# Data Center Efficiency & Renewable Energy

Christopher Peplin (peplin@cmu.edu)

May 29, 2011

## Abstract

Cloud-based software applications have spurred a renewed interest in thin clients - laptops, smartphones and tablet PCs. The resulting increase in the demand for servers in data centers challenges the current electrical grid, but provides an opportunity to pioneer renewable energy, demand response and distributed generation. Organizational issues hinder immediate improvement, but as the infrastructure costs of information technology firms rise, cooperation with utilities will be an increasingly attractive idea.

## 1 Implications of Thin Clients

The software as a service (SaaS) business model is relatively new for application developers and IT firms. Instead of selling packaged software to customers to run on their local computer hardware, SaaS companies provide access to their applications exclusively via the Internet. Any computation required is done by the vendor, on servers they provide. As broadband Internet connection speeds reach more of the population, SaaS has become increasingly popular, and the effect is clear in the hardware purchasing decisions of both businesses and consumers. After years of increasing power in home computers, SaaS shifts the heavy lifting to servers.

In a few ways, SaaS is a throwback to the push for "thin clients" in the 1990s. Thin clients were proposed as low-cost, energy efficient, underpowered machines that would serve as a gateway to applications hosted and run by third parties on Internet servers, or "in the cloud". For whatever reason it failed to take off in the first iteration, the latest push has spurred laptop and smartphone sales, and encouraged new devices like Apple's iPad. Their minimal energy demands are even more attractive in today's energy concious society.

A consequence of the shift is a dramatic increase in the number of servers required to support clients that previously ran their own applications. The type of hardware used for these servers is also quite different that in the last decade. Whereas IBM and Sun Microsystems built their companies with large mainframe servers, they are now sustained by sales and support of volume servers. These are small, inexpensive servers built with commodity parts. Instead of a massive mainframe (a single point of failure for a company), data centers house thousands of rack mounted volume server replacements. A recent report by The Climate Group (2008) found that "[i]f growth continues in line with demand, the world will be using 122 million servers in 2020, up from 18 million today." (The Climate Group 2008, p. 21)

An increase in the number of servers implies a change in the distribution of demand for energy. Processing power is being concentrated in data centers, instead of distributed evenly among all customers. Unless new data centers are strategically planned to optimize energy efficiency, they will burden IT firms with increasing costs, utilities with demand spikes and the environment with increasing emissions. Regulatory agencies, IT firms and public power utilities can cooperate to find an optimal solution that maximizes profit, minimizes emissions, and alleviates some of the problems with young renewable energy sources.

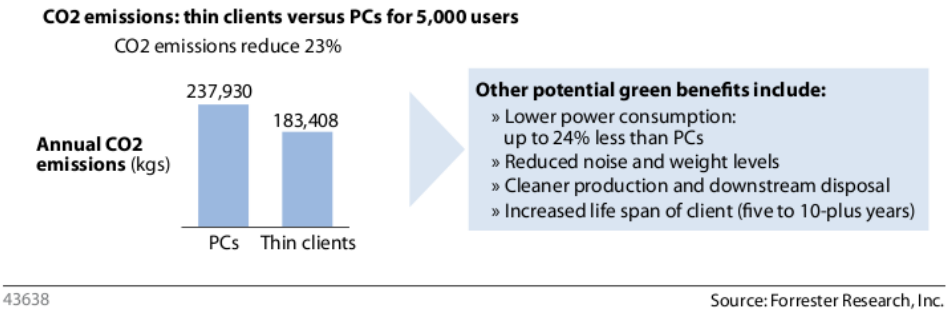


Figure 1: Thin clients use up to 24% less electricity than desktop computers, but some of that efficiency is lost in in the increased demand for servers. Davis (2008)

## 2 Data Center Efficiency

The costs and demands of data centers can be first viewed at the level of an individual deployment. The site infrastructure capital costs of a data center, separate from any application development, alone accounts for 2/3 of IT costs of a data center. (Koomey 2007). These costs as well as emissions are increasing on par with the total number of servers in operation. Figure 2 shows the expected increase in emissions by data centers, broken down by the type of server. Figure 3 shows that the total amortized cost of operating a volume server (1U server) is quickly passing the cost of hardware.

