Computación sostenible: optimización energética en procesamiento de señales y datos

ALBERTO PRIETO ESPINOSA
Profesor Emérito del Departamento de Ingeniería de Computadores, Automática y Robótica (Univ. de Granada)

17 de abril 2024

Máster Oficial en Telemática y Redes de Telecomunicación

E.T.S. DE INGENIERÍA DE TELECOMUNICACIÓN UNIVERSIDAD DE MÁLAGA

ICAR INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

1

Contenido:

ICAR INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- 1. Contribución de la informática y las comunicaciones digitales al consumo energético.
- 2. Evolución del consumo energético de computación.
- 3. El consumo energético por tráfico de datos digitales.
- 4. Consumo energético de aplicaciones de Inteligencia Artificial.
- 5. Procedimientos y técnicas para reducir la energía requerida por las TIC.
- 6. Conclusiones.

A. Prieto

2

Contribución de las TIC al consumo energético

Alberto Prieto
Departamento de Ingeniería de Computadores,
Automática y Robótica.
Universidad de Granada

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

3

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- La sostenibilidad en las TIC se enmarca dentro de uno de los mayores retos de la sociedad actual, consistente fundamentalmente en **reducir el consumo energético.**
- En general, la sociedad desconoce que las TIC, y en particular la IA, constituyen un ámbito relevante en el consumo de energía eléctrica, teniendo un gran impacto en las emisiones de gases de efecto invernadero.
- Todos debemos 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

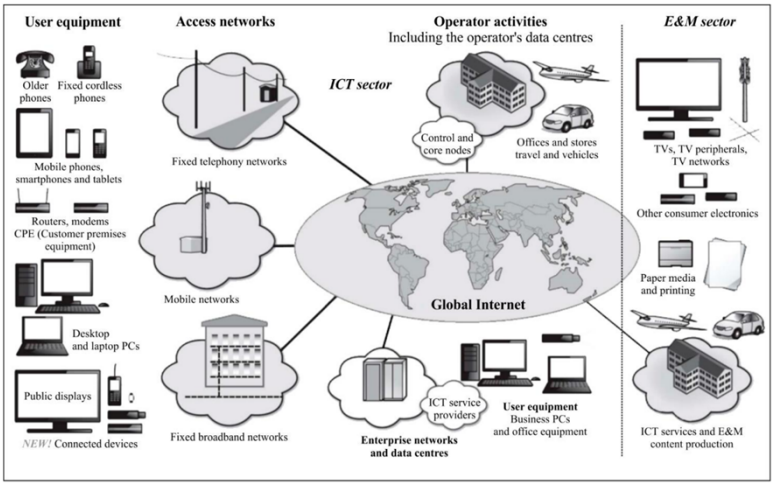
A. Prieto

4

Previsiones del consumo de energía eléctrica (2015)

- 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**, no siendo suficiente, por lo tanto, para alimentar todos los computadores del mundo.

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

Cómo afectan las TIC al medio ambiente

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

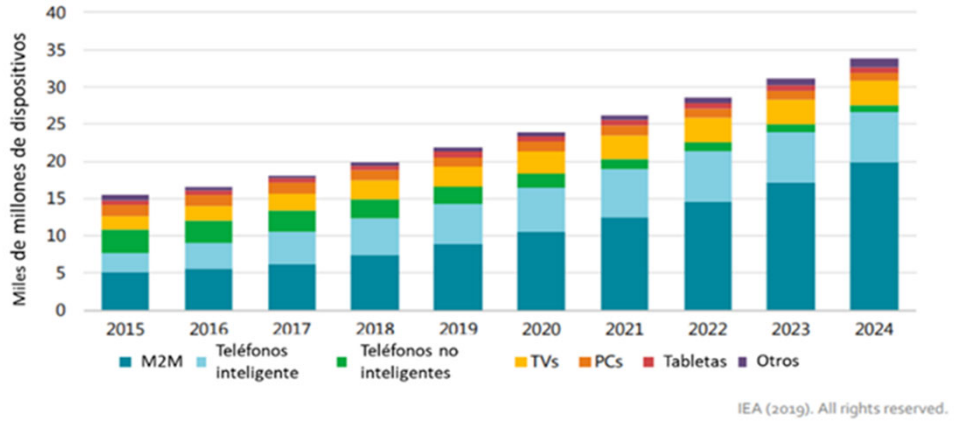
A. Prieto

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

7

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.



A. Prieto

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

8

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.

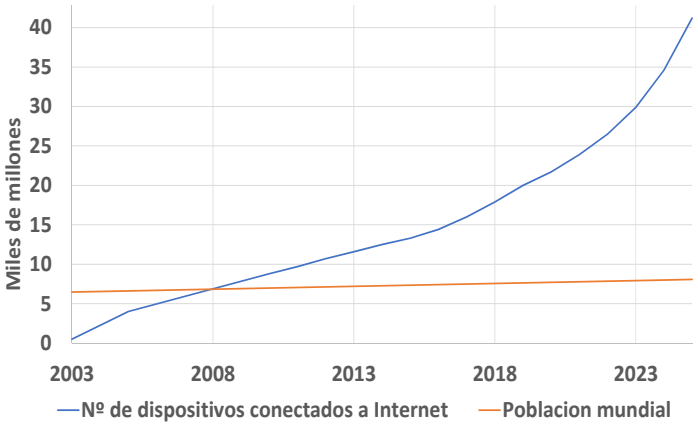
A. Prieto

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

9

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)



A. Prieto

10

Efecto indirecto

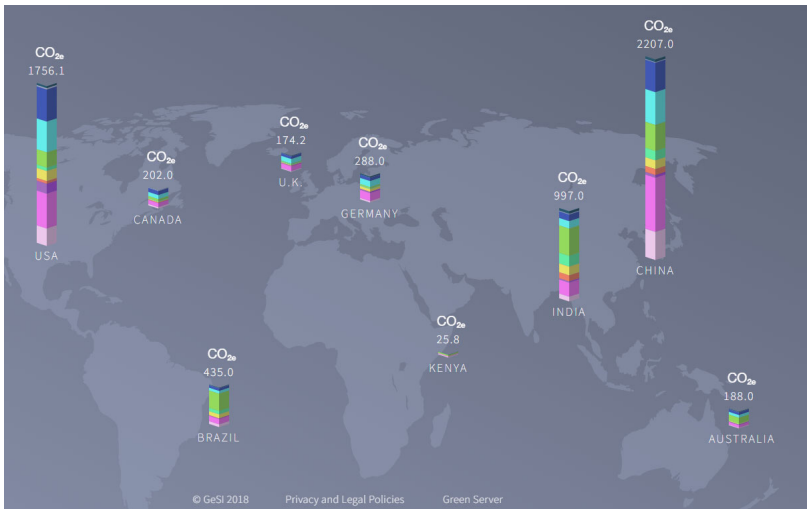
- Está provocada por aplicaciones TIC que facilitan la mejora de la eficiencia y la **reducción del consumo primario de energía** en sectores muy diversos como: construcción, industria, transporte y comercio, aportando soluciones inteligentes.
- **Es bueno para el medio ambiente**.
- En otras palabras, 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.

