

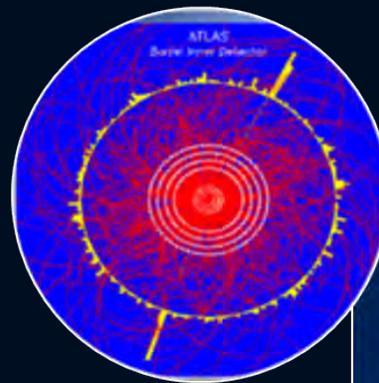
Статус и перспективы Многофункционального информационно- вычислительного комплекса ОИЯИ

Т.А. Стриж
Зам. научного руководителя ЛИТ ОИЯИ

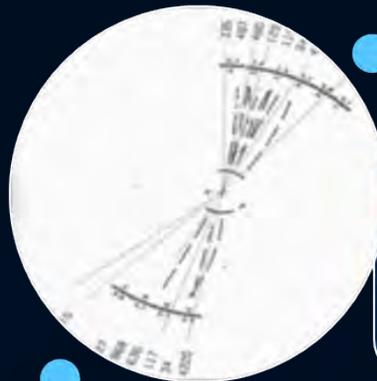
Мировое научное сообщество переходит к новой парадигме проведения научных исследований – значимые научные результаты могут быть получены только на основе анализа огромных массивов, накопленных в конкретных предметных областях данных, которые в настоящее время приобретают статус одного из важнейших стратегических ресурсов.



История: от малых данных к большим данным (пример физики элементарных частиц)



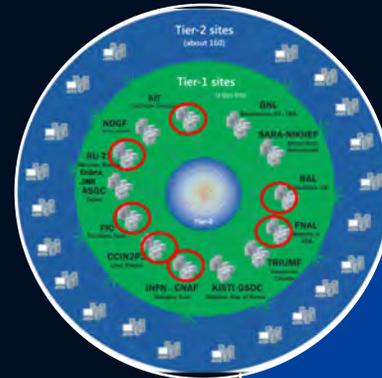
Открытие сегодня
в основном инклюзивные измерения
~2000 учёных в ~100 странах
сотни серверов Linux, суперкомпьютеров,
сетей, облаков и т. д.



Открытие 1970-х годов
более инклюзивные измерения
~200 учёных в ~10 странах
мейнфреймы



Открытие 1930-х годов
эксклюзивные измерения ~2
учёных в 1 стране ручка и
бумага



Грид = кластеры + облака
+ суперкомпьютеры



ПК Фермы, рабочие
станции

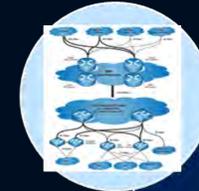
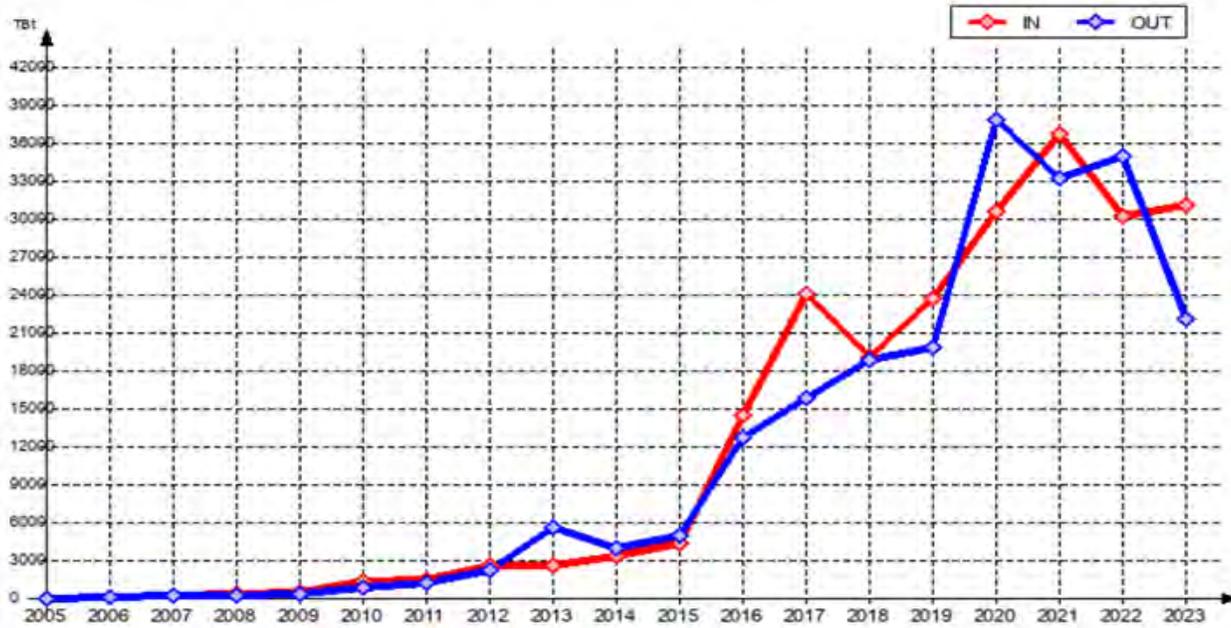


Мэйнфреймы EC 1060,
VAX, CONVEX

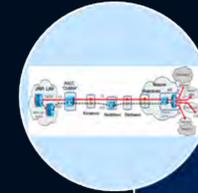
Телекоммуникационные каналы и локальная сеть



Общая статистика по годам.



3x100 Гбит/с



20 Гбит/с



1 Гбит/с

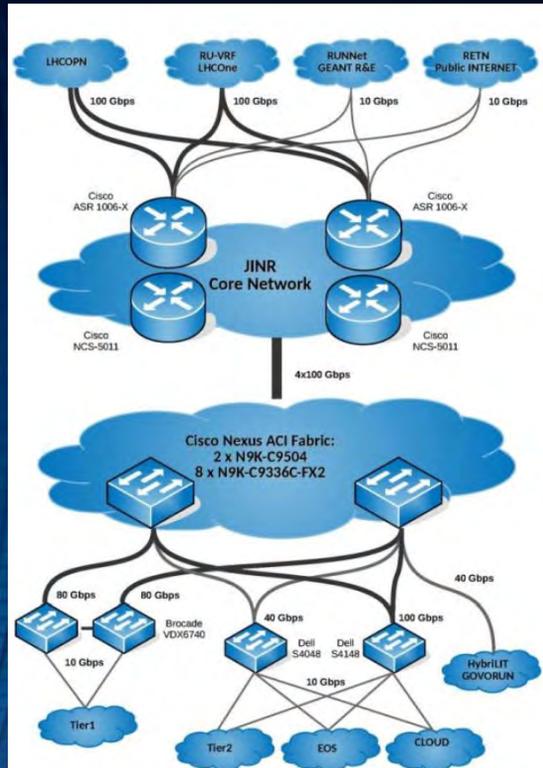


155 Мбит/с

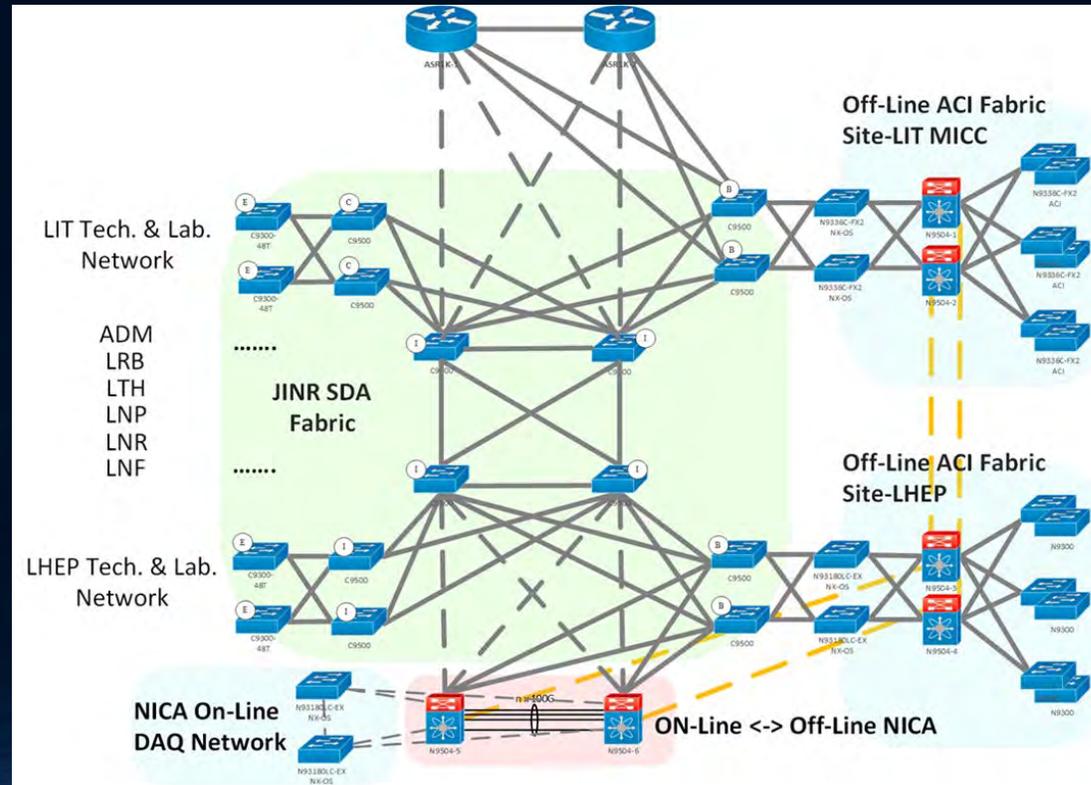
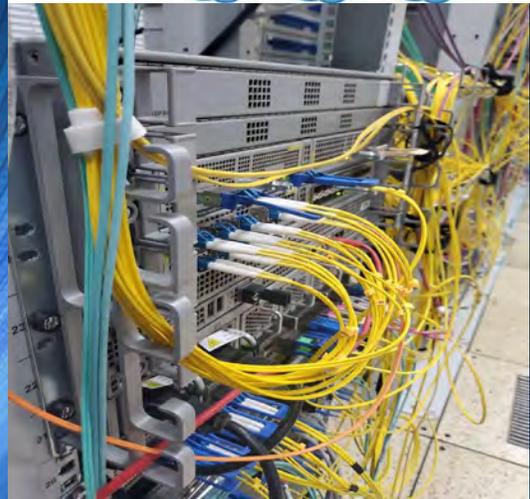


