SATA, SAS, SSD, CAM, GEOM, ...
The Block Storage Subsystem in FreeBSD

Alexander Motin <mav@FreeBSD.org>
iXsystems, Inc.

EuroBSDCon 2013

«A long time ago» … in our own galaxy … appeared block storages ...

FreeBSD 3: struct cdevsw

FreeBSD 4: struct cdevsw + early disk(9) KPI

FreeBSD 5: disk(9) KPI + GEOM

Block storage above disk(9)

- Data operations:
  - Read
  - Write
- Properties
  - Block size
  - Capacity

Block storage KPI

- Data operations:
  - Read
  - Write
- Properties
  - Block size
  - Capacity

- start(struct bio *)
  - BIO_READ
  - BIO_WRITE
  - sectorsize
  - mediasize

Removable block storage

- Media lock/notify
- Data operations:
  - Read
  - Write
- Properties
  - Block size
  - Capacity

- access(), spoiled()
- start(struct bio *)
  - BIO_READ
  - BIO_WRITE

  - sectorsize
  - mediasize

Write-caching block storage

- Media lock/notify
- Data operations:
  - Read
  - Write
  - Cache flush
- Properties
  - Block size
  - Capacity

- access(), spoiled()
- start(struct bio *)
  - BIO_READ
  - BIO_WRITE
  - BIO_FLUSH

  - sectorsize
  - mediasize

Thin-provisioned block storage

- Media lock/notify
- Data operations:
  - Read
  - Write
  - Cache flush
  - Unmap / Trim
- Properties
  - Block size
  - Capacity

- access(), spoiled()
- start(struct bio *)
  - BIO_READ
  - BIO_WRITE
  - BIO_FLUSH
  - BIO_DELETE

  - sectorsize
  - mediasize

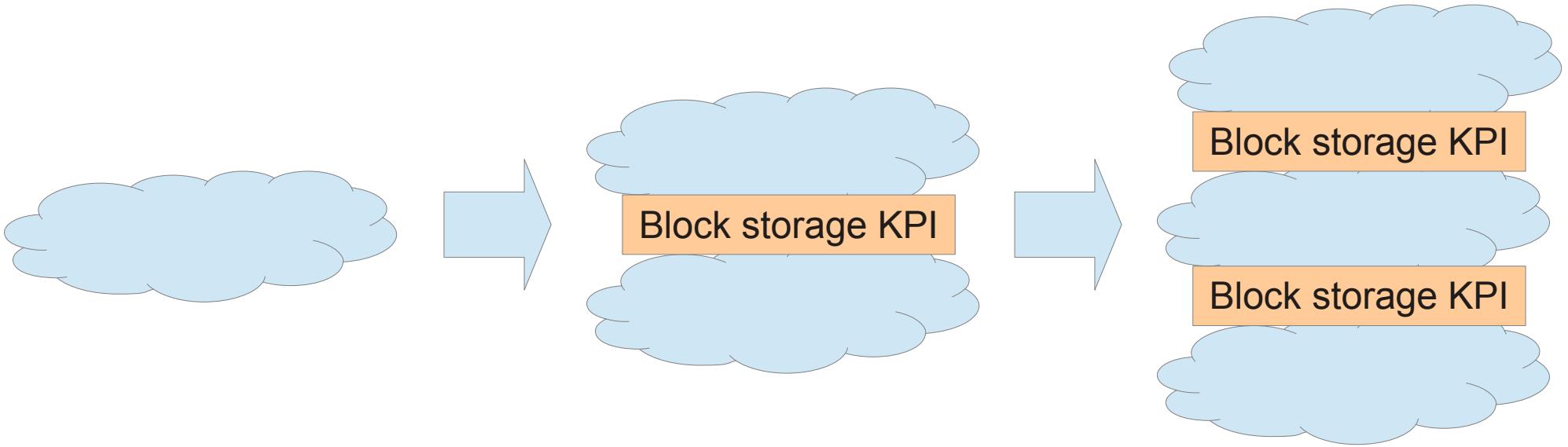
Addtional attributes

- Media lock/notify
- Data operations:
  - Read
  - Write
  - Cache flush
  - Unmap / Trim
- Properties
  - Block size
  - Capacity
  - C/H/S, physical sector size, serial number, ...

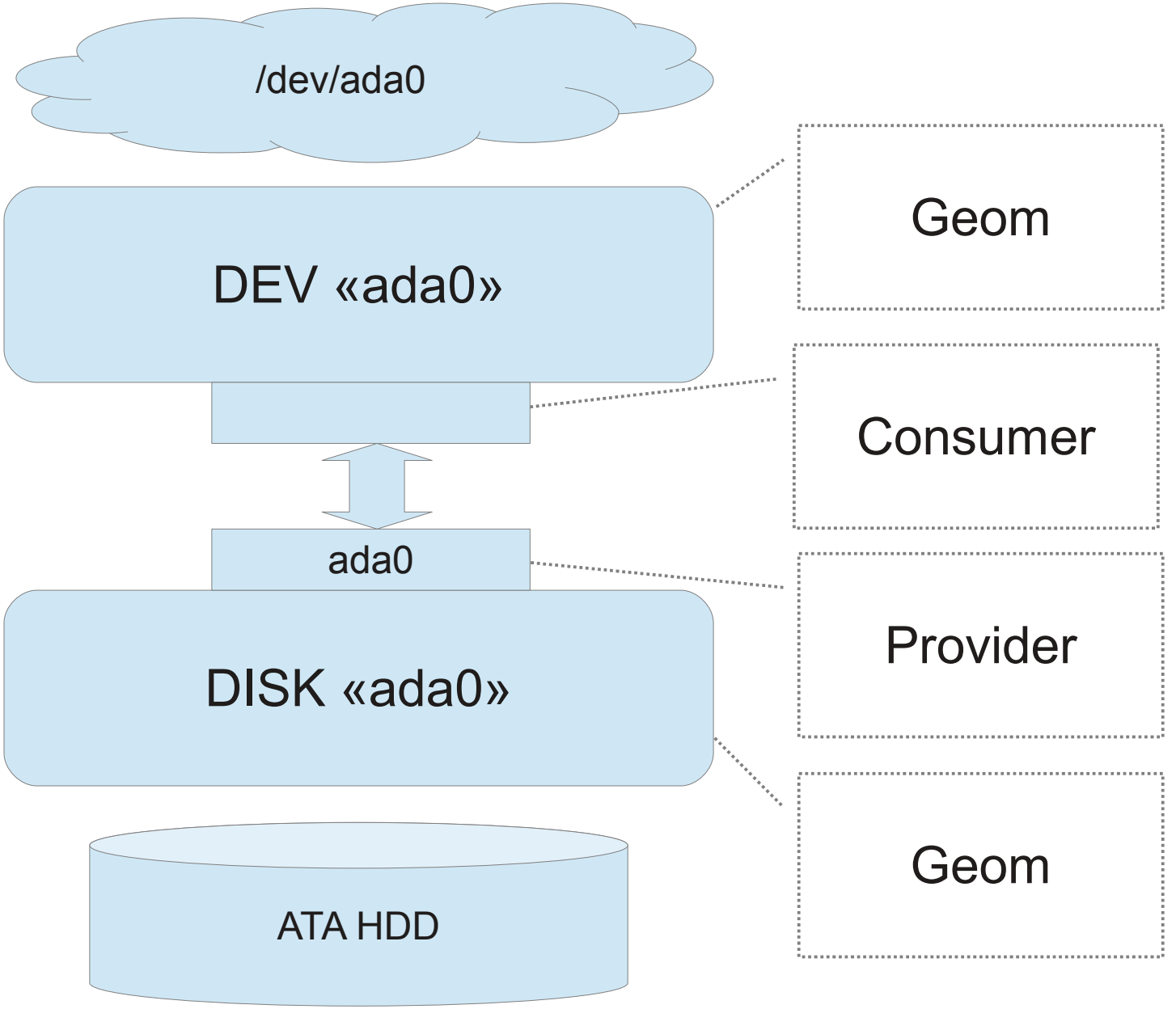
- access(), spoiled()
- start(struct bio *)
  - BIO_READ
  - BIO_WRITE
  - BIO_FLUSH
  - BIO_DELETE

