spawned by far-traveling neutrinos.

But Haeshim Lee, an astrophysicist at Chungnam University in Taejon and head of the telescope's theoretical division, doubted that HANUL would work, and some experimentalists worried that the magnets would be too costly to build. Haeshim Lee aired his doubts in an angry letter circulated among scientists and government officials in June 1998 and later quit the project. (A toned-down version of his critique was published in the Korean Physical Society's monthly journal in September 1999, prompting coverage of the affair last month in an online newsletter, *Korean American Science and Technology News*, published by Moo Young Han, a physicist at Duke University.)

Worried about the status of the project, KOSEF called an emergency meeting in April 1999. Lee and other division heads urged agency officials to replace Song and to give the project greater flexibility. KOSEF declined to act on Song's status, with one official explaining that "we manage the research budget, not the team itself." But in June it cut off funding for HANUL, noting that the scientists had not chosen where to assemble the prototype and could not meet an August deadline for its completion.

"I don't know what to say. I'm just so disappointed," says Jewan Kim, a physicist at Seoul National University who had helped build support for HANUL. "We had many meetings, but people just don't agree. There's nothing you can do about it." Song blames the project's failure on disagreements over physics and cost, on stifling bureaucratic requirements, and on a "lack of warm personalities."

Others regret the loss of a chance to explore neutrino energies in a range between those covered by two other major experiments, the massive Super-Kamiokande underground water detector in Japan and the larger but less acute AMANDA project in Antarctica, which monitors a huge volume of ice. "The HANUL project was trying to make a bridge between these two techniques. It certainly was worthwhile," says Francis Halzen, a physicist at the University of Wisconsin, Madison, and a co-PI for AMANDA.

Five months after losing funding, Lee still hopes to resurrect HANUL elsewhere and somehow include Korea in it. He says that the experience leaves him eager to find international collaborations that can improve Korea's academic environment: "Korea needs more pure research projects so that young people can learn to think for themselves. I thought that, in a small way, I could accomplish that. But I guess the project came a little too early."

–MICHAEL BAKER

Michael Baker is a writer in Seoul.

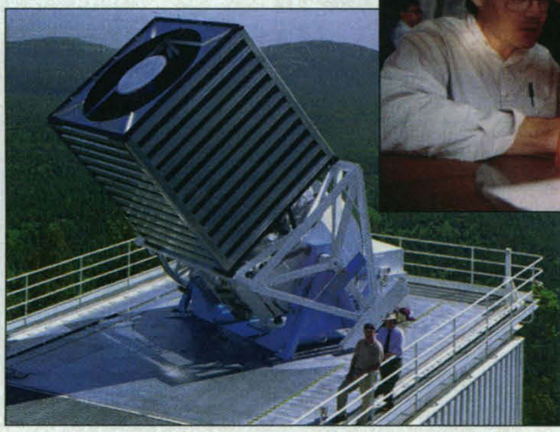
COMPUTER SCIENCE

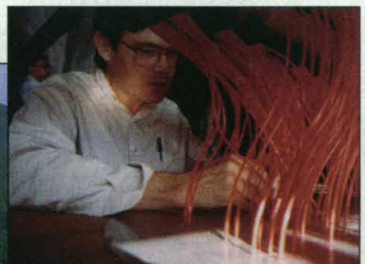
Physicists and Astronomers Prepare for a Data Flood

New accelerators and sky surveys that will spew data by the terabyte are spurring a search for new ways to store and disseminate the flow

The end of a millennium is a time for warnings, and some scientists are joining in: They are predicting a flood. But unlike most millennial doomsayers, the scientists are looking forward to being inundated. Their flood is a torrent of data from new physics and astronomy experiments, and they hope it will sweep some long-awaited treasures within reach, such as the Higgs boson, a hypothetical particle that endows everything else with mass, and a glimpse of life-supporting planets in other solar systems. The greater the torrent of data, the better the chance that scientists will pull these and other prizes from it—providing they can find ways to store and channel the flow.