128 Кбит/с

Сетевая инфраструктура

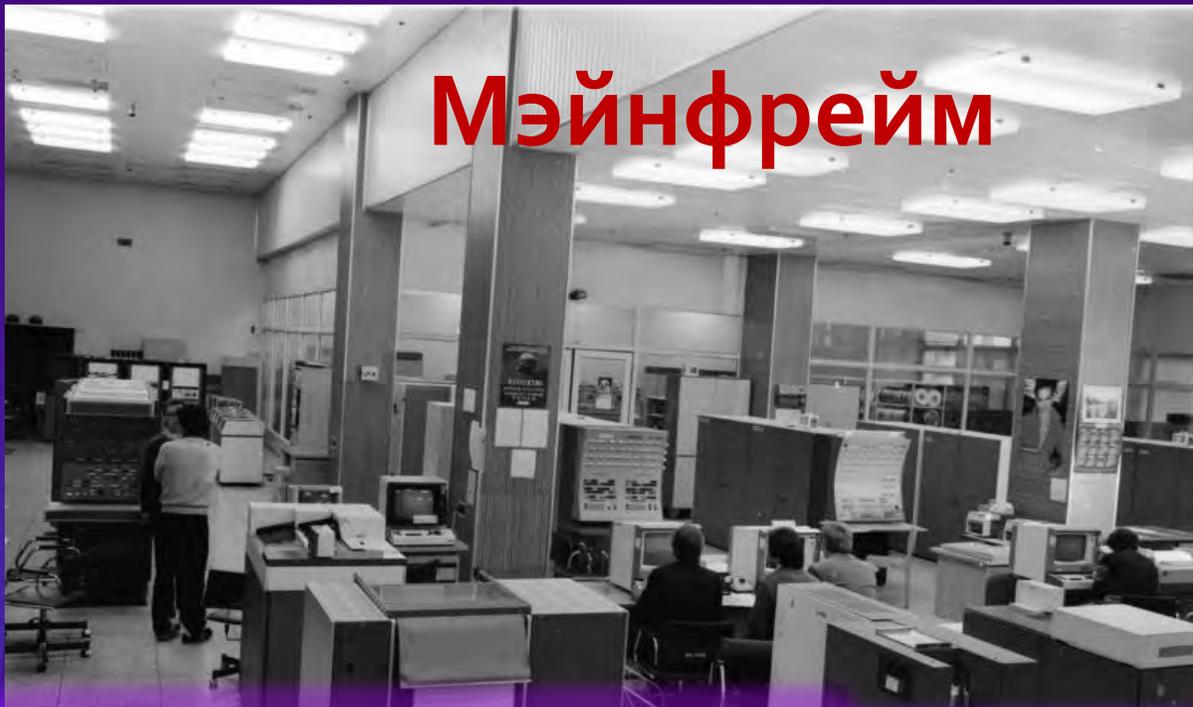


- ОИЯИ- Москва **3x100 Gbit/s** (+ 2x100 Gbit/s)
- ОИЯИ - ЦЕРН - **100 Gbit/s** и ОИЯИ - Амстердам **100 Gbit/s** для сетей LHCOPN, LHCONE, GEANT
- Прямые каналы связи до **100 Gbit/s** для связи с RUHEP центрами и сетями Runnet, ReTN
- Мультикластерная сеть **4x100 Gbit/s** между ЛФВЭ и ЛИТ



- Локальная сеть ОИЯИ:**
- 12816** сетевых элементов
 - 21764** IP-адресов
 - 5757** зарегистрированных пользователей
 - 4489** пользователей сервиса *.jinr.ru
 - 1159** пользователей электронных библиотек
 - 837** пользователей VPN и EDUROAM
- Сетевой трафик в 2023 году
- **30.5 PB** - входящий
 - **21.73 PB** - исходящий

Мэйнфрейм



Первые шаги ЛНС – PC фермы



CCIC JINR
130 CPU →
17TB RAID-5

10 – Interactive & UI
 32 – Common PC-farm
 30 – LHC
 14 – MYRINET (Parallel)
 20 – LCG
 24 – servers

3. Creation of a distributed high-performance computing infrastructure and mass storage resources

- Development of the JINR CICC as a core of the distributed infrastructure.
- Development of the hard- and software multipurpose infrastructure of the JINR CICC according to the requirements of collaborations and users of JINR and its member states as tabulated:

Year	2005	2006	2007	2010	2015
CPU(kSI2000)	100	660	1000	4000	10000
Disk Space(TB)	50	200	400	800	4000
Tape(TB)	1.5	50	450	1000	6000



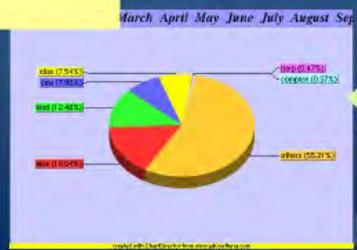
Total 501 users

LIT - 171
 DLNP - 104
 LPP - 53
 VBLHE - 44
 FLNR - 28
 NOJINR - 29
 BLTP - 14
 FLNP - 12
 Adm. - 9

Total 17 experiments

ATLAS - 44
 CMS - 24
 ALICE - 24
 HARP - 9
 COMPASS - 7
 DIRAC - 6
 DO - 3
 NEMO - 6
 OPERA - 3

Special groups for ATLAS, CMS, ALICE, LHCb, HARP, COMPASS, DIRAC, D0, NEMO, OPERA, HERMES, H1, NA48, HERA-B, IREN, STAR, KLOD



Group statistics (9 months 2005)

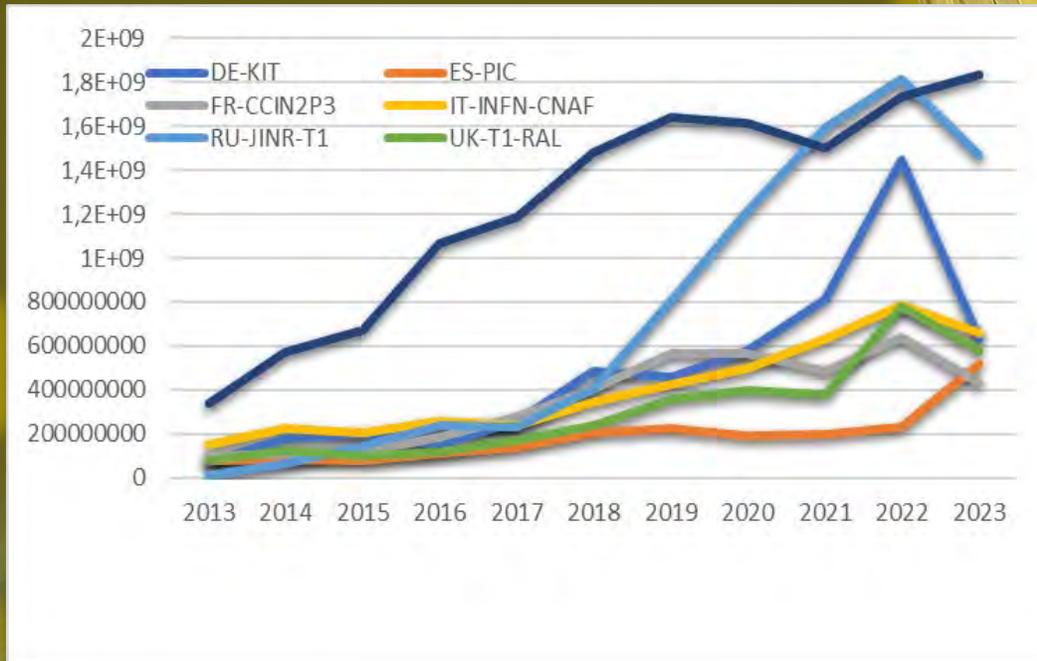
Tier2 - RDIG



eGee
Enabling Grids for E-science



2012 – Tier1 для CMS - прототип 1200 ядер, 720 ТВ диски, 72 ТВ ленты



Tier1

- Получение необработанных (RAW) экспериментальных данных от Tier0 в объеме, определенном соглашением WLCG
- Архивирование и ответственное хранение полученных экспериментальных данных.
- Последовательная и непрерывная обработка данных
- Дополнительная обработка (скимминг) данных RAW, RECO (RECO_nstructed) и AOD (данные объекта анализа).
- Повторная обработка данных с использованием нового программного обеспечения или новых констант калибровки и юстировки установки CMS.
- Обеспечение доступности наборов данных AOD
- Передача наборов данных RECO и AOD на другие сайты уровней 1/2/3 для их дублированного хранения (репликации) и физического анализа.
- Проведение производственной переработки с использованием нового программного обеспечения и новых калибровочных и юстировочных констант частей установки CMS, защищенное хранение моделируемых событий.
- Получение смоделированных данных и анализ данных, записанных в ходе эксперимента CMS.
- **Производство смоделированных данных и их анализ для экспериментов NICA (MPD, BM@N, SPD)**



Tier2

- Производство смоделированных данных и анализ данных для всех виртуальных организаций, зарегистрированных в РДИГ и всех экспериментов с участием ОИЯИ, использующих грид.
- **Производство смоделированных данных и их анализ для экспериментов NICA (MPD, BM@N, SPD)**



Инфраструктура и сервисы Tier1 (JINR-T1) и Tier2 (JINR-LCG2) обеспечивают работу:

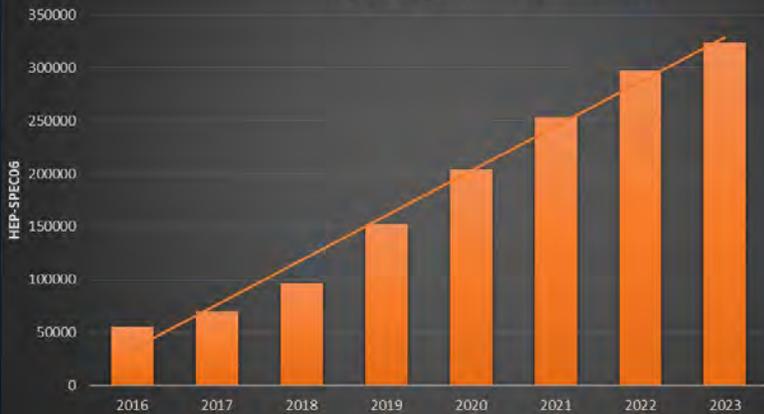
- вычислительного сервиса,
- сервиса хранения данных,
- сервиса доступа к домашним каталогам пользователей,
- сервиса доступа к версиям пользовательского ПО,
- сервисов поддержки грид,
- сервиса передачи данных,
- сервиса управления распределенными вычислительными системами,
- информационных сервисов (мониторинг, информационные сайты).

Общие сервисы для большинства компонент МИВК:

- kerberos, VOMS — аутентификация и авторизация доступа;
- AFS — домашние каталоги пользователей, установка и доступ к пользовательскому и групповому программному обеспечению, доступному по всему миру, как локальная файловая система с доступом POSIX;
- Серверы CVMFS (CernVM-File System) (stratum0/1) — установка и хранение программного обеспечения для совместной работы со многими версиями программного обеспечения, доступными по всему миру, например, локальная файловая система с доступом POSIX.
- Клиенты CVMFS и кеширование — доступ к программному обеспечению для совместной работы (только для чтения), используемому для доступа к локальным CVMFS и глобальным репозиториям со всего мира;
- EOS — хранение и доступ к большим объемам данных, доступных на интерактивных и вычислительных машинах, таких как локальная FS с доступом POSIX, доступ по всему миру через протоколы xroot и http;
- GIT — сервис для сборки и тестирования программного обеспечения для совместной работы с последующей установкой в CVMFS.

