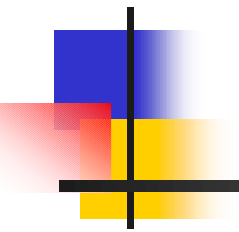
SCSI Device Memory Export Protocol (DMEP) T10 Presentation



Andrew Barry, Kenneth Preslan, Matthew O'Keefe
Sistina Software
1313 5th St. S.E. Suite 111 Minneapolis, MN 55414

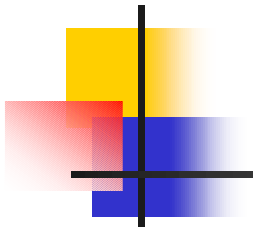
Gerry Johnson – Ciprico Inc

Burn Alting – Comptex Pty.Ltd

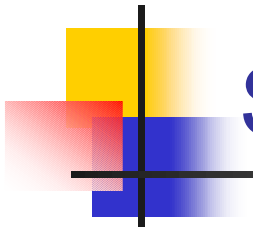
Jim Wayda – Dot Hill Systems Corp.

September 2000

Why Device Memory Export Protocol (DMEP)?



- Device Memory Export Protocol defines the implementation of a generic space of memory buffers on SCSI storage devices
- DMEP can be used to provide a Global Lock Space for a clustered file system
- In order to understand the need for DMEP, it is necessary to first have a basic understanding of Shared Disk File Systems…

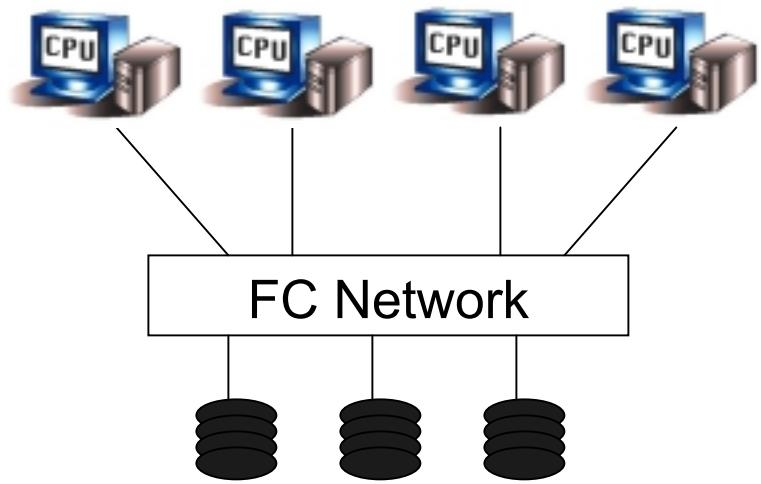


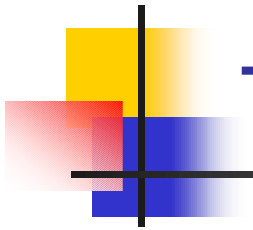
Shared Disk File Systems

- The real power of Storage Area Networks (SAN) is realized through shared disk file systems (SDFS).
- SDFS's allow multiple computers to access the same storage at the same time. This provides faster access and greater availability.
- SDFS's provide better performance than NFS file servers because there is no single server bottleneck.
- Data flows directly between each client and the storage devices, eliminating the need for a file server.

Symmetric Cluster Implementation

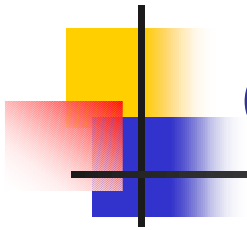
- Machines share disks containing data and metadata
- Metadata is managed by each machine as it is accessed
- Synchronization is achieved using global locks (DMEP or a Distributed Locking Manager (DLM))
- A local file system with inter-machine locking
- Example: DEC VAXcluster, Global File System (GFS)





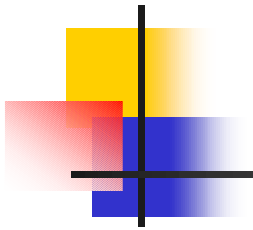
The Global File System

- The Global File System (GFS) is a shared disk file system.
- It is open source (GNU GPL)
- The October 2000 release of GFS will provide journaling support for reliability and fast file system recovery.
- 64-bit files and file system
- High performance
- Runs on Linux and is poised to be ported to other OS's like FreeBSD
- All SDFS's need two supporting technologies
  - storage consolidation
  - locks that can be used by all machines



