

Adatelemzés és számítástechnika

Krasznahorkay Attila

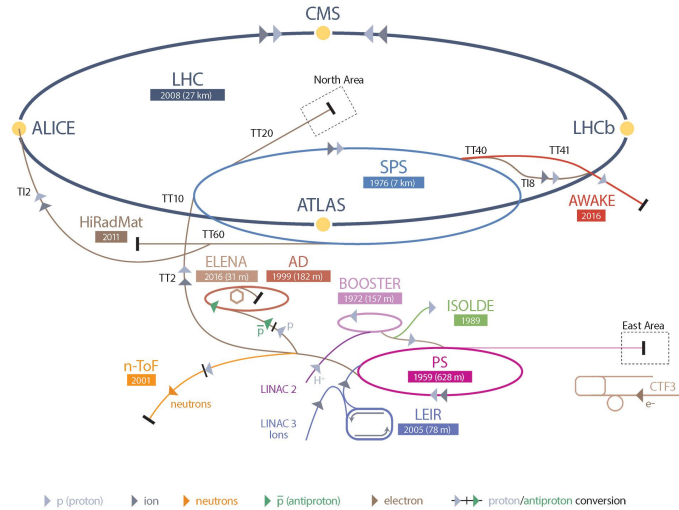


Hungarian Teacher Programme

Emlékeztető

A madártávlat

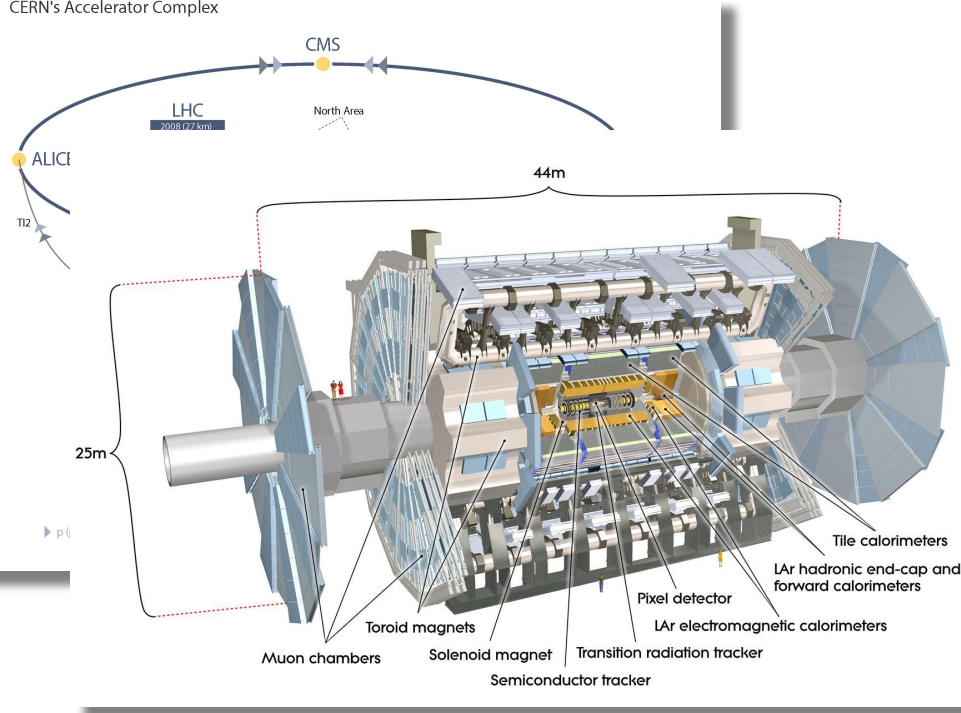
CERN's Accelerator Complex



Létrehozzuk az “érdekes” reakciókat
(Barna Dániel előadása)

A madártásvlat

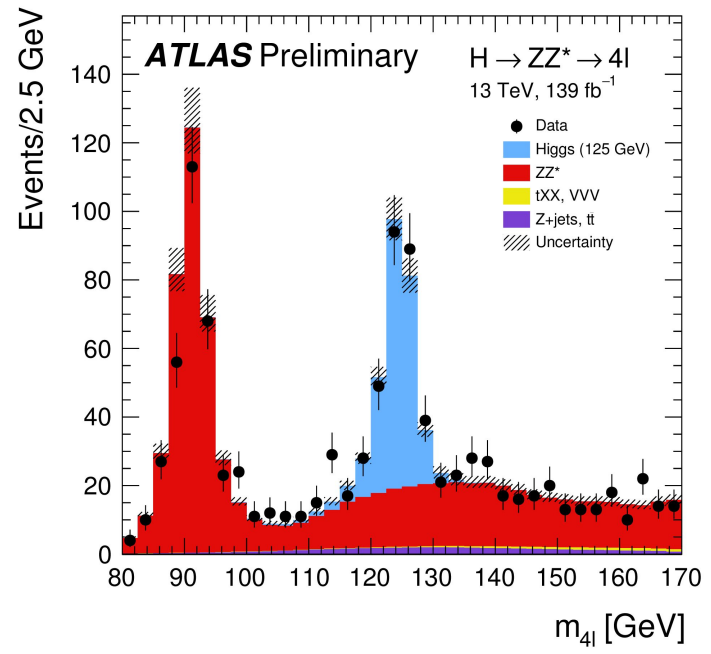
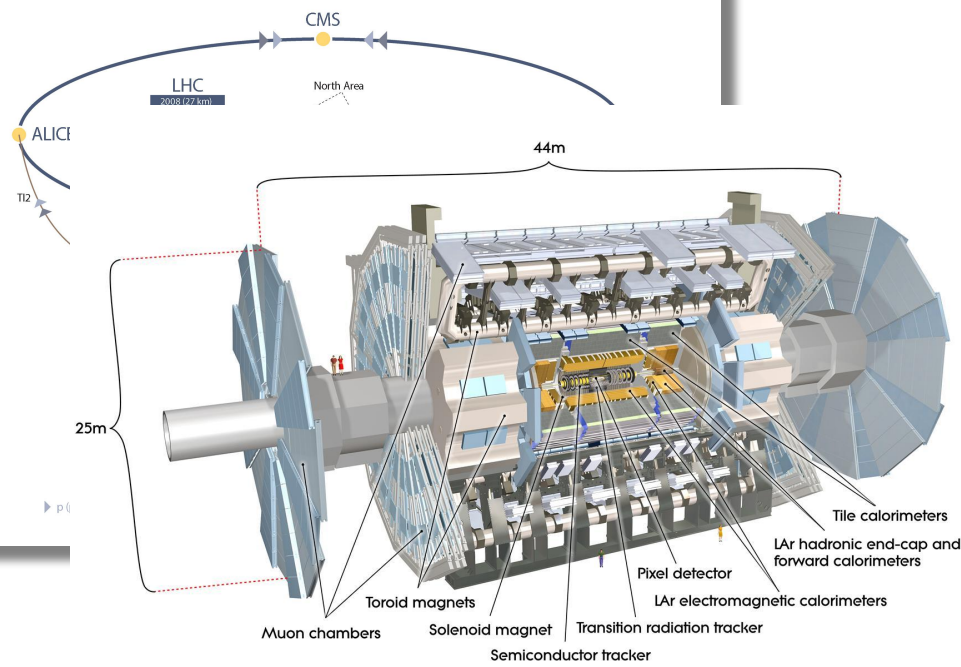
CERN's Accelerator Complex



