BSC~Microsoft Research Centre

High Performance Computing for Artificial Intelligence

Dr. Tassadaq Hussain

Assistant Professor Riphah International University

Collaborations:

Microsoft Research and Barcelona Supercomputing Center

Barcelona, Spain

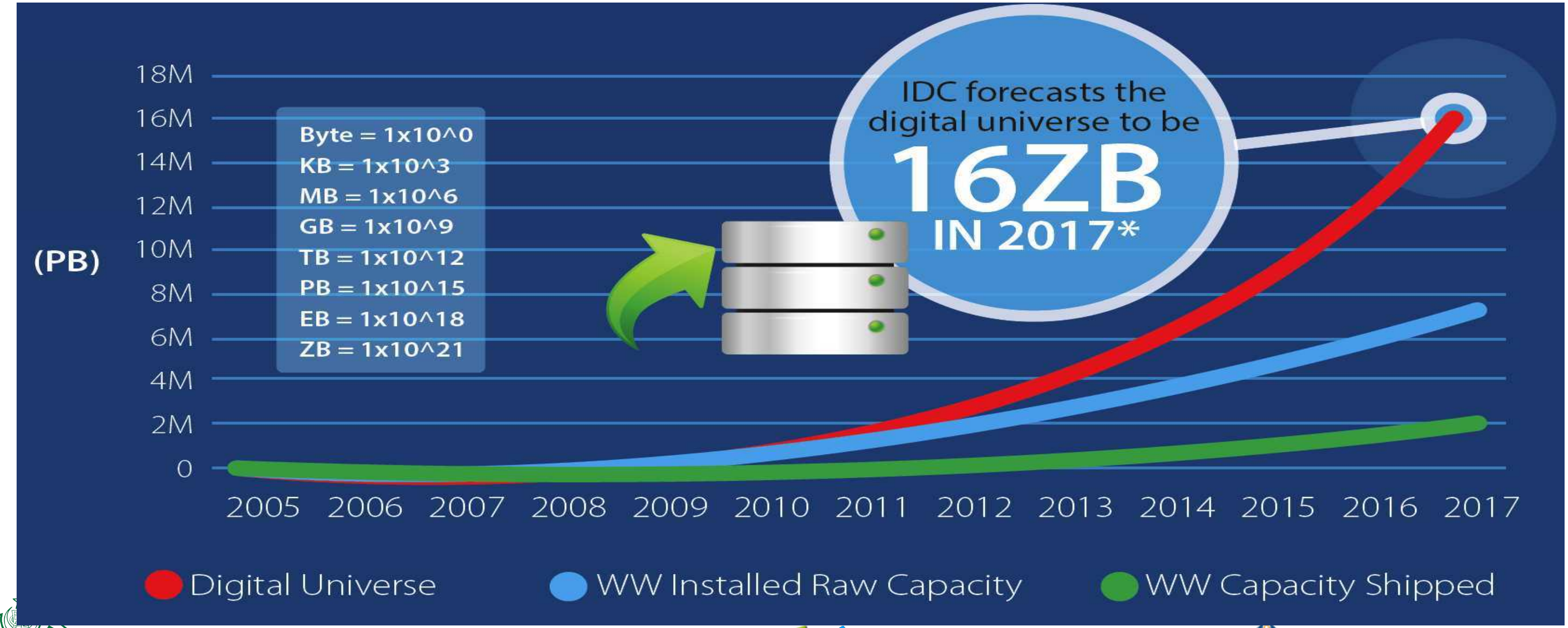
UCERD Pvt Ltd Islamabad

Unal Center of Education Research and Development
www.ucerd.com

- **HPC: Past Present and Future**
- HPC System
- HPC Applications

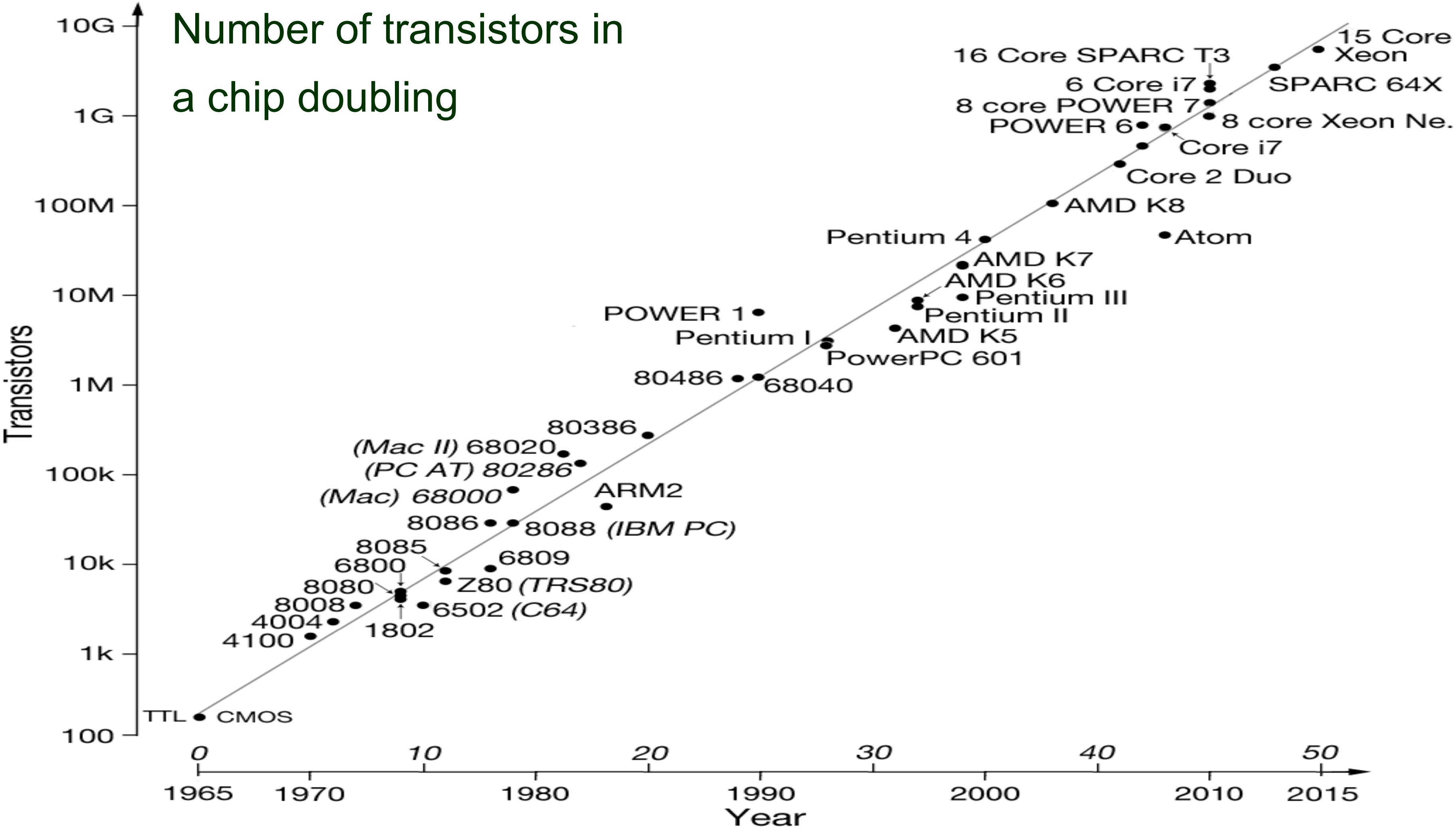
Information Future Trend

- Information Age
- Information doubling after every 18 months



Technology Research Gartner Inc. states that information data volume doubles after every 18 months.

