HITACHI
Inspire the Next

High-Definition Multi-Room DVRs and HDDs

Hitachi Global Storage Technologies

Donald Molaro and Jorge Campello

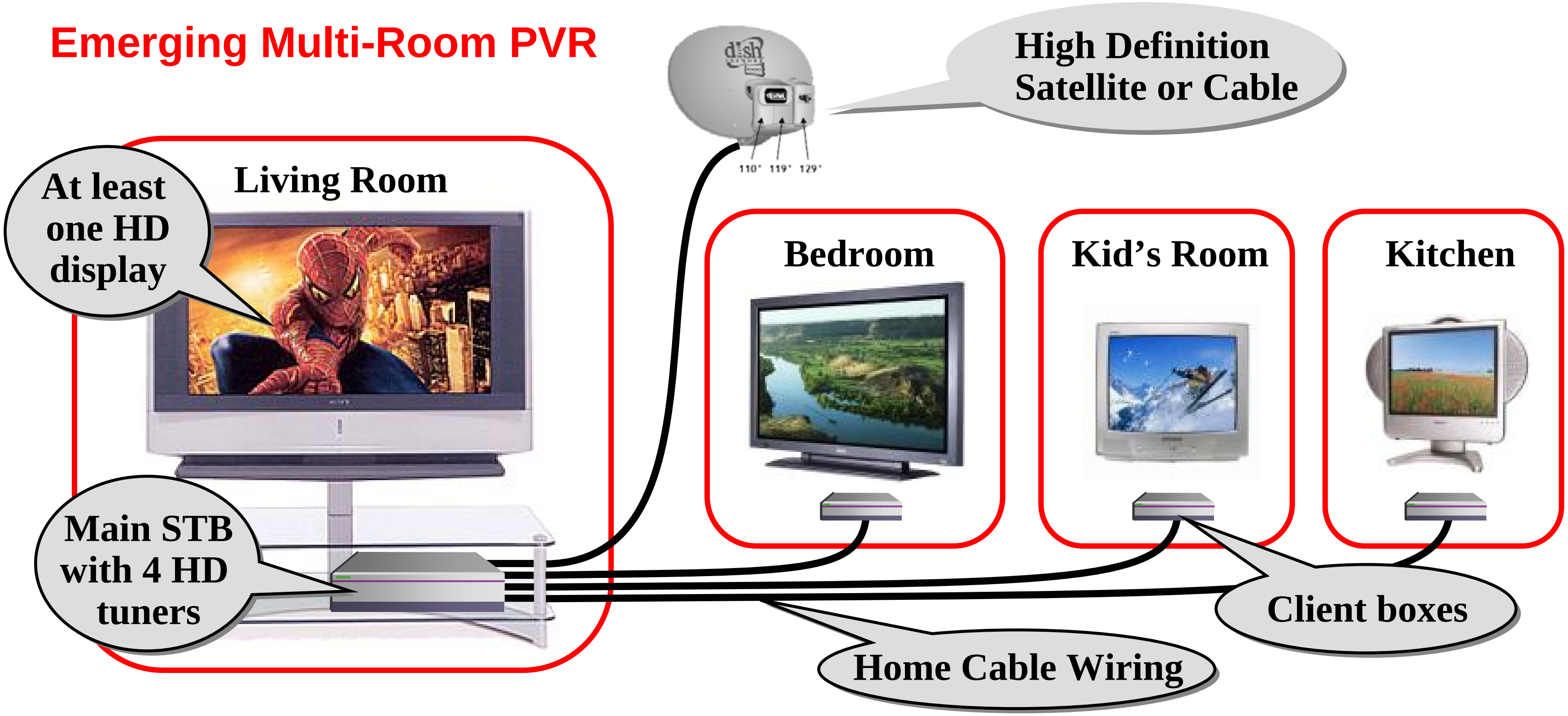
San Jose Research Center

Hitachi GST

VCRs vs DVRs (a.k.a. PVR)

- What is difference between DVR and VCR?
  - Digital vs analogue
  - Easy to program? What about VCR+
  - Tape vs HDD
    - No need to change the tape
    - Stores many weeks
    - Not for archival
    - Simultaneously view and record.
    - Record multiple
    - Playback multiple

Emerging Multi-Room PVR

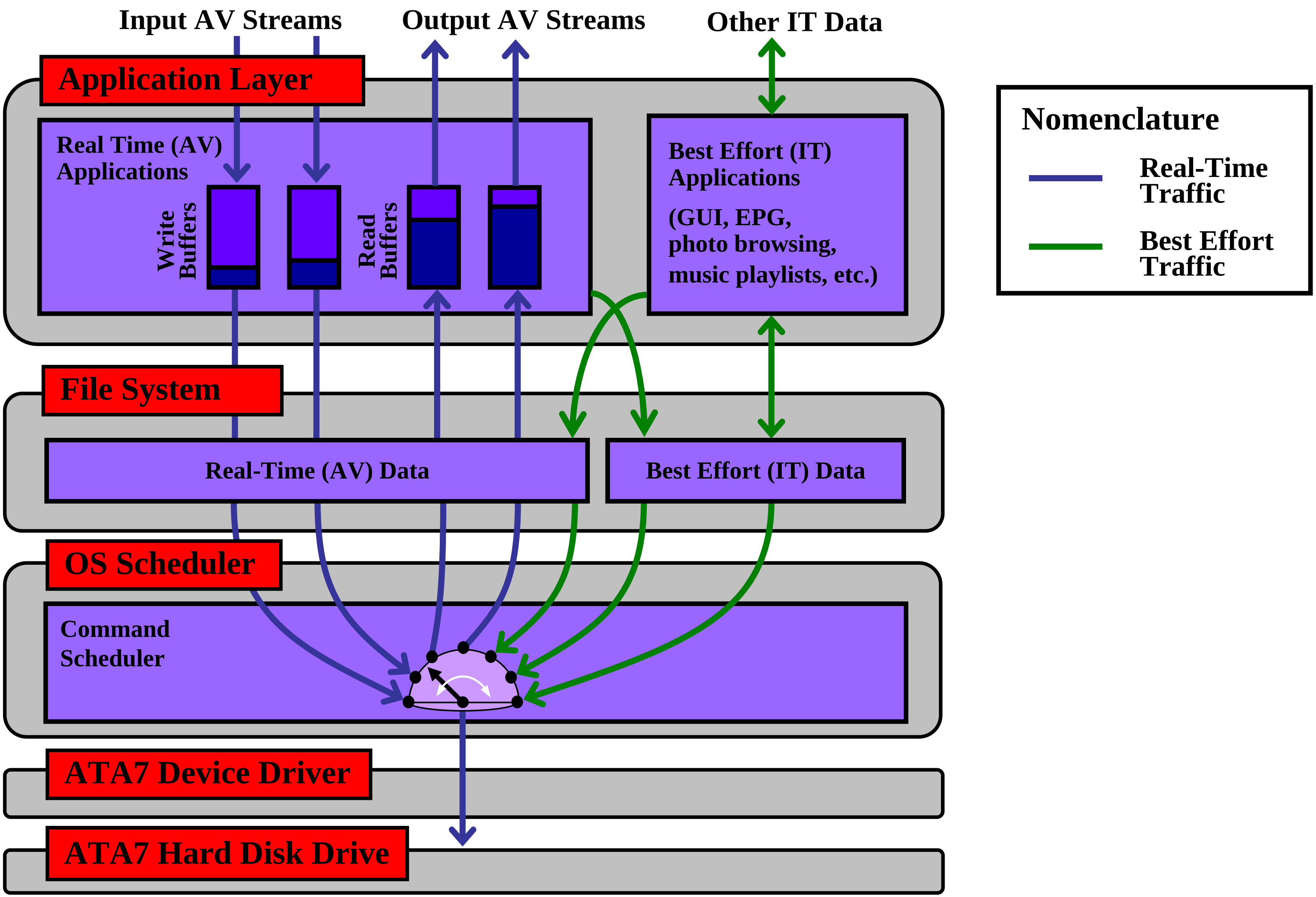


System Requirements

- Support for 4 write streams and 4 read streams, with trick play capability (fast forward and rewind) on each of the read streams.
- Additional support for other "best effort" applications such as: EPG, photo browsing, music playlists, web browsing, CD ripping, IP TV downloads, etc.

QoS For AV Streaming Applications

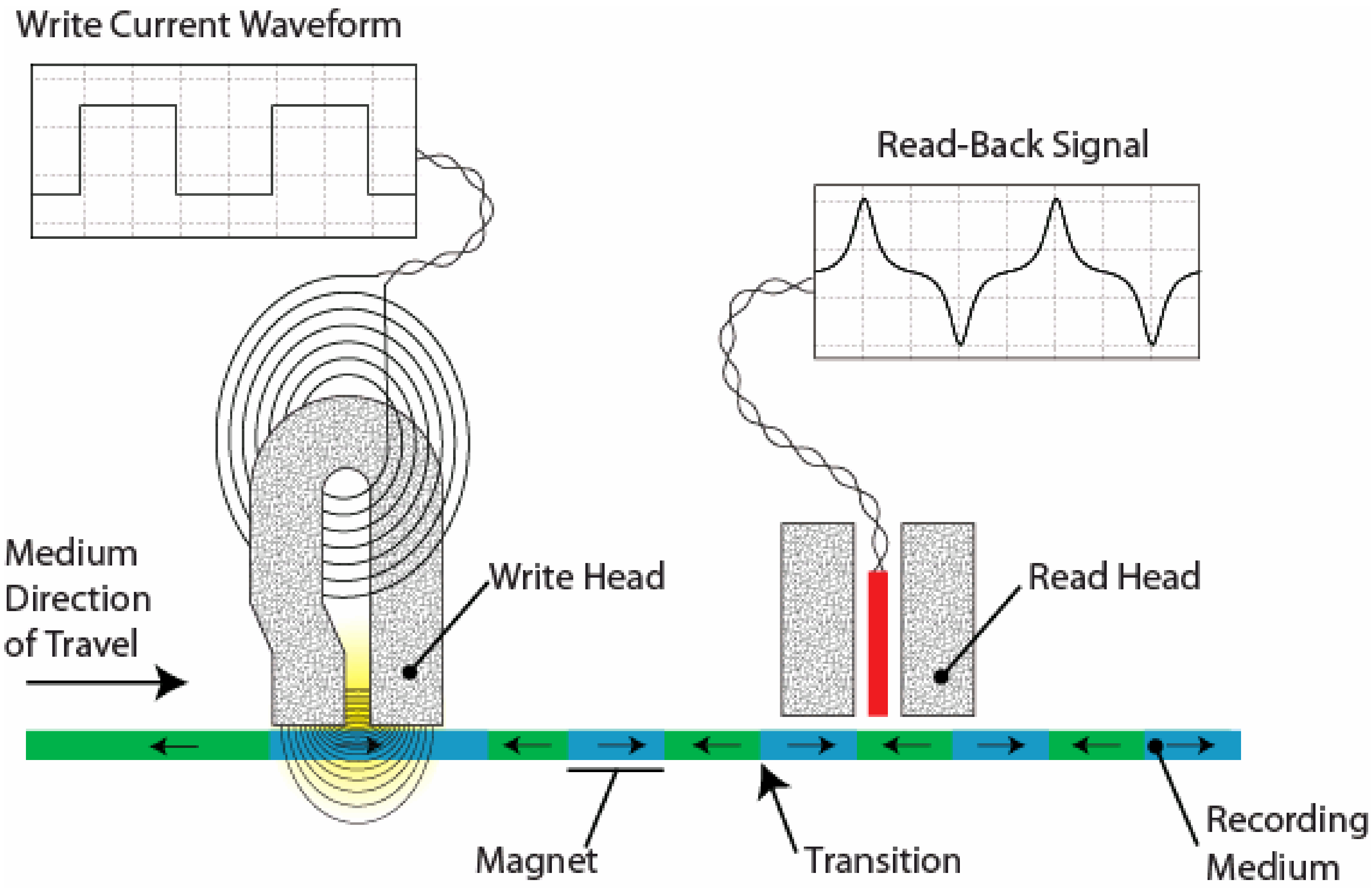
HITACHI
Inspire the Next



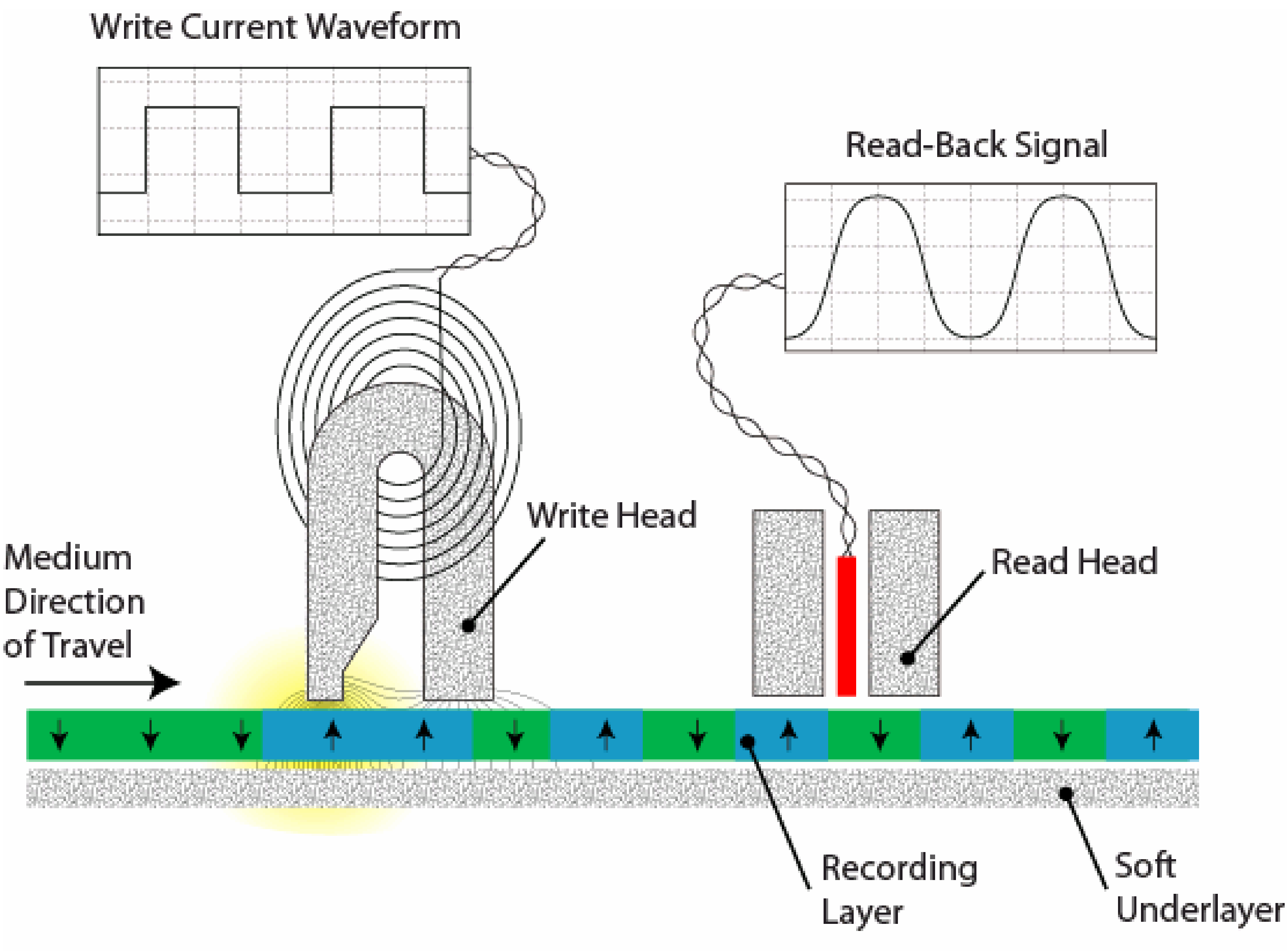
HITACHI
Inspire the Next

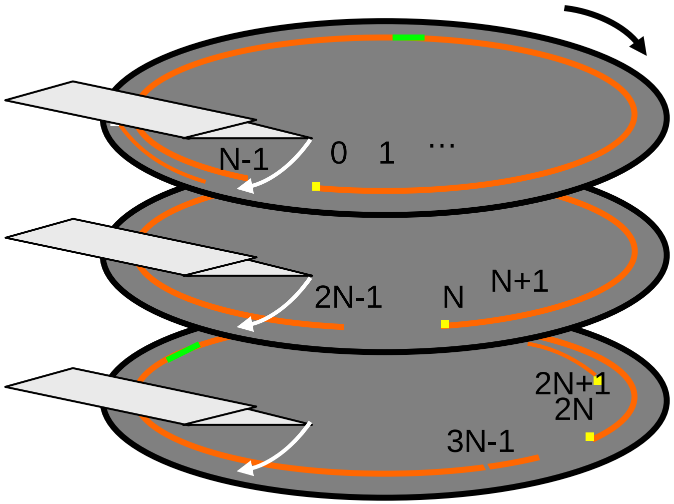
- HDDs as Block devices
- Sequential access and skew
- Zones and data layout
- Command Queues
- Write Cache
- Read Cache

