**MAIN PROJECT REPORT –SYSTOLIC PROCESSOR ARCHITECTURE IMPLEMENTION USING FPGA**

**PROJECT GUIDE : SUBMITTED BY :**

Deepak J Puthukkaden (20)

Giridhar A K (31)

Govindh B (32)

Rohit Sreekumar (53)

Mr. Rashid P.E Shine Ali (59)

**Chapter 1**

**INTRODUCTION**

The most widely known mode of information processing is sequential processing in which the operations that have to be performed to obtain the solution of a problem are carried out in a sequence by a single processing unit. In case the processor is computer it is termed sequential computing. The most common computer design is still based on a von Neumann architecture, which is composed of a single processor executing single sequence of instructions on a single stream of data known also under SISD machine. In von Neumann architecture both instructions and data come from a common memory, while in a Harvard architecture data and instructions are stored in different memories and are fed to the processor by two different paths . The conceptual simplicity of the sequential approach and associated technology at the time led to digital computer revolution.

For most applications sequential computers present an efficient solution providing the required processing power. Nevertheless, there were, are, and will be problems that require computational power that exceeds the resources of the most powerful sequential computers at the time for several orders of magnitude. These problems are not only those referred to as large-scale scientific computing problems, but also several real-time signal and image processing problems. Some of them were covered in the previous chapter.

In order to achieve radically higher computational throughput new approaches must be considered. One of already prominent approaches is solving problems in parallel. As opposite to sequential processing, parallel processing could be regarded as a mode of information processing in which at least two processing units cooperate while carrying out in parallel the information processing operations that belong to a problem. During the process each processing unit works on a different part of a problem.

Usually, multiprocessing and parallel computing are used as synonyms referring to a number of electronic processing units involved in parallel computing of a particular computational problem. If number of these processing units is large, in order of hundreds or thousands, the attribute massively is added.

The main reason for parallel execution of a problem is the acceleration of computations. If single processing unit can accomplish the task in time *T* then *n* processing units could ideally accomplish it in time *T/n*. This is termed a linear speed-up. In most cases the actual speed-up achieved on parallel computers is considerably smaller then the desired linear speed-up.

In parallel computer architectures, a systolicarray is a homogeneous network of tightly coupled Data Processing Units (DPUs) called cells or nodes. Each node or DPU independently computes a partial result as a function of the data received from its upstream neighbors, stores the result within itself and passes it downstream.

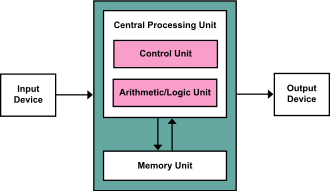
The parallel input data flows through a network of hard-wired processor nodes, resembling the human brain which combine, process, merge or sort the input data into a derived result. Because the wave - like propagation of data through a systolic array resembles the pulse of the human circulatory system, the name systolic was coined from medical terminology. The name is derived from Systole (medicine) as an analogy to the regular pumping of blood by the heart.

**Chapter 2**

**LITERATURE REVIEW**

**2.1 Comparison with Von Neumann Architecture**

This design architecture consists of a processing unit containing an arithmetic logic unit and processor registers, a control unit containing an instruction register and program counter, a memory to store both data and instructions, external mass storage, and input and output mechanisms. The meaning has evolved to be any stored-program computer in which an instruction fetch and a data operation cannot occur at the same time because they share a common bus.



**Fig 1 : Von Neumann Architecture**

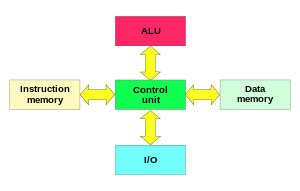
**Limitations and Overcoming Methods :-**

The main limitation of the von Neumann architecture is known as the "von Neumann bottleneck". This is due to the fact that all instructions and all data must pass through the same shared common multiplexed bus to get in or out of the processor, sooner or later things have to wait for other things to get access to this multiplexed bus and the processor gets starved for instructions and/or data. The result is the processor is unable to maintain its designed performance but waits idle instead of doing work.

A major benefit of systolic arrays is that all operand data and partial results are stored within (passing through) the processor array. There is no need to access external buses, main memory or internal caches during each operation as is the case with Von Neumann or sequential machines.

**2.2 Comparison with Harvard Architecture**

The Harvardarchitecture is a computer architecture with physically separate storage and signal pathways for instructions and data.  These machines had data storage entirely contained within the central processing unit, and provided no access to the instruction storage as data. Programs needed to be loaded by an operator; the processor could not initialize itself.



**Fig 2 : Harvard Architecture**

**Limitations and Overcoming Methods :-**

Free data memory cant be used for instruction and vice- versa. Production of a computer with two buses is more expensive and needs more time. Moreover this implementation requires more number of pins and it is not commonly used.