Supercomputación, inteligencia artificial, chips y autonomía europea

// Mateo Valero

LECCIONESCAJAL

LECCIONESCAJAL // 7

Supercomputación, inteligencia artificial, chips y autonomía europea

// **Mateo Valero**

Vicerrectorado de Cultura y Patrimonio
Universidad de Zaragoza

LECCIONESCAJAL // 7

5 de mayo de 2026

La **Lección Cajal** es una conferencia anual dictada desde 2019 en la Universidad de Zaragoza por una figura académica relevante en su campo del saber, impulsada por el Vicerrectorado de Cultura y Patrimonio para conmemorar el 150 aniversario de la entrada de Santiago Ramón y Cajal en esta universidad, su «venerada *alma mater*».

Edita: Vicerrectorado de Cultura y Patrimonio
Prensas de la Universidad de Zaragoza

Diseño: Fernando Lasheras / M. Á. Pérez Arteaga

Compuesto con la tipografía «Carmen» de Andreu Balius

Imprime: Servicio de Publicaciones. Universidad de Zaragoza

ISBN 979-13-7014-121-9

MATEO VALERO[1] es profesor de Arquitectura de Computadores en la Universidad Politécnica de Cataluña (UPC) y es el director fundador del Barcelona Supercomputing Center (BSC). Su investigación se centra en las arquitecturas de los supercomputadores, o computadores de alto rendimiento (HPC, High Performance Computing). Ha publicado más de 700 artículos, ha participado en la organización de más de 300 congresos internacionales y ha impartido más de 800 conferencias como invitado. El Prof. Valero ha recibido numerosos premios, incluyendo los tres más importantes en su área de investigación: el Eckert-Mauchly Award en 2007, del Institute of Electrical and Electronics Engineers (IEEE) y la Association for Computing Machinery (ACM), en Arquitectura de Computadores; el Seymour Cray Award en 2015, de la IEEE en Supercomputadores, y el Premio Charles Babbage Award en 2017 del IEEE, en Computadores Paralelos. Entre otros más, el Harry Goode Award 2009 del IEEE, y el Distinguish Service Award de la ACM. El Prof. Valero es un miembro del «Hall of Fame» del ICT European Program, y, en noviembre de 2008, fue seleccionado como uno de los 25 investigadores más influyentes en Tecnologías

1 <http://www.bsc.es/cv-mateo/>.

de la Información (IT) durante el periodo 1983-2008. En 2020 recibió el HPC Wire Reader's Choice Award «por su liderazgo excepcional a nivel mundial en HPC» y por «ser un pionero en HPC desde 1990 y la fuerza conductora tras el renacimiento de la independencia europea en HPC». En 2023, SCALAC (the Advanced Computing System for Latin America and the Caribbean) le otorgó una distinción «in recognition of his Outstanding Collaboration in Advanced Computing Between Latin America and Europe» e institucionalizó el «Mateo Valero Prize for the Outstanding Collaboration with Latin America and the Caribbean Partners in HPC», que se otorga cada año. Ha recibido dos de los diez premios nacionales de investigación de España: el Julio Rey Pastor en Informática y Matemáticas en 2001, y el Leonardo Torres Quevedo en Ingeniería, en 2007. Le fue concedido el Premio Rey Jaime I en investigación básica en 1997. Recibió el Premio Aragón en 2008 y la Creu de Sant Jordi en 2016, que son los reconocimientos más importantes otorgados por los Gobiernos de Aragón y Cataluña; el Premio Ciudad de Barcelona en 1994 y el Premio Narcís Monturiol, concedido por la Generalitat de Catalunya, en 1994; el Premio de Investigación de la Fundació catalana per a la Recerca i la Innovació, máximo reconocimiento a la investigación en Cataluña, otorgado por el Gobierno de la Generalitat, en 2006. Ha sido honrado con la Condecoración de la Orden Mexicana del Águila Azteca, en 2018, el reconocimiento más importante del Gobierno mexicano a una persona no mexicana, también ha sido reconocido en la primera edición del Barceloní de l'Any por *El Periódico de Catalunya* y el Premio Vanguardia, en la categoría Innovación y Ciencia, ambos en septiembre de 2024. Fue seleccionado como uno de los HPCwire's 35 Legends – Class of 2025. El Prof. Valero tiene doctorados *honoris causa* por las catorce universidades siguientes: Universidad de Chalmers (Göteborg, Suecia, 2008), Universidad de Belgrado (Serbia, 2008), Universidad de Las Palmas de Gran Canarias

