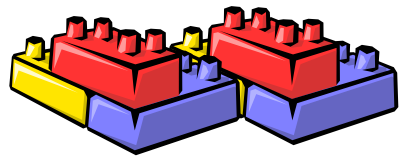
Object-based Storage Devices (OSD) T10 Standard

Erik Riedel
Seagate Research
November 2003

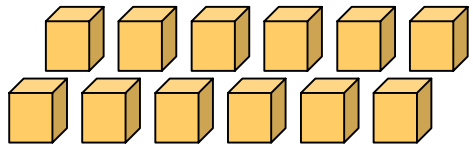
Motivation for OSD

– Improved device and *data sharing*
  - Platform-dependent metadata moved to device
  - Systems need only agree on naming
– Improved *scalability & security*
  - Devices directly handle client requests
  - Object security w/ application-level granularity
  - Finer granularity than LUN-based security
– Improved *performance*
  - Hints, QoS, Differentiated Services
– Improved *storage management*
  - Self-managed, policy-driven storage
  - Storage devices more autonomous

Volumes

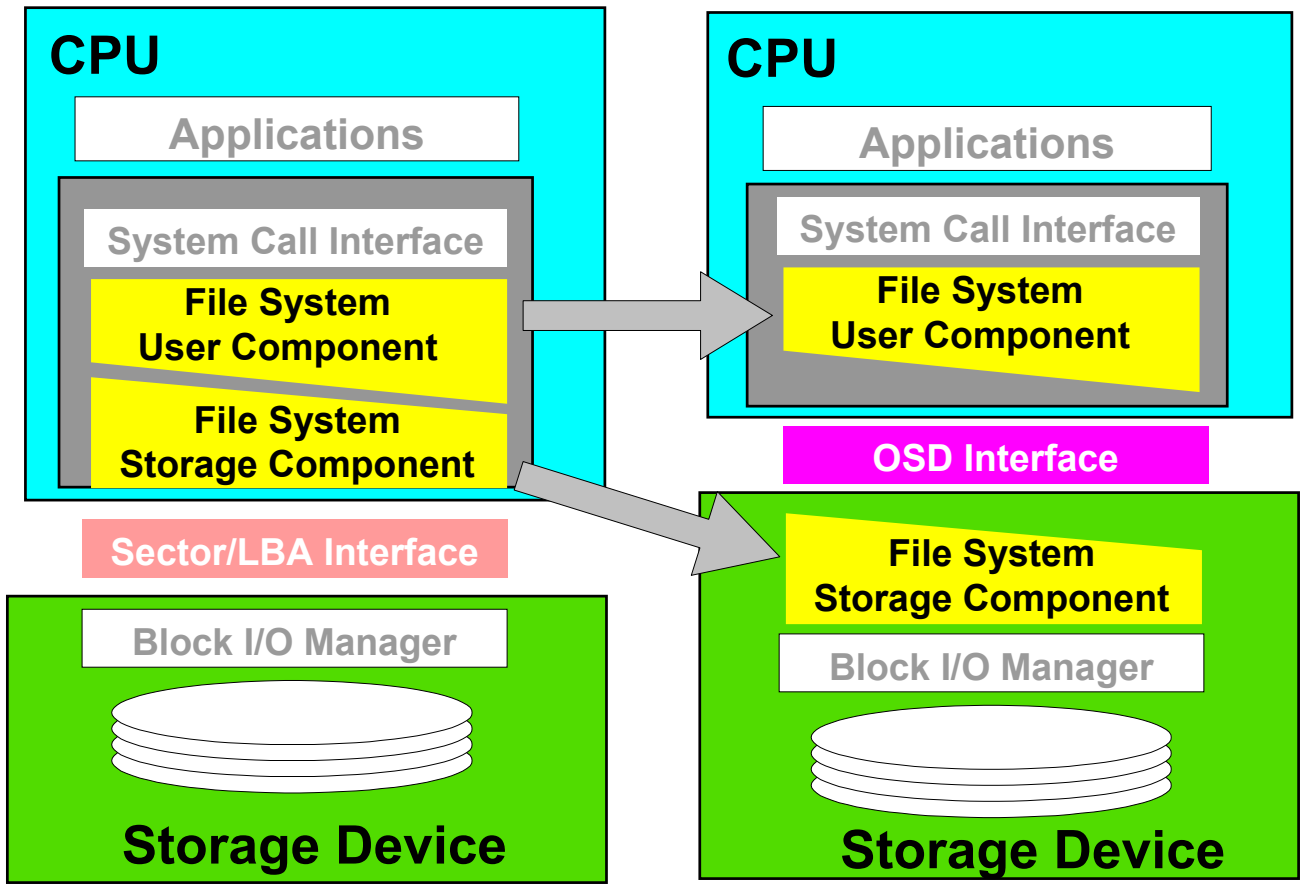


Objects



Blocks

OSD Interface



OSD overview

Basic Protocol

- READ
- WRITE

} Very Basic

- CREATE
- REMOVE

} Space Mgmt

- GET ATTR
- SET ATTR

} Attributes

- timestamps
- vendor-specific
  - shared, opaque

Security

- Authorization – on each request
- Integrity – for args & data
- SET KEY
- SET MASTER KEY

} shared secrets

Groups

- CREATE COLLECTION
- REMOVE COLLECTION
- LIST COLLECTION

Specialized

- APPEND – write w/o offset
- CREATE & WRITE – save msg
- FLUSH OBJ – force to media
- LIST – recovery of objects

Management

- FORMAT OSD
- CREATE PARTITION
- REMOVE PARTITION

Read (8805h) – parameters

Bit Byte	7	6	5	4	3	2	1	0
8	(MSB)							
9				SERVICE ACTION (8805h)				(LSB)
10				OPTIONS BYTE				
11	Reserved			GET/SET CDBFMT		PRIORITY		
12	(MSB)							
19				PARTITION_ID				(LSB)
20	(MSB)							
27				USER_OBJECT_ID				(LSB)
28								
