CS 447/647

Storage
Disk Management

Overview

- Hardware
- /sys files
- /dev files
- Partitioning
- **LVM**
- **mdraid**
- **drbd (light overview)**

References

Clausen, Andrew. “Parted User’s Manual.” *Parted User's Manual*, 2019, www.gnu.org/software/parted/manual/parted.html.

Nemeth, Evi, et al. *UNIX and Linux System Administration Handbook*. Addison-Wesley, 2018.

Hardware

- Interfaces
  - SATA - Consumer 6Gb/s
    - Consumer
  - SAS - Enterprise 24Gb/s, i2c sideband interface
    - Higher Reliability, Higher Cost, ‘Enterprise’
  - **PCIe - NVMe, Speed**
  - M.2 - Single connector for SATA+NVMe
  - U.2 - Connector for 2.5” enterprise NVMe drives
  - EDSFF - Latest form factor for enterprise NVMe drives
    - Connectors are E1 and E3
  - Fiber Channel - Optical Fiber, 128Gb/s
    - Used in high-end network attached storage appliances (NetApp, Pure, etc.)
- Disk Type
  - Hard Drive - Rotational 7,200RPMs
    - Rotational is prone to mechanical failure
  - Solid-State Disk - NAND Flash Memory Storage
  - eMMC - Embedded NAND Flash Memory Storage (Single Board Computers, SBCs)
  - USB
  - 3D XPoint (pronounced crosspoint)

Disk Sector Formats

- 512n - Native 512-byte sectors
- 4Kn - Native 4KiB sectors
- 512e - 4KiB sectors but 512-byte emulation
  - Most Common

4K block = 8 × 512-byte blocks								OS File System
0	1	2	3	4	5	6	7	Logical Blocks
4K Physical Sector #1								Physical Sectors

Figure 4: 512-byte Emulated Device Sector Size

Logical Blocks [0 to 7]								Logical Blocks [8 to 15]							
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4K Physical Sector #1								4K Physical Sector #2							

Figure 5: 512-byte Emulated Device Sector Size—Two "aligned" sectors

Partition Alignment

Since most modern operating systems will write in 4K blocks, it is important that each 4K logical block is aligned to a physical 4K block on the disk (see Figure 5). This is especially important because the 512e feature of the drive cannot prevent a partitioning utility from creating a misaligned partition. When misalignment occurs, a logical 4K block will reside on two physical sectors. In this case, a single read or write of a 4K block will result in a read/write of two physical sectors. The impact of a "read" is minimal, whereas a single write will cause two "Read-Modify-Writes" to occur, potentially impacting performance (see Figure 6)..

				4K logical block = 8 × 512-byte blocks [4:11]											
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4K Physical Sector #1								4K Physical Sector #2							

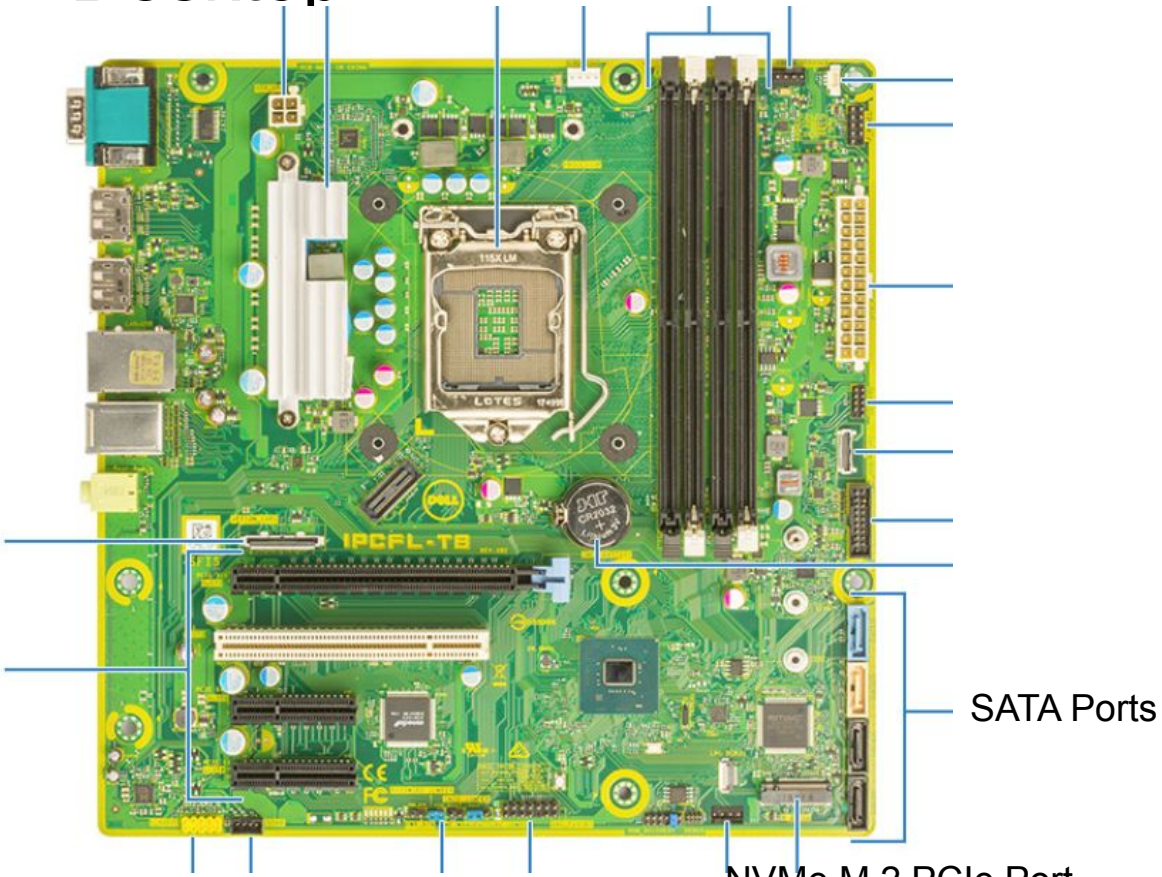
Figure 6: 512-byte Emulated Device Sector Size (MISALIGNED)

https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/white-paper/white-paper-advanced-format.pdf

Optimal sectors

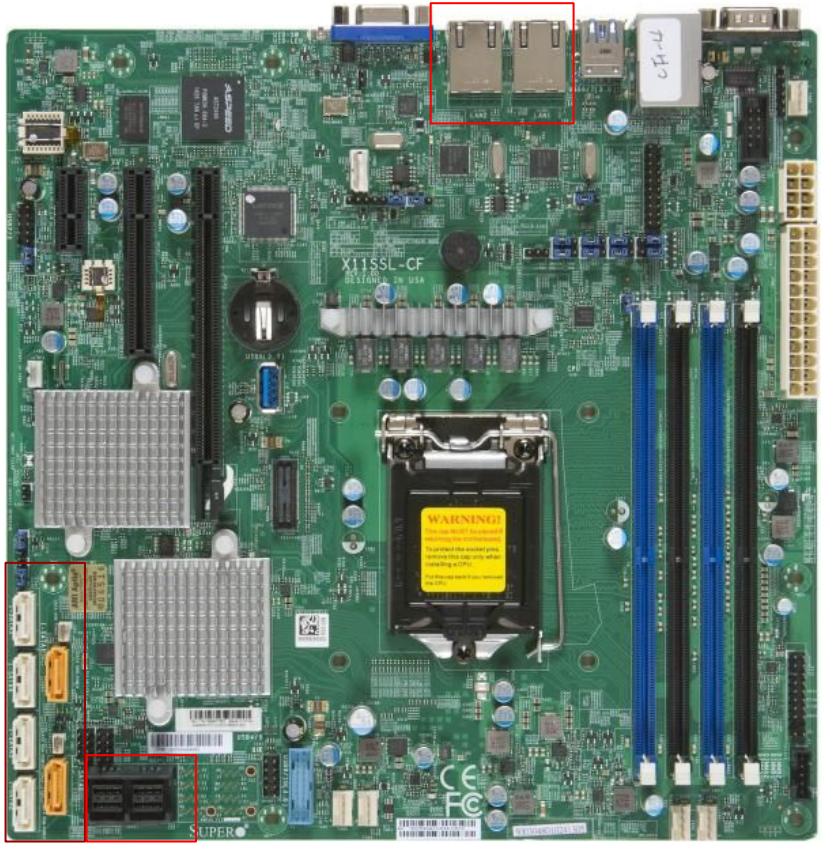
GNU Parted 2.1+ use “-a optimal” or “-a minimal” options

