The Sustainable Information Technology Ecosystem

Cullen E. Bash, Chandrakant D. Patel, Amip J. Shah, Ratnesh K. Sharma
Hewlett-Packard Laboratories
Palo Alto, CA
Phone 650.236.2748, Fax 650.857.7029, {firstname.lastname}@hp.com

ABSTRACT

The Information Technology industry, and consumers of the products and services stemming from the industry, has benefited from rapidly increasing computational performance for all classes of devices over the past several decades. With increased performance, the industry has also experienced a sharp increase in device-level energy consumption and power density. These trends are expected to result in a decline of fully passive thermal management solutions in favor of active solutions. However, fully active solutions will further tax energy resources and are therefore not in the best interest of the IT industry nor the consumer. In this paper, we postulate that the management of available energy as a key resource throughout the IT ecosystem, from hand held devices through data centers, will be a requirement from an economic and sustainability standpoint. We propose that hybrid active-passive solutions with rich sensor networks and multi-tiered management systems that can be scaled across the ecosystem, are required to meet future power dissipation needs while also addressing sustainability concerns. The paper will provide such examples that will span the IT spectrum and consider both the supply side and the demand side aspects of the proposed solutions.

Keywords: Data center cooling, dynamic smart cooling, thermal management, sustainability, exergy

NOMENCLATURE

A_d = Available Energy (exergy) Destroyed
COP = Q/W
Ψ = Exergy

INTRODUCTION

The drivers of the next generation of information technology (IT) services are the teeming millions in emerging economies who want to use IT to improve their quality of life and business competitiveness. Indeed, the limitation in physical infrastructure necessitates IT services as a means to improve productivity. The acceptance of mobile telephony and the emergence of innovative applications in countries such as India provides a vivid example of the magnitude of the growth that is possible. The authors have observed examples of a variety of professionals, shopkeepers and semi-skilled workforce, who would like to avail an IT service for $1 to $2 per month using an $80 service oriented client device. We believe that the inflections in compute, networking and storage technologies will enable the necessary price point. However, to achieve this price point and to minimize the energy consumption of the global IT footprint, key thermo-mechanical challenges must be addressed by the packaging, thermal and systems engineering communities. Of particular interest in this paper are the cooling challenges that must be addressed both from a thermal management and energy management point of view along with lifecycle optimization challenges that could significantly impact the footprint.

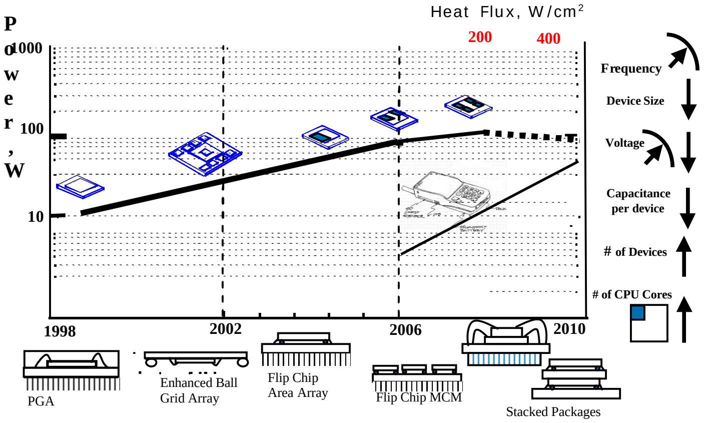


Figure 1. Technology Trends

With reference to Figure 1, we believe the thermo-mechanical challenges arising from the confluence of the key technology advances in semiconductor technologies, chip packaging and optical interconnects will require "active" cooling solutions. Indeed, the passive solutions used today, in devices such as handhelds, will not suffice. The key drivers of this change are the following:

- Integration of multiple functionalities on a single chip and high power density arising from this integration will require active cooling solutions to address local power densities in excess of 4000 W/cm^2 [1].
- Evolution of optical interconnects will necessitate strict temperature control for the optical devices in view of high power density arising from small

surface areas. Small scale input-output drivers will require control of temperature to maintain the integrity of the optical interconnects.

