

THERMAL CRISIS: CHALLENGES AND POTENTIAL SOLUTIONS

LI SHANG AND ROBERT P. DICK

FOR DECADES, DIGITAL INTEGRATED circuits (ICs) have benefited from the relentless progress predicted by Moore's Law. With each new technology generation, ICs become denser, faster, and cheaper. However, with increasing system integration, as well as aggressive technology scaling, IC thermal issues now stand in the way of improved IC performance.

Many of us may have watched a video from <tomshardware.com> called *What happens when the CPU cooler is removed?* This two-minute video clip shows the effect of removing a heatsink from a running 1,400 MHz AMD Athlon processor: the whole system crashes, the CPU temperature shoots up to 698 °F, and smoke starts coming out of the chip within a few seconds. Today's microprocessors can get extremely hot.

IC temperature is a strong function of IC power density. With each technology generation, the integration density of semiconductor devices doubles. As a result of reduction in both device feature size and supply voltage, the power consumption of each transistor decreases. However, increasing transistor counts and aggressive frequency scaling result in a significant increase in chip-power density, hence temperature.

Increasing chip temperature has significant impact on other design metrics including reliability, performance, cost, and power consumption. Circuit reliability depends exponentially upon operating temperature. When temperature increases, major reliability issues, such as electromigration (which affects

metal wires) and time-dependent dielectric breakdown (which affects transistors), become more important. Currently, temperature effects account for over 50% of electronic failures. IC temperature also affects circuit speed. Reduction of charge carrier mobility in

perature result in significant increases to overall system cost.

In addition to reliability and cost, chip temperature also affects power consumption. High-chip power consumption increases cooling power consumption, e.g., from fans. Leakage power consumption, a major problem in nanometer scale devices, also increases with chip temperature. Therefore, chip power consumption and temperature form a positive feedback loop, further increasing IC temperature. As projected by the International Technology Roadmap for Semiconductors, further process scaling for high-performance microprocessors will be bounded by power consumption and heat dissipation below 65 nm: it is critical to address energy and thermal issues to meet the urgent needs of the semiconductor industry and enable future technology scaling.

In the past, thermal issues were widely ignored during circuit design and solely addressed by packaging and cooling designs based on worst-case chip power consumption and thermal profile. As a result of increases to chip power densities and cooling costs, worst-case cooling design now faces fundamental difficulties. To avoid thermal crises, we must simultaneously and cooperatively address thermal issues at each stage of the design process: during packaging and cooling design, architecture design, circuit design, fabrication, and the development of novel devices.

Packaging and cooling

IC cooling solutions can be roughly partitioned into two categories: passive and active. Passive cooling



© ISWOOP & COREL

transistors and increased interconnect latency resulting from high IC temperature degrade performance. Thermal variations can also lead to significant timing uncertainty, requiring wider time margins, and resulting in poor performance. To guarantee reliable run-time operation, cooling systems need to handle the worst-case chip temperatures. Chip packaging and cooling currently accounts for 30% of IC cost and cooling cost increases superlinearly with power consumption. Thus, increasing chip power and tem-

