Advanced Computer Architecture
(0630561)

Lecture 11

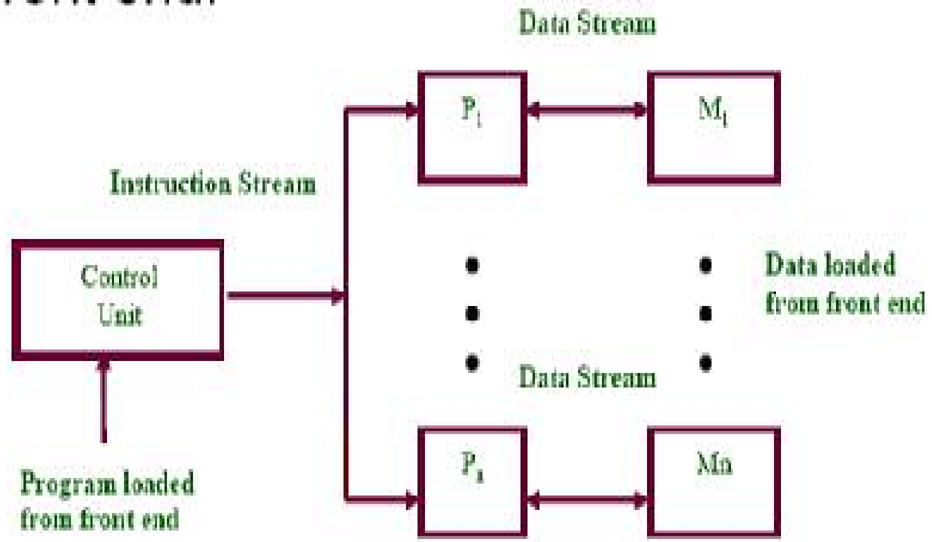
Vector Processing

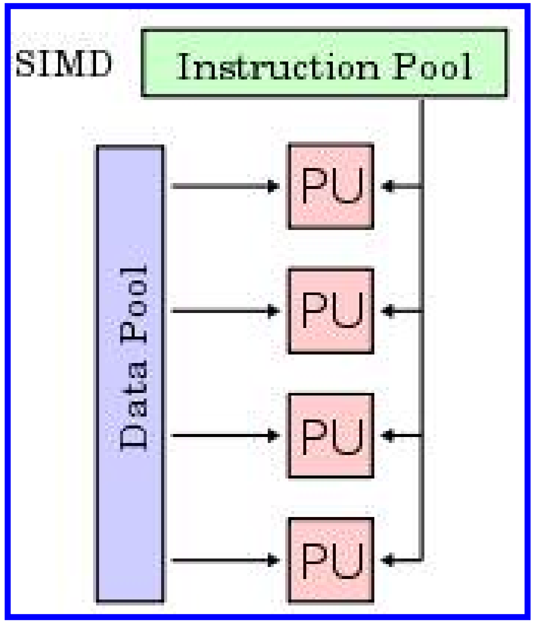
Prof. Kasim M. Al-Aubidy

Computer Eng. Dept.

Single Instruction Multiple Data (SIMD)

- Consists of 2 parts:
  - a front-end Von Neumann computer.
  - A processor array: connected to the memory bus of the front end.



