Cluster Computing (2025)28:42
https://doi.org/10.1007/s10586-024-04767-y

Evolution of computing energy efficiency: Koomey's law revisited

Alberto Prieto[1] · Beatriz Prieto[1] · Juan José Escobar[2] · Thomas Lampert[3]

Received: 23 May 2024 / Revised: 10 July 2024 / Accepted: 5 August 2024

Abstract

For information and communication technology power consumption to be sustainable, the energy efficiency of computing systems must grow at least as fast as the demand for computing services. It is therefore crucial to understand how energy efficiency is evolving and how it will trend in the future, in order to take appropriate measures where possible. This article analyses the evolution of this parameter by analysing high-performance computers from 2008 to 2023, contrasting the results with those from Koomey's Law. It is concluded, after comparing the two that in the studied period and in the near future, energy efficiency continues to grow exponentially but at a slower rate than that established by Koomey's Law (maximum energy efficiency doubles every 2.29 years instead of every 1.57 years). Another interesting result is that energy efficiency grows at a slower rate (doubling every 2.29 years) than performance (doubling every 1.85 years).

Keywords Koomey's Law · Green computing · Green500 · Energy efficiency · High-performance computing (HPC)

1 Introduction

In the context of the environmental implications and relevance of the increasing energy consumption of computer systems, this paper presents a study on the evolution of the energy efficiency in such systems. It has to be considered that the overall energy consumption of ICTs depends not only on the growing number of devices of different nature (mobiles, PC, etc.) and the remarkable use made of them due to the constant increase of new applications, but also on the energy efficiency of these devices. Clearly, in order to contribute to the sustainability of the planet, the interest of manufacturers and engineers is not in reducing the use of ICT, but in increasing the energy efficiency of devices at least as fast as their demand. In addition to reducing greenhouse gas emissions, increasing energy efficiency is of interest to decrease power supply costs (from large data centres to mobile devices) and to extend the life of batteries. Therefore, it is of great interest to analyse the evolution of the energy efficiency of computer systems and to make estimates for the future, which is the objective of this paper.

Koomey's Law, created by analysing data from different systems from 1946 to 2009, established that the energy efficiency of computers doubled every 1.57 years [1, 2]. However, forecasts in the ICT field need to be updated frequently as technological changes occur at a very fast rate. The aim of this paper is to update the Koomey results, using real, public and verified data such as those presented in the TOP and Green500 lists [3, 4] for high-performance computers (HPC). It should be noted that also systems with much lower computing performances, such as personal computers, have energy efficiencies of the same order of magnitude [5], so the results obtained are easily generalisable to this type of systems.

The paper focuses on Koomey's Law, which is of great relevance for engineers and manufacturers. As Erik Brynjolfsson pointed out back in 2011 as a professor at MIT, in

✉ Alberto Prieto
aprieto@ugr.es

Beatriz Prieto
beap@ugr.es

Juan José Escobar
jjescobar@ugr.es

Thomas Lampert
lampert@unistra.fr

[1] Department of Computer Engineering, Automatics and Robotic, CITIC, University of Granada, 18071 Granada, Spain

[2] Department of Software Engineering, CITIC, University of Granada, 18071 Granada, Spain

[3] ICube, University of Strasbourg, 67081 Strasbourg, France

Published online: 24 October 2024

a certain sense this law may eclipse Moore's Law due to the importance that users increasingly place on power consumption in many applications, as is the case, for example, in mobile applications [6–8].

Table 1 summarises the terminology and symbols used in this article to facilitate the readers' understanding.

The rest of this paper is organised as follows. First, in order to frame the context of the paper and to present some basic concepts and terminology, a background section is introduced (Sect. 2). The methodology and data used are justified in Sect. 3. The numerical and graphical results are provided in Sect. 4. The analysis of the above results is presented in Sect. 5, and finally, a discussion and the main conclusions are summarised in Sect. 6.

2 Background

Section 2.1 describes the implications and relevance of the role of ICT in energy consumption. Section 2.2 justifies the use of the data provided by the TOP500 lists, based on the Linpack benchmark, for carrying out the present study. Section 2.3 briefly outlines the contributions of various works related to the topic presented, and finally, Sect. 2.4 defines a number of concepts necessary for a proper understanding of the rest of this study.

Table 1 Notations and symbols used

Acronym	Meaning
CE	Computing efficiency
E	Energy (Watts · hours or Joules)
EE	Energy efficiency
GE	Global energy
GPU	Graphics processing unit
FLOP	Floating-point operations
FLOPS	Floating-point operations per second
HPC	High-performance computing
HPL	High-performance Linpack
ICT	Information and communication technology
IT	Information technology
NB	Number of bits
NBI	Number of bits per instruction
NC	Number of computations
NPU	Neural processing unit
NS	Number of states
NI	Number of instructions
P	Power (Watts)
PC	Personal computer
R	Computing performance
Rmax	Maximum performance
Rpeak	Peak performance
r^2	Determination coefficient
t	Time
TPU	Tensor processing unit

2.1 ICT energy demand

One of the major challenges of today's society is to reduce energy demand, and ICT is a relevant field of electrical energy consumption, having a major impact on greenhouse gas emissions [9, 10]. Indeed, the US Semiconductor Industry Association [11] states that while global energy production grows linearly, electricity demand from computers does so exponentially. Other studies indicate that, in a worst-case scenario, ICTs could contribute up to 23% of global greenhouse gas emissions by 2030 [12]. If the trend continues, the electrical energy consumption of the vast amount of technological equipment will exceed the world's electrical energy production by 2040, which means that there would not be enough to power all the computers in the world [13].

The environmental implications of ICTs are of a different nature, not always harmful, and can be grouped into three types of effects [14–16]:

1. Direct effect. This is mainly due to the large proliferation and global increase in the number of electronic devices, communications networks and data centres connected to the Internet. This effect is also influenced by the increase of applications that are constantly used both in routine tasks (smartphones, e-mails, social networks, etc.) and in traditional computing systems (from PC to HPC applications). It is also necessary to consider the emergence of new applications, which, as in the case of the Internet of the Things (IoT), require new devices that, although individually have a very low consumption, given their enormous quantity, their overall contribution to consumption is very significant.
2. Indirect effect. It is caused by ICT applications that facilitate efficiency improvements and the reduction of primary energy consumption in very diverse sectors such as: construction, industry, transport and commerce, by providing intelligent solutions. It is good for the environment as the increase in ICT consumption comes largely from its reduction in other sectors, moderating, on balance, overall consumption. Among the main sectors benefiting are [17, 18]:
    - E-Work
    - E-Health
    - Smart Grid
    - Smart Agriculture
    - E-Learning

- Connected private transport
- Traffic control and optimisation
- E-Commerce
- E-Banking
- Smart manufacturing
- Smart logistics

3. Rebound effect. This is a phenomenon that occurs as ICT services become more useful, cheaper and more energy efficient. This increases the digital lifestyle of the society, leading to a rebound effect: ICT equipment consumes less, but is used much more. Overall, this has a negative consequence. Estimates show that possible rebound effects due to digitisation range from 10 to 30% higher energy consumption, varying by sector, technology and end-use [18].

The predominant factor in the increase of energy consumption in computing is to a large extent determined by the increasing amount of instruction processing that takes place. The results of some studies' forecasts of energy consumption per instruction or bit processed are not valid. This is because they erroneously consider without further analysis that exponentially growing computing demand translates into exponentially growing energy requirements. The demand can be measured simply by the number of computations performed (NC) but, for valid studies, it is necessary to consider also the energy consumed by each of them (EC). In short, the global energy consumption (GE) in computation is a function of both the processing demand (represented, for example, by the total number of computations executed) and the average energy consumed per computation (EC), giving:

$$GE = NC \cdot EC \tag{1}$$

Computations (NC) can refer to the number of instructions (NI) or bits (NB) executed and energy can be expressed in Joules or KWh.

The energy efficiency of computation, also called electrical efficiency, EE, is a parameter representing the number of executable computations (instructions or bits) per unit of energy (Joule or KWh), such that:

$$EE = \frac{Number\ of\ computations}{Energy\ consumed\ by\ those\ computations} = \frac{NC}{E} \tag{2}$$

where E represents the energy consumed in performing the NC computations (number of instructions or bits) indicated in the numerator. The energy consumed per computation (EC) will be:

$$EC = \frac{E}{NC} = \frac{1}{EE} \tag{3}$$

Thus, substituting the value of EC in Eq. (1), the global consumption (GE) can be expressed as a function of efficiency:

$$GE = \frac{NC}{EE} \tag{4}$$

It is deduced from Eq. (4) that, in order to reduce overall consumption (GE), either user demand for computing (represented by NC) is reduced or energy efficiency (EE) is improved. In other words, in order to prevent an overall increase in computing energy consumption, the denominator of GE in Eq. (2) (efficiency) must grow at least as fast as the numerator (demand). Many forecasts of energy consumption are flawed by considering only estimates of the increase in the numerator without considering the denominator. According to the above reasoning, and as stated in Sect. 1, computer architects and designers should focus on improving (increasing) energy efficiency. The present study addresses this by focussing on the analysis of the evolution over time of this parameter by making as rigorous estimates as possible for the future.

2.2 The TOP500 lists and the Linpack benchmark

The aim of this paper is to analyse the evolution of the energy efficiency (EE) of computers over the last three decades. To do so, it is necessary to start from the knowledge of their computing performance (R), expressed as the number of instructions executed per second, and the electrical power (P) consumed when executing those instructions. At present, it is practically impossible to have access to computers from all the years included in this study to be able to take appropriate measurements. However, such data are available in the TOP500 and Green500 lists [1], released twice a year. These data are widely recognized by the scientific community, since from 2020 to May 2024 more than 5,000 papers that make use of them appear in the literature. Indeed, the TOP500 computer ranking follows a clear and transparent methodology, being validated and presented for discussion in the open forums of the International Supercomputer Conferences on High Performance (ISC HPC), and the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) [2, 3].

Estimating the performance of a computer is a complex task as it depends on many different interrelated factors. These factors, among others, include the compiler's ability to optimise the high-level programs, the operating system, the architecture and the hardware characteristics of the computer. The desired objective of the Linpack and the TOP500 is to know, with a single parameter, how fast a computer will perform when solving real problems. Nevertheless, the applications run on computers in general, and high-performance computers in particular, are very diverse. Finding a single parameter that measures the overall performance of the computer, is a complicated issue since no single computational task can reflect the overall performance of a computer system running a wide range of programs. In order to measure the number of instructions executed per second, the maximum performance (Rmax) is used as a metric, but it must be considered that not all types of instructions consume the same time, so it is necessary to use benchmark programs to obtain measures to objectively compare the performance of different computers. These programs are established by the scientific or industrial communities [4]. Some examples are Whetstone [5, 6], Dhrystone [7], Linpack [3] or SPEC [8–10]. Numerous benchmarks and standards exist to measure other characteristics of computers in addition to computing performance [11].

The tasks that more closely match a diverse and broad set of important applications in the field of high-performance computing (HPC) are based on primitives such as vector, vector–matrix and matrix–matrix operations. These operations are fundamental in scientific applications (weather and climate prediction, for example), engineering, biotechnology, cryptoanalysis, graphics applications and in various fields of Artificial Intelligence (e.g. deep learning). In the case of Linpack benchmark, it focuses on the above-mentioned operations, as it consists in solving a random dense system of n linear equations *(A • x = b),* in double precision arithmetic (64 bits), and determines the amount of time spent factoring and resolving the system, using that time as a measure of computing performance [3].

Linpack is widely used and performance values are available for almost all relevant systems; for example, the TOP500 lists [12] have used it as a benchmark since its beginnings, as it fits reasonably well in most HPC application areas. The TOP500 lists attempt to select and rank the 500 most powerful computers by estimating their processing speed (Rmax). The Green500 ranking, associated with the TOP500 since 2013, uses energy efficiency (EE) as a ranking parameter, instead of maximum performance as the TOP500 does. Both rankings include, for each computer system, other parameters such as processor model, total number of cores, accelerator/co-processor (number of cores and model), architecture (cluster or MPP), processor speed (MHz), interconnection family and location site.

Over time, several versions of Linpack have been developed with different problem sizes. Initially (1977), the matrices associated with the system of linear equations were of the order $n = 100$. Later (1986) it was extended to $n = 1000$, with an additional version for parallel processing. This version gives greater versatility for optimising Linpack implementations as hardware architectures began to include matrix–vector and matrix–matrix operations. The fourth version (1991) was the Highly Parallel Computing Benchmark, or HPLinpack, more appropriate for testing parallel computers. In HPLinpack, the size n of the problem can be as large as necessary to optimise the performance results of the machine. This was the version adopted as a benchmark in the TOP500 in 1993 [1] and allows the user to scale the problem size and optimise the software in order to achieve the best performance for a given machine. A portable and freely available implementation of HPLinpack written in C, called High-Performance Linpack (HPL) and oriented for distributed-memory computers, was also developed and it is considered as a benchmark implementation [3]. The HPL package provides a testing and timing program to quantify the accuracy of the obtained solution as well as the time taken to compute it. The algorithm, depending on the interconnection network, can be scalable in the sense that its parallel efficiency remains constant with respect to the memory usage per processor [13].

