Storage & Performance Graphs

System and Network Administration

Revision 2 (2020/21)

Pierre-Philipp Braun <pbraun@nethence.com>

Table of contents

RAID & Volume Managers

What matters more regarding the DISK resource?…

==> DISK I/O – the true bottleneck

some kind of a resource

- Input/output operations per second (IOPS)
- I/O PERFORMANCE is *the* resource one cannot easily scale
- Old SAN == fixed performance

Solutions

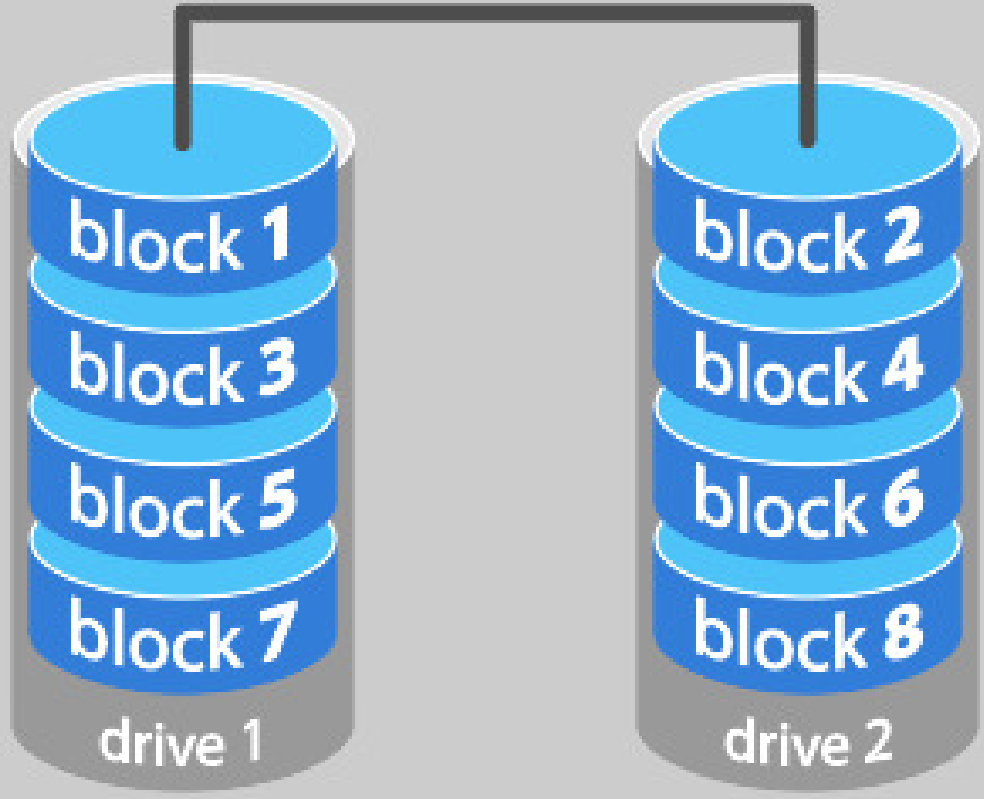
- –> NVMe/SSD or Hybrid SAN
- –> Hybrid SAN
- –> Software Defined Storage (SDS)

RAID TYPES

- RAID-0 stripping
- RAID-1 mirroring
- RAID-5 parity
- RAID-6 double distributed parity
- Nested RAID arrays e.g. RAID-10
- (non-raid stuff)

RAID 0

striping

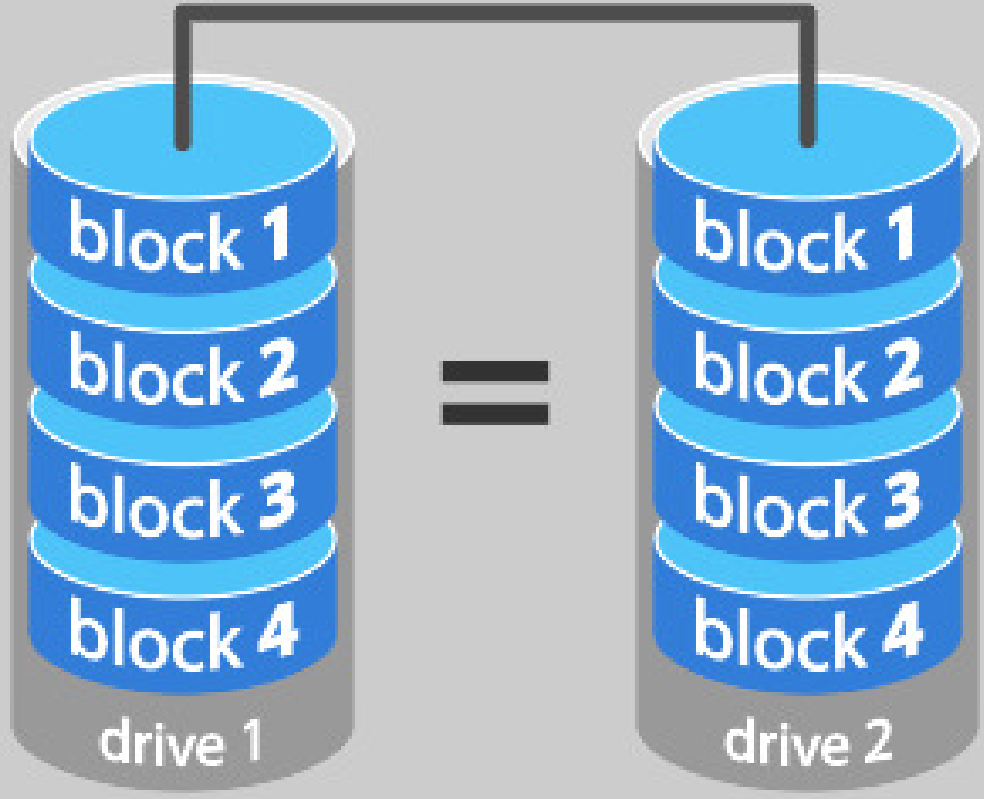


==> RAID-0 stripping

- The fastest array ever (multiply)
- The best resource usage / capacity ever (multiply)
- But absolutely no redundancy / fault-tolerance (divide)

