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

European Commission

Information Society Technologies

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

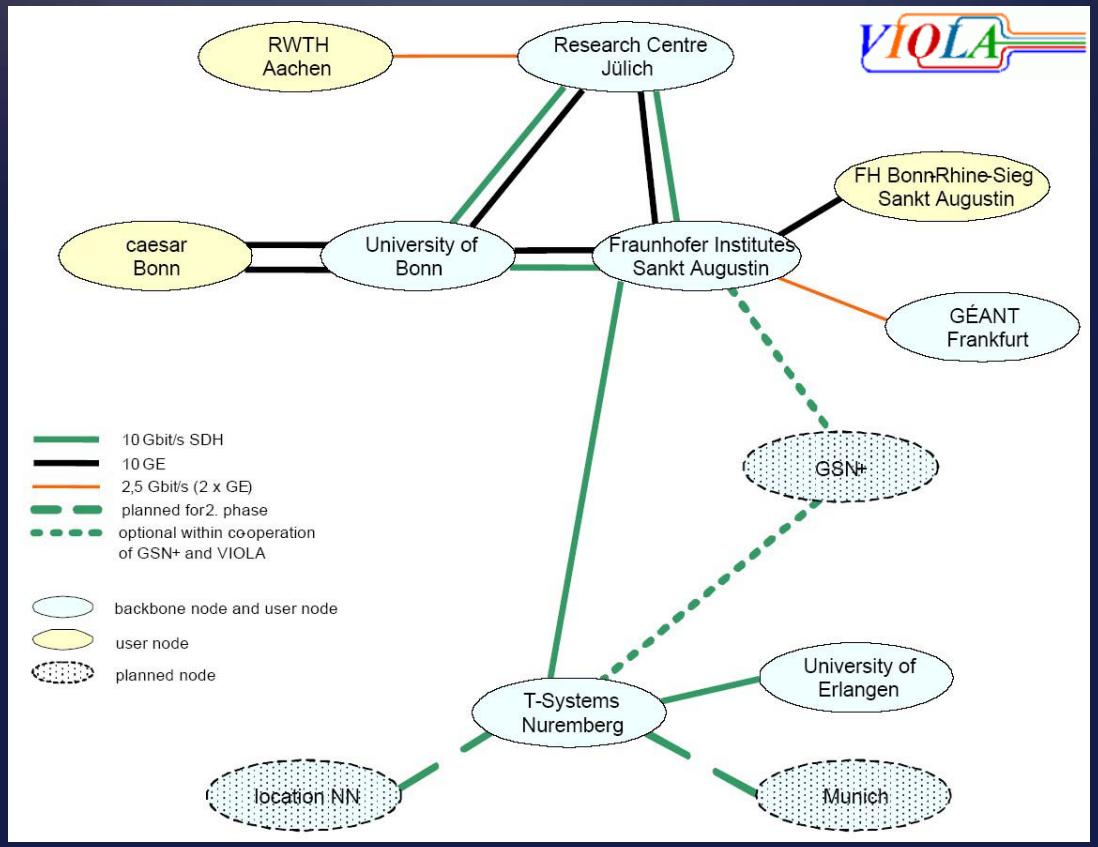
Christoph Barz, Wolfgang Moll, Markus Pilz - University Bonn

GRAAP-WG, GSA-RG, OGSA-RSS - OGF

VIOLA Testbed

VIOLA-Networking

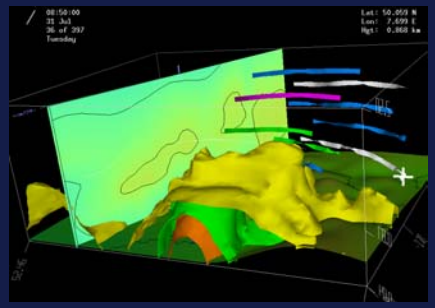
- Deployment and operation of the testbed,
  test of advanced network equipment
- Signaling and reservation
  - bandwidth- and QoS-reservations in the network
  - interfaces for user-driven reservation: immediate and in advance

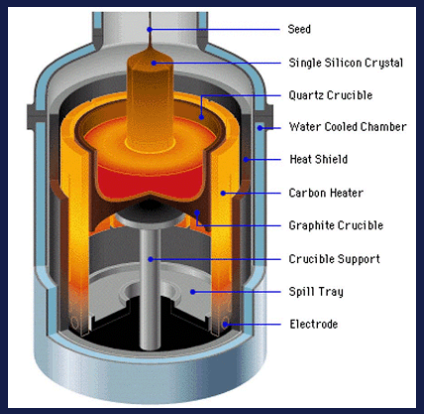


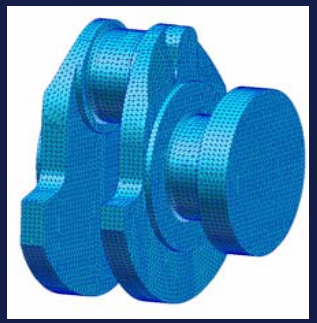
VIOLA Applications: Distributed Parallel Simulations

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

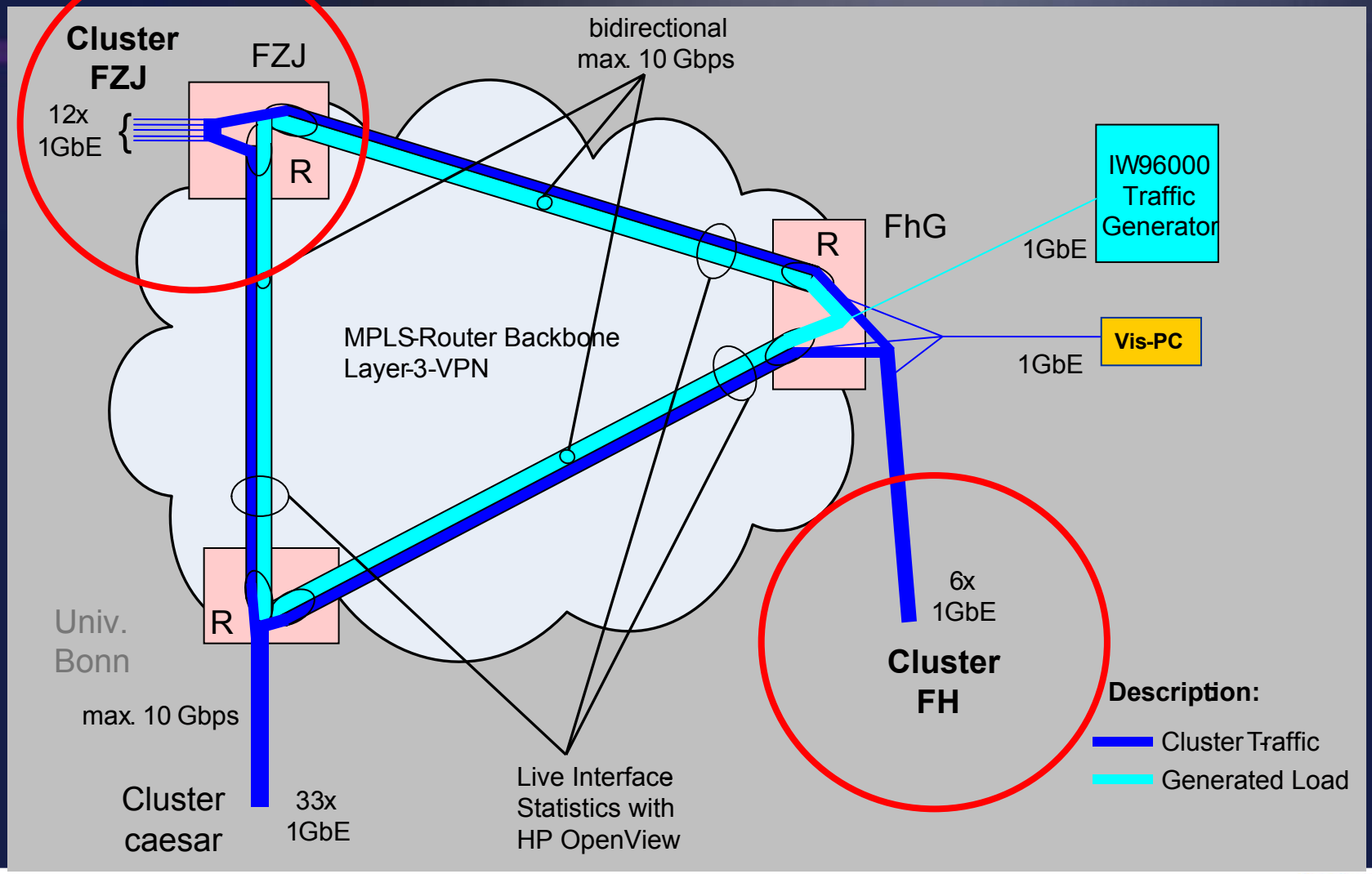
- KoDaVis-Atmo: Collaborative visualization of huge atmospheric datasets (FZJ)
- TechSim: Distributed simulation of crystal growth and biosensors (Caesar)
- AMG-OPT: Parameter optimization and optimal algebraic solvers – Mechanical structure (SCAI)
- MetaTrace: Simulation of pollutant transport in groundwater with distributed SMP-Clusters (FZJ)