- Opto-electronic devices for high intensity illumination purposes will likewise require heat removal while maintaining a given temperature to enable reliable operation.
- 3d stacked packages will result in difficulty in accessing the heat source embedded within a chip. In this context, it is probably that even a handheld device such as a mobile phone will require an active cooling mechanism to remove the heat from the source.

The growth of active cooling mechanisms will tax the ecosystem made up of billions of handhelds and thousands of data centers. In order to achieve the price point that would result in wide spread use of the IT Ecosystem and have a net positive impact on the environment, we must develop a holistic view. The remainder of this paper will introduce the idea of an integrated IT Ecosystem along with the concepts of supply and demand side management of resources within that ecosystem. Examples will be provided on how to address the expected forthcoming resource challenges from both a supply and demand side perspective.

THE IT ECOSYSTEM

We define the IT Ecosystem as an interconnected system within which computing services are requested and delivered. Components of the ecosystem include any and all items that are required to conduct these service-based transactions including, but not limited to, handhelds (cell phones, PDAs, laptops, etc), desktop computers, in-home networked appliances, networked printers, servers and storage devices, networking gear and data centers. Defining the IT Ecosystem in such a way highlights the interconnectedness and interdependence of the components within the system.

As an example, when a handheld device is used to connect to a web site, portions of the entire IT Ecosystem are required to respond to the request. The request is passed from the handheld device through various layers of networking gear into a server sitting in a data center (or into numerous servers in numerous data centers) and the response is routed back. Each stage of the process consumes energy that, in totality, could significantly exceed the energy consumed by the handheld device alone.

As device energy consumption and power densities increase, the impact on the IT Ecosystem will be profound. Furthermore, the impact will not be limited to servers and data center operators but will effect all users of IT services. Handheld devices, in particular, provide a striking example of the potential issues involved with increased power density. Figure 2 shows an example of a future handheld device as envisioned by the authors. As 3D stacked packaging is utilized in devices like cell phones, package-level power dissipation could increase from less than 1 W today to on the order of multiple Watts. Should this happen, active cooling solutions would be required to remove the heat. Given the space constraints in handhelds, thermo-electric devices (TECs) offer one potential solution path but currently suffer from poor efficiency, with coefficients of performance (COP) around 1. Utilizing such a solution in the example handheld with 5 W of package-level heat dissipation would require an additional 5W to power the TEC, doubling the energy requirements of the device.

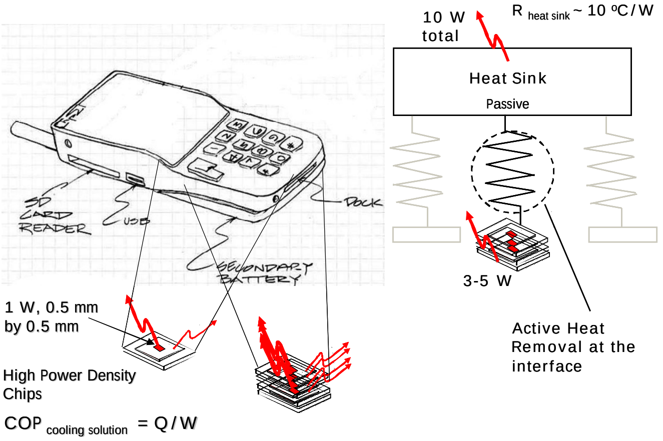


Figure 2: Handhelds with 3D Packaging

The problem can be illustrated further by considering another example of a chip housed in a desktop or server platform as shown in Figure [2]. In this example chip level power dissipation is held to a modest 100W. However, the multi-core architecture results in a single 50 W core with a power density of 200 W/cm^2. A finite element model was built using Mechanica to investigate the impact of this density with a fixed temperature boundary condition at the package lid. The results indicated a temperature rise of 46 C from the lid to the core. If the die is to be held to a maximum of 90 C, the lid, in this case, cannot exceed 44 C which would necessitate an active cooling solution. The result of this is that server level power requirements for the cooling system will increase from 10 to 15% of total system power consumption (fan power only) to over 30% assuming a COP of 3 for the active solution.

