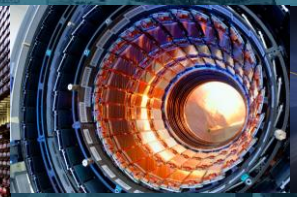
Разпределени изчисления в ATLAS

Част 1: GRID








Д-р Иван Глушков
BNL / АТЛАС
Българска инженерна учителска програма
ЦЕРН, септември 2024

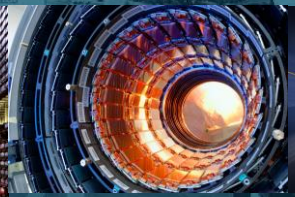
@BrookhavenLab

Разпределени изчисления в ATLAS

Част 1: !GRID

(суперкомпютри, облаци, доброволчески изчисления, графични карти)



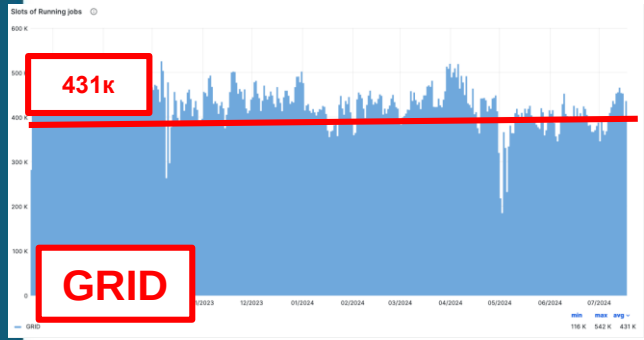


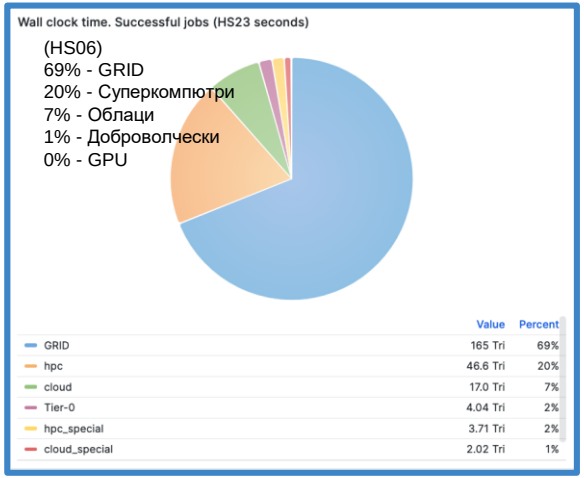


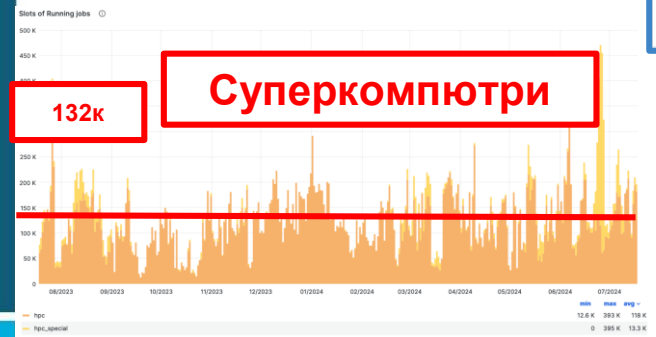

Д-р Иван Глушков
BNL / АТЛАС

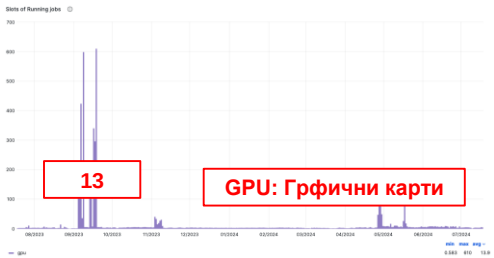
Българска инженерна учителска програма
ЦЕРН, септември 2024

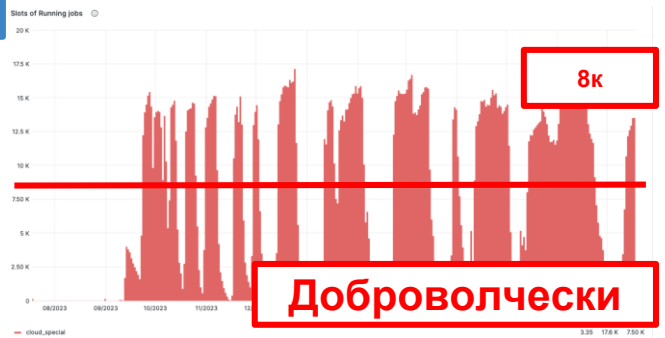
Видове ресурси











Brookhaven National Laboratory

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 изчислителна 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?
  - Гарантирани ресурси в сигурна и изолирана среда без нужда от поддръжка на хардуер. API достъп

