Интеграция ATA в CAM

Alexander Motin
mav@FreeBSD.org

- Предыстория: ISA ATA
  - Существующая ATA(4) подсистема была создана более 10 лет назад. Она ориентировалась на классические ISA ATA контроллеры, известные со времен PC/AT.
  - Классический ATA контроллер - это полуметровое продолжение 16-битной ISA шины, обеспечивающее прямой доступ процессора к регистрам ATA устройств.
  - PIO доступ является прямым чтением/записью процессором регистров ATA устройства. ISA DMA контроллер делае то-же, но аппаратно.
  - Так как протокол взаимодействия DMA отличен от PIO, для DMA вводится отдельный набор ATA команд.
  - С появлением PCI, ATA обзавелся поддержкой BusMastering вместо ISA DMA и возможностью 32-битного доступа к регистрам, для ускорения PIO, но принципиальных изменений не произошло.

- Предыстория: PCI ATA
  - Большая часть классических PCI ATA контроллеров отличается только расположением регистров, процедурами инициализации и управления режимами передачи (таймингами).
  - Все задачи по инициализации шины и управления ею ложатся на процессор. Таких понятий, как общий доступ к шине, арбитраж и т.п. в ATA просто не существует.
  - Спецификация ATAPI добавила возможность передачи SCSI комманд поверх ATA транспорта, за счет введения нескольких дополнительный команд.
  - На закате классического ATA, в него была добавлена поддержка очередей комманд (TCQ), но так-как контроллер по-прежнему был просто удинителем шины, то вся нагрузка по обслуживанию очереди и переключению контекста ложилась на драйвер, что не позволяло достичь желаемых результатов.

- Ближе к реальности: SATA 1.x
  - SATA вводит новый протокол работы шины, не связанный с ISA наследием ATA. Протокол симметричен и строится вокруг передачи абстрактных пакетов, называемых FIS (Frame Information Structure).
  - С этого момента ATA контроллеры получают набор «теневых» регистров, в которых накапливается команда для последующей пересылки в одном FIS. Это стало значительным шагом вперед, так как теперь контроллер оперирует командами, а не просто набором абстрактных регистров. API контроллера отделен от протокола шины.
  - К сожалению, несмотря на значительное внутреннее усовершенствование, внешне сохраняется прежний регистровый метод доступа. Это устраивает на данном этапе и сильно упрощает миграцию, но создает сложности в дальнейшем.

- Ближе к реальности: SATA 1.x NCQ
  - SATA 1.x вводит новый (пока необязательный) механизм очередей - NCQ. Этот механизм предполагает участие контроллера в обслуживании очереди. В частности, контроллер хранит параметры DMA для каждой команды (до 32), переключая их по запросу устройства.
  - Однако при этом сохраняется (расширяется) старый регистровый протокол взаимодействия, что усложняет обработку ошибок.
  - NCQ реализуется посредством введения двух новых ATA команд. По этой причине NCQ (в отличие от TCQ) не применим к ATAPI устройствам.
  - Обычные и NCQ команды не могут пересекаться во времени.

- Ближе к реальности: AHCI
  - AHCI изначально проектировался для поддержки SATA.
  - AHCI отказывается от эмуляции «теневых» регистров, окончательно переходя к оперированию командами.
  - AHCI отказывается от PIO. PIO режимы по-прежнему поддерживаются, но данные передает BusMastering.
  - Каждый SATA порт является независимым. Нет понятия master/slave устройтв.
  - AHCI аппаратно обслуживает очередь из 32 комманд, причем как NCQ, так и обычных. Последние просто исполняются последовательно, сокращая число обращений к драйверу.
  - AHCI стал фактическим стандартом для SATA контроллеров, однако большинство из них по-прежнему поддердивают режим эмуляции классического ATA.
  - В режиме эмуляции преимущества AHCI недоступны.