HITACHI
Inspire the Next



HITACHI
Inspire the Next

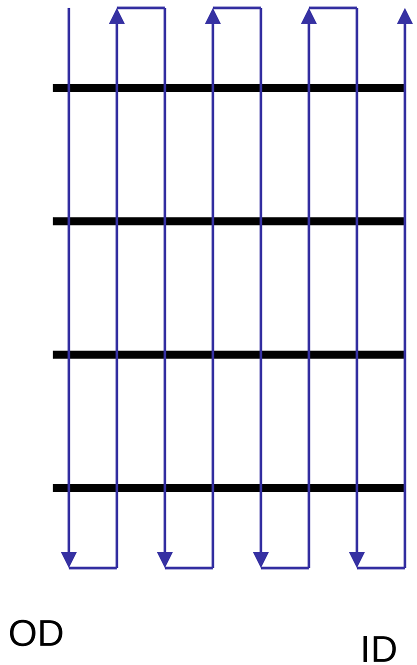


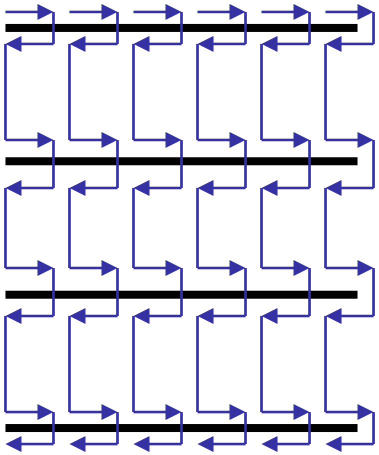


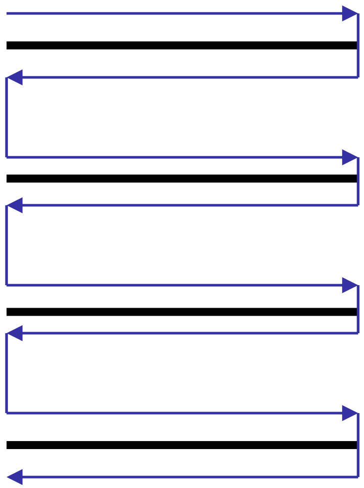
- HDDs are “block devices”
  - Access requests are for blocks of data, as opposed to individual bits or bytes.
  - HDD blocks are called sectors and are (almost always) 512 bytes in size.

The sectors in the HDD are represented by a linear address space. The addresses are often called Logical Block Address or LBA.

Ordering the Tracks







HDD Throughput Optimization

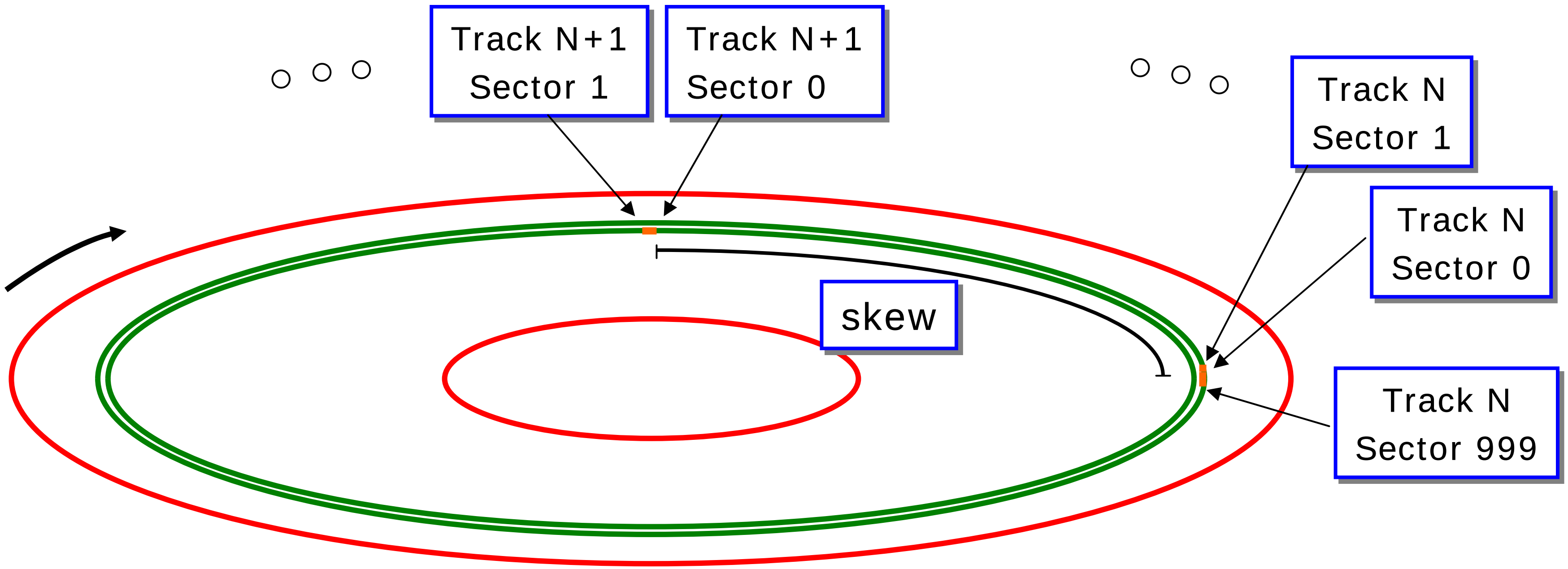
HDDs are designed to optimize sequential access. This is done by varying the position of the first sector in the track by an amount (called skew) proportional to the time required to switch tracks.

Example: (these are illustrative simplified numbers for easy computation)

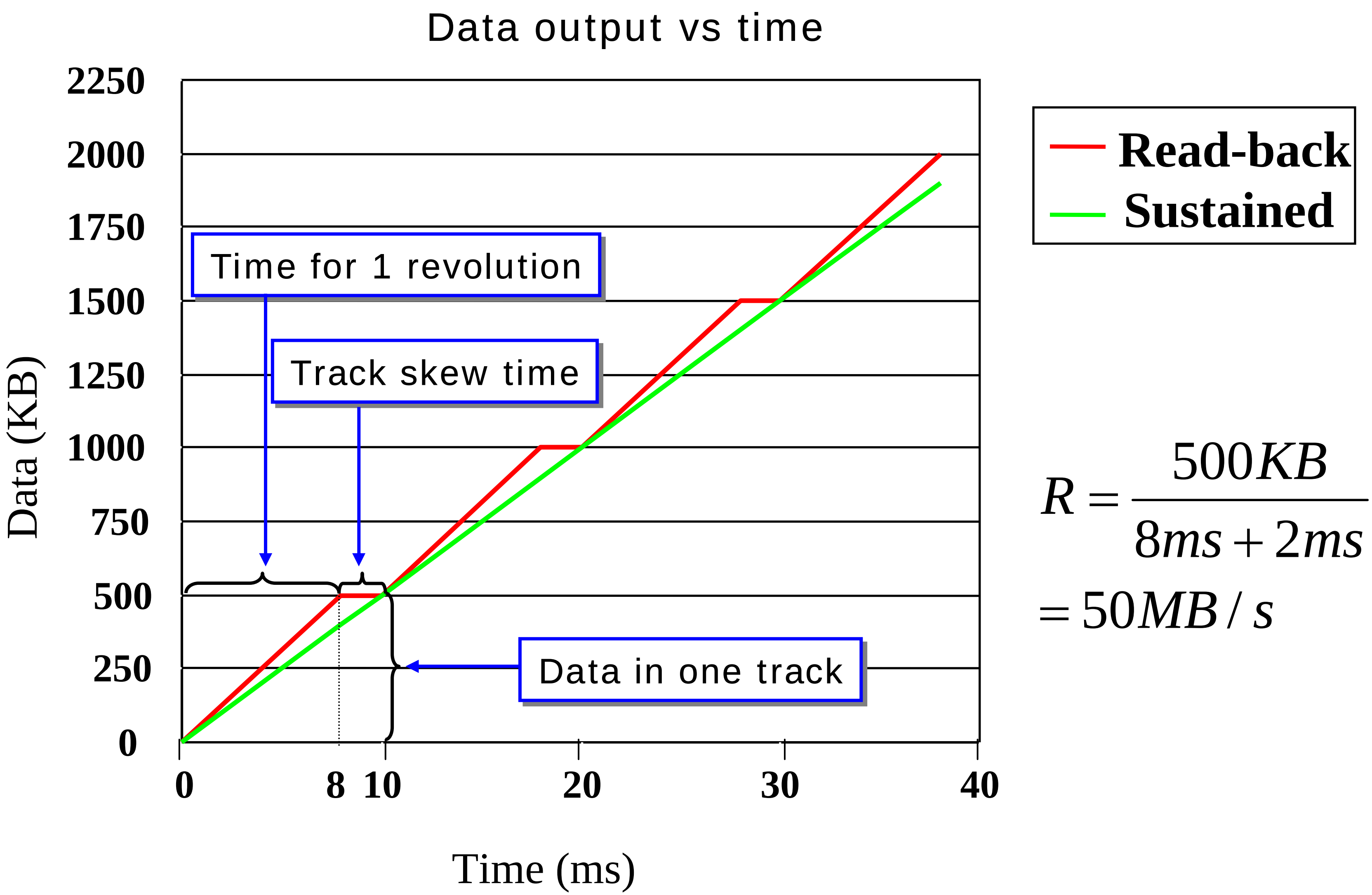
One revolution -> 8ms (7500 rpm).

Track switch time -> 1.6-1.9 ms. Use 2ms = ¼ rev.

Track size -> 1000 sectors.



Sequential Access



$$R = \frac{500KB}{8ms + 2ms}$$

$$= 50MB/s$$

HITACHI Inspire the Next

In an HDD, the number of sectors in a track varies from zone to zone. Hence, for each zone, z, we have S(z) sectors per track and therefore

$$R(z) = \frac{S(z)}{Tr + T_s(1)}$$

The average data-rate for the entire HDD with Nz zones and C(z) cylinders in zone z we have

$$\overline{R} = \frac{1}{D}\sum_{i=1}^{N_z} C(z_i)R(z_i) = \frac{\dfrac{\sum_{i=1}^{N_z} C(z_i)S(z_i)}{D}}{T_r + T_s(1)}$$

$$\therefore \quad \boxed{\overline{R} = \frac{\overline{S}}{T_r + T_s(1)}}$$

where $\overline{S}$ is the average # of sectors per track and D is the total # of cylinders.

We will be interested in the special case where all the zones have the same number of cylinders:

$$C(0) = C(1) = \cdots = C(Nz - 1) = C$$

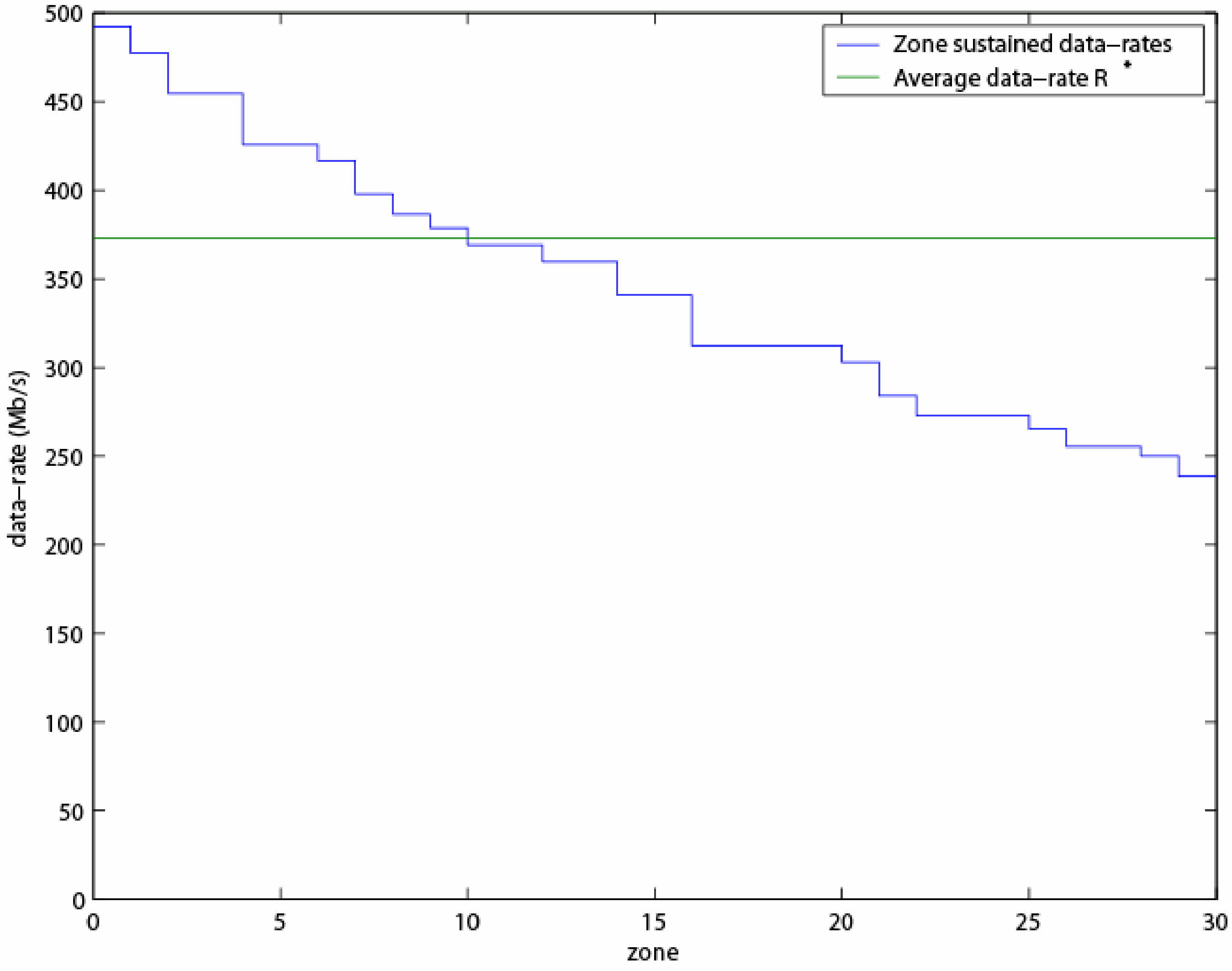
For this special case

$$\overline{S} = \frac{\sum_{i=1}^{N_z} S(z_i)}{N_z}$$

For a system with M streams, the per stream data-rate bound will be

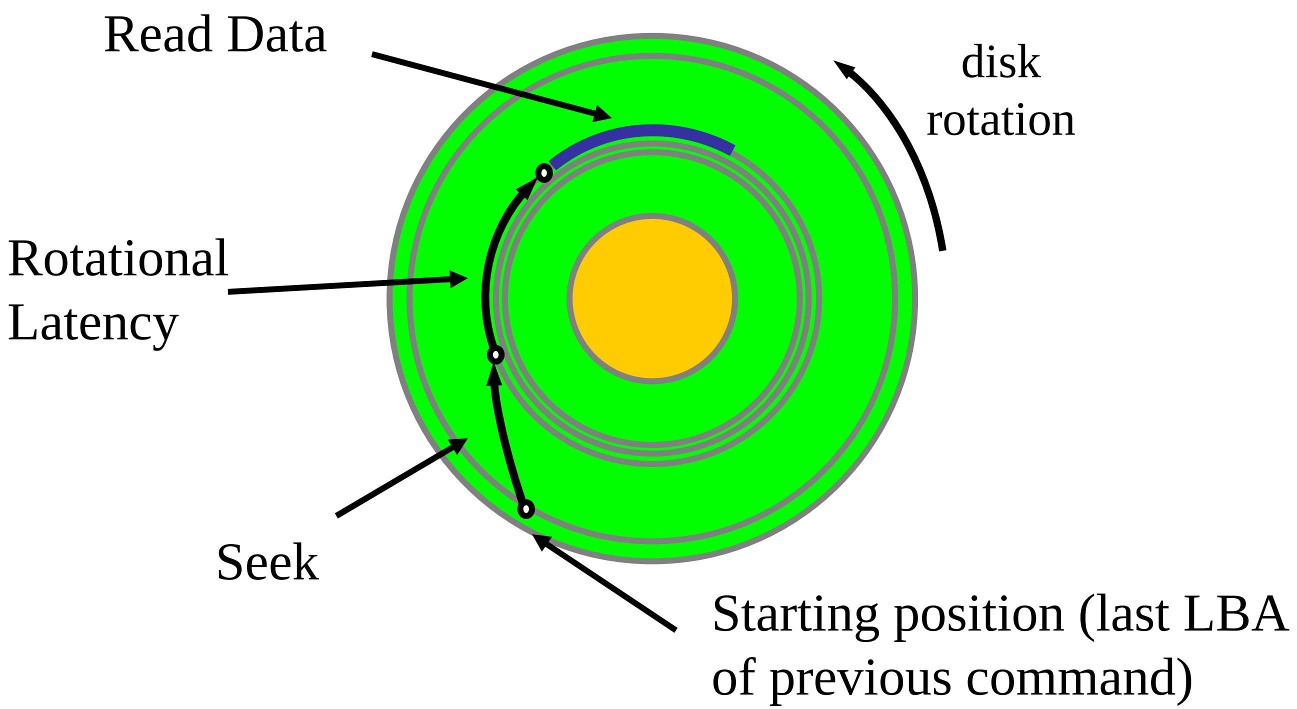
$$R_M^* = \frac{\overline{R}}{M}$$

Range of HDD Sustained Data-Rates for an Example 400GB HDD

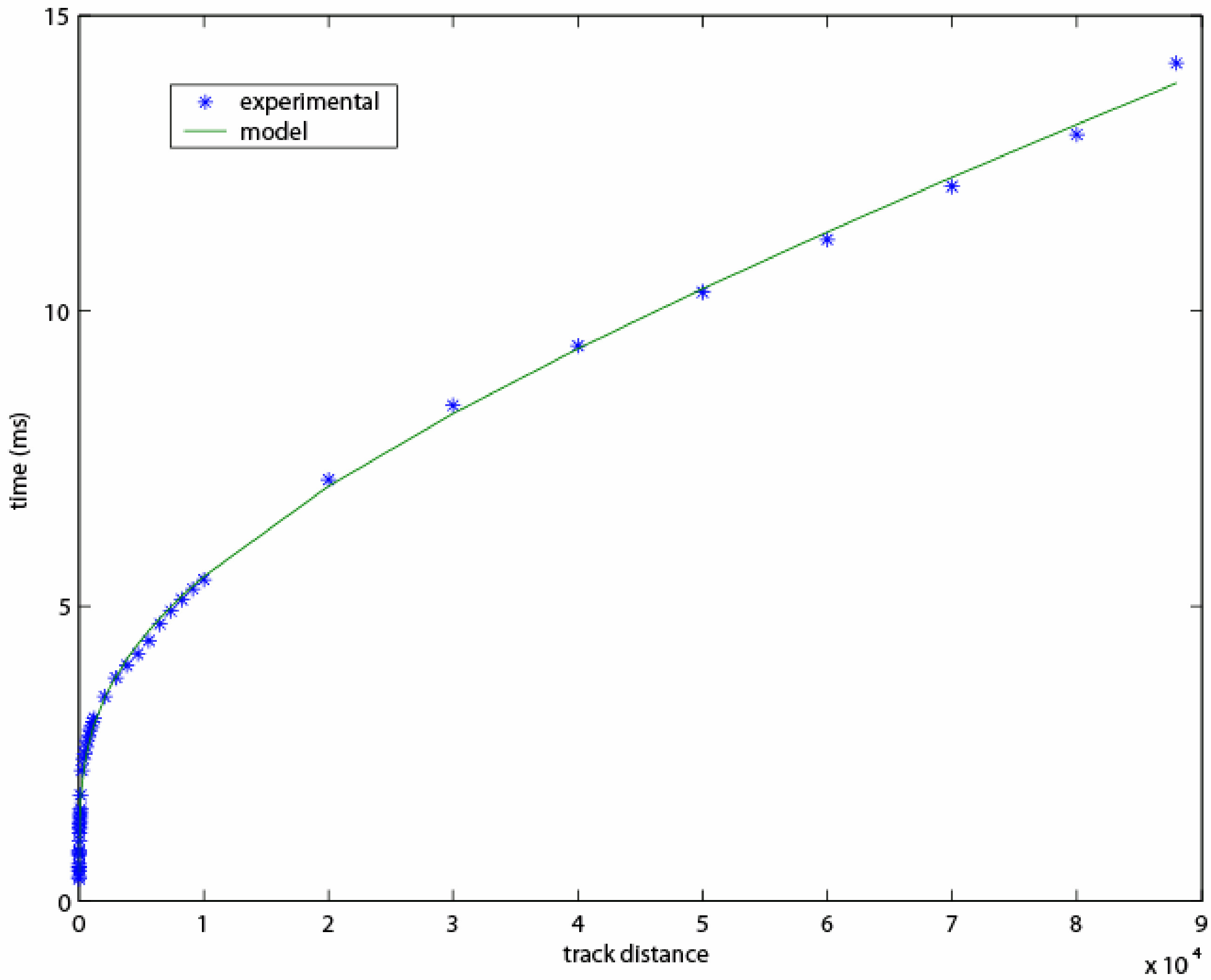


The total time for a read command can be expressed as

Tc = CMD Setup + Seek + Rotational Latency + Read + Re-read

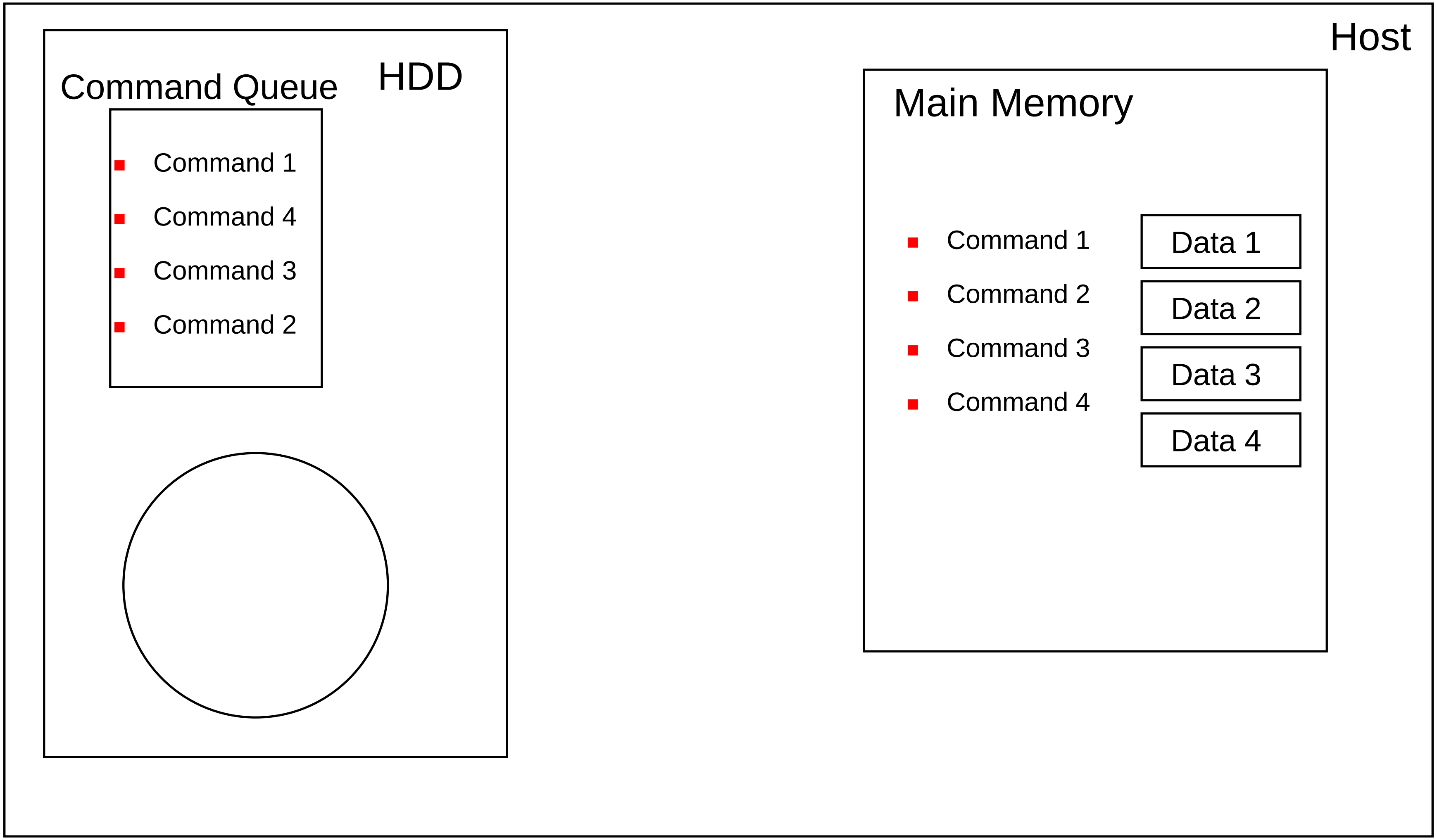


Seek Curve



experimental
model
time (ms)
15
10
5
0
0
1
2
3
4
5
6
7
8
9
track distance
x 10^4

Command Queuing

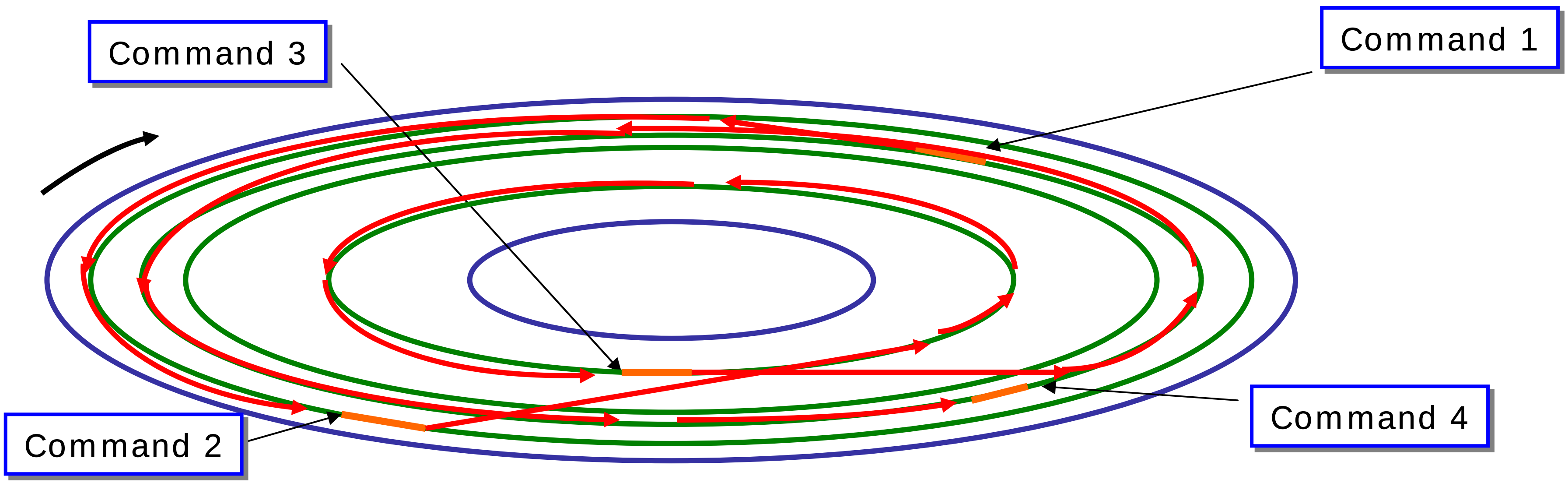


Queued Commands

Most modern 3.5” HDDs have queues that can be used to improve throughput. Imagine that the commands are received in sequential order, that is, command 1, command 2, command 3 and command4. Only the commands themselves are sent to the HDD. The data for any write commands is still in the host waiting for the command to start executing.

