

# О стратегии развития IT-технологий и научного компьютеринга в ОИЯИ

**Кореньков Владимир Васильевич**

**директор ЛИТ ОИЯИ**

**НТС ЛИТ, 6 июня 2019**

# Программы в области развития цифровых технологий

Согласно федеральному проекту «Цифровые технологии», к девяти прорывным направлениям были отнесены: нейротехнологии и искусственный интеллект, технологии виртуальной и дополненной реальности, компоненты робототехники и сенсорики, технологии беспроводной связи, большие данные, системы распределенного реестра (блокчейн), промышленный интернет, новые производственные технологии и квантовые технологии.

- Центры Tier1 и Tier2 российского сегмента Глобальной вычислительной системы Грид БАК
- Прогноз развития суперкомпьютерных и грид-технологий в Российской Федерации на долгосрочный период
- Концепции создания и обеспечения функционирования национальной суперкомпьютерной инфраструктуры.

# The Worldwide LHC Computing Grid

**Tier-0 (CERN):**  
data recording,  
reconstruction and  
distribution

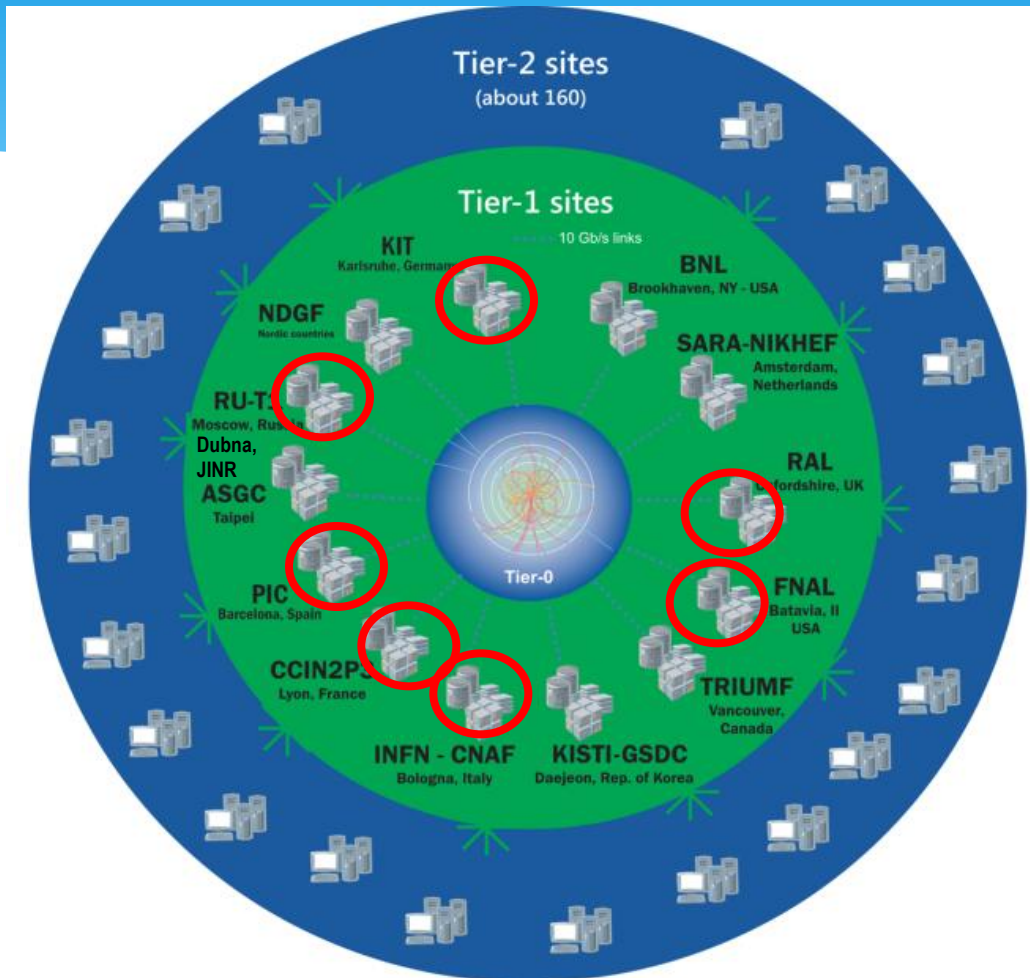
**Tier-1:**  
permanent storage,  
re-processing,  
analysis

