SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

OSD Technical Work Group

Object Storage and Applications

Erik Riedel and Sami Iren

Seagate Technology

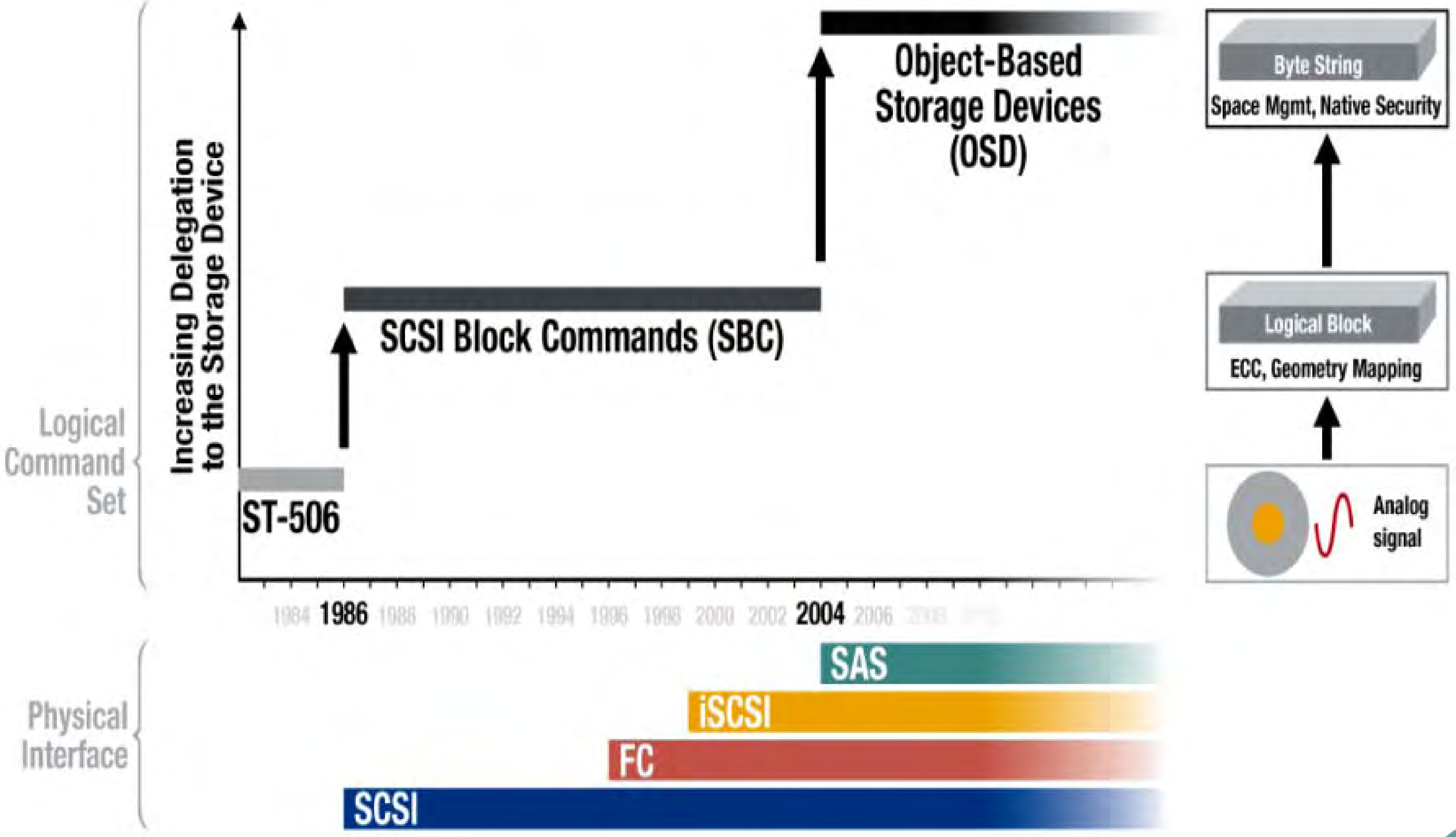
February 2007

Seagate

We turn on ideas

Object-Based Storage—Delegate Deeper

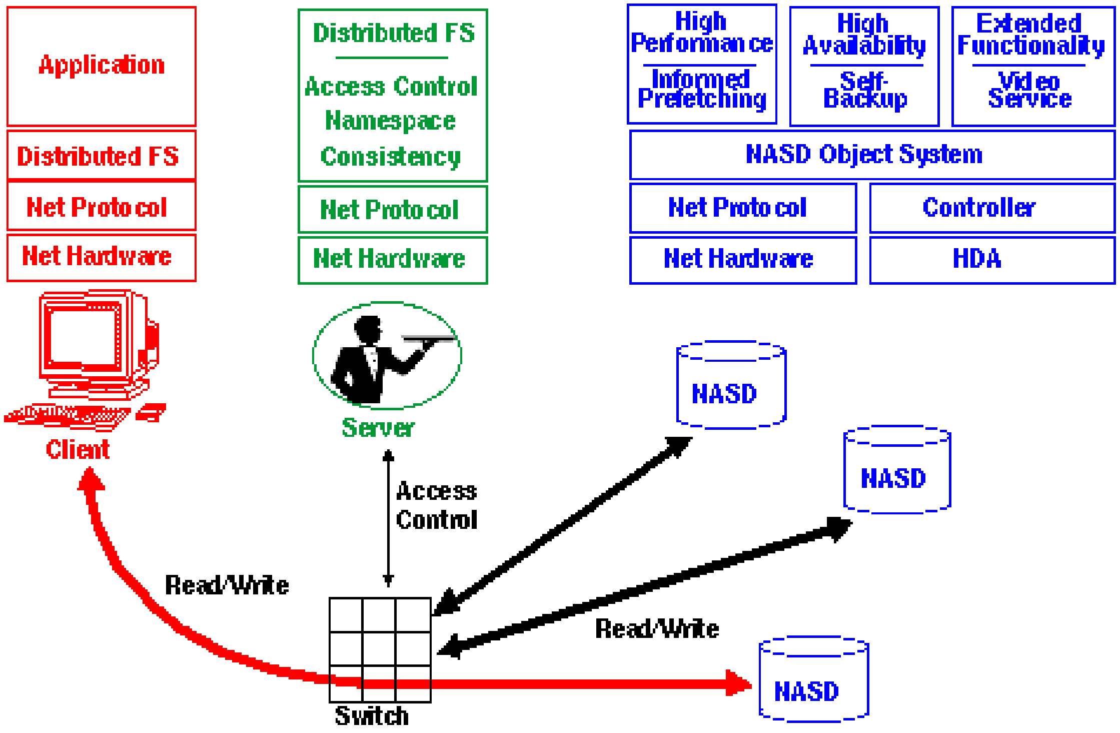
A Fundamental Change in Technology



Seagate
We turn on ideas

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 Commands

OSD-1 r10, as ratified

- Basic Protocol
  - READ
  - WRITE

  very basic

  - CREATE
  - REMOVE

  space mgmt

  - GET ATTR
  - SET ATTR

  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

OSD Systems – 2006

A variety of Object-based Storage Devices being built today

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

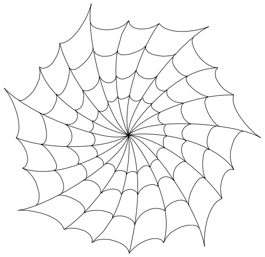
- "Smart" disk for objects
- E.g. Panasas storage blade

