

Parallel Architectures in Biotechnology

Nina Baron-Hionis
CIT595
University of Pennsylvania
Spring 2007

1 Introduction: Parallel Architectures

Parallel computers use the idea of running many processors concurrently to solve problems. This is a cost effective way to get higher speeds using technology that is not highly expensive or not even available. In general parallel computer architectures use this idea but there are many different design configurations that are optimal depending on the specific type of programming the computer will be used for. Forty years ago, Michael Flynn, theorized a taxonomy for parallel architectures, which is still applicable today. His classification is based on parallelism in instruction and data streams. There are four classifications as follows:

1. Single Instruction stream Single Data stream, SISD- one processor with a single data stream.

2. Single Instruction stream Multiple Data streams, SIMD- a single instruction is processed and several processors using multiple data streams.
3. Multiple Instruction streams Single Data stream, MISD- a single data stream is processed by several processors
4. Multiple Instruction streams Multiple Data streams, MIMD- several instructions, several data streams, multiple processors that each fetch and execute their own instructions using their own data stream.

Computers in the past were mostly SISD architectures while today most computers are of the MIMD variety. Two major considerations with parallel computer architectures are how the processors and memory communicate and the structure of memory. With this concern, MIMD architecture computers can then be further classified based on the number of processors they have, which also determines how memory is interconnected. There are centralized shared memory architectures and distributed memory architectures.(3)

Centralized shared memory architectures usually have fewer than three dozen processors and use a shared single memory. (See Figure 1.1. Simple Shared Memory from Wikipedia)

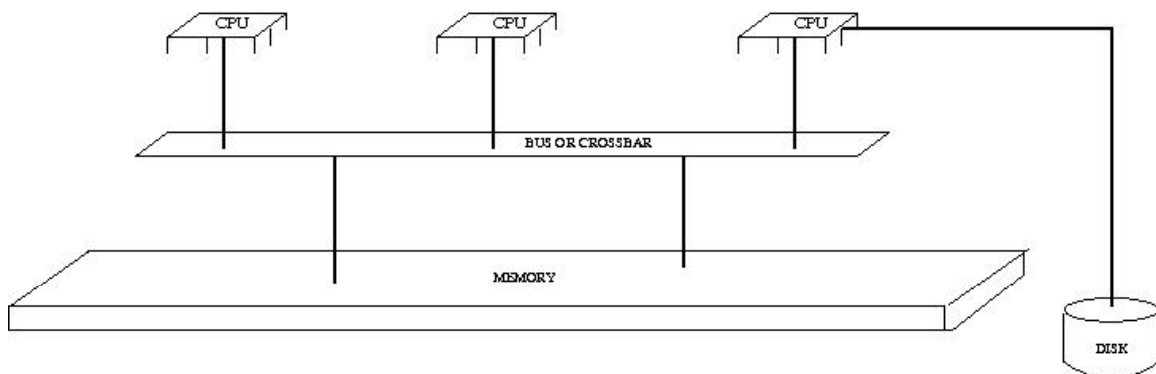
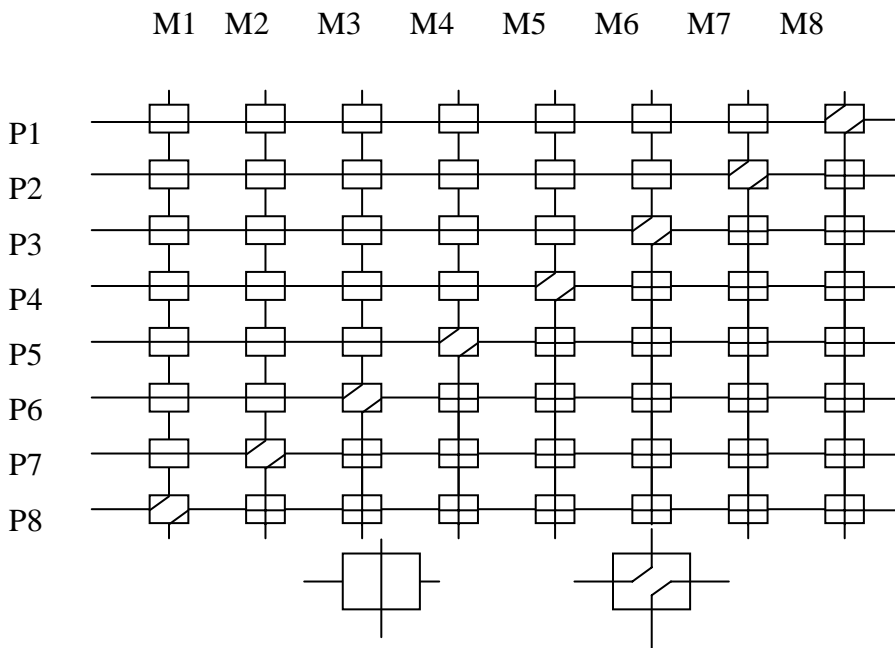


