CS 451 / 551 / ECE 541

ADVANCED
COMPUTER ARCHITECTURE

SESSION no. 22

University of Idaho

EXAM #2

IN CLASS

OPEN BOOK & NOTES

100 PTS, 1 HR + 15 MINS.

FORMAT LIKE EXAM #1

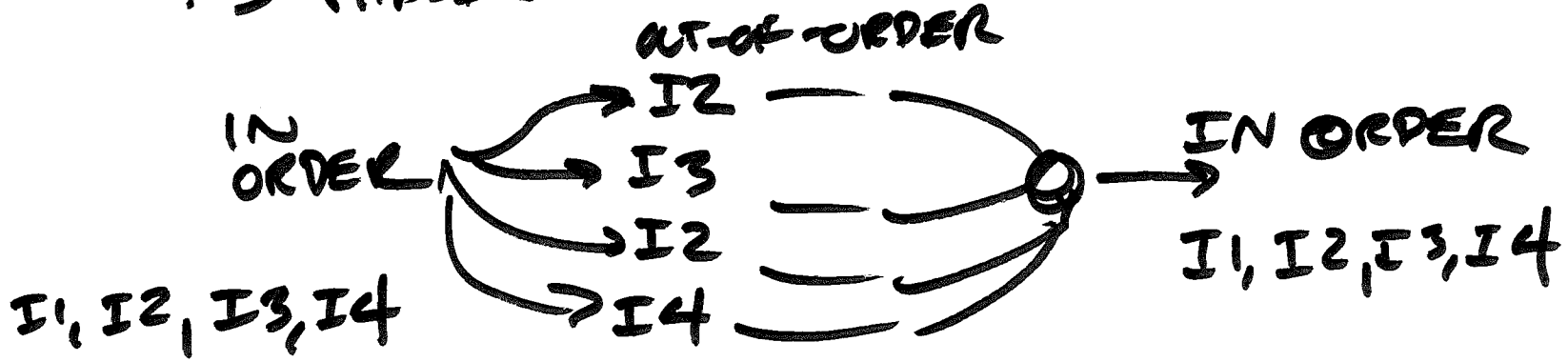
EMPHASIS - DATA PARALLELISM (SIMD)