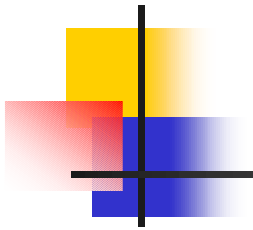
GFS Supporting Technology

- Storage consolidation is necessary if a file system is to be made on more than one device. This can be accomplished in software (GFS's own Pool driver, LVM, etc) or via hardware RAID.
- Global locks can be placed on storage devices to provide the file locking functionality necessary in a shared disk file system environment .
- The faster the lock operations and the more plentiful the locks, the better the SDFS performs



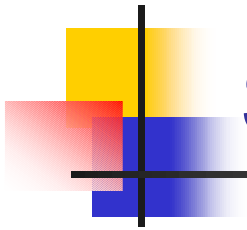
Global Locks On Storage Devices

- Disks:
  - No additional hardware
  - Disks are extremely reliable
  - Lock space automatically scales with storage space
  - Many disks, allow locks to be distributed
- RAID Controllers:
  - Extremely fast
  - RAID controllers have an abundance of memory to provide a large number of locks
  - Very large number of locks = improved file system performance
  - Global Locks can be redundant and fault tolerant (mirrored)



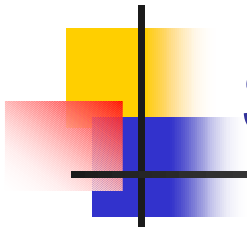
SCSI Memory Export Overview

- A simple generic protocol implemented within a SCSI device to provide cluster wide locking semantics
- Provides a generic space of memory buffers on SCSI storage devices
- Buffers are opaque to the storage device. The storage device does not care about the data in the memory buffers
- Buffers are readable and writeable by clients using a Load/Store conditional mechanism enforced by sequence numbers to ensure atomicity



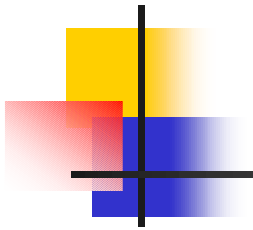
SCSI Memory Export Overview

- Buffers are identified with a 72-bit Logical Buffer Identifier that is mapped to a physical memory buffer on the SCSI device
- Buffers are implemented in a sparse space. Therefore there are many more possible Logical Buffer Identifiers that there are physical memory buffers
- Optionally supports multiple independent buffer spaces (Buffer Segments) on a single storage device. Useful to accommodate multiple GFS file systems or independent GFS clusters.
- Number of buffers and buffer size can be specified on a per buffer segment basis



SCSI Memory Export Overview

- Since the buffers are opaque, all of the complexities of locking are implemented in the initiator rather than in the SCSI storage device
- Future changes to locking semantics are implemented by modifying client software rather than SCSI storage device firmware

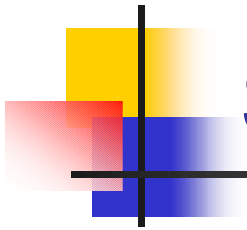


SCSI Memory Export Overview

- Not limited to cluster file system locking applications
- DMEP can be used as a generic memory sharing protocol between SCSI initiators

SCSI Memory Export Concepts

- Memory Export Buffers: The SCSI storage device provides buffers of memory that can be accessed by all initiators.
  - Buffer ID (BID): 9 Byte integer that is used to identify a buffer. The buffer ID is opaque to the device.
  - Physical Buffer Number: Index associated with each buffer that never changes. Buffer numbers range from 0 to N-1 where N is the number of buffers presented by the storage device.
  - Sparse Buffer Space: The Buffer ID is dynamically mapped to a physical buffer number when a load (read) operation is performed.

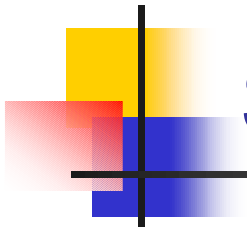


SCSI Memory Export Concepts

- Conditional Stores
  - Initiators perform atomic read-modify-write operations
  - A cluster member reads in a buffer from the storage device, modifies the buffer, and stores the buffer back to the storage device.
  - Two commands are required:
    - Load Command to read the data from the device
    - Store command to write the data to the device

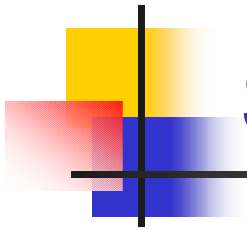
SCSI Memory Export Concepts

- Sequence number ensures atomicity of read-modify write operation
  - Load command returns a 64 bit sequence number and physical buffer number
  - Store command supplies the 64 bit sequence number and physical buffer number
  - If the sequence number and physical buffer number in the Store Command match the sequence number and physical buffer number of the memory buffer, the data is updated in the memory buffer on the storage device.
  - Else, a SCSI check condition is returned and the data is not updated.
  - The Sequence number is incremented after each successful store operation



