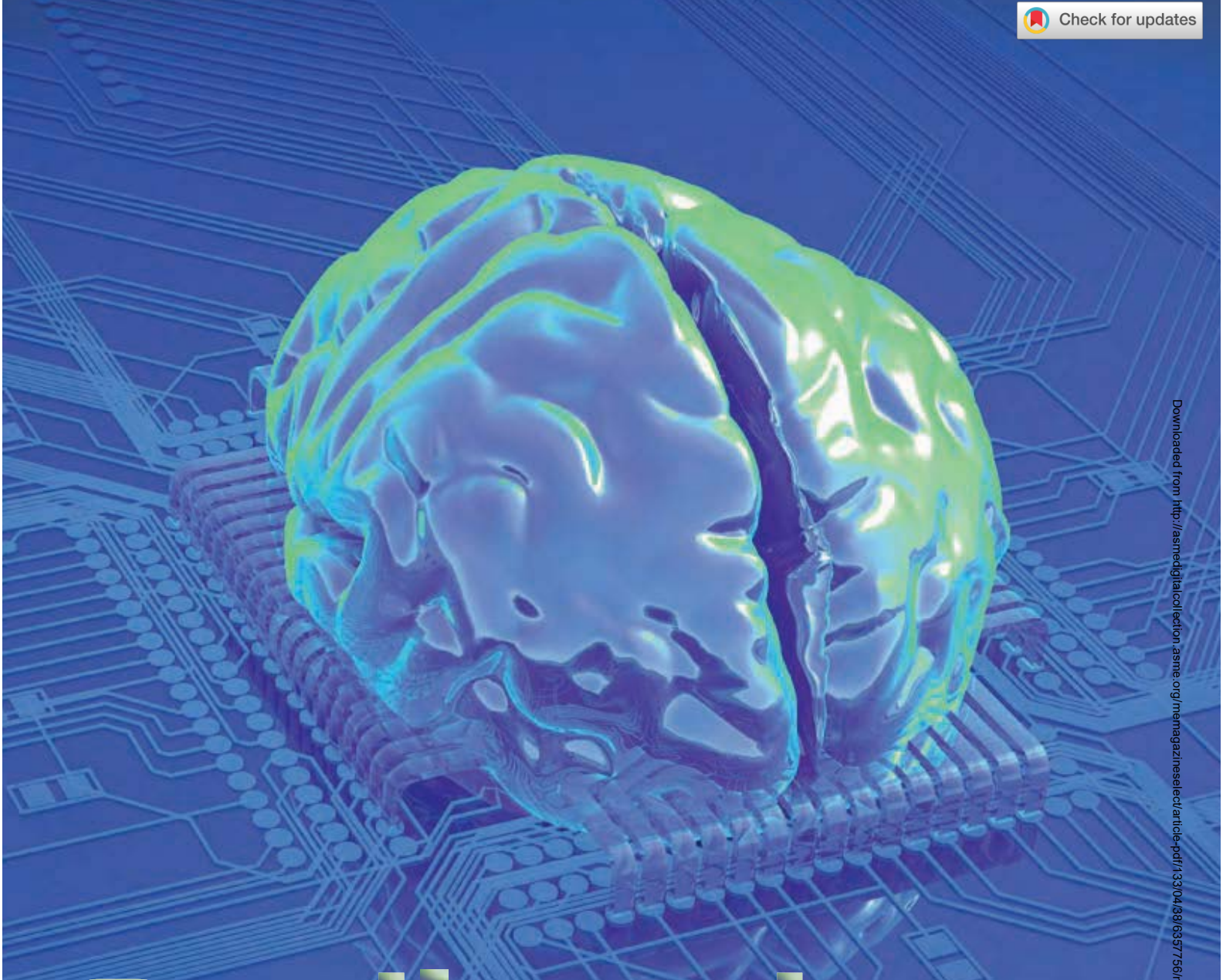


Cooling the Electronic Brain

Stacking processing chips can make for more compact computers. But it will take advances in microfluidics to make such dense number crunching practical.