- CH4 TEXT
- STUDY NOTES
  - LECTURES POSTED ON CLASS WEB SITE

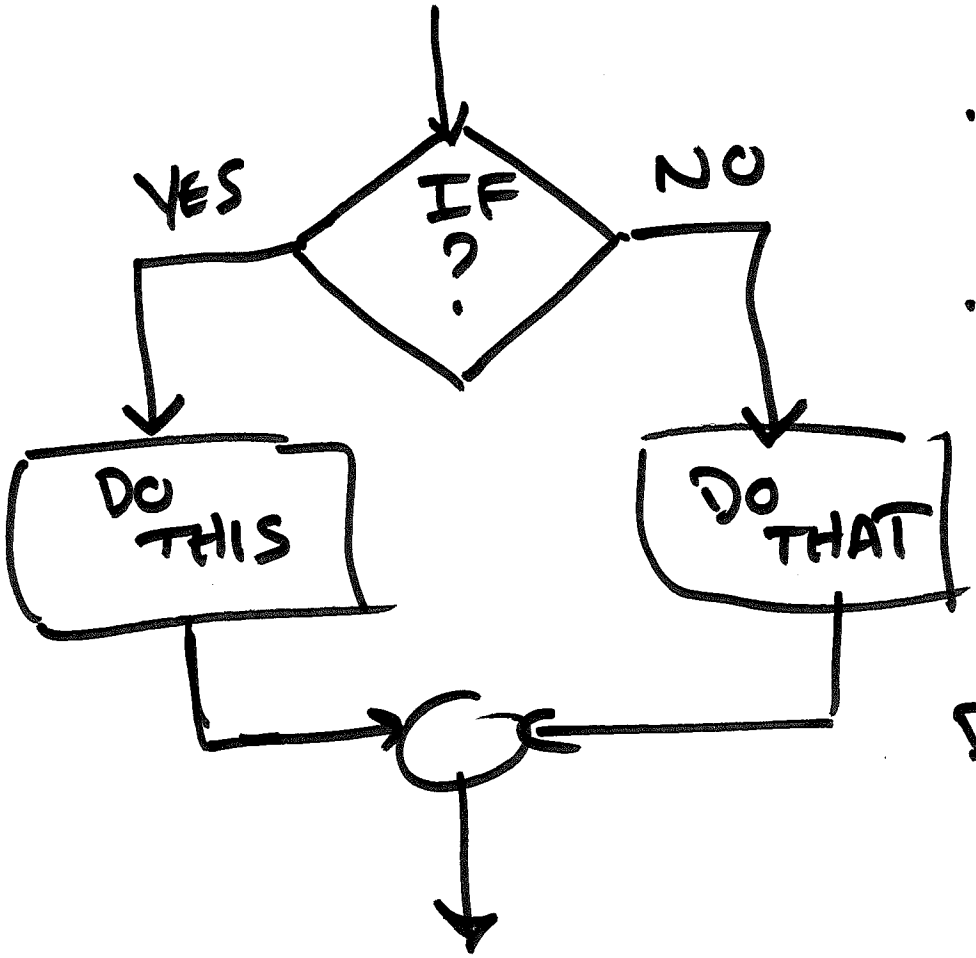
THREAD LEVEL PARALLELISM (TASIC) - MULTITASKING

BUT FIRST— MULTIPLE ISSUE I.L.P.

- MULTIPLE EXECUTION UNITS
  INTEGER, F. P., LOAD/STORE
- 3 PHASES



SPECULATIVE EXECUTION



- AVOID BRANCH HAZARDS
- EXECUTE BOTH BRANCHES: "THIS" & "THAT"
- COMMIT UNIT DECIDES WHICH ONE WAS CORRECT

MULTITHREADING ADVANTAGES

(CONCURRENCY)

- KEEP PROCESSOR ELEMENTS BUSY (WORKING)
- PROCESSOR STALLS - SWITCH TO ANOTHER THREAD
  - SHORT STALLS - FEW CLOCK CYCLES
    - CACHE MISS (1L).
    - HAZARDS -
  - LONG STALLS - MANY CLOCK CYCLES
    - VM PAGE FAULT - 10^6 CYCLES
    - I/O REQUEST, USER INPUT ∞ CLOCK CYCLES

University of Idaho

LONG STALLS. PROCESS SWITCH

SHORT STALLS - THREAD SWITCH

. PROGRAMMER USED THREADS!

~~HARDWARE SUPPORT~~

GRANULARITY

. HOW MANY INSTRUCTIONS IN A "CHUNK" THAT GET EXECUTED BEFORE THREAD SWITCH?

. COARSE - RUN THREAD FOR A WHILE, STOP, RUN ANOTHER

. FINE - INTERLEVE INSTRUCTIONS FROM A THREAD.

PROCESS

- OWN MEMORY, OWN CONTEXT
- PROCESS SWITCH: SAVE A LOT OF "STATE"
  - PC, REG, COND. CODES, MEM PROTECTION REGISTERS, CACHE, ...
  - 10^2 - 10^4 INSTRUCTIONS

THREAD - LIGHTWEIGHT PROCESS

- SHARED MEMORY
- SHARE STATE
- THREAD SWITCH: 1-2 CLOCK CYCLES

COMBINE WITH MULTIPLE ISSUE

- INSTRUCTIONS FROM DIFFERENT THREADS INTERLEAVED AMONG EXECUTION UNITS SIMULTANEOUSLY.

("SIMULTANEOUS MULTITHREADING")

- KEEPS ALL PARTS OF PROCESSOR AS BUSY AS POSSIBLE

WITHOUT HARDWARE SUPPORT:

- INCREASE UTILIZATION
- KEEP BUSY WHEN THREADS STALL

WITH H/W SUPPORT

- SUPER-FAST SWITCHING
- INTERLEAVE INSTRUCTIONS

. SIMULTAEOUS MULTITHREADING
MAX. UTILIZATION OF MULTIPLE ISSUE ARCHITECTURE.