31				Reserved				
32	(MSB)							
39				LENGTH				(LSB)
40	(MSB)							
47				STARTING BYTE ADDRESS				(LSB)

64 bits

64 bits

byte addressable

List (8803h) – parameters

Table 38 — LIST service action

Bit / Byte	7	6	5	4	3	2	1	0
8	(MSB)							
9				SERVICE ACTION (8803h)				(LSB)
10				Reserved				
11	Reserved			GET/SET CDBFMT		SORT ORDER		
12	(MSB)							
19				PARTITION_ID				
20	(MSB)							
27				USER_OBJECT_ID TAG				(LSB)
28								
31				Reserved				
32	(MSB)							
39				ALLOCATION LENGTH				
40	(MSB)							
43				LIST TAG				

only one option – ascending object id

buffer size available

continuation across commands

Objects

Object names

Table 8 — Partition_ID and User_Object_ID value assignments

Partition_ID	User_Object_ID	Description
0h	0h	Root object
0h	1h - FFFF FFFF FFFF FFFFh	Reserved
1h - FFFFh	0h - FFFF FFFF FFFF FFFFh	Reserved
10000h - FFFF FFFF FFFF FFFFh	0h	Partition (assigned by OSD)
10000h - FFFF FFFF FFFF FFFFh	1h - FFFFh	Reserved
10000h - FFFF FFFF FFFF FFFFh	10000h - FFFF FFFF FFFF FFFFh	User object (assigned by OSD)

Partition IDs assigned by device

- primary usage case assumes one manager per partition

Object IDs assigned by device OR by host

- collection IDs share namespace with objects

Attributes

Attributes

range for each object type

Table 9 — Object attribute page numbers

Page number	OSD object type with which the object attributes page is associated
0h - 2FFF FFFFh	User
3000 0000h - 5FFF FFFFh	Partition
6000 0000h - 8FFF FFFFh	Collection
9000 0000h - BFFF FFFFh	Root
C000 0000h - FFFF FFFEh	Reserved
FFFF FFFFh	All attributes pages