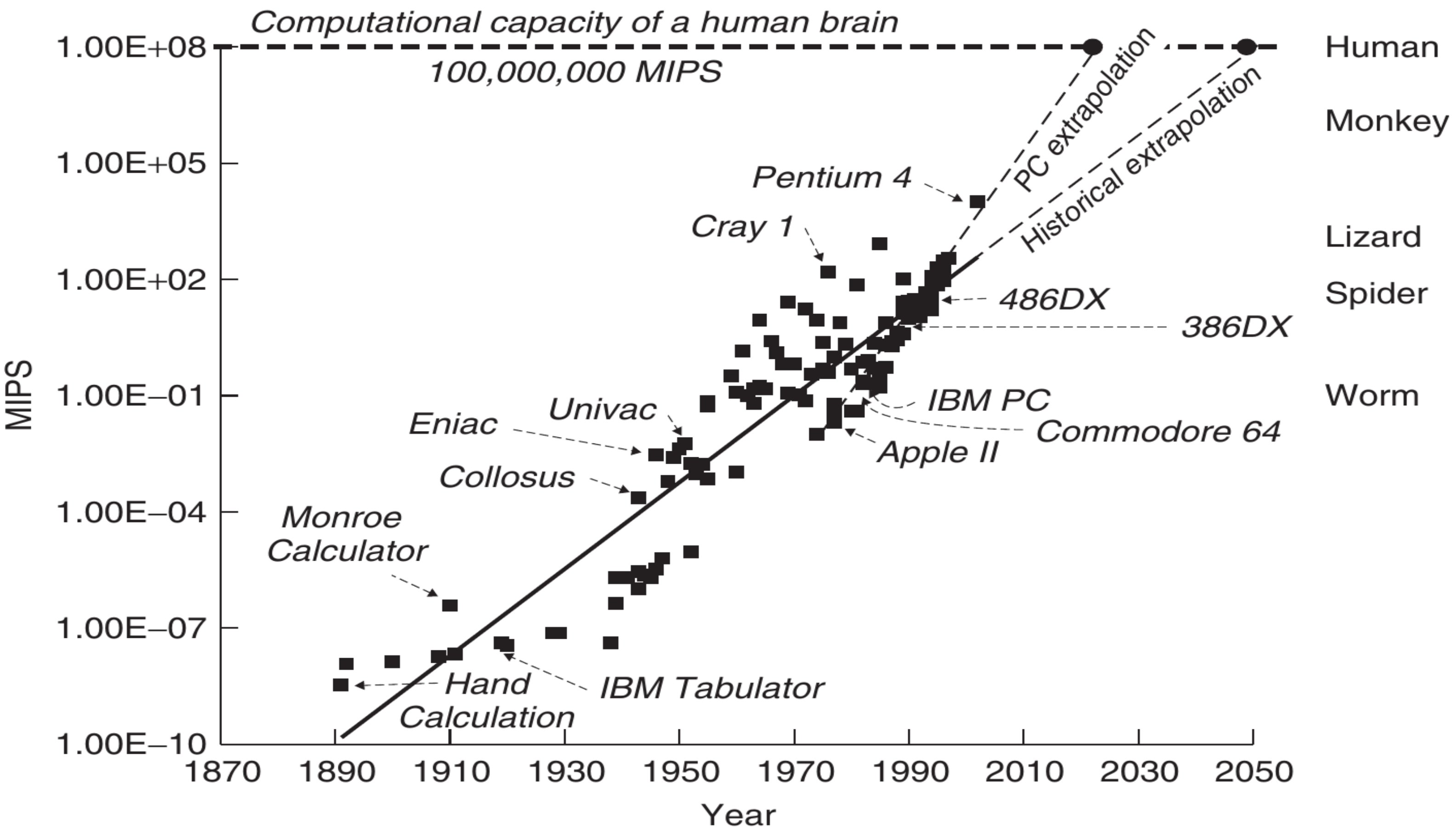
Moore’s Law: Transistor Count



Number of transistors in a chip doubling

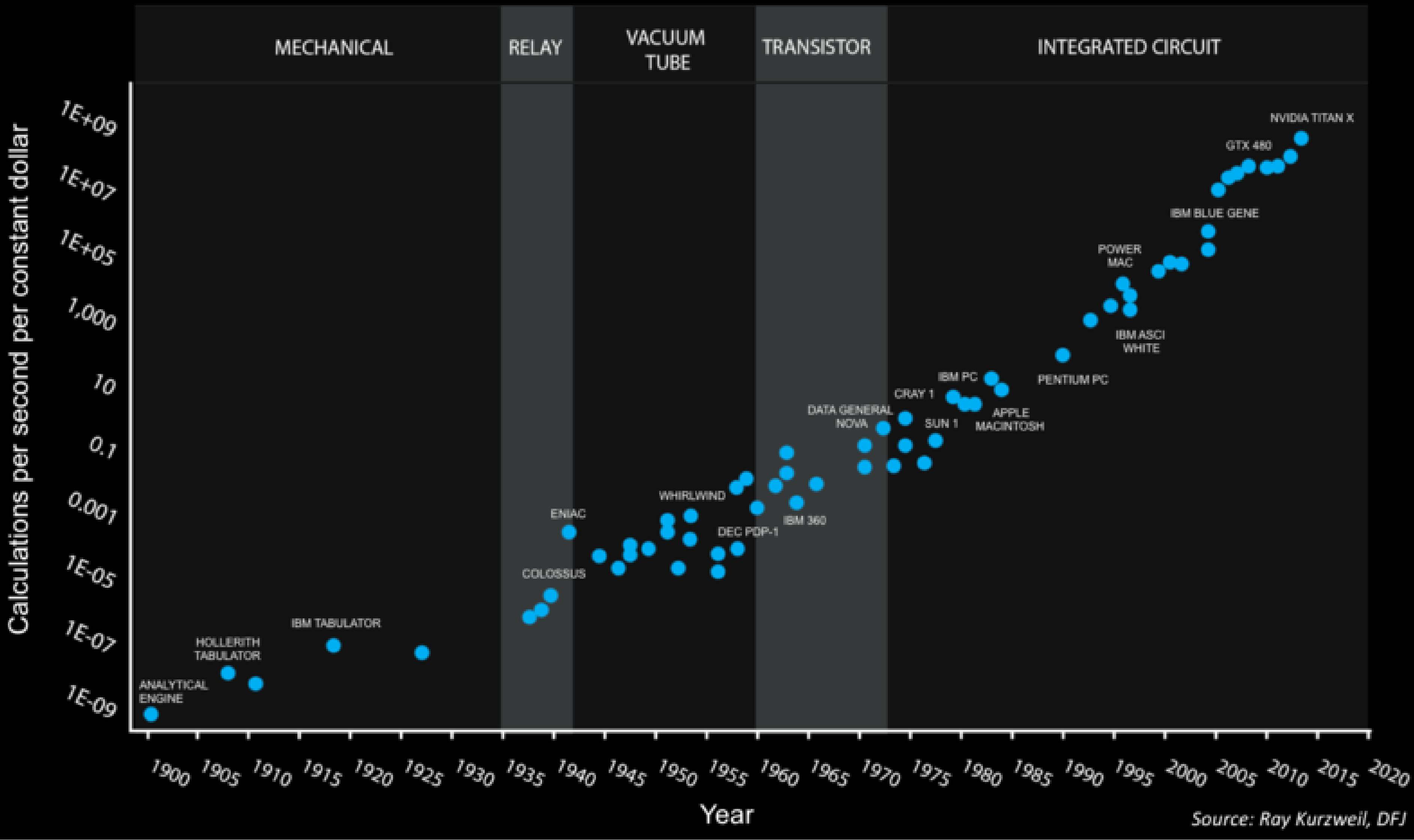
[1] G. E. Moore, “Cramming more components onto integrated circuits,” Electronics, vol. 38, no. 8, April 1965.

Performance Improvement



It is estimated that sometime between the years **2025 and 2050**, a **personal computers** will exceed the calculation power of a human brain.

120 Years of Moore's Law

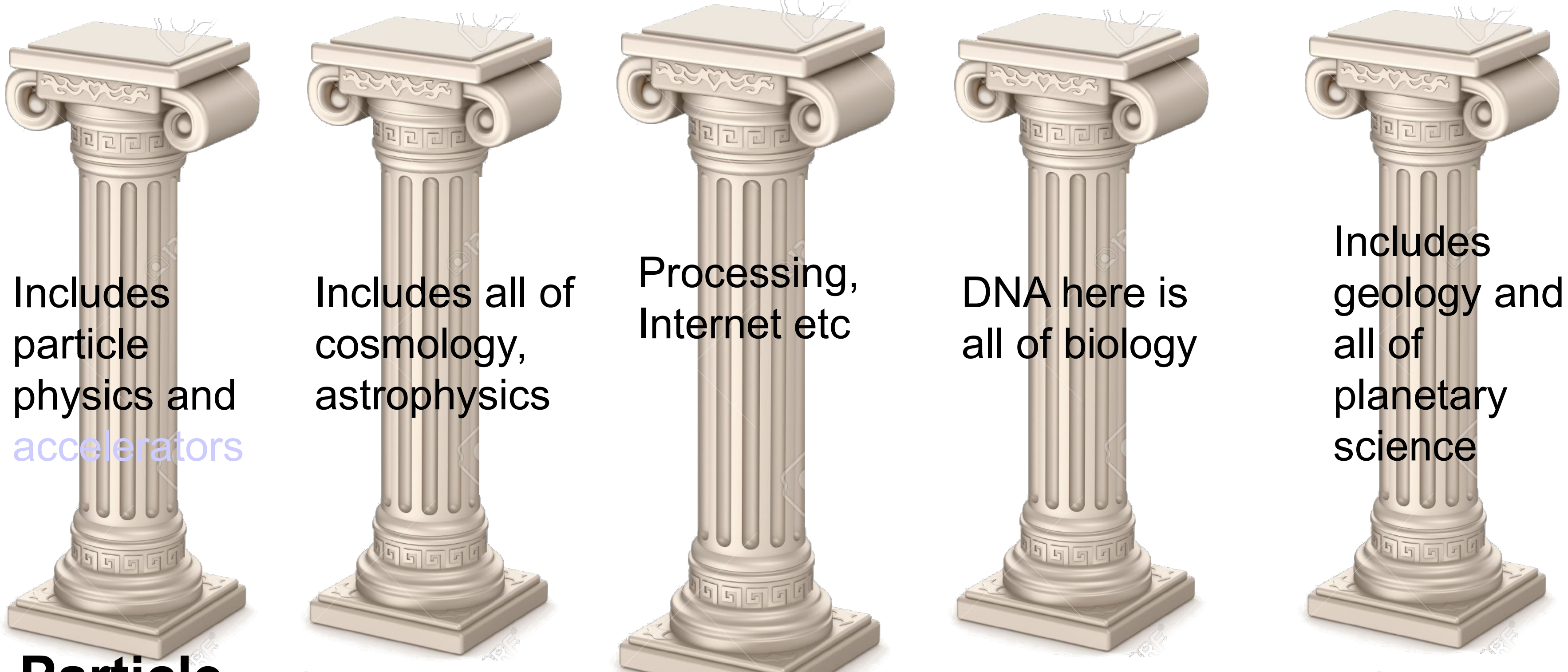


Source: Ray Kurzweil, DFJ

Pillars of Science

Fermi National
Accelerator Laboratory

Science



Includes particle physics and accelerators

Includes all of cosmology, astrophysics

Processing, Internet etc

DNA here is all of biology

Includes geology and all of planetary science

Particle Physics

Cosmology

Computing

Biology

Space

QUARKS

BIG BANG

Cloud Computing

DNA

SPACE

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

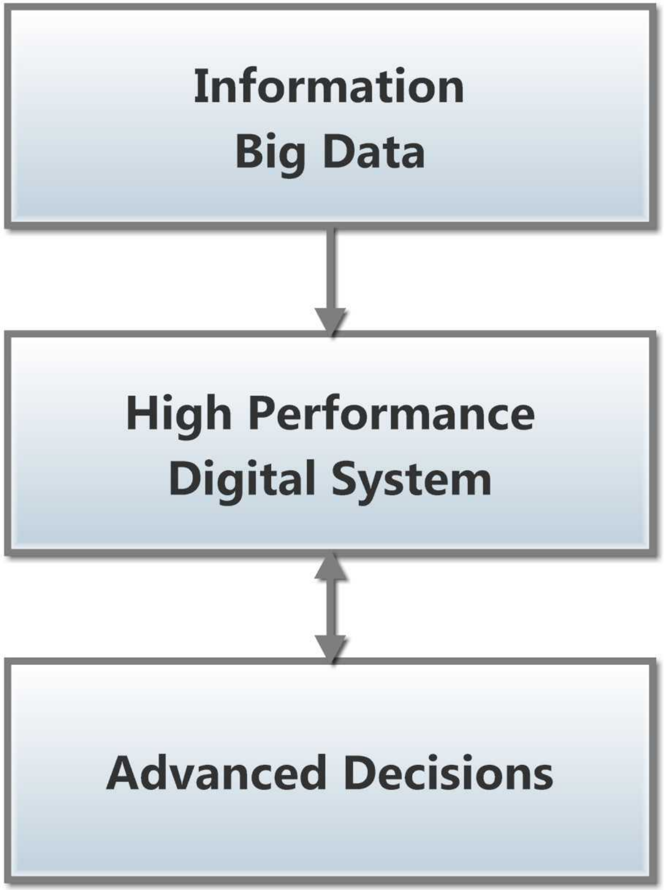
Université de Valenciennes et du Hainaut-Cambrésis

Education Research and Development
www.ucerd.com

RIPHAH INTERNATIONAL UNIVERSITY

Importance of HPC System

- The information data volume doubles after every 18 months.
- The performance of digital system get improved after every 18 months.
- High Performance, Low Cost and Low Power Computer Systems.



