nytimes.com

Data Centers Waste Vast Amounts of Energy, Belying Industry Image

James Glanz

27-35 minutes

Data centers are filled with servers, which are like bulked-up desktop computers, minus screens and keyboards, that contain chips to process data. Ethan Pines for The New York Times

— Jeff Rothschild's machines at Facebook had a problem he knew he had to solve immediately. They were about to melt.

The company had been packing a 40-by-60-foot rental space here with racks of computer servers that were needed to store and process information from members' accounts. The electricity

pouring into the computers was overheating Ethernet sockets and other crucial components.

Thinking fast, Mr. Rothschild, the company's engineering chief, took some employees on an expedition to buy every fan they could find — "We cleaned out all of the Walgreens in the area," he said — to blast cool air at the equipment and prevent the Web site from going down.

That was in early 2006, when Facebook had a quaint 10 million or so users and the one main server site. Today, the information generated by nearly one billion people requires outsize versions of these facilities, called data centers, with rows and rows of servers spread over hundreds of thousands of square feet, and all with industrial cooling systems.

They are a mere fraction of the tens of thousands of data centers that now exist to support the overall explosion of digital information. Stupendous amounts of data are set in motion each day as, with an innocuous click or tap, people download movies on iTunes, check credit card balances through Visa's Web site, send Yahoo e-mail with files attached, buy products on Amazon, post on Twitter or read newspapers online.

A yearlong examination by The New York Times has revealed that this foundation of the information industry is sharply at odds with its image of sleek efficiency and environmental friendliness.

Most data centers, by design, consume vast amounts of energy in an incongruously wasteful manner, interviews and documents show. Online companies typically run their facilities at maximum capacity around the clock, whatever the demand. As a result, data centers can waste 90 percent or more of the electricity they pull off the grid, The Times found.

To guard against a power failure, they further rely on banks of generators that emit diesel exhaust. The pollution from data centers has increasingly been cited by the authorities for violating clean air regulations, documents show. In Silicon Valley, many data centers appear on the state government's Toxic Air Contaminant Inventory, a roster of the area's top stationary diesel polluters.

Worldwide, the digital warehouses use about 30 billion watts of electricity, roughly equivalent to the output of 30 nuclear power plants, according to estimates industry experts compiled for The Times. Data centers in the United States account for one-quarter to one-third of that load, the estimates show.

"It's staggering for most people, even people in the industry, to understand the numbers, the sheer size of these systems," said Peter Gross, who helped design hundreds of data centers. "A single data center can take more power than a medium-size town."

Energy efficiency varies widely from company to company. But at the request of The Times, the consulting firm McKinsey & Company analyzed energy use by data centers and found that, on average, they were using only 6 percent to 12 percent of the electricity powering their servers to perform computations. The rest was essentially used to keep servers idling and ready in case of a surge in activity that could slow or crash their operations.

A server is a sort of bulked-up desktop computer, minus a screen and keyboard, that contains chips to process data. The study sampled about 20,000 servers in about 70 large data centers spanning the commercial gamut: drug companies, military contractors, banks, media companies and government agencies.

