Power Proportional Computing for "Green" Servers

Earl McCune, Jr.

Panasonic Technology Fellow (ret.), Santa Clara, CA emc2@wirelessandhighspeed.com

Abstract - Achieving energy-efficient "Green" operations within data centers used for applications such as Cloud Computing requires matching the power consumed to the data processed at each server in real-time. Beyond having a high efficiency in the server power supplies, it is vitally important to only draw power when the server is actually processing data. To operate at maximum energy efficiency, in the times when a server is idle it needs to draw no power for the server farm. By matching power draw to actual data processing activity at logic speeds, the average energy draw of the server farm drops by 50% or more with no reduction in throughput. Drawing on technology developed for efficient radio transmitters, an agile power supply, able to provide tight voltage regulation and still transition between power-off and power-on (or the other way) in nanoseconds without transition overshoot is described. With this nanosecond agility, this also solves the objective for elastic computing. Additionally, the supply pin pairing required by this energy management method provides benefits toward reducing electromagnetic interference (EMI). Proportional reduction in processor operating temperature improves reliability, along with reducing facility cooling loads.

Index Terms — cloud computing; switching power supply; energy efficiency; data-dynamic power; electromagnetic interference

I. INTRODUCTION

The need to achieve energy efficiency in cloud computing is becoming more important now that the energy consumed by data centers are on the order of 2% of worldwide electricity generation [1] [2]. Because large scale computing facilities are expensive to build and operate, and because computing power use is responsible for a substantial part of data center capital and operating expenses, there is increasing pressure to improve the efficiency of energy use in data centers [3]. Of greatest importance to this goal is to draw energy only when active computing is happening, which can save half of the energy which is used today.

Energy flow in a data center facility is unlike that in a wireless communication system, where a high power output signal carries information away from the transmitter. The wireless communication system has a top level power flow of

$$P_{DC} = P_{OUT} + P_D \quad (1)$$

where P_{DC} is the input power, P_{OUT} is the signal output power, and P_D is the dissipated power from this process. Efficiency of this process (η) is readily evaluated as

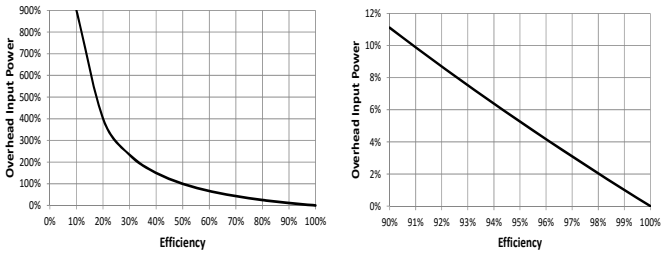
$$\eta \equiv \frac{P_{OUT}}{P_{DC}} = 1 - \frac{P_D}{P_{DC}} \quad (2)$$

which shows that efficiency is maximized only when the dissipated power is minimized.

In a computing facility, a large propagating output signal (P_{OUT}) does not exist, so the efficiency from (2) evaluates to zero because all of the input energy gets converted to heat and is dissipated. This is a lot of heat energy that must be managed carefully to not have a correspondingly large facility temperature increase. Even still, the cooling system within a computing facility must eventually dispose of all the heat from this converted input electrical energy. This cooling system itself requires energy to operate, further increasing the input electrical energy requirements of the computing facility.

Energy is not supplied to a computing facility in the form that it is used by the computing units in that facility, so a conversion process is required. Converting electrical energy from its input form, likely utility alternating current (AC), to its output form of DC near 1 volt, has a finite efficiency. Additional input energy is therefore required to support this conversion loss, as shown in Fig. 1. Having low conversion efficiency is wasteful, and therefore inherently costly, due to this additional overhead input power that is directly converted to heat and does no useful work.

