Cooling and Power Considerations for Semiconductors Into the Next Century

Christian Belady
Hewlett-Packard Company
Richardson, Texas
972-497-4049 / belady@rsn.hp.com

Introduction

With the insatiable desire for higher computer or switch performance comes the undesirable side effect of higher power especially with the pervasiveness of CMOS technology. As a result, cooling and power delivery have become integral in the design of electronics. Figure 1 shows the National/International Technology Roadmap For Semiconductors' projection for processor chip power.

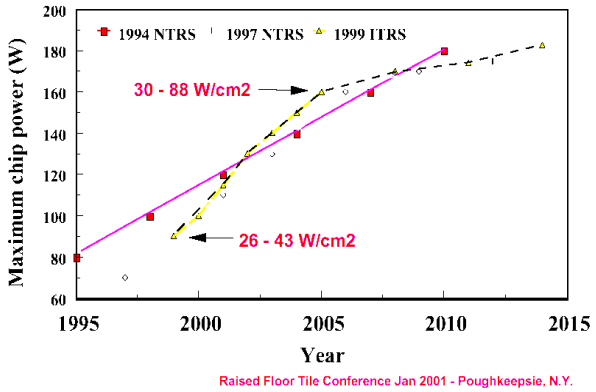

Figure 1. Projection of processor power by the National/International Technology Roadmap For Semiconductors

Note that between the year 2000 and 2005 that the total power of the chip is expected to increase 60%, which will put additional emphasis on the power and cooling systems of our electronics. Further inspection of this figure also shows that the heat flux will more than double during this period. The increases in power and heat flux are driven by two factors, higher frequency and reduced feature sizes.

The objective of this paper is to provide a high level review of some of the cooling and power challenges that our industry will be facing in the coming years. They are:

1) Package Level Cooling. If trends continue, processor cooling will become the design limiter. This is evidenced even further by Pat Gelsinger, VP and CTO of Intel, who said "Looking forward, we think power and power density become fundamental limitations that we have to fully address in our technology bag of tricks". This part of the paper will review issues of semiconductor cooling as well as some of the current and emerging cooling solutions. A related issue is power delivery to the semiconductor package (i.e. current step load and di/dt), but is outside the scope of this paper and will not be discussed here.
2) Infrastructure Level Power and Cooling. If trends continue, the burden on data center and central office owners may prohibit them from upgrading their systems. Taking things a step further, even if they did upgrade their systems it's questionable whether the utility infrastructure can keep up with the growing demand of electronics. This is evidenced with what we are currently seeing with the Utility fiasco in California. This part of the paper will review some of the room infrastructure technologies as well as look at global power issues.

Keep in mind that this paper is not meant to be comprehensive but rather is to promote discussion on this topic and highlight an opportunity for our industry to attack proactively.

Package Level Cooling

Package level cooling has emerged as one of the key core competencies for most electronics companies. In 1990, only a very few large computer manufacturers or defense contractors actually employed thermal engineers. At the turn of the century, every major player in the computer and telecom business had a significant staff of thermal engineers for package level cooling issues. In fact, most computer and telecom start-ups today actively recruit thermal engineers prior to mechanical design engineers.

So what has caused the shift in competency focus for companies? There are two primary factors: increasing power and increasing heat flux. Thermal designers focus on both of these. Their designs need to have the capacity to deal with moving the heat from point A to point B, but they also need to extract the heat effectively from the source. As power goes up feature sizes go down. Going from a feature size of 0.25 micron to 0.1 micron would increase the heat flux by six times if power is held constant. Figure 2 shows a power map

ISLPED'01, August 6-7, 2001, Huntington Beach, California, USA.

(Figure 2a) of a typical processor die with its accompanying temperature distribution (Figure2b). Note that only a quarter of the die is actually dissipating almost all of the power. This is a result of the fact that ¾ of the die is actually on-chip-cache. Typically, in these situations, heat spreaders such as diamond and copper have been used on the die to lower gradients.

