West-Life report

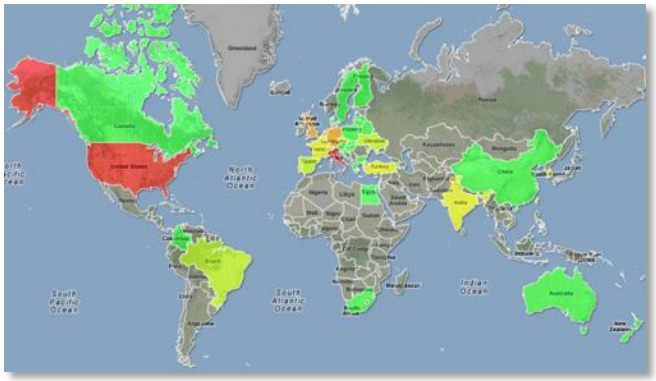
Personale coinvolto nel progetto:

- TI: Marco Verlato, Massimo Sgaravatto
- TD: Paolo Andreetto, Sergio Traldi
- AR: Matteo Segatta, Nicola Tritto

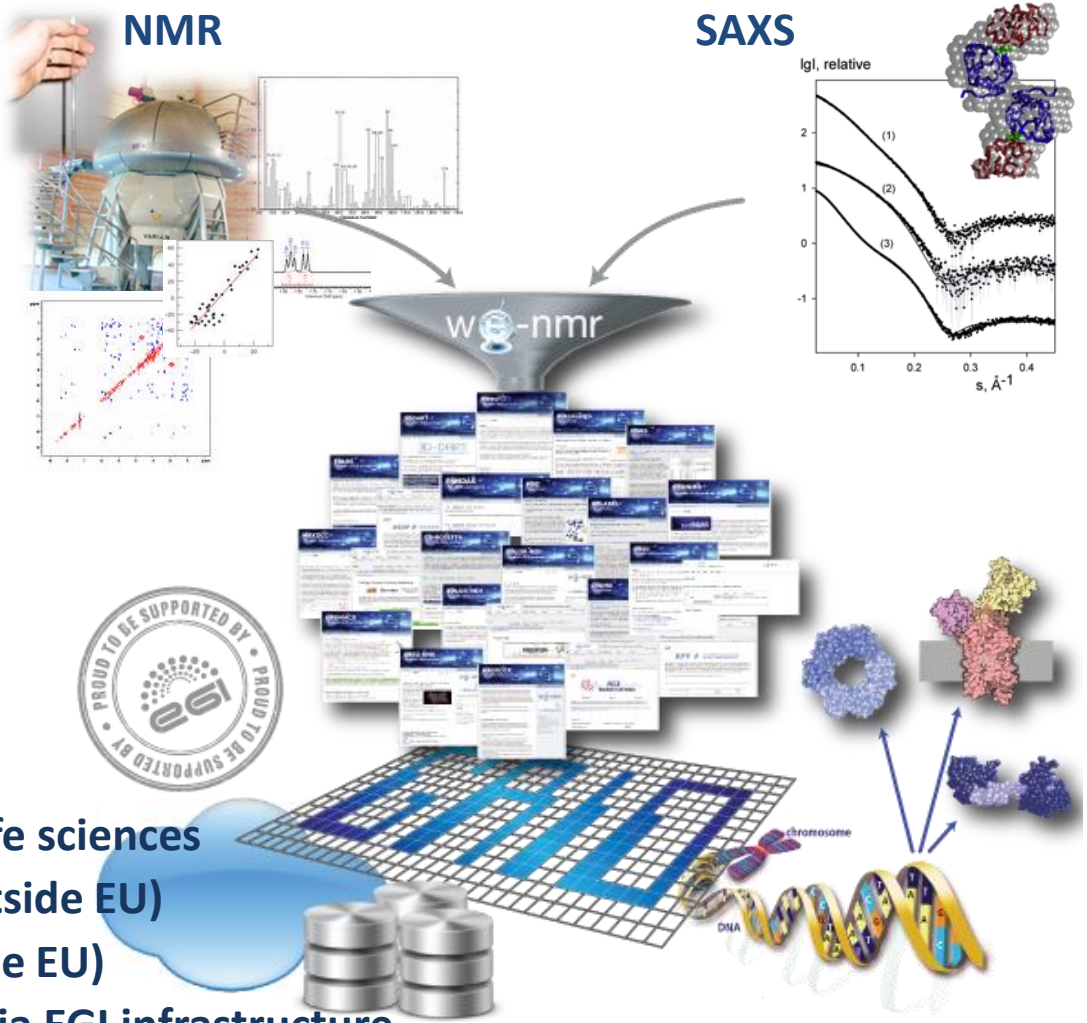
Incontro sul calcolo a Padova

26/9/2016

A worldwide e-Infrastructure for NMR and structural biology



WeNMR VRC (December 2015)



- One of the largest (#users) VO in life sciences
- > 720 VO registered users (36% outside EU)
- > 2250 VRC members (>60% outside EU)
- ~ 41 sites for >142 000 CPU cores via EGI infrastructure
- > 3M jobs/year, > 30M HS06.hrs/year
- User-friendly access to Grid via web portals

