НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "КУРЧАТОВСКИЙ ИНСТИТУТ"

CERN

Р.Ю. Машинистов

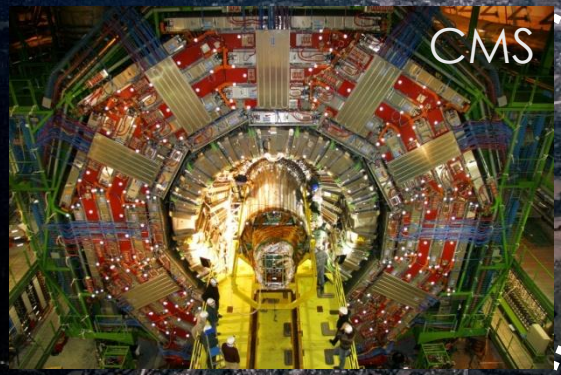
Архитектура PanDA WMS и установка PanDA в КИ

Вступление

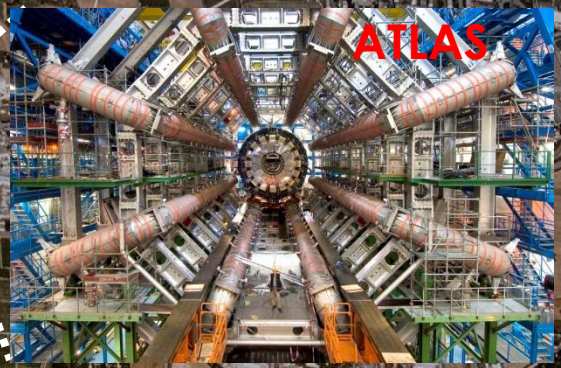
- Системы распределения и управления данными (Workload and Data Management System (WDMS))
- PanDA (акроним для Production and Distributed Analysis – система управления распределенной обработкой и анализом данных)
- Одна из самых успешных систем, разработанных в области физики высоких энергий
  - Проект получил свое начало в 2005 г. Благодаря усилиям групп Брукхейвенской Национальной лаборатории (BNL) и Техасского университета в Арлингтоне (UTA)
  - Система спроектирована для эксперимента ATLAS на ускорителе БАК.

Новая эра фундаментальных исследований

Работа, проводимая экспериментом АТЛАС, это передний край современной науки. Научный прорыв 2012 года— открытие бозона Хиггса, был триумфом научного мегапроекта Большой адронный коллайдер (БАК), выполняющегося в международной Лаборатории ЦЕРН в Женеве, Швейцария.



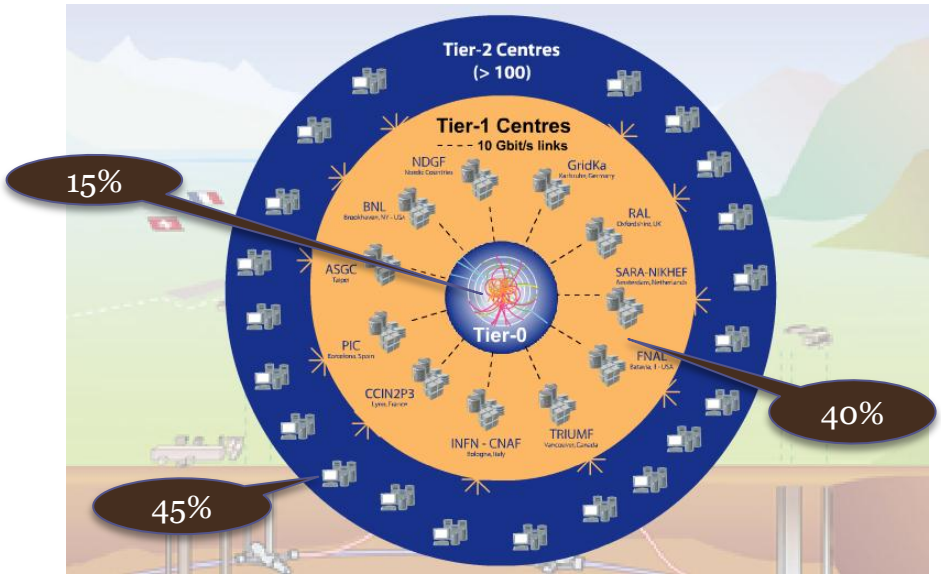
Энергия столкновения пучков протонов достигнет 14 ТэВ в системе центра масс



Вычислительная инфраструктура WLCG

- Чтобы решить беспрецедентную проблему обработки мультипетабайтных данных, эксперимент ATLAS использует вычислительную инфраструктуру грид, развернутую в рамках проекта Worldwide LHC Computing Grid (WLCG)
- Вычислительные средства WLCG в ATLAS организованы по уровням (Tiers). ЦЕРН - источник всех первичных данных, называемых Уровнем 0. Существует 10 центров уровня 1. Каждый центр уровня 1 иерархически поддерживает 5-20 центров уровня 2. PanDA может работать со во всеми центрами ATLAS уровней 1 и 2.
- Обработка и анализ Петебайт данных

- Эксперимент ATLAS располагает объемом данных ~160 PB, распределенных по O(100) компьютерным центрам по всему миру и анализируемых O(1000) физиками
- Детектор ATLAS генерирует порядка 1PB сырых данных в секунду

