An Overview of Version 0.9.5 Proposed SCSI Device Locks

Andrew Barry, Mike Tilstra, and Matthew O'Keefe
Parallel Computer Systems Laboratory, University of Minnesota
200 Union Street SE, Minneapolis, MN 55455
+1-612-626-7180, {barry, okeefe, tilstra}@borg.umn.edu
Kenneth Preslan - Sistina Software
Gerry Houlder and Jim Coomes - Seagate Technology
James Wayda - Dot Hill Systems Corp.

Abstract

Symmetric shared disks file systems are functional, versatile, and just plain cool. For these file systems to work, they require a global lock space, acessable to all clients. This paper describes a proposal for the implementation of the SCSI device lock command which provides a global lock space on storage devices. This paper details how Dlocks behave. This paper is a condensed overview of the 0.9.5 version of the proposed SCSI device locks specification. The full paper includes a thorough description of Dlock implementation, optional features of Dlocks, and a discussion of how Dlocks are used by the Global File System. For a postscript version of the full 0.9.5 Dlock specification, please visit our web site http://www.globalfilesystem.org . This paper is an evolution of the 0.9.4 Specification of SCSI Device Locks, which can also be found at the GFS web site.

1 Dlock Concepts

1.1 Expiration

In a shared disk environment, a failed client may not be allowed to indefinitely hold whatever locks it held when it failed. Therefore, each holder must continually update a timer on the disc. If this timer expires, other lock holders may begin error recovery functions to eventually free the lock. Expiration is alternately refered to as 'timing-out', and the act of updating the timer is often refered to as 'heartbeating'.

1.2 Client IDs

The Client ID is a unique identifier for each client. The client id is completely opaque to the Dlock device. In GFS the client ID is used both as an identifier and to store the IP address of the client, allowing inter-machine communication. The Client ID can be any arbitrary 32-bit number that uniquely identifies a machine.

1.3 Version Numbers

Associated with every lock is a version number. Whenever the data associated with a lock is changed, the version number is incremented. Clients may use cached data instead of re-reading from the disk as long as the version number on the dlock is unchanged since the data was read. This is reported by the lock device, so that the clients know not to use cached data.

1.4 Conversion Locks

The conversion lock is a simple- single stage queue used to prevent writer starvation. There is an awkward case where one client is trying to acquire an exclusive lock and can't because other clients are constantly acquiring and dropping the lock shared. If there is never a gap where no client is holding the lock shared, the

writer starves. To correct this, when a client unsuccessfully tries to acquire a lock, and no other client holds that lock's conversion, the conversion is granted to the unsuccessful client. Once the conversion is acquired, no other clients can acquire the lock. All the current holders eventually unlock, and the conversion holder can get the lock. All of a client's conversions are lost if the client expires.

1.5 Enable

In the event that a lock device is turned off, and comes back on, all the locks on the device could be lost. Though it would be nice if the locks were stored in some form of persistent storage, it is unreasonable to require it. Therefore, lock devices should not accept dlock commands when they are first powered up. The devices should return failure results, with the enabled bit of the dlock rdf cleared, to all dlock actions except refresh_timer, until the dlock action enable is issued to the drive. In this way, clients of the lock device are made aware that the locks on the lock device have been cleared, and can take action to deal with the situation.

2 The SCSI Dlock Command

The SCSI dlock command is a method of data synchronization. The Dlock interface is defined by three main parts: the CDB, the return data format, and the mode page.

2.1 The Dlock CDB

The Dlock command has a 16 byte Command Descriptor Block (CDB). It is shown in Table 1. Its fields are:

Operation Code The SCSI Operation code for Dlock is 83h.

Action This describes the action being requested. The possible values of this field are shown in Table 2.

Lock Number This is the number of the lock on which to operate.

Client ID An application-defined 32-bit number that identifies the client issuing the Dlock command.

