SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

EDUCATION

Object-based Storage (OSD) Architecture and Systems

Erik Riedel, Seagate Technology

April 2007

Abstract

Object-based Storage (OSD) Architecture and Systems

The Object-based Storage Device interface standard was created to integrate chosen low-level storage, space management, and security functions into storage devices (disks, subsystems, appliances) to enable the creation of scalable, self-managed, protected and heterogeneous shared storage for storage networks.

This tutorial will describe how OSD storage devices will be integrated into today's most popular storage systems and how systems using OSD-enabled devices are being designed and built.

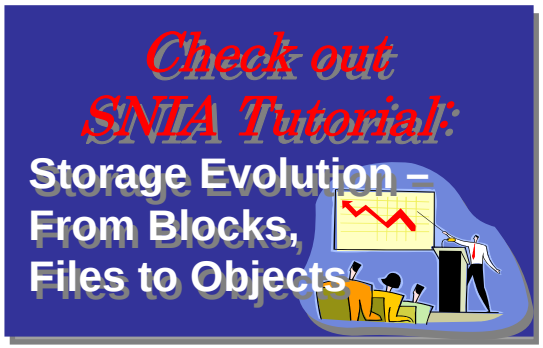
Outline

- Overview
  - motivation, structure, systems
  - history (see appendix)
- Interface
  - commands
  - objects, attributes, security (see 2005 tutorial)
- Architecture
  - scalable NAS enabled by OSD
  - CAS with OSD

  } physical view

  - objects = data + metadata
  - ILM with OSD
  - security

  } logical view
- Status & Next Steps
  - extension work continues today

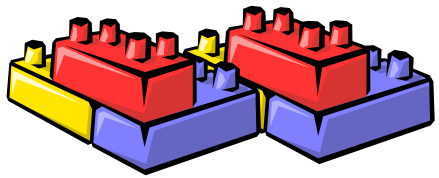


Overview

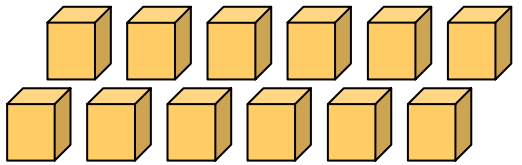
Motivation for OSD

- Improved device and data sharing
  - Platform-dependent metadata at the device
  - Systems need only agree on naming
- Improved scalability & security
  - Devices directly handle client requests
  - Object security w/ application-level granularity
- Improved performance
  - Applications can provide hints, QoS, policy
  - Data types can be differentiated at the device
- Improved storage management
  - Self-managed, policy-driven storage
  - Storage devices become more autonomous

Volumes

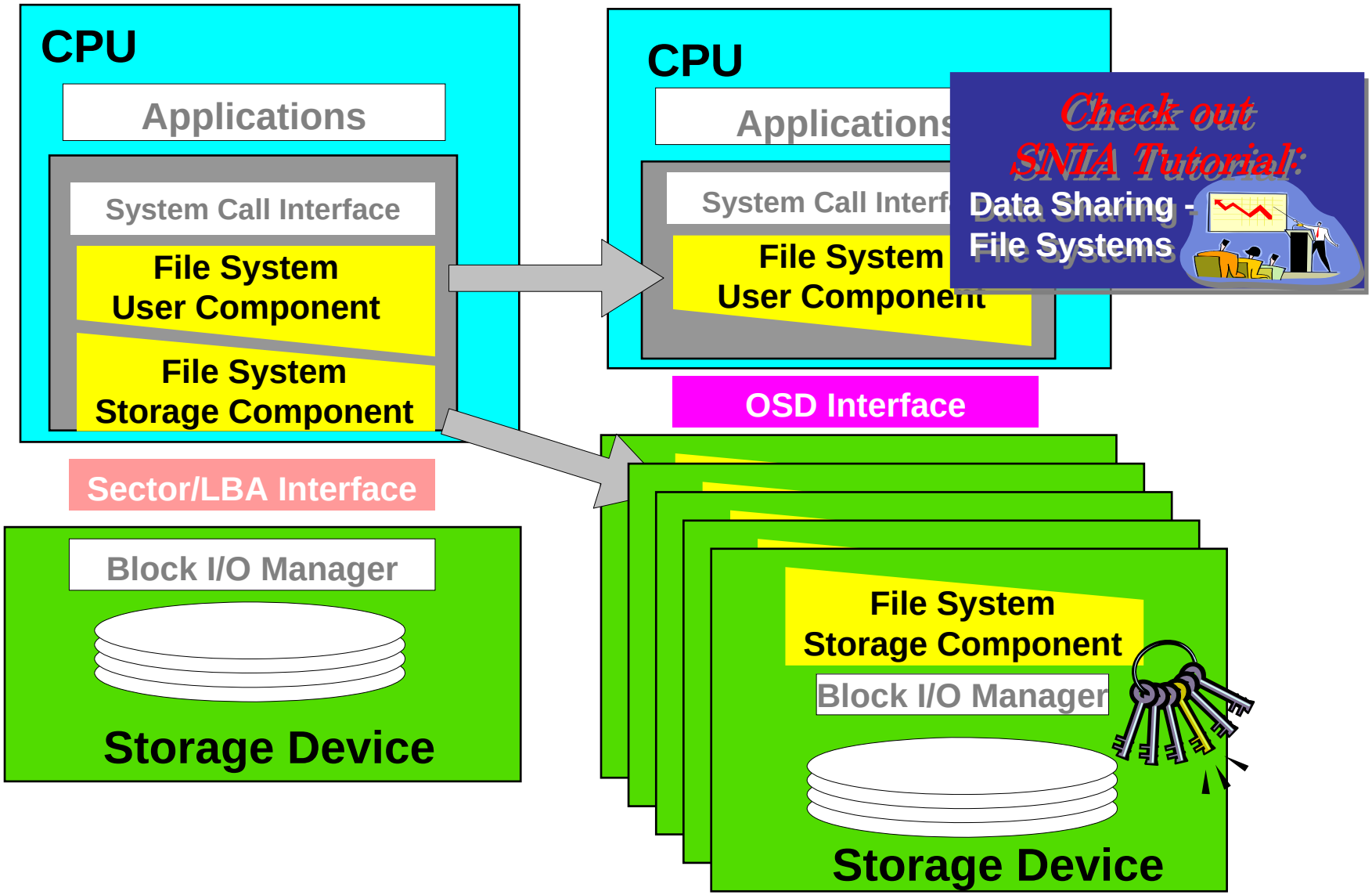


Objects



Blocks

OSD Structure



OSD Systems – 2007

A variety of Object-based Storage Devices being built today

- Disk array/server subsystem
- E.g. LLNL units with Lustre

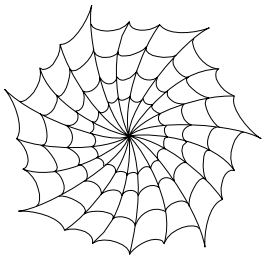
- "Smart" disk for objects
- E.g. Panasas storage blade

