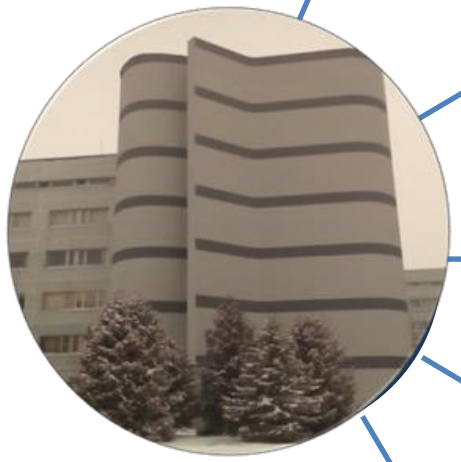

# Гиперконвергентная HPC система, многоуровневая система хранения данных от сверх-горячих до сверх-холодных

**Д.В. Подгайный,**
**Руководитель группы по гетерогенным вычислениям**
**Объединенный институт ядерных исследований**
**Лаборатория информационных технологий**

**2019г.**

# Основные направления деятельности

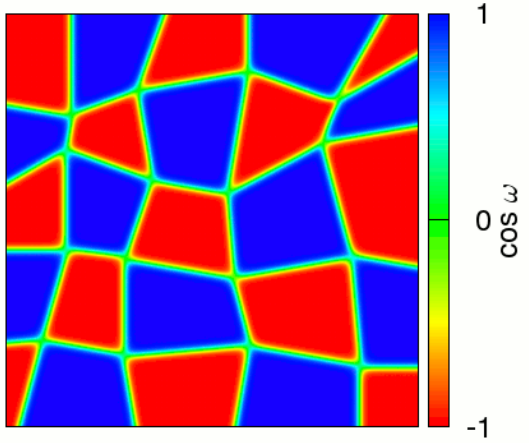
- Обеспечение необходимыми IT – сервисами научной и административной деятельности ОИЯИ
- Формирование компетенции мирового уровня в области IT и вычислительной физики
- Поддержка 24/7 вычислительной и сервисной инфраструктуры Института



- Пользовательская политика
- Вычислительная инфраструктура
- Математическая и программная поддержка
- Корпоративная информационная система
- Обучение и поддержка пользователей

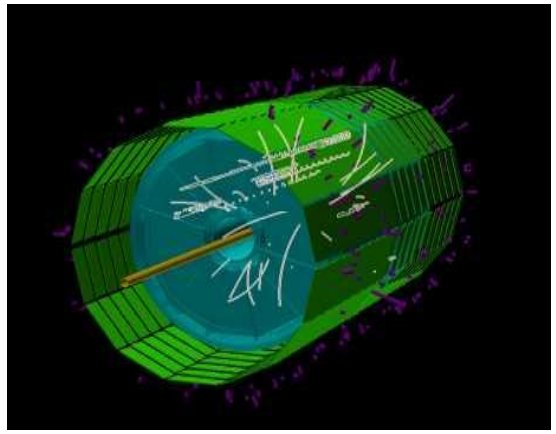
**IT-инфраструктура одна из базовых установок ОИЯИ**

# Основные темы проекта NICA

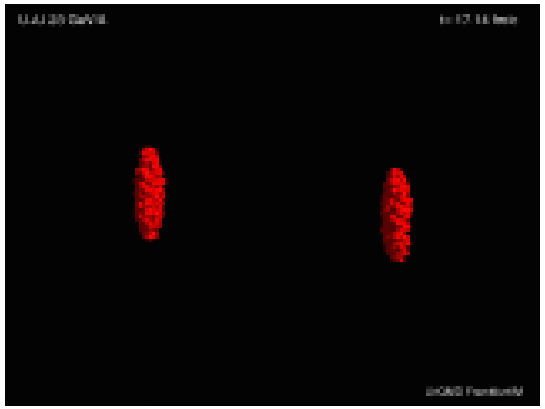



- Теория адронной материи
- Параллельные вычисления КХД
- Модели КХД в вакууме
- Симулирование событий
  - 5 генераторов событий (моделей)
- Реконструкция событий
- Необходимо иметь ~ 100M событий/генератор
  - ~ 4 GB данных на событие
  - Симуляция → 5 ядро-минут (UrQMD).
  - Реконструкция ← 3 ядро-часа (UrQMD)
- Масштабируемая compute|storage архитектура

