
# CERN: ґрід та інформаційні технології

Свірін Павло
ALICE, CERN
Інститут теоретичної фізики ім. Боголюбова НАН України
Національний технічний університет України “КПІ”

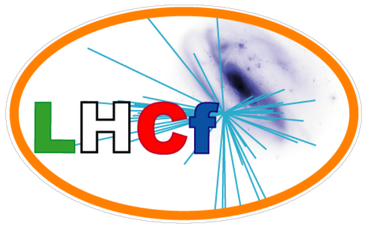
# Великі експерименти CERN



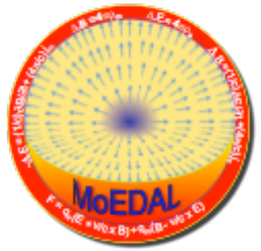


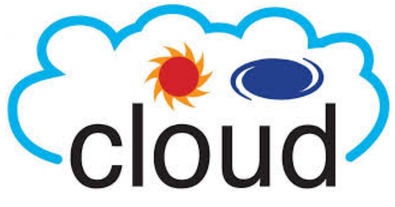
Найбільші експерименти CERN, які генерують переважну кількість даних.

В колаборацiях - близько 10 тисяч чоловік











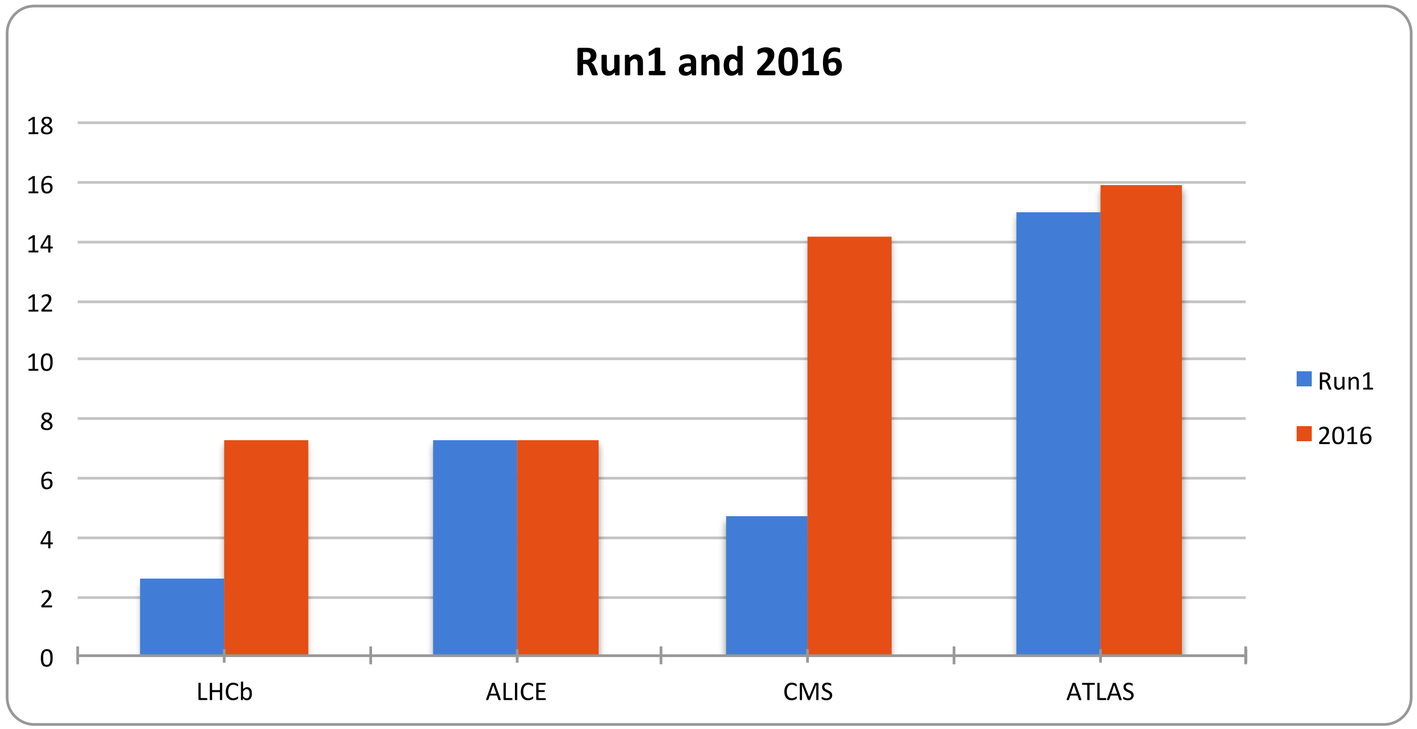


# CERN: малі експерименти

- 21 малий експеримент
- Невелика кількість вчених у колабораціях (по декілька сотень вчених)
- Не всі знаходяться на LHC, тому деякі з них генерують дані щороку
- Генерують відносно незначні об’єми даних

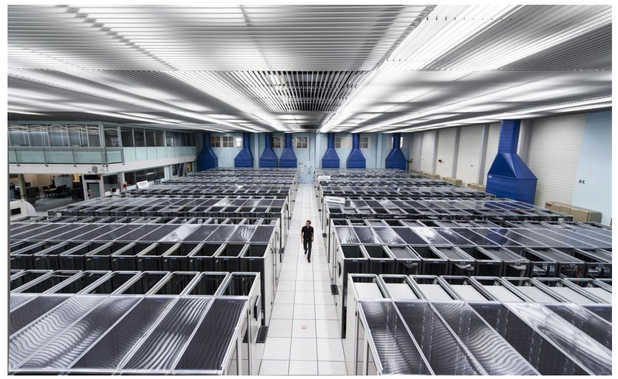
# Об’єм даних на експерименті ALICE

- Об’єм даних, що отримано лише 2015 року дорівнює об’єму даних за перший сезон роботи LHC
- Протягом 2016 року було отримано 45 Пб даних (9 млн. DVD)
- На наступний запуск (2020-2022 pp) прогнозується 25-кратне зростання об’єму отриманих даних



# Датацентр CERN

- Забезпечує функціонування основних сервісів (електронна пошта, відеоконференції, керування даними)
- 10 тис. серверів, загалом 110 тис. процесорів
- Щодня обробляє близько 1 Пб інформації (бл. 210,000 DVD)
- Для обміну даними використовується швидкісний оптичний кабель загальною довжиною 35 тис. км.

# Датацентр Wigner

- Відкрито у червні 2013 року в Будапешті, Угорщина
- Надає близько 30% потужностей порівняно з основним датацентром, до 2020 року планується довести потужність до рівня аналогічного до CERN Datacentre.
- Об’єднано з основним датацентром мережевим каналом 100 Гб/сек (дозволяє передавати на секунду об’єм даних, аналогічний 5 DVD)


# Ґрід

# Ґрід-обчислення

Ґрід є формою розподілених обчислень, в якому багато комп'ютерів об'єднані в один потужний віртуальний комп'ютер, і які працюють разом для виконання трудомістких завдань. Для певних додатків, «ґрід» обчислення можна розлядати як спеціальний тип паралельних обчислень які покладаються на цілі комп'ютери(обладнані процесорами, пам'ятю, живленням, мережевим інтерфейсом і тд.), під'єднані до комп'ютерної мережі(приватної або публічної) звичайним мережевим інтерфейсом, таким як Ethernet.

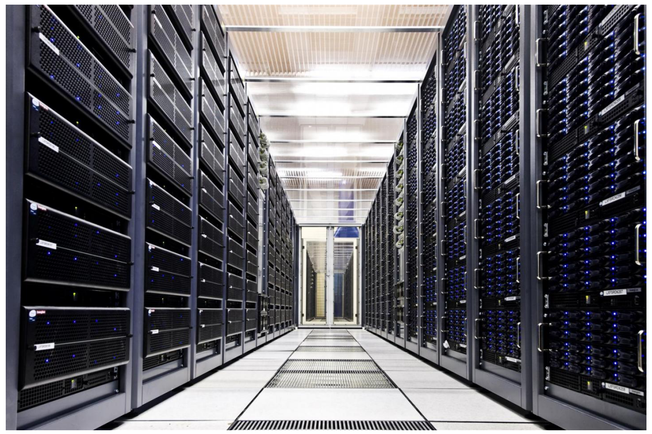
Термін з'явився на початку 1990-х років, як метафора, що демонструє можливість простого доступу до обчислювальних ресурсів як і до електричної мережі (англ. Power grid)

# WLCG: The Worldwide LHC Computing Grid

До 2006 року - LCG (LHC Computing Grid).
Міжнародна колаборація, яка підтримує ґрід-інфраструктуру зі 170 обчислювальних центрів у 36 країнах (дані 2012 року). Інфраструктура розроблена спеціально для обробки даних з Великого адронного колайдера.