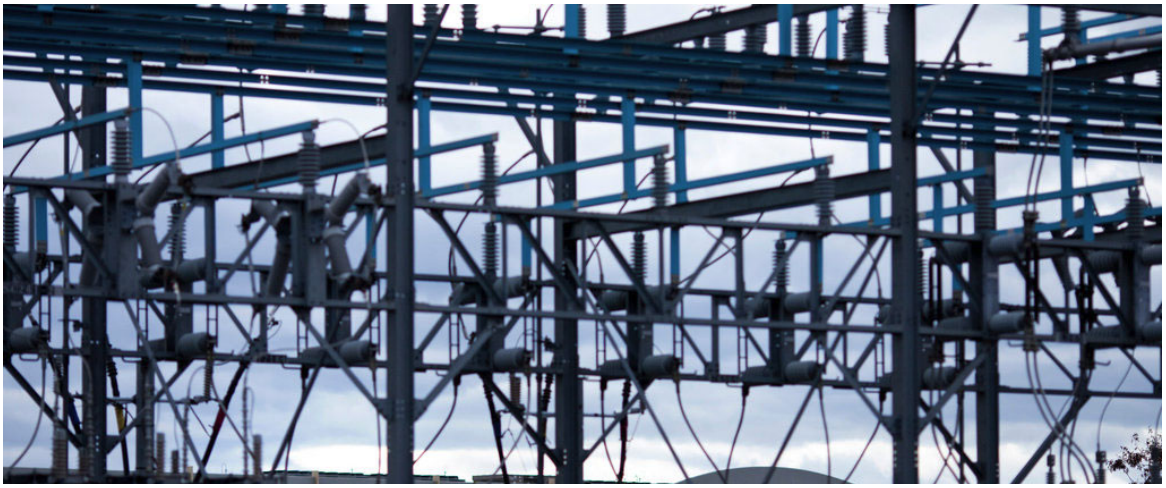
“This is an industry dirty secret, and no one wants to be the first to say mea culpa,” said a senior industry executive who asked not to be identified to protect his company’s reputation. “If we were a manufacturing industry, we’d be out of business straightaway.”

These physical realities of data are far from the mythology of the Internet: where lives are lived in the “virtual” world and all manner of memory is stored in “the cloud.”

The inefficient use of power is largely driven by a symbiotic relationship between users who demand an instantaneous response to the click of a mouse and companies that put their business at risk if they fail to meet that expectation.

Even running electricity at full throttle has not been enough to satisfy the industry. In addition to generators, most large data centers contain banks of huge, spinning flywheels or thousands of lead-acid batteries — many of them similar to automobile batteries — to power the computers in case of a grid failure as brief as a few hundredths of a second, an interruption that could crash the servers.

“It’s a waste,” said Dennis P. Symanski, a senior researcher at the Electric Power Research Institute, a nonprofit industry group. “It’s too many insurance policies.”



A NATURAL PAIRING *A data center in Ashburn, Va., seen past a Dominion Virginia Power substation serving it. Worldwide, such centers use the rough equivalent of the output of 30 nuclear power plants. Brendan Smialowski for The New York Times*

At least a dozen major data centers have been cited for violations of air quality regulations in Virginia and Illinois alone, according to state records. Amazon was cited with more than 24 violations over a three-year period in Northern Virginia, including running some of its generators without a basic environmental permit.

A few companies say they are using extensively re-engineered software and cooling systems to decrease wasted power. Among them are Facebook and Google, which also have redesigned their hardware. Still, according to recent disclosures, Google's data centers consume nearly 300 million watts and Facebook's about 60 million watts.

Many of these solutions are readily available, but in a risk-averse industry, most companies have been reluctant to make wholesale change, according to industry experts.

Improving or even assessing the field is complicated by the secretive nature of an industry that is largely built around accessing other people's personal data.

For security reasons, companies typically do not even reveal the locations of their data centers, which are housed in anonymous buildings and vigilantly protected. Companies also guard their technology for competitive reasons, said Michael Manos, a longtime industry executive. "All of those things play into each other to foster this closed, members-only kind of group," he said.

That secrecy often extends to energy use. To further complicate any assessment, no single government agency has the authority to track the industry. In fact, the federal government was unable to determine how much energy its own data centers consume, according to officials involved in a survey completed last year.

The survey did discover that the number of federal data centers grew from 432 in 1998 to 2,094 in 2010.

To investigate the industry, The Times obtained thousands of pages of local, state and federal records, some through freedom of information laws, that are kept on industrial facilities that use large amounts of energy. Copies of permits for generators and information about their emissions were obtained from environmental agencies, which helped pinpoint some data center locations and details of their operations.

In addition to reviewing records from electrical utilities, The Times also visited data centers across the country and conducted hundreds of interviews with current and former employees and contractors.

Some analysts warn that as the amount of data and energy use continue to rise, companies that do not alter their practices could eventually face a shake-up in an industry that has been prone to major upheavals, including the bursting of the first Internet bubble in the late 1990s.