Hardware - Desktop



Hardware - Server

2 x Ethernet



SATA

SAS3

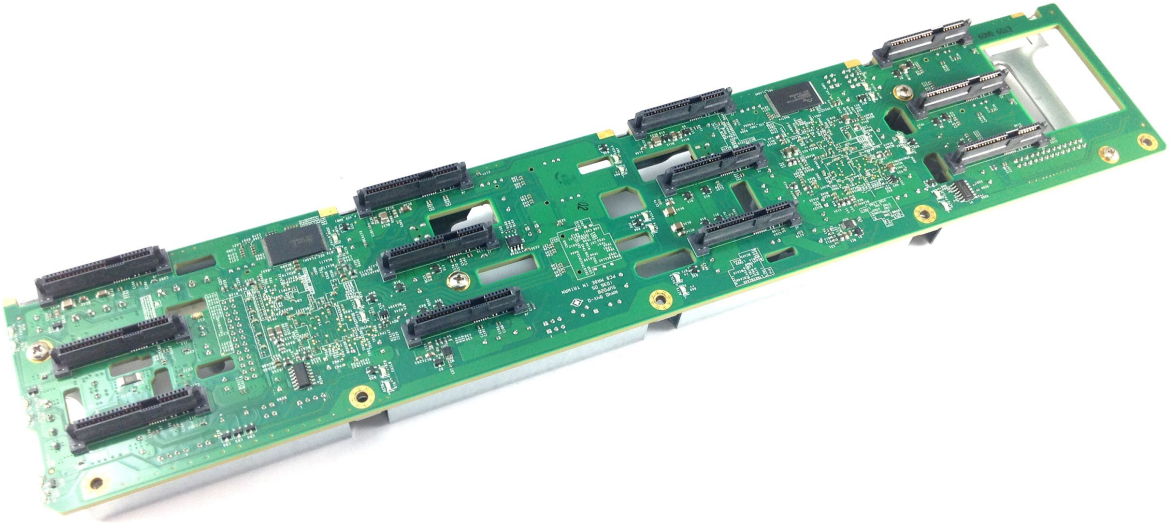
Supermicro

Hardware - Host Based Adapter

- 16 Ports
- 12Gb/s per port
- Just a bunch of disks (JBOD) 240 disks

Supermicro

Hardware Server



Supermicro

Hardware Server

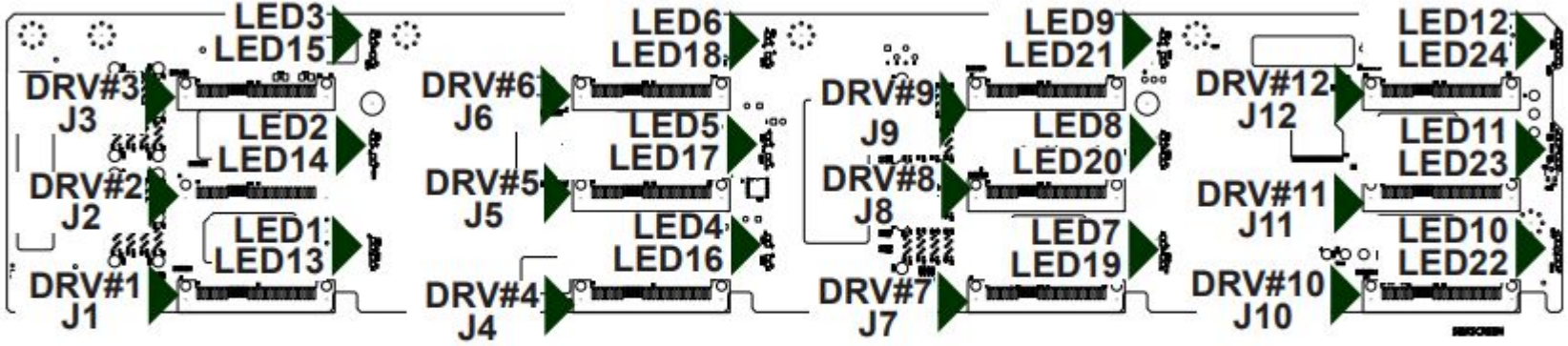
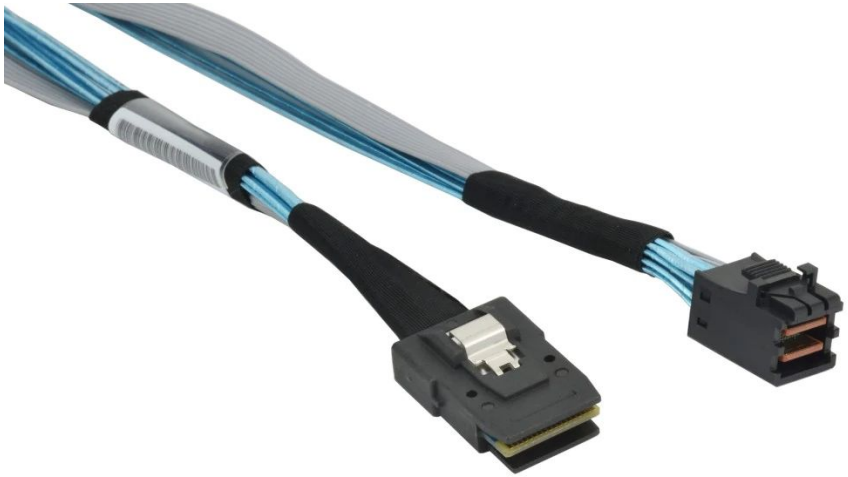