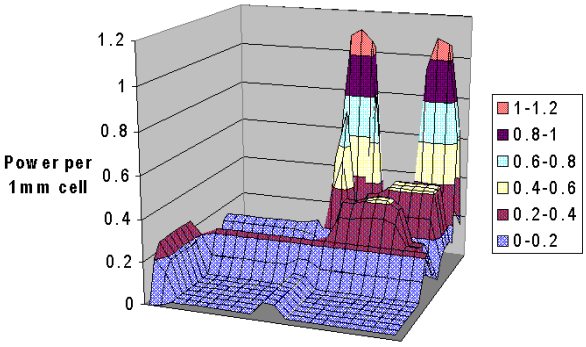

Figure 2a. Processor die power map (Courtesy J. Deeney)

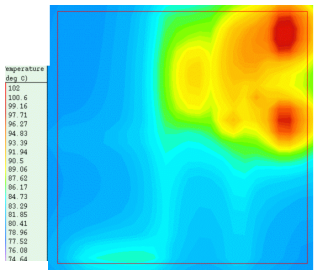

Figure 2b. Processor temperature distribution (Courtesy J. Deeney)

Interface Issues

As a result of this high heat flux, die-attach and interface techniques have been an area of significant attention. Figure 3 shows a typical high performance processor package.

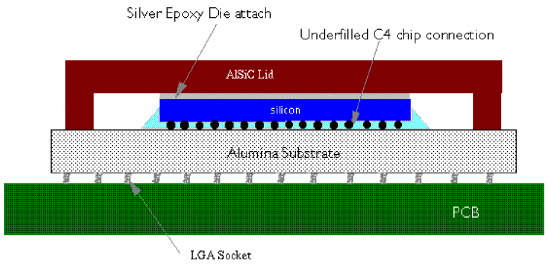

Figure 3. Typical high performance processor package.

Heat flows from the die through the die attach layer, into the lid, through another interface such as grease, into an air-cooled or liquid cooled heat sink. The silver epoxy die attach drives the thermal performance of the package.

So this brings us to one of the key problems, heat fluxes are going up faster than interface material and die attach improvements. There are countless companies who are trying to come out with the miracle material such as conductive greases, epoxies, phase change materials and pads to solve this problem. None of these have provided major breakthroughs to keep pace with shrinking feature sizes. There have been some promising studies in the past such as the work by Dolbear [1] in the area of metal pastes which provide an order of magnitude improvement in interface performance but have many issues at this point that need to be resolved before they can be commercialized.

Package Cooling Technologies

There is number of techniques used for cooling semiconductor packages that have become quite widespread in the industry. But as with everything else, the cooling demands of semiconductors are forcing the need for more aggressive solutions.

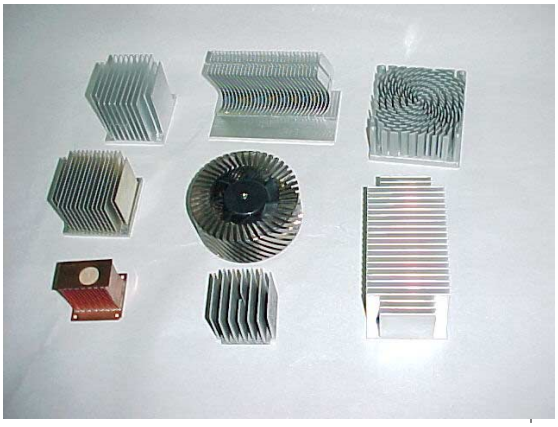

Figure 4. Various heat sink technologies.

The most common cooling technique is the use of heat sinks as shown in Figure 4. These devices are typically attached to the package lid or directly to the die. Their primary purpose is to increase the area for heat rejection to the air. In the figure all of the heat sinks are considered passive except for the center fan-sink, which is considered an active heat sink. These examples show heat sinks that were produced using various techniques such as brazed copper folded fin (lower left), bonded fin(left middle), machined (upper left), skived (upper middle), cold forged (upper right) and extruded (lower right). A more detailed survey was completed by Chu et al [2].