Figure 1.1

Each processor has equal accessibility to memory thus this type of architecture is sometimes referred to as uniform memory access. This type of computer does not try to execute multiple programs at one time but attempts to complete one process using all processors in a smaller amount of time than if only one processor was executing the process. A single bus or a crossbar typically interconnects the processors and memory.

Using a single bus is the simplest and least expensive design but it is also the limiting factor when it comes to the performance of the overall system. Regardless of how fast the processors may be executing processes if the bus has too much traffic it will slow the overall system down. Using a crossbar can alleviate this performance bottleneck. A crossbar hardware interconnection design is similar to using multiple buses. (1) (See Figure 1.2, An 8 x 8 crossbar)

Figure 1.2



Every processor is connected to memory and/or peripheral through multiple paths and all paths can be active at the same time. This design allow for several processors to be accessing memory at the same time. Where as with the single bus, the bus is solely dedicated to the processor that is using it and all the other processor have to sit idle waiting for their turn. Although the crossbar design allows for more processors to access memory at one time there is a saturation point where the system will slow down just as with a single bus. Crossbars are also more expensive and there is higher degree of complexity associated with building and programming this type of design.(1)

Distributed memory architectures are the design solution for systems with more then three dozen processors and high memory bandwidth demands. (3)(See Figure 1.3. Simple distributed Memory, from Wikipedia).

