



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

# KIẾN TRÚC MÁY TÍNH

## Computer Architecture

Course ID: IT3030

Nguyễn Kim Khánh

# Nội dung học phần

Chương 1. Giới thiệu chung

Chương 2. Hệ thống máy tính

Chương 3. Số học và logic máy tính

Chương 4. Kiến trúc tập lệnh

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



- 8.1. Phân loại kiến trúc máy tính
- 8.2. Đa xử lý bộ nhớ dùng chung
- 8.3. Đa xử lý bộ nhớ phân tán
- 8.4. Bộ xử lý đồ họa đa dụng

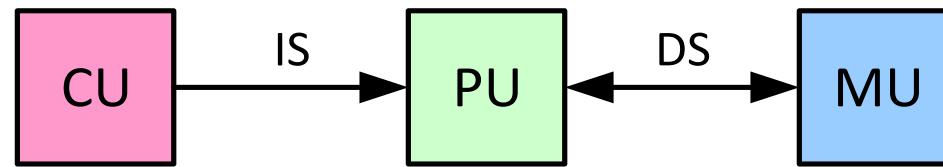


## 8.1. Phân loại kiến trúc máy tính

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



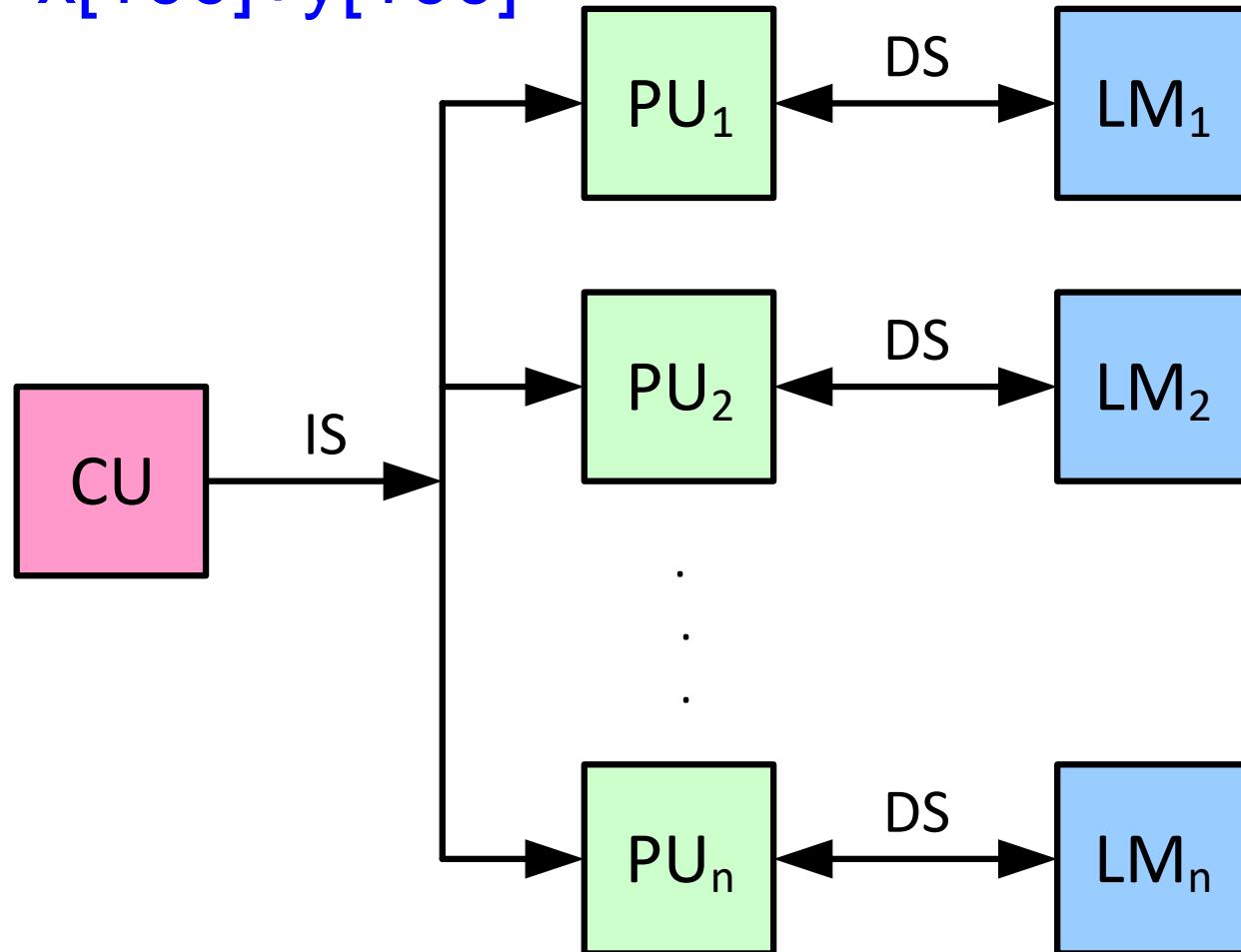


- CU: Control Unit
- PU: Processing Unit
- MU: Memory Unit
- Một bộ xử lý
- Đơn dòng lệnh
- Dữ liệu được lưu trữ trong một bộ nhớ
- Chính là Kiến trúc von Neumann (tuần tự)

$$Z = x + y$$

# SIMD

$$z[100] = x[100] + y[100]$$



# 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 Computer
  - Array processor



- 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

luồng 2 lệnh được nạp vào 2 BXL để tính cùng lúc?

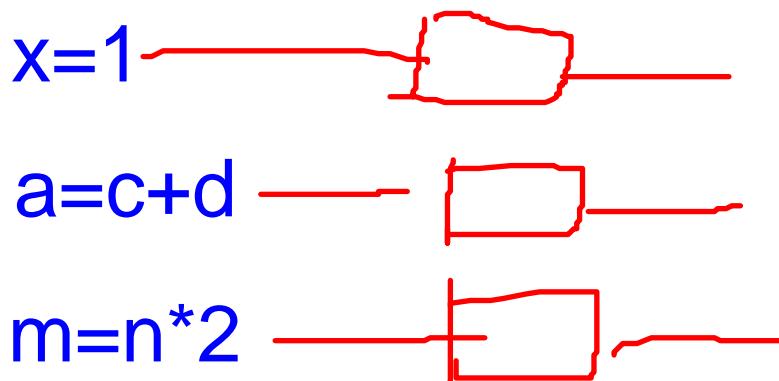
$$x = 1$$

$$x = y+2$$



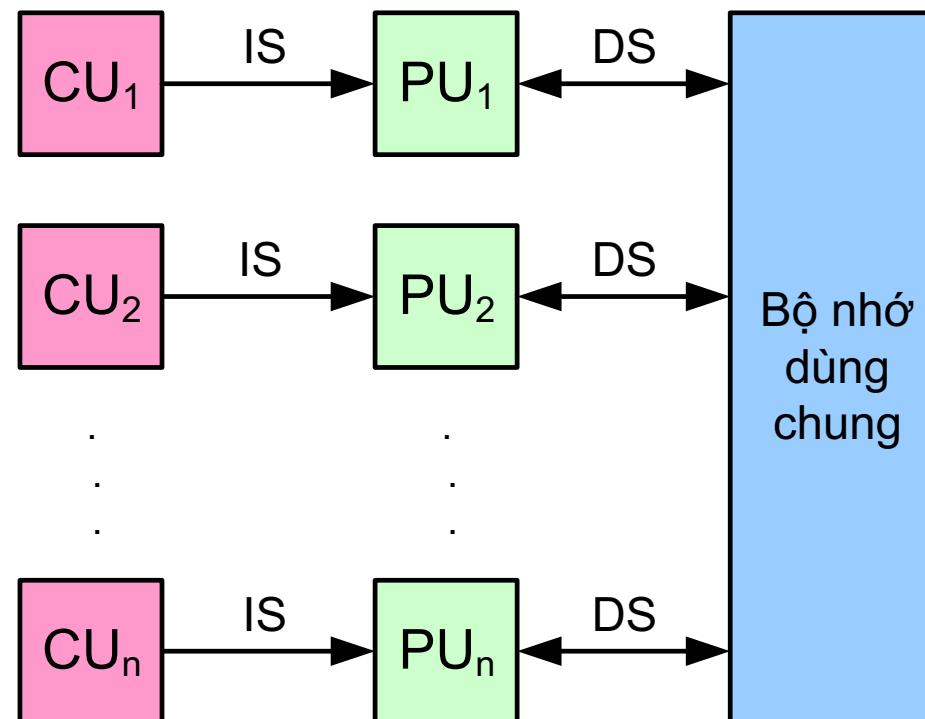
# MIMD

- 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)
  - Multicomputers (Distributed Memory)