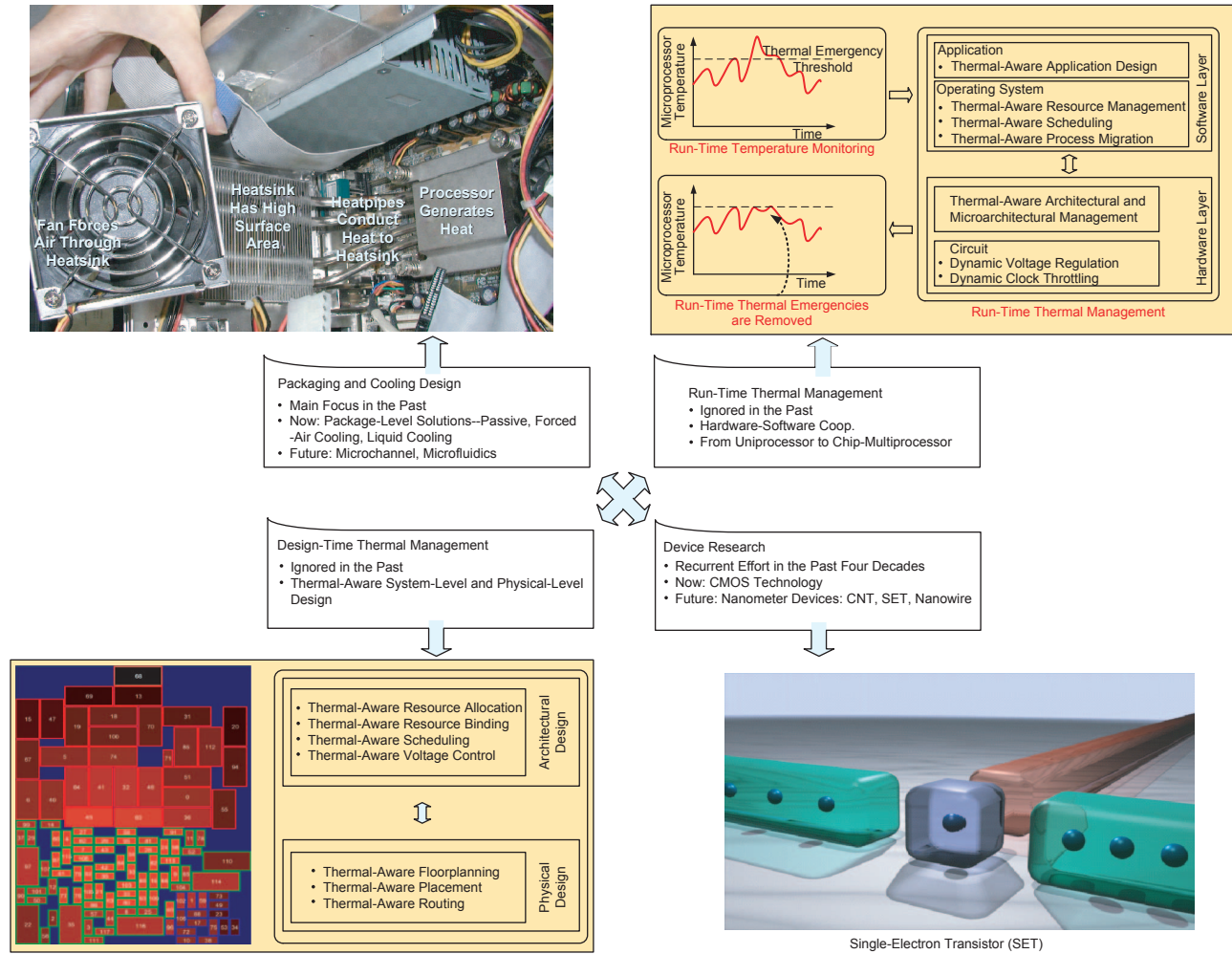


Fig. 1 Potential solutions to thermal crisis

solutions mainly rely on conduction and natural convection. Active cooling uses external forces, such as forced convection. Active cooling is generally much more effective than passive cooling. Various factors need to be considered during cooling design, e.g., cooling efficiency, cost, space, and noise.

The next time you help your parents clean up the garage, you can examine their old 80386 computer. In modern computers, the microprocessor cooling consists of a huge heatsink, heatpipes, and a cooling fan. In your parents' old toy, you may only see a small plastic package without extra cooling. Fundamentally, both the shabby plastic package of the 80386 and the fancy heatsink-heatpipe-cooling fan solution serve

the same thermal purpose—facilitating transfer of heat produced by the microprocessor to the surrounding environment. However, the power consumption of the 80386 is only around 2 W, while modern microprocessors easily consume 100 W. As chip power consumption increases, cooling solutions naturally become more complicated and costly.

Cooling design for today's high-power microprocessors requires detailed thermal modeling and analysis using complex numerical analysis methods. However, the basic idea of cooling design is quite simple. There is a well-known duality between heat transfer and electric phenomena. Heat flows through thermal resistors, charges and discharges thermal capacitors, and causes temperature varia-

tion. In fact, our modeling method for electrical circuits was first proposed by Jean Fourier in 1822 for thermal modeling, and later used to model electrical circuits by Georg Simon Ohm in 1827.