When the conversion efficiency is greater than 90%, further savings in input power from additional improvements in conversion efficiency are very small. Any such improvements are not material to the top level energy supply, and the focus on how to reduce the input energy demand must shift to reducing the energy used by the computing units themselves and their support devices, including the cooling system, all while maintaining or improving throughput.



a b

Fig. 1. Efficiency of the power conversion performance affects the main input power required: a) input power slowly approaches the output power when efficiency exceeds 70%; b) close-in detail when conversion efficiency exceeds 90%.

The solution is to realize true power proportional computing in the sense that power is drawn when useful computing is happening, and no power is drawn during all idle times. This includes realizing zero static power draw within idle intervals,

978-3-901882-85-2 (c) 2016 IFIP

something that only happens when the logic supply goes to zero volts. Recovery back to full operating speed is equally fast. This technique is called zero-power idle (ZPI), and the initial report on the power supply design that meets this switching objective is in this paper.

Within this paper, Section II highlights some of the energy consumption properties of data centers. Section III introduces data-dynamic power supply operation for ZPI and how this minimizes server energy consumption. Measurements from a data-dynamic power supply are presented in Section IV. Finally in the summary conclusions are drawn.

II. Computing energy consumption

In any computing facility the energy consumed results in power dissipation at many locations. Obvious power dissipation occurs in the processing units, the memory, the message input and output (I/O) units, and the cooling system. Of these, the memory and the I/O units generally must operate continuously. Continuous operation is intentionally not designed for the processor units because of important results from queueing theory and the need for low latency responses to messages from the cloud computing facility. Measured data center performance in [4], which shows that these processors are busy primarily between 10% and 50% of the time, is completely consistent with queuing theory for these applications [5].

Power dissipated by server processors nominally follows the relationship

$$P_{D,\text{Processor}} = P_{DYNAMIC}\left(f,V\right) + P_{STATIC}\left(T,V\right) \qquad (3)$$

which has a dynamic component with operating variables of clock frequency f and operating voltage V, and a static power component that is also a function of V and of the temperature T. The dynamic power component follows the well-known relationship based on capacitor charge/discharge cycles

$$P_{DYNAMIC}\left(f,V\right) = \alpha NCV^2 f \qquad (4)$$

where α is the average fraction of circuits that are switching, N is the number of switching circuits available, and C is the effective switching capacitance for each switching circuit (a process metric). To minimize (4) one needs a minimum operating voltage, the lowest practical clock frequency, a CMOS process with small C value, and a design having the least amount of circuitry (a small N value). The dynamic voltage and frequency scaling (DVFS) technique is commonly used to manage (4) by scaling V and f to operate near a local power minimum for the compute job requirements of the moment. Within an Advanced Configuration and Power Interface (ACPI) implementation, this control corresponds to the processor states P0 through P16 [11].

In nanometer CMOS processes there also is a significant static power component. Mathematical models for this are not well published, but the following model is derived from the information provided in [6]

$$P_{STATIC}\left(T,V\right) = K \cdot 1.04^T \cdot V^3 \qquad (5)$$

where the operating temperature T is in Kelvin and the core voltage V is normalized to its maximum value V_{MAX}. The proportionality constant K depends on the specific circuit design, which for the Intel Ivy Bridge is 0.185. This static power remains, even when the processor dynamic power (4) is stopped by setting $f = 0$ (corresponding to the ACPI state C2). To stop this static power, (5) shows that bringing the device voltage V to zero is the only option.

DVFS is an optimum power management strategy when $P_{DYNAMIC}$ dominates. In the presence of significant P_{STATIC} this conclusion no longer holds. DVFS has several limits on its available performance which therefore also limit the ability of DVFS to reach the desired power dissipation minima when the load current can have large "sudden" changes. First, frequency scaling (FS) requires a frequency synthesizer, usually implemented using a phase locked loop (PLL), which has settling times between frequency settings that depend on the PLL loop dynamics. From long experience with communications applications, a well-designed PLL exhibits a settling time T_{SW} governed by

$$T_{SW} > \frac{0.5}{BW_L} \qquad (6)$$

where BW_L is the PLL loop bandwidth in Hz. The value in (6) is a lower bound, and the quality of the PLL design is measured by how closely T_{SW} does get to this lower bound. For example a PLL with BW_L = 200 kHz may exhibit T_{SW} = 10 microseconds, which is a factor of 4 greater than the bound from (6). This amount of time is far slower than logic times, meaning that changing frequency through PLL tuning cannot meet full elastic computing objectives. Holding the clock frequency at a fixed value removes this possible bottleneck.