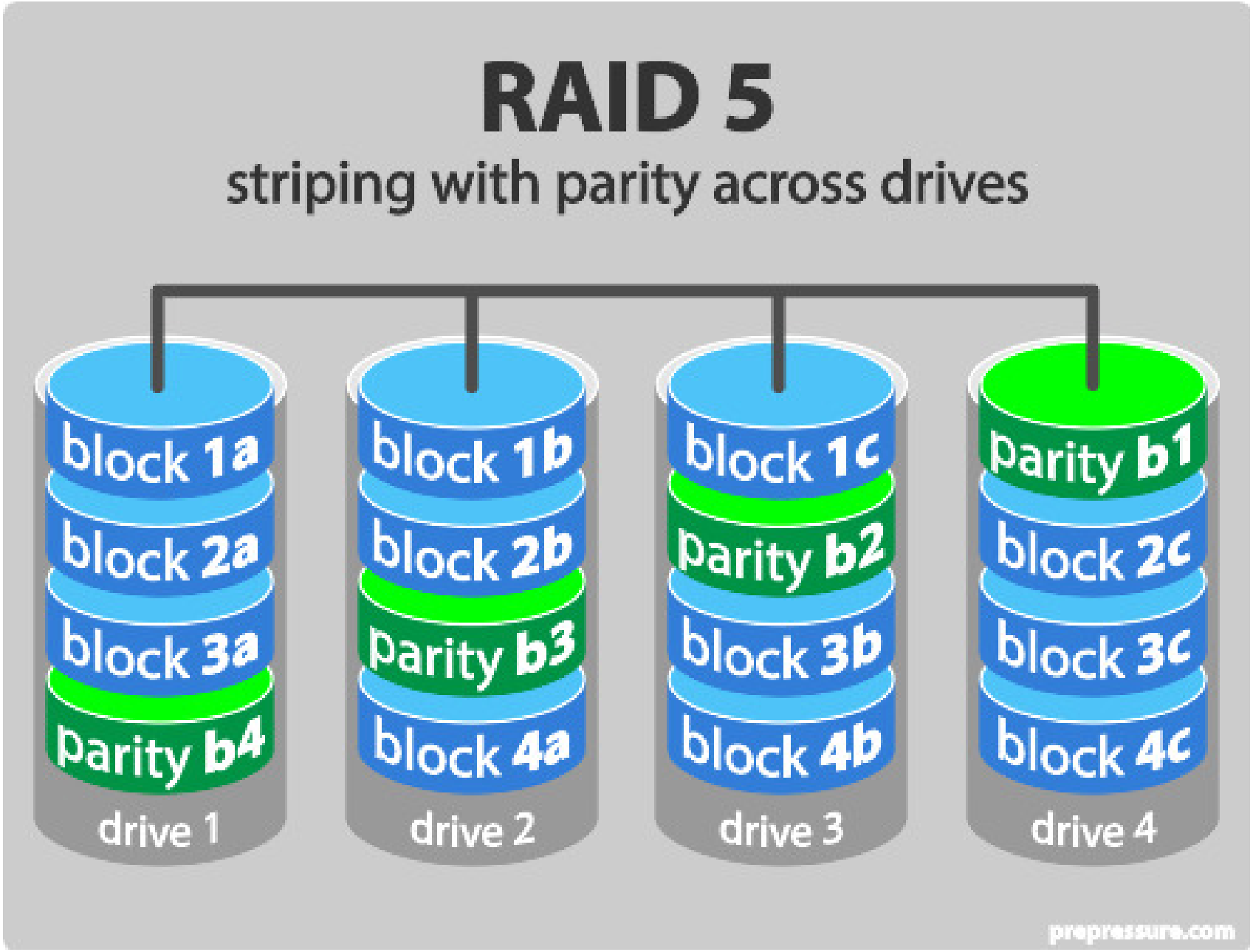
RAID 1

mirroring



==> RAID-1 mirroring

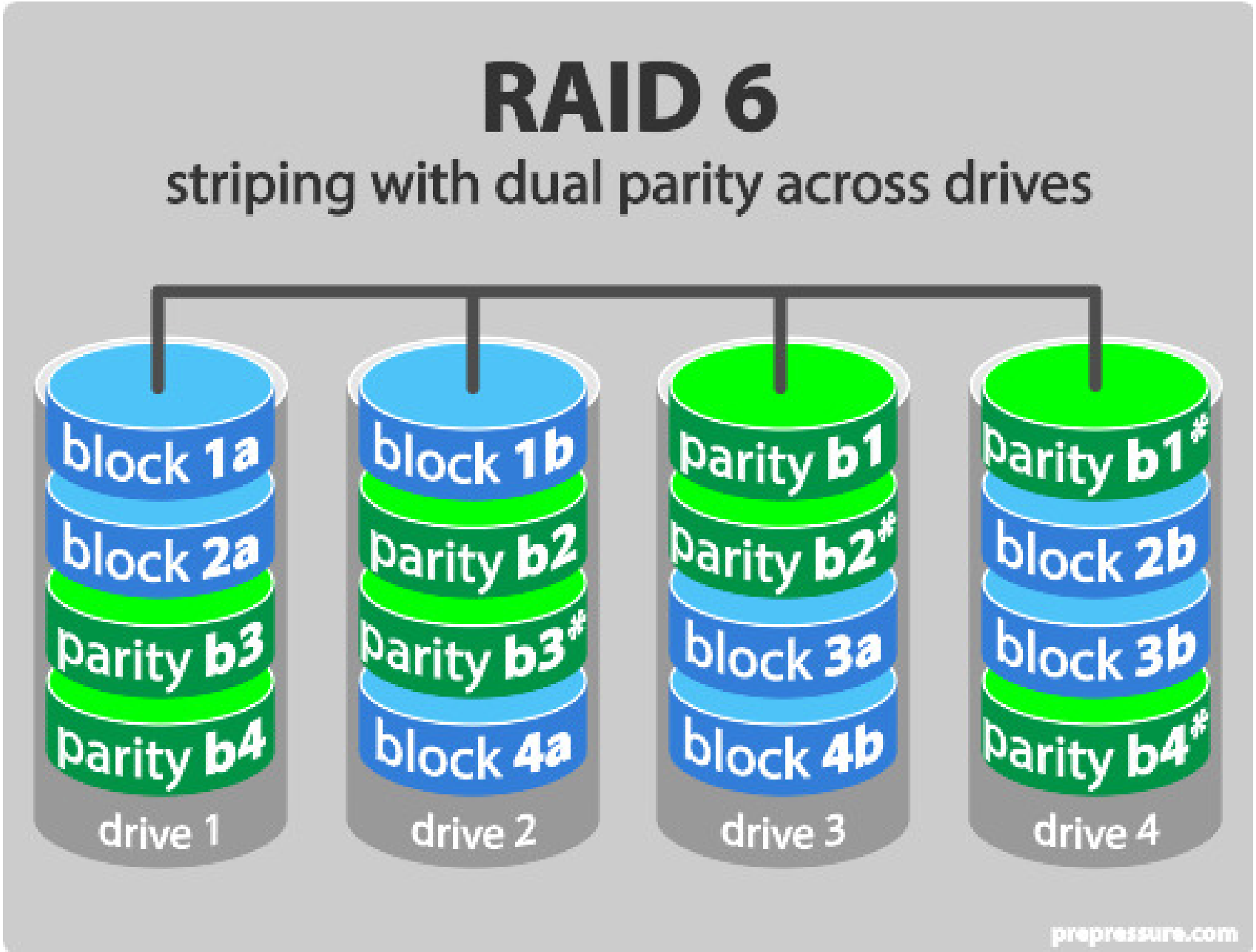
- Faster read, normal writes
- Capacity `/2`
- Fault-tolerant



RAID-5 parity

==> RAID-5 parity

- Faster read, normal writes
- Capacity `N-1`
- Fault-tolerance 1 disk (3 disks min)

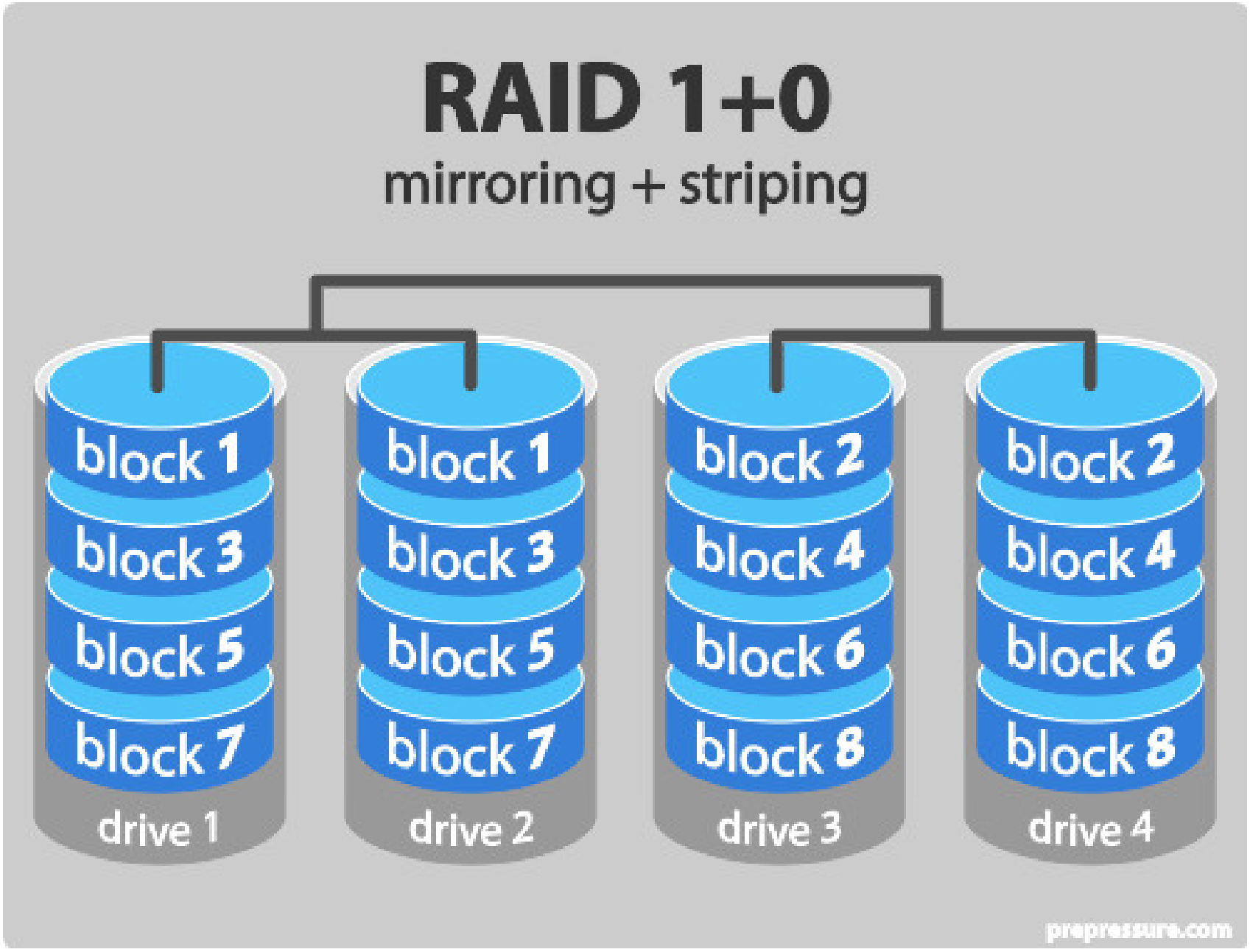


RAID-6 double distributed parity

==> RAID-6 double distributed parity

- Faster read, slower writes
- Capacity `N-2`
- Fault-tolerance 2 disks (4 disks min)

RAID 1+0 aka 10 (nested)



==> RAID-10

- Numbers in order from the root to leaves
- Got advantages of both RAID-1 then RAID-0
- The other way around is called RAID-0+1

Hardware RAID controllers

- Hardware RAID ~ volume manager
- System will think it's a disk while it's not

Bad performance w/o write cache

- Need for *batteries so outstanding requests can be flushed in case of power failures*
- Write-cache gets DISABLED if battery not present…

7Y37A01085, Контроллер Lenovo ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter

Ном. номер: 8000262838
PartNumber: 7Y37A01085
Производитель: Lenovo

lenovo

* Изображения служат только для ознакомления, см. техническую документацию

Бесплатная доставка по всей России

67 290 руб. **4 шт.** со склада г.Москва, срок 9-11 рабочих дней

– 1 +

Добавить в корзину 1 шт. на сумму 67 290 руб. **В корзину**

Купить в 1 клик

Описание | Сроки доставки

Lenovo 7Y37A01085. Поддерживаемые интерфейсы носителя: SAS, Serial ATA III, О интерфейс: PCI Express x8, Формат PCI карты: Полной высоты (низкопрофильный). 1, 10, 5, 50, 6, 60, Оперативная память: 4000 МБ, Чипсет: LSI SAS3516. Семейство LSI. Совместимые операционные системы: Microsoft Windows Server 2012 R2 Micros Server 2016 Red Hat Enterprise Linux 6 Server...

LSI SAS3516

- PCIe 3.0 x8 –> SAS, SATA III
- RAID: 0, 1, 10, 5, 50, 6, 60

210746 | Контроллер HP 820834-B21

RAID-контроллер для шины PCIe 3 x8 с последовательным подключением SCSI (SAS), скорость на порт 12 Гбит/с

в наличии

Доставка по всей России / самовывоз

Добавить к сравнению

Версия для печати

для частных лиц и организаций
любая форма оплаты

35 450 руб.

Добавить в корзину

и продолжить покупки

Калькулятор срока и стоимости доставки

P440/2

263755 | RAID-контроллер HP 804405-B21

RAID-контроллер P408e-p, 8 внешних портов, два разъема SFF8644, интерфейс PCI Express 3.0 x8, ОЗУ 4Гбайт, RAID 0/1/10/50/6/60, пропускная способность интерфейса 12 Гбит/с, требуется аккумуляторная батарея HPE Smart Storage 875241-B21 (приобр. отдельно)

в наличии

Доставка по всей России / самовывоз

Добавить к сравнению

Версия для печати

для частных лиц и организаций
любая форма оплаты

54 840 руб.

