Title
Exploring Total Power Saving from High Temperature of Server Operations

Authors
Lai, Liangzhen
Chang, Chia-Hao
Gupta, Puneet

Publication Date
2014-06-02

Exploring Total Power Saving from High Temperature of Server Operations

Liangzhen Lai Chia-Hao Chang Puneet Gupta

***Abstract*—The inlet air temperature of a data center is directly related to the efficiency of the cooling system. If the temperature can be raised without significant negative impacts on the power consumption and lifetime reliability of the IT equipments, the data center cooling can be more efficient and the cost of operation can be reduced.**

In this work, we analyze the power overhead and reliability loss of a server system by raising its operating temperature. By making changes to the processor implementation, our simulation results show that the total power consumption of data center can be reduced by 10% to 30% with 10°C to 15°C increase in inlet air temperature, while achieving the same reliability and performance goals.

I. Introduction

THE proliferation of cloud computing has resulted in the growing contribution of data centers to the total energy consumption. In 2005, data centers accounted for 1.2% of total electricity consumption of the United States [1]. With semiconductor, circuit and system technology improvements, significant improvements have been made in server system energy efficiency [2]. Meanwhile, the server system power is only a part of the power consumption in data centers. According to [3], the cooling system accounts for 25% to 67% of the total power consumption in a data center. This makes it essential to co-optimize both the computing system and cooling system.

Several contributions have been made to optimize data center cooling efficiency. For example, the use of an economizer has been proposed [3] to directly dissipate the heat to the ambient world. However, free cooling from economizers can be used only when the ambient temperature is sufficiently low. Current cooling objectives for the inlet air temperature range from $20\,^{\circ}\mathrm{C}$ to $25\,^{\circ}\mathrm{C}$ in most data centers. They follow the standards set by the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). However, the standard is supported by relatively dated studies [4]–[6].

For all cooling methods, energy consumption can be significantly reduced if the tolerable inlet air temperature can be increased [7]. This is usually undesirable due to the increase in server operating temperature (junction temperature). High junction temperature can result in reduced component lifetime reliability, degraded performance and increased power consumption. However, for modern silicon technologies, we make the following observations:

- Performance does not depend significantly on temperature and may even improve with increasing temperature for some future technologies [8].
- Server system lifetime may not strongly depend on temperature. Though server system energy consumption is dominated by the processor [9], failures in server systems are dominated by disk storage [10]. Disk failure mechanisms are not strongly temperature-dependent [11].

Based on these two observations, in this paper we explore the possibility of saving total data center energy consumption by *designing* processors to work at higher temperatures. The rest of the paper is organized as follows: Section II describes the effect of temperature on circuits. Section III provides the analysis of data center power. Potential power savings are explored in Section IV through simulation, and Section V concludes the paper.

II. Temperature dependence of circuit metrics

A server system power profile depends on the exact configuration and applications. Measured server system power breakdown examples from [12] are shown in Figure 1. Given that the processor can account for over half of server power consumption, in this section we study the temperature dependence of power, performance and reliability of a processor.

A. Power and Delay Analysis

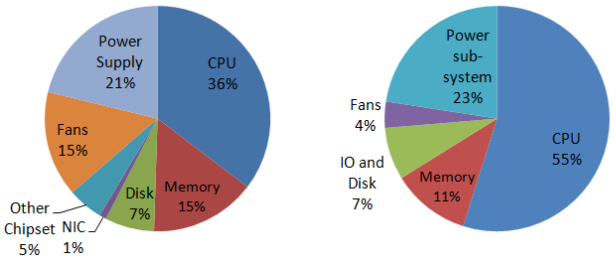
Switched capacitance, and hence dynamic power, has a weak linear dependence on temperature, while leakage power increases exponentially with temperature. For a commercial 45nm technology, the leakage of a inverter gate doubles every $55\,^{\circ}\mathrm{C}$ rise of junction temperature, as shown in Figure 2(a). Figure 2(b) shows that inverter chain delay degrades at the rate of only $0.042\%/\,^{\circ}\mathrm{C}$ in the operating temperature range of $50\,^{\circ}\mathrm{C}$ to $150\,^{\circ}\mathrm{C}$[1] According to [13], elevated temperature will result in decreased electric mobility and threshold voltage. Circuit delay can be modeled using alpha-power model [14] as:

$$Delay \propto \frac{V_{dd}}{\mu(T)(V_{dd} - V_{th}(T))^{\alpha}}$$

where V_{dd} is the supply voltage, T is the junction temperature, μ is the electric mobility, V_{th} is the threshold voltage and α is a technology-dependent parameter which is usually between 1 and 2. Technology scaling and lowered supply voltages (i.e. lowered $V_{dd} - V_{th}(T)$) can result in counter-intuitive $improved$ performance with increasing temperature.

