Big Data jellegű adatok feldolgozása a WLCG segítségével

Házi András

Wigner FK

Big Data Mafihe Téli Iskola 2015

Bevezetés

- Mi a Big Data?
- Mi a Grid?
- Mi a WLCG?
- Mi a Big Data és a WLCG kapcsolata?

Big Data

- Big Data = gyűjtőfogalom, többféle definíció
- 3V:
  - Volume: az adatok mennyisége
  - Velocity: az adatok keletkezésének/feldolgozásának sebessége
  - Variety: az adatok sokszínűsége
  - +1 Veracity: az adatok minősége
- Technológia és szemléletmód is egyben

Big Data

- Szemléletmód:
  - Minél több adatunk van, azoknak fényében jobb döntéseket hozhatunk, ezért:
  - Gyűjts minél több adatot, később még jól jöhet
    - Redundancia! Ehelyett:
  - Gyűjts adatokat többféle forrásból, ezeknek a korrelációs vizsgálata jobb döntésekhez vezethet
  - Mivel az adatmennyiség mindig is relatív értelmű, ezért fontosabb a mennyiségnél a vizsgálati módszer

Big Data

- Technológia:
  - massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud based infrastructure (applications, storage and computing resources) and the Internet

Big Data

- Technológia:
  - Elosztott rendszerek alkalmazása struktúrált/nem struktúrált adatok gyors feldolgozásához (parallel processing)
  - MapReduce architekúra (Google, 2004), később open source klónja Hadoop (Apache)
  - Összetett architektúra előnyös, absztrakt rétegek közbeiktatásával a biztonságosabb, gyorsabb feldolgozás érdekében

Párhuzamos feldolgozás

- Elosztott rendszerek alappillérei:
  - Elosztott hardver
    - Dedikált memória és tárhely (szemben a párhuzamos koncepcióval, ahol ezek közösek)
  - Nagyon gyors Ethernet halózat
  - Elosztott fájlrendszerek
    - AFS
    - NFS
    - Hadoop FS
    - CVMFS
    - GFS
    - Lustre
  - Batch ütemező
    - Condor
    - LSF
    - Maui
    - SGE
    - Torque

  + multithreading kód, algoritmusok

Párhuzamos feldolgozás

Masszívan párhuzamos feldolgozás klaszter típusa:

- egymáshoz közeli számítógépek nagyon gyors halózattal
- klaszterek kialakítása
- klaszterek összeköttetése
- grid infrastruktúra
- sok processzor, cpu-ként/magonként dedikált memória
- közös tárhely klaszterenként
- elosztott fájlrendszer

Párhuzamos feldolgozás

- Grid computing
  - Elosztott rendszer típus
  - Heterogén HW környezet
  - OS szerint is lehet heterogén
  - Földrajzilag távoli klaszterek összeköttetése 1-10 GB halózattal (Gigabit, Infiniband)
  - Klaszterek, azokon belül a gépek egy összetett szoftver rétegen keresztül kommunikálnak
  - Cél: egy adott számítási feladat elvégzése ‘közös’ számítástechnikai erőforrás-halmazzal

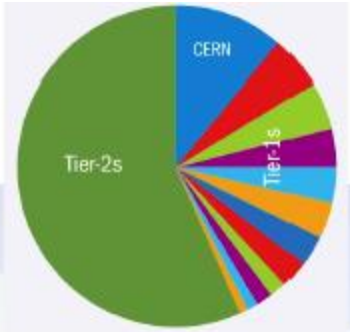
WLCG

- World LHC Computing Grid motívációja:
  - az LHC hatalmas adatmennyiséget generál
  - Helyileg képtelenség feldolgozni
  - Az adatok válogatása több lépcsőben (trigger HW/SW)
  - A válogatott adat továbbküldése az elosztott rendszer többi pontja felé
  - Gyors internet összeköttetés
  - A felhasználóknak nem kell tudniuk, hol van a keresett adat, amivel dolgozni szeretnének
  - Felhasználói feladat futtatásakor a middleware eldönti, hogy amaz melyik klaszteren, melyik site-on fusson