Relevance to industry and academia

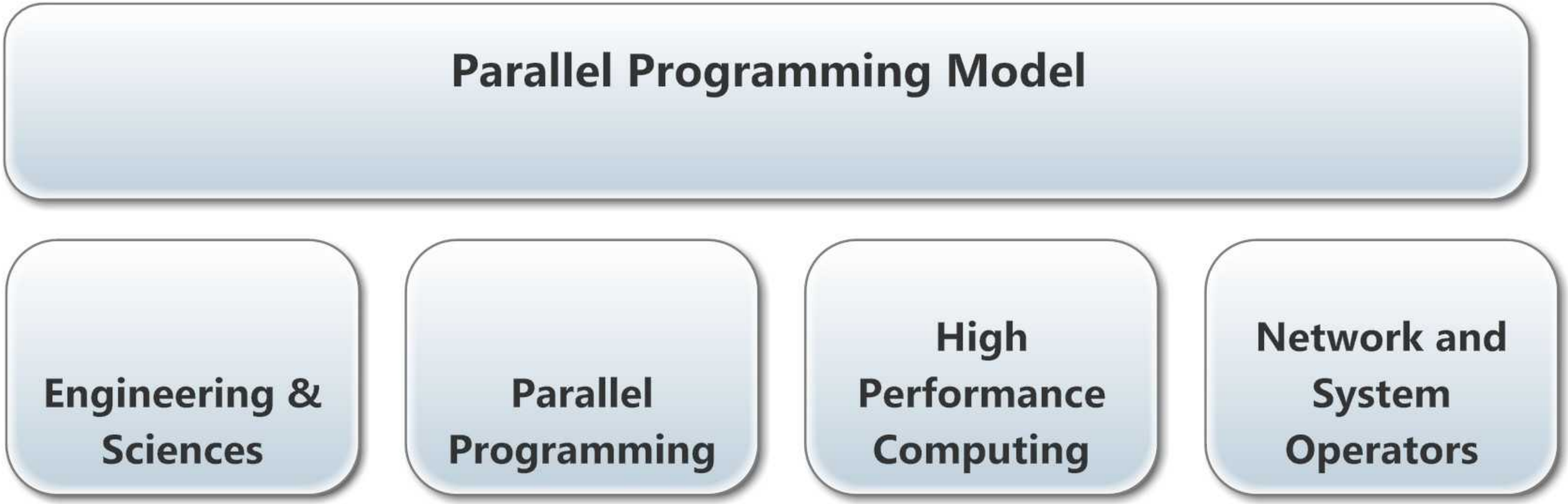
- High performance computing is the need of the day.
- Equally important for high tech industry.
- Optimum resource utilization.
- Less time to solve complex compute intensive problems.

Challenges

- Restrictions on High Performance Target Technologies.
- Limited availability of High Performance Advance hardware.
- Even having high performance hardware does not guaranty its optimum usage.
- High end expertise are required to utilize high performance hardware/software.

- HPC: Past Present and Future
- **HPC System**
- HPC Applications

Generic Proposal: High Performance Supercomputing



Parallel Programming Model

Engineering & Sciences

Parallel Programming

High Performance Computing

Network and System Operators

High performance system architectures for Artificial Intelligence, Embedded Real-time Systems etc.

PAKISTAN ENGINEERING COUNCIL

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Université de Valenciennes et du Hainaut-Cambrésis

Unal Center of Education Research and Development
www.ucerd.com

RIPHAH INTERNATIONAL UNIVERSITY

Objectives

- Executes Engineering and Sciences Applications
  - Compute and Data Intensive
  - Complex and irregular data structures
  - e.g. Artificial Intelligence, Fluid Dynamics, Structural Analysis, 3D/4D Imaging.
- Handle Information in Big Data
  - Support local memory, main memory and external memory systems
  - Perform memory read/write operations in parallel with processing unit
- Multiple Heterogeneous Cores
  - RISC (SSP), vector processor (VP) and application specific hardware accelerator (ASHA)
- Provides Programming support
  - Provide standard C/C++ parallel programming languages for real-time and standalone applications.
  - Support Tools for Visual Analysis, Modeling and Simulations e.g Ansys HPC.

PAKISTAN ENGINEERING COUNCIL

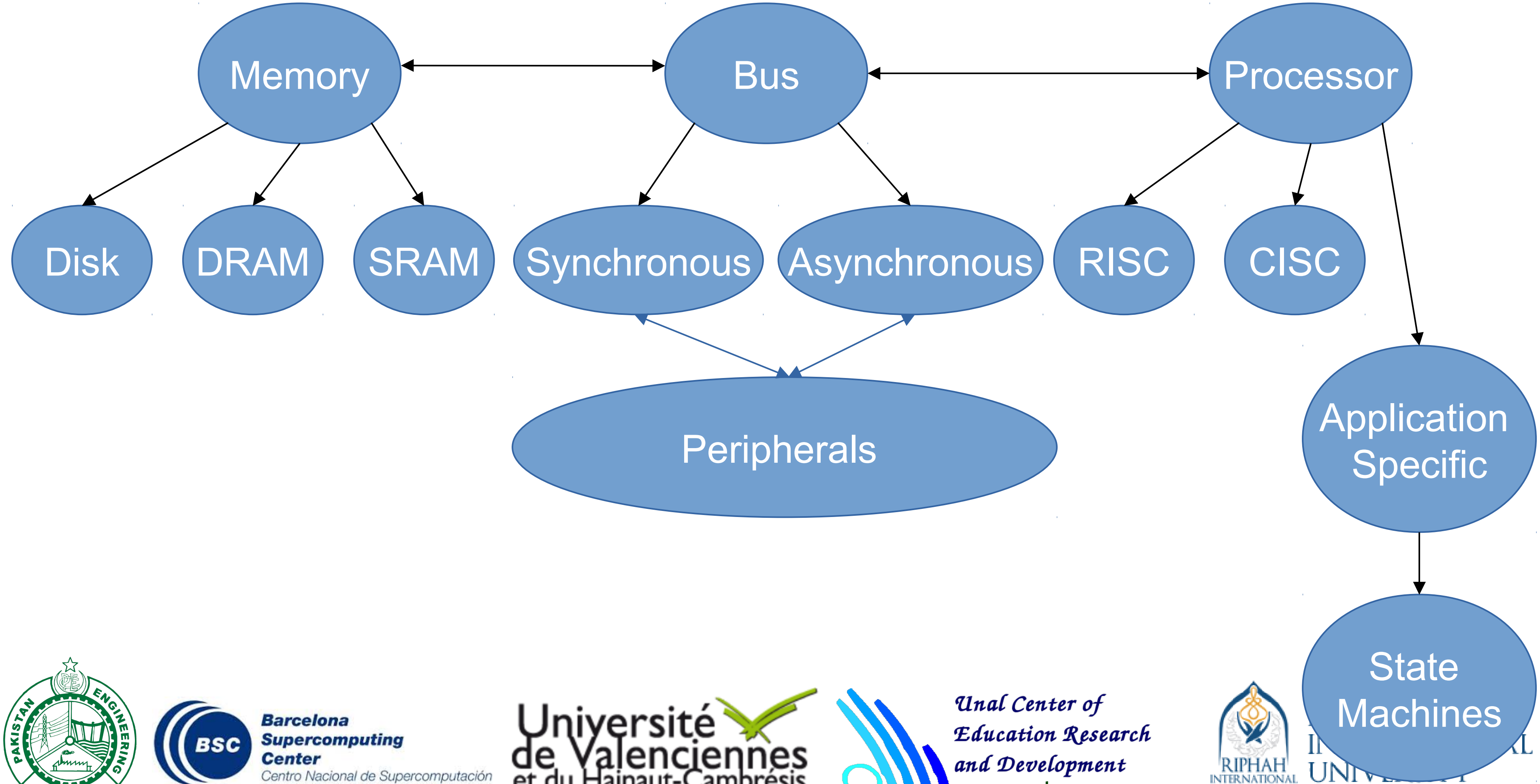
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Université de Valenciennes et du Hainaut-Cambrésis

Unal Center of Education Research and Development
www.ucerd.com

RIPHAH INTERNATIONAL UNIVERSITY

Digital System Components



Processor System Architecture

- Hardware
  - Processor
  - Bus
  - Memory
  - Peripherals

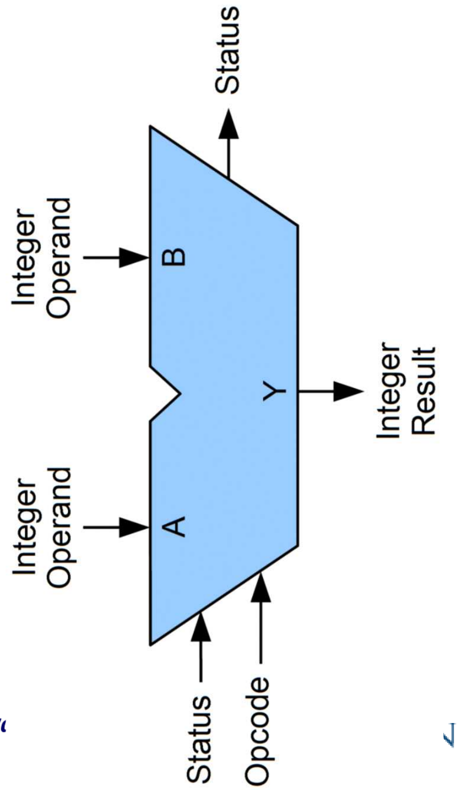
Processor

A simple processor takes a single instruction and generate results in a given time called instruction cycles.

An instruction includes two values (operands) and an arithmetic or a logic operation (operator).

Values (operands) can be from memory

or peripherals.



