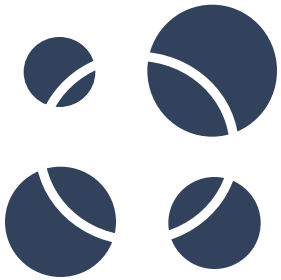
IberGrid 2007
Santiago de Compostela, Spain
May 15, 2007



Advanced Strategies for Efficient Workflow Management with GridWay

José Luis Vázquez-Poletti
Eduardo Huedo Cuesta
Rubén Santiago Montero
Ignacio Martín Llorente
http://asds.dacya.ucm.es

Distributed Systems Architecture Group
Universidad Complutense de Madrid

Contents

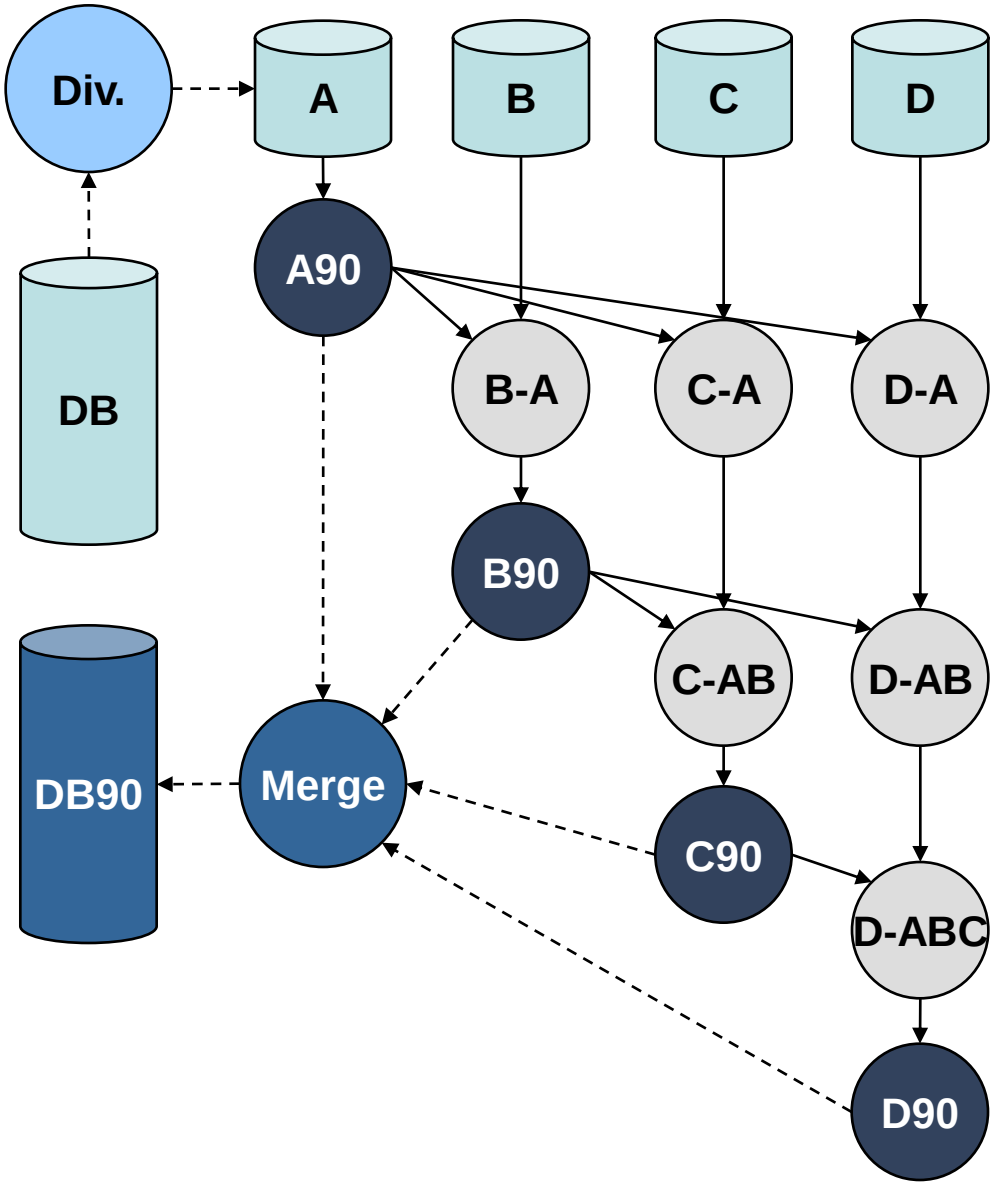
1. Protein Clustering Application (CD-HIT)

