Real-Time Graphics Architecture

Kurt Akeley

Pat Hanrahan

http://www.graphics.stanford.edu/courses/cs448a-01-fall

Nuts and Bolts

Some hardware background

Nuts and Bolts

Fundamentals

- Circuit elements
- Elements of logic and computer architecture
- Integrated circuits, esp. DRAM

Motivate further investigation

- *Digital Systems Engineering*, Dally and Poulton
- *DRAM Design Overview*, Jinji Ogawa, 1998
- *Computer Architecture*, Hennessy and Patterson?

We'll omit the very gory details

- I don't know them anyway

Basic Elements of Circuits

Transistors

- You hear about these all the time

Wires

- You hear much less about these

Others

- E.g. resistors, capacitors, inductors, diodes, ….
- These are statistically uncommon
- They are *properties* of transistors and wires

Electrical Properties

Voltage (v)

- Electrical pressure

Current (i)

- Flow of electrical charge

Resistance (R)

- Dissipates power when current is flowing

$$v = iR \qquad \textit{(Ohm's Law)}$$

$$\textit{Power} = vi = i^2R = v^2/R$$

More Electrical Properties

Resistance (R)

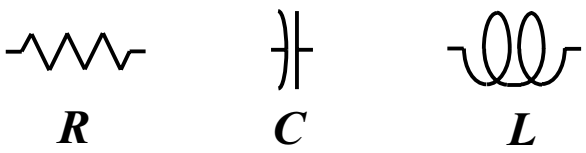
- Dissipates power when current is flowing

Capacitance (C)

- Stores charge, resists change in voltage

Inductance (L)

- Resists change in current



Wires are Interesting

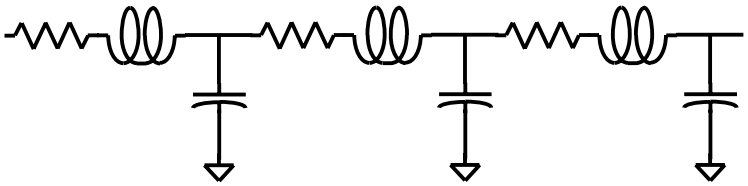
Transistors get the attention, but wires make a design

Signals *propagate* along wires

- Wires are not perfect conductors
- They have resistance, capacitance, & inductance

Ideal

Actual

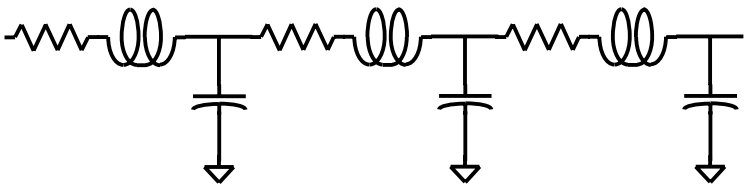


Wires are Interesting (continued)

Signal propagation:

- Takes time (2ns per foot)
- Consumes energy (charging C through R)
- Generates emissions (L, noise)
- Is affected by emissions (interference, cross talk)
- Is affected by changes in *impedance* (LC)

Actual



Wires are Interesting (continued)

Wires consume resources

- On the die (of an integrated circuit chip)
- On the chip (its pins)
- On the circuit board (traces, connectors)

Transistors

Transistors amplify signals

- Transistor is an active gain element
- Without gain, no useful circuit is possible

Transistors invert signals

- Gain can be arranged to be negative
- Without inversion, no useful logic circuit is possible

Transistors are analog devices

- They are not perfect switches
- They have resistance, capacitance, & inductance

Transistors (continued)

Switching a transistor:

- Takes time (now measured in ps)
- Dissipates energy (charging C through R)
- Generates minimal emissions (L, noise)
- Is slightly affected by emissions (L, interference)

Transistors consume resources

- On the die
- (But wires sometimes dominate)

Summary – Circuit Elements

Good digital designers are adequate *analog* designers

- Respect the properties of transistors and wires

Who ever heard of "wire count"?