Figure 2-5: Rear Connectors and LEDs

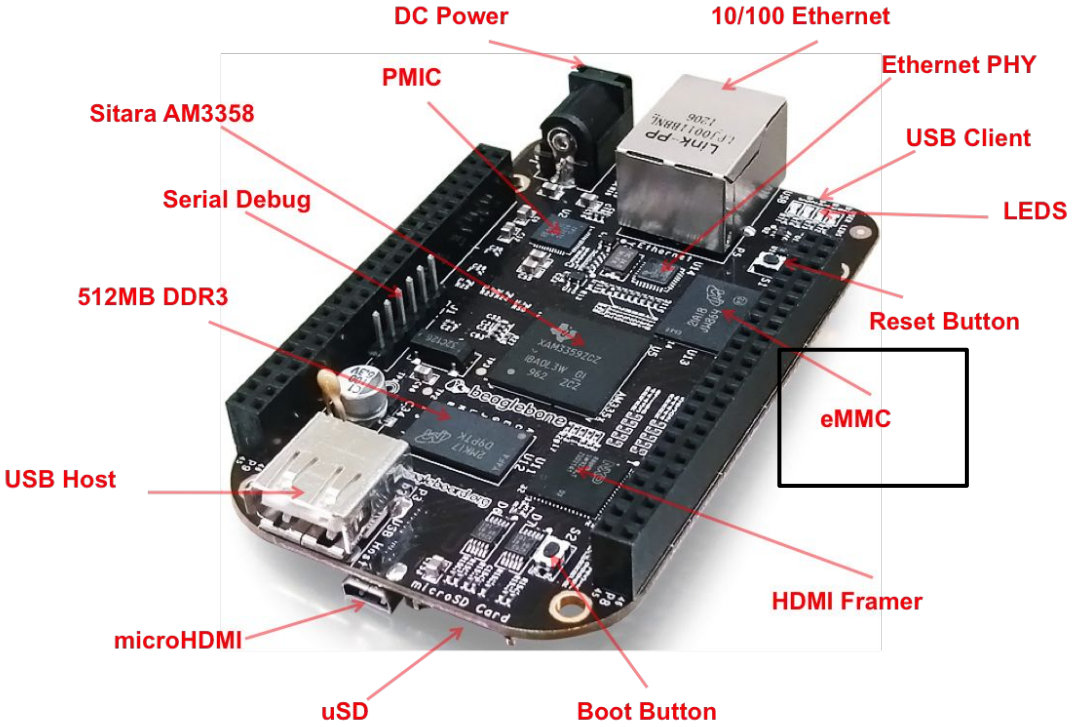
Hardware - Server Cables

Internal iPass (Mini-SAS) to HD (Mini-SAS)




SUPERMICRO

Hardware SBC



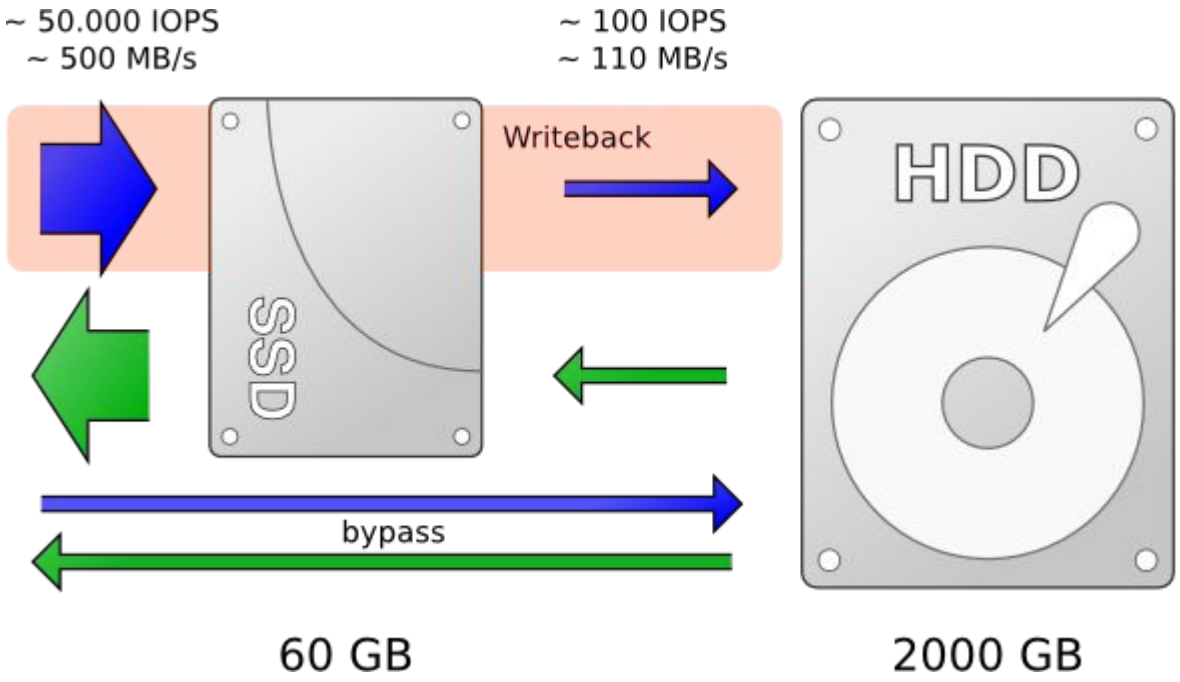
http://beagleboard.org/static/images/black_hardware_details.png

Hardware

Characteristic	HDD	SSD
Typical size	< 16TB	< 2TB
Random access time [a]	8ms	0.25ms
Sequential read	200 MB/s	450 MB/s
Random read	2 MB/s	450 MB/s
IOPS [b]	150 ops/s	100,000 ops/s
Cost	$0.03/GB	$0.26/GB
Reliability	Poor	Poor [c]
Limited writes	No	In theory

a. Performance and cost values are as of mid 2017
b. I/O operations per second
c. Fewer whole-device failures than HDD, but more data loss

SSD + HDD bcache



https://coelhorjc.wordpress.com/2015/03/25/how-to-cache-hdds-with-ssds-using-bcache/

Hardware Endurance

- Mean Time Between Failures (MTBF)
  - Widely Used in Industrial Engineering
    - Robotics
    - Aerospace
    - Electrical
  - Denominated in hours
    - HDD - WD Ultrastar® DC HC520 12TB, 2.5M hours MTBF
    - SSD - Samsung MZ-PZA960BW 2TB, 2M hours MTBF
      - MTBF less important for SSD because there are no mechanical components
  - Temperature and Power matters
    - HDD Designed for 86°F
      - Higher temperature decreases reliability
    - SSD for 0 to 133°F

Hardware Endurance

- SSD endurance and reliability better measured with TBW
- Terabytes Written (TBW)
  - Used for Flash Memory Storage
  - MTBF less relevant
  - Consumer drives - **~2,400 TBW** for a 4 TiB SSD
  - Enterprise Drives - **7,000 to 50,000 TBW** for a 4 TiB SSD
  - Optane Drives - Up to **584,000 TBW** for a 3.2 TiB SSD

Hardware Endurance

The product shall achieve an Annualized Failure Rate - AFR - of 0.73% (Mean Time Between Failures - MTBF - of 1.2 Million hrs) when operated in an environment that ensures the HDA case temperatures do not exceed 40°C. Operation at case temperatures outside the specifications in Section 2.9 may increase the product Annualized Failure Rate (decrease MTBF). AFR and MTBF are population statistics that are ***not relevant to individual units.***

AFR and MTBF specifications are based on the following assumptions for business critical storage system environments:

- ***8,760 power-on-hours per year. (365 days)***
- ***250 average motor start/stop cycles per year. (on/off)***
- ***Operations at nominal voltages. (Power supply and facility outages)***
- *Systems will provide adequate cooling to ensure the case* ***temperatures do not exceed 40°C****. Temperatures outside the specifications in Section 2.9 will increase the product AFR and* ***decrease MTBF.***

Seagate MTBF Reliability Guidance

Specifications

	SATA Models	SAS Models
Model No.	HUH721212ALE60y HUH721212ALN60y	HUH721212AL420y HUH721212AL520y
Configuration		
Interface	SATA 6Gb/s	SAS 12Gb/s
Capacity[1] (TB)	12TB	←
Format: Sector size[3] (bytes)	4Kn: 4096 512e: 512	4Kn: 4096, 4112, 4160, 4224 512e: 512, 520, 528
Max. Areal density (Gbits/sq. in.)	864	←
Performance		
Data buffer[4] (MB)	256	←
Rotational speed (RPM)	7200	←
Latency average (ms)	4.16	←
Interface transfer rate (MB/s, max)	600	1200
Sustained transfer rate[5]		
(MiB/s, typical)	243	←
(MB/s, typical)	255	←
Reliability		
Error rate (non-recoverable, bits read)	1 in 10^{15}	←
Load/Unload cycles (at 40°C)	600,000	←
Availability (hrs/day x days/wk)	24×7	←
MTBF[2] (M hours)	2.5	←
Annualized Failure Rate[2] (AFR)	0.35%	←
Warranty (yrs)	5	←