The quantities of data expected in the next decade will be staggering. Planned experiments at the Large Hadron Collider (LHC), a giant particle accelerator due to be up and running in 2005 at CERN, the European particle physics center near Geneva, "will write data to a disk-based database at a rate of 100 megabytes per second," says Julian Bunn of the California Institute of Technology's (Caltech's) Center for Advanced Computing Research, "and we expect these experiments to run for 10 to 15 years." That is over 100 petabytes of data, roughly the equivalent of 10 million personal computer hard disks. (A petabyte is 10^{15} bytes.) RHIC, an accelerator at Brookhaven National Laboratory in Upton, New York, that collides heavy nuclei to create a primordial state of matter called quark-gluon plasma, is already spewing out data at a rate of nearly a petabyte a year—about 1000 times the volume of data in the largest biological databases.





Data maw. The Sloan Digital Sky Survey's 2.5-meter telescope at the Apache Point Observatory in New Mexico. In the inset, astrophysicist Rich Kron adjusts optical fibers that feed light to instruments analyzing light from many objects at once.

CREDITS (LEFT TO RIGHT): FERMILAB VISUAL MEDIA SERVICES/SDSS; THOMAS NASH/SDSS; A. SZALAY

Astronomy is contributing to the torrent as well. Johns Hopkins University astrophysicist Alex Szalay expects the Sloan Digital Sky Survey (SDSS), which aims to image 200 million galaxies and measure distances to a million of them, to produce about 40 terabytes of information. (A terabyte is 10^{12} bytes.) Several planned sky surveys at other wavelengths, such as radio and infrared, will contribute tens of terabytes more.

Organizing the data and making them available to the global community of scientists without swamping computers or networks will require rethinking the ways data are stored and disseminated. Researchers at institutions including Johns Hopkins, the Fermi National Accelerator Laboratory (Fermilab), Caltech, and Microsoft Corp. are now doing just that. By sorting the data as they flood in and dynamically reorganizing the database to reflect demand, they hope to provide prompt, universal access to the full data archives. "The volume and complexity of the data are unprecedented," says Caltech particle physicist Harvey Newman. "We need a worldwide effort to get the computing capacity."

Two trends have converged to create the database challenge. New particle detectors and telescopes are starting to rake in data at an unprecedented rate. And the experiments themselves have ever larger numbers of far-flung, data-hungry collaborators. The full data sets will have to be stored in central repositories because of their volume: It could take months to transmit a full copy of a petabyte-sized data set over the fastest affordable Internet connection. But to make a mammoth central reservoir usable, says Szalay, "we need a double paradigm shift," encompassing both data storage and dissemination.

He and colleagues at Johns Hopkins, Caltech, and Fermilab are tackling the first step—organizing databases to make them easier to search. Current scientific databases often store data sequentially, as it is churned out by the experiment, which makes retrieving a specific subset of data (all blue galaxies in the SDSS, for example) very slow. Says Szalay, "We have to divide and conquer the data."

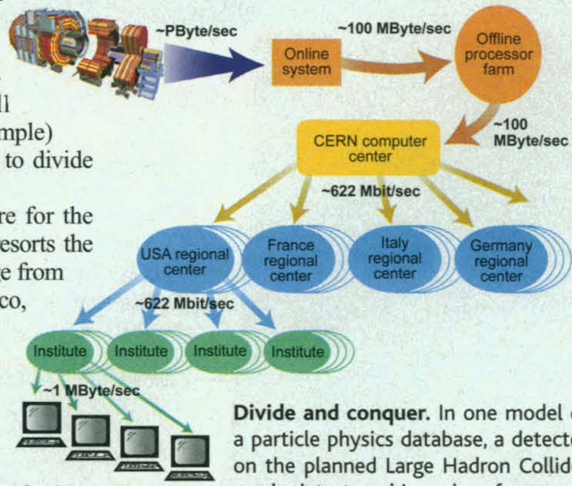
His team has designed software for the Sloan Survey that automatically presorts the stars and galaxies in each new image from the survey's telescope in New Mexico, putting them into separate buckets called objects. Electronic "labels"—"blue galaxies" or "11th magnitude stars"—indicate the contents of the buckets, and the database software will link each bucket to the other buckets to form what's called an object-oriented database. The SDSS data spigot is already open, and Szalay's software is hard at work distributing data into the appropriate buckets.

Presorting the data in this way dramatically decreases the computer time needed to find relevant information. "It's like the difference between picking a song from a cassette tape and one from a compact disc," says physicist Bruce Allen of the University of Wisconsin, Madison. To find a song on a tape, you have to fast-forward through all the other songs, but a CD player can skip over "Twinkle, Twinkle, Little Star" and go directly to "Blue Moon." Similarly, when the whole SDSS object-oriented database is complete, the computer will be able to respond to a request for blue galaxies by going straight to the appropriate bucket.

