Future miscellanea

- Moore's Law has held for 50+ years
  - Geometrical (linear) shrink by $\sqrt{2}$ every 18-24 months
  - Self-fulfilling prophecy: competing manufacturers 'believe' it
  - Growth forecast to *slow down* fairly soon
    New generation every 3(?) years

Eventually this *must* stop!

Limits:

- design — someone has to want/occupy the resource
- physical — it's made from real atoms
- manufacturing — it's got to *be* made (and work) …
- economics — … for an affordable price
- …

Limits to Moore's Law

Design

Producing devices is all very well but they need to do something useful. There is an increasing 'design gap' between what can be made and how it is designed. Whilst CAD tools and synthesis make designers more productive much of the transistor budget is devoted to repetitive structures such as memories. The expanding number of 'cores' on processors is an obvious symptom.

Physical

Within an order of magnitude, all atoms are about the same size which is one angstrom (Å) or 10^{-10} m (or 0.1 nm) in radius. A 22 nm gate length is therefore about 100 atomic diameters across. The dielectric thickness is already much less, perhaps ~1 nm so only a few atoms thick

Clearly a single atom is a hard limit to anything physical. However, at least using CMOS technology, the material must also retain its 'bulk' properties.

Manufacturing

Chips are made using photolithography and the feature size (22 nm) is already much smaller than the wavelength of the light used (193 nm) which is towards the middle of the ultraviolet spectrum. Using shorter wavelength light is increasingly difficult as optics is hard – think about trying to make lenses for X-rays!

It is possible to achieve much higher resolutions using electron or ion beams; compare what is visible under a light and an electron microscope. However a beam needs to be scanned whereas optical exposure can print over large areas.

Therefore making structures photolithographically is akin to printing whereas using direct beams is more like handwriting. It is clear which technique is amenable for mass production!

Economics

As chips get smaller they get harder to make. The machinery needs to be more precise (so it gets more expensive) and there are more manufacturing steps (which increase the cost). The resulting products still need to be cheap; if the cost rises more than the benefit then shrinking becomes uneconomic.

International Technology Roadmap for Semiconductors

The ITRS is a set of internationally produced documents which attempts to predict the future of semiconductor devices. It is produced by leading experts and recognised as the best possible forecast of what should be possible, and when.

This is probably the definitive source for information in this area. The chapter headings are given here to illustrate the scope of the work.

- System Drivers
- Design
- Test and Test Equipment
- Process Integration, Devices, and Structures
- RF and A/MS Technologies for Wireless Communications
- Emerging Research Devices
- Emerging Research Materials
- Front End Processes
- Lithography
- Interconnect
- Factory Integration
- Assembly & Packaging
- Environment, Safety & Health
- Yield Enhancement
- Metrology
- Modeling & Simulation

Much of the information is available from: `http://www.itrs.net/`

Power

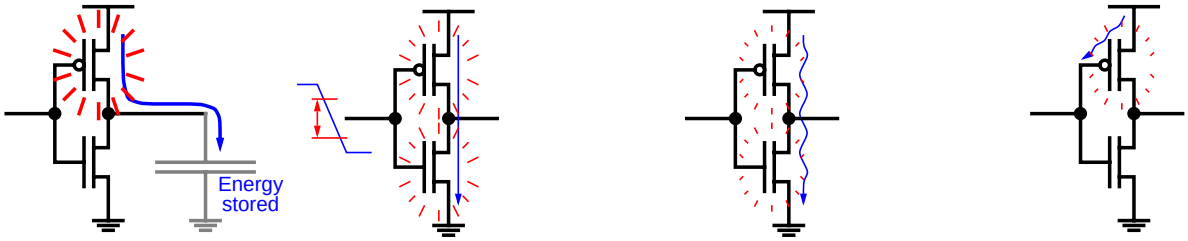
Power dissipation must be controlled for two reasons:

- Battery life
- Cooling

For decades power dissipation was small in CMOS circuits: it has been rising steadily.

Energy is ‘lost’ when some charge finds its way to a lower potential. It reappears as heat.

There are various causes of energy dissipation in CMOS



Dynamic Static

Leakage is an increasing problem.

Power

Dynamic dissipation

When a gate switches its output voltage moves from one rail to the other, either charging or discharging a load capacitance. The charge on a capacitor at voltage V is $Q = C \cdot V$ and, over a complete cycle, this is moved through a potential V (i.e. from ‘power’ to ‘ground’). The energy dissipated over the cycle is therefore: $E = C \cdot V^2$

The energy *stored* on a capacitor is $E = \frac{1}{2} \cdot C \cdot V^2$; it is thus apparent that half the energy is dissipated when charging the load, the other half is stored and dissipated when the node is discharged.