	SATA Models	SAS Models
Acoustics		
Idle (Bels, typical)	2.0/3.6	←
Power		
Requirement	+5 VDC, +12VDC	←
Operating[7]	6.9	10.1
Idle[8] (W)	5.0	6.1
Power consumption efficiency at Idle (W/TB)		
(Watts/TB)	0.42	0.51
(Watts/GB)	0.00042	0.00051
Physical size		
z-height (mm)	26.1	←
Dimensions (width x depth, mm)	101.6 (+/-0.25) x 147	←
Weight (g, max)	660	←
Environmental (Operating)		
Ambient temperature	5° to 60° C	←
Shock half-sine wave 2 ms, G)	70	←
Vibration (G RMS 5 to 500 Hz)	0.67 (XYZ)	←
Environmental (Non-Operating)		
Ambient temperature	-40° to 70° C	←
Shock (half-sine wave, G)	300 (2ms) / 150 (11ms)	←
Random vibration (G RMS 2 to 200 Hz)	1.04 (XYZ)	←

NOTE: See "How to read the Ultrastar model number" below for possible values for xx and y.

https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/product/data-center-drives/ultrastar-dc-hc500-series/data-sheet-ultrastar-dc-hc520.pdf

		MZ-PZA960BW	MZ-PZA480BW
Capacity[1]		960GB	480GB
Form Factor		Half-height Half-length (HHHL)	
Dimensions (WxDxH)		167.7 x 69.9 x 18.8 (mm)	
Weight		Max. 330g	
NAND type		Samsung Low Latency V-NAND	
Interface		PCI Express Gen3 x4, NVMe 1.2	
Performance[2]	Seq. Red (128KB)	up to 3,400 MB/s	
	Seq. Write (128KB)	up to 3,000 MB/s	
	Rand. Read (4KB, QD32)	up to 750,000 IOPS	
	Rand. Write (4KB, QD32)	up to 75,000 IOPS	up to 60,000 IOPS
	QoS Read (99.99%, 4KB, QD1)	up to 0.03 ms	
	QoS Write (99.99%, 4KB, QD1)	up to 0.03 ms	
Encryption Support		AES 256-bit Encryption Engine, TCG/Opal Compliant	
Average Power Consumption[3]		Active Read (Typ.) up to 8.5W, Active Write (Typ.) up to 9.0W, Idle up to 5.5W	
Allowable Voltage		12.0V ± 10%	
MTBF[4]		2,000,000 Hours	
UBER[5]		1 sector per 10^{17} bits read	
Operating Temperature		0-55°C	
Shock		1500G, duration 0.5 ms, Half Sine Wave	
Warranty		5-year limited warranty or 10 DWPD, whichever comes first	5-year limited warranty or 8.5 DWPD, whichever comes first

https://www.samsung.com/semiconductor/global.semi.static/Data_Center_SSD_983_ZET.Product_Brief.pdf

/sys

- Pseudo-Filesystem that presents kernel objects in a filesystem hierarchy
  - Hardware
- Provides a common interface to devices
  - Scan
  - Power
- Devices
  - SATA - Disks
  - PCIe - GPUs, NVMe, Ethernet
  - i2c - Fans, LEDs, Temperature Sensors, IoT
  - SPI - Similar to i2c
  - GPIO - PWM, LEDs, etc.
- `lspci` - list pci devices

/sys/block	This subdirectory contains one symbolic link for each block device that has been discovered on the system.
/sys/bus	This directory contains one subdirectory for each of the bus types in the kernel.
/sys/class	Device classes, terminals, network devices, block devices, graphics devices, sound devices
/sys/class/net	Symbolic links representing one of the real or virtual networking devices
/sys/dev	Block and character devices *major-ID*:*minor-ID*
/sys/devices	Kernel Device Tree
/sys/fs/cgroup	Mount point for cgroups
/sys/module	Loaded kernel modules
/sys/power	Not documented

/sys

```
root@cs447:/# tree -L 1 /sys
/sys
├── block
├── bus
├── class
├── dev
├── devices
├── firmware
├── fs
├── hypervisor
├── kernel
├── module
└── power

11 directories, 0 files
```

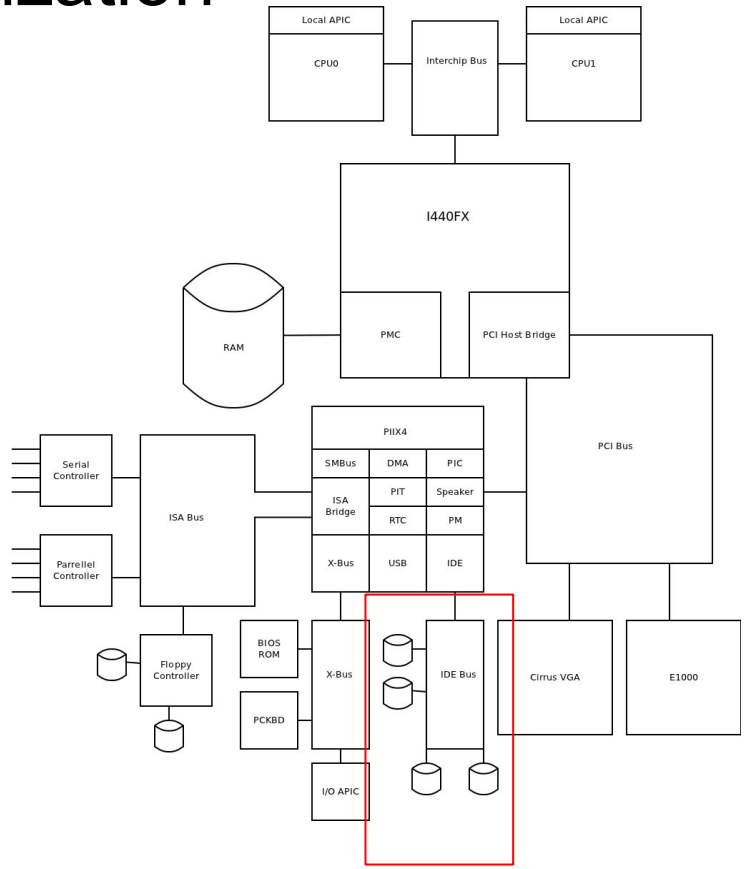
/sys/block

```
root@cs447:/# tree -L 1 /sys/block
/sys/block
├── fd0 -> ../devices/platform/floppy.0/block/fd0
├── loop0 -> ../devices/virtual/block/loop0
├── loop1 -> ../devices/virtual/block/loop1
├── loop2 -> ../devices/virtual/block/loop2
├── loop3 -> ../devices/virtual/block/loop3
├── loop4 -> ../devices/virtual/block/loop4
├── loop5 -> ../devices/virtual/block/loop5
├── loop6 -> ../devices/virtual/block/loop6
├── loop7 -> ../devices/virtual/block/loop7
├── sr0 -> ../devices/pci0000:00/0000:00:01.1/ata2/host1/target1:0:0/1:0:0:0/block/sr0
├── vda -> ../devices/pci0000:00/0000:00:04.0/virtio1/block/vda
└── vdb -> ../devices/pci0000:00/0000:00:05.0/virtio2/block/vdb

12 directories, 0 files
```

Virtual Drive - Why is attached to a PCI bus?

