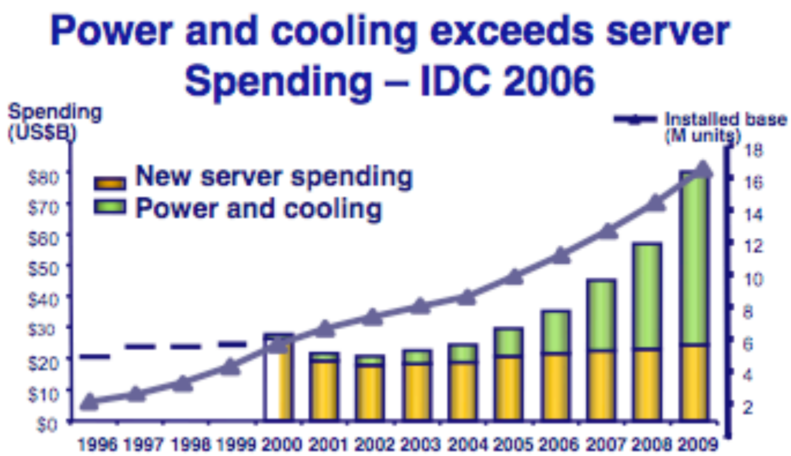
Everywhere Energy-Efficient E-Computing

Wu Feng, feng@cs.vt.edu, 540-267-6545
Department of Computer Science and
Electrical & Computer Engineering
Virginia Tech
Blacksburg, VA 24060

This document outlines a vision for "**green** computing for a **clean** tomorrow" [Feng06]. The first piece of the vision is a bit pedestrian – *holistic energy-efficient computing "in a box"* – but serves as a foundation to a more audacious (tongue-in-cheek) vision of *holistic energy-efficient computing "in a world."*

As recently noted by IDC in an IBM presentation at the Gartner Data Center Summit, December 2006, the annual spending for power and cooling would match the annual budget for new server spending in 2007, as shown in the figure below. In addition to cost, energy-efficient (power-aware) computing can enhance the reliability and availability of ever-increasingly dense computing systems, such as blades; it can also provide additional computational headroom when an institution has reached the limits of its power and cooling infrastructure, particularly when the infrastructure cannot be expanded any further [Feng08].



Holistic Energy-Efficient Computing *in a Box*

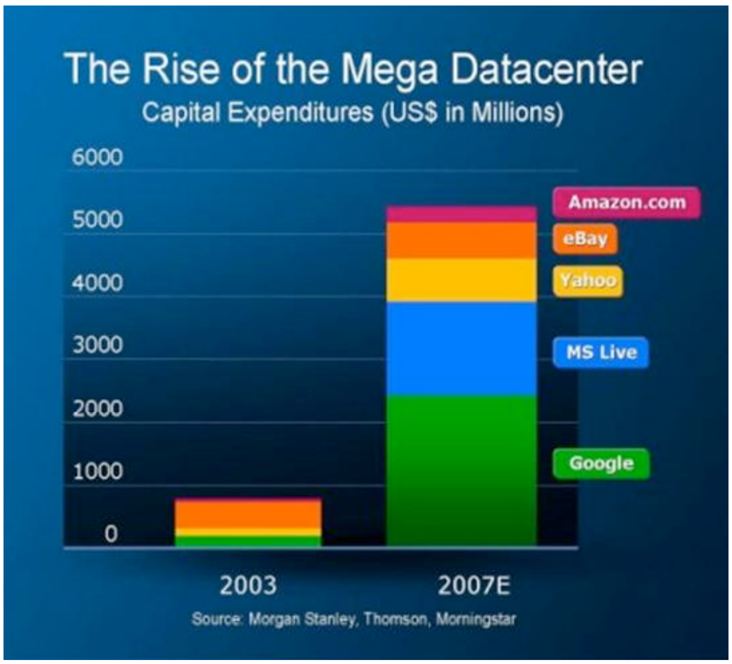
For decades now, we have conducted research in power-aware systems, ranging from mobile systems and laptops to servers and supercomputers. However, for each of the above system "boxes," most of the innovations have focused on optimizing specific subsystems. For example, dynamic voltage & frequency scaling (DVFS) is a processor-centric mechanism for power awareness; for the memory subsystem, there is run-time off-lining & on-lining of memory; storage systems will leverage variable RPM levels to save energy; and so on.

Rather than focus on optimizing a specific subsystem (and observing its effect on the entire system or "box"), we need to take a more holistic approach to dealing with power and co-schedule the various subsystems in an appropriate manner. For instance, if we can automatically identify (coarse) processor-bound phases where the processor should operate at the highest frequency and voltage, we may infer that the memory subsystem, storage subsystem, network subsystem, and video subsystem, for example, can operate at lower power levels.[1]

[1] This does *not* mean that we should eschew optimizations within a specific subsystem. For instance, our preliminary research shows that certain benchmarks *run faster* and *consume less energy* when running on two collocated cores of a dual-core, dual-processor than when running on all four cores.

Holistic Energy-Efficient Computing *in a World*

There are now over *one billion* general-purpose microprocessors in the world (Source: Intel). If we conservatively assume that each microprocessor sits in a "box" that consumes a total of 50 watts of power at idle (after also including the power contributions of a video card, memory, storage, network, and so on), the aggregate power consumption of general-purpose computing boxes sits at 50 GW, or roughly the equivalent to the power output of *25 Hoover Dams* or enough to power approximately 42M homes.[2] Microsoft, Yahoo!, and Google are further accelerating the above power growth by building large datacenters in Quincy, WA (and San Antonio, TX), Wenatchee, WA, and The Dalles, OR, respectively, that will power more than 200,000 (with growth to 800,000 by 2011), 100,000, and 450,000 servers, respectively, as implicitly evidenced in the figure below.



The "Pie in the Sky" Vision: Rather than continue this escalation in "server wars" (or perhaps, more appropriately, "energy wars" due to the excessive power *and* cooling requirements), we should seek to leverage a fraction of the above computational horsepower to create a (distributed) "compute cloud."

OK, so maybe retroactively deploying an overlay network across the above resources is infeasible. Is there another way to ride the electrical power draw of existing computing infrastructure and not have to create an elaborate data-center where for every watt of electric power consumed, 0.7 watts are needed for cooling?

I would argue that there is – the 12,000,000-plus *Xbox LIVE* members and the 19,000,000-plus *Xbox 360* owners, particularly in light of Bill Gates' controversial statement of January 2007 that the Xbox 360 is a general-purpose computer. Having gamers leave their game consoles on for massively distributed cloud computing is easy to incentivize and does not require the massive cooling resources that the aforementioned datacenters require. The challenge will lie in the enabling middleware to harness the above resources.

References

[Feng06] W. Feng, "Green Computing for a Clean Tomorrow," *Dean's Forum on Energy Security and Sustainability*, Virginia Tech, October 2006.

[Feng08] W. Feng, "The Power in Being Low Power," *EDS Innovation Workshop on Next-Generation Efficient Data Centers*, Keynote Talk, Dallas, TX, February 2008.

---

[2] If we were to add in the power consumed by only the embedded processors in embedded systems, we would add 10 GW, or another *5 Hoover Dams*, as ARM is solely responsible for *ten billion* embedded processors, which consume as little as 1 W. Annual unit sales are projected to be *three billion* for 2008.