Simple, fault tolerant, lightweight grid computing approach for bag-of-tasks applications

Yiannis Georgiou, Nicolas Capit, Bruno Bzeznik and Olivier Richard

3rd EGEE User Forum



Grenoble, FRANCE Mescal Project









1 CIGRI: lightweight grid computing approach

- Pault Treatment and Turnaround time Optimisations
- Evaluating CIGRI upon Grid5000 platform

Conclusions

Plan

CIGRI: lightweight grid computing approach

- Motivations and Related Work
- CIGRI lightweight grid system: Principal Concepts
- Global Architecture

2) Fault Treatment and Turnaround time Optimisations

3 Evaluating CIGRI upon Grid5000 platform

4 Conclusions

Motivations and Related Work

CIMENT project: **Mutualise the computing power** of Rhone-Alpes regional private laboratories **cluster resources** from different disciplines (environment, chemistry, physics, astronomy, medecine, ...) to effectuate **larger scale computations**



- Option of Globus: complicated, expensive (ex. Condor-G, Nimrod-G,...)
- Emergence of desktop grid systems and the idea of cycle stealing technologies provided the good bases for the CIGRI approach
- Alternative grid solutions: Condor (..low security,.. parallel applications), OurGrid (..high security, .."BoT" applications)

CIGRI approach for grid computing

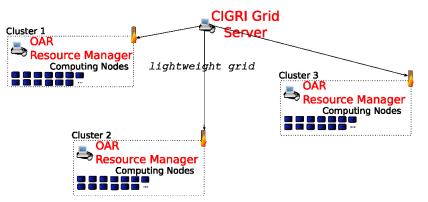
- CIGRI: simpler, lightweight approach of grid platform (low security,..only "Bag-of-Tasks" applications (set of independent tasks))
- Using the method of aggregation of idle cluster resources
- Platform focuses on research and development of problems that come along with the execution of tasks (scheduling, fault tolerance,...)
- Choices of simple solutions on important classic grid issues like security, authentication mechanisms , resource location...





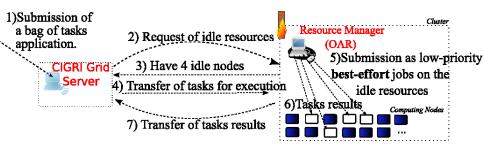
CIGRI: Lightweight grid

 Homogenisation of services and administration procedures between clusters



CIGRI: Cluster utilisation policy

- Works discreetly along with the interconnected clusters: No specific CIGRI software installed on clusters
- **Besteffort jobs**: job type provided by the cluster resource management system (introduced in OAR)
 - Lowest priority jobs submitted only if there is a free resource



• Killed when local cluster job requests the resource

CIGRI Global Architecture

- **High level componennts**: MySQL Database, Perl programming language
- No specific software installed on clusters: Based on linux system commands (bash, ssh, scp, tar, rsync, ...)
- Integrated to function with OAR resource management system
- Modular Architecture: Easy for development and research

OAR: A cluster resource management system with high level components

- Simple, robust and scalable
- High level tools (scripting language, relational database engine)
- Designed to be usable as a production resource management system (ex: GRID5000)
- Performance similar with complex resource management systems (LSF,Torque+Maui,SGE,...)

[Ref: Nicolas Capit et Al.(2005)

A batch scheduler with high level components.(CCGrid05)]