QEMU - Virtualization



https://wiki.qemu.org/Features/Q35

/sys/block - real hardware

```
newellz2@banyan:/$ tree -L 1 /sys/block
/sys/block
├── dm-0 -> ../devices/virtual/block/dm-0
├── dm-1 -> ../devices/virtual/block/dm-1
├── dm-2 -> ../devices/virtual/block/dm-2
├── dm-3 -> ../devices/virtual/block/dm-3
├── loop0 -> ../devices/virtual/block/loop0
├── loop1 -> ../devices/virtual/block/loop1
├── loop2 -> ../devices/virtual/block/loop2
├── loop3 -> ../devices/virtual/block/loop3
├── loop4 -> ../devices/virtual/block/loop4
├── loop5 -> ../devices/virtual/block/loop5
├── loop6 -> ../devices/virtual/block/loop6
├── loop7 -> ../devices/virtual/block/loop7
├── sda -> ../devices/pci0000:00/0000:00:11.4/ata1/host0/target0:0:0/0:0:0:0/block/sda
├── sdb -> ../devices/pci0000:00/0000:00:11.4/ata2/host1/target1:0:0/1:0:0:0/block/sdb
├── sdc -> ../devices/pci0000:00/0000:00:11.4/ata3/host2/target2:0:0/2:0:0:0/block/sdc
├── sdd -> ../devices/pci0000:00/0000:00:11.4/ata4/host3/target3:0:0/3:0:0:0/block/sdd
└── sr0 -> ../devices/pci0000:00/0000:00:1f.2/ata10/host9/target9:0:0/9:0:0:0/block/sr0

17 directories, 0 files
```

```
newellz2@banyan:~$ tree -L 1 /sys/block/sda
/sys/block/sda
├── alignment_offset
├── bdi -> ../../../../../../../../virtual/bdi/8:0
├── capability
├── dev
├── device -> ../../../0:0:0:0
├── discard_alignment
├── events
├── events_async
├── events_poll_msecs
├── ext_range
├── hidden
├── holders
├── inflight
├── integrity
├── power
├── queue
├── range
├── removable
├── ro
├── sda1
├── sda2
├── sda3
├── size
├── slaves
├── stat
├── subsystem -> ../../../../../../../../../class/block
├── trace
└── uevent

12 directories, 16 files
```

```
newellz2@banyan:~$ tree -L 1 /sys/block/sda/device
/sys/block/sda/device
├── blacklist
├── block
├── bsg
├── delete
├── device_blocked
├── device_busy
├── dh_state
├── driver -> ../../../../../../../bus/scsi/drivers/sd
├── eh_timeout
├── evt_capacity_change_reported
├── evt_inquiry_change_reported
├── evt_lun_change_reported
├── evt_media_change
├── evt_mode_parameter_change_reported
├── evt_soft_threshold_reached
├── generic -> scsi_generic/sg0
├── inquiry
├── iocounterbits
├── iodone_cnt
├── ioerr_cnt
├── iorequest_cnt
├── modalias
├── model
├── ncq_prio_enable
├── power
├── queue_depth
├── queue_ramp_up_period
├── queue_type
├── rescan
├── rev
├── scsi_device
├── scsi_disk
├── scsi_generic
├── scsi_level
├── state
├── subsystem -> ../../../../../../../bus/scsi
├── sw_activity
├── timeout
├── type
├── uevent
├── unload_heads
├── vendor
├── vpd_pg80
├── vpd_pg83
└── wwid

9 directories, 36 files
```

Removing a disk

```
echo 1 > /sys/block/sda/device/delete #Remove a disk
```

```
newellz2@banyan:~$ tree -L 1 /sys/devices/pci0000:00/0000:00:11.4/ata1/host0/scsi_host/host0
/sys/devices/pci0000:00/0000:00:11.4/ata1/host0/scsi_host/host0
├── active_mode
├── ahci_host_cap2
├── ahci_host_caps
├── ahci_host_version
├── ahci_port_cmd
├── can_queue
├── cmd_per_lun
├── device -> ../../../host0
├── eh_deadline
├── em_buffer
├── em_message
├── em_message_supported
├── em_message_type
├── host_busy
├── host_reset
├── link_power_management_policy
├── power
├── proc_name
├── prot_capabilities
├── prot_guard_type
├── scan
├── sg_prot_tablesize
├── sg_tablesize
├── state
├── subsystem -> ../../../../../../../class/scsi_host
├── supported_mode
├── uevent
├── unchecked_isa_dma
├── unique_id
└── use_blk_mq

3 directories, 27 files
```

/sys

- Scan a SATA port and find a disk.
  - echo “0 0 0” > /sys/devices/pci0000:00/0000:00:11.4/ata1/host0/scsi_host/host0
- Remove a USB Device
  - echo 1 > /sys/devices/pci0000:00/0000:00:1a.0/usb1/1-1/remove
- Remove PCI device
  - echo 1 > /sys/devices/pci0000:00/0000:00:1d.0/remove

```
newellz2@banyan:~$ lspci -v -nn -s 03:00.0
03:00.0 VGA compatible controller [0300]: NVIDIA Corporation GP104 [GeForce GTX 1080] [10de:1b80] (rev a1) (pro
g-if 00 [VGA controller])
        Subsystem: eVga.com. Corp. GP104 [GeForce GTX 1080] [3842:6180]
        Flags: bus master, fast devsel, latency 0, IRQ 84, NUMA node 0
        Memory at 91000000 (32-bit, non-prefetchable) [size=16M]
        Memory at 33fe0000000 (64-bit, prefetchable) [size=256M]
        Memory at 33ff0000000 (64-bit, prefetchable) [size=32M]
        I/O ports at 2000 [size=128]
        [virtual] Expansion ROM at 92080000 [disabled] [size=512K]
        Capabilities: <access denied>
        Kernel driver in use: nvidia
        Kernel modules: nvidiafb, nouveau, nvidia_drm, nvidia
```

```
newellz2@banyan:~$ tree -L 1 /sys/bus/pci/devices/0000:03:00.0
/sys/bus/pci/devices/0000:03:00.0
├── boot_vga
├── broken_parity_status
├── class
├── config
├── consistent_dma_mask_bits
├── current_link_speed
├── current_link_width
├── d3cold_allowed
├── device
├── dma_mask_bits
├── driver -> ../../../../bus/pci/drivers/nvidia
├── driver_override
├── drm
├── enable
├── i2c-10
├── i2c-11
├── i2c-5
├── i2c-6
├── i2c-7
├── i2c-8
├── i2c-9
├── irq
├── local_cpulist
├── local_cpus
├── max_link_speed
├── max_link_width
├── modalias
├── msi_bus
├── msi_irqs
├── numa_node
├── power
├── remove
├── rescan
├── resource
├── resource0
├── resource1
├── resource1_wc
├── resource3
├── resource3_wc
├── resource5
├── revision
├── rom
├── subsystem -> ../../../../bus/pci
├── subsystem_device
├── subsystem_vendor
├── uevent
└── vendor

12 directories, 35 files
```

```
newellz2@banyan:~$ cat /sys/bus/pci/devices/0000:03:00.0/vendor
0x10de
```

/sys

- Remove a device’s driver
  - `echo 1 > /sys/bus/pci/devices/0000:03:00.0/driver/unbind`
  - Useful for VFIO passthrough