Improvements and add-ons have been and are constantly being introduced in order to use computational and communication data patterns that more closely match a different and broad set of applications. Among other projects highlight the High-Performance Conjugate Gradients (HPCG) Benchmark, which stresses the system's main memory bandwidth and its influence on the overall performance of the system [14]. In addition, it includes a larger set of tasks to be executed than the initial version of the Linpack, including: sparse matrix–vector multiplication, vector updates, global dot products and local symmetric Gauss–Seidel smoother [15, 16].

One of the fundamental characteristics that distinguishes HPL is that it offers full freedom to implement the test and can be optimised for each type of computer or architecture. Indeed, it allows hand optimisations of the program, so that the problem size and its implementation can be adapted and adjusted to use most of the available hardware resources and achieve the best possible performance when executing the benchmark. The methodology used to improve the metric results for a particular platform can subsequently be used to obtain better performance in real applications [17, 18]. Also, the great efforts to obtain the best possible result are made because the inclusion of a computer in leading positions in the TOP500, means great prestige for the institution that owns the computer. One of the objectives of hand optimisations is to use additional resources available in the execution of the benchmark such as accelerators, coprocessors, and specialised hardware [19–22]. It should be noted that, in general, these devices are specialised in vector or matrix processing and can therefore perform the basic operations on which the Linpack focuses. Thus, the latest editions of the TOP500 implicitly reflect the performance of systems with heterogeneous computing resources, such as those using multiple GPUs (Graphics Processing Units), TPUs (Tensor Processing Units) or other specialised accelerators. Table 2 includes a list of the different models of processors used by the computers included in the TOP500 list for November 2023. In Table 3, the co-processors or accelerators (vector processors, matrix processors, GPUs, TPUs, NPUs, etc.) available in the systems of that same edition are referenced. The top place in the TOP500 list is taken by the Frontier exascale system, located at the Oak Ridge National Laboratory in Tennessee, USA, which has a total of 8,699,904 combined CPU and GPU cores, achieving a performance of *Rmax* = 1.194 EFLOPS, and an excellent power efficiency of *EE* = *52.59* GFLOPS/Watt [23].

Table 2 Families or models of processors of the TOP500 computers (November 2023 edition)

Processor family or type	# Of systems that contain it	
Intel Xeon Gold		164
Xeon Gold 62xx (Cascade Lake)	90	
Xeon Gold (Skylake)	71	
Xeon Gold (Sapphire Rapids)	1	
Xeon Gold 42xx (Cascade Lake)	1	
Xeon Gold 63xx (Ice Lake)	1	
AMD Zen		140
AMD Zen-2 (Rome)	69	
AMD Zen-3 (Milan)	66	
AMD Zen-4 (Genoa)	5	
Intel Xeon Platinum		121
Xeon Platinum (Sapphire Rapids)	19	
Xeon Platinum (Skylake)	21	
Xeon Platinum 82xx (Cascade Lake)	40	
Xeon Platinum 83xx (Ice Lake)	35	
Xeon Platinum 92xx (Cascade Lake)	6	
Intel Xeon E5		37
Intel Xeon E5 (Broadwell)	18	
Intel Xeon E5 (Haswell)	11	
Intel Xeon E5 (IvyBridge)	7	
Intel Xeon E5 (SandyBridge)	1	
Fujitsu A64FX		8
IBM Power9		7
Intel Xeon Phi		7
Intel Xeon Max		5
Vector Engine		5
Xeon Silver (Skylake)		3
Hygon Dhyana		1
Sunway		1
Xeon 5600-series (Westmere-EP)		1
Total		500

Table 3 Families and models of coprocessors or accelerators of the TOP500 computers (November 2023 edition)

Coprocessor or accelerator	# of systems that contain it
NVIDIA Tesla V100	60
NVIDIA Tesla A100	47
NVIDIA A100 SXM4	30
AMD Instinct MI	11
NVIDIA H100	10
NVIDIA Tesla K	6
NVIDIA Tesla P	6
NVIDIA Volta	5
Intel Data Center GPU Max	4
Intel Xeon Phi	2
Deep Computing Processor	1
Matrix-2000	1
NVIDIA 2050	1
NVIDIA HGX A100 80 GB 500W	1
Preferid Networks MN-Core	1

As an example, in [24] a novel device-centric High-Performance Linpack (HPL) approach is proposed and experimentally tested for current main-stream multi General-Purpose Graphics Processing Unit (GPGPU) platforms, where each process can make full use of the resources of a node, including accelerators, CPU sockets, PCI-e buses and memory/network bandwidth, etc. In this way, parallel processing can be achieved by combining the Single Instruction, Multiple Data (SIMD) technique with multithreading, thus obtaining SIMT (Single Instruction Multiple Thread) processing. As a result, the workload on the CPU-end and the inter-process communication are greatly enhanced due to higher system utilisation, while the computation on the device-end remains efficient. This approach can serve as a competitive basis for optimisations on future heterogeneous platforms.

As with other benchmarking programs, it should be noted that the results obtained by Linpack have limitations as it is not rigorous to measure the execution time of a single program to determine the computational power of a computer system. Despite its limitations, as it only measures how fast a computer will perform [6], today Linpack is still considered the main reference tool used by scientists, engineers, manufacturers and the Internet community to compare between the performances of the different HPC systems [3].

Linpack is clearly the standard for comparative studies on the performance of parallel computing systems [3, 22]. The set of 62 TOP500 lists brings together valuable information on the evolution of supercomputers over the last 31 years (1993–2023). A systematic, controlled and transparent methodology has been used to compile these lists. Moreover, this information is unique, as there is no other resource that provides such data and allows studies to be carried out on such a large number of computer systems.

2.3 Related works

There are numerous studies on the evolution and prediction models of energy consumption in the field of ICT, some more pessimistic than others, and among them are those referenced chronologically below.

In 2009, Feng and Scogland [25] analysed the first three lists of the Green500 (November 2007 to November 2008), comparing the evolution in the maximum and average energy efficiency, the energy efficiency versus speed (measured as the position within of the TOP500 rank), and the relationship between total power and energy efficiency. Among other conclusions, they indicate that the overall energy efficiency (on average) has improved in a manner that tracks with Moore’s Law, i.e., the average energy efficiency of the Green500 doubles every 18 months.

In 2009 and 2011 Koomey et al. presented a study on the evolution of the energy efficiency of 80 general-purpose computers (like mainframes, minicomputers and PCs) existing between the years 1946 to 2009. They concluded that during that period of time, the computations per KWh doubled every 1.57 years [26, 27]. This relationship is known in the scientific and engineering communities as “Koomey’s Law”. The details of this study, as well as others derived from it, will be analysed throughout this article.

Cameron, in his 2010 article [28], analyses the evolution, from November 2007 to May 2010, of the average values of energy efficiency and electrical power of all computers, the first 10 and the last 10 of each of the Green500 lists. He concludes that the top 10 supercomputers are about three times more efficient than the average system on each list and, despite this result in energy efficiency, the overall energy required for most systems on average is increasing, although the rate of this increase is slowing.

The 2011 article by Hinton et al. [29] shows that the importance of the Internet and ICT is continually increasing both in terms of economic growth and as a source of greenhouse gas production. In this context, the authors propose a network-based model of energy consumption in Internet infrastructure. This model aims to identify the elements of the Internet that dominate its energy consumption as access increases over time. This knowledge is essential to define strategies to improve the energy efficiency of the Internet. They believe that the energy consumption of data centres and content delivery networks is dominated by the energy consumption of data storage for infrequently downloaded material and by data transport for frequently downloaded material.

Deng et al. [30], using data from the TOP500 and Green500, relate the energy efficiency to the Linpack efficiency in the year 2012, and compare their evolution from 2007 to 2012 of these parameters considering various types of networks (Gigabit, Infiniband, proprietary, and custom/others), architectures (MPP, cluster), leading vendors, and processor families.

In 2013, the JASON group published a highly interesting report on the technical challenges and technological implications of supercomputing from 1 PFLOPS (10^{15} FLOPS) to 1 EFLOPS (10^{18} FLOPS) [31]. This study analysed the evolution and extrapolation to future years of various parameters of high-performance computers such as peak performance (1993–2009), energy costs for computational operations (2012 and 2020), relationship between memory bandwidth and energy, and energy consumed per FLOP (1996–2024). They conclude that, while a six-fold reduction in energy consumption for floating-point operations was achieved by 2020, the improvement is more modest (half as much) for on-chip communication. Finally, they show that there is a large disparity between the energy cost of floating-point computing and access to off-chip memory. In 2012, a DRAM access, with 64-bit words, required 1.2 nJ, and in 2020 it is reduced by a factor of 4 (to 320 pJ).

Subramaniam et al. [32] analyse the 2008 DARPA project to build an exascale supercomputer (10^{18} FLOPS) by 2020 with a maximum power consumption of 20 MW to make it economically feasible [33]. They conclude that, given the parameters of the moment in which they wrote their article (2013), a 56.8-fold improvement in computational performance would be required with only a 2.4-fold increase in energy consumption, which would be unachievable by 2020 if energy efficiency were to be increased in line with Koomey’s Law. Using data from the Green500 from 2007 to 2012, they project the trend in HPC energy efficiency for 2020, concluding that unfortunately it would be 7.2 times below the efficiency needed to meet DARPA’s 20 MW EFLOPS target. Also, in their paper they showed that heterogeneous computers (i.e., systems using GPUs or other co-processors) and custom-built systems continue to have a better overall energy efficiency than their conventional counterparts.

Van Heddeghem et al., in a 2014 article [34], evaluated how the electricity consumption caused by the use of ICT evolved from 2007 to 2012. They analysed three main domains of ICT: communication networks, personal computers and data centres. They provided a detailed description of how they obtained the results for the evolution of electricity consumption in each domain. Their estimates show that the annual growth in each of the areas (10%, 5% and 4%, respectively) was greater than the growth in global electricity consumption in the same period (3%). The relative share of this subset of ICT products and services in total global electricity consumption increased from around 3.9% in 2007 to 4.6% in 2012. The contribution to absolute electricity consumption of each of the areas turned out to be approximately the same. It follows that research should be carried out on increasing energy efficiency in all these areas, instead of focusing on just one of them.

Victor Zhirnov and collaborators published a very interesting work in 2014 [35], in which they present various models to estimate the minimum computing energy consumption in computer systems. They obtain from their models and using real data, the energy efficiency of different binary elements (logic devices and memory elements) considering the evolution of the consumption of individual transistors and microprocessors over time and the dynamics of the physical processes that take place in the different components (capacitive and resistive effects, etc.). They state that while world energy production has grown linearly, the demand for electricity from computers has grown exponentially. In typical situations, the minimum amount of energy required per bit is considered to be around 10^{-14} J, with this figure being used for laptops and PCs as well as supercomputers. Furthermore, Victor Zhirnov in his article estimates that in practice improvements are possible to reach a practical lower bound for system-level power consumption, of approximately 10^{-17} J/bit, which can be considered as a challenge to achieve. Another of the conclusions of the report is that, if the trend continues upwards, the consumption of all this huge technological equipment could exceed the world's electricity production by year 2040. Therefore, a radical improvement in the energy efficiency of the IT equipment is needed. Zhirnov's conclusions were collected a year later (2015) in a report published by the U.S. Semiconductor Industry Association in collaboration with the Semiconductor Research Corporation (SRC) and the National Science Foundation [36].

In 2015 and 2019, Andrae and Edler analysed and modelled the electric power use for ICT, making forecasts until 2030. The 2015 study [37] considers three different scenarios for the use and production of consumer devices, communication networks and data centres: the best, the expected, and the worst. One of the conclusions of the study is that, in the worst case, ICT could consume up to 51% of electricity in 2030, generating up to 23% of the greenhouse gas emissions released worldwide that year. In the 2019 work [38], they estimated that consumption from 2019 to 2030 has been lower than the data and expectations they made in 2015. Although these studies project energy consumption over 15 and 11 years and obtain very spectacular figures, they do not sufficiently appreciate the importance of improvements in the energy efficiency of the devices. Furthermore, given the changing nature of computer technology, making predictions over so many years is not reasonable.

Pangrle in his 2015 work [39], with data obtained from the November 2014 list of the Green500, relates energy efficiency to computing performance, and makes estimates of the total power consumed by supercomputers until 2022. The conclusion is that it will reach approximately *20* MW.

Gao and Zhang in 2016 [40] present and analyse the correlations of the Linpack and power efficiencies from 2011 to 2015 in the TOP500 and Green500 lists. They group the supercomputers on the lists according to their architecture: homogeneous or heterogeneous, depending on whether they use a single or various type of processor or core, and including as a subset within each class the type of interconnection (InfiniBand, Gigabit Ethernet, and custom). Within each group they analyse the performance and power behaviours. They conclude that heterogeneous systems improve performance or energy efficiency not by adding the same type of processors, but by adding different processors or coprocessors, which usually have specialised capabilities to speed up massive parallel tasks.