Allocation Length The number of bytes that the initiator has allocated for data returned from the command. Note that the Allocation Length field can be used to control how much of the Reply Data is returned to the initiator. If the Allocation Length is too small for the amount of Reply Data that needs to be returned, only Allocation Length bytes of the Reply Data are returned. The Dlock command completes just as it would have if there was more space allocated. This is not an error.

2.2 Dlock Actions

The possible actions that each Dlock command performs are listed in Table 2.

Nop Return Holders Do not change the lock specified in the CDB, but report its state. The Client ID List in the reply data contains the Client IDs of the current non-expired holders of the specified Dlock.

Nop Return Expired Do not change the lock specified in the CDB, but report its state. The Client ID List in the reply data contains the IDs of the clients which expired while holding the specified Dlock.

Nop Return Conversion Do not change the lock specified in the CDB, but report its state. The Client ID List in the reply data contains the ID of the current holder of the Conversion lock for the specified Dlock.

Lock Shared Acquire the specified Dlock in the Shared state. If the the Client ID specified in the CDB does not hold the conversion for the lock, and there is a conversion holder, the action fails. If the lock is already held exclussively by another client, the action fails. If the action is unsuccessful in acquiring the Dlock, the conversion for the Dlock is acquired if available.

Byte,Bit	7	6	5	4	3	2	1	0
0	Operation Code (83h)							
1	Reserved			Action				
2	(MSB)							
3				Lock Number				
4								
5								(LSB)
6	(MSB)							
7				Client ID				
8								
9								(LSB)
10	(MSB)							
11				Allocation Length				
12								
13								(LSB)
14	Reserved							
15	Control							

Table 1: Dlock CDB

Code	Action	Description	Acts On	Client ID List
00h	Nop Return Holders	No change, return list of live holders	Lock Number	Holders
01h	Nop Return Expired	No change, return list of expired holders	Lock Number	Expired
02h	Nop Return Conversion	No change, return Client ID of conversion holder	Lock Number	Conversion
03h	Lock Shared	Acquire shared lock	Lock Number	Holders
04h	Lock Exclusive	Acquire exclusive lock	Lock Number	Holders
05h	Promote	Promote a shared lock to exclusive	Lock Number	Holders
06h	Unlock	Release lock	Lock Number	Holders
07h	Unlock Increment	Release lock and increment version number	Lock Number	Holders
08h	Demote	Demote an exclusive lock to shared	Lock Number	Holders
09h	Demote Increment	Demote an exclusive lock to shared and increment version number	Lock Number	Holders
0Ah	Refresh Timer	Refresh timer for Client	Client ID	None
0Bh	Reset Expired	Reset Expiration flags for a given Client	Client ID	None
0Ch	Report Expired	Report which Clients have expired	Whole Device	Expired
0Dh	Enable	Enable Dlock operation	Whole Device	None
0Eh	Drop Conversion	Removes the conversion on a lock	Lock Number	Holders
0Fh–1Fh	Reserved	Unused		

Table 2: Dlock Actions

Lock Exclusive Acquire the specified Dlock in the Exclusive state. If the lock is in the unlocked state and the Client ID in the CDB is the holder of the conversion (if any), the lock is acquired in the exclusive state. If other clients hold the lock, the action is unsuccessful. If the action is unsuccessful in acquiring the Dlockand the conversion for the Dlock is available, the conversion for the Dlock is acquired.

Promote Promote the specified Dlock from the shared state to the Exclusive state. If 1) the Dlock is in the shared state, 2) the Dlock is held by only the Client ID specified in the CDB, and 3) Client ID specified in the CDB is holder of the converion lock for this Dlock (if it is held at all), then the lock is promoted to the exclusive state. If the action is unsuccessful in acquiring the Dlock, the conversion lock for the Dlock is acquired if it is available.

Unlock Unlock the Dlock specified in the CDB. If the lock is held in a shared state and there are other holders, the state remains Lock Shared (but the holder count and Client ID list is changed).

Unlock Increment Behaves like the Unlock action, but also increments the Version Number for the specified Dlock.

