

Clash of the Titans

I/O System Performance

- mag. **Sergej Rožman**; Abakus plus d.o.o.
- The latest version of this document is available at:
<http://www.abakus.si/>





As na disku.

Clash of the Titans I/O System Performance

mag. Sergej Rožman

sergej.rozman@abakus.si



17. Strokovno srečanje

SIOUG 2012

Kongresni center Hotel Mons Ljubljana, 15. - 17. oktober

SIOUG

Slovensko
društvo Oracle
uporabnikov

ORACLE Gold
Partner



Mestna občina Ljubljana





Clash of the Titans I/O System Performance

mag. Sergej Rožman

sergej.rozman@abakus.si



OO17. KONFERENCIJA
HRVATSKE UDRUGE
ORACLE KORISNIKA

ORACLE Gold
Partner



Mestna občina Ljubljana



MESTNA OBČINA KOPER
COMUNE CITTA DI CAPODISTRIA



Aerodrom Ljubljana



Mercator



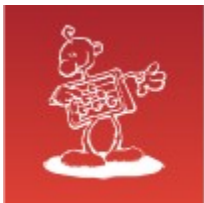
IskraSistemi



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA FINANCE

**BANKA
SLOVENIJE**
EVROSISTEM





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

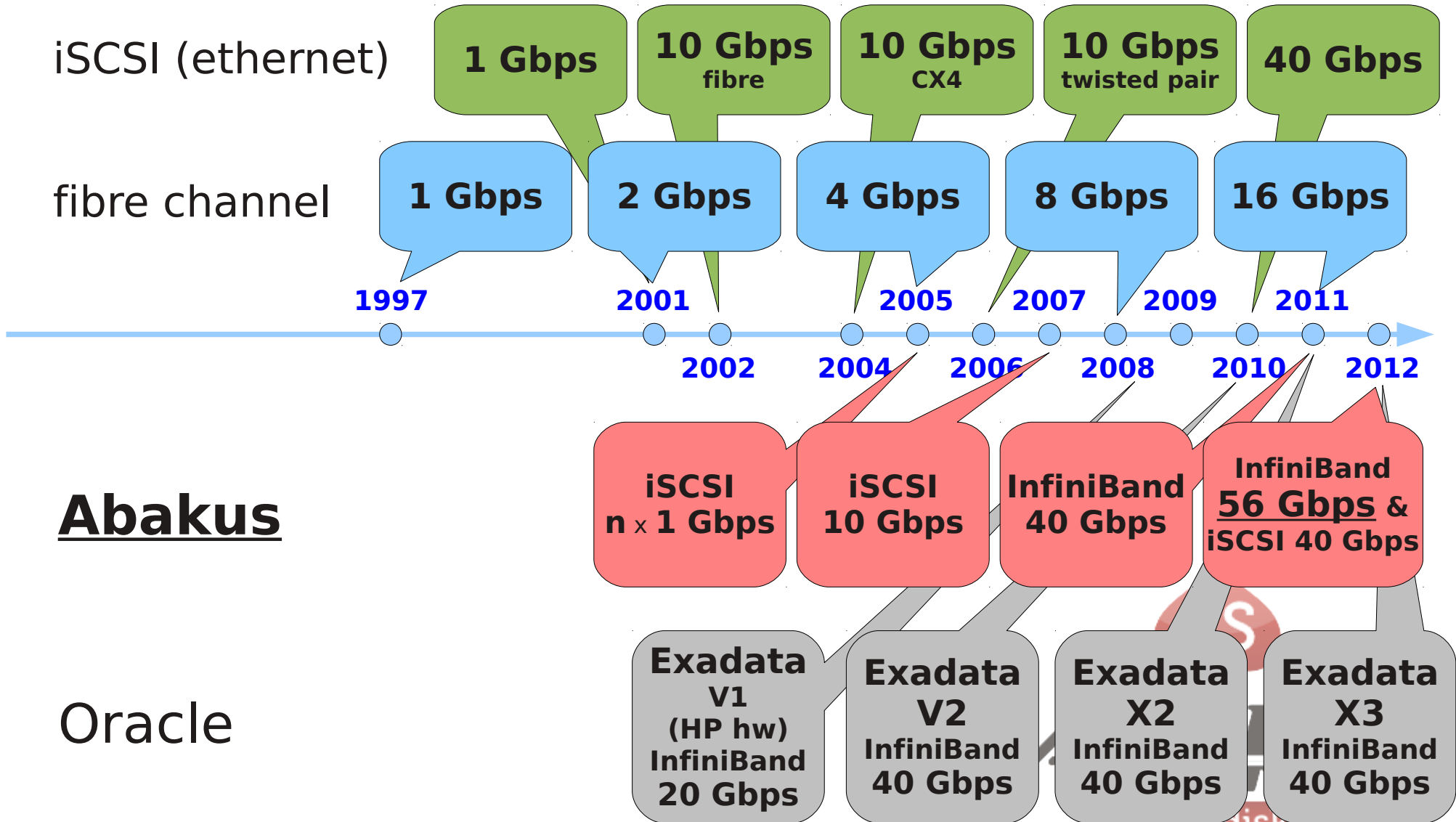
Infrastructure:

- from 1995 GNU/Linux **(17 years of experience !)**
- Oracle on GNU/Linux: since RDBMS 7.1.5 & Forms 3.0 **(before Oracle !)**
- **20 years of experience with High-Availability !**





Timeline



Abakus

Oracle



Oracle Exadata

Advantages

- Oracle Cell Server
 - storage indexes
- »State of the art software«
- performance

Disadvantages

- closed design,
no customization allowed
- Oracle 11g only
- not so »State of the art
hardware«
- price





The Most Expensive SAN Features

- technology (fibre channel vs. others)
- performance (# of IOPS)
- size of cache
- write-back cache with battery backup
- deduplication





Trick Questions

- How much disk space do you need for your database?
- Disks have become faster over time! Really?
- Are SSD drives really very expensive and have short life spans?
- Is write-through cache really faster than write-back cache? Is battery backup unit really necessary?
- Is currently popular deduplication technology safe and useful?





Don't use RAID5!

RAID5 write

- read old data block, read old cksum block
- subtract old data from old cksum
- add new data to cksum
- write new data block, write new cksum block

RAID10 write

- write new data block to disk1, write new data block to disk2

SAN – Sample Specification

RAID	IOPS
Random Writes RAID10	14.399
Random Writes RAID5	2.703
Random Writes RAID6	1.942





How much disk space?

SAN Admin

- How much disk space do you need for your database?

DBA

- About 500 GB.

SAN Admin

- I have one mirrored 3 TB SATA disk in the SAN with two databases on it already. But it has more than enough free space for your database.

DBA

- One disk!?





What about (physical) 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

IO time

- $2 \text{ ms} + 3 \text{ ms} + (4\text{kB}) / (100\text{MB/s}) = 5,04 \text{ ms}$
- $1 / 5,04 \text{ ms} = \mathbf{198 \text{ IOPS}}$

IOPS on SANs are usually limited by

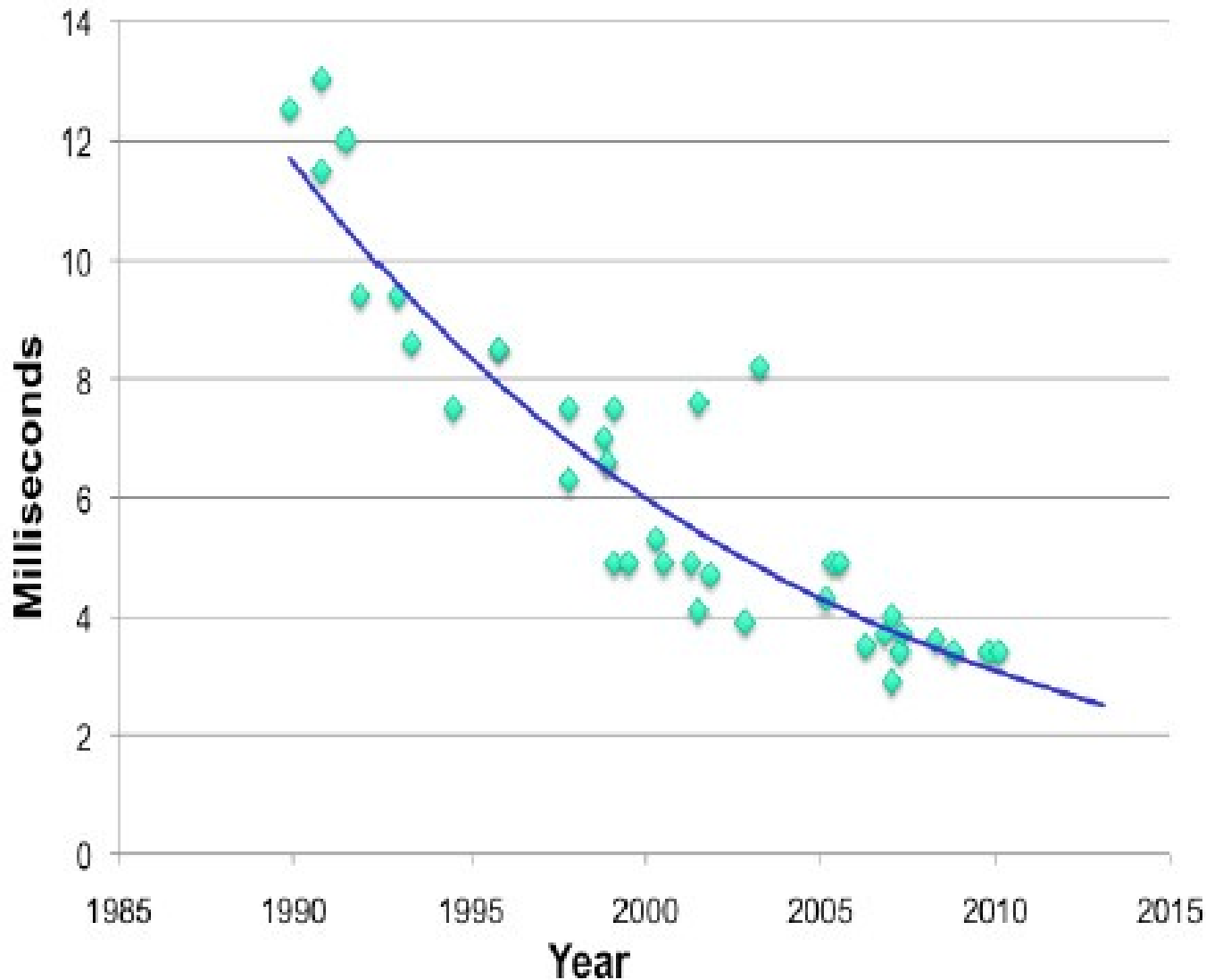
- number of disks!
- not by amount of SAN cache
- not by SAN model nor by manufacturer

