Operációs Rendszerek MSc

Szerverek, Adatközpontok

2019/2020/I.

Dr. Vincze Dávid
Miskolci Egyetem, IIT
vincze.david@iit.uni-miskolc.hu

Szerverek

- Mitől szerver a szerver?

Szerverek

- Mitől szerver a szerver?

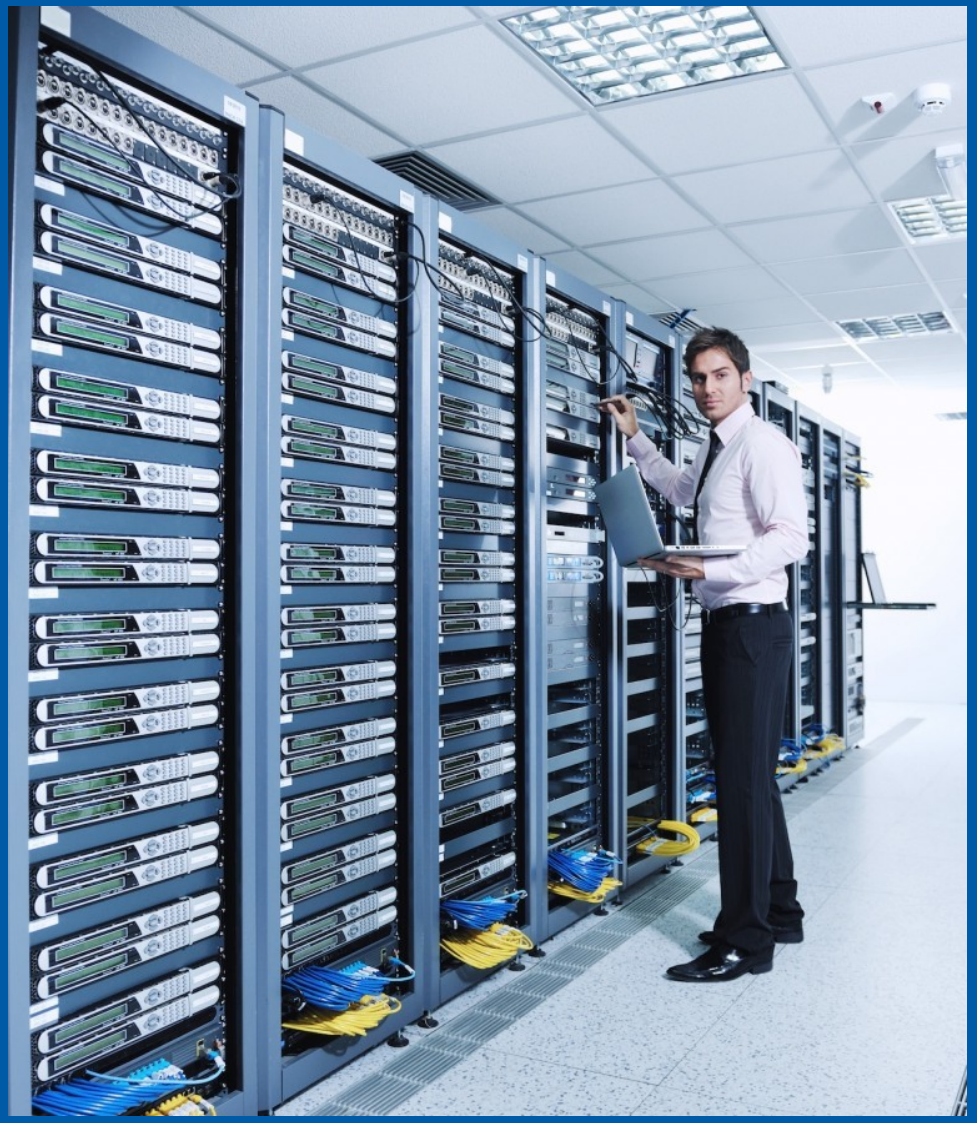
- Mert az a feladata...
- SoftWare
  - a rajta futó alkalmazások teszik azzá
- HardWare
  - mindenből sok... (sok diszk, sok memória, turbo proci...)
  - mindent kibírjon (diszk hiba, memória hiba, tápegység hiba, hálózati hiba, stb.)

Szerverek

- Software ?
  - file
  - web
  - adatbázis
  - név
  - címtár
  - alkalmazás
  - beléptető (autentikációs)
  - log (napló)
  - shell
  - stb.

Szerverek

- Hardware ?
  - erőforrás
  - megbízhatóság
  - menedzselhetőség



Szerverek

- Architektúrák
  - Intel x86, x86_64
  - IBM POWER
    - POWER, PowerPC
  - Sun SPARC – vége? (Oracle)
  - Intel IA64
  - (HP PA-RISC, Digital Alpha, Digital VAX, etc.)
  - ARM, MIPS
  - Egyéb mainframe-ek

  - → uniformitás

Szerverek

- Erőforrások (sok sok sok...)
  - sok diszk
  - sok memória (RAM)
  - sok CPU
  - sok (hálózati) sávszélesség

Diszkek

- Diszk (háttértár)
  - IDE -> SCSI
  - SATA -> SAS
  - SAN
    - FibreChannel (sic!)
    - iSCSI
    - AoE
    - stb.
  - RAID 0,1,10,0+1,1+0,5,5E,5EE,6,50,stb..
    - HW / SW
  - HDD/SSD
  - S.M.A.R.T.

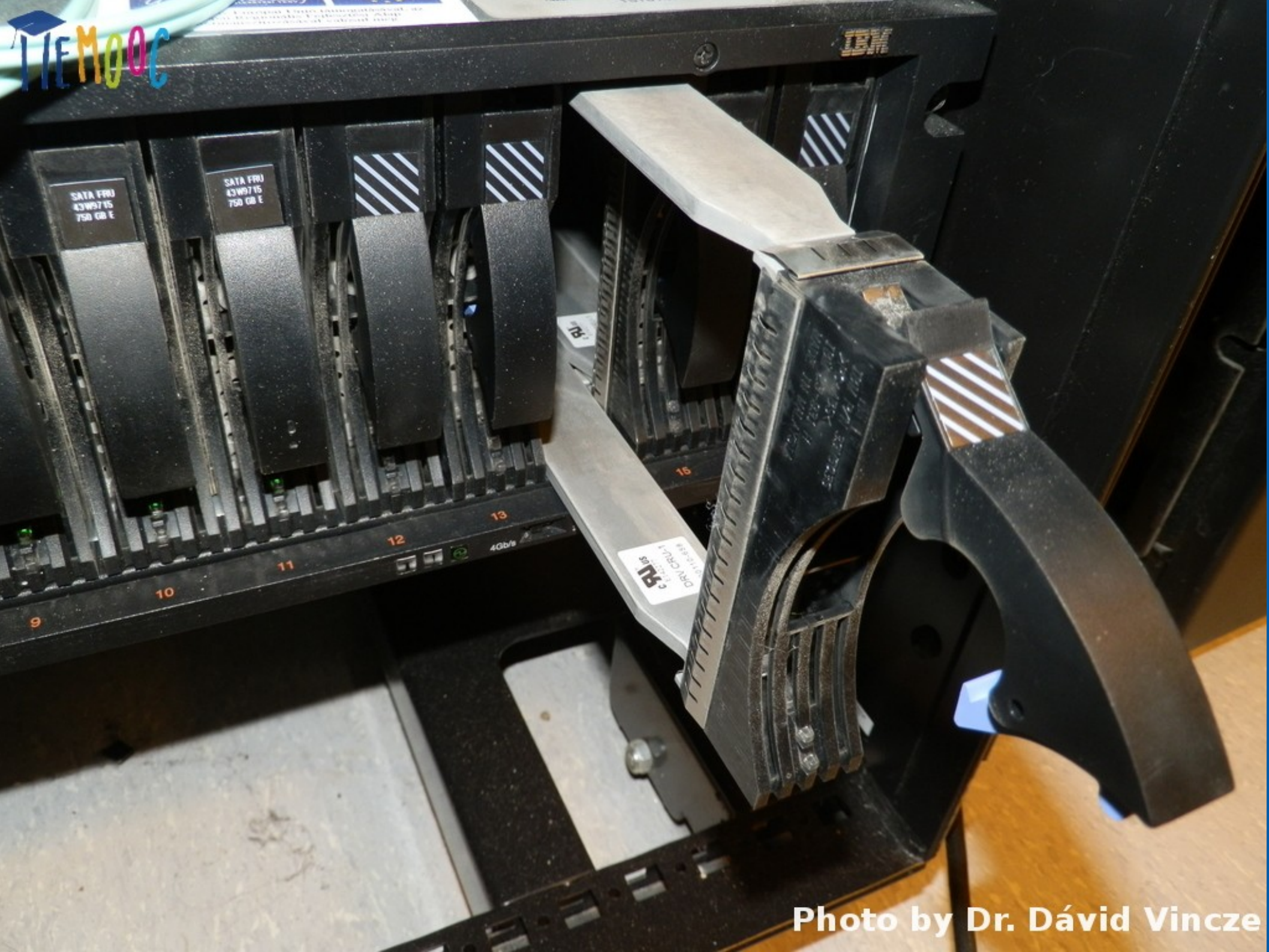
Diszkek

- Diszk (háttértár)
  - ES vs. PS
    - Enterprise Storage vs. Personal Storage
    - ES: SCSI/SAS:
      - Multi CPU támogatás
        - (failover és osztott hozzáférés)
        - 10k-15k RPM
        - sok diszk egy rendszerben
        - jobb „seek time” (fejpozicionálás ideje)
        - Környezeti tényezőknek jobban ellenáll (hőmérséklet, vibráció, stb.)
    - PS: IDE/SATA
      - Cél az alacsony ár
      - Akkor kap új feature-t, ha nem kerül semmibe :)
        - (Régi IDE pl. csak PIO, a DMA később lett bevezetve)
      - Általában 1-2 db van egy rendszerben
    - Vibráció „teszt”:
      YouTube https://www.youtube.com/watch?v=tDacjrSCeq4

Diszkek

- Diszk (háttértár)
  - S.M.A.R.T.
    - Self Monitoring, Analysis and Reporting Technology
    - Lehetséges hiba előrejelzés
      - Semmire sem biztosíték...
    - Néhány a mért adatok közül:
      - Start/Stop count
      - Power-On Hours (POH)
      - Soft Read Error Rate
      - Reallocation Event Count
      - Temperature
      - SSD Life Left
    - …

Storage


IBM
SATA FRU
43W9715
750 GB E
SATA FRU
43W9715
750 GB E
9
10
11
12
13
15
4Gb/s
Photo by Dr. Dávid Vincze