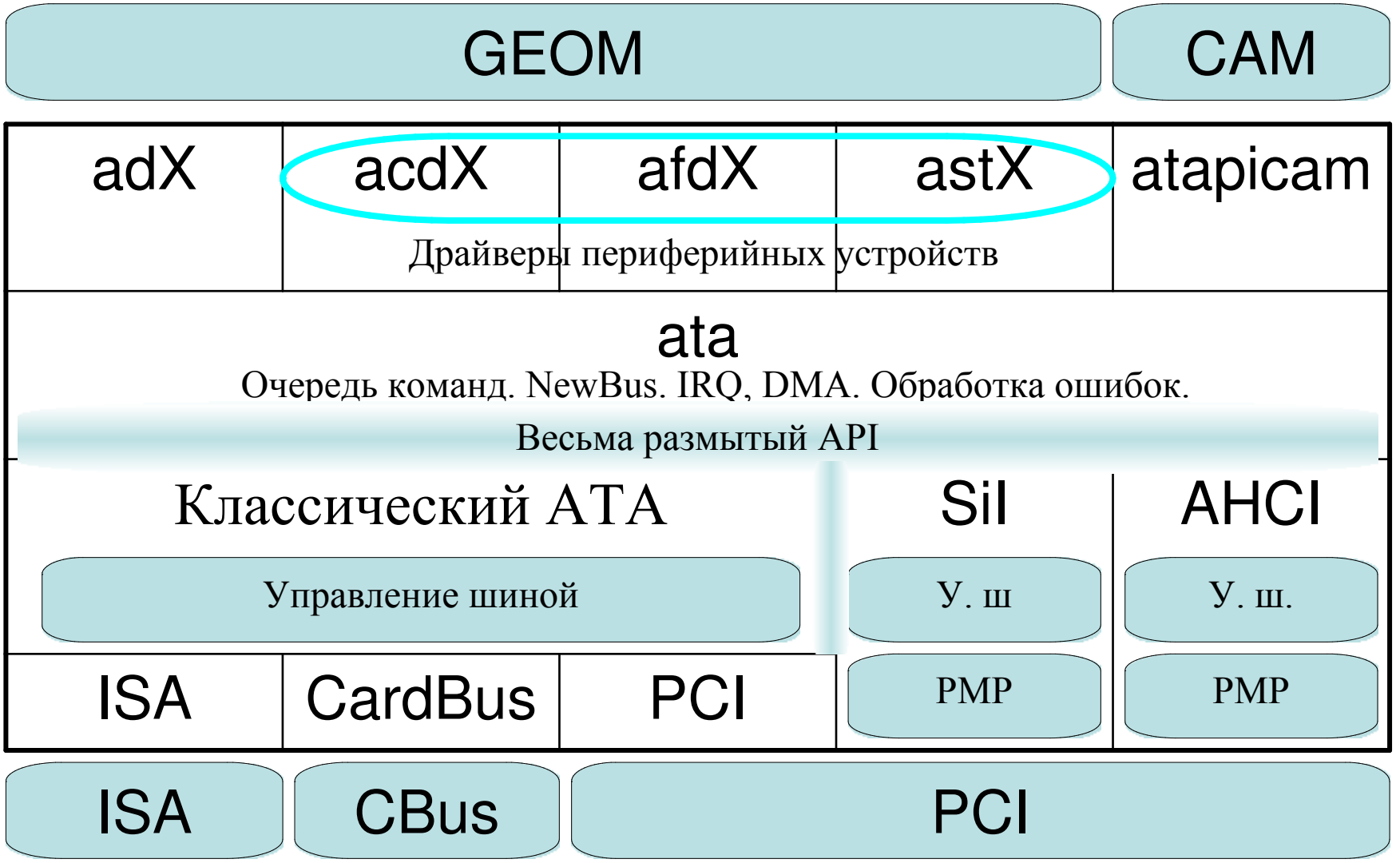
- Реальность: SATA 2.x
  - Помимо повышения скорости новый стандарт делает обязательной поддержку NCQ и вводит возможность подключения к одному порту контроллера до 15 устройств с использованием мультипликатора портов.
  - PMP работает как мультиплексор, разделяя FIS между устройствами по 4-х битному полю в заголовке.
  - Так как поддержка PMP не закладывалась в первоначальную спецификацию SATA, она требует дополнительного усовершенствования контроллера.
  - Для одновременной работы нескольких устройств за PMP контроллеру необходимо отслеживать статус каждого из них, чего AHCI изначально не умел.
  - Поддержка FIS-based switching введена в спецификацию AHCI 1.2, однако реально пока не поддерживается.
  - SATA 2.x контроллеры SiliconImage имеют собственный API и поддерживают FIS-based switching.

