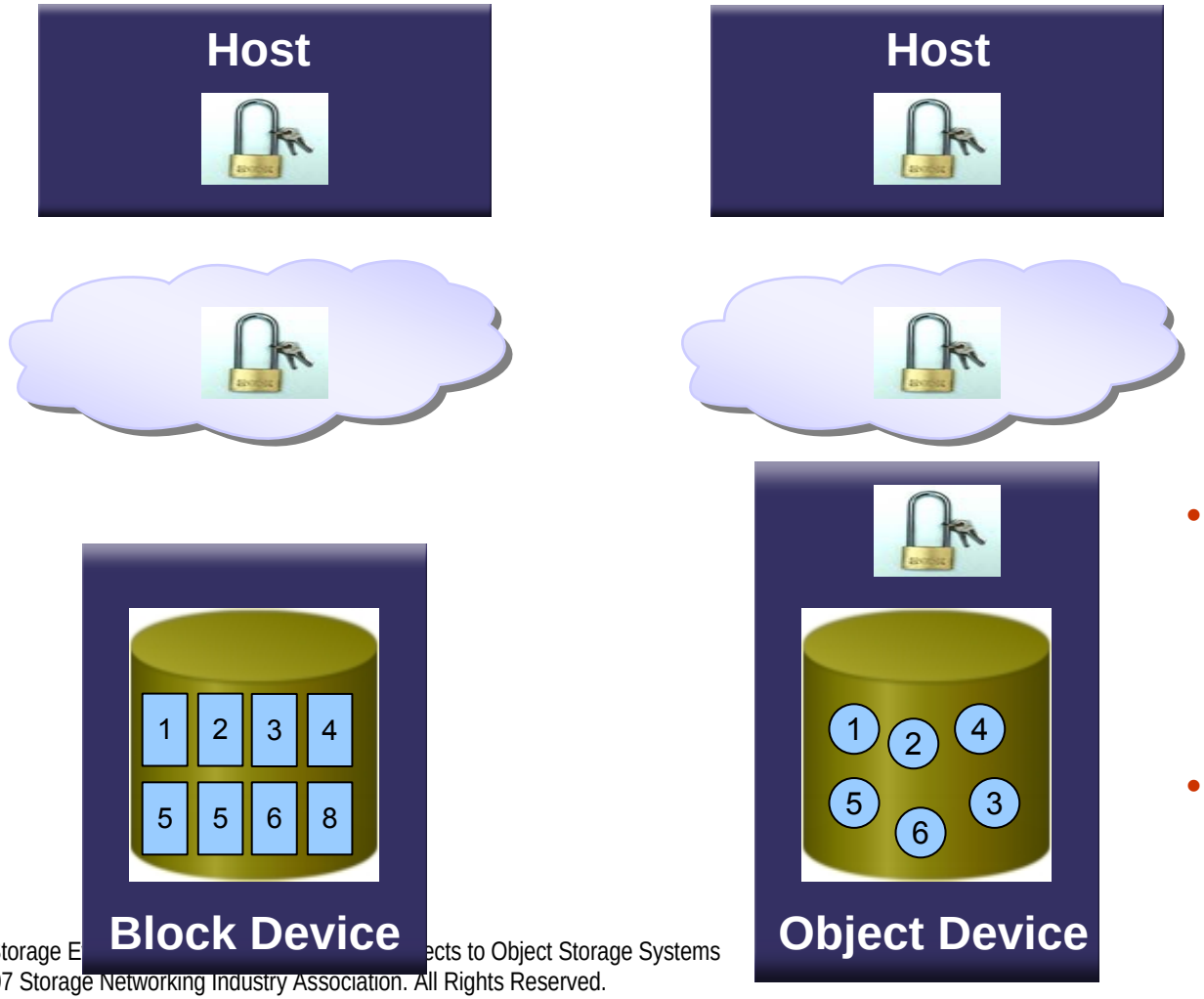


Data Migration - ILM

Homogeneous/Heterogeneous



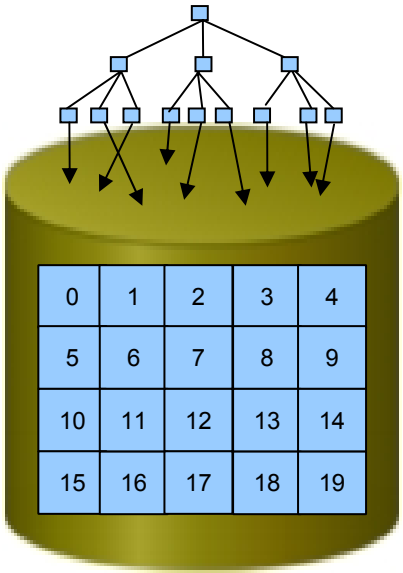
Additional Layer of Security



- strong security via external service
  - authentication
  - authorization
  - NIS, LDAP....
- fine granularity
  - per object