- Wires are critical to IC and system design
- Wires can make or break an architecture

Integrated Circuits

Multum in Parvo (MIPchip ;-)

- Much in a little place
- Exponentially more over time (Moore's Law)

Enabled by photo lithography

- Opto-chemical vector processing
- Operates at wafer (not die) scale

Yield determines economics

- Measured in operating die per wafer
- Larger die → fewer candidates
- More circuitry per die → higher failure rate

Integrated Circuits (continued)

Pins are big and slow

- Wires are bonded to pads, usually on edge of die
- *Pad limited* → circuitry doesn't fill the die

Prioritized design limiting factors:

1. Complexity management
2. Power dissipation
3. Off-die wiring
4. On-die wiring
5. Available transistors

Families of Integrated Circuits

Was once a rich field:

- RTL → DTL → TTL → STTL → FTTL
- ECL
- Gallium Arsenide

The war is nearly over - CMOS is cleaning up

Exceptions are few:

- DRAM memory
- CCD image arrays (but CMOS is closing here too)
- Bipolar (FTTL, ECL) in high-end products
- Vacuum tubes (transmission towers)

CMOS Technology

Complementary Metal Oxide Silicon (CMOS)

MOS Transistor (a.k.a. Field Effect Transistor)

- Metal *gate* is Oxide-insulated from Silicon channel
- Oxide insulator has near infinite resistance
  - No DC current flow at gate (unlike *bi-polar*)
- Switches very nearly *on* or o*ff*

Complementary

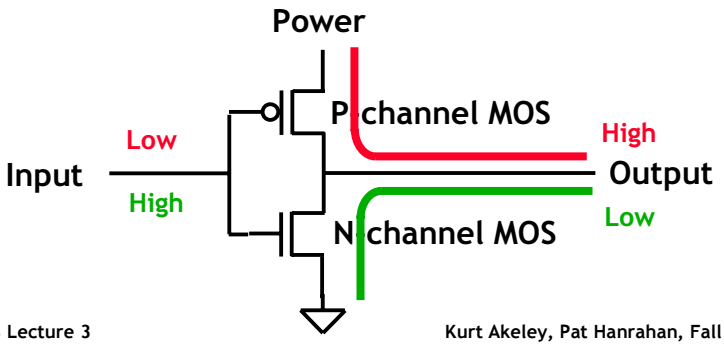
- Two types of MOS transistors
  - P-channel - *on* when gate is low voltage
  - N-channel - o*n* when gate is high voltage

CMOS Inverter Circuit

No DC current flow

- No gate current (MOS transistors)
- No path from ground to power

Dissipates power *only* during transitions



Categories of Integrated Circuits

Fixed purpose (off-the-shelf)

- Memory (SRAM, DRAM, …)
- Microprocessor, DSP, GPU
- Special purpose (DAC, voltage regulator, …)

Programmable

- Field Programmable Logic Array (FPLA)
- Amazingly capable

Custom

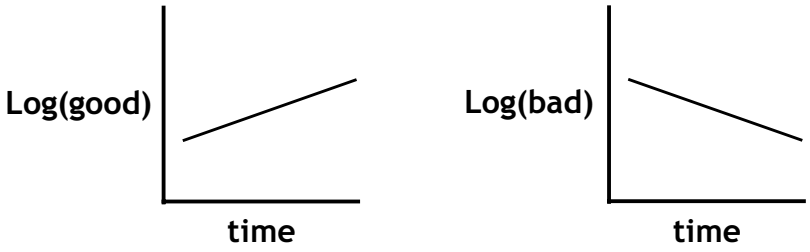
- Effort ranges from large to extraordinary
- (Expensive) software design tools are a must

Scaling of IC Technology

Purely exponential for 30+ years

Should continue at least through 2010

- Perhaps much longer
- Perhaps forever? (*The Age of Spiritual Machines*, Ray Kurzweil, 1999)



