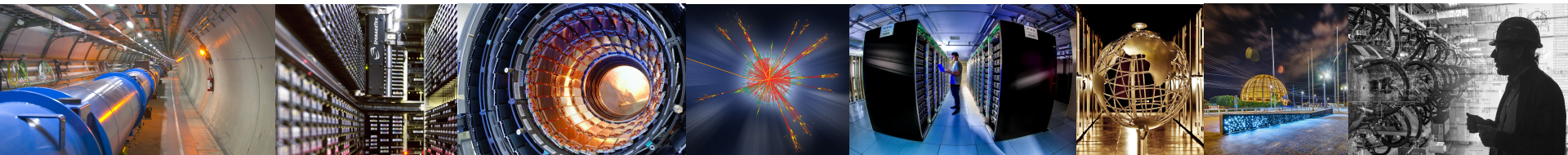


# Разпределени изчисления в ATLAS

## Част 1: GRID

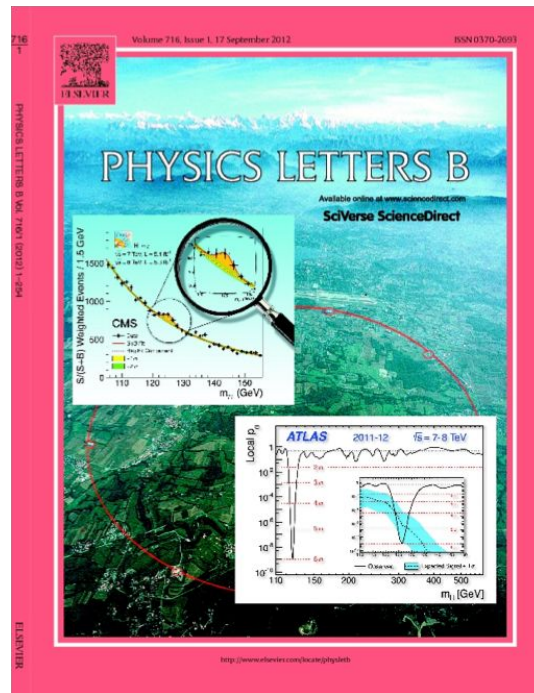
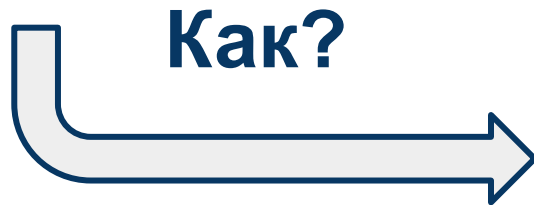
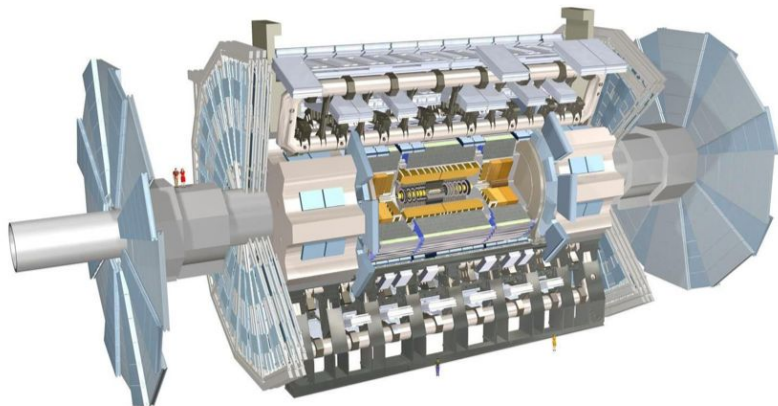


*Др. Иван Глушков  
Тексаски университет / АТЛАС  
Българска инженерна учителска програма  
ЦЕРН, октомври 2019*

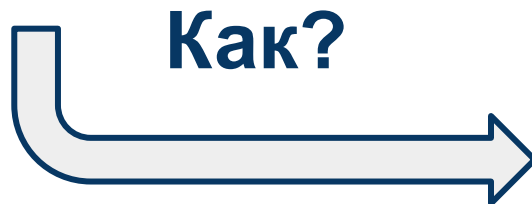
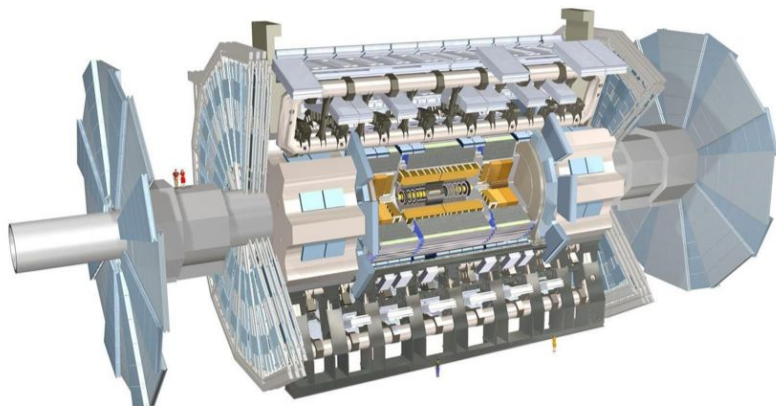