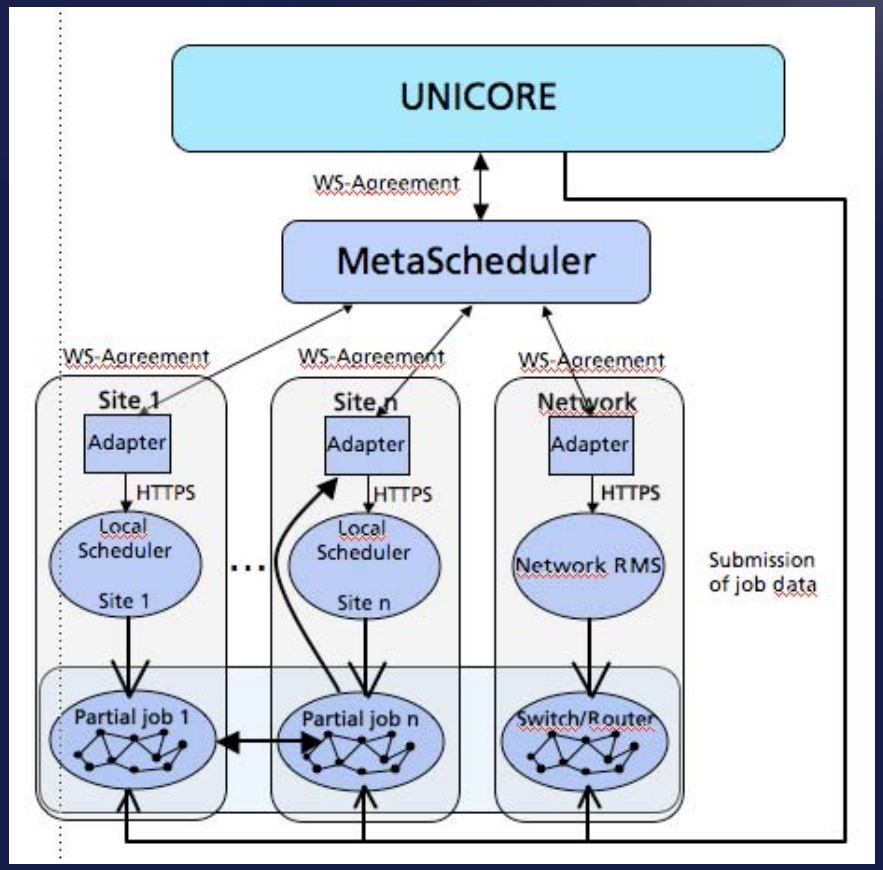




Network Details

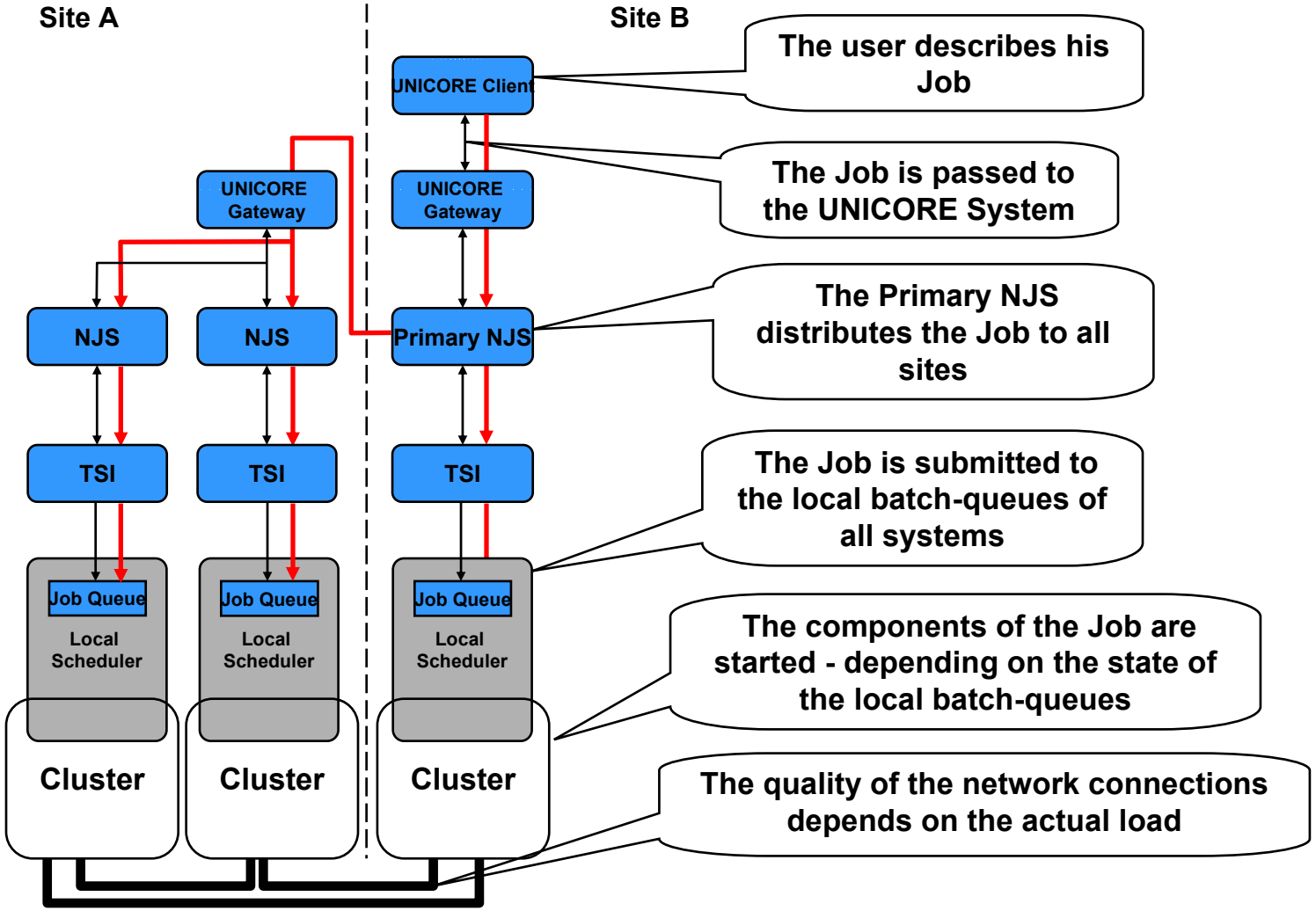


VIOLA Grid Middleware



- 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



Demonstration 1

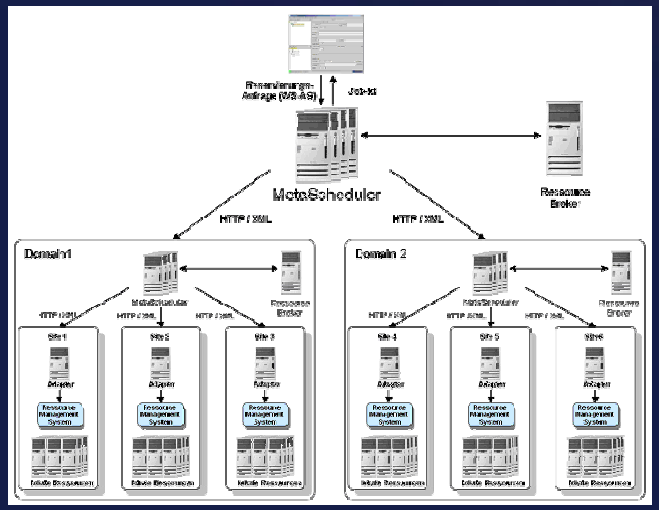
Running the Application without Bandwidth Reservation ...