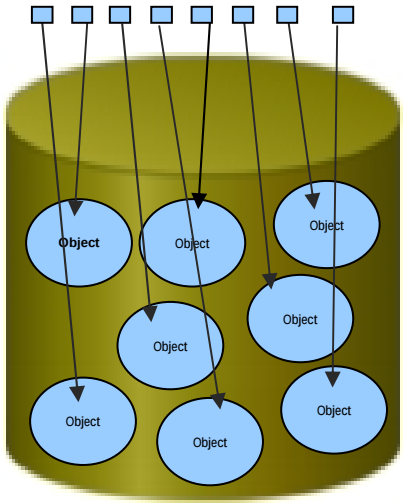
Living in a Flat Namespace

File names / inodes



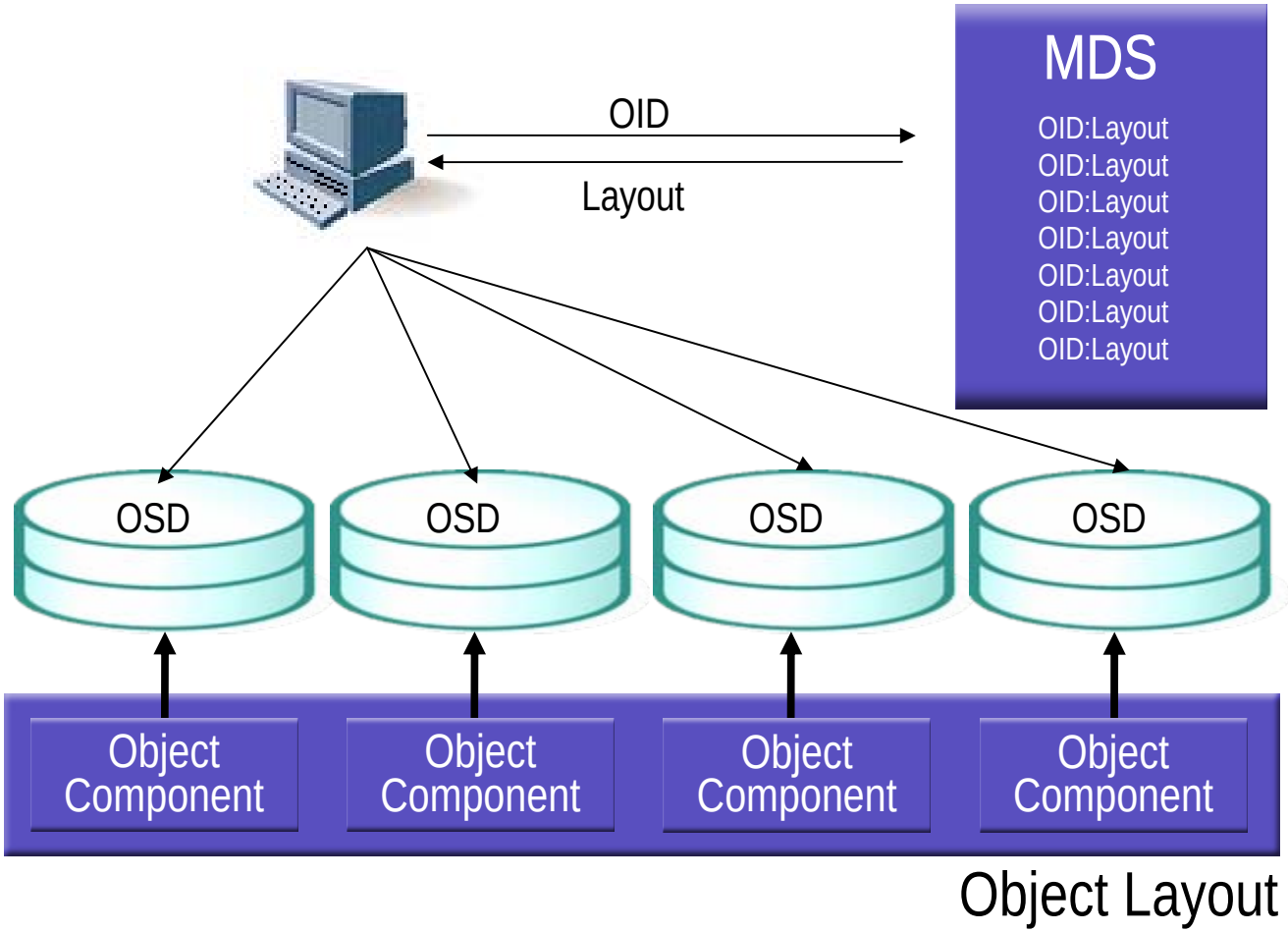
Traditional
Hierarchical

Objects / OIDs

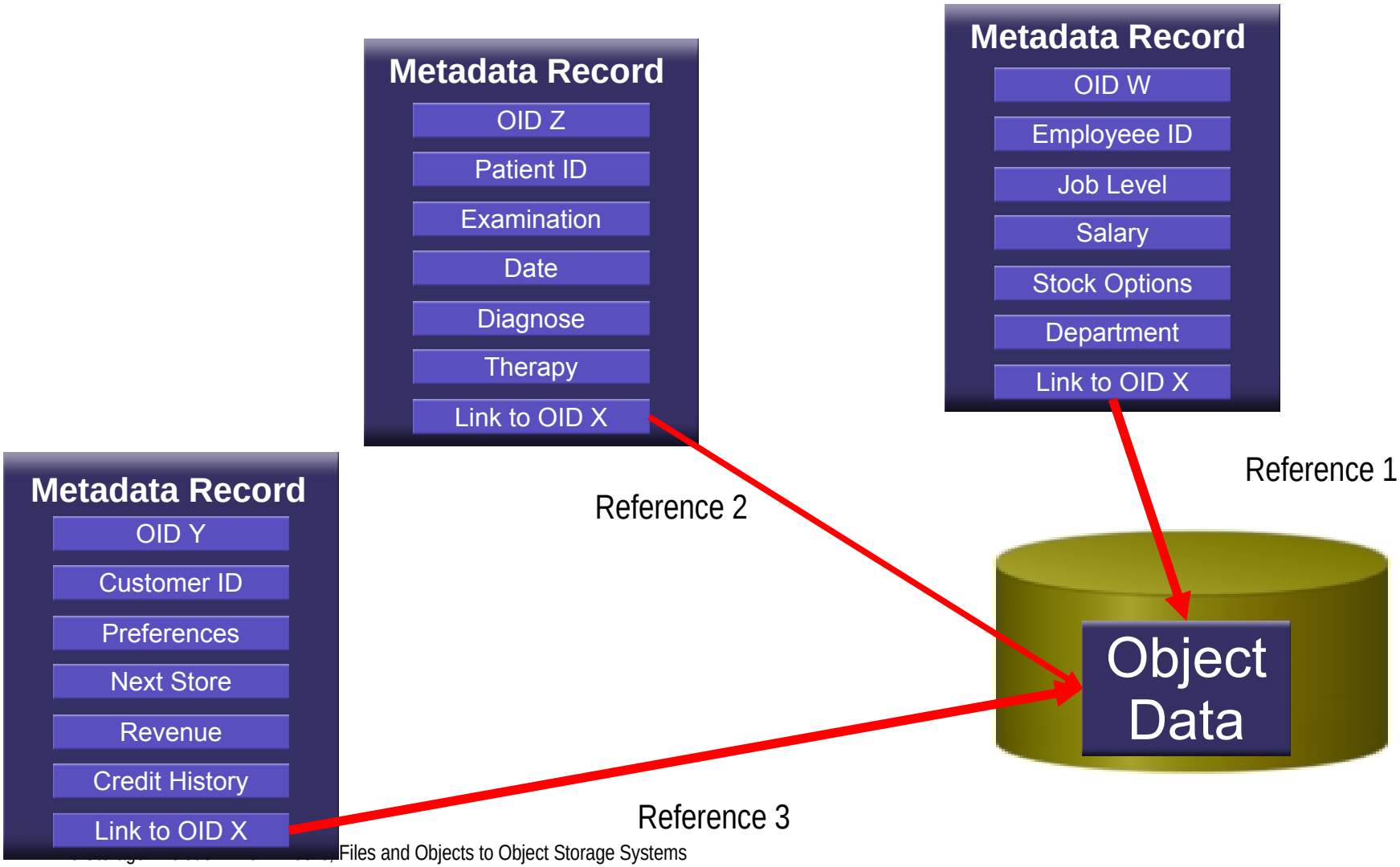


Flat

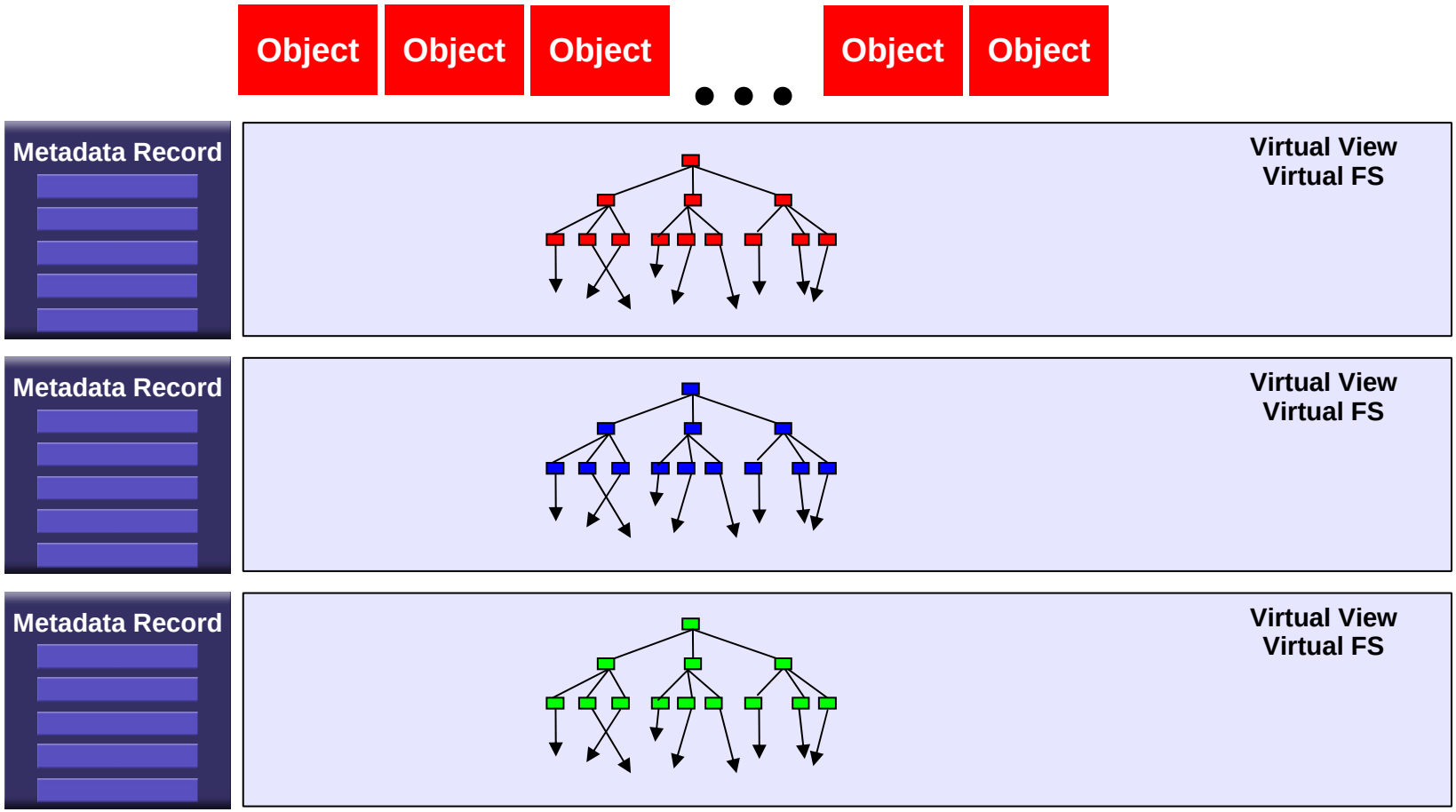
Object Decomposition



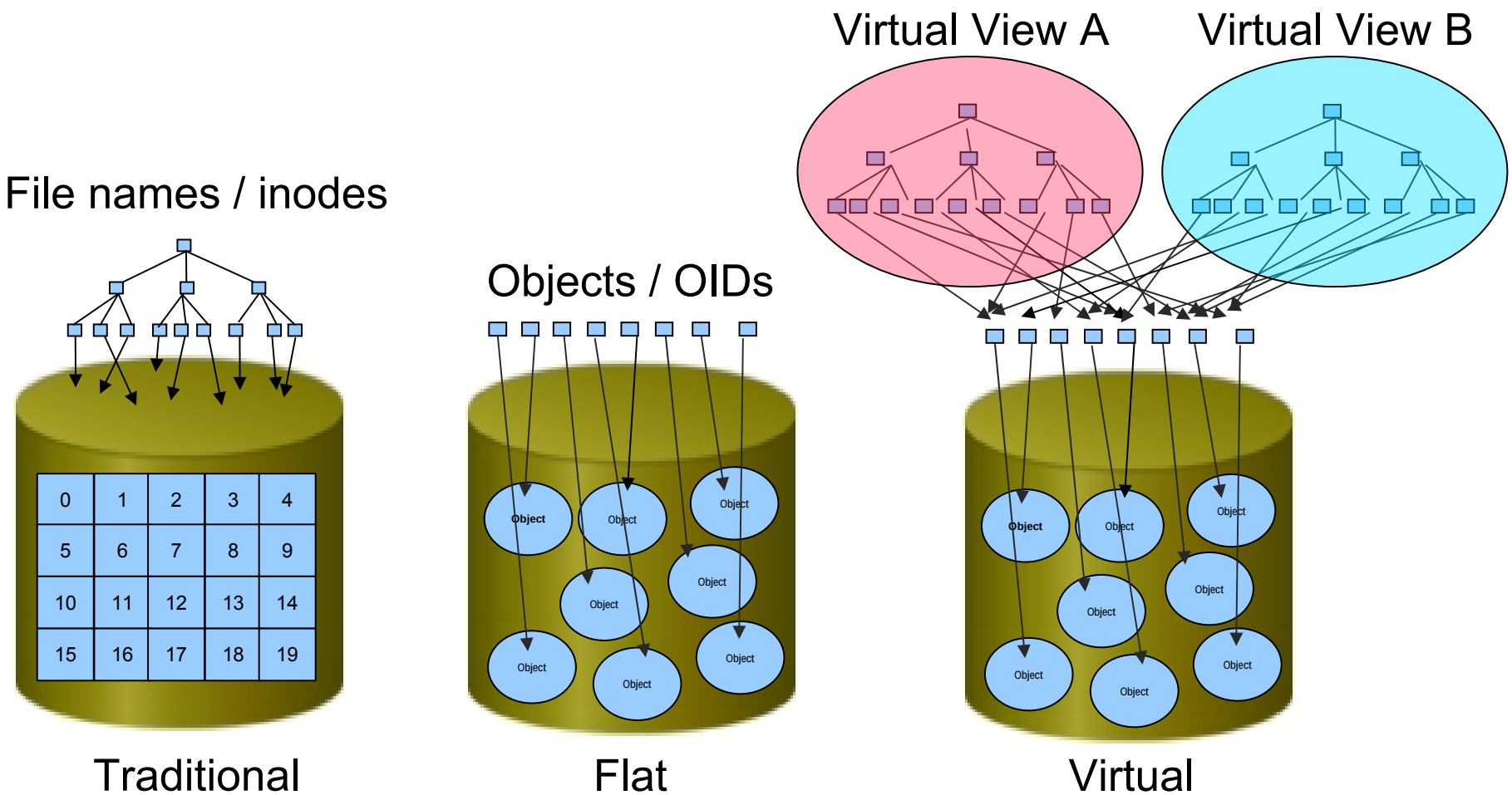
Multiple Referenced Objects



Virtual View / Virtual File Systems



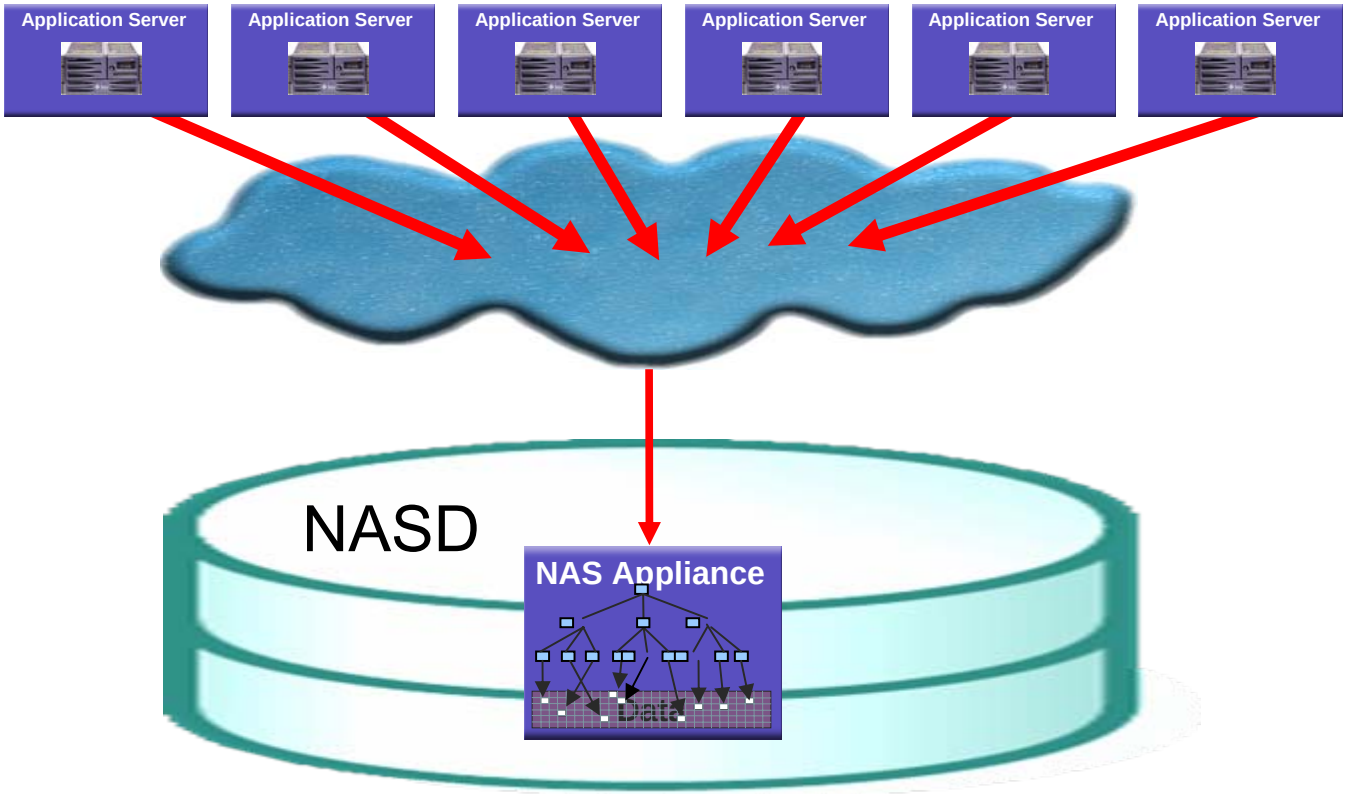
Virtual View / Virtual File Systems (cont'd)



Topics

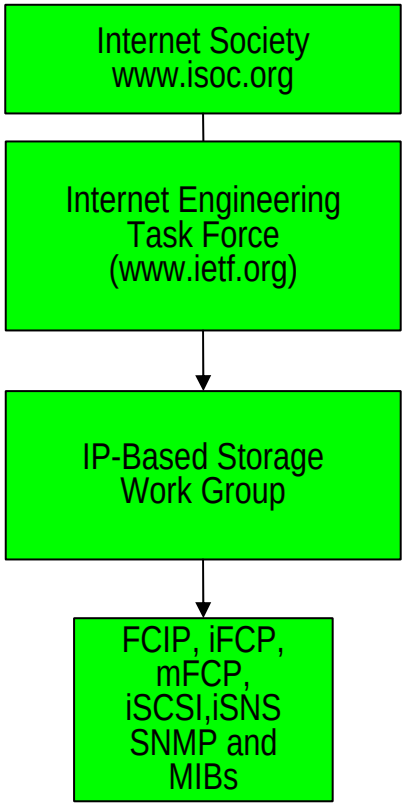
- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

