Date: March 11, 2004

To: T10 Committee (SCSI)

From: George Penokie (IBM/Tivoli), David Chambliss (IBM Almaden Research Center)

Subject: SAM-3 Per-Command Priority Tagging

1 Overview

The following proposed wording represents changes to SAM-3 to enable the transmission of priority information on a per-command basis.

This proposal standardizes the interface by which device servers can offer differentiated quality of service to different applications associated with the same initiator. Examples of its use would include offering lower priority on IO associated with background destage writes within a storage controller or on IO associated with background applications, so that response time may be reduced for those IO operations that directly affect the responsiveness offered to the end user.

The method defined in this proposal to accomplish this involves changes to the protocol standards to accommodate an extension to the task attribute field to allow different priorities to be assigned to simple task attributes.

2 Additions to SAM-3

2.1 The Execute Command procedure call

An application client requests the processing of a SCSI command by invoking the SCSI transport protocol services described in 2.4, the collective operation of which is conceptually modeled in the following procedure call:

Service Response = Execute Command (IN (I_T_L_Q Nexus, CDB, Task Attribute, [Data-In Buffer Size], [Data-Out Buffer], [Data-Out Buffer Size], [Command Reference Number], [Priority]),
OUT ([Data-In Buffer], [Sense Data], [Sense Data Length], Status))

Input Arguments:

I_T_L_Q Nexus: The I_T_L_Q nexus identifying the task (see 4.12).

CDB: Command descriptor block (see 5.2).

Task Attribute: A value specifying one of the task attributes defined in 8.6. SCSI transport protocols may or may not provide the ability to specify a different task attribute for each task (see 8.6.1). For a task that processes linked commands, the Task Attribute shall be that specified for the first command in the sequence of linked commands. The Task Attribute specified for the second and subsequent commands shall be ignored.

Data-In Buffer Size: The number of bytes available for data transfers to the Data-In Buffer (see 5.4.3).

Data-Out Buffer: A buffer containing command specific information to be sent to the logical unit, such as data or parameter lists needed to process the command. The buffer size is indicated by the Data-Out Buffer Size argument. The content of the Data-Out Buffer shall not change during the lifetime of the command (see 5.5) as viewed by the application client.

Data-Out Buffer Size: The number of bytes available for data transfers from the Data-Out Buffer (see 5.4.3).

Command Reference Number (CRN): When this argument is used, all sequential commands of an I_T_L nexus shall include a CRN argument that is incremented by one. The CRN shall be set to one for each I_T_L nexus involving the SCSI port after the SCSI port receives a hard reset or detects I_T nexus loss. The CRN shall be set to one after it reaches the maximum CRN value supported by the protocol. The CRN value zero shall be reserved for use as defined by the SCSI transport protocol. It is not an error for the application client to provide this argument when CRN is not supported by the SCSI transport protocol or logical unit.

Priority: The priority assigned to the task. For specific requirements on the Priority argument see 3.2

Output Arguments:

Data-In Buffer: A buffer to contain command specific information returned by the logical unit by the time of command completion. The **Execute Command** procedure call shall not return a status of GOOD, CONDITION MET, INTERMEDIATE, or INTERMEDIATE-CONDITION MET unless the buffer contents are valid. The application client shall not assume that the buffer contents are valid unless the command completes with a status of GOOD, CONDITION MET, INTERMEDIATE, or INTERMEDIATE-CONDITION MET. While some valid data may be present for other values of status, the application client should rely on additional information from the logical unit, such as sense data, to determine the state of the buffer contents. If the command ends with a service response of SERVICE DELIVERY OR TARGET FAILURE, the application client shall consider this argument to be undefined.

Sense Data: A buffer containing sense data returned in the same I_T_L_Q nexus transaction (see 3.1.46) as a CHECK CONDITION status (see 5.9.6). The buffer length is indicated by the Sense Data Length argument. If the command ends with a service response of SERVICE DELIVERY OR TARGET FAILURE, the application client shall consider this argument to be undefined.

Sense Data Length: The length in bytes of the Sense Data.

Status: A one-byte field containing command completion status (see 5.3). If the command ends with a service response of SERVICE DELIVERY OR TARGET FAILURE, the application client shall consider this argument to be undefined.

Service Response assumes one of the following values:

TASK COMPLETE: A logical unit response indicating that the task has ended. The Status argument shall have one of the values specified in 5.3 other than INTERMEDIATE or INTERMEDIATE-CONDITION MET.

LINKED COMMAND COMPLETE: Logical unit responses indicating that the task has not ended and that a linked command has completed successfully. As specified in 5.3, the Status argument shall have a value of INTERMEDIATE or INTERMEDIATE-CONDITION MET.

SERVICE DELIVERY OR TARGET FAILURE: The command has been ended due to a service delivery failure (see 3.1.113) or SCSI target device malfunction. All output parameters are invalid.

2.2 Command descriptor block (CDB)

2.3 Status

2.4 SCSI transport protocol services in support of Execute Command

2.4.1 Overview

The SCSI transport protocol services that support the **Execute Command** procedure call are described in 2.4. Two groups of SCSI transport protocol services are described. The SCSI transport protocol services that support the request and confirmation for the **Execute Command** procedure call are described in 2.4.2. The SCSI transport protocol services that support the data transfers associated with processing a SCSI command are described in 5.4.3.