CD-HIT

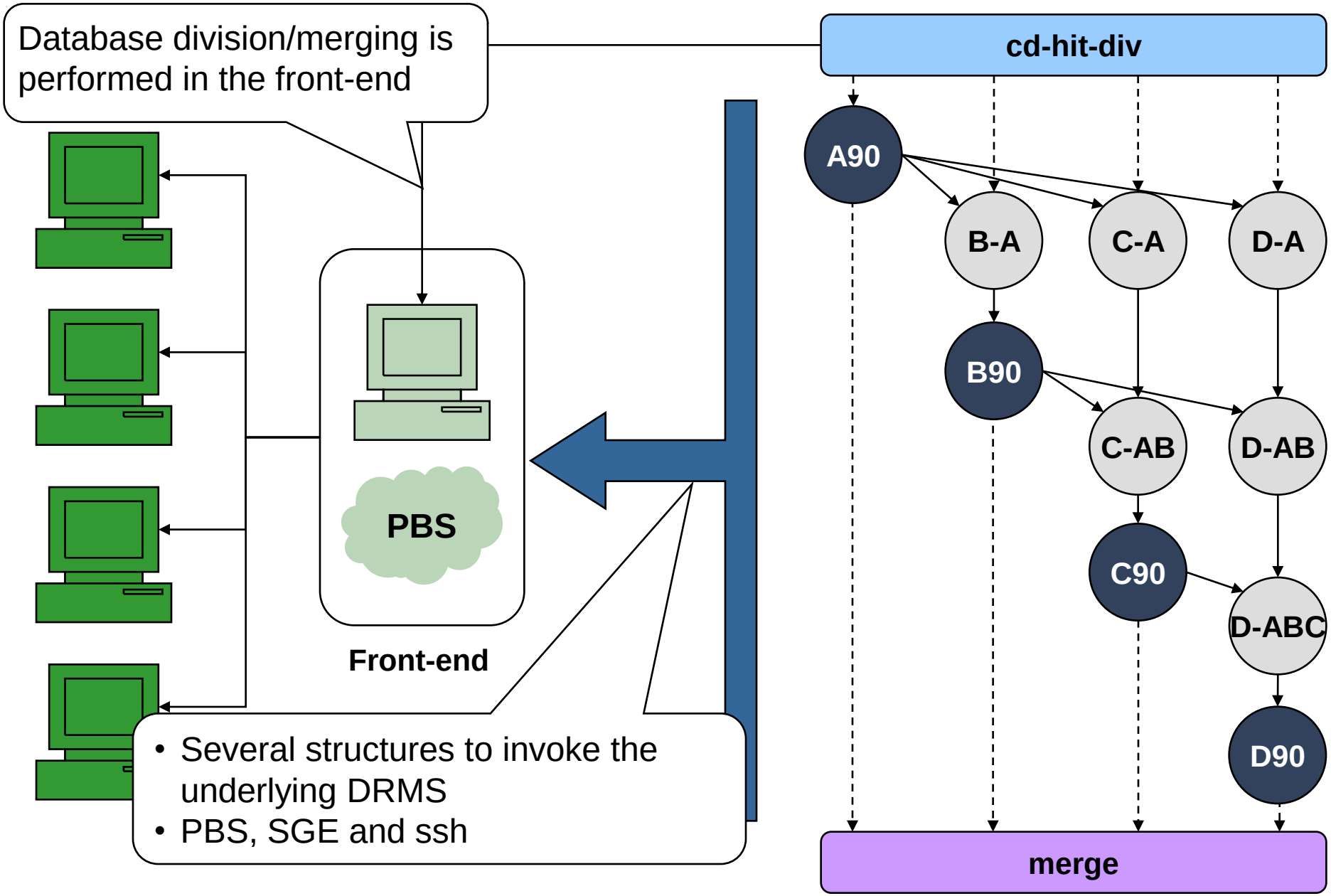
- *"Cluster Database at High Identity with Tolerance"*
- Protein (and also DNA) clustering
  - Compares protein DB entries
  - Eliminates redundancies
- Example: Used in UniProt for generating UniRef data sets
- Our case: Widely used in the Spanish National Oncology Research Center (CNIO)
  - Input DB: 504,876 proteins / 435MB
- Infeasible to be executed on single machine
  - **Memory requirements**
  - Total execution time

