ITANIUM® SOLUTIONS ALLIANCE

Case Study

Parallelizing Biomedical Applications

Interactive Supercomputing (ISC) Star-P™ software enables 200 times speedup of genomic profiling without significant software re-programming

"If your goal is to take the same interactive environment and transfer it to a parallel processing system with a lot more memory, then you'll look for the easiest way to get there. NCI is accustomed to working in MATLAB and with certain formatted files, and this approach retains that environment."

Dr. Mark Potts
Principal of HPC Applications, Inc
Consulting for the National Cancer Institute

Advances in programming tools are speeding up research and being applied to some of the most serious and complex problems, such as finding cures to cancer. These tools simplify the task of adapting research software to massively-parallel computing environments, providing results in hours instead of days. Before the availability of user-friendly parallel computing tools, researchers had to painstakingly rewrite their models to take advantage of affordable, high-performance computers (HPC).

Researchers are working hard to find more effective methods to detect, treat and prevent cancer. At the National Cancer Institute (NCI) researchers are trying to understand how genetics and cancer are linked. The goal is to personalize medicine by customizing the treatment for individual patients. Researchers run matrix operations across large databases to correlate the chromosomes of thousands of cancer patients - genomic data – with risk factors and genetic profiles of tumors. But an explosion in the amount of genomic data available to NCI researchers makes their database mining very time consuming.

Cancer researchers and other technical computing users invest a lot of time developing software models, algorithms and simulations. They often use programming tools, such as MATLAB® from TheMathWorks, to perform computationally intensive tasks faster than with traditional programming languages.

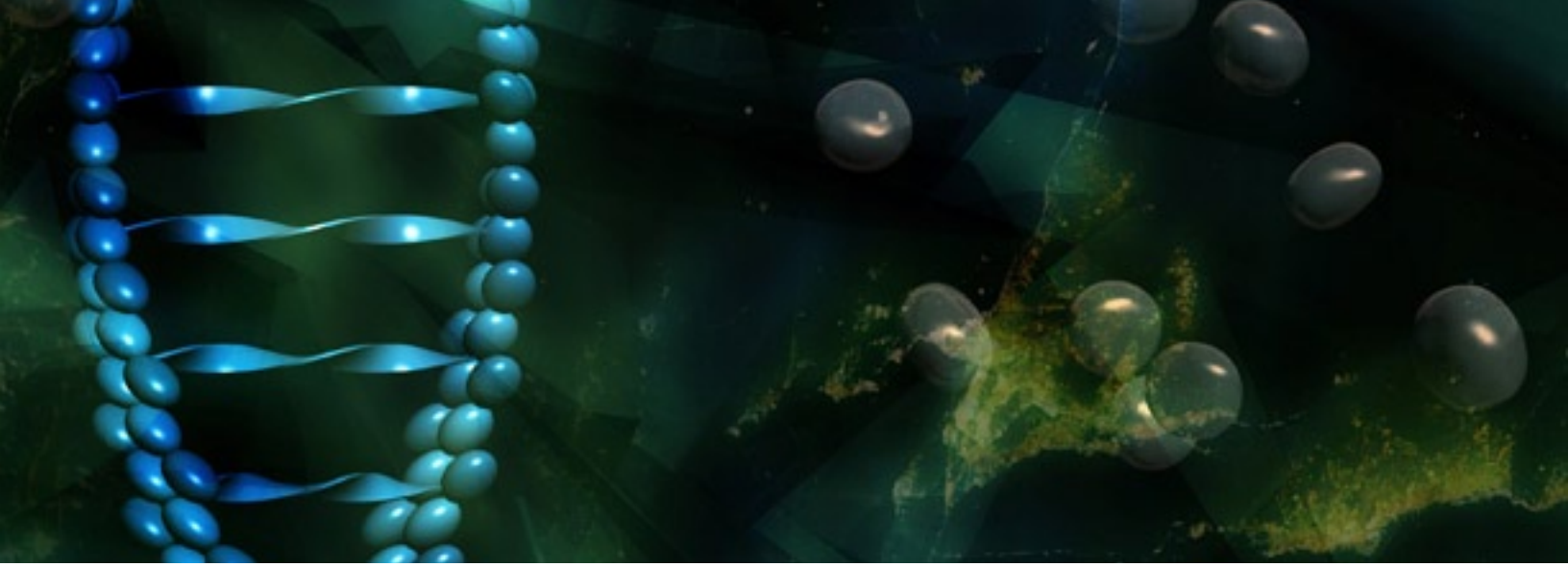
There is a critical issue: many interactive applications like MATLAB don't run on parallel HPC systems, which means some scientists are forced to reprogram software for parallelization, often rewriting it in C++ or Fortran. NCI sought the help of Interactive Supercomputing to provide a tool which would give them the performance benefits from parallel processing without requiring significant changes to their code or MATLAB development environment.