Still another common technique that has emerged in the past decade is the use of heat pipes. Their primary purpose is to aid in the transport heat from the source to the sink. Heat pipes are hollow vessels that contain a small amount of working fluid such as water with a wicking structure. Figure 5 shows an illustration of how a heat pipe works. The fluid evaporates at the heat source and the vapor travels up the center of the pipe and condenses at the sink. The condensed fluid wicks back to the evaporator. Since this cycle takes

advantage of latent heat, the operating temperature is virtually uniform along the pipe and thus, effectively creating an infinite conductor. Figure 6 shows various applications of the use of heat pipes. The heat pipe to the left is a tube in fin heat pipe while the heat pipe to the right is an embedded heat pipe both of which where used in HP's V-Class servers. The heat pipe in the top center is a tower heat pipe used in Convex's C3 server with a sectioned sample just below. At the bottom center of the figure is a typical laptop heat pipe/spreader plate.

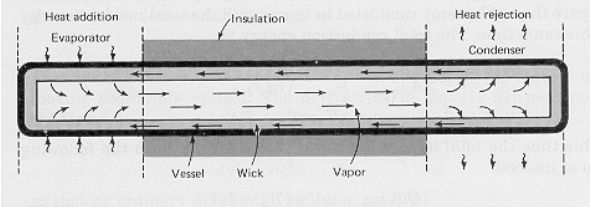

Figure 5. Heat pipe illustration [3].

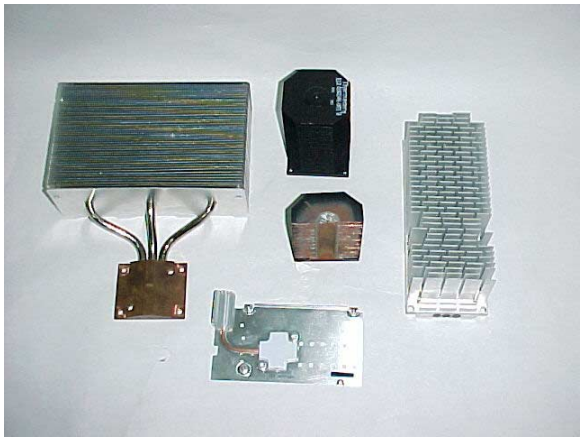

Figure 6. Heat pipe examples.

Although heat sink and heat pipe technology has allowed conventional air-cooled techniques for system thermal management, it becomes obvious that if power trends continue more aggressive cooling techniques will be required. There are many promising technologies emerging but the most promising in the near term are liquid cooling, refrigeration and spray cooling.

Although liquid cooling has been used in the past in mainframe computing until the early 90s, liquid cooling disappeared with the adoption of CMOS. Figure 7 shows an example of a liquid cooled system. Typically, these systems are made up of five major components, a pump, a coldplate, a heat exchanger, pipes and depending on the design, an expansion tank. Note that the example in the figure has five coldplates. The advantage of liquid is that water, for example, it has about 5300 times the heat capacity of air for a given volume with 1 to 2 orders of magnitude higher heat transfer coefficients. This allows significantly more compact cooling solutions. But there are issues that need to be resolved such as pump reliability, cost, weight and the hazard of leaks.

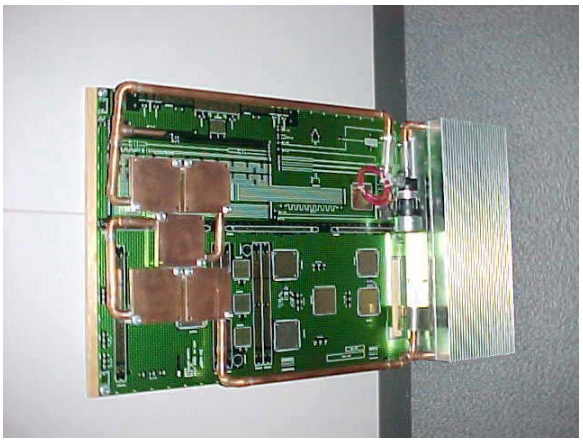

Figure 7. Liquid cooled boards.