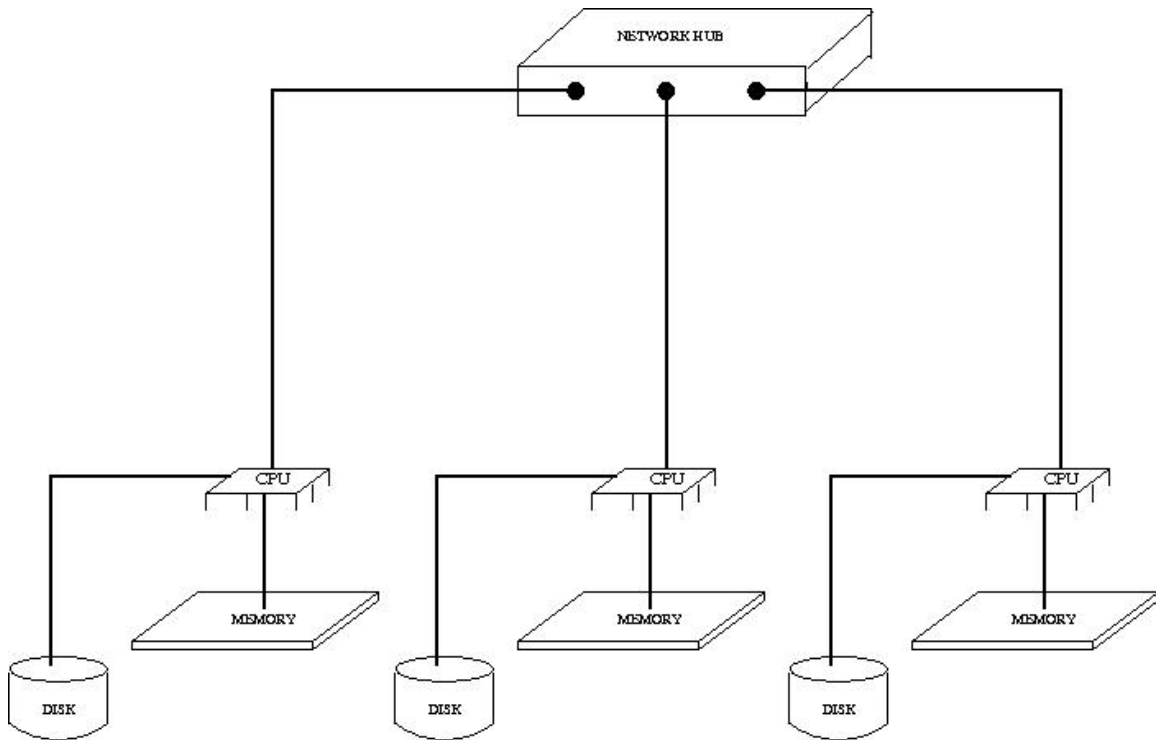


Figure 1.3

There are two types of memory architectures found in this type of distributed memory. One is a distributed shared memory architecture, where the separate memory spaces are accessed as if they were one physically shared memory space. The other type is a when each processor has a private memory address space and other processors do not have access. Identical memory address associated with different processors map to different locations. Communication among the processors is done directly with distributed shared memory design and is done indirectly through messages between the processors in the private address design. This type of system requires a high bandwidth interconnection that usually employs switches. Switching networks commonly use either crossbar switches or two by