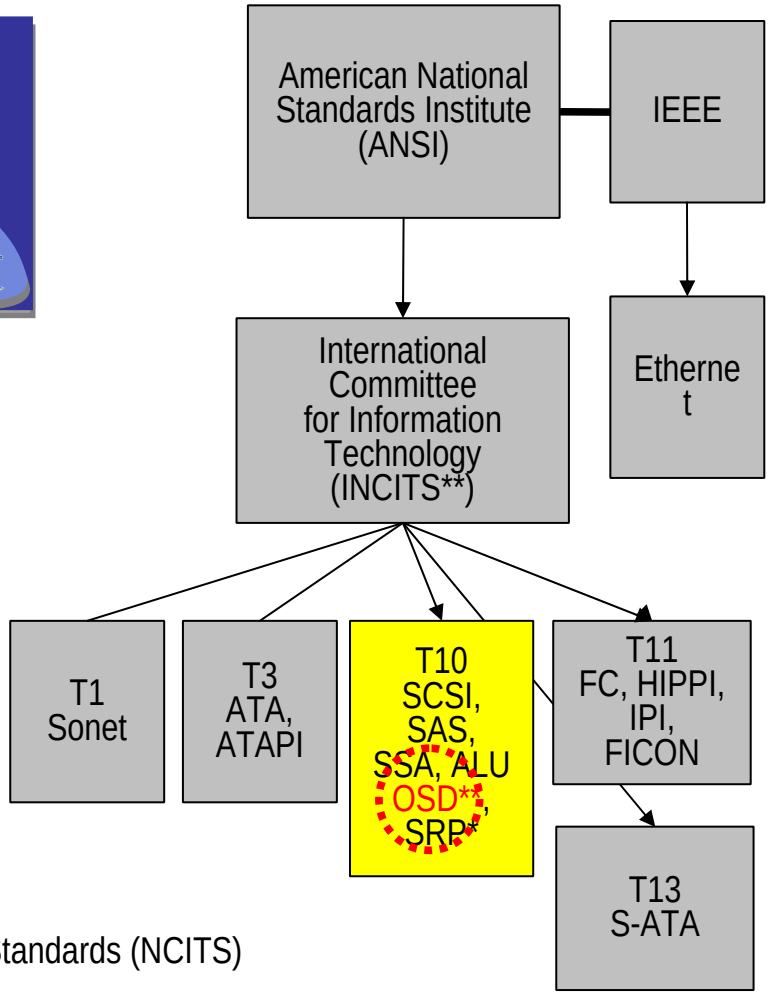
NASD
Network Attached Secure Device



- bring the whole functionality of a NAS device down to a SCSI devices

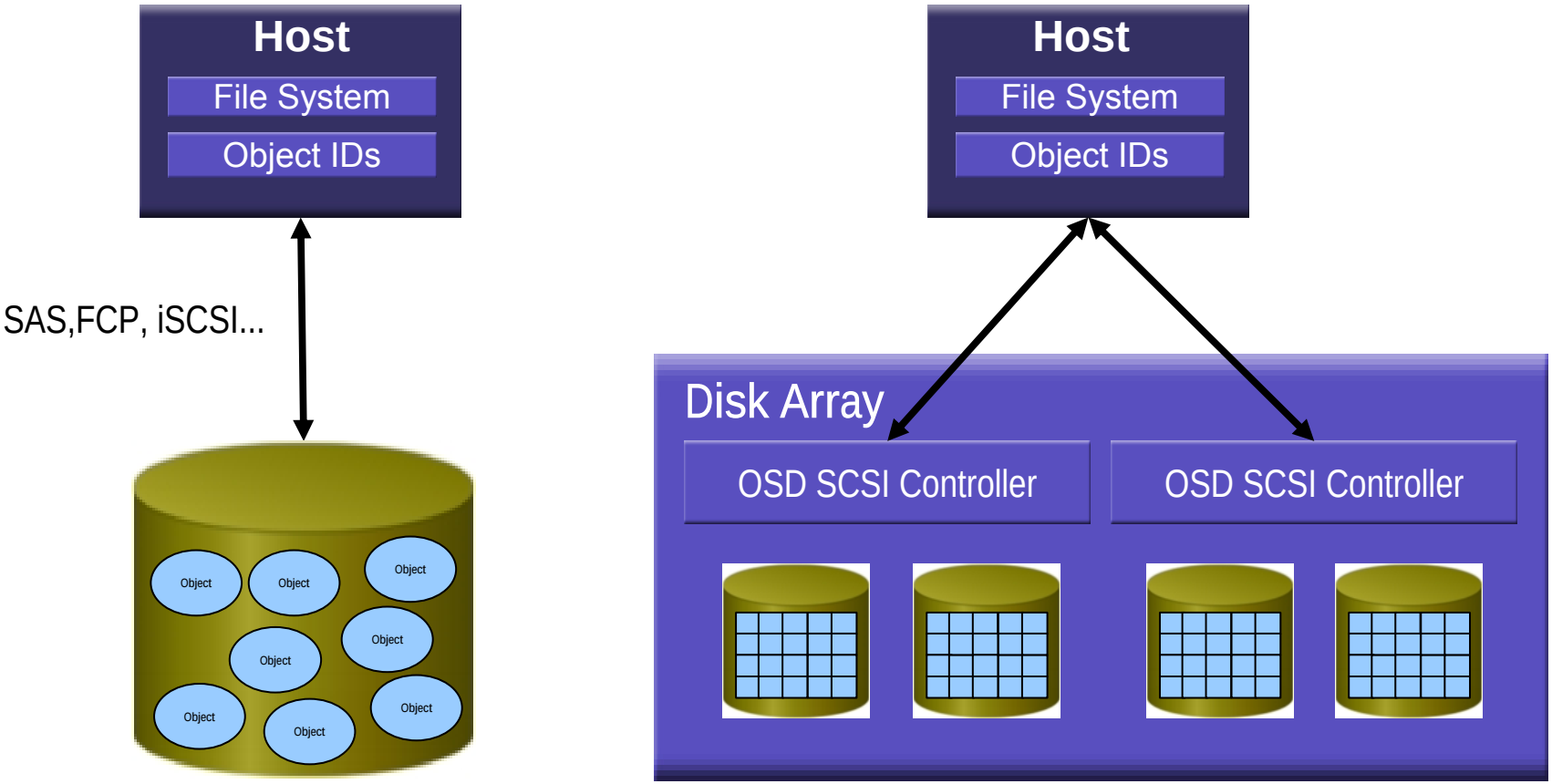
The World of Standards





Note*SRP: SCSI RDMA Protocol
Note**OSD: Object-based Storage Devices
Note*** INCITS – formerly National Committee for Information Technology Standards (NCITS)

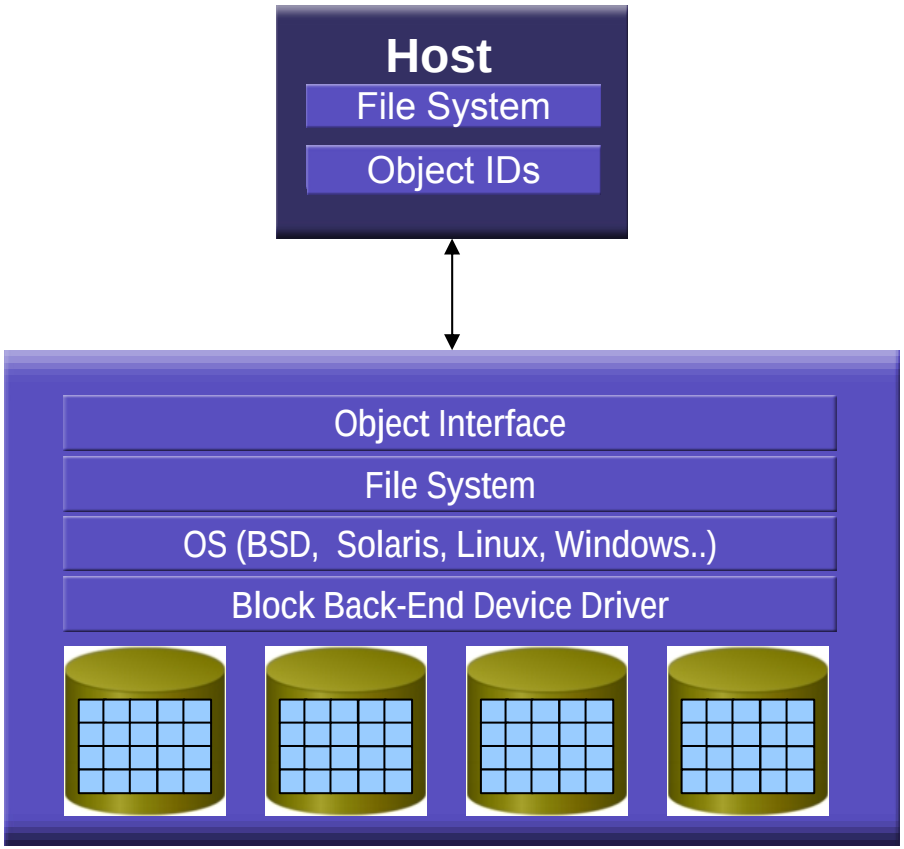
ANSI T10 OSD SCSI Targets



Topics

- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

Object Storage Server - OSS



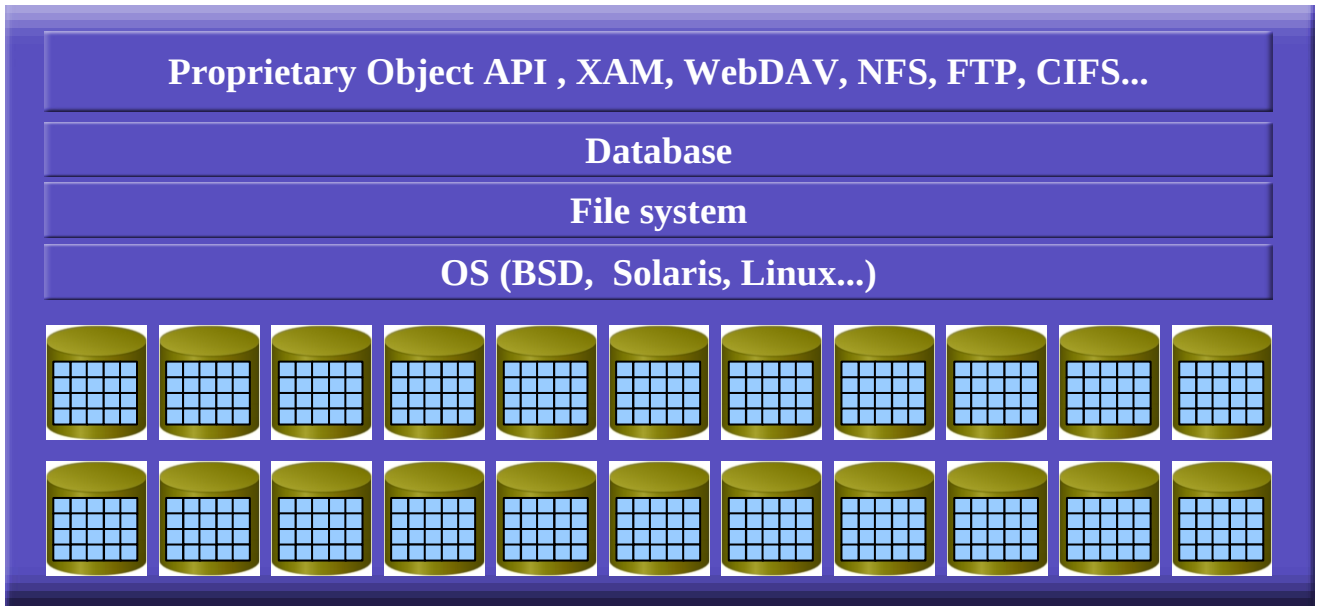
OSS could be a migration path to provide object technologies to legacy block devices

Topics

- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

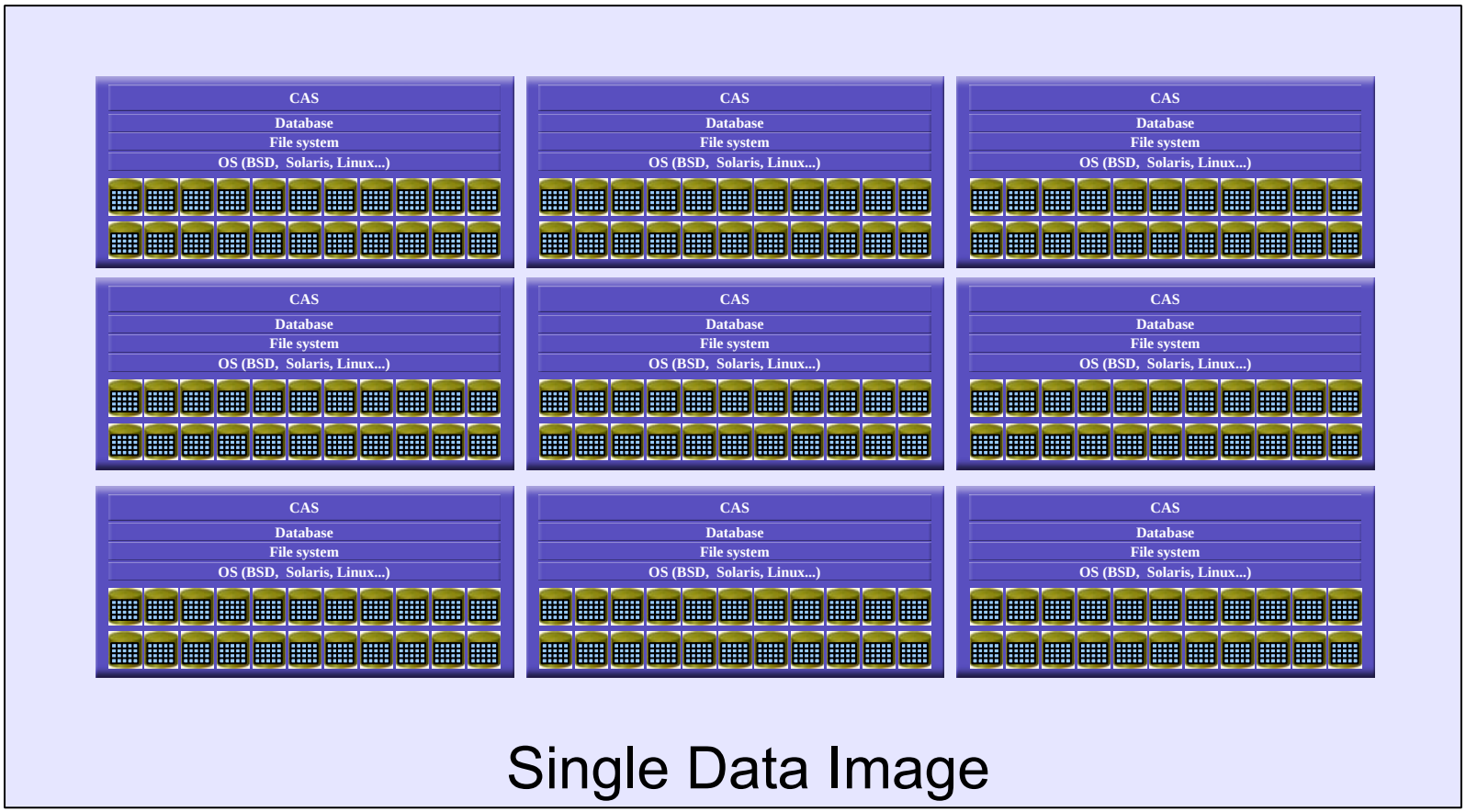
Content Addressable Storage (CAS)

- OIDs are hash values derived from the objects' content
- Used as digital archive systems for long-term fixed content data
- ECM applications used as data injection machines



