CIS 501: Computer Architecture

Unit 3: Technology & Energy

Slides developed by Milo Martin & Amir Roth at the University of Pennsylvania
with sources that included University of Wisconsin slides
by Mark Hill, Guri Sohi, Jim Smith, and David Wood

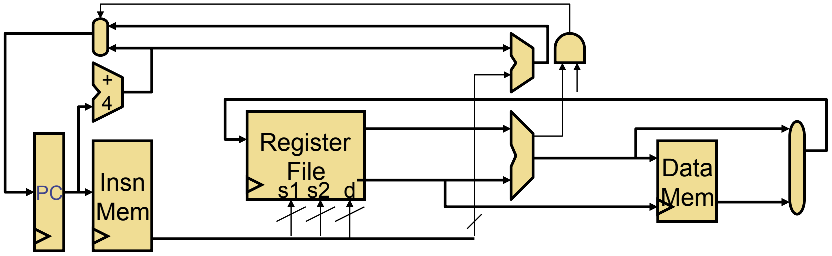
This Unit: Technology & Energy

- Technology basis
  - Fabrication (manufacturing) & cost
  - Transistors & wires
  - Implications of transistor scaling (Moore's Law)
- Energy & power

Readings

- MA:FSPTCM
  - Section 1.1 (technology)
  - Section 9.1 (power & energy)
- Paper
  - G. Moore, "Cramming More Components onto Integrated Circuits"

Review: Simple Datapath



- How are instruction executed?
  - Fetch instruction (Program counter into instruction memory)
  - Read registers
  - Calculate values (adds, subtracts, address generation, etc.)
  - Access memory (optional)
  - Calculate next program counter (PC)
  - **Repeat**
- **Clock period = longest delay through datapath**

Recall: Processor Performance

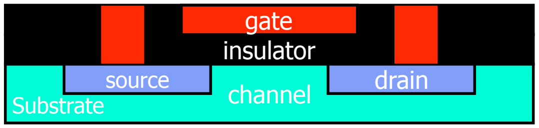
- Programs consist of simple operations (instructions)
  - Add two numbers, fetch data value from memory, etc.
- Program runtime = "seconds per program" =

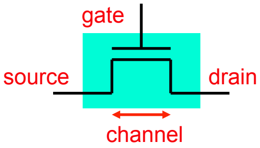
(instructions/program) * (cycles/instruction) * (seconds/cycle)

- **Instructions per program**: "dynamic instruction count"
  - Runtime count of instructions executed by the program
  - Determined by program, compiler, instruction set architecture (ISA)
- **Cycles per instruction**: "CPI" (typical range: 2 to 0.5)
  - On average, how many *cycles* does an instruction take to execute?
  - Determined by program, compiler, ISA, micro-architecture
- **Seconds per cycle**: clock period, length of each cycle
  - Inverse metric: cycles per second (Hertz) or cycles per ns (Ghz)
  - Determined by micro-architecture, **technology parameters**
- **This unit: transistors & semiconductor technology**

Technology & Fabrication

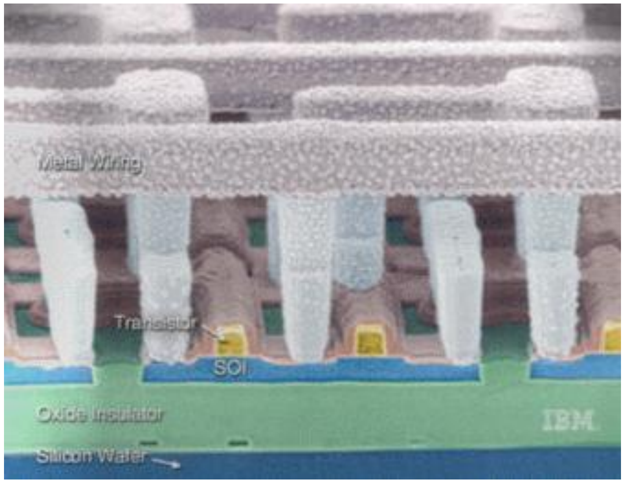
Semiconductor Technology





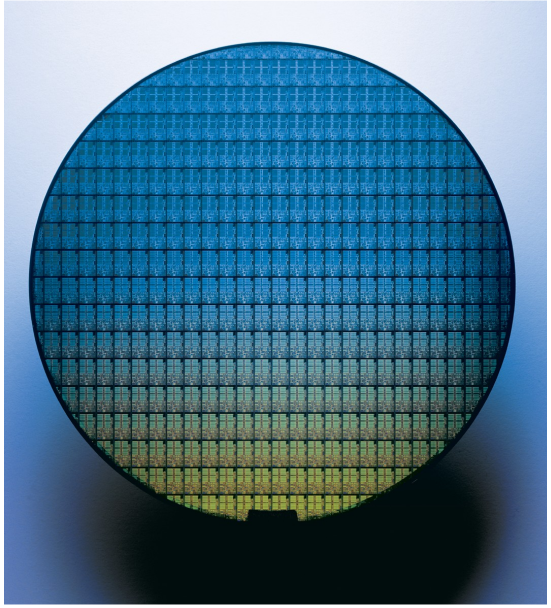
- Basic technology element: **MOSFET**
  - Solid-state component acts like electrical switch
  - **MOS**: metal-oxide-semiconductor
    - Conductor, insulator, semi-conductor
- **FET**: field-effect transistor
  - Channel conducts source→drain only when voltage applied to gate
- **Channel length**: characteristic parameter (short → fast)
  - Aka "feature size" or "technology"
  - Currently: 0.022 micron (μm), 22 nanometers (nm)
  - Continued miniaturization (scaling) known as "**Moore's Law**"
    - Won't last forever, physical limits approaching (or are they?)

