Грід-системи

Судаков О.О.

“Паралельні і розподілені обчислення”
Лекція 6

План

- Структура грід
- Авторизація
- Реалізації проміжного програмного забезпечення грід
- Робота в Українській грід-інфраструктурі

Метакомп'ютери

- Метакомп'ютери – використання існуючих (простоюючих) комп'ютерних ресурсів для рішення задач
  - комп'ютерний клас
  - комп'ютери в межах Інтернет
- Потенційно висока продуктивність – мільйони процесорів

Використання потужності існуючих комп'ютерів

- В нічний час комп'ютери часто простоюють
  - Потенціальна потужність простоюючих комп'ютеров може бути дуже велика
  - Дуже дешеві ресурси
  - Можливість забезпечити надлишковість ресурсів
- Недоліки
  - Надійність каналів передачі невелика
  - Через малі швидкості передачі даних в межах Інтернет неможливо виконувати паралельні розрахунки

Що таке грід?

- Грід
  - Стандарти по з’єднанню обчислювальних ресурсів через інтернет в одну велику систему
- Визначення
  - Грід - Спеціальна форма розподілених обчислень направлена на спільне використання великої кількості георгафічно разподілених ресурсів у віртуальних організація
- Походження
  - Грід - Computing Grid ⇔ Power Grid - мережа

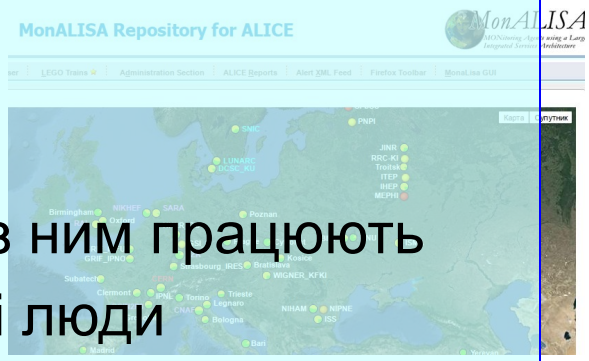
Віртуальні організації (ВО)

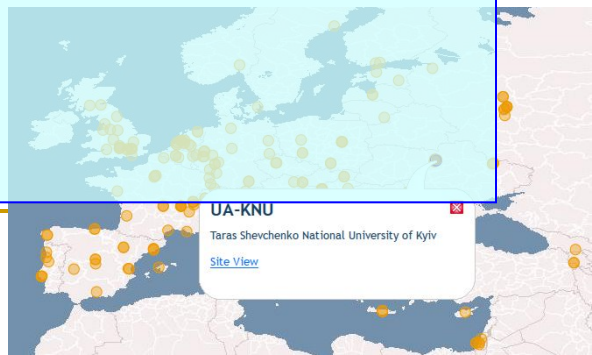
- ВО – добровільне об’єднання людей чи організацій які спільно використовують частину ресурсів грід
- Ресурси грід
  - Обчислювальні елементи – кластери, суперкомп’ютери
  - Елементи збереження – файлові сервери, бази даних
  - Джерела даних (датчики, сенсори) – установки, які видають експериментальні дані
  - Інтерфейси користувачів – портали, термінальні станції
  - Служби – програми, які виконують певні інфраструктурні функції
  - Прикладні програми – наукові, промислові …