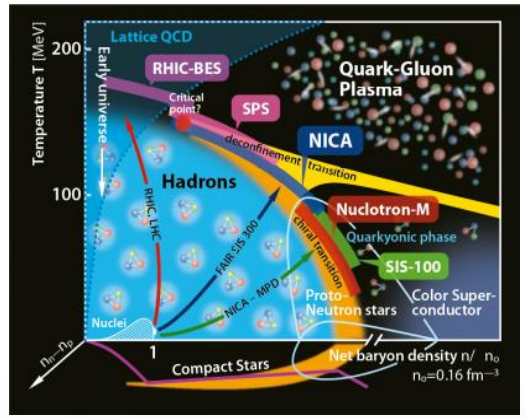
~20 PB “сырых” данных каждый год



Реконструкция событий



Симуляция




Фазовая диаграмма КХД

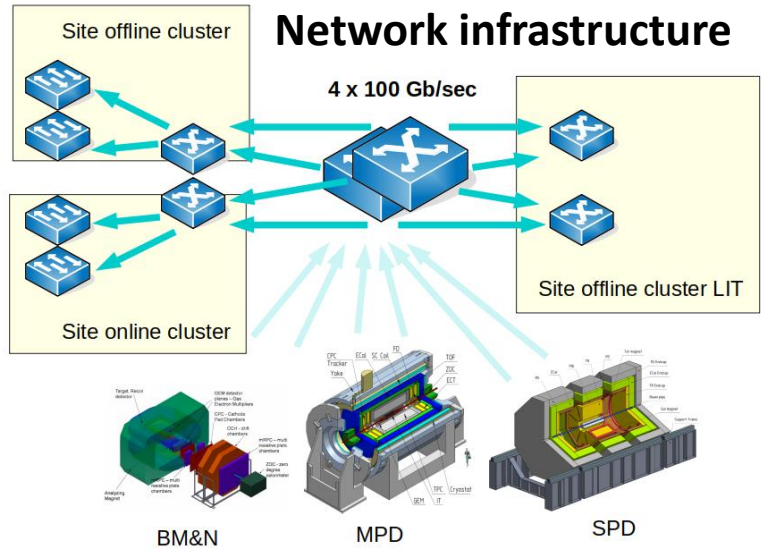
# ИТ инфраструктура ЛИТ для проекта NICA

