A sound barrier for silicon?

DAVID A. MULLER is at the School of Applied and Engineering Physics, Cornell University, Ithaca New York 14853, USA.

e-mail: dm24@cornell.edu

For the first time in thirty five years, the clockspeed of the fastest commercial computer chips has not increased. Is the semiconductor industry just pausing for breath or about to suffer a fate similar to that of aerospace?

This year marks the fortieth anniversary of Gordon Moore's still accurate observation that the number of transistors in an integrated circuit doubles roughly every eighteen months[1]. Moore, and the company he later co-founded (Intel), expect this trend to continue for at least another ten years[2,3]. We have taken almost for granted the apparent corollary: as the transistor count increases each transistor also gets smaller, faster and cheaper. The corollary follows from a consequence of a transistor scaling mechanism first proposed in 1974[4]. If it were not for this scaled shrinkage, computer chips built today using the same technology as the first microprocessor would be about a metre across and run at least 40,000 times slower than today's machines. However, the 'faster' part of the 'smaller, faster, cheaper' triad seems to have hit a speedbump over the past two years (Fig. 1). In response, microprocessor manufacturers have undergone a major shift from increasing the switching speed of a single processor to instead increasing the number of processors that can work in parallel.

To better understand the causes and consequences of the decoupling of Moore's law from one of its three apparent drivers (speed) requires an appreciation for the underlying scaling and materials physics. At its most basic, a field-effect transistor (Fig. 2a) is a three-terminal device where the flow of electrons (or holes) from source to drain through a semiconducting channel is controlled by changing the number of charge carriers in the channel[5]. Carriers can be swept into, or out of, the channel by the electric field produced at the channel/gate dielectric interface by applying a voltage to the gate electrode. The gate dielectric is nothing more than a capacitor, and the strength of the field is determined by the applied voltage, thickness (which today is a few atoms[6]), and the dielectric constant of the (hopefully insulating) gate dielectric layer. The output drain current can be used to drive another transistor, and the switching speed will depend on the time it takes to charge the gate 'capacitor' of the next device. This drive-current (and hence speed) increases with increasing drain voltage, decreasing channel length and decreasing gate dielectric thickness. Furthermore, the next device can be charged more rapidly as the device area is reduced. Combined, these last three effects all suggest a potentially large increase in speed as the device size is reduced. Even ignoring the speed increase, there was a strong economic incentive to shrink transistor sizes: the cost to process a wafer is not very sensitive to the number of transistors on the wafer (as the transistors are patterned lithographically in parallel) so it is almost as easy to print one transistor at a time as one million.

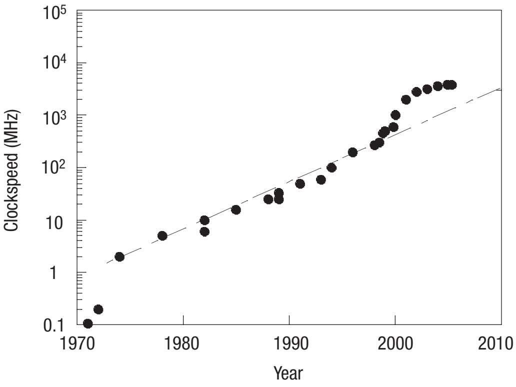

Figure 1 **A 'sound barrier' for silicon? The top clockspeed of Intel microprocessors as a function of time (data summarized from Intel's website). Is this curve levelling off, or just waiting to catch up with the long-term trend?**

In shrinking down the transistor size, all dimensions of the device must be scaled for it to work. The simplest approach to scaling is just to shrink all the linear physical dimensions of each transistor by the same factor while leaving the supply voltage unaltered. This purely geometric approach is known as constant voltage scaling[5]. The speed and packing density of transistors is increased quadratically, whereas the current per transistor remains constant. Because the thickness of the dielectric layer is decreased, the electric field in the channel is increased by the same factor. At first this was a good thing, as the carrier mobility increased with electric field. However, when electric fields exceed 1 $\mathrm{MV\,cm^{-1}}$, the mobility starts to decrease again, offsetting any gains in drive current.

