Operációs Rendszerek MSc

Szuperszámítógépek

2019/2020/I.

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

Operációs Rendszerek MSc

- Miről lesz szó?
- Szuperszámítógépek
  - Symmetric MulitProcessing (SMP)
  - Single System Image (SSI)
  - Massive(ly) Parallel Processing (MPP)
  - Interconnect
- Klaszterek

Operációs Rendszerek MSc

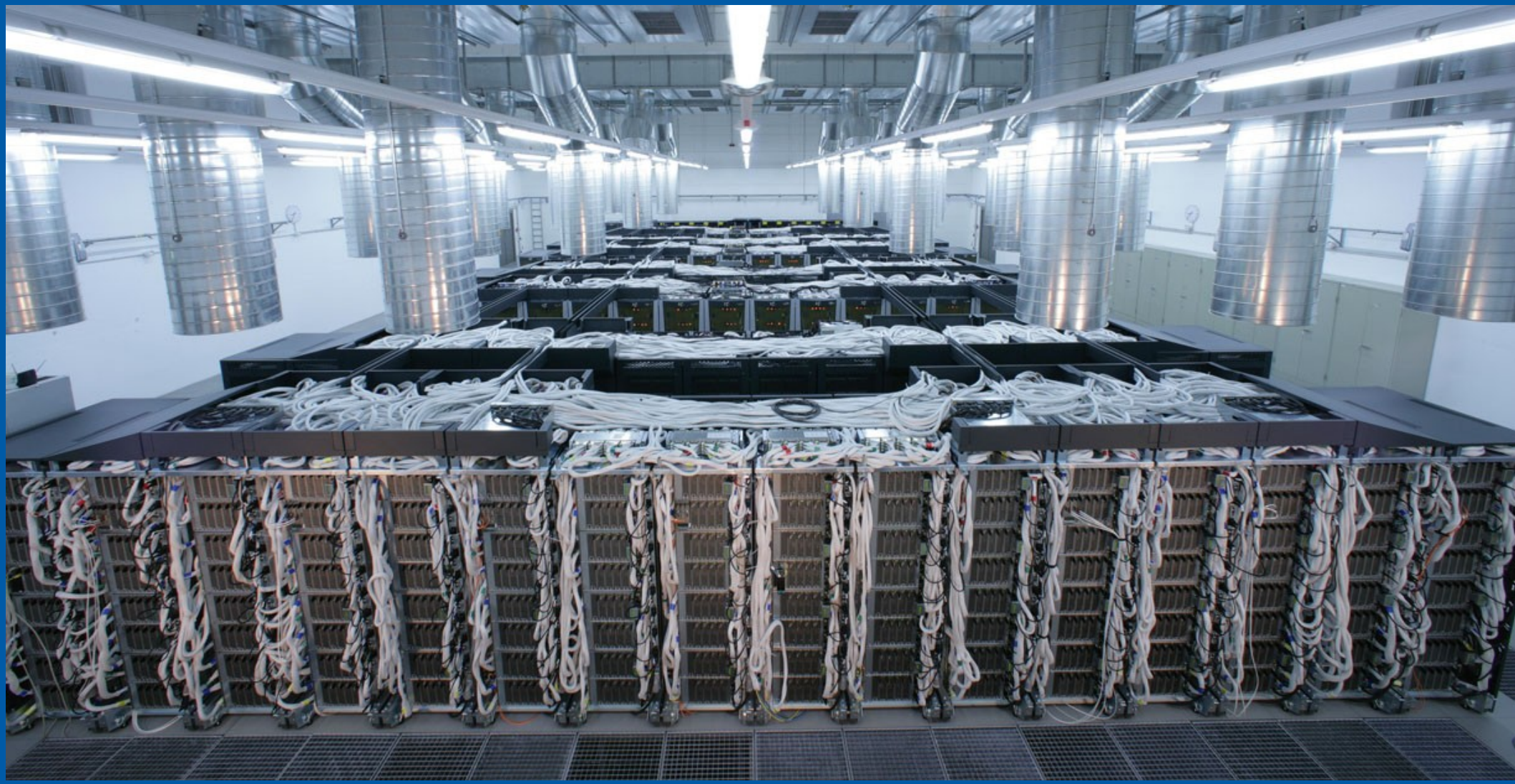
- Szuperszámítógépek
- Egy helyre koncentrált hatalmas erőforrás
  - Sok CPU, sok RAM, nagy I/O teljesítmény
  - Speciális igényekhez módosított architektúra
    - IBM zSeries
    - Sun Enterprise
    - stb.
  - Egy backplane-re csatlakoznak a modulok
    - általában passzív, és nem redundáns
      - de fizikailag tönkre tehető... :)
  - Szervízprocesszor felügyel
  - Magas rendelkezésre állás, redundancia

Operációs Rendszerek MSc

- Szuperszámítógépek
  - Több fizikai gépből felépítve
  - Speciális nagy sávszélességű interconnect
    - pl. InfiniBand, Omni-Path, Tofu, NUMALink (CrayLink)
  - Ez a *Massively Parallel Processing* (MPP)
    - Manapság többnyire ilyenek a mai szuperszámítógépek
  - GPU (Graphical Processing Unit)
    - pl. NVIDIA Tesla / Volta
  - MIC (Many Integrated Core)
    - pl. Xeon Phi
  - NUMA architektúra
    - Non-Uniform Memory Access

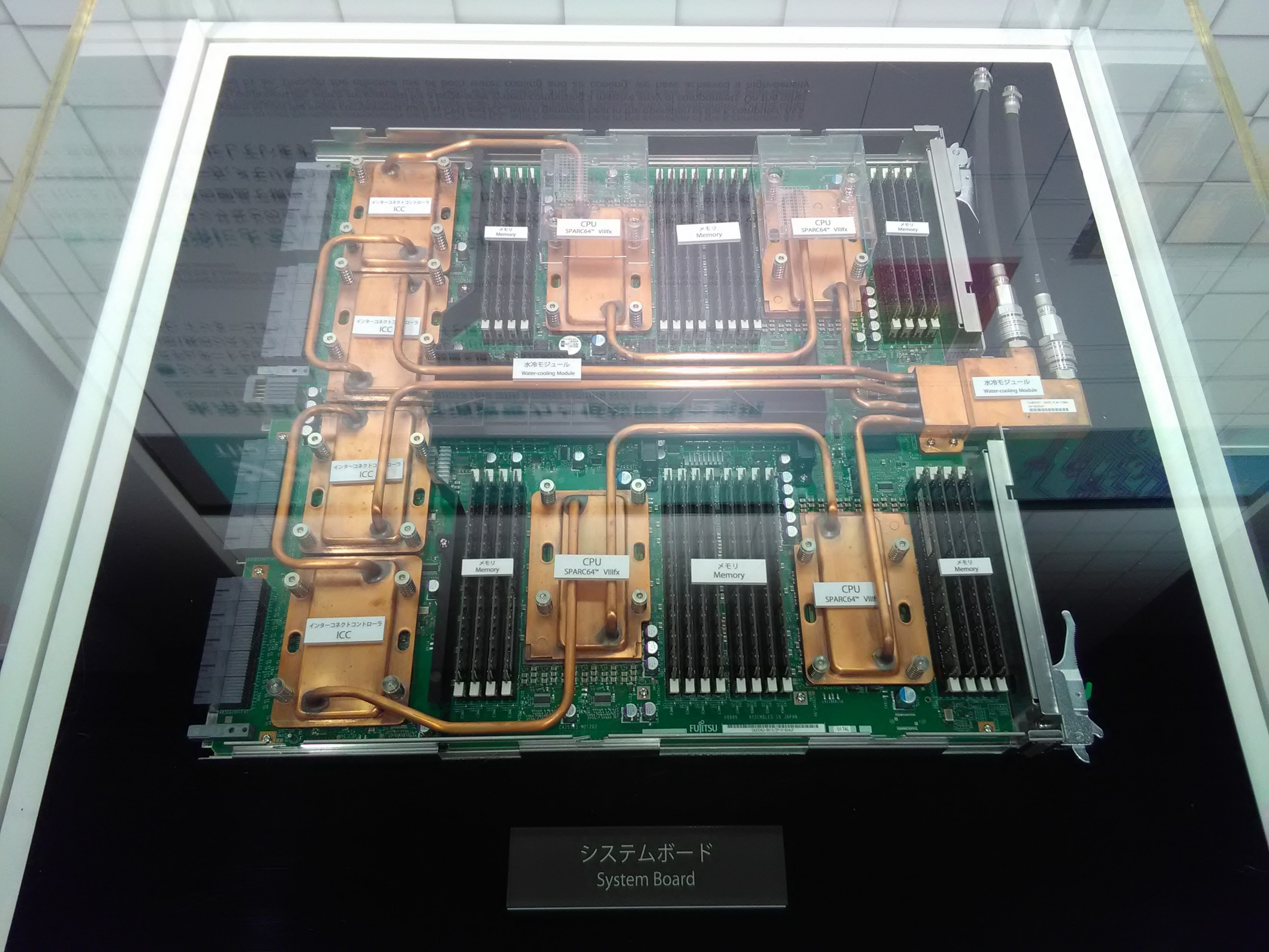

IBM
Blue Gene/P

Operációs Rendszerek MSc



K Computer

https://www.youtube.com/watch?v=UJPsIu9OaTc



システムボード
System Board


世界最高水準のスーパーコンピュータ施設として
ランニングコストと環境負荷の低減化

https://www.google.com/maps/place/K%C3%B3be,+Hj%C3%B3go+prefekt%C3%BAra,+Jap%C3%A1n/@34.6527583,135.2229209,279a,35y,292.43h,39.3t/data=!3m1!1e3!4m5!3m4!1s0x60007d812aed89d9:0xc7126106c2f670f4!8m2!3d34.690083!4d135.1955112


天河
天河二号
TH-2 High Performance Computer System
2014/04/22

每秒 5.49 亿亿次、持续计算速度每秒 3.39 亿亿次、能效比每比特 19 亿次双精度浮点运
值计算速度、持续计算速度以及综合技术水平处于国际领先地位，是我国超级计算技术发展取得的重大
进展。

The MilkyWay II supercomputer system, developed by the National University of Defense Technology (NUDT), is the outstanding achievement of the National 863 Program. Partly funded by Guangdong province and Guangzhou city, the MilkyWay II will be installed in the Guangzhou Supercomputing Center. Capable of a peak performance of 54.9PFlops, the MilkyWay II achieves a sustained performance of 33.9PFlops with a performance-per-watt of 1.9GFlops/W. The peak performance, the sustained performance and the level of comprehensive technology of the MilkyWay II all are in the lead. The MilkyWay II makes a great advance in the research and development of supercomputing technology in China.

天河二号由 170 个机柜组成，包括 125 个计算机柜、8 个服务机柜、13 个通信机柜和 24 个存储机柜，占地面积 720 平方米。内存总容量 1.4PB，存储总容量 12.4PB，最大运行功耗 17.8 兆瓦。

The MilkyWay II consists of a total sum of 170 cabinets including 125 compute cabinets, 8 service cabinets, 13 communication cabinets and 24 storage cabinets, covering 720 square meters. It has 1.4PB memory, 12.4PB storage capacity, and the power consumption of 17.8MW at its peak.