Demote If 1) the specified Dlock is in the Lock Exclusive state and 2) the Dlock is held by the Client ID specified in the CDB, then lock is demoted to the Lock Shared state.

Demote Increment If 1) the specified Dlock is in the Lock Exclusive state and 2) the Dlock is held by the Client ID specified in the CDB, then lock is demoted to the Lock Shared state and the Version Number is incremented.

Refresh Timer This action heartbeats the device and prevents the specified Client ID (and all the locks it holds) from expiring.

Reset Expired Take the specified Client ID out of the Expired Lists of all Dlocks. The Client ID shouldn't show up in the results of any Nop (Expired) or Report Expired actions, unless the client expires again.

Report Expired Return a list of the IDs of all the clients which have expired while holding Dlocks on this device.

Enable This action allows the lock device to accept dlock commands. Until this command is issued to the drive, all commands fail. Enable is summarized more explicitly in the Dlock Concepts section.

Drop Conversion Remove any conversion on the Dlock specified in the CDB. This simplifies error recovery.

2.3 The Dlock Reply Data Format

The Reply Data Format for the Dlock command is shown it Table 3. Its parts are:

Version Number This is the version number of the lock.

Result This bit is one if the action succeeded and zero if the action failed.

Enabled This bit is zero when the lock device is powered on. It remains zero until the enable action is issued to the lock device.

List Type This field describes the type of clients returned in the Client ID list. The possible values of the field are *None*, *Holders*, *Expired*, and *Conversion*. The numerical representations of these values can be found in Table 5. The value returned in this field is dictated by the Action set in the Dlock CDB. Table 2 indicates which Actions result in which list types.

Have Conversion This bit is a one if the Client ID issuing the Dlock command possesses the conversion lock for this Dlock. It's zero otherwise.

Conversion This bit is a one if any client possesses the conversion lock for this Dlock.

State The values of the state of the lock are shown in Table 4.

Byte,Bit	7	6	5	4	3	2	1	0
0	(MSB)							
1				Version Number				
2								
3								(LSB)
4	Result	Enabled	List Type		Hv Conv	Conv	State	
5	Reserved							
6	(MSB)							
7				Number of Live Holders				(LSB)
8	(MSB)							
9				Number of Expired Holders				(LSB)
10	(MSB)							
11				Client ID List Length ($n - 11$)				(LSB)
List of Client IDs								
12	(MSB)							
13				Client ID (first)				
14								
15								(LSB)
...	(MSB)							
...				...				
...								
...								(LSB)
$n-3$	(MSB)							
$n-2$				Client ID (last)				
$n-1$								
n								(LSB)

Table 3: Dlock Reply Data Format

Code	Description
0h	Unlocked
1h	Locked Shared
2h	Locked Exclusive
3h	Reserved

Table 4: Values of the State field

Code	Name	Description
0h	None	No list is returned
1h	Holders	List of clients holding the lock is returned
2h	Expired	List of expired clients is returned
3h	Conversion	The ID of the client holding the conversion lock is returned

Table 5: Values of the List Type field

Byte,Bit	7	6	5	4	3	2	1	0
0	PS	Resvd	Page Code (29h)					
1	Page Length (0Ah)							
2	(MSB)	Maximum clients per lock						
3								(LSB)
4	(MSB)	Number of locks						
5								
6								
7								(LSB)
8	(MSB)	Client Timeout Interval (ms)						
9								
10								
11								(LSB)

Table 6: Mode Page Data

Number of Live Holders This is the number of un-expired clients currently holding this lock.

Number of Expired Holders This is the number of clients that expired when holding this lock.

Client ID List Length The length in bytes of the returned Client ID List.

Client ID List This is a list of Client IDs. The meaning of the list is specified by the List Type field (and the Dlock Action issued). If the value of the List Type field is *None*, this list is empty. If List Type is *Holders*, the list is made up of the current (un-expired) holders of the lock. If the List Type is *Expired*, the list is made up of the IDs of the clients which expired while holding this lock. (In the case of the "Report Expired" action, the list is made up of the IDs of all the clients which expired while holding any lock on this device.) If the List Type is *Conversion*, the list contains the ID of the client (if there is one) who holds the Conversion lock for this Dlock.