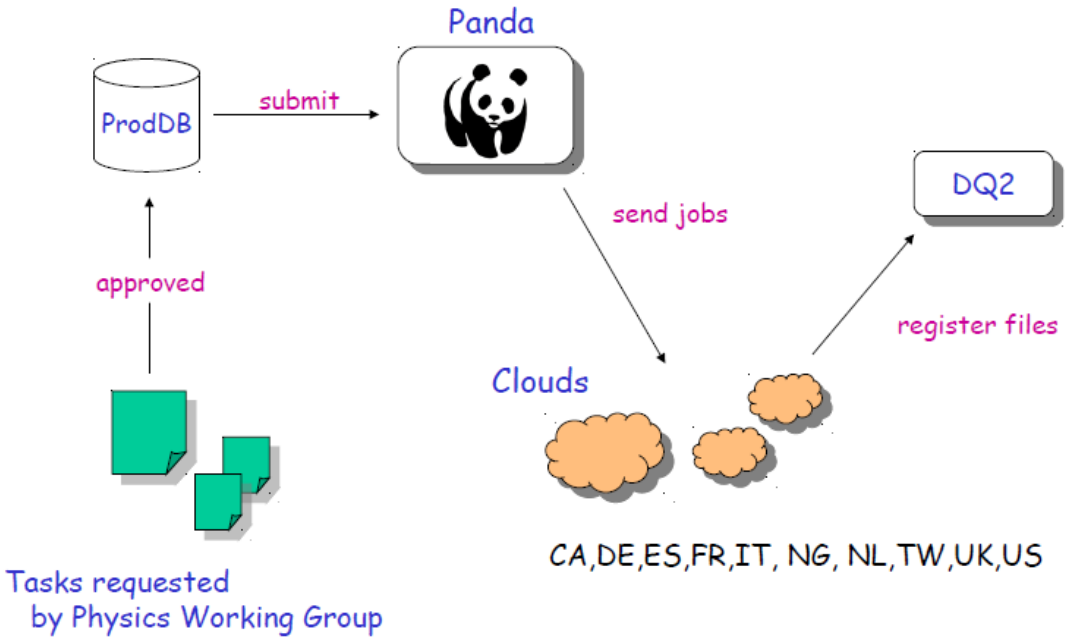


Основные характеристики

- Основная задача системы – это предоставление пользователям простого инструмента для распределенных вычислений.
- Пользователи изолированы от гетерогенности инфраструктуры и промежуточного программного обеспечения
- Управляемое взвешаное разделение ресурсов между тысячами пользователей
- Единый интерфейс для маленьких и больших задач, отдельных пользователей и групп

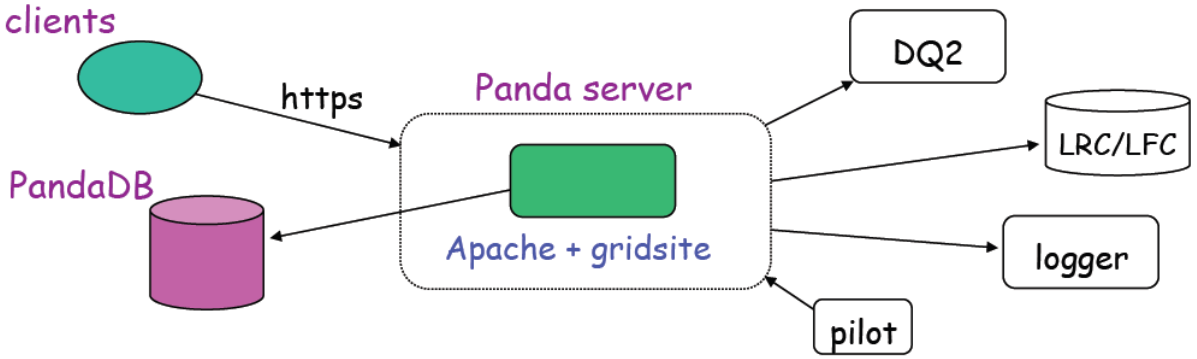
- *С точки зрения пользователей, PanDA предоставляет единое вычислительное устройство, которое используется для обработки всех данных эксперимента, в том числе посредством дата-центров по всему миру.*