Transistors and Wires



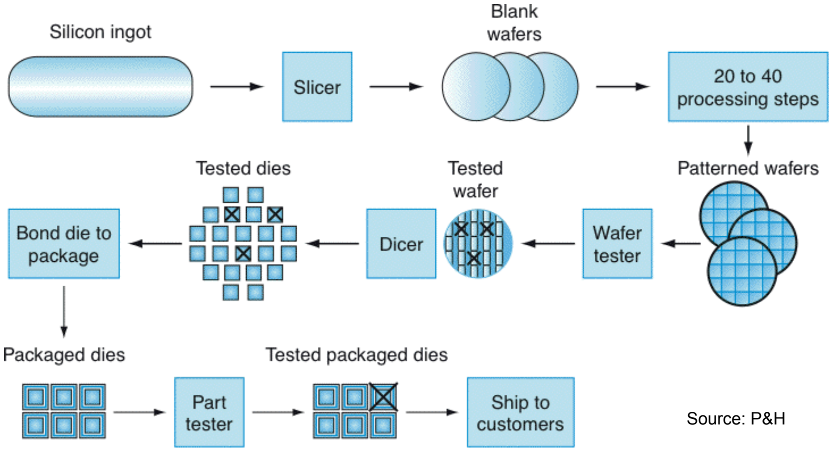
©IBM

From slides © Krste Asanović, MIT



Intel Pentium M Wafer

Manufacturing Steps

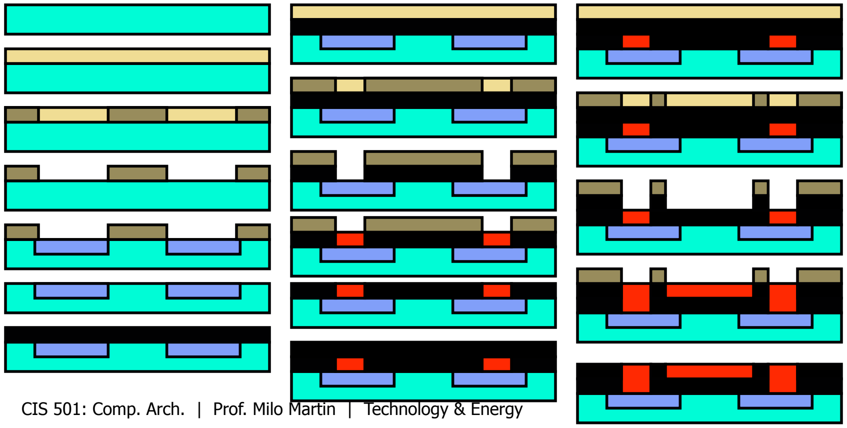


Source: P&H

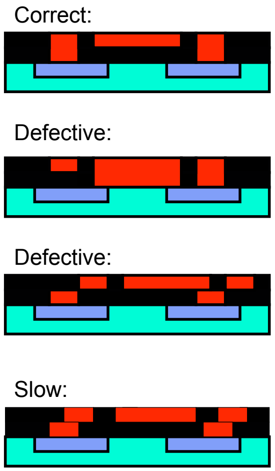
Manufacturing Steps

- Multi-step photo-/electro-chemical process
  - More steps, higher unit cost

+ Fixed cost mass production ($1 million+ for "mask set")



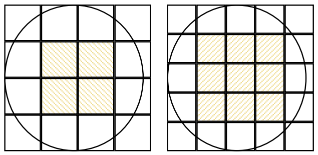
Manufacturing Defects

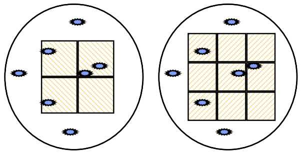


- Defects can arise
  - Under-/over-doping
  - Over-/under-dissolved insulator
  - Mask mis-alignment
  - Particle contaminants
- Try to minimize defects
  - Process margins
  - Design rules
    - Minimal transistor size, separation
- Or, tolerate defects
  - Redundant or "spare" memory cells
  - Can substantially improve yield

Cost Implications of Defects

- Chips built in multi-step chemical processes on **wafers**
  - Cost / wafer is constant, f(wafer size, number of steps)
- Chip (die) cost is related to **area**
  - Larger chips means fewer of them
- Cost is **more than** linear in area
  - Why? random defects
  - Larger chip, more chance of defect
    - Result: lower "yield" (fewer working chips)





- **Wafer yield**: % wafer that is chips
- **Die yield**: % chips that work
- Yield is increasingly non-binary - fast vs slow chips

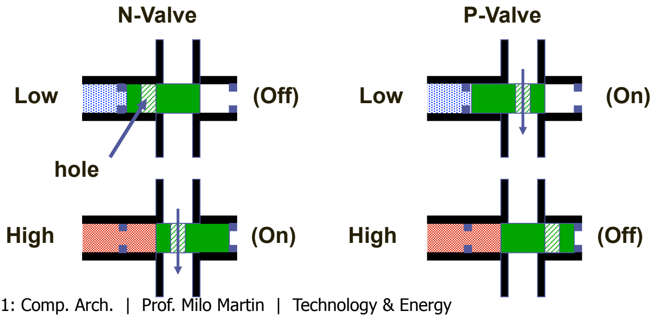
Manufacturing Cost

- **Chip cost vs system cost**
  - Cost of memory, storage, display, battery, etc.
- **Cost vs price**
  - Relationship complicated; **microprocessors not commodities**
  - Specialization, compatibility, different cost/performance/power
  - Economies of scale
- **Unit costs**: die manufacturing, testing, packaging, burn-in
  - Die cost based on area & defect rate (yield)
  - Package cost related to heat dissipation & number of pins
- **Fixed costs**: design & verification, fab cost
  - Amortized over "proliferations", e.g., Core i3, i5, i7 variants
  - Building new "fab" costs billions of dollars today
  - Both getting worse; trend toward "foundry" & "fabless" models

