

E

Energy Implications of Big Data



Behrooz Parhami
Department of Electrical and Computer
Engineering, University of California, Santa
Barbara, CA, USA

Synonyms

[Energy wall](#); [Energy-efficient data storage and processing](#); [Green big data](#)

Definition

Data storage, communication, and processing consume energy, and big data requires a correspondingly big energy budget, necessitating more attention and effort to ensure energy efficiency.

Overview

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, en-

ergy 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.

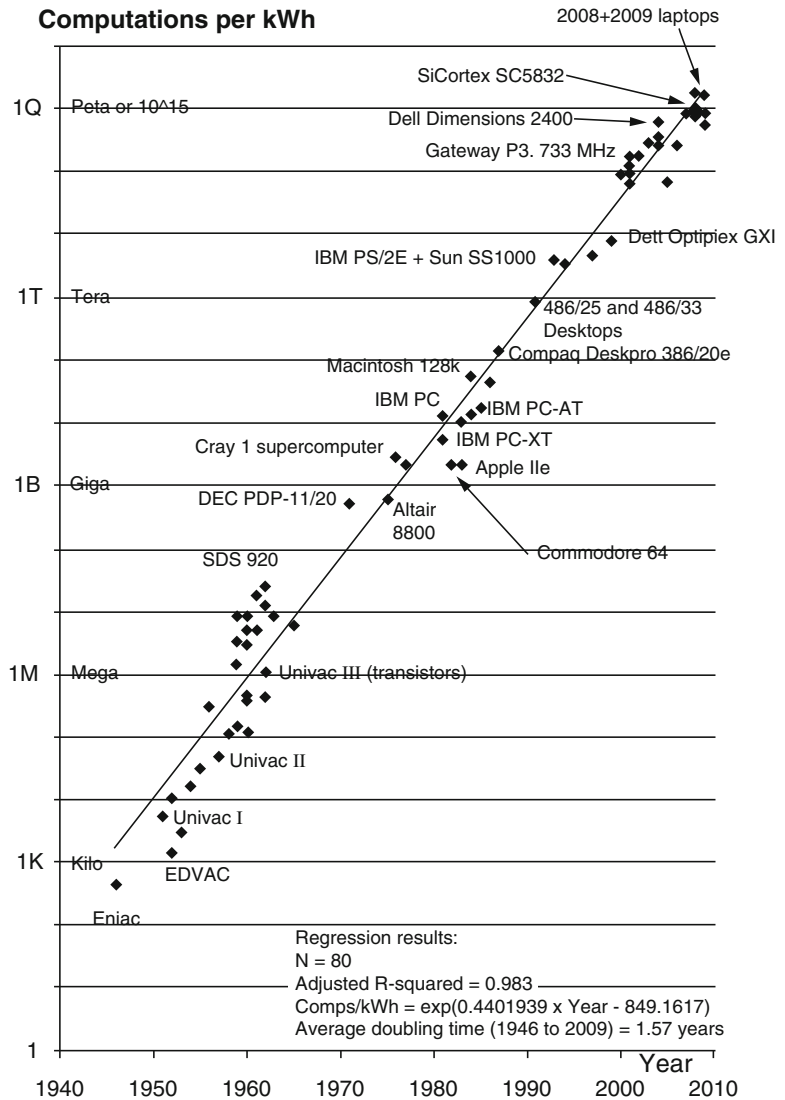
Energy-Efficiency Trends

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

Energy Implications of Big Data, Fig. 1

Trends in energy efficiency, expressed as computations per kWh (Kookey et al. 2011)



kWh, energy requirements of big-data processing will be prohibitive.

Causes of Energy Dissipation

As energy efficiency of mainstream processors that control servers, desktops, laptops, and smartphones has improved, research into ultralow-power digital circuits has paved the way for even greater reduction in energy requirements per computation. Physical laws dictate that any nonlinear 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 $k \cong 1.38 \times 10^{-23}$ 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 not clocking the 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 speedup 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 Silano 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 nonvolatility. 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 re-computation.

