Grid Workflow Job Execution Service 'Pilot'

Lev Shamardin

Scobeltsyn Institute of Nuclear Physics, Moscow State University (SINP MSU)

XtreemOS Summer School 2010,
Reisensburg, Germany

Background: the GridNNN project

The grid for the participants of the National Nanotechnology Network project in Russia.

- A small number of computer centers, but running supercomputers.
- Unobtrusive integration with the existing supercomputer software.

ГРИД ННС

Background: the GridNNN project

The grid for the participants of the National Nanotechnology Network project in Russia.

- A small number of computer centers, but running supercomputers.
- Unobtrusive integration with the existing supercomputer software.
  - `http://tinyurl.com/msu1nodes`

ГРИД ННС

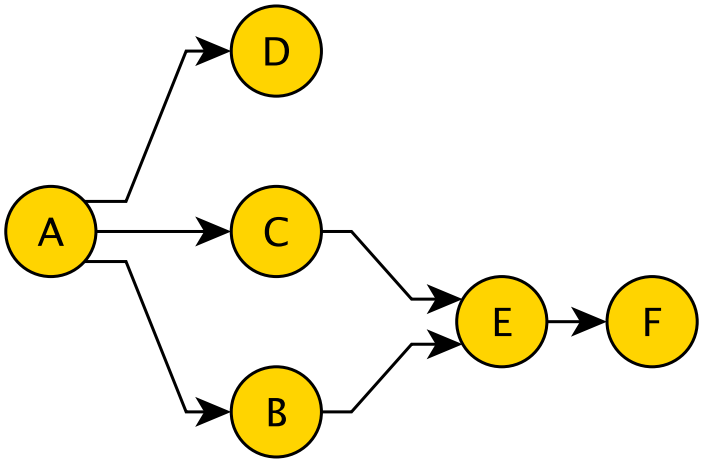
Background: the GridNNN project

The grid for the participants of the National Nanotechnology Network project in Russia.

- A small number of computer centers, but running supercomputers.
- Unobtrusive integration with the existing supercomputer software.
  - `http://tinyurl.com/msu1nodes`
- Multistage workflow jobs.
- Simple interface for the user.
- Big number of VOs.

ГРИД ННС

Workflow jobs

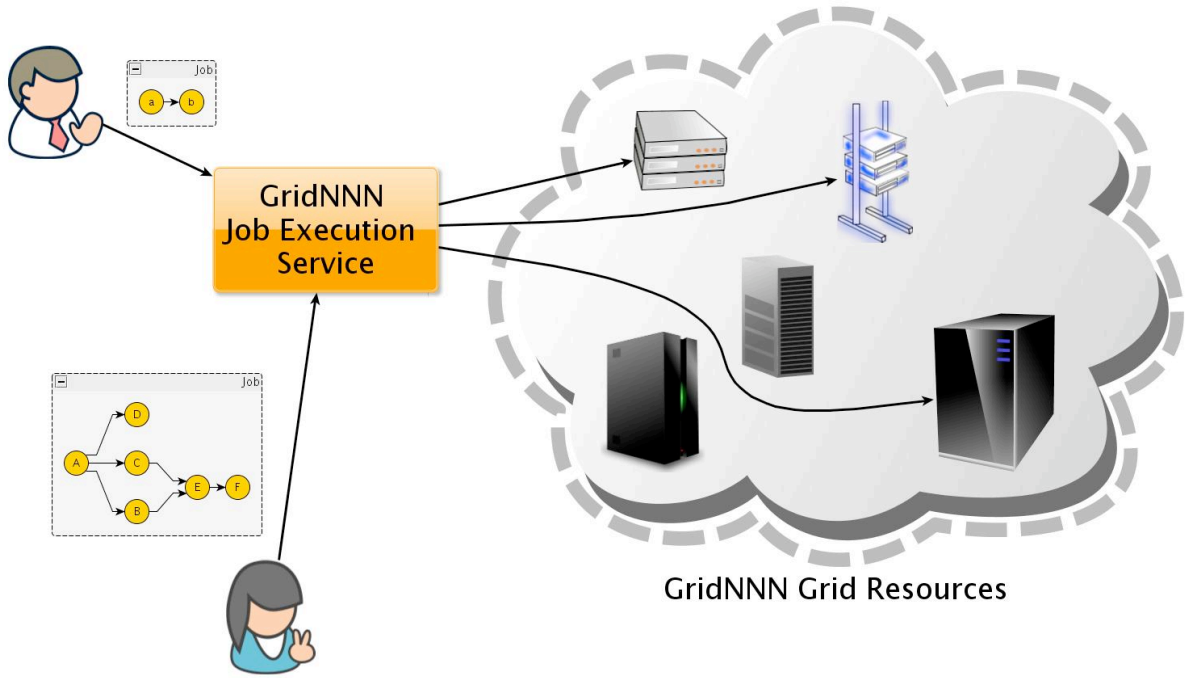


Job A directed acyclic graph. Each node corresponds to a task, edges define the order of execution.

Task Minimal basic executable element.
For example: running a process using a WS-GRAM Service.

ГРИД ННС

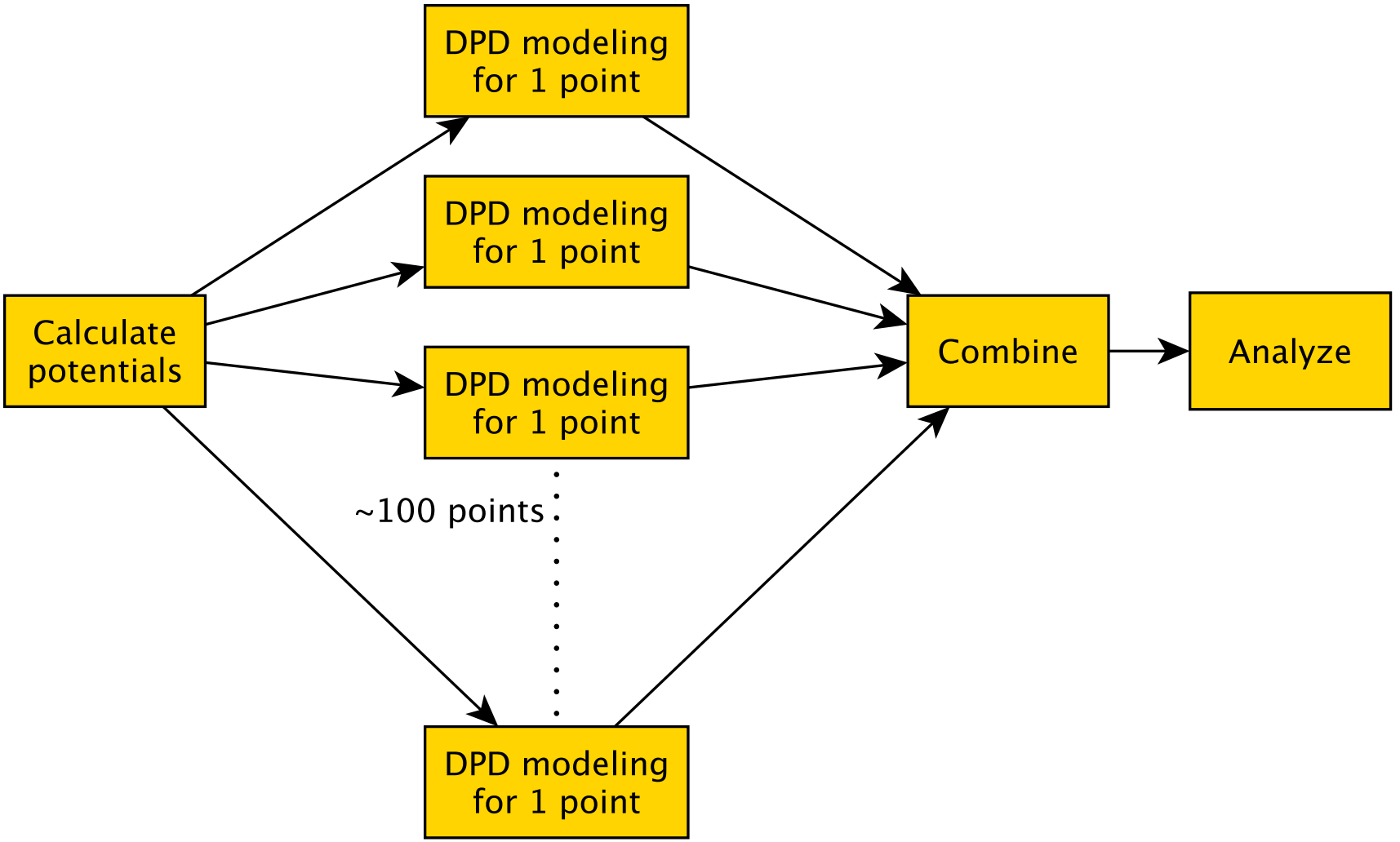
GridNNN Job Execution Service



- Execute jobs submitted by users, discover resources, etc
- Simple API
- Project name: “Pilot”

ГРИД ННС

Example workflow: material properties modeling



ГРИД ННС

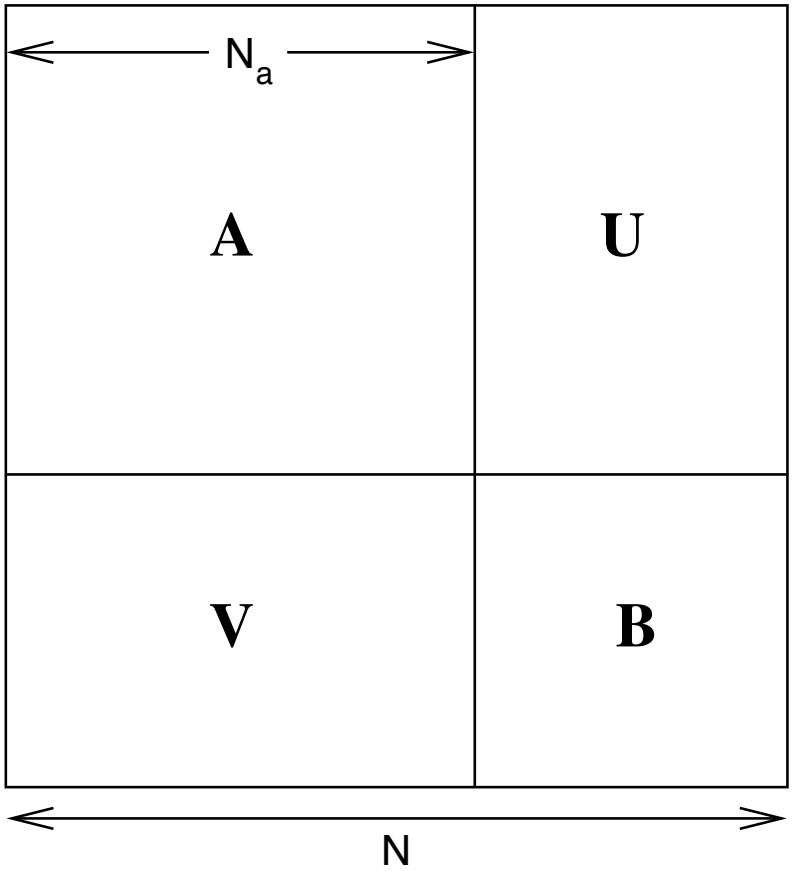
Example: matrix inversion using Schur complement[1]

$$M\,[N \times N],\ M = \begin{bmatrix} A & U \\ V & B \end{bmatrix}$$

$$A = [N_A \times N_A]$$
$$B = [(N - N_A) \times (N - N_A)]$$

$$S = B - VA^{-1}U$$

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}US^{-1}VA^{-1} & -A^{-1}US^{-1} \\ -S^{-1}VA^{-1} & S^{-1} \end{bmatrix}$$

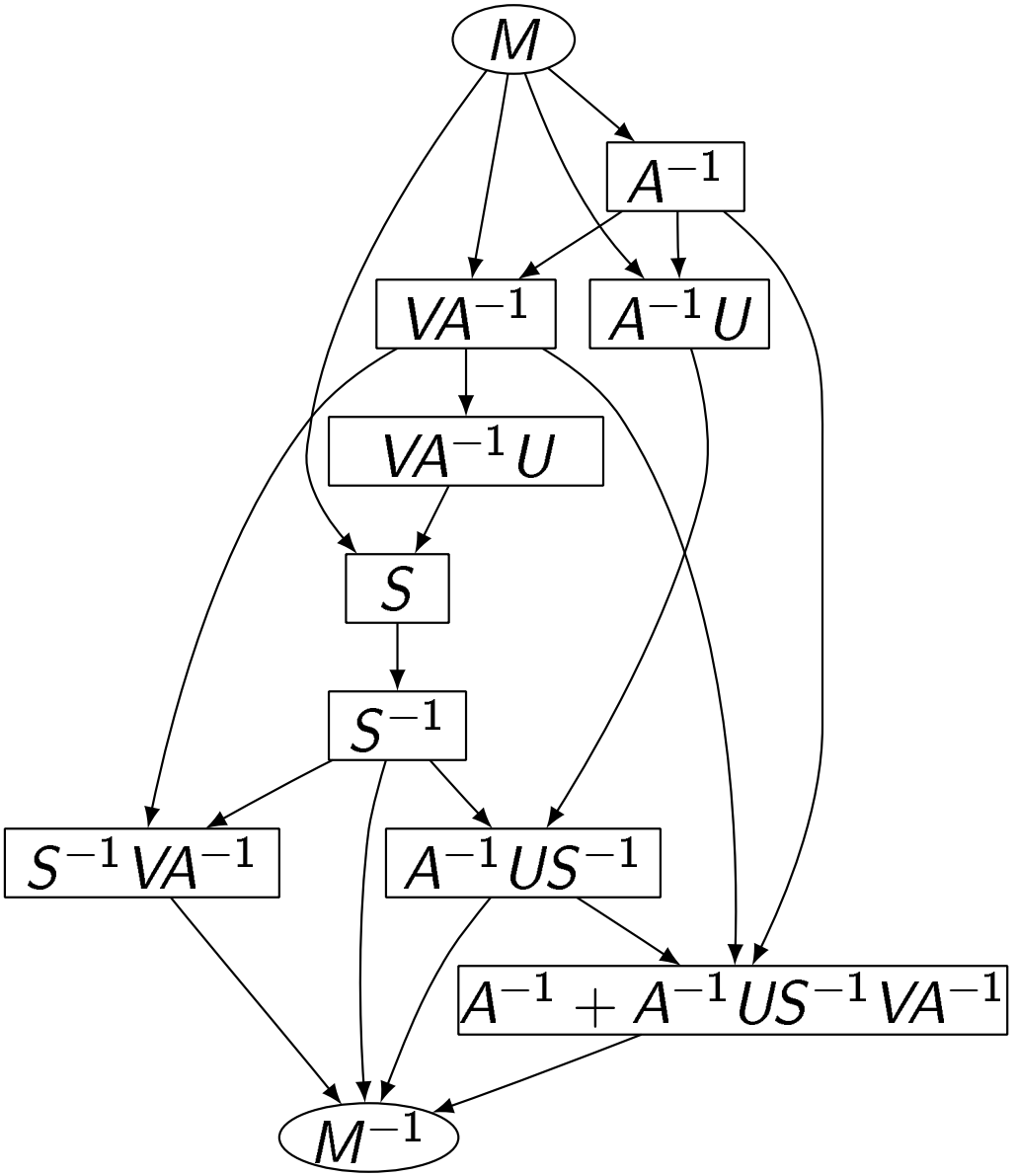


T. Cormen, C. Leiserson, R. Rivest and C. Stein, "Introduction to Algorithms".

[1] This example is courtesy of V. Voloshinov and S. Smirnov

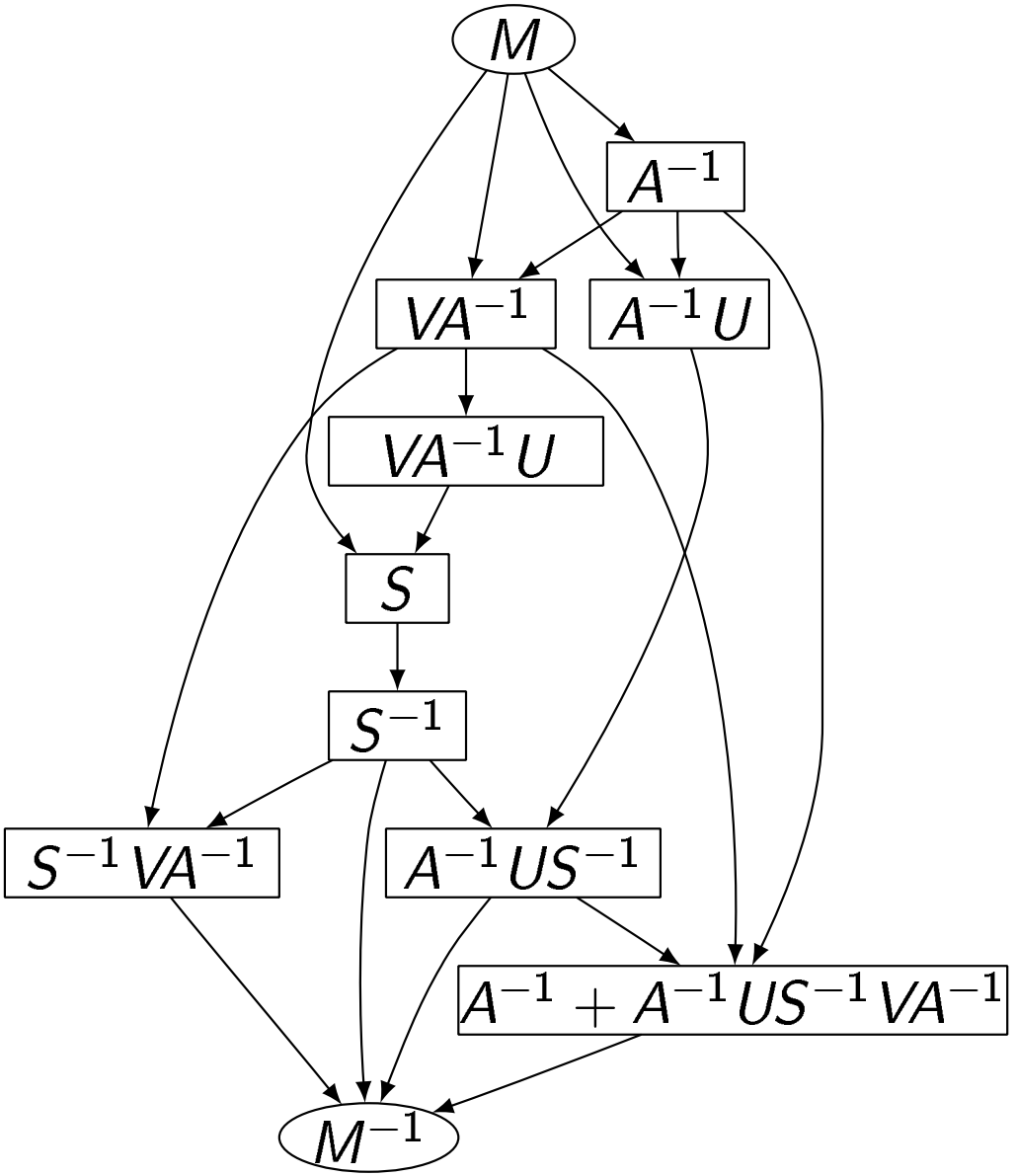
ГРИД ННС

Example workflow: matrix inversion using Schur complement



ГРИД ННС

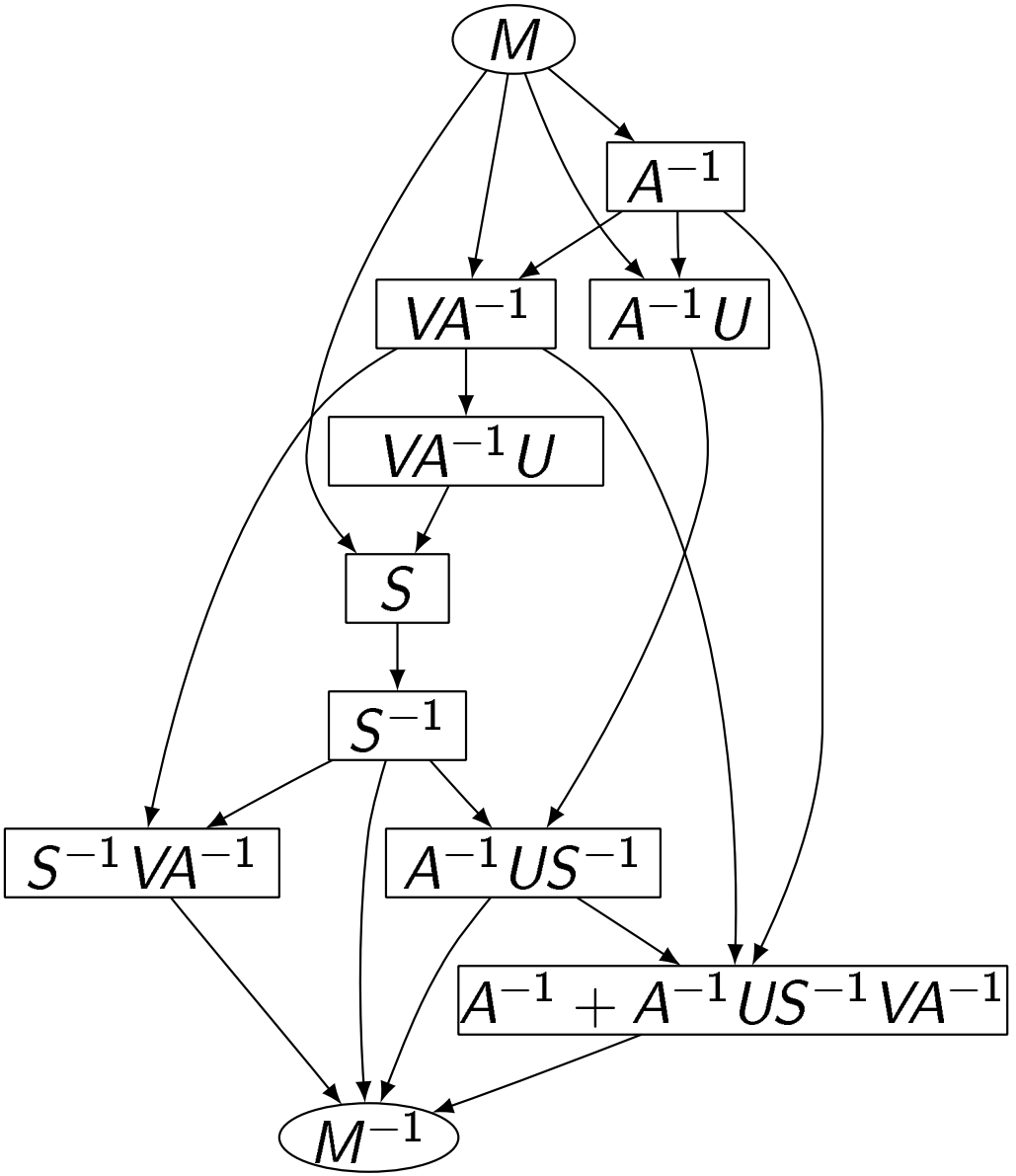
Example workflow: matrix inversion using Schur complement



The result: speedup of 1.3x compared to regular inversion for fixed-precision evaluation.
Even more for symbolic evaluation for large values of N.

ГРИД ННС

Example workflow: matrix inversion using Schur complement



The result: speedup of 1.3x compared to regular inversion for fixed-precision evaluation.
Even more for symbolic evaluation for large values of N.

$N \sim 500$ is large enough for current CPUs.

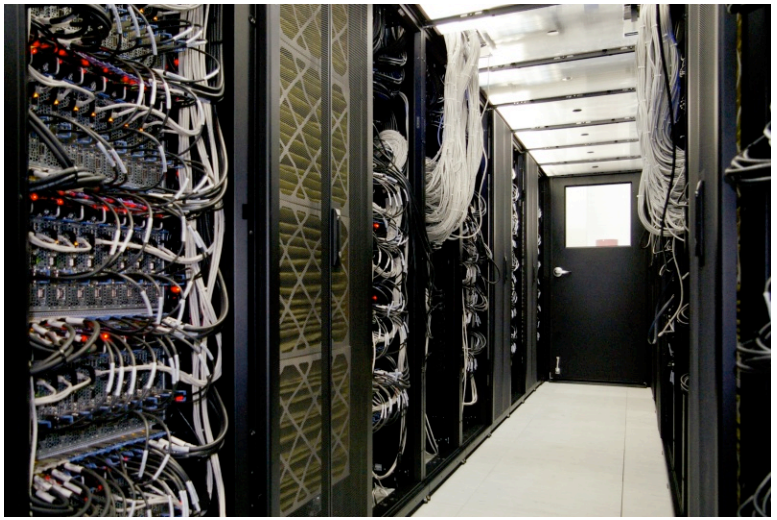
ГРИД ННС

Goals for the Pilot service

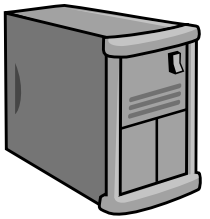
- Orchestrate the execute of jobs on a set of grid resources.
- Select appropriate resources at task execution time.
- Provide a simple job and task description language.
- Provide a simple API.
  - Also provide a CLI.

ГРИД ННС

Grid resources



+



Grid gateway

Grid gateway:

- Globus Toolkit 4 WS-based services
  - WS-GRAM
  - WS-MDS
  - RFT
- GridFTP

ГРИД ННС

Job description

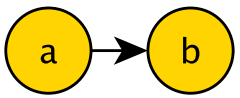
A JSON object containing:

- A list of tasks;
- Task dependencies;
- Common task options;
- Default resource requirements.

ГРИД ННС

Job description: Example

```
{ "version": 2,
  "description": "test job",
  "default_storage_base": "gsiftp://tb01.ngrid.ru/home/shamardin/jt/"
  "tasks": [ { "id": "a",
               "description": "task #1",
               "filename": "task_a.js",
               "children": ["b"]
             },
             { "id": "b",
               "filename": "task_b.js"
             } ]
}
```



Task definition may be inlined to the job definition.

ГРИД НИС

Task description

- Executable file, command-line arguments, environment variables.
- stdin/stdout/stderr.
- Files to be staged in before job execution and staged out after execution.
  - May include the job executable.
  - May be specified as paths relative to some common URL.
  - Must be located in Grid FTP servers.
- Resource requirements.
- Extensions
  - Will be passed to WS-GRAM as is (almost).

ГРИД ННС

Task description: Example

- task_a.js:

```
{ "version": 2,
  "executable": "/usr/bin/whoami",
  "stdout": "task_a.txt"
}
```

- task_b.js:

```
{ "version": 2,
  "executable": "my_wc",
  "arguments": ["-l"],
  "stdin": "task_a.txt",
  "input_files": {
    "my_wc": "gsiftp://tb01.ngrid.ru/usr/bin/wc"
  },
  "stdout": "task_b.txt"
}
```

ГРИД НИС

Resource requirements

- May be specified both in job and task descriptions.
- Host OS:
  - name, release, version;
  - platform;
  - CPU instruction set.
- Node parameters:
  - SMP size;
  - RAM, Virtual memory;
  - CPU speed.
- Software requirements:
  - "mvapich, abinit $>$ 6, gcc==3.5.5"
- Low-level requirements (host name, lrms type, queue name).
- Task resource requirements precede the job resource requirements.

ГРИД ННС

Pilot API

Jobs and tasks

- Create, modify, delete jobs and tasks.
  Tasks are created indirectly from the job descriptions, but may be modified. Any job or task may be modified if it is not running already.
- Get job/task status information.
- Get resource matchmaking information.
- Control the jobs
  Jobs and tasks have state. The fact that the job was created does not mean that it will be (immediately) executed.

ГРИД ННС

Pilot API

Accounting

- Get accounting records for the specified time frame.
  - Records about operations on jobs/tasks.
- Some basic access control
  - User can view his own accounting information.
  - VO manager can view accounting information for his VO.
  - Supervisor access for a set of predefined identities.

ГРИД ННС

Pilot API

REST and JSON

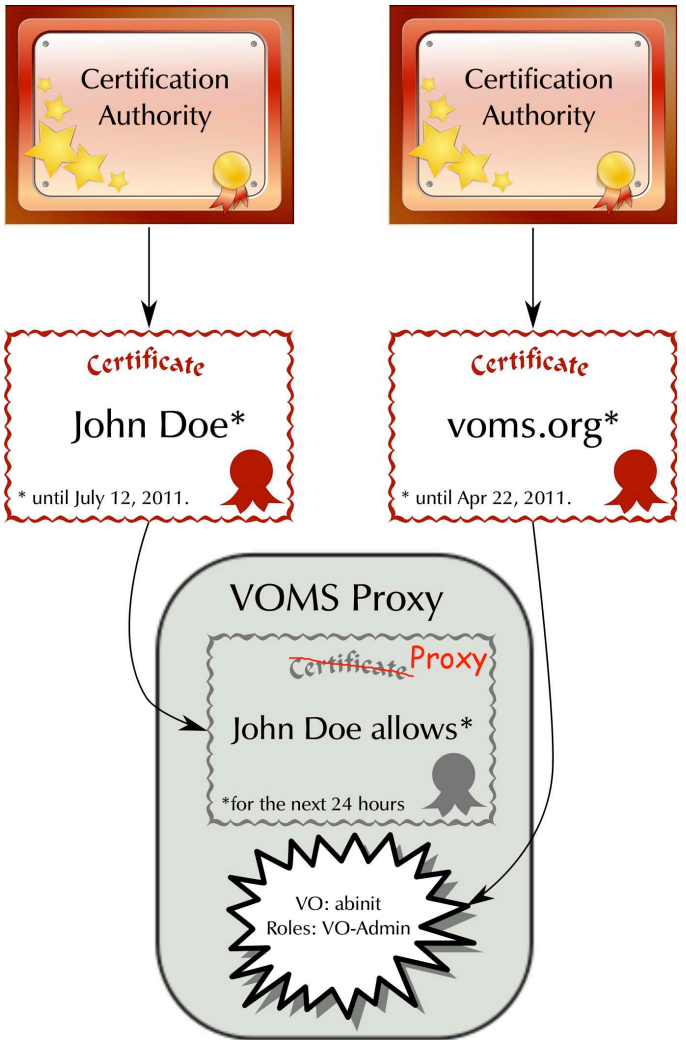
- API designed with simplicity in mind.
- May operate using only idempotent operations (i.e. no `POST`).
  - Support fallback to `POST` for non-trivial operations for simple clients.
- JSON is the main data representation format.
- JSON Schema for all service inputs and outputs.
- Pilot is a RESTful grid service:
  - Resources have a life cycle.
  - Some resources have internal state.
  - Resources have a limited life time.

ГРИД ННС

Pilot API

Authentication and Authorization

- X.509 certificates.
- RFC proxies.
- VOMS attribute certificates



ГРИД НИС

Pilot CLI

Standard set of pilot-*something* commands providing batch-like interface, including:

- job-submit, job-status, job-info (this one includes status of all job tasks), job-cancel
- task-status
- job-matchmake
- query-jobs

ГРИД ННС

Pilot without CLI

$HOME/.curlrc

```
cert = "/tmp/x509up_u500"
cacert = "/home/shamardin/.globus/usercert.pem"
capath = "/etc/grid-security/certificates"
```

Submit a job

```
curl --data @simple_job.js -S -i -H "Content-Type: application/json" \
     -X POST https://tb01.ngrid.ru:5053/jobs/ | awk \
     '$1=="Location:" { print $2 }' > job_uri
```