(2009), Universidad de Veracruz (México, 2010), Universidad de Zaragoza (2011), Universidad Complutense de Madrid (2013), Universidad de Cantabria (2015), Universidad de Granada (2015), el Cinvestav de México (2017), Universidad Cristóbal Colón de Veracruz (México, 2022), la Universidad de Murcia (2024), la Universidad de Santiago de Chile (2024) y la Universidad Illes Balears (2024), y es también profesor emérito por la Universidad Autónoma de Paraguay (2025). Es miembro de las diez academias siguientes: académico fundador de la Real Academia de Ingeniería de España (1994), académico correspondiente de la Real Academia de Ciencias Exactas, Físicas y Naturales de España (2005), académico de la Real Academia de Ciencias y Artes de Barcelona (2006), académico de la Academia Europea «Academy of Europe» (2010), académico correspondiente de la Academia de Ciencias de México (2012), académico de la Academia de Gastronomía de Murcia (2018), académico de honor de la Real Academia Europea de Doctores (2018), académico correspondiente de la Academia de Ingeniería de México (2018), académico de honor de la Real Academia de Medicina de Zaragoza (2021) y académico correspondiente de la Academia de Ciencias de Cuba (2022). Es «Fellow» del Institute of Electrical and Electronics Engineers (IEEE), de la Association for Computing Machinery (ACM), de la Asia-Pacific Artificial Intelligence Association (AAIA) y de la International Artificial Intelligence Industry Alliance (IAIIA). En 1998, fue elegido hijo predilecto de su pueblo, y en el año 2006, la Asociación de Madres y Padres de Alumnos de Alfamén, decidió poner su nombre al colegio público donde el profesor Valero había estudiado.

ES UN GRAN HONOR PARA MÍ haber sido invitado a impartir la Lección Cajal del año 2026, organizada por la Universidad de Zaragoza. Como aragonés y como científico, ¿qué más podía soñar? A la hora de elegir el tema tenía dos opciones extremas. Por una parte, impartir una charla sobre el diseño de arquitecturas de computadores de altas prestaciones para supercomputadores, que ha sido mi tema de investigación durante toda mi vida académica, y por otra, impartir una conferencia sobre divulgación acerca de las actividades que realizamos en el Barcelona Supercomputing Center. Me decidí por la segunda opción.

Así pues, hablaré sobre supercomputación y su relación con la inteligencia artificial (IA). Describiré el Barcelona Supercomputing Center y un resumen de las actividades que llevamos a cabo. Hablaré de los Large Language Models (LLM) y su influencia en el diseño de las arquitecturas de computadores orientadas a ejecutar estos modelos de la manera más eficiente. Hablaremos de los centros de datos, así como de las necesidades de agua y electricidad que requieren. También de las *foundries,* como fábricas de chips. Describiremos la importancia estratégica a nivel mundial del diseño y de la fabricación de chips de altas prestaciones. Y veremos la situación en la que Europa se

encuentra. Deseo que el lector/oyente no experto en estos temas pueda disfrutar y obtener informaciones muy actuales sobre la supercomputación y la inteligencia artificial; ese es mi objetivo fundamental.

1. La investigación es muy importante para el avance de un país. ¿Cómo han evolucionado el método para investigar y los instrumentos utilizados?

Estoy seguro de que compartimos la idea de que la investigación es el motor esencial del desarrollo de un país, y más en particular, de aquellos países que no disponen de grandes recursos naturales. La capacidad de generar conocimiento, transformar ideas en soluciones y aprovechar el talento de las personas es, sin duda, la vía más sólida para resolver los retos de la sociedad y construir una riqueza sostenible basada en la inteligencia humana y la innovación. En los países más ricos y avanzados del mundo, se está haciendo una investigación de muchísima calidad y se aplican sus resultados. Pero el camino hasta llegar a este punto de desarrollo ha sido duro y aún queda mucho para poder llevar a la ciencia a sus límites, si esto fuese posible. Sobre la pregunta de dónde están los límites de la ciencia podríamos hablar muchísimo, sobre todo si creemos que la ciencia puede resolver, de manera razonable, cualquier problema.

Podríamos decir que la humanidad empezó a desarrollar el método científico con los primeros filósofos griegos, como Aristóteles y Platón, y que se basaron en ideas y resultados antes obtenidos en China, Egipto, India y Mesopotamia. Empezaron a plantearse el porqué de las cosas, aunque en aquel momento pudieron dar pocas respuestas útiles para explicar los fenómenos naturales y cosmológicos. Aprender a medir mejor las obser-

vaciones fue el primer paso. El primer interés de la humanidad fue intentar explicar aspectos de la astronomía y sin apenas instrumentos, salvo reglas y compás, crearon diferentes sistemas para medir el tiempo o para posicionar astros y estrellas. Esto me parece ahora increíble que sucediera.

Para avanzar más, se necesitaba la ayuda de las matemáticas, de la física, de la química y de instrumentos asociados. Como ejemplos concretos, Galileo Galilei, con la invención del teles-copio, pudo realizar varios descubrimientos en el campo de la astronomía, pero sin la óptica eso no hubiese sido posible. Teníamos fenómenos que observábamos y necesitábamos un método científico para explicarlos. Fue a partir de Francis Bacon (1561-1626), cuando se estableció un método científico para el avance de la ciencia.