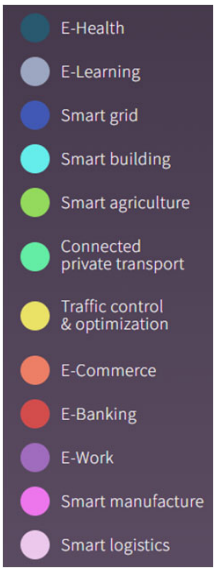
Según datos de la iniciativa GeSI los dominios TIC que están consiguiendo mayores volúmenes de reducción de CO2 son:

- E-salud
- E-enseñanza (clases virtuales, etc.)
- Redes eléctricas inteligentes
- Edificios inteligentes
- Agricultura inteligente
- Transporte privado interconectado
- Control y optimización de tráfico
- E-comercio (transacciones comerciales)
- E-bancario
- E-trabajo (teletrabajo)
- Fabricación inteligente
- Logística Inteligente

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

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





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

Como conclusión podemos decir que:

- Las TIC, globalmente por un lado son perjudiciales para el medio ambiente, pero por otro están contribuyendo a la mejora en distintos sectores.
- Su gran uso (nº de dispositivos y diversidad de aplicaciones) nos reporta grandes beneficios mejorando nuestra productividad y calidad de vida; pero como contrapartida tiene un **efecto nocivo para el medio ambiente**.
- Debe analizarse rigurosamente la evolución del consumo energético, y tomarse las medidas oportunas para contribuir a su sostenibilidad o reducción.

15

V2. Evolución del consumo energético de las TIC

Alberto Prieto
Departamento de Ingeniería de Computadores, Automática y Robótica.
Universidad de Granada

16

Repaso

- Energía (E): Julios
- Energía = Potencia·tiempo; E=P·t
- Energía → [julios] o [Watios·segundo]
- 1 Julio= 1W·s; 1 W·h = 3.600 julios
- Eficiencia energética (EE) en computadores (nº computaciones por julio de energía)
- FLOP: Operaciones con nº reales ("punto o coma flotante")

10^n	Prefijo	Símbolo	Equivalencia decimal
10^{18}	exa	E	1 000 000 000 000 000 000
10^{15}	peta	P	1 000 000 000 000 000
10^{12}	tera	T	1 000 000 000 000
10^{9}	giga	G	1 000 000 000
10^{6}	mega	M	1 000 000
10^{3}	kilo	k	1 000
10^{2}	hecto	h	100
10^{1}	deca	da	10
10^{0}	-	-	1

$$EE = \frac{Instrucciones}{Energía} = \frac{Instrucciones/s}{Energía/s} = \frac{Instrucciones/s}{Vatio} = \frac{FLOP/s}{W}$$

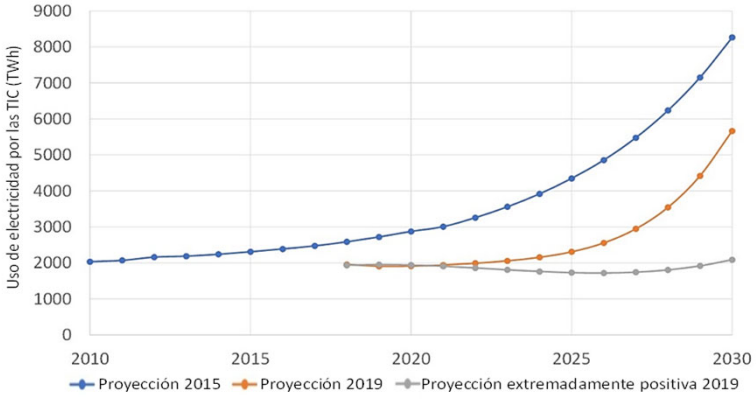
A. Prieto 17

Evolución y previsiones de consumo

- En general, las previsiones sobre el incremento del consumo energético de las TIC 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.

A. Prieto 18

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



Tera → 10E12 → billones

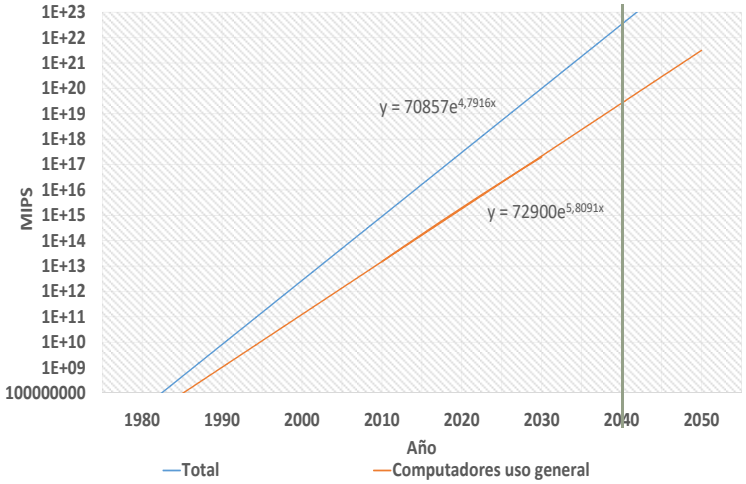
- A.S. Andrae y T. Edler. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157. https://doi.org/10.3390/challe6010117
- A.S. Andrae (2019). Comparison of several simplistic high-level approaches for estimating the global energy and electricity use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

Para medir el consumo energético, por lo general se mide o estima:

- El nº de operaciones que se hacen en un tiempo determinado y el nº de bits implicados en cada una de ellas
- El consumo energético medio en computación por bit en ese tiempo.

- El tráfico de información (nº de bits) que se producen en un tiempo determinado.
- El consumo energético medio en la transmisión de un bit

Volumen de computaciones (Hilbert 2011)



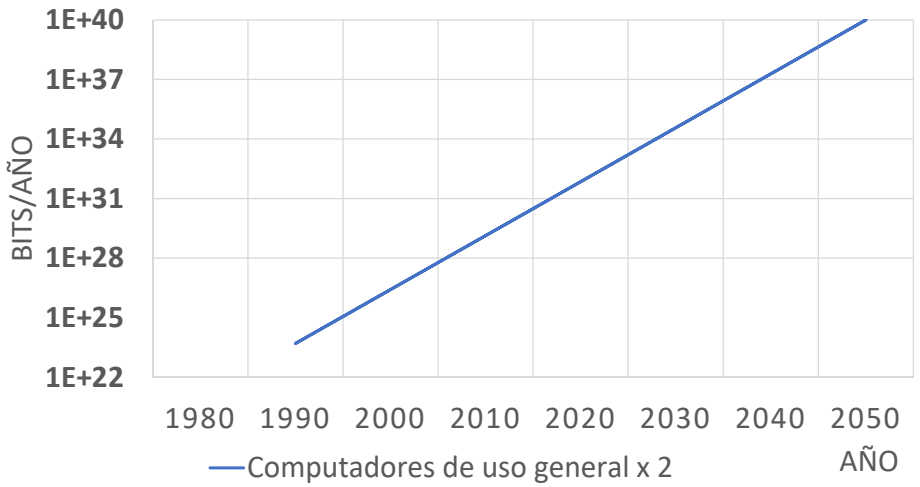
- En el año 2040 se estima llegar a una potencia de procesamiento de $\approx 5x10^{22}$ MIPS, una gran parte de los cuales ($\approx 5x10^{19}$ MIPS) corresponderán a procesamiento de uso general.

