Feeling the heat

The more that microcircuits are shrunk, the hotter they get. Engineers are on the hunt for ways to cool off computing.

BY PHILIP BALL

D. SAELINGER/CORBIS

© 2012 Macmillan Publishers Limited. All rights reserved

A laptop computer can double as an effective portable knee-warmer — pleasant in a cold office. But a bigger desktop machine needs a fan. A data centre as large as those used by Google needs a high-volume flow of cooling water. And with cutting-edge supercomputers, the trick is to keep them from melting. A world-class machine at the Leibniz Supercomputing Centre in Munich, for example, operates at 3 petaflops (3×10^{15} operations per second), and the heat it produces warms some of the centre's buildings. Current trends suggest that the next milestone in computing — an exaflop machine performing at 10^{18} flops — would consume hundreds of megawatts of power (equivalent to the output of a small nuclear plant) and turn virtually all of that energy into heat.

Increasingly, heat looms as the single largest obstacle to computing's continued advancement[1]. The problem is fundamental: the smaller and more densely packed circuits become, the hotter they get. "The heat flux generated by today's microprocessors is loosely comparable to that on the Sun's surface," says Suresh Garimella, a specialist in computer-energy management at Purdue University in West Lafayette, Indiana. "But unlike the Sun, the devices must be cooled to temperatures lower than 100 °C" to function properly, he says.

To achieve that ever more difficult goal, engineers are exploring new ways of cooling — by pumping liquid coolants directly on to chips, for example, rather than circulating air around them. In a more radical vein, researchers are also seeking to reduce heat flux by exploring ways to package the circuitry. Instead of being confined to two-dimensional (2D) slabs, for example, circuits might be arrayed in 3D grids and networks inspired by the architecture of the brain, which manages to carry out massive computations without any special cooling gear. Perhaps future supercomputers will not even be powered by electrical currents borne along metal wires, but driven electrochemically by ions in the coolant flow.

This is not the most glamorous work in computing — certainly not compared to much-publicized efforts to make electronic devices ever smaller and faster. But those high-profile innovations will count for little unless engineers crack the problem of heat.

GO WITH THE FLOW

The problem is as old as computers. The first modern electronic computer — a 30-tonne machine called ENIAC that was built at the University of Pennsylvania in Philadelphia at the end of the Second World War — used 18,000 vacuum tubes, which had to be cooled by an array of fans. The transition to solid-state silicon devices in the 1960s offered some respite, but the need for cooling returned as device densities climbed. In the early 1990s, a shift from earlier 'bipolar' transistor technology to complementary metal oxide semiconductor (CMOS) devices offered another respite by greatly reducing the power dissipation per device. But chip-level computing power doubles roughly every 18 months, as famously described by Moore's Law, and this exponential growth has brought the problem to the fore yet again[2] (see 'Rising temperatures'). Some of today's microprocessors pump out heat from more than one billion transistors. If a typical desktop machine let its chips simply radiate their heat into a vacuum, its interior would reach several thousand degrees Celsius.

That is why desktop computers (and some laptops) have fans. Air that has been warmed by the chips carries some heat away by convection, but not enough: the fan circulates enough air to keep temperatures at a workable 75 °C or so.

But a fan also consumes power — for a laptop, that is an extra drain on the battery. And fans alone are not always sufficient to cool the computer arrays used in data centres, many of which rely on heat exchangers that use liquid to cool the air flowing over the hot chips.

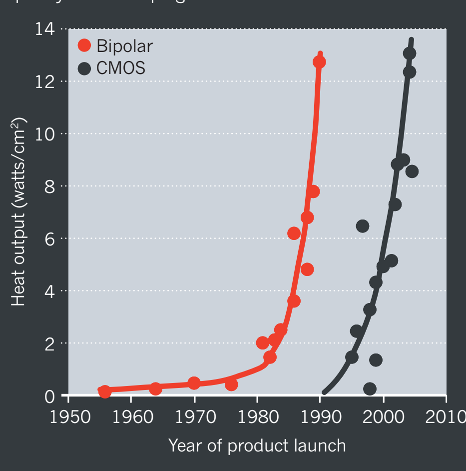
Still larger machines demand more drastic measures. As Bruno Michel, manager of the advanced thermal packaging group at IBM in Rüschlikon, Switzerland, explains: "An advanced supercomputer would need a few cubic kilometres of air for cooling per day." That simply is not practical, so computer engineers must resort to liquid cooling instead[3].

