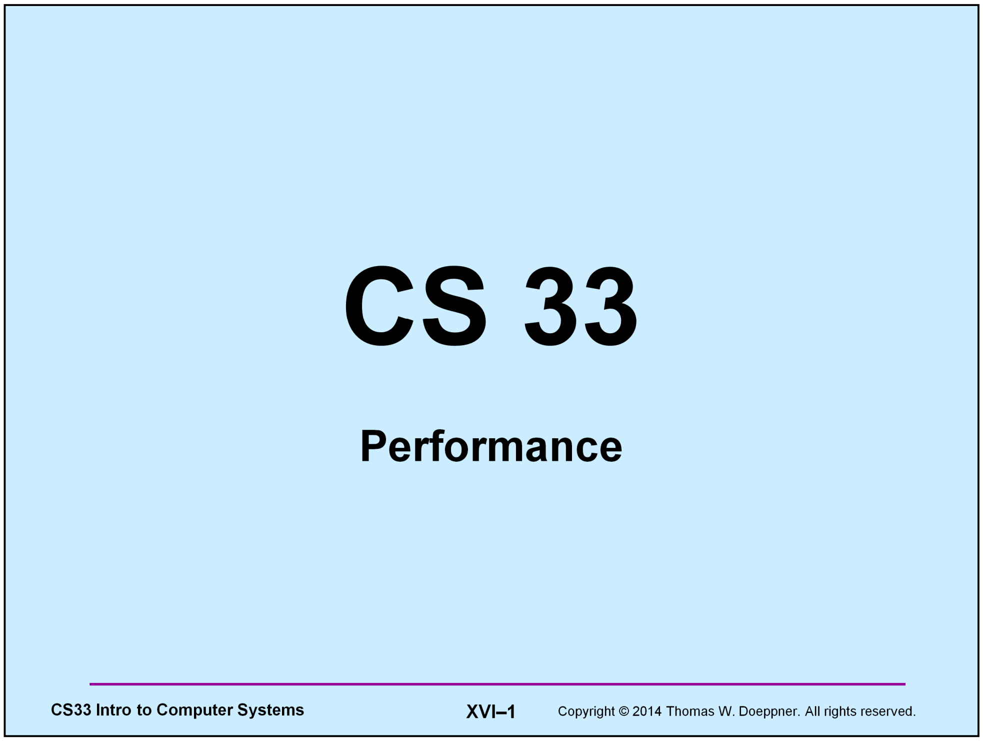


# CS 33

## Performance

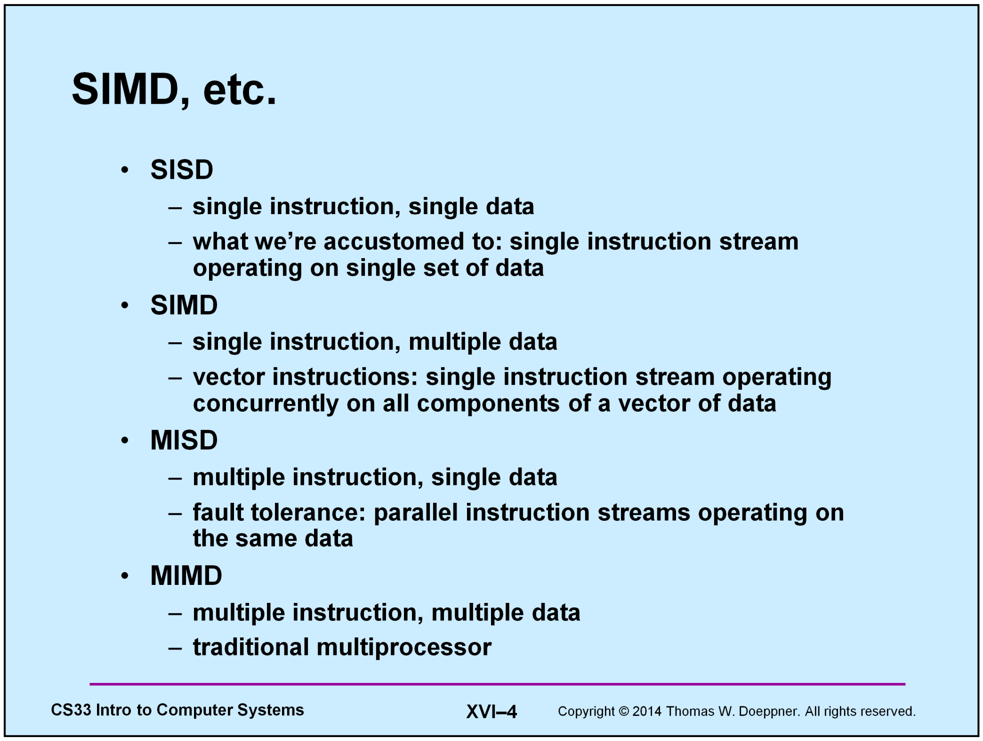
# Floating Point

- 8086 had no floating point instructions
  - – simulated in software

  or
  - – used separate 8087 *coprocessor*
    - » x87 instruction set
- 80486 and later processors subsume x87
  - – additional register set (8 FP registers)

# Multimedia Extensions (MMX)

- Vector integer arithmetic instructions
  - eight new register names (mm0 – mm7), but aliased to floating point registers
  - supports SIMD parallelism
    - for integers
- Introduced with Pentium processors in 1997



## SIMD, etc.

- SISD
  - single instruction, single data
  - what we’re accustomed to: single instruction stream operating on single set of data
- SIMD
  - single instruction, multiple data
