Data centers: How to cut carbon emissions *and* costs

Article at a glance
Every large organization depends on vast arrays of servers to run applications, support electronic communications, and provide productivity tools. But building and operating the data center facilities required consumes ever-larger portions of technology budgets and contributes to greenhouse gas emissions. For some information-intensive businesses, data centers represent half of the corporate carbon footprint.

McKinsey's work in this area suggests that companies can double the efficiency of their data centers through more disciplined management, reducing both costs and greenhouse gas emissions. Specifically, companies need to manage their technology assets more aggressively so existing servers can work at much higher utilization levels. They also need to make significant improvements in forward planning of data center needs in order to get the most from their capital spending.

The demand for data center capacity worldwide has led to a sharp rise in IT costs and a steady increase in carbon emissions. A new efficiency metric provides companies with a clear yardstick for measuring progress.

**William Forrest,
James M. Kaplan, and
Noah Kindler**

The modern corporation runs on data. Data centers house the thousands of servers that power applications, provide information, and automate a range of processes. There has been no letup in the demand for data center capacity, and the power consumed as thousands of servers churn away is responsible for rising operating costs and steady growth in worldwide greenhouse gases.

Our work suggests that companies can double the energy efficiency of their data centers through more disciplined management, reducing both costs and greenhouse gas emissions. In particular, companies need to manage technology assets more aggressively so existing servers can work at much higher utilization levels; they also need to improve forecasting of how business demand drives application, server, and data center–facility capacity so they can curb unnecessary capital and operating spending.

Data center efficiency is a strategic issue. Building and operating these centers consumes ever-larger portions of corporate IT budgets, leaving less available for high-priority technology projects. Data center build programs are board-level decisions. At the same time, regulators and external stakeholders are taking keen interest in how companies manage their carbon footprints. Adopting best practices will not only help companies reduce pollution but could also enhance their image as good corporate citizens.

A costly problem

Companies are performing more complex analyses, customers are demanding real-time access to accounts, and employees are finding new, technology-intensive ways to collaborate. As a result, demand for computing, storage, and networking capacity continues to increase even as the economy slows. To cope, IT departments are adding more computing resources, with the number of servers in data centers in the United States growing by about 10 percent a year. At the same time, the number of data centers is rising even more swiftly in emerging markets such as China and India, where organizations are becoming more complex and automating more operations and where, increasingly, outsourced data operations are located. This inexorable demand for computing resources has led to the steady rise of data center capacity worldwide. The growth shows no sign of ending soon, and typically it only moderates during economic down cycles.

This growth has led to a sharp rise in IT costs (Exhibit 1). Data centers typically account for 25 percent of total corporate IT budgets when the costs of facilities, storage devices, servers, and staffing are included. That share will only increase as the number of servers grows and the price of electricity continues its climb faster than revenues and other IT costs. The cost of running these facilities is rising by as much as 20 percent a year, far outpacing overall IT spending, which is increasing at a rate of 6 percent.

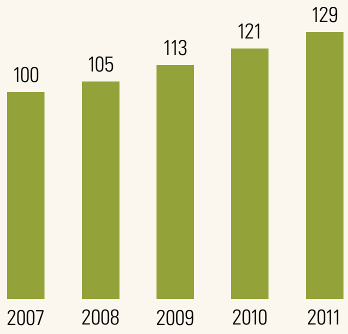
Exhibit 1

Growth threatens profits

For information-intensive industries, growth in data center costs threatens to have a material impact on profitability.

Growth of data center costs, disguised example, $ million

Growth drivers
- 40% growth in transaction volume
- 16% growth in volume of database records
- Aggressive new product-development plans in multiple business units
- High regional demand in Asia

