CS-ROSETTA

Problem Statement

Bob is a UNL chemist who works to derive the structure and properties of many proteins. To do so, he gathers the chemical shift information from thousands of proteins and wants to use the CS-ROSETTA package to predict the structure.

CS-ROSETTA has three parts:

1. **Fragment selection**: Based on the input chemical shifts, select candidate protein fragments from a 80GB database. This takes about 4 hours in one process.
2. **Fragment assembly**: Given the input protein fragments (about 20MB), run ROSETTA to assemble them into candidate molecular models. This is a random process, and we repeat this 2,000 times to generate many candidates. We end up with 2,000 molecular models of 100KB each. ROSETTA assembly takes 1-10 hours.
3. **Structure evaluation**: All the molecular models are gathered into one place. They are scored according to how well they match the input chemical shifts. An ordered list of the top candidates is returned to the user. The

Classification

How does this problem fit into our classification scheme? Is this workflow appropriate for the OSG?

Job Needs

What are the input/output/runtime needs of each step?

Solution method and Diagram

Describe how to solve this problem using a **prestaging scheme**. You do not need to specify any technologies. Draw a diagram showing the location of the input, output, and runtime files. Show the hardware involved, its location, and estimates for the query rate.

BLAST Queries

Problem Statement

Xin runs a bioinformatics lab and studies how virus DNA can become embedded in the DNA of hosts. He takes many sub-sequences from the virus DNA, and would like to see if they are embedded anywhere in the DNA of various animals. For this particular study, he's taking 10,000 sequences and running a BLAST query against a large database containing several species' genomes. The database is 100MB, and each query takes 1 minute.

Classification

How does this problem fit into our classification scheme? Is this workflow appropriate for the OSG?

Job Needs

What are the input/output/runtime needs of each step?

Solution Method and Diagram

Describe how to solve this problem using a **caching scheme**. You do not need to specify any technologies. Draw a diagram showing the location of the input, output, and runtime files. Show the hardware involved, its location, and estimates for the data volume, transfer rate and access rate.

HEP Analysis

Problem Statement

Ken is a physicist on the CMS project (a detector on the LHC particle accelerator), and a specialist in the silicon sub-detector. He wants to study the performance of the sub-detector. To do this, he has to access a dataset totaling 5TB, broken into 2,500 files of 2GB. Each job accesses 2 files and produces 100MB of output in a smaller, more specific data format for Ken's analysis. The output needs to be copied back to Nebraska so he can do his work from his desktop.

Classification

How does this problem fit into our classification scheme? Is this workflow appropriate for the OSG?

Job Needs

What are the input/output/runtime needs of each step?

Solution Method and Diagram

Describe how to solve this problem using a **caching scheme**. You do not need to specify any technologies. Draw a diagram showing the location of the input, output, and runtime files. Show the hardware involved, its location, and estimates for the data volume, transfer rate and access rate.

Weather Forecast Modeling

Problem Statement

Cindy studies the effects of climate change in the state of Nebraska. She uses the global climate models generated by a large supercomputer at Oak Ridge National Lab (50 models each in a 1TB file) for an input to higher-resolution regional models. The regional models are run on a local super-computer, Firefly; the output is again 1TB files, one per input module. The 1TB files each contain a simulation of the weather for the next 50 years. Cindy then takes the weather simulations (which shows the number of days a rain for each county) to build a map showing the probability of a severe flood in the next 50 years.

Classification

How does this problem fit into our classification scheme? Is this workflow appropriate for the OSG?

Job Needs

What are the input/output/runtime needs of each step?

Solution Method and Diagram

Describe how to solve this problem using a data management scheme of your choice. Draw a diagram showing the location of the input, output, and runtime files. Show the hardware involved, its location, and estimates for the data volume, transfer rate and access rate.