

A CERN, mint a Big Data egyik bölcsője...

Barnaföldi Gergely Gábor

Berényi Dániel & Biró Gábor & Nagy-Egri Máté Ferenc & Andrew Lowe

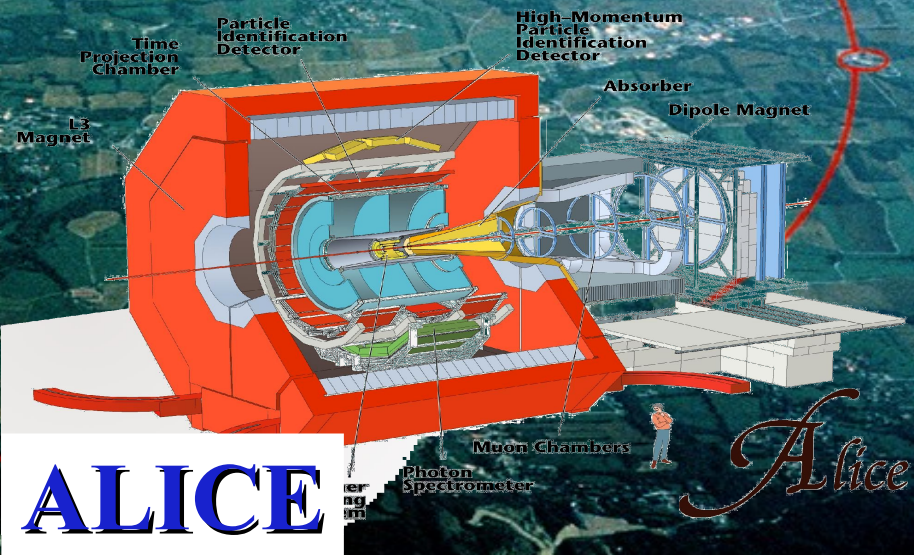
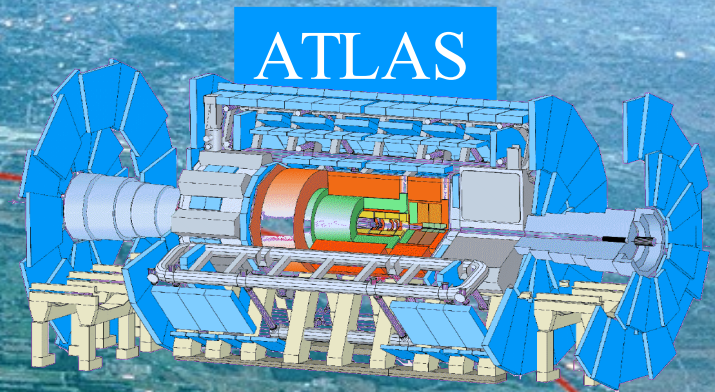
MTA Wigner FK Részecske- és Magfizikai Intézet & Wigner GPU Laboratórium



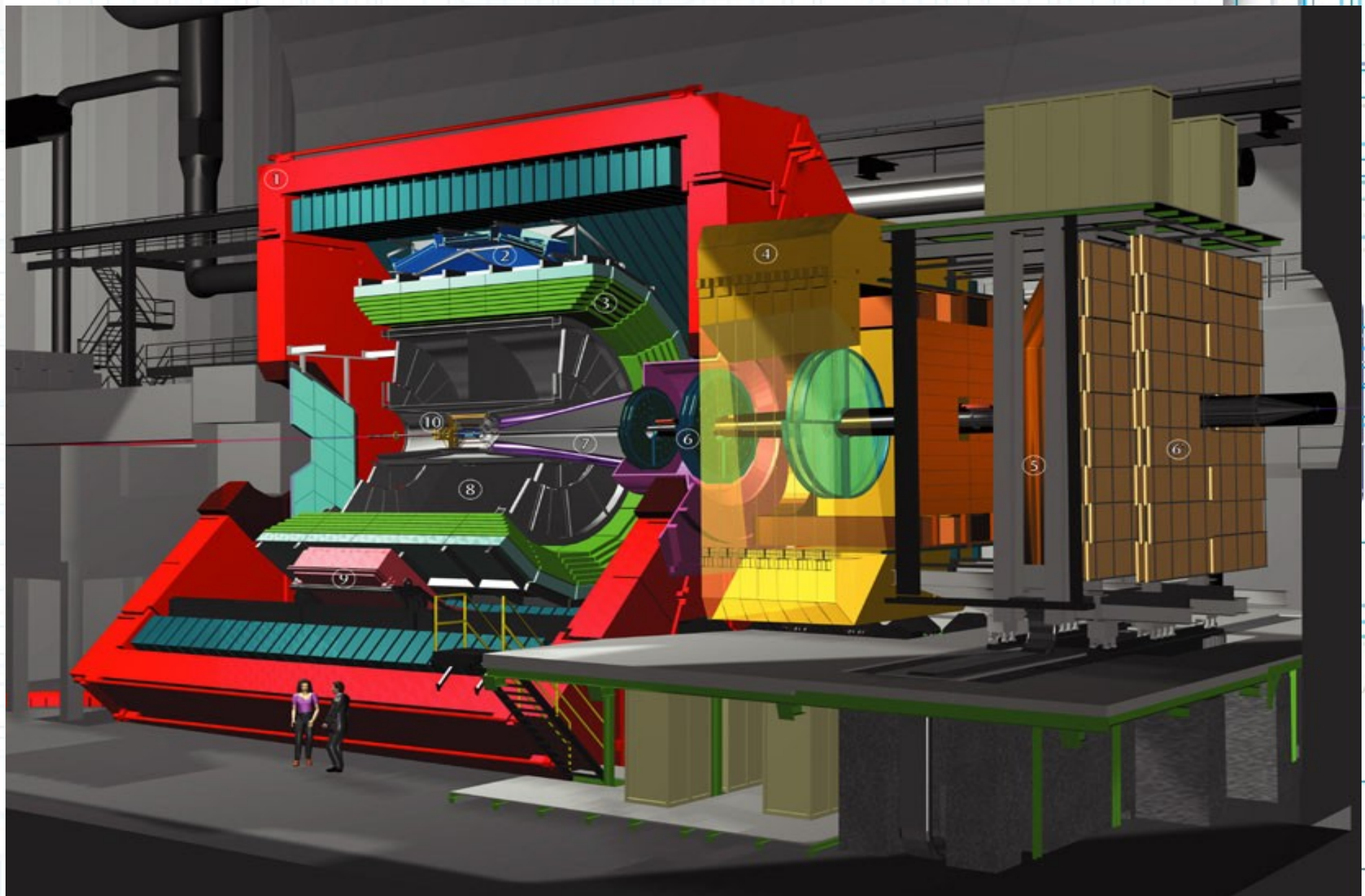
K I V O N A T

- A CERN és az LHC: Hol születik az adat?
- A legtöbb nyersadat feldolgozása LHC & WLCG
- Még több adat?
- Kutatási irányok az MTA Wigner FK
 - GPU alapú Monte Carlo generátorok (MC@GPU)
 - Vektorizált részecsketranszport (GeantV)
 - Részecskekeltés nagyenergián (HIJING)
 - Paramétertér redukció: jó-e fizikai paraméter?
- Mik az erőforrások?

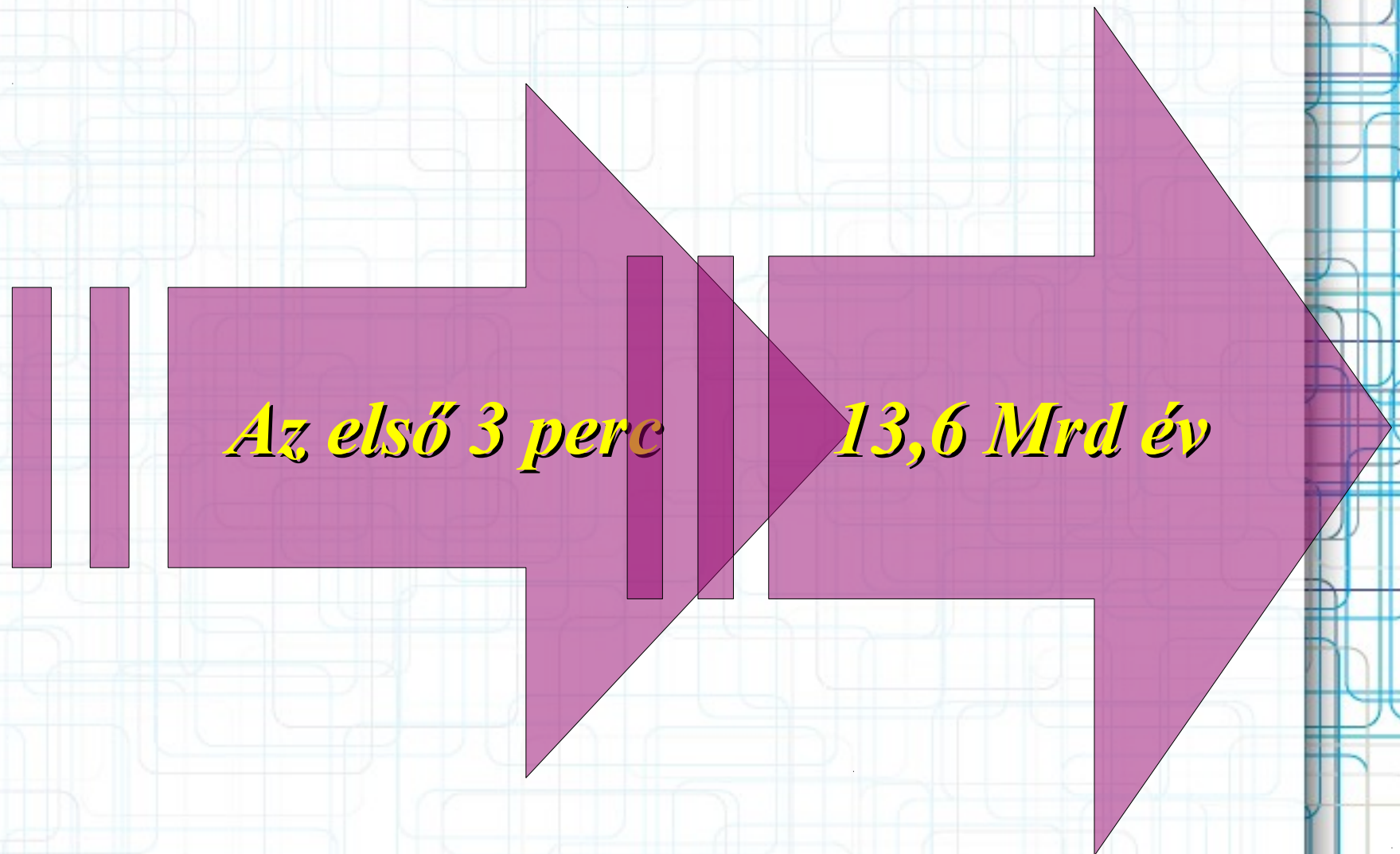
Az LHC és legnagyobb kísérletei



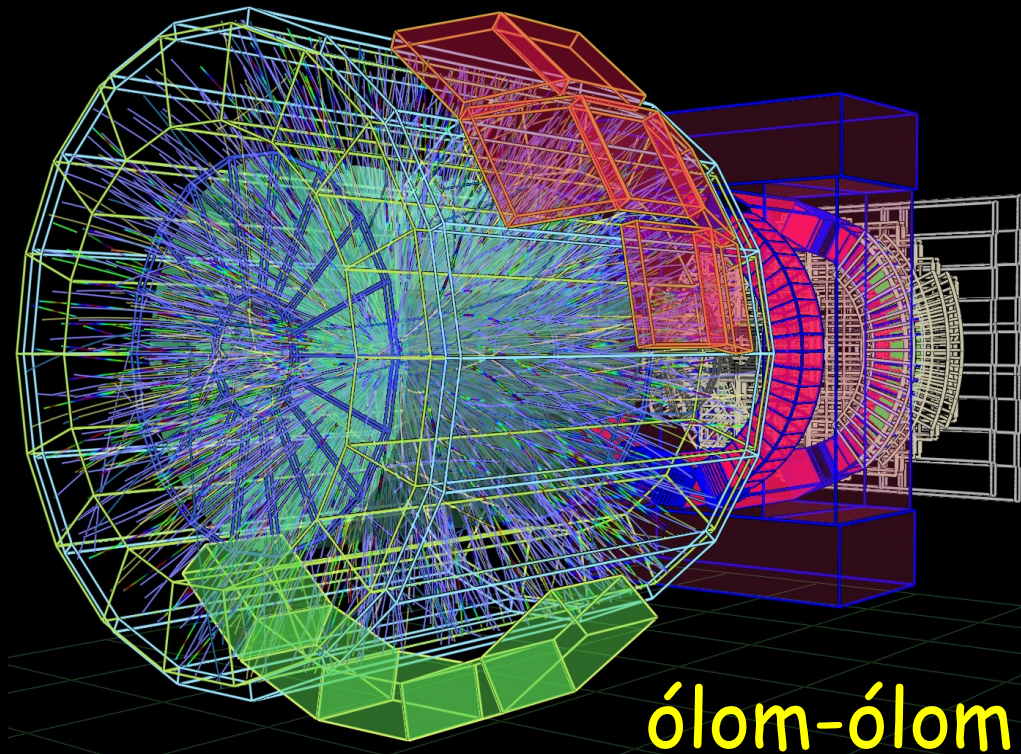
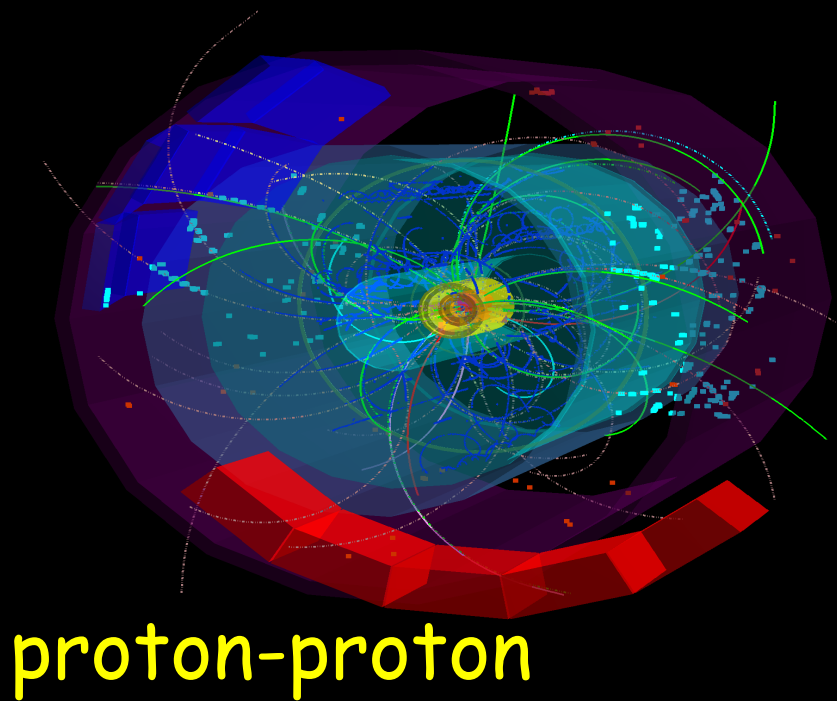
Az űsanyag-vizsgáló: ALICE detektor



A korai Univerzum anyagát keressük!



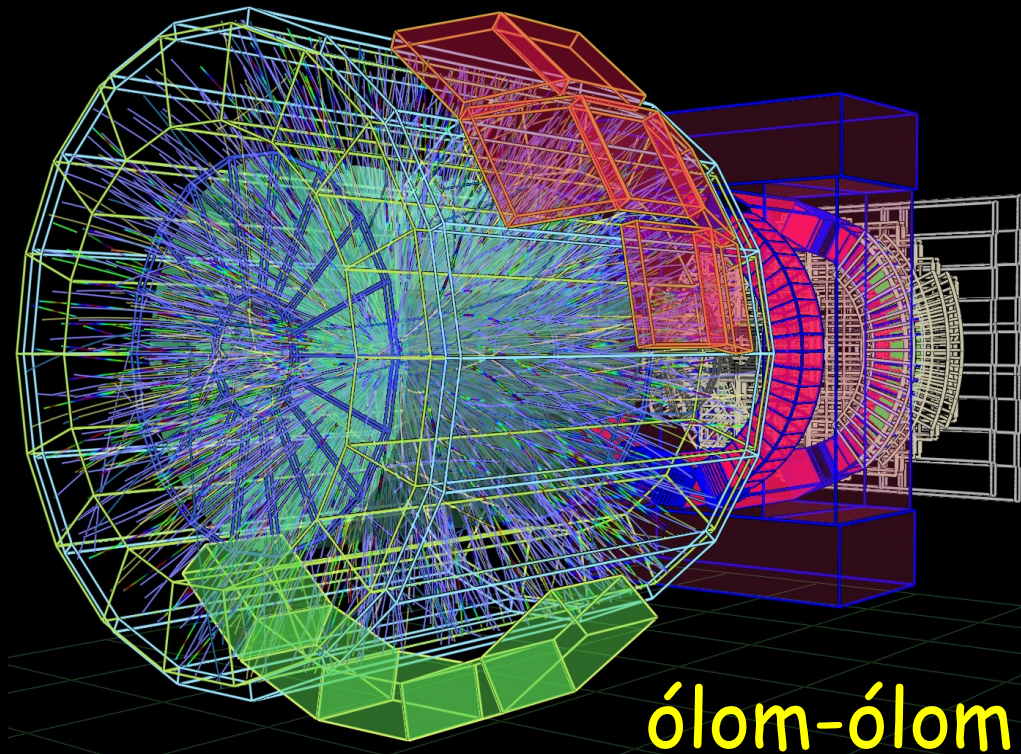
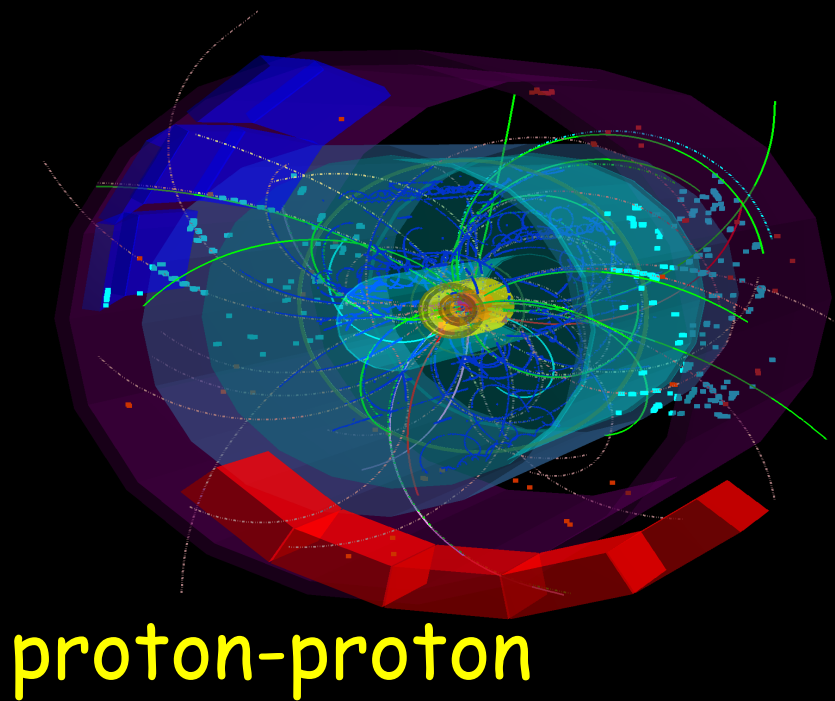
A korai Univerzum anyaga: forró sűrű űsanyag



Kvark Gluon Plazma (QGP):

- proton-proton vs. ólom-ólom
- forró, színes (kvark+gluon)
- a „tökéletes folyadék”...