4	3	2	1

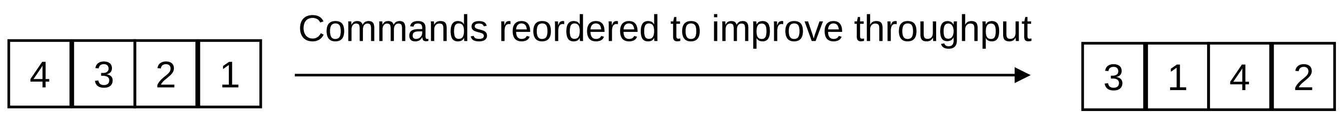
The figure below illustrates the execution of the commands in the order in which they were received, that is, it is a FIFO queue



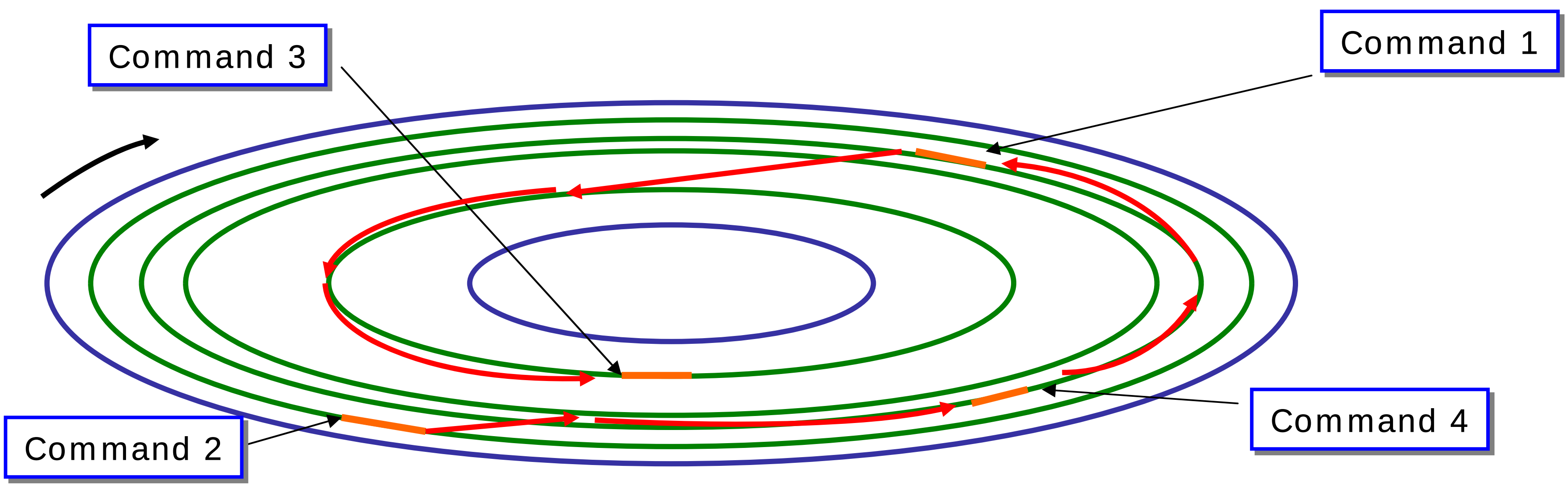
HDD Throughput Optimization

HITACHI
Inspire the Next

In order to improve throughput, the HDD typically will reorder the execution of the commands in an attempt to minimize the overall time required to execute all the commands.



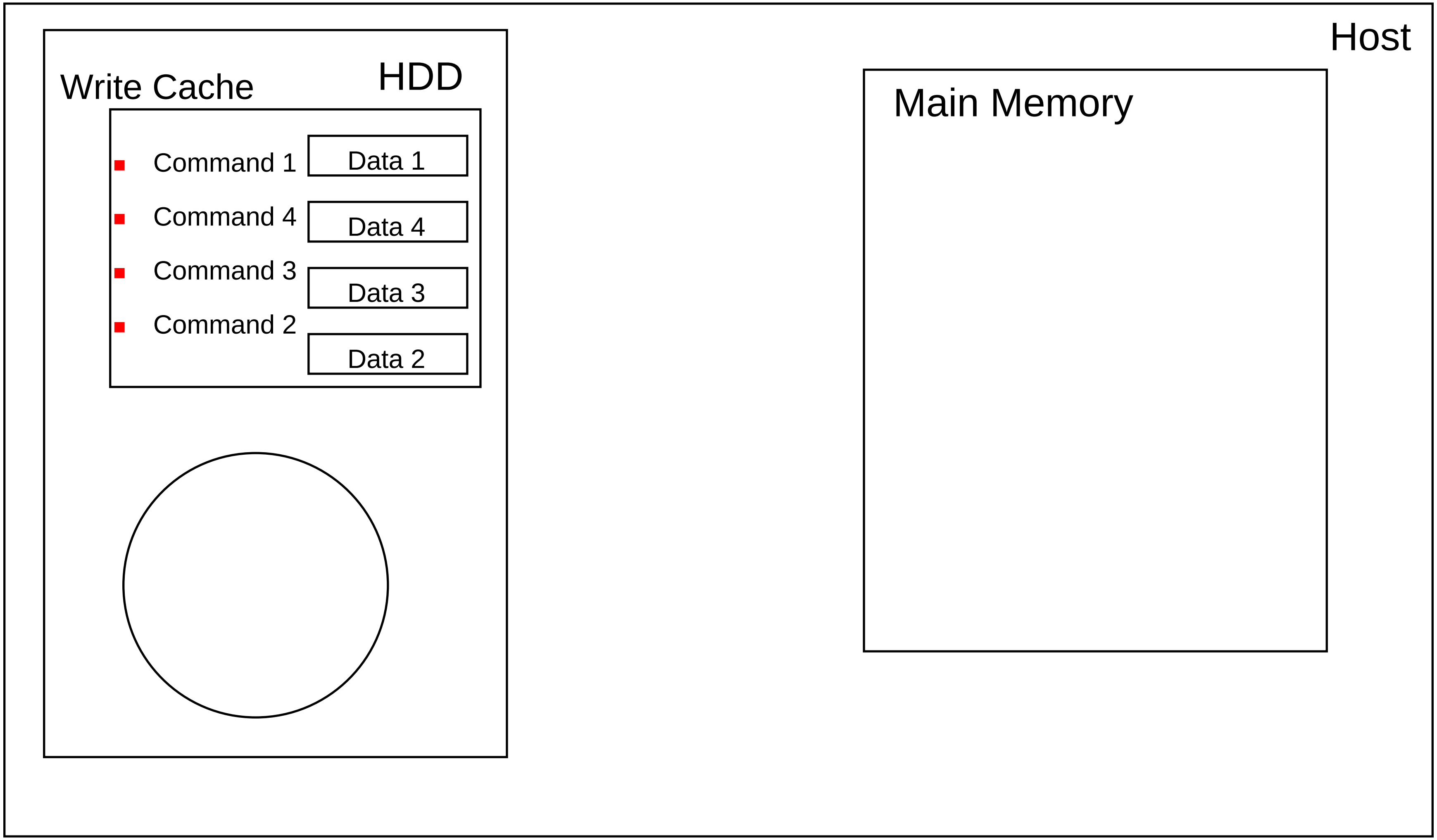
The figure below illustrates the execution of the commands in the throughput optimized order



Write Cache

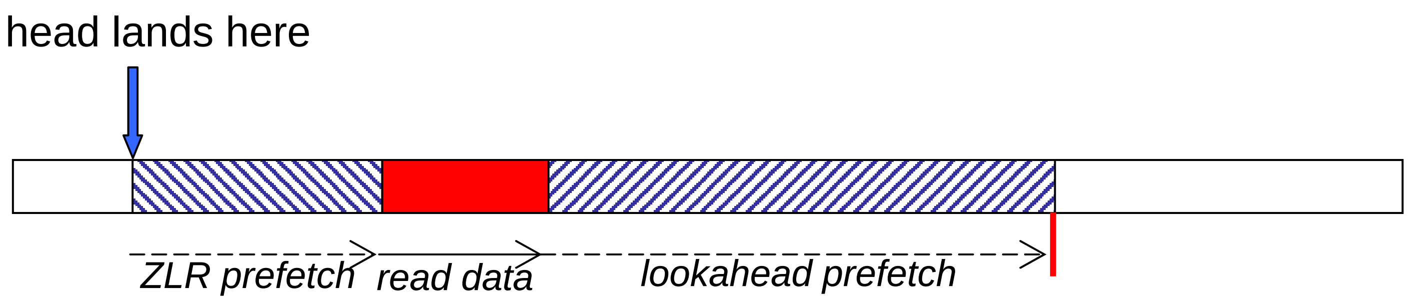
- Write cache functions similar to a Queue in some respects, but is completely transparent to the host.

- When write cache is enabled,
  - Write commands cause the data to be transferred to the HDD's buffer and the command returns "immediately".
  - Several commands can be queued up in the HDD's write cache.
  - After some a number of commands have been accumulated, or there is some free time, the HDD transfers (de-stages) the data to the disk.
  - The Host system has no control over the time when this will happen, other than being able to send a flush cache command that forces the HDD to de-stage the pending write commands immediately.
  - The write commands can be (and often are) reordered in order to improve throughput.


Host
HDD
Write Cache
Command 1
Data 1
Command 4
Data 4
Command 3
Data 3
Command 2
Data 2
Main Memory

Read Cache

- The read cache in an HDD serves the same purpose as the memory caches in microprocessors. In addition, the cache in the HDD is also used for some opportunistic predictive reading.
- As shown in the figure below, the HDD can read (configurable) a certain amount of data located immediately before and after the data being requested.



Part III: HDD Performance Issues for AV Applications

- Measures of Performance
- Parameters affecting performance
- Methods of access and their impact on performance
- Methods of data storage allocation and impact on performance

DVR vs PC

Requirements from a user's point of view:

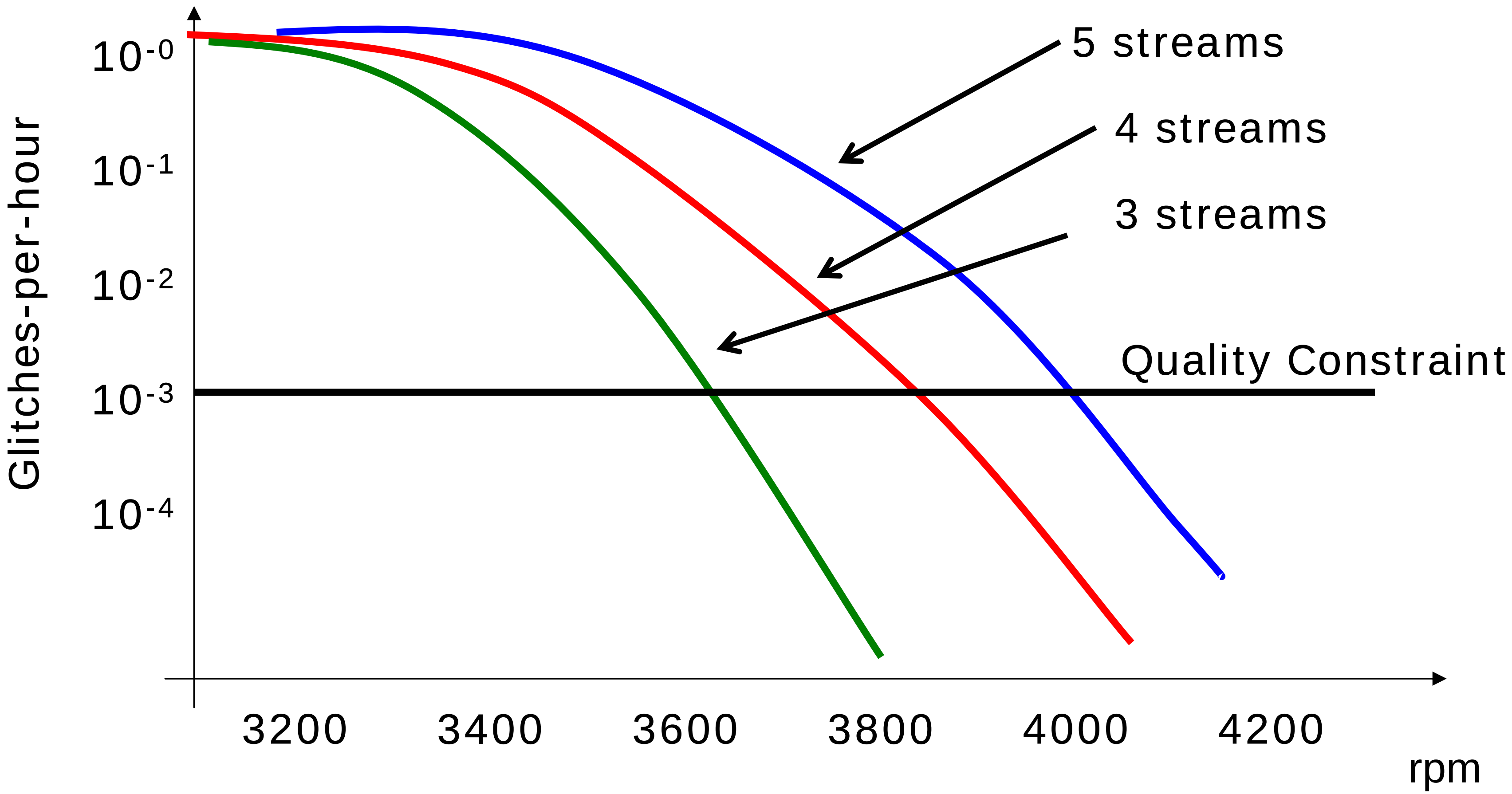
Fixed hard constraints

HDD Parameter	PC	DVR
Capacity	How much data (files, programs) can I store?	How many videos can I store?
Performance	Average data-rate. Application load time Time to boot the OS.	How many streams can it support?
Reliability	Likelihood of data loss. Likelihood of complete drive failure.	How often visible glitches occur?

Measure of Quality

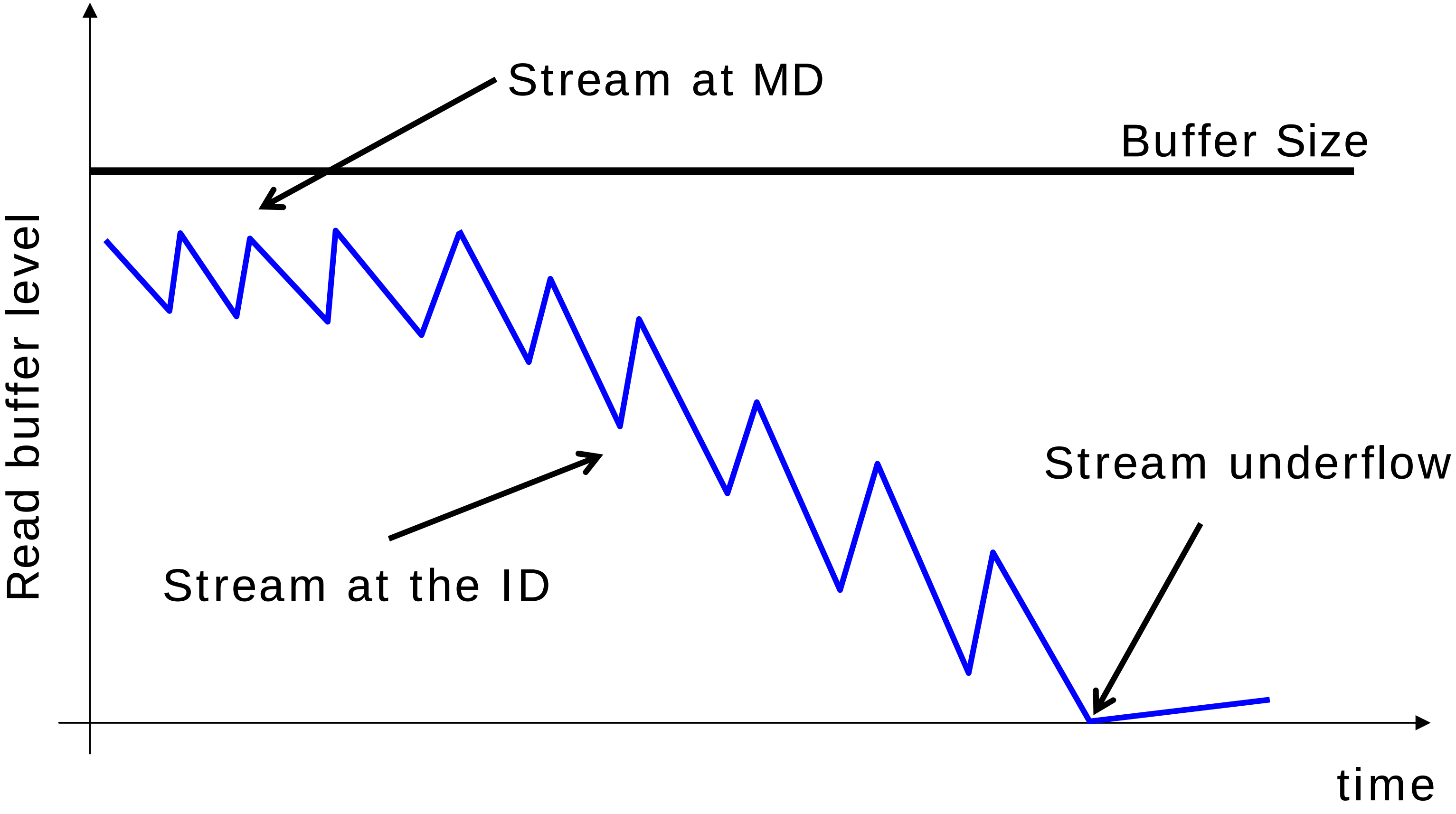
System Performance

Example: How does the system behave in terms of rpm?