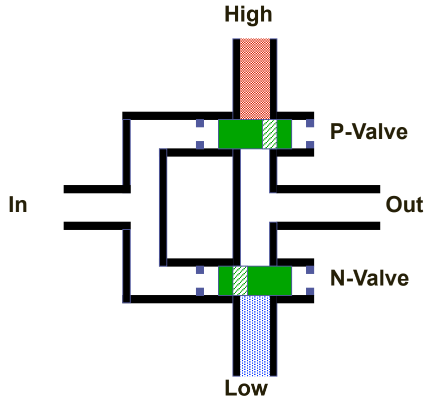
Transistor Switching Speed

A Transistor Analogy: Computing with Air

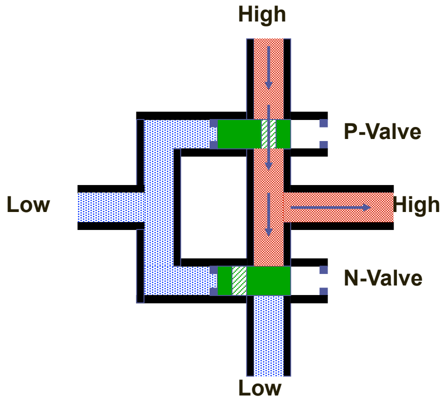
- Use air pressure to encode values
  - High pressure represents a "1" (blow)
  - Low pressure represents a "0" (suck)
- Valve can allow or disallow the flow of air
  - Two types of valves



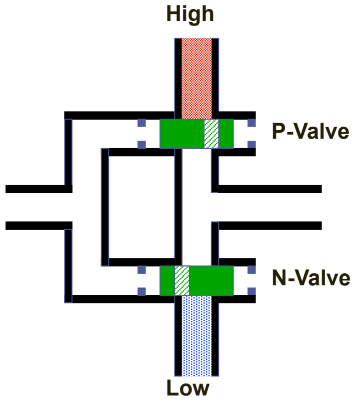
Pressure Inverter



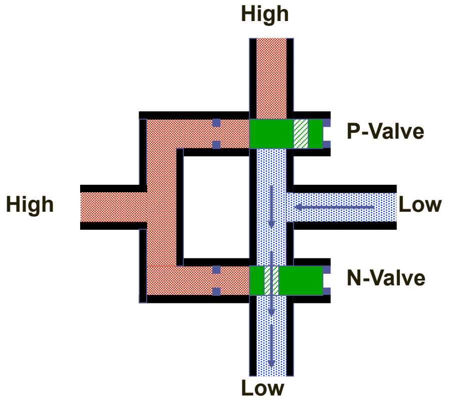
Pressure Inverter (Low to High)



Pressure Inverter



Pressure Inverter (High to Low)

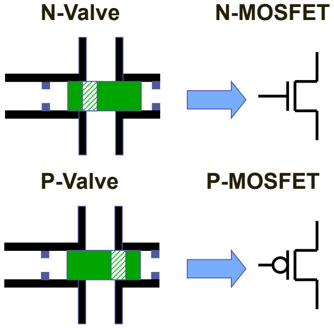


Analogy Explained

- Pressure differential → electrical potential (voltage)
  - Air molecules → electrons
  - Pressure (molecules per volume) → voltage
  - High pressure → high voltage
  - Low pressure → low voltage
- Air flow → electrical current
  - Pipes → wires
  - Air only flows from high to low pressure
  - Electrons only flow from high to low voltage
  - Flow only occurs when changing from 1 to 0 or 0 to 1
- Valve → transistor
  - The transistor: one of the century’s most important inventions

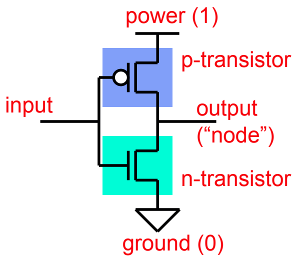
Transistors as Switches

- Two types
  - N-type
  - P-type
- Properties
  - Solid state (no moving parts)
  - Reliable (low failure rate)
  - Small (45nm channel length)
  - Fast (<0.1ns switch latency)



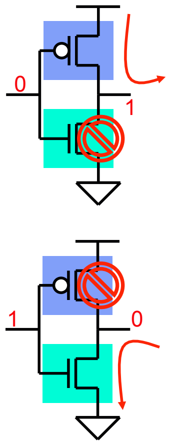
Complementary MOS (CMOS)

- Voltages as values
  - Power (V_{DD}) = “1”, Ground = “0”
- Two kinds of MOSFETs
  - **N-transistors**
    - Conduct when gate voltage is 1
    - Good at passing 0s
  - **P-transistors**
    - Conduct when gate voltage is 0
    - Good at passing 1s
- **CMOS**
  - Complementary n-/p- networks form boolean logic (i.e., gates)
  - And some non-gate elements too (important example: RAMs)



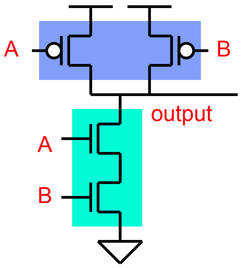
Basic CMOS Logic Gate

- **Inverter**: NOT gate
  - One p-transistor, one n-transistor
  - Basic operation
  - Input = 0
    - P-transistor closed, n-transistor open
    - Power charges output (1)
  - Input = 1
    - P-transistor open, n-transistor closed
    - Output discharges to ground (0)

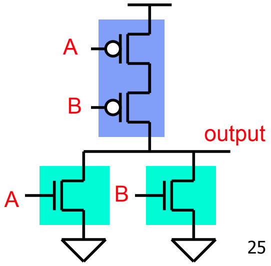


Another CMOS Gate Example

- What is this? Look at **truth table**
  - 0, 0 → 1
  - 0, 1 → 1
  - 1, 0 → 1
  - 1, 1 → 0
  - Result: **NAND** (NOT AND)
  - NAND is "universal"



  - What function is this?