Forced-air cooling is the most widely used cooling technology for personal computers. Heat generated by transistors and metal interconnects flows through the silicon die (with thermal resistance R_{die}) as well as cooling and packaging layers (with thermal resistance R_{colpkg}) by heat conduction, and then dissipates to the ambient environment by heat convection. Heat convection cooling efficiency is proportional to A_{colpkg} , the interface area between cooling package and ambient environment. Coefficient b is a function of air flow rate determined by the cooling fan

$$T_{\text{junction}} = (R_{\text{die}} + R_{\text{colpkg}}) \times P + \frac{h}{A_{\text{colpkg}}} \times P + T_{\text{ambient}}$$

This equation implies a tradeoff between cooling via conduction and convection. As shown in this equation, extra packaging and cooling material layers are detrimental to thermal conductivity. The purpose of these cooling layers, such as the heatsink, is to improve convection efficiency by increasing the interface area with the ambient environment, A_{colpkg} .

To achieve high cooling efficiency within limited space, e.g., heatsink-heatpipe-cooling fan based designs have been widely used. The upper-left quadrant of Fig. 1 shows a photograph of one such design used to cool an AMD Athlon processor within a compact Shuttle XPC computer. Heatpipes use a two-phase heat transfer mechanism that provides order of magnitude better cooling efficiency than an equivalent piece of copper. A heatpipe is a two-layer, closed, evacuated cylindrical vessel. Inside a heatpipe, coolant vaporizes at a high-temperature zone, flows along the pipe, and condenses at a low temperature zone. The condensed coolant is then returned to the high-temperature zone via a wick.

Power consumption in modern microprocessors is approaching the point at which forced-air cooling becomes impractical. As a result, liquid cooling, which was widely used in high-end servers before the complementary metal oxide semiconductor (CMOS) era, has started appearing in the personal computer market. One example is the Microsoft Xbox 360. Its reliability has been questioned due to overheating. Various liquid cooling solutions were developed to replace the default forced-air cooling configuration. Liquid cooling is analogous to the use of a radiator in a car. Liquid coolant circulates through the heatsink, absorbs heat from the microprocessor, and dissipates heat to the ambient environment. Compared to forced-air cooling, liquid cooling is much more efficient and quieter, but has other disadvantages, e.g., bulk, cost, complexity, and risk of leaking.

As semiconductor technology scales further, due to high power density and nonuniform heat generation, existing package-level cooling tech-

nologies face two main challenges. First, local hotspots cannot be effectively handled by package-level cooling. One recent example is the data corruption problem in AMD Opteron processors that resulted from overheating floating-point units. Second, the thermal contact resistance between silicon die and heatsink affects the efficiency of heat conduction. To address these challenges, much attention has been focused on micro and nano cooling technologies.

Since first introduced in 1986 by Tuckerman and Pease, microchannel cooling has been one of the most effective cooling solutions. Numerous channels with diameters in the range of millimeters to micrometers are embedded within a thin layer of cooling material or directly etched into a silicon die. Within these channels, coolant transfers heat through either single-phase (liquid only) or two-phase (liquid and gas) flow. It has shown that a microchannel heatsink can handle heat density in the range of unit $1,000 \text{ W/cm}^2$. In addition, by controlling the operation of microchannels at hotspot regions, local thermal emergencies can be eliminated. Furthermore, since microchannels can be integrated directly into the silicon substrate, thermal contact resistance can be avoided. Therefore, the two main cooling objectives, reduction of overall chip temperature and elimination of local hotspots, can be achieved by microchannel-based cooling solutions.

Run-time thermal management

As dynamic power management techniques, such as voltage scaling and clock gating, become more widely used in ICs, run-time power and thermal variations increase. The worst-case chip power consumption, used to guide cooling design, is seldom observed. Cooling cost increases superlinearly with chip power consumption. Therefore, conservative cooling design targeting the worst-case power and thermal scenario becomes increasingly excessive and infeasible.