This work is supported in part by NSF Variability Expedition grant CCF-1029030.

[1]Note that this is the junction temperature which is much higher than the inlet air temperature.



(a) Big science system (b) Blade server system

Fig. 1. Examples of server system power breakdown [12].

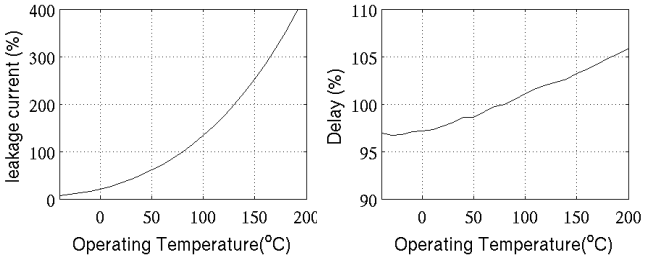


Fig. 2. SPICE simulation results of temperature dependence of power and performance(normalized to 60°C) for an inverter chain in a commercial 45nm technology: (a) normalized leakage power; (b) normalized delay.

For instance, Figure 3 compares 45nm and 32nm commercial CMOS technology temperature-delay dependence for an inverter chain. For low-power (i.e., low Vdd) servers in future technologies, we can observe this temperature reversal. For other types of technologies (e.g. non-planar devices), the trend may be different.

B. Circuit Lifetime Reliability Analysis

There are many temperature-dependent failure mechanisms in modern CMOS technologies: delay degradation caused by hot carrier injection (HCI) and negative bias temperature instability (NBTI), and hard failures caused by electromigration (EM), time-dependent dielectric breakdown (TDDB) and stress migration (SM).

HCI slows down transistors when they are switching, while NBTI increases the threshold voltage of PMOS transistors (and hence slows them down) when they are on. Drive current degradation due to HCI and threshold voltage degradation due to NBTI are plotted in Figure 4, with different expected lifetimes.

Based on mean-time-to-failure (MTTF) expressions in [15] and fitted values in [15]–[17], we can conclude that MTTF halves every 10 °C for both SM and EM and it halves every 20 °C for TDDB. Comparing the different hard failure mechanisms, TDDB is projected as the largest challenge with technology scaling [15]. In this work we consider TDDB as the hard failure reliability constraint of raising temperature. For TDDB, voltage and temperature dependence are plotted in Figure 5.

Although the processor is commonly the most power consuming component, it is not the most critical component for server reliability. Hard drives, as shown in Table I [10], are the most commonly replaced components. Moreover, based on the study [11] by Google, the disk drive lifetime has no obvious relationship with the server inlet air temperature. This

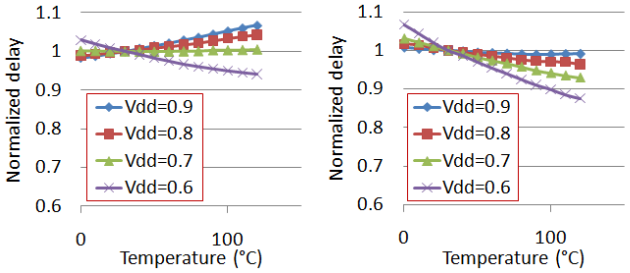


Fig. 3. Delay vs. temperature at 45nm (left) and 32nm (right) technologies using SPICE simulation. Note that the inverse dependence starts at higher Vdd for 32nm than 45nm.

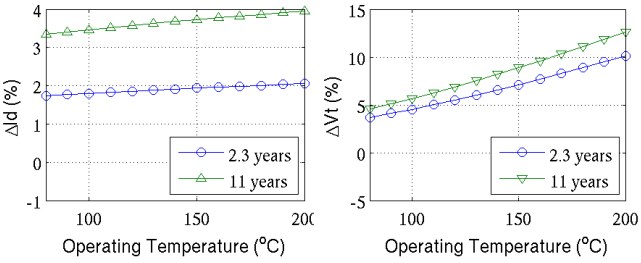


Fig. 4. (a) Drive current degradation due to HCI; (b) Threshold voltage degradation due to NBTI. The data is derived based on design manual of a commercial 45nm technology

shows that a higher operating temperature might not harm the overall reliability too much even for current server products. The conclusion matches the proof of concept experiment by Intel [18].

III. Data Center Power Analysis

To demonstrate the potential of raising the operating temperature, we perform analysis on both an industrial level processor design and a cooling system model calibrated under real operating conditions.