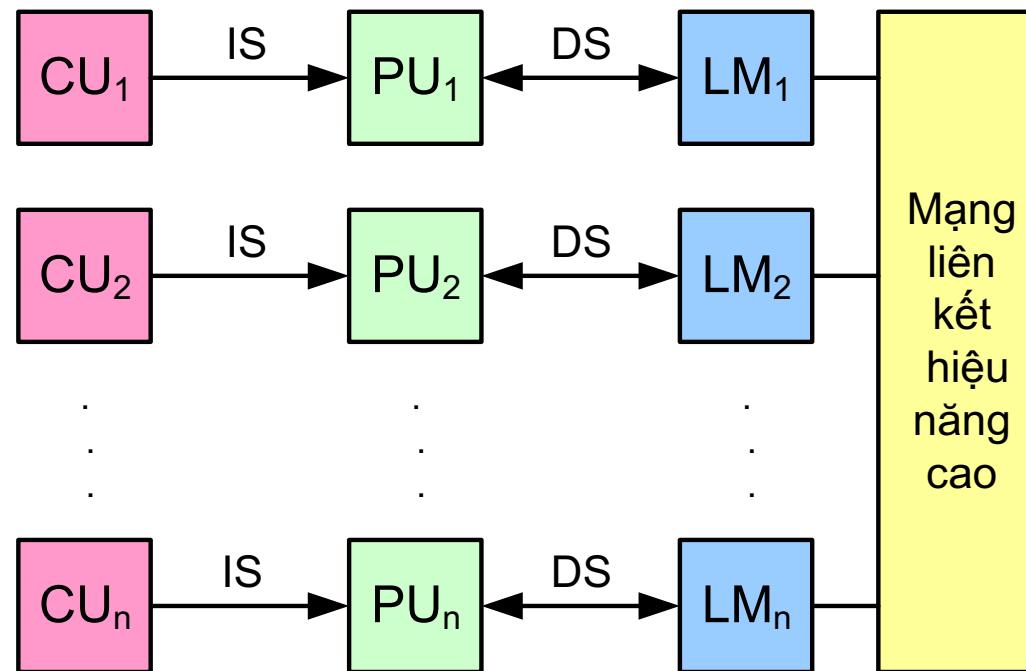
# MIMD - Shared Memory

Đa xử lý bộ nhớ dùng chung  
(shared memory multiprocessors)



# MIMD - Distributed Memory

Đa xử lý bộ nhớ phân tán  
(distributed memory multiprocessors or  
multicomputers)



# Phân loại các kỹ thuật song song

- Song song mức lệnh
  - pipeline
  - superscalar
- Song song mức dữ liệu
  - SIMD
- Song song mức luồng
  - MIMD
- Song song mức yêu cầu
  - Cloud computing

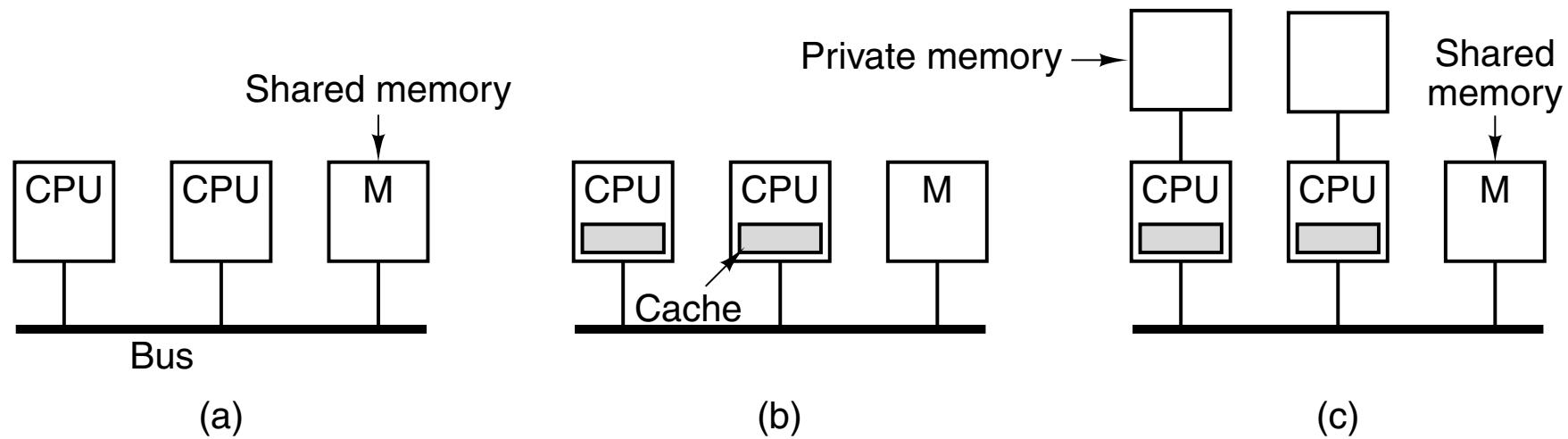


## 8.2. Đa xử lý bộ nhớ dùng chung

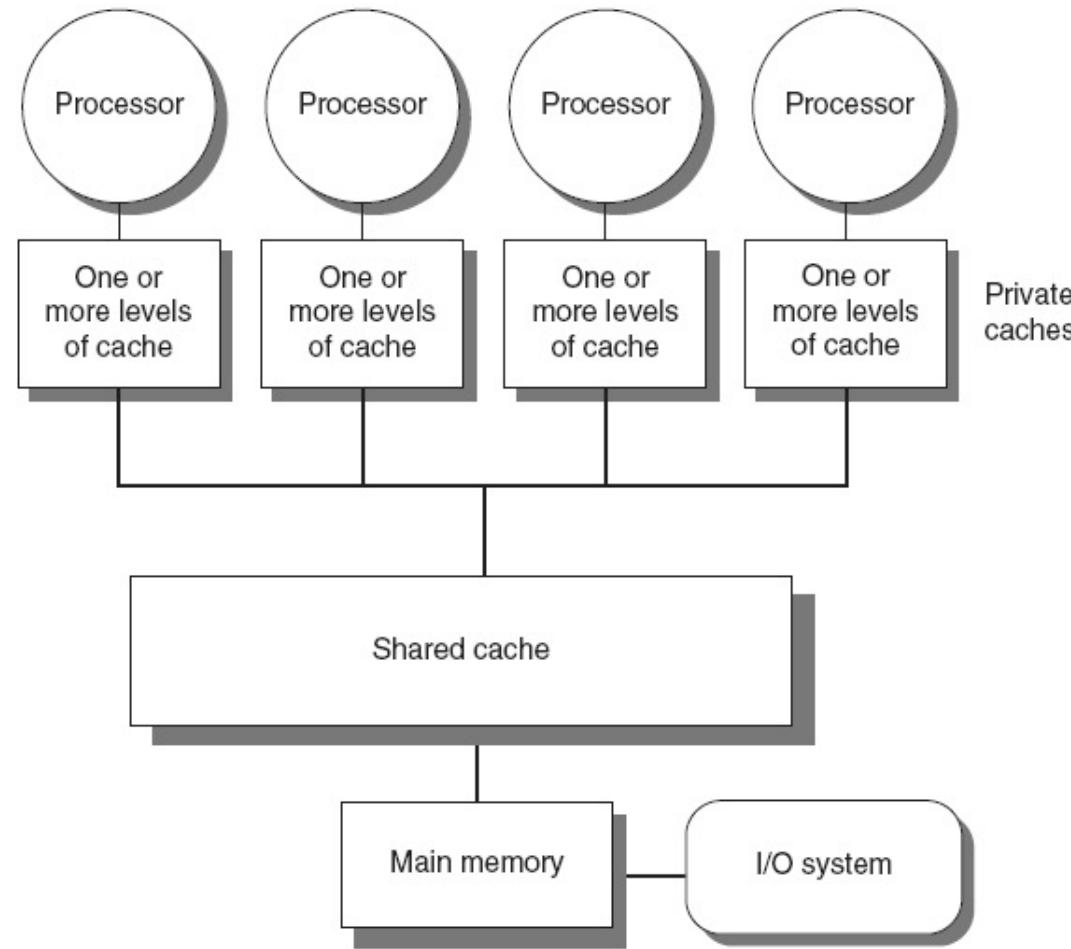
- Hệ thống đa xử lý đối xứng (SMP- Symmetric Multiprocessors)
- Hệ thống đa xử lý không đối xứng (NUMA - Non-Uniform Memory Access)
- Bộ xử lý đa lõi (Multicore Processors)