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

- **File/ Security Manager**

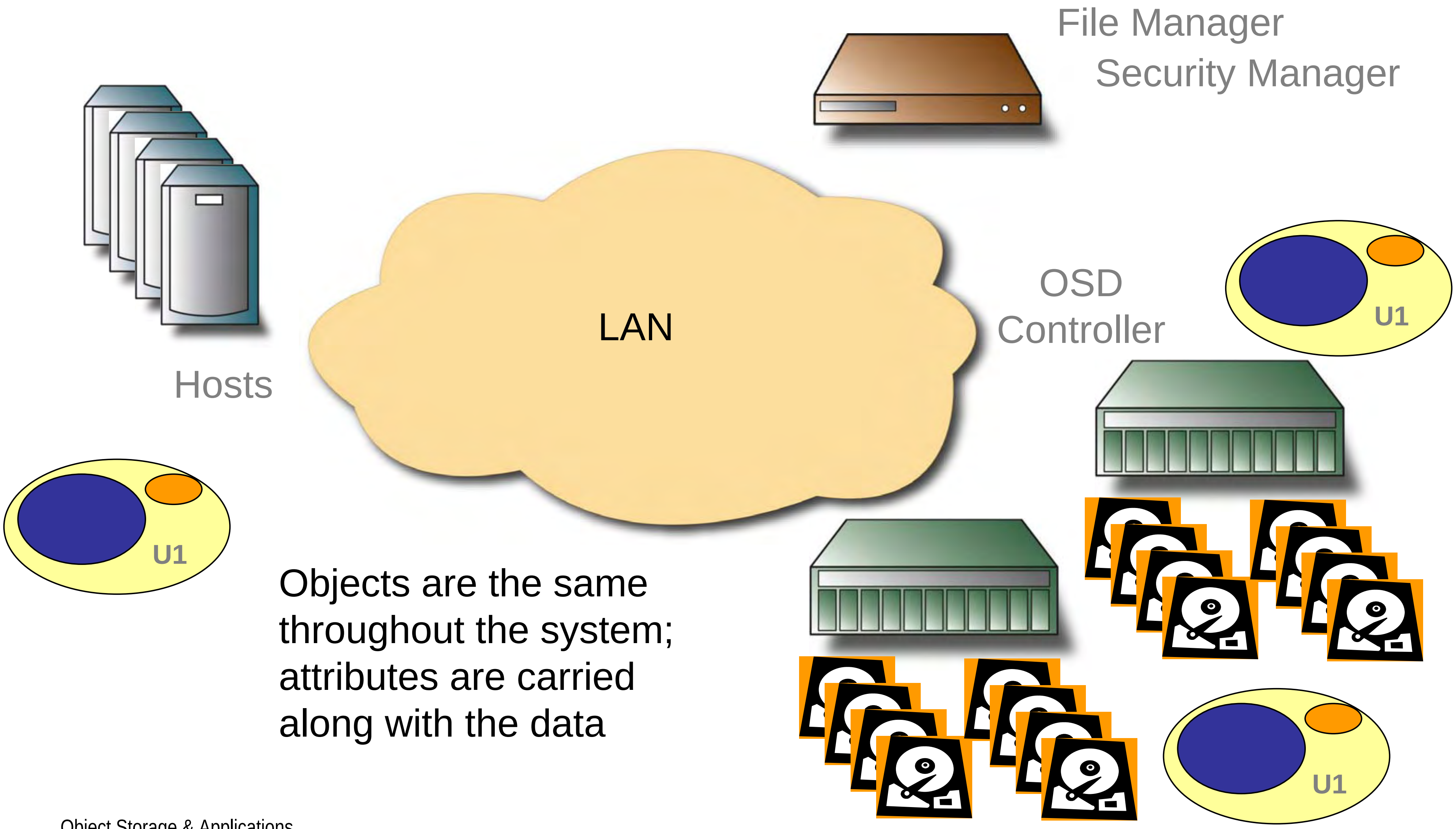
- Orchestrates system activity
- Balances objects across OSDs
- Called clustered MDS in Lustre
- Called Mgmt Blade by Panasas
- Called ST server cluster by IBM

