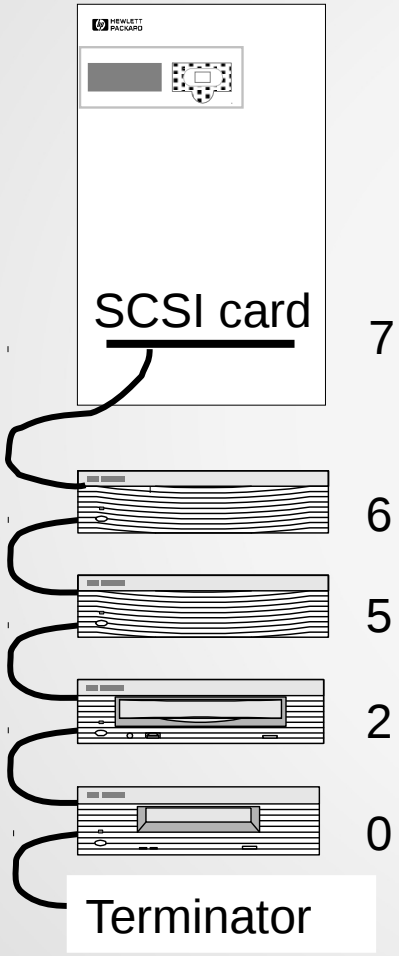
Типы дисков, используемые в СХД

Протокол	Описание
IDE/ATA	• Популярный интерфейс, используемый для подключения жестких и оптических дисков • Версия Ultra DMA/133 протокола ATA поддерживает пропускную способность 133 Мбайт/с
Serial ATA	• Последовательная версия спецификации IDE/ATA обычно используется для внутренних подключений • Обеспечивает скорость передачи данных до 16 Гбит/с (стандарт 3.2)
SCSI	• Популярный стандарт, используемый для подключения вычислительной системы к системе хранения • Поддерживает до 16 устройств на одной шине • Версия Ultra-640 обеспечивает скорость передачи данных до 640 Мбайт/с
SAS	• Последовательный протокол «точка-точка», заменяющий параллельный протокол SCSI • Поддерживает скорость передачи данных до 12 Гбит/с (SAS 3.0)
FC	• Широко используемый протокол для высокоскоростного обмена данными между вычислительной системой и системой хранения • Обеспечивает последовательную передачу данных, осуществляемую по медному и/или волоконно-оптическому кабелю • Последняя версия интерфейса Fibre Channel «16FC» позволяет передавать данные со скоростью до 16 Гбит/с
IP	• Существующая сеть на основе протокола IP используется для обмена данными между системами хранения • Примеры: протоколы iSCSI и FCIP

SCSI

SCSI (англ. Small Computer Systems Interface, произносится скази) — интерфейс, разработанный для объединения на одной шине различных по своему назначению устройств, таких как жёсткие диски, накопители на магнитооптических дисках, приводы CD, DVD, стримеры

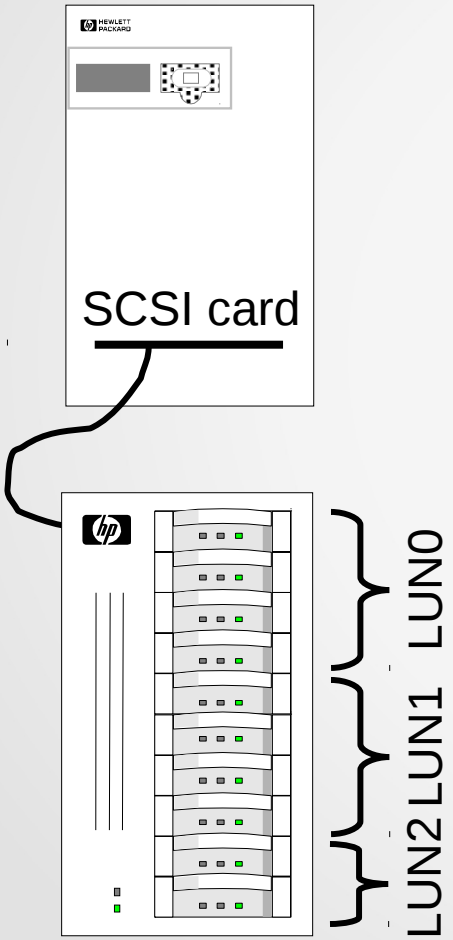
Концепции и адресация SCSI устройств



- стандарта SCSI
  - SE single-ended,
  - LVD low-voltage-differential — интерфейс дифференциальной шины низкого напряжения (+большая скорость)
  - HVD high-voltage-differential — интерфейс дифференциальной шины высокого напряжения (+большое растояние)
- Типы шины
  - Узкий" ("Narrow") 8-битные данные
  - "Широкий" ("Wide") 16-битные данные
- **SCSI цепочки**
- **SCSI терминаторы**
- **SCSI таргет адреса**

7 (более высокий приоритет) ----------> 0 ----> 15 ----------> 8 (более низкий приоритет)

Адресация SCSI устройств



HOST . CHANNEL . TARGET . LUN

SCSI Lun (0x00 … 0xff)

SCSI Target (0x00 … 0xff)

SCSI-Bus per HBA (0x00/0x01 for FC)

SCSI/FC HBA (0x00, …) sometimes called SCSI host scsi0...scsiX used in various commands

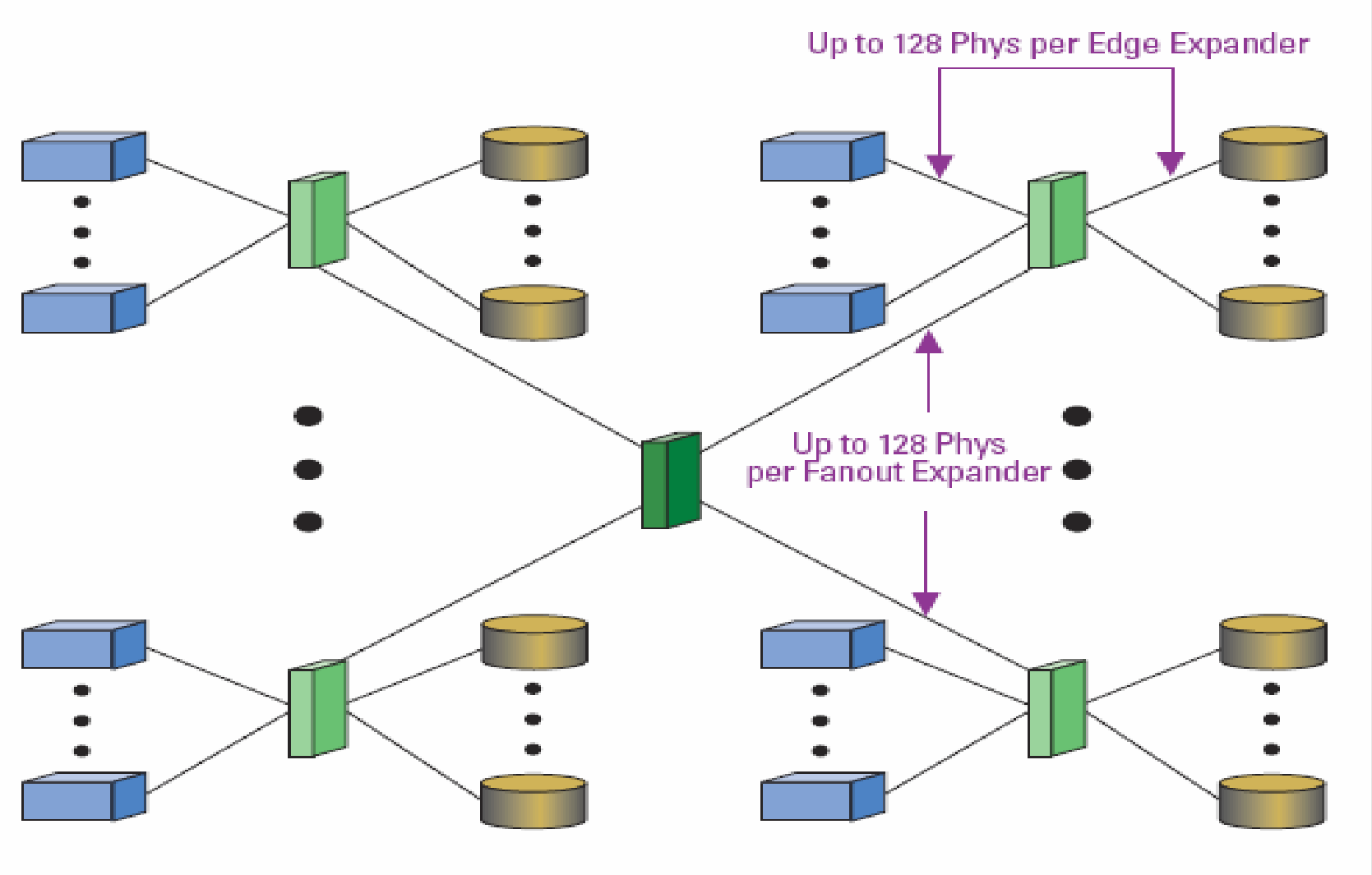
Serial Attached SCSI (SAS)

- компьютерный интерфейс, разработанный для обмена данными с такими устройствами, как жёсткие диски, накопители на оптическом диске и т. д.
- SAS использует последовательный интерфейс для работы с непосредственно подключаемыми накопителями (англ. Direct Attached Storage (DAS) devices).
- SAS разработан для замены параллельного интерфейса SCSI и позволяет достичь более высокой пропускной способности, чем SCSI.
- управления SAS-устройствами используются команды SCSI.