系统特点 System Features

（1）**高性能**：天河二号计算速度世界领先，排名世界超级计算机 500 强第一位。

High Performance: The performance of the MilkyWay II is world leading, which is ranking No.1 on the TOP500 list.

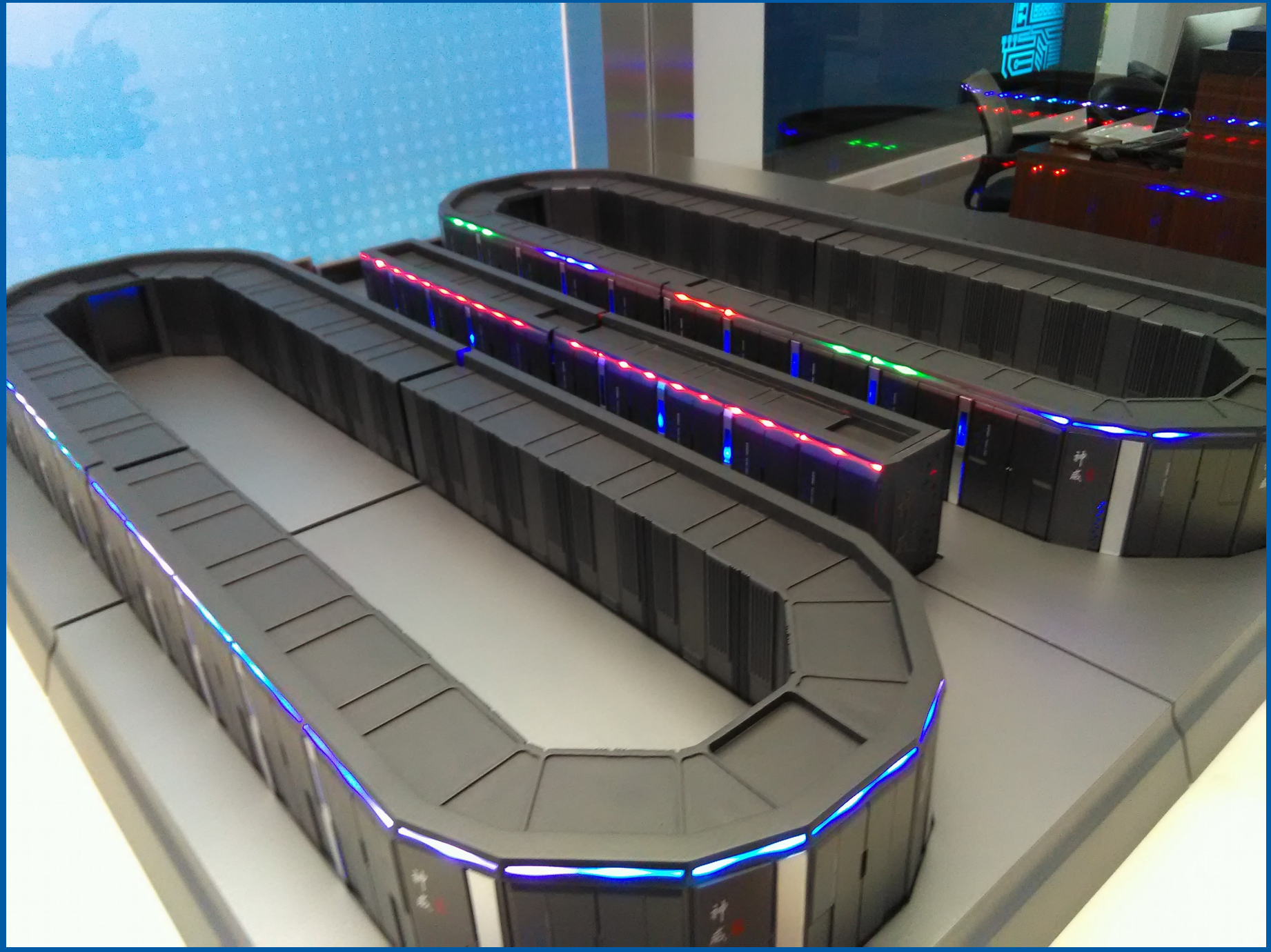
（2）**高能效**：天河二号节能降耗水平处于国际先进行列。

High Energy Efficiency: The energy efficiency of the MilkyWay II is at advanced level in the world.

（3）**应用面广**：天河二号采用新型异构多态体系结构，可有效满足科学工程计算、大数据处理、高吞吐率和高安全的信息服务等多类应用需求。

Massive Applications: A Multipurpose-Heterogeneous Architecture is adopted, which effectively meets the needs of big data processing, high throughput and high security information service as well as scientific and engineering computing.

（4）**易用性好**：天河二号采用新型微异构计算降低了应用软件编程难度



YouTube https://www.youtube.com/watch?v=KEdsrT1mFAU

YouTube https://www.youtube.com/watch?v=OU68MstXsrI

https://www.youtube.com/watch?v=RIKZyF9WIH4

HP cluster

Két szuperszámítógép a Miskolci Egyetemen (SGI UV2000, HP cluster)

26-node HP cluster (312-core, 1,2 TB RAM)

- 24 node
  - 2x 6-core Intel Xeon X5650 2.67GHz
  - 48GB ECC RAM
- 2 node
  - 2x 6-core Intel Xeon X5670 2.93GHz
  - 48GB ECC RAM
  - NVIDIA GPU
- InfiniBand

SGI UV2000

Két szuperszámítógép a Miskolci Egyetemen (SGI UV2000, HP cluster)

SGI UV2000 – SSI

- 3x8x2x8 mag (Intel Xeon E5-4627 v2 @ 3.33 GHz)
- 1.4TB RAM
- 240TB HDD
- NUMALink 6 (4x 6.5Gbyte/s)


SGI UV 2000
sgi
Altix UV 1000
sgi


sgi

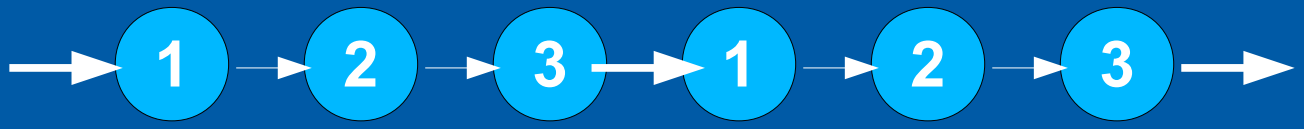
Alkalmazás

- Mire lehet használni?
  - Folyamat modellje

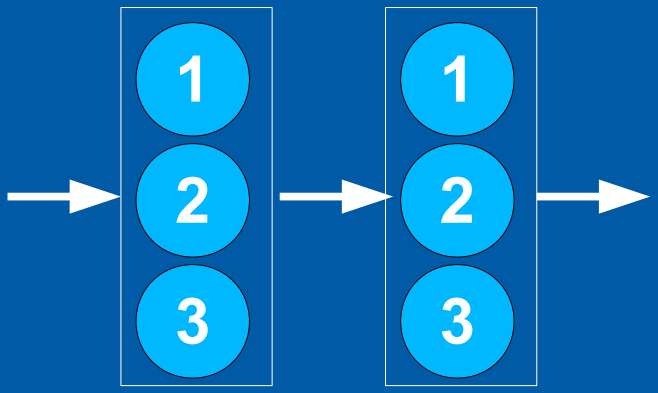
$$q^{k+1}_{i_1 i_2 \ldots i_N u} = q^{k}_{i_1 i_2 \ldots i_N u} + \prod_{n=1}^{N} \mu_{i_n,n}(s_n) \cdot \mu_u(a) \cdot \Delta \tilde{Q}^{k+1}_{i,u} =$$

$$= q^{k}_{i_1 i_2 \ldots i_N u} + \prod_{n=1}^{N} \mu_{i_n,n}(s_n) \cdot \mu_u(a) \cdot \alpha^{k}_{i,u} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v \in U} \tilde{Q}^{k+1}_{j,v} - \tilde{Q}^{k}_{i,u} \right)$$

  - Szekvenciális probléma



  - Párhuzamosítható probléma



Alkalmazás

- Deep Blue vs. Garry Kasparov

  - IBM Deep Blue
  - 1996/1997
  - 11.38 GFLOPS
    - TOP500: 259.
  - Sakkra specializált segédprocesszorok
  - 1996: 4-2 GK – DB
  - 1997: 3½–2½ DB – GK

Alkalmazás

- Filmgyártás / animáció
  - Képkockák végleges renderelése
  - Képkocka szinten párhuzamosítható
  - Maguk a képkockák részei is egymástól függetlenül előállíthatóak
  - Monsters University (2013):
    - 24 000 processzormagos sz.sz.

Alkalmazás

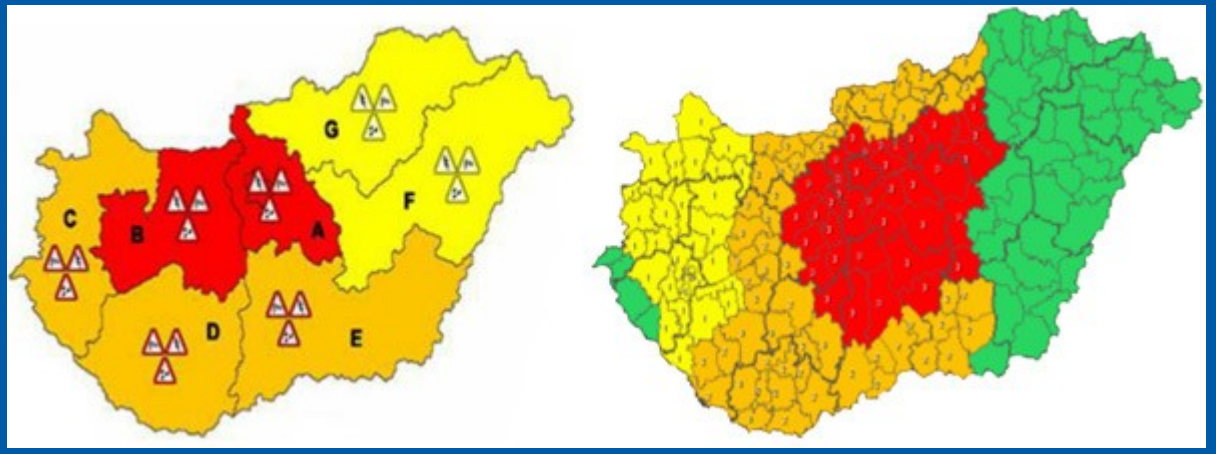
- SETI at home

- Search for Extra-Terrestrial Intelligence
  „földön kívüli intelligencia keresése”

- Elosztott rendszer
  - „virtuális sz.sz.”
- Bárki csatlakozhat
- Rádióteleszkópok által vett jelek feldolgozása, mintázatok keresése