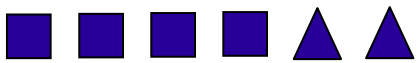
RAIN

Redundant Array of Inexpensive/Independent Nodes

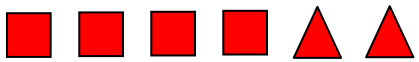


Data Placement

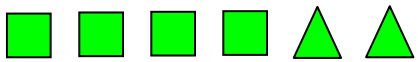
Object 1

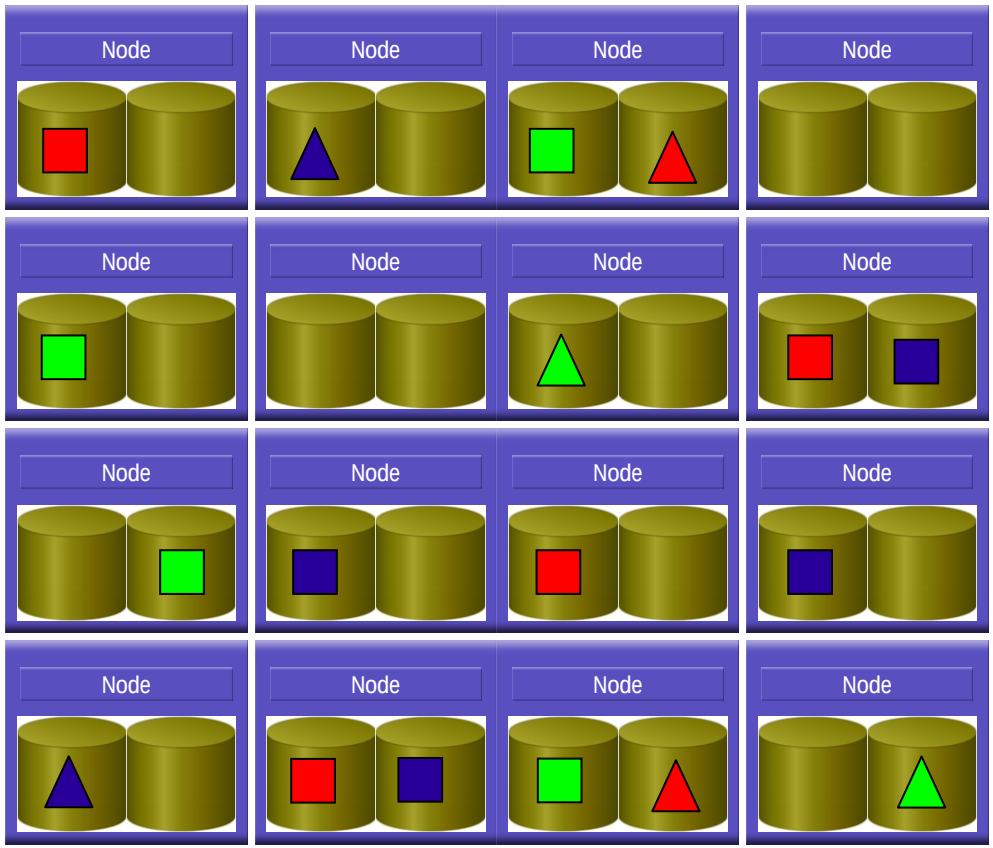


Object 2



Object 3





□ = Data

△ = Parity

Archiving vs Protection

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- Data Protection is about Data Recovery
  - e.g. RAID, snapshot, replication, backup...

- Data Archiving is about Data Discovery – Archiving requires data protection
  - e.g. index, search, aggregate

Archiving: allow near instantaneous retrieval of images, and do it at tape-like prices

The New Challenges

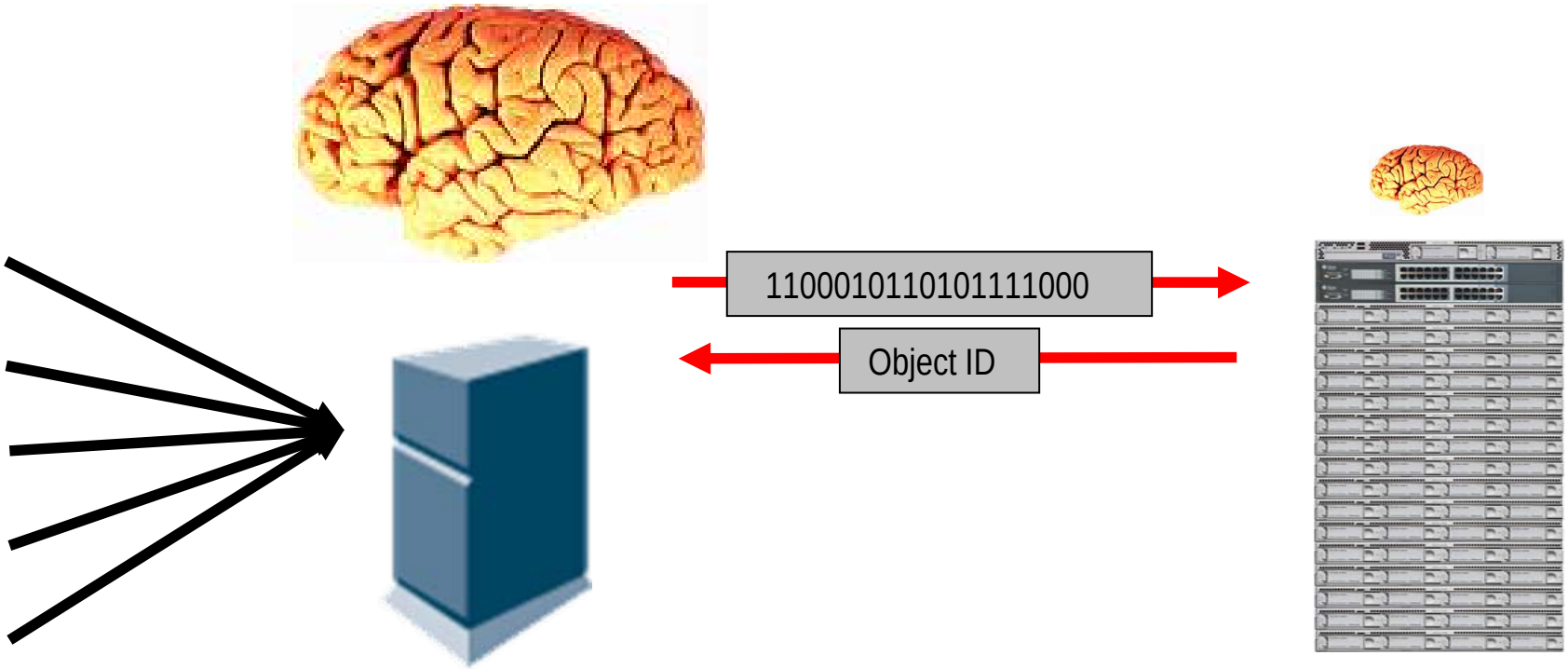
- Fast store and retrieval
- Availability
- Reliability
- Easiness to store, organize, retrieve and dispose New!
- Complex data operations New!
  - aggregate, join, view, sort, convert, encrypt... New!
- Enhanced search operations New!
- Flexibility to present data New!
- Customized storage behavior New!
- Reduced administration costs New!

Topics

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

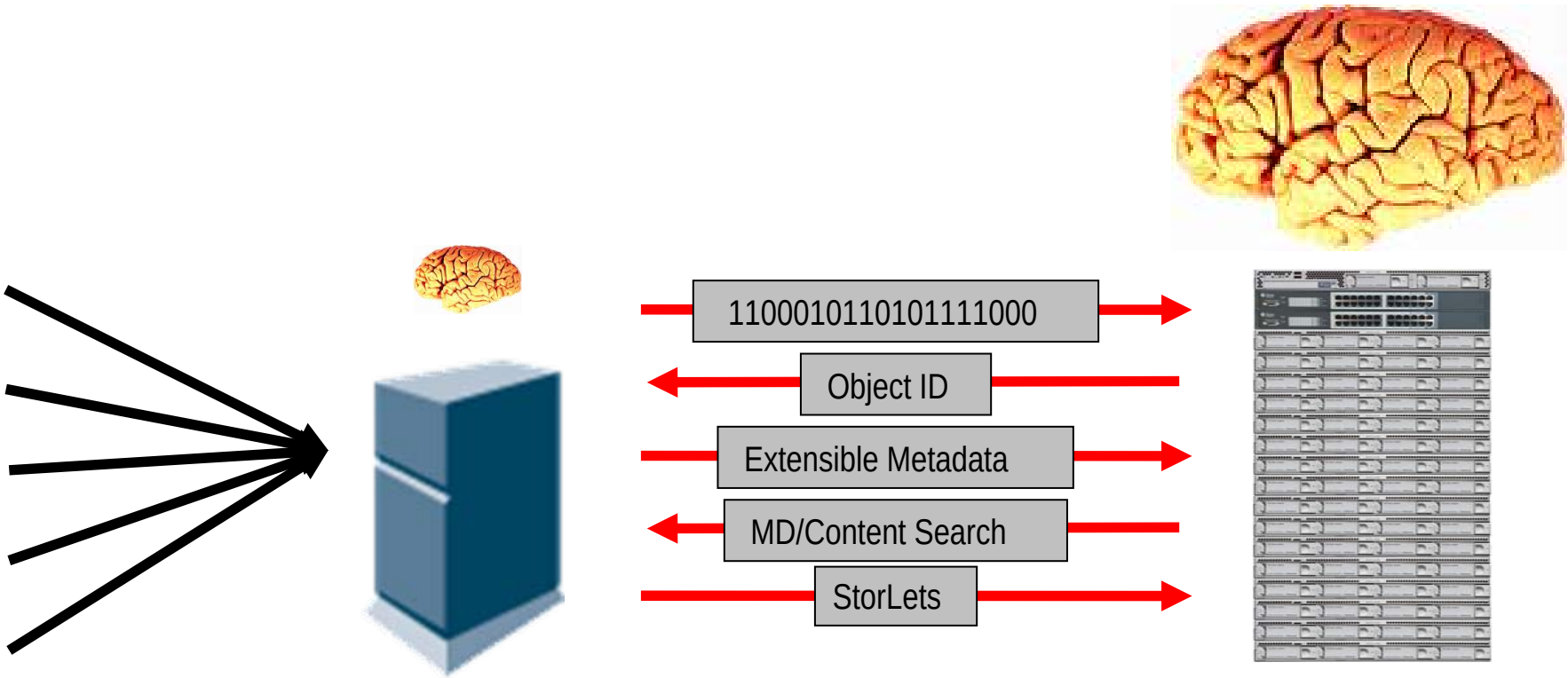
- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

CAS: “Content Addressable Storage”



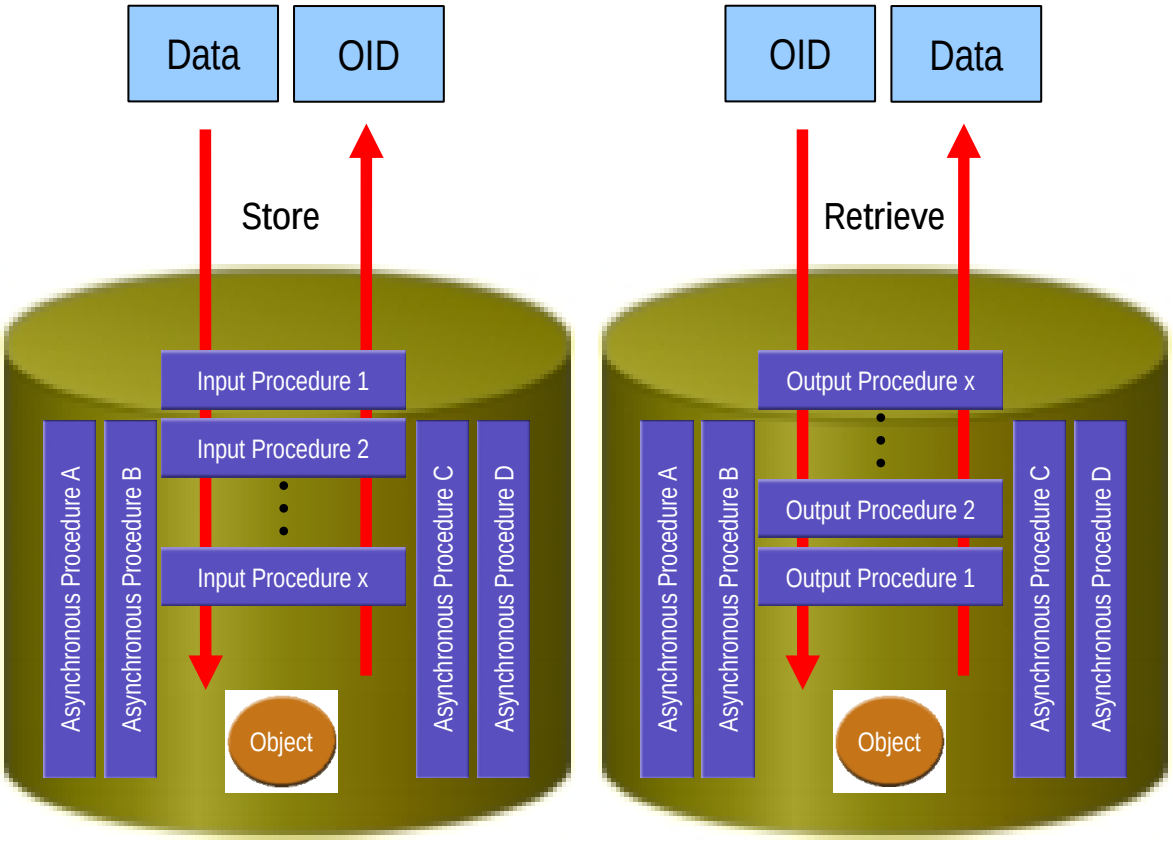
Enterprise Content Management
Injection Engine

CAS: “Content Aware Storage”



Enterprise Content Management
Injection Engine

Content Aware Storage
Flexibility



- user-developed trigger apps
- synchronous:
  - modify the behavior of store/retrieve/query/delete
  - e.g. transcode, downsample, filter, watermark, extract metadata from file, headers, encrypt, audit log...
- Asynchronous:
  - process data at rest
  - e.g. capacity optimization, scrubbing, migration, sanity check...

Topics

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- Block-Based Data Access
- File-Based Data Access
- Object-Based Data Access
  - Object-Based Storage Devices (OSD)
  - Object Storage Systems
    - Object Storage Server (OSS)
    - Content Addressable Storage (CAS)
    - Content Aware Storage (CAS)
- Intelligent Storage Nodes (ISN)

The Evolution of Data Processing

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

Storage Management

configuration, topology, RAID...