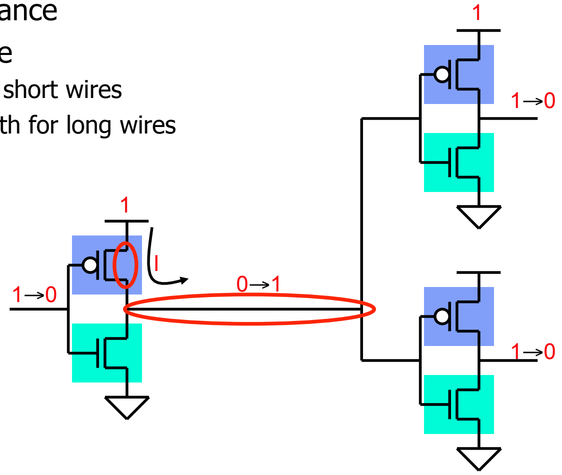
Technology Basis of Transistor Speed

- Physics 101: delay through an electrical component ~ **RC**
  - **Resistance (R)** ~ length / cross-section area
    - Slows rate of charge flow
  - **Capacitance (C)** ~ length * area / distance-to-other-plate
    - Stores charge
  - **Voltage (V)**
    - Electrical pressure
  - **Threshold Voltage (V_t)**
    - Voltage at which a transistor turns "on"
    - Property of transistor based on fabrication technology
  - **Switching time ~ to (R * C) / (V – V_t)**

- Two kinds of electrical components
  - CMOS transistors (gates, sources, drains)
  - Wires

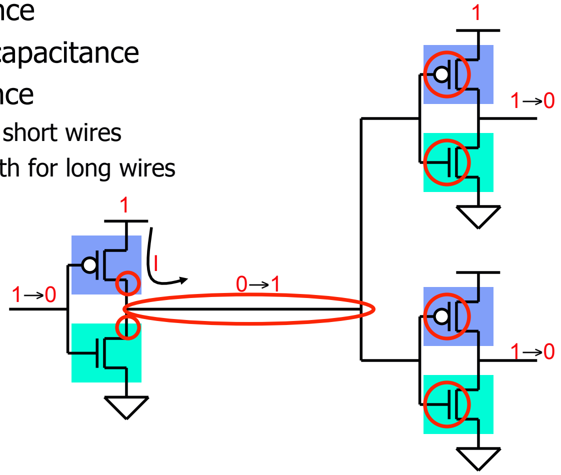
Resistance

- Channel resistance
- Wire resistance
  - Negligible for short wires
  - Linear in length for long wires

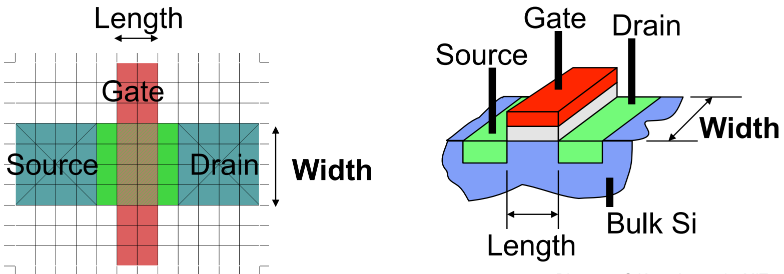


Capacitance

- Gate capacitance
- Source/drain capacitance
- Wire capacitance
  - Negligible for short wires
  - Linear in length for long wires



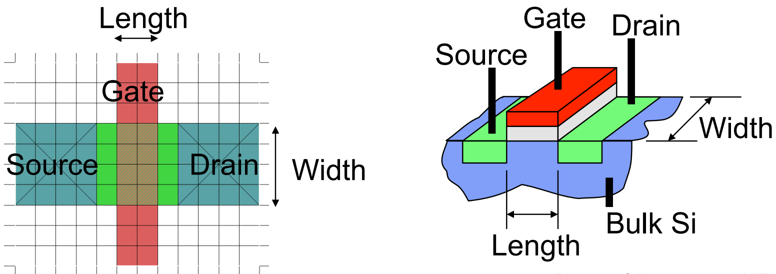
Transistor Geometry: Width



Diagrams © Krste Asanovic, MIT

- **Transistor width**, set by designer for each transistor
- Wider transistors:
  - **Lower resistance** of channel (increases drive strength) – good!
  - But, **increases capacitance** of gate/source/drain – bad!
- Result: set width to balance these conflicting effects

