The International Symposium on Micro-Mechanical Engineering
December 1-3, 2003

ISMME2003-K15

A VISION OF ENERGY AWARE COMPUTING FROM CHIPS TO DATA CENTERS

Chandrakant D. Patel

Hewlett Packard Laboratories, 1501 Page Mill Road, Palo Alto, California 94304-1126, U.S.A
chandrakant.patel@hp.com

The miniaturization of silicon devices, and the integration of functionalities on a single chip, has resulted in high power density chips, systems and data centers. The increase in power density in all these three areas necessitates a holistic examination – following the path of the heat flux from the chip, through the system enclosure to the room and out to the environment. Furthermore, computing has become pervasive and will soon account for a large portion of global energy use, particularly with respect to distribution of high power data centers around the world. In this context, future thermo-mechanical solutions have two clear objectives - to facilitate effective heat transfer from high power density chips and systems in order to maintain specified temperature on the device, and to facilitate the heat removal efficiently by minimizing the energy used to remove the dissipated heat. Energy management plays a lead role in data centers - machine rooms that aggregate hundreds of computers to provide useful computing services and can reach 10 MW of power dissipation from the hardware. In high power density chips, heat transfer solutions that maintain specified chip temperature while minimizing the energy used to affect the thermal management play a central role. This paper examines an energy aware thermal management approach from chips to data centers and proposes second law analysis as a measure of overall management of energy consumption.

Keywords*: energy, efficiency, thermal management, cooling, micro-mechanical, data center*

INTRODUCTION

The rapid growth in information technology services fueled by low cost standardized computing hardware will result in a computing utility with billions of users and trillions of services [1][2]. The services will be delivered from data centers that will house compute, storage and networking hardware. The hardware will be composed of compact, highly integrated, low cost, high power density components. The compact high performance hardware will enable unparalleled performance per unit area. It will also result in very high power density installations with power density per unit area increasing from current 500 W/m^2 to 3000 W/m^2. These high power density data centers will pose a tremendous challenge in thermal and energy management. This paper is a holistic examination of the thermal path from the key sources of heat generation – microprocessor, memory and input-output devices through the computer system and the data center to the environment.

The microprocessor of today is a highly integrated device with one or more functional units and a large memory cache. The next five years will enable much greater integration of functionalities – multiple central processing unit (CPU) cores, memory controller, network controller, graphics controller, several megabytes of cache – all on a single 20 mm by 20 mm chip. The power density from a CPU core, occupying fraction of the 20 mm square chip, will reach 200 W/cm^2. The power dissipation from all the building blocks on the chip will aggregate to approximately 125 W. While the total power of 125 W from the 20 mm by 20 mm chip is of the same order of magnitude as high performance chips of the mid 1990s, the power density based on localized heat load is two orders of magnitude higher than the chips of that era.

A computer system will contain several of these highly integrated microprocessors, additional memory, input-output (I/O) devices, mass storage and power supply. Such a computer system, occupying a volume of approximately 425 mm wide by 44 mm high by 700 mm deep, will dissipate 400 W. Other types of computer system form factors will include highly dense “blade” type enclosures. These enclosures will contain multiple microprocessor boards and I/O boards, or “blades”, connected with a standardized physical connector. As an example, a single processor “blade” assembly, occupying a volume of approximately 308 mm by 400 mm by 44 mm, will contain multiple microprocessors and memory devices and dissipate 250 W. Useful services will be

provided by aggregation of thousands of blades in enclosures housed in data centers.

A modern data center consists of multiple racks of compute, networking and storage hardware. Standard 2 meter high EIA (Electronics Industries Association) racks, arranged in rows at a pitch of approximately 2 meters, occupy the data center. The maximum power dissipation from a high performance, fully utilized, rack stands at approximately 10 KW. While the racks in most of current data centers are not fully utilized, the next five years will result in consolidation of data centers and full utilization of rack space across a given data center. The consolidation will be driven by savings in real estate, operation and capital costs. This compaction and consolidation will enable a given 2754 m^2 (30,000 ft^2) data center to house approximately thousand racks of compute, networking and storage hardware. The power delivered to the hardware will aggregate to 10 MW – all dissipated as heat. Finally, the delivery of envisioned services – computing and information on demand to billions of users – will result in a global distribution of such high power data centers[1][2]. Thus, from a thermo-fluids perspective, the control volume will now span geographic distribution of data centers around the world. With respect to energy aware global computing, the reservoir for heat rejection - the external environs - and internal thermo-fluids behavior in a data center will play a critical role[3][4].

NOMENCLATURE

P: Pressure, Pa
T: Temperature, ^oC
m: Fluid mass flow rate, kg/s
Q : Rate of heat exchange, W
W: Characterized compute workload; assumed synonymous with heat load in this paper

Subscripts

in: inlet
out: outlet
i,j: i^{th} row, j^{th} rack
k: index of geographic physical location of a data center
l: l^{th} system
o: o^{th} board (processor, I/O, etc) in a system
region-p: index of region on a chip

Need for Energy Aware Computing and Thermal Management

The emergence of the global compute utility, and the ensuing high power needs of the utility will require energy aware "smart" design approaches at all levels – specifically with regard to dissipation of power by electronic equipment and in limiting power used to facilitate the heat removal. Indeed, government regulations such as Japan's top runner program [4] will require such heed in design of systems. According to this program, energy consumption in buildings has to be reduced by 10% by 2008. In addition to this, reductions in industrial energy consumption will lead to a total reduction of energy consumption by 14%. Furthermore, an improvement in computer energy efficiency of 83% by 2005 (base year of 1997) is sought[4].