A. Prieto

- M. Hilbert y P. Lopez, "The world's technological capacity to store, communicate, and compute information," Science 332 (2011) 60-65

21

De lo anterior se deduce la evolución anual del número total de transiciones binarias implicadas en computación

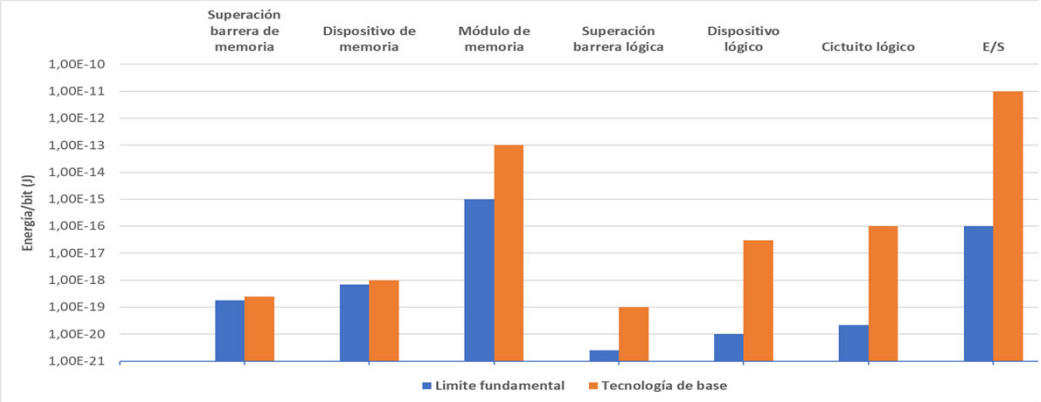


A. Prieto

22

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)



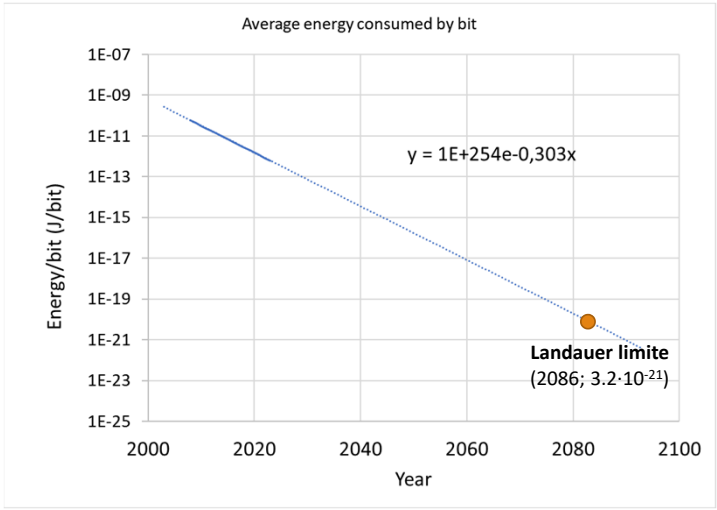
- V. Zhirnov, R. Cavin y L. Gammaitoni. (2014). Minimum energy of computing, fundamental considerations. In ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology. IntechOpen
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: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

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

- 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.

Evolución de la energía consumida por bit, con nuestros resultados, obtenidos a partir de la eficiencia energética:



≈ 10^{-13} J/bit

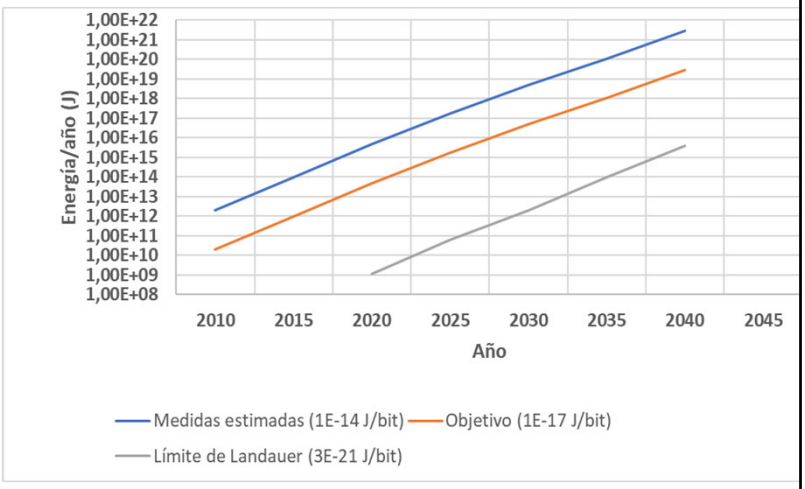
A. Prieto

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

25

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

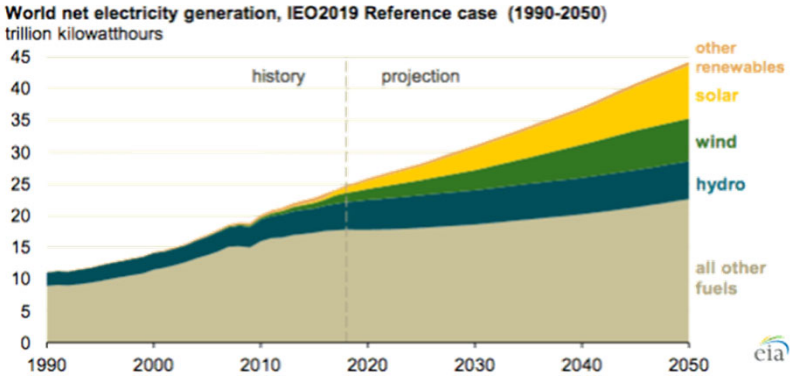


A. Prieto

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

26

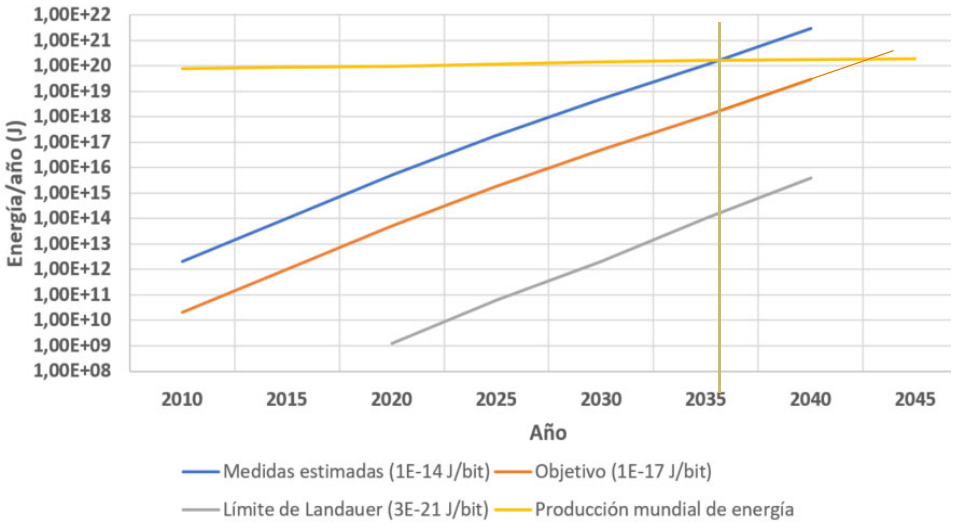
La evolución del incremento de generación de electricidad es lineal



