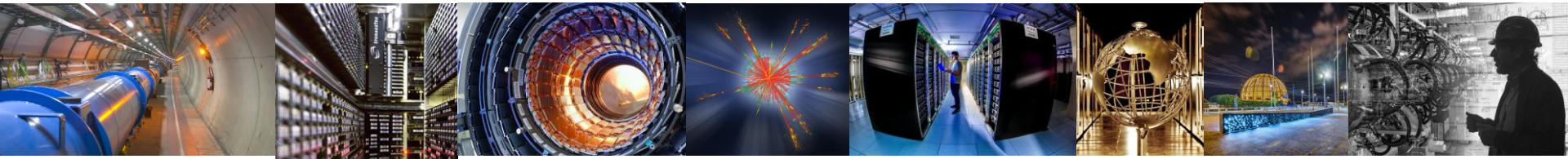
ИТ в ЦЕРН

Разпределени изчисления в ATLAS



Др. Иван Глушков
Тексаски университет / АТЛАС
Българска учителска програма
ЦЕРН, юли 2019

ИТ в ЦЕРН

- Обслужва ~15000 потребители в цял свят
  - Физици, инжинери, програмисти, администрация, финансисти
  - Поддържа пълен набор от ИТ решения за съответните заддачи
  - (когато е необходимо) Разработва специфични за ЦЕРНрешения
  - Оптимална цена
  - Изследване на нови насоки в ИТ

Технологии в ЦЕРН..

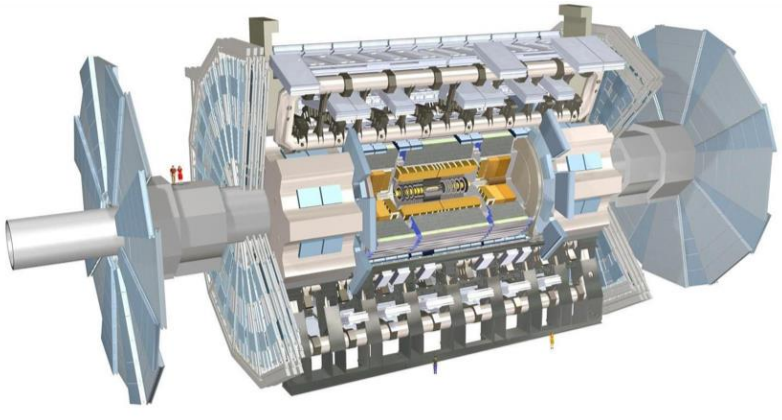
... за които ще си говорим само ако вие искате

Разпределени изчисления

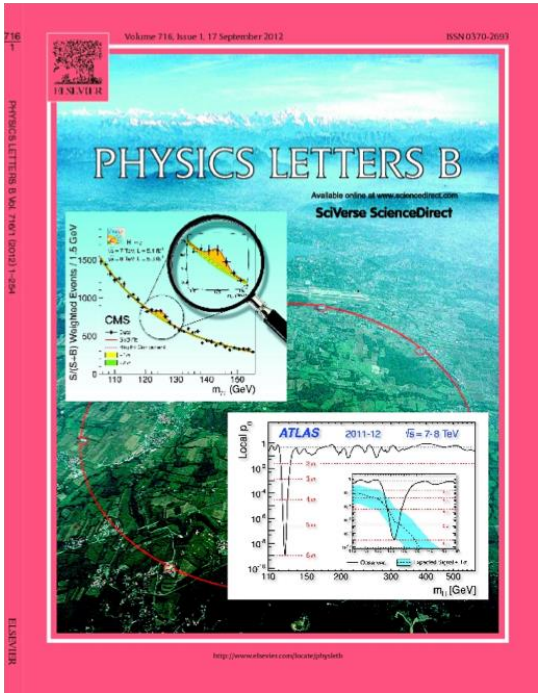
(в АТЛАС)

Какъв е проблема?

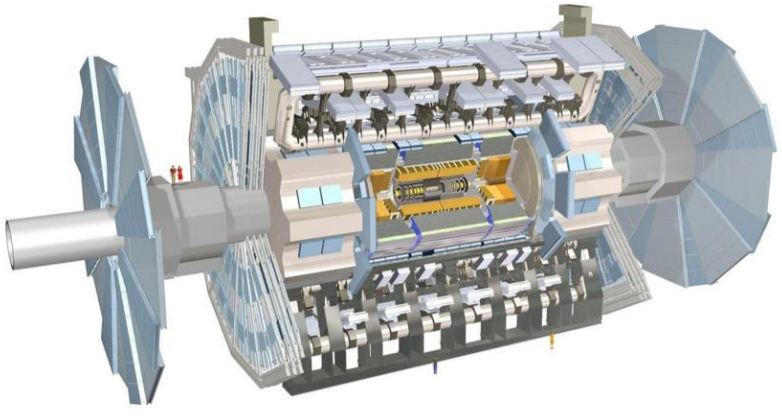
От ATLAS до научна публикация



Как?

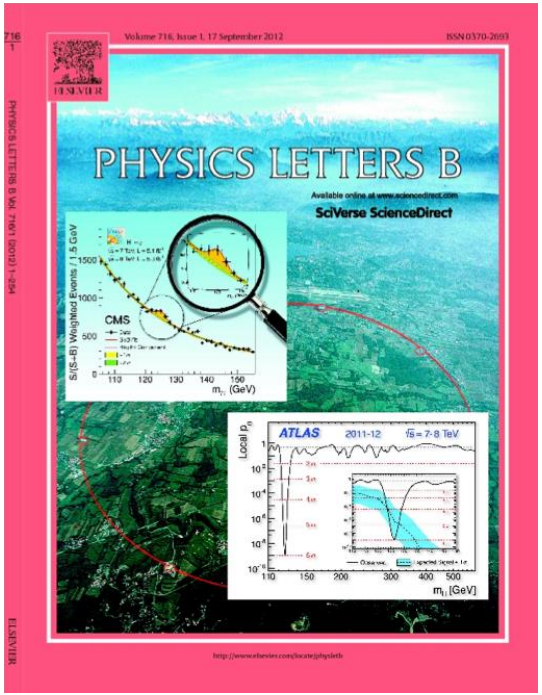


От ATLAS до научна публикация

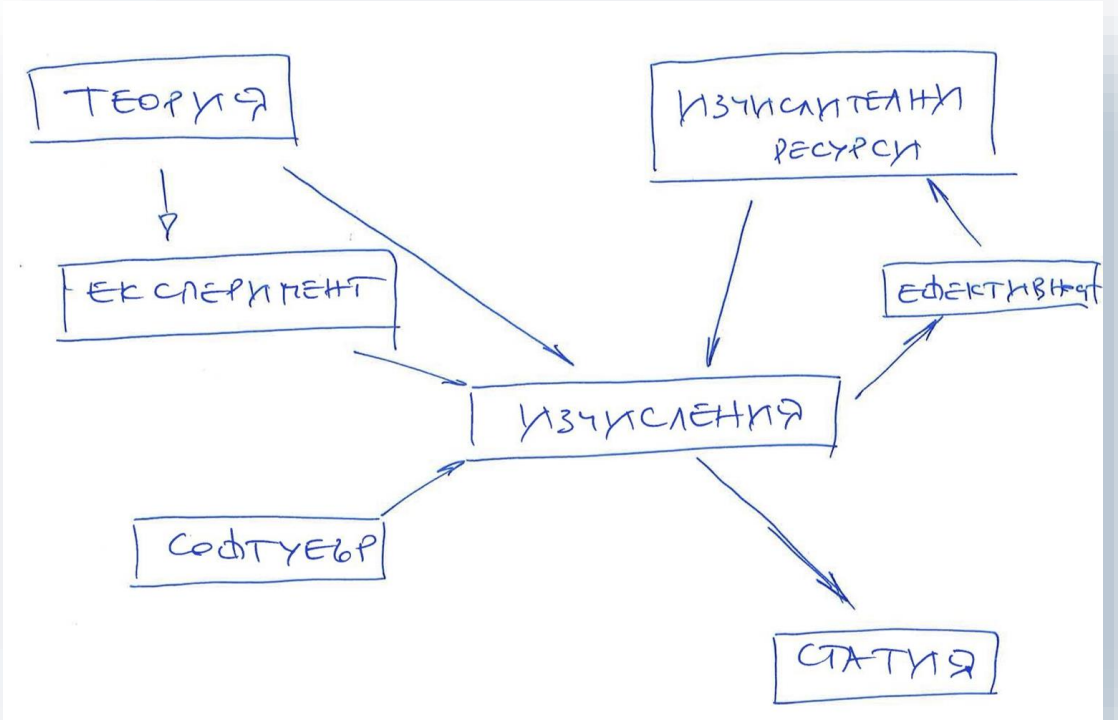


Как?

.. и бързо моля

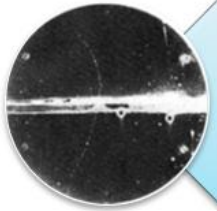


Парадигма на науката



Какво е нужно за откритие?

(обработка на данни)

