El reto de la computación verde:
¿Hasta cuándo habrá energía suficiente para hacer todo lo que queramos con la informática?

Alberto Prieto Espinosa
Miembro de la Academia de Ciencias Matemáticas, Físico Químicas y Naturales de Granada.
Profesor Emérito del Departamento de Ingeniería de Computadores

Faculta de Ciencias (Universidad de Granada)
3 de noviembre 2023

Computación verde o sostenibilidad en las TIC

- Es el estudio y la práctica de diseñar, fabricar, usar y eliminar computadores, servidores y su hardware asociado para consumir energía de manera eficiente y efectiva con un impacto mínimo o nulo en la salud y el medio ambiente.
  - No voy a considerar los aspectos medioambientales relacionados con la fabricación, reciclaje y eliminación.

- Dhaini, M., Jaber, M., Fakhereldine, A., Hamdan, S., & Haraty, R. A. (2021). Green computing approaches-A survey. Informatica, 45(1)
- IBM (2022) Learn how green computing reduces energy consumption. https://www.ibm.com/cloud/blog/green-computing#:~:text=Green%20computing%20(also%20known%20as,consumed%20by%20manufacturers%2C%20data%20ce

2

Contexto

- Green Computing se enmarca dentro de uno de los mayores retos de la sociedad actual, consistente en **reducir el consumo energético.**
- En general, la sociedad desconoce que las TIC constituyen un campo relevante en el consumo de energía eléctrica, teniendo un gran impacto en las emisiones de gases de efecto invernadero.
- Científicos, ingenieros y profesionales deben participar activamente en el reto de reducirlo.
- Además de las razones medioambientales, reducir el consumo de energía:
  - tiene fuertes implicaciones económicas y
  - mejora la autonomía de muchos dispositivos que utilizan baterías, como teléfonos inteligentes, dispositivos móviles y elementos del Internet de las Cosas

3

Contenido

- Análisis de la contribución de las TIC al consumo energético
- Cómo afectan las TIC al medio ambiente
- Estimaciones sobre la evolución del consumo energético de computación.
- Consumo del tráfico de datos digitales
- Algunos datos y situaciones prácticas
- Procedimientos para reducir el consumo energético en TIC
- Conclusiones

4

Análisis de la contribución de las TIC al consumo energético

5

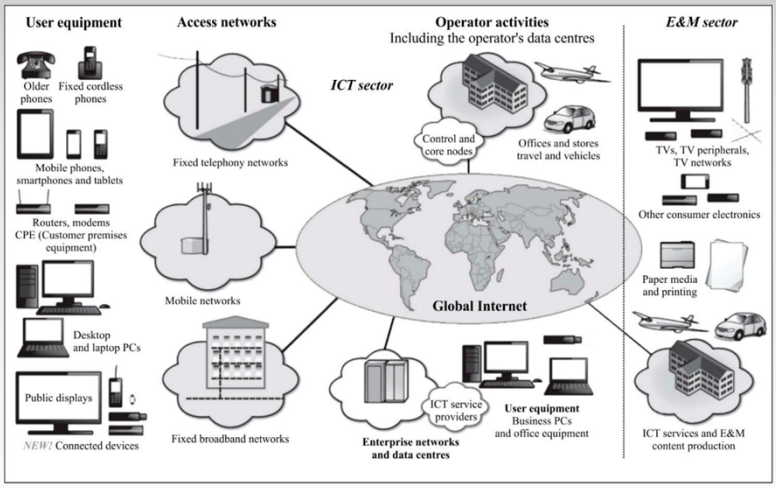
Análisis de la contribución de las TIC en el consume de energía eléctrica

- La Asociación de la Industria de Semiconductores de EE. UU. afirma:
  - Aunque la producción mundial de energía crece linealmente, la demanda de electricidad procedente de ordenadores lo hace de forma exponencial
- En el peor de los casos, las TIC podrían contribuir hasta el **23% de las emisiones globales de gases de efecto invernadero para 2030.**
- De continuar la tendencia, el consumo de energía eléctrica de la gran cantidad de equipos tecnológicos **superará la producción mundial de energía eléctrica en 2040**, lo que no sería suficiente para alimentar todos las computadores del mundo.

- V. Zhirnov ,Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution. (2015)
- A.S. Andrae y T. Edler. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157
- A. Burgess, T. Brown. By 2040 there may not be enough power for all our computers, Manufacturer, 17 Aug 2016
- C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. Blair, A. Friday. The climate impact of ICT: A review of estimates, trends and regulations. arXiv preprint arXiv:2102.02622. (2022)
- J. Malmodin y D. Lundén. (2018). The energy and carbon footprint of the global ICT and E&M sectors 2010–2015. Sustainability, 10(9), 3027

6

Sectores de las TIC involucrados en el consumo de energía



Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement. International Telecommunication Union (ITU). Recommendation ITU-T L.1470

7

Cómo afectan las TIC al medio ambiente

8

Cómo afectan las TIC al medio ambiente

- Efecto directo
- Efecto indirecto
- Efecto terciario (o de rebote)

J. Desjardins. (2018). What happens in an internet minute in 2018. Visual Capitalist. https://www. visualcapitalist. com/internetminute-2018

9

El efecto directo, en primer lugar, es debido a:

- La gran proliferación e **incremento global del número** de dispositivos electrónicos, redes de transmission y centros de datos conectados a Internet.

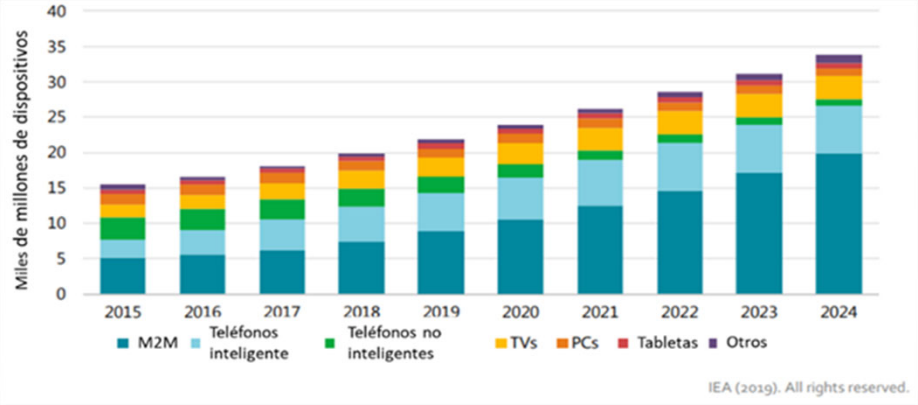


Grafico realizado por *International Energy Agency* en base al trabajo de T. Barnett y colaboradores (2019) and Cisco (2016)

10

El efecto directo también es debido a:

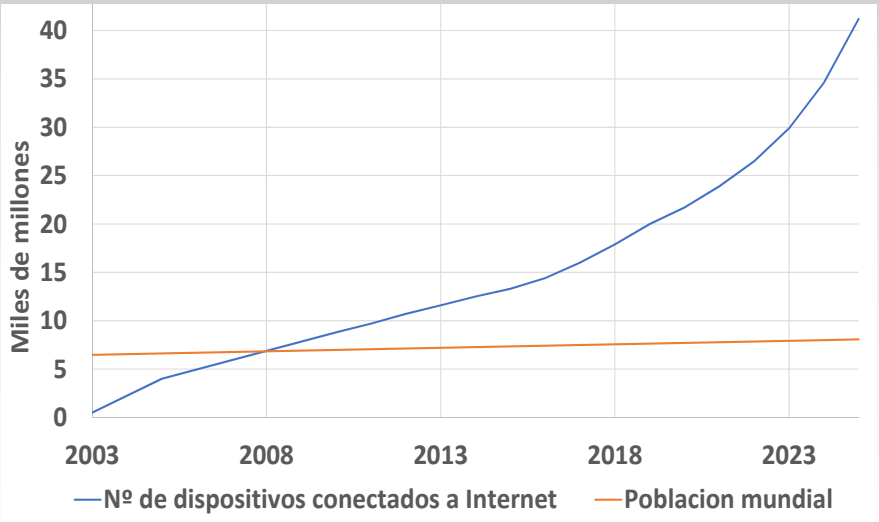
- El **incremento de aplicaciones** que constantemente usamos tanto para tareas rutinarias (telefonos inteligentes, emails, social redes sociales, ...), como para programar tareas tradicionales de computación (PCs → HPC).
- La aparición de **nuevas aplicaciones** que requieren nuevos dispositivos que, aunque individualmente consumen muy poca energía, dada su enorme cantidad su contribución global al consume es muy significativa.

J. Desjardins. (2018). What happens in an internet minute in 2018. Visual Capitalist. https://www. visualcapitalist. com/internetminute-2018.

