User Tools and Languages for Graph-based Grid Workflows

Fraunhofer

Institut
Rechnerarchitektur
und Softwaretechnik

User Tools and Languages for Graph-based Grid Workflows

Global Grid Forum 10 – Berlin, Germany

Grid Workflow Workshop

Andreas Hoheisel
(andreas.hoheisel@first.fraunhofer.de)

Fraunhofer Institute for Computer Architecture and Software Technology – Berlin, Germany

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Outline

Fraunhofer Resource Grid

Grid Job Orchestration
- Petri nets
- Grid Job Builder

Workflow Enactment
- Dynamic Workflow Refinement
- Fault management
- Parameter Studies
- Grid Job Handler

Conclusions and future work

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Fraunhofer Resource Grid (FhRG)

Objectives

Development and implementation of a stable and robust Grid infrastructure

Software layer on top of Globus to enable fast realizations of distributed applications (computational science & engineering)

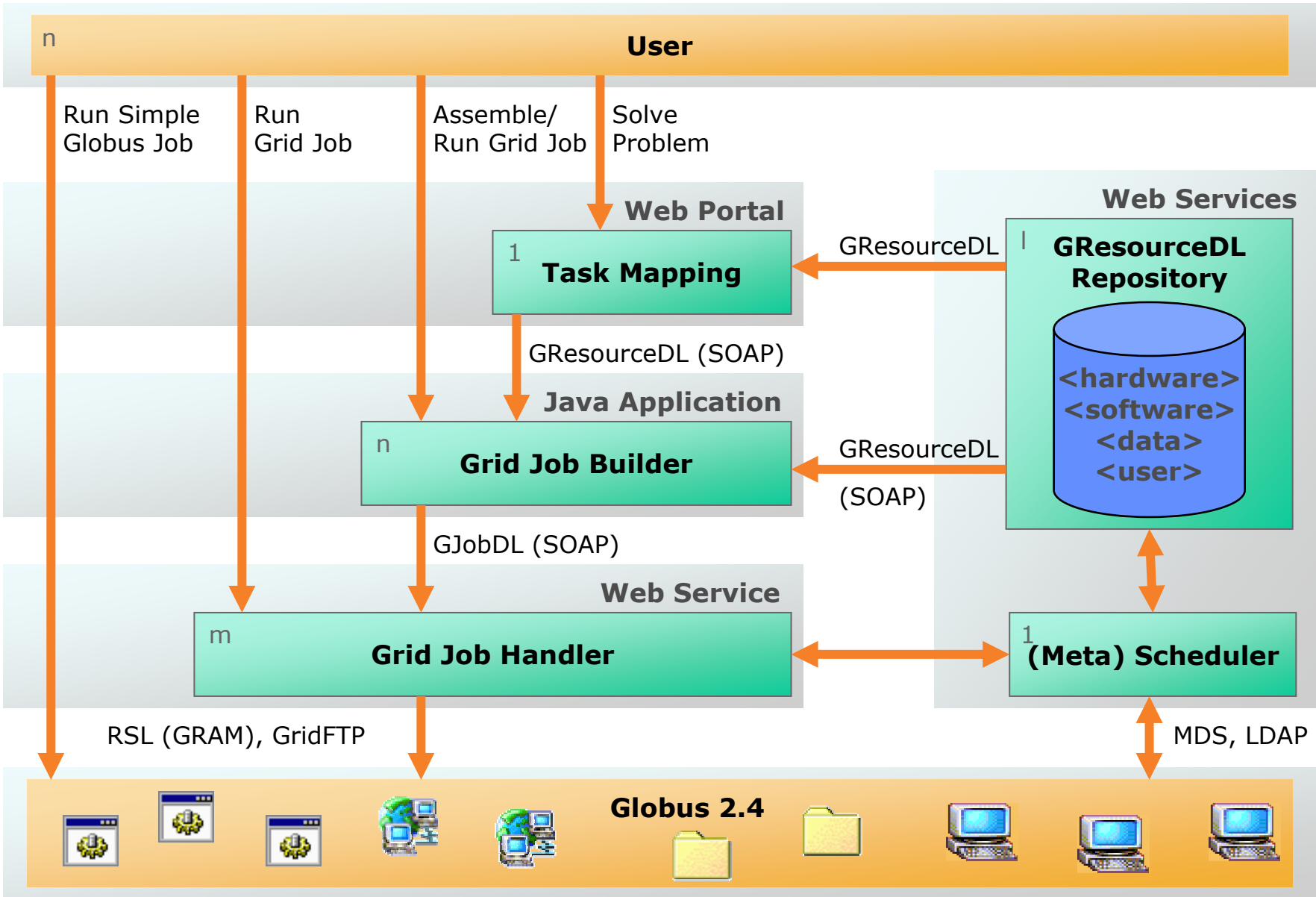
Integration of available resources

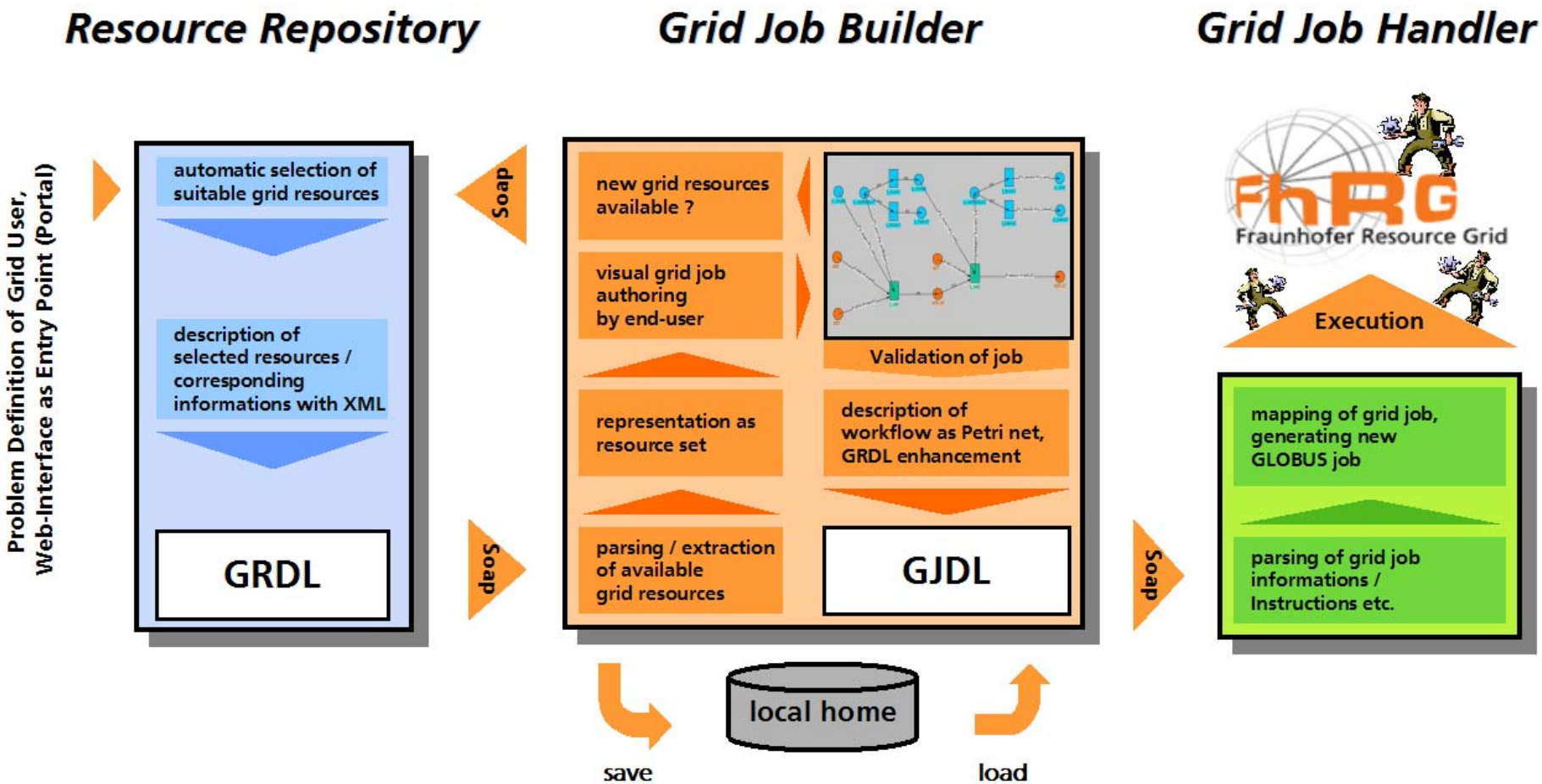
eXeGrid
execution grid

Open Source Software:
http://www.eXeGrid.net/

FhRG Fraunhofer Resource Grid

Fraunhofer FIRST Institut Rechnerarchitektur und Softwaretechnik


n
User
Run Simple Globus Job