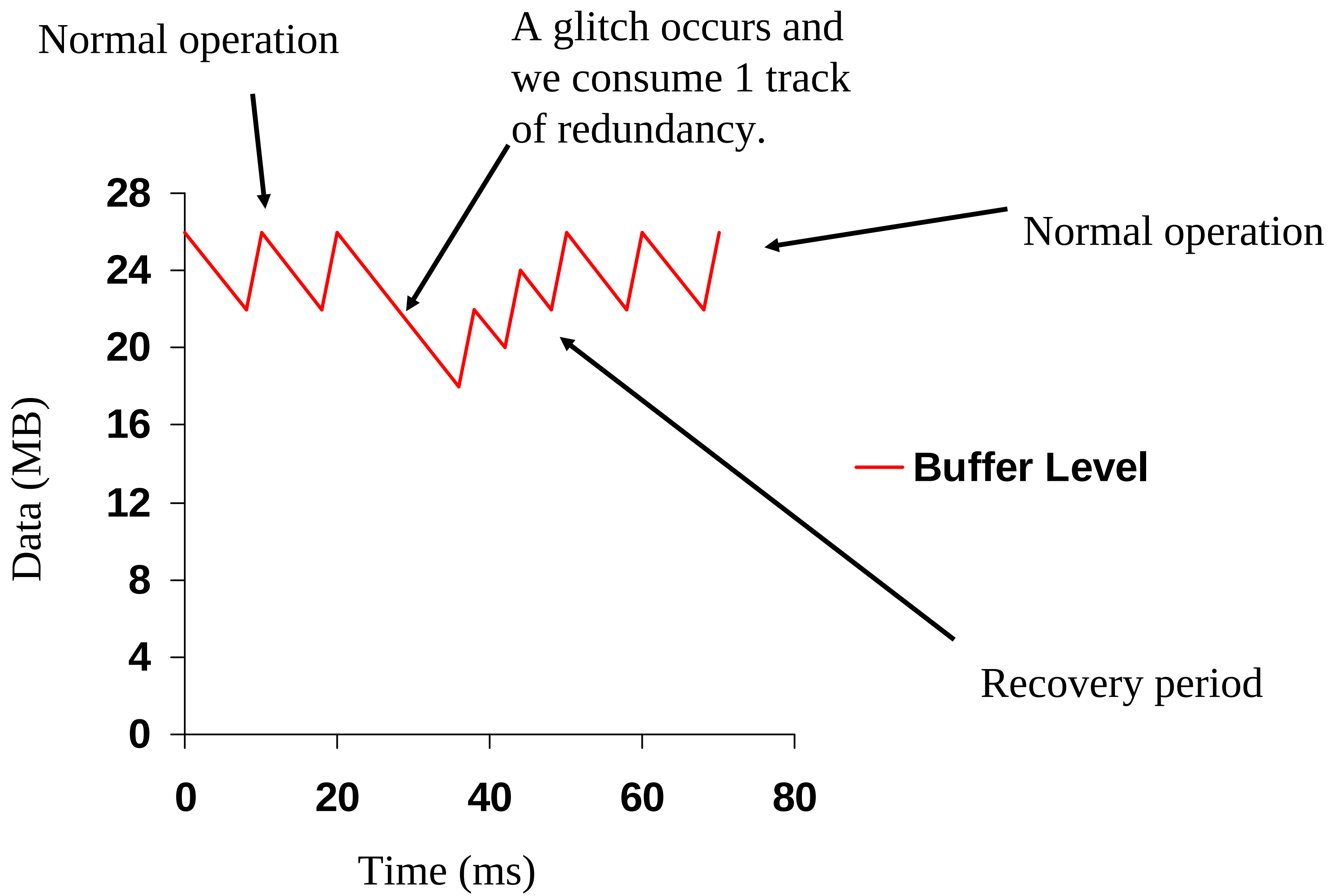


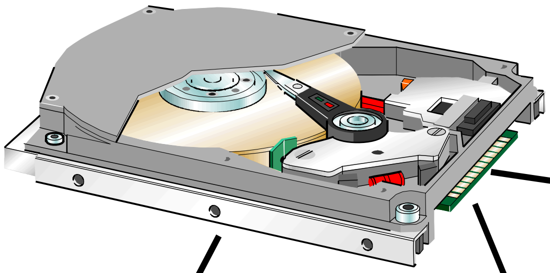
Effect of Storing Algorithm

Depending on how files are stored, the stream might get “stuck” at the ID where it is more vulnerable to glitches.



Buffer Level





Parameters affecting mainly Latency / consistency of response

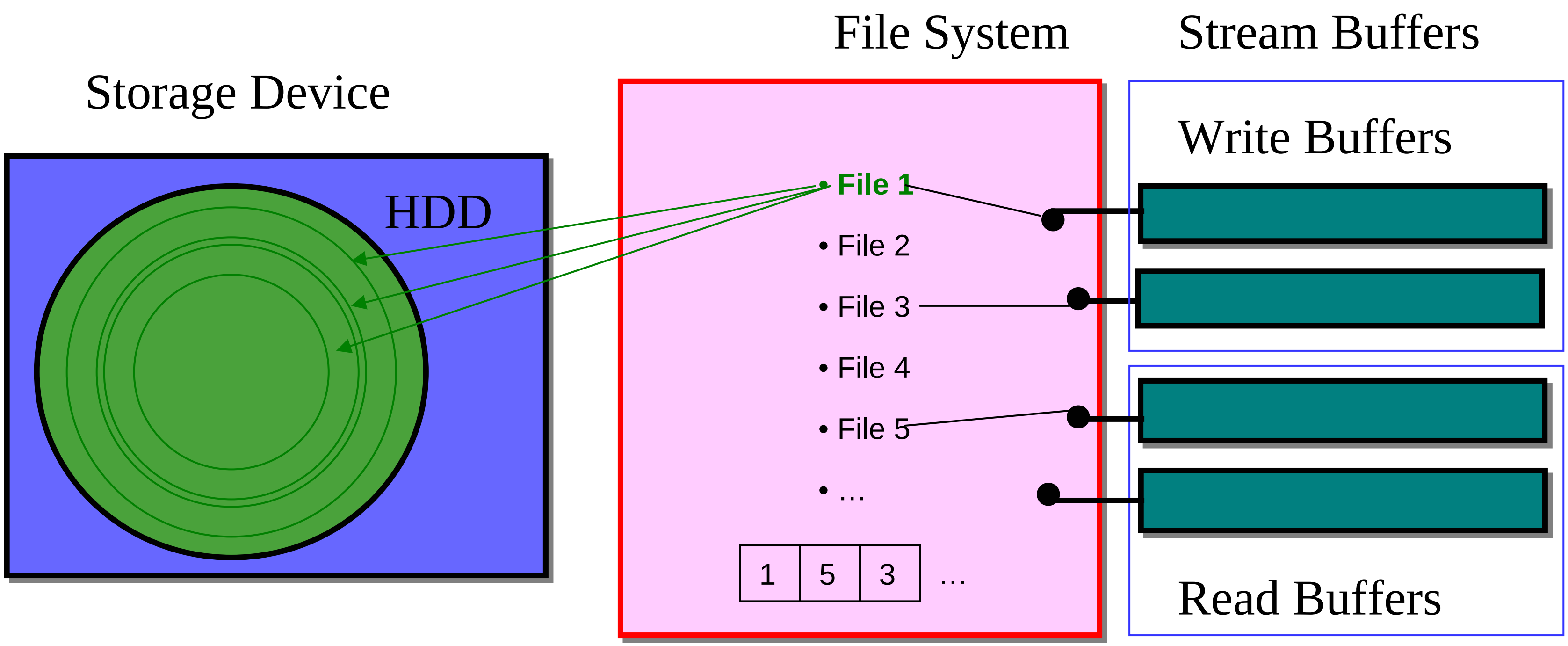
- Defect management
- Error recovery algorithms

Parameters affecting both throughput and Latency / consistency of response

- Write cache
- Queuing
- Read cache
- Physical layout

Parameters affecting mainly throughput

- RPM
- Seek curve



Data Layout

- •How to organize data on the surface of the disk.
- •For each new file, decide which free blocks to use for storing the new file.

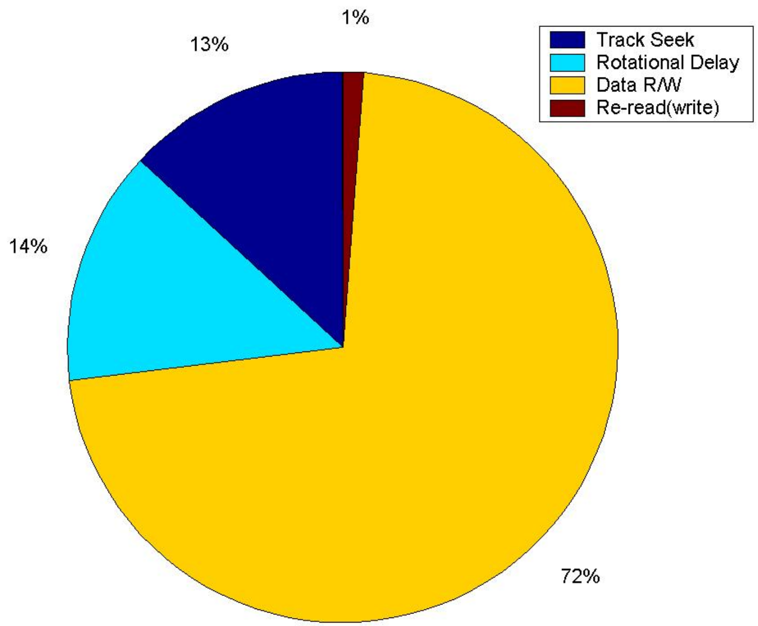
Scheduling Algorithms

- •Determines in what order access requests should be satisfied.
- •Algorithms will control the trade-off between QoS and throughput.

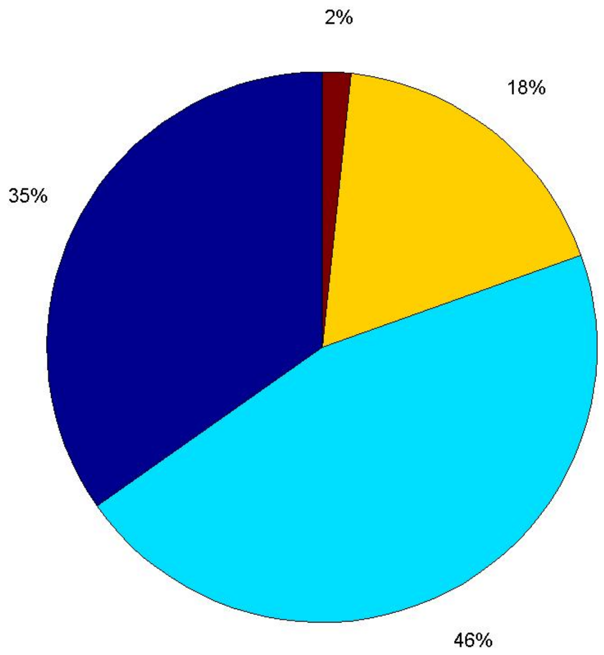
Metadata, File Structures and Interface

- •How should files be represented.
- •Allocation unit size
- •Size of block access

(3 streams HDTV)