Device	IOPS
SATA drive 7.200 rpm	~100
SAS drive 10k rpm	~150
SAS drive 15k rpm	~200
SSD drive SATA/SAS	5.000 – 120.000
SSD drive PCI-E	up to 1.200.000



Disks have become faster over time! Really?

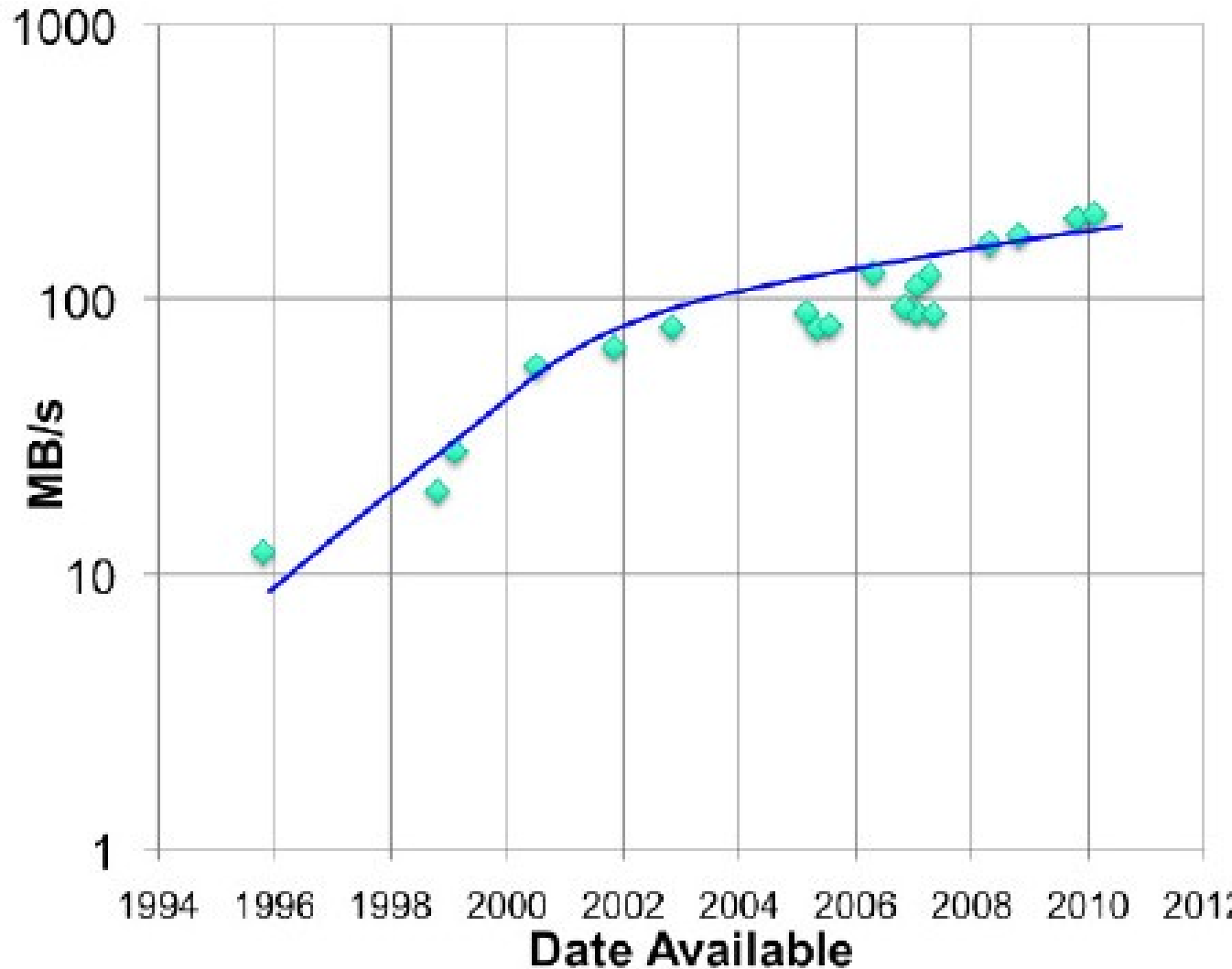
Average seek time over the years

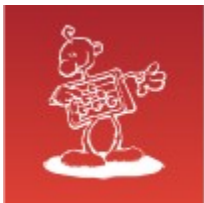




Disks have become faster over time! Really?