- **Scalable Network**



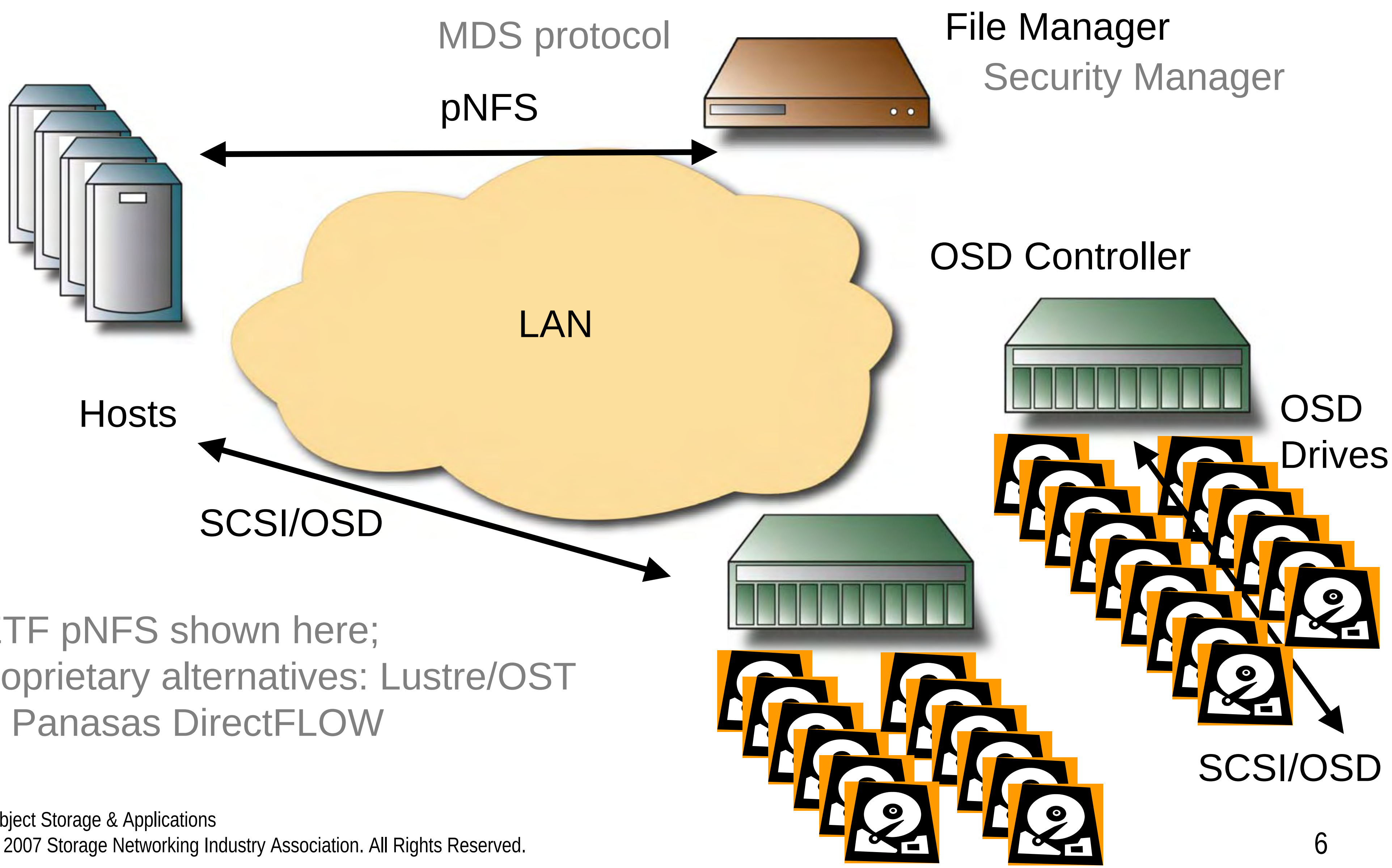
- Connectivity among clients, managers, and devices
- Shelf-based GigE (Panasas)
- Specialized cluster-wide high-performance network (Lustre)
- Storage network (IBM)

