

ENGINEERING

Cooling Energy-Hungry Data Centers

G. I. Meijer

Soon after the Internet took off in the mid-1990s, enterprise computing infrastructures with warehouses full of servers, known as data centers, became commonplace. The energy consumption challenges posed by such data centers are considerable. The power dissipation of servers has to be managed skillfully. Perhaps surprisingly, the power consumption of the cooling infrastructure that is required to keep the microelectronic components from overheating is on a par with that of the servers themselves.

In 2009, an estimated 330 terawatt-hours of energy—about 2% of the global electricity production—was consumed to operate data centers worldwide (1). Apart from the sheer economic impact (U.S.\$ 30 billion), there are also considerable ecological consequences. The International Energy Agency estimated

sistors, leakage currents consume more power than the actual computational processes. To alleviate this burden, new materials were introduced in the late 2000s. Most notably, the replacement of the SiO₂ gate oxide, which is only a few atomic layers thick, with a physically thicker layer of a hafnium-based oxide enabled an appreciable reduction of the gate tunneling currents while maintaining the electrical performance of the transistor (5, 6). Nevertheless, keeping the gate leakage power per unit area below 1 kW cm⁻² will remain a particularly difficult issue, especially for scaled-down devices beyond the 2013 horizon.

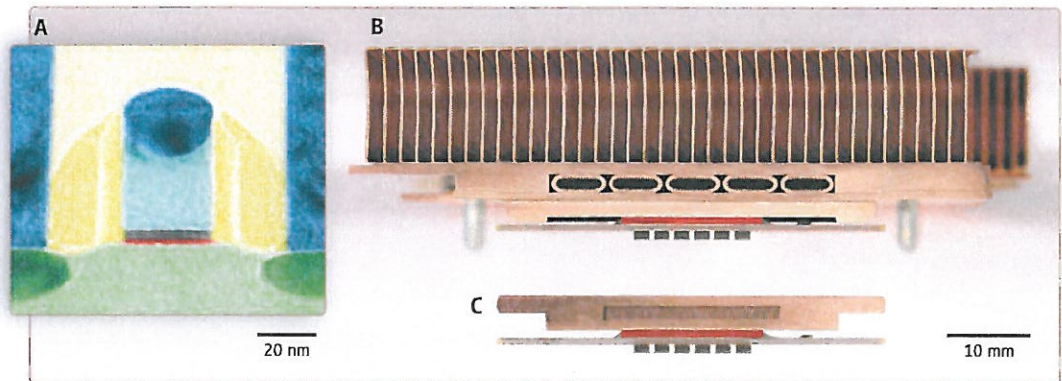
Consequently, amid the new microprocessor generations, the prospect of restraining the power dissipation at peak performance still looks grim. For high-performance microprocessor circuitry, the power dissipation is pro-

The information technology industry is focusing on approaches to hot-water cooling for the design of energy-efficient data centers.

the electricity consumption currently goes toward powering this cooling infrastructure. It therefore appears ironic that Moore's law (a projection of microprocessor performance) is widely known and often cited, while increasingly critical physical laws of thermodynamics receive little popular attention.

The first law of thermodynamics states that energy is conserved. The electrical energy that is supplied to the computer system is eventually entirely converted into thermal energy. The standard method to remove this heat is by forced circulation of large amounts of chilled air. Massive heat sinks with long and fine-pitched fins protruding from a heat-spreader base are used to enhance the convective heat transfer from the hot modules to the cold air (see the figure, panel B). The collected heat is then expelled to the outside.

Heat generation and cooling in the data center. (A) Transmission electron microscopy image of a transistor. Heat is primarily generated near the gate oxide (marked red). (B) Cross-section of microprocessor module with air-cooled heat sink. Area in red indicates location of microprocessor. (C) Cross section of microprocessor module with liquid-cooled microchannel heat sink.



that the greenhouse-gas emissions associated with the electricity production was around 200 million metric tons of CO₂ and accounted for 0.7% of the global energy-related CO₂ emissions. This prompted increased scrutiny from regulators (2, 3).