The state of the art methods in heat removal at all levels - chip, system and data center are based on maximum power dissipation and on maintenance of a fixed temperature at a given location. There is little variability in the cooling resources, and the power consumed by the cooling resources remains at peak levels regardless of the state of the chip, system or data center. As an example, design of thermal management solution for chips is based on maximum power dissipation, say 100 W, and maintenance of 85 ^oC at a specified location on the chip, irrespective of the state of the chip e.g. idle, busy, etc. The thermal management solution consumes power at fixed level irrespective of the state of the chip. This results in gross over-provisioning for high power density chip floor plans and waste of energy. There is a need for scaling and provisioning of cooling resources with respect to power dissipation from the chip. Moreover, there is a need to devise much more granular "fine tuned" provisioning of cooling based on heat dissipation map on a chip; and on managing the power dissipation map based on most efficient cooling configuration.

Similarly, for systems used in data centers such as industry standard server platforms, the heat removal techniques are based on maximum power dissipation. The systems are forced air cooled with air movers that provide adequate mass flow to keep the system components at a given temperature. The air movers consume 10% of maximum power – 1.0 KW in a 10 KW rack – largely as flow work. In non optimized systems, with improper matching of efficiency curves, the power consumed by the air movers can be higher. Speed control of air movers is applied in a rudimentary way by sensing a temperature at a single ad hoc location in a system. However, this level of sensing does not have the granularity for management of energy used by the cooling resources. A more local and global understanding through distributed sensing and control is needed. A "fine tuned" provisioning and scaling of power used by the cooling resources based on heat dissipation level of a system is needed.

Furthermore, the state of art computer systems do not have the ability to set various power dissipation levels e.g. multiple power states by scaling the microprocessor and other component power. Current systems operate at high fixed power dissipation level while "busy" or "idle". The typical range of power between these two states is maximum at busy to approximately 60% of maximum at idle. Systems with multiple power states that substantially

lower the power dissipation when idle, or when executing low performance workload, are needed. The availability of multiple power states for microprocessors, and systems, is inevitable [6][7]. Future "smart" design should have the ability to scale power based on the type of compute workload, priority and availability of cooling resources. Such component level energy scale down has been shown to be effective in the context of battery life of mobile devices[8]. The ability to scale power, through voltage and frequency scaling for a microprocessor, can be a great advantage in devising energy efficient utilization of cooling resources. It enables energy aware design where combination of balanced cooling with respect to the heat load can be achieved by scaling of power in systems based on efficient availability of cooling resources, and provisioning of cooling resources based on dynamic and precise demands of systems.

Air conditioners are, by far, the largest consumer of power for cooling purposes in a data center - for every 1 W removed, an additional 0.5 W is used by the cooling equipment[9]. The power consumed is composed of flow and thermodynamic work. Flow work is the work performed in moving the fluid in the data center and through external "roof-top" heat exchanger or cooling tower. Thermodynamic work is work performed in extracting the heat and reducing the temperature of exhaust air from the computers. The immense consumption of power by cooling resources, 5 MW for a 10 MW compute installation, requires an energy aware data center design. Energy has to be managed as a resource in a data center. Thus, there is a need for a global management system that dynamically deploys the cooling resources in the data center based on the heat load distribution, and deploys the heat loads or compute workloads based on the most energy efficient cooling configuration in the room. The complex thermo-fluids behavior in a high power density data center necessitates creation of "thermo-fluids" policies that can be used by the data center management system to allocate compute workloads and dynamically provision the cooling resources[7]. The systems not in use are turned "off' or otherwise scaled back.

From a thermodynamics perspective, heat dissipation and energy efficiency can be easily optimized for an isolated system. However, it is difficult to optimize these for open systems, where mixing of cooling streams and heat sources at different temperatures complicate the heat transfer and fluid mechanics. Such scenarios occur more often than not, from chip-scale to data center level. Applying closed system methods to an open system design leads to over sizing of cooling capacity and affects control of cooling resources. A second law approach based on exergy (essergy) or available energy is required to analyze and optimize these systems. The exergy approach, based on minimization of irreversibilities, will depend on the distributed sensing and control of the relevant infrastructure to optimize heat dissipation and energy or exergy efficiency. Prior work has shown that this approach can provide a significant advantage for enabling energy aware cooling[10][3].

Thus, a philosophy of dynamic balancing of cooling resources and heat loads, using the following salient points, is needed:

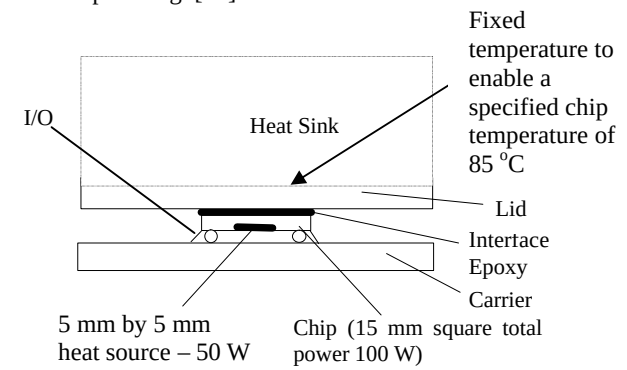
- Heat removal technique for chips and systems that has the flexibility to adapt to the real time heat flux distribution and afford the highest efficiency for the range of power dissipated. The flexibility and adaptability is exploited through a distributed sensing and control system that dynamically provisions the cooling resources for varying heat dissipation levels – operating at all times at levels that minimize energy use. Lastly, a thermal management solution optimized for nominal heat dissipation, rather than worst case heat dissipation. When demand exceeds available cooling (maximum power from all components) workloads are shifted and/or systems shut down.
- A global assessment of energy use in data centers, and placing workloads and using resources at locations that offers the best energy efficiency at a given time. Thus, the workload placement decision would be a function of data center thermo-fluids coefficient comprised of efficiency in internal thermo-fluids behavior and efficiency of the vapor compression cycle based on the heat rejection temperature at a given geographic location. Similarly, within a particular data center, distribution of compute workloads to various computer systems would be based on most energy efficient cooling configuration. The cooling resources in the data center would be provisioned based on the heat load distribution[11]. Systems not in use are turned off or scaled down in power dissipation.
- An evaluation and control mechanism based on development of dimensionless parameters for a unified approach to scale-up, scale-down of cooling design and control of cooling resources[12]. Development of an evaluation technique using the second law of thermodynamics with the intent of identifying and minimizing irreversibilities in the heat removal system and determination of overall exergy efficiency.

