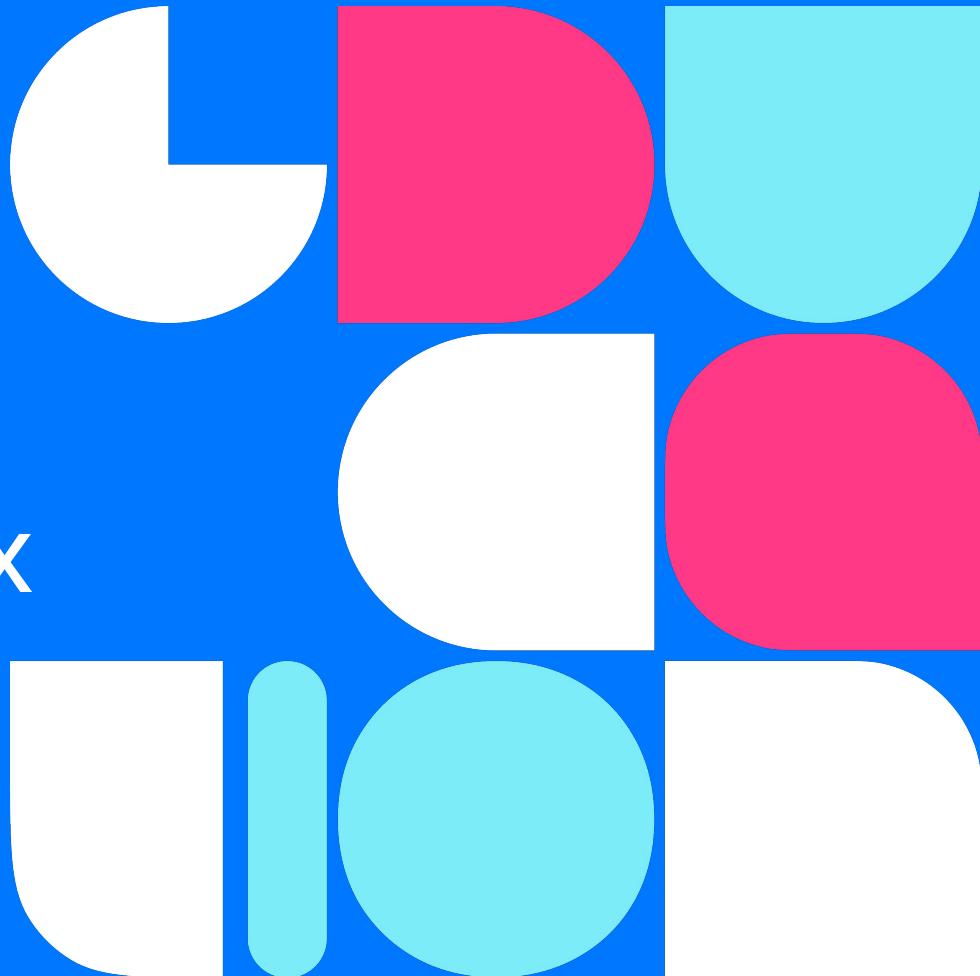




Управление ресурсами linux

Эксплуатация высоконагруженных систем



Александр Фатеев

Системный администратор



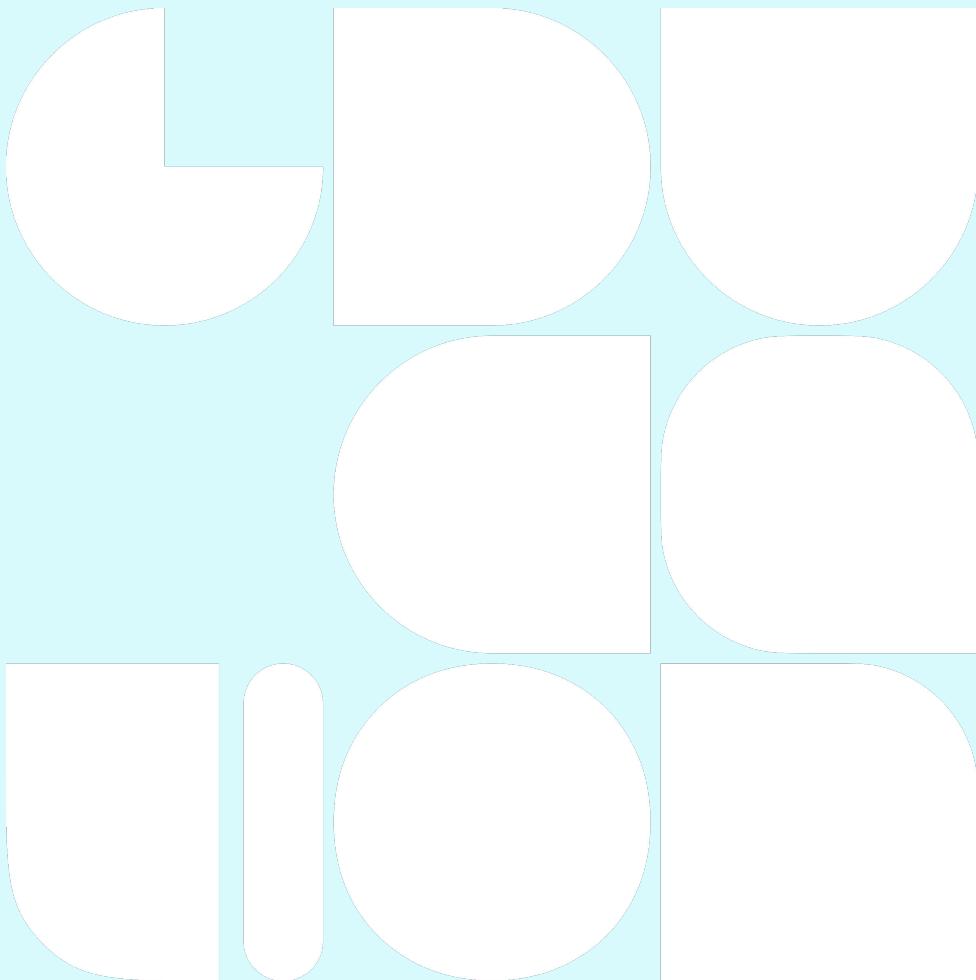
Цель занятия:

На занятии вы научитесь управлять ресурсами Linux, такими как CPU, память, хранилища данных.

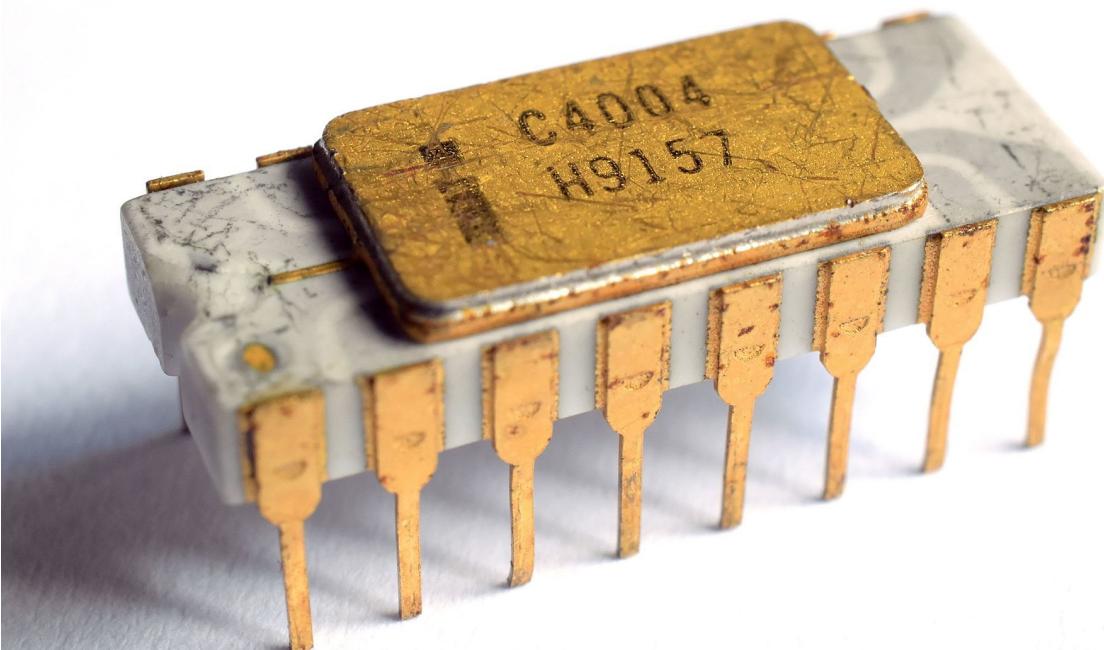
О чём поговорим?

- CPU
- Память
- Дисковая подсистема
- Cgroups

CPU

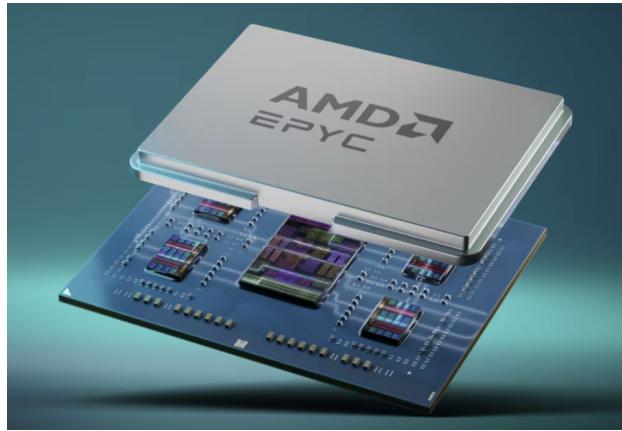
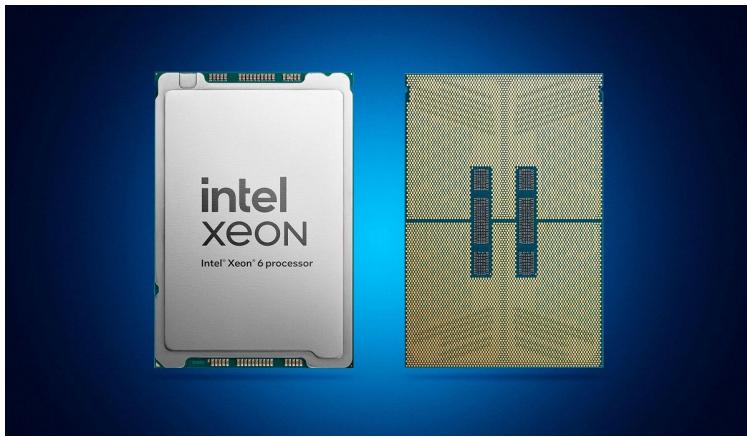


CPU



- частота до 740 КГц
- техпроцесс 10 мкм

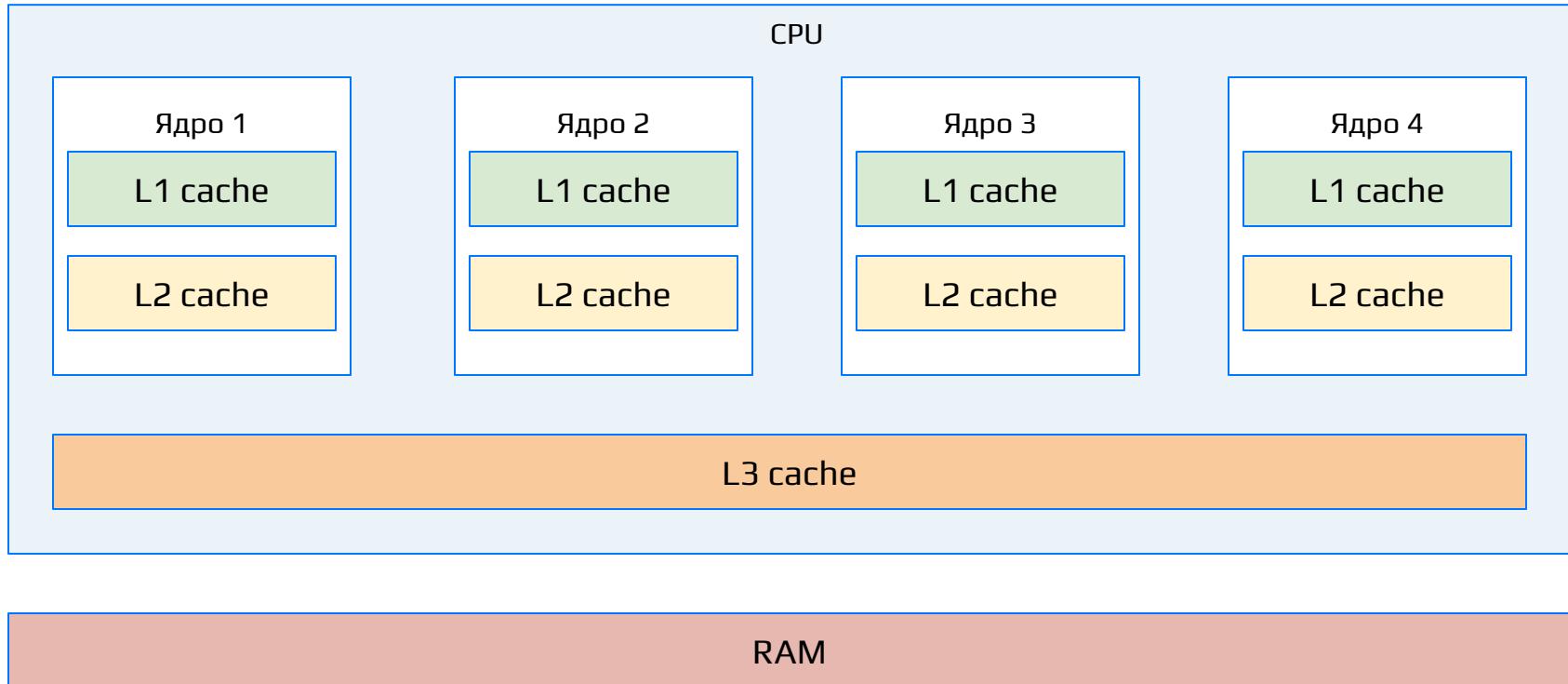
CPU



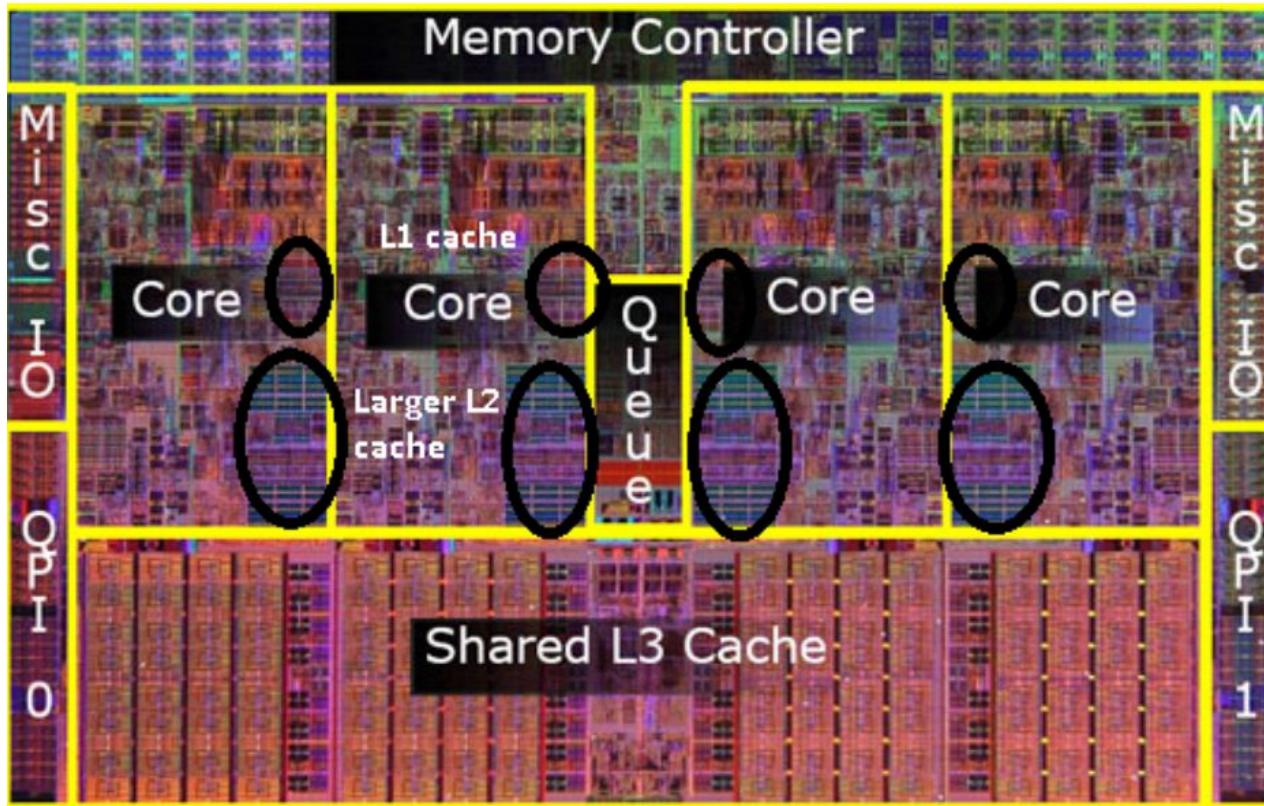
- до 128 ядер
- L3 кэш до 504 мб
- TDP до 500 Вт
- частота 2 ГГц (до 3,9)
- техпроцесс 3 нм

- до 192 ядер
- L3 кэш ??? (до 1152 мб в 4 поколении)
- TDP ??? (до 400 Вт в 4 поколении)
- частота ??? (от 2,25 до 4,1 ГГц в 4 поколении)
- техпроцесс 3 нм

CPU: кэш



CPU: кэш



CPU: кэш

```
PROCESSOR PHYSICAL: 2
PROCESSOR LOGICAL: 128
PROCESSOR MODEL: Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz
                  Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz

MEMORY TOTAL:      527978092 kB
```

```
[root@srvh999 ~]# lscpu | grep cache
```

```
L1d cache:          3 MiB (64 instances)
L1i cache:          2 MiB (64 instances)
L2 cache:           80 MiB (64 instances)
L3 cache:           96 MiB (2 instances)
```

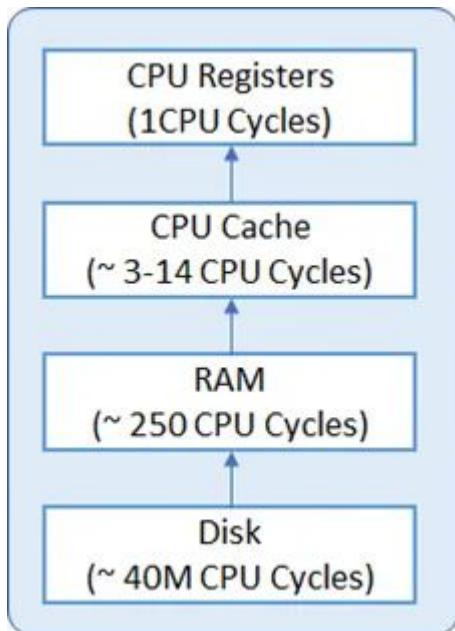
CPU: кэш

```
PROCESSOR PHYSICAL: 2
PROCESSOR LOGICAL: 256
PROCESSOR MODEL:      AMD EPYC 9534 64-Core Processor
                      AMD EPYC 9534 64-Core Processor
MEMORY TOTAL:        792305188 kB
```

```
[root@srvm2080 ~]# lscpu | grep cache
```

```
L1d cache:          4 MiB (128 instances)
L1i cache:          4 MiB (128 instances)
L2 cache:           128 MiB (128 instances)
L3 cache:           512 MiB (16 instances)
```

CPU: кэш



CPU: кэш

L1 cache reference	0.5	ns			
L2 cache reference	7	ns			14x L1 cache
Main memory reference	100	ns			20x L2 cache, 200x L1 cache
Read 4K randomly from SSD*	150,000	ns	150	us	~1GB/sec SSD
Read 1 MB sequentially from memory	250,000	ns	250	us	
Read 1 MB sequentially from SSD*	1,000,000	ns	1,000	us	1 ms ~1GB/sec SSD, 4X memory
Disk seek	10,000,000	ns	10,000	us	10 ms
Read 1 MB sequentially from disk	20,000,000	ns	20,000	us	20 ms 80x memory, 20X SSD

CPU

```
# dmidecode -type=4
```

Handle 0x0061, DMI type 4, 48 bytes

Processor Information

Socket Designation: CPU1

Type: Central Processor

Family: Xeon

Manufacturer: Intel(R) Corporation

ID: A6 06 06 00 FF FB EB BF

Signature: Type 0, Family 6, Model 106, Stepping 6

Flags:

FPU (Floating-point unit on-chip)

...