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 aggregating 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 (Bolla 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 users' laptops 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, backup 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 centers 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 2012).

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.

Cross-References

- ▶ [Big Data and Exascale Computing](#)
- ▶ [Computer Architecture for Big Data](#)
- ▶ [Emerging Hardware Technologies](#)
- ▶ [Parallel Processing with Big Data](#)

References

- Abts D, Marty MR, Wells PM, Klausler P, Liu H (2010) Energy proportional datacenter networks. *ACM SIGARCH Comput Archit News* 38(3):338–347
- Assuncao MD, Calheiros RN, Bianchi S, Netto MAS, Buyya R (2015) Big data computing and clouds: trends and future directions. *J Parallel Distrib Comput* 79:3–15
- Barroso LA, Holzle U (2007) The case for energy-proportional computing. *IEEE Comput* 40(12):33–37
- Barroso LA, Clidara J, Holzle U (2013) The data center as a computer: an introduction to the design of warehouse-scale machines. *Synth Lect Comput Archit* 8(3):1–154
- Benini L, De Micheli G (2002) Networks on chips: a new SoC paradigm. *IEEE Comput* 35(1):70–78
- Benini L, De Micheli G, Macii E (2001) Designing low-power circuits: practical recipes. *IEEE Circuits Syst Mag* 1(1):6–25
- Bennett CH (2003) Notes on Landauer’s principle, reversible computation and Maxwell’s Demon. *Stud Hist Philos Mod Phys* 34(3):501–510
- Bolla R, Bruschi R, Davoli F, Cucchietti F (2011) Energy efficiency in the future internet: a survey of existing approaches and trends in energy-aware fixed network infrastructures. *IEEE Commun Surv Tutor* 13(2):223–244
- Brock DC, Moore GE (eds) (2006) Understanding Moore’s law: four decades of innovation. Chemical Heritage Foundation, London
- Denning PJ, Lewis YG (2017) Exponential laws of computing growth. *Commun ACM* 60(1):54–65
- Devadas S, Keutzer K, White J (1992) Estimation of power dissipation in CMOS combinational circuits using boolean function manipulation. *IEEE Trans Comput Aided Des Integr Circuits Syst* 11(3):373–383
- Dhiman G, Rosing TS (2006) Dynamic power management using machine learning. In: *Proceedings of IEEE/ACM international conference on computer-aided design*, pp 747–754
- Dreslinski RG, Wieckowski M, Blaauw D, Sylvester D, Mudge T (2010) Near-threshold computing: reclaiming Moore’s law through energy efficient integrated circuits. *Proc IEEE* 98(2):253–266
- Eisenberg A (2010) Bye-Bye batteries: radio waves as a low-power source. *New York Times*, July 18, p BU3. <http://www.nytimes.com/2010/07/18/business/18novel.html>
- Feller E, Ramakrishnan L, Morin C (2015) Performance and energy efficiency of big data applications in cloud environments: a Hadoop case study. *J Parallel Distrib Comput* 79:80–89
- Gautschi M (2017) Design of energy-efficient processing elements for near-threshold parallel computing, doctoral thesis, ETH Zurich
- Gepner P, Kowalik MF (2006) Multi-core processors: new way to achieve high system performance. In: *Proceedings of IEEE international symposium on parallel computing in electrical engineering*, pp 9–13
- Hammadi A, Mhamdi L (2014) A survey on architectures and energy efficiency in data center networks. *Comput Commun* 40:1–21
- Hardavellas N, Ferdman M, Falsafi B, Ailamaki A (2011) Toward dark silicon in servers. *IEEE Micro* 31(4):6–15
- Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S (2015) The rise of ‘Big Data’ on cloud computing: review and open research issues. *Inf Syst* 47:98–115
- Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: *Proceedings of 33rd Hawaii international conference on information sciences, IEEE*, pp 10
- Holzle U (2010) Brawny cores still beat wimpy cores, most of the time. *IEEE Micro* 30(4):23–24
- Holzle U (2017) Advances in energy efficiency through cloud and ML, energy leadership lecture, University of California, Santa Barbara
- Hubner M, Silano C (eds) (2016) Near threshold computing: technology, methods, and applications. Springer, Cham
- Jaberipur G, Parhami B, Abedi D (2018) Adapting computer arithmetic structures to sustainable supercomputing in low-power, majority-logic nanotechnologies. *IEEE Trans Sustain Comput*. <https://doi.org/10.1109/TSUSC.2018.2811181>

