ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

KIẾN TRÚC MÁY TÍNH

Computer Architecture

Course ID: IT3030, IT3034, IT3283

Version: CA.RISCV.2025

Nội dung học phần

Chương 1. Giới thiệu chung

Chương 2. Hệ thống máy tính

Chương 3. Kiến trúc tập lệnh

Chương 4. Số học máy tính

Chương 5. Bộ xử lý

Chương 6. Bộ nhớ máy tính

Chương 7. Hệ thống vào-ra

Chương 8. Các kiến trúc song song

Chương 8
CÁC KIẾN TRÚC SONG SONG

Nội dung

8.1. Phân loại kiến trúc máy tính

8.2. Bộ xử lý đồ họa đa dụng

8.3. Đa xử lý bộ nhớ dùng chung

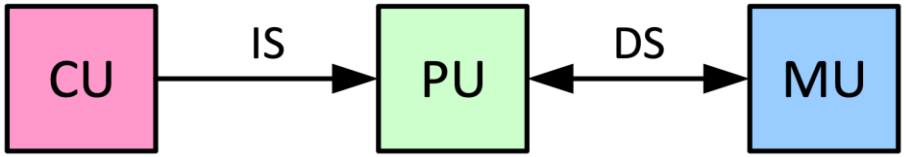
8.4. Đa xử lý bộ nhớ phân tán

8.1. Phân loại kiến trúc máy tính

Phân loại kiến trúc máy tính (Michael Flynn -1966)

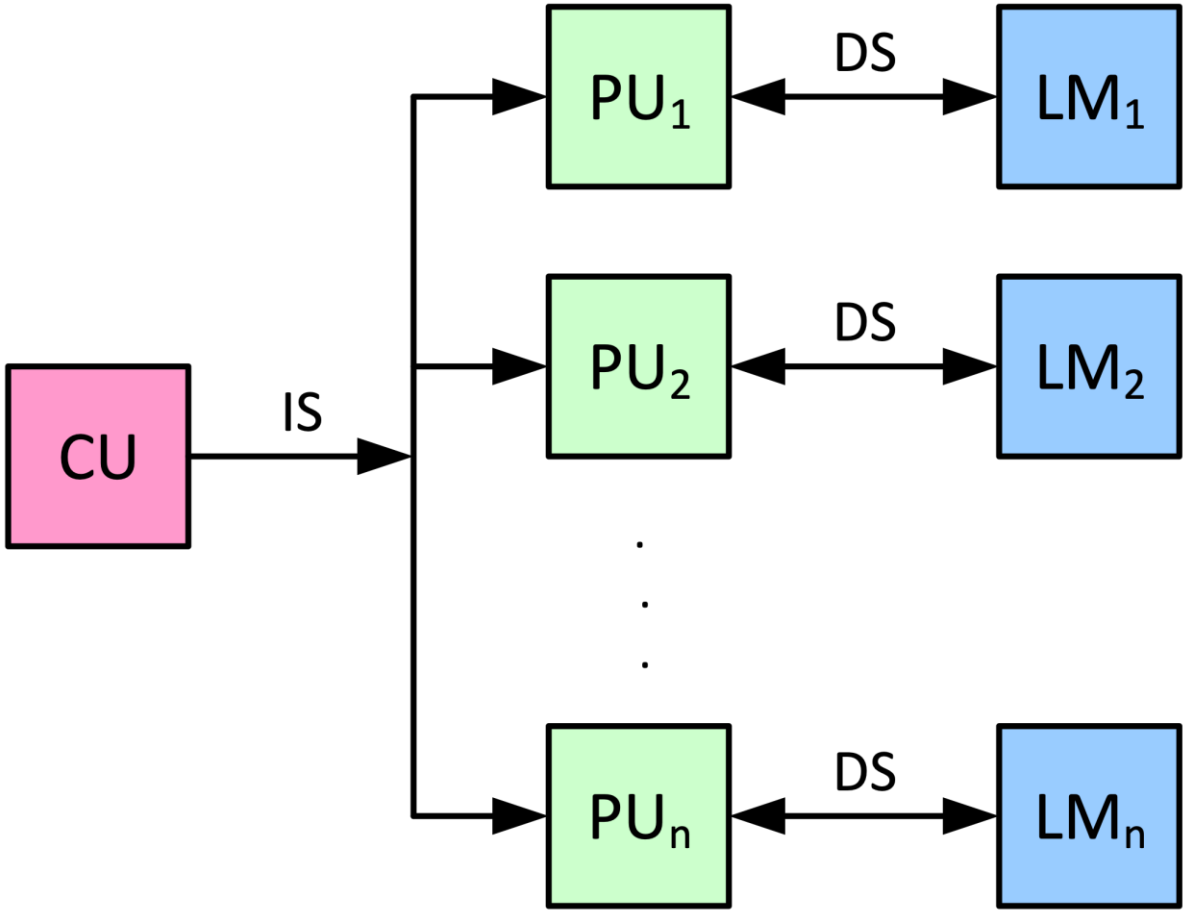
- SISD - Single Instruction Stream, Single Data Stream
- SIMD - Single Instruction Stream, Multiple Data Stream
- MISD - Multiple Instruction Stream, Single Data Stream
- MIMD - Multiple Instruction Stream, Multiple Data Stream

SISD



- CU: Control Unit
- PU: Processing Unit
- MU: Memory Unit
- Một bộ xử lý
- Đơn dòng lệnh
- Dữ liệu được lưu trữ trong một bộ nhớ
- Chính là Kiến trúc von Neumann (tuần tự)

SIMD



SIMD (tiếp)

- Đơn dòng lệnh điều khiển đồng thời các đơn vị xử lý PUs
- Mỗi đơn vị xử lý có một bộ nhớ dữ liệu riêng LM (local memory)
- Mỗi lệnh được thực hiện trên một tập các dữ liệu khác nhau
- Các mô hình SIMD
  - Vector Computers
  - GPU (Graphic Processing Unit)

MISD

- Một luồng dữ liệu cùng được truyền đến một tập các bộ xử lý
- Mỗi bộ xử lý thực hiện một dãy lệnh khác nhau.
- Chưa tồn tại máy tính thực tế
- Có thể có trong tương lai

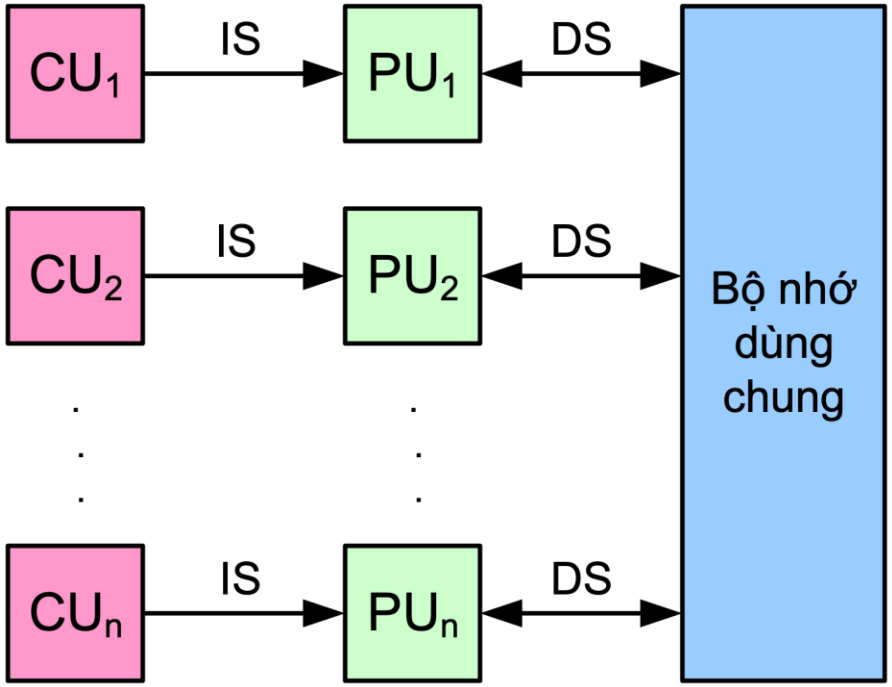
MIMD

- Tập các bộ xử lý
- Các bộ xử lý đồng thời thực hiện các dãy lệnh khác nhau trên các dữ liệu khác nhau
- Các mô hình MIMD
  - Multiprocessors (Shared Memory Multiprocessors)
  - Multicomputers (Distributed Memory Multiprocessors)

MIMD - Shared Memory

Đa xử lý bộ nhớ dùng chung

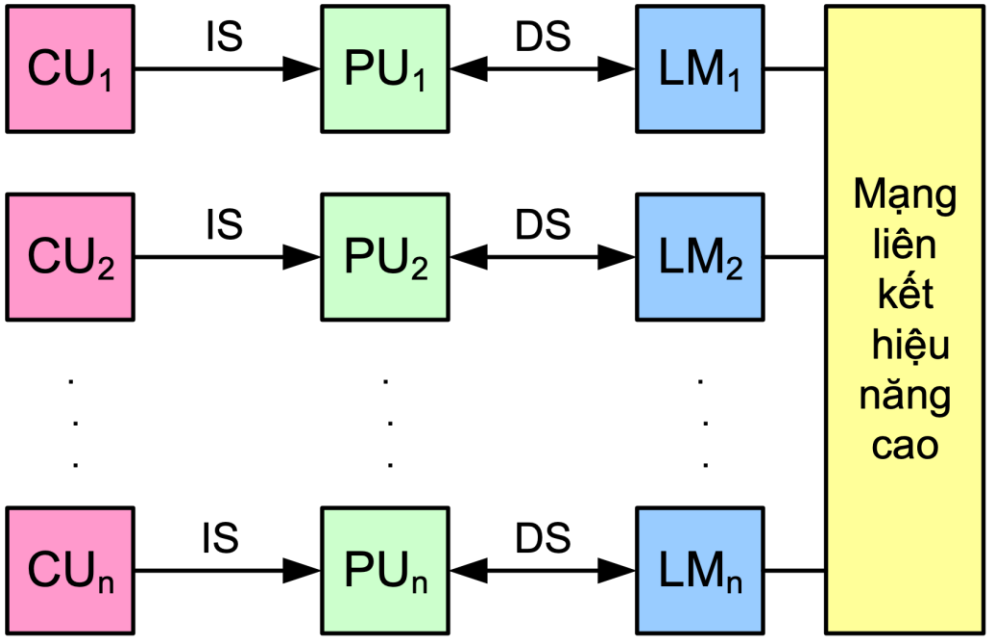
(shared memory mutiprocessors)



MIMD - Distributed Memory

Đa xử lý bộ nhớ phân tán

(distributed memory mutiprocessors or multicomputers)



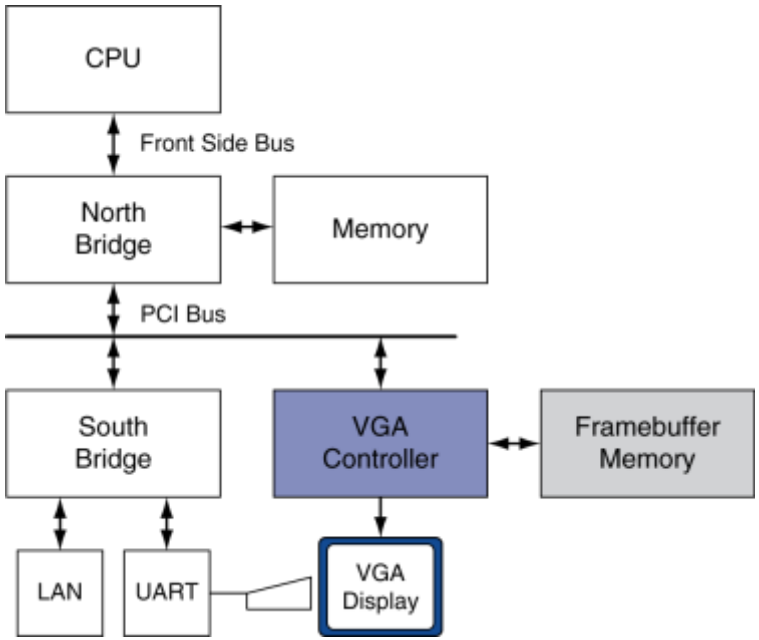
Phân loại các kỹ thuật song song

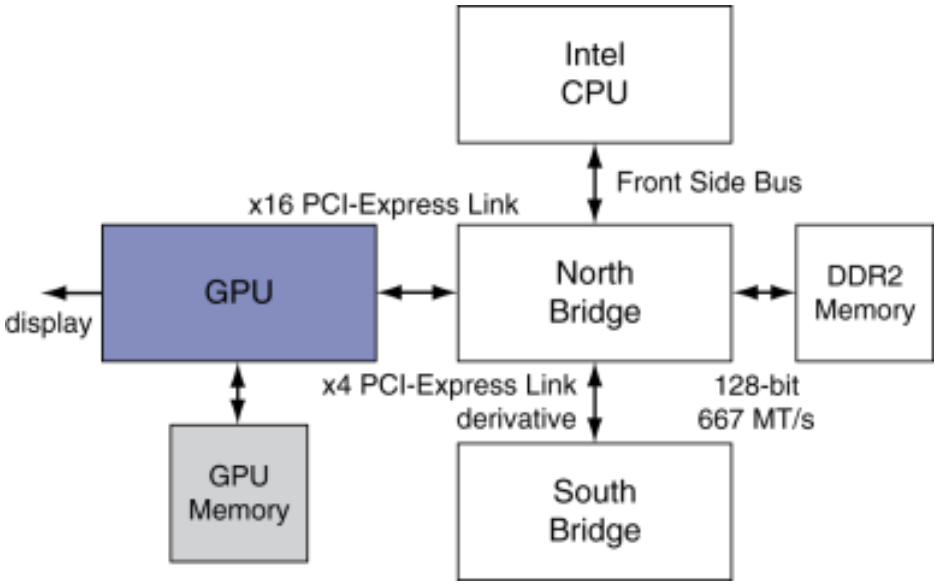
- Song song mức lệnh (Instruction-Level Parallelism)
  - pipeline
  - superscalar
- Song song mức dữ liệu (Data-Level Parallelism)
  - SIMD
- Song song mức luồng (Thread-Level Parallelism)
  - MIMD
- Song song mức yêu cầu (Request-Level Parallelism)
  - Cloud computing

8.2. Bộ xử lý đồ họa đa dụng (GPU)

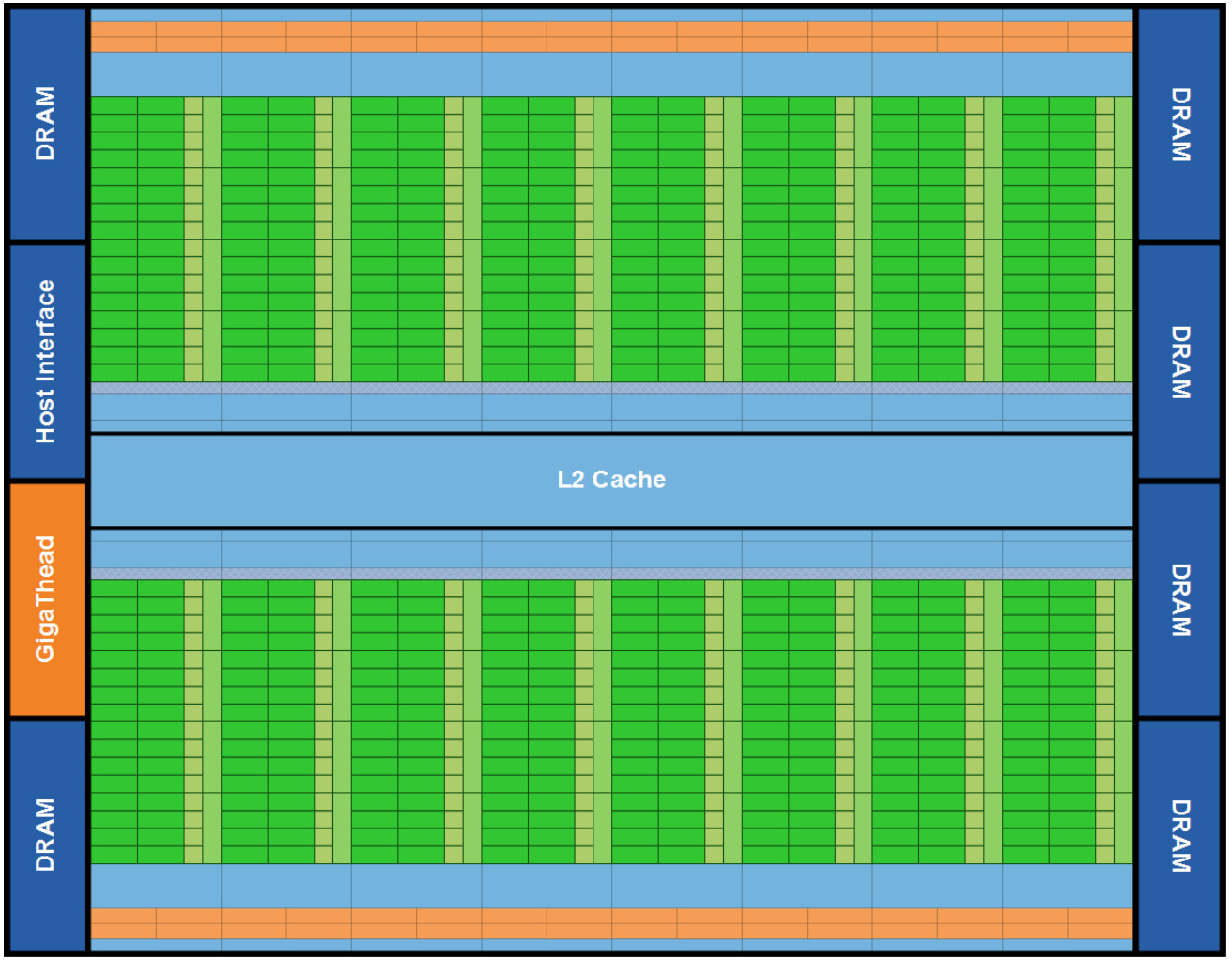
- Kiến trúc SIMD
- Xuất phát từ bộ xử lý đồ họa GPU (Graphic Processing Unit) hỗ trợ xử lý đồ họa 2D và 3D: xử lý dữ liệu song song
- GPGPU – General purpose Graphic Processing Unit
- Hệ thống lai CPU/GPGPU
  - CPU là host: thực hiện theo tuần tự
  - GPGPU: tính toán song song

Bộ xử lý đồ họa trong máy tính



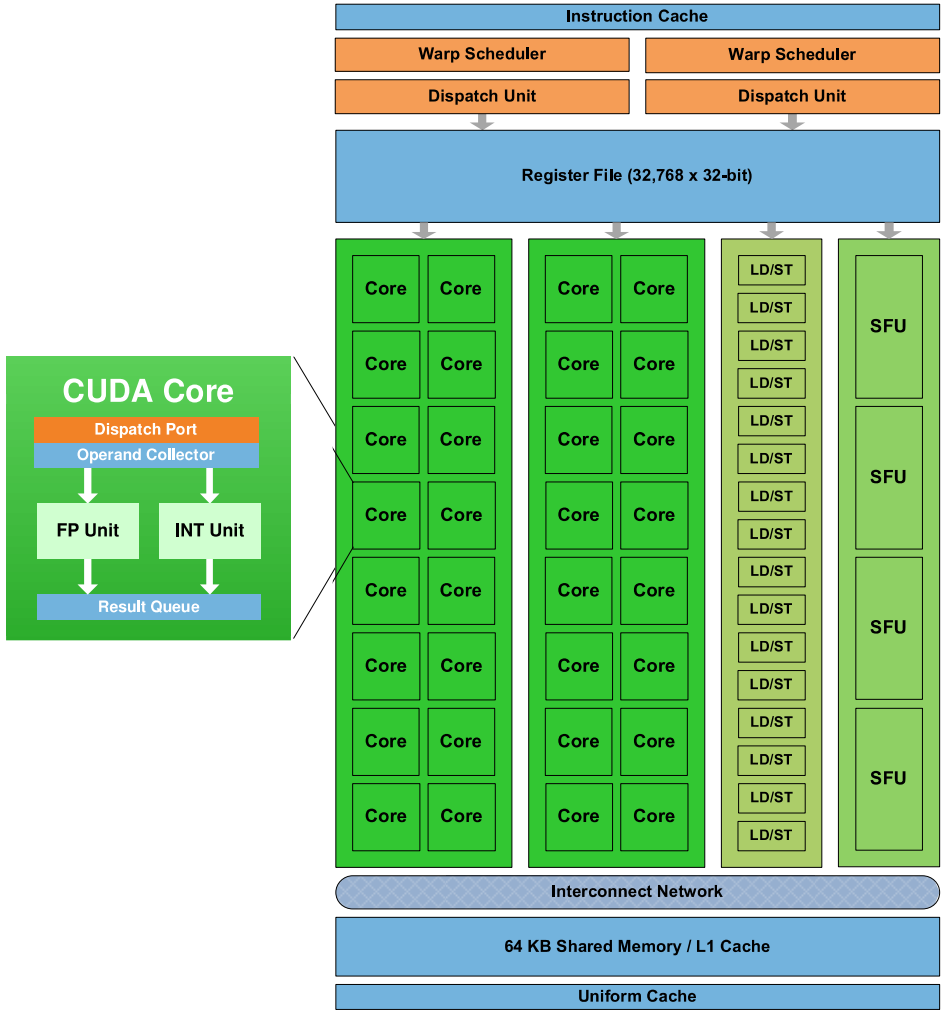


GPGPU: NVIDIA Fermi



NVIDIA Fermi

- Có 16 Streaming Multiprocessors (SM)
- Mỗi SM có 32 CUDA cores.
- Mỗi CUDA core (Compute Unified Device Architecture) có 01 FPU và 01 IU



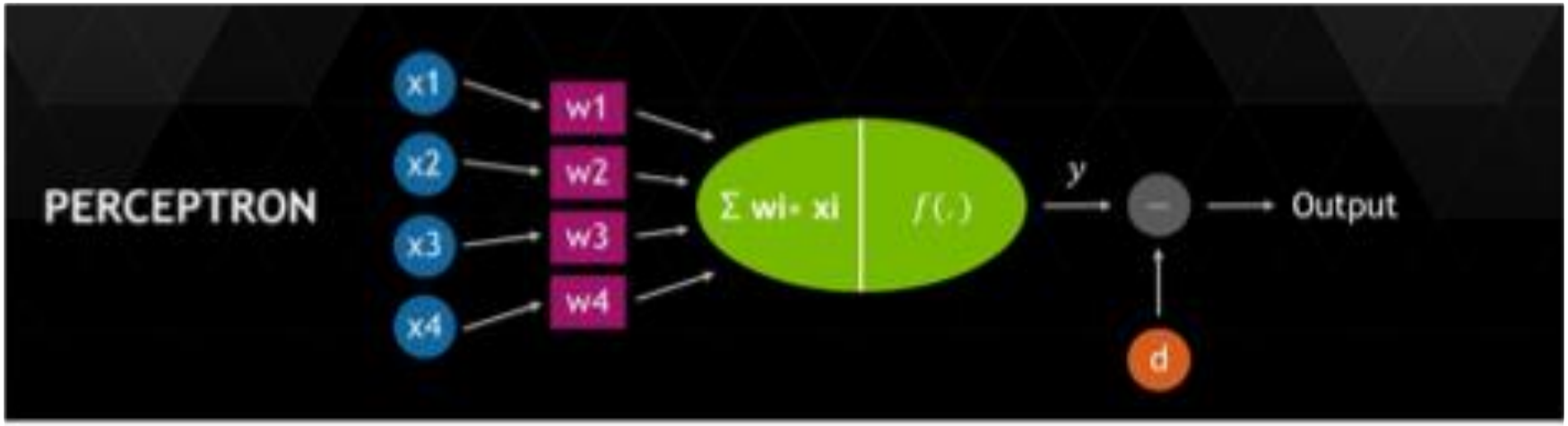
Volta GPU (2017)

So sánh các NVIDIA Tesla GPU

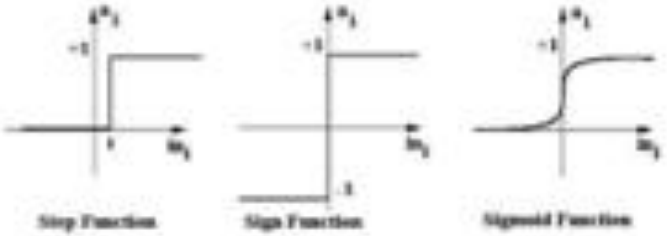
Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1530 MHz
Peak FP32 TFLOPS[1]	5	6.8	10.6	15.7
Peak FP64 TFLOPS[1]	1.7	.21	5.3	7.8
Peak Tensor TFLOPS[1]	NA	NA	NA	125
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2	815 mm^2
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN
[1] Peak TFLOPS rates are based on GPU Boost Clock				

Volta GV100 Streaming Multiprocessor (SM)

Perceptron



ACTIVATION FUNCTIONS:



LEARNING:

$$y^{(t)} = f\left\{\sum_i w_i^{(t)} x_i^{(t)}\right\}$$

Update:

$$\Delta w_i^{(t)} = \varepsilon(d^{(t)} - y^{(t)}) x_i^{(t)}$$

$$w_i^{(t+1)} = w_i^{(t)} + \Delta w_i^{(t)}$$

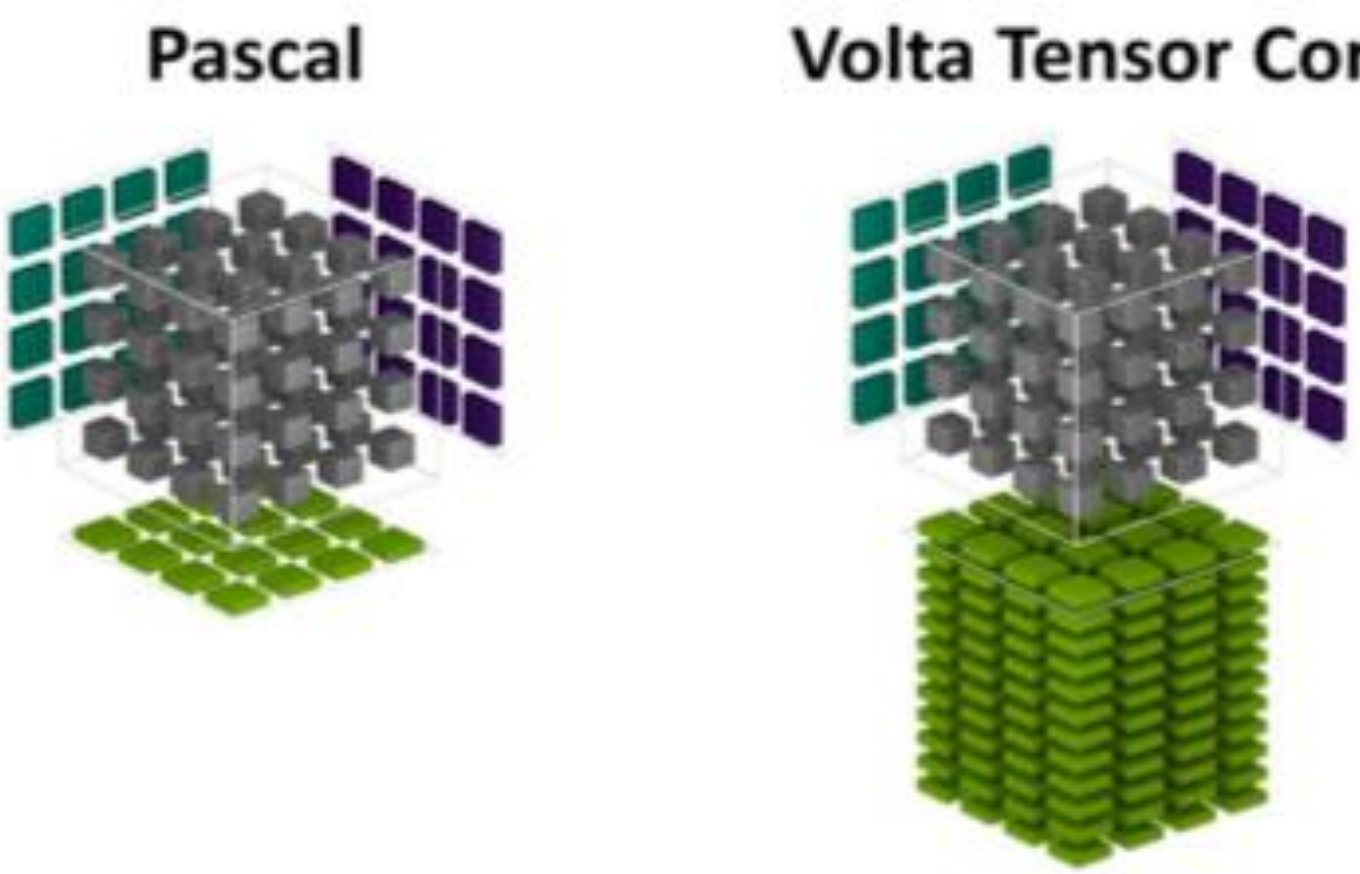
Tensor Core

Phép toán cốt lõi của neural networks:

$$D = A \times B + C$$

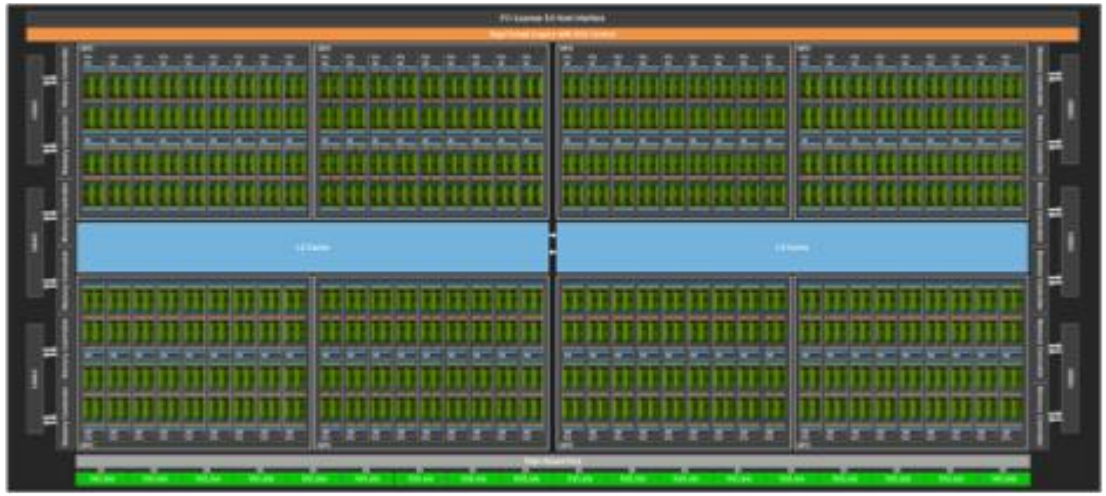
A, B, C, D là các ma trận 4x4

Nhân ma trận 4x4 trên hai GPU: Pascal và Volta



NVIDIA Ampere GPU A100 (2020)

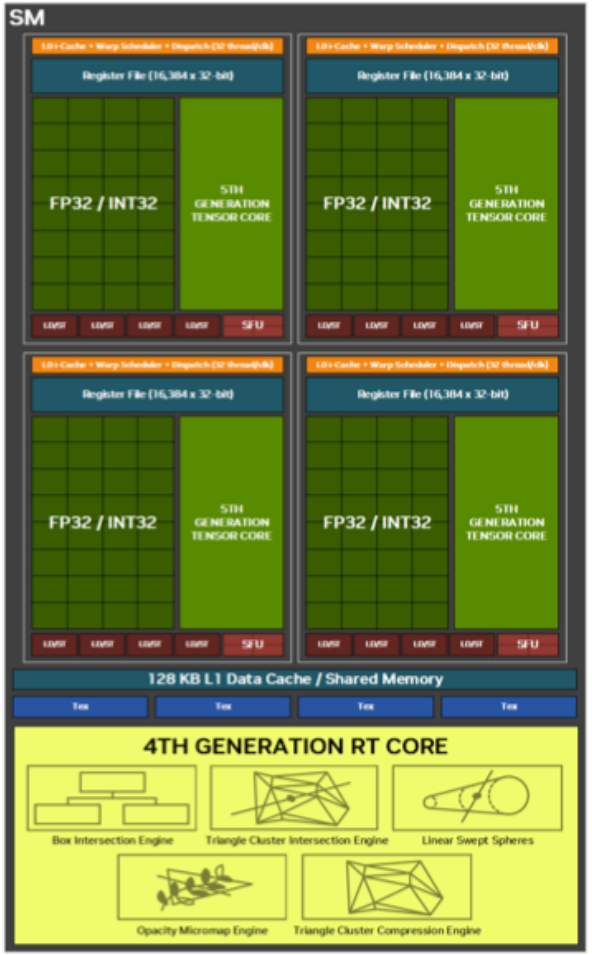
Hopper GPU (2022)



Blackwell GPU (2024)

The full GB202 GPU includes: 192 SMs

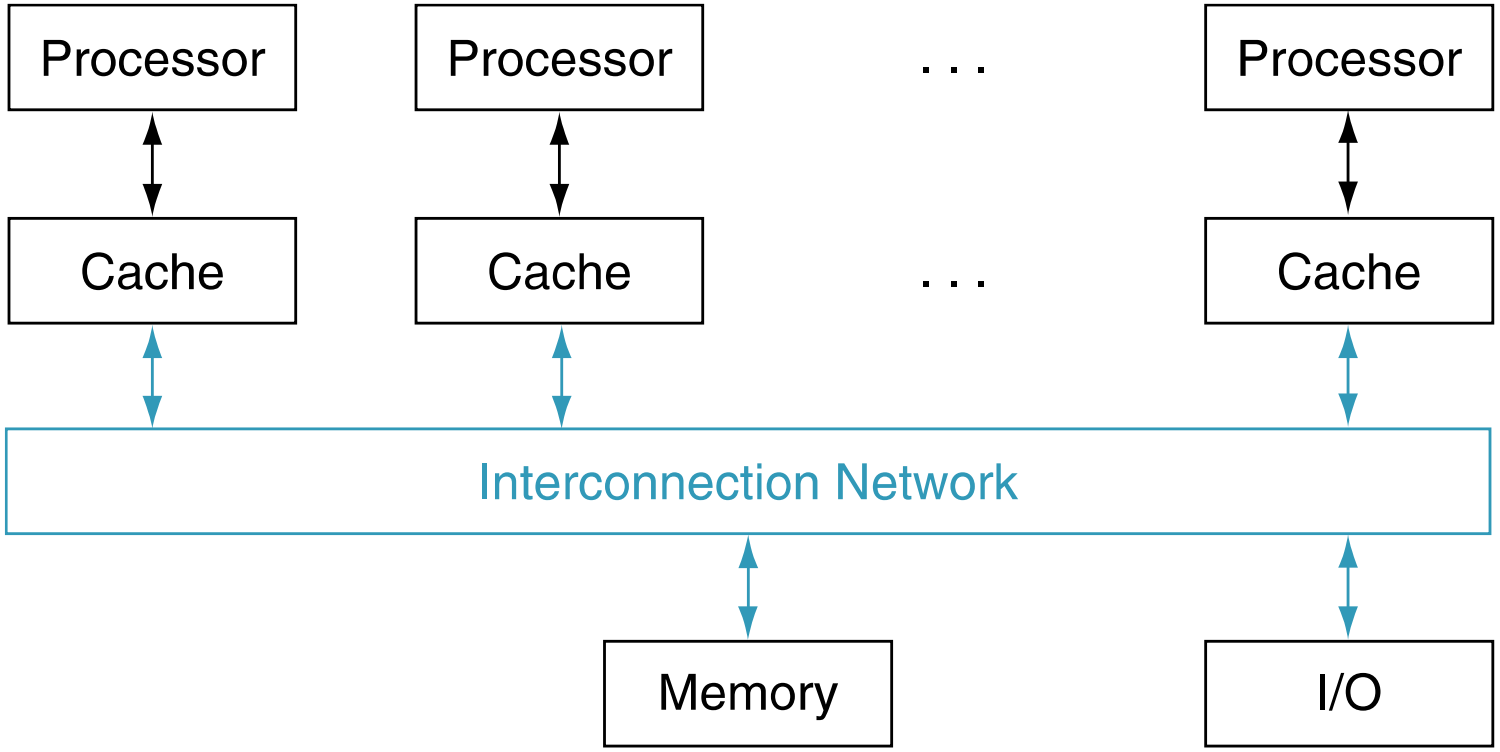
- 24576 CUDA Cores
- 192 RT Cores
- 768 Tensor Cores
- 768 Texture Units



8.3. Đa xử lý bộ nhớ dùng chung

- Hệ thống đa xử lý đối xứng (SMP- Symmetric Multiprocessors)
- Hệ thống đa xử lý không đối xứng (NUMA – Non-Uniform Memory Access)

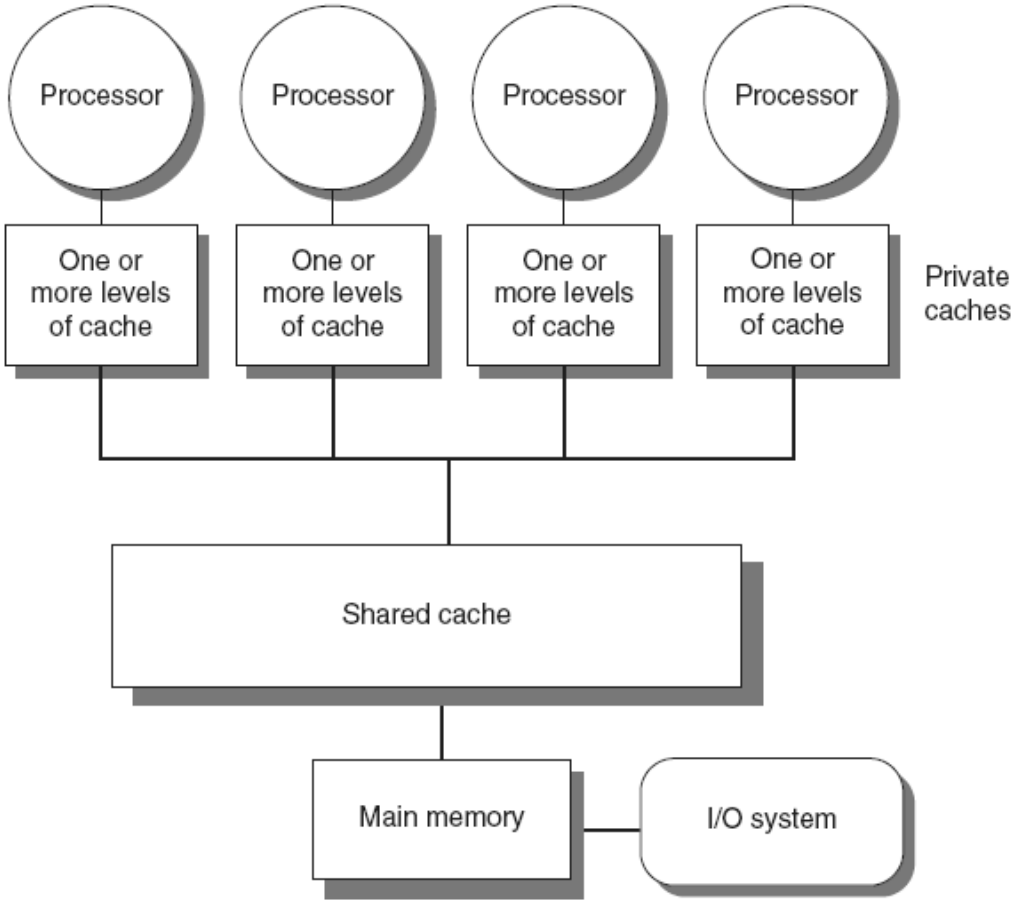
Đa xử lý đối xứng SMP (Symmetric Multiprocessors)



SMP (tiếp)

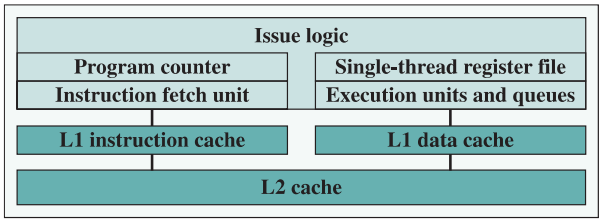
- Một máy tính có n >= 2 bộ xử lý giống nhau
- Các bộ xử lý dùng chung bộ nhớ và hệ thống vào-ra
- Thời gian truy cập bộ nhớ là bằng nhau với các bộ xử lý
- Các bộ xử lý có thể thực hiện chức năng giống nhau
- Hệ thống được điều khiển bởi một hệ điều hành phân tán
- Hiệu năng: Các công việc có thể thực hiện song song
- Khả năng chịu lỗi

Máy tính SMP dùng bộ xử lý đa lõi

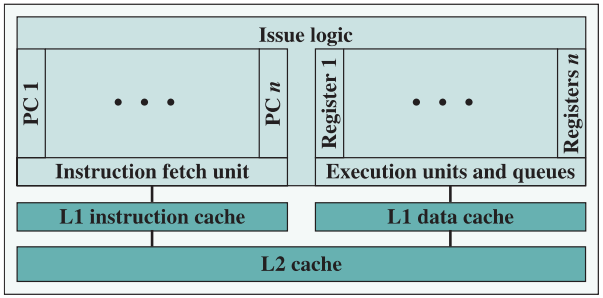


Thay đổi của bộ xử lý

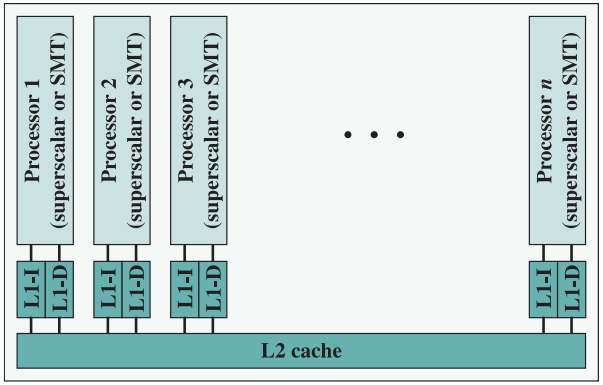
- Tuần tự
- Pipeline
- Siêu vô hướng
- Đa luồng đồng thời
- Đa lõi: nhiều CPU trên một chip
- Đa lõi CPU/GPU
- Đa lõi CPU/DSP
- Tensor core/TPU/NPU



(a) Superscalar

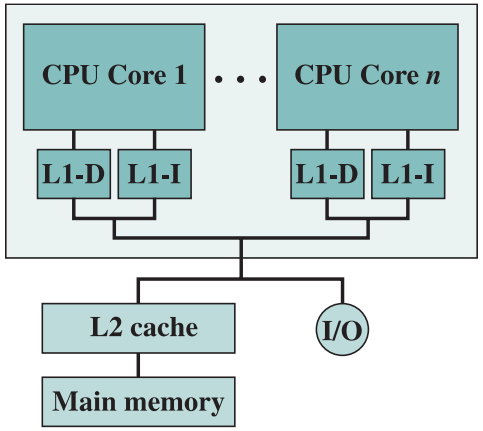


(b) Simultaneous multithreading

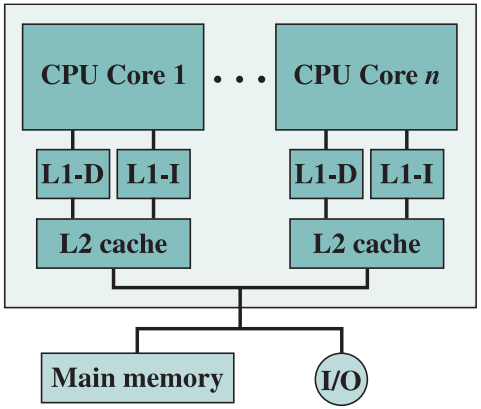


(c) Multicore

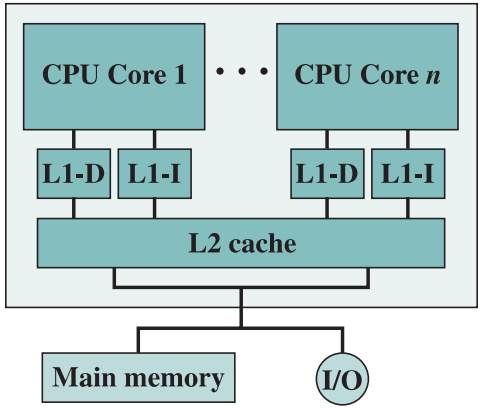
Các dạng tổ chức bộ xử lý đa lõi



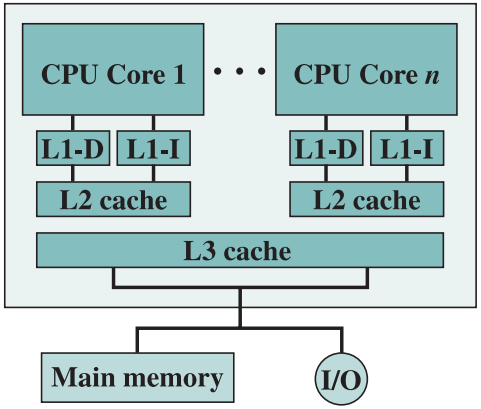
(a) Dedicated L1 cache



(b) Dedicated L2 cache

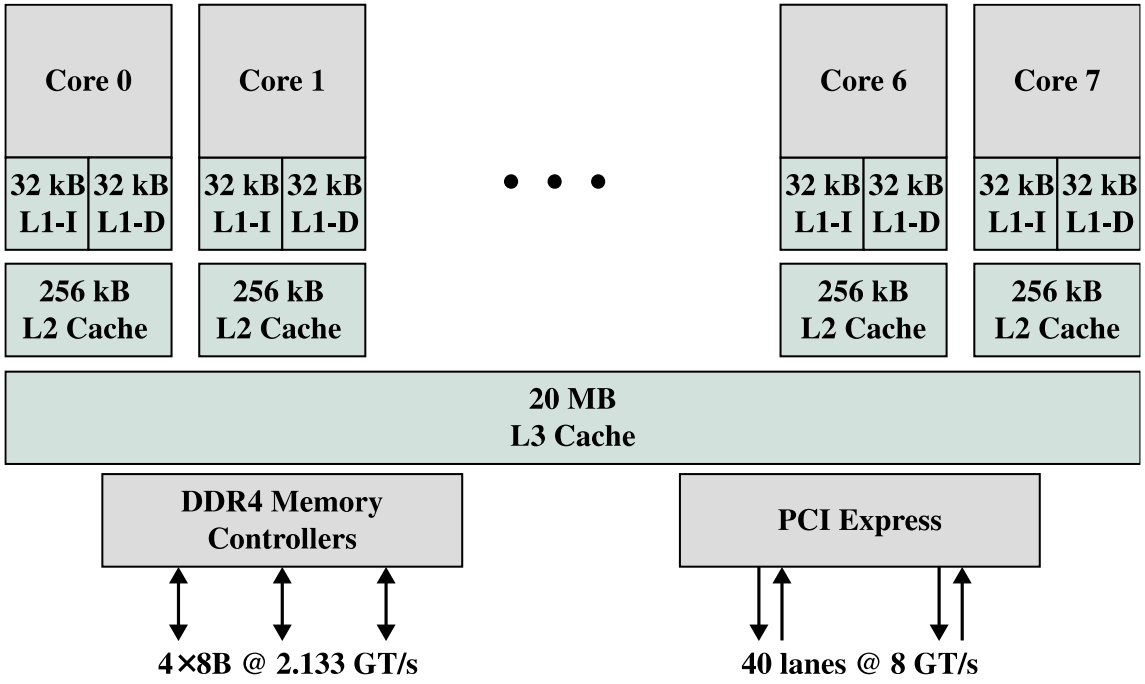


(c) Shared L2 cache



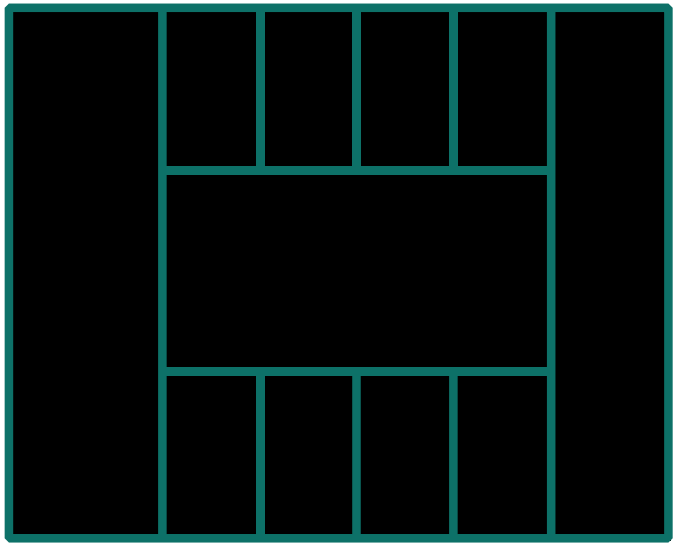
(d) Shared L3 cache

Intel Core i7-5960X

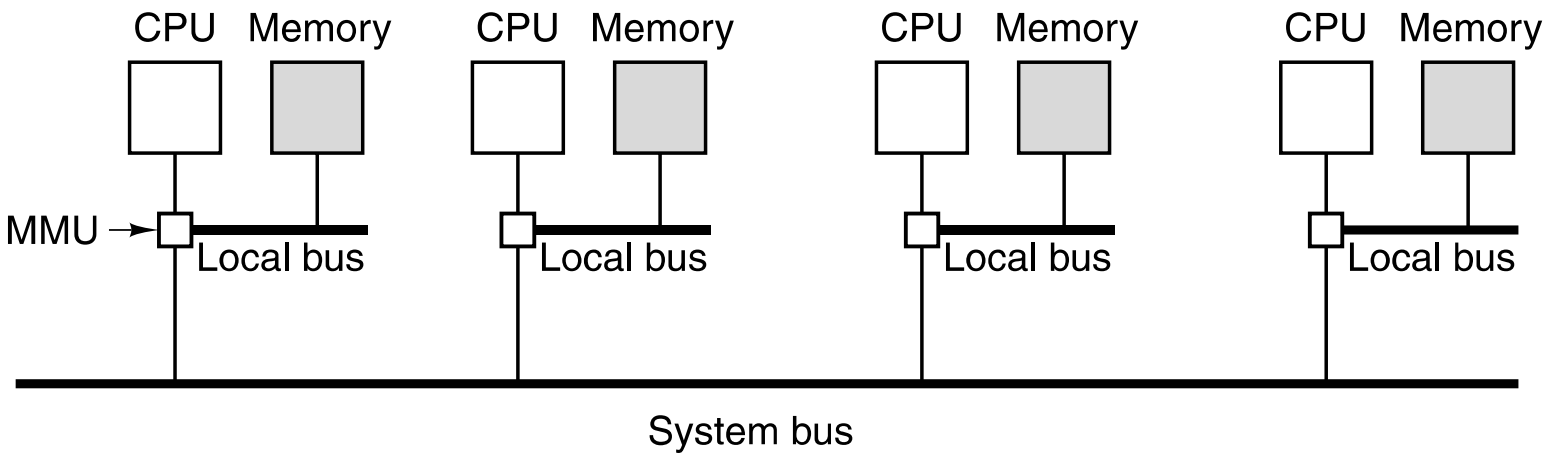


Block diagram

Physical layout on chip

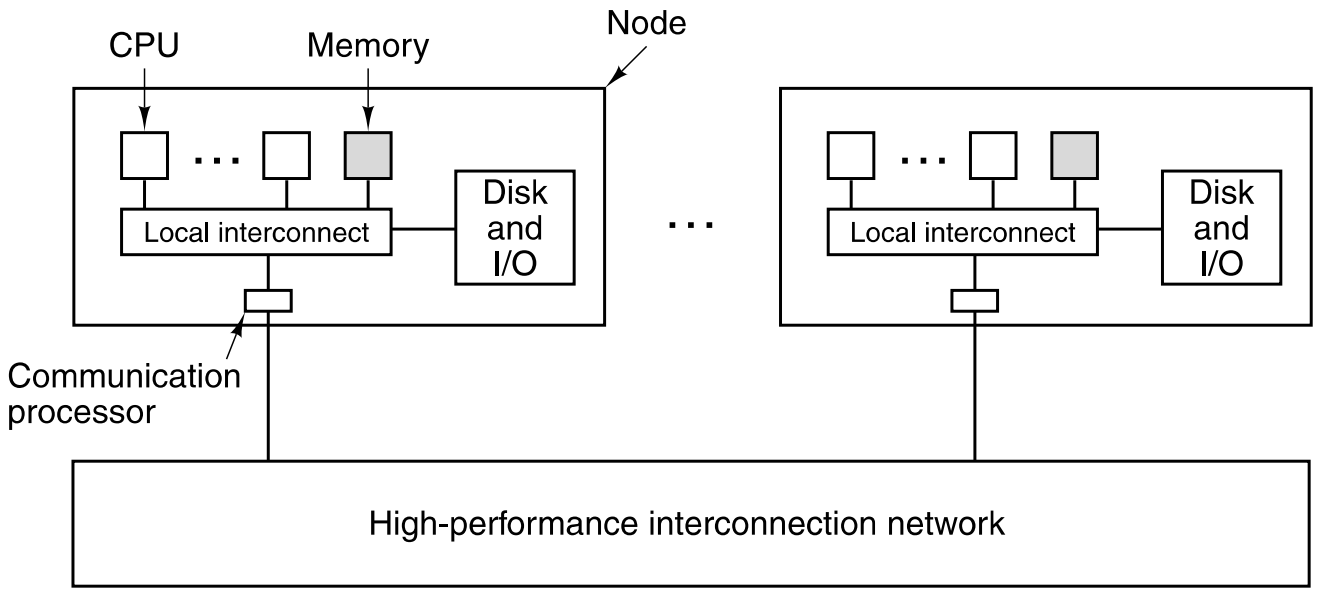


NUMA (Non-Uniform Memory Access)



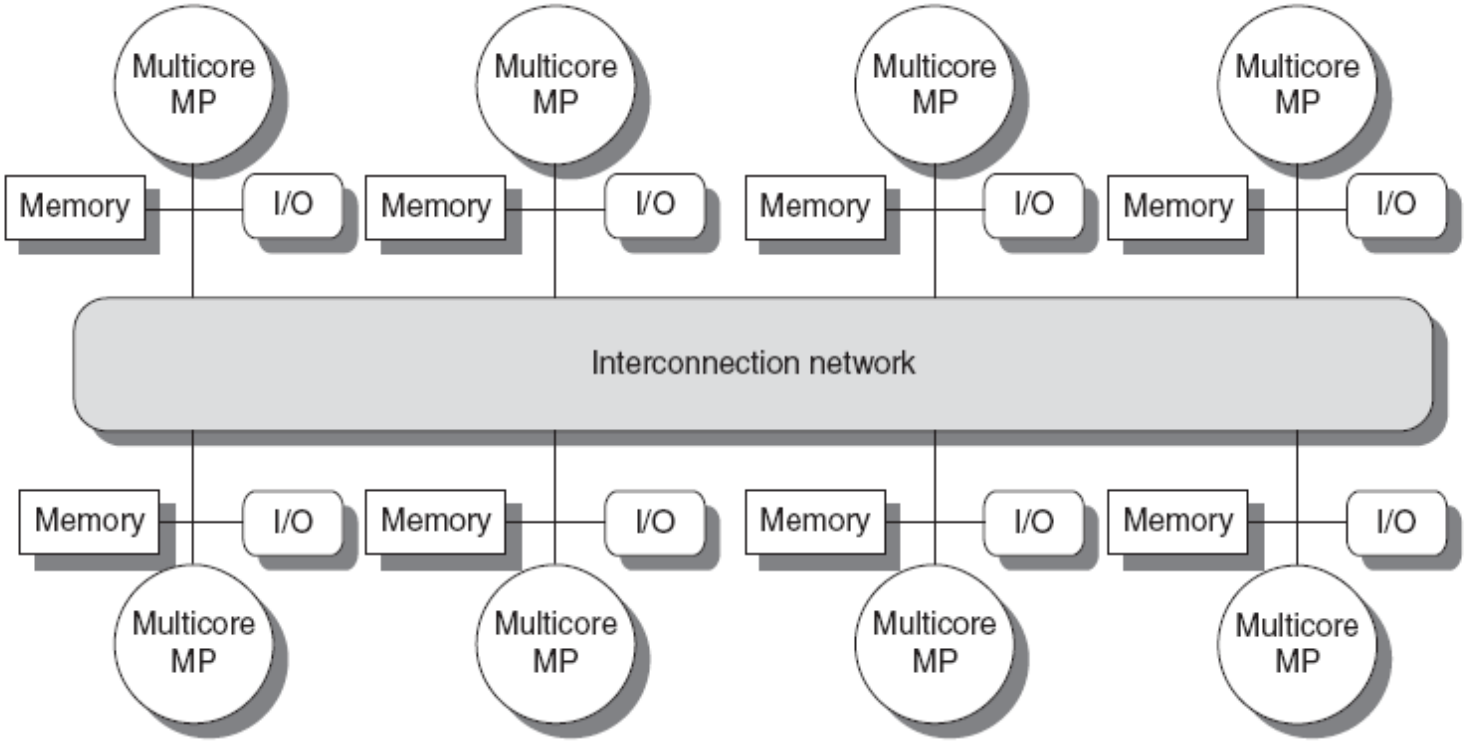
- Có một không gian địa chỉ chung cho tất cả CPU
- Mỗi CPU có thể truy cập từ xa sang bộ nhớ của CPU khác
- Truy nhập bộ nhớ từ xa chậm hơn truy nhập bộ nhớ cục bộ

8.4. Đa xử lý bộ nhớ phân tán

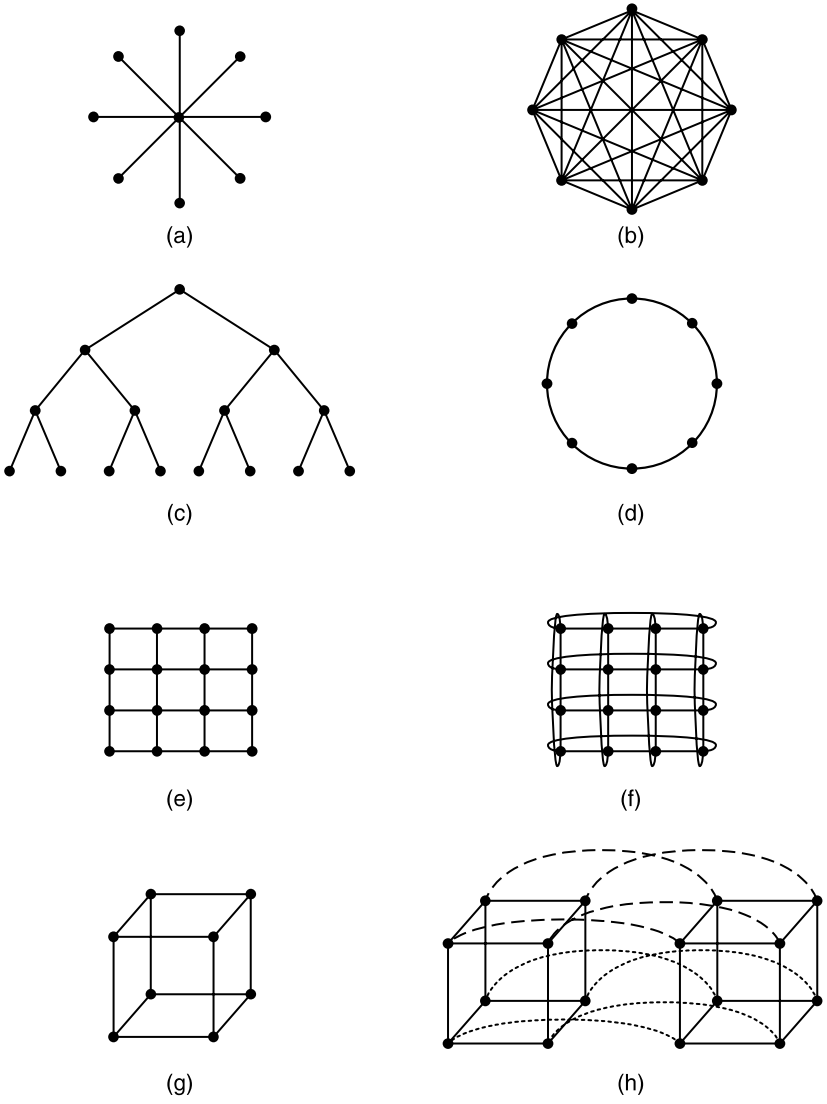


- Máy tính qui mô lớn (Warehouse Scale Computers)
- Máy tính cụm (clusters)

Đa xử lý bộ nhớ phân tán



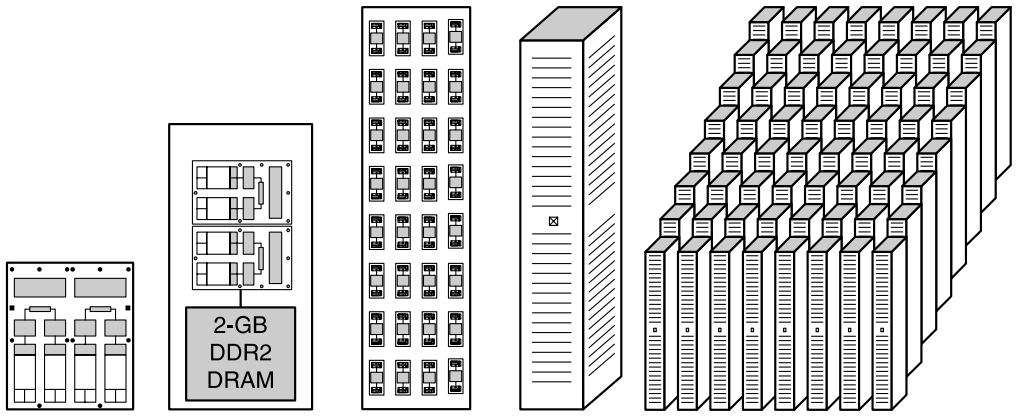
Mạng liên kết



Warehouse Scale Computers

- Hệ thống qui mô lớn
- Đắt tiền: nhiều triệu USD
- Dùng cho tính toán khoa học và các bài toán có số phép toán và dữ liệu rất lớn
- Siêu máy tính

IBM Blue Gene/P



Chip:	Card	Board	Cabinet	System
4 processors 8-MB L3 cache	1 Chip 4 CPUs 2 GB	32 Cards 32 Chips 128 CPUs 64 GB	32 Boards 1024 Cards 1024 Chips 4096 CPUs 2 TB	72 Cabinets 73728 Cards 73728 Chips 294912 CPUs 144 TB
(a)	(b)	(c)	(d)	(e)

Cluster

- Nhiều máy tính được kết nối với nhau bằng mạng liên kết tốc độ cao (~ Gbps)
- Mỗi máy tính có thể làm việc độc lập (PC hoặc SMP)
- Mỗi máy tính được gọi là một node
- Các máy tính có thể được quản lý làm việc song song theo nhóm (cluster)
- Toàn bộ hệ thống có thể coi như là một máy tính song song
- Tính sẵn sàng cao
- Khả năng chịu lỗi lớn

Hết chương 8