MPI WG Proceedings

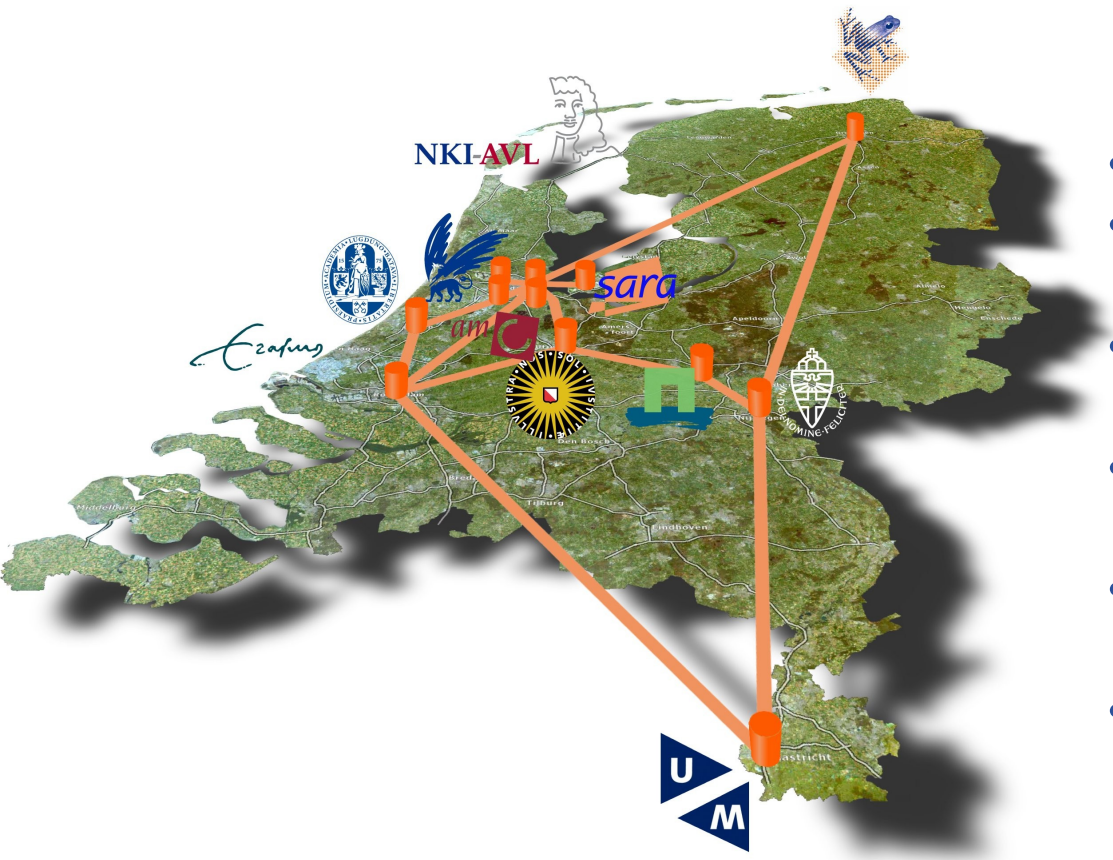
Jeroen Engelberts
SARA Reken- en Netwerkdiensten
Amsterdam, The Netherlands

Agenda

Agenda

- Introduction
- Actions of the MPI WG
- Goals/achievements previous MPI WG
- Goals of current MPI WG
- Find out why MPI not abundant on EGEE
- Increasing versatility MPI in gLite
- Current package
- On the ease of installation
- Recommendations

LifeScience Grid



- Consultant at SARA
- Computational chemist by education
- Assisting users on SARA infrastructure
- Technical project leader LifeScience Grid
- Interested in MPI because of request from LSG community
- Became the new chair man

What did we do?

- Assembled a team of MPI WG members
- Organized two phone confs
- Organized a face to face meeting
- Sent out a survey both to users and sys admins
- Produced a rough draft
- Sent additional ideas to TMB via mail
- Presented the ideas at the Multi-core Workshop at CERN, June 2009

-> What was already there when we started?