Run-time thermal management techniques have recently been proposed by the computer architecture research society to address the drawbacks of conventional cooling design. Using these techniques, run-time chip

power consumption and thermal profile are monitored. ICs remain in normal operation as long as their temperatures are within the reliable operational range. When run-time thermal emergencies occur due to prolonged high power consumption, thermal-aware control techniques are engaged to eliminate thermal emergencies by reducing the power consumption, and performance, of the microprocessor. The upper-right quadrant of the figure shows how run-time thermal management uses a control-feedback loop to prevent thermal emergencies in an IC. Run-time thermal management allows cooling solutions to target average-case, instead of worst-case, chip power consumption, thereby significantly reducing cooling cost. As worst-case power consumption occurs rarely, the performance penalty is low.

Recently, various run-time thermal management techniques have been developed. The Intel Pentium 4 is one of the first microprocessors with built-in run-time thermal control. The processor contains one on-die thermal sensor and one on-die thermal diode. The on-die thermal sensor protects the processor in the event of a catastrophic cooling failure by throttling the system clock or completely shutting down the microprocessor. The on-die thermal diode enables off-chip temperature measurement. Processor run-time temperature gathered by the thermal diode can be fed back to system software layers, i.e., operating system and applications, thus providing the opportunity for efficient hardware-software cooperative run-time thermal control. Microprocessors from AMD also provide similar thermal monitoring and management capabilities. The video from <tomshardware.com> is somewhat misleading. The thermal failure is mainly due to the motherboard used in the test, which does not support an on-board thermal diode.

Spatial thermal variation and local thermal hotspots are becoming increasingly important. A distributed thermal monitoring and management solution is used in the IBM POWER5 processor. The POWER5 employs 24 on-chip thermal sensors to track the run-time temperature of each power-hungry component. Each component can be throttled individually. This approach is more effective at handling local thermal hotspots with lower

performance penalty than the global throttling approach used in the Pentium 4.

Power consumption and wire delay constraints are pushing micro-processor design toward the chip-multiprocessor (CMP) paradigm. For instance, the recently developed Cell processor for the Sony PlayStation 3 contains nine on-chip processing elements. In CMPs, the chip temperature profile results from thermal interactions among all the processors. In addition, each individual on-chip processor may operate independently at run-time. This distributed scenario imposes unique challenges on run-time thermal management. Distributed, cooperative management techniques are essential for efficient run-time temperature control. The Foxton technology recently proposed by Intel serves as a good starting point for CMP thermal management. Foxton allows run-time control of both supply voltage and frequency of each individual on-chip processor, which can potentially lead to an efficient distributed run-time thermal management solution. Software layers, such as the operating system, may also play an important role for run-time thermal management of CMPs in the future.

Design-time thermal optimization

Design techniques can be used to improve the power and thermal characteristics of integrated circuits. Some of the most promising techniques are voltage scaling, thermal-aware architectural planning, and thermal-aware physical design.

Voltage scaling has allowed dramatic reduction of power consumption in CMOS, the dominant IC technology today, and for years to come. In static CMOS circuits, every output node is driven through transistors, some of which are connected, in series or parallel, to a high voltage (V_{DD}) and some of which are connected to ground. Under normal operation, either the path to V_{DD} or ground is closed and the other path is broken. Output nodes are connected to the insulated gate of the next logic device. Therefore, if a CMOS device is not changing values, its power consumption is low. If we assume that a CMOS device switches with some frequency, f , its power consumption can be expressed as

$$P = fCV_{DD}^2$$

where C is the capacitance of the subsequent interconnect and logic device. From this equation, it is clear that we can improve power by reducing f , C , or V_{DD} . Unfortunately, performance reduces in proportion to f so the energy required to finish a given amount of work will remain constant even if f is reduced. C can be reduced by decreasing the sizes of devices and interconnects. However, existing fabrication technologies impose limits on minimum feature sizes. Reducing V_{DD} turns out to be one of the most promising methods of reducing power consumption and temperature.

