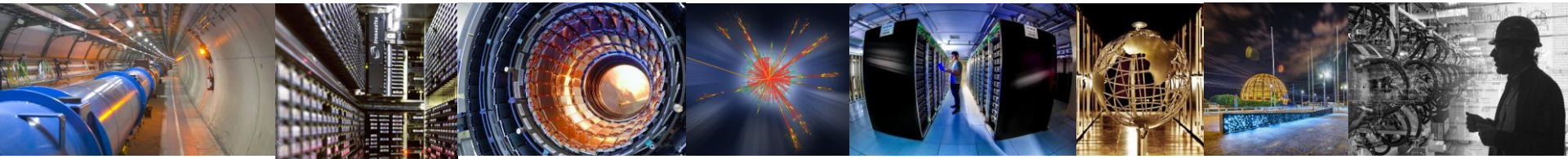
Разпределени изчисления в ATLAS

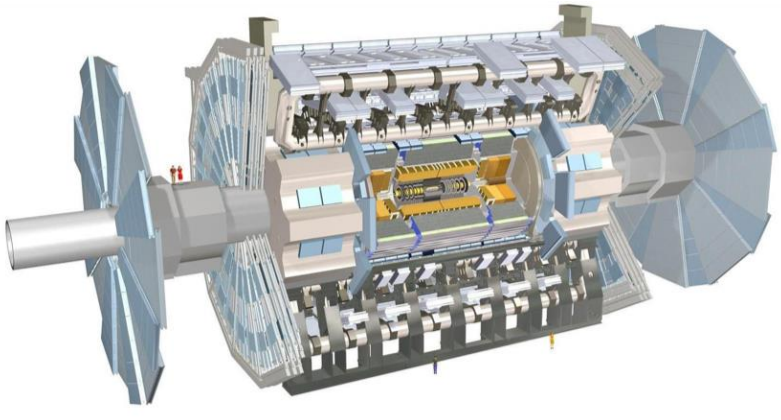
Част 1: GRID



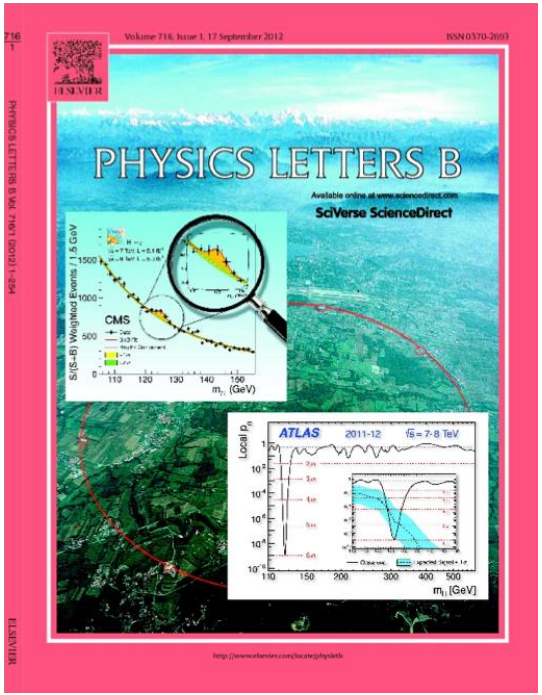
Др. Иван Глушков
Тексаски университет / АТЛАС
Българска инженерна учителска програма
ЦЕРН, септември 2022

Какъв е проблема?

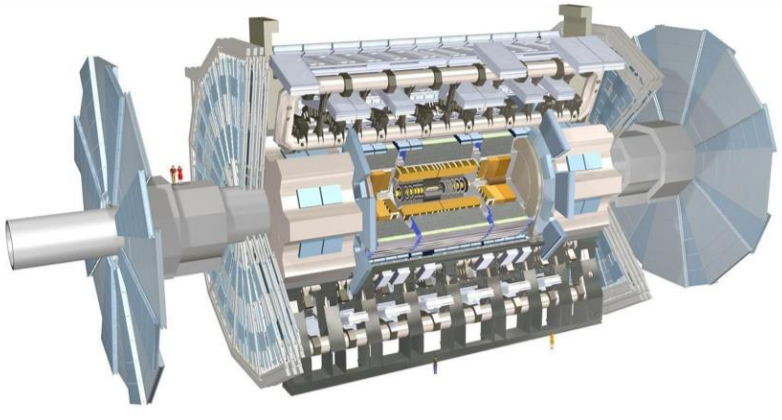
От ATLAS до научна публикация



Как?

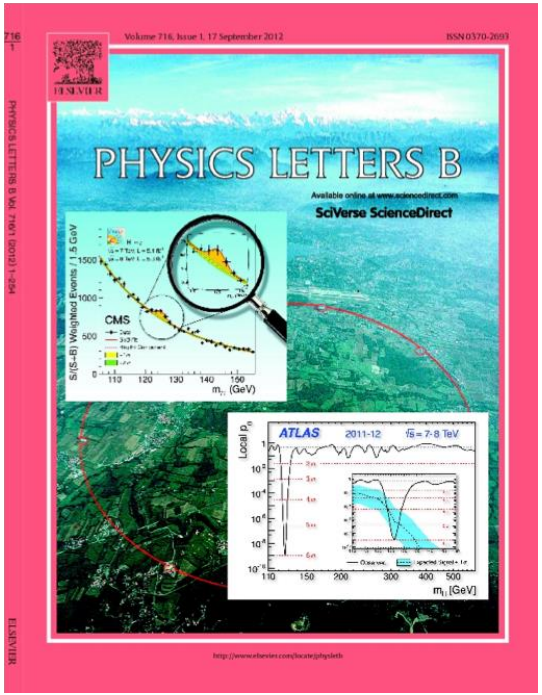


От ATLAS до научна публикация

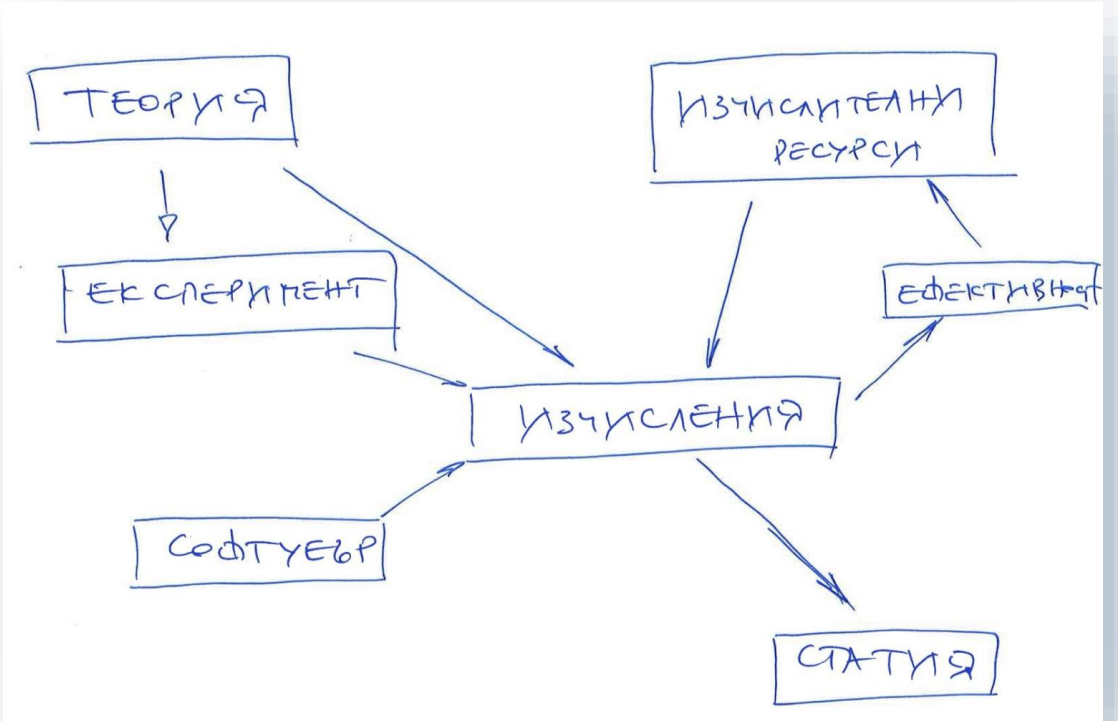


Как?