A. Server power analysis

1) Server power distribution: As the temperature of operation increases, the power consumption of the server hardware also increases. As shown in power breakdown examples in Figure 1, the processor consumes the most power in the server systems.

In this analysis, we only consider the temperature dependence of the processor. We assume that the other server components have constant power consumption. A similar assumption is also adopted by [19] and supported by their empirical observations. Memory components such as DRAMs are also reported to be quite consistent across different temperatures as in [20].

TABLE I
Relative frequency of hardware component failures that require replacement for an Internet service providing server system

Component	%	Component	%
Hard drive	49.1	SCSI cable	2.2
Motherboard	23.4	Fan	2.2
Power supply	10.1	CPU	2.2
RAID card	4.1	CD-ROM	0.6
Memory	3.4	Raid Controller	0.6

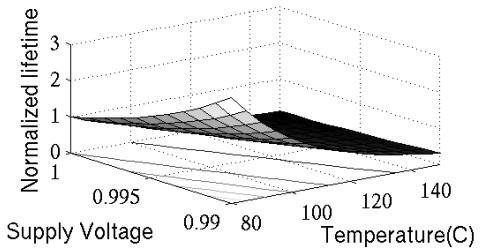


Fig. 5. Normalized TDDB lifetime v.s. temperatures and supply voltages. The values are derived from [15]

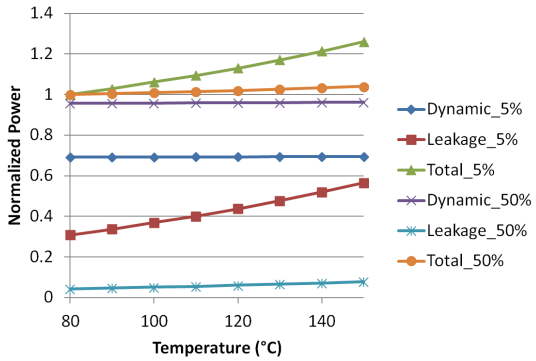


Fig. 6. The breakdown of total power, dynamic power and leakage power at different temperature points and utilization rates (50% and 5%): leakage power increases with temperature, while dynamic power remains almost constant

2) Processor power analysis: To analyze processor power under different temperatures, we implement the Leon3 processor [21] using commercial tools [22] with a commercial 45nm technology. Power and delay are calculated using standard cell libraries (.lib) which are characterized using SPICE under different temperatures. The processor power at different temperatures and utilization rates is plotted in Figure 6.

B. Data center cooling system

The basic function of a cooling system is to move the heat from the cooling objective to the ambient world through coolant circulation. The cooling efficiency is mainly controlled through adjusting the coolant flow rate and coolant temperature.

To model the temperature dependence of the cooling system, the analytical model in [23] for the data center cooling efficiency is used. Each component in a cooling system has a characteristic formula to simulate its behavior. The scenario of the model in [23] is shown in Figure 7(a).

To further model the scenario when the chiller is turned off (as shown in Figure 7(b)), we remove the central chiller, and directly use the coolant from cooling tower.

To simulate the cooling power consumption at different target inlet temperatures, we modified the configurations of the cooling system. For the first scenario with a chiller, we changed the chiller output water temperature while keeping coolant flow rates as constants to meet target inlet temperature. For the second scenario, we tuned the coolant flow rates, as they are the only factors related to inlet temperature. The initial configuration data is the same as that in [23].

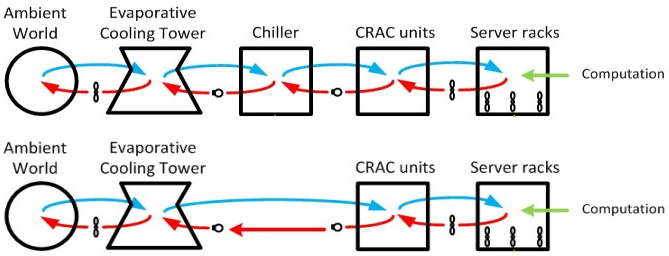


Fig. 7. Cooling scenarios: with chiller (top) [23] and with free cooling (bottom). A chiller is used to further reduce the coolant temperature if it is required for higher cooling efficiency. CRAC stands for the computer room air-conditioning.

IV. Experiment Setup and Results

To demonstrate the potential benefit of raising the server's target operating temperature, we also changed the processor configuration and design to compensate the impacts of temperature on reliability and delay.