Добавить в корзину

и продолжить покупки

Калькулятор срока и стоимости доставки

P408e-p

Батарея HPE 96W Smart Storage (до 20 устр./145мм каб., аналог 875241-B21) (P01366-B21)

В НАЛИЧИИ

1

Люсиновская	Дорогобужская	Удаленный склад
(Отсутствует)	(Отсутствует)	(Очень мало)

Цена: 6 760 ₽

Software RAID

- Also bad performance w/o write cache
- Need for an UPS -or-
- 2N & redundant PSUs

Any software RAID products in mind?…

==> Software RAID products

- Linux raidtools/raidtools2/mdadm
- Linux LVM2 mirror (not bootable)
- ZFS pool on Solaris, FreeBSD, NetBSD, Linux
- NetBSD RAIDframe
- *got more?*

–> If bootable media, enable firmware to boot on both disks…

Non-raid stuff

Unraid

- data stored and available by independent disks
- parity done separately

// Questions so far on RAID?

How a systems gets to know about (virtual) disk changes

some magic vmm@ needs to know

Ack a new/deleted disk

```
ls -lF /sys/class/scsi_host/
echo "- - -" > /sys/class/scsi_host/host0/scan
```

Ack a larger/smaller vdisk

```
ls -lF /sys/class/scsi_device/scsi=0:0:0:0
echo 1 > "/sys/class/scsi_device/$scsi/device/rescan"
```

Software volume managers

far more flexible than RAID controllers

- LVM2
- XVM
- ZFS
- Veritas VxVM
- Microsoft® Logical Disk Manager

LVM2 basics

MBR partition type

```
8e  Linux LVM
```

GPT partition type

```
8E00  Linux LVM
```

Don’t even bother if it is on a secondary disk!

```
pvcreate /dev/sdb
vgcreate datavg /dev/sdb
lvcreate -n datalv -l 100%FREE datavg
```

Check

```
pvscan / pvs
vgscan / vgs
lvscan / lvs
```

Flexible you say?…

Increase VG/LV/FS with new disk

check available VG size after its extend

```
pvcreate /dev/sdb
vgextend vg_rhel6vm1 /dev/sdb
vgdisplay vg_rhel6vm1 | grep Free
```

hot-grow the LV, even for a system volume

```
lvextend -l +2559 /dev/vg_rhel6vm1/lv_root
# -r
# 100%FREE?
```

finally extend the FS

```
resize2fs /dev/vg_rhel6vm1/lv_root
xfs_growfs /dev/mapper/vg_rhel7-lv_root
```

Hot-shrink a data volume

before you plan to reduce virtual disk size…

```
fsck -f /dev/mapper/vg_data-lv_data
...144315055/199750656 blocks

echo $((199750656 * 4096))
818178686976
```

à la louche à 1GiB prêt

```
resize2fs /dev/mapper/vg_data-lv_data 650G
...170393600 (4k) blocks.

echo $((170393600 * 4096))
697932185600
```

avec la même louche

```
lvdisplay --units B /dev/vg_data/lv_data

lvreduce --size 650G /dev/vg_data/lv_data
# --resizefs

lvdisplay --units B /dev/vg_data/lv_data
```

then finally shrink the virtual disk or remove some PV

```
pvs
pvremove /dev/sdb
# /dev/sdb1
```

Even more flexiblility you say?…

THE GOLDEN HORDE

MIGRATE FROM ONE SAN TO ANOTHER WITHOUT INTERRUPTION

```
pvcreate /dev/sdb
vgextend vgdata /dev/sdb
pvmove -v --interval 5 /dev/sdc1 /dev/sdb
```

AND POSSIBLY WITH PAUSE/RESUME

```
^C
pvmove
```

When volume fully migrated, clean-up and get rid of the old SAN

```
vgreduce vgdata /dev/sdc1
pvremove /dev/sdc1
```

// Questions on volume managers?

File-systems

What do we have for DOS & Windows vs Linux?…

DOS & Windows

- FAT12 (old floppy)
- FAT16
- FAT32 (DOS/Windows/UEFI)
- NTFS

Linux

- btrfs (has compression)
- ext2,3,4
- xfs (SGI)
- jfs (IBM) big files
- reiserfs (conflict with community AND killed his wife) small files
- reiser4 (has compression)

Any exotic systems? Are there other file-systems out there?…

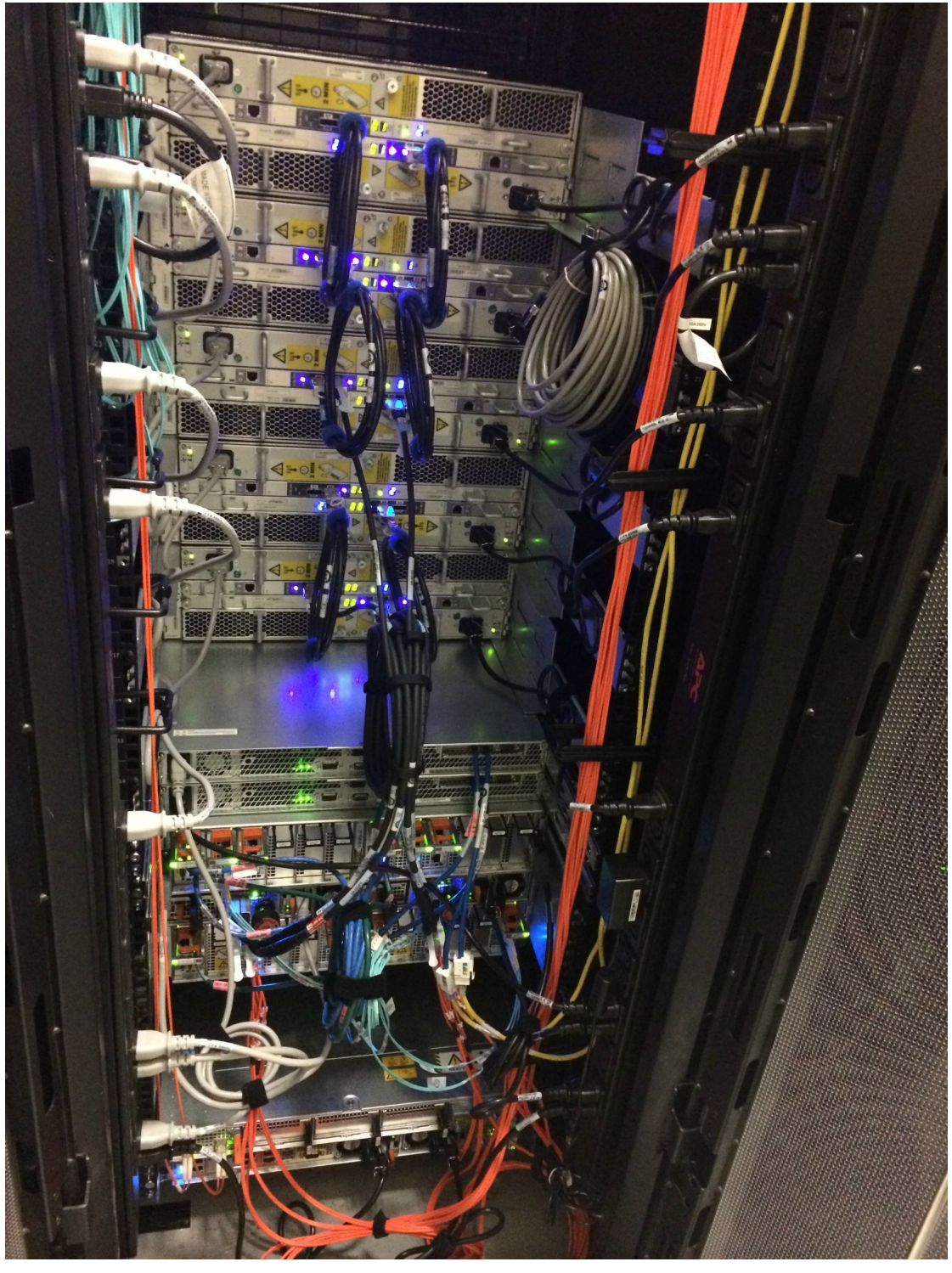
Other UNICES

- FreeBSD - UFS/UFS2, EXT2FS
- OpenBSD/NetBSD - FFS/FFS2, EXT2FS
- MINIX file-system base for Linus' ext (later ext2)
- LAB // what about BeOS's and more exotic systems' FS?

// Questions on file-systems?

Old-school SAN


VNX5400
EMC
VNX 5400



Protocols

- Fibre Channel Protocol
- iSCSI
- Network Block Device (NBD)

Terminology

- Target
- Fabric Switch / Fibre Switch
- GBICs / Mini-GBICs
- Initiator

Architecture

- Rackmounts full of disks
- Redundant SAN controller
- Split into multiple RAID groups e.g. RAID-50 arrays or whatever
- Then spliting those again into LUNs
- Delivering those through SAN Switch Fabric(s)

Storage bays

- DELL EMC hybrid-flash
- DELL EMC SC series
- DELL EMC VMAX
- HPE 3PAR StoreServ
- Netapp NVMe-over-Fabrics
- IBM goes flash too
- Got competition! Lenovo Storage S2200 and S3200
- Hitachi Vantara
- Huawei OceanStor (all-flash vs hybrid)
- Oracle ZFS Storage (software RAID!)
- Fujitsu ETERNUS (SDS!)
- WD Ultrastar - Hitachi Global Storage Technologies (HGST)

Open Source storage

take a server with loads of disks

- XigmaNAS (formerly NAS4Free) - early FreeNAS fork
- iXsystems TrueNAS - all OpenZFS powered * CORE (formerly FreeNAS) * ENTERPRISE (commercial) - hybrid-capable * SCALE (Linux-based) - convergence-capable
- ESOS - Enterprise Storage OS - got iSCSI & FC targets

and others…

- OpenMediaVault (OMV) - Debian-based, mostly NAS
- Openfilter - CentOS-based, got iSCSI & FC targets

Got dedicated hardware?

somehow yes, as many SAN/NAS vendors embbed Linux

- QNAP (Taiwan)?
- Synology (Taiwan)?
- Thecus (Taiwan)

not really…

- ODROID-HC2 / ODROID-XU4 with an LCD
- LG N2R1 / N2B1 / N1T1 – Tantalus' Firmwares
- *idem* with NetBSD

HA-capable

compete with storage bay controllers…

- XigmaNAS – HA thanks to HAST
- TrueNAS – HA only with commercial version
- ESOS – HA thanks to DRBD