Table 10 — Object

Page Number within range	Description	Page Number	Attributes List
0h - 7Fh	Defined by this standard	Yes	Yes
80h - 7FFFh	Reserved		
8000h - EFFFh	Defined by other standards	Yes	Yes
F000h - FFFFh	Defined by OSD manufacturer product specifications	Yes	Yes
1 0000h - 2FFF FFFFh	Assigned by the OSD logical unit **ephemeral**	No	Yes
2000 0000h - 2FFF FFFFh	Vendor specific	n/a	n/a

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
- …

Table 26 — Page format get and set attributes CDB parameters format

Bit / Byte	7	6	5	4	3	2	1	0
79				⋮ Other CDB fields				
80	(MSB)			GET ATTRIBUTES PAGE				
83								(LSB)
84				GET ATTRIBUTES ALLOCATION LENGTH				
87								(LSB)
88				RETRIEVED ATTRIBUTES OFFSET				
91								
92	(MSB)			SET ATTRIBUTE PAGE				
95								(LSB)
96	(MSB)			SET ATTRIBUTE NUMBER				
99								(LSB)
100				SET ATTRIBUTE LENGTH				
103								(LSB)
104				SET ATTRIBUTES OFFSET				
107								
108				⋮ Other CDB fields				

which attrib

how much buffer host has available

which attrib

how much attribs I am sending

Object attributes

Table 64 — User Object Information attributes page contents

Attribute Number	Bytes	Attribute	May be set	OSD Provided
0h	40	Page identification	No	Yes
1h	8	Partition_ID	No	Yes
2h	8	User_Object_ID	No	Yes
3h - 7h		Reserved	No	
8h	8	Used capacity **size**	No	Yes
9h - Bh		Reserved	No	
Ch	variable	Username	Yes	No
Dh - Eh		Reserved **length**	No	
Fh	8	Object logical length	No	Yes
10h - FFFF FFFFh		Reserved	No	

Table 69 — User Object Resources attributes page contents

Attribute Number	Bytes	Attribute	May be set	OSD Provided
0h	40	Page identification	No	Yes
1h	8	Capacity quota **quota**	Yes	No
2h - 7h		Reserved	No	
8h	8	Starting byte address of write or append	No	Yes
9h - FFFF FFFFh		Reserved	No	

Object attributes (2)

Table 77 — User Object Timestamps attributes page contents

Attribute Number	Bytes	Attribute	May be set	OSD Provided
0h	40	Page identification	No	Yes
1h	6	Created time	No	Yes
2h	6	Attributes accessed time	No	Yes
3h	6	Attributes modified time	No	Yes
4h	6	Data accessed time	No	Yes
5h	6	Data modified time	No	Yes
6h - FFFF FFFFh		Reserved	No	

Table 79 — Collections attributes page contents

Attribute Number	Bytes	Attribute	May be set	OSD Provided
0h	40	Page identification	No	Yes
1h - FFFF FF00h	0 or 8	Collection pointer	Yes	No
FFFF FF01h - FFFF FFFFh		Reserved	No	

set of collections an object belongs to

Security

Read – security

48 59	protect arguments	REQUEST INTEGRITY CHECK VALUE (LSB)
60 71	protect replays	REQUEST NONCE
72 75	protect attributes and data	DATA-IN INTEGRITY CHECK VALUE OFFSET
76 79		DATA-OUT INTEGRITY CHECK VALUE OFFSET
80 107		Get and set attributes parameters (see 5.1.1.3)
108 163		Capability (see 4.6.4.4.3)

How to get integrity values

Table 81 — Current Command attributes page contents

Attribute Number	Bytes	Attribute	May be set	OSD Provided
0h	40	Page identification	No	Yes
1h		Reserved	No	
2h	8	Created User_Object_ID	No	Yes
3h	12	Response integrity check value	No	Yes
4h - FFFF FFFFh		Reserved	No	