Version: Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz

Voltage: 1.6 V

External Clock: 100 MHz

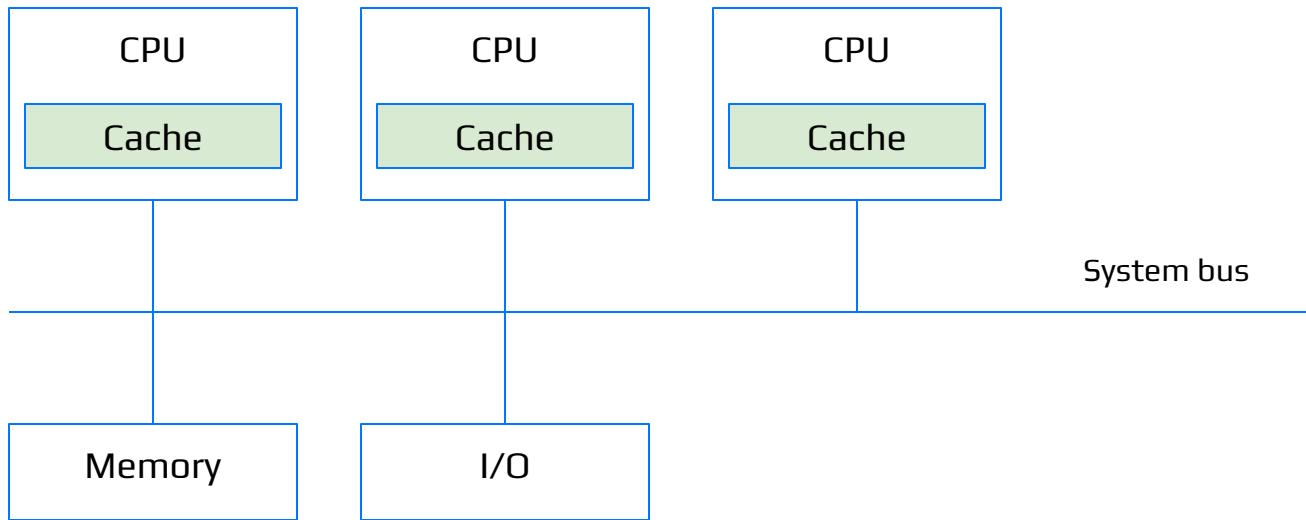
Max Speed: 4000 MHz

Current Speed: 2000 MHz

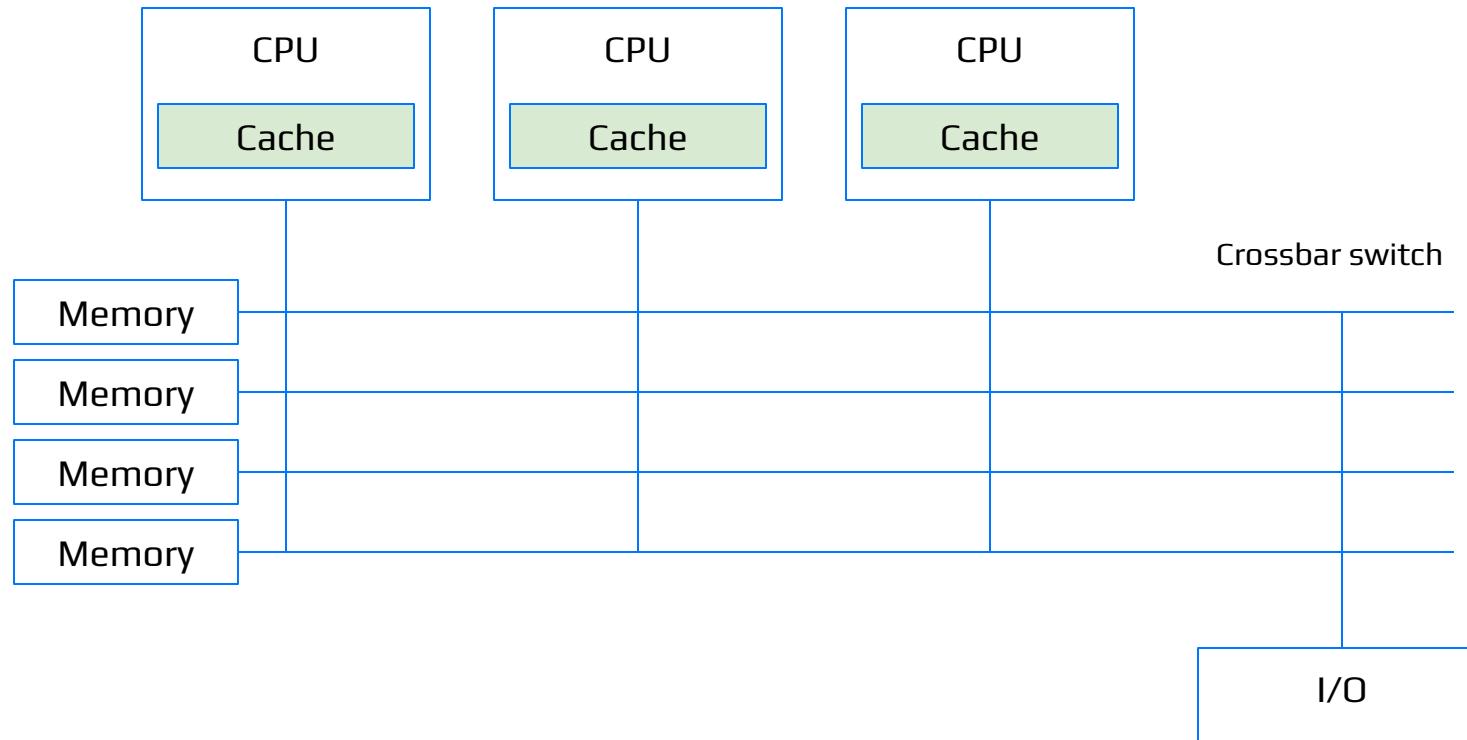
Status: Populated, Enabled

Upgrade: Socket LGA4189

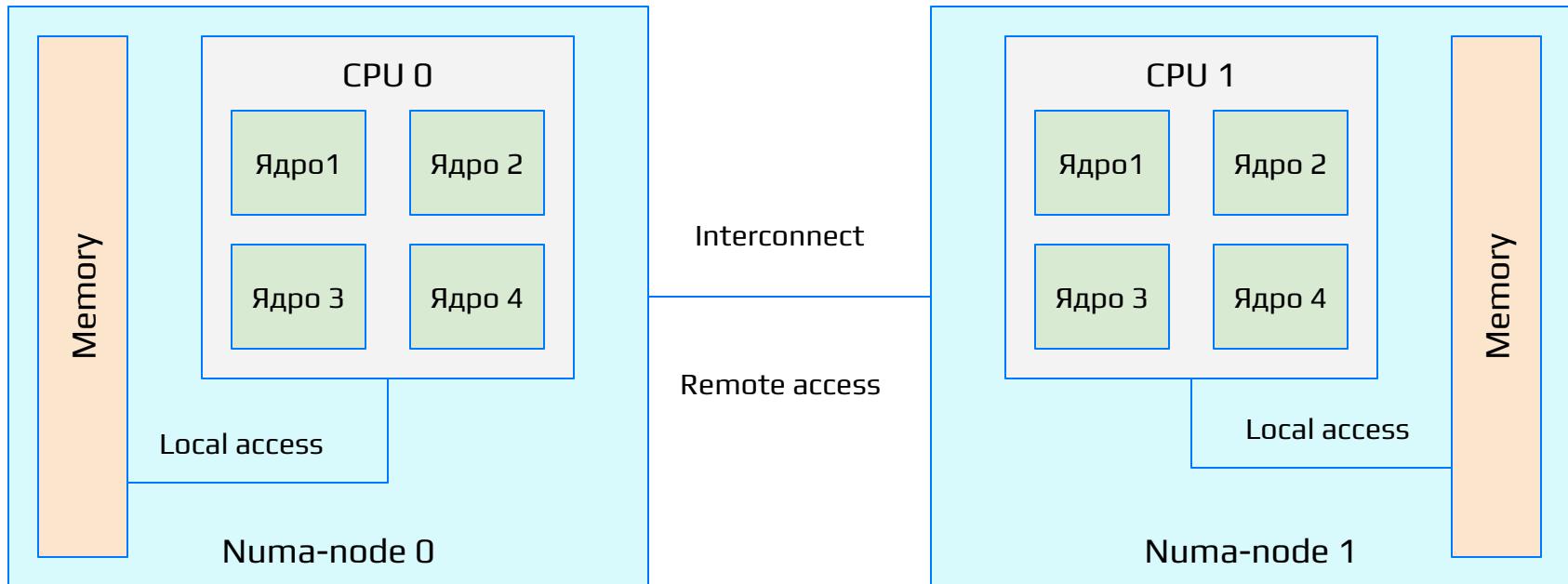
CPU: SMP



CPU: SMP



CPU: NUMA



CPU: NUMA

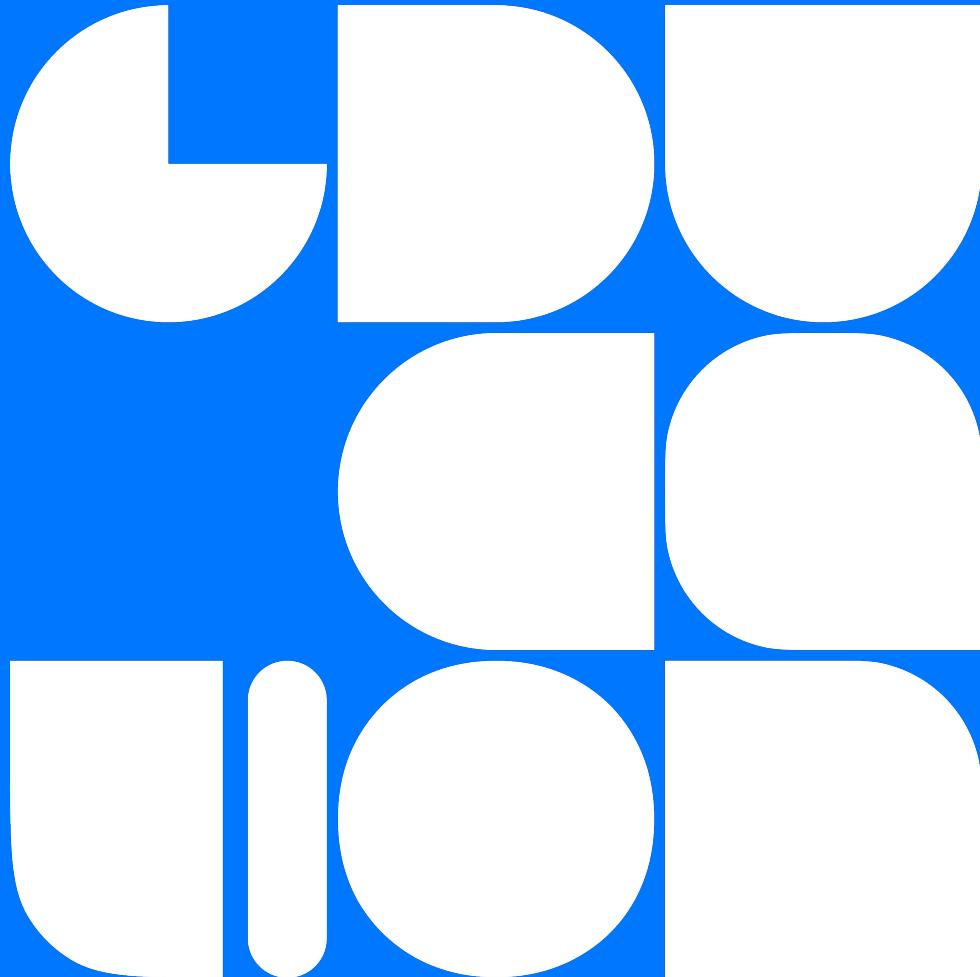
15.11.2021 - numa off

16.11.2021 - numa on и контейнеры по локальным нодам.

Данные по 1.ejb.web-group8.dc

param	15.11.2021	16.11.2021
Average duration of EJB event	100.00%	-3%
Daily_RemoteOperationsHosts	100.00%	+1%
Perf Counters (cycles)	100.00%	-18%
Perf Counters (cache-misses)	100.00%	-16%
CPU load percent from quota	100.00%	-20%
vCPU peak millicores (only total)	100.00%	-10%

Практика



CPU: NUMA

```
# cat /proc/cpuinfo  
# lscpu  
# numactl --hardware  
# numactl --cpunodebind=0 <program>  
# numactl --physcpubind=2 <program>  
# cat /proc/<PID>/status  
# cat /proc/<PID>/numa_maps
```

CPU: NUMA

```
# numactl --interleave=all <programm>
```

CPU: NUMA

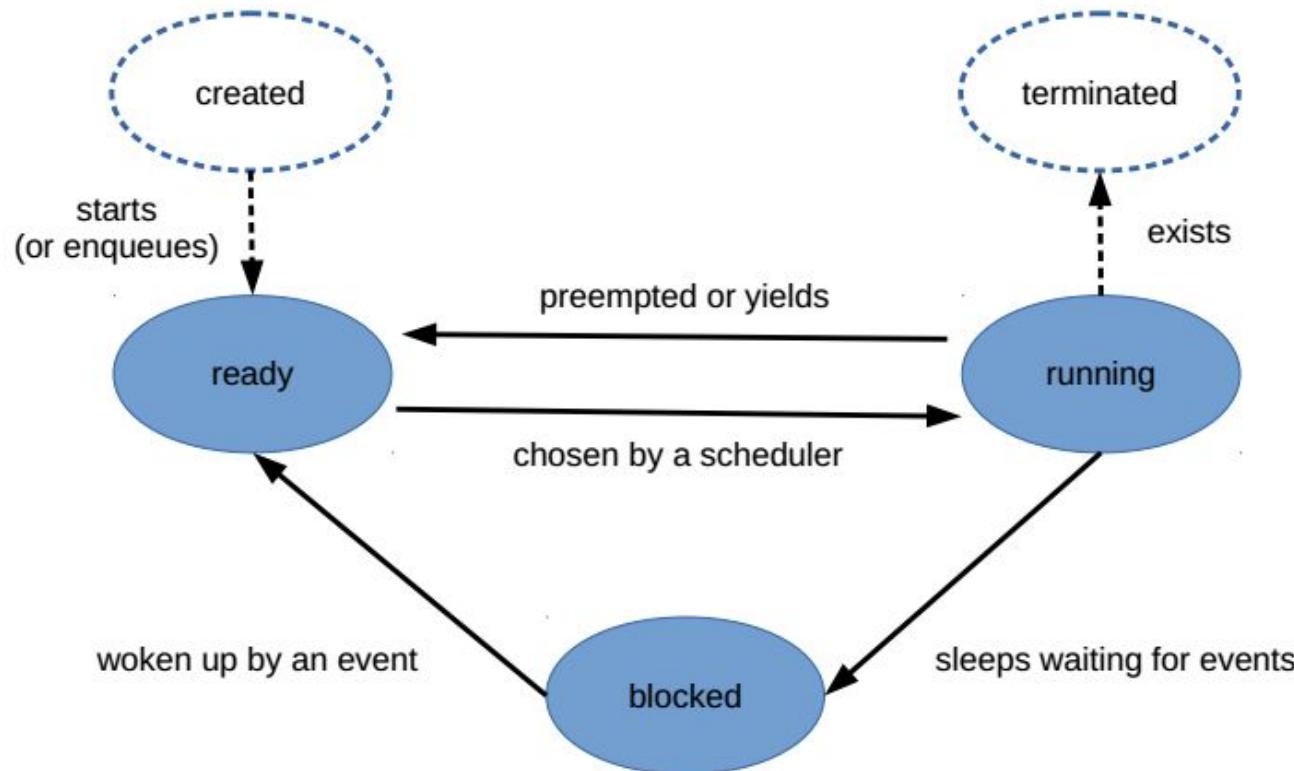
```
# numactl --interleave=all <programm>
```

```
# numactl --interleave=0,1 <program>
```

CPU: NUMA

```
/etc/default/grub:  
+ numa=off b GRUB_CMDLINE_LINUX  
  
# grub2-mkconfig -o /boot/grub2/grub.cfg  
# reboot
```

CPU: linux kernel scheduler



CPU: linux kernel scheduler

Normal policies (Completely Fair Scheduler):

- SCHED_OTHER
- SCHED_BATCH
- SCHED_IDLE

CPU: linux kernel scheduler

Normal policies (Completely Fair Scheduler):

- SCHED_OTHER
- SCHED_BATCH
- SCHED_IDLE

Realtime policies:

- SCHED_FIFO
- SCHED_RR

CPU: linux kernel scheduler

Normal policies (Completely Fair Scheduler):

- SCHED_OTHER
- SCHED_BATCH
- SCHED_IDLE

Realtime policies:

- SCHED_FIFO
- SCHED_RR

Deadline scheduler:

RUNTIME <= DEADLINE <= PERIOD

CPU: linux kernel scheduler

Normal policies (Completely Fair Scheduler):

- SCHED_OTHER
- SCHED_BATCH
- SCHED_IDLE

Realtime policies:

- SCHED_FIFO
- SCHED_RR

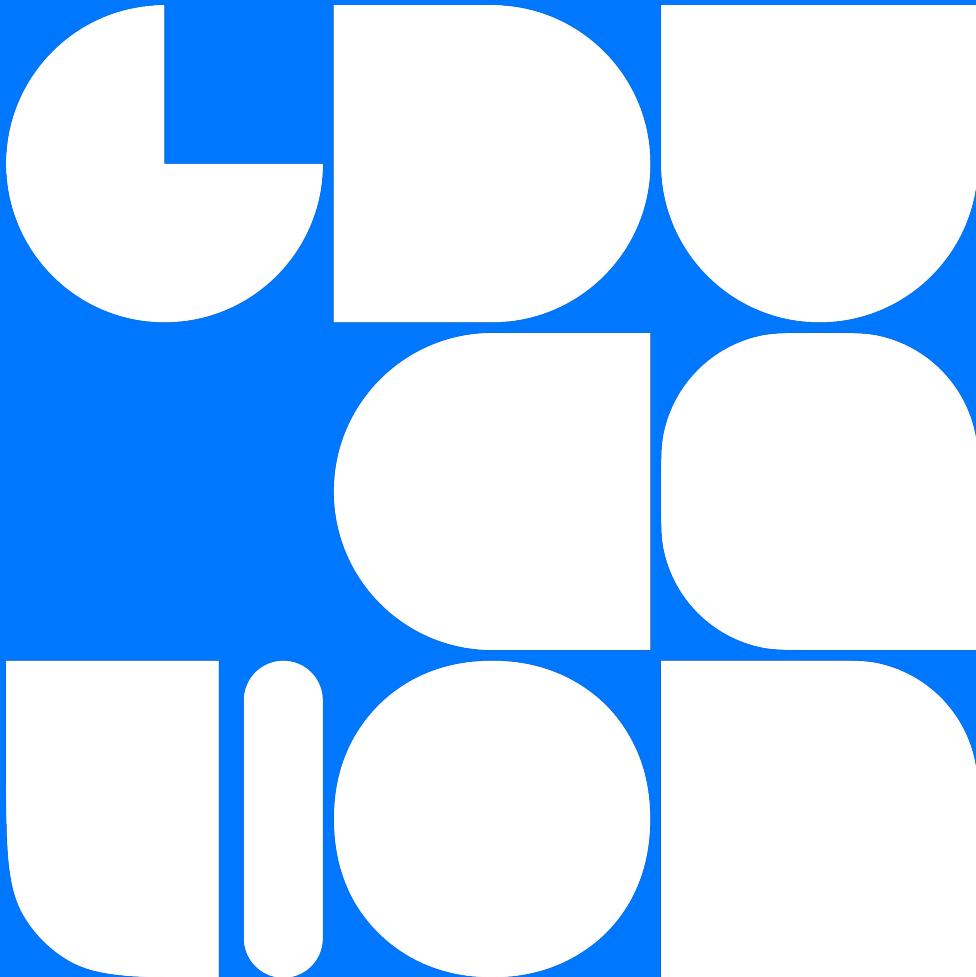
Deadline scheduler:

RUNTIME <= DEADLINE <= PERIOD

EEVDF (Earliest Eligible Virtual Deadline First)

Kernel 6.6

Практика



CPU: linux kernel scheduler

```
# ps -c -p <PID>
# top
# chrt -b 0 <programm> - запустить с политикой shed_batch
# chrt -p -i 0 <PID> - поменять приоритет на лету, в данном примере на sched_idle
# /proc/<PID>/sched
# /proc/<PID>/schedstat
# chrt -m - посмотреть диапазоны sched_prio для разных политик
# sysctl
# /etc/sysctl.conf
# /proc/sys/kernel/sched_rt_runtime_us
```