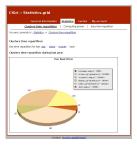




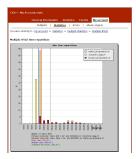
Global Architecture

CIGRI Grid Functionalities

• Web Portal for grid monitoring (+statistics)



| your leges log-grid | | without water | | | | | |
|--|--|---|--|--|--|--|--|
| Details of the Multijobs 171 | | | | | | | |
| Parameters in execution | | | | | | | |
| | | | | | | | |
| country 104 | aler harre (e) | increasing. | | | | | |
| bearing into | | | | | | | |
| econdary bey 1.4 | - 8 | Increasing decreasing | | | | | |
| | VHE | | | | | | |
| | | | | | | | |
| the cumber of perso | sunder of personners in encursion, 21 | | | | | | |
| | | | | | | | |
| | | | | | | | |
| CLOP Mane | 10-114T | 307400 | | | | | |
| bioineos inea fr | 2008-08-15 | CribitOT04_15,mat 04395 satisficT04_15,mat 04395 | | | | | |
| | 2009/06/15 | GENERAL ILLING ONCO | | | | | |
| DESIGN/P.37A3.2 | 89.31.95 | | | | | | |
| tertimestr. | 2004-06-15 | orbitO714_13.mat 04337 orbitO714_13.mat 04238 | | | | | |
| terne ut-prenetie. | 10.20.10 | centering 13, national on the | | | | | |
| | | | | | | | |
| DODADA JOANST | 2004-08-15 | SIDERTS A TRUNK ONDER | | | | | |
| | 10-3812 | info@0754_13.mail.06343 info@0754_13.mail.06343 | | | | | |
| | | | | | | | |
| | | | | | | | |
| and much | 2004-06-15 | | | | | | |
| | 2004-08-15 | brbBOP14_13.mat 04345 brbBDP14_15.mat 04345 | | | | | |
| toot map in | 2004-08-15 30-28-20 2004-08-15 30-28-21 | 0/58/214_13.mat (4045 0/58/214_13.mat (4045 0/58/214_13.mat (4047 0/58/214_13.mat (4047 | | | | | |
| (000.300a).3' | 2004-06-15 m-28-20 2004-06-15 m-28-21 2004-06-15 2004-06-18 | brbsK/14,13,mat 04345 brbst/174,13,mat 04345 brbst/174,13,mat 04345 brbst/174,13,mat 04345 brbst/174,13,mat 04348 brbst/174,13,mat 04348 | | | | | |
| Cost image? List image? | 2008-08-15 10-28-20 2008-08-15 10-28-21 2008-08-15 10-38-21 2008-08-18 | between the second seco | | | | | |
| (000.300a).3' | 2004-06-15 30-26-27 2004-06-15 30-26-21 2004-06-15 30-36-21 2004-06-15 | orbell/164_13.mat (4045 orbell/164_13.mat (4045 orbell/164_13.mat (4046 orbell/164_13.mat (4046 orbell/164_13.mat (4056 orbell/164_13.mat (4056 orbell/164_13.mat (4056 | | | | | |
| Lost Imap 3' Lost Imap 3' | 2008-08-15 30-28-20 2008-08-15 3008-08-15 3008-08-15 3008-08-15 3008-08-15 | EMBNYOL, J. Kural Galad GREETYA, J. Kural Galad | | | | | |
| Cost image? List image? | 2004-08-15 30-28-22 2004-08-15 30-38-21 2004-08-13 30-38-21 2004-08-13 30-28-21 2004-08-15 30-28-25 2004-08-15 30-28-25 | beth/Tel.3 kmid (e045 bith/Tel.3 kmid (e045 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046 bith/Tel.3 kmid (e046) bith/Tel.3 kmi | | | | | |
| Lost Imap 3' Lost Imap 3' | 2004-08-15 m-26-29 2004-08-15 30-36-21 2004-08-15 30-36-21 2004-08-15 30-36-28 2004-08-15 30-36-28 2004-08-15 30-36-28 | Interview, James 40445 solutions, James 40445 solutions, James 40445 solutions, James 40445 solutions, James 40455 solutions, James 40455 soluti | | | | | |
| Lost Imag 3* Lost Imag 3* Lost Imag 3* Lost Imag 3* Lost Imag 3* | 2004-08-15 m-26-29 2004-08-15 so-26-21 2004-08-15 so-26-21 2004-08-15 so-26-25 2004-08-15 so-26-25 2004-08-15 so-26-25 2004-08-15 so-26-25 2004-08-15 so-26-25 2004-08-15 | eth/Vis_3 hms/6445 priot/Vis_3 hms/6445 priot/Vis_3 hms/6446 priot/Vis_3 hms/6446 priot/Vis_3 hms/6426 priot/Vis_3 | | | | | |
| Lost Imap 3* Lost Imap 3* Lost Imap 3* Lost Imap 3* | 2004-08-15 30-28-29 2004-08-15 30-38-21 2004-08-13 30-38-21 2004-08-13 30-38-21 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-28 2004-08-15 30-38-21 2004-08-15 30-38-20 2004-08-15 30-30-20 2004-08-15 30-30-15 30-30-100-100-100-100-100-100-100-100-10 | Intervice, James (1444) colorrow, James (1444) color | | | | | |