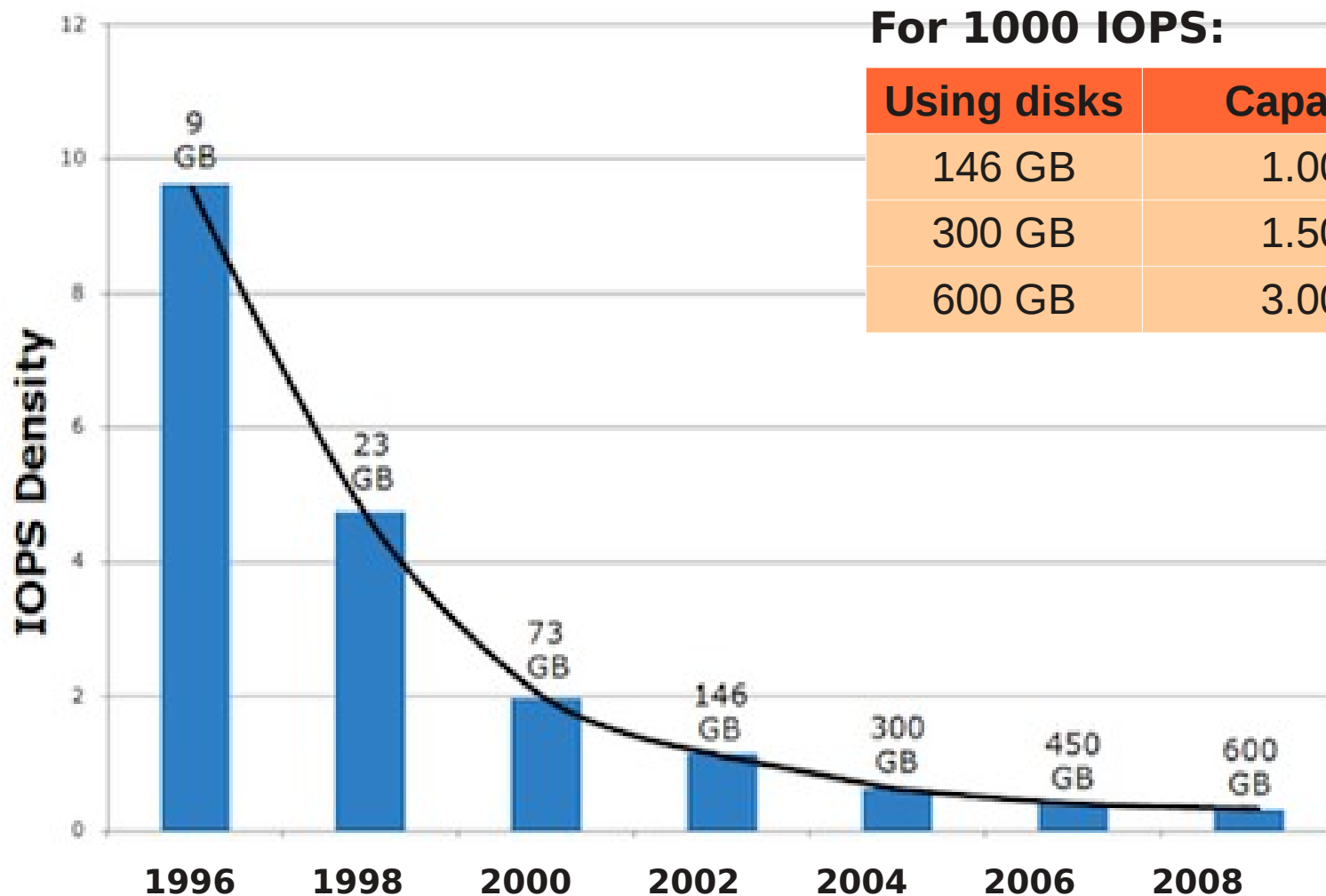
Bandwidth over the years





Disks have become faster over time! Really?

IOPS per GB over the years



For 1000 IOPS:

Using disks	Capacity	# of disks
146 GB	1.000	7
300 GB	1.500	5
600 GB	3.000	5





Are SSD drives really very expensive?