- Highly integrated, single disk
- E.g. prototype Seagate OSD

- **File/ Security Manager**

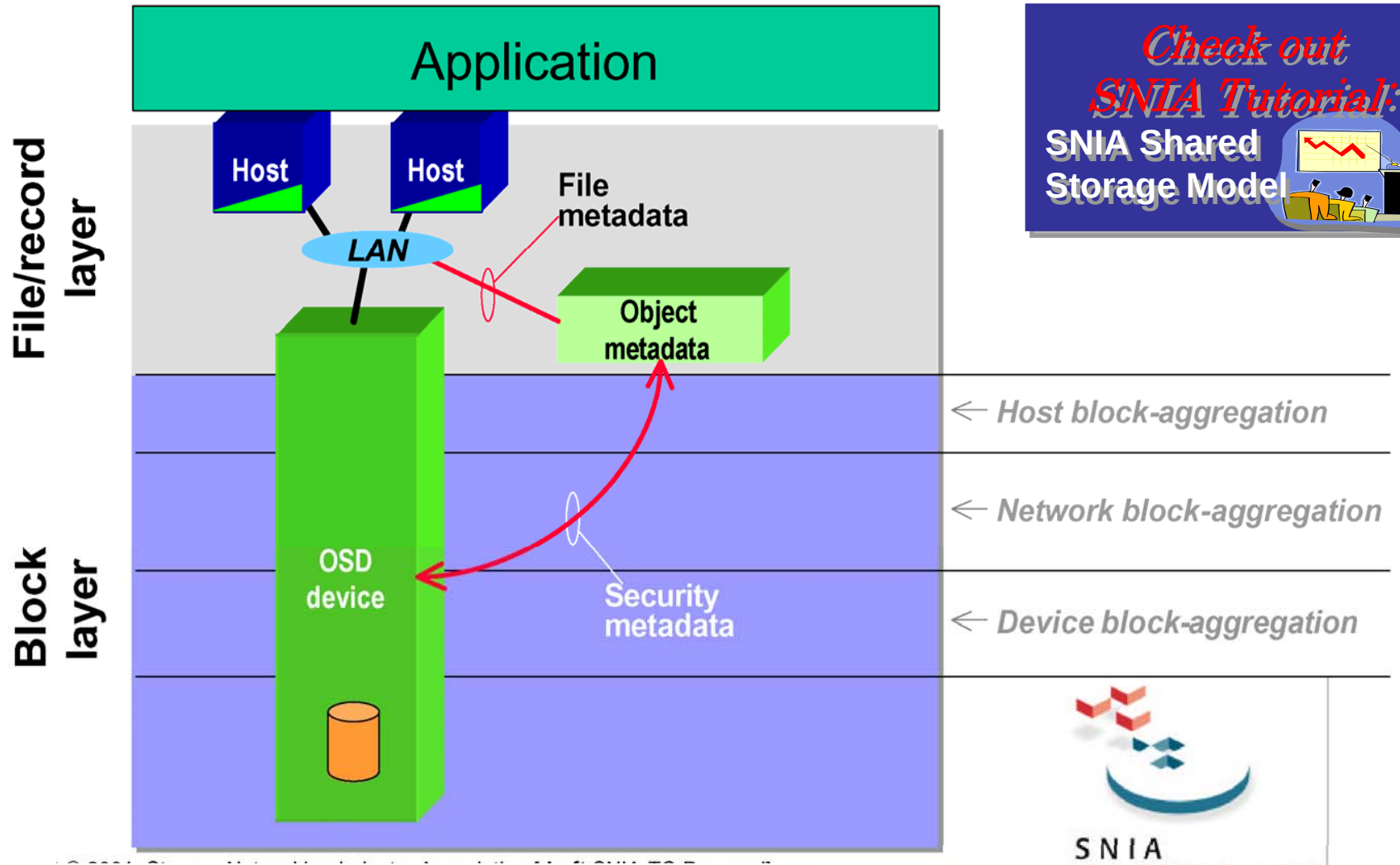
  - Orchestrates system activity
  - Balances objects across OSDs
  - Called clustered MDS in Lustre
  - Called Mgmt Blade by Panasas
  - Called ST server cluster by IBM

- **Scalable Network**



  - Connectivity among clients, managers, and devices
  - Shelf-based GigE (Panasas)
  - Specialized cluster-wide high-performance network (Lustre)
  - Storage network (IBM)

Object-based Storage Device (OSD), CMU NASD



SNIA
Storage Networking Industry Association

Interface

OSD Commands

OSD-1 r10, as ratified

- Basic Protocol
  - READ
  - WRITE
  - CREATE
  - REMOVE
  - GET ATTR
  - SET ATTR

  very basic

  space mgmt

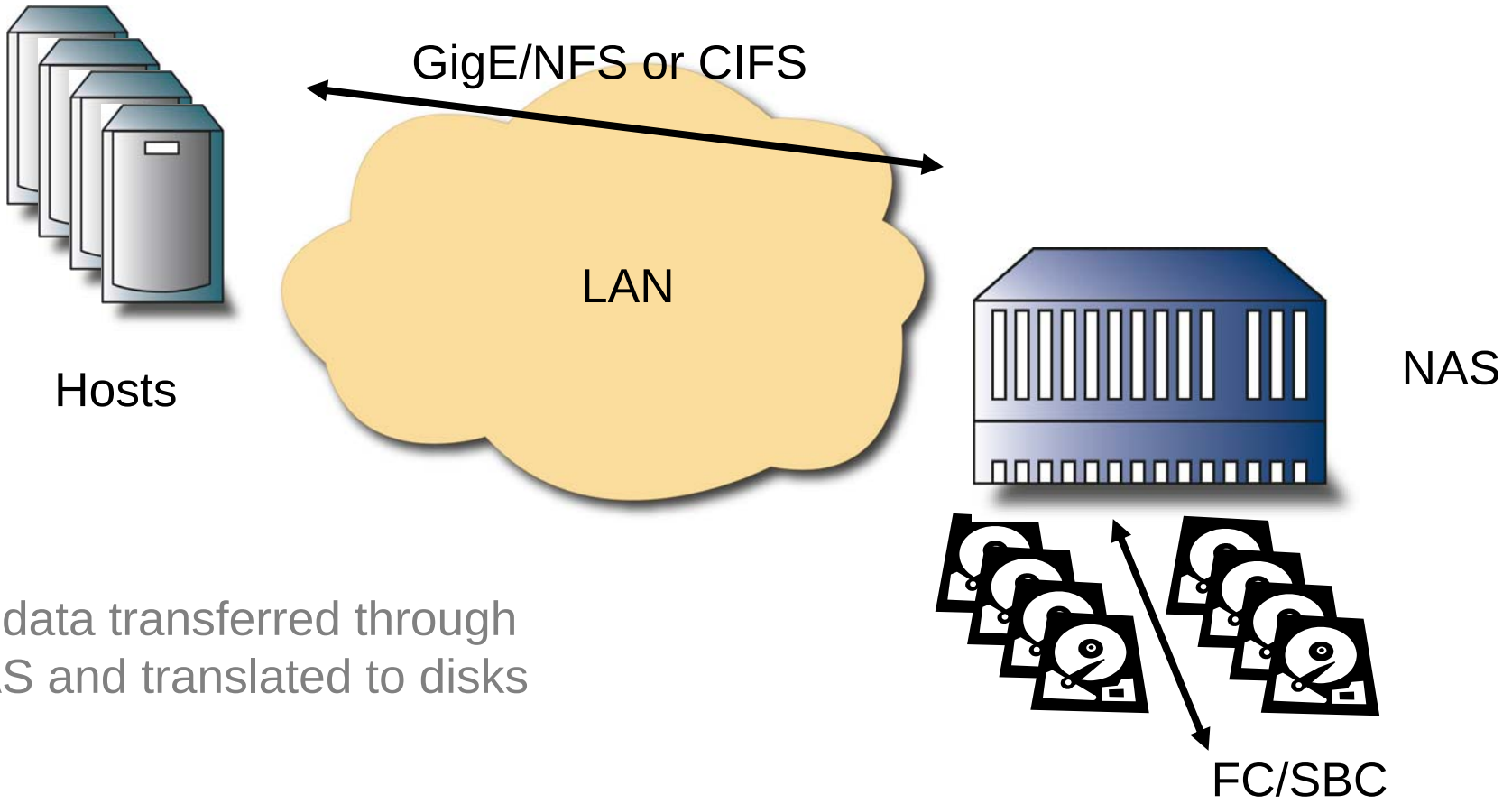
  attributes
  - timestamps
  - vendor-specific
    - opaque
    - shared
- Specialized
  - FORMAT OSD
  - APPEND – write w/o offset
  - CREATE & WRITE – save msg
  - FLUSH – force to media
  - FLUSH OSD – device-wide
  - LIST – recovery of objects
- Security
  - Authorization – each request
  - Integrity – for args & data
  - SET KEY
  - SET MASTER KEY

  shared secrets
- Groups
  - CREATE COLLECTION
  - REMOVE COLLECTION
  - LIST COLLECTION
  - FLUSH COLLECTION
- Management
  - CREATE PARTITION
  - REMOVE PARTITION
  - FLUSH PARTITION
  - PERFORM SCSI COMMAND
  - PERFORM TASK MGMT