The information-technology industry therefore needs efficacious concepts to reduce the energy consumption of data centers. The key culprits are the server microprocessors, or more precisely, the transistors inside these microprocessors (see the figure, panel A). Currently, transistors with 45-nm lateral features are in volume production, and the pace of miniaturization continues unabated (4). It is a formidable challenge to keep the power dissipation of these transistors within acceptable limits. With shrinking dimensions of the tran-

sistors, leakage currents consume more power than the actual computational processes. To alleviate this burden, new materials were introduced in the late 2000s. Most notably, the replacement of the SiO₂ gate oxide, which is only a few atomic layers thick, with a physically thicker layer of a hafnium-based oxide enabled an appreciable reduction of the gate tunneling currents while maintaining the electrical performance of the transistor (5, 6). Nevertheless, keeping the gate leakage power per unit area below 1 kW cm⁻² will remain a particularly difficult issue, especially for scaled-down devices beyond the 2013 horizon.

Consequently, amid the new microprocessor generations, the prospect of restraining the power dissipation at peak performance still looks grim. For high-performance microprocessor circuitry, the power dissipation is projected to stay around 50 W cm⁻², a value that is several times greater than that of a stove's hot plate. Hence, if the data-center power-dissipation issue cannot be mitigated at its source, what can be done to reduce the total energy consumption and carbon footprint? For the newest members of microprocessor families, sophisticated circuit architectures have been introduced, which allow the power associated with computational processes and also the leakage power to be adapted (7, 8). The microprocessor frequency can be adjusted and circuit blocks can be temporarily powered down completely when not in use. These innovations lead to energy savings for a computational load that comes in bursts or that is bound to memory latency or input/output operations.

An alternative approach is to tackle the problem at the cooling infrastructure (9). Approximately 50% (industry average) of

A straightforward approach to reducing the energy consumption of the cooling infrastructure is through a careful segregation of the chilled- and the hot-air flows.

The real key to ratcheting down the energy consumption of a computing facility is liquid cooling. The reason is that thermodynamically liquid cooling is much more efficient than air cooling because the heat capacity of liquids is orders of magnitude larger than that of air (for example, for water it is 4 MJ m⁻³ K⁻¹ versus 1 kJ m⁻³ K⁻¹ for air). Once the heat has been transferred to the liquid, it can be handled very efficiently. Critics of liquid cooling might contend that it comes at the price of increased mechanical complexity. True, but this can be managed as computers were once equipped with liquid cooling when the power density of bipolar-transistor-based computer systems reached its peak during the 1980s. For example, the

Downloaded from www.sciencemag.org on April 21, 2010

Cray-2 supercomputer used liquid immersion cooling (10).

Recently, chilled-liquid cooling was reintroduced in high-end mainframes and densely packed servers to cope with the high heat fluxes. Yet, liquid cooling can be taken further if we consider a microfluidic heat sink (11) (see the figure, panel C). Microchannel heat sinks can be designed such that the thermal resistance between the transistor and the fluid is reduced to the extent that even cooling-water temperatures of 60° to 70°C ensure no overheating of the microprocessors. This hot-water cooling has compelling advantages. First, chillers are no longer required year-round and thus the data-center energy consumption plummets by up to 50%. Second, and perhaps most important, direct utiliza-

tion of the collected thermal energy becomes feasible, either using synergies with district heating or specific industrial applications. With such an appealing waste-heat recovery system, the green diligence of data centers would be upheld substantially.

Reducing the energy consumption of data centers and concomitantly restraining costs, while curtailing carbon emission, is achievable. Despite power dissipation in microprocessors continuing to be a source of severe concern, liquid cooling and deploying waste heat appear to become imperative in the drive for improving the data-center energy efficiency.

References

1. International Data Corporation, Document No. 221346 (2009), www.idc.com.

2. U.S. Environmental Protection Agency, Report to Congress on Server and Data Center Energy Efficiency (2007).
3. European Commission, Code of Conduct on Data Centres Energy Efficiency (2008).
4. International Technology Roadmap for Semiconductors, 2009 Edition, Executive Summary; www.itrs.net/Links/2009ITRS/Home2009.htm.
5. K. Mistry *et al.*, *IEEE IEDM 2007 Tech. Digest*, 10.2 (2007).
6. M. Chudzik *et al.*, *IEEE VLSI 2007 Tech. Digest*, 11A-1 (2007).
7. N. A. Kurd *et al.*, *IEEE ISSCC 2010 Tech. Digest*, 5.1 (2010).
8. M. Ware *et al.*, *IEEE HPCA 2010 Tech. Digest*, 6.4 (2010).
9. L. A. Barroso, U. Hölzle, *The Datacenter as a Computer* (Morgan and Claypool, 2009); www.google.com/corporate/green/datacenters.
10. S. R. Cray Jr., U.S. Patent 4590538 (1986).
11. D. B. Tuckerman, R. F. W. Pease, *Electron Device Lett.* 2, 126 (1981).