Run Grid Job
Assemble/ Run Grid Job
Solve Problem
Web Portal
1
Task Mapping
GResourceDL
Web Services
l
GResourceDL Repository
<hardware>
<software>
<data>
<user>
GResourceDL (SOAP)
Java Application
n
Grid Job Builder
GResourceDL
(SOAP)
GJobDL (SOAP)
Web Service
m
Grid Job Handler
1
(Meta) Scheduler
RSL (GRAM), GridFTP
MDS, LDAP
Globus 2.4



FhRG Fraunhofer Resource Grid

FIRST

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Grid Job Orchestration

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

What is a Grid Job?

Grid job	Composition of Grid resources forming grid applications
Grid resource	Software, hardware, data, (people)
Atomic job	Single task, indivisible component of a Grid job Execution of a software component with input data Future: Invocation of a WebService method call
Component Model	Loosely coupled software components Executables that read input files and write output files Communication via files and GridFTP
GJobDL	Description of Grid jobs on an abstract level Independent from Grid infrastructure Connecting software components and data Based on Petri Nets XML

FhRG Fraunhofer Resource Grid

FIRST

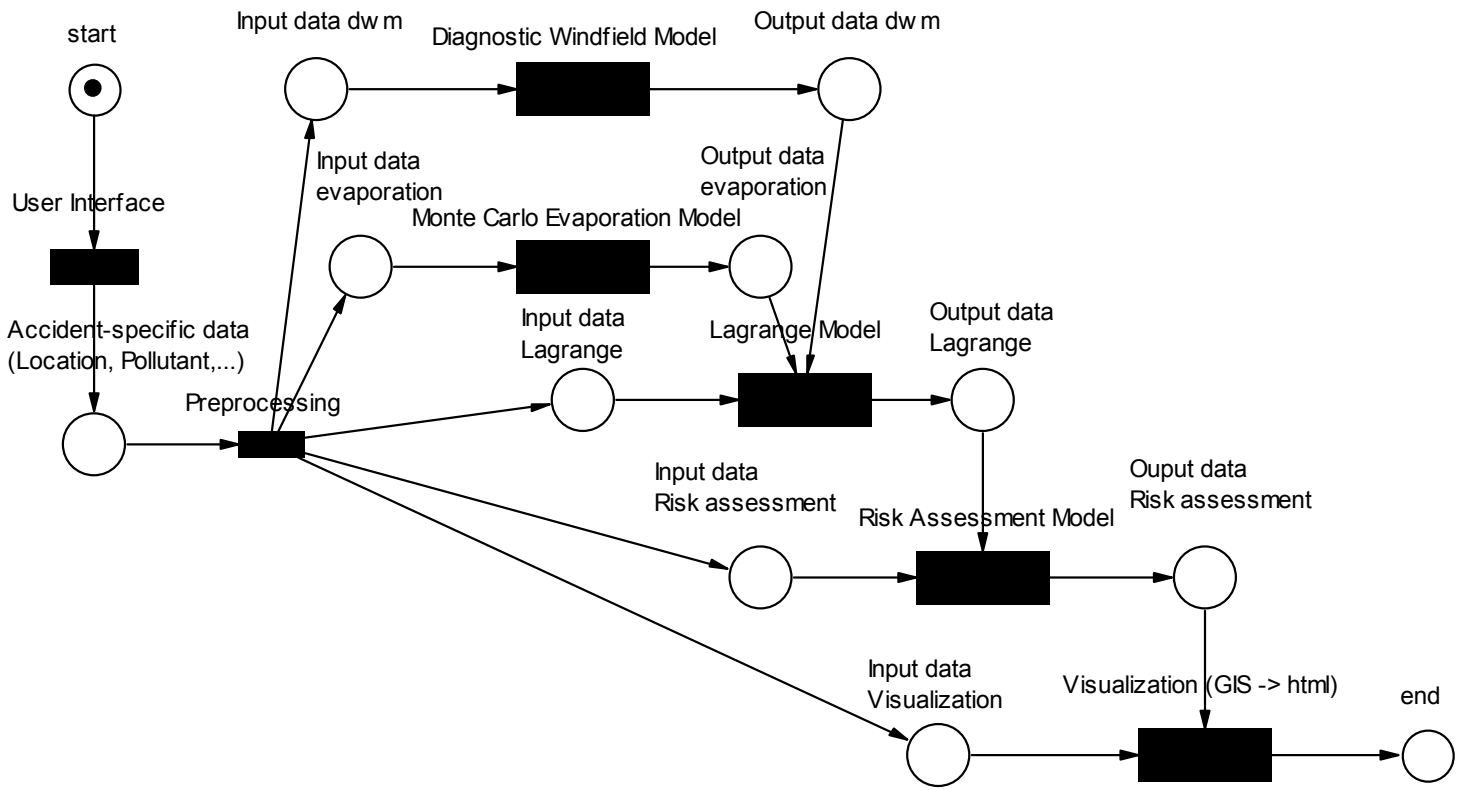
Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Different Grid Workflow Approaches