Поток заданий



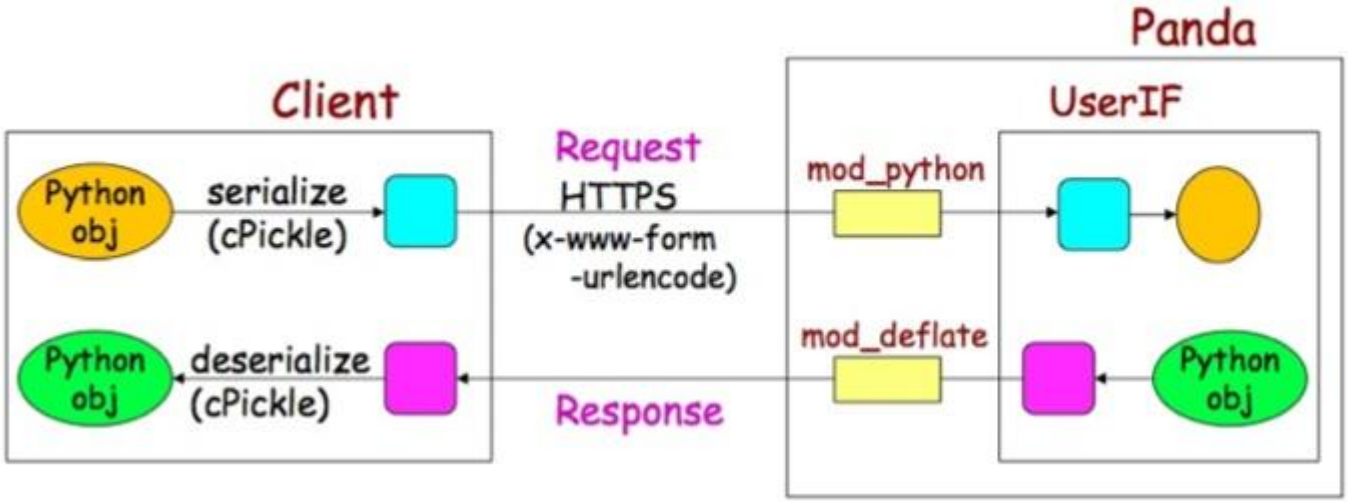
- PanDA поддерживает собственную центральную базу данных, обеспечивающую интегрированное представление всех подконтрольных ресурсов, и центральную очередь всех заданий
- Каждый центр PanDA обеспечивает доступный в гриде Вычислительный элемент (Compute Element (CE)) и Элемент хранения данных (Storage Element (SE)).
- Программный комплекс DQ2 используется системой для управления файлами ATLAS

Основные компоненты PanDA



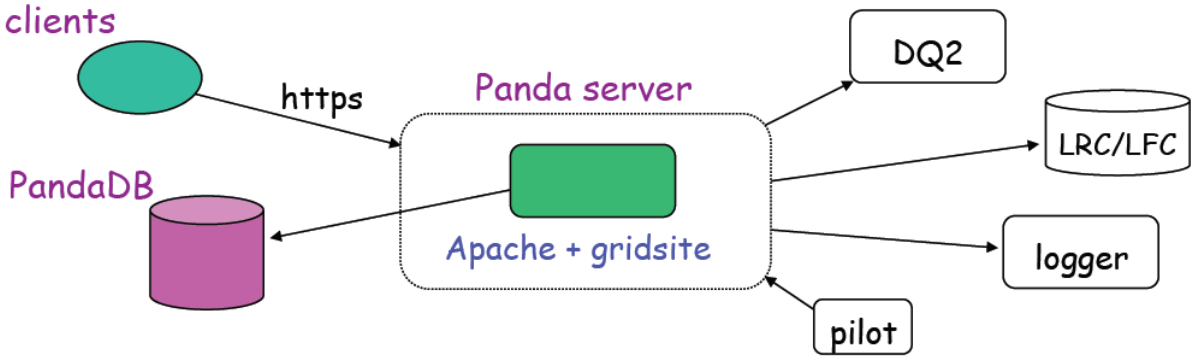
- Простой, основанный на языке Python, пользовательский интерфейс обеспечивает интеграцию с разнообразными средствами запуска заданий
- Пользователи определяют набор заданий (jobs), соответствующие датасеты, входные/выходные файлы.

Взаиможействие клинта с сервером



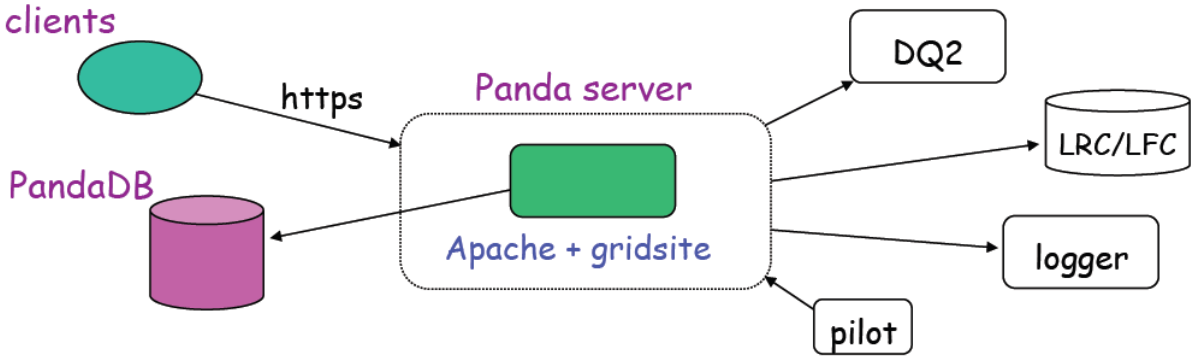
- Коммуникация на основе протокола HTTP/S (curl+grid proxy+python)
- Аутентификация GSI посредством mod_gridsite
- Сервер выполняет инструкции python сразу после получения HTTP запроса и незамедлительно дает ответ.