Fibre switch w/o GBICs

Mini-GBICs

XFR vs SFP+

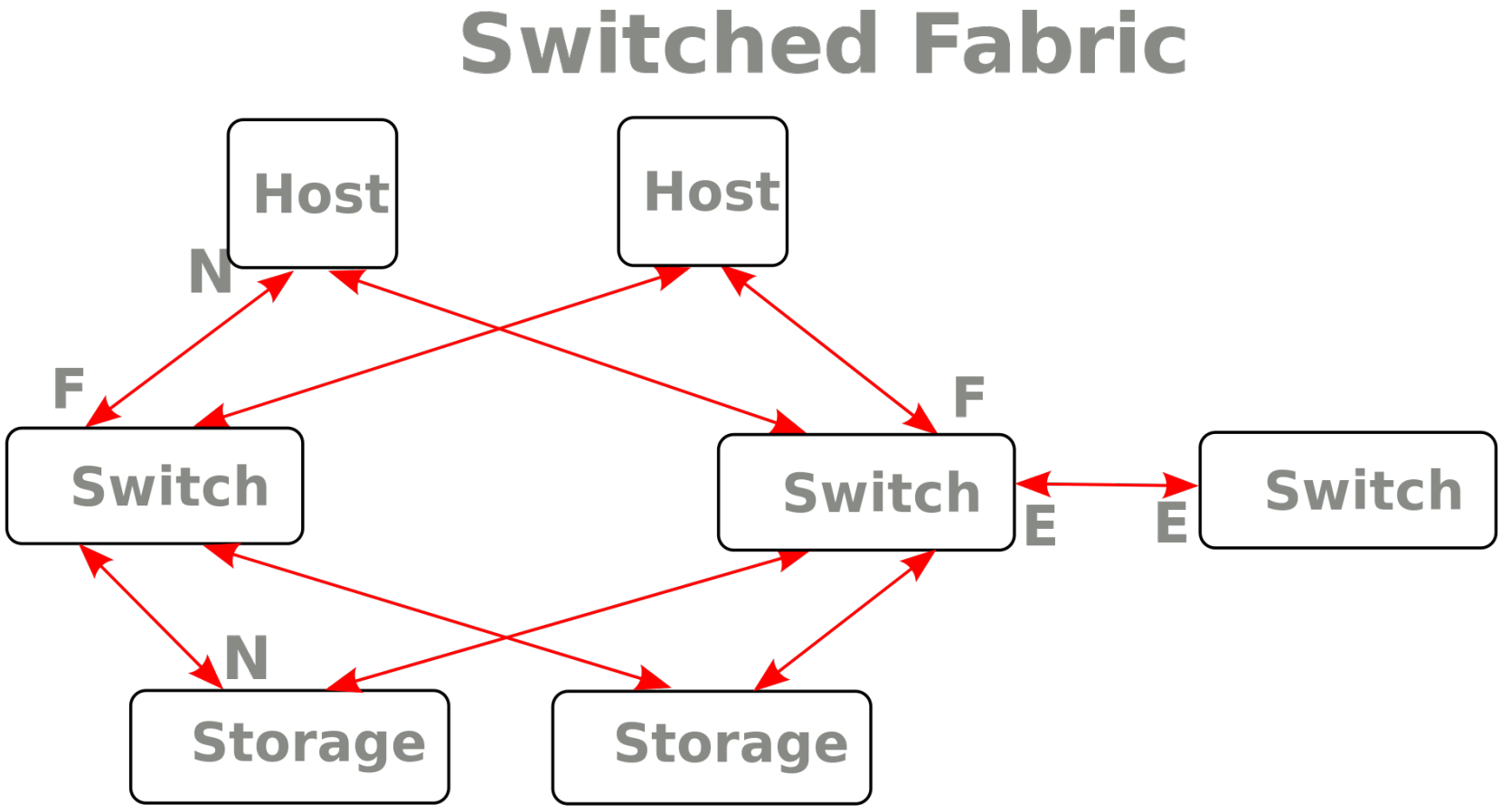
Fibre switch w/ GBICs

Brocade 5120 40x 8Gb Active SFP+

Switch Fabric(s)

- Redundant Fibre Switches
- Redundant Switch Fabrics (optional)
- Setting up `zoning`
- Setting up `LUN masking`

Single Fabric



SAN initiators

- Redundant Host Bus Adapters (HBA)
- Obtaining `sdX` for every path through the Fabric(s)

How to deal with the redundant disk devices?…

Linux DM Multipath

CentOS/RHEL

```
yum install device-mapper-multipath
vi /etc/multipath.conf
systemctl start multipathd
```

Ubuntu

```
apt install multipath-tools
vi /etc/multipath.conf
systemctl start multipath-tools
```

(all)

```
multipath -ll
ls -lF /dev/mapper/mpath*
ls -lF /dev/mpath/mpath*
```

Mixing up protocols

- FC over Ethernet (FCoE)
- FC over IP
- Internet Fibre Channel Protocol (iFCP)
- iSCSI (SCSI over TCP/IP)
- Ethernet over Infiniband

iSCSI

Specifics about iSCSI vs. FC

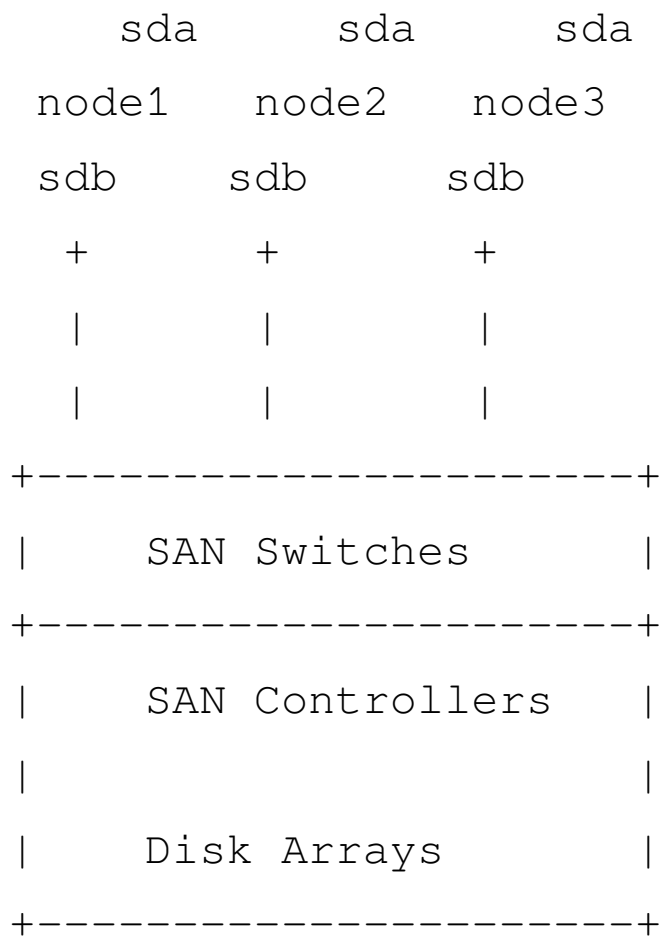
- Same target / initiator convention
- Fibre Channel World Wide Names (WWN)
- But without FC Switch Fabric nor zoning
- iSCSI Qualified Name (IQN)

So why is it cheaper?

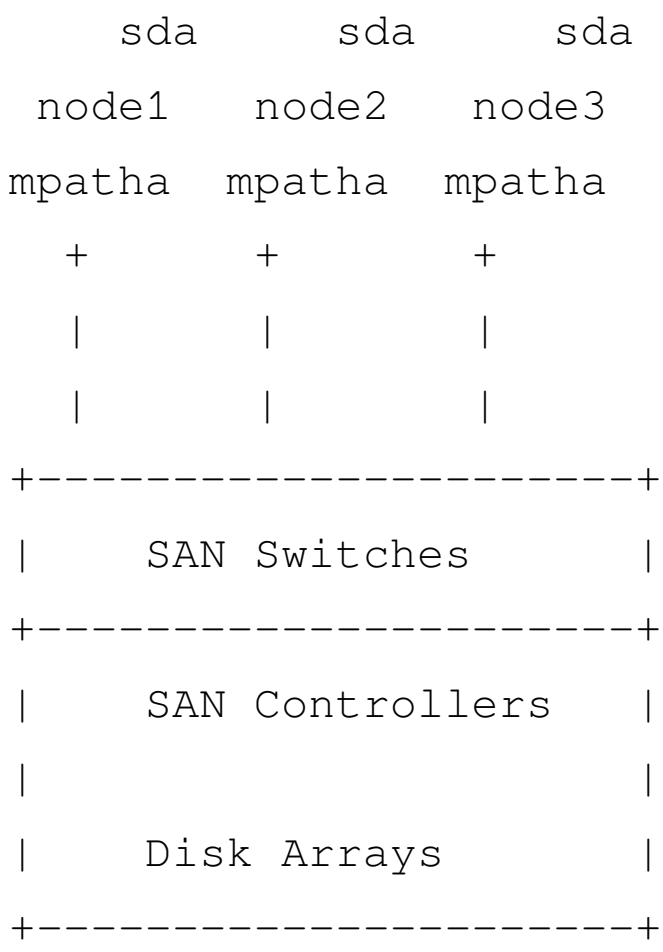
==>

- No SAN-specific switches and cabling (no Fibre Channel)
- Can be used on the same network switches with physical VLANs (untagged/access)
- Dedicated NICs are recommended (ideally 10Gbit/s)

// Questions on SAN targets and initiators?



RAID is here and there…



Can I use a casual file-system on `mpatha`*?*

==> NO – casual file-systems are single mount

Deal with LUNs

- don’t mount it
- –or– monut it only once
- –or–… *how to deal with multiple mounts?…*

==> Shared-disk file-system

aka clustered file-system

- Redhat GFS2
- OCFS2
- VxFS (Veritas Cluster File System)
- IBM GPFS
- SGI CXFS
- Lustre

==> Eventually volumes as well

- lvmlockd (preferred)
- CLVM
- VxVM (Veritas Volume Manager)

// Questions on shared-disk file-systems?

SAN vs NAS

Storage Area Network

- you get disks (LUNs)…

Network Attached Storage

- you get a network file-system (CIFS, NFS, …)

Network file-system flavors

- NFS v2 (need re-enable it for some old network boot-loaders)
- NFS v3
- NFS v4
- CIFS *(fine for end-users…)*

File ownerships & permissions

- Got an NT4 domain? Plug Samba 3 to the DC.
- Got an AD domain? Plug Samba 4 to the AD.

==> enable ACLs in `fstab` and reboot

- Got NIS or LDAP domain? Simply plug the NAS or NFS server to it.

