CS 152
Computer Architecture and Engineering
Lecture 25: The Final Chapter

Dec 5, 1995

Dave Patterson (patterson@cs)

lecture slides: http://www-inst.eecs.berkeley.edu/~cs152/

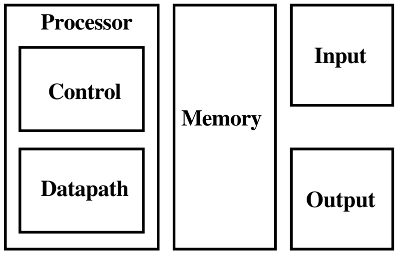
Outline of Today's Lecture

- Recap: What was covered in lectures (15 minutes)
- Questions and Administrative Matters (2 minutes)
- Future of Computer Architecture and Engineering (15 minutes)
- Lessons from CS 152 (10 minutes)
- Your Cal Cultural Heritage (20 minutes)
- HKN evaluation of teaching staff (15 minutes)

Where have we been?

The Big Picture

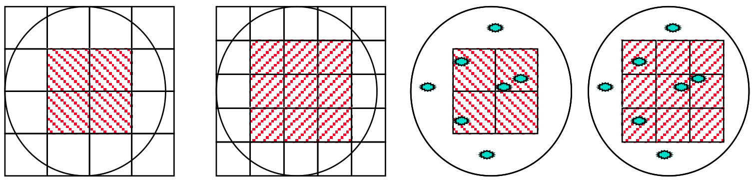
- Since 1946 all computers have had 5 components



Integrated Circuits Costs

$$\text{Die cost} = \frac{\text{Wafer cost}}{\text{Dies per Wafer} * \text{Die yield}}$$

$$\text{Dies per wafer} = \frac{\pi * (\text{Wafer\_diam} / 2)^2}{\text{Die Area}} - \frac{\pi * \text{Wafer\_diam}}{\sqrt{2 * \text{Die Area}}} - \text{Test dies} \approx \frac{\text{Wafer Area}}{\text{Die Area}}$$



$$\text{Die Yield} = \text{Wafer yield} * \left\{1 + \frac{\text{Defects\_per\_unit\_area} * \text{Die\_Area}}{\alpha}\right\}^{-\alpha}$$

Die Cost is goes roughly with the cube of the area.

Performance Evaluation Summary

$$\text{CPU time} = \frac{\text{Seconds}}{\text{Program}} = \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}}$$

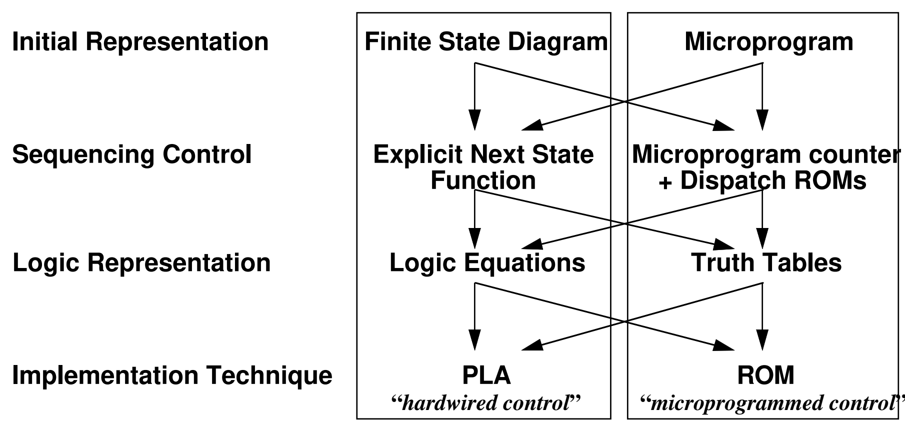
- Time is the measure of computer performance!
- Remember Amdahl's Law:
  Speedup is limited by unimproved part of program
- Good products created when have:
  - Good benchmarks
  - Good ways to summarize performance
- If NOT good benchmarks and summary, then choice between
  1) improving product for real programs
  2) changing product to get more sales (sales almost always wins)

Arithmetic

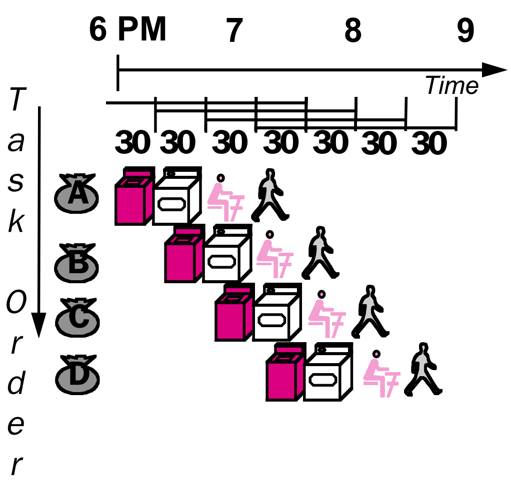
- Bits have no inherent meaning: operations determine whether really ASCII characters, integers, floating point numbers
- Divide uses same hardware as multiply (Hi & Lo registers in MIPS)
- Floating point follows paper & pencil method of scientific notation
  – using integer algorithms for multiply/divide of significands
- Pentium: Difference between bugs that board designers must know about and bugs that potentially affect all users
  – $200,000 cost in June to repair design
  – $400,000,000 loss in December in profits to replace bad parts
  – How much to repair Intel's reputation?
  – Make public complete description of bugs in later category?
- What is technologist's and company's responsibility to disclose bugs?