# SMP hay UMA (Uniform Memory Access)



# SMP hay UMA (Uniform Memory Access)

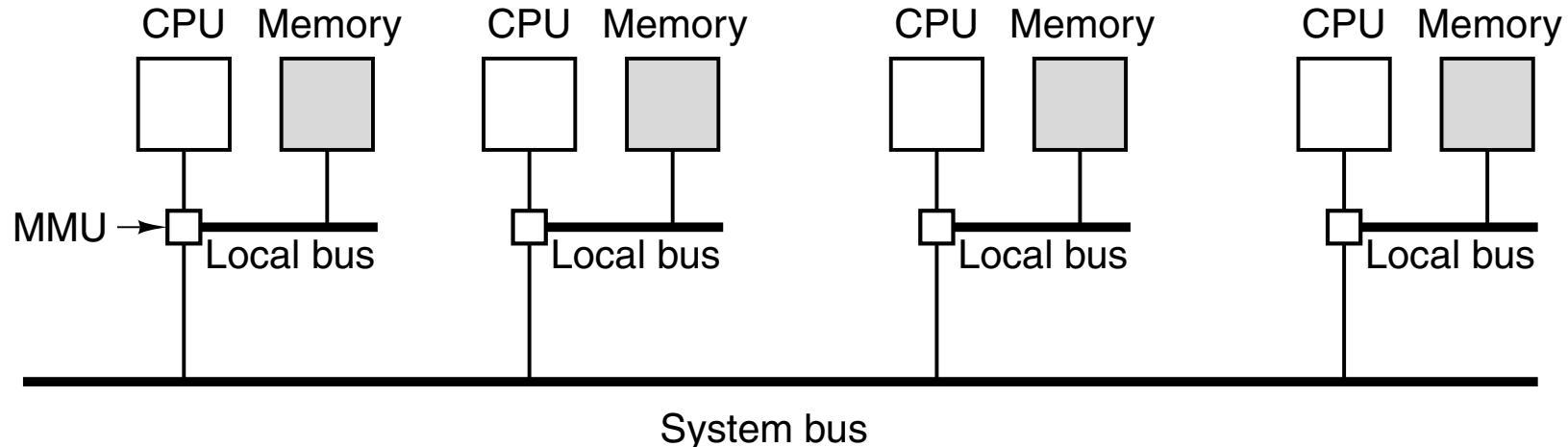


## SMP (tiếp)

- Một máy tính có  $n \geq 2$  bộ xử lý giống nhau
- Các bộ xử lý dùng chung bộ nhớ và hệ thống vào-rà
- 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



# NUMA (Non-Uniform Memory Access)



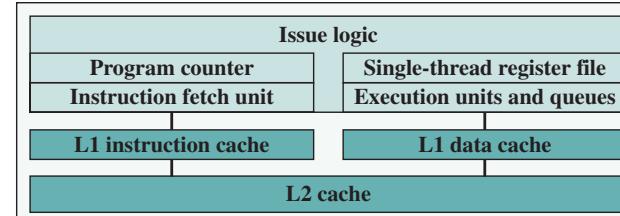
- Có một không gian địa chỉ chung cho tất cả CPU
- Mỗi CPU có thể truy cập từ xa sang bộ nhớ của CPU khác
- Truy nhập bộ nhớ từ xa chậm hơn truy nhập bộ nhớ cục bộ



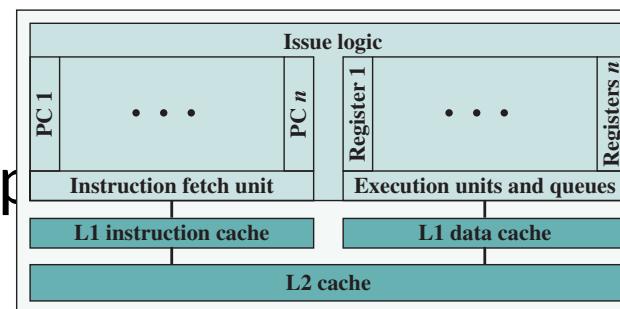
# Bộ xử lý đa lõi (multicores)

## ▪ Thay đổi của bộ xử lý:

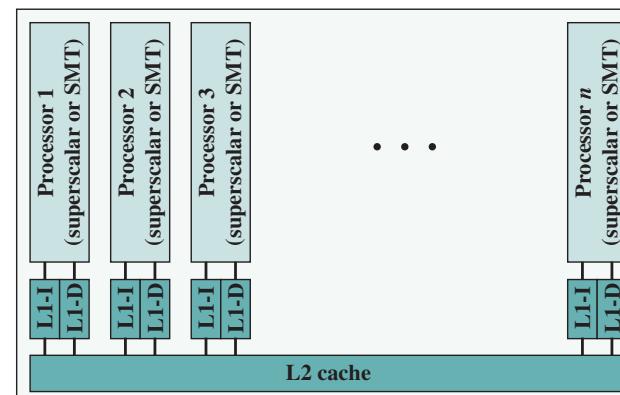
- Tuần tự
- Pipeline
- Siêu vô hướng
- Đa luồng
- Đa lõi: nhiều CPU trên một chip