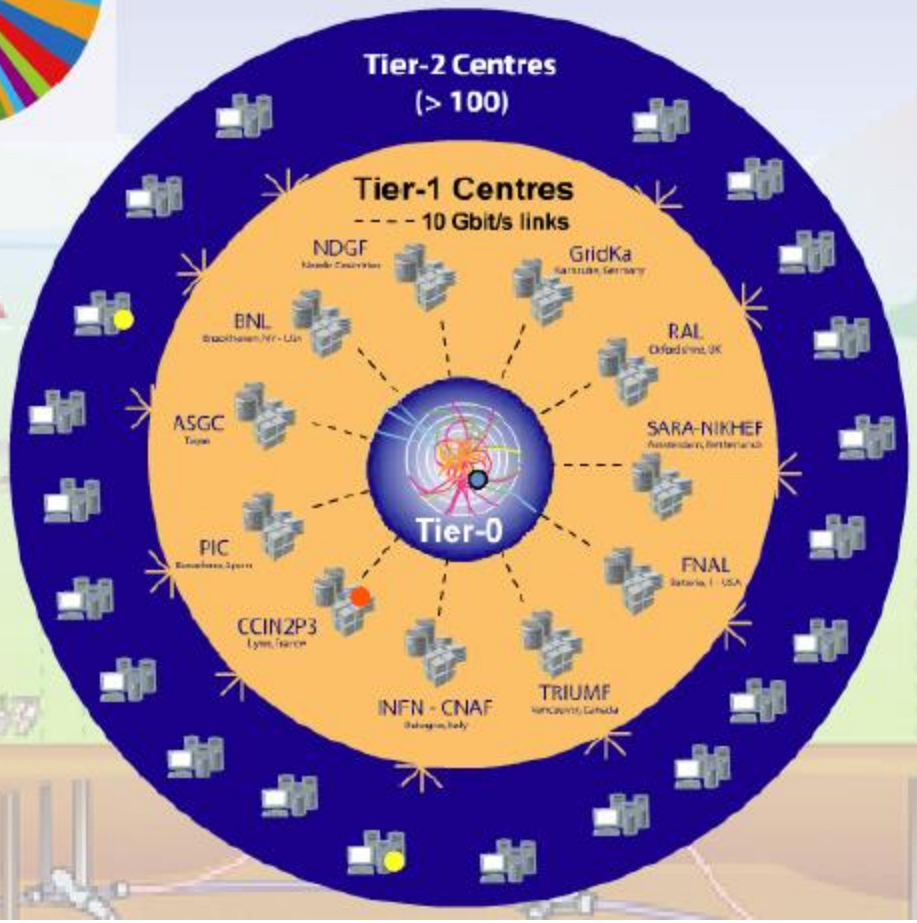
WLCG

- World LHC Computing Grid felépítése:
  - Trigger által válogatott adatok továbbküldése elosztott klaszter csomópontokhoz
  - ezek az ún. Tier 0/1/2/3 grid site-ok
    - Tier-0
      - CERN, Wigner Data Center
    - Tier-1
      - USA, Németo., Olaszo., Franciao., Anglia, Spanyolo. …
    - Tier-2
      - 148 világszerte, köztük a budapesti T2 site
    - Tier-3