  - sectorsize
  - mediasize
  - stripesize, stripeoffset, BIO_GETATTR

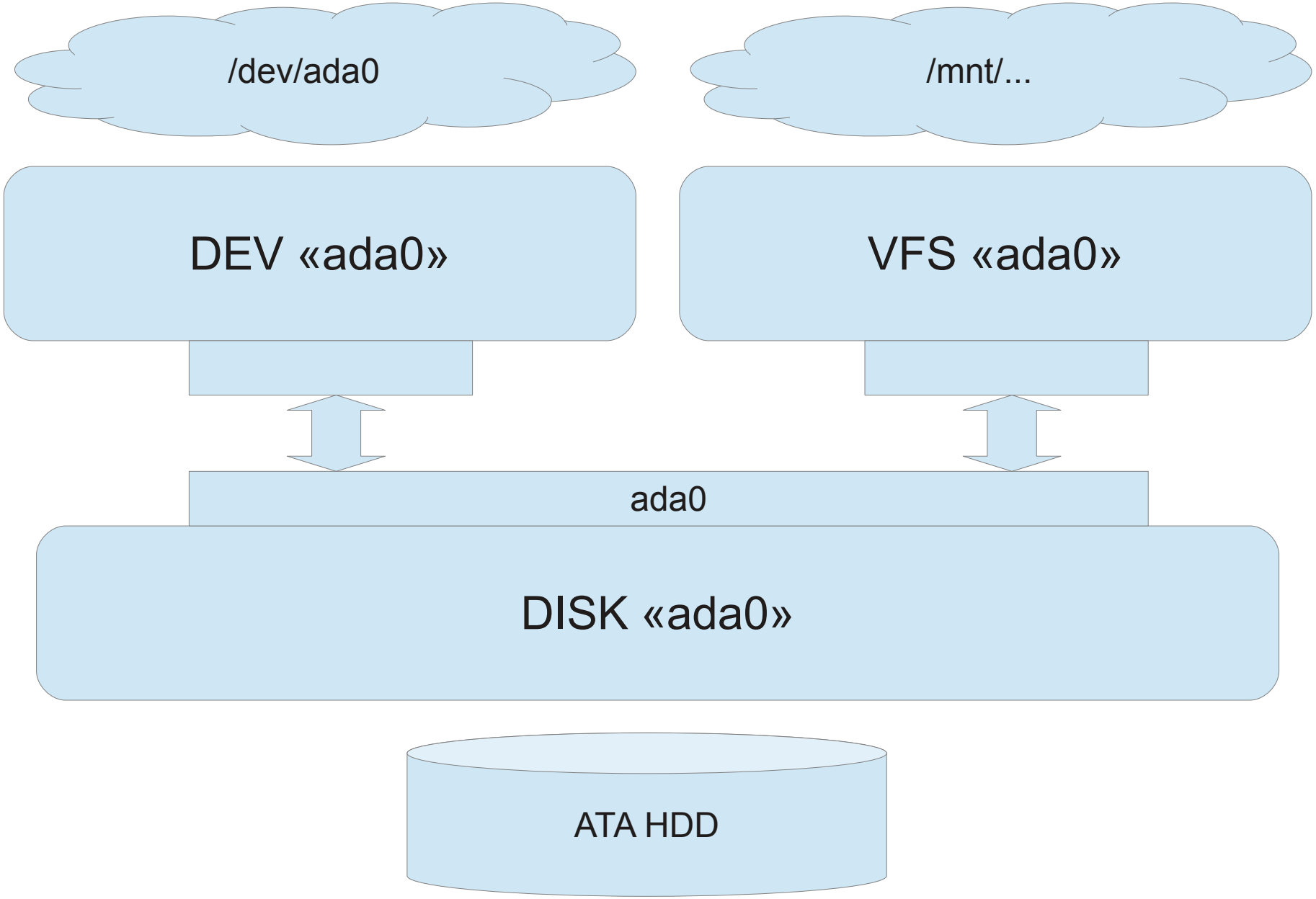
From one layer to many – GEOM



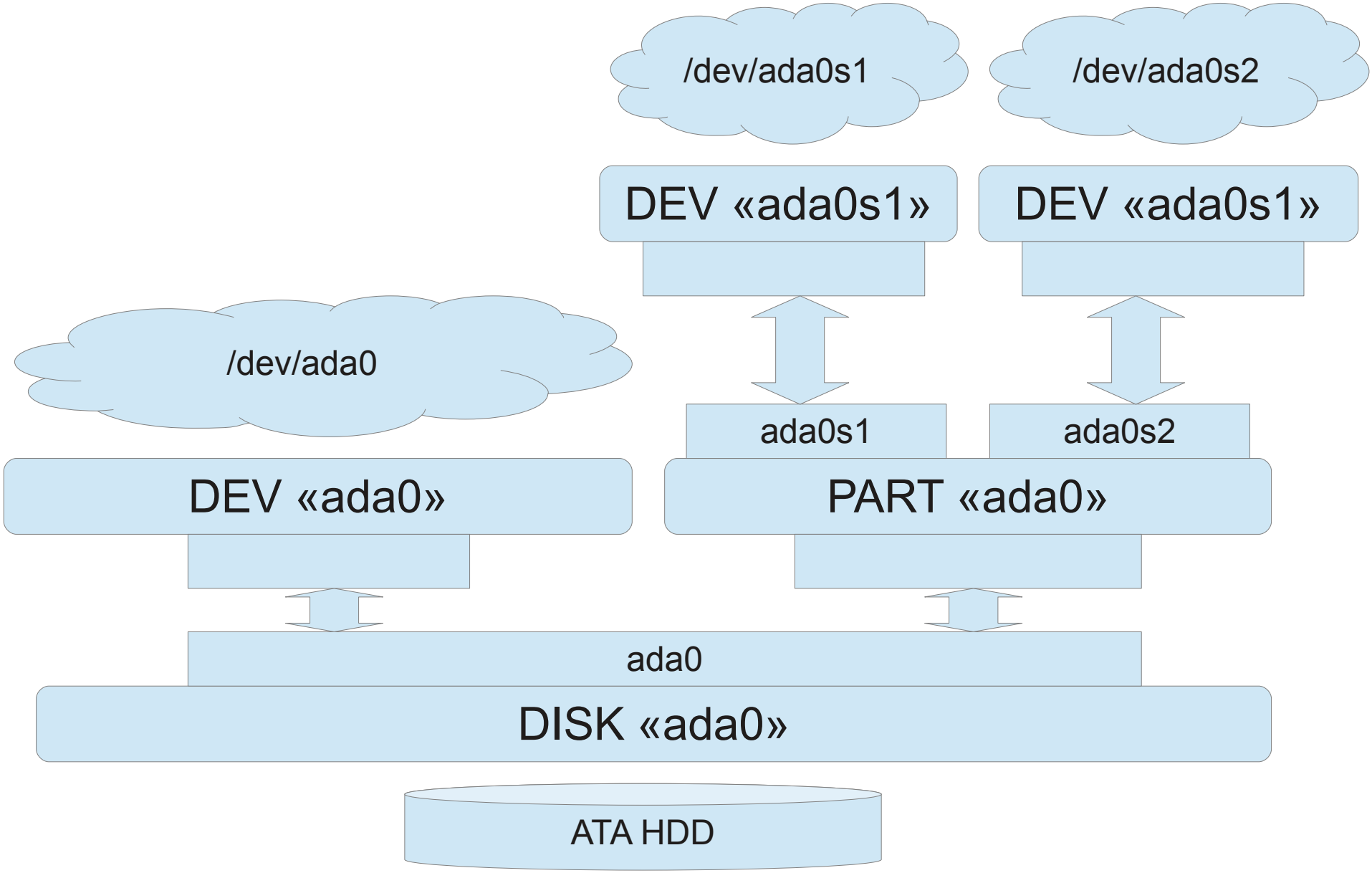
GEOM topology



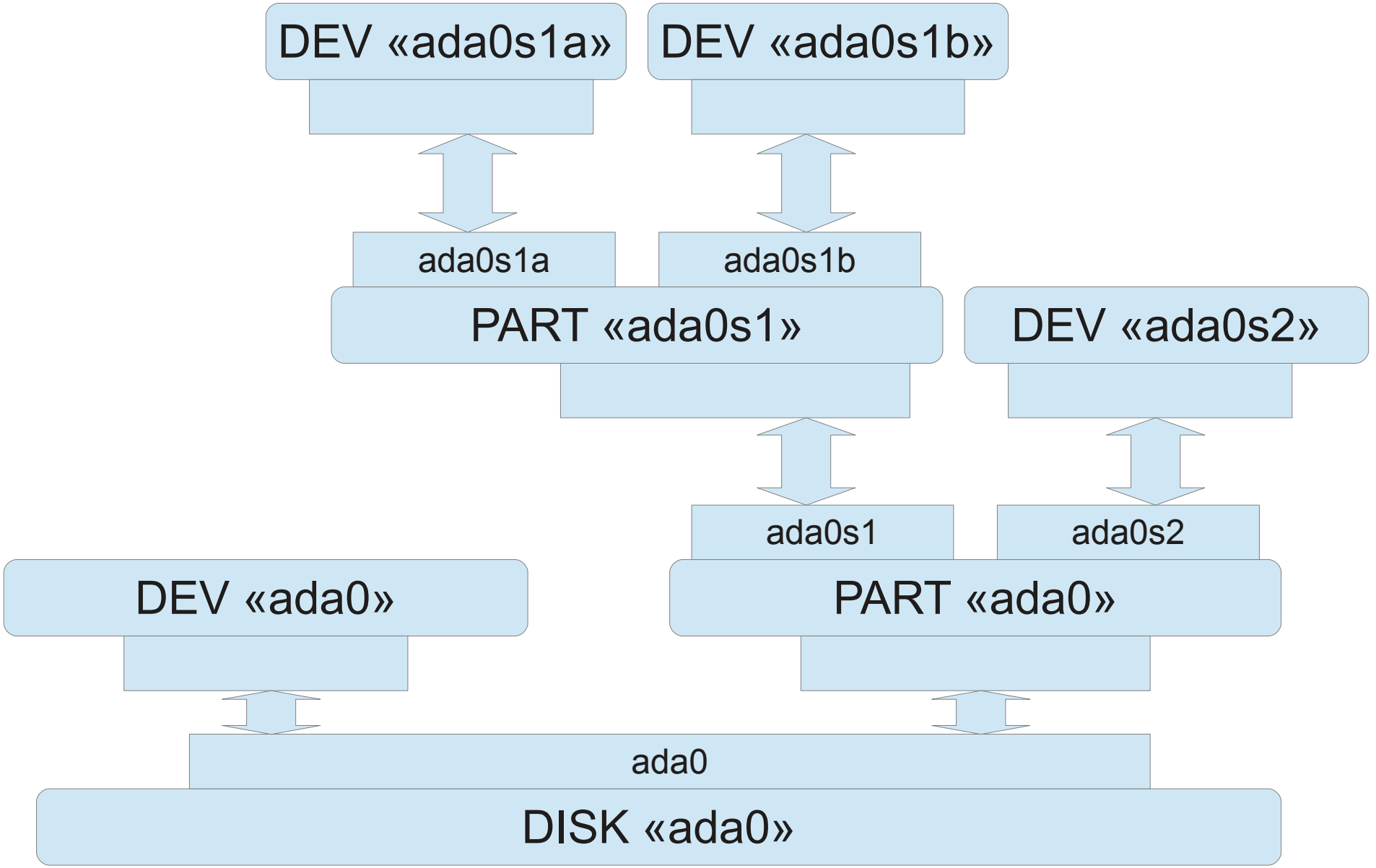
Mounted UFS in GEOM



Disk partitioning in GEOM



Cascaded disk partitioning



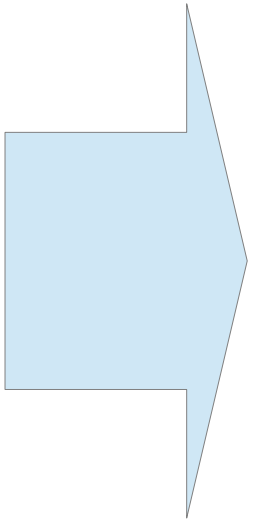
GEOM functionality

- Tasting
- Orphanization
- Spoiling
- Configuration

- I/O procesing

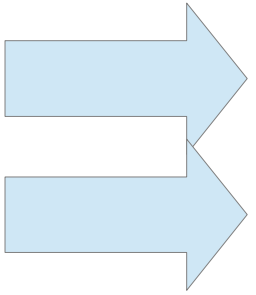
GEOM in threads

- Tasting
- Orphanization
- Spoiling
- Configuration



g_event

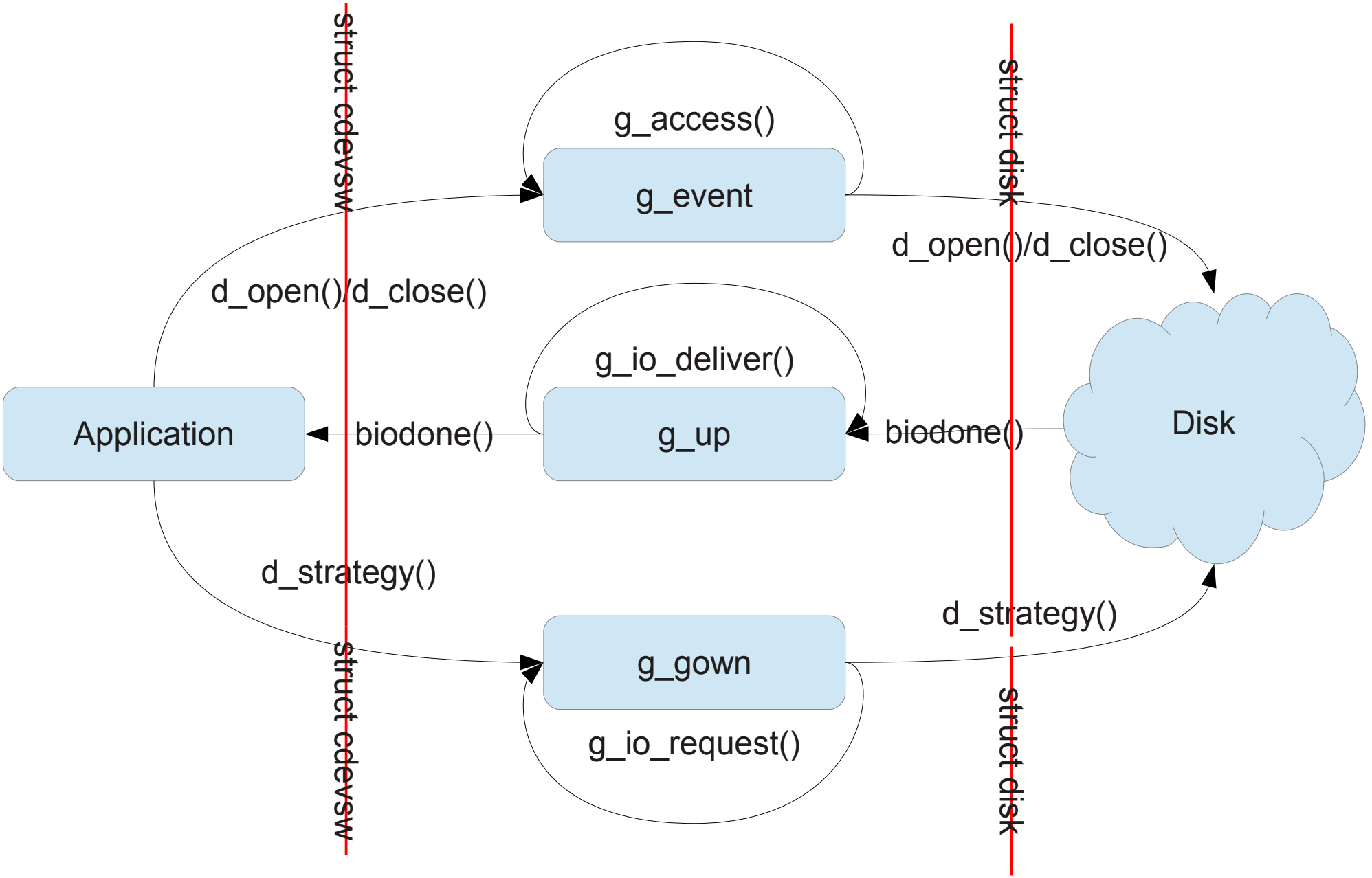