11

Ejemplo de ámbito de aplicación: Internet de las Cosas (IoT)

- Se puede considerar que IoT nació cuando el número de dispositivos conectados a Internet superó el número total de habitantes de la Tierra (finales de 2008)



12

Efecto indirecto

- Está provocada por aplicaciones TIC que facilitan la mejora de la eficiencia y la **reducción del consumo primario de energía** en sectores muy diversos como: construcción, industria, transporte y comercio, aportando soluciones inteligentes.
- **Es bueno para el medio ambiente**.
- En otras palabras, el aumento del consumo de TIC proviene en gran medida de su **reducción en otros sectores**, moderándose, como saldo total, el consumo global.
- **Objetivo**: identificar las diferentes aplicaciones TIC en la edificación, el transporte y la industria que redundan en una reducción del consumo energético.

13

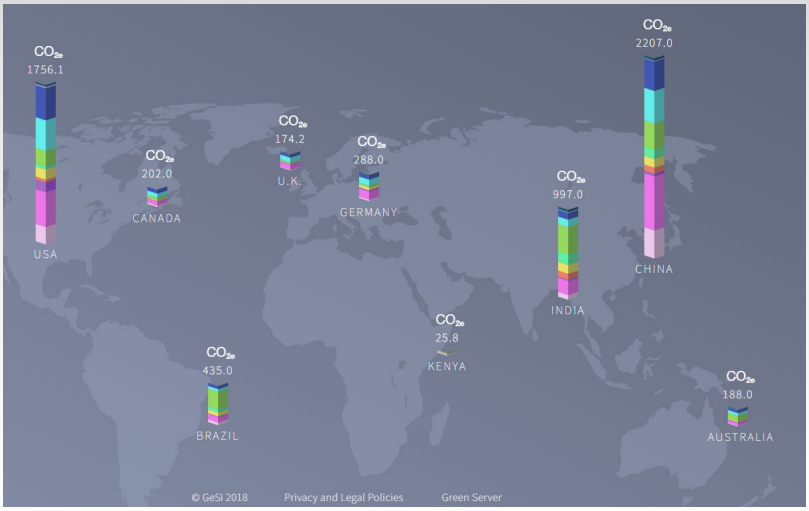
Según datos de la iniciativa GeSI los dominios TIC que están consiguiendo mayores volúmenes de reducción de CO2 son:

- E-salud
- E-enseñanza
- Redes eléctricas inteligentes
- Edificios inteligentes
- Agricultura inteligente
- Transporte privado interconectado
- Control y optimización de tráfico.
- E-comercio
- E-bancario
- **E-trabajo**
- Fabricación inteligente
- Logística Inteligente.

14

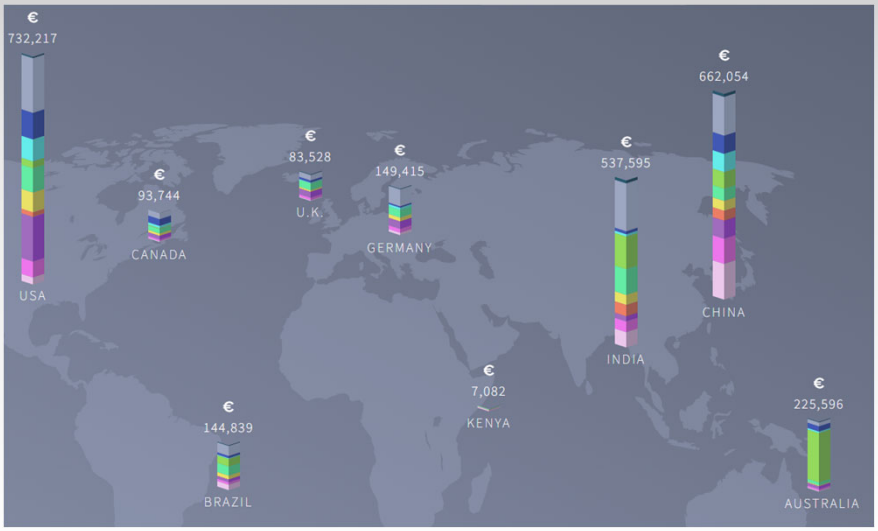
Efectos indirectos: reducción de emisiones de CO2 en millones de toneladas, gracias a las TIC

- https://smarter2030.gesi.org/explore-the-data/



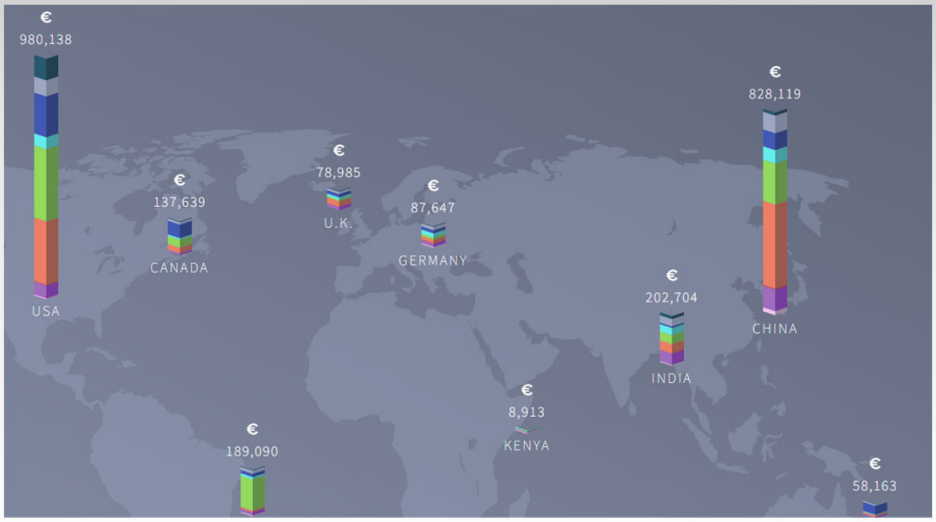
15

Ahorro de costes producido (in millones de €) (no se especifica para cada ámbito)



16

Ingresos producidos para las TIC (retorno) (en millones €)



17

Efecto terciario (o efecto rebote)

- Es un fenómeno que se produce a medida que los servicios TIC son más útiles, más baratos y eficientes energéticamente, ya que esto aumenta nuestro estilo de vida digital, lo que produce un efecto rebote: **los equipos** TIC consumen menos, pero **se utilizan mucho más**.
- A nivel mundial tiene una consecuencia **negativa.**
- Las estimaciones muestran que los posibles efectos rebote por la digitalización varían entre un **10% y un 30% de mayor consumo eléctrico**, dato que varía según el sector, la tecnología y el uso final.

GeSI. Global e-Sustainability Initiative. Accenture strategy SMARTer2030-ICT solutions. (2015) https://smarter2030.gesi.org/downloads/Full_report.pdf

18

Estimaciones sobre la evolución del consumo

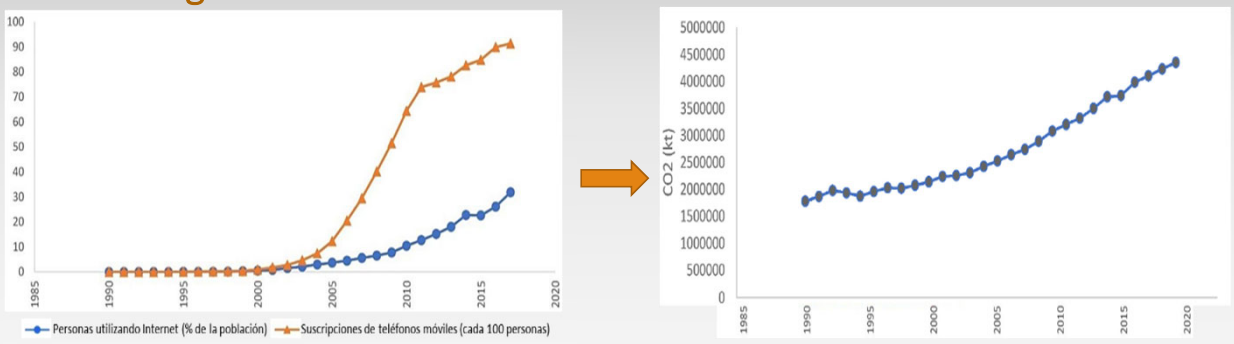
19

Previsiones sobre el consumo energético de las TIC

- Hay multitud de artículos sobre las previsiones del consumo energético originado por las TIC.
- Suelen referirse a países o grupos de países concretos y no son globales, y por lo general, utilizan modelos de regresión utilizando diversos parámetros, a veces muy reducidos, para hacer las previsiones.
- En general, las previsiones realizadas hace dos décadas eran excesivamente pesimistas, pero han servido para crear conciencia en las comunidades científica y tecnológica sobre el problema.
- No eran rigurosas ya que partían de hipótesis muy parciales como, en algún caso considerar, sin más, que el consumo crece proporcionalmente con el número de dispositivos, usuarios o tráfico de datos; obviando incluir otros parámetros como que la eficiencia energética de estos dispositivos es cada vez mejor.