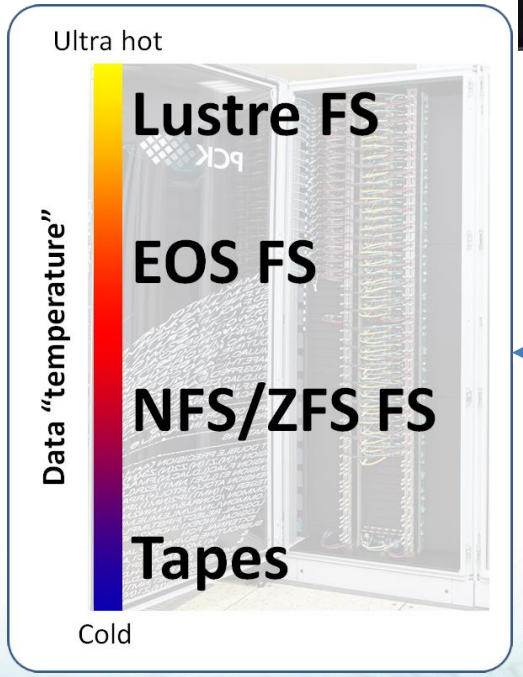
DIRAC THE INTERWARE

**Суперкомпьютер Говорун**

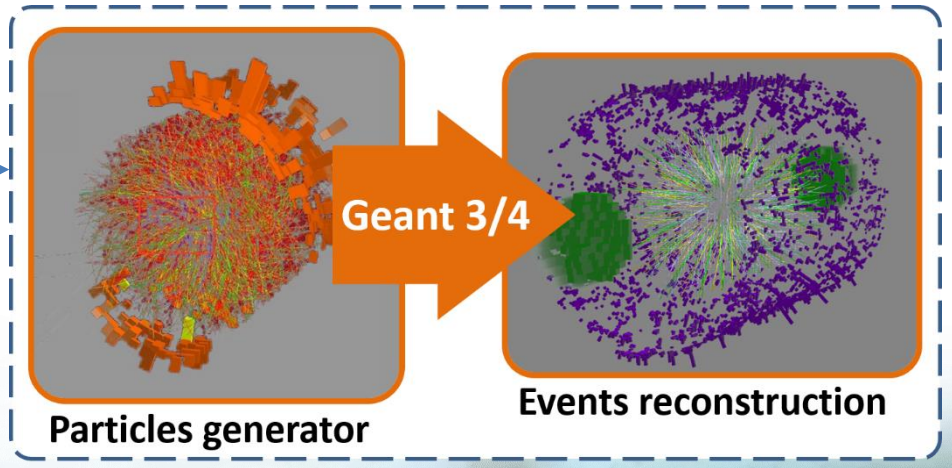
**LHC Tier1/Tier2**

**Network infrastructure**

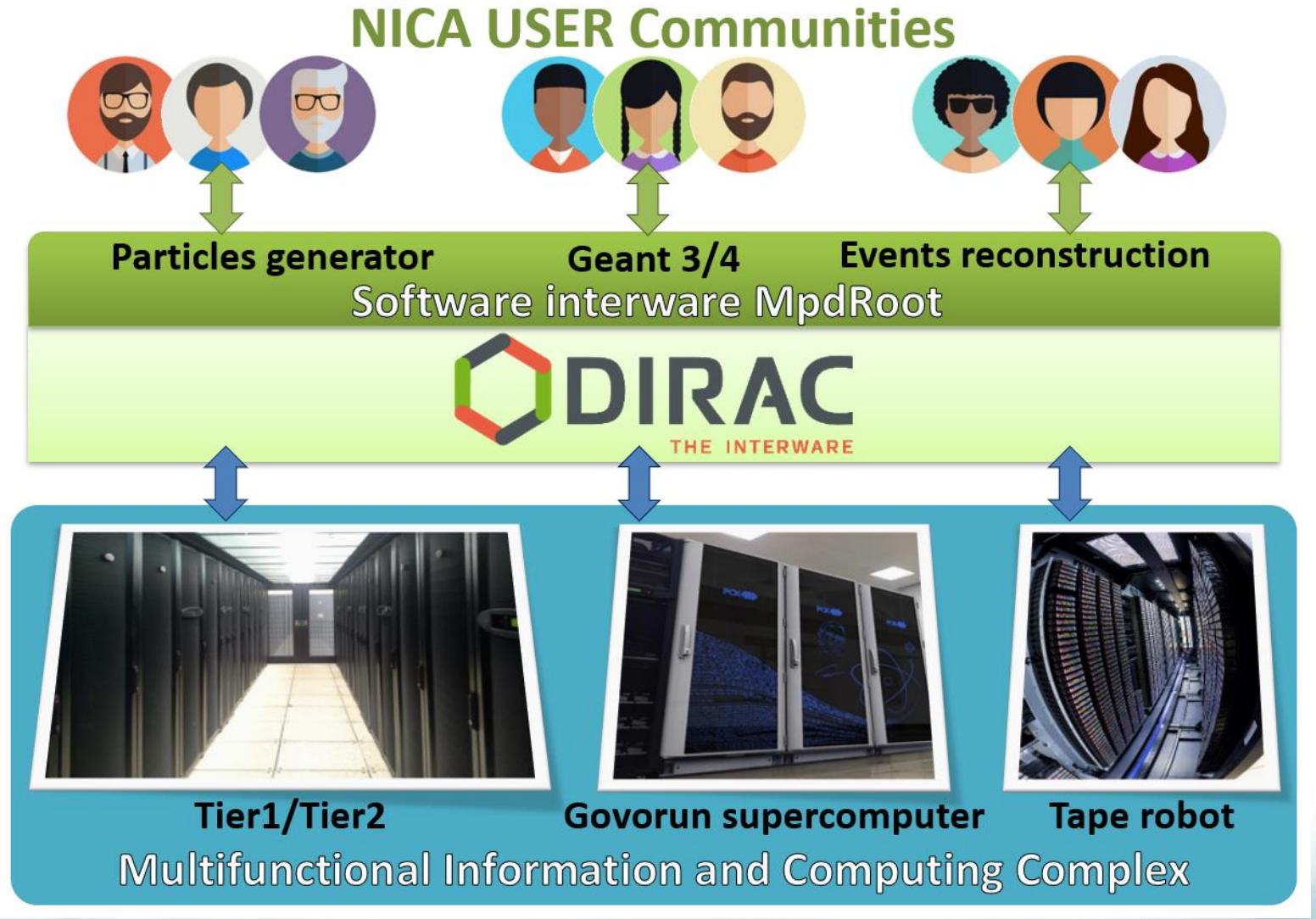



DIRAC THE INTERWARE



# DIRAC. The INTERWARE



# Обновление суперкомпьютера Говорун в ОИЯИ


- **535TFLOPS** пиковой производительности - **#10** в Top50
- Программно-определяема архитектура системы
- **#1** в производитель-ти систем хранения в России **>300GB/s**
- Масштабируемое решение **Storage-on-demand**
- Многоуровневая система хранения для максимальной эфф-ти
- Охлаждение горячей водой (compute, storage, interconnect)
- Наиболее энергоэффективный центр в России (**PUE = 1,027**)

