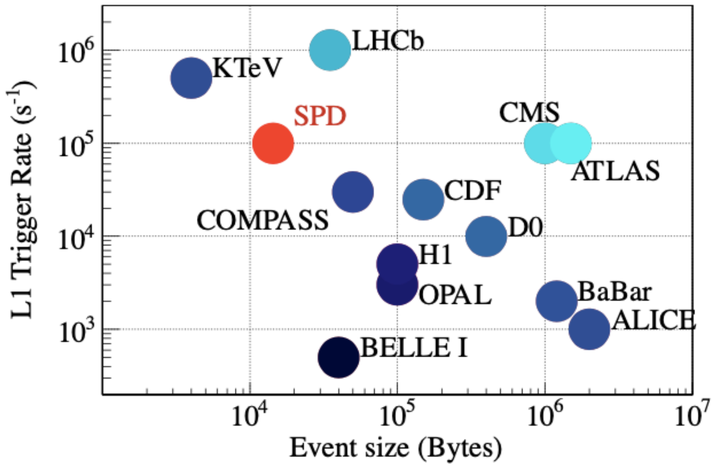
Среда распределённого хранения и обработки данных эксперимента SPD

Артём Петросян, ЛИТ ОИЯИ
20 декабря 2023

Introduction

The expected event rate of the SPD experiment is about 3 MHz (pp collisions at $\sqrt{s}$ = 27 GeV and 10^{32} $cm^{-2}s^{-1}$ design luminosity). This is equivalent to a raw data rate of 20 GB/s or 200 PB/year, assuming a detector duty cycle is 0.3, while the signal-to-background ratio is expected to be on the order of 10^{-5}. Taking into account the bunch-crossing rate of 12.5 MHz, one may conclude that pile-up probability cannot be neglected.

- SPD TDR



The goal of the online filter is at least to decrease the data rate by a factor of 20, so that the annual growth of data, including the simulated samples, stays within 10 PB. Then, data are transferred to the Tier-1 facility, where a full reconstruction takes place and the data is stored permanently. The data analysis and Monte-Carlo simulation will likely run at the remote computing centers (Tier-2s). Given the large data volume, a thorough optimization of the event model and performance of the reconstruction and simulation algorithms are necessary.

SPD как источник данных

- Набор данных ~10 петабайт данных каждый год
- Размер события 10-15 килобайт
- Целевое время на обработку одного события — 1 секунда
- Необходимо контролировать размеры файлов — слишком маленькие создадут избыточную нагрузку на каталоги и системы управления нагрузкой (1 файл = 1 запись в базе и 1 задача), слишком большие сложно передавать, хранить и обрабатывать (оптимум — 6-8 часов)
- Итого: около 60000 ЦПУ для обработки и прирост данных на 10 ПБ в год
- ОИЯИ обеспечит 25-30% необходимых эксперименту ресурсов