Water-cooled computers were commercially available as early as 1964, and several generations of mainframe computers built in the 1980s and 1990s were cooled by water. Today, non-aqueous, non-reactive liquid coolants such as fluorocarbons are sometimes used, often coming into direct contact with the chips. These substances generally cool by boiling — they absorb heat and the vapour carries it away. Other systems involve liquid sprays or refrigeration of the circuitry.

SuperMUC, an IBM-built supercomputer housed at the Leibniz centre, became operational in 2012. The 3-petaflop machine is one of the world's most powerful supercomputers. It has a water-based cooling system, but the water is warm — around 45 °C. The water is pumped through microchannels carved into a customized copper heat sink above the central processing unit, which concentrates cooling in the parts of the system where it is most needed. The use of warm water may seem odd, but it consumes less energy than other cooling methods, because the hot water that emerges from the system requires less chilling before it is reintroduced. And the use of hot-water outflow for heating nearby office buildings results in further energy savings.

Michel and his colleagues at IBM believe that flowing water could be used not just to extract heat, but also to provide power for the circuitry in the first place, by carrying dissolved ions that engage in electrochemical reactions at energy-harvesting electrodes. In effect, the coolant doubles as an electrolyte 'fuel'. The idea is not entirely new, says Yogendra Joshi, a mechanical engineer at the Georgia Institute of Technology in Atlanta. "It has been used for many years in thermal management of aircraft electronics", which are cooled by jet fuel, he says.

Delivering electrical power with an electrolyte flow is already a burgeoning technology. In a type of fuel cell known as a redox flow battery, for example, two electrolyte solutions are pumped into an electrochemical cell, where they are kept separate by a membrane that ions can flow through. Electrons travel between ions in the solutions in a process known as a reduction–oxidation (redox) reaction — but they are forced to do so through an external circuit, generating energy that can be tapped to provide electrical power.



SOURCE: REF. 3

SALTY LOGIC

Redox-flow cells can be miniaturized using microfluidic technology, in which the fluid flows are confined to microscopic channels etched into a substrate such as silicon[4]. At such small scales, the liquids can flow past each other without mixing, so there is no need for a membrane to separate them. With this simplification, the devices are easier and cheaper to make, and they are compatible with silicon-chip technology.

Michel and his colleagues have begun to develop microfluidic cells for powering microprocessors, using a redox process based on vanadium ions. The electrolyte is pumped along microchannels that are 100–200 micrometres wide and similar to those used to carry coolant flows around some chips. Power is harvested at electrodes spaced along the channel, then distributed to individual devices by conventional

© 2012 Macmillan Publishers Limited. All rights reserved

Steam carries heat away from Google's data centre in The Dalles, Oregon.

C. ZHOU/GOOGLE/ZUMAPRESS/EYEVINE

metal wiring. The researchers unveiled their preliminary results in August, at a meeting of the International Society of Electrochemistry in Prague[5].

But they remain some way from actually powering circuits this way. At present, the power density of microfluidic redox-flow cells is less than 1 watt per square centimetre at 1 volt — two or three orders of magnitude too low to drive today's microprocessors. However, Michel believes that future processors will have significantly lower power requirements. And, he says, delivering power with microfluidic electrochemical cells should at least halve the power losses that occur with conventional metal wiring, which squanders around 50% of the electrical energy it carries as resistive heating.

BECOMING BRAINIER

Electrochemical powering could help to reduce processors' heat dissipation, but there is a way to make a much bigger difference. Most of the heat from a chip is generated not by the switching of transistors, but by resistance in the wires that carry signals between them. The problem is not the logic, then, but the legwork. During the late 1990s, when transistors were about 250 nanometres across, 'logic' and 'legwork' accounted for roughly equal amounts of dissipation. But today, says Michel, "wire energy losses are now more than ten times larger than the transistor-switching energy losses". In fact, he says, "because all components have to stay active while waiting for information to arrive, transport-induced power loss accounts for as much as 99% of the total".