Основные компоненты PanDA



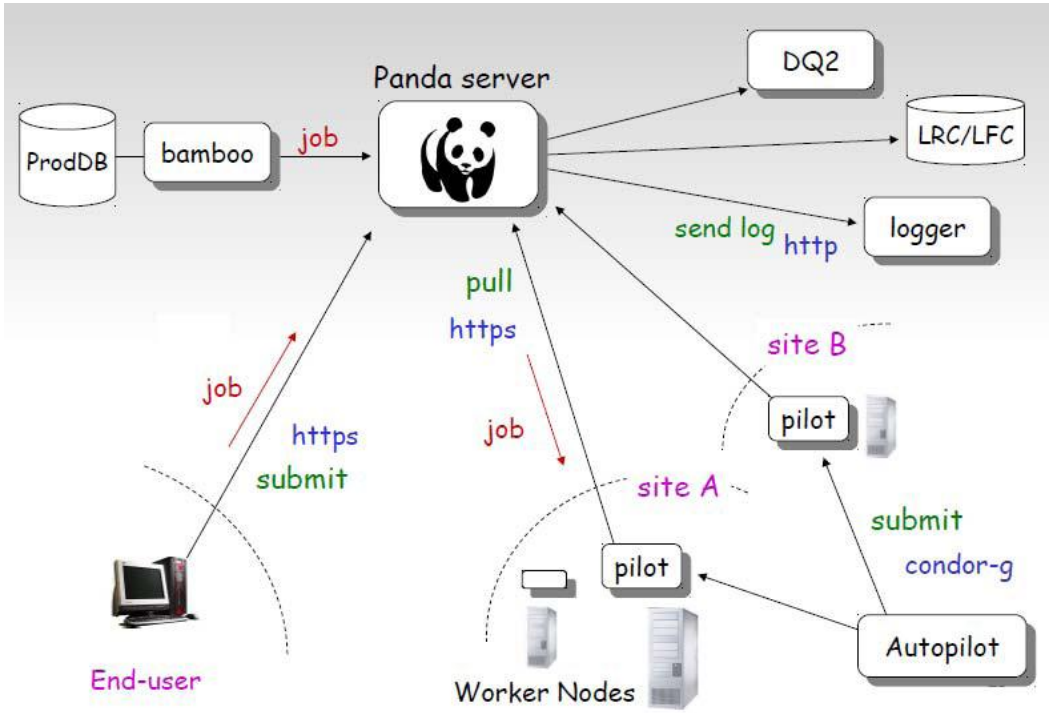
- *База данных PanDA* хранит всестороннюю статическую и динамическую информацию обо всех заданиях в системе.
  - Oracle и MySQL поддерживается как альтернативные бэкэнды базы данных
  - Для пользователей и для самой PanDA база данных заданий представляется по существу как единая многопараметрическая очередь к глобальному ресурсу обработки.

Основные компоненты PanDA



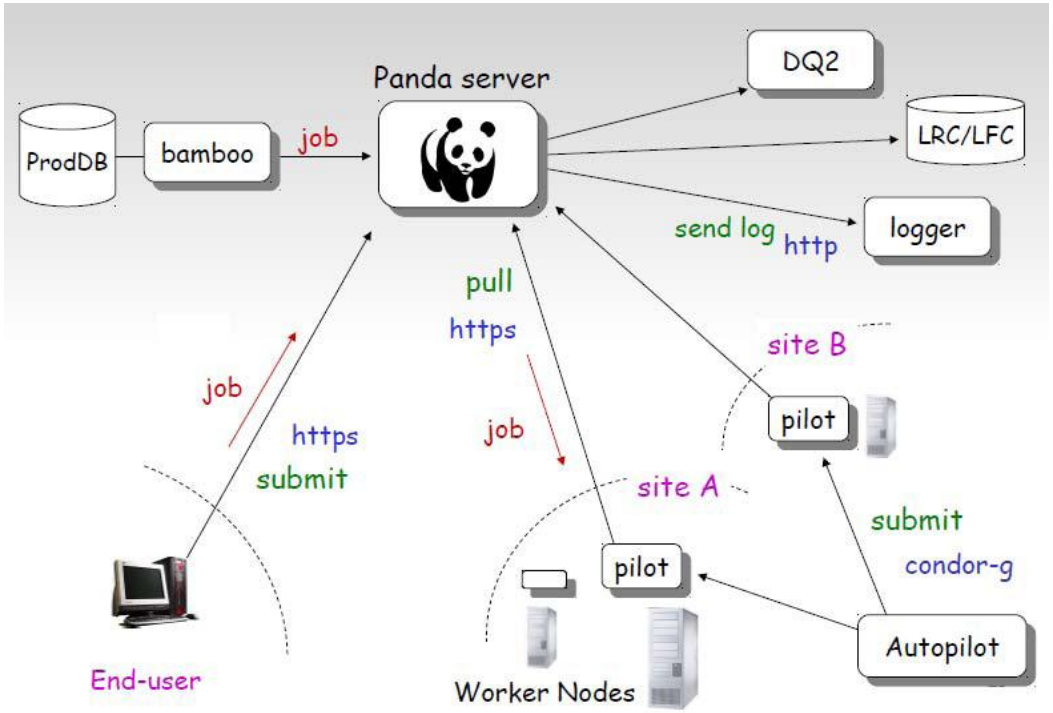
- *Пилоты* используются для сбора информации о вычислительных ресурсах и запуска рабочих задач
  - Рабочие задачи передаются сервером успешно активированным и проверенным пилотам на основе критериев выбора ресурса
  - 'Поздняя привязка' рабочих задач к месту вычислений предотвращает задержки и отказы, и максимизирует гибкость выделения ресурсов на основе динамического состояния обрабатывающих ресурсов и приоритетов задач