20

Tendencias del porcentaje de población que usa Internet y móviles celulares, y de las emisiones de CO2 (KT), países emergentes

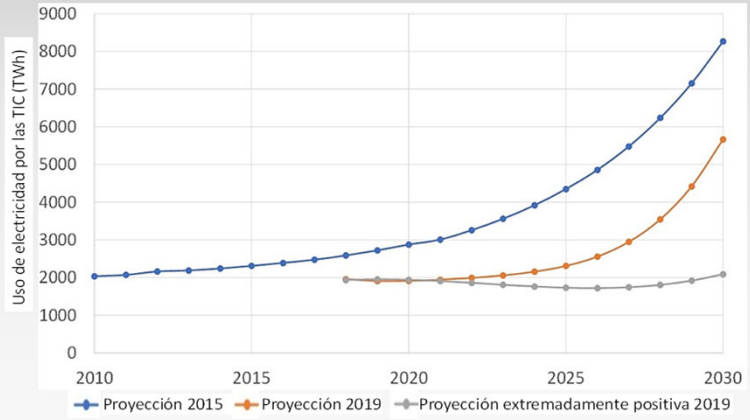


- Observación: no considera otros parámetros relevantes como número y tipos de dispositivos, ni mejoras previstas en eficiencia energética

A. Haldar y N. Sethi. (2022). Environmental effects of Information and Communication Technology-Exploring the roles of renewable energy, innovation, trade and financial development. Renewable and Sustainable Energy Reviews, 153, 111754

21

Proyecciones de Andrae and Edler sobre el uso de energía eléctrica por las TIC en TW·h por año (2015 y 2019)



$$E = P \cdot t$$

1 W·h = 3,600 Joules

- A.S. Andrae y T. Edler. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157. https://doi.org/10.3390/challe6010117
- A.S. Andrae (2019). Comparison of several simplistic high-level approaches for estimating the global energy and electricity use of ICT networks and data centers. International Journal, 5, 51. DOI: 10.30634/2414-2077.2019.05.06

22

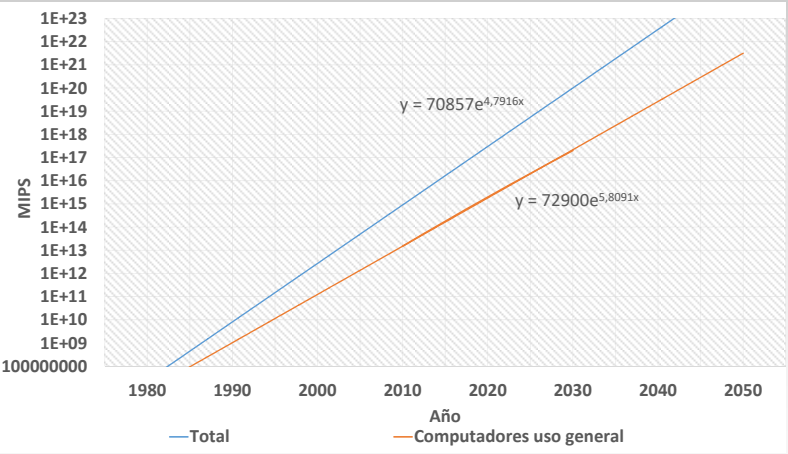
Para medir el consumo de energía. Generalmente se estima:

- En computación durante un tiempo *t* (día, año, etc.):
  - Volumen de cálculos (bits, instrucciones, etc.) x consumo medio de cada cálculo
- En tráfico de información
  - Volumen de información transmitida (bits) x Consumo medio por bit
- Eficiencia energética (EE) en computadores

$$EE = \frac{Instrucciones}{Energía\ para\ ejecutar\ esas\ instrucciones} = \frac{Instrucciones/s}{Energía/s} = \frac{R_{max}}{P} = \frac{FLOPS}{Vatios}$$

23

Volumen de computaciones, partiendo de datos de diversos autores



- En el año 2040 se estima llegar a una potencia de procesamiento de ≈$5x10^{22}$ MIPS, una gran parte de los cuales (≈ $5x10^{19}$ MIPS) corresponderán a procesamiento de uso general.

- M. Hilbert y P. Lopez, "The world's technological capacity to store, communicate, and compute information," Science 332 (2011) 60-65
- F. Roccaforte, F. Giannazzo y G. Greco. (2022, Enero). Ion Implantation Doping in Silicon Carbide and Gallium Nitride Electronic Devices. Micro 2(1) pp. 23-53)
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015)

24

Consumo por bit.
Energía mínima teórica consumida por bit. Principio de Landauer

- A temperatura constante en un proceso en un sistema aislado se verifica:

$$S_2 - S_1 \geq \frac{Q_{1\to 2}}{T}$$

- S_2: entropía final; S_1: entropía inicial, $Q_{1\to 2}$: cantidad de calor intercambiada entre el sistema y el entorno; T: temperatura absoluta
  - El cambio de entropía es igual al calor intercambiada partido por T
  - La entropía mide el desorden o la incertidumbre
  - Procesamiento de un bit → Disminución de entropía → disipación de calor
- Fórmula de la entropía (S) de Boltzmann-Plank:

$$S = k_B \cdot ln\, W$$

  - $k_B \approx 1.38\times10^{-23}$ J/K; W es el número de alternativas posibles de estado. S= Q/T

$$Q \geq k_B \cdot T \cdot ln\, 2$$

- Con temperatura = 20^0; **energía mínima disipada en forma de calor por bit procesado ≈ *$3\cdot10^{-21}$ J/bit***.

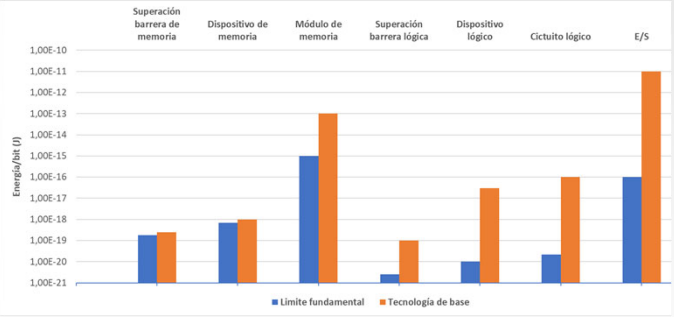
25

- En el caso de computaciones reversibles (computación cuántica) se debe utilizar como límite inferior del consumo energético el establecido por el Teorema de Margolus–Levitin: $6\cdot10^{33}$ operaciones por segundo por julio de energía.

26

V. Zhirnov et al. publican (2014) un artículo donde:

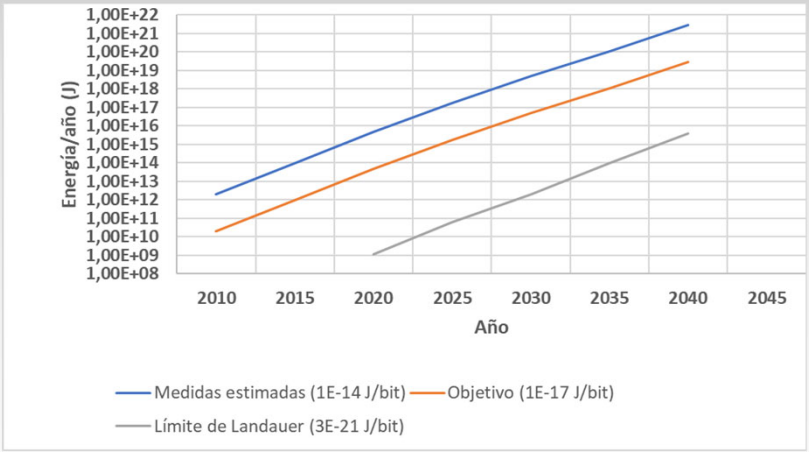
- Con datos reales obtienen la evolución del consumo por bit de diferentes elementos binarios (dispositivos lógicos y elementos de memoria) considerando el consumo de transistores y microprocesadores individuales y la dinámica de los procesos físicos en los diferentes componentes (capacitivos, resistivos, etc.).
- Llegan a la conclusión de que, en situaciones típicas;
  - La energía mínima requerida por transición (conmutación) de un bit es de alrededor de ≈ **10^{-14} J/bit.**
  - Valor estimado como un objetivo alcanzable ≈ **10^{-17} J/bit.**



- V. Zhirnov, R. Cavin y L. Gammaitoni. (2014). Minimum energy of computing, fundamental considerations. In ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology. IntechOpen
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015). Append A4.