CPU: linux kernel scheduler

TS - SCHED_OTHER

B - SCHED_BATCH

IDL - SCHED_IDLE

FF - SCHED_FIFO

RR - SCHED_RR

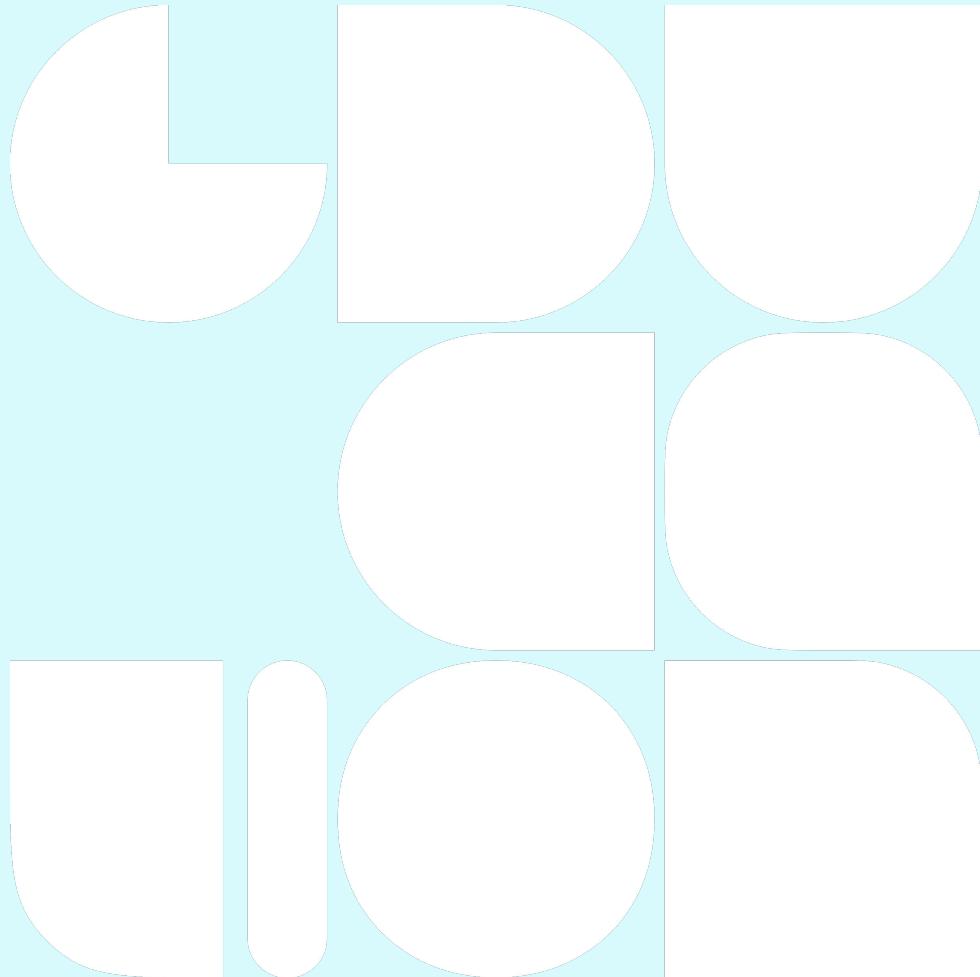
DLN - SCHED_DEADLINE



Вопросы?



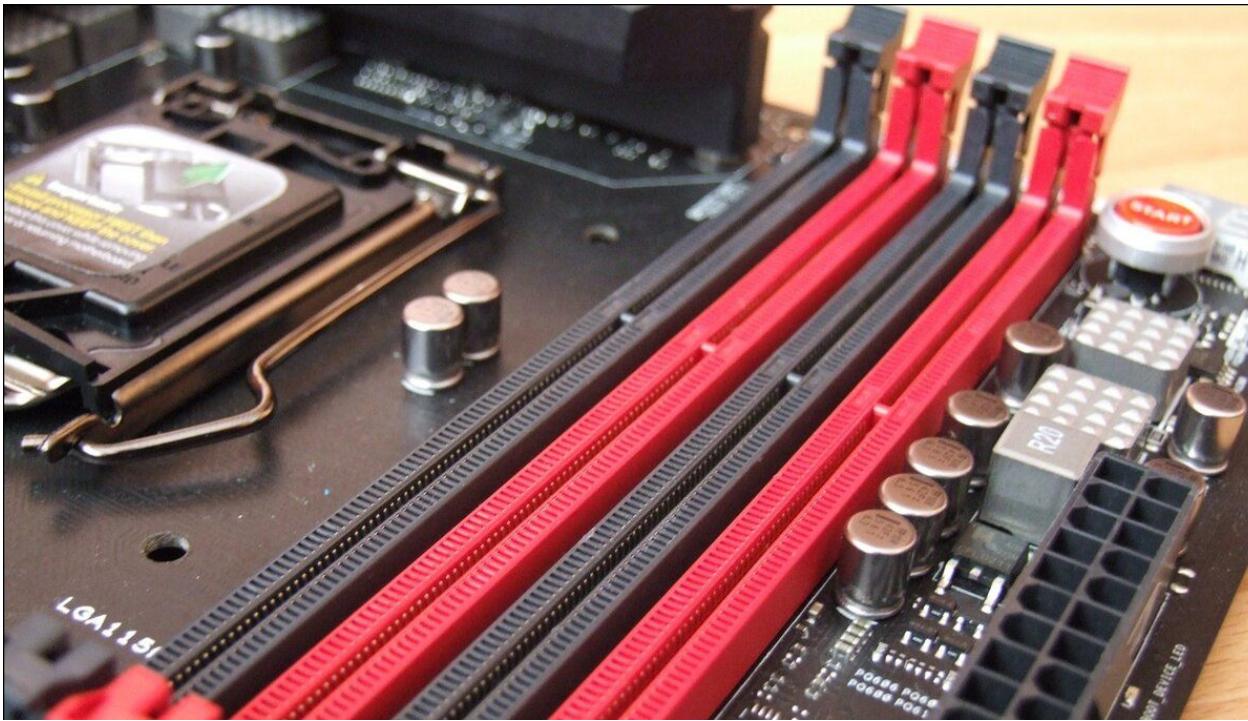
Память



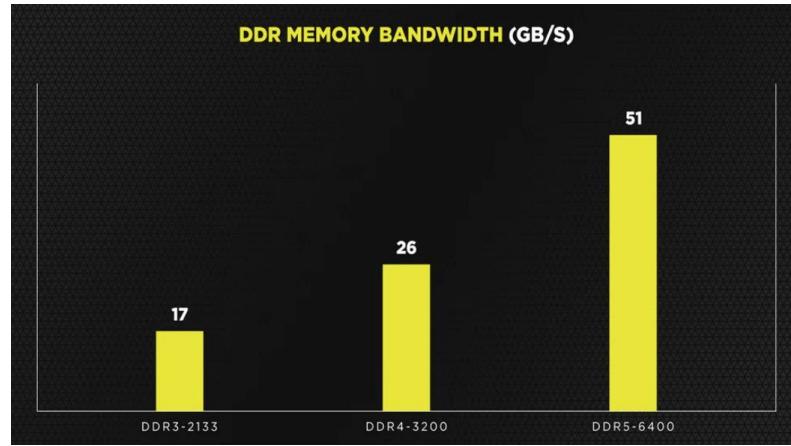
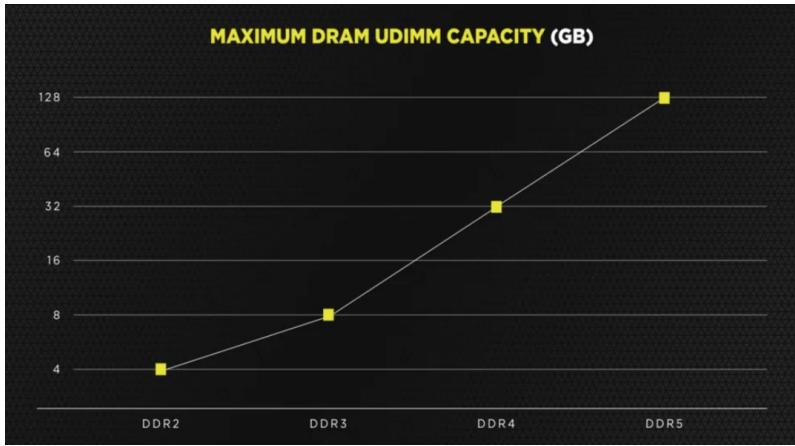
Память



Память



Память: DDR



Память

```
[root@1 ~]# free -m
      total    used    free   shared  buff/cache  available
Mem:      3174     478    2455      12      453     2695
Swap:    1638       0    1638
```

Память

```
# dmidecode --type=17
```

```
Handle 0x0051, DMI type 17, 92 bytes
Memory Device
    Array Handle: 0x0035
    Error Information Handle: Not Provided
    Total Width: Unknown
    Data Width: Unknown
    Size: No Module Installed
    Form Factor: DIMM
    Set: None
    Locator: CPU1_DIMM_F1
    Bank Locator: NODE 6
    Type: Unknown
    Type Detail: Unknown

Handle 0x0052, DMI type 17, 92 bytes
Memory Device
```

Память

```
# dmidecode --type=17
```

```
Handle 0x0052, DMI type 17, 92 bytes
Memory Device
  Array Handle: 0x0035
  Error Information Handle: Not Provided
  Total Width: 72 bits
  Data Width: 64 bits
  Size: 32 GB
  Form Factor: DIMM
  Set: None
  Locator: CPU1_DIMM_G0
  Bank Locator: NODE 7
  Type: DDR4
  Type Detail: Synchronous Registered (Buffered)
  Speed: 3200 MT/s
  Manufacturer: Samsung
  Serial Number: 2141-481F3B9C
  Asset Tag: CPU1_DIMM_G0_AssetTag
  Part Number: M393A4K40EB3-CWE
  Rank: 2
  Configured Memory Speed: 3200 MT/s
  Minimum Voltage: 1.2 V
  Maximum Voltage: 1.2 V
  Configured Voltage: 1.2 V
  Memory Technology: DRAM
  Memory Operating Mode Capability: Volatile memory
  Firmware Version: 0000
  Module Manufacturer ID: Bank 1, Hex 0xCE
  Module Product ID: Unknown
  Memory Subsystem Controller Manufacturer ID: Unknown
  Memory Subsystem Controller Product ID: Unknown
  Non-Volatile Size: None
  Volatile Size: 32 GB
  Cache Size: None
  Logical Size: None
```

Память: virtual memory

Process A

VPFN 7
VPFN 6
VPFN 5
VPFN 4
VPFN 3
VPFN 2
VPFN 1
VPFN 0

Process A
page tables

PFN 4
PFN 3
PFN 2
PFN 1
PFN 0

Process B

VPFN 7
VPFN 6
VPFN 5
VPFN 4
VPFN 3
VPFN 2
VPFN 1
VPFN 0

Virtual memory

Physical memory

Virtual memory

Память: зоны

```
[root@1 ~]# grep -P '^Node\s\d,\szone' /proc/zoneinfo
```

Node 0, zone DMA

Node 0, zone DMA32

Node 0, zone Normal

Node 0, zone Movable

Node 0, zone Device

Node 1, zone DMA

Node 1, zone DMA32

Node 1, zone Normal

Node 1, zone Movable

Node 1, zone Device

https://docs.kernel.org/mm/physical_memory.html

Память: что показывает top

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
897	root	20	0	9400	7076	4516	R	99.3	0.2	1:03.05	numa_test.py
1	root	20	0	171468	12668	9612	S	0.0	0.4	0:02.32	systemd
2	root	20	0	0	0	0	S	0.0	0.0	0:00.01	kthreadd
3	root	0	-20	0	0	0	I	0.0	0.0	0:00.00	rcu_gp
4	root	0	-20	0	0	0	T	0.0	0.0	0:00.00	rcu_tasks_kick

- VIRT (VSZ) - выделенная процессу виртуальная память
- RES (RSS) - потребляемая процессом физическая память (RAM)
- SHR - разделяемая резидентная (физическая) память

Память: /proc/meminfo

```
[root@l ~]# cat /proc/meminfo
MemTotal:      3250404 kB
MemFree:       2962260 kB
MemAvailable:  2889304 kB
Buffers:        0 kB
Cached:        52368 kB
SwapCached:     0 kB
Active:        23820 kB
Inactive:      126972 kB
Active(anon):   96 kB
Inactive(anon): 111032 kB
Active(file):   23724 kB
Inactive(file): 15940 kB
Unevictable:    4000 kB
Mlocked:        0 kB
SwapTotal:     1677684 kB
SwapFree:      1677684 kB
Dirty:          4 kB
Writeback:      0 kB
AnonPages:     102424 kB
Mapped:         36848 kB
Shmem:          12704 kB
KReclaimable:  29908 kB
Slab:           57504 kB
SReclaimable:  29908 kB
SUnreclaim:    27596 kB
KernelStack:    2080 kB
PageTables:    1652 kB
NFS_Unstable:   0 kB
Bounce:         0 kB
WritebackTmp:   0 kB
CommitLimit:   3302884 kB
Committed_AS:  244752 kB
```

Память: madvise

MADV_NORMAL

No special treatment. This is the default.

MADV_RANDOM

Expect page references in random order. (Hence, read ahead may be less useful than normally.)

MADV_SEQUENTIAL

Expect page references in sequential order. (Hence, pages in the given range can be aggressively read ahead, and may be freed soon after they are accessed.)

MADV_WILLNEED

Expect access in the near future. (Hence, it might be a good idea to read some pages ahead.)

Память: paging

Paging - это процесс перемещения страниц памяти между RAM и стораджем.

Память: swap

Хорошо, или плохо?

<https://chrисdown.name/2018/01/02/in-defence-of-swap.html>

Память: swap

Отключение swap

/etc/fstab:

```
#/dev/mapper/<swap-partition> swap      swap  defaults    0 0
```

Память: swap

Отключение swap

/etc/fstab:

```
#/dev/mapper/<swap-partition> swap      swap  defaults    0 0
```

```
# swapoff -a
```

Память: swap

Отключение swap

/etc/fstab:

```
#/dev/mapper/<swap-partition> swap      swap  defaults  0 0
```

```
# swapoff -a
```

```
# free -g | grep -i swap
```

```
Swap:      0      0      0
```

Память: page cache

Page cache - это кэш, используемый для чтения/записи файлов.

Память: page cache

```
# free -g
      total    used    free   shared  buff/cache   available
Mem:      15       6       0       1       9       8
```

Память: page cache

```
# free -g
      total    used    free   shared  buff/cache   available
Mem:       15       6       0       1       9       8
```

```
# sync; echo 1 > /proc/sys/vm/drop_caches
```

```
# vmtouch
```

Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

- при достижении временного интервала, по умолчанию это 30 сек (vm.dirty_expire_centisecs)

Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

- при достижении временного интервала, по умолчанию это 30 сек (vm.dirty_expire_centisecs)
- при системных вызовах sync(), fsync(), msync()

Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

- при достижении временного интервала, по умолчанию это 30 сек (vm.dirty_expire_centisecs)
- при системных вызовах sync(), fsync(), msync()
- при достижении слишком большого количества грязных страниц - dirty_ratio

Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

- при достижении временного интервала, по умолчанию это 30 сек (vm.dirty_expire_centisecs)
- при системных вызовах sync(), fsync(), msync()
- при достижении слишком большого количества грязных страниц - dirty_ratio
- при отсутствии доступных страниц в page cache

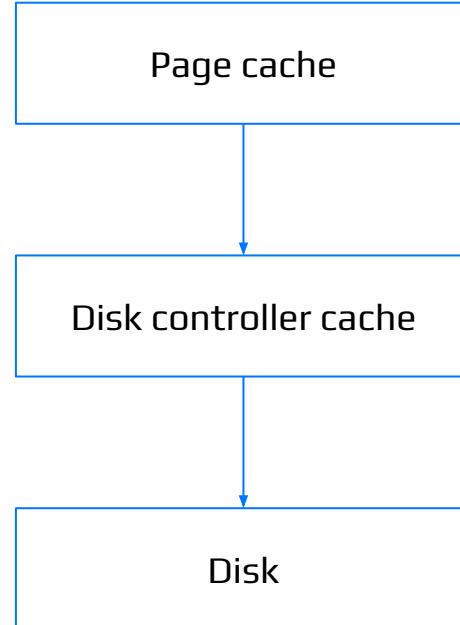
Память: page cache

Dirty pages флашатся на диск при выполнении одного из условий:

- при достижении временного интервала, по умолчанию это 30 сек (vm.dirty_expire_centisecs)
- при системных вызовах sync(), fsync(), msync()
- при достижении слишком большого количества грязных страниц - dirty_ratio
- при отсутствии доступных страниц в page cache

Но гарантирует ли это запись на диск?

Память: page cache



Память: page cache

Приблизительно, мы можем посчитать количество доступной нам памяти как:

$$\text{MemAvailable} = (\text{MemFree} - \text{low_watermark}) + (\text{Active(file)} + \text{Inactive(file)}) - \text{low_watermark} + (\text{SReclaimable} - \text{low_watermark})$$

`low_watermark` значения мы берем из `/proc/zoneinfo` (`low * 4`), остальные из `/proc/meminfo`

Память: OOM killer

OOM killer - Out Of Memory killer

Память: OOM killer

OOM killer - Out Of Memory killer

```
# echo -1000 > /proc/40060/oom_score_adj
```

Память: OOM killer

OOM killer - Out Of Memory killer

```
# echo -1000 > /proc/40060/oom_score_adj
```

Значения могут быть в диапазоне от -1000 до 1000.