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

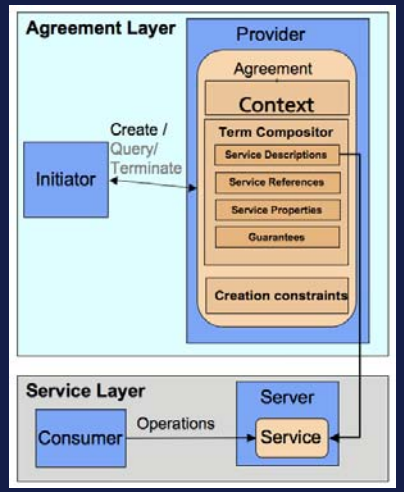
Reference to CoreGRID

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

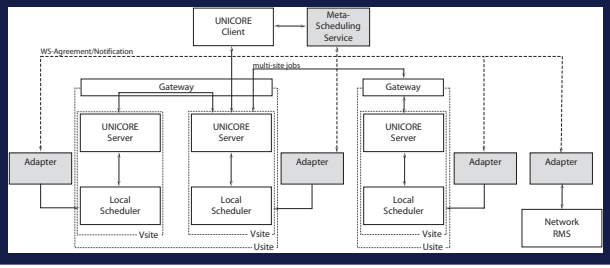


Definition of the components of a Grid scheduling architecture and their interaction

Service Level Agreements



Multi-Level Scheduling Strategies

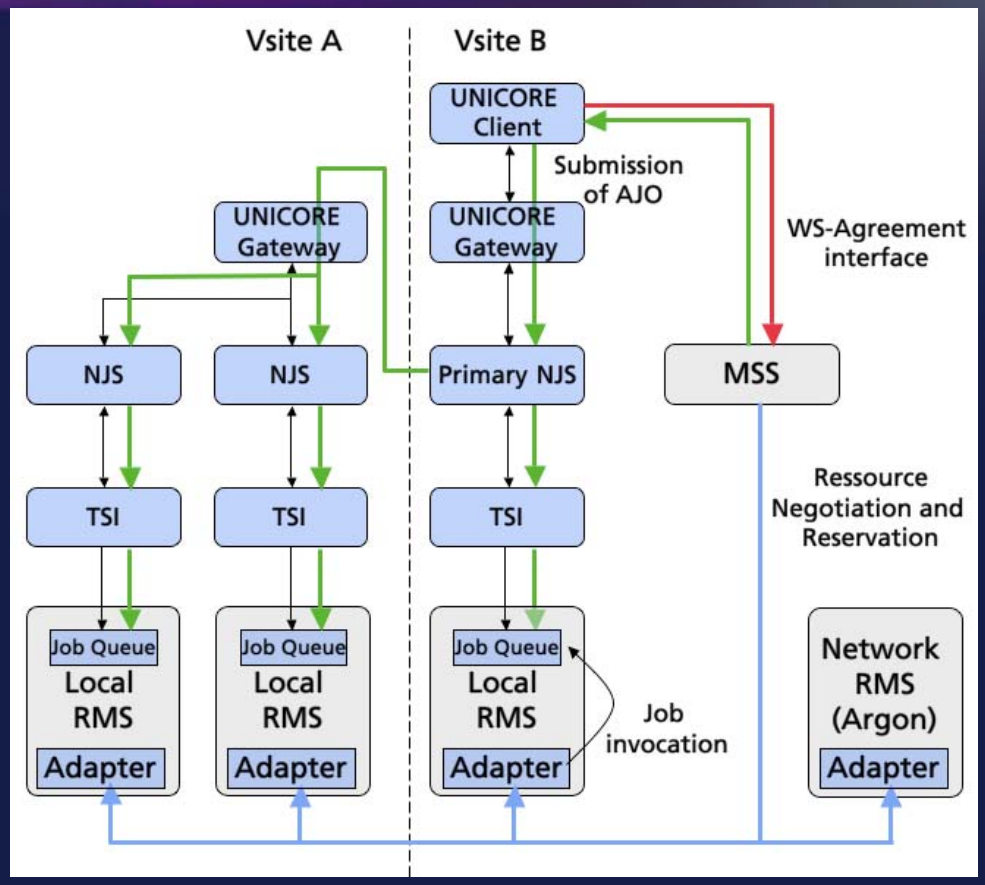


The MetaScheduling Service MSS

Key features

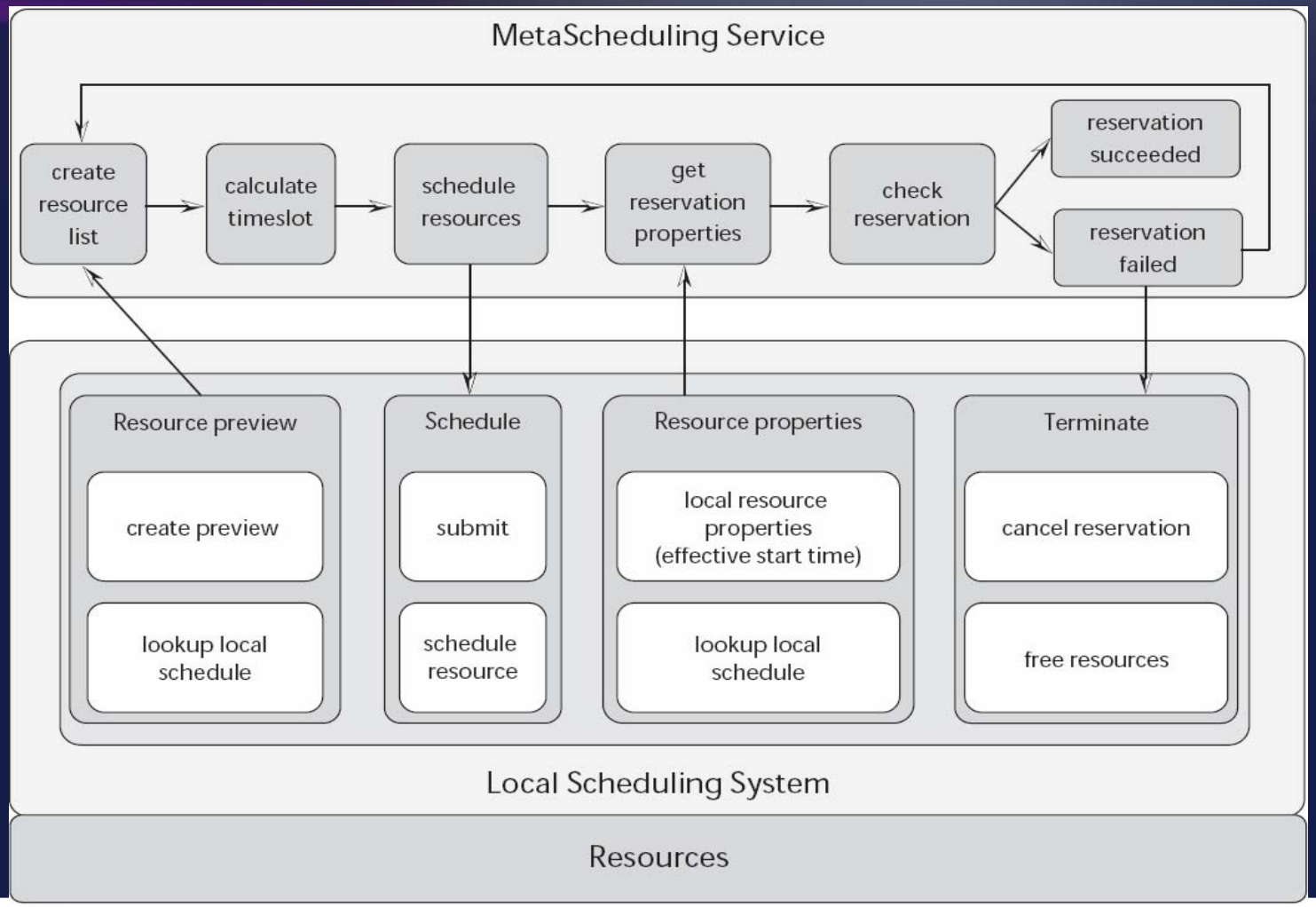
- Operating across multiple administrative domains
- Advance Reservation
  - through capabilities of the local schedulers
- Co-allocation of arbitrary resources
  - Abstraction through adapter pattern
  - Negotiation algorithm changes depending on the degree of information exposed
- Support for Workflow 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