- I/O submission
- I/O completion



g_down

g_up

GEOM calls and threads



Block storages below disk(9)

- SCSI disks/CD/DVD
- ATA/ATAPI disks/CD/DVD
- MMC/SD cards
- NAND flash
- Proprietary block devices:
  - nvme(4)/nvd(4)
  - mfi(4)
  - aac(4)
  - ...

ATA/SCSI block devices before 9.0

ATA – ata(4)

- ad: disk(9) → ATA
- afd: disk(9) → SCSI
- acd: disk(9) → SCSI
- atapicam: wrapper
- ATA bus
- ATA command queue
- ATA HBA drivers

SCSI – CAM

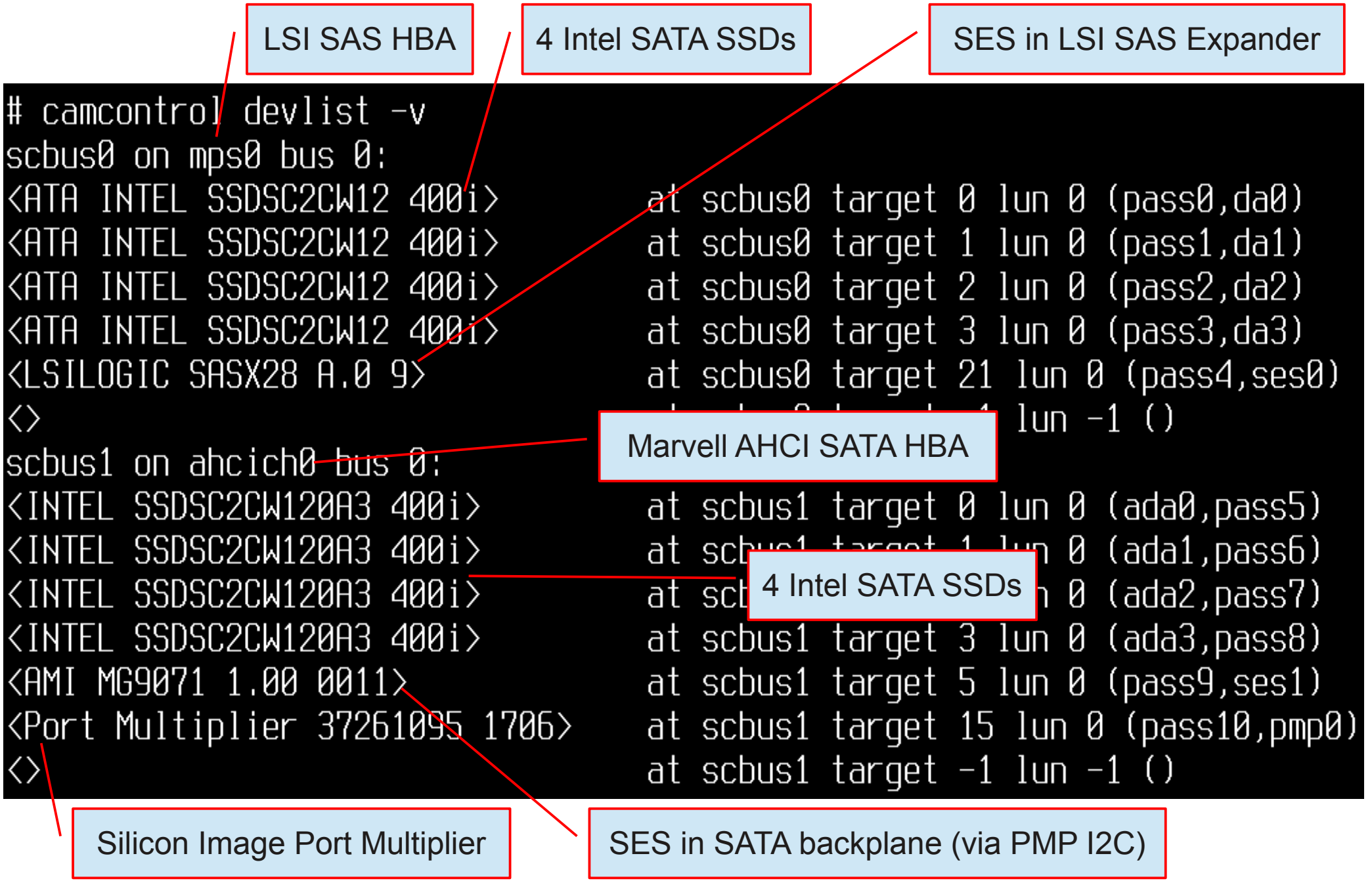
- da: disk(9) → SCSI
- cd: disk(9) → SCSI
- SPI bus
- SCSI command queue
- SCSI HBA drivers