WLCG



Tier 0 – Tier 1 – Tier 2

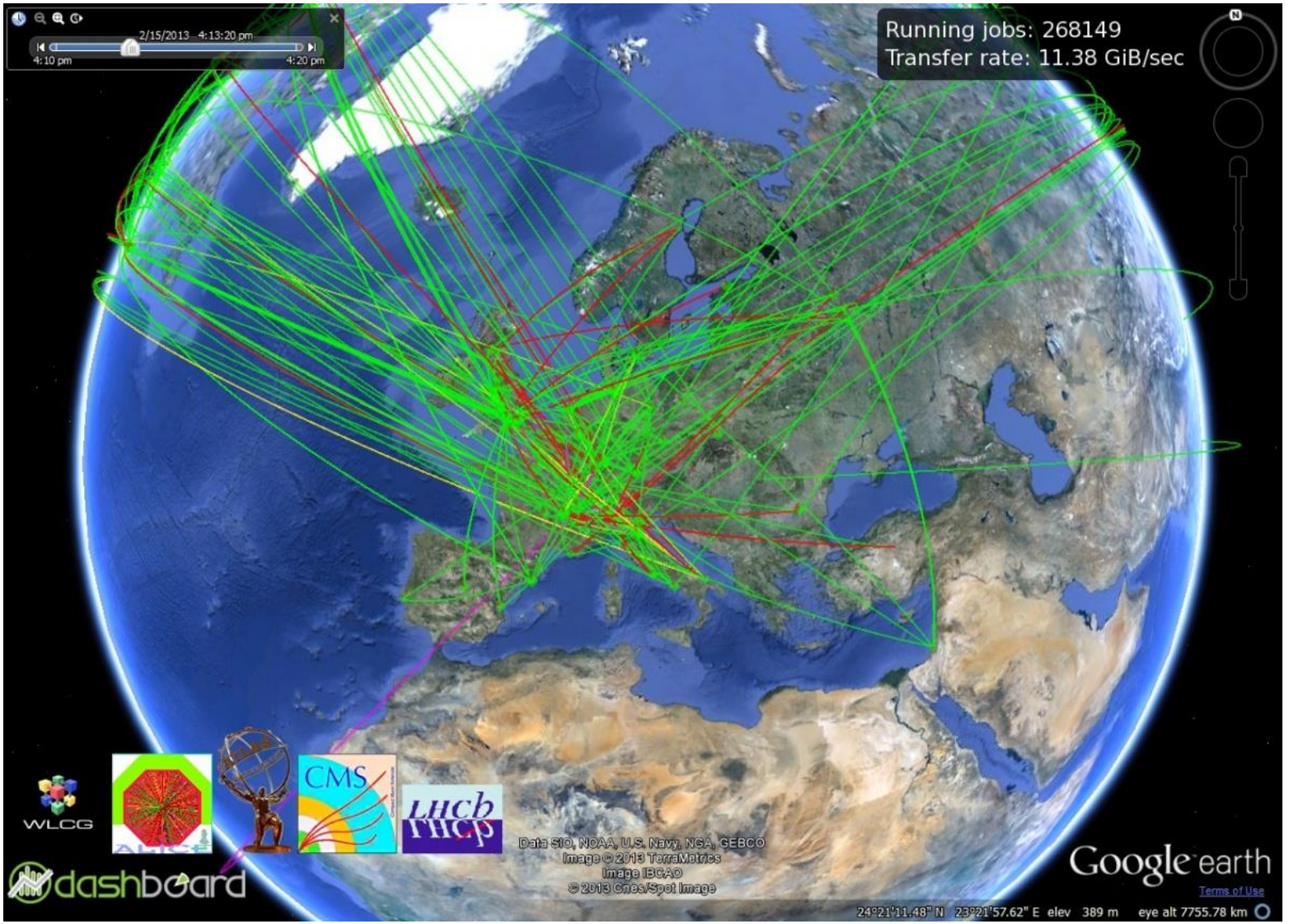


Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

WLCG


2/15/2013 4:13:20 pm
4:10 pm
4:20 pm
Running jobs: 268149
Transfer rate: 11.38 GiB/sec
WLCG
CMS
LHCb
dashboard
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2013 TerraMetrics
Image IBCAO
© 2013 Cnes/Spot Image
Google earth
Terms of Use
24°21'11.48" N 23°21'57.62" E elev 389 m eye alt 7755.78 km

WLCG

- World LHC Computing Grid middleware:
  - Sok komponensű csomag, a griden való párhuzamos futtatáshoz szükséges saját fejlesztésű és third-party, biztonsági, hálózati, batch, broker, információ szolgáltató, monitorozó szoftverelemek összesége
  - Eredetileg gLite néven futott
  - A bővebb EMI (European Middleware Initiative) része lett (jövőben UMD)
  - Az EGI projekten belül fejlesztik (Enabling Grids for E-sciencE (EGEE) jogutódja)

WLCG middleware

- Biztonsági réteg:
  - VOMS: Virtual Organisation Membership Service, user CA hitelesítetett tanúsítványt kap, ha valamelyik VO tagja. A tanúsítvánnyal tud hozzáférni a WLCG erőforrásaihoz (GSI, SSL). A VOMS rendszer felügyeli, hogy a user jogosultságainak megfelelő proxy kulcsot kapjon.
  - MyProxy: érvényes CA tanúsítvány mellett a usernek átmeneti X.509 proxyt biztosít, amit a lokális gép tárol, s job futáskor felhasználásra kerül. Általában 12 óráig érvényes a kapott kulcs, ezzel igazolja jogosultságát (roles, privileges) a user a jobok futtatásákor, passwordless elérés.

WLCG middleware

- Biztonsági réteg:
  - ARGUS: hitelesítő keretrendszer, több szintet ölel fel (PEP, PDP, PAP), eldönti egy felhasználóról, hogy jogosult-e egy adott erőforrás vagy szolgáltatás használatára (user mapping, LCMAPS engedélyezése), finomhangolás (policy definition), konfiguráláshoz XACML
  - Glexec: hitelesítési komponens, ARGUS-hoz kapcsolódik. Lehetővé teszi egy worker node-on egy processzus számára, hogy megváltoztassa a hozzárendelt user ID-t.

WLCG middleware

- Információs réteg:
  - APEL publisher: Accounting Processor for Event Logs, összegyűjti a batch és blah accounting logok alapján a jobokhoz tartozó információkat (CPU idő stb.) és beszúrja ezeket a lokális MySQL adatbázisba. Ezután elküldi (publishing) ezeket egy központi gyűjtő repoba.
  - BDII (site, top): az Information Service része, összegyűjti site-ra vonatkozó információkat, azt továbbítja a top-BDII felé, LDAP protokoll használata, DN attributumokkal, hierarchikus fa struktúra (kép). Site-ra vonatkozó pillanatnyi állapotról, felépítésről adatok lekérdezhetők (lcg-infosites, ldapsearch parancsokkal)

WLCG middleware

- Computing Element réteg:
  - CREAM CE: Computing Resource Execution And Management, lokális erőforrások halmazának (klaszter) irányító eleme, kapcsolatot létesít az ütemező (LRMS), a biztonsági, az információs, a tároló rétegek és a worker node-ok között. Része a Grid Gateway, ami az adott klaszter általános interfészeként működik. Rajta keresztül érkeznek meg a job requestek az LRMS rendszer felé. Elfogadja a requesteket közvetlenül, és a WMS oldaláról is egyaránt.

WLCG middleware

- Computing Element réteg:
  - Torque: LRMS, batch ütemező, a grid ‘motorja’. Közvetlenül ez a komponens osztja szét az erőforrások közt a beküldött feldolgoznivaló jobokat. Szorosan együttműködik a CREAM-el. Vannak más ütemezők is, akár third-party fejlesztésűek is, amit WLCG site-ok használnak: LFS, Condor, SGE/GE. A CREAM az adott ütemezőkkel az azok által értelmezhető query-k által kommunikál, ebből tudjuk folyamatosan követni, a futó jobok melyik állapotban vannak.
  - WN: worker node, maga a jobot futtató egység és annak szoftvercsomagja
  - UI: User Interface, a user által grid eléréshez használt node és annak szoftvercsomagja

WLCG middleware

- Workload Management réteg:
  - WMS: Workload Management System, a user számára transzparenssé teszi a job beküldést, egyszerű parancsokkal felügyelheti a user a job sorsát. A JDL nyelven irt konfigot beolvassa, s az ott meghatározott paraméterek alapján kiválasztja a jobnak leginkább megfelelő CE célpontot (match-making a megadott paraméterek alapján, továbbá a CE rank figyelembe vétele (pl. futó jobok és sorban várakozó jobok aránya)).
  - LB: Logging and Bookkeeping,a WMS által kezelt jobok nyomon követése.