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

A. Prieto https://www.powermag.com/eia-renewables-will-account-for-half-of-global-power-generation-by-2050/ 27

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



A. Prieto 28

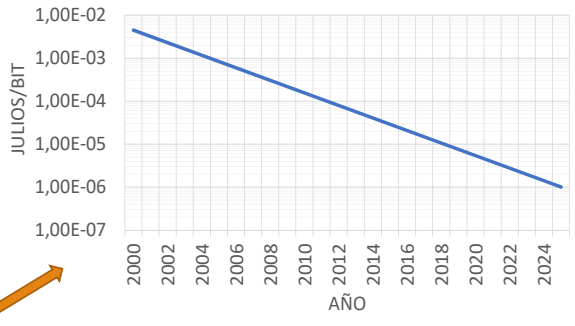
V3. Consumo en tráfico de información digital

Alberto Prieto
Departamento de Ingeniería de Computadores, Automática y Robótica.
Universidad de Granada

29

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).



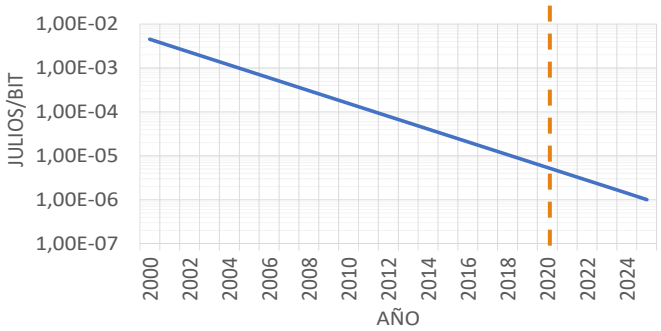
• J. Aslan, K. Mayers, J. G. Koomey y C. France. (2018). Electricity intensity of internet data transmission: Untangling the estimates. Journal of industrial ecology, 22(4), 785-798.

A. Prieto

30

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:
  - $\approx 2{,}8 \cdot 10^{11}$ 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)



A. Prieto

31

Capacidad de datos transmitidos: Los datos nunca duermen, de forma que en 2021 en cada minuto

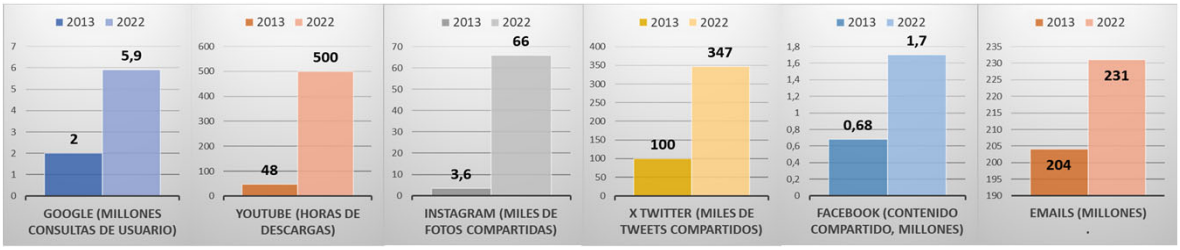
NETFLIX	• 28.000 subscriptores viendo contenido
INSTAGRAM	• 695.000 fotos o vídeos compartidos
LinkedIn	• 9.132 conexiones hechas
WhatsApp y Messanger	• 69 millones de mensajes enviados
Tik Tok	• 5.000 descargas
Twitch	• 2 millones de visualizaciones
E-comercio	• 1,6 millones de € compras en línea
Tinder/X	• 2 millones de tuits
Email	• 197,6 millones de emails enviados
YouTube	• 500 horas de vídeos subidos

- Gráfica original, con datos tomados de Lori Lewis via ALLAcess, STATISTA.

A. Prieto

32

En cada minuto las principals platafromas presentan la siguiente evolución del tráfico

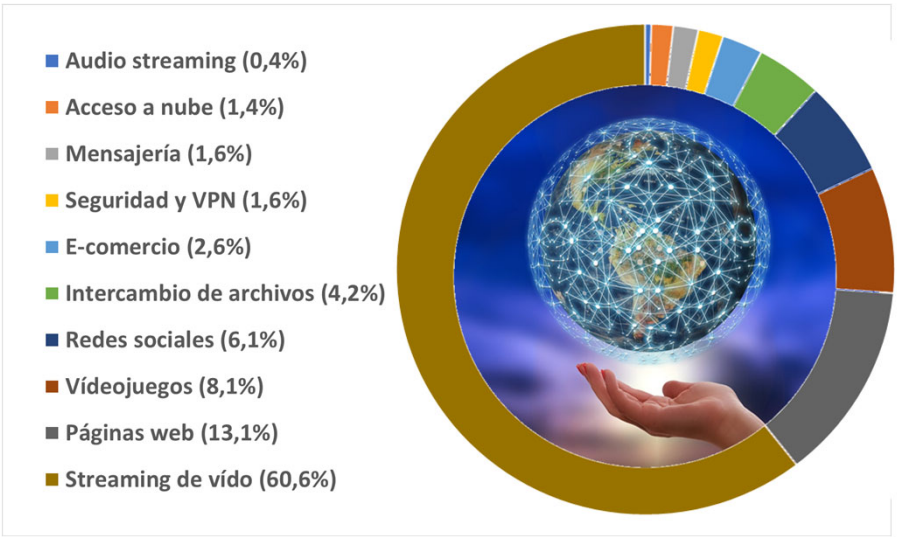


A. Prieto

Diagrama original realizado com datos obtenidos de: https://www.domo.com/data-never-sleeps

33

Orígenes principales del tráfico en Internet a nivel mundial (2019)

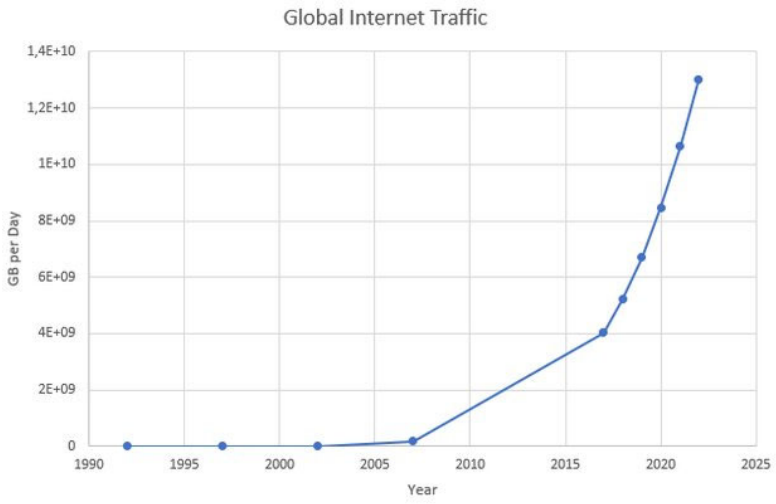


A. Prieto

Diagrama original realizado com datos obtenidos de Statista - Sandvine

34

Evolución del trafico por Internet (GB/día vs año)