Semiconductor Scaling Rates

From: *Digital Systems Engineering*, Dally and Poulton

Parameter	Current Value	Yearly Factor	Years to Double (Half)
Moore's Law (grids on a die)**	1 B	1.49	1.75
Gate Delay	150 pS	0.87	(5)
Capability (grids / gate delay)		1.71	1.3
Device-length wire delay		1.00	
Die-length wire delay / gate delay		1.71	1.3
Pins per package	750	1.11	7
Aggregate off-chip bandwidth		1.28	3

** Ignores multi-layer metal, 8-layers in 2001

Elements of Logic

(Recall that transistors and wires are circuit elements)

Gates

- Combinational logic (e.g. AND, NAND, NOR)
- Transistors and wires

Registers

- Sample and store state bits at clock events
- Transistors and wires

Signals

- Connect gates and registers
- Transistors and wires

Elements of Computer Architecture

State machines

- Gates, registers, and signals
- Examples: FPU, memory controller

Memory

- Gates, registers, and signals
- Examples: DRAM, SRAM, on-die cache memory

Communication (data movement)

- Gates, registers, and signals

Communication (data where needed when needed)

- Gates, registers (memory), and signals

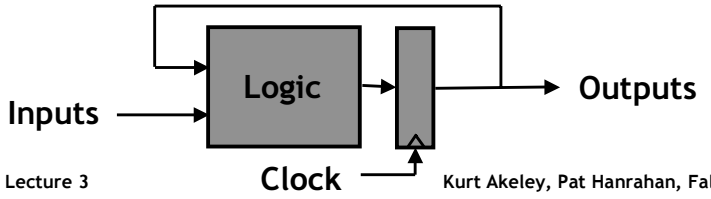
State Machines

Synchronous logic

- Single clock signal distributed through region
- This is the die-length wire!
- Asynchronous design option never achieved

Performance determined by maximum delay path

- Measured in gates
- Signal delays matter too (especially the clock)



Random Access Memory (RAM)

Ideal: addressed array of registers

- Read: present address, contents are output
- Write: present address and data, strobe
- Random → equal access to all locations

Non-random technologies

- Cyclical: Delay lines, shift registers, disks, drums
- Hierarchical: cache

IC technology has bifurcated:

- Static RAM (SRAM) – much like ideal
- Dynamic RAM (DRAM) – *not* ideal

DRAM Fundamentals

Highest RAM density

- Smallest memory cell
- Used for all "large" memories

Highest RAM complexity

- Driven by density goal
- DRAM is "high-maintenance" memory

Different process technology

- Difficult to provide on CMOS logic die

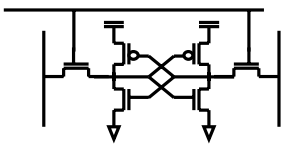
Commodity business

- Everyone playing the "value add" game

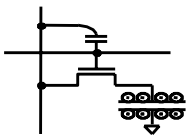
DRAM Fundamentals (continued)

Cell is as small as possible

- One transistor, one capacitor (real)
- Inherently Dynamic
  - No gain element in cell
  - Requires refresh to maintain state reliably
- Inherently slow
  - Minimal bit energy, huge amplification required