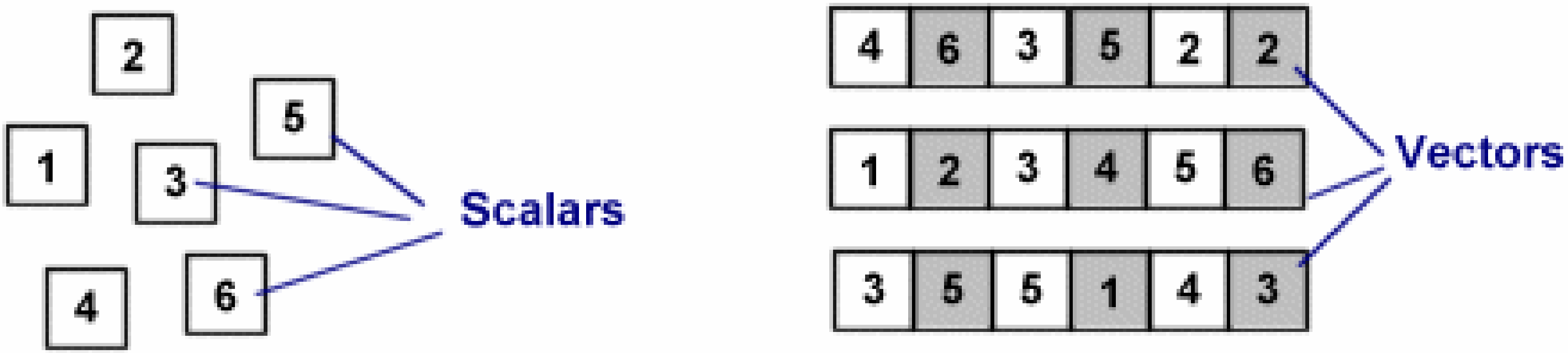


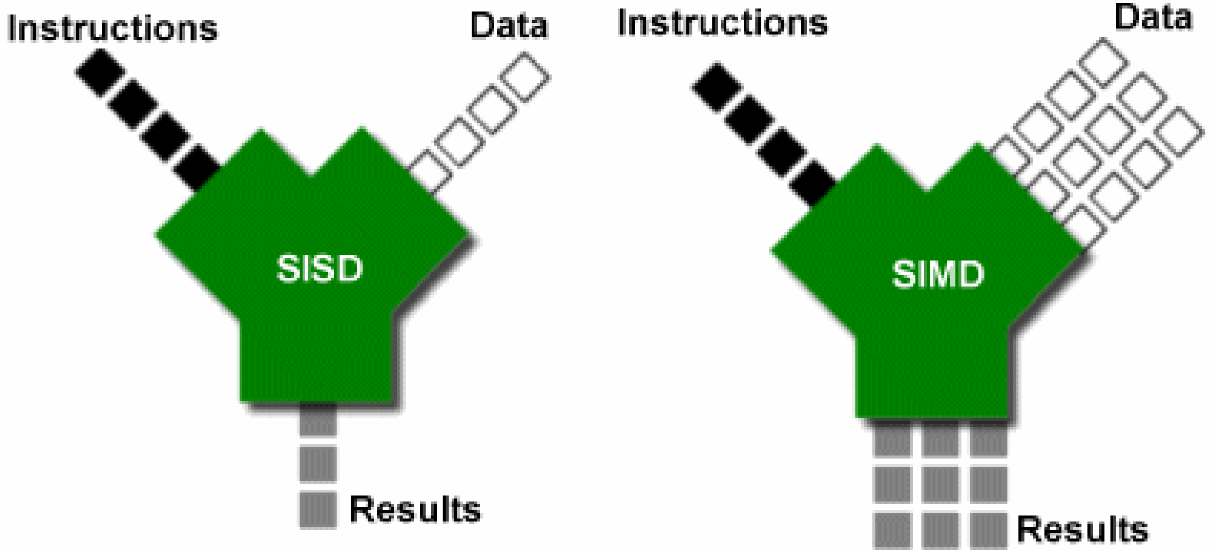
- Vector computers are equipped with scalar and vector hardware or appear as SIMD machines.

SIMD operations

The basic unit of SIMD love is the *vector*, which is why SIMD computing is also known as vector processing.

A vector is nothing more than a row of individual numbers, or scalars.