We need 500 GB and 1000 IOPS:

Using disks	Capacity	# of disks	Unit price*	Total cost
146 GB SAS	1.000	7 (+7)	327 USD	2.289 (+2.289)
300 GB SAS	1.500	5 (+5)	200 USD	1.000 (+1.000)
600 GB SAS	3.000	5 (+5)	380 USD	1.900 (+1.900)
512 GB SSD	512	1 (+1)	430 USD	430 (+430)

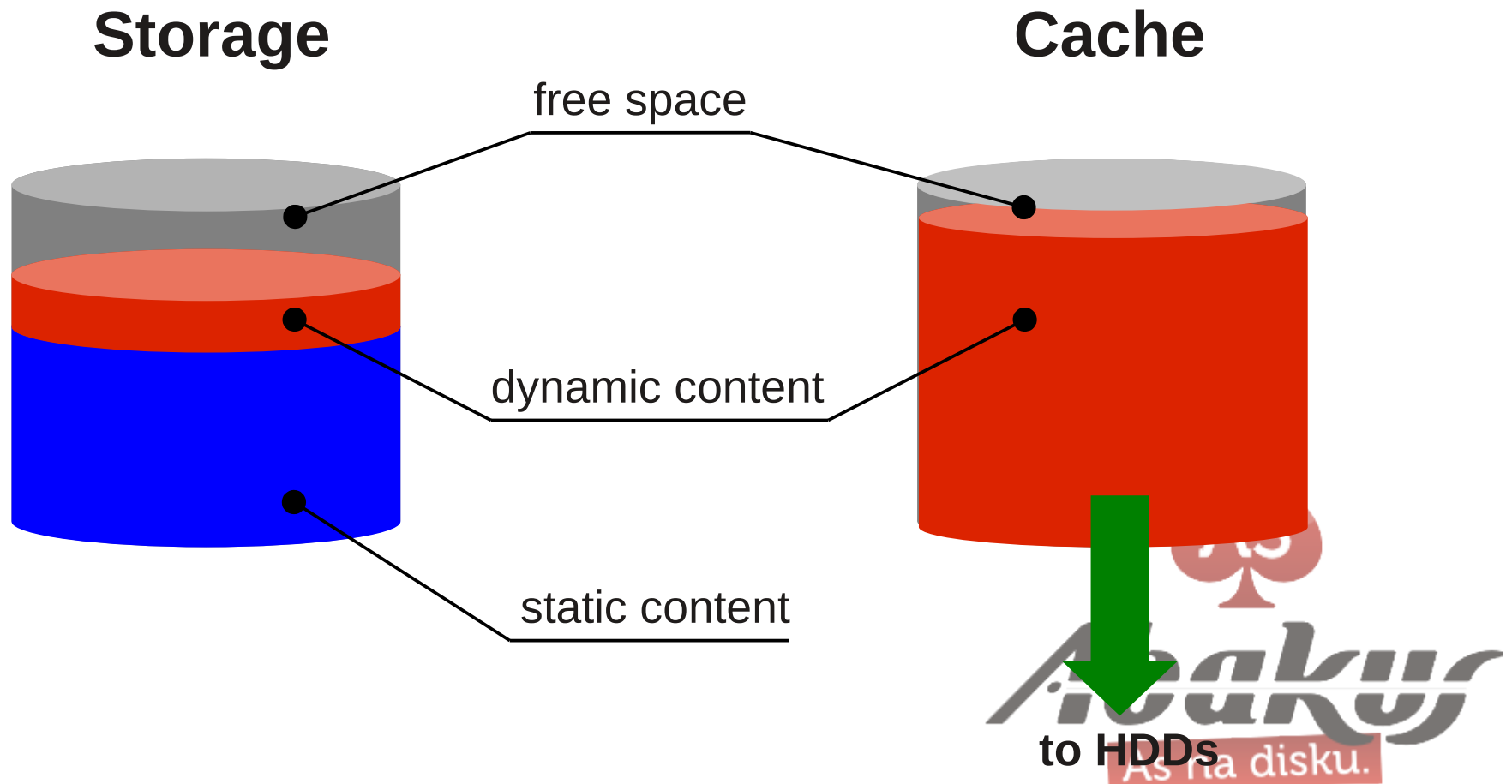
* price from <http://www.newegg.com>





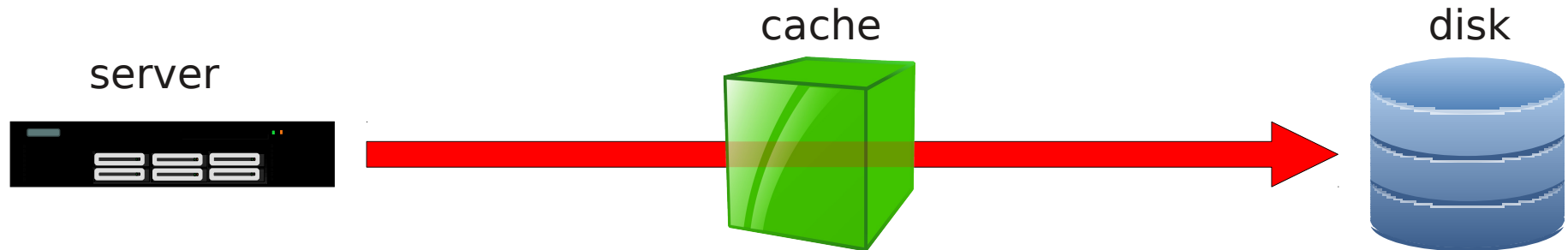
What about life span?