ATA/SCSI block devices after 9.0

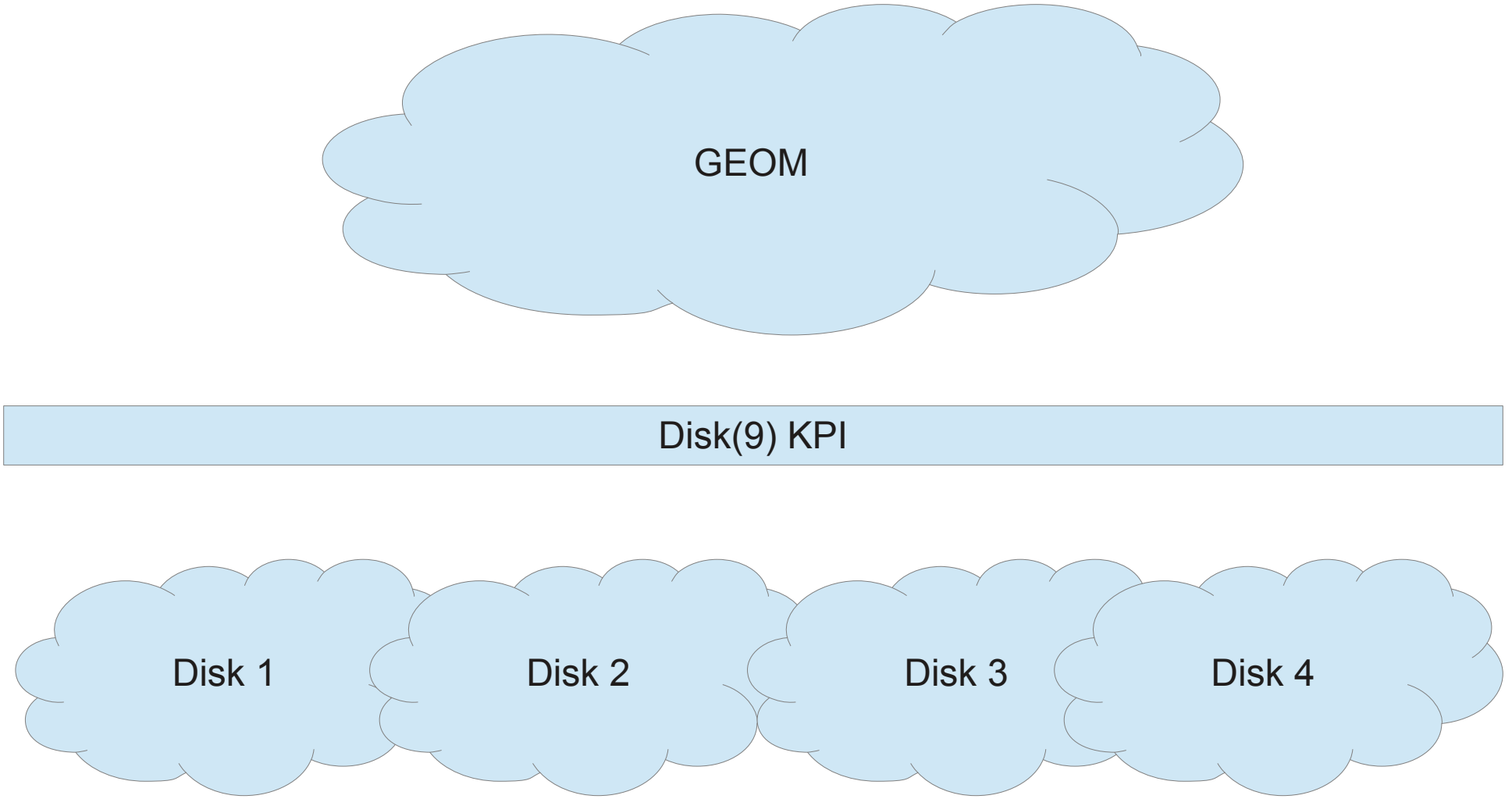
CAM handling both ATA and SCSI

- ada: disk(9) → ATA
- da: disk(9) → SCSI
- cd: disk(9) → SCSI
- Virtualized bus: ATA, SATA, SPI, SAS, ...
- Unified ATA/SCSI command queue
- Unified ATA/SCSI HBA drivers

Unified diversity



Back to a wider view

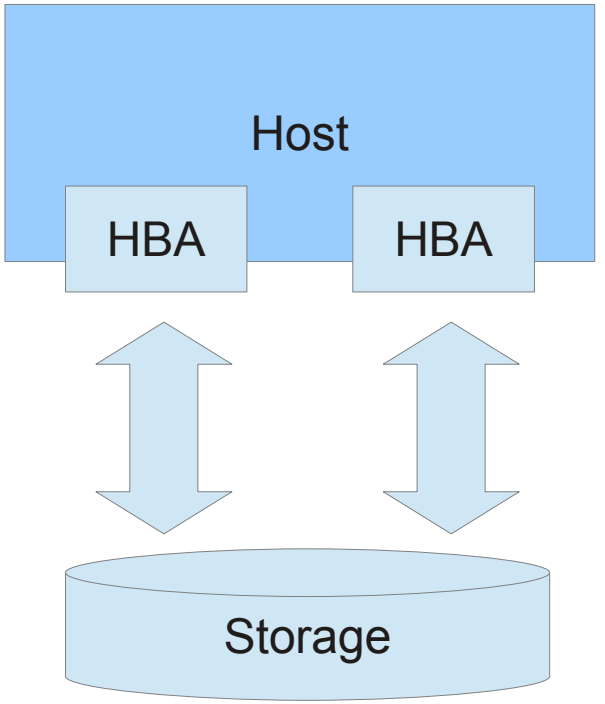


Disk multipath

- 2+ SAS HBAs + dual-expander JBOD + SAS disks;
- 2+ FC HBAs + storage with several FC ports;
- iSCSI initiator and target with 2+ NICs each;
- ...

=

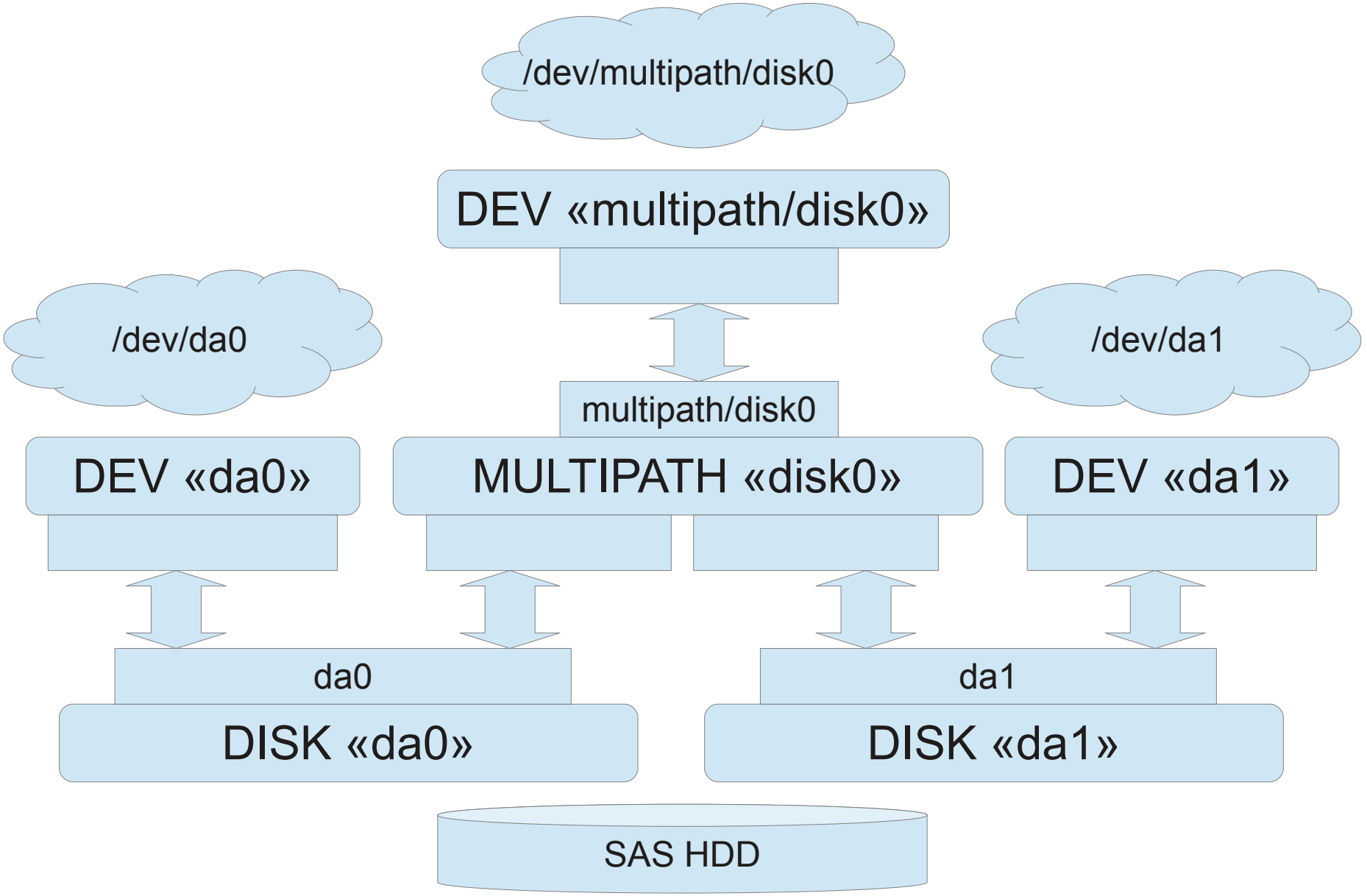
- Improved reliability
- Improved performance



```
# geom disk list
Geom name: da0
Providers:
1. Name: da0
   Mediasize: 750156374016 (698G)
   Sectorsize: 512
   Mode: r0w0e0
   descr: SEAGATE ST3750630SS
   lunid: 5000c50006812d23
   ident: 3QK0A63P00009832U6PM
   fwsectors: 63
   fwheads: 255

Geom name: da1
Providers:
1. Name: da1
   Mediasize: 750156374016 (698G)
   Sectorsize: 512
   Mode: r0w0e0
   descr: SEAGATE ST3750630SS
   lunid: 5000c50006812d23
   ident: 3QK0A63P00009832U6PM
   fwsectors: 63
   fwheads: 255
```