OTHER:

- REDUCE LATENCY FOR REQUEST-RESPONSE
- USEFUL SOFTWARE ABSTRACTION
- NATURALLY LEAD TO MULTICORE

LIBRARIES

C++ MULTITHREADING LIBARARIES

- pthreads
- GNU portable threads (POSIX)
- C++ STANDARD LIB,
- BOOST, ROGUE WAVE...

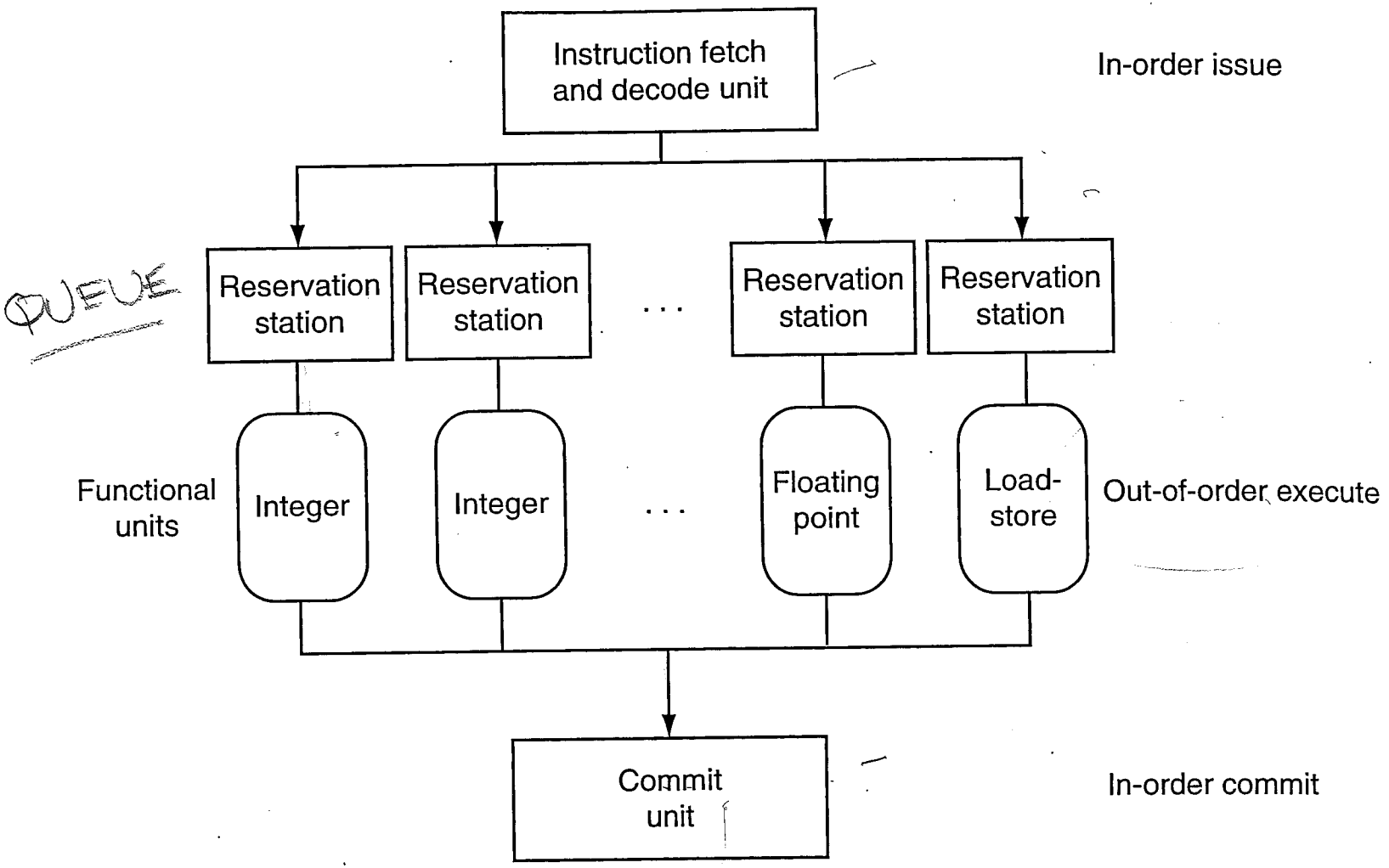


FIGURE 4.72 The three primary units of a dynamically scheduled pipeline. The final step of updating the state is also called retirement or graduation.