**Tier-2:**  
Simulation,  
end-user analysis

**WLCG:**

An International collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists



nearly 180 sites,  
45 countries

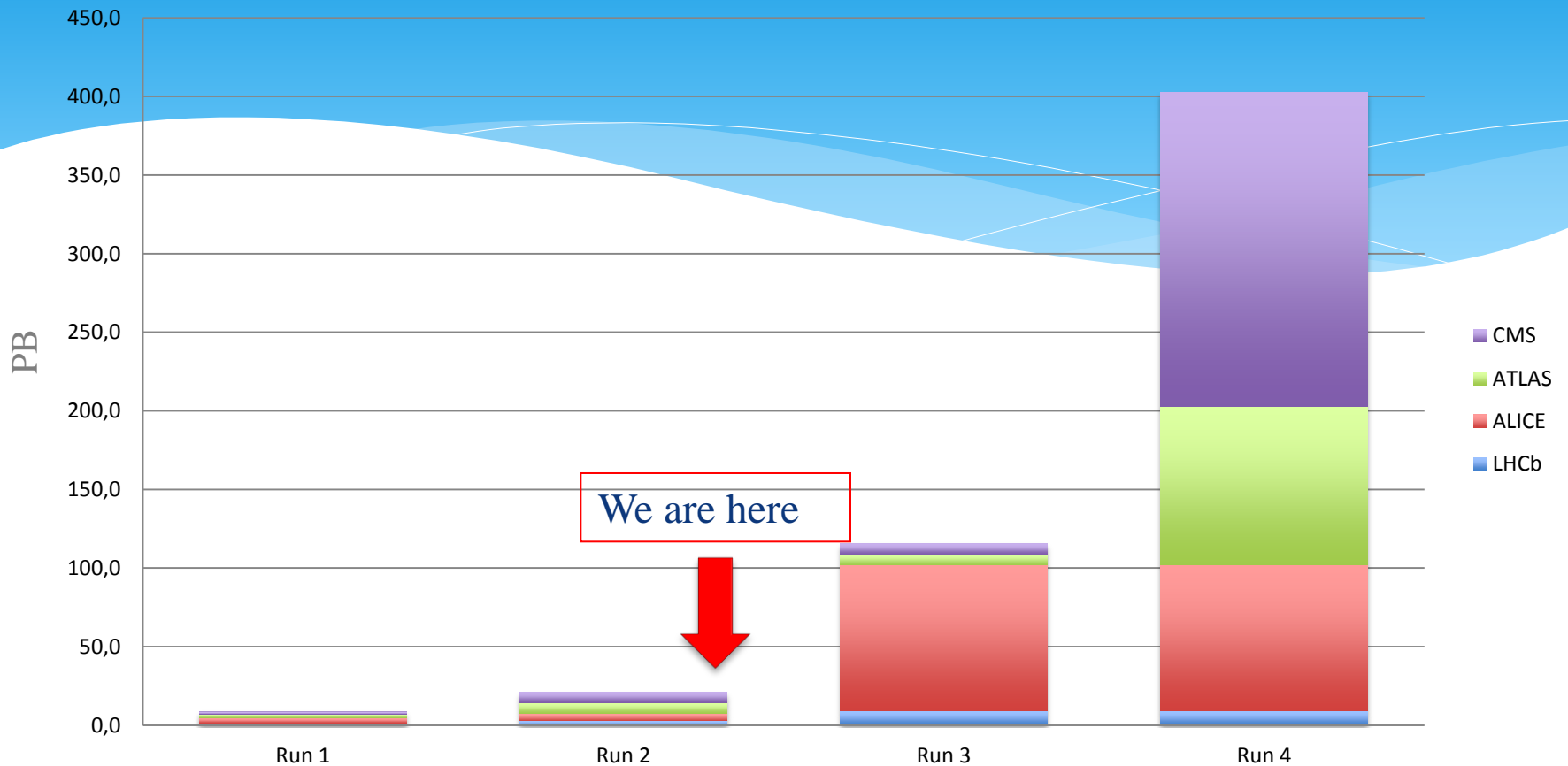
1 000 000 cores

1 EB of storage

> 3 million jobs/day

10-100 Gb links

# Data: Outlook for HL-LHC



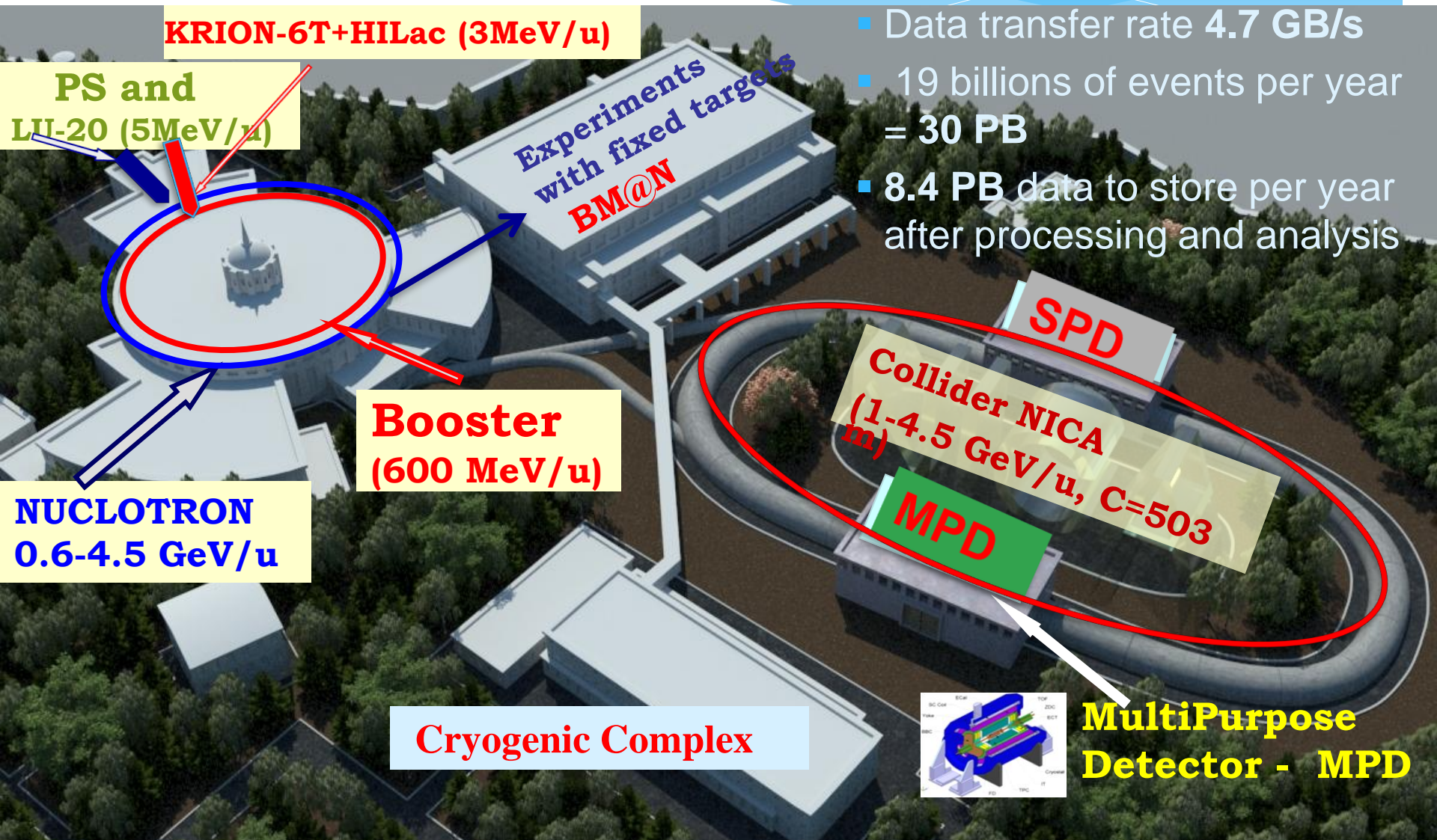
- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
  - To be added: derived data (ESD, AOD), simulation, user data...

# NICA Complex: New era in the hot dense matter science

## Collider basic parameters:

$\sqrt{s_{NN}} = 4-11$  GeV; beams: from p to Au;  $L \sim 10^{27}$  cm<sup>-2</sup> c<sup>-1</sup> (Au),  $\sim 10^{32}$  cm<sup>-2</sup> c<sup>-1</sup> (p)

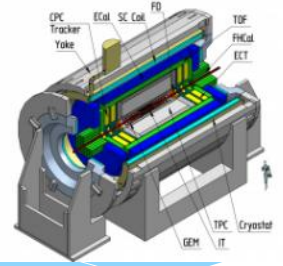
- Data transfer rate 4.7 GB/s
- 19 billions of events per year = 30 PB
- 8.4 PB data to store per year after processing and analysis