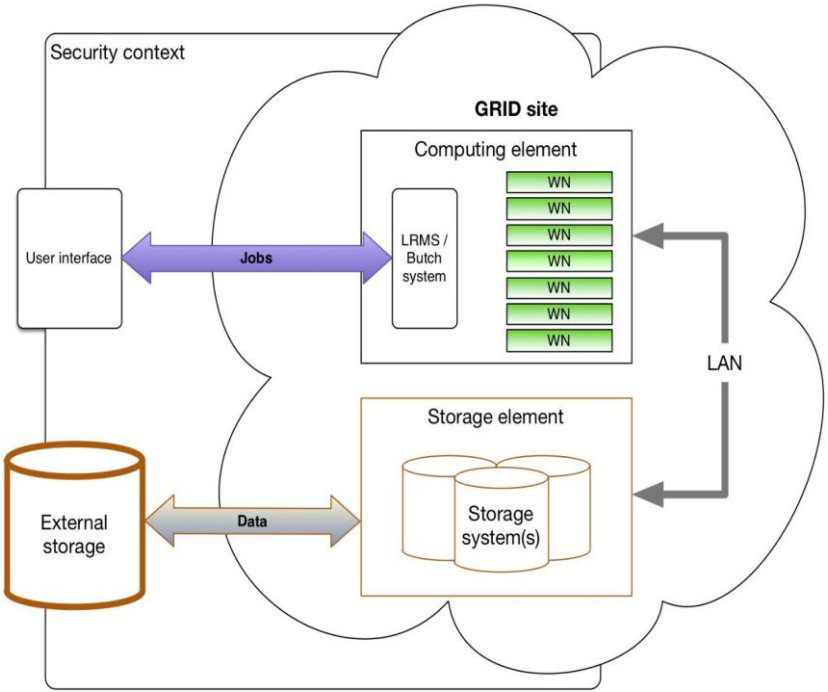
Доброволчески изчисления

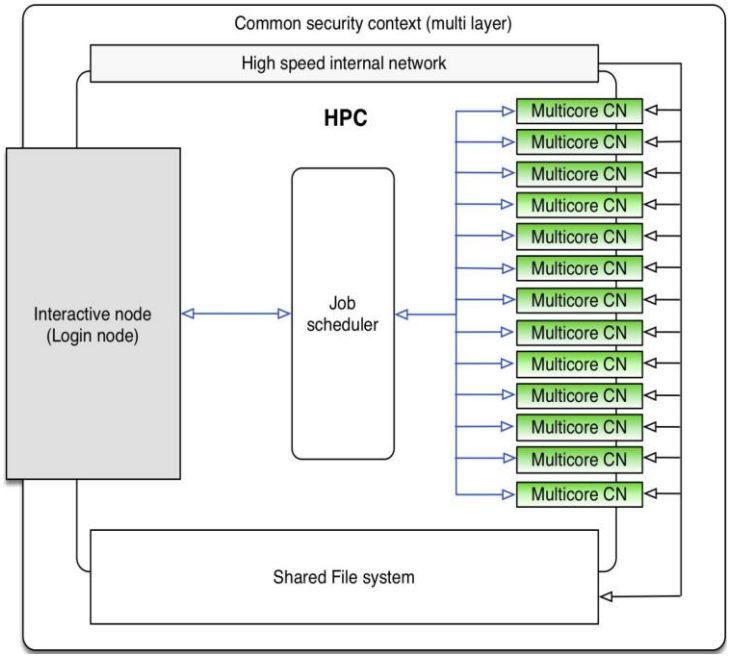
- Какво е?
  - Централизирани изчисления на ресурси предоставени от доброволци
- За какво служи?
  - Оползотворяване на ресурси които иначе биха били прахосани

Brookhaven National Laboratory

Суперкомпютри

GRID и Суперкомпютри





Brookhaven National Laboratory

GRID и Суперкомпютри

- GRID се състои от компютърни 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

Мащаб на задачите на един суперкомпютър

(на днешните машини)

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

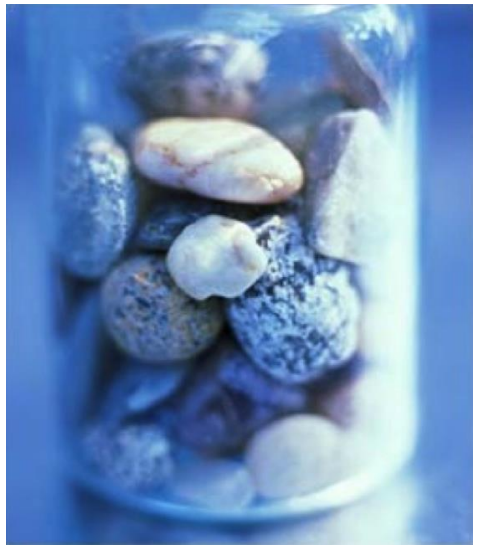
Два начина на ползване на суперкомпютър: квота и backfill

Brookhaven National Laboratory

HPC: Backfill

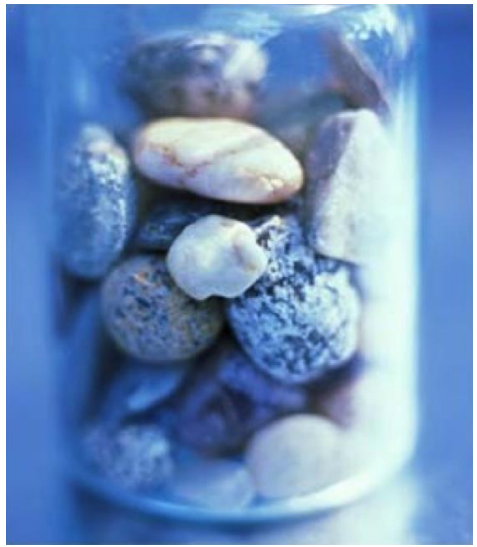
HPC: Backfill

Източник: David Cameron



Суперкомпютри: Backfill

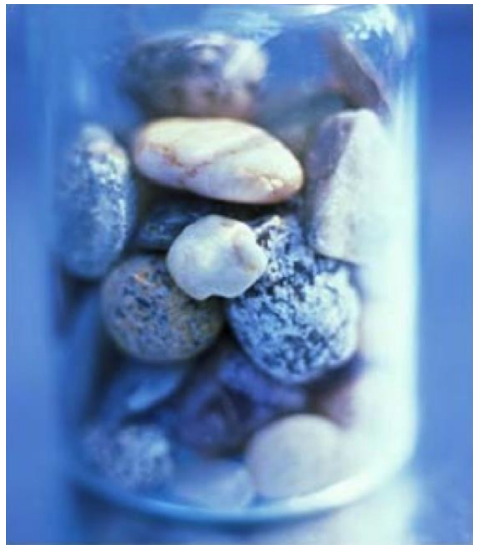
Източник: David Cameron



Суперкомпютри: Backfill

(пълним дупки всякакви)

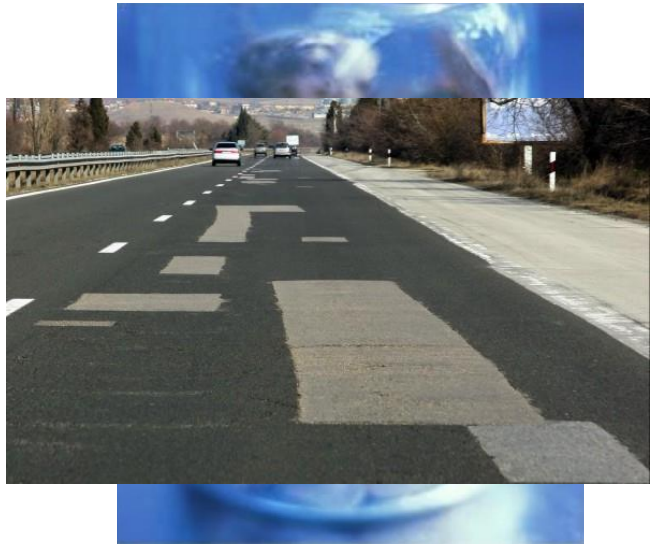
Източник: David Cameron



Суперкомпютри: Backfill

(пълним дупки всякакви)

Източник: David Cameron



Суперкомпютри: Backfill

(пълним дупки всякакви)