- Реальность: SAS
  - SAS электрически совместим с SATA,
  - SAS контроллеры могут эмулировать работу SATA контроллеров для подключения SATA устройств,
  - SAS в родном режиме использует другой логический уровень на базе SCSI комманд,
  - SAS поддерживает более длинные очереди комманд,
  - SAS в отличие от SATA позволяет объединять порты для достижения большей пропускной способности,
  - SAS поддерживает подключение большего числа устройств с применением экспандеров. Экспандеры, в отличие от PMP, могут каскадироваться.
  - SAS экспандеры могут эмулировать работу SATA контроллеров для подключения SATA устройств.

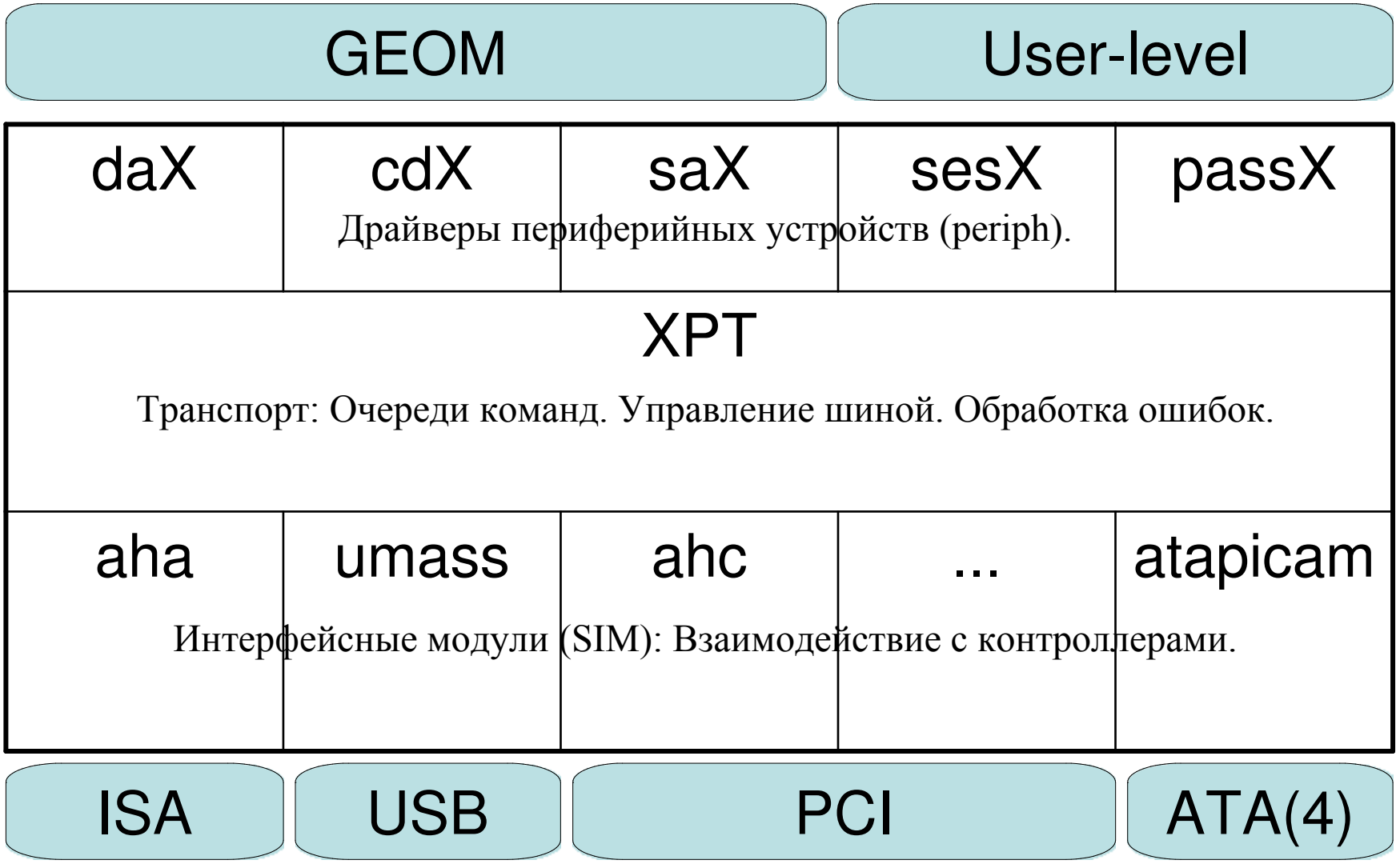
- Основные нововведения:
  - новые контроллеры сильно отличаются от классических,
  - SATA контроллеры поддерживают очереди команд,
  - SATA контроллеры и диски поддерживают NCQ,
  - SATA 2.x контроллеры поддерживают PMP,
  - SAS позволяет подключать SATA диски,
  - SAS экспандеры позволяют подключать SATA диски, тунелируя SATA поверх SAS,
  - ATA при помощи ATAPI позволяет подключать SCSI устройства, тунелируя SCSI поверх ATA.