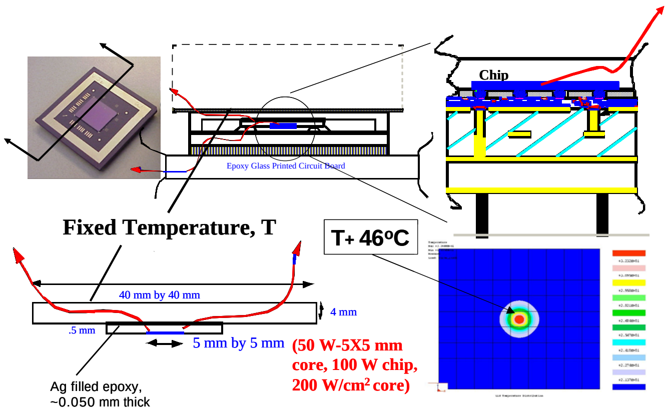


Figure 3: Impact of Chip Power Density

As discussed previously, the impact of increased power dissipation and density will not be felt at the component level alone. Proliferation of IT services will result in proliferation of IT Ecosystem components and, ultimately, data centers. Figure 4 is an estimate by the authors of the energy consumption profile of the IT Ecosystem and the impact of that profile on sustainability – as defined by the annualized destruction of coal in the generation of electricity from the burning of coal and the resulting release of CO_2 to the environment.

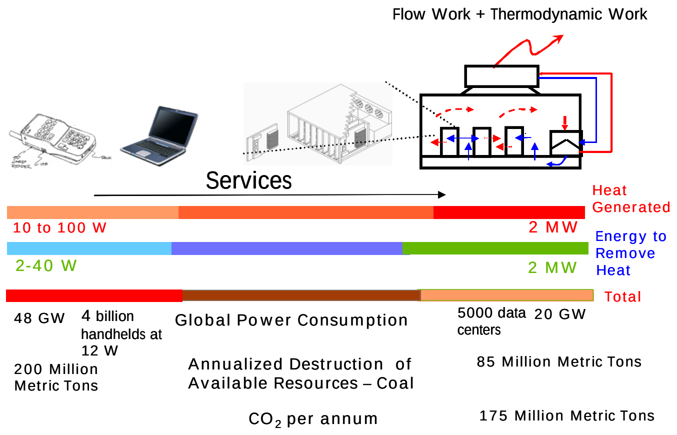


Figure 4: Impact on Global Economics and Sustainability

The figure shows that the energy required to remove the heat generated by components within the ecosystem is a large percentage of the total ecosystem energy requirement and that the primary contributor of this is on the component side of the spectrum. This represents a shift from the current view that most the energy consumption concerns are centered around the data center end of the spectrum [3].

SUSTAINABILITY AND THE IT ECOSYSTEM

The information in Fig. 4 raises two primary issues; a) energy consumption and supply concerns will expand into the future resulting in continuing scrutiny being placed on the IT industry, and b) emissions of greenhouse gases like CO_2 resulting directly from the means of energy generation will further exacerbate global warming with negative impacts on environmental sustainability.

These issues, however, address only part of the problem in that they highlight only energy consumption issues associated with *demand-side* management of resources. An equally important factor in overall IT sustainability is the *supply-side* management of resources. The supply-side includes all aspects of the component lifecycle, from raw material extraction to end-of-life and reclamation - except operations, which is covered by demand-side management - and will be elucidated in the subsequent section.

In recent years, efforts have begun to reduce overall energy consumption, particularly at the data center level [4]. Total cost of ownership models have been introduced that quantify where energy is being consumed in the operational cycle and how to optimize operations accordingly [5][6]. Computer servers have been designed to reduce their energy footprint when idle while research into improved data center cooling solutions has resulted in products that are reducing the energy footprint of large scale facilities [7]. Research has also begun on the movement and placement of IT workload within data centers to optimize the use of cooling resources, resulting in significant improvements in overall data center operational efficiencies [8][9].

