SSD & Oracle

- **mag. Sergej Rožman**; Abakus plus d.o.o.

- The latest version of this document is available at: http://www.abakus.si/

Dilemma

Abakus
As na disku.

SSD & Oracle

mag. Sergej Rožman

sergej.rozman@abakus.si

Strokovno srečanje SIOUG

SIOUG 2017

ORACLE Gold Partner

Municipality of Koper · Harta · Milenijum Osiguranje · futuraplus · Lonia · Banka Slovenije Evrosistem · Gorenjska Banka · Goodyear

City of Ljubljana · Iskra · Sava Savatech, d.o.o. · Delo Prodaja · Ljubljana Airport · Infond vzajemni skladi · Generali Group · Kontrola zračnega prometa Slovenije · Mercator

Abakus plus d.o.o.

ORACLE Gold Partner

History
- from 1992, ~20 employees

Applications:
- special (DB – Newspaper Distribution, FIS – Flight Information System)
- **ARBITER – the ultimate tool in audit trailing**
- **APPM - Abakus Plus Performance Monitoring Tool**

Services:
- DBA, OS administration , programming (MediaWiki, Oracle)
- networks (services, VPN, QoS, security)
- open source, monitoring (Nagios, OCS, Wiki)

Hardware:
- servers, **SAN storage**, firewalls, **Backup Server**

Experience:
- from 1995 GNU/Linux ***(22 years of experience !)***
- Oracle on GNU/Linux: since RDBMS 7.1.5 & Forms 3.0 ***(before Oracle !)***
- **>25 years of experience with High-Availability !**

Municipality of Koper · Harfa · Milenijum Osiguranje · futuraplus · Lonia · Banka Slovenije Evrosistem · Gorenjska Banka · Goodyear · City of Ljubljana · Iskra · Sava Savatech d.o.o. · Delo Prodaja · Ljubljana Airport · Infond vzajemni skladi · Generali Group · Kontrola zračnega prometa Slovenije · Mercator

SSD Experience

Manufacturer	Quantity	First Use
Intel (160 GB)	10	Nov. 2009
Kingston (120 – 480 GB)	~50	Oct. 2011
OCZ (120 – 512 GB)	~200	Jun. 2012
Intenso (256 GB)	9	Sep. 2015
Toshiba (240 – 960 GB)	6	Feb. 2016
Crucial (1 – 2 TB)	28	Nov. 2016
Samsung (250 GB – 4 TB)	~800	Feb. 2014

HDD : SSD Comparison

HDD

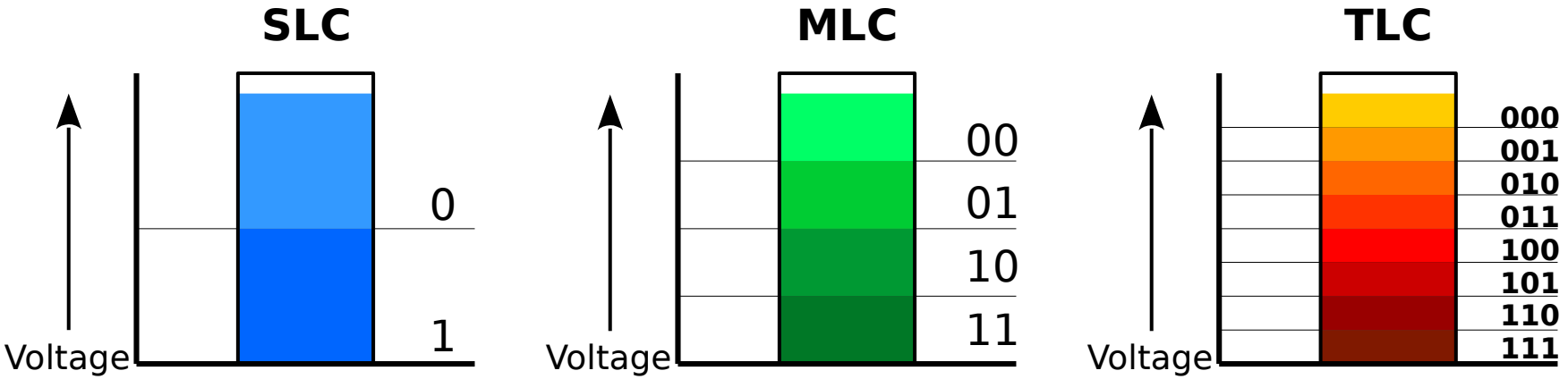
- slower
- bulky
- heavy
- noisy
- damaged when dropped
- life span?
  (mechanical structures wearing)