**Какъв е проблема?**

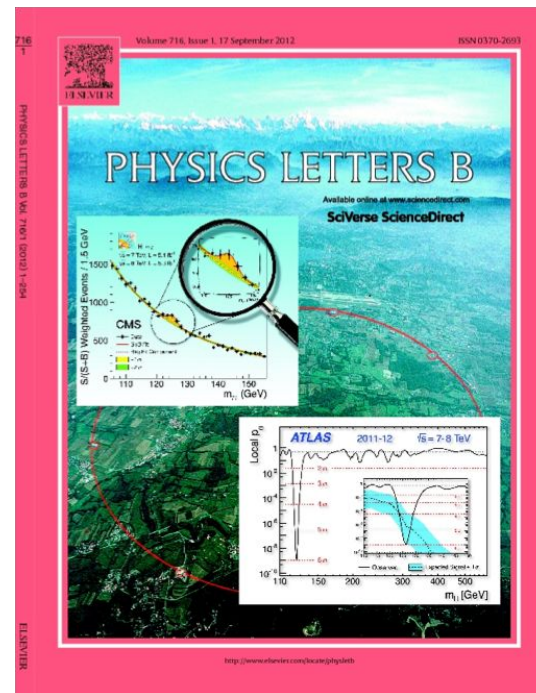
# От ATLAS до научна публикация



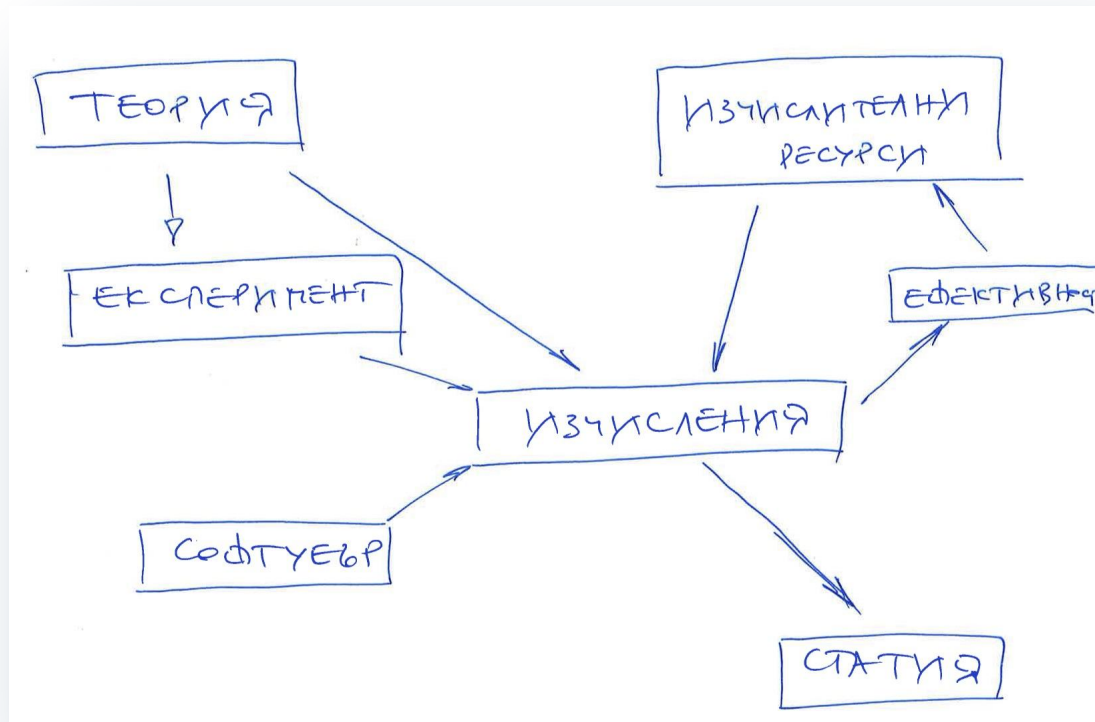
# От ATLAS до научна публикация



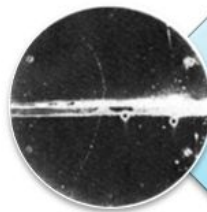
.. и бързо моля



# Парадигма на науката



# Какво ви е (било) нужно за откритие?



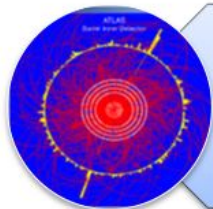
През 1930-те

- ~2 учени от една държава
- Лист и химикал



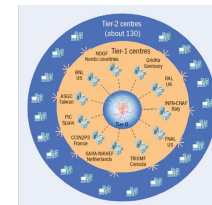
През 1970-те

- ~200 учени ~10 държави
- Мейнфрейм



Днес

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



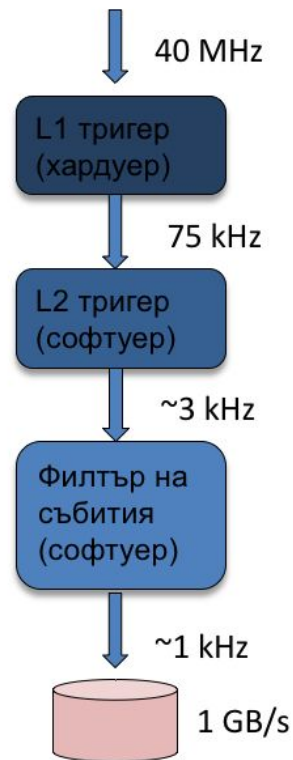
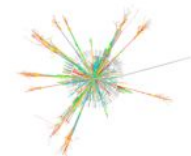