- Большая часть перечисленного не поддерживается ATA(4).

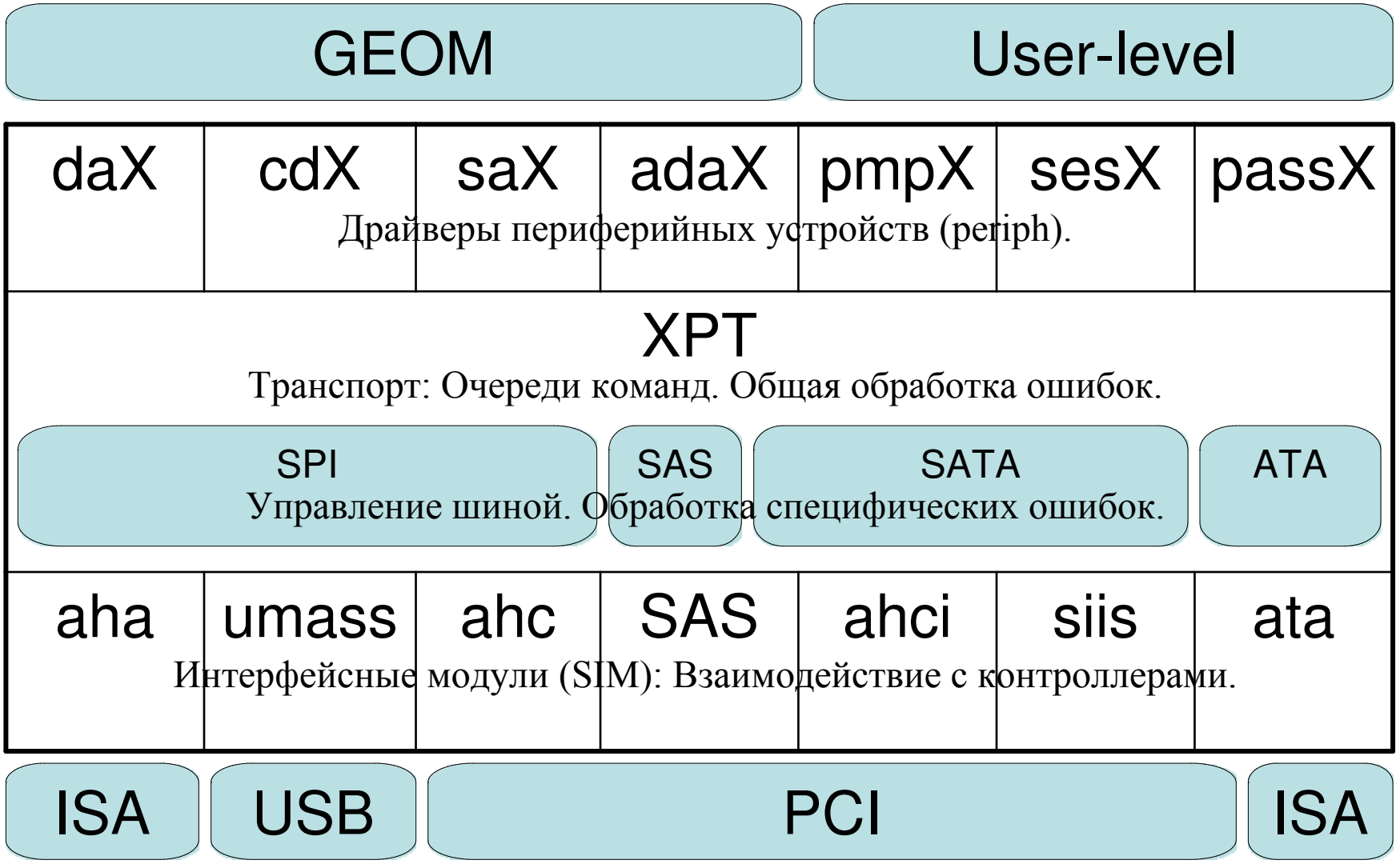
- Структура ATA(4)