27

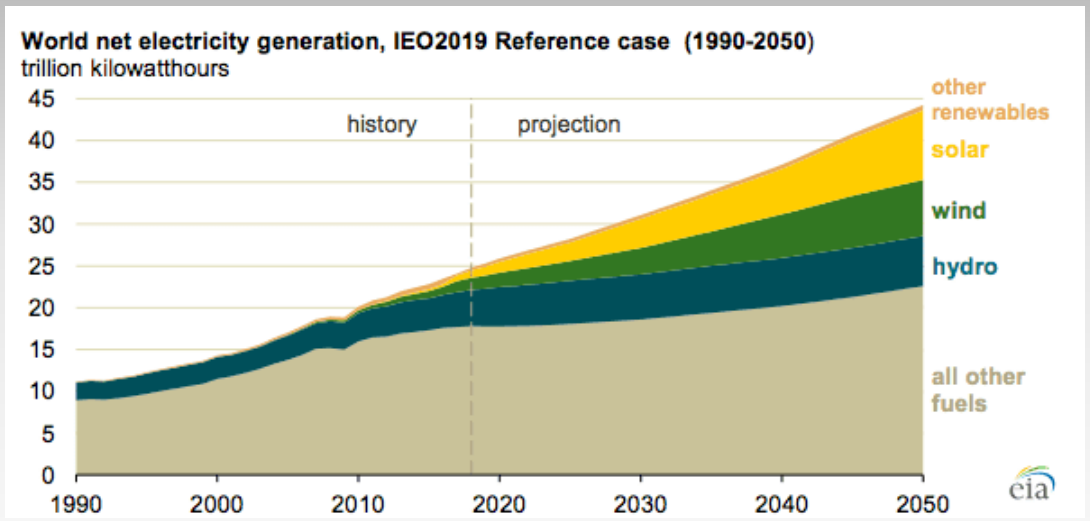
Energía eléctrica global consumida por la informática en un año



- Valor calculado a partir de datos reales (2014): ≈**10^{-14} J/bit.**
- Valor estimado por Zhirnov, como un objetivo a lograr ≈ **10^{-17} J/bit**.
- Límite de Landauer ≈ **$3 \cdot 10^{-21}$ J/bit.**

- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015)

28

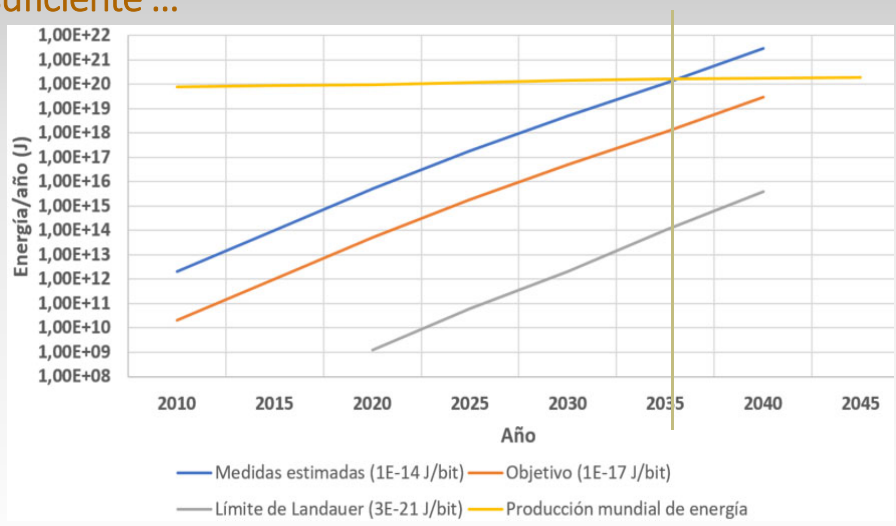


Source: U.S. Energy Information Administration, International Energy Outlook 2019

https://www.powermag.com/eia-renewables-will-account-for-half-of-global-power-generation-by-2050/

29

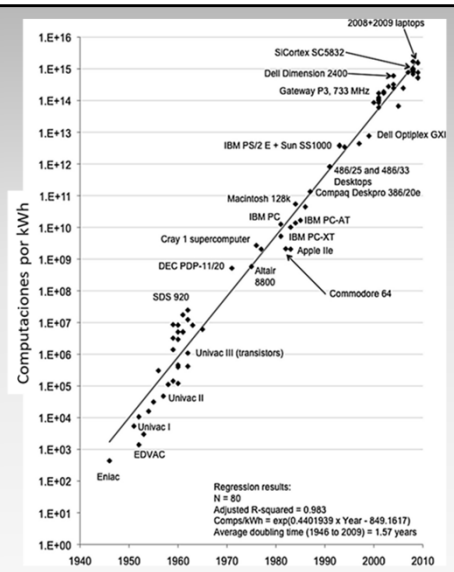
En el peor de los casos, en 2035 no habrá energía eléctrica suficiente …



30

Ley de Koomey (2010)

- Estima que el nº de computaciones por julio de energía disipada se dobla aproximadamente cada **1,57** (2016→ 2,6).
  - Esta Ley se cumplía con una gran precisión (R^2=98%) con datos tomados entre los años 1946 y 2009.



Fuente: Jonathan Koomey

- J. Koomey, S. Naffziger. (2015). Moore's Law might be slowing down, but not energy efficiency. IEEE Spectrum, 52(4), 35
- J. Koomey, S. Berard, M. Sanchezn y H. Wong. (2010). Implications of historical trends in the electrical efficiency of computing. IEEE Annals of the History of Computing, 33(3), 46-54
- Koomey, J., & Naffziger, S. (2016). Energy efficiency of computing: What's next. Electronic Design, 28.

31

El TOP500 y GREEN500

- TOP500, junio 1993
- GREEN500,junio 2013


TOP500 CERTIFICATE

n Fire x2200 M2/X4600 M2 Cluster, Opteron 2.2 Ghz, Infiniband, Sun Microsyst

Universidad de Granada, Spain

is ranked

No. 467

among the world's TOP500 Supercomputers with

4.222 TFlop/s Linpack Performance

The 29th TOP500 list was published at ISC'07 in Dresden, Germany June 23rd 2007

Congratulations from The TOP500 Editors

Hans Meuer, University of Mannheim
Erich Strohmaier, NERSC/Berkeley Lab
Jack Dongarra, University of Tennessee
Horst Simon, NERSC/Berkeley Lab

- https://www.top500.org/lists/green500/

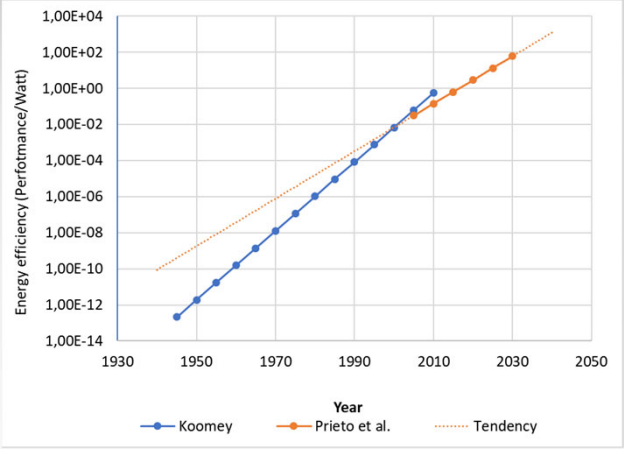
32

Nuestros resultados sobre eficiencia energética

- Considerando las listas TOP500 y GREEN500, desde junio 2008 a junio de 2023 (30 listas con un total de 9,682 HPC).
- Concluimos que el crecimiento es exponencial:

$$EE = 2 \cdot 10^{-265} \cdot e^{0.3026 \cdot year}$$

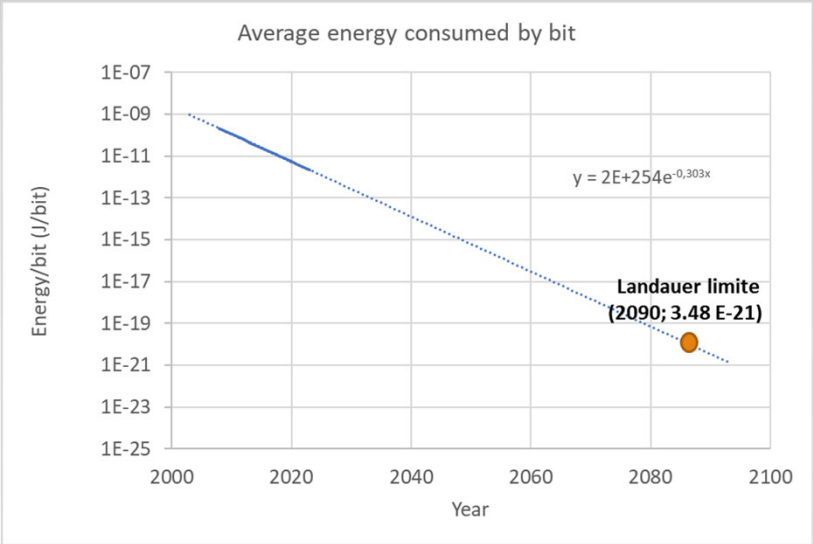
  - coeficiente de determinación r^2 = 0.9907
- **La eficiencia energetica dobla cada 2.29 años**
  - Koomey 2011 → 1,57; 2016→ 2,6.