e-nmr → w-nmr → MoBrain → West-Life

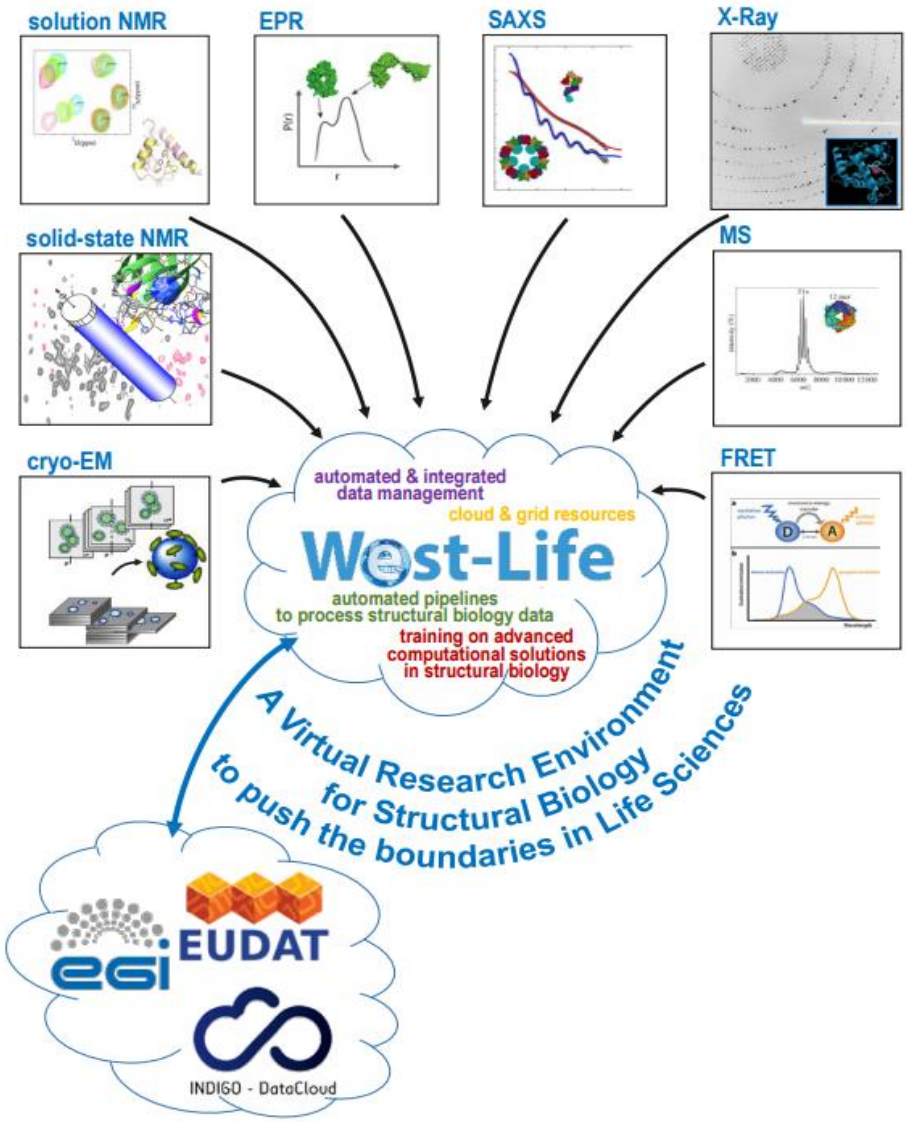
2007 2010 2013 2015 2018

West-Life World-wide E-infrastructure for structural biology

Budget: € 4 000 000 (344 k€ x INFN)
Duration: 36 months
Started: 1 Nov 2015
Proposal ID: 675858

1. Science and Technology Facilities Council, UK
2. Instruct Academic Services, UK (ESFRI for SB)
3. LUNA, France
4. Masarykova Univerzita, Czech Republic
5. Agencia Estatal Consejo Superior de Investigaciones Cientificas, Spain
6. Netherlands Cancer Institute – Antoni van Leeuwenhoek Ziekenhuis (NKI AVL)
7. **Universiteit Utrecht, Netherlands**
8. **EMBL, Germany**
9. **Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine, Italy**
10. **Istituto Nazionale di Fisica Nucleare, Italy**

Networking: 15%
Services: 63%
Joint research: 22%



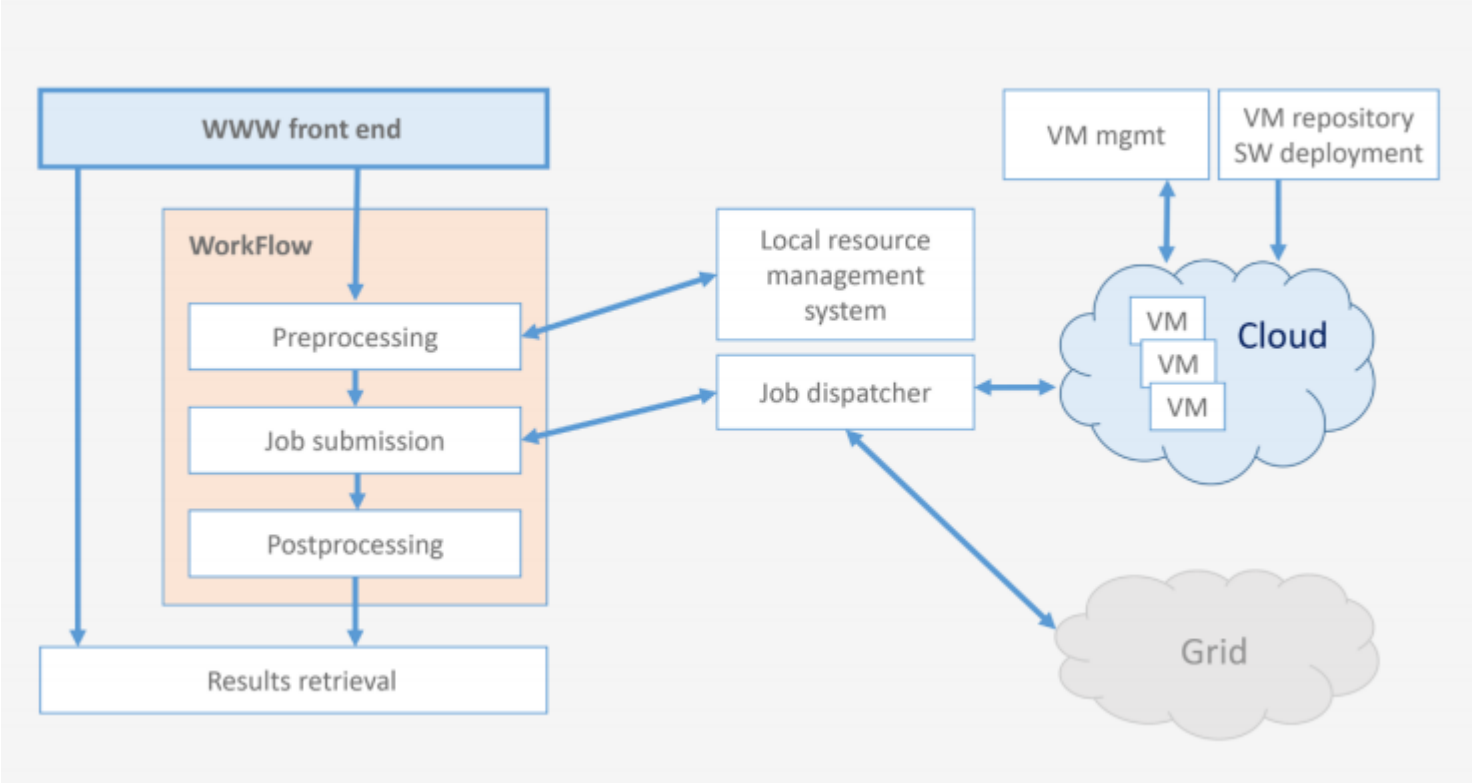
INFN role/contributions:

- Providing in kind computing power and storage from their grid and cloud resource centers federated with EGI
- Consolidating, operating and evolving the previous WeNMR computational platform into the new West-Life computational platform and services provisioned from the existing grid and cloud based e-infrastructures and projects
  - ✓ Consolidation of the job management mechanism
  - ✓ Programmatic access to structural biology datasets
  - ✓ Unified security and accounting model
- Providing technical support for:
  - ✓ the integration of the existing and newly developed scientific portals with the underlying computational platform
  - ✓ the creation of customized end-user VMs and/or containers encapsulating the applications
- Providing knowledge base and user support for the grid an cloud related issues.

First deliverables in July 2016

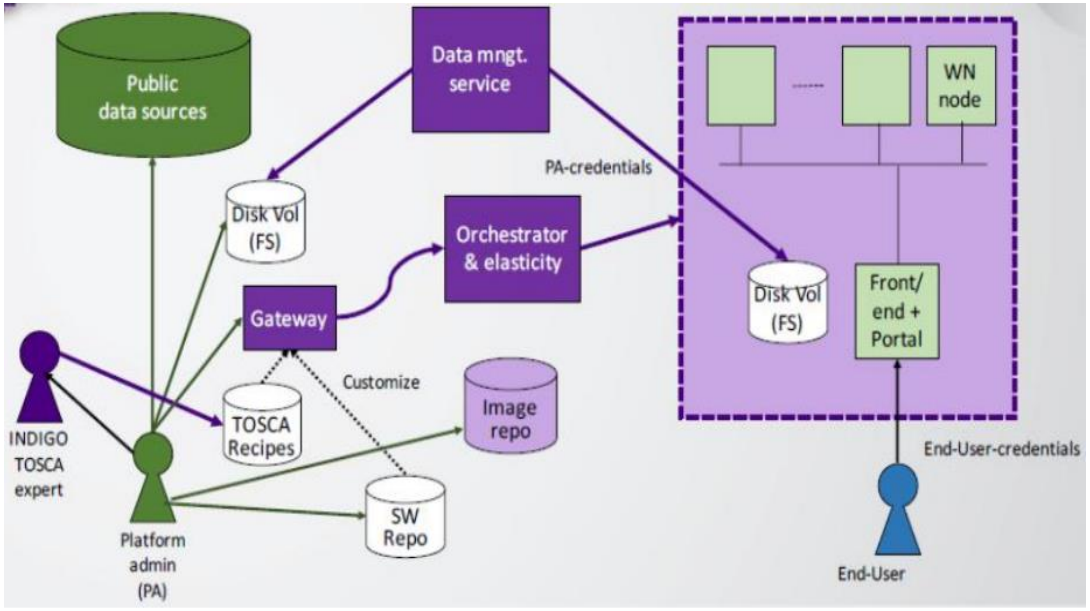
D4.1: Consolidated architecture of job submission and interaction with infrastructure

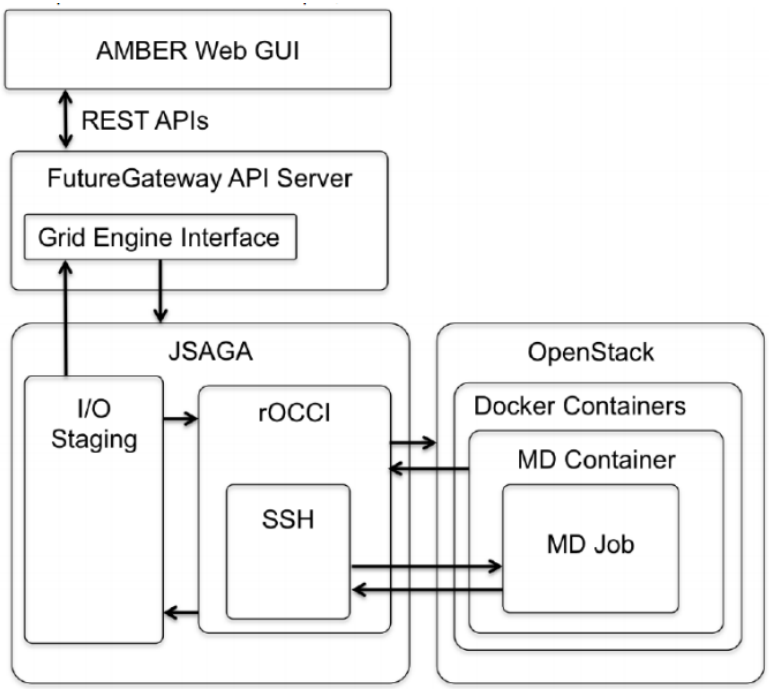
D4.2: Common security model design (based on trends from AARC, INDIGO and EGI)