Data Management

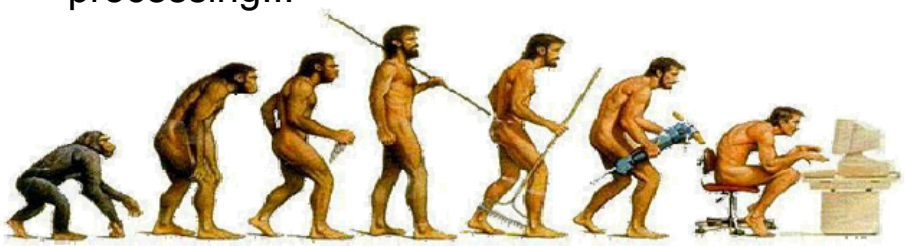
data services: volume mgmt mirror, replication snapshot, backup, DR...

Content Management

ILM, QoS, SLA, content/full text search, content aware processing...

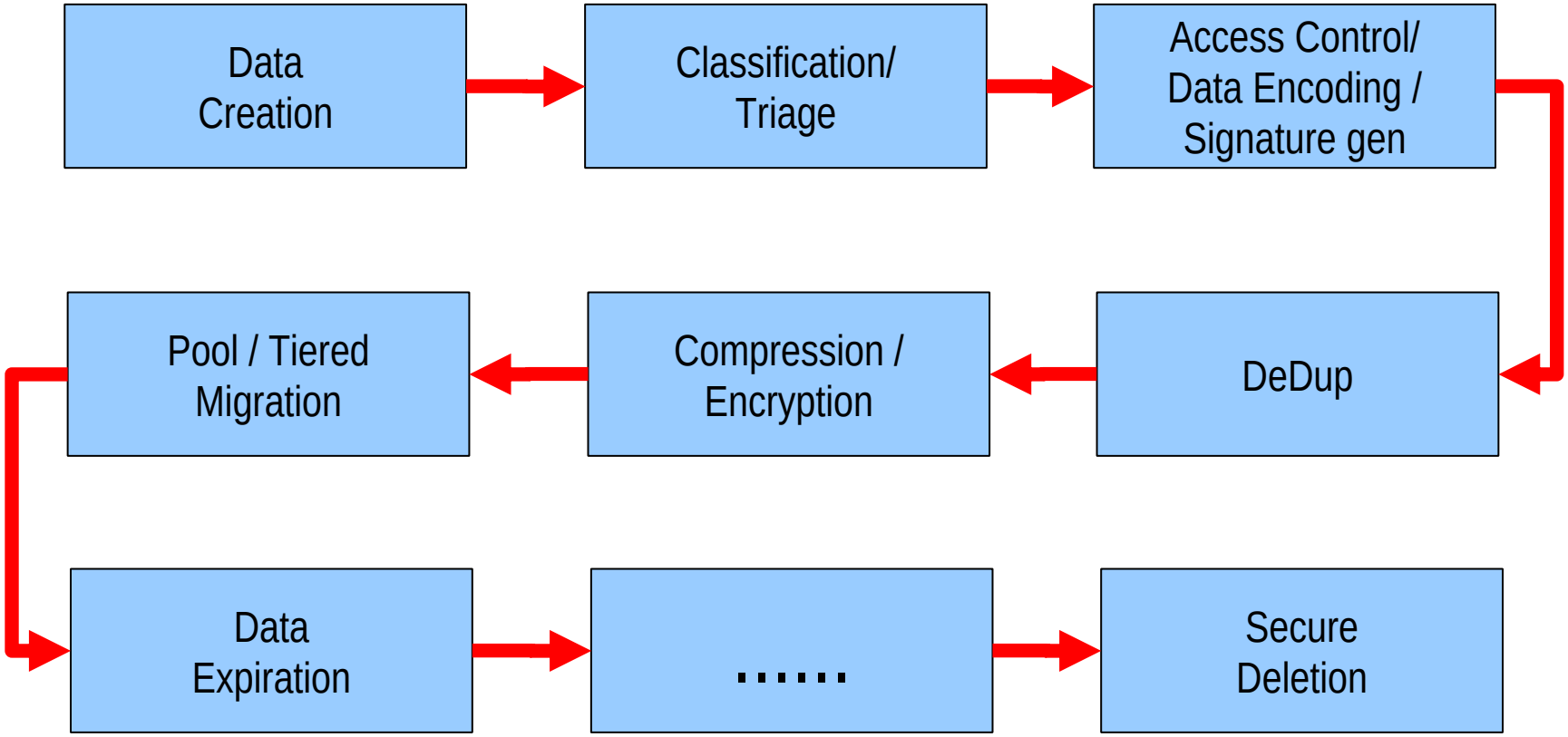
Knowledge/ Information Management

semantics, heuristic, learning, problem-solving...



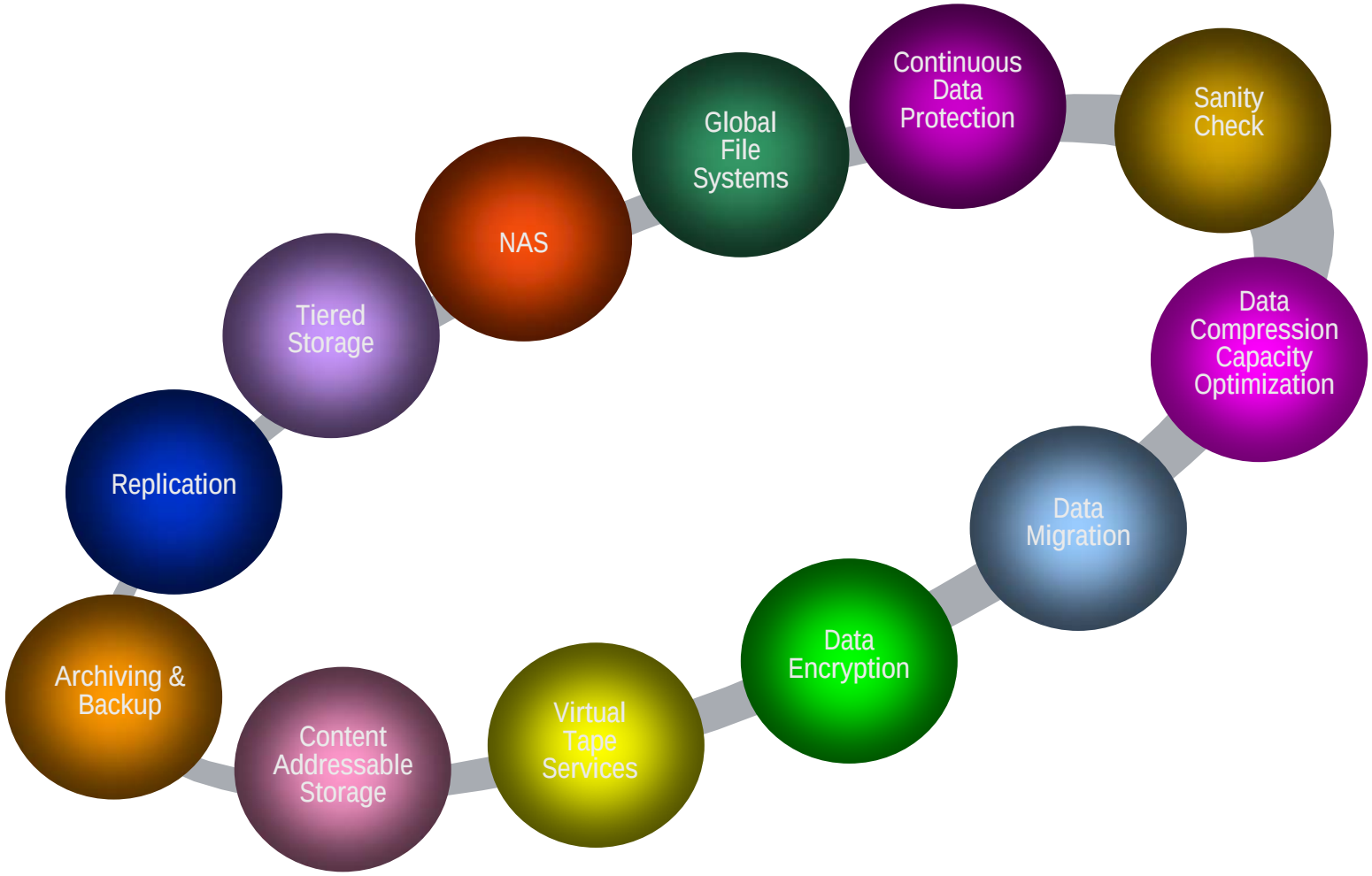
The Active Digital Archive

Archival Process Flow View



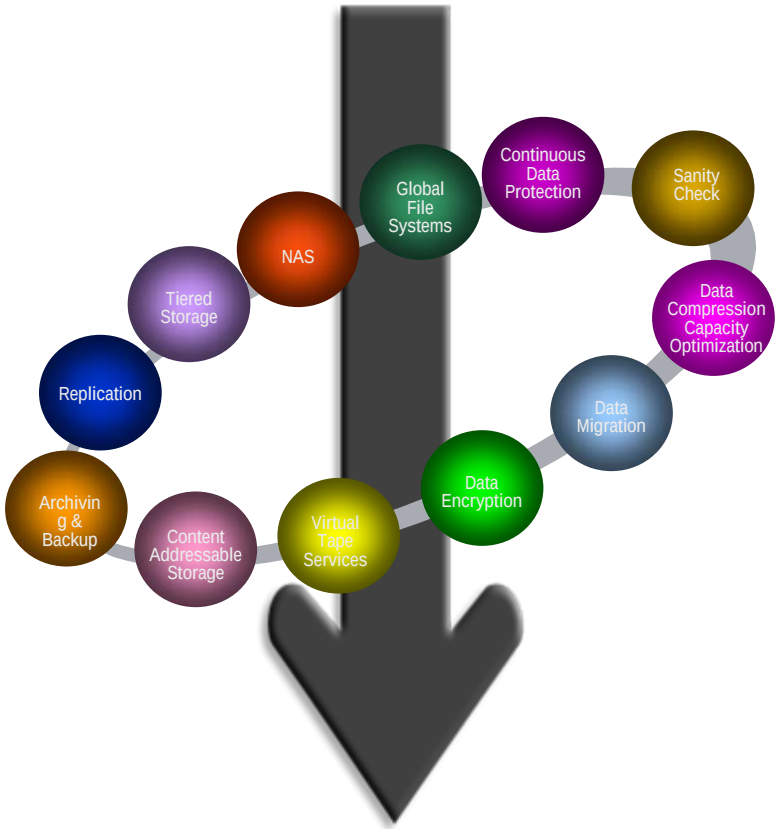
Note:
- Some steps can be done more than once
- Not all steps are needed
- Some steps can not be done out of order

Storage Applications



Migration of Storage Applications

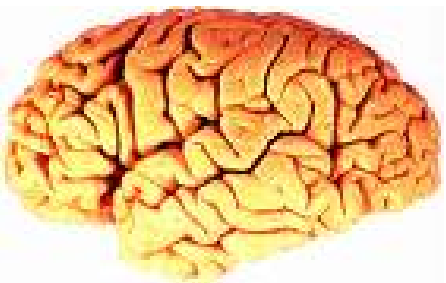
- Process the data where it lives...