SSD

- faster
- lighter
- quieter
- shock-resistant
- limited life span (P/E)

Abakus
As na disku.

Flash Cell



AS Abakus
As na disku.

SSD – Internal Structure

	SLC	MLC	TLC
Bits per cell	1	2	3
P/E Cycles	100.000	3.000	1.000
Read Time	25 µs	50 µs	75 µs
Program Time	200-300 µs	600-900 µs	900-1350 µs
Erase Time	1,5-2 ms	3 ms	4,5 ms

block

512kB – 2MB

erase operations

page

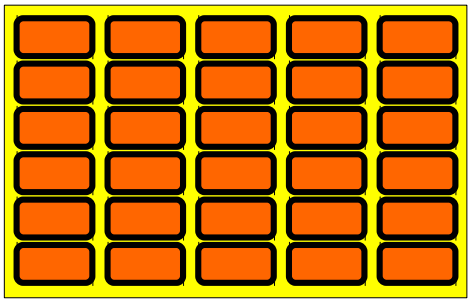
4 - 8 kB

read/write operations

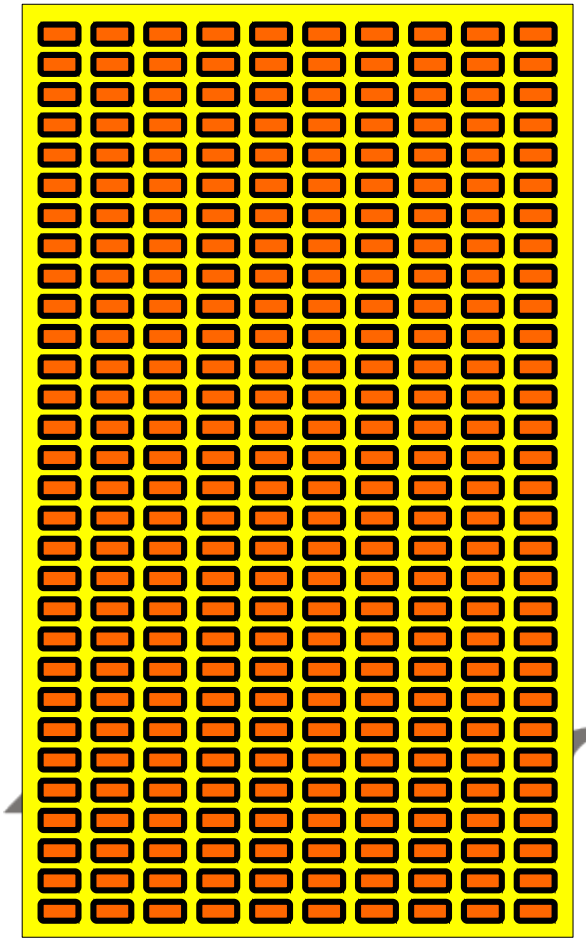
cell

1 - 3 bit

group



group



IOPS

How many IOPS per disk?

