Integrated data and jobs management system for heterogeneous computing environment

Единая система управления данными и задачами в гетерогенной компьютерной среде

А. Новиков, В. Аулов, Д. Дрижук, А. Климентов,
Р. Машинистов, А. Пойда, И. Тертычный.

НИЦ «Курчатовскский Институт», НБИКС-центр
Лаборатория технологий больших данных для проектов в области мега-сайенс

26/05/2015, РИВС-XXI, Дубна

Содержание

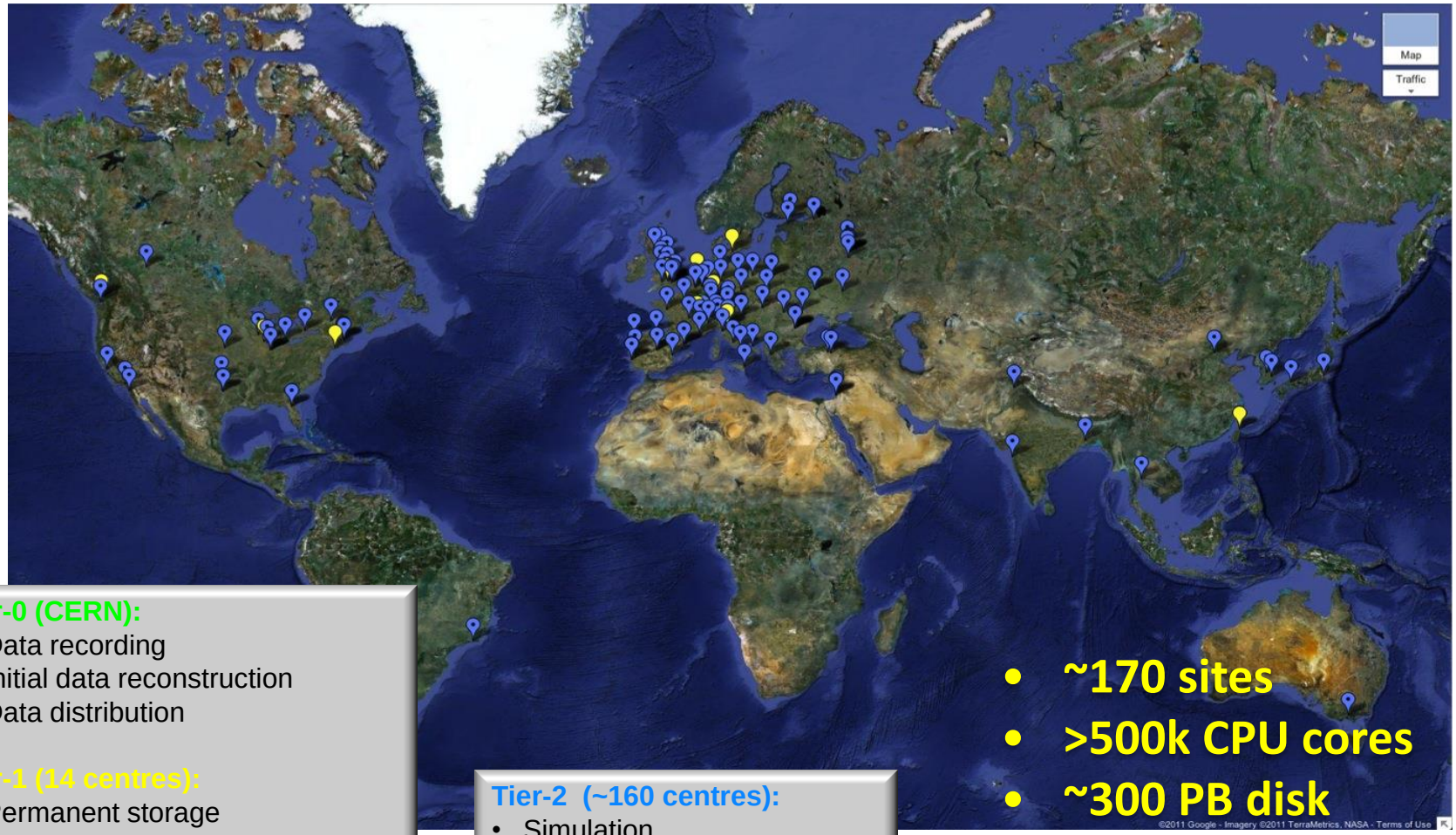
1. Введение
   - bigData, грид, облака и суперкомпьютеры
2. Система управления рабочим потоком PanDA для эксперимента ATLAS LHC. Особенности и архитектура.
3. Проект MegaPanDA, основные цели.
4. Деятельность НИЦ «Курчатовский институт» (НИЦ КИ) в рамках MegaPanDA:
   - Адаптация PanDA к HPC НИЦ КИ [1]
   - Поддержка не HEP областей наук, на примере, биоинформатики [2]
5. Заключение

Подробнее в докладах:

1 - Интеграция суперкомпьютеров в Грид-инфраструктуру на примере Курчатовского института. Д.Дрижук, А.Аулов, А.Климентов, Р.Машинистов, А.Новиков, А.Пойда, И.Тертычный

2 - Интерфейс системы управления данными и задачами, требующими высокоинтенсивных и высокопроизводительных вычислений. В. Аулов, Д.Дрижук, А.Климентов, Р.Машинистов, А.Новиков, А.Пойда, И.Тертычный

Мировые Грид ресурсы



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

Tier-1 (14 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~160 centres):
- Simulation
- End-user analysis

- **~170 sites**
- **>500k CPU cores**
- **~300 PB disk**

http://wlcg.web.cern.ch/

Мировые центры данных Грид, Amazon, Google

- LHC Computing (WLCG)
  - 170 сайтов
  - 500k CPU ядер
  - ~5000 пользователей
- Amazon : 9 больших сайтов/зон
  - до ~2M CPU ядер/сайт, всего ~4M - в 10 раз больше ядер на в 1/10 сайтов чем в грид
  - 500 тысяч пользователей
- Google: ~13-40? центров, стоимостью ~$600M, 100-200МВт каждый (Ломоносов МГУ 52к ядер ~$63M, 2.6-3.0МВт т.1.7Pflops; Tianhe2 Китай 3.1M ядер 24МВт т.27Tflops)