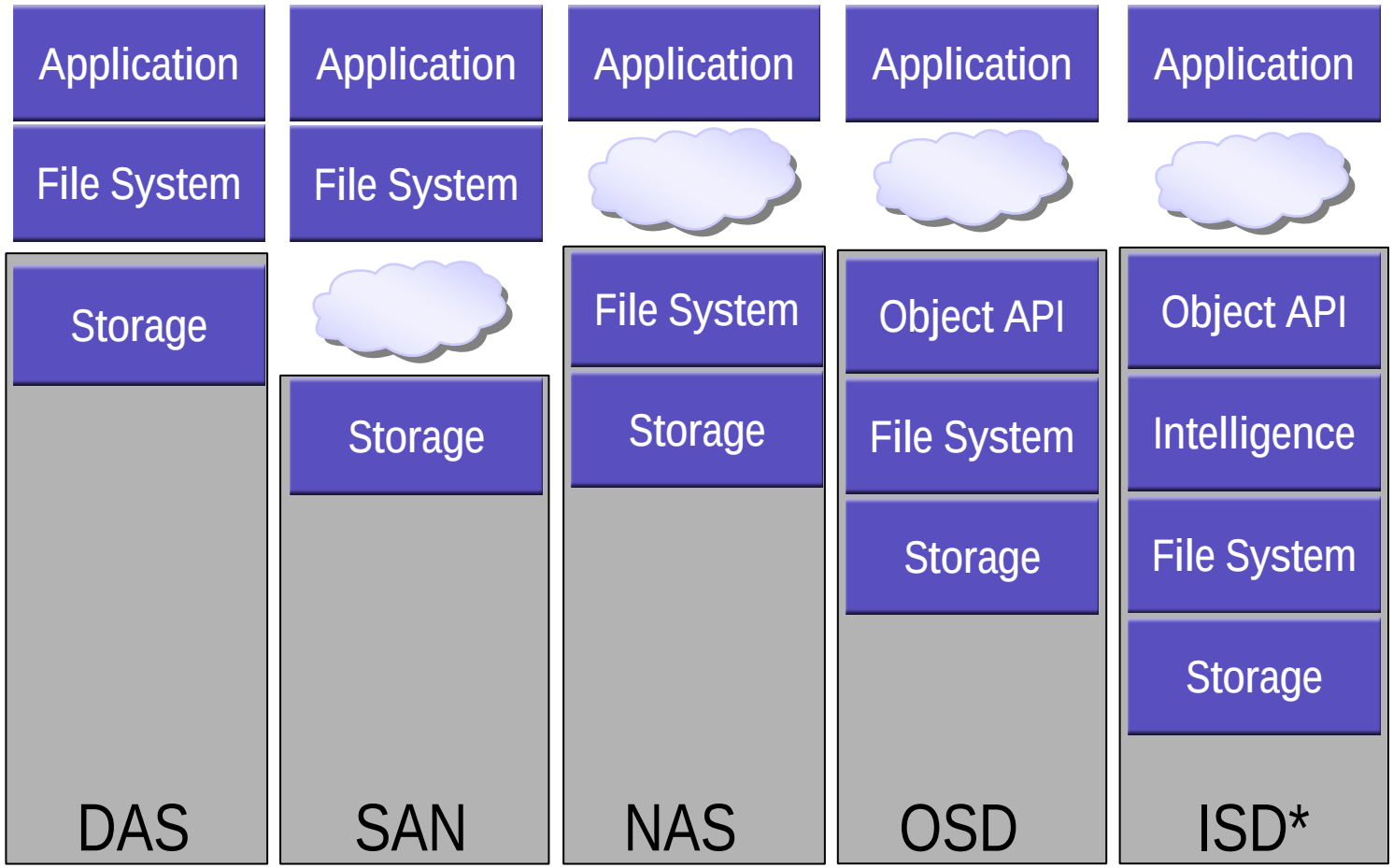
Server

Network

Storage



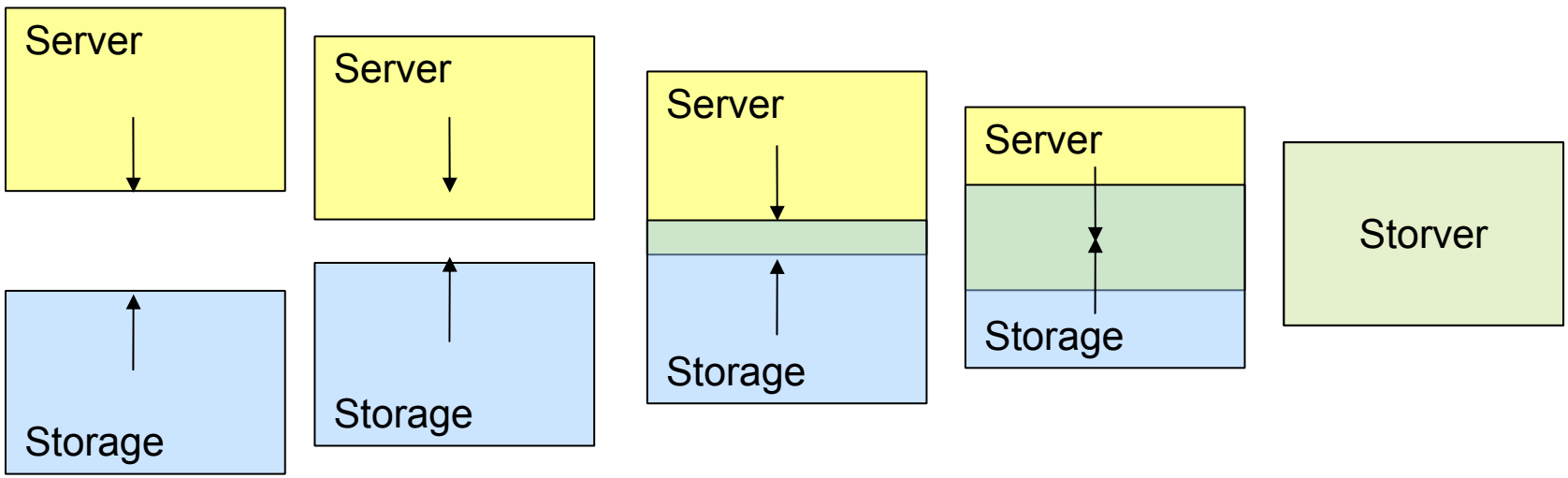
The Evolution of Storage



Note*: Intelligent Storage Node

The Vertical Consolidation

- Storage and server
- Migration of data processing applications
- No I/O is best I/O



The Intelligent Storage Node

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

Presentation Layer (Storage Personality)

Block I/F	File I/F	Object I/F
SCSI, FCP, iSCSI, SAS, SRP, SATA, VTL ...	FTP, NFS, CIFS, WebDAV, HTTP...	API, ANSI T10 OSD, CAS, XAM...

Storage Application Layer (Data Services)

Data Management	Compliance Services	Scheduled Services
Backup, Classification, Data Mining, Application Acceleration, Aggregation, Virtualization, Sort, Views, Search, Indexing, Data Movers, Distributed Namespace...	WORM, Audit, Retention, Secure Deletion...	De-Duplication, Hash Generation, Verification, Snapshot, Replication...

Operating System

Files systems, HSM, Device Drivers, Clustering, Security....

Persistent Storage

Disk, Tap, Library...

External Services

- LDAP
- NIS
- Active Directory
- NTP
- MDS Global Namespace
- others

Further Reference

- http://www.snia.org/tech_activities/workgroups
- http://www.snia.org/apps/org/workgroup/osd/
- http://www.snia.org/apps/org/workgroup/fcastwg/
- http://www.snia-dmf.org/
- http://www.t10.org/ftp/t10/drafts/osd
- http://www.t10.org/ftp/t10/drafts/osd2
- http://ietf.org/html.charters/webdav-charter.html
- http://ietf.org/html.charters/nfsv4-charter.html
- http://www.snia.org/education/tutorials/

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: trackstorage@snia.org

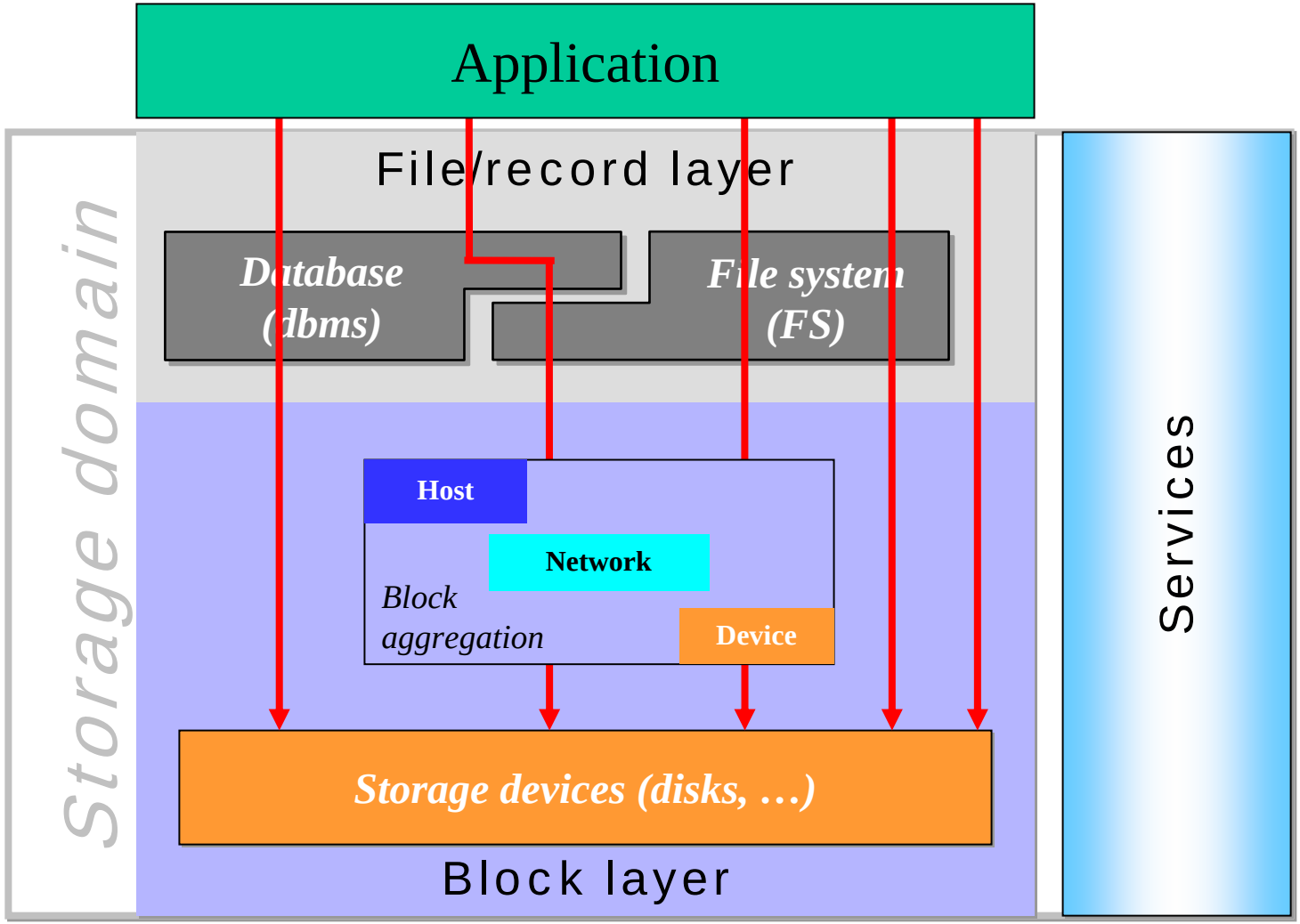
Many thanks to the following individuals for their contributions to this tutorial.

SNIA Education Committee

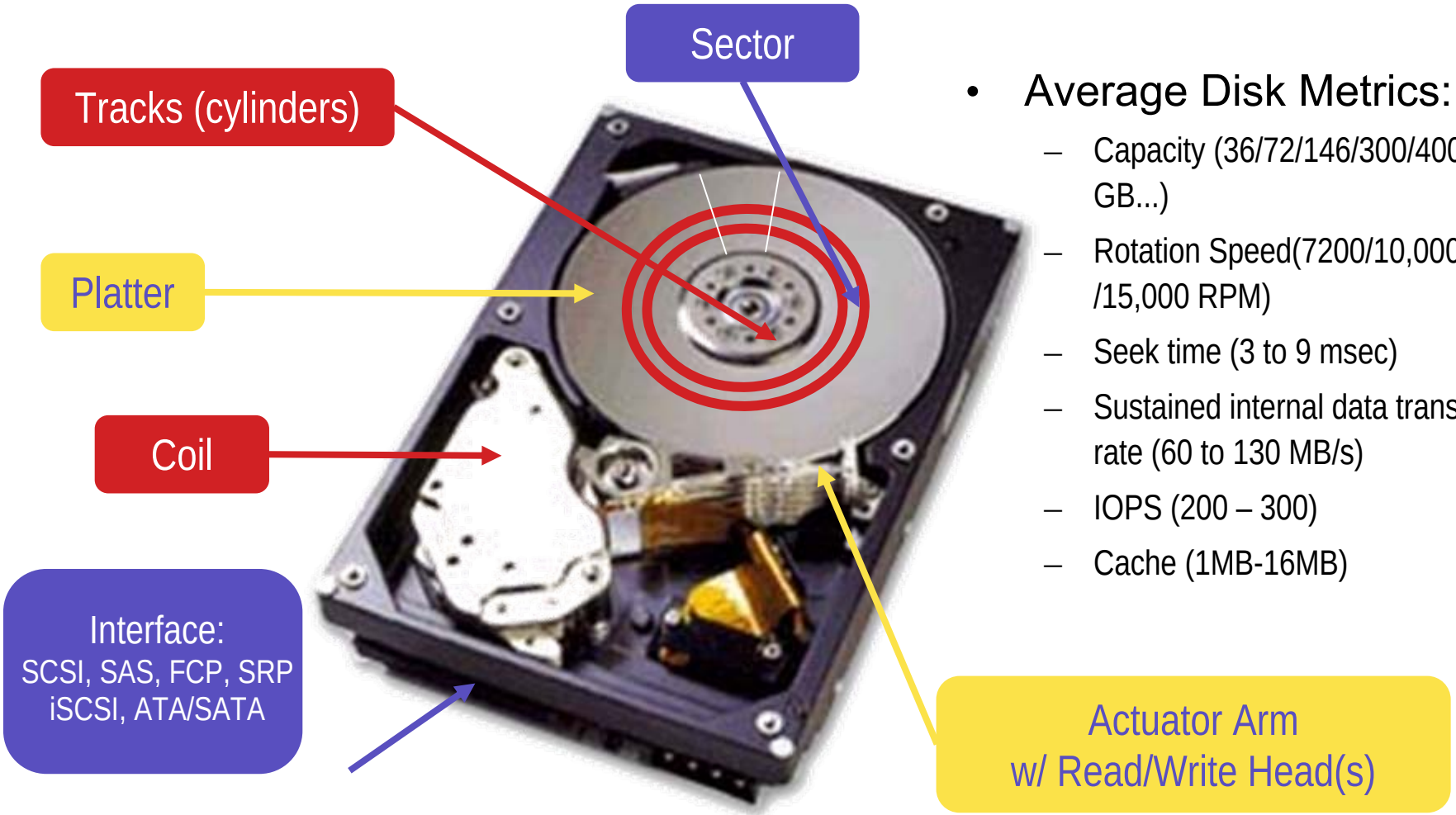
Christian Bandulet, Sun Microsystems

Appendix

The SNIA Shared Storage Model



Disk Drive Anatomy



- Average Disk Metrics:
  - Capacity (36/72/146/300/400/500 GB...)
  - Rotation Speed(7200/10,000 /15,000 RPM)
  - Seek time (3 to 9 msec)
  - Sustained internal data transfer rate (60 to 130 MB/s)
  - IOPS (200 – 300)
  - Cache (1MB-16MB)

Technology Improvements

~1956 first spinning hard drive (IBM RAMAC)

1956: 5 MB – 2000 bits/in^2

2006: 500 GB ~ 200 Gb/in^2

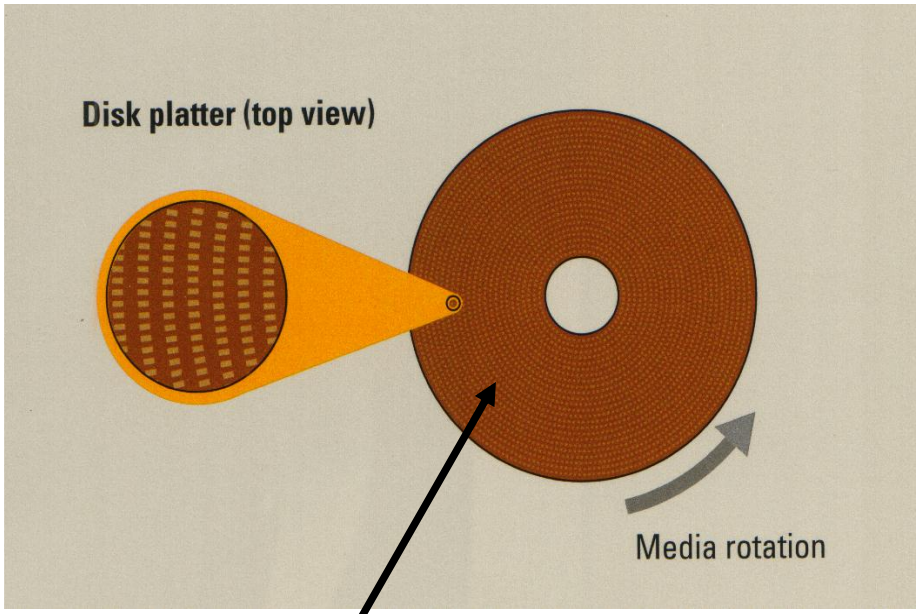
100.000.000 x areal density

100.000 x capacity

Improvement in 50 years !