SCSI Memory Export Concepts

- In-Use and Just-Created Buffers
  - Buffers that have been successfully stored are “In-Use”
  - A load command on a buffer that is not In-Use will cause the buffer to take on the characteristic of “Just-Created”.
  - The Buffer ID of a Just-Created buffer will map to a physical memory buffer

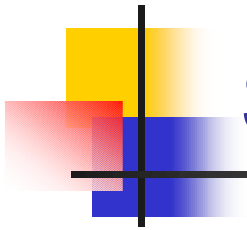


SCSI Memory Export Concepts

- Buffer Segments
  - Storage devices are required to implement at least one buffer segment
  - Storage devices may optionally implement up to 256 independent buffer spaces
- Full Buffer Space
  - Indicates with every load operation the fullness level of the buffer space
  - Fullness indication is on a per buffer segment basis

SCSI Memory Export Concepts

- Enable
  - If the storage device is powered off, the contents of memory buffers are lost
  - Load and store commands will fail with a check condition indicating that the segment is not configured/enabled.
  - Before a segment will accept load or store commands, the segment must be enabled with an “Enable” command.



SCSI Memory Export Concepts

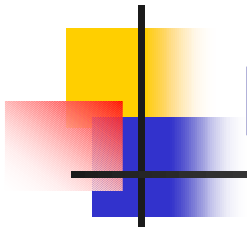
- Dump Buffers
  - A dump buffers command is provided to read a specified number of in-use buffers from the storage device starting at a specified physical buffer number.
  - Useful for file system recovery operations

SCSI Memory Export Concepts

- Select Configuration
  - Select Configuration provides for configuration of each buffer segment
  - Number of memory buffers and size of each memory buffer are configurable on a per segment basis
- Sense Configuration
  - Obtains configuration information for the specified buffer segment

DMEP Commands

- DMEP commands are sent to a SCSI Storage Device using a 16 byte SCSI CDB
- Two new SCSI Commands are defined using Service Actions:
  - Memory Export In (Proposed Operation Code 85h)
    - Reads data from the SCSI device
  - Memory Export Out (Proposed Operation Code 89h)
    - Writes data to the SCSI device



Memory Export In Service Actions

Code	Service Action	Description
0	LOAD BUFFER	Read the state of a single Memory Export buffer off of the target
1	DUMP BUFFERS	Read a large chunk of the Memory Export buffer state off of the target
2	SENSE CONFIG	Read a segment Memory Export Configuration off of the target
3-31	Reserved	

Table 1: MEMORY EXPORT IN Service Actions

Memory Export In CDB

Byte,Bit	7	6	5	4	3	2	1	0
0	Operation Code (85h)							
1	Reserved				Service Action			
2	Segment Number							
3	(MSB)							
4								
5								
6				Buffer Number				
7								
8								
9								
10								
11								(LSB)
12								
13				Allocation Length				
14								
15	Control							

Table 2: Memory Export In CDB

Memory Export In CDB Fields

- Operation Code – The proposed SCSI operation code for Memory Export In is 85h
- Service Action – Describes the Memory Export In function that is to be performed
- Segment Number – The segment number in which the memory export buffer resides
- Buffer Number – 72 bit buffer number on which to operate
  - For the Load Buffer service action this represents the logical buffer number
  - For the Dump Buffers service action, this field represents the 64 bit physical buffer number
- Allocation Length – Maximum number of bytes that the memory export device should return to the initiator

Load Buffer CDB Format

- Service Action – The Service Action for Load Buffer is 0
- Buffer Number – The ID of the buffer that is to be returned to the initiator
- Segment Number – The number of the segment in which the addressed buffer resides
- Allocation length – The maximum number of bytes to be returned.