- SAN producers claim that SSDs have too short life spans for using them in the enterprise environment. They can only be used as a flash cache.





Write-through cache

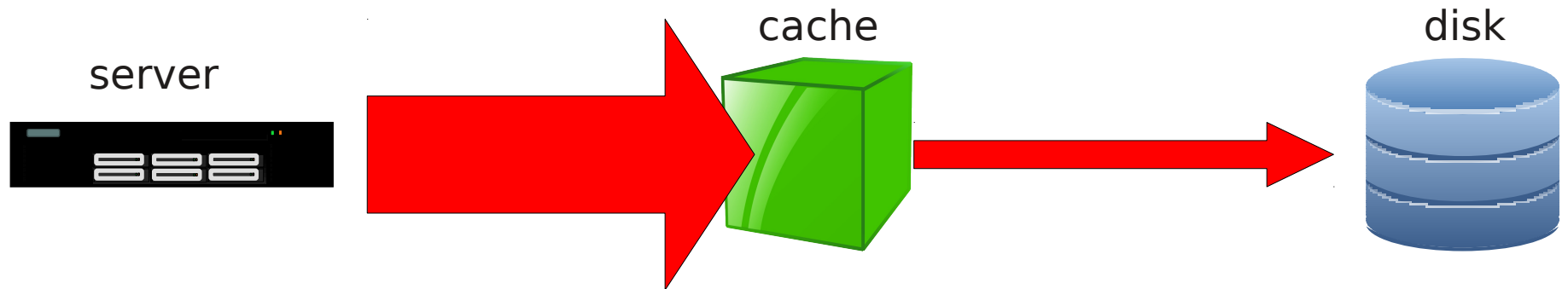


data flow:

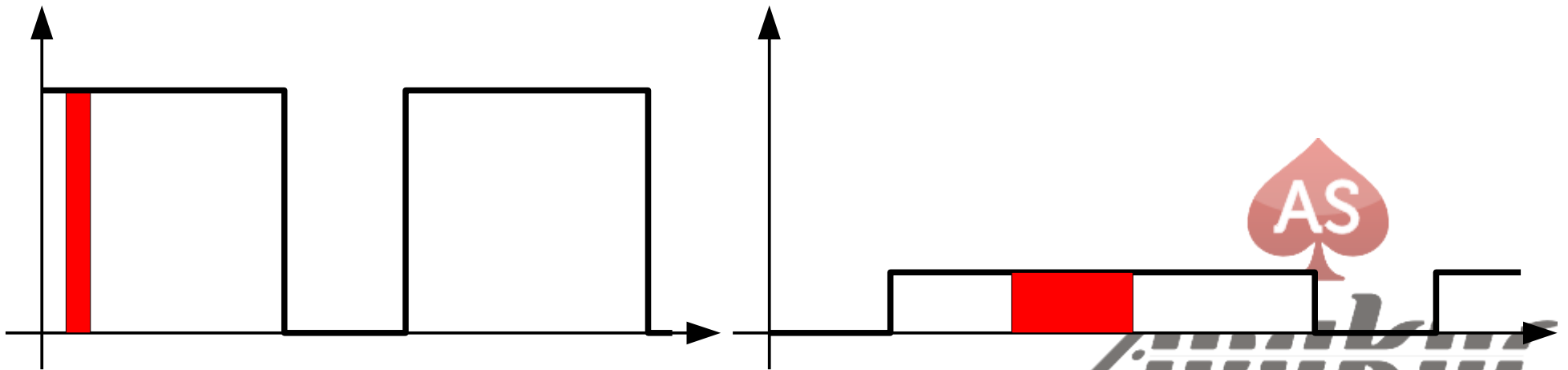




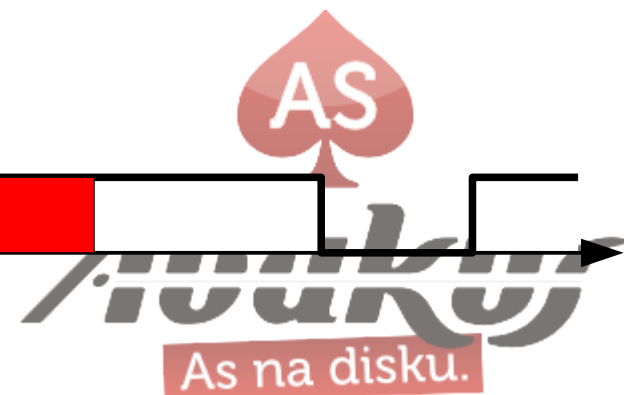
Write-back cache



data flow:



Battery backup?



