Vrhunsko zmogljiv hardver na odprtokodni osnovi

- **Sergej Rožman**, univ.dipl.inž.; Abakus plus d.o.o.

- Zadnja verzija z morebitnimi spremembami je na naslovu: http://www.abakus.si/

Tax Department found inefficient

Abakus
As na disku.

Vrhunsko zmogljiv hardver na odprtokodni osnovi

Sergej Rožman, univ.dipl.inž.

sergej.rozman@abakus.si

Sava
Gorenjska Banka
Banka s poslabom
Aerodrom Ljubljana
Mercator
GOODYEAR
BANKA SLOVENIJE
EVROSISTEM
MERKUR
SPL
ELES
futuraplus
Iskra
Iskra MIS
DELO PRODAJA
jata emona

O podjetju

Zgodovina
- od 1992, 20 zaposlenih

Lastne aplikacije:
- posebne (letalski prometni sistem, distribucija časopisov)

Storitve:
- DBA, vzdrževanje operacijskih sistemov, programiranje (MediaWiki, Oracle)
- omrežja (omr. storitve, VPN, QoS, varnost)
- odportokodne rešitve, nadzorni sistemi (nagios, OCS, wiki)

Okolje:
- od 1995 GNU/linux ***(14-let izkušenj !)***
- prenos Oracle na GNU/linux: RDBMS 7.1.5 in forms 3.0 ***(pred Oraclom !)***
- **skoraj 20 let izkušenj s sistemi za visoko razpoložljivost !**

Aerodrom Ljubljana · BANKA SLOVENIJE · datalab · DELO PRODAJA · futuraplus · gea · Gorenjska Banka d.d. Kranj · Iskra Iskra MIS, d.d. · jata emona · LASERLINE · Mercator · MERKUR · Sava · SPL LJUBLJANA d.d.

O podjetju

Slika: ekspedit v podjetju FuturaPlus v Beogradu

Aplikativna rešitev za distribucijo časopisnih edicij

AS Abakus As na disku.

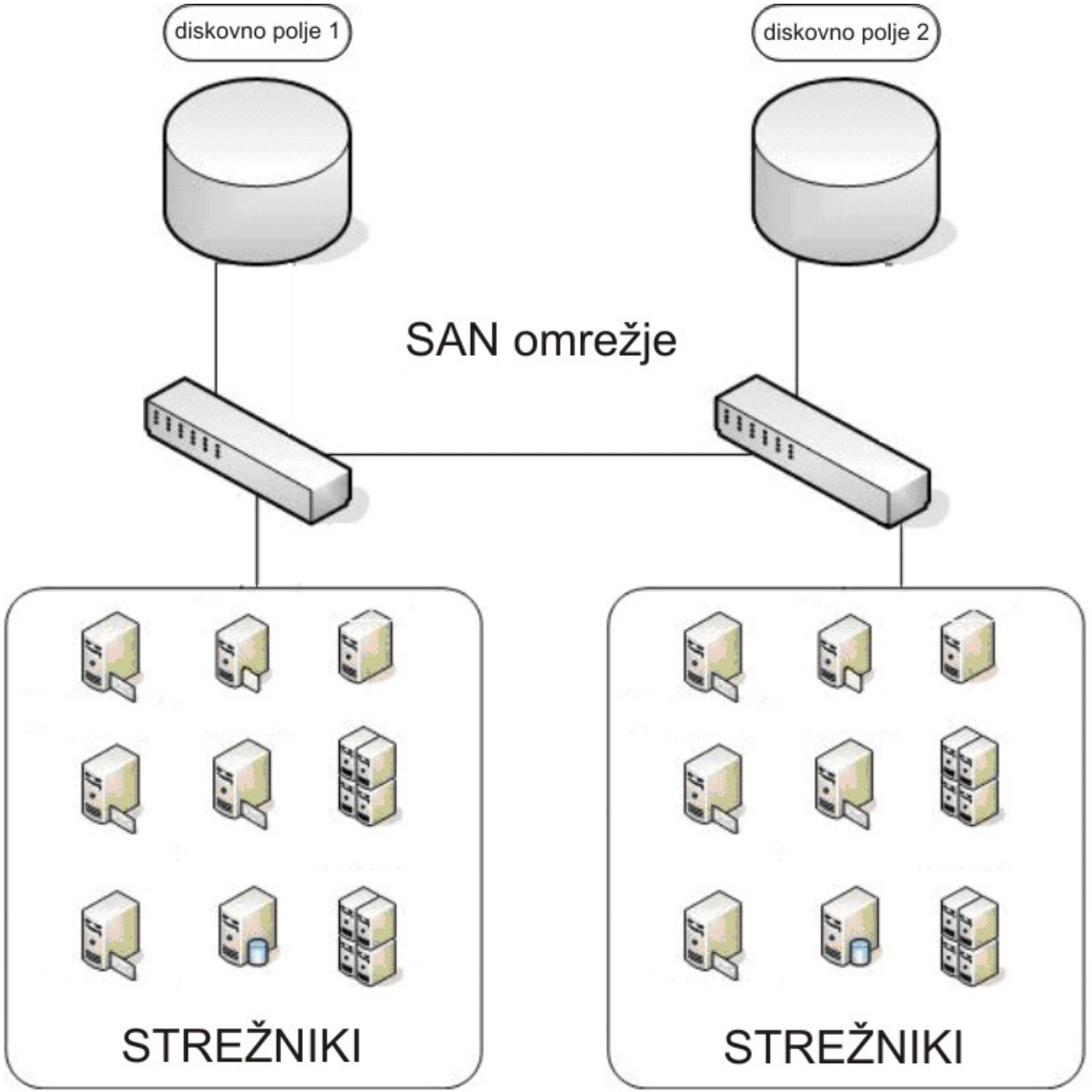
O podjetju

Prometni sistem za letališča (slika iz aplikacije)

Srebrno priznanje za inovacijo

Zanimivost: strežniki na Aerodromu Ljubljana neprekinjeno delujejo že **1221** in **1136** dni (dne: 11. 9. 2009)