In the late 1990s, refrigeration of semiconductors gained some popularity and found its way into some products such as IBM's S/390, DEC's AlphaStation 600/800 and KryoTech's K6 PC [4]. In these systems, refrigeration has been used as an active thermal management system to move heat from point A to point B and also to refrigerate the semiconductor below ambient temperatures to improve performance of the processor. Figure 8 shows how frequency can be increased as a function of processor temperature. It is obvious that the performance benefits for sub-cooling processors can be huge but there are concerns as well. First, condensation is a catastrophic problem that requires meticulous evaporator insulation design. In addition, reliable compressors are bulky. Figure 9 shows an example of a positive displacement vapor-cycle refrigerated PC. Note that for this type of system, the primary components are the condenser, compressor, insulated evaporator (KryoCavity) and accumulator (not shown).

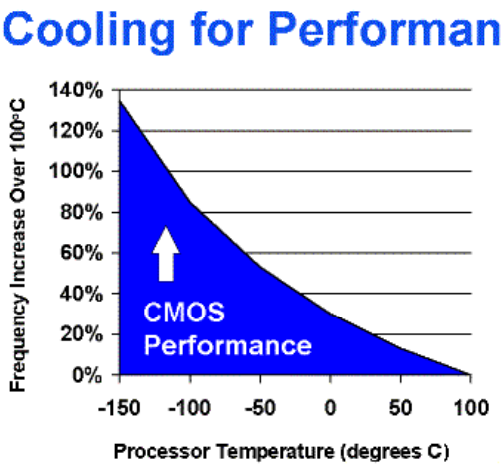

Figure 8. Frequency improvement as a function of temperature (courtesy of KryoTech)

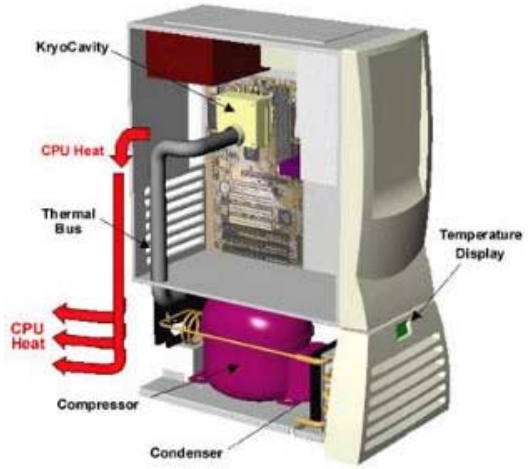

Figure 9. PC with vapor-cycle refrigeration system (courtesy of KryoTech)

All of the systems above use positive displacement pumps but there are a variety of other approaches such as acoustic compressors that vibrate a tuned cavity to create a standing way in the vessel as shown in Figure 10. This shows promise because there are no moving parts and thus, reliability should be excellent though there are noise and vibration issues. Still another interesting refrigeration technique is the use of thermoelectric coolers, which are solid state heat pumps that have no moving parts and operate at low pressure. Figure 11 shows a liquid to liquid thermoelectric cooler but liquid to air

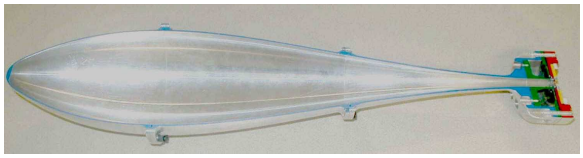

Figure 10. Cross section of acoustic compressor (courtesy of MacroSonix)