- A. Prieto, B. Prieto, JJ Escobar, T. Lampert (2024). Koomey's Law on the Evolution of Computing Energy Efficiency Revisited, *Remitido.*

33

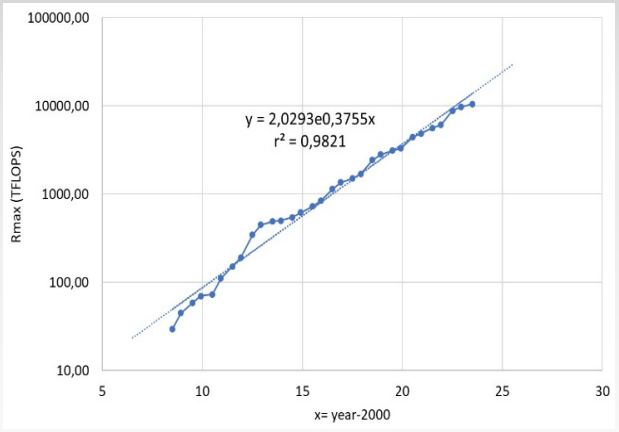
Con nuestros resultados el límite de Landauer se alcanzaría en 2090



34

Hemos obtenido también la evolución el rendimiento de computación (últimos 15 años)

- Dobla cada 1,85 años
- PROBLEMA:
  - **La eficiencia energética crece más lentamente (2.29 años)**
  - Aunque las diferencias parecen pequeñas, doblar la eficiencia energética cada **1.85 años** significa incrementarla aproximadamente 43 veces en una década mientras que doblar cada **2.29 años** supone incrementar por 21 por década.



35

Objetivo para ingenieros en computadores

- Consumo global de energía.
  - GE = Nº de computaciones · energía consumida en cada computación = NC · EC
- Eficiencia energética.
  - $EE = \frac{Number\ of\ computatios}{Energy\ consumed\ by\ those\ computations} = \frac{NC}{E}$
- Consumo global de energía
  - $\mathbf{GE} = \frac{NC}{EE} = \frac{Demanda}{Eficiencia\ energética}$
- Los científicos e ingenieros en computadores tenemos que centrarnos en el denominador (**incrementar la eficiencia energética**)

36

Los resultados se pueden generalizer a PCs

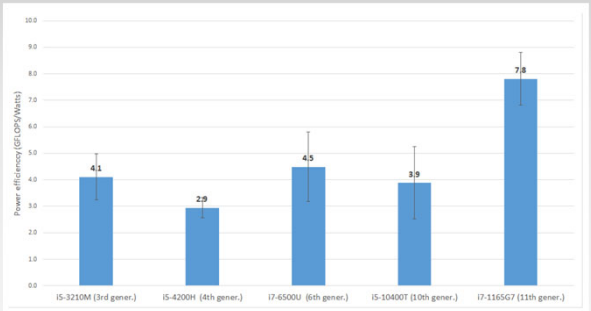


Figure 15. Power efficiency of the microprocessors measured during the Linpack run.

Table. 12. Position, in terms of power efficiency, of the under test systems within the Green500 list [34].

Platform	Position in Green500	Pos. in TOP500	Rmax TFlops	Power (KW)	Power Efficiency (GFlops/W)
MN-3 - MN-Core Server, Xeon	1	301	2,181.2	55	39.379
SSC-21 Scalable Module - Apollo 6500 Gen10 plus, AMD.	2	291	2,274.1	103	33.983
Etc.					
***SUT5. ASUS 2**			**0.148**	**0.033**	**4.434**
JOLIOT-CURIE SKL - Bull Sequana.	79	113	4,065.6	917	4.434
Etc.					
***SUT2. Toshiba 1**			**0.064**	**0.036**	**1.780**
Nuri - Cray XC40, Xeon	155	251	2,395.7	1,359	1.762
Etc.					
***SUT4. HP**			**0.112**	**0.069**	**1.633**
HKVDPSystem – Sugon TC6000, Xeon	162	355	1,979.0	1,216	1.627
***SUT3. ASUS**			**0.075**	**0.093**	**0.799**
***SUT1. SONY**			**0.036**	**0.046**	**0.792**
Thunder – SGI ICE X, Xeon	173	156	3,126.2	4,820	0.649
Etc.					

* System Under Test

- Prieto, B., Escobar, J. J., Gómez-López, J. C., Díaz, A. F., & Lampert, T. (2022). Energy Efficiency of Personal Computers: A Comparative Analysis. Sustainability, 14(19), 12829

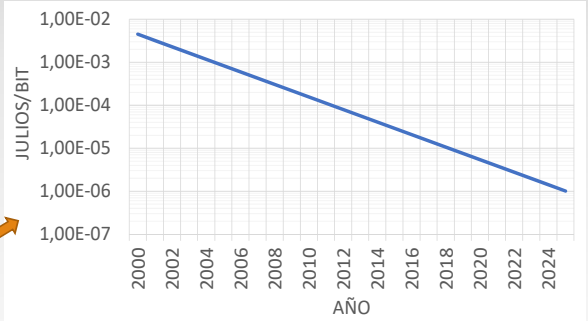
37

Consumo en tráfico de información digital

38

Estimación del consumo de energía en el tráfico por Internet

- Es difícil ya que este depende de muy diversos factores:
  - El canal de transmisión (atmosfera, cable, fibra óptica, etc.)
  - La distancia entre emisor y receptor.
  - El caudal de datos (velocidad de transmisión).
  - En el caso de un mensaje transmitido, depende además de codificación utilizada, tipo de modulación, etc.
- Estimaciones de Aslan y colbs.:
  - El consumo por bit se reduce a la mitad aproximadamente cada 2 años (en procesamiento cada 2,6 años).

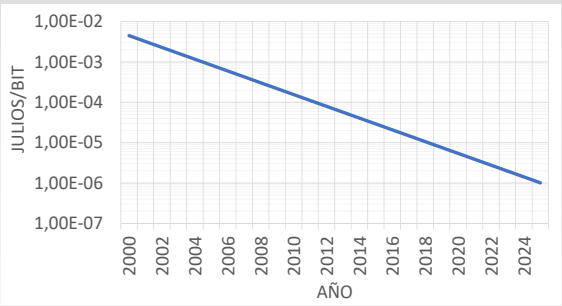


- J. Aslan, K. Mayers, J. G. Koomey y C. France. (2018). Electricity intensity of internet data transmission: Untangling the estimates. Journal of industrial ecology, 22(4), 785-798.

39

Estimación del consumo de energía en el tráfico de 1 bit por Internet

- La energía media de transmisión de 1 bit en 2020 a través de Internet ha sido del orden de **$2{,}77 \cdot 10^{-6}$ J/bit**, lo que supone:
  - ≈ $2{,}8 \cdot 10^{11}$ veces mayor que la del procesamiento de un bit (10^{-17} J/b), si se tienen en cuenta las estimaciones de Zhirnov o de
  - 6 ordenes de magnitud mayor si se tiene en cuenta la eficiencia energética de los supercomputadores obtenida por nosotros ($5{,}5 \cdot 10^{-12}$ J/bit)



40

Algunos datos de situaciones prácticas

41

Ejemplo y ejercicio interesante: colas de emails UGR

- Texto oficial establecido por la Secretaría General de a UGR:
  - *Este mensaje ha sido generado desde una cuenta de la Universidad de Granada para los fines propios de la institución. Su contenido se considera información confidencial, por lo que queda informado de que su utilización, divulgación o copia sin autorización no está permitida. Si usted ha recibido indebidamente el correo le rogamos que advierta de ello por esta misma vía al remitente y proceda a su eliminación. Cualquier incidencia relacionada con la recepción de nuestros correos electrónicos y en particular las relativas a la seguridad y confidencialidad pueden ser comunicadas a protecciondedatos@ugr.es. Para más información al respecto, puede consultar nuestra política de privacidad .*
  - *This message has been generated from an e-mail address of the University of Granada for the institution's own purposes. Its content is considered confidential information, so it is informed that its unauthorized use, disclosure or copying is not permitted. If you have improperly received the email please warn the sender of this, same way and proceed to its removal. Any incident related to the receipt of our emails and in particular those related to security and confidentiality may be communicated to protecciondedatos@ugr.es. For more information, please refer to our privacy policy.*
- 1.285 caracteres (sin contabilizar los enlaces)

42

Cálculos sobre emails de la UGR

UGR emails (2021-2022):

- 6,819,020 mensajes enviados a Internet desde estafetas centrales o departamentales.
- 4,671,665 mensajes provenientes de Internet y entregados en buzones UGR de PAS/PDI.
- 22,278,138 mensajes provenientes de Internet y entregados en buzones UGR de Alumno.

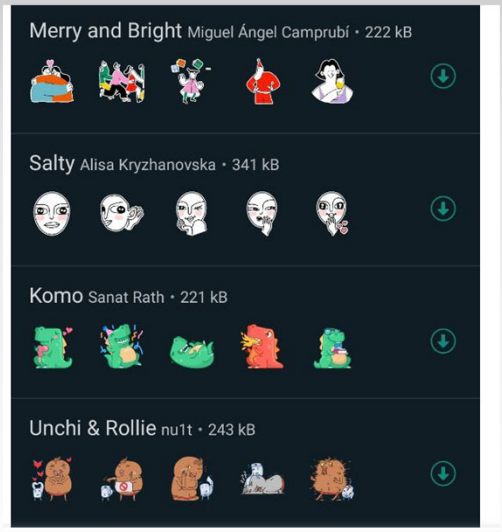
		Mensajes enviados	Mensajes recibidos
Nº caracteres/email		1.285,00	1.285,00
Nº bits (UTF8)/carácter		16	16
Mensajes anuales		6.819.020	56.949.803
Nº de bits anuales		1,40199E+11	1,17089E+12
Energia bit transmitido (2022)	Julios/bit	3,00E-06	3,00E-06
Energia total	Julios	4,21E+05	3,51E+06
Energía total	KWh	1,17E+02	9,76E+02
MIX eléctrico	Kg CO2/KWh	0,351	0,351
Huella de carbono cola emails	KgCO2	4,10E+01	3,42E+02

¡Media Tonelada de CO2 al año!

