The HADDOCK WeNMR portal: Combing gLite, DIRAC4EGI and crowd computing. Cloud next?

ISGC 2015, Taipei

Alexandre M.J.J. Bonvin
Bijvoet Center for Biomolecular Research
Faculty of Science, Utrecht University
the Netherlands
a.m.j.j.bonvin@uu.nl
A worldwide e-Infrastructure for NMR and structural biology

WeNMR VRC (Jan 2015)

• The 1st VRC officially recognized by EGI
• One of the largest (#users) VO in life sciences
• > 1600 VRC members (>60% outside EU)
• of which >680 VO-registered users (with certificates)
• ~ 100 000 CPU cores (from 41 sites)
• > 2000 SI2K CPU years
• > 5M jobs
• User-friendly access to Grid via web portals

www.wenmr.eu
Serving ESFRI INSTRUCT
A glimpse of the WeNMR services portfolio

User-friendly solutions

Predefined applications and workflows

Added value to the data

User support (also via EGI Distributed Competence Center and GGUS portal)

Make use of robot certificates

VRC-specific SSO solutions (direct access to portals via VRC credentials)
WeNMR VRC users distribution

Over 1600 VRC (www.wenmr.eu) members (64% / 36% outside Europe)!
(Jan 2015)

Grid Usage statistics

Enabled via EGI federated operation

Statistics http://accounting.egi.eu
Molecular Docking

\[ E_{\text{elec}} = \frac{q_i q_j}{4 \pi \varepsilon_0 r} \]

\[ d_{\text{eff}}^{AB} = \left( \sum_{n_k=1}^{\text{N atoms}} \sum_{k=1}^{\text{N res A}} \sum_{n_{AB}=1}^{\text{N atoms}} \frac{1}{d_{nm_{AB}}^6} \right)^{-\frac{1}{6}} \]

\[ E_{\text{vdW}} = 4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \]
HADDOCK web portal

- Was existing before the VRC, running on local resources only
- > 5000 registered users (both from WeNMR VRC and HADDOCK direct registration) (heavily cited!)
- > 92000 served runs since June 2008
- > 20% on the GRID
- Increased usage enabled through EGI resources!

HADDOCK (High Ambiguity Driven protein-protein DOCKing) is an information-driven flexible docking approach for the modeling of biomolecular complexes. HADDOCK distinguishes itself from ab-initio docking methods in the fact that it encodes information from identified or predicted protein interfaces in ambiguous interaction restraints (AIRs) to drive the docking process. HADDOCK can deal with a large class of modeling problems including protein-protein, protein-nucleic acids and protein-ligand complexes.

More information about HADDOCK can be found on the HADDOCK website

HADDOCK WEBSERVER
To use the HADDOCK eNMR GRID-enabled docking server you must:
- have registered for a GRID-enabled HADDOCK account
- have registered with the eNMR grid infrastructure.

Note: registration does require a valid grid certificate!!!

- HADDOCK server: the easy interface
- HADDOCK server: the prediction interface
- HADDOCK server: the expert interface
- HADDOCK server: the guru interface
- HADDOCK server: the multi-body interface
- HADDOCK server: the refinement interface
- HADDOCK server: the file upload interface
- HADDOCK server tool: generate AIR files for multibody docking

De Vries et al. Nature Prot. 2010
HADDOCK server users distribution

Over 5250 registered (250 via WeNMR SSO)
(now all have access to the grid-server via robot cert)
What is happening behind the scene?
The deamons behind the portal

WEB

SERVER

UI

GRID

Service interface:
Input validation
Job packaging

Status page:
Status report
Presentation of results

Executable file
Pre-processing
Job packaging

Job dropping
Result

Executable file
Parameter files
Library files

Executable file
Pre-processing
Job packaging

Job dropping
Result

Executable file
Pre-processing
Job packaging

Job dropping
Result

Executable file
Pre-processing
Job packaging

Job dropping
Result

WEB SERVER UI GRID
**Data / jobs volume**

**Data transfer / jobs**

**User level** – data transfer via web interface:
- **Data input:** ~10 kB to 10 MB
- **Data output:** ~100 MB to a few GB

**Local cluster** – data/jobs transfer:
- **Job submission:** ~10 to 50 single jobs
- **Data:** a few GB

**Grid level** – data/jobs transfer  
(UI<->WMS<->CE<->WN):
- **Job submission:** ~250 to 2500 single jobs
- **Data transfer:** ~1 to 10 MB
- **Output:** ~1-20 MB per job
HADDOCK Grid submission

- Each job in the pool directory consists of:
  - Shell script to be executed on WN
  - Compressed tar archive containing the data/directory structure for the job
  - File specifying the location where results should be copied
  - JDL script with requirements/ranks for selecting suitable sites, e.g.:

  ```
  Requirements = (Member("OSG_VERSION", other.GlueHostApplicationSoftwareRunTimeEnvironment) ||
                  (other.GlueCEPolicyMaxWALLTime < 720 && other.GlueCEPolicyMaxWALLTime > 110 &&
                   Member("VO-enmr.eu-CNS1.2", other.GlueHostApplicationSoftwareRunTimeEnvironment)))
  
  Rank = ( other.GlueCEStateFreeJobSlots > 0 ? other.GlueCEStateFreeJobSlots :
           (-other.GlueCEStateWaitingJobs * 4 / ( other.GlueCEStateRunningJobs + 1 )) - 1 );
  ```

- Software is remotely deployed and CE tagged with a software tag:
  - For HADDOCK: VO-enmr.eu-CNS1.2