Inherent model	Workflow is defined inside the software components (e.g. MPI, CORBA, Cactus)
External model	Workflow is defined on top of software components Complete view of workflow (e.g. scripts or graphs)
Scripts	GridAnt, JPython (XCAT)
Combined scripts + graphs	WSFL, XLANG, BPEL4WS, UNICORE, GSFL
Graphs	DAG: Condor DAGman, Symphony Petri net: GJobDL (Fraunhofer Resource Grid)

ERAMAS – Pollutant Transport in the Atmosphere:

Accident → Source → Atmospheric Transport → Exposure



Fraunhofer Resource Grid

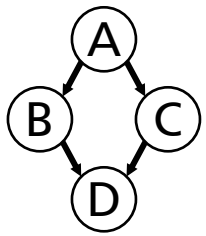
Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Why Petri Nets?

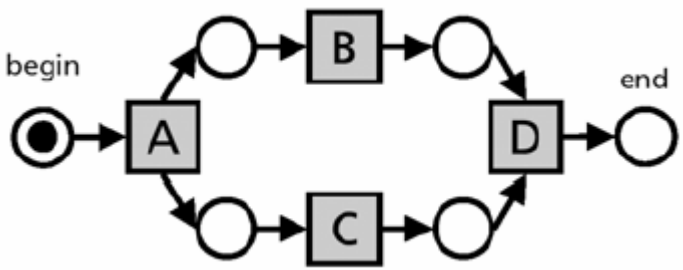
Problem — Description of complex workflows of grid jobs

DAG — Directed Acyclic Graph (see Condor, Cactus, UNICORE)
no loops

```
PARENT A CHILD B C
PARENT B C CHILD D
```



Petri Nets — Graphical flow control of discrete systems
Directed graph, can be made Turing complete



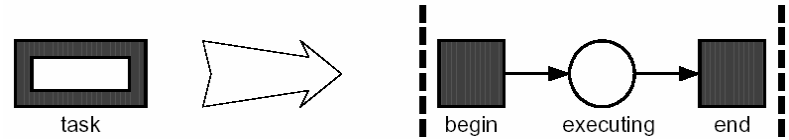
Petri Nets

○	**Places**	Files, buffers, control places
□	**Transitions**	Software components, control transitions
○→□	**Arcs from places to transitions** (Place is input place of transition)	
□→○	**Arcs from transitions to places** (Place is output place of transition)	
●	**Tokens**	Data, State (done, failed)
Rule		A transition is activated if all input places are filled with tokens and all output places have not reached their maximum capacity of tokens
Refinement		A single part of a Petri Net can be replaced by a sub Petri Net
Description of state		A Petri Net describes workflow <u>and</u> state of a system

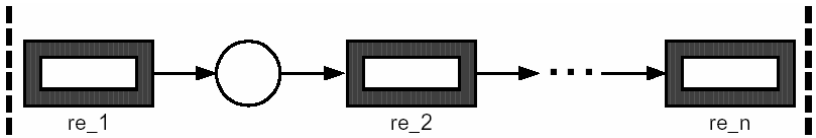
Petri Nets

(from *van der Aalst und Kumar, 2000*)