WLCG middleware

- Tárolási réteg:
  - DPM: Storage Element annak a komponesnek a gyűjtőneve, mely univerzális hozzáférést biztosít egy grid site-on tárolt adatokhoz (diszk szerverek/tömbök, szalagos rendszerek irányítása). Többféle adatelérési interfészt és protokollt támogat (GSIFTP, RFIO, xroot, NFS). Ennek a központi eleme a Storage Resource Manager, aminek sokféle implementációja létezik, site-onként ezek különbözhetnek. A legfontosabbak: CASTOR, dCache, StoRM, és a DPM. A Disk Pool Manager csak a diszk alapú szervereknél alkalmazható.

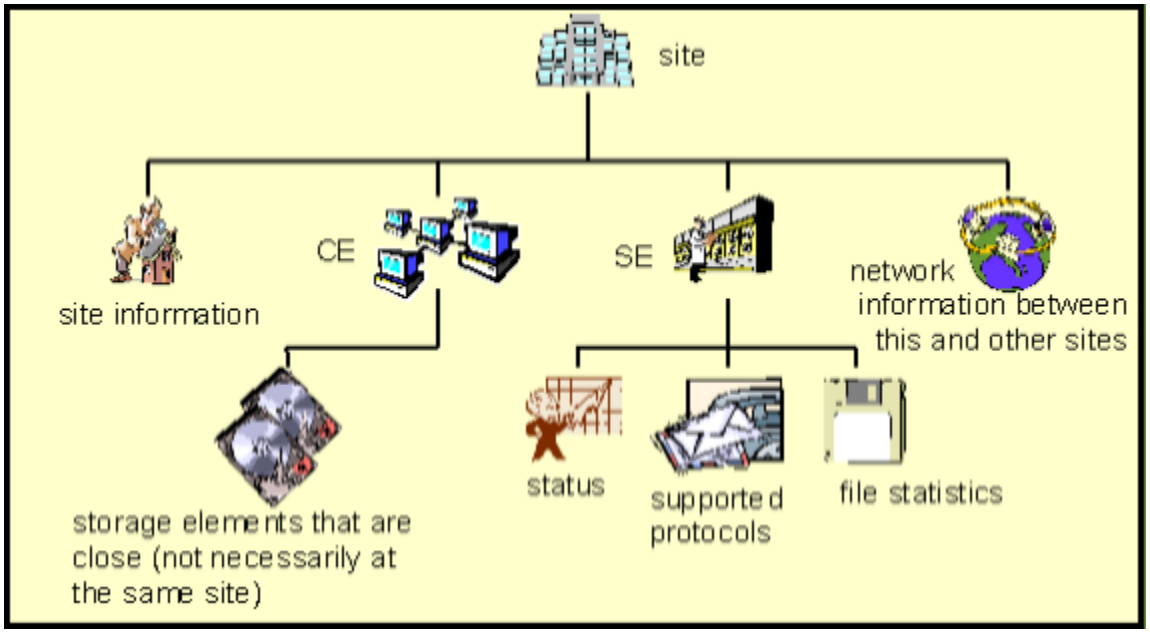
WLCG middleware

- Tárolási réteg:
  - LFC: LHC File Catalogue. Tárolásnál fontos szempont, hogy user elől elrejtsük a tárolási-transzfer rétegek komplexitását. Ebben segítenek a Data Management eszközök, melyek által kezelhetők az adatok replikái is. Az elsődleges egység itt a fájl. Egy-egy fájl másolata létezhet különböző site-ok diszkjein. Az ezek közti keveredést hivatott megakadályozni az egyedi azonosítók (GUID), logikai nevek (LFN) használata.

WLCG middleware

- Tárolási réteg:
  - LFC (folyt.): Továbbá a megkülönböztetést elősegíti a kétféle elérési útvonal, SURL és a TURL alkalmazása. Az első az SRM támogatott SE esetében használatos, a második tartalmazza az elvégzendő fájlművelethez tartozó egyéb információkat (pl. protokoll stb.). A fentiek egymáshoz való hozzárendelése a File Catalogue végzi, a WLCG-n belül ez az LCG File Catalogue.
  - FTS: File Transfer System, az SE pontok közti fájlátvitelek ütemezéséért, indításáért, fogadásáért felelős service. Kiválóan alkalmas a nagy volumenű adatátvitelek lebonyolítására.