Another difficult issue for achieving the DVFS operating goal is a large difference between compute-idle interval transitions of the processor and of the power supply's ability to shift its output voltage value. Processors are capable of transitioning into and out of idle states in nanoseconds. Power supplies take milliseconds to power down and back up. This ~10^6:1 difference results in many server designs keeping the processor powered up as long as the idle operation gaps are shorter than a few 10's of milliseconds. This long response time for power supplies is a constraint imposed from Conservation of Energy. All switch-mode power supplies (SMPS) have energy stored in their output networks: inductor current driving its magnetic field, and capacitor charge forming its electric field. As the supply output changes, both the inductor current and the capacitor charge must correspondingly change.

To affect an output voltage change ΔV_O, say from V_{O1} to V_{O2}, any switching power supply must correspondingly change the energy contained in its output inductor L (E_L) and capacitor C (E_C). This total stored energy change in the output network is dependent on the two output voltages and the load resistance R_{LD} by

$$\Delta E = \Delta E_L + \Delta E_C = \frac{1}{2}\left(\frac{L}{R_{LD}{}^2} + C\right)\left(V_{O2}{}^2 - V_{O1}{}^2\right) \quad (7a)$$

$$= \left(\frac{L}{R_{LD}{}^2} + C\right)\Delta V_O \cdot avg\left(V_{O2}, V_{O1}\right) \quad \text{joules} \quad , \quad (7b)$$

where the average voltage across the voltage change is an additional scaling parameter. This energy change must occur within the output voltage transition time Δt leading to a *power* requirement from the switching supply of ΔE/Δt to do this change. As Δt gets shorter, down to logic speeds (which is required to achieve fully elastic computing), this power increases rapidly. For example, with L = 0.47 microhenry, C = 450 microfarad, R_{LD} = 0.1Ω, V_{O1} = 0, and V_{O2} = 1.7V (the turn-on transition), from (7) the energy change in this supply is 0.72 millijoule. Effecting this energy change in 1 millisecond requires 0.72 watts from this 30 watt supply. Faster transitions require larger power bursts: for a 1 microsecond transition 720 watts is required, which increases to 72 kilowatts for a 10 nanosecond transition. This is a major reason why we have soft-start functions, and why elastic computing is difficult to achieve. This transition-driving power is not useful to the computing load, and puts a limit on how fast any DVS supply can react to a change in computing load requirements. Therefore in DVS, voltage scaling speed is fundamentally limited by conservation of energy considerations of the supply output network, along with dynamics of the supply control loop.

When using normal power supplies, therefore only when the processor is known to be heading for a long duration idle state is it placed into a deep sleep or hibernation state (ACPI C3 state), or actually powered OFF (ACPI state D3-cold). Otherwise the power must be maintained on the processor, though the voltage may be adjusted down if the response time of the power supply to slight voltage changes is fast enough (10's of microseconds) in both going down and, importantly, in recovering back to processor operating-state voltage.

In a server application the computing and idle times are typically measured in microseconds. Switching power supplies cannot stop this fast, and soft-start functions do not allow them to start this fast either. The six decade difference in needed vs. available response times shows that new techniques are needed to match power supply performance to this 'Green' processor operation need, which at the same time achieves elastic computing objectives. Specifically, any new technique must not be bound by the conservation of energy constraint (7) in order to match supply response time to compute load demand variation times.