The energy is dissipated chiefly in the channel(s) of the (dis)charging transistor(s) although some will be lost moving charge through resistive wires.

Nodes can switch at the clock frequency (f), thus more energy is dissipated if the clock is faster. Except for the clock, nodes typically switch rather slower than this; for example the fastest one might expect switching is at half the clock rate (i.e. one edge per *active* clock edge) and most switch less than this. The typical switching rate is represented by an *activity factor* (α).

The major dynamic power is therefore: $P = \alpha \cdot f \cdot C \cdot V^2$

Note: as processes advance the node capacitance shrinks and the voltage reduces – which is a significant gain because power is proportional to voltage *squared*. However the frequency may increase and the *number* of nodes increases.

When a gate switches from driving high to driving low (say), the PMOS stack turns off and the NMOS stack turns on. This does not happen instantaneously; the input edge has a finite speed. Depending on the transistor thresholds, it is possible that both the stacks are ‘on’ at the same time, permitting a pulse of ‘short circuit’ or ‘crowbar’ current to run directly between the power rails. This wastes energy. It can be alleviated by:

- keeping the edges fast
  made harder by increasingly resistive wiring
- using high threshold transistors
  which increases the propagation delay of the gate

Static dissipation

There is a small (but increasing) power dissipation in CMOS logic due to **leakage** currents. Unlike the dynamic (switching) dissipation this is continuous. Several sources can be identified.

- Subthreshold leakage
  A transistor is not an on/off switch although it is used as such; it operates in a continuous domain. Therefore it can only be thought of as ‘nearly off’ (or ‘quite on’). Therefore there is always some small current leaking through the transistor stacks. As supply voltages are reduced the ‘distance’ between these points is reduced and it is not possible to ‘as off’ as it used to be, thus this current is becoming a more serious concern.
- Gate tunnelling
  The transistor gate is insulated from the substrate by a thin ‘oxide’[1] spacer. Electrons can *tunnel* through this insulator, the probability rising swiftly as it becomes very thin. This is a loss of charge which needs replenishing, thus another power waste.
- Substrate leakage
  All the transistors are built on the same silicon substrate and there is some leakage of charge into this. This is quite small in modern processes.

1. May no longer be SiO_2, but principle still applies.

Power control

Power is now **the limiting factor** in high-performance ICs.

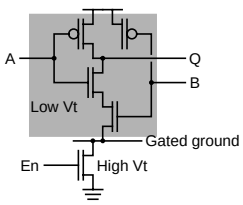
- **Clock gating** is now quite common. It isolates parts of a clock tree which are not in current use.
- **Glitch reduction**. Switching loads is expensive. A net which switches unnecessarily (twice!) due to a glitch generated by logic races is a Bad Thing. It is quite difficult to design out; in detail this may be a CAD issue.
- **Latching values** – especially long distance buses – so they don't 'flap about' when unused is a Good Thing. This is superior to gating them (e.g. to '0') because that could introduce extra switching.
- **Dynamic voltage scaling**. Slow the clock down then reduce the supply voltage. Particularly effective on V^2.
- **Leakage**: lower supply voltage reduces leakage. Techniques include **power gating** and the judicious employment of **multi-threshold** transistors.

Reducing Leakage

Transistors with a **high threshold** voltage (Vt) will be further below that threshold when turned 'off'; this will reduce the subthreshold leakage. This will also mean that the 'overlap' of the P and N stacks being 'on' at the same time can be reduced or 'eliminated'.

The disadvantage in raising the Vt is that the driving transistors will turn on later during their input edge, thus reducing the switching speed of the gate.

Subthreshold leakage may be reduced geometrically by reducing the width of a transistor but this reduces its drive. A similar effect is possibly by *lengthening* the gate of a transistor (i.e. the longer channel will have a higher impedance). The transistor will also be correspondingly weaker when 'on' however, and its capacitance will be raised. Long gates may be used in non-driving applications however where specifically weakened transistors may be desirable.



If more than one 'speed' of transistor is available a compromise between speed and power can be achieved by **power gating**. The adjacent figure shows a 2-input NAND gate with a power gate. (This is sometimes called Multi-Threshold CMOS.)

The gate is made from high-speed (low Vt) transistors but a series high Vt transistor is included. This last transistor is switched 'on' whilst the circuit is operating but is turned 'off' (by some means, hardware or software) when the circuit is idle. This series transistor reduces subthreshold leakage when the circuit is 'powered down'. Switching on and off imposes some delay, however, especially if the gated ground supplies many (related) gates. This technique is also possible using PMOS transistors, or even both.

