By Sam Putnam
Written to support Dr. Sean Meyn's successful proposal for a 2014 Faculty Google Research Award in Computer Science

Academia

Title: Opportunities and Challenges for Data Center Demand Response (http://smart.caltech.edu/papers/dcdrsurvey.pdf)
Authors: Wierman, Liu, Liu, Mohsenian-Rad
Bottom Line: Even without speed scaling, data center power usage can be toggled. Recent empirical studies by Lawrence Berkeley National Lab (LBNL) found that 5% of the load can typically be shed in 5 minutes and 10% of the load can be shed in 15 minutes; and that these can be achieved without changes to how the IT workload is handled, i.e. via temperature adjustment and other building management approaches.

Title: Computational Sprinting (http://acg.cis.upenn.edu/papers/hpca12_sprint.pdf)
Authors: Raghavan, Luo, Chandawalla, Papaefthymiou, Pipe, Wenisch, Martin
Bottom Line: Processor chips have the potential to exceed their temperature rating for sub-second bursts of parallel computation. Hence, data centers could fulfill the sub-second frequency band of ancillary service.

Title: Power-Aware Speed Scaling in Processor Sharing Systems (http://users.cms.caltech.edu/~adamw/papers/power-2.pdf)
Authors: Wierman, Andrew, Tang
Bottom line: A simple on/off scheme (static speed when busy and speed 0 when idle) performs within a factor of 2 to a dynamic speed scaling scheme. However, the dynamic speed scaling scheme provides increased robustness to bursty requests and mis-estimation of expected workload. Hence, even if speed scaling proves too costly to processor chips, simple on/off control can provide significant ancillary service, but if one is to consider dynamic speed scaling, robustness should be the goal, not cost minimization.

Title: On the Interaction between Load Balancing and Speed Scaling (http://web.mit.edu/na_li/www/loadbalancing.pdf)
Authors: Chen, Li, Low
Bottom line: As long as the heterogeneity of the system is small, the design of load balancing can be decoupled from speed scaling. However, if the heterogeneity of the system is large, one has to do energy-aware load balancing if energy consumption is the main concern. The number of servers in the system does not affect this equilibrium between speed scaling (ramping the speed of chips to achieve higher performance) and load balancing (distributing jobs across processing resources to increase performance). Hence, if it can be shown that data centers have similar processors, the control problem gets simpler.

Title: Speed Scaling: An Algorithmic Perspective (http://users.cms.caltech.edu/~adamw/papers/speedscaling-chapterdraft.pdf)
Authors: Wierman, Andrew, Lin
Bottom line: Speed scaling can be decoupled from scheduling, because energy-proportional speed scaling provides near optimal performance for three commonly used scheduling policies. However, designers need to consider fairness, in that certain jobs with certain properties will get served more frequently with dynamic speed scaling, especially during long queue times, when the scheduler has its pick of jobs. Simply put, there is a trade-off between fairness and robustness!

Industry

A smart blogger – (http://perspectives.mvdirona.com/):

Bottom line: The U.S. yearly electricity consumption is 100 quadrillion BTU (EIA), and electricity makes up 40% of that, so call it 40 quadrillion BTU = 12k TWh.

Google's energy use - (http://www.nytimes.com/2011/09/09/technology/google-details-and-defends-its-use-of-electricity.html?_r=4&):

Bottom line: Data centers consume less than 0.05% of U.S. electricity consumption. It's still a lot, but definitely under the 2% stated in a couple papers so far. Google's data centers consume 260 MW - about the quarter of the output of a Nuclear plant, working out to only 2 TWh/year of energy usage, less than 1/1000 of a % of the U.S. energy consumption.

Quora - (http://www.quora.com/Is-there-or-will-there-ever-be-a-data-center-to-rival-Googles):

Bottom line: Google owns over a million servers, while the next closest rival is Facebook, perhaps pushing 200k.

Quora (http://www.quora.com/How-many-data-centers-does-Google-have):

Bottom line: We could be getting fooled. Google could have many more data centers than we know about, which are simply undisclosed. Estimates are that Google has at least 10 data centers, with more like 36 data-center-like data centers, and potentially 1400 undisclosed tiny edge data centers

Quora (http://www.quora.com/How-many-people-does-a-typical-Google-data-center-employ-What-sorts-of-skills-do-they-have):

Bottom line: It's not unreasonable to expect that Google would have enough manpower to work the DR system. Each data center employs about 100 people.

Quora (http://www.quora.com/Google/Why-did-not-Google-pick-India-or-China-for-its-first-data-centers-in-Asia):

Bottom line: Stick with studying data centers in America. Google has a small market share and direct investment in data centers in India from foreign investors are not allowed.