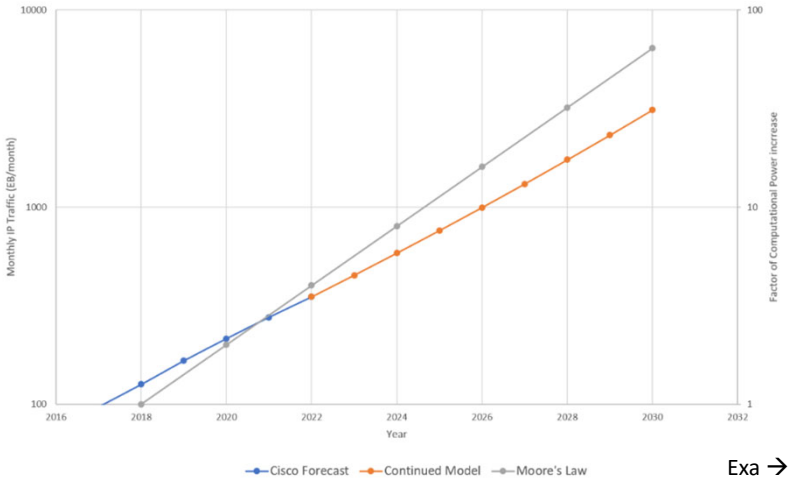


A. Prieto

Cisco, VNI (2018). "Cisco visual networking index: Forecast and trends, 2017–2022". In: White Paper 1

35

Evolución de la movilidad de tráfico IP (EB/mes) e incremento del rendimiento computacional (en gris). 2018



Exa → 10E18 → trillones

A. Prieto

FIGURE 4.7: Forecasted and Modelled IP Traffic vs. Computational Power Density (data from [Cisco, 2018], and based on [Moore, 1965])

36

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)

A. Prieto 37

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

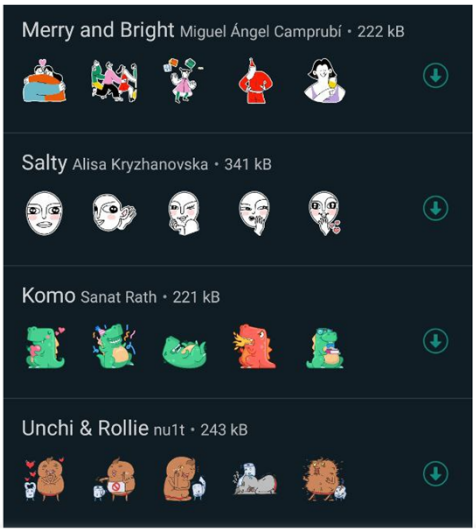
¡Casi media Tonelada de CO2 al año!

A. Prieto 38

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).



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

V4. Consumo energético en aplicaciones de Inteligencia Artificial

Alberto Prieto
Departamento de Ingeniería de Computadores, Automática y Robótica.
Universidad de Granada

41

El uso de la Inteligencia Artificial implica un gran consumo energético…

- Varía según varios factores:
  - tipo de algoritmos utilizados,
  - escala de los modelos,
  - cantidad de datos procesados, y
  - infraestructura de computación empleada.
- Los modelos de IA más grandes, como lo es el de lenguaje GPT *(Generative Pre-trained Transformer)* de OpenAI, consumen una cantidad considerable de energía durante el entrenamiento.
- Las infraestructuras de computación utilizadas para entrenar modelos de IA, cada vez con más frecuencia incluyen servidores de alto rendimiento, unidades de procesamiento gráfico (GPU) y unidades de procesamiento tensorial (TPU) diseñadas específicamente para cargas de trabajo de aprendizaje automático.
- **En 2021**, el consumo total de Google fue de 18,3 TWh, y la IA de la compañía se llevó aproximadamente 2,3 de ellos (el 13%). Esto supone **el equivalente a una población de unos 6 millones de habitantes**, como es la Comunidad de Madrid, durante un año.

Algunos dominios donde se utiliza aprendizaje profundo

(y en los que se puede conocer fácilmente si la respuesta es correcta o no)

- **Clasificación de imágenes**, ImageNet (CIFAR-10, CIFAR-100)
- **Detección de objetos**, MC COCO
- **Comprensión lectora**. Question answering, SQuAD 1.1 (sobre artículos Wikipedia)
- **Reconocimieto de entidades** (nombres de personas, organismos, lugares,, paises, ciudades, nº de telefonos, etc.). Named Entity Recognition CoNLL 2003.
- **Traducción automática**, WMT 2014 (EN-FR y EN-DE),
- **Reconocimiento del habla** (ASR SWB Hub500),
- **Detección de caras** (WIDER Face Hard),
- **Generación de imágenes** (CIFAR10), y
- **Estimación de postura humana** (MPII Human Pose)

A. Prieto

Thompson, N., Greenewald, K., Lee, K., & Manso, G. F. (2023, June). The Computational Limits of Deep Learning. In Ninth Computing within Limits 2023. LIMITS.

43

Polución CO2 generadas por la computación de un programa

- $PCO2 = \#Computaciones \cdot \frac{energía}{Computación} \cdot \frac{CO2}{energía}$

- **PCO2:** Polución de CO2 (Toneladas o Kilogramos)
- **NC**: Nº de computaciones (FLOP).
- **EE**: eficiencia energética (#computaciones por segundo realizables por vatio)
- **E/C**: energía consumida por cada computación (inversa de la eficiencia energética).
- **FFE**: Factor de emisión de energía eléctrica (MIX Eléctrico) Kg CO2/KWh

- $PCO2 = NC \cdot \frac{E}{C} \cdot FEE = NC \cdot \frac{P \cdot t}{Cs \cdot t} \cdot FEE = NC \cdot \frac{1}{EE} \cdot FEE = \left[n^{\underline{o}} \cdot \frac{W}{n^{\underline{o}}/s} \cdot \frac{Kg}{W \cdot s}\right] = [Kg]$

A. Prieto

44

Parámetros utilizados en nuestros cálculos:

- $PCO2 = NC \cdot \frac{1}{EE} \cdot FEE = \left[nº \cdot \frac{W}{nº/s} \cdot \frac{Kg}{W \cdot s}\right] = [Kg]$

- NC → Trabajo de Thompson et als.
- EE= 65,4 GFLOPS/W;
- FEE=0,351 Kg CO2/Kwh
- Rmax= 1.194E15 FLOP/s;

- Emisión anual hogar español: 0,56 T CO2

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

A. Prieto 45

Consumo computacional y energético de aprendizaje profundo en algunas aplicaciones de referencia (proyección polinómica).