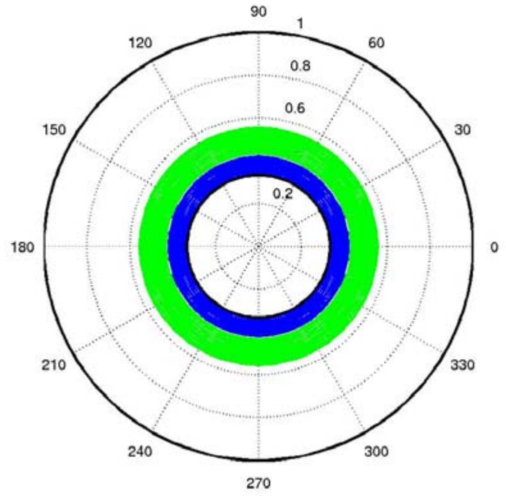
Block Size = 1024K

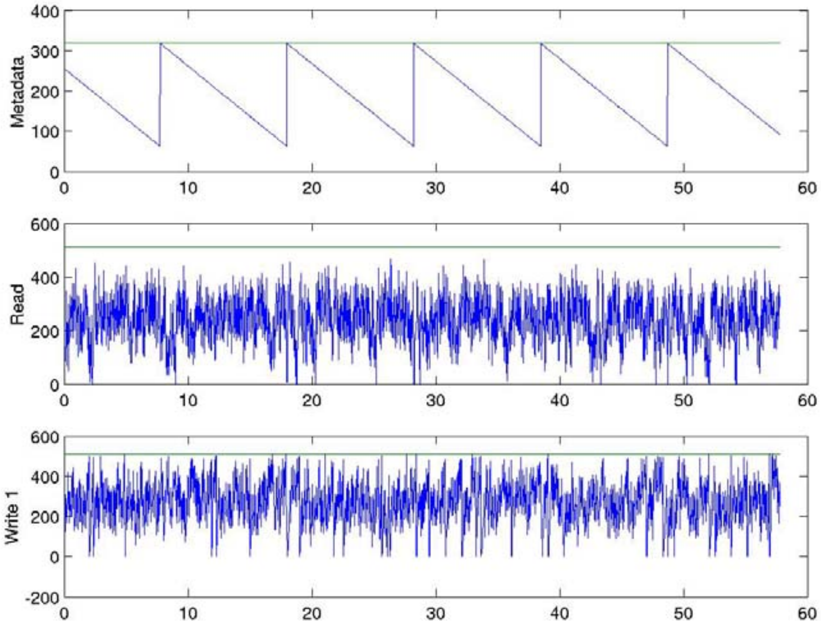


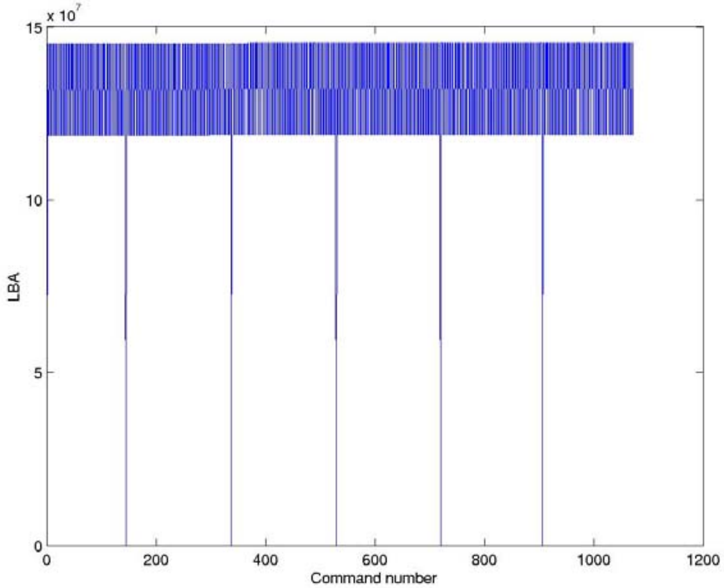
Block Size = 64K

HDTV 2 Streams ID

HITACHI
Inspire the Next

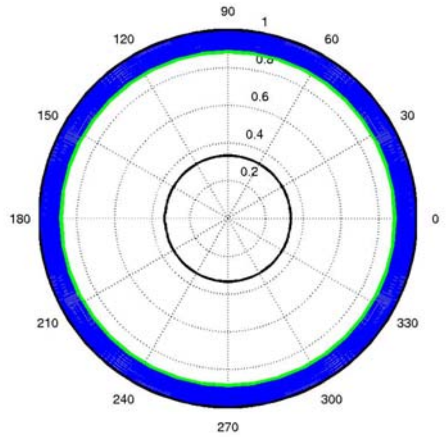


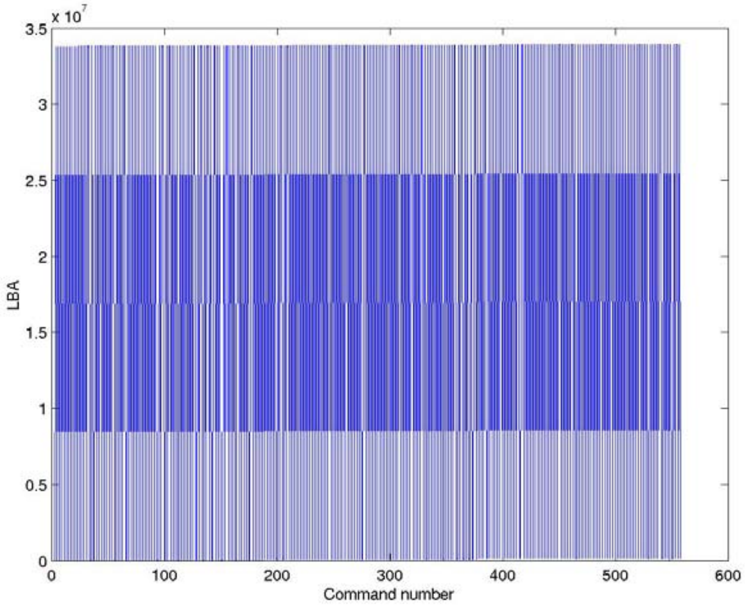


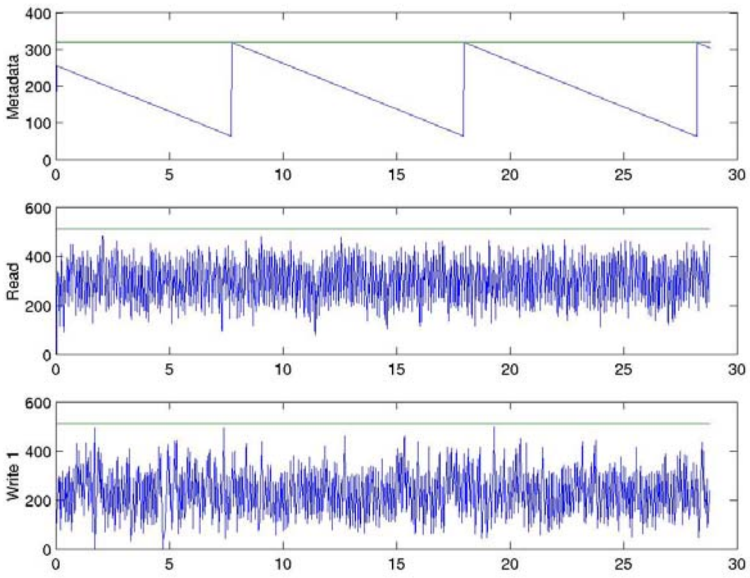


- Rpm 3600, Sb = 512(sectors)
- Poor performance
- High number of glitches

HDTV 2 Streams OD

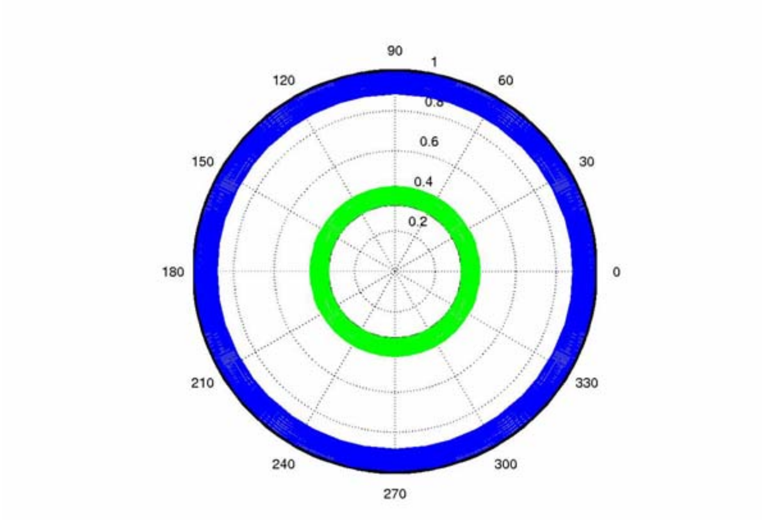


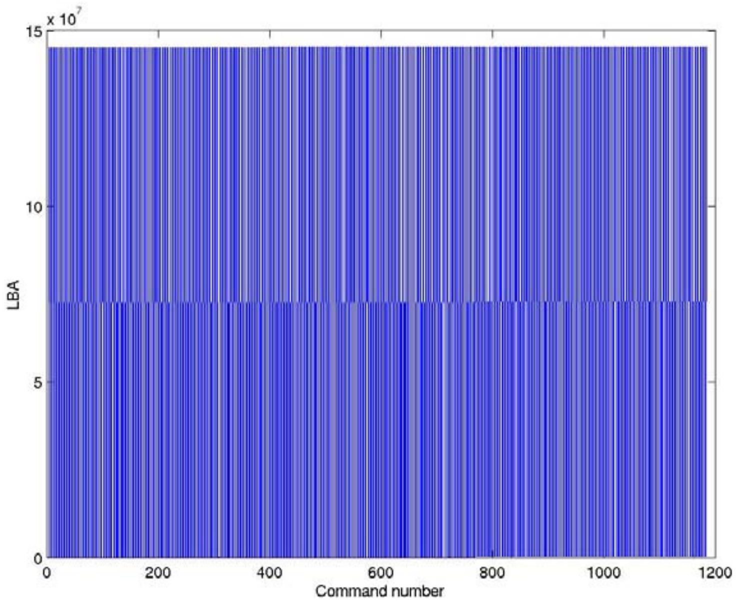


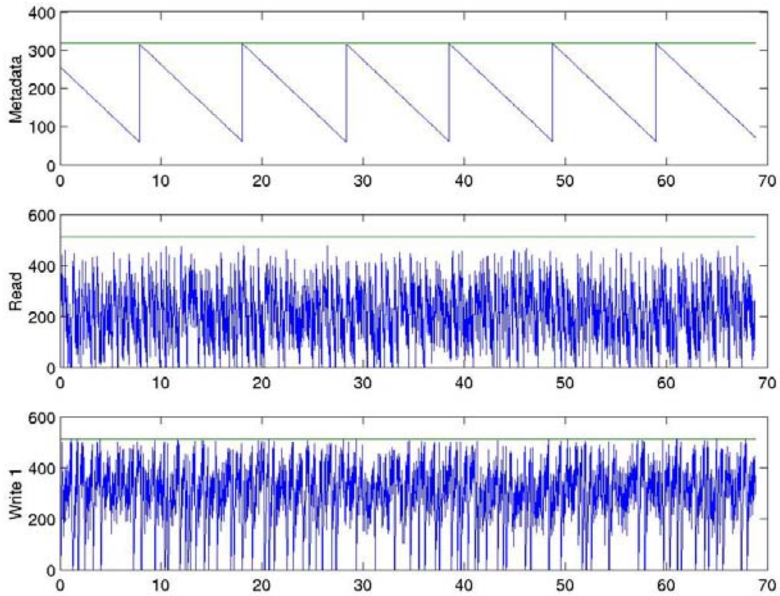


- Rpm 3600, Sb = 512(sectors)
- Reasonable performance
- No Glitches

HDTV 2 Streams ID - OD







- Rpm 3600, Sb = 512(sectors)
- Very Poor performance
- Very large number of glitches

- In order to maximize throughput stream requests can handled in the order that minimizes the intra-stream switching time.
  - That is, the block requests are sorted in terms of LBA and accesses are made in this order.
  - This scheduling is called SCAN.
  - The latency for a particular stream can be very large and therefore large buffers may be required.
- In order to minimize latency requests can be handled in the order of priority
  - That is, the stream who's buffer is about to overflow (underflow) is serviced first, and so on.
  - This is called EDF (earliest deadline first).
  - Since large seeks can be necessary for guaranteeing low latency, this solution may have low throughput.
- There are several algorithms in between the two extremes.
  - SCAN-EDF
  - Group Sweep Scheduling (GSS)
  - Round-robin
- ATA-7 can help handle mixed traffic
  - Put deadlines for best-effort traffic so as to prevent it from disrupting real-time traffic.
  - Put deadlines in real-time traffic to avoid one stream depleting all streams.

HITACHI Inspire the Next

In an HDD, the number of sectors in a track varies from zone to zone. Hence, for each zone, z, we have S(z) sectors per track and therefore

$$R(z) = \frac{S(z)}{Tr + T_s(1)}$$

The average data-rate for reading the entire HDD from OD to ID. If Nz is the number of zones and C(z) the # of cylinders in zone z we have

$$\overline{R} = \frac{1}{D}\sum_{i=1}^{N_z} C(z_i)R(z_i) = \frac{\frac{\sum_{i=1}^{N_z} C(z_i)S(z_i)}{D}}{T_r + T_s(1)} = \frac{\overline{S}}{T_r + T_s(1)}$$

where $\overline{S}$ is the average # of sectors per track and D is the # of cylinders.

The data-rate of the slowest zone (ID) is often a practical bound on the performance of several systems.

$$\overline{R}_{ID} = \frac{S(z_{ID})}{T_r + T_s(1)}$$

HITACHI
Inspire the Next

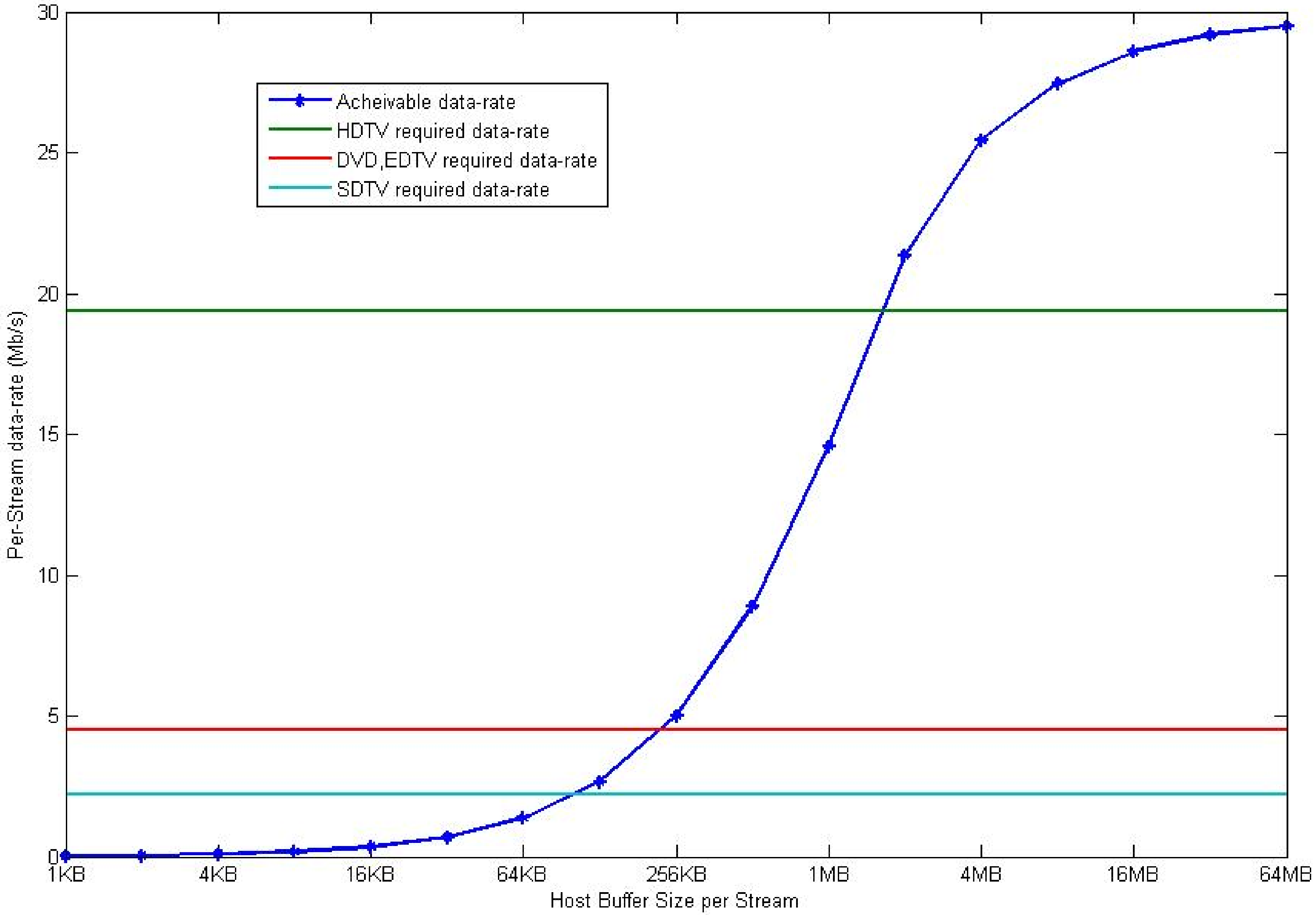
If we use the HDD in a “regular” file system, but with the read/write requests always for S_B bytes, then the following worst-case data-rate is achieved

$$\underline{R}(S_B) = \frac{S_B}{T_{CMD} + T_s(d_{z_0^*}) + \tau + \frac{1}{2}\left(\frac{S_B}{R(z_0^*)} + \frac{S_B}{R_{N_z-1}}\right)}$$

where

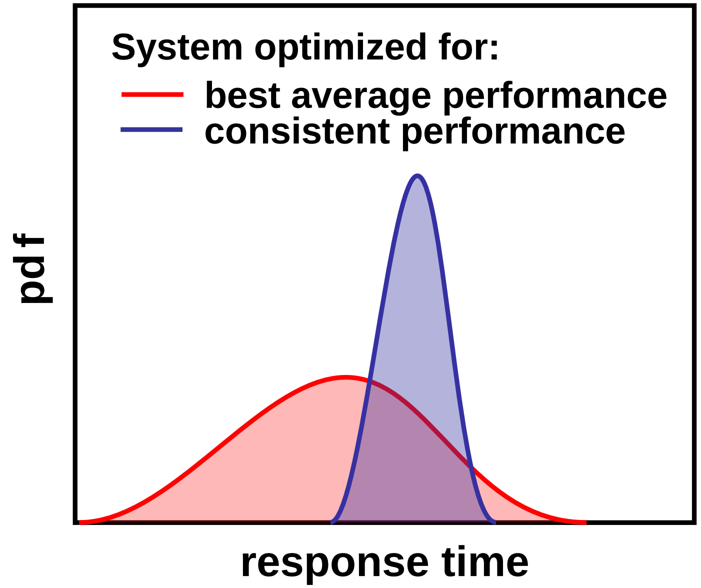
$$z_0^* = \underset{z_0 \in \{0,1,\ldots,N_z-1\}}{\arg\max} \; 2T_s(d_{z_0}) + \frac{S_B}{R_{z_0}}$$