Память: huge pages

grep pse /proc/cpuinfo - проверка поддержки страниц 2mb

grep pdpe1gb /proc/cpuinfo - проверка поддержки страниц 1gb

```
[root@1 ~]# grep hugetlbfs /proc/filesystems
nodev      hugetlbfs
```

<https://docs.kernel.org/admin-guide/mm/hugetlbp.html>

Память: huge pages

```
[root@1 ~]# sysctl -w vm.nr_hugepages=10
```

```
[root@1 ~]# cat /proc/sys/vm/nr_hugepages  
10
```

```
[root@1 ~]# grep -P "HugePages_Total|HugePages_Free" /proc/meminfo  
HugePages_Total:    10  
HugePages_Free:    10
```

```
[root@1 ~]# grep HugePages_Total /sys/devices/system/node/node*/meminfo  
/sys/devices/system/node/node0/meminfo:Node 0 HugePages_Total:  5  
/sys/devices/system/node/node1/meminfo:Node 1 HugePages_Total:  5
```

Память: huge pages

```
[root@1 ~]# mkdir /mnt/huge
```

```
[root@1 ~]# mount -t hugetlbfs none /mnt/huge
```

```
[root@1 huge]# mount | grep /mnt/huge
none on /mnt/huge type hugetlbfs (rw,relatime,pagesize=2M)
```

<input checked="" type="checkbox"/>	video-history.db.video.prod/cdb/64	memory	memdisk	MOUNTED	hugetlbfs1	40G	31.79G	79%
◀								

Память: transparent huge pages

khugepaged

/sys/kernel/mm/transparent_hugepage/enabled

```
# grep AnonHugePages /proc/meminfo
```

```
AnonHugePages: 2873344 kB
```

Память: полезные утилиты

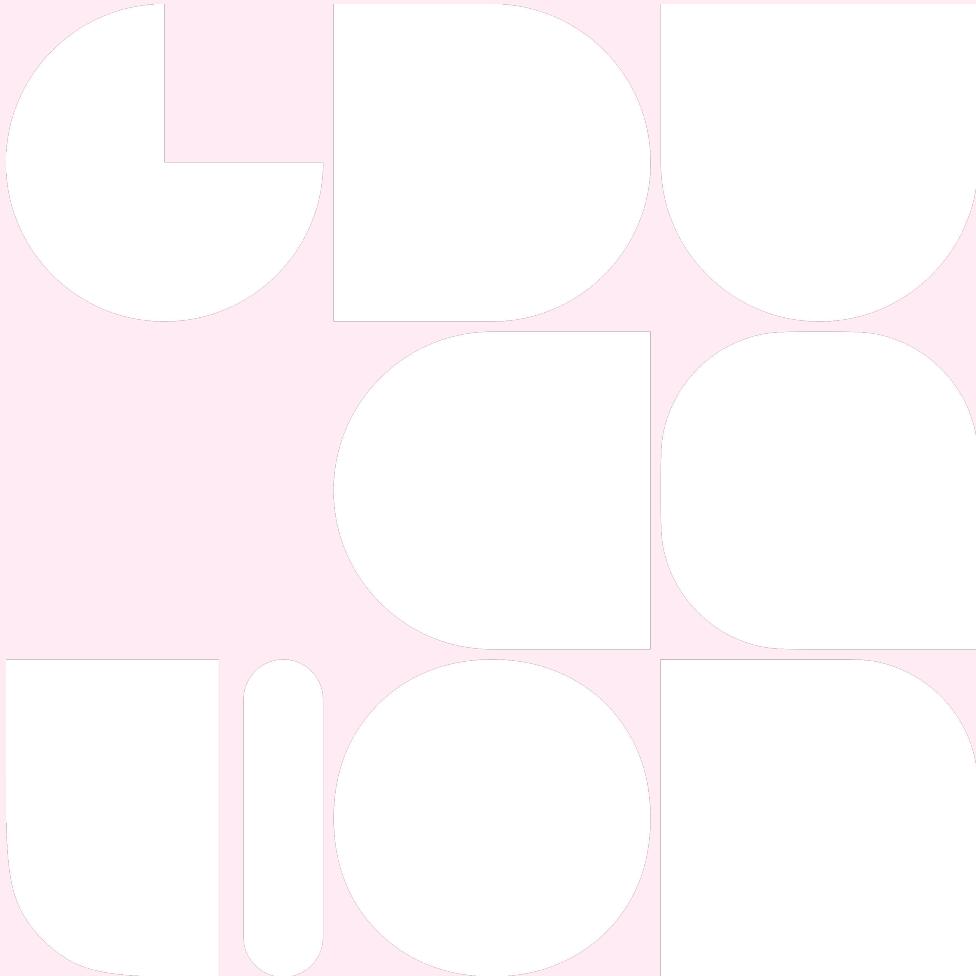
- free
- /proc/meminfo
- numactl
- numastat
- ps
- top
- /proc/<PID>/maps
- /proc/<PID>/smaps
- /proc/<PID>/numa_maps



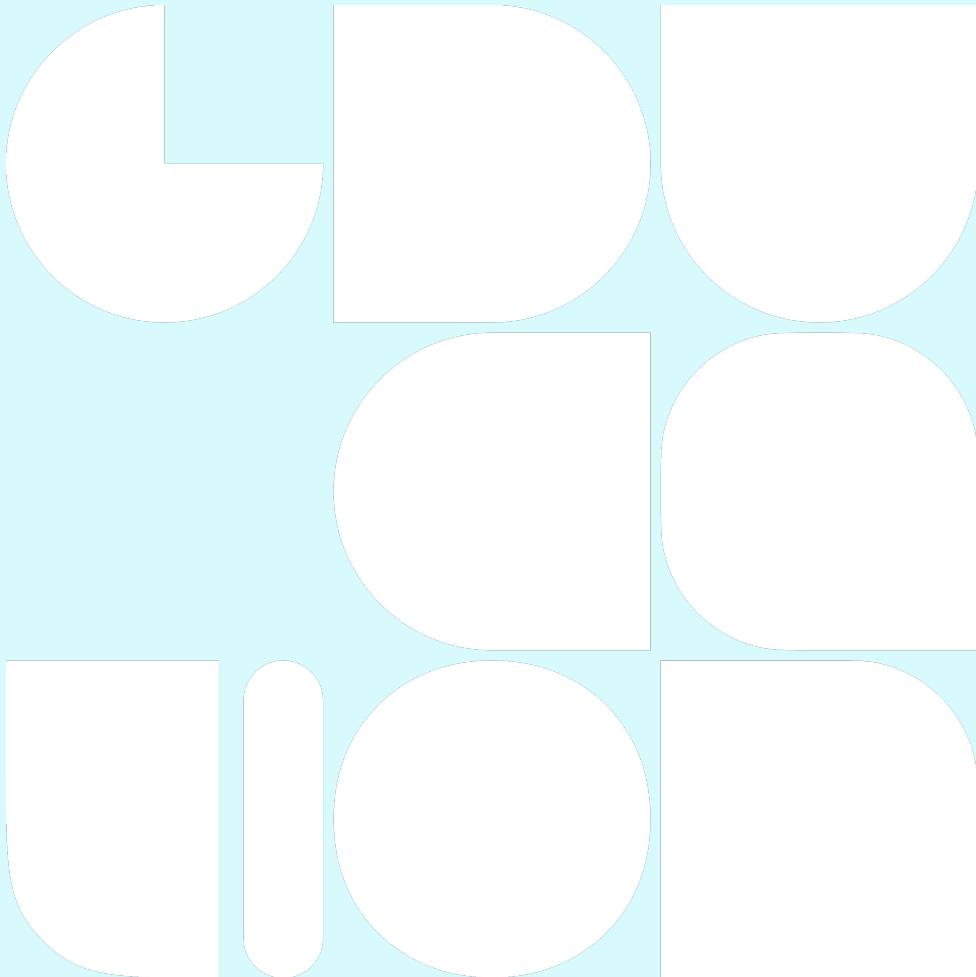
Вопросы?



Перерыв



Дисковая подсистема



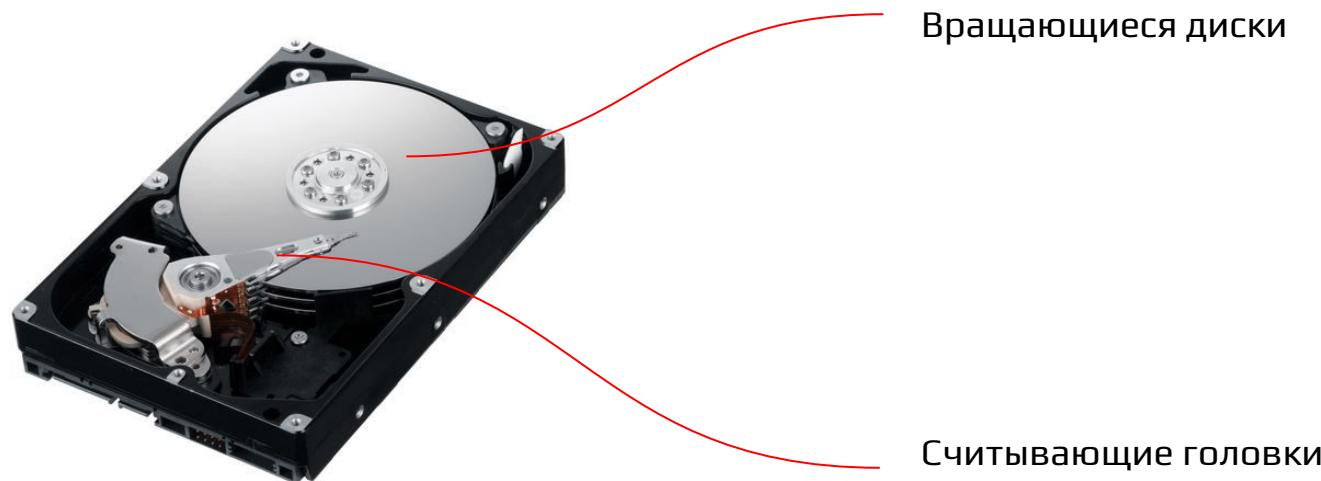
Дисковая подсистема



Дисковая подсистема: виды интерфейсов

- **SCSI**: Small Computer System Interface
- **ATA** (он же **IDE**): Advanced Technology Attachment
- **SATA**: Serial ATA
- **SAS**: Serial Attached SCSI

Дисковая подсистема: HDD



Дисковая подсистема: HDD



Hard Disk Drive: что влияет на скорость?

Дисковая подсистема: HDD



Hard Disk Drive: что влияет на скорость?

- скорость вращения шпинделя
- физическое расположение данных на диске

Дисковая подсистема: HDD



Hard Disk Drive: что влияет на скорость?

- скорость вращения шпинделя
- физическое расположение данных на диске
- плох для случайного чтения/записи

Дисковая подсистема: HDD



Hard Disk Drive: что влияет на скорость?

- скорость вращения шпинделя
- физическое расположение данных на диске
- плох для случайного чтения/записи
- неплох для последовательного чтения/записи

Дисковая подсистема: HDD



Hard Disk Drive: достоинства

- цена
- долговечность

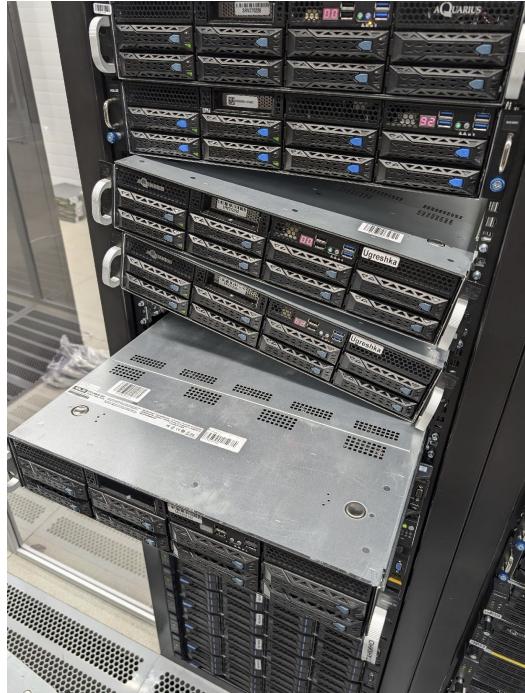
Дисковая подсистема: HDD



Hard Disk Drive: недостатки

- медленный, особенно на случайном доступе
- не более 200 iops
- потребляет много энергии
- шумит
- имеет подвижные части => подвержен механическим повреждениям

Дисковая подсистема: HDD



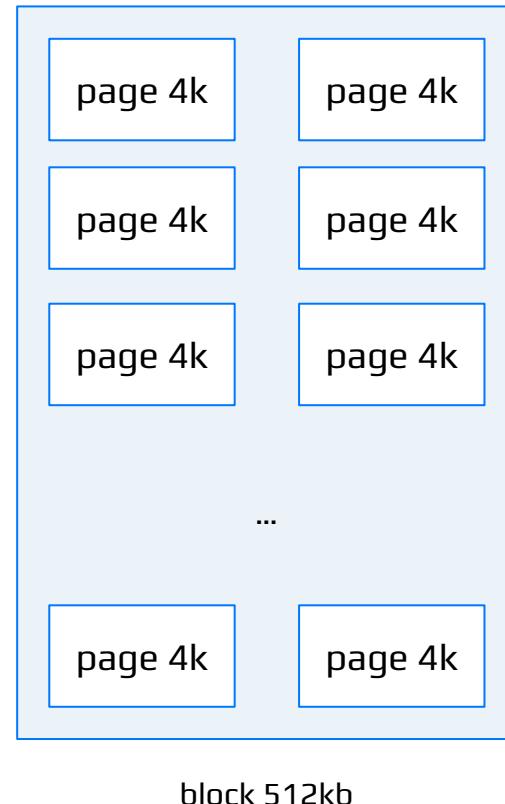
Бывает и так :)

Дисковая подсистема: SSD

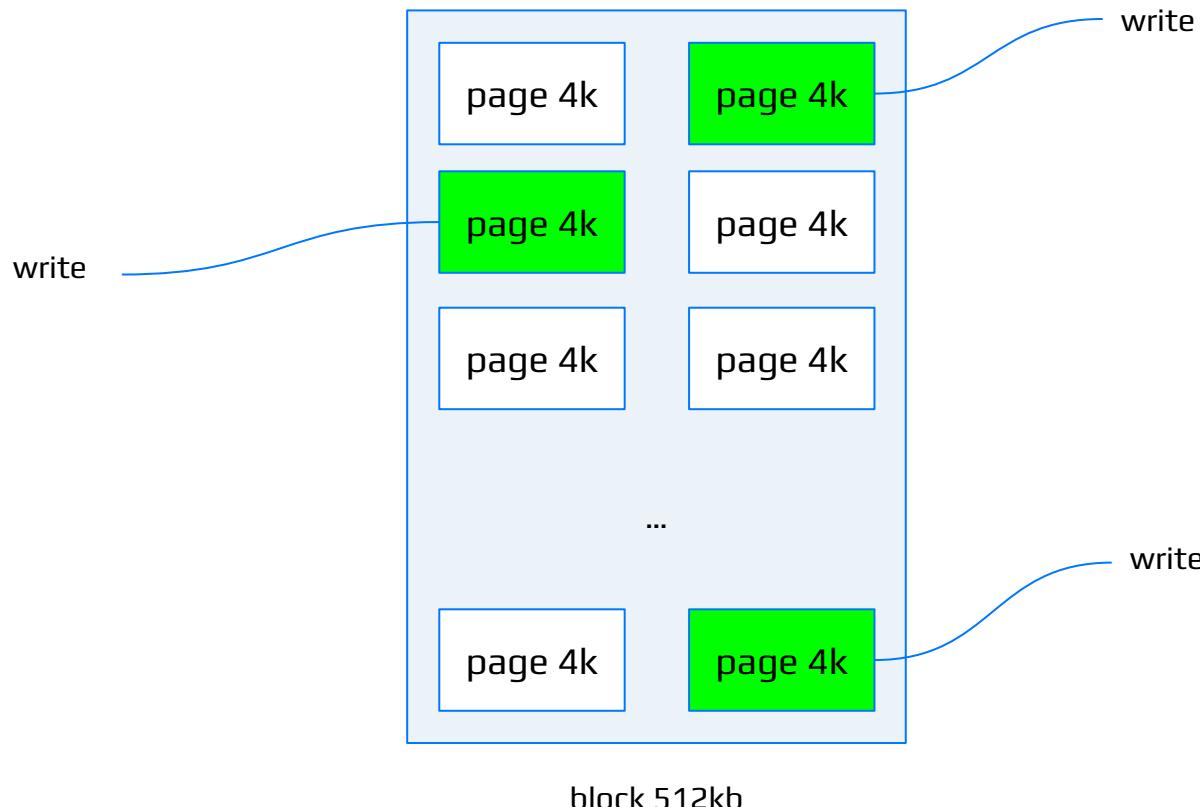


Solid State Drive

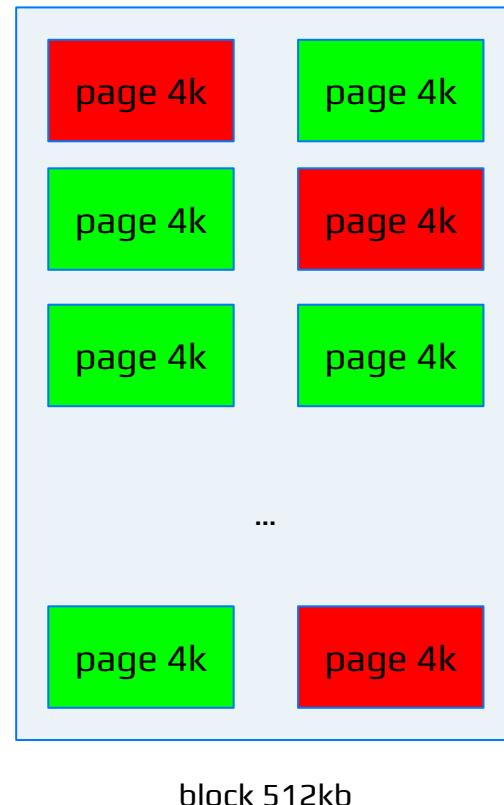
Дисковая подсистема: SSD



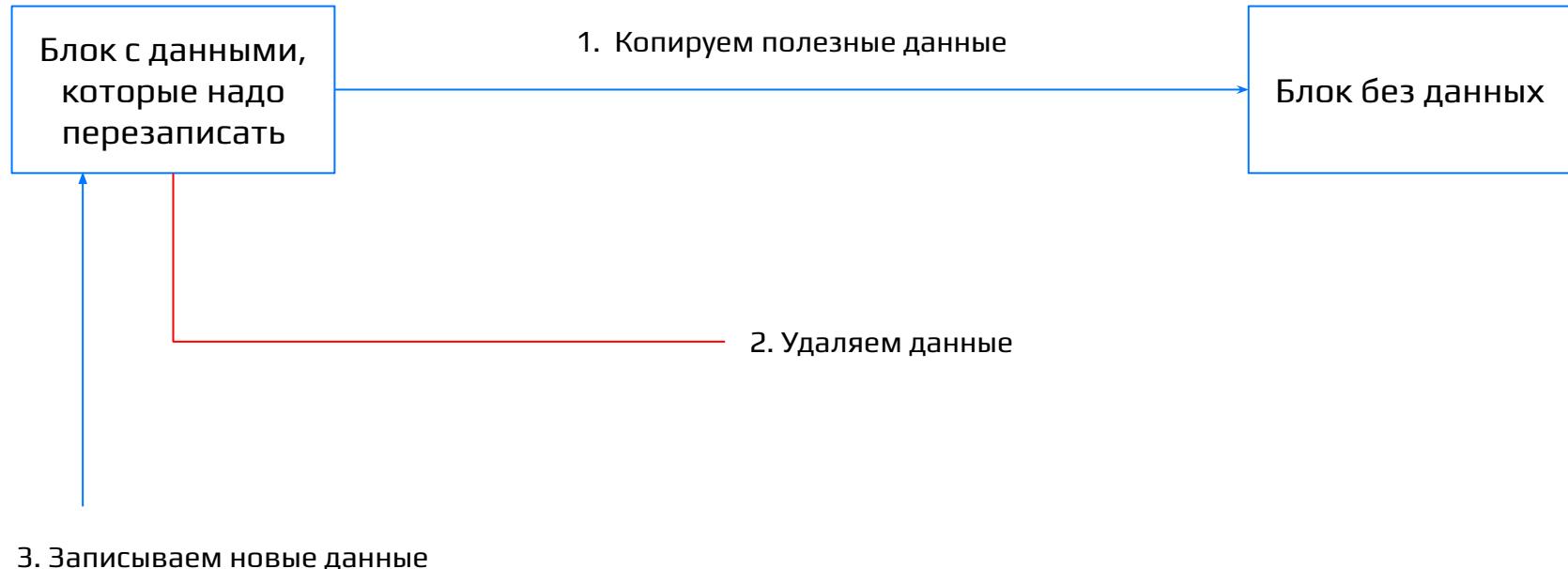
Дисковая подсистема: SSD



Дисковая подсистема: SSD



Дисковая подсистема: SSD



Дисковая подсистема: SSD

discard/fstrim

```
[root@l ~]# fstrim --help
Usage:
  fstrim [options] <mount point>
Discard unused blocks on a mounted filesystem.
```

Дисковая подсистема: SSD wearout

```
wearout_parser() {
    awk '
        BEGIN {
            wearout = -1
        }

        /Device Model:/ {
            device_model = substr($0, index($0,$3))
        }

        /^A mandatory SMART command failed/ {
            wearout = -146
        }

        /^173/ {
            if (device_model ~ /^TOSHIBA THNSN(H|J|8)/) wearout = $4/2
        }

        /^177/ {
            if (device_model ~ /MTFDDAK480TDS/) wearout = $4
            if (device_model ~ /MZ7L3960HCJR/) wearout = $4
            if (device_model ~ /MZ7LH480HAHQ/) wearout = $4
            if (device_model ~ /MZ7L(3|H)(240|480|960|1T9|3T8|7T6)HC/) wearout = $4
            if (device_model ~ /SAMSUNG MZ7LH(240|480|960|1T9|3T8|7T6)H[AM](HQ|JR|LT|LA)-.*/) wearout = $4
            if (device_model ~ /SAMSUNG MZ7KH(240|480|960|1T9|3T8)HA(HQ|JR|LS)-.*/) wearout = $4
            if (device_model ~ /SAMSUNG MZ7LM(120|240|480|960|1T9|3T8)HC(M|HP|GR|FD)-.*/) wearout = $4
            if (device_model ~ /(SAMSUNG )?MZ7LM(240|480|960|1T9|3T8)HM(JP|HQ|LP)(-.*|0D3)/) wearout = $4
        }

        /^233 / {
            if (device_model ~ /^INTEL SSD/) wearout = $4
            if (device_model ~ /IT3RSK41MT300/) wearout = $4 # OCZ Intrepid
            if (device_model ~ /OCZ INTREPID 360/) wearout = $4 # OCZ Intrepid
            if (device_model ~ /OCZ-SABER1000/) wearout = $4 # OCZ Saber1000
            if (device_model ~ /OCZ-VERTEX4/) wearout = $4 # OCZ Vertex4
            if (device_model ~ /SDLF1(DAR|CRR)/) wearout = $4 # Sandisk eco gen II
            if (device_model ~ /^CORSAIR CMFSSD/) wearout = $4
        }

        /^231/ {
            if (device_model ~ /^KINGSTON / ) wearout = $4
            if (device_model ~ /D2CSTK251M3T/) wearout = $4 # OCZ Deneva 2 C
            if (device_model == "Benetek SSD" ) wearout = $4
            if (device_model == "RunCore SSD" ) wearout = $4
            if (device_model ~ /^HYNIX HFS/ ) wearout = $4
            if (device_model ~ /XA(240|480|960|1920|3840)[LM]E10(00|02|04|06|08|10)3/) wearout = $4
        }
    '
}
```