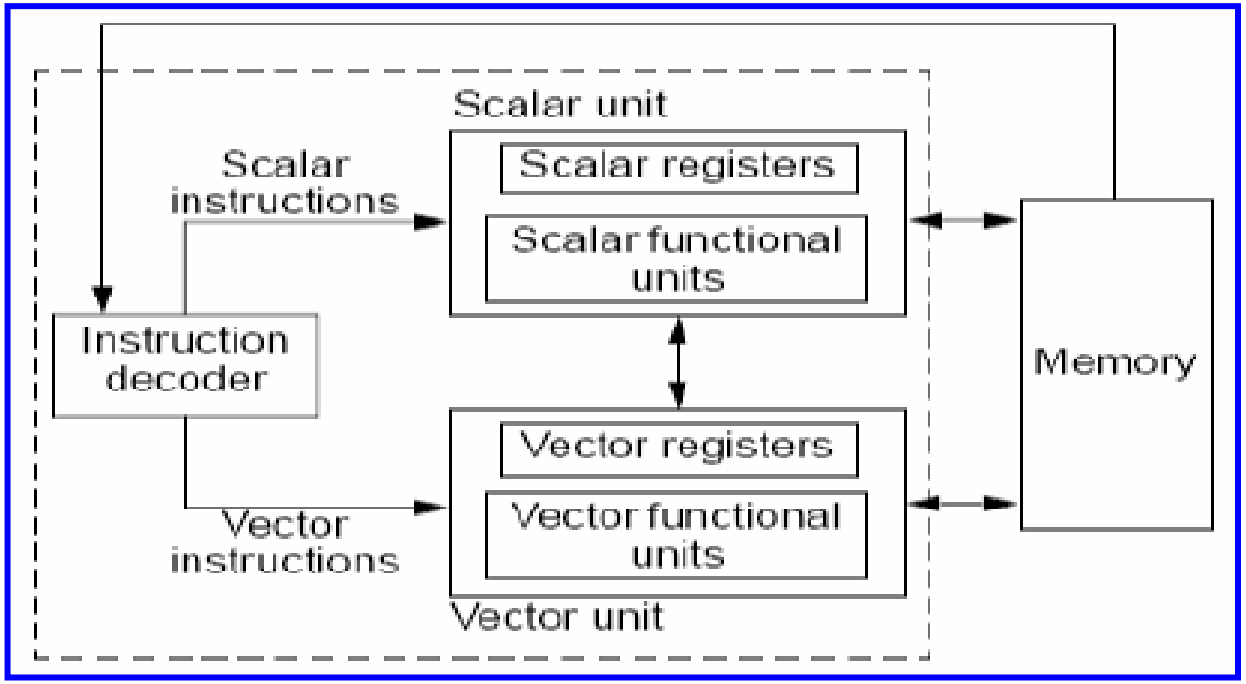


Vector Processors:

Vector processors are SISD processors which include in their instruction set instructions operating on vectors. They are implemented using pipelined functional units.

Several computer architectures have implemented vector operations using the parallelism provided by pipelined functional units. Such architectures are called *vector processors.*

A vector unit typically consists of pipelined functional units and vector registers



- Vector processors are not parallel processors; there are not several CPUs running in parallel. They are SISD processors which have implemented vector instructions executed on pipelined functional units.
- Vector computers usually have vector registers which can store each 64 up to 128 words.
- Vector processors include in their instruction set, beside scalar instructions, also instructions operating on vectors.
- Vector instructions:
  - load vector from memory into vector register
  - store vector into memory
  - arithmetic and logic operations between vectors
  - operations between vectors and scalars
  - etc.

Operation code	Base address source 1	Base address source 2	Base address destination	Vector length

Vector Processing:

- Science and Engineering Applications
  - Long-range weather forecasting, Petroleum explorations, Seismic data analysis, Medical diagnosis, Aerodynamics and space flight simulations, Artificial intelligence and expert systems, Mapping the human genome, Image processing
- Vector Operations
  - Arithmetic operations on large arrays of numbers
  - Conventional scalar processor
    - » Machine language

```
    Initialize I = 0
20  Read A(I)
    Read B(I)
    Store C(I) = A(I) + B(I)
    Increment I = I + 1
    If  I ≤ 100 go to 20
    Continue
```

    - » Fortran language

```
    DO  20  I = 1, 100
20  C(I) = A(I) + B(I)
```

  - Vector processor
    - » Single vector instruction

```
C(1:100) = A(1:100) + B(1:100)
```

Vector Instruction Format :

Operation code	Base address source 1	Base address source 2	Base address destination	Vector length
ADD	*A*	*B*	*C*	*100*