ENERGY AWARE DESIGNS FROM CHIPS TO DATA CENTERS

Energy Aware Chip Design

Figure 1 shows a simplistic example of a microprocessor organization, one containing a 50 W, 5 mm by 5 mm CPU core. The balance of the chip is assumed to contain memory and other functionalities that add up to 50 W for a chip total power dissipation of 100 W. A flip chip single chip package is assumed. The

network of vias on the I/O side, and the poor thermal conductivity of the chip carrier assumed to be laminate or low thermal conductivity ceramic, results in a very high thermal resistance path for the heat flux. Thus, the main thermal path is assumed to be from the non I/O side of the chip, through an epoxy or thermal compound interface layer, to the package lid. A finite element model, with a constant temperature on the flip chip package lid shows a temperature rise of approximately 46 ^{o}C for the example design[13].



Interface Epoxy: 1.5 W/m-^{o}C thermal conductivity, 0.025 mm thick
Thermal Conductivity of Lid: 210 W/m-^{o}C
Ambient of 40 ^{o}C

surface at a fixed temperature – modeling boundary condition

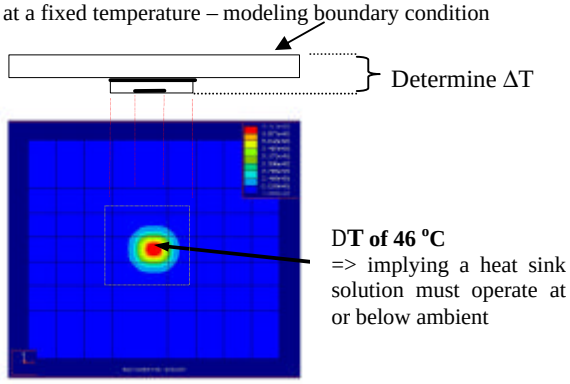
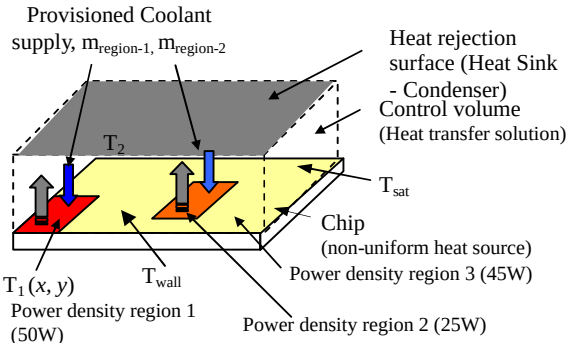


Figure 1. Chip to Heat Sink Temperature Gradient

The same design, with uniform heat flux of 100 W over the entire chip surface area of 20 mm by 20 mm, shows a temperature rise of approximately 7 ^{o}C. As the allowable overall recommended temperature rise is approximately 45 ^{o}C (85 ^{o}C max on the chip for an ambient of 40 ^{o}C) from the chip to ambient, there is no margin left for the heat sink solution. Indeed, failing any marked improvement in the interface and silicon conductivity, an active cooling solution that results in the heat sink operating at or below ambient is sought. A detailed examination of non uniform power distribution on the chip has been investigated [14]. The result of the work suggests the inevitability of liquid cooling with sensible heat gain and other active cooling alternatives in computer applications. Indeed, there are many innovative solutions that strive to reduce the temperature of the heat sink by active cooling means[15][16]. However, in all cases the approach is to design for maximum power, and drive down the heat sink or package (lid) temperature low enough to account for the temperature rise due to highly integrated, compact microprocessor. Furthermore, the energy used by the active cooling solution remains unchanged irrespective of the state of the chip e.g. idle, busy, etc.

Challenge to the Micro-mechanical Community

The challenge to the micro-mechanical community is to devise a solution that dynamically adapts to the non uniform heat distribution on the chip. The solution provisions the coolant supply commensurate with the heat load distribution. As an example, for a two phase solution, the implication is that the mass flow is tuned to the dynamic and non uniform heat load – just sufficient to utilize the latent heat of vaporization – grams to micrograms of flow per second based on regional power distribution. Thus, in such instance, an energy aware design is one that is tuned to the dynamic heat distribution on the chip and is able to regulate the energy used to move the heat e.g. turning “off” or changing flow rate based on the heat load. Furthermore, the phase change phenomena, from an evaporative film boiling perspective, results in minimal excess temperature rise – excess temperature being the difference between the chip wall and the saturation temperature of the fluid (T_{wall} – T_{sat}) for a given ambient pressure in the vessel. Thus, for the chip shown in fig. 2, the dynamic provisioning of mass flows, $m_{region\text{-}p}$, based on the dynamic power map will result in excess temperature across the chip, in all regions, to remain at a constant level.