- Collection of results of the executed jobs from the clusters on a centralised server (scp)
- Automatic Data synchronization of clusters (rsync)
- Support of diskless PCs environment (http://computemode.imag.fr)



Plan

CIGRI: lightweight grid computing approach

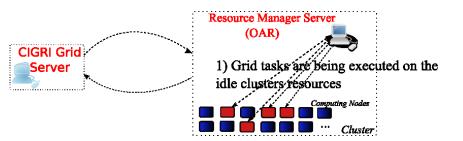


Fault Treatment and Turnaround time Optimisations

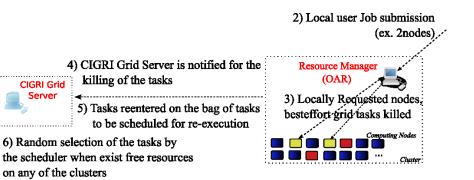
- CIGRI Fault Treatment
- Checkpoint/Restart techniques for turnaround time optimisation
- 3 Evaluating CIGRI upon Grid5000 platform

4 Conclusions

CIGRI Fault treatment



Fault treatment - No Checkpoints Strategy



CIGRI guarantees the complete execution of the application

Motivations for optimisation

Why optimise?

- Valuabe computation (of hours) can become completely useless just after an interference failure (tasks have to start from the beggining)
- CIGRI default Fault Treatment Strategy cannot guarantee a fast turnaround time.

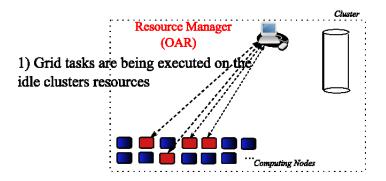
Solution:

- Checkpoint/Restart Technique: to cope with the high failure rate of computing nodes.
- Types: Application level/System level (BLCR library)

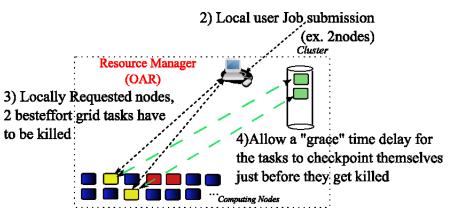
Constraints:

- Application level: Need to change the application code
- System level: BLCR limitations (TCP/UDP Sockets, open file locks, asynchronous I/O not supported)

Triggered checkpoints strategy



Triggered checkpoints strategy



- Used only for interference failures
- Drawback: "grace" time delay

Plan

- 1 CIGRI: lightweight grid computing approach
- Pault Treatment and Turnaround time Optimisations
- 3 Evaluating CIGRI upon Grid5000 platform

4 Conclusions

Real-life reproducible experiments upon Grid5000

- A CIGRI grid deployed upon Grid5000.
- Local cluster jobs based on obtained grid traces (sleep jobs, no computation just interference).
- One real-life multiparameter application send to CIGRI (real computation).
- Evaluate OAR/CIGRI platform (ex. benchmark the Fault-tolerance mechanism and optimisations).
- Sollection of results and a posteriori treatment.

