SSD v serveru

Pavel Šnajdr
InstallFest 2015

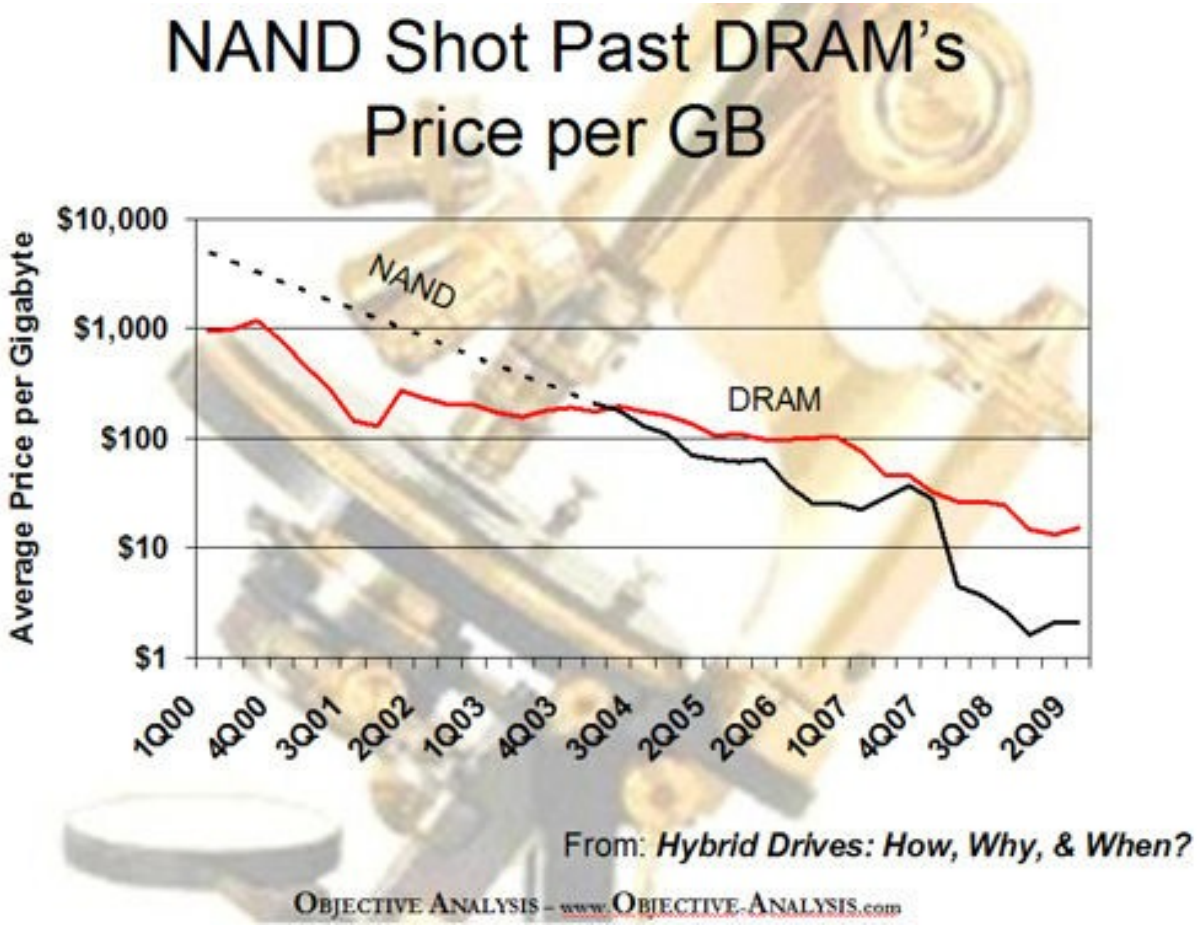
Obsah

- Historie moderních SSD
- Jak funguje SSD
- Vybíráme SSD
- SSD v serveru pod Linuxem
- Vlastní zkušenost
- Diskuze