WLCG middleware



Hagyományos job flow

- Digitális tanúsítvány megszerzése → regisztrálás egy VO-ban → belépés egy UI-ra → proxy gyártása a Grid eléréséhez
- Job beküldés WMS által → job leíró JDL fájl átkerül az UI-ról a WN-re (InputSandbox) → egy eseményrekord bekerül az LB nyilvántartásba → job státusz: SUBMITTED

Hagyományos job flow

- A WMS megkeresi a legjobb elérhető CE-t (ISM, a BDII outputból beolvasott belső cache segítségével) → megkeresi a hardver és tárolási erőforrásokat (LFC) → job státusza: WAITING
- A WMS előkészíti a jobot a futáshoz → az utasítások script-jei átkerülnek a kiválasztott CE-re → job státusza: READY

Hagyományos job flow

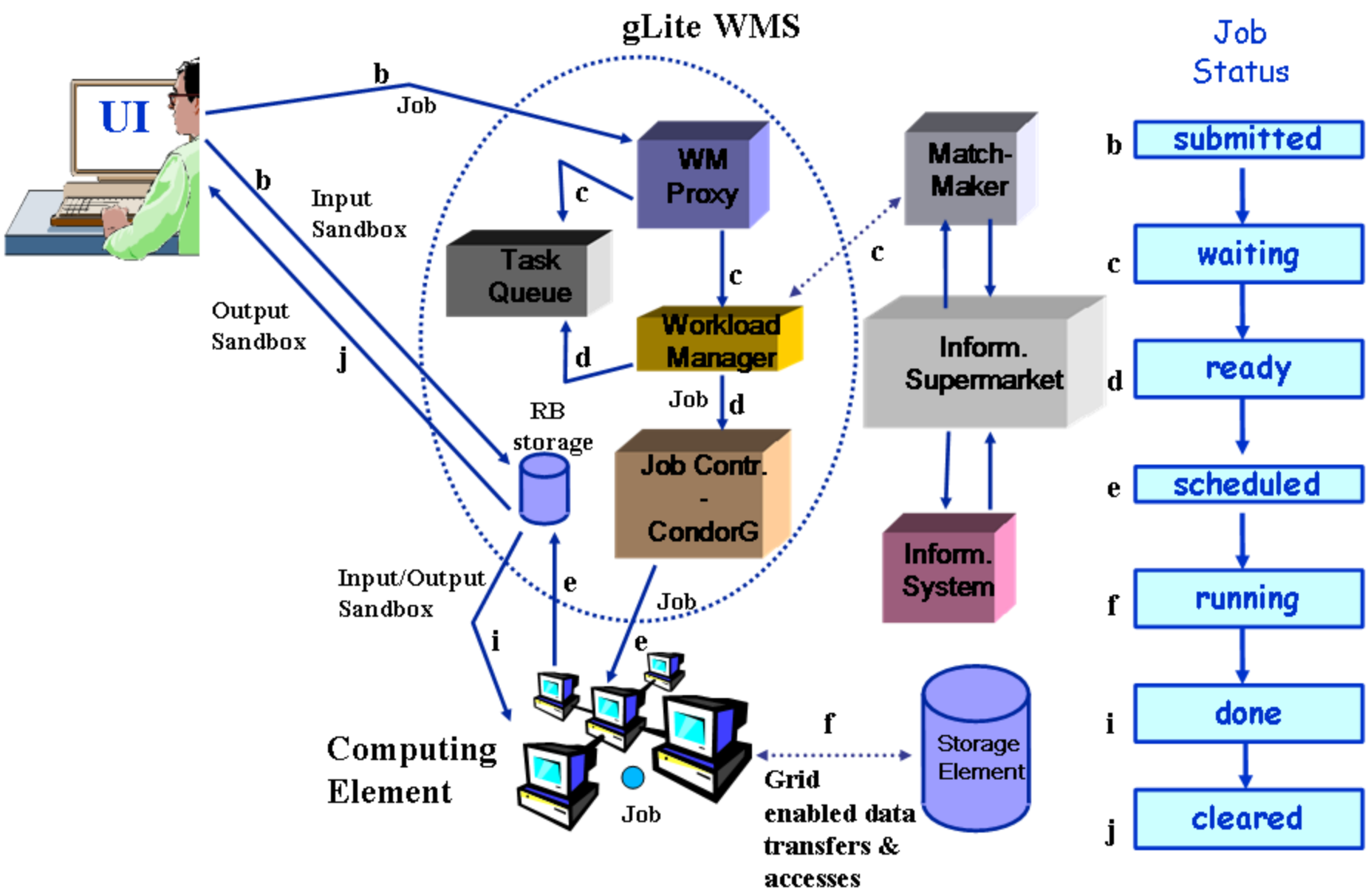
A CE megkapja az utasításokat és elküldi a job instrukciókat végrehajtásra az LRMS-hez → job státusza: SCHEDULED

Az LRMS kezeli a job futást a WN-eken → az InputSandbox tartalma átkerül a WN-ekre → job státusza: RUNNING

Amíg a jobok futnak, a grid fájlok elérhetők az SE-n → ezek átmásolhatók a WN lokális fájlrendszerére

Hagyományos job flow

- A job létrehozhat újabb output fájlokat, melyek elérhetőkké tehetők más Grid userek számára → a futás eredménye felmasolható az SE-re → regisztrálűdik a fájl katalógusban (LFC)
- Ha a job lefut hiba nélkül → kisebb méretű fájlok másolásra kerülnek a WMS node-ra → job státusza: DONE
- A futás eredménye felmásolható az UI-ra → job státusza: CLEARED


gLite WMS
UI
b
Job
b
Input
Sandbox
Output
Sandbox
j
WM
Proxy
c
Task
Queue
c
Workload
Manager
d
Job
d
RB
storage
Job Contr.
-
CondorG
Input/Output
Sandbox
e
Job
i
e
Computing
Element
Job
f
Grid
enabled data
transfers &
accesses
Storage
Element
Match-
Maker
c
Inform.
Supermarket
Inform.
System
Job
Status
b
submitted
c
waiting
d
ready
e
scheduled
f
running
i
done
j
cleared