Most works on the impact of ICT on the global production of greenhouse gas emissions only refer to the electricity produced by the use of the devices. However, the work of Belkhir and Elmeligi [41] estimates the energy necessary for the manufacture of ICT components, that is, the energy of the production phase, which is a fixed value per device produced, and does not overlook the energy costs of the use phase which is a variable value. The authors also analyse a third parameter consisting of estimating the increase in energy consumption caused by the shortening of the useful life (lifecycle) of the devices. Reducing this leads to more frequent resales and, therefore, more purchases of new products, thus increasing production energy consumption occurs. These and other effects are analysed in that work, where they also make a forecast of ICT footprint as a percentage of global footprint projected to 2040 using both an exponential and linear fits.

Morley et al. [42] make a controversial approach by proposing that the growing reduction in electrical consumption caused by digital infrastructures, rather than the improvement of technological efficiency (efficient servers and cooling technologies), requires limiting the growth of digital traffic. Their study focuses on determining the maximum daily data demand and, therefore, the peak electricity consumption of data centres. These peaks are primarily due to the large volume of data transfer for the transmission of streaming video and interactive video, that is, IP traffic from users to data centres.

Hintemann and Hinterholzer in 2019 [43] collect data from various sources on energy consumption of servers and data centres worldwide and show how the various studies presented differ significantly. They briefly analyse possible scenarios for consumption to 2030, which give various results ranging from, in the best case, keeping energy consumption constant, to, in the worst scenario, an increase by a factor of *40* by 2030 (compared to 2015).

Koot and Wijnhoven [44] presented a forecasting model of data centre electricity needs based on understanding usage growth. To make their forecasts, they use data from, among other sources, Cisco Systems from 2010 to 2021 [45], determining the energy needs until 2023 using their model. The simulation shows exponential growths of data centre usage. Their results are compared with the projections obtained in 2020, independently by Andrae [37] and Masanet et al. [46] between 2016 and 2030. The conclusion of this article is that future energy demands of global data centres remain constant due to technological innovations, even as both consumer and enterprise workloads appear to grow exponentially over the next decade. However, the end of Moore's law is likely to cause exponential growth in data centre electricity consumption, while uncertainty in both technological and behavioural evolution explains the discrepancies found in the current literature.

In 2022, Hadlar and Sethi [47] developed an ICT consumption model for *16* countries in emerging economies that uses, as basic data, Internet penetration and the number of mobile subscriptions in relation to CO_2 emissions per person. They analysed the period from 2000 to 2018 and conclude that both the number of mobile phones and the use of the Internet increase constantly. Nevertheless, the slope of growth of gas emissions is gentler. This indicates that $CO_{2,}$ in these emerging countries, increased at a slower rate than the use of ICT.

Katal et al. in 2023 [48] wrote a survey paper concerning software-based technologies that can be used for building green data centres and that include power management at the software level. They describe the existence of new green cloud computing approaches at the virtualisation level, operating system level and application level. They also recommend the use of container technology to reduce energy consumption and achieve the challenge of obtaining more sustainable data centres.

In the article by Fatima et al. [49], the environmental impact of data centres is evaluated and the factors that cause CO_2 emissions are identified. Common strategies that can help make data centres more sustainable are discussed. They also analyse three data centres that have claimed to be green, to identify how they achieve their sustainability goals. For example, reduced carbon emissions, types of energy resources used, and how they restrained e-waste production. They conclude by suggesting a consumption reduction framework based on the concepts described.

In a recent article, Malmodin et al. [50] presented a study in which they estimate, for the year 2020, that the total electricity consumption and greenhouse gas emissions produced by the use of the ICT sector divided into three parts: user devices including the internet of things, networks and data centres. They conclude that globally the ICT sector consumed around 4% of the world's electricity with the use of computing and digital communications equipment, accounting for around 1.4% of global greenhouse gas emissions in 2020. In absolute terms, total greenhouse gas emissions were 5% higher than in 2015. According to these results, emissions from the ICT sector have evolved in line with the rest of the world. However, despite the challenges of many companies, the ICT sector did not reduce its emissions between 2015 and 2020 to meet the decarbonisation targets set by organisations such as the ITU, GSMA, GESI and SBTi, so efforts to reduce ICT energy consumption need to be scaled up and increased.

The above-mentioned works describe the evolution of energy efficiency and electricity consumption in ICT, but changes are constantly occurring in this field. They use different data sources and cover very diverse objectives and approaches such as consumption of data centres, servers, information traffic over the Internet or mobile phones. They also focus on various aspects such as the consumption of computer subsystems or how to limit e-waste production. In any case, considering the continuous technological advances in computer architecture, these analyses and projections need to be reviewed frequently to be valid, so the focus of this work is to achieve this. One of the characteristics of the present work is to use as a base data obtained from experimental measurements that are public and validated by the scientific community. Specifically, those obtained from the Green500 and TOP500 lists. The study carried out uses a large amount of data, unlike the articles referenced above. These data cover from June 2008 to November 2023 and includes a total of 8364 computers, giving great validity to the results obtained. Many of the computers appear repeated in the lists, edition by edition, but, in general, their configurations and features are updated year by year. On the other hand, the methodology used to measure the energy consumed by the systems used is uniform and public [51] and considers the energy consumption of all the elements that make up the computer system and those of the installation where it is located (air conditioning, energy transformation, lighting, etc.).

The energy consumption of ICT is determined by the use of computing and telecommunications equipment (through the execution of applications) and the energy efficiency of said equipment. From a commercial and technological point of view, it does not make sense to limit the use of resources demanded by users, so efforts should focus on improving energy efficiency, this being the parameter mainly analysed in this paper.

2.4 Efficiency and performance in computing

As indicated in Sect. 2.2, usually, and in particular with Linpack, the Rmax value is considered as a measure of performance (processing speed). This parameter indicates the maximal double precision (64 bits) floating-point instructions processed per second (MFLOP/s, or, in short, MFLOPS).

The theoretical peak performance (Rpeak) is also used to measure computing speed. The value of this parameter is determined by the particular architecture of the computer system as it depends on the total number of cores acting in parallel, the processor speed (clock frequency), and the number of additions and multiplications in floating-point full precision that can be performed in one clock cycle. A distributed system, in general, is structured in racks, each of which is composed of nodes, where in each node there are CPU sockets containing multiple cores (CPUs). In this way, the theoretical peak performance can be expressed as [52]:

$$Rpeak = racks \cdot \frac{nodes}{rack} \cdot \frac{sockets}{node} \cdot \frac{cores}{socket} \cdot \frac{cycles}{second} \cdot \frac{FLOP\,instructions}{cycle} \tag{5}$$

where FLOP instructions/cycle represents the average number of instructions executed per cycle in each of the cores, considering the implicit instruction-level parallelism.

Another parameter of interest is the computing efficiency (*CE*), which is defined as the ratio between the maximum measured performance and the peak performance:

$$CE = \frac{Rmax}{Rpeak} \tag{6}$$

The computing efficiency measures the utilisation rate of system's computation resources during the execution of a program. This parameter tries to assess how the integration and coordination between all the elements of a computer (cores, memory storage subsystem, interconnect subsystem, etc.) affect the overall utilisation of computation resources. To calculate the value of *CE* in the TOP500 lists, Rmax is measured by executing the Linpack tool, so the parameter CE is often referred to as Linpack efficiency.

The energy efficiency (EE) value defined by Eq. (2), that is, the number of executable computations per unit of energy consumed, can also be obtained as the quotient between the system performance (R) and the average power (P) consumed by the system to deliver the measured performance. Indeed, considering that the number of computations performed in a time t is $NC = R \cdot t$ and the energy consumed during its execution is $E = P \cdot t$, EE can be calculated as:

$$EE = \frac{NC}{E} = \frac{R \cdot t}{P \cdot t} = \frac{R}{P} = \frac{Performance}{Power} \rightarrow \frac{\text{FLOPS}}{\text{Watt}} \tag{7}$$

In other words, energy efficiency also represents the performances per watt. If each computation is considered to consist of the execution of a floating-point instruction (FLOP), the performance will be expressed in FLOPS and the energy efficiency in FLOPS/W.

Knowing the energy efficiency in instructions/W, it is possible to obtain it in bits/W by simply considering the average number of bits per instruction. Indeed, the number of bits (NB) can be expressed as the number of instructions (NI) multiplied by the average number of data bits per instruction (NBI):

$$NB = NI \cdot NBI \tag{8}$$

In HPCs, it is common to operate with double precision data so, in these cases, $NBI = 64$ bits. In order to summarise the evolution over time of some parameters, and to be able to easily make comparisons, the time necessary to achieve a certain objective, for example, doubling its value, is used. If y represents the value of the parameter and t the time, the slope of the curve $y = f(t)$ at each point represents the instantaneous growth rate (m). If the function $y = f(t)$ were exponential, its logarithmic representation, *ln(y)* versus t, would correspond to a straight line, being the slope:

$$m = \frac{\Delta[\ln(y)]}{\Delta t} = \frac{ln(y_2) - ln(y_1)}{t_2 - t_1} = \frac{ln\left(\frac{y_2}{y_1}\right)}{\Delta t} \tag{9}$$

To find the time interval (Δt) required for the value of the parameter y to double, simply substitute $y_2 = 2 \cdot y_1$ to the above equation, so that:

$$m = \frac{ln(2)}{\Delta t} \rightarrow \Delta t = \frac{ln(2)}{m} = \frac{0.6931}{m} \tag{10}$$

That is, the time required for a doubling of the value of y can be obtained by dividing 0.6931 by the value of the slope (m).

3 Methodology and data

The present study is based on the original data from Koomey [26, 27] and the TOP500 and Green500 lists [1] released twice a year.

The Koomey's law complements Moore's law, according to which the number of transistors in an Integrated Circuit (IC) doubles about every two years [53–56]. The great relevance of Moore's Law should be noted due to the fact that several measures of digital technology are improving by exponential rates, in line with its prediction, including density, size, memory capacity (RAM and flash), component speed, cost and, in particular, the computing performance (R).

On the other hand, regarding the present work, the TOP500 regulations dictate that for a system to be included in the list, performance *(*Rmax*)* and energy efficiency (EE) measurements are carried out by executing the Linpack program. The precision of the results is evaluated and assessed using the partial pivoting method [57]. The operations for solving the system of equations consist of additions and multiplications with 64-bit floating-point data.

The Green500 list, for its part, was associated as a complement to the TOP500 supercomputing list since November 2007 [58–60], allowing scientists, engineers and manufacturers to consider the effects of both performance and energy efficiency in evaluating HPC systems. The methodology for establishing the list was precisely defined by a working group made up of the Energy Efficient High-Performance Computing Working Group (EEHPC WG), the TOP500, the Green500, and the Green Grid [51]. The measurement protocols involve measuring the power consumption while Linpack is running using an external meter or adding the measurements from a mix of multi-meter probes. Green500 provides, among other results, a list of the supercomputers in the world based on energy efficiency (EE) obtained by dividing the performance (FLOPS) by the electrical energy consumed according to Eq. (7).

In short, the sources used in this work on the characteristics of different computers are the following:

- Years 1946–2009: original Koomey data corresponding to a total of 80 computers ranging from PCs to mainframes. It provides the energy efficiency measured in computations per KWh and considers the peak performance of the analysed systems.
- Years 2008–2012: Rmax (TFLOPS), Rpeak (TFLOPS) and Power (KW) data from the TOP500 lists have been considered.
- Years 2013–2023: Rmax (TFLOPS), Rpeak (TFLOPS), Power (KW) and energy efficiency (GFlops/Watts) data from the lists of the June and November editions of the Green500 have been considered.

The total number of computer systems included in the present study, excluding those considered by Koomey, is 9.682.

From the data obtained from the TOP500 and Green500 lists, and using the definitions described in Sect. 2.3, the following results have been derived, which are described in the next section:

- Energy efficiencies described by Koomey, in Gigacomputations/watt versus year (from 1946 to 2008).
- Power efficiencies (Rmax/P) of the average and the first system of each of the TOP500 or Green500 lists (2008–2023).
- Computing performances (Rmax) of the average values of each TOP500 list (2008–2023).
- Relationship of energy efficiency with performance observed with data from the years 2008–2023.
- Bar charts showing the position in the TOP500 of the first systems in the Green500 lists, and vice versa, and position in the Green500 of the first systems in the TOP500 (2013–2023).

In all cases, a regression analysis has been carried out for estimating the mathematical relationships between the parameter considered and the corresponding year, as an independent variable. With these analyses, the function that relates the parameter to the year, the determination coefficient (r^2) and the number of months that must elapse to double the value of the parameter are obtained.