Figure 11. Thermoelectric refrigerator (courtesy of ThermoTek)

designs are available as well. This technique is not as efficient as the vapor phase approach. Figure 11 shows an example of such a cooler.

technique that takes advantage of the latent heat of evaporation by spraying Flourocarbons directly on the chip or the board. Thus far, almost all of the applications have been in government-funded programs but is on the verge of being commercialized. The advantages are huge if one could resolve some of the material compatibility issues that are being tested by current projects. Some key advantages are the ability to have uniform temperature surfaces throughout the system as well a elimination of interfaces by evaporating off of the back side of the die. Figure 12 shows an illustration on how spray cooling works. Figure 13 shows some early prototype multi-chip modules as well as the fan/pump assembly.

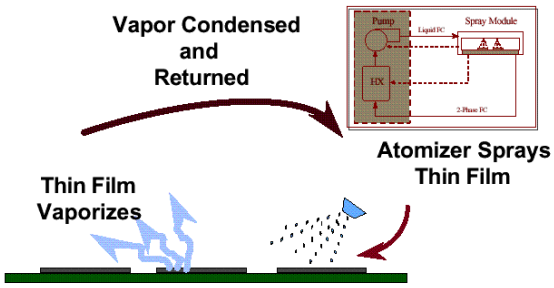

Figure 12. How spray cooling works (courtesy of Isothermal Systems Research)

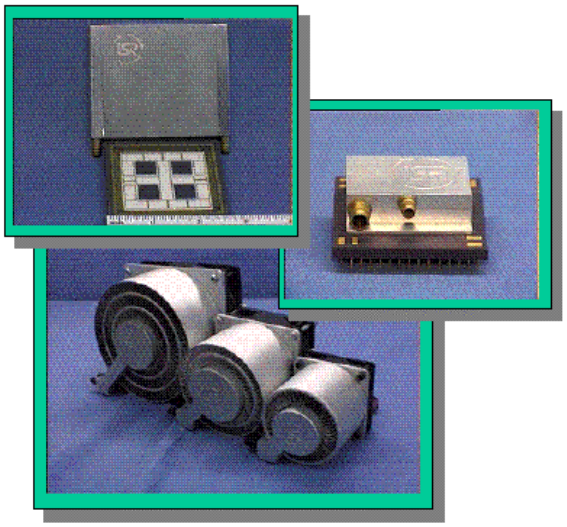

Figure 13. Prototype spray cooling units (courtesy of Isothermal Systems Research)

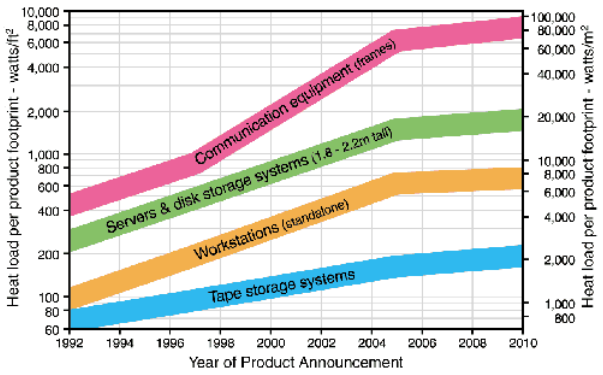

Figure 14. Heat-Density Trends in Data Processing, Computer Systems, and Telecommunications Equipment [5]

Infrastructure Level Power and Cooling

Ultimately, all of this heat that is generated by the chips does impact the electronics system but even worse has direct impact on our ultimate customers who are owners of large data centers such as banks, internet service providers and government labs. In 1998, HP completed a study on data center capacity including all of the major manufacturers current generation servers. The findings showed that most data centers had a capacity of 40 to 70 Watts/ft^2 but at that time the industry was shipping computers that ultimately would require capacities of 75 to 225 W/ft^2 and projections showed that in the next 5 years 500 W/ft^2 capacity will be required. At the time this was alarming because the industry as a whole was not aware of this disconnect. Since then this has become a major issue and a group of manufacturers developed a projection of power for the next decade[5]. Figure 14 shows the projections of the collaborating manufacturers, which included Amdahl, Cisco, Compaq, Cray, Dell, EMC, Hewlett-Packard, IBM, Intel, Lucent, Motorola, Nokia, Nortel, Sun and Unisys.