Remember the VMM farm architecture with its oblivious resource?…

So SAN or NAS, which is best for guests?…

==> Block devices (disks through SAN) are faster

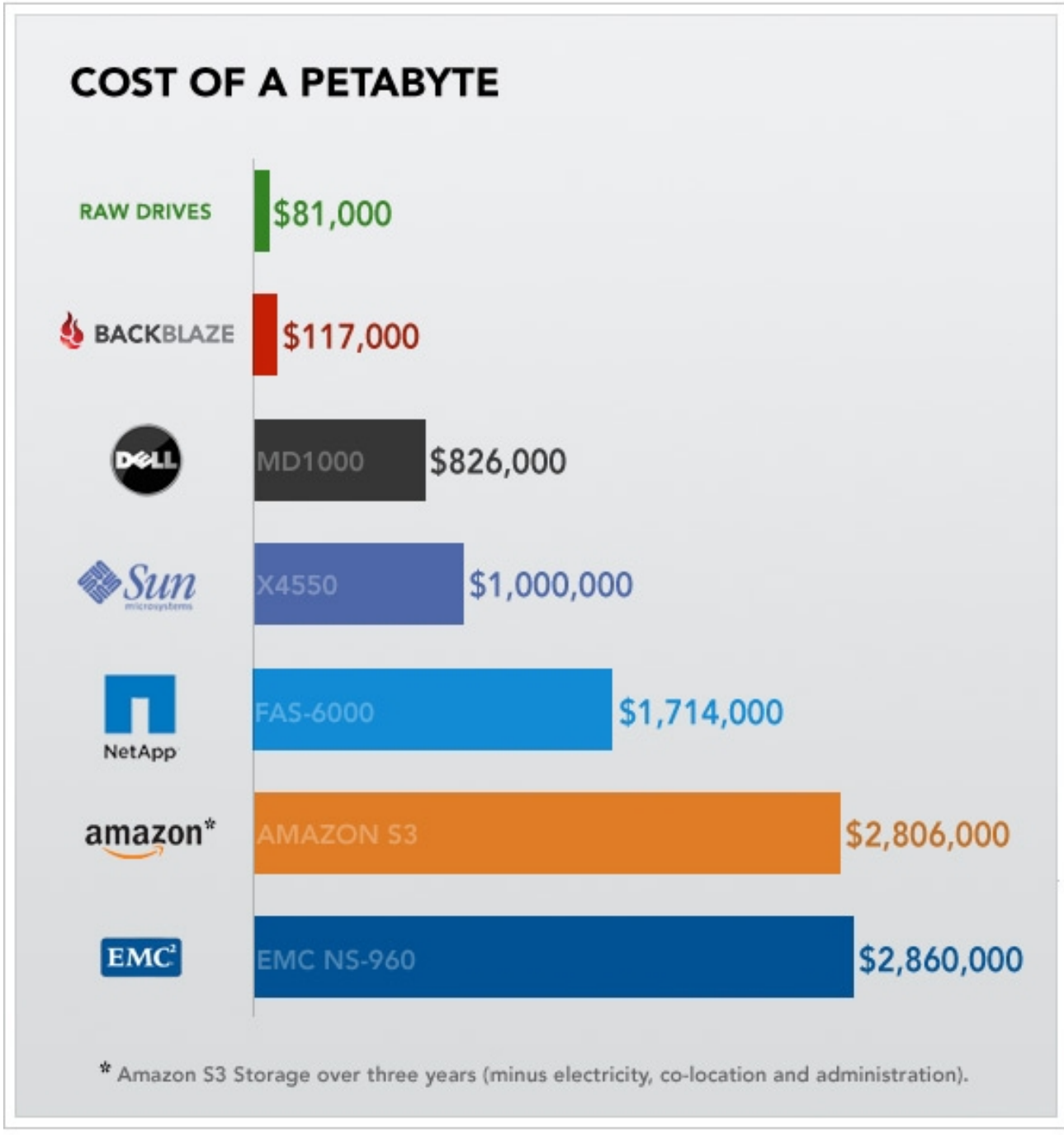
- **Because there is one less layer of file-system**
- **Because that lower-layer goes through *some* network protocol**
- But NFS and vdisk on QCOW2 or sparse files makes life easy
- Just like Oracle RDBMS datafiles vs ASM…

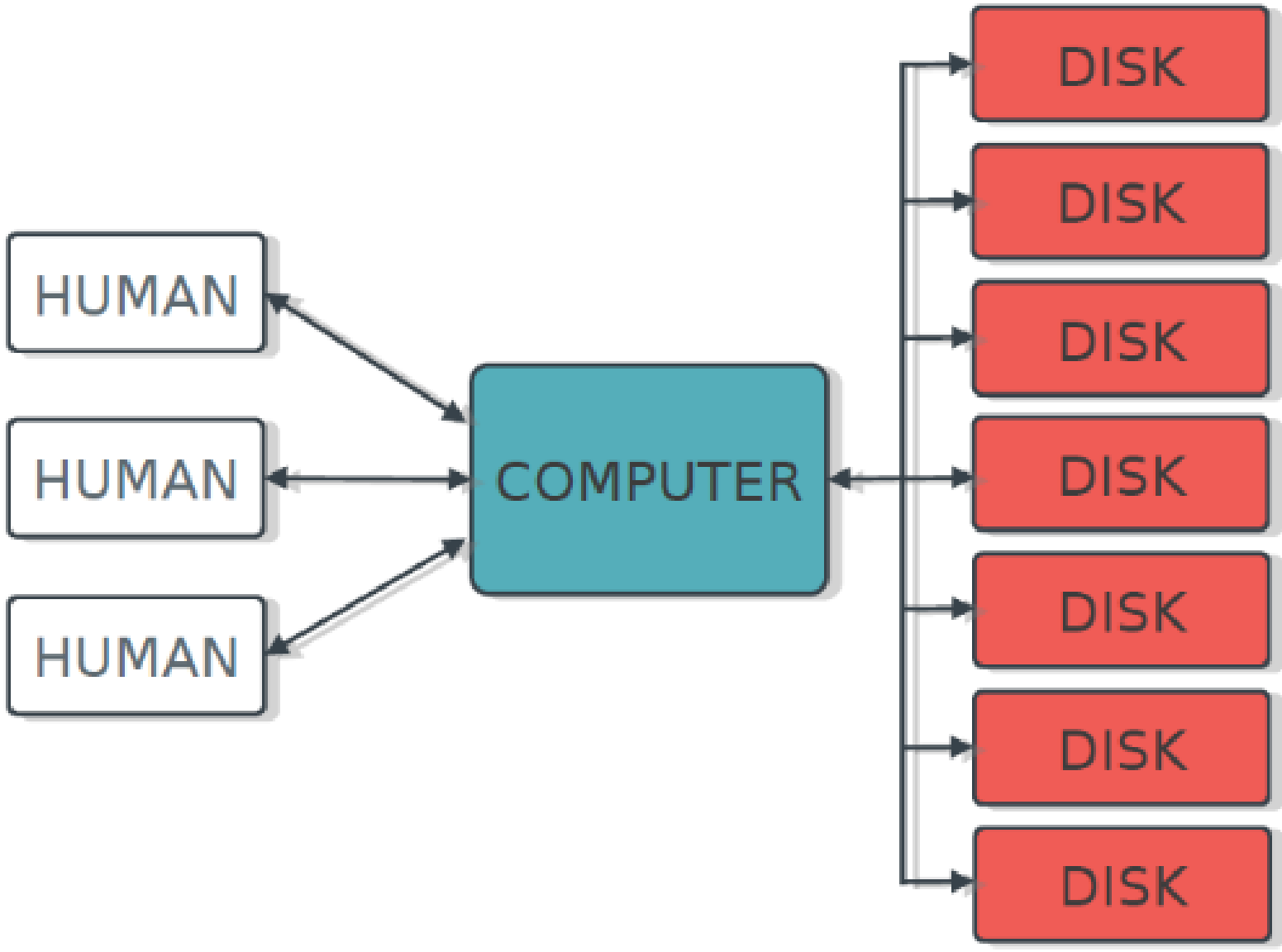
Tricky sploiler - vdisks on shared-disk file-system are also possible

// Questions on NAS?