Write back vs. Write through?

- **Common wisdom is WB is always faster?**
 - Not always so!

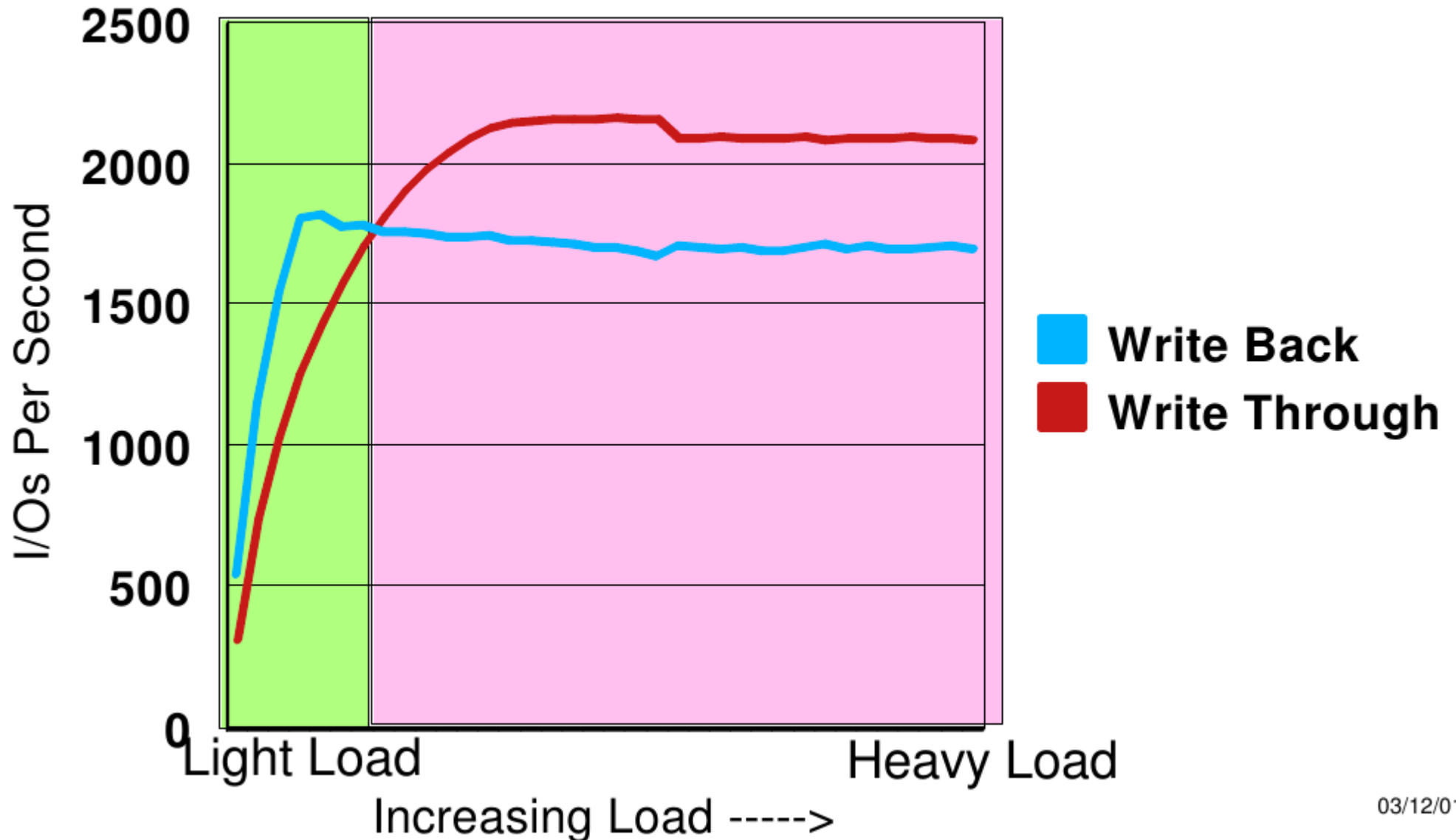
- **WT is usually faster for heavy loads**
 - ▶ select WT
 - ▶ RAID-5 may be best with WB
 - Only when performing sequential loads

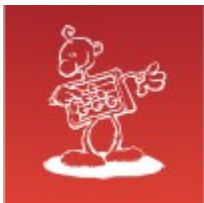
- **WB is usually faster for light loads**