SRAM cell

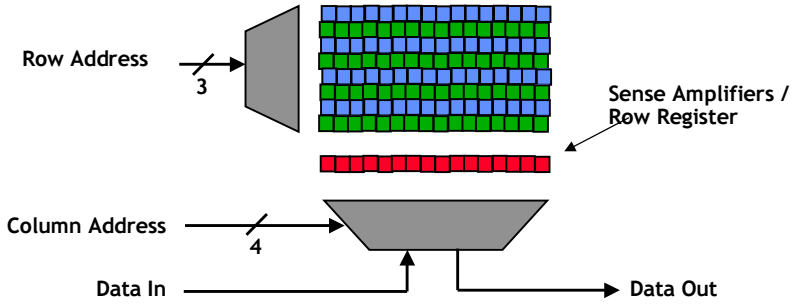


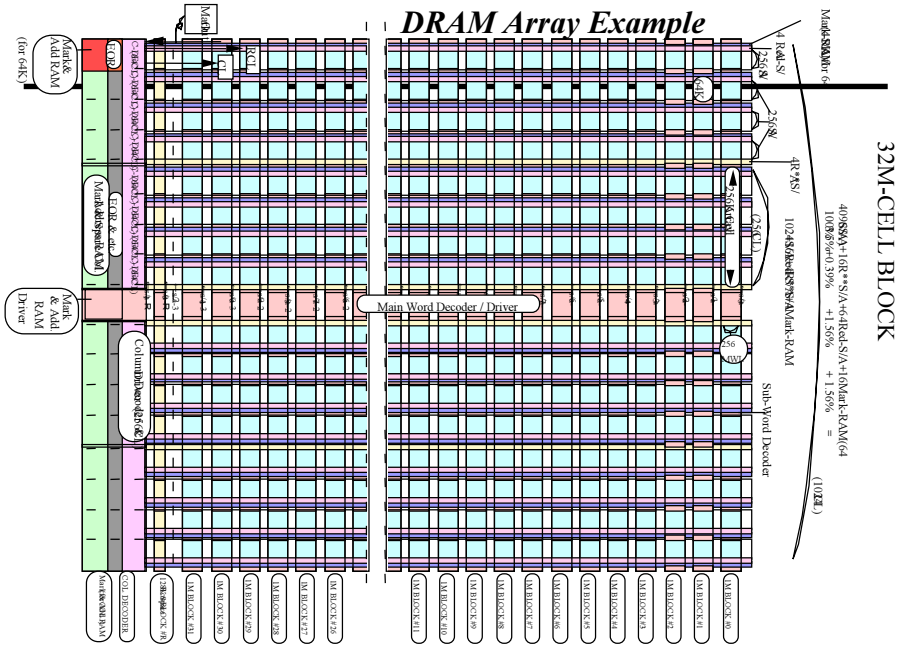
DRAM cell

DRAM Fundamentals (continued)

Memory array is 2D structure

- Row and column addresses presented sequentially
- Row "locality" can be exploited





Slide courtesy of Mark Horowitz, from Junji Ogawa 1998 presentation

DRAM Fundamentals (continued)

Complex state machine

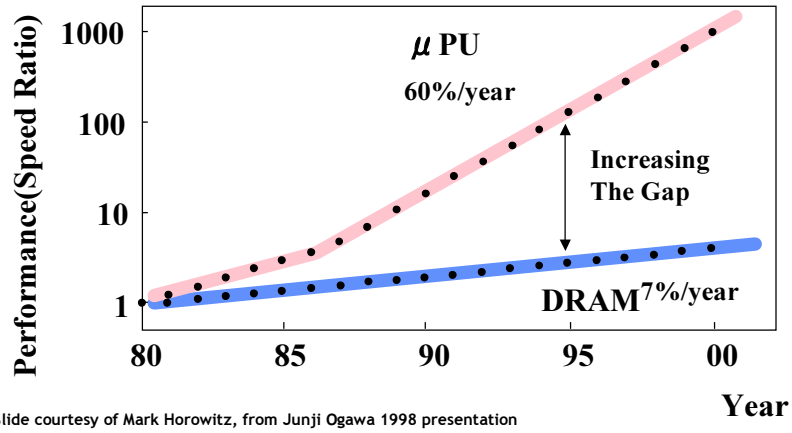
- Multiple active rows
- Split read/write
- Automatic/hidden refresh cycles

Complex interface

- Multiplexed row/column addresses
- Strobe signals (not clocked)
- Hundreds of timing parameters
- Required external memory controller is a complex state machine

DRAM Fundamentals (continued)

Speed Gap between DRAM and CPU
- Memory Wall -



Slide courtesy of Mark Horowitz, from Junji Ogawa 1998 presentation

Signaling and Communication

Bus

- Multi-conductor signal
- Historically multi-drop (like PCI?)