Regions on the chip are various functionalities such as multiple CPU cores, memory, etc

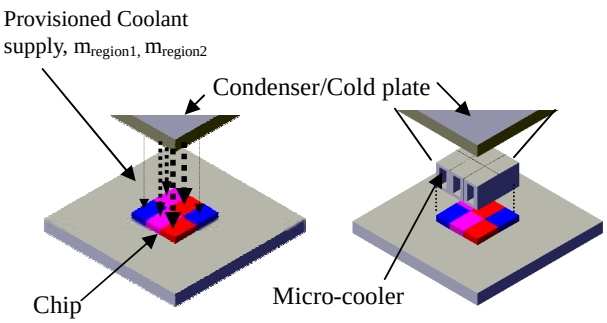
Figure 2. Power and cooling distribution on the chip

Minimization of excess temperature also leads to reduced irreversibilities and is an energy efficient phenomena. For the chip represented in fig. 2, regions 1, 2 and 3 have heat dissipation of 50, 25 and 45W, respectively, amounting to a total power dissipation of 120W. Regions 1 and 2 are 5mm sq. while the chip is 20mm sq. Considering the phase change fluid to be Fluorinert FC72, the mass flow rate distribution for regions 1, 2 and 3 are 2.3, 1.2 and 0.15 gm/s.cm^2, respectively. An example of such implementation is the

use of printer ink jet heads for precision dynamic spray cooling[17]. Bash et. al discuss preliminary results obtained from an ink jet assisted spray mechanism with a 512 nozzle ink jet head. The extension of this approach, using a non-uniform spray, with a given nozzle mass flow, $m_{region-p,}$ enables a "tuned" two phase system. Furthermore, applying a closed loop system to the ink jet cooler – one that follows the variable chip heat load is an example of an energy aware two phase cooler.

The heat rejection surface, the condenser, is maintained at a condensing temperature (T_2) below the saturation temperature (T_{sat}). It is designed for overall heat load of 120 W for the chip shown in fig. 2. While, it does not have to reflect the non uniformity of power distribution on the chip, it has to take into account the total power change from zero to 120 W and be energy aware. As an example, the air movers on the heat rejection surface have a variable speed fan that can be shut off, or speed reduced when the heat load drops. This combination of a micromechanical phase change mechanism that modulates the mass flow being applied to the chip based on the dynamic power distribution, and a heat rejecting system which modulates use of energy based on the total heat load, is an example of energy aware and tuned cooling system.

In addition to provisioning of cooling resources with reference to the heat load, a microprocessor design that allocates the heat dissipation by setting various power and performance states based on most efficient and available cooling resources is needed. A chip design with power considerations is shown by Kumar et. al [18]. A multi CPU core microprocessor design, with an architecture that operates various cores at variable power or performance levels based on the available cooling resources, is an example of an energy aware chip design of the future (fig. 3). Indeed, the power is reduced locally if the cooling resource enters a non-efficient region of operation and the service level agreement (SLA) with the user allows the reduction.



a. Tuned 2-phase cooler with variable fluid flow rate

b. Tuned solid state micro-cooler

Colors are varying power/ performance levels of various functionalities based on most efficient available cooling

Figure 3. Tunable Chip Cooling Mechanisms

In the context of micromechanical design, the creation of microscale refrigeration systems, based on thin-film thermo ionic emission, that is tunable to the power dissipation profile can result in efficient solid state active cooling system[19]. The inverse relationship between length of thermo-element and cooling power density makes it an ideal candidate for integrated microscale chip cooling. As an example, the segmented or segregated micro cooling devices can be constructed for interfacing with silicon or integrated in silicon. In essence, these micro-refrigerators will have multiple solid state thermo-ionic "evaporators" that can be provisioned based on the segregated heat loads. In operation, the reduction in temperature on the "evaporator" side can be based on the heat dissipated in each area of the chip – the regions shown in Figure 2. The power consumed in driving the cold side of the solid state devices to a given temperature, based on regional power dissipation, will result in a tuned local coefficient of performance (COP). The COP being the ratio of heat removed to the energy used by each of the segregated microcooler with respect to localized heat load. Furthermore, as with the two phase approach, the heat load dissipated by the microprocessor itself can work hand in hand with the cooling system i.e. the chip architectural design can scale power if the cooling system cannot maintain a high threshold COP.

Energy Aware System Design

System design should enable an effective heat transfer solution by efficient transport of rejected heat to the external surroundings. Figure 4 shows an example of a typical high density "blade" type system in the market. Multiple single board processor blades and I/O blades are mounted to a backplane. The air movers in the system provide adequate mass flow for a given temperature rise, typically 15 ^{o}C, across the system based on maximum power dissipation from all the blades. The key objective