Load Buffer Parameter Data Format

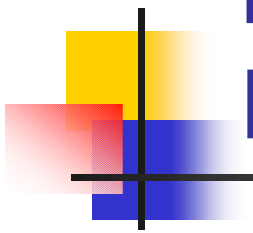
Byte,Bit	7	6	5	4	3	2	1	0
0	(MSB)							
1				Length (n)				
2								(LSB)
3	Reserved				Service Action (0)			
4	In Use	Reserved						
5	Fullness							
6				Reserved				
7								
8	(MSB)							
...				Sequence Number				
15								(LSB)
16	(MSB)							
...				Physical Buffer Number				
23								(LSB)
24								
...				Data				
n								

Table 3: Memory Export Load Parameter Data Format

Load Buffer Parameter Data Format Fields

- Length – The number of byte that are returned by a Load Buffer Action. (24 + segment data size)
- Service Action – The Service Action code for Load Buffer is 0
- In-Use – Indicates that the buffer is in use.
- Fullness – The amount of buffer space that is filled scaled to fit into an 8 bit quantity.
  - 0x00 = All buffer space is free
  - 0xFF = Buffer space is completely full

Load Buffer Parameter Data Format Fields



- Physical Buffer Number – Physical buffer number on the storage device in which the data is stored
- Sequence Number – This number is incremented after each successful store operation to the buffer. Ensures atomicity of read-modify-write operation
- Data – Actual data that is read from the memory buffer on the storage device

Dump Buffers CDB Format

- Service Action – The Service Action for Dump Buffers is 1
- Buffer Number – Starting physical buffer number at which the dump should start. (The first byte is ignored)
- Segment Number – The number of the segment in which the addressed buffer resides
- Allocation Length – The maximum number of bytes that the device should return.

Dump Buffers Parameter Data Format

Byte,Bit	7	6	5	4	3	2	1	0
0	Returned Byte Count							
2								
3	Reserved				Service Action (1)			
4	More	Reserved						
5	Reserved							
7								
8	Reserved							
10								
11	Buffer ID 1							
19								
20	Sequence Number 1							
27								
28	Physical Buffer Number 1							
35								
36	Data 1							
...								
n								
n+1	Reserved							
n+3								
n+4	Buffer ID 2							
n+12								
...	...							
...								
...	Data M							
...								
m*n								

Table 4: Dump Buffer PDF

Dump Buffers Parameter Data Format

- Returned Byte Count
  - Allocation length rounded down to match the size of the returned buffers
  - Could be far smaller than the allocation length if there are not enough In-Use buffers to fill the allocation length
- Service Action – The Service Action code for Dump Buffers is 1

Dump Buffers Parameter Data Format

- More – Indicates that there are more In-Use physical buffers beyond the last buffer returned by this Dump Buffers command.
- Buffer ID 1 – The Buffer ID of the first In-Use physical buffer returned
- Sequence Number 1 – The sequence number of the first physical buffer returned
- Physical Buffer Number – The Physical Buffer Number of the first physical buffer returned
- Data – The buffer data of the first physical buffer returned

Sense Configuration CDB Format

- Service Action – The service Action code for Sense Configuration is 2
- Buffer Number – Ignored
- Segment Number – The segment number for which configuration information should be returned
- Allocation Length – Maximum number of bytes that the device should return

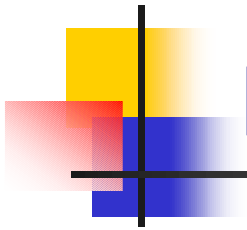
Sense Configuration Parameter Data Format

Byte,Bit	7	6	5	4	3	2	1	0
0	(MSB)							
1				Length				
2								(LSB)
3	Reserved				Action (2)			
4	Number of Segments							
5	Maximum Supported Segments							
6				Reserved				
7								
8	(MSB)							
...				Number of Buffers for Segment				
15								(LSB)
16	(MSB)							
17				Data Size for Segment				
18								(LSB)
19	Reserved							

Table 5: Memory Export Sense Configuration Parameter Data Format

Sense Configuration Parameter Data Format

- Length – Length of the Sense Configuration PDF
- Service Action – The Service Action for Sense Configuration is 2
- Number of Segments – The number of segments currently configured on the Memory Export Device
- Maximum supported Segments – The maximum number of segments supported by the Memory Export Device minus one
- Number of Buffers – The number of Memory Export Buffers in the segment identified by the Segment Number in the CDB
- Data Size – The length in bytes of each Memory Export Buffer identified by the Segment Number in the CDB.