En la figura 1.1 he intentado representar a grandes rasgos, aspectos importantes relacionados con la investigación. Y hemos

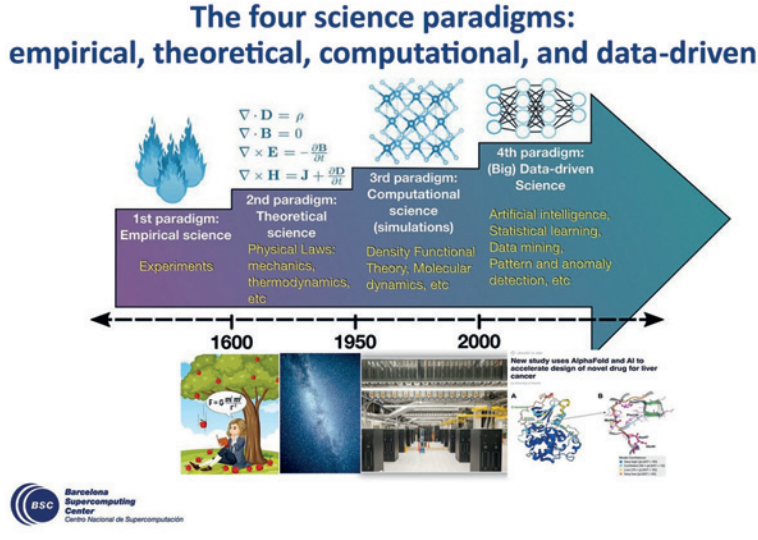


Figura 1.1. Evolución de los métodos usados para investigar.

dividido el tiempo en cuatro grandes periodos. Un ejemplo concreto fue la ley de la gravedad. La tradición cuenta que, observando la caída de las manzanas de los árboles, Isaac Newton (1643-1727) propuso unas fórmulas matemáticas que explicaban el movimiento de la manzana, atraída por la masa de la Tierra. Estableció las leyes de la gravitación universal y también las de la mecánica clásica.

Desde aquel periodo, y mucho antes, se establecieron muchos laboratorios como mecanismo para comprobar las ideas teóricas o como medio para buscar teorías que explicaran los experimentos de los laboratorios. Uno de los primeros laboratorios reconocidos como tal fue el de Pitágoras. Los instrumentos asociados a los laboratorios jugaban un papel fundamental. Y tanto es así que *sir* Humphrey Davy (1778-1829) dijo aquello de que «nada avanza más la ciencia que la creación de un instrumento eficiente». Él lo vivió aplicando la electroquímica al descubrimiento de elementos de la tabla periódica como el calcio, el sodio, el potasio y otros.

En este contexto de progreso tecnológico, resulta inevitable recordar la tradición científica española que lo precede, encarnada en la figura de Santiago Ramón y Cajal.[2] A finales del siglo XIX y comienzos del XX, este aragonés, pionero en sus temas de investigación, desveló la arquitectura microscópica del sistema nervioso humano —compuesto por aproximadamente 86 000 millones de neuronas interconectadas mediante unos 100 billones de sinapsis físicas, según estimaciones modernas—[3]

---

2 Rodolfo R. Llinás (2003), «The contribution of Santiago Ramón y Cajal to functional neuroscience», *Nature Reviews Neuroscience,* vol. 4, pp. 77-80.

3 Congping Lin, Fan Xu y Yiwei Zhang (2003), «Brain-wide dendrites in a near-optimal performance of dynamic range and information transmission», *Sci. Rep.* 13, 7488.

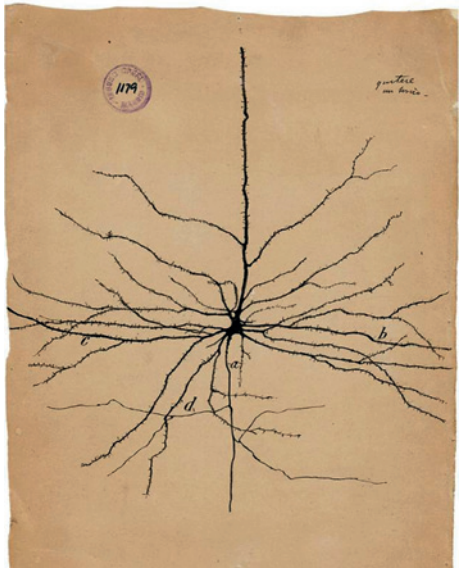


Figura 1.2. En la imagen, la neurona piramidal, a la que Cajal llamó «la noble y enigmática célula del pensamiento».

y formuló la doctrina de la neurona junto con la plasticidad sináptica como base del aprendizaje biológico. Sus intuiciones sobre el refuerzo de conexiones anticiparon las reglas hebbianas —como esta de que las neuronas que se activan juntas se conectan juntas— que subyacen a muchas redes neuronales en la inteligencia artificial.

Sin embargo, cualquier intento de comparar esta complejidad biológica colosal con los modelos masivos de lenguaje (LLM) resulta vano: mientras el cerebro humano integra 86×10^9 neuronas y 10^{14} sinapsis en un sistema dinámico, eficiente y multimodal (con unos 5-12 dendritas primarias por neurona piramidal cortical para la integración masiva de señales), los LLM más avanzados como a finales del año 2025, GPT-4 o Llama 3.1 405B operan con meros 1-2 billones (europeos) de parámetros

(nodos artificiales), equivalentes a pesos sinápticos estáticos, sin la riqueza de la plasticidad biológica ni el procesamiento paralelo optimizado del cerebro; es óptimo en coste energético y de acceso a memoria. El trabajo de Ramón y Cajal no solo fundó la neurociencia moderna y simboliza el esplendor científico español, sino que permanece como antecedente conceptual inalcanzable para los modelos computacionales actuales, subrayando la brecha infranqueable entre la inteligencia situada y sus simulaciones digitales.