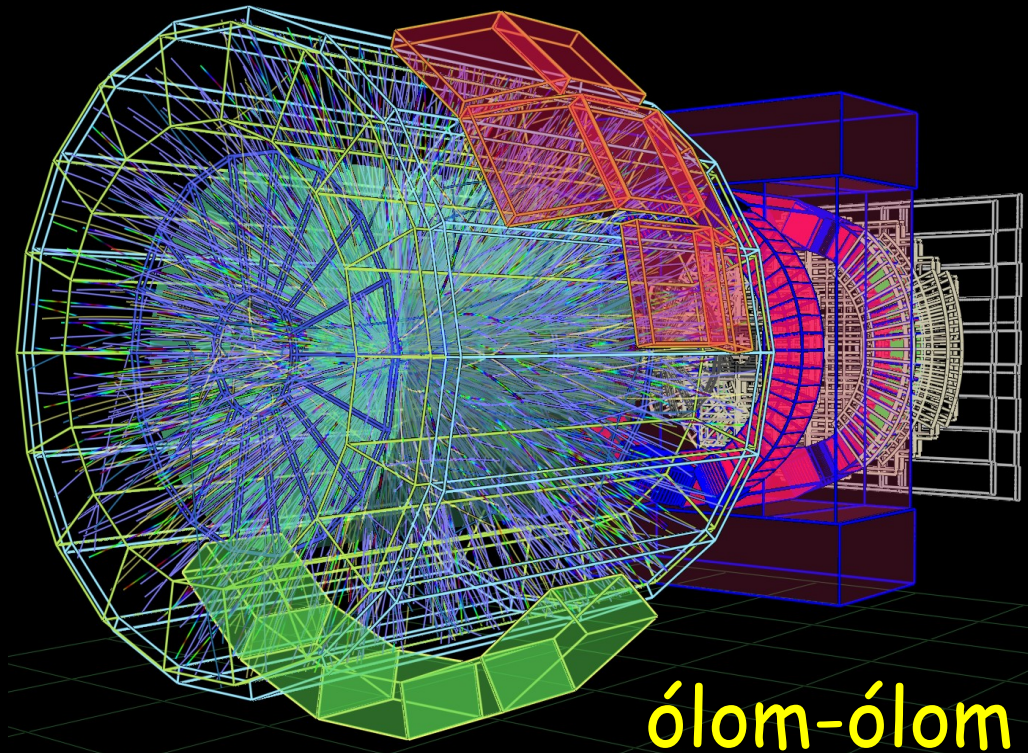
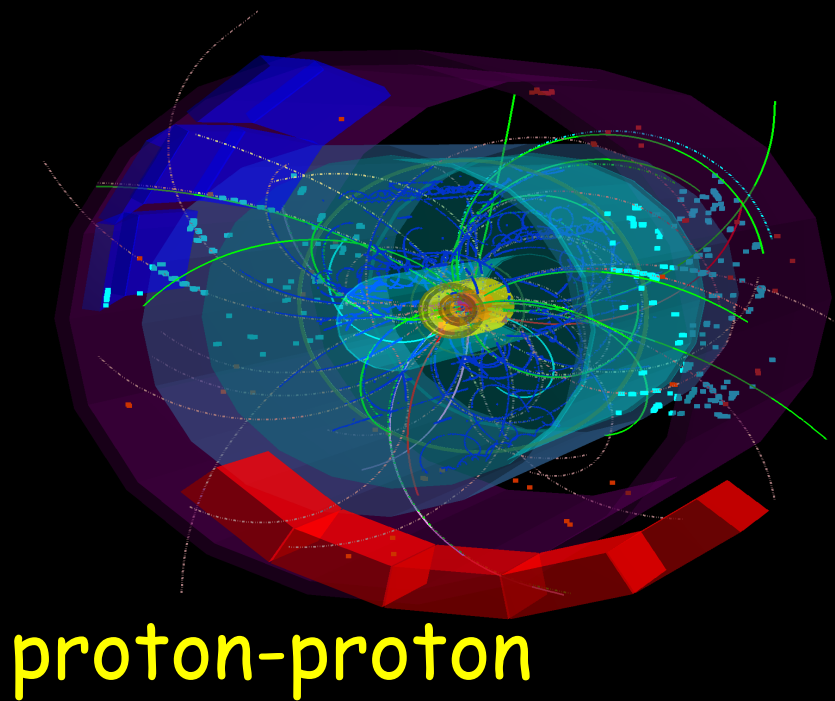
A korai Univerzum anyaga: forró sűrű őszanyag



Kvark Gluon Plazma (QGP):

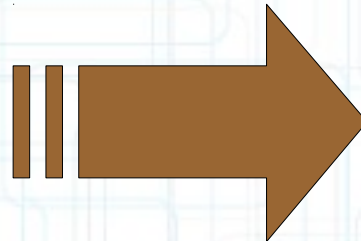
- proton-proton vs. ólom-ólom
- forró, színes (kvark+gluon)
- a „tökéletes folyadék”...

A korai Univerzum anyaga: forró sűrű őszanyag

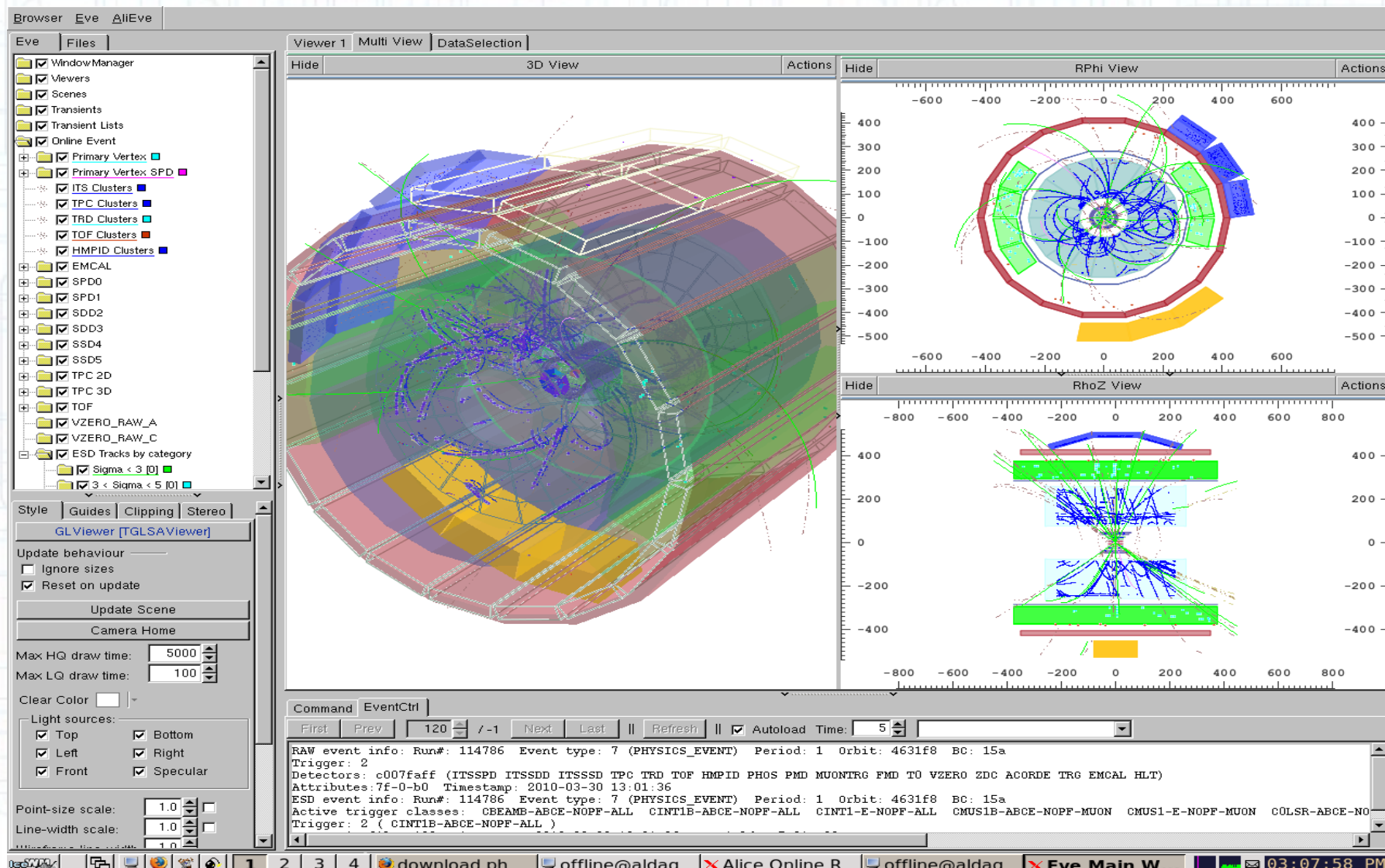


Kvark Gluon Plazma (QGP):

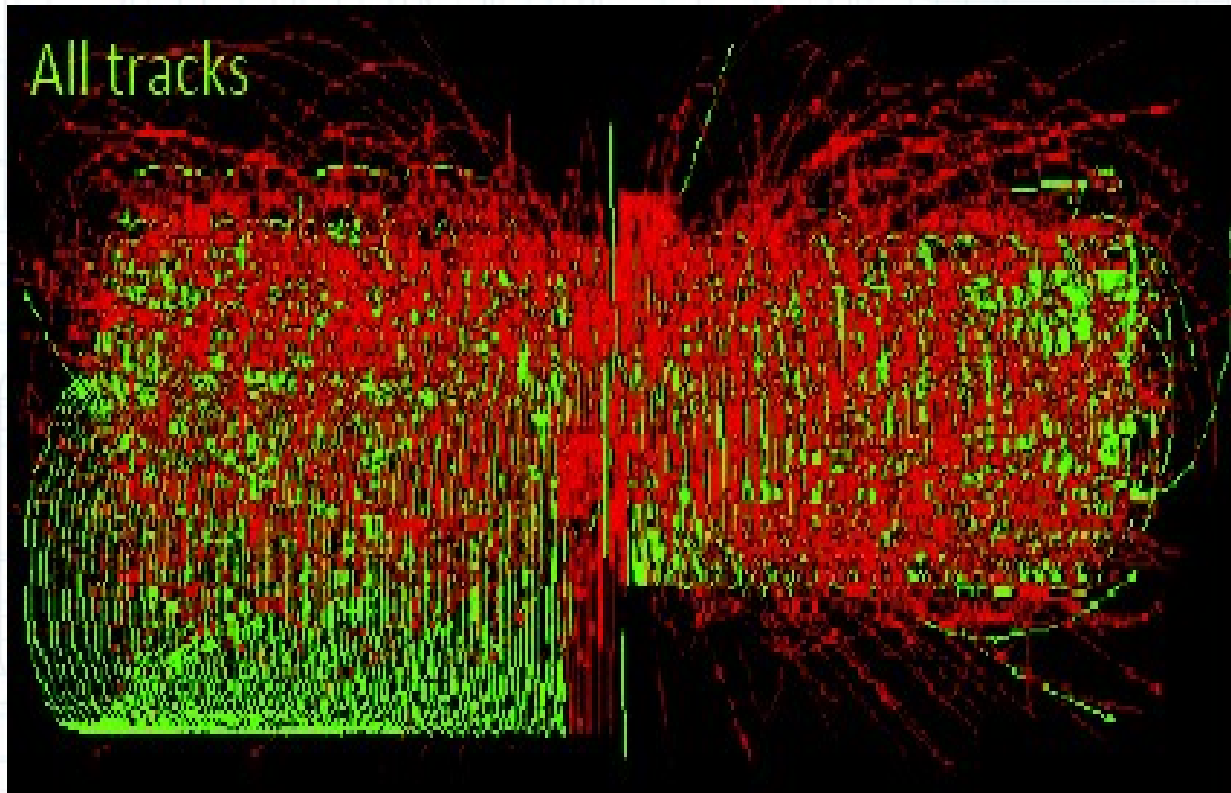
- proton-proton vs. ólom-ólom
- forró, színes (kvark+gluon)
- a „tökéletes folyadék”...



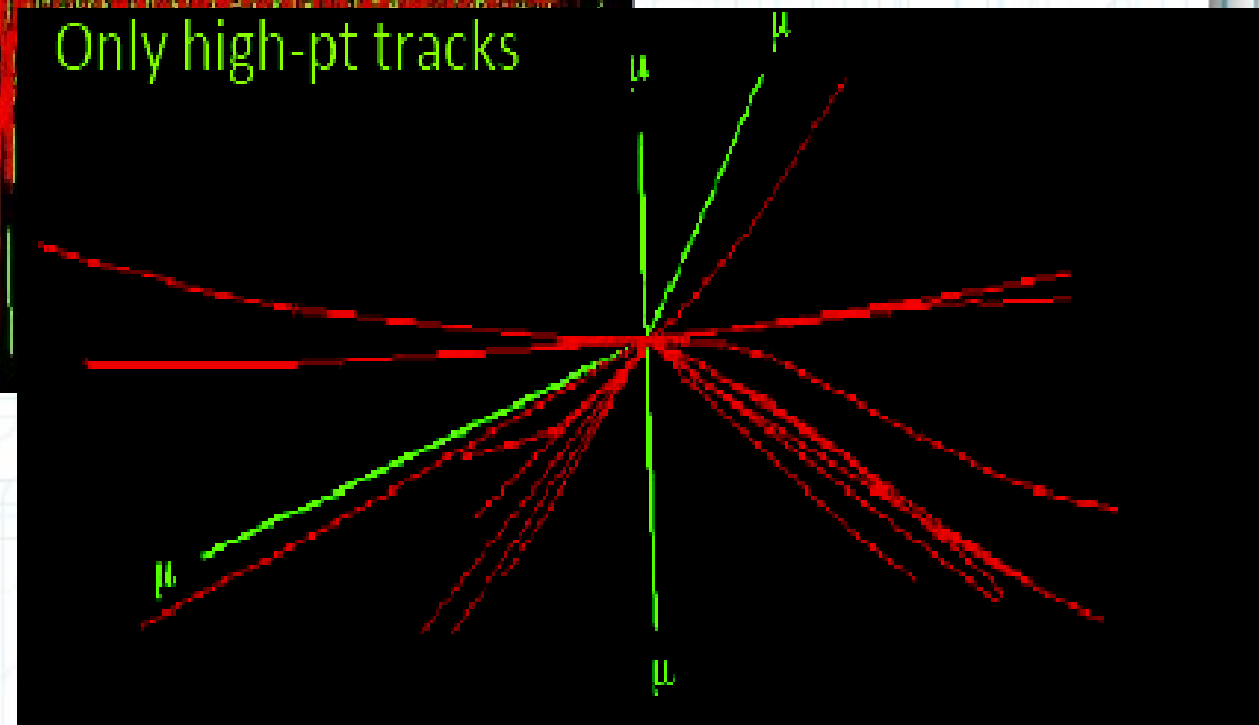
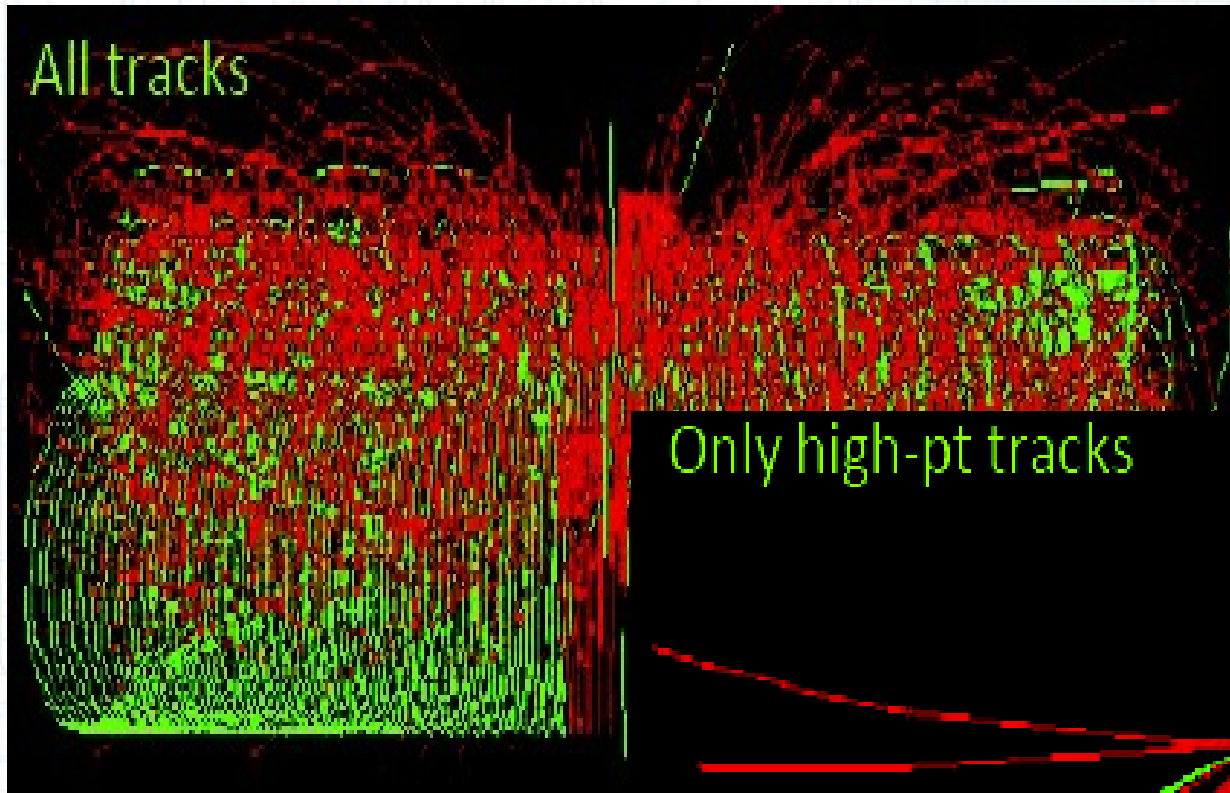
Mit látunk a „pilótafülkéből”?



