

# High Availability in Numbers

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



## High Availability in Numbers

mag. Sergej Rožman

sergej.rozman@abakus.si

18. Strokovno srečanje SIOUG



**SIOUG 2013**

15.-16. oktober 2013

**ORACLE** Gold Partner



Mestna občina Ljubljana



MESTNA OBČINA KOPER  
COMUNE CITTA DI CAPODISTRIA



Aerodrom Ljubljana



REPUBLIKA SLOVENIJA  
MINISTRSTVO ZA FINANCE



Mercator



IskraSistemi



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

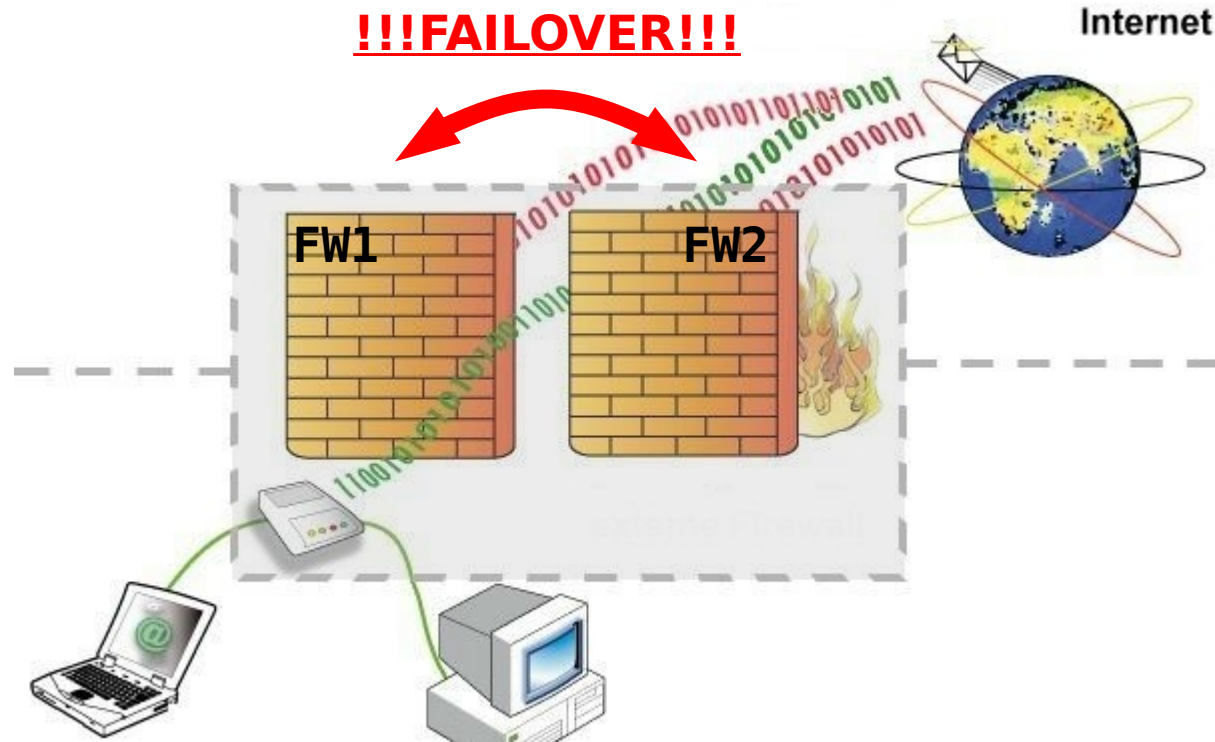


Mestna občina Ljubljana





# Router/Firewall





# HA Customers

**BANKA  
SLOVENIJE**  
EVROSISTEM



Aerodrom Ljubljana



 **Iskra**  
Iskra Sistemi, d.d.

**Gorenjska Banka**  
*Banka s poslubom*



Mestna občina  
Ljubljana



**Hidria**

 **HRANILNICA  
LON d.d., Kranj**



MESTNA OBČINA KOPER  
COMUNE CITTA DI CAPODISTRIA

  
**St. BERNARDIN**  
ADRIATIC RESORT & CONVENTION CENTER  
PORTOROŽ - SLOVENIJA







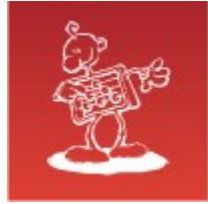
# IDC's Availability Spectrum

Availability level	Impact of Component Failure
AL 1	Work stops; uncontrolled shutdown; data integrity ensured
AL 2	User interrupted, but can quickly log on again; may need to rerun some transactions from journal; may experience performance degradation
AL 3	User stays online; current transaction may need restarting; may experience performance degradation
AL 4	Transparent to user; no interruption of work; no transactions lost; no degradation in performance