“It’s just not sustainable,” said Mark Bramfitt, a former utility executive who now consults for the power and information technology industries. “They’re going to hit a brick wall.”

Bytes by the Billions

Wearing an FC Barcelona T-shirt and plaid Bermuda shorts, Andre Tran strode through a Yahoo data center in Santa Clara

where he was the site operations manager. Mr. Tran's domain — there were servers assigned to fantasy sports and photo sharing, among other things — was a fair sample of the countless computer rooms where the planet's sloshing tides of data pass through or come to rest.

Aisle after aisle of servers, with amber, blue and green lights flashing silently, sat on a white floor punctured with small round holes that spit out cold air. Within each server were the spinning hard drives that store the data. The only hint that the center was run by Yahoo, whose name was nowhere in sight, could be found in a tangle of cables colored in the company's signature purple and yellow.

"There could be thousands of people's e-mails on these," Mr. Tran said, pointing to one storage aisle. "People keep old e-mails and attachments forever, so you need a lot of space."

This is the mundane face of digital information — player statistics flowing into servers that calculate fantasy points and league rankings, snapshots from nearly forgotten vacations kept forever in storage devices. It is only when the repetitions of those and similar transactions are added up that they start to become impressive.

Each year, chips in servers get faster, and storage media get denser and cheaper, but the furious rate of data production goes a notch higher.

Jeremy Burton, an expert in data storage, said that when he worked at a computer technology company 10 years ago, the most data-intensive customer he dealt with had about 50,000 gigabytes in its entire database. (Data storage is measured in bytes. The letter N, for example, takes 1 byte to store. A gigabyte is a billion bytes of information.)

Today, roughly a million gigabytes are processed and stored in a data center during the creation of a single 3-D animated movie, said Mr. Burton, now at EMC, a company focused on the management and storage of data.

INSURANCE *A row of backup generators, inside white housings, lines the back exterior of the Facebook data center in Prineville, Ore. They are to ensure service even in the event of a power failure. Steve Dykes for The New York Times*

Just one of the company's clients, the New York Stock Exchange, produces up to 2,000 gigabytes of data per day that must be stored for years, he added.

EMC and the International Data Corporation together estimated that more than 1.8 trillion gigabytes of digital information were created globally last year.

"It is absolutely a race between our ability to create data and our ability to store and manage data," Mr. Burton said.

About three-quarters of that data, EMC estimated, was created by ordinary consumers.

With no sense that data is physical or that storing it uses up space and energy, those consumers have developed the habit of sending huge data files back and forth, like videos and mass e-mails with photo attachments. Even the seemingly mundane actions like running an app to find an Italian restaurant in Manhattan or a taxi in Dallas requires servers to be turned on and ready to process the information instantaneously.

The complexity of a basic transaction is a mystery to most users: Sending a message with photographs to a neighbor could involve a trip through hundreds or thousands of miles of Internet conduits and multiple data centers before the e-mail arrives across the street.

“If you tell somebody they can’t access YouTube or download from Netflix, they’ll tell you it’s a God-given right,” said Bruce Taylor, vice president of the Uptime Institute, a professional organization for companies that use data centers.

To support all that digital activity, there are now more than three million data centers of widely varying sizes worldwide, according to figures from the International Data Corporation.

Nationwide, data centers used about 76 billion kilowatt-hours in 2010, or roughly 2 percent of all electricity used in the country that year, based on an analysis by Jonathan G. Koomey, a research fellow at Stanford University who has been studying data center energy use for more than a decade. DatacenterDynamics, a London-based firm, derived similar figures.

The industry has long argued that computerizing business transactions and everyday tasks like banking and reading library books has the net effect of saving energy and resources. But the paper industry, which some predicted would be replaced by the

computer age, consumed 67 billion kilowatt-hours from the grid in 2010, according to Census Bureau figures reviewed by the Electric Power Research Institute for The Times.