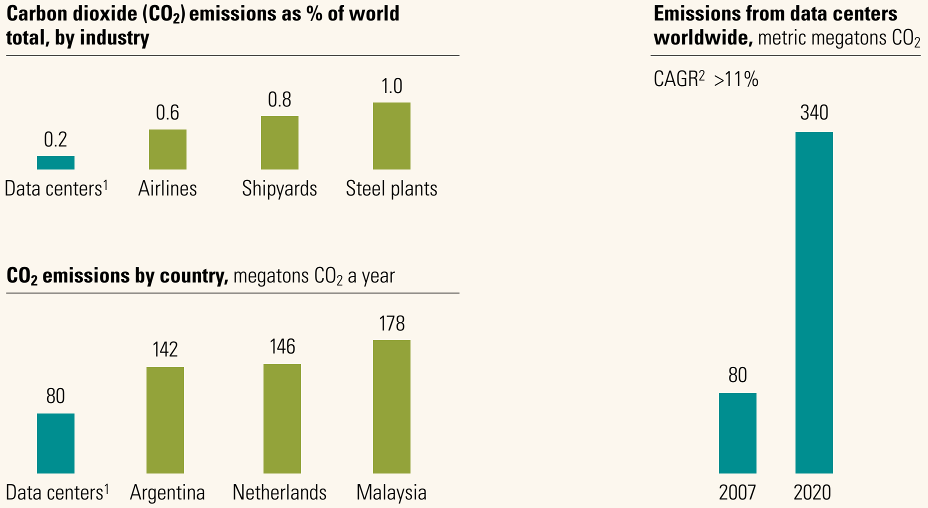


Spending increases on data centers are reshaping the economics of many businesses, particularly those that are intensive users of information such as finance, information services, media, and telecom. The investment required to launch a large-enterprise data center has risen to $500 million, from $150 million, over the past five years. The price tag for the biggest facilities at IT-intensive businesses is approaching $1 billion. This spending is diverting capital from new product development, making some data-intensive products uneconomic, and squeezing margins. The environmental consequences also are stark, as rising power consumption creates a large and expanding carbon footprint. For most service sectors, data centers are a business's number-one source of greenhouse gas emissions. Between 2000 and 2006, the amount of energy used to store and handle data doubled, with the average data facility using as much energy as 25,000 households.

Exhibit 2

Data centers' large carbon footprint

Data centers emissions are now approaching those of Argentina or the Netherlands.


Carbon dioxide (CO_2) emissions as % of world total, by industry

CO_2 emissions by country, megatons CO_2 a year

Emissions from data centers worldwide, metric megatons CO_2

[1] Including custom-designed servers (eg, Google, Yahoo), consumed and embedded carbon.
[2] Compound annual growth rate.

Source: Advanced Micro Devices; *Financial Times*; Gartner; Stanford University; Uptime Institute; McKinsey analysis

Already, the world's 44 million servers consume 0.5 percent of all electricity, with data center emissions now approaching those of countries such as Argentina or the Netherlands. In the United States alone, growth in electricity used by data centers between now and 2010 will be the equivalent of ten new power plants. Without efforts to curb demand, current projections show worldwide carbon emissions from data centers will quadruple by 2020 (Exhibit 2).

Regulators have taken note of these developments and are pressing companies for solutions. In the United States, the Environmental Protection Agency (EPA) has proposed that large data centers use energy meters as a first step toward creating operating-efficiency standards. The European Union, meanwhile, has issued a voluntary code of conduct laying out best practices for running data centers at higher levels of energy efficiency. Government pressure to reduce emissions will likely increase as data center emissions continue to rise.

Far-reaching challenges

In information-intensive organizations, decisions affecting the efficiency of data center operations are made at many levels. Financial traders choose to run complex Monte Carlo analyses, while pharmaceutical researchers decide how much imaging data from clinical trials they want to store. Managers who develop applications decide on how much programming it will take to meet these demands. Those managing server infrastructure decide on equipment purchases. Facilities directors decide on data center locations, power supplies, and the time frame for installing equipment ahead of predicted demand (Exhibit 3).

Exhibit 3

Unintended consequences

Future business decisions may have a large impact on data center demand.