Инфраструктура и сервисы(Tier1 2023)

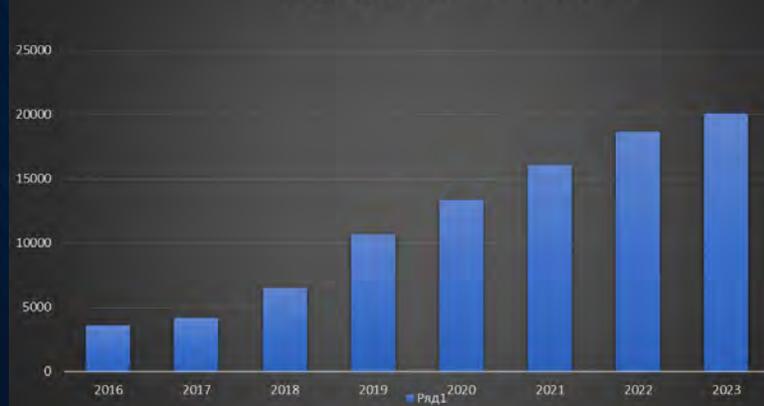
T1_RU_JINR Performance



Вычислительный ресурс (CE)

323820.54 HEP-SPEC06, 20096 ядер
 Среднее HEP-SPEC06 на ядро = 16.11
 468 машин
 CMS (пилоты по 16 ядер):
 Мах: 20096 ядер
 NICA (через DIRAC)
 Мах: 4000 ядер

T1_RU_JINR, number of cores



Системы хранения (SE)

dCache: SE disks: 11763.44 PB
 CMS @ dcache mss Total: 2642.24 TB
 Tapes@Enstore: 35562,00 TB
 Ленточные роботы: 51.5PB, IBM
 TS3500(11.5PB) + IBM T4500(40PB)
 EOS: 21829.01 TB
 CVMFS
 2 squid servers cache CVMFS

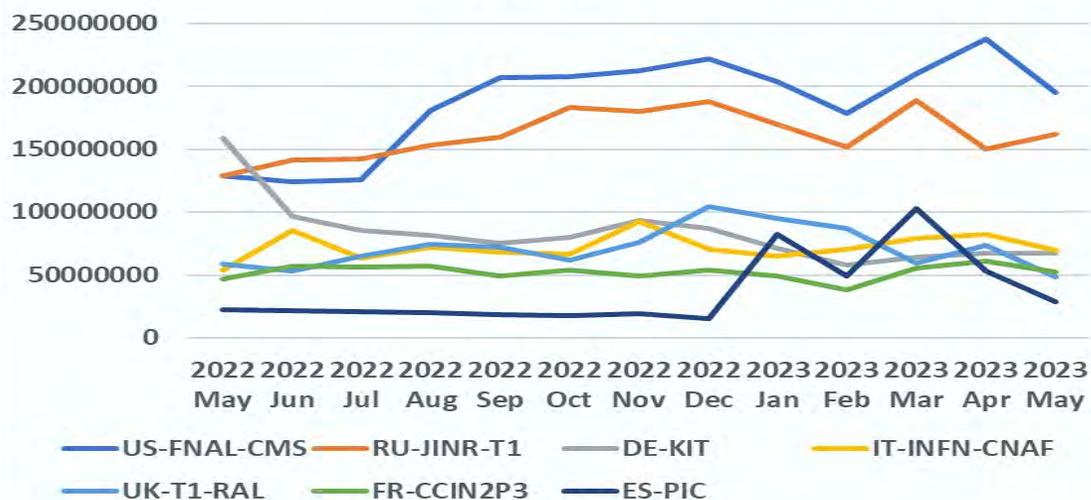
Программное обеспечение:

ОС: Scientific Linux версия 7.9.
 EOS 5.1.23
 dCache 8.2,
 Enstore 6.3.
 Slurm 20.11.
 grid UMD4 + EPEL (текущая версия)
 ARC-CE
 FairSoft
 FairRoot
 MPDroot

JINR Tier1

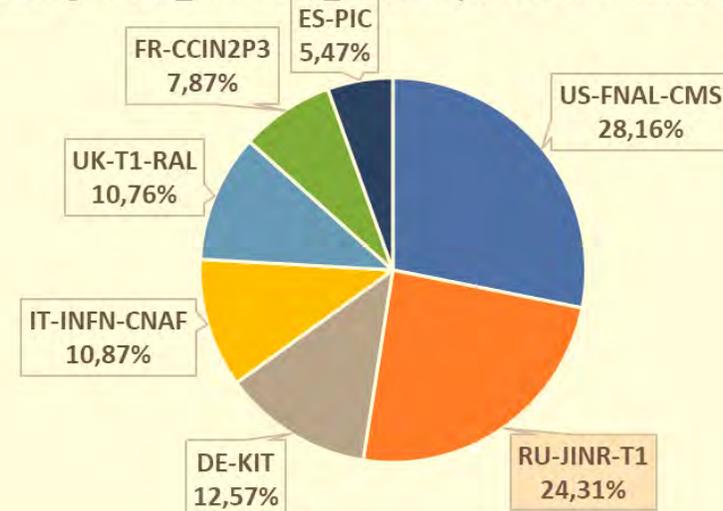


Accounting - 2022_5 to 2023_5 normcpu for CMS TIER1 and DATE

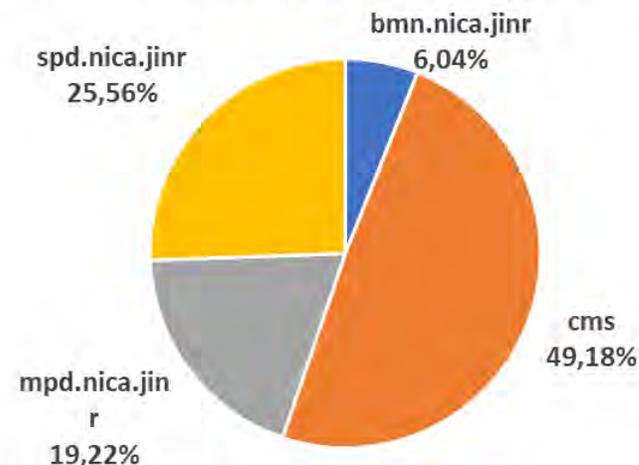


Наш Tier 1 регулярно занимает лидирующее место среди Tier1, обрабатывающих данные эксперимента CMS на БАК.

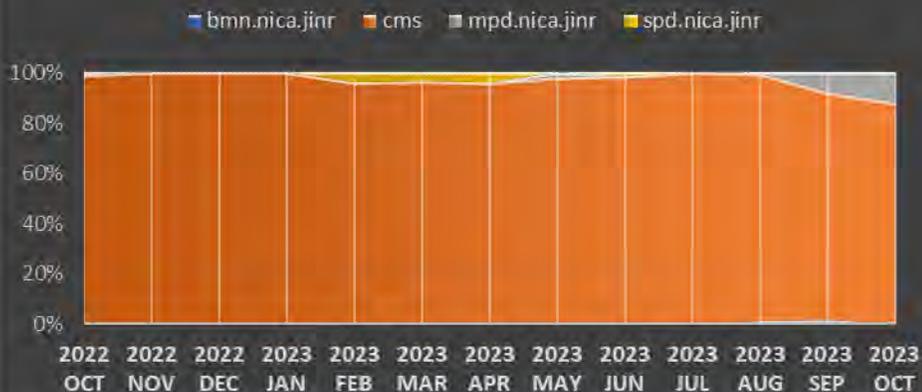
Accounting - 2022_5 to 2023_5 normcpu for CMS TIER1 and DATE



Accounting - 2022_5 to 2023_5 njobs at JINR Tier1 for VO



Tier1 accounting - 2022_10 to 2023_10 normcpu for VO and DATE



С 2019 года Tier1 ОИЯИ демонстрирует стабильную работу не только для CMS (БАК), но и для экспериментов NICA.

Detailed Monthly Site Reliability

Site	Jul-2022	Aug-2022	Sep-2022	Oct-2022	Nov-2022	Dec-2022
T0_CH_CERN	97%	98%	97%	97%	99%	99%
T1_DE_KIT	99%	100%	95%	100%	100%	96%
T1_ES_PIC	100%	99%	98%	99%	99%	99%
T1_FR_CCIN2P3	99%	99%	96%	98%	97%	94%
T1_IT_CNAF	100%	99%	90%	100%	100%	99%
T1_RU_JINR	98%	98%	98%	99%	98%	99%
T1_UK_RAL	98%	94%	95%	86%	99%	99%
T1_US_FNAL	99%	96%	96%	96%	96%	96%
Target	97%	97%	97%	97%	97%	97%

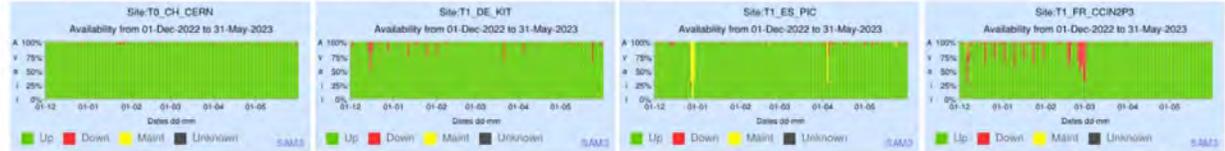


Availability of WLCG Tier-0 + Tier-1 Sites CMS

Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Dec-2022 - May-2023

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM



T0_CH_CERN Avail: 99% Unkn: 0% **T1_DE_KIT** Avail: 98% Unkn: 1% **T1_ES_PIC** Avail: 97% Unkn: 0% **T1_FR_CCIN2P3** Avail: 95% Unkn: 3%



T1_IT_CNAF Avail: 99% Unkn: 0% **T1_RU_JINR** Avail: 98% Unkn: 0% **T1_UK_RAL** Avail: 95% Unkn: 0% **T1_US_FNAL** Avail: 97% Unkn: 1%



Availability of WLCG Tier-0 + Tier-1 Sites CMS

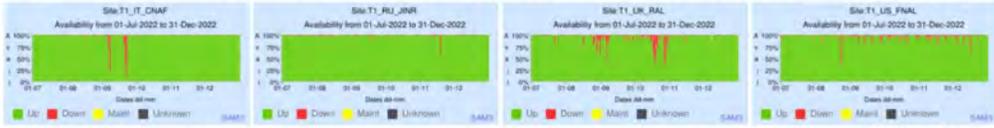
Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

Jul-2022 - Dec-2022

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM



T0_CH_CERN Avail: 98% Unkn: 0% **T1_DE_KIT** Avail: 98% Unkn: 1% **T1_ES_PIC** Avail: 98% Unkn: 0% **T1_FR_CCIN2P3** Avail: 97% Unkn: 2%



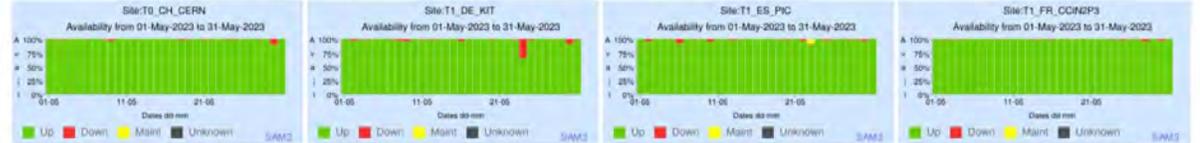
T1_IT_CNAF Avail: 98% Unkn: 0% **T1_RU_JINR** Avail: 98% Unkn: 0% **T1_UK_RAL** Avail: 95% Unkn: 0% **T1_US_FNAL** Avail: 96% Unkn: 2%