# Availability Classes

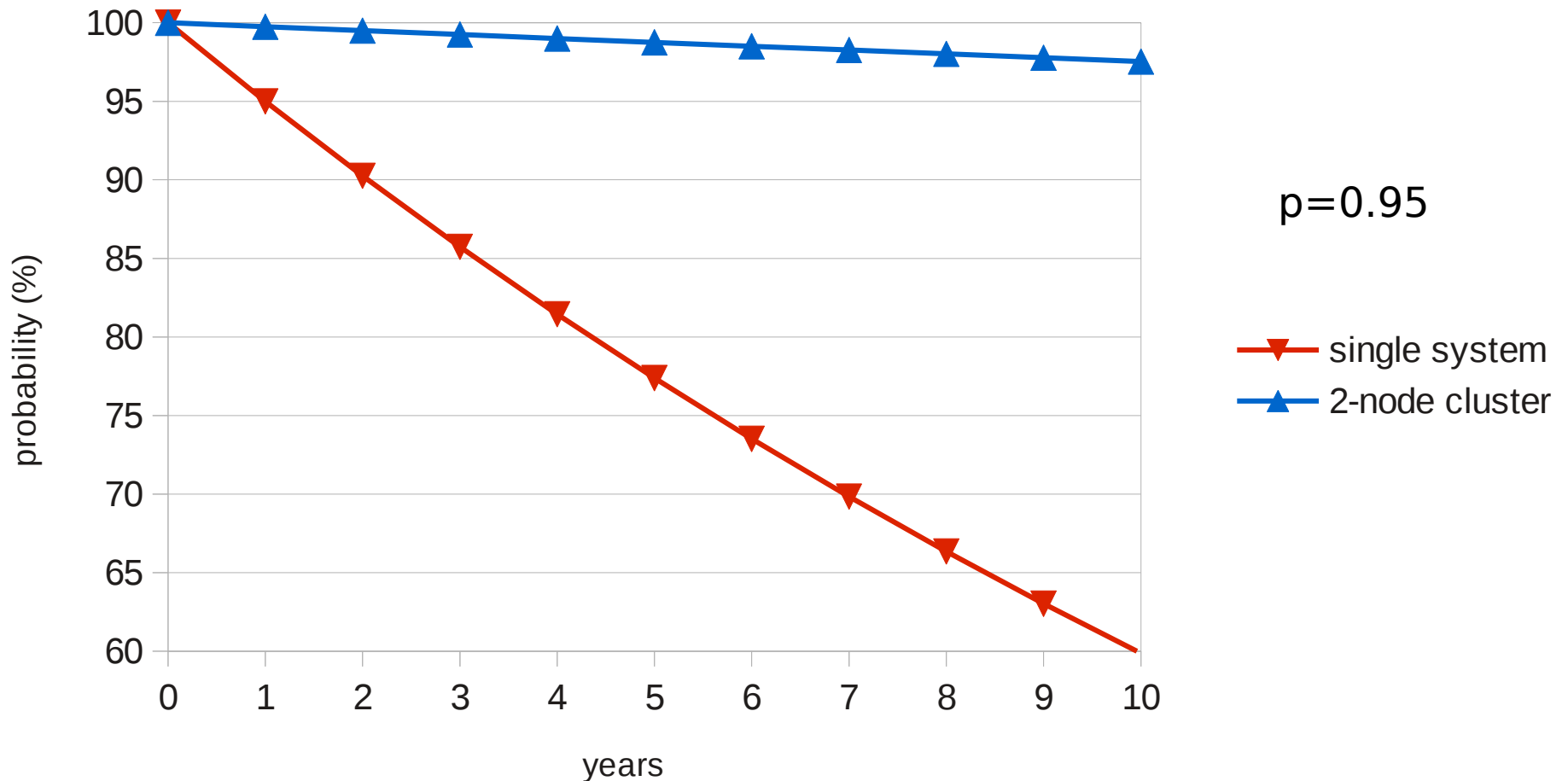
Availability Class	Availability %	Downtime per Year
1	90%	< 36.5 days
2	99%	< 3.65 days
3	99.9%	< 8.77 hours
<b>4</b>	<b>99.99%</b>	<b>&lt; 52.6 minutes</b>
5	99.999%	< 5.26 minutes
6	99.9999%	< 31.6 seconds
7	99.99999%	< 3.16 seconds