Despite improvements at the data center level, and due to the interconnected nature of the IT Ecosystem, more work is needed at the component level on the demand side as well as the supply side. Examples of possible solution paths in each category will be discussed in the following section.

EXAMPLES AND SOLUTIONS

Demand-Driven Resource Management

Adequate provisioning of resources is critical to creating a sustainable IT ecosystem. In the past, resource management has been based on a fixed demand evaluated at design phase. Architectures for delivery of power and extraction of heat have been designed for fixed demand without evaluating the efficiency of the complete process during varying operational states. Cooling resources like fan power and cooling fluid supply are not scaled to the utilization levels of microprocessors or other performance indicators. In most cases demands on resources are not characterized and assumed to be constant or several orders of magnitude higher than normal. The resulting over-design can add to service delivery costs and gate the development of sustainable businesses for IT services.

Key to this problem is the absence of actuation mechanisms in existing thermal management solutions for modulation. In most cases, actuators run full speed during normal operation. Lack of adequate sensor infrastructure also prevents monitoring and aggregation of important data necessary for control. Although utilization levels at the processors, memory, and network and storage devices can vary, the overlying power and cooling architecture is not agile enough to scale commensurately. This lack of scalability leads to deterioration of performance efficiency of such systems at varying utilization levels.

Thermal management mechanisms that can be actively modulated over time and space, like inkjet assisted spray cooling and thermo-electric coolers, among other technologies, offer a unique opportunity to provide a scalable cooling solution that is demand-driven and resource-conscious. In inkjet assisted cooling, fluid can be sprayed in precise volumes in complex spray configurations to dissipate heat at specified locations [10][11][12]. A resource-conscious system can be developed to manage the total fluid delivery cycle from spray to fluid recovery. Figure 5 shows a high momentum spray configuration while figures 6 (a) and (b) show the turn down in flow rates between different operational states of the thermal inkjet. A three orders of magnitude scaling in flow can be achieved while maintaining stable heat transfer performance over small and large areas. A layer of IC sensors provides temperature and flow data. A system of fluidic valves and firing nozzles can be actuated based on built-in logic, triggered by aggregated data from embedded sensors. Optimization algorithms can also be implemented in the hardware to provide operational limits or profiles for control of actuators. These can be tied to the system-level management and the datacenter manager using SLAs or other metrics that capture supply-side constraints at the datacenter level.

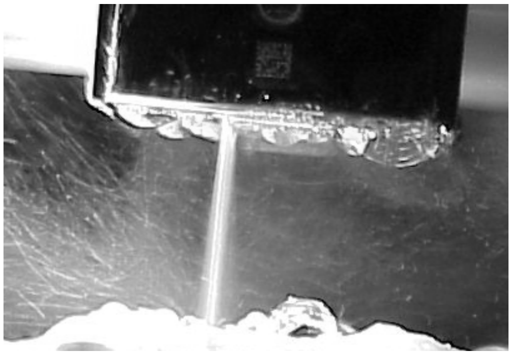
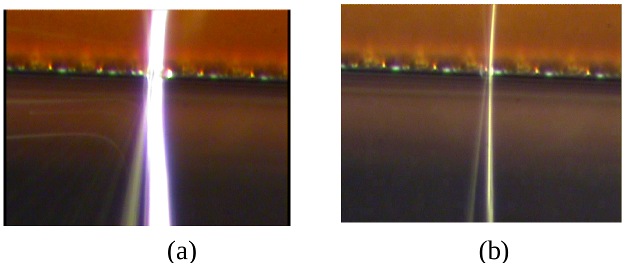


Figure 5: Inkjet Assisted Cooling



(a) (b)

Figure 6: (a) High flow Spray; (b) Low flow Spray

For a system level example, consider the ecosystem made up of billions of handheld devices with 3d stacked packages that require active cooling as described in Fig. 2. Design of an active cooling system for such a handheld device based on fixed demand would result in high energy draw at the ecosystem level and reduction in battery life - by way of example a 1 W active cooling mechanism in each handheld would result in a 1 GW impact at the ecosystem level. A demand based "active-passive" solution that uses a combination of passive phase change cooling and an active cooling solution can address both these issues. As an example, at low power levels, and for short duration conversations, the phase change material in the handheld device can absorb the heat and the active mechanism is triggered only when needed based on a policy subscribed by the user.