Los computadores han sido, tal vez, los instrumentos que más han influido, durante los últimos setenta y cinco años, en la vida de las personas. Y de manera muy clara en la investigación. Nos están permitiendo manejar grandes cantidades de datos y hacer cálculos rápidos sobre ellos. Gracias a los computadores, la ciencia ha dado pasos de gigante en todos los campos del saber humano.

En los últimos años, la captura y generación de muchos datos y la existencia de grandes supercomputadores han hecho resurgir el interés en usar la inteligencia artificial de forma que sus aplicaciones se han posicionado como herramientas que, junto con los datos y los computadores de altas prestaciones, están produciendo un movimiento cámbrico en la ciencia; cada día nos sorprenden con resultados que hace unos pocos años eran difíciles de predecir.

2. ¿Qué es un centro de supercomputación?, ¿qué son los supercomputadores?

Los centros de supercomputación son centros que hacen investigación o desarrollan tecnología, utilizando como herramienta fundamental los supercomputadores. Estos son los computadores más rápidos del mundo utilizados para hacer ciencia.

Los *data centers* más grandes tienen muchos más procesadores que los supercomputadores más rápidos, pero están dedicados a otros objetivos tales como hacer ganar mucho dinero a sus propietarios. ¿Y cómo se construyen estas supermáquinas? Los supercomputadores se construyen utilizando centenares de miles de procesadores, muy potentes, junto con sus memorias asociadas. Los procesadores se conectan entre ellos mediante una red de interconexión muy especial, que permite enviar bits de un procesador a cualquier otro con latencia mínima y gran ancho de banda.

Una vez hemos construido un supercomputador, lo que queremos es ejecutar en ellos programas que necesiten, normalmente, gran cantidad de datos y gran número de operaciones sobre esos datos. Un ejemplo concreto sería simular la aerodinámica de un avión. Si, pongamos por caso, el programa lo hemos dividido en 100 000 partes y asignamos cada una de ellas a un procesador, tendremos que el supercomputador lo ejecutará 100 000 veces más rápido: en otras palabras, podrá ejecutar en una hora lo que un procesador tardaría 100 000; es decir, más de 10 años.

La supercomputación ha transformado profundamente la capacidad de los científicos para abordar problemas complejos en diversas áreas tales como simular fenómenos naturales a gran escala, analizar volúmenes masivos de datos en tiempo real y realizar experimentos virtuales que de otro modo serían imposibles en el mundo físico. La combinación de inteligencia artificial, en particular el aprendizaje automático, y supercomputación está revolucionando campos como el descubrimiento de fármacos o el cambio climático. Desde hace décadas el avance científico está ligado a la supercomputación. Algunos ejemplos son: predecir la estructura tridimensional de las proteínas, el

análisis genómico o producir estimaciones realistas sobre el cambio climático.

La velocidad de estas máquinas ha aumentado muchísimo debido a que, cada vez, los procesadores han sido más rápidos y a que, cada vez, los supercomputadores tienen más y más procesadores, hasta millones de procesadores en la actualidad, trabajando conjuntamente.

La ingeniería moderna y las ciencias no pueden avanzar sin el uso de la supercomputación. El uso intensivo de la IA ha acelerado esta tendencia. La supercomputación actual permite realizar simulaciones muy realistas. Usando estos datos sintéticos realistas para entrenar sistemas basados en la IA, podemos reducir el tiempo de diseño en un factor >10, y esto haciendo diseños mucho mejores que los actuales. Ejemplos claros son la aeronáutica o la descarbonización del transporte y la industria.

3. ¿Cómo ha evolucionado la velocidad de los supercomputadores?

Cada seis meses, se establecen unos *rankings* para clasificarlos. Para ello, resuelven un sistema de ecuaciones con una matriz densa y, según la velocidad que alcanzan, así se les clasifica. Se construye una tabla denominada Top-500[4] con los quinientos supercomputadores más rápidos del momento. Ya hemos comentado que hay computadores mucho más rápidos que los que ocupan las primeras posiciones de la tabla, pero que los dueños (militares, chinos y empresas como Microsoft, Meta o Amazon) no quieren descubrir las características de sus máquinas.

4 <https://top500.org>.