Matrix Multiplication

- 3 x 3 matrices **multiplication** : n² = 9 inner product

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

» $c_{11} = a_{11}\,b_{11} + a_{12}\,b_{21} + a_{13}\,b_{31}$

- Cumulative **multiply-add** operation : n^3 = 27 multiply-add

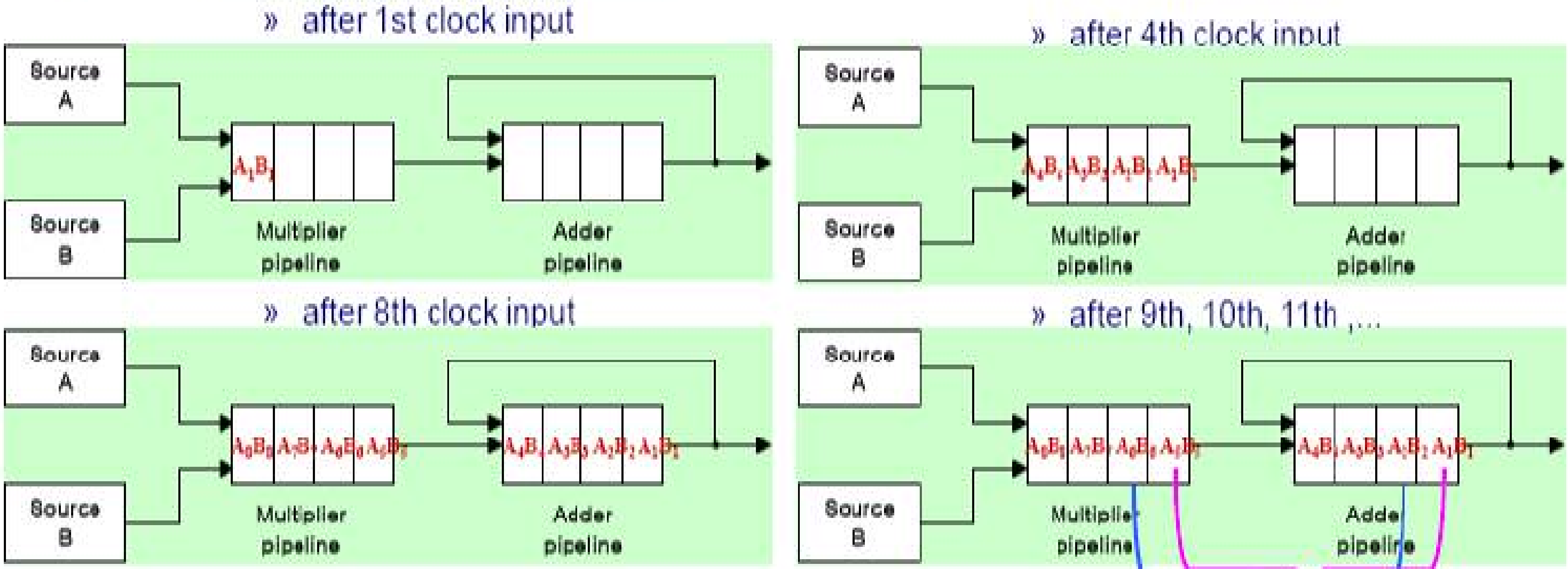
$$c \;= c + a \times b$$

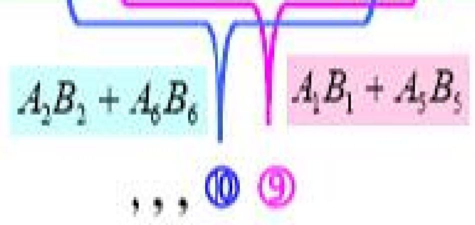
» $c_{11} = c_{11} + a_{11}\,b_{11} + a_{12}\,b_{21} + a_{13}\,b_{31}$

① ① ② ② ③ ③

Pipeline for calculating an inner product :

- Floating point multiplier pipeline : 4 segment
- Floating point adder pipeline : 4 segment
- $C = A_1B_1 + A_2B_2 + A_3B_3 + \cdots + A_kB_k$