III. Data-dynamic power management

In 2009 a technique named PowerNap [8] was proposed, and intended to eliminate server idle power dissipation. This definitely is a step in the right direction, but for whatever reason this technique did not get widely implemented. One can speculate that one problem is that the PowerNap technique does not address static power (5).

To solve this problem for the general case of (3), here we apply techniques used in efficient, switch-based, high power radio transmitters to this processor power supply application, bringing technology developed for very different applications to bear on this problem. Switching radio power amplifiers operate at gigahertz frequencies and multi-ampere currents, with switching transition times well below 1 nanosecond. Having already solved the switching power problem for radio transmitters, application of this RF-based technology to the management of processor power is straightforward.

In order to have the processor power supply provide OFF and ON intervals that match server processor activity, it is necessary to interrupt the 10's of amperes of supply current in nanosecond timescales with no transient to damage either the processor load or the supply itself. Momentum of the inductance does not allow the inductor current to change rapidly. The solution is to keep this inductor current flowing, but to "instantaneously" redirect it away from the load as shown in Fig. 2. Simultaneously the output capacitor has its 'ground' terminal disconnected, and the synchronous rectifier is stopped in an open state. To do this, two switches are added to a conventional switching buck power converter, one to manage the capacitor voltage, and the other to manage the inductor current [7]. All of these switches are implemented with transistors that switch in less than 10 nanoseconds upon command from an external supervising/scheduling processor.

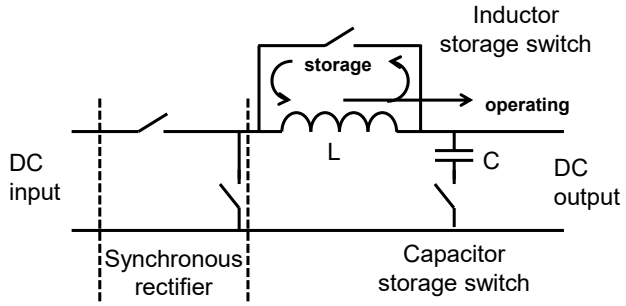


Fig. 2. Nanosecond transitions can occur on the load current when the energy in the power supply output filter is maintained, but with the stored voltage and current held for later use. When the output is OFF the charge stored in the floating output capacitance remains, and the inductor current is looped through the inductor.

To re-route the load current, both switches in Fig. 2 in the output network are operated simultaneously. While the output current is routed away from the load, it is equally important to interrupt current drawn from the supply input. This is done by holding open the synchronous rectifier switches between the storage inductor and the supply input. With this arrangement, the input current goes to zero when the output current is zero because there is no connection between the input supply and the regulator, much less the load. This procedure maintains the energy stored in the SMPS output network while the

output is OFF so that it remains available to quickly restore the output supply on command. This is fundamentally different from turning the supply off, where the internal energy is dissipated away and must be re-stored when the supply is reactivated.

An example of the transition speed available from this approach is provided by the measurement in Fig. 3. Here the supply to the processor is zeroed for 40 microseconds and then recovered, with transition times appearing to be instantaneous at the time scale of 50 microseconds per division. More detailed measurements do show that the actual transition time is less than 20 nanoseconds for each transition direction. It is also evident that there is no overshoot on either turn-OFF or turn-ON transients to damage the processor. This transition behavior is very close to ideal. In essence, this shows that the transition time between ACPI states C0 and C2 (or C3) becomes nearly 'frictionless', here taking less than 20 nanoseconds each. Achieving logic speed for these transitions is a significant enabler of maximally elastic computing.

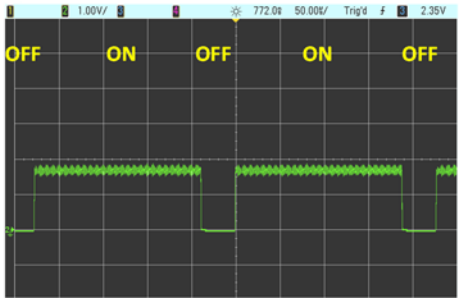


Fig. 3. Dynamic power supply output operation for a 17 ampere load at an 83% duty cycle. Horizontal scale is 50 microseconds per division, showing the OFF time is 40 microseconds with 20 nanosecond transition times in either direction.