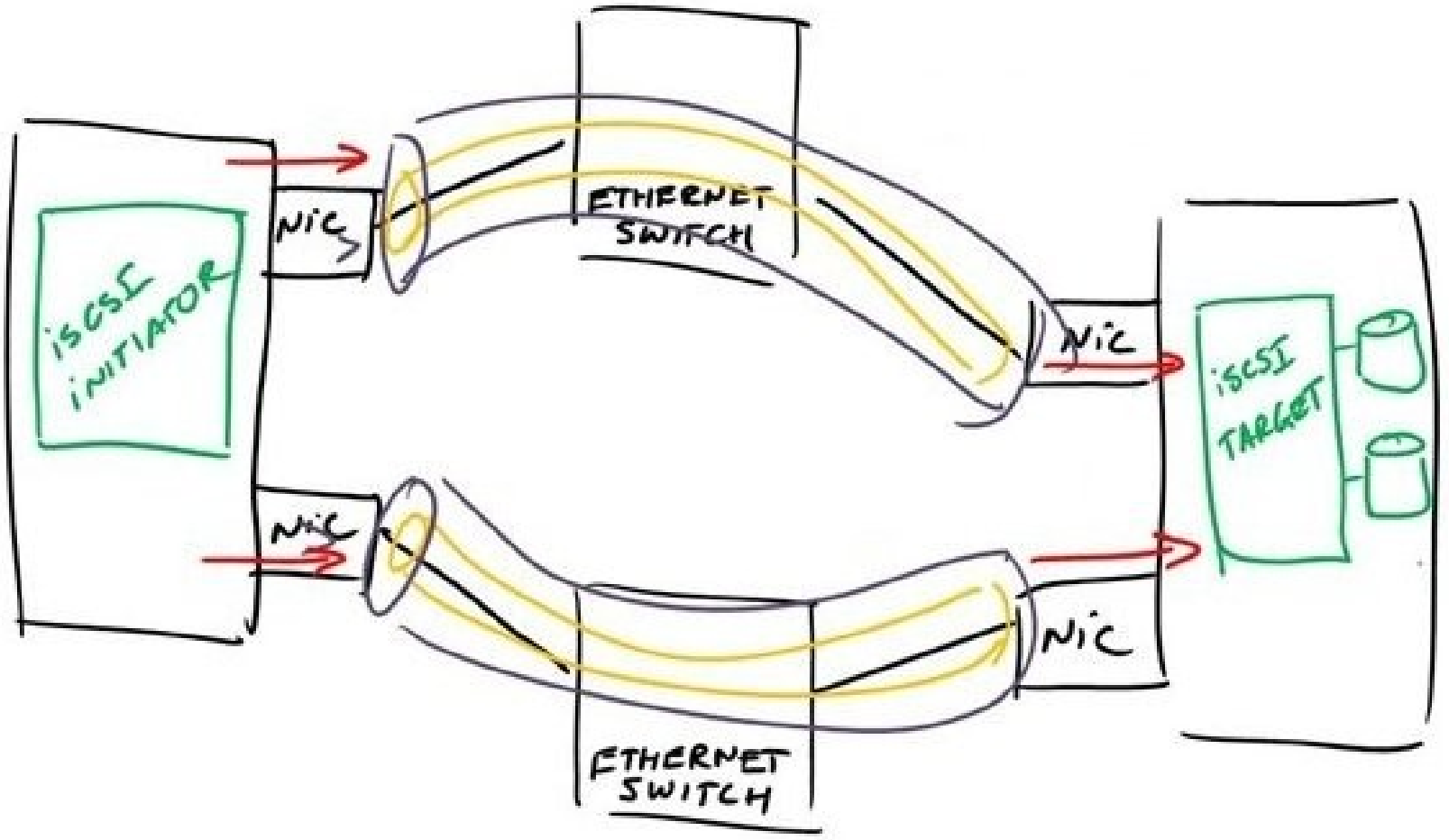
Cilj – zmogljiv hardver zasnovan na odprtih standardih in odprtokodnih tehnologijah

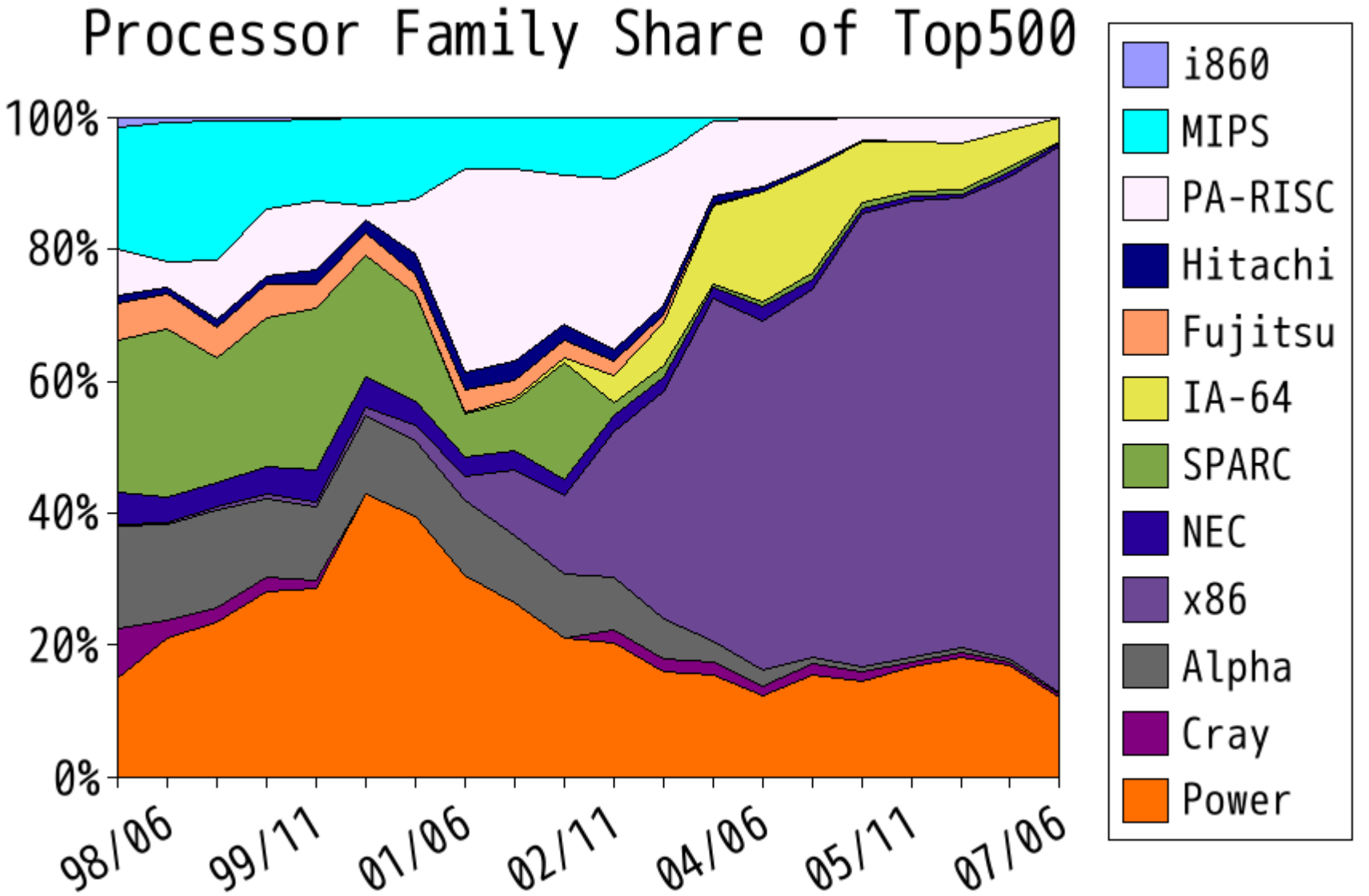


Uporabljene tehnologije:

- operacijski sistem GNU/linux,
- iSCSI enterprise target – IET (iSCSI strežnik),
- LVM,
- mrežno zrcaljenje – DRBD,
- iSCSI initiator (iSCSI odjemalec),
- device mapper,
- gručni datotečni sistem – OCFS, GFS,
- udev.