West-Life Synergy with INDIGO-DataCloud

- CIRMMP and University of Utrecht are active partners of INDIGO too
- First prototypes with INDIGO pre-release (May 2016)
  - GROMACS, HADDOCK, PowerFit, DisVis, AMBER applications
  - Demos with Mesos/Marathon/Chronos and FutureGateway API



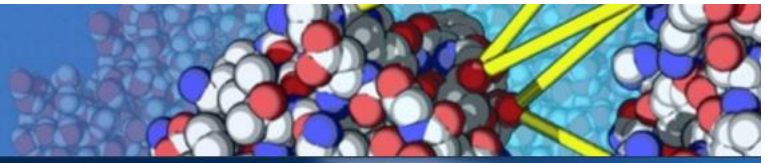


West-Life

Synergy with EGI-Engage

- EGI-Engage has a Competence Center to Serve Translational Research from Molecule to Brain: MoBrain (CIRMMP, UU, CSIC and MU are involved, + INFN-PD)
  - Dedicated task for “GPU portals for biomolecular simulations”
  - Molecular Dynamics (AMBER and GROMACS) and newly developed application (DisVis and PowerFit) exploiting GPGPU resources and executed through user-friendly web portals
- EGI-Engage has a JRA task led by INFN-PD for implementing an “Accelerated Computing Platform”:
  - GPGPU-enabled CREAM-CE
  - PCI-passthrough virtualization of GPGPUs in OpenStack and OpenNebula
- INDIGO has a task (led by LIP) for optimizing the use of Docker containers on GPGPU resources

DISVIS
@BonvinLab

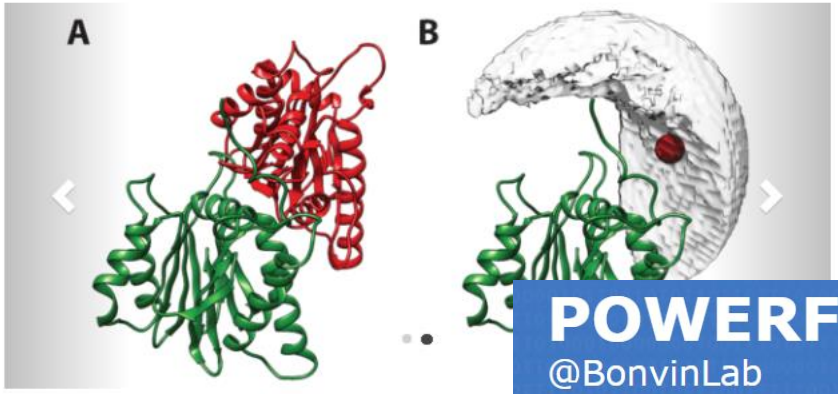


HADDOCK | CPORT | **DISVIS** | POWERFIT | PRODIGY | 3D-DART | BONVIN

About | Submit | Register | Example | Help/Manual | Support

WELCOME TO THE DISVIS WEBSERVER! >>

DisVis visualizes the accessible interaction space!



DisVis allows you to visualize and quantify the information content of dista complexes.

It performs a full and systematic 6 dimensional search of the three translational and number of complexes consistent with the restraints. In addition, it outputs the perce represents the center-of-mass position of the scanning chain corresponding to the hi in space.

DisVis webserver

REGISTRATION: To use the DisVis server you must have registered for an account. **here**

Submit your job to:

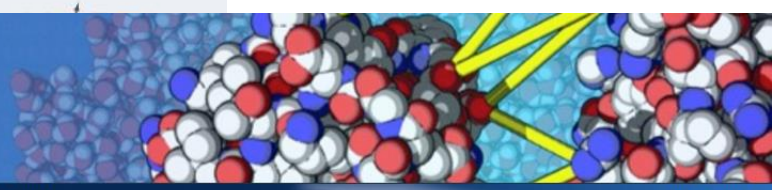
- DISVIS GPU accelarated Grid server
- DISVIS server

Open to users in August 2016

West-Life

Case	Machine	TimeGPU (sec)	TimeCPU 1 core	CPU1/GPU	
B-K40	Baremetal	674	7928	11.8	
K-K40	KVM	671	7996	11.9	<- Cloud
B-K20	Baremetal	830	11839	14.3	
D- K20	Docker	837	11926	14.3	<- Grid

POWERFIT
@BonvinLab



HADDOCK | CPORT | DISVIS | **POWERFIT** | PRODIGY | 3D-DART | BONVIN LAB

About | Submit | Register | Example | Help/Manual | Support

WELCOME TO THE POWERFIT WEBSERVER! >>

PowerFit fits your 3D structures in any map!

PowerFit automatically fits high-resolution atomic structures into cryo-EM densities.
To this end it performs a full-exhaustive 6-dimensional cross-correlation search between the atomic structure and the density. It takes as input an atomic structure in PDB-format and a cryo-EM density with its resolution; and outputs positions and rotations of the atomic structure corresponding to high correlation values. PowerFit uses the local cross-correlation function as its base score. The score can optionally be enhanced by a Laplace pre-filter and/or a core-weighted version to minimize overlapping densities from neighboring subunits.