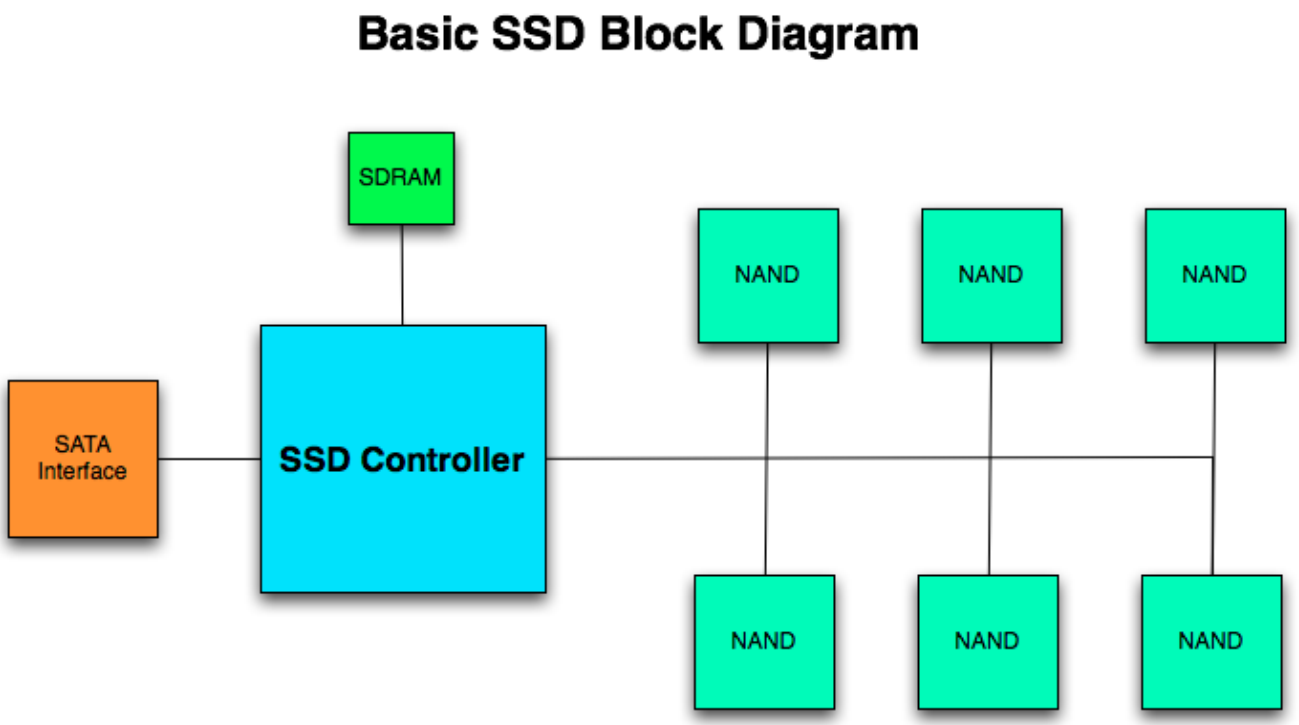
Historie moderních SSD

- SSD založené na NAND flash
  - 1989 Toshiba
  - Průkopníkem v oboru SanDisk, první SSD v r. 1991
  - 1995 Simple Technology začíná vyrábět SSD, v r. 2000 se přejmenovává na sTec, Inc., 2007 $1B
  - 1999 První 3.5” vlaštovky
  - 2002 Založení OCZ
  - 2008 Intel X25-M, X25-E,
  - 2009 gen. 2, o 60% levnější