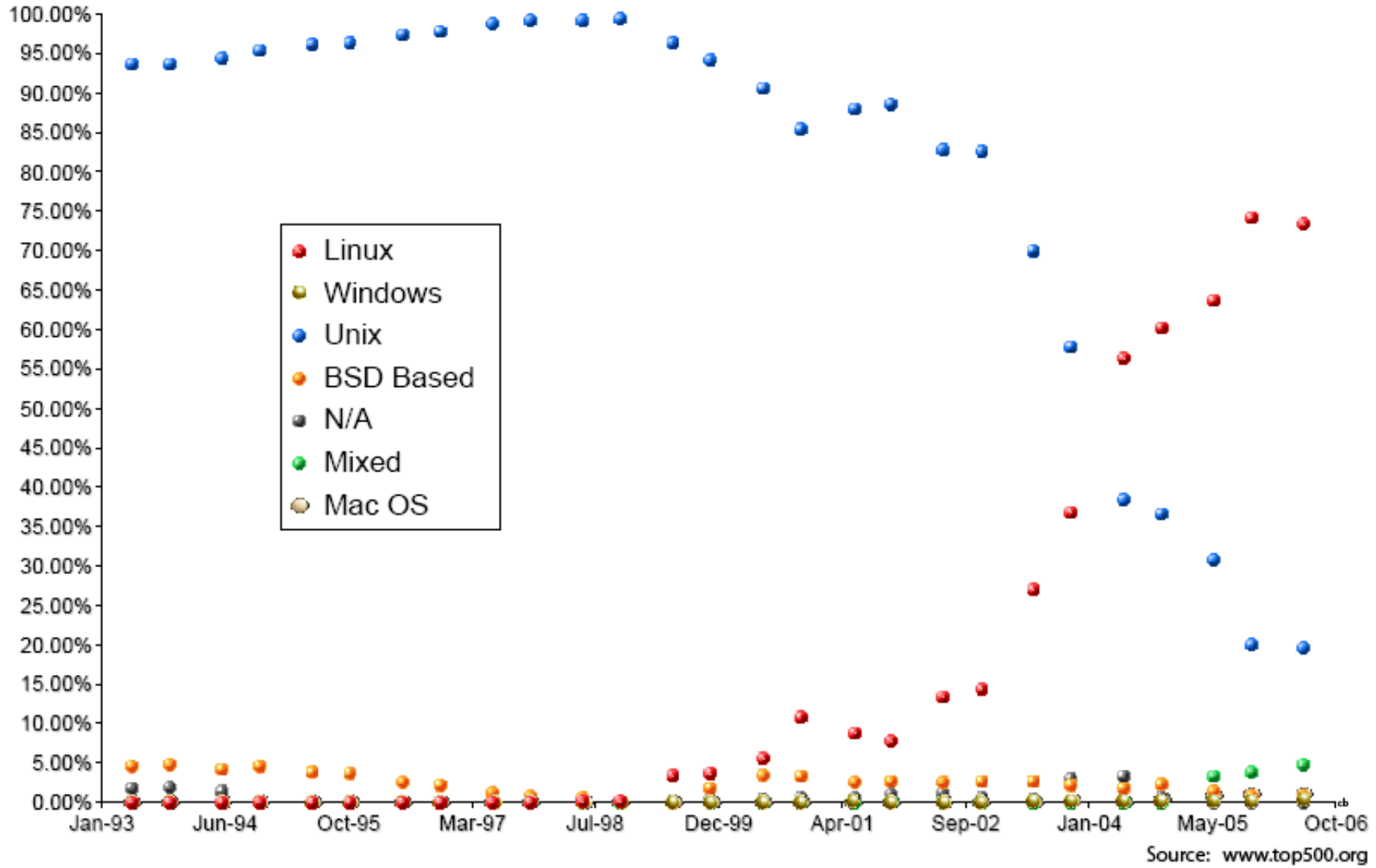
iSCSI





Source: http://www.top500.org

Operating Systems Used On Top500 Supercomputers



Source: www.top500.org

Primerjava lastnosti: fibre channel - iSCSI

fibre channel

- posebna omrežna oprema
- ponavadi izvedba z namensko strojno opremo
- slabša združljivost naprav (ohlapni standardi)
- običajna izvedba z optičnimi vodi
- **počasnejši prenos**
- srednje razdalje na lastniškem omrežju
- draga rešitev
- v Sloveniji zelo razširjeno (HP EVA)

iSCSI (ethernet)

- standardna omrežna oprema
- v strojni ali programski izvedbi
- z združljivostjo ni težav
- običajna izvedba z bakrenimi vodi
- **hitrejši prenos**
- tudi velike razdalje, tudi po javnem omrežju
- poceni rešitev
- v Sloveniji ni poznano in razširjeno

Primerjava zmogljivosti: fibre channel - iSCSI

fibre channel

- 1 Gb/s (1997)
- 2 Gb/s (2001)
- 4 Gb/s (2005)
- 8 Gb/s (2008)
- 16 Gb/s (2011 plan.)

Vir: http://www.fibrechannel.org/OVERVIEW/Roadmap.html

iSCSI (ethernet)

- 10 Mb/s (1980)
- 100 Mb/s (1995)
- 1 Gb/s (f:1998/tp:1999)
- 10 Gb/s (f:2002/cx4:2004/tp:2006)
- 40/100 Gb/s (2010 plan.)

Vir: wikipedia

10 Gb ethernet

Bakreni vodi

- 10GBASE-CX4 (podoben kot InfiniBand) – do 15 m
- 10GBASE-T (cat. 6a) – do 100 m

Optični vodi

- od 300 m do 80 km, odvisno od vodov (SM, MM) in optičnih modulov

Proizvajalci

- Myricom
- Chelsio
- Neterion
- NetXen

AS Abakus
As na disku.

10 Gb ethernet

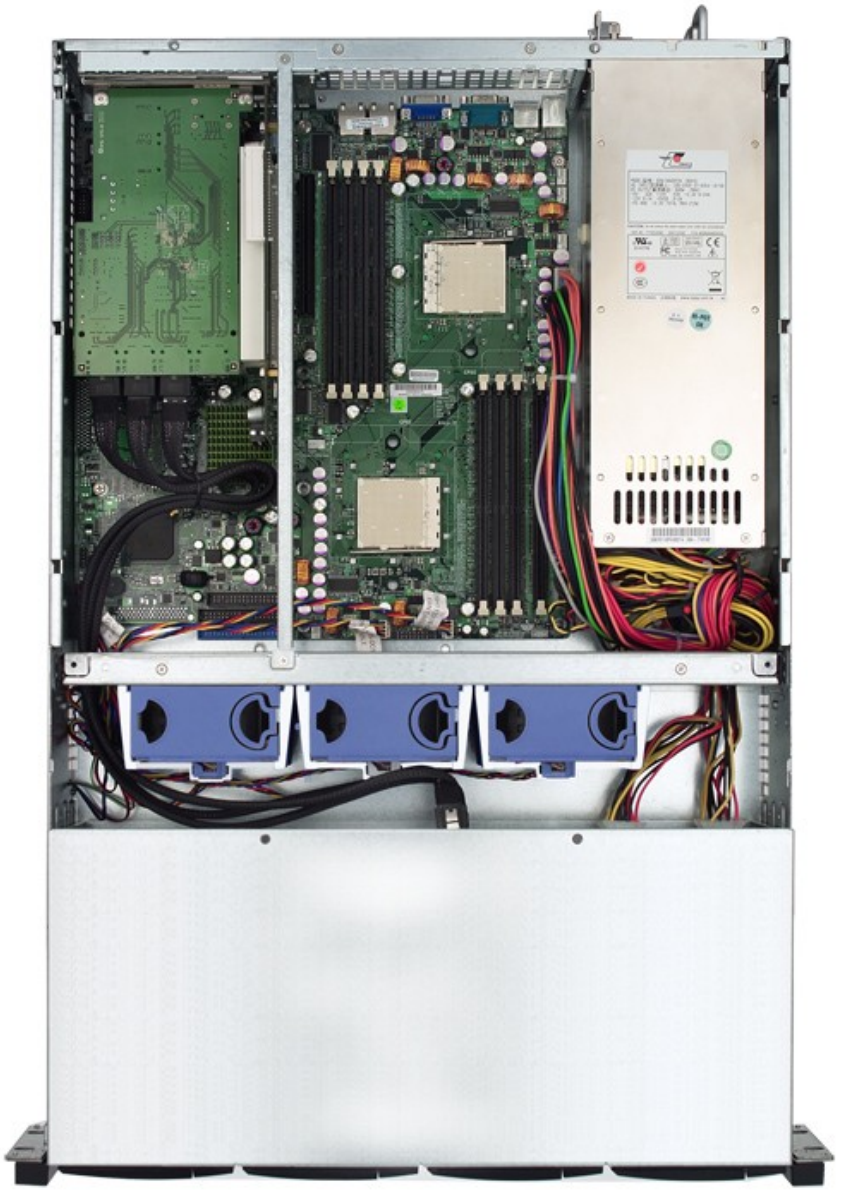
Izkušnje:
10 GbE iSCSI pri polni hitrosti polno obremeni dva procesorja ali dve procesorski jedri