Сравнение SAS и параллельного SCSI

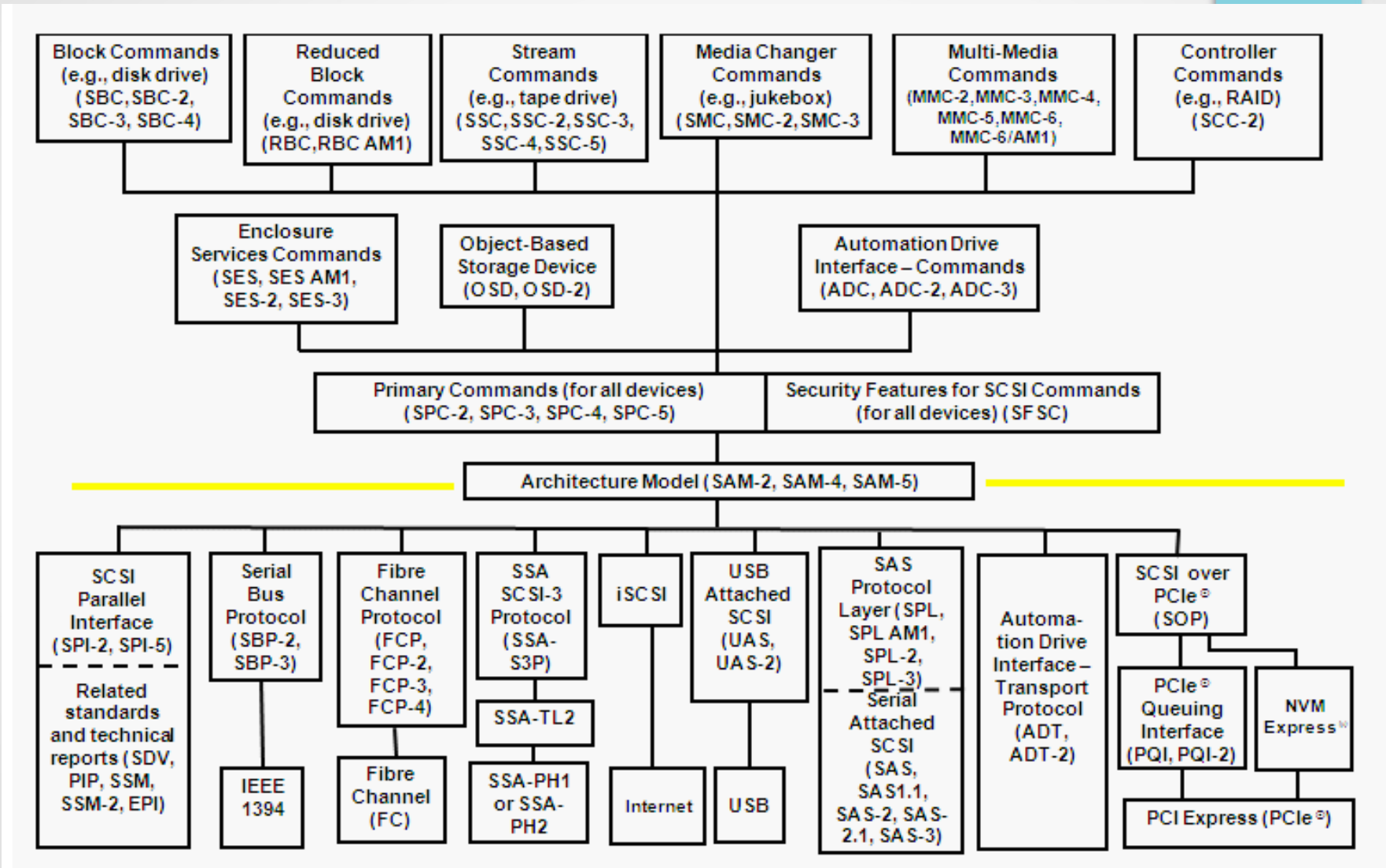
- SAS использует последовательный протокол (меньшее количество сигнальных линий)
- Интерфейс SCSI использует общую шину. SAS использует соединения точка-точка
- SAS не нуждается в терминации шины
- SAS поддерживает большое количество устройств (> 16384)
- SAS поддерживает высокие скорости передачи данных (1,5, 3,0 6,0 или 12 Гбит/с)