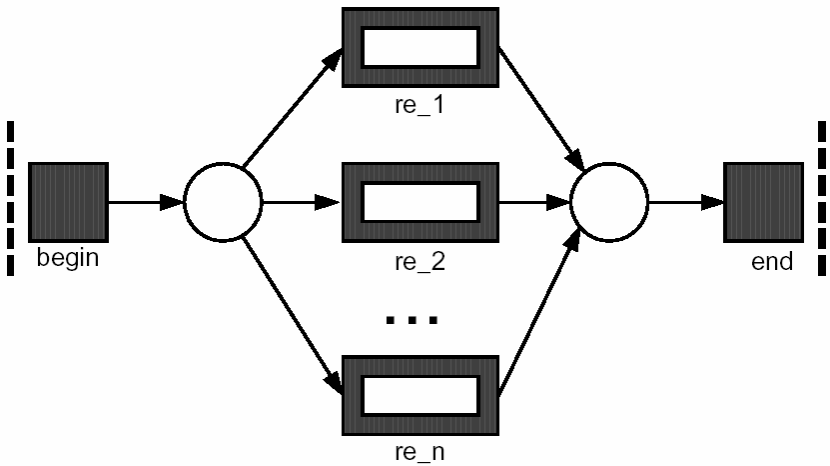
Task



Sequence

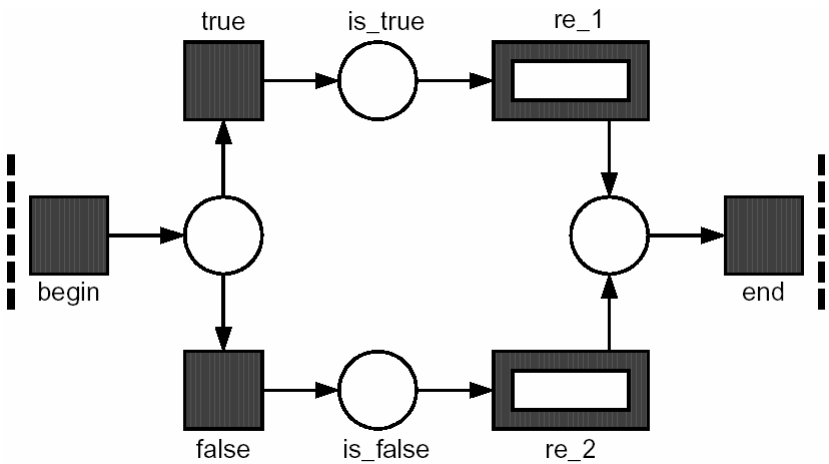


Choice



Petri Nets

Condition

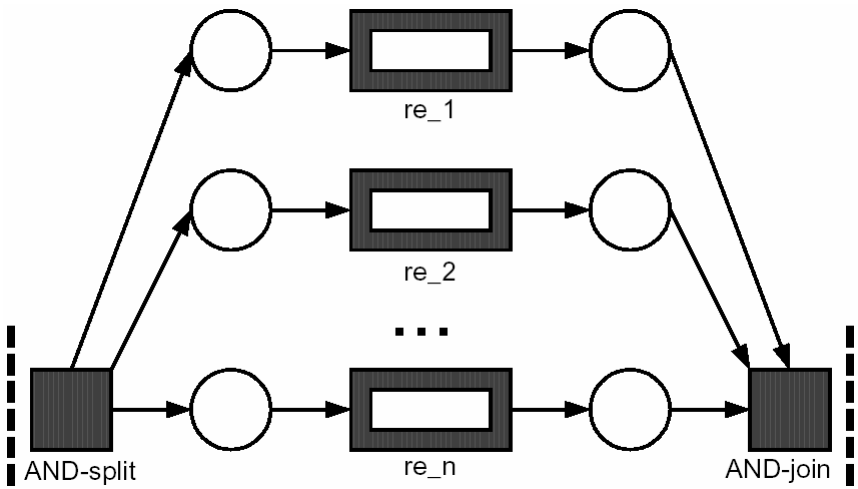


Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Petri Nets

Parallel execution with synchronization

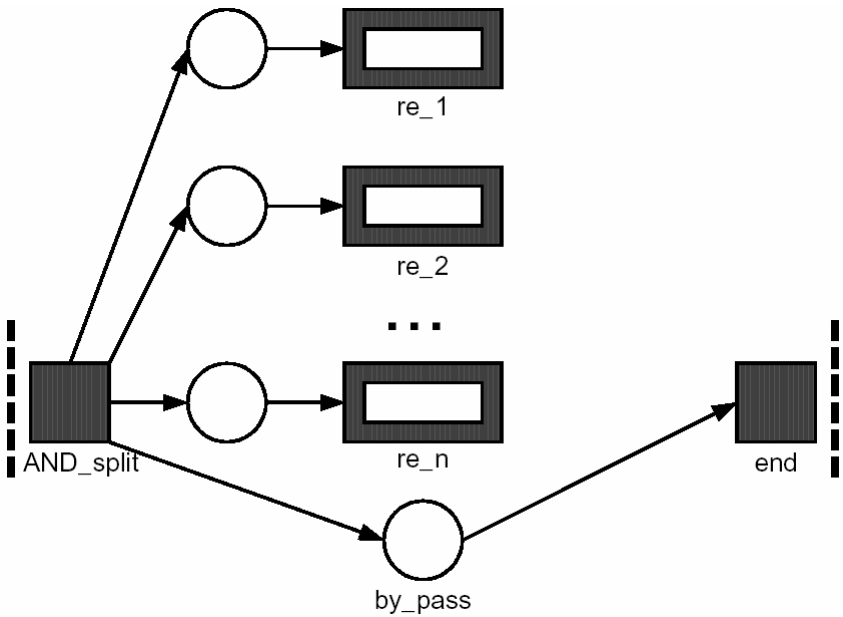


Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

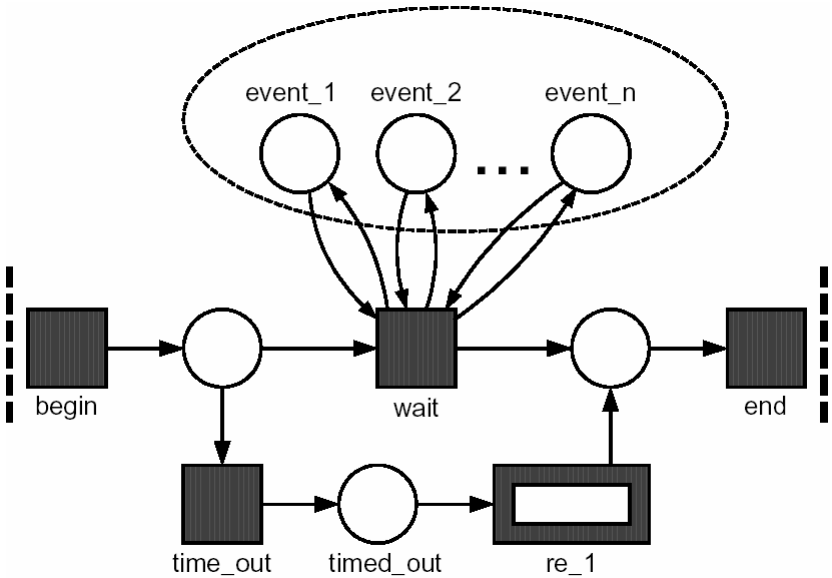
Petri Nets

Parallel execution without synchronization



Petri Nets

Wait all with time out

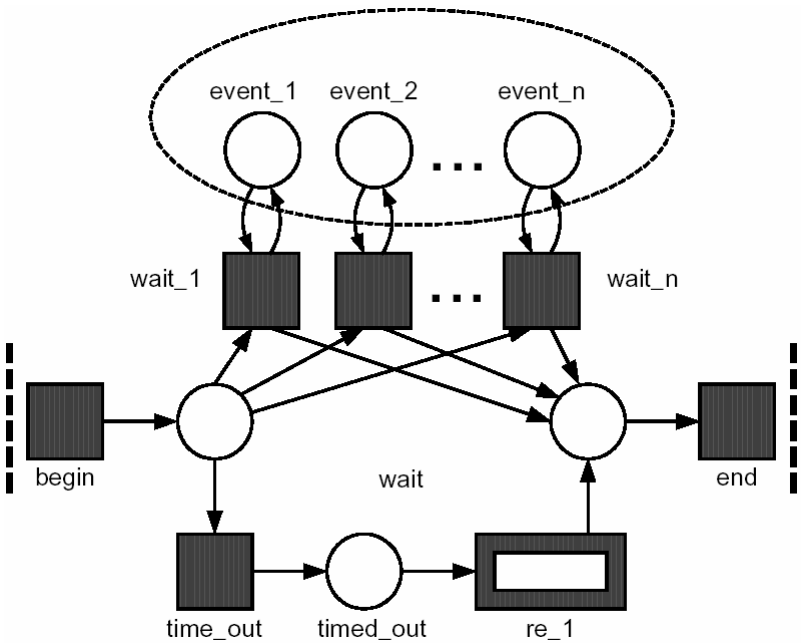


Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

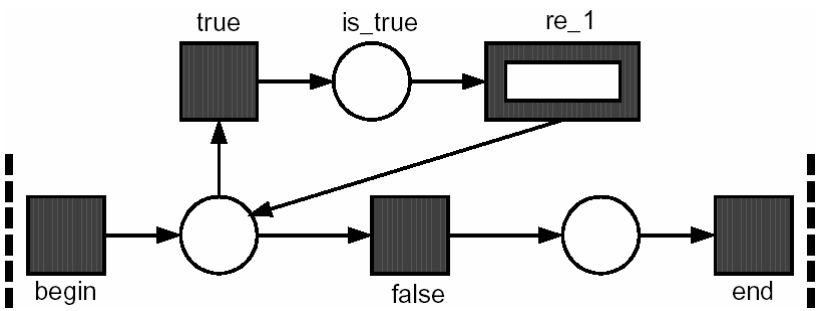
Petri Nets

Wait any with time out



Petri Nets

While … do



Grid Job Builder - Visual Grid Job Authoring

File Edit View Compose Options Help

Construction Kit

Components

Grid Resources | Templates

Available Resources

- data
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
  - http://pollux.swm.iao.fhg.de
- software
  - lagrange model
  - Preprocessing for FRAM
  - dia
  - eva

Grid Job Definition Area

p_mit95Tmp
p_lagrangeEnd
lagrangeFailed
lagrangeDone
p_latim
p_dwmFailed
p_line
p_vis5dTmp
p_arc060
p_param
p_arc045
p_arc030
P:p_mit95Tmp->P:source
P:p_latim->P:latim
P:lagrange->P:p_lagrangeEnd
P:p_arc060->P:in060
P:p_param->P:param
P:p_arc045->P:in045
P:p_arc030->P:in030
P:p_arc015->P:in015

Grid Job Inspector

Petri Net Elements

Grid Job Resources

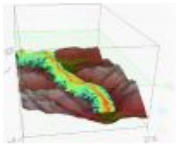
- Requested Resources
  - data
  - software

lagrange model

```
type software