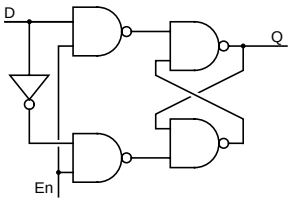
Memory

Memory cells rarely switch yet memories have high transistor density. Leakage is therefore relatively more significant in memories. Gating the power off a CMOS memory will cause it to lose its contents. However it is feasible to 'power-down' memories by reducing their supply voltage which will reduce leakage current.

The power must be ramped up again before reading, which takes some time.

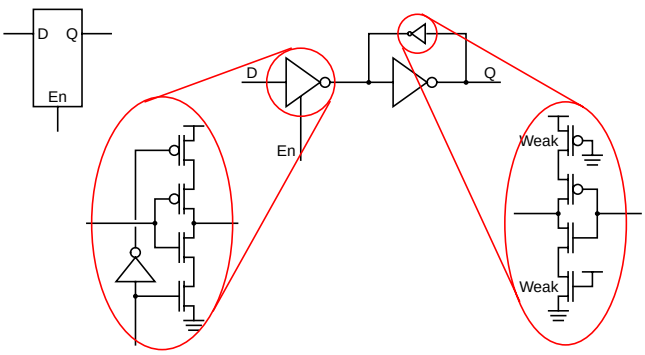
Example circuit: transparent latch

In detail, VLSI circuits may not be what you might draw with gates. The adjacent figure shows a transparent latch made from NAND gates.



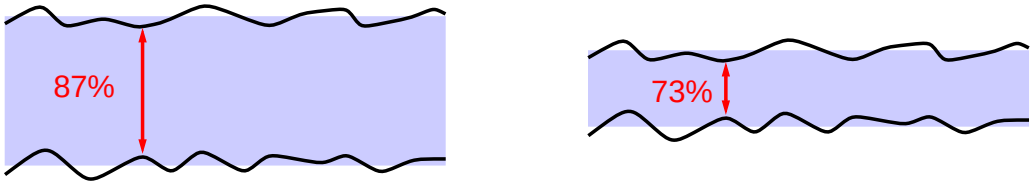
This circuit requires 18 transistors. It has a two-input NAND driving its output so the transistors in the final N stack need to be enlarged. (The equivalent using NOR gates is inferior; *can you remember why?*)

The second example shows a more probable VLSI implementation. It uses 12 transistors. The input stage is a tristate inverter. The output is driven by an inverter (which is the best form of driver with only one P and N transistor). There are two outputs connected together, which is normally a very bad idea, but the feedback is deliberately weak as it does not drive any output load. The input drive is much stronger and, in any contest, quickly overpowers the feedback which thereafter switches. The feedback inverter replaces any leakage when the input is made high impedance. By weakening the feedback with unswitched transistors the capacitive load is reduced.



Variation

As features become narrower the effect of ‘rough edges’ becomes more significant.



As areas shrink the statistical variation in doping becomes more significant.

Channel doping density currently may be ~100 atoms.

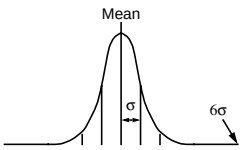
The effect is less control over transistor drive strength, wire resistance, (yield) …

The first price is larger ‘safety margins’ in the clock frequency.

Variation

At time of writing (Sept. 2013) the leading edge *sample* chips have gate lengths ~14 nm – ma be available to end users 2014. How much further can features be shrunk? Probably to ~10 nm (another generation since a doubling of transistor density (over an *area*) equates to a linear shrink of $\sqrt{2}$). Current predictions for this come about around 2017. Further than that is … more challenging: maybe 7 nm in 2020?

ICs are mass-produced. They are subject to manufacturing fluctuations such as **line-edge roughness** and **random dopant fluctuations**. These characteristics mean that feature widths and transistor strengths vary randomly with some probability.



Statistically, a normal (Gaussian) distribution represents random fluctuations from a mean value and is a reasonable representation of what might occur by chance in manufacture. The distribution is often called a ‘bell curve’ for obvious reasons. The width of the bell is specified as a standard deviation (σ). It is expected that about 68% of the population are within σ of the mean value. As the process gets more difficult to control the standard deviation will increase.

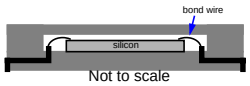
A manufacturing goal is 6σ – the target of ensuring that a manufactured article does not deviate from the average by more than six standard deviations. In a Gaussian distribution this should include 99.99966% of the product.

The chance of lying outside this range is about two in a billion. With perhaps three billion transistors on a chip …

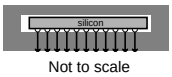
System assembly

[These notes accompany slide 5.]

The ‘traditional’ assembly is to package each chip separately, **bonding** its peripheral pads to the package using **bond wires**. The package contains connections to the outside world (‘**pins**’).



