Energy Implications of Big Data

(An Article in the Springer Encyclopedia of Big Data Technologies)
[Version of February 20, 2018]

Behrooz Parhami
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560, USA
parhami@ece.ucsb.edu

Synonyms — Energy-Efficient Data Storage and Processing; Energy Wall; Green Big Data

Definition of entry — Data storage, communication, and processing consumes energy and big data requires a correspondingly big energy budget, necessitating more attention and effort to ensure energy-efficiency

Cross-references — Big Data and Exascale Computing; Computer Architecture for Big Data; Emerging Hardware Technologies; Parallel Processing with Big Data

Until fairly recently, developments in the field of computer architecture (Parhami 2005) were focused around computation running time and hardware complexity, and how the two can be traded off against each other in various designs. With the advent of mobile devices and terascale, petascale, and, soon, exascale computing, energy consumption emerged as a major design factor that overshadowed the older concerns to some extent. Key driving forces for research into energy efficiency were battery life in compact mobile devices with smallish batteries, energy costs for operators of supercomputer centers, and the difficulties of heat dissipation in both mobile devices and mainframe/data-center installations. We are thus motivated to seek extreme energy economy measures to make the processing of large data sets feasible within power budgets that are practical for both cloud data centers and mobile devices used at the cloud’s edges.

Energy-Efficiency Trends

Early digital computers of the 1940s performed about one operation per Watt-hour (Wh) of energy consumed. Those machines, which resembled factory equipment, were bulky, unreliable, and user-unfriendly, problems that overshadowed their energy-inefficiency. The exponential decline of the per-computation energy requirement, or exponential rise of the number of operations performed per
unit of energy (Denning and Lewis 2017), over the decades is known as Koomey’s law (Koomey et al. 2011). The exponential rise of computational capability per kWh of energy is depicted in Fig. 1.

A simple back-of-the-envelope calculation based on the expected $10^{25}$ bytes of data produced per day by 2020 leads to the conclusion that even at the optimistic $10^{18}$ computations per kWh, energy requirements of big-data processing will be prohibitive.

Fig. 1. Trends in energy efficiency, expressed as computations per kWh (Koomey et al. 2011).
Causes of Energy Dissipation

As energy efficiency of mainstream processors that control servers, desktops, laptops, and smartphones has improved, research into ultra-low-power digital circuits has paved the way for even greater reduction in energy requirements per computation. Physical laws dictate that any non-linear transformation of the kinds performed by standard AND-OR logic gates must necessarily expend energy and a lower bound for this energy is known to be on the order of $kT$, where $k \approx 1.38 \times 10^{-23}$ is the Boltzmann constant and $T$ is the operating temperature in Kelvin (Landauer 1961). This lower bound has come to be known as Landauer’s principle (Bennett 2003). Although, in principle, this minimum should be reachable, current computing devices are a great deal less efficient in the amount of energy they dissipate.

The energy used by a computing device is the product of average power drawn and the computation time. This relationship suggests two ways of reducing the energy consumption: Using lower-power technologies (e.g., Jaberipur et al. 2018) and using faster algorithms. Use of faster algorithms is, of course, highly desirable, as it has performance implications as well. So, there is a great deal of ongoing research to find ever-faster algorithms for computational problems of interest. Power consumption per device has been going down due to Dennard scaling, a law devised in 1974 that maintains the relative constancy of MOSFET power density, as transistors get smaller, a trend predicted by Moore’s law (Brock and Moore 2006, McMenamin 2013).

The power dissipated by computing devices consists of static and dynamic components. Static power is the power used by devices that are in the off state and ideally should not be drawing any power; however, leakage and other factors lead to some power waste, which is increasing in significance as we use smaller and denser circuits. Dynamic power (Benini et al. 2001) is proportional to the operating voltage squared, the operating frequency, circuit capacitance, and activity (prevalence of signal-value changes). Low-voltage circuits use much less power, but the closer we get to the threshold voltage, the slower and less reliable circuit operation becomes (Kaul et al. 2012). Thus, there is a limit to power savings by reducing the operating voltage.
Reasons for energy waste include glitching (Devadas et al. 1992), unnecessary signal-value transitions that can be avoided by suitable encodings (e.g., Musoll et al. 1998), and clocking of circuitry that is not involved in performing useful computations (Wu et al. 2000). Practical notions in this regard include energy-proportional computing, that is, the desirable property that the energy used is commensurate with the amount of useful work performed (Barroso and Holzle 2007), and dynamic frequency and/or voltage scaling (Nakai et al 2005) in an effort to save energy when application requirements do not demand the highest performance level.