At the other end of the spectrum, consider a data center cooling component built with flexibility in cooling – the ability to independently scale flow and temperature. A pervasive sensing and policy based control system can enable demand based provisioning of power, compute and cooling resources throughout the data center space [4][8]. Thus, the ecosystem now becomes demand driven from chips to handhelds to data centers.

Supply-Driven Resource Management

Beyond demand-side management, given the expected proliferation of components in the IT Ecosystem, supply-side management that considers the overall component lifecycle, from inception to reclamation must be examined. Many existing design-for-environment approaches in the IT industry focus on either material management (e.g. by reducing the use of hazardous materials) or energy management (e.g. by reducing the power consumption of system components). However, in these approaches, the optimal combination of a least-materials solution with the least-energy solution is often not obvious: for example, a large microprocessor provides the maximum surface area for heat transfer and therefore may be energetically optimal, but this solution may also deplete the maximum amount of silicon and therefore be materially sub-optimal. A combined supply-side and demand-side approach to resource management is required to minimize the consumption of available resources in the IT Ecosystem.

Previous research suggests that the thermodynamic metric of 'exergy' provides the basis for achieving supply-side management. This concept has successfully been used to simultaneously optimize material and energy flows in power plants [13], chemical manufacturing [14,15], machining [16,17], and manufacturing [18] processes in a supply-side context. More recently, within IT systems, a second-law analysis has been used to examine the operational energy consumption of chips [19] and data centers [20] in a demand-side context.

Thus, as a first step in developing a combined supply- and demand-side framework for next-generation IT products, Hannemann *et al.* [21] have explored an exergy-based life cycle analysis of an enterprise server. The framework is illustrated in Figure 7. First, an inventory is compiled using the production bill-of-materials. The origin of these materials is then traced back to the extraction process, and the successive transformations of the natural resources is then evaluated across resource extraction, manufacturing and assembly, operation, and end-of-life. At each of these stages, basic thermodynamics provides that:

$$A_d = \sum \Psi_{in} - \sum \Psi_{out} - \Delta\Psi \qquad (1)$$

where A represents available energy, Ψ represents exergy (both are used equivalently in this work), and the subscripts are defined as follows: d = destroyed, in = input to the system, out = output from the system. The last term represents the change of exergy within the system. The magnitude of exergy destroyed within each stage provides an estimate of the theoretical potential for improvement in a given phase, since a reversible (sustainable) process would not destroy any exergy. Having applied the formulation of Eq. (1) across all the stages of the product life-cycle, the total exergy loss of the product can be estimated as the sum of exergy losses within each stage.

To illustrate the relevance of such a framework, consider a hypothetical laptop which consumes 75 W of power. Assuming an average of 8 hours per day peak use at an average 65% of peak power consumption over a 3-year lifetime, the total electricity consumed during product use would be around 430 kWh. Demand-side management approaches, such as reduced power states of the processor or more efficient power supplies, could help reduce this operational energy consumption. However, significant amounts of energy and material are also consumed during the extraction, fabrication and disposal of the laptop. By applying the above lifetime exergy framework, we find that the material extraction consumes about 400 kWh of available energy – almost the same amount as electricity consumed during the entire operating lifetime of the system. Further, Figure 8 discusses the material flows associated with a hypothetical laptop computer. Even though aluminum constitutes less than 25% of the total material in the laptop, more than 60% of the exergy costs encountered upstream of use are associated with the extraction of this material. Thus, reducing the amount of aluminum in the product by half could potentially reduce the destruction of available resources by as much as a new power-saving feature on the laptop that reduces the duty cycle by 20%. The optimal system design would require a combination of both approaches.