Размер на дупките

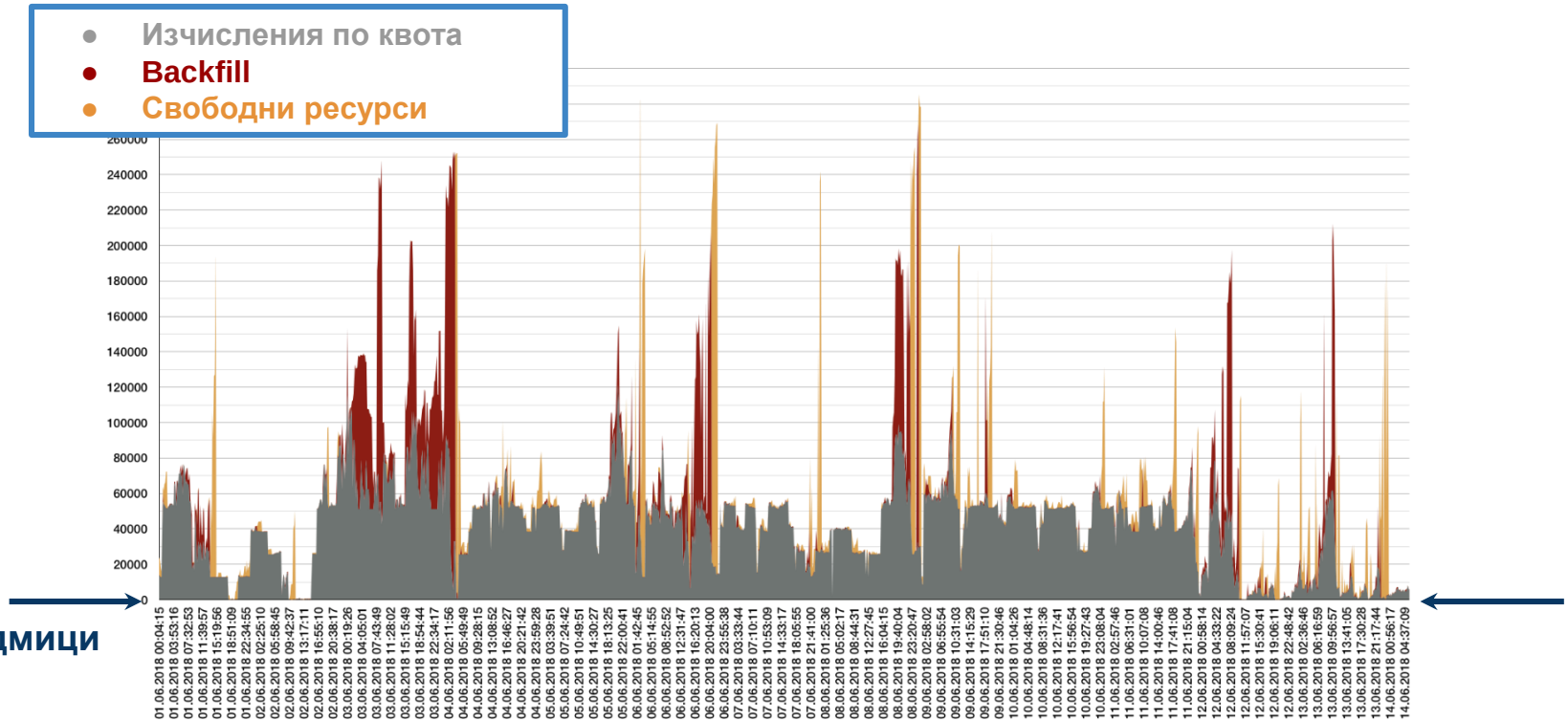
- 400M CPU*часа на година (Titan)
- Обичайна ефективност на използване на суперкомпютър във времето - 90%.
- 10% == 400M CPU*часа на година (Titan)
- Всяка неефективност е нежелана!

Запълване

- С какво?
  - Много, с различен размер, но малки задачи (т.е. - ние)
- На каква цена?
  - Безплатно