LGC – приклад грід системи

LGC – Large Hagron Collider Computing Grid

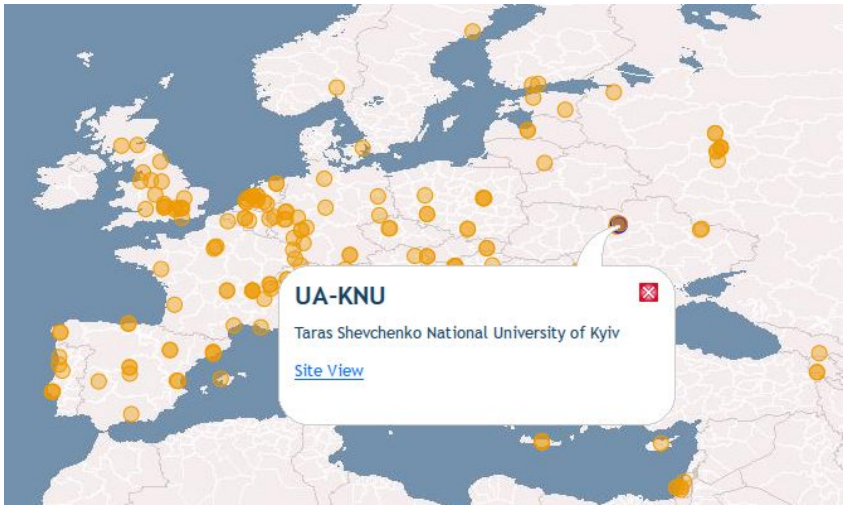
- LCG – джерело даних
  - **Дані - 20 мильйонів CD щорічно!**
  - **обробка ~ 100 000 ПК**
- Кластери по світу – обчислювальні елементи і елементи збереження даних
- Віртуальні організації
  - ATLAS – один з детекторів і люди які з ним працюють
  - ALICE - ще один детектор і відповідні люди
  - CMS – ще один детектор
  - …