Érzékeljük a kijövő részecskéket
(Barna Dániel előadása)

A madártávlat

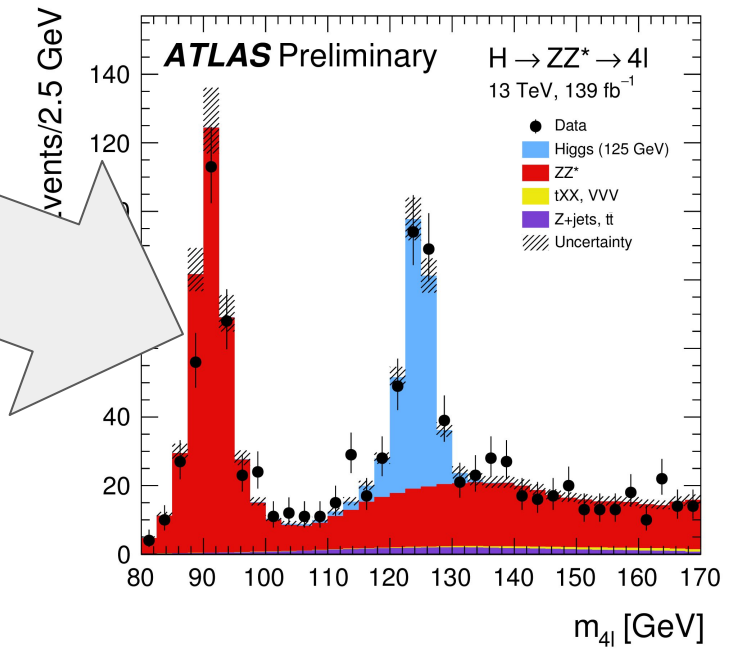
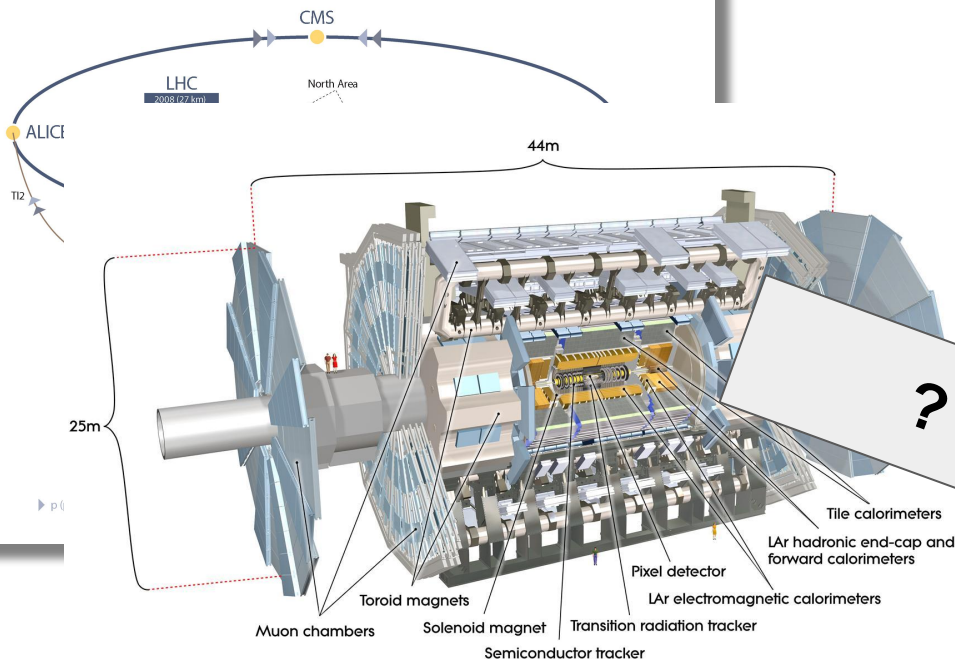
CERN's Accelerator Complex



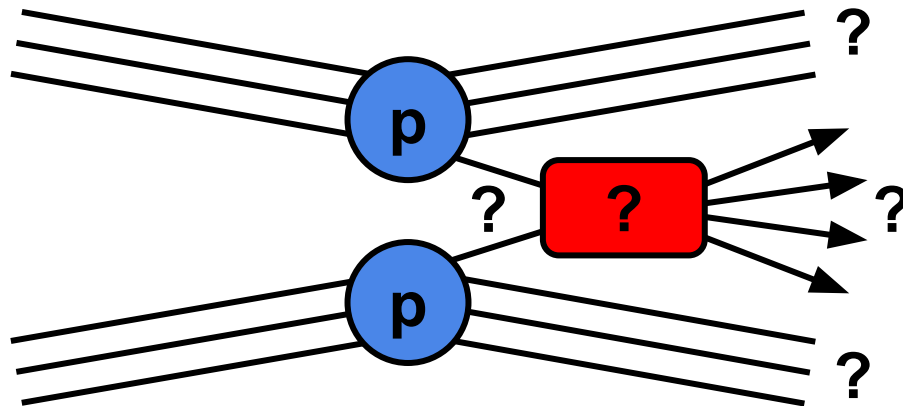
Levonjuk a fizikára vonatkozó következtetéseket
 (Újvári Balázs előadásai)