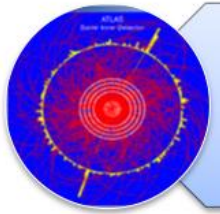


През 1930-те

- ~2 учени от една държава
- Лист и химикал

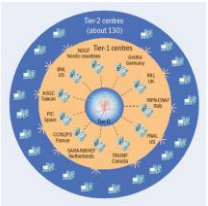
През 1970-те

- ~200 учени ~10 държави
- Мейнфрейм



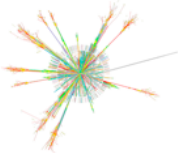
Днес

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

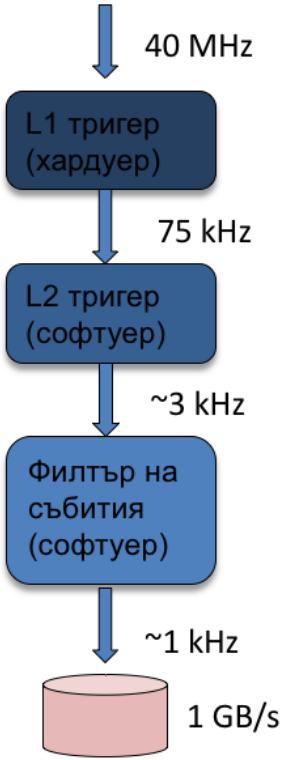


Данните

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'
```

uery Result

SQL | All Rows Fetched: 1 in 68,69 seconds

	SUM(BYTES)/1000/1000/1000/1000/1000
1	58.43299170582995

58 PB!

КАК?!

58 PB!?

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

Колаборация и оптимизация

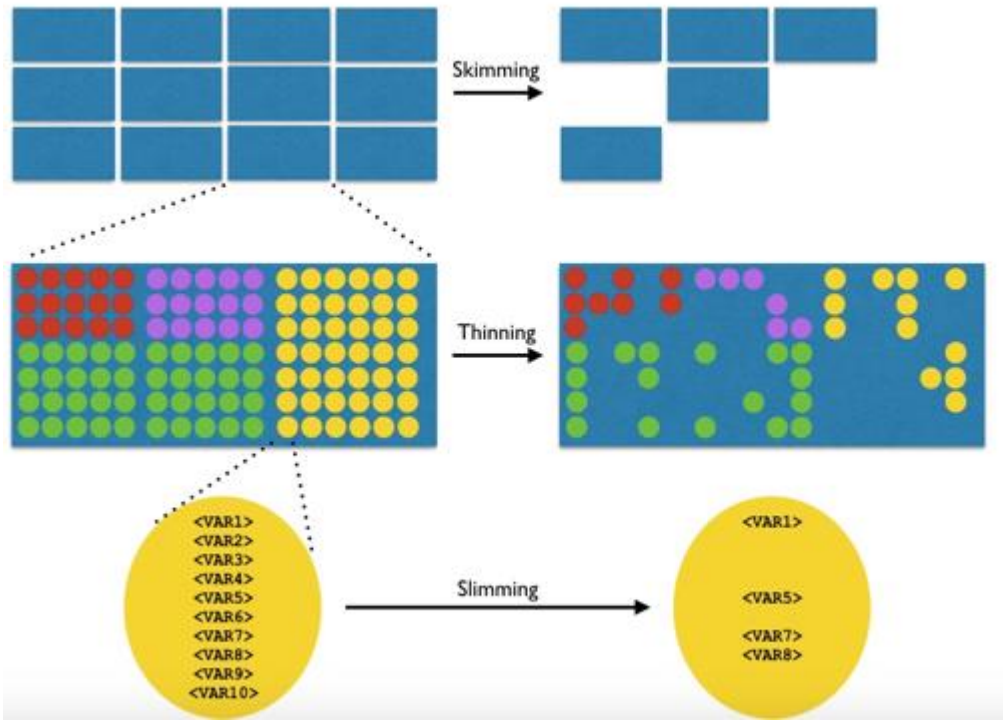
58 PB!?

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)



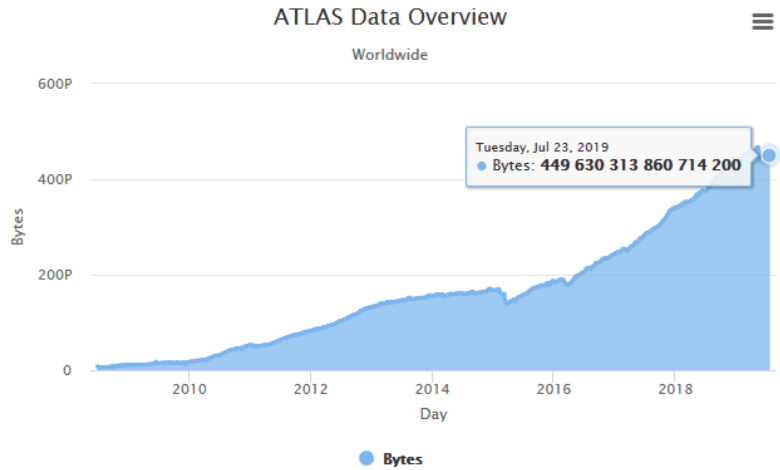
Графика: James Catmore

Премахване на ненужната информация

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

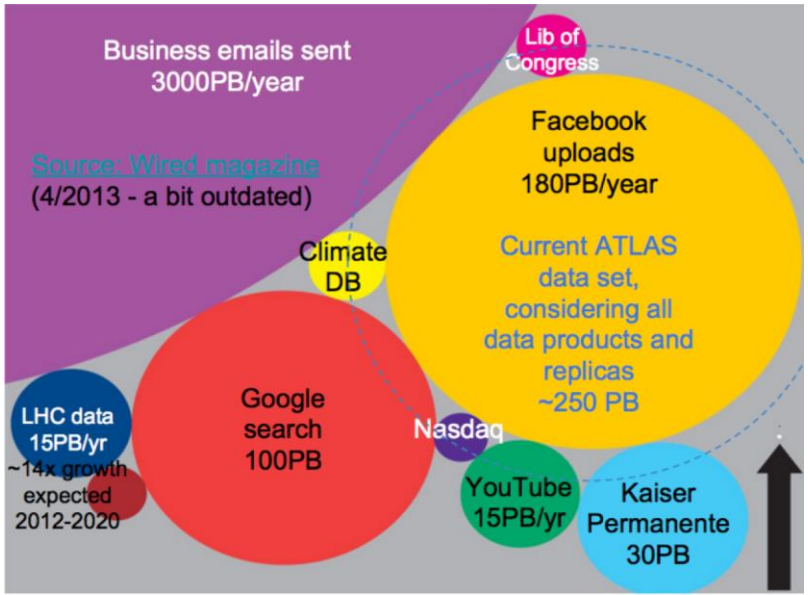
Истина и теория

Всичко общо..



450 PB!

(420 PB преди 6 месеца)



Изчисления

(колко и какви)

Колко изчислителни ресурси ни трябват?

“Колкото повече, толкова повече!”

Мечо Пух, 1966

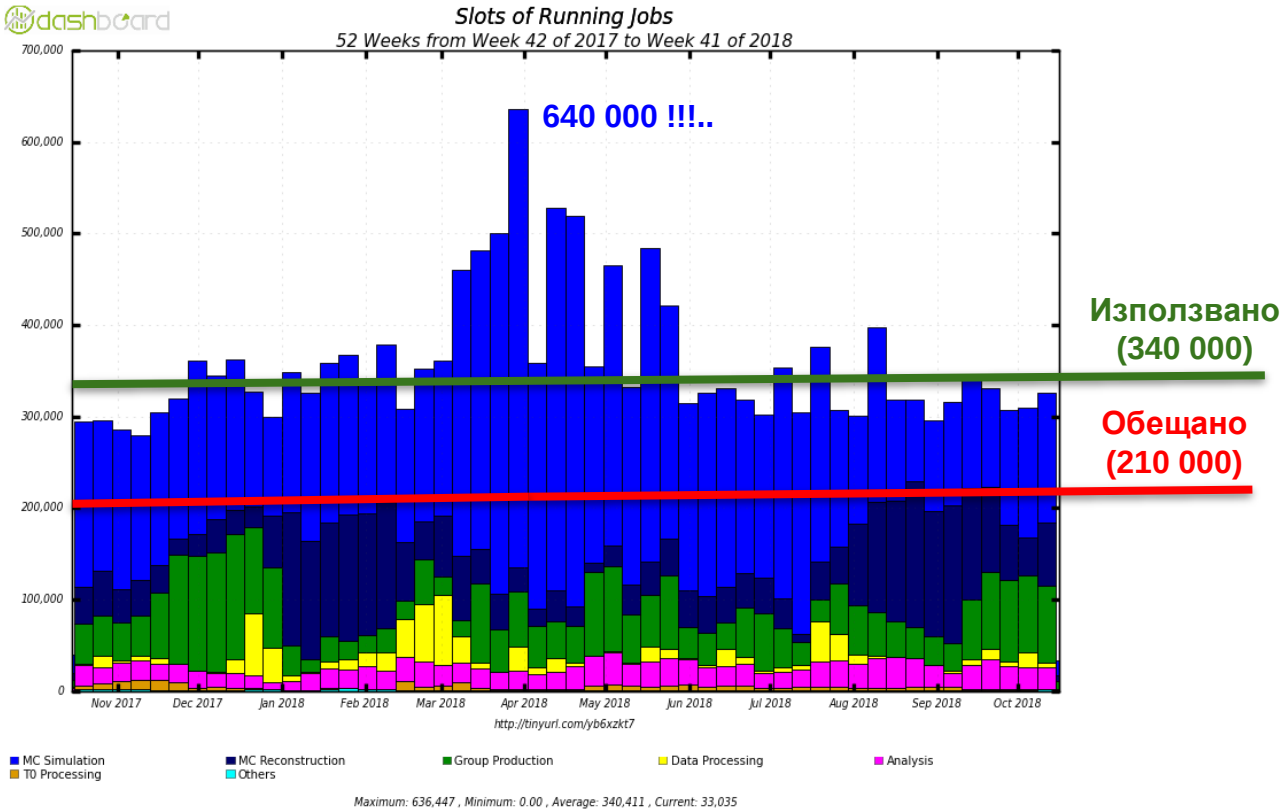
“Колкото повече, толкова повече!”

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?



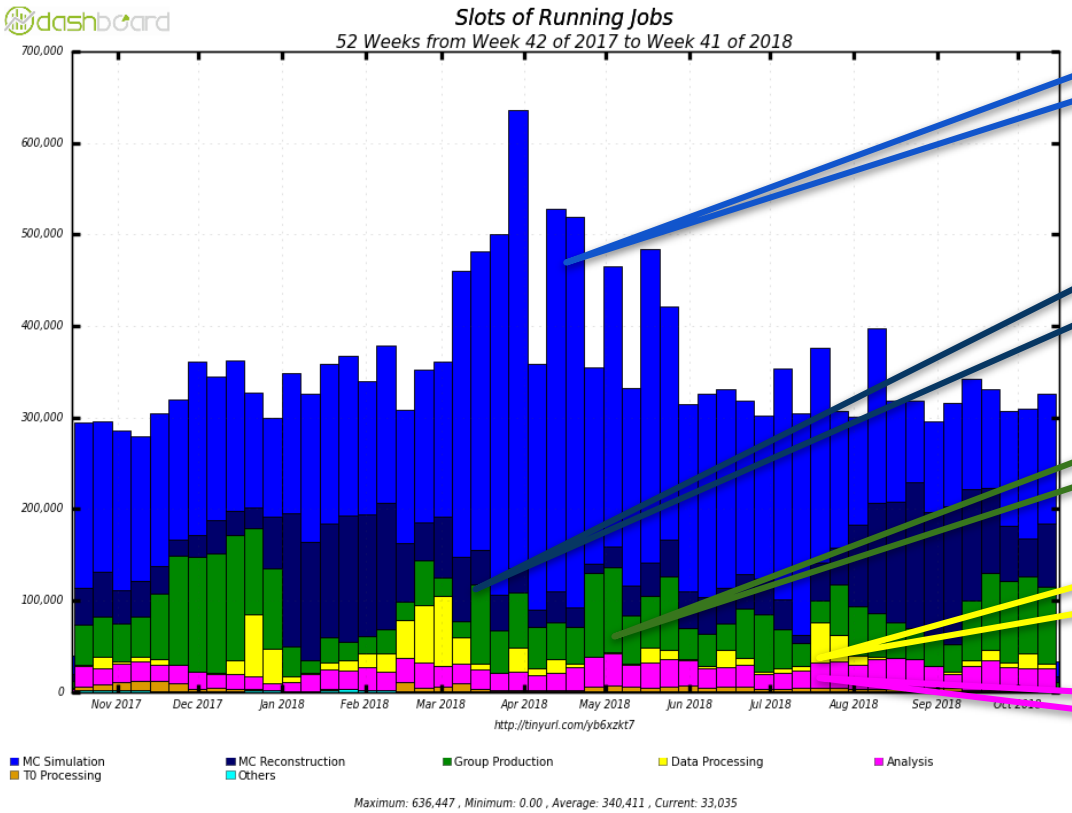
Речник

- 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?



Детекторна Симулация

Реконструкция

Пресяване

Преработка

Анализ

Изчисления

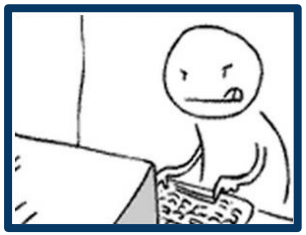
(къде)

Разпрпределени изчисления в ATLAS: Обзор

- Експерти: > 200 (в различна степен на заетост)
  - Системата не може да работи без тях
  - Разработка и интеграция на нововъведения
  - Ескплоатиране / 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
  - 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

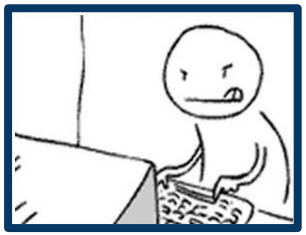
Система за разпределяне на данни (Rucio)

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

Изкам да изпълниш моя код върху данните X

Система за разпределяне на задачи (PanDA)