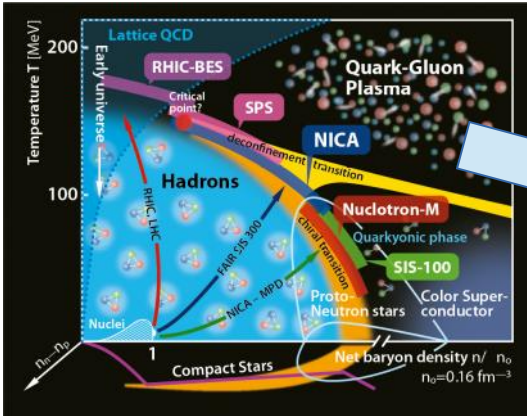
**MultiPurpose Detector - MPD**



# Вызовы компьютеринга для NICA

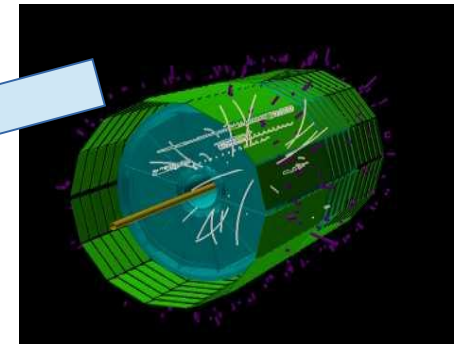


Эксперимент MPD

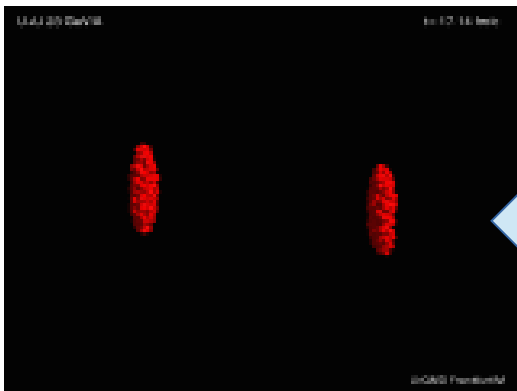


Фазовая диаграмма КХД

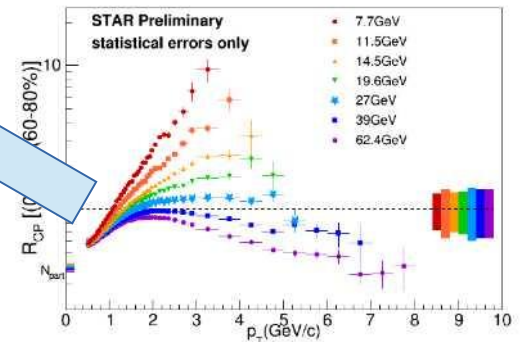
## Суперкомпьютер имени Н.Н.Говоруна



Реконструкция событий

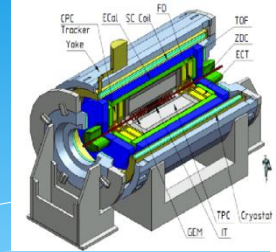


Моделирование

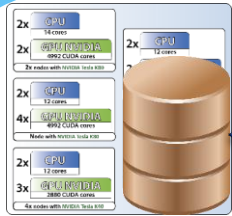


Физический анализ

# Computing for NICA Megaproject on GOVORUN



GOVORUN



“Hot” Data Storage  
Lustre FS

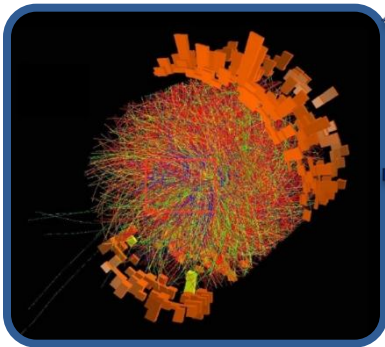


“Cold” Data Storage  
EOS FS

“Cold” Data Storage  
ZFS FS



MpdRoot

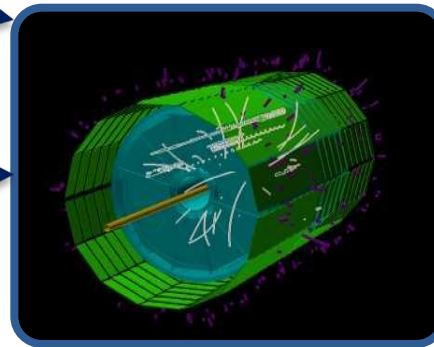


Particles Generator

- internal
- external

More than 60 million events was generated for MPD experiment

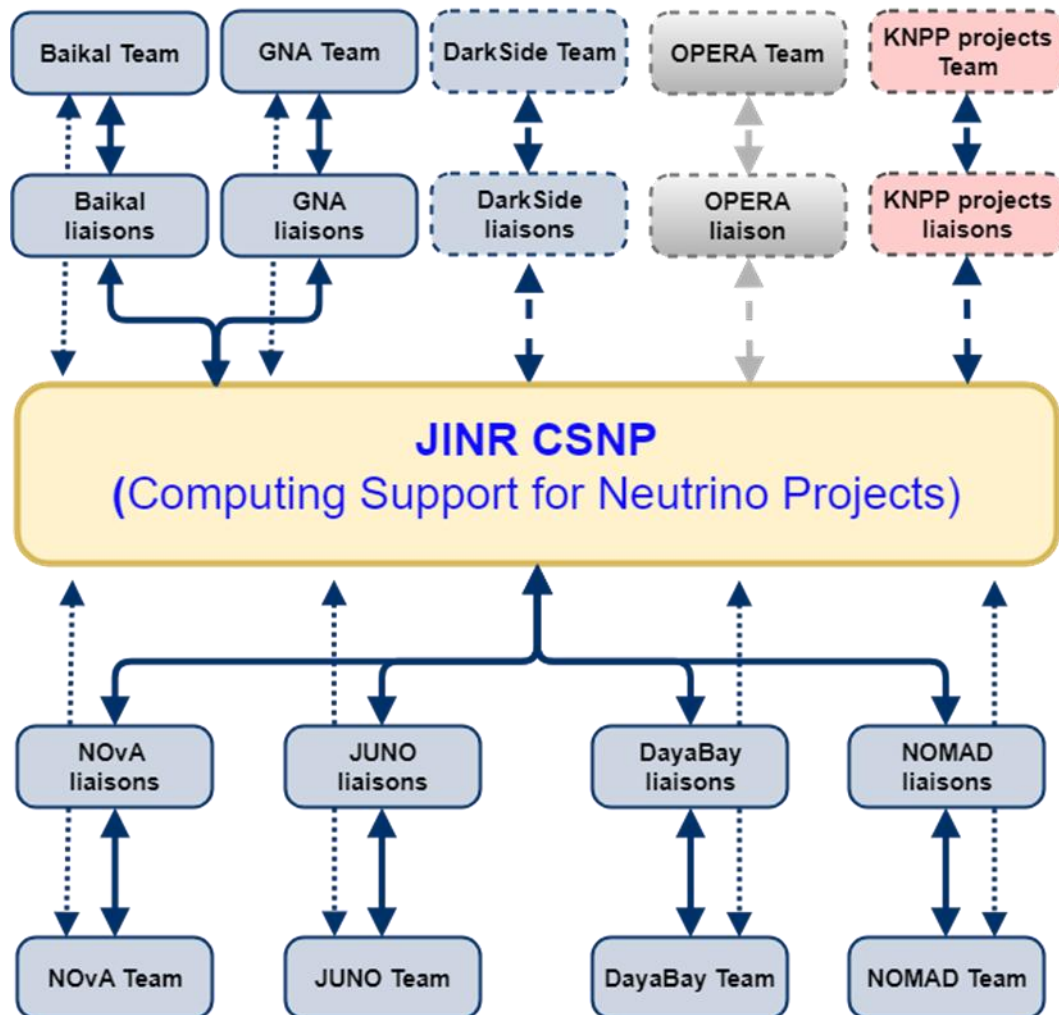
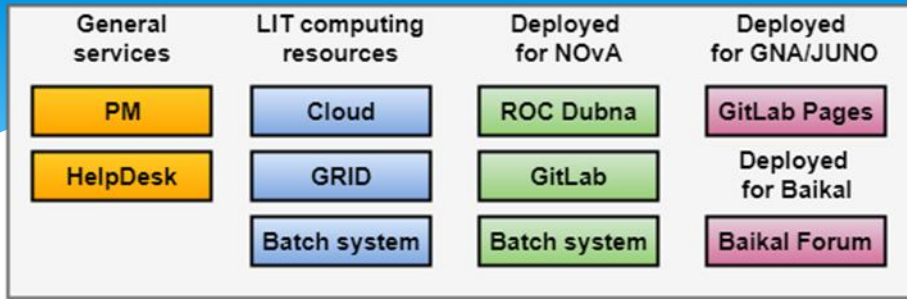
Geant4  
A simulation toolkit



Events reconstruction

- Modeling:
  - physical events
  - receiving data from detectors.
- Data storage organization.

# CSNP Group

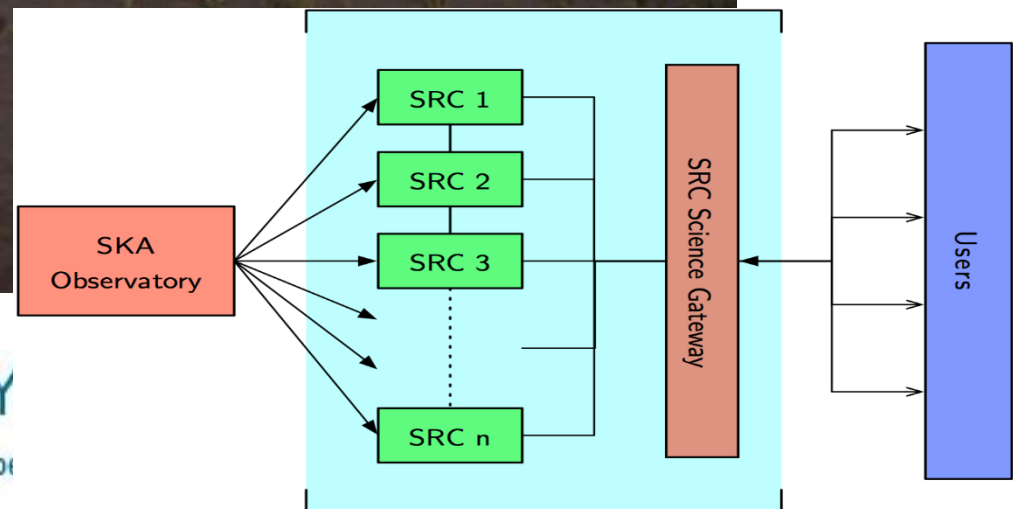
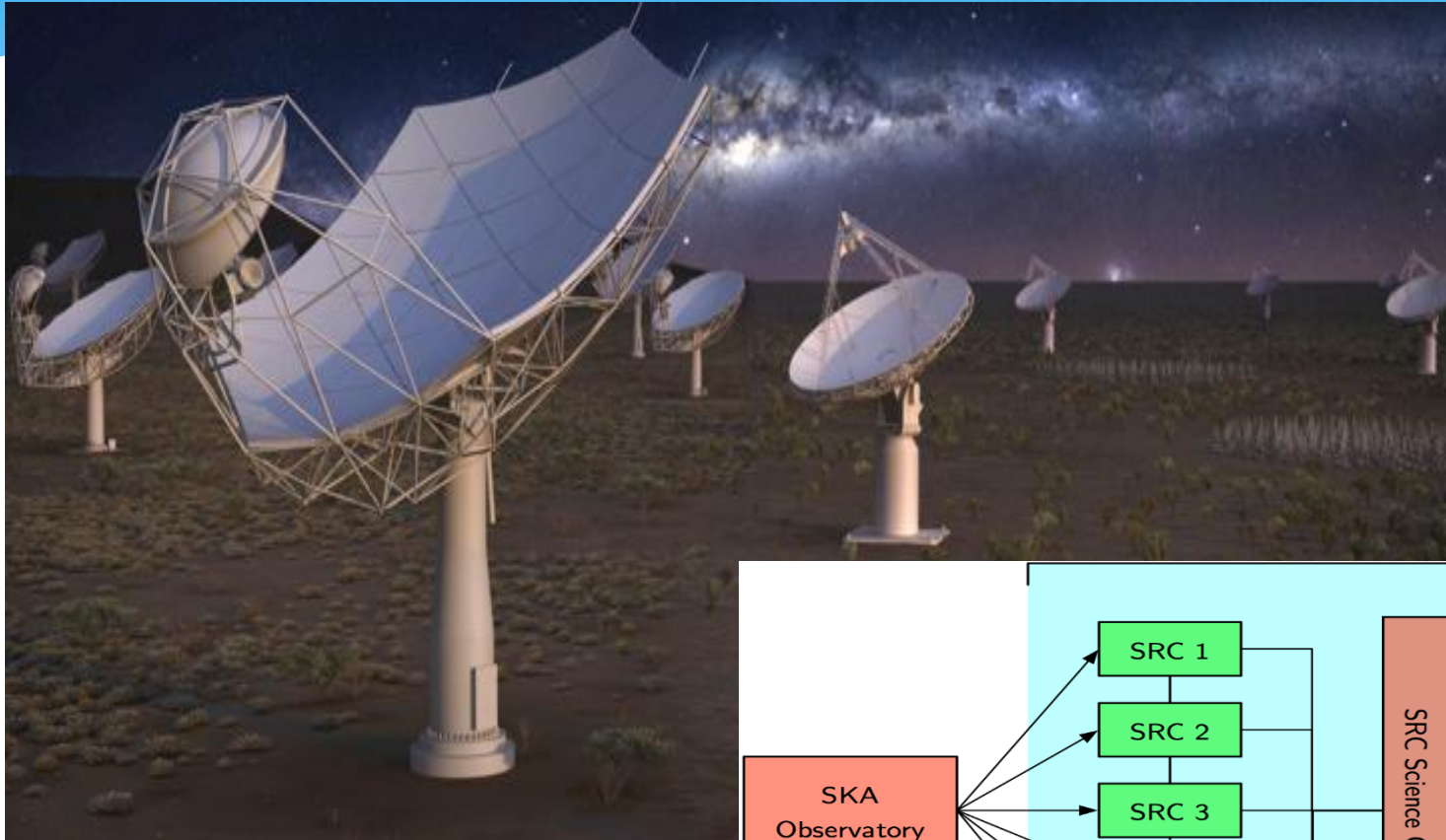


- Supports a set of common IT services
- Supports a number of computing infrastructures
- Provides trainings
- Helps physicists organize their computing
- Deploys and supports additional IT services on demand.
- Helps with computer hardware procurement



# Square Kilometre Array

## SKA Regional Centers



# CHALLENGE: R&D of software to acquire, manage, process, and analyse the big amounts of data to be recorded



## Physics Generators



Detector Simulation

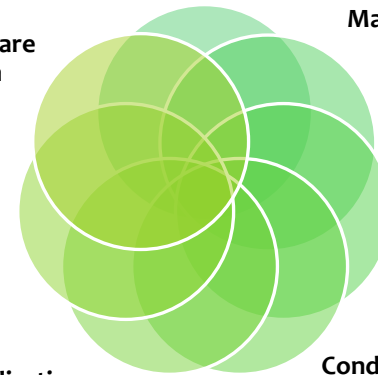
Software Trigger and Event Reconstruction