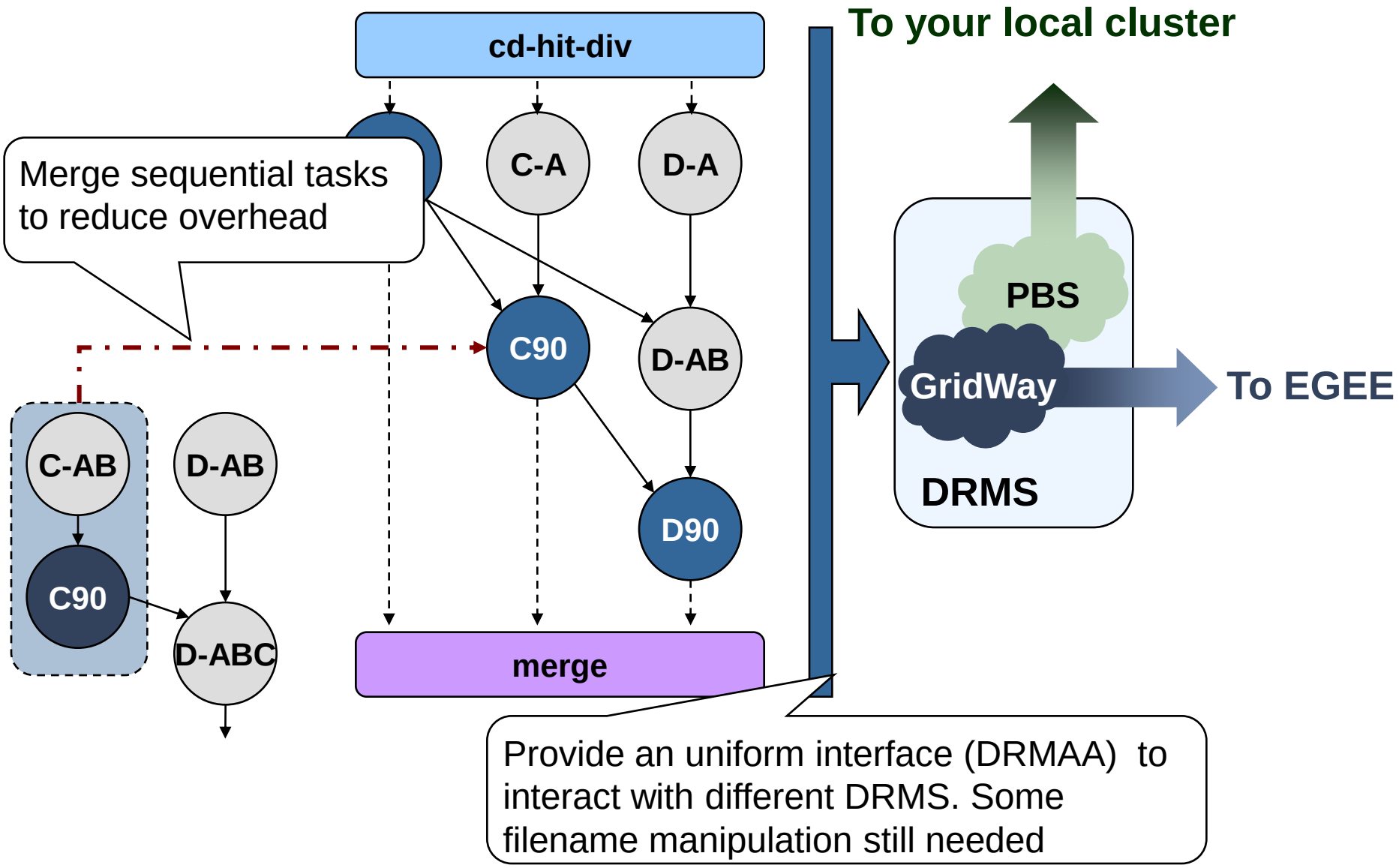
cd-hit-para

- 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. Parallel Execution of CD-HIT (cd-hit-para)