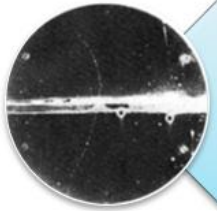
.. и бързо моля



Парадигма на науката



Какво ви е (било) нужно за откритие?

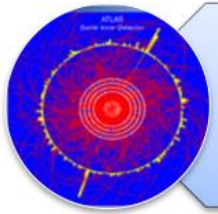


През 1930-те

- ~2 учени от една държава
- Лист и химикал

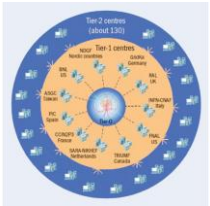
През 1970-те

- ~200 учени ~10 държави
- Мейнфрейм



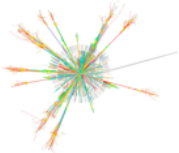
Днес

- ~3000 учени ~100 държави
- **Разпределени изчисления**



Данните

Колко са данните от ATLAS



- Произвеждаме: 40 MHz x 1.5 MB = 60 TB/s
- Данни == C++ обекти
- Време за обработка: ~дни т.е. ~ O(MB - GB)
- Пресяване в реално време:
  - L1 тригер - на експеримента
  - L2 тригер - ~120 000 ядра

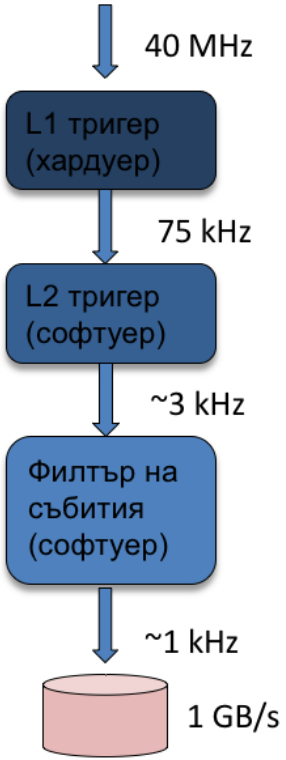
```
SELECT SUM(bytes)/1000/1000/1000/1000/1000 FROM atlas_rucio.dids
WHERE did_type='D'
and datatype = 'RAW'
```

ery Result

SQL | All Rows Fetched: 1 in 68.054 seconds

SUM(BYTES)/1000/1000/1000/1000/1000
64.33564873843434

64 PB!



КАК?!

64 PB!?

Колко време?
Ами ако сбъркам?
Ами ако колегата направи по добър алгоритъм за мюони?
Всичко отначало?!

Колаборация и оптимизация

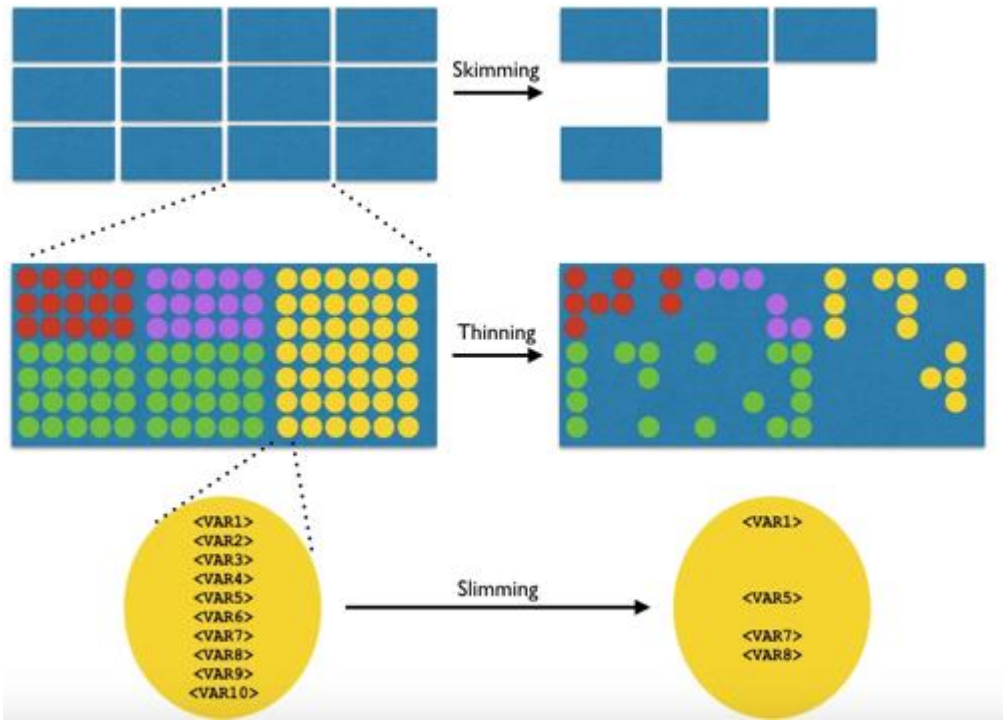
58 PB!?

Колко време?
Ами ако сбъркам?
Ами ако колегата направи по добър алгоритъм за мюони?
Всичко отначало?!

- Групиране на анализите по теми - Хигс, Суперсиметрии, Екзотики
- Селектиране само на данните които са релевантни за дадения анализ

GB / MB

Селектиране на данни
(ATLAS Derivation Framework)