Задачи

ГРИД

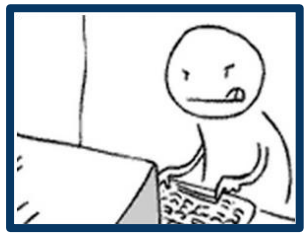
ATLAS сайт

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

Изкам да изпълниш моя код върху данните X

Задачи

Резултати?

Резултати / Още не е готово

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.

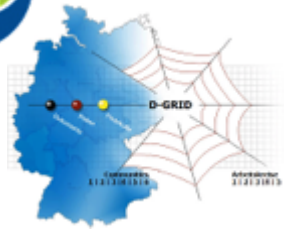
Какво е GRID II

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

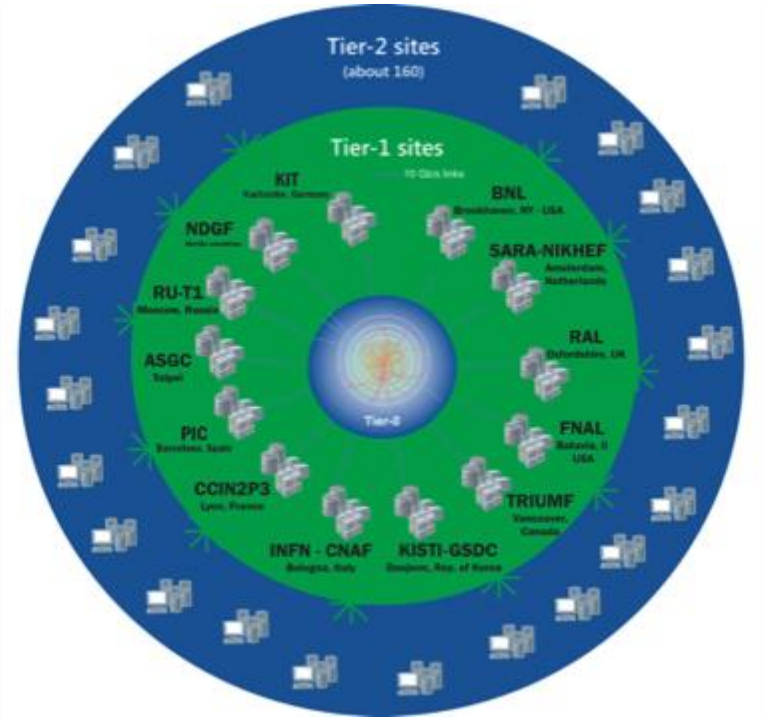
Какво е GRID III

- 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

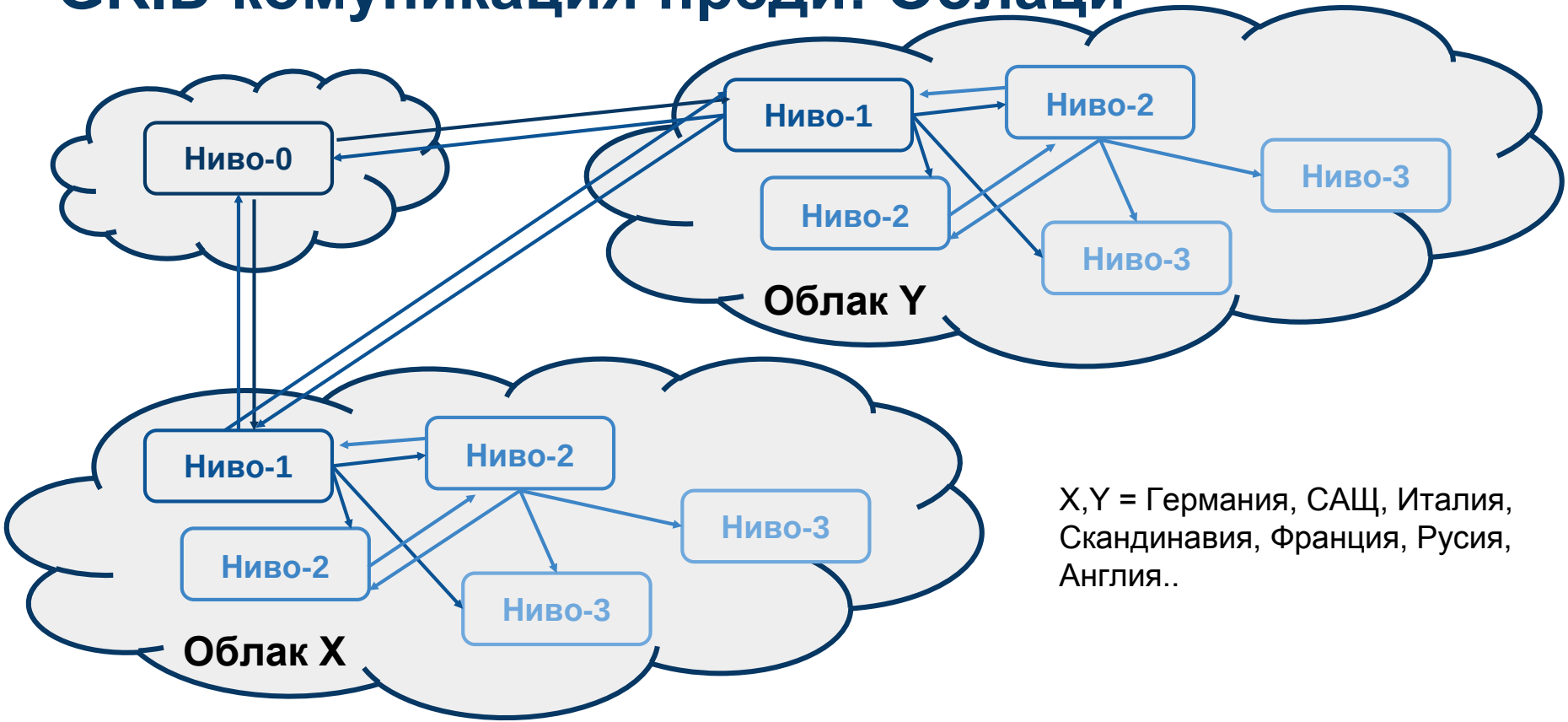
- Всеки сайт (обикновено) се състои от ресурси за съхранение на данни + изчислителни мощности
  - Squid proxy cache - за резпространяване на софтуера и “detector conditions” база данни

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

ATLAS Облак

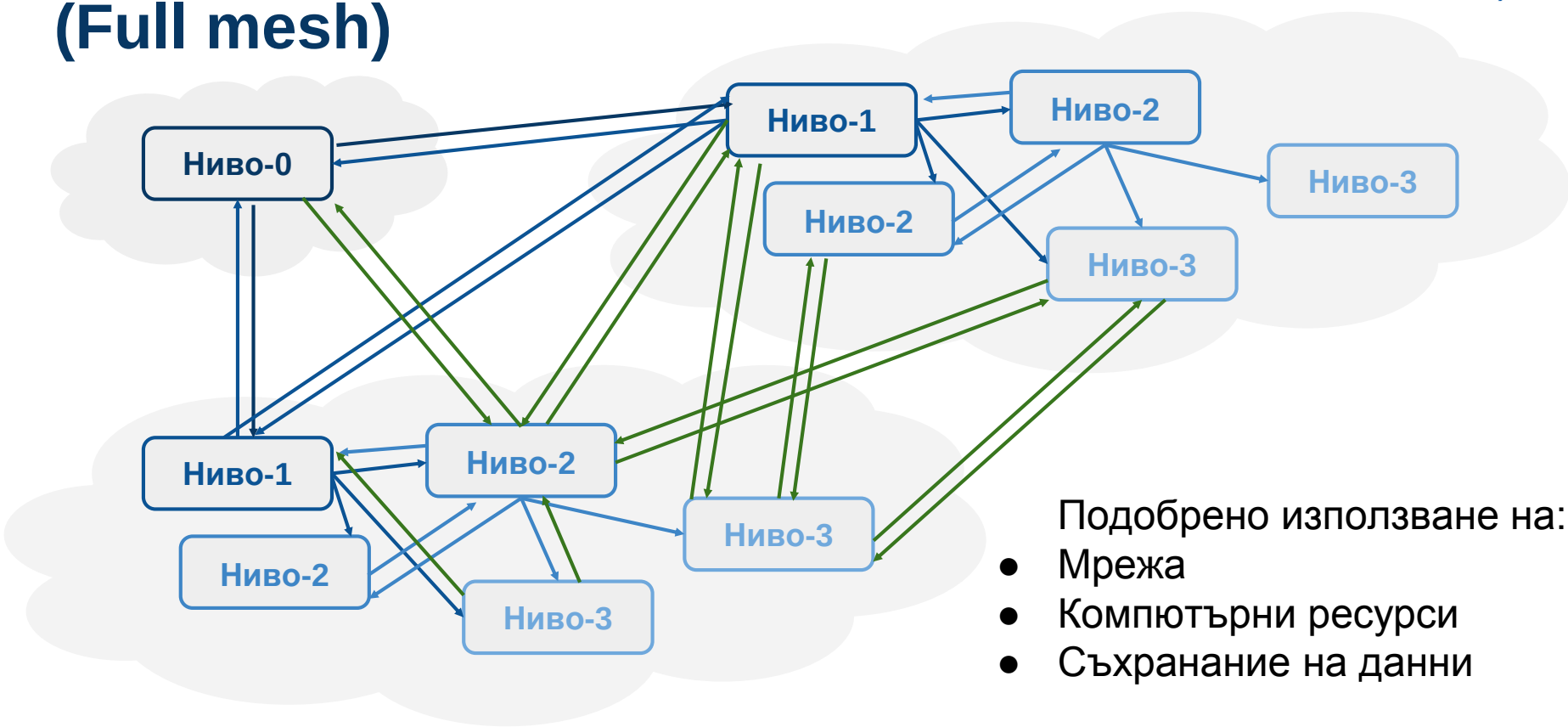
- Няма нищо общо с “облачни изчисления” (AWS, Google Cloud, и т.н.)
- Логическа група от сайтове
  - Един сайт от ниво-1 и няколко ниво-2 и ниво-3
  - Или в един географски регион, или от един източник на финансиране
- Поддръжката е предоставена от екипите на отделните облаци
  - Близо до сайтовете и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

GRID комуникация преди: Облаци



X,Y = Германия, САЩ, Италия, Скандинавия, Франция, Русия, Англия..

GRID комуникация сега: Тотална мрежа (Full mesh)



Подобрено използване на:

- Мрежа
- Компютърни ресурси
- Съхранание 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

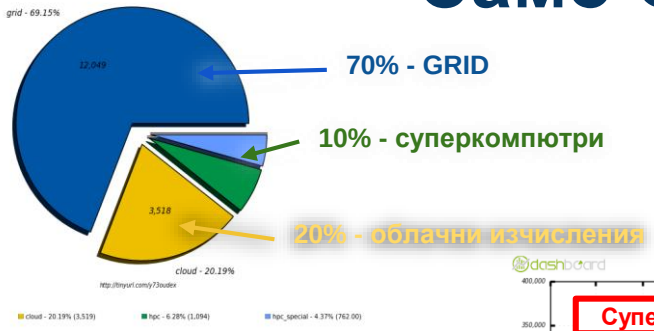
- Централни услуги в ЦЕРН (стабилност)
  - Разпределяне на задачи - PanDA
  - Разпределяне на данни - Rucio
  - База данни - Oracle
  - Прехвърляне на файлове - FTS (WLCG)
  - Кешова инфраструктура - Frontier
  - Разпределяне на софтуер - CVMFS - Stratum 0 (ниво-0)

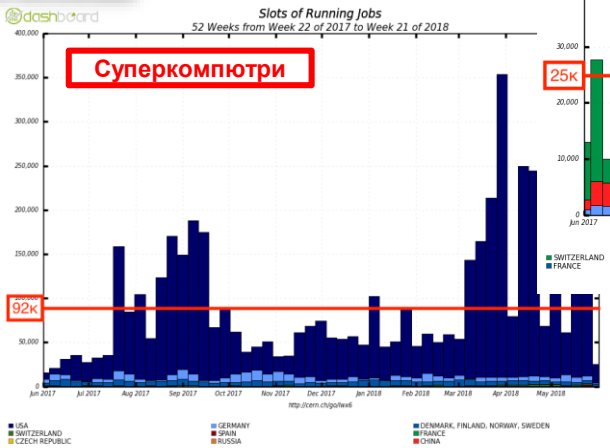
Роля на на нивата: Нива 1, 2, 3

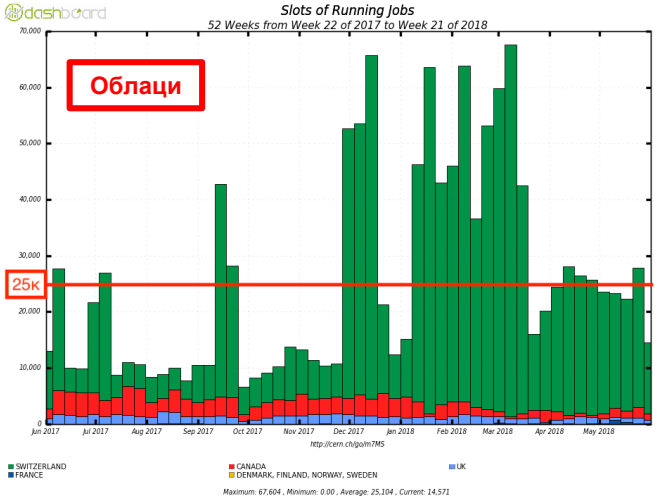
- Нива 1 и 2 - договорени (pledged) компютърни ресурси и дискови и лентови носители
  - Съхранение копие на първичните данни и на вторични данни