A madártávlat

CERN's Accelerator Complex



A kísérlet



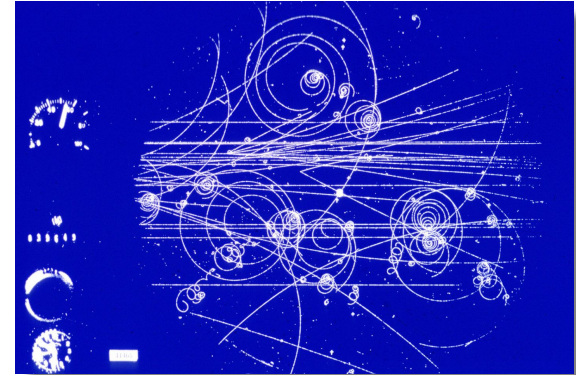
- Amire igazán kíváncsiak vagyunk az LHC-nél az a piros dobozban végbemenő folyamat
- Az alapján amit a reakcióból kijönni látunk, következtetünk vissza, hogy:
 - Milyen proton-proton ütközést figyeltünk meg
 - Milyen reakció zajlott le pontosan

- A detektorokkal egymástól független “eseményeket” rögzítünk
 - Ahogyan Dani bemutatta, minden detektor típus rá jellemző módon tudja érzékelni azt ha különböző részecskék kölcsönhatásba lépnek vele
 - A rögzített információt, mint egy fényképezőgép CCD kameráját, kiolvassuk
- Az LHC másodpercenként >1 milliárd proton-proton ütközést hoz létre, amit mindet nem rögzíthetünk
 - Kb. 40 millió proton csomag ütközés jön létre másodpercenként, mindegyikben 40-60 proton-proton ütközéssel
 - Egy válogató (trigger) rendszerrel választjuk ki a fizika szempontjából legérdekesebb eseményeket, kb. 1000 darabot másodpercenként. Ez másodpercenként néhány GB adat gyűjtéséhez vezet.
- A “nyers” adatokat a CERN számítóközpontjába küldjük, ahonnan biztonsági másolatok készülnek róla a világ több pontjára
 - Évente több milliárd eseményt rögzítünk így

- A nyers adatokat erre fejlesztett célprogramokkal “rekonstruáljuk”. Így állapítva meg, hogy milyen részecskék milyen tulajdonságokkal repültek át (és a legtöbb esetben semmisültek meg) a detektoron/detektorban
 - Megkapjuk, hogy mennyi elektron, müon, hadronzáró, stb. keletkezett az egyes eseményekben, és milyen tulajdonságokkal
- Az egységesen rekonstruált adatokat több lépésen keresztül analizáljuk, az egyes analízisek kívánalmainak megfelelően
 - Például kereshetünk két leptonos eseményeket kevés “hiányzó energiával” ha Z bozonos eseményeket akarunk vizsgálni
 - A kiválasztott eseményekben egyenként kiszámíthatunk különböző komplex paramétereket (például az eseményben feltételezett Higgs bozon tömegét), amiknek az eloszlásából statisztikai analízissel a háttérben zajló fizikára tudunk visszakövetkeztetni.

Történelmi kontextus

- **Érdeemes látni, hogy hogyan is jutottunk idáig**
 - A jelenlegi komplex adatfeldolgozás nem egyik percről a másikra lett kitalálva...
- **Milyenek voltak a kísérletek a múltban?**
 - **Legkorábban:** Fényképek és besugárzott minták szemmel tanulmányozása
 - Gáz és buborékkamra képeket ellenőriztek egyenként, betanított munkával
 - **Később:** Kevés paraméter (pl. energialeadás, pozíció) mérése, és közvetlen analizálása
 - Képesek voltak a különböző detektorokból kijövő elektromos jeleket olyan formára hozni, hogy azokból a detektor által érzékelt részecskék milyenségére tudjanak következtetni
 - De a jeleket lehetetlen volt egyesével rögzíteni, a kor számítástechnikája ezt nem engedte meg
 - [Sokcsatornás analizátorokkal](#) elektronikusan készítették eloszlásokat a mért paraméterekből
 - **Végül:** Amint a mért elektronikus paramétereket egyenként rögzíteni lehetett, az adatok feldolgozása a kísérlet elvégzésétől későbbre tolódhatott
 - Lehetővé téve, hogy kifinomult adatfeldolgozással pontosabb méréseket lehessen elvégezni



- A CERN első “valódi” számítógépe 1958-ban állt üzembe
 - Programokat és adatokat lyukszalagon tudott kapni, és természetesen mai szemmel nagyon lassú volt
- Az évek folyamán egyre nagyobb és nagyobb teljesítményű szinguláris gépeket helyeztek üzembe
 - <https://cerncourier.com/a/computing-at-cern-the-mainframe-era>
- A nagy változást a személyi számítógépek megjelenése hozta
 - Ezeket használni lehetett az adatgyűjtés irányításától egészen az adatfeldolgozás végső lépéséig
 - Azóta is a személyi számítógépek uralják a fizika kísérleteket
 - Bár a “szuperszámítógépek” mostanra újra népszerűek lettek



A grid létrejötte

- A személyi számítógépek előretörésével hamar megjelentek az ezekből épített klaszterek
 - Ezek arra lettek kitalálva, hogy ugyanazt a feladatot, egyenként kis változtatásokkal, sok gépen/processzoron tudják egyszerre futtatni
 - A LEP adatait is ilyen klasztereken dolgozták fel annak idején
- Az LHC tervezésénél hamar világossá vált, hogy nem célszerű egy nagy klaszterre bízni az adatfeldolgozást → A világ sok egyéni klaszterét kell inkább összekötni egy egységes rendszerré
 - Amire úgy lehet feladatokat küldeni, hogy nem kell tudjuk az adatok hol vannak, és a programunk pontosan hol fog lefutni



Modern adatfeldolgozás

Az ATLAS adatgyűjtése



Vistars

op-webtools.web.cern.ch/vistar/vistars.php?usr=LHC1

ATOMKI WW... CERN Users' p... ATLAS Repositories ATLAS Nightli... Other bookmarks