Diszk beépítő keret

Operációs Rendszerek MSc

Külső SCSI

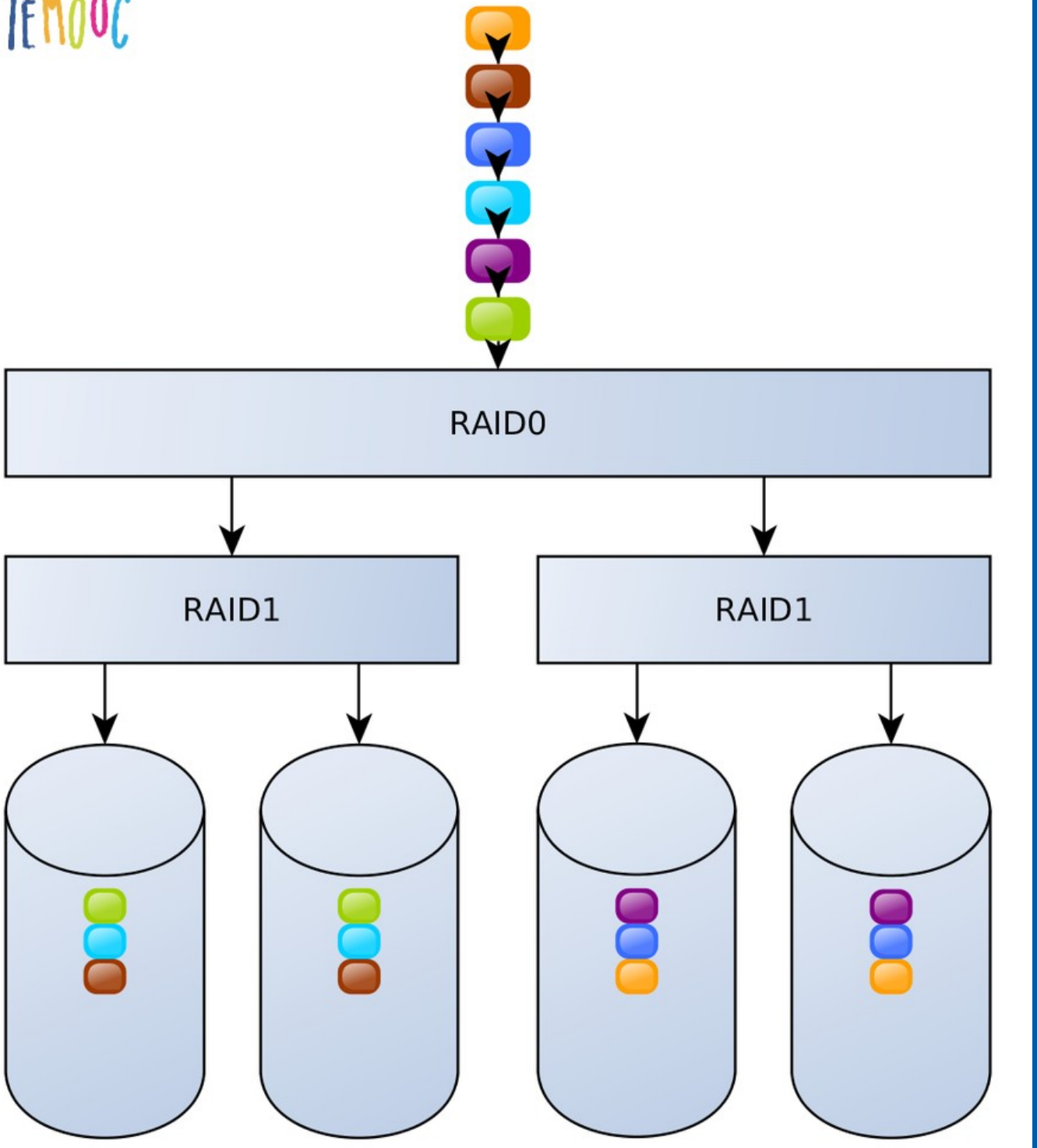
SCSI átalakító

SAS/SATA interfész


Photo by Dr. Dávid Vincze

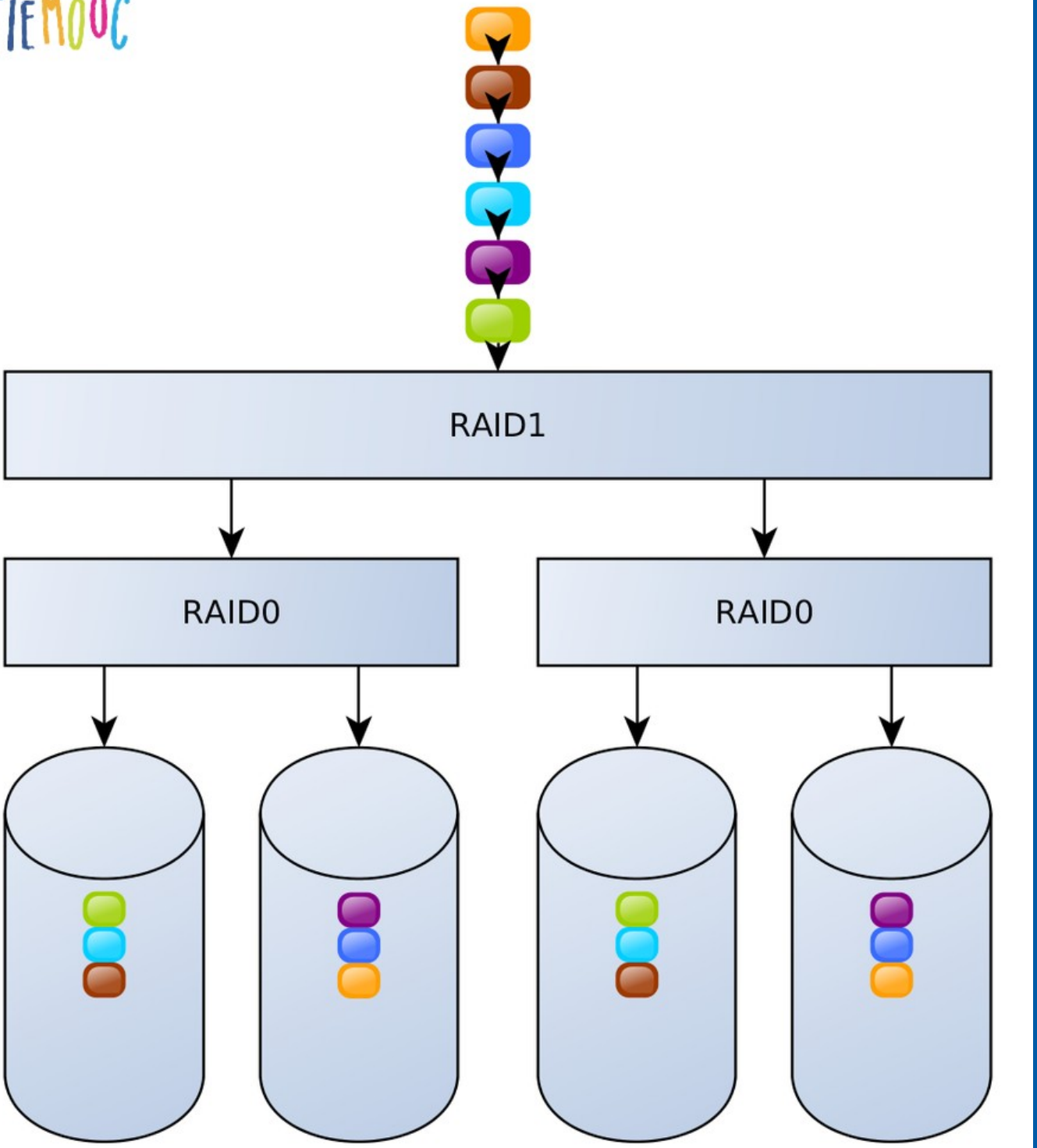
RAID

- RAID
  - Redundant Array of Inexpensive/Independent Disks
  - JBOD: diszkek összefűzve
    - Just a Bunch Of Disks
  - 0: stripe
  - 1: mirror
  - 10: mirror+stripe (min. 4 diszk)
  - 5: parity, xor (min. 3 diszk)
  - 6: 2x parity (min. 4 diszk)
  - egyéb szintek, alverziók, kombinációk is léteznek, de ezek a legelterjedtebbek

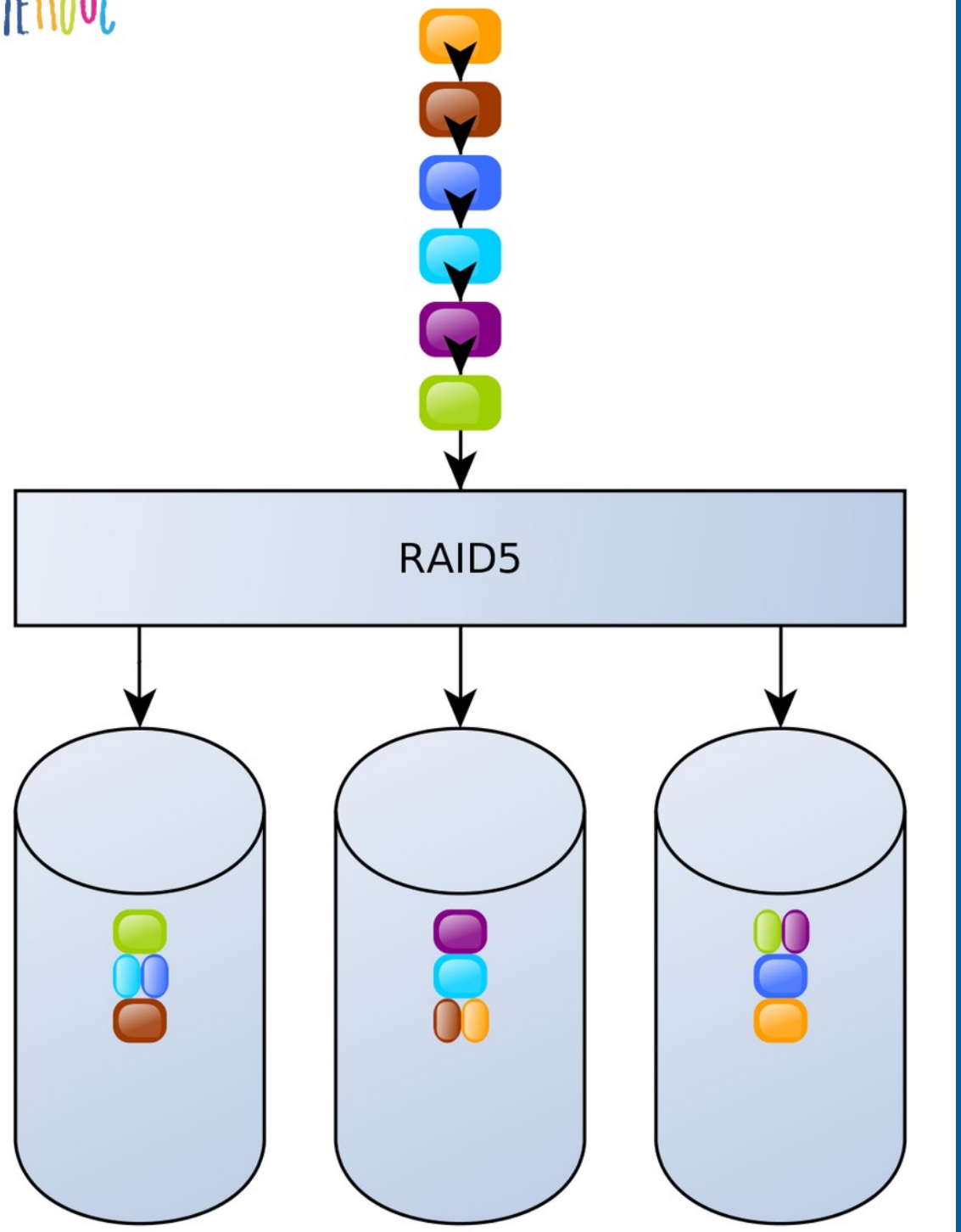


by Dr. Dávid Vincze

0+1

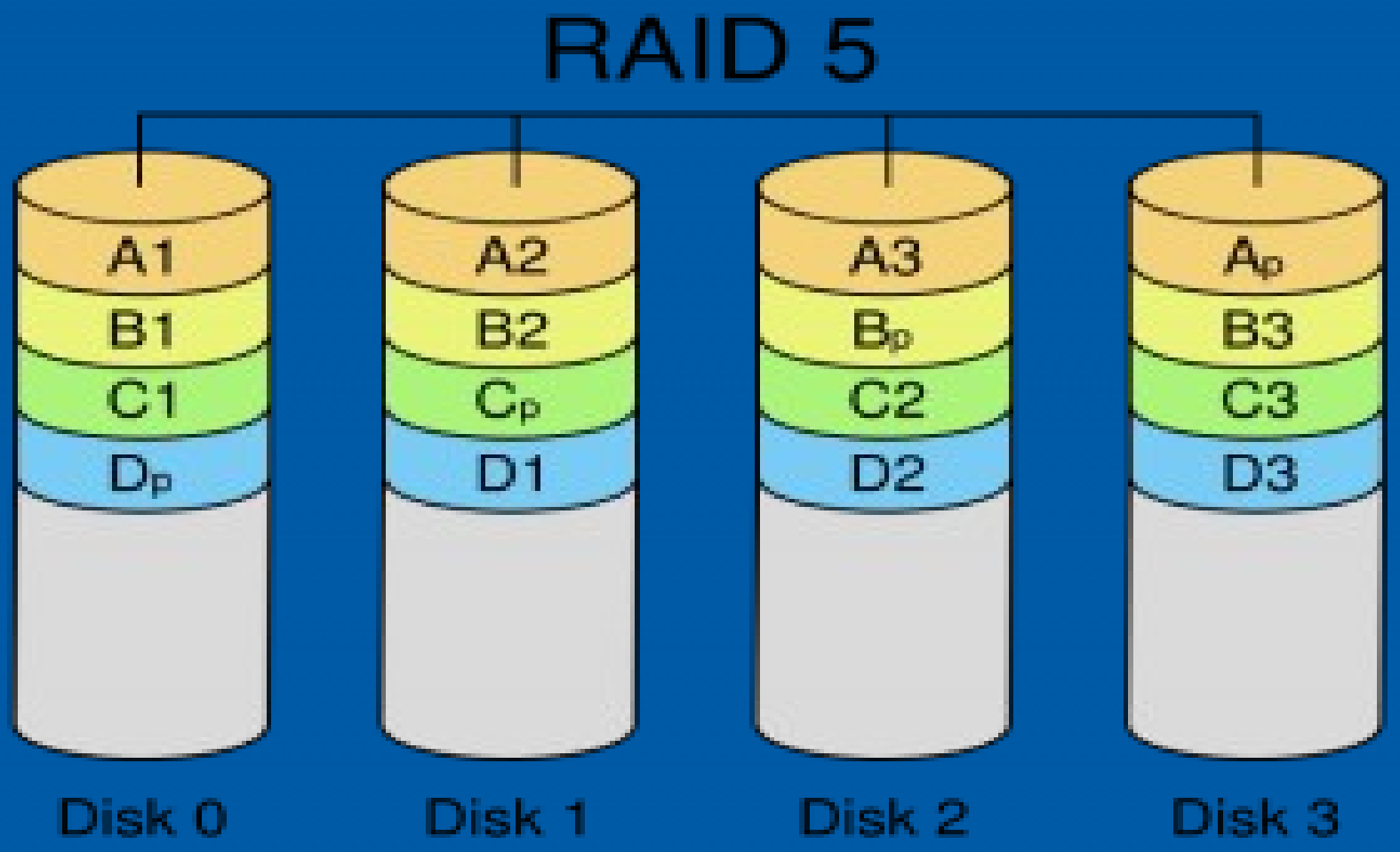


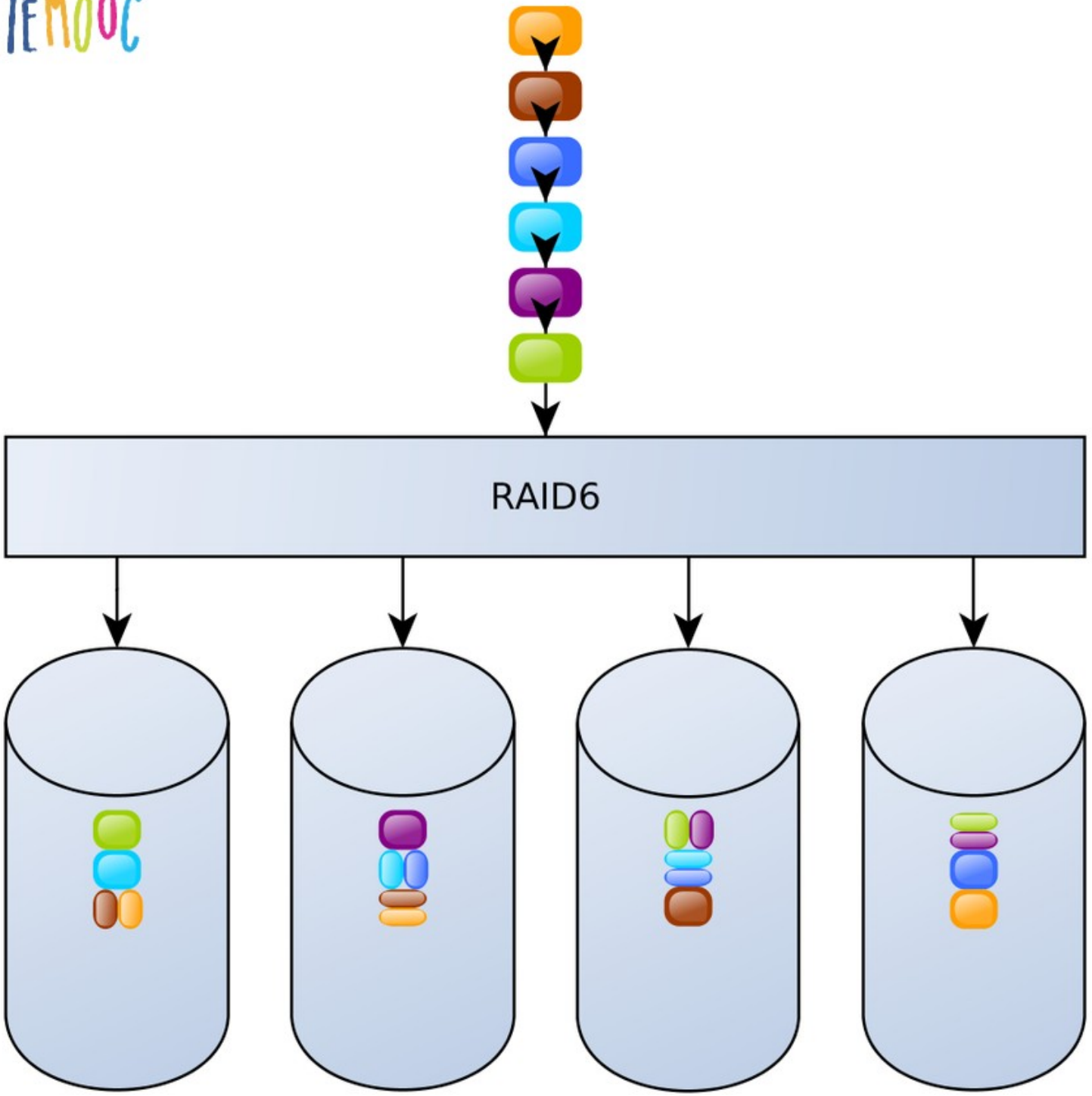
5



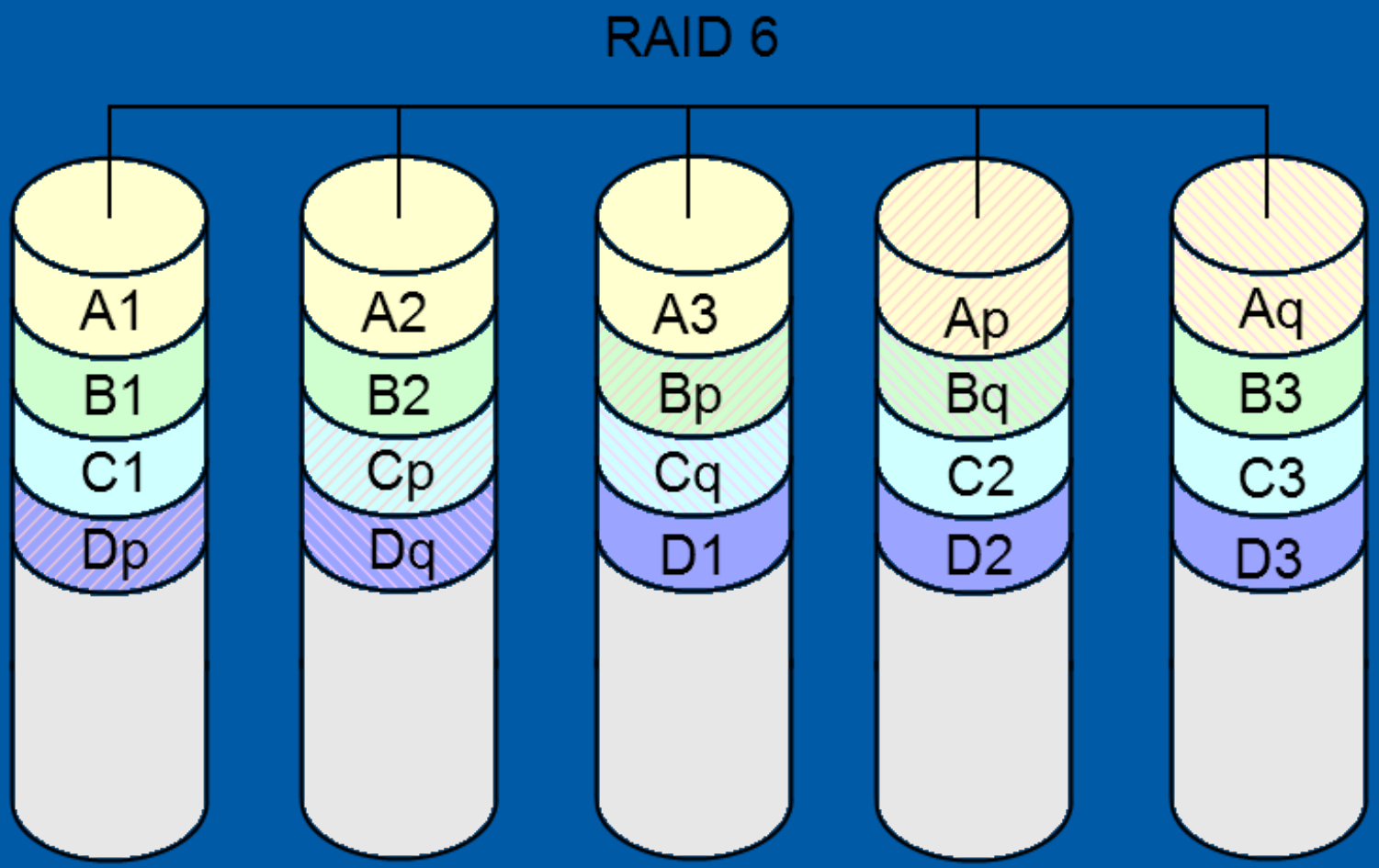
by Dr. Dávid Vincze

RAID


RAID 5
A1
B1
C1
Dp
A2
B2
Cp
D1
A3
Bp
C2
D2
Ap
B3
C3
D3
Disk 0
Disk 1
Disk 2
Disk 3