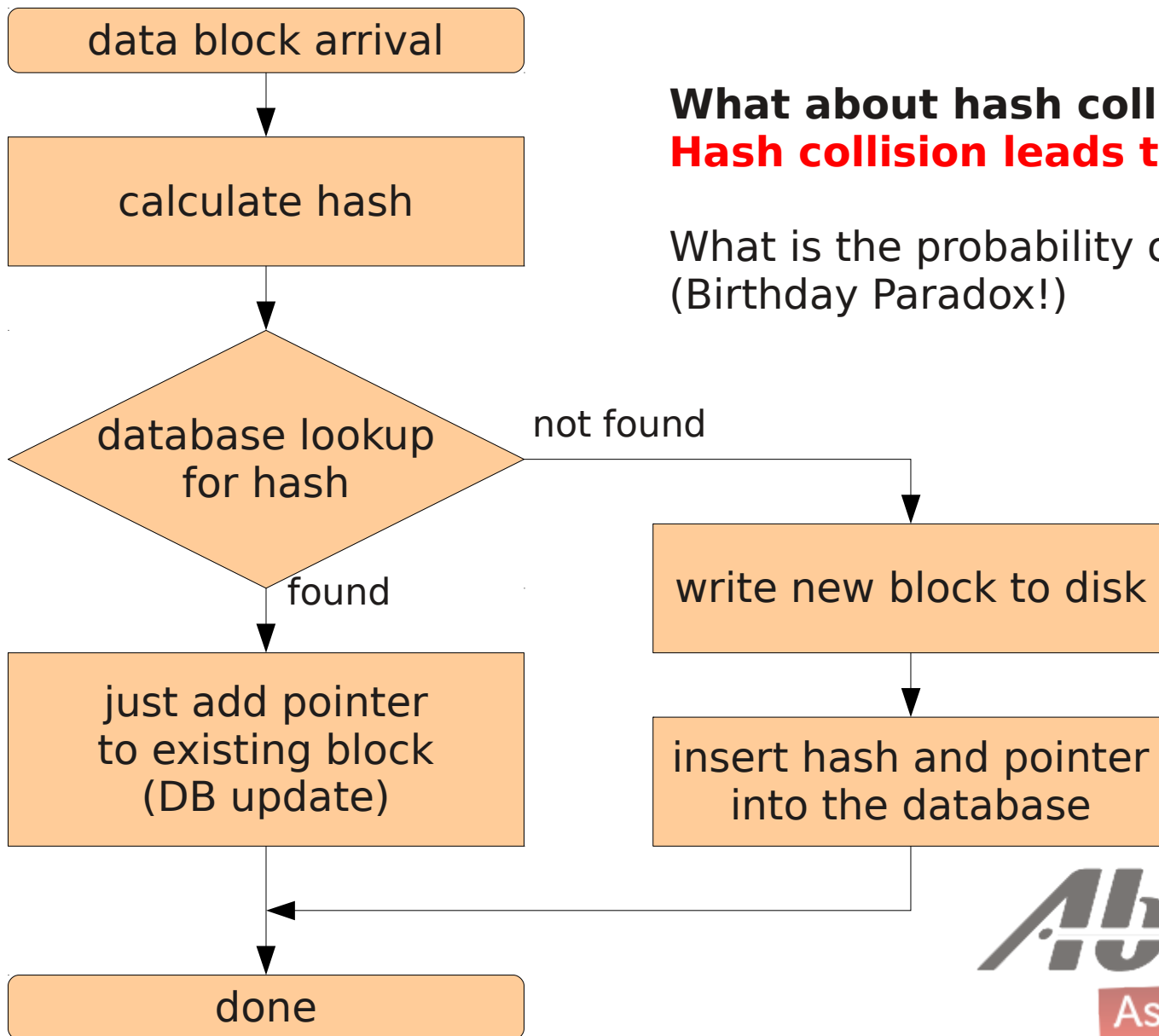
Write-through vs. write-back cache

OLTP 8K Workload





In-line Deduplication



What about hash collision?

Hash collision leads to data loss!

What is the probability of a hash collision?
(Birthday Paradox!)





Hash collision probability

Birthday paradox

$$p(n) = 1 - \frac{n! \binom{2^h}{n}}{2^{hn}}$$

h ... size of hash (bits)

n ... # of data blocks

- unfeasible to compute for large numbers
- approximation using Taylor series

$$p(n) \approx 1 - e^{-\frac{n^2}{2^{h+1}}}$$

# of blocks	hash size	probability
1.000.000.000	64 bit	2,67%
1.000.000.000.000	96 bit	0,0006%
1.000.000.000.000	128 bit	1,4E-13%
23	365	50,73%



Summary

- Don't use RAID5! (no need for RAID at all, use Oracle ASM)
- Use SSDs! They are fast and reliable.
- Use SSDs! You need speed, not space.
- No need for write-back cache & battery backup with Oracle.
- Don't use deduplication. It is slow!





References

References

- Write back vs. Write through? – IBM
- HDD Technology Trends, Storage Newsletter.com
- SAN Stories – IO Performance, Anjo Kolk, Symantec
- Wikipedia – IOPS
- Frits Hoogland Weblog – <http://fritshoogland.wordpress.com>





Clash of the Titans

I/O System Performance

The life is good!

(Piet de Visser)



mag. Sergej Rožman

ABAKUS plus d.o.o.
Ljubljanska c. 24a
Kranj

e-mail: sergej.rozman@abakus.si

phone: +386 4 287 11 14