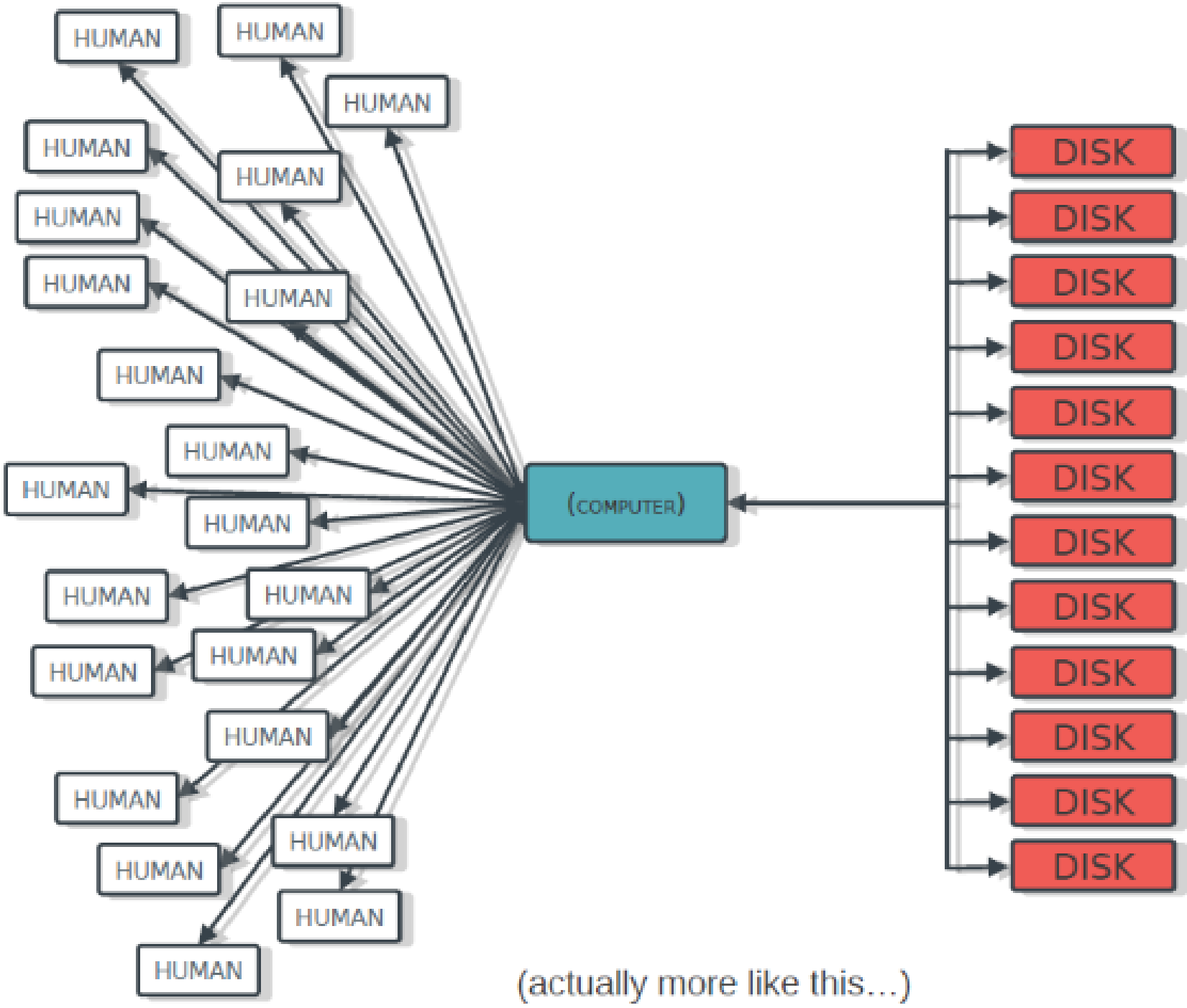
Distributed Storage

What kind of business market is this?…


COST OF A PETABYTE
RAW DRIVES
$81,000
BACKBLAZE
$117,000
DELL
MD1000
$826,000
Sun microsystems
X4550
$1,000,000
NetApp
FAS-6000
$1,714,000
amazon*
AMAZON S3
$2,806,000
EMC²
EMC NS-960
$2,860,000
* Amazon S3 Storage over three years (minus electricity, co-location and administration).



Everything is fine

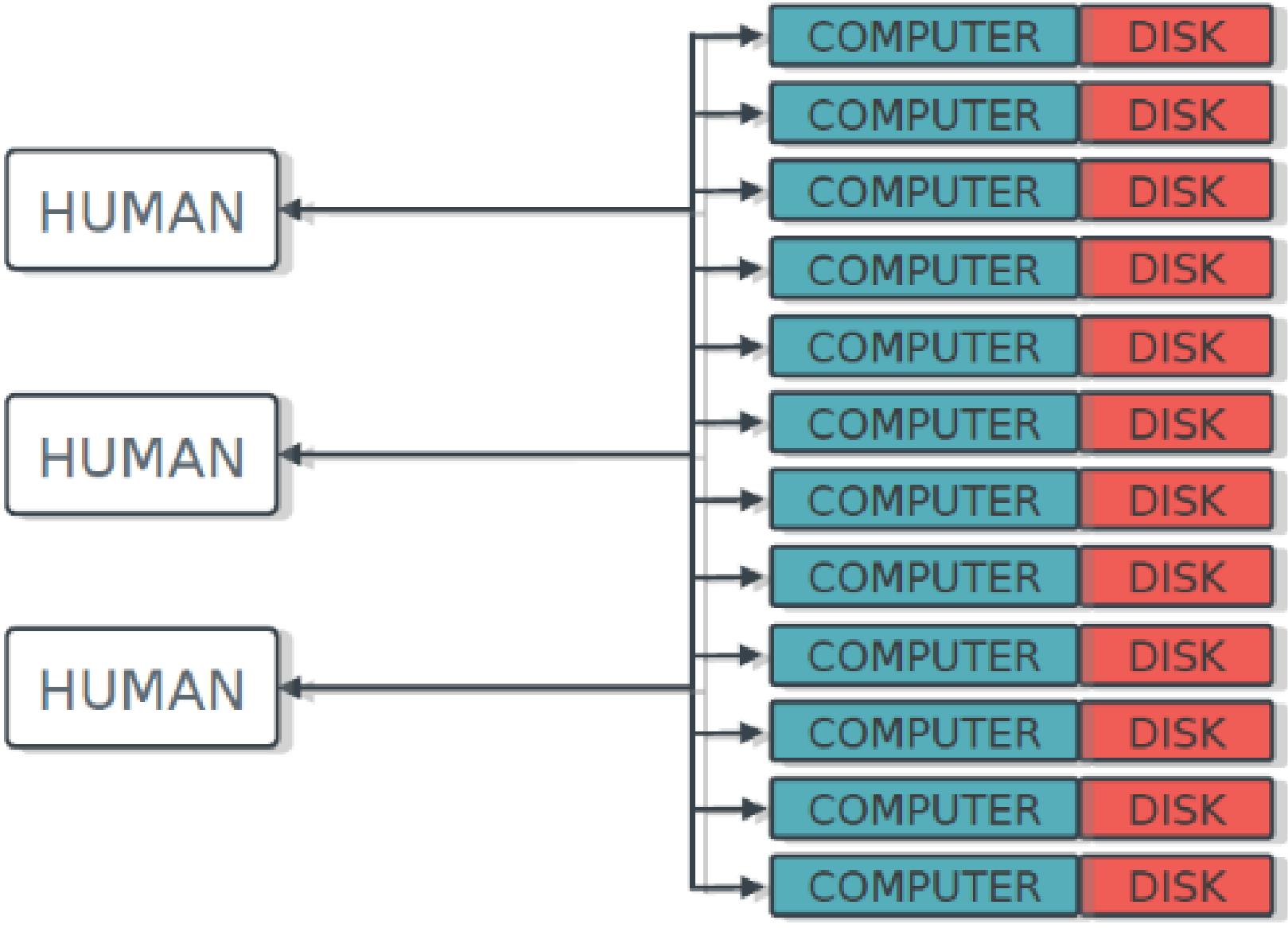

HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
(COMPUTER)
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
(actually more like this…)

How do you design a storage system that scales?…

New-school POWAAAAAAA

This is the new school: servers full of disks as a cluster

- *Smart* RAID over the network
- –or– full-blown Distributed Systems


HUMAN
HUMAN
HUMAN
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK

Storage layers

- Distributed Block Devices
- Distributed File-Systems
- Distributed Object Storage

Distributed Block Devices

- Ceph Block Device (FreeBSD + Linux)
- Gluster Block Device
- DRBD v8 (Linux)
- DRBD v9 (Linux)
- DRBD v9 + Linstor (Linux)
- HAST (FreeBSD)

Distributed File-Systems

- CephFS
- GlusterFS
- HDFS
- Google File-System (this time it is Google's GFS…)

Distributed Object Storage

- Ceph Object Storage
- Hadoop? (on top of HDFS)
- Bigtable? (on top of GFS?)
- MinIO

But you always got NoSQL otherwise

- Apache Cassandra –> Scylla is much faster (C vs Java…)
- MongoDB, CrateDB, etc.

Distributed Storage Architectures

- network mirror(s)
- distributed system algorithm

Network mirror(s)

- DRBD (Linux) primary / secondary (got write-cache!)
- DRBD (Linux) dual-primaries (slower during guest migration only, no write-cache)
- HAST (FreeBSD)

–> That works, but only for the equivalent of RAID-1

DRBD setup types

- DRBD against physical disks
- DRBD against physical volumes (possibly hardware RAID)
- DRBD against MBR/GPT partitions
- DRBD against LVM2 logical volumes
- DRBD against LVM2 LVs w/ thin provisioning (no CLVM/lvmlockd needed!)

Distributed system algorithm

- GFS/bigtable –> hash-table
- HDFS/Hadoop –> hash-table
- Ceph –> CRUSH algo

Hash-table

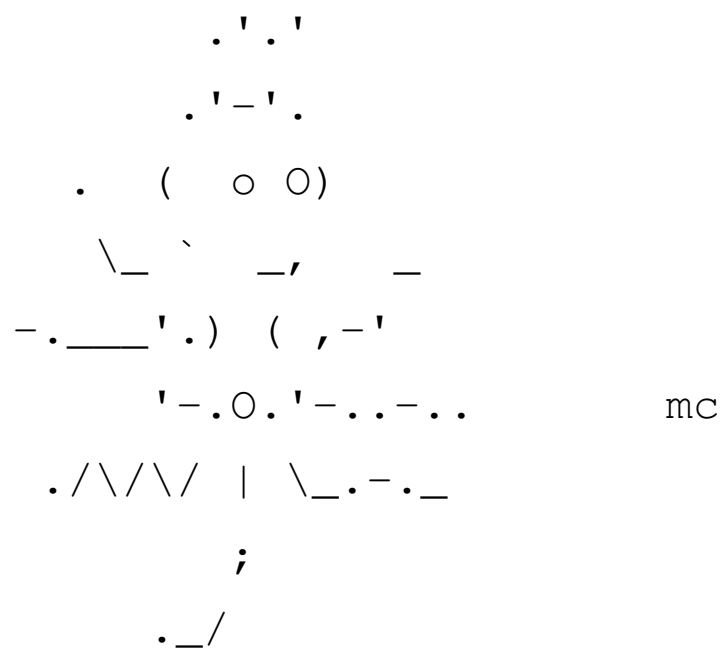
- fast lookups
- not necessarily time-efficient while redistributing data