Volume servers are substantially more difficult to run efficiently, primarily due to their number. Tracking the energy use of 5 high-end machines is a simpler affair than tracking the use of thousands of volume servers. It is also common for a volume server to run with a small amount of load if it is dedicated to a task with an uneven workload. Similar to power utilities overprovisioning for reliability, servers are overprovisioned to meet their peak demand. Unfortunately, a volume server at 20% load still uses 60-90% of its peak power consumption. In the previous decade, an expensive mainframe would be made to always run at optimal efficiency, as it represented a significant cost and investment. A single inefficient volume server is easy to write off because of the tiny impact, but the combined cost in a data center is high.

The high energy demands and fast growth of data centers prompted the United States Congress to pass Public Law 109-431 (Congress 2006) requiring the Environmental Protection Agency to report on the current state of energy efficiency in data centers. Among the EPA's findings is the surprising distribution of power within a single data center. Another recent report (The Climate Group 2008) put data to this fact, not spoken of much outside of IT circles:

> Only about half of the energy used by data centres powers the servers and storage; the rest is needed to run back-up, uninterruptible power supplies (UPS) (5%) and cooling systems (45%). (The Climate Group 2008, p. 22)

Despite the large increase in the number of servers, their individually low power requirements and strides in server energy efficiency mostly mitigate any large increase in power demand. The true effect is the indirect power required to cool the server rooms. This puts data centers more in line with industrial loads, where demand is largely for reactive power. In more detail,

> If the power and cooling overhead needed to support the IT equipment are factored in, only about half the power entering the data center is used by the IT equipment. The rest is expended for power conversions, backup power, and cooling. Peak power usage

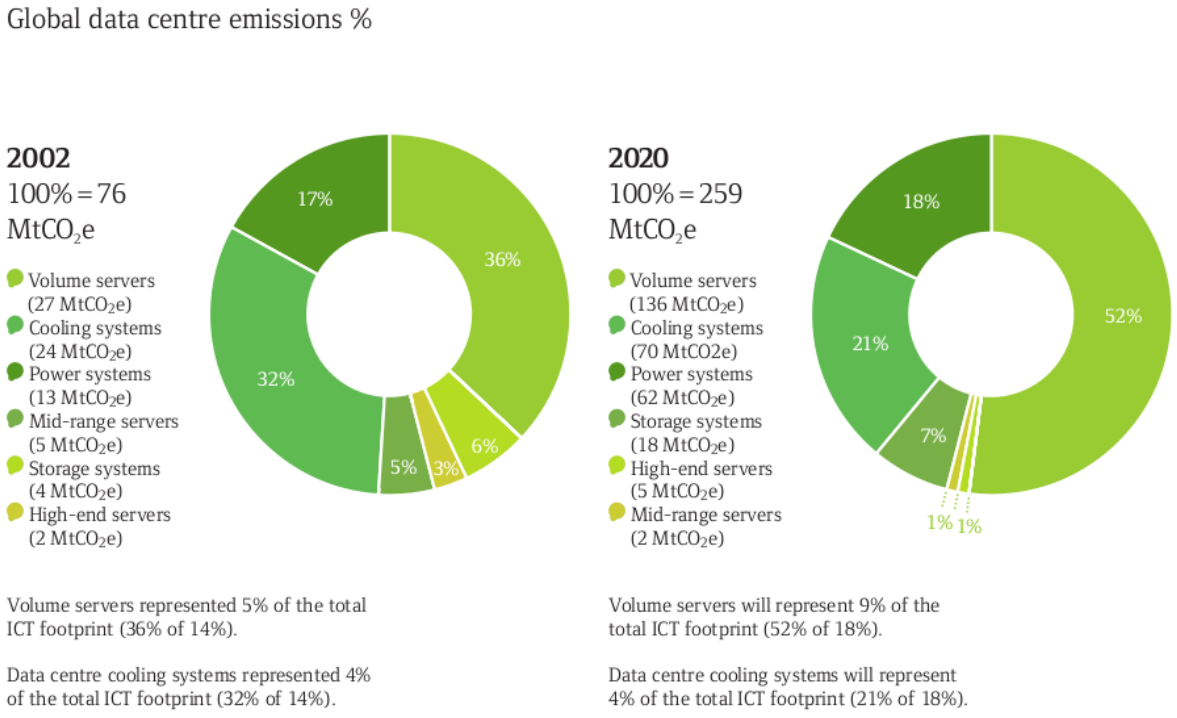


Figure 2: Diagram from The Climate Group (2008) showing projected data center emissions. Volume servers represent the fastest growing segment.