Posebne nastavitve (nekaj primerov)

```
Parametri linux jedra (/proc)
   net.ipv4.tcp_timestamps = 0      (izklop časovnih oznak)
   net.ipv4.tcp_sack = 0            (izklop selektivnega potrjevanja)
   net.*.*mem* = 10000000           (dovolj veliki izravnalni pomnilniki)
   net.core.netdev_max_backlog = 300000

Mrežne nastavitve (ethtool)
   več procesiranja na prekinitev   (ethtool -C $DEVNAME rx-usecs $VALUE)
   MTU = 9000                       »jumbo« okvirji


Sistemske nastavitve
   TCP segmentation offload         (nastavitev gonilnika)
   izklop irq_balancerja            (prekinitve vedno obravnava isti procesor)

Druge nastavitve po priporočilu proizvajalca kartic
```

HARDVER

Osnovni model



Modularna zasnova

Moduli:

< 2U – 12 diskov

1U – krmilnik >

< 3U – 16 diskov

< 5U – 24 diskov

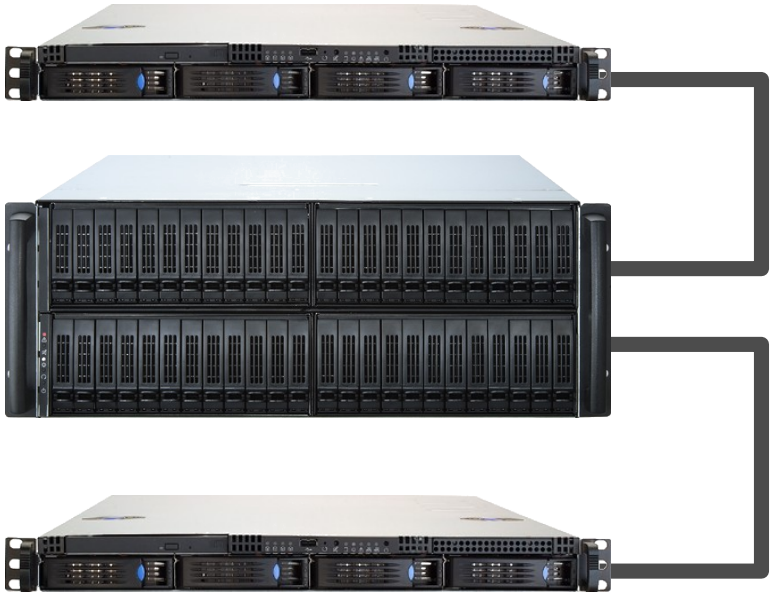
2U – 24 2,5“ diskov >

4U – 48 2,5“ diskov >

< 9U – 50 diskov

As na disku.

Redundantna zasnova

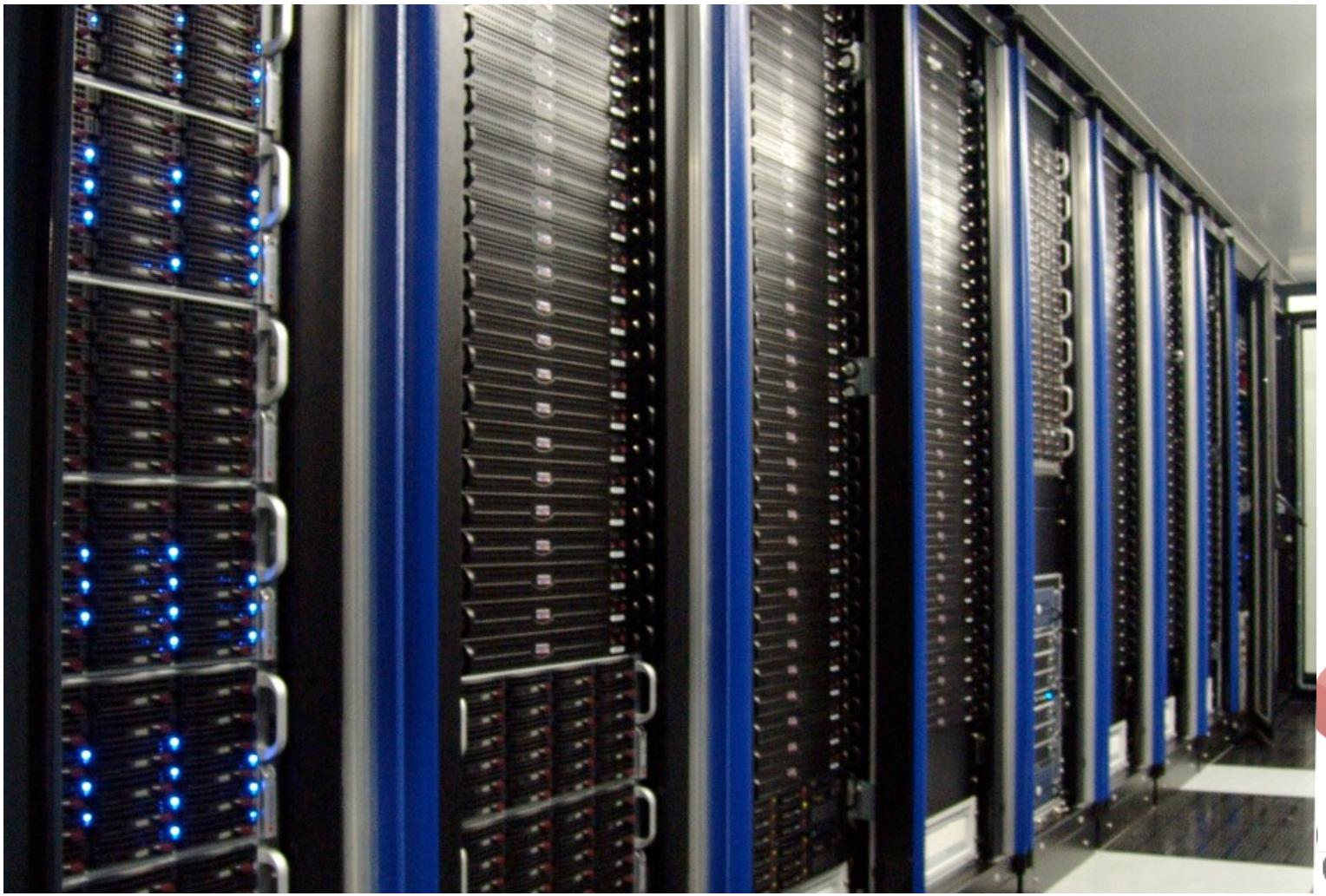


Dvojna povezava do istih diskov

- dvokanalni SAS
- multi-initiator SCSI

Abakus
As na disku.

Sestavljeno skupaj



AS

aku

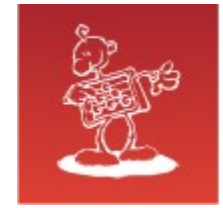
As na disku.


SOFTVER
SOFTPEDIA
AS
Abaky
As na disku.