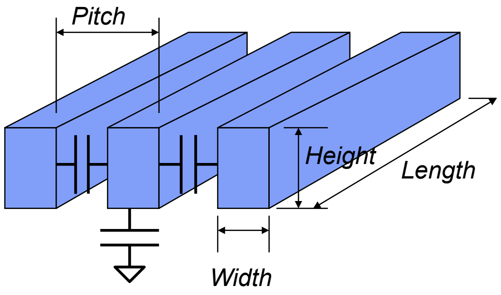
Transistor Geometry: Length & Scaling

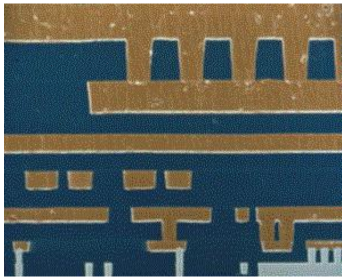


Diagrams © Krste Asanovic, MIT

- **Transistor length**: characteristic of "process generation"
  - "22nm" refers to the transistor gate length
- Each process generation shrinks transistor length by 1.4x
  - "Moore's law" -> roughly 2x improvement transistor density
  - Roughly linear improvement in switching speeds (lower resistance)

Wire Geometry





IBM CMOS7, 6 layers of copper wiring

- Transistors 1-dimensional for design purposes: **width**
- Wires 4-dimensional: **length**, **width**, **height**, **"pitch"**
  - Longer wires have more resistance (slower)
  - "Thinner" wires have more resistance (slower)
  - Closer wire spacing ("pitch") increases capacitance (slower)

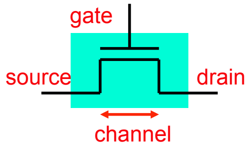
From slides © Krste Asanovic, MIT

Increasing Problem: Wire Delay

- RC Delay of wires
  - **Resistance** proportional to: resistivity * length / (cross section)
    - Wires with smaller cross section have higher resistance
    - Resistivity (type of metal, copper vs aluminum)
  - **Capacitance** proportional to length
    - And wire spacing (closer wires have large capacitance)
    - Permittivity or "dielectric constant" (of material between wires)
- Result: delay of a wire is **quadratic** in length
  - Insert "inverter" repeaters for long wires
  - Why? To bring it back to linear delay… but repeaters still add delay
- Long wires are getting relatively slow to transistors
  - And relatively longer time to cross relatively larger chips

Technology Scaling Trends

Moore's Law: Technology Scaling



- **Moore's Law**: aka "technology scaling"
  - Continued miniaturization (esp. reduction in channel length)
  - + Improves switching speed, power/transistor, area(cost)/transistor
  - – Reduces transistor reliability
  - Literally: DRAM density (transistors/area) doubles every 18 months
  - Public interpretation: performance doubles every 18 months
    - Not quite right, but helps performance in several ways…

Moore's Effect #1: Transistor Count

- Linear shrink in each dimension
  - 180nm, 130nm, 90nm, 65nm, 45nm, 32nm, 22nm, …
  - Each generation is a 1.414 linear shrink
    - Shrink each dimension (2D)
  - Results in 2x more transistors (1.414*1.414) per area

- Generally reduces cost per transistor

- More transistors can increase performance
  - Job of a computer architect: use the ever-increasing number of transistors
  - Today, desktop/laptop processor chips have ~1 billion transistors

Moore's Effect #2: RC Delay

- **First-order: speed scales proportional to gate length**
  - Has provided much of the performance gains in the past
- Scaling helps wire and gate delays in some ways…
  - + Transistors become shorter (Resistance↓), narrower (Capacitance↓)
  - + Wires become shorter (Length↓ → Resistance↓)
  - + Wire "surface areas" become smaller (Capacitance↓)
- Hurts in others…
  - – Transistors become narrower (Resistance↑)
  - – Gate insulator thickness becomes smaller (Capacitance↑)
  - – Wires becomes thinner (Resistance↑)
- What to do?
  - Take the good, use wire/transistor sizing to counter the bad
  - Exploit new materials: Aluminum → Copper, metal gate, high-K

Moore's Effect #3: Cost

- Mixed impact on unit integrated circuit cost
  - + Either lower cost for same functionality…
  - + Or same cost for more functionality
  - – Difficult to achieve high yields

– Increases startup cost

- More expensive fabrication equipment
- Takes longer to design, verify, and test chips

– Process variation across chip increasing

- Some transistors slow, some fast
- Increasingly active research area: dealing with this problem

Moore's Effect #4: Psychological

- **Moore's Curve**: common interpretation of Moore's Law
  - "CPU performance doubles every 18 months"
  - Self fulfilling prophecy: 2X every 18 months is ~1% per week
    - Q: Would you add a feature that improved performance 20% if it would delay the chip 8 months?
  - Processors under Moore's Curve (arrive too late) fail spectacularly
    - E.g., Intel's Itanium, Sun's Millennium

Moore's Law in the Future

- Won't last forever, approaching physical limits
  - "If something must eventually stop, it can't go on forever"
  - But betting against it has proved foolish in the past
  - Perhaps will "slow" rather than stop abruptly

- Transistor count will likely continue to scale
  - "Die stacking" is on the cusp of becoming main stream
  - Uses the third dimension to increase transistor count

- But transistor performance scaling?
  - Running into physical limits
  - Example: gate oxide is less than 10 silicon atoms thick!
    - Can't decrease it much further
  - Power is becoming the limiting factor

Power & Energy

Power/Energy Are Increasingly Important

- **Battery life** for mobile devices
  - Laptops, phones, cameras
- **Tolerable temperature** for devices without active cooling
  - Power means temperature, active cooling means **cost**
  - No room for a fan in a cell phone, no market for a hot cell phone
- **Electric bill** for compute/data centers
  - Pay for power twice: once in, once out (to cool)
- **Environmental concerns**
  - "Computers" account for growing fraction of energy consumption