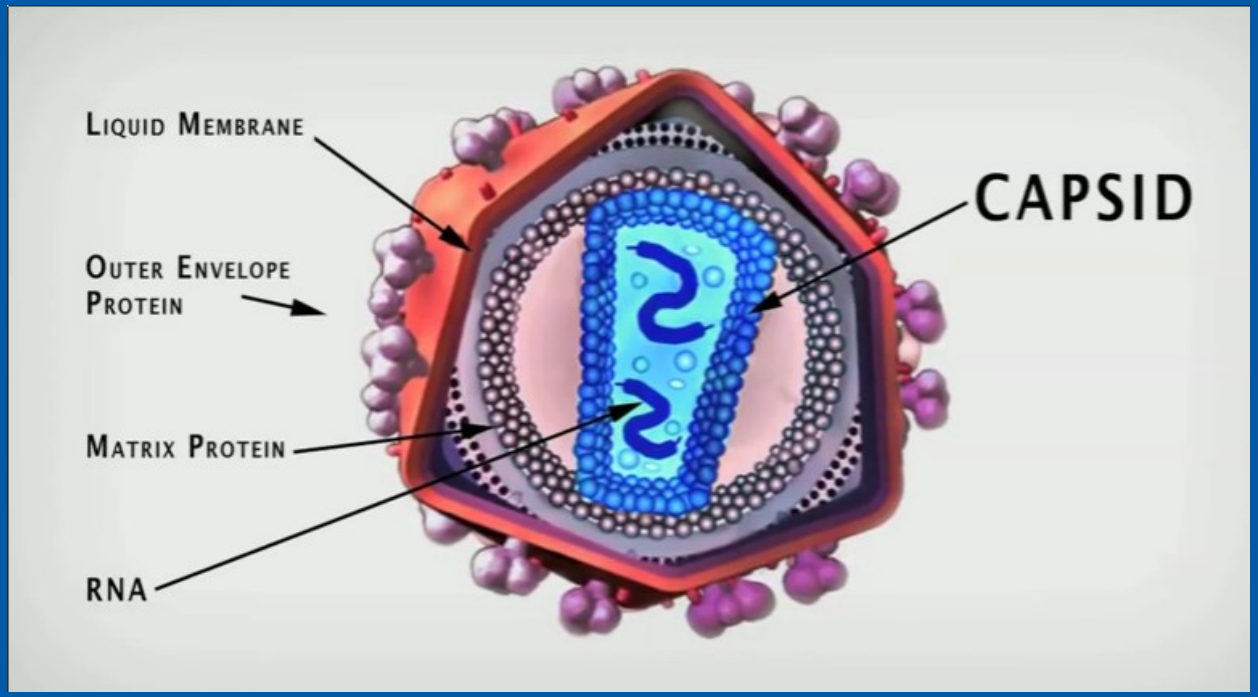
Alkalmazás

- Időjárás előrejelzés
  Adatok: műholdak, repülés, földi mérőállomások + Modell
- OMSZ
- 2010-ben új szuperszámítógép:
  - 7 régió helyett 174 kistérség
- Riasztások kiadása
- 2018.01.10. - új gép – HP Apollo 6000



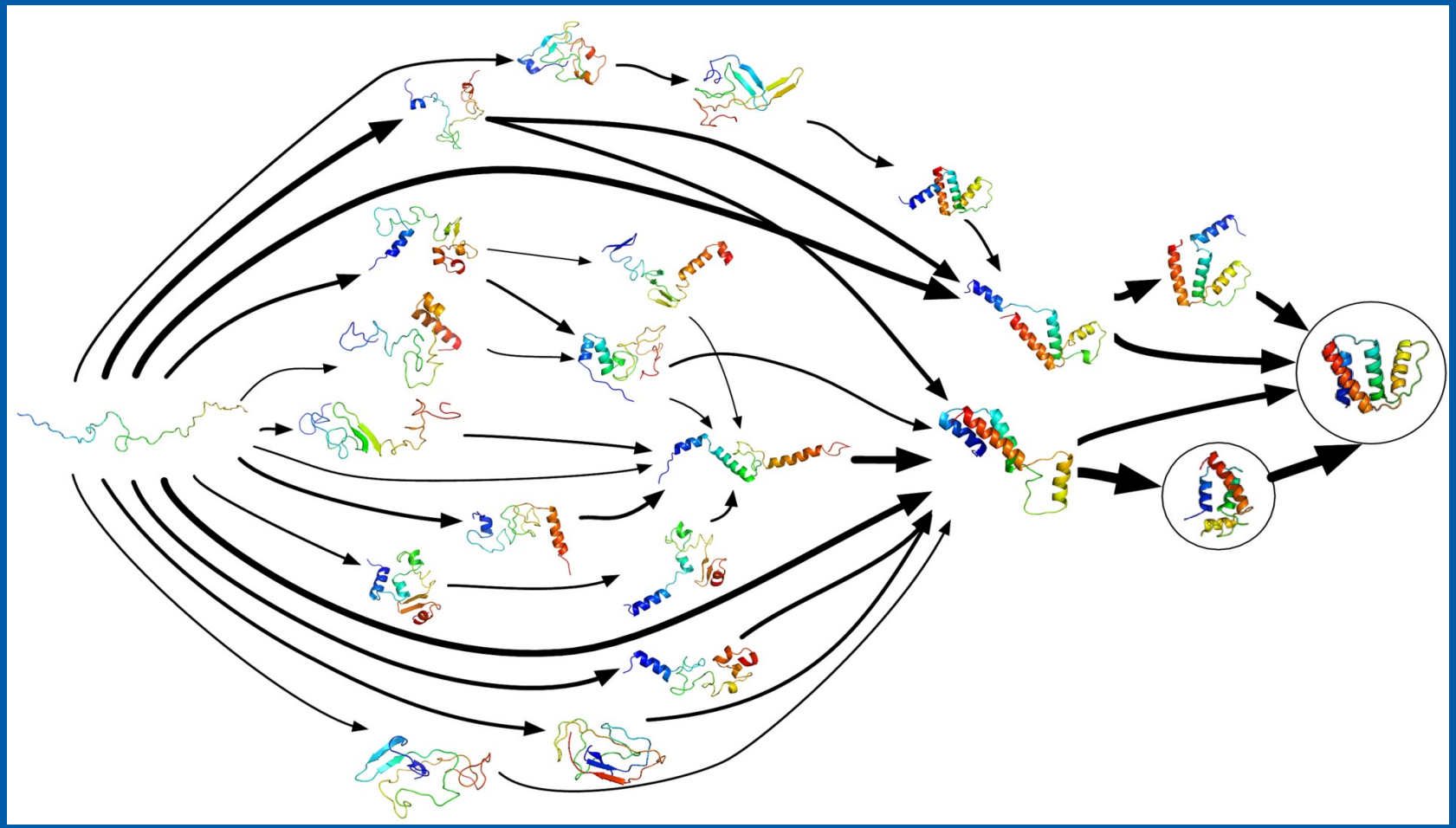
Alkalmazás

- Betegségek kutatása
  - Blue Waters sz.sz.
  - 13,3 PFLOPS; 1,5 PB mem
  - HIV-1 vírus fehérjeburkának feltérképezése
  - 1300 különböző alakú protein
  - 64 millió atom interakciójának szimulációja