Disk multipath in GEOM



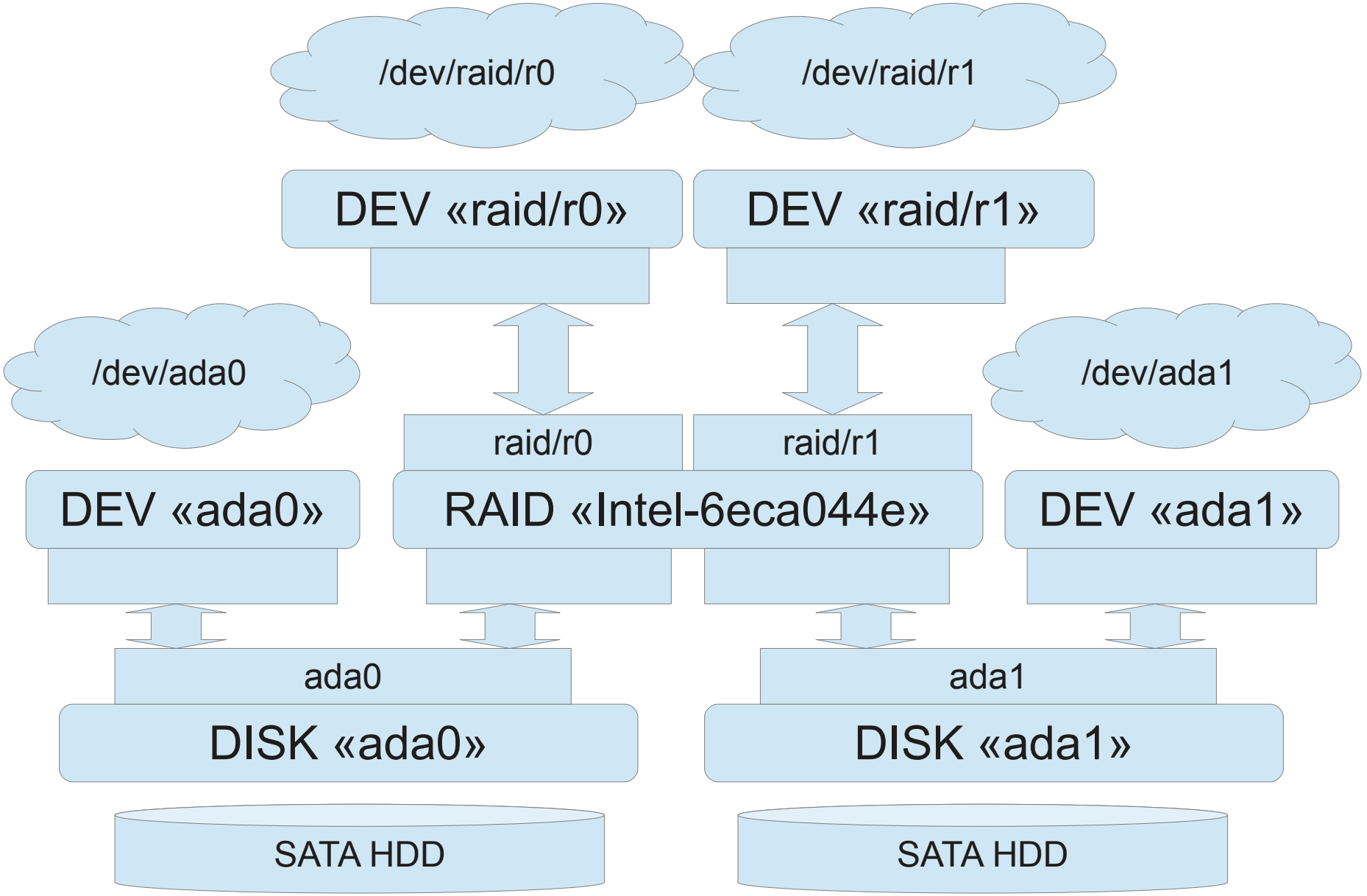
```
# gmultipath list
Geom name: disk0
Type: AUTOMATIC
Mode: Active/Passive
UUID: 5b5b69c0-1d3e-11e3-a992-00259062ec50
State: OPTIMAL
Providers:
1. Name: multipath/disk0
   Mediasize: 750156373504 (698G)
   Sectorsize: 512
   Mode: r0w0e0
   State: OPTIMAL
Consumers:
1. Name: da0
   Mediasize: 750156374016 (698G)
   Sectorsize: 512
   Mode: r1w1e1
   State: ACTIVE
2. Name: da1
   Mediasize: 750156374016 (698G)
   Sectorsize: 512
   Mode: r1w1e1
   State: PASSIVE
```

BIOS-assisted «Fake» RAID

```
# geom disk list
Geom name: ada0
Providers:
1. Name: ada0
   Mediasize: 60022480896 (55G)
   Sectorsize: 512
   Mode: r1w1e1
   descr: D2CSTK251A10-0060
   lunid: 5e83a97010017a69
   ident: A1290011215000012
   fwsectors: 63
   fwheads: 16

Geom name: ada1
Providers:
1. Name: ada1
   Mediasize: 60022480896 (55G)
   Sectorsize: 512
   Mode: r1w1e1
   descr: D2CSTK251A10-0060
   lunid: 5e83a9701001433a
   ident: A1290011214000258
   fwsectors: 63
   fwheads: 16
```

BIOS-assisted RAID in GEOM



```
# graid list
Geom name: Intel-6eca044e
State: OPTIMAL
Metadata: Intel
Providers:
1. Name: raid/r0
   Mediasize: 32212254720 (30G)
   Sectorsize: 512
   Mode: r0w0e0
   Subdisks: ada0 (ACTIVE), ada1 (ACTIVE)
   Dirty: No
   State: OPTIMAL
   Strip: 65536
   Components: 2
   Transformation: RAID1
   RAIDLevel: RAID1
   Label: Volume0
   descr: Intel RAID1 volume
2. Name: raid/r1
   Mediasize: 55610179584 (51G)
   Sectorsize: 512
   Stripesize: 65536
   Stripeoffset: 0
   Mode: r0w0e0
   Subdisks: ada0 (ACTIVE), ada1 (ACTIVE)
   Dirty: No
   State: OPTIMAL
   Strip: 65536
   Components: 2
   Transformation: RAID0
   RAIDLevel: RAID0
   Label: Volume1
   descr: Intel RAID0 volume
```

BIOS-assisted RAID in GEOM

```
Consumers:
1. Name: ada0
   Mediasize: 60022480896 (55G)
   Sectorsize: 512
   Mode: r1w1e1
   ReadErrors: 0
   Subdisks: r0(Volume0):0@0, r1(Volume1):0@32214614016
   State: ACTIVE (ACTIVE, ACTIVE)
2. Name: ada1
   Mediasize: 60022480896 (55G)
   Sectorsize: 512
   Mode: r1w1e1
   ReadErrors: 0
   Subdisks: r0(Volume0):1@0, r1(Volume1):1@32214614016
   State: ACTIVE (ACTIVE, ACTIVE)

# graid status
   Name   Status  Components
raid/r0  OPTIMAL  ada0 (ACTIVE (ACTIVE, ACTIVE))
                  ada1 (ACTIVE (ACTIVE, ACTIVE))
raid/r1  OPTIMAL  ada0 (ACTIVE (ACTIVE, ACTIVE))
                  ada1 (ACTIVE (ACTIVE, ACTIVE))
```

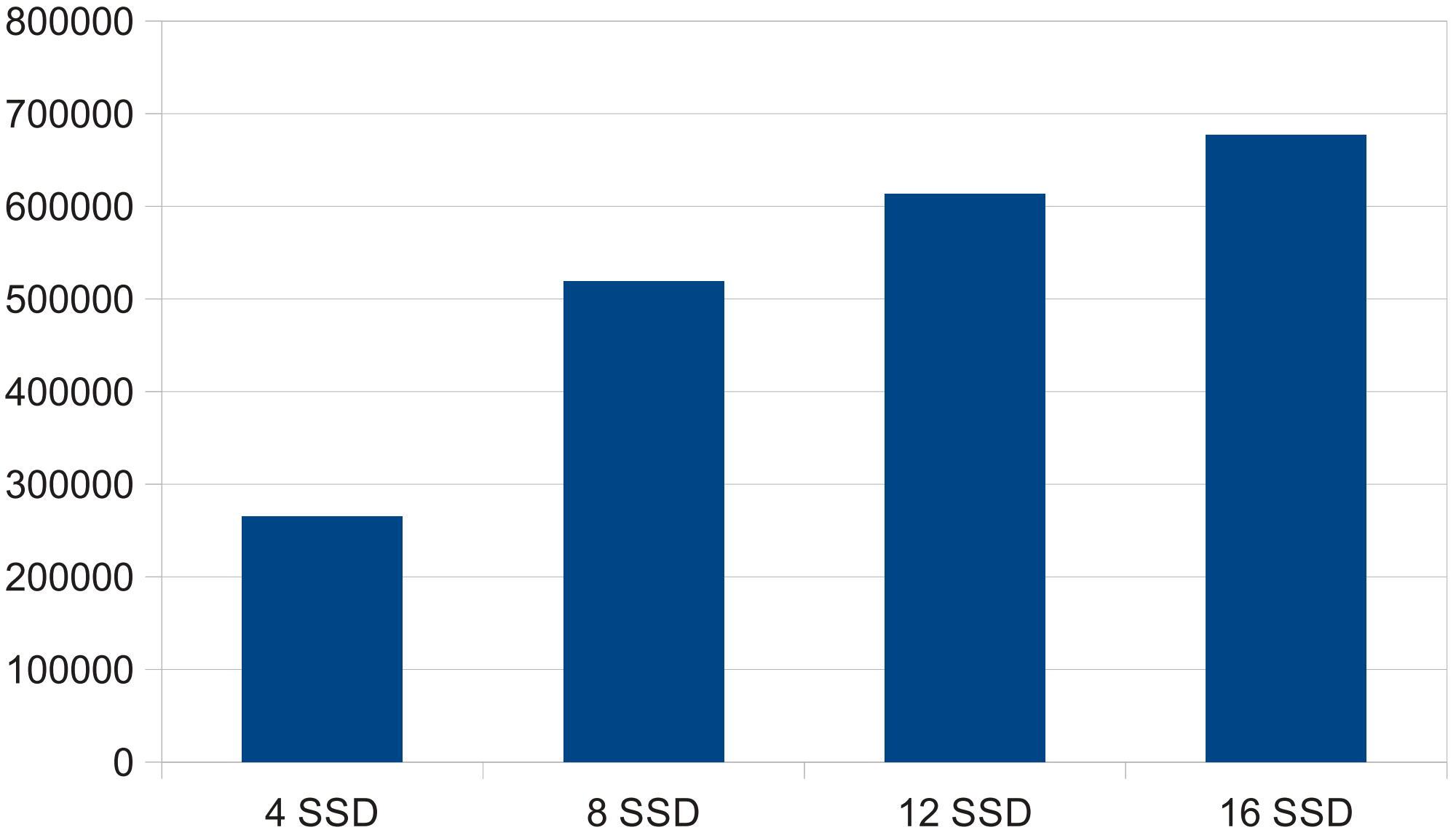
Is GEOM fast?