Amazon

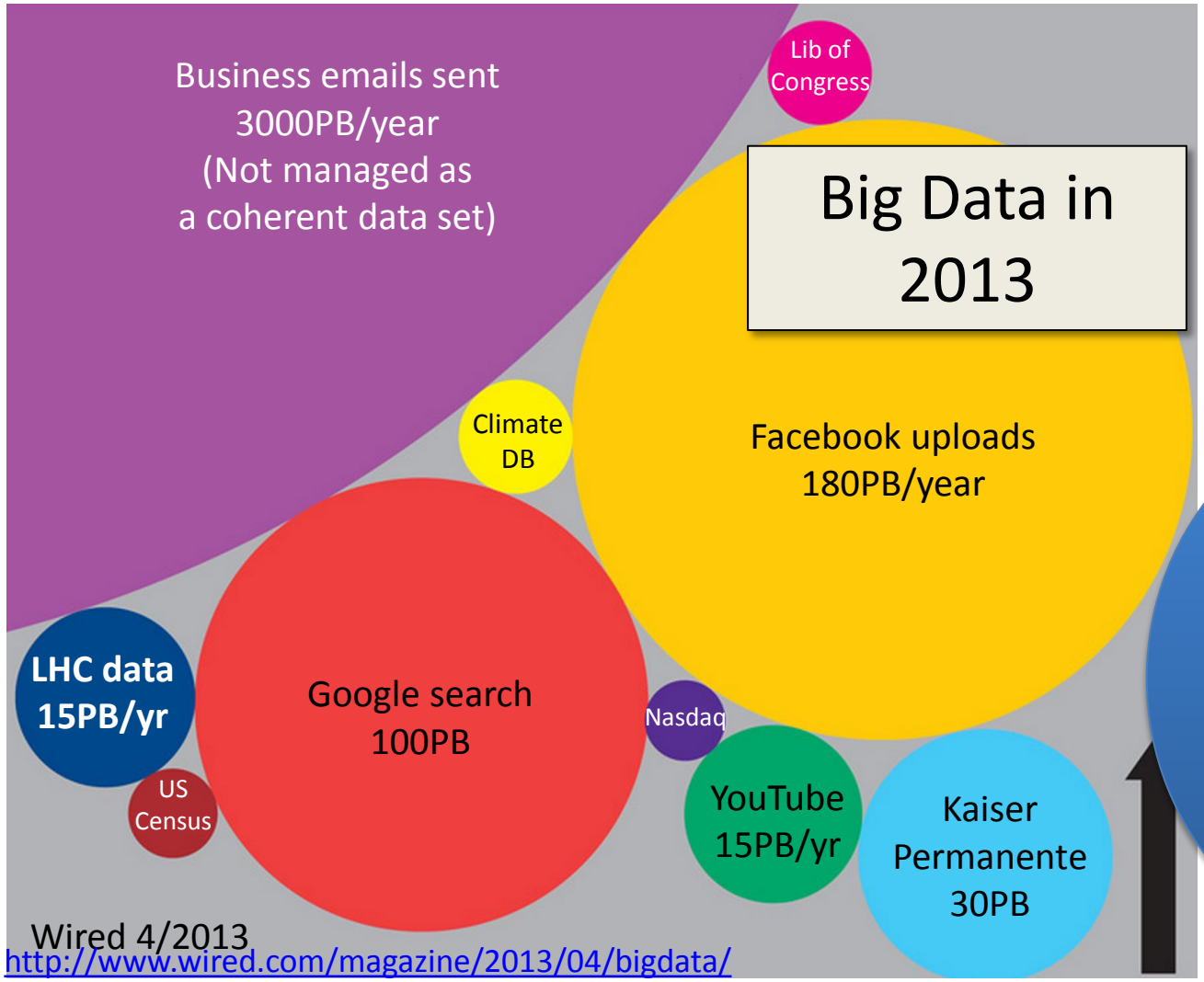
Google

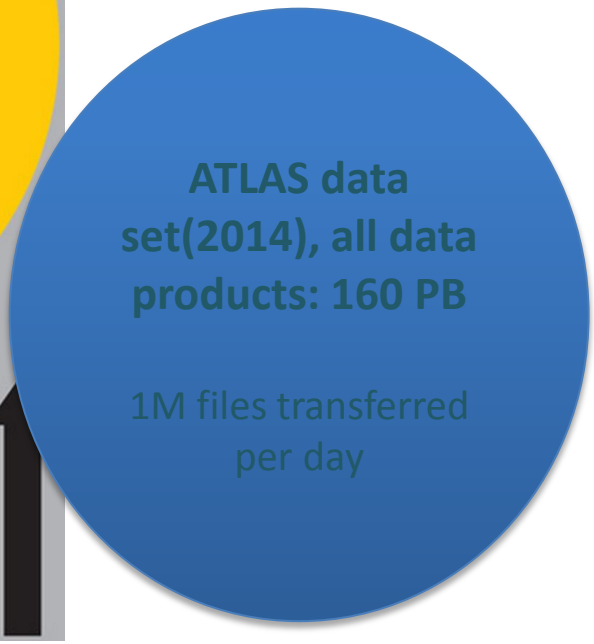
Суперкомпьютеры top500 Nov2014

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510

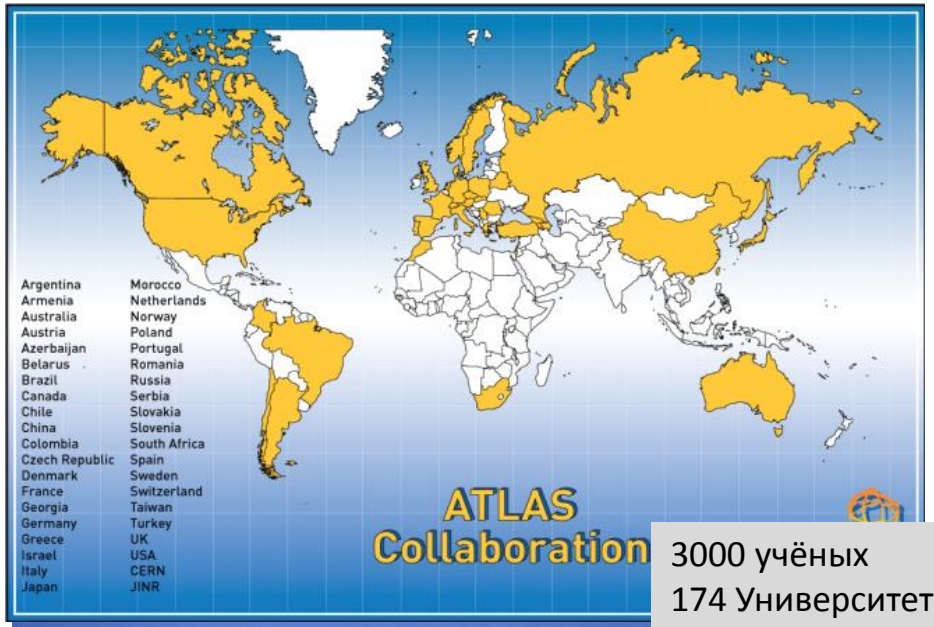
http://www.top500.org/lists/2014/11/

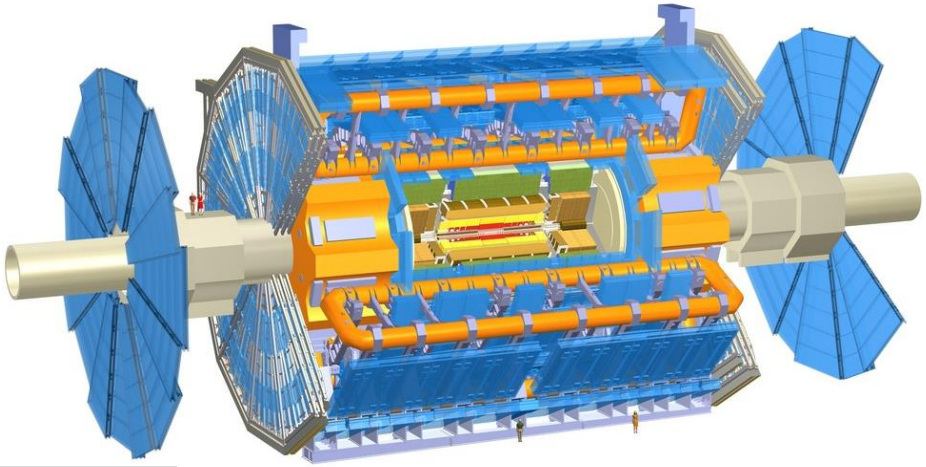
BigData


Business emails sent
3000PB/year
(Not managed as
a coherent data set)
Lib of Congress
Big Data in 2013
Climate DB
Facebook uploads
180PB/year
LHC data
15PB/yr
Google search
100PB
Nasdaq
US Census
YouTube
15PB/yr
Kaiser
Permanente
30PB
Wired 4/2013
http://www.wired.com/magazine/2013/04/bigdata/