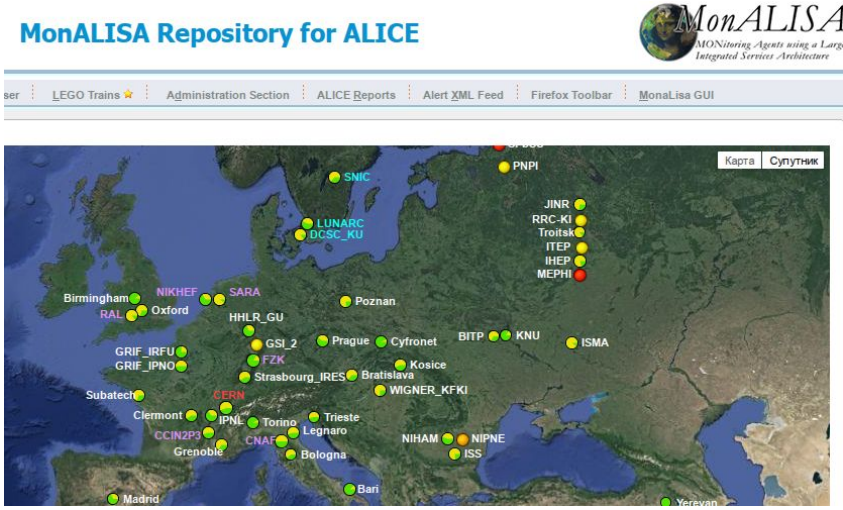


Приклад ВО

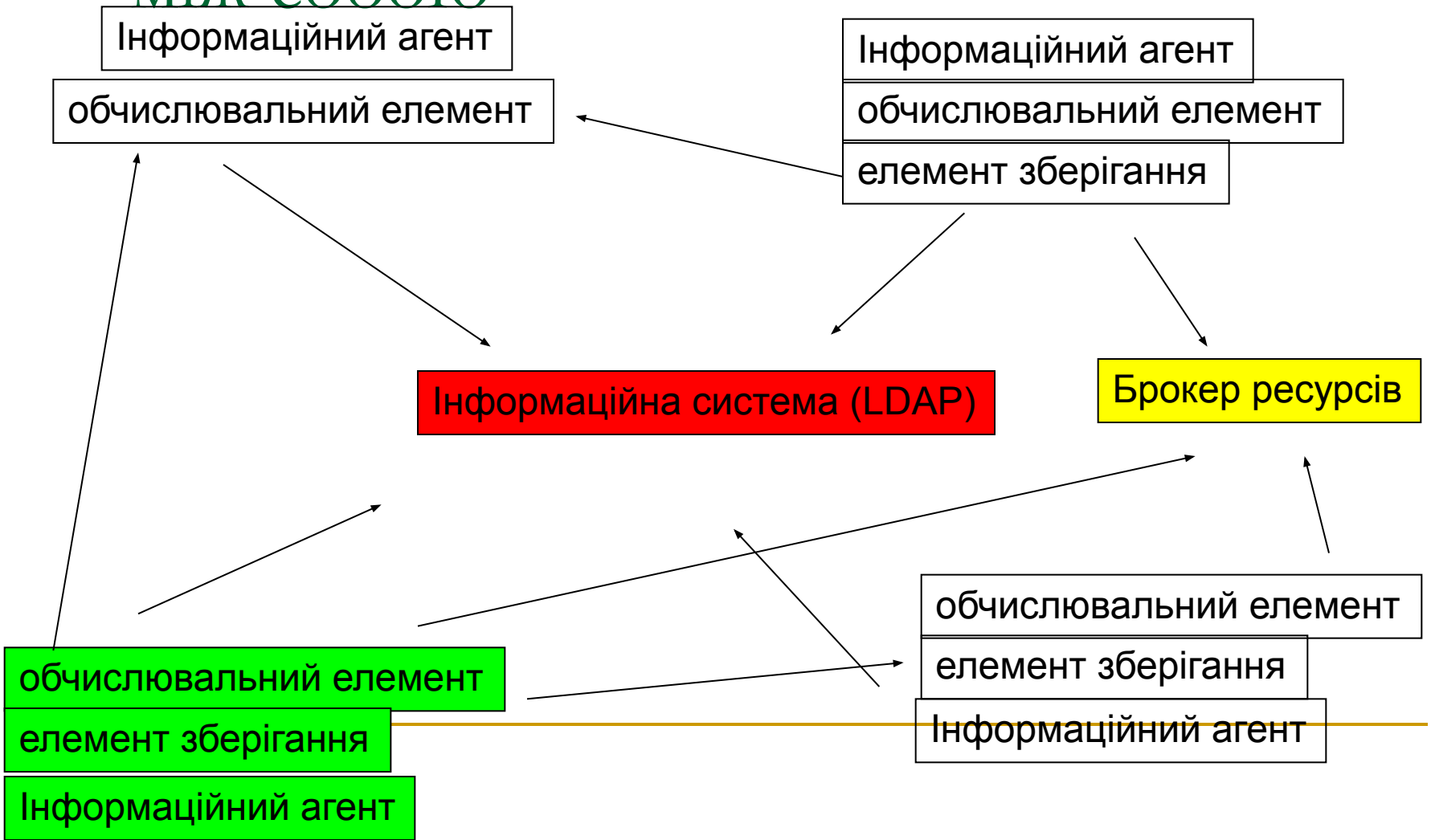
- Кластери європейської грід-інфраструктури
  - prod.cern.ch/gstat



- Частина кластерів рахує для проекту Alice
  - http://alimonitor.cern.ch



Структурна схема – набір географічно розподілених служб, які взаємодіють між собою



Служби грід

- Інфраструктурні служби (1-2 на весь грід)
  - Інформаційна система – база даних
    - Які кластери, задачі, елементи збереження…
    - Всі мають доступ до цієї служби
    - Приклад - Grid Index Information Service (GIIS)
  - Брокер ресурсів – планувальник і запуск задач на обчислювальних елементах
- Служби ресурсів (на всіх кластерах)
  - Інформаційний агент – оновлення інформації в інформаційній службі (кожні 5 хвилин)
    - Приклад Grid Resource Information Service (GRIS)
  - Обчислювальний елемент (CE) – інтерфейс до системи пакетного режиму
  - Елемент збереження даних (SE) – інтерфейс до системи збереження даних
- Багато різних несумісних релізацій програм і протоколів

Приклад роботи інформаційної системи

- Інформаційний агент кожні 5 хвилин
  - З'єднується з інформаційною системою
  - Передає стан всіх задач
  - Передає наявність і стан всіх служб на кластері
    - CE, SE…
- Брокер ресурсів
  - Для запуску задачі вибирає з інфосистеми кластери, які задовольняють вимогам
  - Знаходить найкращий кластер
  - Звертається до SE і CE кластера для запуску задачі
  - Перевіряє стан задач, рестартує …
  - Видаляє інформацію про завершені задачі з інфосистеми