Найбільший в світі обчислювальний ґрід.
Складається з інфраструктур:

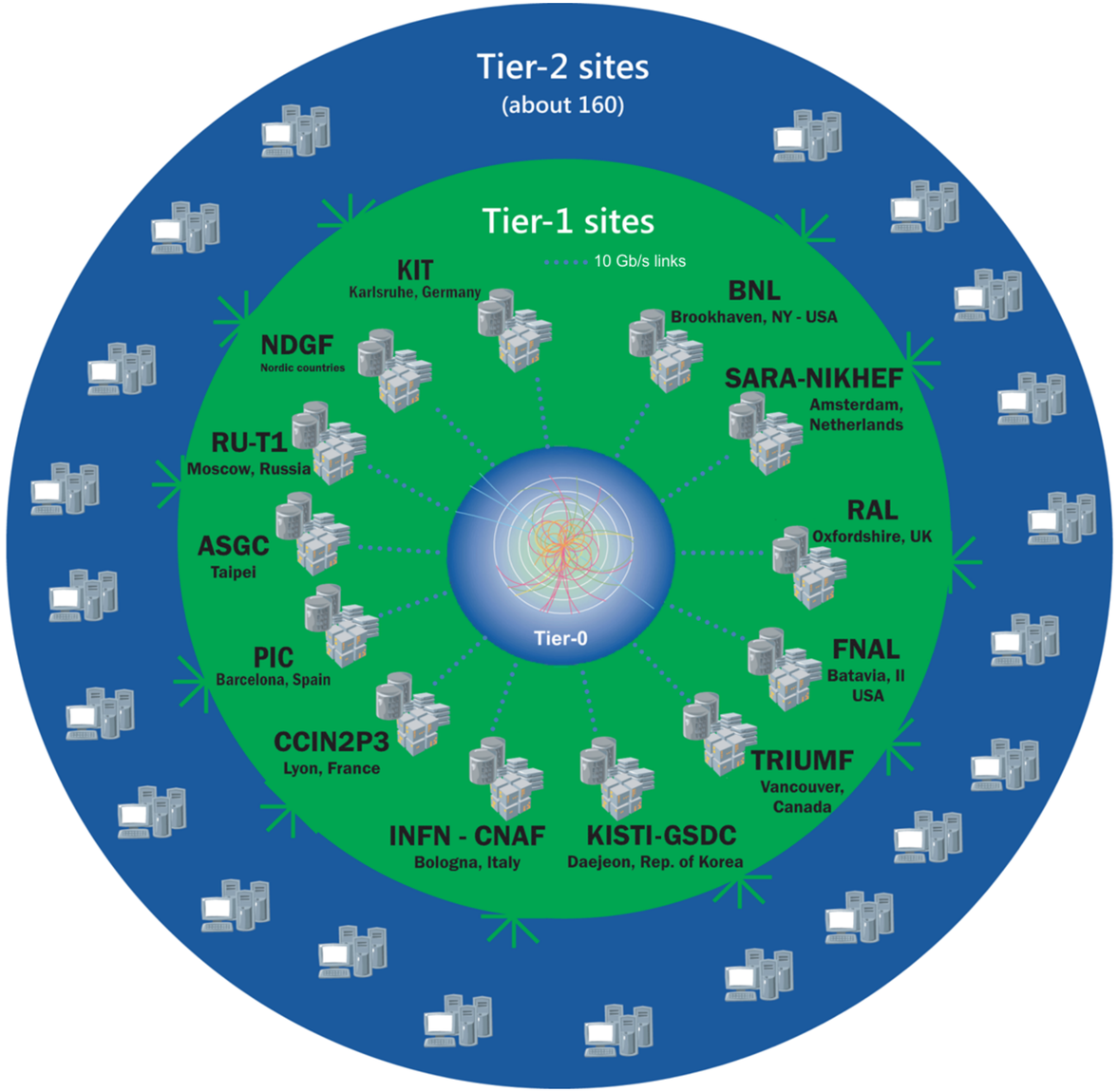
- European Grid Infrastructure
- Open Science Grid (США)
- TWGrid (Тайвань)
- EU-IndiaGrid (ґрід-інфраструктури Європи та Азії)
- NorduGrid (скандинавські країни)



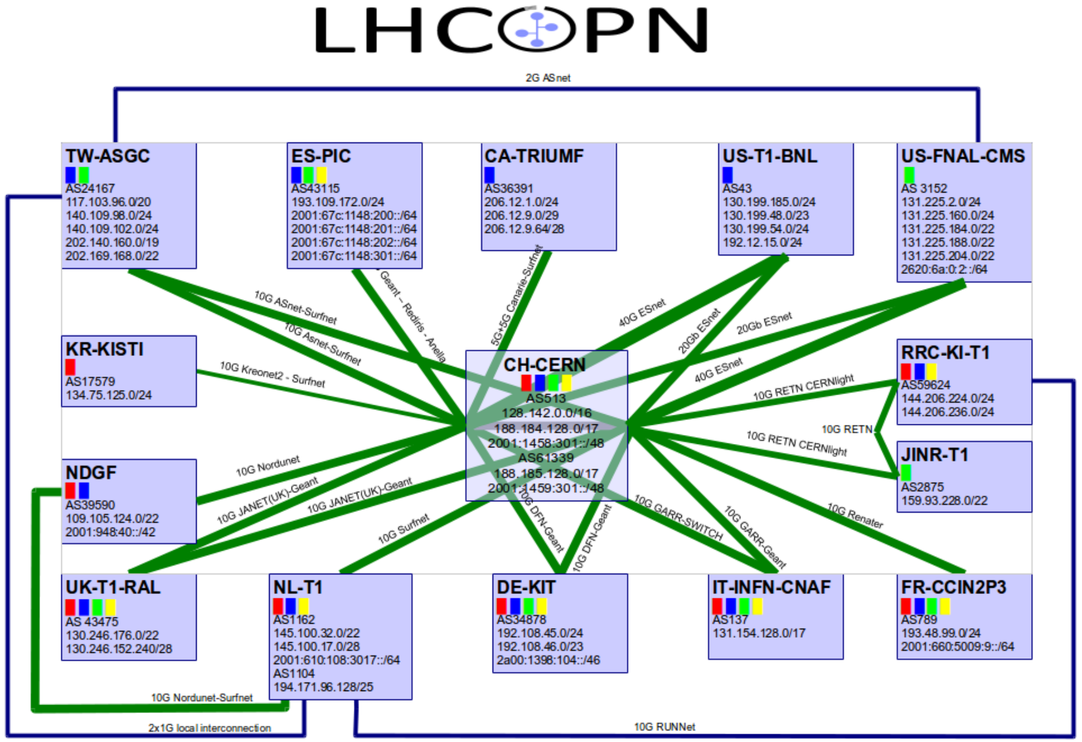
Надає єдиний доступ до обчислювальних та дискових ресурсів, інструментарію візуалізації, тощо. Розробляє вимоги до роботи ресурсів, аутентифікації користувачів.

