Investigation of Alternative Disk Drives for Mark5C Modules

Walter Brisken, Frank Schinzel & Bob McGoldrick
9 May 2012

Abstract This memo is mostly a collection of performance data collected for three hard disk drive models, including the standard 2 TB drive used in 16 TB Mark5C modules. Poor performance is seen for two drive models that suggests limitations other than raw drive speed. A particularly worrying possibility is that the speeds achieved with Western Digital drives is due to use of an obsolete drive feature (TCQ) that may not be supported in the future.

1 Drive models tested

In this memo, three disk drives were tested. The following table summarizes their key aspects. The last two lines indicate transfer speeds as determined by the standard benchmark program hdtune for data at outer radii (where performance should be maximal) and inner radii. For all drives, these numbers hint that the raw drive performance is not the limiting aspect to their performance in a Mark5 module.

<table>
<thead>
<tr>
<th>Drive Model</th>
<th>Western Digital</th>
<th>Hitachi</th>
<th>Seagate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>WD2003FYYYS-02W0B0</td>
<td>HDS723030ALA640</td>
<td>ST32000641AS</td>
</tr>
<tr>
<td>Size (TB)</td>
<td>2</td>
<td>3TB</td>
<td>2TB</td>
</tr>
<tr>
<td>Price ($)</td>
<td>246</td>
<td>325</td>
<td>198</td>
</tr>
<tr>
<td>Spindle speed (rpm)</td>
<td>7200</td>
<td>7200</td>
<td>7200</td>
</tr>
<tr>
<td>hdtune outer track rate (Mbps)</td>
<td>1120</td>
<td>1240</td>
<td>1104</td>
</tr>
<tr>
<td>hdtune inner track rate (Mbps)</td>
<td>560</td>
<td>600</td>
<td>504</td>
</tr>
</tbody>
</table>

2 Conditioning performance tests

A fully populated (8 drive) SATA Mark5 module was built for each drive model. The mk5erase program was run in conditioning mode. Every 10 seconds the progress and average data rate were saved to a file. These rates are plotted against module “write location” in figures 1, 2 & 3, with smaller numbers (worse results) corresponding to inner platter radii.

It is unclear why the performance of the Hitachi and Seagate drives is poor relative to the Western Digital drive. The raw transfer speeds as shown in above table are much greater than the observed performance. The Linux command hdparm -I can be used to determine the features available on an attached hard drive. Full output from this program for the three drives is shown below in section 4. With one drive known to perform well (the Western Digital one) and two that don’t, one can look for capabilities that are either present (or absent) on the drive Western Digital drive but absent (or present) in both of the others. The only reported feature matching this pattern is “Device-initiated power management” which is supported by both Hitachi and Seagate drives but not Western Digital ones. It is unclear if this is related or not.

3 Tagged Command Queuing

One additional option that could explain the difference in performance is the presence or absence of a drive feature called Tagged Command Queuing (TCQ). Western Digital was the first (and perhaps only?) hard drive manufacturer to introduce the SCSI concept of Tagged Command Queuing to their line of SATA drives. This was one feature distinguishing their early enterprise grade SATA drives from the others. TCQ allows

Figure 1: Conditioning speed as a function of module write location for a “standard” Mark5C module built from Western Digital hard drives. Write speed is shown in green and read speed is shown in red. For a module to be useful in a Mark5C system, the write speed must exceed 2050 Mbps across the entire module. It may appear that there are two curves of each color. This is due to the quantized nature of the position measurement made during conditioning.

Figure 2: Conditioning speed as a function of module write location for a Mark5C module built from Hitachi hard drives. Write speed is shown in green and read speed is shown in red.

improved performance by allowing commands (reads or writes) to be executed out of order in a manner that is optimal given the spin phase of the drive platters. This feature has since been superceded by Native
Figure 3: Conditioning speed as a function of module write location for a Mark5C module built from Seagate hard drives. Write speed is shown in green and read speed is shown in red.

Command Queuing (NCQ) which has superior performance. TCQ is of particular interest because it is supported by the Marvell 88i8030 PATA-SATA bridge used in the Conduant Mark5C modules. NCQ, on the other hand, is not supported. It is plausible that the performance of the Western Digital drives is due to latent support for TCQ in the Western Digital drive. If this is the case, then the prospects of long term product availability for Mark5C would be troubling.