Performance Lower Bound for 8 Streams on an Example 400GB HDD



HITACHI Inspire the Next

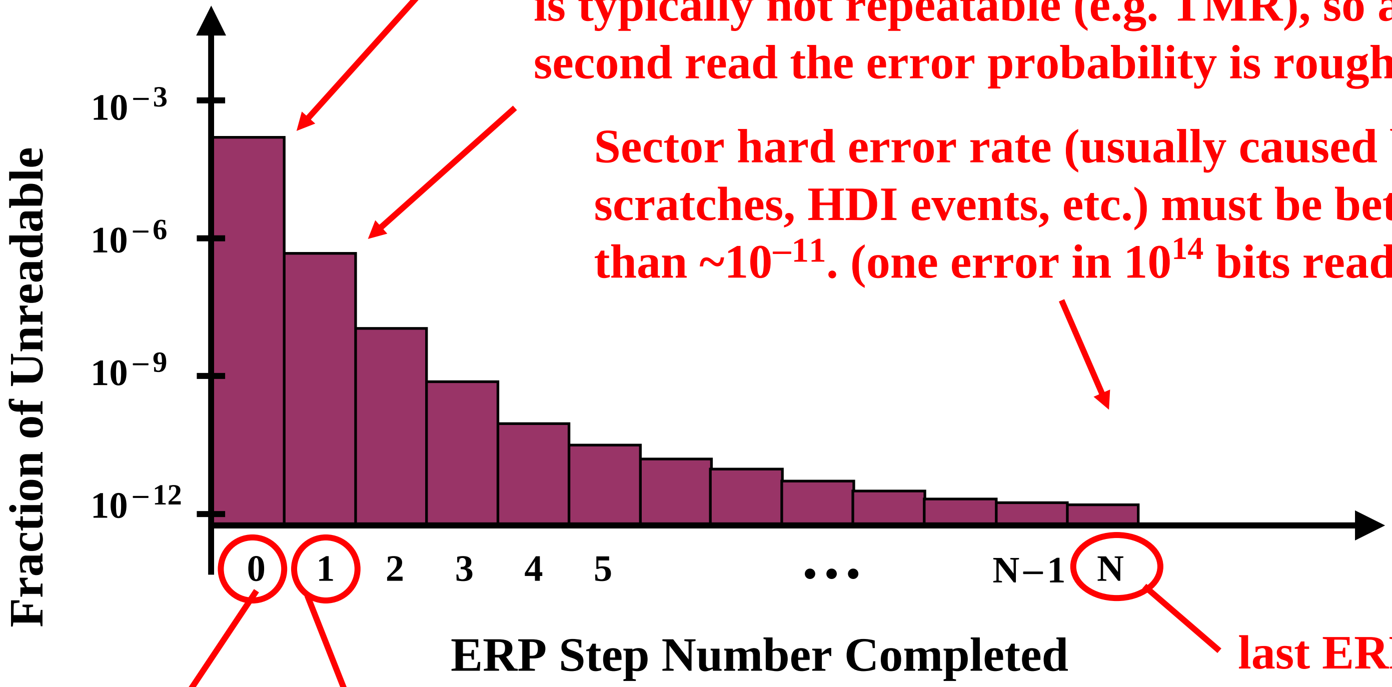
- For desktop HDDs, reliability is usually specified through:
  - A Hard Error Rate (HER).
  - A Mean Time Between Failures (MTBF).
- The hard error rate indicates the probability of losing a small amount of data (usually a sector) after completion of an error recovery procedure. e.g. One error per 10^{14} bits read after ERP.
- The MTBF specifies the catastrophic failure rate for a population of HDDs.
- Another important metric for AV performance is the command completion time distribution.
  - IT applications require good average performance, but can have a wide variance in command completion times.
  - AV applications may trade off average performance to achieve more consistent response times.



Sector failure rate on the first read depends on seek settling time.

The principal failure mechanism on the 1st read is typically not repeatable (e.g. TMR), so after the second read the error probability is roughly

Sector hard error rate (usually caused by scratches, HDI events, etc.) must be better than ~10^{-11}. (one error in 10^{14} bits read)



last ERP step

first

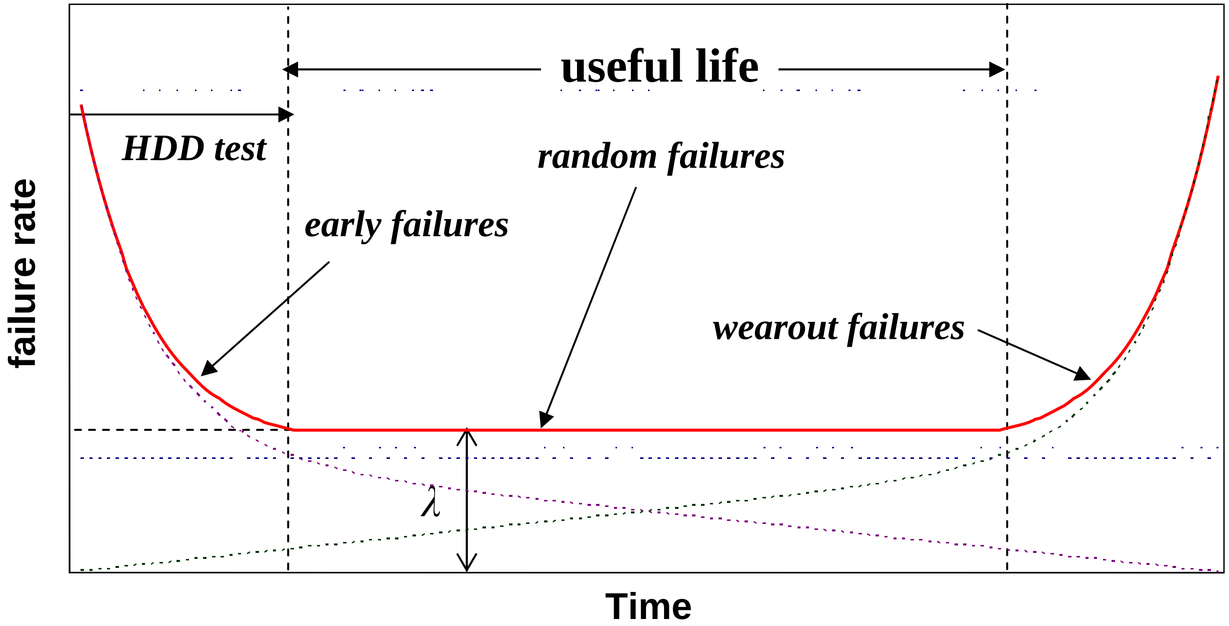
second read

Note: This is an illustration only, not real

Mean Time Between Failure (MTBF)

- **What does it mean to have a MTBF of 1.2M hours (137 years)?**
- **During the period of useful life of the HDD, we assume an exponential failure density function *f(t)* = λ *exp (–λx)* with a constant failure rate λ.**
- **The average time to failure is MTBF = 1/λ = 1.2M hours.**
- **In reality, we do not expect the HDD to be in use for 1.2 M hrs, so the MTBF is used to provide the expected rate for a large population of HDDs.**
- **For example, with 1000 HDDs with 1.2M hour MTBF running for one year, the expected number of failures is 1000 x 365 x 24 / 1.2M = 7.3.**

Reliability "Bathtub" Curve



Power Consumption And Thermal Management

HITACHI
Inspire the Next

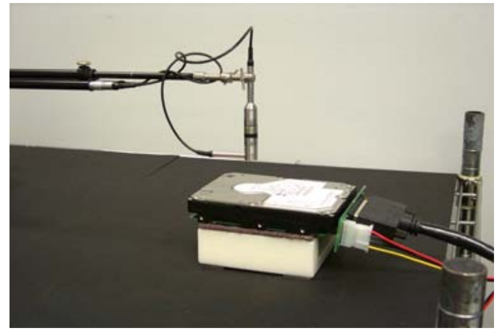
- **Additional power states:**
  - **Unload Idle**
    - 7200 rpm, head unloaded
    - Recovers in ~0.7 sec.
  - **Low RPM Idle**
    - 4500 rpm, head unloaded
    - Recovers in ~4 sec.
- **Quiet seek mode saves about 2.4 W.**
- **SATA power consumption is worse than PATA by about 0.6W. Expected improvement through SATA link power management.**

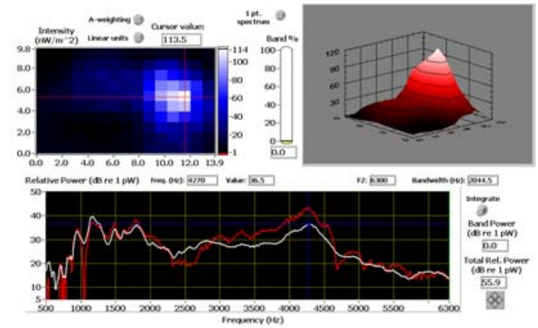
Deskstar 7K250 (PATA) Power	
Power Mode	**Power (W)**
Idle	5.24
Low RPM Idle	2.72
Unload Idle	4.04
Random R/W 30% seek, 45% R/W, 25% idle	9.7
Silent Random R/W	7.3
Standby	0.93
Sleep	0.68

- **Thermal management will impact HDD reliability.**
  - Thermal specification typically 5 to 55 degrees C (operating). Top cover can be at 60 degrees C.
  - Mounting and airflow are extremely important.

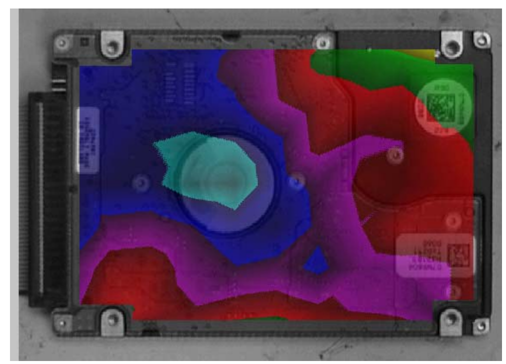
HDD Acoustics Measurements

Anechoic sound chamber





Sound intensity mapping locates noise



Scanning LDV for mode analysis

Binaural head for sound quality

HITACHI
Inspire the Next

Typical HDD Acoustic Specifications (A-Weighted Power)

	Mobile	Desktop	Server
Idle	23 dBA	31 dBA	34 dBA
Seek	27 dBA	35 dBA	44 dBA

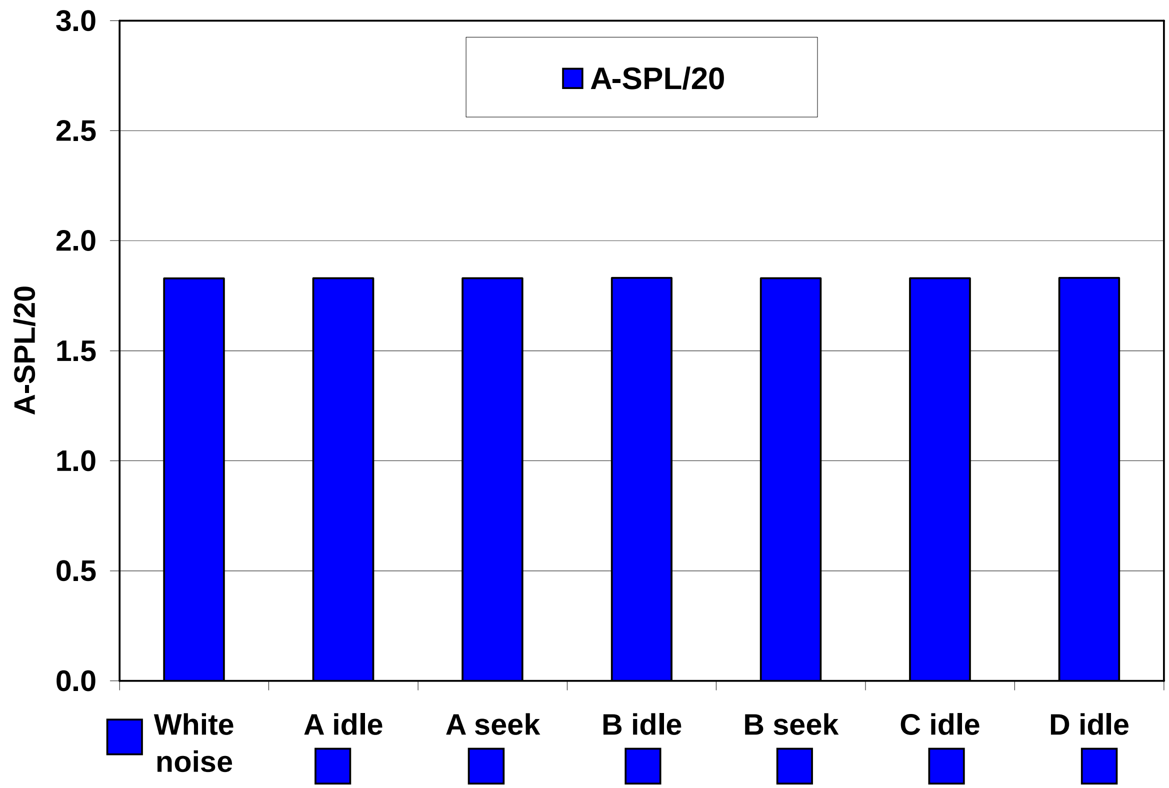
Typical Sound Levels (A-Weighted Pressure)

Acoustic Environment	Sound Pressure
Threshold of hearing	0 dBA
Rustling Leaves	10 dBA
Quiet Library	35 dBA
Business Office	65 dBA
Busy Street	85 dBA
Threshold of Pain	140 dBA

In a hemispheric anechoic chamber at 1m from the noise source, subtract 8 dBA to get the sound pressure

Sound Quality

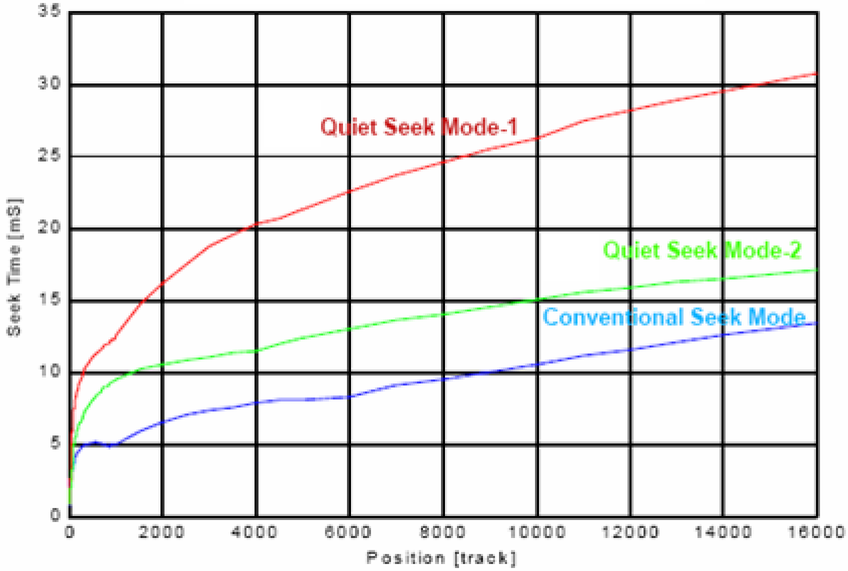
- Current specification uses A-weighted sound power with a tone penalty.
- However, customer preference is more dependent on perception of disk drive noise rather than simply volume
- Sound quality (SQ) addresses many aspects of acoustic annoyance: loudness, tones, sharpness, roughness, etc.