ATLAS data set(2014), all data products: 160 PB
1M files transferred per day

Эксперимент ATLAS LHC





3000 учёных
174 Университетов и лабораторий из
38 стран
>1200 студентов

The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

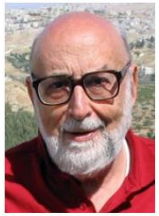


Photo: Pnicolet via Wikimedia Commons
François Englert

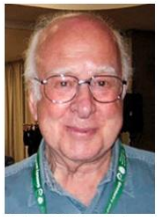


Photo: G-M Greuel via Wikimedia Commons
Peter W. Higgs

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

Эксперимент ATLAS LHC. «Обещанные»(pledged) ресурсы

По всем Tiers доступно:

Physical CPU	Logical CPU	HEPSPEC06	CPU Pledge
179,625	532,910	5,716,182	3,083,512
Total Online Storage,Gb	**Disk Pledge, Gb**	**Total Nearline Storage, Gb**	**Tape Pledge, Gb**
325,934,747	249,432,000	240,447,948	274,592,000

Из них для ATLAS:

	CPU, HEP-SPEC06	Disk, Tbytes	Tape, Tbytes
Требуется	1175000	103000	98000
В наличии	1275226	110293	103190
% от общих	22	34	43

http://wlcg-rebus.cern.ch/apps/pledges/resources/

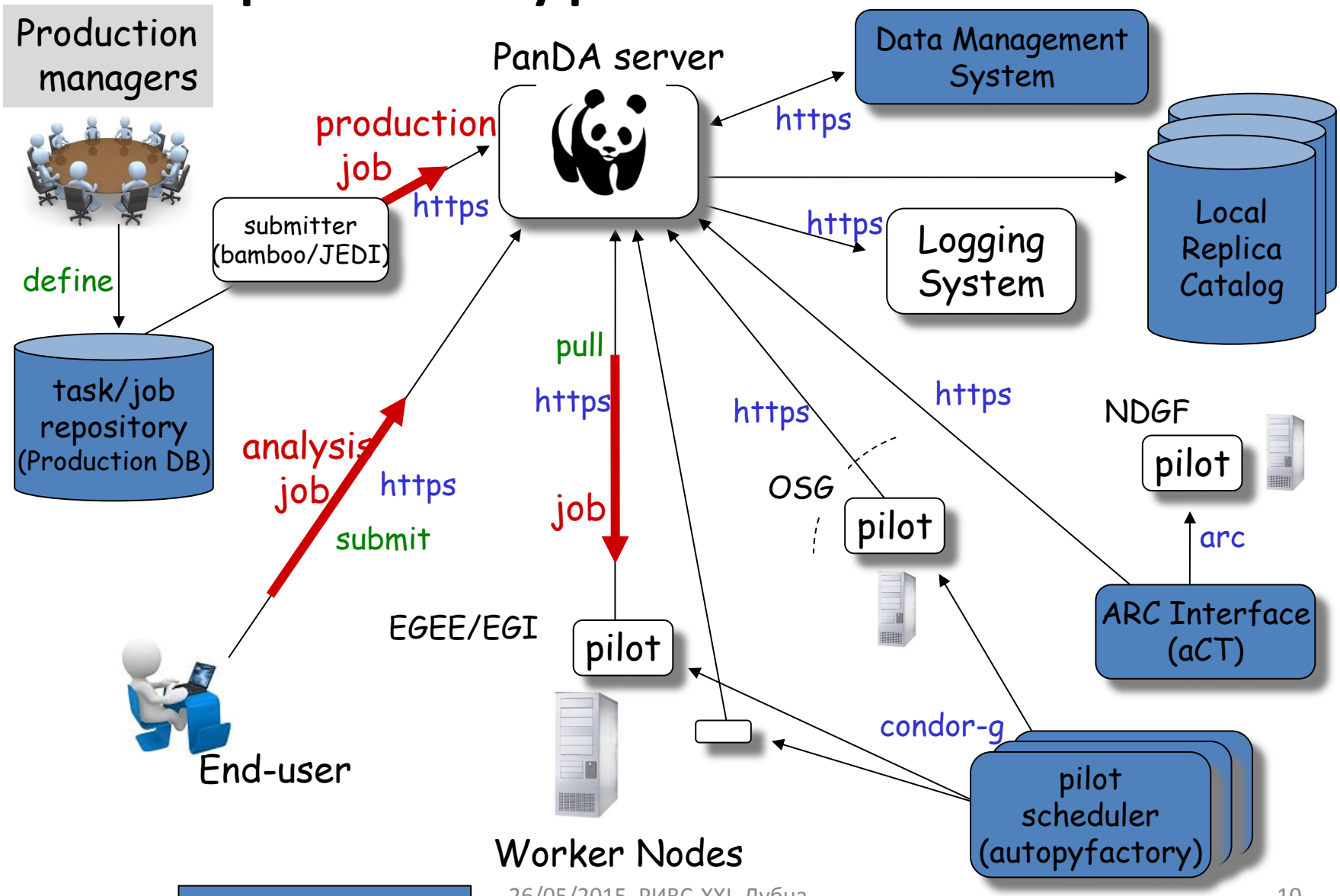
Что такое PanDA?

PanDA - Production and Distributed Analysis, проект, разработанный для поддержки рабочего потока задач и данных эксперимента ATLAS (LHC CERN) и др.

Ключевыми идеями явились:

1. Обеспечение для пользователей единой центральной очереди (из сотен сайтов).
2. Снижение ошибок и издержек на уровне поддержки сайта – с помощью системы запуска пилотных заданий (pilots).
3. Поддержка различных сред ППО (middleware, и их версий) с унифицированным представлением высокоуровневого рабочего потока для пользователей и использованием системы.
4. Скрыть средства автоматизации управления рабочего потока от пользователей. Например, пользователю предоставляется возможность запустить задание над набором входных данных, а система автоматически определит, как разбить задание на задачи, где и когда запускать и перезапускать их.
5. Поддерживать возможность простой интеграции грид и облачных сайтов, а так же HPC.
6. Использовать единую систему управления рабочим потоком(PanDA WMS) для задач генерации событий (аналитических), обработки данных (экспериментальных) и пользовательских задач анализа.

Архитектура PanDA WMS



NonPanDA components

Что такое PanDA Pilot

- PanDA Pilot – легковесное приложение для управления выполнением рабочего потока на некоторых вычислительных ресурсах.
- PanDA Pilot обеспечивает:
  - проверку актуально доступных ресурсов, рабочего окружения, установленного ПО в соответствии с абстрактным описанием ресурса в информационной системе и поддерживаемым экспериментом;
  - запрос информации о рабочем потоке(задачах) от сервера PanDA в режиме поздней привязки к ресурсу;
  - настройку окружения, специфического для ВО(проекта);
  - загрузку входных данных и выгрузку результатов (в т.ч. логов работы пилота и приложения);
  - отслеживание выполнения рабочего потока на узле, включая обновление его статуса на сервере PanDA;
  - возобновление ошибочных задач.

PanDA Pilot

PanDA Pilot имеет модульную архитектуру, что позволяет ему гибко поддерживать различные инфраструктуры (и вычислительные среды) с помощью плагинов, в т.ч. возможно использование различных технологий передачи файлов.