RAID6

RAID



RAID

- RAID
  - kis adategységek (blokkok): **chunk**
  - **0: stripe (csíkozás)**
    - chunk-okat szétszórja a diszkek között
    - nincs redundancia → adatvesztés ellen nem véd
    - több diszk dolgozik egyszerre → nagyobb teljesítmény
  - **1: mirror (tükrözés)**
    - minden diszken ugyan azok a chunk-ok
    - írás költségesebb (minden diszken tárolni kell)
    - olvasás lehet gyorsabb (több diszkről olvashat egyszerre)
    - van redundancia → csökken a kapacitás

RAID

- RAID
  - **10**: mirror+stripe (min. 4 diszk)
    - 1 és 0 kombinációja
  - **5**: parity, xor (min. 3 diszk)
    - az XOR művelet egyik tulajdonságát használja ki
    - a chunk-okról el van tárolva egy a chunk-okban tárolt adatokból számított érték (pont chunk méretű)
    - ebből 1 chunk hiányában még számítható az eredeti chunk adattartalma
    - 1 diszk teljes meghibásodását átvészeli (N+1 red)
  - **6**: 2x parity (min. 4 diszk)
    - bonyolultabb algoritmus
    - 2 diszk teljes meghibásodását átvészeli (N+2 red)
  - egyéb szintek, alverziók, kombinációk is léteznek, de ezek a legelterjedtebbek

RAID

- RAID5

  ^ : XOR

  A ^ B = C → A ^ C = B → C ^ B = A
  pl. 5 ^ 1 = 4 → 5 ^ 4 = 1 → 4 ^ 1 = 5

  A ^ B ^ C = D → A ^ B ^ D = C →
  A ^ D ^ C = B → D ^ B ^ C = A

  A ^ B ^ C ^ D = E → A ^ B ^ C ^ E = D →
  A ^ B ^ E ^ D = C → A ^ E ^ C ^ D = B →
  E ^ B ^ C ^ D = A

  s.í.t.

RAID

- RAID5

A = [0x01, 0x42, 0xA8, 0xBB];
B = [0x08, 0xF2, 0xA8, 0x00];

P = [0x09, 0xB0, 0x00, 0xBB];

Hiányzó A esetén, A = B ^ P
Hiányzó B esetén, B = A ^ P
Hiányzó P esetén nincs teendő (kivéve diszk csere :))

Mik vannak most?

- HDD: 14TB (Hitachi/Western Digital)
  - Ultrastar DC HC620 (Hs14)
  - 8x 1.75TB diszk
  - Helium
  - 7200 RPM
  - SATA/SAS
  - 4.16 ms average latency
  - Seek time 7.7 (r) / 12.0 (w) ms
  - Szekvencialis elérés:
    223 MiB/s – 233 MB/s
  - $500-$600 ???

http://www.hgst.com/sites/default/files/resources/Ultrastar-Hs14-DS.pdf

Mik vannak most?

- HDD: 900GB (pl. Seagate Enterprise Perf.)
  - 3 lemez
  - 15 000 RPM
  - SAS 12Gbit/s
  - 2 ms average latency
  - Seek time: 2.9 ms
  - Szekvencialis elérés:
    300 MB/s
  - ~$560

Mik vannak most?

- SSD: 15.36TB (pl. Seagate Nytro)
  - 2x SAS 12Gbit/s
  - 120 µs average latency
  - ~~Seek time~~
  - Szekvencialis elérés:
    2100 MB/s (r) - 1690/1780 MB/s (w)
  - IOPS ?
  - $??? ...

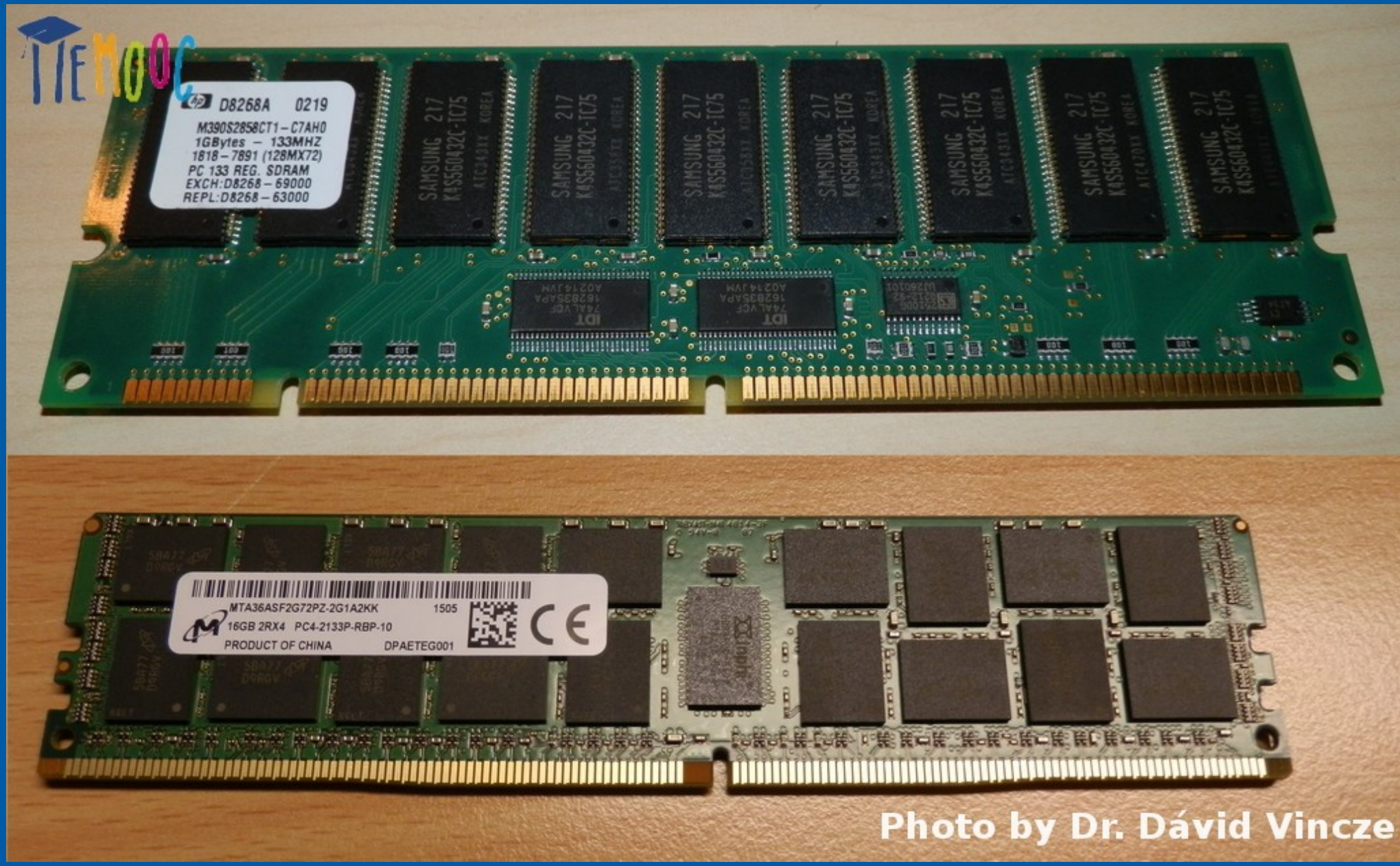
Mik vannak most?

- SSD: 7.7TB
  - (pl. Seagate Nytro XP7200)
  - PCIe x16
  - ? µs average latency
  - ~~Seek time~~
  - Szekvencialis elérés:
    10 GB/s (r) - 2300 MB/s (w)
  - IOPS !
  - $??? ...

RAM

- Memória (RAM)
  - ECC
    - Error Checking and Correcting / Error Correcting Code
    - (több fajta, 1 bit hiba detektálás, 2 bit hiba detektálás + 1 bit javítás, stb.)
    - memória IC (chip ha úgy tetszik)
  - registered / buffered
    - vonali erősítő
    - hogy minél többet lehessen egy memória buszra tenni (ne a vezérlő meghajtó áramköre legyen terhelve)
    - 1 órajelet késleltet
      - nem tehető vegyesen nem regiszteressel!
  - ChipKill (IBM)
  - RAIM

Reg. ECC – SDRAM – DDR



ECC RAM – hibák?

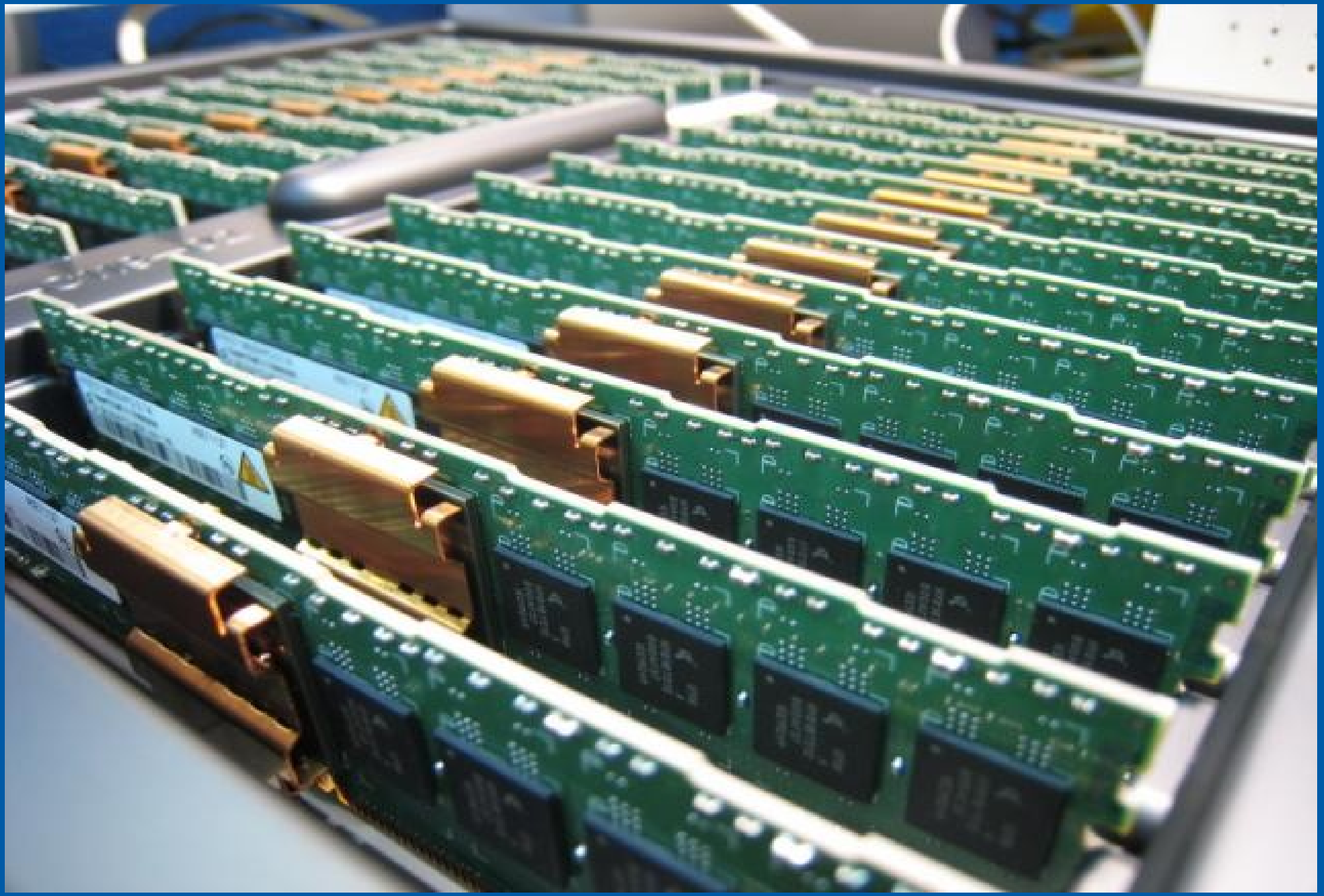
- Memória (RAM)
  - ECC
    - **https://spectrum.ieee.org/computing/hardware/how-to-kill-a-supercomputer-dirty-power-cosmic-rays-and-bad-solder**

„Jaguar had 360 terabytes of main memory, all protected by ECC. I and others at the lab set it up to log every time a bit was flipped incorrectly in main memory. When I asked my computing colleagues elsewhere to guess how often Jaguar saw such a bit spontaneously change state, the typical estimate was about a hundred times a day. In fact, Jaguar was logging ECC errors at a rate of 350 per minute.”