Performance reduces approximately in proportion to V_{DD} . However, unlike frequency scaling, the power consumption improvement resulting from voltage scaling is quadratic in voltage reduction. Therefore, if we halve V_{DD} and halve f to compensate for the increased delay, the resulting power consumption is 1/8 its original value. Of course, with f halved, it takes twice as long to complete a task. However, we can use other design techniques to bring the performance back to the original level, e.g., by doing more work in parallel with the help of additional hardware. This results in a circuit that runs as fast as the original circuit but only consumes 1/4 the power of the original circuit. Voltage scaling has been one of the most useful tools available to low-power designers for many years. However, it will soon reach its limit. V_{DD} values are now approaching the threshold voltages of transistors.

Despite the use of power-aware design techniques, it is often necessary to consider thermal effects. Some portions of ICs produce more heat in less space than others. For example, the floating-point units and instruction issue units in microprocessors typically have higher power densities than caches. Heat flows slowly sideways in integrated circuits because the silicon from which the circuit is built has low thermal conductivity. However, once the heat makes it through the silicon vertically, it has a fast path to the ambient via the high thermal conductivity heatsink. As a result, the thermal resistance between a region of an integrated circuit and the ambient is largely a function of its area. If one

region of an integrated circuit produces heat more rapidly than another, all other things being equal, its temperature will be higher.

One can use thermal-aware physical design techniques to reduce the problem of high power density. The lower-left quadrant in Fig. 1 illustrates an IC floorplan. In this figure, each rectangle is a functional unit that can carry out operations such as multiplication and addition, temperature is indicated by functional unit brightness, and different voltages are indicated by the color of functional unit outlines. The design was automatically produced by a thermal-aware synthesis algorithm. This algorithm automatically partitioned the functional units into different voltage islands. This reduced power density, allowing a reduction in peak temperature from 90.8 °C to 81.4 °C. In addition, thermal-aware floorplanning was used to move high power density functional units apart, further reducing peak temperature from 81.4 °C to 76.2 °C.

Interactions among thermal optimization techniques

Thermal problems can no longer be efficiently solved at one level of the design process. It is now necessary to consider, and optimize, the thermal impact of cooling and packaging, run-time thermal management techniques, and thermal-aware design techniques. Neglecting thermal problems at any level of the design process forces them to be solved at other levels, potentially at much greater cost. On the other hand, synergistic techniques that span multiple design levels can potentially yield highly efficient thermal solutions. Thermal-aware design combined with chip and package cooling design make thermal emergencies rare and run-time thermal management prevents these, now rare, thermal emergencies from causing reliability problems.

Novel devices

Thermal issues constrain further improvements in CMOS integrated circuit performance even when a global approach to thermal optimization is used. Thermal crisis has been a recurring theme during the past four decades. Device innovations have been the most effective method of dealing with increasing power densities.

ENIAC was one of the first electronic digital computers. Its construction started in 1944. ENIAC's 18,000 vacuum tubes consumed 140 kW of power. Compared to vacuum tubes, semiconductor devices are much more power-efficient. For instance, using CMOS technology, an ENIAC-on-a-chip only takes 40 mm² silicon area with 0.5 W power consumption. Power and thermal concerns were some of the main motivations for replacing vacuum tubes with semiconductor devices in the 1960s.

History repeated itself 30 years later. Before the 1990s, high-end servers from IBM used bipolar circuitry, which is faster but also consumes more power than CMOS. The high power consumption of bipolar devices forced IBM to move to CMOS technology in early 1990s. Using CMOS technology, Z9, today's biggest IBM mainframe, provides more than an 80-times speedup over, and consumes only half of the power of, the ES/9021 used at the end of the bipolar era.

History has shown us that, when vacuum tubes were replaced by semiconductor devices, and bipolar replaced by CMOS, circuit power efficiency improved by orders of magnitude. CMOS is the mainstream fabrication technology used today. As technology and system integration scale further, CMOS technology is also approaching fabrication, power consumption, and thermal limits. Researchers have been pursuing new solutions to replace CMOS for decades. Recent breakthroughs in nanotechnology, e.g., carbon nanotube transistors, nanowire logic, and single-electron devices, may provide alternatives with better power consumption and thermal characteristics.