Control: Hardware vs. Microprogrammed

- Control may be designed using one of several initial representations. The choice of sequence control, and how logic is represented, can then be determined independently; the control can then be implemented with one of several methods using a structured logic technique.



Recap: Pipelining Lessons (its intuitive!)



- Pipelining doesn’t help latency of single task, it helps throughput of entire workload
- Multiple tasks operating simultaneously using different resources
- Potential speedup = Number pipe stages
- Pipeline rate limited by slowest pipeline stage
- Unbalanced lengths of pipe stages reduces speedup
- Time to “fill” pipeline and time to “drain” it reduces speedup
- Stall for Dependences

Pipeline Summary

- Pipelines pass control information down the pipe just as data moves down pipe
- Forwarding/Stalls handled by local control
- Exceptions stop the pipeline
- MIPS I instruction set architecture made pipeline visible (delayed branch, delayed load)
- More performance from deeper pipelines, parallelism

First Generation RISC Pipelines (1990)

- All instructions follow same pipeline order (“static schedule”).
- Register write in last stage
  - Avoid WAW hazards
- All register reads performed in first stage after issue.
  - Avoid WAR hazards
- Memory access in stage 4
  - Avoid all memory hazards
- Control hazards resolved by delayed branch (with fast path)
- RAW hazards resolved by bypass, except on load results which are resolved by fiat (delayed load).

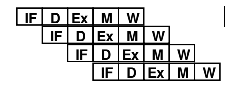
Substantial pipelining with very little cost or complexity.
Machine organization is (slightly) exposed!
Relies very heavily on "hit assumption"of memory accesses in cache

CS 152 project

How can the machine exploit available ILP?