Архитектура PanDA



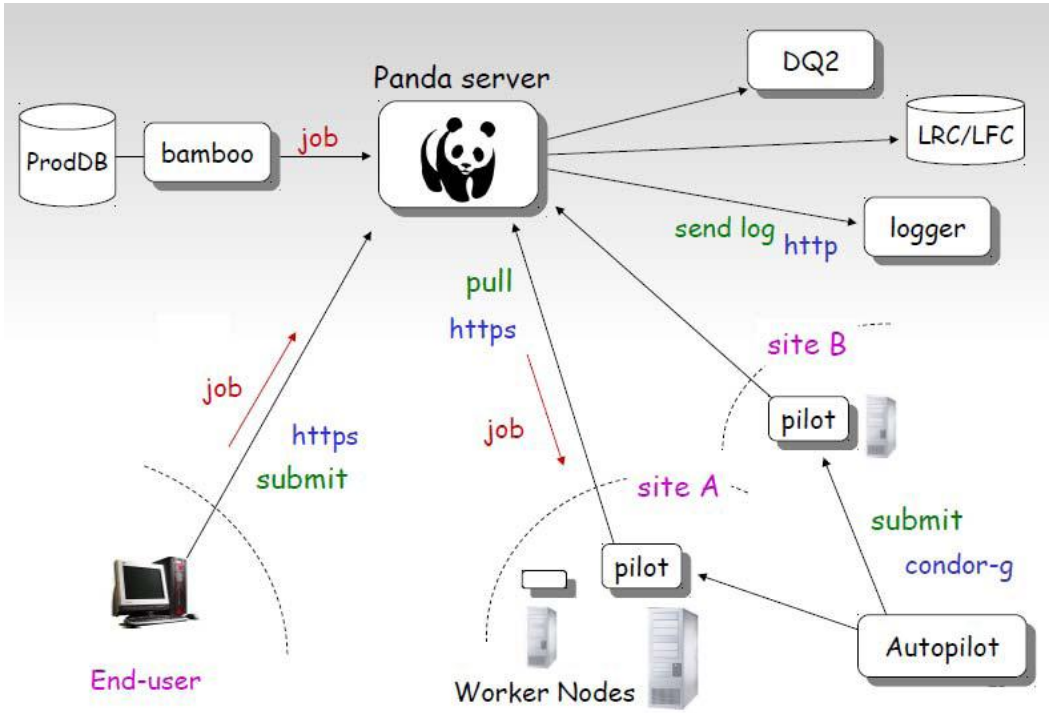
- Задания передаются на PanDA-сервер по защищенному протоколу https с аутентификацией по гридовскому сертификату.
- PanDA-сервер принимает задания и помещает их в глобальную очередь.

Архитектура PanDA



- *PanDA сервер*
  - *Система распределения задач (брокер) PanDA* выполняет выбор подходящего ресурса на основе типа и приоритета задачи, наличия программного обеспечения, входных данных и их местоположения, доступного ЦПУ и ресурсов хранения
  - *Диспетчер PanDA* получает запросы на задания от пилотов и диспетчеризирует задания, используя приоритеты и политику распределения ресурсов