Привязка PanDA к проекту ATLAS

ATLAS специфичность в:

- Pilot – модульный, но много включений, характерных для проекта.
- Работа с планировщиком CondorG – стабильность, но не для всех ресурсов.
- Рассчитан на схему 1 пилот – 1 задача – 1 ядро или узел, а не на поддержку MPI задач.
- Database – центральная или локальная, апдейты привязаны к Oracle(не полностью public).
- Сертификат X.509 с VO role=atlas для запуска пилотов.
- Storage – система метаданных RUCIO (центральный MQ сервер с очередями для перемещения данных).

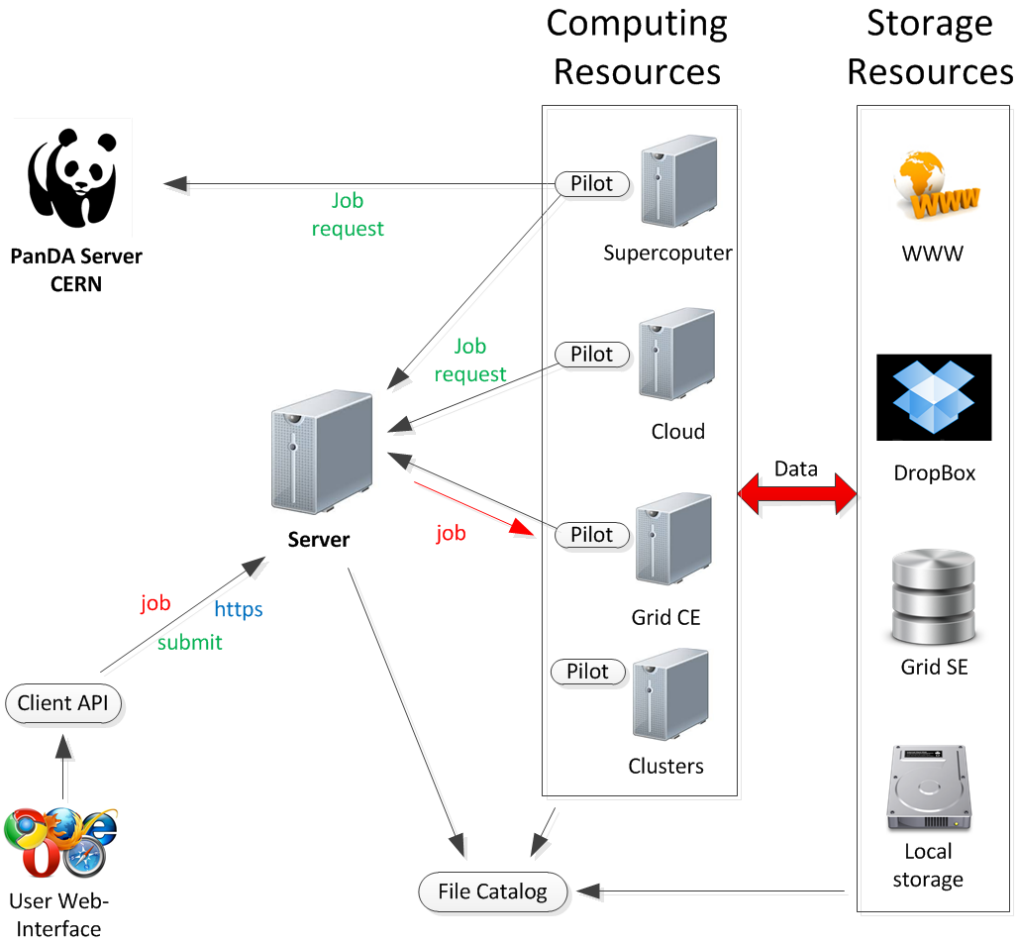
Проект MegaPanDA

Поддержка областей наук за пределами HEP (ATLAS, etc), например, задачи биоинформатики, астрофизики и др..

Участие в:

- **Адаптация запуска PanDA задач на суперкомпьютерах**
- Поддержка и развитие «mysql» версии сервера PanDA.
- Адаптация центральной базы к NoSQL-решениям.
- Рефакторинг кода модулей PanDA.
- Пакетирование.

Архитектура реализуемой системы



Суперкомпьютер НИЦ КИ

Высокопроизводительный вычислительный кластер HPC2 второго поколения с пиковой производительностью 122,9 TFLOPS сдан в эксплуатацию с сентября 2011 года. В 15-ой редакции российского рейтинга суперкомпьютеров top50 он занимает позицию #2

Адаптация к суперкомпьютеру НИЦ КИ

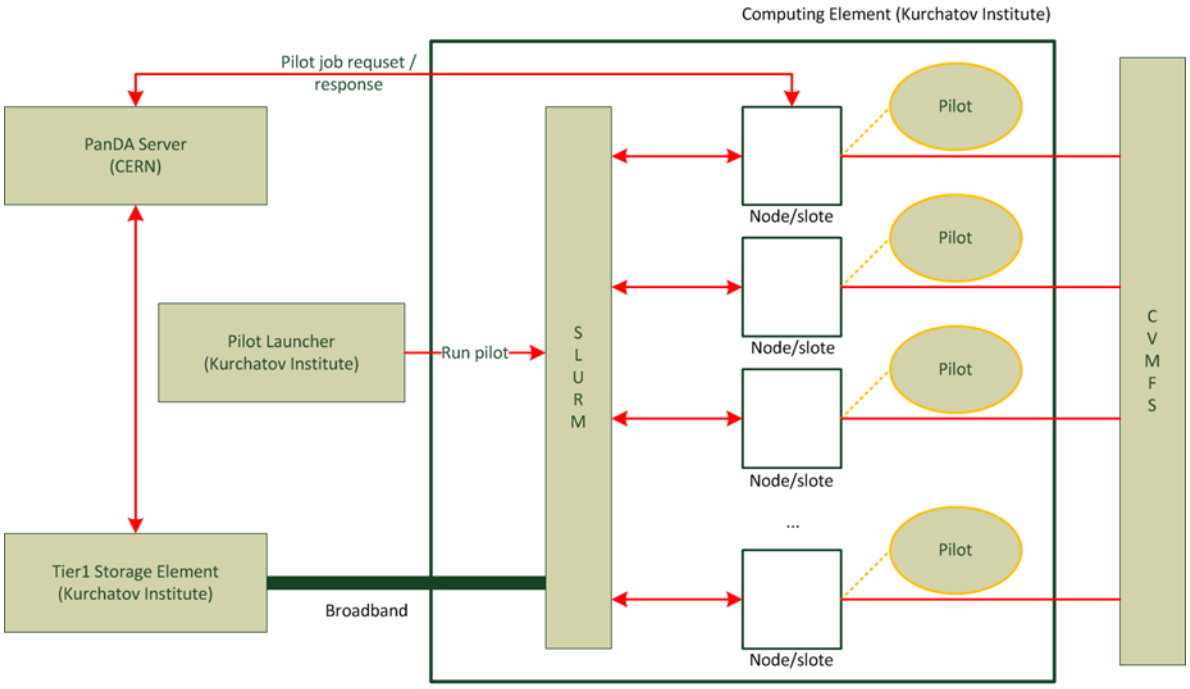
Особенности суперкомпьютера:

- 10240 ядер = 1280 узлов 2x Intel Xeon E5450 3,00ГГц 4 ядра 16 Гб RAM; 122.9 Tflops.
- Точка входа на суперкомпьютер(UI) рассчитана только на запуск задач в очередь или компиляцию приложений.
- Общая файловая система(Lustre) для рабочих узлов(WNs) доступна на UI.
- У всех рабочих узлов есть прямой выход в WAN.
- К WNs подключена CVMFS .
- Организован скоростной доступ к хранилищу Tier-1 SE грид сайта ANALY_RRC-KI.
- Система запуска задач SLURM (не CondorG).