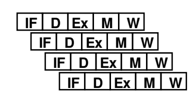
Technique — **Limitation**

- Pipelining



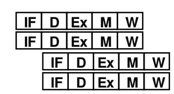
Issue rate, FU stalls, FU depth

- Super-pipeline
  - Issue 1 instr. / (fast) cycle
  - IF takes multiple cycles



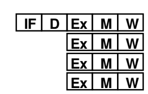
Clock skew, FU stalls, FU depth

- Super-scalar
  - Issue multiple scalar instructions per cycle



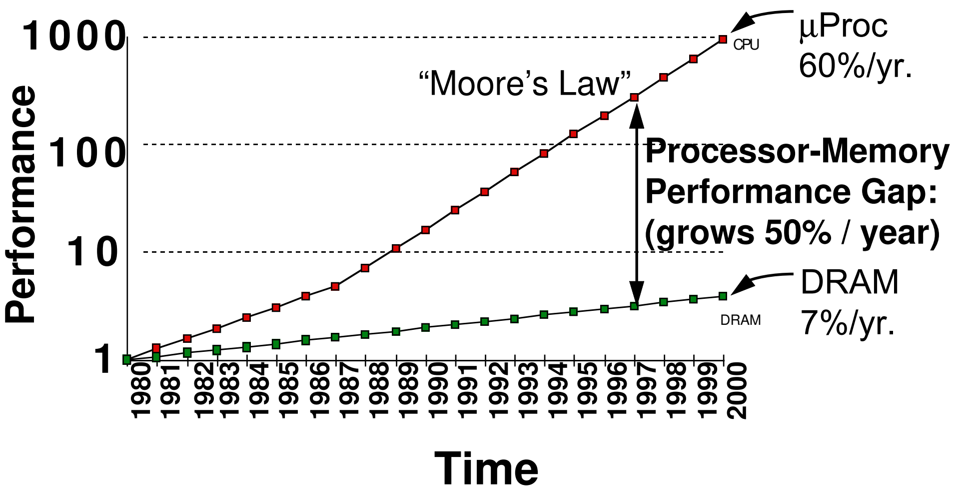
Hazard resolution

- VLIW
  - Each instruction specifies multiple scalar operations

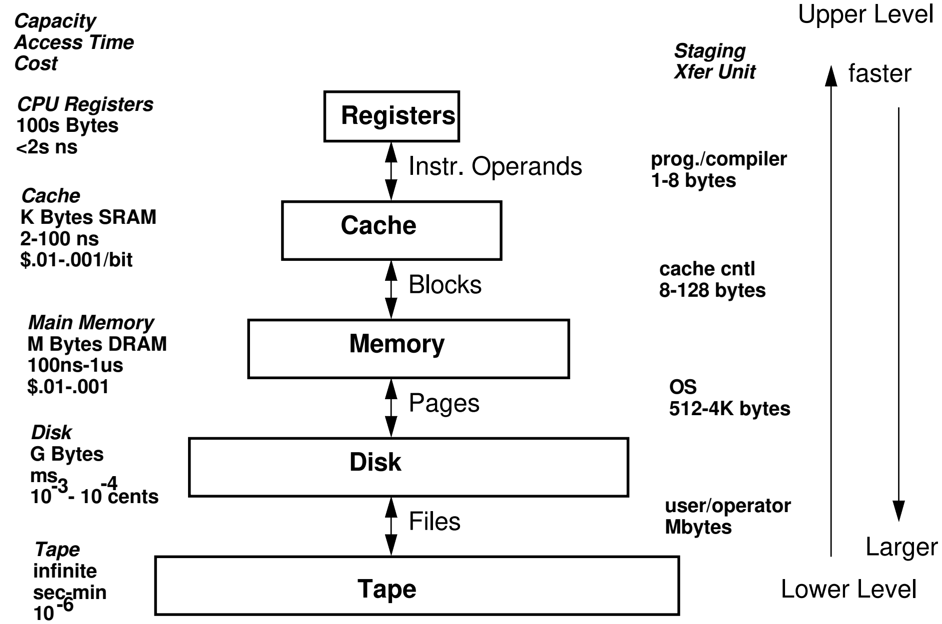


Packing

Processor-DRAM Gap (latency)



Levels of the Memory Hierarchy

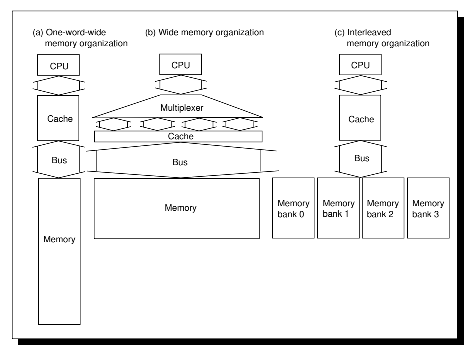


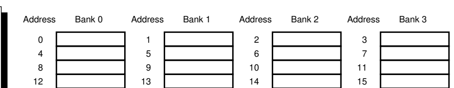
Memory Hierarchy

- **The Principle of Locality:**
  - **Program access a relatively small portion of the address space at any instant of time.**
    - **Temporal Locality: Locality in Time**
    - **Spatial Locality: Locality in Space**
- **Three Major Categories of Cache Misses:**
  - **Compulsory Misses: sad facts of life. Example: cold start misses.**
  - **Conflict Misses: increase cache size and/or associativity.**
  - **Capacity Misses: increase cache size**
- **Virtual Memory invented as another level of the hierarchy**
  - **Today VM allows many processes to share single memory without having to swap all processes to disk, protection more important**
  - **TLBs are important for fast translation/checking**

Main Memory Performance

- ***Simple*: CPU, Cache, Bus, Memory same width (32 bits)**
- ***Wide*: CPU/Mux 1 word; Mux/Cache, Bus, Memory N words (Alpha: 64 bits & 256 bits)**
- ***Interleaved*: CPU, Cache, Bus 1 word: Memory N Modules (4 Modules); example is *word interleaved***



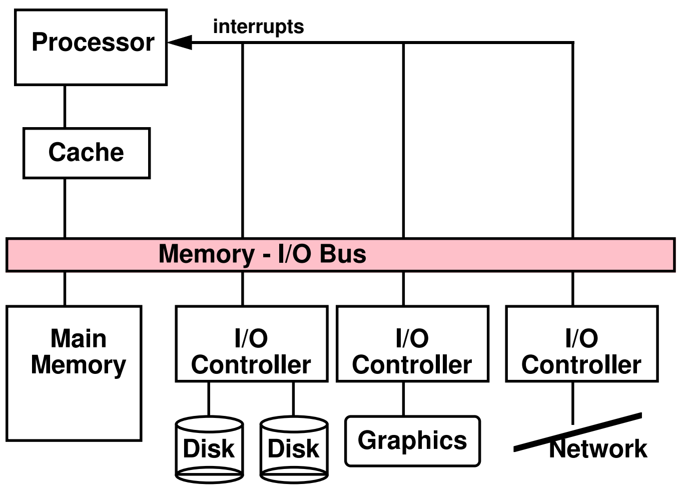


Timing model: 1 to send address, 6 access time, 1 to send data
Cache Block is 4 words

***Simple M.P.* = 4 x (1+6+1) = 32**
***Wide M.P.* = 1 + 6 + 1 = 8**
***Interleaved M.P.* = 1 + 6 + 4x1 = 11**

I/O System Design Issues

- Systems have a hierarchy of busses as well (PC: memory,PCI,ESA)