Архитектура PanDA



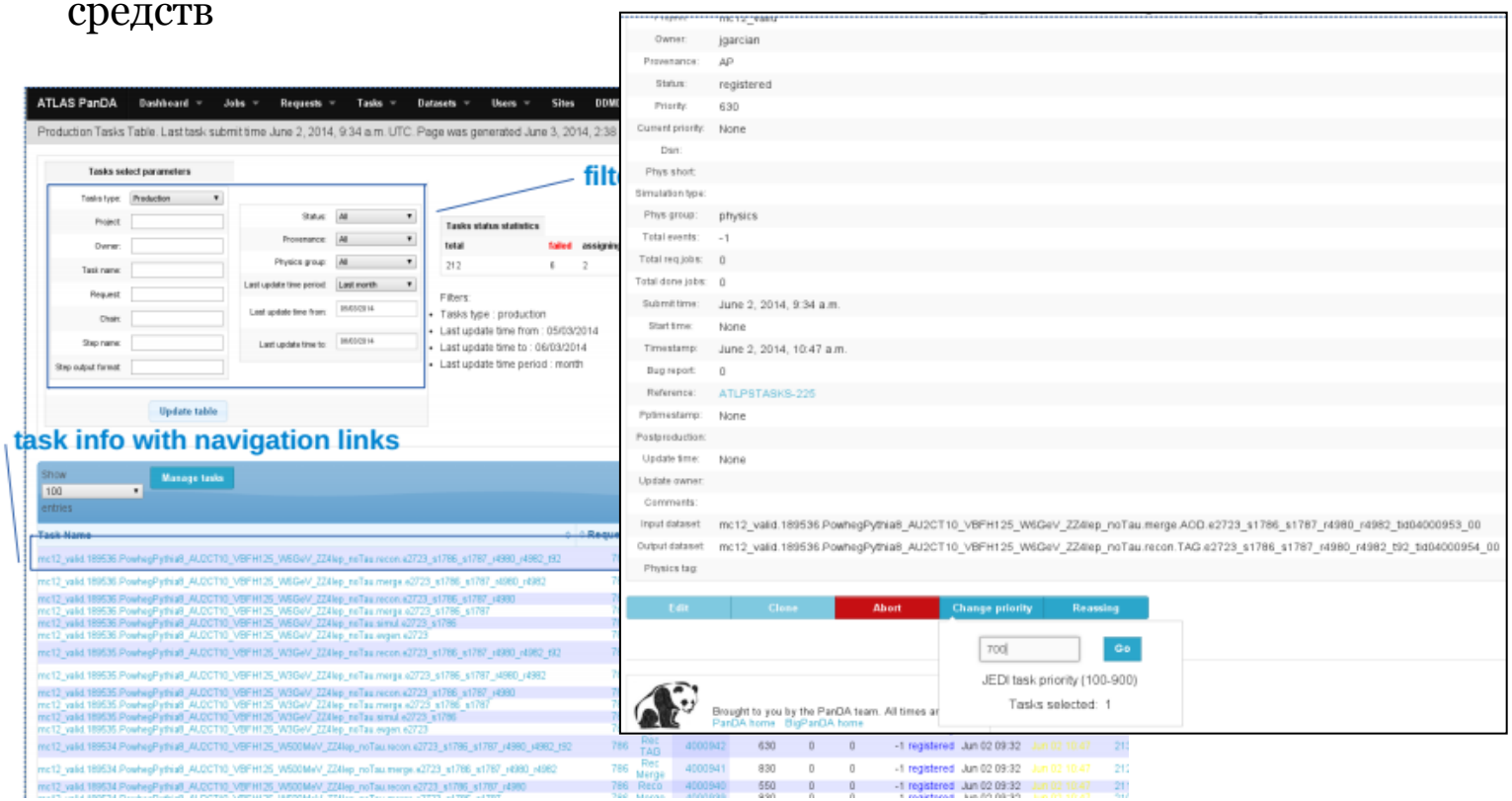
- *Автоматическая фабрика пилотных задач*
  - Независимая подсистема, управляющая поставкой пилотов к рабочим узлам
  - Пилот, запущенный на рабочем узле, связывается с диспетчером и получает доступное задание, приписанное сайту.
  - Важным свойством этой схемы является то, что диспетчеризация пилотов обеспечивает устранение любых задержек в системе планирования
  - Рабочие задачи предоставляются на сайт только после успешного запуска пилота

Как работает pilot

- Отправляет несколько параметров на сервер для получения подходящего задания (HTTP запрос)
  - Скорость CPU
  - Доступная память на WN
  - Список доступных релизов ПО ATLAS
- Плучает задание, находящееся в состоянии активации (HTTP ответ на запрос)
- Заданию на сервесе присваевается новый статус. Активированное –> Выполняется (activated -> running)
- Задание начинает выполняться незамедлительно, т.к. входные данные должны уже быть доступны на сайте
- Pilot каждые 30 мин посылает сигнал «сердцебиения»
- Pilot копирует выходной файл на локальный SE и регистрирует его в каталог - Local Replica Catalog

Мониторинг

- Монитор PanDA обеспечивает всесторонний мониторинг заданий (и задач), как общего, так и индивидуального значения
- Предоставляет подробную информацию о заданиях и сайтах для диагностики их состояния и возможных проблем
- Отображает информацию об использовании, правильности работы и производительности подсистем PanDA и используемых вычислительных средств



PanDA @ NRC-KI

- Для установки PanDA в НИЦ КИ были использованы 4 виртуальные машины:
- Физический ЦПУ Intel(R) Xeon(R) CPU E5450@3.00GHz
- 3x 1 CPU, 1 Gb
  - PanDA сервер
  - PanDA БД
  - PanDa млнитор
- 1x 2 CPU, 2 Gb
  - Фабрика пилотов
- В качестве рабочего узла выступает узел СК

Высокопроизводительный вычислительный кластер HPC2

- 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.

Высокопроизводительный вычислительный кластер HPC2

- Кластер состоит из 1280 счётных двухпроцессорных узлов, объединенных высокопроизводительной сетью передачи данных и сообщений InfiniBand DDR, имеет суммарную оперативную память 20,5 Тбайт и систему хранения данных на 144 Тбайт.
- На счётных узлах кластера установлена операционная система Linux (CentOS). Система хранения данных построена на параллельной файловой системе Lustre 2.0. Для управления распределением ресурсов и выполнением счетных заданий используется менеджер ресурсов SLURM.