Реализована система удалённого запуска через ssh, с поддержкой синхронизации выполнения задач, входных и выходных файлов.

Архитектура PanDA@HPC_KI для запуска ATLAS задач

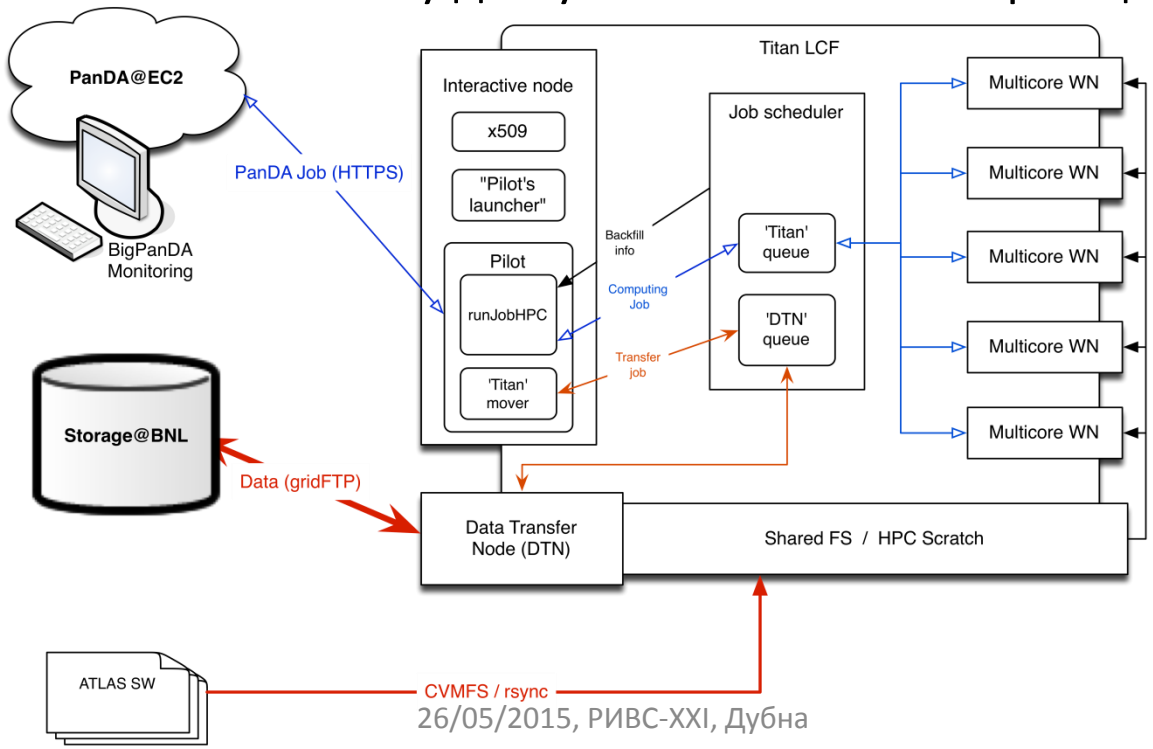
- Пилоты запускаются на каждом узле
- Число пилотов равно числу свободных ядер по числу виртуальных слотов CondorLocal



Архитектура PanDA@HPC_Titan

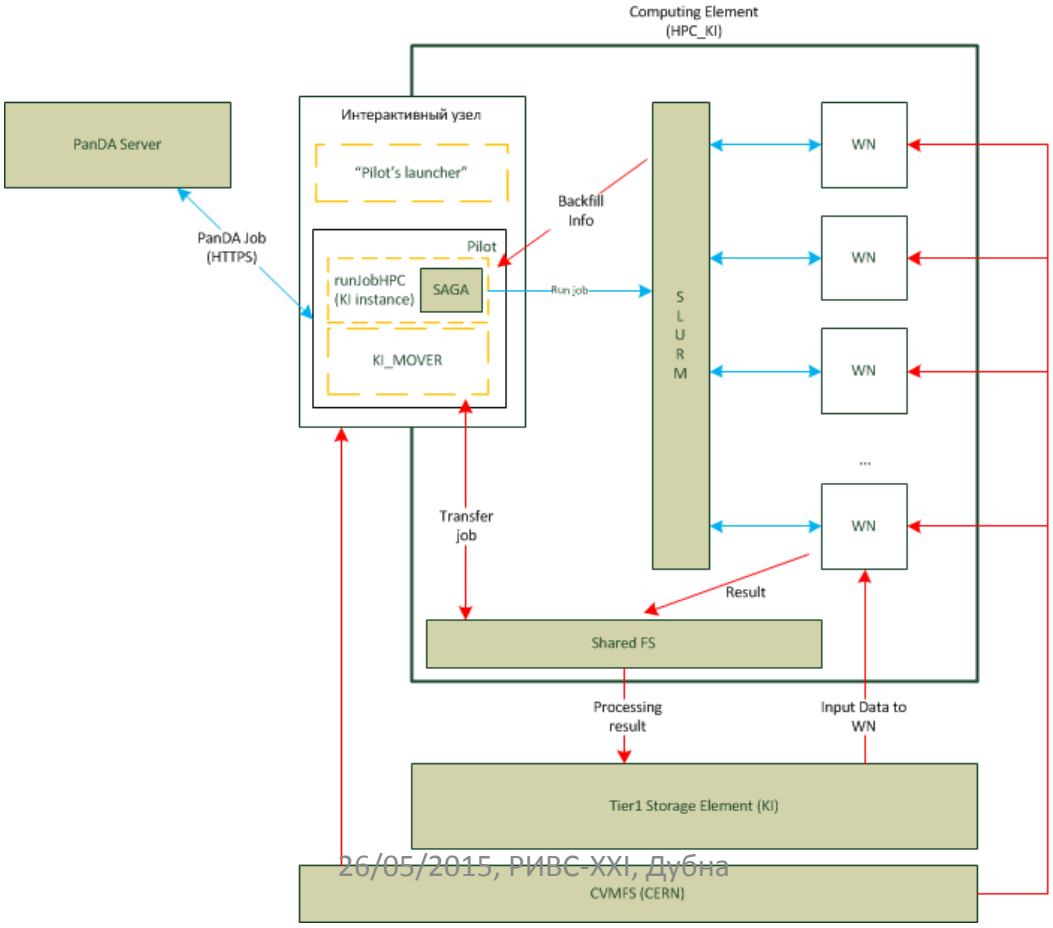
Адаптация PanDA для Titan OLCF, Oak Ridge (Д.Олейник, BNL):

- Пилоты запускаются на интерактивном узле суперкомпьютера
- Пилот взаимодействует с локальным планировщиком для управления задачей
- Число пилотов = числу доступных слотов планировщика