Architecture – Physical View

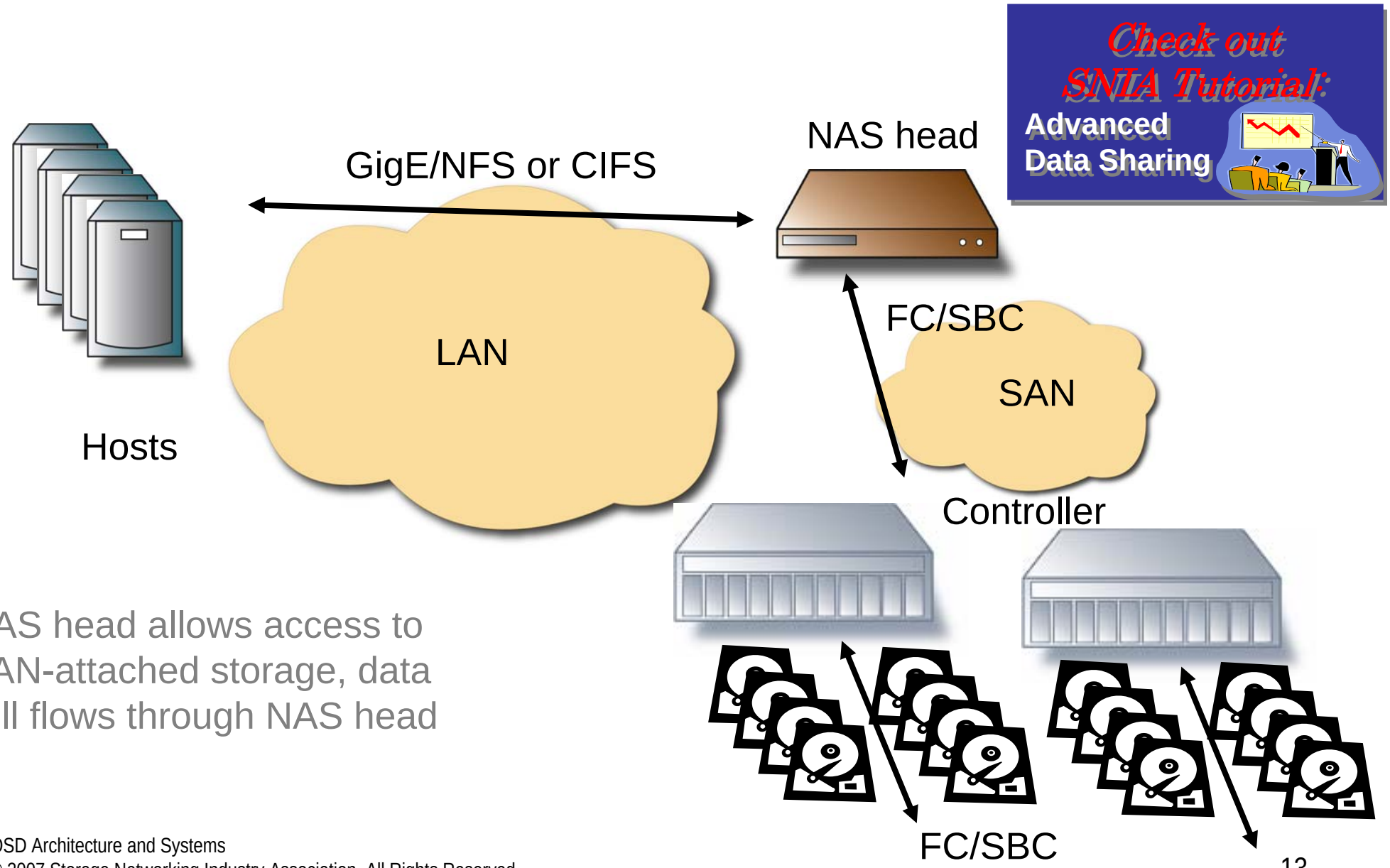
NAS

- Network-Attached Storage (NAS) today



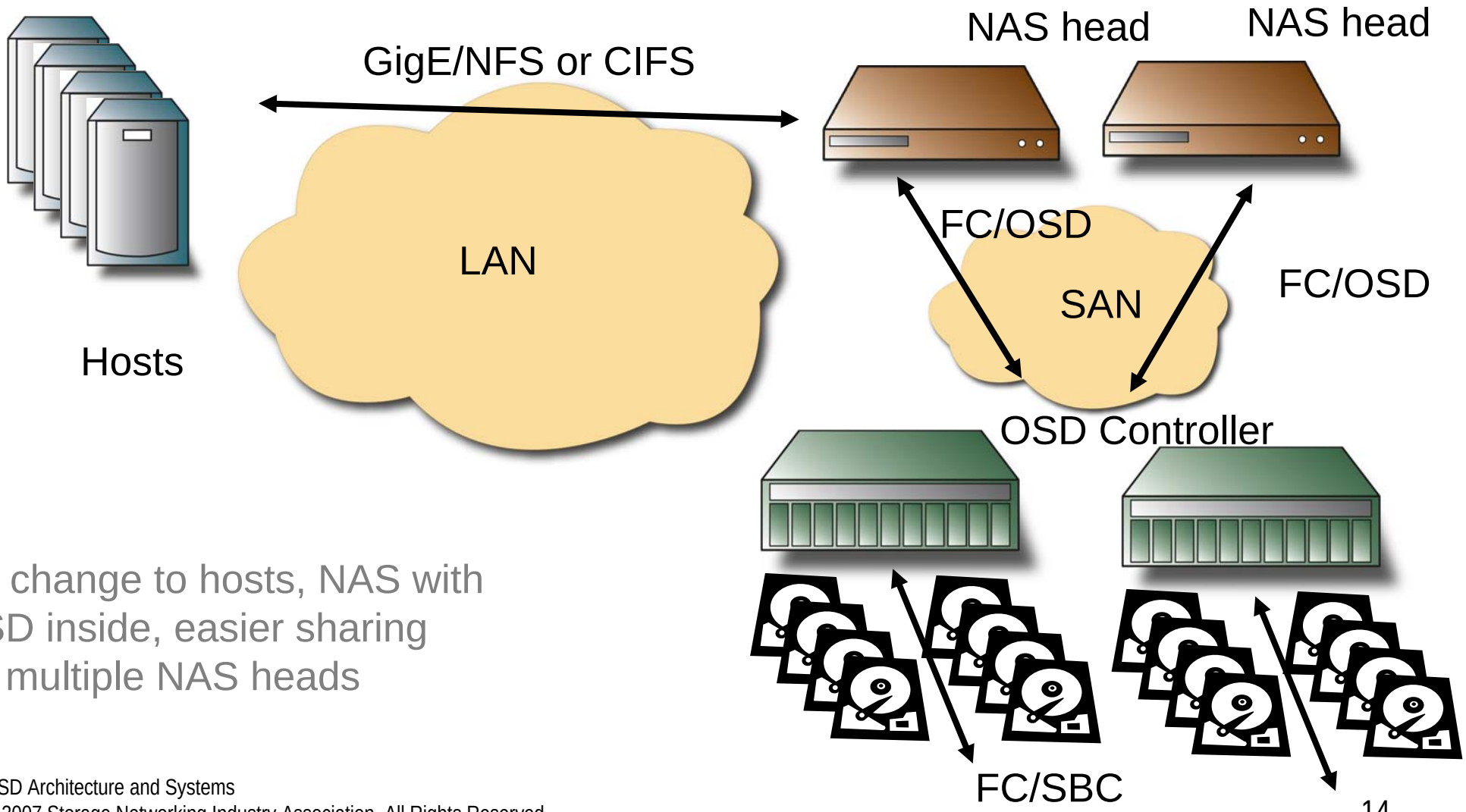
All data transferred through
NAS and translated to disks

NAS Heads



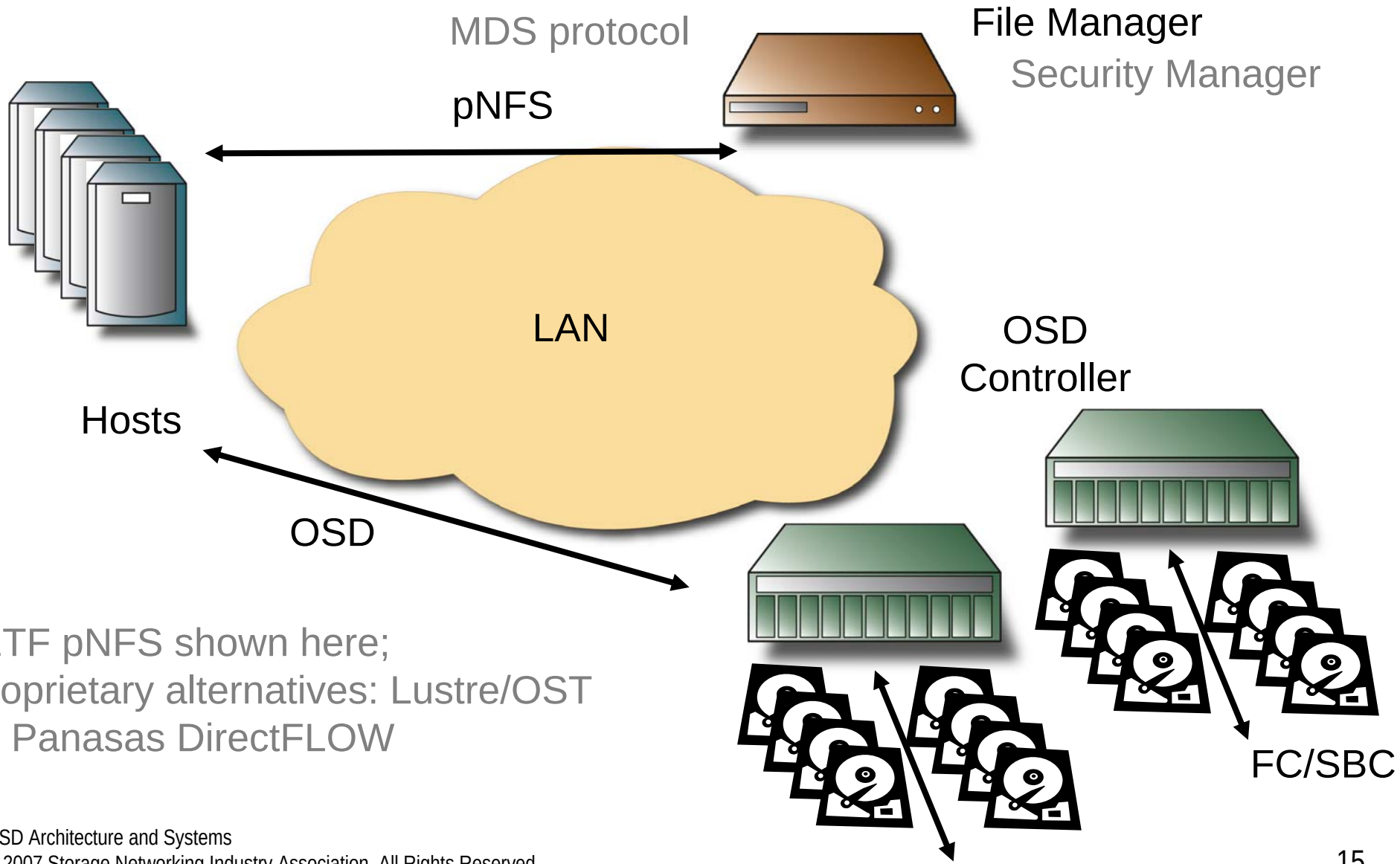
NAS head allows access to SAN-attached storage, data still flows through NAS head