Что это за данные

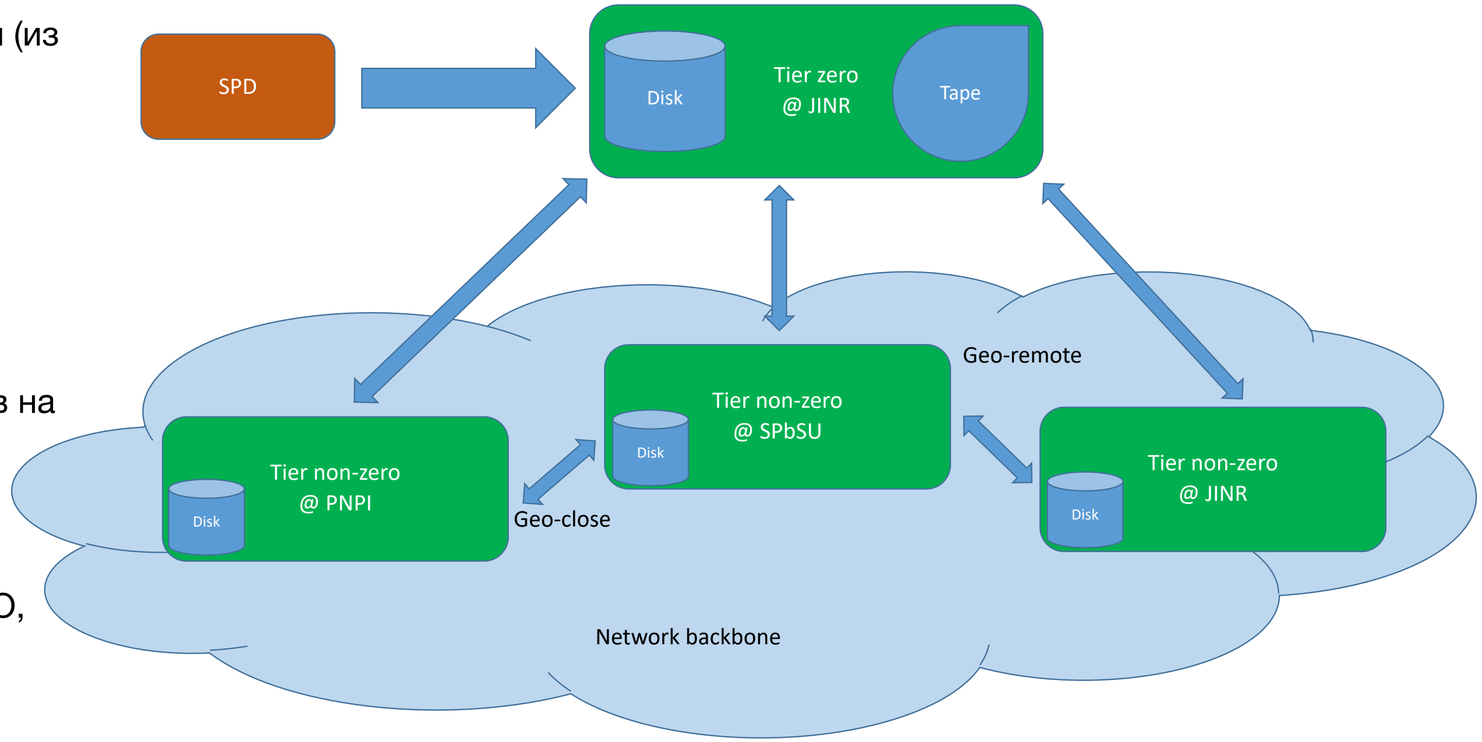
- Результаты кампаний массового моделирования
- Данные с детектора
- Данные различных промежуточных форматов по пути от “сырых” до готовых к анализу физическими группами
- Модели для нейронных сетей
- Логи задач и логи работы промежуточного программного обеспечения

Ценность данных

- Данные неоднородны, и, в зависимости от типа, должны храниться различное время
  - Ценные: полученные в результате работы детектора — хранить “вечно”
    - Их практически невозможно воспроизвести в случае утраты
  - Являются источником физических результатов — как долго сохранять решает коллаборация в каждом конкретном случае
    - Могут быть получены заново, но это достаточно долго и дорого
  - Временные — нужные для проведения какого-то этапа вычислений, после завершения которого могут быть удалены
    - Основная проблема при работе с ними это вовремя и быстро их (и только их) удалять