id   http://pollux.swm.iao.fhg.de/grdl/sw/lagrange-20031105.xml
```

In the **Lagrangian approach** of the *dispersion calculation*, concentrations of *gases* are represented by particles as well as real particulate matter. The model uses *winds* and *fluctuation* computed by the **Diagnostic Windfield Model (DWM)** to ***predict the pathways of individual particles or air volumes***. As long as an output time step has not been reached, the individual particles can be considered as fully independent. This makes the model well suited for *numerical parallelization*.

The software provides the following ports as connections:



INCOMING		OUTGOING	
latim	dataset (latim.dat)	**stdout**	stream for output
line	dataset (line.dat)	**stderr**	stream for error messages
lapos	dataset (lapos.dat)	**vis5d**	dataset (mcconc01.v5d)
dwm	dataset (dwm.out)		
oro	dataset (oro.in)		

(...)

ofer.de/fhrg/gjdl1_1.xsd"

Console | Grid | Preview

Workflow Enactment

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

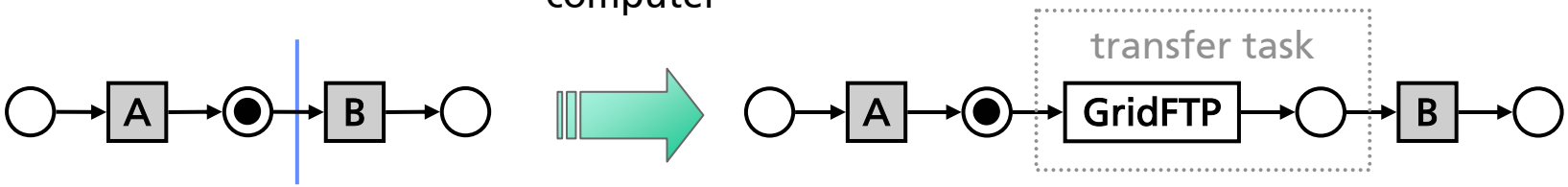
Dynamic Workflows

Refinement

The Grid Job Handler supplements the workflow during runtime by introducing additional tasks

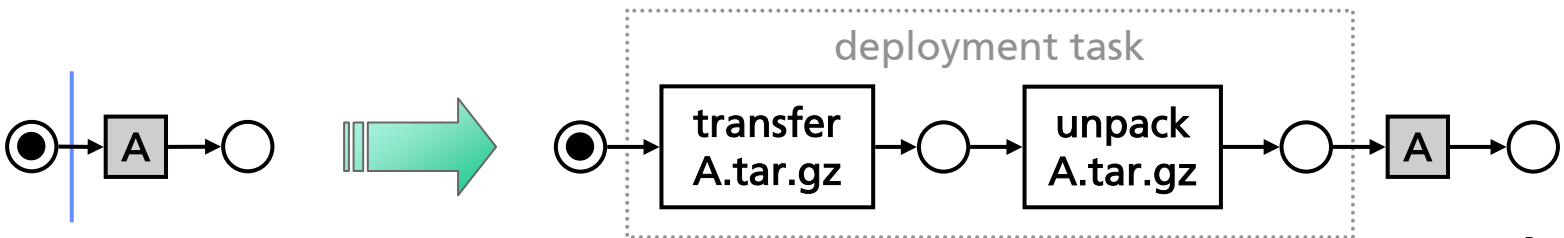
File transfer

A GridFTP task may be introduced automatically to transfer files that are not available on the remote computer



Software deployment

A software deployment task may be introduced to install software components on a remote computer



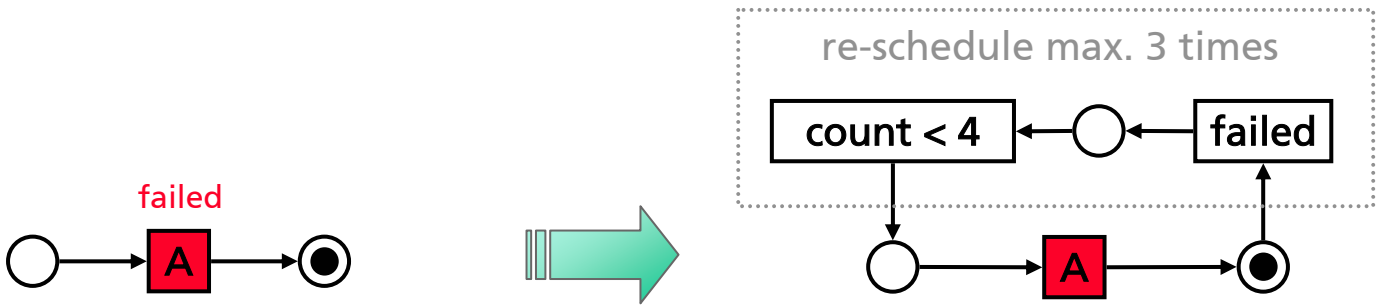
Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Implicit Fault Management