Energy & Power

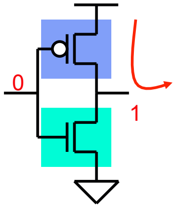
- **Energy**: measured in Joules or Watt-seconds
  - Total amount of energy stored/used
  - Battery life, electric bill, environmental impact
  - Instructions per Joule (car analogy: miles per gallon)
- **Power**: energy per unit time (measured in Watts)
  - Related to "performance" (which is also a "per unit time" metric)
  - Power impacts power supply and cooling requirements (cost)
    - Power-density (Watt/mm^2): important related metric
  - Peak power vs average power
    - E.g., camera, power "spikes" when you actually take a picture
  - Joules per second (car analogy: gallons per hour)
- Two sources:
  - **Dynamic power**: active switching of transistors
  - **Static power**: leakage of transistors even while inactive

Recall: Tech. Basis of Transistor Speed

- Physics 101: delay through an electrical component ~ **RC**
  - **Resistance (R)** ~ length / cross-section area
    - Slows rate of charge flow
  - **Capacitance (C)** ~ length * area / distance-to-other-plate
    - Stores charge
  - **Voltage (V)**
    - Electrical pressure
  - **Threshold Voltage (V_t)**
    - Voltage at which a transistor turns "on"
    - Property of transistor based on fabrication technology
  - **Switching time ~ to (R * C) / (V – V_t)**

Dynamic Power

- **Dynamic power ($P_{dynamic}$)**: aka switching or active power
  - Energy to switch a gate (0 to 1, 1 to 0)
  - Each gate has capacitance (C)
    - Charge stored is ∝ C * V
    - Energy to charge/discharge a capacitor is ∝ to C * V^2
    - Time to charge/discharge a capacitor is ∝ to V
      - Result: frequency ~ to V
  - **$P_{dynamic}$ ≈ N * C * V^2 * f * A**
    - N: number of transistors
    - C: capacitance per transistor (size of transistors)
    - V: voltage (supply voltage for gate)
    - f: frequency (transistor switching freq. is ∝ to clock freq.)
    - A: activity factor (not all transistors may switch this cycle)

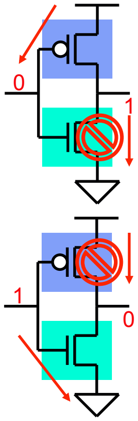


Reducing Dynamic Power

- Target each component: $\mathbf{P_{dynamic} \approx N * C * V^2 * f * A}$
- **Reduce number of transistors** (N)
  - Use fewer transistors/gates
- **Reduce capacitance** (C)
  - Smaller transistors (Moore's law)
- **Reduce voltage** (V)
  - Quadratic reduction in energy consumption!
  - But also slows transistors (transistor speed is ~ to V)
- **Reduce frequency** (f)
  - Slower clock frequency (reduces power but not energy) Why?
- **Reduce activity** (A)
  - "Clock gating" disable clocks to unused parts of chip
  - Don't switch gates unnecessarily

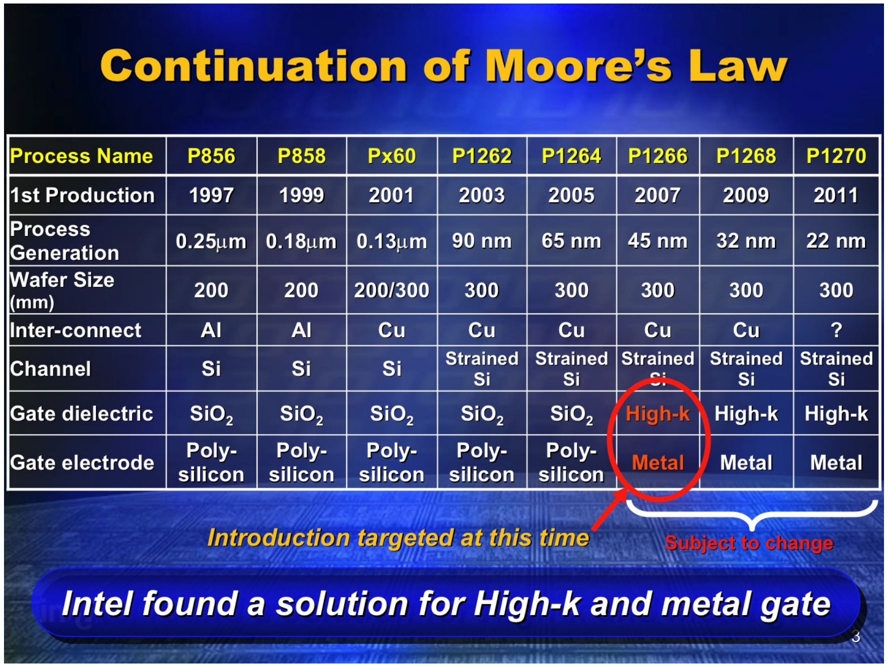
Static Power

- **Static power (P_{static})**: aka idle or leakage power
  - Transistors don't turn off all the way
  - Transistors "leak"
    - Analogy: leaky valve
  - $\mathbf{P_{static} \approx N * V * e^{-V_t}}$
  - N: number of transistors
  - V: voltage
  - **V_t (threshold voltage)**: voltage at which transistor conducts (begins to switch)
- Switching speed vs leakage trade-off
- The lower the V_t:
  - Faster transistors (linear)
    - Transistor speed ∝ to V – V_t
  - Leakier transistors (exponential)