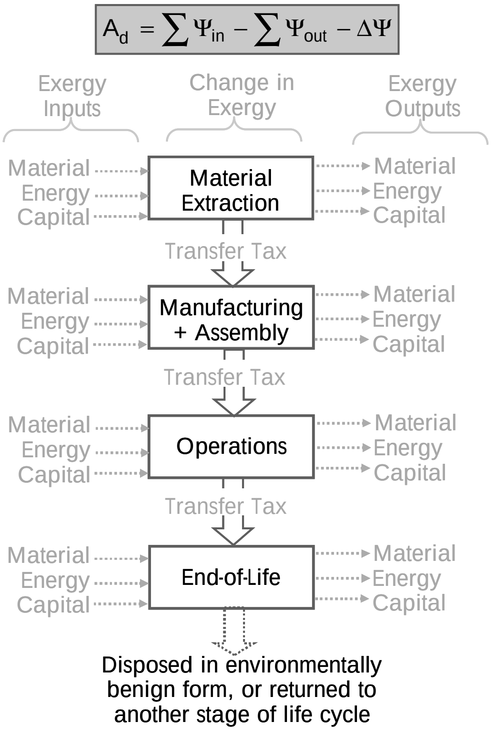


Figure 7. Framework for assessing the exergy loss across the life-cycle of a product. The 'transfer tax' implies a penalty incurred while shifting resources from one stage to another, e.g. the exergy destroyed during transportation.

Thus, the type of insights derived from lifetime exergy analysis will be critical to maximizing the sustainability of the IT ecosystem.

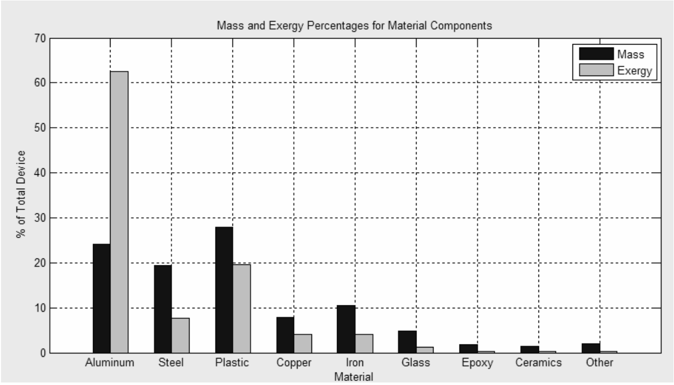


Figure 8. Exergy Loss associated with Material Flows in a Sample Laptop. (Figure by C. Hannemann, UC Berkeley)

SUMMARY AND NEXT STEPS

The confluence of user demand for IT services in emerging markets, increasing chip-level power densities resulting from multi-functional processor technology and multi-core integration, 3D stacked packaging integrated into handhelds that will push the limits of power dissipation and accessibility, and energy demand increases at the appliance to data center level, is creating the makings of a "perfect storm" that has the potential to drive the emerging IT Ecosystem into recession. Demand and supply side management of critical resources, from electricity usage to raw material and end of life disposal and reclamation processes will be a critical aspect of insuring the sustainability and longevity of the future IT Ecosystem. In this paper we address the need for demand-side resource management practices that scale with the needs of the IT components and thereby reduce overall energy demand during the operational portion of the component lifecycle. We also suggest that the supply-side aspect of the lifecycle that governs material extraction through reclamation can have as large a resource impact as the operational side, indicating optimization should not be limited to operations only.

The making of the aforementioned perfect storm by virtue of this confluence of technologies requires the chip, packaging and systems community to develop a library of active-passive cooling solutions. Such a foundation of solutions can be overlaid with sensing networks and policy based control systems to modulate resources as needed and, along with supply side optimization processes, ensure the longevity of the IT Ecosystem.

REFERENCES

[1] Huddle, et. Al., "Thermal Management of Diode Laser Arrays", Sixteenth IEEE SEMI-THERM Symposium, 2000.

[2] Bash, C., Patel, C., Beitelmal, M., Burr, R., “Acoustic Compression for the Thermal Management of Multi-Load Electronic Systems”, ITHERM, San Diego, CA, 2002.

[3] C. Patel, et al., “Energy Flow in the Information Technology Stack: Introducing the Coefficient of Performance of the Ensemble,” Paper No. IMECE2006-14830, Proc. ASME International Mechanical Engineering Congress and Exposition, Chicago, IL, November 2006.