4 Results

4.1 Energy efficiency obtained by Koomey with data from 1946 to 2008

Koomey represents energy efficiency data (in computations per KWh) over the time collected from $N = 80$ different systems, including from PCs to mainframe computers. That is, the energy efficiency value represented is:

$$EE = \frac{computations}{KWh} \tag{11}$$

From the regression analysis carried out, a coefficient of determination $r^2 = 0.983$ is obtained. The resulting function is:

$$EE = e^{0.4401939 \cdot year - 849.1617} \mathrm{computations/KWh} \tag{12}$$

which, on a logarithmic scale, turns out to be a straight line defined by:

$$ln(EE) = 0.4401939 \cdot year - 849.1617 \tag{13}$$

Applying Eq. (10):

$$\Delta t = \frac{0.6931472}{m} = \frac{0.6931472}{0.4401939} = 1.57\,\mathrm{years} \tag{14}$$

Thus, from 1946 to 2009, the energy efficiency doubled every 1.57 years.

Koomey also analysed the evolution of the energy efficiency of personal computers, from which he concluded that the energy efficiency of this type of systems increased exponentially, doubling every 1.52 years from 1975 up to 2010. In this case, he considers $N = 34$ systems and in the regression analysis he obtains a coefficient of determination of $r^2 = 0.970$. The resulting equation is:

$$EE = e^{0.4564139 \cdot year - 881.61658} \text{ computations/KWh} \tag{15}$$

which, on a logarithmic scale, turns out to be the straight line:

$$ln(EE) = 0.4564139 \cdot year - 881.61658 \tag{16}$$

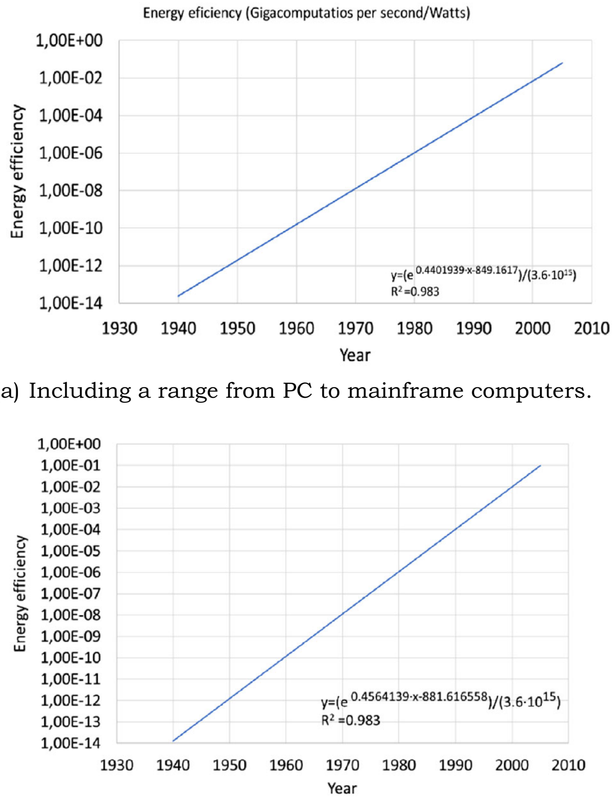
Applying Eq. (10) again:

$$\Delta t = \frac{0.6931472}{m} = \frac{0.6931472}{0.4564139} = 1.52 \text{ years} \tag{17}$$

It follows that, from 1975 to 2009, the energy efficiency for PCs doubled every 1.52 years, growing, therefore, a little faster than the global set of systems.

To better compare the results obtained with data from the TOP500 and Green500, Fig. 1 shows the Koomey regression lines using performance (R) per watt (W) as dimensions of energy efficiency instead of computations/KWh. That is, calling v the value of the energy efficiency in computations/KWh, the following transformation is carried out:

$$EE = v\frac{comp.}{\text{KWh}} = \frac{v}{3.600 \cdot 10^3}\frac{comp.}{\text{W} \cdot \text{s}} = \frac{v}{3.6 \cdot 10^{15}}\frac{Gigacomp./\text{s}}{\text{W}} \tag{18}$$



(a) Including a range from PC to mainframe computers.

(b) Including only PCs.

Fig. 1 Representation of the regression lines for energy efficiency obtained by Koomey in performance (Gigacomputations/s) per watt

The objective of this work is to compare and extend over time the results obtained by Koomey. These functions have been carried out with data obtained from the TOP500 and Green500 lists, which, as mentioned in Sect. 3, correspond to real data. The Appendix includes Tables 6, 7 and 8 that present data collected from both the TOP500 and Green500 lists and some calculations according to the definitions given in Sect. 2.3.

4.2 Energy efficiency obtained with data from 2008 to 2023

The evolution of energy efficiency, considering its maximum performance (Rmax) measured with the Linpack has been obtained:

$$EE = \frac{Performance}{Electrical\,power} = \frac{Rmax}{P} \rightarrow \frac{GFLOPS}{Watt} \tag{19}$$

The graphic results considering the average energy efficiencies of each edition of the TOP500 or Green500, are shown in Fig. 2a. The evolution of said parameter for the computer systems that occupy the first place in the Green500 is depicted in Fig. 2b.

The regression analysis carried out gives the following results for the evolution of the average values of energy efficiency (Fig. 2a):

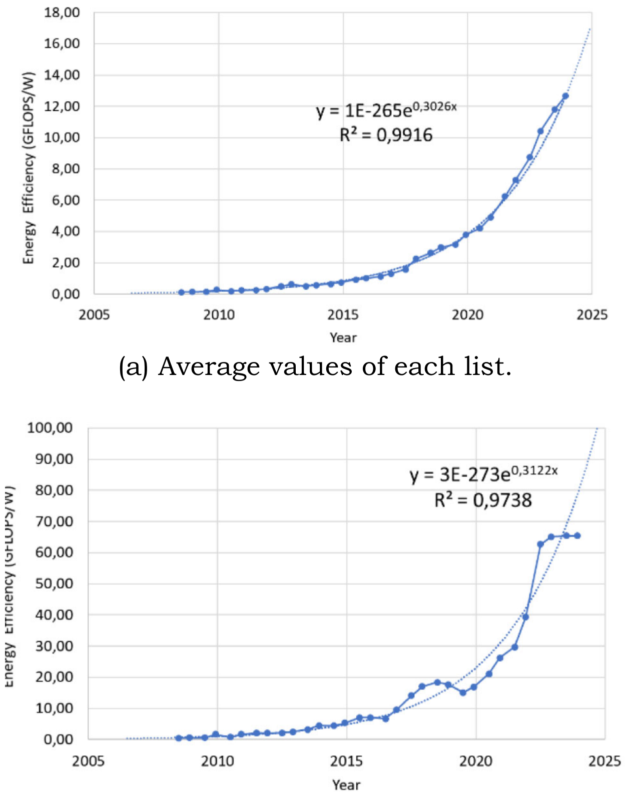
$$EE = 10^{-265} e^{0.3026 \cdot year} \text{ GFLOPS/W} \tag{20}$$

with a coefficient of determination $r^2 = 0.9916$, which, on a logarithmic scale, gives rise to a straight line with the slope $m = 0.3026$. Applying now Eq. (10):

$$\Delta t = \frac{0.6931472}{m} = \frac{0.6931472}{0.3026} = 2.29 \text{ years} \tag{21}$$

It follows that, from June 2008 to June 2023, the energy efficiency doubled every 2.29 years.

Figure 3 compares the regression lines obtained by Koomey and the one presented in this work. From the results, it is concluded that in recent decades the growth of energy efficiency occurs at a slower rate than predicted by Koomey's Law, doubling every 2.29 years, instead of every 1.57 years. It is worth considering that the observed differences may be partly due, in addition to the different data sources used, to the fact that Koomey uses general-



(a) Average values of each list.

(b) Most efficient computer in each of the lists.

Fig. 2 Maximum energy efficiencies *(Rmax/P)* of the supercomputers of each of the editions of the TOP500 or Green500 between the years 2008 to 2023

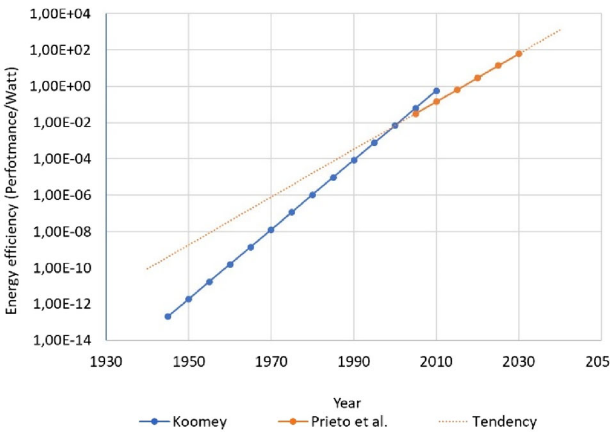


Fig. 3 Comparison of the regression straight lines for the energy efficiency obtained by Koomey (1946–2009) and in the present work (2008–2023)

purpose computers and the present work uses supercomputers. The latter can be, to some extent, considered as special-purpose computers because they are specifically designed for high-performance computing with, in general, few input/outputs operations.

The evolution of energy efficiency for the computer systems that occupy the first place in the Green500 (Fig. 2b), can be represented by the following exponential function, with a determination coefficient r^2 = *0.9738,*

$$EE = 3 \cdot 10^{-273} e^{0.3122 \cdot year} \quad GFLOPS/W \tag{22}$$

It follows that, from June 2008 to June 2023, the energy efficiency of the first place in the Green500 doubled every 2.22 years, as the following is verified:

$$\Delta t = \frac{0.6931472}{m} = \frac{0.6931472}{0.3122} = 2.22 \text{ years.} \tag{23}$$

4.3 Computing performance with data obtained from 2008 to 2023.

Figure 4 represents the regression line obtained for the evolution of the average value of the maximum computing performance (*Rmax*) of the systems included in the Green500 lists, measured, in TFLOPS. The results obtained are the following:

$$Rmax = 2.0724 \cdot e^{0.3739 \cdot (year - 2000)} \text{ FLOPS} \tag{24}$$

with a coefficient of determination r^2 = *0.9834,* which, on a logarithmic scale, gives rise to a straight line with the slope *m* = *0.3739.* Now applying Eq. (10):

$$\Delta t = \frac{0.6931472}{m} = \frac{0.6931472}{0.3739} = 1.85 \text{ years} \tag{25}$$

It can be concluded that, from June 2008 to June 2023, the computing performance doubled every 1.85 years.

Another observation of interest that has been obtained with the data from 2008 to 2023 is that, as Fig. 5 shows, there is a linear relationship between energy efficiency and performance, which can be defined by the following equation:

$$EE = 10^{-6} \cdot Rmax + 0.181 \text{ GFLOPS/W} \tag{26}$$

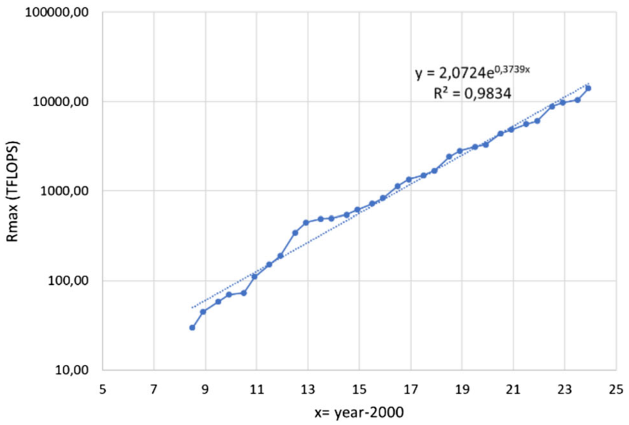


Fig. 4 Annual evolution of average computing performance

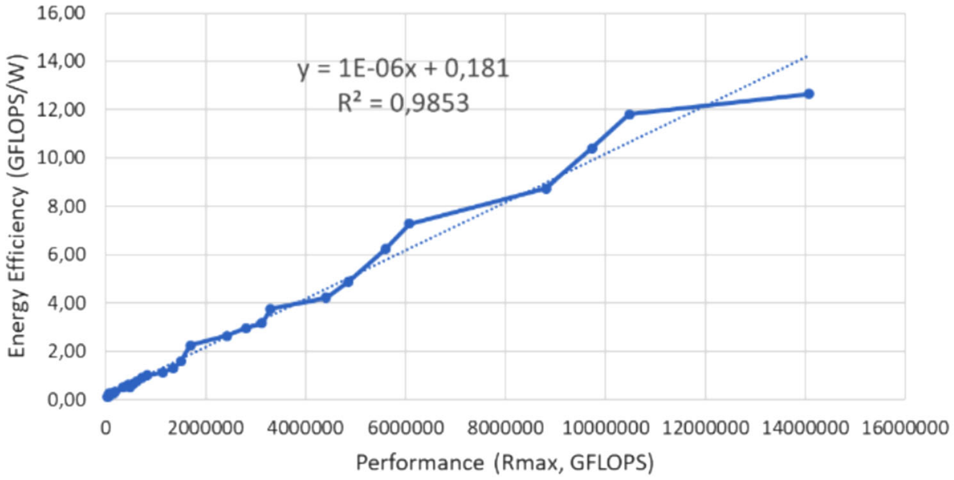


Fig. 5 Relationship of energy efficiency with performance observed with data from the years 2008 to 2023

with a coefficient of determination $r^2 = 0.9853$.