(a) Superscalar

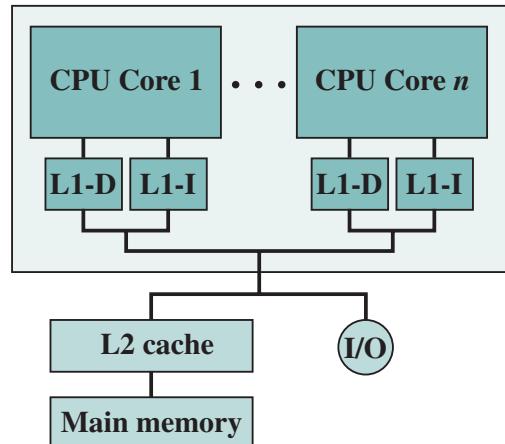


(b) Simultaneous multithreading

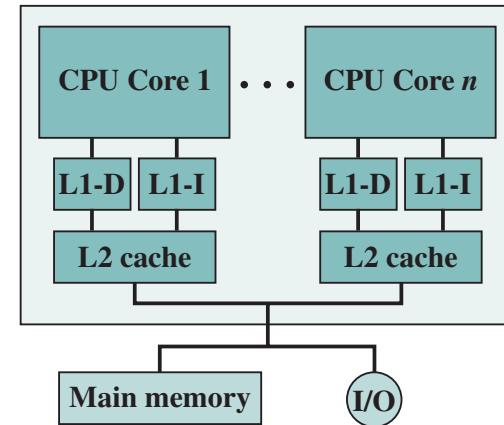


(c) Multicore

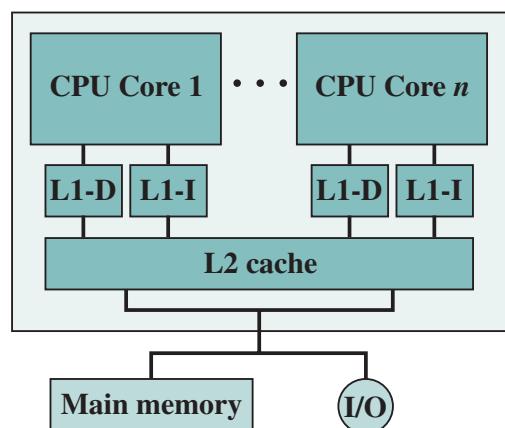
# Các dạng tổ chức bộ xử lý đa lõi



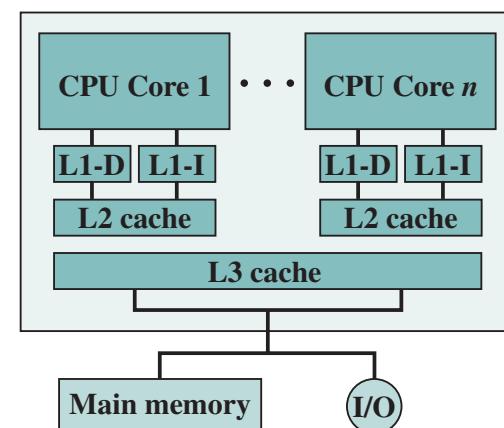
(a) Dedicated L1 cache



(b) Dedicated L2 cache



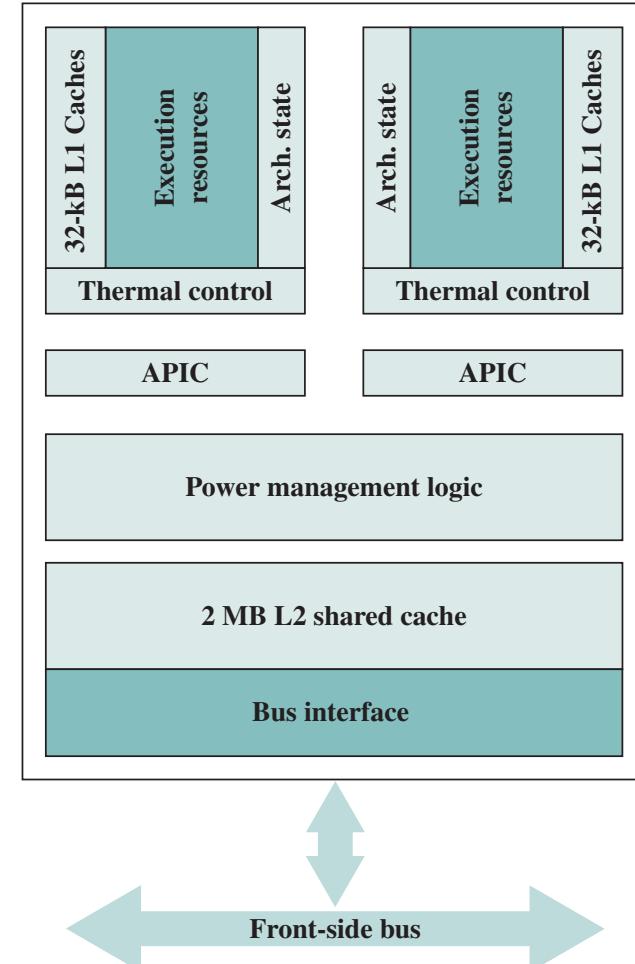
(c) Shared L2 cache



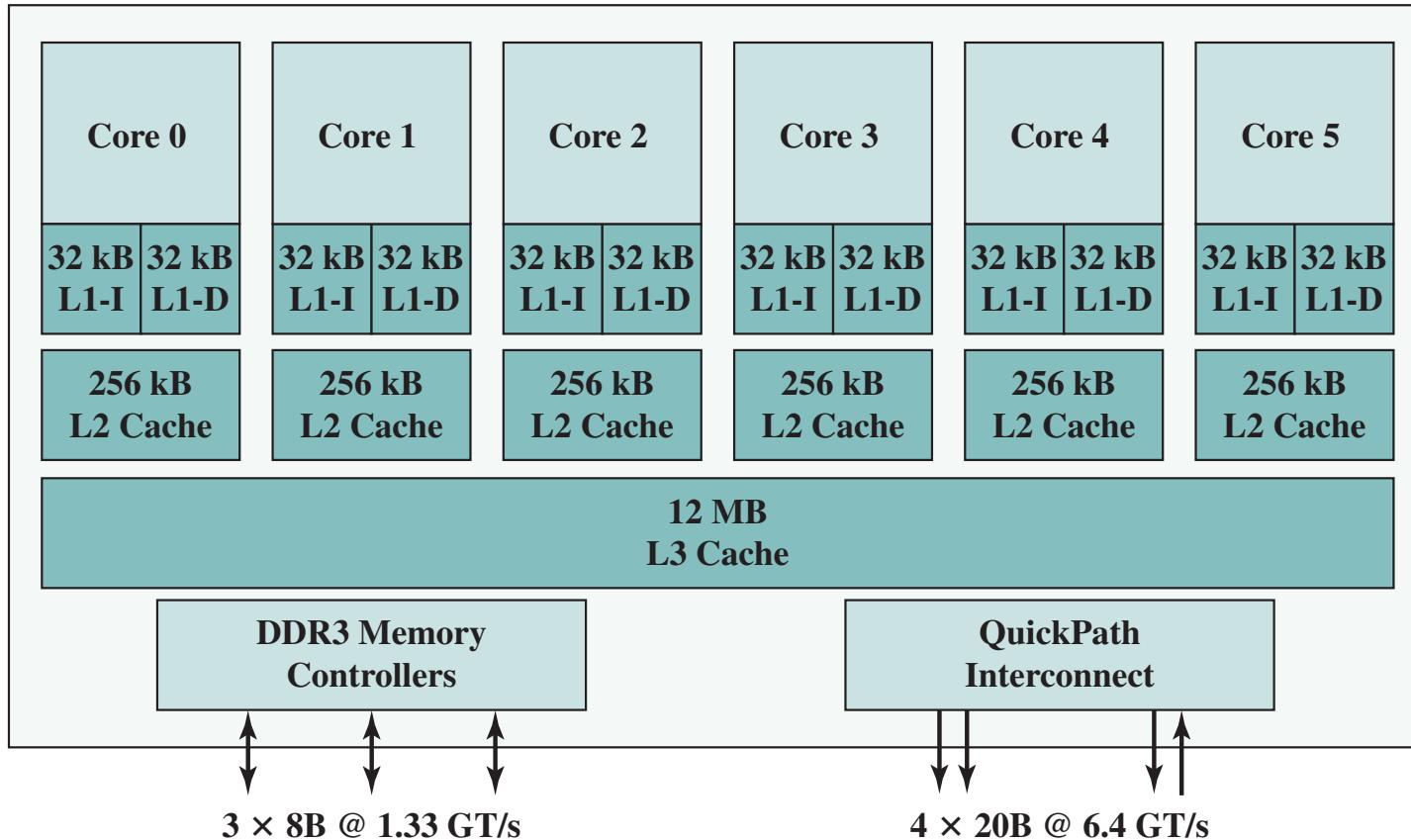
(d) Shared L3 cache

# Intel - Core Duo

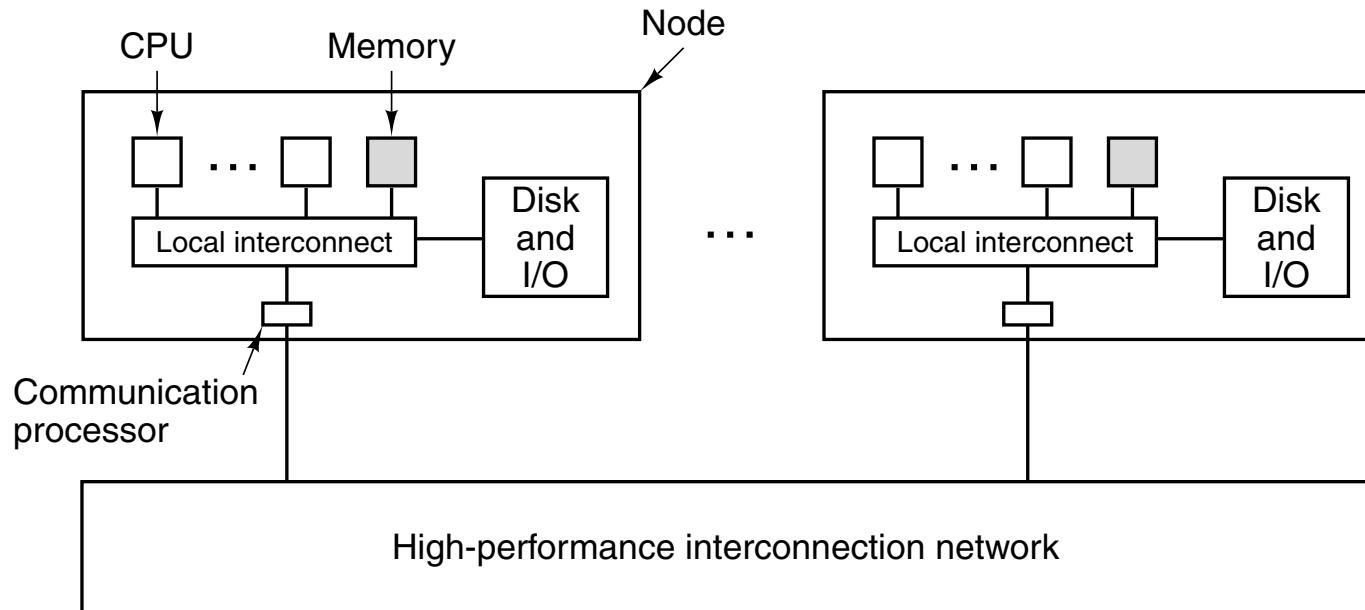
- 2006
- Two x86 superscalar, shared L2 cache
- Dedicated L1 cache per core