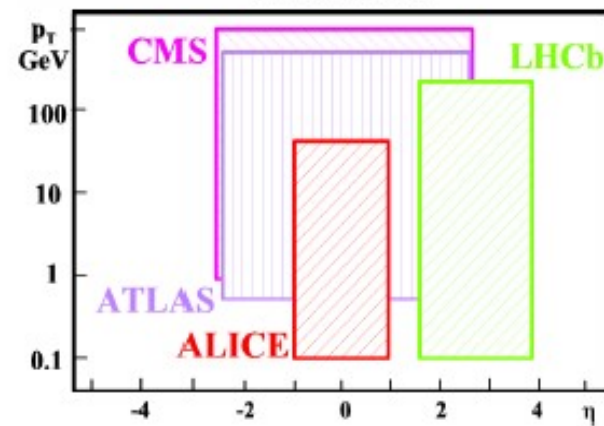
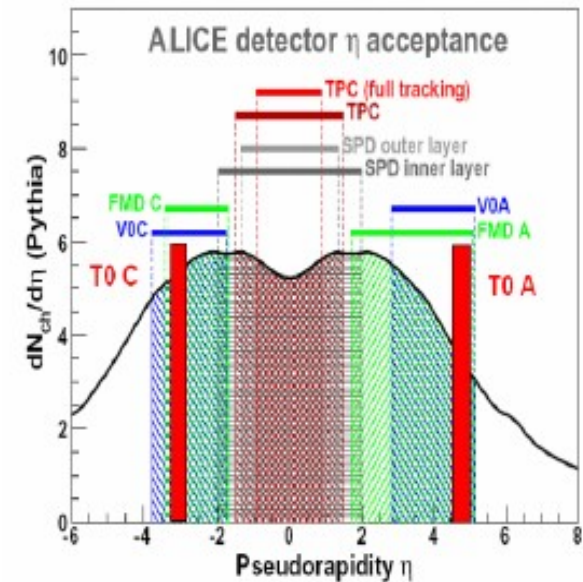
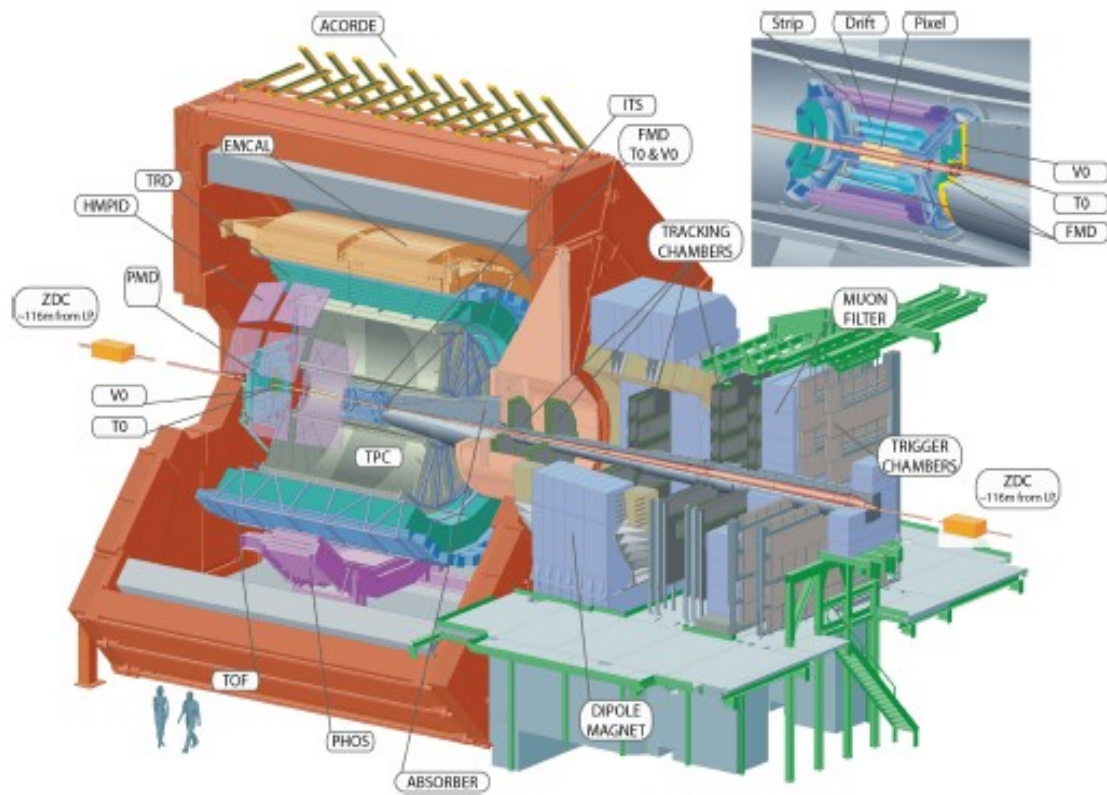
A track - amit keresünk?



A track - amit találunk!

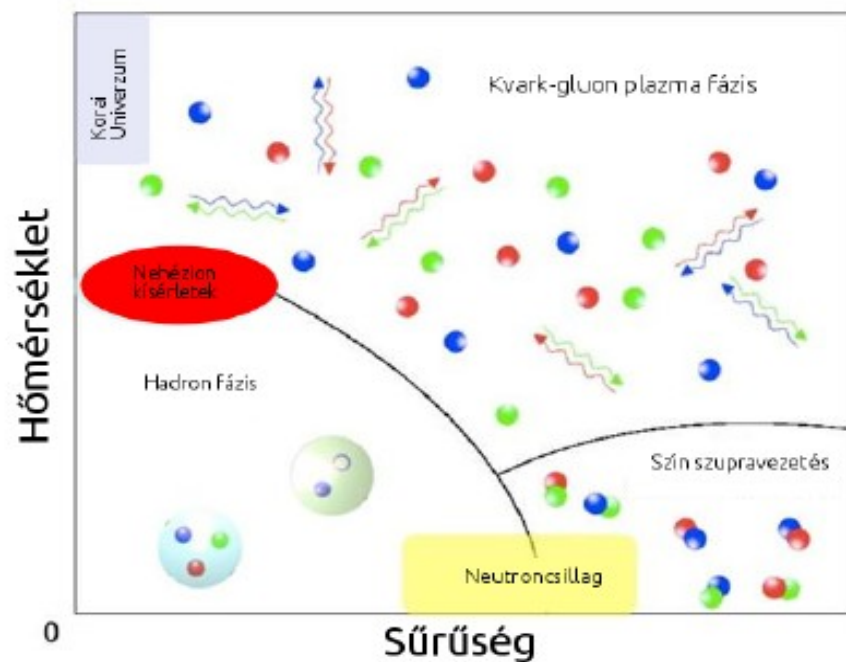


...akkor mindez az ALICE kísérletben



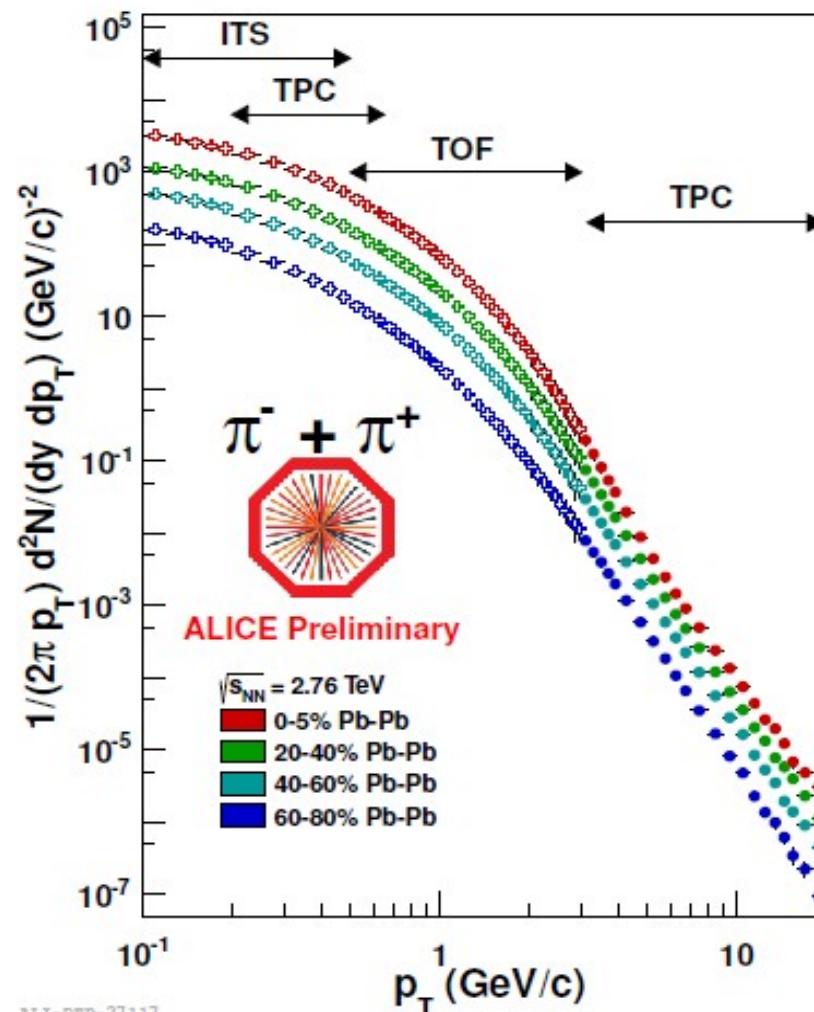
- ▶ ITS, TPC, TOF, HMPID...
- ▶ Részecskeazonosítás: $|\eta| < 0.9$

Fizika az ALICE-szal



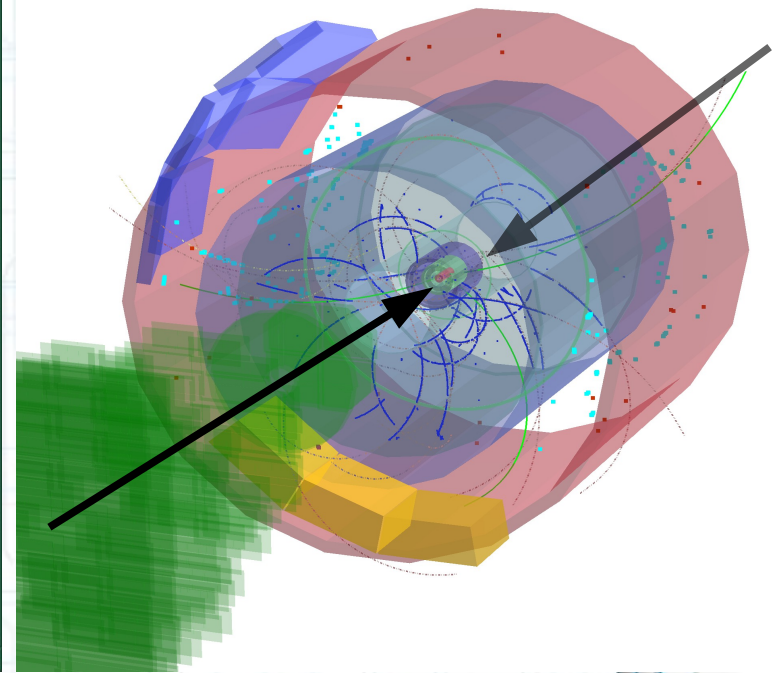
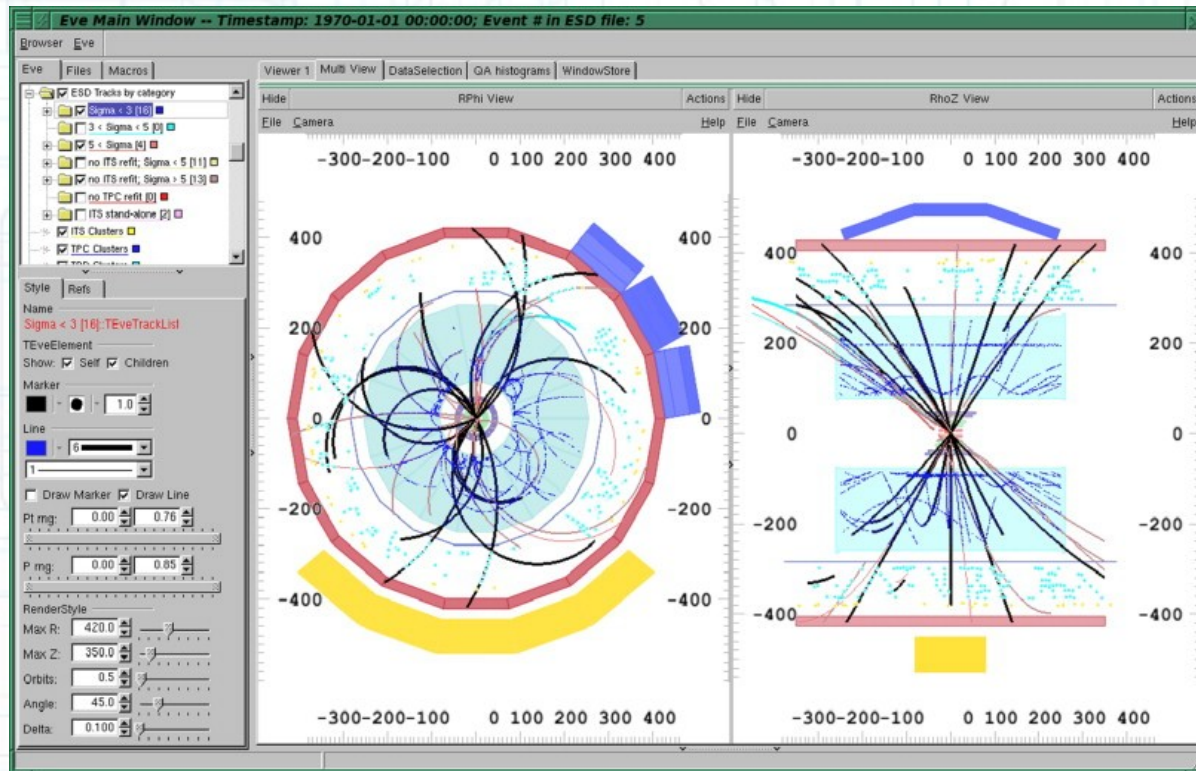
Azonosított hadronspektrumok alapján:

- ▶ A **kvark-gluon plazma** vizsgálata
- ▶ A **hadronizáció** modellje vákuumban és anyagban

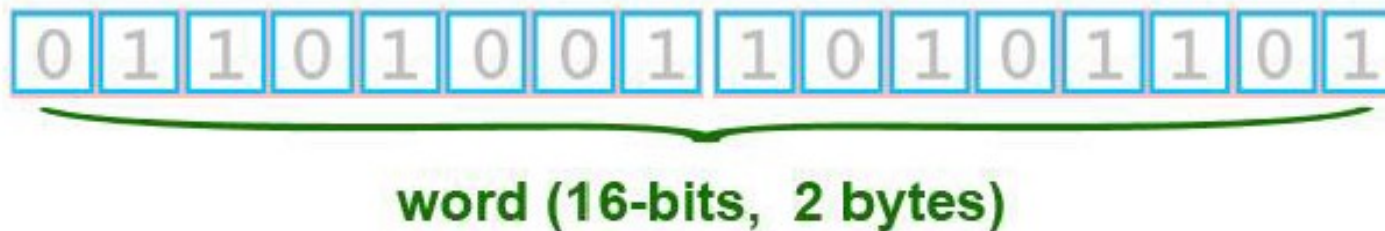


Egyszerű példa: proton-proton ütközés

- Egy rekonstruált esemény az ALICE kísérletben



Big Data - Bigger Data



Byte < KiloB < MegaB < GigaB < TeraB < PetaB < ExaB < ZettaB
1 10^3 10^6 10^9 10^{12} 10^{15} 10^{18} 10^{21}

Avogadro number: 6×10^{23}

Big Data - Bigger Data

SO WHAT IS A PETABYTE ANYWAY?

Source – www.mozy.com

WHAT IS A PETABYTE?

TO UNDERSTAND A PETABYTE WE MUST FIRST UNDERSTAND A GIGABYTE.

- 1 GIGABYTE = 7 MINUTES OF HD-TV VIDEO
- 2 GIGABYTES = 20 YARDS OF BOOKS ON A SHELF
- 4.7 GIGABYTES = SIZE OF A STANDARD DVD-R

THERE ARE A MILLION GIGABYTES IN A PETABYTE

"Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003." (That's something like 5 Exabytes of Data). - Eric Schmidt – Google 8/10

A PETABYTE IS A LOT OF DATA

- 1 PETABYTE = 20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
- 1 PETABYTE = 13.3 YEARS OF HD-TV VIDEO
- 1.5 PETABYTES = SIZE OF THE 10 BILLION PHOTOS ON FACEBOOK
- 15+ PETABYTES = INTERNET USER'S DATA BACKED UP ON MOZY.COM
- 20 PETABYTES = THE AMOUNT OF DATA PROCESSED BY GOOGLE PER DAY
- 20 PETABYTES = TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995
- 50 PETABYTES = THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

Twitter:
Over 7TB a Day in Tweets.

A ZETABYTE IS ONE MILLION PETABYTES!

