This multiprocessor architecture is crucial to the analysis application because of the large database involved. Davison expects that the application running on a four-processor SPARC system will execute between 3 and 3.5 times faster than on a uniprocessor machine.

The Solution

"The major advantage of SPARCworks/iMPact is its ability to provide complete diagnostics," Davison said. "It doesn't just tell you whether a piece of code can or cannot be parallelized - it tells you why. To my knowledge, this capability has not been available on any other MP compiler, and has greatly sped up the development process."

During a given run, the compiler might indicate that a module could not be parallelized because it didn't loop a problem that might be solved by re-ordering the loops. "With a conventional multiprocessor compiler, you wouldn't have any idea what the problem was other than you were getting poor performance," said Davison. "Diagnosis would require a separate debugger and internal timing statements. With SPARCworks/iMPact, this analysis is done directly from within the compiler and requires no temporary modification of the code."

Davison also praises SunSoft's accompanying tools particularly the SPARCworks performance analyzer. "These tools alone are worth their weight in gold because of the depth of information they provide," Davison said. "This code has the property of turning anybody's computer into an I/O-bound machine. So it is very important to be able to see the way data structures are laid out in memory. The tools let me see every nook and cranny of the code's performance, including every kind of I/O, paging and memory use. If a problem arises anywhere, I can spot it quickly and correct it."

Benefits

SPARCworks/iMPact delivered the needed performance while significantly shortening development time by taking the guesswork out of parallelizing code and providing a complete set of graphical tools for analyzing performance. "In research, development time is not just money, it stands between a graduate student and a doctorate."

Davison expects to have the first version of his code completed by June 1994, considerably faster than if he were to have used a conventional MP product. The code will be placed on a machine he felicitously calls the Gene-Server, where it will be available via FTP.

###

SunSoft
User Profile
Customer: University of Houston
Product: SPARCworks™/iMPact

"The major advantage of SPARCworks/iMPact is its ability to provide complete diagnostics. It doesn't just tell you whether a piece of code can or cannot be parallelized - it tells you why."

— Dr. Dan Davison, Assistant Professor, University of Houston

Profile

Researchers in the University of Houston's Biochemical and Biophysical Sciences Department are using SPARCworks™/iMPact, SunSoft's new development toolset that adds automatic parallelization to the FORTRAN compiler and helps developers debug and understand parallel applications to gain a better understanding of the relationship between human genes and disease. The multiprocessor application will play a role in the vast Human Genome Project, which seeks to map the entire genetic complement of a human being, with particular focus on regions that affect health. The program will help researchers identify unknown strands of DNA by searching a massive database that represents the sum of knowledge of the human genome.

The application starts with a symbolic representation of a gene sequence, consisting of a string of letters that stand for the four major chemical components of a DNA molecule - adenine (A), thymine (T), guanine (G) and cytosine (C). The application will match these long strings with a massive database consisting of about 200 million letters grouped in 200,000 separate sequences.

"The goal is to match your string with one or more segments of the database, which is composed of all sequences that have been released by researchers throughout the world," said Dr. Dan Davison, an assistant professor at the department. "This lookup is difficult not just because of the size of the database, but because there are gaps where the sequence has not yet been fully fleshed out. Researchers ultimately use this information to help determine the structure and function of a given gene. By locating it within the context of the known genome, they can get a much better idea of how to structure their experiments."

The Challenge

The analysis application has previously run only on computers that were both costly and scarce. Originally developed at Los Alamos for a CDC 6600 computer, the application has since been ported to such high-end machines as the Hypercube, the CM2, and the Cray supercomputer.

Davison wanted to port the application to SPARC® multiprocessor (MP) systems, which deliver high performance at a fraction of the hardware cost. He believes the SPARC MP machines have a good chance of becoming a de facto standard within the scientific community. Nevertheless, the porting process is extensive because the algorithms must be adapted for two- and four-processor systems.