Machine Learning

Security

Facilities and Distributed Computing

## Data Analysis and Interpretation



Data Organisation, Management and Access

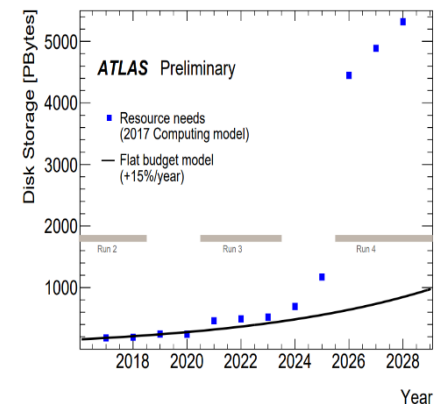
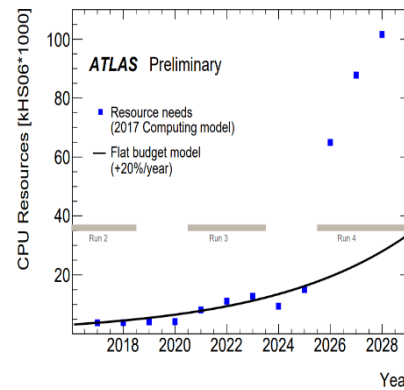
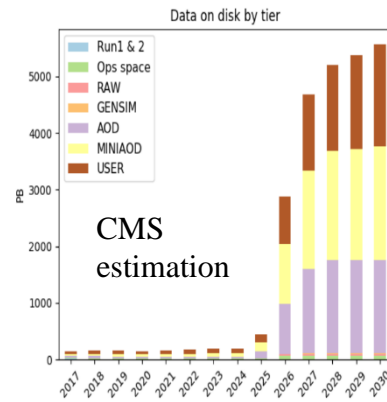
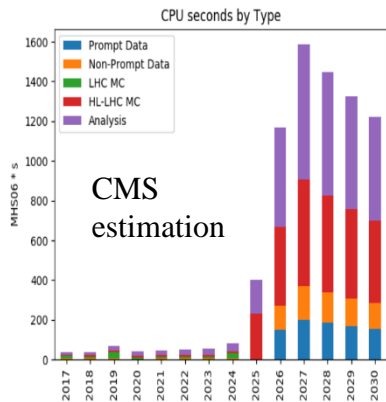
Data-Flow Processing Framework

Conditions Data

Data and Software Preservation

Software Development, Deployment, Validation and Verification

Visualisation



# CHALLENGES: distributed data storage evolution: DATALAKES

## GOAL:

- to provide a computing infrastructure to the experiments and the community to store and analyze data,
- to achieve storage consolidation where geographically distributed storage centers (potentially deploying different storage technologies) are operated and accessed as a single entity.



EOS - a CERN open-source storage software solution to manage multi PB storage.

XRootD - core of the implementation framework providing a feature-rich remote access protocol.

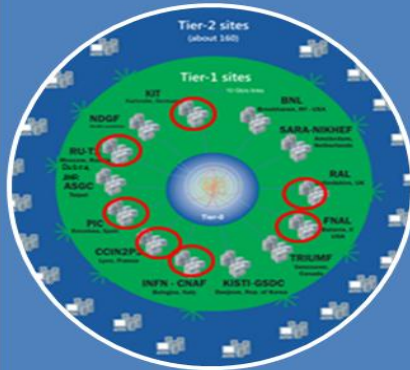


Improvement of already existing production quality Data Management services.

Scalable technologies for federating storage resources and managing data in highly distributed computing environments.

# Multifunctional Information and Computing Complex

## Main components



Grid-Tier1:  
10688 cores  
8.3 PB disk  
11 PB tape



Grid-Tier2  
CICC:  
4128 cores  
2.7 PB disk



Cloud:  
1572 CPU  
8.142 TB RAM  
1.1 PB disk



HPC Govorun  
Peak ~0.5 Pflops  
HybriLIT:  
~70 Tflops

**NICA** off-line cluster and storage system

**DATALAKE**

3740TB

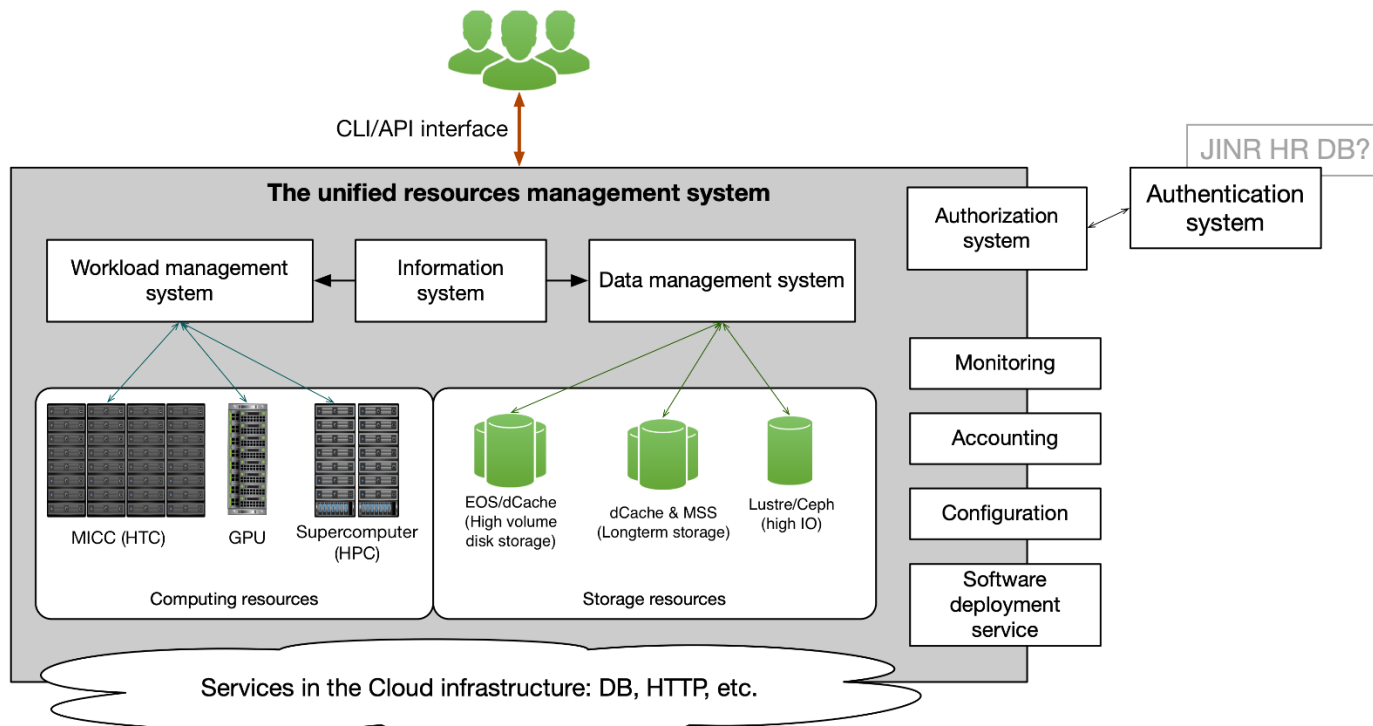
Network infrastructure: LAN: 100 Gbps WAN: 2x100 Gbps

Engineering infrastructure

# Унифицированная система управления ресурсами МИВК

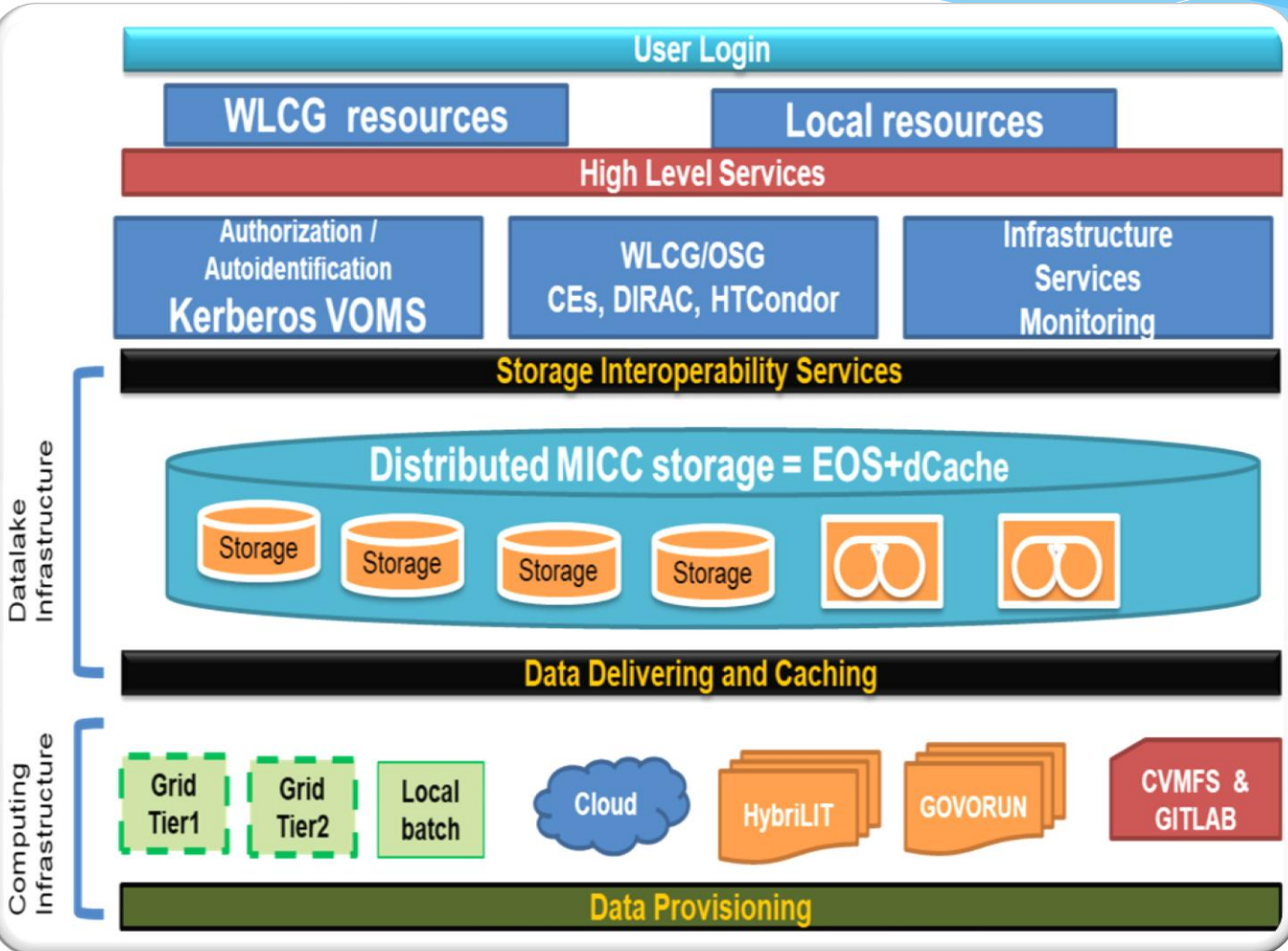
Основными назначениями унифицированной системы управления ресурсами являются:

- предоставление возможности обработки больших объемов данных;
- обеспечение возможности организации массивных вычислительных задач;
- оптимизация эффективности использования вычислительных ресурсов и ресурсов хранения;
- эффективный мониторинг загрузки ресурсов;
- консолидация учета использования ресурсов;
- обеспечение единого интерфейса доступа к ресурсам.