Query job status

```
curl -S -i https://tb01.ngrid.ru:5053/jobs/3VQLHL8E/
```

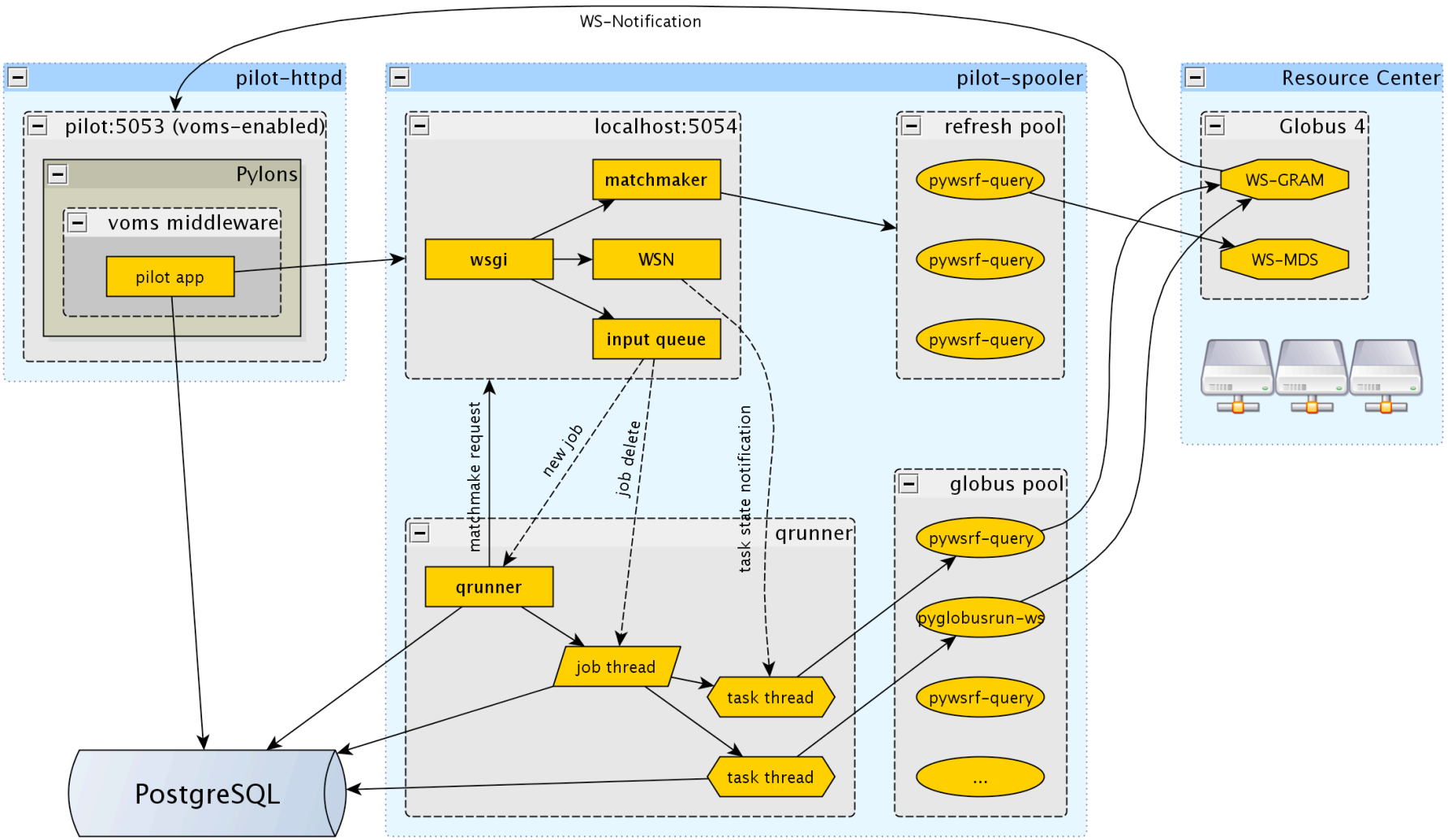
ГРИД НИС

Implementation: Server

- Python
- M2Crypto, libvomsc (using ctypes)
- RDBMS. PostgreSQL and SQLite are supported.
- Frontend:
  - Custom WSGI server
  - Pylons WSGI application
- Backend:
  - Lightweight threads (Eventlet framework)
- Python WSRF libraries (derived from pyGridware)
  - wsrf-query
  - globusrun-ws
- Supports WS-Notification

ГРИД ННС

Server Internals



ГРИД ННС

Resource discovery and matchmaking

Resource information provider can operate in two modes:

- Collect information from a single federated WS-MDS Index Service
- Gather information from a set of WS-MDS Services.
  - List of available services may be obtained automatically from GridNNN Central Service Registry.

Matchmaking:

- JIT Dynamic planning with performance-driven strategy.
- Task-level fault tolerance.

ГРИД ННС

Implementation: CLI

- Python
- Minimal dependencies
- Prepackaged RPMs for CentOS 5 / RHEL 5 are available
  - Server components are also available in RPMs for CentOS 5 / RHEL 5.

ГРИД ННС

Questions?

Thanks for your attention!

ГРИД ННС