  - vector instructions: single instruction stream operating concurrently on all components of a vector of data
- MISD
  - multiple instruction, single data
  - fault tolerance: parallel instruction streams operating on the same data
- MIMD
  - multiple instruction, multiple data
  - traditional multiprocessor

This taxonomy was introduced by Michael Flynn in 1966.

# Streaming SIMD Extensions (SSE)

- Extends x86 SIMD support to floating point
- Adds additional control over caching
- Introduced with Pentium III in 1999
- Adds completely new set of 128-bit registers
  - XMM0 – XMM7 (32-bit architectures)
  - XMM0 – XMM15 (64-bit architectures)
- Further ongoing extensions
  - SSE2
  - SSE3
  - SSE4

# XMM registers

vector of four floats

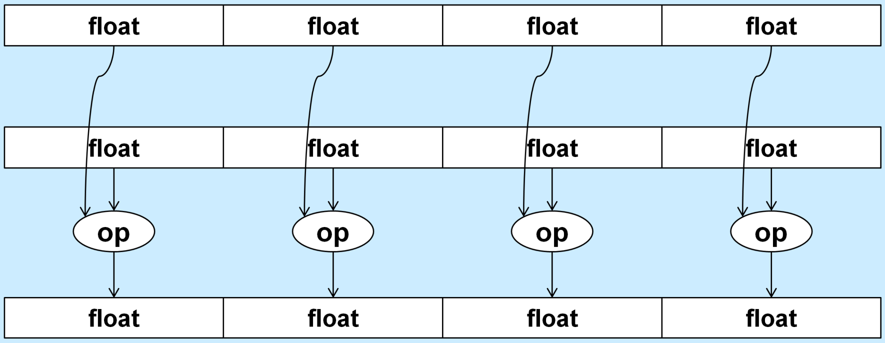
| | | | |
|---|---|---|---|
| float | float | float | float |