Availability of WLCG Tier-0 + Tier-1 Sites CMS

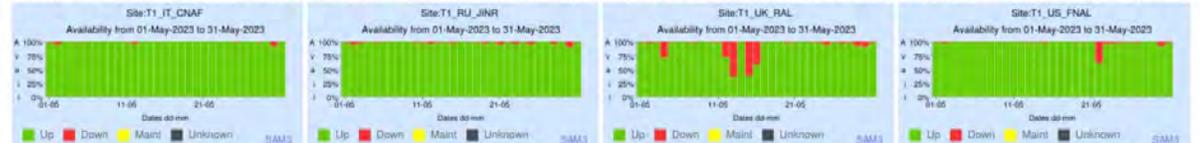
Target Availability for each site is 97.0%. Target for 8 best sites is 98.0%

May 2023

Availability Algorithm: (CREAM-CE + ARC-CE + HTCONDOR-CE) * all SRM



T0_CH_CERN Avail: 99% Unkn: 0% **T1_DE_KIT** Avail: 98% Unkn: 0% **T1_ES_PIC** Avail: 99% Unkn: 0% **T1_FR_CCIN2P3** Avail: 99% Unkn: 0%



T1_IT_CNAF Avail: 99% Unkn: 0% **T1_RU_JINR** Avail: 98% Unkn: 0% **T1_UK_RAL** Avail: 91% Unkn: 1% **T1_US_FNAL** Avail: 98% Unkn: 0%

Инфраструктура и сервисы (Tier2 2023)



Вычислительные ресурсы(CE):

Интерактивный кластер: lxpub [01-05] .jinr.ru

Интерфейс пользователей lxui [01-04] .jinr.ru (шлюз для внешнего соединения)

Вычислительный кластер.

485 машин

10356 ядер

166788.4 HEP-SPEC06

16.11 HEP-SPEC06 среднее на ядро

Система хранения(SE)

EOS=21829.04 TB

ALICE @ EOS 1653.24 TB

AFS: ~12.5TB (пользовательские директории)

CVMFS: 3 машины: 1 stratum0, 2 stratum1
2 squid servers cache CVMFS (VOs: NICA (MPD, B@MN, SPD), dstau, jjnano, juno, baikalgvd).

dCache : SE disks = 3753,69 TB

for CMS: 1903.2695 TB

for ATLAS: 1850.4248 TB

Local & EGI @ dcache2 Total: 256.91 TB

Программное обеспечение:

OS: Scientific Linux release 7.9.

EOS 5.1.23

dCache 8.2

BATCH: Slurm 20.11 адаптированный к kerberos и AFS

grid UMD4 + EPEL (текущие версии)

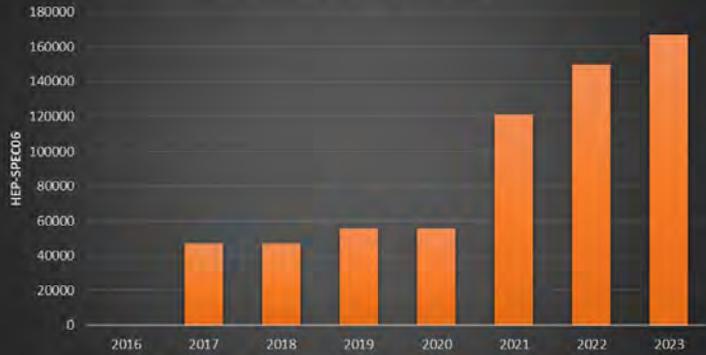
ARC-CE

FairSoft

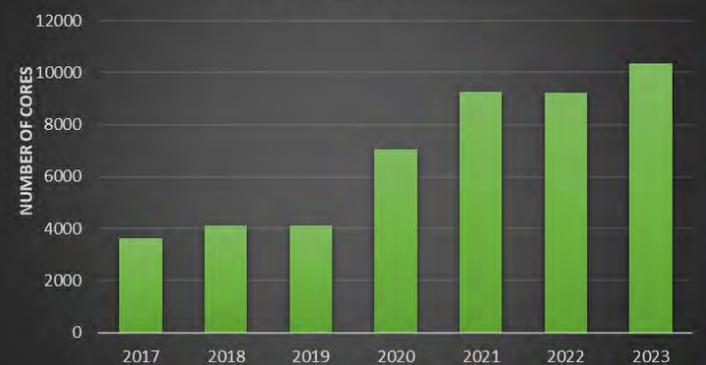
FairRoot

MPDroot

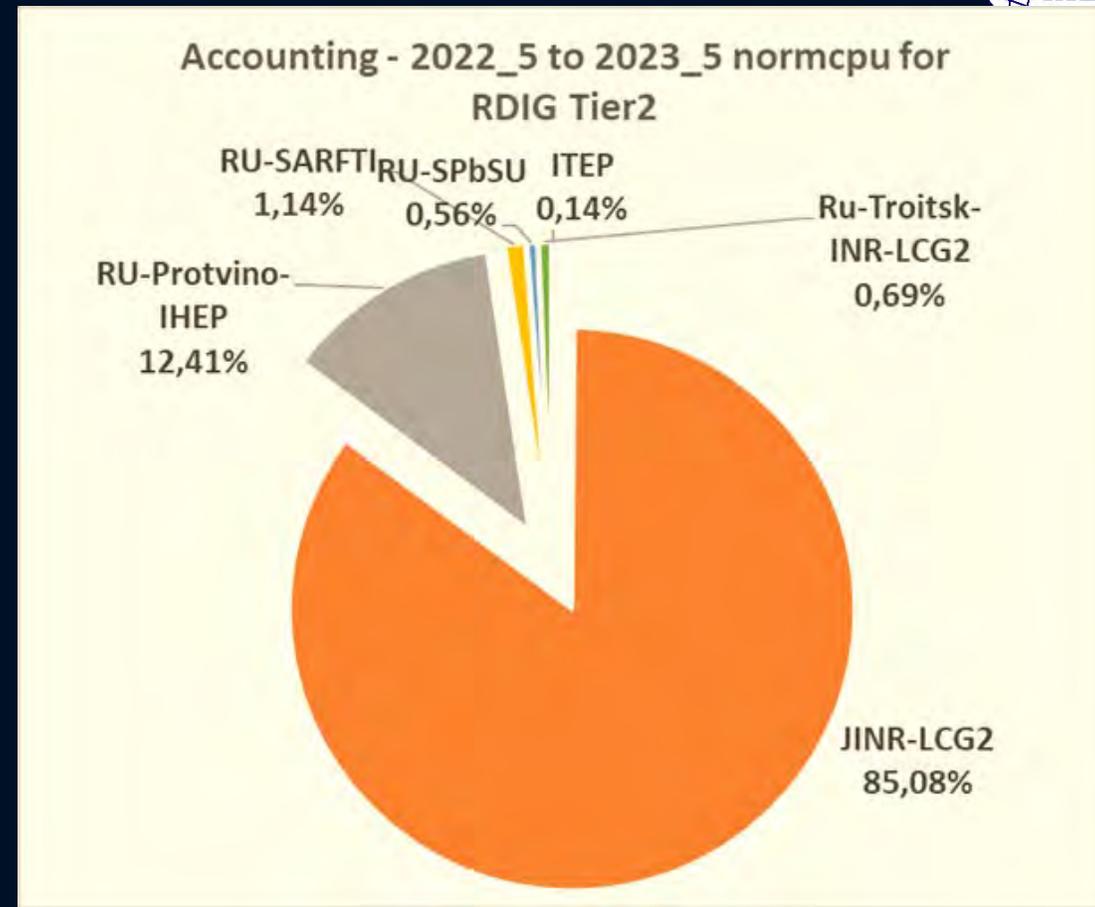
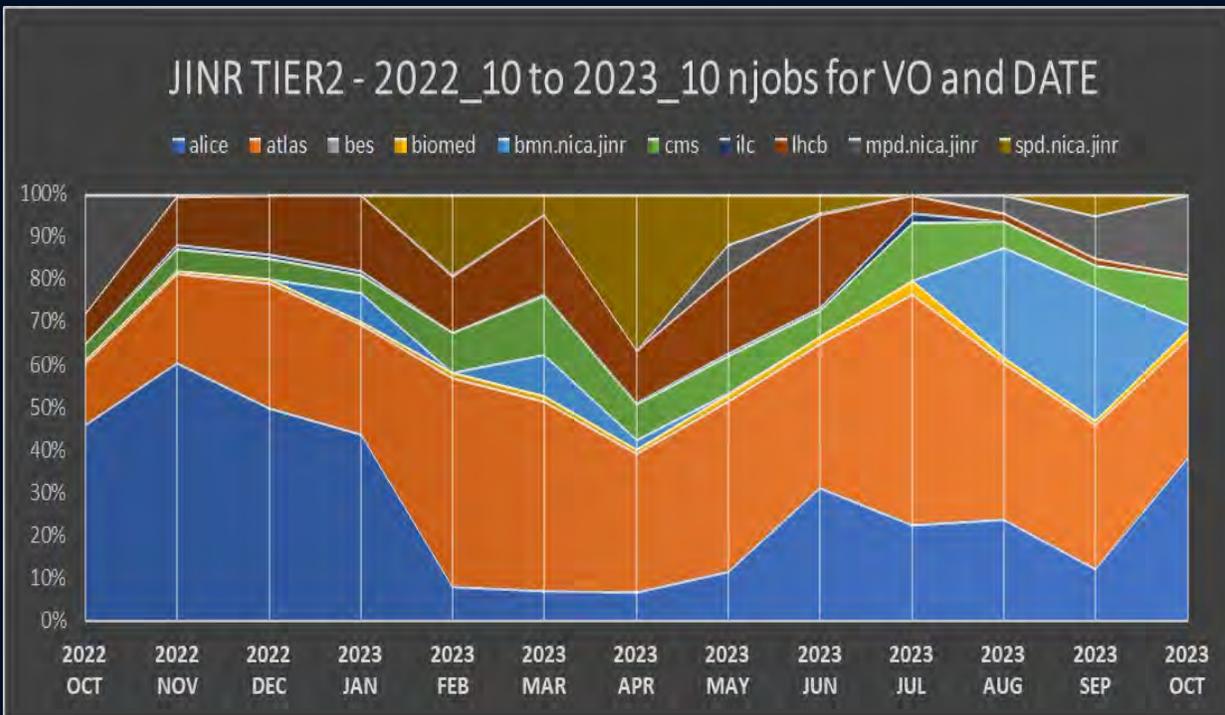
JINR Tier2 Performance



JINR Tier2 cores



JINR Tier2



Tier2 в ОИЯИ предоставляет вычислительные мощности, системы хранения данных и доступа к ним для большинства пользователей ОИЯИ и групп пользователей, а также для пользователей виртуальных организаций (VO) грид-среды (LHC, NICA и др.).