Guest Lectures

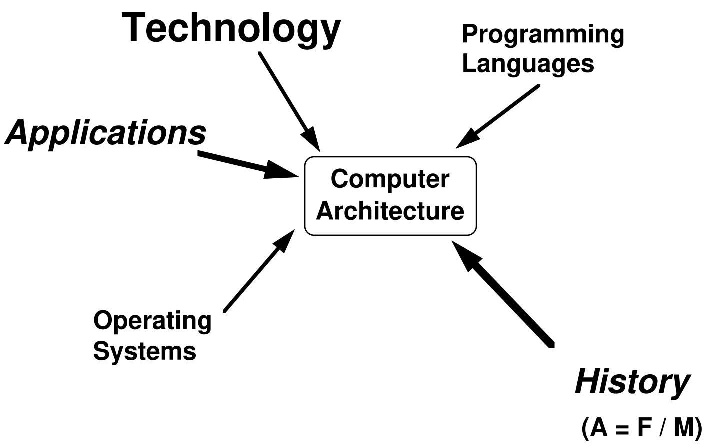
- CMOS power: capacitance x Vdd^2 x frequency
  - Power vs. Energy
- Disk I/O: RAID + hot spare reliable, high BW
- DSP: low power, low cost for 1 program
  - Hard real time performance, continuous I/O
  - Algorithms are king (IIR, FIR, FFT,convolutions)
    - Multiply-accumulates for bragging rites
- IA-64: explicit parallelism with 4X registers, 4X wider instructions, multiple functional units
  - conditional execution to reduce branches; more surprises in store

Questions and Administrative Matters

- Projects due 4PM on Monday Dec 8 1995 in 634 Soda (NOT THE BOX)
- Fix grade problems on assignments so far by Monday (score is wrong)
- grades posted Dec 15
- CS 152 questionnaire: help us improve CS 152
  - Arithmetic in prerequisties vs. lectures
  - Number of guest lectures? Which preferred?
  - Field trips?
  - Your good idea goes on questionnaire
    - e.g., pace of class
    - e.g., reduced number of assignments, requirements
    - e.g., increased disk space, licenses, computers

What does the future hold?

Forces on Computer Architecture



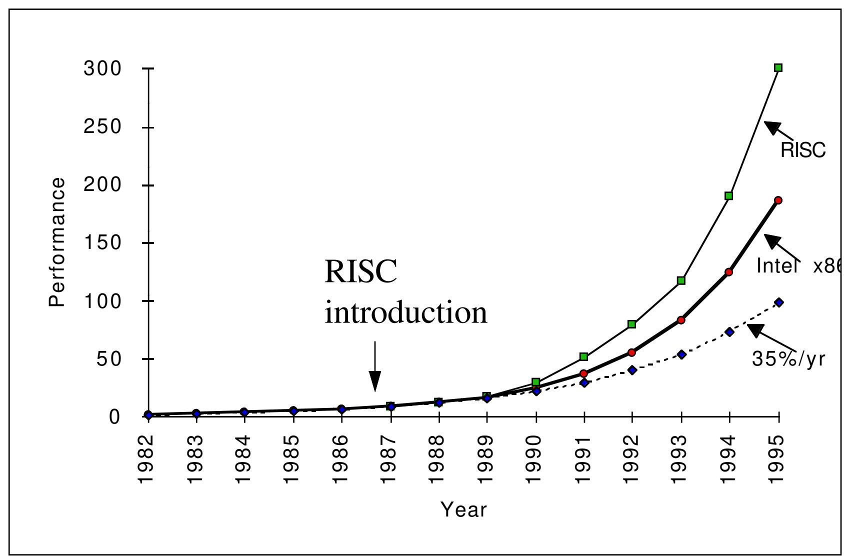
Key Technologies

- Fast, cheap, highly integrated “computers-on-a-chip”
  - IDT R4640, NEC VR4300, StrongARM, Superchips
  -
- Affordable access to fast networks
  - ISDN, Cable Modems, ATM, . . .
  -
- Platform independent programming languages
  - Java, JavaScript, Visual Basic Script
  -
- Lightweight Operating Systems
  - GEOS, NCOS, RISCOS
- ???

Future of Computer Architecture and Engineering

- Performance
- High Level Computer Architecture
- Multiprocessors
- “IRAM”