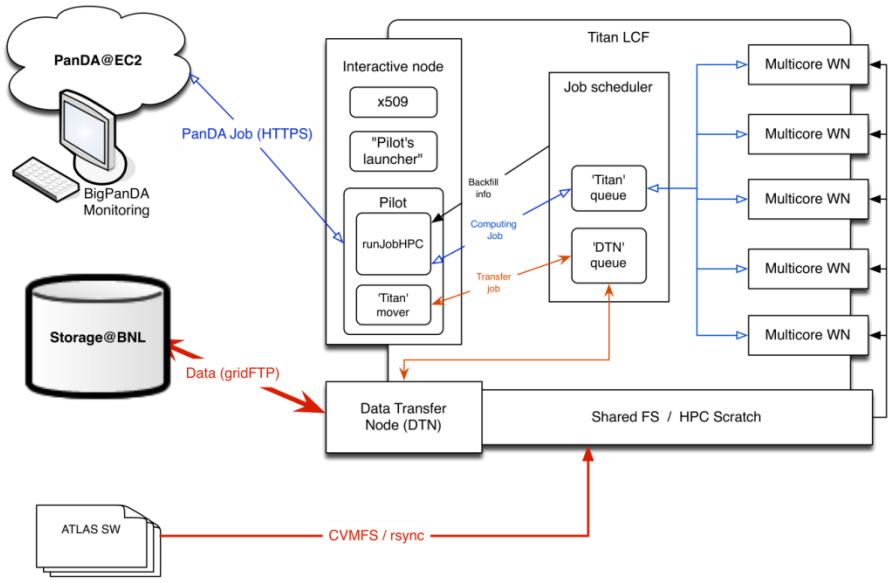
Альтернативная архитектура PanDA@HPC_KI

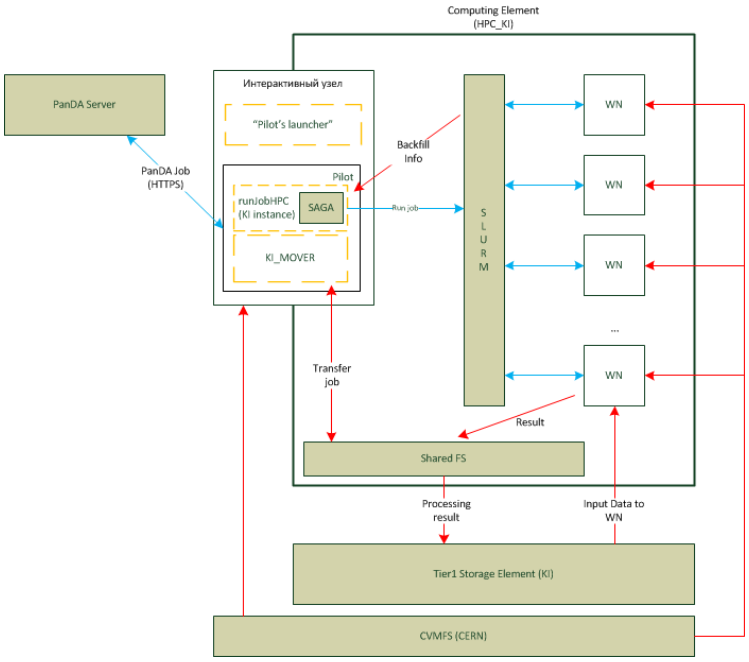
- Пилоты запускаются на внешнем узле или интерактивном узле
- Пилот запускает задачу на узле / узлах через SLURM
- Во время выполнения задача имеет доступ к Tier-1 с данными и к CVMFS



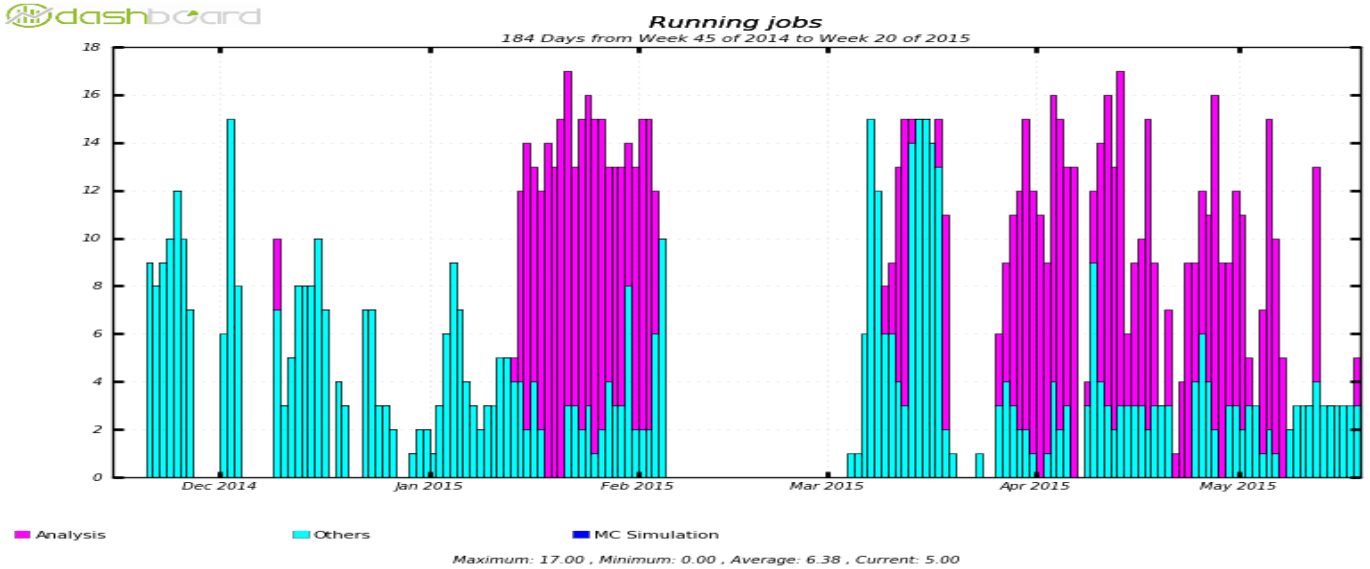
Сравнение двух архитектур

- Архитектура КИ использует HPC интерфейс пилота (реализован свой класс runJobHPC).
- В архитектуре КИ задача, запущенная на WN имеет доступ к CVMFS и к данным, хранящимся на Tier-1 КИ.