BY AVRAM BAR-COHEN AND KARL J. L. GEISLER

The inexorable march towards smaller, faster, and more capable electronic systems has been breathtaking. In 1946, ENIAC, the first programmable computer, housed 50,000 vacuum tubes in 80 feet of cabinetry, drew 150 kilowatts of power, and performed 5,000 operations per second. Today's off-the-shelf Pentium microprocessor jams 2 billion transistors onto a 2.2 square centimeter sliver of silicon roughly 0.3-to-0.7-millimeter thick, draws 80 watts of power, and can perform 3.2 billion operations per second.

The development of the integrated circuit, independently invented by Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Semiconductor, in the late 1950s placed the information industry on this torrid pace of innovation. For the past 35 years, the number of transistors on integrated circuits has doubled approximately every two years. This doubling effect is the so-called Moore's Law, named after Intel cofounder Gordon Moore.

These thousands, then millions, and now billions of transistors switching on and off generate heat. In today's most advanced systems, silicon chips—called dies by their manufacturers—operating at 85 °C can generate average heat fluxes of more than 100 W/cm^2 and produce localized, submillimeter hot spots often exceeding 1 kW/cm^2. This is within an order of magnitude of the heat released into space by the surface of the sun. Without our ability to remove ever-greater heat fluxes from the surfaces of integrated circuits and other electronic components, we would never realize the benefits of their prodigious computational capability.

Engineers often specify air-cooled heat sinks or liquid-cooled cold plates to stabilize high-flux chips thermally. They are attached by successive layers of heat spreaders (usually copper plates that diffuse heat over a greater area) and thermal interface materials (often thermally conductive particle-filled silicones or greases).

CHIP STACKS

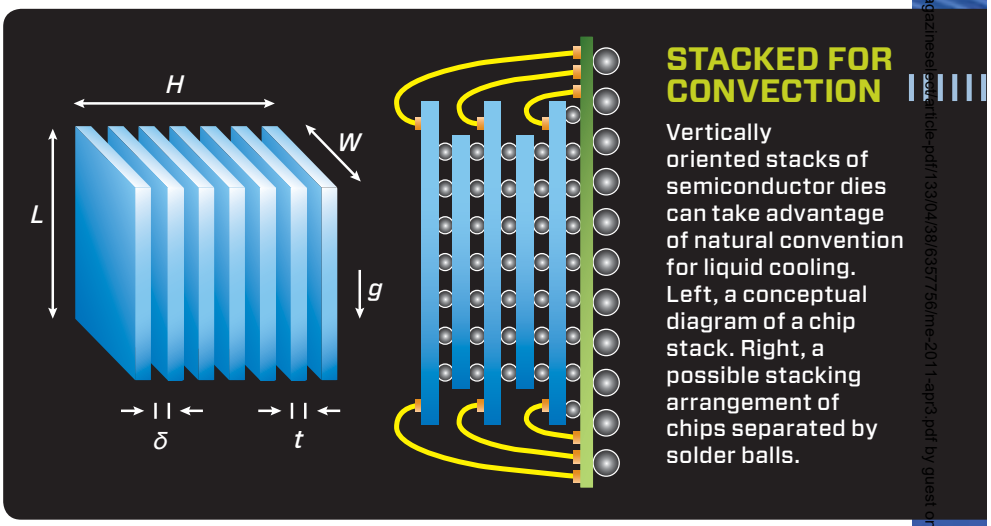
As the performance of microprocessors has begun to approach the complexity of the human brain, the three-dimensional architecture of nature's most powerful biological computer has inspired new ways to organize dies. One promising approach is the three-dimensional chip stack. Here, adjacent chips are piled directly above one another, typically separated by 10 to 50 micrometers, rather than located next to one another and separated laterally by 10 to 50 millimeters on a printed circuit board.

There are several reasons why chip stacks will help us maintain the cadence of Moore's Law far into the future. The vertical placement of one chip on top of the other provides a third dimension of interconnected transistors and functional electronic macrocells. Such close proximity—micrometers rather than millimeters—nearly eliminates significant time delays as signals travel between chips.

Equally important, fusing chips with different functional capabilities—processing, memory, power, communications, and environmental sensing—into a single chip stack could lead to compact microsystems of unrivaled capability and truly ubiquitous computing.