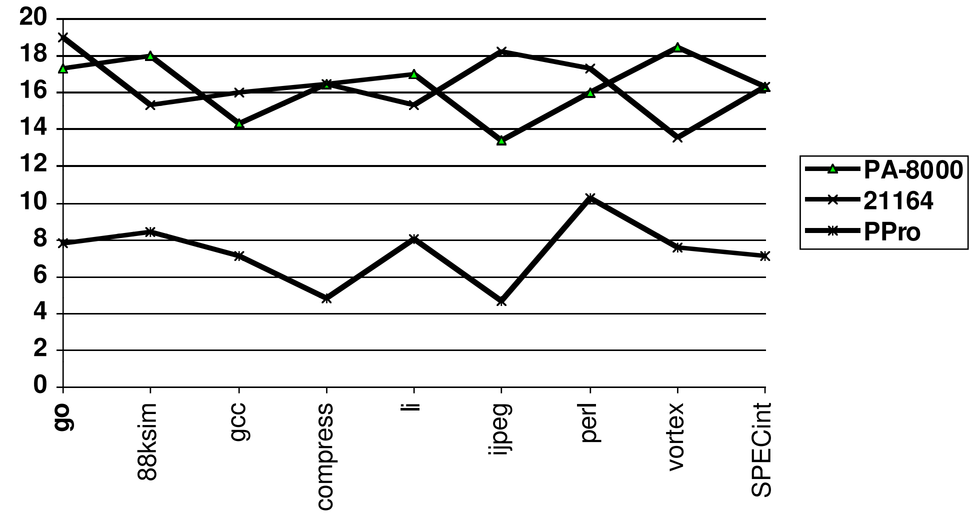
Processor Performance



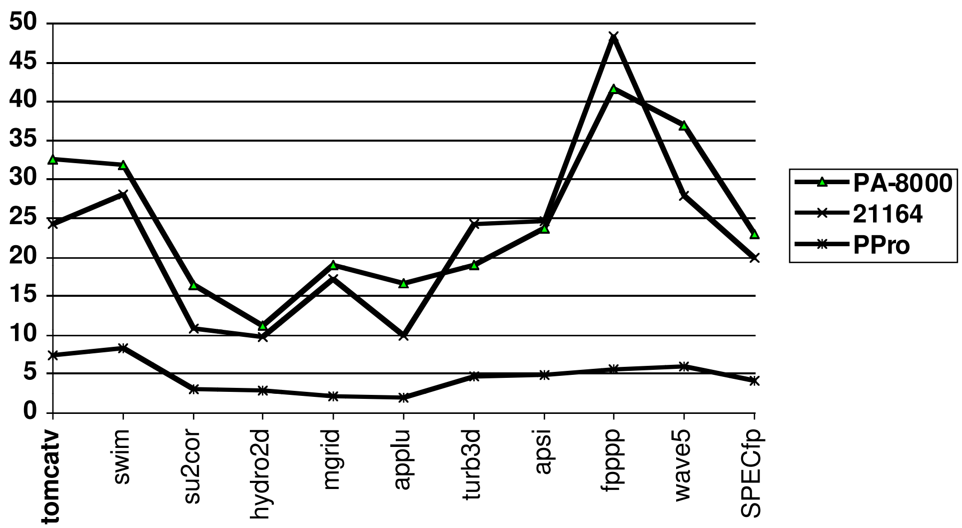
3 Recent Machines

	"Speed Demon"		"Braniac"
	Alpha 21164	Pentium II	HP PA-8000
Year	1995	1996	1996
Clock	600 MHz ('97)	300 MHz ('97)	236 MHz ('97)
Cache	8K/8K/96K/2M	16K/16K/0.5M	0/0/4M
Issue rate	2int+2FP	3 instr (x86)	4 instr
Pipe stages	7-9	12-14	7-9
Out-of-Order	6 loads	40 instr (μop)	56 instr
Rename regs	none	40	56

SPECint95base Performance (Oct. 1997)



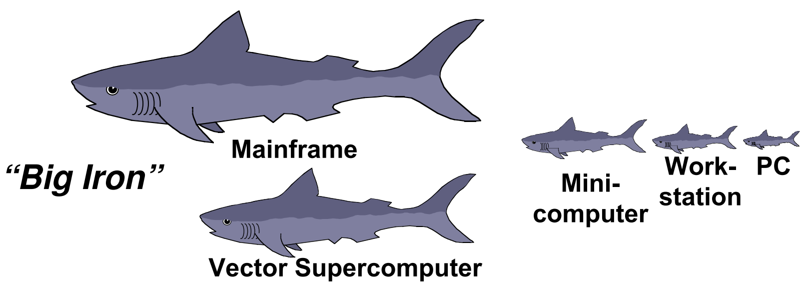
SPECfp95base Performance (Oct. 1997)



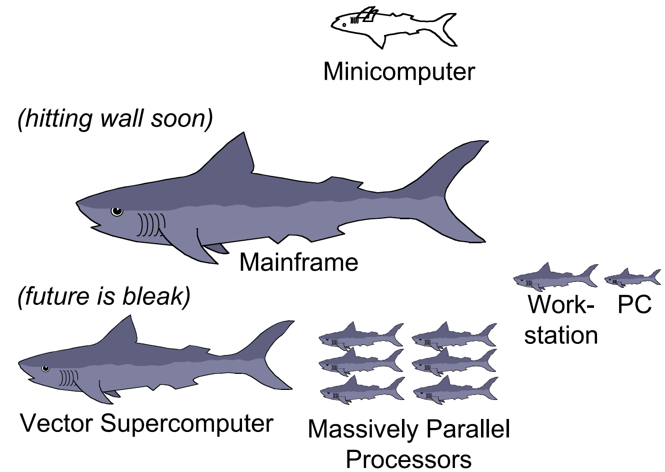
Performance Retrospective

- **Theory of Algorithms & Compilers based on number of operations**
- **Compiler remove operations and "simplify" ops:**
  Integer adds << Integer multiplies << FP adds << FP multiplies
  - **Advanced pipelines => these operations take similar time (FP multiply faster than integer multiply)**
- **As Clock rates get higher and pipelines are longer, instructions take less time but DRAMs only slightly faster (although much larger)**
- **Today time is a function of (ops, cache misses);**
- How do you tune performance on Pentium Pro? Random?
- **Given importance of caches, what does this mean to:**
  - **Compilers?**
  - **Data structures?**
  - **Algorithms?**