Grid middleware — Fault management that is included in the Grid middleware

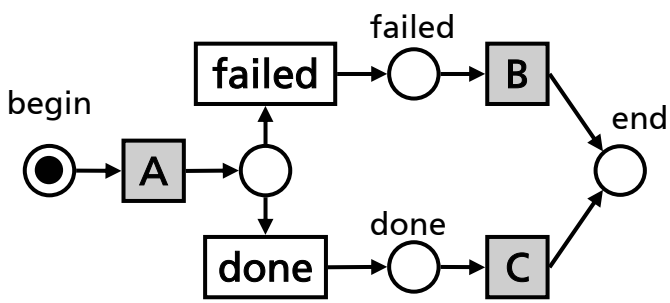
Petri Net refinement — Fault management tasks are introduced automatically if the submission or execution of an atomic task fails



Explicit Fault Management

Petri Net workflow model

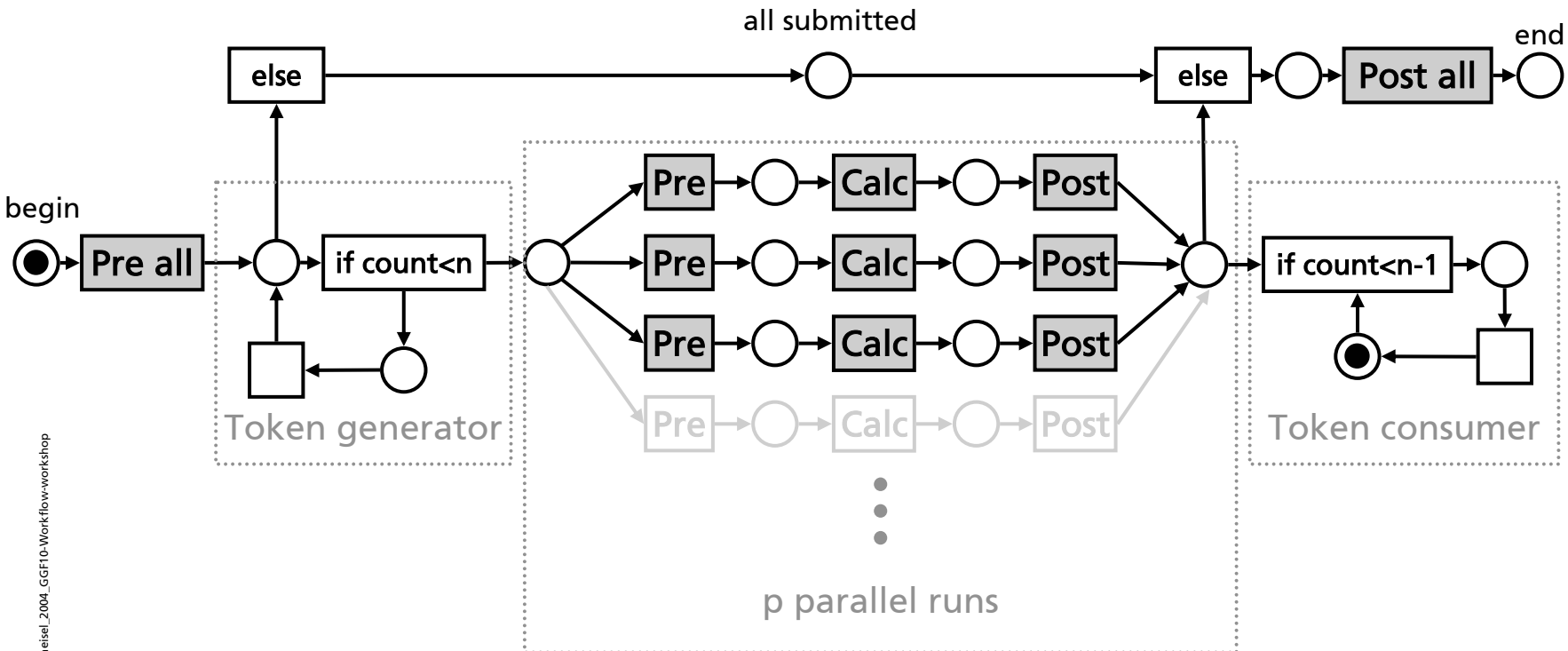
The user defines the fault management explicitly by including user-defined fault management tasks in the Petri Net of his Grid job



Parameter Study

n = total number of runs (e.g. 10,000)

p = number of parallel runs (e.g. 128)

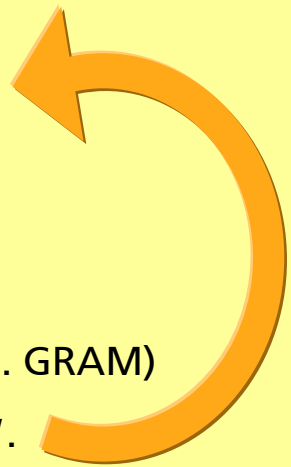


Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Application flow of Grid Job Handler

- Read the GJobDL document
- Create Petri Net from this job description
- Verify the Petri Net (well-formedness, liveliness, deadlocks, pits, ...)
- Start the Grid Job (own thread)
- Collect all activated transitions
- Evaluate conditions
- Invoke resource mapping → repository, (meta-)scheduler
- **Refine the Petri Net**
  → insert GridFTPs, fault management, etc. if necessary
- Create and submit atomic jobs using grid middleware (e.g. GRAM)
- The transition fires, if atomic job is "done" or has "failed".