NAS Heads with OSD



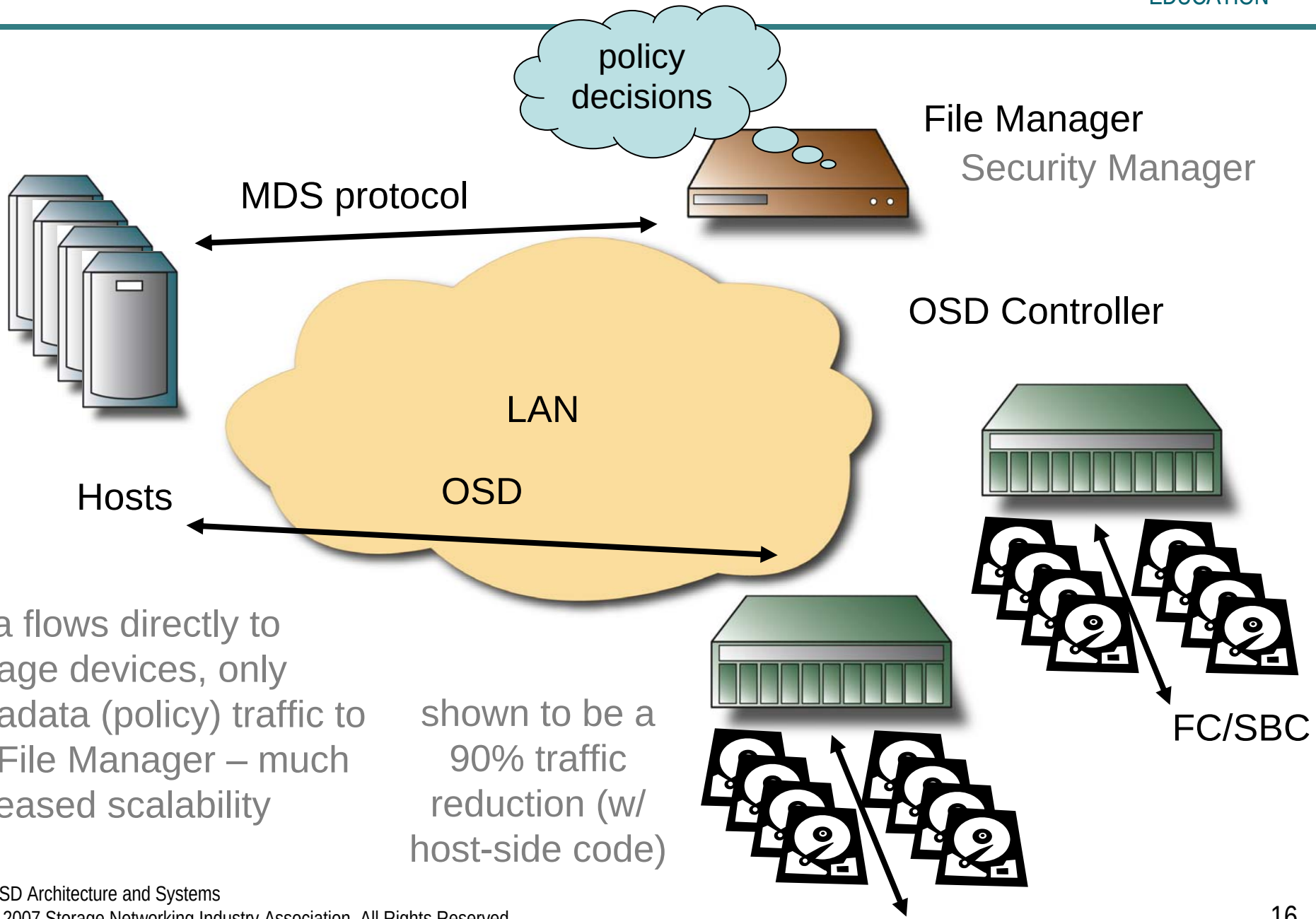
No change to hosts, NAS with OSD inside, easier sharing for multiple NAS heads

Scalable NAS with OSD



IETF pNFS shown here;
proprietary alternatives: Lustre/OST
or Panasas DirectFLOW

Scalable NAS with OSD (2)



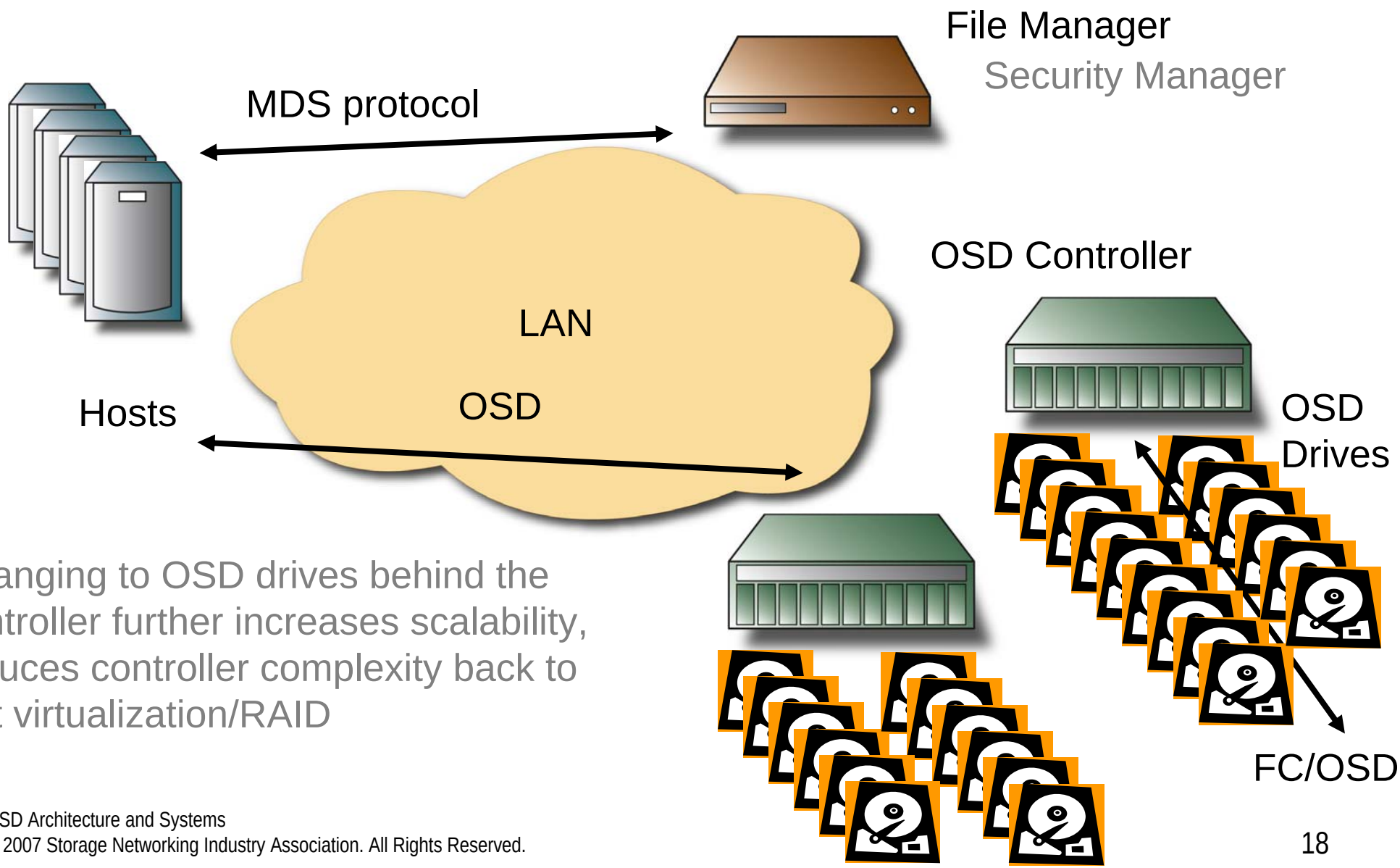
Data flows directly to storage devices, only metadata (policy) traffic to the File Manager – much increased scalability

shown to be a 90% traffic reduction (w/ host-side code)