LHC Page 1 Vistar

LHC Page1 Fill: 8136 E: 6800 GeV t(SB): 01:04:23 17-08-22 08:38:05

PROTON PHYSICS: STABLE BEAMS

Energy: 6800 GeV I B1: 2.03e+14 I B2: 2.03e+14

Beta* IP1: 0.30 m Beta* IP2: 10.00 m Beta* IP5: 0.30 m Beta* IP8: 2.00 m

Inst. Lumi [(ub.s)^-1] IP1: 13958.80 IP2: 8.34 IP5: 13560.06 IP8: 327.24

FBCT Intensity and Beam Energy Updated: 08:38:04

Instantaneous Luminosity Updated: 08:38:03

Comments (17-Aug-2022 07:34:15)
STABLE BEAMS with 1935b (192bpi)
IP2 and IP8 sep. levelling
IP1 and IP5 B* levelling
XRPs IN

BIS status and SMP flags

	B1	B2
Link Status of Beam Permits	true	true
Global Beam Permit	true	true
Setup Beam	false	false
Beam Presence	true	true
Moveable Devices Allowed In	true	true
Stable Beams	true	true

AFS: 25ns_1935b_1922_1602_1672_192bpi_14inj_3INDIV:PM Status B1 **ENABLED** PM Status B2 **ENABLED**

Az ATLAS adatgyűjtése



LHC Page 1

LHC Page1 **Fill: 8136** **E:**

PROTON PHYS

Energy: **6800 GeV** **IB1**

Beta* IP1: **0.30 m** **Beta* IP2:** **10**

Inst. Lumi [(ub.s)⁻¹] **IP1: 13958**

FBCT Intensity and Beam Energy Updated: 0

Comments (17-Aug-2022 07:34:15)
 STABLE BEAMS with 1935b (192bpi)
 IP2 and IP8 sep. levelling
 IP1 and IP5 B* levelling
 XRPs IN

AFS: 25ns_1935b_1922_1602_1672_192bpi_14inj

ATLAS Detector Systems ATLAS Operation Systems

Inner Detector

- PIX
Pixel Detector
- SCT 1
Semiconductor Tracker
- TRT 1
Transition Radiation Tracker
- IDE 13
Inner Detector

Calorimeters

- LAR
Liquid Argon Calorimeter
- TIL
Tile Calorimeter

Muon Spectrometer

- MDT
Monitored Drift Tubes

ATLAS DETECTOR CONTROL

17-08-22 08:33:57

Inner Detector	
BARREL B LAYER DISKS	OK
INF	OK
BARREL ENDCAP A	OK
ENDCAP C	OK
INF	OK
BARREL A	OK
BARREL C	OK
ENDCAP A	OK
ENDCAP C	OK
INF	OK
EMBA	OK
EMBC	OK
EMECC	OK
HEFCAL A	OK
HEFCAL C	OK
INF	OK
LBA	OK
LBC	OK
EBA	OK
ERC	OK
INF	OK
BARREL A	OK
BARREL C	OK
ENDCAP A	OK
ENDCAP C	OK
INF	OK
RPC SIDE A	OK
RPC SIDE C	OK
RPC INF	OK
RPC-BIS78	OK
RPC INF	OK
TGC SIDE A	OK
TGC SIDE C	OK
TGC INF	OK
MMG SIDE A	OK
MMG SIDE C	OK
MMG INF	OK
STG SIDE A	OK
STG SIDE C	OK

FMD

Inner Detector

Calorimeter

Muon Spectrometer

LHC

STABLE BEAMS

Stable	Physics
Beam	Run# 431493
Standby	N
N	1905.1 10 ¹¹
I	1905.1 10 ¹¹
E	6799 GeV
L	8472.2 10 ¹⁰
ATLAS-Q23	1.023

RUNNING

Type	Physics
Run#	431493
LB#	309
Physics	TRUE
7720	A
2020	A
5693	Hz

Az ATLAS adatgyűjtése



LHC Page1 Fill: 8136

PROTON PHYS

Energy: 6800 GeV

Beta* IP1: 0.30 m Beta* IP2: 10

Inst. Lumi [(ub.s)^-1] IP1: 13958

Comments (17-Aug-2022 07:34:15)
 STABLE BEAMS with 1935b (192bpi)
 IP2 and IP8 sep. levelling
 IP1 and IP5 B* levelling
 XRPs IN

AFS: 25ns_1935b_1922_1602_1672_192bpi_14inj

ATLAS Detector Systems

Inner Detector

- PIX Pixel Detector
- SCT Semiconductor Tracker
- TRT Transition Radiation Tracker
- IDE Inner Detector