Fraunhofer Resource Grid Control Desktop

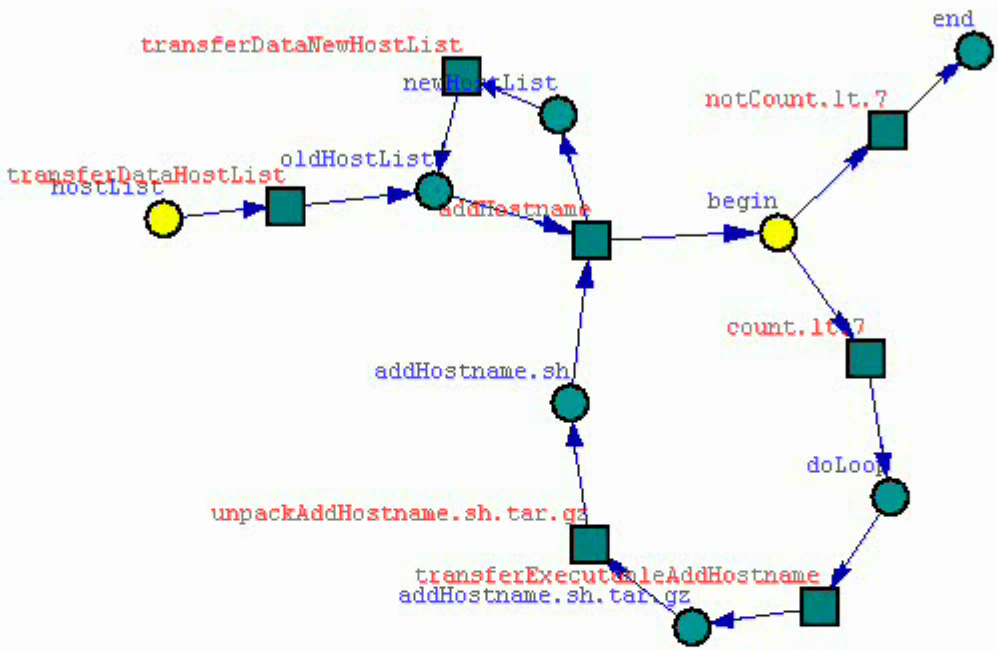
File Grid Job Security Help

Fraunhofer Resource Grid Control Desktop

File Grid Job Security Help

touchgraphed Petri-Net for GridJob #3

Navigate Edit Zoom



Grid Job #3

Run Stop Show Workflow Show new Workflow Show Map

status	start time	end time	ID	resourceMa...	executable	arguments	stdout	stderr

Fraunhofer Resource Grid Control Desktop

File Grid Job Security Help

touchgraphed Petri-Net for GridJob #5

Navigate Edit Zoom

p_failed1
t_failed1
p_exitStatus1
t_done1
d25
d26
t_cat2_transferExecutable
p_cat1_exe
t_cat1_unpack
p_cat1_exe.tar.gz
t_cat1_transferExecutable
p_begin
p_cat2_exe.tar.gz
t_cat2_unpack
p_cat2_exe
d27
d25-27
p_exitStatus2
t_done2
p_end

Grid Job #5

Run Stop Show W

status st... e... ID r... e... ... ...

Fraunhofer Resource Grid Control Desktop

File Grid Job Security Help

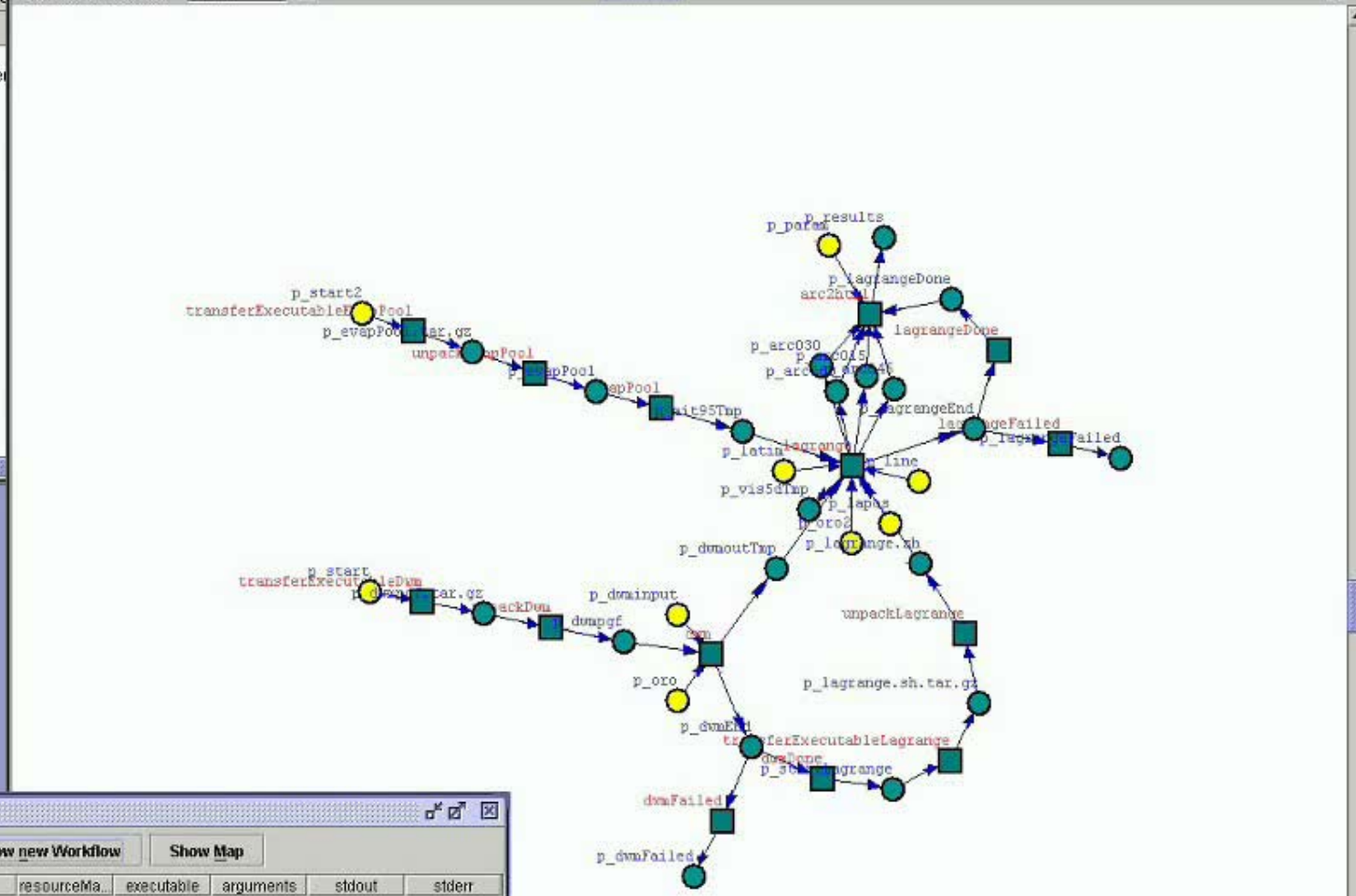
GJobDL: C:\hoheisel\I-Lab\executor\examples\era