Banking example

New growth		Potential impact
New products, services	Real-time balance information	• Movement from batch processing to real-time processing requires large increase in processing power • Need to upgrade to new and more powerful servers, as well as to change transaction processes
	Increase in online transactions	• Increased small-transaction flows and higher number of interbank transactions • Reduced credit card transaction volume • New applications to process nontraditional payments
	Value-added services for commercial cards	• New applications (ie, integrate transactions in real time, process payroll, checks, interbank transactions) • Increased transaction volume and online processing requiring increase in processing power
	Change in geographic portfolio	• Increase in foreign-currency transactions • New applications to incorporate geographic legal procedures • Increase in round-the-clock utilization of servers, thereby reducing maintenance window
M&A activity	Acquiring a competitor	• Increased transactions • Potential new supply of data centers from acquired card provider; need to integrate various data center technologies
Macroeconomic shock	Economic downturn	• Idle or underutilized capacity in data centers • Increased number of collections agencies and terminals • Increase in cost per transaction due to lower volume • Need to reduce cost

These decisions are usually made in isolation. A sales manager may choose to change transactions from overnight to real-time clearing, or a financial analyst may want to store multiple copies of historical data—without thinking about the impact on data center costs. Applications developers rarely think of fine-tuning their work to use the fewest number of servers, or of creating design applications that can be shared across servers. Managers buying them may select those with the lowest prices or those with which they're most familiar. But these servers may waste electricity or space in data centers. Frequently, managers purchase excess devices to guarantee capacity in the most extreme usage scenarios, creating large amounts of excess capacity. And managers often build facilities with excess floor space and high cooling capacity to meet extreme demands or all expansion contingencies.

Multiplied across an organization, these decisions result in both costs and environmental implications. In many cases, existing servers could be decommissioned and plans for new ones shelved without diminishing the ability of companies to manage data. This can be accomplished using well-known techniques, including virtualization, which in effect share capacity by seeking unused portions of servers to run pieces of applications. But this doesn't always happen, since no one executive has end-to-end accountability. Within the organization, managers optimize for their own interests, resulting in the inefficiency observed in most data centers. In many instances, only a single software application runs on a server.

Within one media company, almost a third of the nearly 500 servers we analyzed had utilization rates below 3 percent, and nearly two-thirds were below 10 percent. This company used none of the number of readily available management tools for tracking use. On a global basis, we estimate daily server utilization generally tops out at 5 to 10 percent, wasting both energy and employed capital. A common response from data center managers is that the servers exist to provide capacity for extreme situations, such as the shopping crunch on the day before Christmas. However, the data generally don't support this assertion: when average utilization is very low, so is peak usage. Furthermore, sprawling data facilities are sometimes only half occupied by servers and related equipment, suggesting hundreds of millions of dollars in wasted capital spending. Even in data centers that companies report as full, a walk down the aisles often reveals significant gaps within racks of servers, where equipment has been decommissioned.

In the absence of true cost analysis, ***overbuilding, overdesign, and inefficiency*** *become the rule*

These mismatches arise in part from the difficulty of forecasting data center requirements. Operating time frames are one problem. Data centers take 2 years or more to design and build and are expected to last at least 12 years, so capacity is added well in advance of the actual needs of business units. At the same time there is an incomplete understanding of how business decisions affect one another, how they translate into the need for new applications, and how much server capacity is needed to meet demand. Many companies, for example, would have difficulty forecasting whether a 50 percent increase in customer demand would require 25 percent or 100 percent more server and data center capacity. In the extreme, we have seen some facilities lie half empty years after opening; other companies complete one data center only to find they need to build a new one almost immediately.

Considering that data centers have become a costly asset, accountability for financial performance is poor. Financial and management responsibility for facilities often falls to real-estate managers who have little technical expertise and few insights into how IT relates to core business issues. Those managing server operations, meanwhile, rarely see data on crucial operating spending such as electricity consumption or the true cost of the real estate occupied by the IT equipment. Conversely, when IT managers decide on additional applications or new servers, they sometimes use only basic metrics such as initial hardware costs or software licenses. Figuring out the real costs requires consideration of facilities operations and leases, electricity use, support, and depreciation. These charges can multiply the initial purchase cost of a server by a factor of four or five. Combined with the siloed decision making and accountability issues discussed above, extra servers are often added as insurance with little discussion of cost trade-offs or the needs of the business. In the absence of true cost analysis, overbuilding, overdesign, and inefficiency become the rule.