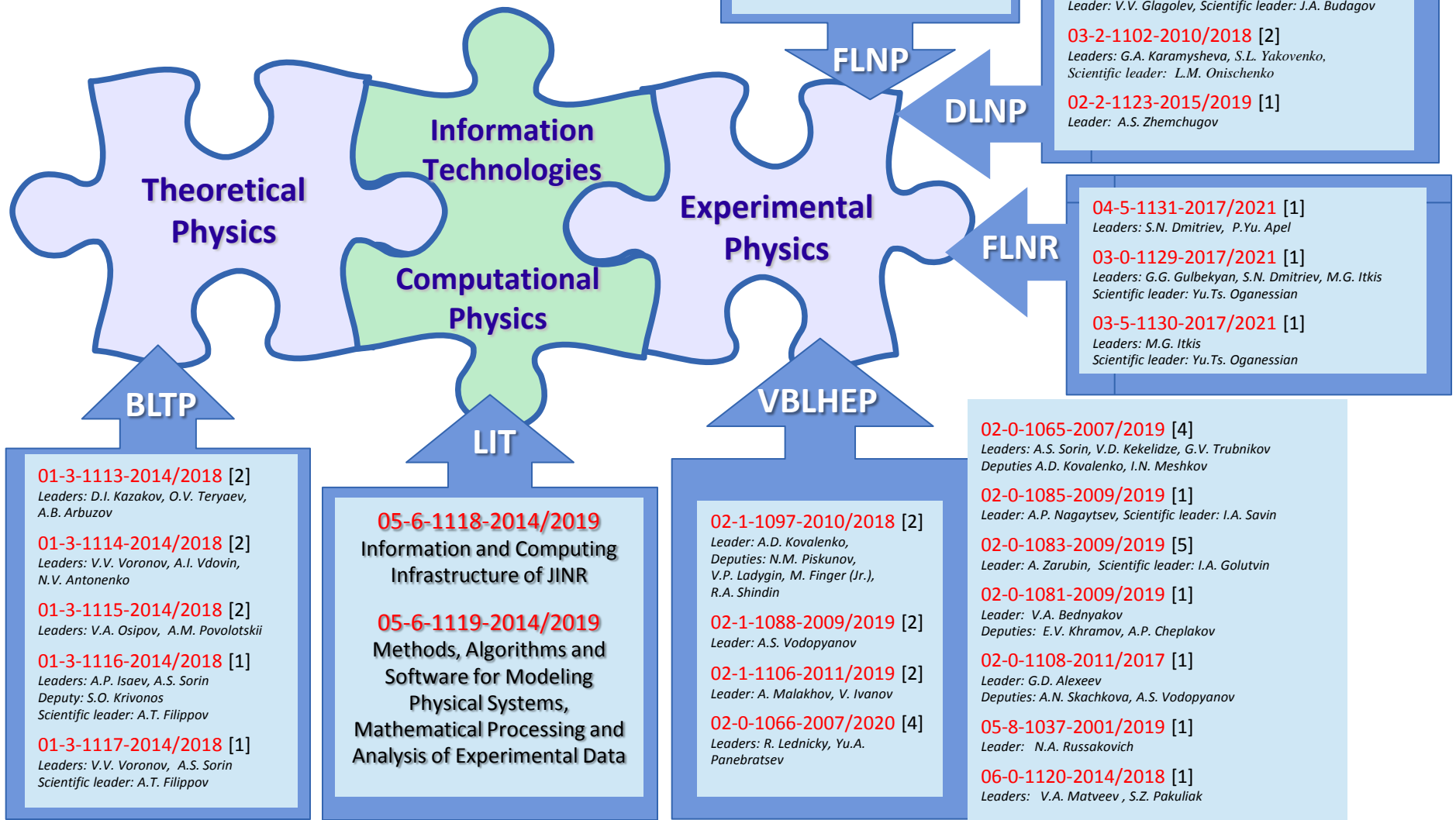


# JINR in Data(Lakes)

We start to develop the common EOS based data storage for MICC components.  
Total space: 3740TB  
User space: 1870TB (2 replicas)



# LIT participates in 48 projects of 30 JINR topics of the 2017 Topical Plan of JINR



04-4-1122-2015/2017 [1]  
Leaders: S.A. Kulikov, V.I. Prikhodko

03-4-1128-2017/2019 [1]  
Leader: V.N. Shvetsov,  
Deputies: Yu.N. Kopatch, E.V. Lychagin,  
P.V. Sedyshev

04-4-1121-2015/2017 [1]  
Leaders: D.P. Kozlenko, V.L. Aksenov,  
A.M. Balagurov

02-2-1099-2010/2018 [1]  
Leaders: D.V. Naumov, A.G. Olshevskiy

02-2-1125-2015/2017 [2]  
Leader: L.G. Tkatchev, Deputy: V.M. Grebenyuk

03-2-1101-2010/2017 [2]  
Leader: A.V. Kulikov, Deputy: Z.Tsamalaidze

02-2-1124-2015/2017 [1]  
Leader: V.V. Glagolev, Scientific leader: J.A. Budagov

03-2-1102-2010/2018 [2]  
Leaders: G.A. Karamysheva, S.L. Yakovenko,  
Scientific leader: L.M. Onischenko

02-2-1123-2015/2019 [1]  
Leader: A.S. Zhemchugov

04-5-1131-2017/2021 [1]  
Leaders: S.N. Dmitriev, P.Yu. Apel

03-0-1129-2017/2021 [1]  
Leaders: G.G. Gulbekyan, S.N. Dmitriev, M.G. Itkis  
Scientific leader: Yu.Ts. Oganessian

03-5-1130-2017/2021 [1]  
Leaders: M.G. Itkis  
Scientific leader: Yu.Ts. Oganessian

01-3-1113-2014/2018 [2]  
Leaders: D.I. Kazakov, O.V. Teryaev,  
A.B. Arbuzov

01-3-1114-2014/2018 [2]  
Leaders: V.V. Voronov, A.I. Vdovin,  
N.V. Antonenko

01-3-1115-2014/2018 [2]  
Leaders: V.A. Osipov, A.M. Povolotskii

01-3-1116-2014/2018 [1]  
Leaders: A.P. Isaev, A.S. Sorin  
Deputy: S.O. Krivonos  
Scientific leader: A.T. Filippov

01-3-1117-2014/2018 [1]  
Leaders: V.V. Voronov, A.S. Sorin  
Scientific leader: A.T. Filippov

05-6-1118-2014/2019  
Information and Computing  
Infrastructure of JINR

05-6-1119-2014/2019  
Methods, Algorithms and  
Software for Modeling  
Physical Systems,  
Mathematical Processing and  
Analysis of Experimental Data

02-1-1097-2010/2018 [2]  
Leader: A.D. Kovalenko,  
Deputies: N.M. Piskunov,  
V.P. Ladygin, M. Finger (Jr.),  
R.A. Shindin

02-1-1088-2009/2019 [2]  
Leader: A.S. Vodopyanov

02-1-1106-2011/2019 [2]  
Leader: A. Malakhov, V. Ivanov

02-0-1066-2007/2020 [4]  
Leaders: R. Lednicky, Yu.A.  
Panebratsev

02-0-1065-2007/2019 [4]  
Leaders: A.S. Sorin, V.D. Kekelidze, G.V. Trubnikov  
Deputies A.D. Kovalenko, I.N. Meshkov

02-0-1085-2009/2019 [1]  
Leader: A.P. Nagaytsev, Scientific leader: I.A. Savin

02-0-1083-2009/2019 [5]  
Leader: A. Zarubin, Scientific leader: I.A. Golutvin

02-0-1081-2009/2019 [1]  
Leader: V.A. Bednyakov  
Deputies: E.V. Khramov, A.P. Cheplakov

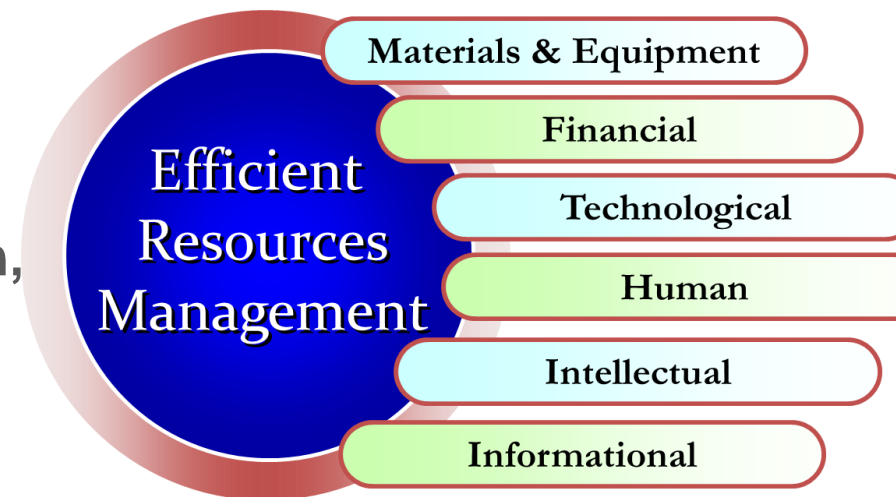
02-0-1108-2011/2017 [1]  
Leader: G.D. Alexeev  
Deputies: A.N. Skachkova, A.S. Vodopyanov