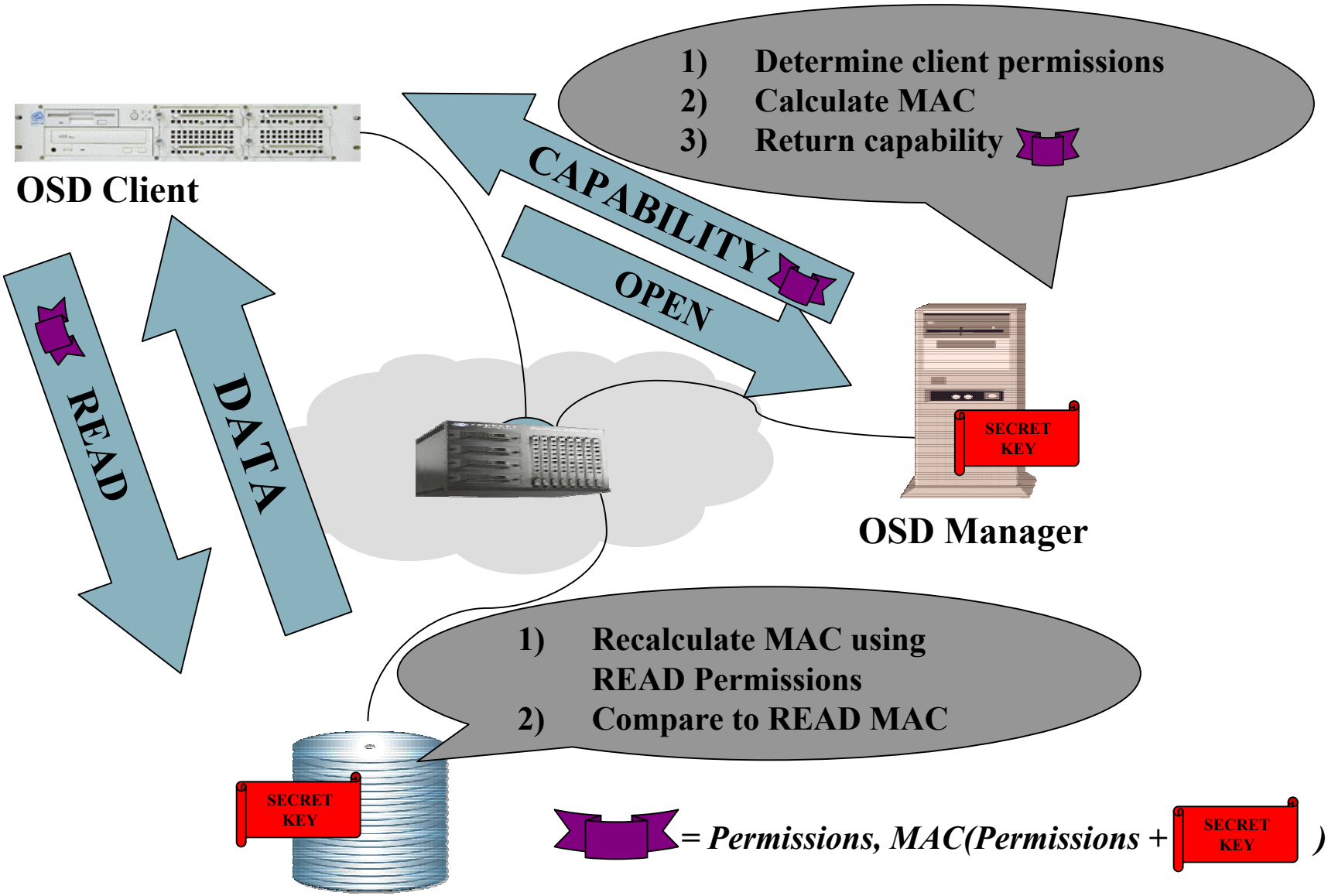
Special attribute to read the integrity value

overall structure

Table 1 — OSD Data-In Buffer and Data-Out Buffer model

Bit / Byte	7	6	5	4	3	2	1	0
0	Traditional command or parameter							
m-1								
m	Meta data							
n-1								
n	Integrity check value							
n+11								