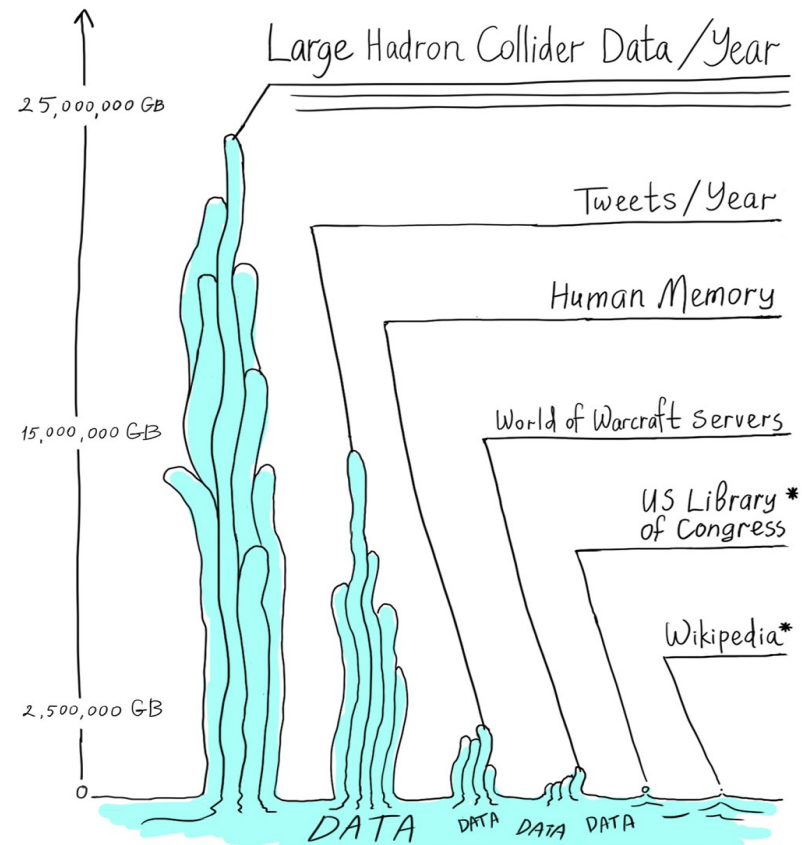
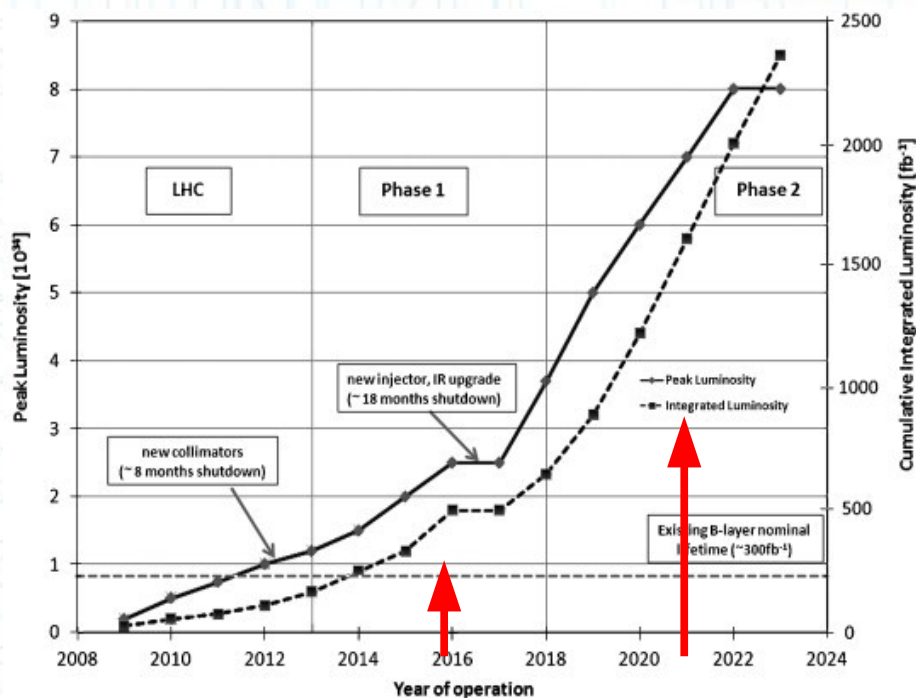
Facebook:
More than 750 Million Users.
Average user creates 90 Pieces of content each month.
More than 30B pieces of content shared each month.

9

Big Data - Bigger Data @ LHC

WLCG - Worldwide LHC Computing GRID:

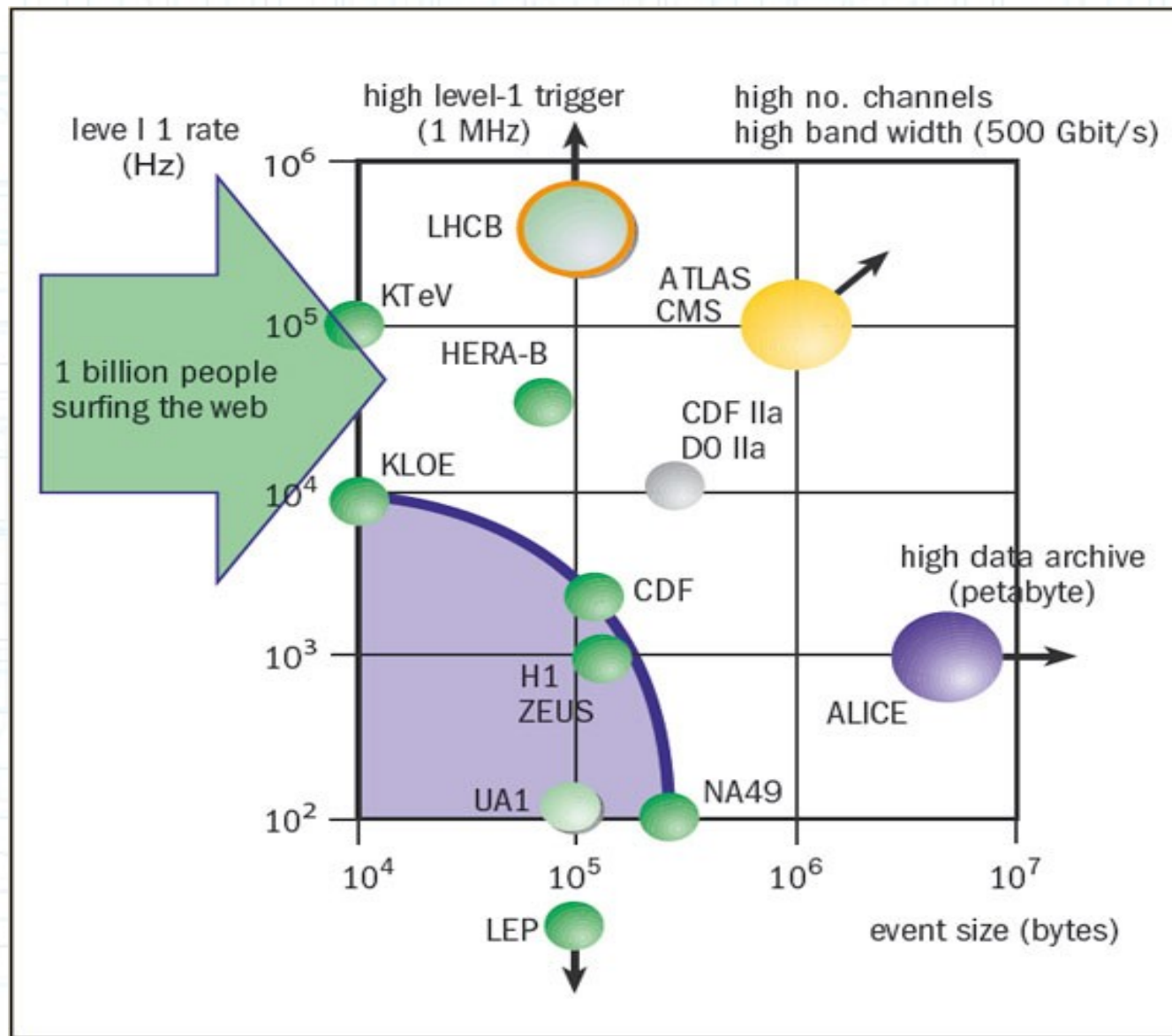
- 15-20 Petabytes data évente
- ...és még több...



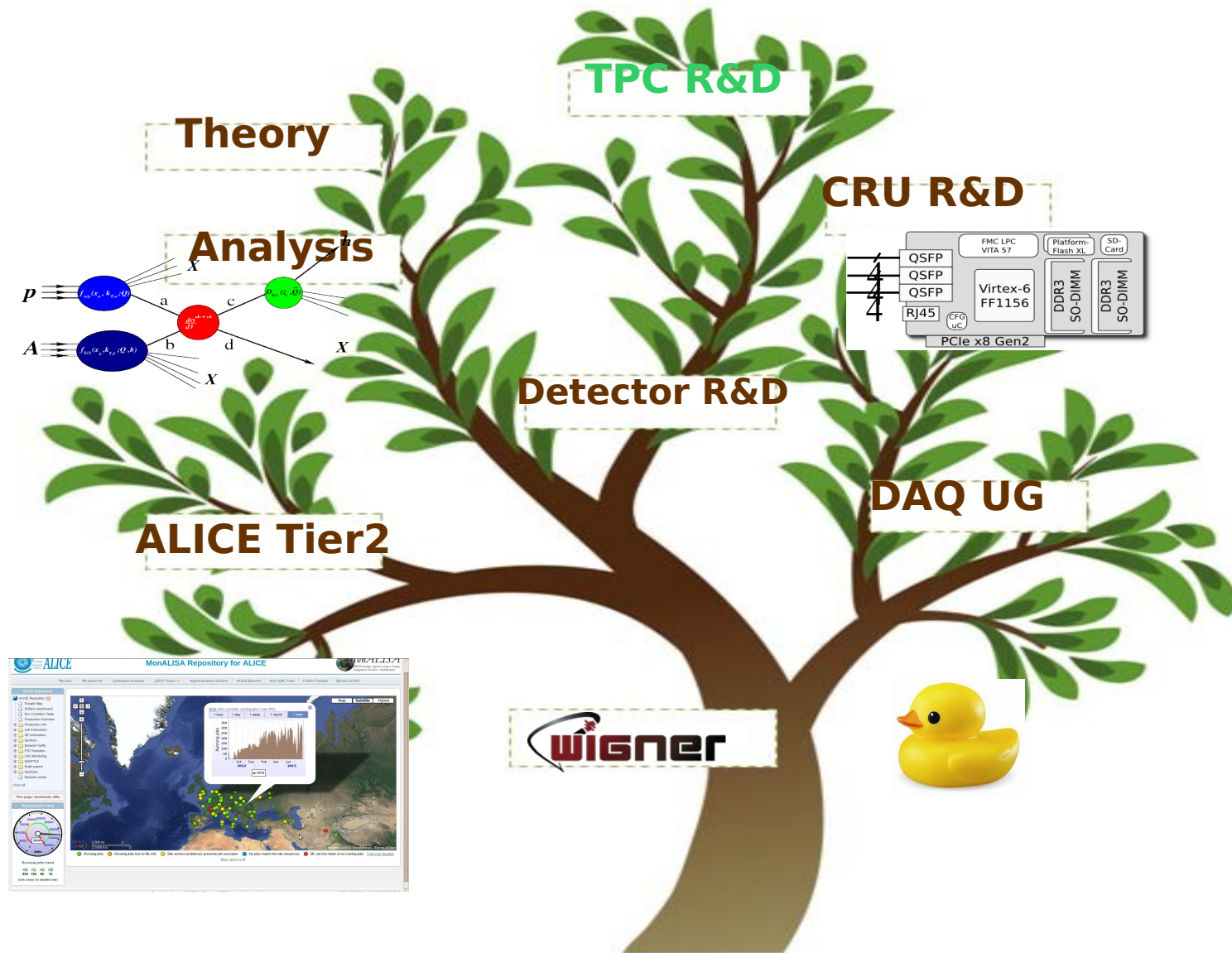
All numbers approximate.

* Binary Data

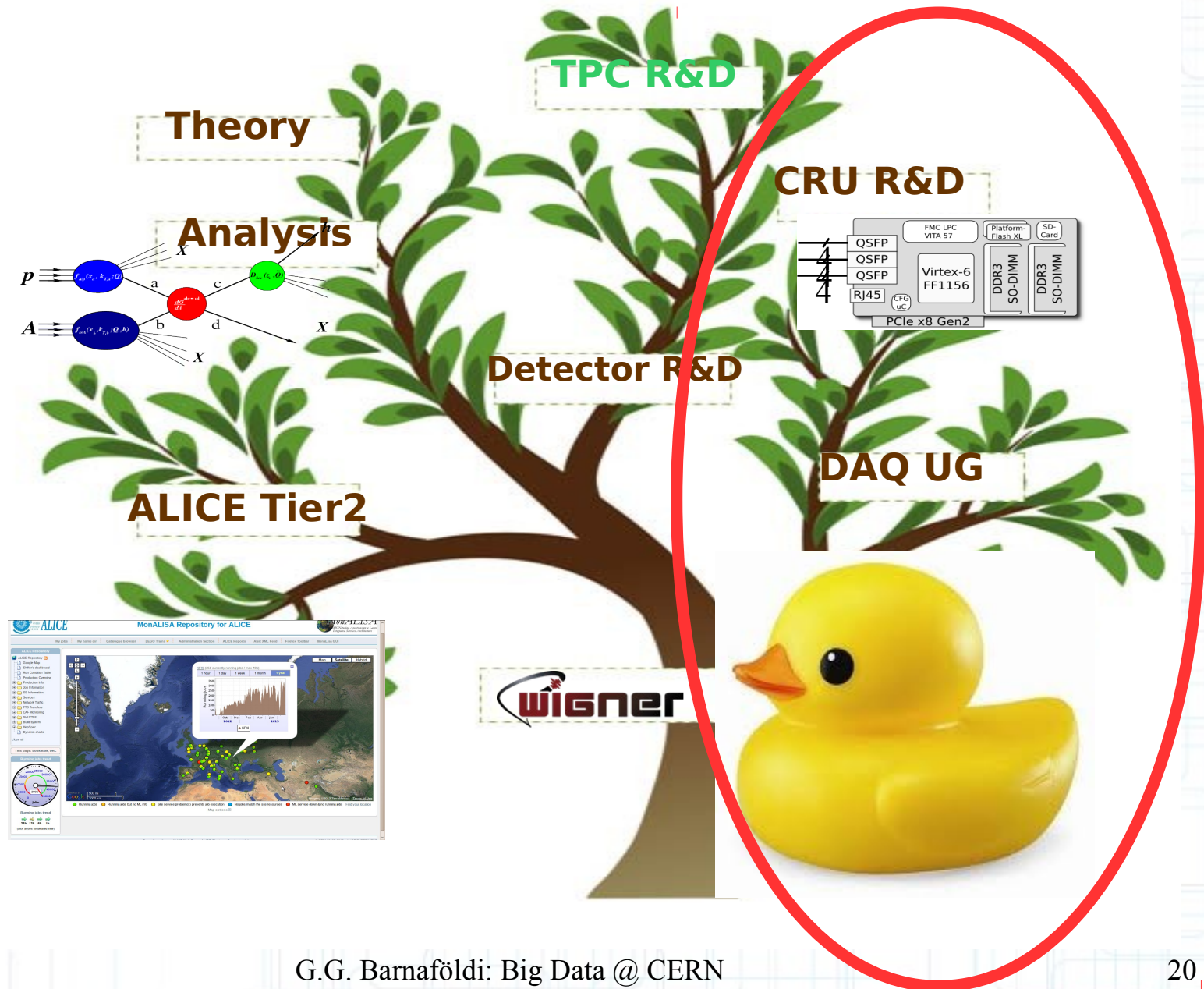
Big Data - Bigger Data @ LHC



Magyar részvétel az ALICE kísérletben



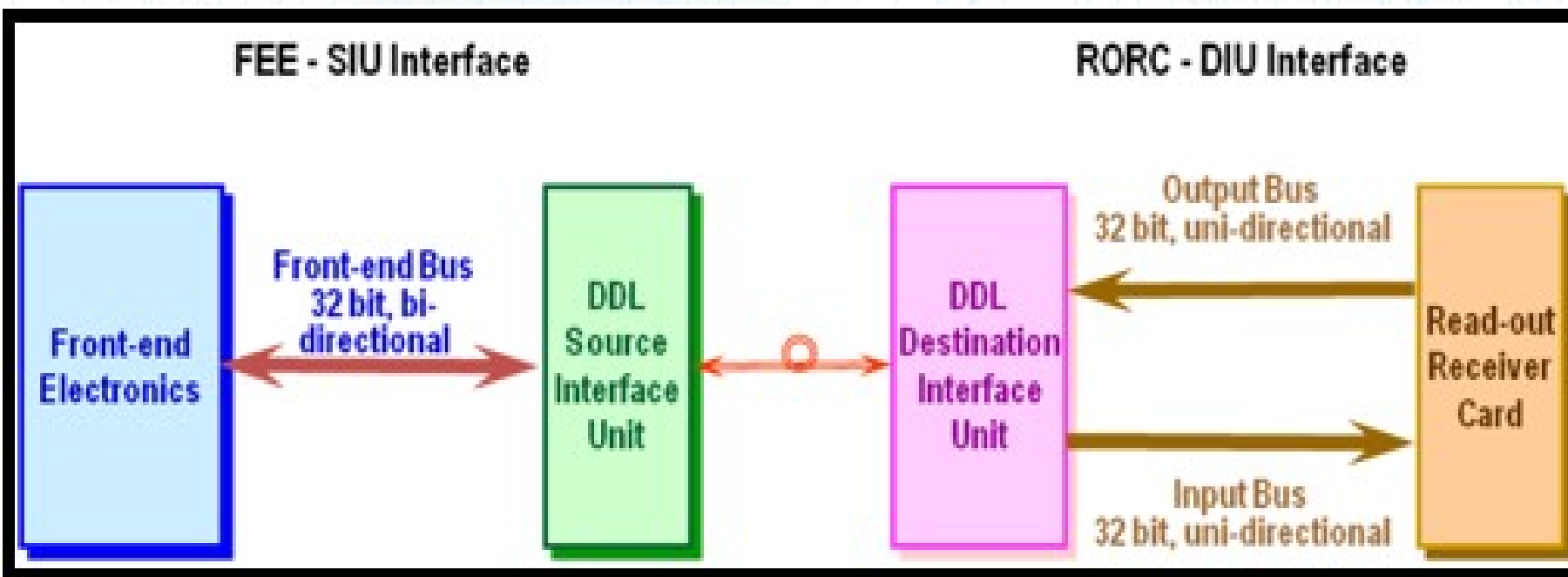
Magyar részvétel DAQ fejlesztésekben



ALICE DAQ: Információs versenypálya

ALICE DAQ/DDL adatgyűjtő/továbbító rendszer

A frontend elektronikák (FEE) és a adatgyűjtő számítógépek közötti kapcsolat a Detektor Data Link (DDL) és a Read-Out Receiver Card (RORC)



ALICE DAQ: Információs versenypálya

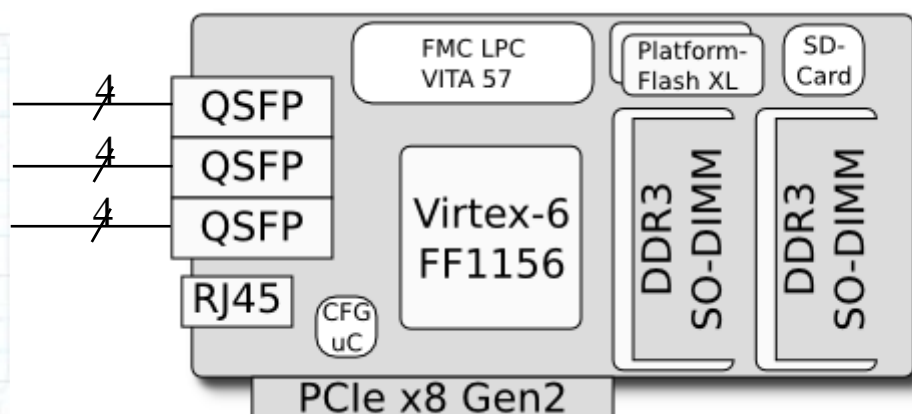
ALICE DAQ/DDL

adatgyűjtő/továbbító rendszer

- 500 db adat-link kártya: DDLs
- 450 db fogadó kártya: D-RORCs
- 2 PB/év adat kezelése
- Óriási sugárzási háttér (kRad)!
- Minden detektor, de legfőbb a TPC
- Más CERN kísérletek is használják
- Sikertörténet reloaded...