# Данните



# 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'
```

Query Result x

SQL | All Rows Fetched: 1 in 51,936 seconds

SUM(BYTES)/1000/1000/1000/1000/1000
1 58.44298773921483

## 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?!

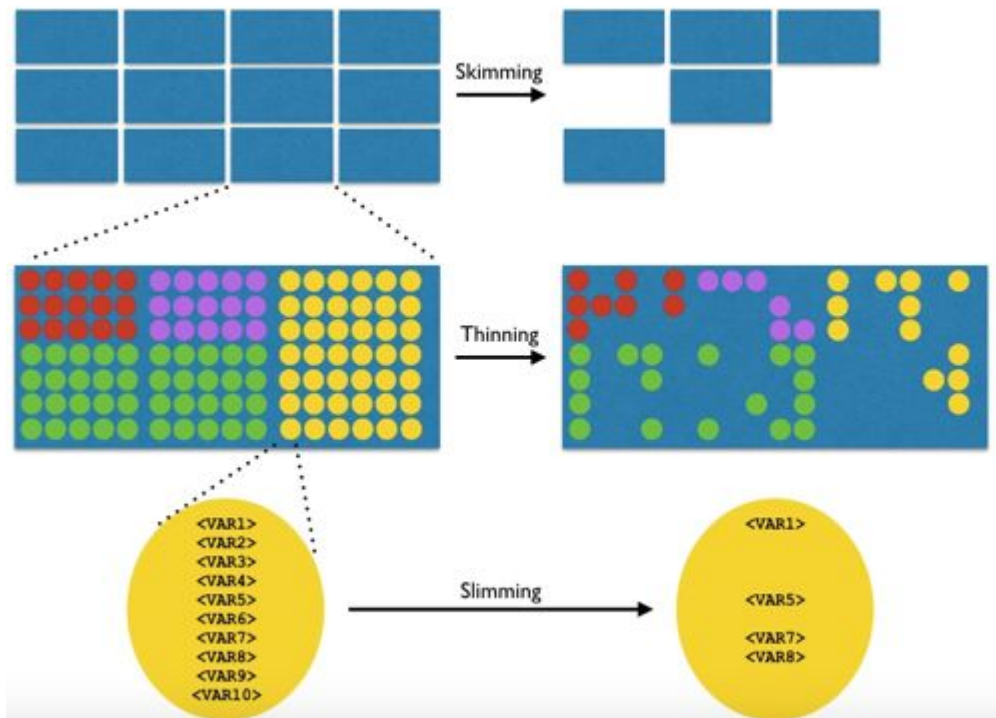


- Групиране на анализите по теми - Хигс, Суперсиметрии, Екзотики
- Селектиране само на данните които са релевантни за дадения анализ

**GB / MB**

# Селектиране на данни

## (ATLAS Derivation Framework)



## Премахване на ненужната информация

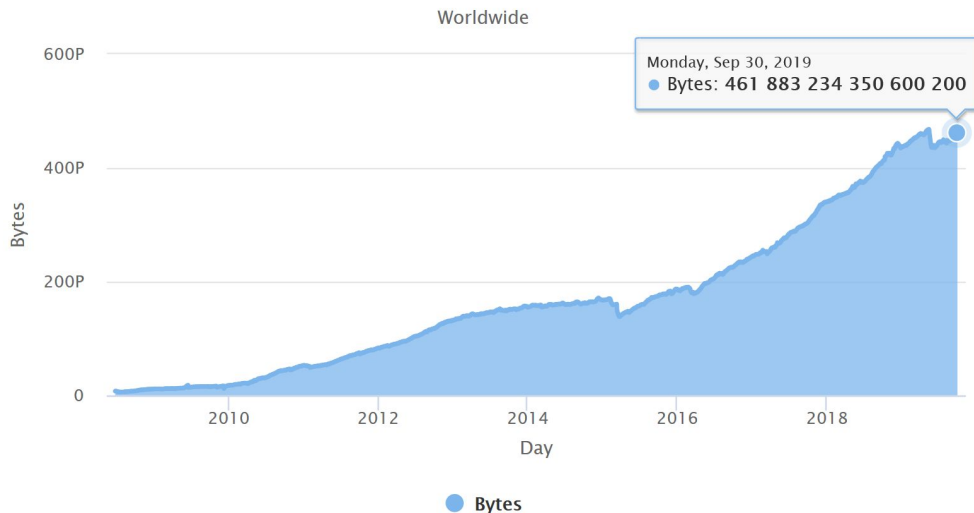
- Цели събития
- Части от събитията
- Характеристики на събитията
- За всеки анализ - отделни данни

Графика: James Catmore

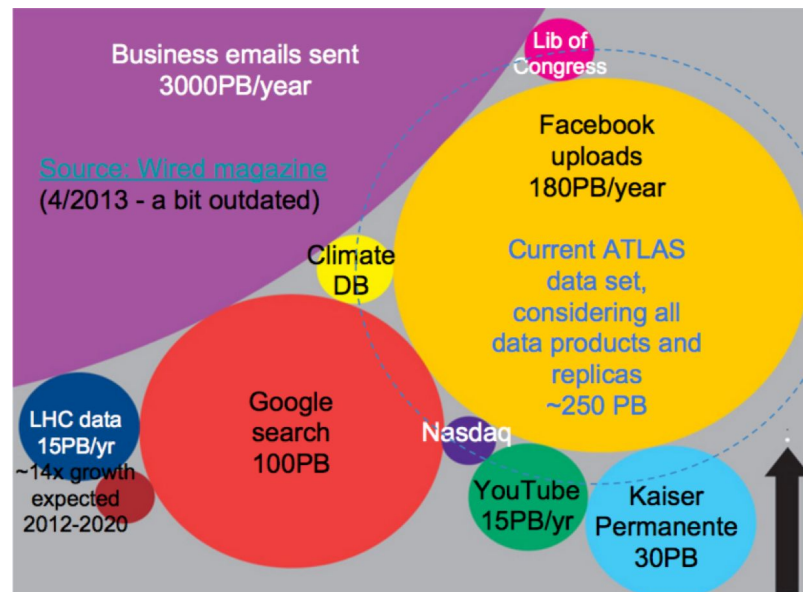
# Истина и теория



# Всичко общо..



## 462 PB!





# Изчисления

(колко и какви)



**Колко изчислителни  
ресурси ни трябвават?**