Back to disks...

```
newellz2@banyan:~$ dmesg | grep sda
[    1.969202] sd 0:0:0:0: [sda] 1875385008 512-byte logical blocks: (960 GB/894 GiB)
[    1.969207] sd 0:0:0:0: [sda] Write Protect is off
[    1.969208] sd 0:0:0:0: [sda] Mode Sense: 00 3a 00 00
[    1.969215] sd 0:0:0:0: [sda] Write cache: enabled, read cache: enabled, doesn't support DPO or FUA
[    1.970573]  sda: sda1 sda2 sda3
[    1.970794] sd 0:0:0:0: [sda] Attached SCSI disk
[    4.010421] EXT4-fs (sda2): mounted filesystem with ordered data mode. Opts: errors=remount-ro
```

udev - creates /dev device files

```
root@banyan:~# udevadm info /dev/sda
P: /devices/pci0000:00/0000:00:11.4/ata1/host0/target0:0:0/0:0:0:0/block/sda
N: sda
S: disk/by-id/ata-SAMSUNG_MZ7KM960HAHP-0E005_S2NFNXAG901025B
S: disk/by-id/wwn-0x5002538c4006676a
S: disk/by-path/pci-0000:00:11.4-ata-1
E: DEVLINKS=/dev/disk/by-id/ata-SAMSUNG_MZ7KM960HAHP-0E005_S2NFNXAG901025B /dev/disk/by-path/pci-0000:0
E: DEVNAME=/dev/sda
E: DEVPATH=/devices/pci0000:00/0000:00:11.4/ata1/host0/target0:0:0/0:0:0:0/block/sda
E: DEVTYPE=disk
E: ID_ATA=1
E: ID_ATA_DOWNLOAD_MICROCODE=1
E: ID_ATA_FEATURE_SET_HPA=1
E: ID_ATA_FEATURE_SET_HPA_ENABLED=1
E: ID_ATA_FEATURE_SET_PM=1
E: ID_ATA_FEATURE_SET_PM_ENABLED=1
E: ID_ATA_FEATURE_SET_SECURITY=1
E: ID_ATA_FEATURE_SET_SECURITY_ENABLED=0
E: ID_ATA_FEATURE_SET_SECURITY_ENHANCED_ERASE_UNIT_MIN=32
E: ID_ATA_FEATURE_SET_SECURITY_ERASE_UNIT_MIN=32
E: ID_ATA_FEATURE_SET_SMART=1
E: ID_ATA_FEATURE_SET_SMART_ENABLED=1
E: ID_ATA_ROTATION_RATE_RPM=0
E: ID_ATA_SATA=1
E: ID_ATA_SATA_SIGNAL_RATE_GEN1=1
E: ID_ATA_SATA_SIGNAL_RATE_GEN2=1
E: ID_ATA_WRITE_CACHE=1
E: ID_ATA_WRITE_CACHE_ENABLED=1
E: ID_BUS=ata
E: ID_MODEL=SAMSUNG_MZ7KM960HAHP-0E005
E: ID_MODEL_ENC=SAMSUNG\x20MZ7KM960HAHP-0E005\x20\x20\x20\x20\x20\x20\x20\x20\x20\x20\x20\x20\x20\x20
E: ID_PART_TABLE_TYPE=gpt
E: ID_PART_TABLE_UUID=d3cef820-c225-4406-9abe-e722d4a036ea
E: ID_PATH=pci-0000:00:11.4-ata-1
E: ID_PATH_TAG=pci-0000_00_11_4-ata-1
E: ID_REVISION=GXM1003Q
E: ID_SERIAL=SAMSUNG_MZ7KM960HAHP-0E005_S2NFNXAG901025B
E: ID_SERIAL_SHORT=S2NFNXAG901025B
E: ID_TYPE=disk
E: ID_WWN=0x5002538c4006676a
E: ID_WWN_WITH_EXTENSION=0x5002538c4006676a
E: MAJOR=8
E: MINOR=0
E: SUBSYSTEM=block
E: TAGS=:systemd:
E: USEC_INITIALIZED=1980290
```

Disk Information

```
hdparm -i /dev/sda #Get disk information
```

```
root@banyan:~# hdparm -i /dev/sda

/dev/sda:

 Model=SAMSUNG MZ7KM960HAHP-0E005, FwRev=GXM1003Q, SerialNo=S2NFNXAG901025B
 Config={ Fixed }
 RawCHS=16383/16/63, TrkSize=0, SectSize=0, ECCbytes=0
 BuffType=unknown, BuffSize=unknown, MaxMultSect=16, MultSect=off
 CurCHS=16383/16/63, CurSects=16514064, LBA=yes, LBAsects=1875385008
 IORDY=on/off, tPIO={min:120,w/IORDY:120}, tDMA={min:120,rec:120}
 PIO modes:  pio0 pio1 pio2 pio3 pio4
 DMA modes:  mdma0 mdma1 mdma2
 UDMA modes: udma0 udma1 udma2 udma3 udma4 udma5 *udma6
 AdvancedPM=no WriteCache=enabled
 Drive conforms to: unknown:  ATA/ATAPI-2,3,4,5,6,7

 * signifies the current active mode
```

NVME

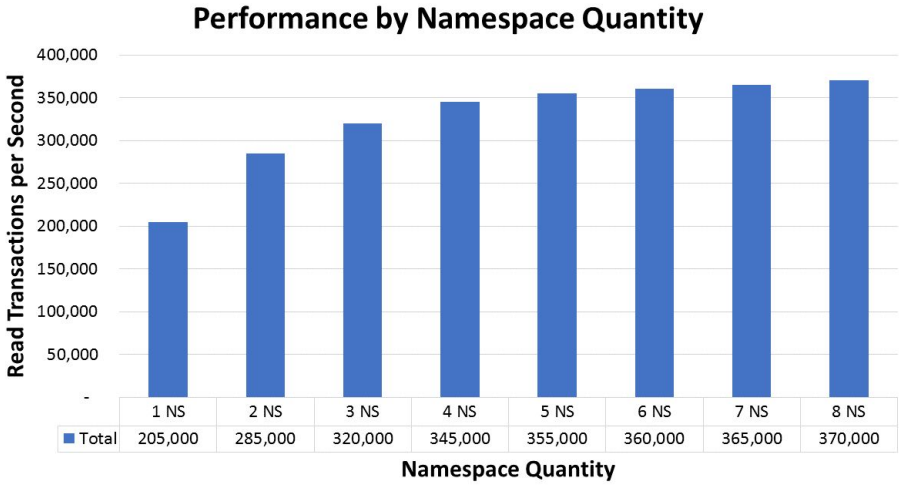
hdparm -i /dev/sda #Get disk information

```
root@ncr-0:~# nvme list
Node             SN                   Model                                    Namespace Usage                      Form
at           FW Rev
---------------- -------------------- ---------------------------------------- --------- -------------------------- ----
------------ --------
/dev/nvme0n1     201527CCA9E7         Micron_9300_MTFDHAL3T2TDR                1           3.20  TB /   3.20  TB    512
  B +  0 B   11300DG0
/dev/nvme1n1     201727D05E86         Micron_9300_MTFDHAL3T2TDR                1           3.20  TB /   3.20  TB    512
  B +  0 B   11300DG0
/dev/nvme2n1     201527CCAA5B         Micron_9300_MTFDHAL3T2TDR                1           3.20  TB /   3.20  TB    512
  B +  0 B   11300DG0
/dev/nvme3n1     201727D05E58         Micron_9300_MTFDHAL3T2TDR                1           3.20  TB /   3.20  TB    512
  B +  0 B   11300DG0
```

https://www.micron.com/about/blog/2019/june/using-namespaces-on-the-micron-9300-nvme-ssd-to-improve-application-performance

NVMe Namespaces

- Present a single physical NVMe device as multiple logical NVMe devices.
  - Similar to partitioning
- Used for Virtualization
- Performance
  - More threads
  - Legacy Software