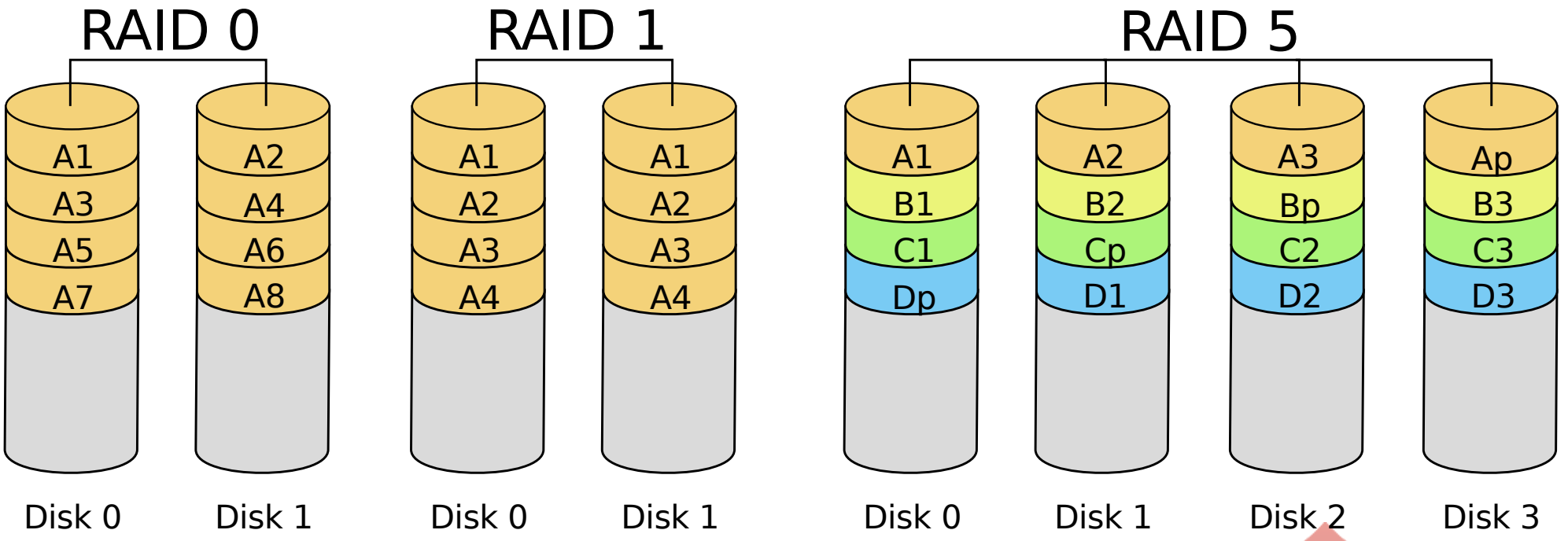
iSCSI enterprise target (IET)

The iSCSI Enterprise Target Project

- RFC 3720
- http://iscsitarget.sourceforge.net/
- dinamično konfiguriranje (CLI)
- nima grafičnega vmesnika (GUI) **(še!)**



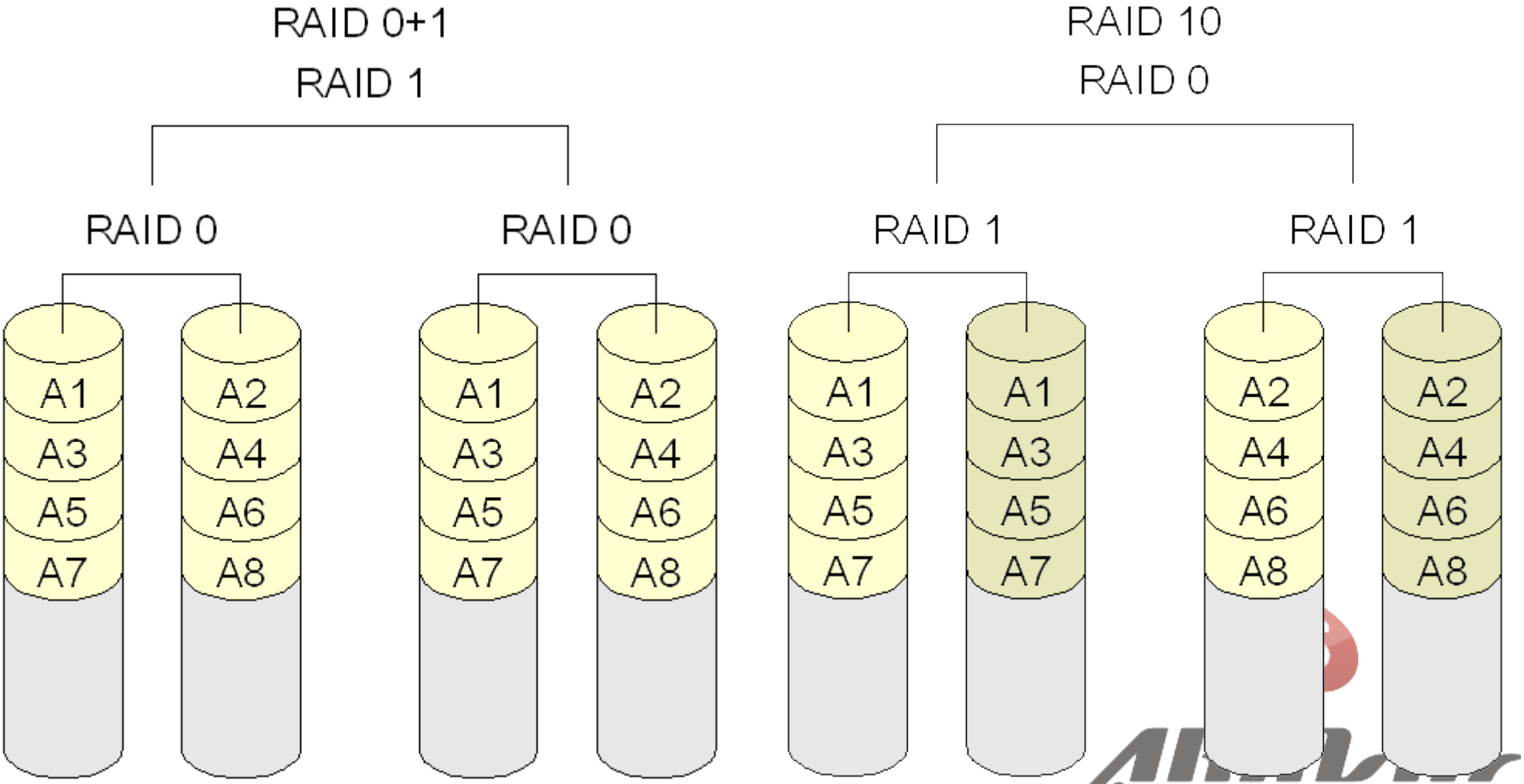
RAID (Redundant Array of Inexpensive Disks)



Vir: wikipedia

BAARF – http://www.baarf.com/
Battle Against Any Raid Five (Four, Free)

RAID – sestavljene stopnje



Vir: wikipedia

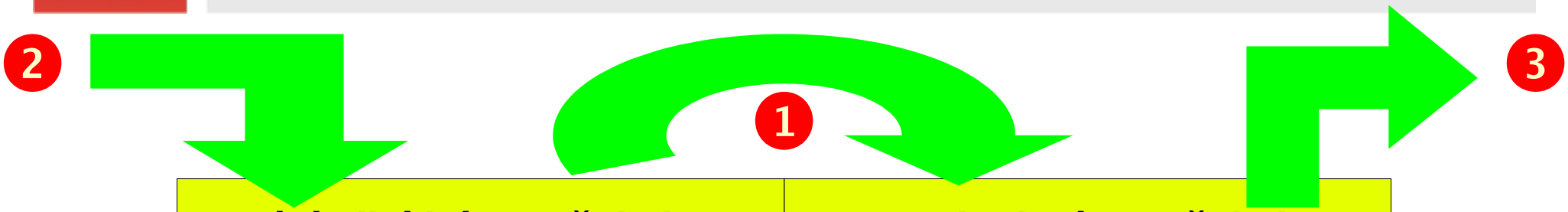
LVM (Logical Volume Management)

Katera funkcionalnost LVM je najbolj uporabna pri SAN?