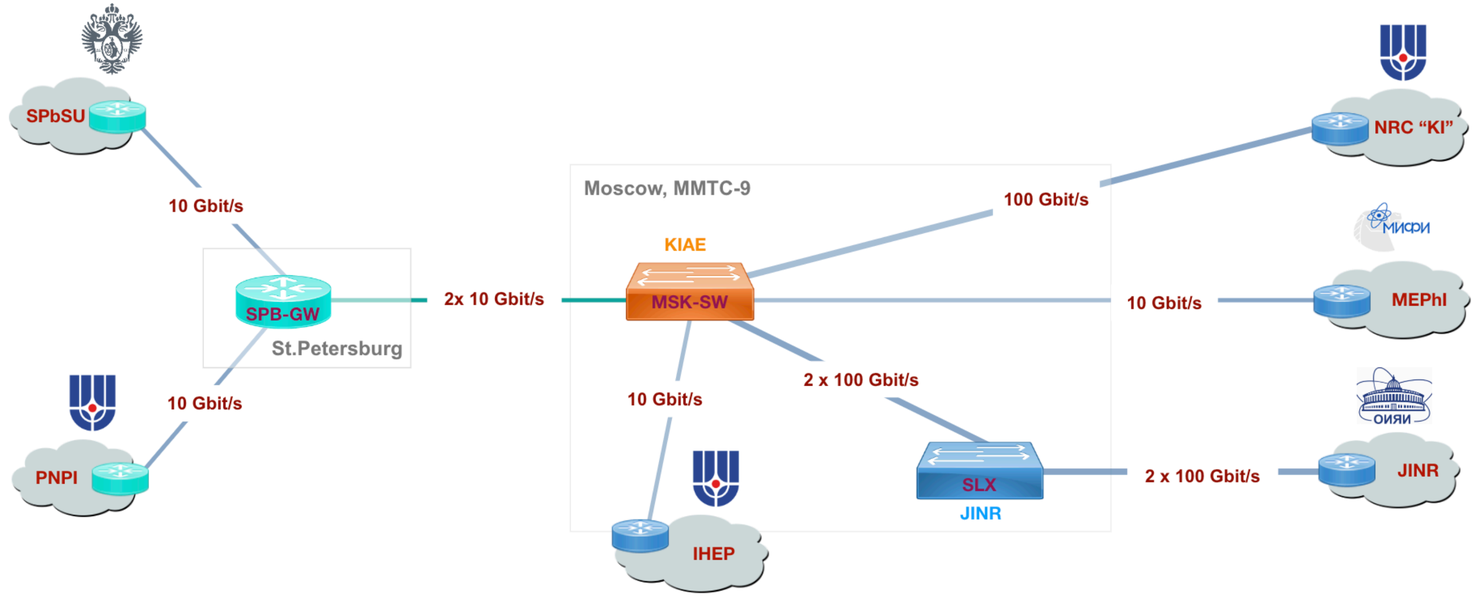
Вычислительное облако SPD

- Сформулированы требования к центрам, желающим участвовать в обработке данных эксперимента
  - >10 Гб/сек пропускная способность канала связи (из технического проекта)
  - >500 ТБ объем системы хранения (пока не в техпроекте, но ведется дискуссия о добавлении)
- По возможности максимально использовать уже существующее ПО
  - Опыт участия в обработке данных экспериментов на Большом адронном коллайдере
- Оптимизация усилий по управлению и эксплуатации
  - Не использовать самодельные, редкие сетапы ПО, отличающиеся от сайта к сайту
  - Предоставить разумные рекомендации по взаимодействию физических ресурсов с централизованными службами управления данными