#	Site	Manufacturer	TOP10 Computer of the TOP500	Country	Cores	Rmax [Pflop/s]	Power [MW]
1	Lawrence Livermore National Laboratory	HPE	El Capitan HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	11,340,000	1,809	29.7
2	Oak Ridge National Laboratory	HPE	Frontier HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	9,066,176	1,353	24.6
3	Argonne National Laboratory	Intel	Aurora HPE Cray EX/Intel Exascale Compute Blade, Xeon Max 9470, Data Center GPU Max, Slingshot-11	USA	9,264,128	1,012	38.7
4	EuroHPC / FZJ	EVIDEN	JUPITER Booster BullSequana XH3000, NVIDIA GH200 Superchip, InfiniBand NDR	Germany	4,801,344	1.000	15.8
5	Microsoft Azure	Microsoft	Eagle Microsoft NDv5, Xeon Platinum 8480C, NVIDIA H100, Infiniband NDR	USA	1,123,200	561.2	
6	Eni S.p.A. Center for Computational Science	HPE	HPC6 HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Italy	3,143,520	477.9	8.5
7	RIKEN Center for Computational Science	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
8	Swiss National Supercomputing Centre (CSCS)	HPE	Alps HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, GH200, Slingshot-11	Switzerland	2,121,600	434.9	7.1
9	EuroHPC / CSC	HPE	LUMI HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,752,704	379.7	7.1
10	EuroHPC / CINECA	EVIDEN	Leonardo Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,824,768	241.2	7.5

Figura 2. Última tabla publicada de los diez supercomputadores más rápidos del mundo, que corresponde a noviembre de 2025.

Si personas pioneras en el campo de la computación levantaran la cabeza no darían crédito a sus ojos. Me refiero, por ejemplo, a Blaise Pascal (1623-1662), que diseñó calculadoras mecánicas mediante ruedas dentadas, denominadas pascalinas, usando el sistema de numeración decimal, que podían hacer operaciones básicas de suma y resta. O a George Boole (1815-1864), creador del Álgebra de Boole,[5] base de la computación digital, y Charles Babbage (1791-1871), que diseñó una calculadora mecánica capaz de calcular tablas numéricas mediante el método de diferencias finitas. Tanto Pascal como Babbage habían utilizado el sistema de numeración decimal y ruedas dentadas para diseñar sus máquinas. George Boole tenía una lógica binaria, pero no había tecnología apropiada para implementarla. Aparecieron las vál-

5 George Boole (1854), *An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities by George Boole,* Londres, Walton and Maberly.

vulas, pero no eran muy adecuadas por su enorme tamaño, consumo eléctrico y fiabilidad,

Tuvimos que esperar al invento del transistor en el año 1947, para contar con la tecnología que ha permitido esos avances enormes en la informática. El transistor fue inventado por John Bardeen, Walter Brattain y William Schockley, quienes recibieron el Premio Nobel de Física en 1956.[6] También fue muy importante la invención del circuito integrado atribuida a Jack S. Kilby en el año 1958.

Como vemos en la figura 2, el supercomputador más rápido del mundo en noviembre de 2025, se llama El Capitán y está instalado en el «Lawrence Livermore National Laboratory». Tiene más de 11 millones de procesadores y cuando ejecuta la prueba que los clasifica, denominada HPL, obtiene una velocidad de 1,809 exaflops (un exaflop es 10^{18} operaciones por segundo). Hay que decir que la velocidad máxima es de 2,74 exaflops, por lo que la eficiencia ejecutando ese programa sencillo y muy paralelo es del 66 %. A título informativo y de comparación: la primera clasificación de los supercomputadores se hizo en junio de 1993. La máquina más rápida en esa primera tabla era una Connection Machine, desarrollada en el MIT e instalada en el «Alamos National Laboratory». La máquina tenía 1024 procesadores y alcanzaba una velocidad de 59,70 gigaflops. Es decir, El Capitán es treinta millones de veces más rápido que la Connection Machine, y eso ha sido posible en treinta y tres años.

---

6 Nobel Prize in Physics (1956). <https://www.nobelprize.org/prizes/physics/1956/summary/>.

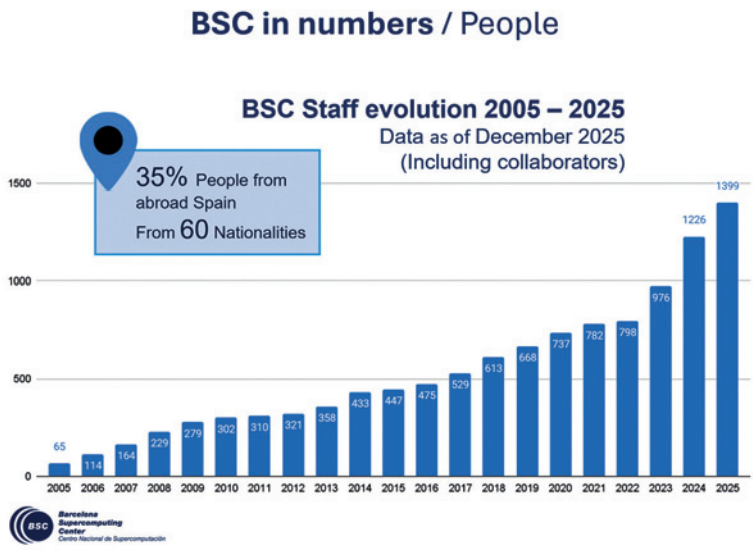


Figura 3. Evolución del *staff* en el BSC.

4. ¿Qué es el BSC, Barcelona Supercomputing Center – Centro Nacional de Supercomputación?