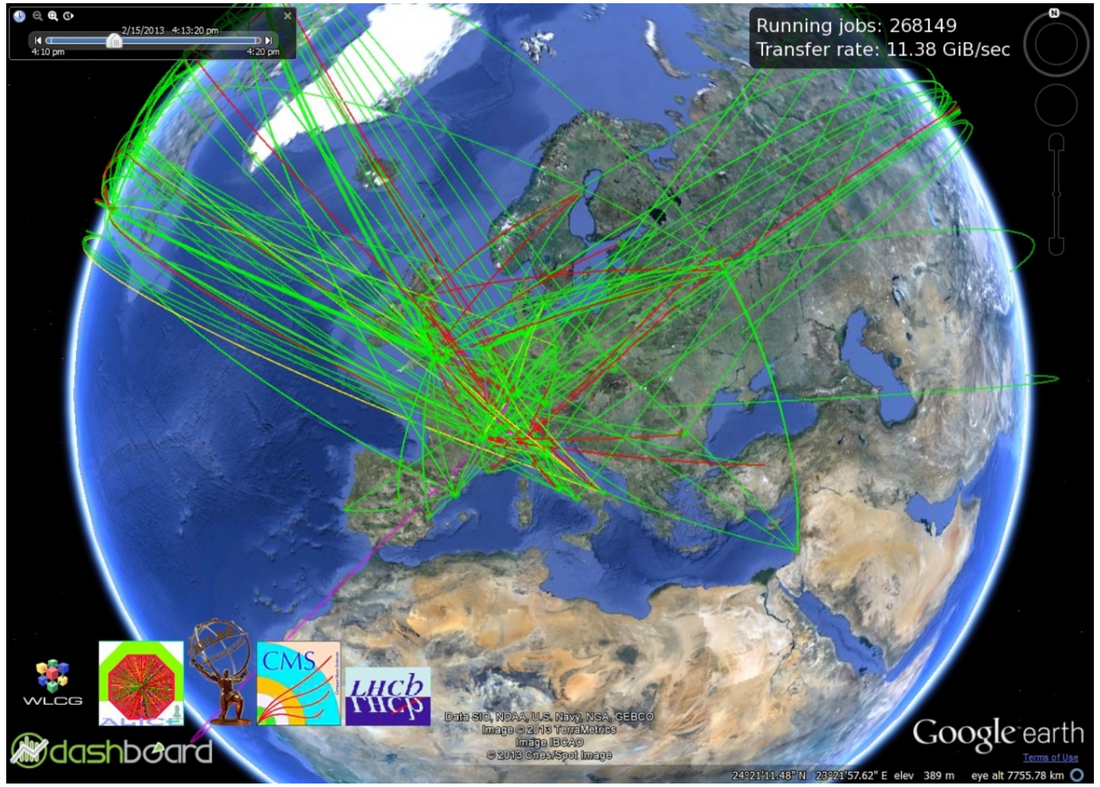
# TIER-центри

- Tier-0: знаходяться в CERN та в датацентрі Wigner. Відповідають за збереження "сирих даних" (перша копія даних), перший прохід реконструкції та взаємодію з Tier-1. В періоди простою ВАК беруть участь у загальній оброботці даних.
- Tier 1: великі комп’ютерні центри з відповідними обчислювальними можливостями, також зберігають великі об’єми даних. Відповідають за взаємодію з обчислювальними ресурсами Tier-2.
- Tier 2: університети чи наукові інститути, що збрерігають достатньо інформації та надають обчислювальні потужності для виконання необхідних задач з аналізу даних. (близько 160)
- Tier 3: окремі комп’ютери чи локальні кластери.



# Обмін даними між Tier1-Tier2



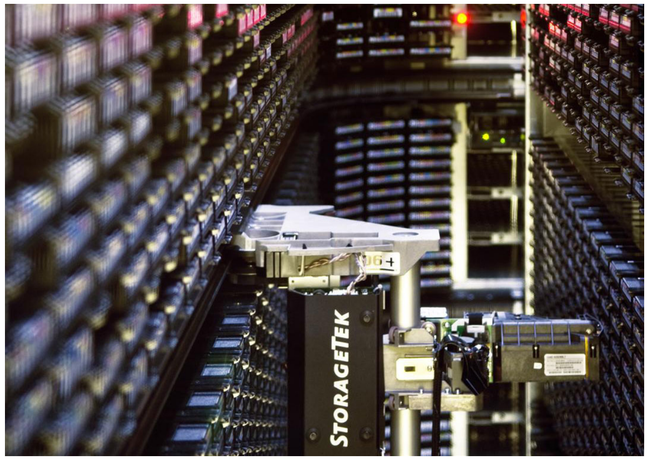

**OUTDATED!!!**

# Системи зберігання даних

Для зберігання даних використовуються наступні системи:

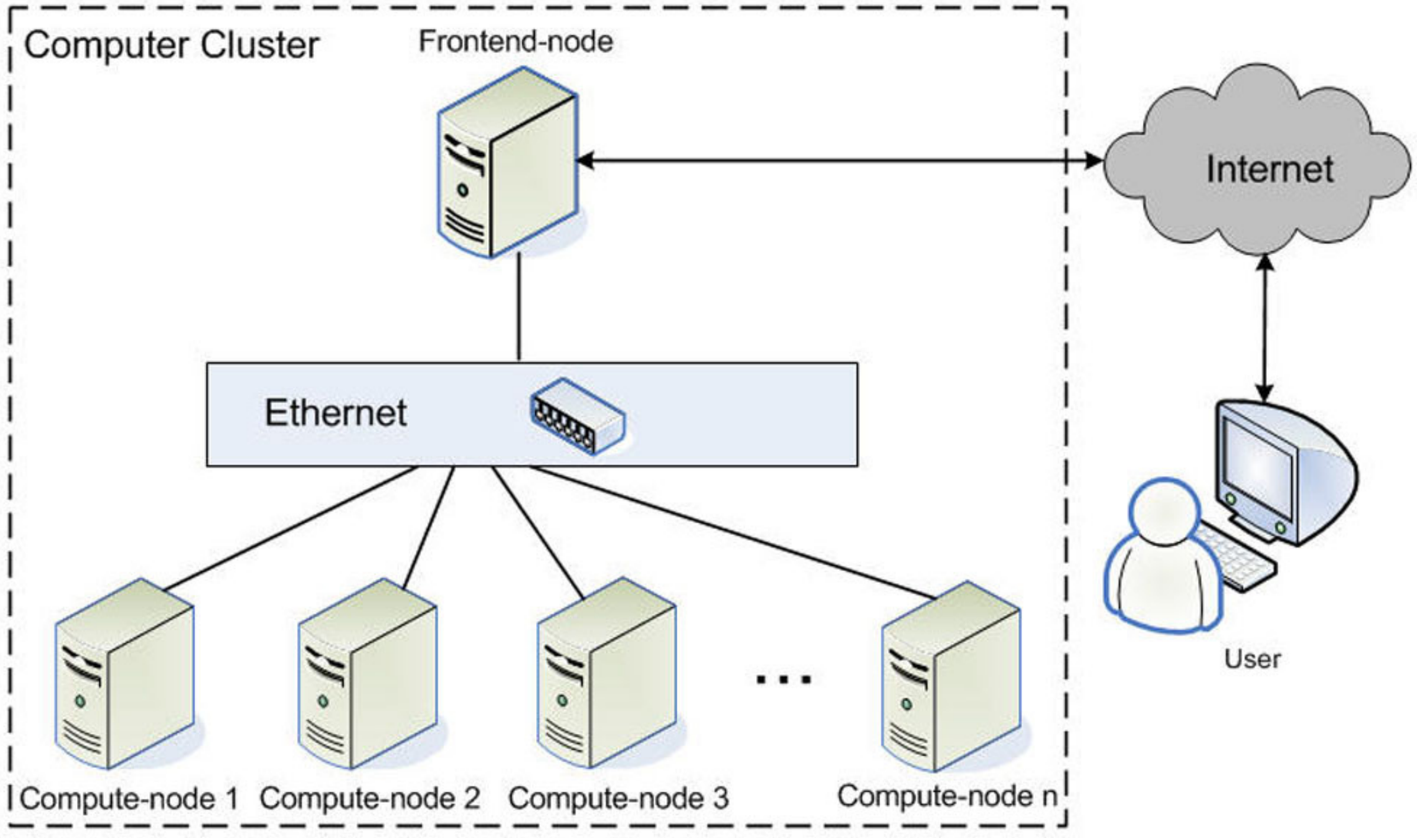
- EOS - на жорстких дисках, працює за протоколом XROOTD
- CASTOR - використовуються плівкові носії. Переваги: економія електроенергії, дешевизна. Недоліки: швидкість доступу.

Загальний об’єм плівкових накопичувачів у CERN - близько 100 Пб (на січень 2013 року)


# Обчислювальні кластери

- Користувачі подають завдання на виконання використовуючи командний або графічний інтерфейс
- Головний вузел (frontend node) оцінює та контролює завантаження кластера, запускає задачі із врахуванням пріоритету користувача
- Може групувати ресурси відповідно до певних вимог



# Обчислювальні кластери

- Велика кількість систем керування завданнями  на кластерах
- Кожна - зі своїм інтерфейсом
- Як їх об’єднати в одну обчислювальну систему?





# Middleware


globus toolkit


“Проміжне” програмне забезпечення:

- об’єднує різнорідні ресурси в єдину мережу, надає ним однаковий інтерфейс
- проводить аутентифікацію користувачів, визначає квоти для користувачів
- передає виконання завдання до кластера, який воно представляє
- контролює виконання завдань
- у деяких випадках може надавати інтерфейс для групи кластерів


Lightweight Middleware for Grid Computing

HTCondor

High Throughput Computing

# Condor та DreamWorks


- використовує кластер під управлінням Condor для рендерингу 3D графіки
- більше 65 мільйонів процесорних годин на виробництво одного фільму
- пікове завантаження: 15,000 ядер з 22,000 доступних
- обробляє під час рендерингу фільму більше за 500 млн файлів (200 Тб)

# Безпека в Ґрід

Існують центри сертифікації (Certificate Authorities, CA), які видають цифрові сертифікати за стандартом X509 для користувачів та обчислювальних ресурсів. Ґрід-сертифікат є цифровою карткою, яка ідентифікує користувача (чи хост) і надає інформацію про CA, який його видав.

“Картка” складається з приватного ключа і сертифіката - “публічного ключа”. Ці складові використовуються для асиметричного шифрування трафіку.

Також використовуються так звані “проксі-сертифікати” - тимчасові перепустки у Ґрід. Звичайно “виписуються” на 1 добу.

# Суперкомп’ютери

Комп’ютери з великими обчислювальними потужностями порівняно зі звичайними. Швидкодія розраховується у кількості операцій над числами з плаваючою точкою на секунду (FLOPS).

Вперше побудовані у 1960-х роках компанією Cray Research.

З 1990-х років суперкомп’ютери використовують тисячі процесорів, в 2000-х - вже десятки тисяч.

Використовуються для розрахунків прогнозів погоди, молекулярної динаміки, ядерних реакцій, комп’ютерної безпеки.