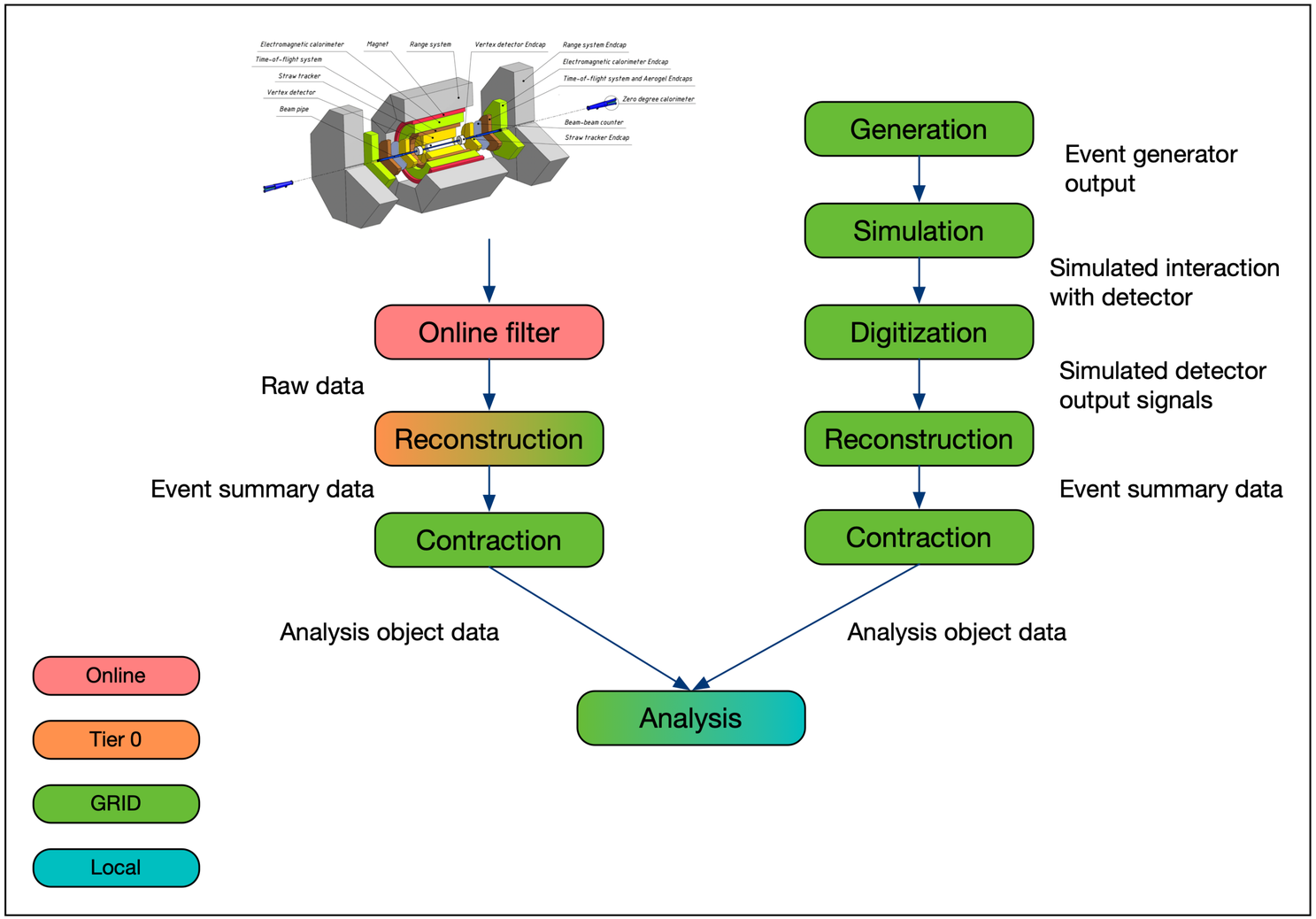
Наиболее очевидные кандидаты на участие в хранении и обработке данных эксперимента

- Участники коллаборации, уже предоставляющие вычислительные ресурсы: СПбГУ, ПИЯФ, НИИЯП БГУ
- Работаем над тем, чтоб расширить список участников
- Подключение центров без опыта участия в WLCG это долгая и трудоёмкая работа