El BSC es el Centro Nacional Español de Supercomputación. Fue establecido en el año 2004 como continuación de otro centro que creamos en la Universidad Politécnica de Cataluña (UPC), en 1985, y que se llamó Centro Europeo de Paralelismo de Barcelona (CEPBA). Ambos centros nacieron fruto de la colaboración entre los Gobiernos de España y Cataluña y la UPC.

El BSC es el resultado de la colaboración entre las instituciones mencionadas anteriormente, y es también un punto de encuentro entre ciencia y sociedad. Nuestro objetivo es hacer ciencia excelente, que se publica en las mejores revistas y con-

gresos del mundo, y hacer ciencia relevante que ayude a resolver los problemas de la sociedad.

Creado para llegar a tener entre sesenta y setenta personas, en diciembre del año 2025 alcanzamos las mil cuatrocientas. Es importante resaltar que el 35 % de nuestro personal es de sesenta y tres países: somos capaces de atraer talento internacional.

El BSC está organizado en seis departamentos: cuatro de investigación, el de operaciones para gestionar las máquinas y el de gestión. Los cuatro departamentos de investigación son el de Ciencia de los Computadores, con casi quinientos investigadores, el de Ciencias de la Vida, con doscientos cincuenta, el de Ciencias de la Tierra, con otros doscientos cincuenta, y el de Ingeniería, con ciento cincuenta. Somos el centro más grande en supercomputación en Europa y tal vez, el centro más grande de investigación en España.

5. ¿Qué tipo de investigación se hace en el BSC?

En el BSC se hace investigación muy diversa, como indica el hecho de que tenemos investigadoras e investigadores de más de treinta carreras diferentes. Por otra parte, los investigadores del BSC tienen colaboraciones con centros de investigación, hospitales, universidades y empresas de todo el mundo. La investigación está organizada en grupos y tenemos más de setenta y cinco grupos activos de investigación. Por ello, es muy difícil resumir la investigación que se hace y, por otra parte, si describiéramos la investigación realizada por un número pequeño de grupos, no seríamos justos con el resto

Dicho esto, diremos que el Departamento de Ciencias de los Computadores realiza investigaciones en todos los aspectos relacionados con los diseños *hardware* y *software* de los supercomputadores, así como en temas de la inteligencia artificial y ahorro de energía, entre otros.

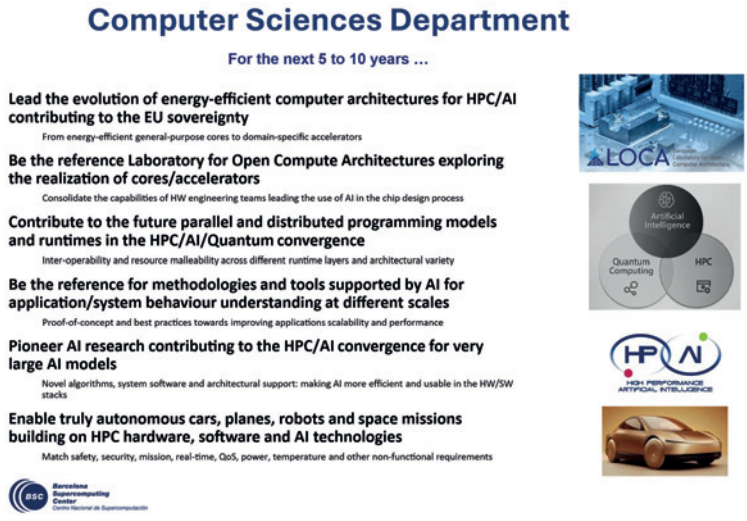


Figura 4. Resumen de las investigaciones del Departamento de Computer Science.

Los otros tres departamentos son de aplicaciones y lo que realizan, en general, son gemelos digitales o *digital twins,* que no son sino representaciones virtuales de algo que queremos conocer mejor (predicción del tiempo) o conocer por primera vez (el diseño de un nuevo material o el comportamiento del plasma en el reactor de fusión ITER, de Kadarache).

Así, un gran proyecto en el Departamento de Ciencias de la Tierra (figura 5.1) es la realización de un gemelo digital de la tierra, agua, aire y polos que nos dé una precisión nunca obtenida hasta ahora. Esto es posible, ya que los supercomputadores tienen memorias cada vez más grandes donde se pueden almacenar los datos del modelo y procesadores con velocidades enormes para operar sobre esos datos. El proyecto europeo en esa línea se denomina *Destination Earth*.

Figura 5.1. Resumen de las investigaciones del Earth Sciences Department.

El gran reto del Departamento de Ciencias de la Vida (figura 5.2) es desarrollar un gemelo digital individualizado del cuerpo humano. Tal vez es el reto más grande que tenemos hoy en día en la investigación a nivel mundial. Tales gemelos permitirían avances en la medicina nunca vistos y podríamos predecir, prevenir y curar las enfermedades de manera personalizada.