# Class Feasibility

Binomial distribution  
(Bernoulli trials)

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$







# DEFINITION: Highly Available System

The computer system is highly available when it has two following properties:

- no component is a single point of failure;
- overall, it is reliable enough that it can be repaired before something else breaks.

*Tone Vidmar; Informacijsko-komunikacijski sistem*





# Mean Time Between Failures

$$MTBF = \frac{\sum \text{operating time}}{\text{failures}}$$

Component	MTBF (h)	MTBF (years)
Disk drive; e.g., Seagate ST3450857SS	1 600 000	183
Interface card; e.g., LSI MegaRAID SAS 9280-16i4e SGL	919 964	105
Cooling fan	250 000	29



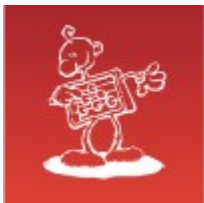


# MTBF of a Single System

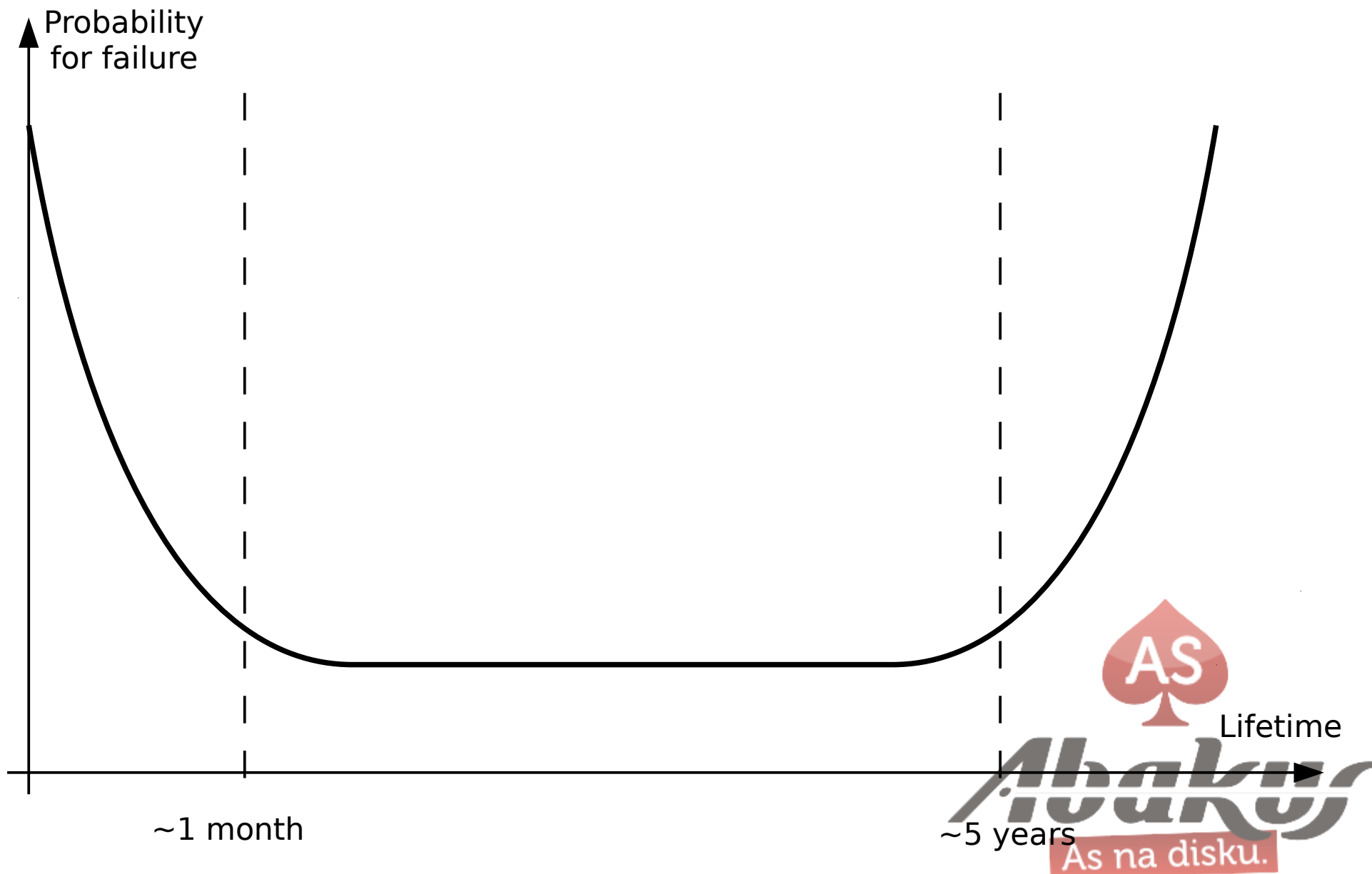
Component	MTBF (h)	MTBF (years)
Server board	208 000	24
Backplane board	760 000	87
Power supply	400 000	46
Cooling fans	100 000	11
PCI riser card	4 100 000	468
Front panel board	16 000 000	1825
All together	52 800	6