- »SNAPSHOT«

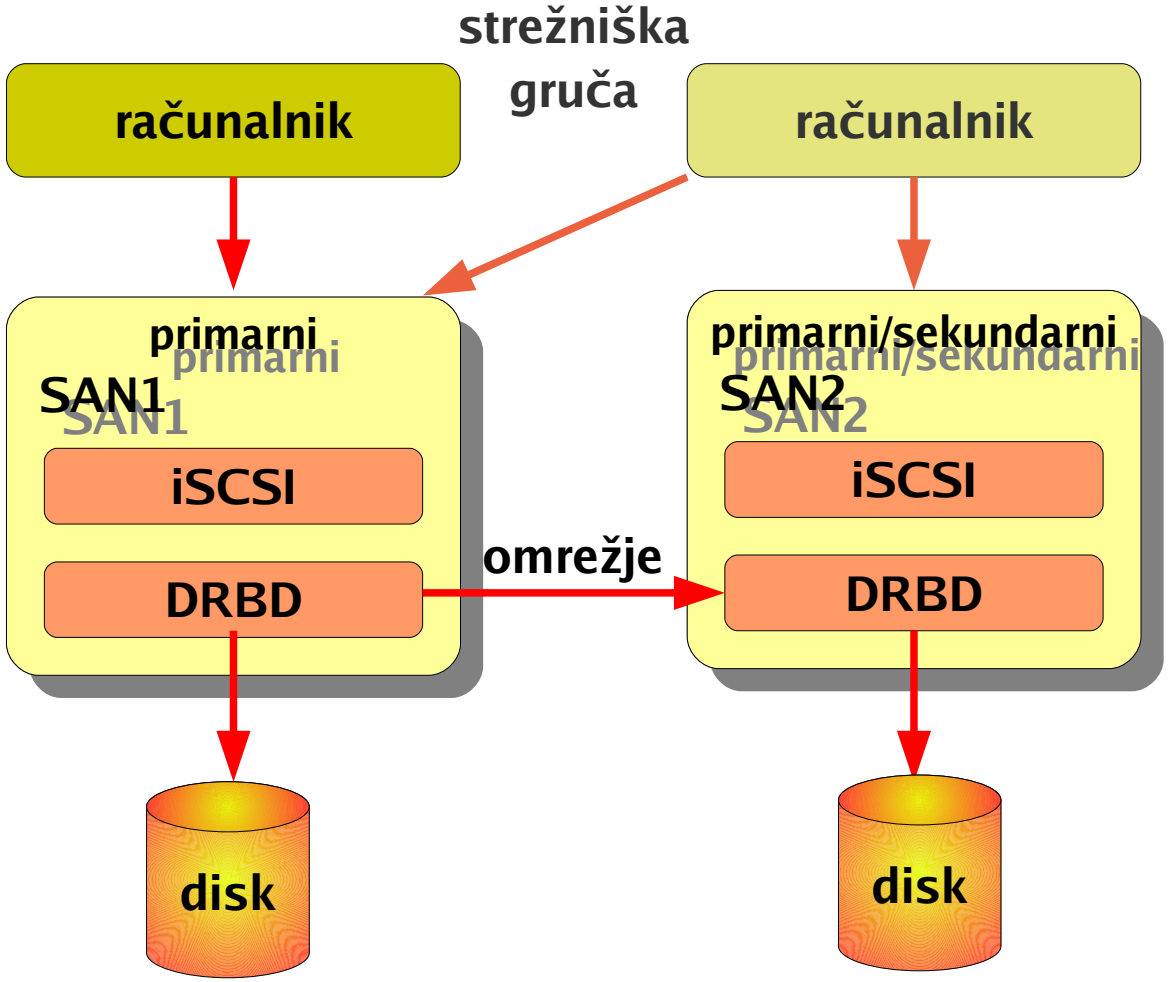
LVM »snapshot« navidezno v trenutku zamrzne stanje particije, ki jo zato lahko konzistentno prekopiramo/shranimo

LVM (Logical Volume Management)



produkcijski datotečni sistem /disk (OCFS2)	»zamrznjeni« datotečni sistem /backup (OCFS2)
logični disk/particija (LV) 500 GB	»snapshot« logični disk/particija (LV) 200 GB
logična grupa (LG) diskov 3 x 300 GB	
fizični disk (PV) 300 GB	fizični disk (PV) 300 GB	fizični disk (PV) 300 GB
disk1	disk2	disk3

DRBD – mrežno zrcaljenje
(Distributed Replicated Block Device – LINBIT)



- sinhrona replikacija
- asinhrona replikacija
- dvosmerna sinhrona replikacija!
  (od DRBD v. 8.0 dalje)

Podobna funkcionalnost: Oracle ASM (Automatic Storage Management) na oddaljen SAN – samo za potrebe Oracle zbirk podatkov

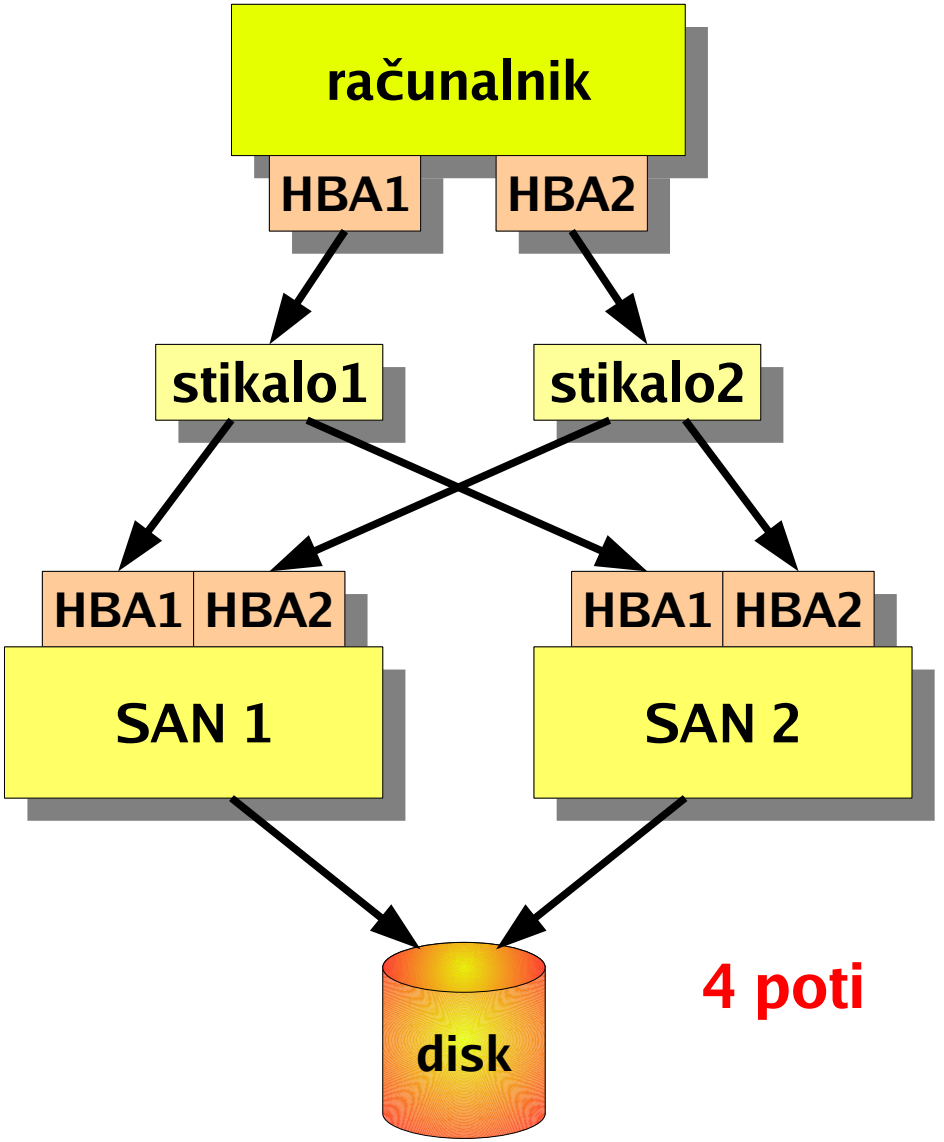
iSCSI initiator

iSCSI klient

- Open-iSCSI (http://www.open-iscsi.org/)
- Linux-iSCSI(sfnet) (http://linux-iscsi.sourceforge.net/)
  (starejši – performančne težave pri branju)