ALICE DDL/DAQ: adatfeldolgozó sztráda

- Közös DAQ és trigger DAQ/HLT DDL2, RORC2
 - Prototípus KÉSZ
 - Beszerelés/upgrade a Nagy Leállítás alatt (2014-16) (LS1) and LS2
 - 12 db DDL2 (6 Gb/s) link együtt DAQ LDC (36 Gb/s) forgalom.
 - PCIe V2 8 buszokon (500 MB/s/lane) I/O 32 Gb/s
 - FPGA alapú adatfeldolgozás már trigger DAQ szinten(e.g cluster findina)



- Jelen: 1 PC-ben 5 link (2Gb/s) I/O (10 Gb/s)
- Prototípus paraméterek (fejlesztés alatt)
 - 12 link (6 Gb/s)
 - 6 link DAQ LDC együtt (36 Gb/s).
 - PCIe2 x8 (500 MB/s/sáv) I/O (32 Gb/s)
- Beszereléskor
 - 12 links (10 Gb/s/PC)
 - PCIe3 16 sávon I/O (128 Gb/s)

DDL SIU EVOLÚCIÓ

SIU1



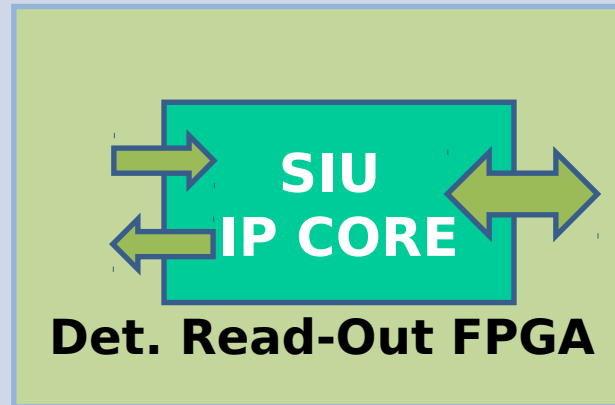
1 ch @ 2 Gb/s
ACTEL FPGA
(CORE cost 560 CHF)

Custom DDL protocol

RUN1

L
S
1

SIU2



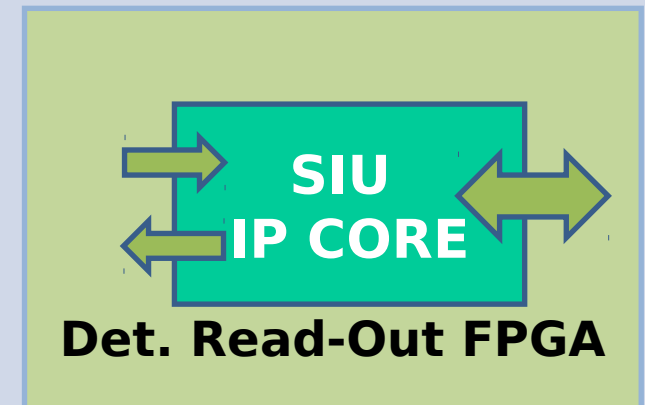
Up to 6 Gb/s
XILINX / ALTERA / ACTEL
FPGA
(CORE cost 0 CHF)

Custom DDL protocol
(same protocol but faster)

RUN2

L
S
2

SIU3



10 Gb/s
XILINX / ALTERA / ACTEL
FPGA
(CORE cost 0 CHF)

- Custom DDL 10 Gb/s
- Ethernet @ 10 Gb/s
- PCIe over fibre

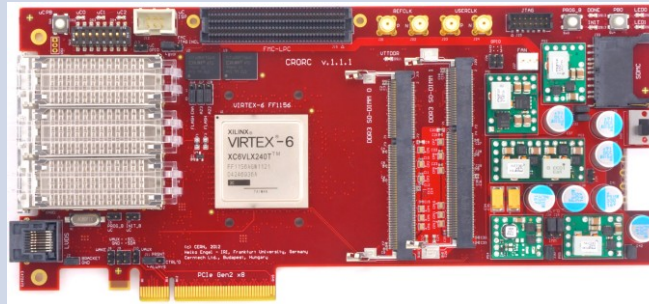
RUN3

RORC EVOLUCIÓ

RORC1

RORC2 (aka C-RORC)

RORC3



TBD

2 ch @ 2 Gb/s
PCIe gen.1 x4 (1 GB/s)
ALTERA FPGA

12 ch @ up to 6 Gb/s
PCIe gen.2 x8 (4 GB/s)
XILINX FPGA

12 ch @ 10 Gb/s
PCIe gen.3
ALTERA / XILINX

Custom DDL protocol

Custom DDL protocol
(same protocol but faster)

- Custom DDL 10 Gb/s
- Ethernet @ 10 Gb/s
- PCIe over fibre

RUN1

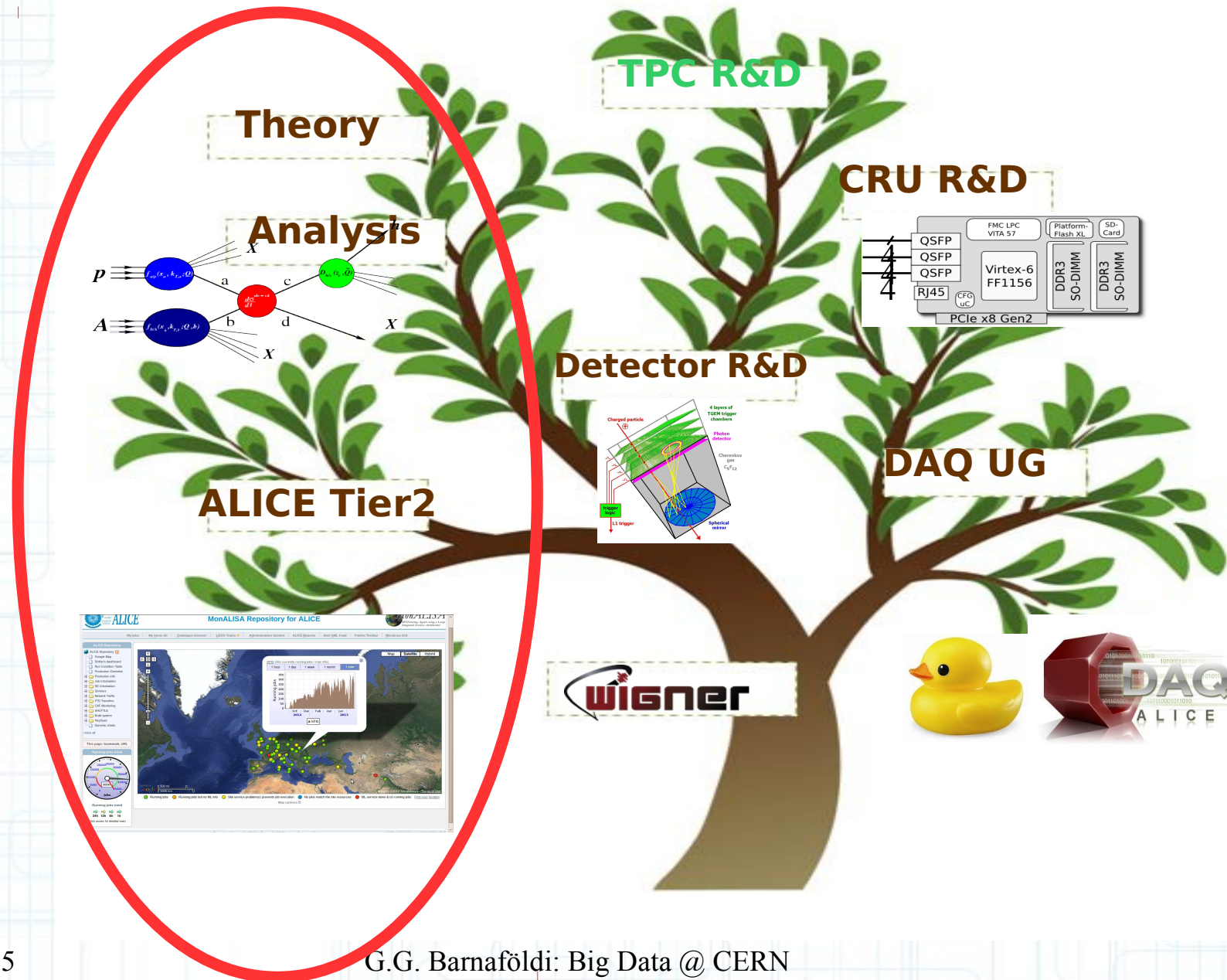
L
S
1

RUN2

L
S
2

RUN3

Magyar részvétel az adatanalízisben

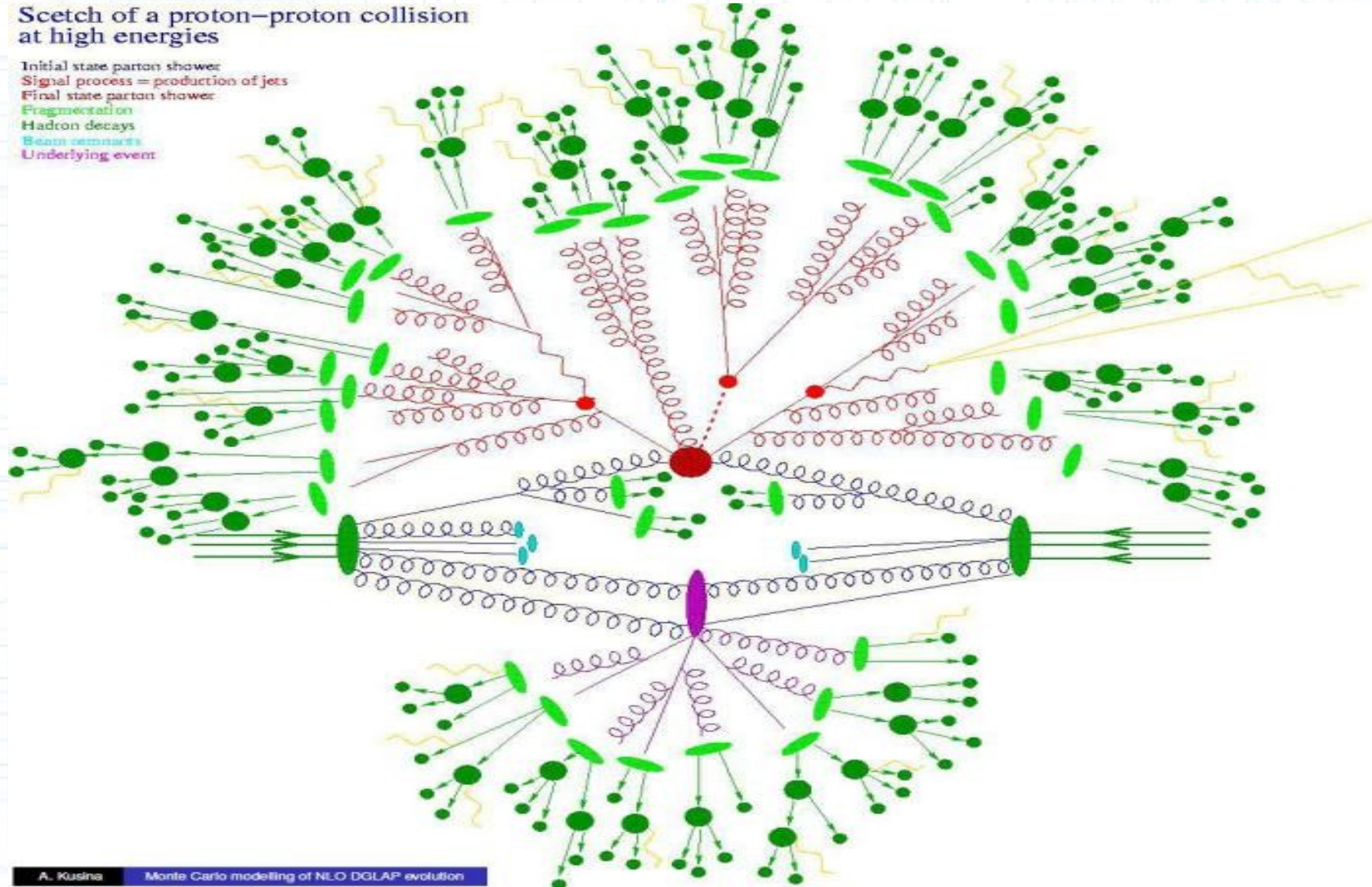


Egy kis fizikai: proton-proton ütközés

- Egyszerű elméleti modell pp ütközésre

Sketch of a proton-proton collision at high energies

Initial state parton shower
Signal process = production of jets
Final state parton shower
Fragmentation
Hadron decays
Beam remnants
Underlying event



A. Kusina Monte Carlo modelling of NLO DGLAP evolution

A mért adat még nem elég

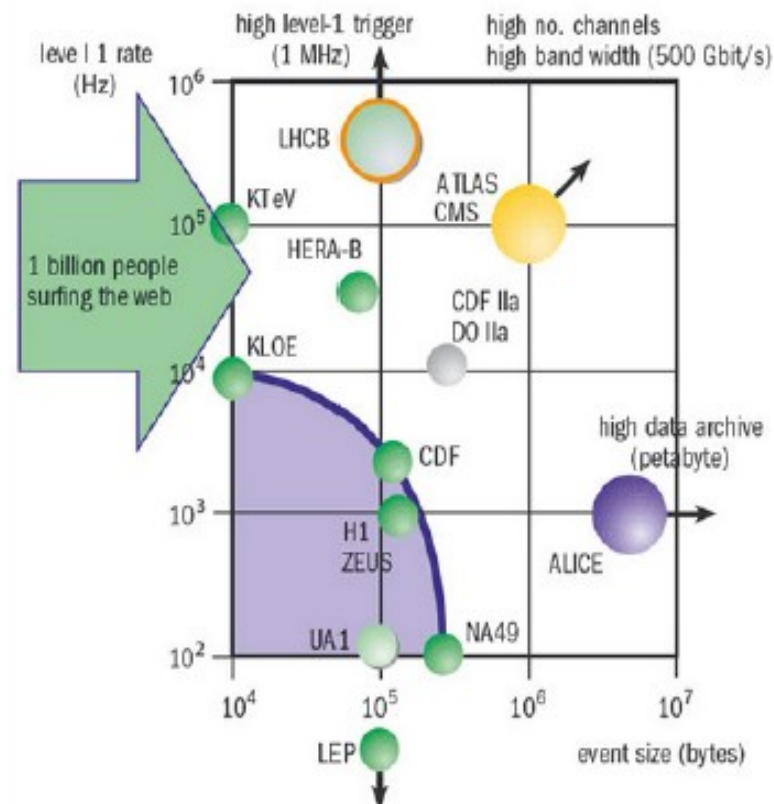


Szimulált esemény



Valódi esemény

- ▶ Minél **több**, annál jobb (pontosság $\sim 1/\sqrt{N}$)
- ▶ **Ideális**: szimulált adatmennyiség $\approx$ valós adatmennyiség
- ▶ **Szükséges** idő az ALICE geometriájának szimulálására: 3.8 ms/pálya
- ▶ **Szükséges** idő az ALICE 1 másodpercének szimulálásához: $\mathcal{O}(\text{napok})$



Részecsketranszport: sok erőforrás

Sok **időre** van szükség...



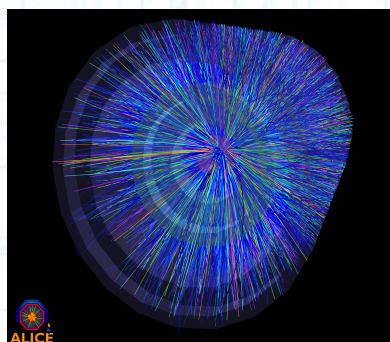
<http://www.kgionline.com/distributed-computing.jpg>

...vagy sok **CPU**-ra...

MC generátorok a nagyenergiás fizikában

Miért van szükség Monte Carlo generátorokra?