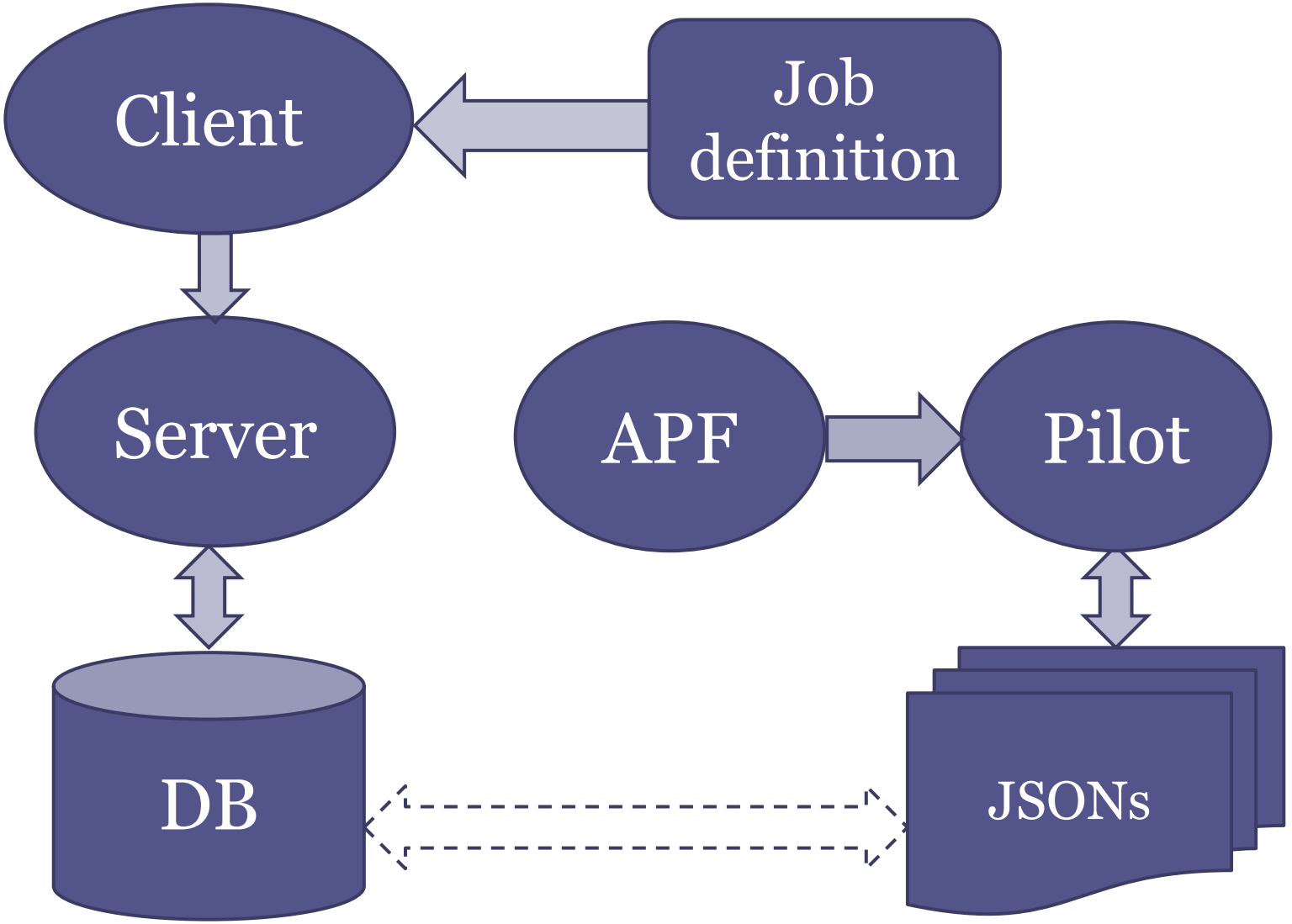
Технические характеристики счётных узлов кластера HPC2

Технические характеристики счётных узлов и кластера HPC2	
Счётные узлы на процессорах Intel Xeon E5450 (3,00 ГГц, 4 ядра)	
Количество процессоров на узел	2
Количество ядер на узел	8
Оперативная память на узел (Гбайт)	16
Оперативная память на ядро (Гбайт)	2
Локальная дисковая память на узел (Гбайт)	120
Общее количество узлов	1280
Общее количество процессоров	2560
Общее количество ядер	10240
Общая пиковая производительность (TFLOPS)	122,9