Test setup:

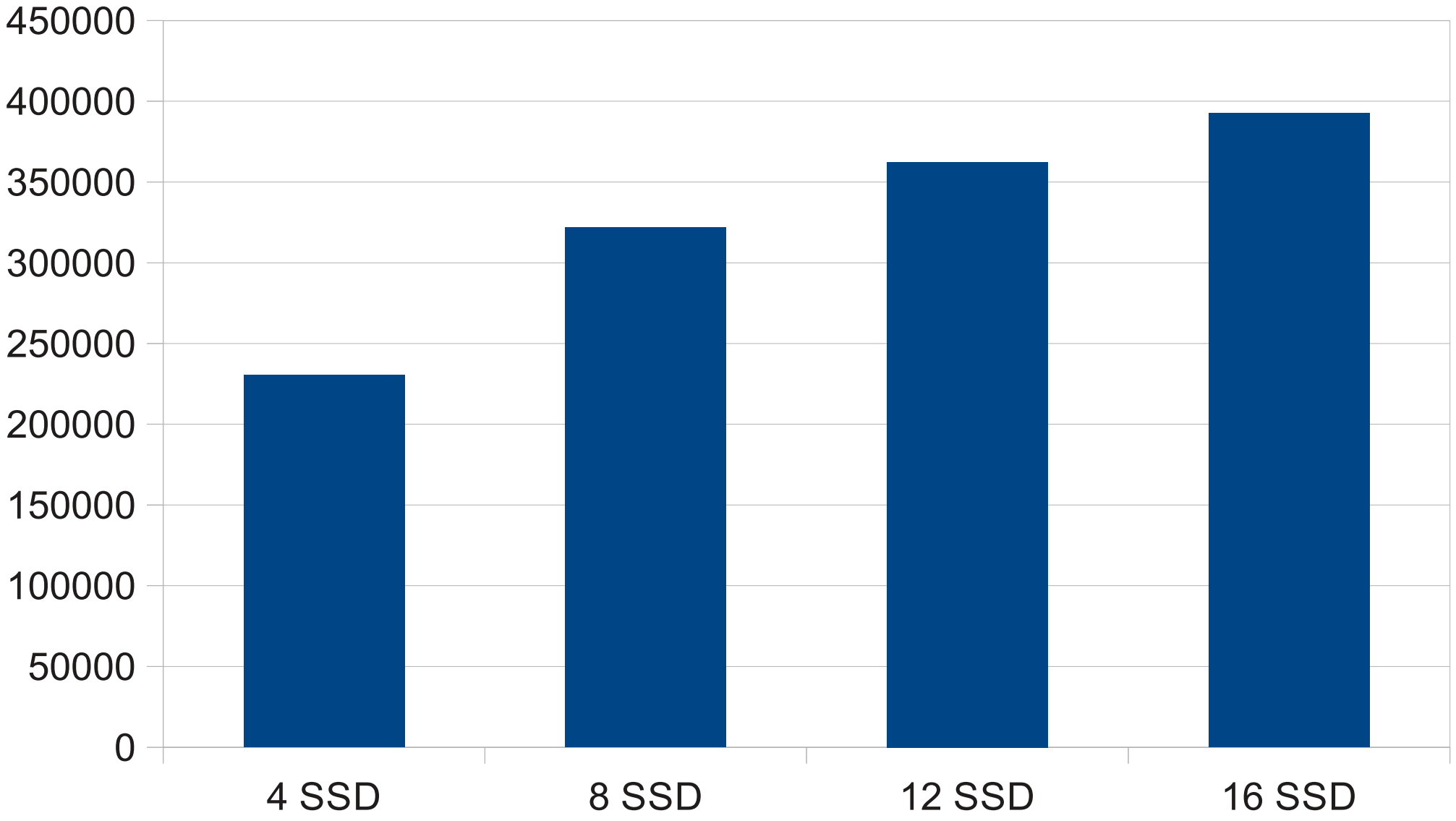
- 4 LSI 6Gbps SAS HBAs
- 16 6Gbps SATA SSDs
- Platform 1:
  - Intel Core i7-3930K, 6x2 cores @ 3.2GHz
  - ASUS P9X79 WS
- Platform 2:
  - 2x Intel Xeon E5645, 2x6x2 cores @ 2.4GHz
  - Supermicro X8DTU

Test: Total number of IOPS from many instances of `dd if=/dev/daX of=/dev/null bs=512`

Platform 1: Core i7-3930K 3.2GHz



Platform 2: 2xXeon E5645 2.4GHz



Can GEOM be made faster? Yes!

```
last pid:  1960;  load averages:  4.69,  2.00,  0.92    up 0+00:07:19  08:43:49
929 processes: 31 running, 840 sleeping, 58 waiting
CPU:  1.4% user,  0.0% nice, 13.8% system,  5.5% interrupt, 79.3% idle
Mem: 162M Active, 54M Inact, 570M Wired, 60K Cache, 35M Buf, 34G Free
ARC: 443K Total, 4K MFU, 356K MRU, 16K Anon, 10K Header, 57K Other
Swap:

  PID USERNAME   PRI NICE   SIZE    RES STATE   C   TIME    WCPU COMMAND
   13 root        -8    -     0K    48K CPU7    7   1:59 100.00% geom{g_down}
   13 root        -8    -     0K    48K CPU8    8   1:11  72.07% geom{g_up}
   12 root       -68    -     0K   960K CPU0    0   1:07  56.88% intr{swi2: cam
   12 root       -88    -     0K   960K WAIT   14   0:10  16.89% intr{irq276: m
   12 root       -88    -     0K   960K WAIT   12   0:37  16.80% intr{irq274: m
   12 root       -88    -     0K   960K CPU13  13   0:18  16.26% intr{irq275: m
   12 root       -88    -     0K   960K WAIT   15   0:03  12.70% intr{irq277: m
 1887 root        20    0 12196K  1952K physrd 18   0:00   0.49% dd
 1890 root        20    0 12196K  1952K physrd 21   0:00   0.49% dd
 1858 root        20    0 12196K  1952K physrd  2   0:00   0.49% dd
 1882 root        20    0 12196K  1952K physrd 20   0:00   0.49% dd
 1829 root        20    0 12196K  1952K physrd 15   0:00   0.49% dd
 1891 root        20    0 12196K  1952K physrd 12   0:00   0.49% dd
 1784 root        20    0 12196K  1952K physrd  6   0:00   0.49% dd
 1951 root        20    0 12196K  1952K physrd  9   0:00   0.49% dd
 1848 root        20    0 12196K  1952K physrd  2   0:00   0.49% dd
 1772 root        20    0 12196K  1952K physrd  9   0:00   0.39% dd
 1690 root        20    0 12196K  1952K physrd  2   0:00   0.39% dd
```

Bottlenecks

Can GEOM be made faster? Yes!