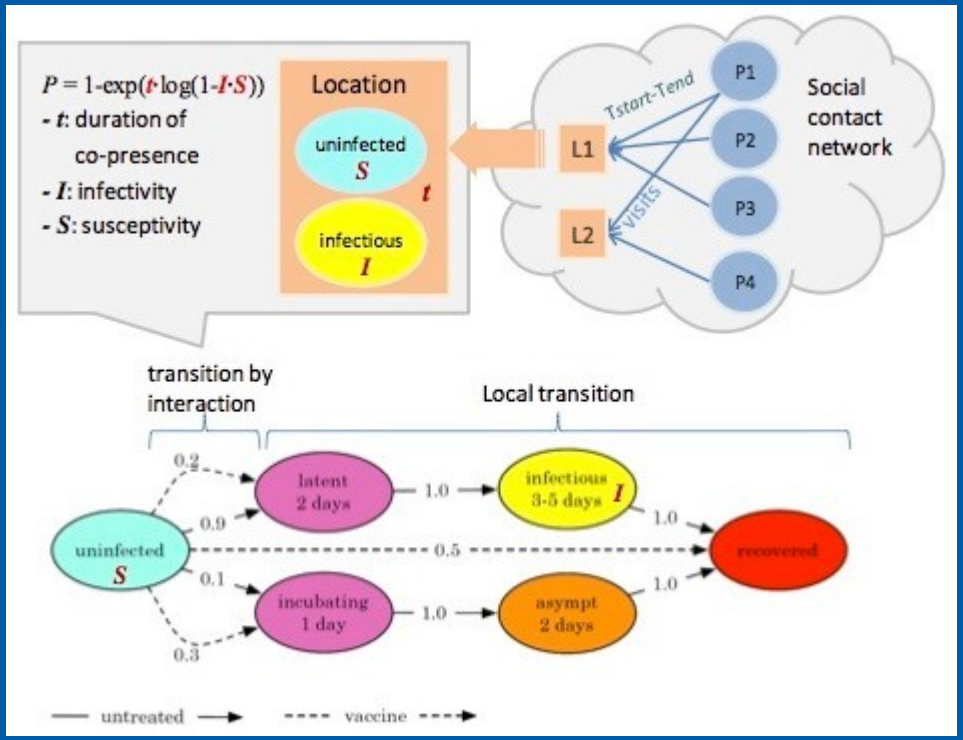
Alkalmazás

- Protein folding
  - Folding at Home
  - kb. 40 PFLOPS, elosztott rendszer

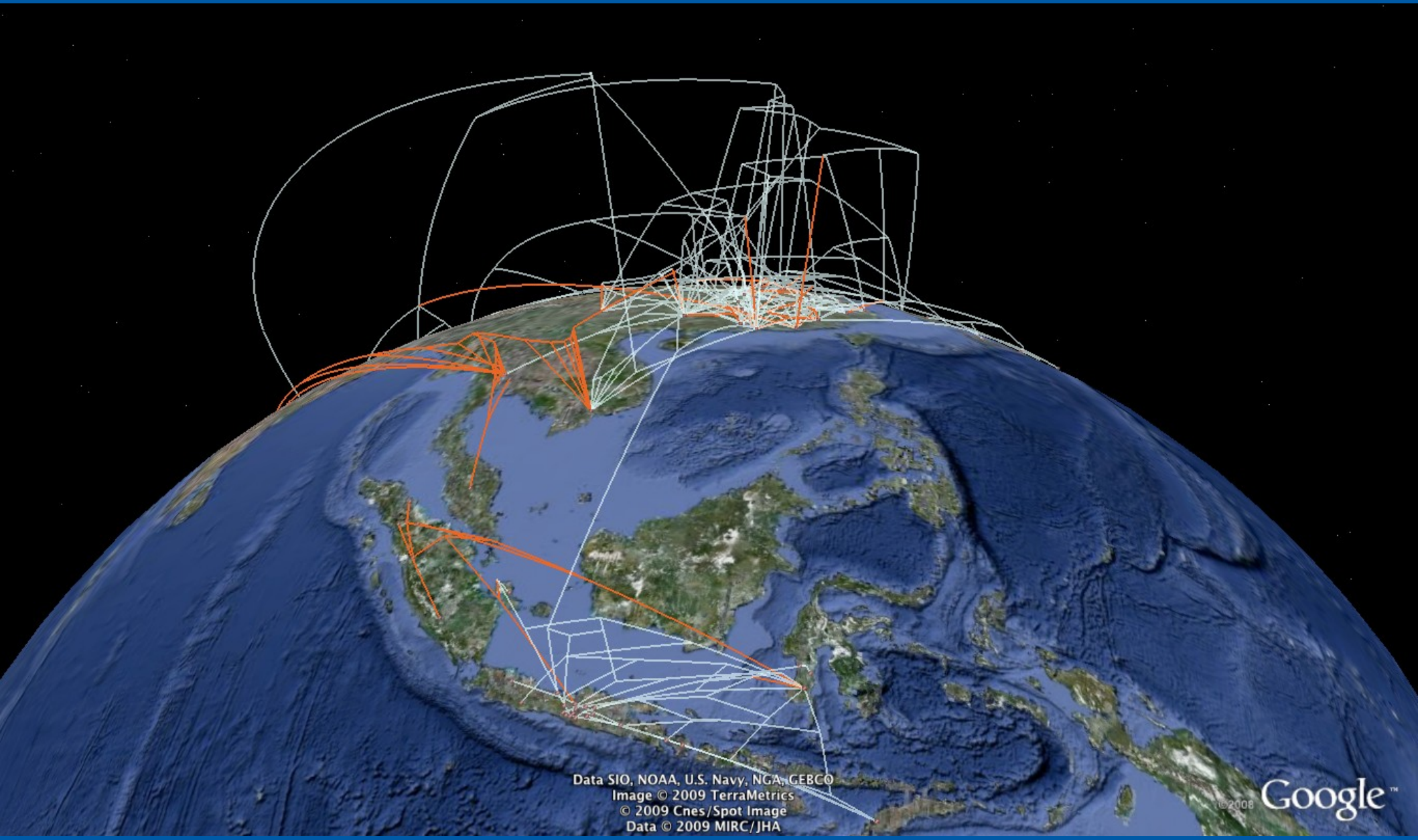


Alkalmazás

- Fertőzések terjedésének modellezése
  - pl. H1N1, H5N1
  - Begyűjtött adatok alapján előrejelzés készítése
  - „EpiSimdemics” programcsomag
    - 100 milliós nagyságrendű szociális kapcsolatrendszer kezelése

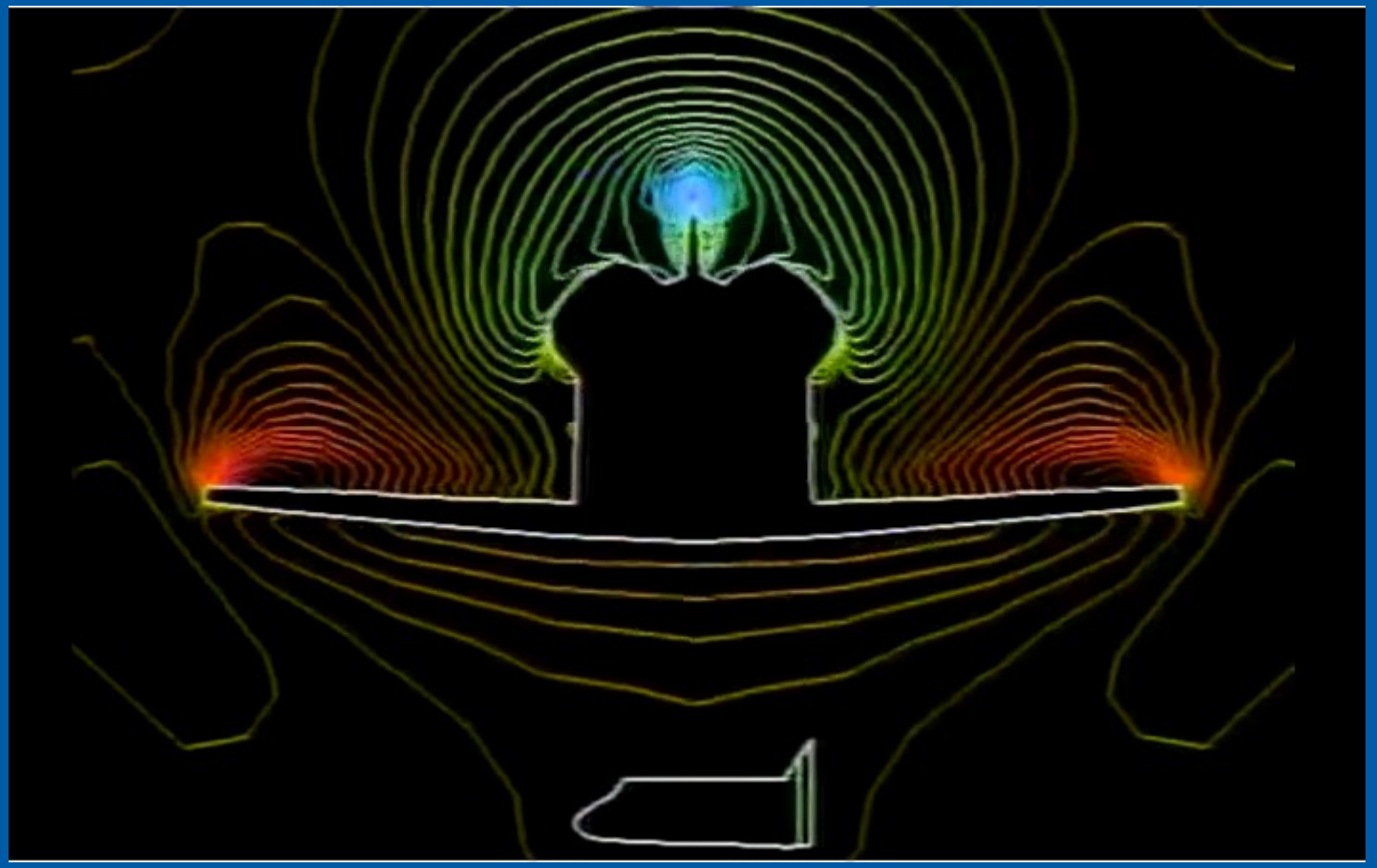


Alkalmazás


Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2009 TerraMetrics
© 2009 Cnes/Spot Image
Data © 2009 MIRC/JHA
Google

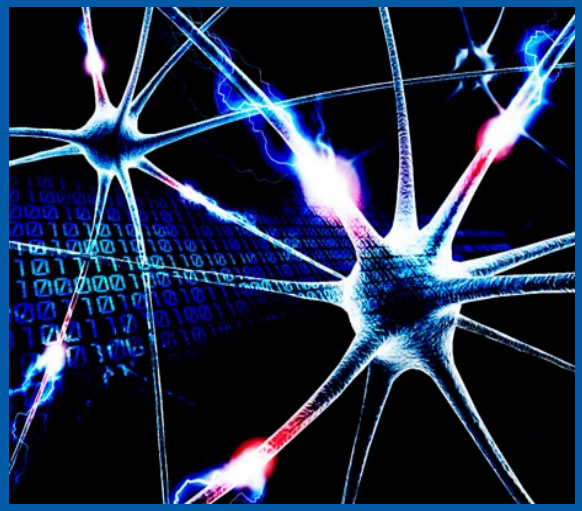
Alkalmazás

- NASA, 1986, CRAY-2, 1.9 GFLOPS
  - Áramlási szimulációk
  - Űrsikló



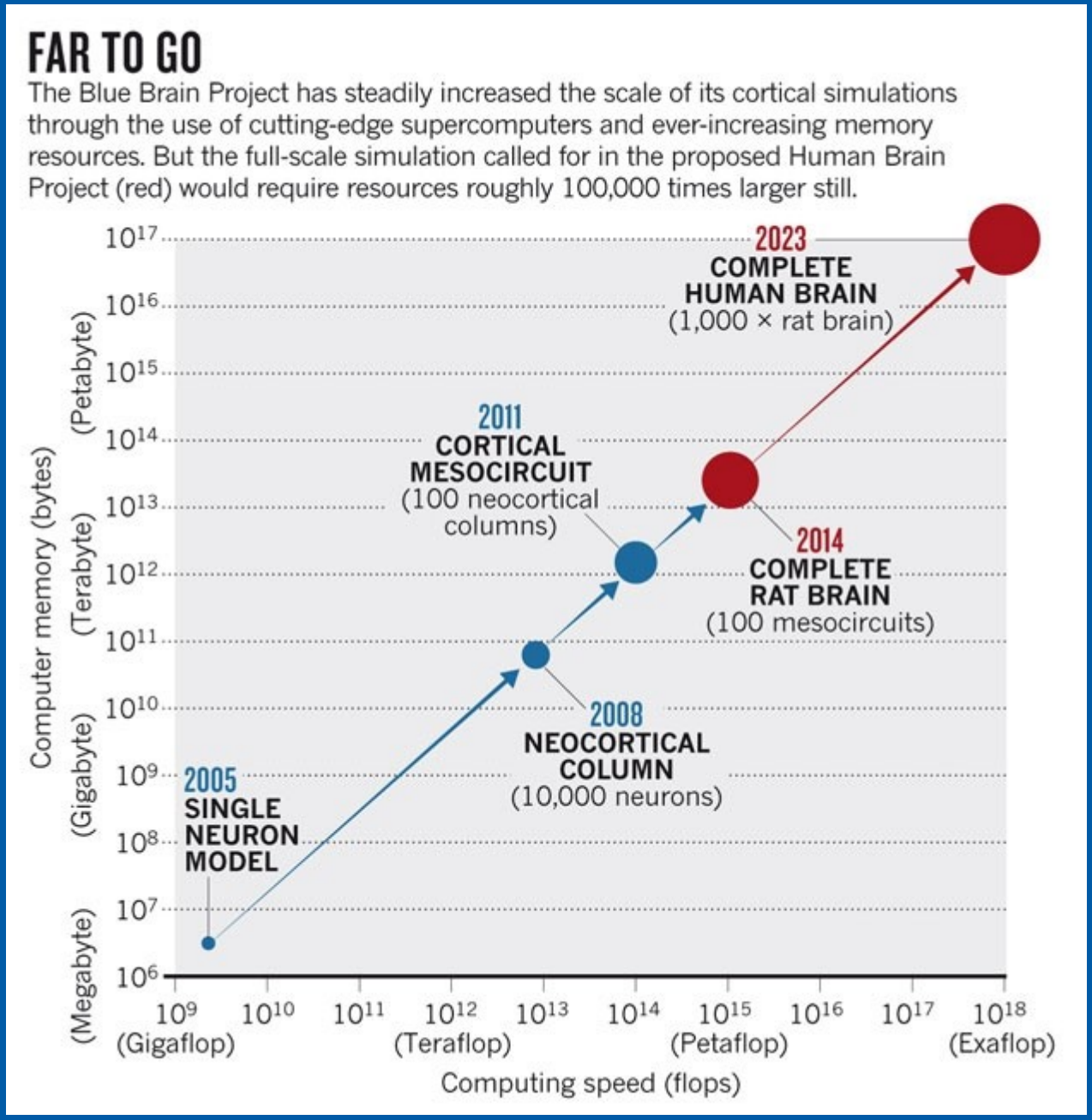
Alkalmazás

- Agy modellezés
- „K Computer”, Japán, Kobe
- 10.5 PFLOPS (TOP500: 4.)