Memory Export Out Service Actions

Code	Service Action	Description
0	STORE BUFFER	Change the state of a single Memory Export buffer on the target
1	Reserved	
2	SELECT CONFIG	Set the Memory Export Configuration of the target
3	ENABLE SEGMENT	Activate a Memory Export segment on the target
4-31	Reserved	

Table 6: MEMORY EXPORT OUT Service Actions

Memory Export Out CDB

Byte,Bit	7	6	5	4	3	2	1	0
0	Operation Code (89h)							
1	Reserved			Service Action				
2	Segment Number							
3	(MSB)							
4								
5								
6				Buffer Number				
7								
8								
9								
10								
11								(LSB)
12								
13				Parameter Length				
14								
15	Control							

Table 7: Memory Export Out CDB

Memory Export Out CDB Fields

- Operation Code – The proposed SCSI operation code for Memory Export In is 89h
- Service Action – Describes the Memory Export Out function that is to be performed
- Segment Number – The segment number in which the memory export buffer resides
- Buffer Number – 72 bit buffer number on which to operate
  - For the Store Buffer service action this represents the logical buffer number
- Parameter Length – The number of bytes that the initiator sends to the Memory Export Device

Store Buffer CDB Format

- Service Action – The Service Action for Store Buffer is 0
- Buffer Number – 72 Bit Buffer Number of which to operate
- Segment Number – The number of the segment in which the addressed buffer resides
- Parameter Length – The length in bytes of the data sent to the Memory Export Device
  - 24 + segment data size when a buffer is being stored
  - 24 when a buffer is being freed

Store Buffer Parameter Data Format

Byte,Bit	7	6	5	4	3	2	1	0
0	(MSB)							
1				Length (n)				
2								(LSB)
3	Reserved				Service Action (0)			
4	In Use	Reserved						
5				Reserved				
7								
8	(MSB)							
...				Sequence Number				
15								(LSB)
16	(MSB)							
...				Physical Buffer Number				
23								(LSB)
24								
...				Data				
n								

Table 8: Memory Export Store Parameter Format

Store Buffer Parameter Data Format

- Length – The number of Bytes sent in the PDF
- Service Action – The Service Action code for store buffer is 0
- In-Use – Indicates whether a buffer is in use. To de-allocate a buffer, issue a store command with the In-Use bit set to zero.
- Physical Buffer Number – The physical buffer number on the Storage Device in which the data is stored.
- Sequence Number – This number is incremented every time a buffer is stored
- Data – The actual data that is to be stored in the buffer on the Storage Device

  Note: The Sequence Number and Physical Buffer Number supplied in the PDF must match the Sequence number and Physical Buffer Number of the Memory Export Buffer for the store to succeed.

Select Configuration CDB Format

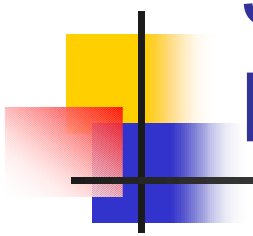
- Service Action – The Service Action for the Select Configuration Command is 2
- Buffer Number – Ignored
- Segment Number – The number of the segment for which the configuration information should be changed
- Parameter Length – The length in bytes of the parameter data

Select Configuration Parameter Data Format

Byte,Bit	7	6	5	4	3	2	1	0
0	(MSB)							
1				Length				
2								(LSB)
3	Reserved				Service Action (2)			
4								
5				Reserved				
6								
7				Reserved				
8	(MSB)							
...				Number of Buffers for Segment				
15								(LSB)
16	(MSB)							
17				Data Size for Segment				
18								(LSB)
19	Reserved							

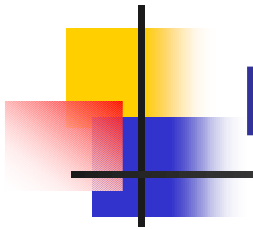
Table 9: Memory Export Select Parameter Data Format

Select Configuration Parameter Data Format



- Length – The number of bytes sent in the PDF
- Service Action – The Service Action for Select Configuration is 2
- Number of Buffers – The number of Memory Export Buffers in the segment identified by the Segment Number in the CDB
- Data Size – The length in bytes of each Memory Export Buffer in the segment identified by the Segment Number in the CDB

Enable Segment CDB Format



- Service Action – The Service Action for Enable Segment is 3
- Buffer Number – Ignored
- Segment Number – The number of the segment which is to be enabled
- Parameter Length – The length in bytes of the Parameter Data. Must be Zero

SCSI Check Conditions for Memory Export Out Commands

Sense Key	Additional Sense	ASCQ	Description
05h	04h	0Ah	Memory Export segment not enabled
05h	26h	10h	BID never loaded
0Eh	26h	0Eh	Sequence Number Error
0Eh	26h	0Fh	Buffer Number Error
06h	2Ah	06h	Memory Export Parameters Have Changed

Table 11: SCSI Check Conditions Sense Qualifiers for Memory Export Out commands

Questions?