# “Колкото повече, толкова повече!”

Мечо Пух, 1966



**“Колкото повече, толкова повече!”**

Мечо Пух, 1966

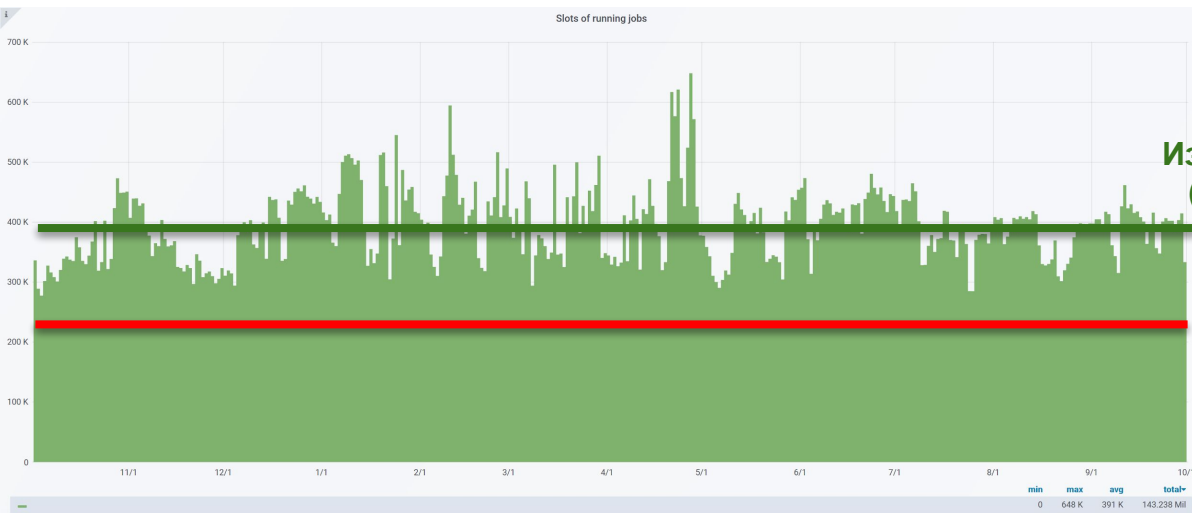


**“Има само едно нещо, което е по-хубаво от гърненце мед.. И това са две гърненца мед.”**

Мечо Пух, 1966

# Колко използваме?

648 000 !!!..



Обещано  
(279 000)

Използвано  
(391 000)

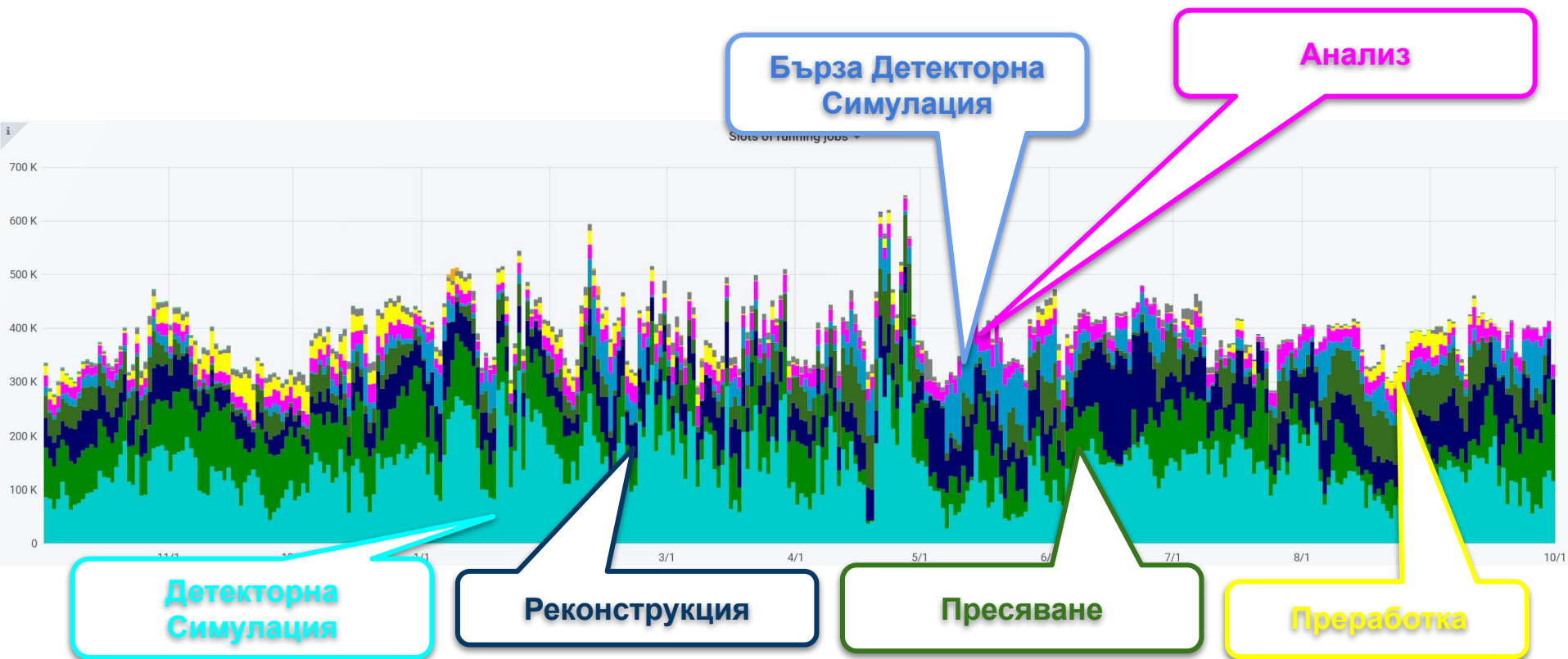
## Речник

- 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?





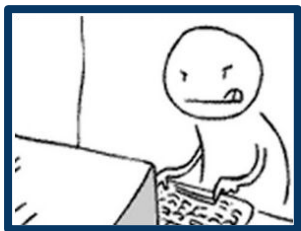
# Изчисления

(къде)

# Разпределени 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: Принцип на работа



Къде / кои са данните които  
искам да обработвам?

Система за  
разпределяне на данни  
(Rucio)

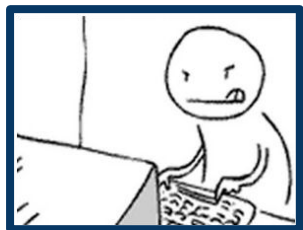
Данни X, които се намират  
на сайт Y

## Речник

- Сайт - (отдалечен)  
компютърен център



# Потребител и разпределени изчисления: Принцип на работа



Къде / кои са данните които  
