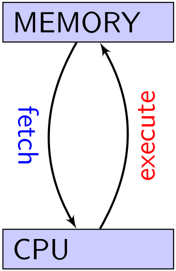
Lecture 5:

Parallel, Distributed, Mobile, Social Computing

Faculty of Computing & Informatics
Multimedia University

von Neumann Architecture

- Common machine model
  – for over 40 years
- Stored-program concept
- CPU executes a stored program
- A sequence of read and write operations on the memory (RAM)
- Order of operations is sequential



CISC von Neumann Computer

- CISC (Complex Instruction Set Computer)
  – with a single bus system
- Harvard (RISC) architecture utilizes two buses
  – separate data bus and address bus
- RISC (Reduced Instruction Set Computer)
- They are SISD machines
  – Single Instruction Stream Single Data Stream

Motivations for Parallel Computing

- Fundamental limits on single processor speed
- Disparity between CPU & memory speed
  – Performance Mismatch Problem
- Distributed data communications
- Need for very large scale computing platforms

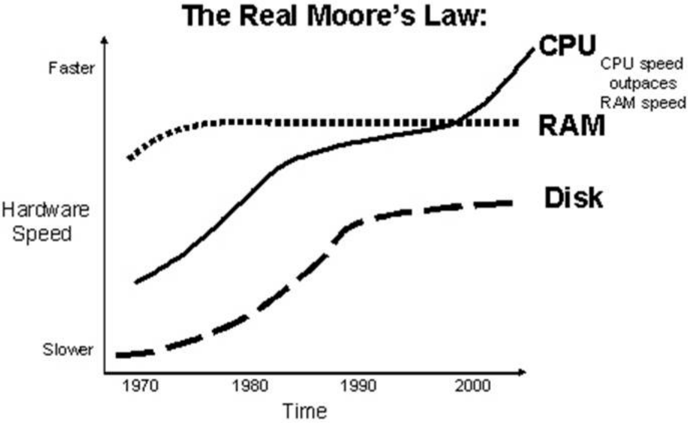
Moore’s Law

- Moore’s observation in 1965:
  – number of transistors per square inch on integrated circuits had doubled every year
- Moor’s revised observation in 1975:
  – the space slowed down a bit, but data density had doubled approximately every 18 months
- How about the future?
  – (price of computing power falls by a half every 18 months?)

CPU and Memory Speeds

- In 20 years, CPU speed (clock rate) has increased by a factor of 1000
- DRAM speed has increased only by a factor of smaller than 4
- How to feed data faster enough to keep CPU busy?
- CPU speed: 1-2 ns
- DRAM speed: 50-60 ns
- Cache: 10 ns

CPU, Memory, and Disk Speed



Why Use Parallel Computing

- Save time - wall clock time - many processors work together
- Solve larger problems - larger than one processor's CPU and memory can handle
- Provide concurrency - do multiple things at the same time: online access to databases, search engine
- Google's 4,000 PC servers are one of the largest in clusters the world

Other Reasons for Parallel Computing

- Taking advantages of non-local resources → using computing resources on a wide area network, or even internet (grid & cloud computing)
  – Remote Access Resources
- Cost savings → using multiple "cheap" computing resources instead of a high-end CPU
- Overcoming memory constraints → for large problems, using memories of multiple computers may overcome the memory constraint obstacle

Types of Parallel Computing Models

- Data Parallel – the same instructions are carried out simultaneously on multiple data items (SIMD)
- Task Parallel - different instructions on different data (MIMD)
  1. SPMD (single program, multiple data) not synchronized at individual operation level
  2. SPMD is equivalent to MIMD since each MIMD program can be made SPMD (similarly for SIMD, but not in practical sense.)

Other Distributed Computing Platforms

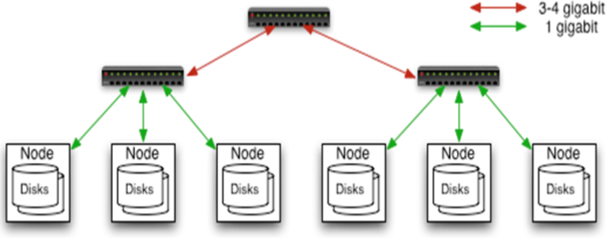
- MapReduce/Hadoop
- Android

What Is hadoop?

- Distributed computing frame work
  - For clusters of computers
  - Thousands of Compute Nodes
  - Petabytes of data
- Open source, Java
- Google's MapReduce inspired Yahoos Hadoop.
- Now part of Apache group

What is Hadoop?

Commodity Hardware



Typically in 2 level architecture:

- Nodes are commodity PCs
- 30-40 nodes/rack
- Uplink from rack is 3-4 gigabit
- Rack-internal is 1 gigabit

What is Hadoop?

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. Hadoop includes:

- Hadoop Common utilities
- Avro: A data serialization system with scripting languages.
- Chukwa: managing large distributed systems.
- HBase: A scalable, distributed database for large tables.
- HDFS: A distributed file system.
- Hive: data summarization and ad hoc querying.
- MapReduce: distributed processing on compute clusters.
- Pig: A high-level data-flow language for parallel computation.
- ZooKeeper: coordination service for distributed applications.

Who uses Hadoop?

Amazon, Facebook, Google, IBM, Joost,Last.fm, New York Times, PowerSet,Veoh, Yahoo!

What is MapReduce used for?

- At Google:
  1. Index construction for Google Search
  2. Article clustering for Google News
  3. Statistical machine translation
- at Yahoo!:
  1. "Web map" powering Yahoo! Search
  2. Spam detection for Yahoo! Mail
- at Facebook:
  1. data mining
  2. Ad optimization
  3. Spam detection

Social Computing: Social systems

- Social network: graph that represents relationships between independent agents.
- Social networks are everywhere and are shaping our lives:
  1. Network of professional contacts (e.g., for finding jobs)
  2. Network of colleagues (e.g., for learning new techniques)
  3. Web 2.0 systems:
     - Online social networks: facebook, myspace, orkut, IM, linkedIn, twitter,
     - Content sharing: flickr, del.icio.us, youtube, weblogs,
     - Content creation: wikipedia,