Engineers have already begun to design products to leverage those advantages. Stacks of memory chips have been in commercial use for more than ten years. Yet rudimentary stacks of two or three low-power logic dies are just now beginning to find application in portable computers and high-end cell phones.

Until now, however, thermal problems have stymied attempts to commercialize high-performance chip stacks. Such future chip stacks will require cooling systems that can handle heat densities above 1,000 W/cm^3 and heat fluxes above 100 W/cm^2 while maintaining chip operating temperatures below 100 °C at ambient temperatures of 40° C.



STACKED FOR CONVECTION

Vertically oriented stacks of semiconductor dies can take advantage of natural convention for liquid cooling. Left, a conceptual diagram of a chip stack. Right, a possible stacking arrangement of chips separated by solder balls.

COOLING OPTIONS

Today's baseline thermal packaging approach for three-dimensional chip stacks involves conduction through solids. Conduction carries heat generated by various chip and interconnect layers either up to the top of the package and its heat sink or down to the underlying printed circuit board and a supporting cold plate.

A more advanced approach to thermal packaging might add solid silicon carbide or diamond thermal conduc-

Avram Bar-Cohen is Distinguished University Professor of Mechanical Engineering at the University of Maryland in College Park. Karl J. L. Geisler is a research scientist/engineer with 3M in Maplewood, Minn.

tors, or interposers, between the bare dies. The interposers facilitate lateral transport of dissipated heat to the package periphery. Unfortunately, thermal contact resistance, which limits conduction between the chip and interposer, together with limitations to effective cooling of the interposer at its edges or the outside surfaces of a chip stack package, erodes the performance of this approach.

The thermal management of the human brain, the body's most concentrated heat source, offers a potential alternative. The brain's thermal management relies on blood flow through the extensive multiscalar network of blood vessels to stabilize its temperature. One near-term way to mimic this system would be to direct cool liquid from manifolds into microchannels machined or etched into the interposers. This solution overcomes edge-cooling limitations, but remains constrained by the resistance to thermal conduction between the chip and the solid body of the interposer. Moreover, engineers would need to drill hundreds or even thousands of through holes, or vias, into the interposers to accommodate electrical interconnections between chips.

A more elegant implementation involves immersing the chip stack in a dielectric liquid, which is electrically insulating. This exploits the gaps between the chips and other naturally occurring passages as channels for fluid flow through the three-dimensional package. It also provides high heat transfer rates directly at the chip surfaces, and avoids the detrimental effects of the contact resistance encountered with interposers.

That sort of embedded microfluidic cooling should prove very efficient in dissipating high rates of volumetric heat generation. It would eliminate hot spots, hold all silicon layers to a nearly uniform temperature, and accommodate Joule heating in high current-density interconnects. It could also eliminate detrimental flow instabilities, such as flow maldistributions, oscillations, and reversals, caused when momentary temperature spikes in one microchannel increase the heat dissipated into neighboring channels.

IMMERSION COOLING

Cooling microelectronic components with liquids began to attract serious attention in the mid-1980s, when IBM, Honeywell, Sperry-Univac, Control Data, and Hitachi all introduced indirectly water-cooled mainframe computers. These indirect techniques involved removing heat conductively from a chip or chip package, followed by convection to a liquid.

Despite their relatively primitive implementation, liquid cooling technology enabled engineers to significantly improve heat removal capacity and create denser, more efficient circuits. The barrier-busting performance achieved by aggressive cooling made these computers successful in the marketplace. It also set the stage for the development of more advanced thermal control techniques that could overcome the inefficiencies associated with conduction through solids and multiple interfaces.

Like indirect cooling, direct cooling of bare chips has a substantial history. Starting in 1985, four supercomputers employed dielectric liquids to cool the external surface of bare high-performance chips. Cray engineers used immersion in a perfluorinated liquid on the Cray-2, and gas-assisted evaporation of a perfluorinated liquid on the Cray-3. Control Data relied on immersion in liquid nitrogen for its ETA-10. Supercomputer Systems' SS-1 used submerged, impinging liquid jets to remove close to 100 W/cm^2 from the chip surface.