## Компоненты:

### Узлы на Intel® Xeon® Scalable gen 2:

- Пиковая производительность – **463**ТФЛОПС
- Intel® Xeon® Platinum 8268 processors (24 cores)
- Intel® Server Board S2600BP
- Intel® SSD DC S4510 (SATA, M.2),
  2 x Intel® SSD DC P4511 (NVMe, M.2) 2TB
- RAM – 192 GB DDR4 2933 ГГц
- Intel® Omni-Path 100 Gbit/s
- 48-port Intel® Omni-Path Edge Switch 100 Series со 100% жидкостным охлаждением

### Hyperconverged:

- 18 узлов с 12-ю NVMe SSD слотами
- 4 узла Optane с 3,4TB IMDT памяти
- 12 узлов OSS с NVMe SSD – **256TB**
- 2 узла MDS с 12-ю **Optane 375GB**
- ПФС Lustre как основная опция
- Storage-on-Demand с **RSC BasIS** на узлах кластера

### Стек ПО “RSC БазИС”

### Intel® Xeon Phi™ nodes:

- Пиковая производительность – **72,576** ТФЛОПС
- Intel® Xeon Phi™ 7190 CPUs (72 cores)
- Intel® Server Board S7200AP
- Intel® SSD DC S3520 (SATA, M.2)
- RAM – 96 GB DDR4 2400 ГГц
- Intel® Omni-Path 100 Гбит/с
- 48-port Intel® Omni-Path Edge Switch 100 Series 100% liquid cooling

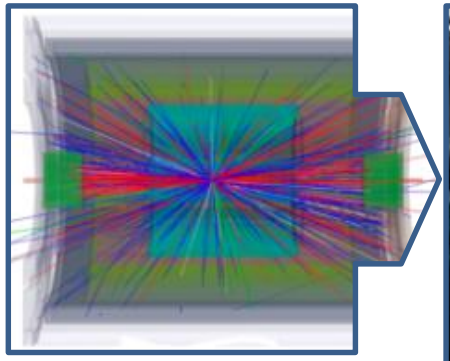
# Различные типы узлов в Hyper-converged системе

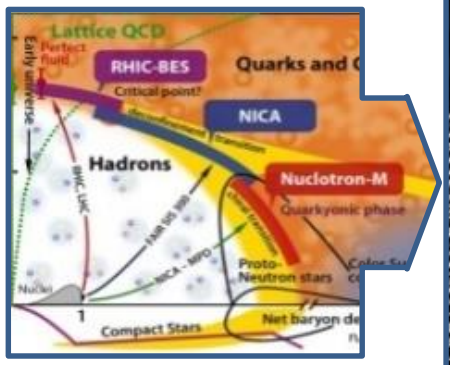
Узлы для создания различных быстрых ПФС (Lustre, BeeGFS, DAOS и др.)