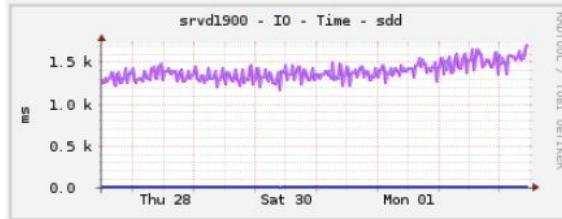
Дисковая подсистема: SSD wearout

```
[root@srvd4061 ~]# smartctl -a /dev/sdb | grep -i Percent_Lifetime_Remain  
202 Percent_Lifetime_Remain 0x0030 077 077 001 Old_age Offline - 23
```

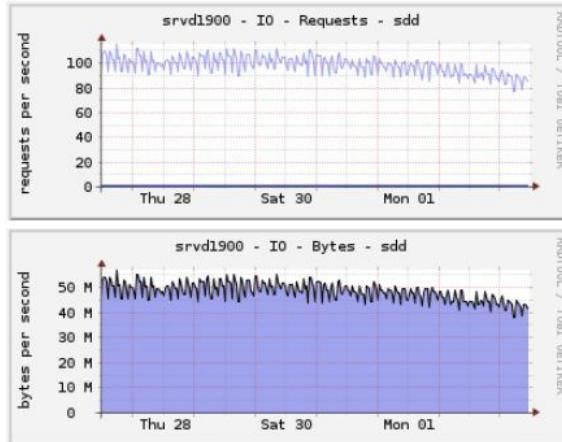
Дисковая подсистема: SSD wearout

По эффектам:

когда wearout дошел до 6 (~3%) начало расти время:

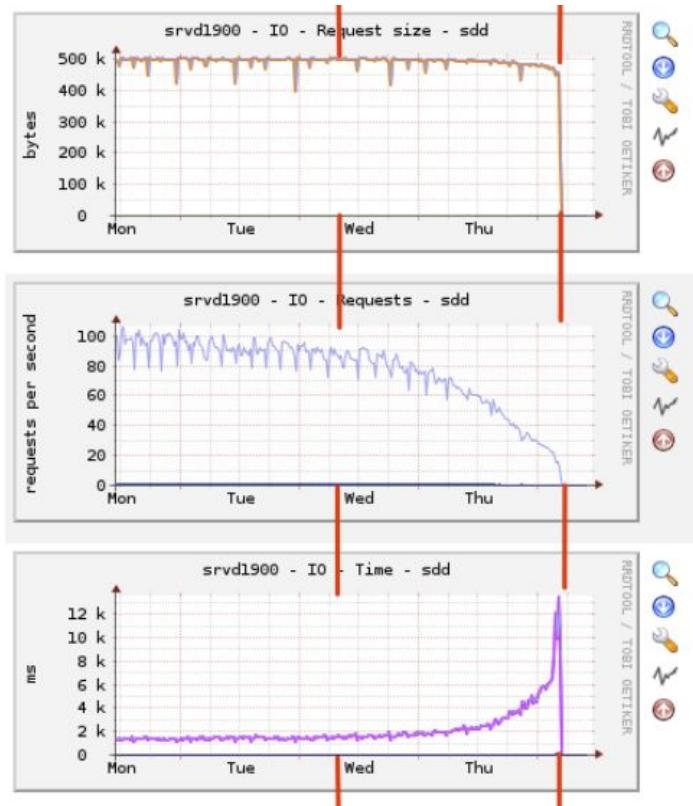
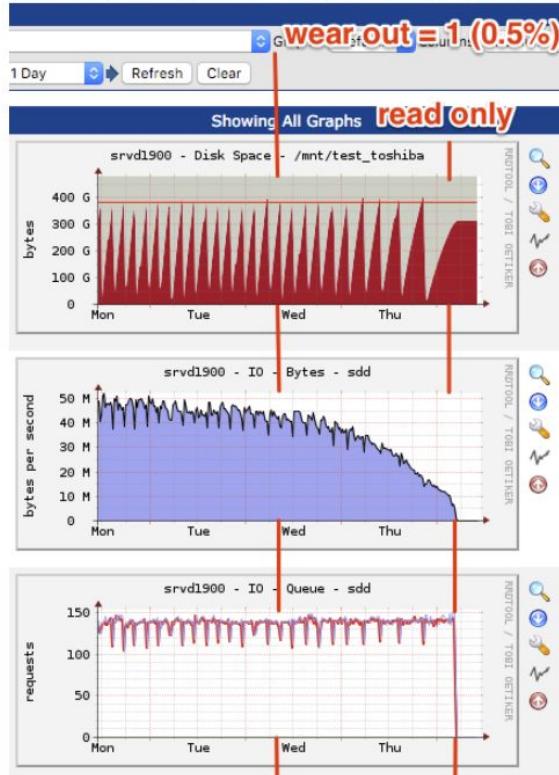


при том что количество запросов и, соответственно трафик на диске начал проседать:



Дисковая подсистема: SSD wearout

После того как диск дошел до 1 wear out, началась очень стремительная деградация:



Дисковая подсистема: SSD wearout

```
Feb  5 04:49:11 srvd1900 kernel: [6859508.637286] ata4.00: exception Emask 0x0 SAct 0x7fffffff SErr 0x0 action 0x6 frozen
Feb  5 04:49:11 srvd1900 kernel: [6859508.647425] ata4.00: failed command: WRITE FPDMA QUEUED
Feb  5 04:49:11 srvd1900 kernel: [6859508.654897] ata4.00: cmd 61/00:00:d0:d1:89/04:00:27:00:00/40 tag 0 ncq 524288 out
Feb  5 04:49:11 srvd1900 kernel: [6859508.654897]          res 40/00:01:00:00:00/00:00:00:00/00 Emask 0x4 (timeout)
Feb  5 04:49:11 srvd1900 kernel: [6859508.674252] ata4.00: status: { DRDY }
Feb  5 04:49:11 srvd1900 kernel: [6859508.680004] ata4.00: failed command: WRITE FPDMA QUEUED
...
Feb  5 04:49:11 srvd1900 kernel: [6859509.525345] ata4.00: failed command: WRITE FPDMA QUEUED
Feb  5 04:49:11 srvd1900 kernel: [6859509.531693] ata4.00: cmd 61/00:f0:d0:cd:89/04:00:27:00:00/40 tag 30 ncq 524288 out
Feb  5 04:49:11 srvd1900 kernel: [6859509.531693]          res 40/00:01:00:00:00/00:00:00:00/00 Emask 0x4 (timeout)
Feb  5 04:49:11 srvd1900 kernel: [6859509.548840] ata4.00: status: { DRDY }
...
Feb  5 04:49:22 srvd1900 kernel: [6859520.475680] XFS (sdd1): Log I/O Error Detected. Shutting down filesystem
Feb  5 04:49:22 srvd1900 kernel: [6859520.475684] XFS (sdd1): Please umount the filesystem and rectify the problem(s)
Feb  5 04:49:22 srvd1900 kernel: [6859520.476188] XFS (sdd1): metadata I/O error: block 0x1dd41fde ("xlog_iodone") error 5 numblks 0
Feb  5 04:49:22 srvd1900 kernel: [6859520.476191] XFS (sdd1): xfs_do_force_shutdown(0x2) called from line 1172 of file fs/xfs/xfs_l
...
Feb  5 04:49:23 srvd1900 kernel: [6859520.477370] lost page write due to I/O error on sdd1
Feb  5 04:49:23 srvd1900 kernel: [6859520.477372] Buffer I/O error on device sdd1, logical block 82932648
...
```

Дисковая подсистема: SSD



Solid State Drive: преимущества

- до 100000 iops
- высокая скорость (до 600 мб/с)
- малые габариты
- низкое энергопотребление
- отсутствие подвижных частей => не шумит, устойчив к механическим нагрузкам

Дисковая подсистема: SSD



Solid State Drive: недостатки

- ограниченное количество циклов перезаписи
- скорость записи зависит от свободного пространства
- высокая стоимость, относительно HDD

Дисковая подсистема: NVMe



Используются операционными системами для обмена данными с устройствами хранения

	AHCI	NVMe	
	Разработано для жестких дисков с технологией вращающихся дисков		Разработано для твердотельных накопителей с технологией флеш-памяти
1	Есть только 1 очередь команд	64 тыс.	Есть 64 тыс. команд на очередь
32	Может отправлять только 32 команды на очередь		Может отправлять 64 тыс. команд на очередь
	Команды используют большое количество циклов ЦП		Команды используют малое количество циклов ЦП
	Задержка 6 микросекунд		Задержка 2,8 микросекунд
	Требуется связь с контроллером SATA		Связывается напрямую с ЦП системы
	IOPs до 100 тыс.		IOPs более 1 миллиона

Дисковая подсистема: NVMe

Форм-факторы SSD-накопителей: форма и размеры твердотельных накопителей

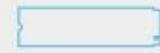
Клиентские
системы



BGA



M.2 2230



M.2 2280



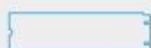
M.2 22110

Серверы

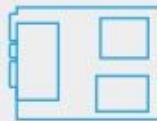
U.2 - 2,5 дюйма,
U.3 - 2,5 дюйма



EDSFF E1.S



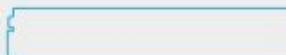
EDSFF E3.S



EDSFF E3.L



EDSFF E1.L



M.2 22110



Дисковая подсистема: NVMe



Non-Volatile Memory Express: преимущества

- свыше 1000000 iops
- очень высокая скорость (до 32000 мб/с)
- энергоэффективность
- очень низкие задержки

Дисковая подсистема: NVMe



Non-Volatile Memory Express: недостатки

- высокая стоимость

Дисковая подсистема: latency

- **NVMe SSD:** задержка около 30 мкс.
- **SATA SSD:** задержка около 100 мкс.
- **HDD:** задержка около 2-5 мс.

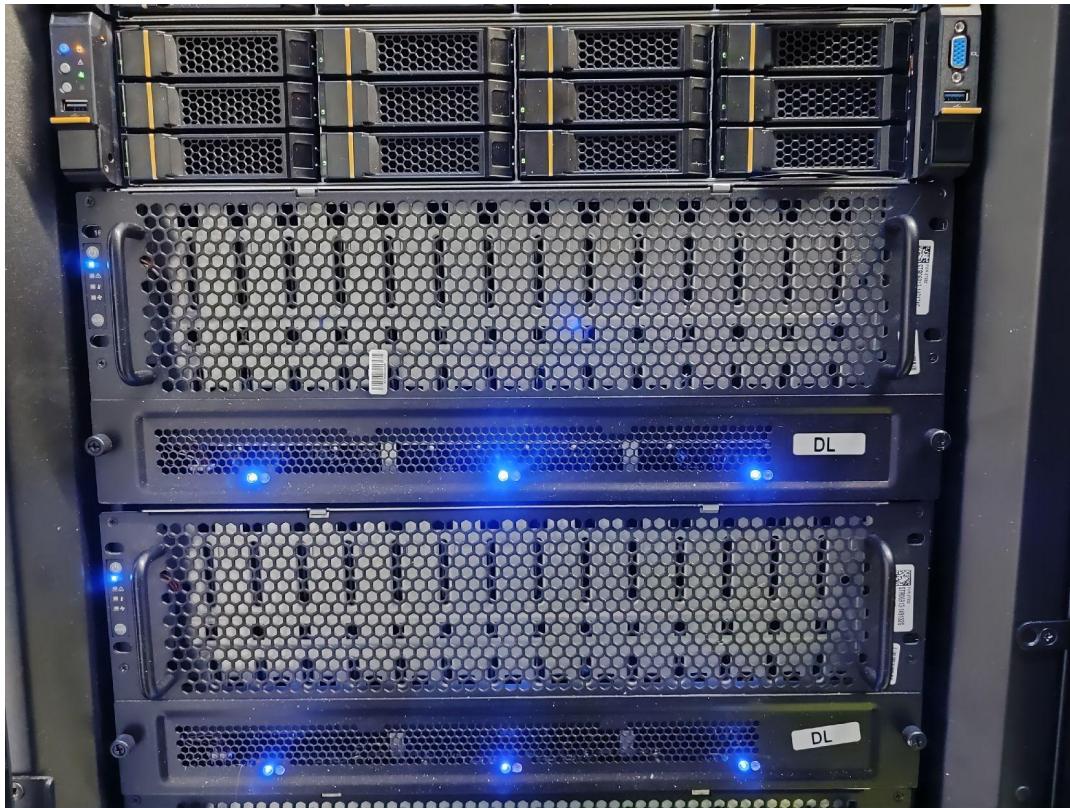
Дисковая подсистема: а что же выбрать?



Дисковая подсистема: а что же выбрать?



Дисковая подсистема: а что же выбрать?



$132 * 20 \text{ TB} = 2,64 \text{ PB}$

Дисковая подсистема: а что же выбрать?



$$264 * 20 \text{ TB} = 5,28 \text{ PB}$$

Дисковая подсистема: а что же выбрать?

```
[root@srvd6068 ~]# lsscsi
[0:0:0:0] disk ATA KINGSTON SEDC500 J2.8 /dev/sda
[1:0:0:0] disk ATA KINGSTON SEDC500 J2.8 /dev/sdj
[2:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdh
[3:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdd
[4:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdc
[5:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdg
[6:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdm
[7:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdi
[8:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdb
[9:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sde
[10:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdk
[11:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdl
[12:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdn
[13:0:0:0] disk ATA ST20000NM007D-3D SN01 /dev/sdf
[14:0:0:0] enclosu AIC 12G 4U0: Hub 0c29 -
[14:0:1:0] disk ATA ST20000NM007D-3D SN01 /dev/sdo
[14:0:2:0] disk ATA ST20000NM007D-3D SN01 /dev/sdp
[14:0:3:0] disk ATA ST20000NM007D-3D SN01 /dev/sdq
[14:0:4:0] disk ATA ST20000NM007D-3D SN01 /dev/sdr
[14:0:5:0] disk ATA ST20000NM007D-3D SN01 /dev/sds
[14:0:6:0] disk ATA ST20000NM007D-3D SN01 /dev/sdt
[14:0:7:0] disk ATA ST20000NM007D-3D SN01 /dev/sdu
[14:0:8:0] disk ATA ST20000NM007D-3D SN01 /dev/sdv
[14:0:9:0] disk ATA ST20000NM007D-3D SN01 /dev/sdw
[14:0:10:0] disk ATA ST20000NM007D-3D SN01 /dev/sdx
[14:0:11:0] disk ATA ST20000NM007D-3D SN01 /dev/sdy
[14:0:12:0] disk ATA ST20000NM007D-3D SN01 /dev/sdz
[14:0:13:0] disk ATA ST20000NM007D-3D SN01 /dev/sdaa
[14:0:14:0] disk ATA ST20000NM007D-3D SN01 /dev/sdab
[14:0:15:0] disk ATA ST20000NM007D-3D SN01 /dev/sdac
[14:0:16:0] disk ATA ST20000NM007D-3D SN01 /dev/sdad
[14:0:17:0] disk ATA ST20000NM007D-3D SN01 /dev/sdae
[14:0:18:0] disk ATA ST20000NM007D-3D SN01 /dev/sdaf
[14:0:19:0] disk ATA ST20000NM007D-3D SN01 /dev/sdag
[14:0:20:0] disk ATA ST20000NM007D-3D SN01 /dev/sdah
[14:0:21:0] enclosu AIC 12G 4U0: Edge-C 0c2a -
[14:0:22:0] disk ATA ST20000NM007D-3D SN01 /dev/sdai
[14:0:23:0] disk ATA ST20000NM007D-3D SN01 /dev/sdaj
[14:0:24:0] disk ATA ST20000NM007D-3D SN01 /dev/sdak
[14:0:25:0] disk ATA ST20000NM007D-3D SN01 /dev/sdal
[14:0:26:0] disk ATA ST20000NM007D-3D SN01 /dev/sdam
[14:0:27:0] disk ATA ST20000NM007D-3D SN01 /dev/sdan
[14:0:28:0] disk ATA ST20000NM007D-3D SN01 /dev/sdao
[14:0:29:0] disk ATA ST20000NM007D-3D SN01 /dev/sdap
[14:0:30:0] disk ATA ST20000NM007D-3D SN01 /dev/sdaq
[14:0:31:0] disk ATA ST20000NM007D-3D SN01 /dev/sdar
[14:0:32:0] disk ATA ST20000NM007D-3D SN01 /dev/sdas
[14:0:33:0] disk ATA ST20000NM007D-3D SN01 /dev/sdat
```

Дисковая подсистема

```
# lsblk  
[0:0:0:0] disk  ATA  KINGSTON SH103S3 BBFO /dev/sda  
[1:0:0:0] disk  ATA  KINGSTON SH103S3 BBFO /dev/sdb  
[2:0:0:0] disk  ATA  KINGSTON SH103S3 BBFO /dev/sdd  
[3:0:0:0] disk  ATA  KINGSTON SH103S3 BBFO /dev/sde  
[6:0:0:0] disk  Promise VTrak E610f  0336 /dev/sdc
```

Дисковая подсистема

```
# lsblk
NAME  MAJ:MIN RM  SIZE RO TYPE MOUNTPOINTS
sr0   11:0    1  374K  0 rom
vda   252:0   0  10G  0 disk
└─vda1 252:1   0  10G  0 part /
vdb   252:16   0 12G  0 disk
vdc   252:32   0 12G  0 disk
vdd   252:48   0 12G  0 disk
vde   252:64   0 1.6G 0 disk
└─vde1 252:65   0 1.6G 0 part [SWAP]
```

Дисковая подсистема: создание раздела

```
# parted /dev/vdb mklabel gpt
```

```
# parted /dev/vdb mkpart primary 1024KiB 100%
```

```
# ls -l /dev/vdb*
```

```
# mkfs.xfs /dev/vdb1
```

/etc/fstab:

```
/dev/vdb1 /mnt xfs defaults 0 0
```

```
# mount -a
```

Дисковая запутанность

```
[root@srvk2835 dev]# lsscsi | grep -P "sd\w+"
[9:0:0:0] disk ATA SAMSUNG MZ7LM480 404Q /dev/sda
[10:0:0:0] disk ATA SAMSUNG MZ7LM480 404Q /dev/sdb
[13:0:0:0] disk ATA ST16000NM001G-2K SN02 /dev/sdc
[13:0:1:0] disk ATA ST16000NM001G-2K SN02 /dev/sdd
[13:0:2:0] disk ATA ST16000NM001G-2K SN02 /dev/sde
[13:0:3:0] disk ATA ST16000NM001G-2K SN02 /dev/sdf
[13:0:4:0] disk ATA ST16000NM001G-2K SN02 /dev/sdg
[13:0:5:0] disk ATA ST16000NM001G-2K SN02 /dev/sdh
[13:0:6:0] disk ATA ST16000NM001G-2K SN02 /dev/sdi
[13:0:7:0] disk ATA ST16000NM001G-2K SN02 /dev/sdj
[13:0:8:0] disk ATA ST16000NM001G-2K SN02 /dev/sdk
[13:0:9:0] disk ATA ST16000NM001G-2K SN02 /dev/sdl
[13:0:10:0] disk ATA ST16000NM001G-2K SN02 /dev/sdm
[13:0:11:0] disk ATA ST16000NM001G-2K SN02 /dev/sdn
[13:0:13:0] disk ATA ST16000NM001G-2K SN02 /dev/sdo
[13:0:14:0] disk ATA ST16000NM001G-2K SN02 /dev/sdp
[13:0:15:0] disk ATA ST16000NM001G-2K SN02 /dev/sdq
[13:0:17:0] disk ATA ST16000NM001G-2K SN02 /dev/sds
[13:0:18:0] disk ATA ST16000NM001G-2K SN02 /dev/sdt
[13:0:19:0] disk ATA ST16000NM001G-2K SN02 /dev/sdu
[13:0:20:0] disk ATA ST16000NM001G-2K SN02 /dev/sdv
[13:0:21:0] disk ATA ST16000NM001G-2K SN02 /dev/sdw
[13:0:22:0] disk ATA ST16000NM001G-2K SN02 /dev/sdx
[13:0:23:0] disk ATA ST16000NM001G-2K SN02 /dev/sdy
[13:0:24:0] disk ATA ST16000NM001G-2K SN02 /dev/sdz
[13:0:25:0] disk ATA ST16000NM001G-2K SN02 /dev/sdaa
[13:0:26:0] disk ATA ST16000NM001G-2K SN02 /dev/sdab
[13:0:27:0] disk ATA ST16000NM001G-2K SN02 /dev/sdac
[13:0:28:0] disk ATA ST16000NM001G-2K SN02 /dev/sdad
[13:0:29:0] disk ATA ST16000NM001G-2K SN02 /dev/sdae
[13:0:30:0] disk ATA ST16000NM001G-2K SN02 /dev/sdaf
[13:0:31:0] disk ATA ST16000NM001G-2K SN02 /dev/sdag
[13:0:32:0] disk ATA ST16000NM001G-2K SN02 /dev/sdah
[13:0:33:0] disk ATA ST16000NM001G-2K SN02 /dev/sdai
[13:0:34:0] disk ATA ST16000NM001G-2K SN02 /dev/sdaj
[13:0:35:0] disk ATA ST16000NM001G-2K SN02 /dev/sdak
[13:0:36:0] disk ATA ST16000NM001G-2K SN02 /dev/sdal
[13:0:38:0] disk ATA TOSHIBA MG10ACA2 4307 /dev/sdr
```