Data from particle physics experiments will be sorted and stored in much the same way. Collision events could be sorted by "the curvature of a particle track in a magnetic field or the energy collected from an electron shower," suggests Bunn. Ongoing projects such as the Particle Physics Data Grid and the Globally Interconnected Object Databases—two Caltech-based projects—are already testing object-oriented database technologies for particle physics.

The scientists are hoping that they can link and search the database objects with inexpensive, off-the-shelf software, such as Oracle or Objectivity. Whereas a high-end, custom-built database costs about a dollar per megabyte of stored information, says Microsoft database expert Jim Gray, a commercial database costs only a penny per megabyte. Preliminary work he and his colleagues did with 100-gigabyte Oracle databases stored on an off-the-shelf PC network looks promising, he says: "We think this is a design for the future."

Szalay's second paradigm shift would affect not the database itself but the computer network, transforming it from the traditional client-server computer network architecture to a hierarchical computer grid. In the client-server model, the database is stored at a single location. The client requests information directly from the server computer, and the server sends it back. But even if the data are presorted, as in an object-oriented database, millions of data requests from thousands of scientists in every corner of the world could quickly bring even the most powerful supercomputer to a screeching halt. "A single computer would be swamped," says Bunn. "We are obliged to do something radically different."



Divide and conquer. In one model of a particle physics database, a detector on the planned Large Hadron Collider sends data to a hierarchy of computers. The lowest tiers place subsets of the data near the users who request them most frequently.

CREDIT: MONARC PROJECT

What's more, even with a projected 1000-fold increase in network bandwidth in the next few years, network access "will remain a scarce resource," says Newman. So a collaboration of particle physicists and astronomers funded by the National Science Foundation's Knowledge Discovery Initiative and headed by Szalay is drawing up plans to distribute both the SDSS and LHC databases over multiple computers, arranged in a hierarchy. At the highest level, one complete copy of the presorted object-oriented database will be split into pieces and distributed among a handful of "Tier-0" centers. To protect against any errors that may creep into the presorted data, a copy of the raw data will also be stored at the Tier-0 centers.

Below the Tier-0 centers will be a series of three increasingly specialized layers of data-storing centers. Every Tier-0 center will be electronically linked to several Tier-1 regional centers, serving a particular region or country. The Tier-1 regional centers in turn will be connected to local universities (Tier-2) and finally to individual researchers (Tier-3). Each tier will house a copy of a progressively smaller piece, or cache, of the total database.

The exact contents of a given cache will change over time. Initially, "we will assess how people might use the system," says Newman, and then load each cache with the information most likely to be used by researchers connected to that branch of the hierarchy. For example, universities with large cosmology groups might choose to store a list of the sky coordinates of all the galaxies observed by the SDSS but ignore all the stars in the data set. Then, when a researcher queries the database for the locations of galaxies, his computer only has to go as far as the next tier for the information. On the other hand, another astronomer at the same university who wants the colors of nearby stars might have to search all the way up to a Tier-0 center. But because "caching will be triggered by access patterns," as Newman puts it, data will constantly be redistributed among the centers to make the system as efficient as possible.

So will it work? Yes, says computer scientist Krystof Sliwa of Tufts University in Medford, Massachusetts. He and his collaborators have constructed a detailed computer program to simulate the behavior of the proposed grids. "It's a classic optimization problem," says Sliwa. "We look at the cost and time required to do a set of jobs." Sliwa's models indicate that the existing Internet could handle the data requests to a layered computer network housing a petabyte-scale database. Newman cautions, however, that bottlenecks might develop as many users try to access the grid at once.

If these schemes succeed in making the giant databases of the future accessible and flexible, many scientists believe that querying will itself become a new research mode, opening a new era of computer-aided discovery. "These databases will be so information-rich, they will enable science that their creators never envisioned," says Caltech astronomer George Djorgovski. One approach is to model the properties of a new object and then go look for it in the database, says Djorgovski, "but that builds in prejudices." He prefers the idea of unleashing software that automatically searches the database for entries with common properties, revealing unsuspected new classes of phenomena. "You will rediscover the old stuff," but now and then, he says, you'll pull something completely new from the floodwaters.

–MARK SINCELL

Mark Sincell is a science writer in Houston.