two switches. Crossbar switches are either open or closed while 2 by 2 switches can have one of four different states: through, cross, upper broadcast and lower broadcast. There are two inputs, upper and lower, and two outputs, upper and lower. Through has upper input to upper output and lower input to lower output. Cross has upper input to lower output and lower input to upper output. Upper broadcast has upper input go to both upper and lower outputs. Lastly, lower broadcast has the lower input go to both upper and lower outputs.(3,1) (See Figure 1.4. 2x 2 switch)

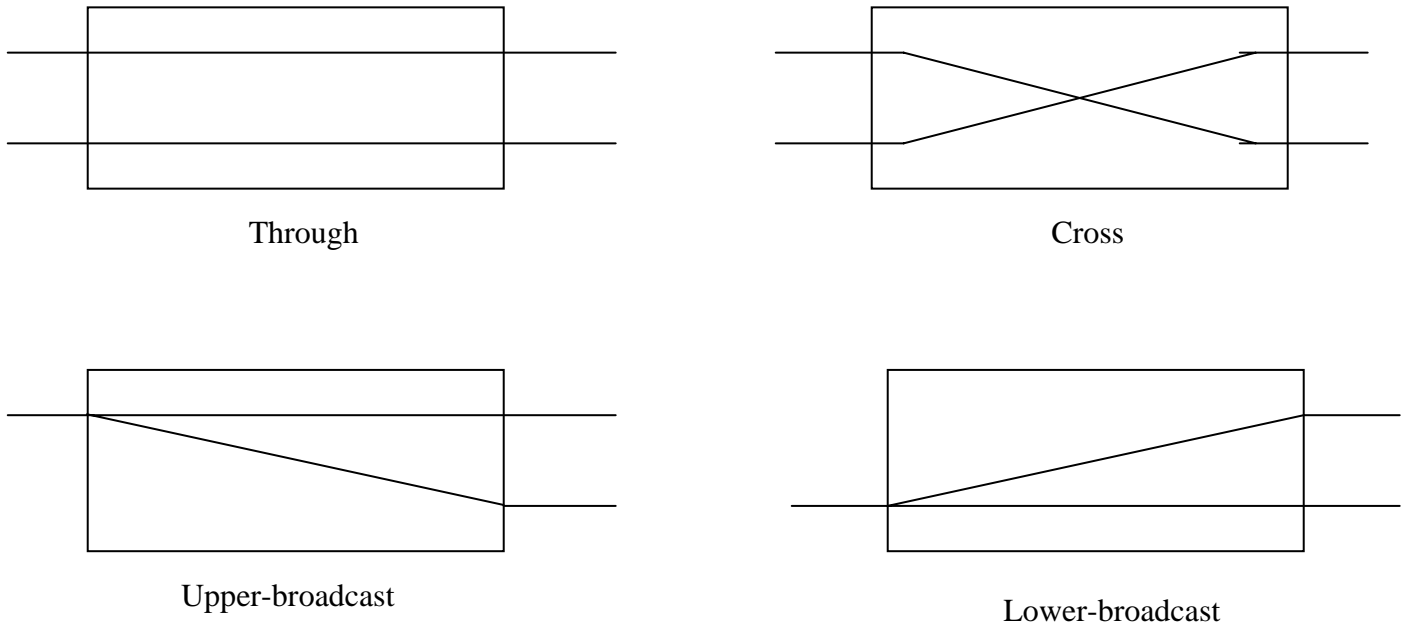


Figure 1.4

Distributed memory architectures also suffer latency problems when processor's are trying to communicate information with other processors because this can only be done across some type of shared interconnection.