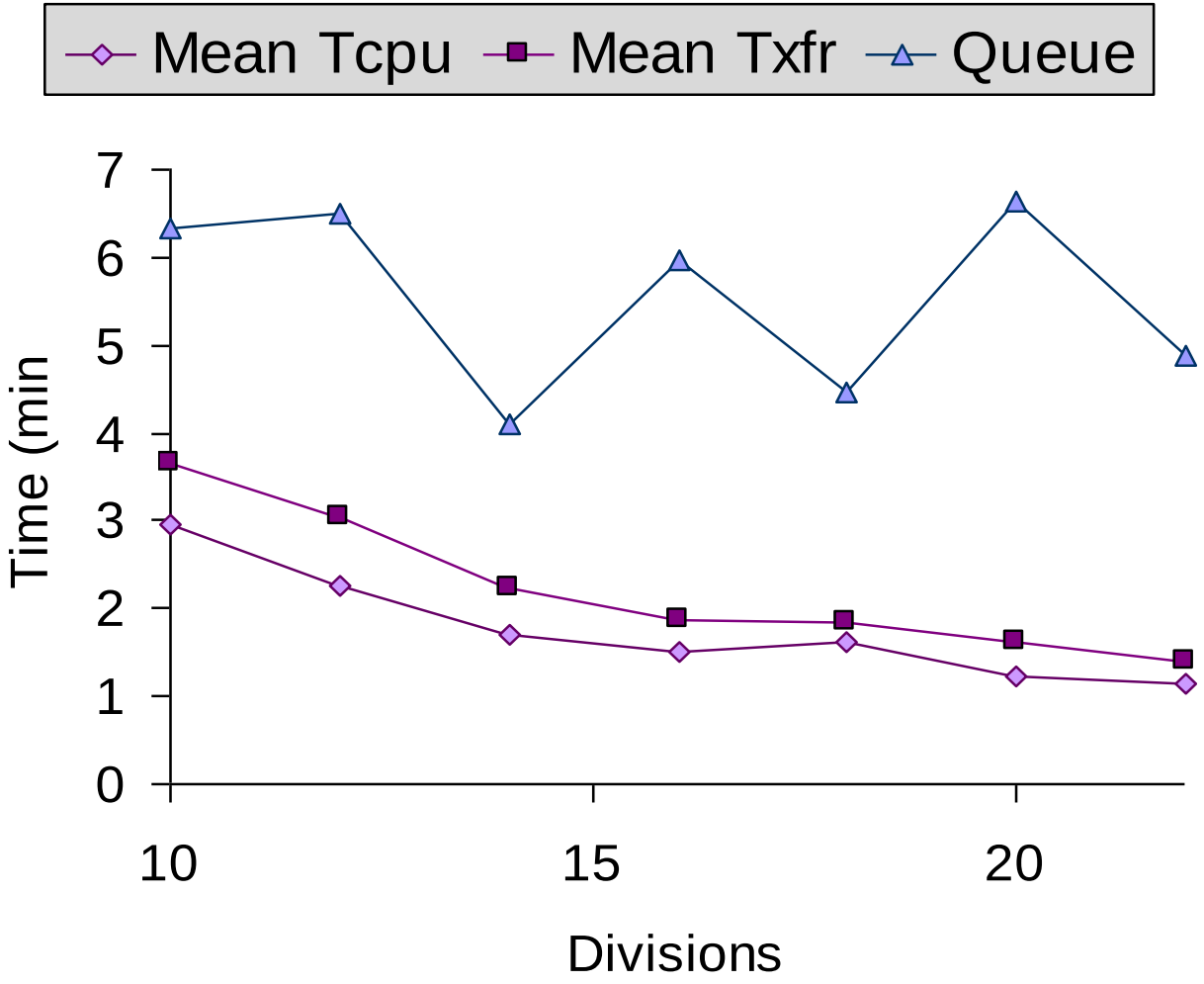

cd-hit-div
C-A
D-A
Merge sequential tasks to reduce overhead
C90
D-AB
C-AB
D-AB
C90
D90
D-ABC
merge
To your local cluster
PBS
GridWay
To EGEE
DRMS
Provide an uniform interface (DRMAA) to interact with different DRMS. Some filename manipulation still needed