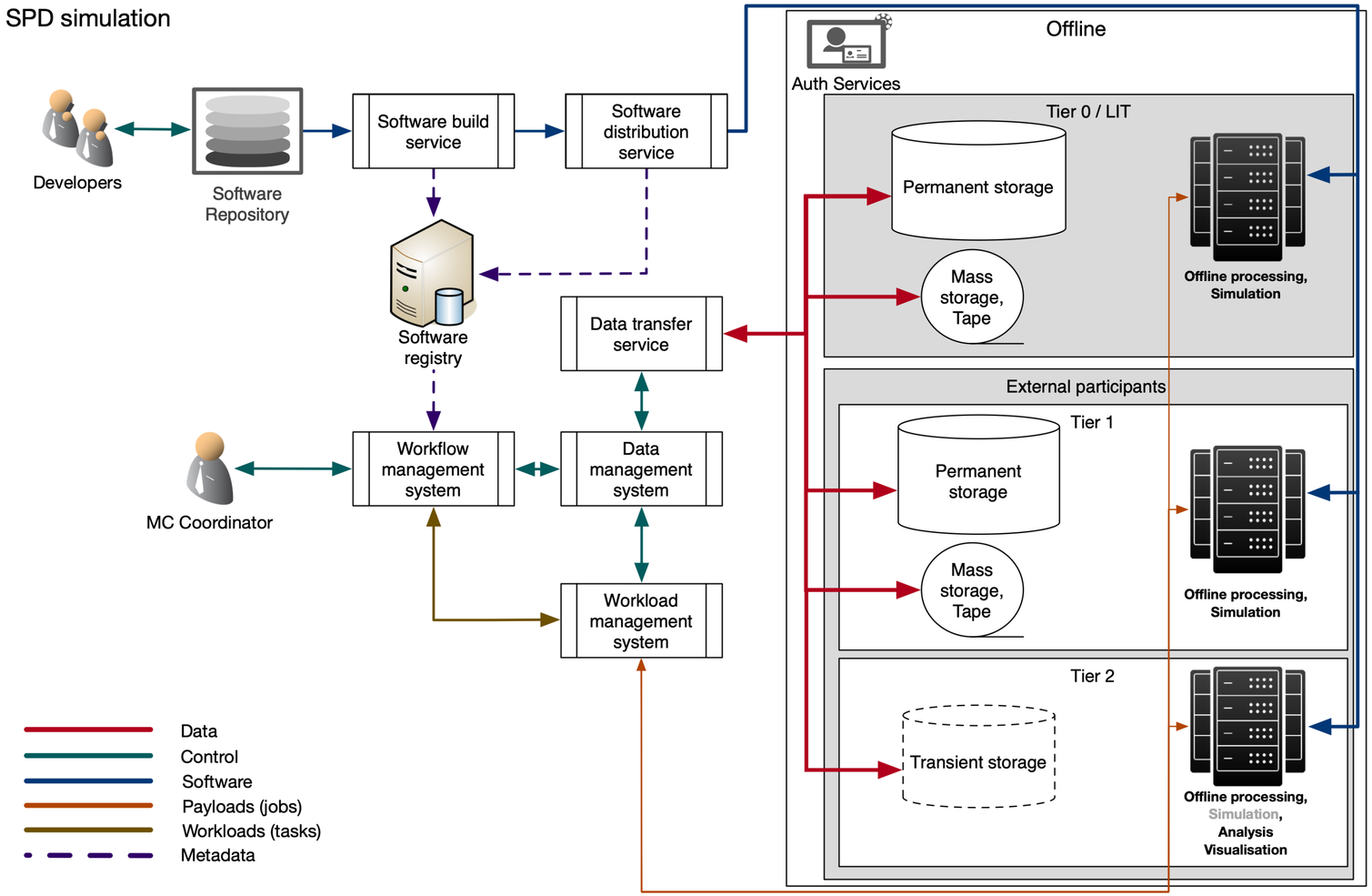
Распределение процессов обработки данных по вычислительным ресурсам

- Выполнение задач реконструкции сопровождаются интенсивными операциями ввода-вывода и будут выполняться преимущественно в ЛИТ ОИЯИ на сайте уровня Tier 0
- Использование сайта уровня Tier 0 диктуется большими объемами исходных данных, получаемых с детектора — они должны быть максимально “уменьшены” прежде чем передавать их для анализа на сайты внешних участников
- Процессы, менее требовательные к ресурсам, например Монте Карло моделирование, могут проводиться на удаленных вычислительных центрах
- Пользовательский анализ планируется проводить на любом подходящем пользователю вычислительном ресурсе



Сервисы распределенной среды хранения и обработки данных эксперимента SPD

- Система аутентификации
- Система авторизации
- Информационная система
- Система управления распределенными данными
- Система управления нагрузкой
- Система управления процессами обработки
- Сервис передачи данных
- Сервис кэширования программного обеспечения



Ключевые управляющие сервисы

- Информационная система Computing Resource Information Catalog была установлена в ЛИТ с некоторыми доработками в 2020 году

- Production and Distributed Analysis System используется в ЛИТ с 2015 года, сначала для эксперимента COMPASS, версия для SPD была установлена в 2020-м, обновлена в 2023-м, в неё постепенно добавляется логика работы с ПО и данными SPD

- Система управления данными Rucio была установлена в ЛИТ в 2022, сейчас постепенно настраивается для работы с данными SPD

- Сервис передачи данных установлен в ЛИТ в 2023 году

Безопасность

- Работа с подобным метахранилищем и метавычислителем требует большого внимания к вопросам безопасности
  - Каждый пользователь должен получить сертификат стандарта X.509 и соответствующую своим обязанностям роль в системе
  - С реализацией в управляющих сервисах поддержки JSON Web Token появилась возможность полностью отказаться от пользовательских сертификатов и напрямую завязать аутентификацию на JINR SSO
  - Разработка регламентов по работе с ассоциированными членами персонала ОИЯИ позволил регистрировать внешних участников в качестве пользователей цифровых сервисов ОИЯИ
  - Таким образом, все пользователи вычислительной среды регистрируются в базе данных пользователей ОИЯИ, получают учётную запись, и в дальнейшем выполняют все действия в системе только с этой учётной записью