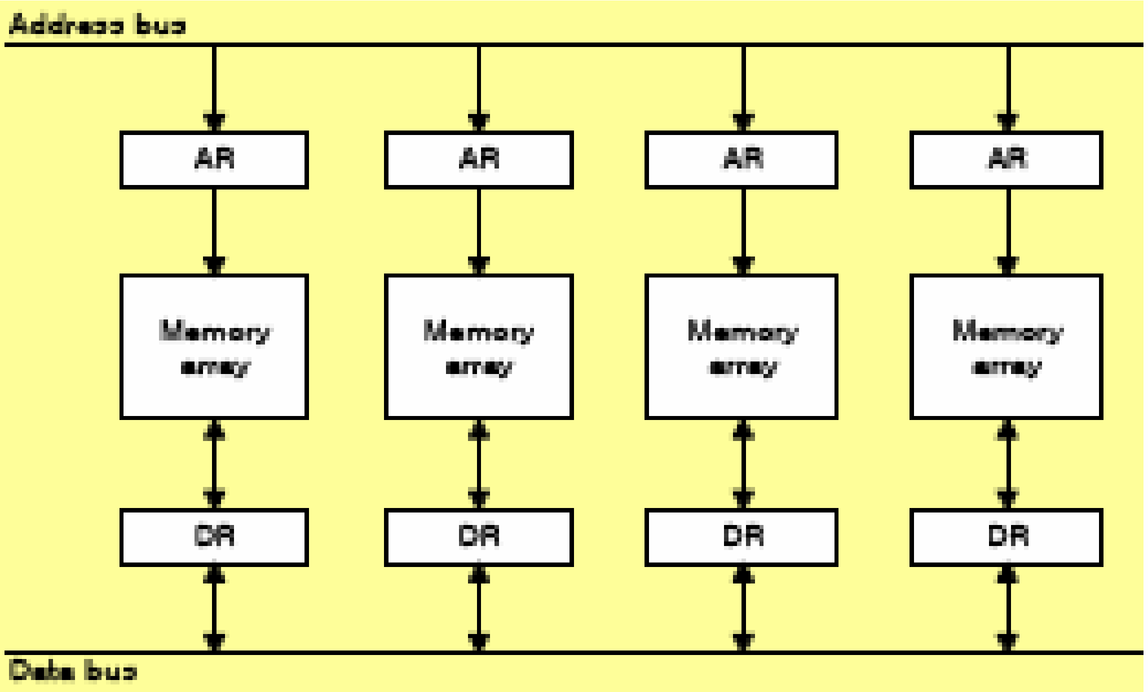


» Four section summation

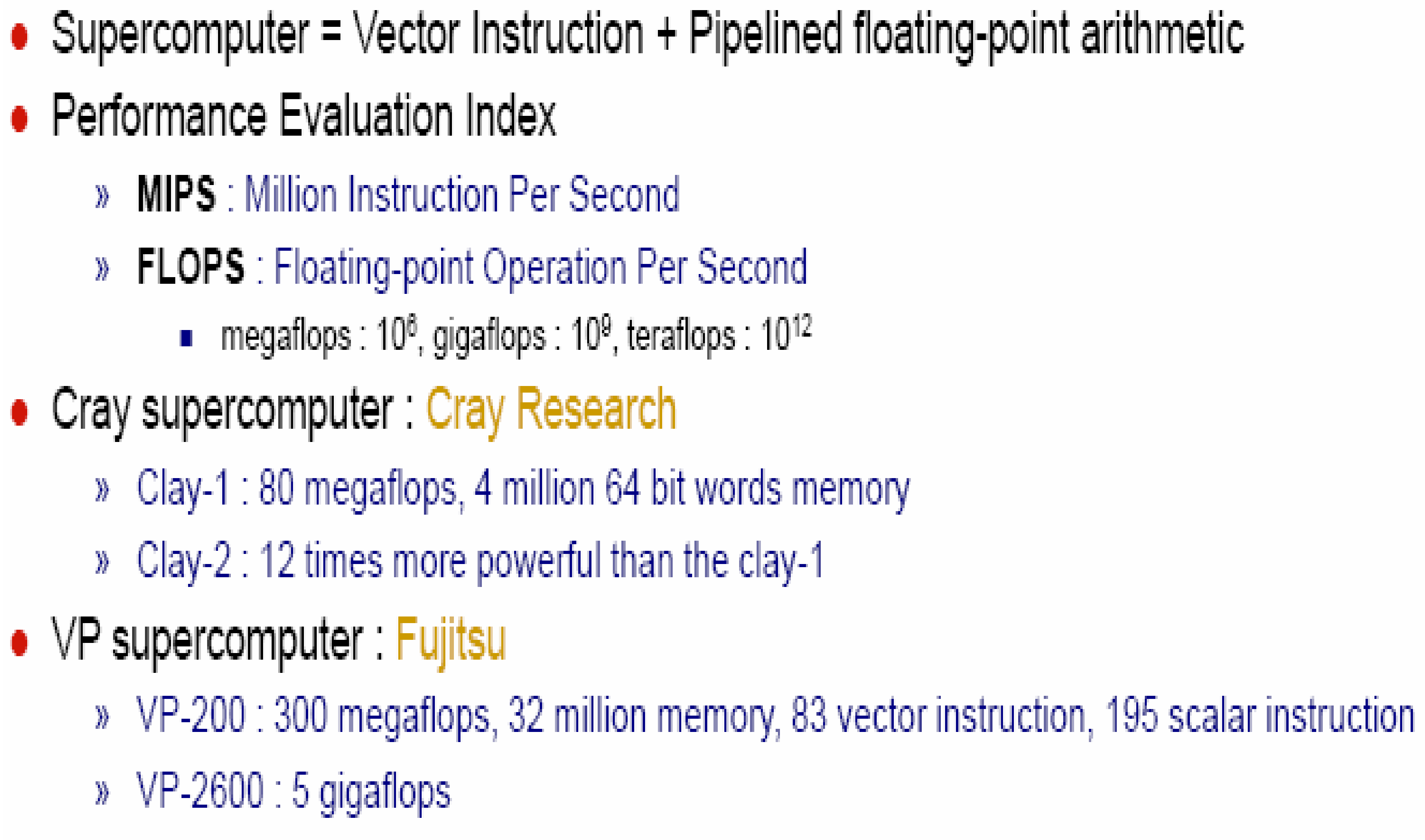
$$
\begin{aligned}
C = {} & \boxed{A_1B_1 + A_5B_5} + A_9B_9 + A_{13}B_{13} + \cdots \\
& + \boxed{A_2B_2 + A_6B_6} + A_{10}B_{10} + A_{14}B_{14} + \cdots \\
& + A_3B_3 + A_7B_7 + A_{11}B_{11} + A_{15}B_{15} + \cdots \\
& + A_4B_4 + A_8B_8 + A_{12}B_{12} + A_{16}B_{16} + \cdots
\end{aligned}
$$