Calorimeters

- LAR Liquid Argon Calorimeter
- TIL Tile Calorimeter

Muon Spectrometer

- MDT Monitored Drift Tubes

ATLAS Detector Operation

ATLAS ONLINE LUMINOSITY CONTROL

ATLAS Data Summary

2022 - pp

Peak Luminosity by Fill

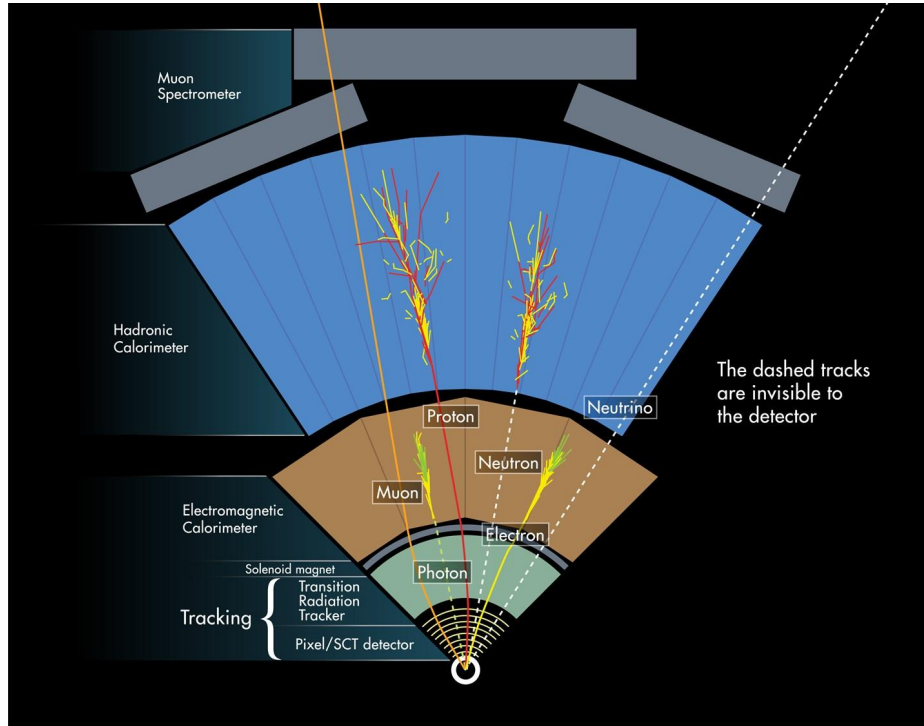
Luminosity by Day

Total Luminosity

Efficiency by Day

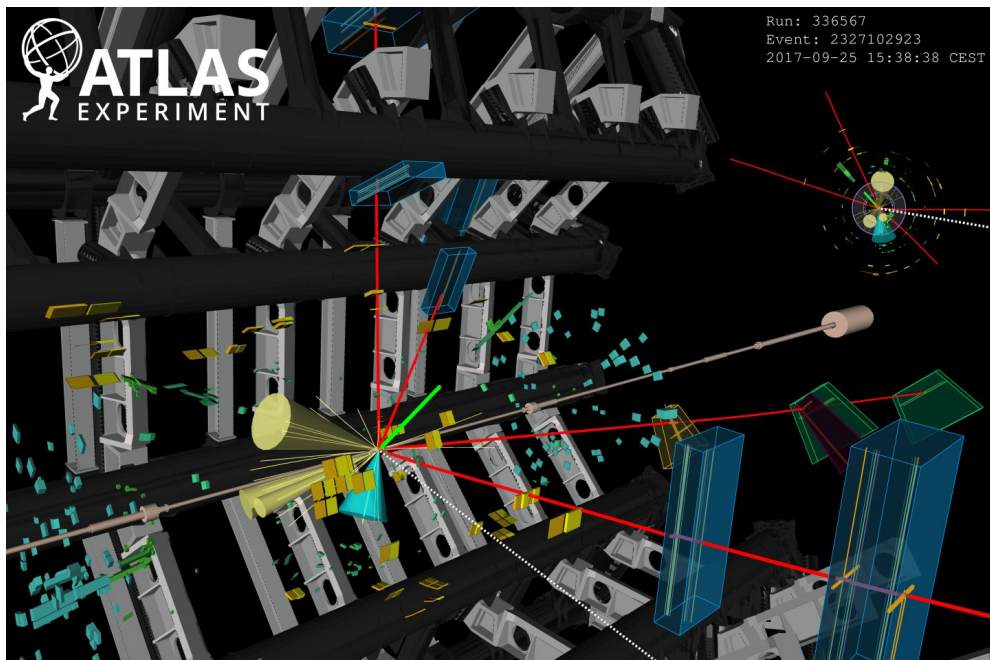
Day	Summary	Luminosity	LHC Status
Wednesday 17 Aug	Peak Stable Lumi: $1.32 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ Peak <Events>/BX: 48.8 Avg <Events>/BX: 39.0 Lumi (pb): 24.81 (100.0%) Physics Beams Del.: 24.81 (100.0%) ATLAS Ready Del.: 18.75 (75.6%) ATLAS Ready Rec.: 18.41 (74.2%) Del. after Warmstop: 0.0 (0.0%)		
Tuesday 16 Aug	Peak Stable Lumi: $1.72 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ Peak <Events>/BX: 50.9 Avg <Events>/BX: 48.9 Lumi (pb): 50.9 (100.0%) Physics Beams Del.: 50.9 (100.0%) ATLAS Ready Del.: 18.75 (75.6%) ATLAS Ready Rec.: 18.41 (74.2%) Del. after Warmstop: 0.0 (0.0%)		

- A detektorokat úgy építjük meg, hogy a nekünk érdekes részecskék egymástól különböző jeleket hagyjanak bennük
- Pl. egy elektron rekonstrukciójához:
 - Keresnünk kell egy töltött részecske nyomát a nyomkövető detektorban
 - Ehhez hozzá kell tudnunk rendelni egy energia-klasztert az elektromágneses kaloriméterben
 - A klaszter és töltött nyom tulajdonságaira sok technikai feltételt is szabunk
 - Nem szabad “aktivitást” találnunk a hadron kaloriméterben az elektron “mögött”