Magnetic Disk Recording
Longitudinal Recording

- Technology is ~50 years old!
- First introduced with IBM RAMAC 5MB in 1956
- areal density increases 100% / year since early 1990s
- Disk areal density progress slowed down in 2003 as recording challenges appeared

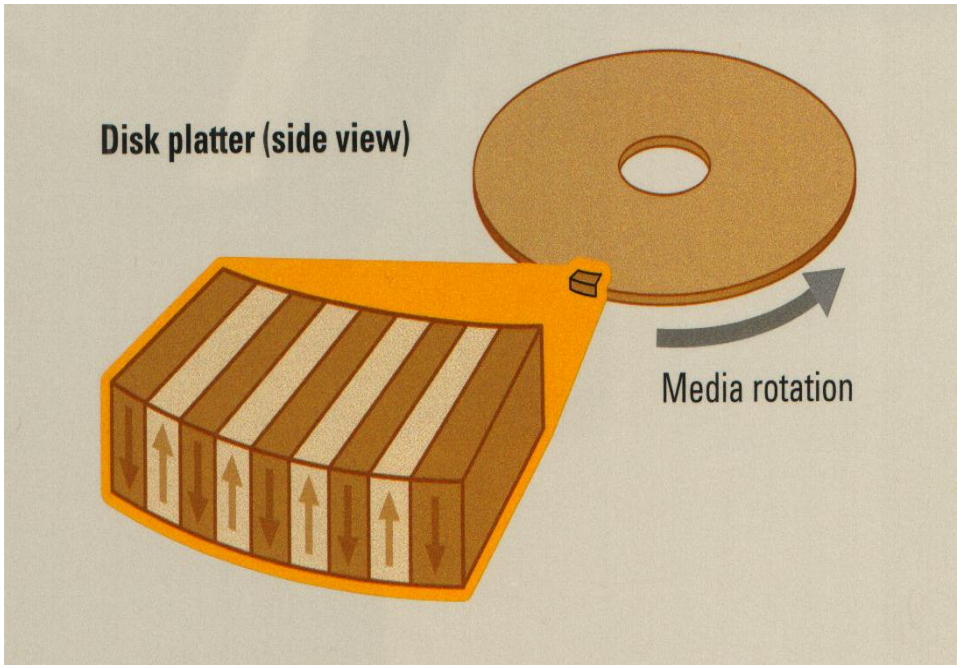


Source: www.horison.com

Blocks, Sectors, Tracks

Magnetic Disk Recording
Perpendicular/Vertical Recording

- Expected to delay Superparamagnetic Effect, not eliminate it...

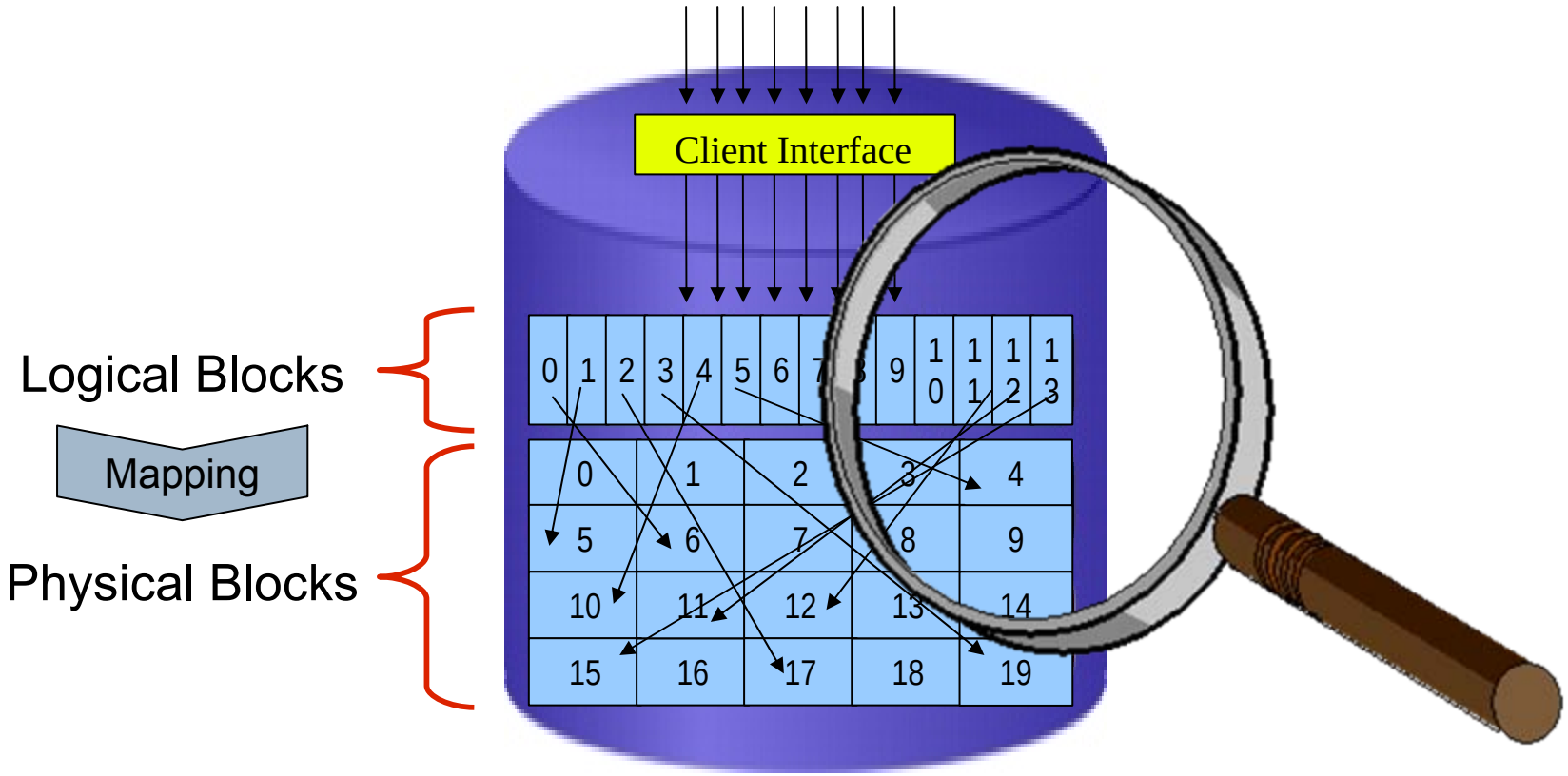


Source: www.horison.com

Logical Blocks & Physical Blocks

Let's have a closer look....

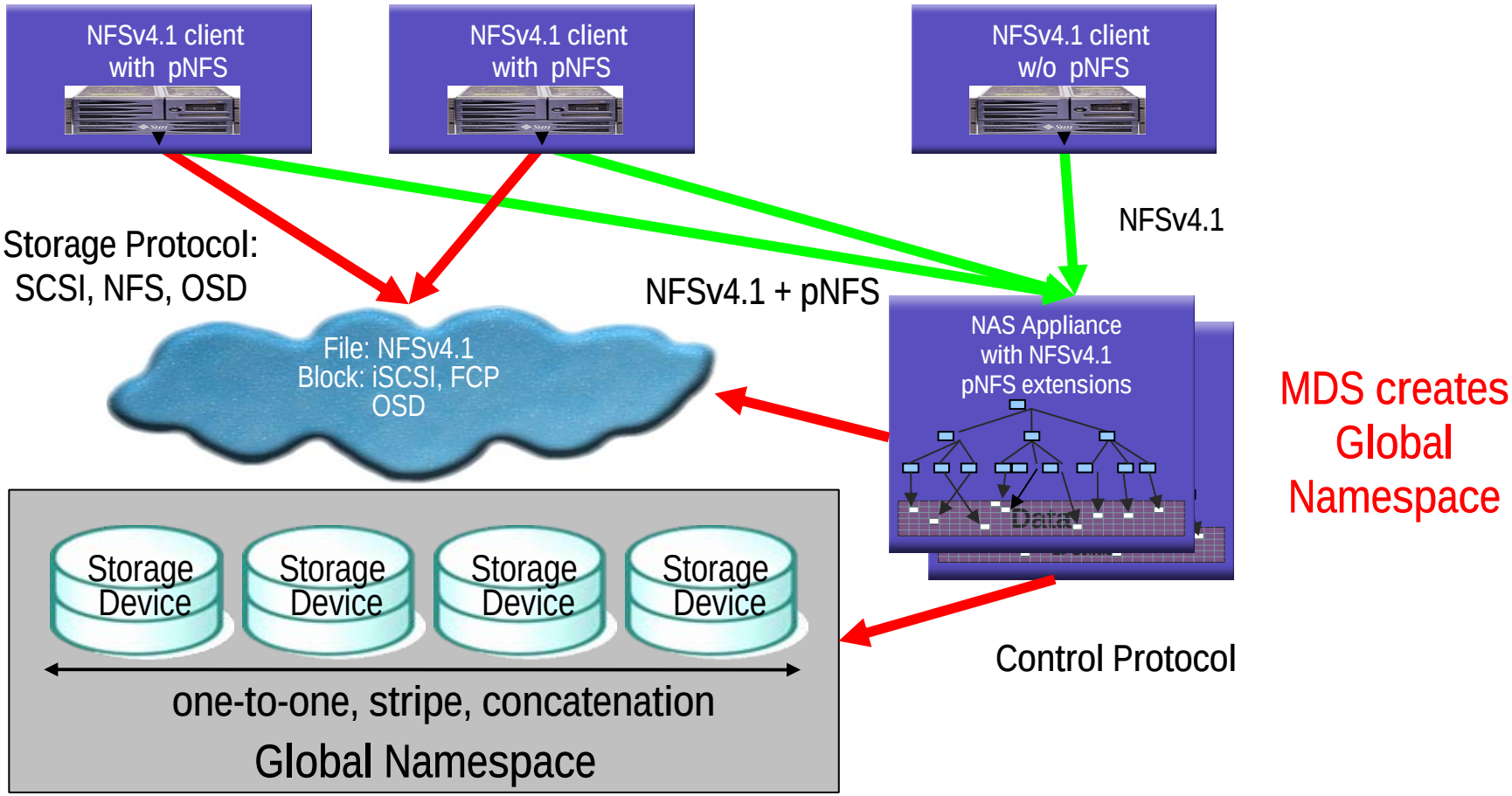
SCSI, SAS, FCP, SRP, iSCSI, ATA, SATA



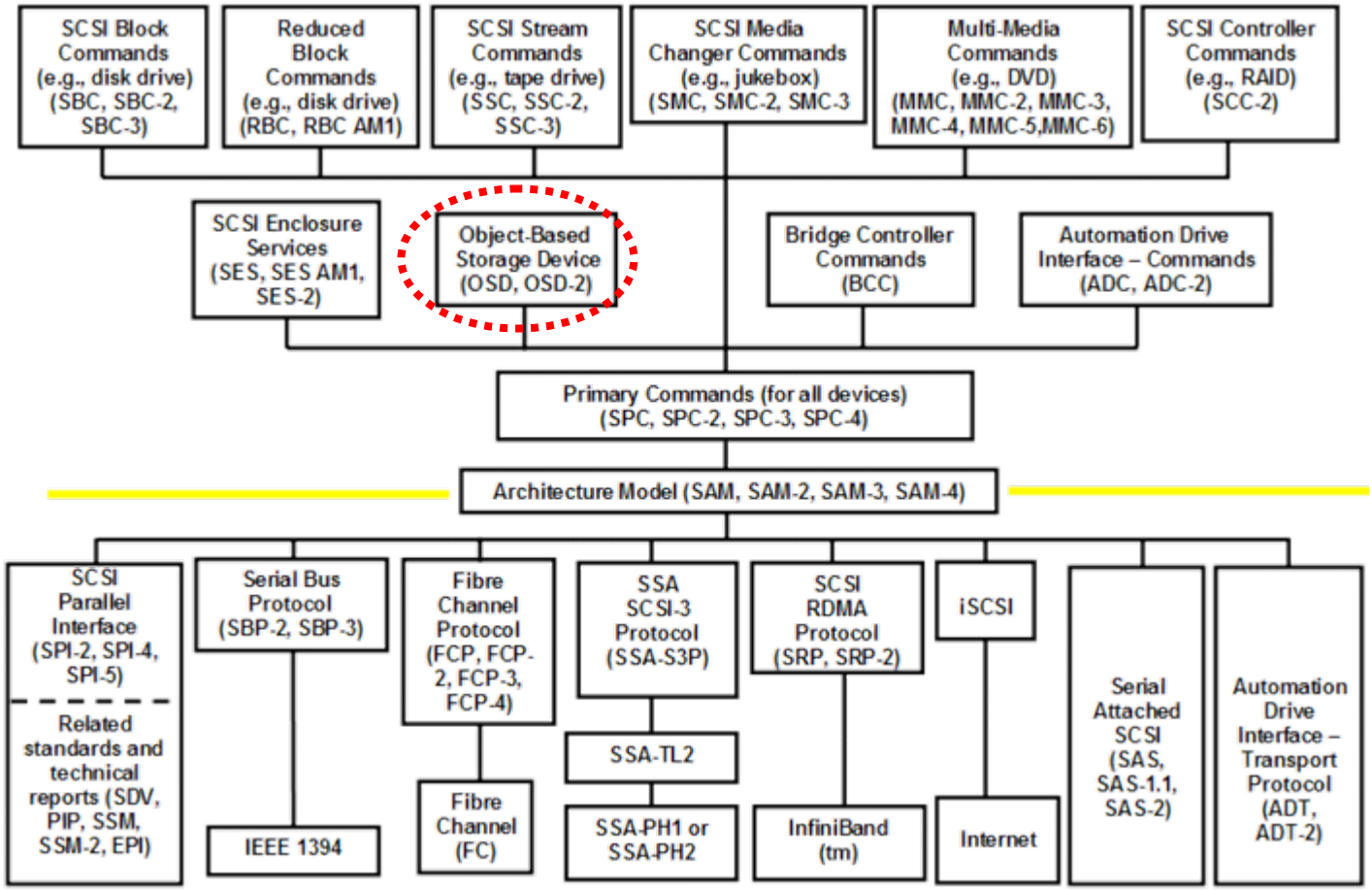
Scalable NAS (cont'd)