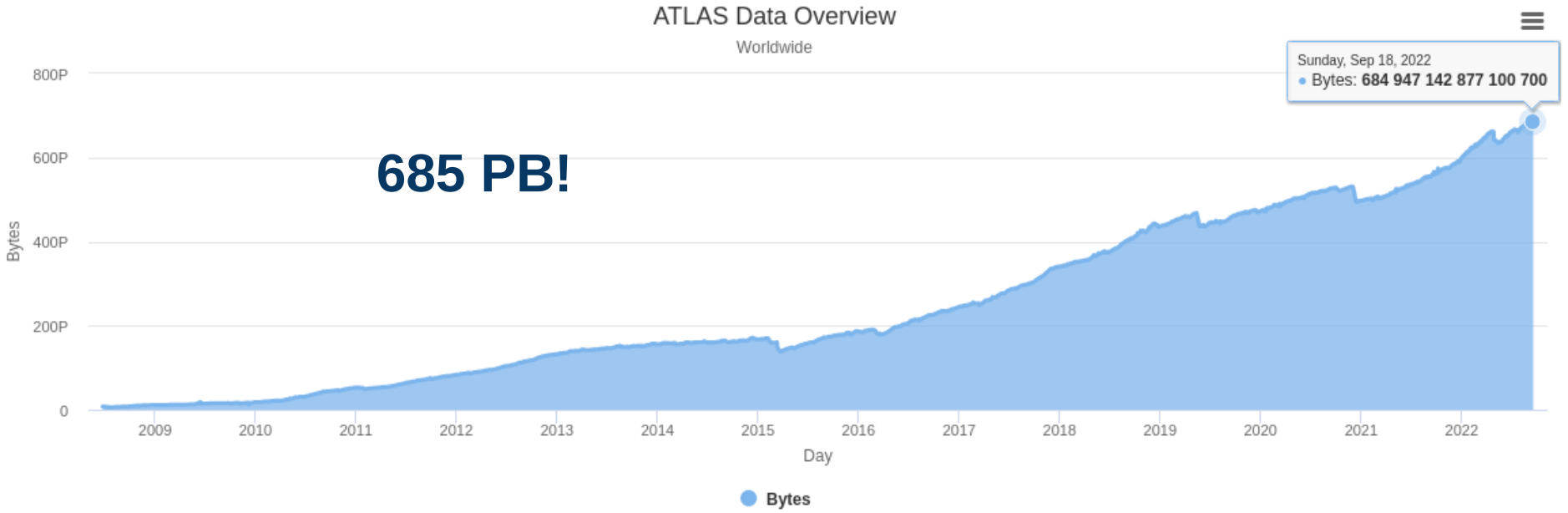
Графика: James Catmore

Премахване на ненужната информация

- Цели събития
- Части от събитията
- Характеристики на събитията
- За всеки анализ - отделни данни

Истина и теория

Всичко общо..



685 PB!

Изчисления

(колко и какви)

Колко изчислителни ресурси ни трябват?

“Колкото повече, толкова повече!”

Мечо Пух, 1966

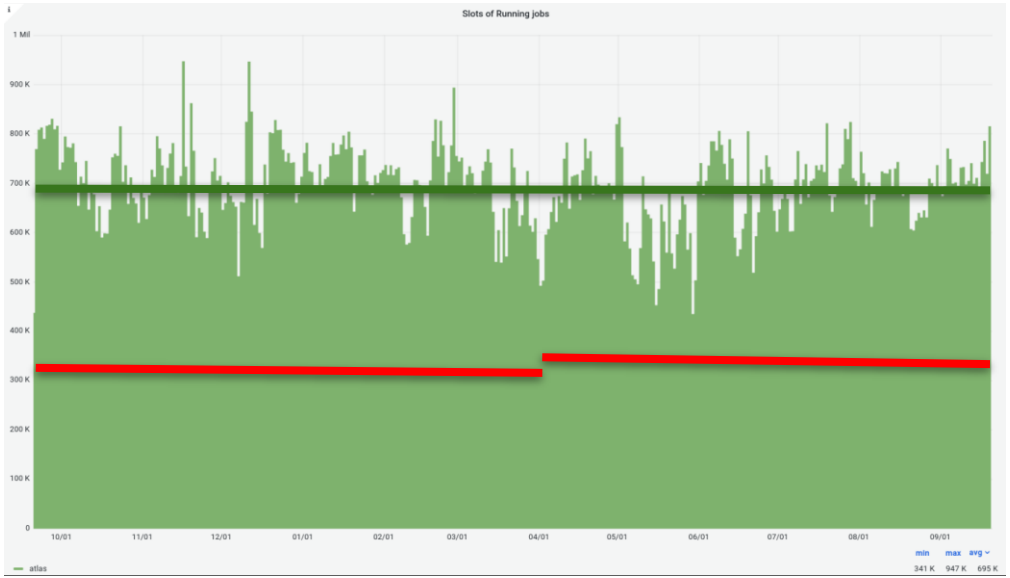
“Колкото повече, толкова повече!”

Мечо Пух, 1966

“Има само едно нещо, което е по-хубаво от гърненце мед.. И това са две гърненца мед.”

Мечо Пух, 1966

Колко използваме?



Използвано (695 000)

Обещано (359 000)

Речник

- Слот - място за един процес заделено от компютърна ферма
- Задача (Job) - количеството “работа” което трябва да се сметне на един слот

Колкото повече, толкова…

- По-бързо
- Изследване на повече нови опции - генератори, конфигурации..

За какво ги използваме?

Изчисления

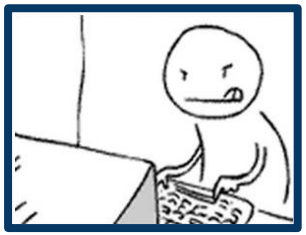
(къде)

Разпрпределени изчисления в ATLAS: Обзор

- Експерти: > 200 (в различна степен на заетост)
  - Системата не може да работи без тях
  - Разработка и интеграция на нововъведения
  - Експлоатиране / опериране на системата
- (Централизирани) услуги
  - Разпределяне на задачи - PanDA (ATLAS)
  - Разпределяне на данни - Rucio (ATLAS)
  - Прехвърляне на файлове - FTS (WLCG)
  - Разпределяне на софтуер - CVMFS (WLCG)
- Отдалечени компютърни центрове (сайтове)
  - GRID
  - Облаци, суперкомпютри, доброволчески изчисления (във втората лекция)
  - Хардуер: ресурси за обработка и съхранение на данни
  - …. и експерти на местно ниво!

Потребител и разпределени изчисления:

Приниципп на работа



Къде / кои са данните които искам да обработвам?

Данни X, които се намират на сайт Y

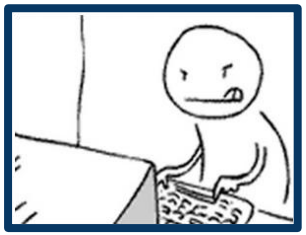
Система за разпределяне на данни (Rucio)

Речник

- Сайт - (отдалечен) компютърен център

Потребител и разпределени изчисления:

Приниципп на работа



Къде / кои са данните които искам да обработвам?

Система за разпределяне на данни (Rucio)

Данни X, които се намират на сайт Y

Искам да изпълниш моя код върху данните X

Система за разпределяне на задачи (PanDA)

Задачи

ГРИД

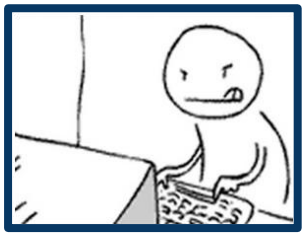
ATLAS сайт

Речник

- Сайт - (отдалечен) компютърен център

Потребител и разпределени изчисления:

Приниципп на работа



Къде / кои са данните които искам да обработвам?

Данни X, които се намират на сайт Y

Искам да изпълниш моя код върху данните X

Резултати?

Резултати / Още не е готово

Задачи

Система за разпределяне на данни (Rucio)

Система за разпределяне на задачи (PanDA)

ГРИД

ATLAS сайт

Речник

- Сайт - (отдалечен) компютърен център

Какво е GRID

GRID е **технология** която позволява оптимизиран, оторизиран достъп до ресурси - компютърни и за съхранение на данни - принадлежащи на различни собственици.