A. Lifetime Reliability Compensation

Increased temperatures can reduce the reliability and lifetime of processors. This can be compensated at the design stage by a variety of techniques. For example, operating at a lower supply voltage will greatly increase the MTTF of TDDB. Figure 5 shows the TDDB MTTF versus both temperature and supply voltage. There are many other architecture/circuit level techniques to improve the reliability [24]–[26].

In our experiments, we lowered the processor operating voltage to achieve the same TDDB MTTF constraints under different junction temperatures as shown in Table II. As a side effect, soft errors can increase with reduced supply voltage which can be compensated through hardware design changes (e.g. [27], [28]). As the maximum voltage reduction is less than 1% of the nominal supply voltage, we expect such overheads to be small. In the experiments, we focus on the lifetime reliability and did not model the additional hardware cost for the soft error compensation.

For other circuit aging effects such as NBTI and HCI, we compensate for the impact of temperature rising by adding extra delay margin to the processor, which will be discussed in Section IV. B.

B. Delay Compensation

To compensate for delay degradation due to a lower operating voltage, higher temperature and larger amount of circuit aging, we improve the processor performance by threshold voltage assignment to replace original gates with the gates that have lower threshold voltages. The extra delay margin is calculated to account for delay degradation due to NBTI and HCI with 3 year expected lifetime, i.e. about 6% for 50°C temperature increase. Table II shows the configuration with same MTTF and delay. D.P. stands for dynamic power.

C. Power Saving Results

We simulate the data center power by combining the power model in Table II and the data center cooling system model

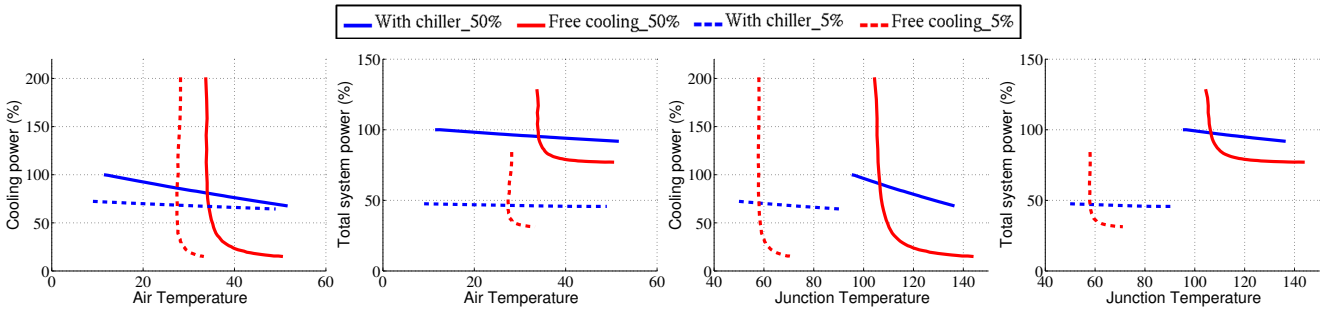


Fig. 8. Cooling system power and total data center power vs. inlet air temperature and junction temperature for different server utilization rates

TABLE II
SERVER POWER AFTER RELIABILITY AND DELAY COMPENSATION VIA LOWERING SUPPLY VOLTAGE AND MULTI-VT DESIGN

Temp.	Supply V.	Core Leakage	Cache Leakage	D.P.
80 °C	1	100%	100%	100%
90 °C	0.9967	109.3%	110.8%	99.4%
100 °C	0.9947	119.7%	123.5%	99.1%
110 °C	0.9907	130.3%	136.5%	98.3%

described in Section III.B. Simulation results are plotted in Figure 8. By raising the maximum tolerable inlet air temperature from 25°C to 35°C, the total data center power is reduced by 2% to 5% without an economizer, while with the economizer the power saving can be 10% to 30% under the typical data center processor utilization rates between 5% to 50% [29].

V. CONCLUSION

In this paper, we started with the observation that circuit performance is becoming less sensitive or even inversely dependent on temperature. To evaluate the benefit of increasing data center operating temperature, we analyzed the temperature-dependence of circuit reliability and cooling system efficiency. With co-optimization of both processor design and cooling system, simulation results showed 10% to 30% of total data center power savings by raising the inlet air temperature by 10°C, while keeping the same performance and reliability targets.

REFERENCES

[1] J. G. Koomey, "Estimating total power consumption by servers in the us and the world," 2007.

[2] E. Masanet *et al.*, "Estimating the energy use and efficiency potential of us data centers," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1440–1453, 2011.

[3] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, and B. Myatt, "Best practices for data centers: Lessons learned from benchmarking 22 data centers," *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August*, vol. 3, pp. 76–87, 2006.