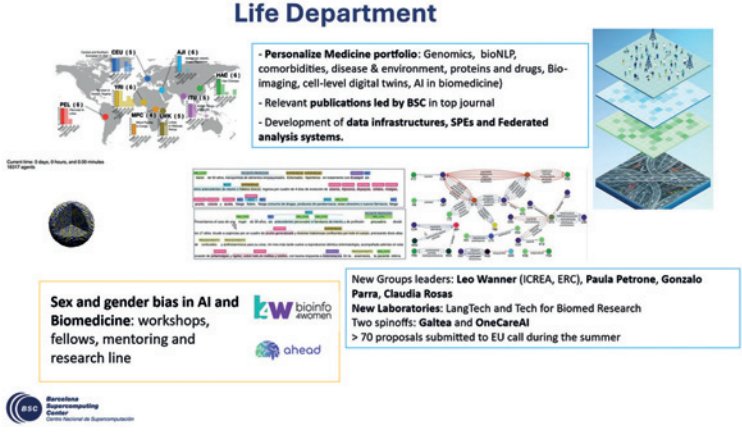


Figura 5.2. Resumen de las investigaciones del Departamento de Life Science.

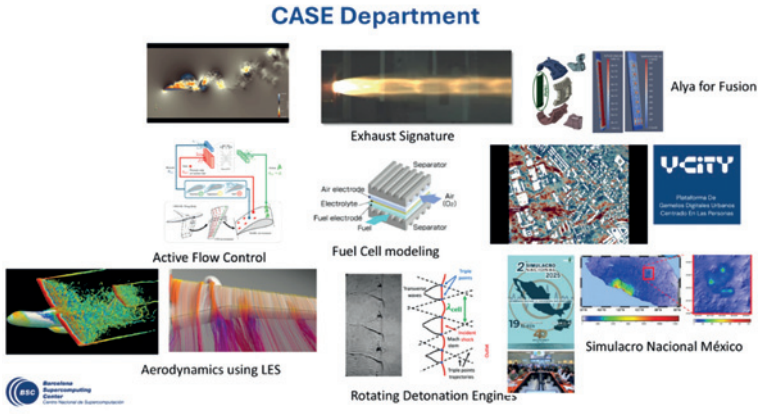


Figura 6. Resumen de las investigaciones del Departamento de Ingeniería.

El Departamento de Ingeniería realiza grandes proyectos de ingeniería, la mayoría de ellos en colaboración con empresas como pueden ser Repsol, Iberdrola, Indra y la Caixa. También realizan gemelos digitales de las ciudades y simulaciones de órganos como el corazón, y se ocupan del desarrollo de computadores cuánticos, entre otras muchas actividades.

La importancia de que en España se apoye la supercomputación no es solo una cuestión de mejora de la competitividad económica, es una cuestión de «soberanía tecnológica». Y esto afecta no solo a la economía, sino también a temas como la defensa y la seguridad de nuestro sistema político de libertades. Es importante que Europa desarrolle tecnología para lo que se ha llamado *Science for Peace and Security* (SPS).

Europa se ha dado cuenta de esto y en el siguiente programa marco de investigación, el término de tecnología dual aparecerá en todas partes.

Figura 7. Resumen de las investigaciones del Departamento de Ingeniería sobre tecnologías duales.

El BSC es un actor activo de esta política en España apoyando a toda la industria nacional y los programas del Gobierno de España. Trabajamos con empresas como Indra, Sener, ITP, GMV, Grupo Oexia, Airbus...

Trabajamos con la Unidad Militar de Emergencias (UME), en el proyecto UDRUME (figura 7). Se trata de desarrollar *software* para cualquier catástrofe natural (desde incendios a terremotos) que permita:

1. La simulación de diferentes escenarios de catástrofes naturales.
2. La toma de decisiones operativas en tiempo real ante una catástrofe.
3. El entrenamiento del personal para la actuación en una catástrofe.

El uso de la supercomputación como herramienta transversal promete avances significativos en ciencia e industria. A corto

plazo, se espera que acelere la investigación en áreas como el desarrollo de fármacos, la modelización climática y la optimización de procesos industriales. A largo plazo, podría revolucionar campos como la física cuántica, la fusión nuclear y la exploración espacial.

En nuestra opinión, resumiendo lo anterior, las oportunidades más importantes incluyen:

1. Medicina de precisión y descubrimiento de fármacos de forma más acelerada y económica.
2. Simulaciones climáticas más precisas y mejora de la estimación del impacto del cambio climático.
3. Diseño de nuevos materiales de forma más acelerada y económica.
4. Optimización de cadenas de suministro y logística, como ya lo hace, por ejemplo, Amazon.
5. Mejora de los diseños complejos en ingeniería, por ejemplo, la aeronáutica.
6. Avances en inteligencia artificial y, en particular, en agentes autónomos y aprendizaje automático. Los ejemplos paradigmáticos son los modelos masivos de lenguaje (LLM) y los modelos fundacionales (FM).

La supercomputación ya contribuye de forma decisiva a todas estas áreas económicas, proporcionando la potencia de cálculo necesaria para procesar enormes cantidades de datos y realizar simulaciones complejas, acelerando así la innovación y el progreso científico y, por supuesto, el desarrollo económico y social al reducir la brecha de uso de estas herramientas que, en el caso del MN5 y de todas las máquinas europeas, son públicas y de uso público y gratuito, y permiten que investigadores

y empresas de toda Europa tengan acceso a los supercomputadores. Además, la colaboración entre supercomputadores a través de redes como la Red Española de Supercomputación (RES) amplifica su impacto, permitiendo abordar problemas aún más complejos y fomentando la cooperación científica internacional, por ejemplo, con Latinoamérica.