Приклад запуску задачі

- Користувач
  - Створює (вибирає) файл опису задачі
  - Створює (вибирає) програми для запуску
  - Викликає команду для відправки задачі брокеру
  - Періодично звертається до брокера дял отримання статусу задачі
  - Після виконання отимує файли результатів
- Брокер
  - Перевіряє опис задачі
  - Запускає задачу

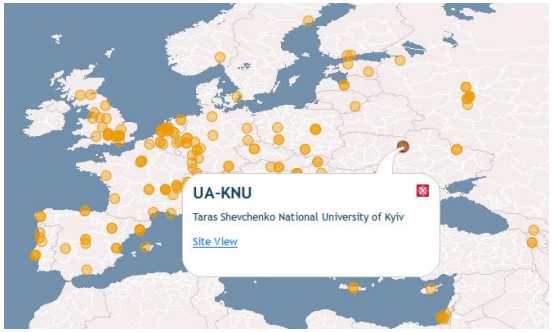
Приклад виконання задачі

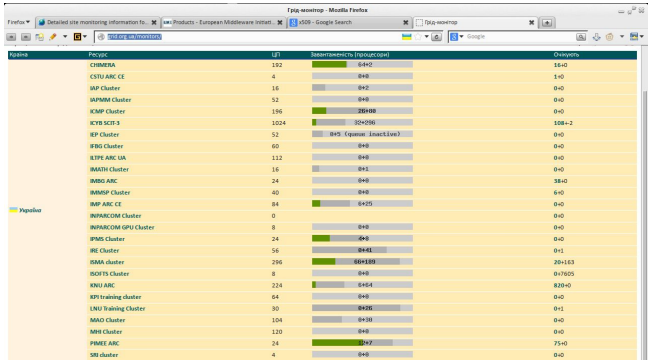
- Брокер
  - вибір кластера з інфосистеми
  - Передача на CE кластера опису задачі
  - Передача на SE кластера вхідних файлів
- CE
  - очікує поки всі вхідні дані будут готові –stagin (з брокера чи інших SE)
  - ставить задачу в чергу на кластері
  - Чекає поки задача дорахується
  - Знищує всі файли задачі, крім вихідних (stagout)
  - Записує вихідні файли на інші SE чи на брокер
  - Реєструє в інфосистемі завершення задачі
- SE
  - Виконує запити на запис, читання файлів

Моніторинг – аналіз стану всіх служб за інформаційною системою

- http://prod.cern.ch/gstat
- http://alimonitor.cern.ch
- http://grid.org.ua/monitors

- Є моніторинг з обмеженим доступом





Необхідність авторизації в грід

- Багато служб і користувачів
  - Якщо всім все можна - легко зловживати довірою
  - Для стабільності і ефективності необхідно давати доступ лише авторизованим користувача і службам
  - Потрібен механізм точно знати хто звертається

- Grid Security Infrastructure (GSI)
  - Кожен користувач має персональний сертифікат
  - Кожна служба має персональний сертифікат
  - Сертифікат – достовірний спосіб авторизації (хто такий)

Сертифікати X509 - стандарт авторизації публічного ключа

- Користувач генерує 2 ключі
  - Закритий (таємний) – для шифрування
  - Відкритий (публічний) - для розшифрування
  - Якщо документ не зашифровано приватним ключем, то публічним його розшифрувати дуже важко

- Електронний підпис даних
  - Генерується контрольна сума даних
  - Контрольна сума підписується приватним ключем
  - Публічний ключ передається разом з даними і зашифрованою контрольною сумою
  - Перевіряється контрольна сума даних і порівнюється з розшифрованою контрольною сумою

Акредитований центр сертифікації ключів (CA, Certification Authority)