CRUSH

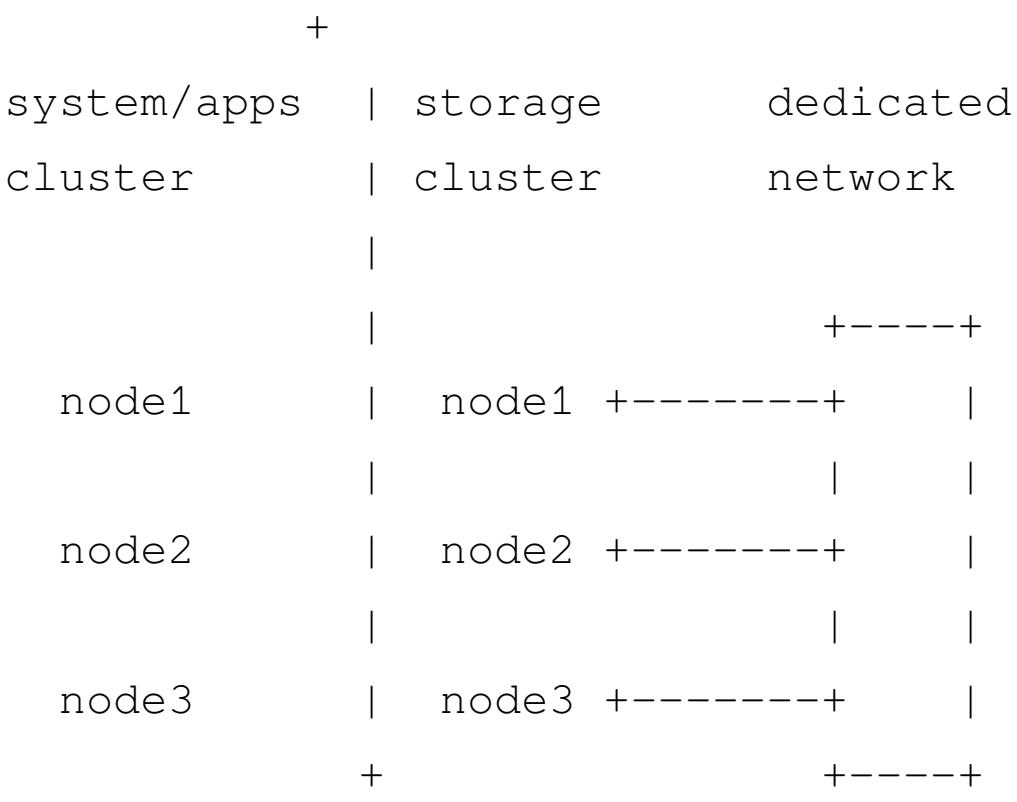
- some kind of hash-table but it is more complex
- efficient re-distribution when you scale your cluster

// Any question on the principles of distributed storage?

Convergence

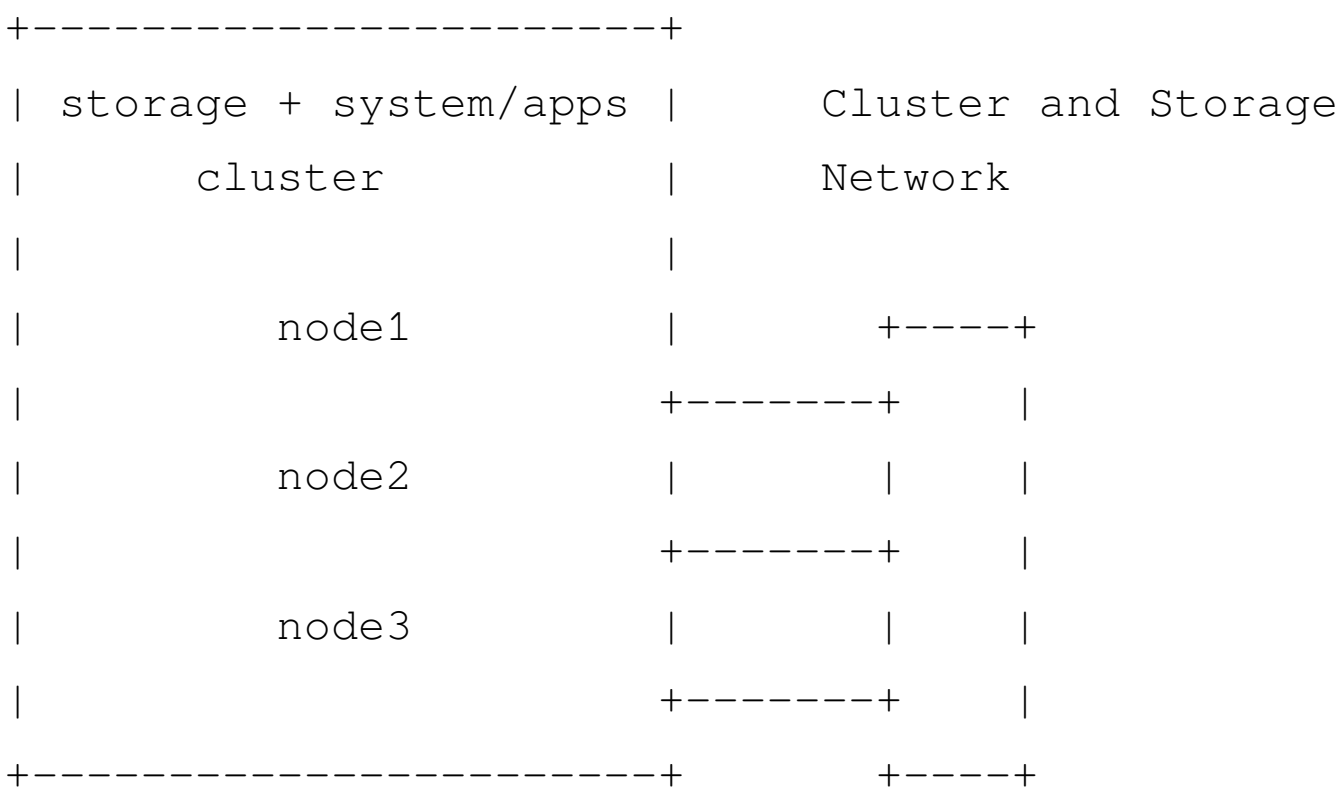


A safe separation

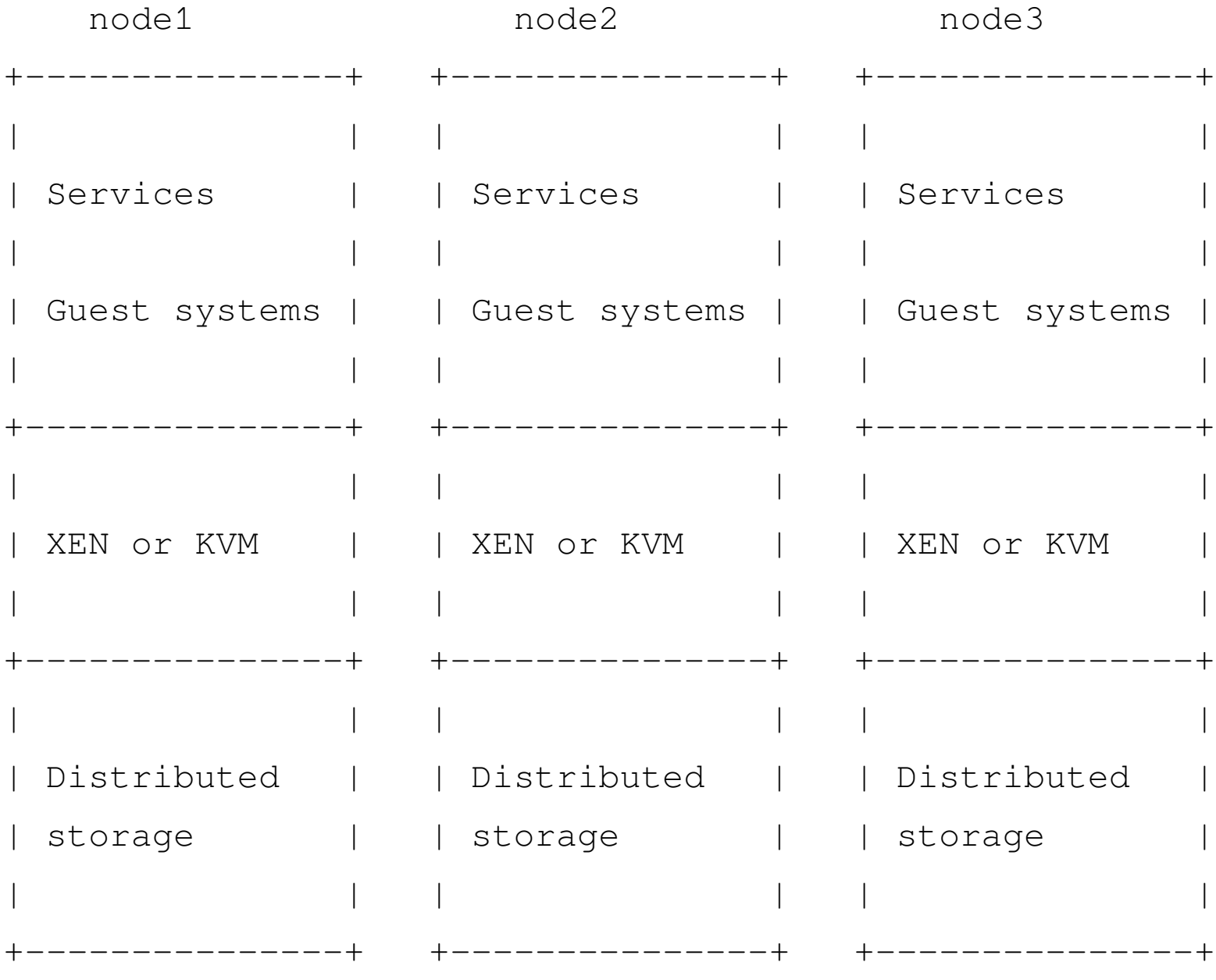


- Can also be virtualization + system/apps
- Dedicated network for storage (phy vlan/access)
- Dedicated network for cluster (phy vlan/access)

Convergence



Converged Infrastructure



–> THREE OR FOUR STACKS on the same hardware

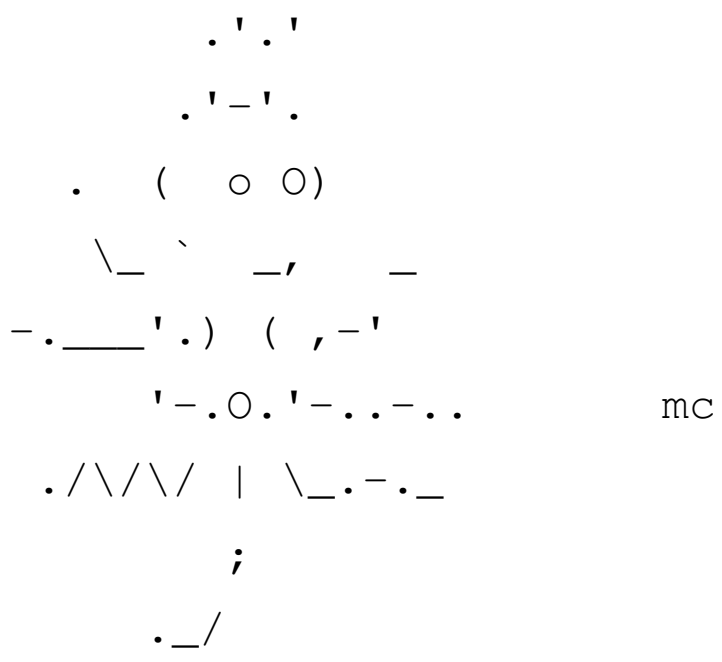
- Less hardware maintenance, less costs and resource optimization
- Everything is distributed –> scale-out capable
- One node dies, every cluster stack deals with it on its own
- VLANs kept separate for cluster vs storage

But…

- PERFORMANCE BOTTNECK MAY HIT YOU HARD
- NO WAY FOR DISTRIBUTED NAS
- THIS IS FOR DISTRIBUTED SAN ONLY

==>

- Plug to DRBD / Ceph RBD / Gluster Block Device cluster
- Not to CephFS nor GlusterFS



// Any question on converged infrastructures?