2.4.2 Execute Command request/confirmation SCSI transport protocol services

All SCSI transport protocol standards shall define the SCSI transport protocol specific requirements for implementing the **Send SCSI Command** SCSI transport protocol service request and the **Command Complete Received** confirmation. Support for the **SCSI Command Received** indication and **Send Command Complete** response by a SCSI transport protocol standard is optional. All SCSI I/O systems shall implement these SCSI transport protocols as defined in the applicable SCSI transport protocol specification.

SCSI Transport Protocol Service Request:

Send SCSI Command (IN (I_T_L_Q Nexus, CDB, Task Attribute, [Data-In Buffer Size], [Data-Out Buffer], [Data-Out Buffer Size], [Command Reference Number], [Priority], [First Burst Enabled]))

Input Arguments:

I_T_L_Q Nexus: The I_T_L_Q nexus identifying the task (see 4.12).

CDB: Command descriptor block (see 5.2).

Task Attribute: A value specifying one of the task attributes defined in 8.6. For specific requirements on the Task Attribute argument see 2.1.

Data-In Buffer Size: The number of bytes available for data transfers to the Data-In Buffer (see 5.4.3).

Data-Out Buffer: A buffer containing command specific information to be sent to the logical unit, such as data or parameter lists needed to process the command (see 2.1). The content of the Data-Out Buffer shall not change during the lifetime of the command (see 5.5) as viewed by the application client.

Data-Out Buffer Size: The number of bytes available for data transfers from the Data-Out Buffer (see 5.4.3).

Command Reference Number (CRN): When this argument is used, all sequential commands of an I_T_L nexus shall include a CRN argument that is incremented by one (see 2.1).

Priority: The priority assigned to the task. For specific requirements on the Priority argument see 3.2

First Burst Enabled: An argument specifying that a SCSI transport protocol specific number of bytes from the Data-Out Buffer shall be delivered to the logical unit without waiting for the device server to invoke the **Receive Data-Out** SCSI transport protocol service.

SCSI Transport Protocol Service Indication:

SCSI Command Received (IN (I_T_L_Q Nexus, CDB, Task Attribute, [Command Reference Number], [Priority], [First Burst Enabled]))

Input Arguments:

I_T_L_Q Nexus: The I_T_L_Q nexus identifying the task (see 4.12).

CDB: Command descriptor block (see 5.2).

Task Attribute: A value specifying one of the task attributes defined in 8.6. For specific requirements on the Task Attribute argument see 2.1.

Command Reference Number (CRN): When this argument is used, all sequential commands of an I_T_L nexus shall include a CRN argument that is incremented by one (see 2.1).

Priority: The priority assigned to the task. For specific requirements on the Priority argument see 3.2

First Burst Enabled: An argument specifying that a SCSI transport protocol specific number of bytes from the Data-Out Buffer are being delivered to the logical unit without waiting for the device server to invoke the **Receive Data-Out** SCSI transport protocol service.

3 Task Set Management

3.1 Introduction to task set management

Clause 3 describes some of the controls application clients have over task set management behaviors (see 8.3). Clause 3 also specifies task set management requirements in terms of:

a) Priority (see 3.2)
b) Task states (see 8.5);
a) Task attributes (see 8.6);
b) The events that cause transitions between task states (see 8.4 and 8.5); and
c) A map of task state transitions (see 8.7).

Clause 3 concludes with several task set management examples (see 8.8).

Task behavior, as specified in clause 3, refers to the functioning of a task as observed by an application client, including the results of command processing and interactions with other tasks.

The requirements for task set management only apply to a task after it has been entered into a task set. A task shall be entered into a task set unless:

a) A condition exists that causes that task to be completed with a status of BUSY, RESERVATION CONFLICT, TASK SET FULL, or ACA ACTIVE;
b) Detection of an overlapped command (see 5.9.3) causes that task to be completed with a CHECK CONDITION status; or
c) SCSI transport protocol specific errors cause that task to be completed with a status other than GOOD.

3.2 Priority

A priority set to a value other than zero specifies the relative scheduling importance of a task having a SIMPLE task attribute in relation to other tasks already in the task set. Priority 1h is the highest priority, with increasing priority values indicating lower scheduling importance.

If the priority is set to zero, or is not contained within the SCSI transport protocol service indication, a priority code assigned to the I_T_L nexus may be used by the task manager to determine an ordering to process tasks with the SIMPLE task attribute in addition to its vendor specific ordering rules. A priority may be assigned to an I_T_L nexus by a SET PRIORITY command (see SPC-3) or by the INITIAL PRIORITY field in the Control Extension mode page (see SPC-3). If no priority has been assigned to the I_T_L nexus using the SET PRIORITY command (see SPC-3) and the logical unit does not support the INITIAL PRIORITY field of the Control Extension mode page the priority assigned to the task is vendor specific.

A difference in priority between tasks does not necessarily override other scheduling considerations (e.g., different times to access different logical block addresses). However, processing of a collection of tasks with different priorities should cause the subset of tasks with the higher priority to return status sooner in aggregate than the same subset would if the same collection of tasks were submitted under the same conditions but with all priorities equal.

For a task that processes linked commands, the priority shall be that specified for the first command in the sequence of linked commands. The priority specified for the second and subsequent commands shall be ignored.

The size of the PRIORITY field shall be four bits.