Table 4 shows a summary of the results obtained. The last column gives the number of years that must pass for the value of each of the parameters to double.

4.4 Landauer's principle

The Landauer ́s principle [61–65] derives from the second law of thermodynamics and the concept of the entropy change associated with information gain. It states that in any logically irreversible manipulation of information, such as in the erasure of a bit (making a bit, 0 or 1, become a 0) or when, for example, as a consequence of a calculation, two bits are logically combined to produce only one (the AND operation, for example), part of the information is lost. The decrease in the amount of information is accompanied by a corresponding increase in the entropy of the processing system and its environment, which is considered as an isolated system.

An irreversible process, in the context of computing, is one in which, when obtaining the output, the information of the input is lost, that is cannot be recovered from the output. For example, an addition operation is irreversible since from the result alone, the values of the input addends cannot be obtained. On the contrary, a NOT logic gate (inverter) can be considered as a reversible operation because the input can be determined from the output.

Landauer's principle can be deduced from the Boltzmann entropy formula:

$$S = k_B \cdot \ln(NS) \tag{27}$$

where S is entropy, $k_B \approx 1.38 \cdot 10^{-23}$ J/K the Boltzmann constant, and NS is the number of states. Since the entropy, assuming a constant temperature T, can be expressed as

Table 4 Summary of functions that determine the variation of the parameters with time (years) obtained through regression analyses

Years and parameter	Values	Function	Coefficient of determination (r^2)	Years to double the value of the parameter
1946 to 2009				
Energy efficiency	Obtained by Koomey	Mainframes and PCs: $EE = e^{0.440 \cdot year - 849.16}$ computations/KWh	0.983	1.57
		Only PCs: $EE = e^{0.4564 \cdot year - 881.62}$ computations/KWh	0.9700	1.52
2008 to 2023				
Energy efficiency	Average value of the #1 of each list	$EE = 10^{-265} \cdot e^{\ 0.3026 \cdot year}$	0.9916	2.29
		$EE = 3 \cdot 10^{-273} \cdot e^{0.3122 \cdot \cdot year}$	0.9738	2.22
Computing performance	Average values of each list	$Rmax = 2.0724 \cdot e^{0.3739 \cdot (year - 2000)}$ FLOPS	0.9834	1.85
Energy efficiency versus performance	Average value of each list	$EE = 10^{-6} \cdot Rmax + 0.181$	0.9853	

S = E/T, where E is the energy (heat dissipated) and *T* is the temperature of the heat sink in absolute degrees. The previous expression turns out to be:

$$E = k_B \cdot t \cdot \ln(NS) \tag{28}$$

The energy loss (consumption) per processed bit can be obtained, considering that in this case there are *NS* = 2 possible states, so the energy associated with the processing of 1 bit turns out to be:

$$E = k_B \cdot t \cdot \ln(2) \tag{29}$$

Considering that *ln(2) ≈ 0.69315*, and an ambient temperature of 20^0 C = 293.1^0 K, it is obtained that the Landauer limit represents an energy of approximately 0.0175 eV ($2.805 \cdot 10^{-21}$ J) per missing bit. That is, approximately $3 \cdot 10^{-21}$ J/bit.

From the results obtained, it is possible to calculate the year in which the Landauer limit will be reached. It must be considered that by knowing the variation of energy efficiency over time (EE), the evolution of the energy consumed per bit (J/bit) can be obtained. Indeed, from Eq. (2), the energy consumed by computing *(E/NC)* can be expressed as:

$$\frac{E}{NC} = \frac{1}{EE} \quad \mathrm{J/FLOP} \tag{30}$$

As the floating-point instructions considered are double precision, each of them contains NBI = 64 bits, so the energy consumed per bit (EB) based on energy efficiency will be:

$$EB = \frac{1}{64 \cdot EE} \quad \mathrm{J/bit} \tag{31}$$

Substituting the formula obtained for energy efficiency (Eq. (20)), into the previous equation, the average energy consumed per bit turns out to be:

$$EB \approx 10^{254} \cdot \mathrm{e}^{-0.303 \cdot \mathrm{year}} \ \mathrm{J/bit} \tag{32}$$

The corresponding function is represented in Fig. 6. The year in which the forecasts will no longer be valid due to the Landauer limit having been reached can be obtained simply by replacing *EB* in the previous expression with the value of the said limit $\approx 3 \cdot 10^{-21}$ J/bit, and isolate the year variable. As shown in Fig. 6, according to the forecasts made in this paper, the Landauer limit will be reached by approximately 2090. It should be noted that this calculation was made assuming a system temperature of T = 20 °C. Obviously, if this calculation were made at temperatures lower than the room temperature, more efficiency could be achieved (see Eq. (29)). Of course, cooling has an energy cost, so this will partly (or perhaps fully) offset the benefits of lower temperatures.

In addition to the Landauer limit, there are other estimates of the theoretical minimum energy possible to perform a binary operation (switching a bit). Among these estimates, it is interesting to highlight the one presented by Feynman in a talk in Tokyo in 1985 [66, 67], based on the theoretical possibility of building transistors with only 3 atoms. With these hypothetical transistors, referred to by Koomey in [68, 69], Feynman estimated that electronic computers could be built that would improve energy consumption by a factor of 10^{11} compared to the computer technology of that time (1985). Performing the appropriate calculations, we obtain that the energy consumed by a switching would be $2.19 \cdot 10^{-18}$ J/bit and, with current trends (2024), this theoretical limit, without the use of 3-atom transistors, would be reached approximately 2070 (Fig. 6).

4.5 Relative positions between the Green500 and the TOP500 lists

The position of the #1 Green500 computers in the TOP500 can be obtained from the data in Table 6, and is represented in Fig. 7a. Analogously, Fig. 7b represents the position occupied by the #1 TOP500 computers in the Green500, obtained from the data in Table 8.

5 Analysis of results

One of the first results obtained is that the evolution of energy efficiency is exponential and varies at a slower rate compared to Koomey's inference. Indeed, results show that from 2008 to 2023 energy efficiency doubled every 2.29 years, while the trend obtained by Koomey from 1946 to 2009 showed that it should double every 1.57 years. In the comparison presented herein, the measurement methodology used should be analogous in order to be as rigorous as possible. The initial data source and the concept of "computations" used by Koomey were taken from the work of Nordhaus [70]. However, the new results presented here (2008–2023) follow the TOP500 protocol, which correspond in all cases to double precision floating-point operations. This lack of uniformity of methodologies should be considered when analysing the results. We must also not forget that, as indicated in Sect. 4, the differences observed may be partly due to the fact that Koomey uses general-purpose computers and the present work uses supercomputers.

Considering the other results obtained, it is observed that the evolution of all the parameters reliably follow exponential functions (Table 4), having in all cases the average values of each edition of the TOP500 or Green500 coefficient of determination greater than 0.97 ($r^2 > 0.97$).

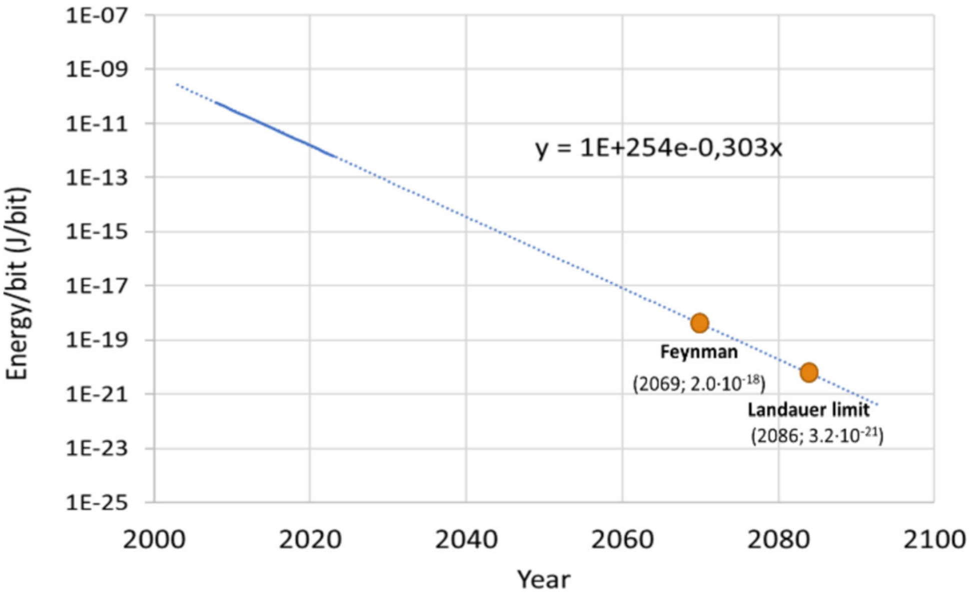


Fig. 6 Evolution over time of the estimates made of energy consumption per bit processed. The point where the Feynman and Landauer limits will be reached are also shown

That is to say, the functions obtained very adequately reflect the behaviour over time of the data considered.

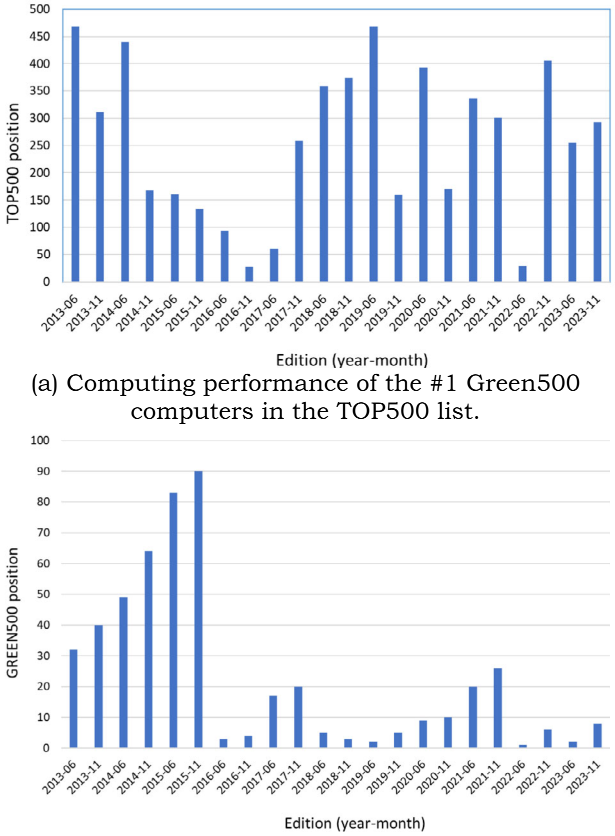
An important aspect to bear in mind about the computational performance measurements in the TOP500 and Green500 is that they are carried out in floating-point double precision (64-bits), FP64. However, in the execution of algorithms for certain workloads, such as those related to machine learning, in contrast to other traditional scientific applications (simulation for modelling phenomena in physics, biology and chemistry, for example), 64-bits precision is not always required. In such cases, mixed-precision arithmetic can be used, so that floating-point operations can be performed, depending on the phase of the algorithm being executed, with 32-bit (FP32), 16-bit (FP16) or even 8-bit (FP8) data, without any deterioration in the quality of the results [71, 72].

There are several accelerators [73, 74] that support the ability to operate on floating-point data of different lengths. The smaller the data size, the higher the performance, the lower the memory usage, and data transfers take less time. Thus, by properly choosing the smallest precision scale at each execution step without sacrificing the precision of the results, it is possible to obtain results of the same quality as with FP64 but in a shorter time. The performance improvement in different arithmetic precisions can be measured with reference to double precision (FP64). For example, in the NVIDIA H100 SMX4 Tensor Core GPU [75], used by several supercomputers included in the 2023 November TOP500, FP64 achieves X = 67 TFLOPS, FP32 attains 15X, FP16 achieves 29X and FP8 reaches 59X.

A new benchmark, HPL-MxP Mixed-Precision [76, 77], has been proposed that attempts to converge traditional HPC workloads with new AI workloads, so that there is a reasonable connection with the measurements made for decades to evaluate the performance of supercomputers. However, it is currently considered appropriate to continue to measure performance in 64-bit FLOPS for the purpose of comparison between different computers. By doing so, it is possible to easily and efficiently compare the performance evolution of HPCs from different eras regardless of the types of accelerators available and the arithmetic precision used at the time.

The new results obtained refer only to HPC computers, but it must be considered that other systems, such as personal computers, have energy efficiencies comparable to those of supercomputers [78] after considering the energy consumption of all elements of the computer, including climate control and following in all cases the methodology established by the TOP500 [51]. In fact, Koomey's original work established that the energy efficiency of PCs doubles at almost the same speed as that of computing systems in general (18 months versus 19, as shown in Table 4). Prieto et al., show in [78] how five Personal Computers (PCs) with Intel processors of different generations have energy efficiencies that are comparable to those of supercomputers within the first 174 positions of the Green500 (November 2021 list), despite the enormous difference in performance *(Rmax)* between PCs and supercomputers.

