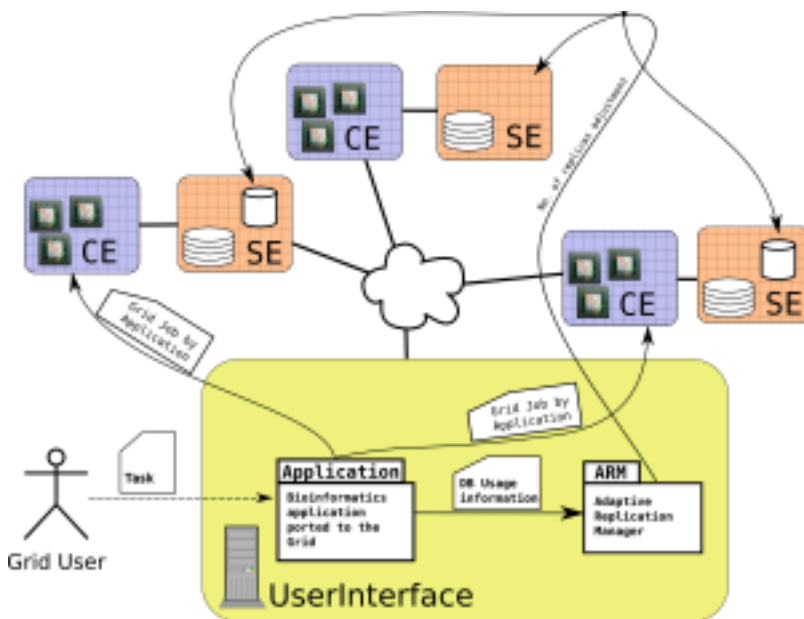


# Database and Functional Genomics Applications

In the scope of BioinfoGRID we are porting to the Grid a large number of bioinformatics applications. Most bioinformatics applications, however, need access to databases of biological data, consisting of large files up to the gigabyte range, which are created and periodically updated by leading institutions such as NCBI, EBI and ExpASY.

While the access from applications to these databases is mostly read-only, these databases still present a challenge for Grid management due to the following factors: 1) the databases need to be present on the worker nodes for the bioinformatics applications to work efficiently, but 2) transferring a database from a remote location before a computation could result time and bandwidth consuming 3) replicating every database on every storage element could not be feasible due to the high storage costs involved



Bioinformatics application use case, using the adaptive database management framework.

In the scope of this activity we created an adaptive database management framework to optimally manage such biological databases on the Grid platform. The framework dynamically adjusts the number of replicas of each database depending on the usage of each particular database in recent times: recently used databases get a high number of replicas so as to increase availability (jobs are constrained to execute near a replica), which in turn reduces job queuing times, while least used databases will get few or only one replica, so as to greatly reduce Grid storage costs for the average case. The number of replicas is constantly updated.

Many applications have since then been ported to the Grid leveraging this framework, and users are benefiting from the dramatically increased throughput which is achievable by the Grid especially on large executions.

Another achievement within this activity is the accomplishment of a 55 CPU-years challenge for the Gene Analogous Finder application. The Gene Analogous Finder goal was to find functionally analogous gene products by comparing gene products according to their description and not to their sequence similarity. For this challenge, one million well-annotated genes were selected from the GODB (Gene Ontology DataBase) and every one of these genes was compared against all the others. Such a comparison would have taken 55 years on a single CPU, but was performed in less than one month on the Grid hence achieving an outstanding 660x average speedup.