Memory Interleaving:

- Pipeline and vector processors often require *simultaneous* access to memory from two or more source using *one memory bus system*



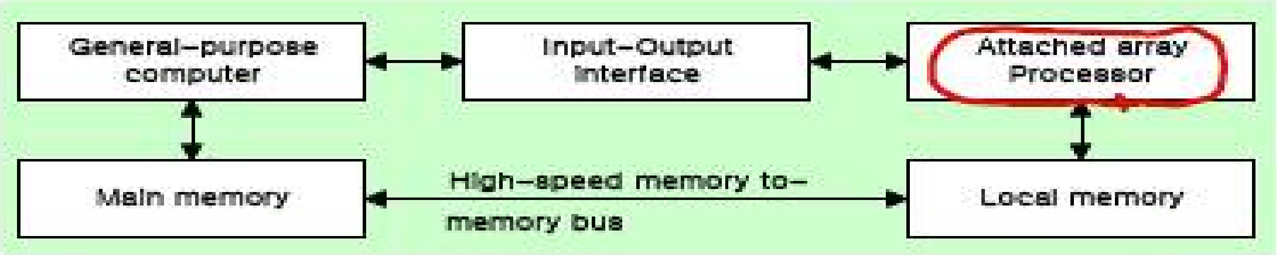
Supercomputers:

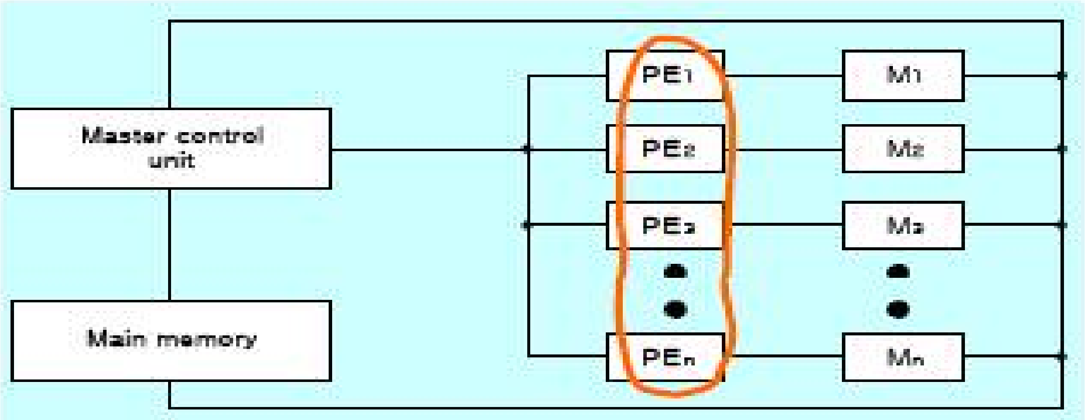


- Supercomputer = Vector Instruction + Pipelined floating-point arithmetic
- Performance Evaluation Index
  - » **MIPS** : Million Instruction Per Second
  - » **FLOPS** : Floating-point Operation Per Second
    - megaflops : 10^6, gigaflops : 10^9, teraflops : 10^{12}
- Cray supercomputer : Cray Research
  - » Clay-1 : 80 megaflops, 4 million 64 bit words memory
  - » Clay-2 : 12 times more powerful than the clay-1
- VP supercomputer : Fujitsu
  - » VP-200 : 300 megaflops, 32 million memory, 83 vector instruction, 195 scalar instruction
  - » VP-2600 : 5 gigaflops