- drugi operacijski sistemi (windows, UNIX, ...)
- VMware ESX Server

Device mapper (multipath)



Uporabna funkcionalnost:

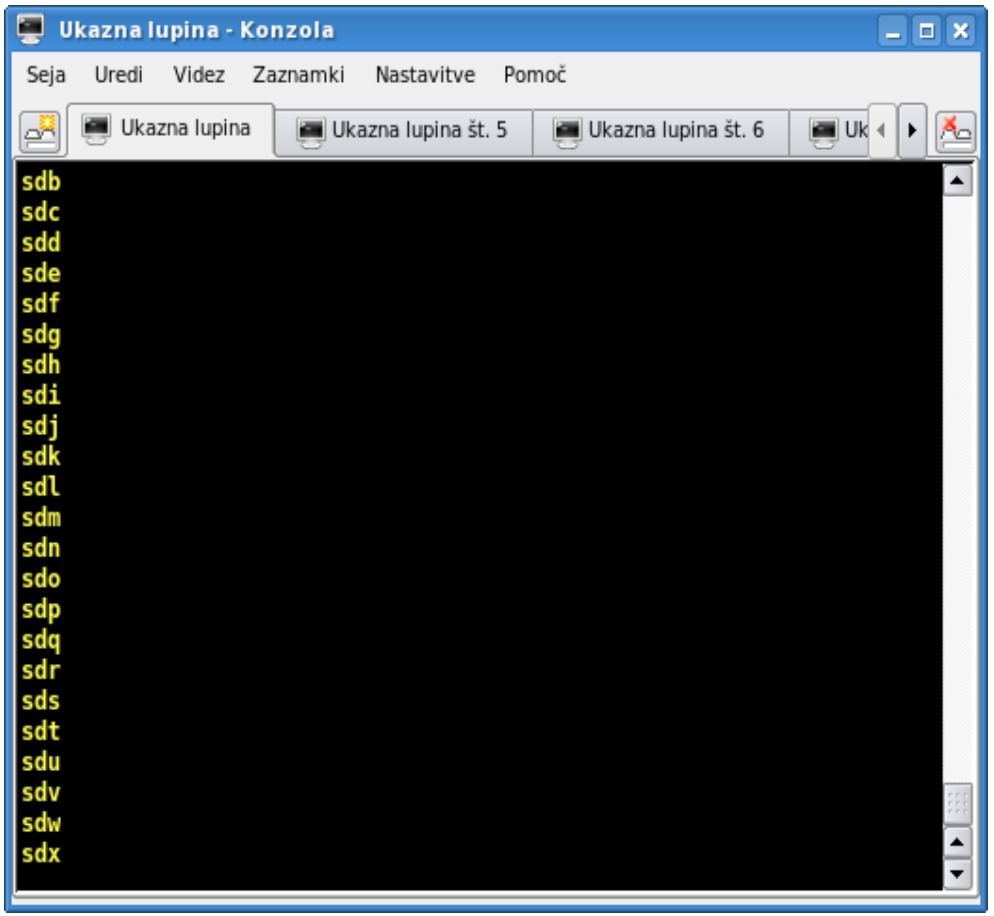
- »MULTIPATH«
  Več vzporednih (redundantnih) poti do istega diska.

Abakus
As na disku.

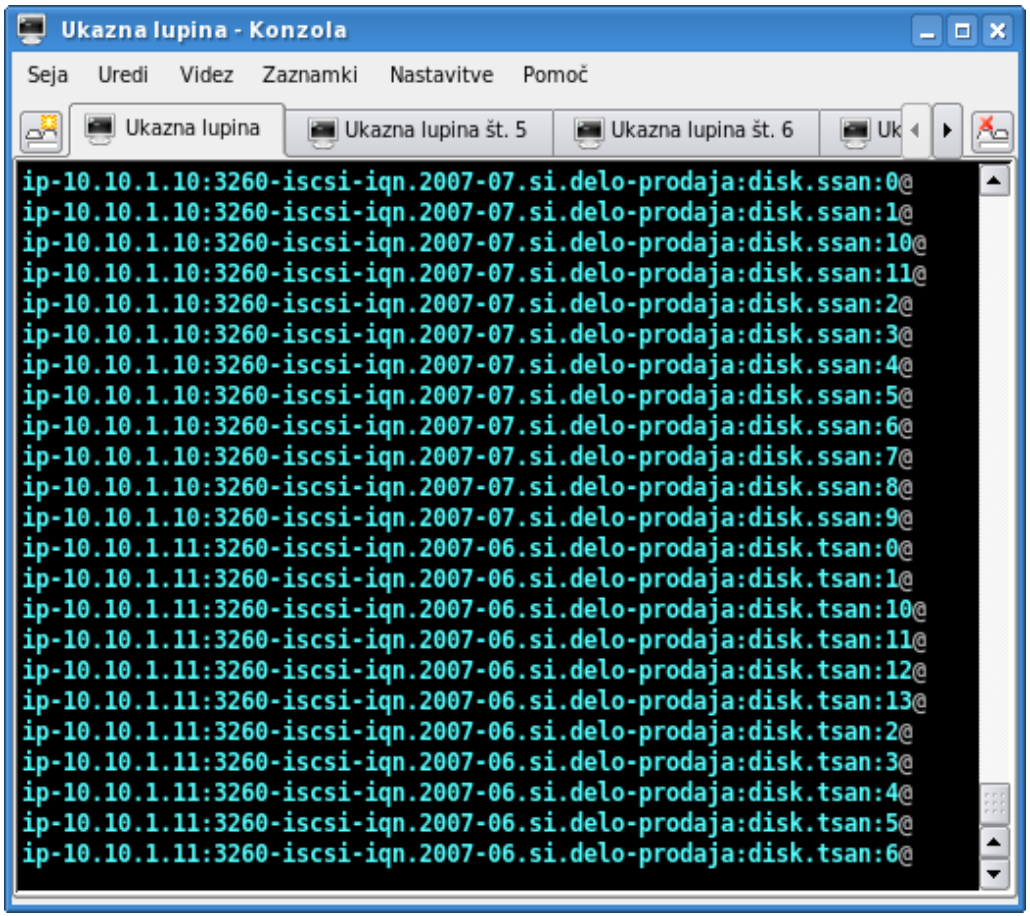
UDEV

- praktično neomejeno število naprav na /dev (dinamično)
- obstojno poimenovanje naprav