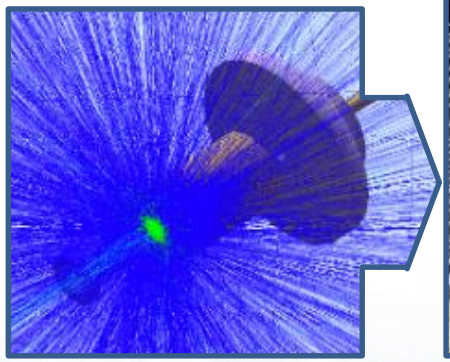
Узлы с большой памятью

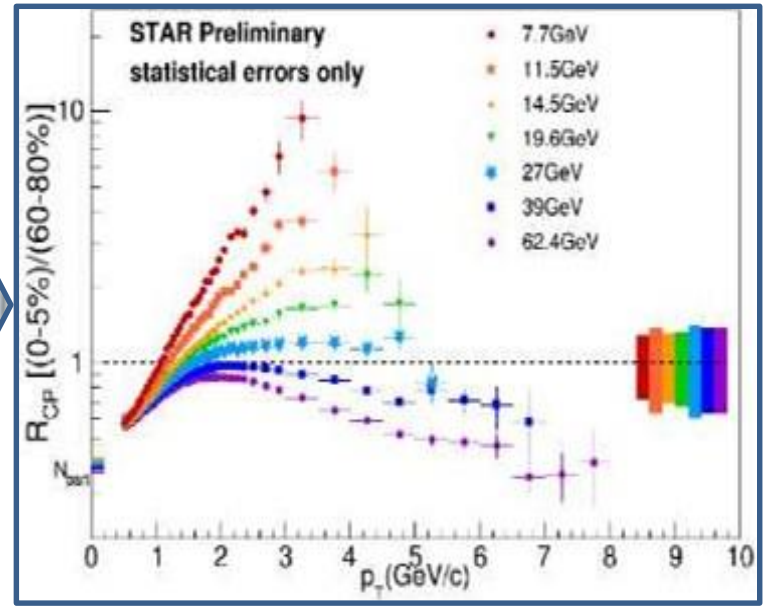
Стандартные вычислительные узлы

# Суперкомпьютер Говорун для проекта NICA

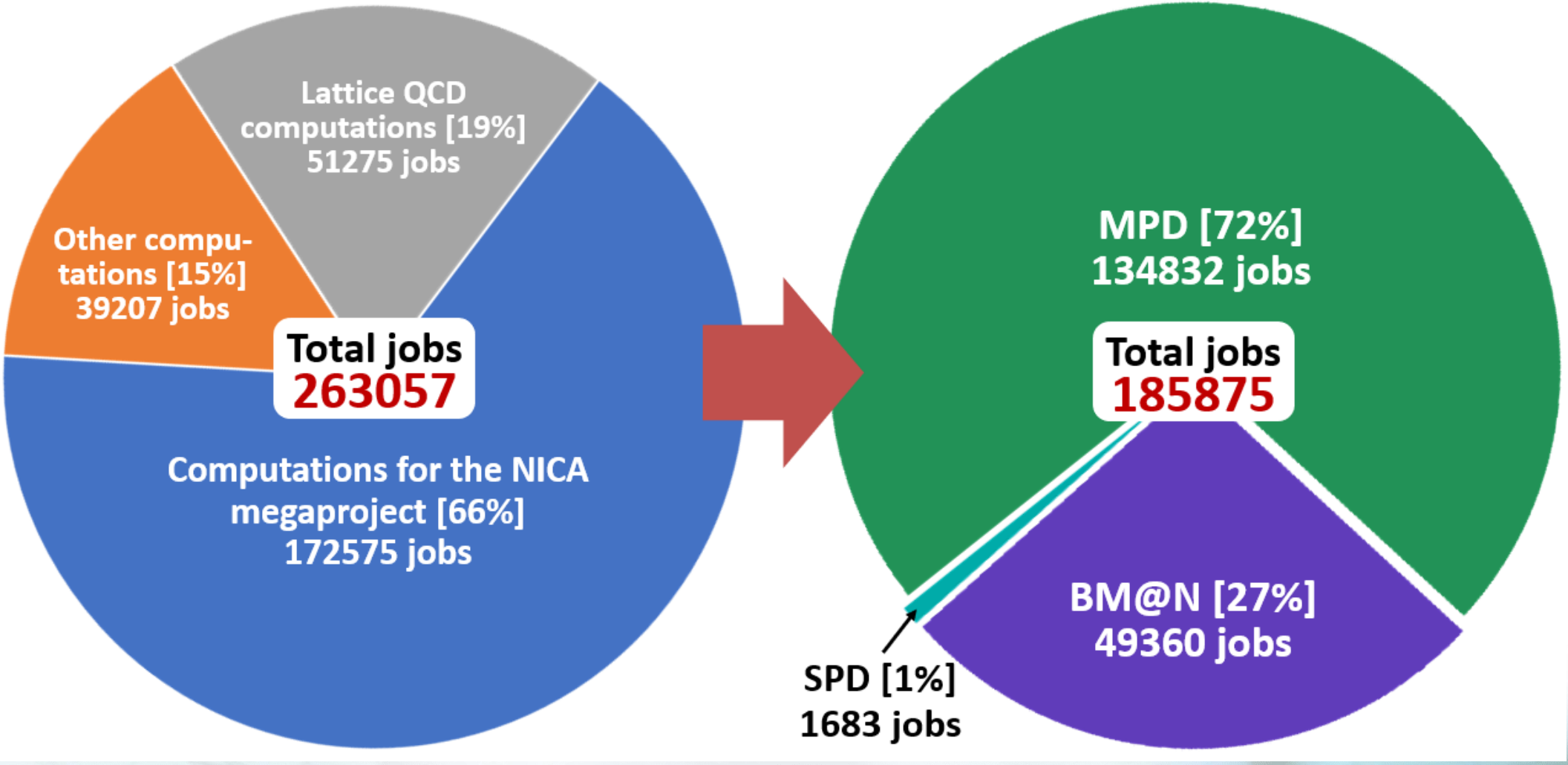




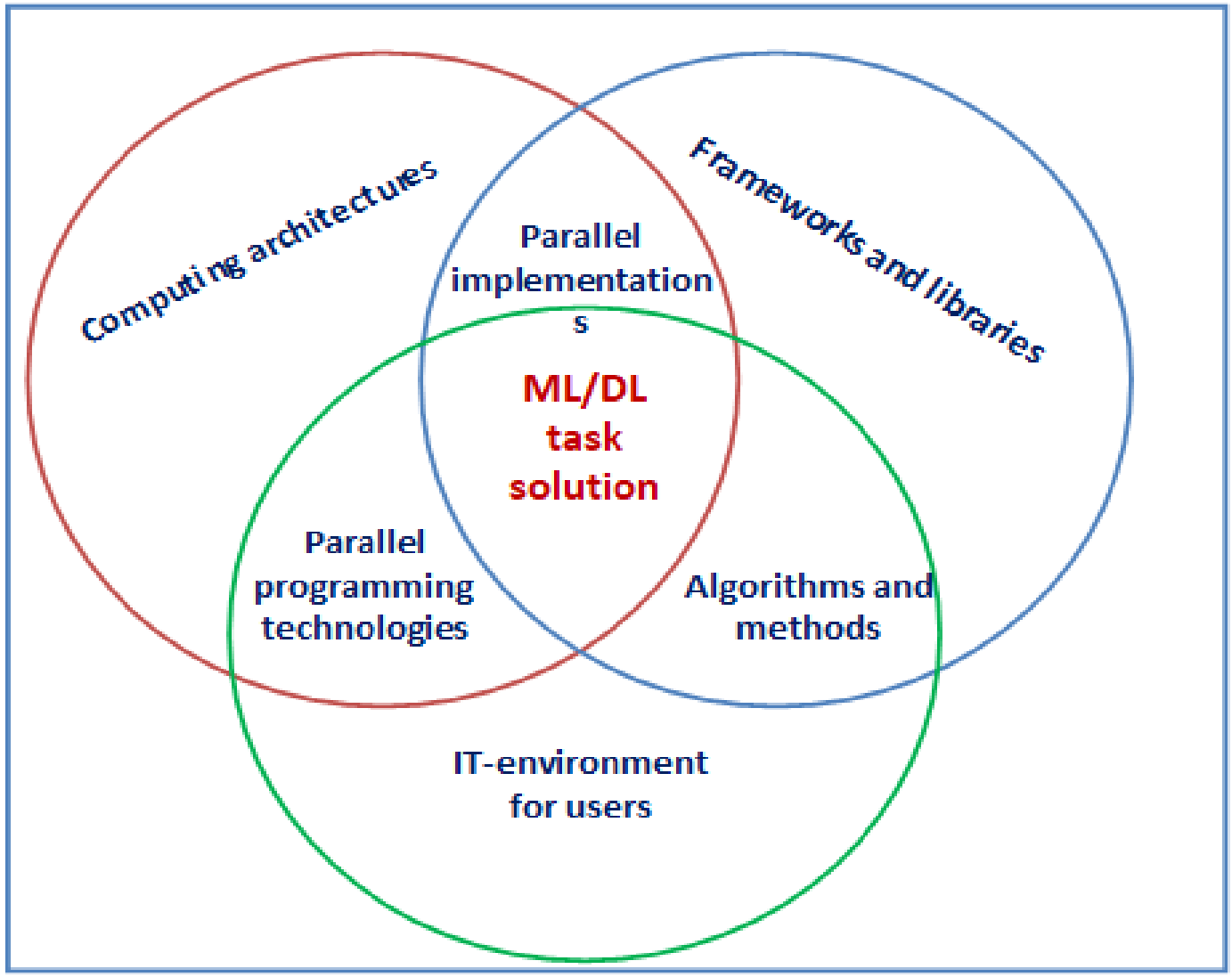





# Статистика использования компонент суперкомпьютера в проекте NICA

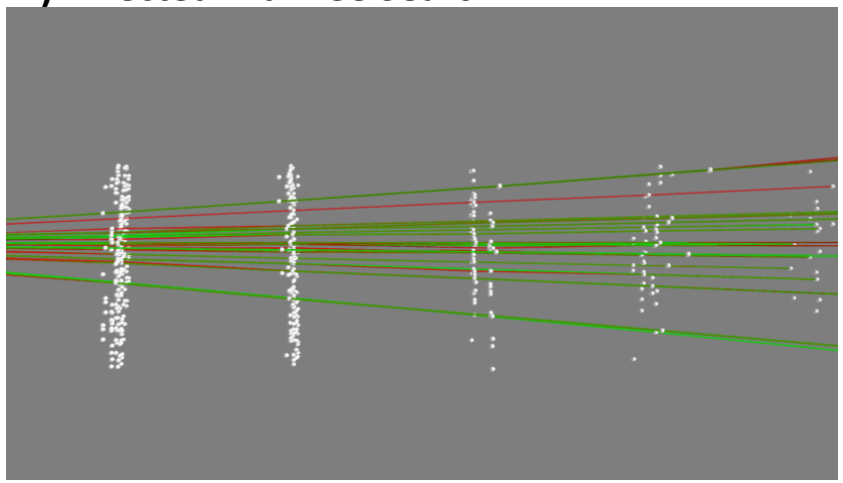