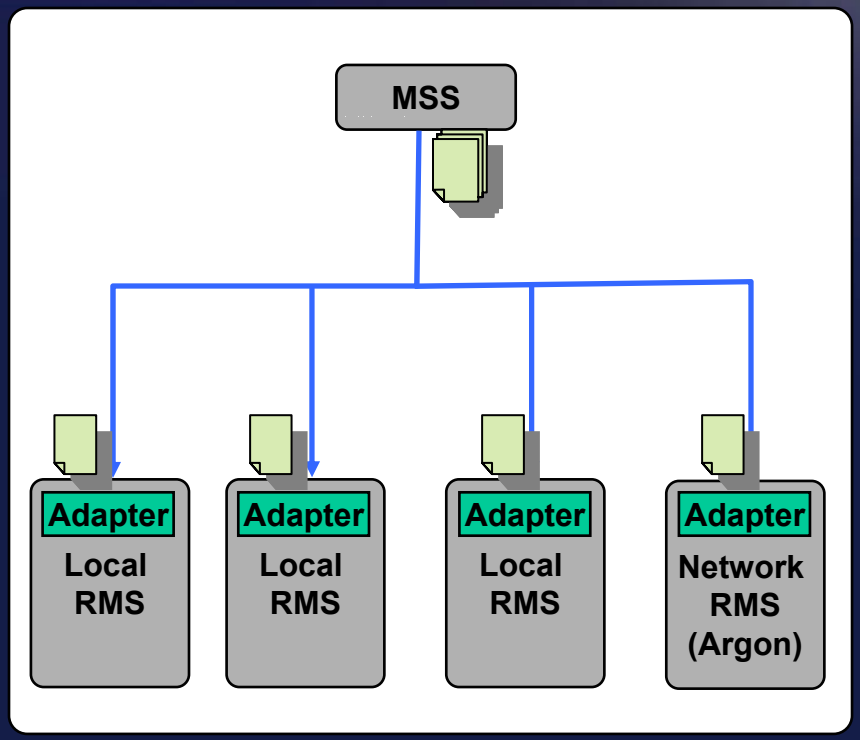


- UNICORE Client sends request to MetaScheduler (WS-Agreement)
- MetaScheduler negotiates earliest time to run this job, requests the reservation of the requested resources and returns the WS-Agreement with additional Status, ID
- UNICORE Client creates Abstract Job Object (AJO) and sends it to the Primary 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

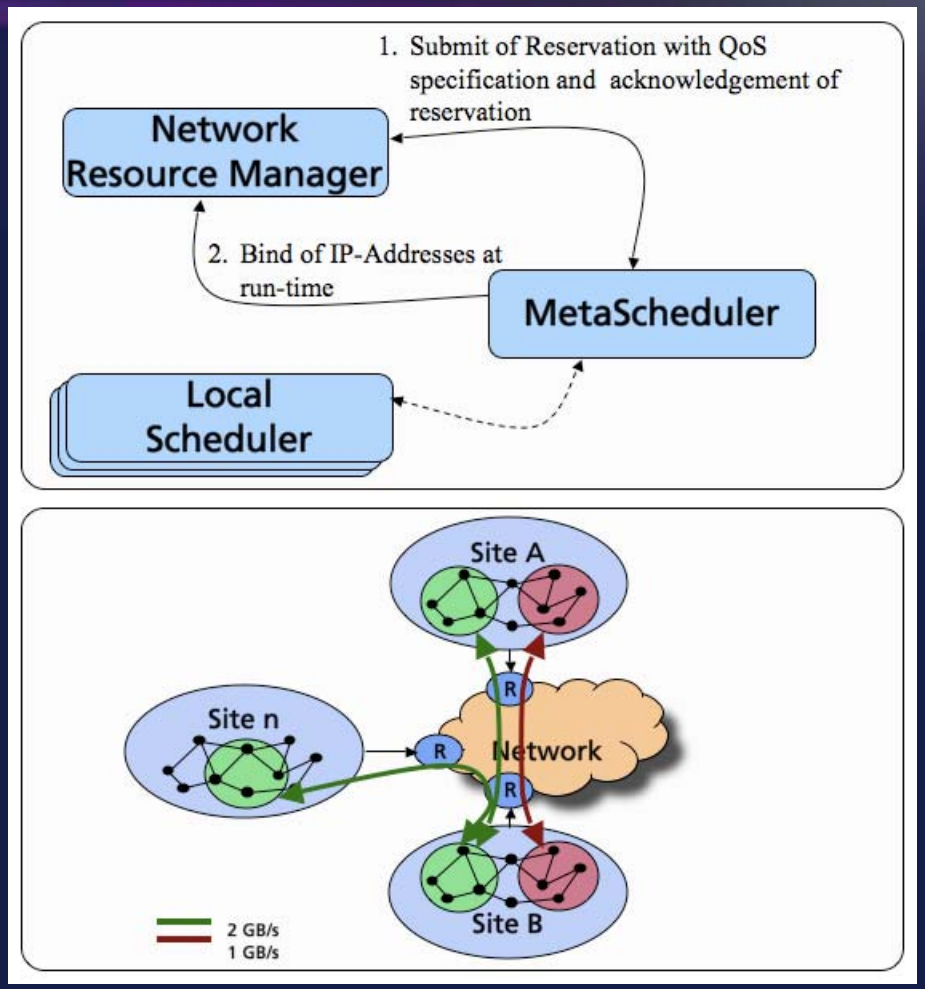


Collecting runtime information



- 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



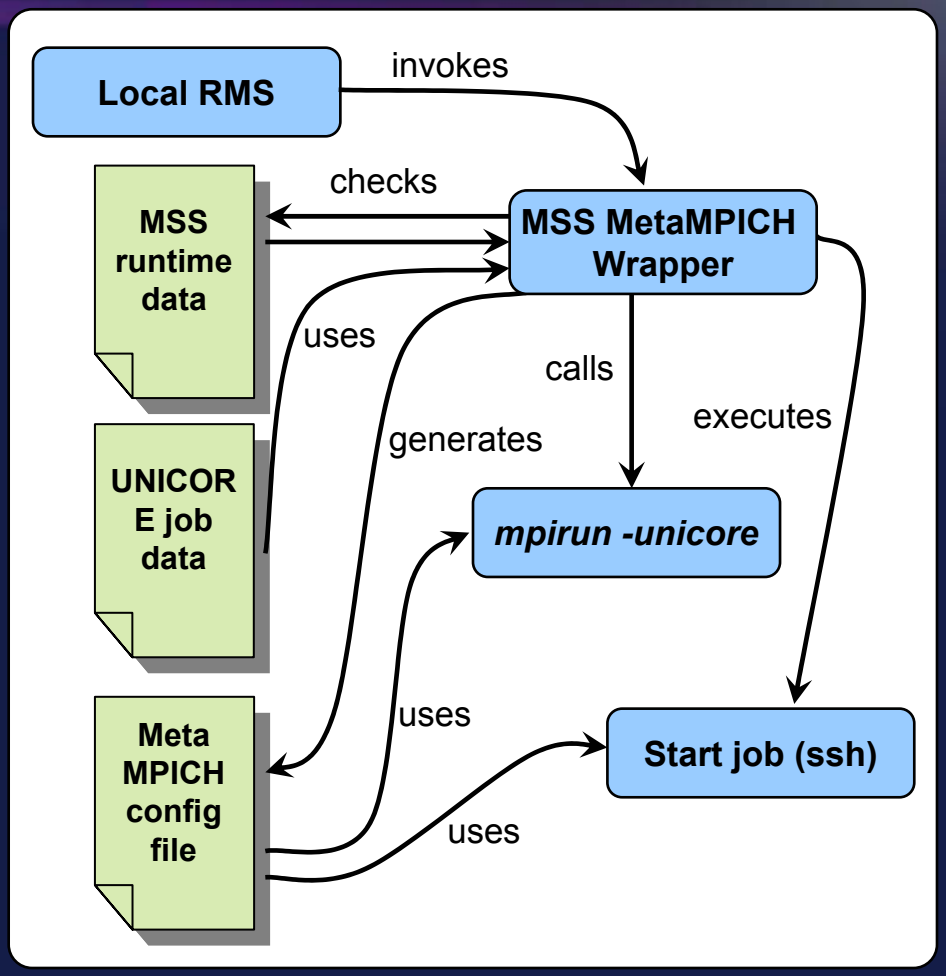
1.) Reservation of required Resources

- Submit of a Reservation to the Network Resource Manager
- Acknowledgement of Reservation