Scalable NAS with OSD

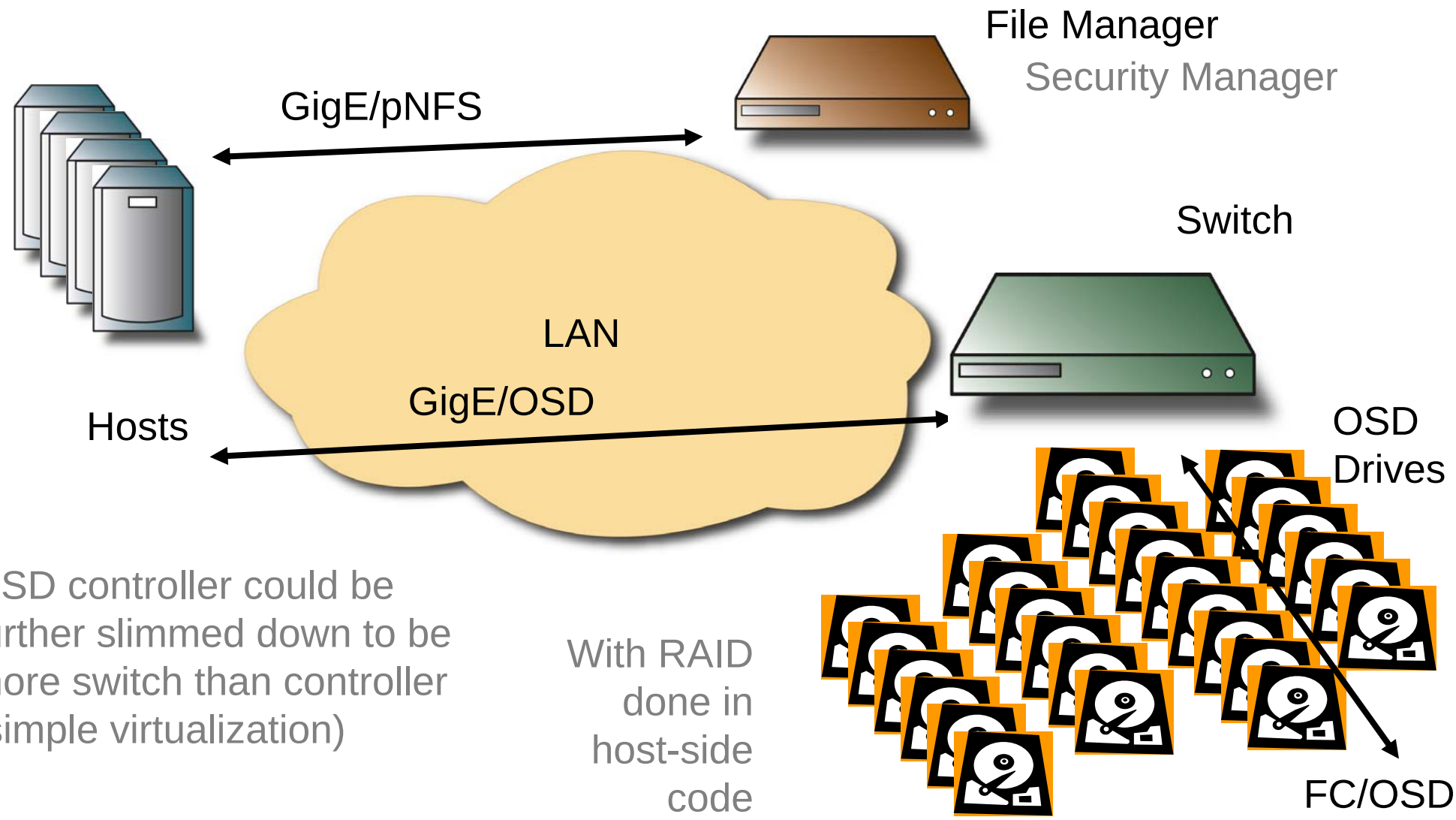
- Protocol to the storage devices is OSD
  - standardized by ANSI T10
  - SCSI/OSD command set on iSCSI, FC, SAS
  - used for all data transfers
- Protocol to the File Manager is FS-specific
  - could be Lustre MDS protocol
  - could be Panasas DirectFLOW
  - could be pNFS server protocol
  - could be application-specific (see next slide)
  - used for policy decisions

Scalable NAS with OSD (3)



Changing to OSD drives behind the controller further increases scalability, reduces controller complexity back to just virtualization/RAID

Scalable NAS with OSD (4)



OSD controller could be further slimmed down to be more switch than controller (simple virtualization)

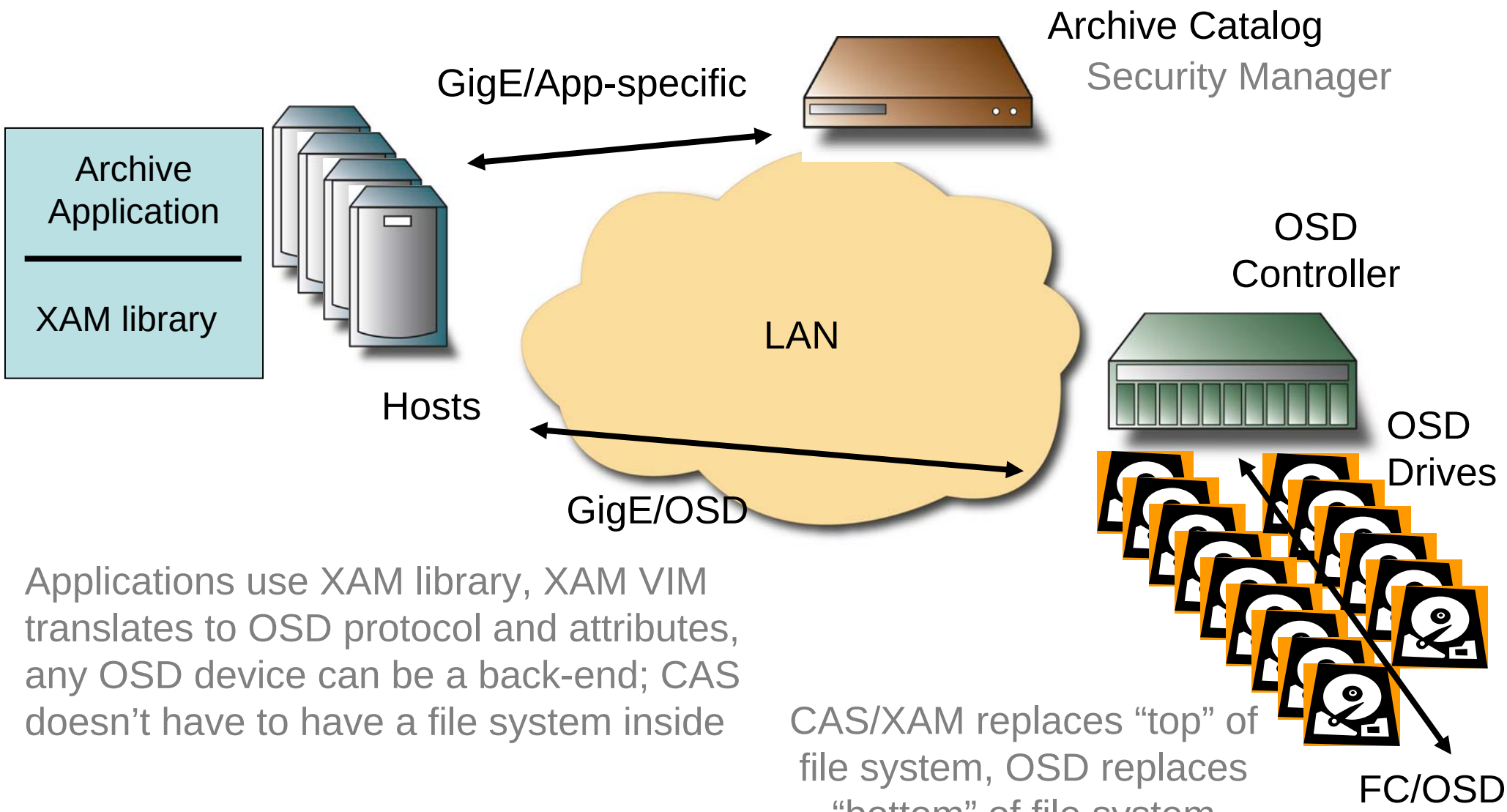
With RAID done in host-side code

Design Choices with OSD

- NAS heads with OSD controllers
  - no host-side changes, shared storage among NAS heads
- Scalable NAS with OSD controllers, block drives
  - benefit of offloading metadata, separating policy management
  - heavier controller, since it must now do space management
- Scalable NAS with OSD controllers, OSD drives
  - controller handles only virtualization and RAID, as it does today
  - benefits of logical objects & native security (more on this later)
- Scalable NAS w/ virtualizing switch, OSD drives
  - most scalable, but requires RAID done in host-side software