Scalable NAS with OSD



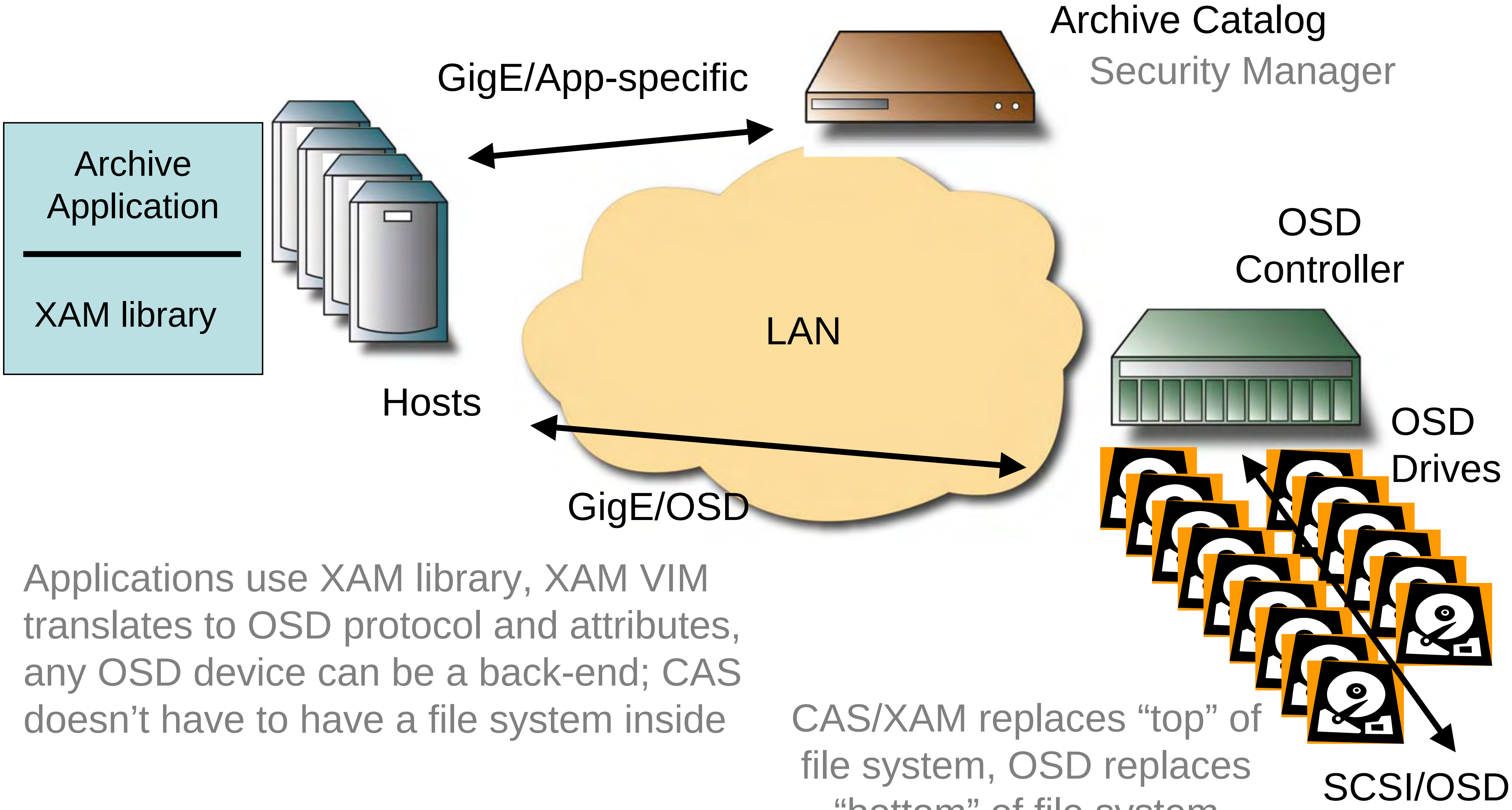
Objects are the same throughout the system; attributes are carried along with the data

