

Data Set Management Commands Proposal for ATA8-ACS2

| Oct. 5th, 2007
Revision 3

Author: Frank Shu /Nathan Obr
Microsoft Corporation
One Microsoft Way
Redmond, WA. 98052-6399
USA

Phone: (425) 707-1779
E-Mail: frankshu@microsoft.com

Phone: (425) 705-9157
E-Mail: natobr@microsoft.com

Table of Contents

1	INTRODUCTION	3
2	DESCRIPTION OF I/O QUALITY OF SERVICE.....	3
3	DESCRIPTION OF THE DATA SET MANAGEMENT COMMAND	43
4	DATA SET MANAGEMENT COMMANDS PROPOSAL.....	4
4.1	DATA SET MANAGEMENT COMMANDS REQUIREMENTS	4
4.2	PROPOSED DEFINITION OF TERMS FOR SOLUTION.....	4
5	PROPOSED CHANGES TO ATA8-ACS	5
5.1	IDENTIFY	5
5.2	DCO	5
5.3	CLARIFICATION OF SECURITY ERASE TO DEALLOCATED LBA(S).....	5
6	PROPOSED NEW COMMANDS FOR ATA8-ACS	65
6.1.1	<i>Data Set Management Ext – xxh, DMA.....</i>	6 5

Revision History		
Date	Revision	Description
2007-04-21	0	Initial Draft
2007-06-04	1	<ol style="list-style-type: none"> 1. Changing the title of the proposal to reflect more advanced feature covered by this proposal. 2. Adding new dataset attributes for device optimization 3. Incorporating feedbacks on “Trim” attribute. <ul style="list-style-type: none"> • Non media specific command • Define the read response after “Trim” • Clarification of relation to Security Erase
2007-09-05	2	<ol style="list-style-type: none"> 1. Moving normal return status of Data Set Management command from Status register to Count register. 2. Attributes Importance 3. Attributes Clear bit
<u>2007-10-05</u>	<u>3</u>	<ol style="list-style-type: none"> 1. Incorporated feedbacks from 9/16/07 T13 ad-hoc meeting 2. Moving “Trim” attribute ahead with space in protocol for extension of other attributes later.

Revision History		
Date	Revision	Description

1 Introduction

The first version of this proposal was called “Trim” command proposal, the “trim” command carries the information related to deleted data blocks to device for optimization. But, more data has showed that the information other than deleted data blocks provided by host can also be useful to device, one example would be the frequency of data access of write or read. A collected set of such information we called “data set attributes”. The proposal now is called “Data Set Management Commands”

This proposal discusses the details necessary for defining and implementing Data Set Management commands on an ATA device. The advantage to Data Set Management commands are having methods to communicate an application’s **I/O access behaviors** within a host to the ATA device. This new feature, when used with applications within a host that accurately provide I/O access behavior information about it’s data, allows an ATA device to be able take internal optimizations to provide the responsiveness needed by the host when later accessing that data.

This document provides a common understanding of the new concepts that Data Set Management introduces, provides a common language to discuss Data Set Management functionality, and proposes a new command that creates a Data Set Management abstraction.

2 Description of I/O Quality of Service

An application layer in a host that is a source for I/O operations may be able to know or predict the nature of the data it is storing on the ATA device. The nature of how this data will be accessed or used in the future is expressed by **context attributes** that identify how the data should to be stored. Context attributes are applied to **sector ranges** in order to identify a **data set**. Together context attributes for a data set instructs the ATA device how that data set will likely be used in the future by the host.

If the host can provide accurate context attributes to the ATA device, then the ATA device can use this context to make informed decisions on how to best provide the **quality of service**, indicated by attributes. By providing context attributes, the host provides Data Set Management information to the ATA device.

3 Description of the Data Set Management Command

For the context attributes provided by the host to be meaningful, the context attributes defined in this document must allow the host reflect aspects of quality of service such as pattern, frequency, latency, and sector ranges read and written by the host.

The primary piece of information provided in a Data Set Management command is the sector ranges that represent the data set that the application stored on the ATA device. Often times the data set stored by the application is not contiguous and consequently it is important that the command be able to provide a series of discontinuous sectors of a flexible length. Since the command structure in ATA does not meet this requirement, the sector ranges for a Data Set Management command must be passed to the ATA device through a data transfer just like the NVC feature set.

The secondary piece of information provided in a Data Set Management command is the context attributes that identify the quality of service requested by the host. The context attributes that the host provides is also valuable for any command that provides a sector range, such as read or write commands.

4 Data Set Management Commands Proposal

The primary focus of the proposed Data Set Management command is to enable the host to share quality of service context attributes for sector ranges with the ATA device, which is a new concept to the ACS standard. Consequently many of the concepts that will be used in discussion and in the remainder of this document are also new. This section establishes the goals and requirements for the Data Set Management proposal and defines terminology for new Data Set Management concepts.

4.1 Data Set Management Commands Requirements

The purpose of Data Set Management command is to create a mechanism by which the host may share quality of service information with the ATA device. Requirements for the Data Set Management proposal break down into the mechanism's ability to deliver the quality of service context information and the abstraction that defines what form the context information takes.