During the ON intervals in Fig. 3 the switching regulator operates normally. Full and normal feedback control is present to assure that the output voltage is exactly what is needed. There is no energy inertia using this design approach, since here the energy is held and not moved, confirming that (7) does not constrain this technique.

Operation of this supply state change is presently implemented with commands from an external controller. This is likely to be from a processor working as the scheduler managing an array of servers, such as that shown in Fig. 4.

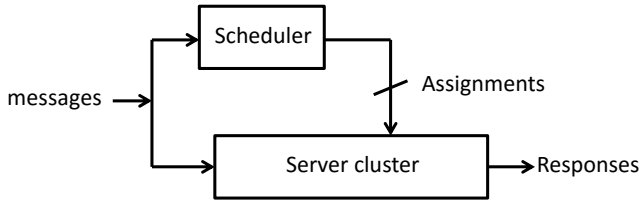


Fig. 4. The scheduler processor for a set of server processors is aware of when each processor has a task or not, and can provide the associated power ON and OFF commands to the power supply for each processor.

The overall power dissipation of a processor operated with this dynamic power supply technique is reduced by the fraction of time that it is powered OFF. The model from (3) therefore now changes to account for the duty cycle operation, giving

$$P_{D,\text{Processor}} = \left(P_{DYNAMIC}\left(f,V\right) + P_{STATIC}\left(T,V\right)\right)\cdot D \quad (8)$$

where D is the average duty cycle of the power ON state. This duty cycle represents the realized energy savings from powering processors completely down when idle, and can approach truly data-proportional power consumption only when the overhead of times needed to shut down and recover from the shutdown are much shorter than the actual processor work time, as shown here in Fig. 5 and with more detail in [9]. Therefore not only must the power supply be able to react to the nanosecond scale load changes, managing state recovery and reaction time to the scheduler's power command must also be short compared to the useful processing time for best energy savings.

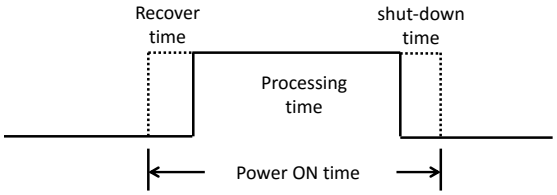


Fig. 5. Matching the power duty cycle D, based on the power ON time, to the processor activity time requires minimizing the overhead times required for safe shutdown and state recovery.

This lower power dissipation in the processor comes with a corresponding drop in average processor temperature. Whether this temperature drop is used to reduce heatsink cost, or to improve reliability, or some other combination is up to the actual computing hardware implementers.

This same arrangement is present in a single-chip multiple core processor array. For best effect this technique should be applied to each core individually, following the design principle of only powering the circuitry that is presently doing useful work.

It is interesting to further observe that this type of energy management forces the power source and return pins for each controlled block to be strictly matched. This is also a requirement for the greatest reduction of electromagnetic interference (EMI) from any CMOS circuit, by minimizing the area of each individual processor current loop [10]. Implementing processor power management with this ZPI technique naturally matches with preferred techniques for minimizing circuit EMI.

IV. Measurements

Measurement of the data-dynamic ZPI power supply prototype for short interruptions in the output current is already shown in Fig. 3. Here the load current of 17 amperes is interrupted for 40 microseconds and then immediately

restored. The transition times to OFF and back to ON appear instantaneous at this 50 microsecond per division timescale.

Under usual conditions with a one second idle time it is likely that the conventional switching power supply would be fully shut down, and then restarted. While there is plenty of time to do this, the result of draining the supply's energy, and then replacing that energy, makes even this overall energy efficiency worse than using the ZPI technique.