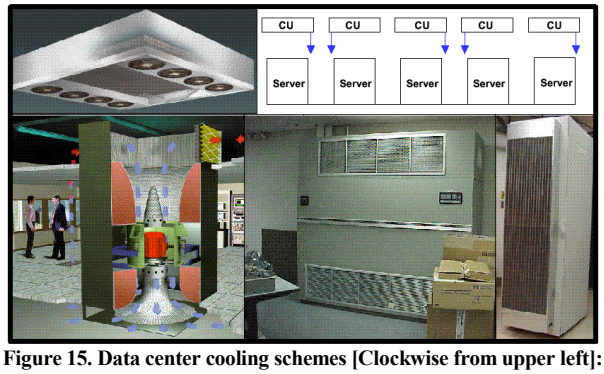

Figure 15. Data center cooling schemes [Clockwise from upper left]: ENP (Emerson Network Power) DataCool, ENP DataCool over head cooling scheme, ENP "Backpack", ENP System 3, M&I Column Fan.

Traditionally data centers have used CRAC (Computer Room Air Conditioning) with air distributed under the floor. Typically, systems like Emerson Network Power's (ENP) System 3 (lower middle in Figure 15) have been used in many of the Data Centers in the Continental US. These have worked very well in the under 100 W/ft^2 but have had some limitations in the above 100/ft^2. For this reason many new technologies have been emerging to address the climbing heat loads in the data center. In the lower right corner of Figure 15 is a novel approach by ENP called the "backpack". It is essentially a liquid cooled heat exchanger that cools the exhaust air on the back of the rack so that air leaving the rack is cooled back down to ambient. The advantage of this approach is that condensation is not an issue and the air-cooling is distributed throughout the data center. This will increase the capacity of the data center although capacity values were not available at the time of this report. In the lower left corner of Figure 15 is an M&I Column Fan with a capacity of 300 W/ft^2. This approach still uses under floor cooling which has certain limitations and thus, needs to be evaluated further. The top left image in Figure 15 shows the ENP DataCool system. This innovative ENP system was developed for HP and was tested at HP in a 500ft^2 prototype lab of know server heat loads. The testing demonstrated an unprecedented 500W/ft^2 of capacity and is currently the only technology known by this author that can handle that level of heat load. The upper right corner shows how the DataCool system is implemented. Note that this is a localized overhead-cooling scheme that blows down on computers and sucks up the waste heat. This approach solves many of the problems associated with the under floor cooling. Some of those issues are: flow distribution/capacity, under floor cable blockage and air flow capacity of tiles/grates. As a final note, with densities as high as they are, liquid cooled data centers with liquid cooled computers may be revisited once again.

So the big picture question is how much power does a high-end data center use? If one looks at some of the large Internet Service providers, it is not uncommon to see data centers that are as large as 100,000 ft^2 to 200,000 ft^2. If one does the math and trends continue, a 200,000 ft^2 data center within the next 5 years could require 100 Megawatts of power. To support 100 Megawatts of power a minimum of 60 Megawatts of power would be needed for the supporting mechanical room for a total of 160MW. This is 16% of the output of a typical nuclear power plant. This in itself is a scary proposition but there are other things to consider. First, the electricity cost would be over $100 million per year. Second, the approximate usage of water in the cooling towers would be about 10 million gallons per week, which may be a significant resource issue in some area of the country.

So, are we surprised about the power shortages in California? Figure 16 shows the demand and supply of electricity in the U.S. It should be no surprise to see the difference between the two diminishing. It is speculated that the reason for this is that nobody expected the increase in power caused by the Internet age and compounded by the fact that nobody wanted a power plant in their back yard.