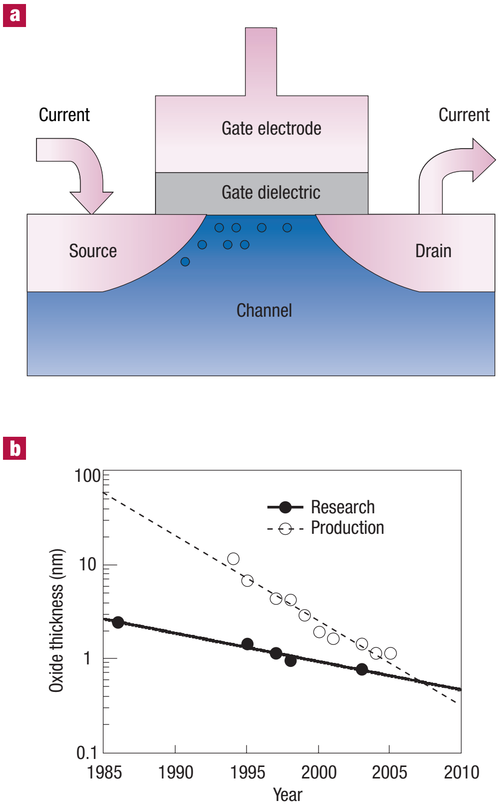

Figure 2 **The importance of the gate-oxide thickness. a, Simplified cartoon of a metal oxide semiconductor field-effect transistor of the type used in modern computer chips. Current flow from source to drain is controlled by a voltage applied across an insulating gate oxide causing carriers to accumulate below the gate oxide, creating a conducting channel. b, Effective gate-oxide thickness as a function of time, both for commercial production, and the best reported research results[6,8,11,15–18]. These results all turn out to be from SiO_2 or nitrided SiO_2 oxides, which outperform their high-*k* would-be replacements. Only those results that were able to demonstrate an improvement in drive current/unit length at realistic operating voltages and leakage currents were considered for the research curve. Results more recent than 2004 are not comprehensive.**

Additionally, reliability and leakage currents worsen exponentially.

An elegant and obvious solution is constant-field scaling[4], where both the supply voltage and geometry are both reduced by the same amount. Now the electric field in the channel remains constant, avoiding the reliability and mobility problems of constant-voltage scaling. In constant-field scaling the packing density of transistors is increased quadratically as before, but the power per transistor now decreases quadratically, so the overall power density remains constant instead of geometrically exploding. The current density also increases linearly rather than quadratically, so ohmic heating losses and electromigration damage in the connecting wires do not grow as rapidly. However, the speed only improves linearly.

Constant-voltage and constant-field scaling represent two extremes, and in practice speed and power increases can be traded off against each other by choosing intermediate scaling rules[7]. From 1973 to 1993, there were long periods when the computer industry essentially followed constant-voltage scaling, reducing the supply voltage only twice. From 1993 to 2003, the supply voltage decreased with each new generation of transistors, although not as rapidly as the ideal constant-field scaling[8,9]. However, it may well be that the rapid increase in clockspeed in Fig. 1 between 2000 and 2003 reflects that the gate thickness was decreased more aggressively than the supply voltage, giving a speed boost at the expense of power consumption. Even so, the gate thickness (for reasons discussed below) was still not being scaled as rapidly as the channel length, adding to the power-density woes[8].

Even small deviations from constant-field scaling add up over time. For instance, Intel saw the power consumption of their chips double roughly every 36 months from 1971 to 2000 (ref. 7), and reaching almost 100 watts by 2004 before they abandoned the goal of further increasing the clockspeed through simple scaling. They cancelled further Pentium-4 designs in favour of a dual-core architecture with more but slower transistors. Since then the operating voltage has crept back up, but the power consumption has dropped five to tenfold.

Constant-field scaling cannot continue indefinitely because the switching voltage soon shrinks to the scale of thermal fluctuations. This is a problem for any device where an electric field is used to control the flow of charge through a channel. Whether the device is made from silicon or a single organic molecule, it is still governed by the same laws of statistical mechanics: if the difference between the on and off voltages is decreased, the difference between leakage current in the off-state and the on-state shrinks exponentially with a slope determined only by temperature. Today, the power losses in the off state are almost comparable to the on-state losses — the transistor acts more like a dimmer than a switch. If we require a factor of ten difference in current between the on and off states of a device, a 60 mV voltage change at room temperature would be sufficient for high-current devices. If the switch is designed to work with small numbers of electrons, a larger voltage change is needed to discriminate against thermal fluctuations. As present supply voltages are around 1 V, it is still theoretically possible to reach lower supply voltages (perhaps for about 15 years at present scaling rates). Very different designs with lower leakage currents will be required to take advantage of this, however, because present structures are already operating very close to their noise margins.