The number of IDs in the list is zero (for List Type *None*), "Number of Live Holders" (for List Type *Holders*), "Number of Expired Holders" (for List Type *Expired*), or zero or one depending on whether or not there is a holder of the conversion lock (for List Type *Conversion*). If the Allocation Length specified in the CDB is too small to hold the whole Client ID list, as much as possible is returned (with the "Client ID List Length" set to the appropriate value). This is not an error.

2.4 Mode Page

The Mode Page returns configuration information about several lock parameters. The page is shown in Table 6.

Maximum number of clients able to share a lock This is a 16-bit field, storing how many clients can simultaneously hold a lock in the lock shared state.

Number of locks on the device Returns the number of Dlocks on the device. If this value is hexidecimal 0xffffffff, the device supports a sparse lock space.

Client Timeout Interval The number of milliseconds after which a client ID will timeout. If the value is zero, client IDs never time out.

All dlock implementations must support changeable Client Timeout Intervals as part of the Mode Select Command. Implementers may optionally implement changeable Max Clients Per Lock and Number of Locks fields. Dlock devices can also support optional sparse lock spaces. Any Mode Select Command which sets one or more of the parameters in this mode page must also clear all locks on the device, and reset the enable bit.

2.5 State Transition Diagrams

The state transition diagrams are shown in Figure 1 and Figure 2.

Figure 1 This figure details the possible state changes possible due to normal lock operations. The conversion states are entered when a lock shared operation fails because the lock is held exclusively by another client, or when a lock exclusive operation fails because the lock is held by another client. If a conversion lock is held (i.e., the lock is in one of the conversion states), most actions fail that would normally succeed, except when issued by the conversion holder. The exceptions are unlock and demote.

Figure 2 This figure demonstrates the state transitions that occur when a client expires. These are not operations issued by a client, but are the actions performed by the lock device when a client ID fails to refresh it's timer within the timeout interval. Each arrow in this diagram represents a client timing out and therefore being removed from the holder list of the lock, and being placed in the expired holder list for the lock. Expired clients also lose any lock conversions.

References

[1] X3T10 SCSI committee. Document T10/ 98-225R1 – Proposed SCSI Device Locks. http:// ftp.symbios.com/ ftp/ pub/ standards/ io/ x3t10/ document.98/ 98-225r1.pdf, October 1998.

[2] Roy G. Davis. *VAXCluster Principles*. Digital Press, 1993.

[3] Andrew Barry et al. *Proposed SCSI Device Locks Version 0.9.5*. University of Minnesota, Parallel Computer Systems Laboratory, http:// www.globalfilesystem.org/ pubs/ dlock-0.9.5.ps, 1999.

[4] Kenneth Preslan et al. A 64-bit, shared disk file system for linux. In *The Sixteenth IEEE Mass Storage Systems Symposium held jointly with the Seventh NASA Goddard Conference on Mass Storage Systems & Technologies*, San Diego, California, March 1999.

[5] Kenneth Preslan et al. *Proposed SCSI Device Locks Version 0.9.4*. University of Minnesota, Parallel Computer Systems Laboratory, http:// www.globalfilesystem.org/ pubs/ dlock-0.9.4.ps, 1999.

[6] Kenneth Preslan, Steven Soltis, Christopher Sabol, and Mat thew O'Keefe. Device locks: Mutual exclusion for storage area networks. In *The Seventh NASA Goddard Conference on Mass Storage Systems and Technologies in cooperation with the Sixteenth IEEE Symposium on Mass Storage Systems*, San Diego, CA, March 1999.