  - 32KiB instruction and 32KiB data
- 2MiB shared L2 cache



# Intel Core i7-990X

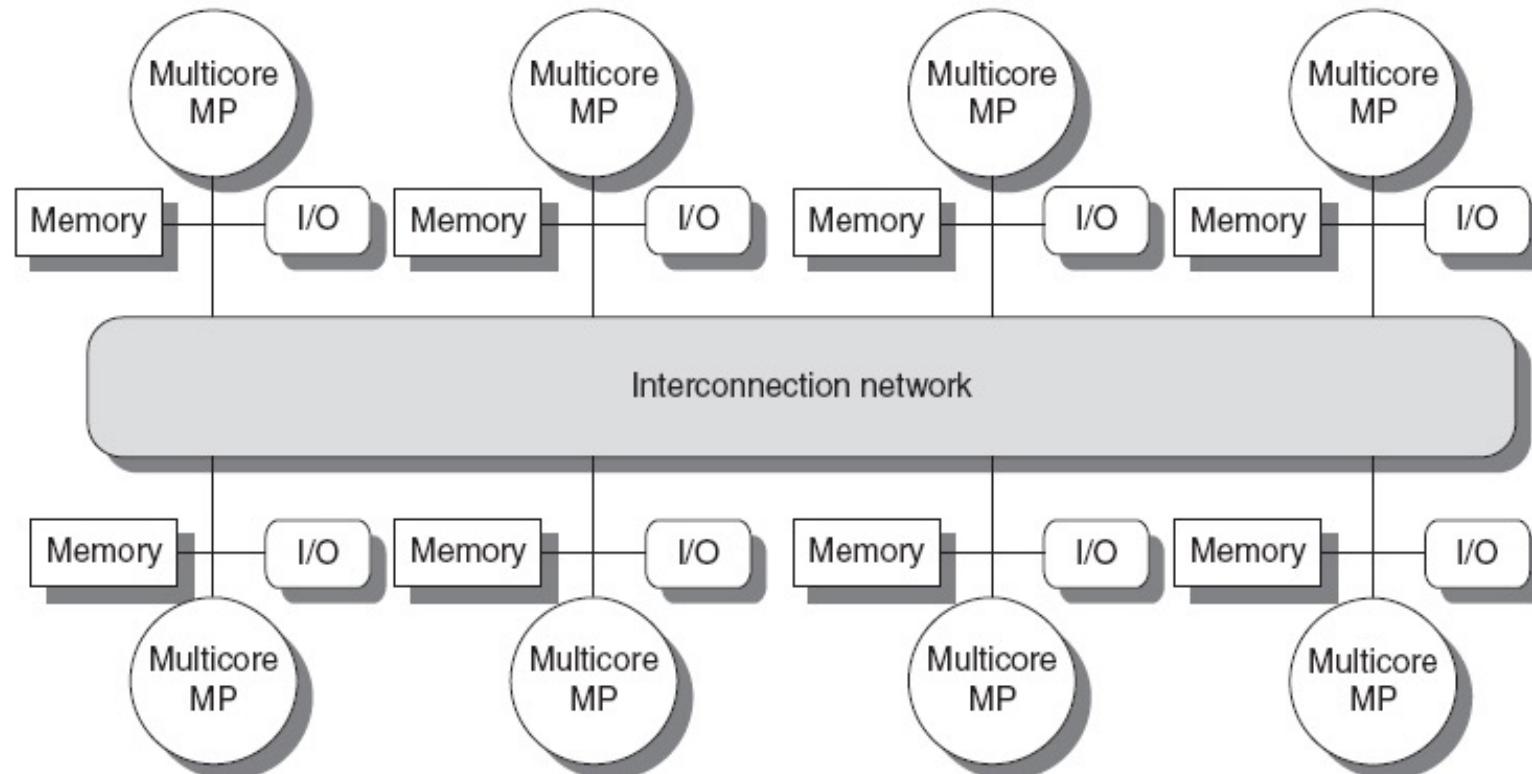


## 8.3. Đa xử lý bộ nhớ phân tán

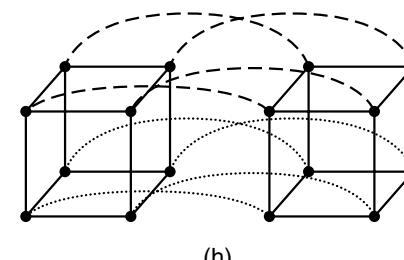
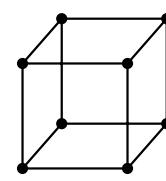
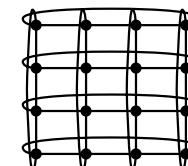
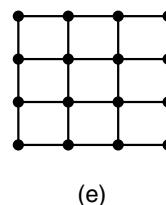
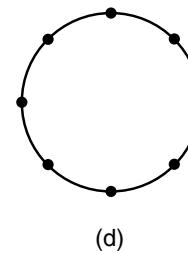
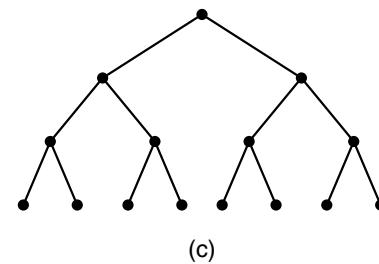
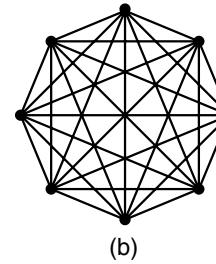
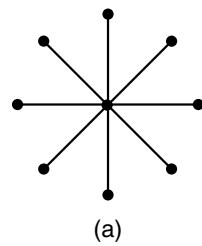


- Máy tính qui mô lớn (Warehouse Scale Computers or Massively Parallel Processors – MPP)
- Máy tính cụm (clusters)