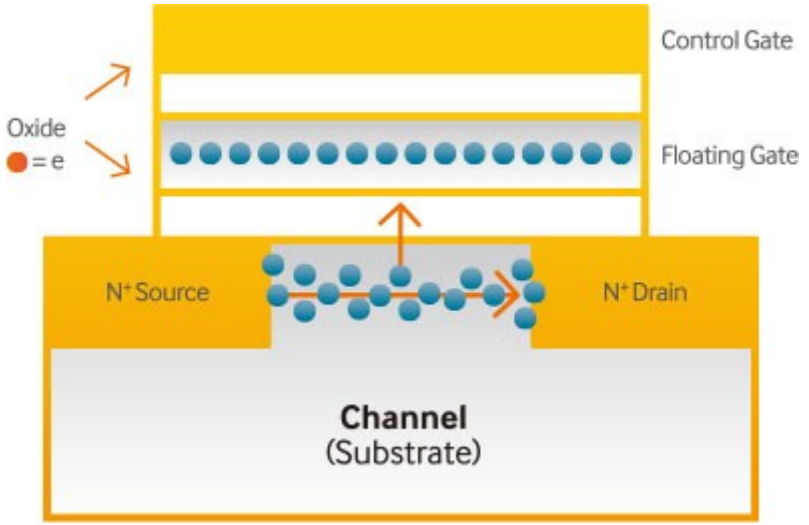
Historie moderních SSD



Jak funguje SSD



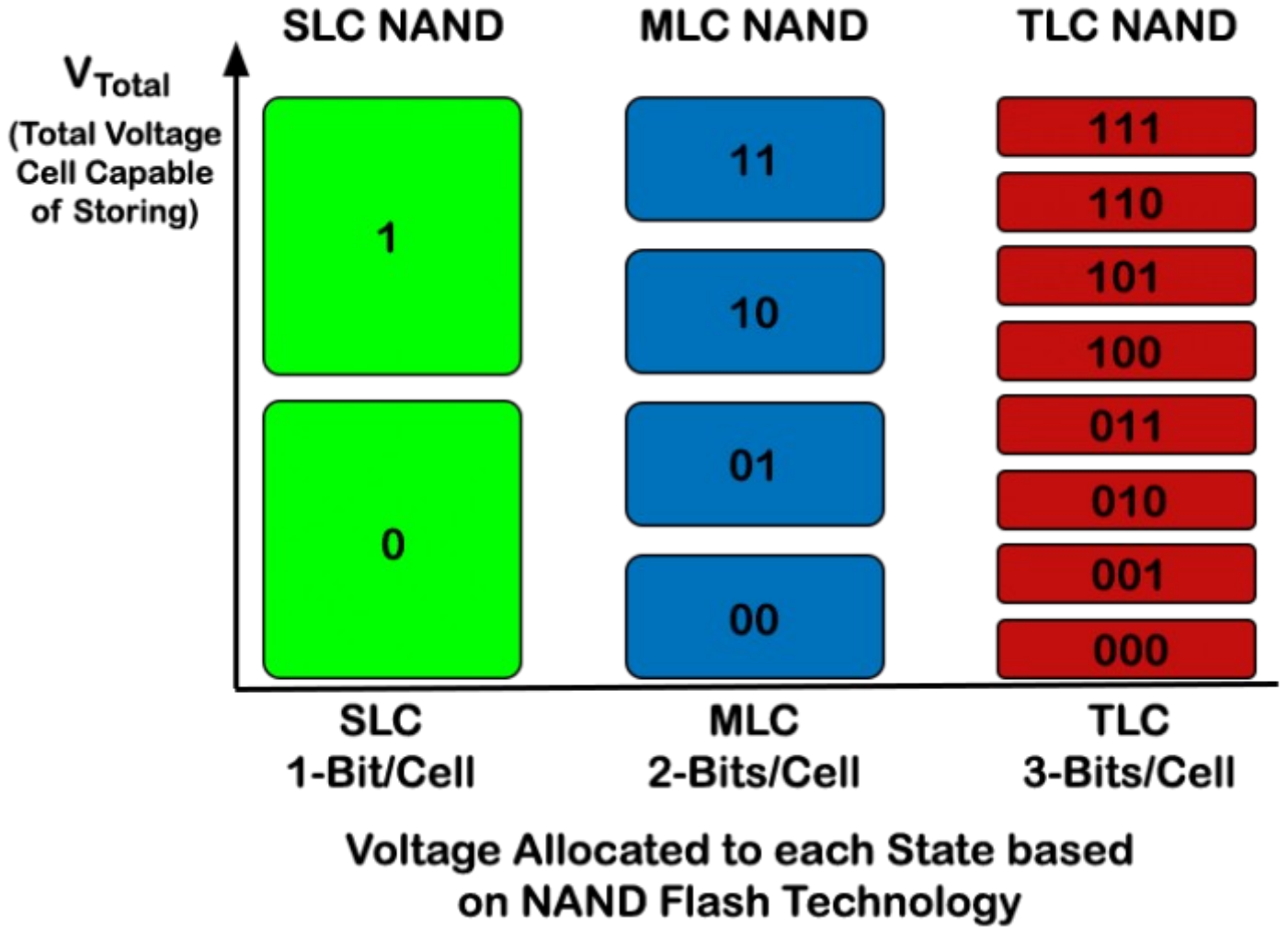
Jak funguje SSD



Jak funguje SSD

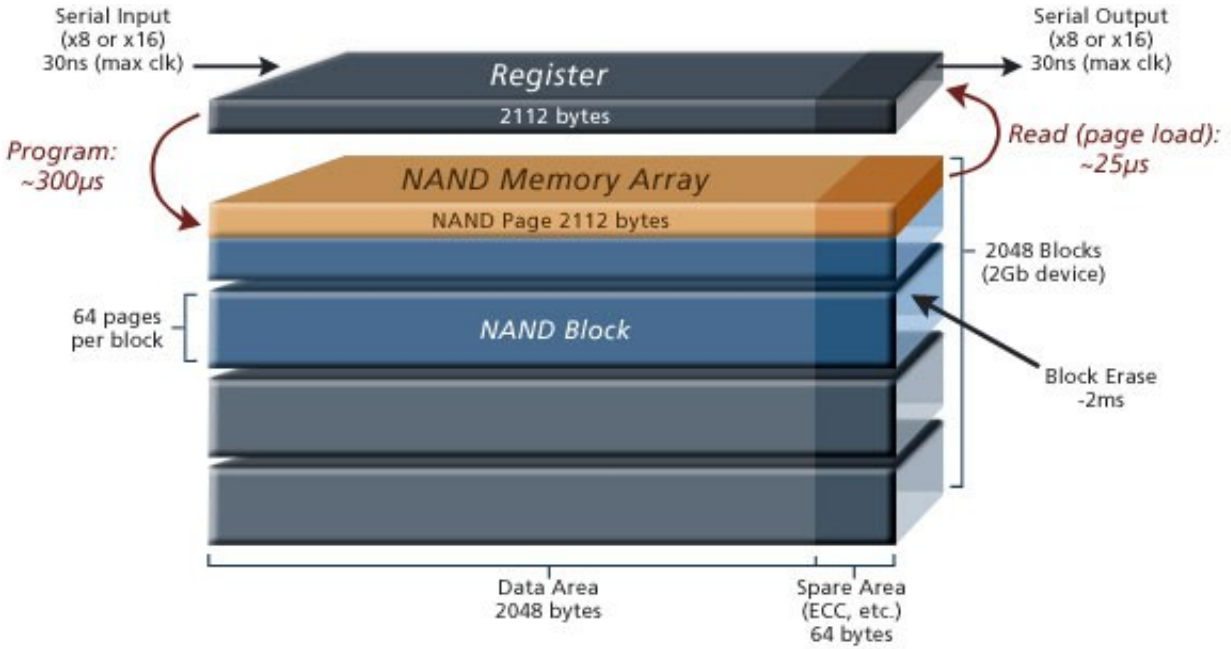
	SLC	MLC	TLC
Bits per cell	1	2	3
P/E cycles (2x nm)	100 000	3 000	1 000
Read time	25us	50us	75us
Program time	200-300us	600-900us	900-1300us
Erase time	1.5-2ms	3ms	4.5ms

Jak funguje SSD



Voltage Allocated to each State based on NAND Flash Technology

Jak funguje SSD



Writing data to a solid-state drive

1. Initial configuration

Block 1000 (data)

PPN	data
0	x
1	y
2	z
3	

Block 2000 (free)

PPN	data
0	
1	
2	
3	

- Initially, block 2000 is free and block 1000 has three used pages at PPN = 0, 1, and 2 (Physical Page Number), and one free page at PPN = 3.