$$\frac{1}{MTBF} = \sum_n \frac{1}{MTBF_n}$$





# Probability of Failure





# Annual Failure Rate

$$AFR \approx \frac{1}{MTBF_{in-years}}$$

Component	MTBF (h)	MTBF (years)	AFR
Disk drive	1.600.000	183	0,0055
Interface card	919.964	105	0,0095
Cooling fan	250.000	29	0,035

AFR can be interpreted as a probability that the component fails within one year.





# Annual Failure Rate

$$AFR \approx \frac{1}{MTBF_{in-years}}$$

Component	MTBF (h)	MTBF (years)	AFR
Disk drive	1.600.000	183	0,0055
Interface card	919.964	105	0,0095
Cooling fan	250.000	29	0,035

$$\text{expected\_failures} = \sum_{\text{components}} AFR * \text{time}_{in-years}$$

For example:

- SAN system with 200 disk drives:  
**expected failure rate is 1.1 disk drives per year!**





# Mean Time to Repair

MTTR of a disk drive in a RAID configuration:

- 8h – with a hot-spare
- 48h – without a hot-spare

$$AFR_{RAID5} \approx \frac{n(n+1)}{2} \frac{MTTR}{MTBF^2} \cdot 8760$$

$$AFR_{RAID10} \approx n \frac{MTTR}{MTBF^2} \cdot 8760$$

\* from [2] but not entirely true, did not consider that neighbouring time slot failures are problematic as well