PowerFit webserver

REGISTRATION: To use the PowerFit server you must have registered for an account. If you do not have an account yet you can **register here**

Submit your job to:

- POWERFIT server
- POWERFIT GPU accelarated Grid server

Reference for use of the server

Powered by

West-Life

bioexcel

INDIGO - DataCloud

MoBrain

egi

DisVis and PowerFit on EGI platforms

Requirements:

- **Basic:**
  - Python2.7
  - NumPy 1.8+
  - SciPy
  - GCC (or another C-compiler)
- **Optional for faster CPU version:**
  - FFTW3
  - pyFFTW
- **Optional for GPU version:**
  - OpenCL1.1+
  - pyopencl
  - clFFT
  - gpyfft

Solution for grid and cloud computing:

INDIGO - DataCloud

Docker containers built with proper libraries and OpenCL support:

Base dockerfile with NVIDIA driver:
https://github.com/LIP-Computing/ansible-role-nvidia

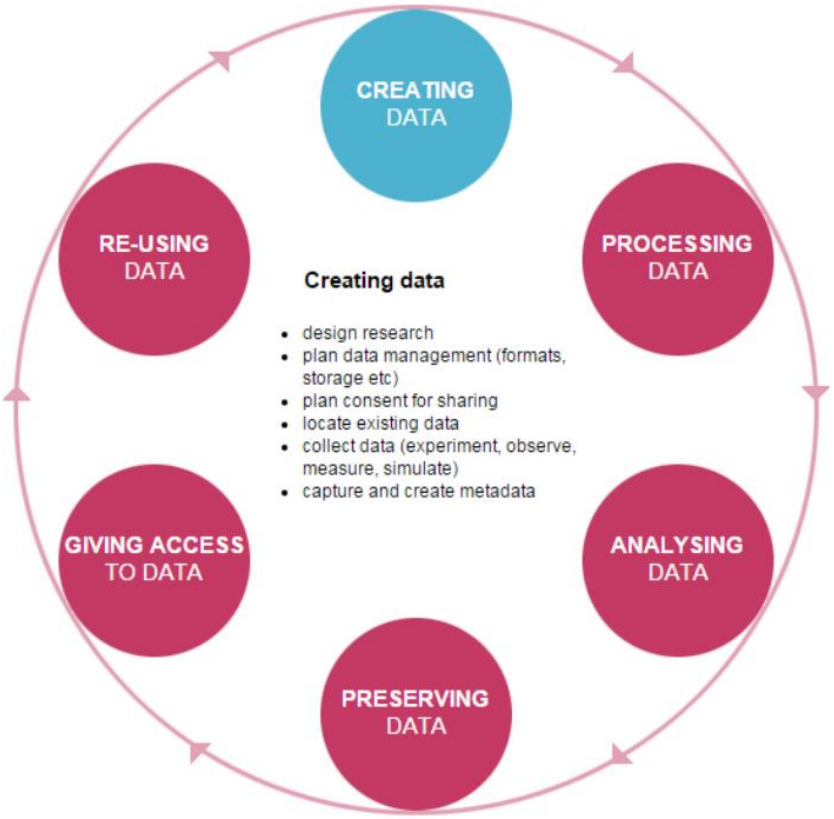
Dockerfile for DisVis and PowerFit:
https://github.com/indigo-dc/ansible-role-disvis-powerfit

Docker engine not required on grid WNs: use udocker tool to run docker containers in user space (https://github.com/indigo-dc/udocker)

GPGPU-enabled CREAM-CE at CIRMMP

Data management and lifecycle

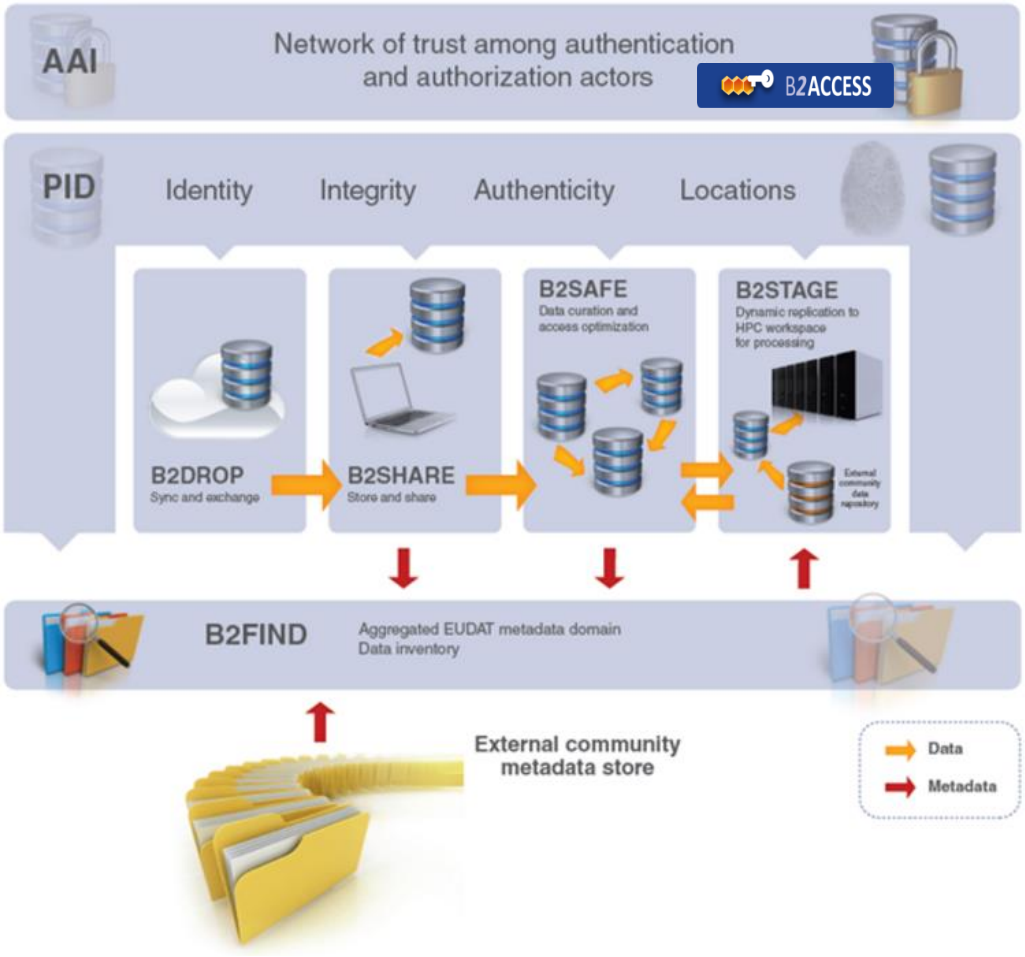
- > 1PB of data in 2015 from Diamond Light Source only
- Combined output of European SB facilities > LHC, and XFEL will double it
- Raw data acquired at experiments:
  - NMR spectroscopy (MBs to GBs)
  - X-ray diffraction at a synchrotron or home source (GBs)
  - Electron microscopy (TBs)
- + metadata describing the experiment and the provenance of the sample (e.g. how the protein was created and purified)
- then a series of processing steps will:
  - reduce experimental data
  - determine the structure
- Paper publication requires reduced data deposition in PDB/EMDB