2. Writing a page

Block 1000 (data)

PPN	data
0	x
1	y
2	z
3	x'

Block 2000 (free)

PPN	data
0	
1	
2	
3	

- The data in block 1000 at PPN = 0 gets updated and becomes x'.
- Since pages cannot be overwritten, the page that contains x becomes stale (PPN = 0), and the new version of the data is stored in a free page, at PPN = 3.

3. Erasing a block (garbage collection)

Block 1000 (free)

PPN	data
0	
1	
2	
3	

Block 2000 (data)

PPN	data
1	y
2	z
3	x'

- The garbage collection process copies all the valid pages from the data block 1000 into the free block 2000, leaving behind the stale pages.
- Block 1000 is erased, which makes it ready to receive new write operations. Blocks can only be erased a limited number of times (P/E cycles) until they wear off and become unusable.

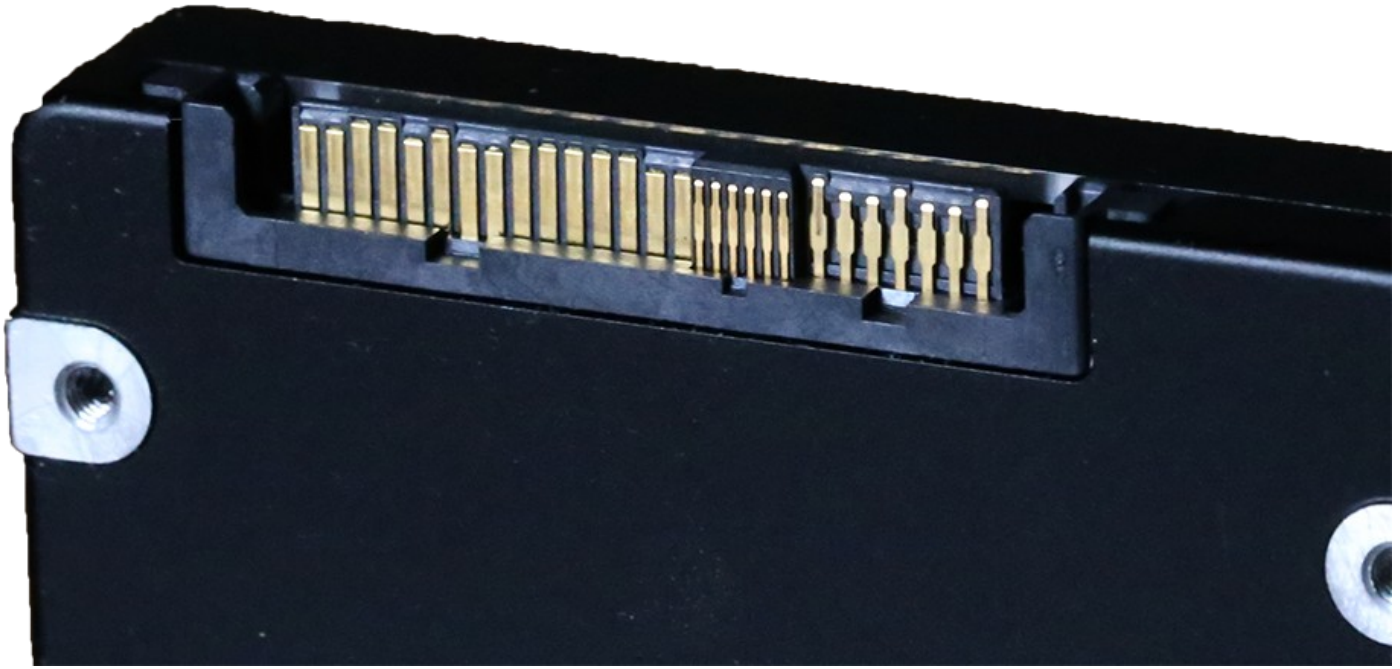
Správa flash paměti

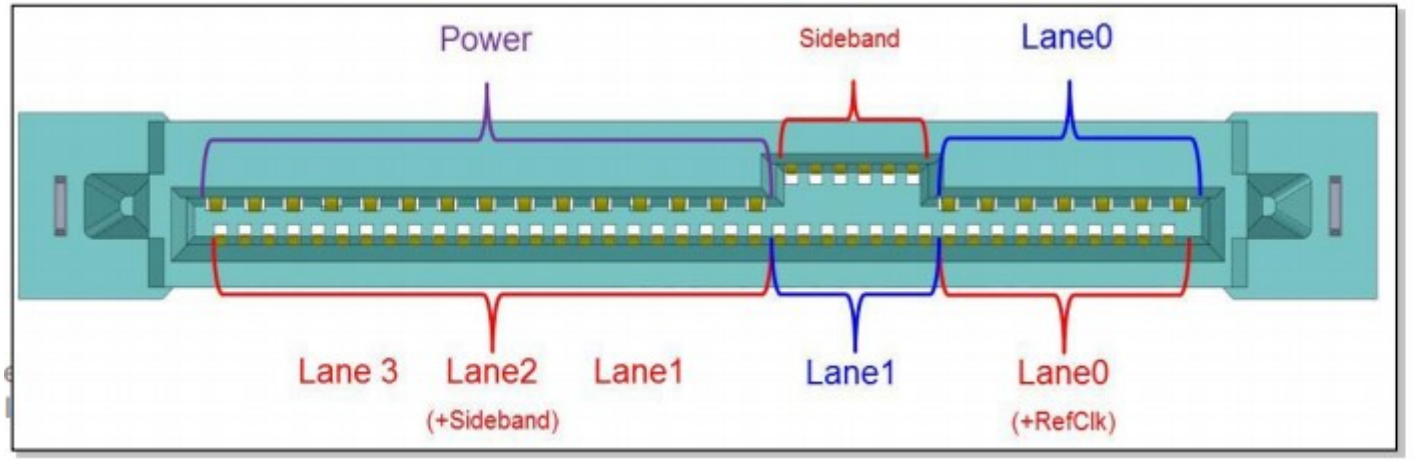
- Správa vadných bloků
  - Náhradní bloky (spare area)
- Wear leveling
- Garbage collection

Provedení a rozhraní

- Provedení a rozhraní
  - PCIe
    - PCIe-to-SAS/SATA – FP/LP karta
    - NVMe – 2.5” disk / karta
  - 3.5”/2.5”/1.8” disk
    - SATA3, SAS2 (6 Gbps)
    - SAS3 (12 Gbps)
  - mSATA
  - M.2

NVMe SFF