~~PR~~ CHALLENGES

- MENTAL SHIFT - CONCURRENCY
- SYNCHRONIZATION - SHARED MEMORY
- DEBUGGING - "HEISENBUGS"

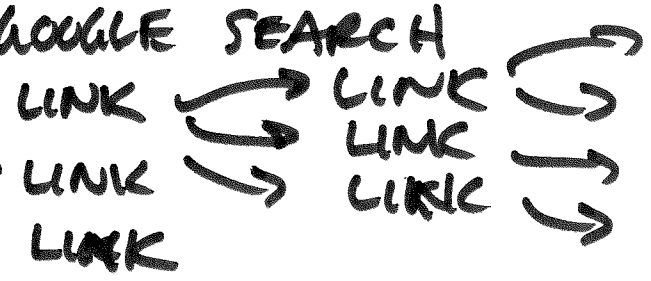
CH5. MOVE TO MULTICORE

2004 - MOORE'S LAW - FREQUENCY SCALING COMES TO END.

MOTIVATION

- LIMITS ILP
- DEMAND
  - BIG SERVERS, CLOUD COMPUTING
- LEARNING HOW TO USE PARALLELISM

Ex GOOGLE SEARCH



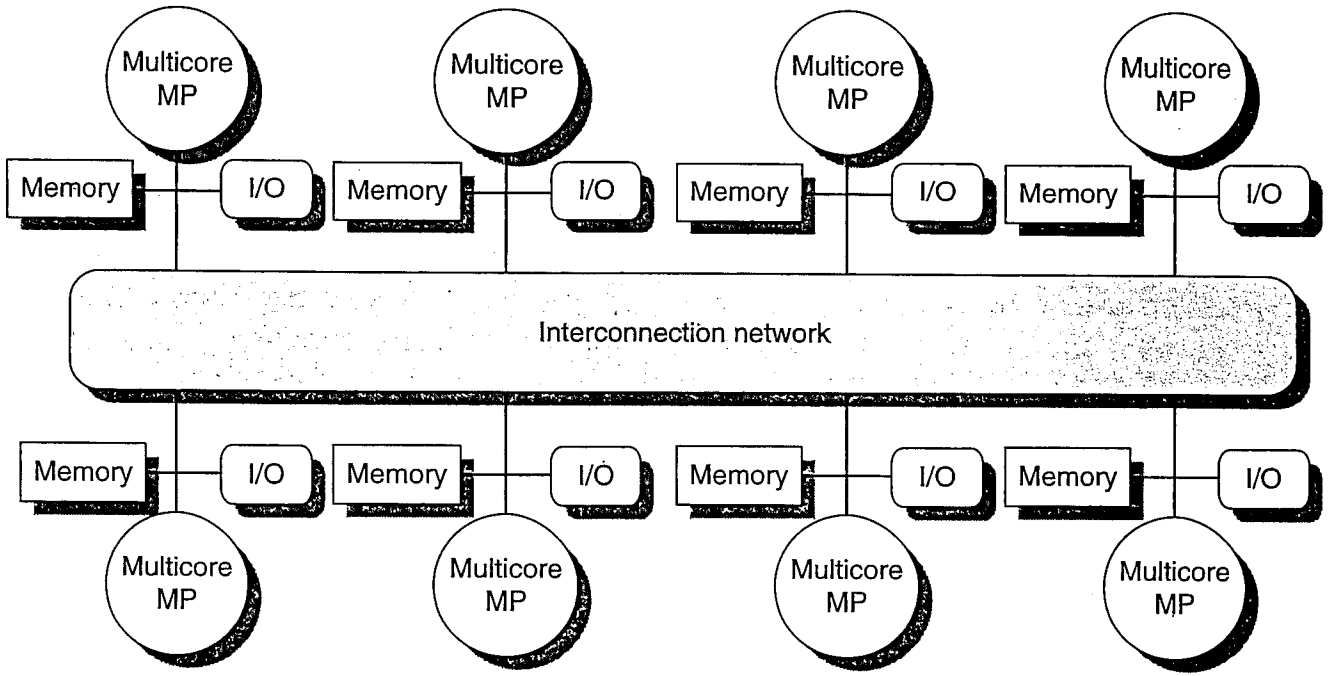


Figure 5.2 The basic architecture of a distributed-memory multiprocessor in 2011 typically consists of a multicore multiprocessor chip with memory and possibly I/O attached and an interface to an interconnection network that connects all the nodes. Each processor core shares the entire memory, although the access time