High-speed signaling requires

- Point-to-point wiring
- Proper termination
- Clever attention to detail
  - (see Dally and Poulton, the whole book)

Example Gate Counts

Circuit Description	Gates	Growth
1-bit register file (1 read, 1 write)**	2	n
24-bit fast integer adder	1000	$n \log(n)$
24-bit integer multiplier	4000	n^2
32-bit IEEE adder	4000	
32-bit IEEE multiplier	5000	
Microprocessor core	~50000	

** Multi-port register is wire limited

Data courtesy of John Montrym, NVIDIA

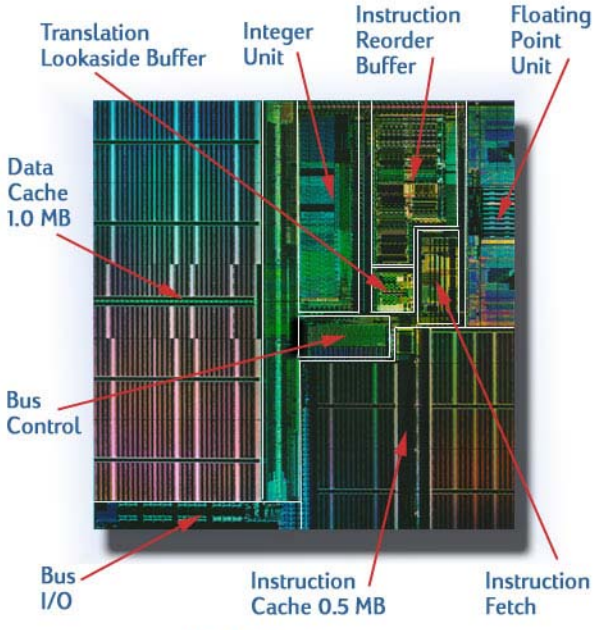
GPU vs. CPU

GPU performance compounding at over 2x per year

- Reference: Introduction slides

CPU's performance compounding at ~1.5x per year

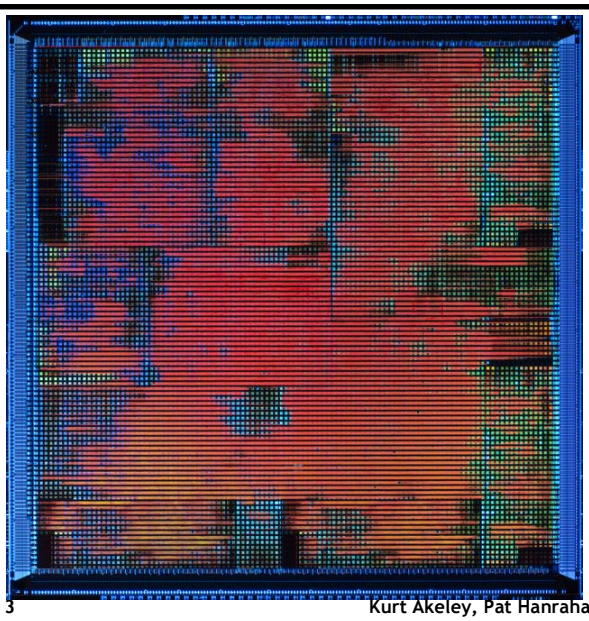
Why is this?



PA-8500 microprocessor

http://www-vlsi.stanford.edu/group/chips_micropro.html

GeForce3 Die Plot (Courtesy NVIDIA)



Real-Time Graphics Architecture

Kurt Akeley

Pat Hanrahan

http://www.graphics.stanford.edu/courses/cs448a-01-fall

Symbols