SFF-8639

Blue = SAS/SATA

Red = Enterprise PCIe

Vybíráme SSD

- Provedení a rozhraní
- Kapacita
- Overprovisioning (spare area, náhradní bloky)
- Množství zapsaných dat, než odejde
- IOPS, latence
- Konzistentnost výkonu v čase
- Ochrana proti ztrátě napájení

- AnandTech reviews (http://anandtech.com)

Consumer-grade SSD v serveru

- Většina není odolná proti ztrátě napájení
  - OK: Crucial MX100, MX200, M500, M550
  - Ideálně: Intel 530
- TLC moc nevydrží
- Spare area bývá okolo 7% při nejlepším
- Špatná konzistence výkonu v čase
- Vhodné max pro méně vytížené servery nebo pro SSD-only konfigurace
- Ideálně nevyužívat celou kapacitu

Datacenter-grade SSD

- 2.5” SATA3
  - Intel 730
  - Intel DC S3500/S3700 (S3610/S3710)

- NVMe SFF
  - Samsung XS1715
  - Intel DC P3600/P3700 (i ve formě PCIe karet)

- SAS2/SAS3
  - Hitachi, Seagate, Toshiba

- Fusion-IO

SSD v serveru pod Linuxem

- IO je častý bottleneck aplikací
  - Jestli i u Vás, už SSD máte mít :)
- Data nativně na SSD vs. použití jako cache pro rotační disky
- Databáze
  - IOPS ~> TPS
- SATA SSD – AHCI mód řadiče (NCQ)

I/O Scheduler

- Default většinou CFQ
  - Přerovnává I/O v závislosti na LBA
  - Má zbytečnou latenci