10.1126/science.1182769

MATERIALS SCIENCE

The Future of Metals

K. Lu

On 15 December 2009, the world's most fuel-efficient commercial jetliner—the Boeing 787 Dreamliner—completed its first flight. The airliner is mostly made from carbon fiber–reinforced polymeric composites (50% by weight, up from 12% in the Boeing 777) (1). Traditional metals are substantially replaced by composites with higher strength/weight ratios; aluminum usage has dropped to 20% (versus 50% in the 777). Ever since the 1950s, when “engineering materials” mainly meant metals (2), the share of metals in engineering materials has been diminishing. What are the reasons behind this trend, and which applications are likely to stay in the domain of metals?

The main property limitation of metals as structural materials is their low specific strength (the strength/weight ratio). Most engineering designs call for structural materials that have high strength, fracture toughness (a measure of the energy required for propagating cracks), and stiffness while minimizing weight. Most metals have high strength and stiffness, but because they are dense (steels are several times as dense as ceramics and polymers), their strength/weight and stiffness/weight ratios are low relative to competing materials (see the figure). This is a key reason for replacing metals in aircraft and sporting goods, where weight is a primary

concern. Some metals such as aluminum and magnesium are light, but they are too soft for many applications and have low toughness and stiffness. Titanium alloys partly overcome these problems: They are about half as dense as steels, have higher strength, and are very tough. Titanium was first used in airliners in the 1960s in the Boeing 707 and its use has increased to 15% in the Boeing 787 (1).

Metals can be strengthened through controlled creation of internal defects and boundaries that obstruct dislocation motion (3). But such strategies compromise ductility and toughness, in contrast to the increasing toughness at higher strength seen with polymeric composites (see the figure). Strengthening may also compromise other metal properties, such as conductivity and corrosion resistance. One method for strengthening metals without losing toughness is grain refinement (grain size reduction), but when the grain sizes fall below ~1 μm, strengthening is usually accompanied by a drop in ductility and toughness (4). A recent study points the way to overcoming this problem: In a low-alloy steel containing ultrafine elongated ferrite grains strengthened with nanosized carbides, toughness and strength both rose when temperature was lowered from 60° to –60°C (5). In contrast, conventional metals become strong but brittle at lower temperatures. The authors attributed the observed toughening to the unique hierarchical anisotropic nanostructures in their steels.

Nanotwinned metals are another example of hierarchical nanostructured metals

Despite advances made in composite materials, metals remain irreplaceable in many important applications.

with extraordinary mechanical properties (3). When a high density of twin boundaries (highly symmetrical interfaces between two grains of the same lattice structure) is incorporated into polycrystalline copper grains, with boundary spacing in the nanometer scale, the material becomes stronger than coarse-grained copper by a factor of 10; it is also very ductile. The ultrastrong nanotwinned copper has an electrical conductivity comparable to that of high-conductivity copper (6) and a much enhanced resistance against electromigration (7). It has great potential for applications in microelectronics.

Corrosion is another headache for metals (8). To protect metals from corrosion, they are commonly coated with a layer of corrosion-resistant material. The Hangzhou Bay Bridge in China is an outstanding example of this technique. This 36-km-long bridge—the world's longest to date, with a design life of 100 years—is supported by several thousand pillars made of concrete-filled steel tubes ~80 m in length. The tubes are protected against corrosion in the harsh ocean environment by a coating of novel polymeric composites combined with cathode attachments.

Metal corrosion can also be resisted by forming a continuous protective passivation layer on the metal surface. For example, Yamamoto *et al.* (9) have added 2.5% Al to conventional austenitic stainless steels, resulting in the formation of a protective aluminum oxide layer that can resist further oxidation at elevated temperatures. Given their enhanced

Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang 110016, China. E-mail: lu@imr.ac.cn