Scaling of the gate dielectric poses an immediate materials challenge. At present, the gate dielectric is made from silicon dioxide (or its nitride derivatives) and the capacitance is increased by decreasing the oxide thickness. This cannot continue indefinitely. Even before one runs outs of atoms in the dielectric layer, the leakage current due to direct tunnelling increases exponentially. A practical limit seems to be around 0.7–1.2 nm for pure SiO_2 — four to five silicon atoms thick[6]. The solution is to find a new material with a higher dielectric constant than SiO_2, but one that also maintains a tunnelling barrier for both the valence and conduction bands. In theory many materials would satisfy these requirements, but in practice one should not expect layers a few atoms thick to behave like a bulk material. Instead, interfacial reaction layers that would not be expected from bulk thermodynamics reduce the capacitance[10]. Point defects also shift the turn-on voltage and reduce the carrier mobility[10]. Figure 2b shows that improvements in research have not kept up with the demands of production. To date, none of the high-dielectric-constant (high-*k*) materials have been shown to improve the drive current of an actual device, and until such a material is found the prospects for long-term speed improvements are limited.

There are, however, a number of one-time 'tricks' that can improve performance[11]. The polysilicon gate electrode could be replaced by metal, removing some series capacitance. The mobility in the channel can be enhanced by using SiGe alloys, straining the channel, or even using different wafer orientations. Silicon-on-insulator structures can control some of the short-channel problems. Combined, these strategies could increase the device speed by another factor of two to four — giving another 18–36 months of breathing room. However, one reason that these methods were not used before is that they were always more expensive than the simpler scaling strategies.

To be clear, the current problems lie in increasing the clockspeed, rather than shrinking the transistor, which can be accomplished simply by trading off speed for size. Currently, the smallest silicon devices that show some switching (or rather dimming) behaviour have 8-nm gate lengths. Below 30 nm, transport is largely ballistic so small silicon devices perform just as well, if not better than carbon nanotube transistors[12].

Long term, it is not so much that the fundamental physical limits on the power dissipation and tunnelling through silicon dioxide present absolute barriers to improvements in clockspeed (any more than the sound barrier prevents supersonic flight), but rather that the price one pays to get around those physical limits may prove to be too high. Figure 3 shows that for almost sixty years, aircraft cruising speeds also had their own form of a Moore's law. Even though most military aircraft are capable of supersonic flight, it has not proven to be commercially viable in the long term and all civilian aircraft today stay below the sound barrier. As the technology matured, the demand for aerospace engineers plummeted. Despite the massive layoffs in the 1970s in the US, the sector shrank by another 50% from 1990 to 2005, and even today the US bureau of labour statistics outlook is for a declining job market in aerospace[5].

Recent press releases have promised processor performance improvements through new system architectures[13,14]. Can clever systems design and parallelism compensate for a slowdown in increased clockspeed? Consider the performance history from the first microprocessor, the Intel 4004 at 0.06 MIPS (millions of instructions per second) to a typical mid-2005 microprocessor at 11,000 MIPS. The clockspeed grew from 100 kHz to just under 4 GHz, and the data buswidth grew from 4 bits to 64. Overall, the number of bits per second grew by about a factor of 3 million. Of this increase, a factor of roughly 40,000 can be attributed to scaling enabled by materials and lithographic improvements, a factor of 16 due to increased buswidth (which also increased the die size), and only a factor of 5 is left for smarter design. Based on past performance, future improvements through software and systems optimizations are likely to be more modest arithmetic corrections, rather than the exponential growth we have become accustomed to. The semiconductor companies have anticipated this, and more recent marketing campaigns have de-emphasized clockspeed and stressed power management and less-tangible features like compatibility and reliability, much as one might choose a car based on fuel economy and maintenance record rather than top speed.