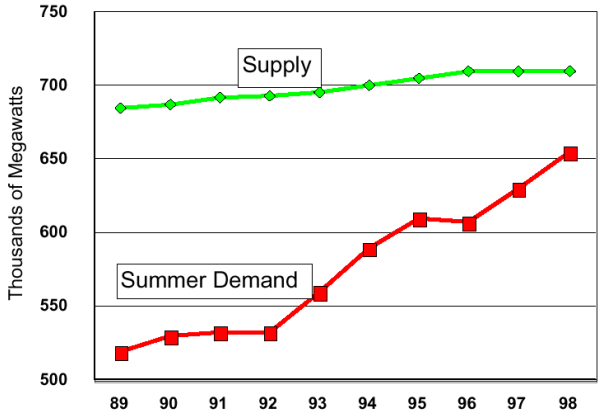

Figure 16. US Electrical and Supply Trends (Courtesy of R. Schmidt)

There has been a lot of debate on what percentage of the electrical supply computers and the internet consume. Kawamoto[6] estimated that in 1999 2% of the grid was used by the industry but the Wall Street Journal[7] estimated 13% in 1999. ITI's projection in Figure 17 corroborates with the latter. Why the difference? The possible reason, Kawamoto only includes power at the plug while the others also include the burden on the power grid and cooling infrastructure. The key thing to note is not the number but the trend, which is growing rapidly.

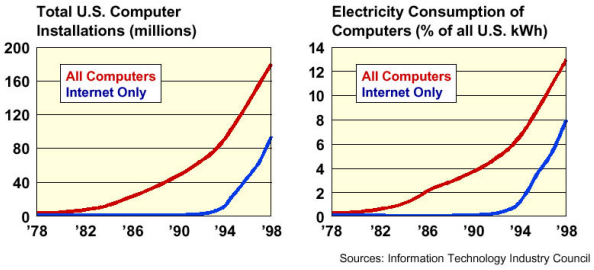

Figure 17. Computer growth and electricity demand (Courtesy of R. Schmidt)

Conclusion

With semiconductor power increasing rapidly, there is no doubt that package level cooling issues are important engineering problems to solve. Some of these techniques were discussed here. Similarly, examples of emerging data center solutions demonstrate the engineering community's ability to solve problems at all levels.

But there is a looming problem ahead of us…With North America using 30% of the world's power for 8% of the population; it becomes clear that as the rest of the world joins the digital economy, the energy demands will accelerate rapidly. Either we let our fate determine its own course or we proactively attack the power issues that we will face. It is obvious that the opportunities in the next century will be in power management and efficiency...let's make it happen!

References

1. Dolbear, T., "Liquid Metal Pastes For Thermal Connections," Proceedings of IEPS International Electronics Packaging Conference, p. 475, Austin, 1992
2. Chu, H., C. Patel, C. Belady, "A Survey of High-Performance, High Aspect Ratio, Air Cooled Heat Sinks," Proceedings of IMAPS International Systems Packaging Symposium, p.71, San Diego, 1999
3. Holman, J.P., "Heat Transfer," McGraw-Hill Inc., Fifth Edition, New York, 1981
4. ClieNT Server NEWS Flash, Issue Number 222.1 News Flash, G-2 Computer Intelligence Inc.,New York, October 20-24 1997
5. "Heat-Density Trends in Data Processing, Computer Systems, and Telecommunications Equipment," The Uptime Institute, http://www.upsite.com/TUIpages/whitepapers/tuiheat1.0.html
6. Kawamoto, K., J. Koomey,. B. Nordman, R. Brown, M. Piette, A. Meier , "Electricity Used by Office Equipment and Network Equipment in the U.S.", ACEEE Summer Study Conference on Energy Efficiency in Buildings, Asilomar, CA, August 2000
7. "Got a Computer? More Power to You," Wall Street Journal, September 7, 2000.