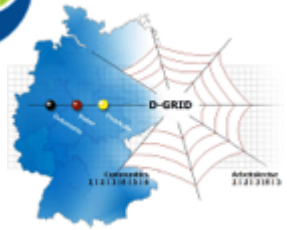
Какво е GRID II

- Идеята се ражда през 90-те
- Аналог на мрежата за електроснабдяване

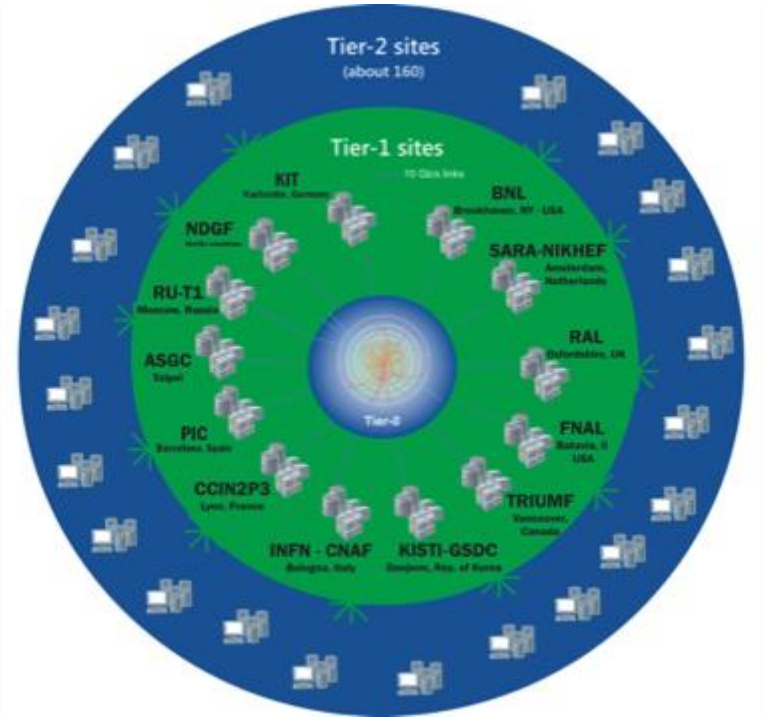
Какво е GRID III

- Софтуер или “GRID middleware” - дава достъп до хетерогенните ресурси на GRID с общ интерфейс
  - CE (Computing Element) - достъп до изчислителни ресурси
  - SE (Storage Element) - достъп до ресурси за съхранение на данни
- Потребителски достъп - идентификация и упълномощаване
  - Няма как да дадем логин и парола на x10000 хора на x100 сайта
  - Дигитални сертификати (x509-базирани)
    - Раздавани от сертифицирани “авторитетни източници” (certification authority)
    - Важи за всеки GRID сайт
    - Локално на сайта всеки сайт съответства на локален акаунт (и дефинира правата)

GRID-ове, GRID-ове..



WLCG (The Worldwide Computing GRID)



Участващи GRID-ове:

- EGI (European GRID Infrastructure) - Европейска GRID инфрастриктура
- OSG (Open Science Grid) - GRID за общодостъпна наука (САЩ)
- NDGF (Nordic Data Grid Facility) - Скандинавски GRID

ATLAS Сайт

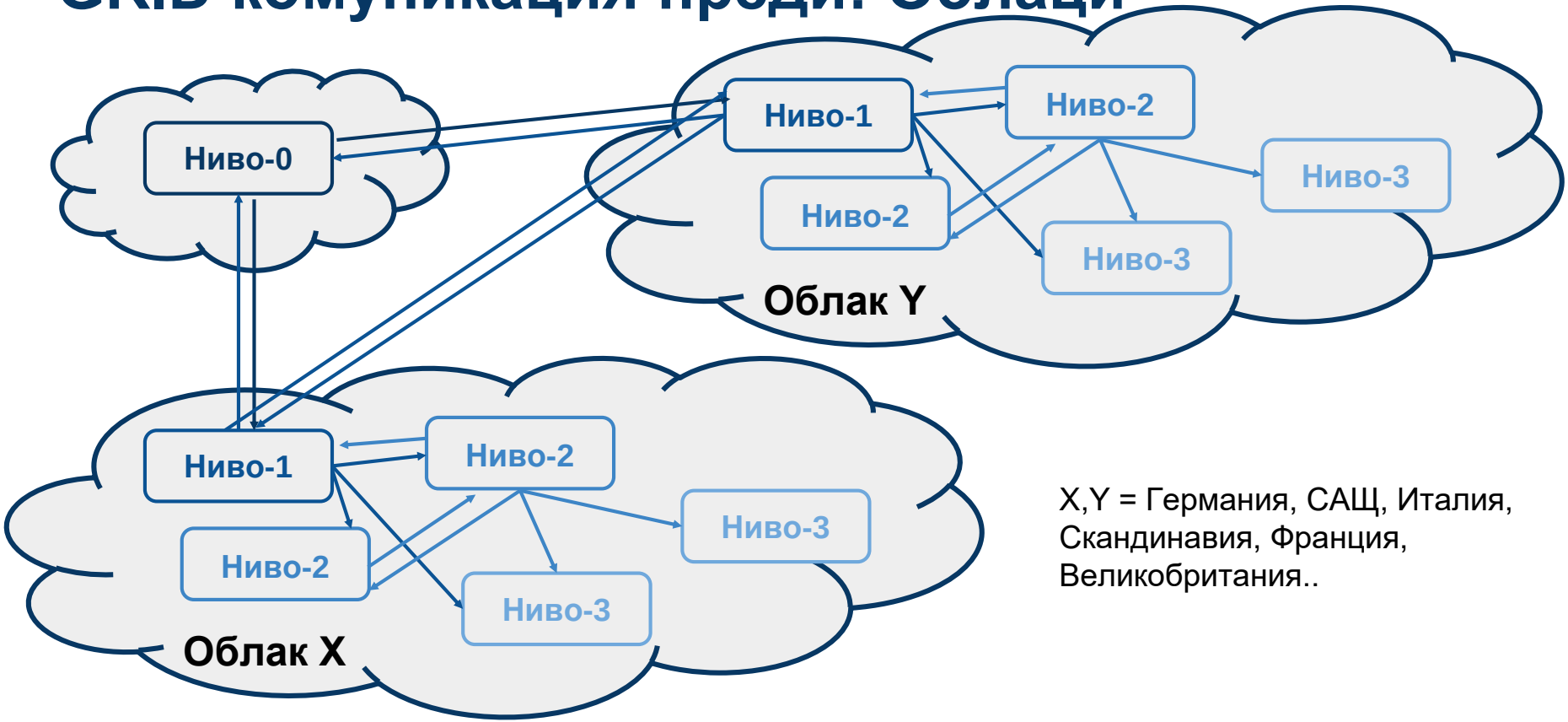
- Всеки сайт (обикновено) се състои от ресурси за съхранение на данни + изчислителни мощности
  - Squid proxy cache - за разпространяване на софтуера и “detector conditions” база данни

- Нива:
  - Ниво-0 (ЦЕРН)
  - Ниво-1 - Лентови носители + някои услуги (FTS))
  - Ниво-2 - Сайтове с подписан “договор” (Memorandum of Understanding - MoU)
  - Ниво-3 - Непостоянни ресурси