Software Defined Storage (SDS)

related but not identical to distributed storage

- ~Storage virtualization
- ~Replication
- ~Data de-duplication
- ~Thin provisioning

And more generally, the ability to manage volumes by an API or some integration plug-in e.g. Proxmox plugs to DRBD (LAB).

// Questions on software defined storage?

Performance Tools & Graphs

- Tools for troubleshooting and sizing
- Performance graphs for daily activity and RCAs

Performance bottleneck troubleshooting

What if the service is up but does not perform well?…

Namely, users and customers are complaning about latencies are are saying "it is slow".

==> need to find the performance bottleneck

Sizing migrations

Know what resources you need

- for P2V & V2V
- for P2C & V2C (cloud migrations)

Note another way to go is to give max power to all guests and closely monitor their consumption (private cloud only).

RESOURCE TYPES TO TRACK

- CPU (usage vs. load queue)
- RAM USAGE (& RAM BUS)
- DISK I/O
- NETWORK TX/RX PER INTERFACE

How to check for CPU usage and load queue manually?

==> CPU

```
uptime
top -b
htop
```

X11

```
xload
conky
gkrellm
```

XEN

```
xentop -b -i 1
#--> CPU(sec) CPU(%)
```

How to check for RAM usage manually?

==> RAM

```
free -m
htop
```

XEN

```
xentop -b -i 1
#--> MEM(k) MEM(%)  MAXMEM(k) MAXMEM(%)
```

How to check for DISK I/O manually?

==> DISK I/O

```
apt install sysstat iotop
iostat -d 30 /dev/sda
iostat -x /dev/sda #--> %util
iotop -b -n 1
```

XEN

```
xentop -b -i 1
#--> VBD_RD     VBD_WR
#--> VBD_RSECT  VBD_WSECT
```

How to check for NETWORK INTERFACE TX/RX manually?

==> NETWORK TX/RX PER INTERFACE

```
iftop -i eth0
iptraf
trafshow
nload
nethogs eth0
vnstat -i eth0
```

XEN

```
xentop -b -i 1
#--> NETTX(k) NETRX(k)
```

And many others…

```
bmon, bwm-ng, cbm, slurm, tcptrack, netload, collectl, speedometer,
    pktstat, netdiag/netwatch, ifstat, dstat
```

Performance graphs

…that was just some UI (Grafana)

Performance graphs

Goals per system

- See how well your bare-metal systems are sized
- *idem* for guests

Spot the waste (and possibly a DoS attack) e.g.

- Who's using 100% ram?
- Who's using 100% disk i/o?

Also useful for

- Root Cause Analysis (RCA)
- Sizing machines and migrations e.g. P2V & V2V

Goals per hypervisor

- See how well your cluster farm is behaving
- (is the orchestrator doing its job?)
- RAM over-commitment vs. TMEM
- –> 70-90% is good (depending on your cluster size)

and beyond the 4 resource types

- Virtual disks' thin-provisioning

Products?…

==> THE COMPETITION

- SolarWinds – *major leakage lately…*
- *Any other proprietary product in mind?*

==> FOSS & commercial

- Nagios XI incl. performance graphs
- Monit & M/Monit (graphs built-in)

LAB // is there anything closed-source in Nagios XI? Use The Source, Luke.

==> OPEN SOURCE ASSETS

- Grafana –> Prometheus
  - vs. –> graphite
  - vs. –> influxdb
- Zabbix (graphs built-in)
- Munin + RRDtool
- Icinga 2 // LAB
- Sentry? // LAB
- Nagios Core?
- *(Monit – agent can send metrics…)*

LAB // try some performance graphs plugin for Nagios Core (without XI) – see resources

Various ways to get the metrics

- Agents (auto-deploy)
- Hypervisors
  - XEN `xentop`
  - XEN light library
  - some KVM equivalent? // LAB
  - possible from VMware ESXi or vSphere? // LAB
- SNMP

App & services' QA

- `ping` ***response time***

Business logic monitoring

Activities e.g.

- How many connections…
- How many users…

Time-Series DBMS (TSDB) engines

- Prometheus (& poor dashboard)
- Graphite (& poor dashboard)
- ElasticSearch as in ELK (& dashboard)
- Cacti?

Dashboards

Full-featured

- Grafana
- Kibana as in ELK

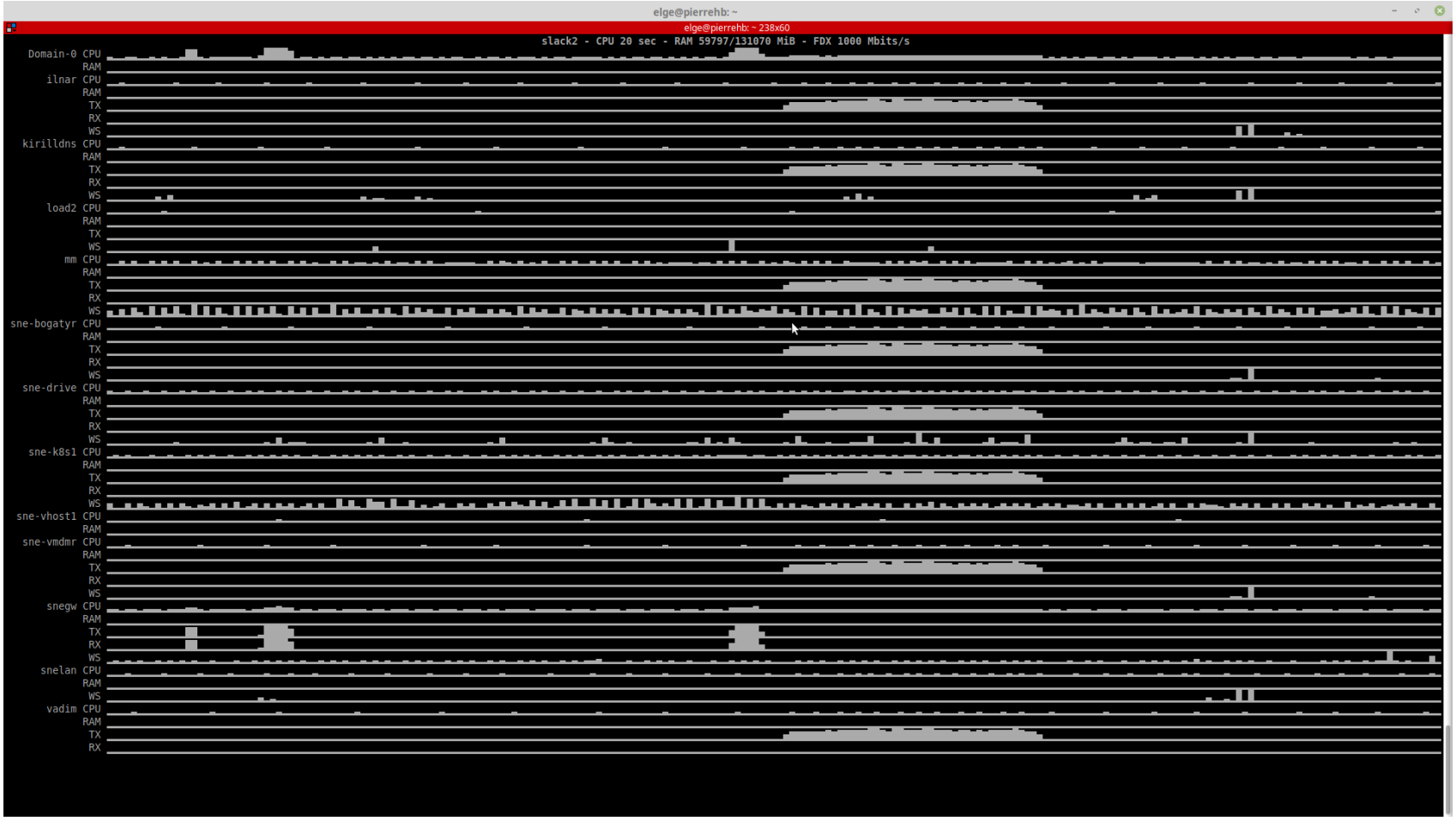
DIY

- MRTG vs. RRDtool
- Highcharts/Highstock
  - Wants JSON
  - Displays LIVE
- Spark (text-based UTF-8 bars)

Data collectors –> TSDB

- Node_exporter –> Prometheus (OpenMetrics)
- Collectd –> Graphite/Carbon
- Statsd –> Graphite/Carbon
- Logstash –> ELK
- Telegraf –> InfluxDB
- Glances –> InfluxDB

No Web Required



No Web Required

LOAD TEST ACCEPTANCE

How to benchmark vs. stress-test?…

==> Benchmarking == dedicated resources (ideally bare-metal)

==> Stress-test == just push-up the volume

LOAD STRESS CPU

assuming 16 cores

```
stress --cpu --cpu 16
openssl speed -multi 16
```

LOAD STRESS RAM

assuming 16 cores

```
stress --vm 16 --vm-keep
```

alternative to avoid OOM

```
mkdir -p ram/
mount -t tmpfs -o size=7168M tmpfs ram/
dd if=/dev/zero of=ram/ramload bs=1M
```

LOAD STRESS DISK

Get some idea about disk's speed

```
hdparm -tT /dev/sda
```

Stress the disk

```
time dd if=/dev/zero of=device-or-file bs=1G count=30
bonnie++ ...
```

assuming 16 cores

```
stress --io 16
```

LOAD STRESS NETWORK INTERFACES

Flood the network in one direction (UDP)

```
iperf3 -c -u x.x.x.x
```

–or– regulate while checking how much packets got there (TCP)

```
iperf3 -c x.x.x.x
```

// Questions on performance monitoring?