- Організація, якій всі довіряють
  - Публічний ключ CA доступний всім
  - Може гарантувати, що підписує сертифікати лише тим кому треба
- Підпис сертифіката
  - Публічний ключ підписується CA
  - Сертифікат – публічний ключ+його контрольна сума, підписана CA
  - Дерево підписів – сертифікат підписаний сертифікатом, підписаний сертифікатом …

Перевірка сертифікатів

- Сертифікат містить
  - Дату видачі
  - Тривалість дії
  - Ідентифікатор
  - Ідентифікатор того, хто підписав
- Перевіряється
  - дійсність всіх полів (час…)
  - Дійснсіть сертифікатів всіх, хто підписав
- Відкликання сертифікатів
  - Можуть всі, хто входить в дерево: CA, користувач…
  - Генерується список відликаних сертифікатів: CRL

Проксі сертифікат

- Проксі
  - Короткодіюча пара приватного і публічного ключів
  - підписана сертифікатом користувача
  - Передається публічно іншим для виконання дії від імені користувача
  - Не може діяти довше, ніж сертифікати, яким підписано
- Використання
  - Передається разом із задачею, щоб кластер (служба) могли підписуватись від імені того, хто запустив задачу

Служба VOMS- Virtual Organization Membership Service

- Авторизує членство у віртуальних організація
  - Список унікальних імен користувачів (DN) і CA
  - /DC=org/DC=ugrid/O=people/O=KNU/OU=ICC/CN=Oleksandr Sudakov
- Користувачі реєструються у віртуальній організації
  - Підписують (шифрують) запит своїм сертифікатом
- VOMS
  - Перевіряє дійсніть підпису
  - Відправляє запит адміністратору
- Адміністратор
  - Повинен мати дійсний сертифікат
  - Права адміністратора на додвання користувачів
- VOMS-проксі
  - Проксі-сертифікат користувача з додатковими атрибутами VOMS
  - Підписаний сертифікатом служби VOMS

Авторизація на грід-ресурсах

- Кожна служба (програма)
  - Має підписаний сертифікат
  - Має список всіх сертифікатів довірених CA
  - Має список всіх довірених ВО
  - Має список CRL
- Кожен кластер (ресурс)
  - Оновлює список всіх сертифікатів довірених CA
  - Періодично оновлює список CRL
  - Періодично оновлює список користувачі довірених ВО
  - Може відображати користувачів ВО на локальних користувачів
- Кожен користувач
  - Генерує пару ключів
  - Підписує сертифікат в CA
  - Реєструэться у ВО
  - Отримує доступ до всіх служб, які які підтримують ВО

Проміжне програмне забезпечення грід

- Проміжне програмне забезпечення (програмне забезпечення середнього рівня, middlware)
  - Бібліотеки, служби, програми – які дають можливість використовувати грід-інфраструктуру
  - Для грід - як правило складне програмне забезпечення, погано структуроване і нестабільне
  - Є різні несумісні реалізації

Реалізації

- Globus toolkit http://www.globus.org
  - мінімальний набір засобів для створення грід-служб
  - Основа багатьох інших middleware
- Unicore http://www.unicore.eu
  - Орігінальний набір інфтрументів для об'єднання суперкомп'ютерів у Німеччині
- LCG http://wlcg.web.cern.ch
  - Розширений набір служб та інструментів на базі globus для LHC
- Glite
  - розширення LCG
  - Одне з найфункціональніших
- Nordugrid-arc
  - Розширення globus для nordugrid
  - Одне з найстабільніших
- EMI www.eu-emi.eu
  - Найновіше
  - Об'єднує користні функції ARC, glite, Unicore

Українська грід-інфраструктура

- 2002 Перший грід-кластер у Харківському фізико-технічному інституті
  - В росйському грід і до цього часу
- 2005 Перші Українські кластери AlieEn-грід за ініціативою Інституту теорфізики НАНУ та Київського університету
  - в КНУ - Перший офіційний український грід-ресурс
- 2007 створення Українського академічного гріду
  - Близько 10 кластерів
- 2009 перший український кластер (КНУ) в EGEE