IBM ChipKill

- ChipKill (IBM) – IBM Whitepaper
  - *„IBM engineers have solved this problem for Netfinity servers by placing a **Redundant Array of Inexpensive DRAM (RAID)** processor chip **directly on the memory DIMM**. The RAID chip calculates an ECC checksum for the contents of the entire set of chips for each memory access and stores the result in extra memory space on the protected DIMM. Thus, when a memory chip on the DIMM fails, the RAID result can be used to "back up" the lost data, allowing the Netfinity server to continue functioning. This RAID technology is similar to the RAID technology used to protect the contents of an array of disk drives. We call this memory technology Chip-kill DRAM.”*

Reg. ECC RAM

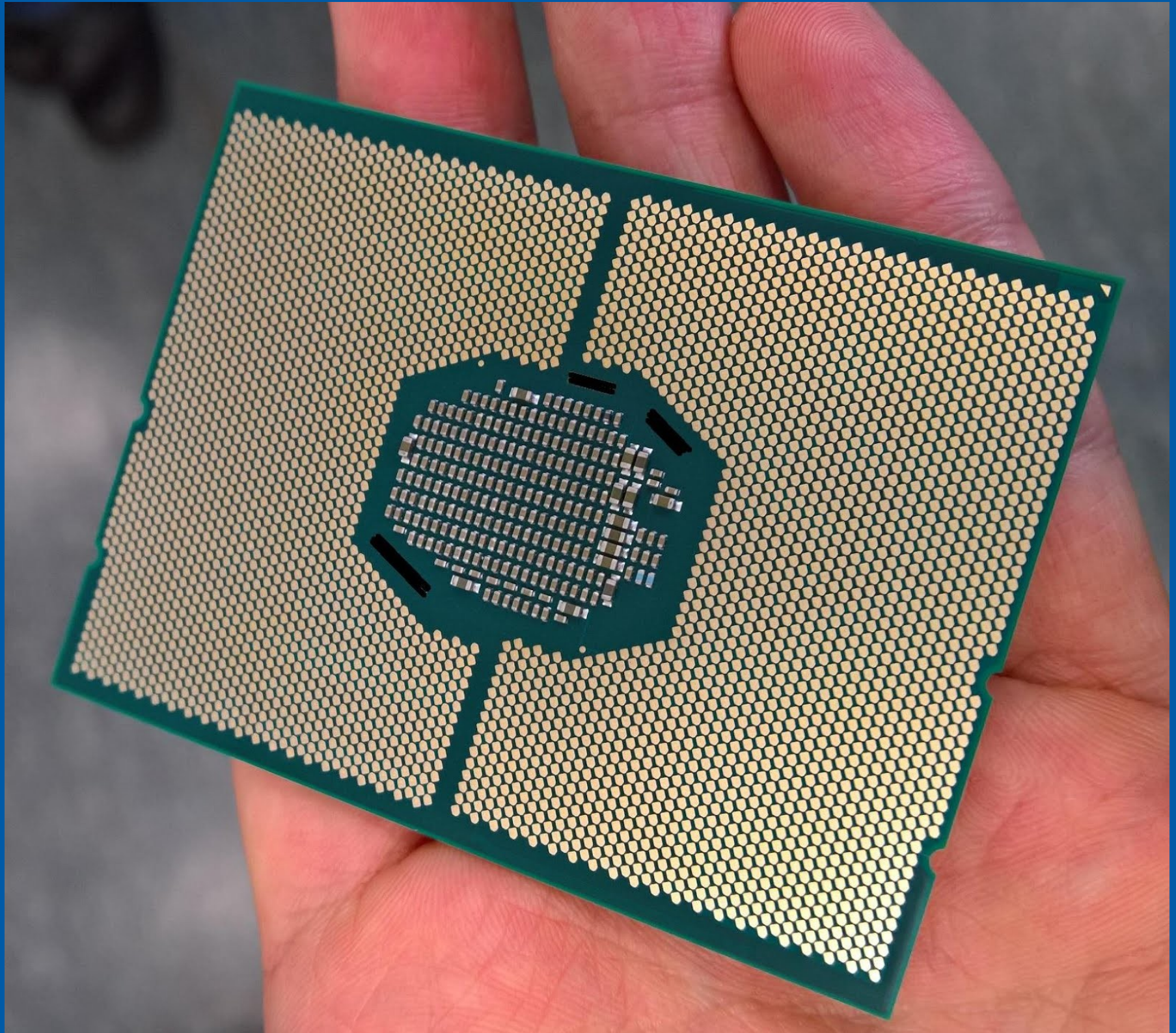


Processzorok

- CPU
  - socket
  - core
  - thread
- Mik vannak most?

Intel x86

- Intel Xeon
  - x86_64 / x64
  - Xeon Platinum 8180
    - 28-core, 2.5GHz, 28x1MB L2 cache, 38.5MB L3
    - 2 thread/core, 14 nm
    - 3x UltraPath Interconnect (UPI)
    - 8 CPU/system
    - AVX-512
    - 2017 Q3 ~ $13000+
  - Xeon E3-1285 v6
    - 4-core, 4.1GHz (4.5 GHz), 4x256KB L2, 8MB L3
    - 2 thread/core, 14 nm
    - 1 CPU/system
    - AVX-512
    - 2017.08. ~ $450+



IBM POWER

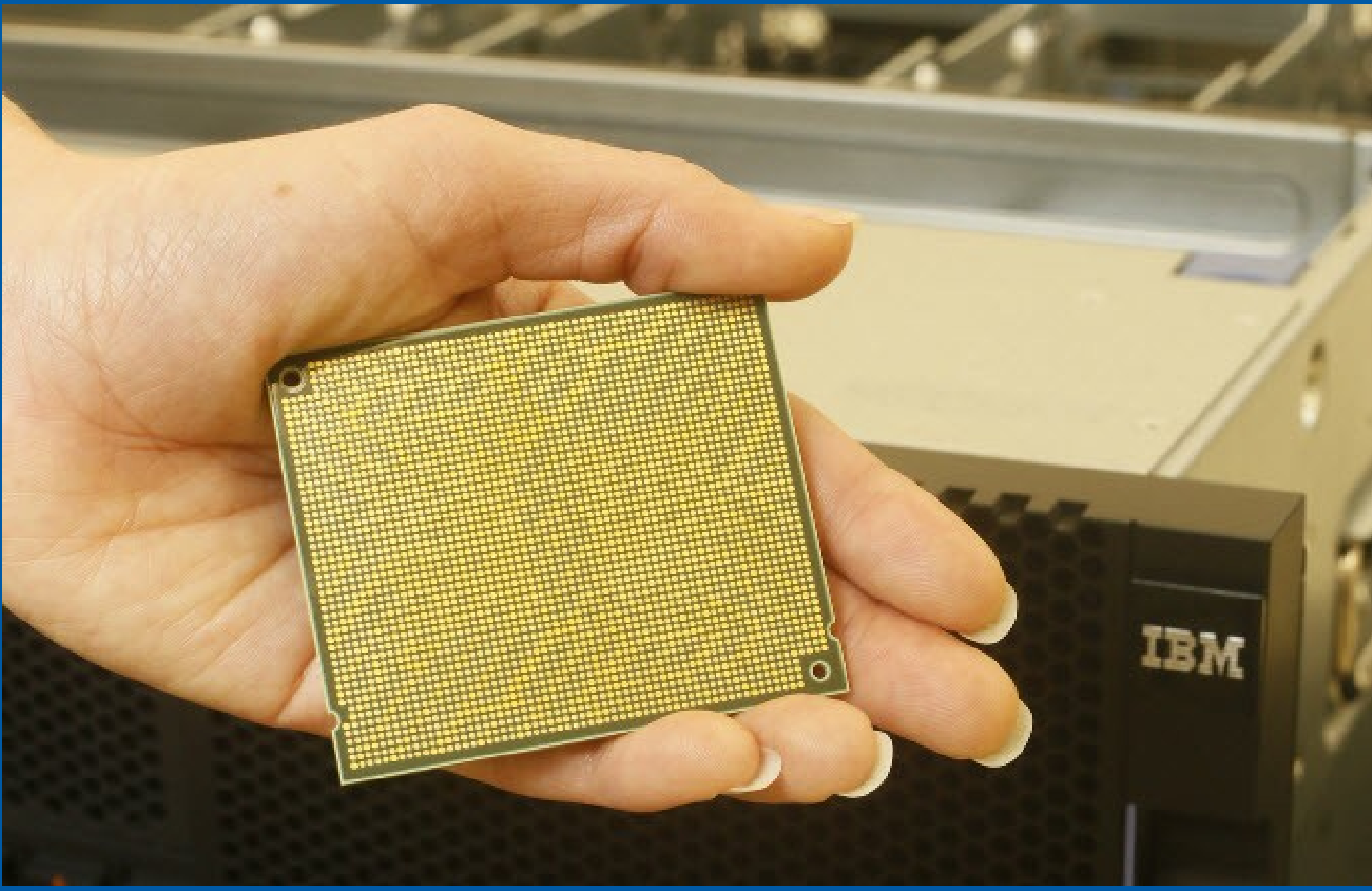
- IBM POWER

  - Performance Optimized With Enhanced RISC
  - POWER8 (with NVLink)
  - 6-core 4.35GHz
  - 8-core 4.02GHz
  - 4-socket/building block (max. 16-socket/system)
  - 512KB L2 cache/core
  - 8MB L3, max 128MB L4
  - 22nm

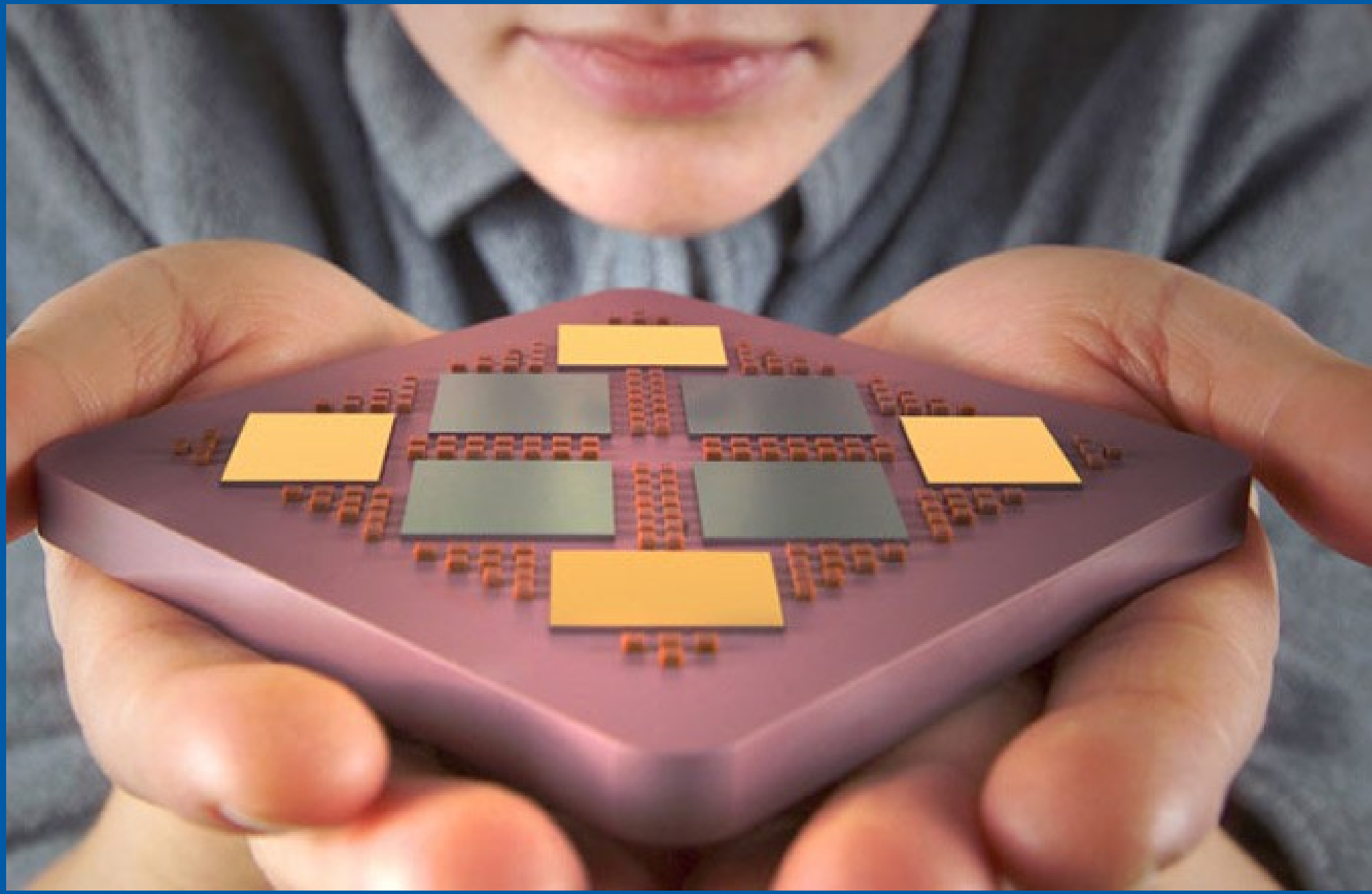
IBM POWER

- IBM POWER

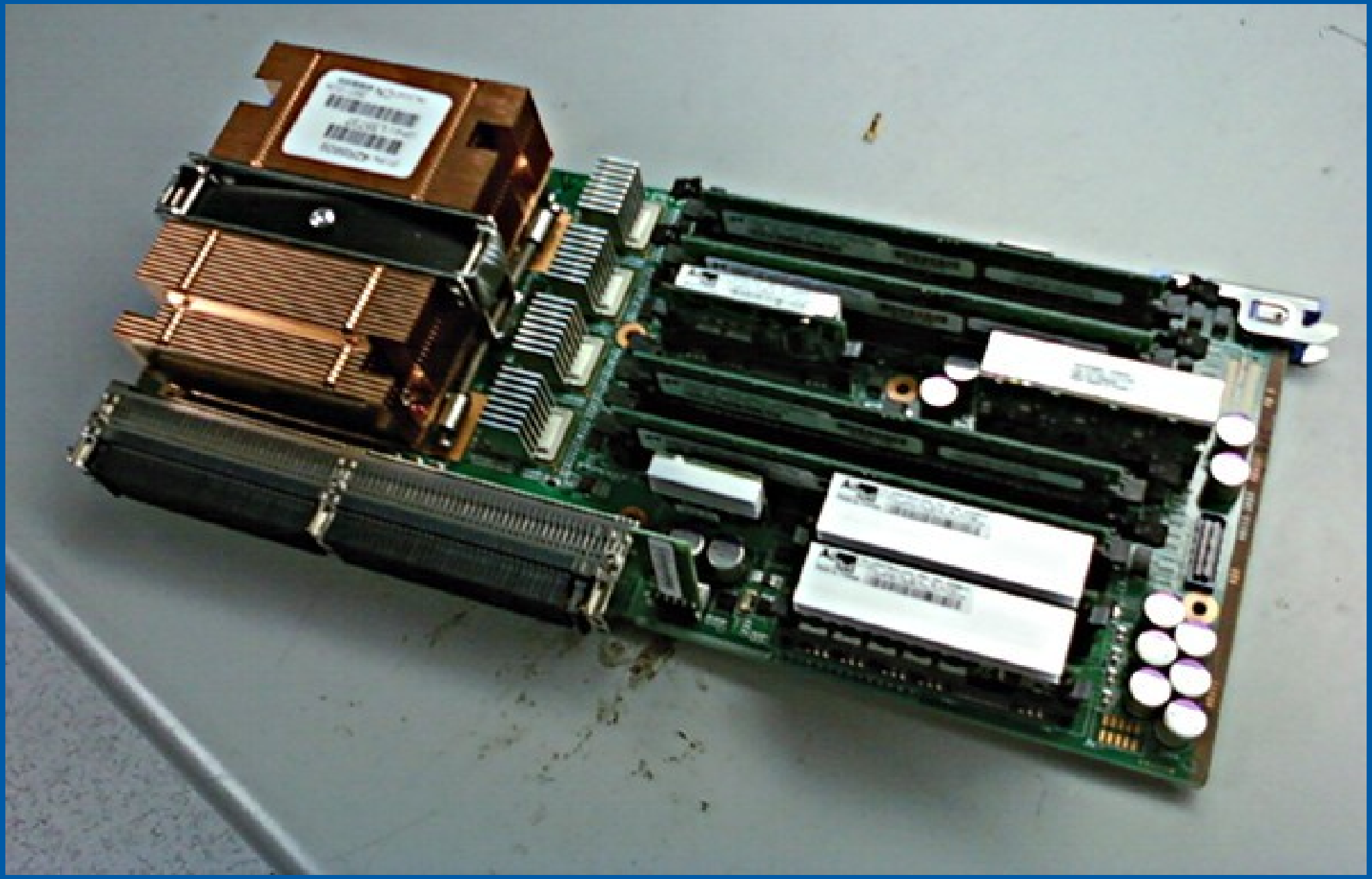
  - Performance Optimized With Enhanced RISC
  - POWER9
  - 14nm
  - 24-core SMT4 / 12-core SMT8
  - L1I + L1D = 32kb + 32kb
  - L2: 512kb/mag
  - L3: 120Mb/socket
  - 7TB/sec on-chip bandwidth
  - ~4GHz órajel


IBM

POWER



POWER



POWER

SPARC

- SPARC
  - Scalable Processor ARChitecture
  - SPARC M7
    - 4.133 GHz, 32-core, 8thread/core → 256 thread CPU
    - 16+16 L1cache (I/D)
    - 256KB/4core+128KB/2core L2 cache, 64MB L3 cache
    - Max 16 CPU/system (16 socket)
    - 20nm, 2015
  - SPARC S7
    - 4.27 GHz, 8-core, 8 thread/core → 64 thread CPU
    - 256x2KB+256x4KB L2 cache, 16MB L3 cache
    - 20nm, 2016
  - Sonoma
    - buta M7 + 2x 56Gbit Infiniband