43

Emoticones, emojis y pegatinas (*stickers*) en WhatsApp

- Emoji: 4 bytes
- Pegatina ≈ 50 KB, <100 KB → 0,120 g de CO2
- 20.000 millones de WhatsApp diarios.
- Si 1 de cada 10 es una pegatina → **87.500 T de CO2 al año**.



Para que os hagáis una idea, **en el caso del *pack* "Chummy Chum Chums", oficial de WhatsApp, sus 16 *stickers* solo pesan 3,7 megas**, por lo que están muy lejos de ese máximo que tenemos disponible (1MB/pegatina).

44

Jueves 17.12.20
IDEAL

La contaminación silenciosa

ALBERTO PRIETO ESPINOSA
Academia de Ciencias de Granada

Todo proceso de transferencia o movimiento de datos, entre móviles, computadores, etc, consume energía, y muchos dispositivos, entre los que se encuentran los supercomputadores, están funcionando las 24 horas del día, siendo los consumos muy altos

Uno de los retos más importantes de la sociedad actual es reducir el consumo de energía con el objeto de mantener o hacer posible la sostenibilidad de nuestro planeta. Por hacer referencia a nuestro contexto, la Unión Europea tiene como una

y de los programas que es mucho mayor. Así, cuando enviamos un correo electrónico, además del consumo inherente a la transmisión de los bits hay que añadir el del programa que me permite editar, enviar, recibir y visualizar los emails.

Hasta ahora los parámetros que se utilizaban para medir las prestaciones de un

A. Prieto. La contaminación silenciosa. Ideal, 17/12/2020, p. 24

45

Aprendizaje en Chat GPT-3 175B

- Contenido:
  - 175 mil millones de parámetros (pesos de redes neuronales).
  - 800 GB de memoria
  - Procesamiento de consultas de hasta 2048 tokens (palabras o subpalabras), se consideran como "ventana contextual". Eso significa que tiene 2.048 pistas a lo largo de las cuales se procesan los tokens sucesivos.
- Aprendizaje
  - Última actualización enero de 2022.
  - **Pre-entrenamiento no supervisado de red neural profunda (96 capas) con un corpus de 570 mil millones de tokens recopilados de Internet en sitios como Wikipedia (3 mil millones), Common Crawl (410 mil millones),** …
- Generación de texto
  - 96 capas de ANN transformers decoders (palabra + posición en la frase, etc.) que generan textos que simulan la redacción humana (concepto de "atención", semántica de las palabras, contexto, etc.)

46

Pre-entrenamiento de GPT-3 750B

- Tiempo total de computación para el pre-entrenamiento:
  - ≈ 3.000 PFLOPS·días

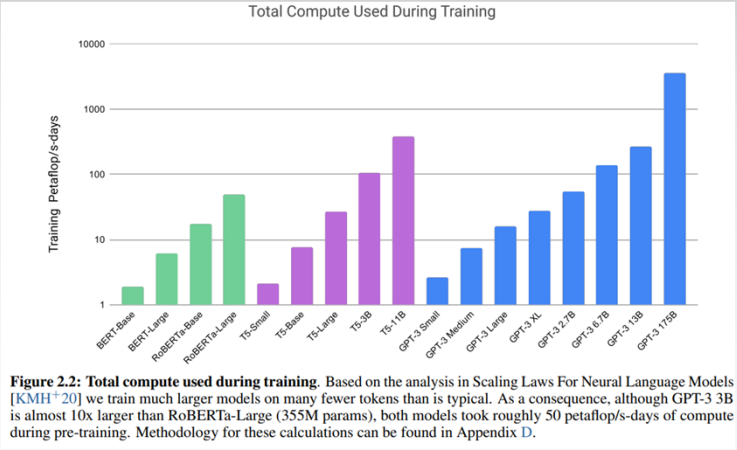


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. Cited by 12,240

47

Supongamos que el entrenamiento lo efectuamos con los #1 del TOP500 y Green500 (Junio 2023) …

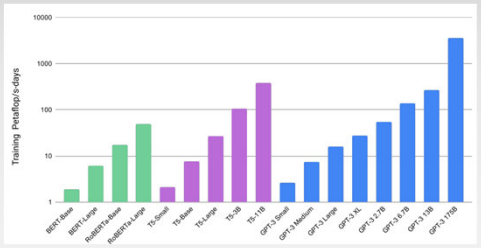
Green500 Data

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
6	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	22,703	52.592

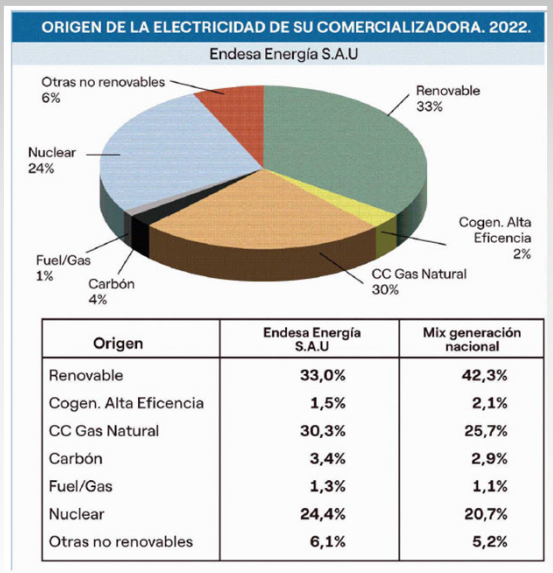
48

El preentrenamiento a gran escala requiere grandes cantidades de cálculo, lo que consume mucha energía (ChatGPT-3 175B)

	1º TOP500 (Junio 2023)	1º Green500 (Junio 2023)
	Frontier	Henri
GPT-3 entrenamiento. PetaFLOPS·dias	3.000,00	3.000,00
Gigaflops-dia	3,00E+09	3,00E+09
Eficiencia energética (GFlops/watts)	52,592	65,396
W·dia	5,70E+07	4,59E+07
KW·h	1,37E+06	1,10E+06
MIX (Kg CO2/KWh)	0,351	0,351
Huella de carbono (Kg CO2)	480.529,36	386.445,65
Prestaciones, Rmax (PFlops)	1194,00	2,88
Tiempo ejecución (s)	2,17E+05	9,00E+07
Tiempo de ejecución (dd:hh:mm:ss)	02:12:18:05	06:16:00:00



480.5 toneladas de CO2 (2 dd 12 hh) vs 386.4 ton (6 dd 16 hh)



Origen	Endesa Energía S.A.U	Mix generación nacional
Renovable	33,0%	42,3%
Cogen. Alta Eficencia	1,5%	2,1%
CC Gas Natural	30,3%	25,7%
Carbón	3,4%	2,9%
Fuel/Gas	1,3%	1,1%
Nuclear	24,4%	20,7%
Otras no renovables	6,1%	5,2%

Emisiones de CO_2 equivalente
Endesa Energía S.A.U

Emisiones CO_2 eq. (g/kWh)	**188**
Media nacional (g/kWh)	**162**

- El consumo de pre-entrenamiento se amortiza según se va usando.
- Una vez entrenado, en su uso normal (inferencia), generar 100 páginas de contenido a partir de un modelo entrenado puede consumir del orden de 0,4 kW-h.

0.4 KW·h x 0.351 = 0.14 Kg CO2 a la hora

51

Procedimientos y técnicas para incrementar la eficiencia energética en las TIC

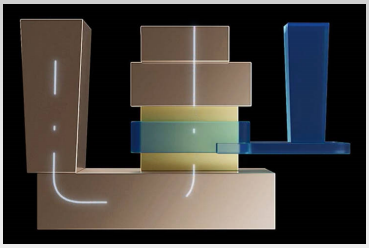
52

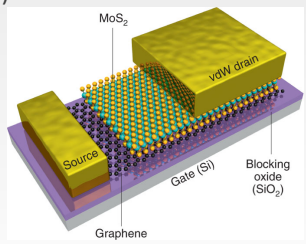
- Afortunadamente las previsiones más pesimistas (mediados de la década de los 2010) sobre el consumo energético asociado a las TIC no se están cumpliendo, siendo menor que el esperado.
  - La industria alrededor de las TIC es consciente del problema y se están invirtiendo recursos financieros y políticas activas para reducir el incremento del consumo, tanto en la fabricación de nuevos productos como en el consumo inherente a su uso.
- Acciones para reducir el consumo:
  A. **Mejoras tecnológicas en los componentes electrónicos y dispositivos**
  B. **Gestión y planificación del uso de los recursos**.
  C. **Cambios de escala**

53

A. Mejoras tecnológicas en los componentes electrónicos y dispositivos …

- Cambios en los dispositivos y en la arquitectura interior de los microchips.
  - Prototipo de IBM de CI que hace posible apilar verticalmente los transistores-
  - Incremento de la densidad de integración y reducción del consumo de energía (se estima que **hasta un 85% menos**).