Direct comparisons between the industries are difficult: paper uses additional energy by burning pulp waste and transporting products. Data centers likewise involve tens of millions of laptops, personal computers and mobile devices.

Chris Crosby, chief executive of the Dallas-based Compass Datacenters, said there was no immediate end in sight to the proliferation of digital infrastructure.

“There are new technologies and improvements,” Mr. Crosby said, “but it still all runs on a power cord.”

‘Comatose’ Power Drains

Engineers at Viridity Software, a start-up that helped companies manage energy resources, were not surprised by what they discovered on the floor of a sprawling data center near Atlanta.

Viridity had been brought on board to conduct basic diagnostic testing. The engineers found that the facility, like dozens of others they had surveyed, was using the majority of its power on servers that were doing little except burning electricity, said Michael Rowan, who was Viridity’s chief technology officer.

A senior official at the data center already suspected that something was amiss. He had previously conducted his own informal survey, putting red stickers on servers he believed to be “comatose” — the term engineers use for servers that are plugged in and using energy even as their processors are doing little if any computational work.

“At the end of that process, what we found was our data center had a case of the measles,” said the official, Martin Stephens,

during a Web seminar with Mr. Rowan. “There were so many red tags out there it was unbelievable.”

The Viridity tests backed up Mr. Stephens’s suspicions: in one sample of 333 servers monitored in 2010, more than half were found to be comatose. All told, nearly three-quarters of the servers in the sample were using less than 10 percent of their computational brainpower, on average, to process data.

The data center’s operator was not some seat-of-the-pants app developer or online gambling company, but LexisNexis, the database giant. And it was hardly unique.

LOW-TECH AMID HIGH-TECH *A backup diesel generator at a large computer data center, one of six in the room. Combined, they could provide enough power for a community of 7,000 homes. Richard Perry/The New York Times*

In many facilities, servers are loaded with applications and left to run indefinitely, even after nearly all users have vanished or new versions of the same programs are running elsewhere.

“You do have to take into account that the explosion of data is what aids and abets this,” said Mr. Taylor of the Uptime Institute. “At a certain point, no one is responsible anymore, because no one, absolutely no one, wants to go in that room and unplug a server.”

Kenneth Brill, an engineer who in 1993 founded the Uptime Institute, said low utilization began with the field’s “original sin.”

In the early 1990s, Mr. Brill explained, software operating systems that would now be considered primitive crashed if they were asked to do too many things, or even if they were turned on and off. In response, computer technicians seldom ran more than one application on each server and kept the machines on around the clock, no matter how sporadically that application might be called upon.

So as government energy watchdogs urged consumers to turn off computers when they were not being used, the prime directive at data centers became running computers at all cost.

A crash or a slowdown could end a career, said Michael Tresh, formerly a senior official at Viridity. A field born of cleverness and audacity is now ruled by something else: fear of failure.

“Data center operators live in fear of losing their jobs on a daily basis,” Mr. Tresh said, “and that’s because the business won’t back them up if there’s a failure.”

In technical terms, the fraction of a computer’s brainpower being used on computations is called “utilization.”

McKinsey & Company, the consulting firm that analyzed utilization figures for The Times, has been monitoring the issue since at least 2008, when it published a report that received little notice outside the field. The figures have remained stubbornly

low: the current findings of 6 percent to 12 percent are only slightly better than those in 2008. Because of confidentiality agreements, McKinsey is unable to name the companies that were sampled.

David Cappuccio, a managing vice president and chief of research at Gartner, a technology research firm, said his own recent survey of a large sample of data centers found that typical utilizations ran from 7 percent to 12 percent.

“That’s how we’ve overprovisioned and run data centers for years,” Mr. Cappuccio said. “ ‘Let’s overbuild just in case we need it’ — that level of comfort costs a lot of money. It costs a lot of energy.”

Servers are not the only components in data centers that consume energy. Industrial cooling systems, circuitry to keep backup batteries charged and simple dissipation in the extensive wiring all consume their share.

In a typical data center, those losses combined with low utilization can mean that the energy wasted is as much as 30 times the amount of electricity used to carry out the basic purpose of the data center.