Scalable NAS with OSD



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

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

Advantages w/ Objects

- Semantics for more sophisticated data mgmt
  - Flexible space management
  - Metadata tags sit alongside object data
  - Error reporting can be done on an object basis
    - Clear hook for reporting damage (e.g. fence bits)
  - Native strong security
    - Authorization directly at devices via capabilities
  - Self-managing devices
    - Offload common activity; scale with devices
  - Differentiate data types via attributes (next slides)

Attributes

range for each object type

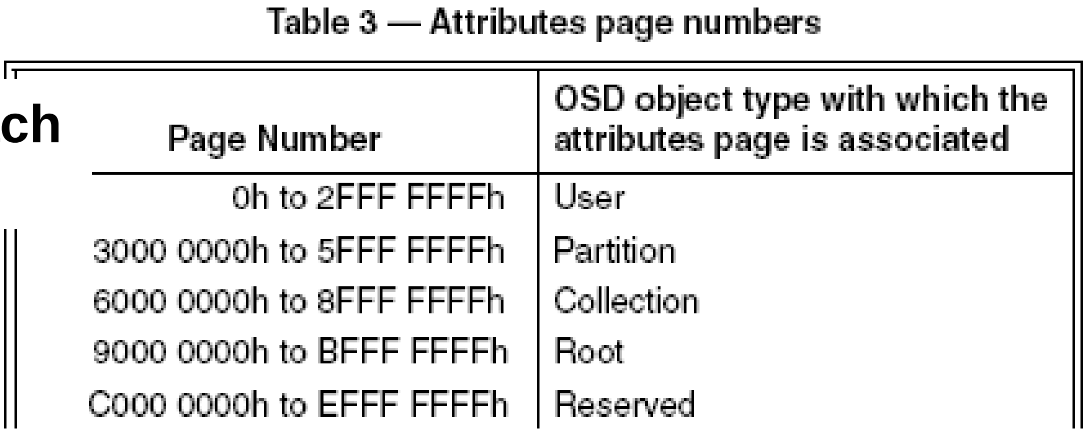


Table 3 — Attributes page numbers

Page Number	OSD object type with which the attributes page is associated
0h to 2FFF FFFFh	User
3000 0000h to 5FFF FFFFh	Partition
6000 0000h to 8FFF FFFFh	Collection
9000 0000h to BFFF FFFFh	Root
C000 0000h to EFFF FFFFh	Reserved