Different design choices apply to different system trade-offs

CAS with OSD



Applications use XAM library, XAM VIM translates to OSD protocol and attributes, any OSD device can be a back-end; CAS doesn't have to have a file system inside

CAS/XAM replaces "top" of file system, OSD replaces "bottom" of file system

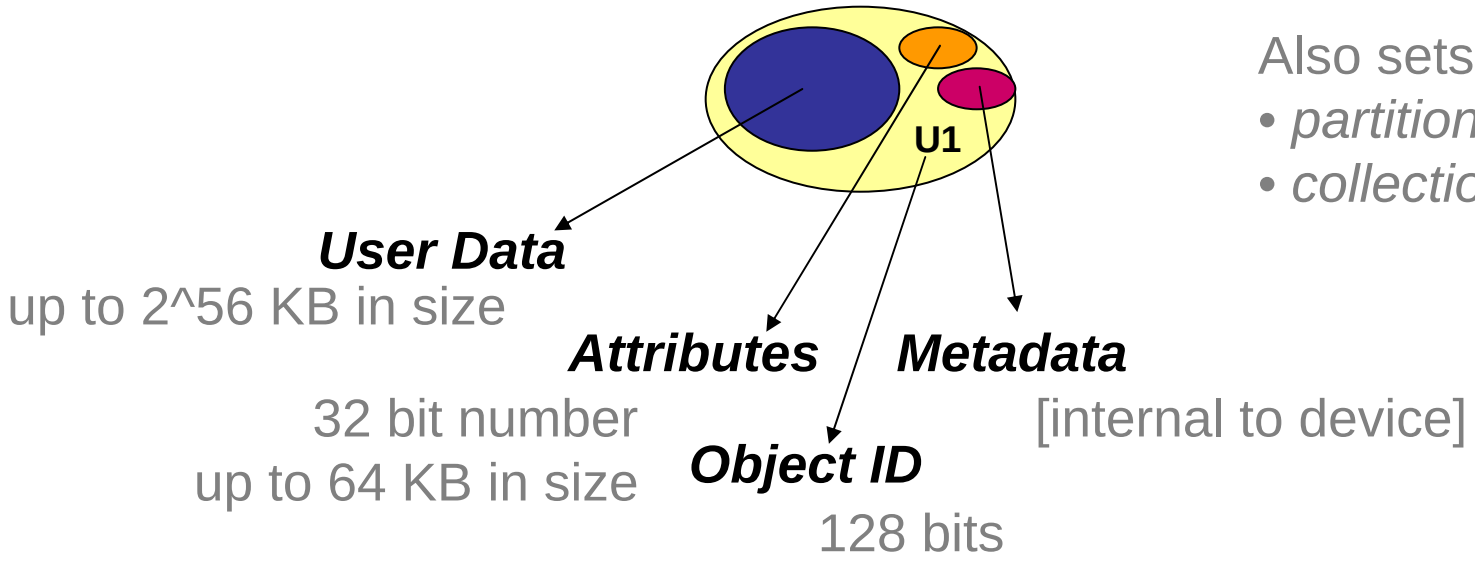
CAS with OSD

- Archive applications are written to XAM API
  - XAM is a host-side library; multiple vendor-specific plug-ins
  - one of those is a mapping to T10/OSD
  - any OSD device can sit behind XAM
  - XAM metadata mapped to OSD attributes
  - metadata + data move through system together (more on this later)
- Protocol to the Catalog or Directory Service is application-specific
  - enterprise content management
  - compliance management (co-located with security manager)
  - ILM policy management for tiered storage (more on this later)

Architecture – Logical View

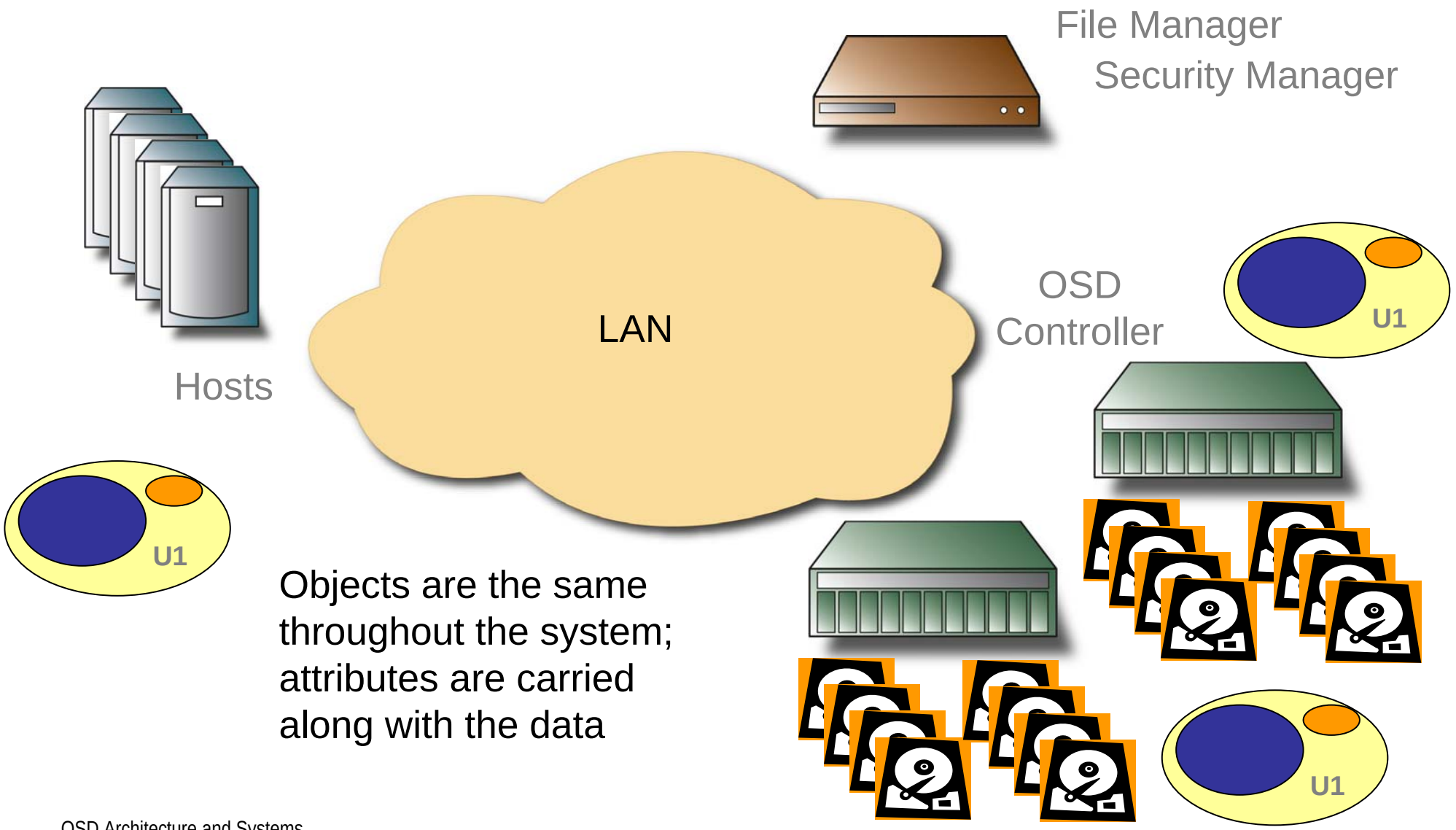
Logical View – Objects

- OSD objects consist of User Data + Attributes
  - device tracks internal metadata (space mgmt)
  - object ID for identification (unique per device)