Reforming data center operations

When we began our research, we expected to find that building new energy-efficient data centers would offer the best hope of reducing their cost and carbon footprint. New facilities could take advantage of current technologies that make use of natural cooling and of power supplies that produce fewer emissions. However, we also learned that the most dramatic reductions in cost and carbon emissions come from improving the low efficiency of data centers that companies already operate. Through better management of assets, more accountable management, and setting clear goals for reducing energy costs and carbon emissions, most companies can double IT energy efficiency by 2012 and halt the growth of their data centers' greenhouse gas emissions. Indeed, the greenest data center is the one that you don't have to build.

Manage IT assets aggressively

One large company's approach illustrates the potential gains from a disciplined use of existing servers and facilities. The company's plans for meeting its 2010 information needs called for increasing the server base and building a new data center to house these servers and other IT equipment. Its board already had approved the plans, which represented a significant amount of the company's capital spending that year. It has since radically revised them. More than 5,000 rarely used servers will be shut down. Virtualization of some 3,700 applications—15 percent of the companywide total—will allow a reduction in the number of active servers to 20,000, from 25,000. The company has also replaced some older servers with those that use electricity 20 percent more efficiently.

These changes enabled the company to shelve its data center expansion plans, saving $305 million in capital investment costs. Projected operating expenses (for fewer servers and less power consumption) are set to fall by $45 million, to $75 million. Taking into account decommissioning and virtualization, the average server will run at 9.1 percent of capacity rather than the current 5.6 percent. The company will still meet its growing data needs, but reduction in power demands means that CO_2 emissions over the next four years will be cut to 341,000 tons, from 591,000 tons.

Companies can also save by better managing rising demand for data. Business units should review policies on how much data should be retained and whether to scale back some intensive data analyses. Some transaction computation can be deferred, thus reducing peak use of servers, and not all corporate information requires extensively backed-up disaster recovery capabilities.

Get better information

Better forecasting and planning is the foundation for data center efficiency gains. Companies should track how their forecasts for data needs vary with real demand and then provide bonuses to those business units that are able to minimize deviations. Data center managers should incorporate the most complete view of future trends in their models, such as organizational growth and business cycles. Input from data centers, applications architects, and facilities operators can be used to improve these models. One

global communications company instituted a planning process that included developing scenarios for data growth for each of its business units. While the company eventually concluded that it needed additional capacity, a large portion of future needs were met using existing assets, saving 35 percent in planned capital expenses.

True accounting for costs

In many organizations, data centers are treated as buckets waiting to be filled, rather than as scarce and expensive resources. To combat this tendency, companies can adopt true cost of ownership (TCO) accounting when estimating costs for new servers or additional applications and data. Lifetime costs of running applications and operating servers are rarely included in spending decisions by business units, software developers, or IT managers. Building them in upfront can help limit excess demand.

One financial institution adopted TCO accounting for all the applications that supported its trading and investment-banking products. It resulted in first-ever discussions with IT managers about which investments in software applications were actually producing adequate returns, providing a road map for reducing areas of overinvestment and IT inefficiency. Multiplying these conversations across business units can bring much-needed discipline to decisions that ultimately have an impact on data center costs.

Centralize responsibility

Managing these kinds of changes may be difficult. Many in large organizations don’t recognize the cost of data. Demands for data center services arrive from across the enterprise. Responsibility for meeting those demands falls across IT departments (including operations and application development), facilities planners, shared services groups, and corporate real-estate functions. There is no single standard for reporting the costs.

We suggest a new governance model for managing data center needs, with full responsibility and accountability falling to the CIO. Under such a regime, the CIO would have much greater visibility into the data demands of business units and could enforce requirements that energy consumption and facilities costs figure into return-on-investment calculations for new data projects requiring additional servers or software applications. We also suggest that CIOs employ a new metric for measuring progress (see sidebar, "Improving data center efficiency"). With sharpened accountability, the CIO will have

Improving data center efficiency

As part of a program for data center improvement, we suggest employing a new metric: corporate average data center efficiency (CADE). Similar to the United States' Corporate Average Fuel Economy (CAFE) mileage standards, CADE takes into account the energy efficiency of facilities, their utilization rates, and the level of utilization of servers in the data center. Multiplying these factors together yields the overall efficiency of the data center, or CADE (exhibit). Companies that reduce costs and emissions will improve their data centers' CADE scores. That's similar to how better mileage bolsters CAFE ratings in the auto industry.