- Vannak fizikai problémák amelyekre nincs zárt, analitikus képlet, vagy determinisztikus leírás:



- Stohasztikus folyamatok (független események)
- numerikus (multi-D) integrálok
- optimalizációk

Megoldási módszerek & és hibák

- Numerikus eredmények véletlen mintavételezése
- Hibabecslés (standard deviation)

Gyors random számok → Computing & IT

Gyors számítások = parallel számítások

- Moore törvénye:



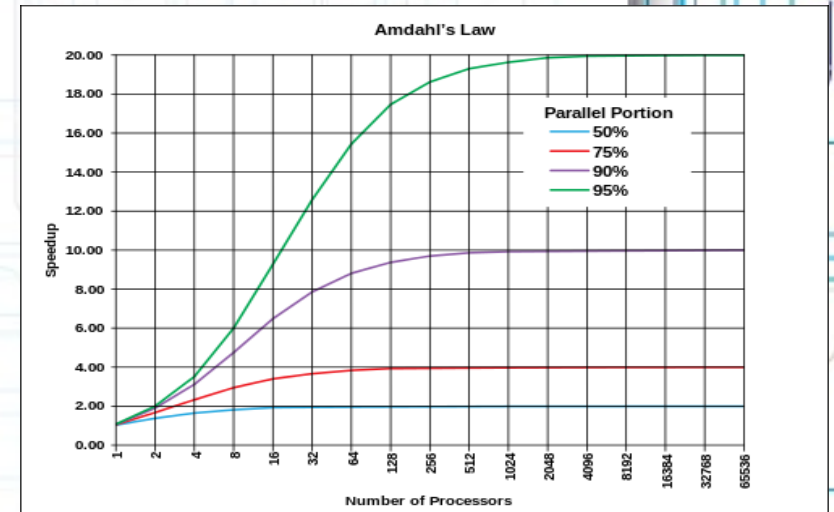
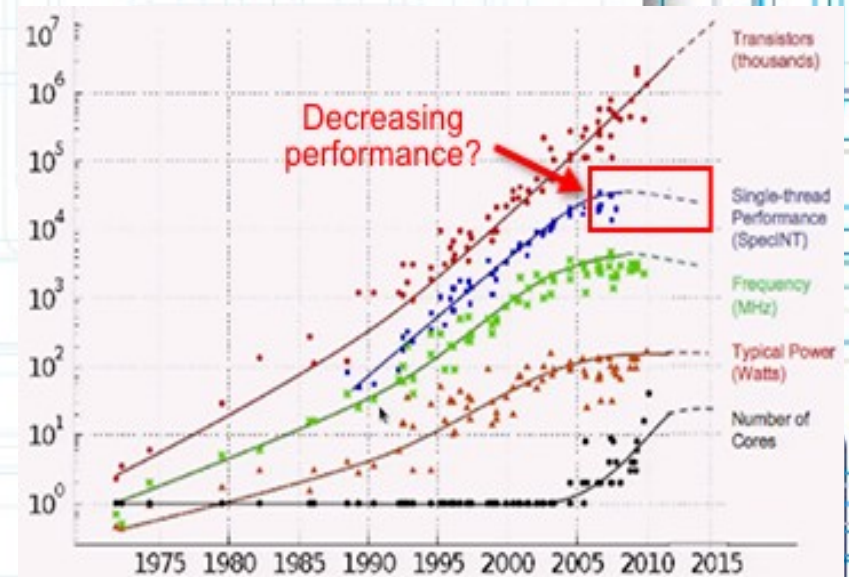
Kétévente a tranzistorok száma megduplázódik az számítástechnikai integrált áramkörökben.

- Amdalh törvénye:



A párhuzamosítás során nyerhető gyorsítás, ha a program p hányada párhuzamosítható N szálon:

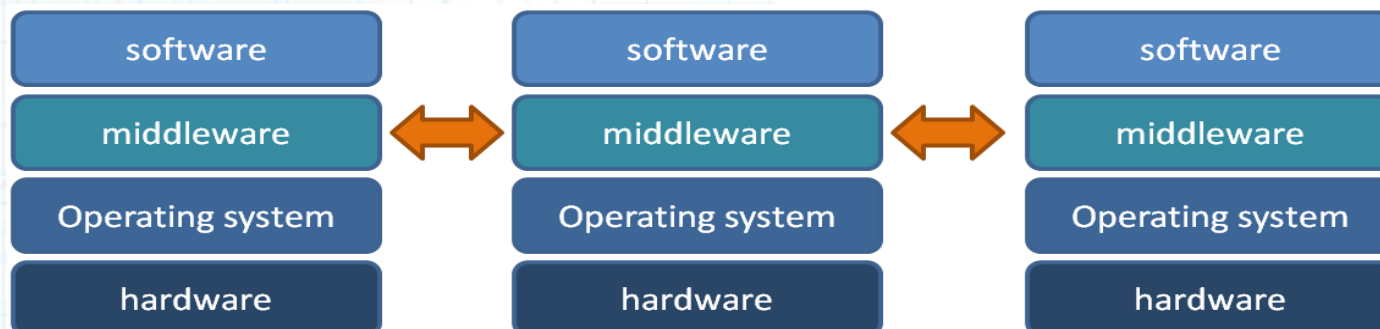
$$S(N) = \frac{1}{(1 - P) + \frac{P}{N}}$$



Hogyan használjuk ki jobban a WLCG-t

WLCG:

- Kritikus pont a WN-ek száma és teljesítménye.
- Multicore gépek, single-thread számolások
- Ha van szabad multi-core kapacitás, akkor software és middleware szintű belépés



Egy lehetőség: Vektorizáció

- ▶ ...**vagy** párhuzamosíthatunk!
- ▶ **Esemény** szintű parallelizáció: a számítási idő szétosztása (Worldwide LHC Computing Grid)
- ▶ **Utasítás** szintű parallelizáció: valódi sebességnövekedés
 - > **Vektorizáció**: SIMD (Single Instruction Multiple Data) utasítások használata

Skalár mód

$$\begin{array}{c} a[i] \\ + \\ b[i] \\ = \\ a[i] + b[i] \end{array}$$

SIMD művelet

$$\begin{array}{cccccccc} a[i] & a[i+1] & a[i+2] & a[i+3] & a[i+4] & a[i+5] & a[i+6] & a[i+7] \\ + \\ b[i] & b[i+1] & b[i+2] & b[i+3] & b[i+4] & b[i+5] & b[i+6] & b[i+7] \\ = \\ a[i] & a[i+1] & a[i+2] & a[i+3] & a[i+4] & a[i+5] & a[i+6] & a[i+7] \\ +b[i] & +b[i+1] & +b[i+2] & +b[i+3] & +b[i+4] & +b[i+5] & +b[i+6] & +b[i+7] \end{array}$$

GeantV fejlesztés

- ▶ Előzetes tanulmányok szerint a részecsketranszport a legtöbb detektorgeometriában erősen **lokalizált**

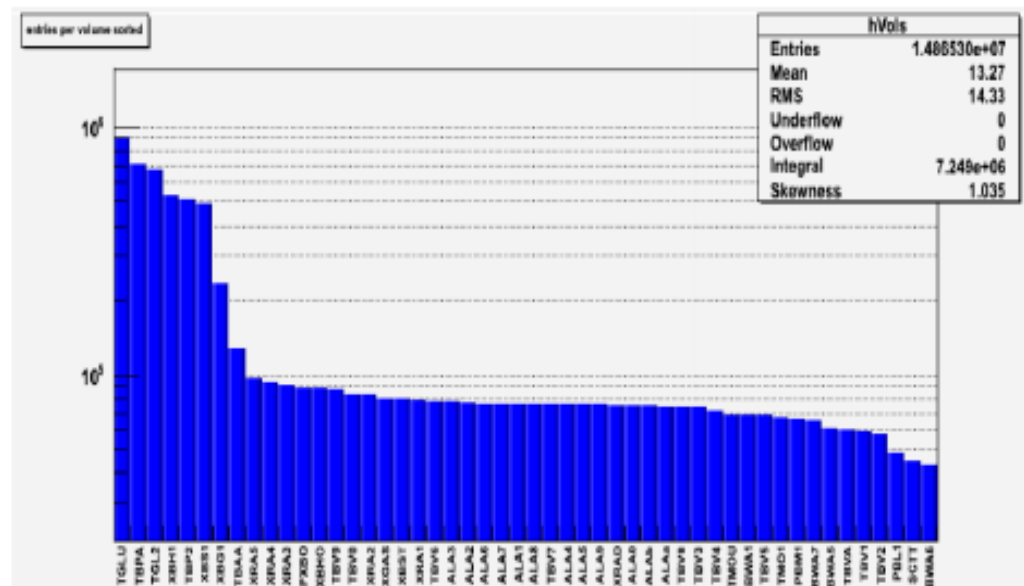
(2012 J. Phys.: Conf. Ser. 396 022014)

- ▶ Geant-V **Vector** Prototype: a részecsketranszport szimulációk új generációja

(2014 J. Phys.: Conf. Ser. 513 052006,

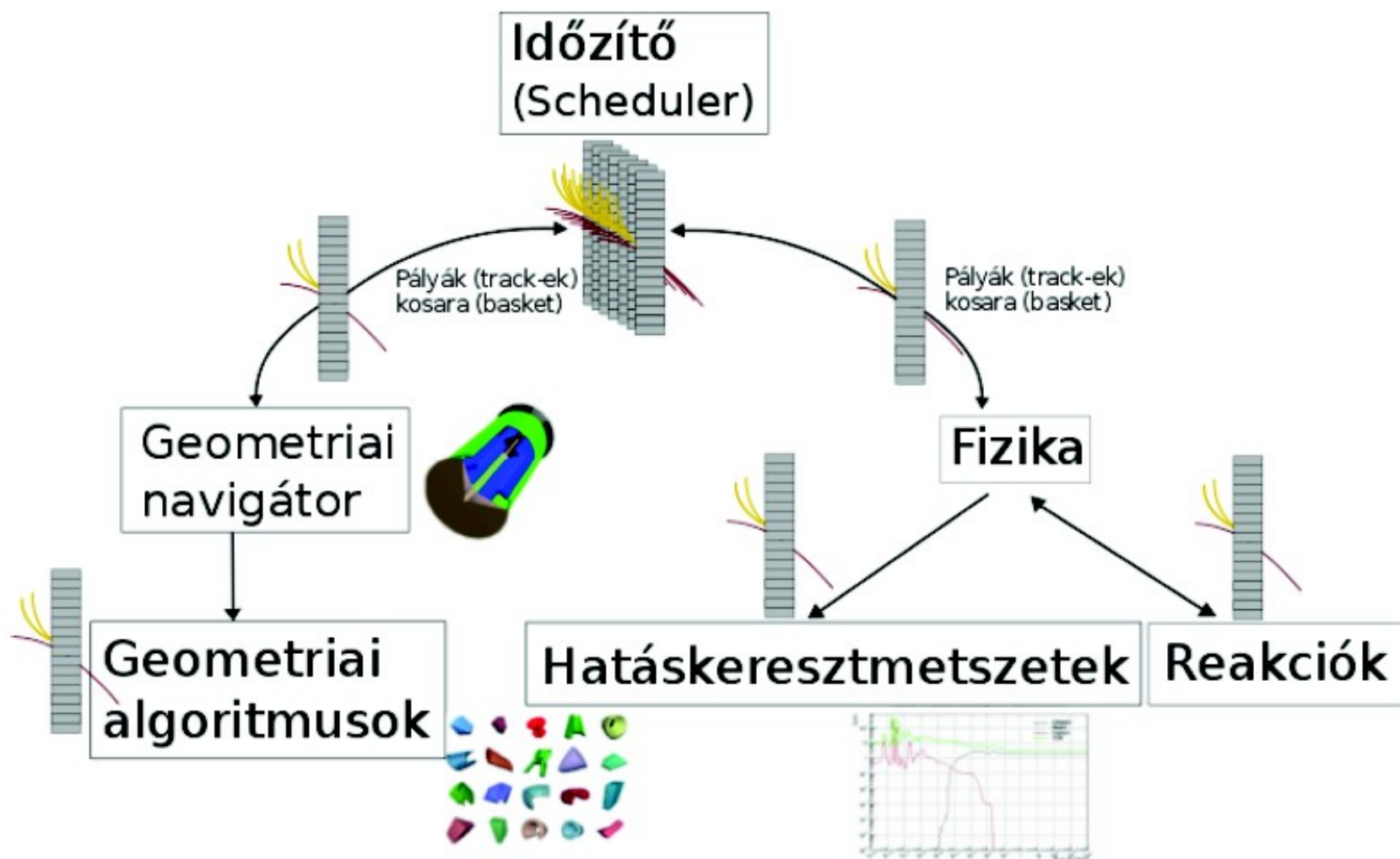
2014 J. Conf. Phys.: Conf Ser. 052013)

- ▶ **Elsődleges** célok: a hardver nyújtotta lehetőségek teljes kiaknázása a hordozhatóság és karbantarthatóság feláldozása nélkül



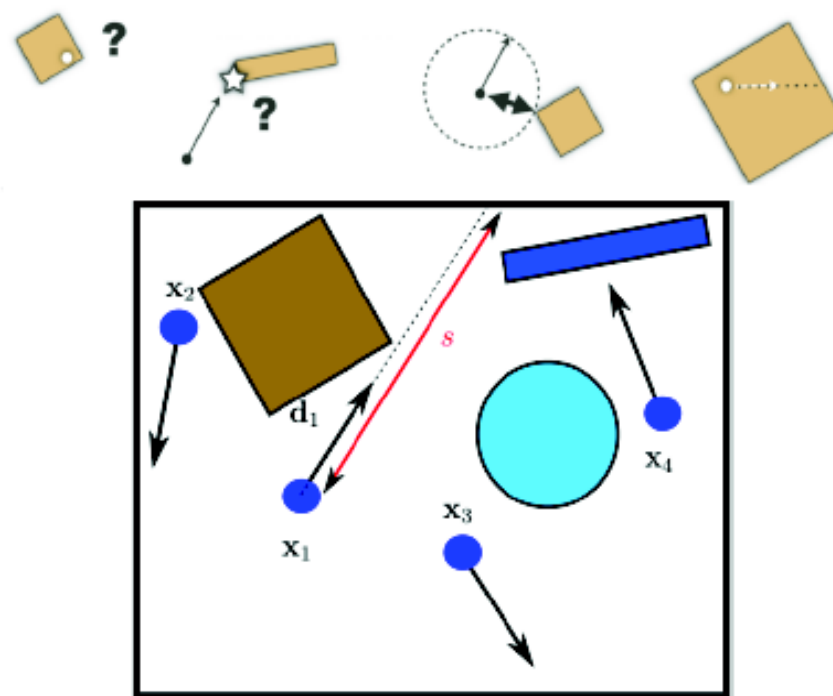
- ▶ Amdahl-törvény és Moore-törvény: a **becsült** sebességnövekedés mértéke: $\mathcal{O}(10)$

Geometriai algoritmusok tömbösítése



Geometriai algoritmusok tömbösítése

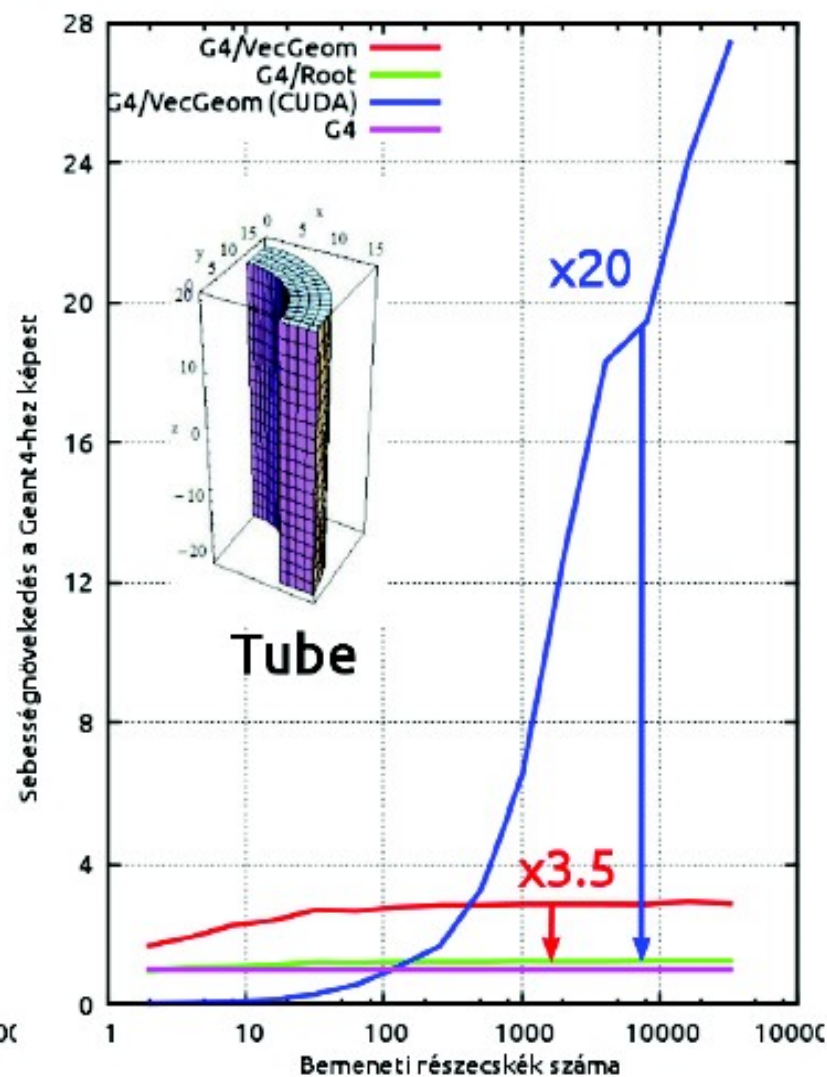
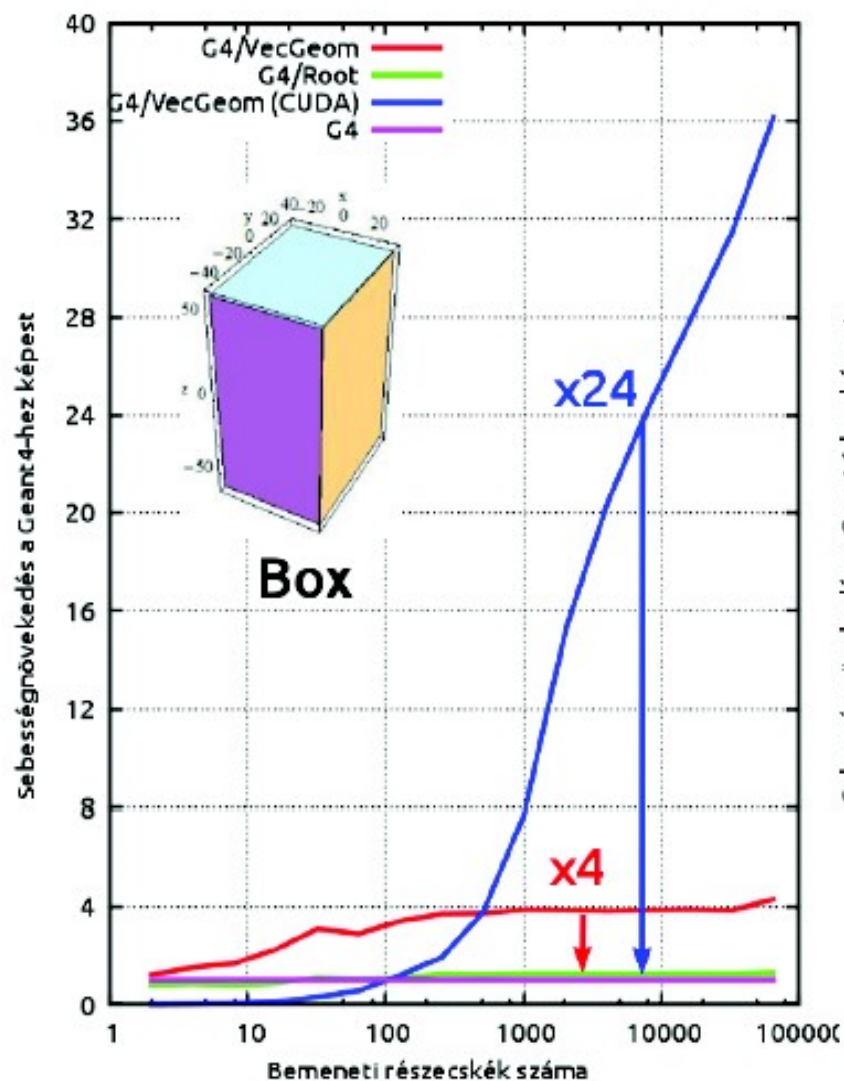
- ▶ **VecGeom:** a Geant-V Vektorizált Geometrája
- ▶ Vektorizált **geometriai** számolások: a részecskék geometriai tulajdonságainak kiszámolása SIMD műveletekkel
- ▶ Nagyfokú **hordozhatóság** backend osztályok segítségével
- ▶ **Egyetlen** kód a skalár, vektor és GPU módban való futtatáshoz