[4] Patel, C., Bash, C., Sharma, R., Beitelmal, M., Friedrich, R., “Smart Cooling of Data Centers”, Proceedings of IPACK'03 International Electronic Packaging Technical Conference and Exhibition, Maui, Hawaii, July 6-11, 2003

[5] C. Patel, et al., “Energy Flow in the Information Technology Stack: Coefficient of Performance of the Ensemble and its Impact on the Total Cost of Ownership,” Technical Report No. HPL-2006-55, Hewlett Packard Laboratories, Palo Alto, CA, March 2006.

[6] Patel, C.D., Shah, A., “Cost Model for Planning, Development and Operation of a Data Center”, HPL Technical Report, HPL-2005-107(R.1), 2005.

[7] Bash, C.E., Patel, C.D., Sharma, R.K., “Dynamic Thermal Management of Air Cooled Data Centers”, Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, San Diego, CA, 2006

[8] Bash, C., Forman, G., “Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-Efficient Locations in the Data Center”, USENIX Annual Technical Conference, San Jose, CA, 2007.

[9] Bash, C., Forman, G., “Data Center Workload Placement for Energy Efficiency”, Proceedings of IPACK2007 ASME InterPACK '07, Vancouver, British Columbia, CANADA, July 8-12, 2007

[10] Bash, C.E., Patel, C.D. and Sharma, R.K., 2003, “Inkjet Assisted spray cooling of electronics”, Proc.IPACK’03 – The PacificRim/ASME Intl. Elect. Pack. Tech. Conf. and Exhibit., ASME IPACK-35058, Maui, HI, July 2003

[11] Sharma, R., Bash, C., Patel, C., 2004, “Experimental investigation of heat transfer characteristics of inkjet assisted spray cooling”,, Proc. 2004 ASME Heat Transfer/Fluids Eng.Conference, HT-FED2004-56183, Charlotte, NC, July 11-15

[12] Escobar-Vargas, et. al., 2007, "High Power density dissipations by spray cooling", Proc. 2007 ASME-JSME Thermal Eng. Summer Heat Transfer Conf.,HT2007-32442, Vancouver, BC, July 8-12 2007

[13] J. Szargut, D.R. Morris, F.R. Steward, Exergy Analysis of Thermal, Chemical and Metallurgical Processes, Hemisphere, New York, 1988.

[14] M. Sorin et al., “Exergy Flows in Chemical Reactors,” Proc. Institute of Chemical Engineers, vol. 76, pp. 389-395, March 1998.

[15] D.R. Morris, “Exergy Analysis and Cumulative Exergy Consumption of Complex Chemical Processes: The Industrial Chlor-Alkali Processes,” Chemical Engineering and Science, vol. 46, no. 2, pp. 459-465, 1991.

[16] J.C. Creyts, V.P. Carey, “Use of Extended Exergy Analysis as a Tool for Assessment of the Environmental Impact of Industrial Processes,” Proc. ASME International Mechanical Engineering Congress and Exposition (IMECE), Dallas, TX, November 1997.

[17] J. Creyts, V. P. Carey, “Use of Extended Exergy Analysis to Evaluate the Environmental Performance of Machining Processes,” Proc. Institute of Mechanical Engineers, vol. 213, no. 4, pp. 247-264, 1999.

[18] T. Gutowski et al., “A Thermodynamic Characterization of Manufacturing Processes,” Proc. IEEE Intl. Symp. On Electronics and the Environment (ISEE), May 2007, pp. 137-142.

[19] A. Shah et al., “An Exergy-Based Figure-of-Merit for Electronic Packages,” ASME J. Electronic Packaging, vol. 127, no. 4, pp. 452-459, 2006.

[20] A. Shah et al., “Exergy Analysis of Data Center Thermal Management Systems,” ASME J. Heat Transfer (in press).

[21] Hannemann et al., “Exergy-Based Life Cycle Assessment of Computer Server,” IEEE Intl. Symp. on Electronics and the Environment (ISEE), May 2008 (submitted).