The entire point of the ZPI technique is to reduce the input current in proportion to the reduction in average output current to get the desired input energy savings. Measurement of input current as the output duty cycle is varied is reported in Fig. 6. Tracking is excellent, and follows a nearly perfect straight line. This line does not pass through the origin, but shows an offset of 11.9 milliamps (mA). This is the bias current of the switching regulator integrated circuit used in this design, and is 0.5% of the design maximum input current.

Efficiency of this power supply as the output duty cycle is varied is presented in Fig. 7. Efficiency of this supply remains above 95% as long as the operating duty cycle exceeds 10%. Below that duty cycle the flow of continuous bias current into this switching regulator becomes significant to the average and the transfer efficiency drops off. This switching regulator is not designed to care about 12 mA when the usual supply input current is 2 amperes. And for a design target of 35% duty cycle this bias current is also not significant enough to matter.

For applications other than servers, such as tablets and laptop computers, the operating system (e.g. Task Manager) often reports that the system is idle 90% of the time, or more. While the original intent of this design is toward server processors, the ZPI approach certainly is applicable to much lower duty cycle applications. The practicality of this is demonstrated in Fig. 6. In battery powered devices this ZPI technique would certainly improve their all-important battery life.

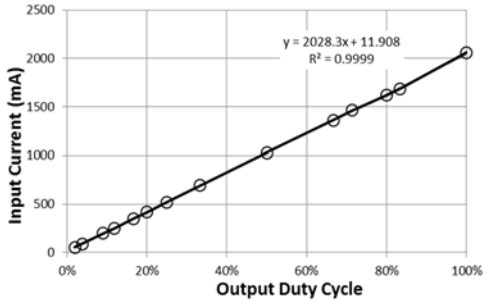


Fig. 6. Input current reduces linearly with the average output current, proving that power proportional computing is realized.

Another view of the power transfer efficiency of this prototype is provided in Fig. 8. Here the OFF time is varied from 5 microseconds out to one second. Effect of the regulator bias current begins to appear when the OFF time exceeds one millisecond. Whether this is important or not depends on the actual application. In all cases this confirms that the straight line characteristic between input current and supply duty cycle is due to the supply efficiency being maintained in the presence of interrupting the current flow to the load and storing the energy in the supply output network, keeping the already stored energy available for rapid restoration to the load on command.

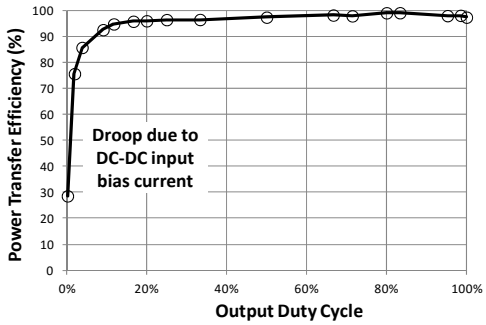


Fig. 7. Measures of power transfer efficiency across a wide range of activity duty cycles. Efficiency remains high until the input current is small enough to be comparable to the regulator bias current.

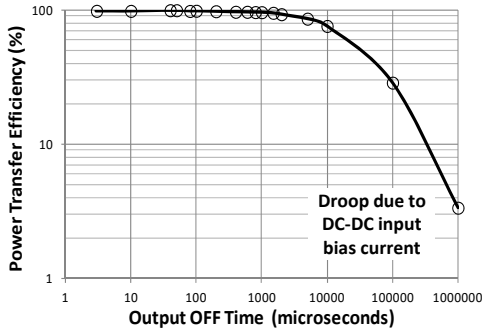


Fig. 8. Power transfer efficiency of the ZPI power supply as the OFF time is increased. Redesign of the regulator IC for reduced bias current will flatten this curve.