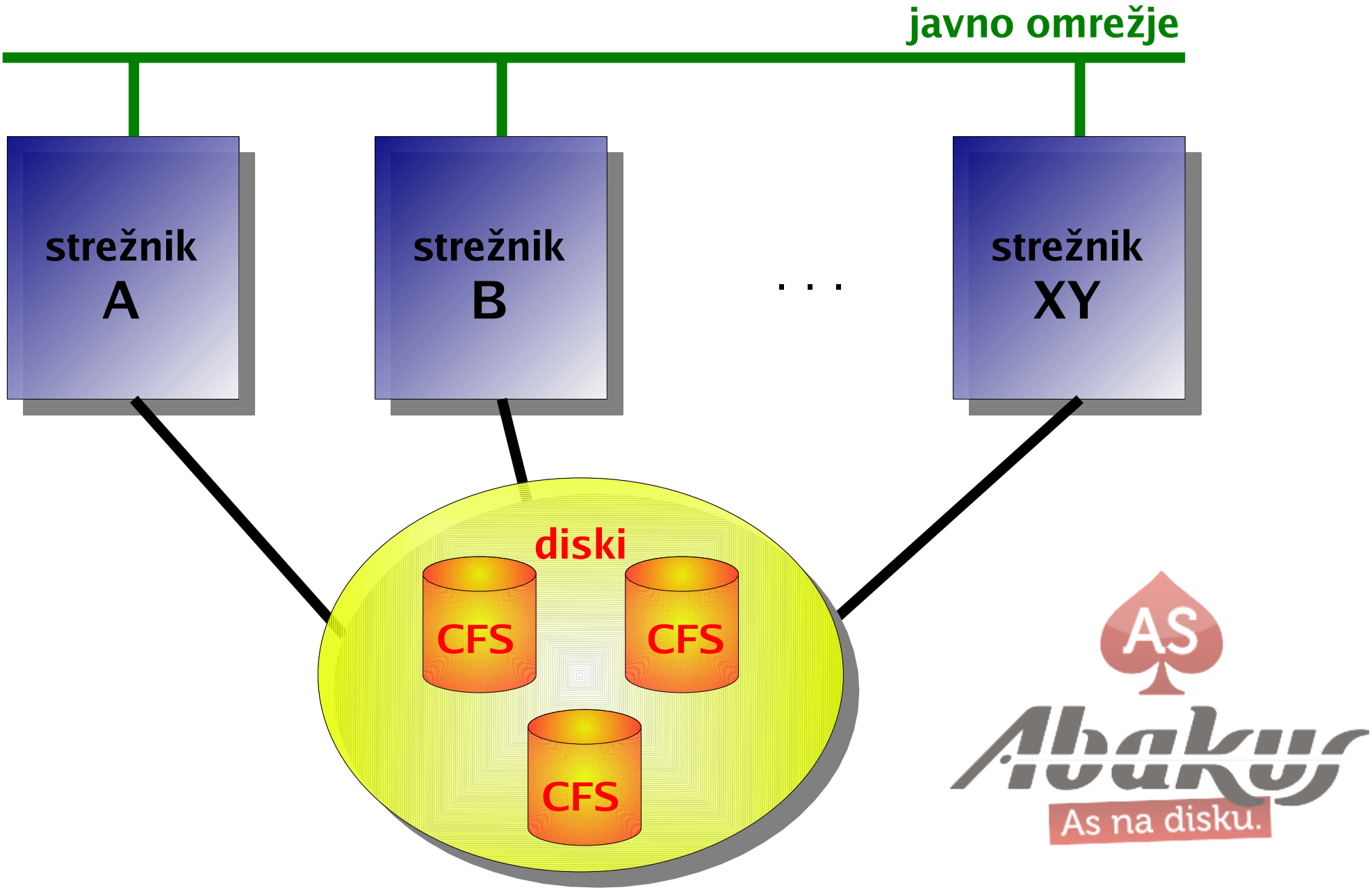
tradicionalno



udev



Gručni datotečni sistem (CFS)



Gručni datotečni sistem (CFS)

OCFS2 (Oracle Cluster File System)

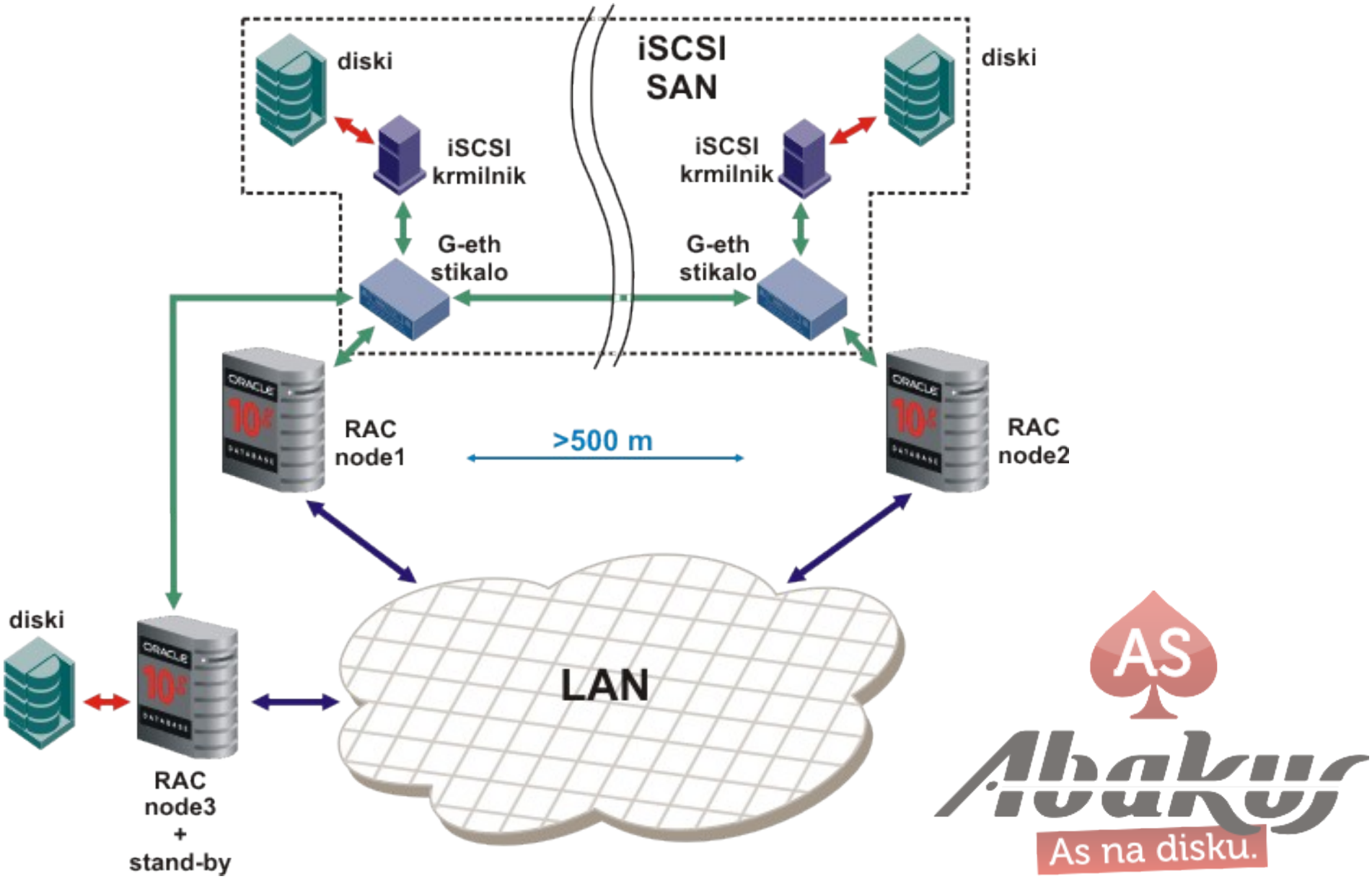
- standardni del GNU/linux jedra
  od verzije 2.6.16 (20. 3. 2006) – ekperimentalno in
  od verzije 2.6.19 (29. 11. 2006) – produkcijsko

GFS (Global File System – RedHat)

- standardni del GNU/linux jedra
  od verzije 2.6.19 (29. 11. 2006)

zasnovano na DEC tehnologiji

Uporaba: distribuirana strežniška gruča



Za konec

- učinkovito hranimo podatke:
  - zanesljivo
  - zmogljivo
  - poceni
- odprtokodne rešitve so preizkušene – na tisoče inštalacij po celem svetu
  (DRBD: >60.000 inštalacij, >20 TB diskovnega prostora)

Kako učinkovito hraniti podatke

Vprašanja

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

e-pošta: sergej.rozman@abakus.si

tel. št.: 04 287 11 14