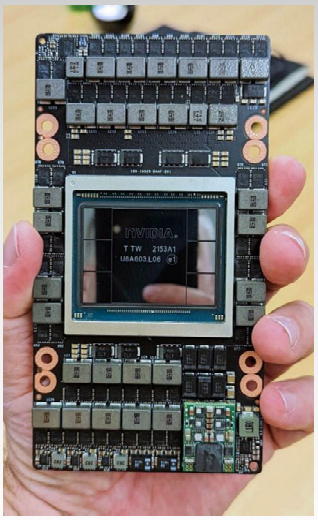
IBM's VTFET with a vertical channel (yellow) and gate-all-around (blue). Contacts are brown and the white line shows current flow.

- Zhang, J., Gao, F., & Hu, P. (2021). A vertical transistor with a sub-1-nm channel. Nature Electronics, 4(5), 325-325.
- Steve Bush, (14 diciembre 2021) IBM beats finFETs with vertical CMOS at IEDM. Electronics Weekly.com

54

… mejoras tecnológicas en los componentes electrónicos y dispositivos …

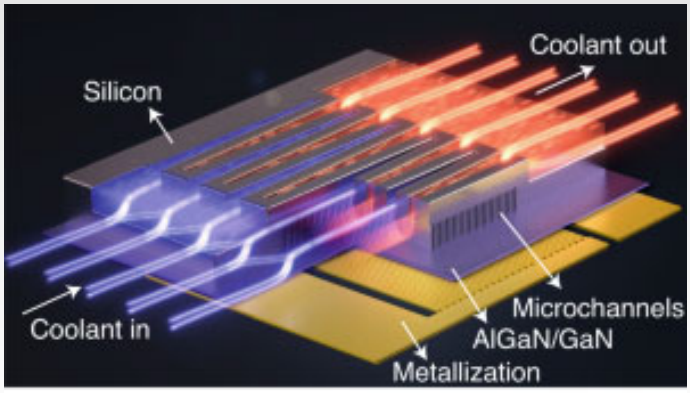
- Inclusión de funciones de gestión de energía dentro de las CPUs con los que, dependiendo de la carga de trabajo, cambian dinámicamente entre diferentes estados de energía (modo de espera, por ejemplo).
- Desarrollo de procesadores de uso específico para ámbitos o funciones concretas, como GPU y TPU.
- Fuentes de alimentación AC/DC conmutadas; introducción de nuevos materiales, como el nitruro de galio y el carburo de silicio, que permiten diseños a más altas frecuencias.



55

… mejoras tecnológicas en los componentes electrónicos y dispositivos …

- Integración directa en los chips de sistemas de refrigeración con microfluidos, que sustituyan a los ventiladores externos.

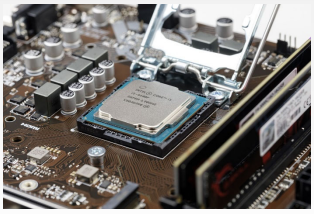


• Varnava, C. Chips cool off with integrated microfluidics. Nat Electron 3, 583 (2020). https://doi.org/10.1038/s41928-020-00494-5

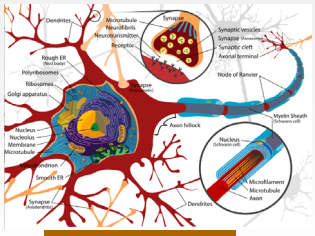
56

… mejoras tecnológicas en componentes y dispositivos electrónicos

- Cambios en la tecnología de otros dispositivos. HDD → SSD, reducción del consumo de energía superior al 50%
- Computación neuromórfica
  - El cerebro humano es uno de los sistemas más eficientes energéticamente ya que consume una potencia de unos 25 vatios (menos de la mitad de un PC portátil) y dispone de 86.000 millones de elementos informáticos (neuronas).
    - Opino que el bajísimo consumo energético se debe más que a la arquitectura a la tecnología subyacente (material) → en lugar de silicio, arseniuro de galio, etc. el del cerebro es de naturaleza biológica (bioquímica, células, tejidos,...)



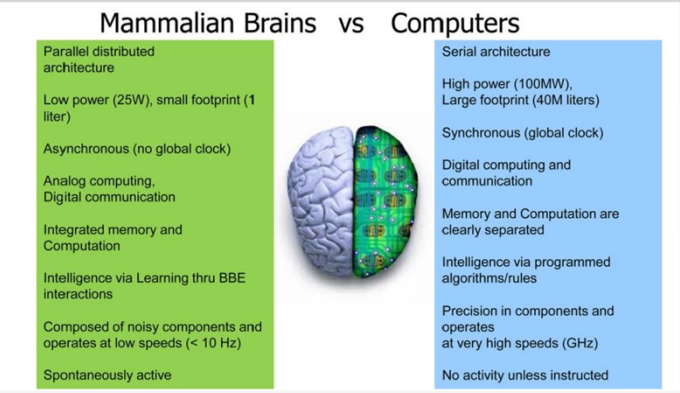
pixabay



Wikimedia Commons

57

¿Por qué no podemos hacer en la actualidad computadores que procesen datos como lo hace el cerebro?



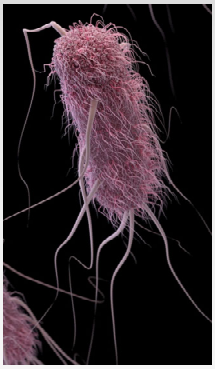
Cerebro completo $86 \cdot 10^9$ neuronas ; 1 chip ELSI → 10^8 transistores ;

- Avram Bar-Cohen (DARPA), "Cognitive computing, Towards the electronic brain," presentation at the Workshopon Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015)

58

Nivel inferior de procesamiento, al de las neuronas: células

- Las unidades más pequeñas de materia viva (**células**), poseen capacidades computacionales asombrosas, siendo los procesadores de información más pequeños conocidos.
  - Las células vivas individuales (bacterias, p.e.) son capaces de realizar procesamientos de uso general con propiedades análogas a los computadores convencionales.
    - La molécula de **ADN actúa como memoria no volátil** del computador celular, mientras que muchas **proteínas en el citoplasma** tienen como principal función la transferencia y **procesamiento** de información, por lo que pueden considerarse como los elementos lógicos del procesador biológico de la célula.
    - Se ha estimación del contenido de información de una célula viva, por ejemplo para la bacteria E. coli, siendo ≈10^{11}-10^{12} bits (Las mediciones experimentales de reducción de entropía del contenido de información de las células bacterianas utilizando técnicas micro-calorimétricas arrojaron resultados muy similares).



- V. Zhirnov, R. Cavin y L. Gammaitoni. (2014). Minimum energy of computing, fundamental considerations. In ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology. IntechOpen

59

	Procesador celular biológico	Tecnologías base de las TIC
Memoria	10^7 bit	10^7
Logic	10^6 bit	10^6
Energía por bit	~10^{-20} J (valor medio a nivel de sistema)	10^{-13} (Memoria) 10^{-16} (Logica)
Rendimeinto binario	10^7 bit/s	10^7 bit/s
Tiempo de tarea	3000 s	3000 s
Energía total consumida por la tarea	10^{-10} J	10^{-6} J

60

B. Gestión y planificación del uso de los recursos:

- Utilizar los distintos sistemas tratando de reducir el consumo energético global, como:
  - Hacer entrar en los **modos de suspensión o de espera** a los recursos (servidores, sistemas de memoria etc.) que en un momento dado no sean necesarios.
  - **Escalado dinámico de la tensión y de la frecuencia** (*Dynamic Voltage and Frequency Scaling,* DVFS).
    - Ejecutar lentamente los programas que no necesiten un tiempo de respuesta muy corto.
    - Se estima que, **si la frecuencia de reloj se reduce a la mitad**, el tiempo de ejecución se duplica, pero **el consumo energético se reduce a una cuarta parte**.
  - Ejecutar, en lo posible, las aplicaciones dentro de "**horas valle**" donde la producción de energía eléctrica procedente de fuentes limpias es mayor por ser más elevada la producción eólica debido al viento u horas donde la radiación solar es mayor. Doble beneficio:
    - se reduce el coste económico de la energía necesaria para la ejecución de los programas
    - se favorece el uso de las energías alternativas.