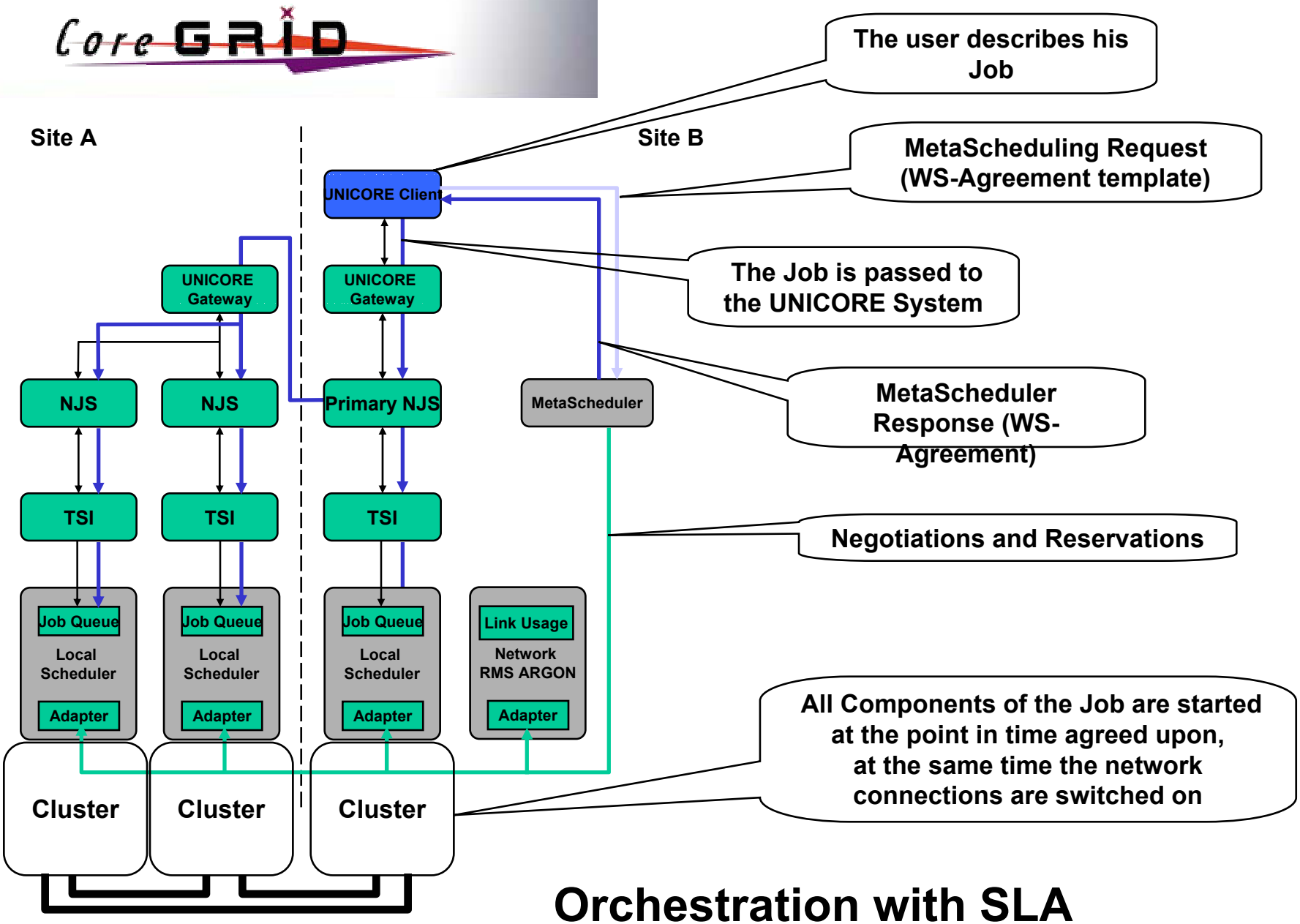
2.) Bind of IP-Addresses at Run-time

- 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



Orchestration with SLA

Demonstration 2

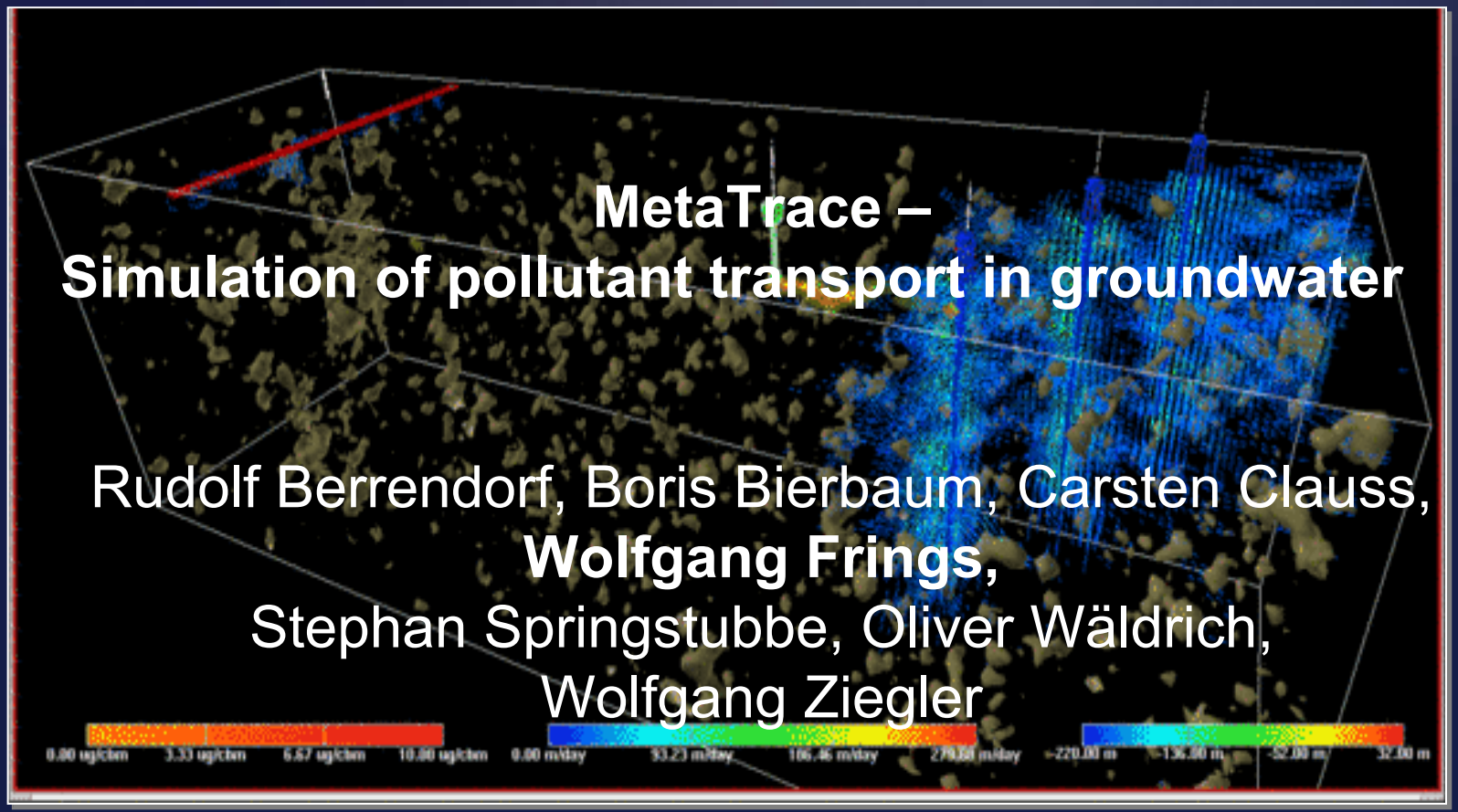
Running the Application with Bandwidth Reservation ...