Brookhaven National Laboratory

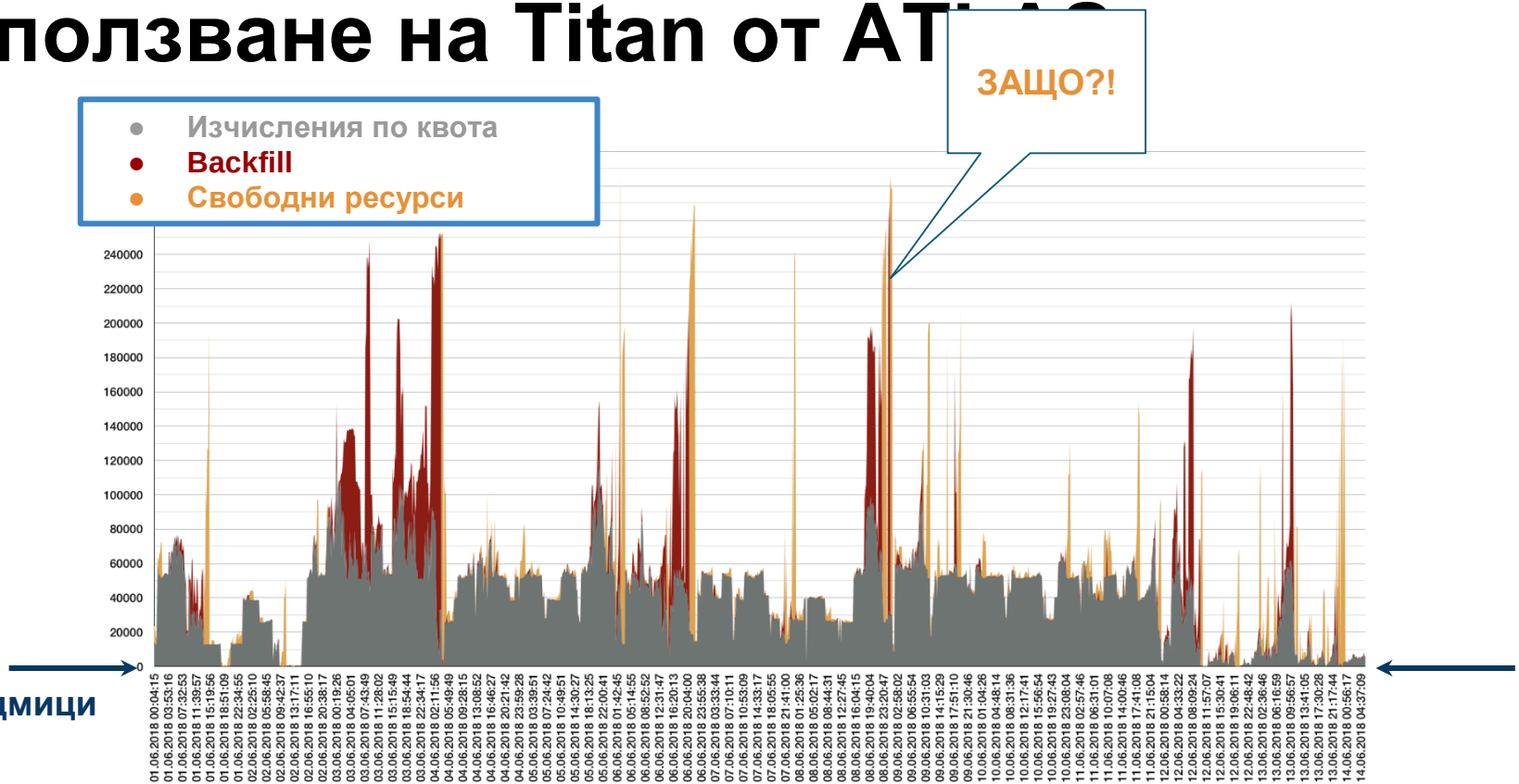
Използване на Titan от ATLAS



две седмици

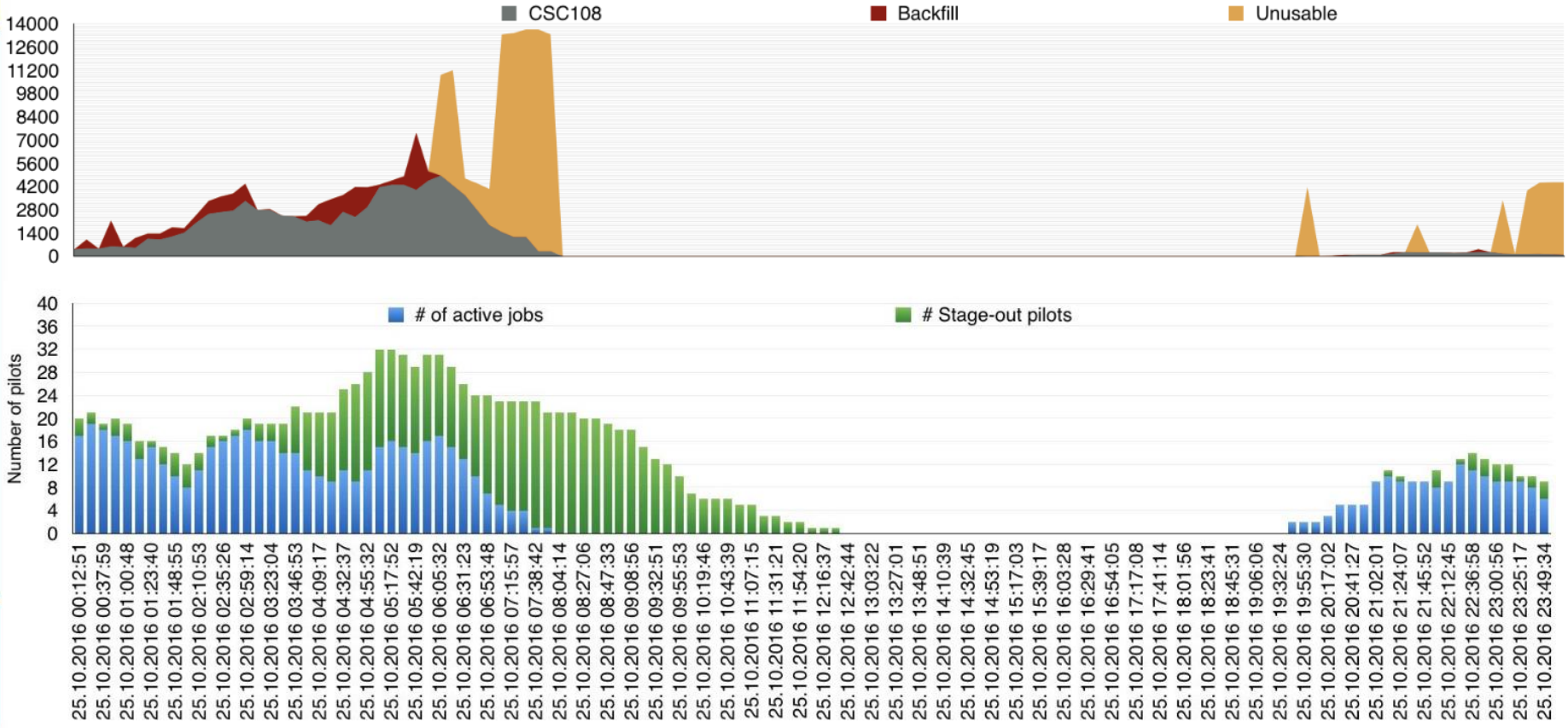
Brookhaven National Laboratory

Използване на Titan от ATLAS



две седмици

Проблемът


CSC108
Backfill
Unusable
14000
12600
11200
9800
8400
7000
5600
4200
2800
1400
0
Nodes
# 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

Американски суперкомпютри използвани от АТЛАС

Cori (САЩ)
Cray/HPE
Intel Xeon Phi 7250 68C 1.4GHz
622336 ядра
Производителност: 14 PFLOP/сек

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

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

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

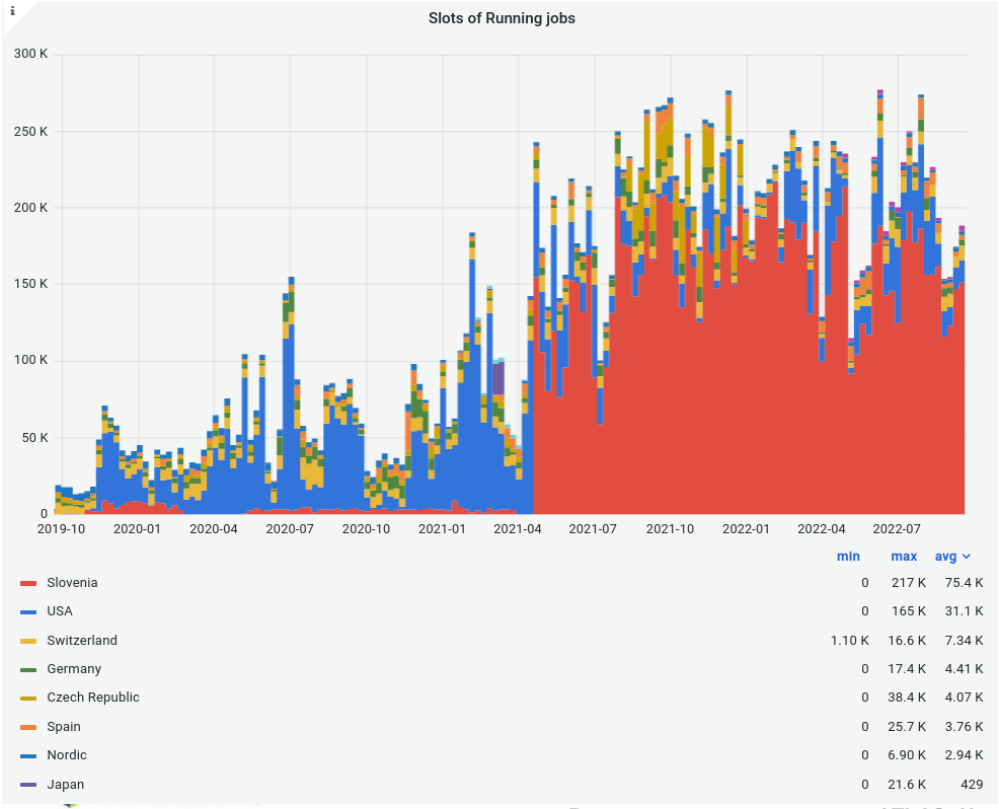
Еднаквото между всички..