Challenge

- Mine databases faster. Reduce the time for correlation software to complete, currently measured in days, and remove the obstacle to attempting more complex search.
- Minimize software reprogramming. Eliminate the need for manual parallelizing of complex software and avoid a major time sink and new software bugs.

Solution

- Deploy Star-P software. Maintain existing MATLAB development environment with only a few simple modifications to software code.
- Upgrade hardware. Run software on 64 Intel Itanium 2 processors with 256 GB of RAM.



Implementing a Parallel Computing Platform

Star-P is an interactive parallel computing platform that lets users continue to work with their preferred desktop tools, like MATLAB, Python* and R*, but run the compute intensive algorithms on HPC systems. This is accomplished by implementing a client-server model, allowing users continue to run their desktop tools on their clients, typically desktop computers, without any functionality limitations. Star-P merges two previously distinct environments - desktop computers and high performance servers - into one.

Star-P software runs on both the client and the server. The client software interfaces to the desktop tools and communicates to the server. The server software controls the resources of the HPC system, processors and memory, and executes the parallelization strategies. Star-P employs libraries to support task and data parallelization and provides an API for users to plug in their own existing serial and parallel code libraries.

Users need an awareness of their software, such as knowing which code portions represent the greatest workload and where there are parallelization opportunities. Task-level parallelization can be done on relatively short tasks whose execution scheduling and data requirements are independent of each other. Data-level parallelization occurs when the same operation is repeated on a large number of data elements. The correlation searches done by NCI are data parallel operations because the same function is done on millions of data elements populating large matrices.

To help users optimize the execution of algorithms, the Star-P performance analyzer displays the amount of time spent on clients, servers and the network. This information can be used to improve parallelization strategies and resolve contention issues that create execution delays.

Two key benefits from using Star-P are that users can continue to optimize their software on familiar tools, and they don't need to manage the complexities of the HPC hardware because these are handled by Star-P. This can help technical users, like the more than one million MATLAB users in industry and academia, dramatically speed up their application execution with significantly less software development effort than before.

Spotlight: Interactive Supercomputing (ISC)

- More than two dozen organizations have already adopted Star-P software for solving some of the world's greatest scientific and engineering challenges.
- Star-P 2006 Most Significant HPC Software Product

  - HPC Wire Editor Choice Awards

"Technical computing users would prefer to continue working with their favorite desktop tools while tapping into the computing muscle of parallel systems. But there is clearly pain involved with re-programming their desktop models to run on parallel systems."

Peter Simon
President
Simon Management Group

“Running a single correlation on a desktop computer could take a week or more to complete. Their (NCI) tasks require more computing power, more system memory, and - all too often - more time. And in the race to understand how genetics and cancer are linked, time is precious.”

Bill Strecker
Chief Technical Officer
Interactive
Supercomputing

“In computing with humans, response time is everything. One’s likelihood of getting the science right falls quickly as one loses the ability to steer the computation on a human time scale.”

Prof. Nick Trefethen
Oxford University

Summary and Metrics

Turn-around times of up to a week had proven to be the practical limit for NCI researchers. And with bioinformatics on a steep growth curve, the problem was only growing worse.

Before migrating to Star-P, researchers ran their correlations on desktop PCs, with some “routine” algorithm processing taking more than two days to complete. In addition to the limitation that the desktop systems were serial, not parallel, system memory was limited to 2 GB. Since NCI operated on large matrices, 30,000 by 30,000, the desktop systems spent a lot of time swapping data between RAM and hard disk drives.