Achievements previous WG

- Recommended an MPI implementation that required very few middleware modifications
- Info systems advertise the available capabilities/flavors
- Environment variables are set to help the user
- A step by step installation procedure has been written for the site administrators
- Use of Jobtype="MPICH" is deprecated. Nowadays, normal jobs can handle multiple cores (Cpunumber=*x*)
- And as a result: of the 455 unique CE's from a top level BDII 120 advertise to be MPI enabled (situation 23-6-2009)

Goals of the MPI WG

- Recommend a method for deploying MPI support for users and site administrators
- Recommend on a method to enable the possibility for a user to request all cores on one physical machine
- And/or, the possibility to request the job to run on as little separate machines as possible
- Reduce the amount of MPI flavors to a minimum (OpenMPI and MPICH-2)
- Recommendations should extend to other methods of parallel computing as well, like OpenMP
- Revive the MPI SAM tests
- Figure out why so few sites have MPI enabled via gLite on their clusters
- Future support of MPI in EGEE/EGI

Current Package Issues

Issues

- Current MPI package has been fixed quickly, but rather *ad hoc*

Recommendations

- Long(er) term support should be arranged (EGI/NGI)
- Before middleware and package updates are released, MPI functionality should be tested and certified

A Single (Meta) Package

Issues

- Large part of sys admins considers installation “difficult” or “extremely hard”
- Currently the installation recipe is quite long

Recommendations

- Create a single metapackage for basic/generic MPI/OpenMP functionality
- Make a single Yaim function for it
- Keep an eye on performance. Benefits of special hardware, *eg* Infiniband, may be lost

JobType and Scheduling

Issues

- User has no control over distribution of cores
- Hence, OpenMP can not be run efficiently or fairly
- On PBS Cpunumber=*x* translates to -lnodes=*x*

Recommendations

- Make an additional JDL keyword, like granularity
- User must be able to request complete nodes without specifying/knowing the total amount of cores
- Fix the problem with Cpunumber in “Normal” jobs, because #nodes ≠ #cores nowadays

MPI SAM Test

Issues

- OK, 120 clusters are MPI enabled via the Grid layer. Is MPI really working?
- Earlier MPI SAM tests indicated sites having an MPI problem, while in fact they didn't.

Recommendations

- Revive the MPI SAM test project
- Make it easier (only check for presence of environment variables and libraries/binaries)
- Encourage, or even enforce MPI enabled sites to run MPI SAM tests

MPI-Start/mpiexec

- MPI-Start contains a set of wrapper scripts which get a job from the WMS onto the WN
- mpiexec is a script replacing the binary to fix the assumption that all MPI jobs start with “mpiexec”

Issues

- For MPI-Start: “Note that Int.EU.Grid officially finishes at the end of April 2008. However, the Savannah server will be maintained also after that.”

Recommendations

- Since these scripts are *vital* for the functionality of MPI on the gLite middleware support (preferably also after EGEE III) must be arranged.

What the users say

Recommendations from the “user”

- Organize workshops about porting typical MPI applications on the Grid
- Write documentation for the Web
- Setup a Specialized Support Center
- Periodically check that MPI is properly working on the sites (Add MPI tests on SAM)
- Make sure that most common MPI versions are supported
- MPI should be installed in standard way on all the sites of the Grid
- Convince Grid sites to support MPI in the first place
- Go on with the MPI-Start wrapper script initiative; make sure MPI jobs launched with mpi-start-wrapper will eventually succeed

Summary

Recommendations

- Make the installation easier (Yaim function)
- Check MPI functionality on a regular basis (SAM test)
- Check MPI functionality *after*, or better, *before* upgrading
- Enable support for the wrapper scripts, installation procedure (preferably also after EGEE III has ended)

Follow our progression

http://www.grid.ie/mpi/wiki

The Future of MPI on EGEE

Recommendations (personal)

- Don't setup another MPI WG with a small scope, but rather an MPI TF (based on diversity of e-mails)
- The MPI TF should be headed by an experienced person

Plans for the MPI WG

- Assemble bits and pieces from drafts and e-mails
- Pre-final version will be sent to MPI WG members
- Have a final phone conf to finalize the recommendation document
- Dissolve the current MPI WG

Acknowledgements

Members of the MPI Working Group (alphabetical)

- Roberto Alfieri (INFN, I)
- Roberto Barbera (INFN, I)
- Ugo Becciani (INAF, I)
- Stefano Cozzini (Democritos, I)
- Dennis van Dok (Nikhef, NL)
- Fokke Dijkstra (Groningen University, NL)
- Karolis Eigilis (Baltic Grid)
- Jeroen Engelberts (SARA, NL)
- Francesco De Giorgi (Democritos, I)
- Oliver Keeble (CERN)
- Vangelis Koustis (PhD student, Greece)
- Alvarez Lopez (CSIC, ES)
- Salvatore Montforte (INFN, I)
- John Ryan (Trinity College Dublin, IRL)
- Mehdi Sheikhalishahi (PhD Student, Iran)
- Steve Traylen (CERN)