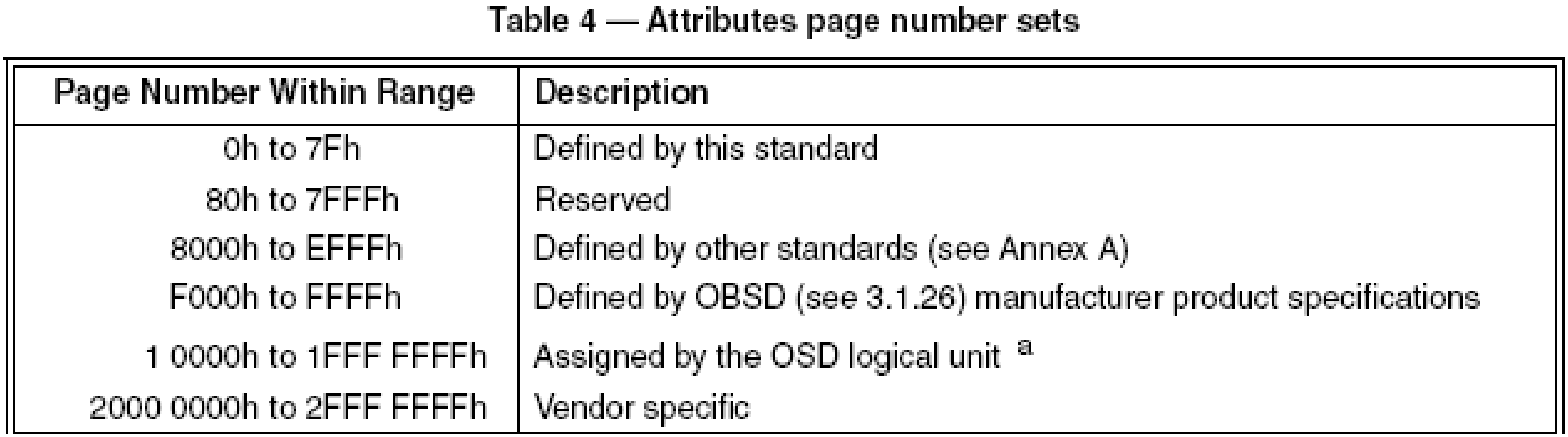


Table 4 — Attributes page number sets

Page Number Within Range	Description
0h to 7Fh	Defined by this standard
80h to 7FFFh	Reserved
8000h to EFFFh	Defined by other standards (see Annex A)
F000h to FFFFh	Defined by OBSD (see 3.1.26) manufacturer product specifications
1 0000h to 1FFF FFFFh	Assigned by the OSD logical unit [a]
2000 0000h to 2FFF FFFFh	Vendor specific

Limited number defined by standard

- length, size, timestamps

Vendor extensions

- opaque – for application use only
- shared – device-interpreted (impacts behavior)

Also used to do device-level params

- security level
- capacity
- …

Extensions w/ Attributes

- Specify additional semantics at per-object level
  - Example – reliability levels
    - <low> vs. <medium> vs. <high>
  - Example – QoS handling
    - <sequential> vs. <random>
    - <bandwidth=x>

      (this may want session-based OPEN/CLOSE)
  - Example – compliance
    - <expiration date> or <write-once>
  - Example – database access
    - <field size> or <layout schema>

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 OSD-2 features
  - Extended exception handling and recovery [draft]
  - Richer collections – multi-object operations [draft]
  - Snapshots – managed on-device [proposal]
  - Mapping of XAM onto OSD [ongoing w/ FCAS TWG]
  - Additional security support [discussion]
  - Quality of Service attributes [discussion]
  - Device-to-device data migration [early discussion]
- expect a new round of T10 standardization in 2007
  - join us – www.snia.org/tech_activities/workgroups/osd/

References

- Standards work
  - www.snia.org/members/twg_ip/ (OSD TWG)
    (if SNIA member, sign up via company account, else email Erik)
  - www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf
  - www.t10.org/ftp/t10/drafts/osd2/osd2r01.pdf
- Tutorials
  - www.snwusa.com/documents/presentations-f06/ErikRiedel.pdf
  - www.snia.org/education/tutorials/spr2005/storage (at bottom)
- Academic research
  - www.pdl.cmu.edu ; www.dtc.umn.edu ; csl.cse.ucsc.edu/obsd.shtml
- Industry research & development
  - www.haifa.ibm.com/projects/storage/objectstore
  - www.lustre.org ; www.panasas.com
  - www.hp.com/techservers/products/sfs.html

Appendix

OSD Standard – to 2006

- Seagate & IBM co-chair OSD Technical Work Group
- EMC, HP, Intel, Panasas, Veritas, Xyratex were the most active participants leading up to OSD-1
  - 35 companies, 5 universities 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 released OSD driver stack for OpenSolaris in December 2006
- Ongoing university work at UC – Santa Cruz, Carnegie Mellon, Univ of Minnesota, Ohio-State and Texas A&M