Regarding the variation over time of the average computing performance of the Green500 computers, it has been found that this parameter doubled every 1.85 years, in line with Moore's Law, which doubles about every 2 years. Another result of interest is that performance improves more quickly (doubles every 1.83 years) than maximum energy efficiency (doubles every 2.29 years). Also, from the slope of the straight line in Fig. 5, it can be deduced



(a) Computing performance of the #1 Green500 computers in the TOP500 list.

(b) Energy efficiency of the #1 TOP500 computers in the Green500 list.

Fig. 7 Position in the TOP500 and Green500 lists of the first Green500 and TOP500 computer, respectively, within the same edition

that each increase in performance of 1 TFLOPS produces only an improvement of 1 MFLOPS/W in energy efficiency.

As shown in Fig. 6, according to the forecasts made in this paper, the Landauer limit will be reached in approximately the year 2090. This means that, if the energy efficiency of irreversible information processing follows the trend of the last 16 years, the limit will be reached around 2090. This is because, as described in Sect. 4.5, according to the second principle of thermodynamics, it is physically impossible to irreversibly process information consuming less than $\approx 3{\cdot}10^{-21}$ J/bit of energy. The processing of a bit is identified as a logical switching or elementary computation. There are other predictions, such as that of Feynman, which assume a three-atom transistor to calculate this limit, setting it at approximately $2.0{\cdot}10^{-18}$ J/bit [66, 68, 69]. It should be noted that, in the case of reversible computations (as occurs in the field of quantum computing), the value deduced by the Margolus-Levitin Theorem should be used as the lower limit of energy consumption, which is $\approx 3.0{\cdot}10^{-34}$ J/bit [79].

For conventional (non-quantum) computing, if the increase in performance follows the trend of the last 15 years (doubling every 1.85 years, in line with Moore's Law) when the Landauer limit is reached, the performance would be of the order of $Rmax \approx 10^{14}$ TFLOPS.

Concerning the position occupied by the #1 Green500 computers in the TOP500 tables of the same editions (Fig. 7a), it is observed that 19% of Green500 winners occupy the first quartile of the TOP500; 24% the second quartile; 33% the third quartile and 24% the fourth quartile. On the contrary, for the case of the position occupied by the #1 TOP500 computers in the Green500 tables (Fig. 7b), it was concluded that from 2013 to 2015 they occupied positions ranging from 30 to 90, gradually decreasing positions of the Green500, until reaching position 90. However, from 2016 to 2023, the energy efficiency substantially improved since the first computer in the TOP500 of each list occupies positions ranging between 1 and 26 of the Green500 (Fig. 7b).

6 Discussion and conclusions

Regarding the data source used in this work, it should be noted that clear protocols on the methodology must be followed to take measurements for computers to be included in the TOP500 and Green500 lists. However, the results are provided by those responsible for the data centres themselves, with little or no independent controls to verify their authenticity. Indeed, those responsible for preparing the lists, in addition to checking different sources of information, limit themselves to randomly selecting a statistical representative sample of the first 500 systems of their database, performing an audit on them. For example, the methodology to be followed in the Green500 measurements [51] establishes that, for the calculation of energy efficiency, the electrical consumption of all computational nodes, any interconnect network the application uses, any head or control nodes, any storage system the application uses, all power conversion losses inside the computer, and any internal cooling devices (self-contained liquid cooling systems and fans), must be included. Nevertheless, no procedures are defined to verify that this is done correctly.

Another issue of interest is to highlight that Linpack is a benchmark aimed at measuring computing power in applications that require intensive calculation (a lot of data including vector and matrix operations), but it may not correlate well with some real workloads of current supercomputers or general-purpose computers (which follow other objectives and trends). In these cases, Linpack would not reflect the hardware improvements designed to obtain greater efficiency in other particular types of workloads.

Notwithstanding what has been said in the previous paragraphs, the TOP500, together with the Green500, constitutes an exceptional and open meeting point for scientists and engineers. In fact, these two rankings are useful to analyse the situation and trends in the evolution of the characteristics of HPC systems, as well as comparing different equipment (always considering the indicated limitations). Moreover, Linpack has been in use for decades and allows consistent comparisons over time and remains a very useful tool.

This work has been carried out based on Koomey data covering the years 1946 to 2010 and the TOP and Green500 lists from 2008 to June 2023. It has been proven that the energy efficiency of HPCs between the last above-mentioned years grew exponentially, doubling every 2.29 years. This conclusion has been obtained through a regression analysis with coefficient of determination of $r^2 = 0.9916$, considering a total of 9,682 HPCs included in the 30 lists used. It must be noted that in successive lists many supercomputers are repeated, although their configurations and characteristics are generally updated list by list.

The result obtained indicate that the growth of energy efficiency is occurring at a slower rate than that obtained by Koomey in 2011 with data between 1946 and 2009, which was doubling every 1.57 years with a coefficient of determination of $r^2 = 0.983$. However, the result has been obtained using as "computations double precision floating point instructions", and in the case of Koomey, as mentioned in Sect. 5, the concept of "computation" is based on the work of Nordhaus [70].

This present work has focused on analysing the evolution of the energy efficiency of the most powerful supercomputers in the world compiled in the TOP500 lists, representing the entirety of each list by their average values. It is worth noting that many supercomputers are repeated throughout editions, but generally their structures are modified, either by simply adding more nodes and racks or by changing some of them for more powerful or energy-efficient computing units. These lists reflect the reality of the computers that operate every year. Another approach of great interest is the one followed by Koomey [80], which tries to reflect the improvements over time of the current technical ability to create new computing devices. To do this, he considers supercomputers only in their year of first operation, so he does not reflect machines beyond this date. In the indicated work by Koomey, both computing power and energy efficiency are analysed. The following conclusions are drawn with respect to energy efficiency in data that is cleaned to include only equipment in its first year of operation:

- The energy efficiency of the supercomputer suite from 2009 to 2019 doubled every 2.14 years with $r^2 = 0.6$.
- The energy efficiency of the first supercomputer from 2009 to 2019 doubled every 2.12 years with $r^2 = 0.86$.
- The energy efficiency of the top 10% of supercomputers from 2009 to 2019 doubled every 2.11 years with $r^2 = 0.7$.

These results are summarised in Table 5.

It has also been shown that, with the trends obtained here, the Landauer limit would be reached approximately in the year 2090, and the energy/bit equivalent to that estimated by Feynman with 3-atom transistors in 2070.

The evolution of other parameters has also been analysed, such as computing performance, which doubles every

Table 5 Comparison of results obtained with those of Koomey and Subramaniam

References	Parameter	Year	Data source	Computers	Analysed years	Doubling years	r^2
Koomey	EE	2009	Diverse	Mainframes, server, general purpose and PCs	1946–2009	1.57	0.983
Koomey	EE	2009	Diverse	PCs	1975–2010	1.52	0.970
Subramaniam	EE	2017	Green500	Top 100 of each list	2007–2012	2,33	0.84
Koomey	EE	2020	TOP500	TOP #1 (Lists of computers in their 1st year of operation)	2009–2019	2.12	0.86
Koomey	EE	2020	TOP500	TOP 10% Lists of computers in their 1st year of operation	2009–2019	2,11	0,7
Koomey	EE	2020	TOP500	Lists of all computers in their 1st year of operation	2009–2019	2.14	0.6
Koomey	Rmax	2020	TOP500	Lists of computers in their 1st year of operation	2009–2019	1.66	0.73
Present work	EE	2024	Green500	Average value of each list	2008–2023	2.29	0.99
Present work	EE	2024	Green 500	TOP #1	2008–2023	2.22	0.97
Present work	Rmax	2024	Green500	Average value of each list	2008–2023	1.85	0.98

1.85 years (in line with Moore's Law). It is assumed that by increasing computing performance, the number of applications and use of computers will increase, so the number of computations (NC) would increase. Under this hypothesis, it is worrying that energy efficiency is growing at a slower rate than performance (doubling every 2.29 years compared to 1.85). Another unfavourable implication is that there might be an eventual negative trend, although rather slow, in energy efficiency, i.e., it is possible that it will decrease further in the future. Although the difference seems small, doubling energy efficiency every 1.85 years means increasing it approximately 43 times in a decade, and doubling it every 2.29 years means increasing it only about 21 times per decade. Therefore, more needs to be done to ensure that energy efficiency grows at least as fast as performance.

The results obtained are of interest to researchers, engineers and manufacturers in order to make forecasts about new products, trying to ensure that the efficiency increase exceeds that of the demand for computing services.

A common goal of institutions owning HPC systems is to be included in the TOP500 list, and within it in leading positions. To this end, Linpack implementations are optimised to take full advantage of the heterogeneity in the systems and the different accelerators, coprocessors, and specialised hardware available. In this way, the measures presented in the TOP500 are constantly adapted to reflect the improvements introduced by new concepts and technologies in computer architecture. However, one must be careful with forecasts, as new ideas and technologies are being researched. This is the case, for example, in the area of reducing consumption in servers, storage, networks, interconnections, power conversion and cooling systems, where the following concepts, among others, can be found [81]:

- Changes in the devices and in the internal architecture of the microchips [82–84].
- Management and planning of resource use, from low to high system levels, such as using the Dynamic Voltage and Frequency Scaling (DVFS) technique [85, 86], Dynamic Power Management (DPM) [86, 87], or even using power capping protocols, establishing a certain power threshold for a device that it cannot exceed [88].
- Scale changes, in order to plan and assign tasks to the available hardware resources considering their energy efficiency. Within this area, virtualisation technologies [48] have acquired great relevance, which have been enhanced by the increase in scale of data centres through the merger or transformation of medium-sized centres to hyperscale centres (Google Cloud, Amazon Web Services, Microsoft Azure, OVHCloud or Rackspace Open Cloud), where energy consumption is better managed [89–92].

An interesting aspect is that the ultimate objective is to reduce the energy consumed in the execution of our programs, a value that can be obtained by applying Eq. (3), where in this case NC would be the number of instructions executed by the program and EE the energy efficiency of the hardware where these instructions are executed. Consequently, to reduce the energy consumed by the program, the energy efficiency of the hardware devices (EE) must be increased and the number of instructions (NC) must be reduced as much as possible, that is, maintaining the response times and precision required for the results. Therefore, from an energy point of view, it is extremely important, not only to increase energy efficiency, as considered in this work, but also to use techniques for efficient algorithm development: HW/SW codesign procedures, compilers, and software, in general, both for general-purpose computers and for specific applications. As Leiserson [93] points out, as miniaturisation approaches its limits, bringing an end to Moore's law, performance improvements will have to come from what might be called the three "top end" technologies: software, algorithms and hardware, to distinguish them from the traditional "bottom end" technologies (semiconductor physics and silicon fabrication technology). These three top technologies have a key role to play in reducing the energy consumption of ICT.

Koomey and Masanet indicate that "IT changes so quickly that most data characterizing it are obsolete in short order" [94], so that in this work we have tried to update some of the forecasts made. However, due to the great improvements that are constantly being introduced, it is advisable that the projections presented should be only considered for a few years.

Appendix

See Tables 6, 7, 8.