05-8-1037-2001/2019 [1]  
Leader: N.A. Russakovich

06-0-1120-2014/2018 [1]  
Leaders: V.A. Matveev, S.Z. Pakuliak

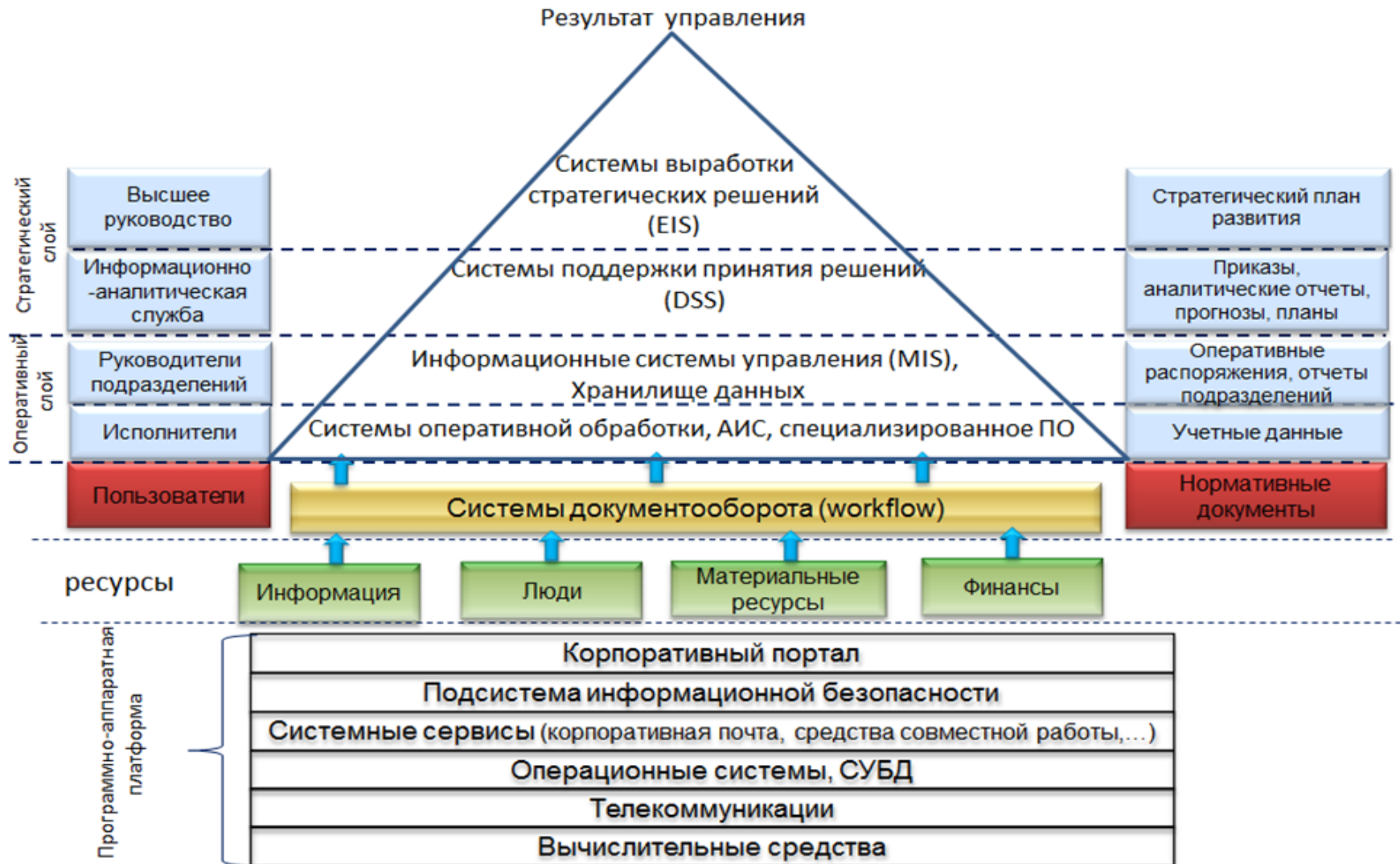
# The JINR corporative information system

- ❑ General 1C:Enterprise platform intended for automation of everyday tasks of economic and management activity,
- ❑ APT EVM system (Activity Planning Tool Earned Value Management) for NICA and future projects management,
- ❑ Electronic document handling system EDH «Dubna»
- ❑ JINR Document Server – electronic open archive-repository of scientific publications and documents,
- ❑ JINR and JINR Member-states access to e-library,
- ❑ PIN – JINR staff personal information,
- ❑ JINR portal





# Структура КИС



# Международная школа по информационным технологиям «Аналитика Больших данных»

Цель Международной школы по информационным технологиям «Аналитика больших данных» – подготовка высококвалифицированных ИТ-специалистов в области Data Science, умеющих формулировать и решать научно-практические задачи с использованием аналитики Больших данных. Программа подготовки будет ориентирована на приобретение глубоких знаний в области математической статистики, машинного обучения, программирования, методов и технологий обработки и анализа данных, понимания бизнес-запросов и задач своей отрасли.

Среди основных направлений подготовки отдельное внимание будет уделено развитию модели компьютинга, программной платформы системы сбора, хранения, обработки и анализа данных экспериментов на установках класса мегасайенс (NICA, PIC, LHC, FAIR, SKA и др.).

# Концепция развития IT-технологий и научного компьютеринга в ОИЯИ (до 2030 г.)

- \* Концепция развития IT-технологий и научного компьютеринга нацелена на обеспечение решения стратегических задач стоящих перед ОИЯИ посредством внедрения и развития целого спектра новейших IT-решений, интегрированных в единую информационно-вычислительную среду — научную IT-экосистему, объединяющую множество различных технологических решений, концепций и методик.

# Научные исследования в развитии высокопроизводительных вычислений и суперкомпьютерных технологий.

## Ключевые задачи

### **ФУНДАМЕНТАЛЬНЫЕ И ПРИКЛАДНЫЕ ИССЛЕДОВАНИЯ:**

- Квантовых суперкомпьютеры и квантовые вычисления
- Фотонные вычислители
- Спецпроцессоры
- программная платформа суперкомпьютерного моделирования, цифровых испытаний и прогнозирования сложных технических систем
- Фундаментальные поисковые исследования в развитии физико-математических моделей и математических методов для экзафлопсных вычислений (суперкомпьютерные двойники)
- Супер-ЭВМ экзафлопсного класса (научно-технические решения, прикладное и системное программное обеспечение и другие)

# Сверхбыстрая система хранения данных (ССХД)

Основными характеристиками ССХД являются:

1. Сверхвысокая производительность системы, обеспечивающая пропускную способность записи и чтения данных в сотни гигабайт в секунду при объемах системы хранения в 1-5 петабайт (больше миллиона файловых операций в секунду).
2. Возможность линейного наращивания производительности (скорости работы с данными) так и объёма хранилища на несколько порядков без изменения принципов архитектурного построения системы.
3. Гиперконвергентность и программно-определяемая архитектура ССХД позволяет обеспечить максимальную гибкость конфигураций системы хранения данных.
4. Реализация ССХД на базе технологии прямого жидкостного охлаждения российской компании РСК позволяет разместить более 900 сверхбыстрых дисков в стандартной монтажной стойке с общим объемом 3,6-7ПБ и скоростью чтения/записи около 1ТБ/с.

# Модель ДОМА

- \* В последнее время тенденция использования и работы с системами хранения данных сместилась от простого подхода, разделяющего организацию, управление, доступ и его иерархичность к модели ДОМА (Data Organisation, Management and Access), комбинирующей все аспекты. Такая модель позволяет создавать более гибкую систему работы с данными, учитывающую как текущие сетевые возможности, в том числе – программную ориентированность сетевых компонентов, так и раскрывающую потенциал оптимизации доступа, размещения и управления данными, поскольку ДОМА учитывает кросс-системные знания и возможности взаимного обмена информацией о процессах, протекающих внутри систем.

# Квантовый компьютер и квантовая информатика

Квантовый компьютер, который развивается в последние годы ускоренными темпами, предлагает новые возможности для обработки данных HL-LHC (ЦЕРН) средствами машинного обучения и, в особенности, глубокого обучения. Скорость роста числа кубитов, характеризующего максимально возможный объем входных данных для квантового компьютера явно выше линейной, что показывает ускоренный прогресс в создании оборудования квантовых компьютеров.

- \* С не меньшей скоростью развивается и программное обеспечение квантовых компьютеров. Созданием и развитием квантовых компьютеров интенсивно занимаются такие ИТ гиганты, как IBM, Google, Microsoft, Intel, каждый из которых разрабатывает свою оригинальную технологию создания оборудования квантового компьютера и оригинальное программное обеспечение, включающее интерфейс пользователя.