OSD Security – Illustrated



Security levels

Table 13 — Security level threats thwarted

Threat	Threat Thwarted by Security Level			
	0	1	2	3
Forgery of credential	No	Yes	Yes	Yes
Alteration of capabilities	No	Yes	Yes	Yes
Replay of command or status	No	No	Yes	Yes
Alteration of command or status	No	No	Yes	Yes
Replay of data	No	No	No	Yes
Alteration of data	No	No	No	Yes
Inspection of command, status or data	No	No	No	No

Level 1 needs to be layered

Level 3 needs streaming SHA-1

Credentials

Table 14 — Credential format

Bit / Byte	7	6	5	4	3	2	1	0
0				Capability (see 4.6.4.4.3)				
55								
56				OSD SYSTEM ID				
73								
74	(MSB)			PARTITION_ID				
81								
82	(MSB)			OBJECT CREATION TIME				
87								(LSB)
88				Reserved				
91								
92	(MSB)			CREDENTIAL INTEGRITY CHECK VALUE				
111								(LSB)

uniquely identify object in time

Credential format (2)

Table 15 — Capability format

Bit / Byte	7	6	5	4	3
0	Reserved				CREDENTIAL/CAPABILITY FORMAT (1h)
1	KEY VERSION				INTEGRITY CHECK VALUE ALGORITHM
2	(MSB)				
			CAPABILITY EXPIRATION TIME		
					(LSB)
			NONCE AUDIT		
12	(MSB)				
			NONCE RANDOM NUMBER		
23					(LSB)
24					
			PERMISSIONS BIT MASK		
29					
30					
31	USER OBJECT DESCRIPTOR TYPE				Reserved
32					
			USER OBJECT DESCRIPTOR		
55					

only one option today

which key

one today – SHA-1

expiration – may vary by manager, by op, by object

provide an optional "host tag"

more complex specification in future

which object

Table 18 — Capabilities Permission Bits

Bit	Allowed Operation	Bit	Allowed Operation
APP	APPEND	RD	READ
CRE	CREATE or CREATE AND WRITE	RMV	REMOVE
CRE_COL	CREATE COLLECTION	RMV_COL	REMOVE COLLECTION
CRE_PART	CREATE PARTITION	RMV_PART	REMOVE PARTITION
FLS_O	FLUSH OBJECT	ST_ATTR	SET ATTRIBUTES
GT_ATTR	GET ATTRIBUTES	ST_KY	SET KEY
LST	LIST	ST_MKY	SET MASTER KEY
LST_COL	LIST COLLECTION	WRT	WRITE
OPNCLS	OPEN or CLOSE		

one bit for each operation

The USER OBJECT DESCRIPTOR TYPE field (see table 19) specifies the format of information that appears in the USER OBJECT DESCRIPTOR field.

Table 19 — User object descriptor types

User Object Descriptor Type	Description
0h	The USER OBJECT DESCRIPTOR field shall be ignored
1h	Single user object
2h - Fh	Reserved

single object

multi-object in future

Nonces – replay protection

Table 21 — Request nonce format

Bit / Byte	7	6	5	4	3	2	1	0
0	(MSB)							
5				TIMESTAMP				(LSB)
6	(MSB)							
11				RANDOM NUMBER				(LSB)

Prevent requests from being captured and re-sent

- at a later point
- overwrite file data
- denial of service

Nonce management

- rough window of time – can age old nonces
- must remember future nonces
- under attack – change key version; or “cut off” a host via audit nonce

Key hierarchy

Table 22 — OSD secret key hierarchy

Key Name	Key Shared Using	Key Used To	Key Update Frequency
Keys Shared Between the Security Manager and the OSD Device Server			
Master	SET MASTER KEY service action	Update Drive key	Change of OSD device owner
Drive	SET KEY service action	Update Partition key	When Partition key may have been compromised (i.e., very infrequently)
Partition [a]	SET KEY service action	Update Working keys	When Working key updates may have been compromised (i.e., infrequently)
Working [a]	SET KEY service action	Create Capability keys	When normal key use affords to much chance that the working key might be reverse engineered (i.e., regularly)
Keys Shared Between the Security Manager and the Application Client [c]			
Capability [d]	Credentials and mecha-nisms not specified in this standard	Secure commands, responses, and data	New with each new Credential

[d] As dual purpose number, the capability key is different from other keys in the hierarchy. The capability key is the credential integrity check value, meaning that even though the security manager computes its the computation is based on values beyond the security manager's control (e.g., the user object to which the credential allows access). The time interval during which the capability key is very short. While changing the working key used to construct the credential integrity check value invalidates the capability key, the credential may expire long before that and thus invalidate the capability key.