Table 6 Data from the computers with the highest energy efficiency, extracted from TOP500 and GREEN500 lists

Source	Green500 edition	TOP500 Rank	Name	Rmax (TFLOPS)	Rpeak (TFLOPS	Rmax/ Rpeak	Power (kW)	Maximum energy efficiency (GFLOPS/ W)	Peak energy efficiency (GFLOPS/W)
TOP500	2008-06	324	BladeCenter QS22 Cluster	11.11	18.28	0.61	22.76	0.49	0.80
TOP500	2008-11	220	BladeCenter QS22 Cluster	18.57	30.46	0.61	34.63	0.54	0.88
TOP500	2009-06	422	BladeCenter QS22 Cluster	18.57	30.46	0.61	34.63	0.54	0.88
TOP500	2009-11	445	GRAPE-DR accelerator Cluster	21.96	84.48	0.26	51.20	1.65	1.65
TOP500	2010-06	131	QPACE SFB TR Cluster	44.50	55.71	0.80	57.54	0.77	0.97
TOP500	2010-11	115	NNSA/SC Blue Gene/Q	65.35	104.86	0.62	38.80	1.68	2.70
TOP500	2011-06	109	NNSA/SC Blue Gene/Q Prot. 2	85.88	104.86	0.82	40.95	2.10	2.56
TOP500	2011-11	64	BlueGene/Q	172.49	209.72	0.82	85.12	2.03	2.46
TOP500	2012-06	252	BlueGene/Q	86.35	104.86	0.82	41.09	2.10	2.55
TOP500	2012-11	253	Beacon	110.50	157.55	0.70	45.11	2.45	3.49
Green500	2013-06	467	Eurora	98.51	175.67	0.56	30.70	3.21	5.72
Green500	2013-11	311	TSUBAME-KFC	125.10	217.66	0.57	27.78	4.50	7.84
Green500	2014-06	439	TSUBAME-KFC	151.80	217.82	0.70	34.58	4.39	6.30
Green500	2014-11	168	L-CSC	301.30	593.60	0.51	57.15	5.27	10.39
Green500	2015-06	160	Shoubu	353.82	842.96	0.42	50.32	7.03	16.75
Green500	2015-11	133	Shoubu	353.82	1535.83	0.23	50.32	7.03	30.52
Green500	2016-06	94	Shoubu	1001.01	1533.46	0.65	149.99	6.67	10.22
Green500	2016-11	28	DGX SaturnV	3307.00	4896.51	0.68	349.50	9.46	14.01
Green500	2017-06	61	TSUBAME3.0	1998.00	3207.63	0.62	141.60	14.11	22.65
Green500	2017-11	259	Shoubu system B	841.96	1127.68	0.75	49.50	17.01	22.78
Green500	2018-06	359	Shoubu system B	857.63	1127.68	0.76	46.60	18.40	24.20
Green500	2018-11	374	Shoubu system B	1063.31	1353.22	0.79	60.40	17.60	22.40
Green500	2019-06	469	DGX SaturnV Volta	1070.00	1819.75	0.59	97.00	15.11	18.76
Green500	2019-11	159	A64FX prototype	1999.50	2359.30	0.85	118.48	16.88	19.91
Green500	2020-06	393	MN-3	1621.10	3922.33	0.41	76.80	21.11	51.07
Green500	2020-11	170	NVIDIA DGX SuperPOD	2356.00	2812.80	0.84	89.94	26.20	31.27
Green500	2021-06	336	MN-3	1822.40	3137.87	0.58	61.36	29.70	51.14
Green500	2021-11	301	MN-3	2181.20	3389.52	0.64	55.39	39.38	61.19
Green500	2022-06	29	Frontier TDS	19,200.00	23,105.54	0.83	308.68	62.68	74.85
Green500	2022-11	405	Henri	2038.00	5417.34	0.38	31.31	65.09	173.02
Green500	2023-06	255	Henri	2882.00	3579.13	0.81	44.07	65.40	81.21
Green500	2023-11	293	Henri	2882.00	3579,13	0.81	44.07	65.40	81.21

Table 7 Average values calculated in each TOP500 and Green500 list

Source	Green500 edition	Rmax (TFLOPS)	Rpeak (TFLOPS)	Computing efficiency (Rmax/Rpeak)	Power (kW)	Number of systems with power data	Maximum energy efficiency (GFLOPS/W)	Peak energy efficiency (GFLOPS/W)
TOP500	2008-06	29.57	44.03	0.63	253.41	247.00	0.12	0.19
TOP500	2008-11	44.70	30.46	0.62	359.83	253.00	0.13	0.21
TOP500	2009-06	57.98	84.14	0.63	387.30	238.00	0.15	0.24
TOP500	2009-11	69.50	98.33	0.66	401.53	238.00	0.27	0.27
TOP500	2010-06	72.75	99.90	0.67	398.42	257.00	0.20	0.28
TOP500	2010-11	110.24	160.56	0.71	476.50	263.00	0.24	0.35
TOP500	2011-06	151.46	211.30	0.68	545.92	274.00	0.25	0.38
TOP500	2011-11	188.52	271.22	0.66	592.86	283.00	0.33	0.54
TOP500	2012-06	343.50	463.35	0.68	667.37	293.00	0.50	0.74
TOP500	2012-11	445.74	613.24	0.70	684.78	281.00	0.63	0.93
Green500	2013-06	489.53	701.51	0.68	988.17	500.00	0.49	0.71
Green500	2013-11	497.77	729.11	0.69	1100.05	500.00	0.58	0.81
Green500	2014-06	546.04	807.00	0.68	1124.37	500.00	0.64	0.91
Green500	2014-11	615.73	907.01	0.69	1184.62	500.00	0.75	1.10
Green500	2015-06	723.21	1026.25	0.72	1222.06	500.00	0.92	1.29
Green500	2015-11	834.18	1277.91	0.68	1321.90	500.00	1.01	1.56
Green500	2016-06	1139.83	1698.32	0.67	1271.98	500.00	1.13	1.68
Green500	2016-11	1350.86	2038.37	0.67	1335.91	500.00	1.29	1.91
Green500	2017-06	1496.74	2264.48	0.66	1305.55	500.00	1.58	2.41
Green500	2017-11	1690.24	2678.68	0.64	1502.84	305.00	2.25	3.35
Green500	2018-06	2421.83	3843.33	0.64	1602.57	263.00	2.64	3.98
Green500	2018-11	2809.84	4396.78	0.64	1755.46	234.00	2.97	4.60
Green500	2019-06	3119.52	4934.48	0.63	1756.60	209.00	3.17	5.12
Green500	2019-11	3294.41	5496.72	0.64	1555.19	214.00	3.77	6.43
Green500	2020-06	4412.27	6957.84	0.64	1673.28	206.00	4.21	7.20
Green500	2020-11	4857.52	7692.07	0.62	1727.63	189.00	4.88	8.24
Green500	2021-06	5593.44	8854.21	0.63	1899.95	182.00	6.24	10.16
Green500	2021-11	6073.72	9577.79	0.62	1753.91	179.00	7.27	12.11
Green500	2022-06	8806.17	13,696.88	0.61	1782.65	191.00	8.74	13.71
Green500	2022-11	9728.77	15,081.39	0.60	1780.59	195.00	10.40	16.32
Green500	2023-06	10,478.05	15,651.84	0.61	1813.99	188.00	11.80	18.08
Green500	2023-11	14,063.68	21,319.25	0,66	2068.71	190.00	12.66	18.92

Table 8 Position in the Green500 lists of the #1 of the TOP500

Edition	Position in Green500
2013-06	32
2013-11	40
2014-06	49
2014-11	64
2015-06	83
2015-11	90
2016-06	3
2016-11	4
2017-06	17
2017-11	20
2018-06	5
2018-11	3
2019-06	2
2019-11	5
2020-06	9
2020-11	10
2021-06	20
2021-11	26
2022-06	1
2022-11	6
2023-06	2
2023-11	8

Acknowledgements The authors appreciate the cooperation of Christian Morillas, Jesús González and Francisco Illeras (from the Department of Computer Engineering, Automatics and Robotic of the University of Granada, Spain) on this work.

Author contributions All authors have contributed to the study conception and development of this work. A.P. performed supervision, conceptualization and methodology. B.P. performed formal analysis and investigation. B.P. and J.J.E. prepared data collection and analysis. J.J.E. and T.L. performed validation of results. B.P. and J.J.E. wrote the first draft of the manuscript. All authors reviewed and approved the final version of manuscript.

Funding This work was partially supported by Grant PID2022-137461NB-C31 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU", Grant PID2022-137461NB-C32 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU", and Project PPJIA2023-025 funded by the University of Granada (Spain).

Data availability Data is provided within the manuscript. The data and results presented in this article can be used freely, always stating its source. The source data can be consulted freely at https://www.top500.org/lists/green500 (Green500 list), and at https://www.top500.org (TOP500 list).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Koomey, J.G., Berard, S., Sanchez, M., Wong, H.: Assessing trends in the electrical efficiency of computation over time. - IEEE Ann. Hist. Comput. 17 (2009)
2. Koomey, J., Berard, S., Sanchez, M., Wong, H.: Implications of historical trends in the electrical efficiency of computing. IEEE Ann. Hist. Comput. **33**(3), 46–54 (2011)
3. TOP500 The list. https://www.top500.org/.
4. Green500 list, https://www.top500.org/lists/green500/
5. Prieto, B., Escobar, J.J., Gómez-López, J.C., Díaz, A.F., Lampert, T.: Energy efficiency of personal computers: a comparative analysis. Sustainability **14**(19), 12829 (2022)
6. Brynjolfsson, E.: Is Koomey's Law eclipsing Moore's Law?. Economics of Information Blog. MIT (2011)
7. Greene, K.: A new and improved Moore's law. MIT Technology Rev, 12. (2011) www.technologyreview.com/2011/09/12/191382/a-new-and-improved-moores-law/
8. Zhang, N.: Moore's Law is dead, long live Moore's Law!. (2022) arXiv preprint arXiv:2205.15011.
9. Malmodin, J., Lundén, D.: The energy and carbon footprint of the global ICT and E&M sectors 2010–2015. Sustainability **10**(9), 3027 (2018)
10. Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G.S., Friday, A.: The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. Patterns, 2(9), (2021)
11. Semiconductor Industry Association and the Semiconductor Research Corporation.: Rebooting the IT Revolution: A call to the action. (2015) https://www.semiconductors.org/resources/rebooting-the-it-revolution-a-call-to-action-2/.
12. Andrae, A.S., Edler, T.: On global electricity usage of communication technology: trends to 2030. Challenges **6**(1), 117–157 (2015)
13. Burgess, A., Brown, T.: By 2040, There may not be enough power for all our computers. HENNIK RESEARCH, 17 (2016)
14. Haldar, A., Sethi, N.: Environmental effects of Information and communication technology-exploring the roles of renewable energy, innovation, trade and financial development. Renew. Sustain. Energy Rev. **153**, 111754 (2022)
15. Li, X., Zhang, C., Zhu, H.: Effect of information and communication technology on CO2 emissions: an analysis based on country heterogeneity perspective. Technol. Forecast. Soc. Chang. **192**, 122599 (2023)
16. Oo, K.T., Jonah, K.A.Z.O.R.A., Thin, M.M.Z.: A systematic review of the pros and cons of digital pollution and its impact on the environment. J. Sustain. Environ. Manage. **2**(1), 61–73 (2023)
17. Global e-Sustainability Initiative. Accenture Strategy (2015)# SMARTer2030-ICT Solutions for 21st Century

Challenges. Brussels: Global e-Sustainability Initiative (GeSI) and AccentureStrategy.