  - Разлики между ниво-1 и ниво-2
    - Лентови носители (само в ниво-1)
    - Поддръжка
      - 24x7 за ниво-1
      - В рамките на работния ден - ниво-2
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

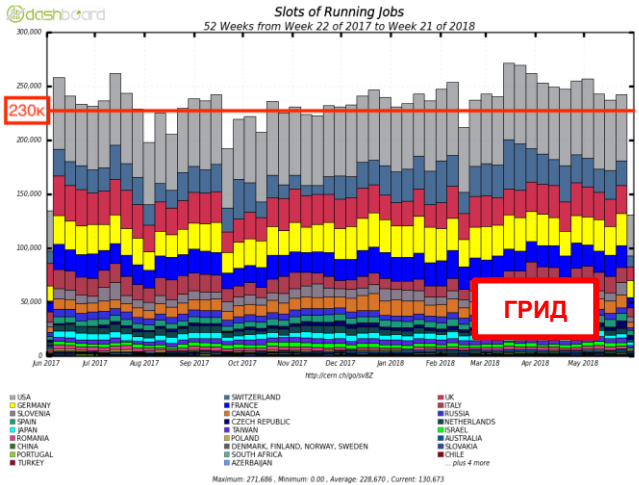
Симулации в ATLAS

Само GRID ли?









Разликата

HTC: High-Throughput Computing (GRID)

- Какво е?
  - Голяма изчислителна мощност за дълго време.
- За какво служи?
  - Ефективното изпълнение на много и слабо свързани задачи.

HPC: High-Performance Computing (Суперкомпютри)

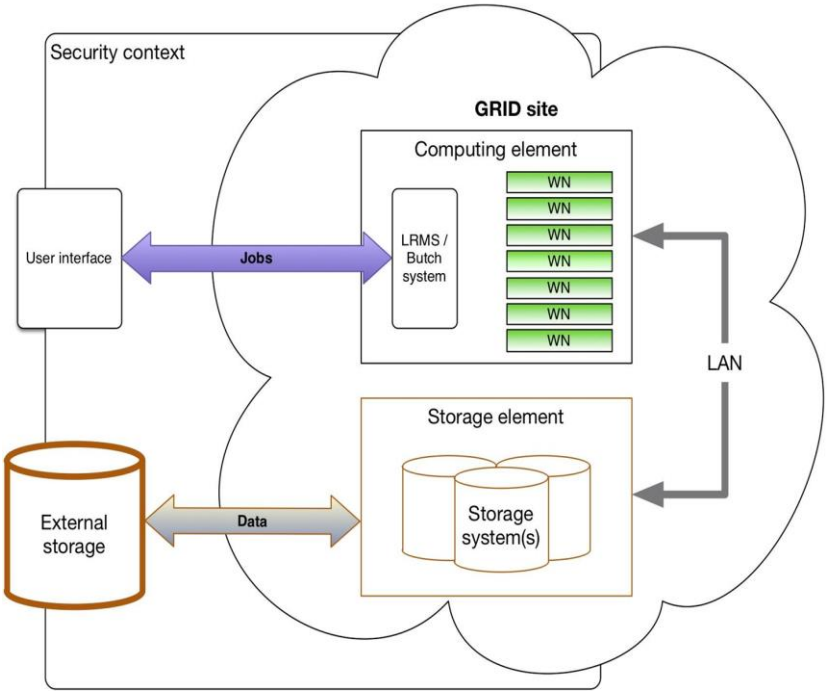
- Какво е?
  - Голяма изчислителан мощност за ограничено време.
- За какво служи?
  - Ефективно изпълнение на много, тясно свързани задачи.

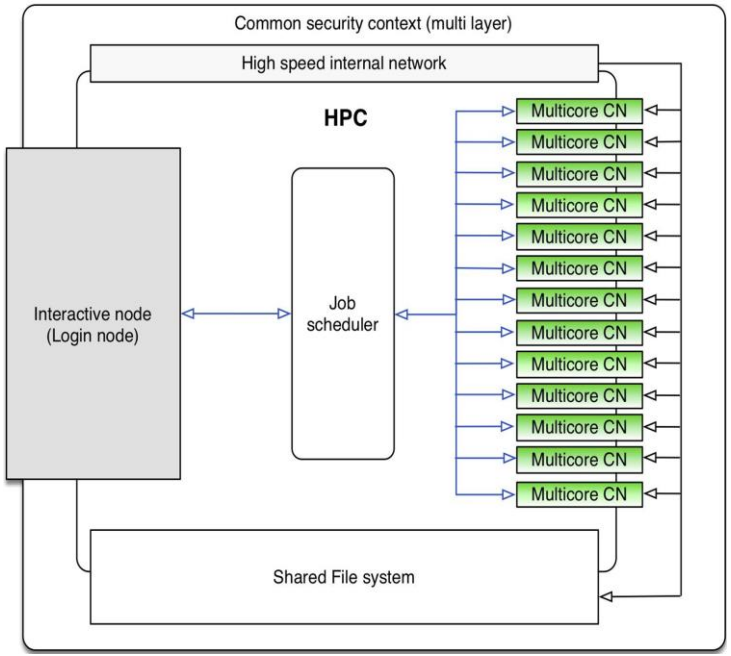
Облачни изчисления

- Какво е?
  - Виртуални машини - колкото и каквито са нужни
- За какво служи?
  - Рарантирани ресурси в сигурна и изолирана среда без нужда от поддръжка на хардуер. API достъп

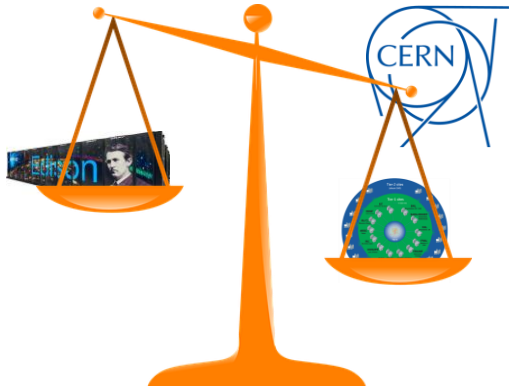
Суперкомпютри

GRID и Суперкомпютри





GRID и Суперкомпютри



- GRID се състои от компюютърни клъстери
- По-голяма част от изчисленията в другите области на науката стават в суперкомпютри.
- Разлики: суперкомпютри <> GRID:
  - Нишки: много (~100 000 нишки) <> една
  - Работа на отделния компютър: част от задача <> една задача
  - Входно - изходни операции: малко <> много
  - Брой на файлове: малко <> много
  - Потребителска идентификация: логин/парола (еднократна!) <> сертификат
  - Процесори / операционни системи: много <> няколко
  - Директна връзка с интернет: не <> да
  - Местонахождение едно <> много

Суперкомпютри изпозвани от АТЛАС

Най-големите...

Titan (САЩ, top500: 7)
Cray XK7
Opteron 6274 16C 2.2 GHz
560640 ядра / 694 TB памет / 40 PB диск
Производителност: 27.11 PFLOP/сек

Theta (САЩ, top500: 21)
Cray XC40
Intel Phi 7230 64C 1.3GHz
280320 ядра / 914 TB памет / 11 PB диск
Производителност: 11.66 PFLOP/сек

Edison (САЩ, top500: 104)
Cray XC30
Intel Xeon E5-2695v2 121C 2.4GHz