1985 Computer Food Chain

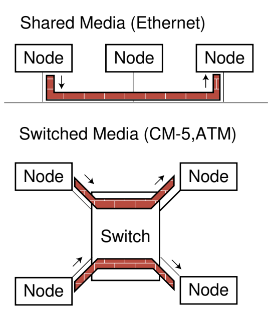


1995 Computer Food Chain

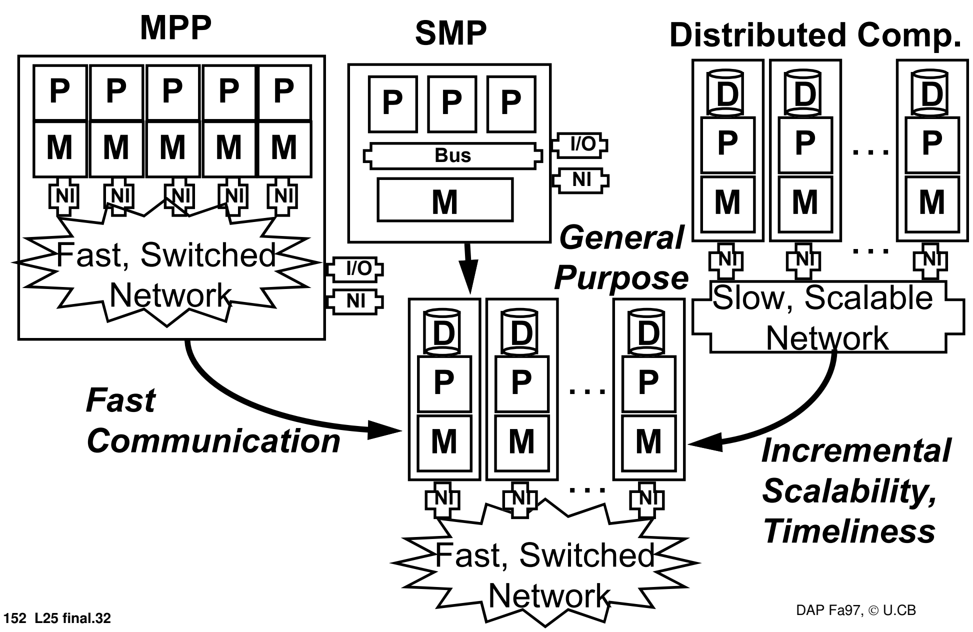


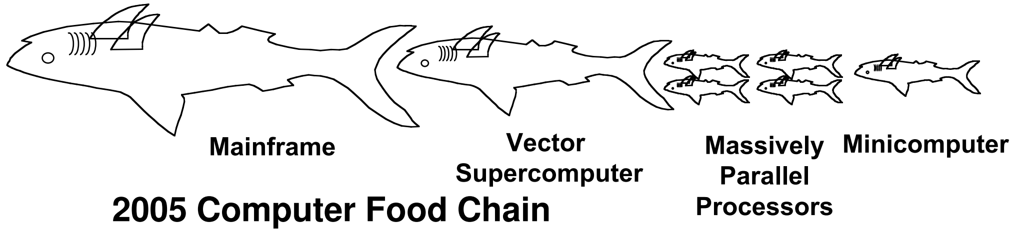
Interconnection Networks

- Switched vs. Shared Media: pairs communicate at same time: "point-to-point" connections

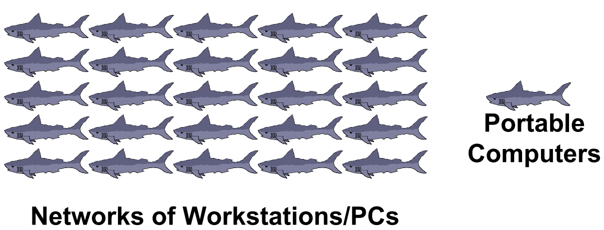


Cluster/Network of Workstations (NOW)





2005 Computer Food Chain

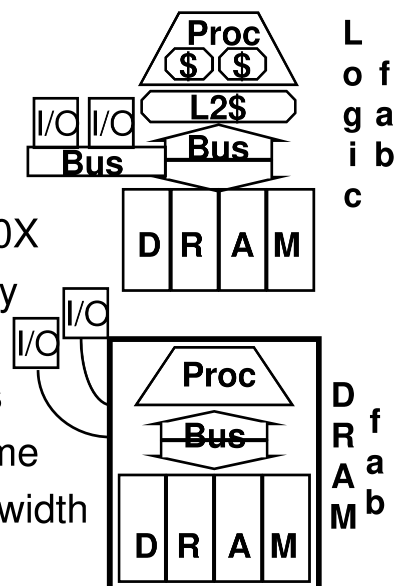


Intelligent DRAM (IRAM)

- IRAM motivation (≈2000 to 2005)
  - 256 Mbit/1Gbit DRAMs in near future (128 MByte)
  - Current CPUs starved for memory BW
  - On chip memory BW = SQRT(Size)/RAS or 80 GB/sec
  - 1% of Gbit DRAM = 10M transistors for µprocessor
  - Even in DRAM process, a 10M trans. CPU is attractive
  - Package could be network interface vs. Addr./Data pins
  - Embedded computers are increasingly important
- Why not re-examine computer design based on separation of memory and processor?
  - Compact code & data?
  - Vector instructions?
  - Operating systems? Compilers? Data Structures?