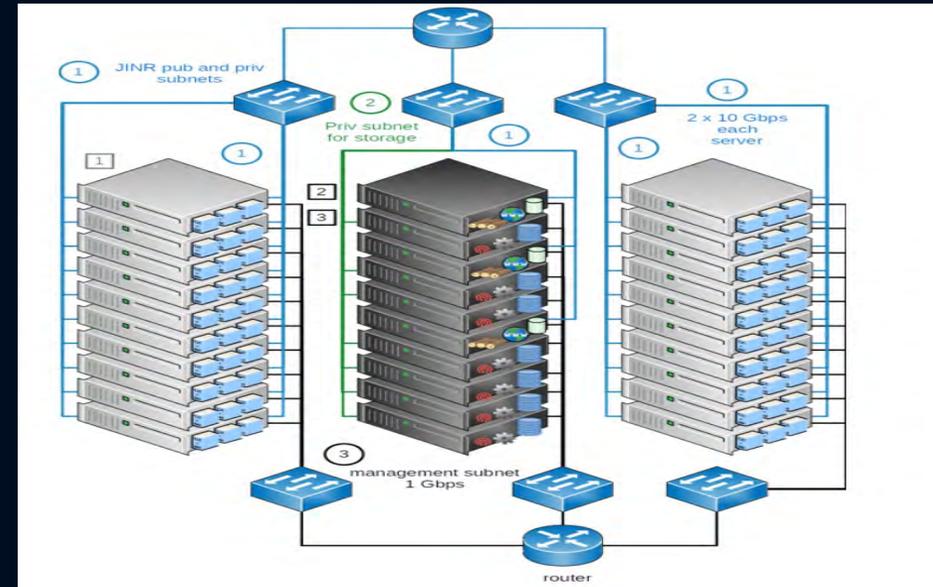
Tier2 ОИЯИ является самым производительным в Российском грид для интенсивных операций с данными (RDIG).
Более 80% общего процессорного времени в RDIG используется для вычислений на нашем сайте.

Облачные вычисления



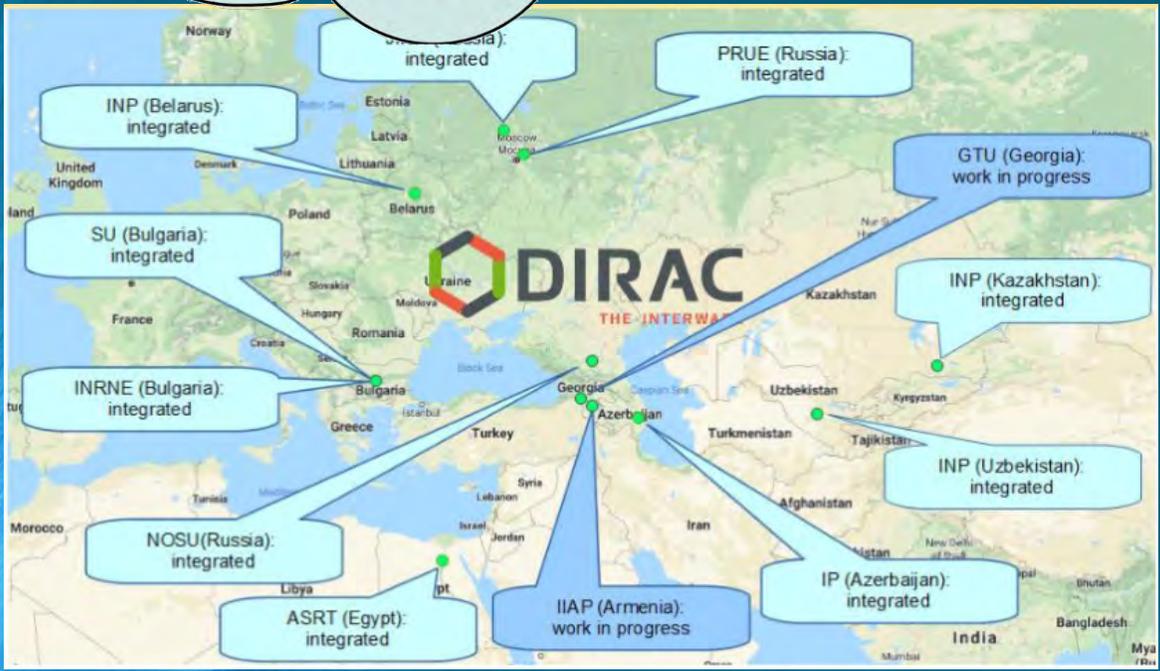
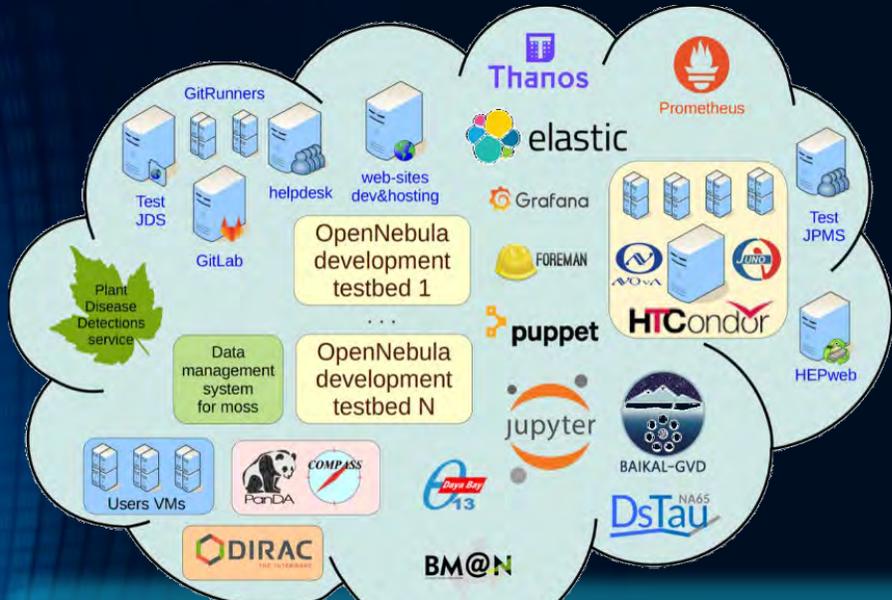
Облачные инфраструктуры ОИЯИ и его организаций-членов основаны на решении с открытым исходным кодом OpenNebula. Облако ОИЯИ является ядром этой инфраструктуры. На нем размещаются службы DIRAC, которые управляют вычислительными задачами и данными с использованием ресурсов ОИЯИ и его организаций-членов.

- Вычислительные ресурсы для нейтринных экспериментов:
- Виртуальные машины для пользователей ОИЯИ
- Испытательные стенды для исследований и разработок в области ИТ.
- Сервисы системы обработки данных эксперимента COMPASS
- Система управления данными МСП ЕЭК ООН по растительности.
- Сервис для визуализации данных, Gitlab и некоторые другие.
- Распределенная информационно-вычислительная среда на базе DIRAC (DICE), которая интегрирует облака организаций стран-членов ОИЯИ.

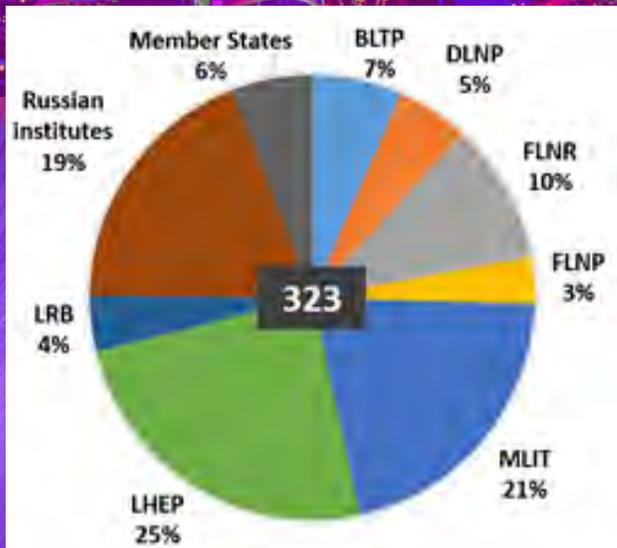


- Облачная платформа: OpenNebula (v5.12.0.4 CE)
- Виртуализация: KVM
- Серверное хранилище для образов виртуальных машин KVM: блочное устройство serph
- Пользовательские интерфейсы: веб-интерфейс и интерфейс командной строки.
- Аутентификация в облачном веб-интерфейсе: центральный пользователь ОИЯИ база данных (LDAP+Kerberos)
- Аппаратное обеспечение: 174 сервера для VM: >5000 ядер ЦП, ОЗУ на каждое ядро ЦП без HT: 5,3–16 ГБ
- 24 сервера для хранилищ Serph 3 ПБ
- URL веб-интерфейса: <http://cloud.jinr.ru>.

Облачные вычисления



HybriLIT



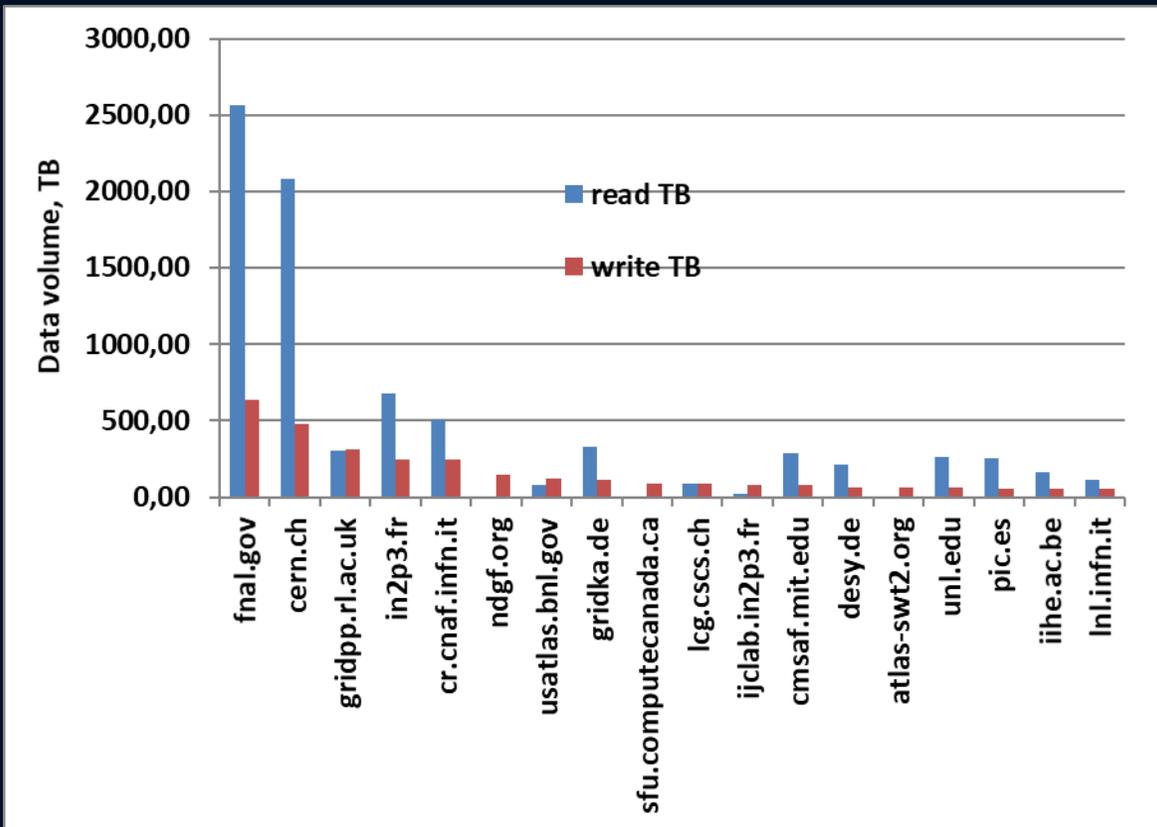
GPU-accelerator

Hyperconverged CPU and Distribut



Система долговременного хранения

Статистика по обмену данными с начала 2023



Системы хранение и данные.