A data center’s energy efficiency is influenced by three factors attributed to the facility, server conversion efficiency, and efficiency at the level of circuits and devices (Barroso et al. 2013). Only about a third of the energy is used for productive computation (Holzle 2017). Generally speaking, large-scale installations provide more opportunities for energy fine-tuning, thus underlining the importance of data centers and cloud-computing in handling big-data loads. More on this later.

**Near-Threshold Multi-Core Computing**

Parallel processing is nearly a necessity for handling of large data volumes, on performance grounds. It is also helpful for reducing power consumption. Using \( m \) processors/cores, each with \( 1/m \) the speed, is more energy-efficient than using a single high-performance processor. Thus, if an application offers enough parallelism to use multiple processors/cores efficiently, it will lead to higher performance and/or lower energy consumption. Beginning with dual-core chips in the mid-2000s, multi-core chips (Gepner and Kowalik 2006) have allowed performance scaling to continue unabated, despite Moore’s Law, in its original formulation, becoming invalid (Mack 2011).

While use of a large number of simple cores carries energy efficiency benefits, one must also pay attention to the effect of inherently sequential parts of a computation which give rise to speed-up limits predicted by Amdahl’s law (Parhami 1999). Perhaps a judicious mix of powerful or “brawny” and simple or “wimpy” cores (Holzle 2010) in a multi-core chip, possibly also including application-specific performance boosters (Hardavellas et al. 2011), can help mitigate this problem.
The ultimate in energy efficiency is achieved when the processors/cores operate very near the threshold voltage (Dreslinski et al 2010, Gautschi 2017, Hubner and Silvano 2016). As operating voltage is scaled down, performance is reduced nearly linearly, whereas active energy dissipation goes down quadratically. Therefore, performance per unit of energy improves. When the static or leakage power is taken into account, the savings become less pronounced, but still significant. Near-threshold computing has pitfalls in terms of computation stability and system reliability, so there is a sensitive trade-off to be performed. There is also the issue of giving up too much in sequential performance in the hopes of regaining some of it via parallel processing (Mudge and Holzle 2010).

**Memory/Storage Energy Requirements**

Dynamic random-access memory (DRAM) chips account for a significant fraction of the power used in a computer’s electronic parts. Static RAM (SRAM) is even more energy-intensive, but because SRAM sizes tend to be relatively much smaller, they do not contribute as much to the energy budget of a system. Precise modeling of DRAM energy consumption has been attempted in an effort to determine the exact share of power used, importance of power dissipation within the memory cells vs. peripheral circuitry for decoding, access, and content refreshing, and expected changes as DRAM technology is scaled down and goes through other generational changes (Vogelsang 2010). These methods will guide energy reduction in main memories in the near future. For the longer term, a variety of main memory technologies, some of which enable inherently lower-power operation, are being investigated (Xie 2011).

Disk memories are both slow and energy-intensive, but have the advantage of permanence, low cost per bit, and non-volatility. For small to modest memory sizes, replacement of disks with solid-state memories is feasible, but for very large data volumes, such a replacement is still impractically expensive. So, in the short term, we have to live with disk memory shortcomings. Even though a single disk unit has a long mean time to failure (MTTF), when many thousands of disks are involved in a storage warehouse, failures are inevitable. Modern data storage facilities make use of redundant
arrays of independent disks, RAID for short (Schroeder and Gibson 2007), to mitigate the slowness and low reliability of disk units. A variety of methods have been devised to make disk memories and attendant units, such as disk caches and RAID controllers, that make a disk array look like a single disk and reconstruct damaged or lost data upon disk failures, more energy-efficient (Pinheiro and Bianchini 2014).

As storage devices are increasingly networked to gain the reliability, availability, and scale benefits of distributed access, communication overhead and its attendant energy requirements must be factored in when comparing different organizations with regard to energy efficiency.

**Communication Energy Requirements**

Like computation and storage, communication also dissipates energy. In fact, one of the most challenging aspects of managing big-data applications is to decide how the energy budget should be allocated to the three aspects of computing, storage, and communication. Data replication often reduces communication costs during access but may impose a nontrivial overhead for keeping replicas consistent, and storage of previously computed results may obviate the need for recomputation.

