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!

Supply voltage reduction
For many years ‘popular’ logic devices used a 5 V supply. Two big problems as devices have shrunk:

- Power dissipation too great (esp. CMOS V²)
- Transistor dielectric breakdown

To combat these, core supply voltages have been gradually reduced, now to 1 V or below; a trend which must continue.

At PCB level, where different parts are assembled (and electromagnetic interference will be greater due to the larger ‘aerials’ (tracks)) the voltage reduction has been more conservative, having largely reduced only to 3.3 V (a.k.a. 3V3). Of course this necessitates level shifting at the chip’s pads.

Some implications

High-K dielectric
‘K’ is the ‘dielectric constant’, more properly called the ‘relative permittivity’ (εᵣ). It is a measure of the field strength in the dielectric. Replacing SiO₂ as a dielectric allows a thicker separator (less leakage) yet greater capacitance (which is what switches the FET).

Materials containing elements such as hafnium or zirconium are typical.
Variability

Maybe the biggest challenge to future devices is random variability in the manufacturing processes.

- Diffusing atoms act statistically:
  - with large numbers there is a strong likelihood that the actual number will be close to the mean
  - with small numbers statistical variations are proportionately larger
  - transistor properties are becoming less predictable \( \Rightarrow \) larger margins allowed

- Etching of tracks (etc.) will experience similar edge roughness, thinning or bridging.

Reliability

Guaranteeing the reliability of future devices may be difficult.

One future challenge is making reliable computing devices from less reliable parts. This type of engineering is common in certain applications areas through the use of redundancy – for example a commercial aeroplane may have two engines but can fly with only one, engines being regarded as one of the less reliable components. (They also have multiple redundant instruments, hydraulics etc.)

Integrated circuits have been regarded as quite reliable (once the manufacturing ‘defects’ have been identified by testing and those chips discarded). The reliability of the tests is threatened (do you want something which ‘just passed’?) as is the production yield.

Two questions for the near future are ‘where it is appropriate to introduce redundancy?’ and ‘how should it be managed?’ It is easy to include more microprocessor systems in a Symmetric Multiprocessor (SMP), assuming that one or more may not work, and the software to handle this \( \text{may} \) be straightforward … or not. Designing in redundancy at circuit level appears to be more difficult although this could be transparent to the software layer if achieved.

One place where redundancy has been used extensively is RAM, especially on server systems. Often error detection and correction codes (e.g. Hamming codes) are employed to increase the memory’s reliability.

Figures from studies of DRAM error rates vary widely. Some of the worst-case numbers suggest perhaps a one-bit ‘drop’ (fault) in a gigabyte of DRAM every two hours. A typical cause is cosmic radiation. (Error rates on spacecraft are significantly higher!) See “ECC memory” for more details.

Smaller dice?

One trend is that the typical physical size of silicon die is shrinking – although capacity is still increasing. This helps in keeping an economic yield.
Three dimensional structures

First ICs had one layer of metal interconnect.
Current chips may have 10+ ... but the logic is still, basically 2D on the silicon surface.
Arguable exceptions e.g. DRAM ‘trench’ capacitor:

Making a truly 3D chip is a really serious challenge – not least in terms of power density

Making a stack of 2D chips is easier, and still:
- saves space from packaging view
- reduces chip interconnection wire length

Problem
- How to connect chips together?

Some other manufacturing techniques

Strained silicon
A term you may come across. Silicon is a semiconductor (you knew that) and the electrons (or holes) can move through the crystal lattice with a certain mobility. Like any crystal structure it has a characteristic atomic spacing.

Straining the silicon involves distorting the lattice distances, which can improve the carrier mobility and thus the transconductance. In short, it makes a transistor of a given size ‘faster’.

Straining may be achieved by overlaying the Si on a SeGe substrate (which has a different spacing). More recently, doping the source and drain of PMOS transistors has been used to compress the channel between then (refer to a cross-section of a MOSFET). To stretch the channel on NMOS FETs a silicon nitride ($\text{N}_4\text{Si}_3$) gate overlay is used.

Interconnect
The ‘traditional’ on-chip wiring has been aluminium which is a good electrical conductor and is easy to etch. More recently this has been (largely?) superseded by copper which is a better conductor; the conductance of wires is increasingly important as the wires become ever thinner. Copper is not as easy to work as aluminium; instead of deposit/plasma etch the chip’s surface (at that stage) has a SiO$_2$ insulating layer which is patterned with etched trenches. Copper is deposited over all and then mechanically polished away, leaving only the deposits in the protected trenches as the wires. The process is repeated for higher layers.

Exotic structures
FinFETs and similar were mentioned in an earlier session. It is likely that physical structures which improve transistors’ switching characteristics will continue to be developed. The objective is to conduct better when ‘on’ yet avoid leakage when ‘off’. See, for example, Intel’s ‘Tri-gate’ transistors.

It is also likely that processes will allow more variety of different doping levels (‘speeds’) within the same chip.

Multigate devices
(See lecture on ‘Power’.)

Foundries are developing transistor structures on top of – and in some cases embedded in – the substrate.

FinFETs are one such structure although the term is used informally for other designs.

Another example is Intel’s tri-gate transistor which features multiple parallel channels with ‘wrap-around’ gates.

The objectives are higher switching speed and reduced leakage current (hence power wastage).
Through-Silicon Vias (TSVs)

- Wafer back-ground to thickness of (maybe) 50 µm
- ‘Deep Reactive-Ion Etching’ used to cut near-vertical holes: maybe ~5-10 µm across
- Holes insulated (SiO₂), then metal filled
- Resulting wires are ‘big’ but not as big as bond wires
- Still require some substantial driver amplifiers

TSVs with spacing as small as 10 µm have been demonstrated; initially practicality may increase this distance.

For contrast, pad spacing for bond wires is limited to around 200 µm.

Allows interfacing such as ‘Wide I/O’ to provide greater off-chip bandwidth:
- faster due to lower load
- more wires so greater parallelism

This may be particularly useful for memory connections.

Chip stacking

In principle a stack of chips can be interconnected with TSVs.

Pros
- Lots of functionality in a small volume
- High-bandwidth interconnection (many wires)
- Chips can be manufactured on different processes
  - Logic
  - DRAM
  - EEPROM
  - Analogue
  - …

Cons
- New and expensive
- Problems with cooling/heat dissipation
- Power supplies
Future SoCs

- Future SoCs are likely to contain many processors
  - Some explicitly will be ‘multiprocessor’ chips
  - Processors too useful not to be used for flexibility (testing, redundancy …)

- Likely to comprise multiple regions using different clocks, power supplies etc.
  - Regions may be asynchronous; communication may be asynchronous rather like computer networks now
  - Some (most?) regions may be powered off much of the time

- Reconfigurable hardware likely
  - Big FPGAs already contain ‘hard’ processors (e.g. see Xilinx ‘Zynq’)
  - *Dynamically* reconfigurable accelerators will allow (parts of) ‘software’ algorithms to be executed in hardware (faster and lower power)
  - Distinction of ‘software’ and ‘hardware’ may be more blurred

Moore’s Law hasn’t run out yet!

‘That’s all folks!’