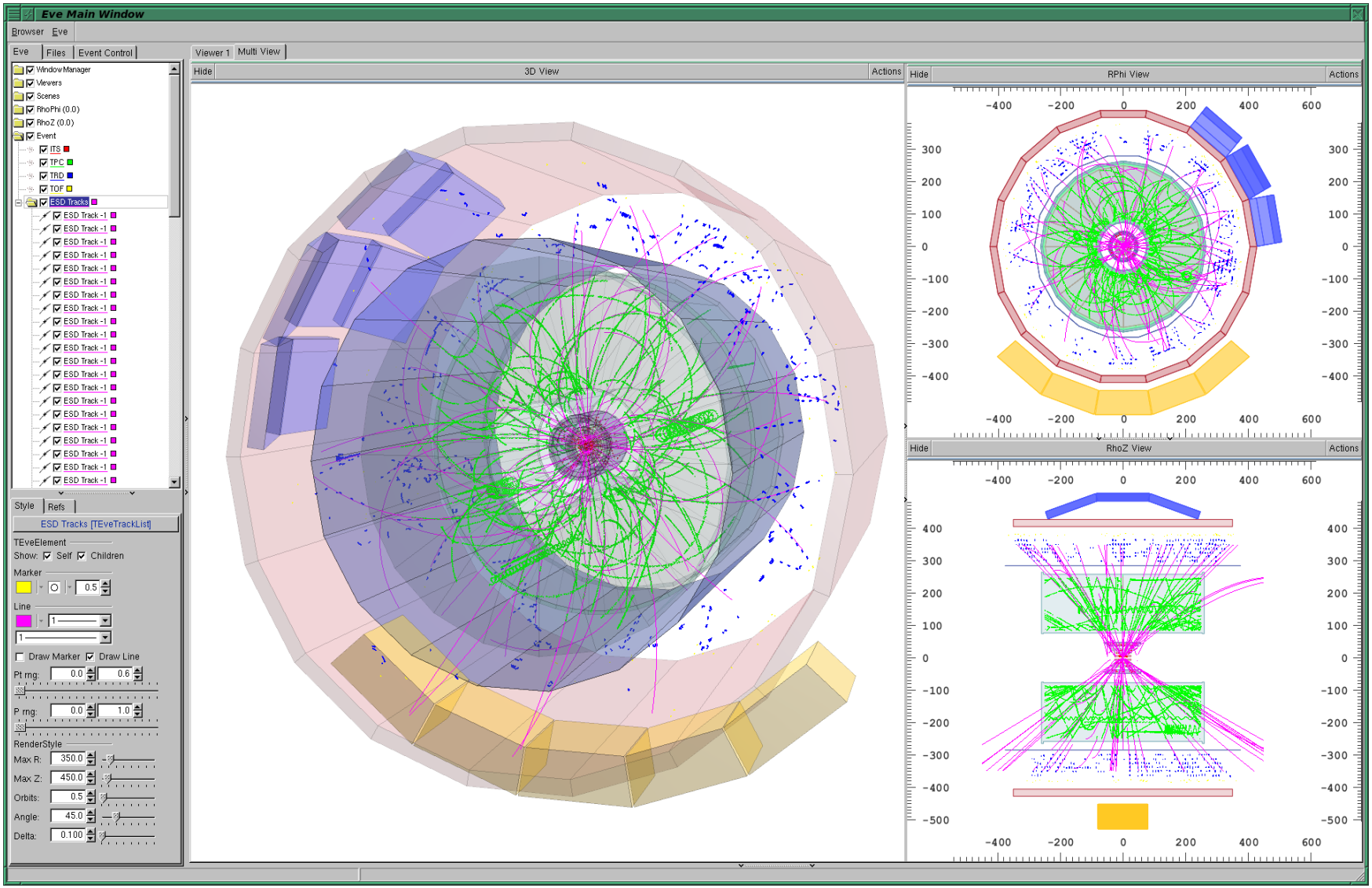
WLCG job szoftverek

- WLCG jobok futásakor használt legfontosabb eszközök:
  - Dirac (LHCb)
  - Panda (ATLAS)
  - Ganga (ATLAS, LHCb)
  - AliEN (ALICE)
  - CRAB (CMS)
  - Geant
  - Root