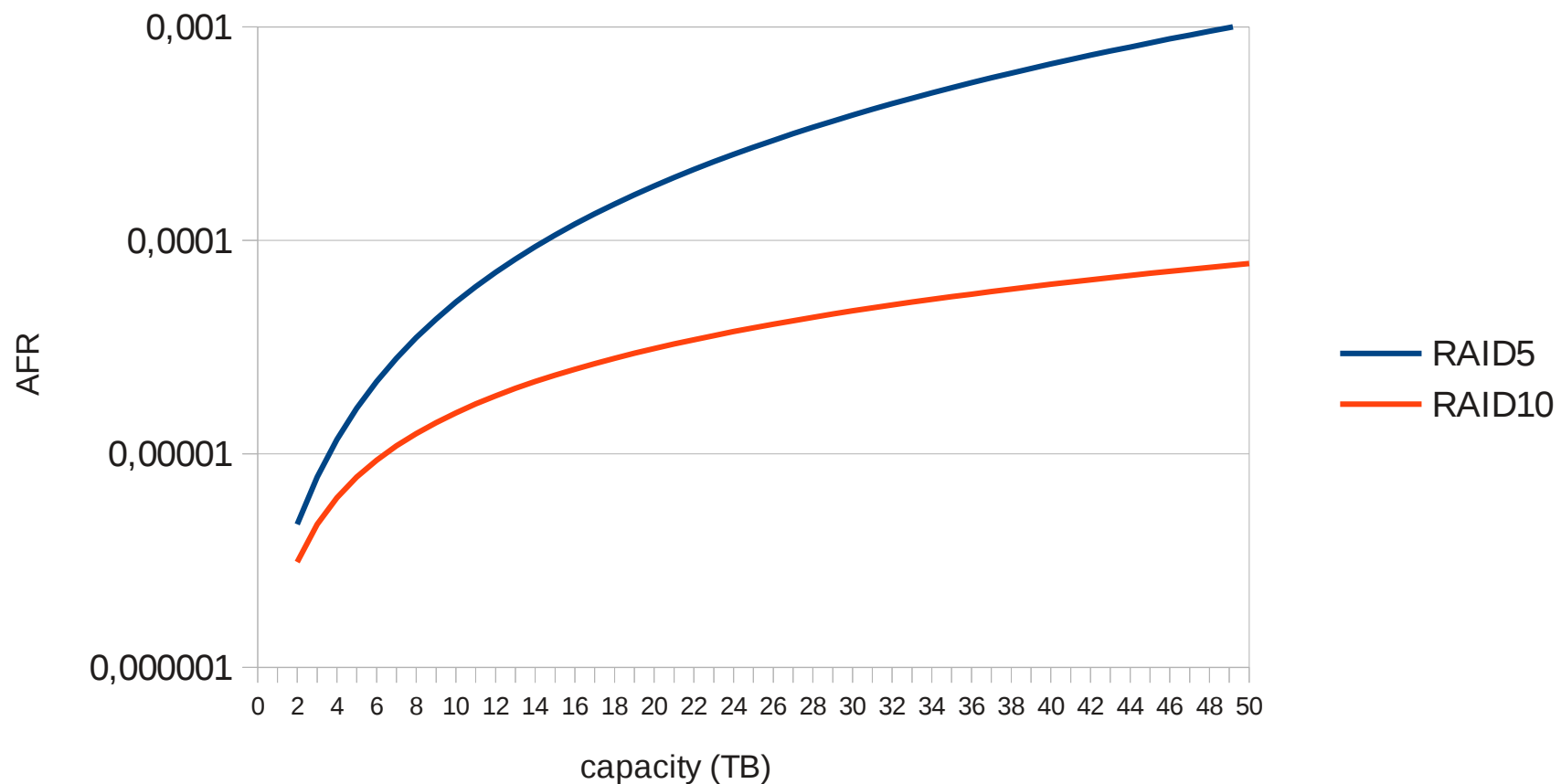


# MTBF, MTTR & AFR

MTBF = 300 000 h

MTTR = 8 h

using 1 TB disk drives





# About Disk Drives

- Array disk failures are highly correlated, making RAID 5 two to four times less safe than assumed
- Enthusiast-oriented reports such as AnandTech, quoting an »8% chance of complete data loss using RAID 5 with 200GB spindles«
- The »father of DEC StorageWorks« (now HP EVA) quoting that »If you have one petabyte of desktop drives with RAID 5, you could lose data twice a year«





# Simultaneous Multiple Failures

## Birthday paradox

$$p(n) = 1 - \frac{n! \binom{t}{n}}{t^n}$$

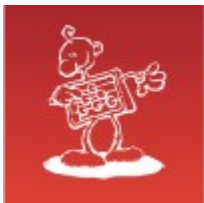
n ... # of components, e.g. disk drives

t ... # of MTTR intervals in lifetime

$$t = MTBF / MTTR$$

lifetime	MTTR	t	n	probability
10 years	48 h	1 826	50	49.2%
10 years	8 h	10 957	100	36.4%
30 years	48 h	5 479	50	20.0%
30 years	8 h	32 872	50	3.66%
birthdays:				
1 year	1 day	365	23 people	50.7%

As na disku.