http://www.data-archive.ac.uk/

Data management and lifecycle

- Link with EUDAT through STFC → West-Life/EUDAT Pilot
- Link with RDA – Structural Biology Interest Group (chaired by STFC and CIRMMP)



EUDAT

West-Life

EUDAT service uptake

The West-Life Pilot will rely on the following EUDAT services:

- **B2DROP for WebDAV**
- **B2ACCESS: integrate with Instruct AAI**
- **B2SHARE, when it supports WebDAV**
- **B2NOTE, including PROV-O support**
- **(If required) B2STAGE for performant transfer of large datasets**

Possible contribution from INFN/INDIGO:

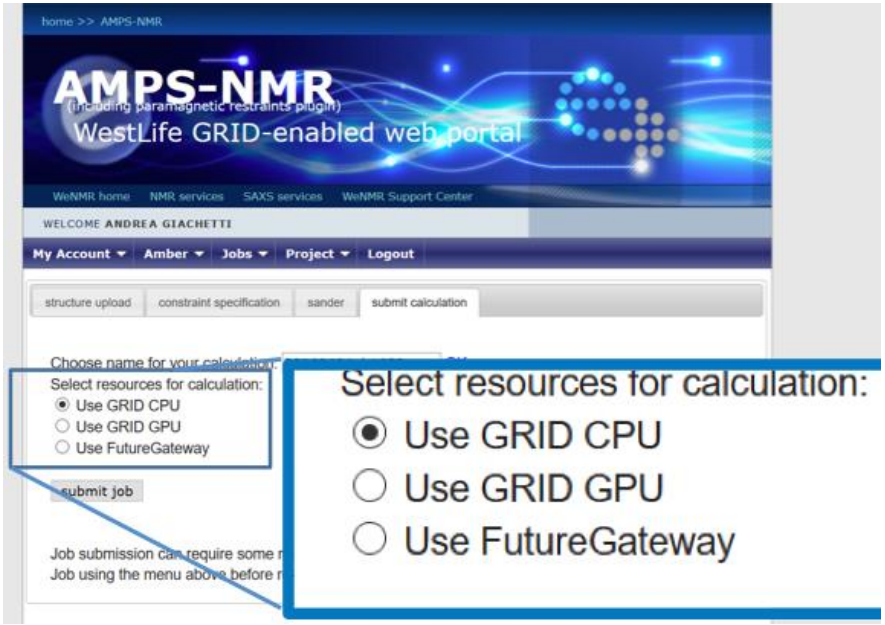
- **OneData comparison with B2DROP/B2SHARE/B2STAGE**