Няма такова..

И така до май 2021..

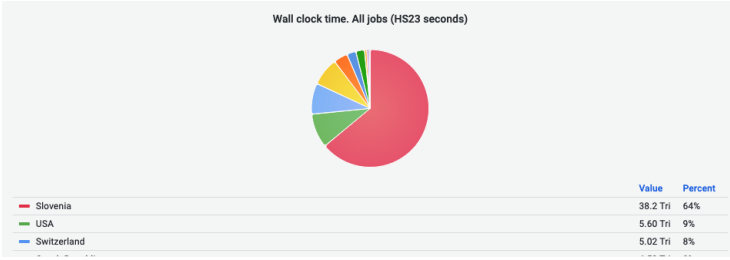
Brookhaven National Laboratory

Vega



Vega е:

- Суперкомпютър / HPC
- Открит на 20.04.2021
- В Словения
- Финансиран от EuroHPC и Словения
- На 166-то място в top500
- Atos/BullSequana XH2000
- AMD EPYC 7H12 2.6GHz
- 122800 ядра
- Производителност: 375 PFLOP/сек
- Не се използва напълно от потребителите
- ATLAS използва ресурсите в режим “backfill”



EuroHPC JU

(European High-Performance Computing Joint Undertaking / Европейска HPC инициатива)

- Инициатива между държавни и частни институции за консолидиране на ресурси на европейско ниво за развитие на HPC
- Членове: ЕС, Австрия, Белгия, България, Хърватска, Кипър, Чахия, Дания, Естония, Финлндия, Франция, Германия, Гърция, Унгария, Исландия, Ирландия, Италия, Латвия, Литва, Люксембург, Малта, Холандия, Северна Македония, Норвегия, Полша, Португалия, Румъния, Сърбия, Словакия, Словения, Испания, Швеция, Турция, представители на 3 частни партньори
- Цел: Разработка на пан-европейста HPC инфраструктура
- Централа: Люксембург
- Начало: 2018
- Финансиране: 7 мрд. Евро за периода 2021-2027
  - 3 мрд. - дългосрочен бюджет на ЕС
  - 3 мрд. - членове на инициативата
  - 1 мрд. - частни донори

Brookhaven National Laboratory

Суперкомпютрите на EuroHPC JU

LUMI (Финландия, top500: 3)
Hewlett Packard/Cray EX
AMD EPYC 64C 2GHz
2,220,280 ядра
Производителност: 428 PFLOP/сек

MELUXINA (Люксембург, top500: 52)
Atos/BullSequana XH2000
AMD EPYC 7452 32C 2.35GHz
172544 ядра
Производителност: 13 PFLOP/сек

LEONARDO (Италия, top500: 4)
Atos/BullSequana XH2000
Intel Xeon Platinum 8358 32C 2.6 GHz
1,824,768 ядра
Производителност: 305 PFLOP/сек

KAROLINA (Чехия, top500: 95)
Hewlett Packard
AMD EPYC 7763 64C 2.45GHz
102272 ядра
Производителност: 9 PFLOP/сек

MARENOSTRUM 5 (ще бъде в Испания)
Bull SAS/Bull Sequana XH3000
Intel Sapphire Rapid
Производителност: 314 PFLOP/сек

DISCOVERER (България, top500: 134)
Atos/BullSequana XH2000
AMD EPYC 7H12 64C 2.6GHz
144384 ядра
Производителност: 6 PFLOP/сек

VEGA (Словения, top500: 131
Atos/BullSequana XH2000
AMD EPYC 7H12 2.6GHz
122800 ядра
Производителност: 7 PFLOP/сек

DEUCALION (Португалия, в строеж)
Fujitsu/PRIMEHPC+BullSequana
A64FX, AMD EPYC
Производителност: 7 PFLOP/сек

Надпреварата

- **Целта:** Пускане в експлоатация на компютър който може да направи 10^{18} изчисления за секунда (exaFLOP)
  - Целта е постигната през юни 2022 в САЩ!
- **Статус:** top500.org
- **Тест:** Linpack Benchmark
  - Задача: Решаване на система от линейни уравнения
  - Класиране: Само за линейни уравнения
  - Резултати:
    - R_{peak}: Теоретична максимална производителност
    - R_{max}: Измерена производителност
- **Резултати:**
  - Top10: Само 2 машини в Европа
  - Топ500:
    - Всички машини в таблицата са с производителност > 1 petaFLOP
    - Повечето машини са в САЩ
    - Китай спря да публикува тестовите си резултати
    - Нито една машина не е в България
- **Защо е необходимо?**

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States				
3	NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States				
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194
7	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
9	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States	485,888	121.40	188.65	

Brookhaven National Laboratory

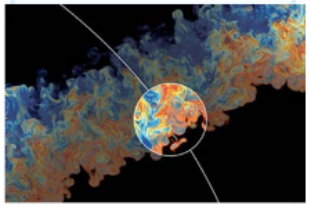
ECP applications target national problems in six areas

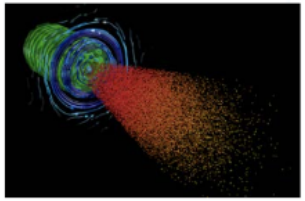
National security

Next-generation, stockpile stewardship codes

Reentry-vehicle-environment simulation

Multi-physics science simulations of high-energy density physics conditions





Energy security

Turbine wind plant efficiency

Design and commercialization of SMRs

Nuclear fission and fusion reactor materials design

Subsurface use for carbon capture, petroleum extraction, waste disposal

High-efficiency, low-emission combustion engine and gas turbine design

Scale up of clean fossil fuel combustion

Biofuel catalyst design

Economic security

Additive manufacturing of qualifiable metal parts

Urban planning

Reliable and efficient planning of the power grid

Seismic hazard risk assessment

Scientific discovery

Cosmological probe of the standard model of particle physics

Validate fundamental laws of nature

Plasma wakefield accelerator design

Light source-enabled analysis of protein and molecular structure and design

Find, predict, and control materials and properties

Predict and control stable ITER operational performance

Demystify origin of chemical elements

Earth system

Accurate regional impact assessments in Earth system models

Stress-resistant crop analysis and catalytic conversion of biomass-derived alcohols

Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Health care

Accelerate and translate cancer research

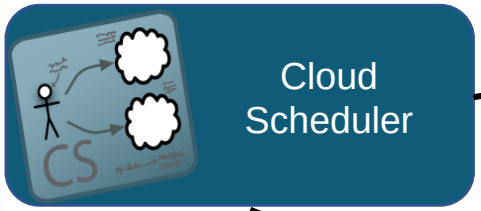
Взето от презентацията на Lori Diachin, зам-директор на "Exascale Computing Project", от конференцията "Supercomputing 18", 14.11.2018

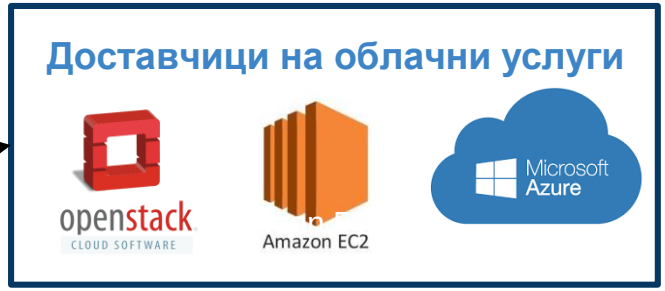
Облаци

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

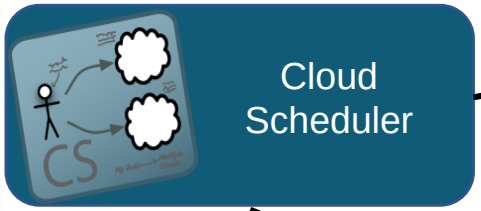




Изпраща
задачи

Brookhaven
National Laboratory

Принцип на работа



Направи ми виртуални машини

Доставчици на облачни услуги

openstack CLOUD SOFTWARE

Amazon EC2

Microsoft Azure