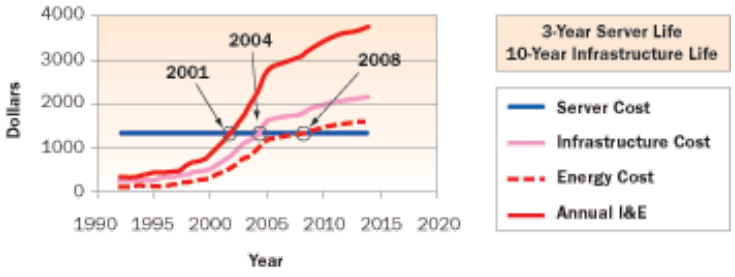


Figure 3: The total cost of ownership of a 1U or volume server is almost four times that of the hardware itself, and it still increasing. EPA (2007)

> for data centers can range from tens of kilowatts for a small facility, to tens of megawatts for the largest data centers. (EPA 2007, p. 21)

Data centers often perform critical functions for business who are loath to experiment with a system that is working well enough to finish the job. The increased costs are bearable with the subsequent increase in profit due to enhanced availability and features. The EPA found resistance to energy efficiency widespread in IT businesses:

> With the increasing importance of digital information, data centers are critical to businesses and government operations. Thus, data center operators are particularly averse to making changes that might increase the risk of down time. Energy efficiency is perceived as a change that, although attractive in principle, is of uncertain value and therefore may not be worth the risk. (EPA 2007, p. 12)

The EPA plans to introduce Energy Star ratings for data centers and their equipment, allowing cost planners to better analyze the total cost of ownership for a server. In combination with increased public pressure for environmental efficiency, the efficiency of data centers should continue to increase.

### 2.1 Virtualization

Even with efficiency improvements, naive data center software will hamper any energy conservation. As consumers are encouraged to shut electrical devices off when not in use, and to hunt down glowing lights and phantom power in their living rooms, a server often burns for them 24/7 to provide on-demand availability.

Server virtualization is a new technique to avoid this symptom. Briefly, virtualization allows many virtual servers be spun up and down dynamically across a smaller number of physical servers. One piece of hardware that was previously leased by a small website can now be used to run 50 small websites with no change for the customer. For example, at night when application activity is low, the servers can be automatically shut down to save power. A small volume website run on a single server can also be scaled up to hundreds in a few seconds if there is a large traffic spike. Utilities such as California's PG&E are offering monetary incentives to remove physical hosts from demand and use virtualization instead. (PG&E 2009) Virtualization allows data centers to aviod the volume server efficiency problem discussed earlier. Virtual servers can be spread across the fewest number of physical machines required, each running at 100% load and thus maximum efficiency.

## 3 Location

Location selection is a critical decision for IT companies, regardless of power, because of the sensitivity and security requirements of data and software running inside. They also must consider the added network latency due to the geographic distance between service and client. At the start of the new Internet age, data centers were typically located closed to metropolitan areas to provide the lowest latency to their prime users. Unfortunately, this proved less than ideal as the number of servers grew:

> It is important to note that two of the cities with the highest concentration of data centers New York City and San Francisco are geographically isolated areas with relatively limited electricity transmission resources. (EPA 2007, p. 64)

The impact of data center location is less important than it was a decade ago. IT firms are increasing interconnecting their networks (peering) to avoid paying transit costs to the Internet backbone providers, and to create direct links to their customers. These changes, both policy wise and physically, have more effect on latency than data center location. The latency within regions of considerable size is