Дисковая запутанность

```
[root@srvk2835 dev]# ls -l /dev/disk/
итого 0
drwxr-xr-x. 2 root root 3340 июл  1 12:14 by-id
drwxr-xr-x. 2 root root  760 июл  1 12:15 by-label
drwxr-xr-x. 2 root root   60 июл  1 12:14 by-partlabel
drwxr-xr-x. 2 root root  840 июл  1 12:14 by-partuuid
drwxr-xr-x. 2 root root 3040 июл  1 12:14 by-path
drwxr-xr-x. 2 root root  820 июл  1 12:15 by-uuid
```

Дисковая запутанность

```
[root@srvk2835 dev]# ls -l /dev/disk/by-uuid | grep -P "sd\w+"
lrwxrwxrwx. 1 root root 10 июл 14 2023 025ad70b-42f5-4a84-9659-79e15a3e195c -> ../../sdy1
lrwxrwxrwx. 1 root root 11 июл 14 2023 05e5865f-70a4-4459-88d3-7b8a89604909 -> ../../sdafl
lrwxrwxrwx. 1 root root 11 июл 14 2023 1183abd6-f7fe-4d1b-b46e-89b4c6a4b9d6 -> ../../sda1
lrwxrwxrwx. 1 root root 10 июл 14 2023 15920725-3435-4296-bcd2-494a8dc44bed -> ../../sdf1
lrwxrwxrwx. 1 root root 10 июл 14 2023 1c8dfbda-6d21-4add-a915-2074d8ac953b -> ../../sds1
lrwxrwxrwx. 1 root root 10 июл 14 2023 1d25b948-499d-4941-afb6-585e5ae7d005 -> ../../sdl1
lrwxrwxrwx. 1 root root 10 июл 14 2023 26a203ca-6526-4aa9-bc41-d11799fb47be -> ../../sdv1
lrwxrwxrwx. 1 root root 11 июл 14 2023 2deae0c7-08c3-4618-81b6-cbd3c176ae71 -> ../../sdacl
lrwxrwxrwx. 1 root root 10 июл 14 2023 368fc348-a3b6-4c2c-8ac0-7195017b451f -> ../../sdo1
lrwxrwxrwx. 1 root root 10 июл 1 12:15 601eddbf-9778-49d4-8ee4-41b4506aceef -> ../../sdr1
lrwxrwxrwx. 1 root root 11 июл 14 2023 612f4892-84a2-417f-95da-d5c74244d7b2 -> ../../sdael
lrwxrwxrwx. 1 root root 11 июл 14 2023 62b473d9-1cb4-48ff-b331-92f117de8acb -> ../../sda1
lrwxrwxrwx. 1 root root 10 июл 14 2023 66369ad8-d8b6-4e2f-9417-67e5e5eb6e11 -> ../../sdl1
lrwxrwxrwx. 1 root root 11 июл 14 2023 7493e48e-441a-4381-bd49-2473e039e279 -> ../../sdaal
lrwxrwxrwx. 1 root root 10 июл 14 2023 7c858b7e-dded-4aab-b981-01bdf1d148def -> ../../sdel
lrwxrwxrwx. 1 root root 11 июл 14 2023 7d268c0c-2311-47b5-8ae5-6bed22e1d08a -> ../../sda1
lrwxrwxrwx. 1 root root 10 июл 14 2023 810aebf2-02b9-4c67-9a30-0e38306a11cf -> ../../sdx1
lrwxrwxrwx. 1 root root 10 июл 14 2023 8607901a-f24e-40a2-b0a3-0608990cd5cc -> ../../sdl1
lrwxrwxrwx. 1 root root 10 июл 14 2023 8f724360-b171-408f-8962-72eea3100699 -> ../../sdl1
lrwxrwxrwx. 1 root root 10 июл 14 2023 9137e2cf-8a67-4e2f-b6d8-042f3ef50768 -> ../../sdp1
lrwxrwxrwx. 1 root root 10 июл 14 2023 999ba07f-cc8b-434c-91d9-3e863bf33191 -> ../../sdc1
lrwxrwxrwx. 1 root root 10 июл 14 2023 999d6364-50d2-4123-86d2-8e5f1250bf8f -> ../../sdh1
lrwxrwxrwx. 1 root root 10 июл 14 2023 9b528ae6-da4f-48d9-a35b-6612b47ef101 -> ../../sdwl
lrwxrwxrwx. 1 root root 10 июл 14 2023 9bb82fdb-8bac-48d0-96b3-5ef7fa423334 -> ../../sdi1
lrwxrwxrwx. 1 root root 10 июл 14 2023 a3885289-d916-49a8-b515-bfce1b19425ec -> ../../sdk1
lrwxrwxrwx. 1 root root 10 июл 14 2023 ac615033-8937-4377-97ad-26d2bdb14927 -> ../../sdi1
lrwxrwxrwx. 1 root root 10 июл 14 2023 bf9b36cf-6462-41eb-88a8-6ed59e971b9a -> ../../sdt1
lrwxrwxrwx. 1 root root 10 июл 14 2023 c48ced11-7813-4006-b092-f50fa18cbefe -> ../../sdd1
lrwxrwxrwx. 1 root root 10 июл 14 2023 c76cb0fb-8d41-4a72-96ff-a9b717498ab3 -> ../../sdu1
lrwxrwxrwx. 1 root root 10 июл 14 2023 d7273131-fc03-484f-97a8-a60ae5belbab -> ../../sdml
lrwxrwxrwx. 1 root root 11 июл 14 2023 d9461839-722b-4064-9d5b-dfa2282d4e58 -> ../../sdad1
lrwxrwxrwx. 1 root root 11 июл 14 2023 dd597c36-7559-4e6e-a906-494e768156f8 -> ../../sdag1
lrwxrwxrwx. 1 root root 11 июл 14 2023 f63966a0-60de-460a-a8c1-5c6ac9517402 -> ../../sda1
lrwxrwxrwx. 1 root root 10 июл 14 2023 f8713142-e007-499b-a50c-0e60fa89526b -> ../../sdn1
lrwxrwxrwx. 1 root root 11 июл 14 2023 fbb1b000-7012-4551-80e4-0183b5444cd4 -> ../../sda1
lrwxrwxrwx. 1 root root 11 июл 14 2023 fc8a3557-da3b-4c86-8fa7-436c4040966f -> ../../sda1
```

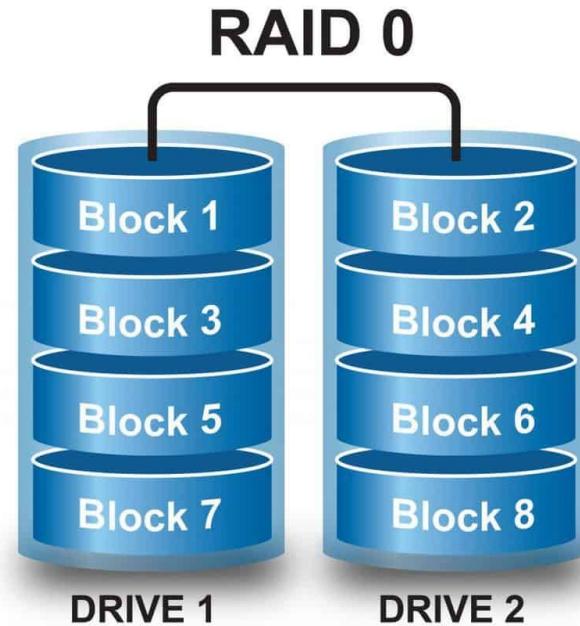
Дисковая запущанность

```
LABEL=srvk2835_0 /mnt/hadoop/0 auto defaults 0 0
LABEL=srvk2835_1 /mnt/hadoop/1 auto defaults 0 0
LABEL=srvk2835_2 /mnt/hadoop/2 auto defaults 0 0
LABEL=srvk2835_3 /mnt/hadoop/3 auto defaults 0 0
LABEL=srvk2835_4 /mnt/hadoop/4 auto defaults 0 0
LABEL=srvk2835_5 /mnt/hadoop/5 auto defaults 0 0
LABEL=srvk2835_6 /mnt/hadoop/6 auto defaults 0 0
LABEL=srvk2835_7 /mnt/hadoop/7 auto defaults 0 0
LABEL=srvk2835_8 /mnt/hadoop/8 auto defaults 0 0
LABEL=srvk2835_9 /mnt/hadoop/9 auto defaults 0 0
LABEL=srvk2835_10 /mnt/hadoop/10 auto defaults 0 0
LABEL=srvk2835_11 /mnt/hadoop/11 auto defaults 0 0
LABEL=srvk2835_12 /mnt/hadoop/12 auto defaults 0 0
LABEL=srvk2835_13 /mnt/hadoop/13 auto defaults 0 0
LABEL=srvk2835_14 /mnt/hadoop/14 auto defaults 0 0
LABEL=srvk2835_15 /mnt/hadoop/15 xfs nofail 0 0
LABEL=srvk2835_16 /mnt/hadoop/16 auto defaults 0 0
LABEL=srvk2835_17 /mnt/hadoop/17 auto defaults 0 0
LABEL=srvk2835_18 /mnt/hadoop/18 auto defaults 0 0
LABEL=srvk2835_19 /mnt/hadoop/19 auto defaults 0 0
LABEL=srvk2835_20 /mnt/hadoop/20 auto defaults 0 0
LABEL=srvk2835_21 /mnt/hadoop/21 auto defaults 0 0
LABEL=srvk2835_22 /mnt/hadoop/22 auto defaults 0 0
LABEL=srvk2835_23 /mnt/hadoop/23 auto defaults 0 0
LABEL=srvk2835_24 /mnt/hadoop/24 auto defaults 0 0
LABEL=srvk2835_25 /mnt/hadoop/25 auto defaults 0 0
LABEL=srvk2835_26 /mnt/hadoop/26 auto defaults 0 0
LABEL=srvk2835_27 /mnt/hadoop/27 auto defaults 0 0
LABEL=srvk2835_28 /mnt/hadoop/28 auto defaults 0 0
LABEL=srvk2835_29 /mnt/hadoop/29 auto defaults 0 0
LABEL=srvk2835_30 /mnt/hadoop/30 auto defaults 0 0
LABEL=srvk2835_31 /mnt/hadoop/31 auto defaults 0 0
LABEL=srvk2835_32 /mnt/hadoop/32 auto defaults 0 0
LABEL=srvk2835_33 /mnt/hadoop/33 auto defaults 0 0
LABEL=srvk2835_34 /mnt/hadoop/34 auto defaults 0 0
LABEL=srvk2835_35 /mnt/hadoop/35 auto defaults 0 0
```

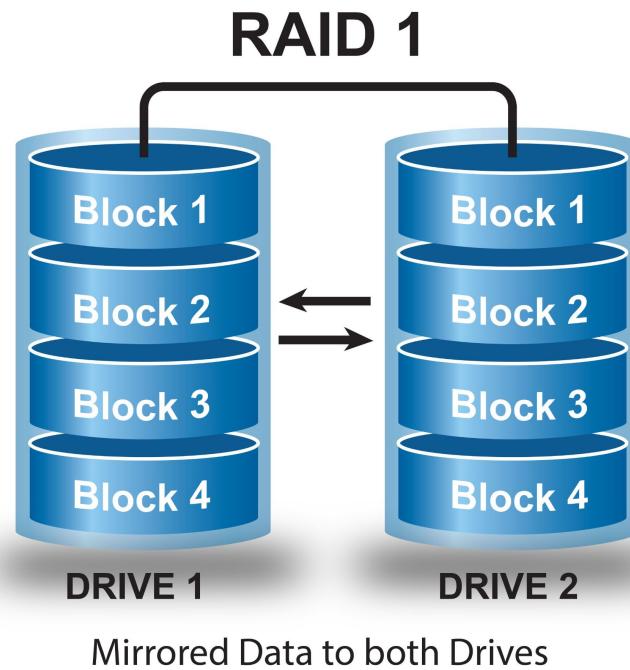
Дисковая подсистема: RAID



Дисковая подсистема: RAID

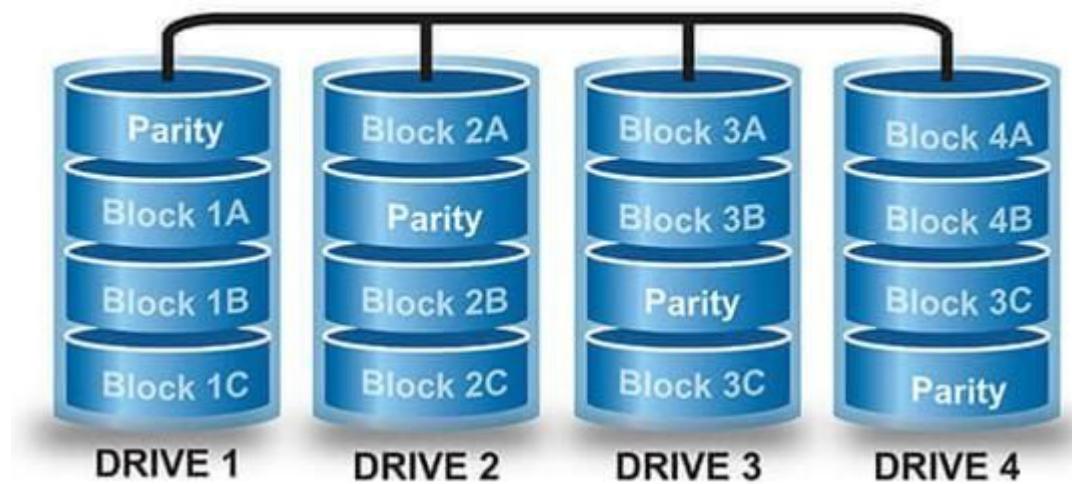


Дисковая подсистема: RAID

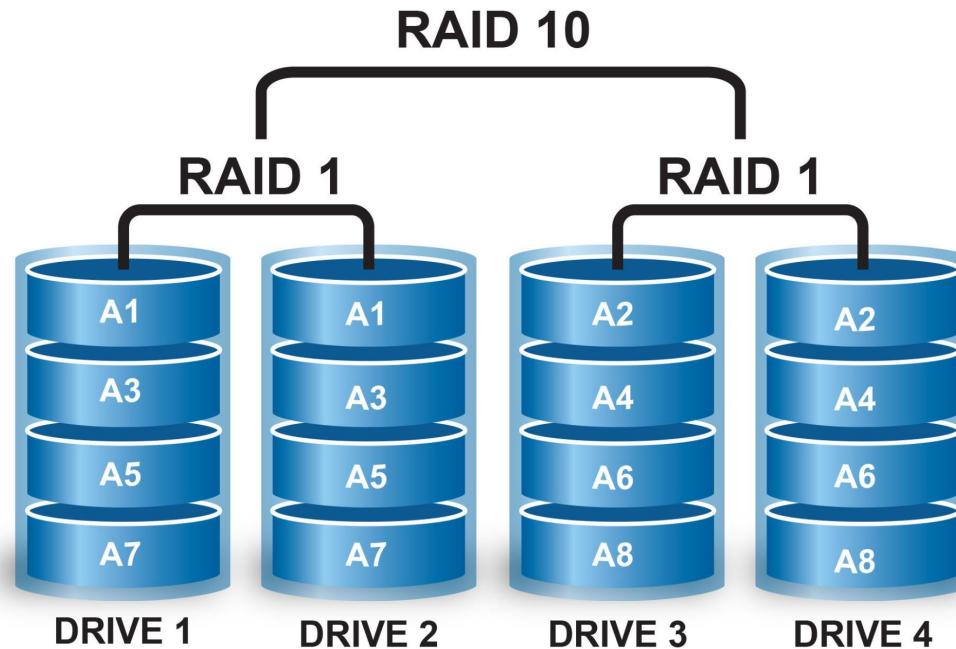


Дисковая подсистема: RAID

RAID 5



Дисковая подсистема: RAID



Дисковая подсистема: RAID

Аппаратный или программный?

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

```
# parted /dev/vdc set 1 raid on
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

```
# parted /dev/vdc set 1 raid on
```

```
# parted /dev/vdd set 1 raid on
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

```
# parted /dev/vdc set 1 raid on
```

```
# parted /dev/vdd set 1 raid on
```

```
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

```
# parted /dev/vdc set 1 raid on
```

```
# parted /dev/vdd set 1 raid on
```

```
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1
```

```
# cat /proc/mdstat
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
```

```
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
```

```
# parted /dev/vdc set 1 raid on
```

```
# parted /dev/vdd set 1 raid on
```

```
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1
```

```
# cat /proc/mdstat
```

```
# mdadm -D /dev/md0
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%  
  
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%  
  
# parted /dev/vdc set 1 raid on  
  
# parted /dev/vdd set 1 raid on  
  
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1  
  
# cat /proc/mdstat  
  
# mdadm -D /dev/md0  
  
# echo 5000 > /sys/block/md0/md/sync_speed_max
```

Дисковая подсистема: RAID

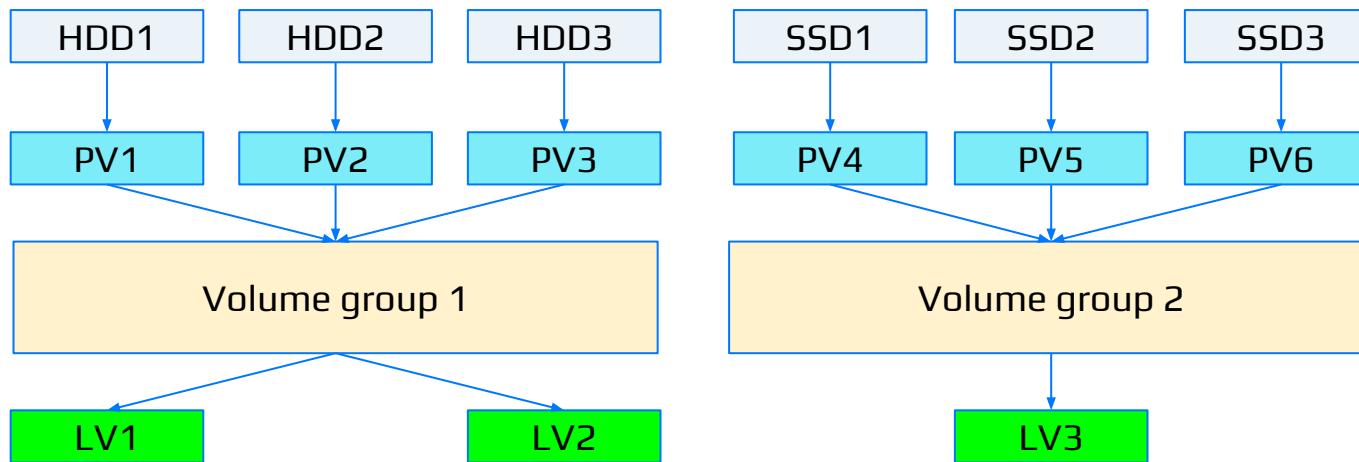
```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%  
  
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%  
  
# parted /dev/vdc set 1 raid on  
  
# parted /dev/vdd set 1 raid on  
  
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1  
  
# cat /proc/mdstat  
  
# mdadm -D /dev/md0  
  
# echo 5000 > /sys/block/md0/md/sync_speed_max  
  
# mkdir /etc/mdadm/
```

Дисковая подсистема: RAID

```
# parted /dev/vdc mklabel gpt && parted /dev/vdc mkpart primary 1024KiB 100%
# parted /dev/vdd mklabel gpt && parted /dev/vdd mkpart primary 1024KiB 100%
# parted /dev/vdc set 1 raid on
# parted /dev/vdd set 1 raid on
# mdadm --create /dev/md0 --level=1 --raid-disks=2 /dev/vdc1 /dev/vdd1
# cat /proc/mdstat
# mdadm -D /dev/md0
# echo 5000 > /sys/block/md0/md/sync_speed_max
# mkdir /etc/mdadm/
# mdadm --detail --sca > /etc/mdadm/mdadm.conf
```

Дисковая подсистема: LVM

LVM - Logical Volume Manager



Дисковая подсистема: LVM