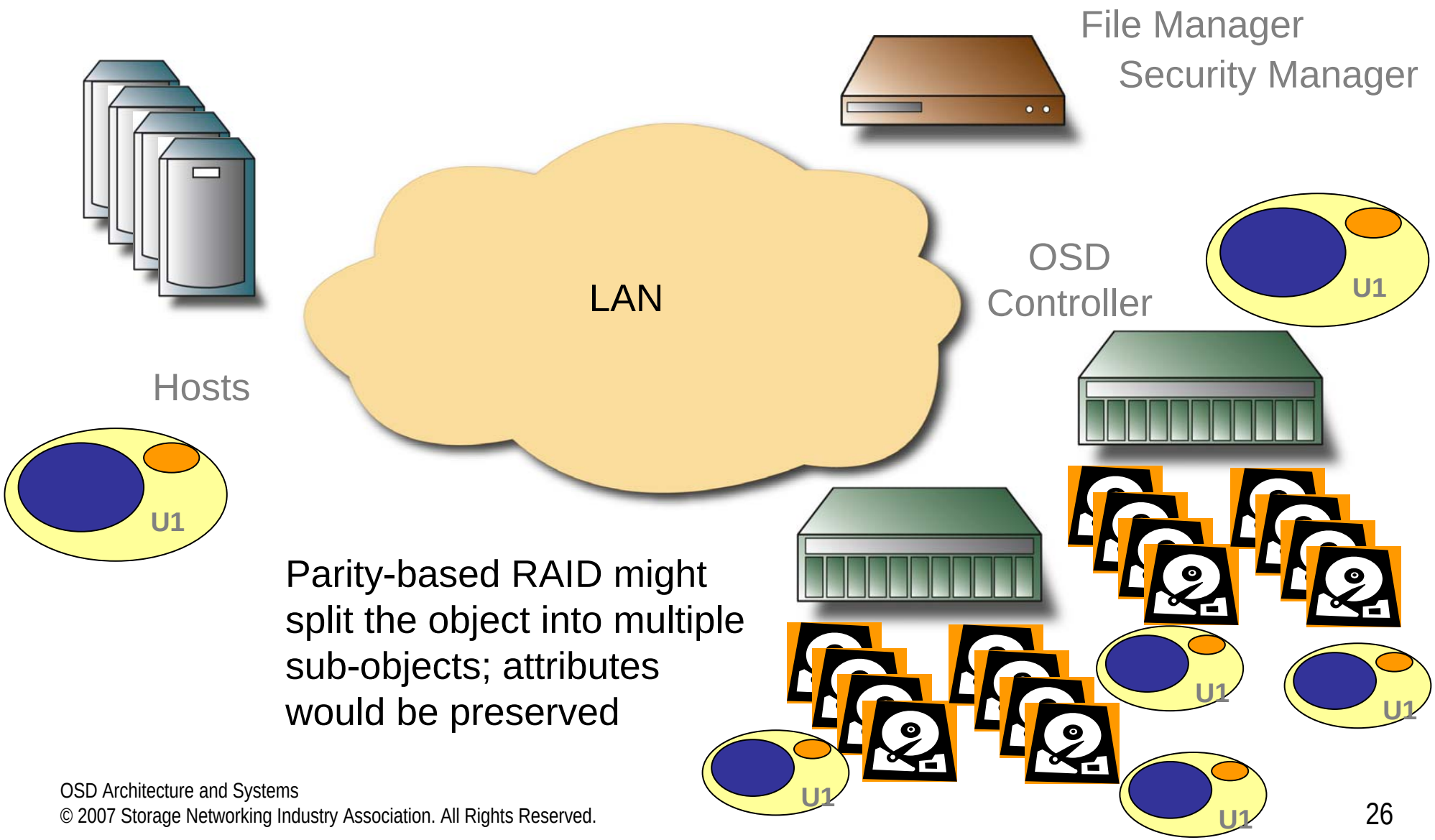
Also sets of objects
- *partitions* – security/quota
- *collections* – grouping

Scalable NAS with OSD



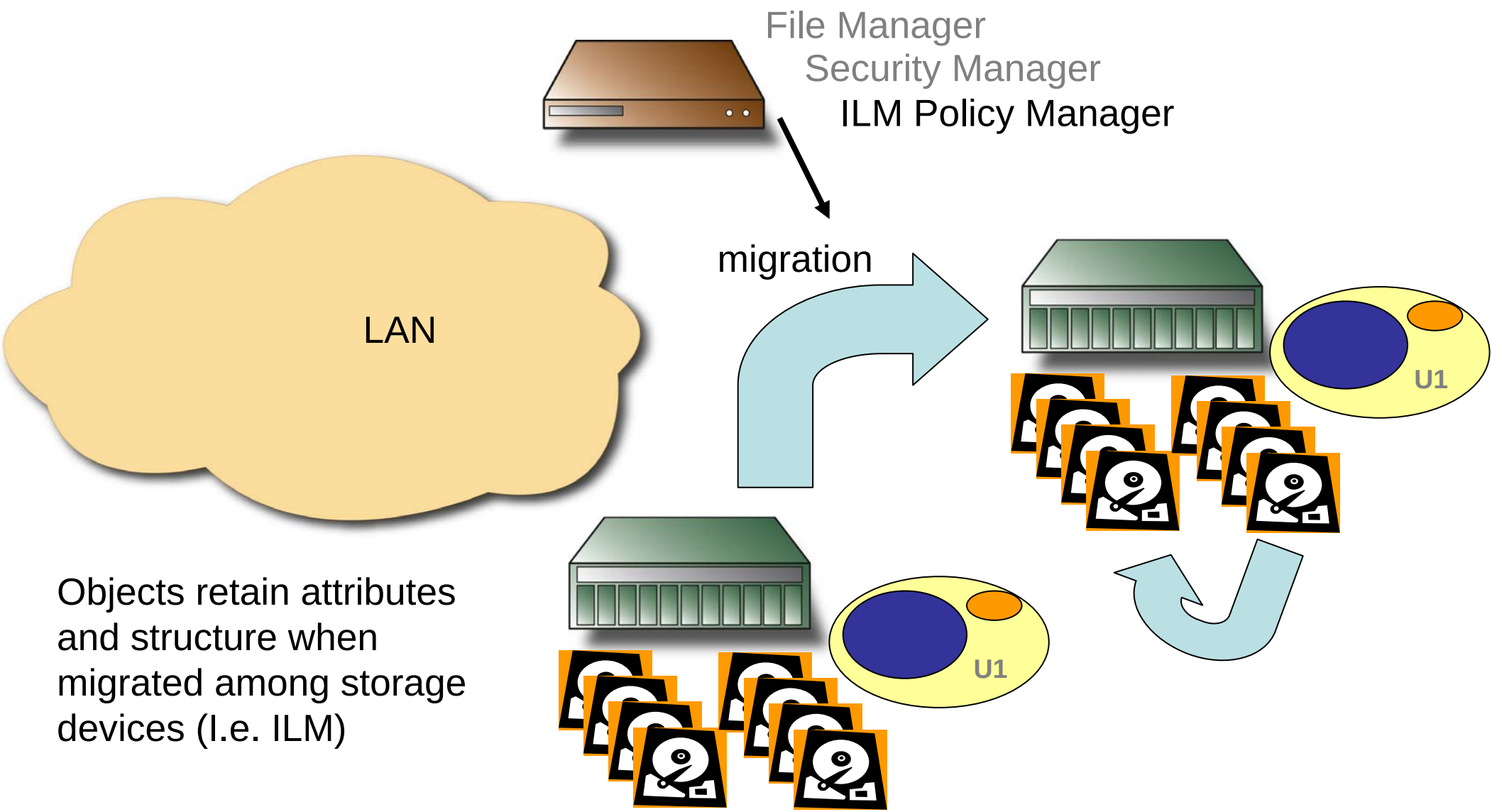
Objects are the same throughout the system; attributes are carried along with the data

Scalable NAS w/ OSD & RAID



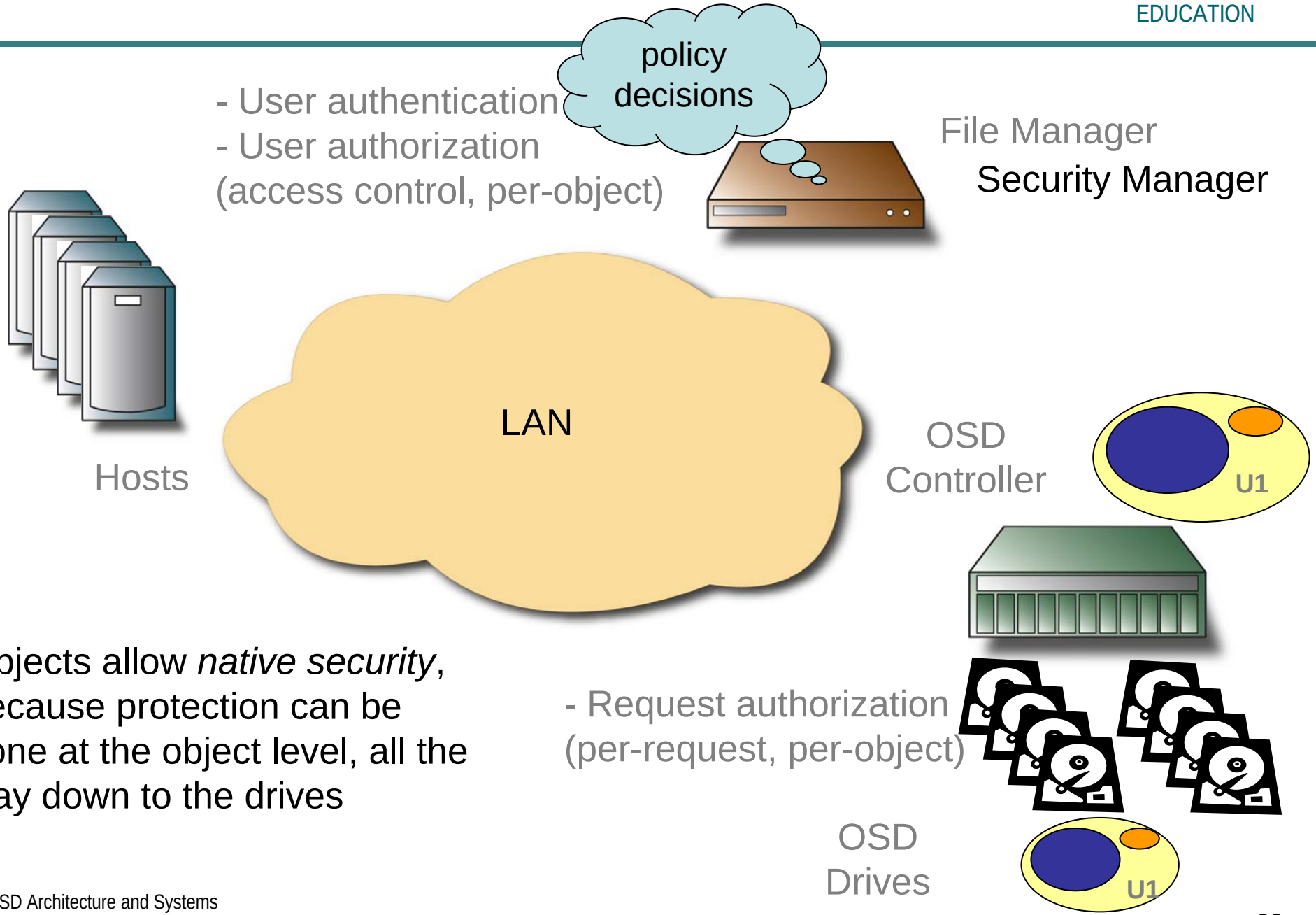
Parity-based RAID might split the object into multiple sub-objects; attributes would be preserved