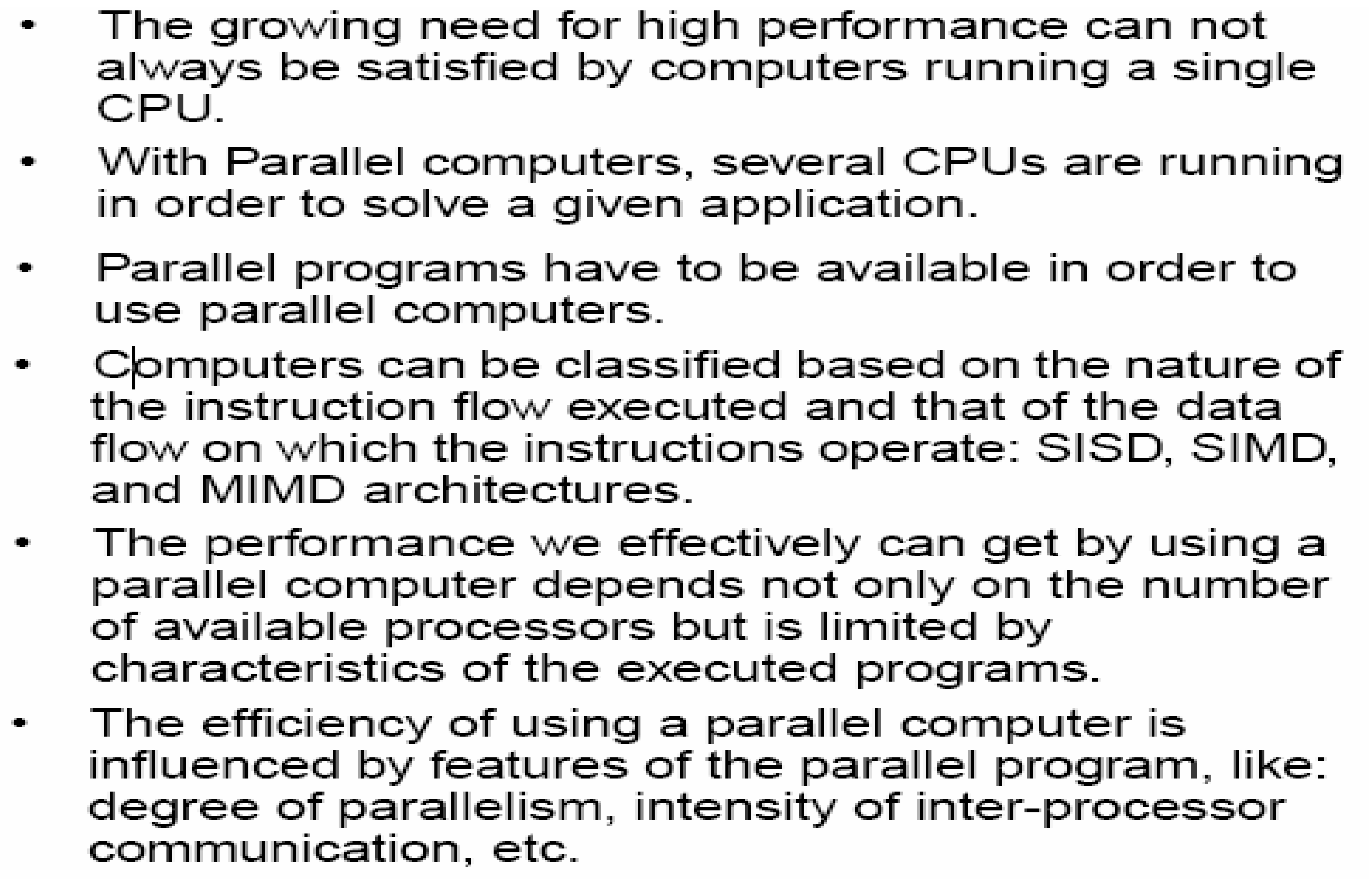
Array Processors:

- Performs computations on large arrays of data
- Array Processing
  - Attached array processor :
    - » Auxiliary processor attached to a general purpose computer
  - SIMD array processor :
    - » Computer with multiple processing units operating in parallel