Create Grid Job

```
<!-- workflow description -->
<job type="petriNet" id="http://mmm.first.fhg.de/gid/er
  <!-- places dwm -->
  <place id="p_start">
    <initialMarking>
      <value type="boolean" op="eq">true</value>
    </initialMarking>
  </place>
  <place id="p_dwmpgf.tar.gz">
    <resourceRef id="tmp1" type="data"/>
  </place>
  <place id="p_dwmpgf">
    <resourceRef id="tmp2" type="data"/>
  </place>
  <place id="p_dwminput">
    <resourceRef id="dwminput" type="data"/>
    <initialMarking>
      <value type="boolean" op="eq">true</value>
    </initialMarking>
  </place>
```

touchgraphed Petri-Net for GridJob #1

Navigate Edit Zoom



Grid Job #1

Run Stop Show Workflow Show new Workflow Show Map

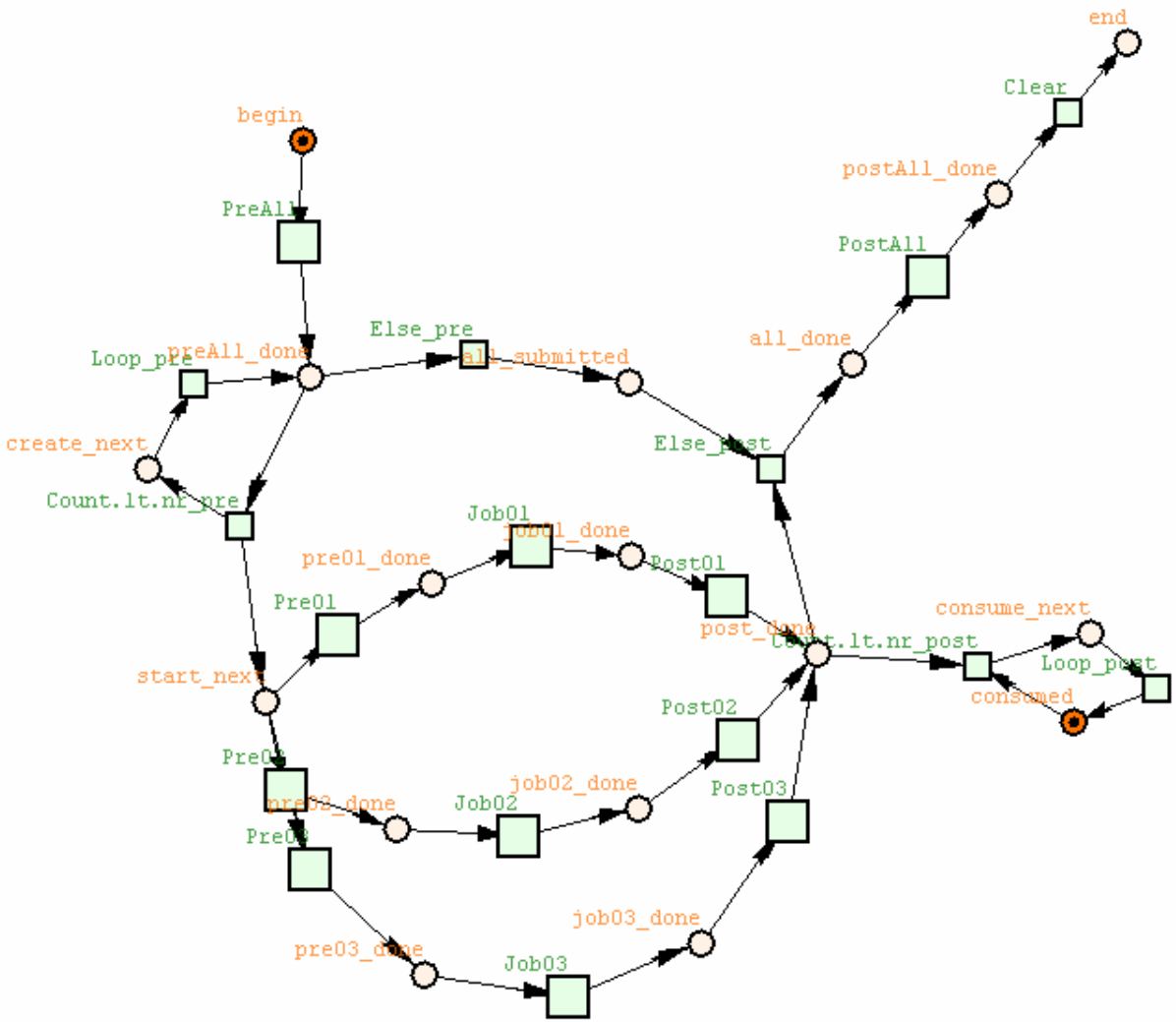
status	start time	end time	ID	resourceMa..	executable	arguments	stdout	stderr

Fraunhofer Resource Grid Control Desktop

File Grid Job Security Help

touchgraphed Petri-Net for GridJob #1

Zoom



Conclusions and Future Work

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Conclusions

Description of workflow	External, graph-based workflow definition GJobDL uses Petri Nets instead of directed acyclic graphs to model workflow of Grid jobs
Petri Nets	Easy orchestration of complex workflows, including conditions and loops
Dynamic workflow model	Petri Nets can be refined and modified during runtime Adding new tasks to the workflow, e.g.: transfer tasks software deployment tasks fault management tasks
Fault Management	implicit → automatic, included in Grid middleware explicit → user-defined, included in workflow model

FhRG Fraunhofer Resource Grid

FIRST

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Future Work

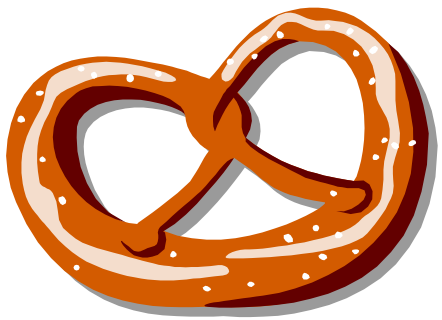
Tight coupling scheme?	Now: one transition → one executable Future: one transition → one method call (?) ➔ Grid Service (OGSA), Web Service
Simulation of Petri Nets	Prediction of Petri Nets for advanced reservation of resources (→ scheduler) based on software und hardware benchmarks
Fault management	Workflow check pointing, recovery of grid jobs

FhRG Fraunhofer Resource Grid

FIRST

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

More Information: http://www.fhrg.fhg.de/
http://www.andreas-hoheisel.de/
andreas.hoheisel@first.fraunhofer.de



Wednesday, March 10th,
5:30pm – 7:00pm

Senate Hall and Senate Hall corridor.

Please join us for some German beer and pretzels sponsored by the Fraunhofer Grid Alliance

Fraunhofer Resource Grid

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik