CD-HIT Workflow Execution on Grids Using Replication Heuristics

José Luis Vázquez-Poletti - Eduardo Huedo - Rubén S. Montero - Ignacio M. Llorente

dsa-research.org

Distributed Systems Architecture Research Group
Universidad Complutense de Madrid

Universidad Complutense Madrid

La Suma de Todos
Consejería de Educación
Comunidad de Madrid
www.madrid.org

Ministerio de Educación y Ciencia

Contents

"A civilization is built on what is required of men, not on that which is provided for them."

Antoine de Saint-Exupéry

1. Starting Point

The Importance of being Workflow… at the Grid

- **Workflow Management Systems:** Allow the execution of complex applications that can be divided in tasks with dependencies.
- **The Grid:** Offers access to a great amount of computing resources.

Job done till now CCGrid 07

- Representative Bioinformatics application was ported onto the Grid.
- Framework running on production environment.
- Not so good performance results due to Grid's nature…

Algorithm needs to be revisited!

(why we are here)

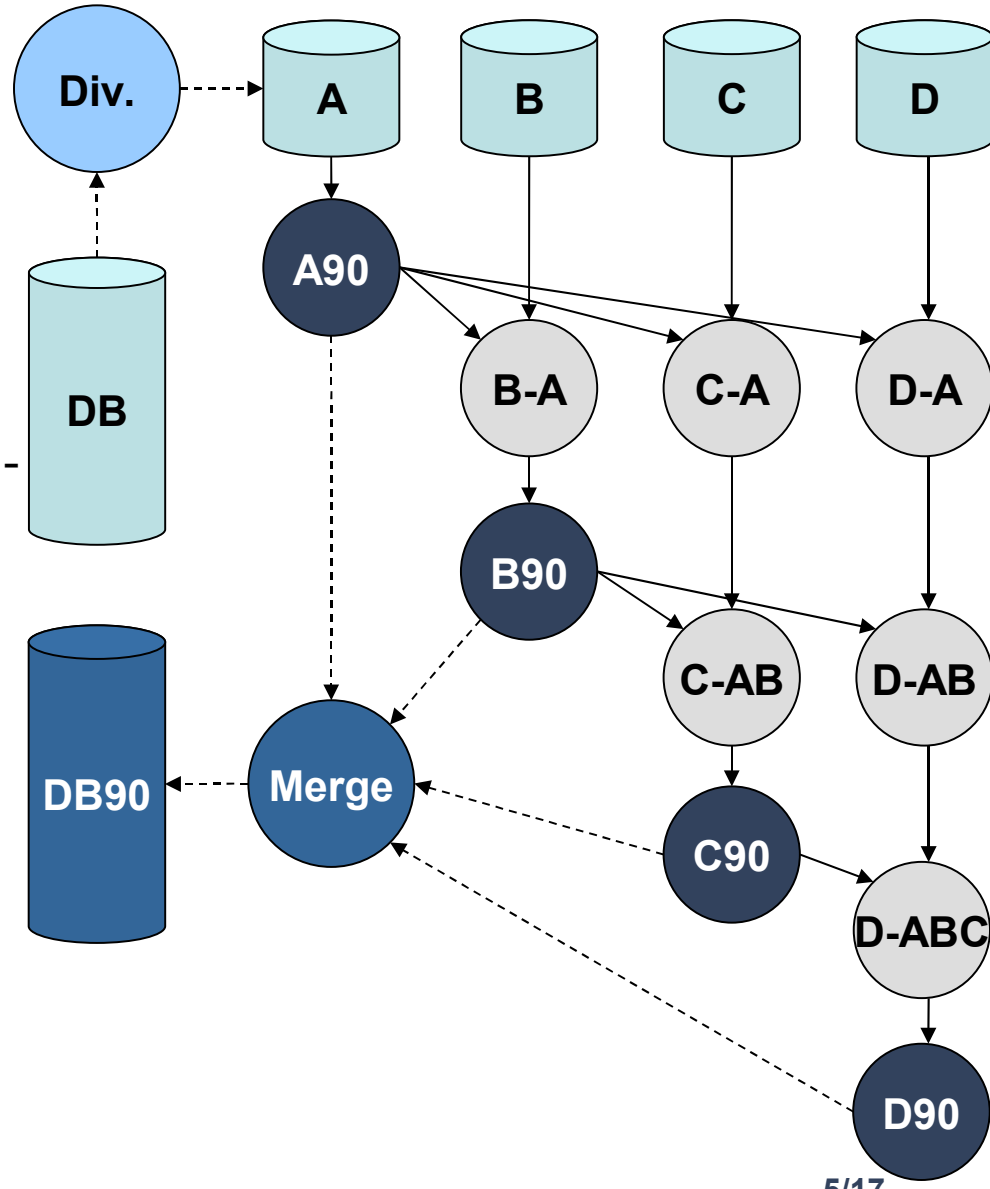
CCGrid 2008

Application Description

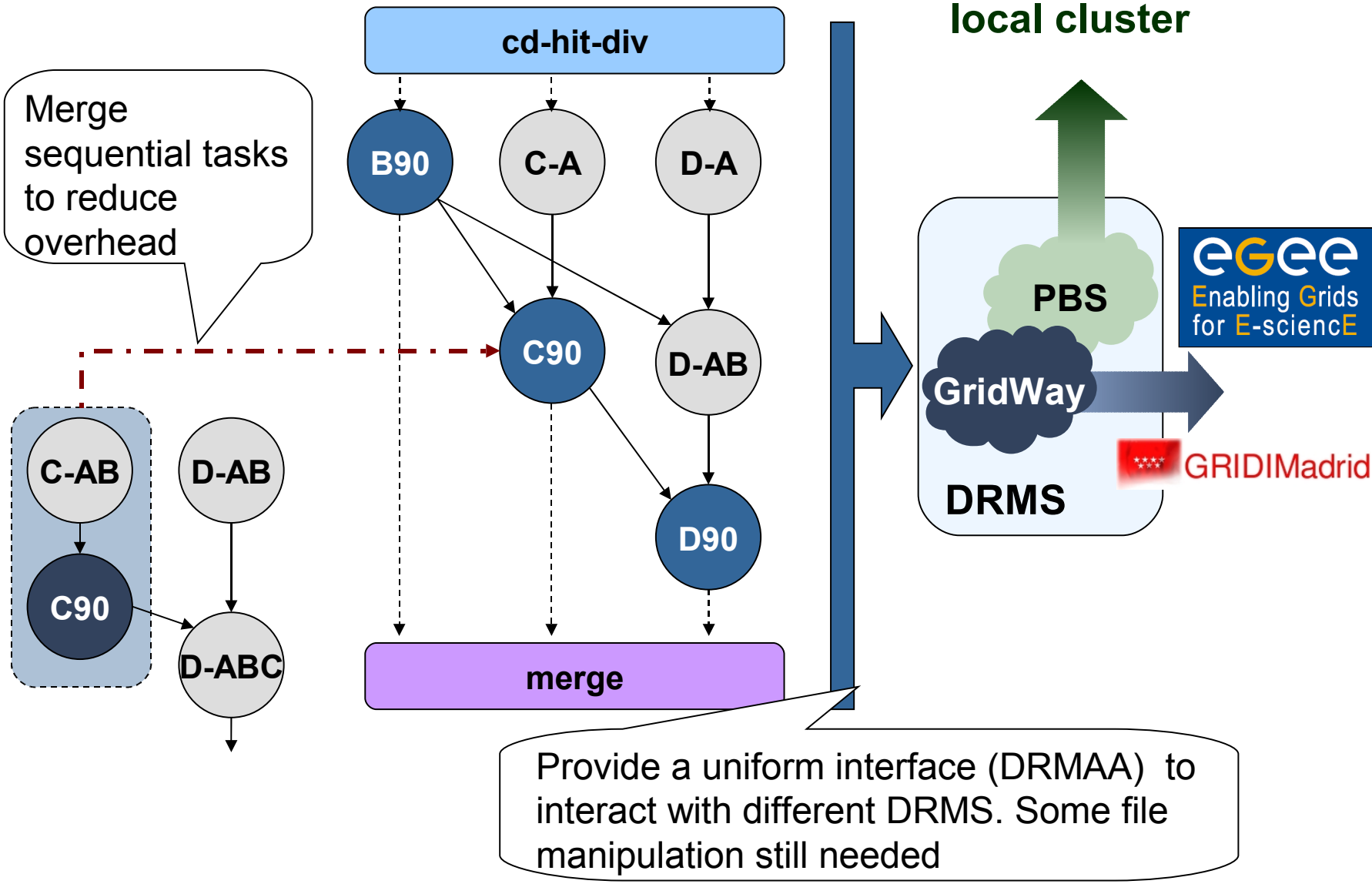
- *"Cluster Database at High Identity with Tolerance"*
- Protein (and also DNA) clustering
  - Compares protein DB entries
  - Eliminates redundancies
- Example: Used in UniProt for generating UniRef data sets
- Our case: Widely used in the Spanish National Oncology Research Center (CNIO)
  - Input DB **now**: 4,186,284 proteins / 1.7GB
- Infeasible to be executed on single machine
  - **Memory requirements**
  - Total execution time
- UniProt is the world's most comprehensive catalog of information on proteins. CD-HIT program is used to generate the UniRef reference data sets, UniRef90 and UniRef50.
- CD-HIT is also used at the PDB to treat redundant sequences

CD-HIT Parallel

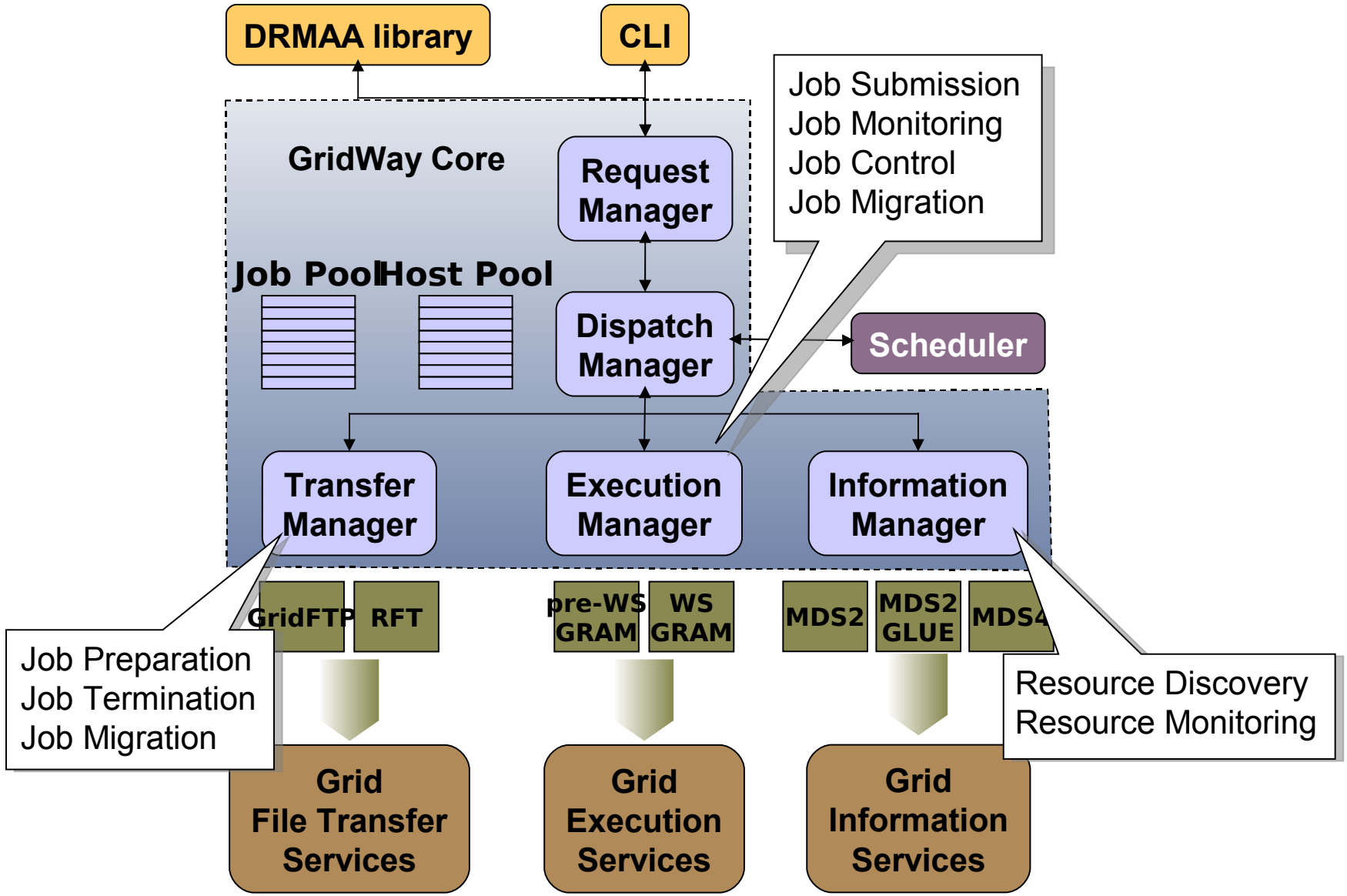
- Execute cd-hit in **parallel mode**
- **Idea:** divide the input database to compare each division in parallel
  - Divide the input db
  - Repeat
    - Cluster the first division (cd-hit)
    - Compare others against this one (cd-hit-2d)
  - Merge results
- Speed-up the process and deal with **larger databases**
- **Computational characteristics**
  - Variable degree of parallelism
  - Grain must be adjusted



2. The CD-HIT Application



GridWay Internals



4. Workflow Optimization Heuristics

List Scheduling

- **Heterogeneous Earliest Finish Time:** schedules tasks minimizing their finishing time in an insertion based manner.
- **Critical Path on a Processor:** detaches a machine just for critical path tasks.
- **Bubble Scheduling and Allocation:** firstly serializes task graph and then inserts tasks to processor.
- **Dynamic Level Scheduling:** delays scheduling when given task is ready.
- **Critical Nodes Parent Trees:** Considers task earliest execution time.
- **Iso-Level Heterogeneous Allocation:** allocates to each processor a number of tasks proportional to its computing power.

Some of them are **implicitly** applied by GridWay
(number of CPU's/Job, scheduling order because of critical path,…)

Agglomeration

Clusters jobs for reducing communication overheads.

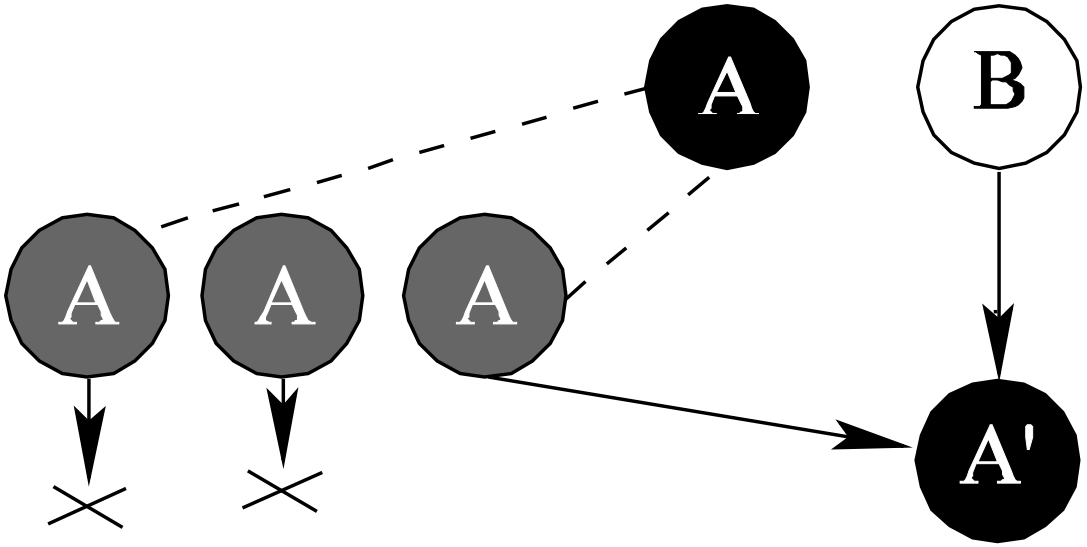
Prototype which executes locally last levels' tasks
developed and being tested

4. Workflow Optimization Heuristics

Replication

- **Heterogeneous Critical Tasks Reverse Duplicator:** replicates the parent-tree or some selected parents of chosen task.
- **Bottom-up Top-down Duplication Heuristic:** replicates all necessary task ancestors.
- **Heterogeneous Critical Parents with Fast Duplicator:** replicates considering idle time left by select task on given machine.
- **Critical Path based Full Duplication Algorithm:** replicates all possible parents of considered task.

5. Applying the Replication Heuristic



Variants

- Replicate all tasks
- Replicate tasks above *blocking threshold:*

$$b_{i,j} = \begin{cases} ST(N-i) & \text{if} \quad i = j \\ (i-1) + ST(N-i) & \text{if} \quad i \neq j \end{cases}$$

i,j = task coordinates
n = workflow levels

Blocking Subtree:

$$ST(n) = \frac{n(1+n)}{2}$$

- **Replicate tasks from Critical Path** (Critical Path based Full Duplication)
  - 3 copies/task.
  - Not all replicated tasks are sent necesarily to same cluster (backfilling mechanisms are avoided).
  - Persistent resource fails make it to be *banned*. Allows *functional* resource discovery at the beginning.

6. Experimental Results

Experiment Resources (April 2007)

- **EGEE:** Production and large scale
- **GRIDIMadrid:** Research and short scale

Site	Count.	Proc.	Speed	Nodes
GRIDIMadrid Resources				
UCM	ES	P4	3216	2
CIEMAT	ES	P4	2392	22
EGEE Resources (BIOMED Virtual Organization)				
BHAM-UNI	UK	PIII	800	128
BRUNEL	UK	P4	2000	5
CGG	FR	PIII	1266	56
CIEMAT	ES	PIII	1001	220
CYF-KR	PL	P4	2800	264
GRID-ACAD	BG	P4	2400	78
HELLASGRID	GR	P4	3400	356
IFCA	ES	P4	3200	96
II	MK	P4	3300	8
IMPERIAL	UK	P4	2000	188
IN2P3	FR	PIII	1001	569
INFN	IT	P4	2400	124
IPP-ACAD	BG	P4	2800	10
JET-EFDA	UK	PIII	1098	66
KELDYSH	RU	P4	3000	14
L.-HEP	UK	P4	3000	380
LIP	PT	P4	2200	52
MAN-UNI	UK	P4	2800	844
PNPI	RU	P4	3000	112
SAVBA	SK	P4	3200	41
SRCE	HR	P4	2193	16
UAM	ES	P4	2566	14
UCL	UK	P4	2800	312
UNI-LINZ	AT	P4	3014	8

GRIDIMadrid

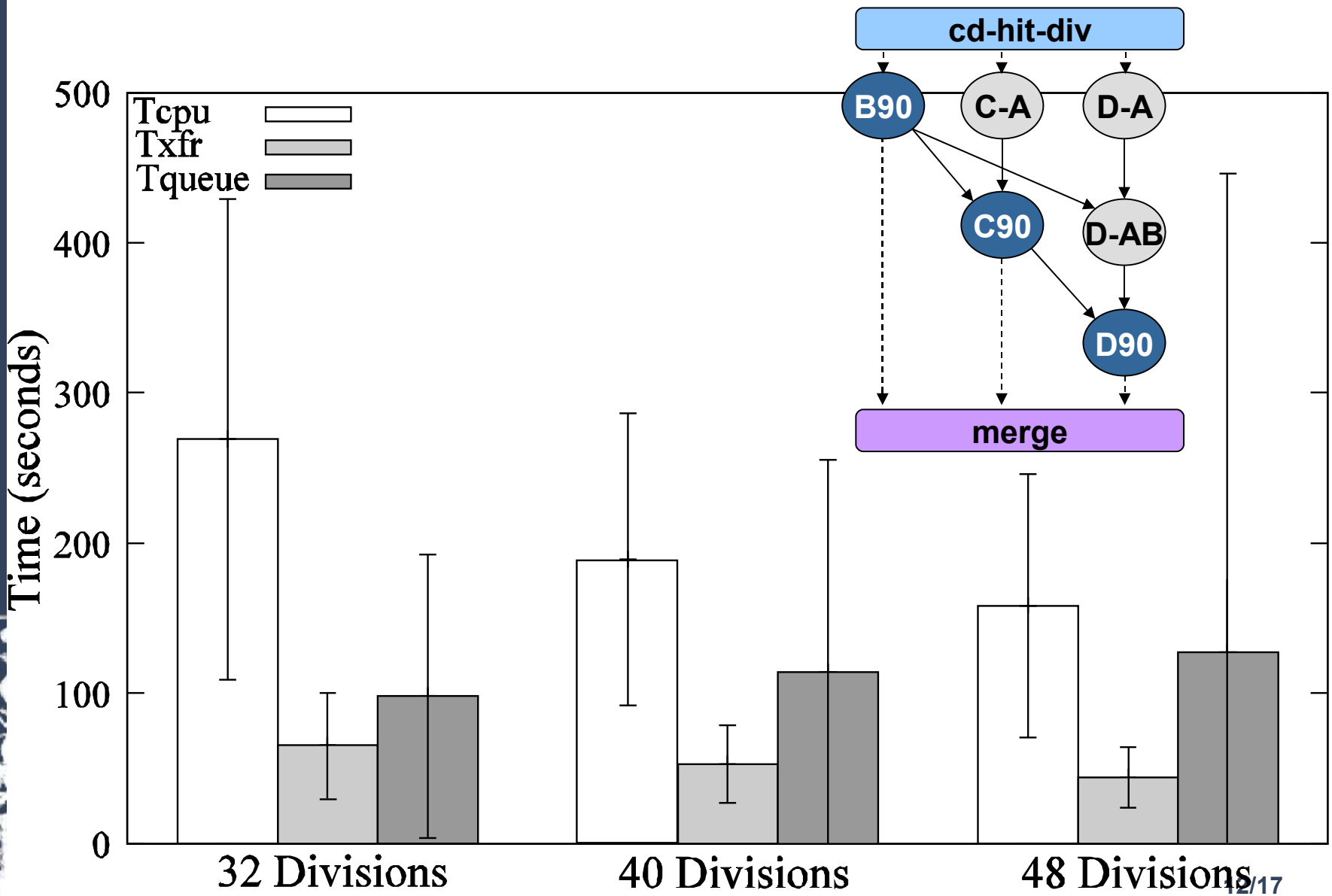
http://www.gridimadrid.org/

Enabling Grids for E-sciencE

http://www.eu-egee.org/

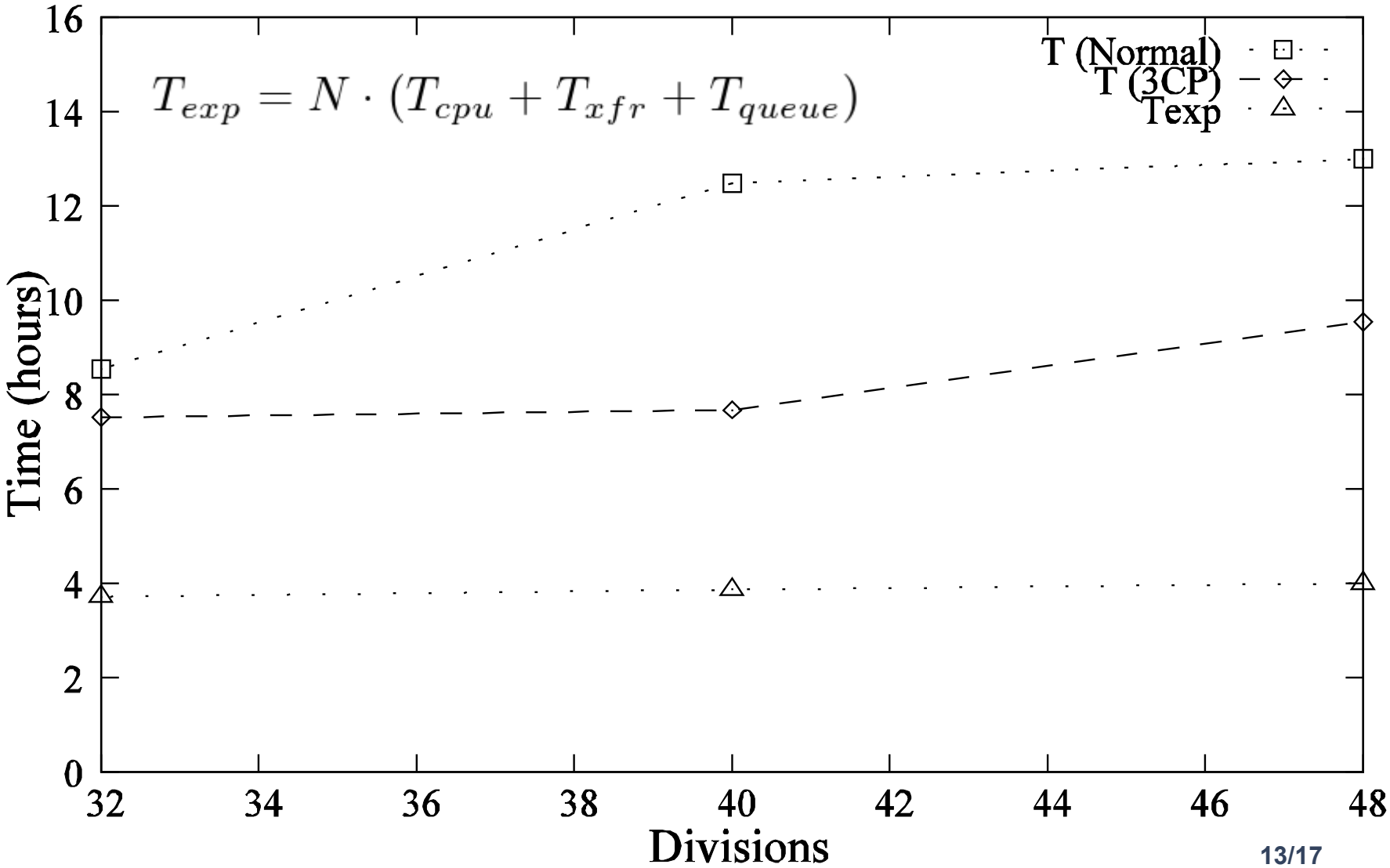
6. Experimental Results

Consolidated Times



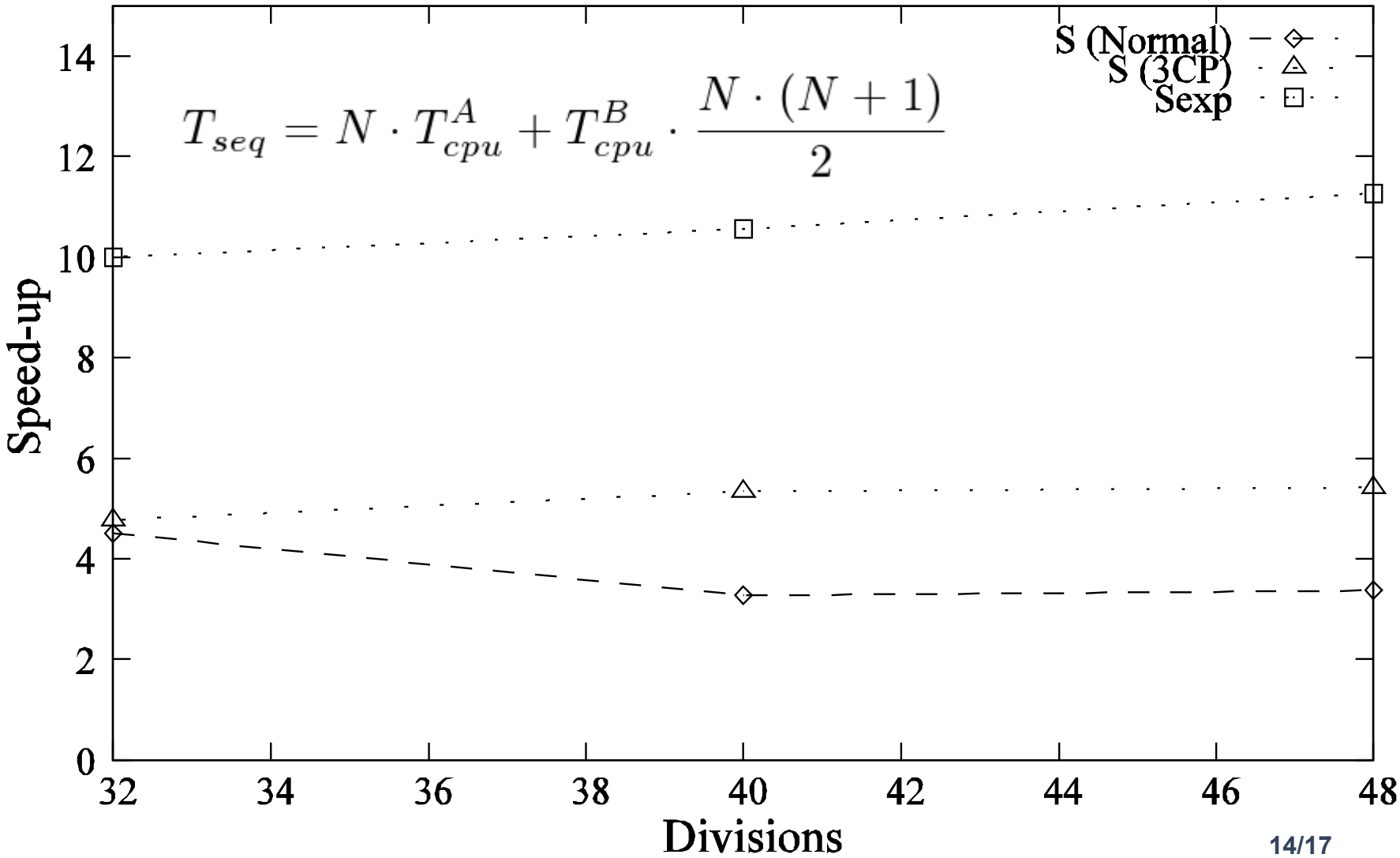
6. Experimental Results

Workflow Total Execution Times



6. Experimental Results

Workflow Speed-up



6. Experimental Results

Reschedules

Divisions	Reschedules		Mean Times	
	Norm.	3CP	Norm.	3CP
32	69	34	16.3’	14.7’
40	170	35	14.6’	11.7’
48	68	27	14.8’	12.1’

“The Dark Side” (Replication Costs)

- **32 Divisions (8h 06’ – Upperbound: 32h)**
  - Txfr: 45’’
  - Tcpu: 5’ 16’’
- **40 Divisions (7h 25’ – Upperbound: 28h)**
  - Txfr: 30’’
  - Tcpu: 3’ 51’’
- **48 Divisions (7h 55’ – Upperbound: 49h)**
  - Txfr: 25’’
  - Tcpu: 2’ 55’’

Additional computational effort done at **spare** resources

No harm to other Grid applications

7. Conclusions and Future Work

We have…

… reviewed 3 approaches to optimization techniques.
… focused in one of them for a specific workflow.
… demonstrated through experiments that *replication* gave high speed-up.
… shown Grid's nature affected the whole process.