Important Parameters of a Processor

Clock

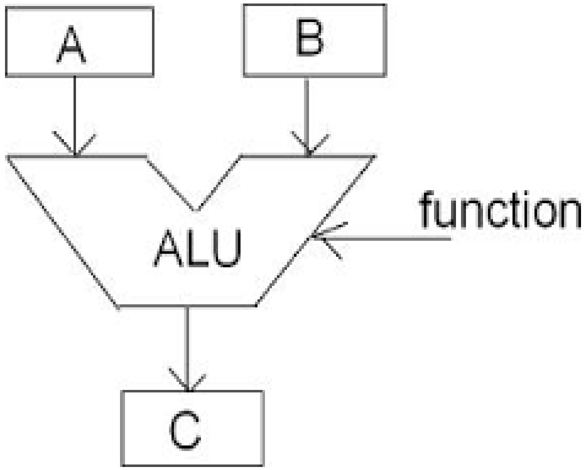
Data Bus

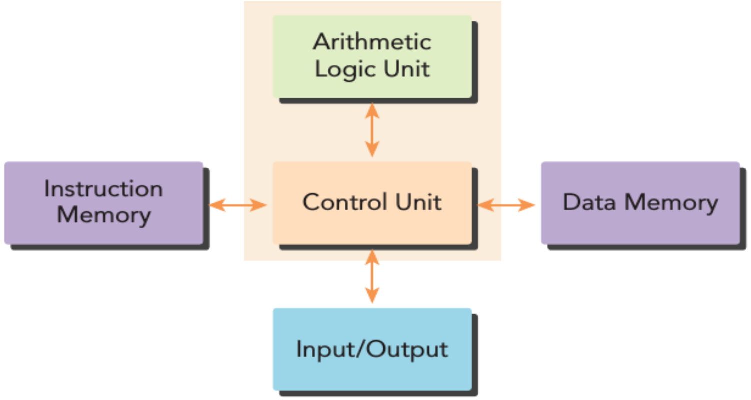
Instruction Bus

Instructions Per Cycles

Pipeline Stage

Basic introduction of Microprocessor

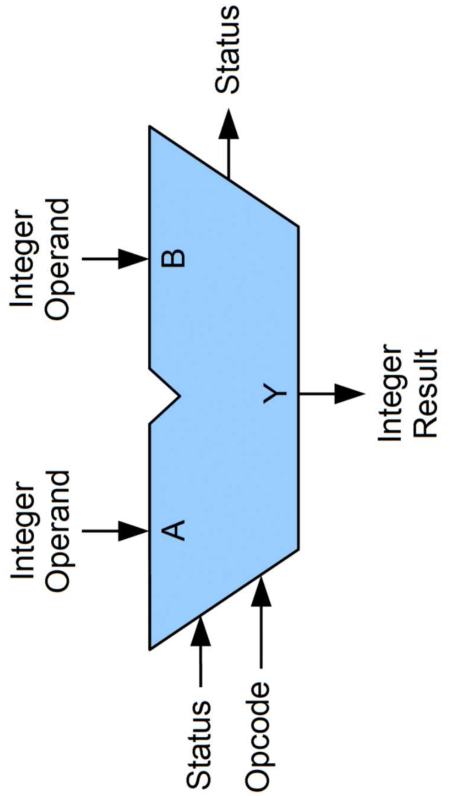


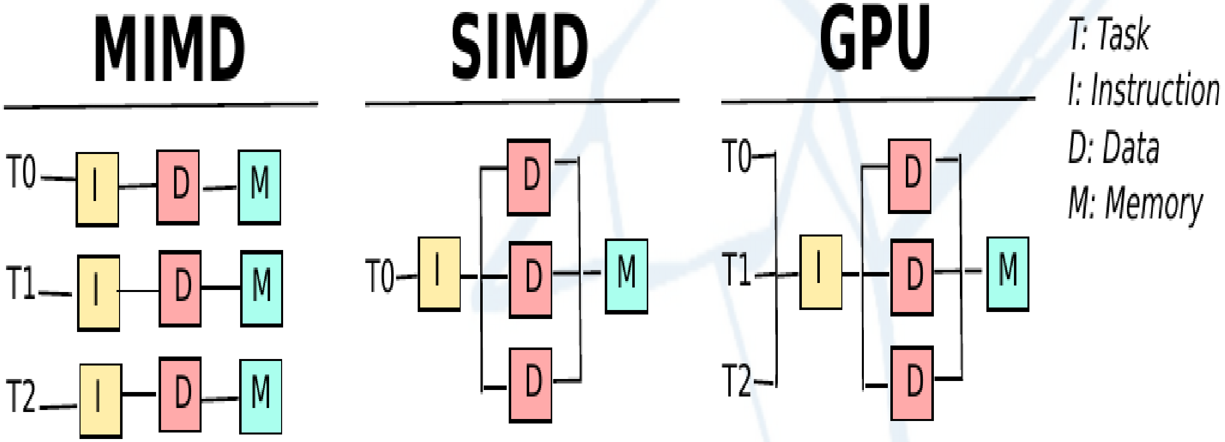


0	A	B	A+B
1	A	B	A-B

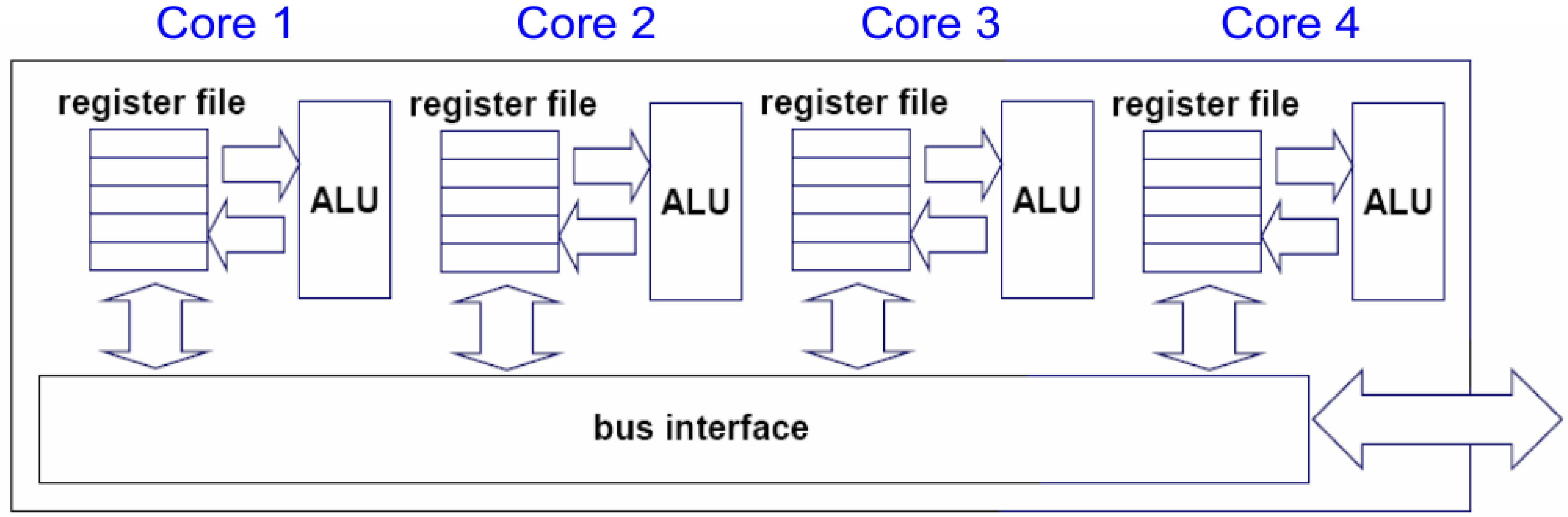
Processor Architectures

- SISD
  - RISC
- SIMD
  - CISC
- MISD
- MIMD
  - Multi-core





Multi-core Processor



Types of MIMD Architecture

Centralized shared-memory architectures or symmetric shared-memory multiprocrssors (SMP) or uniform memory access (UMA) architectures.

Distributed-memory multiprocessors.

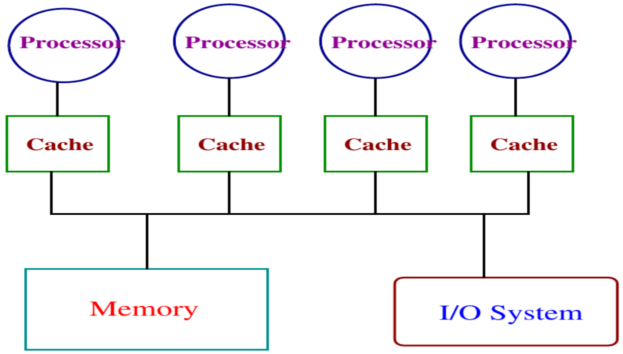
SMP: Shared Memory Processor

Small number of similar processors (at most a few dozen).

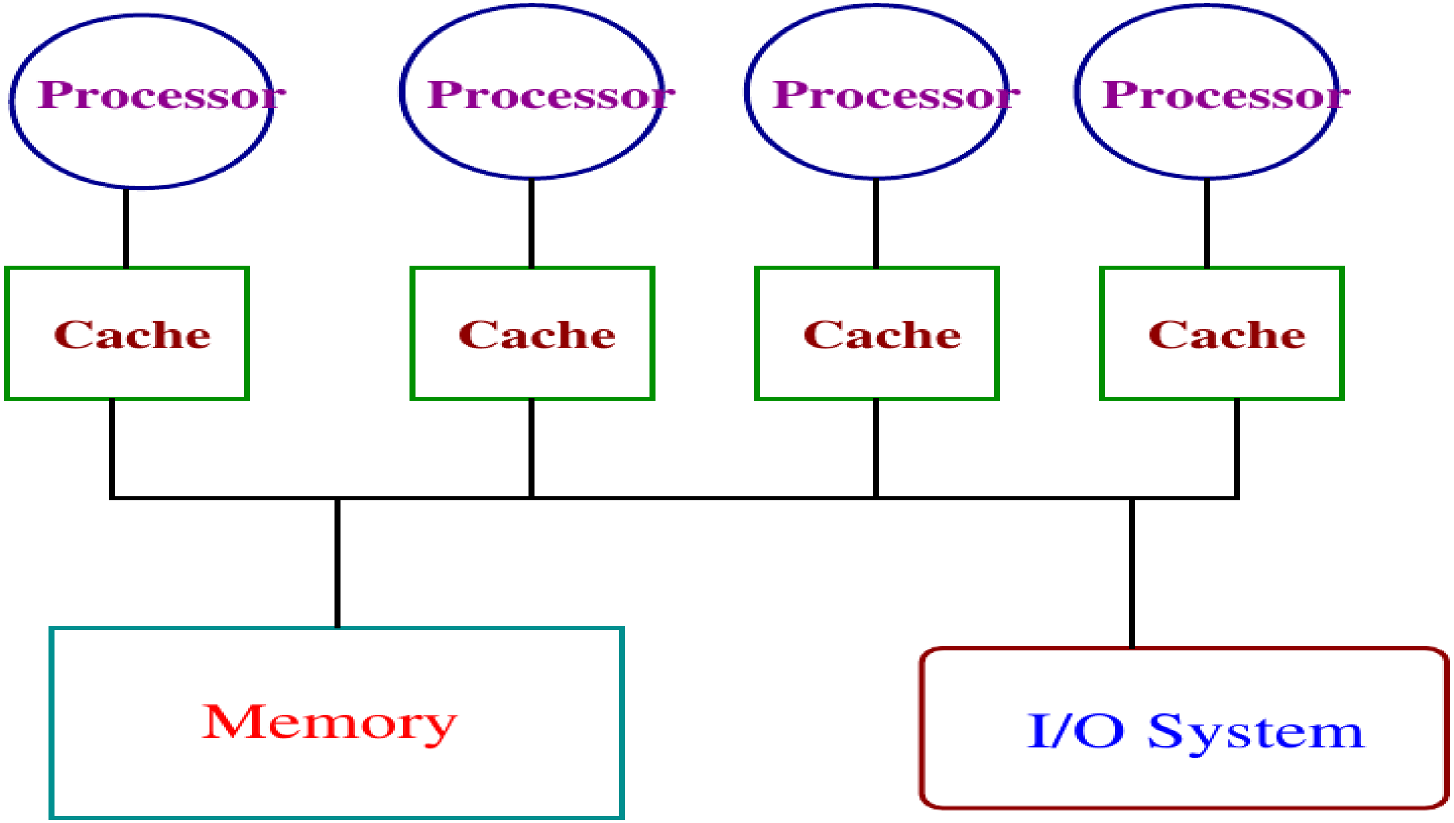
Each processor has a large cache.

A centralized memory (multiple banks) is shared through a memory bus.

Each memory location has identical access time from each processor.



SMP


Processor
Processor
Processor
Processor
Cache
Cache
Cache
Cache
Memory
I/O System

Distributed Memory

Larger processor count.

