OCP Fast Fail @ home in Command Duration Limits

Ralph Weber & Robert Horn
Western Digital

24 January 2019
OCP Storage Workshop

Command Duration Limits
Goal

- ✓ **Provide Better Utilization of HDDs**

  by

- ✓ **Eliminating the *Fear of Queuing***

Fear of Queuing

- ✓ Queuing Maximizes IOPS Performance
  - ✗ Disk Develops a Mind of Its Own
    IOPS become the only answer
  - ✗ No Go for Hosts With Varied Clients
    thus, the *Fear of Queuing*
- ✓ Must Reign-in Over-Zealous Disks
  - ✓ Protocol to Send Host Cares to Disk
    - ✚ Workable for Both Ends
    - ✚ Useful in Many Situations
  - ✓ Development Started in 2016, but …
    … Learning Curve is Steep

❀☆ Presenting ☆❀

OCP Fast Fail for SCSI SAS Devices

✓ Fast Fail proposal (based on OCP spec)

- ➡ appended to this PDF – this group only
- ➡ no confidential information disclosed to T10

  T10/18-089r3 is the T10 document

✺ Outline

- ☆ Quickie tour of three SCSI standards
- ☆ Detour to Log Counters
- ☆ Fast Fail Mode Parameters (the beef at last)
- ☆ Q&A

CDBs to Command Duration

From CDB

Duration Limits Descriptor
DLD (3 bits)
in selected commands

values 1 to 7

SAM-6

From a Mode Page

Command Duration Limits Descriptor

value 0
means
no command duration limits

- ✓ Fast Fail Inactive Time‡
- ✓ Fast Fail Active Time‡

‡ Command Duration Limits Descriptors contain other parameters, but these two are interesting to this group.

Date: 22 January 2019
To: OCP Storage Workshop (& T10)
From: Ralph O. Weber and Robert Horn
Subject: SPC-6, SBC-4, SAM-6: More Specific Command Duration Limits

Introduction

The OCP (Open Compute Project) is on the verge of publishing *Cloud HDD - Fast Fail Read*. The concepts in this document intend that read commands which cannot be completed in a timely fashion be terminated as soon as the inevitability of excessive time usage is detected. Proper operation of this concept depends on capabilities that are outside the scope of SCSI (i.e., the idea that multiple read commands are sent to multiple devices with the intention that the first command to return data completes the requested read and the results from the other read commands are ignored).

In addition, the *"Host-Managed Read Retries"* section of the *"Disks for Data Centers"* white paper from the Fast 2016 Conference describes command latency management mechanisms that are more flexible than those currently defined for SCSI and which have the added advantage of independence from fields that can only be transferred in transport protocol specific information units. Although the latency approach described in this proposal is not included in the proposed solutions, in September 2016, the SNIA published a white paper titled *"Hyperscaler Storage"* described how already-standardized features could be used to accomplish several of the goals described in the "*Disks for Data Centers*" white paper.

This proposal intends to add all these concepts to SPC-6, with minimal corollary changes in SAM-6 and SBC-4.

Revision History

r0 Initial revision
r1 Incorporated comments from September CAP working group, took steps to better align this proposal with an anticipated equivalent proposal in T13, and incorporated changes suggested by discussions with other interested parties.
r2 Incorporated comments from November CAP working group. Added a bit to the RWCDLUNIT field (renamed to T2CDLUNITS in r3) and changed its position in the read write command duration limit descriptor. Added more subclause structure to the SAM-6 *Command duration limit* subclause. Made other, less significant wording changes.
r3 Incorporated changes discussed by the January CAP working group. Updated to latest posted working drafts for SPC-5 and SBC-4. Some subclause and table numbering changes not highlighted with change bars.

Unless otherwise indicated additions are shown in underlined blue, deletions in ~~red strikethrough~~, and comments in green. Differences between this revision and the previous revision are highlighted with change bars.

SPC-5 r20a Model Clause Changes Being Proposed for SPC-6

5.2 Command duration limits

An application client uses a command duration limit to specify the scheduling and processing of a duration limited command (see SAM-5). If a device server determines that it is unable to complete processing of a duration limited

18484 Preston Road, Suite 102 PMB 178, Dallas, TX 75252 214-912-1373 ralph.weber@wdc.com

command before expiration of the duration expiration time, then the device server may terminate the processing of that command (see SAM-5).

Device servers that support command duration limits shall:

a) support one or more commands that are capable of specifying a duration limit descriptor index ~~value~~ set to a nonzero value (see SBC-4);
b) for each command that is capable of specifying a command duration limit descriptor index ~~value~~ set to a nonzero value, report that capability ~~a nonzero value~~ in the RWCDLP bit and the CDLP field (see 6.34.2); and
c) indicate support in the RWCDLP bit and the CDLP field for: ~~the Command Duration Limit A mode page (see 7.5.9) or the Command Duration Limit B mode page (see 7.5.10) as indicated by the CDLP field.~~
   A) the Command Duration Limit A mode page (see 7.5.9);
   B) the Command Duration Limit B mode page (see 7.5.10);
   C) the Command Duration Limit T2A mode page (see 7.5.mpa); or
   D) the Command Duration Limit T2B mode page (see 7.5.mpb).

~~A nonzero value in the CDLP field indicates whether the Command Duration Limit A mode page or the Command Duration Limit B mode page specifies the command duration limit applicable to that command. A duration limit descriptor value set to a nonzero value in the CDB specifies which duration limit descriptor in the indicated mode page specifies the command duration limit.~~

To determine the command duration limit information that applies to a command whose CDB contains a non-zero duration limit descriptor index, the device server uses:

1) the values in the RWCDLP bit and in the CDLP field in the parameter data returned by the REPORT SUPPORTED OPERATION CODES command (see 6.34) for this command to determine which command duration limit mode page, if any, applies the CDB; and
2) the command duration limit descriptor specified by the duration limit descriptor index in the CDB.

EXAMPLE - A duration limit descriptor index of 001b selects the command duration limit information contained in the first command duration limit descriptor. A duration limit descriptor index of 010b selects the command duration limit information contained in the second command limit duration descriptor. A duration limit descriptor index of 111b selects the command duration limit information contained in the seventh command limit duration descriptor.

…

SPC-5 r20a Mode Page Changes Being Proposed for SPC-6

7.5 Mode parameters

7.5.1 Summary of mode page codes

…

Table 440 — Summary of mode page codes

Mode page name	Page code	Subpage code	Reference
Command Duration Limit A	0Ah	03h	7.5.9
Command Duration Limit B	0Ah	04h	7.5.10
Command Duration Limit T2A	0Ah	<<TBD ¿07h?>>	7.5.mpa
Command Duration Limit T2B	0Ah	<<TBD ¿08h?>>	7.5.mpb
…	…	…	…
<<<No other changes proposed in table 440.>>>			

…

7.5.9 Command Duration Limit A mode page

The Command Duration Limit A mode page (see table 443) provides controls for command duration limit (see SAM-5) that are applicable to all device types, for commands for which the REPORT SUPPORTED OPERATION CODES command parameter data RWCDLP bit and CDLP field (see 6.34) indicate ~~indicates~~ the Command Duration Limit A mode page. The mode page policy (see 7.5.2) for this mode page should be per I_T nexus. The mode page policy may be shared. If a field in this mode page is changed while there is a command already in the task set, then the new value of the field shall not apply to that command.

…

The command duration limit descriptor (see table 444) describes the command duration limit corresponding to the duration limit descriptor index ~~value~~ in the CDB if the Command Duration Limit A mode page is indicated (see 5.2) ~~(see appropriate command standard)~~.

<<<No other changes proposed in 7.5.9.>>>

…

7.5.10 Command Duration Limit B mode page

The Command Duration Limit B mode page (see table 446) provides controls for command duration limit (see SAM-5) that are applicable to all device types, for commands for which the REPORT SUPPORTED OPERATION CODES command parameter data RWCDLP bit and CDLP field (see 6.34) indicate ~~indicates~~ the Command Duration Limit B mode page. The mode page policy (see 7.5.2) for this mode page should be per I_T nexus. The mode page policy may be shared. If a field in this mode page is changed while there is a command already in the task set, then the new value of the field shall not apply to that command.

…

The command duration limit descriptor (see table 444) describes the command duration limit corresponding to the duration limit descriptor index ~~value as defined in 7.5.9~~ in the CDB if the Command Duration Limit B mode page is indicated (see 5.2).

<<<No other changes proposed in 7.5.10.>>>

…

7.5.mpa Command Duration Limit T2A mode page

<<<All of 7.5.mpa is new. Use of modification markups suspended.>>>

7.5.mpa.1 Overview

The Command Duration Limit T2A mode page (see table x1) provides controls for command duration limit (see SAM-5) that are applicable to all device types, for commands for which the REPORT SUPPORTED OPERATION CODES command parameter data RWCDLP bit and CDLP field (see 6.34) indicate the Command Duration Limit T2A mode page. The mode page policy (see 7.5.2) for this mode page should be per I_T nexus. The mode page policy may be shared. If a field in this mode page is changed while there is a command already in the task set, then the new value of the field shall not apply to that command.

Table x1 — Command Duration Limit T2A mode page format

Bit / Byte	7	6	5	4	3	2	1	0
0	PS	SPF (1b)	PAGE CODE (0Ah)					
1	SUBPAGE CODE (TBD) <<<requesting 07h>>>							
2	(MSB)	PAGE LENGTH (00E4h)						
3								(LSB)
4	Reserved							
•••								
6								
7	Reserved				PERFORMANCE VERSUS LATENCY CONTROLS			
	T2 command duration limit descriptor list							
8	T2 command duration limit descriptor [first]							
•••								
39								
40	T2 command duration limit descriptor [second]							
•••								
71								
	⋮							
200	T2 command duration limit descriptor [seventh]							
•••								
231								

The PS bit, SPF bit, PAGE CODE field, SUBPAGE CODE field, and PAGE LENGTH field are described in 7.5.7.

The SPF bit, PAGE CODE field, SUBPAGE CODE field, and PAGE LENGTH field shall be set as shown in table x1 for the Command Duration Limit T2A mode page.

The PERFORMANCE VERSUS LATENCY CONTROLS field (see table x2) specifies how much overall performance (e.g., total reads and writes per second) is allowed to be effected by read write latency controls (i.e., the combination of the REQUESTED LATENCY TARGET fields, PREDICTIVE MISS POLICY fields, LATENCY MISS POLICY fields, NONCONFORMANCE POLICY fields, and CONFORMANCE TARGET fields in every T2 command duration limit descriptor (see 7.5.mpa.2) in the Command Duration Limit T2A mode page (see 7.5.mpa) and every T2 command duration limit descriptor in the Command Duration Limit T2B mode page (see 7.5.mpb)).

Table x2 — PERFORMANCE VERSUS LATENCY CONTROLS **field**

Code	Description
0	Read write latency controls have top priority, regardless of their effects on overall performance.
1 to 14	Overall performance may be reduced by up to adjustment percent in order to improve the effectiveness of read write latency controls. where: adjustment = (3 × code) + 50 and code = the value in the PERFORMANCE VERSUS LATENCY CONTROLS field
15	Overall performance has top priority, regardless of its effect on read write latency controls.

The T2 command duration limit descriptor (see 7.5.mpa.2) describes the command duration limit information that corresponds to the duration limit descriptor index in the CDB if the Command Duration Limit T2A mode page is indicated (see 5.2).

7.5.mpa.2 T2 command duration limit descriptor

The T2 command duration limit descriptor (see table x3) describes the command duration limit information that corresponds to one duration limit descriptor index.

Table x3 — T2 command duration limit descriptor format

Bit / Byte	7	6	5	4	3	2	1	0
0	Reserved				T2CDLUNITS			
1	Reserved							
2	(MSB)							
3	FAST FAIL INACTIVE TIME							(LSB)
4	(MSB)							
5	FAST FAIL ACTIVE TIME							(LSB)
6	FAST FAIL INACTIVE TIME POLICY				FAST FAIL ACTIVE TIME POLICY			
7	Reserved							
8	(MSB)							
9	REQUESTED LATENCY TARGET							(LSB)
10	(MSB)							
11	Reserved							(LSB)
12	PREDICTIVE MISS POLICY				LATENCY MISS POLICY			
13	NONCONFORMANCE POLICY				CONFORMANCE TARGET			
14								
•••	Reserved							
31								

The T2CDLUNITS field (see table x4) specifies the time units for the FAST FAIL INACTIVE TIME field, the FAST FAIL ACTIVE TIME field, the REQUESTED LATENCY TARGET field, and the ACHIEVABLE LATENCY TARGET field. The default value for the T2CDLUNITS field is the smallest value that the device server allows for time units. If a MODE SELECT command attempts to set a T2CDLUNITS field to a time unit value that is less than the minimum that the device supports, then the command may be terminated with CHECK CONDITION status, with the sense key set to ILLEGAL REQUEST, and the additional sense code set to INVALID FIELD IN PARAMETER LIST.

Table x4 — T2CDLUNITS field

Code	Description
0h	No value specified
6h	500 nanoseconds
8h	1 microsecond
Ah	10 milliseconds
Eh	500 milliseconds
all others	Reserved

The FAST FAIL INACTIVE TIME field specifies an upper limit on the time that elapses before a command becomes an enabled command (i.e., the interval between the time at which the SCSI Command Received transport protocol service indication is invoked (see SAM-6) to the time at which the command becomes an enabled command). A FAST FAIL INACTIVE TIME field set to a non-zero value specifies the time upper limit in units indicated by the T2CDLUNITS field. A FAST FAIL INACTIVE TIME field set to zero specifies that no time upper limit is specified by this T2 command duration limit descriptor. If the T2CDLUNITS field is set to 0000b, the FAST FAIL INACTIVE TIME field shall be ignored.

The FAST FAIL ACTIVE TIME field specifies an upper limit on the time that elapses during the processing of a command (i.e., the interval between the time at which a command first becomes an enabled command to the time the device server returns status for the command). An FAST FAIL ACTIVE TIME field set to a non-zero value specifies the time upper limit in units specified by the T2CDLUNITS field. An FAST FAIL ACTIVE TIME field set to zero specifies that no time upper limit is specified by this T2 command duration limit descriptor. If the T2CDLUNITS field is set to 0000b, the FAST FAIL ACTIVE TIME field shall be ignored.

The FAST FAIL INACTIVE TIME POLICY field (see table x5) specifies the policy action taken if the fast fail inactive limit is not met (i.e., the time used to cause a command to become an enabled command exceeds the time specified by the FAST FAIL INACTIVE TIME field and the T2CDLUNITS field).

The FAST FAIL ACTIVE TIME POLICY field (see table x5) specifies the policy action taken if the fast fail active time limit is not met (i.e., the time used to process a command exceeds the time specified by the FAST FAIL ACTIVE TIME field and the T2CDLUNITS field).

Table x5 — Fast fail policy actions

Code	Description
0h	The device server shall complete the command at the earliest possible time (i.e, do nothing based on the fast fail time limit not being met).
1h to Dh	Reserved
Eh	The device server shall terminate the command with CHECK CONDITION status, with the sense key set to ABORTED COMMAND and the additional sense code set to COMMAND TIMEOUT DURING PROCESSING or COMMAND TIMEOUT DURING PROCESSING DUE TO ERROR RECOVERY. If the starting LBA contents have been transferred to the application client, then the device server may indicate the largest LBA for which a contiguous range of LBAs have been transferred to the application client starting with the starting LBA of the read command as described in SBC-4.
Fh	The device server shall terminate the command with CHECK CONDITION status, with the sense key set to ABORTED COMMAND and the additional sense code set to COMMAND TIMEOUT BEFORE PROCESSING.

The REQUESTED LATENCY TARGET field specifies a total time for command completion (i.e., the interval between the time at which the SCSI Command Received transport protocol service indication is invoked (see SAM-6) to the time at which the device server returns status for the command). A REQUESTED LATENCY TARGET field set to a non-zero value specifies a requested total time in units indicated by the T2CDLUNITS field. A REQUESTED LATENCY TARGET field set to zero specifies that no requested total time is specified by this T2 command duration limit descriptor. If the T2CDLUNITS field is set to 0000b, the REQUESTED LATENCY TARGET field shall be ignored.

A REQUESTED LATENCY TARGET field set to FFFFh shall specify that the requested latency target is the sum of the contents of the FAST FAIL ACTIVE TIME field and the FAST FAIL INACTIVE TIME field. If a MODE SELECT command specifies a requested latency target that is greater than the sum of the fast fail active time and the fast fail inactive

time, then the device server may use parameter rounding (see 5.10) to set the REQUESTED LATENCY TARGET field to sum of the contents of the FAST FAIL ACTIVE TIME field and the FAST FAIL INACTIVE TIME field.

The time specified by the REQUESTED LATENCY TARGET field and the T2CDLUNITS field (i.e., the requested latency target time) affects the processing of the PREDICTIVE MISS POLICY field, the LATENCY MISS POLICY field, and the NONCONFORMANCE POLICY field.

The PREDICTIVE POLICY field specifies the requested latency policy action (see table x6) that the device server shall apply to a command if predictive methods conclude that the requested latency target time is going to be exceeded in the future.

The LATENCY MISS POLICY field specifies the requested latency policy action (see table x6) that the device server shall apply to a command that the device server determines has:

a) exceeded the requested latency target time; and
b) not exceeded the conformance target specified by the CONFORMANCE TARGET field.

The NONCONFORMANCE POLICY field specifies the requested latency policy action (see table x6) that the device server shall apply to a command that the device server determines has exceeded:

a) the requested latency target time; and
b) the conformance target specified by the CONFORMANCE TARGET field.

Table x6 — Requested latency policy actions

Code	Description
0h	The device server shall complete the command at the earliest possible time (i.e, do nothing based on the fast fail time limit not being met).
1h	The device server shall add one to the duration limit descriptor index in the CDB and process the command using the T2 command duration limit descriptor, if any, indicated by the modified duration limit descriptor index. If the modified duration limit descriptor index is greater than seven or the T2 command duration limit descriptor indicated by the modified duration limit descriptor index does not define a nonzero requested latency target time, then the device server shall process the command using a duration limit descriptor index of 2h.
2h	The device server shall process the command using a duration limit descriptor index of 0h (i.e., no T2 command duration limit descriptor specified), and the device server shall not allow the command to become an enabled command until all of the commands in the task set have a duration limit descriptor index of 0h.
3h to Dh	Reserved

Table x6 — Requested latency policy actions

Code	Description
Eh	The device server shall terminate the command with CHECK CONDITION status, with the sense key set to ABORTED COMMAND and the additional sense code set to COMMAND TIMEOUT DURING PROCESSING or COMMAND TIMEOUT DURING PROCESSING DUE TO ERROR RECOVERY. If the starting LBA contents have been transferred to the application client, then the device server may indicate the largest LBA for which a contiguous range of LBAs have been transferred to the application client starting with the starting LBA of the read command as described in SBC-4.
Fh	The device server shall terminate the command with CHECK CONDITION status, with the sense key set to ABORTED COMMAND and the additional sense code set to COMMAND TIMEOUT BEFORE PROCESSING.

The CONFORMANCE TARGET field (see table x7) specifies the percentage of the total number of commands having a specific duration limit descriptor index in the CDB that shall be counted as having been processed based on the LATENCY MISS POLICY field before command processing based on the NONCONFORMANCE POLICY field begins. After the NONCONFORMANCE POLICY field has been applied to one command for a given duration limit descriptor index in the CDB, the device server shall apply NONCONFORMANCE POLICY field to all commands with the same duration limit descriptor index in the CDB a logical unit reset is processed.

Table x7 — CONFORMANCE TARGET field

Code	Description
0h	The first command that is affected by the LATENCY MISS POLICY field shall be processed based on the NONCONFORMANCE POLICY field instead.
1h	10%
2h	1%
3h	0.1%
4h	0.01%
all others	Reserved

7.5.mpb Command Duration Limit T2B mode page

<<<All of 7.5.mpb is new. Use of modification markups suspended.>>>

The Command Duration Limit T2B mode page (see table x8) provides controls for command duration limit (see SAM-5) that are applicable to all device types, for commands for which the REPORT SUPPORTED OPERATION CODES command parameter data RWCDLP bit and CDLP field (see 6.34) indicate the Command Duration Limit T2B mode page. The mode page policy (see 7.5.2) for this mode page should be per I_T nexus. The mode page policy may be shared. If a field in this mode page is changed while there is a command already in the task set, then the new value of the field shall not apply to that command.

If the Command Duration Limit T2B mode page is supported, the Command Duration Limit T2A mode page (see 7.5.mpa) shall be supported.

Table x8 — Command Duration Limit T2B mode page format

Bit / Byte	7	6	5	4	3	2	1	0
0	PS	SPF (1b)	PAGE CODE (0Ah)					
1	SUBPAGE CODE (TBD) <<<requesting 08h>>>							
2	(MSB)	PAGE LENGTH (00E4h)						
3								(LSB)
4	Reserved							
•••								
7								
	T2 command duration limit descriptor list							
8	T2 command duration limit descriptor [first]							
•••								
39								
40	T2 command duration limit descriptor [second]							
•••								
71								
	⋮							
200	T2 command duration limit descriptor [seventh]							
•••								
231								

The PS bit, SPF bit, PAGE CODE field, SUBPAGE CODE field, and PAGE LENGTH field are described in 7.5.7.

The SPF bit, PAGE CODE field, SUBPAGE CODE field, and PAGE LENGTH field shall be set as shown in table x8 for the Command Duration Limit T2B mode page.

The T2 command duration limit descriptor (see 7.5.mpa.2) describes the command duration limit information that corresponds to the duration limit descriptor index in the CDB if the Command Duration Limit T2B mode page is indicated (see 5.2).

SPC-5 r20a Log Page Changes Being Proposed for SPC-6

7.3 Log parameters

7.3.1 Summary of log page codes

…

Table 320 — Summary of log page codes

Log page name	Page code	Subpage code	Reference
…	…	…	…
Cache Memory Statistics	19h	20h	7.3.6
Command Duration Limits Statistics	19h	<<TBD ¿21h?>>	7.3.cdl
…	…	…	…
<<<No other changes proposed in table 320.>>>			

…

7.3.cdl Command Duration Limits Statistics log page

<<<All of 7.3.cdl is new. Use of modification markups suspended.>>>

7.3.cdl.1 Overview

Using the format shown in table x10, the Command Duration Limits Statistics log page contains log parameters (see table x9) that indicate the effects of the Command Duration Limit T2A mode page (see 7.5.mpa) and the Command Duration Limit T2B mode page (see 7.5.mpb).

Table x9 — Command Duration Limits Statistics log page parameter codes

Parameter code	Description	Resettable or Changeable [a]	Reference	Support
0001h	Achievable Latency Target	Never	7.3.cdl.2	Mandatory
0011h to 0017h	Command Duration Limit T2A (1 to 7)	Reset Only	7.3.cdl.3	see [b]
0021h to 0027h	Command Duration Limit T2B (1 to 7)	Reset Only	7.3.cdl.3	see [c]
all others	Reserved			

[a] The keywords in this column – Always, Reset Only, and Never – are defined in 7.3.3.
[b] If the Command Duration Limit T2A mode page (see 7.5.mpa) is supported, these parameter codes shall be supported.
[c] If the Command Duration Limit T2B mode page (see 7.5.mpb) is supported, these parameter codes shall be supported.

The Command Duration Limits Statistics log page has the format shown in table x10.

Table x10 — Command Duration Limits Statistics log page

Bit / Byte	7	6	5	4	3	2	1	0
0	DS	SPF (1b)	PAGE CODE (19h)					
1	SUBPAGE CODE (<<<TBD ¿21h?>>>)							
2	(MSB)	PAGE LENGTH (n-3)						
3								(LSB)
	Command duration limits statistics log parameters							
4	Command duration limits statistics log parameter (see table x9) [first]							
•••								
	⋮							
	Command duration limits statistics log parameter (see table x9) [last]							
•••								
n								

The DS bit, SPF bit, PAGE CODE field, SUBPAGE CODE field, and PAGE LENGTH field are described in 7.3.2. The SPF bit, PAGE CODE field, and SUBPAGE CODE field shall be set as shown in table x10 for the Cache Memory Statistics log page.

The contents of each cache memory statistics log parameter depends on the value in its PARAMETER CODE field (see table x9).

7.3.cdl.2 Achievable Latency Target log parameter

The Achievable Latency Target log parameter has the format shown in table x11.

Table x11 — Achievable Latency Target log parameter

Bit / Byte	7	6	5	4	3	2	1	0
0	(MSB)	PARAMETER CODE (0001h)						
1								(LSB)
2	Parameter control byte – unbounded data counter log parameter (see 7.3.2.2.2.3)							
	DU	Obsolete	TSD (1b)	Obsolete			FORMAT AND LINKING	
3	PARAMETER LENGTH (04h)							
4	(MSB)	ACHIEVABLE LATENCY TARGET						
•••								
7								(LSB)

The PARAMETER CODE field is described in 7.3.2.2.1, and shall be set as shown in table x11 for the Achievable Latency Target log parameter.

The DU bit, TSD bit, and FORMAT AND LINKING field are described in 7.3.2.2.2.1. The DU bit and FORMAT AND LINKING field shall be set as described for an unbounded data counter log parameter (see 7.3.2.2.2.3) for the Achievable Latency Target parameter. The TSD bit shall be set as shown in table x11 for the Achievable Latency Target log parameter.

The PARAMETER LENGTH field is described in 7.3.2.2.1, and shall be set as shown in table x11 for the Achievable Latency Target parameter.

At the time the LOG SENSE command (see 6.9) is processed, the ACHIEVABLE LATENCY TARGET field indicates the lowest achievable value for any REQUESTED LATENCY TARGET field (see 7.5.mpa.2) in the Command Duration Limit T2A mode page (see 7.5.mpa) and the Command Duration Limit T2B mode page (see 7.5.mpb). The contents of the ACHIEVABLE LATENCY TARGET field are measured in microseconds (i.e., the value contained in the REQUESTED LATENCY TARGET field if the T2CDLUNITS field (see 7.5.mpa.2) is set to 100b).

Factors that affect the ACHIEVABLE LATENCY TARGET field include:

a) the contents of the PERFORMANCE VERSUS LATENCY CONTROLS field (see 7.5.mpa.1);
b) the contents of all non-zero REQUESTED LATENCY TARGET fields and their associated:
   A) PREDICTIVE MISS POLICY fields;
   B) LATENCY MISS POLICY fields;
   C) NONCONFORMANCE POLICY fields; and
   D) CONFORMANCE TARGET fields;
c) the commands being processed by the device server;
d) data transfer errors, if any, processed by the device server; and
e) the background activities, if any, being performed by the device server.

The ACHIEVABLE LATENCY TARGET field may be set to zero, if:

a) all REQUESTED LATENCY TARGET fields are set to zero; or
b) the number data transfer commands processed by the device server are not sufficient to determine an achievable latency target.

7.3.cdl.3 Command Duration Limits log parameter

The Command Duration Limits log parameter has the format shown in table x12.

Table x12 — Command Duration Limits log parameter

Bit / Byte	7	6	5	4	3	2	1	0
0	(MSB)							
1	PARAMETER CODE (see table x9)							(LSB)
2	Parameter control byte – unbounded data counter log parameter (see 7.3.2.2.2.3)							
	DU	Obsolete	TSD (1b)	Obsolete			FORMAT AND LINKING	
3	PARAMETER LENGTH (20h)							
4	(MSB)							
•••	NUMBER OF INACTIVE TARGET MISS COMMANDS							
7								(LSB)
8	(MSB)							
•••	NUMBER OF ACTIVE TARGET MISS COMMANDS							
11								(LSB)
12	(MSB)							
•••	NUMBER OF LATENCY MISS COMMANDS							
15								(LSB)
16	(MSB)							
•••	NUMBER OF NONCONFORMING COMMANDS							
19								(LSB)
20	(MSB)							
•••	NUMBER OF PREDICTIVE LATENCY MISS COMMANDS							
23								(LSB)
24	(MSB)							
•••	NUMBER OF LATENCY MISSES ATTRIBUTABLE TO ERRORS							
27								(LSB)
28	(MSB)							
•••	NUMBER OF LATENCY MISSES ATTRIBUTABLE TO DEFERRED ERRORS							
31								(LSB)
32	(MSB)							
•••	NUMBER OF LATENCY MISSES ATTRIBUTABLE TO BACKGROUND OPERATIONS							
35								(LSB)

The PARAMETER CODE field is described in 7.3.2.2.1, and shall be set as shown in table x12 for the Command Duration Limits log parameter.

The DU bit, TSD bit, and FORMAT AND LINKING field are described in 7.3.2.2.2.1. The DU bit and FORMAT AND LINKING field shall be set as described for an unbounded data counter log parameter (see 7.3.2.2.2.3) for the Command Duration Limits parameter. The TSD bit shall be set as shown in table x12 for the Command Duration Limits log parameter.

If the device server processes a LOG SELECT command (see 6.8) with the TSD bit set to zero in a Command Duration Limits log parameter, the command may be terminated with CHECK CONDITION status, with the sense key set to ILLEGAL REQUEST, and the additional sense code set to INVALID FIELD IN PARAMETER LIST.

The PARAMETER LENGTH field is described in 7.3.2.2.1, and shall be set as shown in table x11 for the Command Duration Limits parameter.

The NUMBER OF INACTIVE TARGET MISS COMMANDS field indicates the number of commands for which the FAST FAIL INACTIVE TIME POLICY field (see 7.5.mpa.2) was processed.

The NUMBER OF ACTIVE TARGET MISS COMMANDS field indicates the number of commands for which the FAST FAIL ACTIVE TIME POLICY field (see 7.5.mpa.2) was processed.

The NUMBER OF LATENCY MISS COMMANDS field indicates the number of commands for which the LATENCY MISS POLICY field (see 7.5.mpa.2) was processed.

The NUMBER OF NONCONFORMING COMMANDS field indicates the number of commands for which the NONCONFORMANCE POLICY field (see 7.5.mpa.2) was processed.

The NUMBER OF PREDICTIVE LATENCY MISS COMMANDS field indicates the number of commands for which the PREDICTIVE MISS POLICY field (see 7.5.mpa.2) was processed.

The NUMBER OF LATENCY MISSES ATTRIBUTABLE TO ERRORS field indicates the number of commands for which the processing of the LATENCY MISS POLICY field was determined to be the result of data transfer error processing for that command.

The NUMBER OF LATENCY MISSES ATTRIBUTABLE TO DEFERRED ERRORS field indicates the number of commands for which the processing of the LATENCY MISS POLICY field was determined to be the result of data transfer error processing for a different command.

The NUMBER OF LATENCY MISSES ATTRIBUTABLE TO BACKGROUND OPERATIONS field indicates the number of commands for which the processing of the LATENCY MISS POLICY field was determined to be the result of background operation processing.

SPC-5 r20a … REPORT SUPPORTED OPCODES Changes Being Proposed for SPC-6

6.34 REPORT SUPPORTED OPERATION CODES command

…

6.34.2 All_commands parameter data format

…

Each command descriptor (see table 260) contains information about a single supported command CDB.

Table 260 — Command descriptor format

Bit / Byte	7	6	5	4	3	2	1	0
0	OPERATION CODE							
1	Reserved							
2	(MSB)			SERVICE ACTION				
3								(LSB)
4	Reserved							
5	Reserved	RWCDLP	MLU		CDLP		CTDP	SERVACTV
6	(MSB)			CDB LENGTH				
7								(LSB)
8				Command timeouts descriptor, if any (see 6.34.4)				
•••								
19								

The OPERATION CODE field indicates the operation code of a command supported by the logical unit.

The SERVICE ACTION field indicates a supported service action of the supported operation code indicated by the OPERATION CODE field. If the operation code indicated in the OPERATION CODE field does not have any service actions, the SERVICE ACTION field shall be set to 00h.

The multiple logical units (MLU) field is described in table 261.

Table 261 — MLU field description

<<<No changes are proposed in table 261.>>>

The read write command duration limits page (RWCDLP) bit and the command duration limit page (CDLP) field (see table 262) indicate ~~indicates~~ the mode page, if any, that specifies the command duration limit for the command.

Table 262 — RWCDLP bit and CDLP field

RWCDLP bit	~~Code~~ CDLP field	Description
0b	00b	No command duration limit mode page is indicated for this command
1b	00b	Reserved
0b [a]	01b [a]	Command Duration Limit A mode page
0b [b]	10b [b]	Command Duration Limit B mode page
1b [c]	01b [c]	Command Duration Limit T2A mode page
1b [d]	10b [d]	Command Duration Limit T2B mode page
0b or 1b	11b	Reserved

[a] If ~~this value is~~ these values are returned, the Command Duration Limit A mode page (see 7.5.8) shall be supported.

[b] If ~~this value is~~ these values are returned, the Command Duration Limit B mode page (see 7.5.9) shall be supported.

[c] If these values are returned, the Command Duration Limit T2A mode page (see 7.5.mpa) shall be supported.

[d] If these values are returned, the Command Duration Limit T2B mode page (see 7.5.mpb) shall be supported.

A command timeouts descriptor present (CTDP) bit set to one indicates that the command timeouts descriptor (see 6.34.4) is included in this command descriptor. A CTDP bit set to zero indicates that the command timeouts descriptor is not included in this command descriptor.

A service action valid (SERVACTV) bit set to zero indicates the operation code indicated by the OPERATION CODE field does not have service actions and the SERVICE ACTION field contents are reserved. A SERVACTV bit set to one indicates the operation code indicated by the OPERATION CODE field has service actions and the contents of the SERVICE ACTION field are valid.

The CDB LENGTH field indicates the length of the command CDB in bytes for the operation code indicated in the OPERATION CODE field, and if the SERVACTV bit is set to one the service action indicated by the SERVICE ACTION field.

The command timeouts descriptor is described in 6.34.4.

6.34.3 One_command parameter data format

The REPORT SUPPORTED OPERATION CODES one_command parameter data format (see table 263) contains information about the CDB and a usage map for bits in the CDB for the command specified by the REPORTING OPTIONS field, REQUESTED OPERATION CODE field, and REQUESTED SERVICE ACTION field in the REPORT SUPPORTED OPERATION CODES CDB.

Table 263 — One_command parameter data

Bit Byte	7	6	5	4	3	2	1	0
0	Reserved							RWCDLP
1	CTDP	MLU		CDLP		SUPPORT		
2	(MSB)							
3		CDB SIZE (n-3)						(LSB)
4								
•••		CDB USAGE DATA						
n								
n+1								
•••		Command timeouts descriptor, if any (see 6.34.4)						
n+12								

A command timeouts descriptor present (CTDP) bit set to one indicates that the command timeouts descriptor (see 6.34.4) is included in the parameter data. A CTDP bit set to zero indicates that the command timeouts descriptor is not included in the parameter data.

The multiple logical units (MLU) field is described in table 261.

The read write command duration limits page (RWCDLP) bit and the command duration limit page (CDLP) field (see table 262) indicate ~~indicates~~ the mode page, if any, that specifies the command duration limit for the command.

The SUPPORT field is defined in table 259.

<<<No other changes are proposed in 6.34.3.>>>

SAM-6 r04 Changes Being Proposed

8.6 Command duration limit

8.6.1 Command duration limit overview

A command duration limit specifies the scheduling and processing of a duration limited command. A duration limited command is a command that:

a) has a SIMPLE task attribute or an ORDERED task attribute;
b) reports a nonzero value in the CDLP field of the REPORT SUPPORTED OPERATION CODES all_commands parameter data or the REPORT SUPPORTED OPERATION CODES one_command parameter data (see SPC-5); and

c) defines a duration limit descriptor index that contains three bits where a nonzero index selects ~~a nonzero~~ command duration limit information in:
   A) the Command Duration Limit A mode page (see SPC-5); ~~or~~
   B) the Command Duration Limit B mode page (see SPC-5);~~.~~
   C) the Command Duration Limit T2A mode page (see SPC-6); or
   D) the Command Duration Limit T2B mode page (see SPC-5).

If the command duration limit mode page selection results from the RWCDLP bit being:

a) set to zero (i.e., the Command Duration Limit A mode page or the Command Duration Limit B mode page is selected), then commands with the SIMPLE task attribute or ORDERED task attribute are processed as described in 8.6.2; and
b) set to one (i.e., the Command Duration Limit T2A mode page or the Command Duration Limit T2B mode page is selected), then commands with the SIMPLE task attribute are processed as described in 8.6.3.

<<<The following black text is to be moved without changes from the location indicated below.>>>
<<<Any change markups shown in the following text represents changes from the text that is being moved.>>>

8.6.2 Command duration scheduling

If the CDLP field is not set to zero and the RWCDLP bit is set to zero, then the processing of duration limited commands is described in this subclause.

If the device server has requests to process duration limited commands with an ORDERED task attribute, then the device server shall process the command as described in 8.7.3.2.

~~A command duration limit provides information~~ The selected command duration limit information (see 8.6.1) is used as part of scheduling of a duration limited command with a SIMPLE task attribute in relation to other commands having SIMPLE task attributes that are already in the task set.

The device server may use the duration expiration time to determine the order of processing commands with the SIMPLE task attribute within the task set. A difference in duration expiration time between commands may override other scheduling considerations (e.g., different times to access different logical block addresses or vendor specific scheduling considerations). Processing of a collection of commands with different command duration limit settings should cause a command with an earlier duration expiration time to complete with status sooner than a command with a later duration expiration time.

<<<The text being moved ends here.>>>

If the device server has requests to process duration limited commands with a SIMPLE task attribute, then the device server should process the commands based on the command:

1) duration scheduling as defined in this subclause ~~8.6.2~~; and
2) priority scheduling as defined in 8.5.

~~If the device server has requests to process duration limited commands with an ORDERED task attribute, then the device server shall process the command as described in 8.7.3.2.~~

If the device server determines that it is unable to complete a duration limited command before the duration expiration time, then the device server shall terminate the command as described in 8.7.3.2.

The duration expiration time is calculated as follows:

duration expiration time = command arrival time + command duration limit

where:

command arrival time	is the time at which the SCSI Command Received transport protocol service indication (see 5.4.2.3) is invoked; and
command duration limit	is as specified in the duration limit descriptor bits in the CDB (see appropriate command standard).