Комутация устройств в SAS

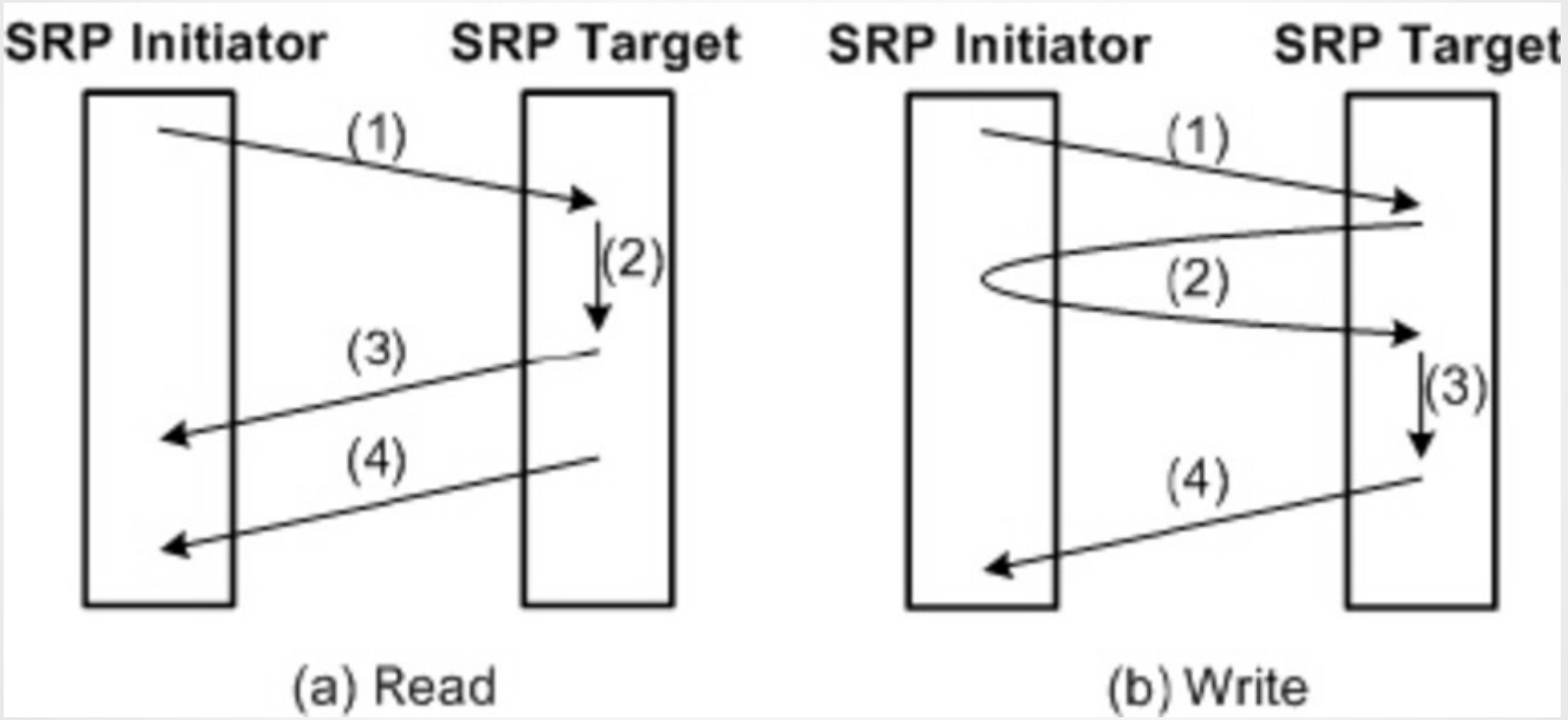


Maximum SAS domain

SCSI-3 Standards Architecture

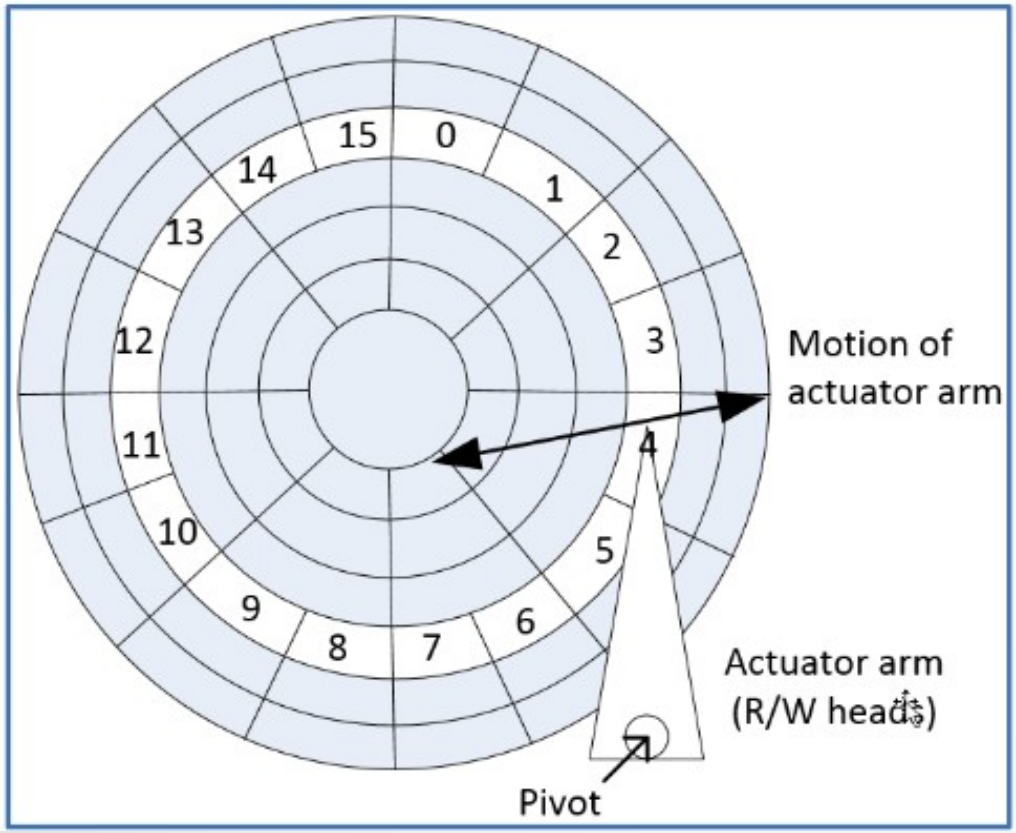


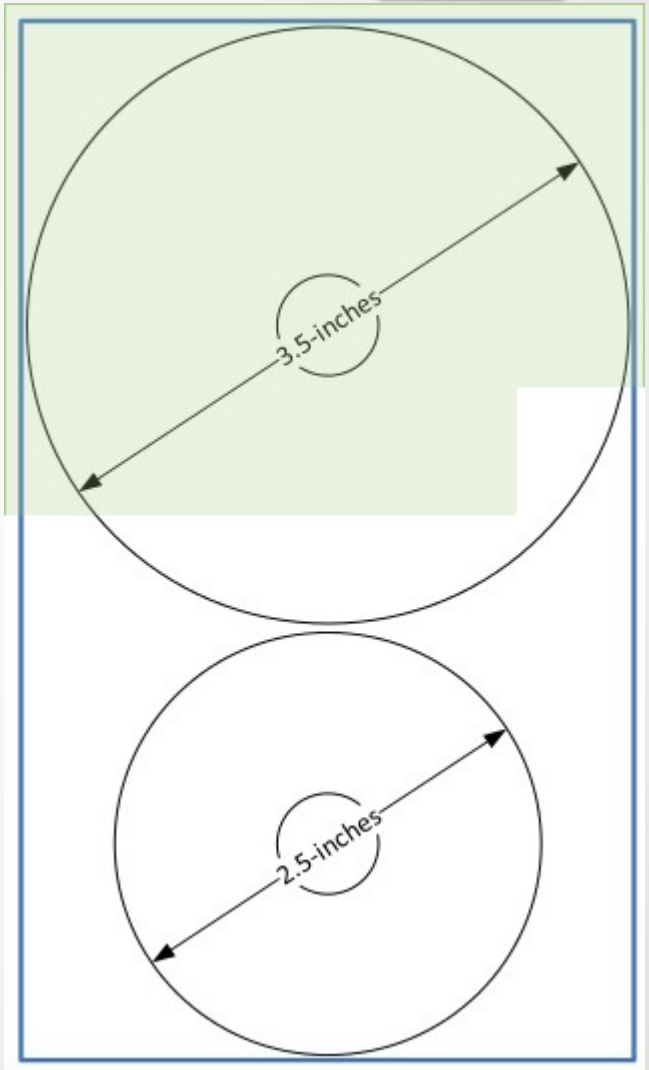
Последовательность команд



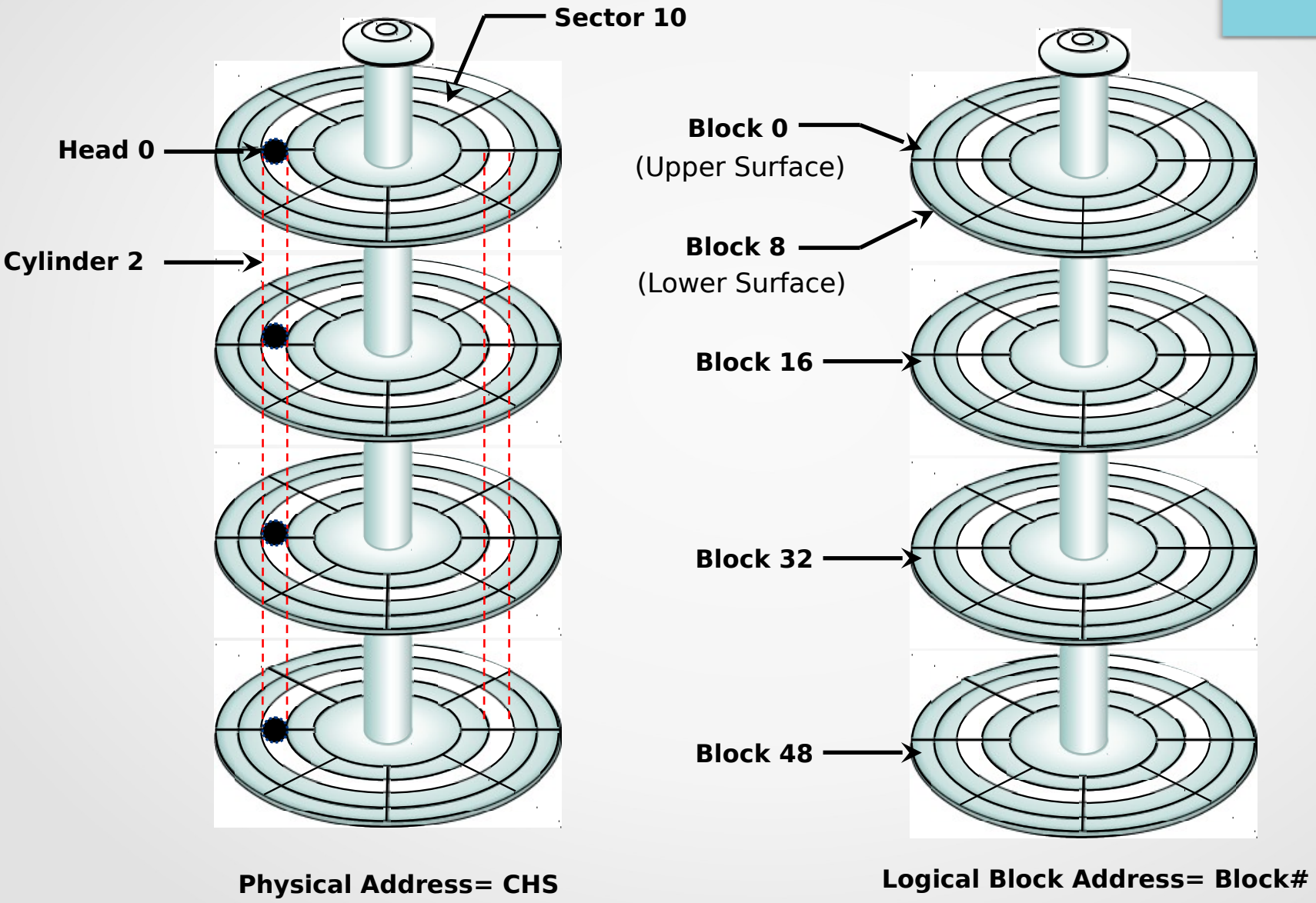
НЖМД


Жёсткий магнитный диск("блин")
Шпиндель
Головка
Коромысло
Ось вращения коромысла
Разъём питания
Блок перемычек
IDE разъём
Сервопривод коромысла


15
0
14
1
13
2
12
3
11
4
10
5
9
8
7
6
Motion of actuator arm
Actuator arm (R/W heads)
Pivot


3.5-inches
2.5-inches

НЖМД


Sector 10
Head 0
Cylinder 2
Block 0
(Upper Surface)
Block 8
(Lower Surface)
Block 16
Block 32
Block 48
Physical Address= CHS
Logical Block Address= Block#

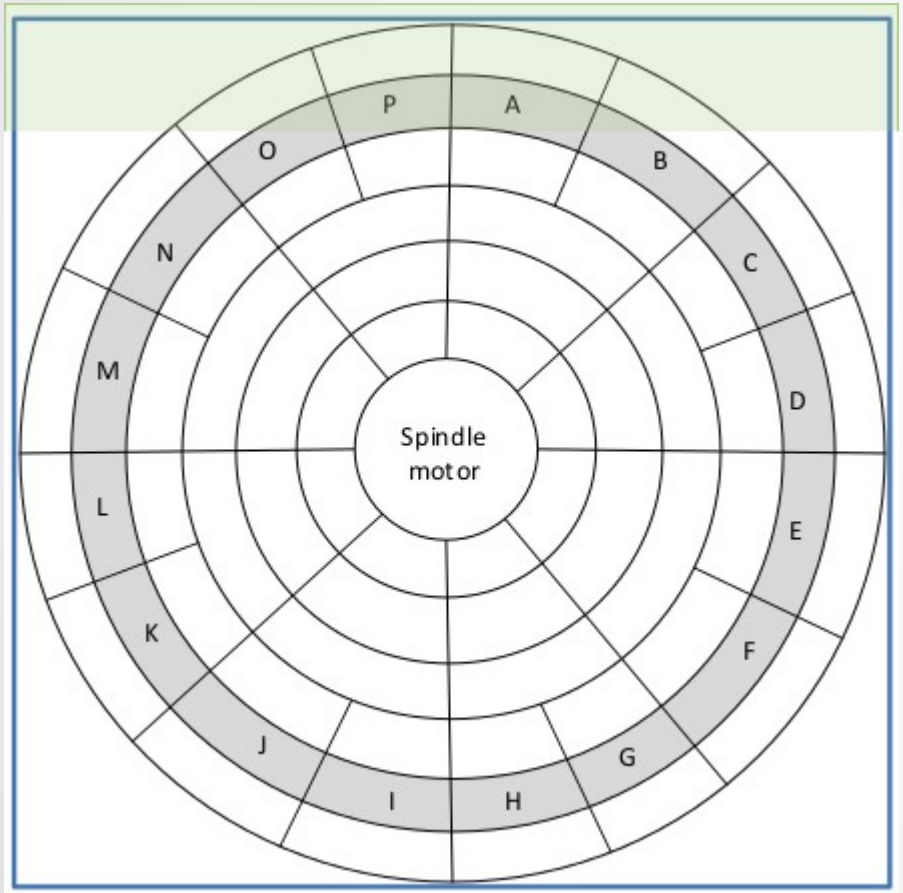
Преобразования между CHS и LBA

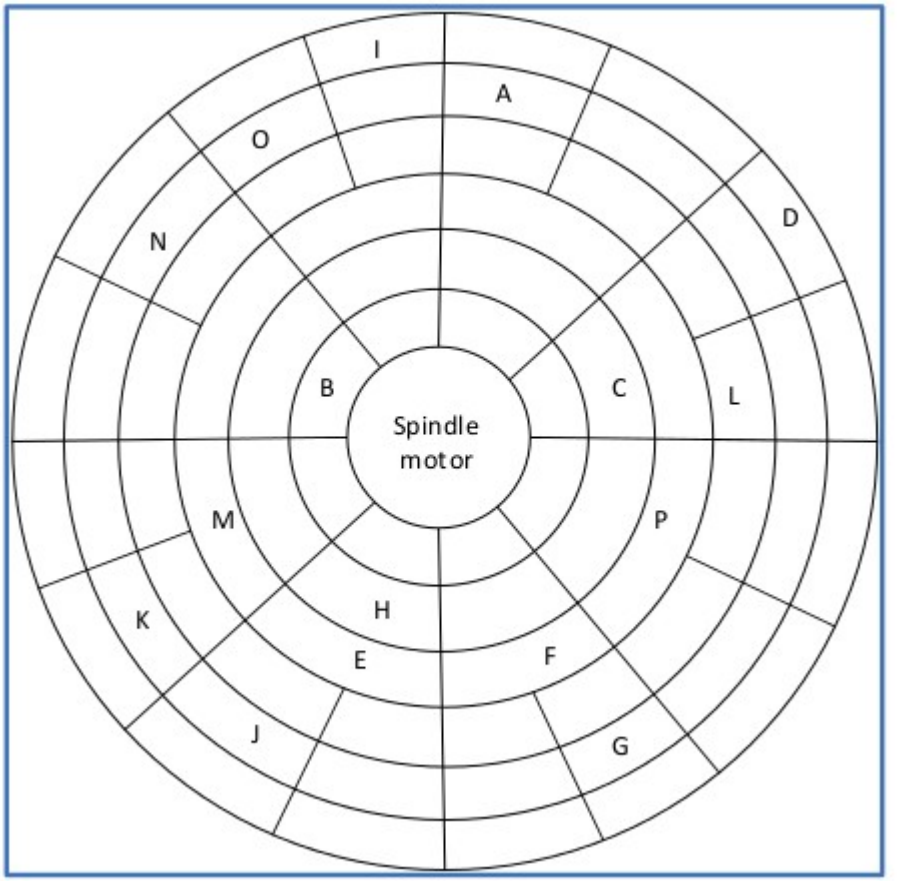
- Кортежи CHS можно преобразовать в адреса LBA и обратно по следующим формулам:
-
$$LBA(c,h,s) = (c \cdot H + h) \cdot S + s - 1$$
-
$$s = (LBA \bmod S) + 1$$
-
$$h = \frac{LBA - (s-1)}{S} \bmod H$$
-
$$c = \frac{LBA - (s-1) - h \cdot S}{H \cdot S}$$
- где c — номер цилиндра, h - номер головки, s - номер сектора, H — число головок, S — число секторов на дорожке, mod — операция взятия остатка от деления.

Если диск и BIOS используют LBA, но в какой-то момент требуется получить адрес в формате C/H/S, то используется схема, зависящая только от размера диска:

Размер диска	Секторов/дорожку	Головки	Цилиндры
1 < X ≤ 504 MiB	63	16	X/(63*16*512)
504 MiB < X ≤ 1008 MiB	63	32	X/(63*32*512)
1008 MiB < X ≤ 2016 MiB	63	64	X/(63*64*512)
2016 MiB < X ≤ 4032 MiB	63	128	X/(63*128*512)
4032 MiB < X ≤ 8032.5 MiB	63	255	X/(63*255*512)

Последовательный и случайный доступ



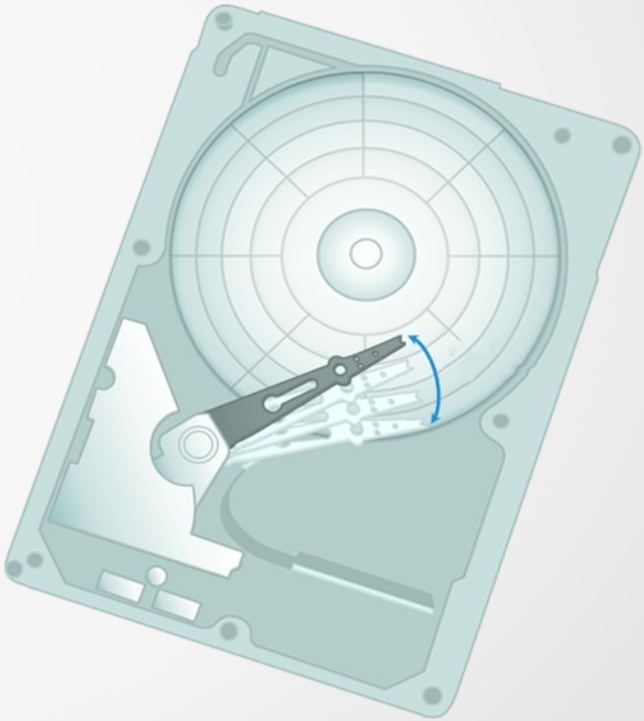


- RPM – Revolutions Per Minute
  - 5,400 RPM (5.4K) - Rotational Latency 5.56 ms
  - 7,200 RPM (7.2K) - Rotational Latency 4.17 ms
  - 10,000 RPM (10K) - Rotational Latency 3.00 ms
  - 15,000 RPM (15K) - Rotational Latency 2.00 ms
- Disk Service Time
  - Время, затраченное на диске, чтобы завершить запрос ввода / вывода
    - времени поиска (Seek Time)
    - задержки из-за вращения диска (Rotational Latency)
    - скорости передачи данных (Data Transfer Rate)

Время обработки диска = время поиска + задержка из-за вращения диска + время передачи данных

Время поиска

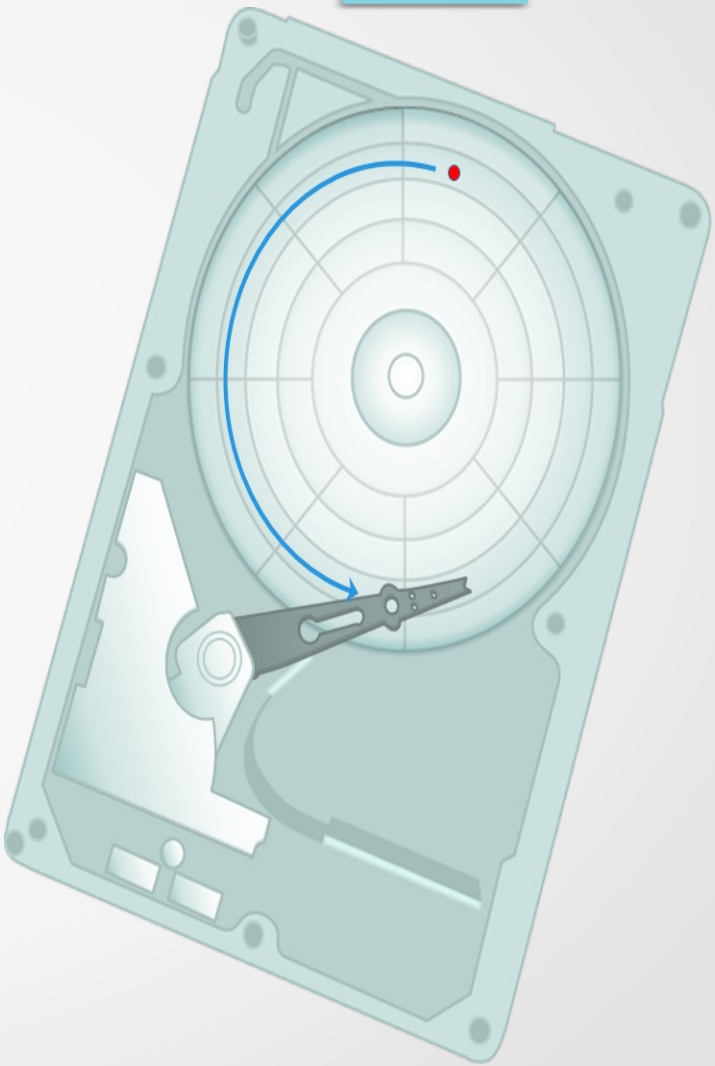
- Время, необходимое для позиционирования головки чтения/записи
- Чем меньше времени занимает поиск, тем быстрей проходят операции ввода-вывода
- Характеристики времени поиска:
  - время для полного оборота;
  - среднее время поиска;
  - время для перехода с дорожки на дорожку.
- Время поиска диска указывается его производителем



Задержка из-за вращения диска

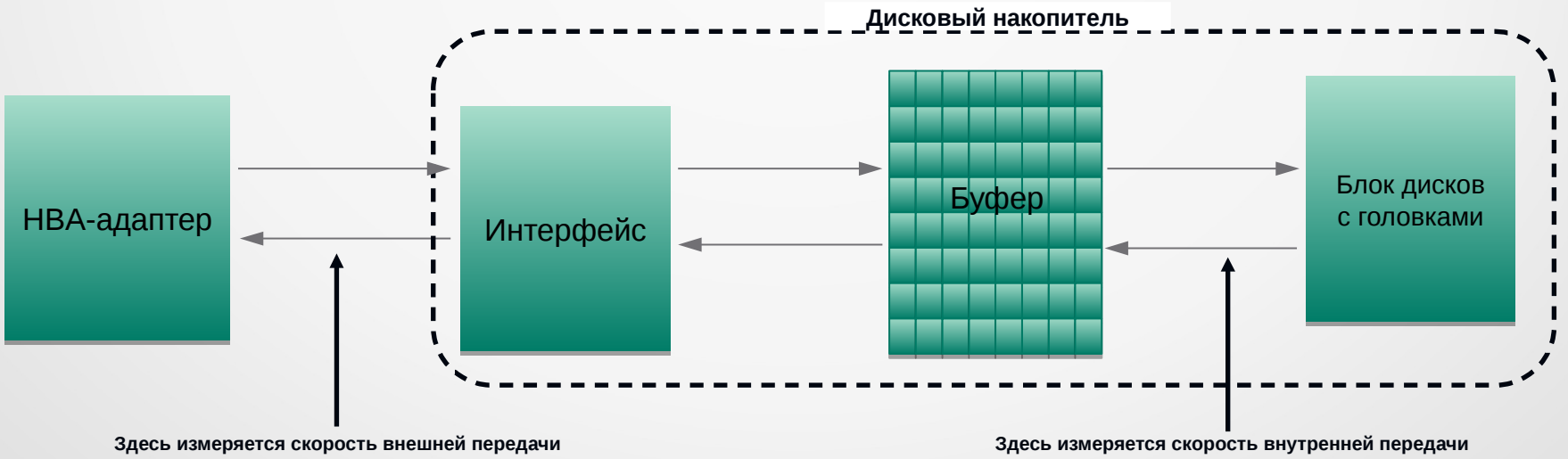
- Время, необходимое пластине для вращения и позиционирования данных в головке чтения/записи
- Зависит от скорости вращения шпинделя
- Средняя задержка из-за вращения диска
  - Половина времени, необходимого для полного оборота
  - Для «*X*» об/мин задержка диска вычисляется в миллисекундах по формуле:

$$= \frac{\left(\frac{1}{2} \times 1000\right)}{\left(\frac{X}{60}\right)} = \frac{500}{\left(\frac{X}{60}\right)} = \frac{30000}{X}$$



Скорость передачи данных

- Среднее количество данных, которое диск может доставить в HBA-адаптер за единицу времени
  - Скорость внутренней передачи: скорость, с которой данные перемещаются с поверхности пластины во внутренний буфер диска
  - Скорость внешней передачи: скорость, с которой данные перемещаются через интерфейс в HBA-адаптер



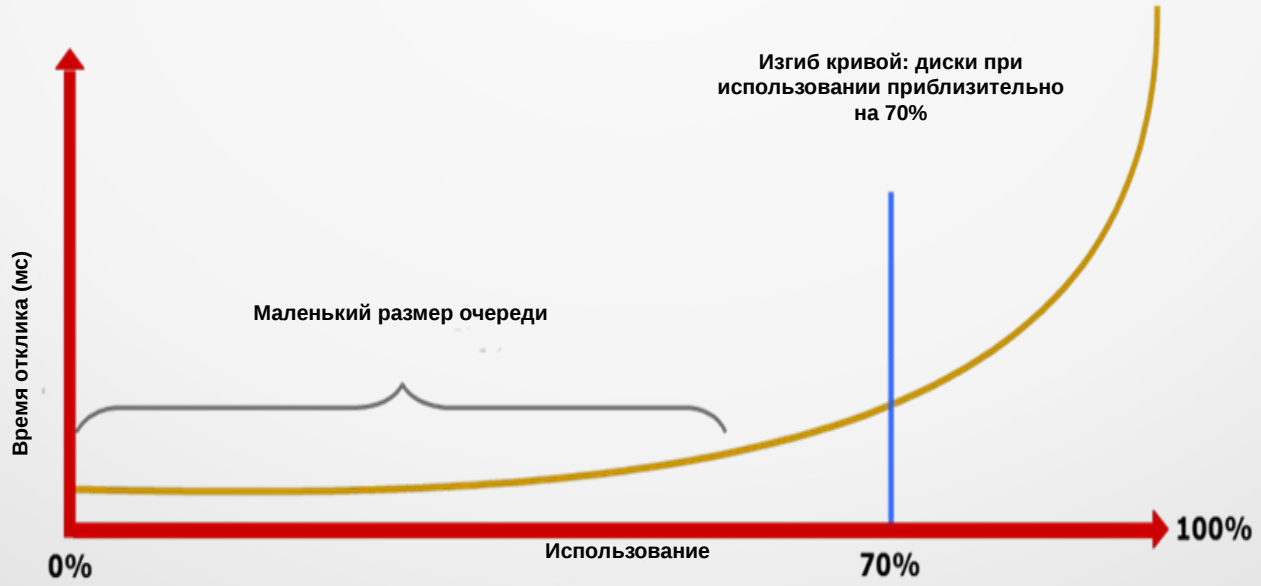
Размер блока	T=5мс+(0.5*250об/с) +РазмерБлока/40МБ	IOPS=1/T	MB/S=IOPS*Размер Блока
4	7,1	141	0,6
8	7,2	139	1,1
16	7,4	135	2,2
32	7,8	128	4,1
64	8,6	116	7,4
128	10,2	98	12,5
256	13,4	75	19,1

Сравнение использования контроллера ввода-вывода и времени отклика

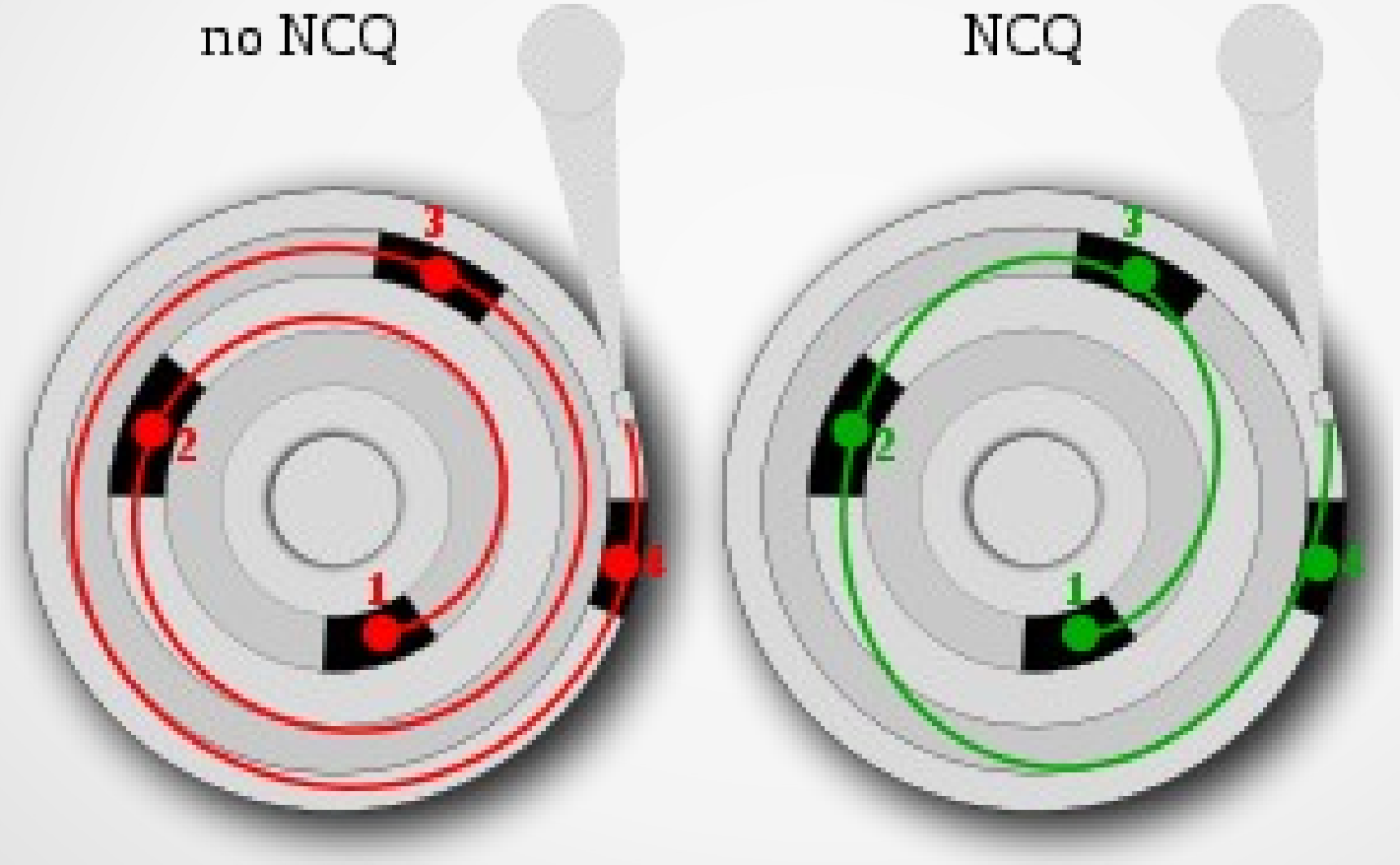
- На базе основополагающих правил производительности жестких дисков:

$$\text{Среднее времяотклика} = \frac{\text{Время обслуживания}}{(1 - \text{Использование})}$$

  - Время, необходимое контроллеру для обработки операций ввода-вывода
- Для приложений с высокими требованиями к производительности дисками обычно используется меньше 70% производительности обслуживания операций ввода-вывода



Последовательность обработки запросов



Конструкция системы хранения на основе требований приложений и производительности жестких дисков

- Количество дисков, необходимое для удовлетворения потребности приложения в емкости (DC):

$$Dc = \frac{\text{Общая необходимая емкость}}{\text{Емкость одного диска}}$$

- Количество дисков, необходимое для удовлетворения потребности приложения в производительности (DP):

$$Dp = \frac{\text{Сгенерированные приложением IOPS при максимальной рабочей нагрузке}}{\text{IOPS, обслуживаемые одним диском}}$$

- Количество операций ввода-вывода в секунду (S), обслуживаемых диском, зависит от времени обслуживания диска (T_S):

$$Ts = \text{Время поиска} + \frac{0.5}{(\text{Скорость вращения диска}/60)} + \frac{\text{Размер блока данных}}{\text{Скорость передачи данных}}$$

- TS — это время на завершение операции ввода-вывода, поэтому количество операций ввода-вывода в секунду (S), обслуживаемых диском, равняется (1/TS)

  - Для приложений, требовательных к производительности (S)= $0.7 \times \frac{1}{T_s}$

Необходимый для приложения диск = Макс. (DC, DP)

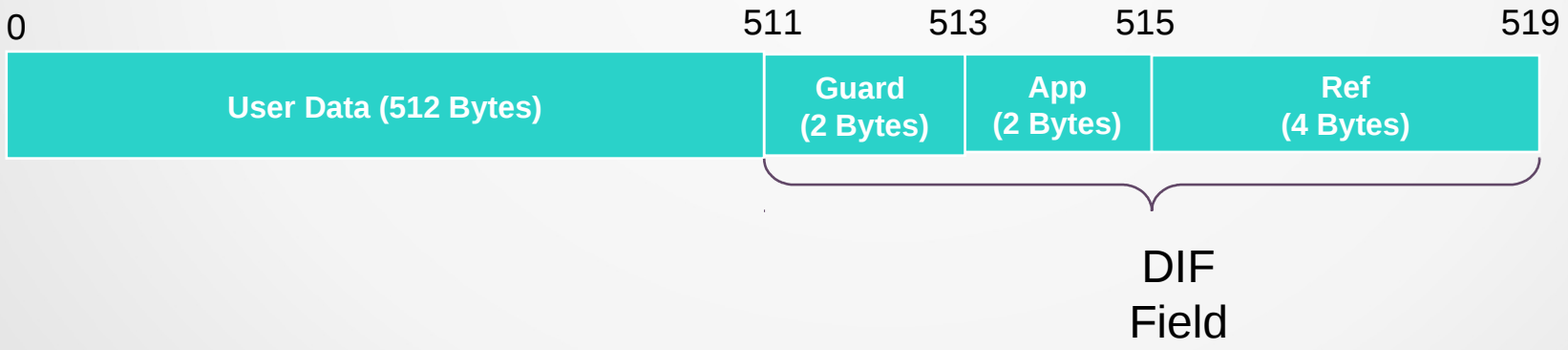
Упрощенная структура сектора жесткого диска

1. Адресный маркер
2. Адрес сектора
3. Контрольная сумма - для проверки целостности адреса
4. 512 байт данных пользователя
5. ECC-код коррекции ошибок данных
6. Контрольная сумма данных
7. Байты пробела

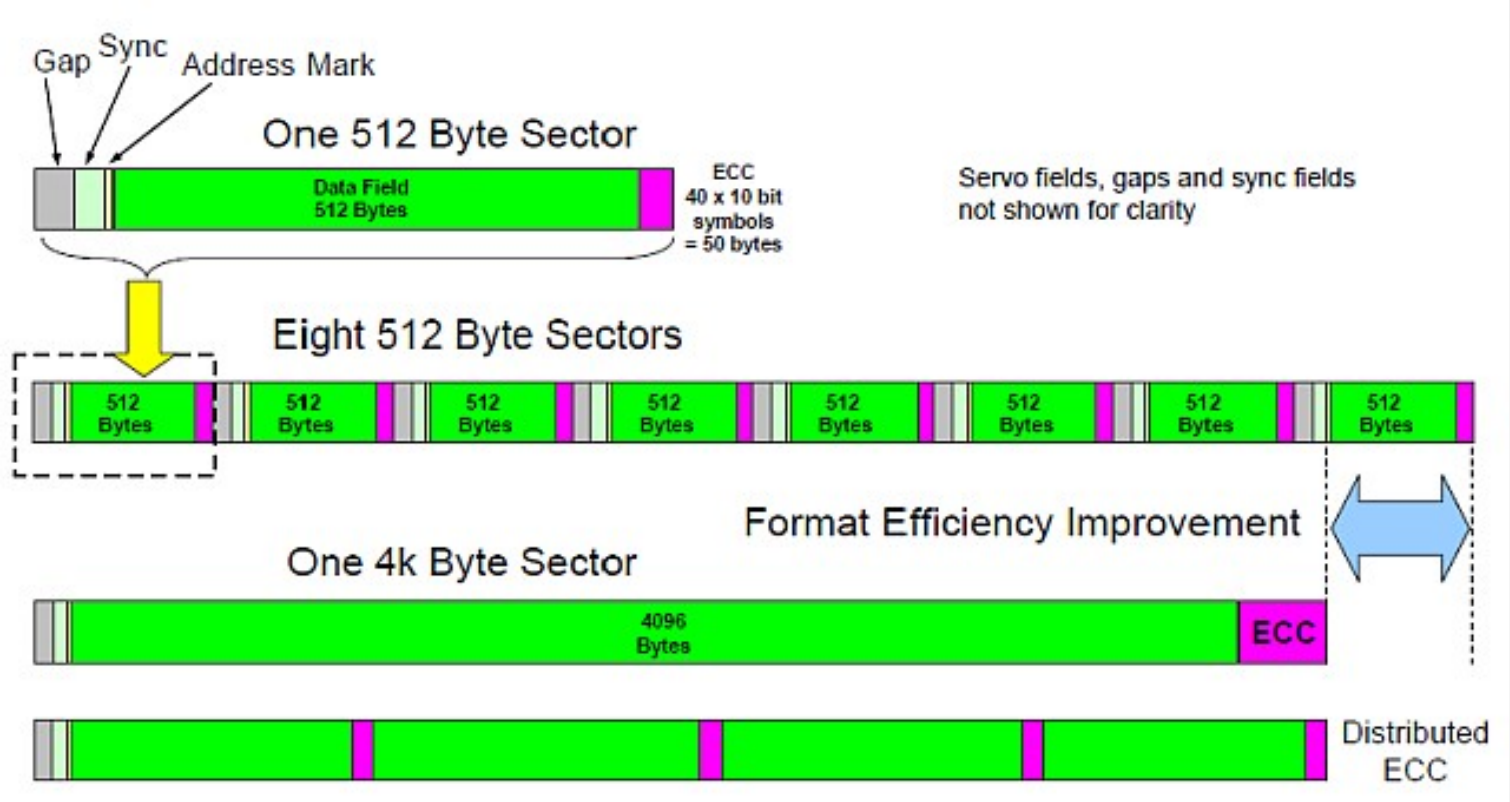
T10-DIF Data Integrity Field

An extra 8-byte Data Integrity Field (DIF) is added to the standard 512-byte disk block

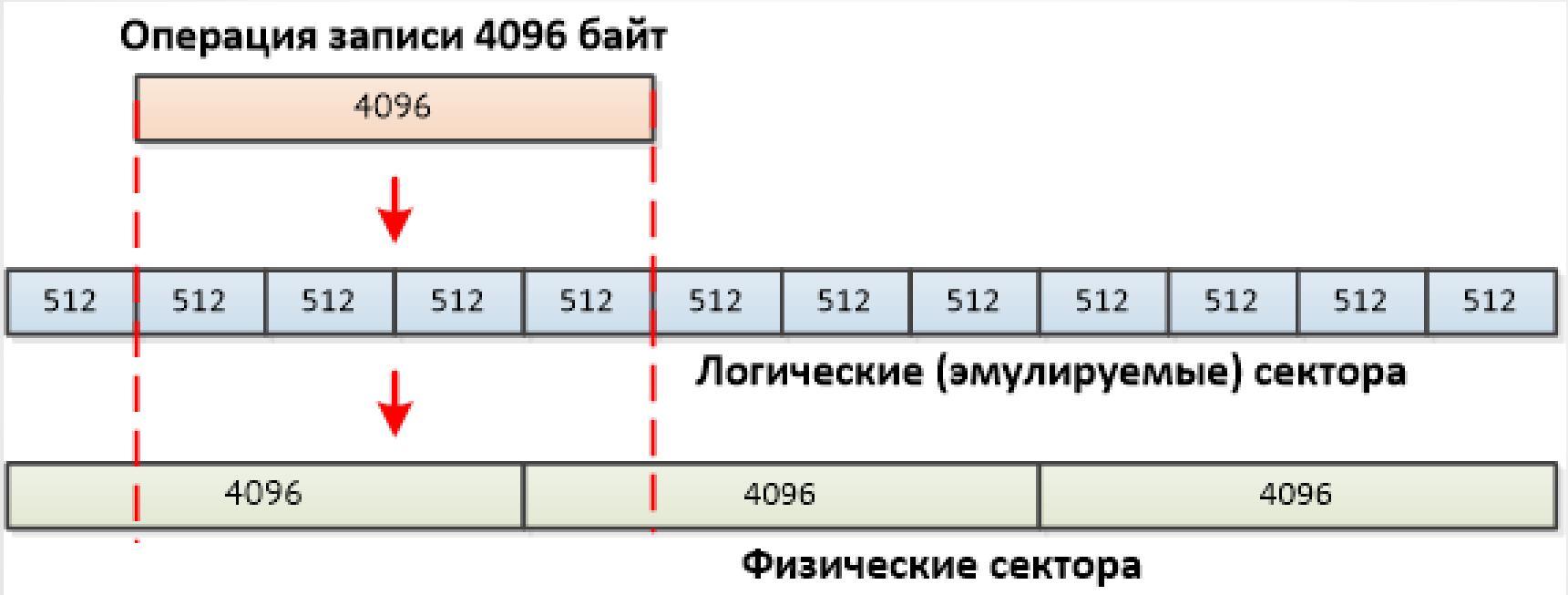
- 2 bytes Guard field: CRC of the data block
- 2 bytes App field: Application specific field
- 4 bytes Ref: Least significant bit of a Logical Block Address (LBA)



Advanced Format



Проблема выравнивания



SSD

SSS – Solid State Storage. All things solid state!

SSC – Solid State Card

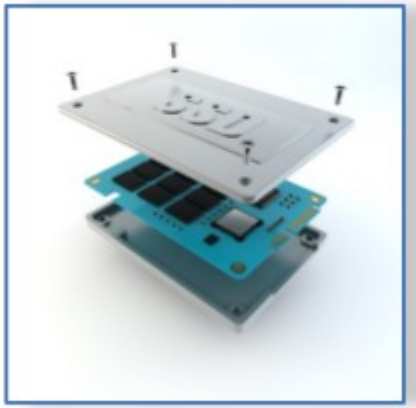
Better known as either

"PCI flash"

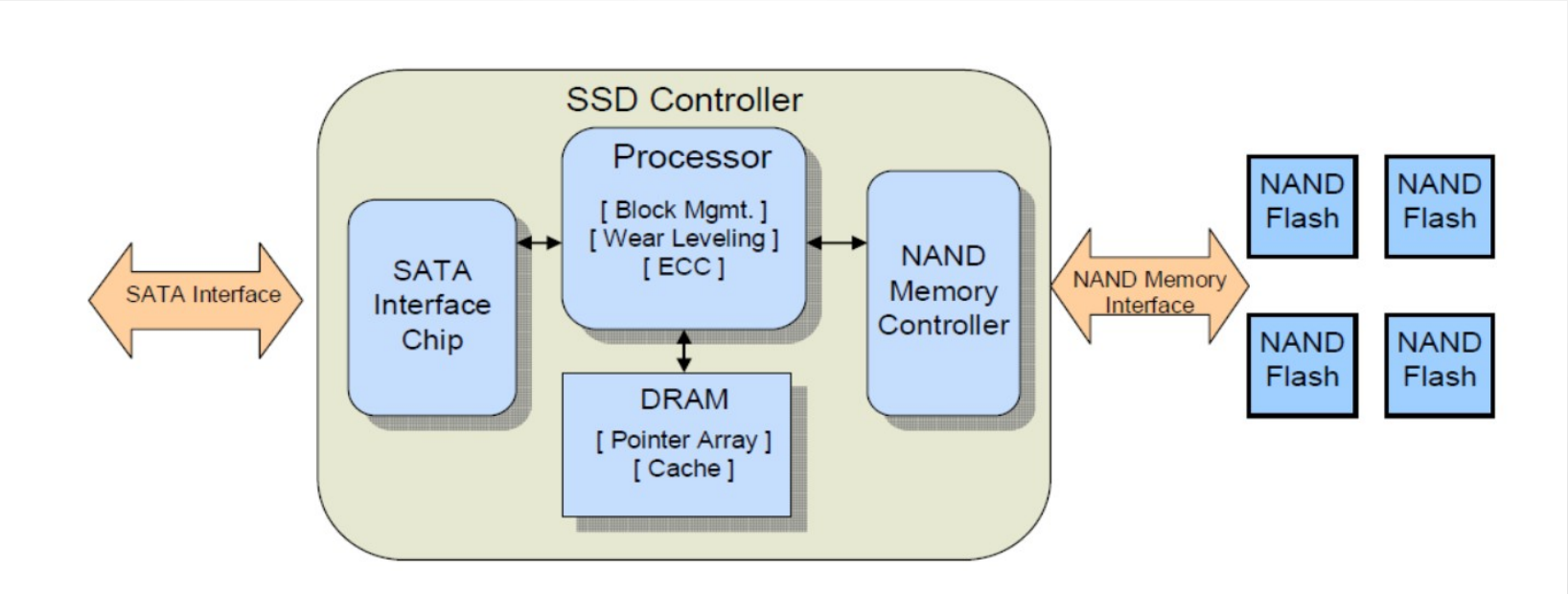
"PCI card flash"

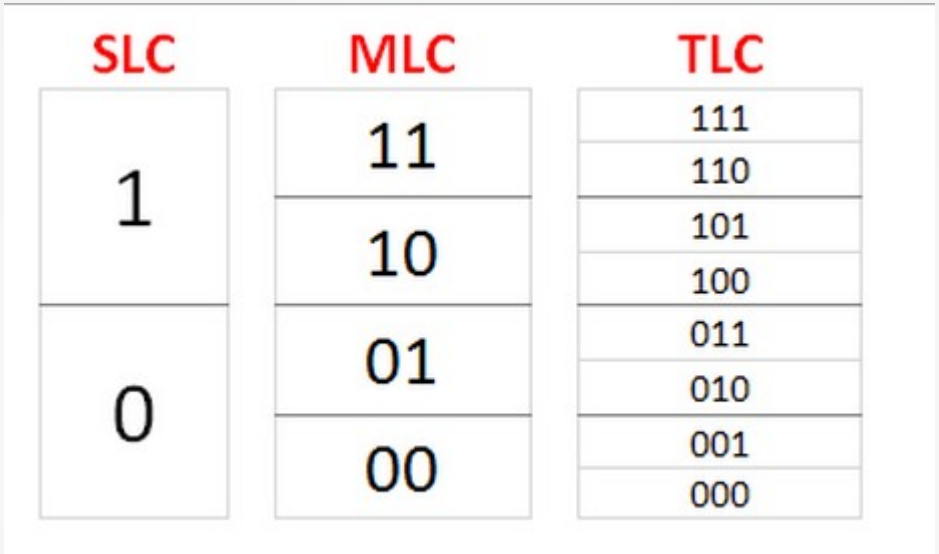
SSD – Solid State Drive

Solid state technology disguised as a spinning disk drive

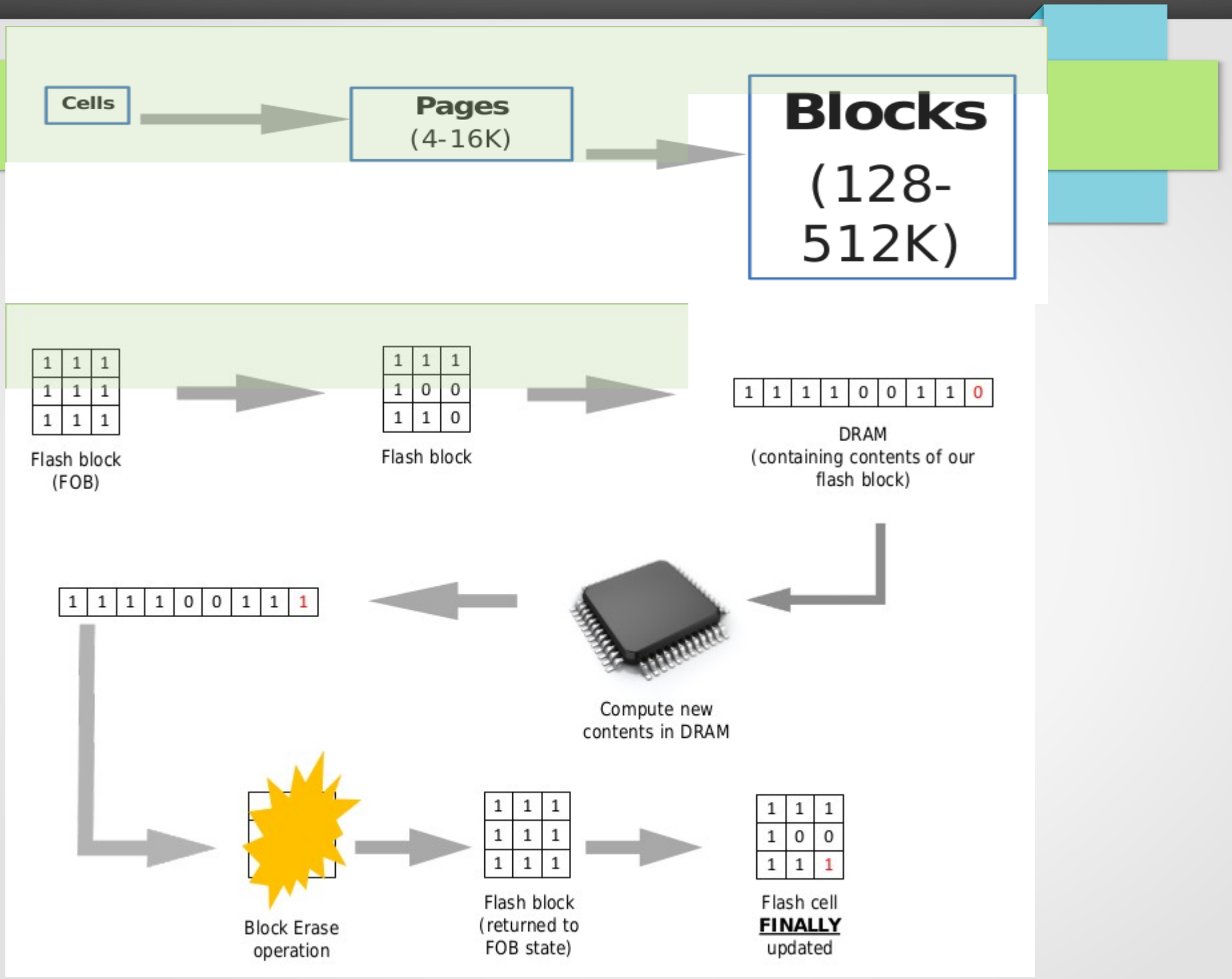


SSD


SSD Controller
Processor
[Block Mgmt.]
[Wear Leveling]
[ECC]
SATA Interface
SATA Interface Chip
NAND Memory Controller
NAND Memory Interface
DRAM
[Pointer Array]
[Cache]
NAND Flash
NAND Flash
NAND Flash
NAND Flash



NAND flash type	SLC	MLC	TLC
Bits per cell	1	2	3
P/E Cycles	100,000	3,000	1,000
Read Time	25us	50us	~75us
Program Time	200-300us	600-900us	~900-1350us
Erase Time	1.5-2ms	3ms	~4.5ms


Cells
Pages
(4-16K)
Blocks
(128-
512K)
1 1 1
1 1 1
1 1 1
Flash block
(FOB)
1 1 1
1 0 0
1 1 0
Flash block
1 1 1 1 0 0 1 1 0
DRAM
(containing contents of our
flash block)
Compute new
contents in DRAM
1 1 1 1 0 0 1 1 1
Block Erase
operation
1 1 1
1 1 1
1 1 1
Flash block
(returned to
FOB state)
1 1 1
1 0 0
1 1 1
Flash cell
FINALLY
updated

	15k rpm SAS HDD	Enterprise SSD
		200 GB or 400 GB
8 KB random reads (1/O per sec or IOPS)	280 at < 25 ms	36,445 at < 3 ms
8 KB random writes (IOPS)	240 at < 25 ms	15,515 at < 3 ms
128 KB sequential reads (MB/s)	138 at < 25 ms	425 at < 25 ms
128 KB sequential writes (MB/s)	26 at < 25 ms	230 at < 25 ms

Self-encrypting drives: SED