# Đa xử lý bộ nhớ phân tán



# Mạng liên kết

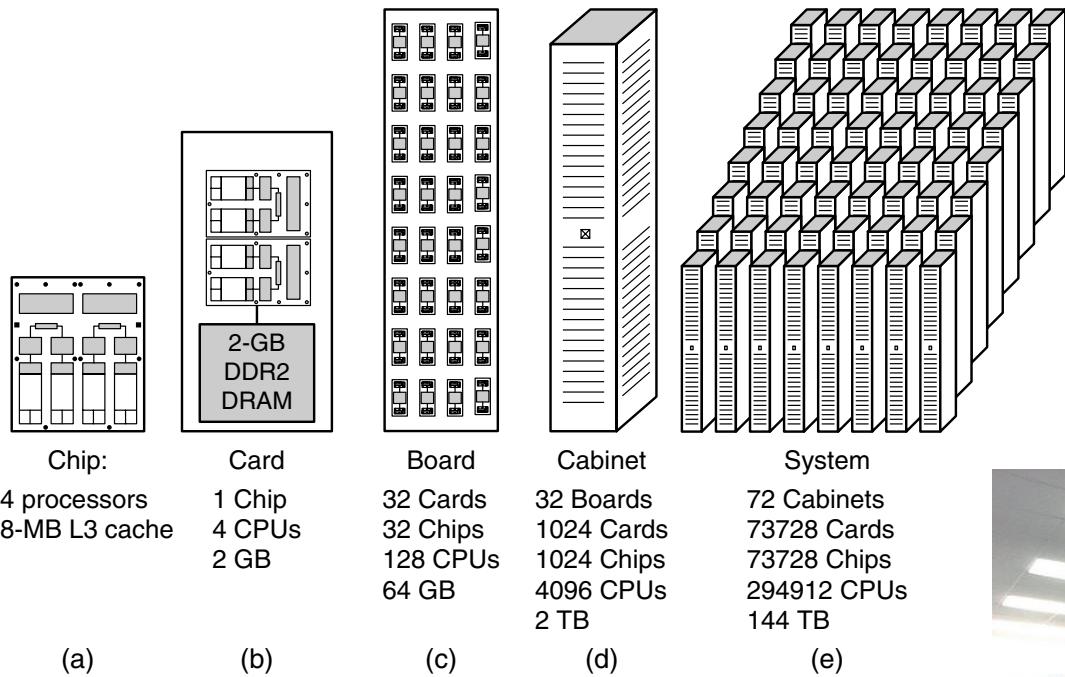


# Massively Parallel Processors

- 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

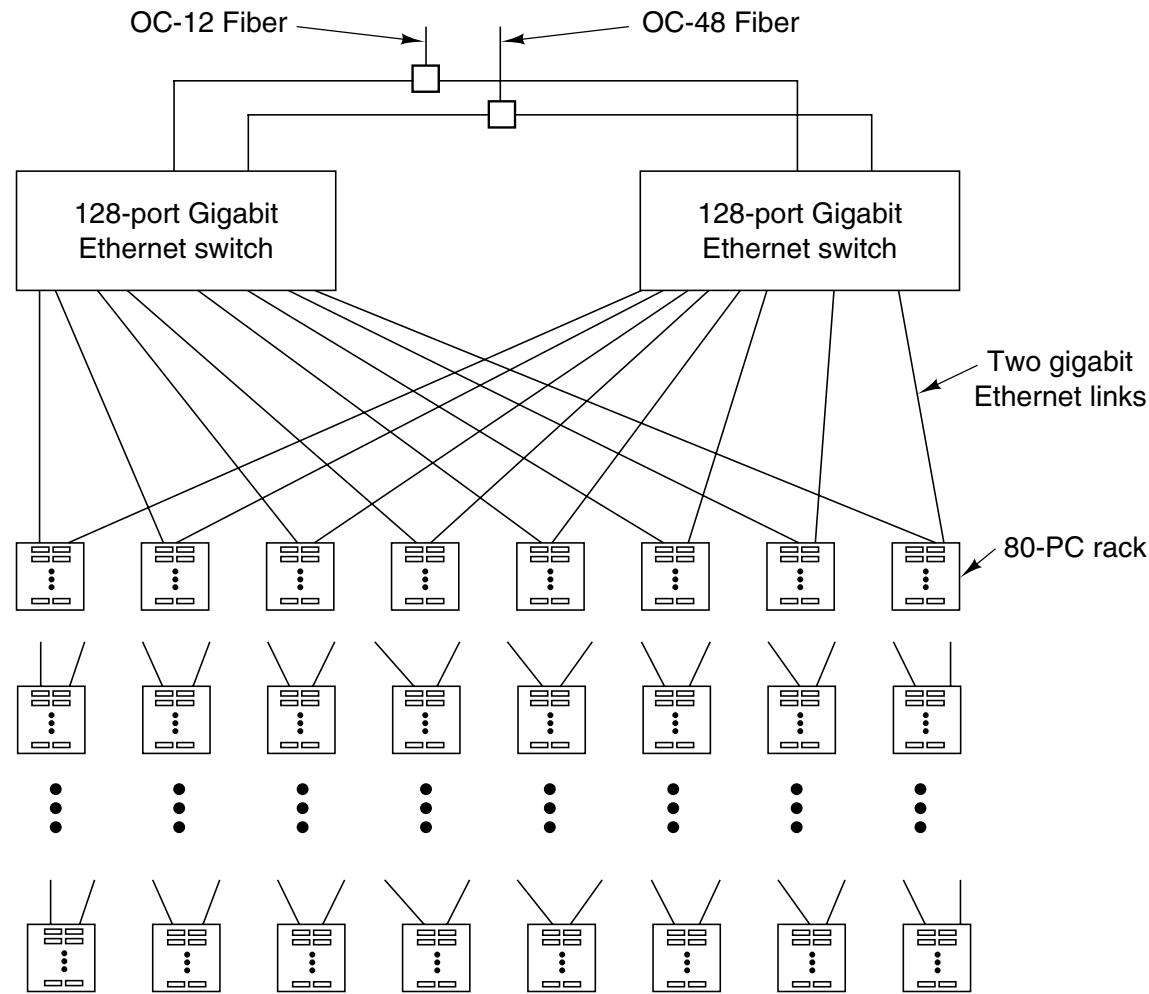


# Cluster

- 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



# PC Cluster của Google

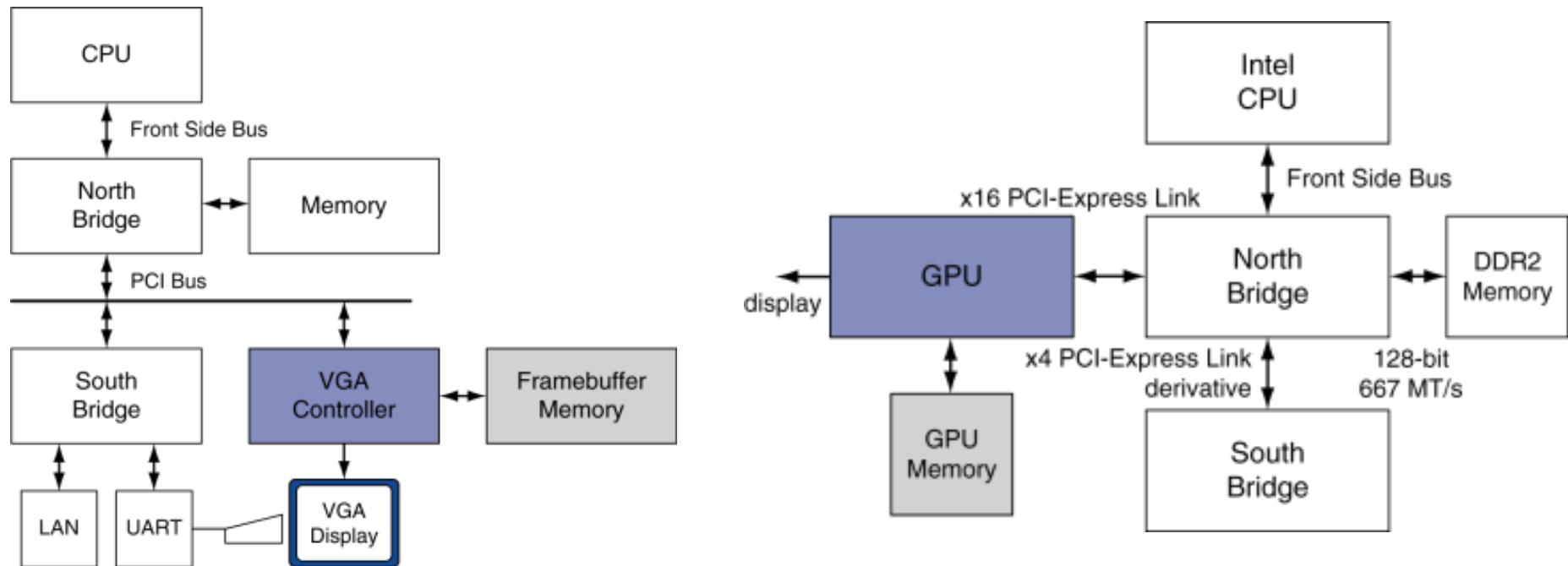