Quiet Seek

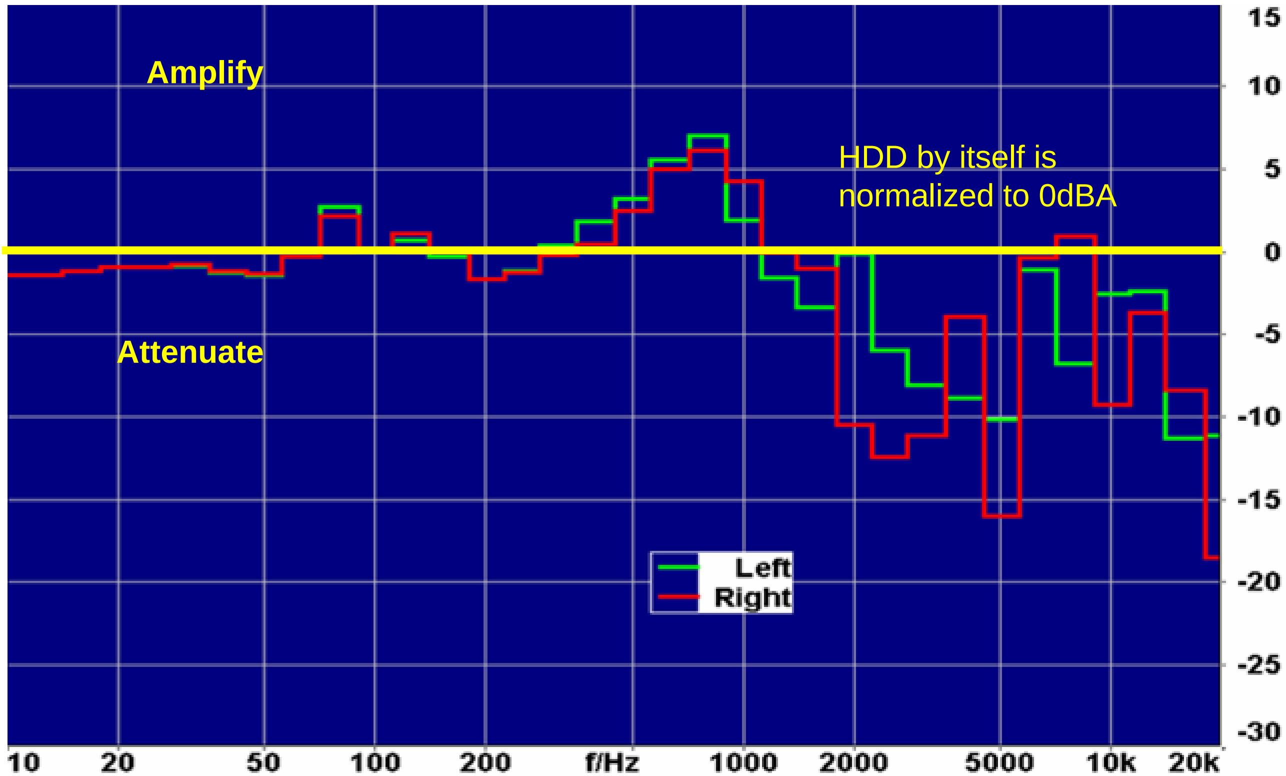
- Depending on the specific product, quiet seek provides an improvement of up to 5 dBA in emitted sound power compared to performance seek mode, in addition to a significant improvement in sound quality.
- Use of quiet seek may also reduce power consumption by ~2.5 W compared to performance seek.
- Quiet seek also increases full volume seek times by about a factor of two.
- For AV applications, with large block IOs, quiet seek provides a significant benefits with minimal reduction in performance.

Seek Length Versus Seek Time
(Data for Illustration Only)



- Incorrect mounting of the HDD can cause the system box to act as a resonator for the HDD acoustic noise.
- Careful mounting and system design are required in order to create a low noise system.

Example: Impact of a system box on sound pressure level.



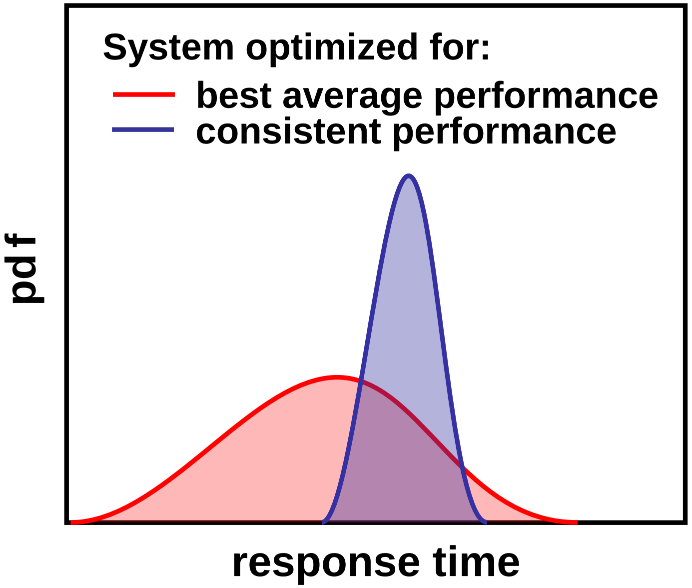
Measurements from typical CE devices in a hemi-anechoic chamber.

Source	Sound Power (dBA)
background noise level	26.2
CD player (playing)	26.4
DVD player (playing)	27.5
VCR (playing)	36.9
PVR (playing)	37
PVR (rewinding at 15x)	37.3
PVR (skipping at 300x)	38.2
VCR slow FF (from play)	40.8
small cheap TV	44
VCR fast FF (from stop)	55.7

Typical desktop HDD 30-35 dBA

Performance For AV Streaming Applications

- **New Emerging High Performance Multi-Stream Applications**
  - Multi head DVR systems.
  - DLNA compliant Digital Media Servers.
  - AV streaming combined with IT traffic.
- **Performance Metrics:**
  - Stream Performance: How many streams can you support, at what bit rate, with what trick-play features?
  - Quality of Service: How frequently do your streams get interrupted (i.e. how often do video glitches occur in the system)?
- **Objectives:**
  - Minimize the probability of stream buffer underflows and overflows.
  - Achieve a reasonable throughput & latency tradeoff with IT traffic, without allowing the IT traffic to interfere with AV traffic.
  - Increase system margin to allow the use of quiet seek.

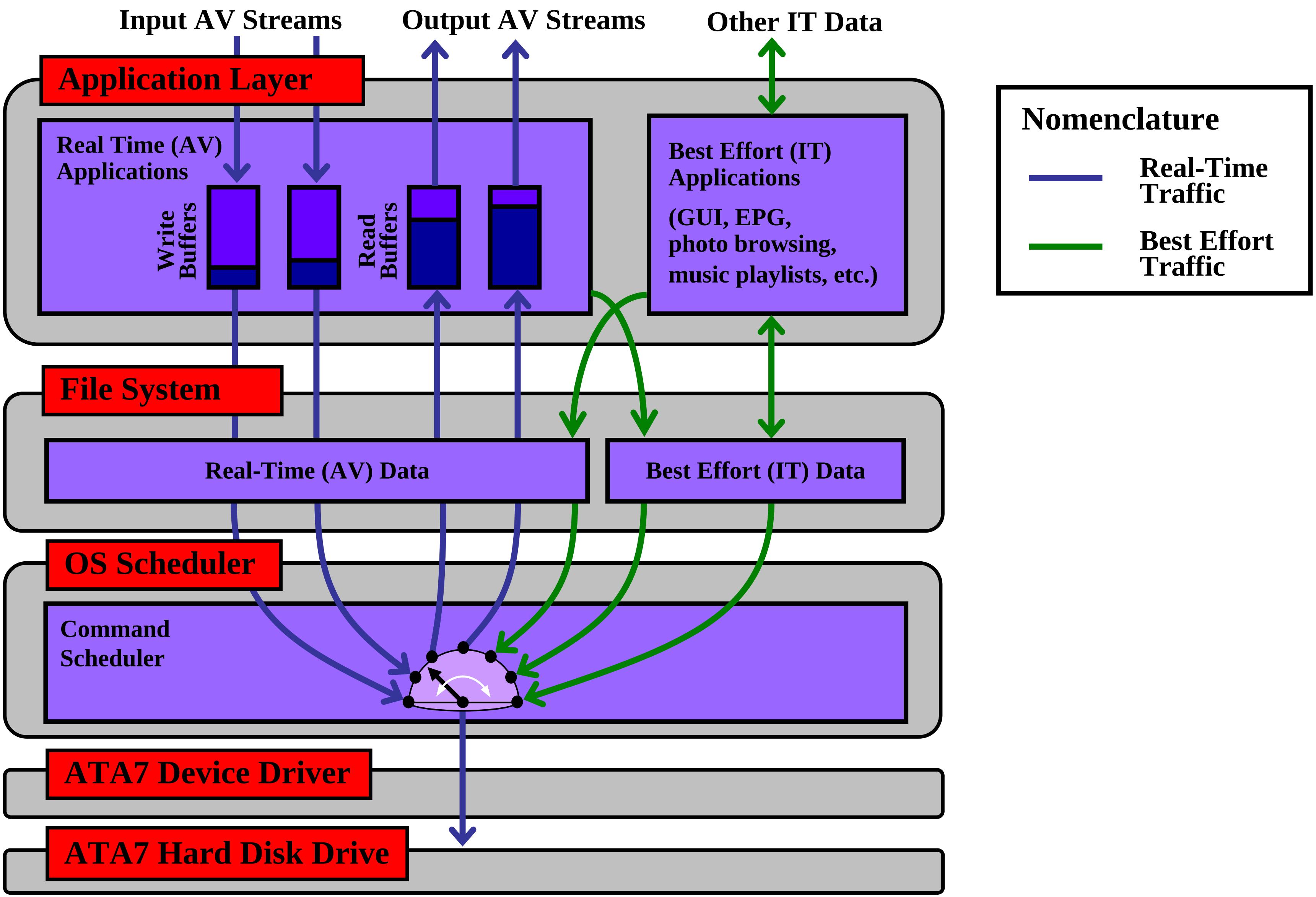


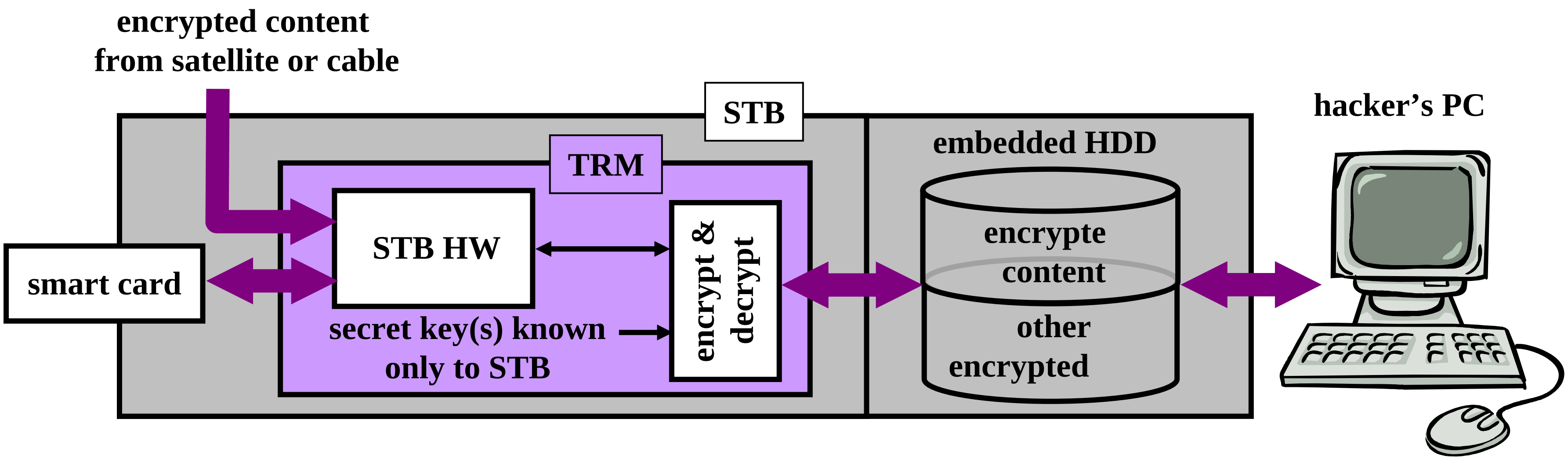
ATA-7 AV Streaming Commands

- **Specify Time Limits on Commands**
  - **Can either abort or suspend the command if the time limit is not met.**
  - **If the command is suspended, it can be resumed at a later time.**
- **Read Continuous Mode**
  - **The HDD will attempt to read data once, and will return with a list of sectors which it read and which it failed to read.**
  - **The system can decide whether to attempt to re-read data which was missed on the previous attempt.**
- **The commands provide a mechanism for ensuring Quality of Service (QoS) in streaming applications.**

QoS For AV Streaming Applications

HITACHI
Inspire the Next





All data (AV content and non-AV data) is stored in encrypted form.

Non AV-data may include:

- Encrypted digital rights information (licenses, permissions, pay-per-view credits).
- Other types of state information, depending on STB features.
- Software modules and system configuration data.

Security concerns

- Attacker may attach HDD to a PC and read or write all of the data.
- Easy to mount offline attack against CA scheme.
- May attack state information on the HDD.

Requirements

- Mutual authentication between HDD and host, and secure session.

Summary