CoreGRID Industrial Conference

Advance Reservation of Compute and Networking Resources in the Grid with QoS Guarantees for MPI-applications

Wolfgang Ziegler
Fraunhofer Institute SCAI - Department on Bioinformatics

www.scai.fraunhofer.de
Wolfgang.Ziegler@scai.fraunhofer.de
AGENDA

- VIOLA Grid testbed
- MetaTrace Application
- Demonstration 1
- MetaScheduling Service
- Demonstration 2
Acknowledgements

Some of the work reported in this presentation is funded by the German Federal Ministry of Education and Research through the VIOLA project under grant #01AK605F. This presentation also includes work carried out jointly within the CoreGRID Network of Excellence funded by the European Commission’s IST programme under grant #004265.

Contributions by:

Stephan Springstubbe, Oliver Wäldrich, Wolfgang Ziegler - FHG.SCAI
Stephan Heisterkamp, Ferdinand Hommes, Carsten Rosche - FHG.IAIS
Thomas Eickermann, Lydia Kirtchakova, Philipp Wieder - FZJ
Ramin Yahyapour, University Dortmund
Ariel Oleksiak - PSNC
Christoph Barz, Wolfgang Moll, Markus Pilz - University Bonn
GRAAP-WG, GSA-RG, OGSA-RSS - OGF
VIOLA Testbed

VIOLA-Networking
- Deployment and operation of the testbed, test of advanced network equipment
- Signaling and reservation
  - bandwidth- and QoS-reservations in the network
  - interfaces for user-driven reservation: immediate and in advance
VIOLA Applications: Distributed Parallel Simulations

VIOLA-Applications (Multi-physics, Tele-collaboration)

- KoDaVis-Atmo: Collaborative visualization of huge atmospheric datasets (FZJ)
- TechSim: Distributed simulation of crystal growth and biosensors (Caesar)
- AMG-OPT: Parameter optimization and optimal algebraic solvers – Mechanical structure (SCAI)
- MetaTrace: Simulation of pollutant transport in groundwater with distributed SMP-Clusters (FZJ)
European research network on foundations, software infrastructures and applications for large-scale, distributed GRID and peer-to-peer technologies.

Network Details

Cluster FZJ

- 12x 1GbE

Cluster caesar

- 33x 1GbE

Live Interface Statistics with HP OpenView

Cluster FH

- 6x 1GbE

MPLS-Router Backbone Layer-3-VPN

bidirectional max. 10 Gbps

Description:

- Cluster Traffic
- Generated Load

IW96000 Traffic Generator

Vis-PC

Univ. Bonn

max. 10 Gbps
VIOLA Grid Middleware

- Negotiation of timeslot & nodes with local schedulers for each job
- UNICORE initiates the reservation and submits the job-data
- UNICORE Client / MetaScheduler Service interface using WS-Agreement protocol
- Interface MetaScheduler / Adapters based on WS-Agreement
- Interface between MetaScheduler Service and local RMS implemented with adapter pattern
- Authentication and Communication of Adapter and local Scheduler through Proxies running at the (remote) sites
- Communication secured by ssl
What might happen without Orchestration

Site A

- UNICORE Client
- Local Scheduler
- Job Queue
- Site A
- Cluster

Site B

- The user describes his Job
- The Job is passed to the UNICORE System
- The Primary NJS distributes the Job to all sites
- The Job is submitted to the local batch-queues of all systems
- The components of the Job are started - depending on the state of the local batch-queues
- The quality of the network connections depends on the actual load
Demonstration 1

Running the Application without Bandwidth Reservation ...
Motivation

• Need to use more than one resource for an application
  – E.g. multi-physics simulations or Collaborative simulations/visualisation needing
    • Multiple compute resources
    • Network between the compute resources
    • Visualisation devices
    • Storage
  – Need for guarantees on the availability of the resources
    – Co-allocation of resources
    – More efficient Workflow Scheduling
Reference to CoreGRID

Several CoreGRID Partners
Integrating the R&D in the CoreGRID Institute on RMS

Definition of the components of a
Grid scheduling architecture and their interaction

Multi-Level Scheduling Strategies

Service Level Agreements
The MetaScheduling Service MSS

Key features

• Operating across multiple administrative domains
• Advance Reservation
  – through capabilities of the local schedulers
• Co-allocation of arbitrary resources
  – Abstraction through adapter pattern
  – Negotiation algorithm changes depending on the degree of information exposed
• Support for Workflow dependency scheduling
  – Decrease/minimise the makespan of a workflow
• Service level Agreement between the client and the MSS
  – Combines the individual SLAs between the MSS and the adapters
  – Based on WS-Agreement with small extensions to support the negotiation between MSS and local schedulers
• Implemented as Webservice
  – Using Apache hosting environment MUSE recently developed
• Middleware Independent
**MetaScheduler - Integration in UNICORE**

- UNICORE Client sends request to MetaScheduler (WS-Agreement)

- MetaScheduler negotiates earliest time to run this job, requests the reservation of the requested resources and returns the WS-Agreement with additional Status, ID

- UNICORE Client creates Abstract Job Object (AJO) and sends it to the Primary Network Job Supervisor (NJS)