ATLAS Облак

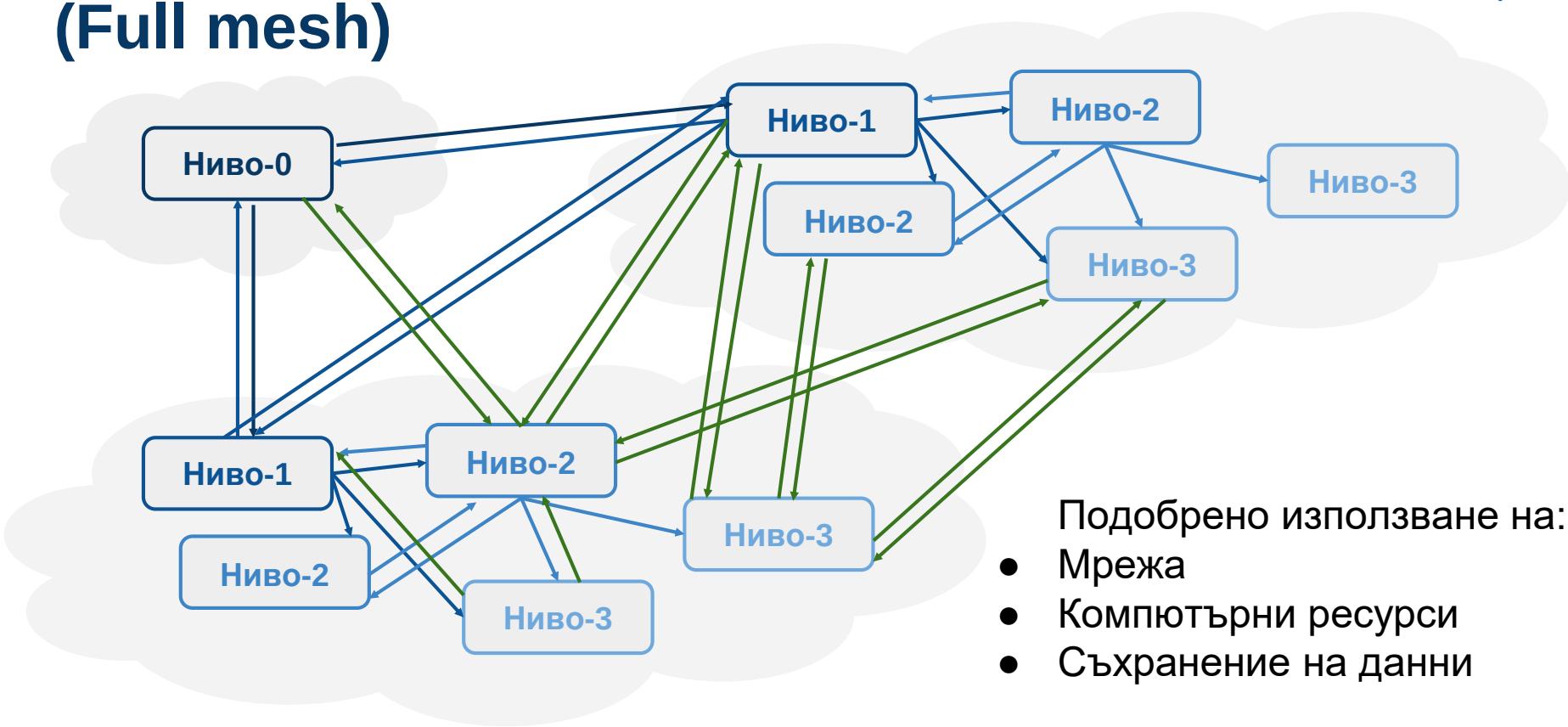
- Няма нищо общо с “облачни изчисления” (AWS, Google Cloud, и т.н.)
- Логическа група от сайтове
  - Един сайт от ниво-1 и няколко ниво-2 и ниво-3
  - Или в един географски регион, или от един източник на финансиране
- Поддръжката е предоставена от екипите на отделните облаци
  - Близо до сайтовете и техните проблеми
  - Най-често - на същия език
- Остаряла концепция
  - Диктувана от ограничения в мрежовите скорости
  - Все още се използва донякъде - поддръжка.

GRID комуникация преди: Облаци



X,Y = Германия, САЩ, Италия, Скандинавия, Франция, Великобритания..

GRID комуникация сега: Тотална мрежа (Full mesh)



Подобрено използване на:

- Мрежа
- Компютърни ресурси
- Съхранение на данни

Роля на нивата: Ниво-0

- Ниво-0
  - Копие на всички първични данни на лентови носители
  - Първо ниво на обработка на данните от LHC

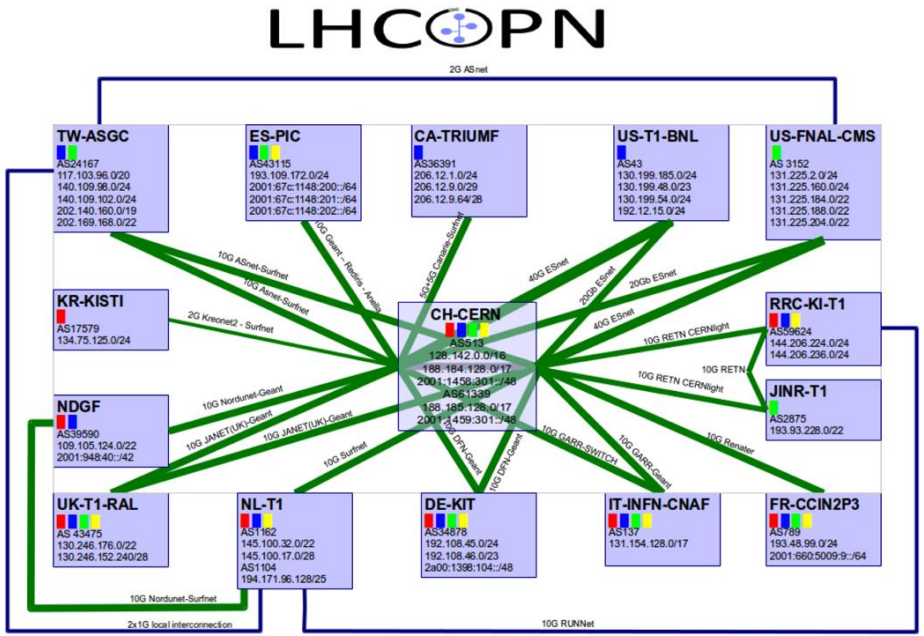
- Централни услуги в ЦЕРН (стабилност)
  - Разпределяне на задачи - PanDA
  - Разпределяне на данни - Rucio
  - База данни - Oracle
  - Прехвърляне на файлове - FTS (WLCG)
  - Кешова инфраструктура - Frontier
  - Разпределяне на софтуер - CVMFS - Stratum 0 (ниво-0)

Роля на на нивата: Нива 1, 2, 3

- Нива 1 и 2 - договорени (pledged) компютърни ресурси и дискови и лентови носители
  - Съхранение копие на първичните данни и на вторични данни
  - Разлики между ниво-1 и ниво-2
    - Лентови носители (само в ниво-1)
    - Поддръжка
      - 24x7 за ниво-1
      - В рамките на работния ден - ниво-2
    - Някои ниво-1 сайтове доставят и централни услуги (Frontier, FTS)
      - Близост до сайтовете
      - Сигурност на услугата

Мрежа

- Ниво-0 и Ниво-1: LHCOPN
- Ниво-0, 1 и 2: LHCONE (LHC Open Network Environment)
  - Колаборация между научни организации и университети
  - Динамично използване на ресурсите, споделяне на разходите
  - Layer3 рутиран VPN