2 Parallel Architectures in Biotechnology

Parallel computers are the ideal cost effective solution for the need for high power computing in biotechnology. Many problems such as modeling, protein folding and sequence analysis require extremely fast computers to do thousands of computations in a reasonable amount of time. Depending on the specific type of problem there are several varieties of parallel architectures, massively parallel architectures (supercomputers), cellular architectures and systolic architectures.

2.1 BlueGene/L: Massively Parallel Computer

BlueGene/L is a massively parallel supercomputer designed by IBM. Currently it is considered to be the fastest computer in the world. BlueGene/L has 65,536 nodes. Each node was built with new on-chip technology, which means all the necessary parts of the

system are on a single integrated circuit. Each of these chips has 2 processors.(See Figure 2.1.1 BlueGene/L Chip)

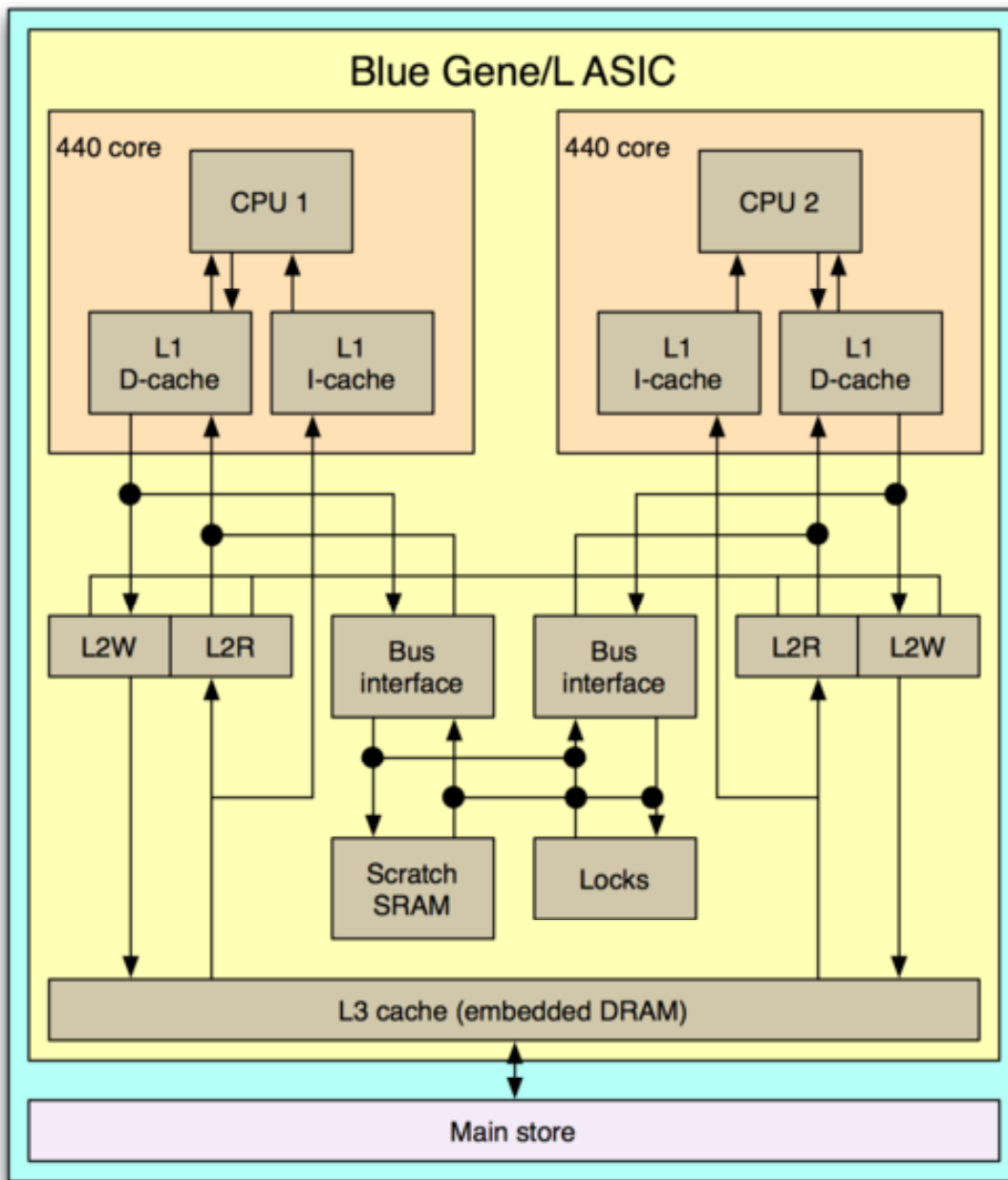


Figure 2.1.1

All the nodes are interconnected through 5 networks with the main network consisting of a 64x32x32 3D torus network. (See Figure 2.1.2 A simple 2 x 2 x 2 torus network) A torus network allows unrestricted node-to-node communications. For global

communications there is a collective network that extends the entire BlueGene/L machine.(See figure 2.1.3. A simple collective network).

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 2.1.2

Figure 2.1.3

The BlueGene/L computer is a “dedicated application specific machine.” It was designed with the specific application purpose of scientific computing. Some of the applications it was designed for are: modeling of biological phenomena, real time data processing and offline data analysis. (6) One of the major projects that it was developed for is the Blue Brain project. The Blue Brain project’s goal is to reverse-engineer the mammalian brain, starting at a cellular level. Given the complexity of the brain a massively parallel computer such as BlueGene/L is exactly the type of computer needed to model the brain. Each node in the computer is going to represent a neuron and the connections will be formed between the nodes/neurons over the networks. The ultimate goal of this project is to further the understanding of the function and the dysfunction of the brain for the better development of therapies for diseases and/or injuries of the brain. (7)

2.2 Cyclops64: Cellular Architecture

Cyclops64 is another IBM designed supercomputer. Cyclops64 is based on a cellular architecture, which is another variation of a parallel architecture. The main features of a cellular architecture are that the programmer has direct access to the hardware configuration so to design and optimize software specifically for the system and the use of on-chip technology or “supercomputer on a chip” with multi thread parallelism in each processor. (See Figure 2.2.1. A Cyclops64 cellular architecture)