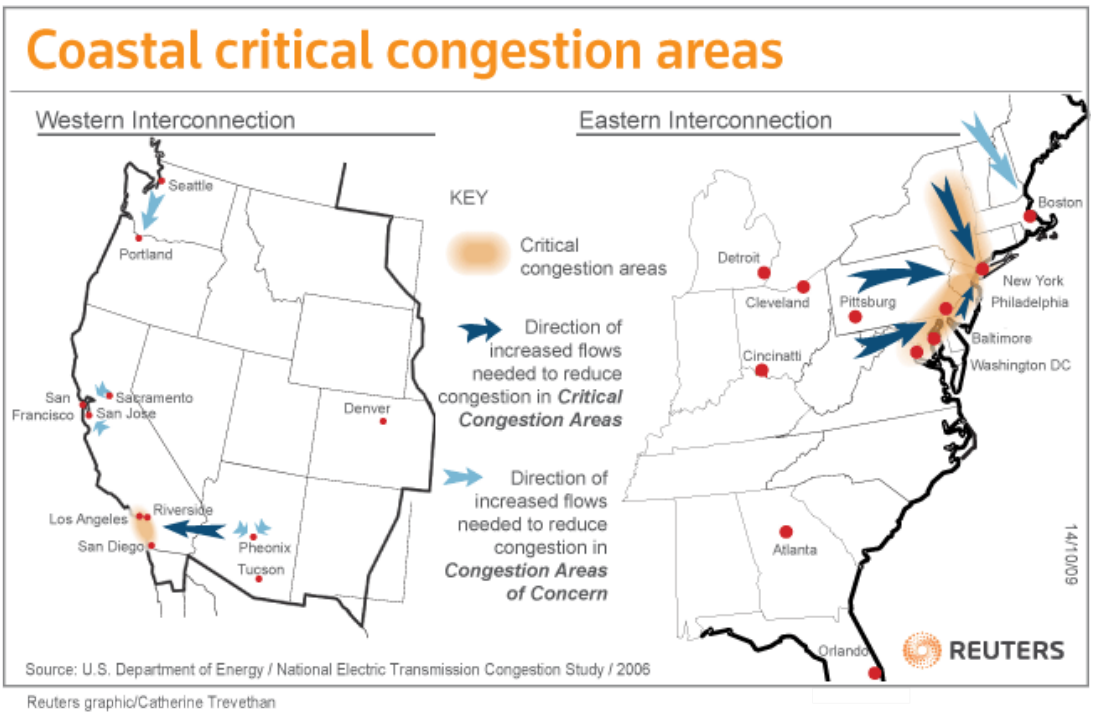


Figure 4: Some of the areas of high congestion concern noted by the U.S. Department of Energy are also popular locations for data centers which contribute additional strain to the grid. EPA (2007)

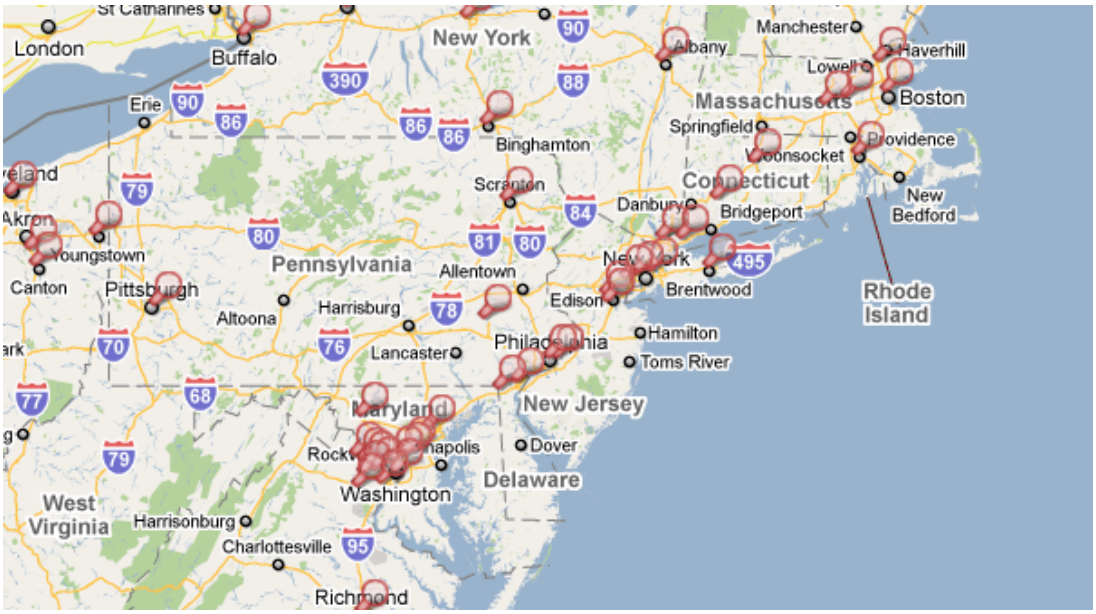


Figure 5: A high density strip of data centers mirrors the strip of transmission congestion in Figure 4. *Map of Data Centers on East Coast* (2010)

comparable enough to allow more flexibility in locating a data center.