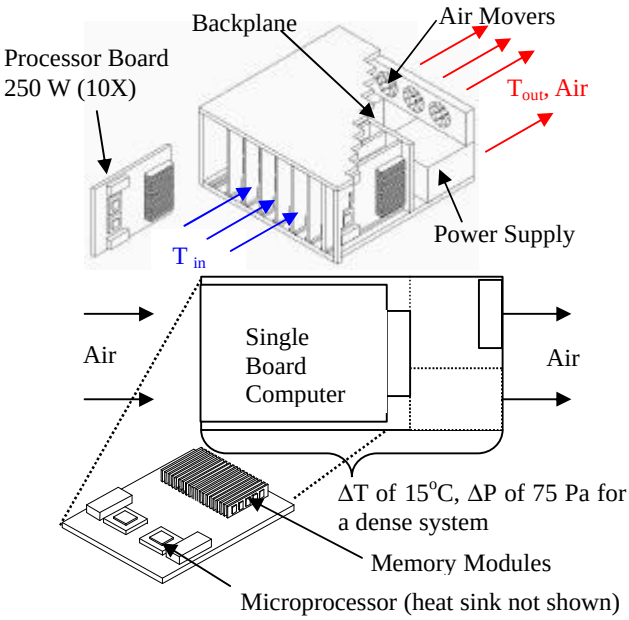


Figure 4. High Density System Enclosure

is to maintain the microprocessor heat sinks such that the CPU core is at specified temperature of 85 ^{o}C to 90 ^{o}C. Other compact systems, especially 1U (44 mm high rack mounted) servers have a similar design philosophy. In general, all slim rack mounted server designs have a high flow resistance resulting from compact design. The air movers have to sustain a high mass flow at a significant pressure drop to enable removal of heat from high power density components with high fin density heat sinks. For the system shown in fig. 4, the flow work performed by the air movers, 0.150 m^3/s at about 75 Pa, is approximately 11 W for a dense 2.5 KW system at sea level. This requirement remains unchanged as the system is not fine tuned to local generation of heat. There is rudimentary speed control of air movers based on a single point temperature measurement, typically the lid of the microprocessor package, but this is not sufficient. Further, in many cases air movers are not optimally sized and the "wire to air efficiency" is very poor. Accounting for overall efficiency, the energy required by air movers could be 110 W for the 2.5 KW system (10% wire to air efficiency applied to the 11W flow work calculated earlier).

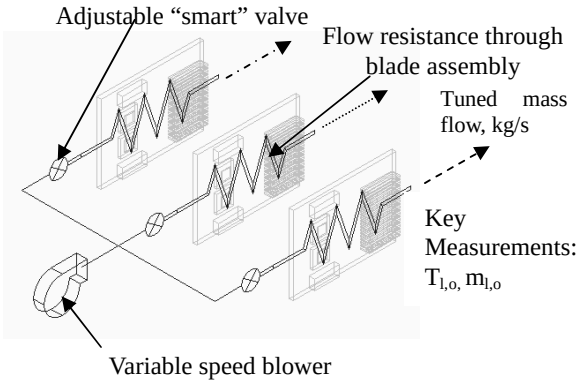


Figure 5. Energy Aware Blades Cooling

Figure 5 is an example of a tuned and variable energy option. The key differences in this approach are the following:

a. The air movers are selected and optimized for nominal flow rate and pressure drop and have the ability to vary speed.
b. The inlet and outlet temperatures in each section, $T_{l,o}$, are measured and used to determine the caloric content using the energy equation – product of mass flow, specific heat capacity and temperature rise. The blower speed is modulated based on the caloric content of each section to produce a mass flow, $m_{l,o.}$
c. The computer architecturally contains a feature in its operating system that contains "smart" power management that instructs the thermal management system about the computer' s state e.g. maximum performance state or low performance state. The "smart" power option in the operating system also modulates the performance and workload characteristics, $W_{l,o}$ based on the most energy efficient mass flow, $m_{l,o,}$ available in each section.
d. "Smart" valves meter the flow; e.g. shut the air flow into compartments that do not have any compute workloads. Furthermore, the compute workloads, $W_{l,o,}$ are assigned based on most efficient cooling available in a given compartment.
e. If all computers in the enclosure are required to operate at maximum heat dissipation level, and the blower cannot maintain adequate mass flow levels, some of the systems are shutdown and the workload migrated to other systems in the data center [7].

The example shown above can also be scaled to a rack design. The same principle, that of designing for nominal heat dissipation, and scaling the mass flow based on the caloric content can be applied. This design philosophy of tuned provisioning of cooling resources based on the heat load, and provisioning of the workloads, $W_{j,l,o}$,based on the most efficient availability of cooling resources also applies for racks in a data center.

Energy Aware Data Center Design

The thermo-mechanically complex data center of tomorrow cannot be designed using current modus operandi - by use of energy balance in sizing of air conditioning resources and by intuitive distribution of air flow[3][20][21]. Indeed, *the data center is a computer*, with its walls akin to the walls of a computer enclosure. The current design approach based on maximum power dissipation by all the compute, networking and storage elements and manual qualitative setting of the air conditioning units leads to mal-provisioning of the cooling resources[3]. In today's data center the air conditioning (AC) unit's "setpoint" is manually set to approximately 18 ^{o}C - the "setpoint" is defined as the temperature of the return air from the data center. This is treatment of the data center as a home – striving for a uniform temperature by promoting mixing of air streams. An energy efficient data center is one where the cold inlet air to all the systems is maintained at a specified temperature, typically 25 ^{o}C, and the exhaust hot air at 40 ^{o}C[3][10]. The exhaust air is prevented from mixing with incoming cold air and is driven back to the AC units. This fluidic separation of hot and cold streams is a necessary condition for efficient use of energy and proper provisioning of AC units. Proper provisioning of the AC units is achieved by three dimensional computational fluid dynamics modeling[3][10][20]. The air conditioning resources are set, by virtue of vent tile openings, and other variable settings, to deliver proper mass flow for a given geometric distribution of heat loads. Alternatively, or in combination, the heat loads are geometrically distributed to enable proper provisioning of the AC units. Data center design is evaluated using data center indices called supply heat index (SHI) and return heat index (RHI)[12]. The indices, a measure of mixing of hot and cold air streams and provisioning of AC resources, are

based on critical dimensionless parameters e.g. ratio of height of the rack to room height, etc[12]. An analysis that enables provisioning of the AC resources for fixed heat loads is called "static" provisioning.