- NJS incarnates the AJO according to the information in the AJO and the UIDB, forwards it to the local Target System Interface (TSI) and sends the AJO to all other NJSs

- TSI creates the entry for the Meta-Job in the UNICORE Job Queue, and stores the job data in the User-directory

- Scheduler triggers job at start time
Negotiation Protocol

MetaScheduling Service

create resource list → calculate timeslot → schedule resources → get reservation properties → check reservation

reservation succeeded → reservation failed

Resource preview
create preview → lookup local schedule

Schedule
submit → schedule resource

Resource properties
local resource properties (effective start time) → lookup local schedule

Terminate
cancel reservation → free resources

Local Scheduling System

Resources

European research network on foundations, software infrastructures and applications for large-scale, distributed GRID and peer-to-peer technologies
Collecting runtime information

- After job submission the MSS monitors the state of all reservations
- When all reservations changed to *active* the MSS gathers the IP addresses of finally assigned nodes
- This information is aggregated by the MSS
- The aggregated runtime information is sent to the subsystems
- Compute RMS use this information to build the MetaMPICH configuration and start the application
- The network RMS completes the network configuration based on this information
Allocate and Configure the Network Resources

1.) Reservation of required Resources
- Submit of a Reservation to the Network Resource Manager
- Acknowledgement of Reservation

2.) Bind of IP-Addresses at Run-time
- IP-Addresses are published at run-time of the job through the local Adapter
- Bind of the IP-Addresses by the Network Resource Manager
- Without explicit Bind the QoS Parameters for the Site-to-Site Interconnection are used
Application Startup

- At the time a reservation becomes active MSS MetaMPICH wrapper script is started
- The wrapper checks periodically if the MSS runtime configuration file is present
- When the MSS runtime was found the MetaMPICH configuration file is generated
- Afterwards it calls the mpirun command with the UNICORE flag (-unicore) using the generated configuration file
- This prints out the ssh calls to start up a MetaMPICH job
- The wrapper only executes the ssh call for the local site in order to start the application at the local site
European research network on foundations, software Infrastructures and applications for large-scale, distributed GRID and peer-to-peer technologies

Orchestration with SLA

The user describes his Job

MetaScheduling Request (WS-Agreement template)

The Job is passed to the UNICORE System

MetaScheduler Response (WS-Agreement)

Negotiations and Reservations

All Components of the Job are started at the point in time agreed upon, at the same time the network connections are switched on

The user describes his Job

Site B

MetaScheduling Request (WS-Agreement template)

The Job is passed to the UNICORE System

MetaScheduler Response (WS-Agreement)

Negotiations and Reservations

All Components of the Job are started at the point in time agreed upon, at the same time the network connections are switched on
Demonstration 2

Running the Application with Bandwidth Reservation ...
MetaTrace – Simulation of pollutant transport in groundwater

Rudolf Berrendorf, Boris Bierbaum, Carsten Clauss, Wolfgang Frings, Stephan Springstubbe, Oliver Wäldrich, Wolfgang Ziegler
MetaTrace Goals (1/2)

Category: Environmental research, Multi-Physics

Application: Simulation of pollutant transport in groundwater

Goal: Long-time simulations of large areas via optimization of the coupled applications on Grids (distr. SMP-Clusters)
Environmental research:
- experiments in a test field
- MPI-based simulations on a Grid

Optimal distribution of work in the Grid:
- calculation of water-flow
- distribution and chemical reactions of pollutants
- exchange of intermediate results:
  “up to 1 GB between two nodes every 10..20 sec.”

Grid computing:
- simulate more accurate models in less time
MetaTrace Components

- Simulation of Groundwater: MPI F90 program “Trace”
- Simulation of Pollutant transport: MPI C++ program “Partrace”

- Visualization: Program “AVS Express”
MetaTrace Coupling and Requirements

- **HPC requirements:** 2-5 PC-Clusters, min. to days
- **Network technology** GbE via optical links
- **Networking requirements:** up to 8 Gbit/s, Latency < 0.1 ms
- **Software requirements:** Message Passing Interface (MPI)
VIOLA-UNICORE GUI (1/3): Job initiation and resources

Initiate a new meta job

Choose your sites

Get information on site resources
Specify nodes and tasks

Specify executable and working directory

Specify local communication device
Specify needed bandwidth

Specify max. runtime

Submit metajob to Metascheduler
MetaTrace Job

- **MetaMPICH-coupled codes: Trace and Partrace**
  
  Simulation of Pollutant transport in the Groundwater

- **VISIT-coupled codes: Partrace and AVS/Express**

  Visualization of Pollutant transport

[Diagram showing the coupling of components and visualization tools]
MetaTrace Model and AVS Visualization

- Simulation area #FE 280.543 (127x47x47)
- River: add +20 cbm/day
- Solid ground with reduced hydrostatic permeability
- 3 wells: drain -150 cbm/day
- Pollution: one injector with 100,000 particles per day (cloud with different concentrations)
- Scales:
  - Pollution µg/cbm
  - Velocity m/day (typ. 1 m/day)
  - Water pressure m
MetaTrace Performance Visualization

Grid network Performance for different message sizes