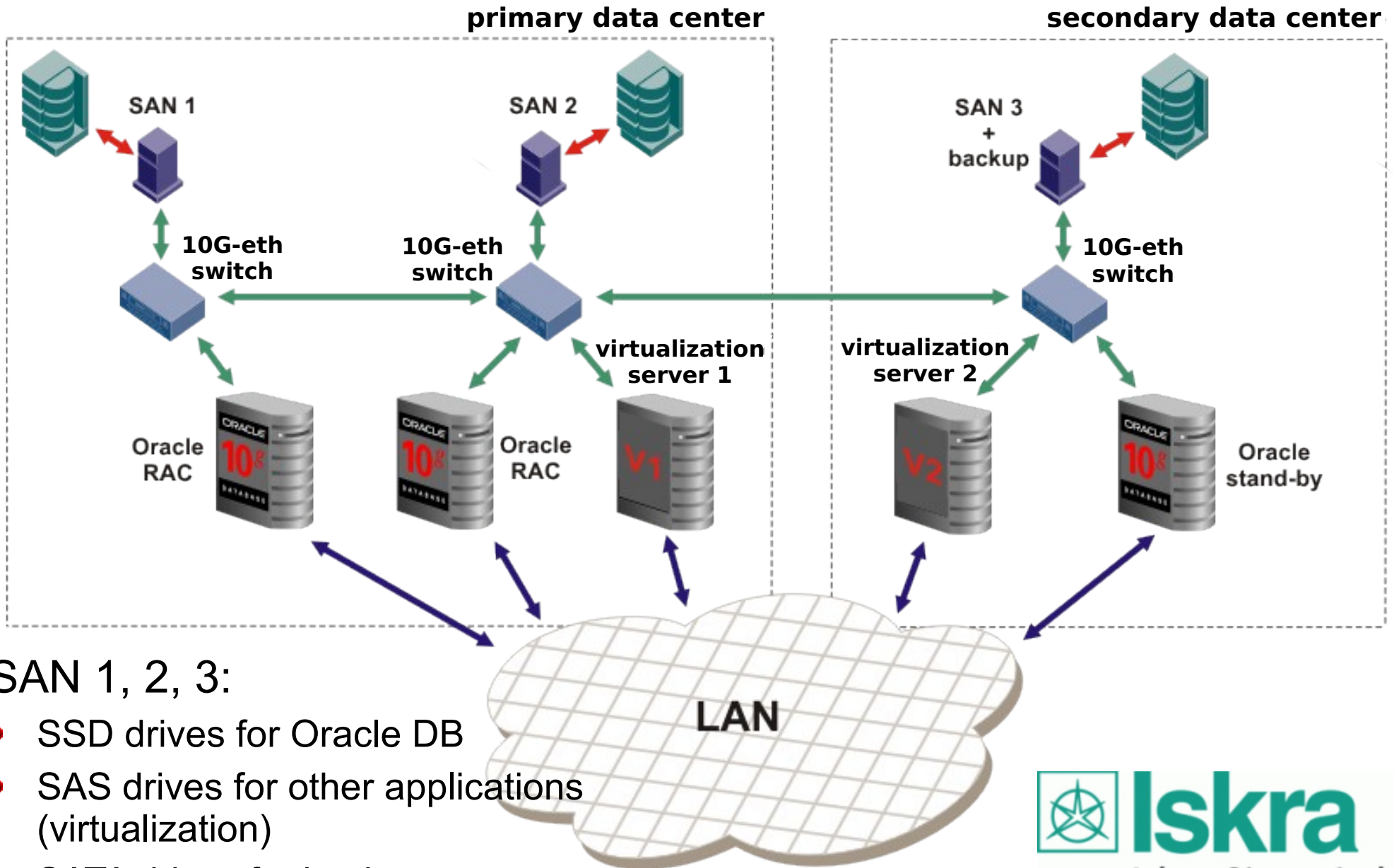
# Guidelines

- Make full failure mode analysis of your system(s).
- The AFR for RAID group should be smaller than  $10^{-4}$ .  
This maximizes the number of disks to 8 (7 + 1) for RAID5 and 16 (8 + 8) for RAID10.  
(Google claims: AFR of a single disk drive is >4%)
- Use hot-spare disks.