to the lock memory attached to the core's chip will be much faster than the access time to remote memories.

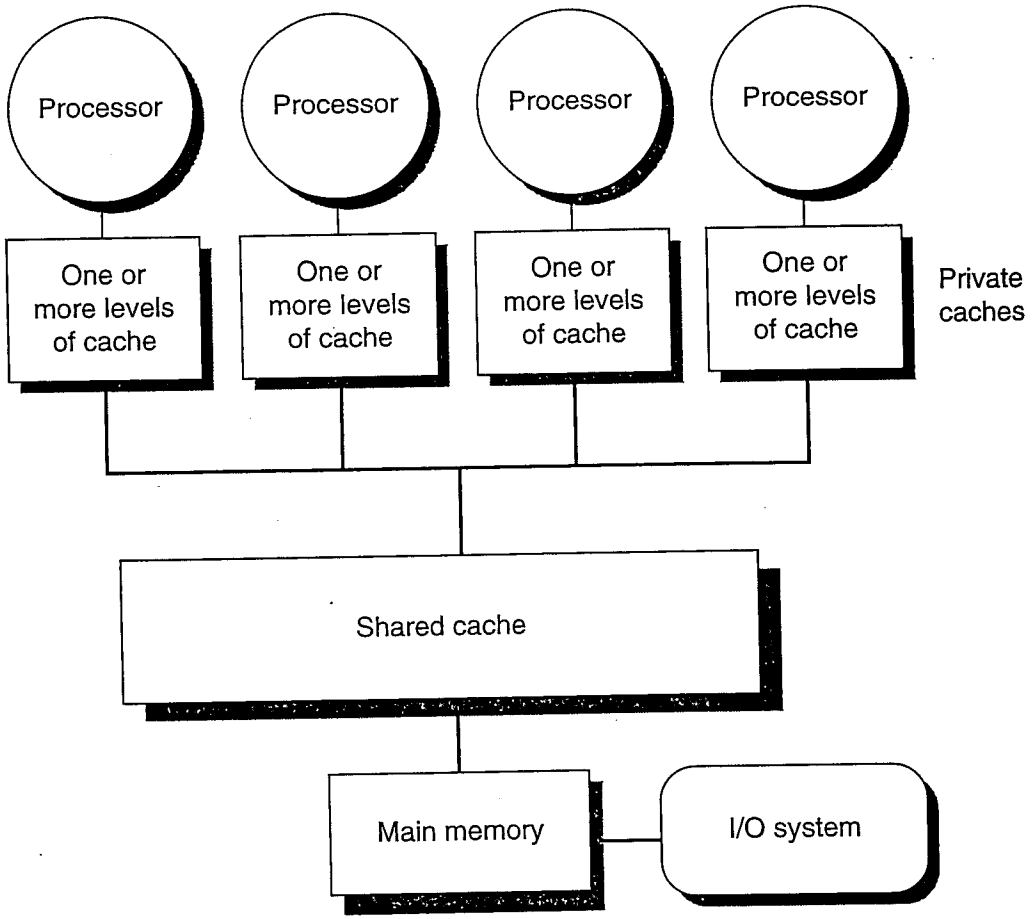


Figure 5.1 Basic structure of a centralized shared-memory multiprocessor based on a multicore chip. Multiple processor–cache subsystems share the same physical memory, typically with one level of shared cache, and one or more levels of private per-core cache. The key architectural property is the uniform access time to all of the memory from all of the processors. In a multichip version the shared cache would be omitted and the bus or interconnection network connecting the processors to memory would run between chips as opposed to within a single chip.