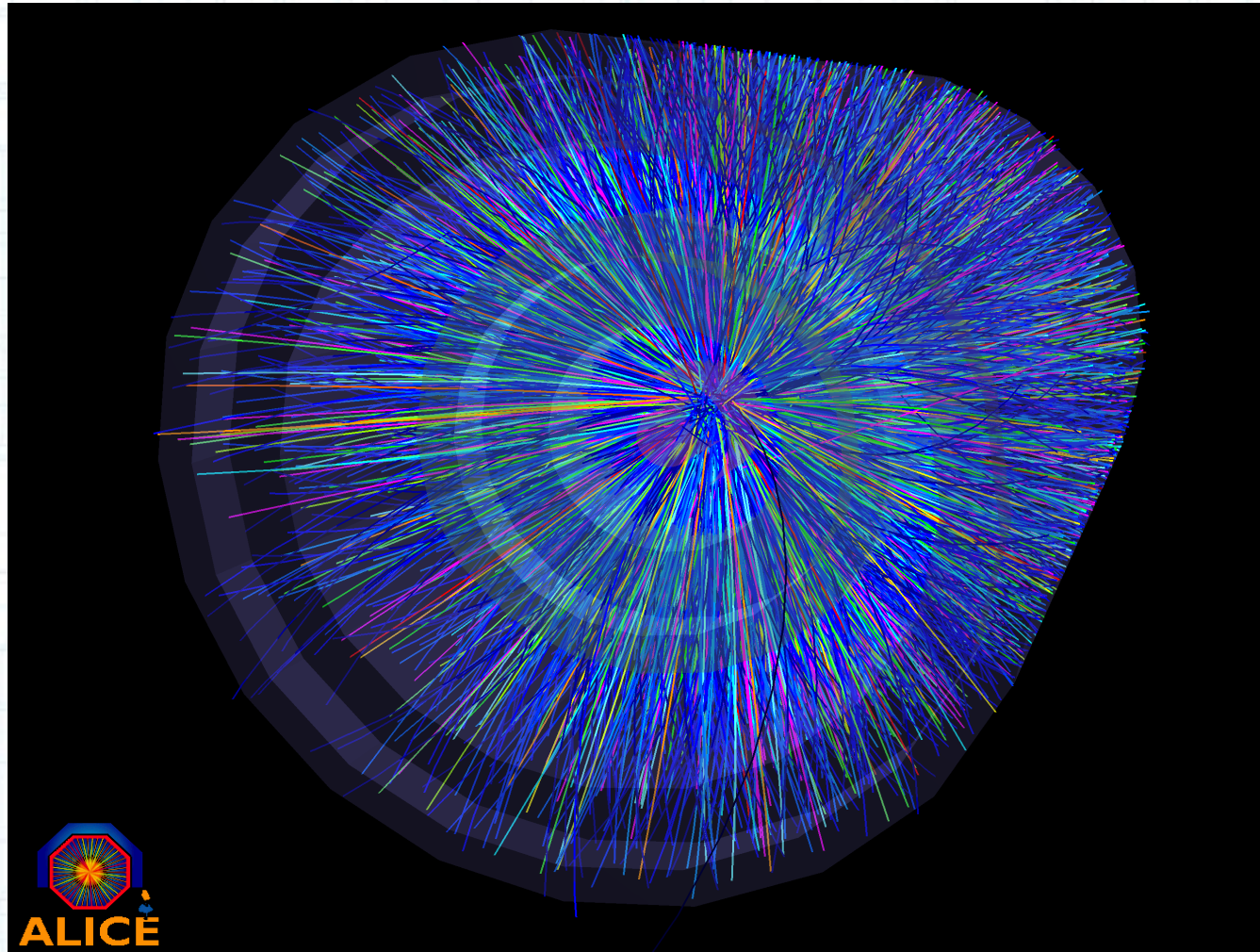
- ▶ **SIMD** műveletek: Vc vektor könyvtár használata (<http://code.compeng.uni-frankfurt.de/projects/vc>)
- ▶ **GPU:** C++ sablonok használatával, generikus programozás révén
 - > **NVidia** CUDA támogatás
 - > **OpenCL** támogatás: folyamatban...

Geometriai algoritmusok tömbösítése

Távolság a határig, 5000 ismételés, Intel® Core™ i7-920 Processor + NVidia Tesla C2050



Egy másik lehetőség: GPU@MC



Mikor jön el a „GPU-pillanat”?

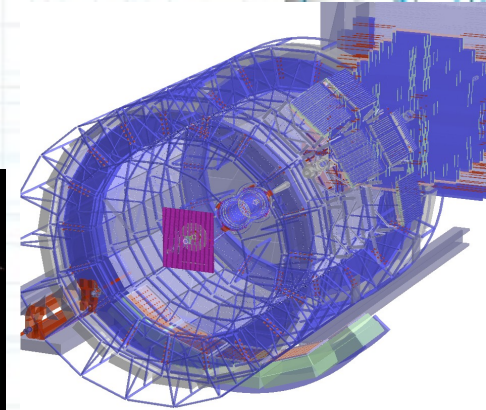
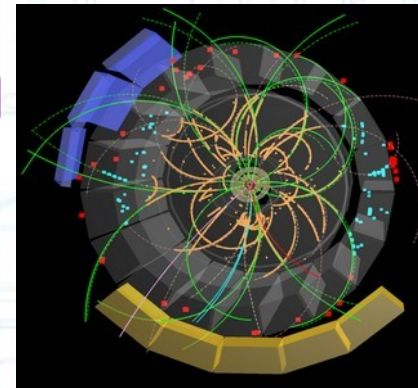
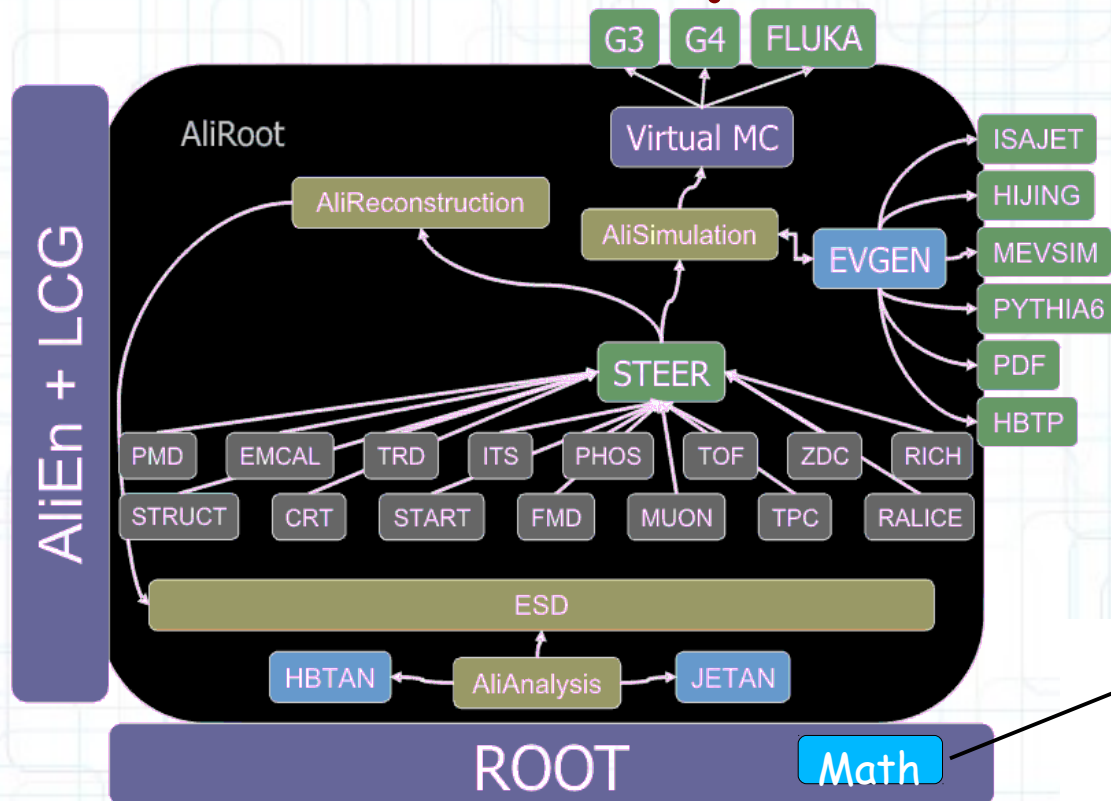
Nincs direkt válasz!

- Pilótatnulmány: opimalizáció
- Szükség van nagyon nagy skálájú számításokra
- 10x több idő kell programozásra, fejlesztésre
- Magas szintű hardverközeli programozás
- \$\$\$\$\$\$

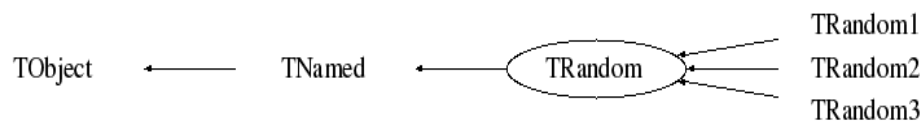
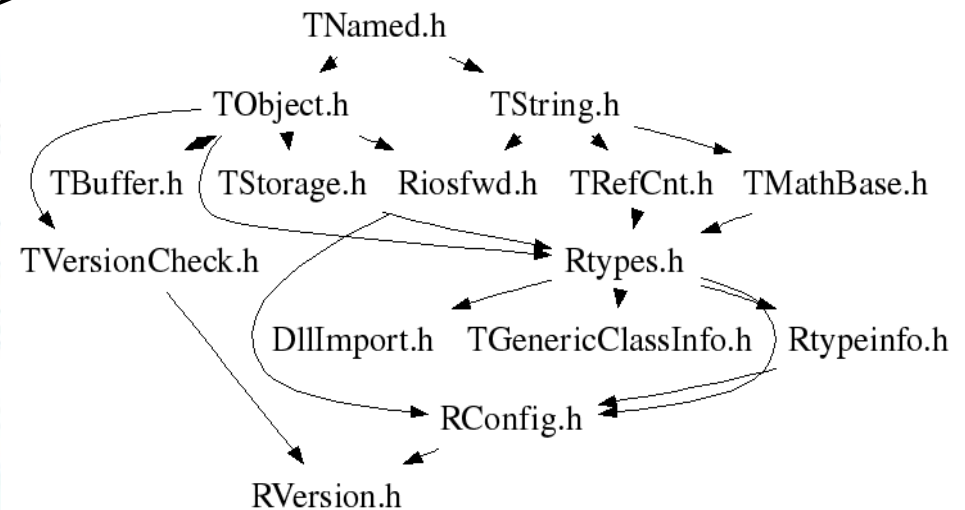
Mit találunk a „piacon”? – (komolyabb CUDA, stb... tudás nélkül)

- Libek & toolkek (BLAS, FFTW, CUBLAS, CUFFT)
- Wrapperek (C, FORTRAN → CUDA)
- OpenCL standardok (Ati, NVidia)
- Mathematica, MatLab (GPU támogatással)

GPU alapú PRNG MC



./math/mathcore/inc/TRandom.h

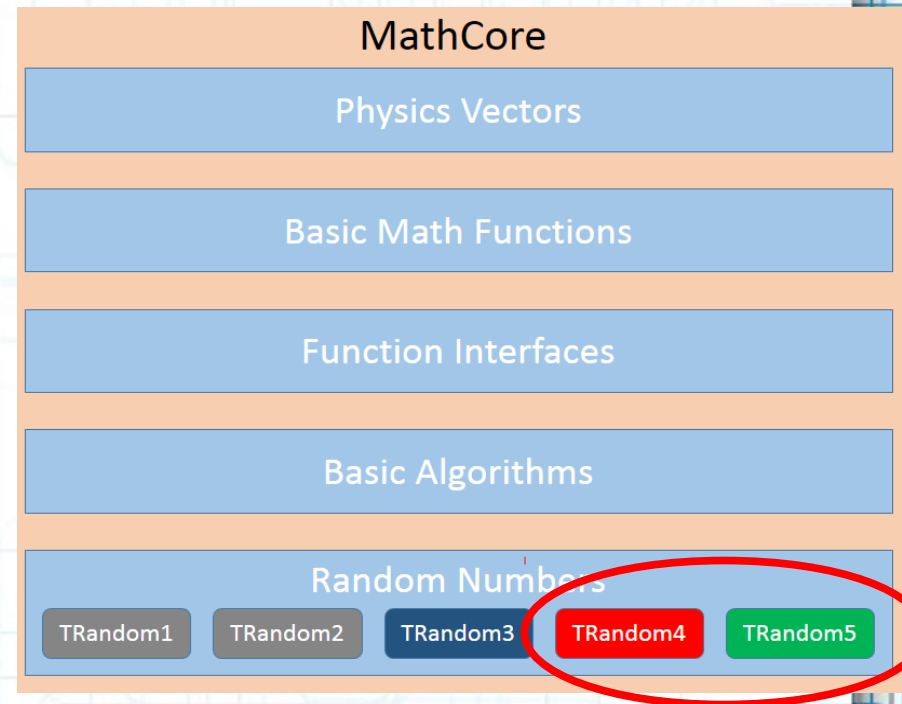


AliRoot az ALICE
adatszimulációs és
rekonstrukciós, elemzője

GPU alapú PRNG MC

Tesztelt PRNG kódok

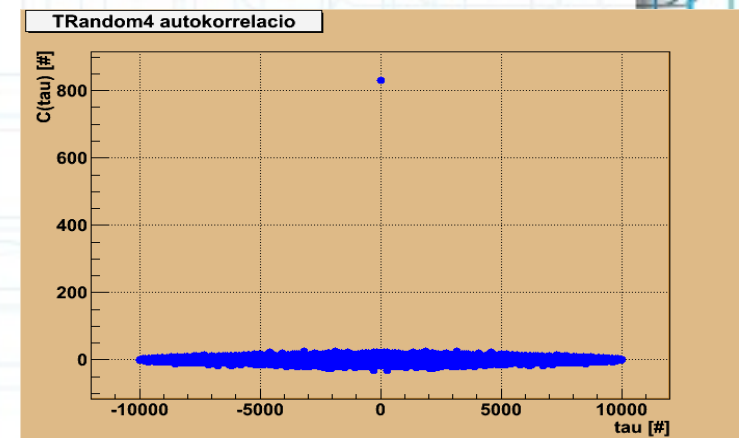
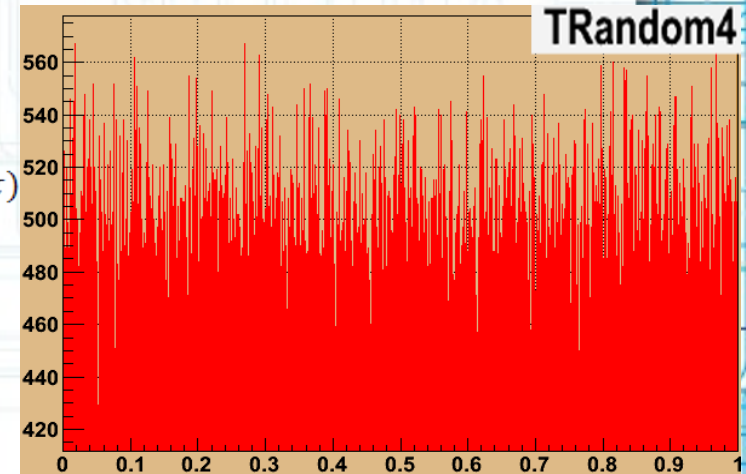
- Trandom1 (RANLUX)
- TRandom2 (Tausworthe)
- Trandom3: Original CPU based Mersenne Twister) algorithm
- Trandom4: CPU/GPU based SFMT (SIMD-oriented Fast Mersenne Twister) algorithm
- Trandom5: CPU/GPU based MWC64X algorithm



A PRNG minősége is számít ám...