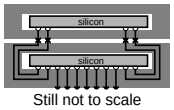
With a high demand for connections, ships may be ‘flipped’ and bonded with bumps. Similarly, packages may have connections across all their lower surface: a Ball Grid Array (BGA).



Packages are then interconnected using a multi-layer Printed Circuit Board (PCB). Although both sides may be used, this is largely a 2D structure.

The packages and, particularly, the PCB tracks (wires) contribute significant capacitance. They also occupy quite an awkward (big, flat) space

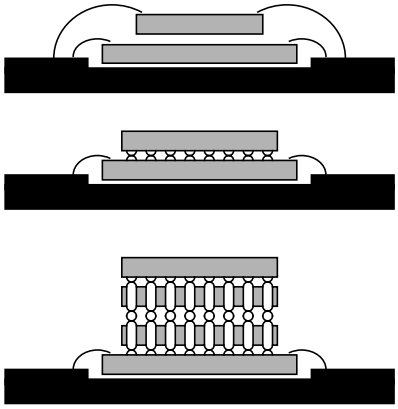
One partial solution is to use Package-on-Package (PoP) technology, where one chip’s package is ‘piggybacked’ on another. Example: Raspberry Pi mounts a memory on top of the ARM SoC.



The space required and the inter-chip capacitance is much reduced. The idea can be taken further …

Another Dimension

- 3D packaging
  - System in package
  - Flip-chip
  - Through-Silicon Vias



- 3D components
  - Attractive idea: infeasible?

Three dimensional assembly

Chips are still basically two-dimensional structures. Another way of increasing computational density – and possibly consequently shortening wire length – is to move to three dimensions.

There are numerous potential benefits including:

- Smaller space
  - lower system cost
- Shorter wires
  - lower delays and power
- More connections (using through silicon vias)
  - particularly attractive for bandwidth to RAM
- Greater flexibility
  - can integrate chips from different *processes* e.g. processor with DRAM

Naturally there are also some drawbacks:

- Complexity – more to go wrong
  - vias need to be built
  - need alignment in assembly
- More testing problems
- Heat dissipation
  - more power/unit volume
  - poorer thermal contact: less heat sinking

3D on one chip

Through Silicon Vias

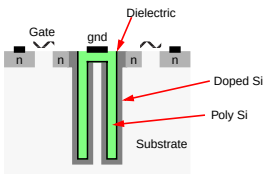
A TSV is small in one sense (maybe 10-100 µm across) but this is a large obstacle compared with the size of standard cells.

The silicon must be mechanically ground thin. Back-grinding wafers to fit them in thinner packages is now standard anyway. Some sort of mechanical support is needed to stop the wafer cracking however; this may be used to thin the silicon to 200-300 µm. For TSVs this may be reduced to maybe 50 µm.

Vias are cut by etching away silicon; a technique called Deep Reactive-Ion Etching can be used to cut near vertically. The holes are then insulated and metal-filled.

They may be made before or after metallisation; the former may be routed over, the latter reach both sides of the chip.

Trench Capacitors



Two DRAM cells (cross section)

DRAMs offer high memory density by using a small component count. Data is stored as charge on a capacitor. To make an adequate capacitor takes a certain area which can reduce the memory density. Thus it is now usual to build the capacitors not *on* but *into* the silicon substrate, thus saving considerable area.

Making a chip involves a large number of processing steps. Each step has a cost and introduces another risk of a fault. It is best to use the feasible minimum number of steps.

The steps involved in DRAM (making trench capacitors, for example) are not used in logic. EEPROM/Flash memory uses a different set of steps.

SRAM is easily manufacturable on a logic process. Thus it is common to integrate SRAM onto logic chips. EEPROM is less common although present on some microcontrollers. Integrating processing and DRAM is unusual. This means there is considerable attraction in being able to assemble stacks of different chips in a small space.

Cooling is, of course, still a potential problem!

Exotic stuff

- Can the synchronous model be maintained?
  - self-timed circuits may tolerate transistor variation more effectively
- Error tolerant circuits?
  - error correcting codes are common in communications and memories
  - can something similar be done in processing?

- New technology: will CMOS be supplanted?
  - Single electron transistors
  - Quantum dots
  - Spintronics
  - Graphene et alia
  - Molecular mechanics
  - …

Nearly all the progress in computing has been driven by the implementation technology.

- Although there is much research in alternative technology there is no clear successor to CMOS at this time.
- Moore's Law will soon start cracking in the sense that progress will slow down.
- Eventually the 'law' will break completely and progress will (largely) halt.
- However, by then it seems certain that there will be (the ability to produce) chips with hundreds – and probably thousands – of computer cores in a fingernail-sized blob.
- Alternatively there may be other things to do with 100 000 000 000 transistors.
- What are *you* going to do with them?