133824 ядра / 357 TB памет / 7.56 PB диск
Производителност: 2.57 PFLOP/сек

……..

И още много в...

САЩ, Германия, Швейцария, Чехия, Испания, Франция, Китай, Скандинавия, Русия..

Еднаквото между всички..

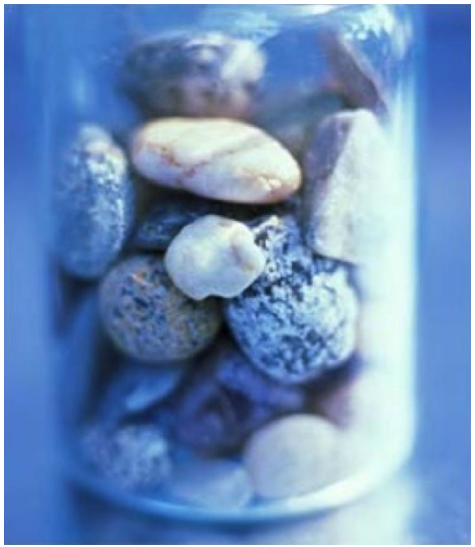
Няма такова..

Мащаб на задачите на един суперкомпютър

Вид задача	Минимален брой компютри	Максимален брой Компютри	Максимална дължина на задачата	Приоритет
1	11,250	—	24.0	15
2	3,750	11,249	24.0	5
3	313	3,749	12.0	0
4	126	312	6.0	0
5	1	125	2.0	0

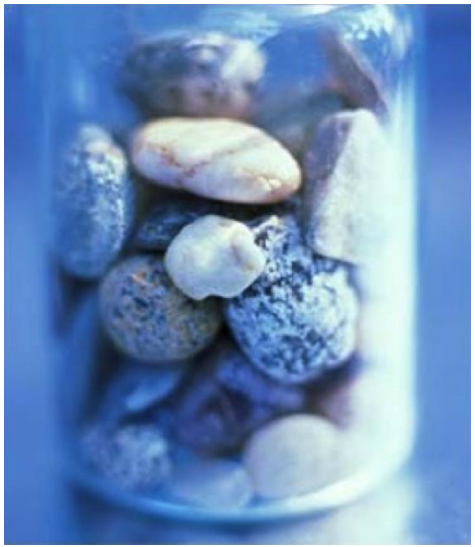
HPC: Backfill

HPC: Backfill



Източник: David Cameron

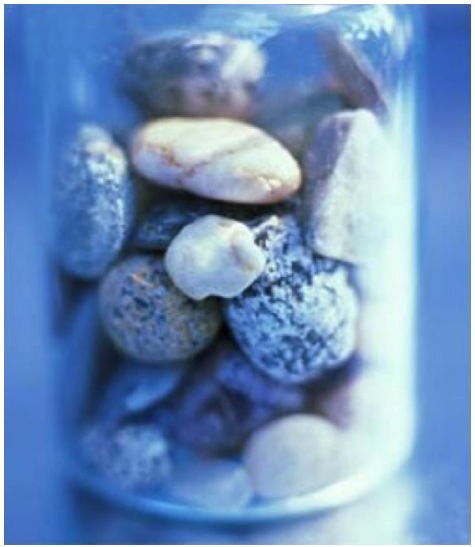
Суперкомпютри: Backfill



Източник: David Cameron

Суперкомпютри: Backfill

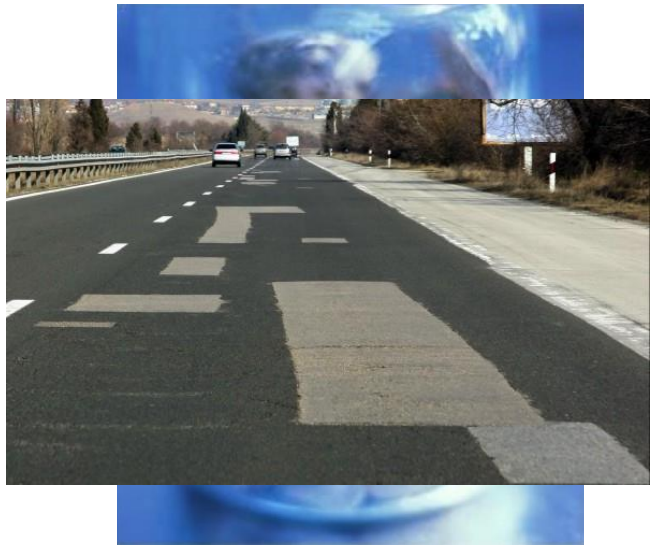
(пълним дупки всекакви)



Източник: David Cameron

Суперкомпютри: Backfill

(пълним дупки всекакви)



Източник: David Cameron

Суперкомпютри: Backfill

(пълним дупки всекакви)

Размер на дупките

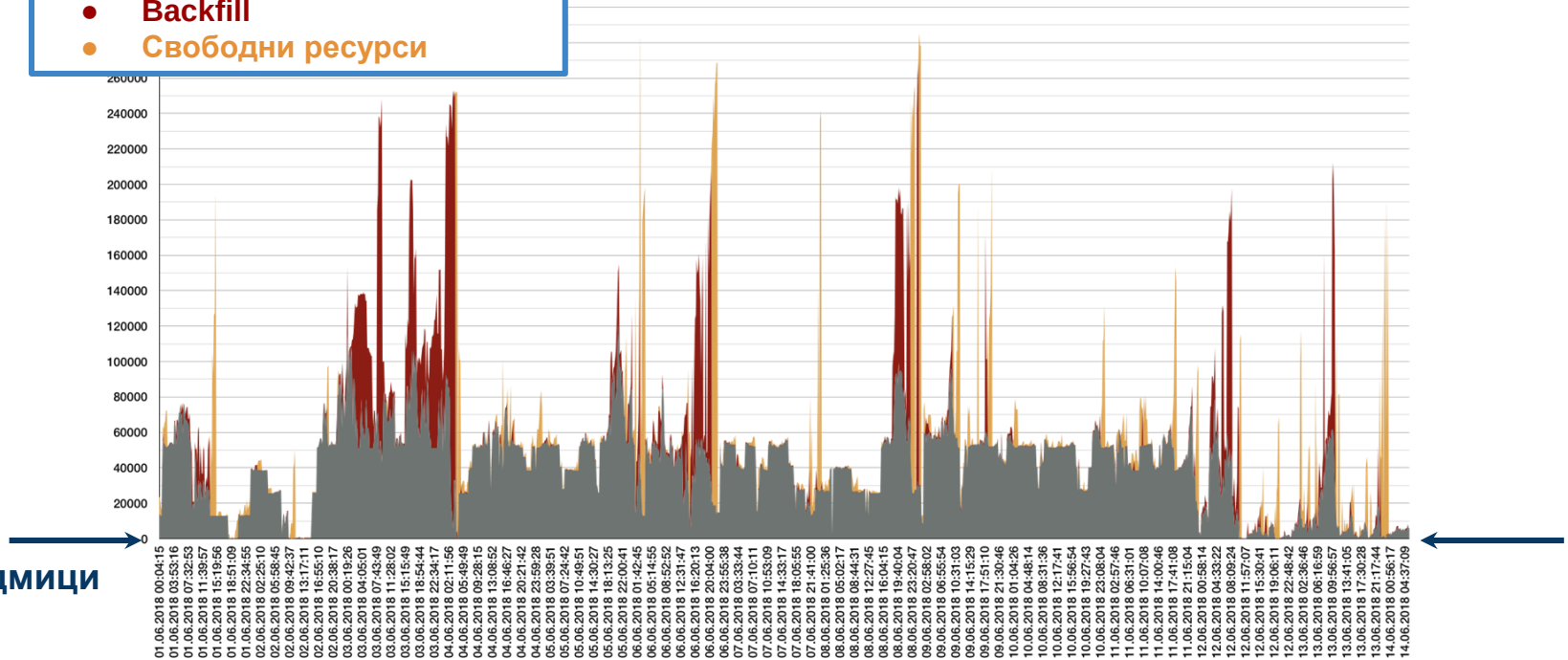
- 400M CPU*часа на година (Titan)
- Обичайна ефективност на използване на суперкомпютър във времето - 90%.
- 10% == 400M CPU*часа на година (Titan)
- Всяка неефективност е нежелана!

Запълване

- С какво?
  - Много, с различен размер, но малки задачи (т.е. - ние)
- На каква цена?
  - Безплатно
- И на края..
  - Всички живели щастливо..

Използване на Titan от ATLAS

- Изчисления по квота
- Backfill
- Свободни ресурси

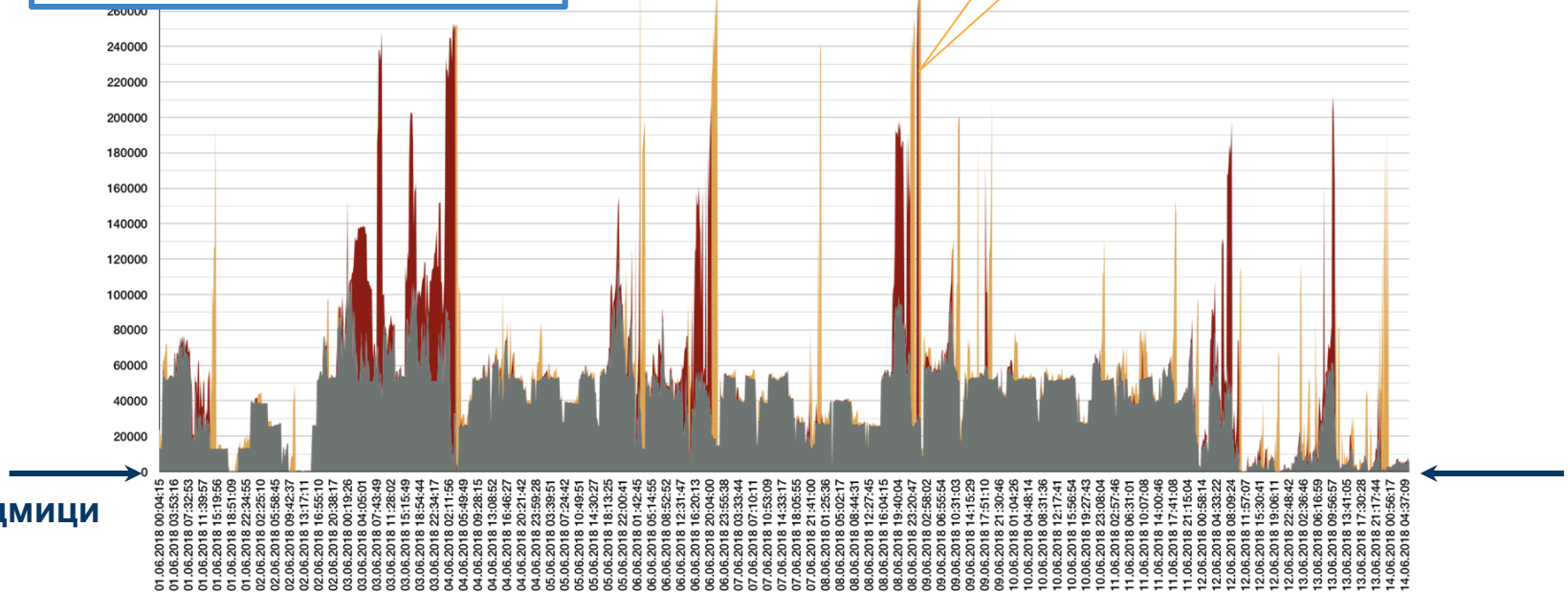


две седмици

Използване на Titan от ATLAS

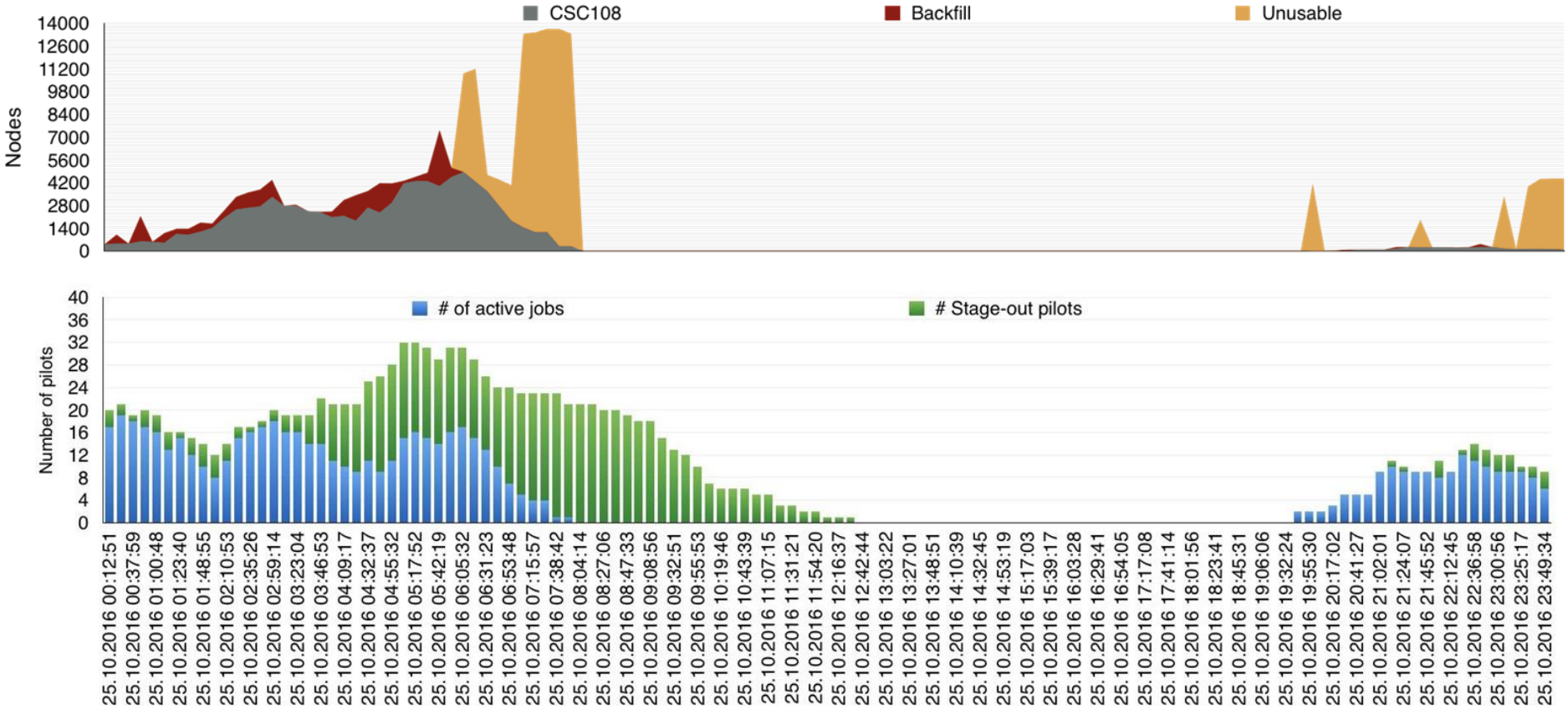
- Изчисления по квота
- Backfill
- Свободни ресурси

ЗАЩО?!



две седмици

Проблемът

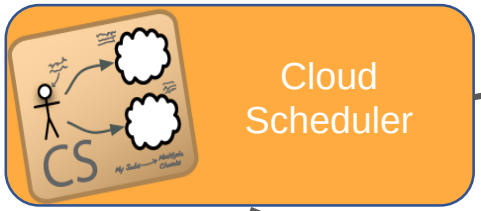

CSC108
Backfill
Unusable
Nodes
14000
12600
11200
9800
8400
7000
5600
4200
2800
1400
0
# of active jobs
Stage-out pilots
Number of pilots
40
36
32
28
24
20
16
12
8
4
0
25.10.2016 00:12:51
25.10.2016 00:37:59
25.10.2016 01:00:48
25.10.2016 01:23:40
25.10.2016 01:48:55
25.10.2016 02:10:53
25.10.2016 02:35:26
25.10.2016 02:59:14
25.10.2016 03:23:04