# Внедрение нейросетевого подхода, методов и алгоритмов ML/DL

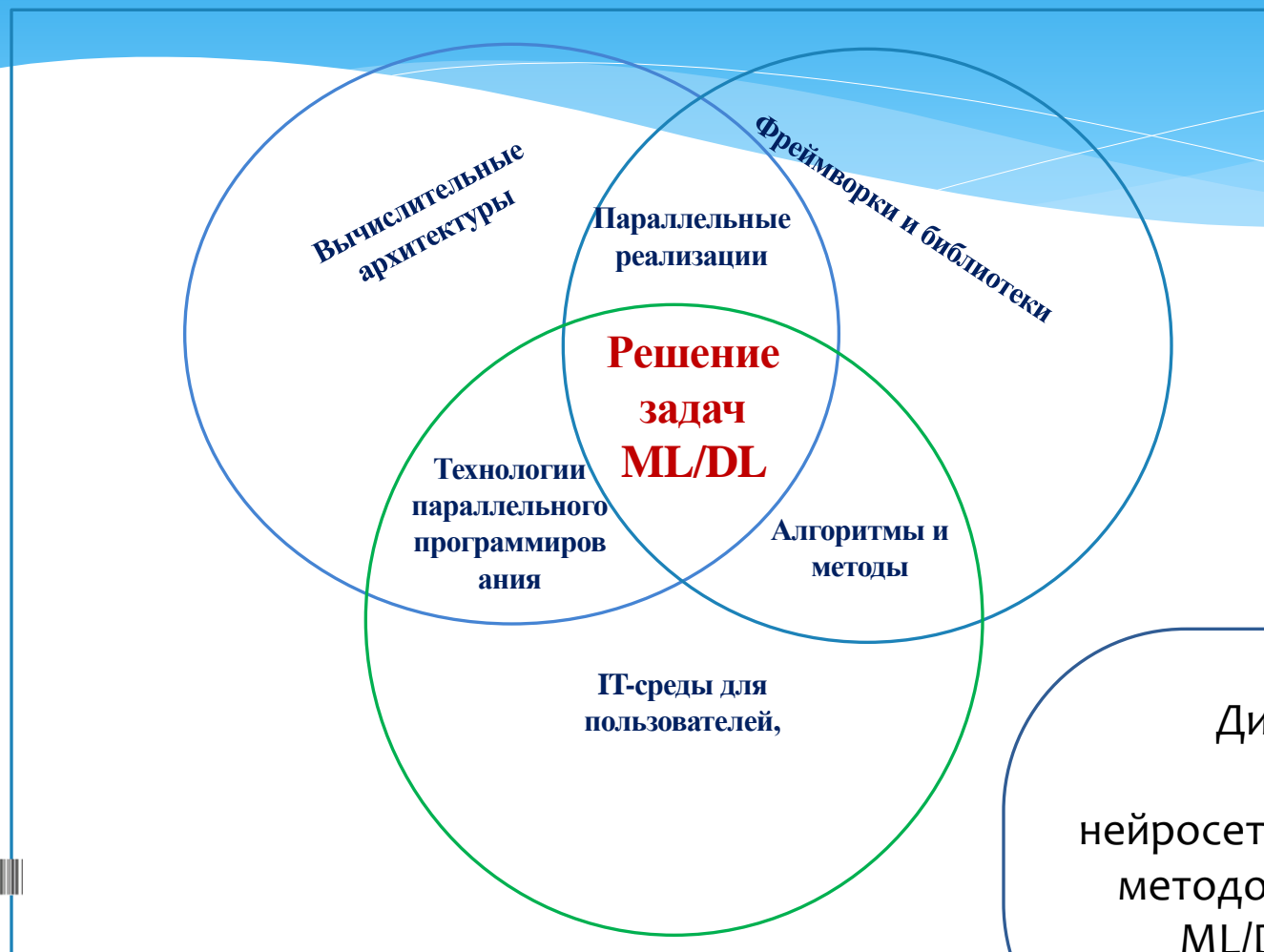


Диаграмма Венна о внедрении нейросетевого подхода, методов и алгоритмов ML/DL для решения прикладных задач.



# Тенденции развития Аналитики Больших данных

## Полный цикл Больших данных. Платформы

62

Y. Zhang et al. / Journal of Cleaner Production 197 (2018) 57–72

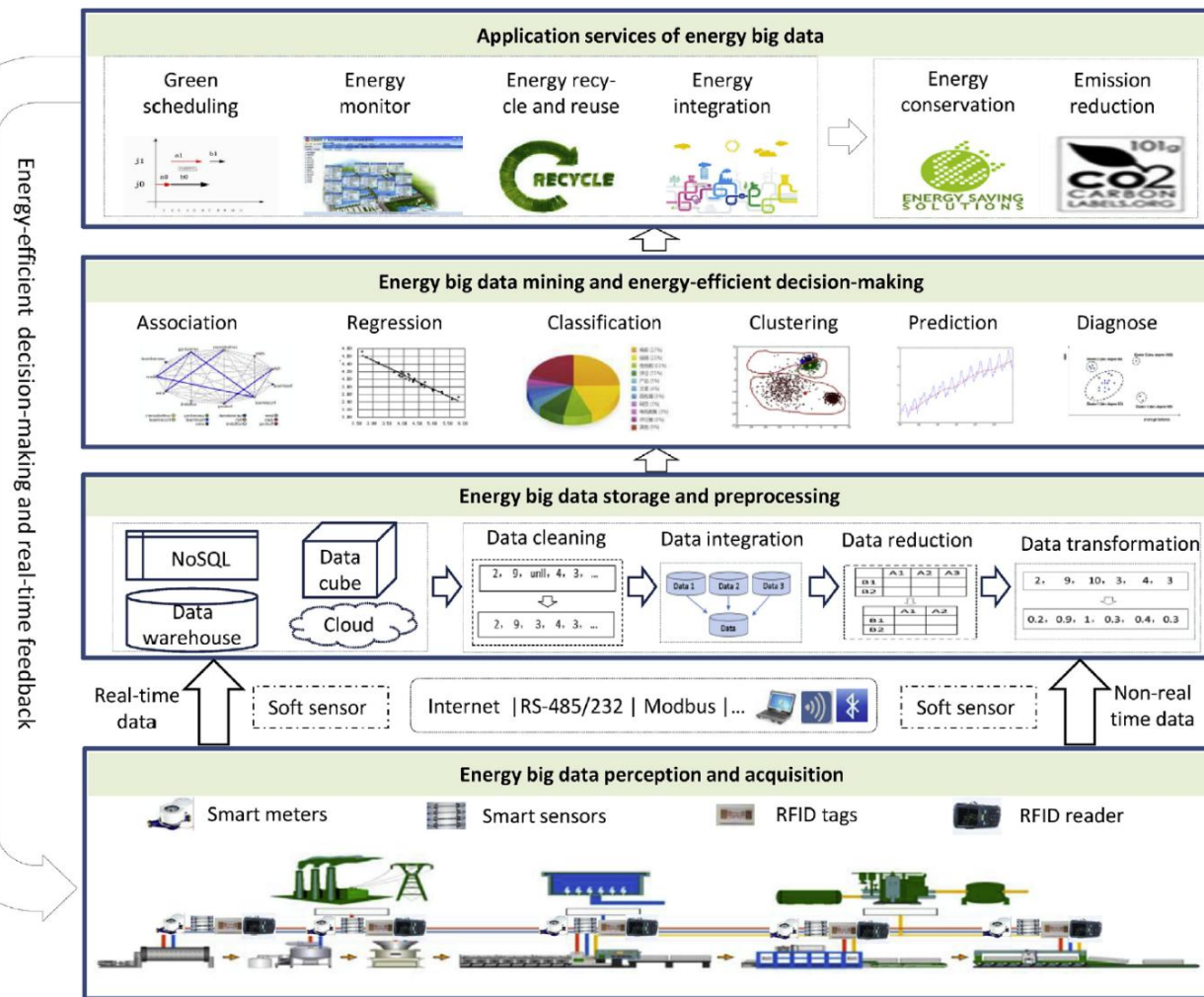


Fig. 3. A big data driven analytical framework for energy-intensive manufacturing industries (based on the example of ceramics industry).

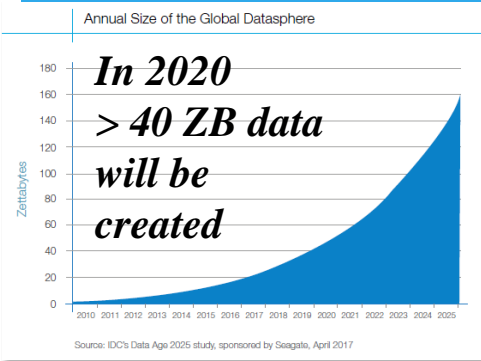
Приложения, сервисы, мониторы, системы принятия решений, ситуационные центры

Анализ – кластеризация, классификация, ассоциация, визуализация, интеллектуальный анализ, моделирование, прогноз.

Первичная обработка, размещение, интеграция, фильтрация, сжатие, обогащение, преобразование

Высокоскоростной автоматизированный сбор и передача разнообразных данных от измерительных приборов, камер наблюдения, умных датчиков, устройств чтения радиочастотных меток и т.д.

# Big Data + HPC (HPDA - High Performance Data Analysis)



*Annual data production follows to exponential law.*

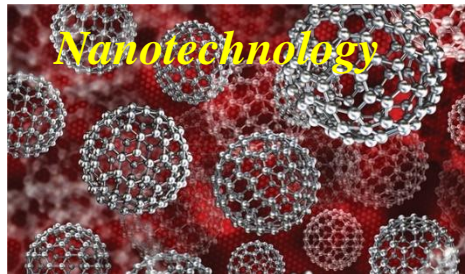
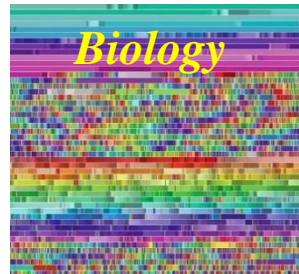
## High Energy Physics



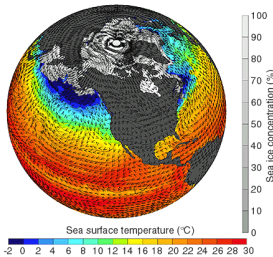
## Science

**CERN Large Hadron Collider**  
**> 20 Pb/Year, > 200 Pb stored**

## Astrophysics



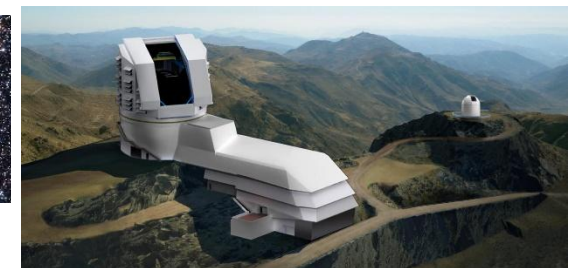
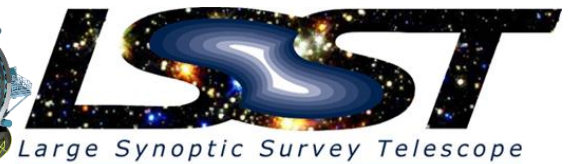
## Climate