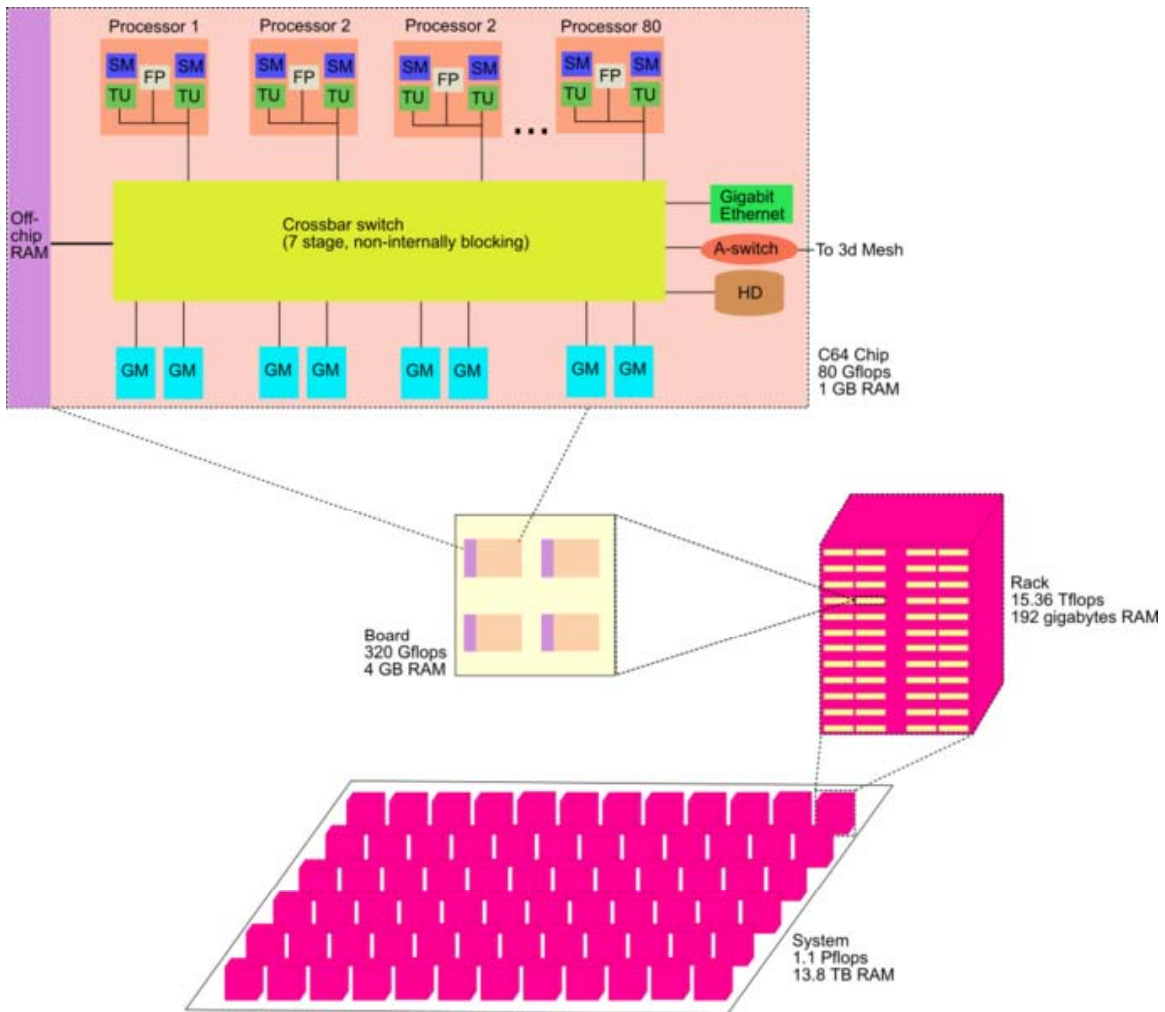


Figure 2.2.1

Each chip has 80 processors with 2 thread units, memory and communication. A crossbar network interconnects the chips. (See Figure 2.2.2 Cyclops 64 node)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 2.2.2

The Cyclops64 supercomputer is being designed for the study of molecular dynamics and protein folding. It is currently not finished with an expected completion date of late 2007.(2)

2.3 Kestrel: A Systolic Linear Array

Systolic arrays are generally single instruction multiple data stream architectures. They have large array of processors that use a vector pipeline for data flow. Data is computed by circulating data through the system of processors, similar to how blood flows through the heart and rest of body, thus the name systolic array. (See Figure 2.3.1 A simple systolic array)