Reducing Static Power

- Target each component: $\mathbf{P_{static} \approx N * V * e^{-V_t}}$
- **Reduce number of transistors** (N)
  - Use fewer transistors/gates
- **Disable transistors** (also targets N)
  - "Power gating" disable power to unused parts (long latency to power up)
  - Power down units (or entire cores) not being used
- **Reduce voltage** (V)
  - Linear reduction in static energy consumption
  - But also slows transistors (transistor speed is ~ to V)
- **Dual V_t** – use a mixture of high and low V_t transistors
  - Use slow, low-leak transistors in SRAM arrays
  - Requires extra fabrication steps (cost)
- **Low-leakage transistors**
  - High-K/Metal-Gates in Intel's 45nm process, "tri-gate" in Intel's 22nm
- Reducing frequency can hurt energy efficiency due to leakage power



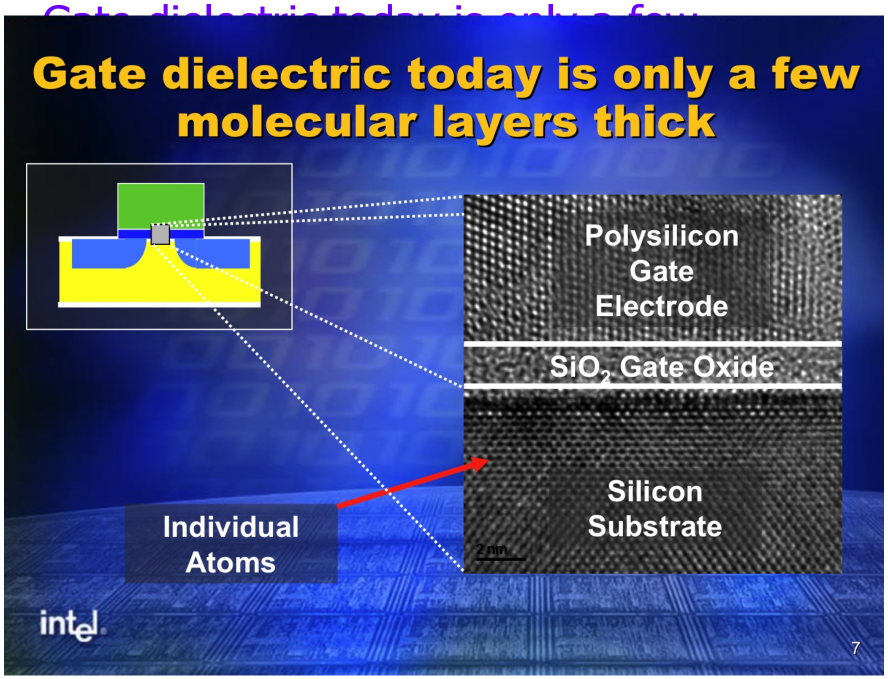
Continuation of Moore's Law

Process Name	P856	P858	Px60	P1262	P1264	P1266	P1268	P1270
1st Production	1997	1999	2001	2003	2005	2007	2009	2011
Process Generation	0.25μm	0.18μm	0.13μm	90 nm	65 nm	45 nm	32 nm	22 nm
Wafer Size (mm)	200	200	200/300	300	300	300	300	300
Inter-connect	Al	Al	Cu	Cu	Cu	Cu	Cu	?
Channel	Si	Si	Si	Strained Si	Strained Si	Strained Si	Strained Si	Strained Si
Gate dielectric	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	High-k	High-k	High-k
Gate electrode	Poly-silicon	Poly-silicon	Poly-silicon	Poly-silicon	Poly-silicon	Metal	Metal	Metal

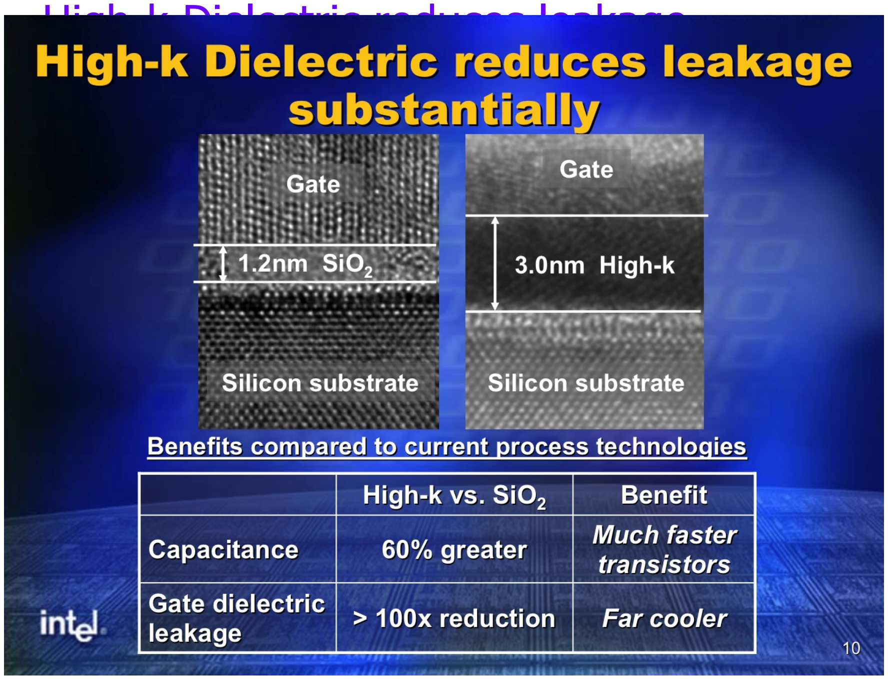
Introduction targeted at this time

Subject to change

Intel found a solution for High-k and metal gate



Gate dielectric today is only a few molecular layers thick



High-k Dielectric reduces leakage substantially

Benefits compared to current process technologies

	High-k vs. SiO$_2$	Benefit
Capacitance	60% greater	*Much faster transistors*
Gate dielectric leakage	> 100x reduction	*Far cooler*