Backup Slides

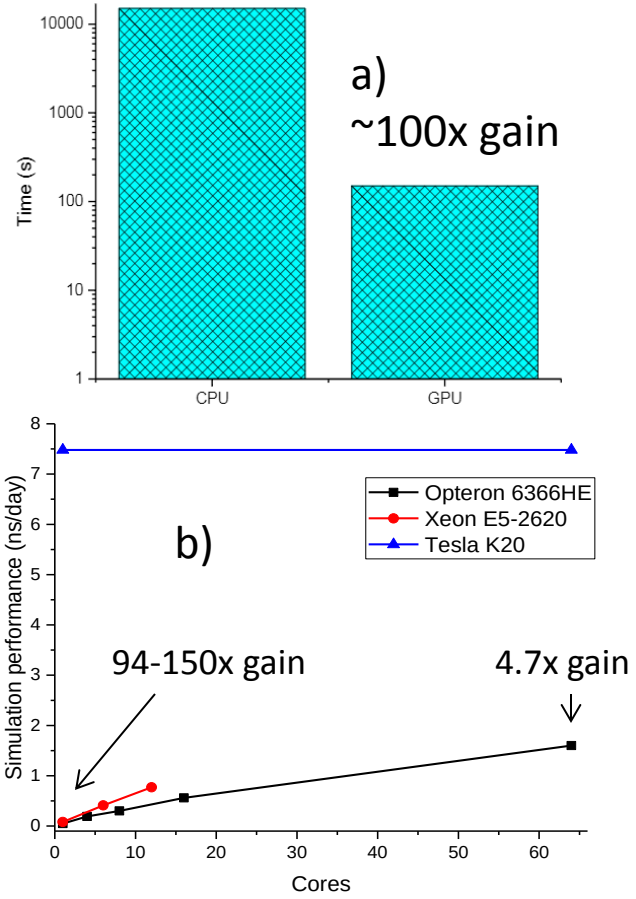
AMBER

a) Restrained (rMD) Energy Minimization on NMR Structures

b) Free MD simulations of ferritin



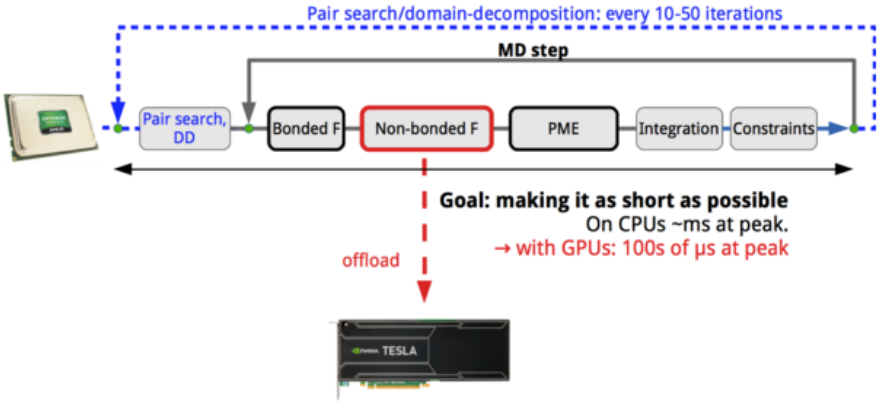
WeNMR/AMBER grid portal can now exploit GPU resources

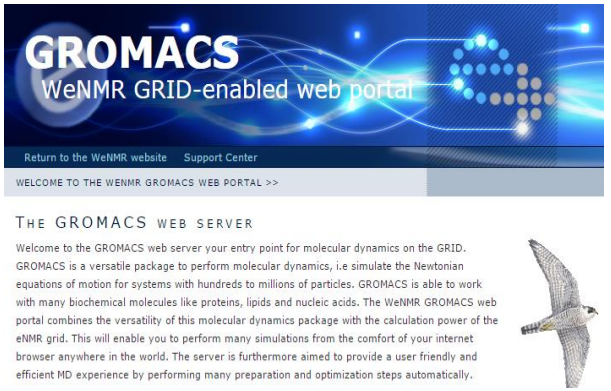


GROMACS

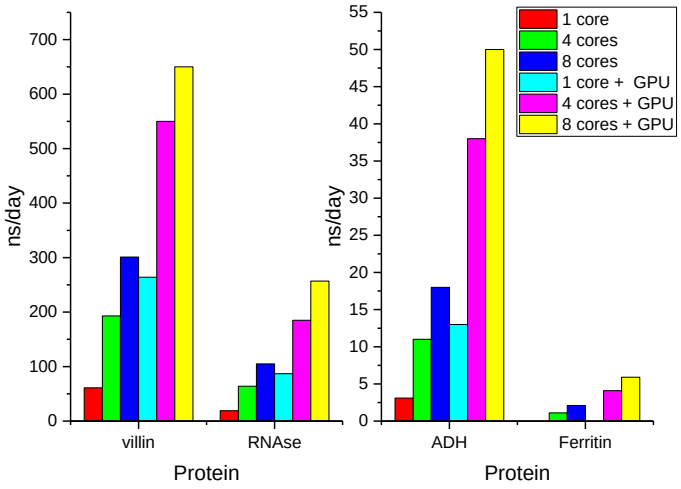
- GPU acceleration introduced in version v4.5
  - Grid portal runs it in multi-threading mode
  - No significant cloud overhead measured