Disk Information - `lsblk`

```
root@banyan:/var/tmp# lsblk
NAME                  MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
loop0                   7:0    0    10G  0 loop
├─loop0p1             259:0    0   953M  0 loop
└─loop0p2             259:1    0   952M  0 loop
sda                     8:0    0 894.3G  0 disk
├─sda1                  8:1    0     1M  0 part
├─sda2                  8:2    0     1G  0 part /boot
└─sda3                  8:3    0 893.3G  0 part
  ├─ubuntu--vg-root 253:0    0   400G  0 lvm  /
  ├─ubuntu--vg-home 253:1    0     2T  0 lvm  /home
  └─ubuntu--vg-tmp  253:3    0   500G  0 lvm
sdb                     8:16   0 894.3G  0 disk
└─sdb1                  8:17   0 894.3G  0 part
  └─ubuntu--vg-home 253:1    0     2T  0 lvm  /home
sdc                     8:32   0 894.3G  0 disk
└─sdc1                  8:33   0 894.3G  0 part
  ├─ubuntu--vg-opt  253:2    0   500G  0 lvm  /opt
  └─ubuntu--vg-tmp  253:3    0   500G  0 lvm
sdd                     8:48   0 894.3G  0 disk
└─sdd1                  8:49   0 894.3G  0 part
  └─ubuntu--vg-home 253:1    0     2T  0 lvm  /home
sr0                    11:0    1  1024M  0 rom
```

lsblk -o +MODEL,SERIAL

```
(base) [newellz2sa@ph-head-0 ~]$ lsblk -o +MODEL,SERIAL
NAME      MAJ:MIN RM   SIZE RO TYPE  MOUNTPOINT MODEL            SERIAL
sda         8:0    0 372.6G  0 disk             INTEL SSDSC2BX40 BTHC71840B92400VGN
├─sda1      8:1    0   512M  0 part
│ └─md127   9:127  0   512M  0 raid1 /boot
├─sda2      8:2    0    16G  0 part
│ └─md125   9:125  0    16G  0 raid1 [SWAP]
└─sda3      8:3    0 356.1G  0 part
  └─md126   9:126  0   356G  0 raid1 /
sdb         8:16   0 372.6G  0 disk             INTEL SSDSC2BX40 BTHC714101YG400VGN
├─sdb1      8:17   0   512M  0 part
│ └─md127   9:127  0   512M  0 raid1 /boot
├─sdb2      8:18   0    16G  0 part
│ └─md125   9:125  0    16G  0 raid1 [SWAP]
└─sdb3      8:19   0 356.1G  0 part
  └─md126   9:126  0   356G  0 raid1 /
sdc         8:32   0   1.5T  0 disk             SSDSC2BB016T7R   PHDV716200ZK1P6EGN
└─md3       9:3    0   8.7T  0 raid6 /apps
sdd         8:48   0   1.5T  0 disk             SSDSC2BB016T7R   PHDV716200XE1P6EGN
└─md3       9:3    0   8.7T  0 raid6 /apps
sde         8:64   0   1.5T  0 disk             SSDSC2BB016T7R   PHDV716200Y91P6EGN
└─md3       9:3    0   8.7T  0 raid6 /apps
sdf         8:80   0   1.5T  0 disk             SSDSC2BB016T7R   PHDV716202FU1P6EGN
```

Disk Information

- Self-Monitoring, Analysis and Reporting Technology (SMART)
  - Built into most ATA/SATA and SCSI/SAS hard drives
  - Monitor the reliability
  - Predict drive failures
  - Run self-tests

```
smartctl -x /dev/sda
```

```
root@banyan:~# smartctl -x /dev/sda
smartctl 6.6 2016-05-31 r4324 [x86_64-linux-4.15.0-65-generic] (local build)
Copyright (C) 2002-16, Bruce Allen, Christian Franke, www.smartmontools.org

=== START OF INFORMATION SECTION ===
Model Family:     Samsung based SSDs
Device Model:     SAMSUNG MZ7KM960HAHP-0E005
Serial Number:    S2NFNXAG901025B
LU WWN Device Id: 5 002538 c4006676a
Firmware Version: GXM1003Q
User Capacity:    960,197,124,096 bytes [960 GB]
Sector Size:      512 bytes logical/physical
Rotation Rate:    Solid State Device
Device is:        In smartctl database [for details use: -P show]
ATA Version is:   ACS-2, ATA8-ACS T13/1699-D revision 4c
SATA Version is:  SATA 3.1, 6.0 Gb/s (current: 6.0 Gb/s)
Local Time is:    Tue Dec 31 18:44:44 2019 PST
SMART support is: Available - device has SMART capability.
SMART support is: Enabled
AAM feature is:   Unavailable
APM feature is:   Unavailable
Rd look-ahead is: Enabled
Write cache is:   Enabled
ATA Security is:  Disabled, NOT FROZEN [SEC1]
Wt Cache Reorder: Enabled
```

```
SMART Attributes Data Structure revision number: 1
Vendor Specific SMART Attributes with Thresholds:
ID# ATTRIBUTE_NAME          FLAGS    VALUE WORST THRESH FAIL RAW_VALUE
  5 Reallocated_Sector_Ct   PO--CK   100   100   010    -    0
  9 Power_On_Hours          -O--CK   093   093   000    -    31185
 12 Power_Cycle_Count       -O--CK   099   099   000    -    29
177 Wear_Leveling_Count     PO--C-   099   099   005    -    51
179 Used_Rsvd_Blk_Cnt_Tot   PO--C-   100   100   010    -    0
180 Unused_Rsvd_Blk_Cnt_Tot PO--C-   100   100   010    -    7721
181 Program_Fail_Cnt_Total  -O--CK   100   100   010    -    0
182 Erase_Fail_Count_Total  -O--CK   100   100   010    -    0
183 Runtime_Bad_Block       PO--C-   100   100   010    -    0
184 End-to-End_Error        PO--CK   100   100   097    -    0
187 Uncorrectable_Error_Cnt -O--CK   100   100   000    -    0
190 Airflow_Temperature_Cel -O--CK   073   049   000    -    27
195 ECC_Error_Rate          -O-RC-   200   200   000    -    0
197 Current_Pending_Sector  -O--CK   100   100   000    -    0
199 CRC_Error_Count         -OSRCK   100   100   000    -    0
202 Exception_Mode_Status   PO--CK   100   100   010    -    0
235 POR_Recovery_Count      -O--C-   099   099   000    -    15
241 Total_LBAs_Written      -O--CK   099   099   000    -    30934598597
242 Total_LBAs_Read         -O--CK   099   099   000    -    56577552073
243 SATA_Downshift_Ct       -O--CK   100   100   000    -    0
244 Thermal_Throttle_St     -O--CK   100   100   000    -    0
245 Timed_Workld_Media_Wear -O--CK   100   100   000    -    65535
246 Timed_Workld_RdWr_Ratio -O--CK   100   100   000    -    65535
247 Timed_Workld_Timer      -O--CK   100   100   000    -    65535
251 NAND_Writes             -O--CK   100   100   000    -    43619670056
                            |||||_ K auto-keep
                            ||||__ C event count
                            |||___ R error rate
                            ||____ S speed/performance
                            |_____ O updated online
                            ______ P prefailure warning
```

How do we use a disk?