4 Drive specifics

The sections below are raw output from the standard linux program `hdparm` using option `-I`. In all cases, one of the drives was directly attached to a motherboard SATA port, bypassing the StreamStor card. It is thus likely that a different set of supported features becomes enabled during Mark5 usage.
Output of `hdparm -I` for a Western Digital drive

ATA device, with non-removable media

- **Model Number:** WDC WD2003FYYS-02W080
- **Serial Number:** WD-WMAY00766109
- **Firmware Revision:** 01.01D01

**Transport:** Serial, SATA 1.0a, SATA II Extensions, SATA Rev 2.5

**Standards:**
- Supported: 8 7 6 5
- Likely used: 8

**Configuration:**

<table>
<thead>
<tr>
<th>Logical max current</th>
<th>cylinders</th>
<th>16383</th>
<th>16383</th>
</tr>
</thead>
<tbody>
<tr>
<td>heads</td>
<td></td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>sectors/track</td>
<td></td>
<td>63</td>
<td>63</td>
</tr>
</tbody>
</table>

- **CHS current addressable sectors:** 16514064
- **LBA user addressable sectors:** 268435455
- **LBA48 user addressable sectors:** 390709168
- **device size with M = 1024×1024:** 1907729 MBytes
- **device size with M = 1000×1000:** 2000398 MBytes (2000 GB)

**Capabilities:**

- LBA, IORDY (can be disabled)
- Queue depth: 32
- Standby timer values: spec’d by Standard, with device specific minimum
- R/W multiple sector transfer: Max = 16 Current = 0
- Advanced power management level: unknown setting (0x0080)
- Recommended acoustic management value: 128, current value: 254
- DMA: mdma0 mdma1 mdma2 udma0 udma1 udma2 udma3 udma4 *udma6

**Cycle time:**
- min=120ns recommended=120ns
- PIO: pio0 pio1 pio2 pio3 pio4
-Cycle time: no flow control=120ns IORDY flow control=120ns

**Commands/features:**

- **Enabled**
  - SMART feature set
  - Security Mode feature set
  - Power Management feature set
  - Write cache
  - Look-ahead
  - Host Protected Area feature set
  - WRITE_BUFFER command
  - READ_BUFFER command
  - NOP cmd
  - DOWNLOAD_MICROCODE
  - Advanced Power Management feature set
  - Power-Up In Standby feature set
  - SET_MAX security extension
  - Automatic Acoustic Management feature set
  - 48-bit Address feature set
  - Device Configuration Overlay feature set
  - Mandatory FLUSH_CACHE
  - FLUSH_CACHE_EXT
  - SMART error logging
  - SMART self-test
  - General Purpose Logging feature set
  - WRITE_(DMA|MULTIPLE)_FUA_EXT
  - 64-bit World wide name
  - IDLE_IMMEDIATE with UNLOAD
  - WRITE_UNCORRECTABLE command
  - (READ,WRITE)_DMA_EXT_GPL commands
  - Segmented DOWNLOAD_MICROCODE
  - SATA-I signaling speed (1.5Gb/s)
  - SATA-II signaling speed (3.0Gb/s)
  - Native Command Queueing (NQ2)
  - Phy event counters
  - unknown 76[11]
  - unknown 76[12]
  - DMA Setup Auto-Activate optimization
  - Software settings preservation

**Security:**

- **Master password revision code = 65534**
  - supported
  - not enabled
  - not locked
  - not frozen
  - not expired: security count
  - supported: enhanced erase

**Checksum:** correct

306min for SECURITY ERASE UNIT. 306min for ENHANCED SECURITY ERASE UNIT.
Output of hdparm -I for a Hitachi drive

Model Number: Hitachi HDS723030ALA640
Serial Number: NK0331YHGKBS2A
Firmware Revision: MSADGASCO
Transport: Serial, ATAS-AST, SATA 1.0a, SATA II Extensions, SATA Rev 2.5, SATA Rev 2.6; Revision: ATAS-AST T13 Project D1697 Revision 0b

Standards:
- Used: unknown (minor revision code 0x0029)
- Supported: 8 7 6 5
- Likely used: 8

Configuration:
- Logical max current cylinders: 16383 16383
- heads: 16 16
- sectors/track: 63 63
- CHS current addressable sectors: 16514064
- LBA user addressable sectors: 268435455
- LBA48 user addressable sectors: 586053136
- Logical Sector size: 512 bytes
- Physical Sector size: 512 bytes
- device size with M = 1024*1024: 2861588 MBytes
- device size with M = 1000*1000: 3000592 MBytes (3000 GB)
- cache/ buffer size = unknown
- Form Factor: 3.5 inch
- Nominal Media Rotation Rate: 7200

Capabilities:
- LBA, IORDY (can be disabled)
- Queue depth: 32
- Standby timer values: spec'd by Standard, no device specific minimum
- R/W multiple sector transfer: Max = 16, Current = 0
- Advanced power management level: disabled
- DMA: mdma0 mdma1 mdma2 udma0 udma1 udma2 udma3 udma4 udma5 *udma6
- Cycle time: min=120ns recommended=120ns
- PIO: pio0 pio1 pio2 pio3 pio4
- Cycle time: no flow control=120ns IORDY flow control=120ns