Aplicación de referencia	Tasa de error		Nº de operaciones (nº reales, flops)	Emisiones CO_2 (Tm)	Consumo anual de nº de hogares en España	Tiempo de ejecución aa:mm:dd:hh
Clasificación de imágenes, ImageNet	2022	9.00%	1,00E+23	1,49E+02	266	00:00:01:23
	Objetivo 1	5%	1,00E+26	1,49E+05	266.235	03:08:09:08
	Objetivo 2	1%	1,00E+33	1,49E+12	2,66E+12	2,69E+07 años
Detección de objetos, MS COCO	2022	38.7%	1,00E+22	1,49E+01	27	00:00:00:02
	Objetivo 1	30%	1,00E+23	1,49E+02	266	00:00:01:23
	Objetivo 2	10%	1,00E+31	1,49E+10	2,66E+10	2,69E+05 años
Comprensión lectora (SQuAD 1.1)	2022	9.4%	1,00E+22	1,49E+01	27	00:00:00:02
	Objetivo 1	2%	1,00E+29	1,49E+08	266.235.156	2,69E+03 años
	Objetivo 2	1%	1,00E+32	1,49E+11	2,66E+11	2,69E+06 años
Reconocimiento de nombres de entidades (CoNLL 2003)	2022	5.4%	1,00E+23	1,49E+02	266	00:00:01:23
	Objetivo 1	2%	1,00E+39	1,49E+18	2,66E+18	2,69E+13 años
	Objetivo 2	1%	1,00E+50	1,49E+29	2,66E+29	2,69E+24 años

Nº de habitantes de la Tierra: 8E+09

A. Prieto

Datos sobre tasa de error y nº de operaciones obtenidos de Thompson-2023

46

Aprendizaje en Chat GPT-3 175B

- Contenido del GPT-3 175:
  - **175 mil millones de parámetros** (pesos y otros valores de las redes neuronales).
  - 800 GB de memoria
  - Procesamiento de consultas de textos de hasta 2048 tokens (palabras o subpalabras). Cada conjunto se considera como "ventana contextual". Eso significa que la arquitectura del sistema tiene 2.048 pistas o cauces a lo largo de las cuales se procesan los tokens sucesivos en paralelo.
- 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.)

A. Prieto 47

Pre-entrenamiento de GPT-3 750B

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

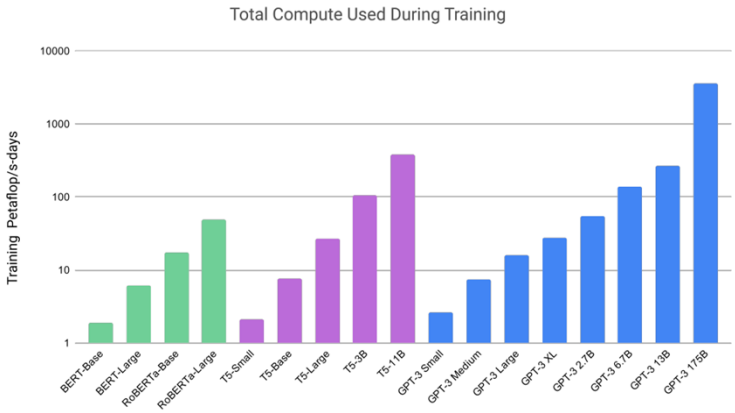


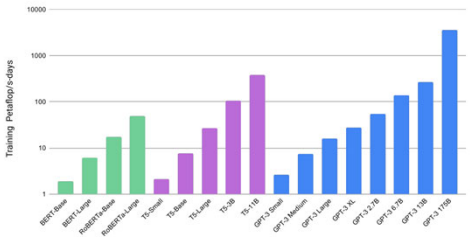
Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 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

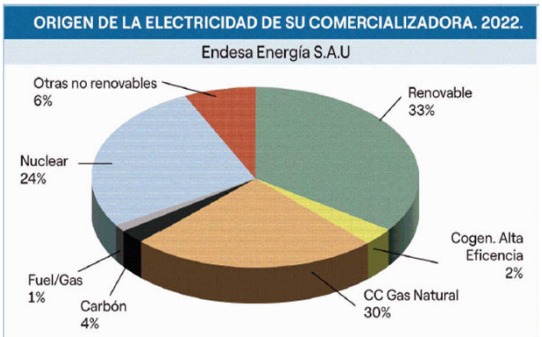
A. Prieto 48

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
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



A. Prieto 49



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%

FFE: Factor de emisión de energía eléctrica (MIX Eléctrico)

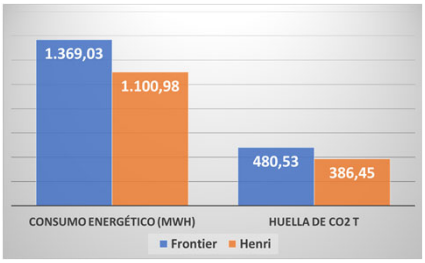
Emisiones de CO_2 equivalente
Endesa Energía S.A.U

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

Consumo anual medio de un hogar en España: 3.487 KWh → 0,56 T CO2

A. Prieto 50

	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
Huella equivalente a hogares españoles (nº):	851	684
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)

A. Prieto

51

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

0.4 KW·h x 0.351 = 0.14 Kg CO2

- Consumo muy pequeño, pero hay millones de consultas:
  - 2,8 millones de accesos (inferencias) consumen lo de un entrenamiento.

A. Prieto

- Brown, T., et als. (2020)

52

Análisis de resultados anteriores:

- Se estima que el **entrenamiento del Chat GPT3–175B**, suponiendo que se realiza en el supercomputador más eficiente del mundo (Henri, 2023), implica una generación de 386,4 T de CO2 lo que equivale al consumo anual de 684 hogares españoles.
- Hemos estimado que el **entrenamiento del modelo de clasificación de imágenes** (ImageNet) para obtener una **precisión del 5%** supondría una generación de unas 149.000 T de CO2, lo que es superior a la generación anual de los **domicilios de los habitantes de Granada** (230.000 vs 260.000)
- Al aumentar los requisitos de **precisión** aumenta muy considerablemente (polinómicamente) la **potencia informática** requerida y las **necesidades energéticas**. Sin el aumento de la eficiencia energética de los equipos informáticos sería imposible la mejora de la precisión de resultados que se está obteniendo.
  - La introducción de GPUs condujo a unas mejoras muy significativas, así ya en el 2009 Raina et als. desarrollaron una implementación del aprendizaje de una red neuronal profunda de cuatro capas con 100 millones de parámetros, paralelizando el proceso para una tarjeta gráfica (GPU) con un total de 240 núcleos (30 multiprocesadores con 8 cauces de procesamiento cada uno). De esta forma consiguieron **reducir el tiempo de ejecución unas 70 veces (de tardar varias semanas a un día) con GPUs frente a un procesador de doble núcleo**.

A. Prieto 53

Conclusión:

- Desde los puntos de vista del tiempos de respuestas y consumo energético, es imposible continuar con el ritmo de progreso utilizando las líneas actuales.
  - El progreso continuo en las aplicaciones de Inteligencia Artificial requiere métodos computacionalmente más eficientes, que tendrán que provenir de cambios en el aprendizaje profundo o de pasar a otros métodos de aprendizaje automático.
  - Obviamente se pone de manifiesto que el desarrollo de la IA ha sido factible gracias a las mejoras de la eficiencia energética que proporciona la electrónica y la Ingeniería de Computadores: Estas mejoras deben continuar con sus excelentes avances para lograr retos inalcanzables en el momento actual.
- A continuación analizaremos procedimientos y técnicas para reducir el consumo energético dentro del ámbito de las TIC.