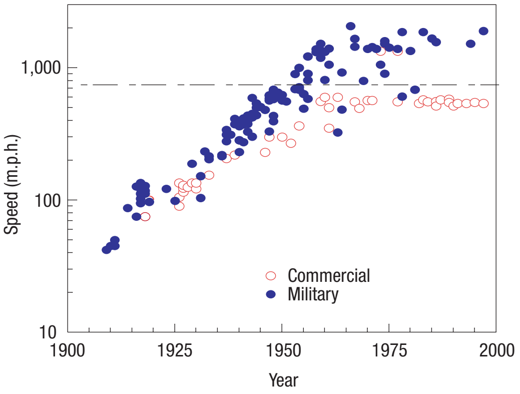

Figure 3 **Cruising speed (in miles per hour) of commercial[19] and military aircraft[20]. The speed of sound is indicated with the dashed line. Although it is possible to go faster than the speed of sound, the commercial curve indicates it has not proven economically sustainable to do so.**

Are consumers willing to pay for more powerful chips? If so, more transistors per chip are needed, they must all work, and manufacturing yield determines winners and losers. Yield in turn requires more sophisticated fabrication facilities and the semiconductor industry is likely to experience consolidation down to a couple of large survivors, similar to that of the aerospace industry. If, on the other hand, consumers are satisfied with computers about two to four times as fast as today's machines, yield is less critical and there is less to differentiate between companies. In this direction lies commoditization and a fate that may resemble the automobile industry, with many more national companies competing on price and reputation.

REFERENCES

1. Moore, G. Cramming more components onto integrated circuits. *Electronics* **38,** 144–117 (1965).
2. Moore, G. in *IEEE International Solid-State Circuits Conference* Vol. 1, 20–23 (IEEE, San Francisco, 2003).
3. http://www.intel.com/technology/silicon/mooreslaw/
4. Dennard, R. *et al.* design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid-State Circuits* **9,** 256–268 (1974).
5. Sze, S. M. *Semiconductor Devices: Physics and Technology* (Wiley, New York, 1985).
6. Muller, D. A. *et al.* The electronic structure at the atomic scale of ultra-thin gate oxides. *Nature* **399,** 758–761 (1999).
7. Mahajan, R. *et al.* Emerging directions for packaging technologies. *Intel Technol. J.* **6,** 62–75 (2002).
8. Nowak, E. J. Maintaining the benefits of CMOS scaling when scaling bogs down. *IBM J. Res. Dev.* **46,** 169–180 (2002).
9. Doyle, B. *et al.* Transistor elements for 30nm physical gate lengths and beyond. *Intel Technol. J.* **6,** 42–54 (2002).
10. Wallace, R. M. & Wilk, G. D. Alternative gate dielectrics for microelectronics. *Mater. Res. Soc. Bull.* **27,** 186–191 (2002).
11. Semiconductor Industry Association *International Technology Roadmap for Semiconductors, Update* (2000); see http://public.itrs.net/Files/2000UpdateFinal/2kUdFinal.htm.
12. Chau, R. *et al.* Benchmarking nanotechnology for high-performance and low-power logic transistor applications. *IEEE Trans. Nanotechnol.* **4,** 153–158 (2005).
13. Hiremane, R. From Moore's law to Intel innovation—prediction to reality. *Intel Mag.* 1–9 (April 2005).
14. Markoff, J. in *New York Times* C3 (New York, 7 February 2005).
15. Horiguchi, S., Kobayashi, T., Miyake, M., Oda, M. & Kiuchi, K. Extremely high transconductance (above 500 mS/mm) MOSFET with 2.5 nm gate oxide. *IEDM Tech. Dig.* 761–773 (1985).
16. Momose, H. S. *et al.* 1.5 nm direct-tunneling gate oxide Si MOSFETs. *IEEE Trans. Elec. Dev.* **43,** 1233–1242 (1996).
17. Chau, R. *et al.* High-k/metal gate stack and its MOSFETs characteristics. *IEEE Elect. Dev. Lett.* **25,** 408–410 (2004).
18. Thompson, S. *et al.* 130nm logic technology featuring 60nm transistors, low-k dielectrics, and Cu interconnects. *Intel Technol. J.* **6,** 5–13 (2002).
19. *The History of United Airlines* http://www.united.com/page/middlepage/0,6823,2286,00.html.
20. National Museum of the United States Airforce; www.wpafb.af.mil.museum.htm.