Bottlenecks:

- 5 threads and up to 10 swiches per request:
  dd, g_down, HBA HWI, CAM SWI, g_up
- GEOM threads are capped at 100% CPU
- Congested per-HBA locks in CAM

Solutions:

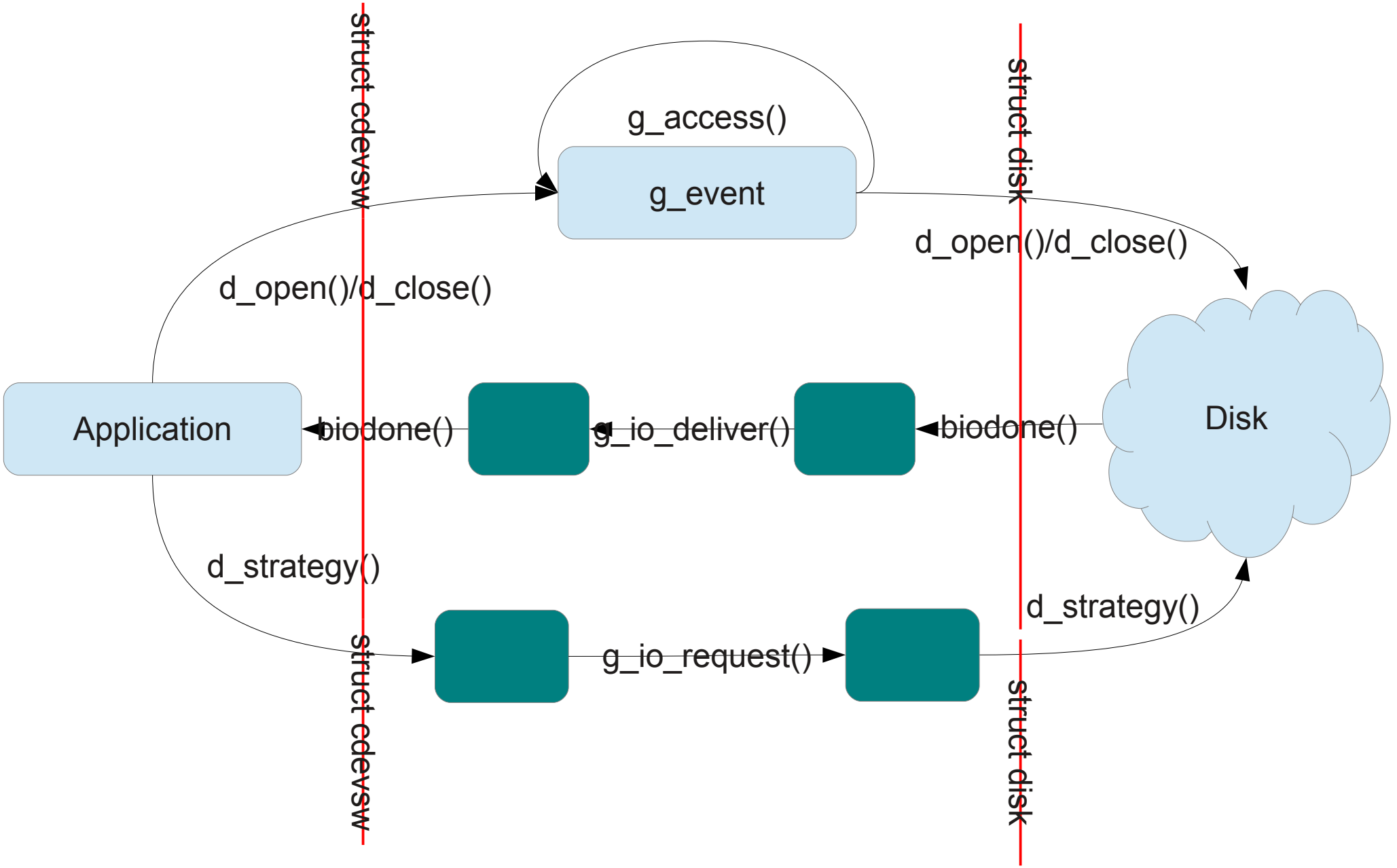
- Direct dispatch in GEOM
- Improved CAM locking
- More completion threads or direct dispatch in CAM

Direct dispatch in GEOM

Requirements:
- Caller should not hold any locks
- Caller should be reenterable
- Callee should not depend on g_up / g_down threads semantics
- Kernel thread stack should not overflow

Implementation:
- Per-consumer/-provider flags to declare caller and callee capabilities
- Kernel thread stack usage estimation

Direct dispatch in GEOM

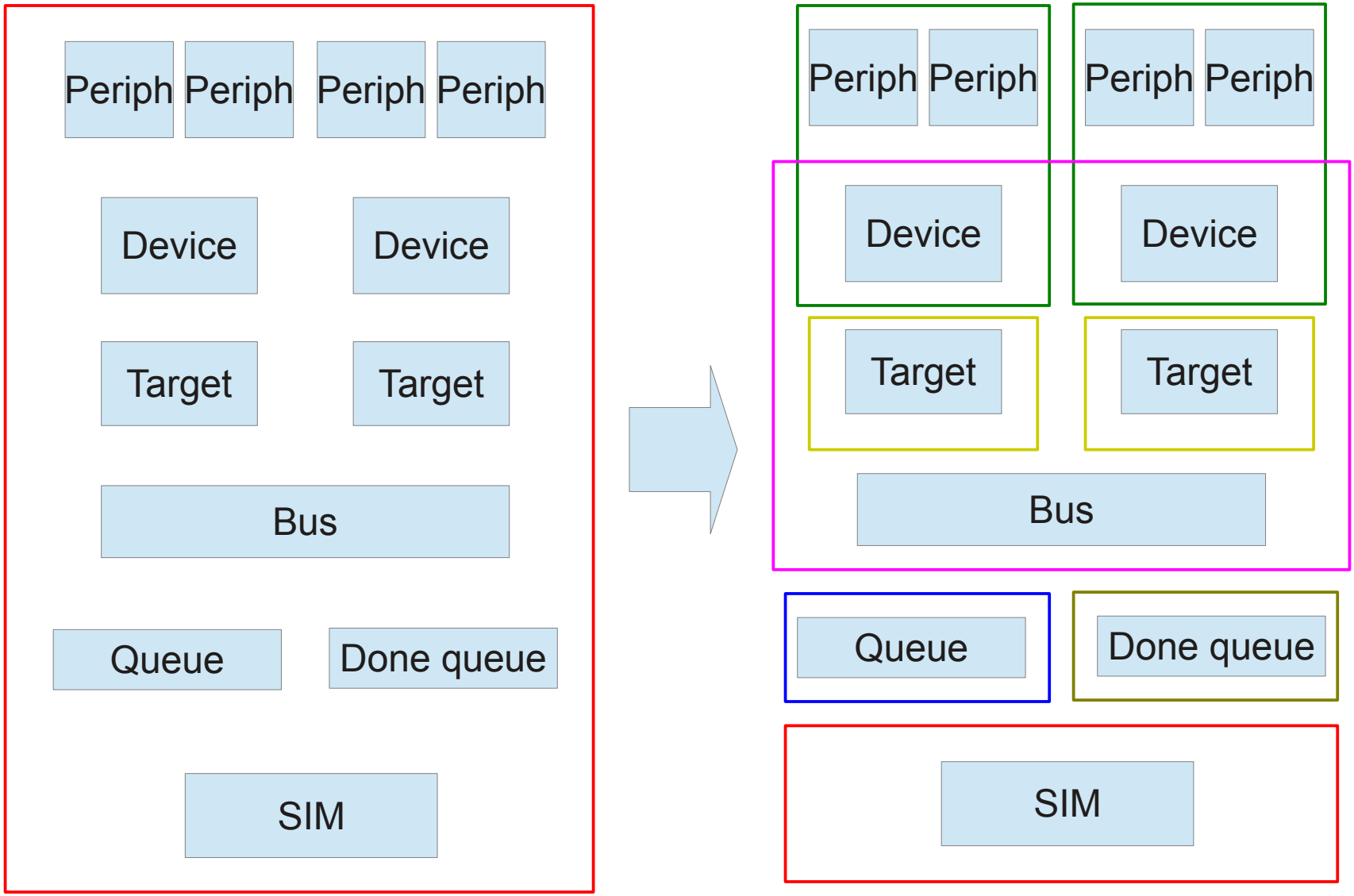


Improved CAM locking

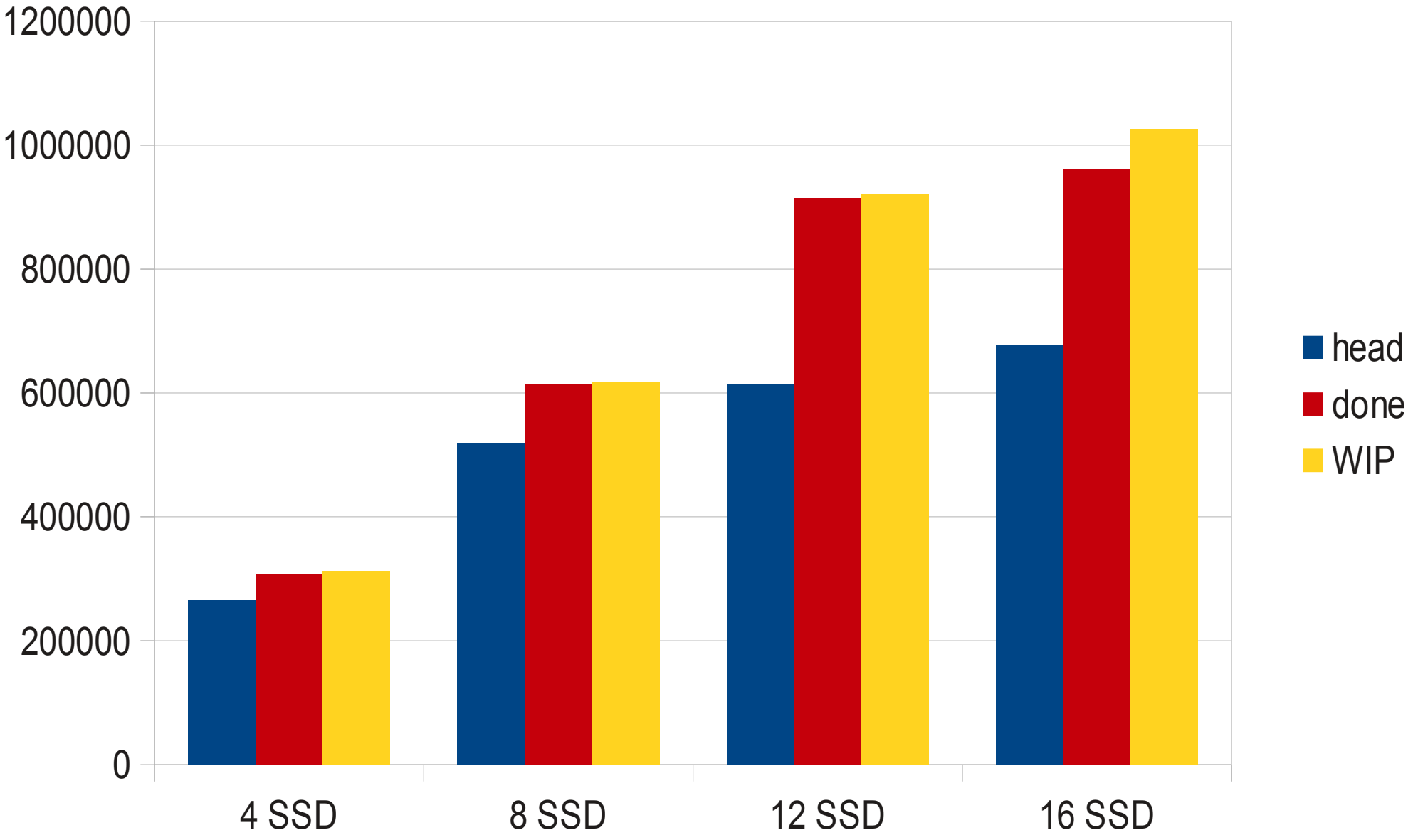
Before:
• Per-SIM locks protect everything for one SIM (HBA) from periph drivers state to HBA hardware access