or

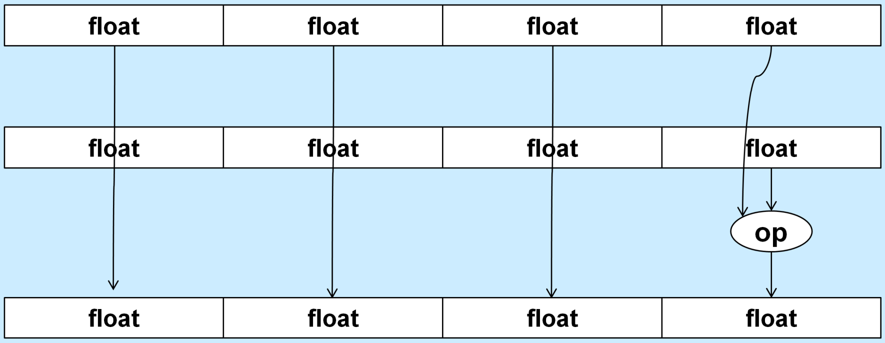
vector of two doubles

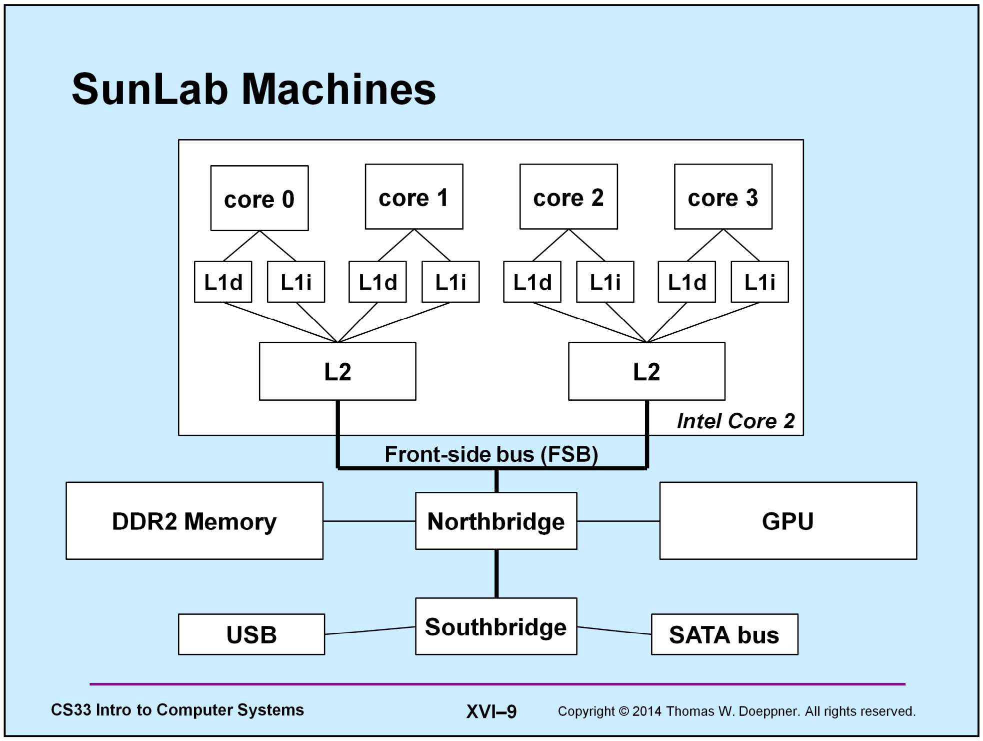
| | |
|---|---|
| double | double |

# Vector Operations



# Scalar Operations

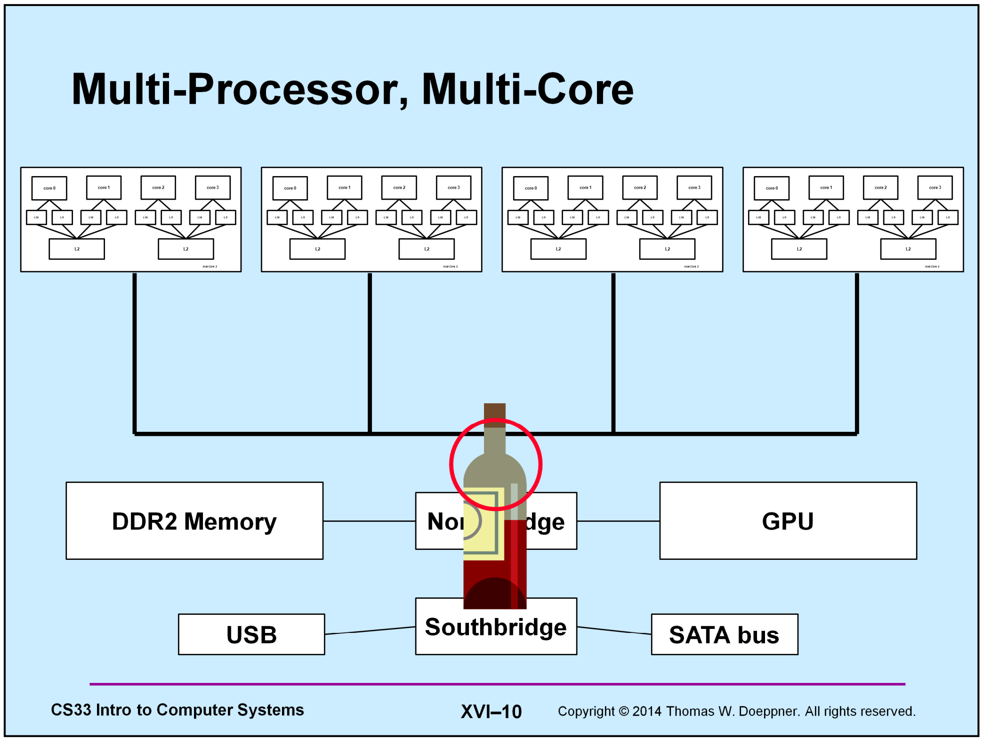




## SunLab Machines

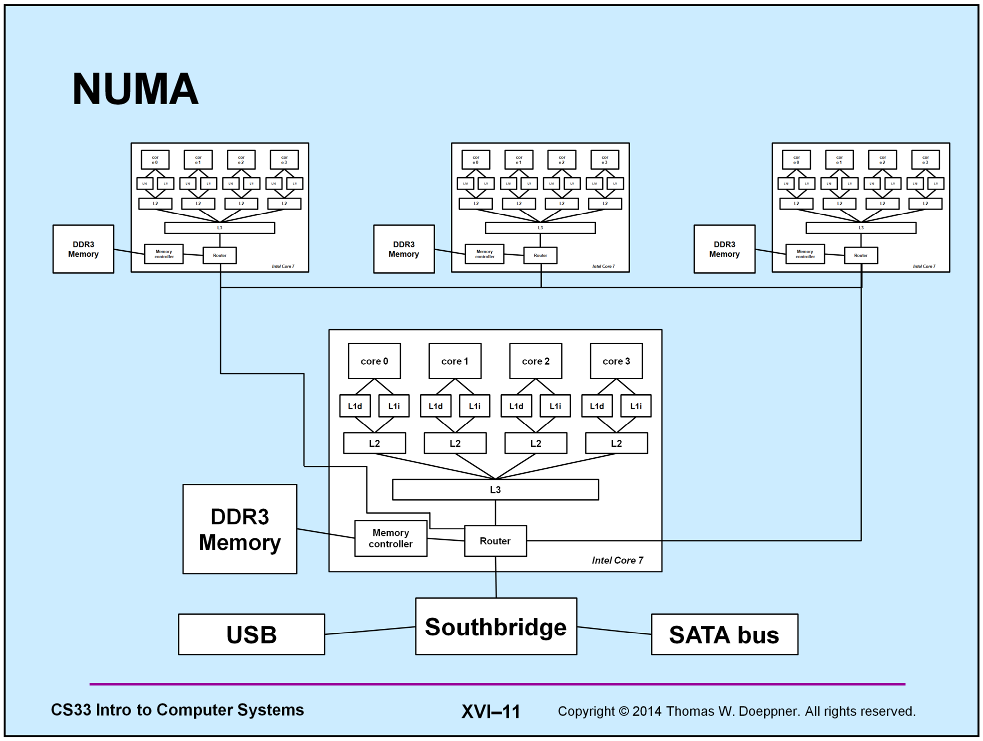
The SunLab machines are powered by Intel Core 2 quad-core processors, running at 2.4 GHz. Each of the L1 data and L1 instruction caches holds 32 k bytes; the L1 data-cache line size is 64 bytes. The L2 caches are 4 M bytes each. The front-side bus, over which all data and instructions from and to memory travel, is 64 bits wide and runs at 1066 MHz, which means, roughly, that the peak transfer rate is 8 GB/sec. But this is not achieved for longer than very brief intervals. The DDR2 memory cannot stream data that quickly, and much of this bandwidth is used for transmitting address information. Thus the effective bandwidth is far less.

Northbridge and Southbridge are standard names for the two parts of the chipset on the motherboard of Intel PCs and servers (though more modern systems have a different arrangement that we cover next). The Northbridge contains the memory controller and thus memory is directly attached to it, as is the graphics processing unit (GPU). Everything else is connected via the Southbridge, in particular the USB (to which keyboard and mouse are connected) and the SATA bus (to which the disk is connected).



# Multi-Processor, Multi-Core

If we try to enhance our system by adding multiple core-2 processors, we run into a problem because all cores on all processors share the bandwidth of the front-side bus as well as the bandwidth of the memory system. Furthermore, the DMA disk devices also compete for Northbridge and memory bandwidth.



# NUMA

More recent designs put a memory controller inside of each processor chip, with separate memory attached to each chip. Thus access to the attached memory is relatively quick. To reach memory attached to other chips from a particular chip, the request must be routed through one or more other processor chips. Thus accessing other memory is more time-consuming than accessing local memory. Such systems are referred to as *non-uniform memory access* (NUMA) systems. What's shown here is the Intel Core 7 with Intel's *QuickPath Interconnect*. More information can be found at http://www.intel.com/content/www/us/en/io/quickpath-technology/quick-path-interconnect-introduction-paper.html.