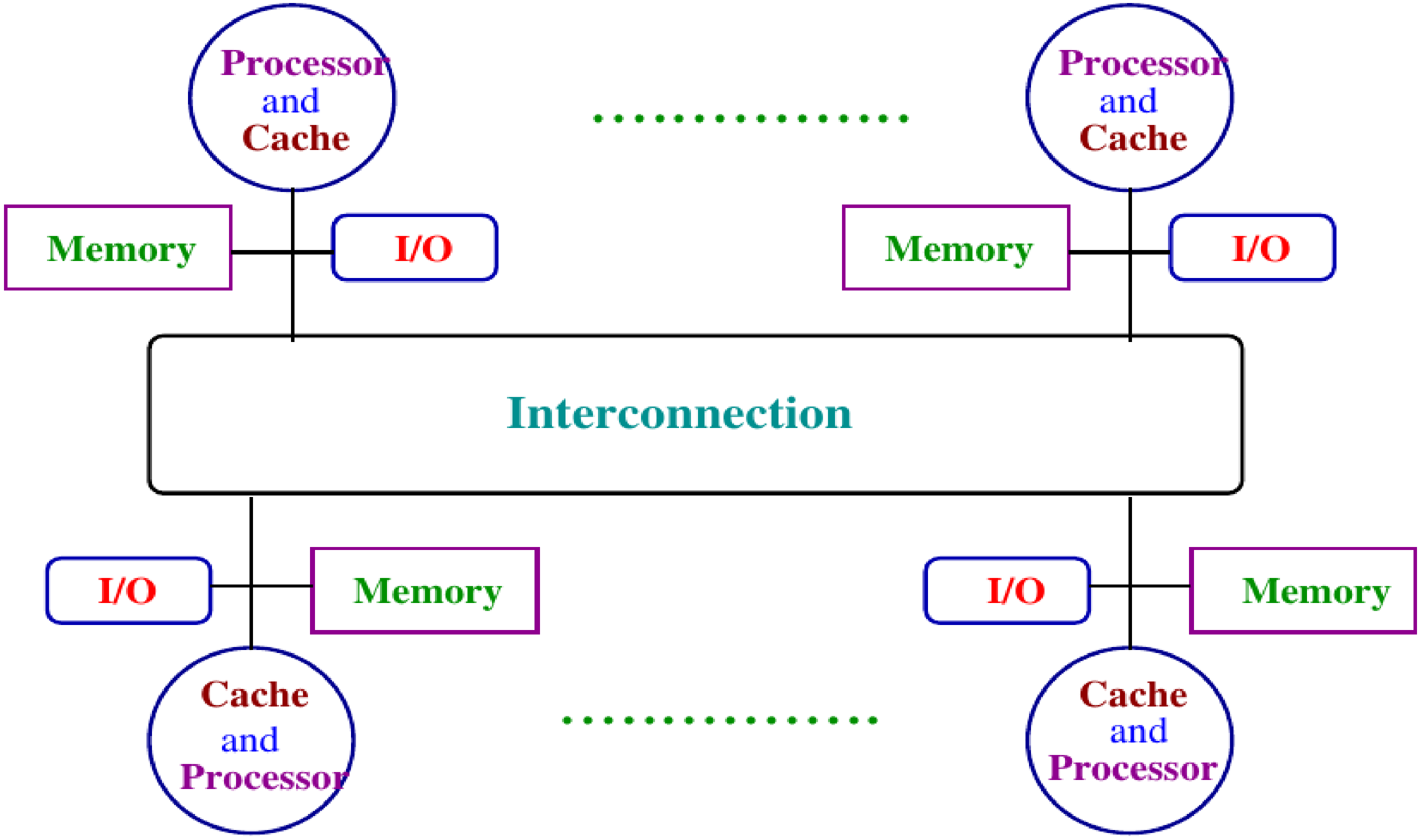
Memory is physically distributed among the processors for better bandwidth.

Connected through high-speed interconnection e.g. switches.

Distributed Memory

The bandwidth for the local memory is high and the latency is low.

But access to data present in the local memory of some other processor is complex and of high latency.


Processor
and
Cache
Processor
and
Cache
Memory
I/O
Memory
I/O
Interconnection
I/O
Memory
I/O
Memory
Cache
and
Processor
Cache
and
Processor

Distributed Shared Memory Architecture

Large-scale multiprocessors have **physically distributed memory** with the processors.

There are essentially **two different models of memory architectures** and the corresponding models of communication.

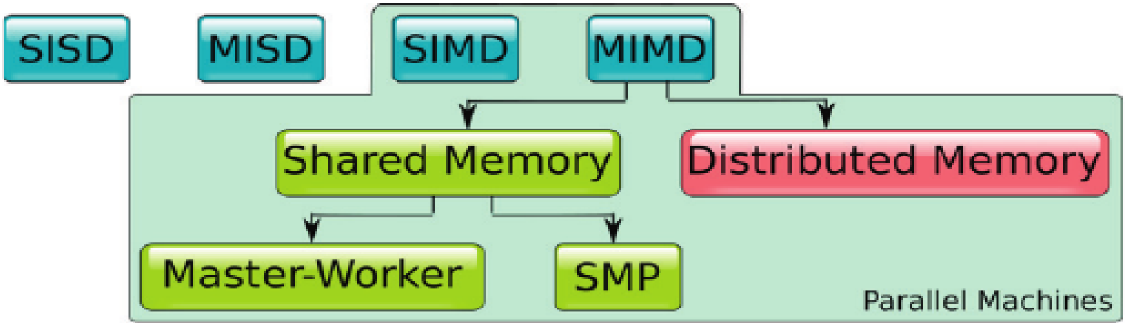
DSM Architecture

- Memory is distributed with different processors to support higher bandwidth demand of larger number of processors.
- Any processor can access a location of physically distributed memory (with proper access permission).
- This is called distributed shared-memory architecture (DSM) also known as NUMA (nonuniform memory access).

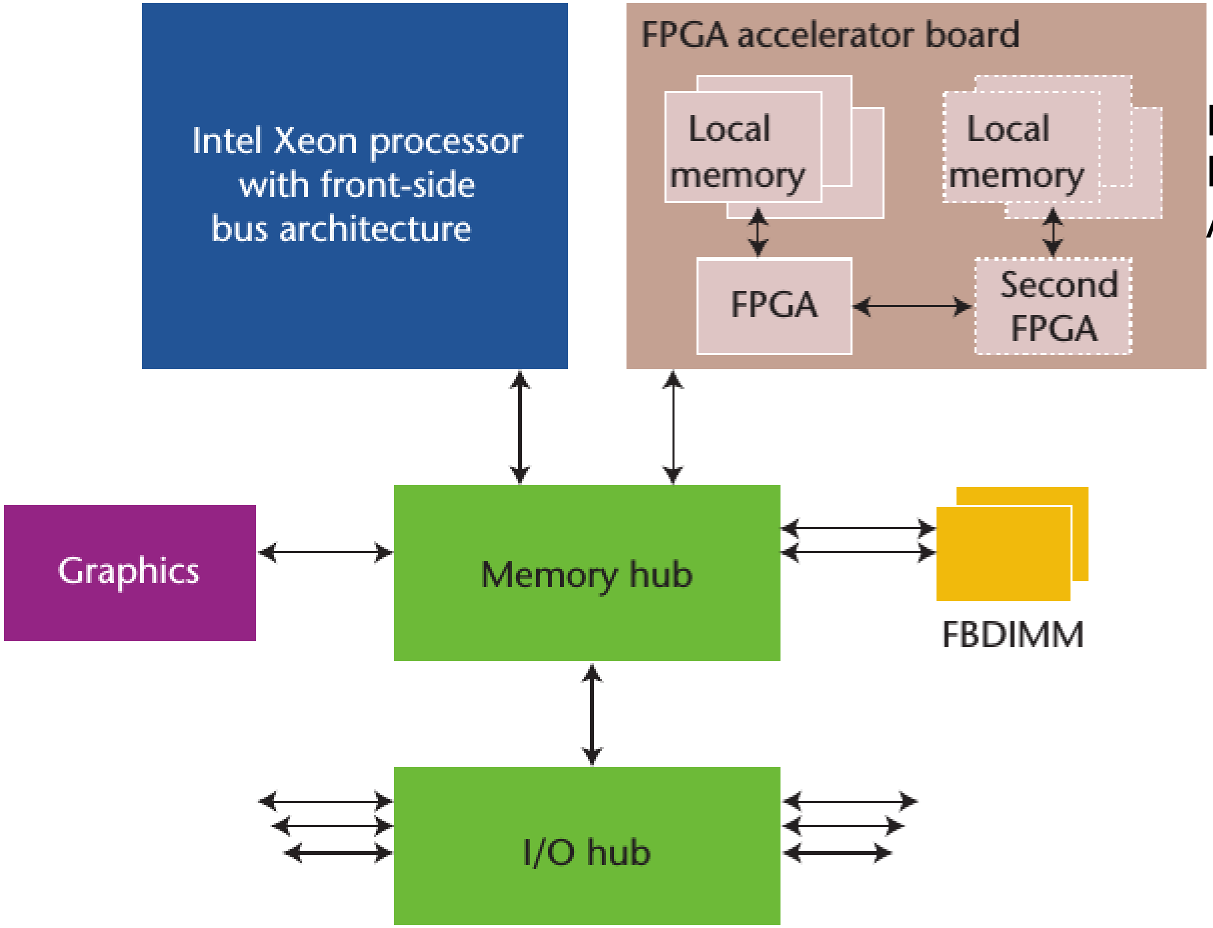
Available Computer Architectures

There are currently two trends in utilizing the increased transistor count afforded by miniaturization and advancements in semiconductor materials:

- Increase the **on-chip core count**,
  - Combined with augmented specialized SIMD instruction sets (e.g., SSE and its subsequent versions, MMX, AESNI, etc.) and larger caches.
  - This is best exemplified by Intel’s x86 line of CPUs and the Intel Xeon Phi coprocessor.
- **Combine heterogeneous cores** in the same package,
  - Typically CPU and GPU ones, each optimized for a different type of task.
  - This is best exemplified by AMD’s line of Accelerated Processing Unit (APU) chips. Intel is also offering OpenCL-based computing on its line of CPUs with integrated graphics chips.