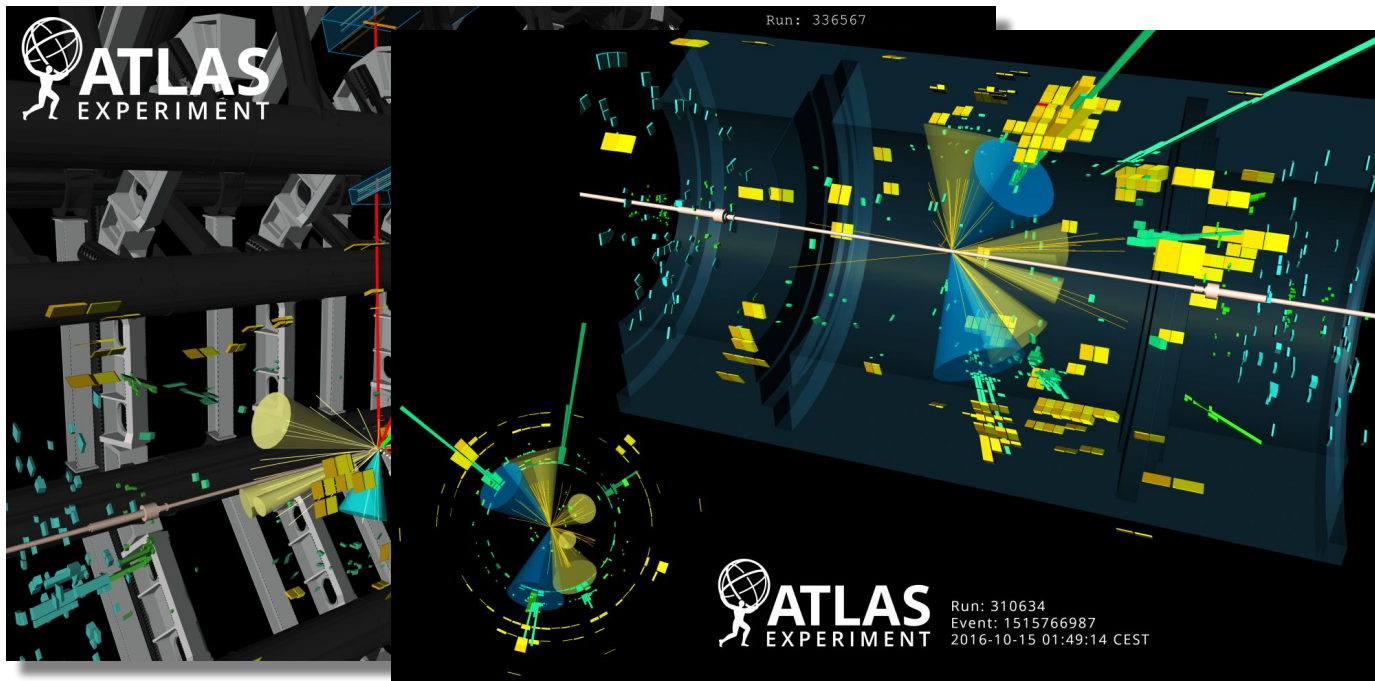
Trigger

- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki



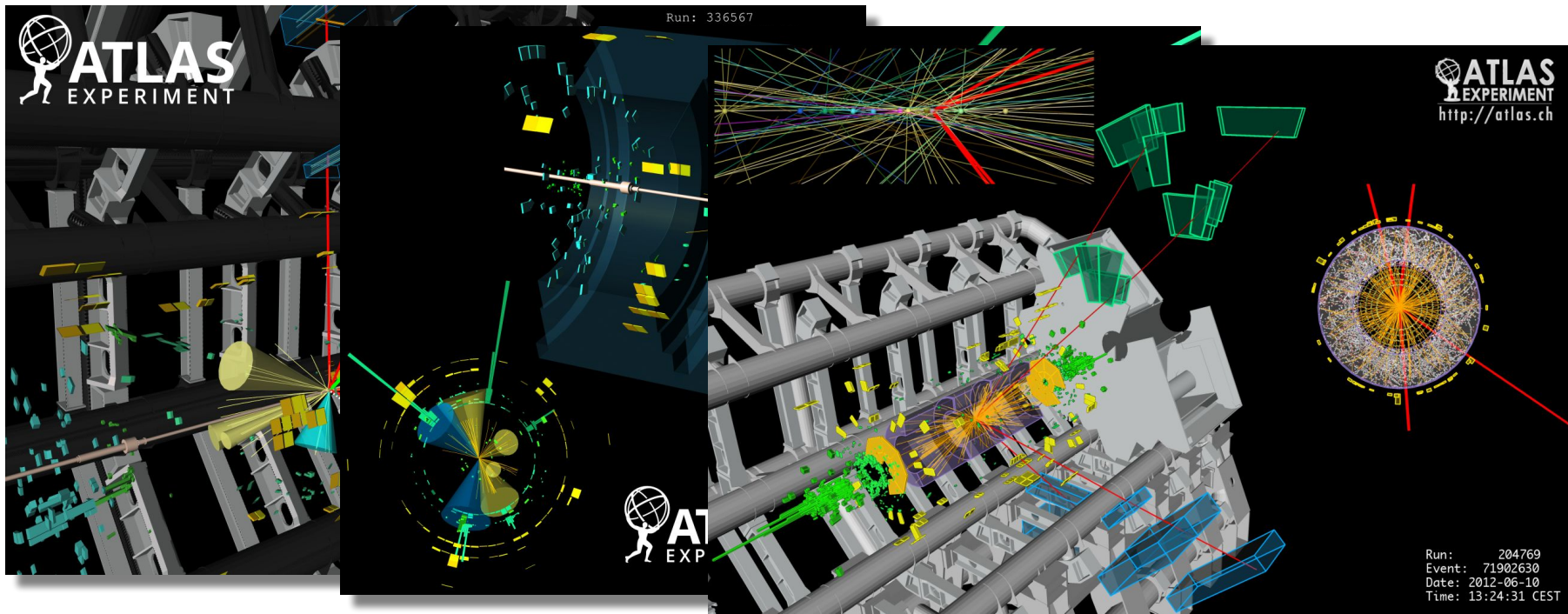
Trigger

- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki



Trigger

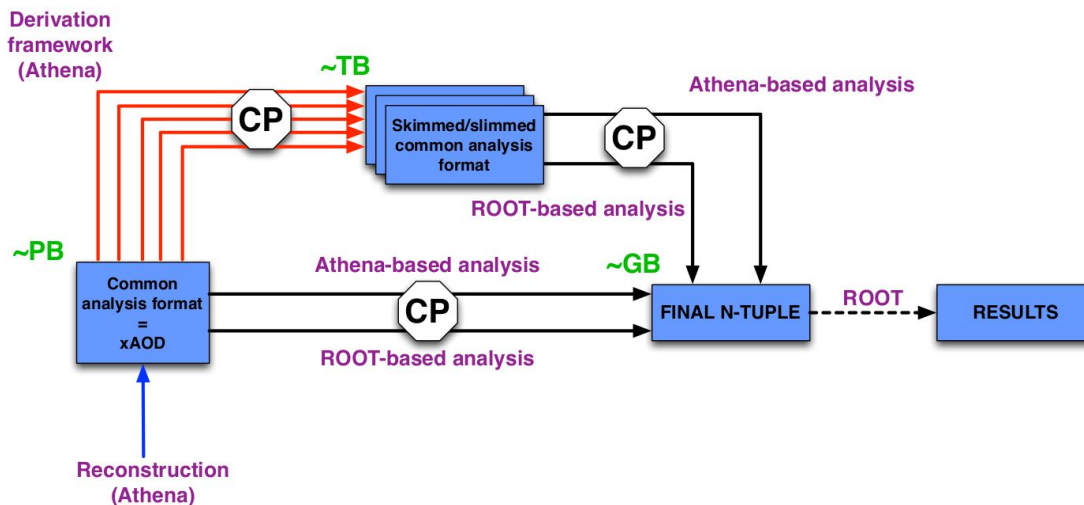
- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki



- A rögzített adatokat a legtöbb esetben csak szimulációk segítségével tudjuk értelmezni
 - Sokszor a legegyszerűbb fizikai folyamatok is annyira bonyolult módon hatnak kölcsön a detektorokkal, hogy csak így tudjuk megbecsülni, mit is mérünk pontosan
- Sok lépésből áll. Az érdekesek ezekből:
 - **Esemény generálás:** Elméleti fizikus kollégák írnak rá szoftvert, hogy véletlenszerű statisztikával valamilyen fizikai folyamathoz “eseményeket” gyártsunk.
 - **Detektor szimuláció:** A generált események “stabil” részecskéinek a kölcsönhatását modellezzük a teljes detektorral. Szimulálva a detektor minden hibáját, és az LHC beállításait is.
 - Itt vesszük figyelembe, hogy mostanra >60 proton-proton ütközés is történhet egyszerre egy-egy eseményben
 - **Trigger és rekonstrukció:** A szoftverünket ugyanúgy futtatjuk a szimuláció kimenetén mint a valódi, a detektorból gyűjtött adatokon

Analízis modell

- Nem engedhetjük meg, hogy minden fizikus $O(100)$ PB adatot dolgozzon fel minden alkalommal amikor változtat valamit az analízisén
 - Ehhez a Föld összes számítási kapacitása sem lenne elég
- Az analízis “elejét” szervezett formában végezzük a legtöbb kísérletben
 - Az eredetnél sokkal kisebb adatmennyiséget adva az egyes analizátoroknak, amiket már néhány nap alatt feldolgozhatnak a grid-en



- Kiválasztja, hogy milyen adatszettet fog feldolgozni
 - Egyúttal azt is, hogy milyen triggeret fog használni, ami hatékonyan őrizte meg a számára érdekes fizikai folyamat eseményeit
- Kiválasztja, hogy milyen szimulációkat kell figyelembe vennie az analízisében
 - Ez magában foglalja az analízis által keresett fizikai folyamat szimulációját, és minden más “háttér” folyamatot amire érzékeny lehet az adatfeldolgozás
- Lefuttatja az analízis szoftverét ugyanolyan beállításokkal minden adatra és szimulációra, (relatív) apró adatfájlokat létrehozva
 - Sokszor ez több lépésben történik, de végül különböző mennyiségek eloszlásait határozzák meg az analízisek a legtöbb esetben
- “Statisztikai analízist” végez a kapott eloszlásokon
 - Megméréndő valamilyen paramétert, vagy felfedezési/kizárási valószínűséget meghatározva
- Elvégzi az egészet még N alkalommal, az analízis minden szisztematikus bizonytalanságát figyelembe véve

Modern számítástechnika

Az ATLAS offline szoftvere



The screenshot displays the GitLab web interface for the ATLAS 'athena' project. The browser address bar shows the URL `gitlab.cern.ch/athena/athena`. The page header includes the GitLab logo and navigation menus for Projects, Groups, Activity, Milestones, and Snippets. The main content area shows the project details for 'athena' (Project ID: 53790), including statistics for commits (42,575), branches (31), tags (2,762), and files (1 GB). A recent commit by Edward Moyses is highlighted, with the message 'Merge branch 'psc_semicolon' into 'master''. Below this, there are buttons for adding README, CHANGELOG, CONTRIBUTING, and setting up DevOps and CI/CD. A table lists the project's subdirectories and their last commit details.

Name	Last commit	Last update
AsgExternal/Asg_Test	Delete CMT requirements files	1 year ago
AtlasGeometryCommon	Rewriting the GeoIDSvc to new config	3 weeks ago
AtlasTest	Merge branch 'make_unique.ControlTest-20190808' into 'master'	1 week ago
Build	Merge remote-tracking branch 'upstream/master' into AnalysisBaseRevival-master-20181218	7 months ago
Calorimeter	new style config of CaloCellMaking: Include tile-cell bulding, pedestal (=BCID) correction and ...	4 days ago
Commission	Merge branch 'EventSelector_CleanUp' into 'master'	10 months ago
Control	fix exception syntax in ComponentAccumulator.py	1 day ago
DataQuality	Convert MuonEDMHelperTool to Service	1 week ago
Database	Merge branch 'test-IOVDbSvc-20190815' into 'master'	5 days ago

Az ATLAS “offline” szoftvere



```
Totals grouped by language (dominant language first):
cpp:      3919711 (69.71%)
python:   1178662 (20.96%)
xml:      341767 (6.08%)
sh:       64815 (1.15%)
fortran:  58333 (1.04%)
ansic:    29040 (0.52%)
f90:      12340 (0.22%)
javascript: 6512 (0.12%)
php:      5487 (0.10%)
perl:     3571 (0.06%)
csh:      759 (0.01%)
pascal:   674 (0.01%)
java:     493 (0.01%)
awk:      343 (0.01%)
vhdl:     103 (0.00%)

Total Physical Source Lines of Code (SLOC)           = 5,622,610
Development Effort Estimate, Person-Years (Person-Months) = 1,731.67 (20,780.03)
(Basic COCOMO model, Person-Months = 2.4 * (KSLOC**1.05))
Schedule Estimate, Years (Months)                   = 9.11 (109.31)
(Basic COCOMO model, Months = 2.5 * (person-months**0.38))
Estimated Average Number of Developers (Effort/Schedule) = 190.11
Total Estimated Cost to Develop                       = $ 233,924,916
(average salary = $56,286/year, overhead = 2.40).
SLOCCount, Copyright (C) 2001-2004 David A. Wheeler
SLOCCount is Open Source Software/Free Software, licensed under the GNU GPL.
SLOCCount comes with ABSOLUTELY NO WARRANTY, and you are welcome to
redistribute it under certain conditions as specified by the GNU GPL license;
see the documentation for details.
Please credit this data as "generated using David A. Wheeler's 'SLOCCount'."
```

- Kb. 4 millió sor C++ és 1 millió sor Python
 - Óvatos becsléssel is >1000 emberi munkaév van benne
- Folyamatosan fejlesztjük, nekem is ez a fő feladatom az ATLAS-ban
- A teljes szoftvert több mint 5 óra megépíteni egy 16 processzoros gépen
 - Így a szoftverfejlesztést is egyedi módon kell végeznünk...
- Egy világméretű elosztott fájlrendszeren tároljuk
 - /cvmfs/atlas.cern.ch



Analízis program-kód



The screenshot shows a GitLab repository page for 'stop11-xaad' (Project ID: 57031). The repository is under the 'atlas-phys-susy-wg' group. It features a 'STOP code' logo and a 'No license' notice. The repository has 3,399 commits, 9 branches, 27 tags, and 1.6 GB of files. A recent merge of branch 'uproot' into 'master' is shown, authored by Javier Montejo Berlingen. The repository includes a README, CHANGELOG, CI/CD configuration, and a Kubernetes cluster. A table of files and their last commit details is provided below.