To establish targets for improvement, we set five CADE tiers. Those centers operating at CADE level one are the weakest in terms of efficiency; most organizations initially are likely to fall within the lower ranges. Shutting down underused servers, employing virtualization, and using space within facilities more efficiently will raise CADE scores. CADE also allows companies to benchmark across data center facilities, or against those of rivals, as well as set and track performance goals for managers.

Exhibit

A new efficiency metric

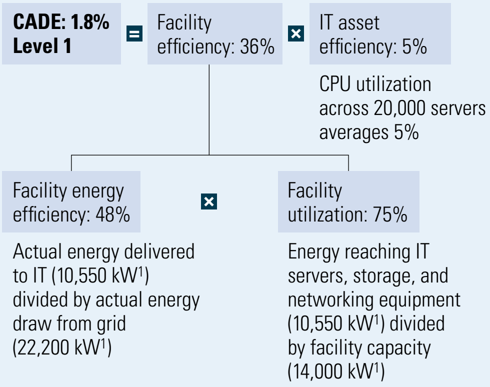
A new metric for data center efficiency provides companies with a clear yardstick for measuring progress.

Corporate average data center efficiency (CADE)

Range of efficiency for data centers (level 1 = least efficient, level 5 = most efficient)

1	0–5%	Typical range today
2	5–10%	Range to target by 2012
3	10–20%	
4	20–40%	
5	>40%	

Example: actual data center currently on level 1 of CADE



CADE: 1.8%
Level 1
=
Facility efficiency: 36%
×
IT asset efficiency: 5%

CPU utilization across 20,000 servers averages 5%

Facility energy efficiency: 48%
×
Facility utilization: 75%

Actual energy delivered to IT (10,550 kW[1]) divided by actual energy draw from grid (22,200 kW[1])

Energy reaching IT servers, storage, and networking equipment (10,550 kW[1]) divided by facility capacity (14,000 kW[1])

Year-1 improvements underway to enable doubling of CADE by 2012

- Remove 4,000 dead servers. Average CPU utilization increases by 10%.
- Virtualize 8,000 servers on 4 to 1 ratio with 50% utilization. Further increase average CPU utilization from 15% to 20%.
- Implement full suite of industry best practices. Facility energy efficiency increases to ~53%.
- Defer new data center construction. Allow 15% annual organic IT growth to increase facility utilization.

[1] Kilowatts.

Source: Uptime Institute; McKinsey analysis greater incentive to seek improvements, such as virtualization and better use of existing facilities. Since this model vests much broader responsibility with the CIO for key business decisions, it needs full support from the CEO and a change in organizational mindset that business unit requests for added data center capacity won't always be met.

In addition, the CIO should publicly commit to the goal of doubling data center energy efficiency as a way of encouraging improvements and of helping the business to get ahead of regulatory or other stakeholder pressures. Our analysis indicates that nearly every company is capable of doubling its data center energy efficiency over the next three or four years using currently available techniques and technology. Achieving this goal requires stronger data center management, better planning, and increased accountability.

Data center inefficiency is widespread, and it has become a major concern worldwide. But there is significant opportunity for improvement. Following the recommendations outlined above can create a virtuous cycle of better data center management leading to more efficient energy use, lower costs, and steady reductions in carbon emissions. MoBT

The authors would like to recognize the important contributions of Kenneth Brill and The Uptime Institute to the development of this article and its recommendations. The Institute provided insight based on many years of experience, as well as proprietary data and analysis.

William Forrest (William_Forrest@McKinsey.com) is an associate principal in McKinsey's Chicago office.

James Kaplan (James_Kaplan@McKinsey.com), a principal in the New York office, leads the technology infrastructure practice for the global IT group.

Noah Kindler (Noah_Kindler@McKinsey.com) is a consultant in the New York office.