After:
• Per-SIM locks protect only HBA, keeping KPI/KBI
• Queue locks protect CCB queues and serialise SIM calls to reduce SIM locks congestions
• Per-bus locks protect reference counting
• Per-target locks protect list of LUNs
• Per-LUN locks protect device and periph

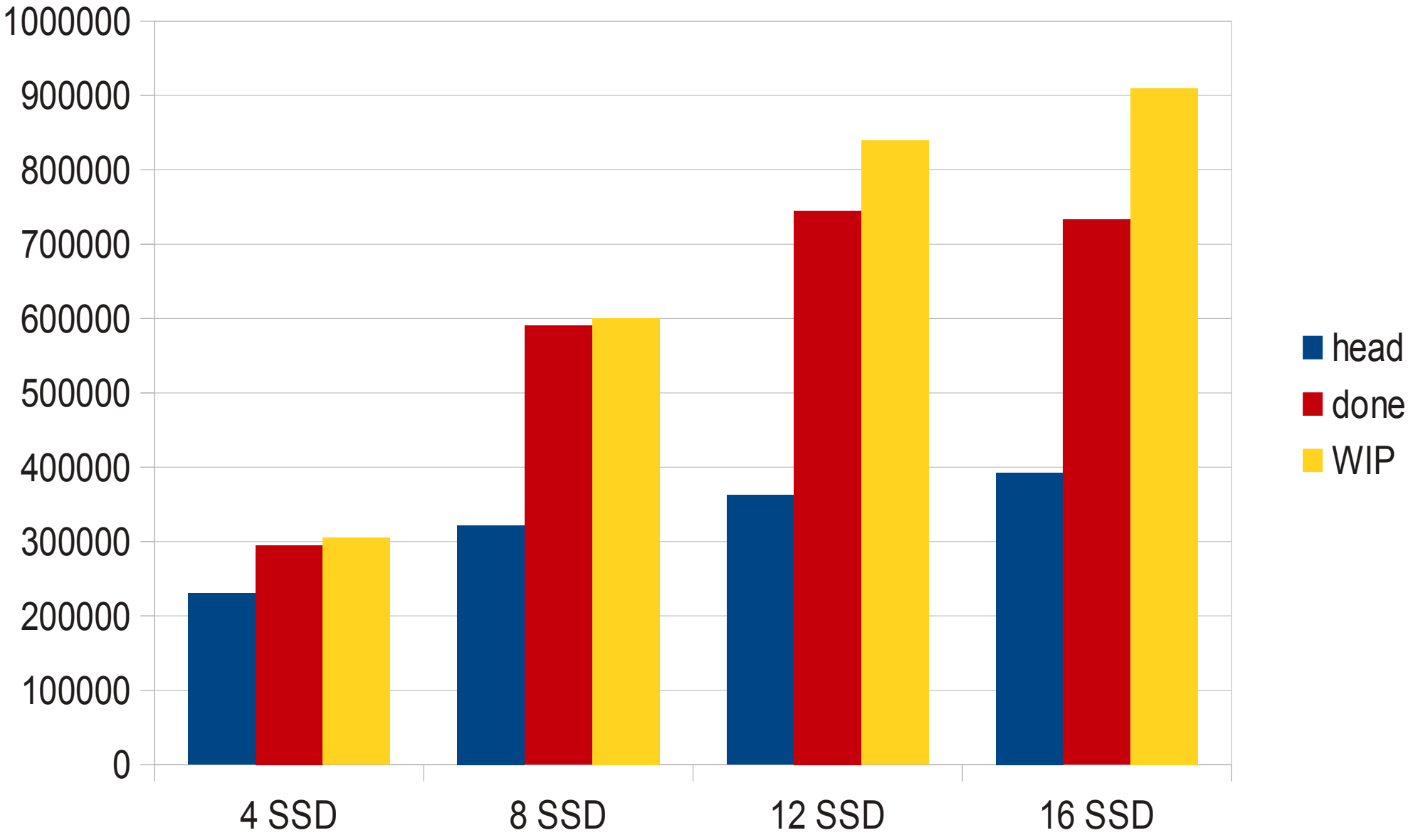
Improved CAM locking



Platform 1: Core i7-3930K 3.2GHz



Platform 2: 2xXeon E5645 2.4GHz



Can we do even more? Possibly!

```
last pid:  1498;  load averages: 24.51, 13.03,  5.66
946 processes: 49 running, 824 sleeping, 73 waiting
CPU:  4.2% user,  0.0% nice, 63.8% system,  6.7% interrupt, 25.3% idle
Mem: 156M Active, 37M Inact, 575M Wired, 2116K Cache, 31M Buf, 34G Free
ARC: 443K Total, 4K MFU, 356K MRU, 16K Anon, 10K Header, 57K Other
Swap:

  PID USERNAME   PRI NICE   SIZE    RES STATE   C   TIME    WCPU COMMAND
    2 root       -16    -     0K   128K CPU0    0   2:44 100.00% cam{doneq4}
    2 root       -16    -     0K   128K CPU4    4   2:44  99.56% cam{doneq6}
    2 root       -16    -     0K   128K CPU15  15   2:31  87.35% cam{doneq5}
    2 root       -16    -     0K   128K CPU2    2   2:30  82.28% cam{doneq0}
    2 root       -16    -     0K   128K CPU8    8   2:25  75.88% cam{doneq3}
    2 root       -16    -     0K   128K RUN    13   1:28  46.09% cam{doneq1}
    2 root       -16    -     0K   128K CPU21  21   1:23  44.09% cam{doneq2}
   12 root       -88    -     0K  1184K CPU22  22   1:12  41.06% intr{irq274: mps0}
   12 root       -88    -     0K  1184K CPU20  20   1:07  40.77% intr{irq277: mps3}
   12 root       -88    -     0K  1184K CPU21  21   1:04  38.28% intr{irq275: mps1}
   12 root       -88    -     0K  1184K WAIT   14   0:58  37.50% intr{irq276: mps2}
 1244 root        24    0 12196K  1952K CPU18  18   0:12   6.40% dd
 1276 root        23    0 12196K  1952K RUN    17   0:13   6.30% dd
 1437 root        23    0 12196K  1952K physrd  8   0:11   5.57% dd
 1214 root        23    0 12196K  1952K physrd  4   0:11   5.47% dd
 1207 root        23    0 12196K  1952K physrd  4   0:11   5.47% dd
 1457 root        23    0 12196K  1952K physrd  1   0:11   5.37% dd
 1250 root        22    0 12196K  1952K physrd  1   0:11   5.37% dd
 1438 root        22    0 12196K  1952K physrd 19   0:10   5.37% dd
 1275 root        23    0 12196K  1952K physrd  8   0:11   5.27% dd
 1447 root        23    0 12196K  1952K CPU17  17   0:11   5.27% dd
 1211 root        22    0 12196K  1952K physrd 10   0:11   5.27% dd
 1439 root        22    0 12196K  1952K physrd 22   0:11   5.27% dd
 1210 root        23    0 12196K  1952K physrd  9   0:11   5.27% dd
 1451 root        23    0 12196K  1952K physrd  8   0:11   5.18% dd
```

Bottlenecks

Context switches

Multiple queues/IRQs support

```
ahci0@pci0:0:31:2:      class=0x010400 card=0x060015d9 chip=0x28228086 rev
    vendor     = 'Intel Corporation'
    device     = '82801 SATA Controller [RAID mode]'
    class      = mass storage
    subclass   = RAID
    cap 05[80] = MSI supports 16 messages enabled with 16 messages
    cap 01[70] = powerspec 3  supports D0 D3  current D0
    cap 12[a8] = SATA Index-Data Pair
    cap 13[b0] = PCI Advanced Features: FLR TP
mps0@pci0:5:0:0:        class=0x010700 card=0x30201000 chip=0x00871000 rev
    vendor     = 'LSI Logic / Symbios Logic'
    device     = 'SAS2308 PCI-Express Fusion-MPT SAS-2'
    class      = mass storage
    subclass   = SAS
    cap 01[50] = powerspec 3  supports D0 D1 D2 D3  current D0
    cap 10[68] = PCI-Express 2 endpoint max data 256(4096) FLR link x8(x8)
                 speed 5.0(8.0) ASPM disabled(L0s)
    cap 03[d0] = VPD
    cap 05[a8] = MSI supports 1 message, 64 bit
    cap 11[c0] = MSI-X supports 16 messages, enabled
                 Table in map 0x14[0xe000], PBA in map 0x14[0xf000]
```

Work In Progress

- Commit the CAM and GEOM changes.
- Add multiple queues support to HBA drivers.
- File systems, schedulers and other places outside block storage also need work to keep up. Join!

Questions?