- 15k rpm (average rotational delay ~ one-half the rotational period = 2 ms),
- 3 ms average seek time
- 100 MB/sec transfer rate
- 4 kB block

SAS HDD

- 2 ms + 3 ms + (4kB) / (100MB/s) = 5,04 ms
- 1 / 5,04 ms = **198 IOPS**

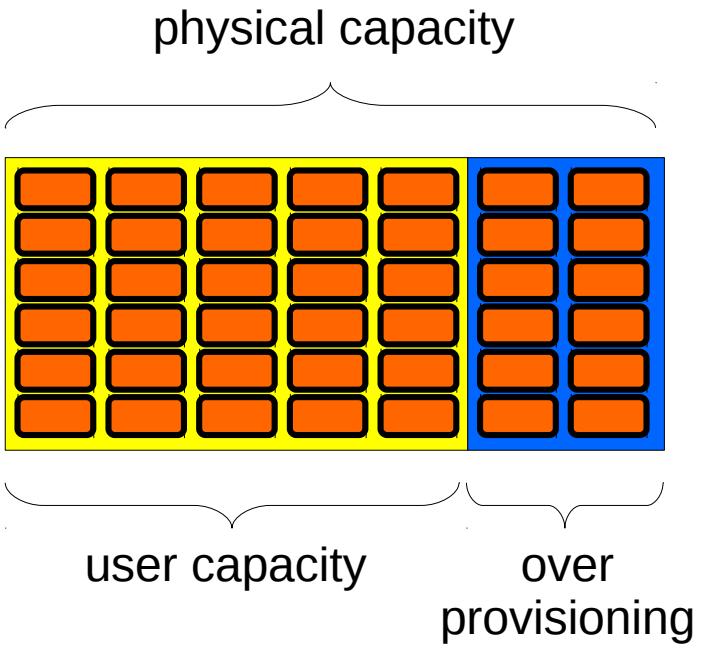
SATA SSD

- 75 µs + (4kB) / (500MB/s) = 82,8 µs
- 1 / 82,8 µs = **12.000 IOPS**
  (for one channel flash access)

Device	IOPS
SATA drive 7.200 rpm	~100
SAS drive 10k rpm	~150
SAS drive 15k rpm	~200
SSD drive SATA	up to 120.000
SSD drive NVMe	up to 1.200.000

Over-provisioning

$$\frac{\text{physical capacity} - \text{user capacity}}{\text{user capacity}} \times 100 = \text{over-provisioning percentage}$$



- typical over-provisioning: 7 – 40%
- can be adjustable

Free Pages

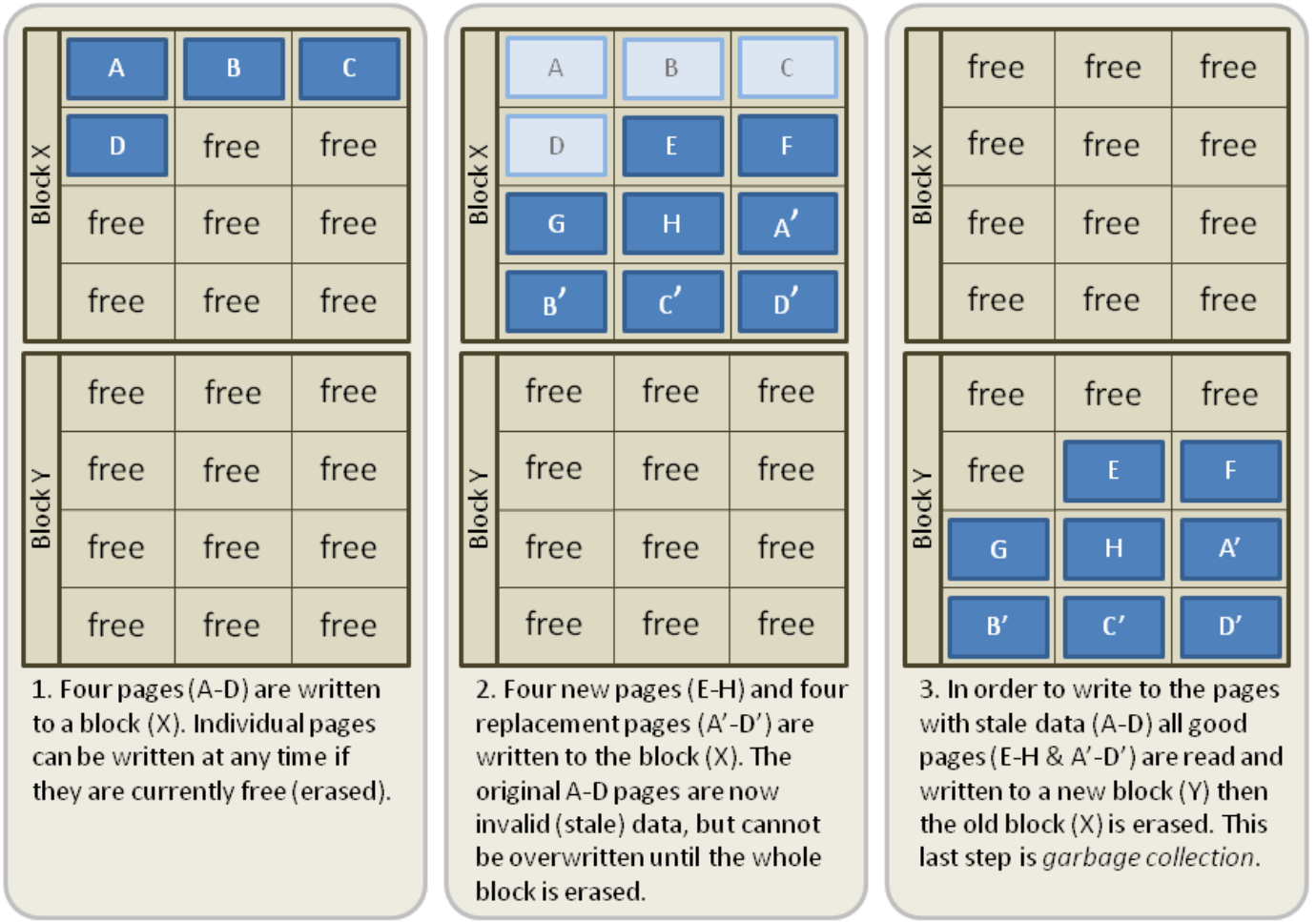
Which pages are considered as free?