Commands/features:
- Enabled
- SMART feature set
- Security Mode feature set
- Power Management feature set
- Write cache
- Look-ahead
- Host Protected Area feature set
- WRITE_BUFFER command
- READ_BUFFER command
- NOP cmd
- DOWNLOAD_MICROCODE
- Advanced Power Management feature set
- Power-Up In Standby feature set
- SET_FEATURES required to spinup after power up
- SET_MAX security extension
- 48-bit Address feature set
- Device Configuration Overlay feature set
- Mandatory FLUSH_CACHE
- FLUSH_CACHE_EXT
- SMART error logging
- SMART self-test
- Media Card Pass-Through
- General Purpose Logging feature set
- WRITE_DMA(MULTIPLE)_FUA_EXT
- 64-bit World wide name
- URG for READ_STREAM(DMA)_EXT
- URG for WRITE_STREAM(DMA)_EXT
- WRITE_UNCORRECTABLE_EXT command
- {READ,WRITE}_DMA_EXT_GPL commands
- Segmented DOWNLOAD_MICROCODE
- unknown 119[7]
- Gen1 signaling speed (1.5Gb/s)
- Gen2 signaling speed (3.0Gb/s)
- unknown 76[3]
- Native Command Queueing (NCQ)
- Host-initiated interface power management
- Phy event counters
- NCQ priority information
- Non-Zero buffer offsets in DMA Setup FIS
- DMA Setup Auto-Activate optimization
- Device-initiated interface power management
- In-order data delivery
- Software settings preservation
- SMART Command Transport (SCT) feature set
- SCT LBA Segment Access (AC2)
- SCT Error Recovery Control (AC3)
- SCT Features Control (AC4)
- SCT Data Tables (AC5)

Security:
- Master password revision code = 65534
- supported
- not enabled
- Logical Unit WWN Device Identifier: 5000cca225c7e665
  - NAA: 5
  - IEEE OUI: 000cca
  - Unique ID: 225c7e665
  - Checksum: correct
ATA device, with non-removable media
Model Number: ST32000641AS
Serial Number: Z2700FPQ
Firmware Revision: CC13
Transport: Serial

Standards:
Used: unknown (minor revision code 0x0029)
Supported: 8 7 6 5
Likely used: 8

Configuration:
Logical cylinders max current
16383 16383
heads 16 16
sectors/track 63 63

---
CNS current addressable sectors: 16514064
LBA user addressable sectors: 268435455
LBA48 user addressable sectors: 3907029168
Logical/Physical Sector size: 512 bytes
device size with M = 1024*1024: 1907729 MBytes
device size with M = 1000*1000: 2000398 MBytes (2000 GB)
cache/buffer size = unknown

Nominal Media Rotation Rate: 7200

Capabilities:
LBA, IORDY(can be disabled)
Queue depth: 32
Standby timer values: spec'd by Standard, no device specific minimum
R/W multiple sector transfer: Max = 16 Current = ?
Recommended acoustic management value: 254, current value: 0
DMA: mdma0 mdma1 mdma2 udma0 udma1 udma2 udma3 udma4 *udma5 udma6
Cycle time: min=120ns recommended=120ns
PIO: pio0 pio1 pio2 pio3 pio4
Cycle time: no flow control=120ns IORDY flow control=120ns

Commands/features:
Enabled Supported:
* SMART feature set
  Security Mode feature set
  Power Management feature set
  Write cache
  Look-around
  Host Protected Area feature set
  WRITE_BUFFER command
  READ_BUFFER command
  DOWNLOAD_MICROCODE
  SET_MAX security extension
  48-bit Address feature set
  Device Configuration Overlay feature set
  Mandatory FLUSH_CACHE
  FLUSH_CACHE_EXT
  SMART error logging
  SMART self-test
  General Purpose Logging feature set
  WRITE_(DMA|MULTIPLE|PUA)_EXT
  64-bit World wide name
  Write-Read-Verify feature set
  WRITE_UNCORRECTABLE_EXT command
  {READ,WRITE}_(DMA|MULTIPLE|PUA)_GPL commands
  Segmented DOWNLOAD_MICROCODE
  Gen1 signaling speed (1.5Gb/s)
  Gen2 signaling speed (3.0Gb/s)
  Gen3 signaling speed (6.0Gb/s)
  Native Command Queuing (NCQ)
  Phy event counters
  Device-initiated interface power management
  Software settings preservation

Security:
Master password revision code = 65534
supported
not enabled
not locked
not frozen
not expired: security count
supported: enhanced erase
324min for SECURITY ERASE UNIT. 524min for ENHANCED SECURITY ERASE UNIT.

Logical Unit WWN Device Identifier: 6000c5003ecf91f
NAA : 5
IEEE OUI : 000c50
Unique ID : 03ecf91f

Checksum: correct