искам да обработвам?

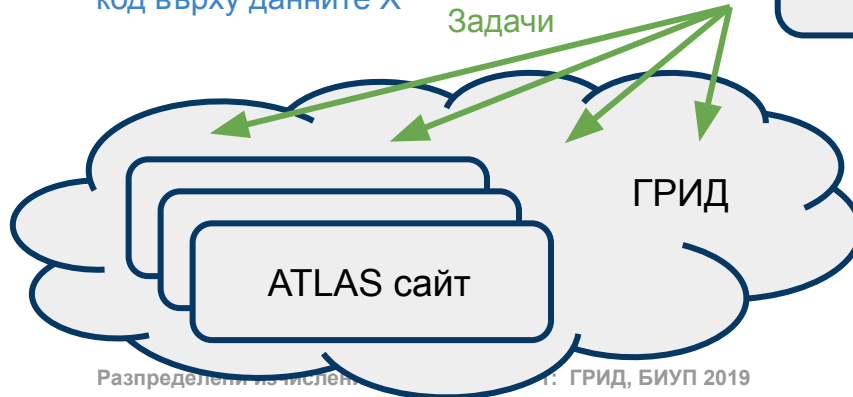
Система за  
разпределяне на данни  
(Rucio)

Данни X, които се намират  
на сайт Y

Искам да изпълниш моя  
код върху данните X

Система за  
разпределяне на задачи  
(PanDA)

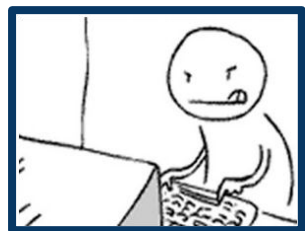
Задачи



## Речник

- Сайт - (отдалечен) компютърен център

# Потребител и разпределени изчисления: Принцип на работа



Къде / кои са данните които  
искам да обработвам?

Система за  
разпределяне на данни  
(Rucio)

Данни X, които се намират  
на сайт Y

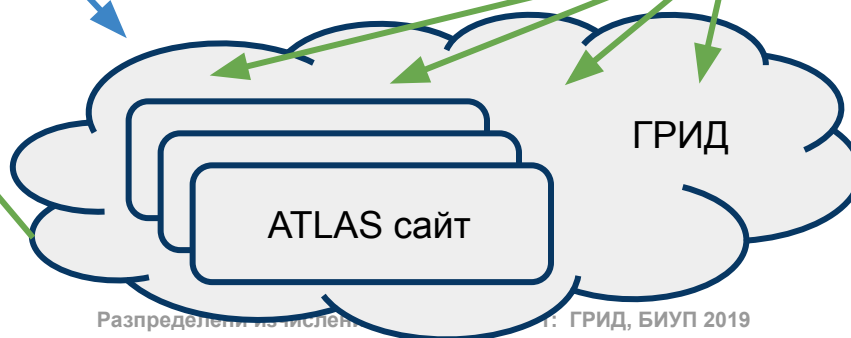
Система за  
разпределяне на задачи  
(PanDA)

Искам да изпълниш моя  
код върху данните X

Задачи

Резултати?

Резултати / Още  
не е готово



## 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

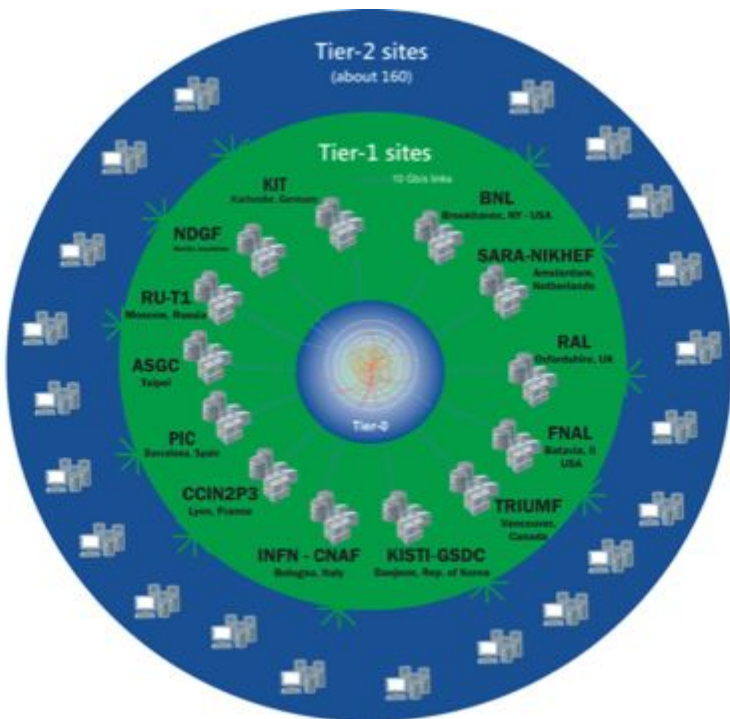
# Какво е GRID III

- 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-ove, GRID-ove..