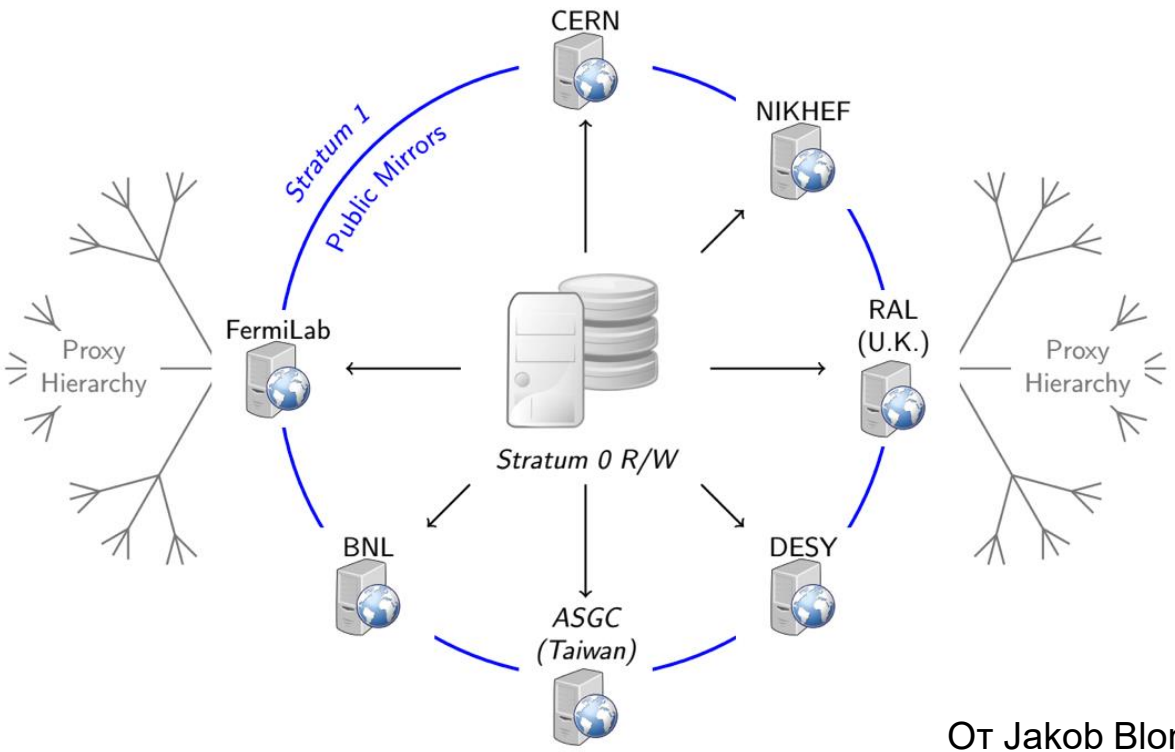


Дистрибуция на софтуера: CVMFS

- CVMFS - CERN VM FileSystem
  - Мрежова файлова система базирана на http и оптимизирана да доставя софтуера на експериментите по сайтовете бързо и сигурно
  - Новия софтуер се поставя на Stratum-0 и автоматично се репликира на всички Spratum-1
  - Най-голямата част от репликираното се поема от кешовете (Squids) на всеки един сайт
    - Компютрите на всеки сайт четат софтуер само от локалния кеш
    - В случай че локалния кеш не работи, компютрите четат от следващото ниво

- Всички стандартни сайтове в АТЛАС използват CVMFS
  - Трябва връзка с външния свят - не работи при повечето суперкомпютри

CernVM-FS Content Distribution

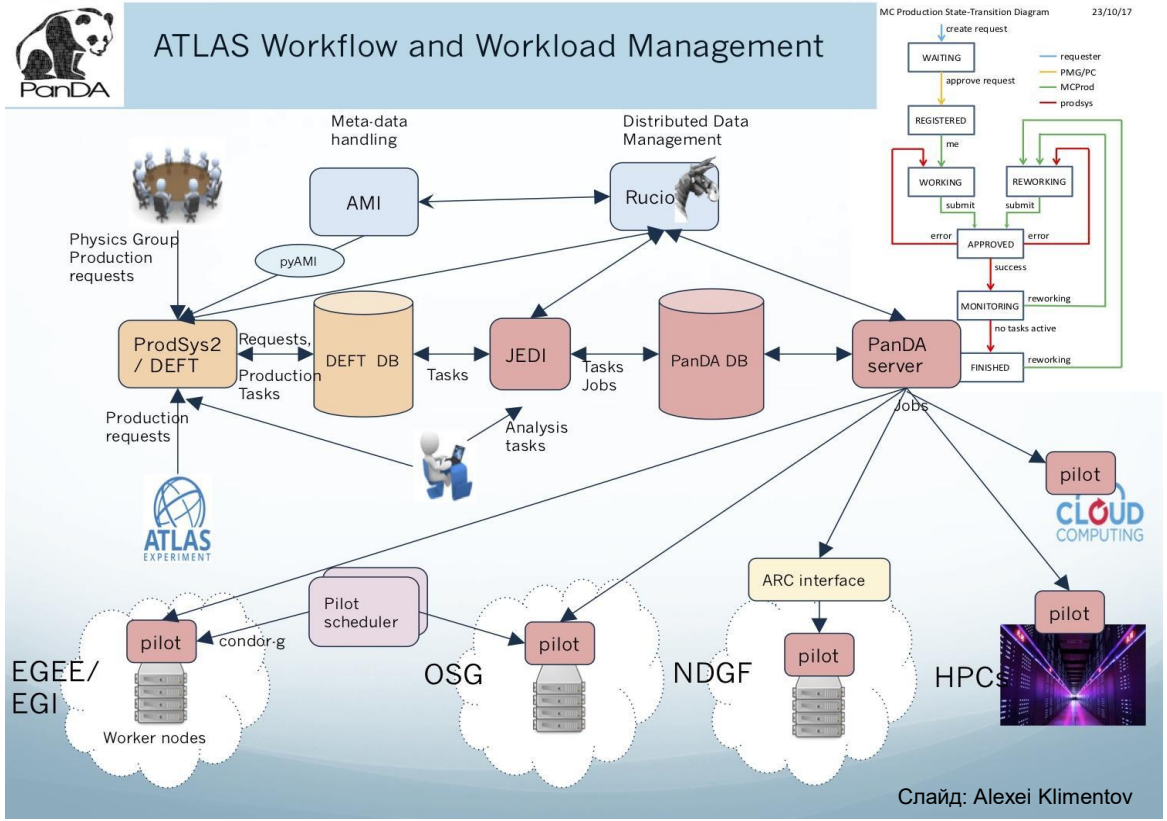


От Jakob Blomer

Изчисления

(как)