# Рейтинг суперкомп’ютерів “Top 500”

Публікується двічи на рік (червень, листопад).

Суперкомп’ютери оцінюються за швидкістю розв’язування великих систем лінійних алгебраїчних рівнять.

Найбільше суперкомп’ютерів з TOP500 у КНР (167), США (165), Японії (29), Німеччині (26), Франції (18), Великобританії (12).

| Rank | Site | Cores | Rmax (TFlop/s) | Rpeak (TFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | Sunway TaihuLight: National Supercomputing Center in Wuxi (China) | 10,649,600 | 93,014.6 | 125,435.9 | 15,371 |
| 2 | TianHe: National Super Computer Center in Guangzhou (China) | 3,120,000 | 33,862.7 | 54,902.4 | 17,808 |
| 3 | Titan: DOE/SC/Oak Ridge National Laboratory (USA) | 560,640 | 17,590.0 | 27,112.5 | 8,209 |
| 4 | DOE/NNSA/LLNL (USA) | 1,572,864 | 17,173.2 | 20,132.7 | 7,890 |
| 5 | RIKEN Advanced Institute for Computational Science (AICS) (Japan) | 705,024 | 10,510.0 | 11,280.4 | 12,660 |


# ORNL Titan

Суперкомп’ютер в лабораторії Oak Ridge National Lab, США. Відкрито 2012 року. На момент пуску займав першу сходинку в світі.

**6 “критичних кодів”:**

- Молекулярна динаміка (LAMMPS)
- Молекулярна фізика (S3D)
- Моделювання ядерних реакцій (Denovo)
- Глобальні атмофсерні моделювання (CAM-SE)
- Астрофізика (NRDF)
- Термодинаміка (WL-LSMS)

# ORNL Titan

- 2 Гб пам’яті на 1 обчислювальне ядро, більша частина обчислювальної потужності знаходиться в графічних прискорювачах
- Теоретично можливо отримати до 10% ресурсів суперкомп’ютеру
- ALICE та ATLAS беруть участь у експлуатації ORNL Titan в рамках проекту CSC108

| | |
|---|---|
| ***Архітектура*** | ***18,688 AMD Opteron 6274 16-core CPUs, 18,688 Nvidia Tesla K20X GPUs*** |
| ***Операційна система*** | ***Традиційна Linux та Cray Linux Environment (модифікована SuSE Linux 11) на робочих вузлах*** |
| ***Пам’ять*** | ***693.5 TiB (584 TiB CPU та 109.5 TiB GPU)*** |
| ***Дисковий простір*** | ***32 PB, 1.4 TB/s IO Lustre filesystem*** |
| ***Пікова продуктивність*** | ***27.1 PF (18,688 обчислювальних вузлів, 24.5 GPU + 2.6 PF CPU)*** |
| ***Вузли вводу-виводу*** | ***512 service and I/O nodes*** |

# Доступ до ORNL Titan

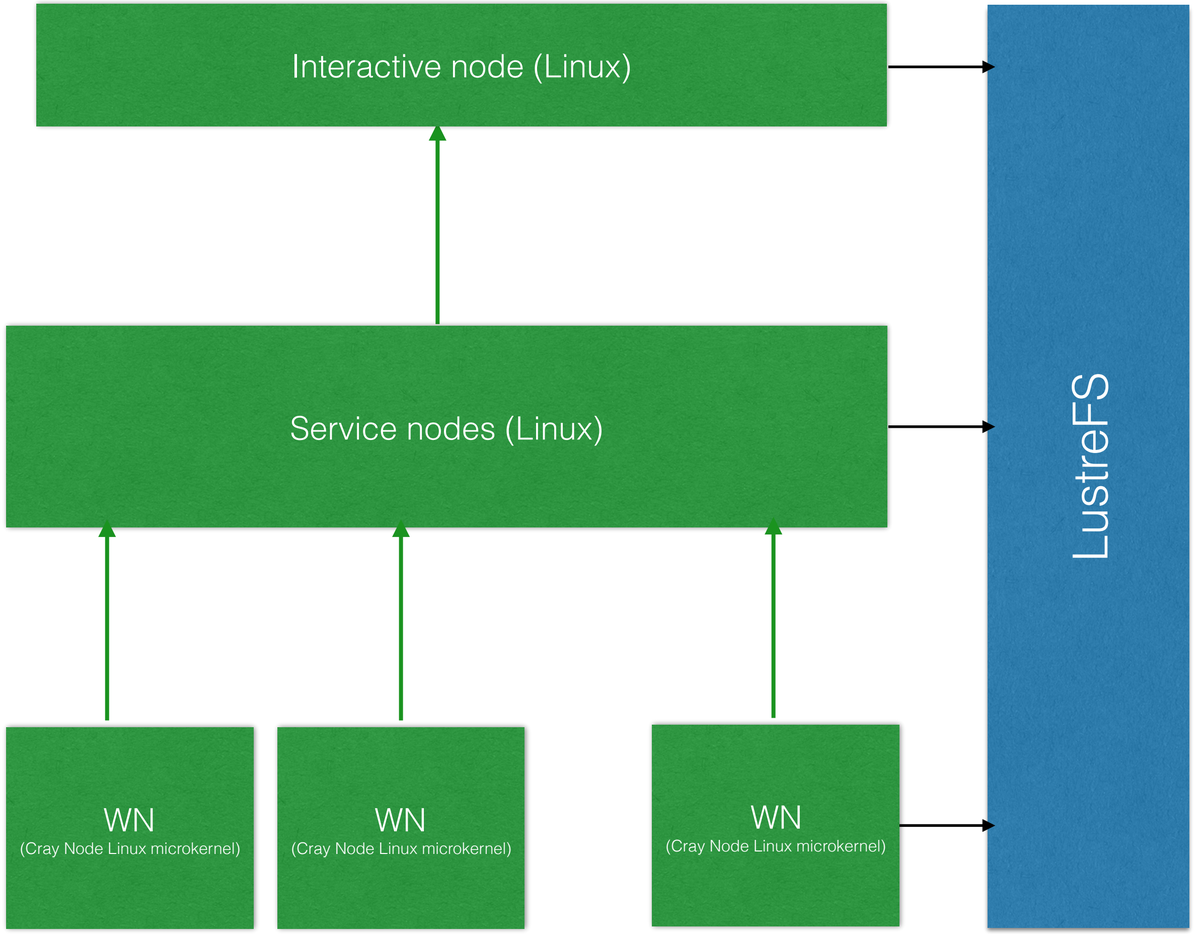
Використовується двохфакторна аутентифікація із застосуванням RSA-токена (пароль користувача + згенерований код з екрану пристроя).

Кожен токен реалізую однаковий алгоритм генерації випадкових чисел, різними є апаратно прошиті числа, на базі яких проходить генерація.

Кожні 60 секунд генерується новий код.

# Архітектура ORNL Titan

- Користувачі можуть проводити підготовку даних для обробки лише на інтерактивних вузлах (interactive nodes)
- Для запуску задач користувач робить запит на отримання комп’ютерного часу.
- Робочі вузли (worker nodes) не мають виходу до зовнішньої комп’ютерної мережі.
- Вся взаємодія користувача з робочими вузлами проходить через спільну файлову систему LustreFS.

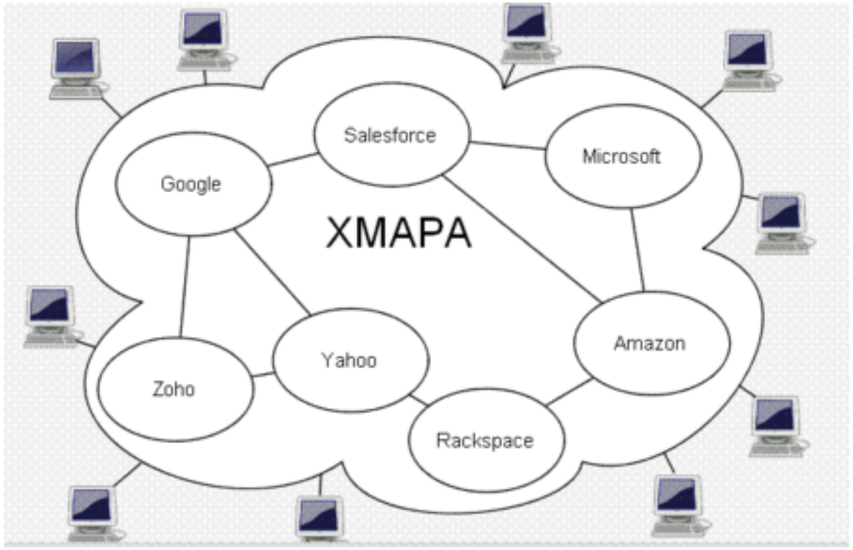


# Українські суперкомп’ютери

- “СКІТ-3” та “СКІТ-4”, встановлені в Інституті кібернетики НАН України.
- На момент запуску (2012 рік) входили до рейтингу “Top-50” на пострадянському просторі (43 TFlops, 170 Тб файлового сховища, пам’ять: 2.5 Тб, 716+448 ядер)
- Залучені до розрахунків у експерименті ALICE
- Серед інших проектів: медична кібернетика, комп’ютерний моніторинг грунтів та підземних вод, математичне моделювання літальних апаратів