Наблюдава опашката от задачи

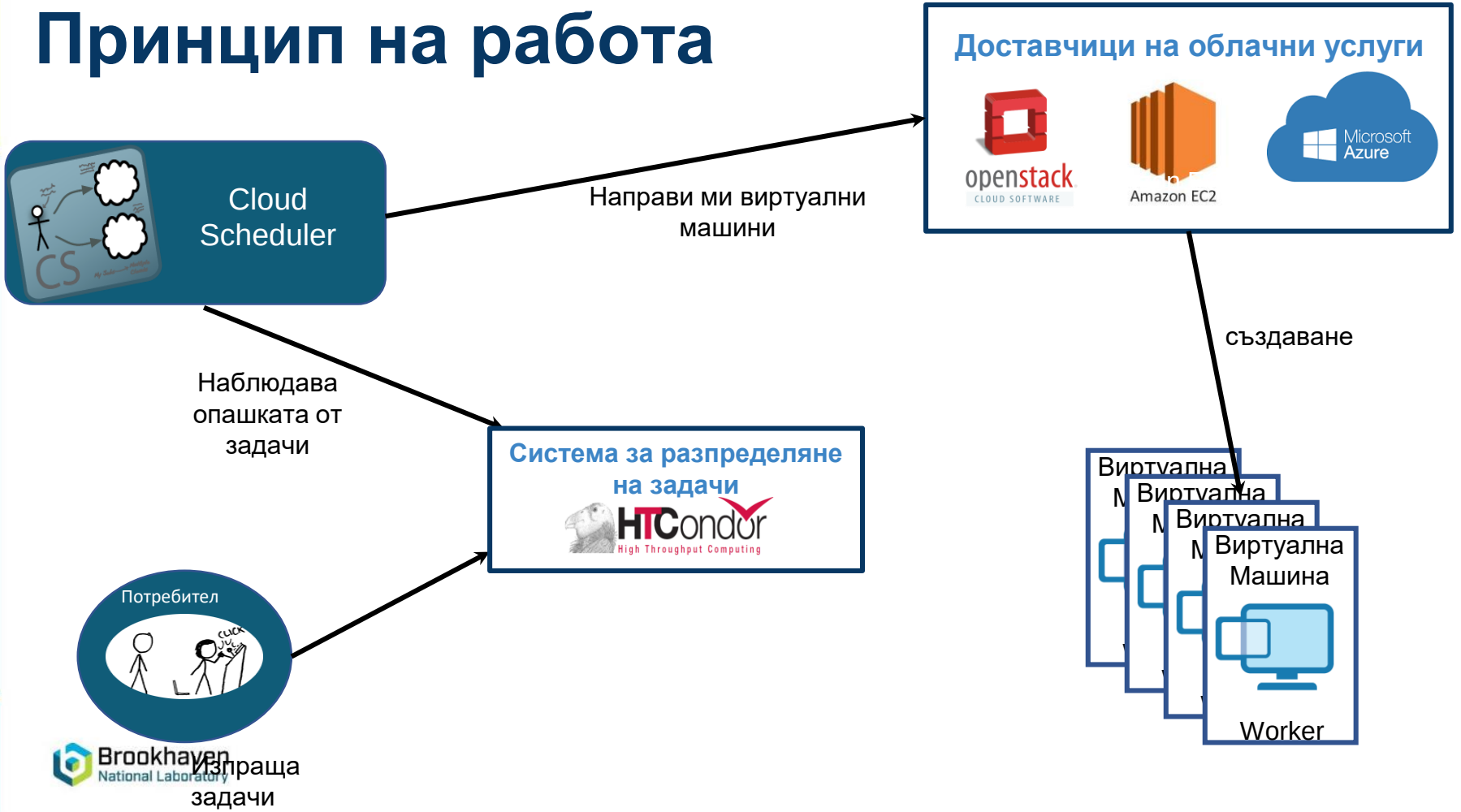
Система за разпределяне на задачи

HTCondor High Throughput Computing

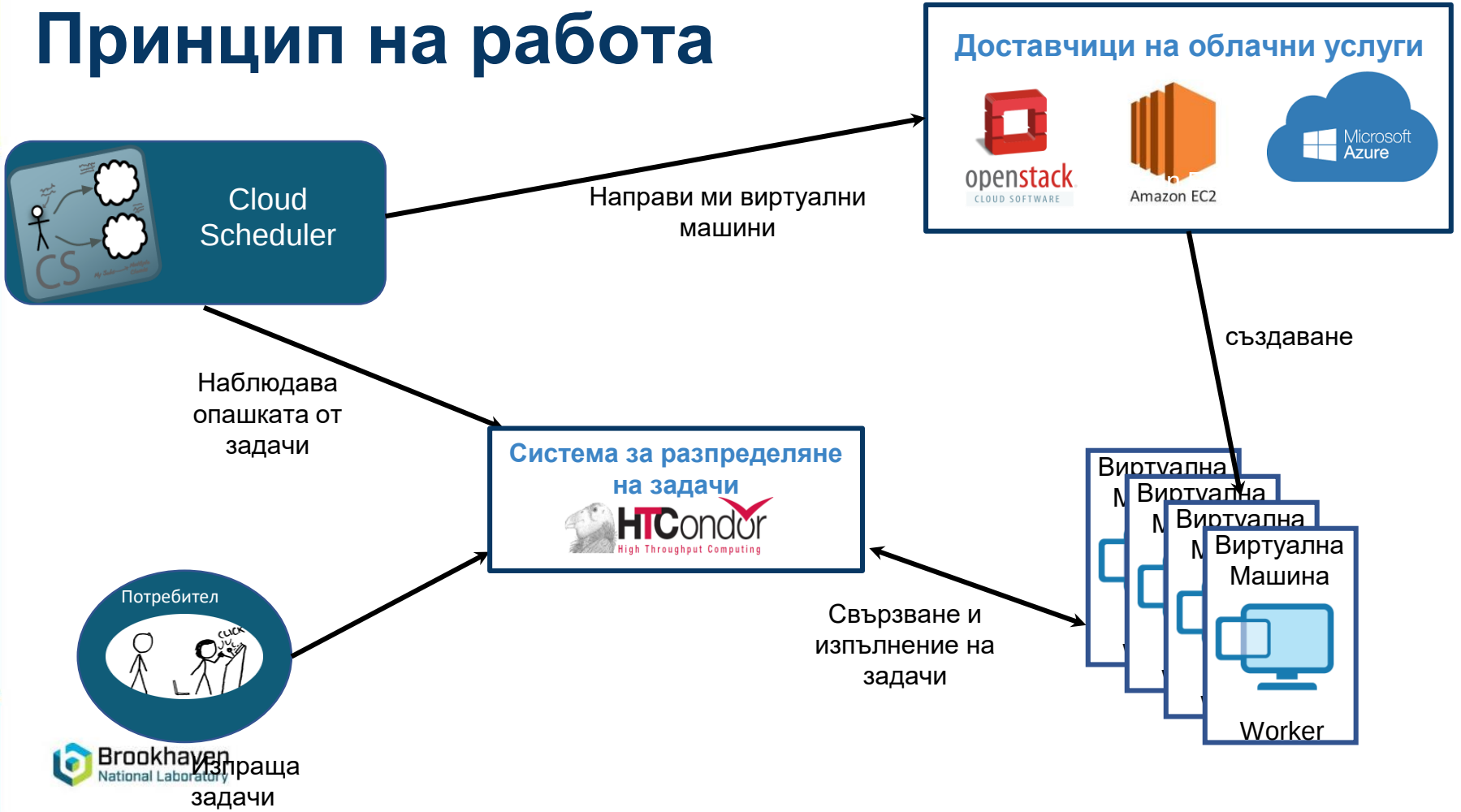
Изпраща задачи

Brookhaven National Laboratory

Принцип на работа

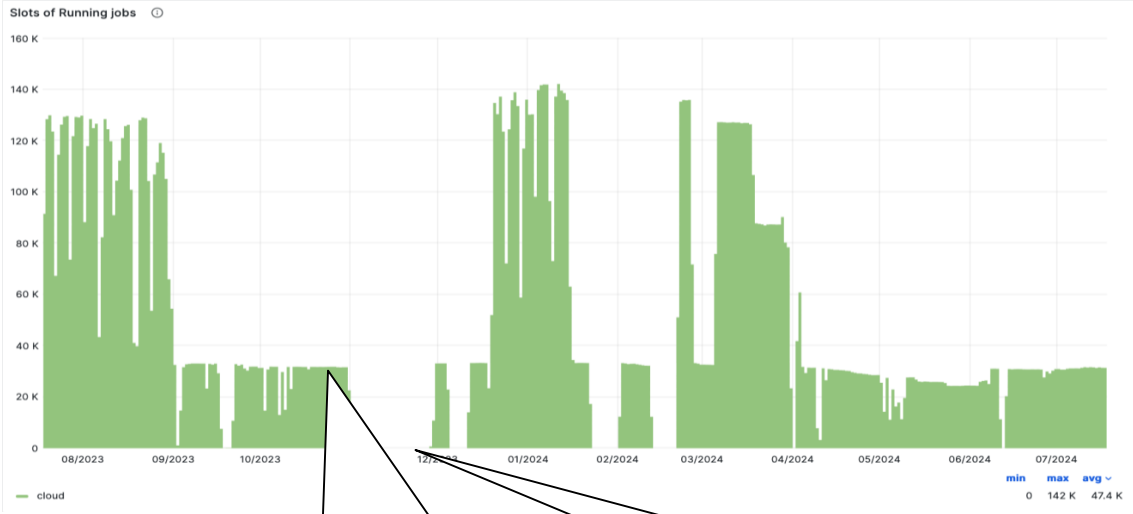


Принцип на работа



Собствени облаци - тригерна ферма

- Ресурс - 133 000 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)
- Оптимизирането на използването на ресурсите не е тривиален проблем



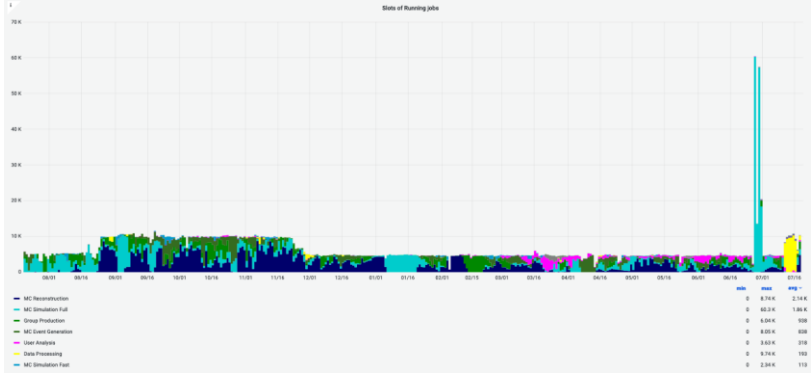
Използва се само когато ресурсите не са нужни за набирането на сурови данни от експеримента.