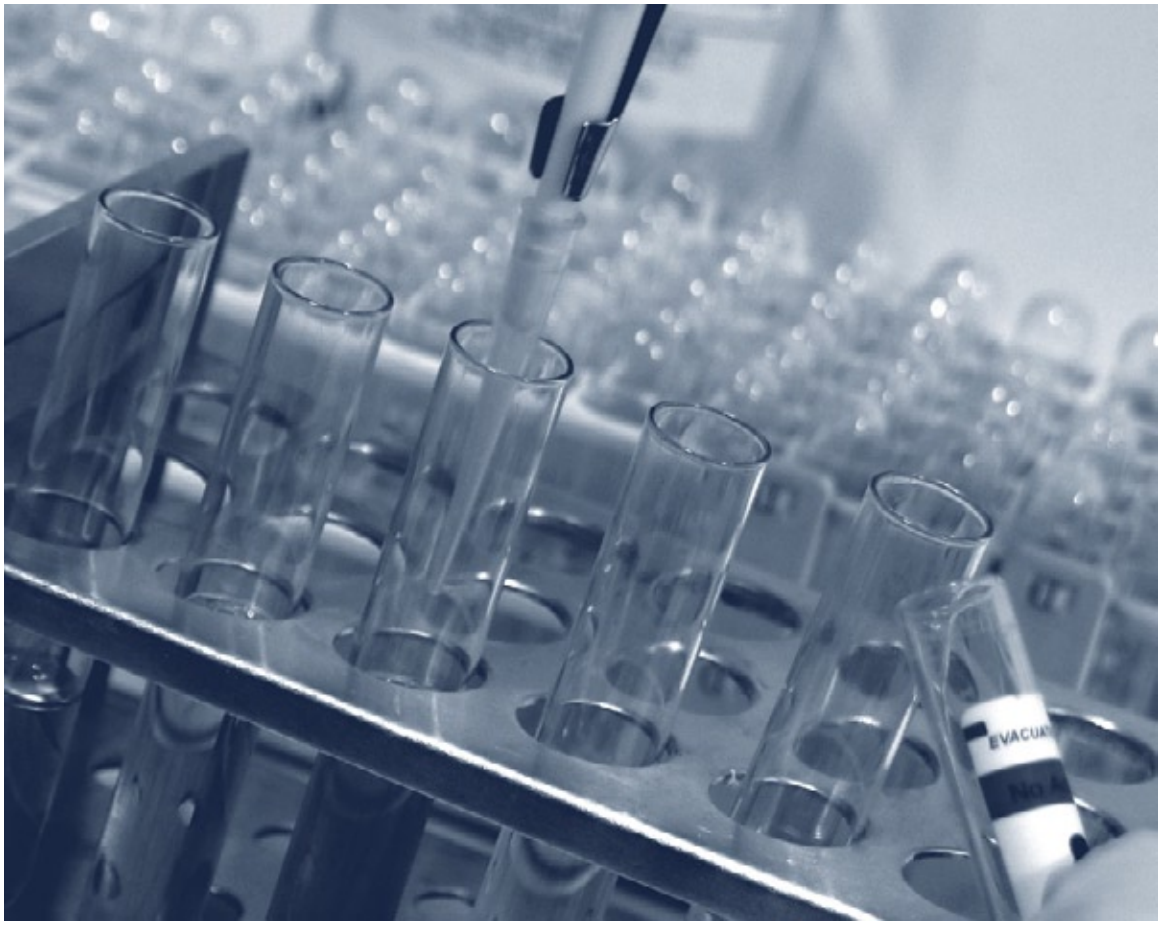
By migrating from a desktop environment to an Intel Itanium-based system, with 64 processors and 256 GB of memory using Star-P, the entire correlation took less than 15 minutes.

Future Cancer Research

The combination of Star-P software and Itanium 2-based severs gives NCI the ability to analyze the data from more patients and possibly approach problems differently than they would have before. With a more powerful parallel system at their disposal, researchers can try even more complex searches that previously weren’t an option. The current sever architecture can scale up to 256 Intel Itanium 2 processors and 1TB RAM, quadrupling the number of processors and system memory.

The research team estimates their computation limit of 30,000 by 30,000 data matrices could be increased to a data matrix of 100,000 by 100,000. By increasing the matrix size, researchers have a better chance to uncover key genetic relationships and unlock the mysteries behind cancer.

.



Key Technologies

- NCI's data mining application employs Star-P software to run on a parallel computing system, maintaining the MATLAB development environment.
- The HPC system integrates SGI* Altix* servers powered by 64 Intel Itanium 2 processors and 256 GB of RAM.

Multiprocessing is managed by the SUSE* Linux* Enterprise Server 9 operating system from Novell supporting high-availability clustering services.

Other Successes

Molecular Simulation: Model the thermodynamic properties of smog to predict its interaction with weather systems.

Radar System Design: Increase image processing capacity as satellite radar replaces land-based and Megabyte data transmissions expand to Terabytes.

Sparse Matrices & Large Graphs: Distribute calculations on large matrices - with dimensions in the millions - across multiple processors used in computational biology and web searches.

Benefits of Using Itanium® 2-based Solutions

Itanium 2-based systems offer significant advantages when running software applications parallelized with Star-P. With single system image node sizes from 8 to 256 processors, along with the ability to support shared memory across cluster nodes, the configuration can be scaled for any HPC application. The unique "expand on demand" capability allows I/O and memory to be scaled independently of processors, so users can precisely tune their system resources for specific application workloads. For example, an Itanium 2-based system can apply multiple processors to a single problem or commit a single processor to handle computational tasks while still accessing large amounts of system memory to work with a large data matrix.

Built on industry-standard components Itanium 2-based platforms are available in a vast range of configurations, from entry-level and mid-range cluster solutions to larger shared memory clusters, servers and supercomputers that scale as users' needs evolve.

ITANIUM® SOLUTIONS ALLIANCE