Name	Last commit	Last update
Analysis/RPV1L	Add systematic variations	2 months ago
LimitSetting	Change yapf formatting style to 'google'	6 months ago
5Wup	uproot enabled plotting	1 week ago
athena @ b292074b	Add systematic variations	2 months ago
tests	set workdir env and test condor submission	6 months ago
.dockerignore	Image build	6 months ago
.gitignore	simple setup.sh script for local dev	6 months ago
.gitlab-ci.yml	Add systematic variations	2 months ago

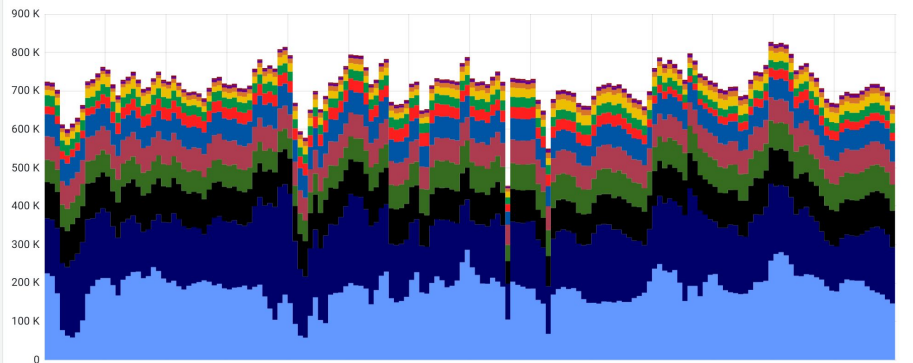
- Nagyon kevés fizika analízis létezik manapság nélküle...
- Részecskefizikában 2 dologra használjuk a legtöbb esetben
 - **Klasszifikáció:** Egy-egy esemény tulajdonságai alapján azt döntjük el, hogy az olyan-e mint amit keresünk
 - A rekonstruált részecskék fajtáját is sokszor így határozzuk meg
 - **Regresszió:** Részecskék / események tulajdonságait számoljuk ki “alacsonyabb rendű” tulajdonságai alapján. Pl. meghatározzuk egy hadronzáró pontos energiáját a kaloriméterekben érzékelt jelekből.
- Ilyen technikákat már nagyon régóta használunk a fizikában.
 - A múltban ez főleg döntési fákat és neurális hálókat jelentett
- Az utóbbi 5-10 évben nagy fejlődésen ment keresztül
 - Mind elméleti, mind gyakorlati szempontból
 - Mostanra “az ipar” által készített kódokat szeretjük a leginkább használni
 - A Google által fényképek azonosítására használt kód nagyon jól használható az LHC eseményeinek megértésére is 😊

Mekkora is az LHC/ATLAS grid?



Slots of Running jobs by Processing Cloud

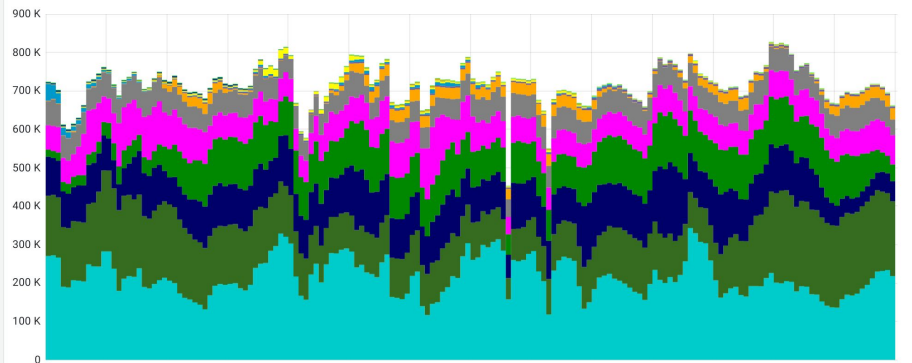
Last 7 days timeshift -1h



	min	max	avg	current
ND	57.8 K	287 K	185 K	157 K
US	92.8 K	302 K	168 K	123 K
DE	59.0 K	98.7 K	89.4 K	98.7 K
UK	42.7 K	71.6 K	63.4 K	67.1 K
CERN	38.7 K	66.3 K	61.1 K	38.7 K
FR	34.1 K	58.4 K	54.8 K	54.3 K
CA	14.5 K	34.9 K	26.8 K	20.9 K
IT	17.2 K	28.7 K	24.5 K	25.3 K

Slots of Running jobs by Activity

Last 7 days timeshift -1h



	min	max	avg	current
MC Simulation Full	117 K	343 K	216 K	211 K
MC Event Generation	55.4 K	241 K	142 K	189 K
MC Reconstruction	38.3 K	194 K	103 K	57.8 K
Group Production	18.3 K	140 K	95.8 K	39.2 K
User Analysis	45.0 K	149 K	73.2 K	74.2 K
Group Analysis	39.1 K	69.0 K	57.2 K	39.1 K
t0_processing	13.8	30.9 K	15.5 K	14.5 K
MC Resimulation	413	5.24 K	3.56 K	4.17 K

- Mostanra a szoftverfejlesztés egy nagyon lényeges elemévé vált a fizikának
 - Gyakorlatilag semmilyen kísérlet nem létezhet saját adatgyűjtő és feldolgozó szoftver nélkül manapság
 - Az adatok feldolgozását speciális programok írásával végezzük
 - Minden LHC fizikusnak muszály legalább “elfogadhatóan” programozni tudnia
- Az LHC több száz petabájt adatát a világméretű grid sok-százezer processzorán dolgozzuk fel



<http://home.cern>