This is why "the industry is moving away from traditional chip architectures, where communication losses drastically hinder performance and efficiency", says Garimella. The solution seems obvious: reduce the distance over which information-carrying pulses of electricity must travel between logic operations. Transistors are already packed onto 2D chips about as densely as they can be. If they were stacked in 3D arrays instead, the energy lost in data transport could be cut drastically. The transport would also be faster. "If you reduce the linear dimension by a factor of ten, you save that much in wire-related energy, and your information arrives almost ten times faster," says Michel. He foresees 3D supercomputers as small as sugar lumps.

What might 3D packaging look like? "We have to look for examples with better communication architecture," Michel says. "The human brain is such an example." The brain's task is demanding: on average, neural tissue consumes roughly ten times more power per unit volume than other human tissues — an energy appetite unmatched even in an Olympic runner's quadriceps. The brain accounts for just 2% of the body's volume, but 20% of its total energy demand.

Yet the brain is fantastically efficient compared to electronic computers. It can achieve five or six orders of magnitude more computation for each joule of energy consumed. Michel is convinced that the brain's efficiency is partly due to its architecture: it is a 3D, hierarchical network of interconnections, not a grid-like arrangement of circuits.

SMART BUILD

This helps the brain to make much more efficient use of space. In a computer, as much as 96% of the machine's volume is used to transport heat, 1% is used for communication (transporting information) and just one-millionth of one per cent is used for transistors and other logic devices. By contrast, the brain uses only 10% of its volume for energy supply and thermal transport, 70% for communication and 20% for computation. Moreover, the brain's memory and computational modules are positioned close together, so that data stored long ago can be recalled in an instant. In computers, by contrast, the two elements are usually separate. "Computers will continue to be poor at fast recall unless architectures become more memory-centric", says Michel. Three-dimensional packaging would bring the respective elements into much closer proximity.

All of this suggests to Michel that, if computers are going to be packaged three-dimensionally, it would be worthwhile to try to emulate the brain's hierarchical architecture[6]. Such a hierarchy is already implicit in some proposed 3D designs: stacks of individual microprocessor chips (on which the transistors themselves could be wired in a branching network) are stacked into towers and interconnected on circuit boards, and these, in turn, are stacked together, enabling vertical communication between them. The result is a kind of 'orderly fractal' structure, a regular subdivision of space that looks the same at every scale.

Michel estimates that 3D packaging could, in principle, reduce computer volume by a factor of 1,000, and power consumption by a factor of 100, compared to current 2D architectures. But the introduction of brain-like, 'bionic' packaging structures, he says, could cut power needs by another factor of 30 or so, and volumes by another factor of 1,000. The heat output would also drop: 1-petaflop computers, which are now large enough to occupy a small warehouse, could be shrunk to a volume of 10 litres.

If computer engineers aspire to the awesome heights of zetaflop computing (10^{21} flops), a brain-like structure will be necessary: with today's architectures, such a device would be larger than Mount Everest and consume more power than the current total global demand. Only with a method such as bionic packaging does zetaflop computing seem remotely feasible. Michel and his colleagues believe that such innovations should enable computers to reach the efficiency — if not necessarily the capability — of the human brain by around 2060. That is something to think about. ■

Philip Ball *is a writer based in London.*

1. Garimella, S. V. *et al. IEEE Trans. Components Packaging Technol.* **31,** 801–815 (2008).
2. Chu, R. C., Simons, R. E., Ellsworth, M. J., Schmidt, R. R. & Cozzolino, V. *IEEE Trans. Device Mater. Reliability* **4,** 568–585 (2004).
3. Ellsworth, M. J. *et al. ITHERM* 266–274 (2008).
4. Shaegh, S. A. M., Nguyen, N.-T. & Chan, S. H. *Int. J. Hydrogen Energ.* **36,** 5675–5694 (2011).
5. Ruch, P. W., Rapp, T., Schmidt, T. J. & Michel, B. Studies of power density in microfluidic redox flow cells. Abstract presented at the 63rd Annual Meeting of the International Society of Electrochemistry (2012).
6. Ruch, P., Brunschwiler, T., Escher, W., Paredes, S. & Michel, B. *IBM J. Res. & Dev.* **55,** 593–605 (2011).

© 2012 Macmillan Publishers Limited. All rights reserved