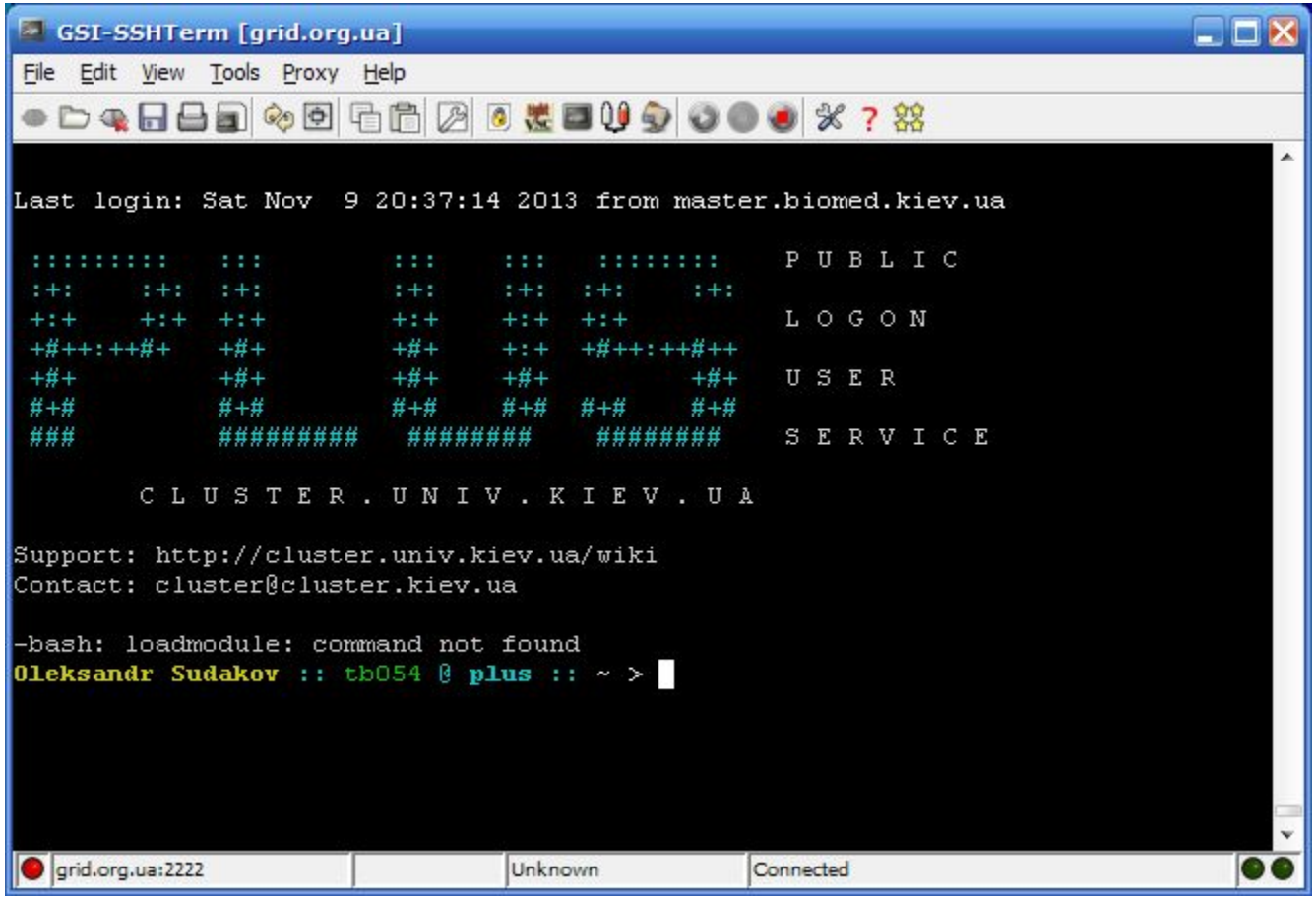
Особливості реалізації

- Middleware nordugrid-arc – більшість кластерів
- Інфосистема
  - Giis.grid.org.ua – КНУ
  - Giis.bitp.kiev.ua – ВІТР
- Віртуальні організації
  - voms.grid.org.ua
- Кластери
  - Grid.org.ua/monitors
  - Nordugrid.org