Thermal control of avionic components by direct immersion in low-boiling-point fluids dates back to the open cycle, pool evaporators developed in the late 1940s. Such systems can dissipate large amounts of heat over long operating periods, though they must condense and recirculate the vapor they generate. They can do this with a remote condenser connected to the immersion module; condensing surfaces placed directly in the vapor space above the boiling liquid; or a submerged heat exchanger in the liquid.

Researchers are now applying the experience gained with direct liquid cooling of macro-sized systems to the small, new, 1 cm^3 stacks of bare chips.

COOLING CAPACITY

In a paper presented at the August 2010 International Heat Transfer Conference (IHTC-14), organized by ASME on behalf of the Assembly for International Heat Transfer Conferences in Washington, D.C., we summarized the results of several of our distinct immersion cooling studies to define the potential of direct liquid cooling of three-dimensional chip stacks. The paper demonstrated that immersion cooling with dielectric liquids could become the cooling technique of choice for this emerging packaging paradigm.

We assessed four possible immersion cooling strategies. Two are passive: natural convection and pool boiling. Passive immersion cooling relies on natural circulation caused when a heated fluid rises to the top of a channel, sucking in cooler fluid behind it. Pool boiling is a two-

NO GUTS, NO GLORY

AVRAM BAR-COHEN discusses issues of electronics, thermal management, and global competitiveness. A video recording of Bar-Cohen's address to a plenary session of the 2010 ASME International Mechanical Engineering Congress & Exposition, in Vancouver, is available at www.memagazine.org.

phase process that occurs when surfaces get hot enough to vaporize the coolant. The phase change from liquid to gas absorbs more heat than convection alone, and the resulting vapor bubbles vigorously pump the liquid coolant through the channels and enclosure.

We also considered two active cooling strategies that augment natural circulation with a pumping mechanism to produce forced flow. This boosts flow rates, thus increasing the performance of both convective and ebullient (pool boiling) cooling methods.

In the study, we determined the cooling densities for all four immersion cooling techniques, assuming the use of Fluorinert FC-72, a commonly used perfluorinated dielectric liquid. For passive systems, the cooling densities ranged from 25 W/cm^3 for natural convection to 200 to 400 W/cm^3 for pool boiling. For active technologies, the densities were 100 to 300 W/cm^3 for forced convection and more than 2,000 W/cm^3 for flow boiling. We found the optimum die spacings for both single and two-phase direct cooling to be in the range of 0.2 mm to 0.6 mm for typical microelectronics geometries, though substantial cooling densities could be achieved at less-than-optimum spacings.

Immersion cooling's combination of thin die gaps and high (and often very high) cooling densities will make it possible to build full-powered stacks and three-dimensional packages comprising many different chips. Passive natural convection and/or pool boiling could provide the thermal management capability needed for the stacks of chips anticipated for use in smart phones, laptop and notebook computers, and many other types of portable equipment. Alternatively, the pumped flow of dielectric liquids through microgaps in three-dimensional stacks could meet many of the most extreme thermal management requirements for high-performance 3-D microsystems.

The continuing miniaturization of semiconductor transistors and the growing importance of three-dimensional packaging necessitate the parallel development of a new thermal management paradigm—embedded microfluidic cooling—that mimics the temperature stabilization technique of the mammalian brain. The first implementations of this new paradigm will likely involve microchanneled interposers, since the wiring in interposes lets engineers connect chips whose inputs and outputs would not ordinarily match up.

Eventually, those inputs and outputs will align. Then chip stack immersion in dielectric liquid, which exploits the gaps between chips created by microfabrication and assembly methods in addition to other passages in the three-dimensional package, can be expected to emerge as the thermal management technique of choice for three-dimensional microsystems.

The inherent advantages of direct liquid cooling are well established: It provides high heat transfer rates directly at the chip surface while avoiding the detrimental effects of contact resistance encountered with interposers. Nevertheless, additional research and development will be required to support the detailed design and optimization of such two-phase microfluidic cooling systems. ■

MANY WAYS OF STACKING

Chip stacks combine semiconductor dies and come in many possible configurations: (a) wire bonded die stacks; (b) wire bonded package stacks; (c) flip chip die stacks; (d) solder ball package stacks; and (e) substrate and solder ball folded stacks. As dies grow more powerful, they generate heat that is difficult to dissipate without liquid cooling.