- Szimuláció:
  - 1.73 milliárd virtuális idegsejt
  - 10.4 billió virtuális szinapszis
  - Kb. 1% egy átlagos emberi agynak
  - 1 másodpercnyi aktivitást 40 percnyi számítást vesz igénybe

Alkalmazás


FAR TO GO
The Blue Brain Project has steadily increased the scale of its cortical simulations through the use of cutting-edge supercomputers and ever-increasing memory resources. But the full-scale simulation called for in the proposed Human Brain Project (red) would require resources roughly 100,000 times larger still.
Computer memory (bytes)
(Megabyte)
(Gigabyte)
(Terabyte)
(Petabyte)
10^{6}
10^{7}
10^{8}
10^{9}
10^{10}
10^{11}
10^{12}
10^{13}
10^{14}
10^{15}
10^{16}
10^{17}
2005
SINGLE NEURON MODEL
2008
NEOCORTICAL COLUMN
(10,000 neurons)
2011
CORTICAL MESOCIRCUIT
(100 neocortical columns)
2014
COMPLETE RAT BRAIN
(100 mesocircuits)
2023
COMPLETE HUMAN BRAIN
(1,000 × rat brain)
10^{9} (Gigaflop)
10^{10}
10^{11}
10^{12} (Teraflop)
10^{13}
10^{14}
10^{15} (Petaflop)
10^{16}
10^{17}
10^{18} (Exaflop)
Computing speed (flops)

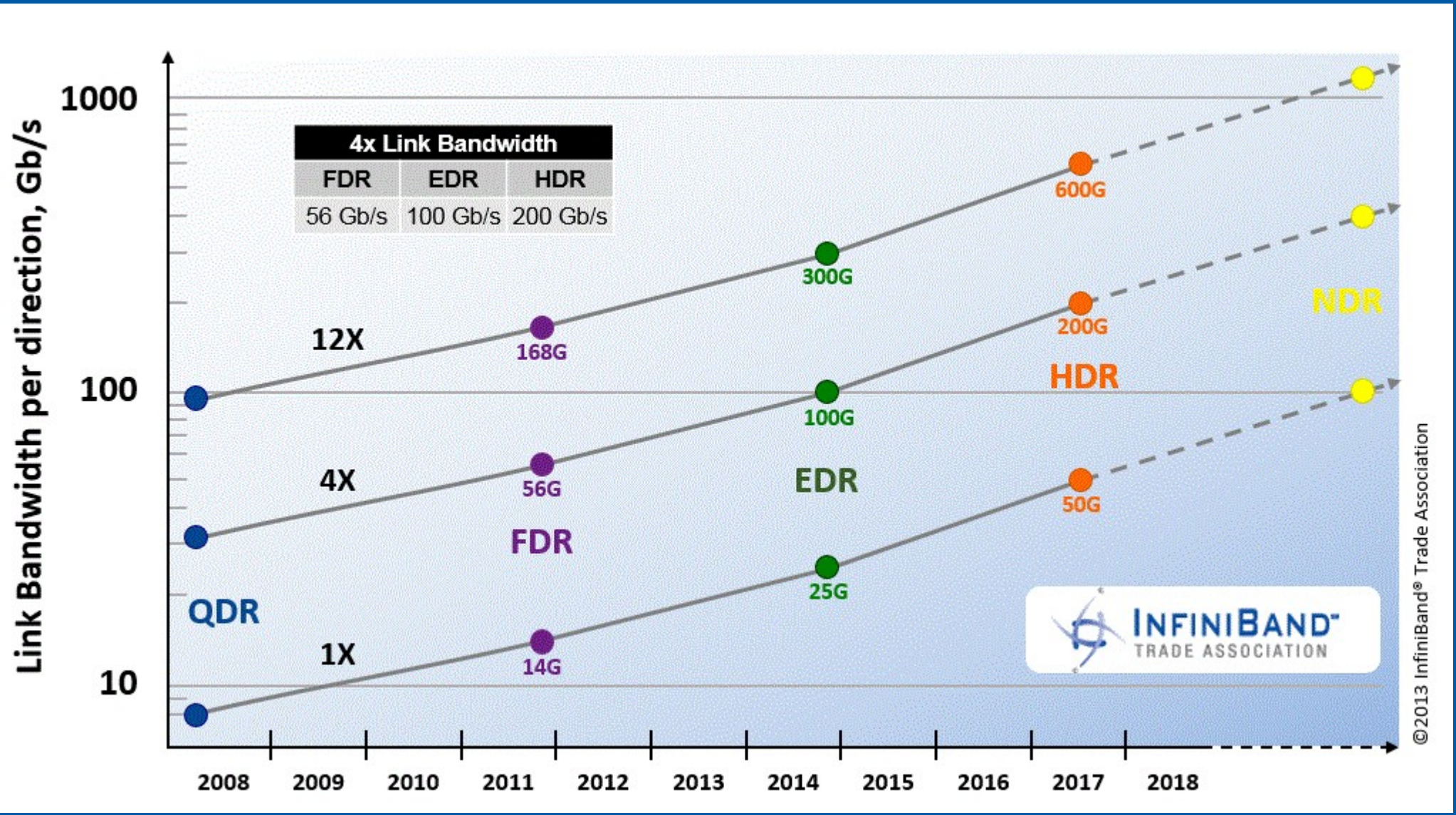
Alkalmazás

- Kőolaj / földgáz lelőhelyek kutatása
- Mérési adatok feldolgozása
  - mesterséges rezgéskeltés (hanghullám)
  - visszaverődések rögzítése (számos ponton)
- Szeizmikus modell
- Kőzetminták tanulmányozása
  - nagy felbontású scanner
  - áramlás

InterConnect - InfiniBand

- Interconnect - InfiniBand
  - sok vállalat kooperációja révén
  - nagyon elterjedt
  - alacsony késleltetés
  - alapvetően soros átvitel (de lehet több *lane*)
  - nagy sávszélesség
    - Eredetileg 2.5Gbit/s az alap sebesség (1db lane)
      - 4 pin (8b/10b kódolás)
    - InfiniBand 4X – 10Gbit/sec, 16 pin
    - InfiniBand 12X – 30Gbit/sec, 48 pin
    - Alapsebesség növekszik
    - Elterjedt manapság az 56Gbit/sec és a 100Gbit/s (4x 14Gbit / 25Gbit – 64b/66b kódolás)
    - De piacon van már a 200 (HDR) Gbit/sec
  - 64k db eszköz címezhető

InterConnect - InfiniBand



InterConnect - InfiniBand

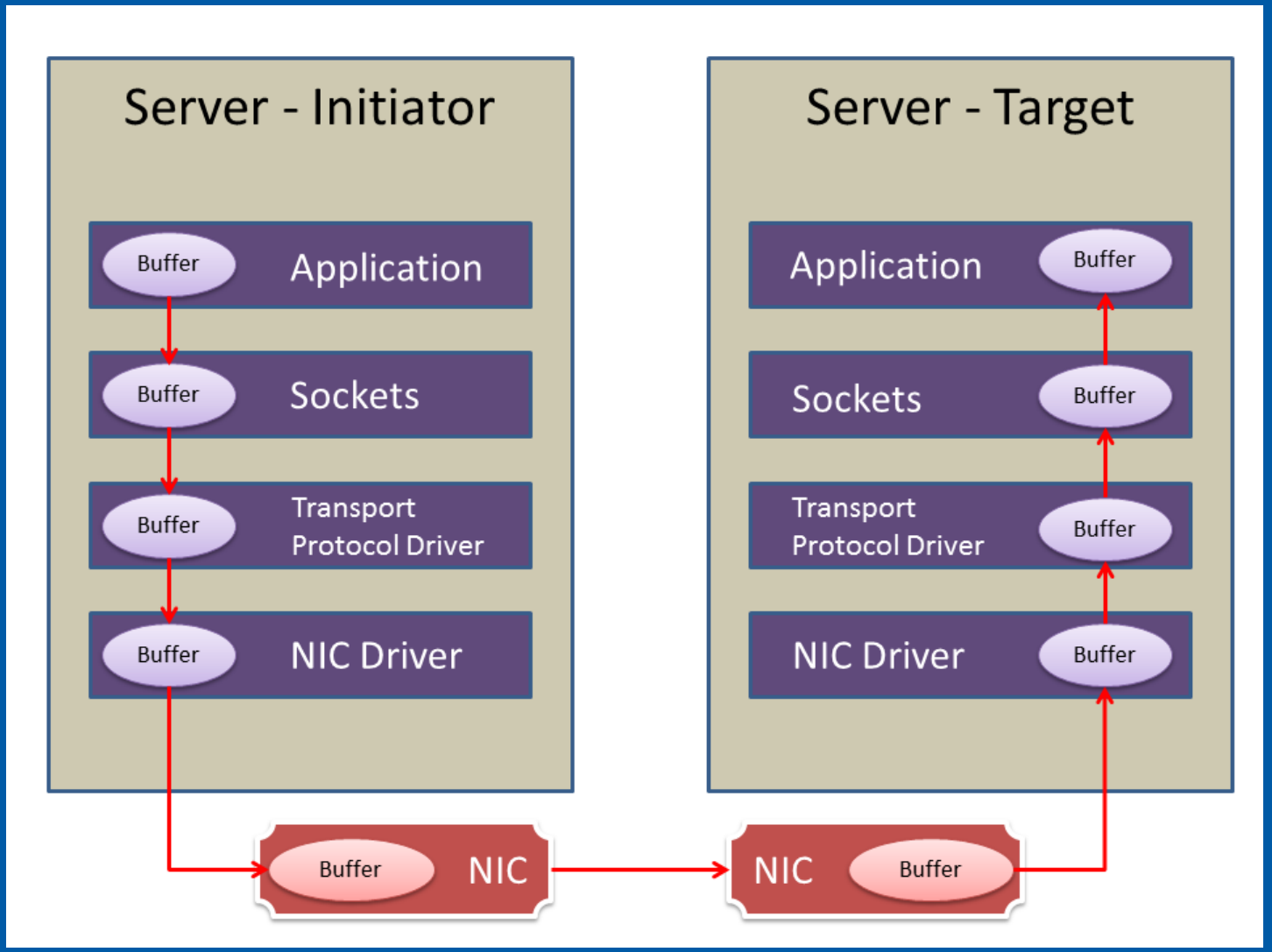
- Interconnect - Infiniband
  - switched-fabric topológia
  - HCA (host channel adapter)
  - TCA (target channel adapter) – I/O eszközökhöz
  - Switch
  - Multipath
  - Multicast
  - Médiák:
    - rézkábel (~17m, gyártófüggő)
    - üvegszál (km-ek)
    - PCB (NYÁK)