Статус и планы работ 1/4

- Аутентификация — ведутся работы по настройке сервиса выдачи JSON Web Token вместо сертификатов стандарта X.509
- CVMFS
  - Развернут раздел для эксперимента SPD
  - Написана базовая документация по подключению на удаленных центрах, отлажена на примере сайта СПбГУ
  - Планируем разместить в CVMFS клиентов для всех используемых сервисов, а также скрипты настройки среды, чтоб у пользователей была возможность в один шаг прописать все необходимые пути
- Ведётся работа по организации CI/CD релизов прикладного ПО из Gitlab, упакованных в контейнеры
- EOS — настроены разделы под данные обычных пользователей и под данные результатов массовой обработки с более строгими правами доступа
- CTA — ждём, когда можно будет настроить и начать пользоваться

Статус и планы работ 2/4

- Информационная система CRIC
  - Развернута, интегрирована с JINR SSO, заполнены базовые элементы вычислительной среды эксперимента: вычислительные очереди, системы хранения, протоколы
  - Планируем развивать в сотрудничестве с А. Анисёнковым из НГУ
- Система управления данными Rucio
  - Развернута тестовая полнофункциональная версия, в которой можно создавать датасеты, загружать файлы на EOS
  - Зарегистрированы как пользователи, так и сервисы, например panda
  - Необходимо настроить взаимодействие с информационной системой CRIC
  - Планируем в 2024 году развернуть “продакшн” версию, связать с FTS, прописать квоты и ввести в опытную эксплуатацию
  - За развитие системы отвечает А. Конак, студент ТулГУ и сотрудник ЛИТ

Статус и планы работ 3/4

- Система управления нагрузкой PanDA
  - Развернута, интегрирована с информационной системой CRIC
  - Отправляет задачи на вычислительные ресурсы ЛИТ, результаты выгружает на EOS
  - Отлажен запуск задач в контейнерах
  - Настроено взаимодействие с системой управления данными Rucio — регистрируются файлы и датасеты, метаданные
- Сервис передачи данных FTS — развернут, проведена начальная настройка, ведется тестирование и наладка
- Подключены вычислительные ресурсы ПИЯФ, СПбГУ, НИИЯП БГУ
  - Сотрудничаем с А. Кирьяновым, А. Зароченцевым и Д. Ермаком
- Ведется работа по подключению системы хранения НИИЯП БГУ

Статус и планы работ 4/4

- Мониторинг
  - Система управления данными Rucio поставляется с мониторингом, он развернут и работает
  - Сервис передачи файлов FTS поставляется с мониторингом, он развернут и работает
  - Система управления нагрузкой PanDA — поставляется в виде отдельного пакета, планируем развернуть
- Система управления процессами обработки SPD ProdSys
  - На этапе проектирования: описано соглашение о наименовании наборов данных (датасетов), описано модель движения данных в процессе обработки, проводятся обработки тестовых наборов данных
  - Планируем в 2024 году выполнить несколько сеансов массовой обработки данных, общим объёмом до 1 ПБ
  - Студент МИФИ Н. Монаков выбрал эту работу в качестве темы своей магистерской

Заключение

- Строящийся детектор SPD будет генерировать большие потоки данных, предполагающие построение распределенной среды их многоуровневого хранения и многоэтапной обработки
- В рамках решения задач по управлению данными мы создаем распределенное комбинированное хранилище из ленточных и дисковых ресурсов, с контролем реплик, управлением временем жизни данных и их целостностью
- Для обработки мы планируем привлекать все возможные вычислительные ресурсы, и будем их использовать наиболее оптимальным образом, распределяя задачи по самым подходящим центрам
- Для управления процессами обработки данных эксперимента SPD мы строим высокоавтоматизированную систему, которая будет учитывать при работе разнообразные параметры: размеры файлов, их местонахождение, подходящие для каждого этапа обработки процессоры и память, параметры сетевых соединений и тд
- Набор управляющих систем и сервисов развернут, ведутся работы по их настройке для работы друг с другом и выводу на режим опытной эксплуатации с постепенным повышением требований к нагрузкам
- Прорабатываются сценарии процессов обработки, ведется работа по их детализации и реализации

Спасибо за внимание!