A. Prieto 54

V5. Procedimientos y técnicas para reducir la energía requerida por las TIC

Alberto Prieto
Departamento de Ingeniería de Computadores, Automática y Robótica.
Universidad de Granada

55

Cómo reducir la generación de CO2 (PCO2↓)

- $PCO2 = \#Computaciones \cdot \frac{energía}{Computación} \cdot \frac{CO2}{energía} = NC \cdot \frac{1}{EE} \cdot FEE$
- NC↓ Reducción la cantidad de computaciones(NC). Mejora de algoritmos y programas
- E/C↑ Aumentar la eficiencia energética del hardware
  - Reducir la cantidad de energía por computación
  - Google estima que de una generación de Unidad de Procesamiento Tensorial (TPU) a otra, obtuvieron un aumento de rendimiento de 2,1 veces y mejoraron la eficiencia energética en 2,7 veces.
- FEE↓ Reducir el Factor de Emisión de Energía Eléctrica (MIX Eléctrico)
  - Centros de datos ubicados en lugares donde se produce energía limpias o renovables.

A. Prieto

56

- 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. Programación eficiente (software).**
  - **B. Mejoras tecnológicas en los componentes electrónicos y dispositivos.**
  - **C. Planificación y gestión del uso de los recursos.**
  - **D. Cambios de escala.**

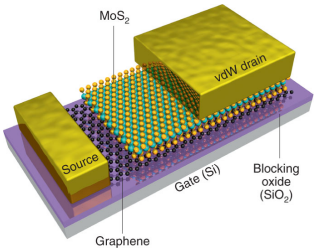
A. Programación eficiente

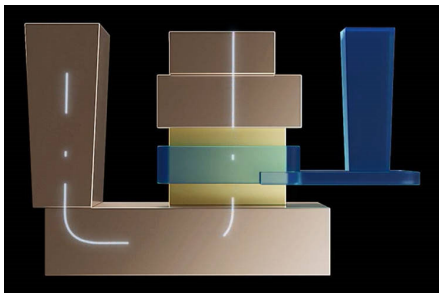
- Caracterizar los módulos del programa de acuerdo con su consumo energético.
- Elegir entre varias alternativas (entre navegadores, por ejemplo), la de menor consumo energético.
- Utilizar estrategias para que los algoritmos sean computacionalmente menos costosos y por tanto su ejecución más rápida → menor consumo de energía.
  - Se trata de "no matar hormigas a cañonazos". Ejemplos:
    - Entrenar las redes neuronales para las tasas de error requeridas y no menor.
    - Podado de pesos, redes cuantizadas, comprensión de bajo rango. Parada temprana.
    - En aprendizaje en máquinas, a veces se pueden obtener resultados plenamente satisfactorios en aplicaciones de ámbito muy próximo y con conjuntos de datos similares, utilizando información de modelos previamente entrenados (meta-aprendizaje y aprendizaje por transferencia).
    - En las fases de aprendizaje no siempre es requerido que los datos sean de doble precisión (64 bits), pudiéndose utilizar en las primeras fases precisiones muy inferiores, incluso de 8 bits. De esta forma se reduce el coste computacional obteniéndose los mismos resultados que si se utiliza permanentemente doble precisión (Sensibilidad de los pesos en los resultados)
  - **¡CUANTO MENOS INSTRUCCIONES TENGAN QUE EJECUTARSE MEJOR!**

Bernier, J. L., Ortega, J., Ros, E., Rojas, I., & Prieto, A. (2000). A quantitative study of fault tolerance, noise immunity, and generalization ability of MLPs. Neural Computation, 12(12), 2941-2964.

B. 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**).



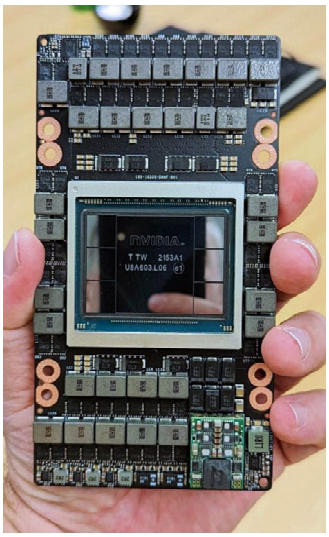


IBM's VTFET with a vertical channel (yellow) and gate-all-around (blue). Contacts are brown and the white line shows current flow.

- 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 dec. 2021) IBM beats finFETs with vertical CMOS at IEDM. Electronics Weekly.com

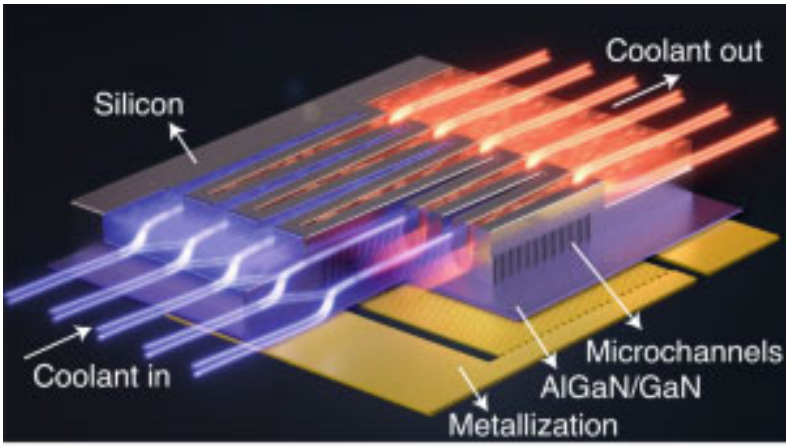
… mejoras tecnológicas en los componentes electrónicos y dispositivos …

- 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.



… 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.



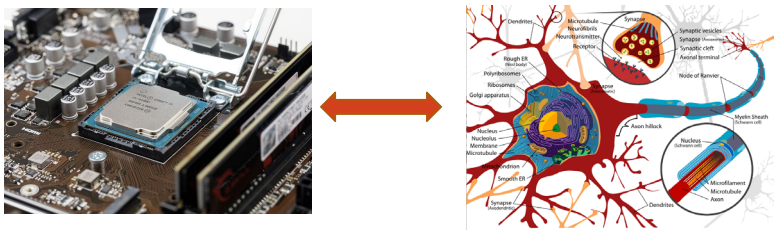
A. Prieto

Varnava, C. Chips cool off with integrated microfluidics. Nat Electron 3, 583 (2020).

61

… 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 del mundo, 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, grafeno, etc. el del cerebro es de naturaleza biológica (bioquímica, células, tejidos,...)