InterConnect - InfiniBand

- InfiniBand
  - RDMA (Remote DMA)
    - Távoli gép memóriájának közvetlen elérése egy processzből! (Driver + OS + VirtMem)
  - QoS
  - Maximum payload: 4k
  - Rétegelt protokoll
    - Érdekesség, hogy IPv6-ot használ a network layer-ben
  - Route-olható

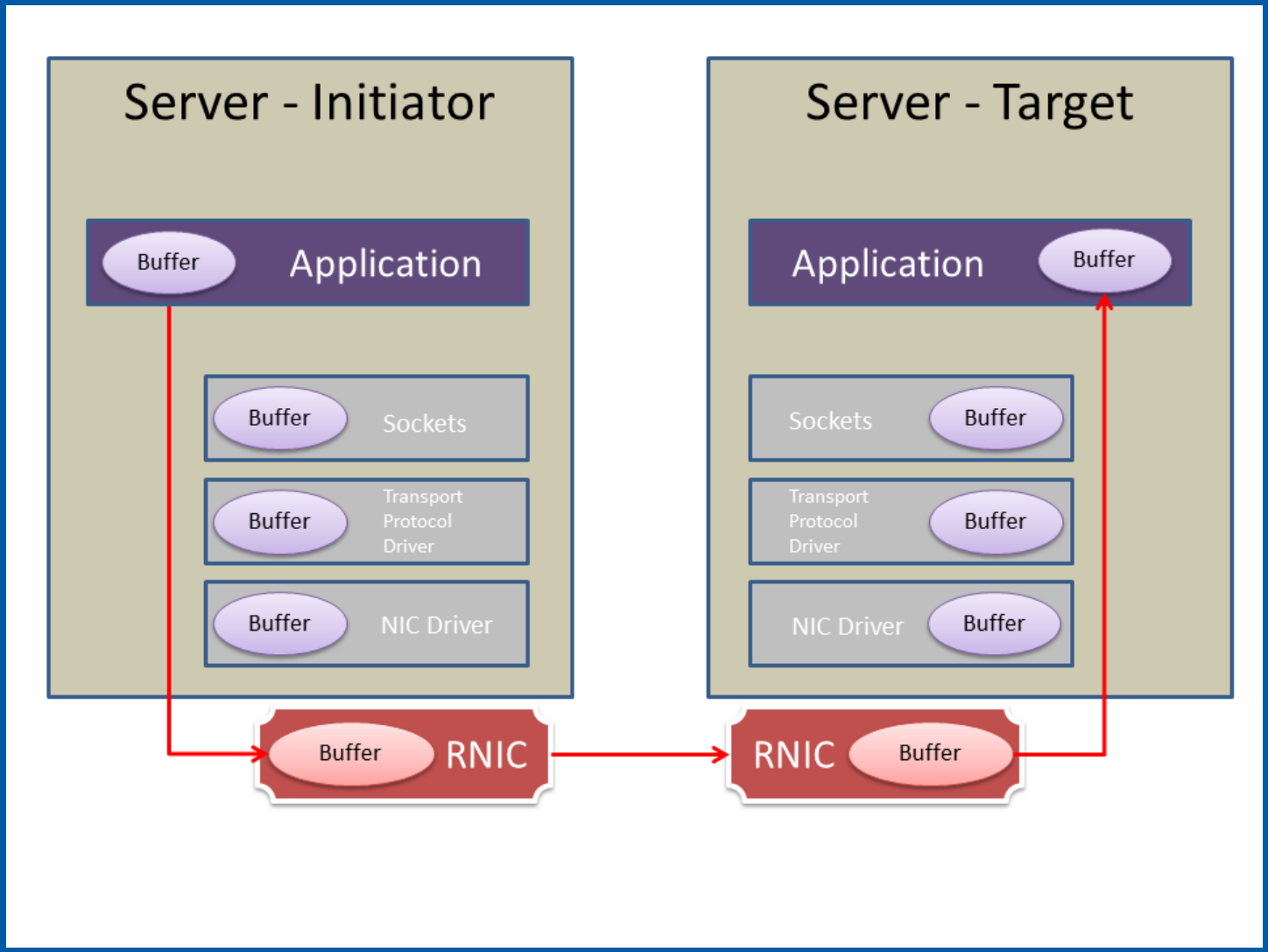
InfiniBand - RDMA

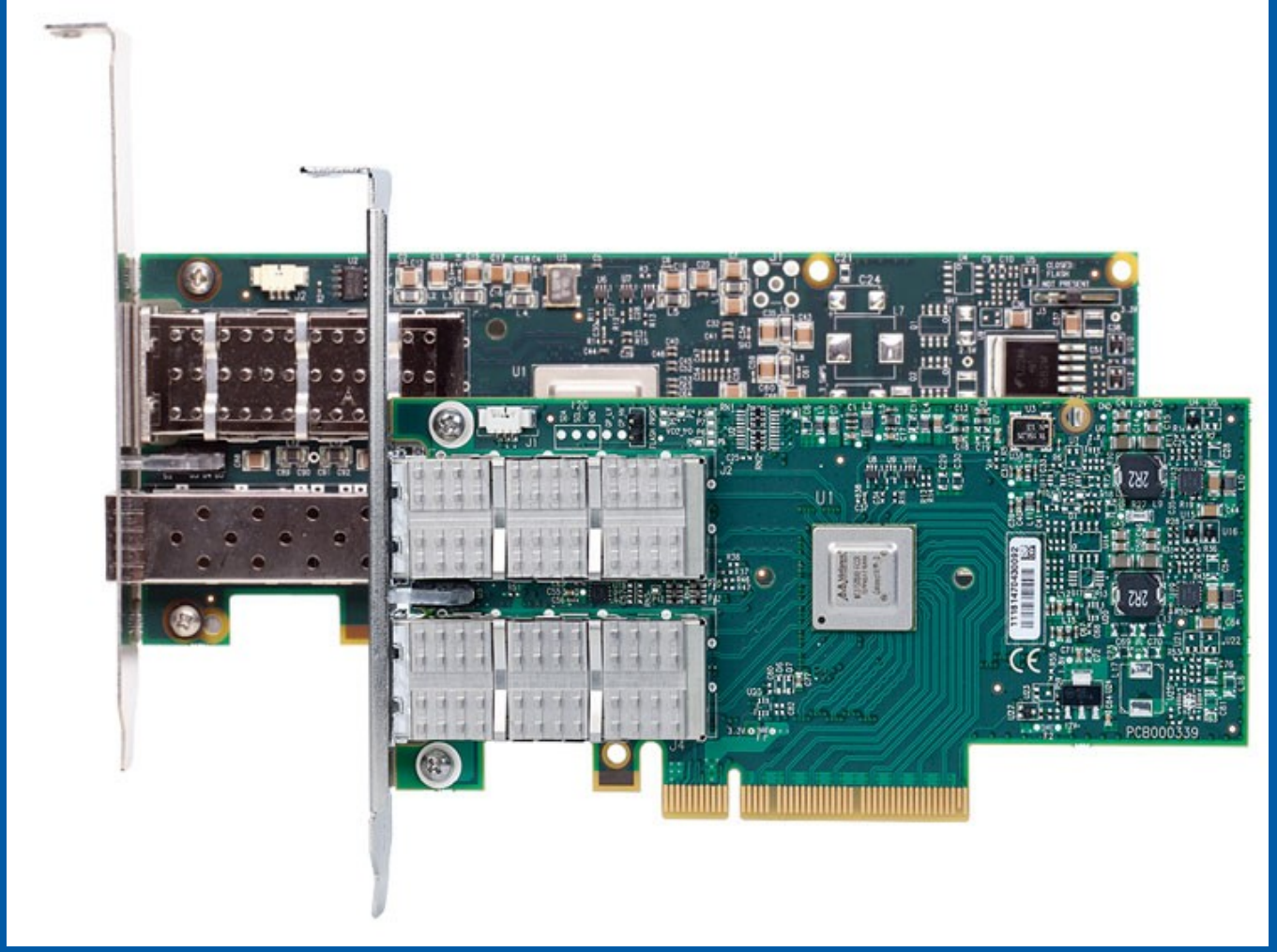
- InfiniBand RDMA
  - „Hagyományos” hálózati átvitel

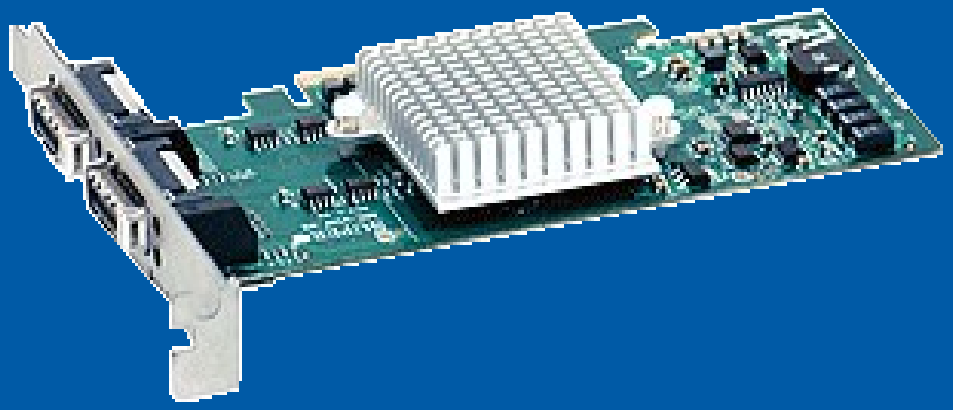


InfiniBand - RDMA

- InfiniBand RDMA
  - RDMA hálózati átvitel




U1
U1
PCB000339



InfiniBand - Switch

InterConnect - Omni-Path

- Interconnect – Egyebek
  - Omni-Path (Intel)
    - 100Gbit/s, 64b/66b kódolás
    - 24-bit címzés
    - Retransmit-kor nem muszáj az egész csomagot újraküldeni
    - Intel: „HPC optimalizált”, „hatékonyabb”,
      - Kevesebb hely, kevesebb energia, olcsóbb, stb.
    - Mellanox: „not competitive”, nincs off-load, stb.
    - 48-port (OP) vs. 36-port (IB) switch
      - Kevesebb switch-re van szükség (mondja az Intel...)

OmniPath vs. InfiniBand

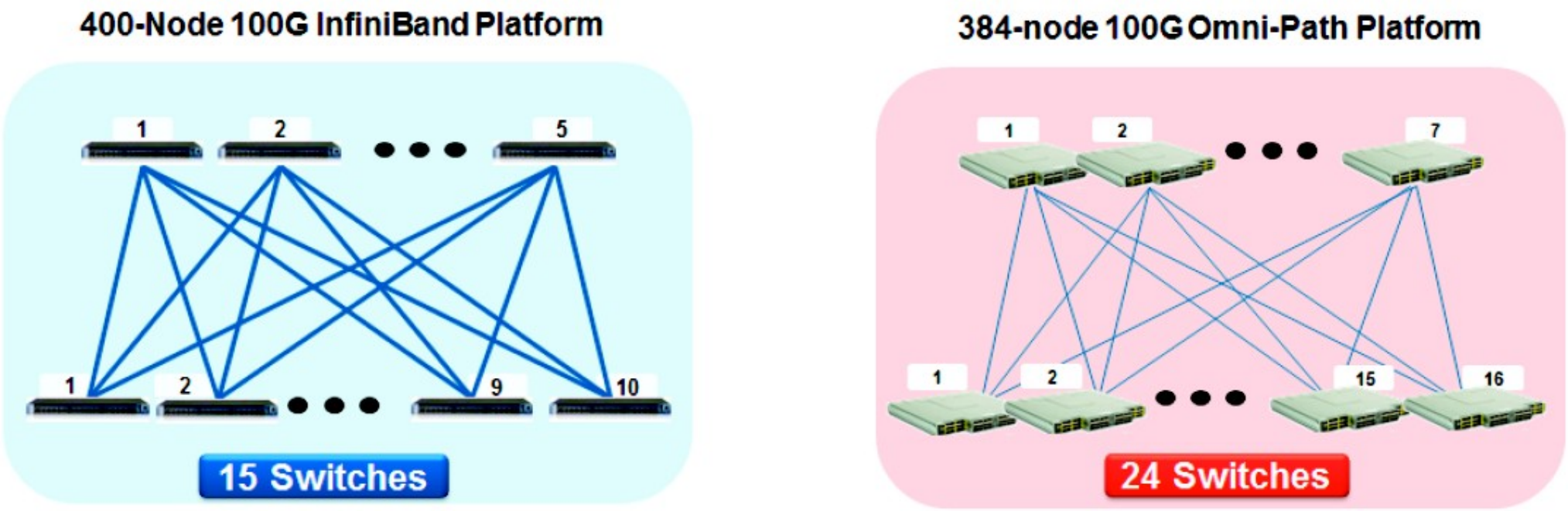
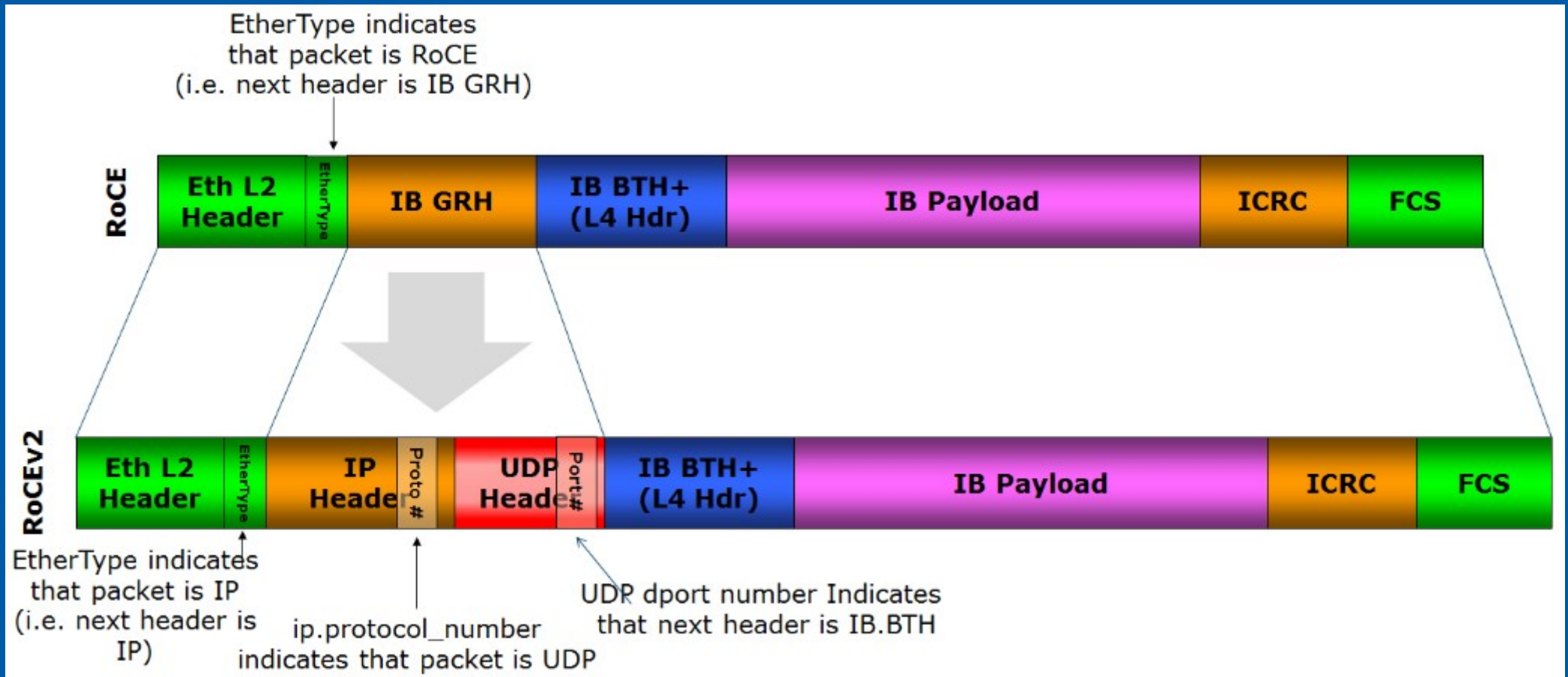


Figure 2. *HDR100 Requires 1.6X Fewer Switches for 400-Nodes*

http://www.mellanox.com/related-docs/whitepapers/WP_Introducing_200G_HDR_InfiniBand_Solutions.pdf

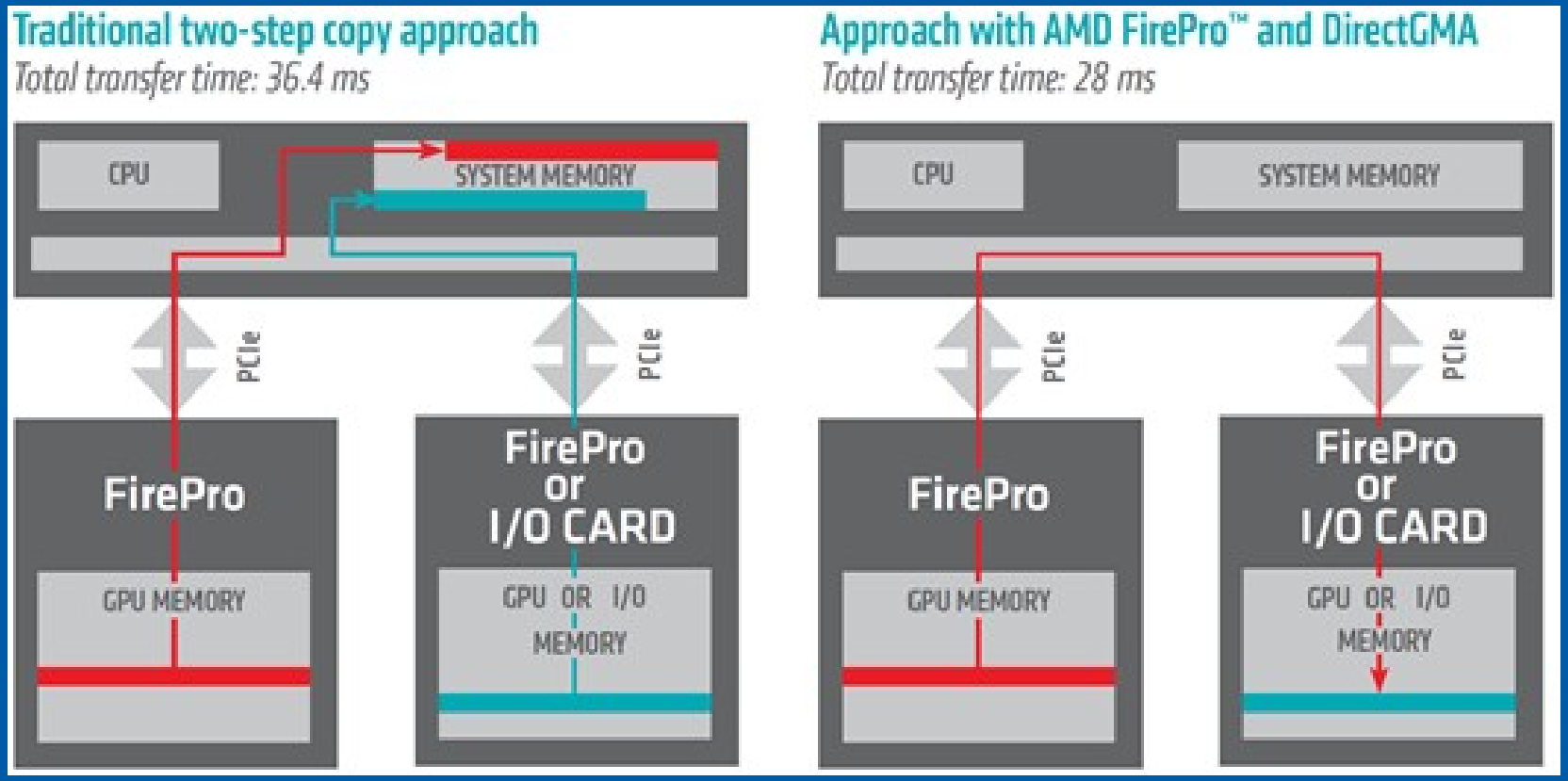
InterConnect - RoCE

- Interconnect – Egyebek
  - RDMA over Converged Ethernet (RoCE)
    - Ethernet felett (pl. 10/25/50/100Gbit/s)
    - RoCEv2
      - IP+UDP
    - IB transport protocol

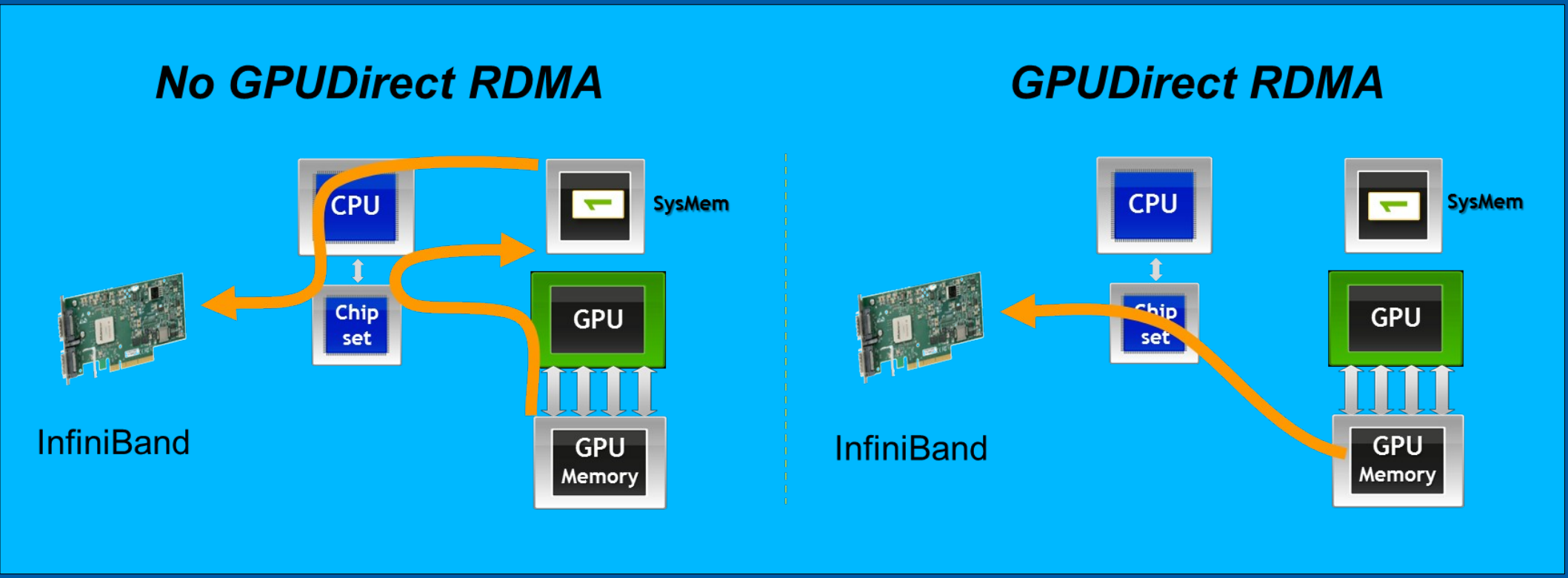


InterConnect - GPU

- Interconnect – Egyebek
  - GPUDirect / DirectGMA
    - Nem a központi memóriát lehet RDMA-zni
    - Hanem perifériák memóriáját. pl. GPU
    - Persze maga a komm. mehet IB-n pl.



InterConnect - GPU



Operációs Rendszerek MSc

- Klaszterek
  - Több (akár heterogén) számítógép „egyesítése” adott feladatra
  - HA-LB
    - High Availability – Load Balancing
  - HA
    - Valamelyik erőforrás kiesik, maradjon működőképes
    - Kiszolgálók, eszközök többszörözése
    - Adatok folyamatos szinkronizációja
    - Master-Slave, Master-Master
    - Hogyan detektáljuk, hogy valami kiesett?
      - biztos, hogy nem „én” estem ki?
      - mi a helyzet, ha menet közben visszajön?
      - kinek friss az adata?
    - Fencing

Operációs Rendszerek MSc

- Klaszterek
  - HA
    - jellemző, hogy több hálózati címük van
      - Minden gépnek egy saját
      - Egy, amin a szolgáltatás fut
      - Az veszi fel a „szolgáltatás” címet, amelyik aktív
  - LB
    - A terhelést valamilyen szempont szerint szétosztani
    - Kiszolgálók, eszközök többszörözése
    - Adatok folyamatos szinkronizációja
    - Master-Slave, Master-Master
      - read only node-ok
    - **Hogyan detektáljuk, hogy valami kiesett?**
      - biztos, hogy nem „én” estem ki?
      - mi a helyzet, ha menet közben visszajön?

Operációs Rendszerek MSc

- Klaszterek
  - HA és LB sokszor együtt
  - Single Point of Failure (SPoF)
  - Elosztás több szempont szerint
    - Alkalmazás
      - Egyik szerver Web, másik SQL
      - DNS Round Robin
      - stb.
    - Hálózati réteg
      - Linux Virtual Server
      - HW Load Balancer
      - stb.
    - OS
    - HW

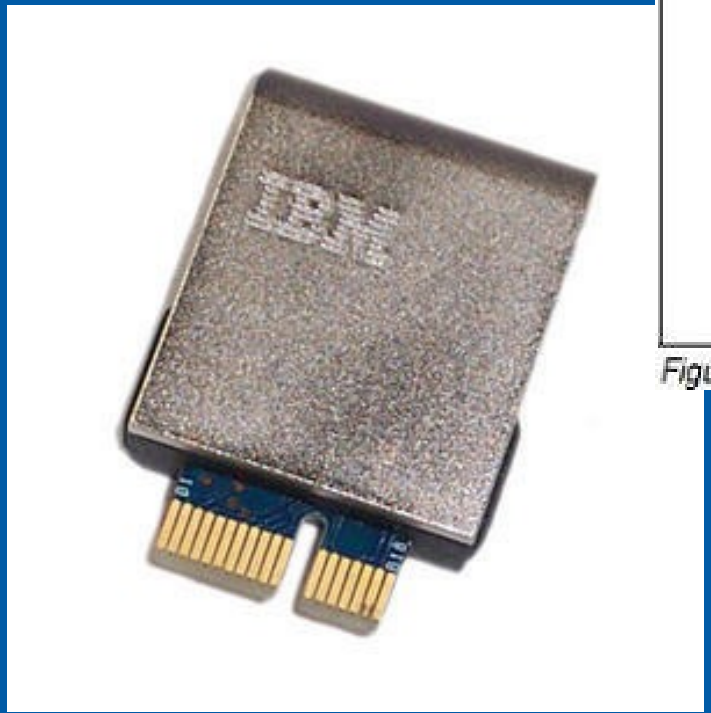
Operációs Rendszerek MSc

- Single System Image (SSI)
  - Erőforrások egyesítése
  - Több fizikai eszköz hardveres vagy szoftveres egyesítése egy logikai erőforrássá
  - Magyarul: több gép egynek látszódik
  - Processzek különböző node-okon
  - Memóriaterület elosztva (NUMA!)
  - Háttértár elosztva
    (jusson eszünkbe a RAID, konzisztencia, SAN...)
  - Checkpoint
  - Processz migráció
  - Szint: HW, OS, middleware, alkalmazás

Operációs Rendszerek MSc

- HW
  - Régen Digital Alpha rendszerek
    - OpenVMS galaxy
  - Korábban pl. IBM x3850M2 / x3950M2
  - 4 node fűzhető össze
    - saját interfész (ScaleXpander)
  - Egy gép 256GB RAM, 4x6-core CPU
  - Max 1TB RAM, 4*4-way, 4*4*6 mag
    - Maga a gép is elég drága + ScaleXpander - $3999
  - SGI UV 2000/3000

Operációs Rendszerek MSc



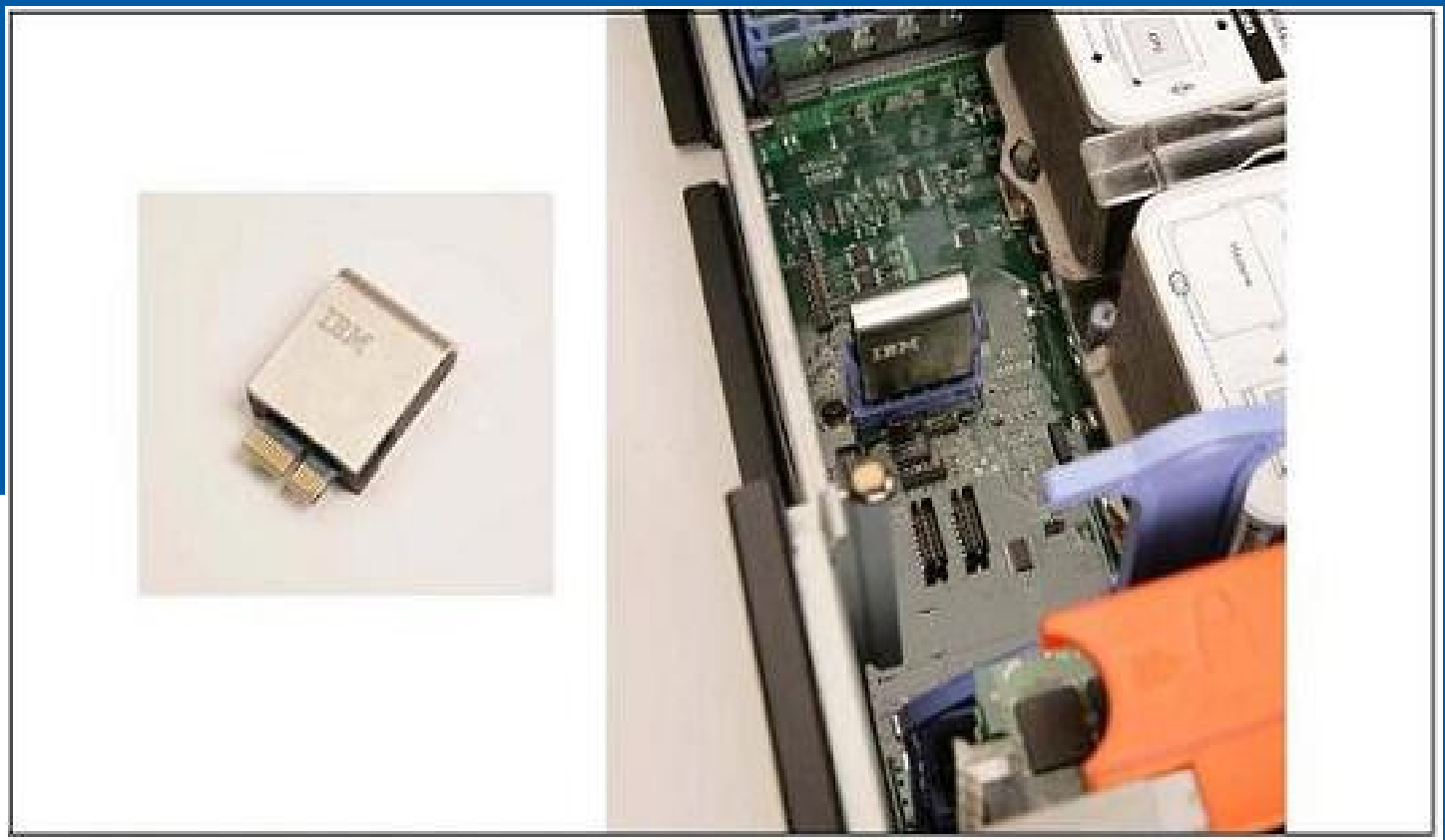
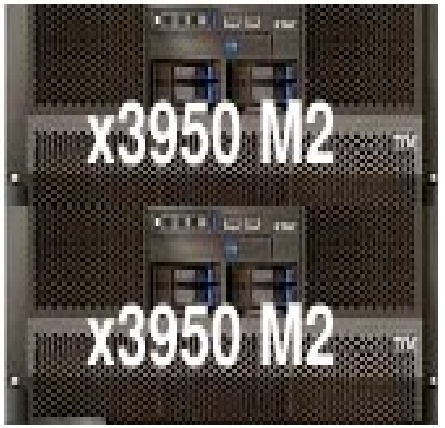


Figure 2 The ScaleXpander chip installed in the x3850 M2 enables the server to scale

One node
2-way or 4-way
Up to 256 GB RAM

Two nodes
4-way or 8-way
(Each node is
2-way or 4-way)
Up to 512 GB RAM



Three nodes*
6-way or 12-way
(Each node is
2-way or 4-way)
Up to 768 GB RAM

Four nodes*
8-way or 16-way
(Each node is
2-way or 4-way)
Up to 1 TB RAM

* 3-node and 4-node configurations are planned to be supported in 2Q08