- Check and note the serial number
  - Why? /dev/disk/by-id
- Insert
- Partition (most of the time)
  - `parted` - GPT and MBR
  - `fdisk` - Master Boot Record
  - `gdisk` - Like fdisk for GPT
- Create Filesystem
  - `mkfs -t ext4 -L myfs /dev/sda1`
- Mount
  - `mount /dev/sda1 /mnt`
- Create\Read\Update\Delete (CRUD) Files

Partitioning

- Disks are broken into segments called partitions

```
+-------------------------------------------------------------+
|            storage device with no partitions                |
+-------------------------------------------------------------+
0 start                                                     end

+--+---------------+----------------+-----------------------+
|PT|  Partition 1  |  Partition 2   |  Partition 3          |
+--+---------------+----------------+-----------------------+
0 start                                                     end
```

https://www.gnu.org/software/parted/manual/parted.html

https://systemd.io/DISCOVERABLE_PARTITIONS/

Common partition types

Partition type	Mountpoint	gdisk's code	Partition type GUID
Linux filesystem	Any	8300	0FC63DAF-8483-4772-8E79-3D69D8477DE4
EFI system partition	Any[1]	ef00	C12A7328-F81F-11D2-BA4B-00A0C93EC93B
BIOS boot partition	None	ef02	21686148-6449-6E6F-744E-656564454649
Linux x86-64 root (/)	/	8304	4F68BCE3-E8CD-4DB1-96E7-FBCAF984B709
Linux swap	[SWAP]	8200	0657FD6D-A4AB-43C4-84E5-0933C84B4F4F
Linux /home	/home	8302	933AC7E1-2EB4-4F13-B844-0E14E2AEF915
Linux /srv	/srv	8306	3B8F8425-20E0-4F3B-907F-1A25A76F98E8
Linux /var	/var [1]	8310	4D21B016-B534-45C2-A9FB-5C16E091FD2D
Linux /var/tmp	/var/tmp [1]	8311	7EC6F557-3BC5-4ACA-B293-16EF5DF639D1
Linux LVM	Any	8e00	E6D6D379-F507-44C2-A23C-238F2A3DF928
Linux RAID	Any	fd00	A19D880F-05FC-4D3B-A006-743F0F84911E
Linux LUKS	Any	8309	CA7D7CCB-63ED-4C53-861C-1742536059CC
Linux dm-crypt	Any	8308	7FFEC5C9-2D00-49B7-8941-3EA10A5586B7

parted - creating and manipulating partition tables.

```
truncate -s 1G /var/tmp/disk.img  #Create a sparse file
losetup --find --show disk.img    #File -> Block Device
parted -s /dev/loop0 'print'      #Blank disk
```

```
Model: Loopback device (loopback)
Disk /dev/loop0: 1GB
Sector size (logical/physical): 512B/512B
Partition Table: unknown
Disk Flags:
```

Partitioning - Loop Device

- Kernel module `loop`
- Block device that maps its data blocks to a file
- Useful for a partitioned disk image stored in a file
- Backups!

```
root@banyan:~# grep LOOP /boot/config-4.15.0-72-generic
CONFIG_BLK_DEV_LOOP=y
CONFIG_BLK_DEV_LOOP_MIN_COUNT=8
CONFIG_BLK_DEV_CRYPTOLOOP=m
CONFIG_NVME_TARGET_LOOP=m
# CONFIG_NVME_TARGET_FCLOOP is not set
CONFIG_LOOPBACK_TARGET=m
# CONFIG_NET_DSA_LOOP is not set
CONFIG_SPI_LOOPBACK_TEST=m
CONFIG_RC_LOOPBACK=m
CONFIG_SND_ALOOP=m
CONFIG_GREYBUS_LOOPBACK=m
CONFIG_IIO_TIGHTLOOP_TRIGGER=m
CONFIG_AUFS_BDEV_LOOP=y
```

```
.config - Linux/x86 4.15.18 Kernel Configuration
> Device Drivers > Block devices
                              Block devices
  Arrow keys navigate the menu.  <Enter> selects submenus ---> (or empty
  submenus ----).  Highlighted letters are hotkeys.  Pressing <Y> includes, <N>
  excludes, <M> modularizes features.  Press <Esc><Esc> to exit, <?> for Help,
  </> for Search.  Legend: [*] built-in  [ ] excluded  <M> module  < > module
      ^(-)
      <M>     DataStor EP-2000 protocol
      <M>     FIT TD-2000 protocol
      <M>     FIT TD-3000 protocol
      <M>     Shuttle EPAT/EPEZ protocol
      [*]       Support c7/c8 chips
      <M>     Shuttle EPIA protocol
      <M>     Freecom IQ ASIC-2 protocol
      <M>     FreeCom power protocol
      <M>     KingByte KBIC-951A/971A protocols
      <M>     KT PHd protocol
      <M>     OnSpec 90c20 protocol
      <M>     OnSpec 90c26 protocol
      <M>   Block Device Driver for Micron PCIe SSDs
      <M>   Compressed RAM block device support
      [*]     Write back incompressible page to backing device
      <M>   Mylex DAC960/DAC1100 PCI RAID Controller support
      <M>   Micro Memory MM5415 Battery Backed RAM support
      <*>   Loopback device support
      (8)     Number of loop devices to pre-create at init time
      <M>     Cryptoloop Support
      v(+)

          <Select>    < Exit >    < Help >    < Save >    < Load >
```

MBR

- Master Boot Record
- Originally from Microsoft's DOS
- Can only be used on disks < 2TiB
- Maximum of 4 Partitions
  - Workaround was to reserve one for ‘logical partitions’
  - Windows must boot from a Primary Partition

parted

```
parted -s /dev/loop0 'help mklabel' #Creates a partition table

parted -s /dev/loop0 'mklabel msdos' #Master Boot Record Partition

parted -s /dev/loop0 'print'

parted -s /dev/loop0 'mkpart primary 1 ext4 1M 200M' #Create Part

parted -s /dev/loop0 'unit G print'

parted -s /dev/loop0 'unit GiB print'

parted -s /dev/loop0 'unit MB print'
```

```
+--+---------------+-----------------------------------------+
|PT|  Partition 1  |                                         |
+--+---------------+-----------------------------------------+
0   1MB start        1000MB end
```

parted -s /dev/loop0 'mkpart primary fat32 200M 400M'

```
+--+---------------+----------------+-------------------------+
|PT|  Partition 1  | Partition 2    |                         |
+--+---------------+----------------+-------------------------+
0  1MB            200MB             400MB
```

```
parted -s /dev/loop0 'unit MB print'
Model: Loopback device (loopback)
Disk /dev/loop0: 10737MB
Sector size (logical/physical): 512B/512B
Partition Table: msdos
Disk Flags:

Number  Start   End     Size   Type     File system  Flags
 1      1.05MB  200MB  200MB  primary
 2      200MB   400MB  200MB  primary               lba
```

GPT - GUID Partition Table

- Described by Extensible Firmware Interface (EFI)
- Overcomes MBR shortcomings
- 64bit disk sector pointers, Max Partition size of 8 ZiB (zebibytes)
  - MBR uses 32bit
  - 8 ZiB is roughly 9.4 billion terabytes / 9.4 trillion gigabytes
- Supports up to 128 partitions
  - No primary, extended or logical partition types
- Partition Name
  - MBR does not support a Partition Name
- Partition Type
  - Autodiscovery

parted

```
parted -s /dev/loop0 'mklabel gpt # GPT Partition

parted -s /dev/loop0 'print'

parted -s /dev/loop0 'mkpart PART1 ext4 1M 200M' #Create Part

parted -a optimal -s /dev/loop0 'mkpart PART2 LVM 200M 400M'

#Create Part with optimal alignment
```

fstab - static information about the filesystems

Each filesystem is described on a separate line. Fields on each line are separated by tabs or spaces. Lines starting with '#' are comments. Blank lines are ignored. /etc/fstab and /etc/mtab

Fields

1. fs_spec - block special device or remote filesystem to be mounted.
2. fs_file - describes the mount point (target) for the filesystem. Can be “none”.
3. fs_vfstype - type of the filesystem.
    a. ext4, xfs,btrfs, f2fs, vfat, ntfs, hfsplus, tmpfs, sysfs, proc, iso9660, udf, squashfs, nfs, cifs, and many more.
4. fs_mntops - mount options associated with the filesystem.
    a. defaults, noauto, user, owner, comment
5. fs_freq - used by dump(8) to determine which filesystems need to be dumped.
6. fs_passno - order in which filesystem checks are done at boot time. / should be 1.

getmntent(3) or libmount