TS3200 используется только для тестов.

TS3500 в режиме ожидания, в данный момент подключен к СТА
 TS4500 работает на CMS, половина емкости зарезервирована для NICA.

TS3500 12 ПБ свободно

TS4500 всего 40 ПБ, из них 20 CMS, 20 резервных

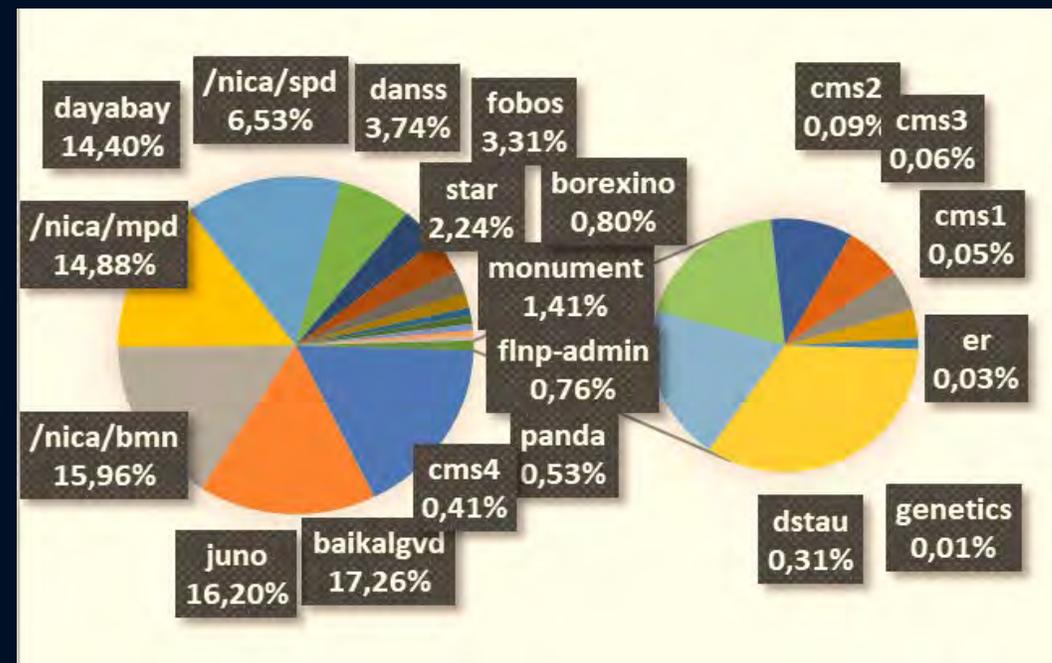
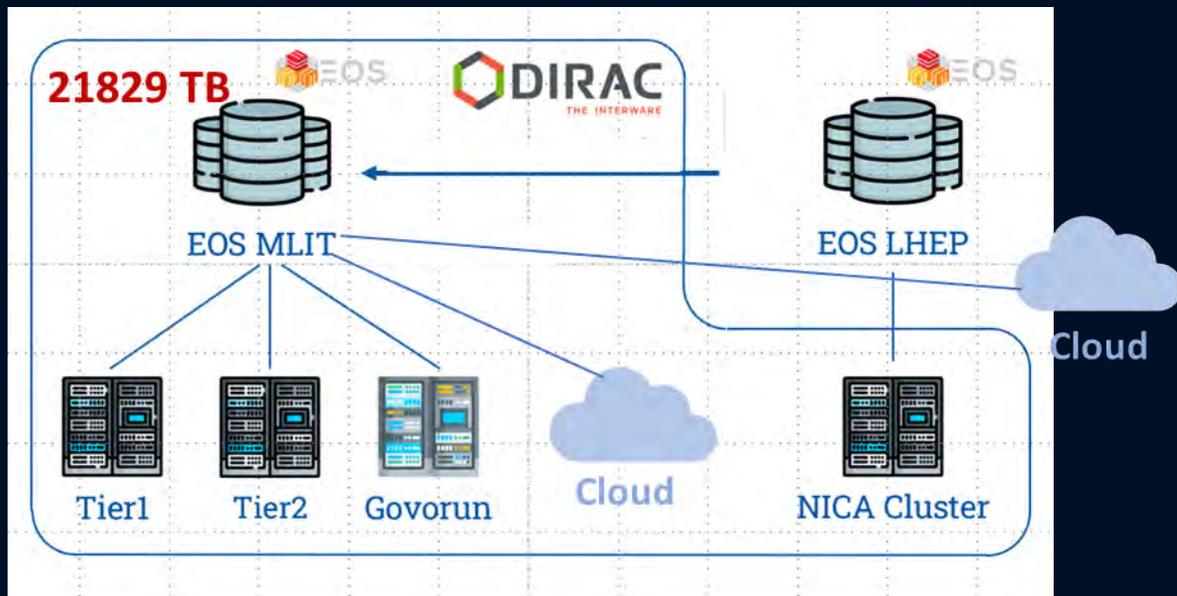
Объем записанных и считанных в 2023 году данных в PB

	T2	T1	T1mss
write	3.0	4.3	2.0
read	8.2	19.8	1.1

Развитие на ближайшие год-два

TS3500 12 PB, 12 LTO6 используются в качестве испытательной площадки для установки EOSСТА. Будет хранилищем для экспериментов, не связанных с WLCG.
 TS4500 40 PB 12 драйвов 3592-60F Jaguar будет разделен на 2 логические библиотеки
 20 PB 6 драйвов под управлением Enstore для CMS
 20 ПБ 6 драйвов под EOSСТА для NICA

Система среднесрочного хранения



EOS является системой хранения очень больших объемов данных.

Оптимален по соотношению стоимость/объем хранения.

Удобна для пользователей почти как локальная файловая система.

Поддерживает множество протоколов доступа: POSIX при установке на пользовательском компьютере; xroot и http для быстрого удаленного доступа.

Высокая надежность хранения данных за счет дублирования на разных серверах, хранения на разных серверах в формате вертикального RAID с контрольными суммами.

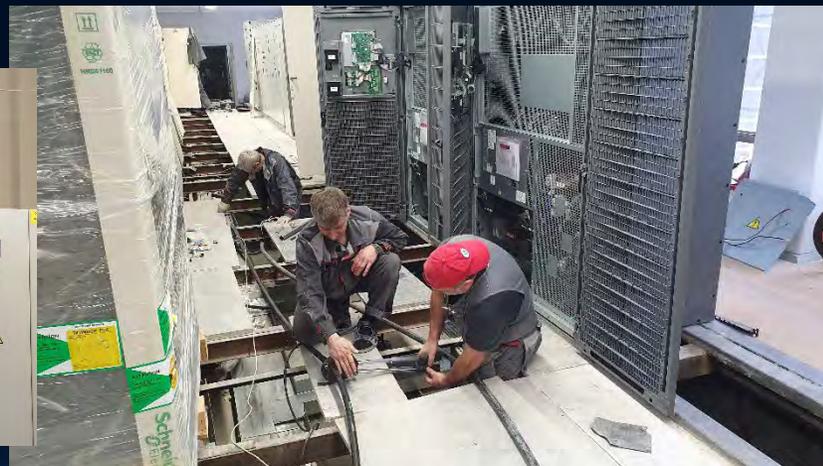
Высокая скорость доступа к данным за счет параллельного копирования с множества серверов.

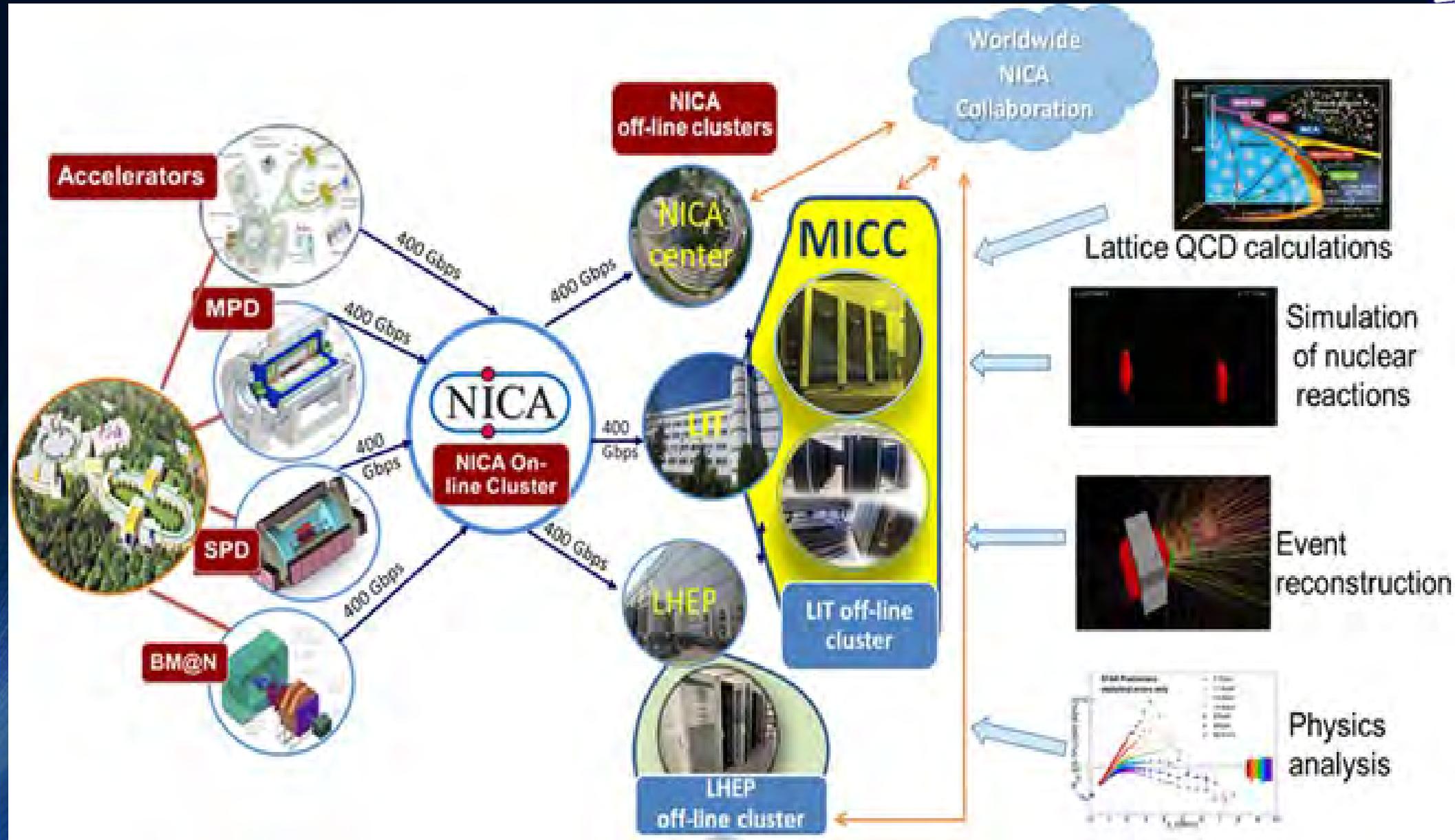
Защита данных с помощью расширенного списка модов доступа. Набор групп и отдельных пользователей.