Beyond "static" provisioning, Patel et. al propose a "smart" data center - a vision of an intelligent data center that dynamically apportions the cooling resources based on the demand and places workloads based on the most efficient use of cooling resources[11]. Furthermore, this "smart" data center is envisaged as one where characterized compute workloads, $W_{i,j,l,o}$ are placed based based on the most energy efficient availability of cooling resources in the data center. The compute resources not in use are put on "standby" or a low power state. The "smart" data center operates through a pervasive sensing layer – a network of hundreds of temperature sensors at the inlet and outlet of the servers in the racks ($T_{i,j,l,o}$) – the data from which is used with high level policies such as SHI [12] to control the variable air conditioning resources. Sensing also includes measurement of pressure in the plenum and power to the servers. The variable air conditioning resources are variable valves for chilled water control, variable or "smart" vents, variable air movers in AC units and variable compressors as shown in fig. 6 [11].

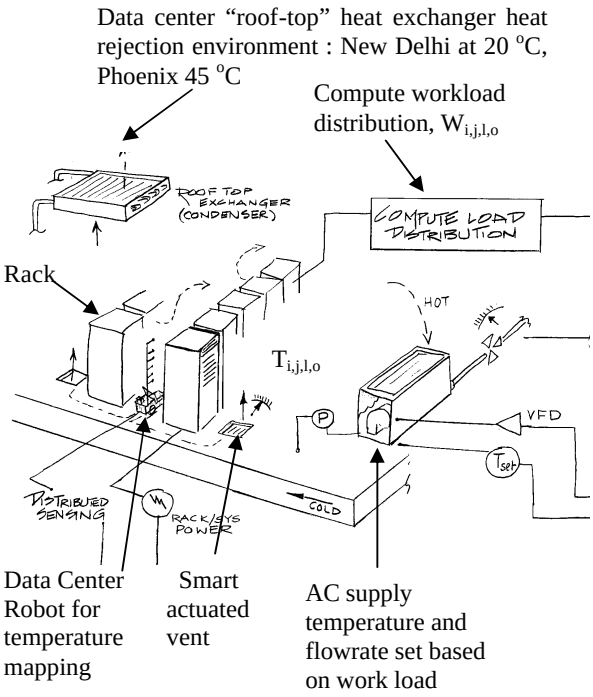


Figure 6. Smart Data Center [11]

A data center management system, based on high level thermo-fluids policies, enables the automated dynamic provisioning of air conditioning resources and distribution of the compute workloads for power management. Thus, the "smart" data center manages energy as a critical resource and maintains the data center in a provisioned state completely in balance with the heat loads. The energy manager is global and places the workloads, $W_{k,i,j,l,o}$ based on the internal data center thermo-fluids characteristics and the external heat regional rejection environment e.g. a lower lift from the evaporator to condenser exists in New Delhi, India at night relative to Phoenix, Arizona day time temperature of 45 °C resulting in a 60% improvement in cycle efficiency [4].

GLOBAL EVALUATION CRITERIA

Exergy Destruction as a Multi-level Evaluation Criteria

For the energy aware design philosophy from chips to data centers, a global criteria is needed that can provide a uniform evaluation and control mechanism. Figure 1 shows the current chip thermal design approach – one where the heat sink temperature is driven down to a point that allows the critical chip temperature to remain at or below a fixed threshold level. The large discontinuity in temperature at the chip to heat sink interface results in irreversibility that manifests itself in destruction of available energy or exergy (essergy). As shown in fig. 3, a micromechanical approach that provisions the cooling fluid and undergoes phase change while minimizing the irreversibility could lead to much lower destruction of exergy. A similar exergy analysis can be applied by drawing a control volume around a system and eventually around a data center. This is shown in fig. 7. The available work reduces with temperature as we go from chips to datacenter. Conversion of electrical power to heat leads to destruction of exergy in chips. The resulting heat is transported across finite temperature differences, finally being rejected at ambient condition. The transport process across multiple thermal resistances leads to further destruction of exergy. Non ideal effects include friction and other dissipative effects.

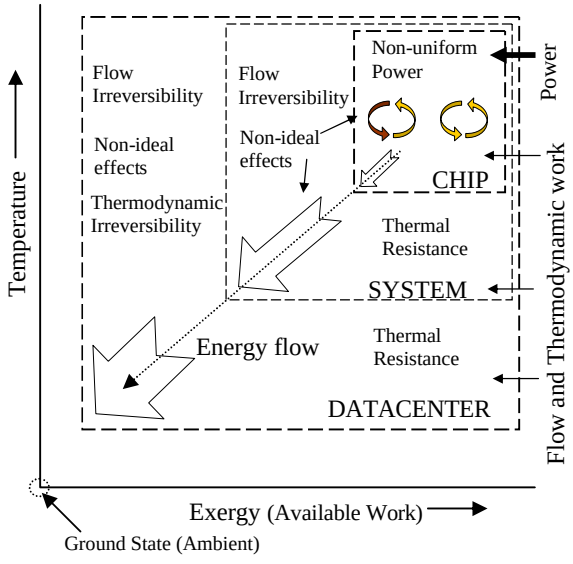
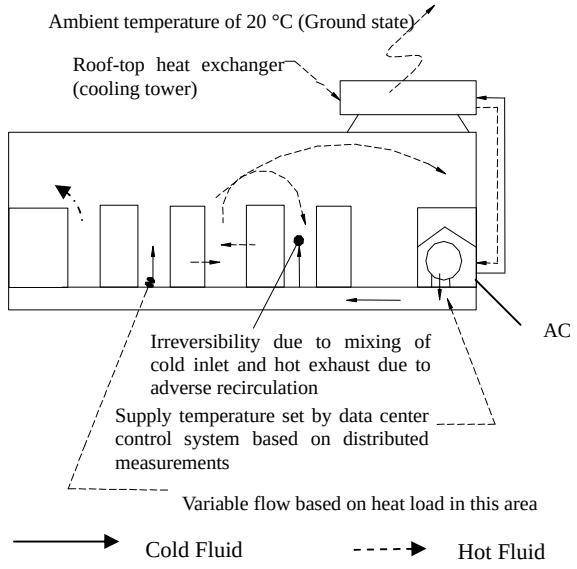