IRAM Vision Statement

Microprocessor & DRAM on a single chip:
- on-chip memory latency 5-10X, bandwidth 50-100X
- improve energy efficiency 2X-4X (no off-chip bus)
- serial I/O 5-10X v. buses
- smaller board area/volume
- adjustable memory size/width



and why not

- multiprocessors on a chip?
- complete systems on a chip?
  - memory + processor + I/O
- computers in your credit card?
- networking in your kitchen? car?
- eye tracking input devices?

Learned from Cal/CS152?

Online Notes

- Guess: Which has more: CS 152 online slides vs. pages in COD (including forward, appendices)?
- Pages in COD 2/e:
  995
- Total CS152 slides online:
  1020

Project summaries

- Problem: VHDL takes no time, logic takes time; mux slower than tristate
- Speed, notes of 14 projects;

40 ns, tristate	77 ns,
40 ns, tristate/WBbypass	100 ns,
50 ns, tristate	120 ns,
60 ns, stall branch hazard	133 ns, cache 1/2 clock
60 ns, WBbypass	200 ns, cache 1/2 clock
66 ns	

- Caches fewer problems than datapaths (learned from mistakes)
- Things done: 64b memory, interlocked loads, 2-way set assoc cache, fully associative, subblock placement for writes, TLB, Branch prediction (initialize to intermediate state)+BTB;
- In report include clock cycles for Quicksort, where cycles go

CS152: So what's in it for me? (from 1st lecture)

- **In-depth understanding of the inner-workings of modern computers, their evolution, and trade-offs present at the hardware/software boundary.**
  - **Insight into fast/slow operations that are easy/hard to implementation hardware**
- **Experience with the *design process* in the context of a large complex (hardware) design.**
  - **Functional Spec --> Control & Datapath --> Physical implementation**
  - **Modern CAD tools**
- **Designer's "Intellectual" toolbox.**

Simulate Industrial Environment (from 1st lecture)

- Project teams must have at least 4 members
  - Managers have value
- Communicate with colleagues (team members)
  - What have you done?
  - What answers you need from others?
  - You must document your work!!!
  - Everyone must keep an on-line notebook
- Communicate with supervisor (TAs)
  - How is the team's plan?
  - Short progress reports are required:
    - What is the team's game plan?
    - What is each member's responsibility?

So let's thanks those TAs

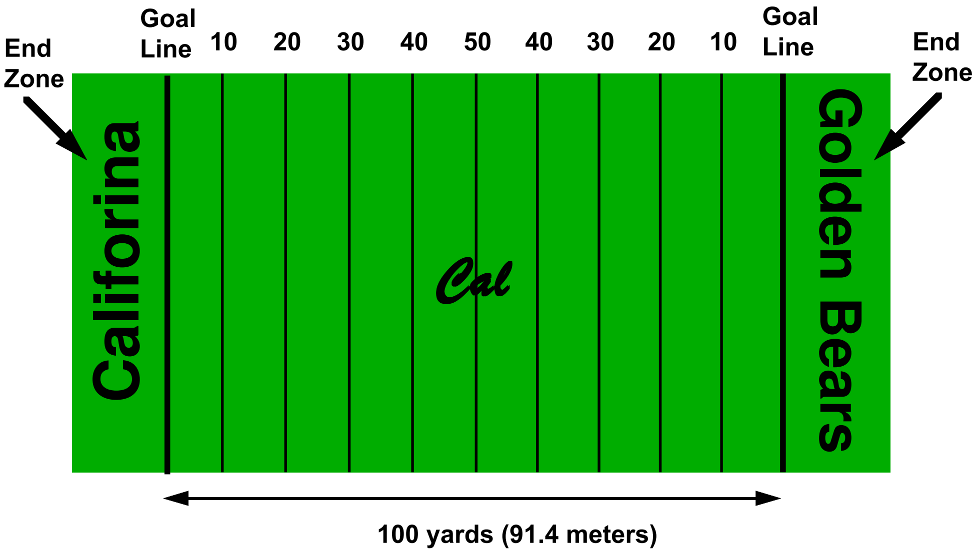
Summary: Things we Hope You Learned from 152 (from 1st lecture)

- Keep it simple and make it work:
  - Fully test everything individually & then together; break when together
  - Retest everything whenever you make any changes
  - Last minute changes are big "no nos"
- Group dynamics. Communication is the key to success:
  - Be open with others of your expectations & your problems (e.g., trip)
  - Everybody should be there on design meetings when key decisions are made and jobs are assigned
- Planning is very important ("plan your life; live your plan"):
  - Promise what you can deliver; deliver more than you promise
  - Murphy's Law: things ***DO*** break at the last minute
    - ***DON'T*** make your plan based on the best case scenarios
    - Freeze you design and don't make last minute changes
- Never give up! It is not over until you give up ("Bear won't die")