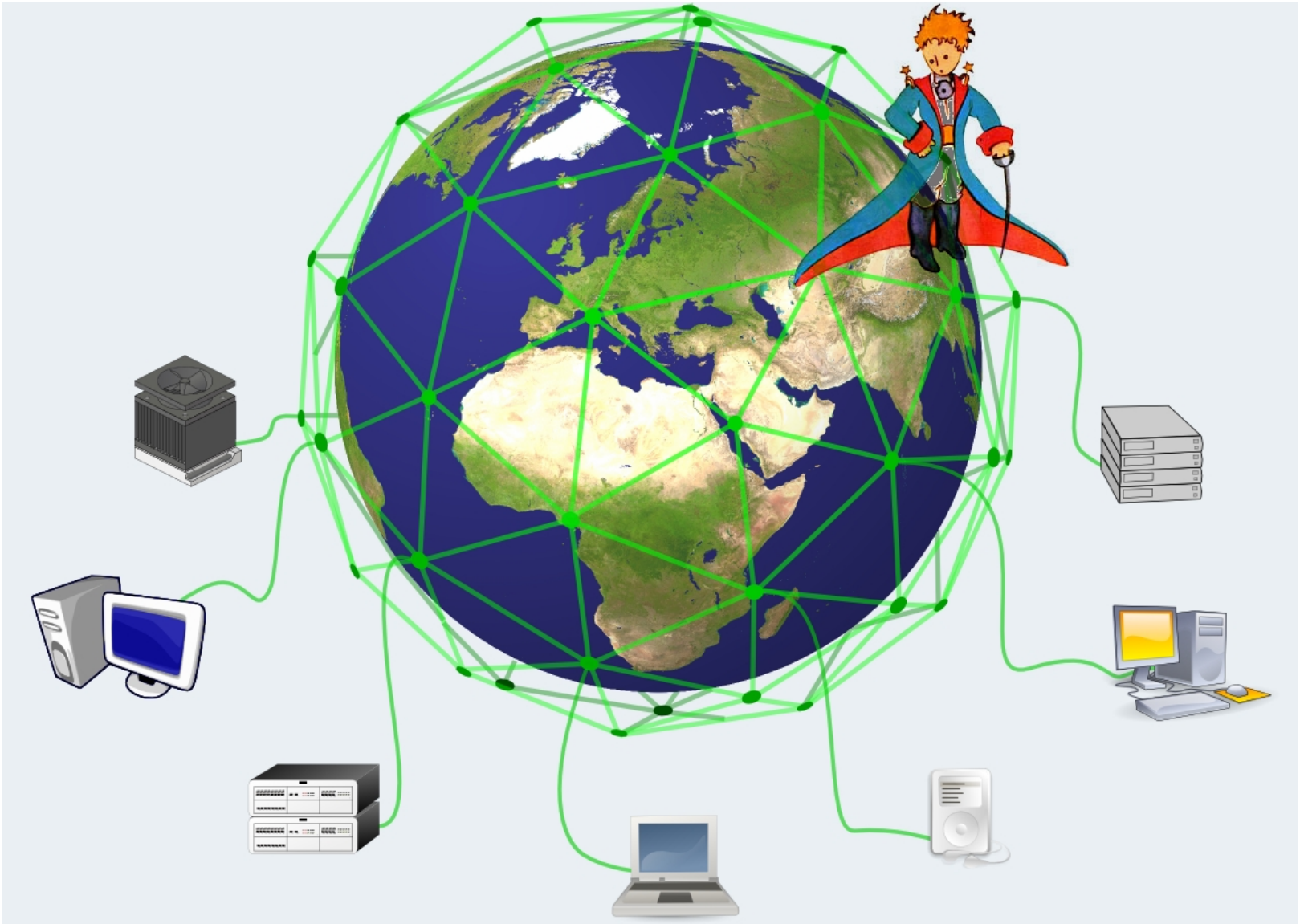
We will…

… study the effect of *agglomeration*.
… provide a model for workflow execution.
… study the effects of optimization heuristics on model.

Want to participate?

Visit http://www.gridway.org/ now!

Questions?



Merci pour votre attention!