The Data Set Management mechanism must:

- specify a mechanism to identify a flexible number of possibly discontinuous sector ranges as a data set
- limit sector ranges to physical sector aligned and physical sector multiple sizes
- provide a mechanism for identifying which quality of service context attributes the ATA device supports for future expansion.

The quality of service context information must:

- reflect the pattern, frequency, latency for one or more data sets
- be able to be contained in a standalone command that specifies a sector range as well as an extension to a read command and a write command.

4.2 Proposed Definition of Terms for Solution

I/O Access Behavior The read and write pattern, dependency between reads and writes and frequency of reads and writes created by an I/O source in a host.

Context Attributes Encodable information that identifies I/O access behavior applied to one or more data sets.

Sector Range A start sector address and length that identifies contiguous sectors that make up part or all of a data set.

Data Set A set of **Sector Ranges** to be treated by device as single group.

Quality of Service The frequency the host expects to access a data set and the required latency expected by the host for any operation on elements of the data set.

LBA Range Entry as defined in ATA-ASC(4.16.3.2 LBA Range Entry)

5 Proposed Changes to ATA8-ACS

5.1 Identify

Determining whether a given ATA HDD supports Data Set Management is a straight forward process. One reserved bit of one of the ATA-8 ACS IDENTIFY DEVICE words is used by the device during device enumeration to indicate support. The chosen IDENTIFY DEVICE word would be augmented with the following description:

Bit [TBD1] when set to one indicates that the device supports the Data Set Management command and performs internal device optimizations based on attributes for I/O ranges specified in Data Set Management commands.

5.2 DCO

Likewise, one reserved bit of "Command set/feature set supported part 2" DCO word is used by device to indicate Data Set Management support. The chosen DCO word would be augmented with the following description:

Bit [TBD2] bit [TBD2] if set to one indicates that support for Data Set Management feature set is changeable.

5.3 Clarification of Security Erase to Deallocated LBA(s)

Deallocated command does not affect existing Security Erase operation. When security erase is required, it shall apply to deallocated LBA(s) as well.

5.4 Clarification of Data Set Attribute Clear

Data Set Attribute has two types; Persistent vs. Temporary, they are indicated by Data Set Management command bit 15 of Feature register. The data sets with persistent attributes are OS data and file system meta data which is intended to be used over time(sessions and power-on cycles). This type of attributes should be maintained by device all time. On the other hand, temporary data attribute only indicates user's data access pattern over session(s), and it likely will change over time. To effectively use temporary data attribute, only data sets with high W/R frequency/Impotency are important to host, the temporary data attributes with low R/W frequency/Impotency can be cleared by device.

6 Proposed New Commands for ATA8-ACS

New commands:

Data Set Management

6.1.1 Data Set Management Ext – xxh, DMA

ATA Command Format for Data Set Management Command

Word	Name	Description
00h	Feature	Bit Description
		15 Reserved <u>Persistent Attribute</u>
		14:12 Reserved
		11 Reserved <u>Deallocate(Trim)</u>
		10 Reserved <u>Write-Prepare</u>
		9:8 Reserved <u>Write-Latency</u>
		7:6 Reserved <u>Read-Latency</u>
		5:4 Reserved <u>Write-Frequency/Importance</u>
		3:2 Reserved <u>Read-Frequency/Importance</u>
		1 Reserved <u>Atomic Write</u>
0 Deallocated(Trim) <u>Atomic Read</u>		
01h	Count	Number of 256 word-blocks of dataset LBA Range Entry to be transferred, 0000h specifies that 65,536 blocks are to be transferred.
02 – 04h	LBA	Bit Description
		47:20 Reserved
		19:16 Reserved <u>Session Length</u>
		15:0 Reserved <u>Common Access Size</u>
05h	Command	XXh(TBD3)

Formatted: Font: Not Bold

LBA Range Entry as defined in ATA-ASC(4.16.3.2 LBA Range Entry)

Atomic Read

Atomic Read is set to one to indicate the data set should be optimized for atomic read access. In the case of a range larger than the maximum transfer size, a read to any portion of the range should imply that reads to all the other ranges are imminent. The host expects to perform read operations on the data set as single object.

Atomic Write

Atomic Write shall be set to one to indicate the data set should be optimized for atomic write access. In the case of a range larger than the maximum transfer size, a write to any portion of the range should imply that writes to all the other ranges are imminent. The host expects to perform write operations on the data set as single object.

Read Frequency/Importance

Read Frequency specifies the expected number of times for a given session time that any or all of the given range will be read by the host. This field can take on 4 separate values defined as:

00 No Read Frequency given. Read Frequency for this range is not being changed or set.

01 Long term storage. Read less than once per Session.

10 Users current working set(Important). Read every Session.

~~11 Dynamic Data(Important). Read more than once per Session.~~

Write Frequency/Importance

~~Write Frequency specifies the expected number of times for a given session time that any or all of the given range will be Written by the host. This field can take on 4 separate values defined as:~~

~~00 No Write Frequency given. Write Frequency for this range is not being changed or set.~~

~~01 Long term storage. Written less than once per Session.~~

~~10 Users current working set(Important). Written every Session.~~

~~11 Dynamic Data(Important). Written more than once per Session.~~

Read Latency

~~Read Latency specifies the required access latency the host expects the device to use for accessing any or all of the given range when read. This field can take on 4 separate values defined as:~~

~~00 No Read Latency given. Read Latency for this range is not being changed or set.~~

~~01 Idle. Should be processed at the same priority as the devices background tasks.~~

~~10 Normal. Should be processed without optimizations.~~

~~11 Urgent. Should be processed at the fastest possible latency even at the expense of other Fast latency I/O operations.~~

Write Latency

~~Write Latency specifies the required access latency the host expects the device to use for accessing any or all of the given range when written. This field can take on 4 separate values defined as:~~

~~00 No Write Latency given. Write Latency for this range is not being changed or set.~~

~~01 Idle. Should be processed at the same priority as the devices background tasks.~~

~~10 Normal. Should be processed without optimizations.~~

~~11 Urgent. Should be processed at the fastest possible latency even at the expense of other Fast latency I/O operations.~~

Write Prepare

~~Allocate shall be set to one to indicate the provided range is being prepared to be written by the host. Any optimizations that the device can take to prepare for an imminent write to this range should be taken. When Allocate is set to one, there is no guarantee that the host will write to any or all of this range. Setting Allocate to one on a range does not remove its deallocated status.~~

Deallocate(Trim)

~~Deallocate shall be set to one to indicate the data set no longer needs to be maintained. The host expects to write this data set before it reads it again. If a read occurs to any part of the data set before it is written, the device shall return the data stored from its previous write or it shall return all 0s.~~

~~Once any sector of a deallocated range has been written, that sector no longer has its deallocated status.~~

Session Length

~~Determines the time frame within which the host expects to take an access operation on the given range.~~

~~0 = No session provided. No Read Frequency or Write Frequency information should be provided. If Session Length is 0 Read Frequency and Write Frequency information shall be ignored.~~

~~1 = 1 minute~~

~~2 = 1 hour~~

~~3 = 1 day~~

~~4 = 1 week~~

~~5 = 1 month~~

~~All other values = Reserved~~

Common Access Size

~~Number of logic sectors expected to be done in a single transfer from this data set. The value of 0 indicates no Common Access Size given.~~

Persistent Attribute

~~When set to "1", it indicates the data set attributes are persistent over sessions and device power-on cycles, when is "0", it indicates temporary attributes. Persistent Attribute does not apply to "trim" attribute.~~

6.1.1.1 Normal Outputs

Word	Name	Description																
00h	Error	00h																
01h	Count	15: 2 Reserved 1 Reserved 0 Unaligned Range Reserved Write Prepare Failed																
02h-04h	LBA	Reserved																
05h	Status	<table border="1"> <thead> <tr> <th>Bit</th> <th>Description</th> </tr> </thead> <tbody> <tr> <td>7:6</td> <td>Transport Dependent - See clause 6.2.11 of ATA8-ACS</td> </tr> <tr> <td>5</td> <td>Device Fault - See clause 6.2.4 of ATA8</td> </tr> <tr> <td>4</td> <td>N/A</td> </tr> <tr> <td>3</td> <td>Transport Dependent - See clause 6.2.11 of ATA8-ACS</td> </tr> <tr> <td>2</td> <td>Reserved</td> </tr> <tr> <td>1</td> <td>Reserved</td> </tr> <tr> <td>0</td> <td>Error - See clause 6.2.3 of ATA8-ACS</td> </tr> </tbody> </table>	Bit	Description	7:6	Transport Dependent - See clause 6.2.11 of ATA8-ACS	5	Device Fault - See clause 6.2.4 of ATA8	4	N/A	3	Transport Dependent - See clause 6.2.11 of ATA8-ACS	2	Reserved	1	Reserved	0	Error - See clause 6.2.3 of ATA8-ACS
Bit	Description																	
7:6	Transport Dependent - See clause 6.2.11 of ATA8-ACS																	
5	Device Fault - See clause 6.2.4 of ATA8																	
4	N/A																	
3	Transport Dependent - See clause 6.2.11 of ATA8-ACS																	
2	Reserved																	
1	Reserved																	
0	Error - See clause 6.2.3 of ATA8-ACS																	

Unaligned Range

~~Unaligned Range shall be set to one if one or more of the range LBA Values were not physical sector aligned or one or more of the range lengths were not physical sector size multiples. When this bit is set to one, the attribute settings provided may not have been accepted by the device.~~

Write Prepare Failed

~~Allocate Failed should be set to one to indicate that the device is becoming full enough to hurt performance and if the data can be stored on another device the host should try to store the data elsewhere. Setting 'Allocate Failed' to one does not give the device permission to fail writes to any or all of the range specified. The host may still issue the write to this range at any time.~~

Error Outputs

See Table 120