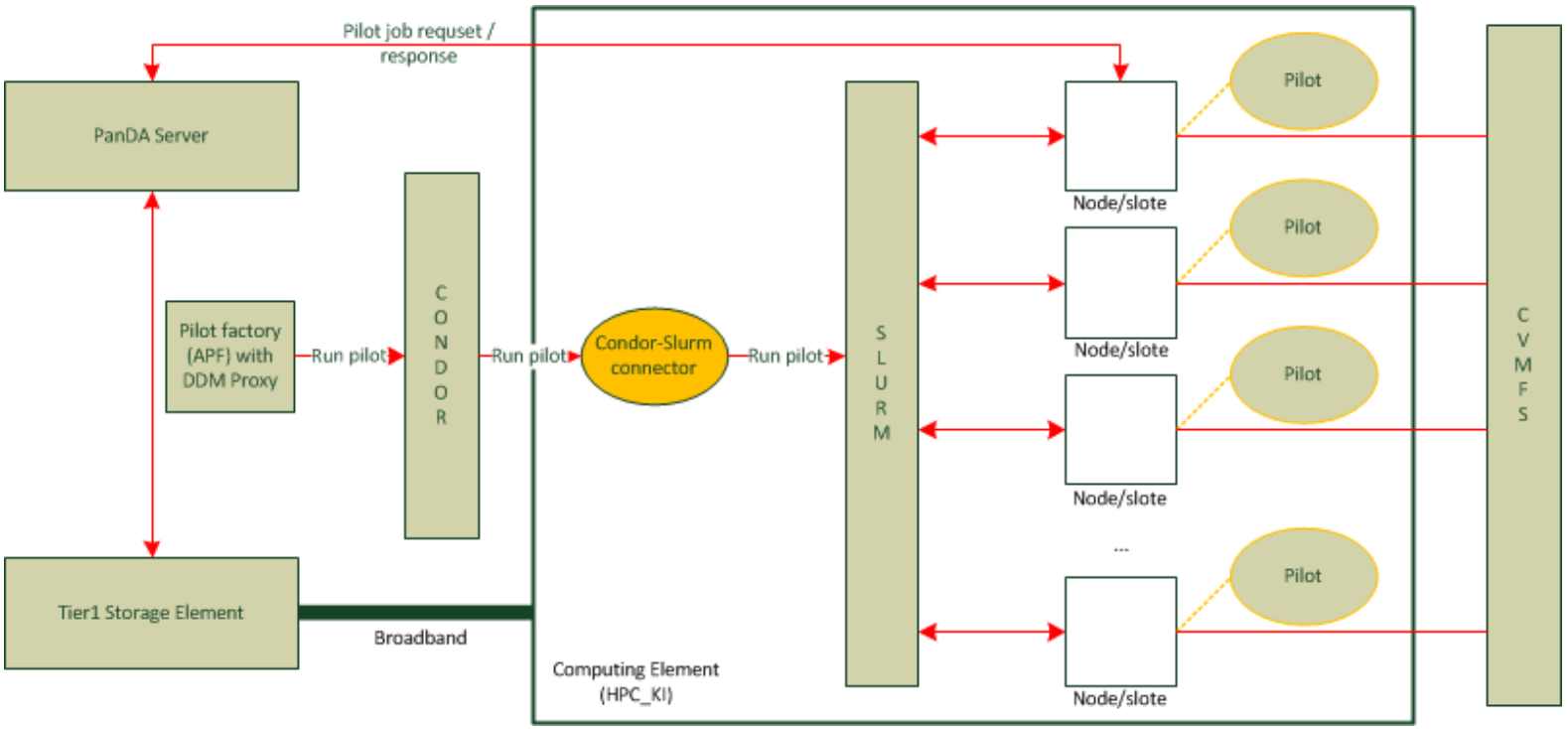
Установка компонентов

- Установлены основные компоненты
  - Server, monitor, DB, APF
  - Установлен клиент DQ2 – системы управления датасетами
  - Установлены зависимые пакеты, настройка среды
  - Выполнена настройка компонентов PanDA
- Определена тестовая PanDA-очередь, соответствующая СК
  - Server: таблица sysconfig
  - Pilot: файл JSON
- Произведено первичное тестирование системы

Тестирование



Интерфейс Condor-SLURM



- Реализация временного решения – интерфейса Condor-SLURM

Тестирование

- Клиент посылает тестовое задание
- Сервер принимает задание и записывает в БД
- APF генерирует пилоты
- Пилот обращается к серверу и получает задание
- Задание выполняется на узле СК
- Каждый шаг отображается на мониторе

Сертификаты

- Пользовательский сертификат д.б. Зарегистрирован в BO
- Сертификат APF д.б. Зарегистрирован в BO и иметь роль – Pilot

Заключение

- Следующим шагом будет выполнена отладка работы сайта на реальной задаче ATLAS с клиентом pathena.
  - Pathena будет использоваться для формирования задач Monte-Carlo моделирования для эксперимента ATLAS
  - Будет выполнена регистрация нашей PanDA-очереди в центральном информационном сервисе. Это позволит получать задания, переданные на центральный сервер ATLAS в ЦЕРНе

Заключение

- Отдельной задачей является изучение механизма формирования задания через простой python-интерфейс
  - Простой интерфейс будет использован для формирования задач вне эксперимента ATLAS

Информационная система (ИС) (Backup slide)

- Информационная база данных сайта/очереди в масштабе всей системы, записывающая статическую и динамическую информацию
- Эта информация используется системой PanDA, чтобы сконфигурировать и контролировать поведение системы от регионального уровня до уровня отдельной очереди
- ИС обеспечивает доступ к информации через http интерфейс.
- Пилоты запрашивают информации от ИС, чтобы сконфигурировать задачу в соответствии с параметрами очереди, в которую их направит PanDA брокер.

Диаграмма выполнения заданий в PanDA (Backup slide)

- PanDA посылает запрос DDM
- DDM перемещает файлы и посылает уведомление назад PanDA
- PanDA и DDM работают ассинхронно
- Доставляет входные файлы на выбраный сайт
- Задания переходят в состоние `активировано` когда все входные данные скопированы и собраны пилотом