Backup Slides

OSD Status

- History
  - Started with NSIC NASD research 1995-1999
  - Carnegie Mellon, HP, IBM, Quantum, STK, Seagate
  - Seagate led NSIC OSD into SNIA in 1999
- Today
  - Intel & IBM leading SNIA OSD effort
  - EMC, HP, Panasas, Seagate, Veritas involved
  - IBM architecting objects into version 2 of StorageTank
  - Lustre – CFS/HP/BlueArc open-source OSD for DoE
    - 1,000 node; 225 TB cluster installed October 2002
  - Panasas shipping OSD-based products today
    - scalable NAS; large-scale systems (300+ devices)

OSD Commands

Table 29 — Commands for OSD devices (part 1 of 2)

Command name	Operation code	Service action [a]	Type	Reference
APPEND	7Fh	8807h	M	6.2
CHANGE ALIASES	A4h	0Bh	O	SPC-3
CREATE	7Fh	8802h	M	6.3
CREATE AND WRITE	7Fh	8812h	M	6.4
CREATE COLLECTION	7Fh	8815h	O	6.5
CREATE PARTITION	7Fh	880Bh	M	6.6
FLUSH OBJECT	7Fh	8808h	M	6.7
FORMAT OSD	7Fh	8801h	M	6.8
GET ATTRIBUTES	7Fh	880Eh	M	6.9
INQUIRY	12h		M	SPC-3
LIST	7Fh	8803h	M	6.10
LIST COLLECTION	7Fh	8817h	O	6.11
LOG SELECT	4Ch		O	SPC-3
LOG SENSE	4Dh		O	SPC-3
MODE SELECT(10)	55h		O	SPC-3
MODE SENSE(10)	5Ah		O	SPC-3

OSD Commands (2)

PERSISTENT RESERVE IN	5Eh		M	SPC-3
PERSISTENT RESERVE OUT	5Fh		M	SPC-3
PREVENT ALLOW MEDIUM REMOVAL	1Eh		O	SPC-3
READ	7Fh	8805h	M	6.12
READ BUFFER	3Ch		O	SPC-3
RECEIVE COPY RESULTS	84h		O	SPC-3
RECEIVE DIAGNOSTIC RESULTS	1Ch		O	SPC-3
REMOVE	7Fh	880Ah	M	6.13
REMOVE COLLECTION	7Fh	8816h	O	6.14
REMOVE PARTITION	7Fh	880Ch	M	6.15
REPORT ALIASES	A3h	0Bh	O	SPC-3
REPORT LUNS	A0h		O	SPC-3
REQUEST SENSE	03h		M	SPC-3

OSD Commands (3)

Table 29 — Commands for OSD devices (part 2 of 2)

Command name	Operation code	Service action [a]	Type	Reference
SEND DIAGNOSTIC	1Dh		M	SPC-3
SET ATTRIBUTES	7Fh	880Fh	M	6.16
SET KEY	7Fh	8818h	M	6.17
SET MASTER KEY	7Fh	8819h	M	6.18
START STOP UNIT	1Bh		O	SBC-2
TEST UNIT READY	00h		M	SPC-3
WRITE	7Fh	8806h	M	6.19
WRITE BUFFER	3Bh		O	SPC-3

Key: M = Command implementation is mandatory.
O = Command implementation is optional.
OB = Command implementation is defined in a previous standard

[a] No entry in the service action column means that the SERVICE ACTION field does not apply to the command. OSD service action codes not listed in this table (i.e., 8800h and 8810h-88FFh) are reserved for future standardization