Some companies, academic organizations and research groups have shown that vastly more efficient practices are possible, although it is difficult to compare different types of tasks.

The National Energy Research Scientific Computing Center, which consists of clusters of servers and mainframe computers at the Lawrence Berkeley National Laboratory in California, ran at 96.4 percent utilization in July, said Jeff Broughton, the director of operations. The efficiency is achieved by queuing up large jobs and scheduling them so that the machines are running nearly full-out, 24 hours a day.

A company called Power Assure, based in Santa Clara, markets a technology that enables commercial data centers to safely power down servers when they are not needed — overnight, for example.

But even with aggressive programs to entice its major customers to save energy, Silicon Valley Power has not been able to persuade a single data center to use the technique in Santa Clara, said Mary Medeiros McEnroe, manager of energy efficiency programs at the utility.

“It’s a nervousness in the I.T. community that something isn’t going to be available when they need it,” Ms. McEnroe said.

ENERGY HUNGRY *Row after row after row of servers, at data centers around the world, perform the functions that constitute the cloud. They consume vast amounts of electricity, often wastefully. Richard Perry/The New York Times*

The streamlining of the data center done by Mr. Stephens for LexisNexis Risk Solutions is an illustration of the savings that are possible.

In the first stage of the project, he said that by consolidating the work in fewer servers and updating hardware, he was able to shrink a 25,000-square-foot facility into 10,000 square feet.

Of course, data centers must have some backup capacity available at all times and achieving 100 percent utilization is not possible. They must be prepared to handle surges in traffic.

Mr. Symanski, of the Electric Power Research Institute, said that such low efficiencies made sense only in the obscure logic of the digital infrastructure.

“You look at it and say, ‘How in the world can you run a business like that,’ ” Mr. Symanski said. The answer is often the same, he said: “They don’t get a bonus for saving on the electric bill. They get a bonus for having the data center available 99.999 percent of the time.”

The Best-Laid Plans

In Manassas, Va., the retailing colossus Amazon runs servers for its cloud amid a truck depot, a defunct grain elevator, a lumberyard and junk-strewn lots where machines compress loads of trash for recycling.

The servers are contained in two Amazon data centers run out of three buildings shaped like bulky warehouses with green, corrugated sides. Air ducts big enough to accommodate industrial cooling systems sprout along the rooftops; huge diesel generators sit in rows around the outside.

The term “cloud” is often generally used to describe a data center’s functions. More specifically, it refers to a service for leasing computing capacity. These facilities are primarily powered from the national grid, but generators and batteries are nearly always present to provide electricity if the grid goes dark.

The Manassas sites are among at least eight major data centers Amazon operates in Northern Virginia, according to records of Virginia's Department of Environmental Quality.

The department is on familiar terms with Amazon. As a result of four inspections beginning in October 2010, the company was told it would be fined $554,476 by the agency for installing and repeatedly running diesel generators without obtaining standard environmental permits required to operate in Virginia.

Even if there are no blackouts, backup generators still emit exhaust because they must be regularly tested.

After months of negotiations, the penalty was reduced to $261,638. In a "degree of culpability" judgment, all 24 violations were given the ranking "high."

Drew Herdener, an Amazon spokesman, agreed that the company "did not get the proper permits" before the generators were turned on. "All of these generators were all subsequently permitted and approved," Mr. Herdener said.

The violations came in addition to a series of lesser infractions at one of Amazon's data centers in Ashburn, Va., in 2009, for which the company paid $3,496, according to the department's records.

Of all the things the Internet was expected to become, it is safe to say that a seed for the proliferation of backup diesel generators was not one of them.

Terry Darton, a former manager at Virginia's environmental agency, said permits had been issued to enough generators for data centers in his 14-county corner of Virginia to nearly match the output of a nuclear power plant.

"It's shocking how much potential power is available," said Mr. Darton, who retired in August.

No national figures on environmental violations by data centers are available, but a check of several environmental districts suggests that the centers are beginning to catch the attention of regulators across the country.