- Kaul H, Anders M, Hsu S, Agarwal A, Krishnamurthy R, Borkar S (2012) Near-threshold voltage (NTV) design – opportunities and challenges In: Proceedings of 49th ACM/EDAC/IEEE design automation conference, pp 1149–1154
- Kaushik RT, Nahrstedt K (2012) T*: a data-centric cooling energy costs reduction approach for big data analytics cloud. In: Proceedings of international conference on high performance computing, networking, storage and analysis, pp 1–11
- Koomey JG, Berard S, Sanchez M, Wong H (2011) Implications of historical trends in the electrical efficiency of computing. *IEEE Ann Hist Comput* 33(3):46–54
- Krishnamachari L, Estrin D, Wicker S (2002) The impact of data aggregation in wireless sensor networks. In: Proceedings of 22nd international conference on distributed computing systems, pp 575–578
- Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM J Res Dev* 5(3):183–191
- Mack CA (2011) Fifty years of Moore’s law. *IEEE Trans Semicond Manuf* 24(2):202–207
- McMenamin A (2013) The end of Dennard scaling, online document. <http://cartesianproduct.wordpress.com/2013/04/15/the-end-of-dennard-scaling/>. Accessed 20 Feb 2018
- Mudge T, Holzle U (2010) Challenges and opportunities for extremely energy-efficient processors. *IEEE Micro* 30(4):20–24
- Musoll E, Lang T, Cortadella J (1998) Working-zone encoding for reducing the energy in microprocessor address buses. *IEEE Trans VLSI Syst* 6(4):568–572
- Nakai M, Akui S, Seno K, Meguro T, Seki T, Kondo T, Hashiguchi A, Kawahara H, Kumano K, Shimura M (2005) Dynamic voltage and frequency management for a low-power embedded microprocessor. *IEEE J Solid State Circuits* 40(1):28–35
- Pande PP, Grecu C, Jones M, Ivanov A, Saleh R (2005) Performance evaluation and design trade-offs for network-on-chip interconnect architectures. *IEEE Trans Comput* 54(8):1025–1040
- Parhami B (1999) Chapter 7, sorting networks. In: Introduction to parallel processing: algorithms and architectures. Plenum Press, Plenum, New York, pp 129–147
- Parhami B (2005) Computer architecture: from microprocessors to supercomputers. Oxford University Press, New York
- Pinheiro E, Bianchini R (2014) Energy conservation techniques for disk array-based servers. In: Proceedings of ACM international conference on supercomputing, pp 369–379
- Schroeder B, Gibson GA (2007) Understanding disk failures rates: what does an MTTF of 1,000,000 hours mean to you? *ACM Trans Storage* 3(3):Article 8
- Vogelsang T (2010) Understanding the energy consumption of dynamic random access memories. In: Proceedings of 43rd IEEE/ACM international symposium on microarchitecture, pp 363–374
- Wu L, Barker RJ, Kim MA, Ross KA (2013) Navigating big data with high-throughput, energy-efficient data partitioning. *ACM SIGARCH Comput Archit News* 41(3):249–260
- Wu Q, Pedram M, Wu X (2000) Clock-gating and its application to low power design of sequential circuits. *IEEE Trans Circuits Syst I* 47(3):415–420
- Xie Y (2011) Modeling, architecture, and applications for emerging memory technologies. *IEEE Des Test Comput* 28(1):44–51