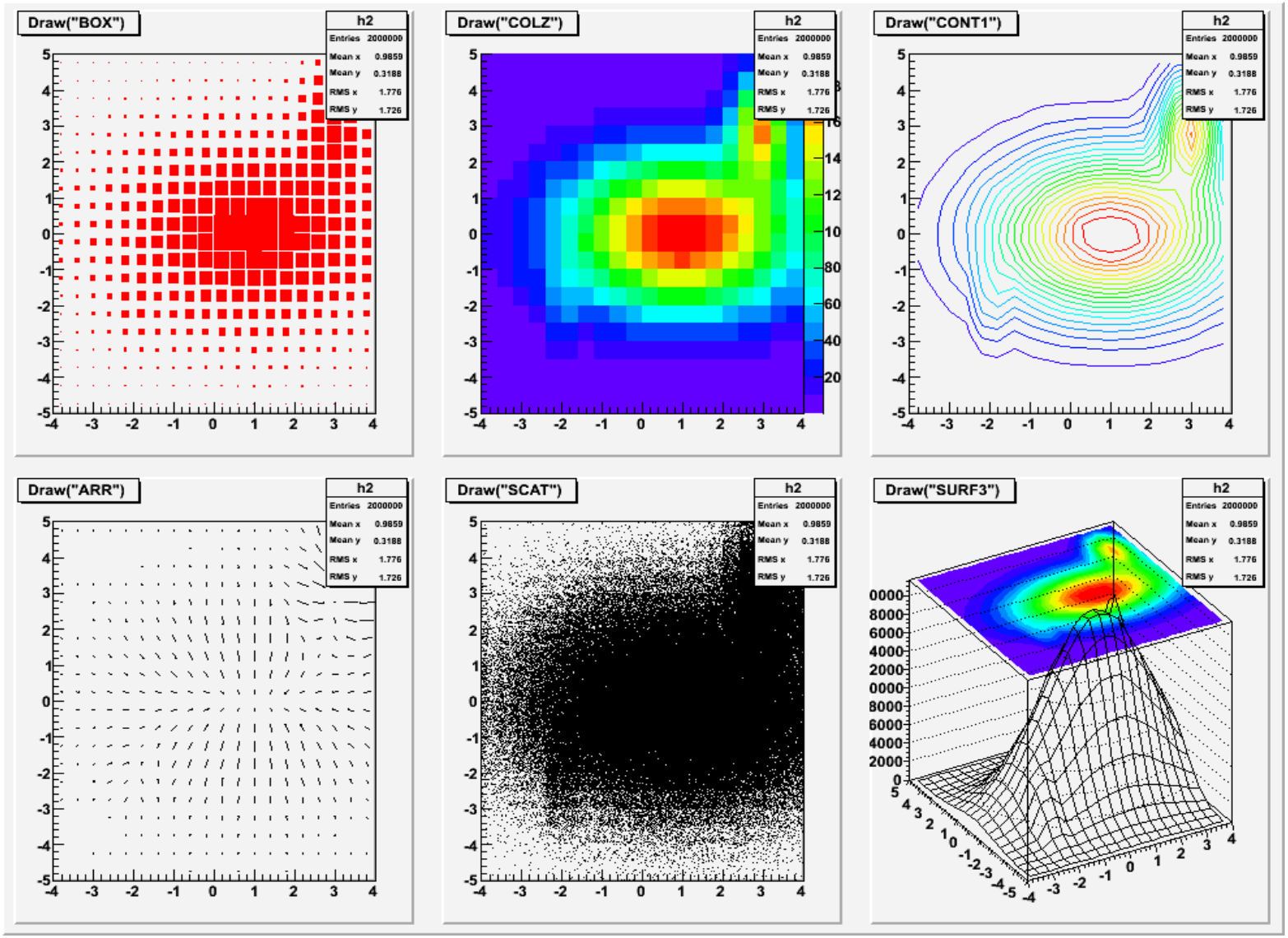
WLCG szoftverek

- ROOT: az adatfeldolgozás, és vizualizáció alapvető keretrendszere, nemcsak a nagy energiás fizikáan használják.
- PROOF: Parallel Root Facility, a ROOT kiterjesztése, melynek segítségével az analízis során használt ROOT fájlok párhuzamosíthatók (klaszter, vagy multicore)