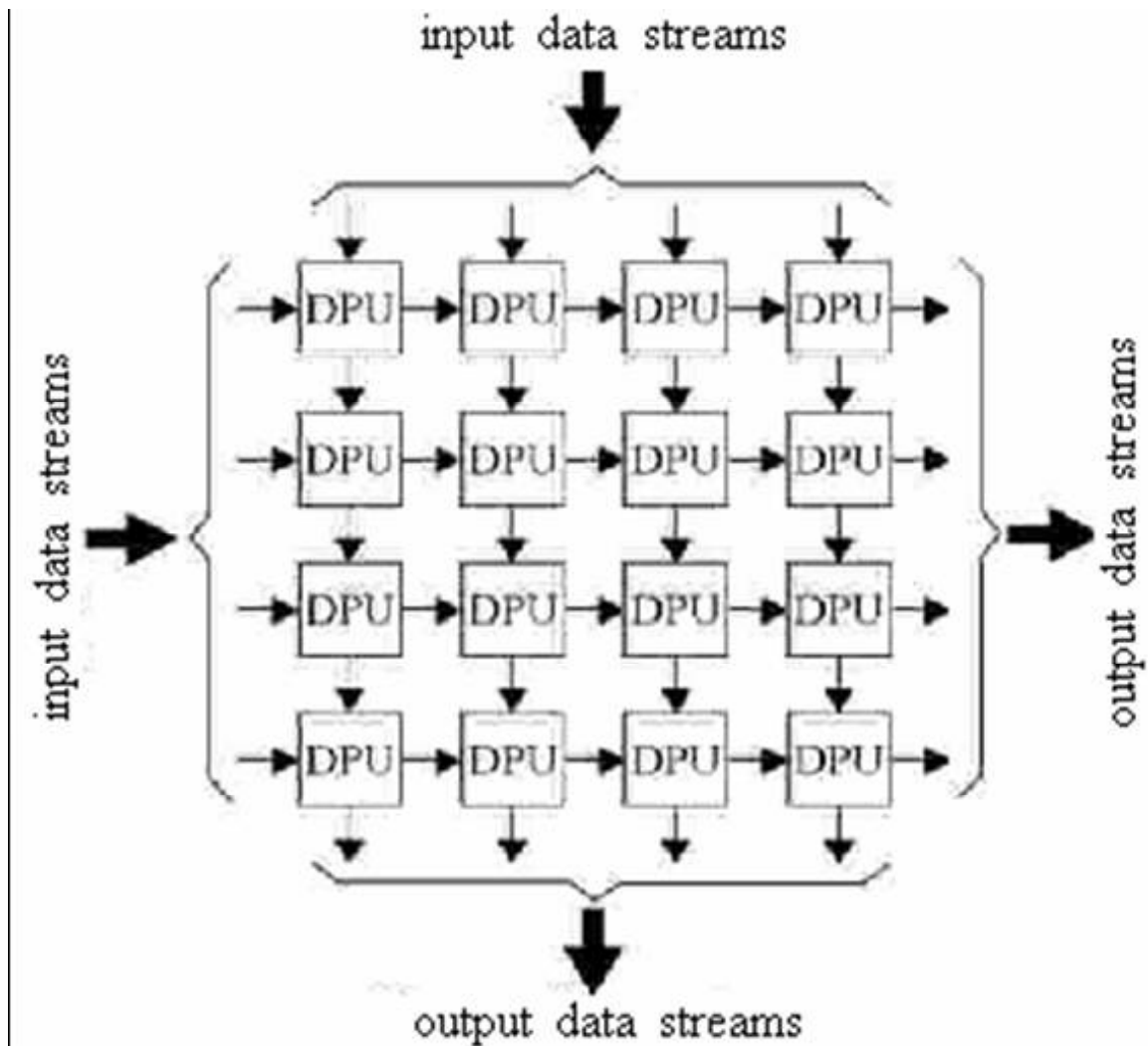


Figure 2.3.1

Kestrel is single purpose machine with dedicated hardware. It uses a SIMD architecture with 512 unit linear array and systolic shared registers. Kestrel was designed for the purpose of DNA and RNA sequence analysis.(4,5)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 2.3.2

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 2.3.3

3 Summary

Parallel computing systems are the ideal answer for University environments which need to find cost effective architectures for high powered computing. Also the need in the biotechnology field is also to find systems that can compute in a reasonable amount of time. It is easy too see with DNA and RNA sequence analysis that some type of high end computer would be needed since there are billions of base pairs to be analyzed. Protein folding is also an extremely math intensive application, as researchers have tried to develop algorithms that will predict protein structure based on a proteins amino acid sequence. Lastly, modeling the mammalian brain. We know so little about the brain but we do know there are millions on neurons that interconnect in a so many ways to pass information to other areas of the brain in fractions on a second. In order to model this type of behavior it is easy see why the BlueGene/l super computer is currently the fastest computer in the world. Only a computer with this type of processing speed could even come close to modeling the behavior of the brain and even still it will never be as fast.

4 References

1. Dowd, Kevin and Severance, Charles. High Performance Computing. O'Reilly and Associates, 1998.
2. Gao, Guang et all. "Toward a Software Infrastructure for the Cyclops Cellular Architecture". *University of Delaware, Department of Electrical and Computer Engineering, Computer Architecture and Parallel Systems Laboratory*.
3. Hennessy, John and Patterson, David. Computer Architecture: A Quantitative Approach. Morgan Kaufmann Publishers, 2003.
4. Hughey, Richard et all. "The UCSC Kestrel Parallel Processor." *IEE Transactions on Parallel and Distributed Systems*. Vol. 16, No. 1, Jan 2005
5. Null, Linda and Lobur, Julia. Computer Organization and Architecture. Jones and Bartlett Publishers, 2006.

6. Vranas, P. et al. "Overview of the BlueGene/L system Architecture". *IBM Journal of Research and Development*. Vol. 49, No. 2/3 March/May 2005.

7. Blue Brain Project
<http://bluebrain.epfl.ch/>