Figure 7: Schematic of exergy destruction and energy flow from chips to data center

Additional work is added from auxiliary equipment as energy transport occurs inside the system and the datacenter. Irreversibilities in flow and heat transfer add up at each stage leading to further destruction of exergy. In a data center, exergy destruction occurs in equipment such as AC units, heat sources in the systems, and from flow related irreversibilities due to mixing of hot return stream with the cold inlet air[3]. Figure 8 shows the irreversibilities due to mixing. A finite volume approach that identifies exergy loss in equipment and at physical locations within the data center due to adverse mixing can be very useful in provisioning the cooling resources efficiently. Besides provisioning of cooling resources, as shown in fig. 8, the heat loads can be distributed in a data center such that mixing of the hot exhaust and the cold inlet air is minimized and AC resources are used efficiently[3]. Detailed work is underway in development of a second law analysis technique and preliminary work is shown in a forthcoming paper [22]

Among global distribution of data centers, overall exergy efficiency with respect to ambient (dry and wet bulb temperature) used to pick a physical geographic location



Row-wise heat distribution and the provisioning of AC equipment

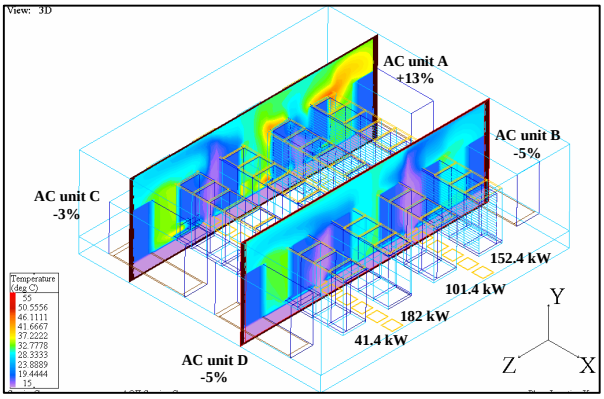


Figure 8: Mixing and Provisioning

Exergy as a Computer System Performance Metric

Lastly, the second law analysis technique can be used to characterize the cost of computing in the computing utility of tomorrow. A computer system performance criteria that quantifies MIPS (million instructions per second) or transactions per second to unit of exergy destroyed is proposed. Such a metric would enable a global evaluation, and selection of the best geographic location and configuration to enable a given computer service in the compute utility of the future.

Reuse of Waste Heat

Second law analysis has been reported for least-energy optimization of forced convection plate-fin heat sink for electronics application[23]. Exergy analysis can also be extended to examine opportunities for energy reuse. Indeed, a holistic analysis of the type shown in fig. 7, can be used to determine available work that can be extracted from the heat rejected from the datacenter. The available work obtainable at each level can be identified and extracted for useful purposes. Waste heat available at different chip, system and data center levels could be used to provide auxiliary cooling. As an example, a data center with local power generation may use flue gases to power an absorption refrigeration cycle[24]. The impact of the thermal loads on the external environment can be assessed by choosing an exergy ground state that reflects the ambient conditions of the geographic region within which the data center operates[4][22]. Thus, an exergy analysis within the data centers, and indeed globally from chips to data centers, can be used to evaluate techniques for energy reuse.

SUMMARY AND CONCLUSIONS

The era of pervasive computing has arrived. Computing will be a utility with a global distribution of data centers. In this context, role of thermal and computer sciences is to examine the globe as a control volume. Indeed, a control volume drawn around a global distribution of data centers, the confines of which maintain balanced compute and cooling loads, will save a world of energy. This paper presented some options for achieving such a balanced global computing utility. The data centers, installations that can utilize 15 MW of power, are the first instance of application of the balanced cooling-heat load philosophy. Preliminary work [3][10][11] has shown that energy consumption of the cooling portion (5MW) of a 15 MW data center can be reduced by 35% using a system that provisions cooling. An additional 20% savings is projected when power is scaled in systems. This is a reduction of over 2.5 MW and is achieved through deployment of air conditioning resources locally when required, and scaling down power in systems that are not being used in a data center. Future global computing utility will have hundreds of high power data centers, and the energy savings mandated by

the world[4] will require management of energy as a key resource.

An important step in global energy management with respect to computing is an evaluation approach based on second law of thermodynamics. A holistic examination of exergy destruction from chips to data centers is introduced in this paper to determine and eliminate irreversiblities in the thermal path. Furthermore, this approach is proposed as a measure for quantifying computer performance – MIPS per unit of exergy destroyed or transactions per unit of exergy destroyed. In the era of Grid computing[25], a user could fulfill a high performance computing need with little effort by accessing the globally distributed computing resources. A criterion that uses destruction of available energy as metric to charge the user and deliver the service will enable the selection of the proper set of compute resources in the proper geographic location. Indeed, an excellent example of global computing in use today are sophisticated search engines [26] that place user query in various geographic locations.

ACKNOWLEDGMENTS

The author acknowledges with great gratitude the work performed by the cool team at Hewlett Packard Laboratories – thanks to Cullen Bash, Ratnesh Sharma, Monem Beitelmal. Thanks also to our collaborators Amip Shah, Tim Boucher and Aaron Wemhoff of the CITRIS program at University of California, Berkeley. Finally, many thanks to the management at Hewlett-Packard for encouraging this type of work in thermo-fluids.