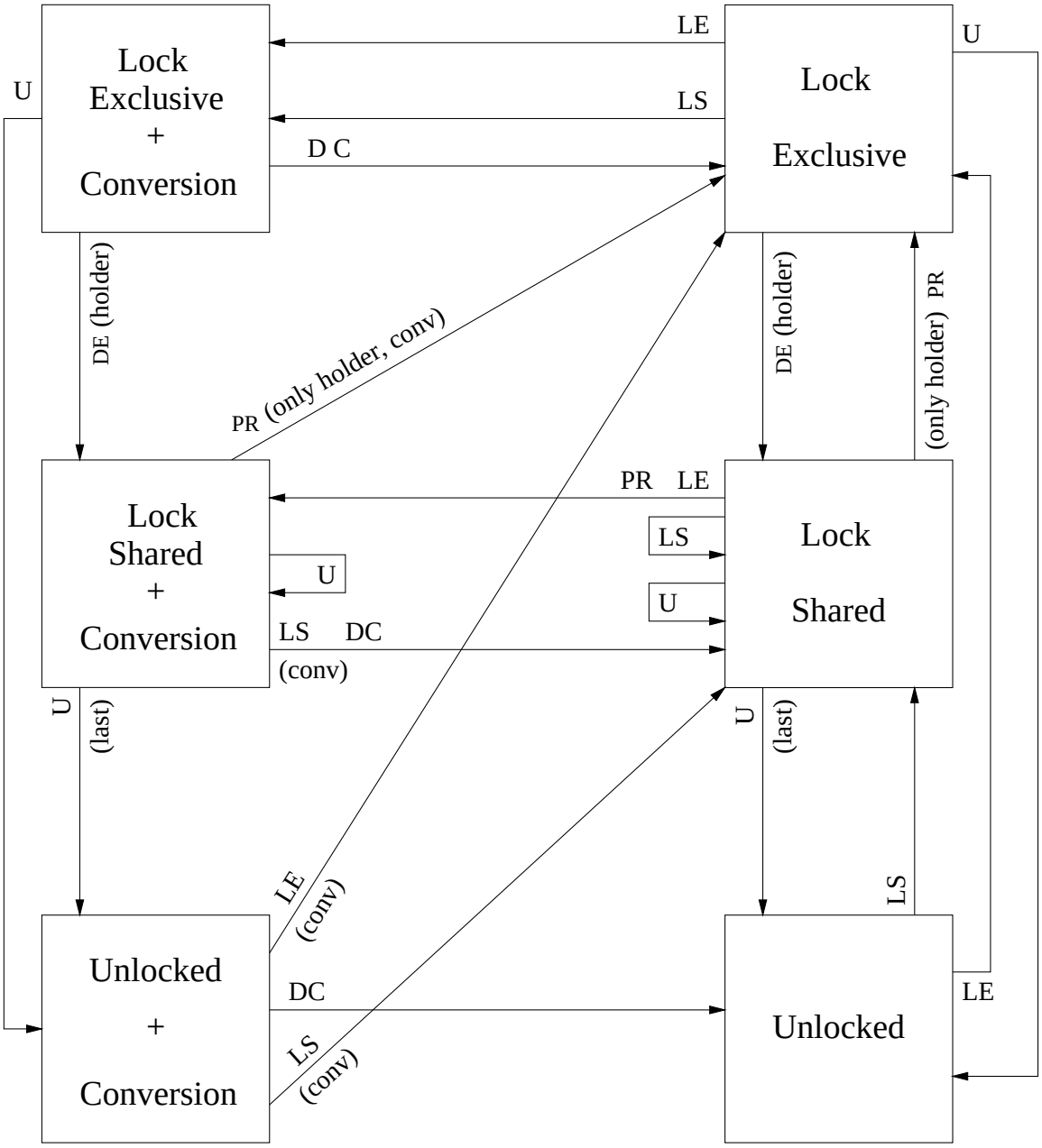

Figure 1: A state transition diagram that describes the possible transitions due to Lock Shared, Lock Exclusive, Unlock, Promote, Demote, and Drop Conversion actions. Unlock Increment has the same transistions as Unlock and Demote Increment has the same transitions as Demote.