# IT Infrastructure



SAN 1, 2, 3:

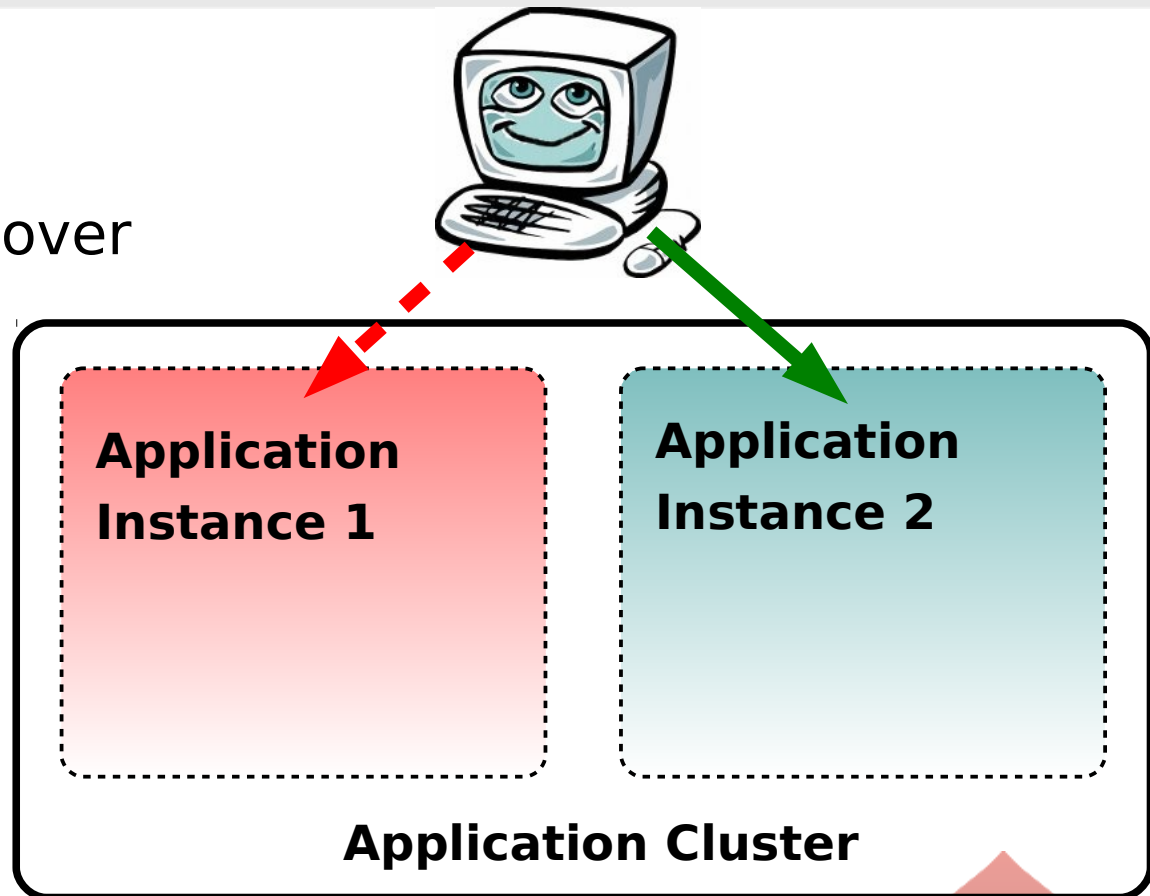
- SSD drives for Oracle DB
- SAS drives for other applications (virtualization)
- SATA drives for backup