Периоди на профилактикла на системата

Brookhaven National Laboratory

Комерсиални облаци - Google

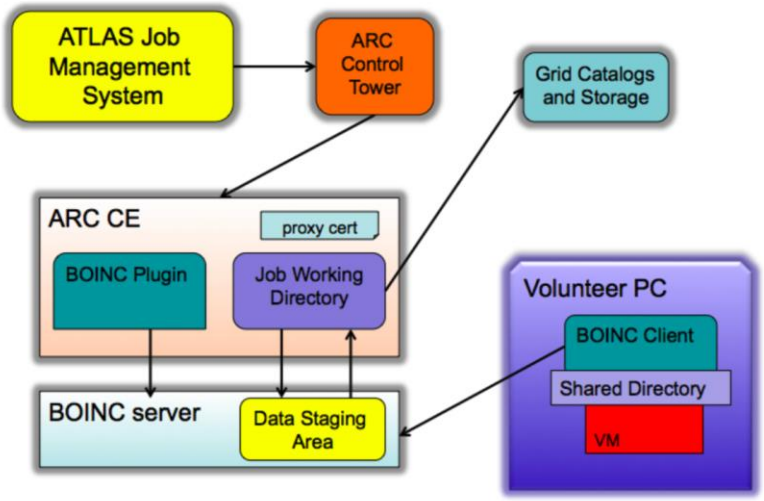
- Тестов договор с Google за да се видят предимствата и недостатъците
  - Цена
  - Производителност
- Включен успешно към системата за разпределени изчисления на АТЛАС

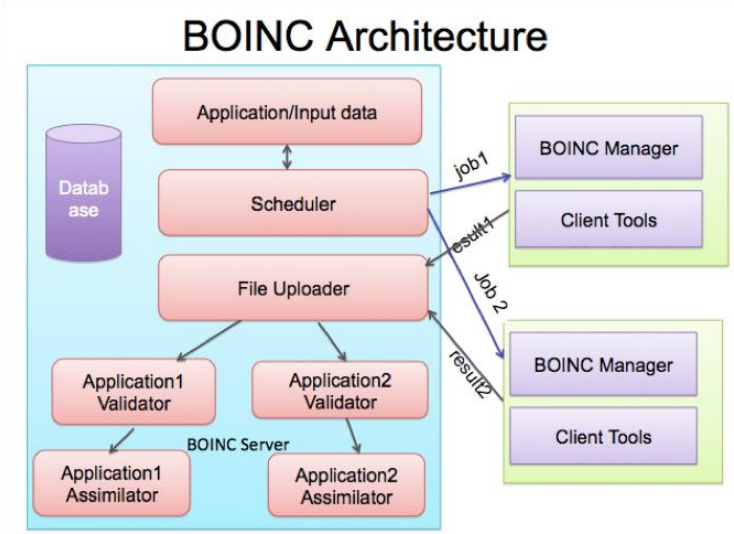


Доброволчески споделени изчисления
(Volunteer Computing)

Brookhaven National Laboratory

BOINC

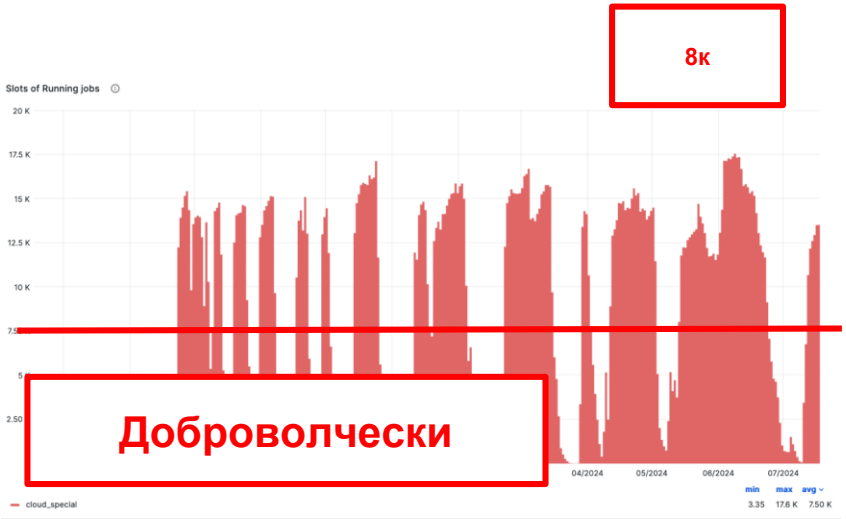




Brookhaven National Laboratory

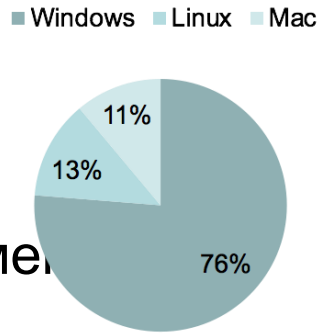
ATLAS@Home

- Обработва ~ 1% от симулираните събития 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

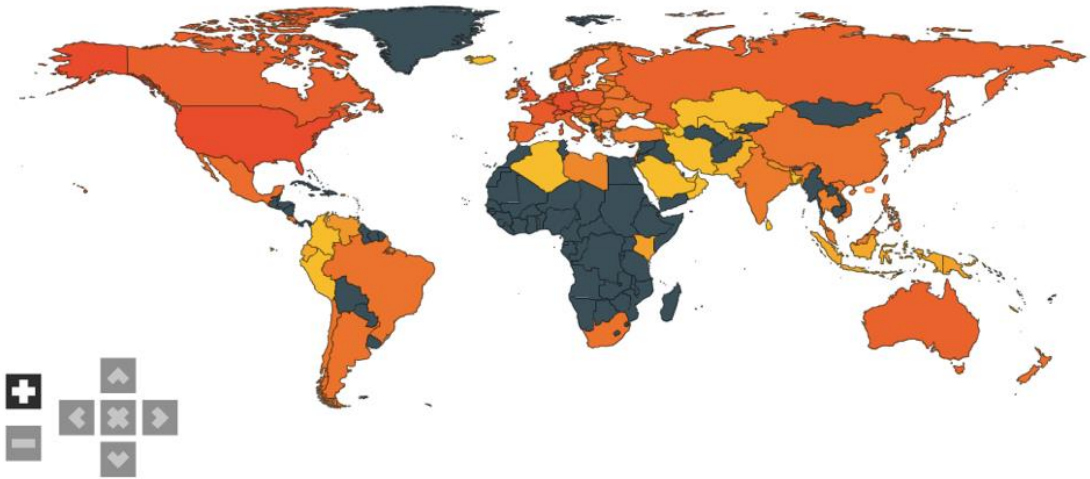


Brookhaven National Laboratory

Кой ни помага



- BOINC е огромно общество от отдадени хора
- Много от тях участват е няколко проекта едновреме



И ние инсталираме BOINC

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

Обобщение

- GRID е все още тук и не се очаква скоро да си тръгне
- Все повече организации и спонсори избират алтернативи на конвенционалните компютърни центрове - суперкомпютри, облаци..
- Колаборацията ATLAS успешно ги интегрира и използва


Brookhaven
National Laboratory
U.S. DEPARTMENT OF
ENERGY
Въпроси?
Ivan.Glushkov@cern.ch
@BrookhavenLab

Технологии в ЦЕРН