# Реализация методов нейронных сетей для NICA



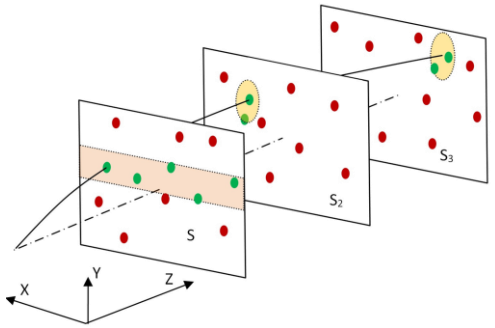
# Реконструкция путей частиц методами ML/DL

## 1) Directed K-d Tree Search



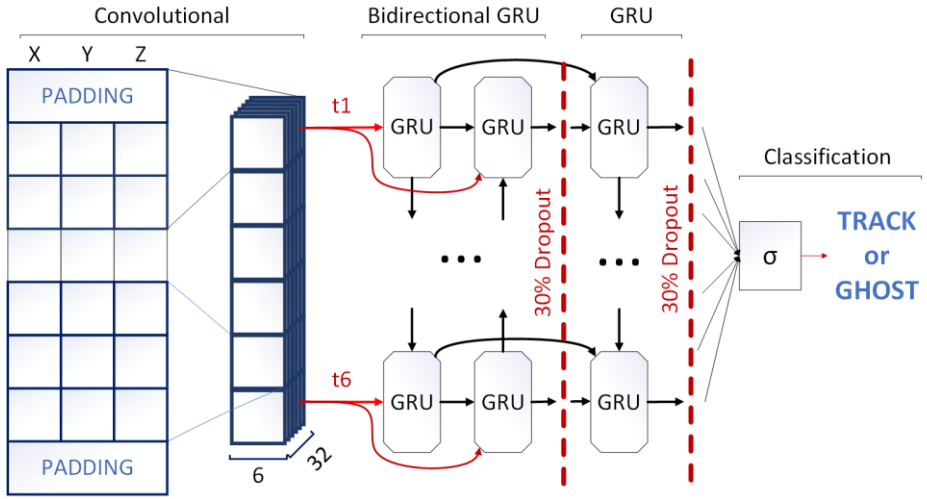
Real track
True found track
Ghost track

White dots are both hits and fakes




Кандидаты в треки

## 2) Deep Recurrent Neural Network Classifier




five hidden layers: convolutional network layer and two Gated Recurrent Unit (GRU) layers alternating with dropout layers

Results on simulated events: **True track recognition efficiency is on the level of 97.5%.**

# Использование данных в LHC Tier1/Tier2

**Задача:** Разделить максимально эффективно данные на несколько уровней: Очень горячий, горячий, теплый, холодный, очень холодный для поддержки эксперимента NICA и других с требуемой скоростью усвоения и обработки данных.