Unfortunately, data center locations are notoriously hard to analyze, as companies are understandably protective of the information. The EPA found that comprehensive data about the locations is not readily available because:

- Organizations are concerned about the physical security of these critical infrastructure facilities
- Private data centers are seen as a confidential, strategic asset
- Many data centers are part of larger commercial buildings and campuses and therefore not separately identified or metered
- Data centers have only recently been seen by government agencies as important infrastructure and an indicator of economic activity

(EPA 2007, p. 63)

While increasing the efficiency of individual data centers reduces costs and increases profits, the environmental impact may be negligible. A recent report found that many new data centers are located next to strong but non-renewable power sources.

> [...] efficiency by itself is not green if you are simply working to maximise output from the cheapest and dirtiest energy source available. The US EPA will soon be expanding its EnergyStar rating system to apply to data centres, but similarly does not factor in the fuel source being used to power the data centre in its rating criteria. (Greenpeace 2010, p. 8)

Acting individually, data center operators are encountering the same issue as consumers regarding renewable power sources. The electricity market and Kirchoff's laws dictate that with a general connection to the grid, it cannot be definitively said where the power originated. Data centers must accept a mix of clean and dirty energy sources. (Greenpeace 2010, p. 7)

Data centers are typically located near baseload, reliable sources of power. They are increasingly being built outside of metropolitan areas partly in response to the congestion encountered with the earliest data centers. The lower cost of land, construction also contribute to the shift. (EPA 2007, p. 64) As long as power and latency are met at an adequate cost level, the location of a data center is relatively flexible. This makes them a good candidate for pairing with power sources that are cost effective only in certain regions. An ongoing issue with wind power is that the wind blows strongest in the central United States, exactly where the demand is the lowest and most spread out. Utilities could work with IT firms to move the load, not the generation.

# 4 Common Goals Among Utilities & ITC

The effect of data center growth and location on the power grid and public utilities was one of the items set to be investigated by Congress (2006). The EPA was required to analyze "[...] the potential cost savings and benefits to the energy supply chain through increasing the energy efficiency of data centers and servers, including reducing demand, enhancing capacity, and reducing strain on existing grid infrastructure [...]." (Congress 2006)

The EPA suggested that electric utilities "consider offering incentives for energy-efficient data center facilities and equipment." (EPA 2007, p. 15) Data centers represent a significant portion of the load in the United States, so having an open communications channel between the two could benefit both parties. According to the EPA,

> Nationwide, the energy use estimate for data centers translates into a peak load of more than 7 GW in 2007 (equivalent to the output of about 15 baseload power plants), growing to

> about 12 GW if current growth trends continue. (EPA 2007, p. 58)

As discussed in Section 3, many data centers are located near metropolitan areas and contribute to the transmission congestion problems threatening the United States. According to the EPA, "the peak-load reductions achievable through energy efficiency improvements could play a significant role in relieving capacity constraints in these grids." (EPA 2007, p. 58)

With server virtualization (see 2.1) and given that most major IT firms have multiple data centers among which work is distributed, the demand of any given location could be made highly flexible. When the servers are operating continuously, the load shape of a data center is currently very flat. Some of the efficiency improvements may change that, including the ability to use outside air for cooling and computational load distribution among multiple data centers. Consider a data center handling exclusively non real-time data processing timed to peak at night, located near a wind turbine installation. A coordinated effort between IT firms and power utilities could pair computation load profiles with the best possible generator time profile. (EPA 2007, pp. 60–61)

This flexibility, and the requirement of steady power for data centers, offer two additional opportunities for cooperation - demand response and distributed generation.