University of Idaho

MOVE TO MIMD

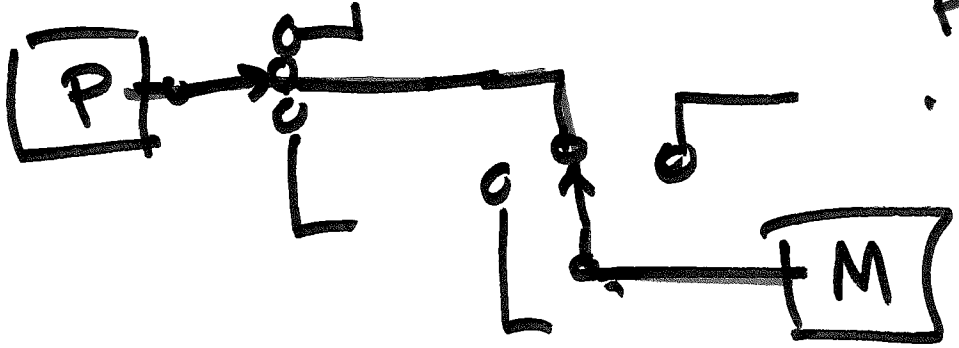
- TWO BASIC MODELS
  - SYMMETRIC (SHARED MEMORY) MULTIPROCESSOR - SMP
  - DISTRIBUTED SHARED MEMORY MULTIPROCESSOR - DSM

SMP

- SMALL NUMBER OF CORES
- SINGLE SHARED MEMORY
- UNIFORM MEMORY LATENCY (UMA)
- LIMITED SCALABILITY
- EASIEST TO PROGRAM

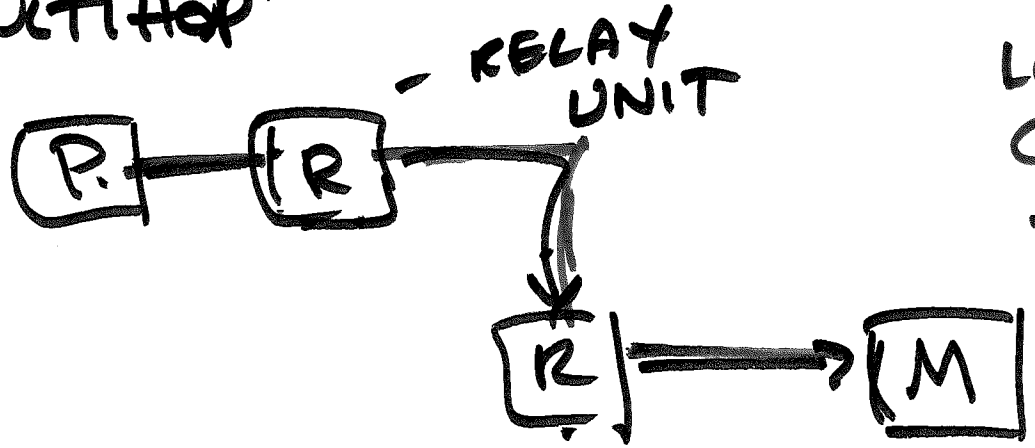
INTERCONNECTION NETWORK

"DIRECT CONNECT"



PHYSICAL "WIRE"
- MOVE ELECTRONS
- 1 CLOCK CYCLE TO TRANSFER A BIT

"MULTI HOP"



- RELAY UNIT

LOGICAL CONNECTION
- MULTIPLE CLOCK CYCLES/BIT

DSM

- LARGE NR. OF CORES
- DISTRIBUTED MEMORY - SHARED OVER A NETWORK
- NON-UNIFORM ACCESS LATENCY (NUMA)
  - EX IBM CELL PROCESSOR
    SONY PLAYSTATION 2
- HARDER TO PROGRAM
- MORE EASILY SCALED TO MANY PROCESSORS