Охлаждение



ЭНЕРГООБЕСПЕЧЕНИЕ

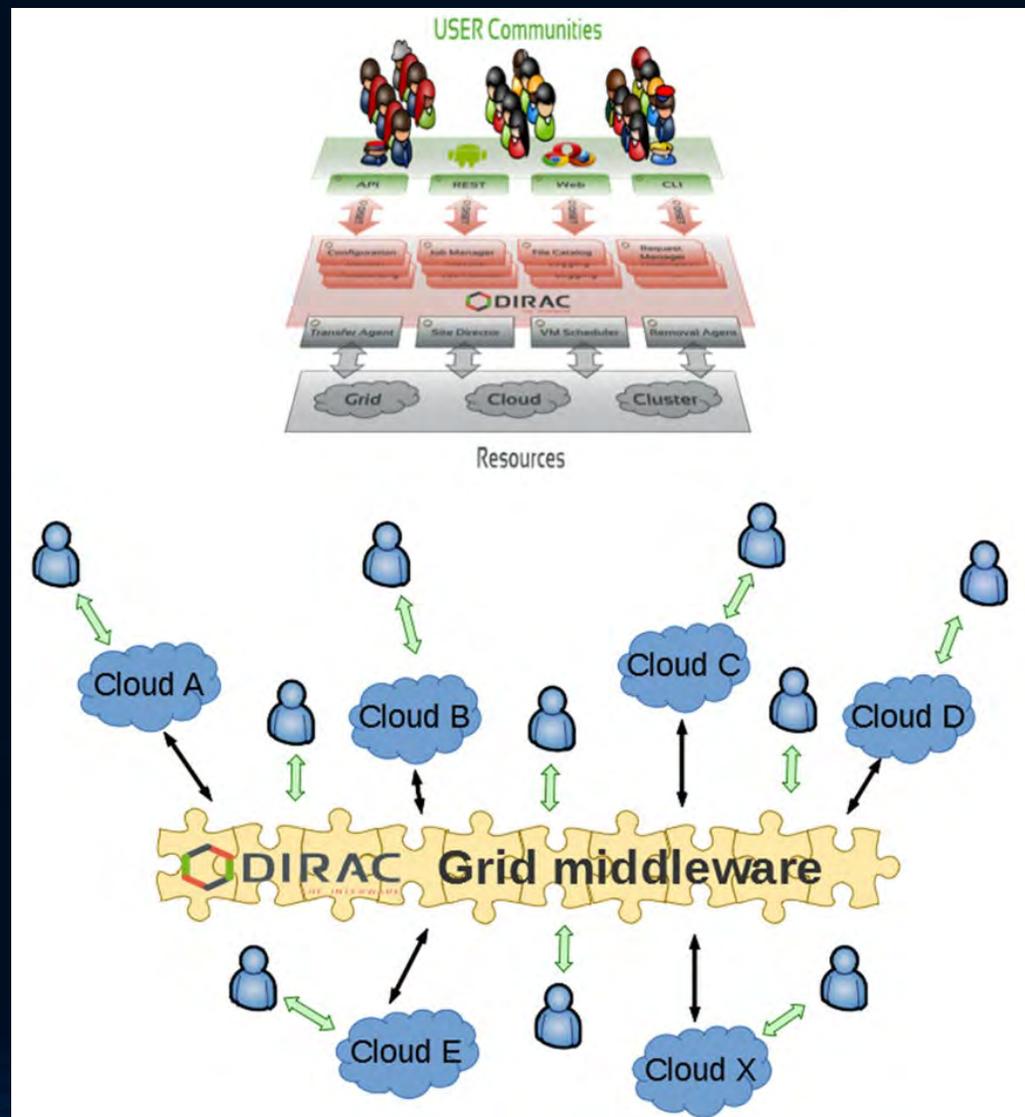




DIRAC @ ОИЯИ



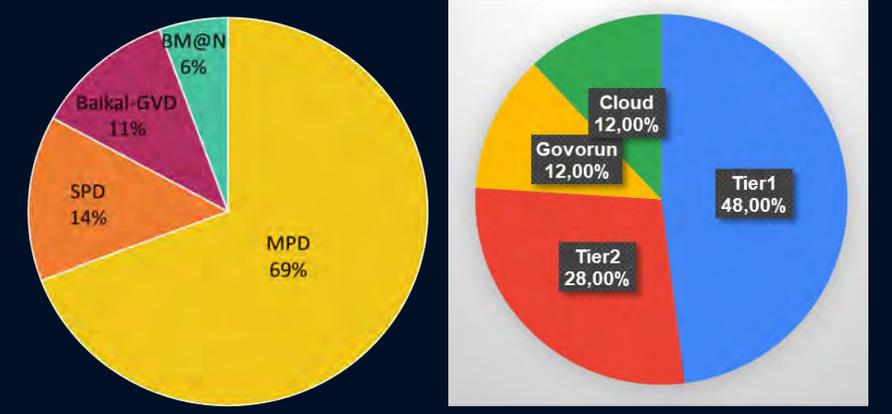
DIRAC (распределенная инфраструктура с контролем удаленного агента) INTERWARE — это программная среда для распределенных вычислений, обеспечивающая полное решение для одного (или более) сообщества пользователей, требующего доступа к распределенным ресурсам. DIRAC создает слой между пользователями и ресурсами, предлагая общий интерфейс для ряда гетерогенных поставщиков, интегрируя их бесшовно, обеспечивая интероперабельность, одновременно с оптимизированным, прозрачным и надежным использованием ресурсов.





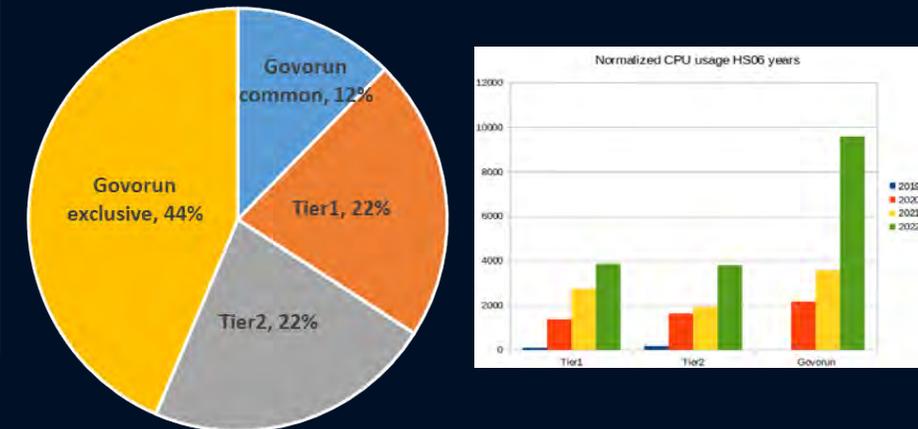
НИКС (Национальная исследовательская компьютерная сеть) — крупнейшая в России научно-образовательная сеть. Массовые испытания с заданиями MPD были успешно проведены в начале 2022 года.

Использование платформы DIRAC экспериментами



Основным пользователем платформы является эксперимент MPD

MPD Monte-Carlo



EGEE Enabling Grids for E-science

RDIG Russian Data Intensive Grid

RDIG monitoring & accounting

<http://rocmon.iinj.ru:8080>

MonALISA
MONitoring Agents using a Large Integrated Services Architecture

onALISA
MONitoring Agents using a Large Integrated Services Architecture

UChb jobs

V.V.Ivanov (LIT) PAC for Particle Physics

МОНИТОРИНГ



МИВК мониторинг и аккаунтинг

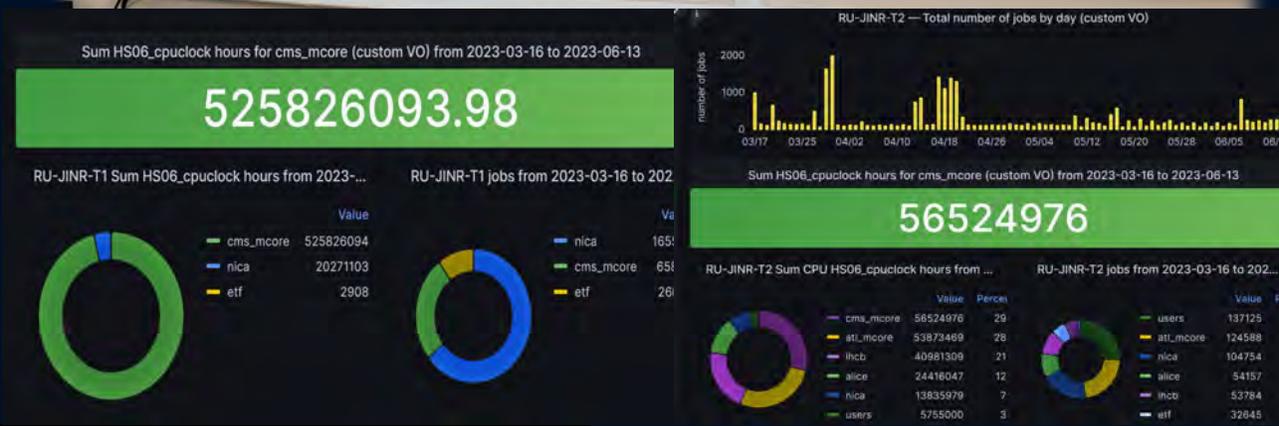


Успешное функционирование вычислительного комплекса обеспечивается системой, которая контролирует все компоненты МИВК.

Необходимо:

- расширить систему мониторинга, интегрировав в нее локальные системы мониторинга систем электроснабжения (дизель-генераторы, блоки распределения электроэнергии, трансформаторы и источники бесперебойного питания).;
- организовать мониторинг системы охлаждения (градирни, насосы, контуры горячей и холодной воды, теплообменники, чиллеры).;
- создать центр управления инженерной инфраструктурой (специальные информационные панели для визуализации всех статусов инженерной инфраструктуры МИВК в единой точке доступа);
- учитывать каждое пользовательское задание в каждом компоненте МИВК.

Требуется разработать интеллектуальные системы, которые позволят обнаруживать аномалии, что приведет к необходимости создания специальной аналитической системы в рамках системы мониторинга для автоматизации процесса.