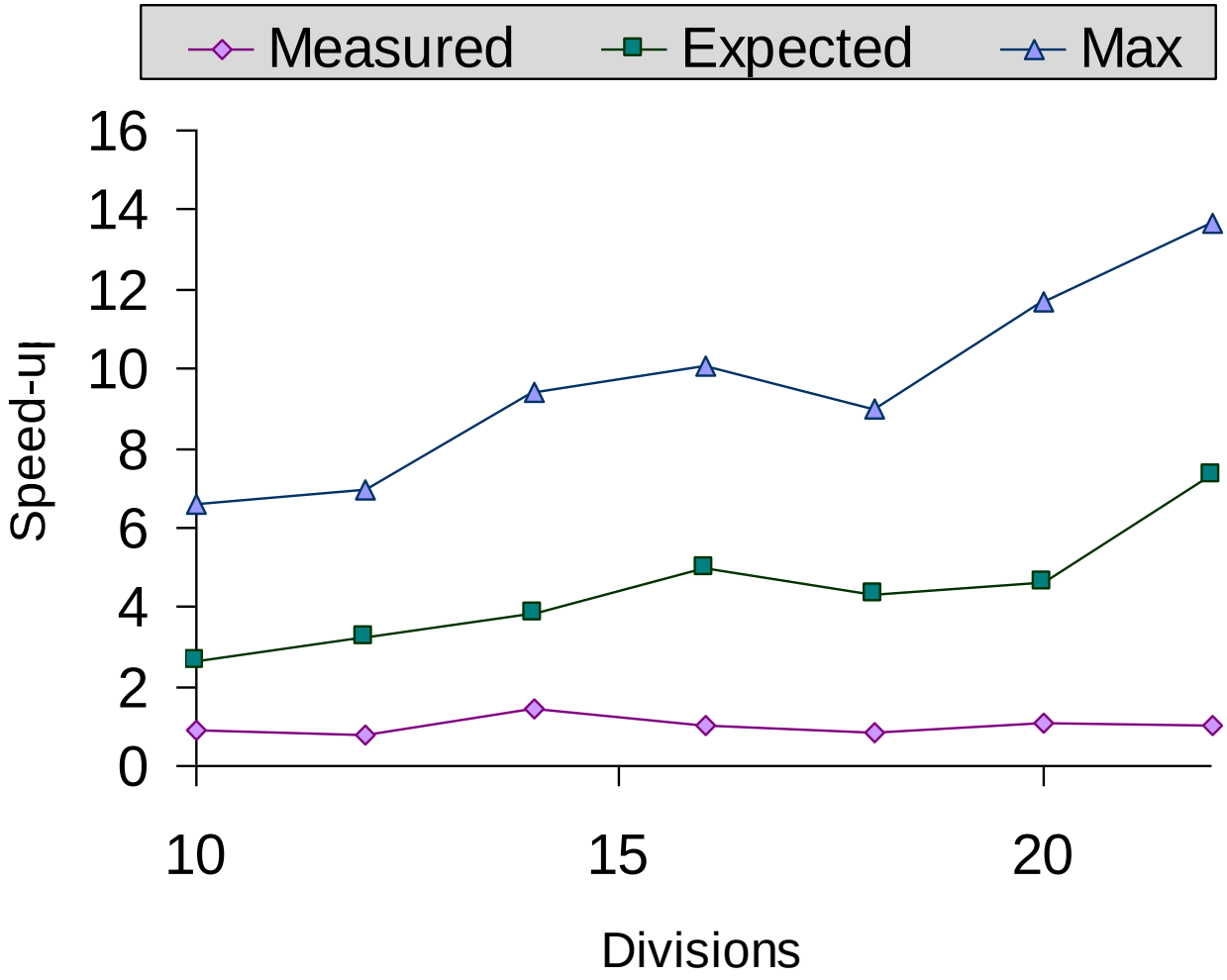
Grid Infrastructure

BIOMED sites

Site	Processor	Nodes	Speed
BIFI	Intel P4	56	3.2GHz
CESGA	Intel PIII	16	500MHz
CGG	Intel PIII	58	1.2GHz
CIEMAT	Intel Xeon	226	3.2GHz
GRIF	Intel P4	14	2.8GHz
JINR	Intel PD	30	2.8GHz
L.-HEP	Intel P4	374	3GHz
PNPI	Intel P4	60	3GHz
RAL	Intel P4	62	2.8GHz
RALPP	Intel PIII	1064	1GHz
ScotGRID	Intel Xeon	6	2.8GHz
SINP	Intel Xeon	94	2.8GHz