Изчислителни задачи: от А до Я



Дефиниции

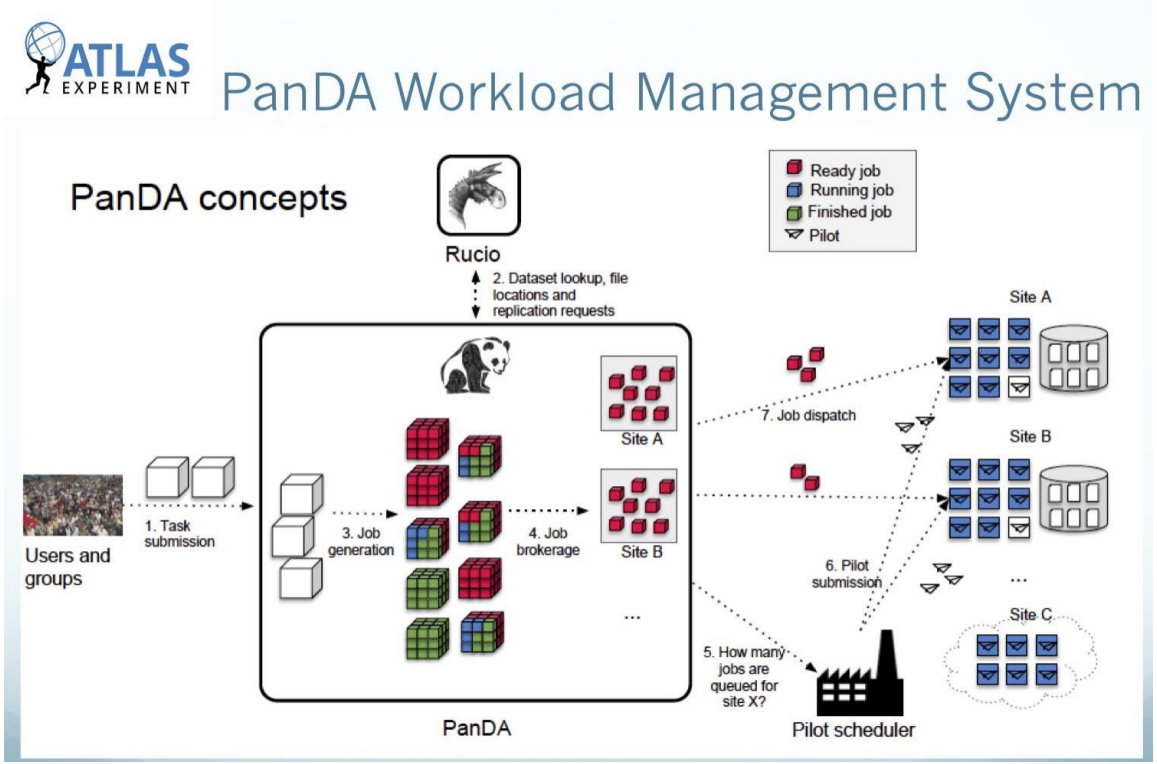
Разпределени изчисления:

- Задача (Job) - какво трябва да се сметне на един “компютър”
- Пилот - малка програма която се изпраща на сайта и вика реалната задача от системата за задачи
- HPC - Суперкомпютър

АТЛАС:

- ProdSys2 / DEFT - системата която взема една заявка (request) и я разпределя
- Rucio - системата за данни - записване, изтриване, преместване, репликиране и т.н.
- Задание (Task) - какво трябва да се пресметне от данни получени при еднакви условия

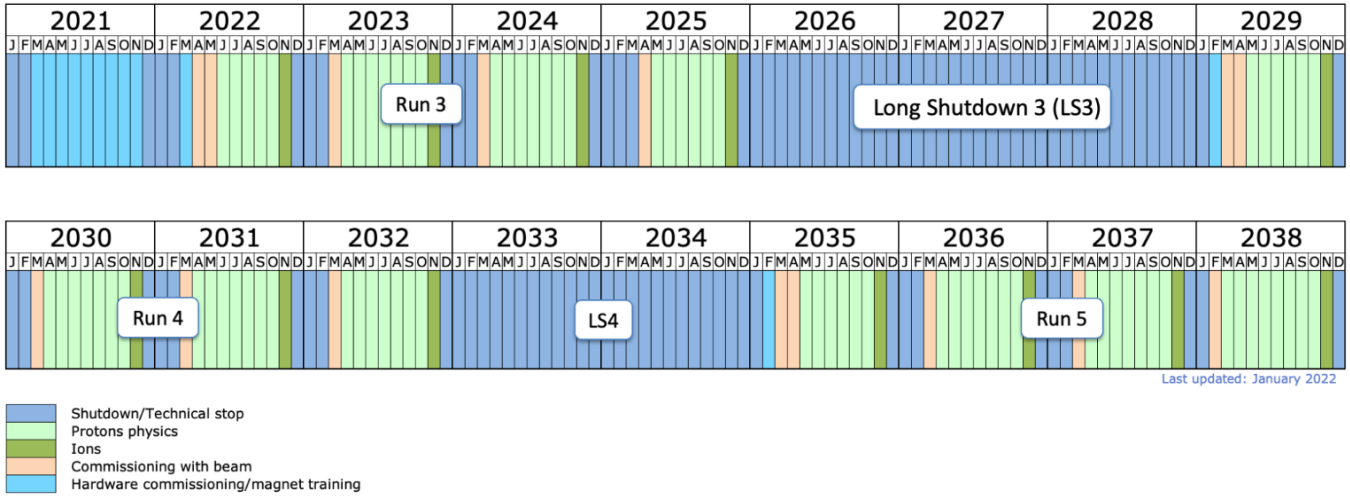
Система за разпределяне на задачи в АТЛАС

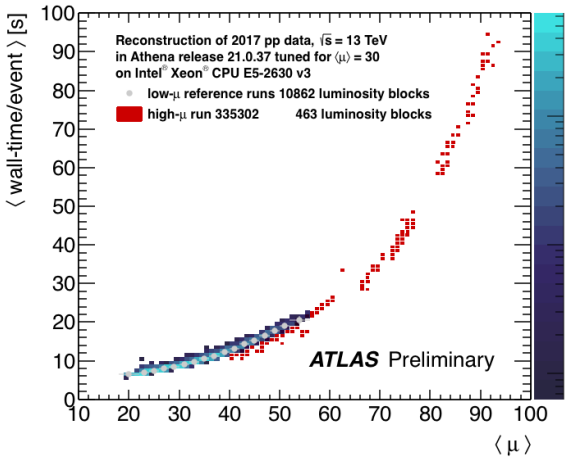


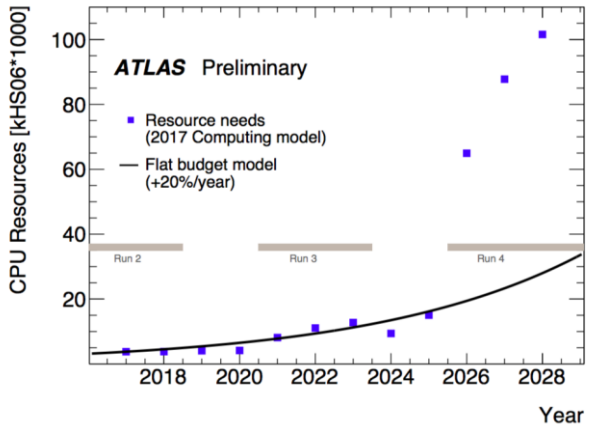
Слайд: Alexei Klimentov

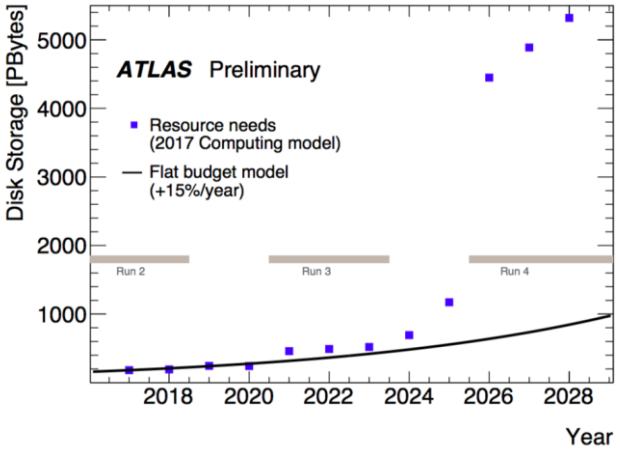
Бъдещето

HL-LHC


2021
2022
2023
2024
2025
2026
2027
2028
2029
Run 3
Long Shutdown 3 (LS3)
2030
2031
2032
2033
2034
2035
2036
2037
2038
Run 4
LS4
Run 5
Last updated: January 2022
Shutdown/Technical stop
Protons physics
Ions
Commissioning with beam
Hardware commissioning/magnet training