  for GPU speedups

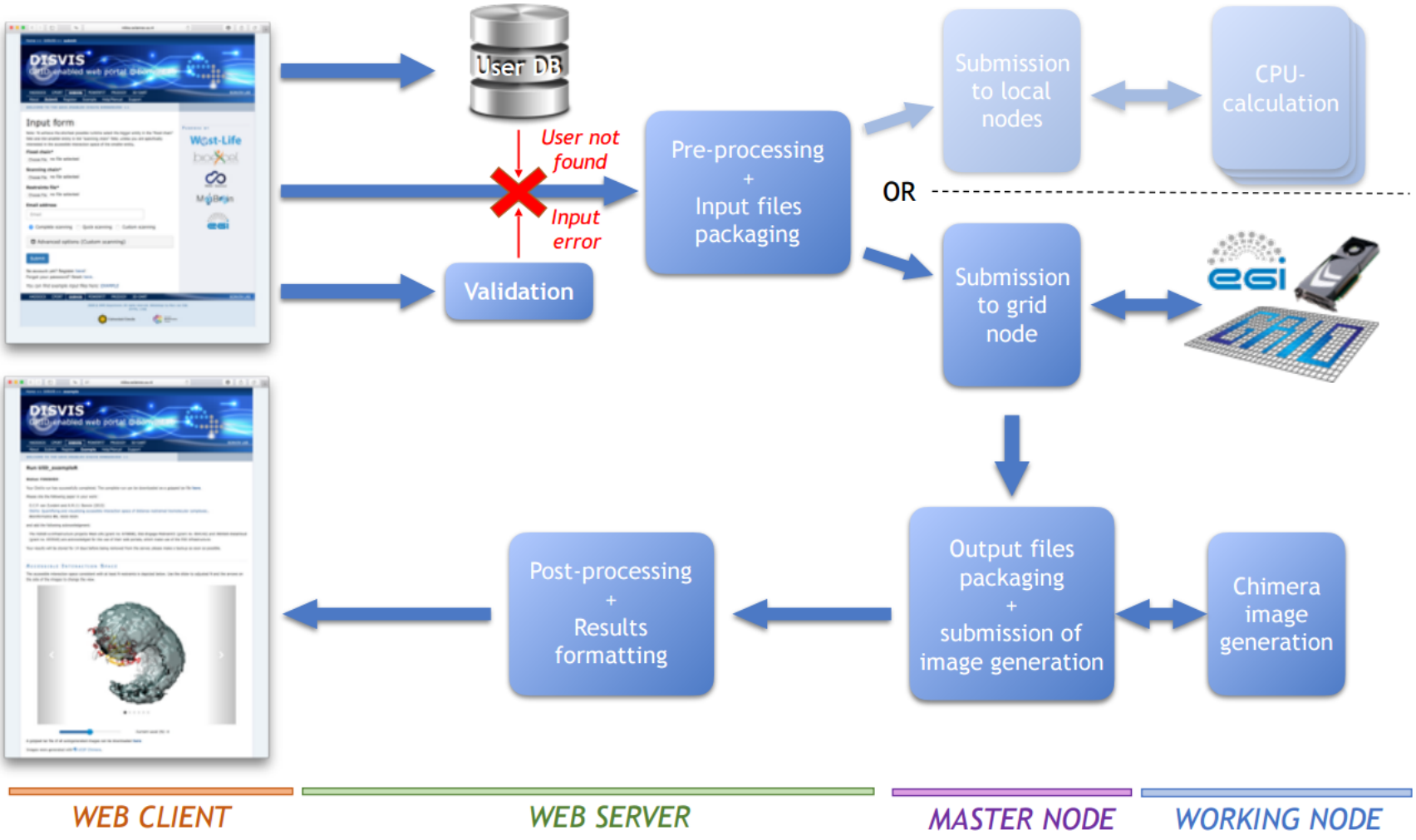




		Simulation performance in ns/day						GPU Acceleration		
Dataset	Protein size (aa)	1 core	4 cores	8 cores	1 core + GPU	4 cores + GPU	8 cores + GPU	1 core	4 cores	8 cores
Villin	35	61	193	301	264	550	650	4.3x	2.8x	2.2x
RNAse	126	19	64	105	87	185	257	4.6x	2.9x	2.4x
ADH	1,408	3.1	11	18	13	38	50	4.2x	3.5x	2.8x
Ferritin	4,200	-	1.1	2.1	-	4.1	5.9	-	3.7x	2.8x



Grid-enabled web portal architecture



DisVis job example

```
#!/bin/sh
driver=$(nvidia-smi | awk '/Driver Version/ {print $6}')
export WDIR=`pwd`
git clone https://github.com/indigo-dc/udocker
cd udocker
./udocker.py pull indigodatacloudapps/disvis:nvdrv_$driver
rnd=$RANDOM
./udocker.py create --name=disvis-$rnd indigodatacloudapps/disvis:nvdrv_$driver
mkdir $WDIR/out
./udocker.py run -v /dev --volume=$WDIR:/home disvis-$rnd "disvis \
/home/O14250.pdb /home/Q9UT97.pdb /home/restraints.dat -g -a 5.27 -vs 1 \
-d /home/out"
./udocker.py rm disvis-$rnd
./udocker.py rmi indigodatacloudapps/disvis:nvdrv_$driver
cd $WDIR
tar zcvf res-gpu.tgz out/
```

Driver identification

Install udocker tool

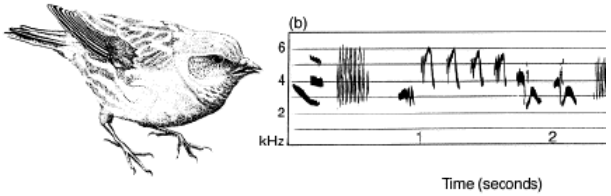
Pull DisVis image

Create the container

Run the container executing DisVis

Biodiversity

- LifeWatch is the European e-Science infrastructure for Biodiversity and Ecosystem Research (ESFRI)
- ANN & Pattern Recognition Tools can be applied in many cases:
  - Bird recognition (by sound)
  - Satellites data (land type, land use, water…)
  - Species classification



- Due to different features, like memory bandwidth or architecture, GPUs get much better performance in training ANN than CPUs
- They adopt Caffe: one of the most popular deep learning frameworks, implemented in pure C++/CUDA

  http://caffe.berkeleyvision.org

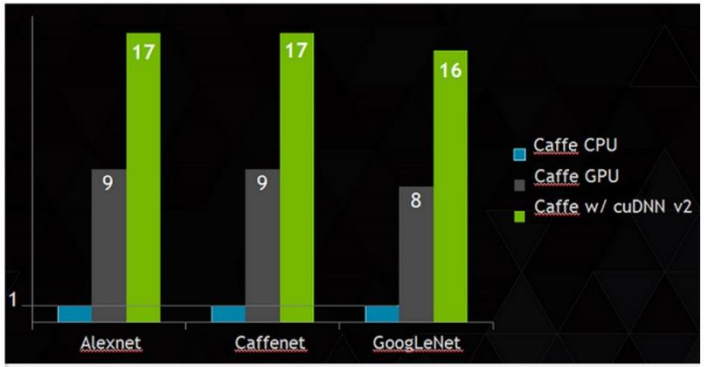


Figure 1: cuDNN performance comparison with CAFFE, using several well known networks. CPU is 16-core Intel Haswell E5-2698 2.3 GHz with 3.6 GHz Turbo. GPU is NVIDIA GeForce GTX TITAN X.