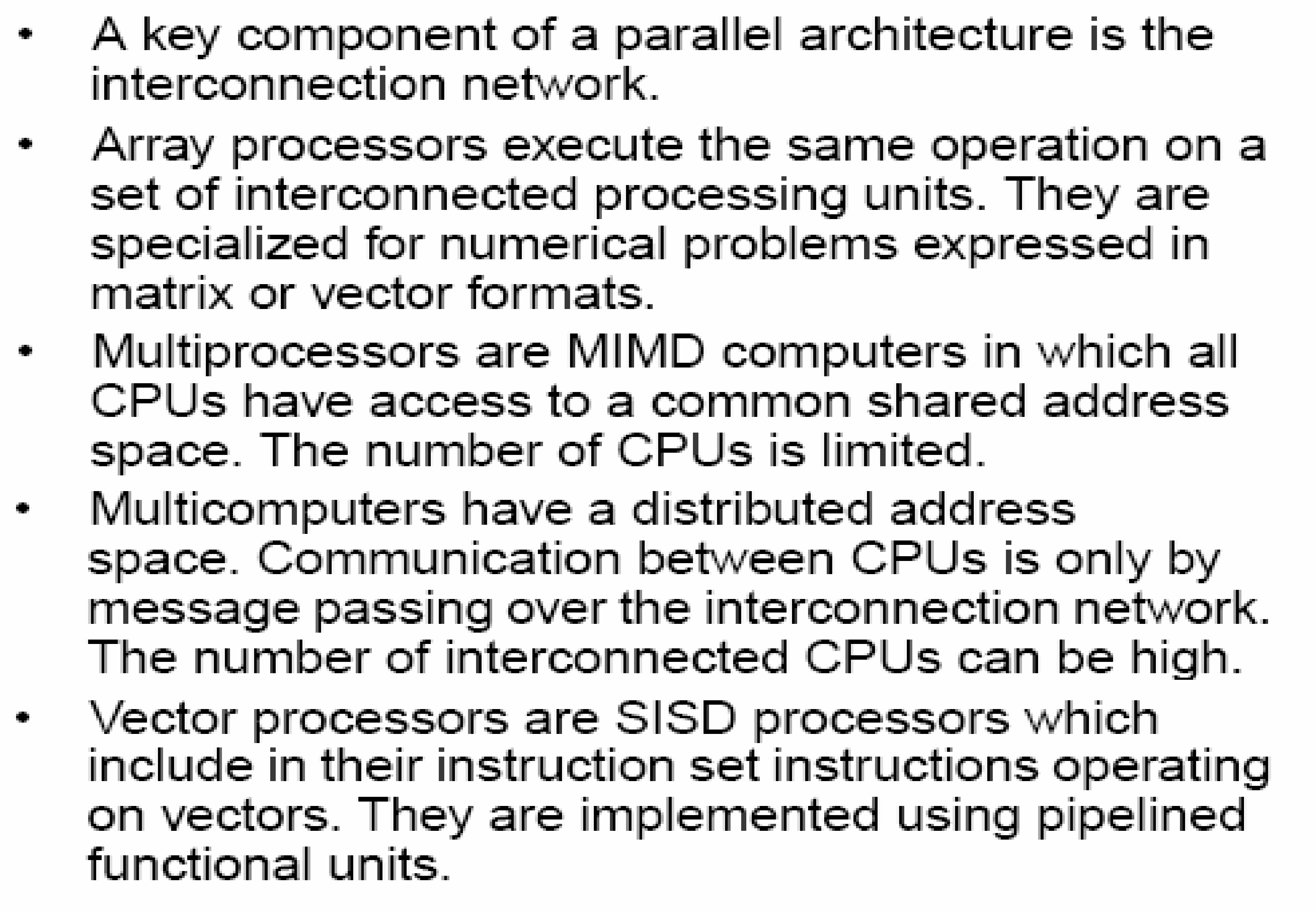




Summary:



- The growing need for high performance can not always be satisfied by computers running a single CPU.
- With Parallel computers, several CPUs are running in order to solve a given application.
- Parallel programs have to be available in order to use parallel computers.
- Computers can be classified based on the nature of the instruction flow executed and that of the data flow on which the instructions operate: SISD, SIMD, and MIMD architectures.
- The performance we effectively can get by using a parallel computer depends not only on the number of available processors but is limited by characteristics of the executed programs.
- The efficiency of using a parallel computer is influenced by features of the parallel program, like: degree of parallelism, intensity of inter-processor communication, etc.



- A key component of a parallel architecture is the interconnection network.
- Array processors execute the same operation on a set of interconnected processing units. They are specialized for numerical problems expressed in matrix or vector formats.
- Multiprocessors are MIMD computers in which all CPUs have access to a common shared address space. The number of CPUs is limited.
- Multicomputers have a distributed address space. Communication between CPUs is only by message passing over the interconnection network. The number of interconnected CPUs can be high.
- Vector processors are SISD processors which include in their instruction set instructions operating on vectors. They are implemented using pipelined functional units.