# High Availability - Traditional Way

Clustering - Grid

Application service failover



Physical Hosts





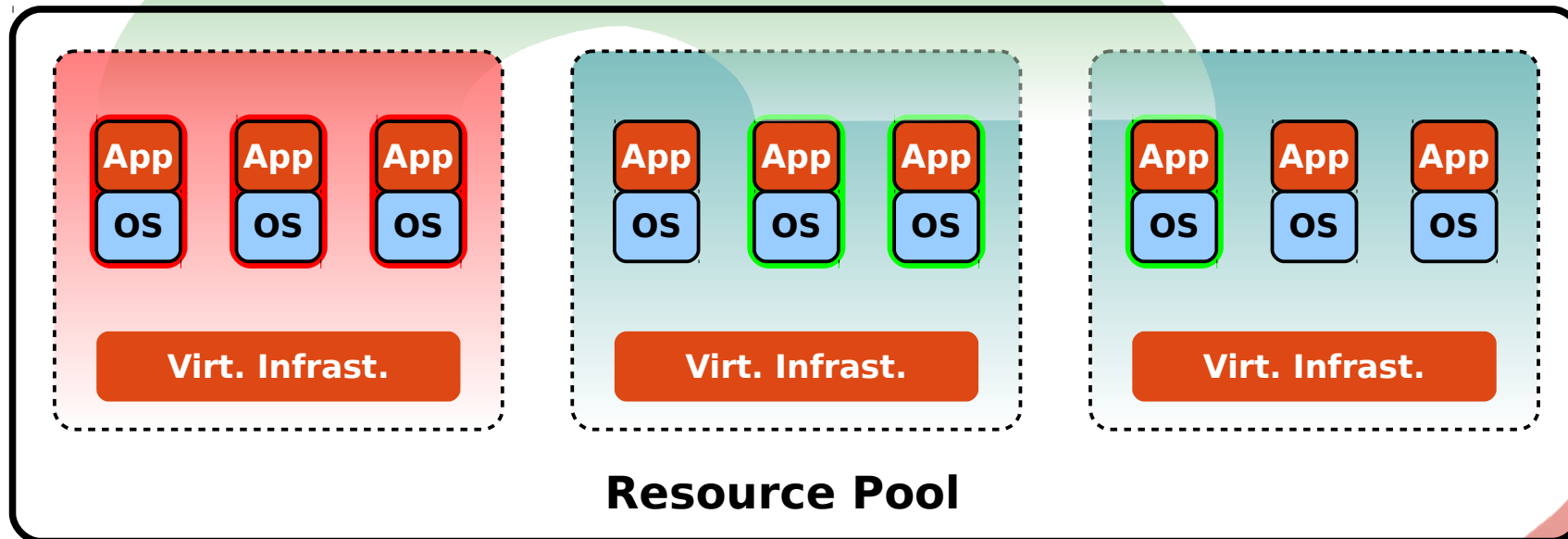
# High Availability – Inventive Way

Virtualization – Cloud

Transparent system failover

- XEN – remus
- KVM – kemari

failover



Physical Hosts







# HA Work in Progress

- massive virtualization – cloud: using low cost commodity hardware
- disaster site: on-line transparent stateful switchover/failover (and switchback again) – as many times as you like 🤪





# References

[1] Intel; *Calculated MTBF Estimates*

(<http://download.intel.com/support/motherboards/server/sb/s3420gpmtbfcaculationrev10.pdf>)

[2] Klaus Schmidt; *High Availability and Disaster Recovery: Concepts, Design, Implementation*

[3] Tone Vidmar; *Informacijsko–komunikacijski sistem*

[4] Eduardo Pinheiro, Wolf-Dietrich Weber, Luiz Andre Barroso – Google Inc; *Failure Trends in a Large Disk Drive Population*

([http://research.google.com/archive/disk\\_failures.pdf](http://research.google.com/archive/disk_failures.pdf))

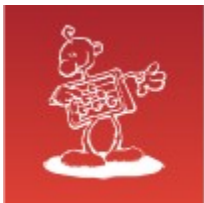
[5] Johan De Gelas; *Server Guide part 2: Affordable and Manageable Storage*

(<http://www.anandtech.com/show/2105/5>)

[6] Robin Harris; *StorageMojo*

(<http://storagemojo.com/>)





# High Availability in Numbers

## Questions

**mag. Sergej Rožman**

ABAKUS plus d.o.o.

Ljubljanska c. 24a

Kranj

e-mail: [sergej.rozman@abakus.si](mailto:sergej.rozman@abakus.si)

phone: 04 287 11 14