18. Global e-Sustainability Initiative.: Available online: http://gesi.org/assets/js/lib/tinymce/jscripts/tiny_mce/plugins/ajaxfilemanager/uploaded/SMARTer 2020 - The Role of ICT in Driving a Sustainable Future - December 2012._2.pdf. (2012)
19. Strohmaier, E., Meuer, H.W., Dongarra, J., Simon, H.D.: The TOP500 list and progress in high-performance computing. Computer **48**(11), 42–49 (2015)
20. Dongarra, J.J., Luszczek, P., Petitet, A.: The LINPACK benchmark: past, present and future. Concurr. Comput.tion **15**(9), 803–820 (2003)
21. Davies, I.N., Ike, A.V.: Overview of Common Parallel Benchmark Applications and Suites. J. Appl. Comput. Sci. Math. 16(34) (2022)
22. Curnow, H.J., Wichmann, B.A.: A synthetic benchmark. Comput. J. **19**(1), 43–49 (1976)
23. Freiburger, D.: Computer benchmarks: 50 years ago and now. ACM SIGHPC Connect **11**(2), 5–7 (2023)
24. Weicker, R.P.: Dhrystone: a synthetic systems programming benchmark. Commun. ACM **27**(10), 1013–1030 (1984)
25. Dixit, K.M.: Overview of the SPEC Benchmarks. The Benchmark Handbook, 7. (1993)
26. Bucek, J., Lange, K. D., Kistowski, J.: SPEC CPU2017: Next-generation compute benchmark. In Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, pp. 41–42. (2018)
27. Brunst, H., Zavala, M.: First experiences in performance benchmarking with the new SPEChpc 2021 suites. In 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pp. 675–684 (2022)
28. Aslanpour, M.S., Gill, S.S., Toosi, A.N.: Performance evaluation metrics for cloud, fog and edge computing: a review, taxonomy, benchmarks and standards for future research. Internet of Things **12**, 100273 (2020)
29. TOP500. The Linpack Benchmark. https://www.top500.org/project/linpack/
30. Petitet, A.: HPL-a portable implementation of the high-performance Linpack benchmark for distributed-memory computers. (2004) https://www.netlib.org/benchmark/hpl/
31. Heroux, M.A., Dongarra, J.: Toward a new metric for ranking high performance computing systems (No. SAND2013-4744). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); University of Tennessee,, Knoxville, TN (2013)
32. Dongarra, J., Heroux, M.A., Luszczek, P.: A new metric for ranking high-performance computing systems. Natl. Sci. Rev. **3**(1), 30–35 (2016)
33. Dongarra, J., Heroux, M.A.: Toward a new metric for ranking high performance computing systems. Sandia Report, SAND2013-4744, 312, 150 (2013)
34. Barrett, R.F., Chan, T.H.F., D'Azevedo, E.F., Jaeger, E.F., Wong, K., Wong, R.Y.: Complex version of high performance computing LINPACK benchmark (HPL). Concurr. Comput. Pract. Exp. **22**(5), 573–587 (2010)
35. Heinecke, A., Dubey, P.: Design and implementation of the linpack benchmark for single and multi-node systems based on intel xeon phi coprocessor. In IEEE 27th International Symposium on Parallel and Distributed Proceedings, pp. 126–137 (2013)
36. Rohr, D., Lindenstruth, V.: A load-distributed linpack implementation for heterogeneous clusters. In 2015 IEEE 17th International Conference on Embedded Software and Sys, pp. 436–443. IEEE (2015)
37. Intel.: Distribution for LINPACK and the Intel Optimized HPL-AI benchmarks. (2023) https://www.intel.com/content/www/us/en/docs/onemkl/developer-guide-windows/2023-2/intel-distribution-for-linpack-benchmark-contents.html
38. Chalmers, N., Bauman, P.: Optimizing high-performance linpack for exascale accelerated architectures. In Proceedings of the International Conference for High Performance Computing, pp. 1–12 (2023)
39. Kim, J., Kwon, H., Kang, J., Park, J., Lee, S., Lee, J.: SnuHPL: high performance LINPACK for heterogeneous GPUs. In Proceedings of the 36th ACM International Conference on Supercomputing, pp. 1–12 (2022)
40. TOP500 The List. Novembre 2023. https://www.top500.org/lists/top500/2023/11/
41. Sun, Q., Ma, W., Sun, J., Li, H.: Evolving the HPL benchmark towards multi-GPGPU clusters. CCF Trans. High Perform. Comput. **5**(1), 84–96 (2023)
42. Feng, W.C., Scogland, T.: The green500 list: Year one. In 2009 IEEE International Symposium on Parallel & Distributed Processing, pp. 1–7. IEEE (2009)
43. Cameron, K.W.: A tale of two green lists. Computer **43**(09), 86–88 (2010)
44. Hinton, K., Baliga, J., Feng, M., Ayre, R., Tucker, R.S.: Power consumption and energy efficiency in the internet. IEEE Netw. **25**(2), 6–12 (2011)
45. Deng, Y., Zhang, P., Marques, C., Powell, R., Zhang, L.: Analysis of Linpack and power efficiencies of the world's TOP500 supercomputers. Parallel Comput. **39**(6–7), 271–279 (2013)
46. JASON: Technical challenges of exascale computing. MITRE Corporation, McLean, VI, USA (2013)
47. Subramaniam, B., Saunders, W., Feng, W.C.: Trends in energy-efficient computing: A perspective from the Green500. In 2013 International Green Computing Conference Proceedings, pp. 1–8. IEEE (2013)
48. Bergman, K., Borkar S., Yelick, K.: Exascale computing study: Technology challenges in achieving exascale systems. Defense Advanced Research Projects Agency Information Processing Techniques Office (DARPA IPTO), Tech. Rep., 15, 181 (2008)
49. Van Heddeghem, W., Lambert, S., Lannoo, B., Colle, D., Pickavet, M., Demeester, P.: Trends in worldwide ICT electricity consumption from 2007 to 2012. Comput. Commun. **50**, 64–76 (2014)
50. Zhirnov, V., Cavin, R., Gammaitoni, L.: Minimum energy of computing, fundamental considerations. In ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology. IntechOpen (2014)
51. Andrae, A.S.: Comparison of several simplistic high-level approaches for estimating the global energy and electricity use of ICT networks and data centers. Int. J. **5**, 51 (2019)
52. Pangrle, B.: News on energy-efficient large-scale computing. In CHIPS 2020 VOL. 2: New Vistas in Nanoelectronics, pp. 165–170. Springer International Publishing, Cham (2015)
53. Gao, Y., Zhang, P.: A survey of homogeneous and heterogeneous system architectures in high performance computing. In 2016 IEEE International Conference on Smart Cloud (SmartCloud), pp. 170–175. IEEE (2016)
54. Belkhir, L., Elmeligi, A.: Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J. Clean. Prod. **177**, 448–463 (2018)
55. Morley, J., Widdicks, K., Hazas, M.: Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption. Energy Res. Soc. Sci. **38**, 128–137 (2018)
56. Hintemann, R., Hinterholzer, S.: Energy consumption of data centers worldwide. Business, Computer Science (ICT4S) (2019)

57. Koot, M., Wijnhoven, F.: Usage impact on data center electricity needs: A system dynamic forecasting model. Appl. Energy **291**, 116798 (2021)
58. Cisco Systems.: Forecast and Methodology, 2012–2017; (2013) https://d2sr9p9v571tfz.cloudfront.net/upload/images/10_2013/131028130134.pdf
59. Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J.: Recalibrating global data center energy-use estimates. Science **367**(6481), 984–986 (2020)
60. Katal, A., Dahiya, S., Choudhury, T.: Energy efficiency in cloud computing data centers: a survey on software technologies. Clust. Comput. **26**(3), 1845–1875 (2023)
61. Fatima, E., Ehsan, S.: Data centers sustainability: approaches to green data centers. In 2023 International Conference on Communication Technologies (ComTech), pp. 105–110. IEEE (2023)
62. Malmodin, J., Lövehagen, N., Bergmark, P., Lundén, D.: ICT sector electricity consumption and greenhouse gas emissions–2020 outcome. Telecommun. Policy **48**(3), 102701 (2024)
63. Tomić, D., Imamagić, E., Gjenero, L.: Towards new energy efficiency limits of High Performance Clusters. In Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces, pp. 89–93. IEEE (2013)
64. Moore, G.: Moore's law. Electron. Mag. **38**(8), 114 (1965)
65. Schaller, R.R.: Moore's law: past, present and future. IEEE Spectr. **34**(6), 52–59 (1997)
66. Mollick, E.: Establishing Moore's law. IEEE Ann. Hist. Comput. **28**(3), 62–75 (2006)
67. Koomey, J., Naffziger, S.: Moore's Law might be slowing down, but not energy efficiency. IEEE Spectr. **52**(4), 35 (2015)
68. Dongarra, J.J., Moler, C.B., Bunch, J.R., Stewart, G.W. LINPACK users' guide. Society for Industrial and Applied Mathematics (1979)
69. Sharma, S., Hsu, C.H., Feng, W.C.: Making a case for a Green500 list. In Proceedings 20th IEEE International Parallel & Distributed Processing Symposium, p 8. IEEE (2006)
70. Ge, R., Feng, X., Pyla, H., Cameron, K., Feng, W. Power measurement tutorial for the Green500 list. The Green500 List: Environmentally Responsible Supercomputing (2007)
71. WG, EEHPC.: Energy Efficient High Performance Computing Power Measurement Methodology. (version 2.0 RC 1.0), p. 33, (2015) https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf
72. Landauer, R.: Irreversibility and heat generation in the computing process. IBM J. Res. Dev. **5**(3), 183–191 (1961)
73. Landauer, R.: Dissipation and noise immunity in computation and communication. Nature **335**(6193), 779–784 (1988)
74. Lloyd, S.: Ultimate physical limits to computation. Nature **406**(6799), 1047–1054 (2000)
75. Bérut, A., Arakelyan, A., Lutz, E.: Experimental verification of Landauer's principle linking information and thermodynamics. Nature **483**(7388), 187–189 (2012)
76. Markov, I.L.: Limits on fundamental limits to computation. Nature **512**(7513), 147–154 (2014)
77. Koomey, J.G., Scott Matthews, H., Williams, E.: Smart everything: Will intelligent systems reduce resource use? Annu. Rev. Environ. Resour. **38**(1), 311–343 (2013)
78. Feynman, R.P.: The computing machines in the future. In: The pleasure of finding things out: The best short works of Richard P. Feynman. Basic Books. Perseus Books (2005)
79. Feynman, R.P.: The computing machines in the future. Lect. Notes Phys. **746**, 99–113 (2008). https://doi.org/10.1007/978-4-431-77056-5
80. Nordhaus, W.D.: Two centuries of productivity growth in computing. J. Econ. Hist. **67**(1), 128–159 (2007)
81. Baboulin, M., Tomov, S.: Accelerating scientific computations with mixed precision algorithms. Comp. Phys. Comm. **180**(12), 2526–2533 (2009)
82. Dörrich, M., Fan, M., Kist, A.M.: Impact of Mixed precision techniques on training and inference efficiency of deep neural networks. IEEE Access (2023)
83. NVIDIA.: Train with Mixed Precision. Userś guide. (2023) https://docs.nvidia.com/deeplearning/performance/pdf/Training-Mixed-Precision-User-Guide.pdf
84. Kahler, S.: TensorFlow 2.12, Intel. (2023) https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/AMX-Support-for-Mixed-precision-Training-and-Inference-Now/post/1484953.
85. NVIDIA.: H100 Tensor Core GP, Data sheet. (2023) https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet
86. HPL-MxP.: HPL-MxP Mixed-Precision benchmark. (2023) https://hpl-mxp.org/
87. Lin, R., Yuan, X., Wang, F.: 5 ExaFlop/s HPL-MxP Benchmark with Linear Scalability. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–13 (2023)
88. Koomey, J.G.: A primer on the energy efficiency of computing. Phys. Sustain. Energy III PSE III Using Energy. Effic. Prod. Renew. **1652**(1), 82–89 (2015)
89. Margolus, N., Levitin, L.B.: The maximum speed of dynamical evolution. Physica D **120**(1–2), 188–195 (1998)
90. Koomey, J., Schmidt, Z., Naffziger, S.: Supercomputing performance and efficiency: an exploration of recent history and near-term projections. Burlingame, CA: Koomey Analytics, for AMD (2020)
91. Bharany, S., Sharma, S., Khalaf, O.I., Abdulsahib, G.M., Al Humaimeedy, A.S., Aldhyani, T.H., Alkahtani, H.: A systematic survey on energy-efficient techniques in sustainable cloud computing. Sustainability **14**(10), 6256 (2022)
92. Zhang, J., Gao, F., Hu, P.: A vertical transistor with a sub-1-nm channel. Nature Electron. **4**(5), 325–325 (2021)
93. Bush, S.: IBM beats finFETs with vertical CMOS at IEDM. Electronics Weekly.com (2021)
94. Varnava, C.: Chips cool off with integrated microfluidics. Nature Electron. **3**(10), 583–583 (2020)
95. Herbert, S., Marculescu, D.: Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In Proceedings of the 2007 International Symposium on Low Power Electronics and Design, pp. 38–43 (2007)
96. Kocot, B., Czarnul, P., Proficz, J.: Energy-aware scheduling for high-performance computing systems: a survey. Energies **16**(2), 890 (2023)
97. Dougherty, B., White, J., Schmidt, D.C.: Model-driven autoscaling of green cloud computing infrastructure. Futur. Gener. Comput. Syst. **28**(2), 371–378 (2012)
98. Benini, L., Bogliolo, A., De Micheli, G.: A survey of design techniques for system-level dynamic power management. IEEE. Trans. Very Large Scale Integr. (VLSI) Syst. **8**, 299–316 (2000)
99. Project Natick Phase 2.: https://natick.research.microsoft.com/
100. Roach, J.: Microsoft finds underwater datacenters are reliable, practical and use energy sustainably. Microsoft Innov. Stories, 14 (2020)
101. Zhang, Y., Shan, K., Li, X., Li, H., Wang, S.: Research and Technologies for next-generation high-temperature data centers. Renew. Sustain. Energy Rev. **171**, 112991 (2023)
102. Capozzoli, A., Primiceri, G.: Cooling systems in data centers: state of art and emerging technologies. Energy Proc. **83**, 484–493 (2015)
103. Leiserson, C.E., Thompson, N.C., Emer, J.S., Kuszmaul, B.C., Lampson, B.W., Sanchez, D., Schardl, T.B.: There's plenty of

room at the Top: What will drive computer performance after Moore's law? Science **368**(6495), eaam9744 (2020)

104. Koomey, J., Masanet, E.: Does not compute: Avoiding pitfalls assessing the Internet's energy and carbon impacts. Joule **5**(7), 1625–1628 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alberto Prieto received the M.Sc. and Ph.D. degrees in Physics (electronics) from University Complutense de Madrid (1968) and the University of Granada (1976), respectively. He is a Professor Emeritus in the Department of Computer Engineering, Automatic Control and Robotics of the University of Granada. His research interests include computer engineering, artificial neural networks and intelligent systems, and most recently green computing.

Beatriz Prieto received the M.Sc. and Ph.D. in Electronic Engineering from the University of Granada. She is Associate Professor at the Department of Department of Computer Engineering, Automatic Control and Robotics of the University of Granada. Her research interests are focused in the fields of intelligent systems for signal processing applied to biomedical applications, and recently, green computing.

Juan José Escobar received the M.Sc. and Ph.D. degrees in Computer Engineering from University of Granada, Spain, in 2014 and 2020, respectively. He is a Permanent Lecturer at the Department of Software Engineering of University of Granada. His research interests include code optimization, energy-efficient parallel computing, and workload balancing strategies on heterogeneous and distributed systems, especially in issues related to evolutionary algorithms and multi-objective feature selection problems.

Thomas Lampert has Ph.D. in Computer Sciences from the University of York. He have held various research positions at the University of York and the University of Strasbourg, where he is now the Chair of Data Science and Artificial Intelligence. His research interests are in the field of Artificial Intelligence, and more specifically in Machine Learning and Image and Time-Series Analysis in various fields of application (most recently in medical imaging and remote sensing).