3 monitoring servers

out 16000 service checks

About 1000 nodes

Семилетний план развития МИВК



1. Семилеткой предусмотрено создание на базе ЛИТ центра долгосрочного хранения данных на ресурсах МИВК.
2. Процесс моделирования, обработки и анализа экспериментальных данных, полученных с детекторов $BM@N$, MPD и SPD , будет реализован в распределенной вычислительной среде на базе МИВК и вычислительных центров ЛФВЭ и стран-участниц коллабораций.
3. Региональный центр обработки данных, предназначенный для производства, хранения и обработки для эксперимента JUNO. Ожидается, что этот центр обработки данных станет одним из трех европейских центров обработки данных JUNO. Ресурсы, необходимые для обработки и хранения данных JUNO, были одобрены сторонами в рамках «Меморандума о взаимопонимании по сотрудничеству в развертывании и эксплуатации вычислительной сети JUNO», подписанного между ИФВЭ и ОИЯИ 1 сентября 2022 г.
4. Продолжение работы в качестве Tier1 и Tier2 для LHC (HL-LHC).
5. Расширение инфраструктуры облачных вычислений.
6. Дальнейшее развитие, наращивание производительности и возможностей суперкомпьютера «Говорун».

Информационно-вычислительный блок комплекса NICA в ОИЯИ включает в себя:

1. онлайн-кластер NICA,
2. автономный кластер NICA в ЛФВЭ,
3. все компоненты МИВК (Tier0, Tier1, Tier2, суперкомпьютер «Говорун», облачные вычисления),
4. многоуровневую систему хранения данных,
5. распределенную вычислительную сеть.

NICA Tier 0,1,2	2024	2025	2026	2027	2028	2029	2030
CPU (PFlops)	2.2	2.6	8.6	8.6	15.6	15.6	15.6
DISK (PB)	17	24	47	75	96	119	142
TAPE (PB)	45	88	170	226	352	444	536
NETWORK (Gbps)	400	400	800	800	800	1000	1000

Следует отметить, что ресурсы, представленные в таблице могут быть покрыты примерно на 20-25% бюджета, выделенного на МИВК.

Развитие серверных залов МИВК

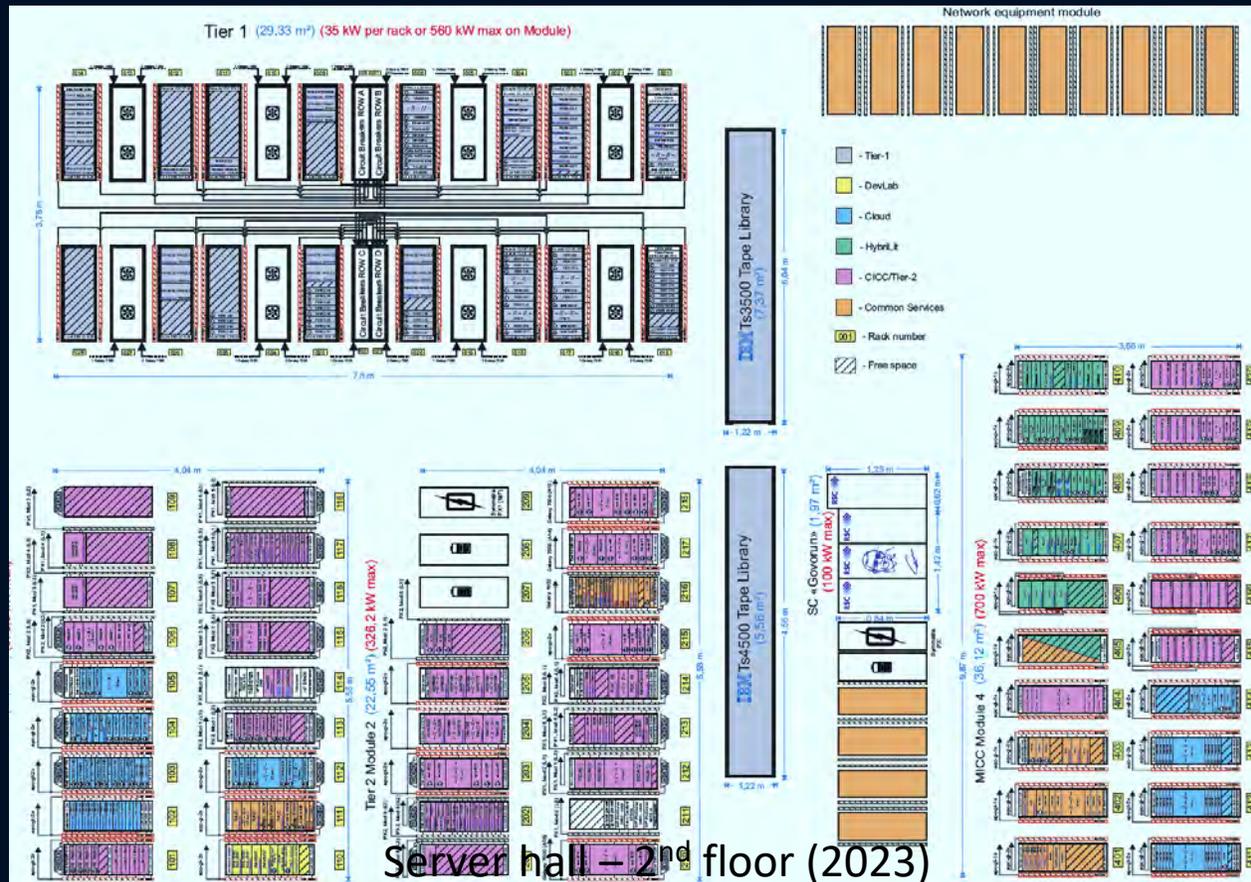


Сегодня (1000 кВт)

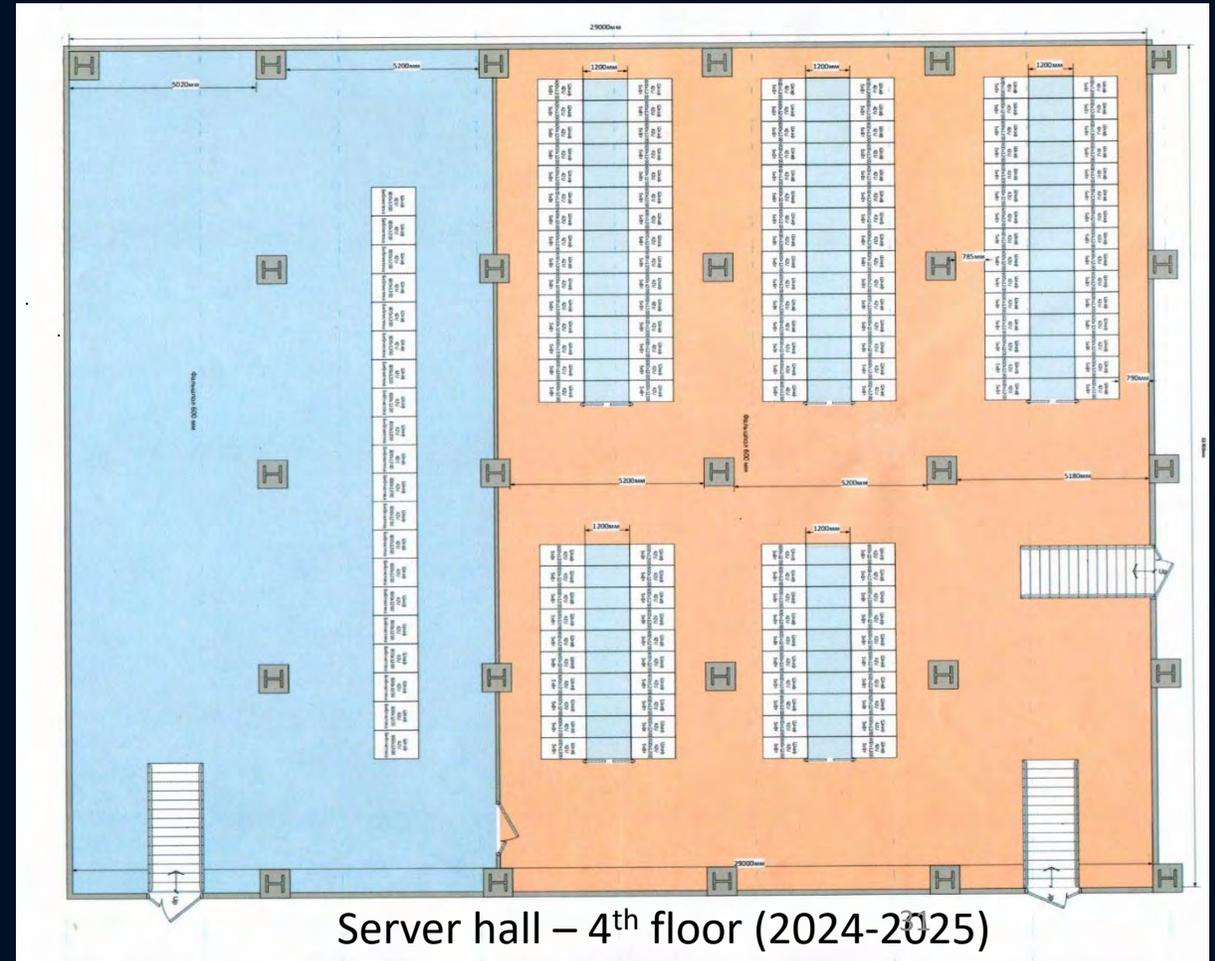
- 69 стоек для серверов
- 4 стойки для СК «Говорун»
- 10 стоек для сетевого оборудования
- 4 стойки для административных сервмсов
- 2 роботизированные ленточные библиотеки

Планируем – новый серверный зал МИВК (600 кВт)

- зона роботизированных ленточных библиотек
- 130 стоек для серверов



Server hall – 2nd floor (2023)



Server hall – 4th floor (2024-2025)



Облачная инфраструктура

Вам нужно больше компьютеров для исследований?

Создайте их в нашем облачном веб интерфейсе. Выберите необходимое Вам количество ядер, ОЗУ и операционную систему для своих целей.



Гетерогенная платформа

Нужны параллельные преимущества современных графических ускорителей?

Используйте 1000 ядер в один момент, чтобы получить результаты так быстро, как это возможно.



Грид-инфраструктура

Нужен анализ данных экспериментов БАК?

Получите доступ к нашему грид кластеру для выполнения анализа.



ЦИВК

Нужны ресурсы для длительных вычислений?

ЦИВК - это набор серверов, которые вы можете загружать своими задачами. Чтобы использовать параллельные функции фермы, используйте MPI задачи.

MICC

DIRAC, PanDA, etc

Tier1
20064
Cores
11,8 PB

Tier2/CCIC
10564
Cores
5,65 PB

GOVORUN
1,7 Pf
8 PB

CLOUD
5152
Cores
3,5 PB

DATA STORAGE 75PB

NETWORK 3x100Gbps

POWER@COOLING 800 kVA@1400kW

Основная цель проекта — обеспечить

- многофункциональность,
- масштабируемость,
- высокую производительность,
- надежность и доступность в режиме 24x7x365 для различных групп пользователей, выполняющих научные исследования в рамках Тематического плана ОИЯИ.



Спасибо за внимание