Dynamic Voltage/Frequency Scaling

- **Dynamically trade-off power for performance**
  - Change the voltage and frequency at runtime
  - Under control of operating system
- Recall: $P_{dynamic} \approx N * C * \mathbf{V^2} * \mathbf{f} * A$
  - Because frequency ∝ to $V - V_t$…
  - $P_{dynamic}$ ∝ to $\mathbf{V^2(V - V_t)} \approx \mathbf{V^3}$
- Reduce both voltage and frequency linearly
  - **Cubic decrease in dynamic power**
  - Linear decrease in performance (actually sub-linear)
    - Thus, only about quadratic in energy
  - Linear decrease in static power
    - Thus, static energy can become dominant
- Newer chips can adjust frequency on a per-core basis

Dynamic Voltage/Frequency Scaling

	Mobile PentiumIII "**SpeedStep**"	Transmeta 5400 "LongRun"	Intel X-Scale (StrongARM2)
f (MHz)	300–1000 (step=50)	200–700 (step=33)	50–800 (step=50)
V (V)	0.9–1.7 (step=0.1)	1.1–1.6V (cont)	0.7–1.65 (cont)
High-speed	3400MIPS @ 34W	1600MIPS @ 2W	800MIPS @ 0.9W
Low-power	1100MIPS @ 4.5W	300MIPS @ 0.25W	62MIPS @ 0.01W

- Dynamic voltage/frequency scaling
  - **Favors parallelism**
- Example: Intel Xscale
  - 1 GHz → 200 MHz reduces energy used by 30x
    - But around 5x slower
  - 5 x 200 MHz in parallel, use **1/6th the energy**
  - Power is driving the trend toward multi-core

Moore's Effect on Power

+ Moore's Law reduces power/transistor…
  - Reduced sizes and surface areas reduce capacitance (C)

– …but increases power density and total power
  - By increasing transistors/area and total transistors
  - Faster transistors → higher frequency → more power
  - Hotter transistors leak more (thermal runaway)

- What to do? Reduce voltage (V)
  + Reduces dynamic power quadratically, static power linearly
    - Already happening: Intel 486 (5V) → Core2 (1.3V)
  - Trade-off: reducing V means either…
    – Keeping V_t the same and reducing frequency (f)
    – Lowering V_t and increasing leakage exponentially
  - Use techniques like high-K and dual-V_T
- The end of voltage scaling & "dark silicon"

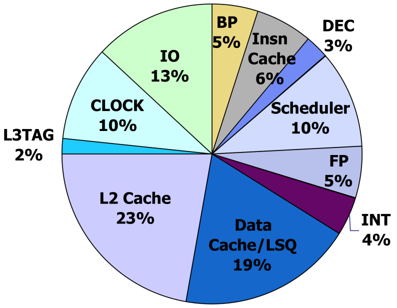
Trends in Power

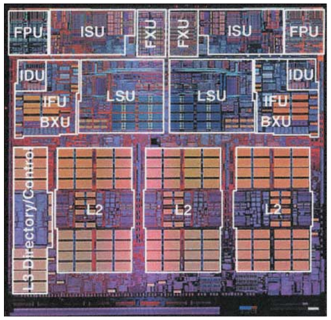
	386	486	Pentium	Pentium II	Pentium4	Core2	Core i7
Year	1985	1989	1993	1998	2001	2006	2009
Technode (nm)	1500	800	350	180	130	65	45
Transistors (M)	0.3	1.2	3.1	5.5	42	291	731
Voltage (V)	**5**	**5**	**3.3**	**2.9**	**1.7**	**1.3**	**1.2**
Clock (MHz)	16	25	66	200	1500	3000	3300
Power (W)	1	5	16	35	**80**	**75**	**130**
Peak MIPS	6	25	132	600	4500	24000	52800
MIPS/W	6	5	8	17	56	320	406

- Supply voltage decreasing over time
  - But "voltage scaling" is perhaps reaching its limits
- Emphasis on power starting around 2000
  - Resulting in slower frequency increases
  - Also note number of cores increasing (2 in Core 2, 4 in Core i7)

Processor Power Breakdown

- Power breakdown for IBM POWER4
  - Two 4-way superscalar, 2-way multi-threaded cores, 1.5MB L2
  - Big power components are L2, data cache, scheduler, clock, I/O
  - Implications on "complicated" versus "simple" cores





Implications on Software

- Software-controlled dynamic voltage/frequency scaling
  - Example: video decoding
    - Too high a clock frequency – wasted energy (battery life)
    - Too low a clock frequency – quality of video suffers
  - "Race to sleep" versus "slow and steady" approaches
- Managing low-power modes
  - Don't want to "wake up" the processor every millisecond
- Tuning software
  - Faster algorithms can be converted to lower-power algorithms
  - Via dynamic voltage/frequency scaling
- Exploiting parallelism & heterogeneous cores
  - NVIDIA Tegra 3: 5 cores (4 "normal" cores & 1 "low power" core)
- Specialized hardware accelerators

Summary

Technology Summary

- Has a first-order impact on computer architecture
  - Performance (transistor delay, wire delay)
  - Cost (die area & defects)
  - **Changing rapidly**
- Most significant trends for architects
  - More and more transistors
    - What to do with them? → integration → **parallelism**
  - Logic is improving faster than memory & cross-chip wires
    - "Memory wall" → caches, more integration

  Rest of course
- Power and energy
  - Voltage vs frequency, parallelism, special-purpose hardware
- This unit: a quick overview, just scratching the surface