- never used pages (new)
- drive secure erase
- over-provisioning
- trim operation
- **overwritten pages**

Abakus
As na disku.

Write Amplification & Garbage Collection

$$\frac{\text{data written to the flash memory}}{\text{data written by the host}} = \text{write amplification factor}$$

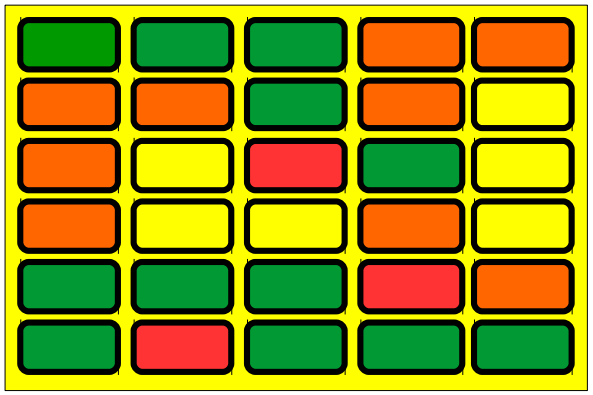


Example write amplification (1–3):

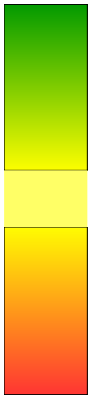
$$\frac{20}{12} = 1{,}67$$

Wear Levelling

without wear levelling

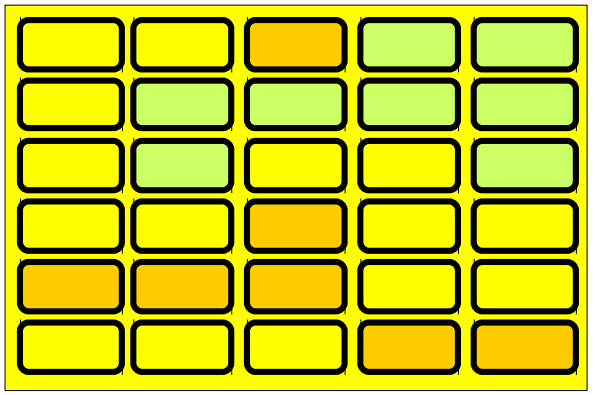


new



worn

with wear levelling



Durability

Example:

- 1 TB SSD drive (10% over-provisioning)
- 1000 max P/E cycles
- 100 GB data written per day
- 3 – average write amplification factor

Expected life span of a single drive:

$$\frac{1{,}1 \times 1\ TB \times 1000\ cycles}{3 \times 100\ GB/day} = 3.667\ days = 10\ years$$

Experience

Interference
&
Errors

Fragmentation

- SSD fragmentation doesn’t matter. Really?
- SSD drive is always internally fragmented (striped)
- How many I/O operations to read a file?
  (can be a lot of work to do)

```
# filefrag zimbra.img
zimbra.img: 150360 extents found
```

- **Nevertheless: don’t defrag SSD drives!**

Fuzzy I/O times

- HDD: I/O times not consistent, predictable
- SSD: I/O times more consistent, nonpredictable

Bad Blocks

- HDD: most common error
- SSD: very uncommon (non-existent?)

AS Abakus
As na disku.

Tired SSD drive

- SSD performance degrades over time. Sometimes significantly.
- A secure erase may help.

Offline Drive

- SSD drive goes offline occasionally, unknown reason

SSD Dementia?

- long-term memory content is preserved
- short-term memory content is lost

A ≠ A'

AS Abakus
As na disku.

ASM (12.2.0.1.0)

ASM alertlog:

```
NOTE: Scrubbing operation is requested by database (gn 13889214760156266498,
fn 410739070249572787 2blk 90597, ftype 2, lblksiz 8192)
NOTE: Scrubbing operation is requested by database (gn 13889214760156266498,
fn 410739070249572787 2blk 90598, ftype 2, lblksiz 8192)
NOTE: Scrubbing operation is requested by database (gn 13889214760156266498,
fn 410739070249572787 2blk 90599, ftype 2, lblksiz 8192)
...
```

DB alertlog:

```
Corrupt block relative dba: 0x070161fe (file 28, block 90622)
Bad header found during user buffer read
Data in bad block:
 type: 110 format: 4 rdba: 0x1a51fee5
 last change scn: 0x7f03.5e9c.b5c3e703 seq: 0xc1 flg: 0xf0
 spare3: 0x4f7
 consistency value in tail: 0x64c57453
 check value in block header: 0x4d0b
 block checksum disabled

Reading datafile '+SRTEST/TTABAKUS/DATAFILE/sr.256.956326421' for corruption
at rdba: 0x070161fe (file 28, block 90622)
Read datafile mirror 'SRTEST_0002' (file 28, block 90622) found same corrupt
data (no logical check)
Read datafile mirror '' (file 28, block 90622) found valid data
Hex dump of (file 28, block 90622) in trace file
/oradmin/ttabakus/diag/rdbms/ttabakus/ttabakus/trace/ttabakus_ora_21484.trc
...
```

Performance

Abakus SAN (2017)
16x SSD - RAID10, QDR Infiniband 40G (single link)
(dbms_resource_manager.calibrate_io)

- max_iops = **149.918**
- latency = **0**
- max_mbps = **3.184**

References

- Chris Buckel; *Understanding Flash: SLC, MLC and TLC*
(https://flashdba.com/2014/07/03/understanding-flash-slc-mlc-and-tlc/)
- English Wikipedia;
(https://en.wikipedia.org/wiki/Write_amplification)
- Leo Bien Durana; *TLC IS BECOMING THE MAINSTREAM FOR SSD IN THE CONSUMER MARKET*
(https://www.techporn.ph/tlc-is-becoming-the-mainstream-for-ssd-in-the-consumer-market/)
- Steve Larrivee; *Solid State Drives 101: Everything You Ever Wanted to Know*
(https://www.cactus-tech.com/resources/blog/details/solid-state-drives-101)

SSD & Oracle

mag. Sergej Rožman
ABAKUS plus d.o.o.
Ljubljanska c. 24a
Kranj

e-mail: sergej.rozman@abakus.si

phone: +386 4 287 11 14