CoreGRID

Managed by

European Research Consortium for Informatics and Mathematics ERCIM

Grids@Work, 30.11./1.12.2006



MetaTrace – Simulation of pollutant transport in groundwater

Rudolf Berrendorf, Boris Bierbaum, Carsten Clauss,
Wolfgang Frings,
Stephan Springstubbe, Oliver Wäldrich,
Wolfgang Ziegler

Fachhochschule Bonn-Rhein-Sieg

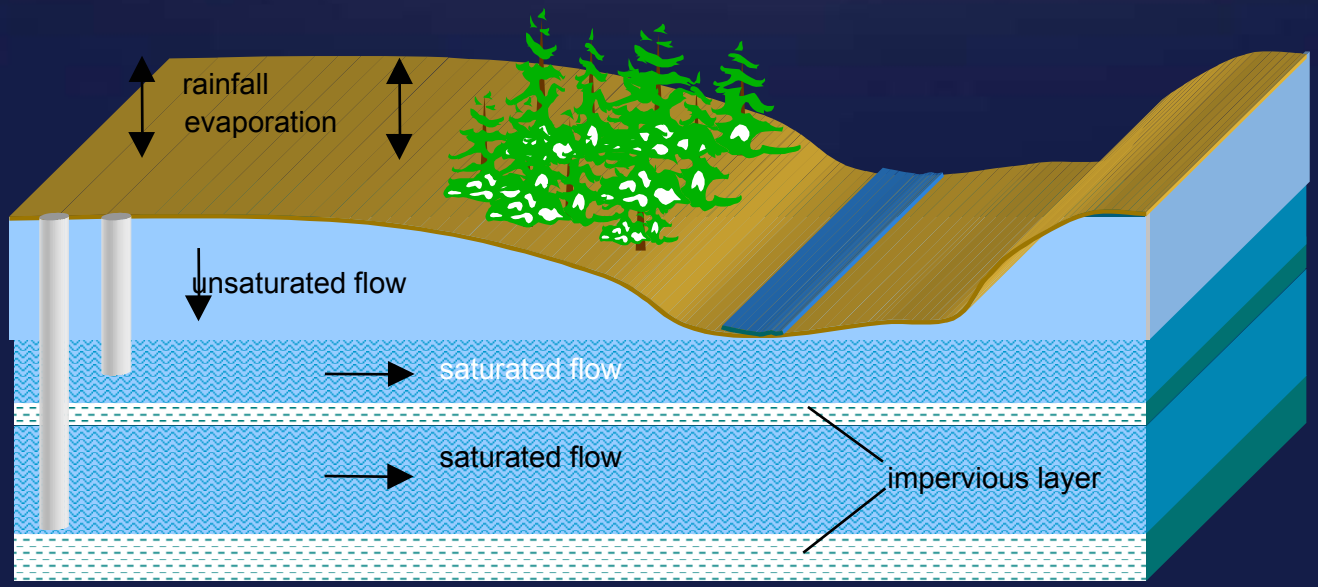
Forschungszentrum Jülich in der Helmholtz-Gemeinschaft

Fraunhofer Institute Algorithms and Scientific Computing

MetaTrace Goals (1/2)

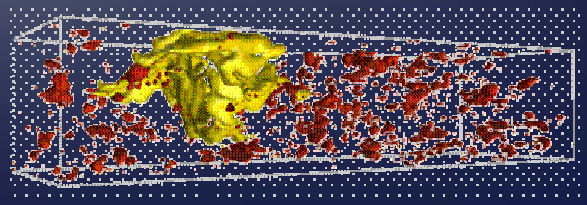
Category: Environmental research, Multi-Physics

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)

MetaTrace Goals (2/2)



Environmental research:

- experiments in a test field
- MPI-based simulations on a Grid

Optimal distribution of work in the Grid:

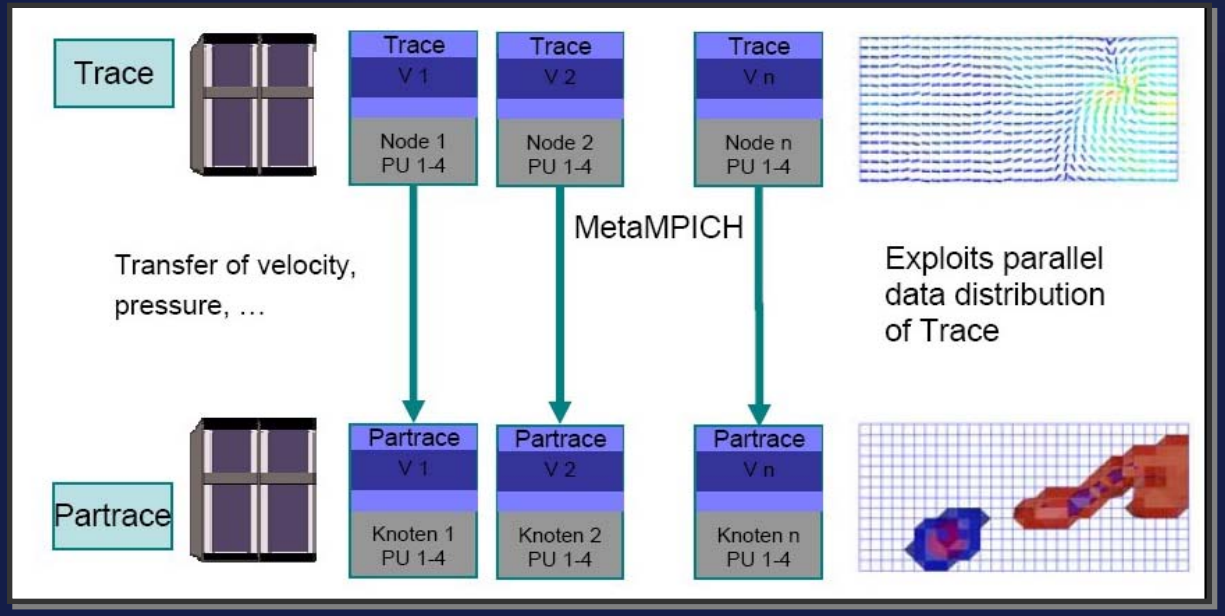
- calculation of water-flow
- distribution and chemical reactions of pollutants
- exchange of intermediate results:
  "up to 1 GB between two nodes every 10..20 sec."

Grid computing:

- simulate more accurate models in less time

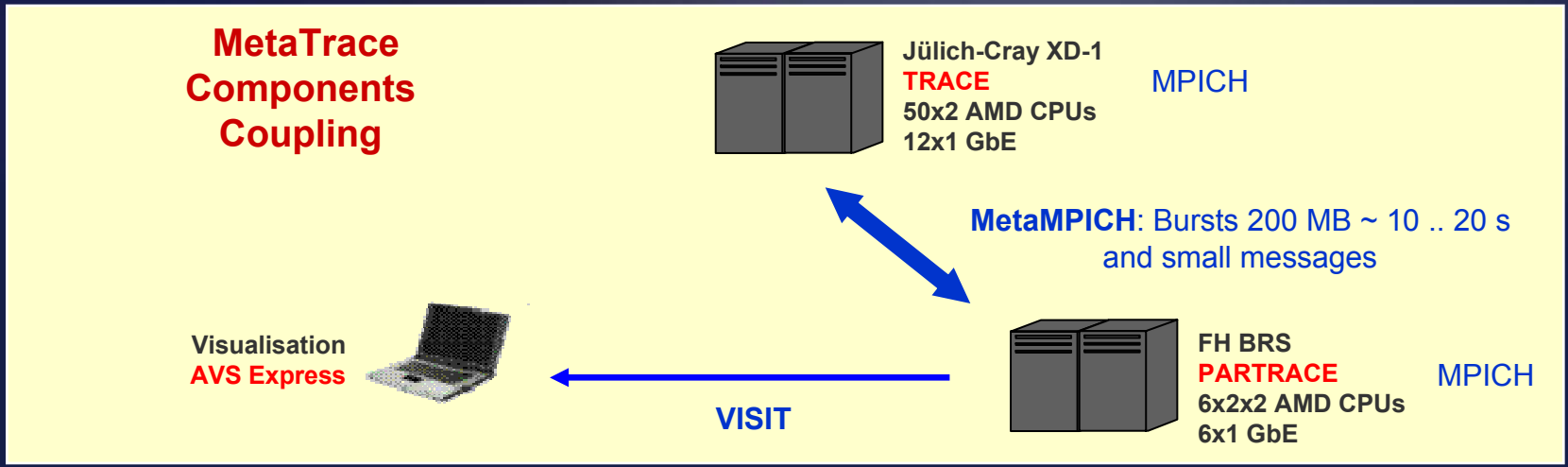
MetaTrace Components

- **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

UNICORE Client

File Job Preparation Job Monitoring Settings Extensions Help

Job Preparation
- New_MetaMPIJob_4 [15:28:56 03/0
  - New_MetaSubjob_0_C
  - New_MetaSubjob_1
- New_MetaMPIJob_3
  - New_MetaSubjob_0_DEMO_NJS
  - New_MetaSubjob_1_PACK
- New_MetaMPIJob_1