A carbon nanotube is a sheet of carbon atoms that has been rolled into a tube a few nanometers in diameter. Using nanotubes, the energy consumption of a single transistor can be reduced to 1×10^{-11} J per switching event. Recently, researchers have demonstrated simple circuit designs using chemically doped semiconductor nanowires, which have power efficiency similar to carbon nanotubes. It is projected that, using carbon nanotubes or nanowires, circuit integration density can be improved by orders of magnitude, with power density remaining within the same range as current CMOS circuits.

A single-electron tunneling transistor is a nanometer scale island with very low capacitance isolated from source and drain leads by tunnel junctions. The lower-right quadrant of Fig. 1 illustrates a single-electron transistor. The green block to the right is the source, the green block to the left is the drain, the blue block in the center is an island, and the red block to the top is the gate. Electrons, illustrated by blue spheres, may tunnel from source to island, and from island to drain. The operation of a single-electron transistor is governed by the Coulomb charging effect, in which the energy to charge or discharge the island is greater than the available thermal energy. By tuning the voltage of a gate, which is capacitively coupled to the island, this charging energy can be changed, thereby controlling the transport of each individual electron. The International Technology Roadmaps for Semiconductors projects that single-electron transistors hold the promise of achieving the lowest power of any currently-known logic device. The energy consumption of a single-electron transistor can be as low as 1×10^{-18} J per switching event.

Conclusions

Novel nano-devices may ultimately be the solution to the power and thermal limits of CMOS transistors. However, nanotechnology researchers have grappled with fabrication challenges for decades. At present, most circuits based on nanodevices cannot be economically manufactured, i.e., most remain in the realm of research laboratories. Nevertheless, many expect the transition from CMOS to a new nanotechnology to occur within the next 15 years.

Read more about it

- International Technology Roadmap for Semiconductors, <http://public.itrs.net>.
- A.E. Bergles, "Evolution of cooling technology for electrical, electronic, and microelectronic equipment," *Trans. Components Packaging Technologies* vol. 26, no. 1, pp. 6–15, Mar. 2003.
- J. Clabes, J. Friedrich, M. Sweet, J. DiLullo, S. Chu, D. Plass, J. Dawson, P. Muench, L. Powell, M. Floyd, B. Sinharoy, M. Lee, M. Goulet, J. Wagoner, N. Schwartz, S. Runyon, G. Gorman, P. Restle, R. Kalla, J. McGill, and S. Dodson, "Design and

implementation of the POWER5 microprocessor," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 2004, pp. 56–57.

- C. Poirier, R. McGowen, C. Bostak, and S. Naffziger, "Power and temperature control on a 90 nm Itanium®-family processor," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 2005, pp. 304–305.

- R.P. Dick, L. Shang, and N.K. Jha, "Power-aware architectural synthesis," in *The VLSI Handbook*, W.-K. Chen, Ed. Boca Raton, FL: CRC Press, 2006.

- K.K. Likharev, "Single-electron devices and their applications," *Proc. IEEE*, vol. 87, no. 4, pp. 606–632, Apr. 1999.

Acknowledgments

We would like to thank Prof. Seda Ogrenci Memik for suggesting improvements to this article.

About the authors

Li Shang (li.shang@ee.queensu.ca) received his B.E. and M.E. degrees from Tsinghua University and his Ph.D. degree from Princeton University. He is currently an assistant professor at the Department of Electrical and Computer Engineering, Queen's University, Canada. He has published in the areas of computer architecture, design automation, thermal/power modeling and optimization, reconfigurable computing, and nanocomputing. He won the Best Paper Award at PDCS'02 and his department's teaching award in 2006. He is the Walter F. Light Scholar of Queen's University.

Robert P. Dick (dickrp@eecs.northwestern.edu) received his B.S. degree from Clarkson University and his Ph.D. degree from Princeton University. He worked as a visiting researcher at NEC Labs America, a visiting professor at Tsinghua University's Department of Electronic Engineering, and is currently an assistant professor at Northwestern University's Department of Electrical Engineering and Computer Science. He received an NSF CAREER award and won his department's Best Teacher of the Year award in 2004. He has published in the areas of embedded system synthesis, mobile ad-hoc network protocols, reliability, behavioral synthesis, data compression, embedded operating systems, and thermal analysis of integrated circuits.