**Как было:** При наличии только двух технологий DRAM и HDD решение состояло из двух типов серверов и ленточных накопителей:

Сервер приложений:
2xIntel Xeon 26xx v2-3/128 DDR3/4x4TB HDD/10Gbs LAN

Сервер хранения данных:
2xIntel Xeon 2620 v2-4/128GB DDR3/12-24x6-8TB HDD/10Gbs LAN

Ленточные накопители (7ПБ)


**Состояние:** Неэффективно. Не позволяет поддержать NICA и другие HPC проекты по скорости обработки данных.


**Сделали:** Провели несколько пилотных проектов по использованию Intel NVMe SSD, Intel Optane.

# Данные для конвергентной HPC системы

**Задача:** Разделить максимально эффективно несколько десятков ПБ данных на несколько уровней: Очень горячий, горячий, теплый, холодный, очень холодный для поддержки эксперимента NICA и других с требуемой скоростью усвоения и обработки данных.

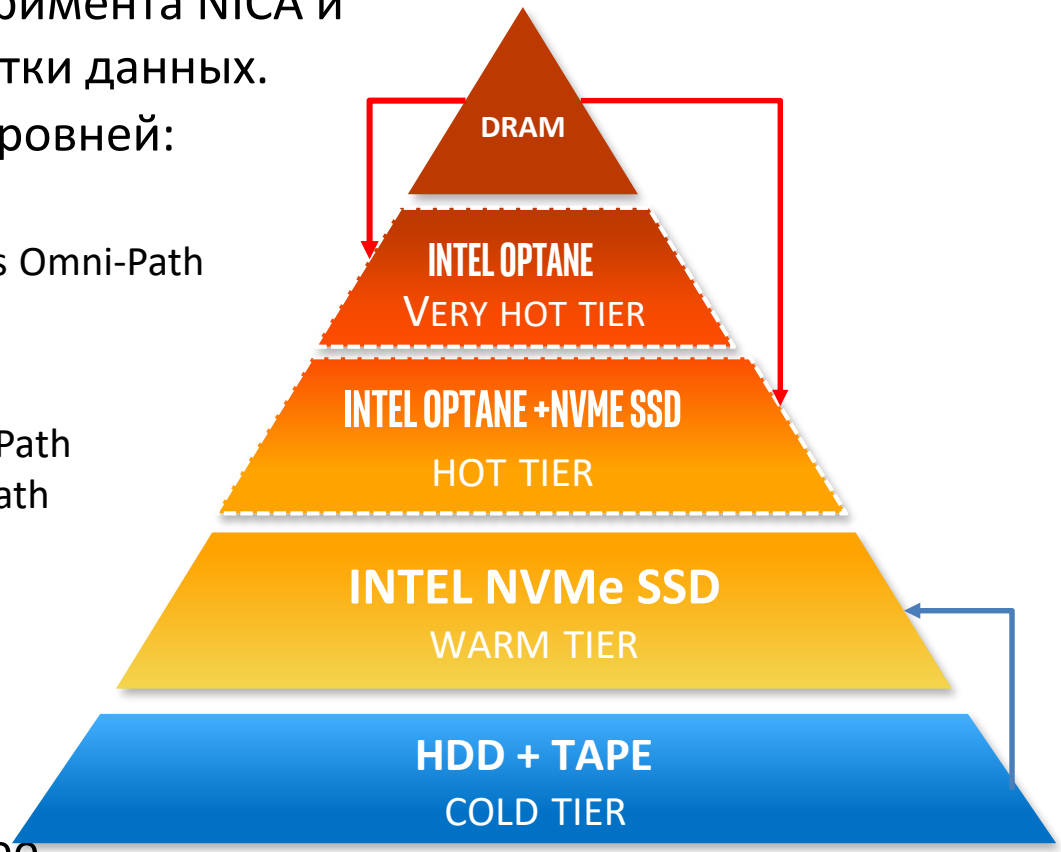
**Результат:** Комплексное решение включает в себя сервера уровней:

**Очень горячий**:
2xIntel Xeon 8268/12x**Intel Optane P4801X 375GB+IMDT**/100Gbs Omni-Path

**Горячий**:
2xIntel Xeon 8268/12x**Optane P4801X 375GB**/100Gbs Omni-Path
2xIntel Xeon 8268/12x**Intel NVMe SSD P4511 2TB**/100Gbs Omni-Path
2xIntel Xeon 8268/2x**Intel NVMe SSD P4511 2TB**/100Gbs Omni-Path

**Теплый**:
В проработке с **NVMe SSD**

**Холодный**:
2xIntel Xeon 6xxx/24x14TB HDD/10Gbs Ethernet и Ленты




**Состояние:** Производительность горячего слоя достигает более чем 300GB/s. Активно работаем над теплым слоем.