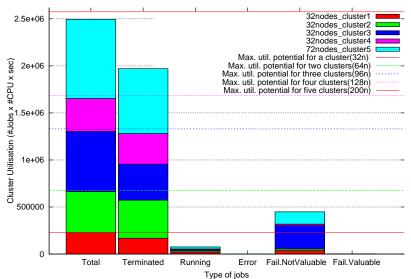




Evaluating CIGRI upon Grid5000 platform

CIGRI grid utilisation for 5 clusters of 200nodes grid, No checkpoints strategy (default)





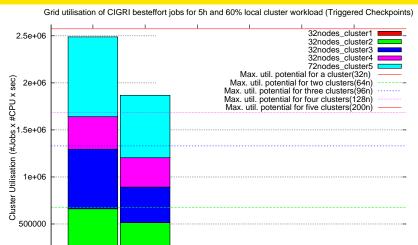
19/26

Evaluating CIGRI upon Grid5000 platform

Total

Terminated

CIGRI grid utilisation for 5 clusters of 200nodes grid, Triggered checkpoints strategy



Error

Fail.NotValuable Fail.Valuable

Running

Type of jobs

CIGRI grid experiment Results

 State of grid jobs for 5hours of experimentation of 5 clusters, 200nodes(DUAL OPTERON 2.0GHz, 2GB RAM), 60% local workload

| Strategies / Jobs | Total | Terminated | Running | Error | Inter. Failures Not Valuable | Inter. Failure Valuable |
|-----------------------|-------|------------|---------|-------|---------------------------------|----------------------------|
| Triggered checkpoints | 1377 | 762 | 74 | 103 | 226 | 212 |
| No checkpoints | 1420 | 739 | 74 | 8 | 581 | 0 |

[Ref: Yiannis Georgiou et Al.(2007)

Evaluations of the Lightweight Grid CIGRI upon the Grid5000 Platform (eScience2007)]

Plan

- 1 CIGRI: lightweight grid computing approach
- 2 Fault Treatment and Turnaround time Optimisations
- 3 Evaluating CIGRI upon Grid5000 platform
- Conclusions

Conclusions

CIGRI Principal contributions

- A Lightweight grid computing approach
- CIGRI Grid jobs: **transparent** execution upon the interconnected clusters.
- Best-effort jobs and the utilisation of unused resources
- Checkpoint/Restart for faster turnaround time

CIGRI-EGEE Interactions

- Transparency of Best-effort grid jobs on EGEE...??Could it be valuable?
- Current project: Cohabitation of CIGRI RhoneAlpes Grid with gLite EGEE Grid
 - Usage: Local cluster jobs, Low priority besteffort grid jobs(CIGRI) and remote grid jobs(EGEE)
 - OAR Resource Management System can cohabitate with any RMS system (SGE,LSF,PBS,..) connected with gLite



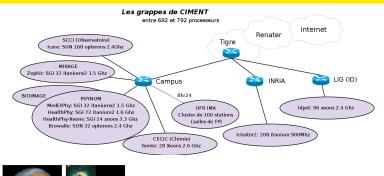


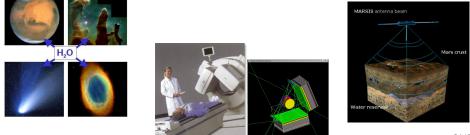




Conclusions

CIMENT: an OAR/CIGRI grid research platform





Links

CiGri

http://cigri.imag.fr
http://cigri.ujf-grenoble.fr



http://oar.imag.fr

CIGRI vs gLite: Principal Differences

CiGri

- Ideal for Regional or Metropolitan Area Grids
- Basic Security
- Simple (ssh, bash, rsync)
- Support of only BoT applications
- Best-effort jobs and the utilisation of unused cluster resources



- World Area Grid
- Strong security measures
- Standards (GRAM, GridFTP)
- Support of BoT + MPI + DAG Dependent applications
- Dedicated cluster resources
- Virtual Organizations