# WLCG (The Worldwide Computing GRID)



Участващи GRID-ове:

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



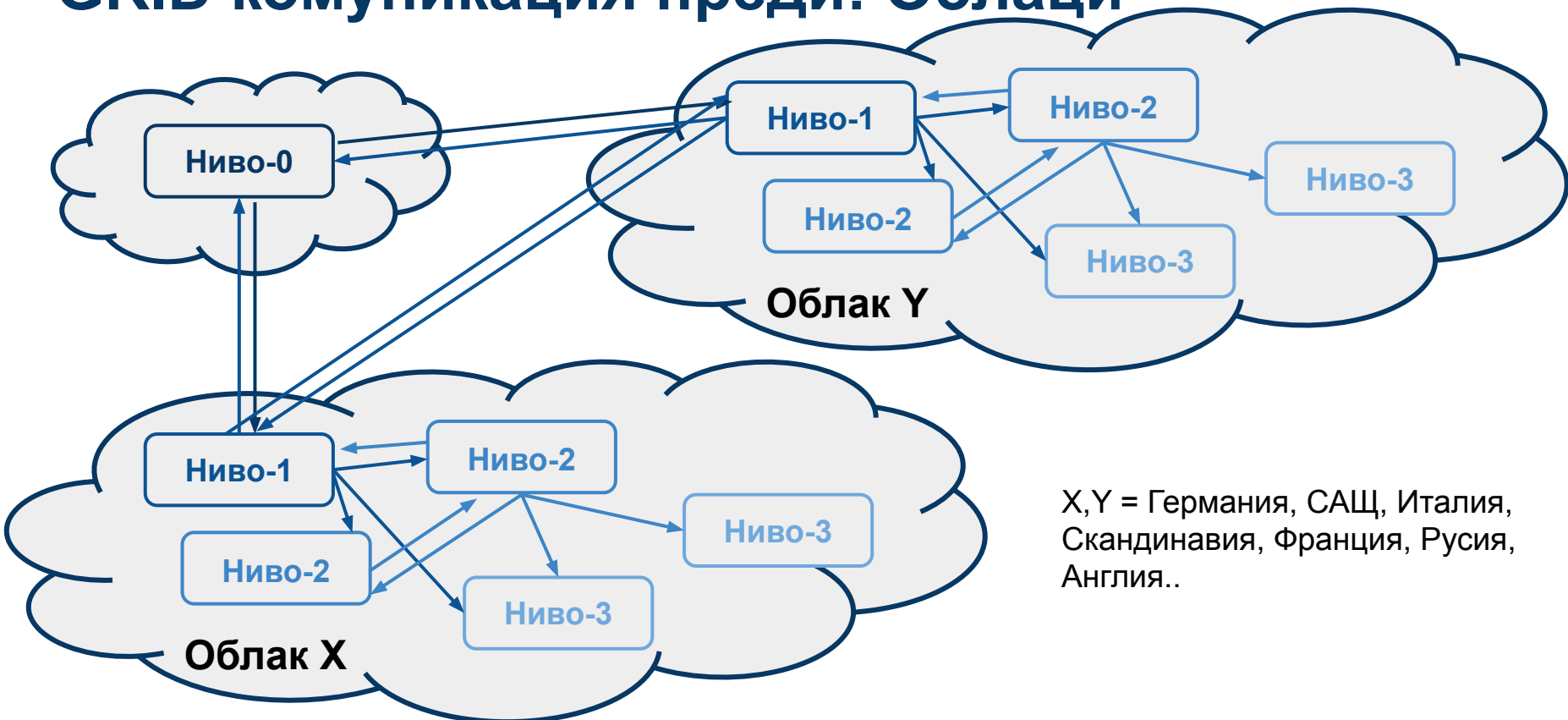
# ATLAS Сайт

- 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

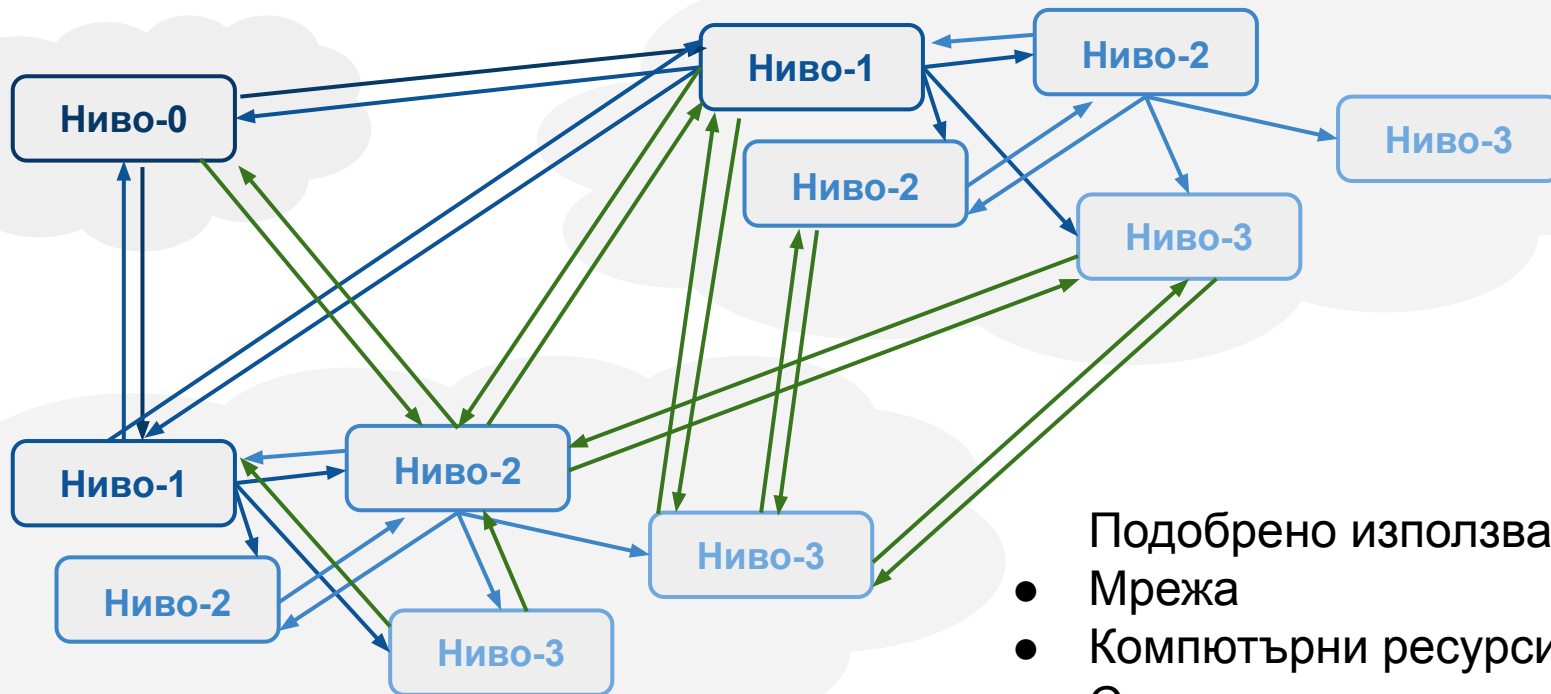
# ATLAS Облак

- 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



# GRID комуникация сега: Тотална мрежа (Full mesh)



Подобрено използване на:

- Мрежа
- Компютърни ресурси
- Съхранение на данни

# Роля на нивата: Ниво-0

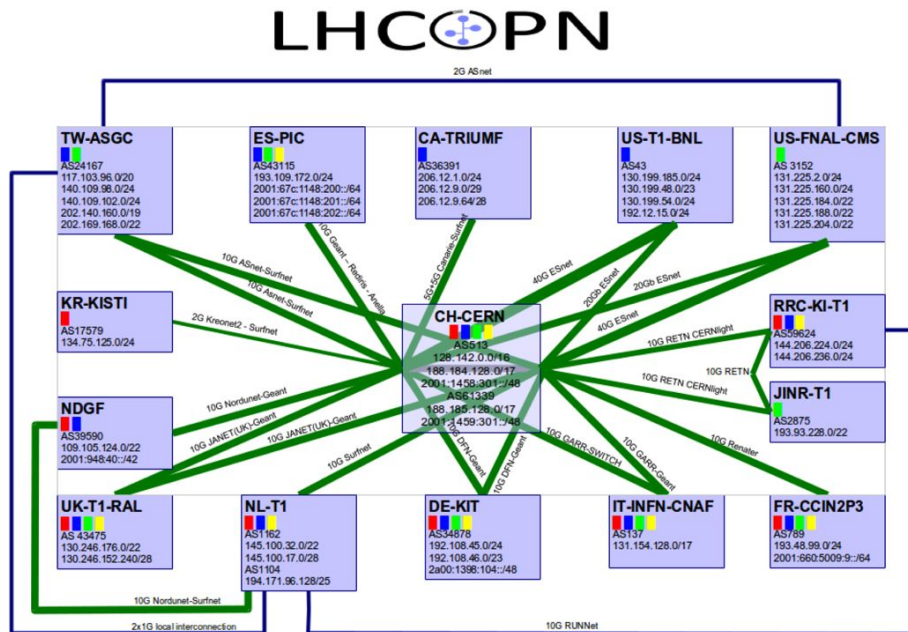
- Ниво-0
  - Копие на всички първични данни на лентови носители
  - Първо ниво на обработка на данните от LHC
  
- Централни услуги в ЦЕРН (стабилност)
  - Разпределяне на задачи - PanDA
  - Разпределяне на данни - Rucio
  - База данни - Oracle
  - Прехвърляне на файлове - FTS (WLCG)
  - Кешова инфраструктура - 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

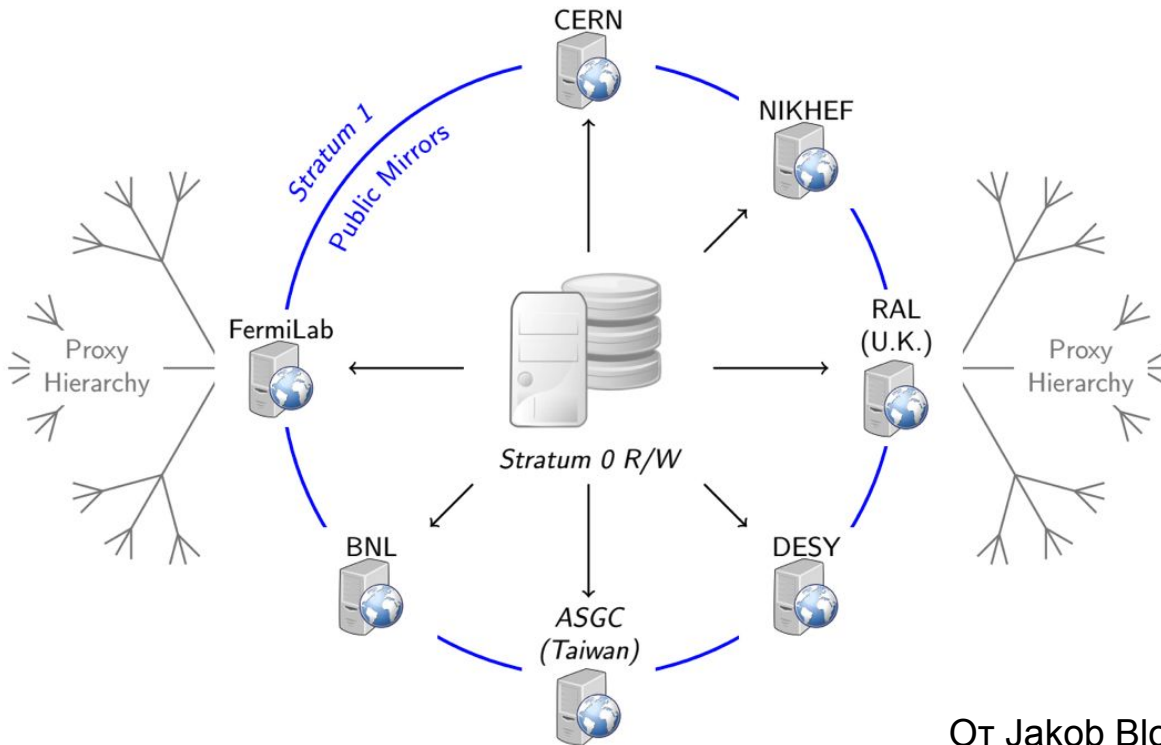
- Ниво-0 и Ниво-1: LHCOPN
- Ниво-0, 1 и 2: LHCONE (LHC Open Network Environment)
  - Колaborация между научни организации и университети
  - Динамично използване на ресурсите, споделяне на разходите
  - Layer3 рутуран VPN



# 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 Stratum-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





От Jakob Blomer



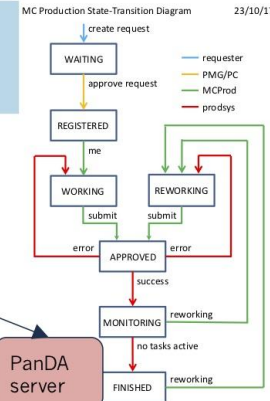
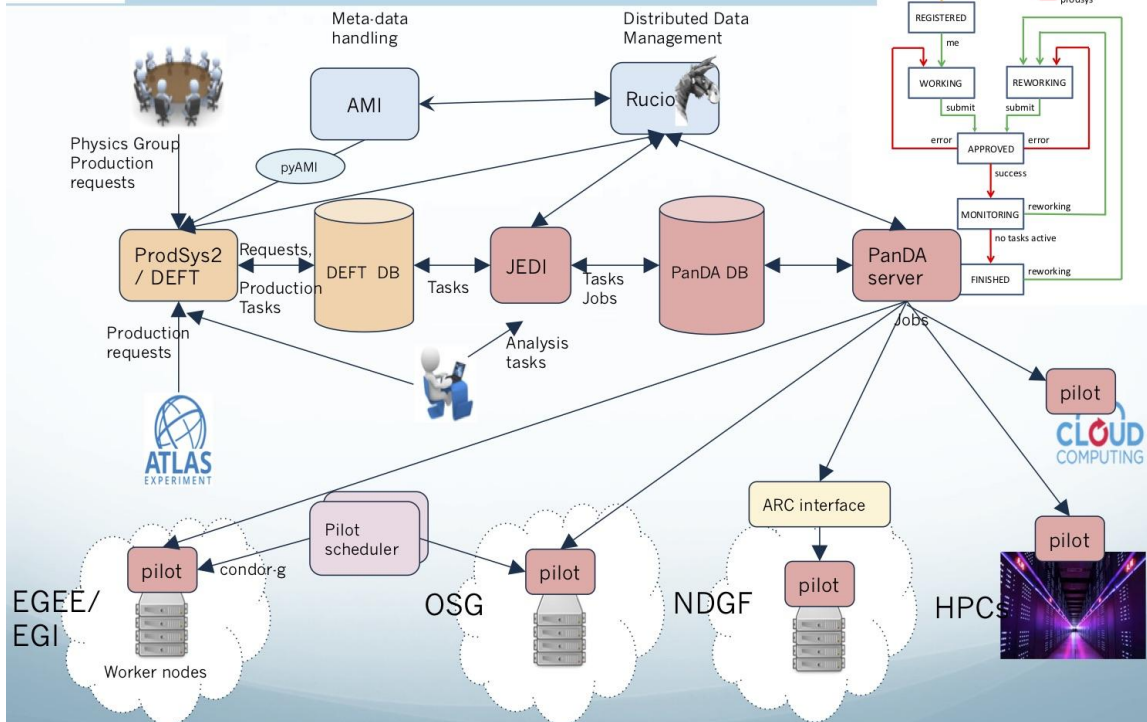
# Изчисления

(как)

# Изчислителни задачи: от А до Я



## ATLAS Workflow and Workload Management



## 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

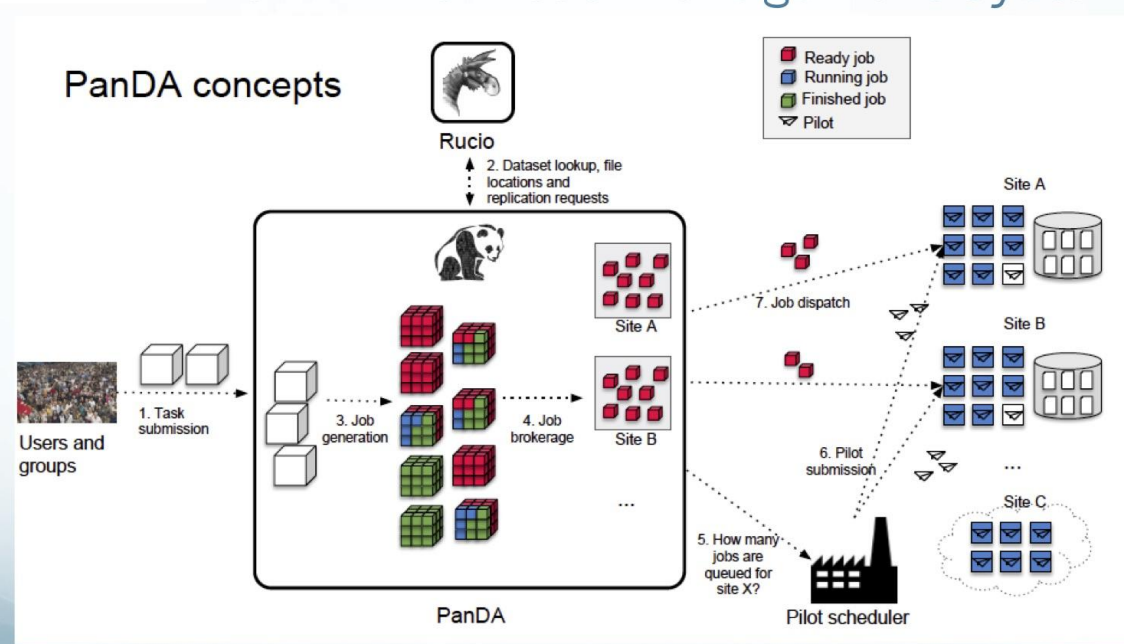
## АТЛАС:

- 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



## PanDA Workload Management System

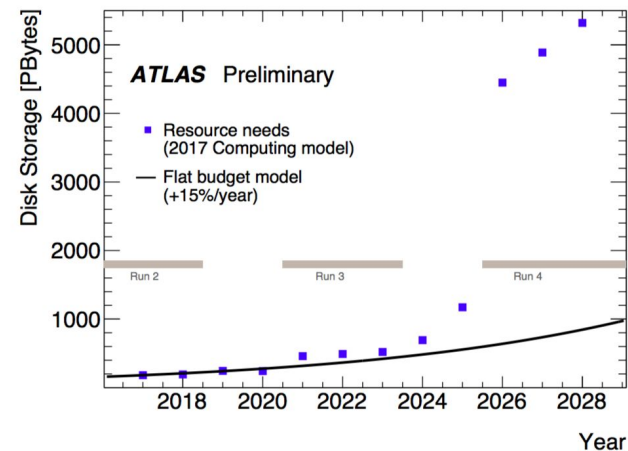
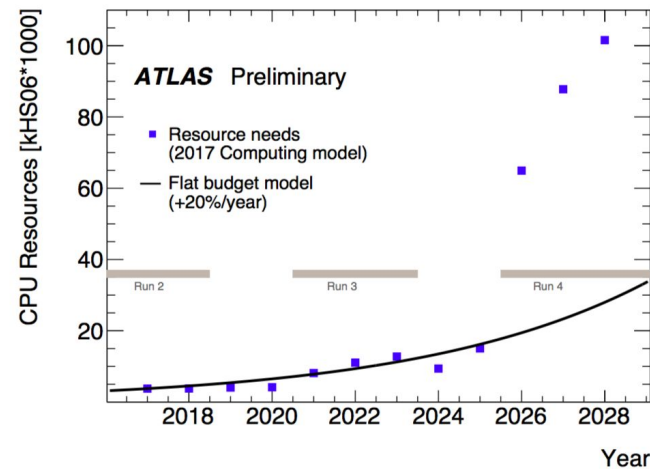
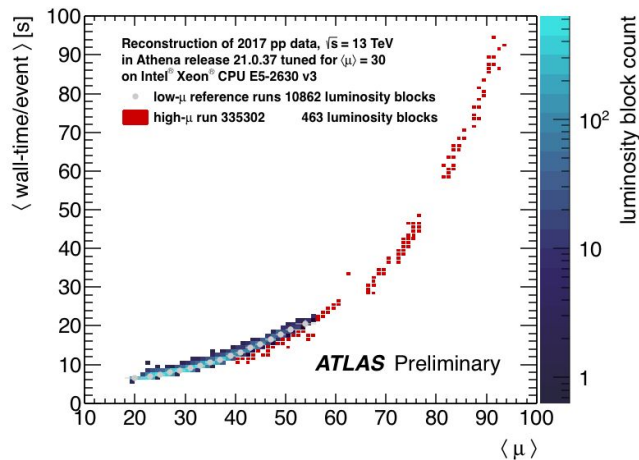
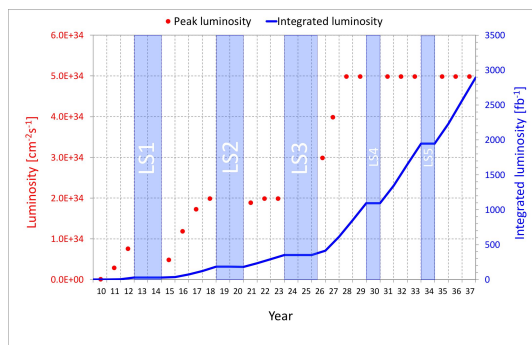


CHEP talk about new PanDA developments: "Harvester: an edge service harvesting heterogeneous resources for ATLAS",  
 Track 3 – Distributed computing, 12 Jul 2018, 11:15  
<https://indico.cern.ch/event/587955/contributions/2937391/>  
 Alexei Klimentov



# Бъдещето

# HL-LHC



# Заклучение

- Ресурсите и данните 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, изкуствен интелект
  - Оптимизиране на използването на хардуера - кеш сайтове, директно четене от ленти,
  - Нов тип ресурси - суперкомпютри, облаци, нови архитектури (следващата лекция)



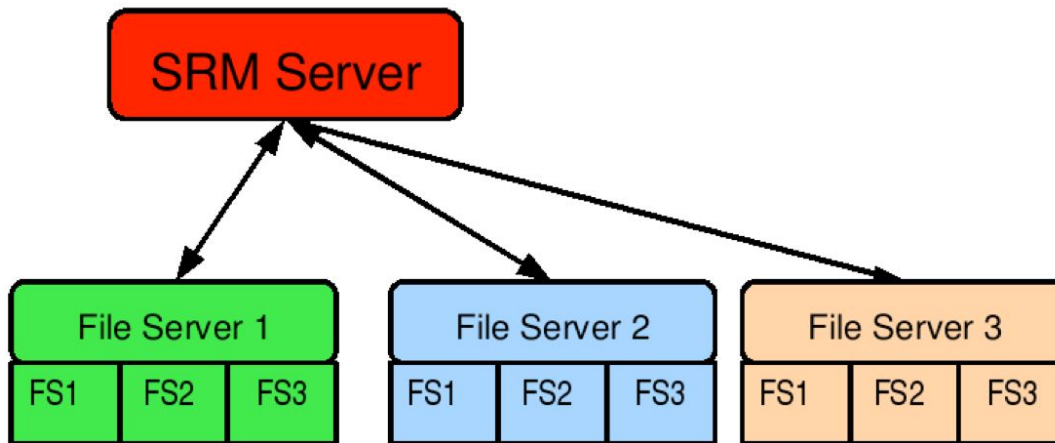
**Въпроси?**



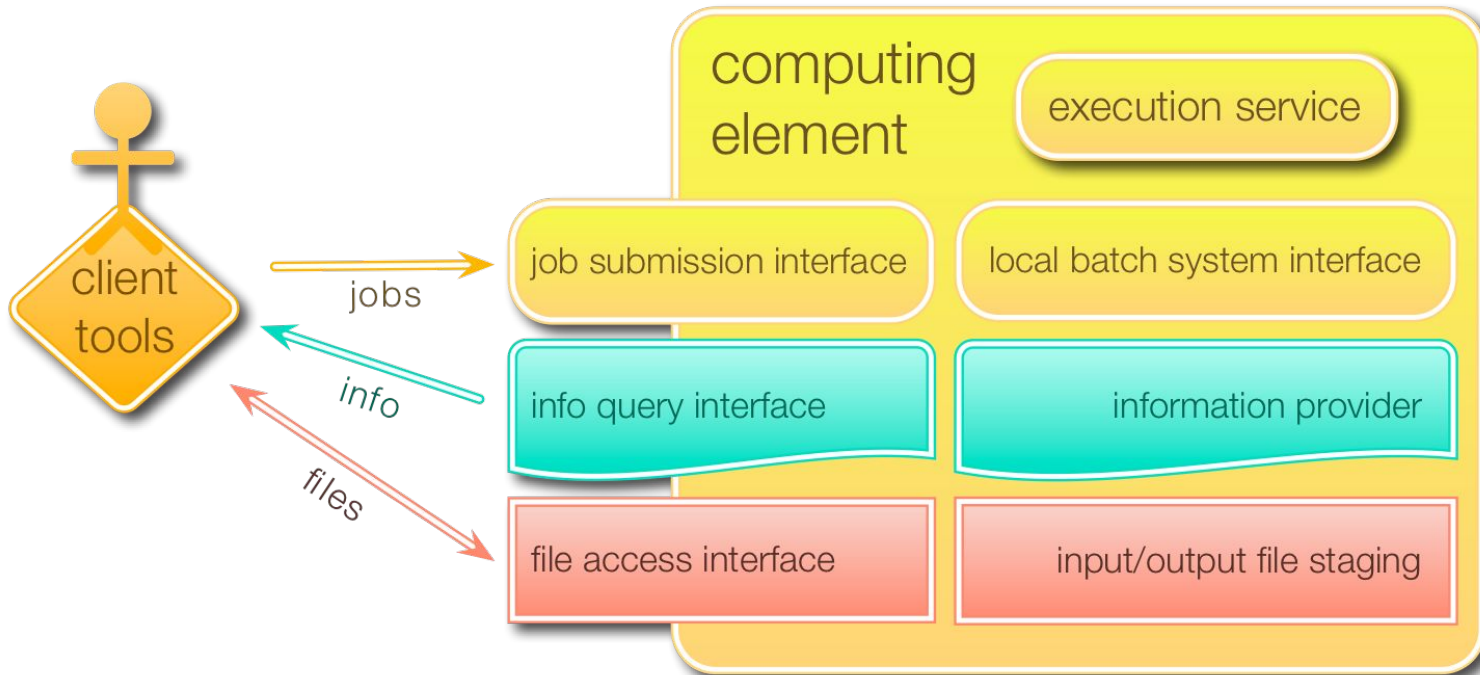


# Подробности

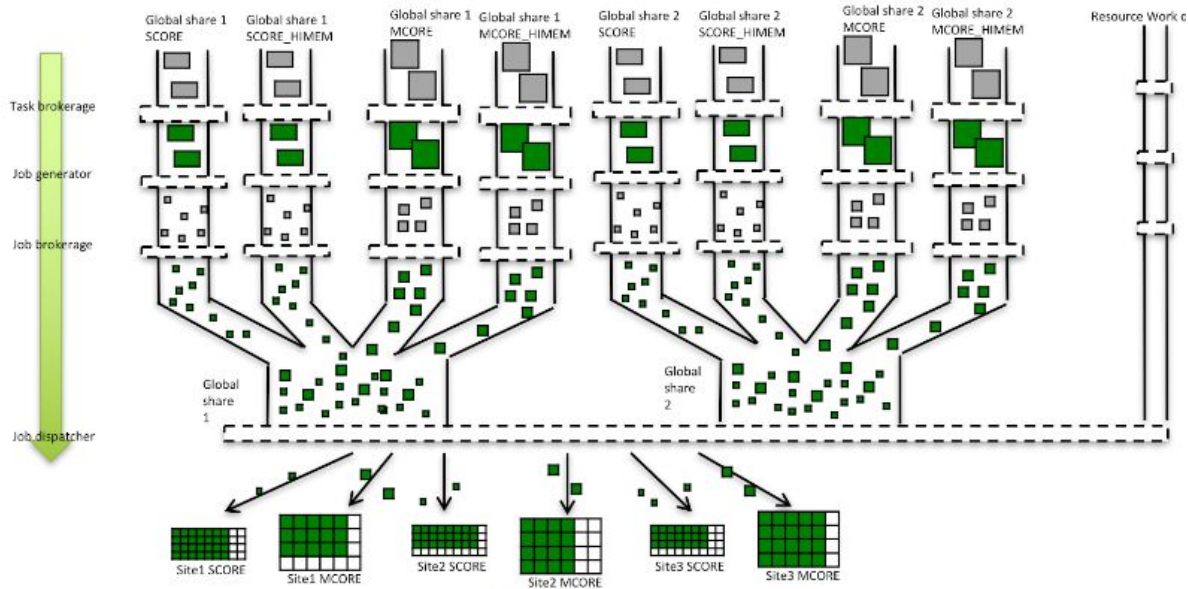
# SE (Storage Element)

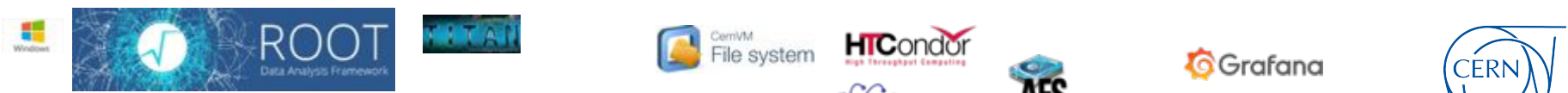


# CE (Computing Element)



# Global Shares





# Технологии в ЦЕРН