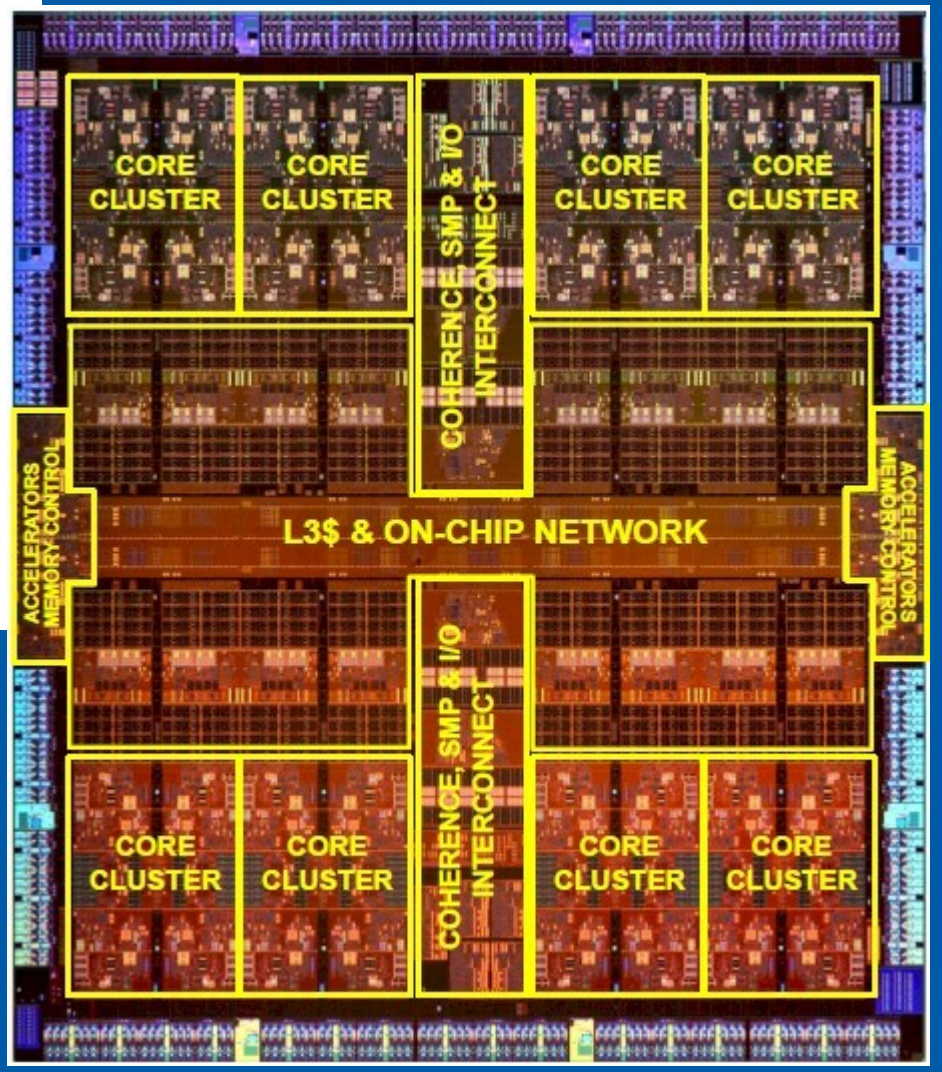
SPARC

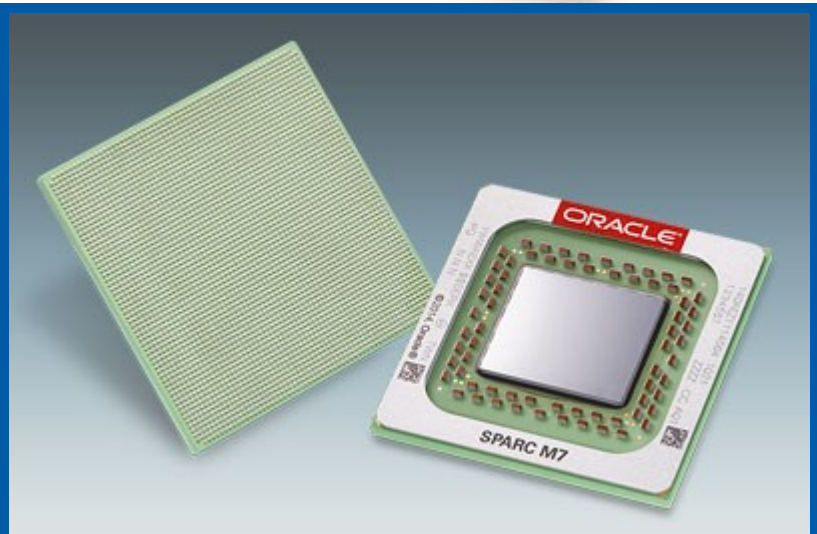
- SPARC
  - Scalable Processor ARChitecture
  - SPARC M8
    - 5.06 GHz, 8-core, 32thread/core → 256 thread CPU
    - 32K+16KB L1 cache (I/D) / core
    - 256KB/4core+128KB/core L2 cache (I/D)
    - 64MB L3 cache / socket
    - Max 8 CPU/system
    - DAX – Data Analytics Accelerator
    - 20nm, **2017.09.18.**

    - **http://www.oracle.com/us/products/servers-storage/sparc-m8-processor-ds-3864282.pdf**

SPARC


ORACLE
SPARC M7
©2015, Oracle®
72GZL6C SERXXX
PG 2.1.0
TWN
140AEZ114564
3600 .6 A01
7048884
1021


CORE CLUSTER
CORE CLUSTER
COHERENCE, SMP & I/O INTERCONNECT
CORE CLUSTER
CORE CLUSTER
ACCELERATORS MEMORY CONTROL
L3$ & ON-CHIP NETWORK
ACCELERATORS MEMORY CONTROL
CORE CLUSTER
CORE CLUSTER
COHERENCE, SMP & I/O INTERCONNECT
CORE CLUSTER
CORE CLUSTER


ORACLE
SPARC M7

Hálózat

- Network
  - Ethernet / Fast Ethernet / GigaBit Ethernet / 10 GigaBit Ethernet / ... (2.5/5/10/25/40/50/100/200/400Gbit/s)
  - (ATM, SONET/SDH, WAN interfészek, stb.)
  - Nem csak az áteresztő képesség a fontos, hanem a válaszidő is!
  - Média (átviteli közeg)
    - rézdrót (copper)
      - (coax, UTP, STP, FTP)
    - fényvezető / üvegszál / száloptika / stb. (fiber optics)
      - EM zajra érzéketlen
      - Multimode (MM), SingleMode (SM) / monomódus, gradiens
      - BWDM, CWDM, DWDM

Hálózat

- Network
  - Channel / Trunk / Bond
    (ahány gyártó annyi elnevezés...)
    - nagyobb sávszélesség
    - redundancia
    - különböző csomag szétosztási algoritmus
  - VLAN
    - szeparáltság
    - link spórolás

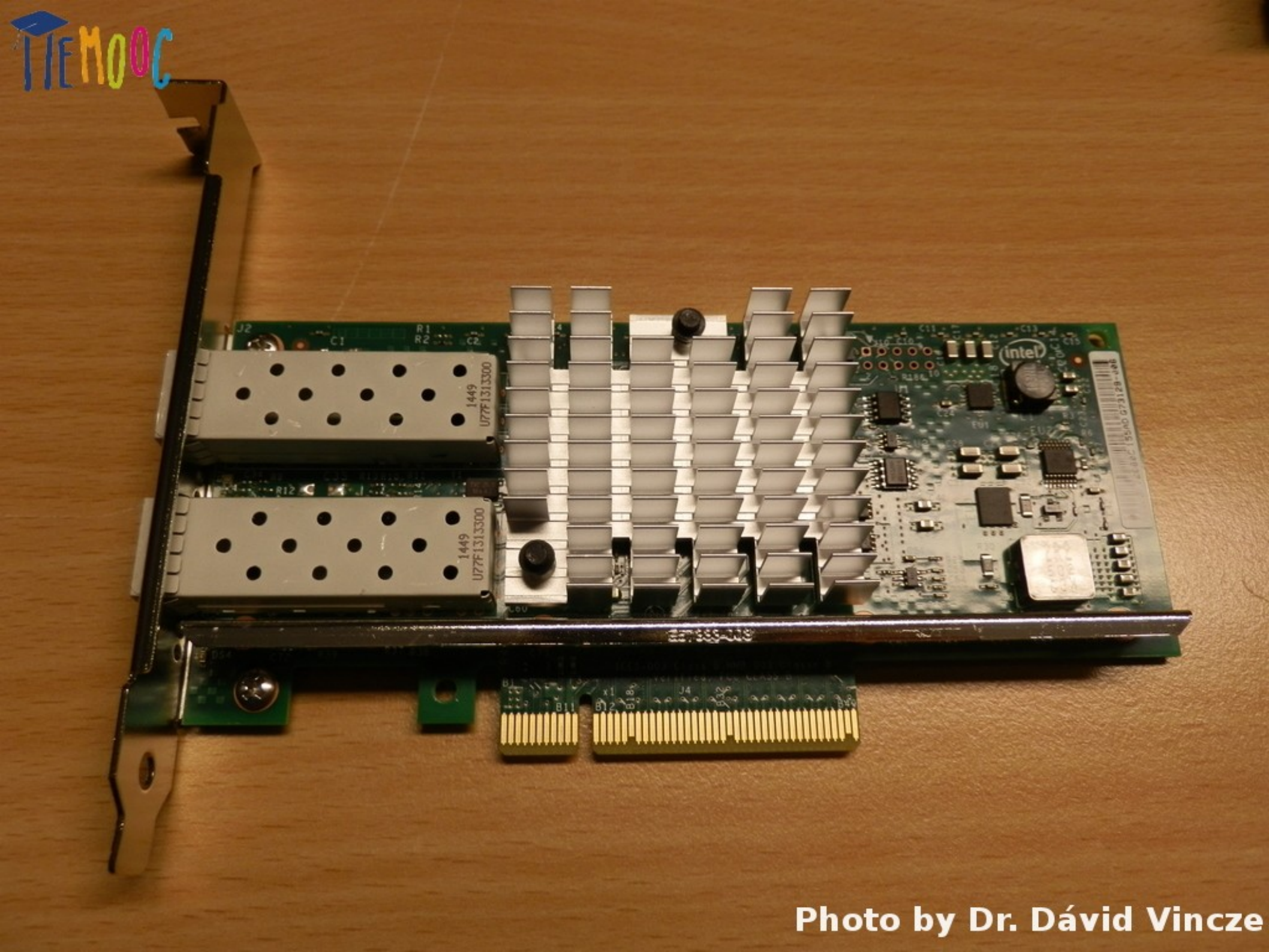


Photo by Dr. Dávid Vincze

Photo by Dr. Dávid Vincze


intel

Management

- Menedzselhetőség
  - moduláris felépítés
  - rack szerelhető (rack-mount)
  - szervízprocesszor (IPMI, HP ILO, IBM ASM/RSA, stb. stb.)
  - RAS - Reliability, Availability and Serviceability

Rendelkezésre állás

- Minimum down-time
  - hot-swap disk
  - hot-swap PSU (Power Supply Unit)
  - hot-swap fan (ventilátor)
  - hot-plug PCI
  - hot-add memory
  - hot-swap memory
  - hot-swap CPU
  - CPU on demand
    - capacity on demand
  - klaszterezés...
    - később részletesebben

Redundáns PSU

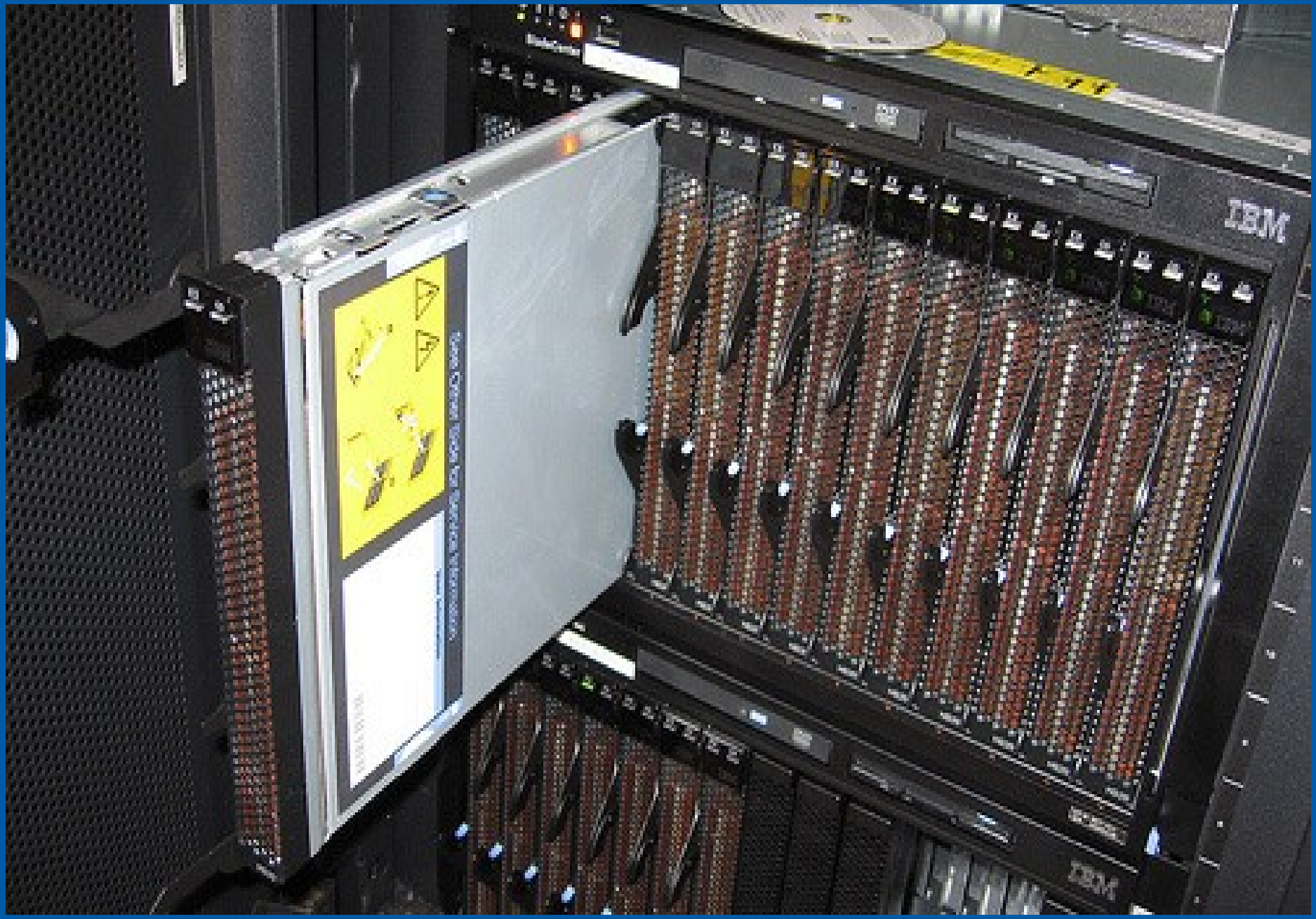
https://www.youtube.com/watch?v=AGaqEnU8sy0