- Použijte deadline/noop

```
# echo deadline > /sys/block/$YOURDRIVE/queue/scheduler
```

TRIM

- Informuje SSD o nepoužívaných LBA
- Na serveru ke zvážení
- Lepší alternativou je zvýšit overprovisioning (spare area)
- Případně `fstrim` v idle periodách (cron)
  - Doporučuje se jednou týdně

Oddíly a LVM

- Zarovnání na stránky
  - Pokud si nejste jistí, zarovnávejte na 1MB (moderní distra to dělají automaticky)
  - Zarovnání na erase block nedává smysl

- Vytvářejte FS s velikostí bloku rovnou stránce
  - `mkfs.ext4/mkfs.xfs -b 8k`
  - `mkfs.btrfs -n 8k`

Oddíly a LVM

- TRIM pro LVM v /etc/lvm/lvm.conf
  - V sekci devices:

```
issue_discards = 1
```

RAID

- RAID0 pro ještě vyšší výkon
- RAID1 jako ochrana proti úmrtí elektroniky
- RAID10

- RAID5/RAID6 nedávají příliš smysl
  (opotřebení NAND bude podobné a začnou umírat najednou)
  - K diskuzi
  - Chunk size != page size

- MDRAID umí TRIM, LVM RAID ne

Swap

```
swapon -o discard[=policy]
 discard=once
 discard=pages
 discard
```

```
/dev/sda1   none   swap   defaults,discard  0 0
```

Filesystemy

- TRIM podporují: Btrfs, Ext4, JFS, VFAT, XFS
  - fstrim nepodporuje VFAT
  - mount flag `-o discard`
  - podpora na cestě pro ZFSonLinux

- Btrfs
  - flag -o ssd
  - flag -o ssd_spread

Flashcache

- https://github.com/facebook/flashcache/
- Facebook, 2010
  - Primárně pro akceleraci InnoDB (MySQL)
- Read i write cache
- 3 módy
  - write-back
  - write-through
  - write-around
- Nastavitelné eviction mechanismy
  - FIFO
  - LRU-2Q, konfigurovatelný middle point

Flashcache

- Device mapper, ze dvou blkdev dělá jedno
- Konfigurace přes sysctl
  - dev.cachedev+slowdev.*
    - skip_seq_thresh_kb

- Nepodporuje TRIM

EnhanceIO

- https://github.com/stec-inc/EnhanceIO
- sTec, Inc., 2012, fork Flashcache
  - Přepsaný write-back režim
    - Komprimovaná metadata
  - Nepoužívá device mapper
    - Cache device se dá odebrat za běhu
  - Nyní pozadu oproti Flashcache
  - Podporuje (zatím) až Linux 3.18

bcache

- V jádře od 3.10
  - RHEL6 ze hry
- 2 módy
  - write-trough
  - Write-back
- Konfigurace v /sys/block/yourcachedev/bcache
  - cache_mode
  - sequential_cutoff
- Cache zařízení manipulovatelná za běhu

dm-cache

- V jádře od 3.9
- Device mapper modul
- Umí separátní zařízení na metadata
- 3 módy
  - write-back
  - write-through
  - pass-through
- Replacement policies
  - multiqueue
  - cleaner

ZFS

- ZFSonLinux (https://github.com/zfsonlinux/zfs)
- L2ARC
- SLOG pro ZIL

Vlastní zkušenost

- vpsFree.cz používá SSD od 2011
  - Začínali jsme s Intel 320 a OCZ Vertex 2
- Dobíhají OCZ Vertex 3 (240 GB in prod)
- Od 2013 kupujeme Intel DC S3700 200 GB
- Nejstarší Vertex 3 120 GB mají zapsáno ~70TB
  - Na nich jela Flashcache
- Flashcache do Q2 2014
  - Větší počet MySQL instancí problém
- ZFS od Q2 2013 na prvních strojích, Q2 2014 plný přechod
  - Největší benefit je dedicated SLOG device
  - L2ARC

Diskuze

- Vaše zkušenosti
- Dotazy
- RAID a SSD