Застосування грід

- Отримання сертифікату
  - https://ca.ugrid.org
  - https://testbed.univ.kiev.ua – тільки для університету
- Реєстрація у ВО
  - https://voms.grid.org.ua – для України
- Встановити UI
  - Отримати доступ до командного рядка кластера
  - Застосувати GSI-SSH https://testbed.univ.kiev.ua

Доступ до кластера КНУ через GSI-SSH



Генерація проксі-сертифікату

- При доступі по GSI-SSH проксі вже є

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > echo $X509_USER_PROXY
/tmp/x509up_p21149.fileHWeC6N.1
Oleksandr Sudakov :: tb054 @ plus :: ~ >
```

- Коли на кластері є сертифікат і ключ у форматі X509

```
[saa@s27 ~]$ loadmodule -i arc-2.0.0
Module arc-2.0.0 allready loaded
[saa@s27 ~]$ loadmodule -i glite-3.2
Module glite-3.2 allready loaded
[saa@s27 ~]$ arcproxy -S testbed.univ.kiev.ua
Enter pass phrase for load private key:
Your identity: /DC=org/DC=ugrid/O=people/O=KNU/OU=ICC/CN=Oleksandr Sudakov
Contacting VOMS server (named testbed.univ.kiev.ua): voms.grid.org.ua on port: 1
5112
Proxy generation succeeded
Your proxy is valid until: 2013-11-10 08:52:00
[saa@s27 ~]$
```

Конвертація сертифікатів

- 3 pkcs12 у X509

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > ls saa.p12
saa.p12
Oleksandr Sudakov :: tb054 @ plus :: ~ >
```

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > mkdir -p .globus
Oleksandr Sudakov :: tb054 @ plus :: ~ > openssl pkcs12 -in saa.p12 -out .globus
/userkey.pem -nocerts
Enter Import Password:
MAC verified OK
Enter PEM pass phrase:
Verifying - Enter PEM pass phrase:
Oleksandr Sudakov :: tb054 @ plus :: ~ > openssl pkcs12 -in saa.p12 -out .globus
/usercert.pem -nokeys
Enter Import Password:
MAC verified OK
Oleksandr Sudakov :: tb054 @ plus :: ~ >
```

- 3 X509 у pkcs12

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > openssl pkcs12 -export -in .globus/user
cert.pem -inkey .globus/userkey.pem -out cert.p12
Enter pass phrase for .globus/userkey.pem:
Enter Export Password:
Verifying - Enter Export Password:
Oleksandr Sudakov :: tb054 @ plus :: ~ > ls cert.p12
cert.p12
Oleksandr Sudakov :: tb054 @ plus :: ~ >
```

Налаштування інформаційної системи на клієнті

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > cat ~/.arc/client.conf


[registry/KNU]
url = ldap://giis.grid.org.ua:2135/Mds-Vo-name=Ukraine,o=grid
registryinterface = org.nordugrid.ldapegiis
default = yes

[registry/BITP]
url = ldap://lcg.bitp.kiev.ua:2135/Mds-Vo-name=Ukraine,o=grid
registryinterface = org.nordugrid.ldapegiis
default = yes
```

Файл опису задачі XRSL

```
Oleksandr Sudakov :: tb054 @ plus :: job_scripts > cat
   nordujob
&
(executable=job.sh)          - сценарій задачі
(executables=a.out)          - виконувані файли
(inputFiles=(a.out ""))        - stagin файли
(stdout="hello.txt")         - стандартний вивід
(stderr="hello.err")          - стандартні помилки
(outputFiles=
    ("hello.txt" "")
    ("hello.err" "")
 )                 - stagout файли
(gmlog="gridlog")            - інформація грід
(jobname="Chimera Jobs")   - ім'я задачі
( walltime="20 minutes")   - walltime
```

Вказування вхідних і вихідних файлів

