Contribution ID: **319** Type: **oral presentation**

Grid Collector: Using an Event Catalog to Speed up User Analysis in Distributed Environment

Thursday, 30 September 2004 17:10 (20 minutes)

Nuclear and High Energy Physics experiments such as STAR at BNL are generating millions of files with PetaBytes of data each year. In most cases, analysis programs have to read all events in a file in order to find the interesting ones.
Since most analyses are only interested in some subsets of events in a number of files, a significant portion of the computer time is wasted on reading the unwanted events. To address this issue, we developed a software system called the Grid Collector. The core of the Grid Collector is an "Event Catalog".
This catalog can be efficiently searched with compressed bitmap indices. Tests show that it can index and search STAR event data much faster than database systems.
It is fully integrated with an existing analysis framework so that a minimal effort is required to use the Grid Collector in an analysis program. In addition, by taking advantage of existing file catalogs, Storage Resource Managers (SRMs) and GridFTP, the Grid Collector automatically downloads the needed files anywhere on the Grid without user intervention.

The Grid Collector can significantly improve user productivity. The improvement in productivity is more significant as users converge toward searching for rare events, because only the rare events are read into memory and the necessary files are automatically located and downloaded through the best available route. For a user that typically performs computation on 50% of the events, using the Grid Collector could reduce the turn around time by a half.

Primary authors: SHOSHANI, A. (Lawrence Berkeley National Lab); WU, K. (LAWRENCE BERKELEY NATIONAL LAB); PEREVOZTCHIKOV, V. (Brookhaven National Lab); ZHANG, W-M. (Kent State University)

Presenter: WU, K. (LAWRENCE BERKELEY NATIONAL LAB)

Session Classification: Distributed Computing Systems and Experiences

Track Classification: Track 5 - Distributed Computing Systems and Experiences