- Submission via robot proxie – fully automated – no user/operator intervention.
HADDOCK goes DIRAC

- DIRAC submission enabled at minimum cost!
  - In one afternoon, thanks to the help of Ricardo and Andrei
  - Clone of the HADDOCK server on a different machine
  - No root access required, no EMI software installation required
- Minimal changes to our submission and polling scripts
  - Requirements and ranking no longer needed, only CPUTime

```plaintext
JobName = "dirac-xxx";
CPUTime = 100000;
Executable = "dirac-xxx.sh";
StdOutput = "dirac-xxx.out";
StdError = "dirac-xxx.err";
InputSandbox = {"dirac-xxx.sh","dirac-xxx.tar.gz"};
OutputSandbox = {"dirac-xxx.out", "dirac-xxx.err","dirac-xxx-result.tar.gz"};
```

- Very efficient submission, high job throughput, currently most used submission mechanism by the HADDOCK portal
HADDOCK goes crowd computing

• IDGF resources enabled at minimum cost!
  – IDGF CE is not (yet) discoverable in normal EGI resources
  – Requires direct submission to CE
  – Simple change to JDL script:

  SubmitTo = "cr2.edgi-grid.eu:8443/cream-pbs-homeboinc"

• Pre-deployed mini Linux VMs on BOINC resources with the software required for running HADDOCK jobs (thanks Jozsef!)

• Currently every 10th HADDOCK grid job is submitted to IDGF resources and duplicated to ensure a high throughput of the HADDOCK server

• In case of failure automatically redirected to normal grid resources
HADDOCK web portal: Making best use of Grid solutions

HADDOCK portal statistics (grid jobs only)

Currently >80% of the submissions are directed transparently to EGI resources

86%/14% distribution DIRAC/gLite submission

~4% of gLite jobs on OSG resources

Direct gLite-based submission

Crowd computing

-> pushing jobs to the grid

Dirac4EGI service

-> pulling jobs to the grid via pilots

Cumulative #jobs

<table>
<thead>
<tr>
<th>Year</th>
<th>#jobs (gLite)</th>
<th>#jobs (IDGF)</th>
<th>#jobs (DIRAC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>1E+00</td>
<td>1E+00</td>
<td>1E+00</td>
</tr>
<tr>
<td>2015</td>
<td>9E+06</td>
<td>8E+06</td>
<td>7E+06</td>
</tr>
<tr>
<td>2016</td>
<td>6E+06</td>
<td>5E+06</td>
<td>4E+06</td>
</tr>
<tr>
<td>2017</td>
<td>3E+06</td>
<td>2E+06</td>
<td>1E+06</td>
</tr>
<tr>
<td>2018</td>
<td>9E+06</td>
<td>8E+06</td>
<td>7E+06</td>
</tr>
<tr>
<td>2019</td>
<td>6E+06</td>
<td>5E+06</td>
<td>4E+06</td>
</tr>
<tr>
<td>2020</td>
<td>3E+06</td>
<td>2E+06</td>
<td>1E+06</td>
</tr>
<tr>
<td>2021</td>
<td>9E+06</td>
<td>8E+06</td>
<td>7E+06</td>
</tr>
</tbody>
</table>

#jobs (gLite)
#jobs (IDGF)
#jobs (DIRAC)

Cumulative #jobs
Reliability

<gLite success %> 78 ± 16  
<IDGF success %> 78 ± 23  
<DIRAC success %> 96 ± 3
Conclusions

- Successful and smooth porting of the HADDOCK portal to both IDGF resources
- DIRAC4EGI submission shows higher performance / reliability than regular gLite-based submission
- Currently almost all HADDOCK grid-enabled portals redirected to DIRAC

- In the future we aim at packaging the complete server in cloud machines and associated compute resources
Conclusions - WeNMR

- Lively, active and still growing community
- Innovative and user-friendly e-Science solutions
- Perfect use of high-throughput data analysis (i.e. Grid)
- Success and sustainability made possible through EGI services (resources, accounting, support, dissemination...)
Future plans

- Bring the micro (WeNMR) and macro (N4U) worlds together into one competence center under future EGI activities:

  MoBRAIN

- With activities toward:
  - Integrating the communities
  - Making best use of cloud resources
  - Bringing data to the cloud (cryo-EM)
  - Exploiting GPGPU resources

- While maintaining the quality of our current services!
Acknowledgments

Arne Visscher

DDSG

VICI

NCF (BigGrid)

BioNMR

WeNMR

HADDOCK Inc.

HADDOCK Inc.

DIRAC4EGI VT

DIRAC4EGI VT

Ricardo Graciani, Andrei Tsaregorodtsev, ...

Ricardo Graciani, Andrei Tsaregorodtsev, ...

HADDOCK Inc.

Gydo van Zundert, Charleen Don, Adrien Melquiond, Ezgi Karaca, Marc van Dijk, Joao Rodrigues, Mikael Trellet, ..., Koen Visscher, Manisha Anandbahadoer, Christophe Schmitz, Panos Kastritis, Jeff Grinstead

Gydo van Zundert, Charleen Don, Adrien Melquiond, Ezgi Karaca, Marc van Dijk, Joao Rodrigues, Mikael Trellet, ..., Koen Visscher, Manisha Anandbahadoer, Christophe Schmitz, Panos Kastritis, Jeff Grinstead
The End

Thank you for your attention!

HADDOCK online:
http://haddock.science.uu.nl
http://bonvinlab.org/software
http://www.wenmr.eu