[4] H. Blanks, "The temperature dependence of component failure rate," *Microelectronics Reliability*, vol. 20, no. 3, pp. 297–307, 1980.

[5] S. H. Charap, P.-L. Lu, and Y. He, "Thermal stability of recorded information at high densities," *Magnetics, IEEE Transactions on*, vol. 33, no. 1, pp. 978–983, 1997.

[6] R. Viswanath, V. Wakharkar, A. Watwe, V. Lebonheur *et al.*, "Thermal performance challenges from silicon to systems," 2000.

[7] T. Breen *et al.*, "From chip to cooling tower data center modeling: Influence of server inlet temperature and temperature rise across cabinet," *Journal of electronic packaging*, vol. 133, no. 1, 2011.

[8] D. Wolpert and P. Ampadu, "Normal and reverse temperature dependence in variation-tolerant nanoscale systems with high-k dielectrics and metal gates," in *Nano-Net*. Springer, 2009, pp. 14–18.

[9] K. Rajamani, C. Lefurgy, S. Ghiasi, J. C. Rubio, H. Hanson, and T. Keller, "Power management for computer systems and datacenters," in *Proceedings of the 13th International Symposium on Low-Power Electronics and Design*, 2008, pp. 11–13.

[10] B. Schroeder and G. A. Gibson, "Disk failures in the real world: What does an mttf of 1,000,000 hours mean to you," in *Proceedings of the 5th USENIX conference on File and Storage Technologies*, vol. 6, 2007, p. 2.

[11] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population," in *Proceedings of the 5th USENIX conference on File and Storage Technologies*, 2007, pp. 2–2.

[12] X. Feng, R. Ge, and K. W. Cameron, "Power and energy profiling of scientific applications on distributed systems," in *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International*. IEEE, 2005, pp. 34–34.

[13] Y. Cheng and C. Hu, *MOSFET modeling and BSIM3 user's guide*. Springer, 1999.

[14] T. Sakurai *et al.*, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *Solid-State Circuits, IEEE Journal of*, vol. 25, no. 2, pp. 584–594, 1990.

[15] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The impact of technology scaling on lifetime reliability," in *Dependable Systems and Networks, 2004 International Conference on*. IEEE, 2004, pp. 177–186.

[16] I. A. Blech, "Electromigration in thin aluminum films on titanium nitride," *Journal of Applied Physics*, vol. 47, no. 4, pp. 1203–1208, 1976.

[17] E. Ogawa, J. McPherson, J. Rosal, K. Dickerson, T.-C. Chiu, L. Tsung, M. Jain, T. Bonifield, J. Ondrusek, and W. McKee, "Stress-induced voiding under vias connected to wide cu metal leads," in *Reliability Physics Symposium Proceedings, 2002. 40th Annual*. IEEE, 2002, pp. 312–321.

[18] D. Atwood and J. Miner, "Reducing data center cost with an air economizer," *White Paper: Intel Corporation*, 2008.

[19] E. Elnozahy *et al.*, "Energy-efficient server clusters," *Power-Aware Computer Systems*, pp. 179–197, 2003.

[20] M. Gottscho *et al.*, "Power variability in contemporary DRAMs," *Embedded Systems Letters, IEEE*, vol. 4, no. 2, pp. 37 –40, June 2012.

[21] [Online]. Available: http://www.gaisler.com

[22] Cadence encounter rtl compiler. [Online]. Available: http://www.cadence.com/products/ld/rtl_compiler/pages/default.aspx

[23] M. Iyengar and R. Schmidt, "Analytical modeling for thermodynamic characterization of data center cooling systems," *Journal of electronic packaging*, vol. 131, no. 2, 2009.

[24] Z. Qi and M. R. Stan, "Nbti resilient circuits using adaptive body biasing," in *Proceedings of the 18th ACM Great Lakes symposium on VLSI*. ACM, 2008, pp. 285–290.

[25] J.-T. Kong, "Cad for nanometer silicon design challenges and success," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, no. 11, pp. 1132–1147, 2004.

[26] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "Exploiting structural duplication for lifetime reliability enhancement," vol. 33, no. 2, pp. 520–531, 2005.

[27] R. Rao *et al.*, "Soft error reduction in combinational logic using gate resizing and flipflop selection," in *Computer-Aided Design, 2006. ICCAD'06. IEEE/ACM International Conference on*. IEEE, 2006, pp. 502–509.

[28] S. Mitra, M. Zhang, N. Seifert, T. Mak, and K. Kim, "Built-in soft error resilience for robust system design," in *Integrated Circuit Design and Technology, 2007. ICICDT'07. IEEE International Conference on*. IEEE, 2007, pp. 1–6.

[29] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.