Reconstruction of 2017 pp data, √s = 13 TeV
in Athena release 21.0.37 tuned for ⟨μ⟩ = 30
on Intel® Xeon® CPU E5-2630 v3
low-μ reference runs 10862 luminosity blocks
high-μ run 335302 463 luminosity blocks
ATLAS Preliminary
⟨ wall-time/event ⟩ [s]
⟨ μ ⟩


ATLAS Preliminary
Resource needs
(2017 Computing model)
Flat budget model
(+20%/year)
Run 2
Run 3
Run 4
CPU Resources [kHS06*1000]
Year


ATLAS Preliminary
Resource needs
(2017 Computing model)
Flat budget model
(+15%/year)
Run 2
Run 3
Run 4
Disk Storage [PBytes]
Year

Заключение

- Ресурсите и данните на физиката на физиката на високите енергии са съизмерими с най-големите в ИТ индустрията
- Без разпределени изчисления съвременната физика на високите енергии не е възможна
- Изискванията към инфраструктурата ще се умножат приблизително 10 пъти за следващите 10 години
- Вероятни решения:
  - Оптимизиране на софтуера
  - Нови технологии - машинно обучение, аналитика, искуствен интелект
  - Оптимизиране на използването на хардуера - кеш сайтове, директно четене от ленти,
  - Нов тип ресурси - суперкомпютри, облаци, нови архитектури (следващата лекция)

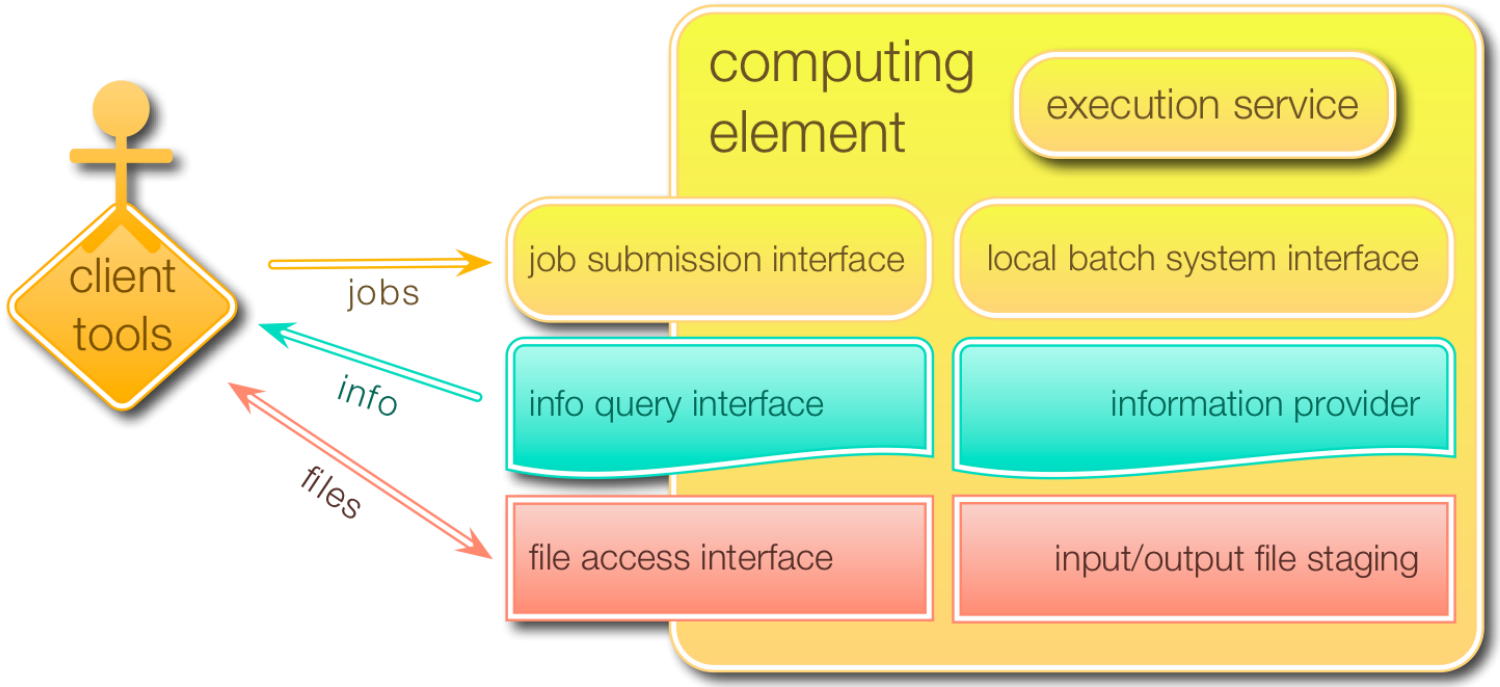
Какво не споменахме..

- Как решаваме кой колко ресурси да получи? Как налагаме и контролираме разпределението?
- Каква е мерната единица за ресурси (HEPSPEC)?
- Как се разпределят изчислителните задачите в един сайт/компютърен център?

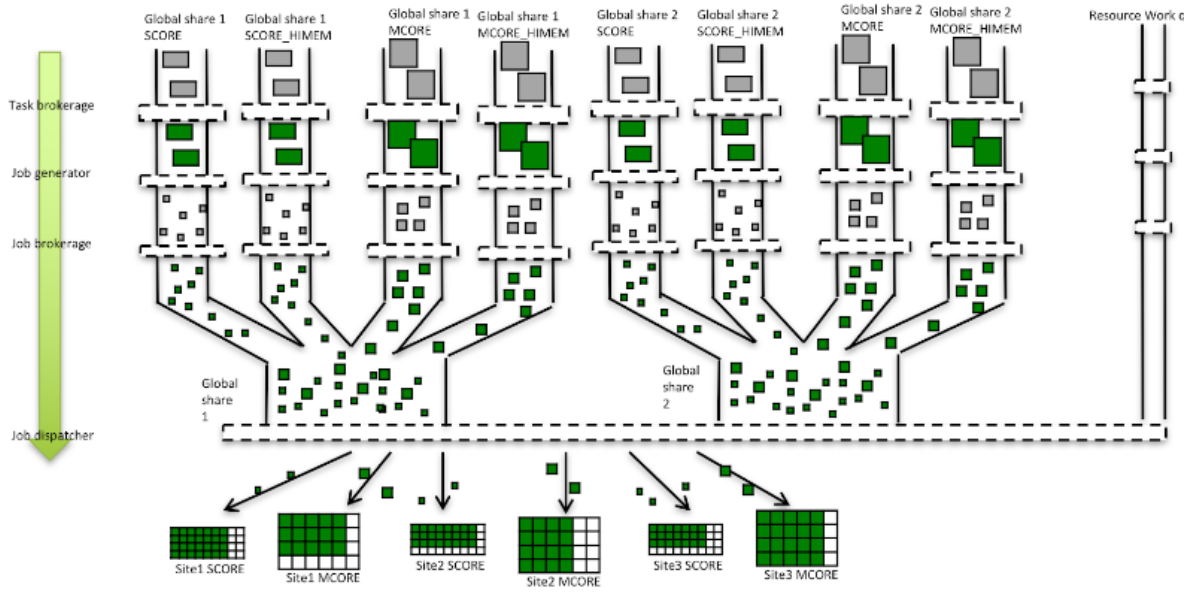
Въпроси?

Подробности

CE (Computing Element)



Global Shares



Технологии в ЦЕРН