Job Monitoring
- VIOLA_CAESAR
  - CAESAR_NJS <NJS>
- VIOLA_FH-BONN
  - DEMO_NJS <NJS>
- VIOLA_FZJ
  - CRAY-XD1 <NJS>

Name: New_MetaMPIJob_4

Job Scheduling

Date: 01.06.06

13:29

job in minutes: 0

Submit Metajob

Resource Requirements

UNICORE Sites
- VIOLA_CAESAR
  - CAESAR_NJS <NJS>
- VIOLA_FH-BONN
- VIOLA_FZJ
- VIOLA_SCAI

Information about the target system

Information

NJS information: Tue Jun 20 12:11:18 CEST 2006
NJS information: 4.6.0 build 2
Execution System: GNU Hurd v1.0 RC336
Architecture: IMD Flopterix Quad-Pumped
Data Representation: IEEE Format
MetaMPICH: /usr/local/viola/mp-mpich
MetaMPICH_Device: ch_usock
Login-Node: pcc3001.viola-testbed.de
Number of CPUs: 2.0
Number of AJOs running on NJS: 3.0

Capacities

Name	Description	Minimum	Maximum	Default	Units
Node	Number of Nodes	1.0	32.0	1.0	Nodes
Processor	Number of PEs per Node	1.0	2.0	1.0	Processor
Memory	Total Amount of Memory	1.0	2056.0	1024.0	Megabytes per node

Storages

Name	Description	Minimum	Maximum	Default	Units
Home	Space in $HOME	0.0	100.0	10.0	MegaBytes
Alternative Job Directory	Job Uspace	0.0	100.0	10.0	MegaBytes
Job Directory	Job Uspace	0.0	100.0	10.0	MegaBytes
Spool	Spool Space	0.0	100.0	10.0	MegaBytes
Root	Absolute paths	0.0	100.0	10.0	MegaBytes

Performance

Name	Description	Minimum	Maximum	Default	Performance	Units
FloatingPoint	Time per Job	2000.0	7200000.0	120000.0	200.0	MegaFlop

Initiate a new meta job

Choose your sites

Get information on site resources

VIOLA-UNICORE GUI (2/3): VSite Spec.

UNICORE Client

File Job Preparation Job Monitoring Settings Extensions Help

Job Preparation
- New_MetaMPIJob_4 [15:28:56 03/06/2006]
  - New_MetaSubjob_0_CAESAR_NJS
  - New_MetaSubjob_1_CAESAR_NJS
- New_MetaMPIJob_3
  - New_MetaSubjob_0_DEMO_NJS
  - New_MetaSubjob_1_PACK
- New_MetaMPIJob_1
  - New_MetaSubjob_0_CAESAR_NJS

Job Monitoring
- VIOLA_CAESAR
  - CAESAR_NJS <NJS>
- VIOLA_FH-BONN
  - DEMO_NJS <NJS>
- VIOLA_FZJ
  - CRAY-XD1 <NJS>
- VIOLA_SCAI
  - PACK <NJS>

Name: w_MetaSubjob_0_DEMO_NJS

Resource Requirements | Special Settings

Compute Settings
- Compute Tasks: 1
- Compute Nodes: 16
- Tasks/Node: 4
- Node usage: shared
- Thread/Task: 1
- Memory:

IO Settings
- IO Tasks: 0
- IO Tasks/Node: 1
- Node usage: shared

Router Settings
- Router Tasks: 0
- Router Tasks/Node: 1
- Node usage: shared

Total nodes: 0 Update

MetaMPICH settings
- MetaMPICH path: /usr/local/viola/mp
- Executable: /home/viola0 ETA-MULTIDEVICE/metatest
- Working directory: /home/viola03/public/META-MULTIDEVIC
- Parameters:
- Device: ch_gm

Browse

Specify nodes and tasks

Specify executable and working directory

Specify local communication device

VIOLA-UNICORE GUI (3/3): Final spec.

UNICORE Client

File Job Preparation Job Monitoring Settings Extensions Help

Job Preparation
- New_MetaMPIJob_3
  - New_MetaSubjob_0_DEMO_NJS
  - New_MetaSubjob_1_PACK
- New_MetaMPIJob_1

Job Monitoring
- VIOLA_CAESAR
  - CAESAR_NJS <NJS>
- VIOLA_FH-BONN
  - DEMO_NJS <NJS>
- VIOLA_FZJ
  - CRAY-XD1 <NJS>
- VIOLA_SCAI
  - PACK <NJS>

Name: New_MetaMPIJob_3

Job Scheduling

Date: 01.06.06

Start time: 13:22

Duration of the job in minutes: 10

Submit Metajob

Resource Requirements

UNICORE Sites
- VIOLA_CAESAR
- VIOLA_FH-BONN
- VIOLA_FZJ
- VIOLA_SCAI

Information about the target syste

Vsite(s)	Total Nodes	Tasks/Node	Threads/Task	DEMO_NJS	PACK
DEMO_NJS <NJS>	0	4	1	-	0 : 1
PACK <NJS>	1	2	1	0 : 1	-

Add Vsite

c2b435fbd6625edcda439c2861989dfc_11d052e5-aa88-4d89-9805-4f776a174856 D:\UNICORE\metatest-wrpack.ajo