Reducing communication costs has been studied extensively in connection with energy-limited mobile and sensor networks (Heinzelman et al. 2000), but many of the techniques are applicable more broadly. For example, expending local computation to aggregate data before transmitting a smaller volume of data is a generally useful method (Krishnamachari et al. 2002).

Given their current prevalence and future importance, energy efficiency consideration for on-chip networks and those used in data centers are of paramount importance. Networks-on-chip (Benini and De Micheli 2002, Pande et al. 2005) are key components of modern multi-core chips and their importance will increase as the number of cores is scaled up. Data-center networks have been subjects of numerous studies, given their role in the energy costs of a typical data center (Hammadi and Mhamdi 2014). One example consists of the proposal for energy-proportional data-center networks (Abts et al. 2010).
Beyond networks-on-chip and data-center networks, we must also be concerned with energy efficiency in broad-area networks, the Internet in particular (Bolta et al. 2011). Whereas the introduction of low-power techniques in the design of routers and other electronic units is necessary for energy-efficient networking, it is not enough. We also need energy savings at the physical network layer and in the algorithms and protocols used for computation, routing, resource management, testing, and fault tolerance.

**Energy Considerations in the Cloud**

The cloud consists of computational, storage, and communication resources. Yet, to achieve energy efficiency in the cloud, we must go beyond separate energy optimizations for the aforementioned components. Such optimizations are challenging, given the scale of the cloud (Assuncao et al. 2015, Feller et al. 2015). As of 2017, servers collectively use about 200 TWh of energy, which is comparable to energy use in all of Mexico. Google alone uses as much energy as the city of San Francisco. As large as this value is, it is dwarfed compared with the energy required by user’s laptop or other computing equipment at the edges of the cloud.

As for the energy used in data-center installations, it consisted until fairly recently of three nearly equal parts devoted to mechanical cooling, IT equipment, and everything else (lighting, back-up power supply, etc.). So, the energy used for the actual computation was multiplied by a factor of 3.0, implying a 200% overhead. More efficient modern data centers reduced this factor to 1.8, for an overhead of 80%. Now, we can go as low as 10% overhead through a variety of energy saving schemes, including the application of machine learning to adjust a building’s cooling strategy based on information about the applicable parameters (Holzle 2017).

Servers have undergone similar efficiency improvements. Earlier, some 50% of energy went to waste, even before power got to the actual circuits. By eliminating this waste, we are now at about 10% overhead relative to the actual energy used by the circuits. The circuit energy has been going down by 20% per year in recent years (post-Moore’s-Law era). Factors leading to this reduction are
smaller circuits, clock-gating (disabling the parts of the circuits not in use, so that they don’t draw energy), frequency scaling, and specialization (tailoring the circuits to computations). In the latter domain, Google’s hardware optimized for machine learning uses 0.2 MW of power, compared with 2.5 MW needed by a general-purpose supercomputer doing the same job (Holzle 2017).

As a whole, the IT industry uses about 2% of the world’s energy, which is of the same order as the amount used by airlines. Because modern data centers are way more efficient than local server installations, moving to the cloud will reduce the energy consumption associated with computations by some 87%. Operating data center with exclusive use of renewable energy is now possible, which constitutes a major benefit. The lower hardware redundancy needed to ensure reliable operation also saves energy. For example, Gmail uses 1% redundancy in hardware resources, whereas a typical local e-mail server installation needs at least duplication to avoid service disruptions. This makes the user-side energy consumption even more important. Fortunately, with the move away from desktops and laptops to tablets and smartphones, user-side energy consumption is also plummeting.

**Future Directions**

Work in proceeding in improving energy-efficiency in all aspects of the big-data environment discussed in the preceding sections. At the circuit and chip levels, ultralow-power technologies are being devised and evaluated. Examples are found in the fields of optical, biologically-inspired, adiabatic, and reversible circuits. The ultimate goal might be doing away with batteries and other power sources altogether (Eisenberg 2010), for a large fraction of the big-data ecosystem. Significant reduction in the dissipated energy will also obviate the need for complex cooling strategies and their associated hardware and software costs (Kaushik and Nahrstedt 2010).

The future of big data is closely tied to the future of cloud computing, as the economy of scale provided by the cloud is necessary for the successful deployment of big-data applications (Hashem et al. 2015, Wu et al. 2013). A promising direction in energy monitoring and optimization, with no need for human intervention, is the use of machine learning strategies. Such techniques have already been
applied to power management at the circuit level (Dhiman and Rosing 2006) and to data-center climate control systems (Holzle 2017), but a great deal more can be done.

References


