We start our discussion of energy use by data centers and its economic and ecological impact with a

brief analysis of the concept of energy-proportional systems. This is a very important concept because a

strategy for resource management in a computing cloud is to concentrate the load on a subset of servers

and switching the rest of the servers to a standby mode whenever possible [7]. This strategy aims to

reduce power consumption and, implicitly, the cost of providing computing and storage services; we

analyze this subject in depth in Chapter 6.

The operating efficiency of a system is captured by an expression of “performance per Watt of

power.” It is widely reported that, during the last two decades, the performance of computing systems

has increased much faster than their operating efficiency; for example, during the period 1998–2007,

the performance of supercomputers increased by 7,000% whereas their operating efficiency increased

by only 2,000%.

In an ideal world, the energy consumed by an idle system should be near zero and should grow

linearly with the system load. In real life, even machines whose power requirements scale linearly, use

more than half the power when idle than they use at full load (see Figure 3.6) [42].

Energy-proportional systems could lead to large savings in energy costs for computing clouds. An

energy-proportional system consumes no power when idle, very little power under a light load, and

gradually more power as the load increases. By definition, an ideal energy-proportional system is always

operating at 100% efficiency. Humans are a good approximation of an ideal energy proportional system;

human energy consumption is about 70W at rest and 120W on average on a daily basis and can go as

high as 1,000–2,000W during a strenuous, short effort [42].

Different subsystems of a computing system behave differently in terms of energy efficiency. Many

processors have reasonably good energy-proportional profiles, but significant improvements in memory and disk subsystems are necessary. The processors used in servers consume less than one-third of their

peak power at very lowload and have a dynamic range15 of more than 70% of peak power; the processors

used in mobile and/or embedded applications are better in this respect. According to [42] the dynamic

power range of other components of a system is much narrower: less than 50% for dynamic random

access memory (DRAM), 25% for disk drives, and 15% for networking switches.

A number of proposals have emerged for energy-proportional networks; the energy consumed by

such networks is proportional to the communication load. For example, in [6] the authors argue that a

data center network based on a flattened butterfly topology is more energy and cost efficient than one

using a different type of interconnect.

High-speed channels typically consist of multiple serial lanes with the same data rate; a physical unit

is stripped across all the active lanes. Channels commonly operate plesiochronously16 and are always

on, regardless of the load, because they must still send idle packets to maintain byte and lane alignment across the multiple lanes. An example of an energy-proportional network is InfiniBand, discussed in

Section 3.1.

Energy saving in large-scale storage systems is also of concern. A strategy to reduce energy consumption

is to concentrate the workload on a small number of disks and allow the others to operate in

a low-power mode. One of the techniques to accomplish this task is based on replication. A replication

strategy based on a sliding window is reported in [364]; measurement results indicate that it performs

better than LRU, MRU, and LFU17 policies for a range of file sizes, file availability, and number of

client nodes, and the power requirements are reduced by as much as 31%.

Another technique is based on data migration. The system in [158] uses data storage in virtual nodes

managed with a distributed hash table; the migration is controlled by two algorithms, a short-term optimization

algorithm, used for gathering or spreading virtual nodes according to the daily variation of the

workload so that the number of active physical nodes is reduced to aminimum, and a long-term optimization

algorithm, used for coping with changes in the popularity of data over a longer period (e.g., a week).

The energy consumption of large-scale data centers and their costs for energy and for cooling are

significant now and are expected to increase substantially in the future. In 2006, the 6,000 data centers

in the United States reportedly consumed 61 109 KWh of energy, 1.5% of all electricity consumption

in the country, at a cost of $4.5 billion [364].

The predictions have been dire: The energy consumed by the data centers was expected to double

from 2006 to 2011; peak instantaneous demand was expected to increase from 7 GW in 2006 to 12 GW

in 2011, requiring the construction of 10 new power plants. The energy consumption of data centers

and the network infrastructure is predicted to reach 10, 300 TWh/year18 in 2030, based on 2010 levels

of efficiency [295]. These increases are expected in spite of the extraordinary reduction in energy

requirements for computing activities; over the past 30 years the energy efficiency per transistor on a

chip has improved by six orders of magnitude.

The effort to reduce energy use is focused on the computing, networking, and storage activities of

a data center. A 2010 report shows that a typical Google cluster spends most of its time within the

10–50% CPU utilization range; there is a mismatch between server workload profile and server energy

efficiency [6]. A similar behavior is also seen in the data center networks; these networks operate in a

very narrow dynamic range, and the power consumed when the network is idle is significant compared

to the power consumed when the network is fully utilized.

Many proposals argue that dynamic resource provisioning is necessary to minimize power consumption.

Twomain issues are critical for energy saving: the amount of resources allocated to each application

and the placement of individual workloads. For example, a resource management framework combining

a utility-based dynamic virtual machine provisioning manager with a dynamic VM placement manager

to minimize power consumption and reduce SLA violations is presented in [358].

The support for network-centric content consumes a very large fraction of the network bandwidth;

according to the CISCO VNI forecast, consumer traffic was responsible for around 80% of bandwidth

use in 2009 and is expected to grow at a faster rate than business traffic. Data intensity for various

activities ranges from 20MB/minute for HDTV streaming to 10MB/minute for standard TV streaming, 1.3MB/minute for music streaming, 0.96 MB/minute for Internet radio, 0.35 MB/minute for Internet

browsing, and 0.0025 MB/minute for ebook reading [295].

The same study reports that if the energy demand for bandwidth is 4 Watts-hour per MB19 and

if the demand for network bandwidth is 3.2 GB/day/person or 2,572 EB/year for the entire world

population, then the energy required for this activity will be 1, 175 GW. These estimates do not count

very high-bandwidth applications that may emerge in the future, such as 3D TV, personalized immersive

entertainment such as Second Life, or massively multiplayer online games.

The power consumption required by different types of human activities is partially responsible for

the world’s greenhouse gas emissions. According to a recent study [295], the greenhouse gas emissions

due to data centers are estimated to increase from 116   106 tons of CO2 in 2007 to 257 tons in 2020,

due primarily to increased consumer demand. Environmentally opportunistic computing is a macroscale

computing idea that exploits the physical and temporal mobility of modern computer processes.

A prototype called a Green Cloud is described in [376].