Replication Heuristics for Efficient Workflow Execution on Grids

J.L. Vazquez-Poletti, E. Huedo, R.S. Montero, I.M. Llorente
DSA Group, Universidad Complutense de Madrid, Spain

On The Move
OTM Federated Conferences and Workshops
GADA

DSA Group

Protein Clustering

Process where protein databases are organized into groups or families in order to capture protein similarity. It can be applied in domain analysis, large protein database organization and search improvement.

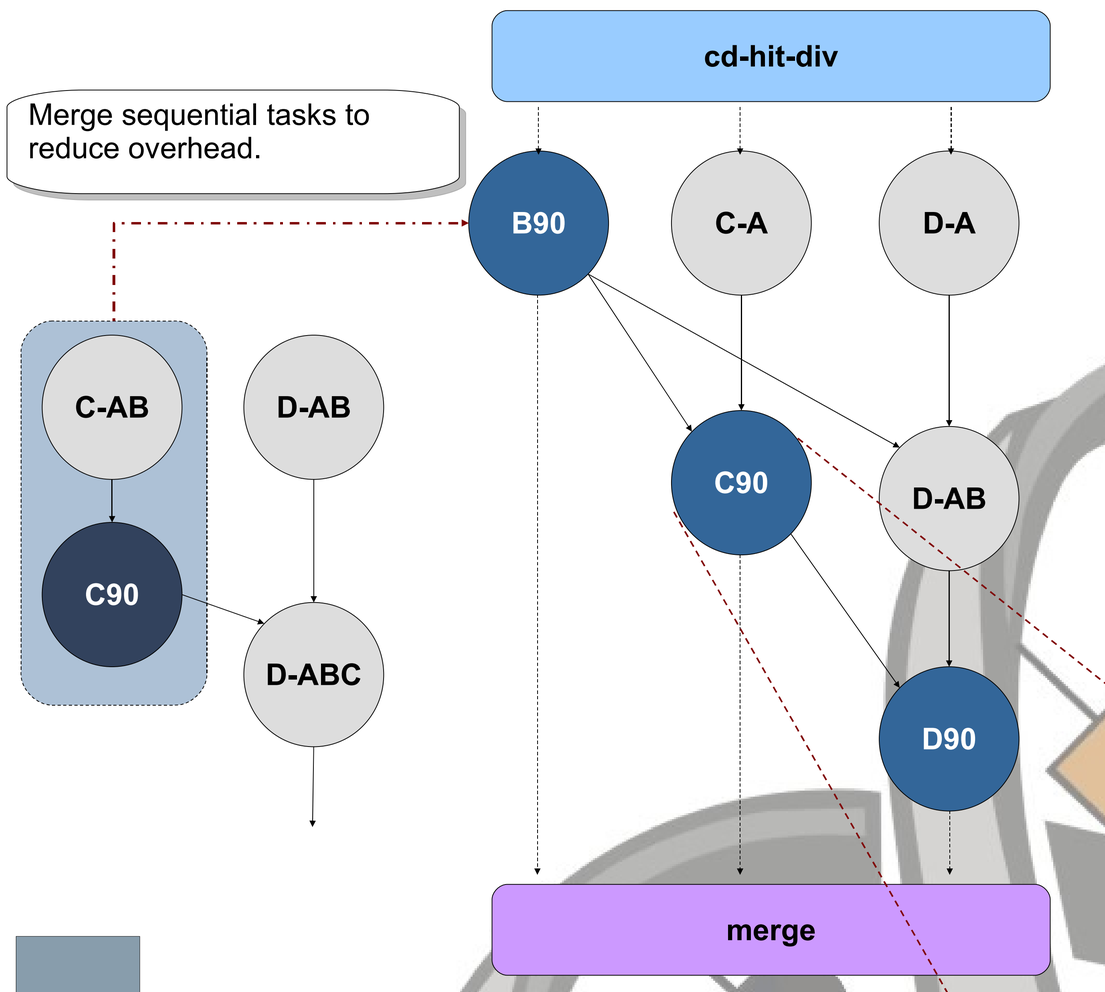
CD-HIT: Cluster Database at High Identity with Tolerance

Toolkit for clustering large protein databases at high sequence identity threshold. Redundant sequences are removed and a database of only representatives is generated.

Input Database

4,186,284 proteins (1.7GB) provided by the National Center for Biotechnology Information

Merge sequential tasks to reduce overhead.



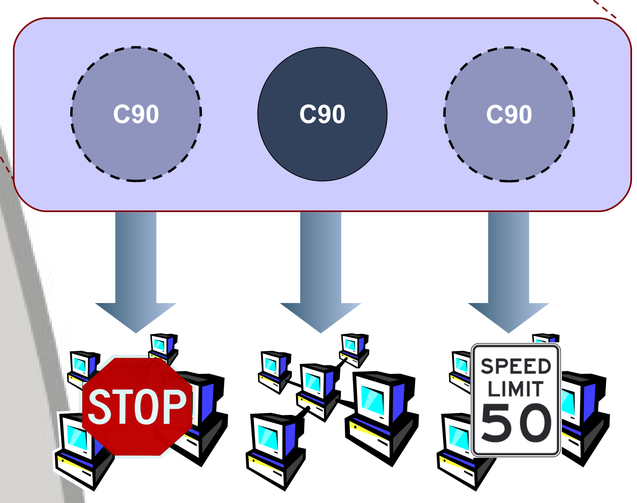
Provide an uniform interface (DRMAA) to interact with different DRMS. Some filename manipulation still needed.

OpenGridForum
OPEN FORUM | OPEN STANDARDS

Optimization

Replication

Supplementary tasks are created for the workflow's *critical path* nodes. When one of these tasks ends, the node is taken as executed and the rest of replicated tasks are killed.



Benefit of Replication

The more replicated tasks are created, the higher the possibility for that node to be executed shortly by reducing the effect of job failures and queue wait times.

GridWay (http://www.gridway.org/)

- Metascheduler part of Globus Toolkit.
- Stands on top of Globus services.
- Handles DAG based workflows.
- Allows advanced flow structures (loops, branches).
- Implements the Distributed Resource Management Application API (DRMAA) which is an OGF Standard.
- Considers static and dynamic resource information.
- Offers automatic staging mechanisms.
- Provides fault tolerance mechanisms (network outage, remote and local machine crash):
  - Tries task execution/file transfer on the same resource.
  - Failed tasks are moved transparently to other resources.

Results