<<<The following strikeout text is the source for the text that is shown above as being moved.>>>

~~**8.6.2 Command duration scheduling**~~

~~A command duration limit provides information used as part of scheduling of a duration limited command with a SIMPLE task attribute in relation to other commands having SIMPLE task attributes that are already in the task set.~~

~~The device server may use the duration expiration time to determine the order of processing commands with the SIMPLE task attribute within the task set. A difference in duration expiration time between commands may override other scheduling considerations (e.g., different times to access different logical block addresses or vendor specific scheduling considerations). Processing of a collection of commands with different command duration limit settings should cause a command with an earlier duration expiration time to complete with status sooner than a command with a later duration expiration time.~~

<<<This ends the text that is the source for the text that is shown above as being moved.>>>

8.6.3 Enhanced command duration limits processing

<<<All of 8.6.3 is new. Use of modification markups suspended.>>>

If the CDLP field is not set to zero and the RWCDLP bit is set to one, then the processing of duration limited commands is described in this subclause.

If the device server has requests to process duration limited commands with a SIMPLE task attribute, then the device server should process the command based on the contents of the selected Command Duration Limit T2A mode page or Command Duration Limit T2B mode page.

8.7 LU (logical unit) state machines

…

8.7.3 LU_DS (device server) state machine

…

8.7.3.2 LU_DS command processing

…

If there are multiple commands requesting to be processed, then the LU_DS determines the command to be processed using the following rules:

a) any command that has a HEAD OF QUEUE task attribute shall be processed before any command that has an ORDERED task attribute or a SIMPLE task attribute; or
b) if there are no commands that have a HEAD OF QUEUE task attribute, then the processing order for commands that have a SIMPLE task attribute is determined using:
   A) the QUEUE ALGORITHM MODIFIER field in the Control mode page (see SPC-4);
   B) the command priority as defined in 8.5; and
   C) the command duration limit as defined in ~~8.6~~ 8.6.2.

…

If during the processing of a duration limited command the LU_DS state determines that it is unable to complete the command prior to the duration expiration time ~~(see 8.6)~~ (see 8.6.2), then:

1) if data is in the process of being transferred from the Data-Out Buffer (see 5.4.3) or into the Data-In Buffer (see 5.4.3), then …

<<<No other changes are proposed in 8.7.3.2.>>>

SBC-4 r16 Changes Being Proposed

5.17 READ (16) command

… <<<SAM-6 defines 'command duration limit' but not 'command duration time'.>>>

The command duration limit ~~time~~ (see SAM-5) is specified by the command duration limit descriptor (see SPC-5). Which command duration limit descriptor, if any, applies to this command is specified by the DLD2 bit, the DLD1 bit, and the DLD0 bit, as shown in table 82. The CDLP field and the RWCDLP bit in the REPORT SUPPORTED OPERATION CODES parameter data (see SPC-5) indicate ~~indicates~~ that the command duration limit descriptor is in: ~~the Command Duration Limit A mode page or the Command Duration Limit B mode page (see SPC-5).~~

a) the Command Duration Limit A mode page (see SPC-5);
b) the Command Duration Limit B mode page (see SPC-5);
c) the Command Duration Limit T2A mode page (see SPC-6); or
d) the Command Duration Limit T2B mode page (see SPC-6).

…

5.40 WRITE (16) command

<<<The January 2019 CAP working group determined that no changes are required in the WRITE (16) command definition. The decision also reduces the changes in the WRITE SCATTERED (16) command to fixing on typo.>>>

…

5.53 WRITE SCATTERED (16) command

…

See the READ (16) command for the definitions of the DLD1 ~~bit,~~ bit and the DLD0 bit.

…

SBC-4 r16 Log Page Changes Being Proposed

6.4 Log parameters

6.4.1 Log parameters overview

6.4.1.1 Summary of log pages

…

Table 176 — Log page codes and subpage codes for direct access block devices

Log page name	Page code	Subpage code	Reference
…	…	…	…
Cache Memory Statistics	19h	20h	SPC-5
Command Duration Limits Statistics	19h	<<TBD ¿21h?>>	SPC-6
…	…	…	…
<<<No other changes proposed in table 176.>>>			

…

SBC-4 r16 Mode Page Changes Being Proposed

6.5 Mode parameters

6.5.1 Summary of mode page codes

…

Table 228 — Mode page codes and subpage codes for direct access block devices

Mode page name	Page code	Subpage code	Reference
Command Duration Limit A	0Ah	03h	SPC-5
Command Duration Limit B	0Ah	04h	SPC-5
Command Duration Limit T2A	0Ah	<<TBD ¿07h?>>	SPC-6
Command Duration Limit T2B	0Ah	<<TBD ¿08h?>>	SPC-6
…	…	…	…
<<<No other changes proposed in table 228.>>>			

<<<No other changes are proposed for SBC-4.>>>

5.17 READ (16) command

The READ (16) command (see table 81) requests that the device server perform the actions defined for the READ (10) command (see 5.15).

Table 81 — READ (16) command

Byte \ Bit	7	6	5	4	3	2	1	0
0	OPERATION CODE (88h)							
1	RDPROTECT			DPO	FUA	RARC	Obsolete	DLD2
2	(MSB)							
•••	LOGICAL BLOCK ADDRESS							
9								(LSB)
10	(MSB)							
•••	TRANSFER LENGTH							
13								(LSB)
14	DLD1	DLD0	GROUP NUMBER					
15	CONTROL							

The OPERATION CODE field is defined in SPC-5 and shall be set to the value shown in table 81 for the READ (16) command.

The command duration time (see SAM-5) is specified by the command duration limit descriptor (see SPC-5) specified by the DLD2 bit, the DLD1 bit, and the DLD0 bit, as shown in table 82. The CDLP field in the REPORT SUPPORTED OPERATION CODES parameter data (see SPC-5) indicates that the command duration limit descriptor is in the Command Duration Limit A mode page or the Command Duration Limit B mode page (see SPC-5).

Table 82 — Duration limit value bits specifying command duration limit descriptor

Duration limit descriptor value bits			Command duration limit descriptor specifying command duration time
DLD2	DLD1	DLD0	
0b	0b	0b	Command is not a duration limited command (see SAM-5)
0b	0b	1b	First command duration limit descriptor
0b	1b	0b	Second command duration limit descriptor
0b	1b	1b	Third command duration limit descriptor
1b	0b	0b	Fourth command duration limit descriptor
1b	0b	1b	Fifth command duration limit descriptor
1b	1b	0b	Sixth command duration limit descriptor
1b	1b	1b	Seventh command duration limit descriptor

The CONTROL byte is defined in SAM-5.

