Dependable resource management in heterogeneous computing infrastructures

Gheorghe Cosmin Silaghi
Babeş-Bolyai University of Cluj-Napoca
Business Information Systems Dept.
Gheorghe.Silaghi@econ.ubbcluj.ro

Timisoara - 29-07-2014
Overview

- MADECIP: Disaster management research infrastructure based on HPC: 2014-2015
- Research activities in distributed systems at BIS@BBU: 2008-2014
MADECIP

- Improve the research infrastructures of BBU for disaster management
- Offers support for regional, national and international decision centers for disaster management lifecycle
- Interdisciplinary approach: mathematics, computer science, physics, chemistry, biology, geography, meteorology, communication science and business/economics
Strategic project for BBU

- Causes and effects of various types of disasters
- Adequate response strategies
- Short and long term effects over society, economy and environment
- Development of global management strategies

Interdisciplinary research project
Computing infrastructure

- Support for workloads of various requirements:
  - Computation intensive
  - Data intensive
- Mix of
  - HPC system
  - Private cloud system
- Main characteristics:
  - Easy interaction with the system
  - Reliability
**Hardware**

- Rpeak: 60Tflops, Rmax: 40Tflops.
- Various sorts of computing nodes used for HPC and private cloud:
  - Nodes with 2 E26XX processors
  - Enhanced nodes with 2 E26XX processors and additional GPU
  - Enhanced nodes with 2 E26XX processors and additional Intel Xeon Phi.
  - Management nodes;
  - Common storage for both HPC and private cloud
  - Tape library to support storage
**Software**

Integrated software for HPC and cloud management:

- SO licences for the whole system,
- cluster management,
- Jobs management and scheduling
- Monitoring and reporting
- MPI compilers and libraries
- Web interface for easy access
- Support for GPU, Intel Phi and virtualization
Specific Software

- Intel Cluster Studio,
- Portland Group,
- PGI Matlab,
- Maple,
- Numerical FDTD,
- Mathematica,
- Ansys,
- CFD Comsol,
- Multi-physics Gaussian,
- Sybyl-X,
- Material Studio
...
Research group on distributed systems @BIS @BBU

Members

- Gheorghe Cosmin Silaghi, Nicolae Tomai
- Experienced researchers: Mircea Moca, Ioan Petri (also at Cardiff Univ.) Cristian Litan (PhD from Carlos III Univ. Spain), Alexandru Stan
- PhD students / young researchers: Alexandru Butoi, Andreea Ilea
- ex PhD students: Gabriela Morar (Raiffeisen/Uniq), Cristina Muntean (CNR Pisa Italy, VLBD group), Cristina Stefanache (Quallysoft Budapest Hungary), Liviu Serban
- Collaborations with Cardiff University, INRIA Lyon, Innsbruck University, TU Delft, Coimbra University

Main group research objective: develop automated tools for dependable resource management in distributed systems based on principles from economics
Research support and computing infrastructure

- 2006-2008: FP6 CoreGrid NoE: activities in grid security transversal activity, 2009-2010 member of the CoreGrid ERCIM working group
- 2008-2013: CNCSIS/UEFISCDI: 2 IDEI and 1 TE projects
- 2008-2015: doctoral and postdoctoral fellowships
- 2014-2016: CNCSIS/UEFISCDI: 1 partnership project
- 2011-2015: Indirect support from FP7 Security IP projects SMART and RESPECT

- 1 small cluster with IBM SystemX M3/M4 servers
- 1 Condor HPC cluster, merging the computing power of about 120 desktop PCs

Timisoara - 29-07-2014
Heterogeneous computing environments

- Grid systems
- Peer-to-peer systems
- Computational grids: desktop grids and volunteer computing

Main research question:
- Develop tools and methods to build dependable distributed systems from heterogeneous resources (possibly over Internet)

Methodology
- Use principles from economic theory towards the above goal
Research contributions

Reputation management

- A utility-based reputation model for collaborative service-based computing

Sabotage tolerance in volunteer computing

- A protocol against collusion in desktop grids
- The result verification mechanism in desktop grids customized for the MapReduce programming model

Resource management

- A time constrained protocol for bilateral negotiation of SLA terms
- A scheduler for mitigating users and providers perspective when integrate various computing infrastructures: best effort grids, Internet desktop grids and public clouds.
Reputation management

- Designed a *generic* reputation service
  - based on utility computing and
  - tailored for service-oriented solutions
  - Applicable for both centralised and decentralised reputation management
- Model usage:
  - Resource brokering in classical grids;
  - Fine-grained access control considering resource usage
Reputation service architecture
Sabotage tolerance in volunteer computing

- Tackle the sabotage problem in volunteer computing, considering the assumption of coordinated attacks (including Sybil attack and whitewashing)
- Proposed a statistical approach to handle the basic collusion in desktop grids
- Extend the approach on the general case, and complemented it with a graph-based analysis solution
- Develop the resource verification mechanism for MapReduce over desktop grids
  - Solution integrated in the prototype developed at INRIA Lyon on top of XtremWeb
Resource management: SLA negotiation

- Applies to computational grids
- A time constrained protocol for bilateral negotiation of SLA terms
- SLA composed of QoS levels, price and penalty
- (Bayesian) agents that learn the opponents profile and propose the next deal such as to maximize the predicted utility of the opponent
- Under time constraints, relax the bidding conditions as the time elapses

- We demonstrated that with such a SLA negotiation strategy one can build a resource management solution to achieve high satisfaction for both providers and consumers
Resource management: scheduling on heterogeneous infrastructures

Mix of infrastructures
- Best effort grids (Grid5000)
- Internet desktop grids (XtreemWeb)
- Public clouds

Different scheduling criteria:
- Expected completion time
- Computing costs
- Expected failure impact (measured with the help of reputation)

Scheduling decision model: based on bilateral evaluations (Promethee method)

Main result: a methodology about how to tune a scheduler to supply best user satisfaction, given the provider established goals
Scheduler architecture

Experimental system

TraceLdr
DCIFactory
EvFactory

EvPQ (priority queue)

EvCtrl

Workload, host availability records
Add new HJ, HLE, RET events

Processed event
Get next event

Check workload completion

Stats

Add new record
HJ, RET events

LogLdr
Records

Builder

WKManager
WKUs
Tasks

SDM
S_1, S_2, ...

Decision req.
Selected task
Update

Task scheduler

Storage

Workload, host availability traces / DCI

Exp. Setup

Run

Write logs

Read logs

Write graphic representation

ExpCrtl

Timisoara - 29-07-2014


Timisoara - 29-07-2014
Thank you