DB div.	Mean size	Tasks
10	44MB	45
12	36.5MB	66
14	31.5MB	91
16	27.5MB	120
18	24.5MB	153
20	22MB	190
22	20MB	231

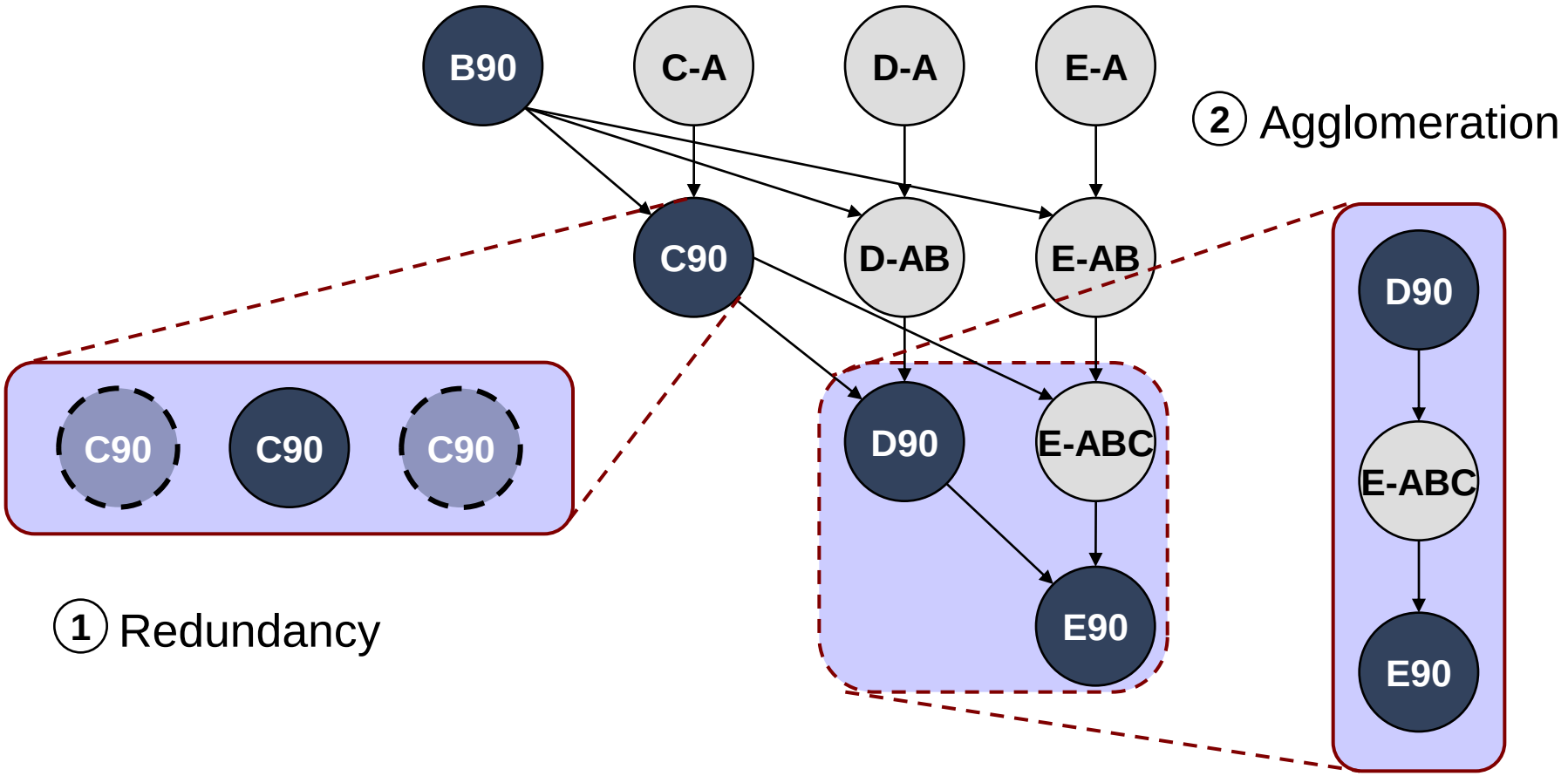




Expected: without considering either queue wait times or job failures
Max: taking into account just the *critical path*

cd-hit-grid

- Larger database 4,186,284 proteins (1.7 GB) from CNIO
- Improve the efficiency of the workflow execution with redundancy and agglomeration



Information and download at http://www.GridWay.org

Thank you
for your attention!