- Структура существующего CAM(4)



- Структура обновленного CAM(4)



- ahci(4) поддерживает:
  - большинство интегрированных и некоторые внешние SATA2 контроллеры,
  - очереди до 32 команд,
  - NCQ,
  - PMP (пока без FIS-based switching, нет железа),
  - MSI (один или несколько векторов),
  - Command Completion Coalescing (по вкусу),
  - эффективное управление питанием SATA,
  - I/O любого размера, вплоть до MAXPHYS.

- siis(4) поддерживает:
  - несколько чипов (3124 - быстрый PCI-X, 3132/3531 - средние PCIe x1),
  - очереди до 31 команды,
  - NCQ,
  - PMP (с поддержкой FIS-based switching),
  - MSI по какой-то причине не работает под нагрузкой,
  - минимальное управление питанием SATA,
  - I/O любого размера, вплоть до MAXPHYS.

- обернутый ata(4) поддерживает:
  - старые чипы,
  - нет очередей команд,
  - нет NCQ,
  - нет PMP,
  - MSI на некоторых контроллерах,
  - минимальное управление питанием SATA,
  - I/O до 64/128K.

- Как это выглядит (список устройств):

```
%atacontrol list
atacontrol: control device not found: No such file or directory
%camcontrol devlist
<Slimtype DVD A  DS8A1P CA11>          at scbus0 target 0 lun 0 (pass0,cd0)
<ST3250620NS 3.AEK>                    at scbus1 target 0 lun 0 (pass0,ada0)
<Optiarc DVD RW AD-7200S 1.0A>         at scbus1 target 1 lun 0 (cd1,pass1)
<Hitachi HTS542525K9SA00 BBFOC31P>  at scbus1 target 2 lun 0 (ada3,pass5)
<Port Multiplier 37261095 1706>        at scbus1 target 15 lun 0 (pass2)
<OCZ-VERTEX 1.30>                      at scbus2 target 0 lun 0 (pass3,ada1)
<ST3250620NS 3.AEK>                    at scbus3 target 0 lun 0 (pass4,ada2)
```