```
# pvcreate /dev/vdb  
  
# pvs  
  
# pvdisplay  
  
# vgcreate vg1 /dev/vdb /dev/vdc  
  
# vgs  
  
# vgdisplay  
  
# lvcreate -L 1g vg1 -n lv1  
  
# lvcreate -L 10G -n lvraid1 --type raid1 vg1
```

Дисковая подсистема: LVM в живой природе

Volumes (19)

	Ns	Shard	Volume	Type	State	Devices	Capacity	Used	%	% Io.~
	odkl	messaging-conversation.db.messaging.prod...	data	ssd	MOUNTED	/dev/sdn	894.3G	448G	50%	8%
	odkl	druid.batch/historical-hot/86	druid-ssd1	ssd	MOUNTED	/dev/sdl	880G	534G	61%	
	odkl	druid.batch/historical-hot/86	druid-ssd2	ssd	MOUNTED	/dev/sdm	880G	532.2G	60%	
	odkl	video-history.db.video.prod/cdb/64	data	ssd	MOUNTED	/dev/sdk	700G	309.7G	44%	8%
	odkl	obs-s3-misc.db.s3.prod/obs/14	data	ssd	MOUNTED	/dev/sdd,/dev/s...	4097G	895.7G	22%	2%
	odkl	obs-s3-misc.db.s3.prod/obs/14	system	ssd	MOUNTED	/dev/sdh	1024M	62.51M	6%	
	odkl	ids-storage.db.ids.prod/cdb/21	data	ssd	MOUNTED	/dev/sdh	200G	18.35G	9%	
	odkl	obs-s3-test.db.s3.prod/obs/10	data	hdd	MOUNTED	/dev/sdc	10T	1985G	19%	
	odkl	obs-s3-test.db.s3.prod/obs/10	system	hdd	MOUNTED	/dev/sdc	1024M	56.5M	6%	
	odkl	messaging-conversation.db.messaging.prod...	commitlog	hdd	MOUNTED	/dev/sde	50G	30.41G	61%	
	odkl	messaging-conversation.db.messaging.prod...	arch	hdd	MOUNTED	/dev/sde	2685G	1818G	68%	
	odkl	video-history.db.video.prod/cdb/64	arch	hdd	MOUNTED	/dev/sde	1434G	574G	40%	
	odkl	video-history.db.video.prod/cdb/64	commitlog	hdd	MOUNTED	/dev/sde	350G	138.4G	40%	
	dzen	telego-vh.app.production.vh.prod/telego/34	telego-data	hdd	MOUNTED	/dev/sde	300G	109.4G	36%	
	odkl	ids-storage.db.ids.prod/cdb/21	arch	hdd	MOUNTED	/dev/sde	500G	280.1G	56%	
	odkl	ids-storage.db.ids.prod/cdb/21	commitlog	hdd	MOUNTED	/dev/sde	100G	71.4G	71%	
	odkl	ids-storage.db.ids.prod/cdb/21	memory	memdisk	MOUNTED	hugetlfs0	50G	49G	98%	
	odkl	video-history.db.video.prod/cdb/64	memory	memdisk	MOUNTED	hugetlfs1	40G	31.79G	79%	
	odkl	messaging-conversation.db.messaging.prod...	memory	memdisk	MOUNTED	hugetlfs1	3072M	3072M	100%	

Дисковая подсистема: LVM в живой природе

```
[root@srvd4821 ~]# pvs
PV          VG      Fmt  Attr PSize   PFree
/dev/md0    srvd4821 lvm2 a--  447,00g     0
/dev/sdc    cloud.hdd lvm2 a--  12,73t   2,73t
/dev/sdd    cloud.ssd lvm2 a--  894,25g  75,00g
/dev/sde    cloud.hdd lvm2 a--  12,73t   7,44t
/dev/sdf    cloud.ssd lvm2 a--  894,25g  75,00g
/dev/sdg    cloud.ssd lvm2 a--  894,25g  75,00g
/dev/sdh    cloud.ssd lvm2 a--  894,25g  693,25g
/dev/sdi    cloud.ssd lvm2 a--  894,25g  75,00g
/dev/sdj    cloud.ssd lvm2 a--  894,25g  75,00g
/dev/sdk    cloud.ssd lvm2 a--  894,25g  194,25g
/dev/sdl    cloud.ssd lvm2 a--  894,25g  14,25g
/dev/sdm    cloud.ssd lvm2 a--  894,25g  14,25g
/dev/sdn    cloud.ssd lvm2 a--  894,25g     0
```

Дисковая подсистема: LVM в живой природе

```
[root@srvd4821 ~]# vgs
  VG          #PV #LV #SN Attr   VSize   VFree
cloud.hdd    2    9    0 wz--n- <25,47t 10,17t
cloud(ssd)  10    7    0 wz--n-  8,73t  1,26t
srvd4821     1    4    0 wz--n- 447,00g      0
```

Дисковая подсистема: LVM в живой природе

```
[root@srvd4821 ~]# lvs
```

LV	VG	Attr	LSize
04706c8f1f0c11efac2f0d998a6bf779	cloud.hdd	-wi-ao---	10,00t
04706c901f0c11efac2f0d998a6bf779	cloud.hdd	-wi-ao---	1,00g
5631f4841e8611efac2f0d998a6bf779	cloud.hdd	-wi-ao---	50,00g
5631f4861e8611efac2f0d998a6bf779	cloud.hdd	-wi-ao---	2,62t
a62a19cece4b11ee9a469b1040addba3	cloud.hdd	-wi-ao---	1,40t
a62a19cfce4b11ee9a469b1040addba3	cloud.hdd	-wi-ao---	350,00g
cf162fac6dc411efa07c2be31a00fdd8	cloud.hdd	-wi-ao---	300,00g
d26ffce14d9211eaae4347151elf0460	cloud.hdd	-wi-ao---	500,00g
d26ffce24d9211eaae4347151elf0460	cloud.hdd	-wi-ao---	100,00g
5631f4851e8611efac2f0d998a6bf779	cloud.ssd	-wi-ao---	894,25g
93889a20fbcb11ee888a0d998a6bf779	cloud.ssd	-wi-ao---	880,00g
93889a21fbcb11ee888a0d998a6bf779	cloud.ssd	-wi-ao---	880,00g
a62a19cdce4b11ee9a469b1040addba3	cloud.ssd	-wi-ao---	700,00g
b9b3d6807fd011eeb1f59b1040addba3	cloud.ssd	-wi-ao---	4,00t
b9b3d6817fd011eeb1f59b1040addba3	cloud.ssd	-wi-ao---	1,00g
d26ffce34d9211eaae4347151elf0460	cloud.ssd	-wi-ao---	200,00g
containers	srvd4821	-wi-ao---	<165,69g
one	srvd4821	-wi-ao---	<234,43g
root	srvd4821	-wi-ao---	<42,89g
swap	srvd4821	-wi-a----	4,00g

Дисковая подсистема: LVM в живой природе

```
[root@srvd4821 ~]# df -h
Файловая система      Размер Использовано  Дост Использовано% Смонтировано в
/devtmpfs                126G      0        126G      0% /dev
tmpfs                     126G     96K      126G      1% /dev/shm
tmpfs                     126G     5,9M     126G      1% /run
tmpfs                     126G      0        126G      0% /sys/fs/cgroup
/dev/mapper/srvd4821-root      43G     23G      21G      52% /
/dev/mapper/srvd4821-one      235G    3,7G     231G      2% /one
/dev/mapper/srvd4821-containers 166G    59G     107G      36% /var/lib/containers
/dev/mapper/cloud.ssd-a62a19cdce4b11ee9a469b1040addba3 700G   310G     390G      45% /run/miniond/storage/video-history.db.video.prod/cdb/64/data#a62a19cdce4b11ee9a469b1040addba3
/dev/mapper/cloud.ssd-b9b3d6807fd011eeb1f59b1040addba3 4,0T   896G     3,2T      22% /run/miniond/storage/obs-s3-misc.db.s3.prod/obs/14/data#b9b3d6807fd011eeb1f59b1040addba3
/dev/mapper/cloud.ssd-b9b3d6817fd011eeb1f59b1040addba3 1014M   63M     952M      7% /run/miniond/storage/obs-s3-misc.db.s3.prod/obs/14/system#b9b3d6817fd011eeb1f59b1040addba3
/dev/mapper/cloud.ssd-d26ffce34d9211eaae4347151elf0460 200G   19G     182G      10% /run/miniond/storage/ids-storage.db.ids.prod/cdb/21/data#d26ffce34d9211eaae4347151elf0460
/dev/mapper/cloud.hdd-a62a19cece4b11ee9a469b1040addba3 1,4T   574G     859G      41% /run/miniond/storage/video-history.db.video.prod/cdb/64/arch#a62a19cece4b11ee9a469b1040addba3
/dev/mapper/cloud.hdd-a62a19cfce4b11ee9a469b1040addba3 350G   139G     212G      40% /run/miniond/storage/video-history.db.video.prod/cdb/64/commitlog#a62a19cfce4b11ee9a469b1040addba3
/dev/mapper/cloud.hdd-d26ffce14d9211eaae4347151elf0460 500G   281G     220G      57% /run/miniond/storage/ids-storage.db.ids.prod/cdb/21/arch#d26ffce14d9211eaae4347151elf0460
/dev/mapper/cloud.hdd-d26ffce24d9211eaae4347151elf0460 100G   72G      29G      72% /run/miniond/storage/ids-storage.db.ids.prod/cdb/21/commitlog#d26ffce24d9211eaae4347151elf0460
none                      50G   49G      1,0G      98% /run/miniond/storage/ids-storage.db.ids.prod/cdb/21/memory#b7deeeaa0e6b611ee8c209b1040addba3
none                      40G   32G      8,3G      80% /run/miniond/storage/video-history.db.video.prod/cdb/64/memory#bbcrafle0e6b611ee8c209b1040addba3
none                      880G   534G     346G      61% /run/miniond/storage/druid.batch/historical-hot/86/druid-ssd1#93889a20fbcb11ee888a0d998a6bf779
none                      880G   534G     347G      61% /run/miniond/storage/druid.batch/historical-hot/86/druid-ssd2#93889a21fbcb11ee888a0d998a6bf779
none                      3,0G   3,0G      0       100% /run/miniond/storage/messaging-conversation.db.messaging.prod/cdb/64/memory#56380ea01e8611efac2f0d998a6bf779
none                      50G   31G      20G      61% /run/miniond/storage/messaging-conversation.db.messaging.prod/cdb/64/commitlog#5631f4841e8611efac2f0d998a6bf779
none                      894G   448G     446G      51% /run/miniond/storage/messaging-conversation.db.messaging.prod/cdb/64/data#5631f4851e8611efac2f0d998a6bf779
```

Дисковая подсистема: LVM в живой природе

```
[root@srvd4821 ~]# ls /dev/mapper/
cloud.hdd-04706c8f1f0c11efac2f0d998a6bf779  cloud.hdd-cf162fac6dc411efa07c2be31a00fdd8  cloud.ssd-a62a19cdce4b11ee9a469b1040addba3  srvd4821-one
cloud.hdd-04706c901f0c11efac2f0d998a6bf779  cloud.hdd-d26ffce14d9211eaae4347151elf0460  cloud.ssd-b9b3d6807fd011eeb1f59b1040addba3  srvd4821-root
cloud.hdd-5631f4841e8611efac2f0d998a6bf779  cloud.hdd-d26ffce24d9211eaae4347151elf0460  cloud.ssd-b9b3d6817fd011eeb1f59b1040addba3  srvd4821-swap
cloud.hdd-5631f4861e8611efac2f0d998a6bf779  cloud.ssd-5631f4851e8611efac2f0d998a6bf779  cloud.ssd-d26ffce34d9211eaae4347151elf0460
cloud.hdd-a62a19cece4b11ee9a469b1040addba3  cloud.ssd-93889a20fbcb11ee888a0d998a6bf779  control
cloud.hdd-a62a19cfce4b11ee9a469b1040addba3  cloud.ssd-93889a21fbcb11ee888a0d998a6bf779  srvd4821-containers
```

Дисковая подсистема: планировщики I/O

- **CFQ** - Complete Fairness Queueing. Он создает отдельную I/O очередь для каждого процесса и разделяет время между ними равномерно.
- **NOOP** (он же NONE) - реализует принцип First In First Out, т.е. этот планировщик не выполняет никакой оптимизации.
- **DEADLINE** - каждый запрос получает таймстамп, являющийся дедлайном, при достижении которого запрос получает наивысший приоритет.

Дисковая подсистема: планировщики I/O

Multi-Queue Block IO Queueing Mechanism (blk-mq)

- **mq-deadline** - реализация deadline для blk-mq.
- **BFQ** - Budget Fair Queueing. BFQ гарантирует низкое время отклика для требовательных к этому приложений, либо же высокую пропускную способность. Имеет довольно много настроек.
- **kyber** - еще один планировщик для работы с быстрыми устройствами. Kyber использует отдельные очереди для чтения и записи и отдает приоритет запросам на чтение. Эффективен для NVME, SSD.

Дисковая подсистема: планировщики I/O

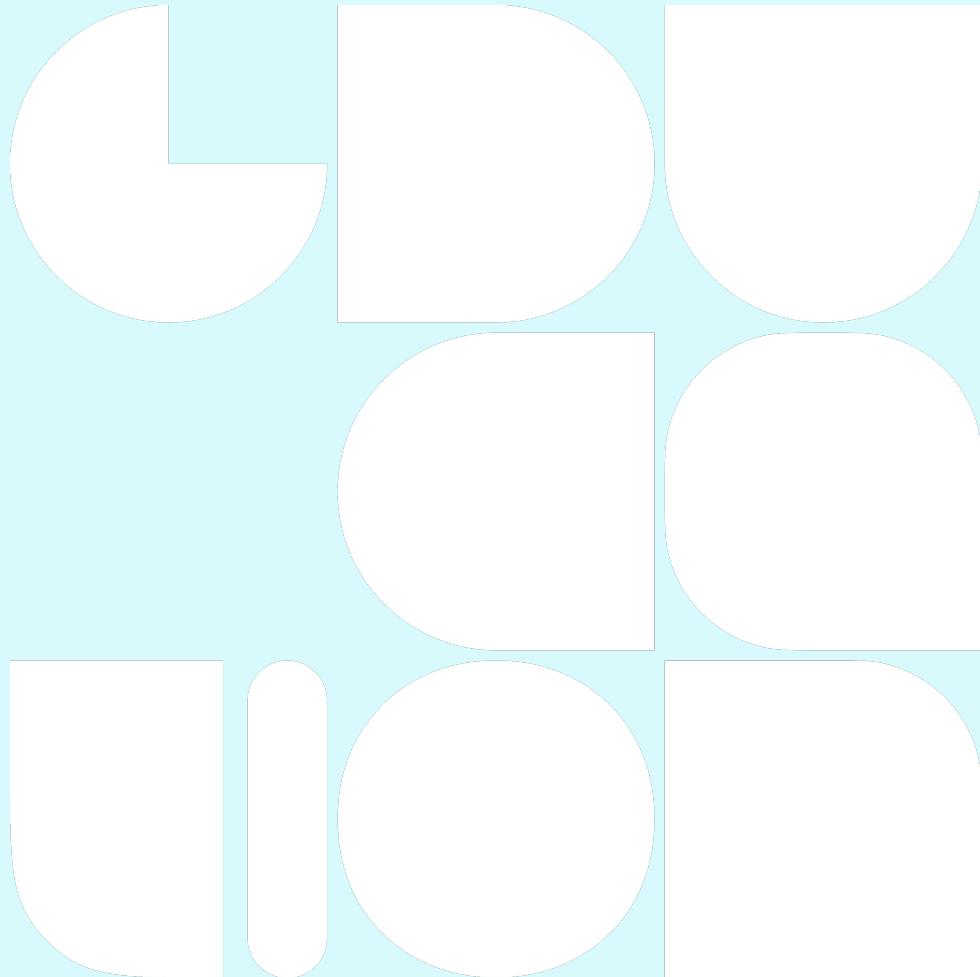
```
# cat /sys/block/vdb/queue/scheduler  
  
# echo bfq > /sys/block/vdb/queue/scheduler  
  
/etc/default/grub:  
GRUB_CMDLINE_LINUX="<текущее значение> scsi_mod.use_blk_mq=1 elevator=mq-deadline"  
  
scsi_mod.use_blk_mq=1 требуется только если мы хотим использовать механизм blk-mq  
  
# grub2-mkconfig -o /boot/grub2/grub.cfg  
  
# reboot  
  
/etc/udev/rules.d/60-io-scheduler.rules:  
KERNEL=="vdb*", ATTR{queue/scheduler}="bfq", GOTO="scheduler_end"
```



Вопросы?



Cgroups



Cgroups

Cgroups (control groups) - это фича ядра линукс, которая позволяет изолировать, учитывать, приоритезировать и ограничивать использование ресурсов (CPU, память, disk IO итд) для определенного набора процессов

Cgroups: v1

/sys/fs/cgroup/

Cgroups: v1

/sys/fs/cgroup/

/sys/fs/cgroup/cpu/my_test_group/

Cgroups: v1

/sys/fs/cgroup/

/sys/fs/cgroup/cpu/my_test_group/

/sys/fs/cgroup/memory/my_test_group/

Cgroups: v1

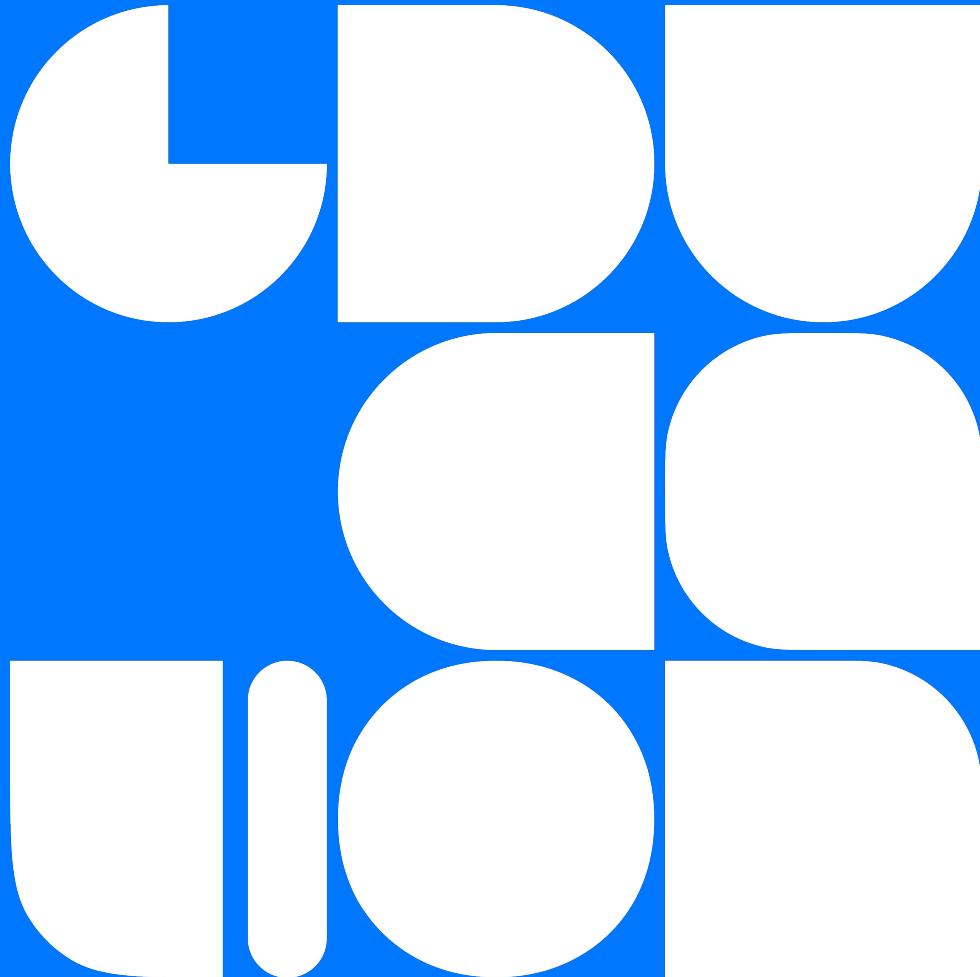
/sys/fs/cgroup/

/sys/fs/cgroup/cpu/my_test_group/

/sys/fs/cgroup/memory/my_test_group/

Для того, чтобы поместить процесс в контрольную группу, нужно добавить его PID в файл cgroup.procs в каждом нужном нам контроллере в группе my_test_group.

Практика



Cgroups: v1

Ограничеваем потребление памяти