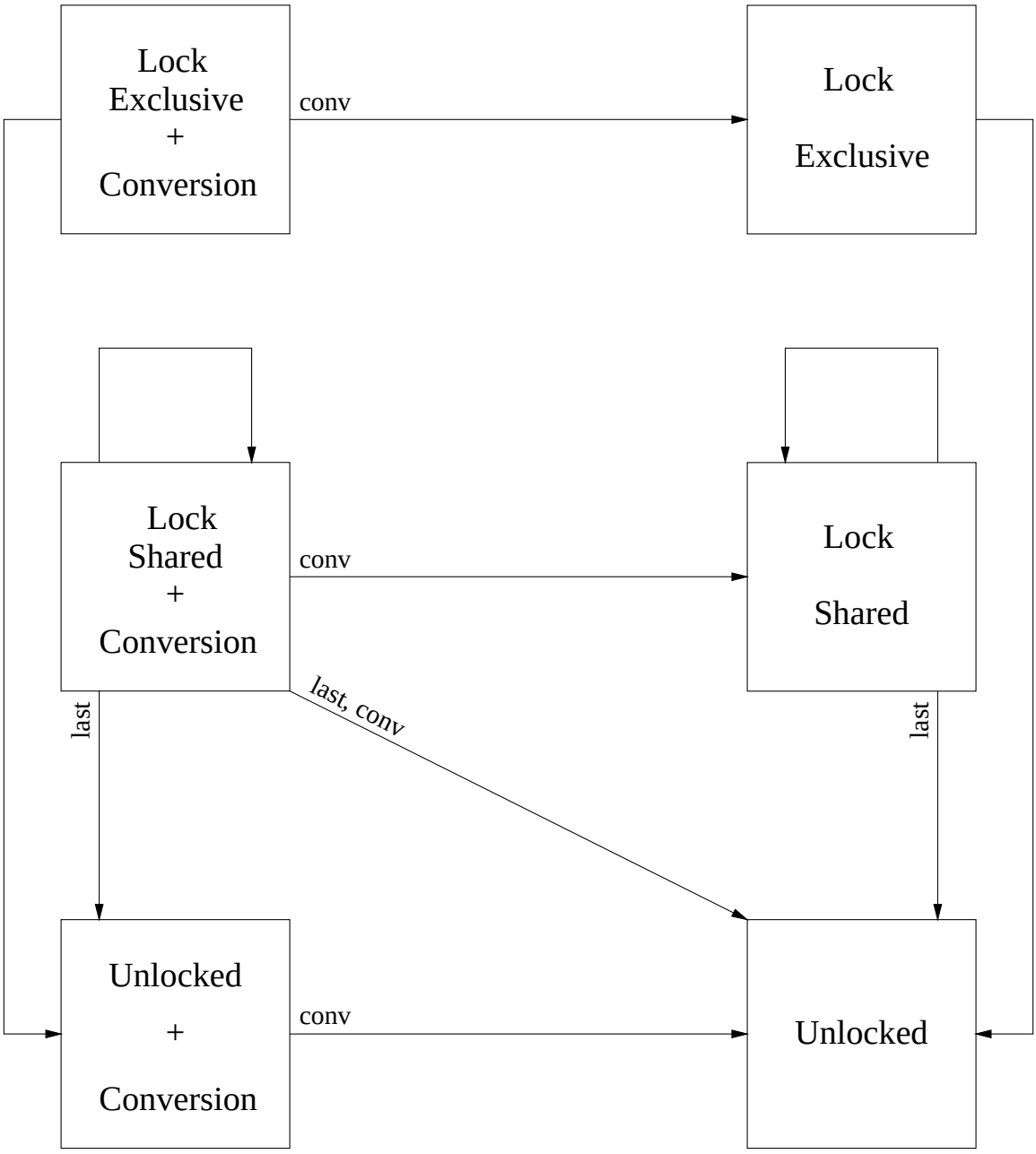

Figure 2: A state transition diagram that describes the possible transitions due to expirations.