aka Loosely Coupled NAS

Global Namespace with NFSv4.1 and pNFS



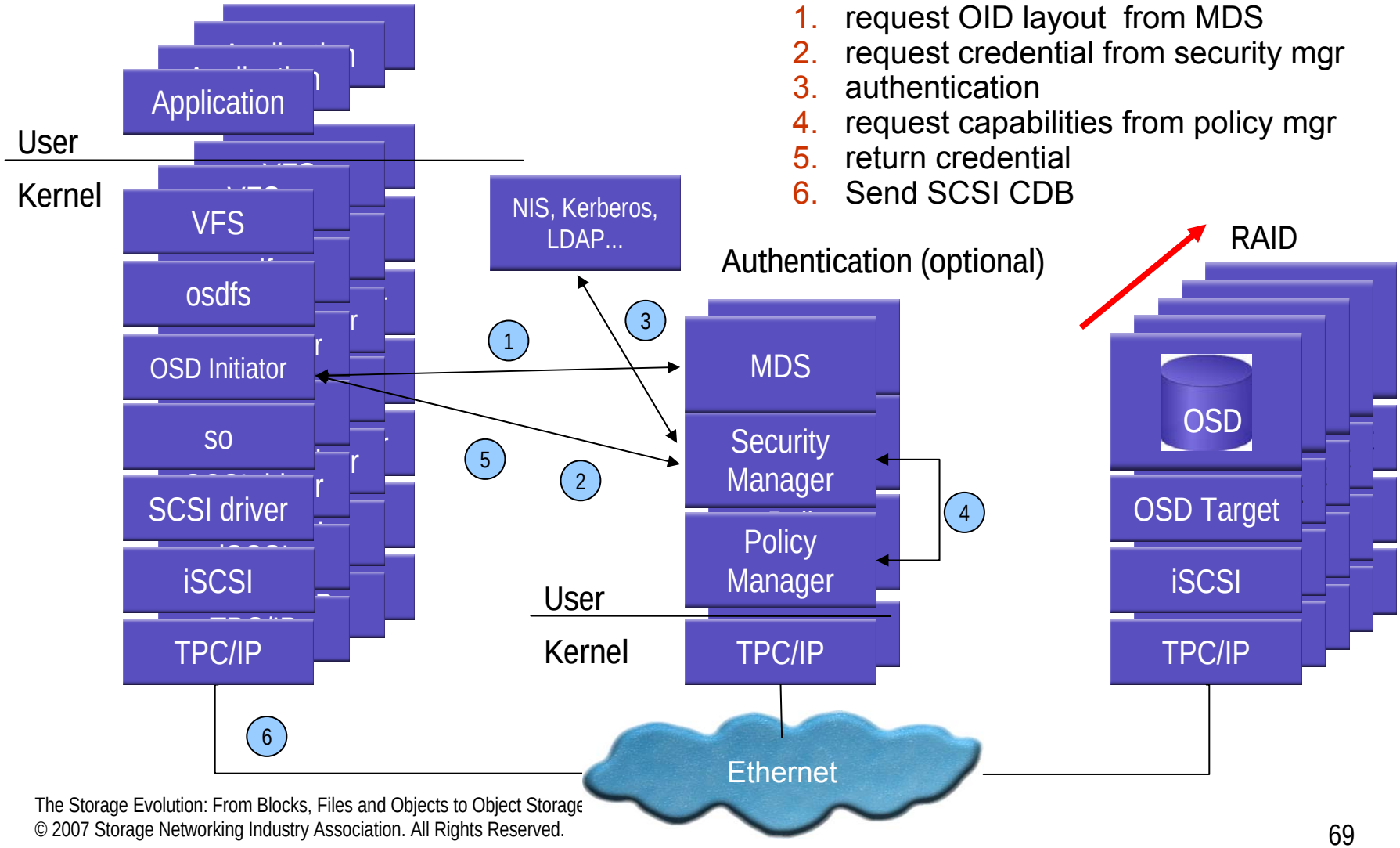
SCSI Standards Architecture



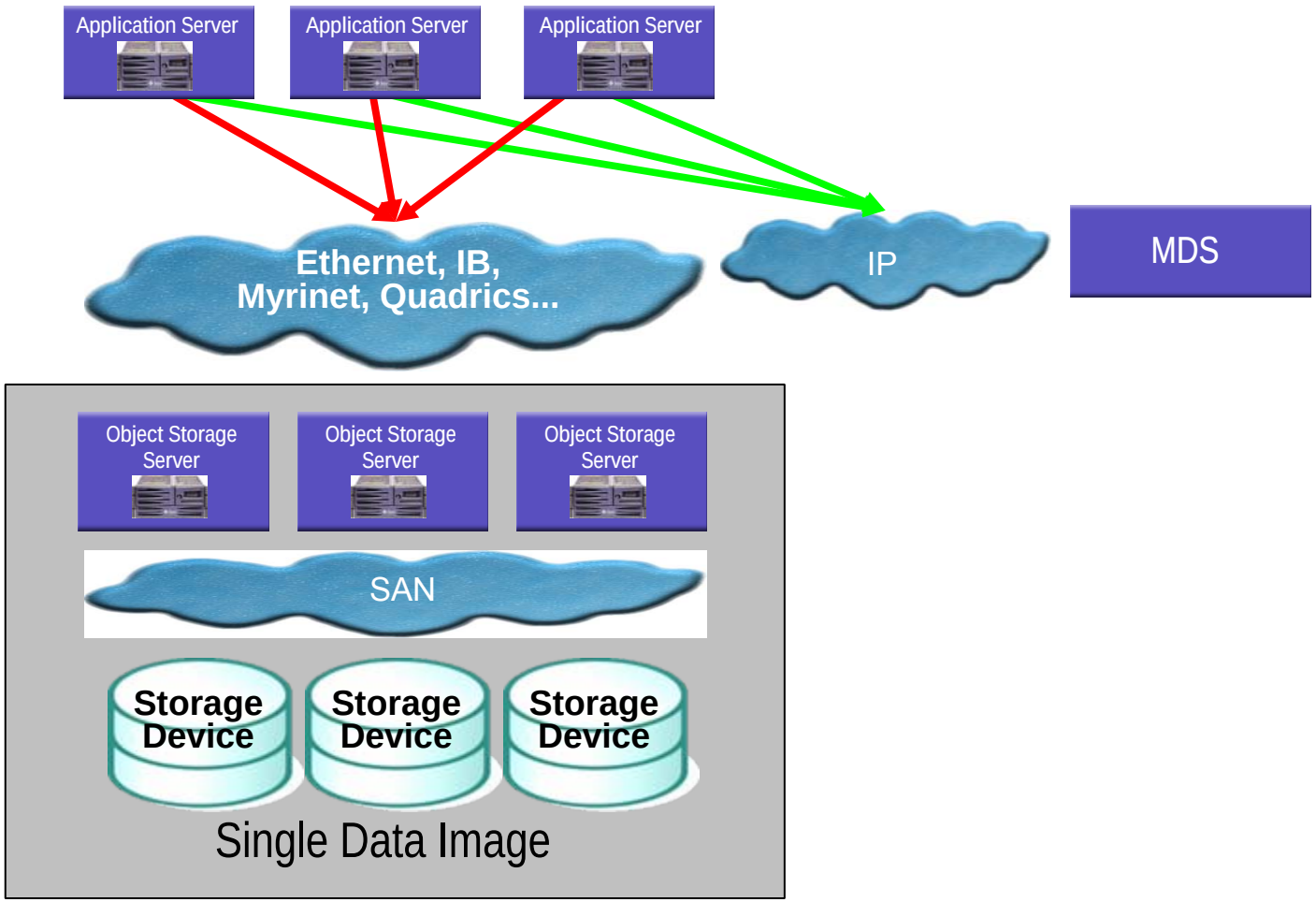
Source: www.t10.org

Files Sharing with OSDs

1. request OID layout from MDS
2. request credential from security mgr
3. authentication
4. request capabilities from policy mgr
5. return credential
6. Send SCSI CDB



Global, Distributed & Parallel FS
With Object Storage Server (OSS)



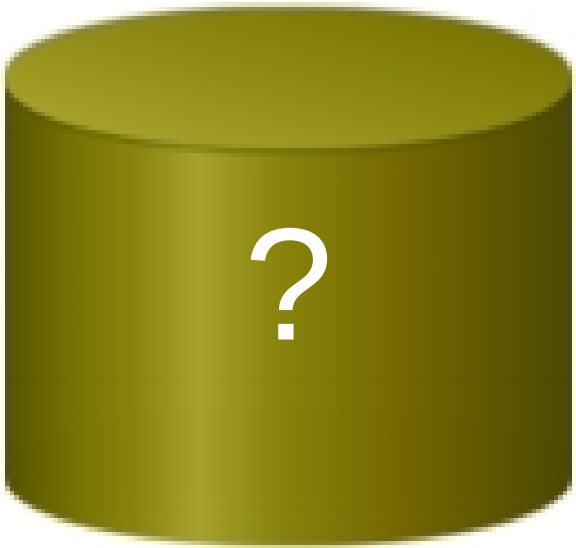
Content Addressable/Aware Storage
aka CAS

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- OIDs are hash values derived from the objects' content
- Objective:
  - Store large amounts of data reliably for long periods of time with fast access time to retrieve data
- Target applications:
  - Homeland security, Satellite imagery, Digital asset management, Medical imaging, Digital photo services, Seismic data archival, Regulatory compliance, Media preservation
- Not used for:
  - Online Transaction Processing (OLTP), Enterprise Resource Planning (ERP), Live database, Small scale file sharing

The Digital Archive Problem

- How do you store and organize 100 million things?



- Issues of:
  - Scale Performance
    - capacity/workload balancing
    - automatic capacity expansion
  - Organize data
    - manage metadata
  - Search
  - Reliability/Availability
    - data rebuild and/or failover
  - Cost (OPEX/CAPEX/TCO)
  - Technology refresh

Content Aware Storage

Attribute Awareness

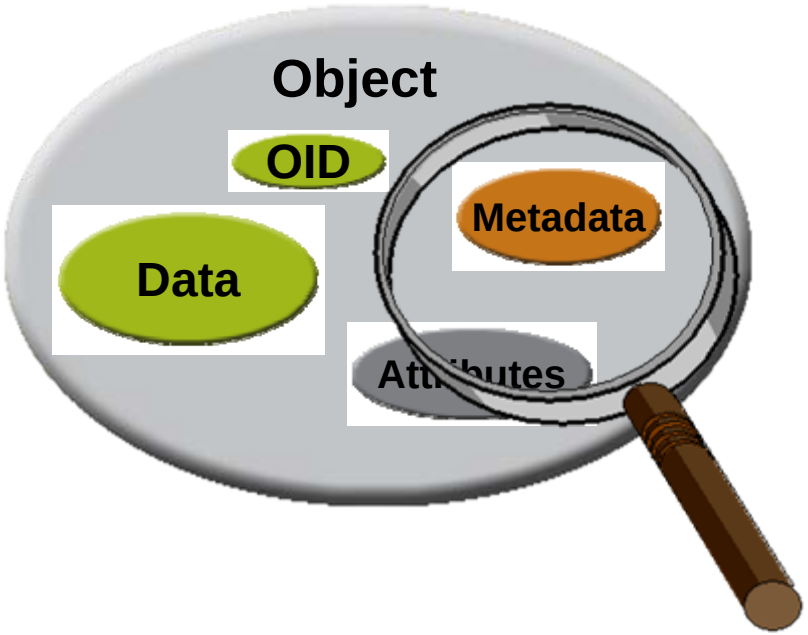
- Object attributes are stored directly with data object by the application
- Attributes are carried automatically between layers and across devices
- When objects pass through a certain system layer or device, that layer can act on the values in the attributes that it understands
- All other attributes are passed along unmodified and not acted upon
- e.g. Objects marked as high-reliability can be treated differently than objects marked as temporary
- Attributes should be dynamically changeable

Layer 3	Attributes Layer 0	Attributes Layer 1	Attributes Layer 2	Attributes Layer 3
Layer 2	Attributes Layer 0	Attributes Layer 1	Attributes Layer 2	Attributes Layer 3
Layer 1	Attributes Layer 0	Attributes Layer 1	Attributes Layer 2	Attributes Layer 3
Layer 0	Attributes Layer 0	Attributes Layer 1	Attributes Layer 2	Attributes Layer 3

Content Aware Storage

Object Discovery

- Searchable metadata
- Name-value based
  - OID
  - Metadata
  - user derived attributes
- Content
  - full text search



Growing Storage Computation

SNIA
STORAGE NETWORKING INDUSTRY ASSOCIATION
EDUCATION

- Database acceleration via offloading
  - health check, multi-level security, db reorganization, image copies, HSM, data mining...
- Business Continuity, Backup, Recovery, D2D2T, CDP...
- Data Reduction
  - Classification, essential vs non essential, single instance, compression...
- Security
  - Authentication, authorization, encryption
- Data Transformation
- Multiple Data Views
  - workflow
- Real-time Data Analytics
  - indexing, search, sort, aggregation
- Business Management
  - Data Life Cycle, migration, compliance