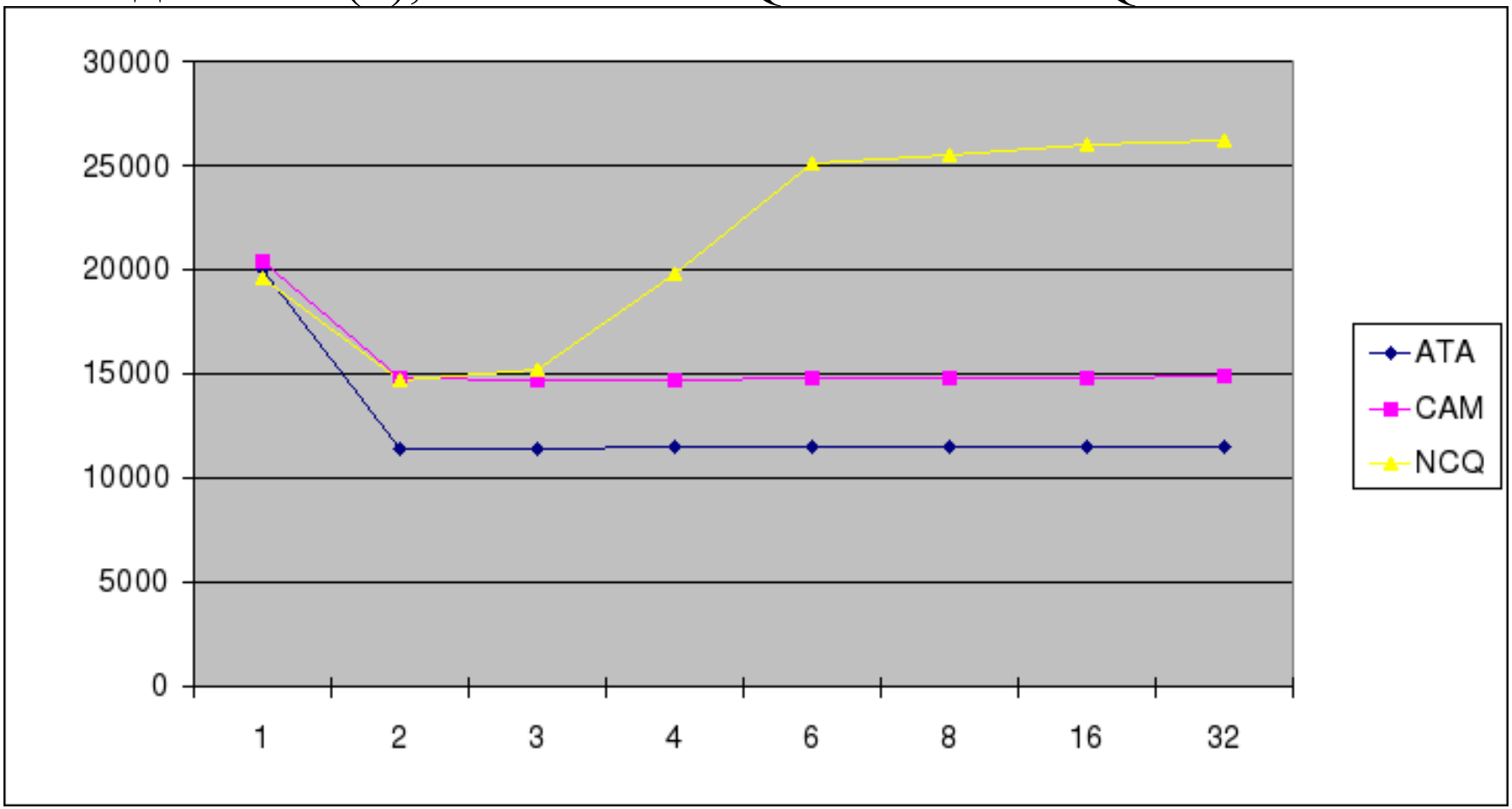
- Как это выглядит (camcontrol identify):

```
%camcontrol identify ada1
pass4: <OCZ-VERTEX 1.30> ATA/ATAPI-7 SATA 2.x device

protocol              ATA/ATAPI-7 SATA 2.x
device model          OCZ-VERTEX
serial number         H262LML036XYSDZVG1JI
firmware revision     1.30
cylinders             16383
heads                 16
sectors/track         63
LBA supported         125045424 sectors
LBA48 supported       125045424 sectors
PIO supported         PIO4
DMA supported         WDMA2 UDMA6
overlap not supported

Feature                      Support  Enable    Value           Vendor
write cache                    yes      yes
read ahead                     yes      yes
Native Command Queuing (NCQ)  yes              31/0x1F
Tagged Command Queuing (TCQ)  no       no       31/0x1F
SMART                          yes      yes
microcode download             yes      yes
security                       yes      no
power management               yes      yes
advanced power management      no       no       0/0x00
automatic acoustic management  no       no       0/0x00  0/0x00
```

- Производительность случайного доступа:
  - Зависимость числа транзакций в секунду от числа параллельних потоков чтения блоками по 512 байт с SSD OCZ Vertex 60GB, подключенного к ICH8 AHCI, для ATA(4), CAM без NCQ и CAM с NCQ.



- Линейная производительность:
  - Две SiI3124 PCI-X карты с 10 дисками, подключенными через 4 PMP, обеспечивают 1ГБ/с линейного чтения. Классический ata(4) в той-же ситуации обеспечивает только около 240МБ/с.

- Вопросы?