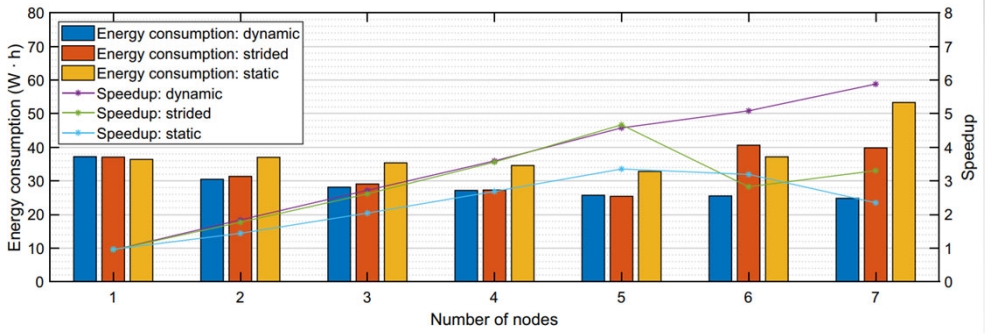
61

C. Cambios de escala

- La proliferación de teléfonos inteligentes y **pequeños dispositivos móviles** da lugar a una reducción del consumo energético ya que cada uno de ellos ofrece multitud de funciones y servicios que antes realizaban dispositivos de consumo independientes.
- Planificación y asignación de tareas a los recursos hardware disponibles teniendo en cuenta su **eficiencia energética**. En particular, debe explotarse el paralelismo de las aplicaciones y de los programas buscando la mejor eficiencia posible.
  - En muchos casos, una asignación eficiente de recursos requiere el rediseño de las aplicaciones y de los algoritmos.
- **Endoso computacional** (*offloading*): los procesos que requieren tareas informáticas intensivas se transfieren (endosan) a una plataforma externa, que puede ser desde un acelerador de hardware hasta un sistema de clúster, o recursos en la nube. **Tecnologías de virtualización**.
  - Sólo es beneficioso cuando se requiere gran volumen de computación con relativamente poca cantidad de comunicación.
- Fusión o transformación de centros de datos medianos a en **centros de datos hiperescala** (mucho mayores) (Google Cloud, Amazon Web Services, Microsoft Azure, OVHCloud, o Rackspace Open Cloud), donde el consumo de energía se gestiona mucho mejor.

62

Ejemplo de considerar el consumo energético: Clasificación de EEG con 3600 características, utilizando mRMR-KNN : los resultados dependen de la distribución de trabajo entre los nodos del clúster. Problema: clasificar 3 movimientos imaginados (mano derecha, mano izquierda, pies)



(a) Speedup and energy consumption when increasing the number of computing nodes

¡La implementación paralela con 7 nodos reduce a un 13,38% el consumo de la secuencial!

- Juan José Escobar, Francisco Rodríguez, Beatriz Prieto, Dragi Kimovski, Andrés Ortiz, Miguel Damas (2023). A distributed and energy efcient KNN for EEG classification with dynamic money saving policy in heterogeneous clusters. Computing. https://doi.org/10.1007/s00607-023-01193-7

63

Centros de datos: Nacional de Supercomputación: Mare Nostrum

- Consumo eléctrico
  - MareNostrum4 → 2 MW
  - MareNostrum5 (en fase de inicio de producción) → 8 MW
- Varios millones de € al año.
- Top500 de junio 2023:
  - MareNostrum 4 (máquina del 2017). Eficiencia energética:
    - partición con procesadores de uso general , 3,96 Gflops/W.
    - maquina basada en Power9 y NVIDIA V100, 14,13 Gflops/W.

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
124	98	**MareNostrum** - Lenovo SD530, Xeon Platinum 8160 24C 2.1GHz, Intel Omni-Path, Lenovo Barcelona Supercomputing Center Spain	153,216	6.47	1,632	3.965

64

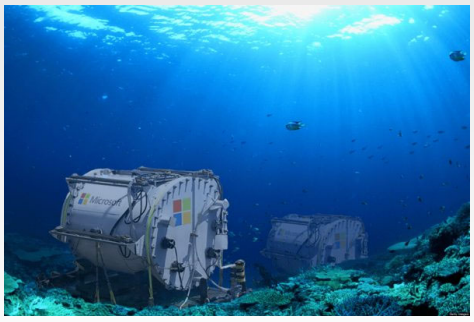
Planta de enfriamiento de Google en Hamina (Finlandia)

- Climatización muy critica y su funcionamiento requiere un gran consumo energético.
- Al ser un país nórdico el coste de la climatización es mucho más bajo que en países más cálidos.
- Este centro utiliza el agua del gélido mar del golfo de Finlandia para refrigerar todas sus instalaciones.

65

Proyecto Natick de Microsoft

- Se sumergieron durante 2 años (2018 a 2020) 864 servidores en un contenedor similar a un submarino.
- Ubicación en las Islas Orcadas, en el norte de Escocia: aguas gélidas y la red eléctrica se abastece al 100% de energía eólica, solar y marina, etc. obtenida en las cercanías. No contaba con refrigeración activa.



- https://news.microsoft.com/es-es/2020/09/15/proyecto-natick-el-futuro-de-los-centros-de-datos-bajo-el-mar-es-fiable-practico-y-sostenible/

66

Proyecto Natick de Microsoft

- Los servidores experimentaron una **tasa de fallos ocho veces inferior** a lo esperado en un Centro de Datos convencional, gracias, entre otras cosas, a la atmósfera de nitrógeno empleada en la cápsula sellada.
- Se rescató del fondo marino, cubierto de algas, percebes y anémonas.
- Se concluyó que **el futuro de los centros de datos bajo el mar es fiable, práctico y sostenible.**

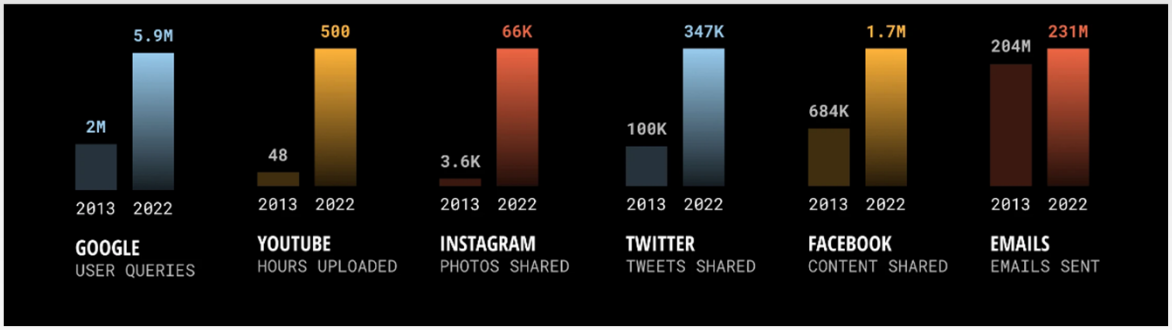
67

Conclusiones

68

Conclusiones …

- Los datos nunca duermen, de forma que **cada minuto**:



Informe Data Never Sleeps 10.0, realizado por Domo. https://www.domo.com/data-never-sleeps

69

… conclusiones …

- Se estima que en el año 2030 las tecnologías de la información consumirán aproximadamente el 13% de la electricidad mundial, y para 2050 el de los centros de datos será unas tres veces mayor que la cantidad total de energía generada en Japón.
- El almacenamiento de datos, por otra parte alcanzaría en el año 2025 los 163 Zettabytes ($163 \cdot 10^{21}$bytes)
- La reducción del consumo energético en el ámbito de las TIC es una cuestión trascendental, y debe ser afrontada desde muy distintos ámbitos (computer engineering, software engineering, dissemination, teaching, etc.)

70

… conclusions.

- En los diseños, arquitecturas, programas y aplicaciones de circuitos de las TIC el **consumo de energía** debe considerarse como una medida de prestaciones tan importante como el **rendimiento computacional.**
- La información sobre la eficiencia energética de los sistemas debe incluirse en publicaciones científicas y técnicas, lo que no suele hacerse.
- Por un lado, hay razones medioambientales y económicas, pero también la necesidad de mejorar la autonomía de los dispositivos que utilizan baterías.
- La sociedad debe estar informada de que el uso de las TIC (sea cual sea su forma), lleva implícito un consumo energético. **¡Todos debemos contribuir, desde nuestros respectivos ámbitos, al reto de lograr la sostenibilidad de nuestro planeta!**

71

Agradecimientos

- Deseo agradecer a las siguientes personas su colaboración en las investigaciones que estamos realizando sobre este tema:
  - Beatriz Prieto
  - Juan José Escobar
  - Miguel Damas
  - Antonio Díaz
  - Christian Morillas
  - Jesús González Peñalver
  - Andrés Ortiz (UMA)
  - Francisco Gil (UAL)
  - Francisco Illeras
- Nuestras investigaciones en este ámbito actualmente se están financiando parcialmente por el Ministerio de Ciencia e Innovación, junto con fondos FEDER de la UE, a través del proyecto PID2022-137461NB-C31.

72

¡Muchas gracias por tu atención!

Alberto Prieto Espinosa. Conferencias
https://icar.ugr.es/informacion/directorio-personal/alberto-prieto-espinosa/web/conferencias