# Хмарні обчислювання

**Хмарні обчислення** (англ. *Cloud Computing*) — це модель забезпечення повсюдного та зручного доступу на вимогу через мережу до спільного пулу обчислювальних ресурсів, що підлягають налаштуванню (наприклад, до комунікаційних мереж, серверів, засобів збереження даних, прикладних програм та сервісів), і які можуть бути оперативно надані та звільнені з мінімальними управлінськими затратами та зверненнями до провайдера



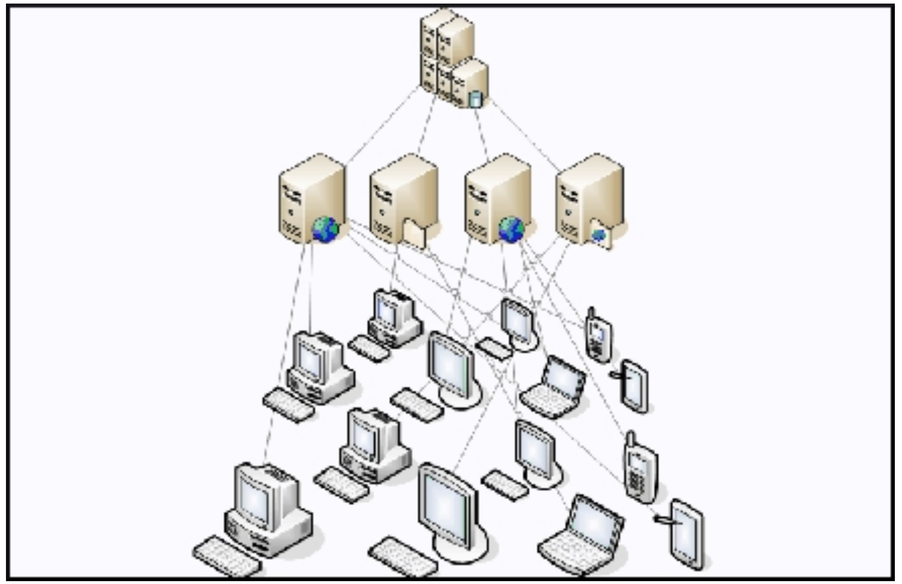

Виділяють наступні моделі надання послуг за допомогою хмари:

- Програмне забезпечення як послуга (SaaS) Прикладами програмного забезпечення як послуги, що працює на основі обчислювальної хмари, є сервіси Gmail та Google docs.
- Платформа як послуга (PaaS) Наприклад, Google Apps надає застосунки для бізнесу в режимі онлайн, доступ до яких відбувається за допомогою Інтернет-браузера тоді як ПЗ і дані зберігаються на серверах Google.
- Інфраструктура як послуга (IaaS) Найбільшими гравцями на ринку інфраструктури як послуги є Amazon, Microsoft, VMWare, Rackspace та Red Hat.

Одніє з найважливіших рис хмарної технології є ***еластичність,*** тобто послуги можуть бути надані, розширені, звужені в будь-який момент часу, без додаткових витрат на взаємодію з постачальником, як правило, в автоматичному режимі

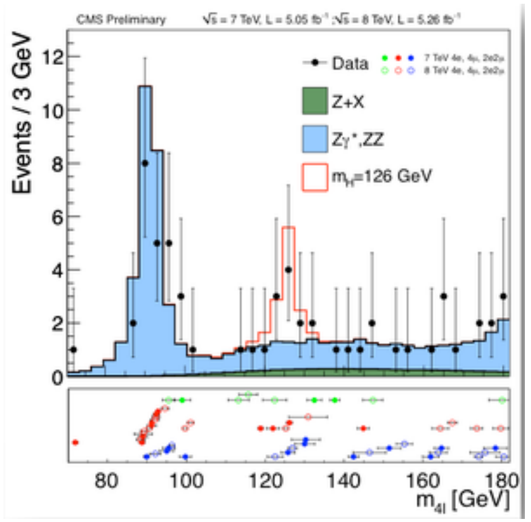
# Волонтерські обчислення

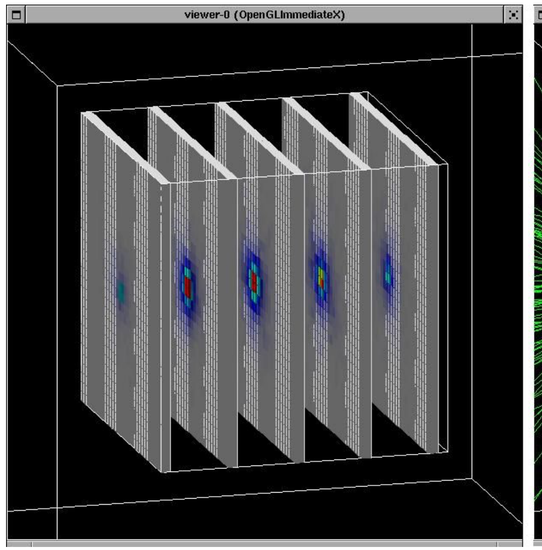
- Комп’ютерні потужності надають користувачі персональних комп’ютерів, тощо.
- Використовуються для математичних розрахунків (наприклад, пошук простих чисел), біологічні розрахунку
- В CERN використовує експеримент ATLAS, хоча на даний момент ефективність такого типу обчислень не є великою
- Учасників треба заохочувати
- Існує критика відносно непрозорості програмного забезпечення через непрозорість програмного забезпечення

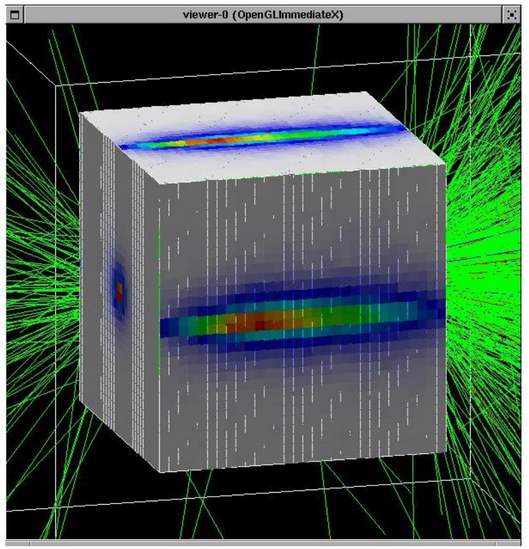


# Програмне забезпечення і CVMFS

- Програмне забезпечення, що найчастіше використовується для аналізу даних:
  - ROOT (статистична обробка даних, матрична алгебра, тощо)
  - GEANT (для опису геометрій детекторів, візуалізації шляхів прольоту частинок, тощо)
- Репозиторій програмного забезпечення ALICE: 1.3 Тб (на початок 2016 р)
- Надається за допомоги розподіленої файлової системи CVMFS

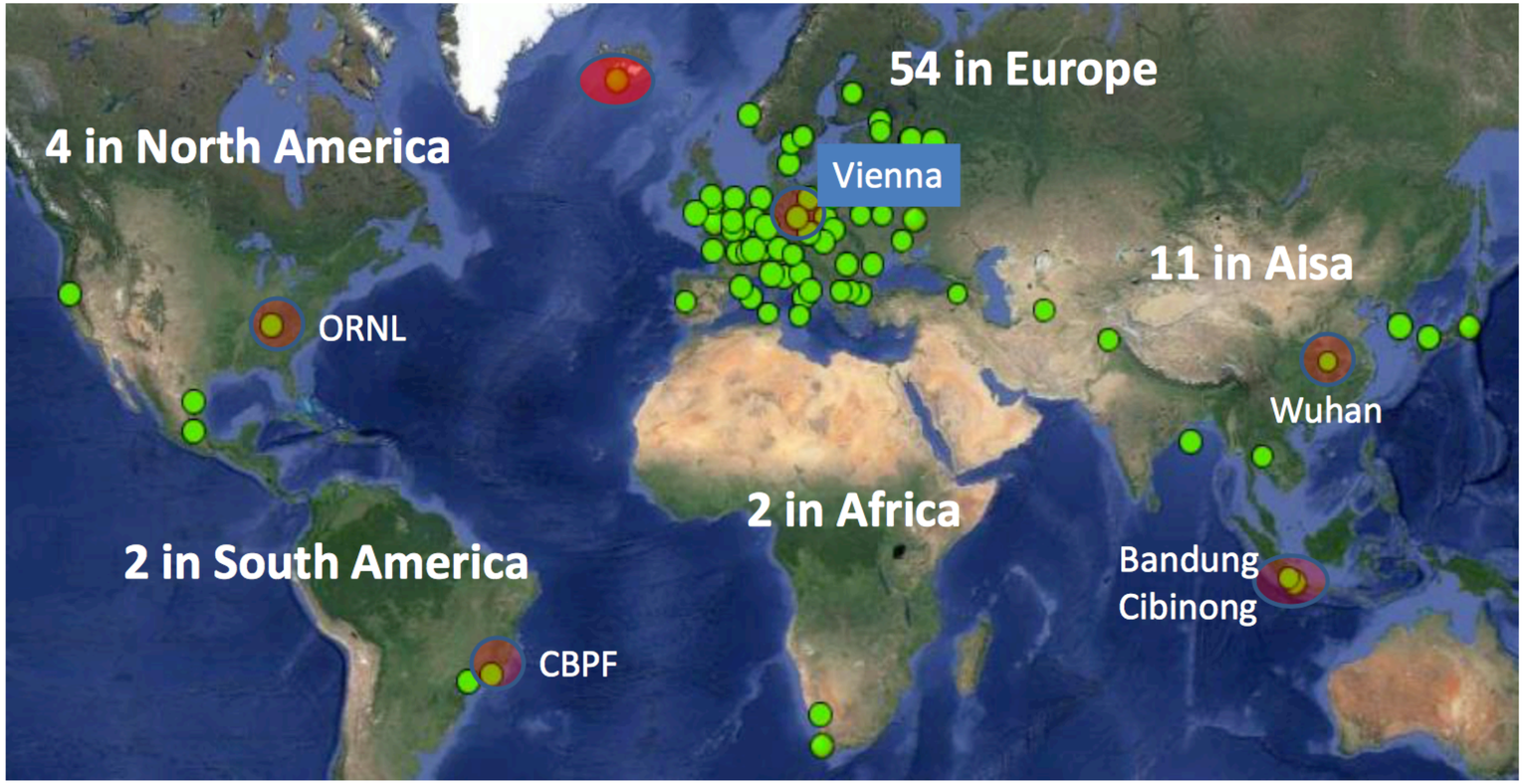



# Грід-інфраструктура ALICE

ALICE
A JOURNEY OF DISCOVERY

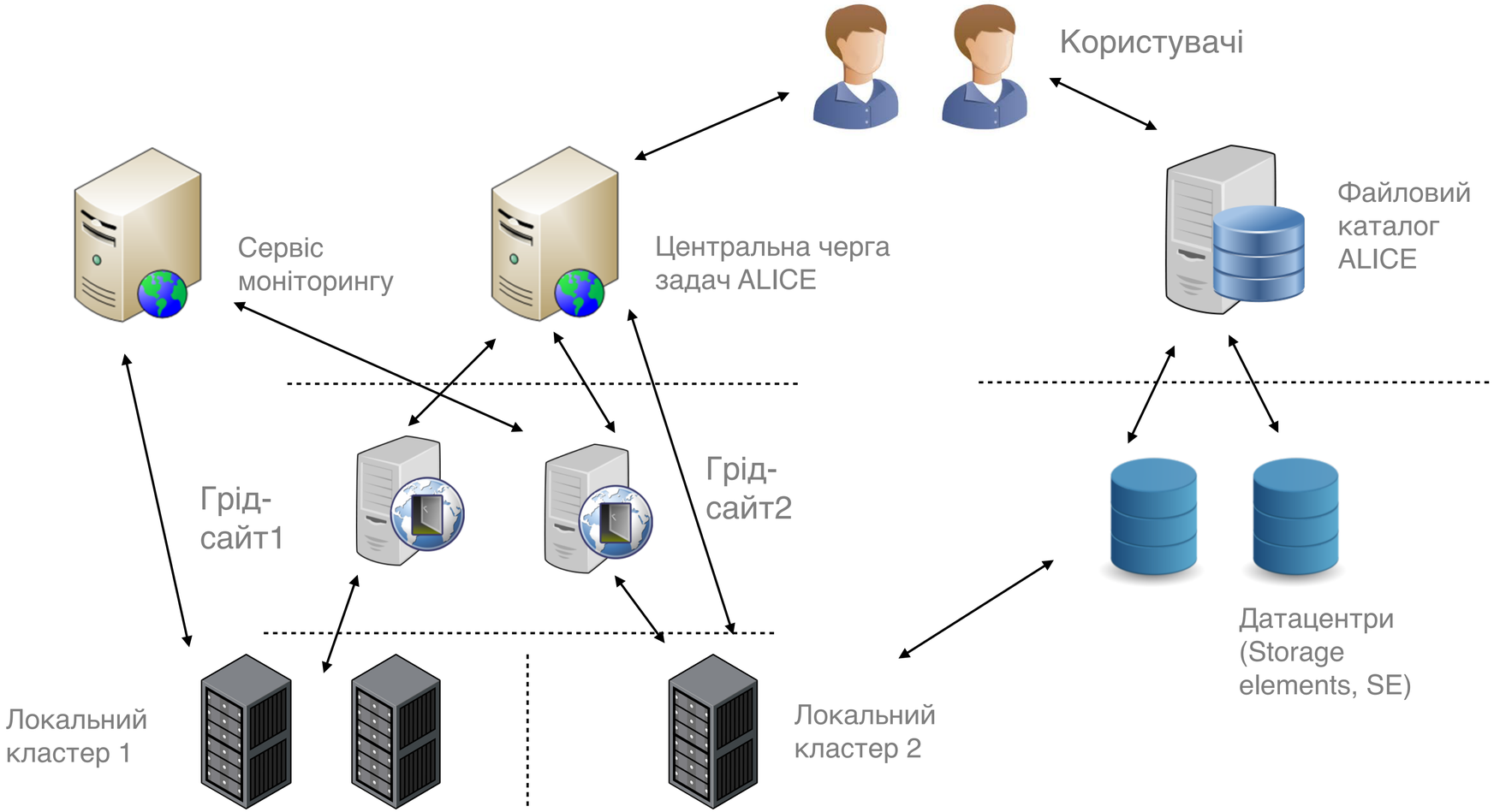
# Ґрід-сайти ALICE

- Всього 73 сайти
- Макс. кількість запущених задач: 110 тис.
- Кількість ядер: 61.4 тис.
- Дисковий об’єм: 26 Пб (5 млн. DVD)
- Кількість файлів у каталозі: 2.14 млрд. (на початок 2016 р.)
- Обмін даними:
  - читання: 20 Гб/сек
  - запис: 3 Гб/сек




ALICE
A JOURNEY OF DISCOVERY

# Структура грід-середовища ALICE




ALICE
A JOURNEY OF DISCOVERY

# Файловий каталог ALICE

aliensh:[alice] [8] /alice/cern.ch/user/p/psvirin/titan-p-p/ >**whereis /alice/cern.ch/user/p/psvirin/titan-p-p/OCDBsim.root**

Dec 6 12:29:37 info The file psvirin/titan-p-p/OCDBsim.root is in
SE => ALICE::CERN::EOS pfn =>root://eosalice.cern.ch:1094//08/12843/1bdabdb0-9c53-11e6-93bc-4bad019c75e6

SE => ALICE::RRC_KI_T1::EOS pfn =>root://io.t1.grid.kiae.ru:1094//08/12843/1bdabdb0-9c53-11e6-93bc-4bad019c75e6

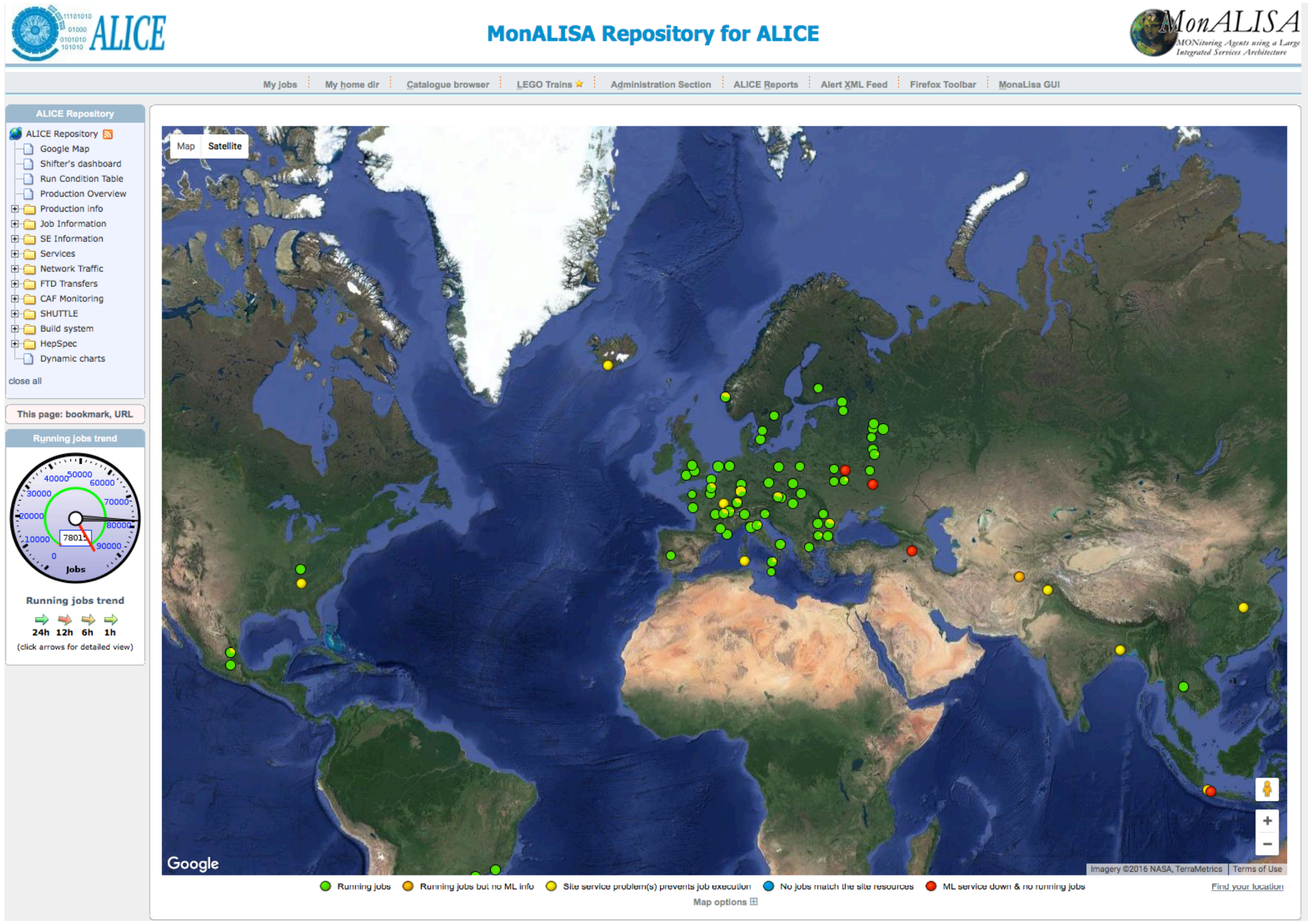
aliensh:[alice] [9] /alice/cern.ch/user/p/psvirin/titan-p-p/ > **whereis /alice/cern.ch/user/p/psvirin/titan-p-p/mc_bitp.jdl**

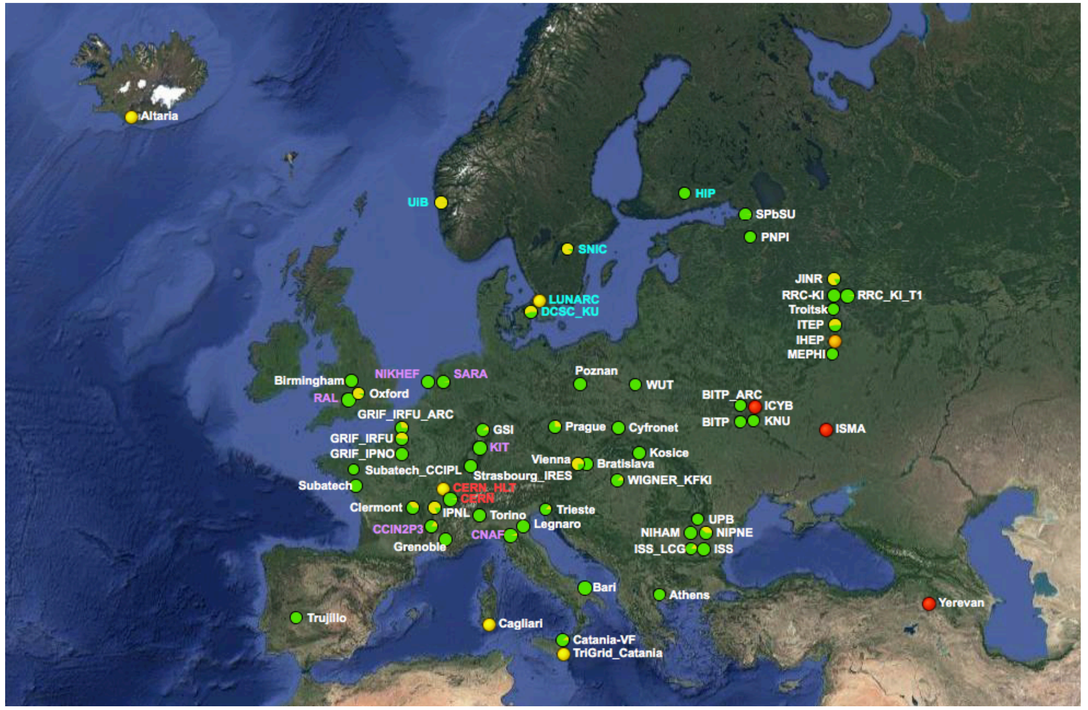
Dec 6 12:29:56 info The file psvirin/titan-p-p/mc_bitp.jdl is in
SE => ALICE::Birmingham::SE pfn =>root://epgsr5.ph.bham.ac.uk:1094//01/00739/d3489068-a1d9-11e6-bf26-001f29eb8d2a

ALICE
A JOURNEY OF DISCOVERY

# Система моніторингу ALICE

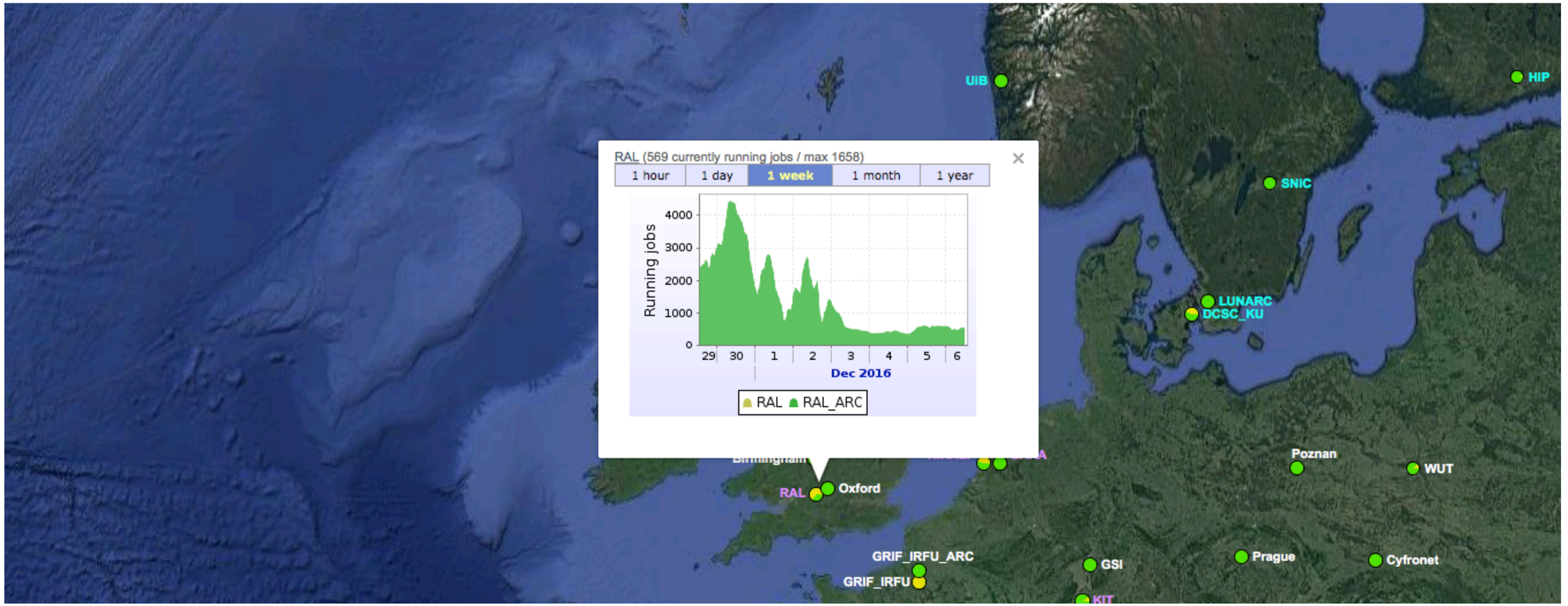
http://alimonitor.cern.ch




ALICE
A JOURNEY OF DISCOVERY

# Система моніторингу ALICE




ALICE
A JOURNEY OF DISCOVERY

# Український грід

# Український Грід

Розпочав роботу 2006 року.

На поточний момент грід інфраструктура України об’єднує 38 кластерів з загальною кількістю ядер більше за  2900 и доступным дисковим об’ємом 250TB.

Кластери працюють під управлінням Nordugrid ARC.

Використовується для досліджень у сферах фізики, біології, медицини та ін.

| | | | | |
|---|---|---|---|---|
| Україна | BITP ARC Training | 11 | 0+0 | 0+0 |
| | BITP Cluster | 88 | 0+34 | 0+0 |
| | CHIMERA | 120 | 0+0 | 116+0 |
| | CSTU ARC CE | 4 | 0+0 | 0+0 |
| | DFTI Cluster | 112 | 0+44 | 1+0 |
| | HPC and FOSS Center | 13 | 0+0 | 0+0 |
| | IAP Cluster | 16 | 0+8 | 0+0 |
| | IAPMM Cluster | 16 | 4+8 | 0+0 |
| | ICMP Cluster | 192 | 0+89 | 0+0 |
| | ICYB SCIT-3 | 1036 | 32+504 | 0+0 |
| | IEP Cluster | 48 | 0+0 (queue inactive) | 0+0 |
| | IFBG Cluster | 72 | 16+41 | 0+0 |
| | ILTPE ARC UA | 88 | 0+0 | 1+0 |
| | ILTPE Cluster | 88 | 0+62 | 0+0 |
| | IMAG cluster | 44 | 0+0 | 0+0 |
| | IMATH Cluster | 16 | 0+2 | 0+0 |
| | IMBG ARC | 100 | 96+0 | 2+0 |
| | IMMSP Cluster | 40 | 0+0 | 6+0 |
| | IMP ARC CE | 84 | 0+0 | 0+0 |
| | INPARCOM Cluster | 8 | 0+0 | 0+0 |
| | INPARCOM GPU Cluster | 8 | 0+0 | 0+0 |
| | IOP Cluster | 104 | 0+0 | 0+0 |
| | IPM Cluster | 44 | 0+5 (queue inactive) | 0+0 |
| | IPMS Cluster | 20 | 0+0 | 0+0 |
| | IRE Cluster | 64 | 0+0 | 0+0 |
| | ISMA cluster | 332 | 0+305 | 0+39 |
| | ISOFTS Cluster | 8 | 0+0 | 0+7605 |
| | KIPT IPP | 2 | 0+0 | 0+0 |
| | KMA Grid Cluster | 0 | | 0+0 |
| | KNU ARC | 40 | 0+37 | 13+11 |
| | KPI training cluster | 24 | 0+0 | 0+0 |
| | LNU Training Cluster | 28 | 0+28 | 0+0 |
| | MAO Cluster | 104 | 0+48 | 0+0 |
| | MHI Cluster | 120 | 0+0 | 0+0 |
| | PIMEE ARC | 24 | 16+0 | 6+0 |
| | RIAN | 1 | 0+0 | 0+0 |
| | SRI cluster | 4 | 0+0 | 0+0 |

# Участь в експериментах CERN


BITP: Інститут теоретичної фізики ім. Боголюбова НАН України

BITP_ARC: Інститут теоретичної фізики ім. Боголюбова НАН України (експериментальний грід-сайт)

KNU: Київський національний університет

ICYB: Інститут кібернетики ім. Глушкова НАН України (CKIT-3/CKIT-4)


UA-ISMA: Інститут сцинтиляційних матеріалів НАН України

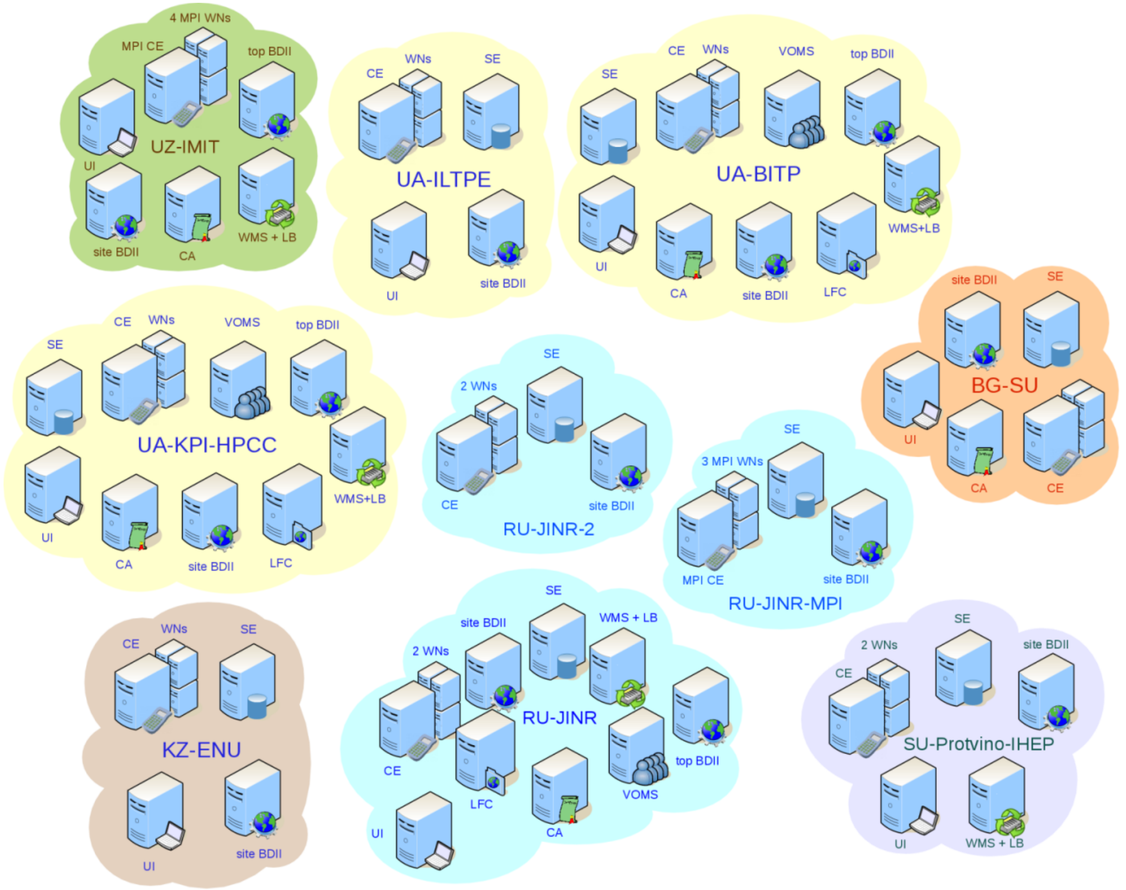
UA-KIPT-LCG2: Харківський фізико-технічний інститут

# Навчальна інфраструктура

Однією з основних проблем Українской грід-спільности є нестача спеціалістів, що вміють використовувати Грід-технології для наукових розрахунків.

На базі НТУУ КПІ організовано дісплейний клас для підготовки грід-адміністраторів.

Інститут теоретичної фізики ім. Боголюбова проводить відеоконференції для грід-користувачів.



# Освітні проекти CERN

# CERN School of Computing (CSC)

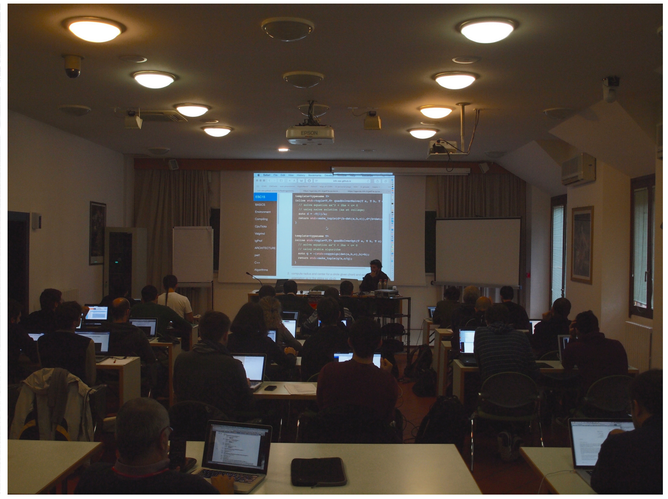
- Кожного року проходить в новому університеті
- Кількість учасників: бл. 60 чол.
- Тематика:
  - паралельне програмування
  - технології зберігання даних
  - ефективне програмування і оптимізація програмного забезпечення
  - безпека
  - машинне навчання
  - аналіз фізичних даних
- Наступна: у Мадриді


# Thematic CSC (tCSC)

- Проходить кожного року у Спліті (Хорватія), організується факультетом електротехніки Сплітського університету
- Тематика:
  - ефективне програмування і оптимізація програмного забезпечення
  - паралельне програмування
  - C++ і його особливості

# ESC INFN

- Проходить кожного року в Бертіноро (Італія)
- Тематика:
  - ефективне програмування і оптимізація програмного забезпечення, ефективне керування пам’яттю
  - паралельне програмування: OpenMPI та OpenMP
  - паралельне програмування на графічних прискорювачах
  - C++ і його особливості



# Висновки

- Україна є учасником обчислень для CERN, очікуємо збільшення долі в обчисленнях
- Підготовка нових кадрів, що вміють адмініструвати ґрід-сайти, та користувачів ґрід
- Очікуємо збільшення рівня участі українських студентів і аспірантів в освітніх програмах CERN, а також в програмах Technical Students/Summer Students

# ДЯКУЮ ЗА УВАГУ!