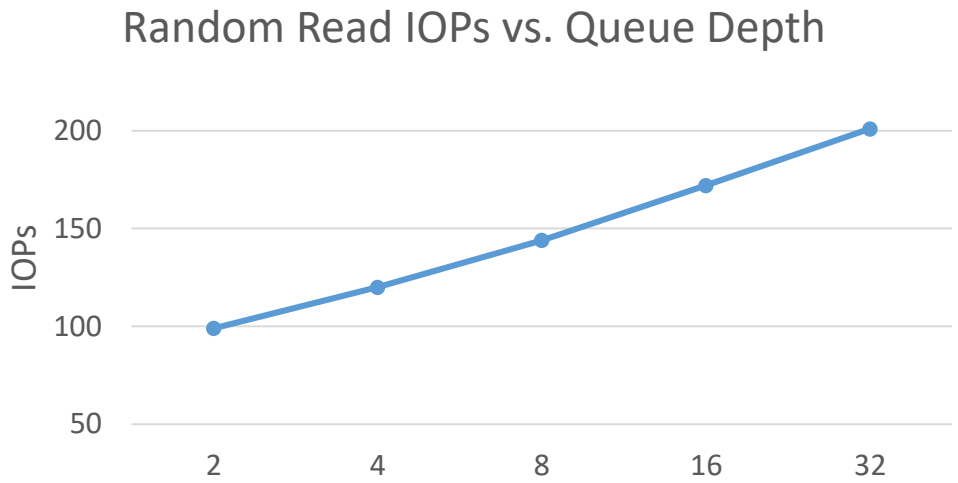
See the READ (10) command (see 5.15) for the definitions of the other fields in this command.

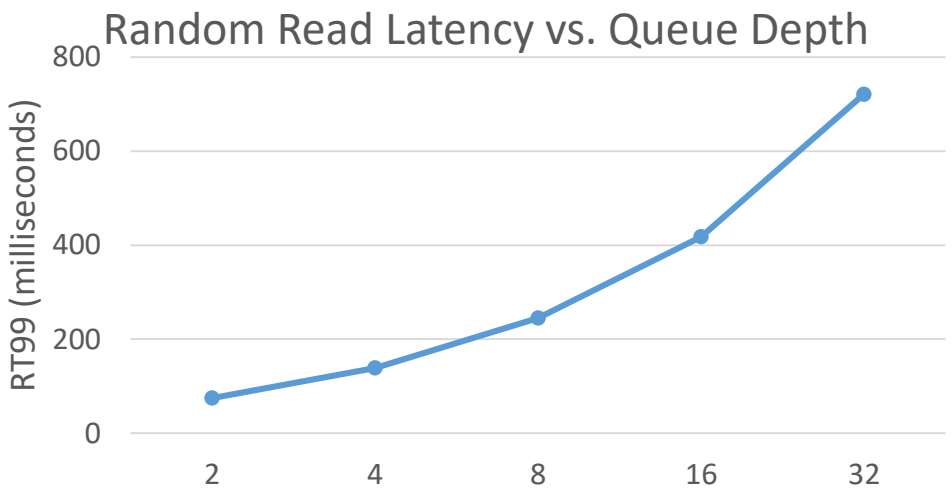
Backups

Command-Level QoS

Background and Problem Statement

- Disk drive performance scales with queue depth as command reordering by the drive reduces seek overhead
- Although IOPs increase, command latencies increase with queue depth
- Typical drive workloads combine IO request from multiple sources
  - User requests that require low latency response
  - Throughput-oriented compute tasks
  - Background operations



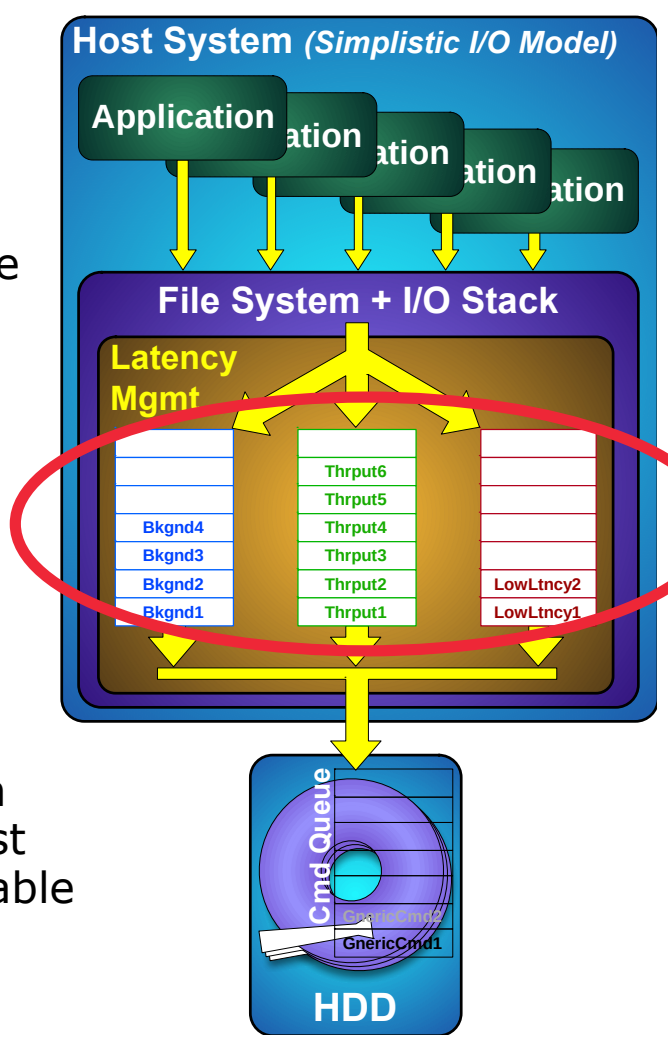


- Because command latencies increase with queue depth, the host must use low queue depth (and sacrifice IOPs) to get low latency when needed

Command-Level QoS

Optimizing Both Latency and IOPs in HDD

- Current Non-Optimal Workaround
  - Applications create multiple requests, host limits HDD queue depth to constrain latency
  - Significant system performance is lost as the drive is unable to optimize seeks



- Proposed Solution
  - Host specifies per-command QoS level and maintains higher drive queue depth
  - Drive reorders commands to optimize IOPs while maintaining latency target for specified low latency commands