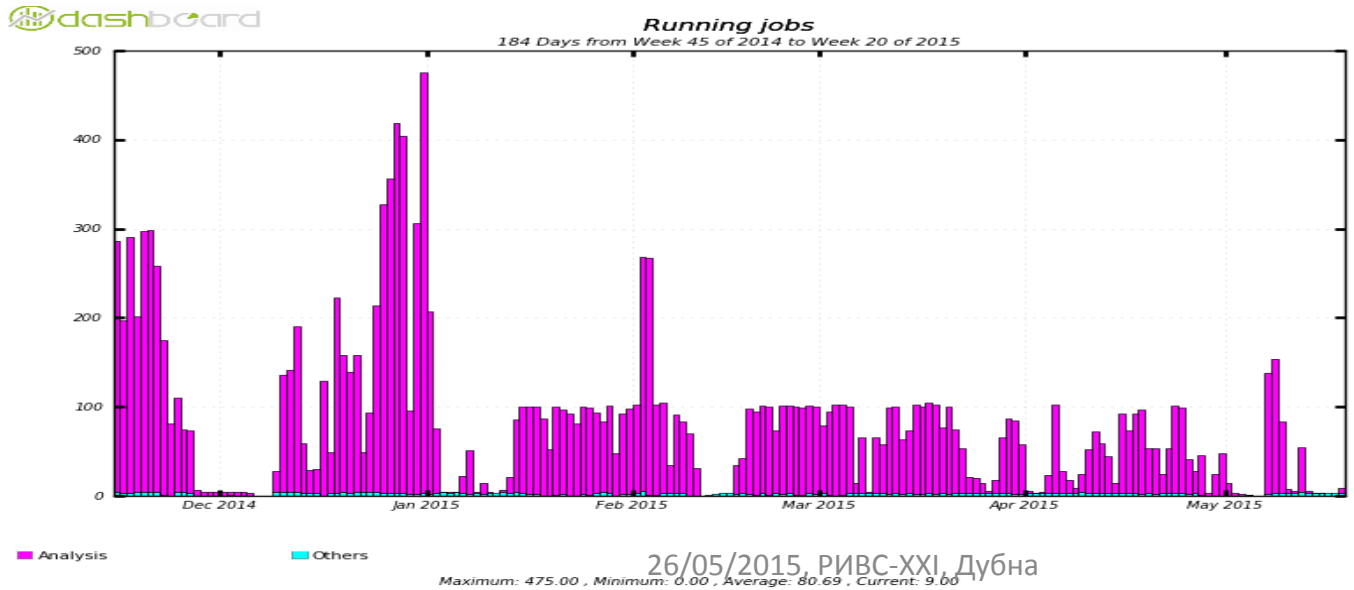




Статистика работы

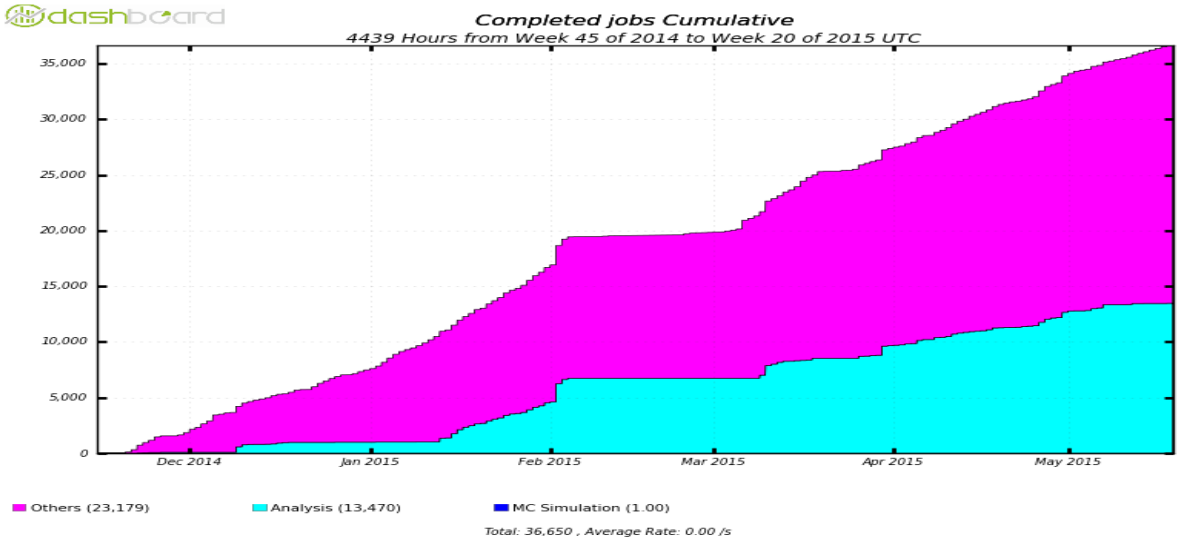


Тестовый PanDA сайт с использованием ресурсов суперкомьютера НИЦ КИ

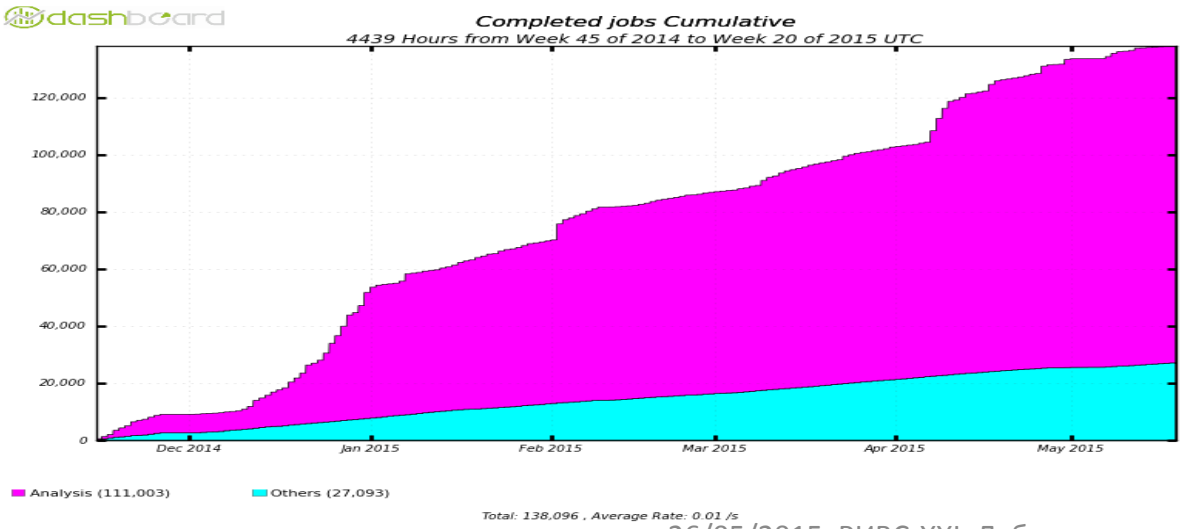


Tier1 грид PanDA ресурс НИЦ КИ

Статистика работы (2)