Cal Cultural History: ABCs of American Football

- Started with soccer; still 11 on a team, 2 teams, 1 ball, on a field; object is to move ball into "goal"; most goals wins
- New World changes the rules to increase scoring:
  - Make goal bigger! (full width of field)
  - Carry ball with hands
  - Can toss ball to another player backwards or laterally (called a "lateral") anytime & forwards ("pass") sometimes
- How to stop players carrying the ball? Grab them & knock them down by making knee hit the ground ("tackle")
  - if drop ball ("fumble"), other players can pick it up and score
- Score by moving ball into goal ("cross the goal line" or "into the end zone") scoring a "touchdown" (6 points), or kicking ball between 2 poles ("goal posts") scoring a "field goal"
  (3, unless after touchdown = 1: "extra point")
- Kick ball to other team after score ("kickoff"); laterals OK
- Game ends when no time left (4 15 min quarters) & person with ball is stopped (Soccer time only:2 45 min halves)

Football Field



The Spectacle of Football

- **Rose Bowl: Prestigious bonus game played January 1 if have a great year ("playoffs")**
  - preceeded by parade
  - national TV coverage
- **1929 Rose Bowl Game**
  - Cal vs. Georgia Tech
  - Cal going left to right (==>), GeorgiaTech right to left (<==)
  - Georgia Tech player fumbles football
  - Cal player, Roy Reigel, picks up football and tries to avoid Georgia Tech players
- **Let's see what happens on video**

The Spectacle of Football

- **Play nearby archrival for last game of season**
- **Cal's archrival is Stanford; stereotype is Private, Elitist, Snobs**
- **The Big Game: Cal vs. Stanford, winner gets a trophy ("The Axe") : Oldest rivalry west of Mississippi; 100th in 1997**
- **American college football is a spectacle**
  - **School colors (Cal: Blue & Gold; Stanford: Red &** White **)**
  - **School nicknames (Cal: Golden Bear; Stanford: Cardinal)**
  - **School mascot (Cal: Oski the bear; Stanford: a tree(!))**
  - **Leaders of cheers ("cheerleaders")**
- **"Bands" (orchestras that march) from both schools at games; before game, at halftime, after game**
  - **Stanford Band more like a drinking club; ≈ "Animal House"**
  - **Plays one song: "All Right Now"**
  - **Stanford used to yell "boring" at band during Cal's performance**

1982 Big Game

- **"There has never been anything in the history of college football to equal it for sheer madness."** ***Sports Illustrated***
- **Cal coach is Joe Kapp, former Cal player; tells team to play 100% for 60 minutes ("40 for 60"; "Bear will not die"); 1st year as coach; lasts 5 years ("Never give up")**
- **Stanford coach is Paul Wiggin, former Stanford player, lots of coaching experience; fired from job next year**
- **Stanford Quarterback is John Elway, who goes on to be a professional All Star football player (still playing today)**
- **Cal Quarterback is Gail Gilbert, who goes on to be a non-starting professional football player (stoped playing 1996)**
- **Stanford lost 4 games in last few minutes of game**
- **Let's see what happens on video**

Notes About "The Play"

- Cal only had 10 men on the field; last second another came on (170 pound Steve Dunn #3) & makes key 1st block
- Kevin Moen #26: 6'1" 190 lb. safety, never scored in 4 years at Cal
  - laterals to Rodgers (and doesn't give up)
- Richard Rodgers #5: 6' 200 lb. safety, "Don't fall with the ball."
  - laterals to Garner
- Dwight Garner #43: 5'9" 185 pound running back
  - almost tackled, 2 legs & 1 arm pinned, laterals to Rodgers
- Richard Rodgers #5 (again): "Give me the ball, Dwight."
  - laterals to Ford
- Mariet Ford #1: 5'9", 165 pound wide receiver
  - leg cramps, overhead blind lateral to Moen & blocks 3 players
- Moen (again) cuts through Stanford band into end zone
- On field for Stanford: 22 football players, 3 Axe committee members, 3 cheerleaders, 144 Stanford band members (172 for Stanford v. 11 for Cal)
- "Weakest part of the Stanford defense was the woodwinds."
- 4 Cal players + Stanford Trombonist (Gary Tyrrell) hold reunion every year at Big Game; Stanford revises history (20-19 on Axe)

Your Cal Cultural History

- Cal students/alumni heritage is the greatest college football play in > 100 years
- Cal students/alumni work hard and play hard
- Cal students/alumni handle adversity
- Cal students/alumni never give up!
- Cal students/alumni triumph over great odds!
- Cal students/alumni take pity on Stanford students/alumni

The Future for Cal people:

- **Better educated than Stanford people**
- Stanford CS/EE undergrads only name 1 or 2 regular CS/EE faculty
- Silicon Valley: more Cal grads than Stanford (Gordon Moore)
- Stanford MS student ≈ Cal BS student (Intel rep)
- Going to grad school Stanford vs. Cal:

  5% vs. 25%
- **Future: What you make it to be**