

## **Grid Collector: Enabling New Science by Accelerating Access from Data Grids**

The Grid Collector is a system that facilitates the effective analysis and spontaneous exploration of high-energy physics scientific data. It combines an efficient indexing technology with a Grid file management technology to speed up common analysis jobs on high-energy physics data and to enable some previously impractical analysis jobs. To analyze a set of high-energy collision events, one typically specifies the files containing the events of interest, reads all the events in the files, and filters out unwanted ones. Since most analysis jobs filter out significant number of events, a considerable amount of time is wasted by reading the unwanted events. The Grid Collector removes this inefficiency by allowing users to specify more precisely what events are of interest and to read only the selected events. This speeds up most analysis jobs. In existing analysis frameworks, the responsibility of bringing files from tertiary storage to disk falls on the users. This forces most of analysis jobs to be performed at centralized computer facilities where commonly used files are kept on disks. The Grid Collector automates file management tasks and makes it easy to perform analyses on data files that are not already on disk. This enables some analysis jobs that were previously too time-consuming, and makes it possible for users to run their analysis jobs on their own workstations instead of running on the centralized facilities.

Rare events in high-energy collisions could reveal important new physics insights. To illustrate the use of the Grid Collector in these cases, we give one example of searching for the evidence of jet quenching. An initial search has been performed on a large number of analysis objects, and a small number of events (about 80) were found to have unusual jet distributions, which could indicate jet quenching. To further investigate these 80 events, they have to be extracted from the files containing them. However, because the events are scattered in many large files and most of the files are on mass storage systems, the analysts are not willing to spend the disk resource or manpower to transfer the files themselves. For this reason, the follow-up analyses were delayed for three years. However, once the analysts learned of the Grid Collector, they were able to extract these 80 events within 15 minutes.