Тестовый PanDA сайт с использованием ресурсов суперкомьютера НИЦ КИ



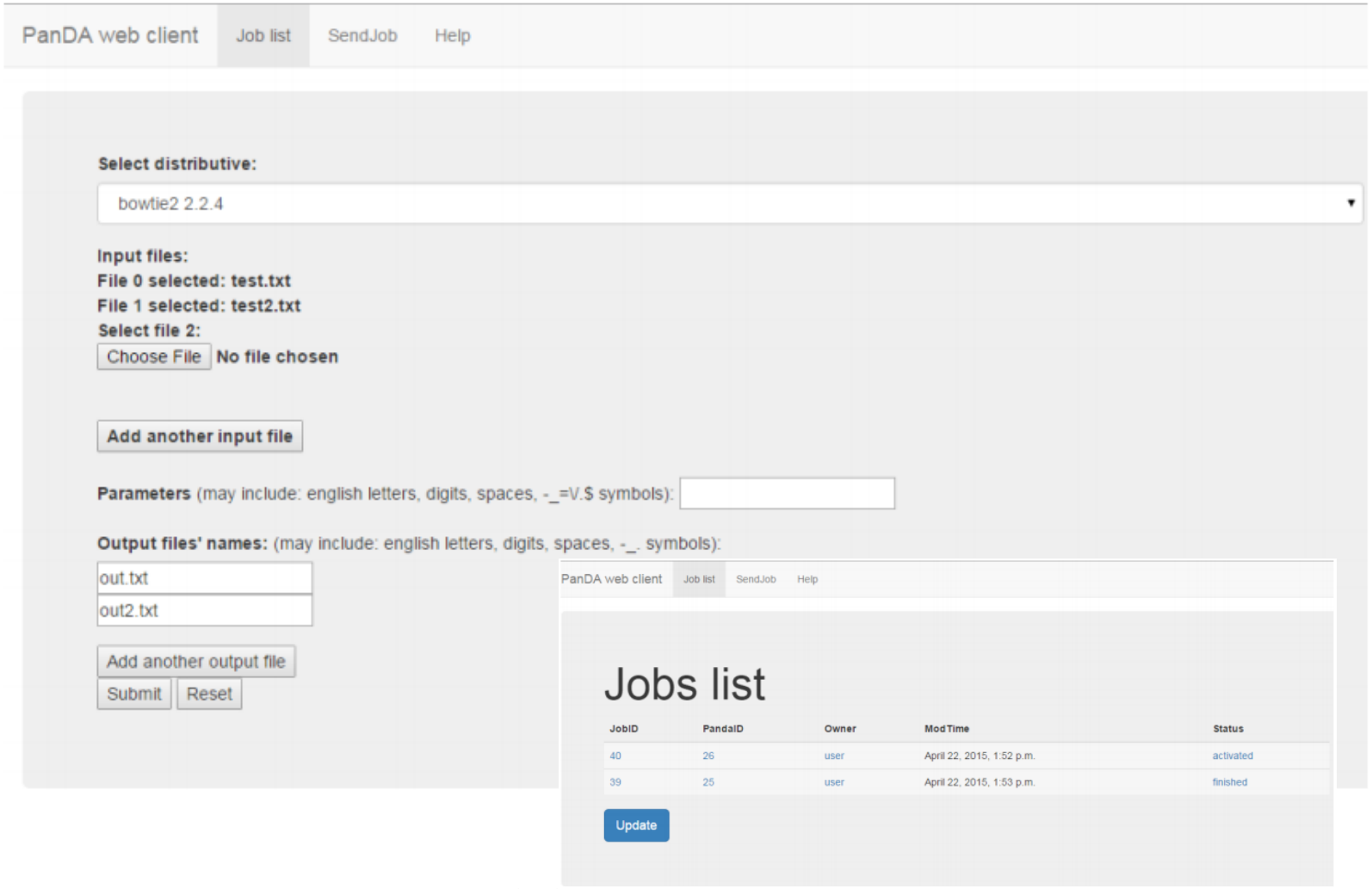
Tier1 грид PanDA ресурс НИЦ КИ

Поддержка задач биоинформатики

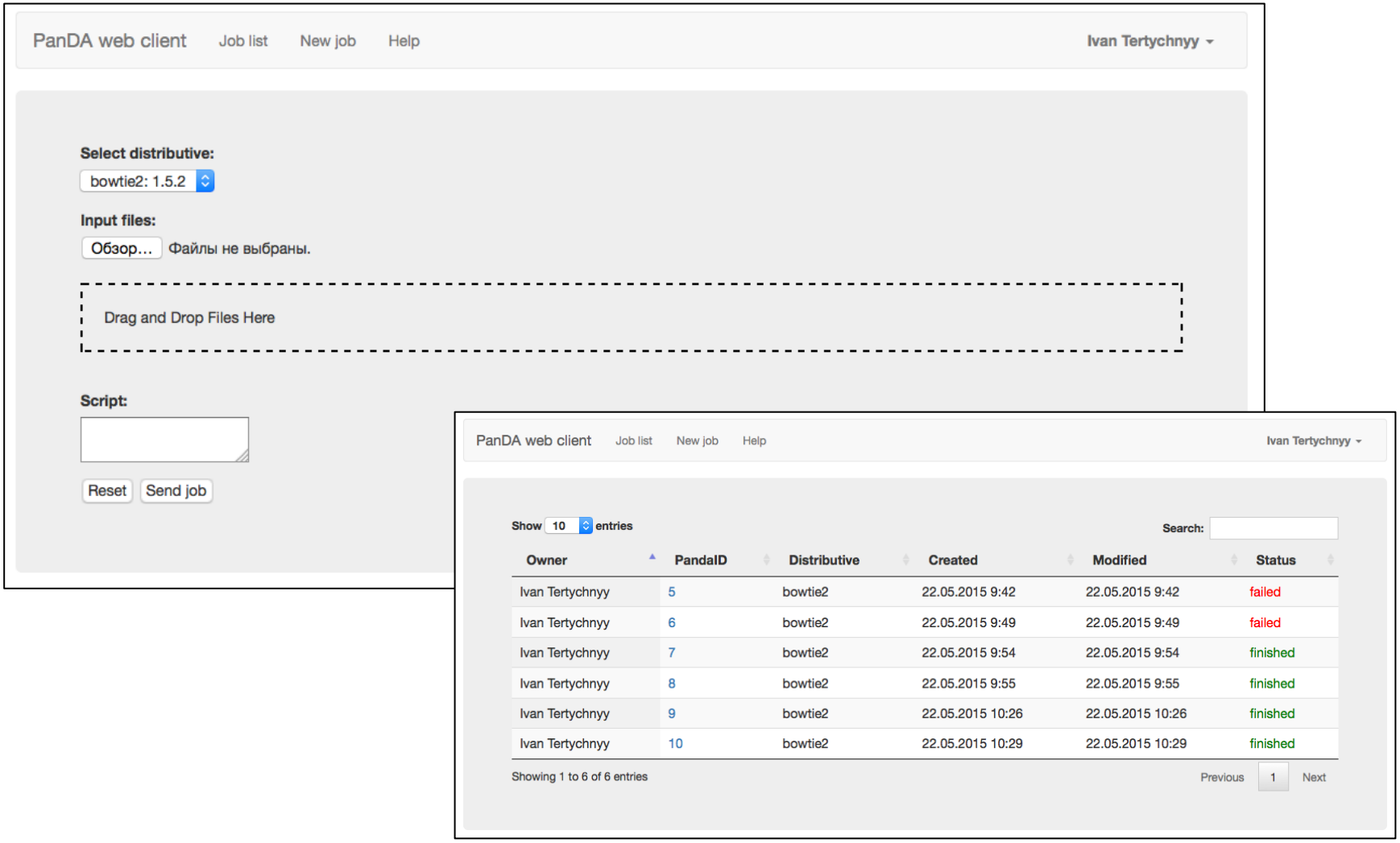
На примере задач выравнивания и сборки генома с помощью программного обеспечения Bowtie2 и Abyss:

- Более ресурсоёмкие(память не 1-2Гб/задача).
- Возможность работы с MPI(т.е. > 1 ядро/задача).
- Пользователи не готовы работать с Грид файловым каталогом (а спец. утилит у них нет – DQ2/RUCIO).
- Проблема грид X.509 сертификата для запуска задач в системе (его нет).
- Совместная обработка задач сторонних пользователей и эксперимента ATLAS (с центрального PanDA сервера).

Пользовательский интерфейс



Пользовательский интерфейс



Заключение и планы

Произведена интеграция PanDA WMS с суперкомпьютером НИЦ “Курчатовский институт”. (в т.ч. с облачной платформой на ресурсах НИЦ КИ).

Разработаны, обобщены и реализованы новые схемы работы пилотных заданий, позволяющих совместную обработку задач эксперимента ATLAS и иных проектов(областей наук).

Рассмотрены проблемы и требования запуска пользовательских задач иных областей наук, на примере задач биоинформатики, разрабатываются интерфейсы и портал для запуска их задач.

Представленный подход позволяет объединять вычислительные мощности различных инфраструктур для проведения масштабных вычислений в высокоинтенсивных областях науки, таких как HEP(Физика Высоких Энергий), биоинформатика, астрофизика и других.

Благодарности

Работа команды лаборатории НИЦ "Курчатовский институт" осуществляется при финансовой поддержке Министерства Образования и Науки РФ в рамках контракта № 14.Z50.31.0024.

Благодарю за внимание!

Вопросы?

novikov@wdcb.ru