```
# mkdir /sys/fs/cgroup/memory/my_test_group  
  
# echo "2097152" > /sys/fs/cgroup/memory/my_test_group/memory.limit_in_bytes  
  
# /sys/fs/cgroup/memory/my_test_group/memory.usage_in_bytes  
  
# /sys/fs/cgroup/memory/my_test_group/memory.max_usage_in_bytes  
  
# /sys/fs/cgroup/memory/my_test_group/memory.oom_control
```

Cgroups: v1

Привязываем к ядрам сри

```
# mkdir /sys/fs/cgroup/cpuset/my_test_group  
  
# /sys/fs/cgroup/cpuset/my_test_group/cgroup.procs  
  
# /sys/fs/cgroup/cpuset/my_test_group/cpuset.cpus  
  
# /sys/fs/cgroup/cpuset/my_test_group/cpuset.mems
```

Cgroups: v1

Ограничиваем процессорное время

```
# mkdir /sys/fs/cgroup/cpu/my_test_group  
  
# /sys/fs/cgroup/cpu/my_test_group/cpu.cfs_quota_us  
  
# /sys/fs/cgroup/cpu/my_test_group/cpu.cfs_period_us  
  
# /sys/fs/cgroup/cpu/my_test_group/cpuacct.stat  
  
# /sys/fs/cgroup/cpu/my_test_group/cpuacct.usage_percpu
```

Cgroups: v1

Делим процессорное время пропорционально

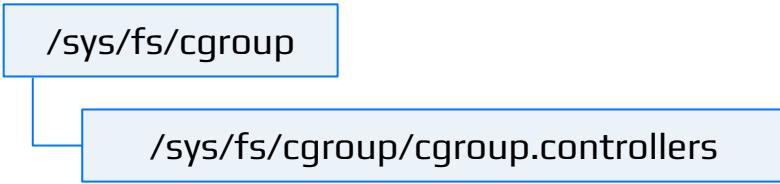
```
# mkdir /sys/fs/cgroup/cpu/my_test_group
```

```
# /sys/fs/cgroup/cpu/my_test_group/cpu.shares
```

Cgroups: v2

/sys/fs/cgroup

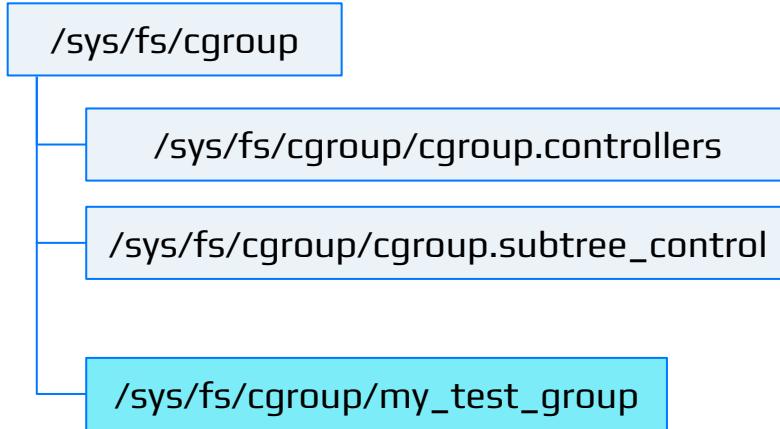
Cgroups: v2



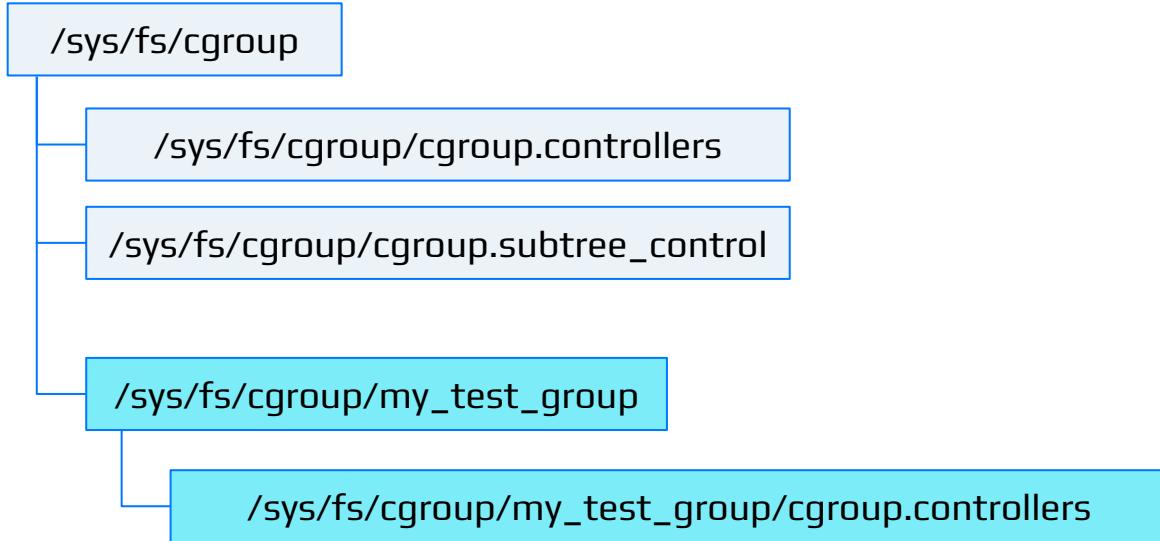
Cgroups: v2



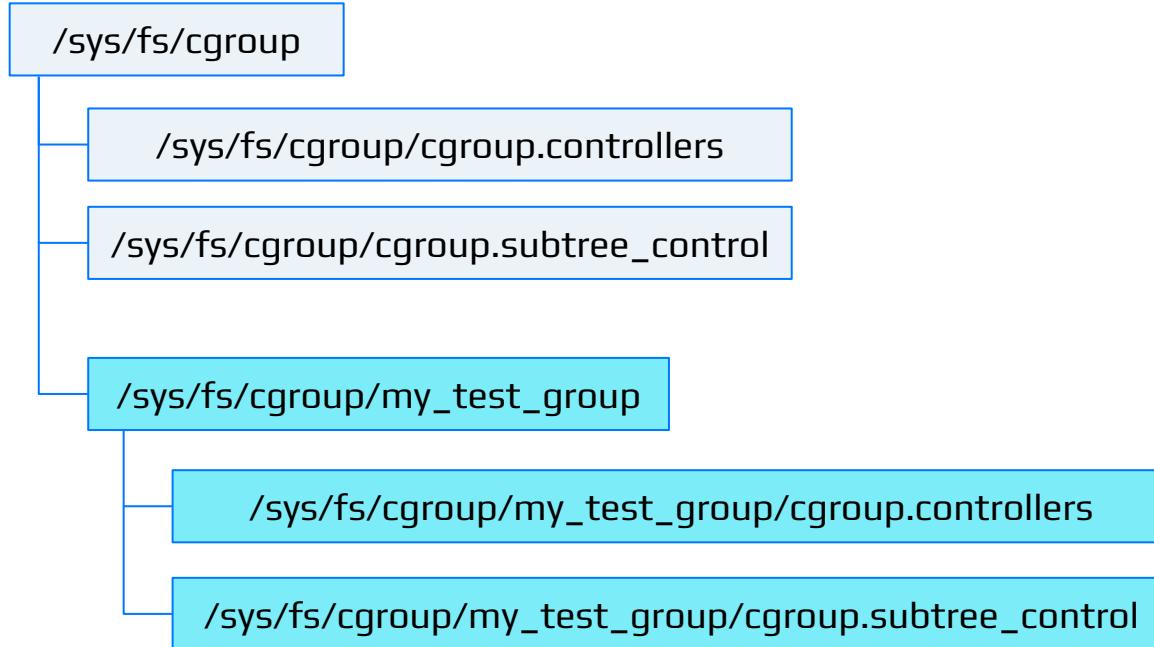
Cgroups: v2



Cgroups: v2



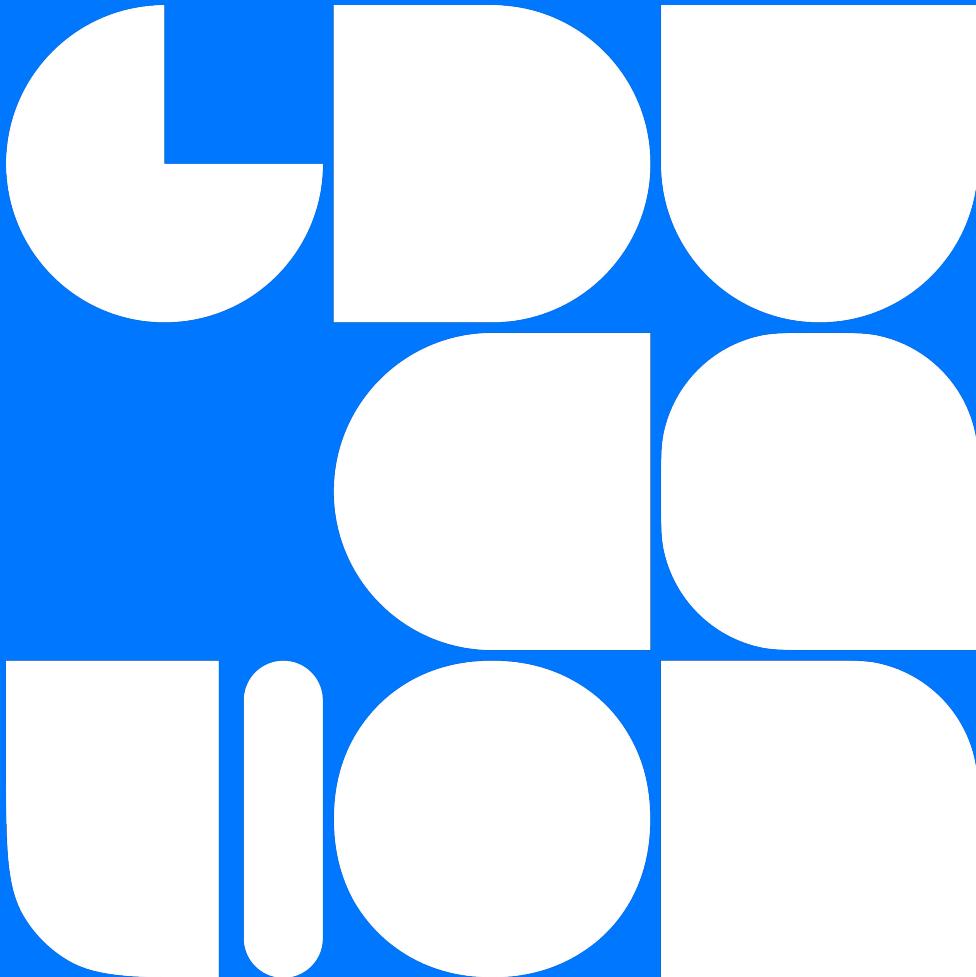
Cgroups: v2



Cgroups: v2

/sys/fs/cgroups/my_test_group/nested_group

Практика



Cgroups: v2

Ограничиваем потребление памяти

```
# mkdir /sys/fs/cgroup/my_test_group
```

```
# cat /sys/fs/cgroup/my_test_group/cgroup.controllers
```

```
# echo "2097152" > /sys/fs/cgroup/my_test_group/memory.max
```

Cgroups: v2

Привязываем к ядрам сри

```
# echo "+cpuset" > /sys/fs/cgroup/cgroup.subtree_control
```

```
# echo 1 > /sys/fs/cgroup/my_test_group/cpuset.cpus
```

```
# echo 0 > /sys/fs/cgroup/my_test_group/cpuset.mems
```

Cgroups: v2

Ограничиваем процессорное время

```
# echo "+cpu" > /sys/fs/cgroup/cgroup.subtree_control
```

```
# echo "50000 100000" > /sys/fs/cgroup/my_test_group/cpu.max
```

Cgroups: v2

Делим процессорное время пропорционально

```
# echo "+cpu" > /sys/fs/cgroup/cgroup.subtree_control
```

```
# echo "25000 100000" > /sys/fs/cgroup/my_test_group/cpu.max
```

Сведем все вместе: one-cloud

Класс изоляции	Пример alloc	Опции Docker run	sched_setscheduler chrt*
Prod	cpu = 4	--cpuquota=400000 --cpuperiod=100000	SCHED_OTHER
Batch	Cpu = [1, *)	--cpushares=1024	SCHED_BATCH
Idle	Cpu= [2, *)	--cpushares=2048	SCHED_IDLE

<https://habr.com/ru/companies/odnoklassniki/articles/346868/>

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

```
alloc:  
vcores: '12'  
mem: 40g  
lan_out: 2G  
lan_in: 2G
```

Лимитыcpu:

```
[root@srvd4821 pids-prod]# cat  
/sys/fs/cgroup/cpu/cloud/prod/64.cdb.video-history.dc-0aac59a0-6067-11ef-8813-9b1040addba3/pids-prod/cpu.cfs_period_us  
100000
```

```
[root@srvd4821 pids-prod]# cat  
/sys/fs/cgroup/cpu/cloud/prod/64.cdb.video-history.dc-0aac59a0-6067-11ef-8813-9b1040addba3/pids-prod/cpu.cfs_quota_us  
1090909
```

Лимиты памяти:

```
[root@srvd4821 pids-prod]# cat  
/sys/fs/cgroup/memory/cloud/prod/64.cdb.video-history.dc-0aac59a0-6067-11ef-8813-9b1040addba3/memory.limit_in_bytes  
42949672960
```

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

Проверяем приоритет планировщика:

```
[root@srvd4821]# cat  
/sys/fs/cgroup/cpu/cloud/prod/64.cdb.video-history.dc-0aac59a0-6067-11ef-8813-9b1040addba3/pids-batch/cgroup.procs  
41188
```

```
[root@srvd4821]# ps -c -p 41188  
 PID CLS PRI TTY      TIME CMD  
41188 TS  19 ?    74-10:17:00 java
```

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

DC instance: 64.cdb.video-history.dc.odkl.ru (64.cdb.video-history.db.video.prod.dc)

[ACL](#) | [Statistics](#) | [Statshouse](#) | [Monitor](#) | [Flows](#) | [Logs](#) | [Links+](#) | [Realtime stats](#) | [JMap](#) | [JStack](#) | [Profile](#) | [Profile History](#) | [Fit mi](#)

state	RUNNING RESERVED*1.50
started	2024-08-22 12:16:07 (1 month 1 week ago)
reported	<i>last: 2024-10-06 11:59:54.729, mean time between: 9210.0ms, phi=0.39 (max 1.0)</i>
service	cdb.video-history
minion	srvd4821/10.4.30.176 numa node1
storages	video-history.db.video.prod/cdb/64 MOUNTED

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

Проверяем NUMA ноду:

```
[root@srvd4821]# lscpu | grep "NUMA node1"
NUMA node1 CPU(s): 20-39,60-79
```

```
[root@srvd4821]# grep -P "Cpus_allowed_list|Mems_allowed_list" /proc/41188/status
Cpus_allowed_list: 20-39,60-79
Mems_allowed_list: 1
```

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

Volumes

	Shard	Volume	Type	State	Devices	Capacity	Used	%	% io..	Message
✓	video-history.db.video.prod/cdb/64	arch	hdd	MOUNTED	/dev/sde	1434G	316.2G	22%		
✓	video-history.db.video.prod/cdb/64	commitlog	hdd	MOUNTED	/dev/sde	350G	121.9G	35%		
✓	video-history.db.video.prod/cdb/64	data	ssd	MOUNTED	/dev/sdk	700G	304.8G	44%	7%	
✓	video-history.db.video.prod/cdb/64	memory	memdisk	MOUNTED	hugetlbfs1	40G	31.79G	79%		
Total Items: 4										

Total Items: 4

```
[root@srvd4821 ~]# df -h | grep video-history
/dev/mapper/cloud.ssd-a62a19cdce4b11ee9a469b1040addba3 700G      305G  395G          44% /run/miniond/storage/video-history.db.video.prod/cdb/64/data#a62a19cdce4b11ee9a469b1040addba3
/dev/mapper/cloud.hdd-a62a19cece4b11ee9a469b1040addba3 1,4T      317G  1,1T          23% /run/miniond/storage/video-history.db.video.prod/cdb/64/arch#a62a19cece4b11ee9a469b1040addba3
/dev/mapper/cloud.hdd-a62a19cfce4b11ee9a469b1040addba3 350G      123G  228G          35% /run/miniond/storage/video-history.db.video.prod/cdb/64/commitlog#a62a19cfce4b11ee9a469b1040addba3
none                                         40G      32G   8,3G          80% /run/miniond/storage/video-history.db.video.prod/cdb/64/memory#bbcfile0e6b611ee8c209b1040addba3
```

Сведем все вместе: one-cloud

64.cdb.video-history.db.video.prod.dc:

```
[root@srvd4821 ~]# lvs | grep -P  
"a62a19cdce4b11ee9a469b1040addba3|a62a19cece4b11ee9a469b1040addba3|a62a19cfce4b11ee9a469b1040addba3"  
  
a62a19cece4b11ee9a469b1040addba3 cloud.hdd -wi-ao---- 1,40t  
a62a19cfce4b11ee9a469b1040addba3 cloud.hdd -wi-ao---- 350,00g  
a62a19cdce4b11ee9a469b1040addba3 cloud.ssd -wi-ao---- 700,00g
```

Сведем все вместе: one-cloud

86.historical-hot.druid.batch.dc:

```
alloc:  
vcores: '2'  
mem: 40g  
lan_out: 100M  
lan_in: 100M  
...
```

Лимиты CPU:

```
[root@srvd4821]# cat  
/sys/fs/cgroup/cpu/cloud/nonprod/node0/batch/86.historical-hot.druid.dc-a4a5736a-70e6-11ef-9d68-9b1040addba3/cpu.shares  
1818
```

Лимиты памяти:

```
[root@srvd4821]# cat  
/sys/fs/cgroup/memory/cloud/nonprod/node0/batch/86.historical-hot.druid.dc-a4a5736a-70e6-11ef-9d68-9b1040addba3/me  
mory.limit_in_bytes  
42949672960
```

Сведем все вместе: one-cloud

86.historical-hot.druid.batch.dc:

Проверяем приоритет планировщика:

```
[root@srvd4821]# cat  
/sys/fs/cgroup/cpu/cloud/nonprod/node0/batch/86.historical-hot.druid.dc-a4a5736a-70e6-11ef-9d68-9b1040addba3/pids-batch/libpo  
d-48f3c4a11e16d1316c99ab42c07287122fde6a7c640deaf1d5ffbec9cba9c82f/cgroup.procs
```

```
10209  
10246  
10248  
10251  
11071  
11120  
11217
```

```
[root@srvd4821]# ps -c -p 11217  
 PID CLS PRI TTY      TIME CMD  
11217 B  0 ?    6-15:02:31 java
```

Сведем все вместе: one-cloud

Instances (260)

	Id	State	Availability	Minion ..	Lan	Cpu	% Cpu
	1	RUNNING	RESERVED*2.00	srvr933	10.82.182.84:1...	58.46	731%
	2	RUNNING	RESERVED*2.00	srvr701	10.82.182.85:1...	25.11	314%
	3	RUNNING	RESERVED*2.00	srvr10748	10.82.182.86:1...	39.84	498%
	4	RUNNING	RESERVED*2.00	srvr846	10.82.182.87:1...	58.87	736%
	5	RUNNING	RESERVED*2.00	srvr419	10.82.182.88:1...	42.49	531%
	6	RUNNING	RESERVED*2.00	srvr972	10.82.182.89:1...	25.92	324%
	7	RUNNING	RESERVED*2.00	srvr838	10.82.182.90:1...	28.75	359%
	8	RUNNING	RESERVED*2.00	srvr932	10.82.182.91:1...	42.7	534%
	9	RUNNING	RESERVED*2.00	srvr707	10.82.182.92:1...	57.26	716%
	10	RUNNING	RESERVED*2.00	srvr874	10.82.182.93:1...	57.93	724%
	11	RUNNING	RESERVED*2.00	srvr587	10.82.182.94:1...	37.05	463%



Вопросы?



Подведём итоги:

1. Компьютеры - это сложно и дорого)
2. Не забывайте про NUMA - это может сильно отразиться на производительности
3. С умом подходите к выбору дисков для задачи - не всегда вам нужен самый дорогой и быстрый
4. Cgroups - удобный и простой способ управлять ресурсами - используйте его!

Спасибо за
внимание!