**Square Kilometer Array radio telescope (SKA) > 20 Pb/Day (estimation)**



*An International radiotelescope for the 21st century*



**Large Synoptic Survey Telescope (LSST) > 10 Pb/Year (estimation)**

*...et cetera*

# Большие данные и искусственный интеллект: текущее состояние

- \* Реконструкция и анализ событий в ФВЭ, нейтринной, ядерной физике
- \* Подходы к решению проблемы размерности в анализе и моделировании квантовомеханических систем
- \* Мониторинг качества данных
- \* Анализ работы распределенной компьютерной инфраструктуры
- \* Компьютерная безопасность

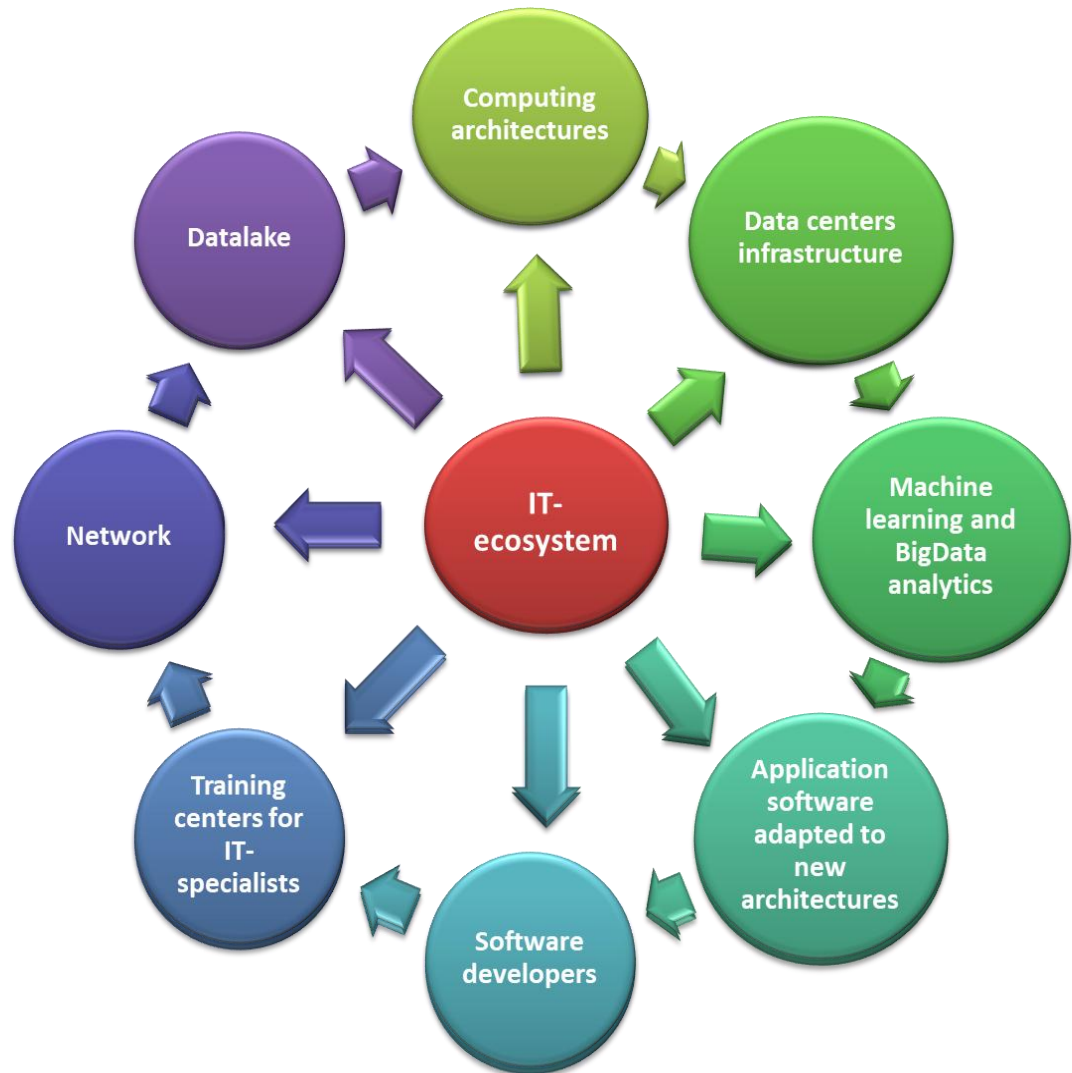
# Большие данные + искусственный интеллект: перспективы

- \* Мониторинг и управление физическими установками
- \* Широкое использование специализированных программируемых микросхем для задач анализа данных и управления
- \* Сжатие и онлайн-фильтрация данных установок мегасайнс
- \* Квантовые и квантовоподобные алгоритмы для решения широкого класса задач (управление, анализ, моделирование)
- \* Перспективные исследования в области биофизики, медицинской физики и генетики
- \* Анализ, валидация и оптимизация научных программных продуктов и систем
- \* Открытые научные данные

# Strategy of Information Technologies and Scientific Computing in JINR

## CONCEPT

of the development of IT-technologies and scientific computing aimed at solving strategic tasks of JINR through the introduction and development of a whole range of advanced IT solutions, integrated into a unified computing environment – **scientific IT ecosystem that combines a variety of technological solutions, concepts and methodologies.**



# Характеристики IT-экосистемы

\* Развиваемая IT-экосистема должна предоставить доступ к информационному пространству, включающему в себя сервисы для организации эффективной работы сотрудников ОИЯИ и программно-аппаратную среду для решения задач различной сложности и тематики, управления и обработки данных различных объемов и структур, обучения и организации научных и исследовательских процессов.

\* Для обеспечения развития этой IT-экосистемы необходимо реализовать информационно-коммуникационное обеспечение, базирующееся на новейших технологических сетевых решениях, обеспечивающих более высокую пропускную способность и надежность функционирования сетей. Постоянно развиваемая сетевая инфраструктура обеспечит возрастающие потребности в эффективной и быстрой обработке и хранении данных, получаемых как с БАК, так и в будущих экспериментальных проектах, реализуемых как в ОИЯИ, так и в мире.

# Характеристики IT-экосистемы

Ключевой основой научной IT-экосистемы является распределенная программно-конфигурируемая HPC-платформа, объединяющая суперкомпьютерные (гетерогенные), грид- и облачные технологии, с целью предоставления оптимальных подходов для решения различных типов научных и прикладных задач, требующих как массивно-параллельных вычислений, так и привлечения методов и технологий Больших данных, данная платформа позволит наиболее эффективно использовать новейшие вычислительные архитектуры.

\* Развитие научной IT-экосистемы требует использование новейших технологий сбора и анализа данных, обмена ими, привлечение новых методов вычислений, такие как квантовые, когнитивные вычисления, методы машинного обучения и интеллектуального анализа данных, а также разработки новой алгоритмической базы.

\* Данная IT-экосистема должна является базовой платформой для обучения IT-специалистов, способных разрабатывать алгоритмические и программные решения в области системного и прикладного программирования, математических, информационных и имитационных моделей, создавать информационные ресурсы глобальных сетей, образовательный контент, прикладные базы данных, решать задачи профессиональной деятельности на основе информационной культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности.

# Стратегическими направлениями являются исследования в области:

## \* телекоммуникационных технологий:

- \* сети с терабитными скоростями передачи данных;
- \* сети, основанные на новых принципах организации, включая когнитивные, гибридные, адаптивные, реконфигурируемые и гетерогенные (программно-конфигурированные сети и т.д.);
- \* сетевые системы с гарантированным динамическим выделением ресурса;
- \* исследовательские системы нового поколения, позволяющие передавать большие объемы данных
- \* сети, позволяющие, выполняя распределенную обработку экзабайтных объемов данных,



# Стратегическими направлениями являются исследования в области:

## \* ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ:

- \* распределенная программно-конфигурируемая HPC-платформа, объединяющая суперкомпьютерные (гетерогенные), грид- и облачные- технологии, с целью наиболее эффективного использовать новейшие вычислительные архитектуры.

## \* алгоритмов и программного обеспечения для:

- \* распределенного решения отдельных классов сложных вычислительных задач,
- \* машинного обучения,
- \* формализации и извлечения знаний из плохо структурированных и неструктурированных данных
- \* вычисления на основе квантового формализма

# Стратегическими направлениями являются исследования в области:

- \* **технологий обработки и анализа данных:**
- \* сервисы для распределенных и параллельных вычислений (metacomputing) для повышения эффективности научных исследований посредством использования суперкомпьютеров;
- \* разработка масштабируемых алгоритмов и программ обработки многопараметрических, многомерных, иерархических и многосерийных наборов данных экзабайтного объема
- \* системы аналитики нового поколения, основанные на эффективных методах и алгоритмах формализации и извлечения знаний и обработки больших данных
- \* прогнозное моделирование компьютерных моделей мегаэкспериментов и функционирования перспективных систем

# Стратегическими направлениями являются исследования в области:

## \* информационной безопасности:

- \* защита компьютерных инфраструктурных систем на основе принципиально новых парадигм, включая квантовую криптографию и вычисления, нейрокогнитивные принципы;
- \* использование перспективных средств и программных систем защиты данных с учетом новых принципов организации данных и взаимодействие объектов данных, в том числе глобальной интеграции информационных систем, повсеместного доступа к приложениям, новых протоколов Интернета, виртуализации, социальных сетей, данных мобильных устройств и геолокации.