## 8.4. Bộ xử lý đồ họa đa dụng

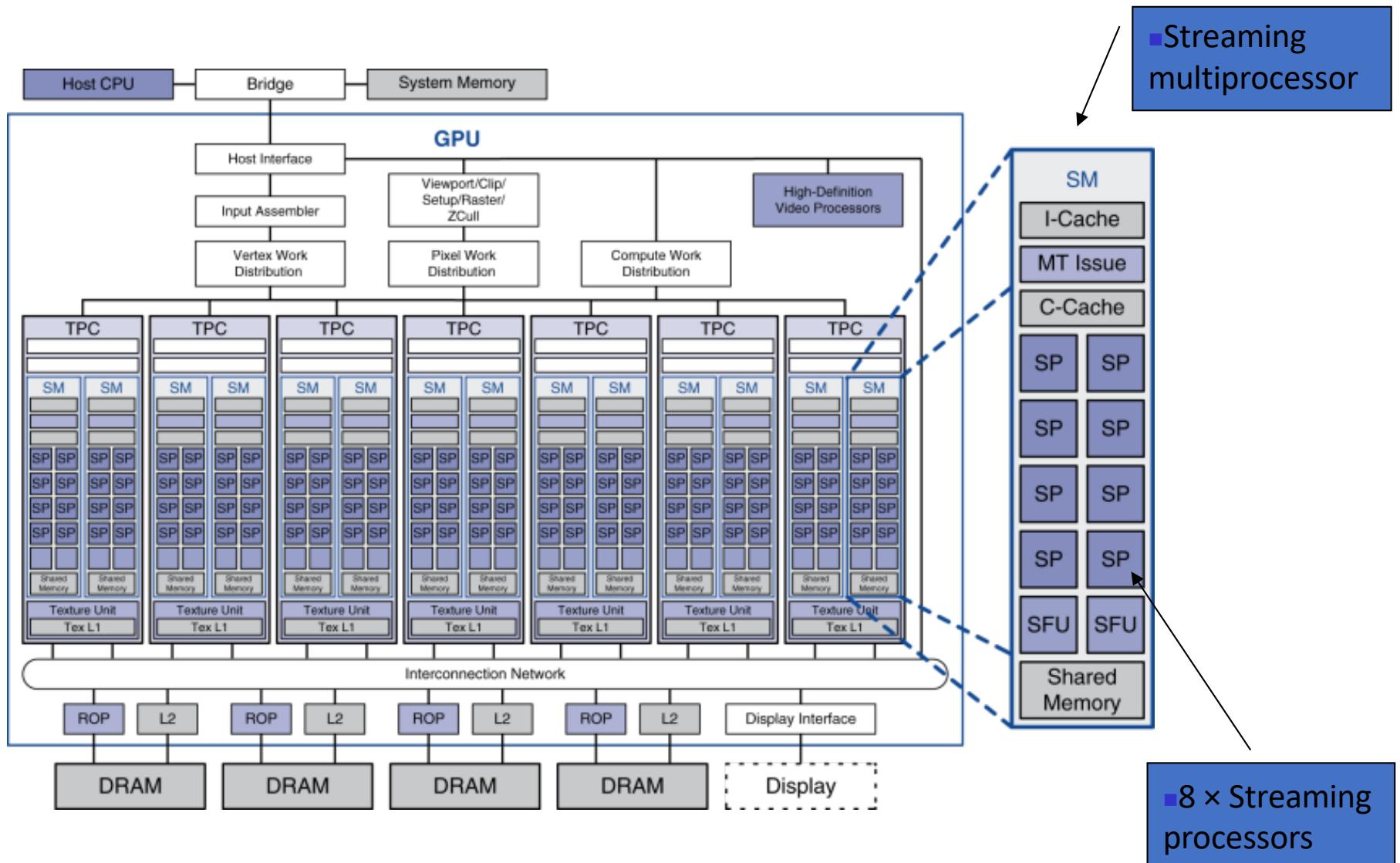
- Kiến trúc SIMD
- Xuất phát từ bộ xử lý đồ họa GPU (Graphic Processing Unit) hỗ trợ xử lý đồ họa 2D và 3D: xử lý dữ liệu song song
- GPGPU – General purpose Graphic Processing Unit
- Hệ thống lai CPU/GPGPU
  - CPU là host: thực hiện theo tuần tự
  - GPGPU: tính toán song song



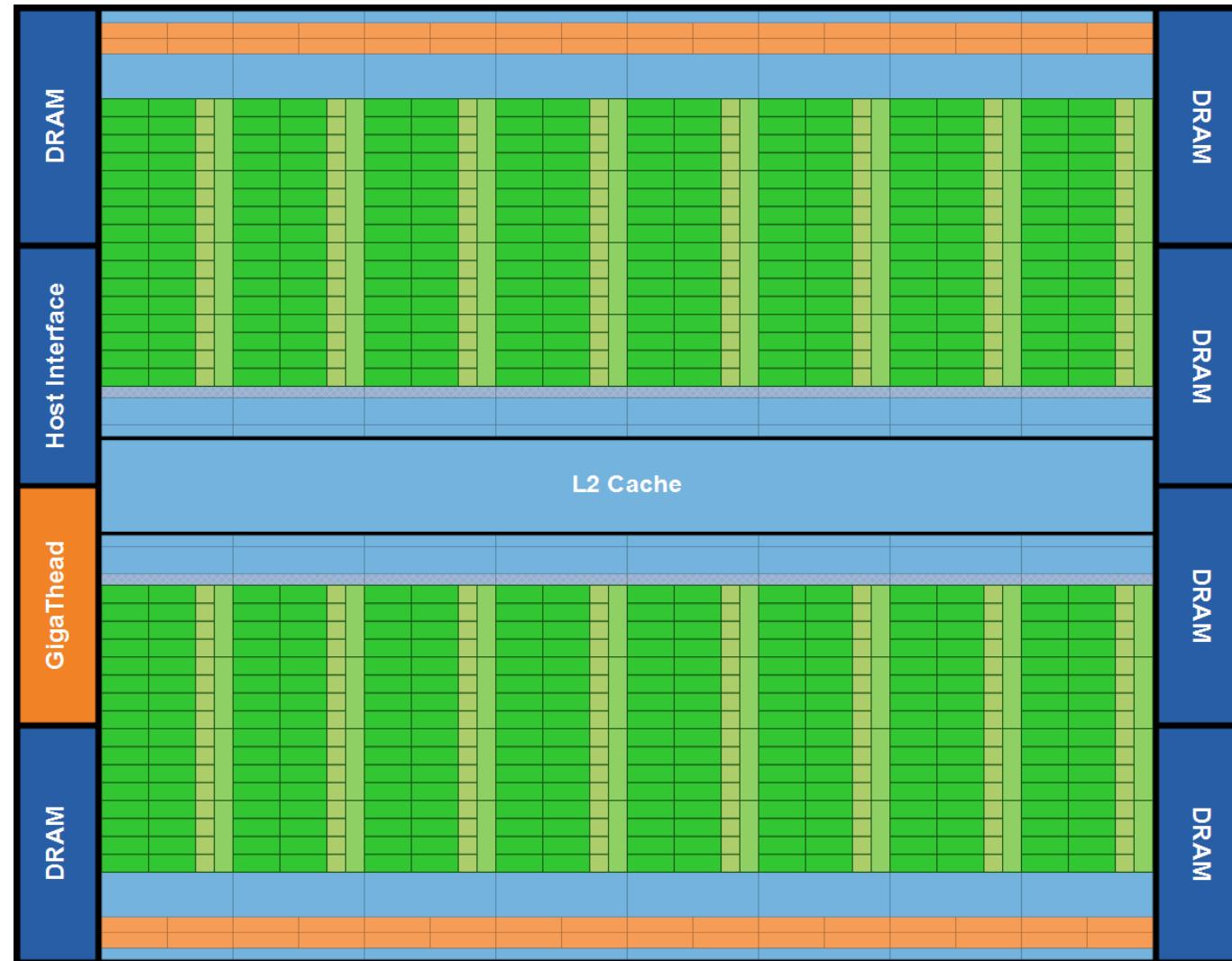
# Bộ xử lý đồ họa trong máy tính



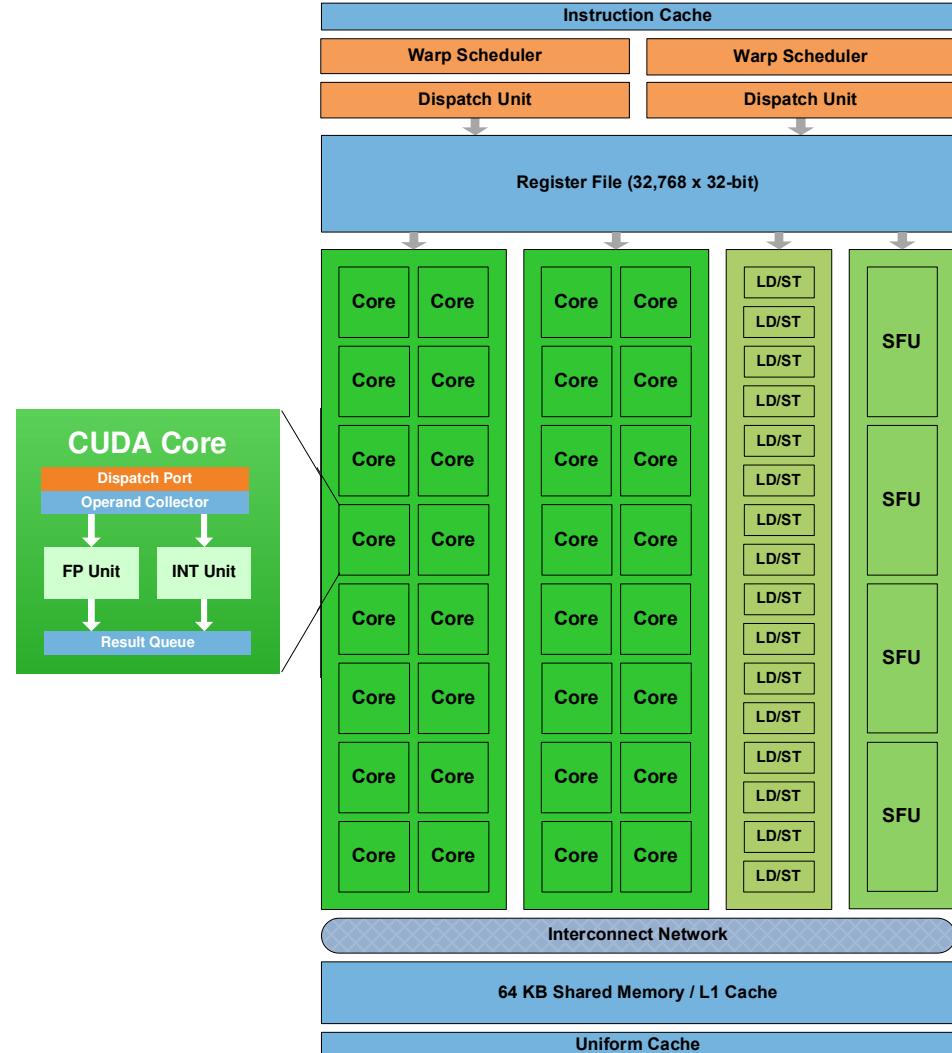
# GPGPU: NVIDIA Tesla



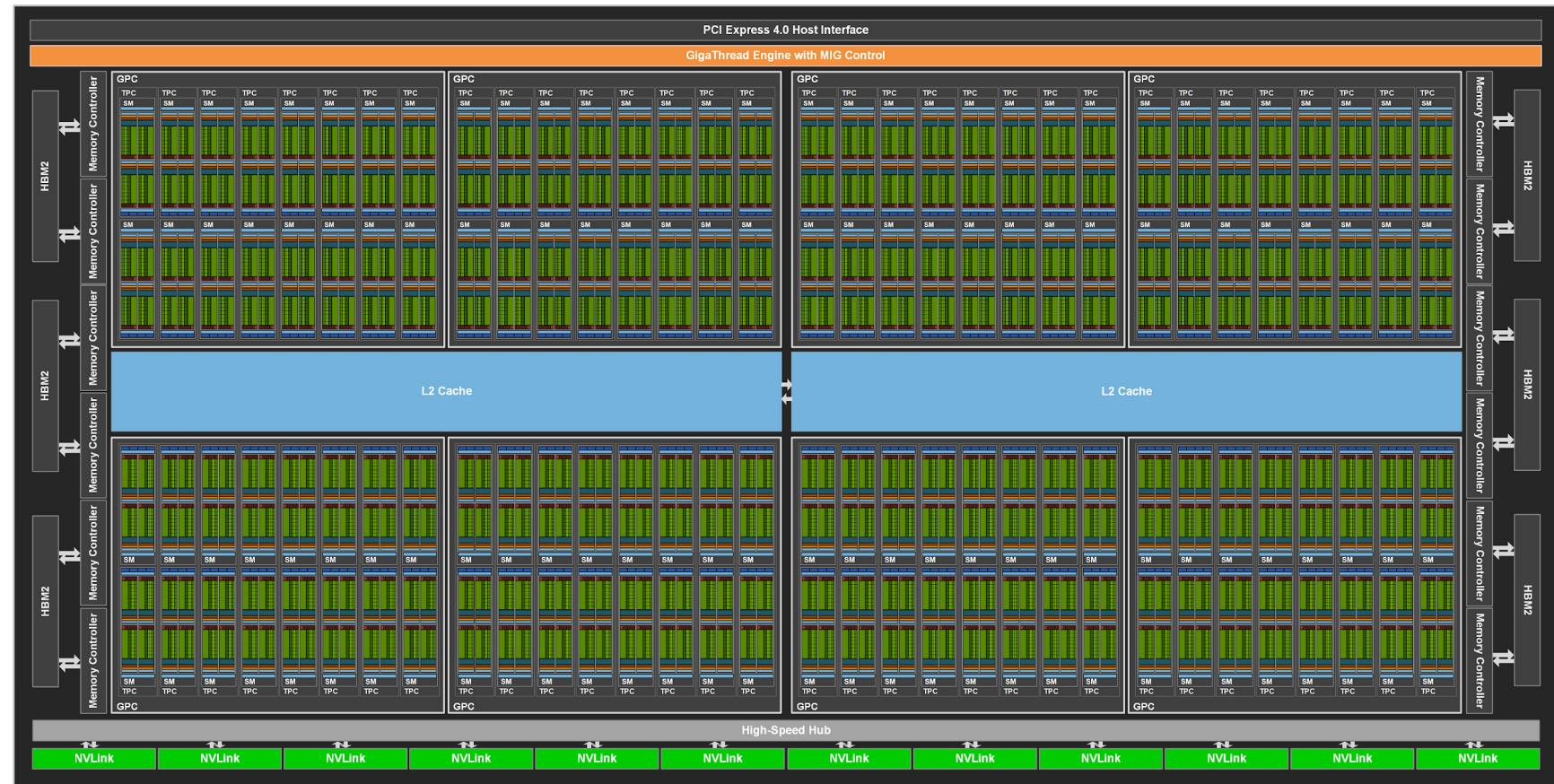
# GPGPU: NVIDIA Fermi



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



# NVIDIA A100 Tensor Core GPU Architecture



# GA100 Streaming Multiprocessor



## Hết chương 8