Redundáns PSU

Blade szerverek

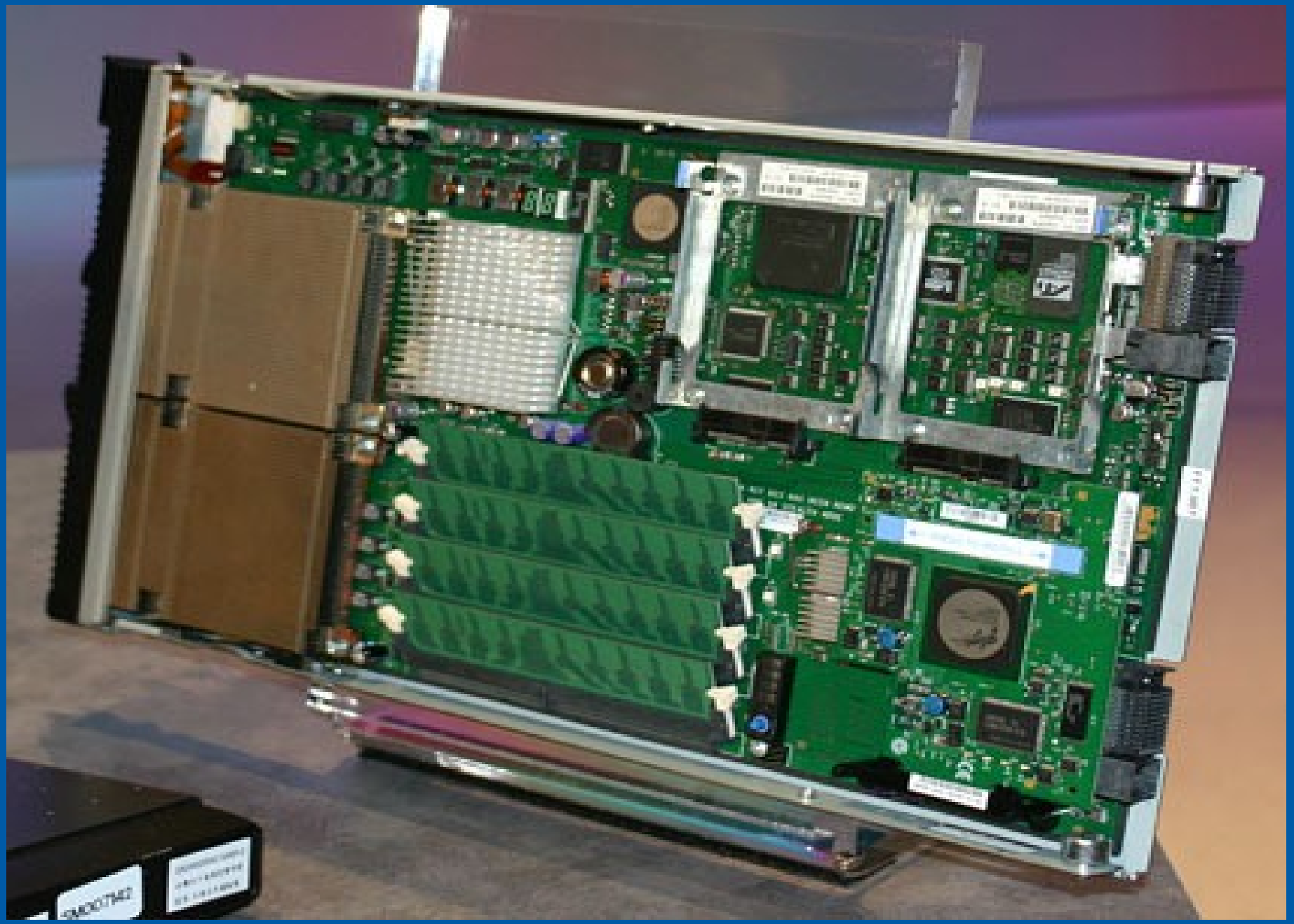
- Blade szerverek
  - Egy nagy keret (chassis), ami közösen biztosít:
    - Áramellátást
    - Hűtést
    - Hálózatot
    - Egyéb interconnect-et
    - Menedzsmentet
  - Maguk a gépek így még kisebbek tudnak lenni
    - Nagyobb sűrűség
  - Összesített fogyasztás csökkenthető
  - Ettől persze ezek különálló független gépek.

Blade szerverek



Blade szerverek

Blade szerverek



SuperBlade SBI-8149P-T8N

(Angled View – Blade Server)

SuperBlade SBI-8149P-T8N

(Rear View – Blade Server)

Redundant 2200 W
Titanium Level Power Supply

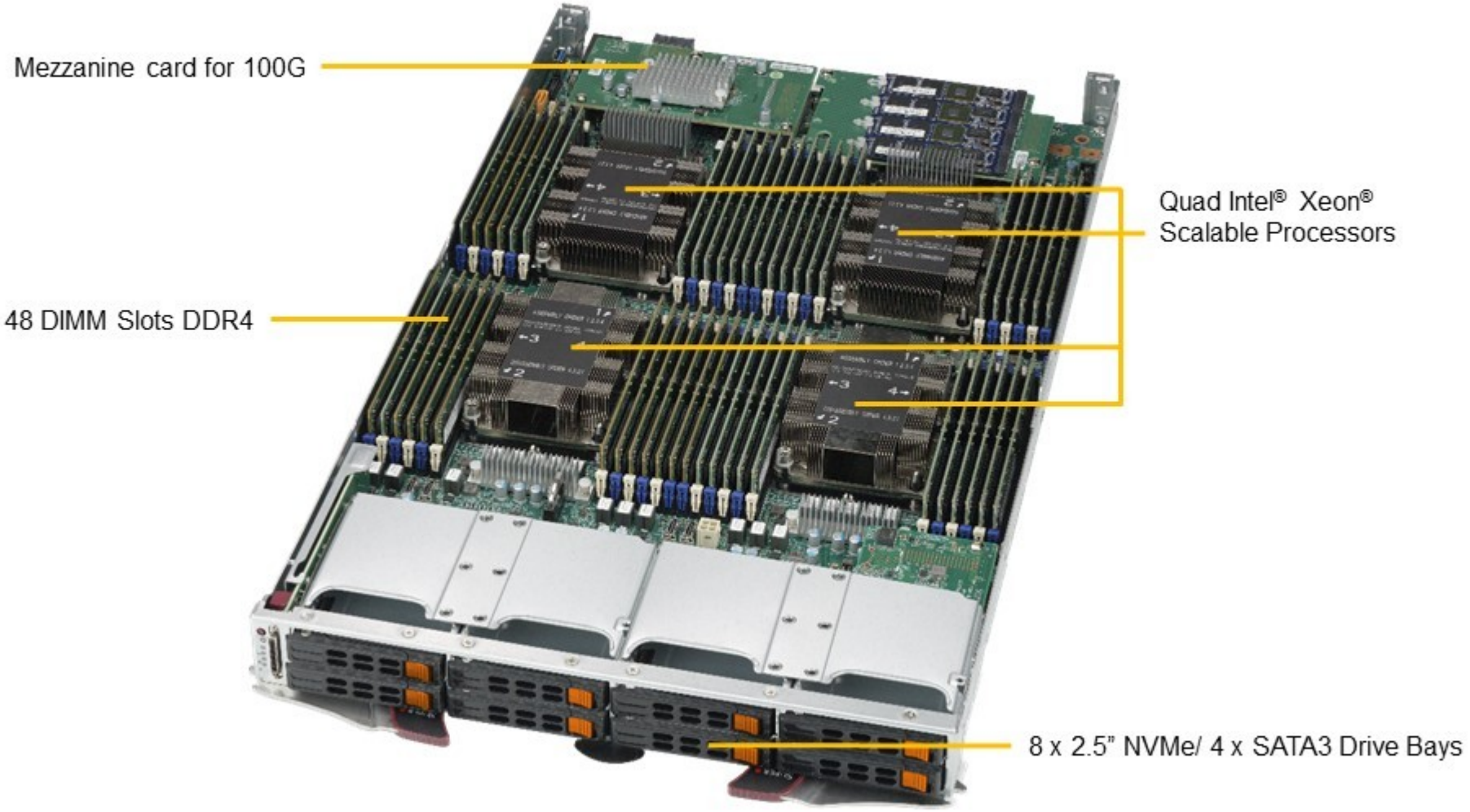
100G OPA/IB Switch

10G Ethernet Switch

Chassis Management Module

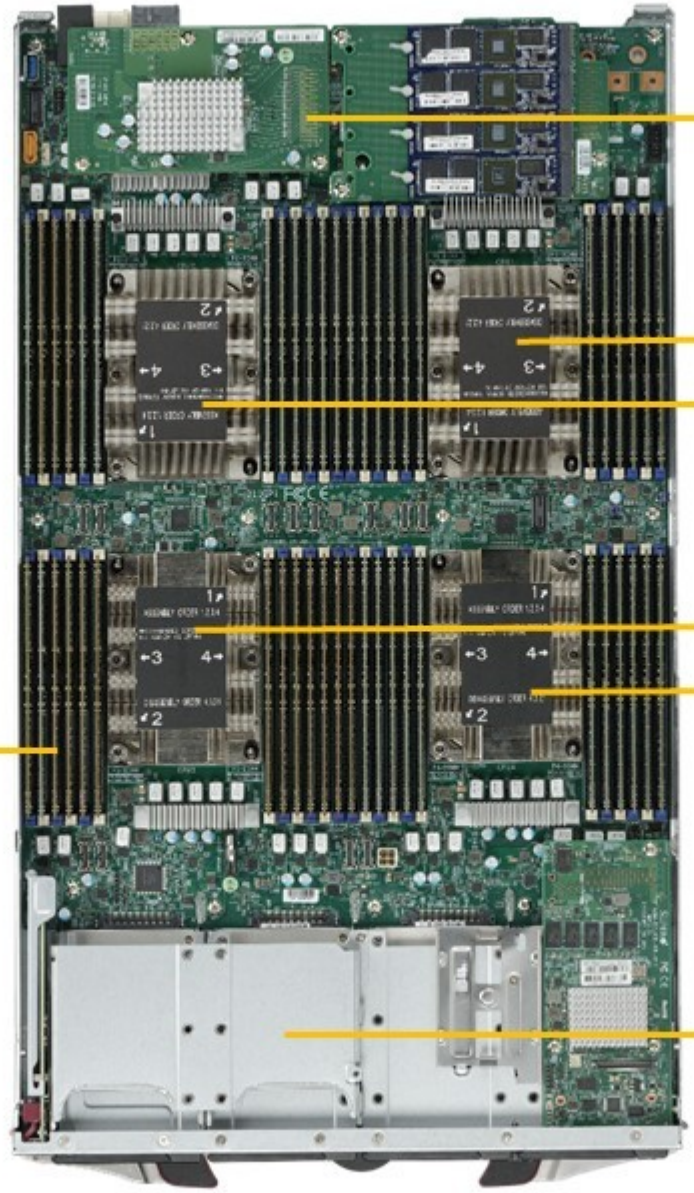
SuperBlade SBI-8149P-T8N

(Angled View – Node)



SuperBlade SBI-8149P-T8N

(Top View – Node)