ILM with OSD



Objects retain attributes and structure when migrated among storage devices (I.e. ILM)

Security with OSD



Objects allow *native security*, because protection can be done at the object level, all the way down to the drives

Security with OSD

- Policy decisions made by the Security Manager
  - *User authentication* (who are you?)
  - *User authorization* (what rights do you have?)
- OSD-enabled devices
  - *Request authorization* (do you have rights now?)
  - [optional] Request or data integrity/confidentiality (in flight)
  - [future] Stored data integrity or confidentiality (encrypt at-rest)
- Comparison to FC-SP and iSCSI security
  - *Device authentication*
  - [optional] In-flight integrity or confidentiality on a session basis
  - No per-request or per-object authorization possible

Status & Next Steps

Status of the Standard

- Standard OSD-1 r10 for Project T10/1355-D (v1) ratified by ANSI in September 2004 after years of SNIA effort
- SNIA TWG working on v2 features
  - Extended exception handling and recovery [draft]
  - Richer collections – multi-object operations [draft]
  - Snapshots – managed on-device [proposal]
  - Mapping of XAM onto OSD [proposal w/ FCAS TWG]
  - Additional security support [discussion]
  - Quality of Service attributes [discussion]
- expect a new round of T10 standardization in mid-2007
  - join us – www.snia.org/tech_activities/workgroups/osd/

Join Us!

SNIA OSD TWG Structure

- Erik Riedel, Seagate (co-Chair)
- Julian Satran, IBM (co-Chair)

Contact us to get involved!

- Error Mgmt & Recovery – C. Mallikarjun, HP
- Snapshots – Oleg Kiselev, Symantec
- Security – Michael Factor, IBM
- Education – <vacant>
- T10 & FCAS Liason – Rich Ramos, Xyratex
- Research – Michael Mesnier, Intel

Further Reference

- Academic research
  - www.pdl.cmu.edu
  - www.dtc.umn.edu
  - ssrc.cse.ucsc.edu/proj/ceph.html
- Standards work
  - www.snia.org/osd
  - www.t10.org/ftp/t10/drafts/osd
- Industry research & development
  - www.lustre.org & www.hp.com/techservers/products/sfs.html
  - www.panasas.com
  - www.seagate.com/docs/pdf/whitepaper/tp_536.pdf
  - www.haifa.il.ibm.com/projects/storage/objectstore
  - www.opensolaris.org/os/project/osd/

SNIA Legal Notice

- 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.

Q&A / Feedback

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

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

SNIA Education Committee

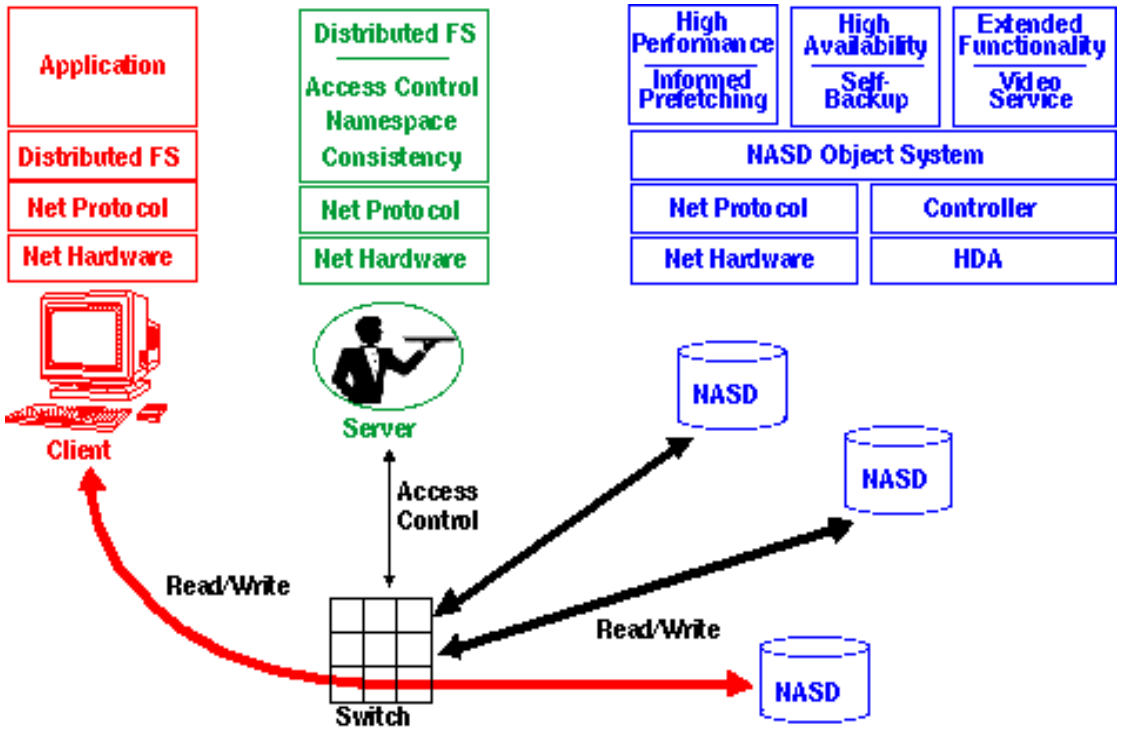
Erik Riedel, Seagate Technology
Dave Nagle, Google
Dave B Anderson, Seagate Technology
Sami Iren, Seagate Technology
Mike Mesnier, Intel & CMU
Elaine Silber, Firefly

Appendix

History

OSD Standard – History

- Started with NSIC NASD research in 1995
  - Network-Attached Storage Devices (NASD)
  - Carnegie Mellon, HP, IBM, Quantum, STK, Seagate
  - Prototypes developed at Carnegie Mellon with funding from DARPA
- Draft standard brought to SNIA in 1999
- Standard ratified by ANSI in 2004



ANSI Project T10/1355-D

revision	date	pages	word count	commands
1	May 2000	77	28,482	14
2	September 2000	84	31,205	15
3	October 2000	94	32,872	16
4	July 2001	111	39,633	15
5	March 2002	116	40,372	16
5t	August 2002	144	51,248	17
6	August 2002	145	51,556	18*
7	June 2003	168	58,405	18
8	September 2003	147	47,614	18
9	February 2004	174	60,736	20
10	July 2004 (ratified)	187	65,216	23

SCSI Object-Based Storage Device Commands (OSD)

OSD Standard – to 2007

- Seagate & IBM co-chair SNIA OSD Technical Work Group
- EMC, HP, Intel, Panasas, Veritas, Xyratex were the most active participants leading up to OSD-1
  - 30 companies, 6 universities/labs paying attention today
- Lustre – CFS/HP open-source OSD for DoE
  - 225 TB cluster installed October 2002; 100+ active sites today
- Panasas shipping OSD-based scalable NAS
  - since October 2003; large-scale systems (300+ device demo)
- IBM, Seagate, and Emulex demo shown at SNW
  - first T10/OSD interoperability demonstration in April 2005
  - with FC/OSD drives, iSCSI/OSD controller, modified SAN file system
- Sun developing drivers and file system for OSD objects
  - OSD drivers for OpenSolaris released in December 2006
- Ongoing university work at UC – Santa Cruz (UCSC), Carnegie Mellon, Univ of Minnesota (UMN), Ohio-State (OSU) and Texas A&M