Specify max. runtime

Submit metajob to Metascheduler

Specify needed bandwidth

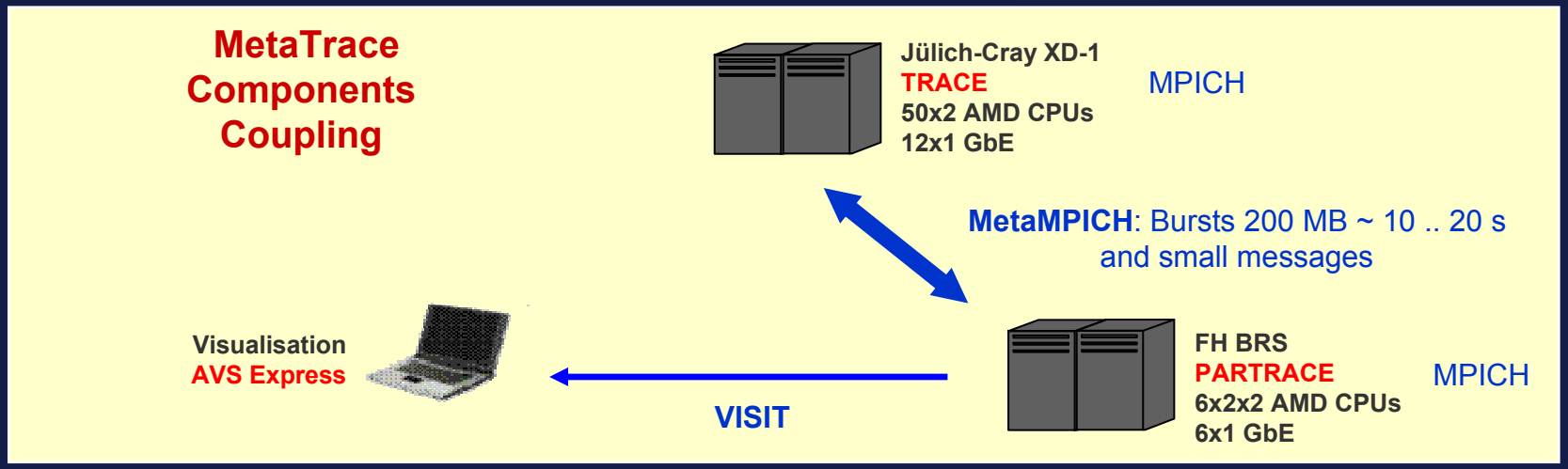
MetaTrace Job

- **MetaMPICH-coupled codes: Trace and Partrace**

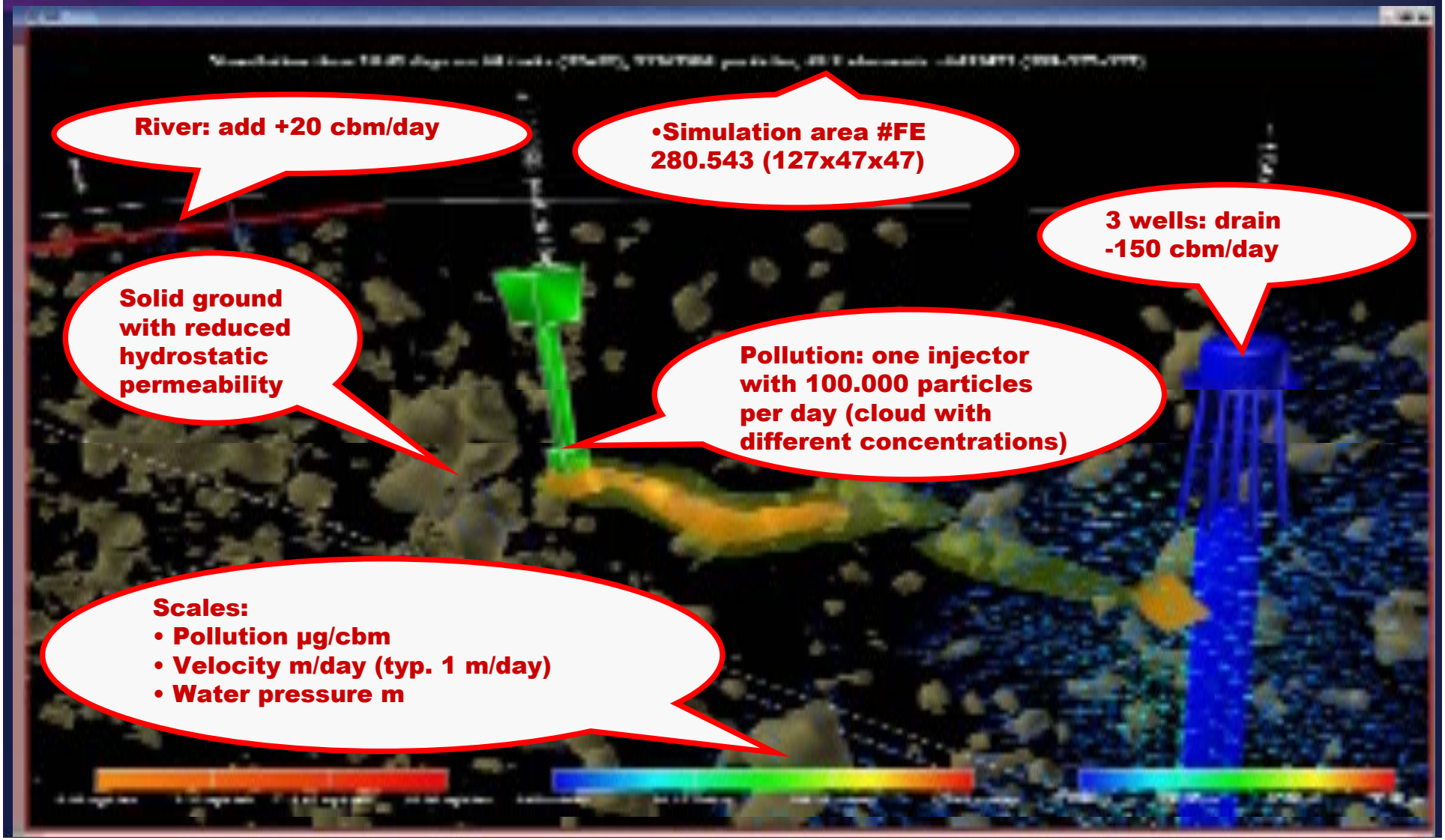
  Simulation of Pollutant transport in the Groundwater

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

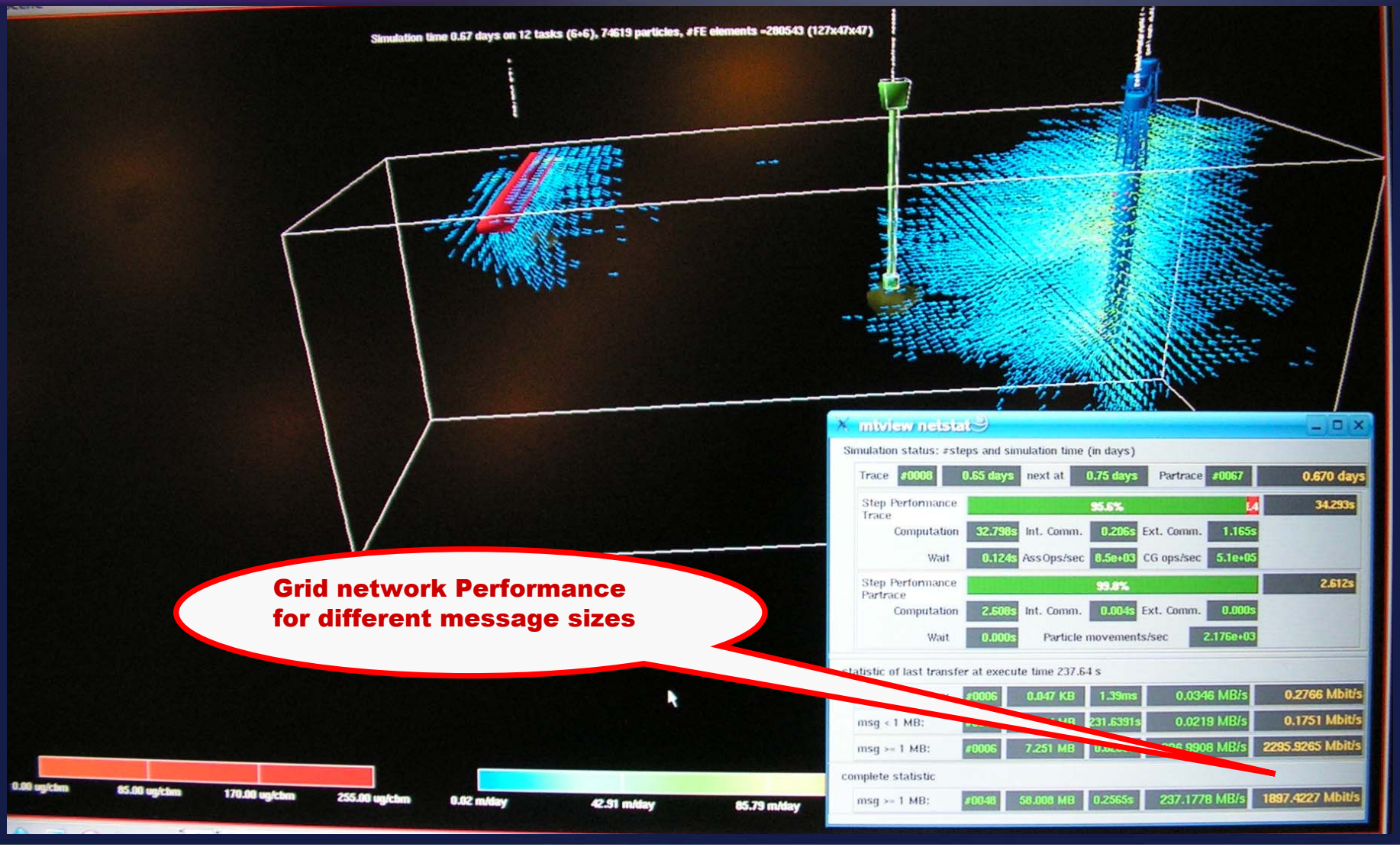
  Visualization of Pollutant transport



MetaTrace Model and AVS Visualization



MetaTrace Performance Visualization



Grid network Performance for different message sizes