WLCG szoftverek



WLCG szoftverek

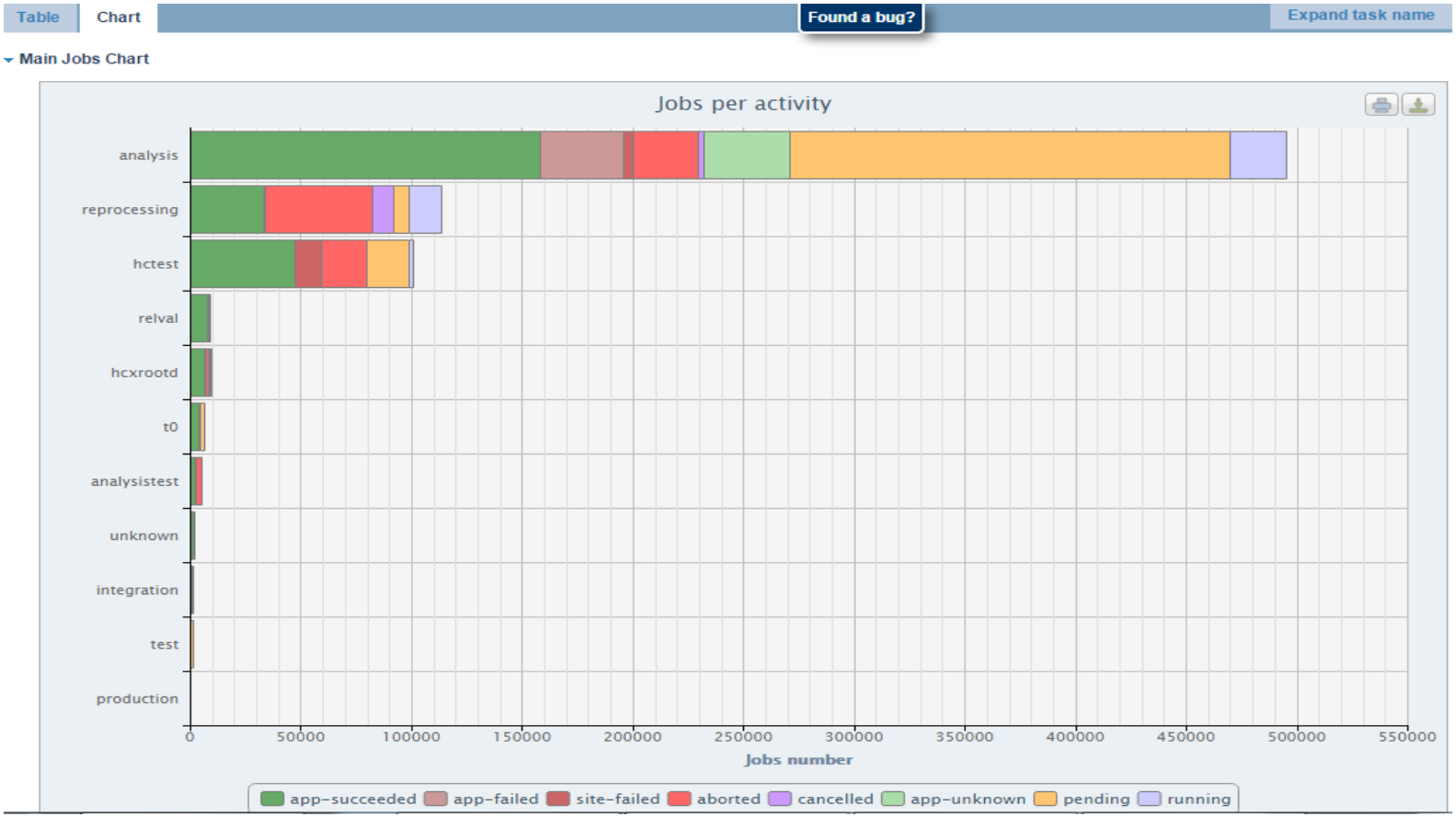


WLCG CMS szoftverek

- CRAB: CMS Remote Analysis Builder egy eszköz, amely megkönnyíti a user számára a grid használatát, ha CMS jellegű munkát (analízis, MC stb.) kíván végezni
- CMSSW: CMS offline simulation framework
- Data Aggregation Service (DAS) egy CMS projekt, mely összekapcsolja a különböző adat szolgáltatásokat (DBS, Phedex, SiteDB, etc), egyszerű query-ken alapuló nyelvet használ. Támogatja az JSON/XML-formatúmokat, biztosít parancssor és webalapú elérést.

WLCG CMS szoftverek

- Dashboard: monitoring framework, mindegyik kísérletnek van saját felülete



WLCG site T2_HU_BUDAPEST

- NGI_HU része
- A CPU core szám 1000 felett (kb. 870 a worker node-ok)
- Storage: 350 TB
- Támogatott VO-k:
  - CMS (67%)
  - ALICE (23%)
  - Hungrid (8%)
  - VIRGO (2%)
- OS : SLC 6, szinte teljesen homogén
- Middleware: EMI
- Batch ütemező: GE (SGE open source verziója)

WLCG site T2_HU_BUDAPEST

Big Data @ WLCG

- Szemléletmód:
  - Az adatok generálása nagyon intenzív
  - Archivált adat 100 PB körüli (pl. oktatás), ez a jövőben elérheti az 5 EB-ot (5.000.000 TB)
  - Pontosabb eredmény ha nagyobb a statisztika, de csak a válogatott adat érdekes
  - Mindig lehet tudni, mit keresünk (fizikai szempontból mi lehet értékes), nem jellemző a redundancia
  - Struktúrált adathalmazok
  - Nem csak a detektorokból jövő adattal kell számolni (online, offline, analízis (Monte Carlo, nyers adat), rekonstrukció), hanem a WLCG állapotát jelző, monitorozáshoz használt információval is

Big Data @ WLCG

- Szemléletmód:
  - Az adatoknak irányítottan csak egy része kerül archiválásra, majd analizálásra; még így is túl sok adat (volume)
  - Tárolt adat elérése kritikus szempont (velocity)
  - Az adat sokféle: az ütközések rekonstrukciójához szükséges paraméterek értékeit tartalmazza, de különböző analízisek, MC-k különféle adatot adnak(variety)
  - Minőség kritikus (veracity)

Big Data @ WLCG

- Technológia:
  - 170 intézmény számítógép központja nagyon gyors internetes összeköttetéssel (globális átlag adatátvitel: 10 GB/s)
  - Napi 1.5-2 millió job futtatása
  - Adatok petabyte-nyi nagyságrendben vándorolnak az elosztott rendszerek központjai közt
  - Sok lépcsős feldolgozás, a middleware architektúra absztrakt rétegekből áll
  - Sok egyéb mellett a piaci trend-diktáló szereplők alkalmazásaihoz hasonló vagy megegyező szoftverek, architektúrák, protokollok (StoRM, dCache, Hadoop, Squid, Maui, SGE, LDAP, MongoDB, SOAP, MPP, Cloud)

Köszönöm a figyelmet!

Linkek, hivatkozások

Data processing at CERN (video):

https://www.youtube.com/watch?v=jDC3-QSiLB4

Open Data Portal:

http://opendata.cern.ch/

Wigner FK Grid T2 web:

grid.wigner.mta.hu