A. Prieto

Imágenes de Pixabay y de Wikimedia Commons

62

C. 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.

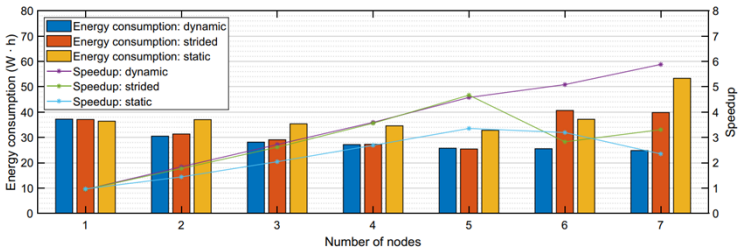
D. Cambios de escala

- 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.
  - Veamos un ejemplo…

- 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.

Clasificación de EEG con 3600 características, utilizando mRMR-KNN

- Los resultados dependen de la distribución de trabajo entre los nodos del clúster. Utilizando paralelismo se reduce el consumo energético.



(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!

- Escobar, J. J., Rodríguez, F., Prieto, B., Kimovski, D., Ortiz, A., & Damas, M. (2023). A distributed and energy-efficient KNN for EEG classification with dynamic money-saving policy in heterogeneous clusters. *Computing*, *105*(11), 2487-2510.
- Escobar, J. J., Rodríguez, F., Kızıltepe, R. S., Prieto, B., Kimovski, D., Ortiz, A., & Damas, M. (2023, June). Energy-Aware KNN for EEG Classification: A Case Study in Heterogeneous Platforms. In International Work-Conference on Artificial Neural Networks (pp. 505-516). Cham: Springer Nature Switzerland.

D. …. cambios de escala.

- **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.
- Se estima que el consumo de energía por los centros de datos contribuye con un porcentaje superior al 30 % de todas las TIC.

Centros de datos sostenibles (energéticamente neutros)

- Los centros de datos son la columna vertebral de Internet y, por lo tanto, del mundo digital que en la actualidad, y poco a poco con más fuerza, gobierna la vida de las personas.
  - En los Centros de Datos se almacenan enormes cantidades de datos que son críticos para el desarrollo de las funciones diarias de los consumidores, de las empresas y de las Administraciones públicas.
- Según las previsiones realizadas por Andrae (2019), en el 2025 el consumo de los Centro de Datos representará el 18 % del consumo total de las TIC, y en el 2030 subirá al 34 %
  - CONSECUENCIA: para reducir el consumo de las TIC hay que considerar el de los Centros de Datos

Políticas de Centros de Datos para mejoras del medio ambiente

- Hacer entrar en los modos de suspensión o de espera a los recursos que en un momento dado no sean necesarios, etc.
- Migrar las aplicaciones a los equipos (servidores) de menor consumo, siempre que cumplan con los tiempos de respuesta requeridos.
- Ubicar los Centros de Datos próximos a lugares donde se generan energías limpias o renovables (Google, Microsoft, Amazon).
  - Las propias empresas están instalando sus generadores de energía:
    - 100% energía renovable propia
    - Incluso venden el exceso de energía producido.
- Ejemplos de centros de Google muy eficientes:

Centro de Google en Oklahoma

- Oklahoma es un estado muy ventoso
- *Eficacia del uso de energia (Power Usage Effectiveness, PUE*) = 1,09
- $PUE = \frac{Energía\ total\ consumida}{Energía\ consumida\ por\ TIC}$

$$= 1 + \frac{Energía\ no\ TIC}{Energía\ TIC}$$

Great Western wind farm in Oklahoma
(225 MW for Google)

A. Prieto

69

Centros Google en Bélgica

Norther Offshore wind farm in Belgium (92 MW for Google)

Google's data center in St. Ghislain, Belgium.

A. Prieto

Centro de datos de Google en Dinamarca

Rødby solar farm in Denmark (55 MW for Google)

A. Prieto 71

Centro de Datos de Google en Hamina (Finlandia)

- Instalado en una antigua fabrica de papel.

A. Prieto Image: YLE / Jani Aarnio 72

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.
- PUE (abril 2024): 1,09

Sistema de refrigeración del Centro de Datos de Changhua (Taiwan)

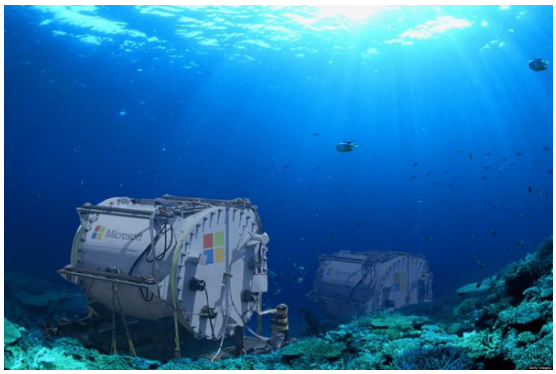
- Photographer: Ashley Pon/Bloomberg

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/

Proyecto Natick de Microsoft



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.
- El "Centro de Datos" 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.**

Conclusiones

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

79

Como conclusiones generales de este vídeo y los anteriores sobre sostenibilidad de las TIC, podemos decir:

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Se estima que en el año 2030 las TIC gastarán aproximadamente el **13% de la electricidad mundial**, y para 2050 el consumo de los centros de datos será unas tres veces mayor que la cantidad total de energía generada en Japón.
- Hay establecidos hitos que hoy día son inalcanzables debido a las extraordinarias necesidades de computo y de energía requeridas.
- El **consumo de energía** debe considerarse por científicos, ingenieros y fabricantes como una medida de prestaciones tan importante como el **rendimiento computacional.**
- Por un lado, hay razones medioambientales y económicas, pero también la necesidad de mejorar la **autonomía de los dispositivos que utilizan baterías**.
- La reducción del consumo energético en el ámbito de las TIC es una cuestión trascendental, y debe ser afrontada desde muy distintos ámbitos (ingeniería de computadores, ingeniería del software, enseñanza, etc.) y trabajar hacia un equilibrio entre innovación y sostenibilidad.
- La sociedad debe estar informada de que el uso de las TIC (sea cual sea su forma) lleva implícito un consumo energético y hacerlo de forma razonable. Hay sencillos procedimientos y técnicas aplicables para reducir el consumo, a las que deberían acostumbrarse todos los usuarios.
- **¡Todos debemos contribuir, desde nuestros respectivos ámbitos, al reto de lograr la sostenibilidad de nuestro planeta!**

A. Prieto

80

Agradecimientos

- Deseo agradecer a las siguientes personas su colaboración en las investigaciones que estamos realizando sobre este tema:
  - Beatriz Prieto Campos (UGR)
  - Juan José Escobar (UGR)
  - Miguel Damas (UGR)
  - Antonio Díaz (UGR)
  - Christian Morillas (UGR)
  - Jesús González Peñalver (UGR)
  - Andrés Ortiz (UMA)
  - Francisco Gil (UAL)
  - Francisco Illeras (UGR)

Nuestras investigaciones en este ámbito actualmente se están financiando parcialmente por:

Proyecto: **PID2022-137461NB-C31** financiado por MCIN/AEI/10.13039/501100011033/ y por FEDER Una manera de hacer Europa

MINISTERIO DE CIENCIA E INNOVACIÓN

UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

AGENCIA ESTATAL DE INVESTIGACIÓN

¡Muchas gracias por tu atención!

- Alberto Prieto Espinosa.
- Conferencias:

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