```
(inputFiles=
(файл1_на_грід "звідки_брати")
(файл2_на_грід " звідки_брати "))
(outputFiles=
 ("файл1_на_грід" "куди_класти")
 ("файл1_на_грід" "куди_класти")
)
```

Шляхи до файлів

- Поточний каталог запуску програми
  - “”
  - `(inputFiles=(a.out ""))`
  - `(outputFiles=("hello.txt" ""))`
- Елемент збереження
  - Повинен бути доступ
  - Різні протоколи
    - `srm://se.biomed.kiev.ua/networkdynamics/hello.txt`

Запуск задачі

- arcsub опис_задачі
- опції arcsub
  - -c кластер
  - -c arc.univ.kiev.ua –c arc.biomed.kiev.ua –c grid.isma.kharkov.ua
  - -d 1|2|3 – відладочна інформація
- Вивід
  - Ідентифікатор задачі

Приклад запуску

```
Oleksandr Sudakov :: tb054 @ plus :: ~ > cp $X509_USER_PROXY ./user_proxy

Oleksandr Sudakov :: tb054 @ plus :: ~ > ssh cluster.univ.kiev.ua

[tb054@s27 ~]$ export X509_USER_PROXY=~/user_proxy
[tb054@s27 ~]$
[tb054@s27 ~]$ loadmodule -i arc-2.0.0
Module arc-2.0.0 allready loaded
[tb054@s27 ~]$ loadmodule -i glite-3.2
Module glite-3.2 allready loaded
[tb054@s27 ~]$
[tb054@s27 ~]$  cd job_scripts/
[tb054@s27 job_scripts]$ ls a.out job.sh nordujob
a.out  job.sh  nordujob
[tb054@s27 job_scripts]$
[tb054@s27 job_scripts]$ arcsub -c arc.univ.kiev.ua -c arc.biomed.kiev.ua
   -c gri
d.isma.kharkov.ua nordujob
Job submitted with jobid:
   gsiftp://arc.univ.kiev.ua:2811/job/LJNODmAPTzinPoP5up3
LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
```

Перевірка стану задачі

```
[tb054@s27 job_scripts]$ arcstat
   gsiftp://arc.univ.kiev.ua:2811/job/LJNODmAPTzin
PoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
WARNING: Job information not found in the information system:
   gsiftp://arc.univ.
kiev.ua:2811/job/LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
No jobs
[tb054@s27 job_scripts]$
```

Чекаємо хвилин 5

```
[tb054@s27 job_scripts]$ arcstat
   gsiftp://arc.univ.kiev.ua:2811/job/LJNODmAPTzin
PoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
Job:
   gsiftp://arc.univ.kiev.ua:2811/job/LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm
   2iJK
DmABFKDm0j6Rkm
 Name: Chimera Jobs
 State: Finished (FINISHED)
 Exit Code: 0

[tb054@s27 job_scripts]$
```

Отримання результатів

- Arcget задача

```
[tb054@s27 job_scripts]$ arcget
   gsiftp://arc.univ.kiev.ua:2811/job/LJNODmAPTzinPoP5up3LFXdqABFKD
   mABFKDm2iJKDmABFKDm0j6Rkm
Results stored at:
   LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
Jobs processed: 1, successfully retrieved: 1, successfully cleaned:
   1
[tb054@s27 job_scripts]$
[tb054@s27 job_scripts]$ cd
   LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm
[tb054@s27 LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm]$
   ls
gridlog  hello.err  hello.txt
[tb054@s27 LJNODmAPTzinPoP5up3LFXdqABFKDmABFKDm2iJKDmABFKDm0j6Rkm]$
```

Особливості застосування грід

- Переваги
  - Дуже велика сумарна обчислювальна потужність
  - Дуже великі елементи збереження даних
- Недоліки
  - Кожна задача може запускатись повільно
  - Ресурси можуть бути недоступні або вимкнутись в процесі роботи
  - Ваша програма може бути несумісна з операційною системою іншого кластера
- Коли грід ефективний
  - Велика кількість слабко залежних задач