### 4.1 Demand Response

Due to their requirements for high reliability, data centers are already equipped to resist short interruptions in power supply. The servers are usually backed with a large, UPS (flywheel or large battery) for the entire data center, or individual batteries on each server. With proper incentives, the utilities could use data centers for demand response by switching servers to their backup power source for short periods of time. The EPA's investigation confirms the feasibility of such an idea:

> Although it is often assumed that data centers are not good candidates for load management because of the critical function they perform, high-reliability data centers are in fact designed to continue operating when the power grid is unavailable using on-site power generation and storage, which suggests that they can also reduce the energy drawn from the grid at times of peak load. (EPA 2007, p. 65)

There would be some additional technical work to implement such a system reliably, as the backup equipment is not intended to be used regularly. The failure rates and repair costs may change if the batteries are discharged often.

These backup batteries also make data centers a good target for renewable power sources like solar and wind, which both have inconsistent time profiles and often generate the most power when it is least needed. Energy storage with compressed air, elevated water or chemical batteries is a promising grid-level solution, and with batteries already in place in data centers, the utility would have smaller startup costs. Fluctuations in generation would matter even less if this was used in combination with a local prime mover generator that was the primary backup for the data center.

### 4.2 Distributed Generation

Some data centers have additional power generation on site for backup and also to alleviate any voltage fluctuations on the grid. Distributed generation is being discussed by utilities, but a large number of private generators are already in operation at data centers. An intelligent connection to the grid is required to bridge between the utilities and these local generators.

Distributed generation applications at data centers include:

- Standby/backup power
- Continuous prime power

- Combined heat and power (CHP) (EPA 2007, p. 81)

  > On-site power generation, whether it is an engine, fuel cell, microturbine, or other prime mover, supports the need for reliable power by protecting against long-term outages beyond what the UPS and battery systems can provide. DG/CHP systems that operate continuously provide additional reliability compared to emergency backup generators that must be started up during a utility outage. (EPA 2007, p. 75)

Renewable generation is often more feasible and less risky at small scale, making distributed generation such as this a good candidate for widespread testing.

# 5 Impediments to Implementation & Conclusion

The biggest impediments to adoption are organizational - the complexity and massive scale of the power grid combined with the unwillingness of private, competitive companies to open up what they consider trade secrets make large scale cooperation unrealistic.

The relationship between data center operators, their financial planners and those interested enough in the environment to start a conversation is difficult to predict.

> Many companies, for example, do not know whether a 50% increase in customer volume would require 25% or 100% more server and data center capacity. As a result, data center facilities can sit half empty, particularly just after construction. In other cases, companies find they complete one data center build program only to find, because of capacity constraints, they must launch a new one almost immediately. (Kaplan, Forrest, and Kindler 2008, p. 7)

Ultimately, what will get companies moving is an opportunity to reduce operating costs of data centers. More and more, they recognize the weight on their balance sheet of the infrastructure demands of SaaS, and only when there is a concise solution to seriously reduce that number will they take action.

# References


Congress, U.S. (Dec. 20, 2006). *An Act, To Study and Promote the Use of Energy Efficient Computer Servers in the United States.* Public Law 109-431, *. 103rd Cong., 2nd sess. H. Doc. H.R. 5646. S.l: S.n., 2007. United States. Cong. House.

Davis, Euan (Mar. 10, 2008). *Green Benefits Put Thin-Client Computing Back On The Desktop Hardware Agenda.* Forrester Research.

EPA (2007). *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431.* U.S. Environmental Protection Agency, Energy Star Program.

Greenpeace (Mar. 30, 2010). *Make IT Green: Cloud Computing and Its Contribution to Climate Change.*

Kaplan, James M., William Forrest, and Noah Kindler (Apr. 17, 2008). *Revolutionizing Data Center Energy Efficiency.* McKinsey & Company.

Koomey, Jonathan G. (Dec. 7, 2007). *Electricity Use and Efficiency of Servers and Data Centers.* YouTube. `http://www.youtube.com/watch?v=s0JoB380xK0`.

*Map of Data Centers on East Coast* (2010). `http://www.datacentermap.com`.

PG&E (Jan. 2009). *Server Virtualization & Consolidation Fact Sheet.* `http://www.pge.com/includes/docs/pdfs/mybusiness/energysavingsrebates/incentivesbyindustry/hightech/C-4166.pdf`.

The Climate Group, Global ESustainability Initiative (June 19, 2008). *SMART 2020: Enabling the Low Carbon Economy in the Information Age.*