REFERENCES

[1] Friedrich, R., Patel, C.D., Jan 2002, "Towards planetary scale computing - technical challenges for next generation Internet computing", *THERMES 2002*, Santa Fe, New Mexico

[2] Friedrich, R, Rolia, J, Patel, C.D, October 2002, "Service Centric Computing – next generation Internet Computing", Performance 2002, Rome

[3] Patel, C.D., Sharma, R.K, Bash, C.E., Beitelmal, A, "Thermal Considerations in Cooling Large Scale High Compute Density Data Centers, " May 2002, ITherm 2002 - Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems" San Diego, California

[4] Patel, C.D., Sharma, R.K, Bash, C.E. and Graupner, S., 2003 "Energy Aware Grid: Global Workload Placement based on Energy Efficiency", IMECE 2003-41443, 2003 International Mechanical Engineering Congress and Exposition, Washington, DC.

[5] Arima, J, Apr 2000, "Top Runner Program", Workshop on Best Practices in Policies and Measures, Copenhagen

[6] Bodas, Deva, "Data Center Power Management and Benefits to Modular Computing", Spring 2003, Intel Developers Forum, http://www.intel.com/idf/us/spr2003/presentations/S03USMODS137_OS.pdf

[7] Sharma, R., Bash, C.E, Patel, C.D, Friedrich, R.S, Chase, J, "Balance of Power: Dynamic Thermal Management for Internet Data Centers", Hewlett-Packard Laboratories Technical Report: HPL-2003-5.

[8] Iyer, S., Luo, A., Mayo, R., and Ranganathan, P., "Energy-Adaptive Display System Design for Future Mobile Environments". Proceedings of the First International Conference on Mobile Systems, Applications, and Services, May 2003.

[9] "Data Center Energy Characterization Study", Pacific Gas and Electric, California, USA, Feb 2001.

[10] Bash, C.E., Patel, C.D., Sharma, R.K., 2003, "Efficient Thermal Management of Data Centers – Immediate and Long-Term Research Needs", Intl. J. Heat, Ventilating, Air-Conditioning and Refrigeration Research, Vol. 9, No. 2, pp137-152

[11] Patel, C.D., Sharma, R.K, Bash, C.E., Beitelmal, A, Friedrich, R., "Smart Cooling of Data Centers", July 2003, IPACK2003-35059, Proceedings of IPACK03-International Electronics Packaging Technical Conference and Exhibition, Maui, Hawaii.

[12] Sharma R, Bash, C.E, Patel, C.D, June 2002, "Dimensionless Parameters for Evaluation of Thermal Design and Performance of Large Scale Data Centers", AIAA-2002-3091, American Institute of Aeronautics and Astronautics Conference, St. Louis, MO

[13] Patel, C.D and Bash, C.E, "Thermal Management of Chips and Systems", Section 7, University of California, Berkeley Extension Course Notes, March 2001

[14] Patel, C.D., "Enabling Pumped Liquid Loop Cooling: Justification and the Key Technology and Cost Barriers", International Conference on High-Density Interconnect and Systems Packaging, Denver, CO, 2000.

[15] Bash, C.E., Patel, C.D., Beitelmal, A., Burr, R., "Acoustic Compression for the Thermal Management of Multi-Load Electronic Systems", ITHERM, San Diego, CA, May 2002

[16] Kakac, S.,Yuncu, H.,Hijikata, K., "Cooling of Electronic Systems"; Dordrecht, Kluwer Academic Publishers; E258;1994

[17] Bash, C.E, Patel, C.D, Sharma. R.K, "InkJet Assisted Spray Cooling of Electronics", IPACK2003-35058, Proceedings of IPACK03- International Electronics Packaging Technical Conference and Exhibition, Maui, Hawaii.

[18] Kumar, R., Farkas, K., Jouppi, N., Ranganathan, R., . Tullsen, D., "Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power Reduction", Proceedings of the 36th International Symposium on Microarchitecture, December 2003

[19] Labounty, C., Shakouri, A., Abraham, P. and Bowers, J.E., Nov 2000, “Monolithic integration of thin-film coolers with optoelectronic devices”, Optical Eng., vol. 39, No.11, pp2847-2852
[20] Patel, C.D., Bash, C.E., Belady, C., Stahl, L., Sullivan, D., “Computational Fluid Dynamics Modeling of High Compute Density Data Centers to Assure System Inlet Air Specifications”, July 2001, Proceedings of IPACK’01 – The PacificRim/ASME International Electronics Packaging Technical Conference and Exhibition, Kauai, Hawaii.
[21] R. Schmidt, July 2001, “Effect of Data Center Characteristics on Data Processing Equipment Inlet Temperatures”, Proceedings of IPACK’01 – The PacificRim/ASME International Electronics Packaging Technical Conference and Exhibition, Kauai, Hawaii.
[22] Shah, A. J., Carey, V. P., Bash, C. E., Patel, C. D., 2003 (submitted), “Exergy Analysis of Data Center Thermal Management Systems”, IMECE 2003-42527, 2003 International Mechanical Engineering Congress and Exposition, Washington, DC.
[23] Iyengar,.M., Bar-Cohen, A., “Least-Energy Optimization of Forced Convection Plate-Fin Heat Sinks”, 2002, Proceedings of the 2002 Inter Society Conference on Thermal Phenomena, pp 792-799
[24] Herold, K., Radermacher, R., “Integrated Power and Cooling Systems for Data Centers”, 2002, Proceedings of the 2002 Inter Society Conference on Thermal Phenomena, pp 808-811
[25] Foster,I., Kesselman, C., (Eds.), “The Grid: Blueprint for a New Computing Infrastructure”, Morgan Kaufmann Publishers, 1999.
[26] Barroso, L., Dean, J., Hölzle, U., 2003, “Web Search for a Planet: The Google Cluster Architecture”, IEEE Micro, IEEE Computer Society, http://computer.org/publications/dlib