Mezzanine card for 100G

Quad Intel® Xeon® Scalable Processors

48 DIMM Slots DDR4

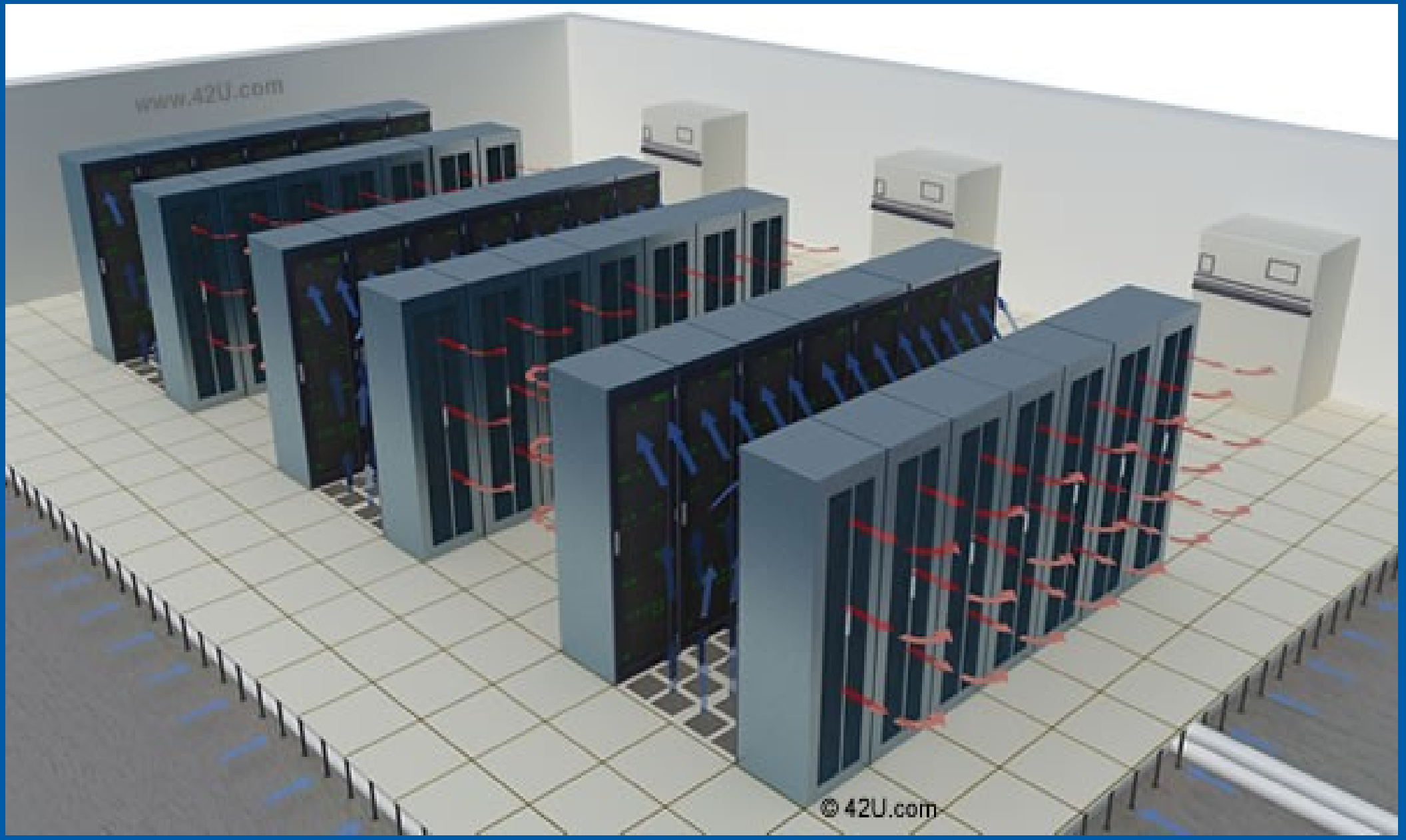
8 x 2.5" NVMe / 4 x SATA3 Drive Bays

Adatközpont környezet

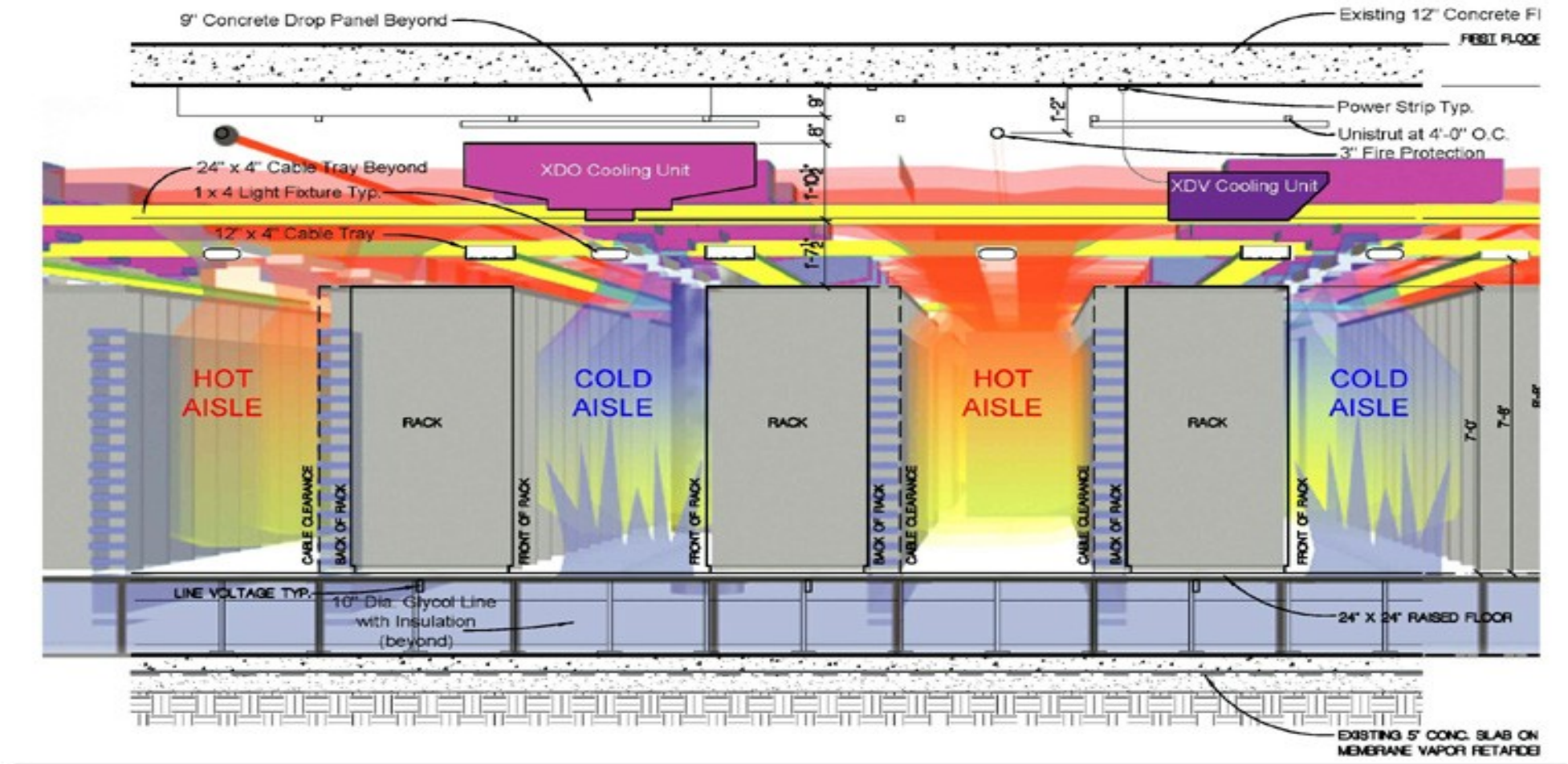
- Megfelelő (fizikai) környezet
  - klimatizált gépterem - Computer Room Air Conditioning (CRAC)
    - hőmérséklet
    - páratartalom
    - hűtött, szűrt levegő, elszívók
  - cold aisle - hot aisle elrendezés
  - szünetmentes tápellátás (CRAC-nak is!)
    - több független betáplálás
    - akkumulátorok
    - generátorok
  - automatikus tűzoltó berendezés
  - álpadló, álmennyezet
  - fizikai biztonság

YouTube https://www.youtube.com/watch?v=OiWYxOi4VyY

Hideg sor – Meleg sor

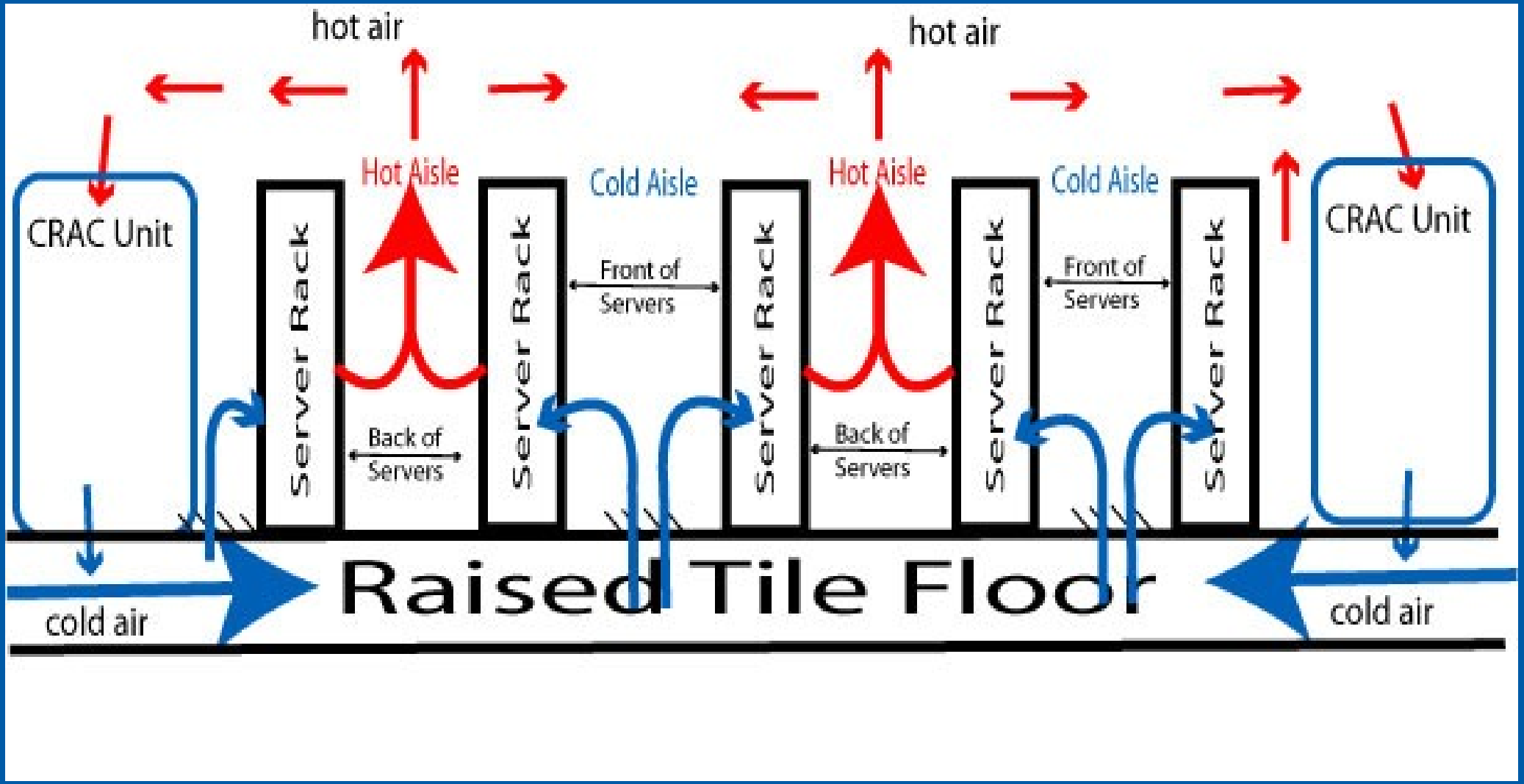


Hideg sor – Meleg sor



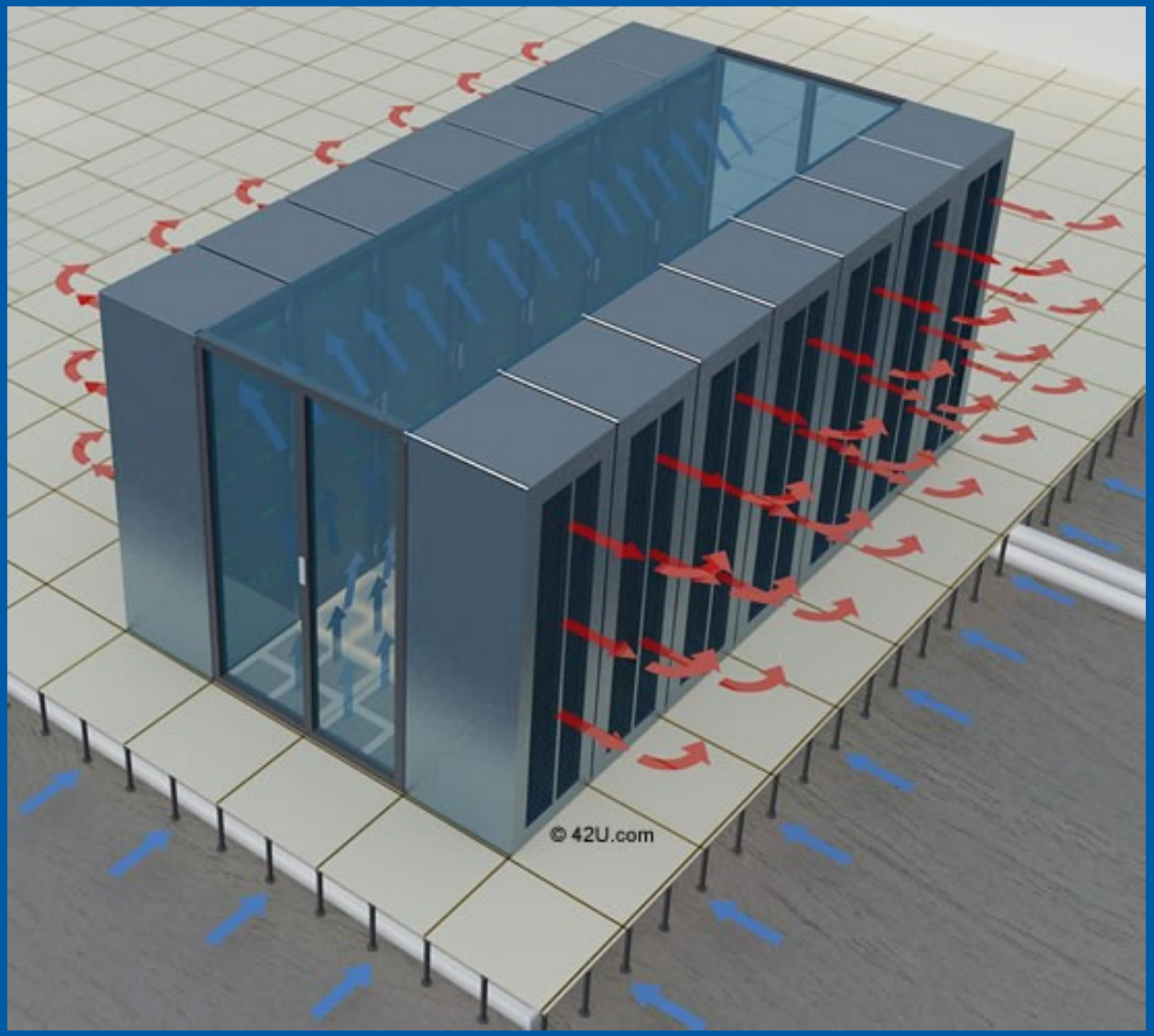
Data Center Floor Plan
Power, Data, and Lighting

Hideg sor – Meleg sor

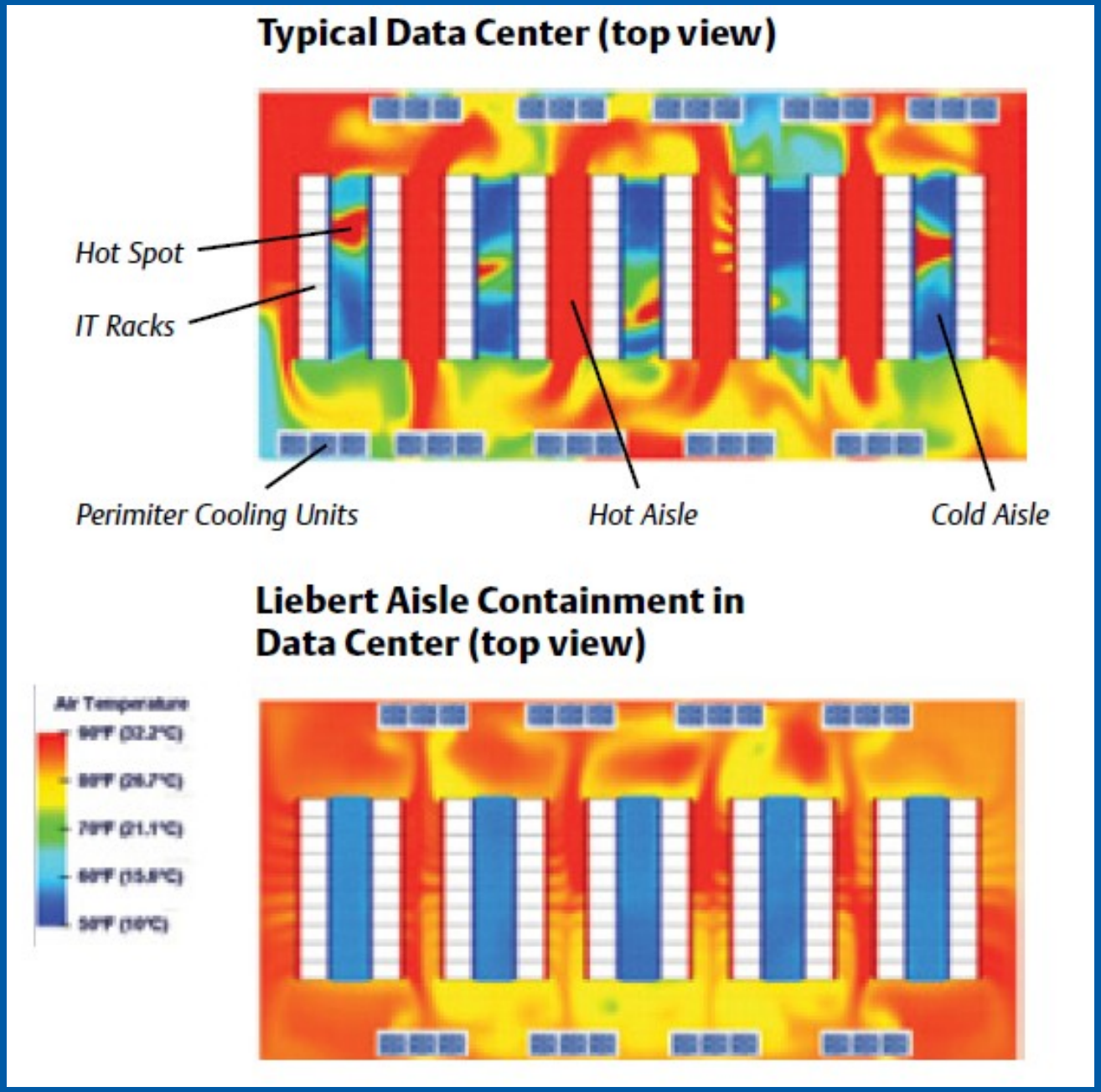


Sorok elkülönítése

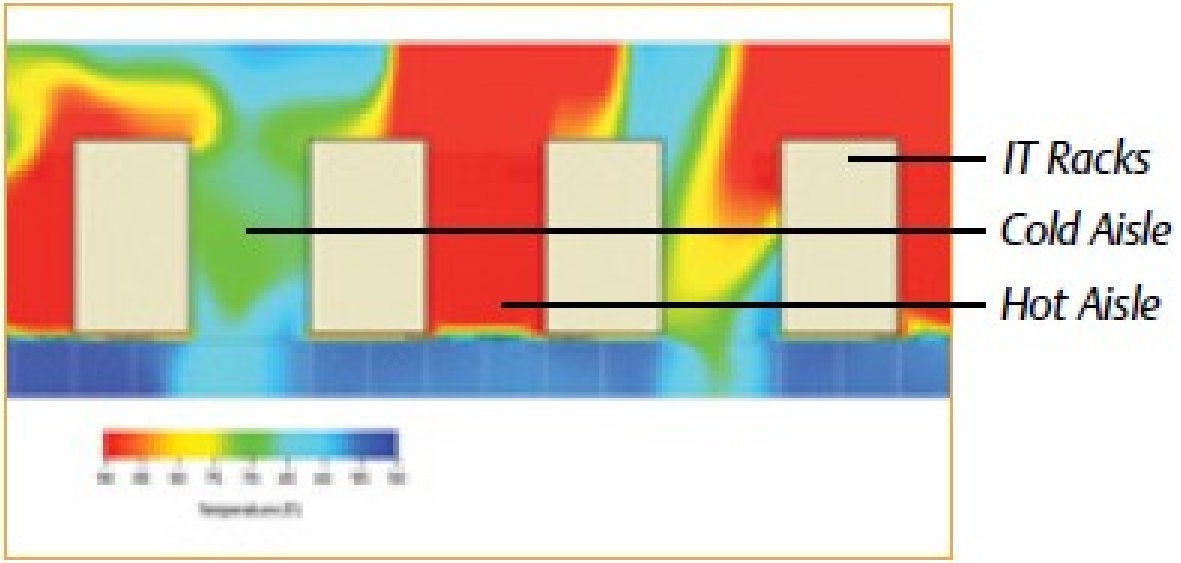
- Aisle Containment



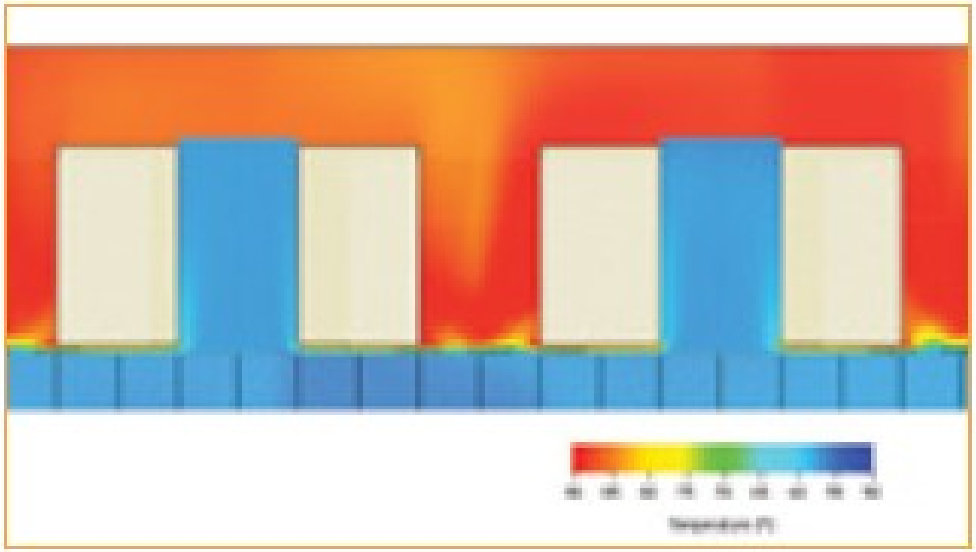
Sorok elkülönítése



Typical Data Center (end of row view)



Data Center with Aisle Containment System / Liebert iCOM controls (end of row view)



Adatmentés

- Backup - Tape / szalag

- https://spectrum.ieee.org/computing/hardware/why-the-future-of-data-storage-is-still-magnetic-tape

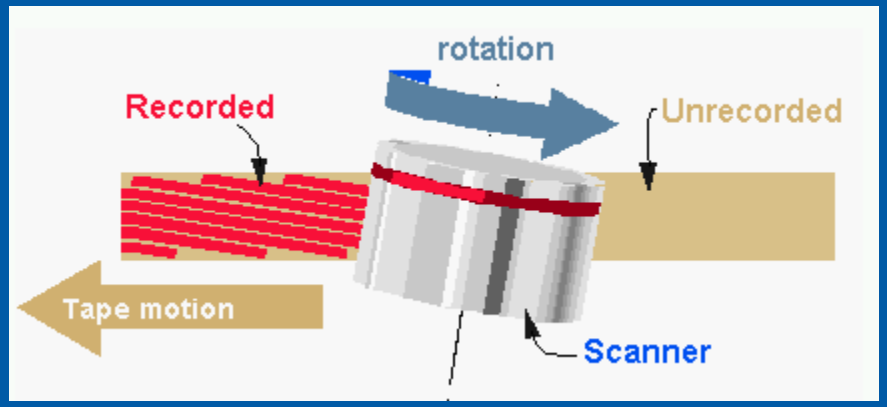
  - **LTO / Ultrium**
    - Linear-Tape Open
    - LTO-8 – 12 TB , 360 MB/sec, 6656 track, 960m (2017 dec)
    - 15-30 év, 5000 betöltés, 32 track/wrap (208x oda-vissza)
    - titkosítás, 2.5:1 arányú tömörítés (marketing...)
    - 5h 50m teleírni

  YouTube https://www.youtube.com/watch?v=75xm3JMxWE0

  YouTube https://www.youtube.com/watch?v=JF-0qPxJ_fw

Adatmentés

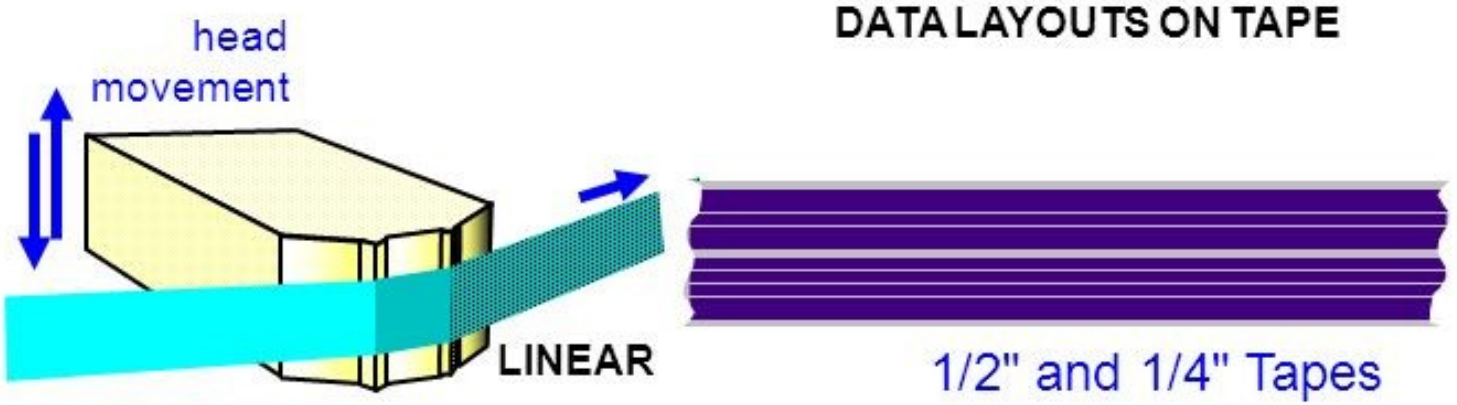
- Backup - Tape / szalag
  - **DAT/DDS**
    - Digital Audio Tape / Digital Data Storage
    - DAT 320 / DDS-7 (2009)
    - 160 GB, 43.2 GB/hr (12MB/s), 150m
    - Kihalófélben (DDS-8 már nem készült el)
    - Helikális rögzítés



https://www.youtube.com/watch?v=CXKyL0HIy8Y
https://www.youtube.com/watch?v=0WYohFfVDk4

  - sok más egyéb formátum (DLT, VXA, Mammoth, stb.)
  - 2:1 (3:1) átlag tömörítés (marketing...)

Linear & Helical Scan Recording Methods



ULTRIUM

DLT

DLT VS80

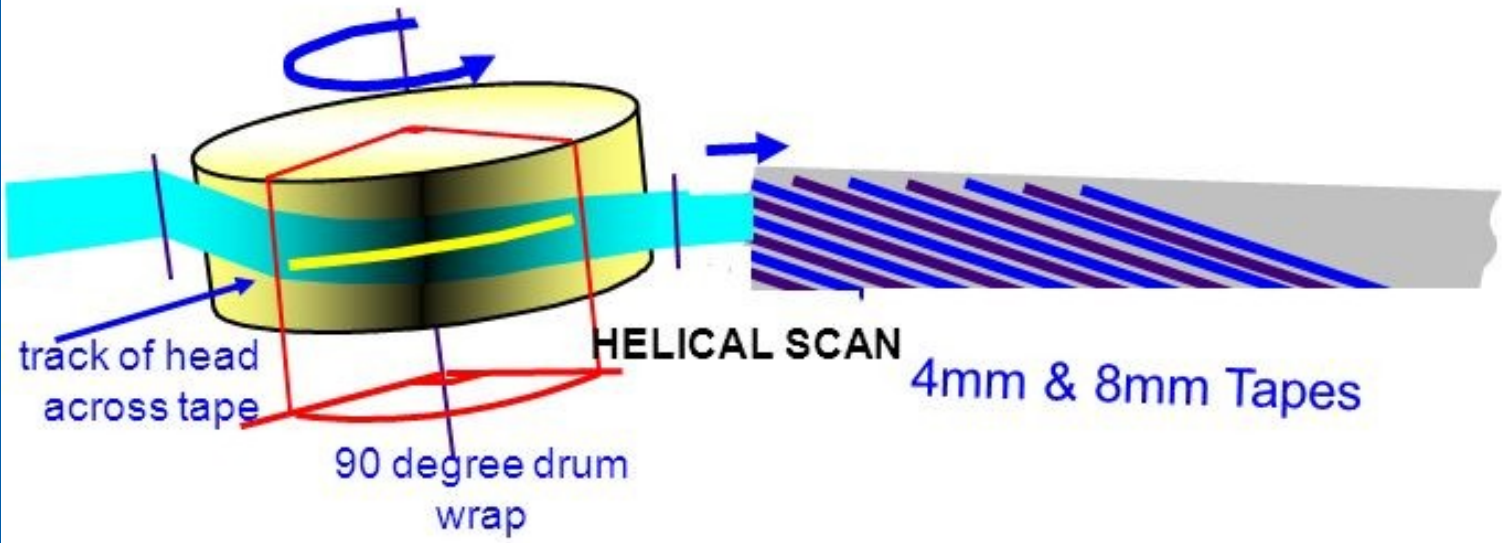
SDLT

STK 9840

SLR/QIC/Travan

IBM 3480/90, 3570, Magstar

½" reel to reel



DDS 2,3,4

Exabyte 8mm

Mammoth

AIT 1, 2

Ecrix VXA

Sony DTF

Adatmentés

- Backup - Tape / szalag
  - StorageTek T10000D
    - 8.5 TB, 252MB/sec
    - T10000E – 2018 – 12/16TB

    https://www.youtube.com/watch?v=AV7Kv6IwPaE
    https://www.oracle.com/storage/tape-storage/t10000d-tape-drive/index.html

  - IBM TS1155 (2017)
    - 15 TB, 360MB/sec
    - FibreChannel / 10G Eth

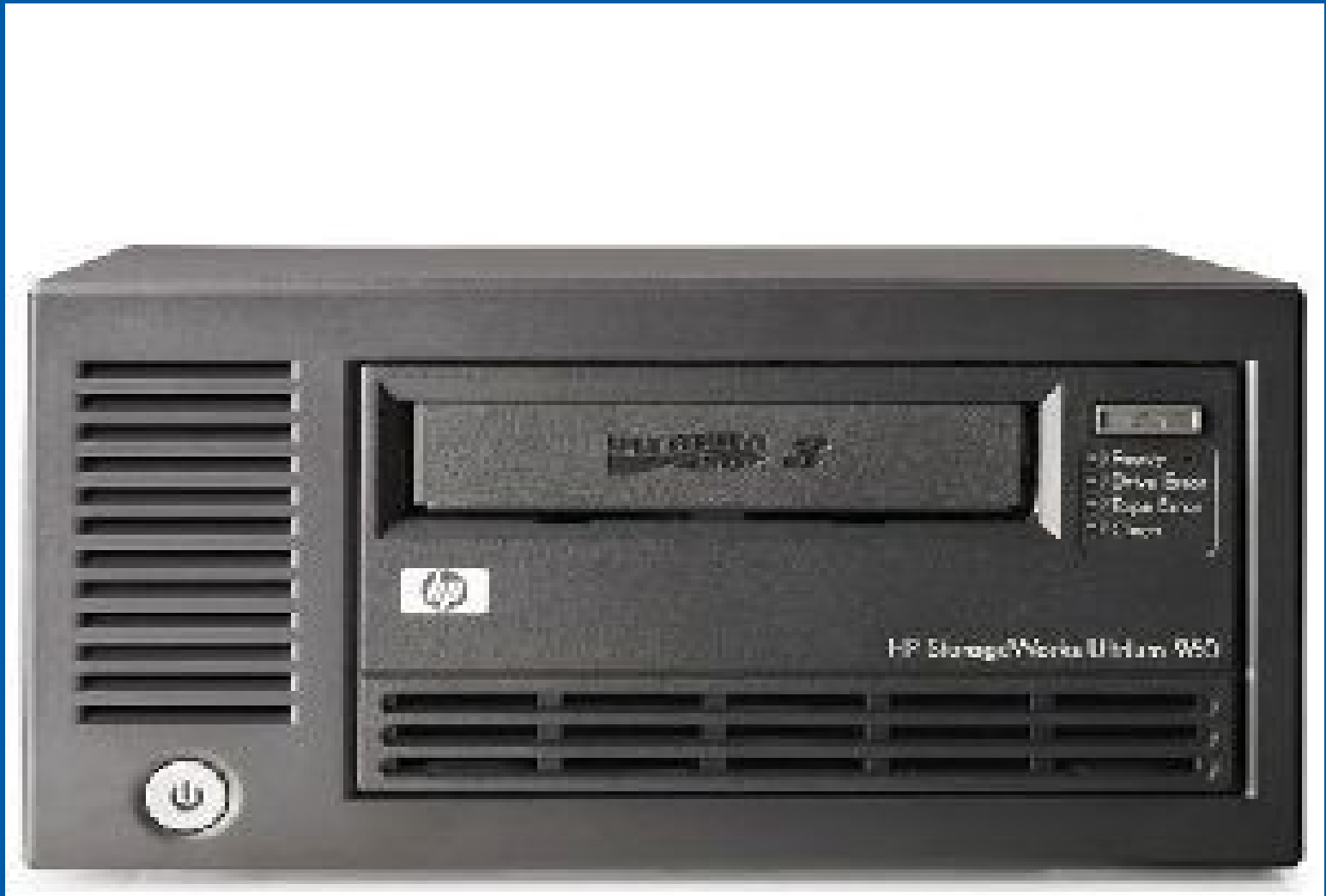
    https://www.youtube.com/watch?v=q9KIUWYJ4-k
    https://www.youtube.com/watch?v=Wm1JiI6CppU
    https://www.ibm.com/hu-en/marketplace/ts1155
    https://www.ibm.com/hu-en/marketplace/3592-tape-cartridge

Photo by Dr. Dávid Vincze


Photo by Dr. Dávid Vincze

LTO Szalag

Szalagos meghajtó (LTO)




3M
60 Mbytes
620 ft (189.0m)
DC 600A DATA CARTRIDGE TAPE
CompacTape II™
3M
200GB
IBM
D8-112 8mm DATA TAPE
3M
DDS-90
SONY
SDX1-25C
Photo by Dr. Dávid Vincze

Photo by Dr. Dávid Vincze

Szalag könyvtár

- Backup
  - Tape Library
    - Bár ez lehet on-line is, nem feltétlen backup
    - http://nava.hu/id/other-robot/

https://www.oracle.com/storage/tape-storage/sl8500-modular-library-system/

https://www.youtube.com/watch?v=xfFfQzELYHE
https://www.youtube.com/watch?v=1yUZ81dCqBg
https://www.youtube.com/watch?v=IDgXa0ioVTs
https://www.youtube.com/watch?v=FYfrC2kYbDc

https://www.ibm.com/it-infrastructure/storage/tape

Elavult mentési módszerek

- Backup
  - optikai lemezek
    - DVD-RAM (mára már ez is elavult)
  - MAID
    - *Backupot milyen gyakran olvassák?*
    - Massive Array of Idle Disks / Inactive Drives
    - nagy kapacitás
    - csak akkor megy ha szükség van rá
      - vagy különböző szinteken üzemel más-más fogyasztással és üzembe helyezési késleltetéssel
    - Virtual Tape-Library (VTL)
      - SGI 400 VTL (Copan-t felvásárolta)
        - Max. 2,688 TB , 4x 8Gb FC
        - Külső interfészben tape library

DVD-RAM


Photo by Dr. David Vincze

DVD-RAM


Panasonic
DVD-RAM
VIDEO RECORDING
120MIN
2~3x
DVD
REWRITABLE

MAID

COPAN 400M MAID Solution

1

2,688TB

Competing Solution

6

2,880TB