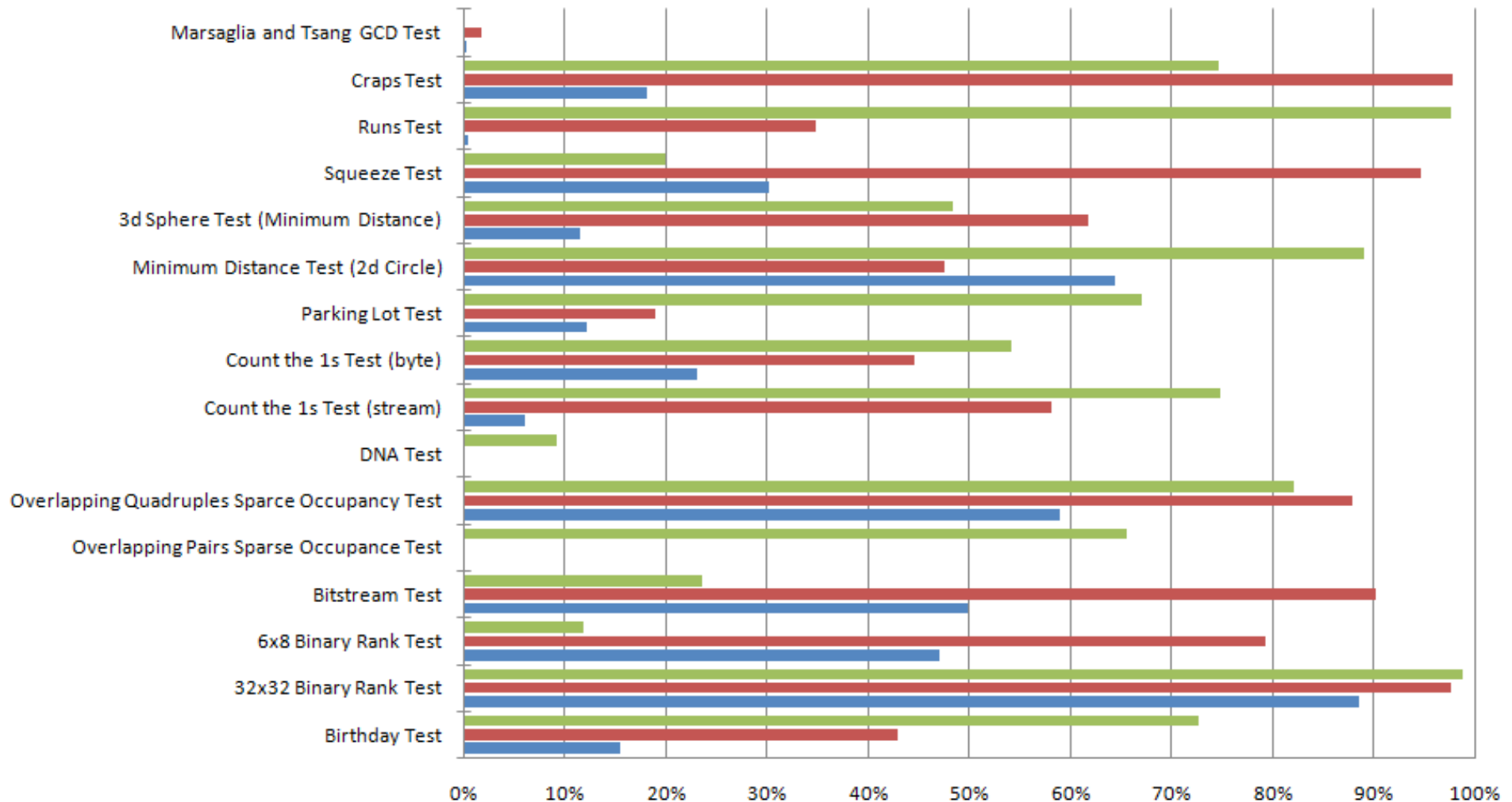
Milyen jó lehet egy PRNG?

- 1) Egyszerű teszt momentumok vizsgálata
- 2) teszt: autokorreláció $C(\tau) = \sum_n f(n)f(n+\tau)$
- 3) PRNG komplex teszt: „Diehard test”:
 - R.G. Brown, D. Eddelbüttel, D. Bauer: Diehard 3.31.1 a Kolmogorov-Smirov teszten alapuló nyílt forráskódú csomag (G. Marsaglia alapján)



A PRNG minősége is számít ám

Results of Dieharder Tests



TRandom3 TRandom4 TRandom5

A PRNG minősége is számít ám

- A PRNG DieHard minőségtesztje

TRandom3 - Original CPU based Mersenne Twister

- TRandom4 - CPU/GPU based SFMT (SIMD-oriented Fast MT)

TRandom5 - CPU/GPU based MWC64X algorithm

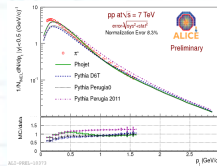
PRNG modules	Platform	Total Kuiper KS p
TRandom3	CPU	29.27 %
TRandom4	CPU/GPU	53.59 %
TRandom5	CPU/GPU	55.56 %

Performance

Egy kis fizikai: proton-proton ütközés

- 400k TRandom5 PRNG
- Transverzális impulzus spektrum

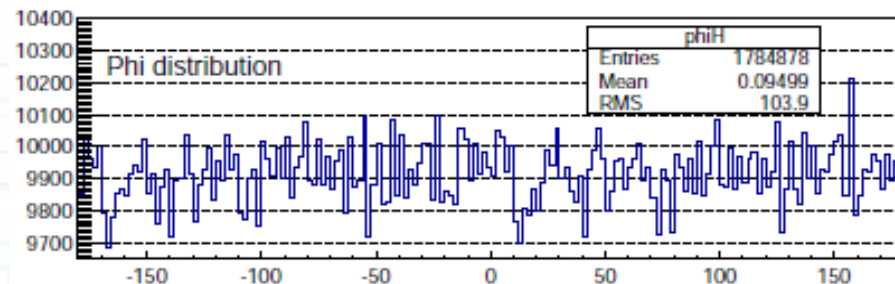
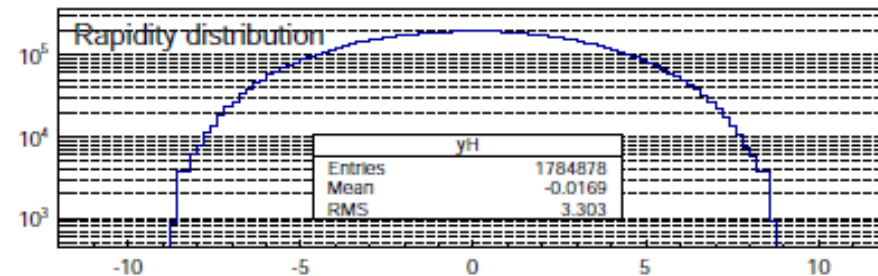
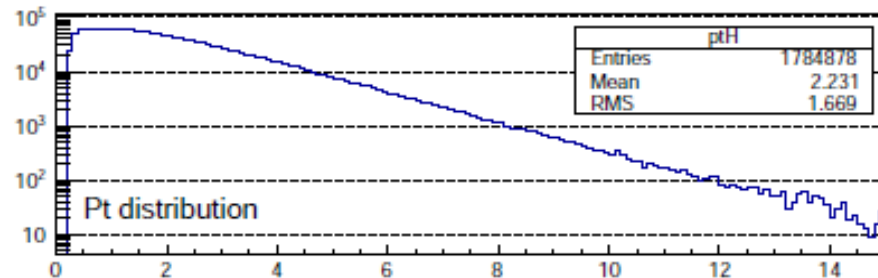
dN/dp_T (Tsallis)



- Rapidity-eloszlás
- dN/dy (Gauss)

Szögeloszlás

$dN/d\phi$ (Izotrópia)

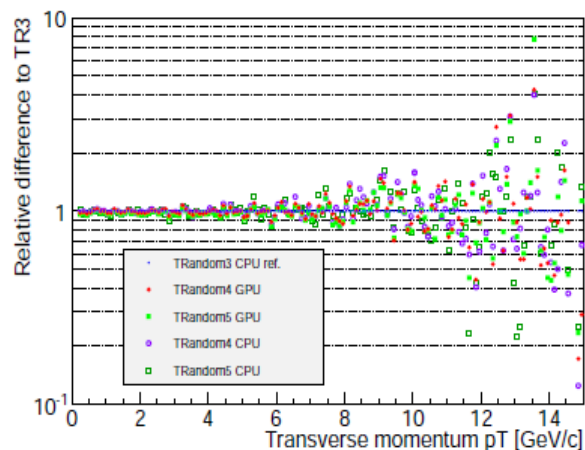


Egy kis fizikai: proton-proton ütközés

- Teszteljük a 'fizikát':

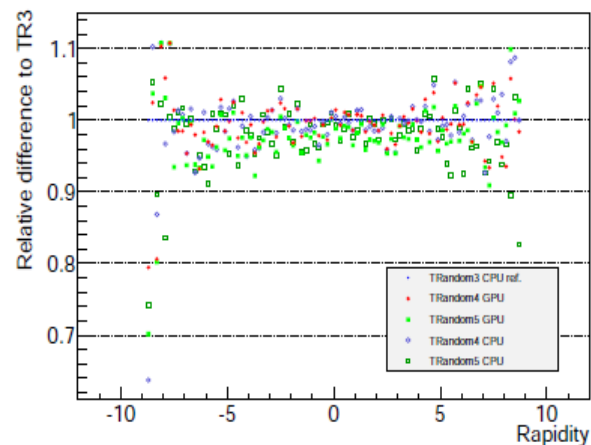
Új/régi Trandom Számolások összehasonlítása

TRandomX/TRandom3 aránynak ~ 1 nek kellene lennie



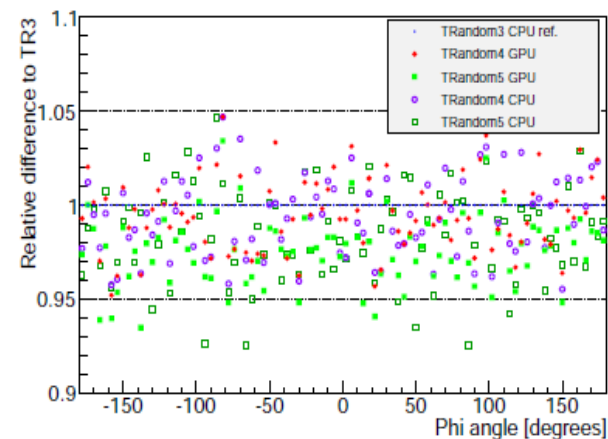
10% egyezés

$p_T < 6$ GeV/c



5% egyezés

$|y| < 5$



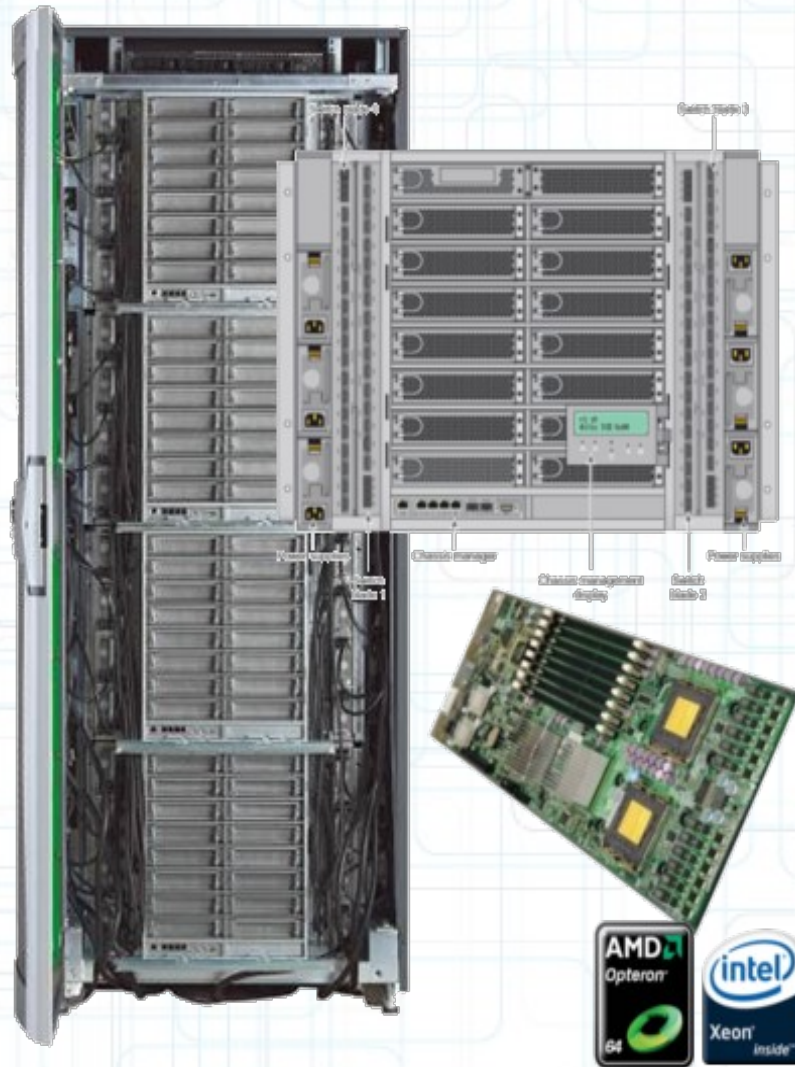
5% egyezés

teljes ϕ tartomány

MTA Wigner FK erőforrások

- **GPU Labor** (Nagy-Egri M. F.)
 - <http://gpu.wigner.mta.hu>
- **Wigner Felhő** (Genagrid, Harangozó Sz., Bíró G.)
 - Genagrid: <http://genagrid.hu>
 - Wigner Felhő (Wigner Cloud):
 - 2015 márciusában indul
 - 800 core, 1-2 PB storage
- **Wigner Adatközpont WDC**
 - CERN Tier0 site
 - 10K (20k) core, 2020-ig 73k core

GenaGrid: Hardware

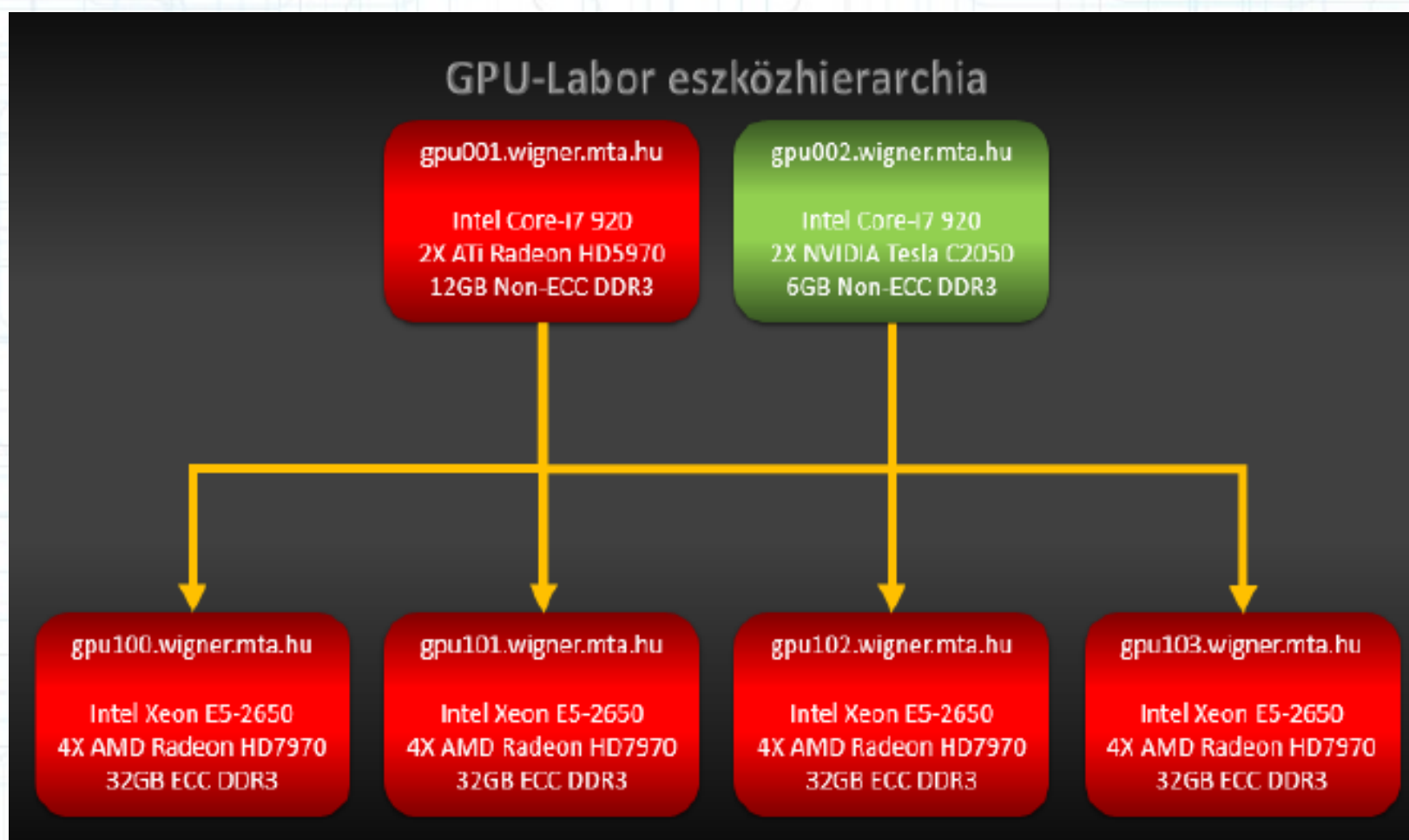


- SGI Altix ICE 8200XE
- 64 blades
- Intel Xeon X5365 3.00GHz x2 (8 core)
- 16GB DDR2 RAM
- InfiniBand

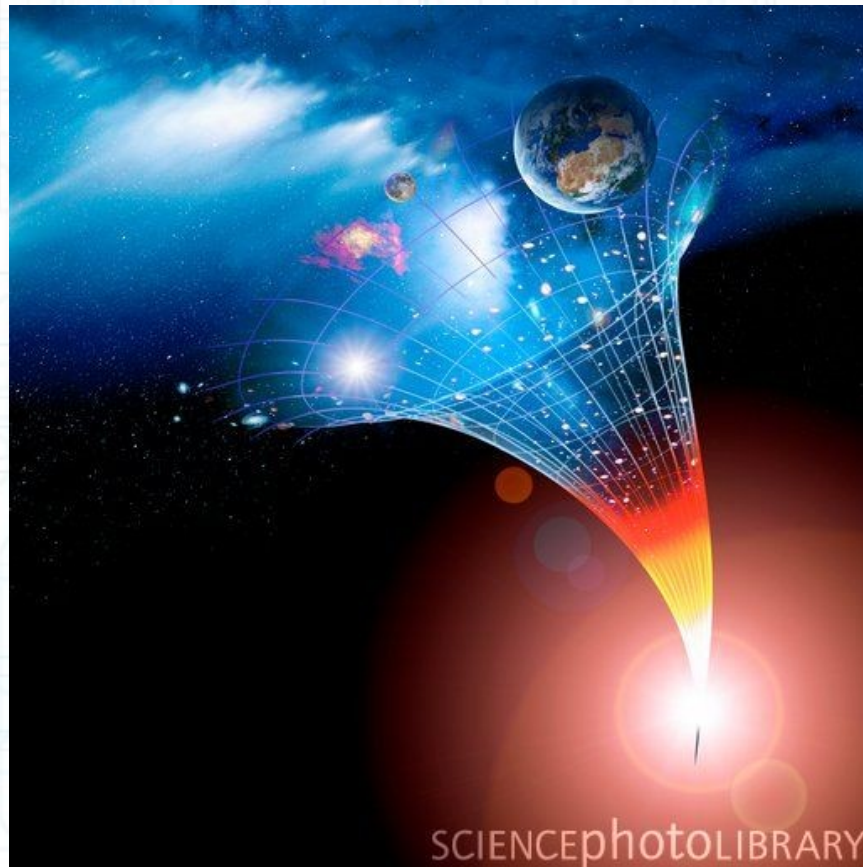
GenaGrid: Software

- Red Hat Enterprise Linux 6
- GCC, G++, GFORTAN ... compilers
- OpenMPI middleware
- AMD OpenCL SDK
- HTCondor que system

Wigner GPU Lab Hardware



Amit ma feldolgozhatsz...



... ne halaszd holnapra!