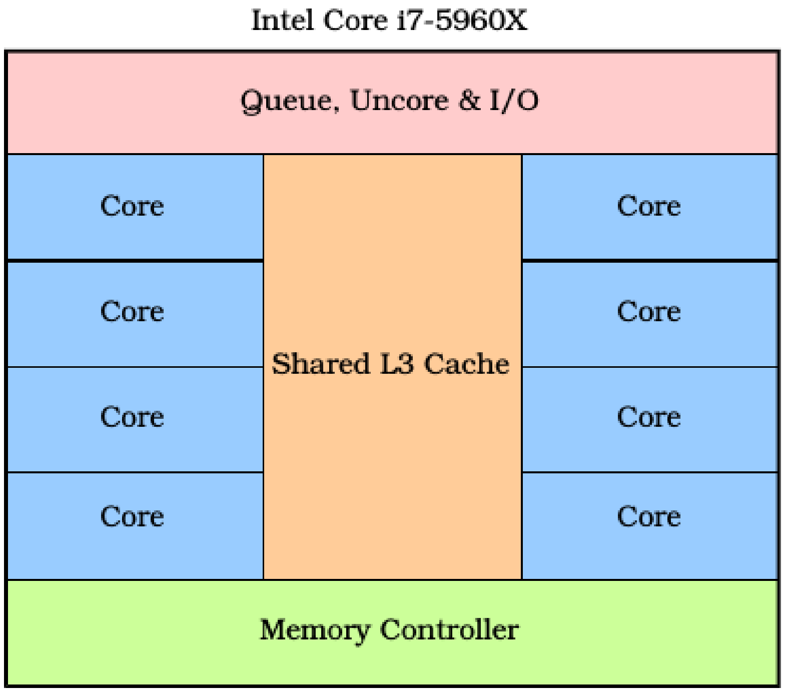
Reconfigurable Accelerators



HPC Reconfigurable Architecture

CPU: Intel Processor

CPUs employ large on-chip (and sometimes multiple) memory caches, few complex (e.g., pipelined) arithmetic and logical processing units (ALUs), and complex instruction decoding and prediction hardware to avoid stalling while waiting for data to arrive from the main memory.



Intel Core i7-5960X

Queue, Uncore & I/O

Core

Core

Core

Core

Shared L3 Cache

Core

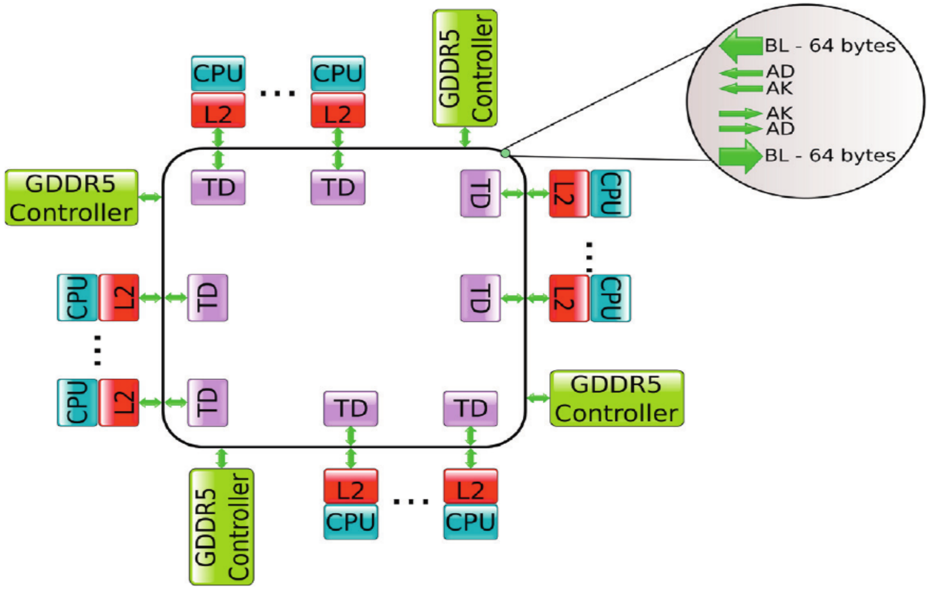
Core

Core

Core

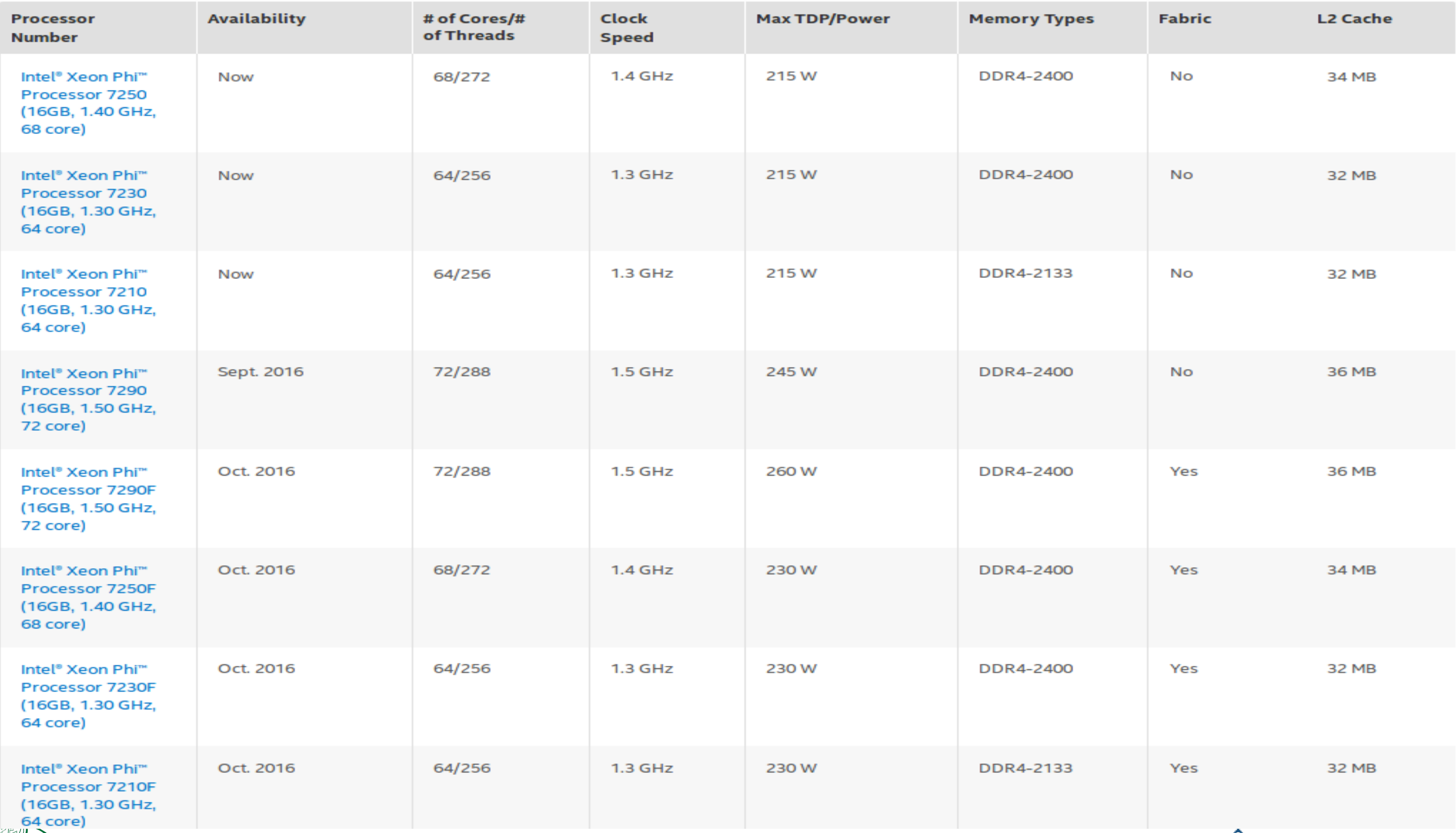
Memory Controller

Intel Xeon Phi



- A Super-scalar Architecture
- Xeon Phi comes equipped upto 72 x86 cores that are heavily customized Pentium cores.
- The customizations include the ability to handle four threads at the same time.
- The coherency is managed by distributed tag directories (TDs)

Intel Super Scalar: A Many Core Architecture



Processor Number	Availability	# of Cores/# of Threads	Clock Speed	Max TDP/Power	Memory Types	Fabric	L2 Cache
Intel® Xeon Phi™ Processor 7250 (16GB, 1.40 GHz, 68 core)	Now	68/272	1.4 GHz	215 W	DDR4-2400	No	34 MB
Intel® Xeon Phi™ Processor 7230 (16GB, 1.30 GHz, 64 core)	Now	64/256	1.3 GHz	215 W	DDR4-2400	No	32 MB
Intel® Xeon Phi™ Processor 7210 (16GB, 1.30 GHz, 64 core)	Now	64/256	1.3 GHz	215 W	DDR4-2133	No	32 MB
Intel® Xeon Phi™ Processor 7290 (16GB, 1.50 GHz, 72 core)	Sept. 2016	72/288	1.5 GHz	245 W	DDR4-2400	No	36 MB
Intel® Xeon Phi™ Processor 7290F (16GB, 1.50 GHz, 72 core)	Oct. 2016	72/288	1.5 GHz	260 W	DDR4-2400	Yes	36 MB
Intel® Xeon Phi™ Processor 7250F (16GB, 1.40 GHz, 68 core)	Oct. 2016	68/272	1.4 GHz	230 W	DDR4-2400	Yes	34 MB
Intel® Xeon Phi™ Processor 7230F (16GB, 1.30 GHz, 64 core)	Oct. 2016	64/256	1.3 GHz	230 W	DDR4-2400	Yes	32 MB
Intel® Xeon Phi™ Processor 7210F (16GB, 1.30 GHz, 64 core)	Oct. 2016	64/256	1.3 GHz	230 W	DDR4-2133	Yes	32 MB

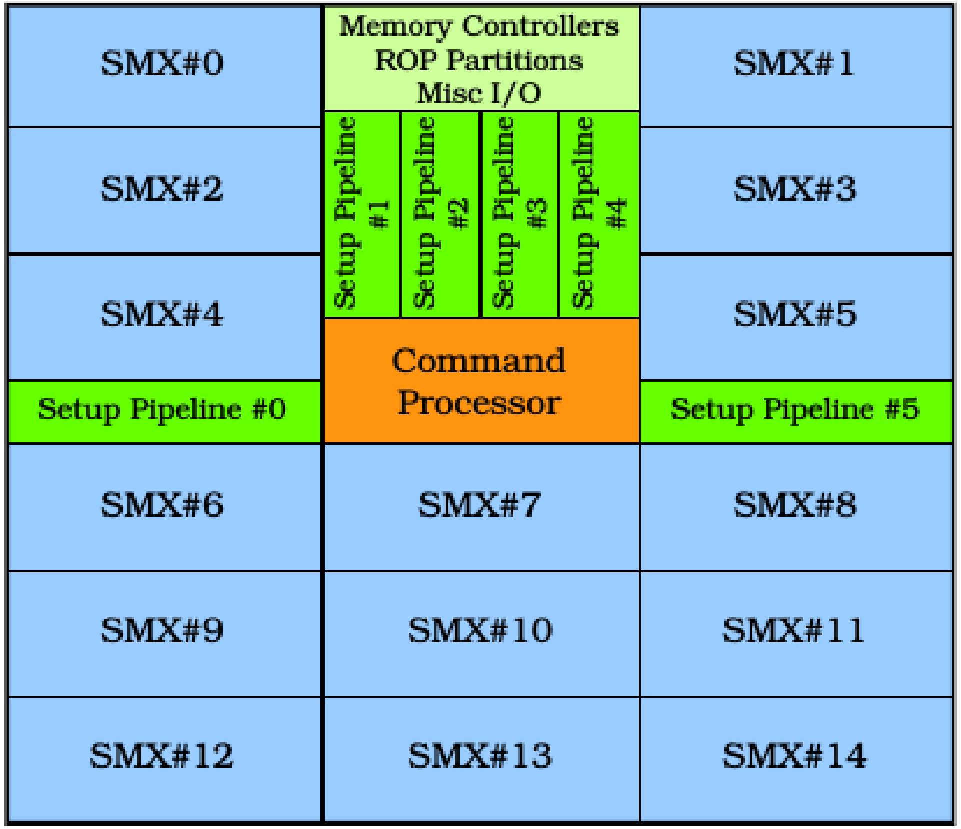
Graphics Processing Unit (GPU) and CPU

- **GPUs** have been developed as a means of processing massive amount of graphics data very quickly, before they are placed in the card's display buffer.
- Their design envelope dictated a layout that departed from the one traditionally used by conventional CPUs.
- **GPU** uses small on-chip caches with a big collection of simple ALUs capable of parallel operation, since data reuse is typically small for graphics processing and programs are relatively simple. In order to feed the multiple cores on a GPU, designers also dedicated very wide, fast memory buses for fetching data from the GPU's main memory.

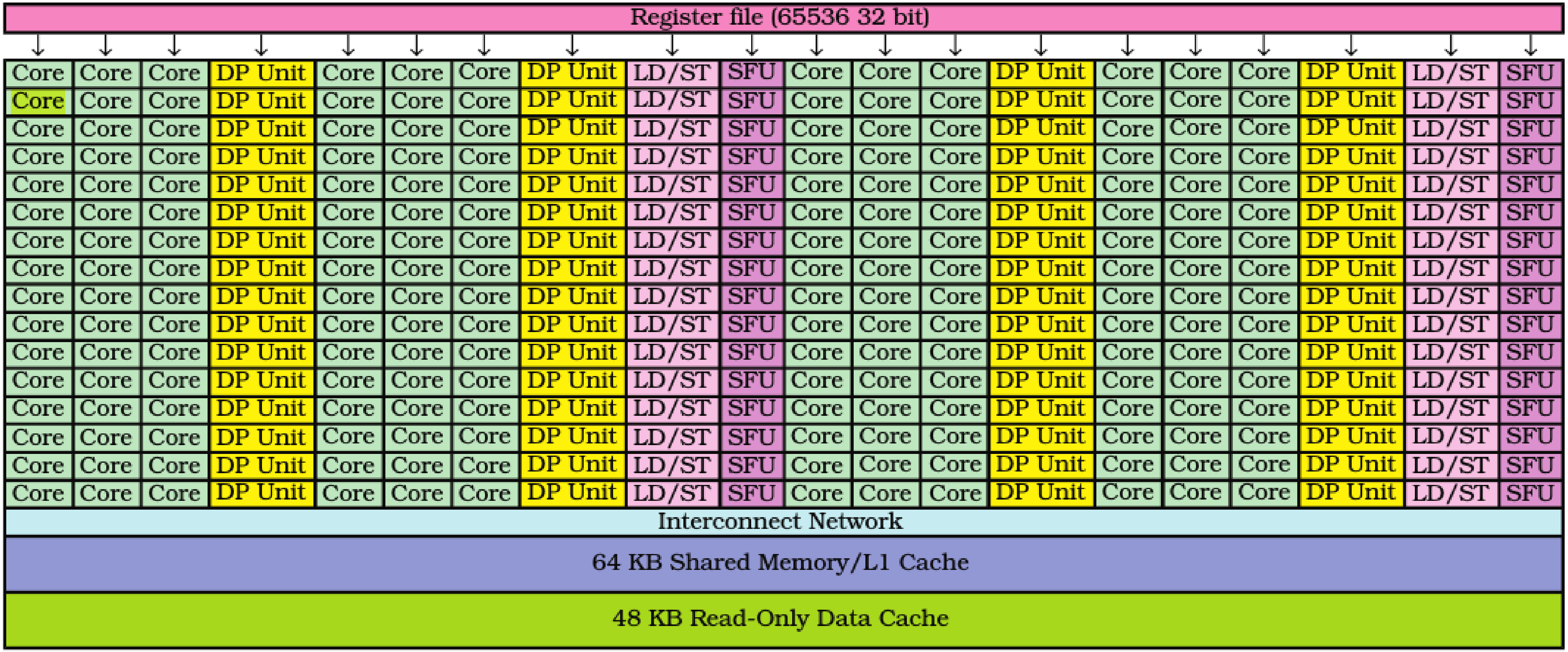
Nvidia Graphics Processing Unit (GPU)

- SM, SMM SMX (Streaming Multiprocessors): Single SMX contains 192 cores executes in SIMD fashion
- Each SMX can run its own program.
- CUDA and OpenACC Programming Models

Nvidia Kepler GK110



GPU SMX Internal Architecture



Register file (65536 32 bit)

Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU

Interconnect Network

64 KB Shared Memory/L1 Cache

48 KB Read-Only Data Cache

192 **Core** : single-precision cores

64 **DP Unit** : double -precision cores

32 **LD/ST** : load/store units

32 **SFU** : Special Function Units

PAKISTAN ENGINEERING COUNCIL

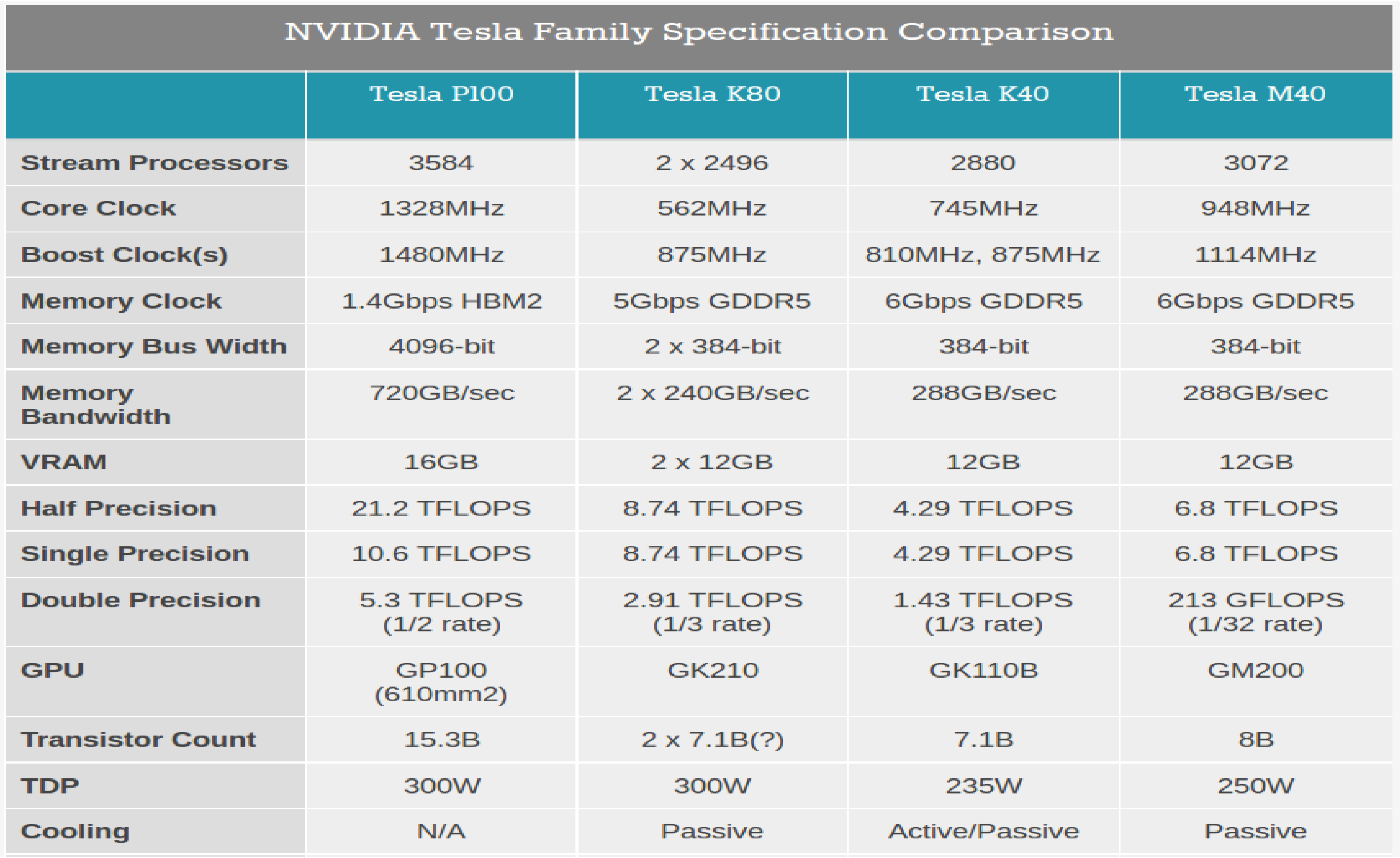
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Université de Valenciennes et du Hainaut-Cambrésis

Unal Center of Education Research and Development
www.ucerd.com

RIPHAH INTERNATIONAL UNIVERSITY

High Performance Accelerators



NVIDIA Tesla Family Specification Comparison

	Tesla P100	Tesla K80	Tesla K40	Tesla M40
Stream Processors	3584	2 x 2496	2880	3072
Core Clock	1328MHz	562MHz	745MHz	948MHz
Boost Clock(s)	1480MHz	875MHz	810MHz, 875MHz	1114MHz
Memory Clock	1.4Gbps HBM2	5Gbps GDDR5	6Gbps GDDR5	6Gbps GDDR5
Memory Bus Width	4096-bit	2 x 384-bit	384-bit	384-bit
Memory Bandwidth	720GB/sec	2 x 240GB/sec	288GB/sec	288GB/sec
VRAM	16GB	2 x 12GB	12GB	12GB
Half Precision	21.2 TFLOPS	8.74 TFLOPS	4.29 TFLOPS	6.8 TFLOPS
Single Precision	10.6 TFLOPS	8.74 TFLOPS	4.29 TFLOPS	6.8 TFLOPS
Double Precision	5.3 TFLOPS (1/2 rate)	2.91 TFLOPS (1/3 rate)	1.43 TFLOPS (1/3 rate)	213 GFLOPS (1/32 rate)
GPU	GP100 (610mm2)	GK210	GK110B	GM200
Transistor Count	15.3B	2 x 7.1B(?)	7.1B	8B
TDP	300W	300W	235W	250W
Cooling	N/A	Passive	Active/Passive	Passive

PAKISTAN ENGINEERING COUNCIL

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

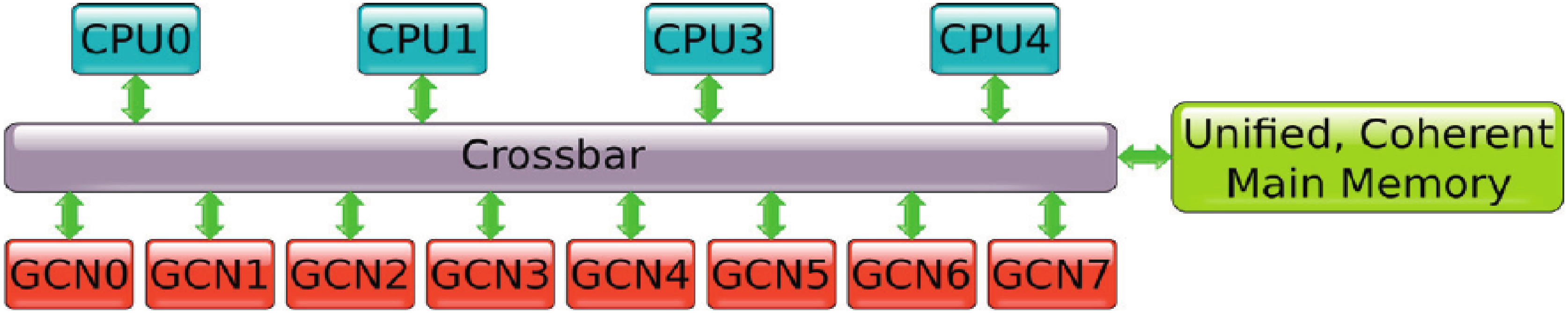
Université de Valenciennes et du Hainaut-Cambrésis

Unal Center of Education Research and Development
www.ucerd.com

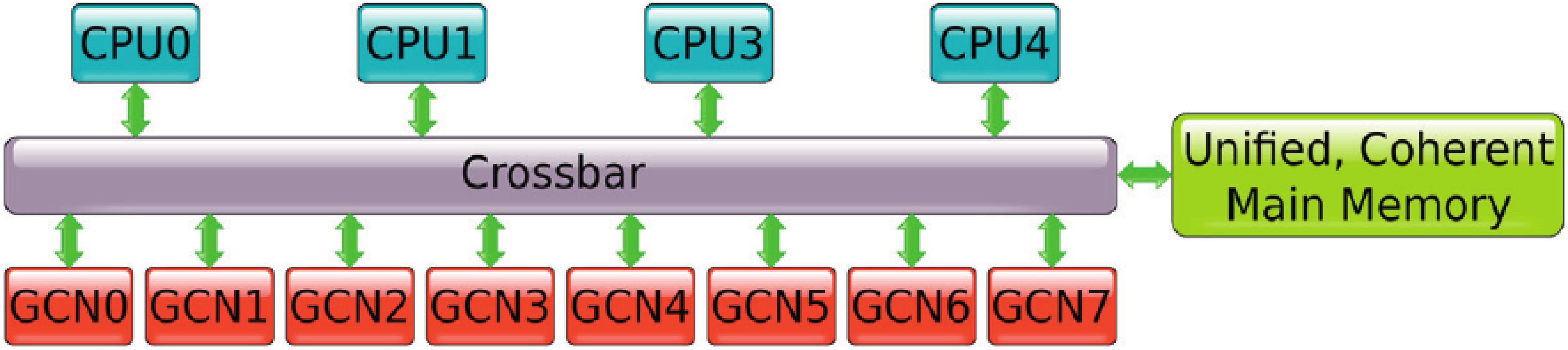
RIPHAH INTERNATIONAL UNIVERSITY

AMD GPU

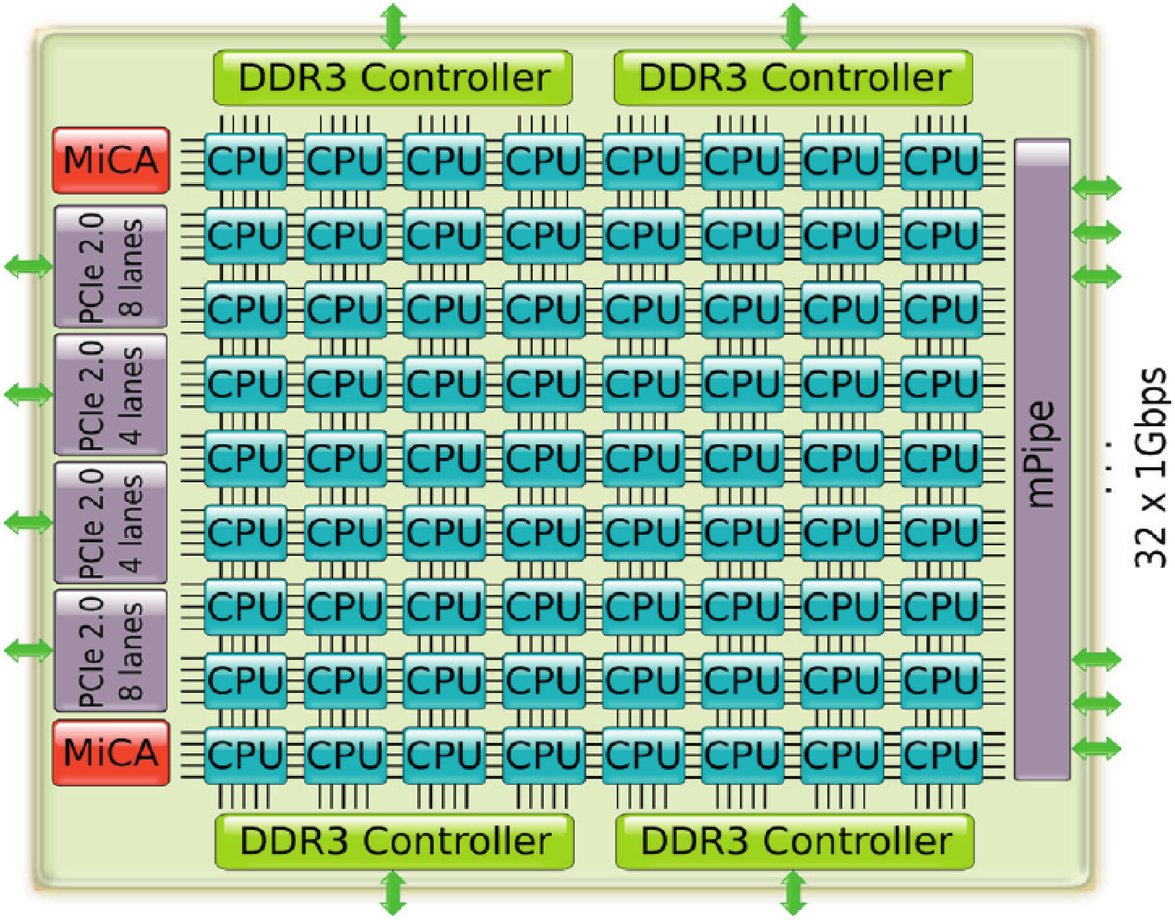
- AMD’s APU chips implement the Heterogeneous System Architecture (HSA).
- The significant of AMD GPU is the unification of the memory spaces of the CPU and GPU cores. This means that there is no communication overhead associated with assigning workload to the GPU cores, nor any delay in getting the results back.
- This also removes one of the major hassles in GPU programming, which is the explicit (or implicit, based on the middleware available) data transfers that need to take place.


CPU0
CPU1
CPU3
CPU4
Crossbar
Unified, Coherent Main Memory
GCN0
GCN1
GCN2
GCN3
GCN4
GCN5
GCN6
GCN7

- The HSA architecture identifies two core types:
- The Latency Compute Unit (LCU), which is a generalization of a CPU. A LCU supports both its native CPU instruction set and the HSA intermediate language (HSAIL) instruction set.
- The Throughput Compute Unit (TCU), which is a generalization of a GPU. A TCU supports only the HSAIL instruction set. TCUs target efficient parallel execution.

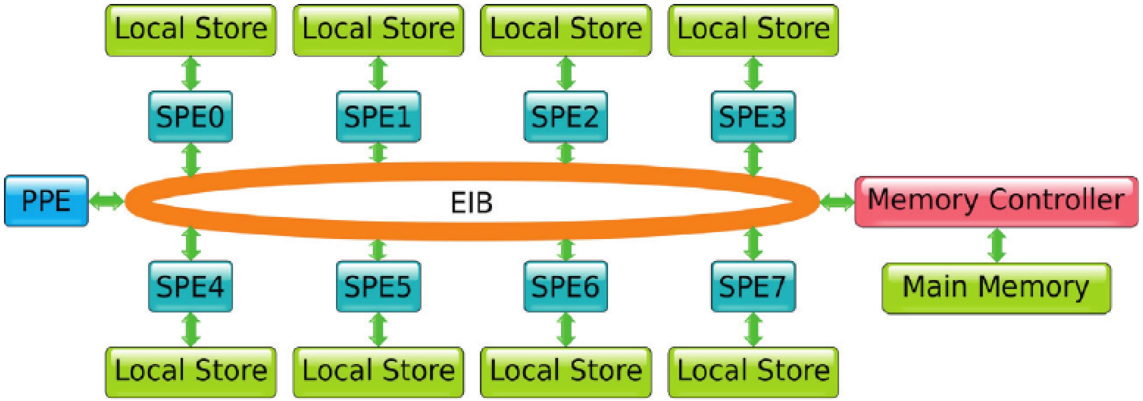


TILERA'S TILE-GX8072


DDR3 Controller
DDR3 Controller
MiCA
PCIe 2.0 8 lanes
PCIe 2.0 4 lanes
PCIe 2.0 4 lanes
PCIe 2.0 8 lanes
MiCA
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU
mPipe
32 x 1Gbps
DDR3 Controller
DDR3 Controller

Power PC

- Master Core: 64-bit PowerPC core also called the Power Processing Element.
- Worker Core: Synergistic Processing Element SPE having 128-bit vector processors.
- Own SIMD instruction set.



Marks Distribution

Evaluation (20 Marks)

Mid-term (35 Marks)

End-term (45 marks)

Task

- Select Data Signals for a Problem/Application e.g. ECG, PPG, EOG, etc.
    - Collect signals at-least 10 objects and label them against the problem. (5 Marks)
    - Write Problem statement after reading at-least 5 papers of your task. (3 Marks)
- Select a Processing Machine and Configure/Install Artificial Intelligence Frameworks. (2 Marks)
    - Install Linux
        - Use Anaconda https://www.anaconda.com/what-is-anaconda/

Example

Problem Selected: An Intelligent ECG based Human Object Identification System.

- Object 1: ECG Digital Signal
- Object 2: ECG Digital Signal
- …………………………………..
- Object 10: ECG Digital Signal

In this work, (ECG) signal for human identification issue has been investigated, and some methods have been suggested. An effective intelligent feature selection method from ECG signals has been proposed.

BSC~Microsoft Research Centre

High Performance Computing for Artificial Intelligence

Dr. Tassadaq Hussain

Assistant Professor Riphah International University

Collaborations:

Microsoft Research and Barcelona Supercomputing Center

Barcelona, Spain

UCERD Pvt Ltd Islamabad

Unal Center of Education Research and Development

www.ucerd.com