25.10.2016 03:46:53
25.10.2016 04:09:17
25.10.2016 04:32:37
25.10.2016 04:55:32
25.10.2016 05:17:52
25.10.2016 05:42:19
25.10.2016 06:05:32
25.10.2016 06:31:23
25.10.2016 06:53:48
25.10.2016 07:15:57
25.10.2016 07:38:42
25.10.2016 08:04:14
25.10.2016 08:27:06
25.10.2016 08:47:33
25.10.2016 09:08:56
25.10.2016 09:32:51
25.10.2016 09:55:53
25.10.2016 10:19:46
25.10.2016 10:43:39
25.10.2016 11:07:15
25.10.2016 11:31:21
25.10.2016 11:54:20
25.10.2016 12:16:37
25.10.2016 12:42:44
25.10.2016 13:03:22
25.10.2016 13:27:01
25.10.2016 13:48:51
25.10.2016 14:10:39
25.10.2016 14:32:45
25.10.2016 14:53:19
25.10.2016 15:17:03
25.10.2016 15:39:17
25.10.2016 16:03:28
25.10.2016 16:29:41
25.10.2016 16:54:05
25.10.2016 17:17:08
25.10.2016 17:41:14
25.10.2016 18:01:56
25.10.2016 18:23:41
25.10.2016 18:45:31
25.10.2016 19:06:06
25.10.2016 19:32:24
25.10.2016 19:55:30
25.10.2016 20:17:02
25.10.2016 20:41:27
25.10.2016 21:02:01
25.10.2016 21:24:07
25.10.2016 21:45:52
25.10.2016 22:12:45
25.10.2016 22:36:58
25.10.2016 23:00:56
25.10.2016 23:25:17
25.10.2016 23:49:34

Облаци

На кратко

- ATLAS използва собствени, научни и комерсиални.
- Предимства
  - Комерсиални: При нужда от много ресурси за кратко време
  - Научни: Споделяне на ресурси и цена между различни институти и различни области на изследване
  - Собствени: Бързо и контролирано усвояване на ресурси
- Недостатъци
  - Комерсиални: Скъпи

Принцип на работа



Cloud Scheduler

Доставчици на облачни услуги

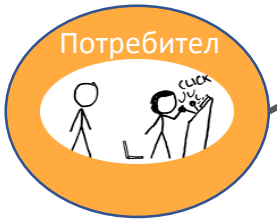
openstack CLOUD SOFTWARE

Amazon EC2

Microsoft Azure

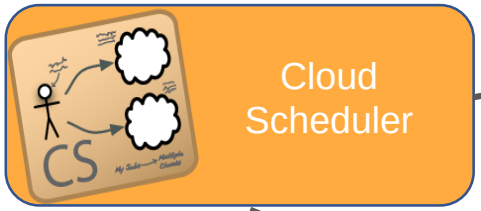
Система за разпределяне на задачи

HTCondor High Throughput Computing



Изпраща задачи

Принцип на работа



Cloud Scheduler

Нарави ми виртуални машини

Доставчици на облачни услуги

openstack CLOUD SOFTWARE

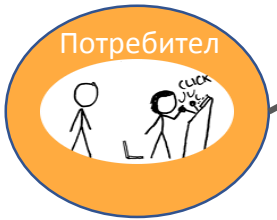
Amazon EC2

Microsoft Azure

Наблюдава опашката от задачи

Система за разпределяне на задачи

HTCondor High Throughput Computing



Изпраща задачи

Принцип на работа

Cloud Scheduler

Нарави ми виртуални машини

Доставчици на облачни услуги

openstack CLOUD SOFTWARE

Amazon EC2

Microsoft Azure

създаване

Наблюдава опашката от задачи

Система за разпределяне на задачи

HTCondor High Throughput Computing

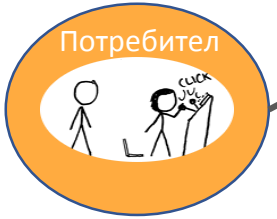
Виртуална Машина

Виртуална Машина

Виртуална Машина

Виртуална Машина

Worker



Изпраща задачи

Принцип на работа

Cloud Scheduler

Нарави ми виртуални машини

Доставчици на облачни услуги

openstack CLOUD SOFTWARE

Amazon EC2

Microsoft Azure

създаване

Наблюдава опашката от задачи

Система за разпределяне на задачи

HTCondor High Throughput Computing

Виртуална Машина

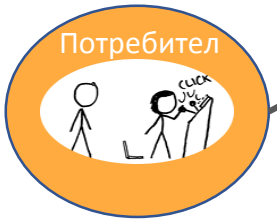
Виртуална Машина

Виртуална Машина

Виртуална Машина

Worker

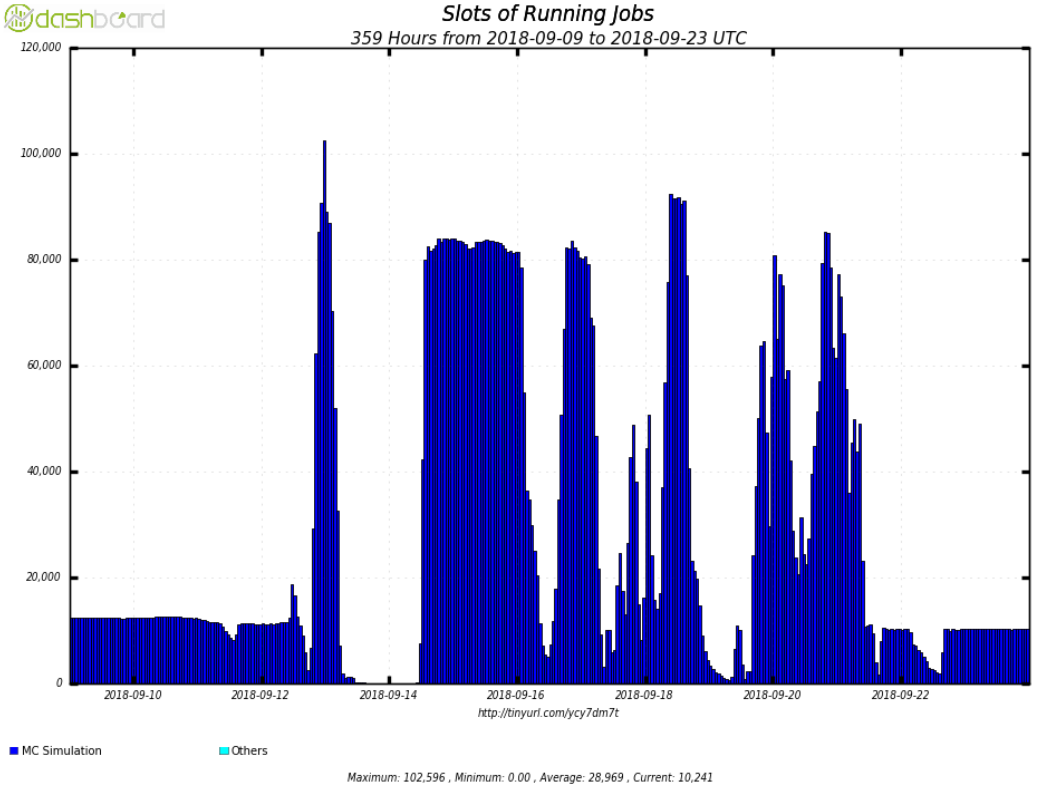
Свързване и изпълнение на задачи



Изпраща задачи

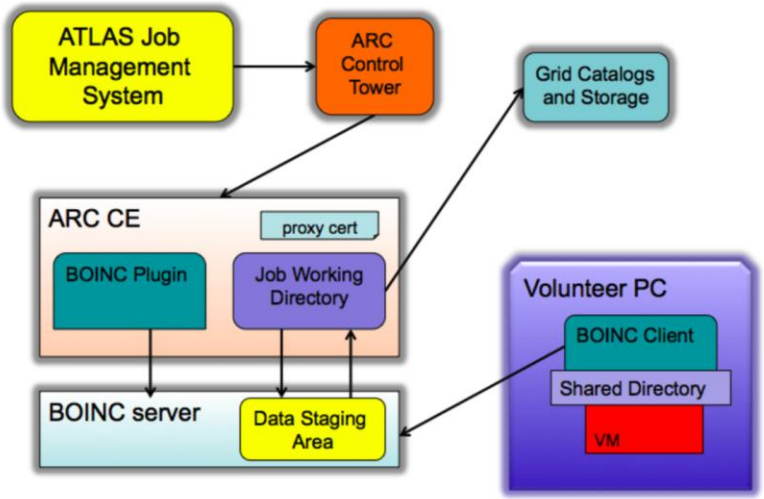
Собствени облаци - тригерна ферма

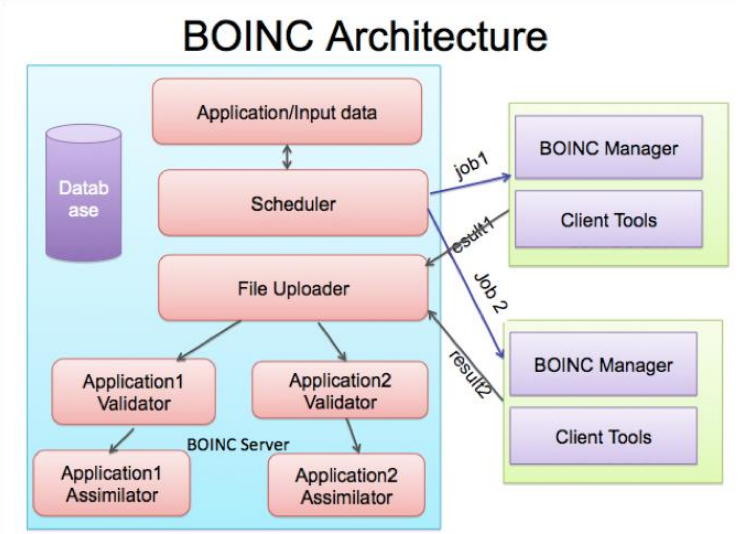
- Ресурс - 120 000 ядра
- Използване - само когато не се набират данни
- Облачно решение:
  - Бързо усвояване на ресурси
  - Няма нужда от инсталиране
  - Светкавично връщане на ресурси при нужда (kill)
- Все още не се използва максимално



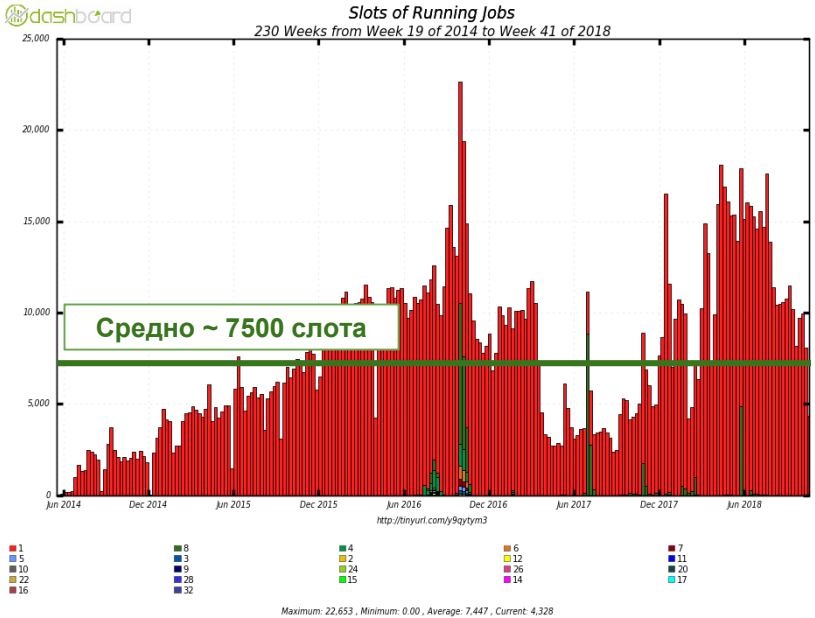
Доброволчески споделени изчисления
(Volunteer Computing)

BOINC


ATLAS Job Management System
ARC Control Tower
Grid Catalogs and Storage
ARC CE
proxy cert
BOINC Plugin
Job Working Directory
BOINC server
Data Staging Area
Volunteer PC
BOINC Client
Shared Directory
VM


BOINC Architecture
Application/Input data
Database
Scheduler
File Uploader
Application1 Validator
Application2 Validator
BOINC Server
Application1 Assimilator
Application2 Assimilator
job1
result1
job 2
result2
BOINC Manager
Client Tools
BOINC Manager
Client Tools

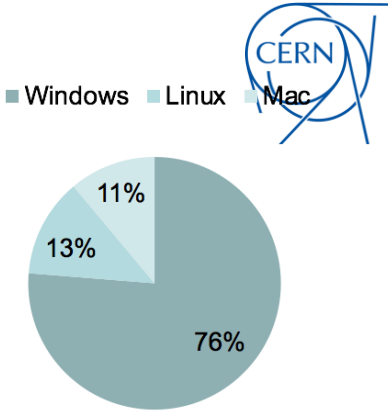
ATLAS@Home

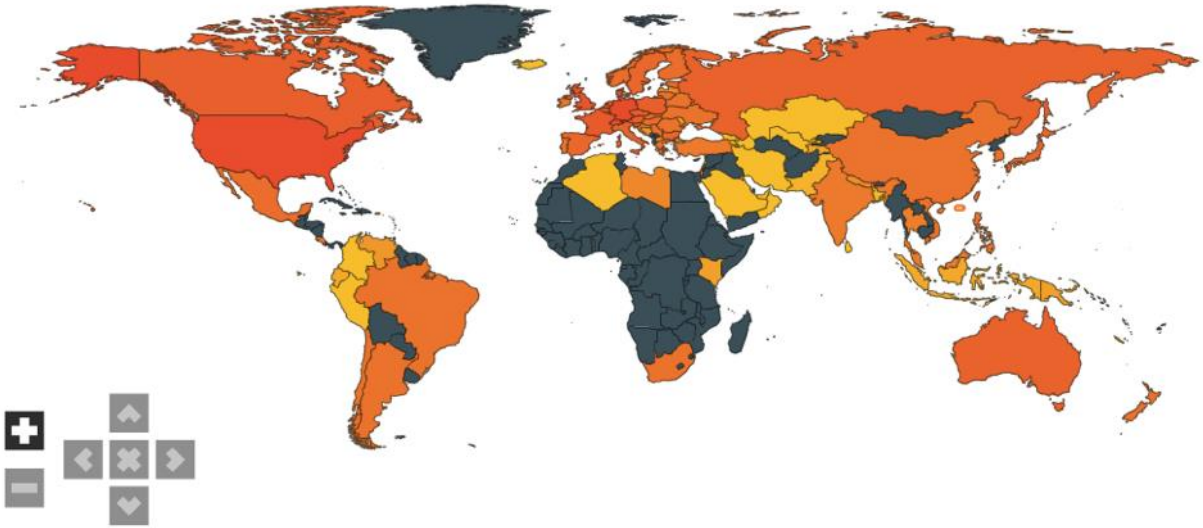


- Обработва ~ 1% от симулираните събитя в ATLAS
- Сравним с голям безплатен компютърен център
- Не точно.. доброволците очакват все пак някаква поддръжка
- Голям потенциал в настолните компютри по учреждения които не се изключват нощно време

Кой ни помага

- BOINC е огромно общество от отдадени хора
- Много от тях очастват е няколко проекта едновременно



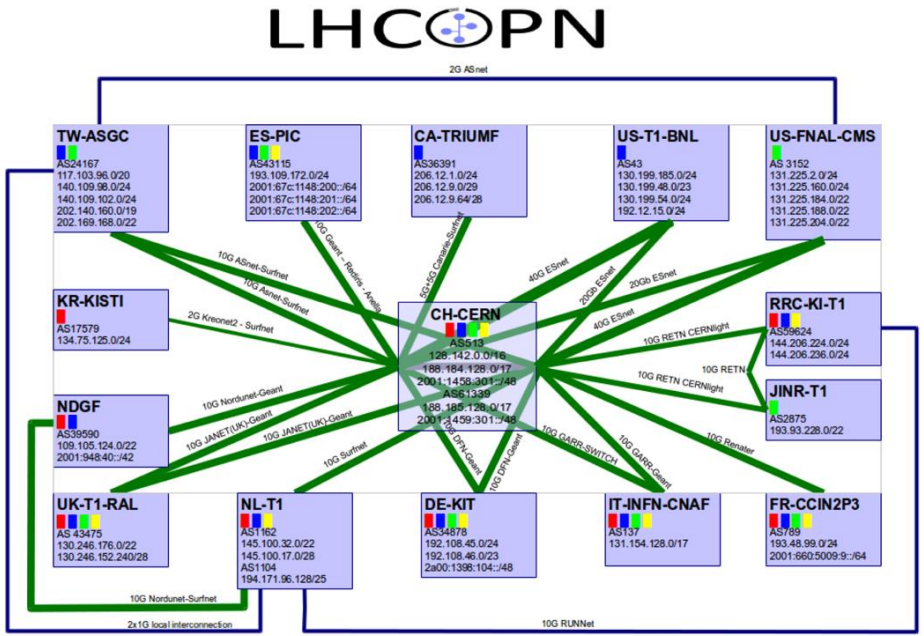


И ние инсталираме BOINC

- На новопристигналите машини в ЦЕРН
- На машини с “неефективни” услуги
- На малки GRID сайтове
- На сайтове с ограничения

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

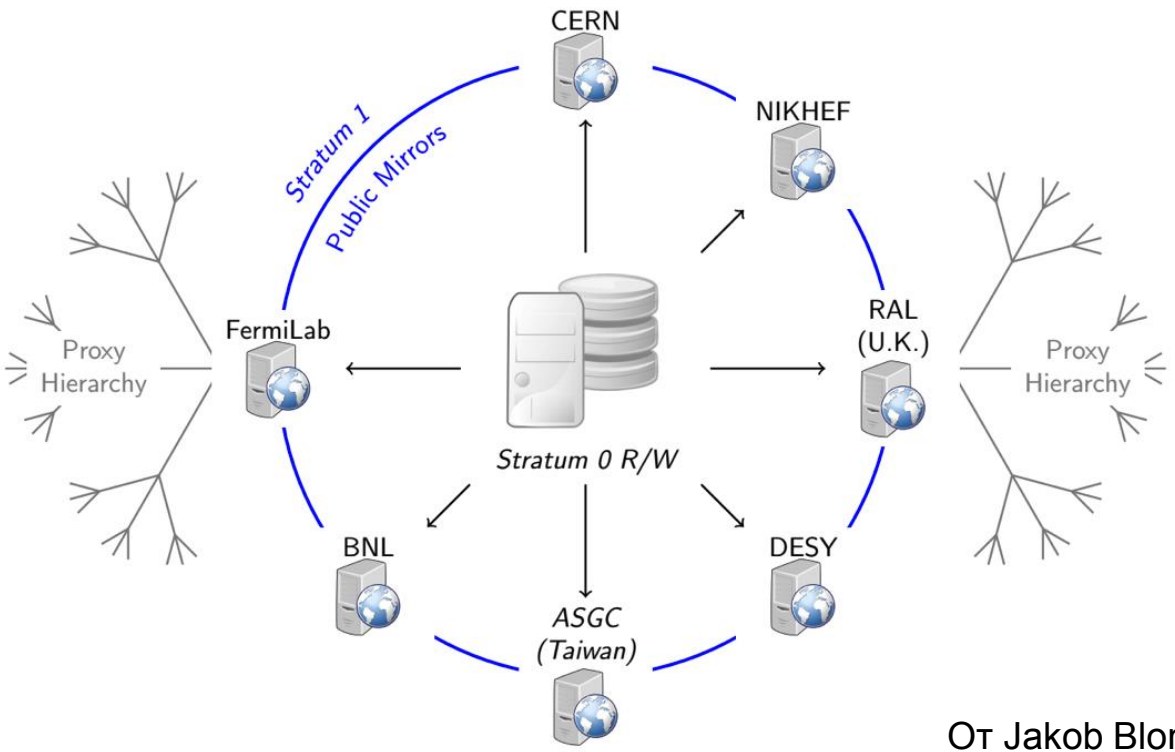


Дистрибуция на софтуера: CVMFS

- 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

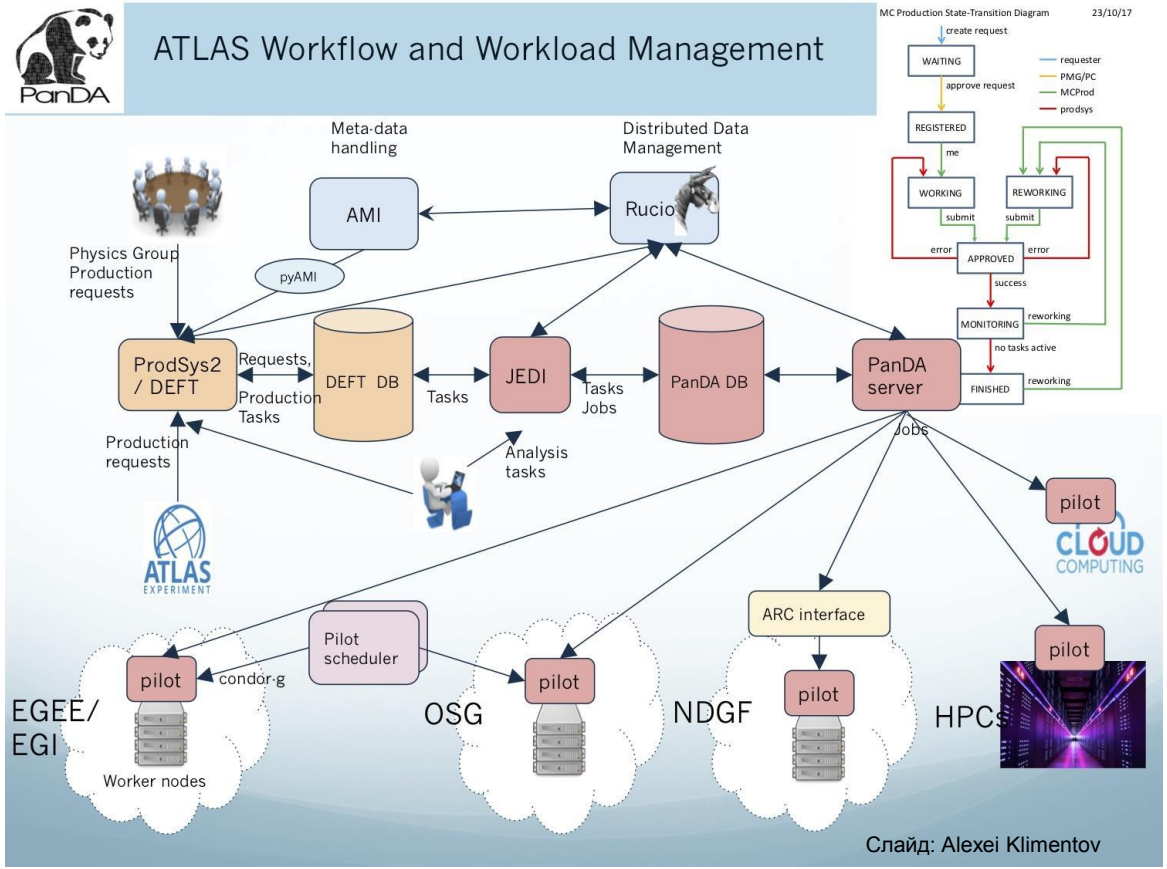


От Jakob Blomer

Изчисления

(как)

Изчислителни задачи: от А до Я



Слайд: Alexei Klimentov

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

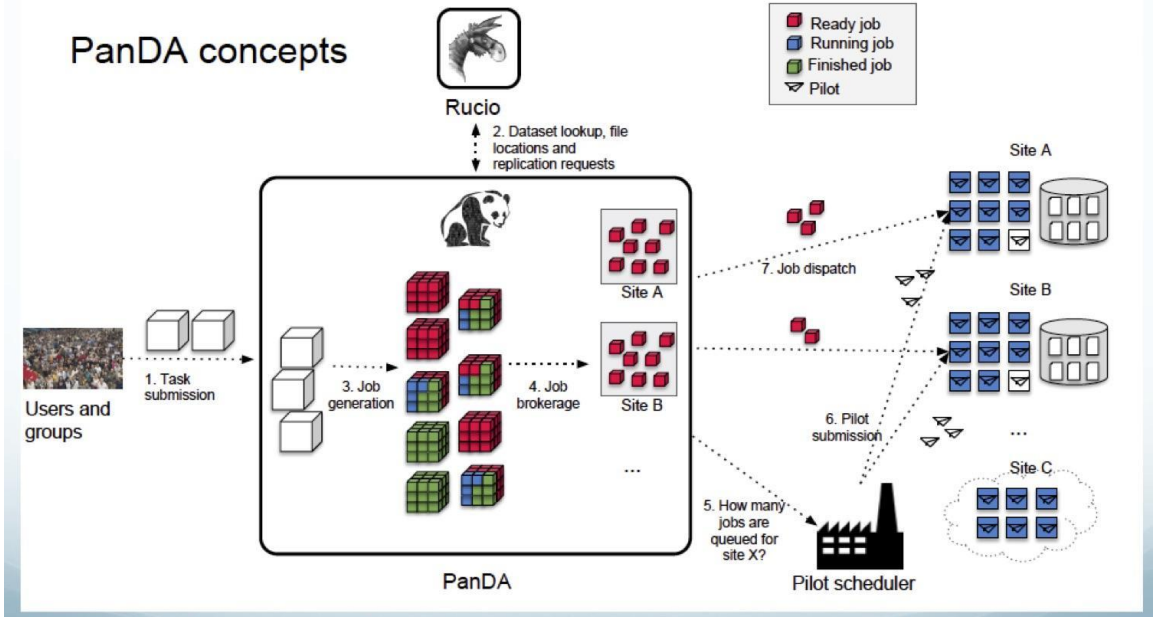
АТЛАС:

- 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 пре еднакви условия

Система за разпределяне на задачи в АТЛАС

ATLAS EXPERIMENT

PanDA Workload Management System

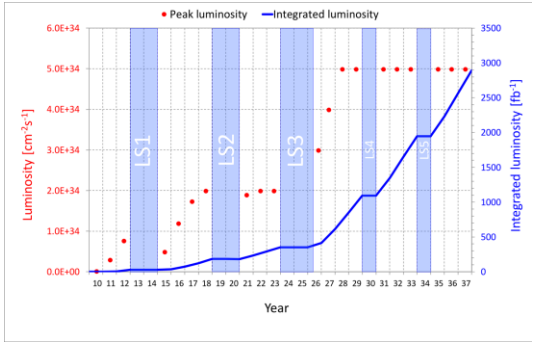


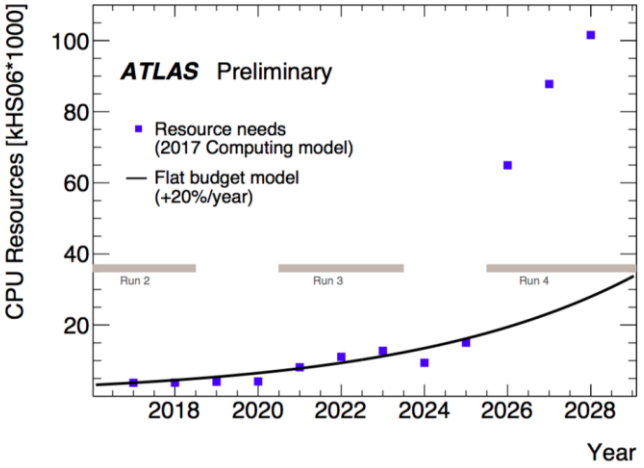
CHEP talk about new PanDA developments : "Harvester : an edge service harvesting heterogeneous resources for ATLAS", Track 3 – Distributed computing, 12 Jul 2018, 11:15
https://indico.cern.ch/event/587955/contributions/2937391/

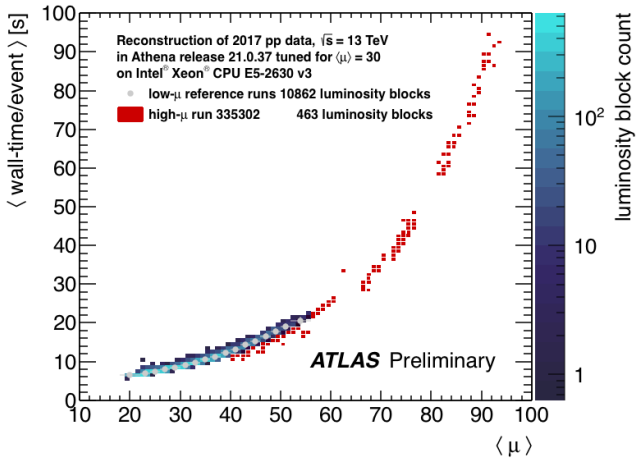
Слайд: Alexei Klimentov

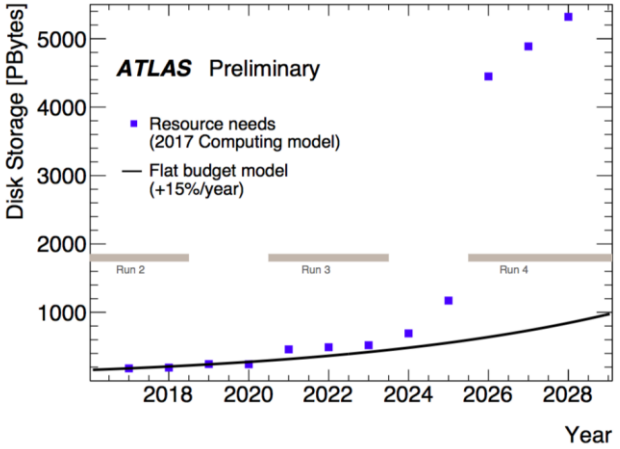
Бъдещето

HL-LHC


Peak luminosity
Integrated luminosity
Luminosity [cm⁻²s⁻¹]
Integrated luminosity [fb⁻¹]
LS1
LS2
LS3
Year


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


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 ⟩
luminosity block count

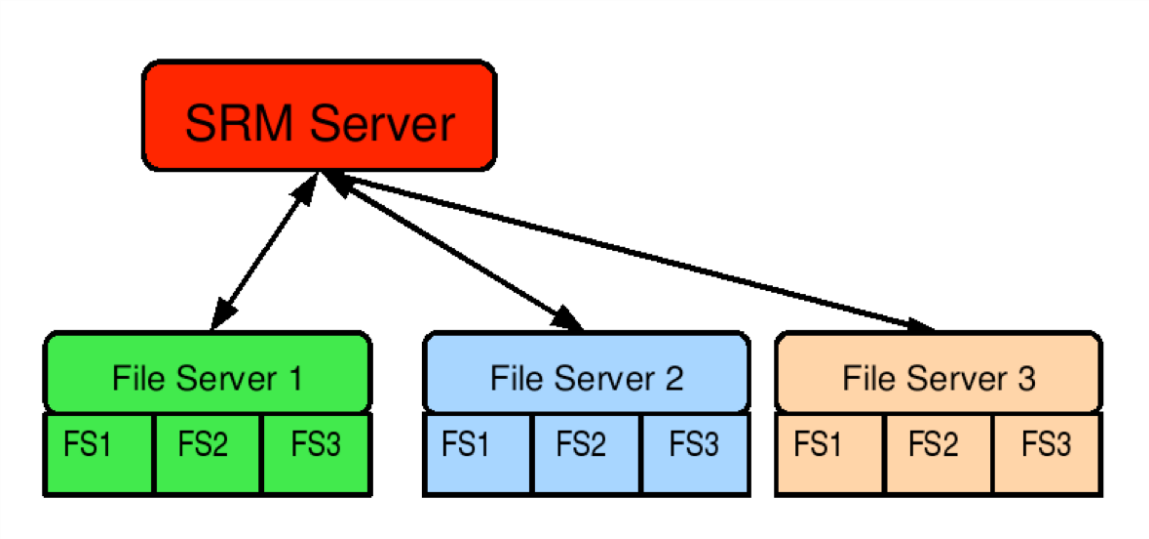

ATLAS Preliminary
Resource needs
(2017 Computing model)
Flat budget model
(+15%/year)
Run 2
Run 3
Run 4
Disk Storage [PBytes]
Year

Въпроси?

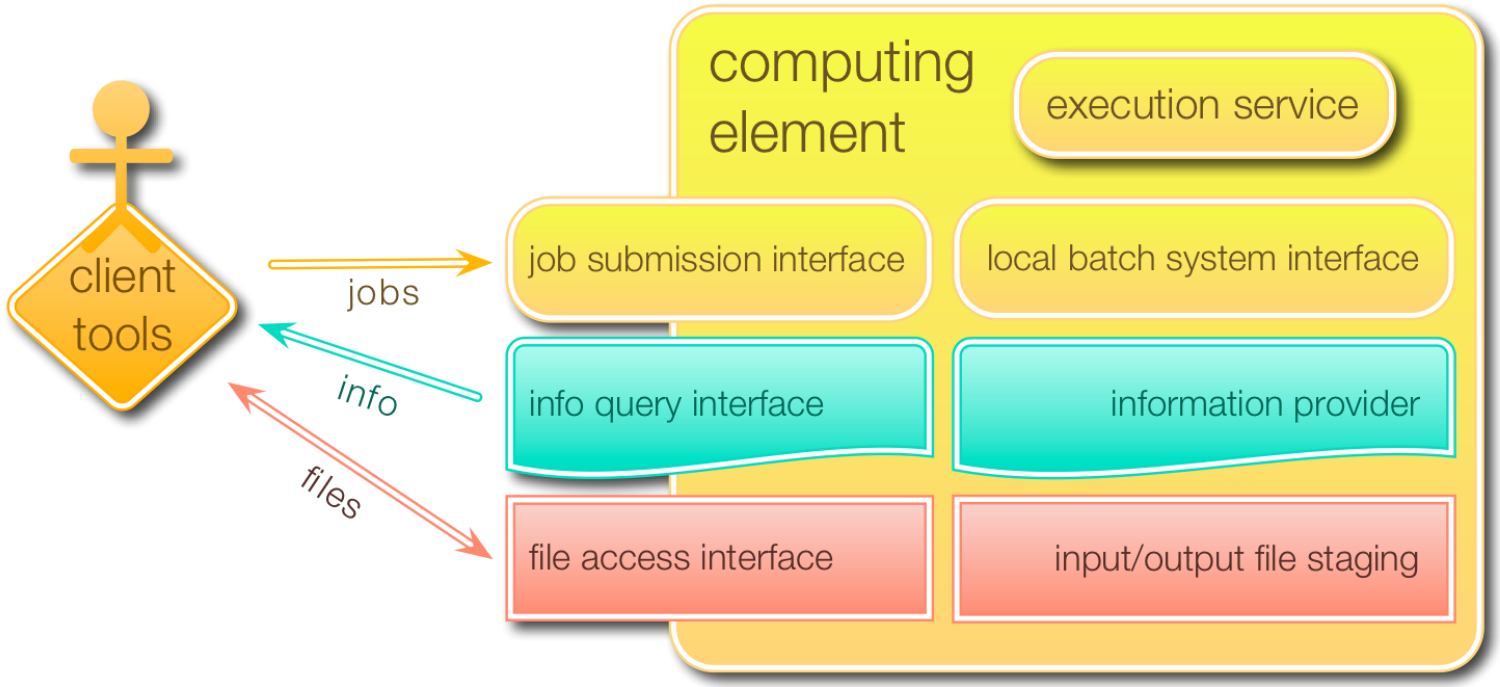
Ivan.Glushkov@cern.ch

Подробности

SE (Storage Element)



CE (Computing Element)



Global Shares