Over the past five years in the Chicago area, for example, the Internet powerhouses Savvis and Equinix received violation notices, according to records from the Illinois Environmental Protection Agency. Aside from Amazon, Northern Virginia officials have also cited data centers run by Qwest, Savvis, VeriSign and NTT America.

Despite all the precautions — the enormous flow of electricity, the banks of batteries and the array of diesel generators — data centers still crash.

Amazon, in particular, has had a series of failures in Northern Virginia over the last several years. One, in May 2010 at a facility in Chantilly, took businesses dependent on Amazon's cloud offline for what the company said was more than an hour — an eternity in the data business.

Pinpointing the cause became its own information glitch.

Amazon announced that the failure "was triggered when a vehicle crashed into a high-voltage utility pole on a road near one of our data centers."

As it turns out, the car accident was mythical, a misunderstanding passed from a local utility lineman to a data center worker to Amazon headquarters. Instead, Amazon said that its backup gear mistakenly shut down part of the data center after what Dominion Virginia Power said was a short on an electrical pole that set off two momentary failures.

Mr. Herdener of Amazon said the backup system had been redesigned, and that “we don’t expect this condition to repeat.”

The Source of the Problem

Last year in the Northeast, a $1 billion feeder line for the national power grid went into operation, snaking roughly 215 miles from southwestern Pennsylvania, through the Allegheny Mountains in West Virginia and terminating in Loudon County, Va.

The work was financed by millions of ordinary ratepayers. Steven R. Herling, a senior official at PJM Interconnection, a regional authority for the grid, said the need to feed the mushrooming data centers in Northern Virginia was the “tipping point” for the project in an otherwise down economy.

Data centers in the area now consume almost 500 million watts of electricity, said Jim Norvelle, a spokesman for Dominion Virginia Power, the major utility there. Dominion estimates that the load could rise to more than a billion watts over the next five years.

Data centers are among utilities’ most prized customers. Many utilities around the country recruit the facilities for their almost unvarying round-the-clock loads. Large, steady consumption is profitable for utilities because it allows them to plan their own power purchases in advance and market their services at night, when demand by other customers plummets.

Mr. Bramfitt, the former utility executive, said he feared that this dynamic was encouraging a wasteful industry to cling to its pedal-to-the-metal habits. Even with all the energy and hardware pouring into the field, others believe it will be a challenge for current methods of storing and processing data to keep up with the digital tsunami.

Some industry experts believe a solution lies in the cloud: centralizing computing among large and well-operated data centers. Those data centers would rely heavily on a technology called virtualization, which in effect allows servers to merge their identities into large, flexible computing resources that can be doled out as needed to users, wherever they are.

One advocate of that approach is Mr. Koomey, the Stanford data center expert. But he said that many companies that try to manage their own data centers, either in-house or in rental spaces, are still unfamiliar with or distrustful of the new cloud technology. Unfortunately, those companies account for the great majority of energy usage by data centers, Mr. Koomey said.

Others express deep skepticism of the cloud, saying that the sometimes mystical-sounding belief in its possibilities is belied by the physicality of the infrastructure needed to support it.

Using the cloud “just changes where the applications are running,” said Hank Seader, managing principal for research and education at the Uptime Institute. “It all goes to a data center somewhere.”

Some wonder if the very language of the Internet is a barrier to understanding how physical it is, and is likely to stay. Take, for example, the issue of storing data, said Randall H. Victora, a professor of electrical engineering at the University of Minnesota who does research on magnetic storage devices.

“When somebody says, ‘I’m going to store something in the cloud, we don’t need disk drives anymore’ — the cloud is disk drives,” Mr. Victora said. “We get them one way or another. We just don’t know it.”

Whatever happens within the companies, it is clear that among consumers, what are now settled expectations largely drive the need for such a formidable infrastructure.

“That’s what’s driving that massive growth — the end-user expectation of anything, anytime, anywhere,” said David Cappuccio, a managing vice president and chief of research at Gartner, the technology research firm. “We’re what’s causing the problem.”