The use of DVFS for logic power management never gets to zero input power and has restricted agility in varying both the clock frequency and the operating voltage, for different reasons. The result of (8) suggests that a more effective power management strategy for nanometer CMOS devices is to duty cycle the power and operation instead, running the process at high speed and therefore minimum time whenever it is needed and then shutting the logic completely off when it is not being used. Of course, these techniques of ZPI and DVFS can be combined since they are independent, should other optima be of interest to particular designers.

V. Conclusions

When evaluating the energy efficiency of a computing facility it is certainly important to maximize the efficiency of each energy conversion step. But even if the conversion efficiency is perfect, the input energy will never get below the

consumption of all parts of the computing facility. Further input energy reductions require solving the next problem, which is to assure that energy is consumed by the facility only when useful work is being done. When any component of the computing facility is idle, it should draw no energy.

Of all the major components in a computing facility, the computing processors are required to operate at a relatively low duty cycle in order to provide low latency response times to randomly arriving processing requests. Through matching the supply provided to each processor with its processing assignments by using a data-dynamic power supply, it is possible to reduce processor energy consumption to the minimum necessary to maintain processing throughput performance. Even if an idle interval is only one microsecond long, the presented data-dynamic power supply allows that energy to be saved. At a typical processing duty cycle of 35%, this corresponds to a 65% reduction of the processing energy consumption of the facility, all else being equal. This amount of facility energy savings is not possible by only improving efficiency of the existing energy conversion processes.

From the entire facility point of view, this energy reduction is weighted by the fraction of computing activity that the processors provide. Similar reductions may also be available in memory and I/O activities, all of which contribute to further "Greening" of the computing facility.

Additional benefits accrue from this type of energy use reduction. The corresponding reduction in power dissipation means that processor temperature reduces proportionately. This reduces the heat load on the cooling system, causing it to also draw less power. By following a strategy of only generating heat when work is being done, benefits to the entire facility compound, improving reliability while also reducing operating expenses. Computing architects get significant energy savings by operating processors with very little overhead power-up and power-down cycles, much like the processing benefits from rapid context switching. Reducing processor power dissipation is one of the most important problems in computing, including cloud computing, to be solved.

ACKNOWLEDGEMENTS

The author would like to thank the reviewers who have provided many productive and useful comments in improving this paper.

REFERENCES

[1] J. Koomey, *Growth in Data Center Electricity Use 2005 to 2010*, Analytics Press, Oakland, CA, Aug. 1, 2011

[2] "Massive energy cost hidden in wireless cloud boom," http://phys.org/news/2013-04-massive-energy-hidden-wireless-cloud.html

[3] J. Glanz, "Power, Pollution, and the Internet," *New York Times*, Sept. 23, 2012, page A1

[4] L.A. Barroso, U. Hölzle, "The Case for Energy-Proportional Computing," *IEEE Computer*, vol. 40, no. 12, Dec. 2007, pp. 33-37

[5] J. L. Hammond, P. J. R. O'Reilly, *Performance Analysis of Local Computer Networks*, Addison-Wesley, 1986

[6] H. Wong, A Comparison of Intel's 32nm and 22nm Core i5 CPUs: Power, Voltage, Temperature, and Frequency, posted at http://blog.stuffedcow.net/2012/10/intel32nm-22nm-core-i5-comparison/

[7] E. McCune, "Rapid-Transition DC-DC Converter," US Patent applications (2012)

[8] D. Meisner, B. Gold, T. Wenisch, "PowerNap: Eliminating Server Idle Power," *Proceedings of the 14th international conference on Architectural support for programming languages and operating systems*, March 07-11, 2009, Washington, DC, USA

[9] J. Koomey, H.S. Matthews, E. Williams, "Smart Everything: Will Intelligent Systems Reduce Resource Use?" The Annual Review of Environment and Resources, vol. 38, October 2013, pp. 311-343

[10] D. Wyskiel, E. McCune, "Low EMI Printed Circuit Board Design for High Frequency Waveforms", *RF Technology International*, August 2012

[11] Advanced Configuration and Power Interface Specification, available at http://www.acpi.info/