6. ¿Qué relación existe entre la supercomputación y la inteligencia artificial?

La supercomputación y la inteligencia artificial son compañeros inseparables en multitud de temas tales como la llamada *inteligencia artificial generativa*. La supercomputación genera muchos datos para entrenar sistemas basados en la IA, por ejemplo, los llamados modelos masivos de lenguaje (LLM), y el proceso de entrenamiento de estos sistemas se hace sobre una infraestructura HPC. Sin HPC no puede existir la IA predictiva, que es la que está cambiando el mundo.

Como he dicho, el transistor fue inventado en el año 1947. Enseguida pasó a ser utilizado en el diseño de los computadores y con ello se abría una nueva época donde los computadores iban a ser más y más rápidos debido a que los transistores iban a ser más y más pequeños (ley de Moore).

Tal vez debido a este hecho, se reunieron en el año 1956, en el Dartmouth College,[7] los siguientes investigadores: John McCarthy (organizador), Marvin Minsky, Trenchard More, Ray Solomonoff y Julian Bigelow, que participaron las ocho sema-

7 «The Research Conference Where AI Began» (1956). <https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth>.

nas. Otros participantes fueron Nathaniel Rochester, Claude Shannon, D. M. Mackay, John Holland, Oliver Selfridge, Ross W. Ashby, Allen Newell y Herbert Simon.

Los objetivos de la reunión fueron descritos por John McCarthy en su propuesta de petición de financiación:

> Proponemos que durante el verano de 1956 tenga lugar en el Dartmouth College en Hanover, Nuevo Hampshire un estudio que dure dos meses, para once personas. El estudio es para proceder sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, ser descrito con tanta precisión que puede fabricarse una máquina para simularlo. Se intentará averiguar cómo fabricar máquinas que utilicen el lenguaje, formen abstracciones y conceptos, resuelvan las clases de problemas ahora reservados para los seres humanos, y mejoren por sí mismas. Creemos que puede llevarse a cabo un avance significativo en uno o más de estos problemas si un grupo de científicos cuidadosamente seleccionados trabajan en ello conjuntamente durante un verano.

Es en esa reunión donde McCarthy propone el término de *inteligencia artificial* para diferenciarlo de la cibernética.

Aunque las personas que asistieron a la reunión, así como sus colaboradores y colegas tenían un potencial investigador increíble, se pusieron cotas temporales a objetivos concretos que no pudieron cumplirse (por ejemplo, el reconocimiento y la traducción automática del lenguaje). Y tal vez la razón fundamental era que los computadores todavía no tenían la potencia de cálculo que la inteligencia artificial está utilizando hoy en día para no dejar de sorprendernos a cada momento.

Durante muchos años, las aplicaciones de la inteligencia artificial estuvieron aparcadas en el invierno polar, ya que no daban los resultados que prometían. Tal vez, una de las primeras demostraciones de que algo estaba cambiando fue cuando

el computador Deep Blue de IBM le ganó al campeón mundial de ajedrez, Garri Kaspárov en 1997. Hay que decir aquí que el Deep Blue era una máquina con apenas mil procesadores y 450 chips especializados en el juego del ajedrez. Sin embargo, algo poco conocido es que la primera partida a nivel mundial entre el Deep Blue y un jugador humano, en este caso Miquel Illescas, que resultó ganador, tuvo lugar en Barcelona en el año 1995, fue organizada por nuestro grupo de investigación durante la celebración del Congreso International Conferencie on Supercomputing (ICS) del ACM.

En febrero de 2011 Watson, otro supercomputador creado por IBM, logró ganar jugando al programa televisivo *Jeopardy* en el que derrotó a los dos máximos campeones en la historia del programa. A diferencia de Deep Blue, debió ser programado con una vasta cantidad de información, y con un motor de inteligencia artificial capaz de decidir por sí mismo la respuesta más acertada a cada pregunta del programa. El computador Watson que jugó al *Jeopardy* tenía noventa nodos donde cada nodo contenía un chip con 8 procesadores Power 750 con un reloj de 3,5 Ghz y un total de 16 terabytes de RAM. En ese momento, el supercomputador más potente del mundo era el K Computer que poseía más de 700 000 procesadores, por lo que el Watson era una máquina mucho menos potente que el K Computer.

Otro gran hito en la intersección entre los juegos y la inteligencia artificial se produjo en 2016, cuando el programa AlphaGo, ejecutado en un conjunto de aceleradores de Nvidia, derrotó al campeón mundial de Go, Lee Sedol, en un histórico enfrentamiento de cinco partidas.

Pero, tal vez, el momento más importante en el que el potencial de la IA llegó al público fue cuando la empresa OpenAI, en 2022, lanzó públicamente el programa ChatGPT. Este programa pertenece a los denominados *large language models*

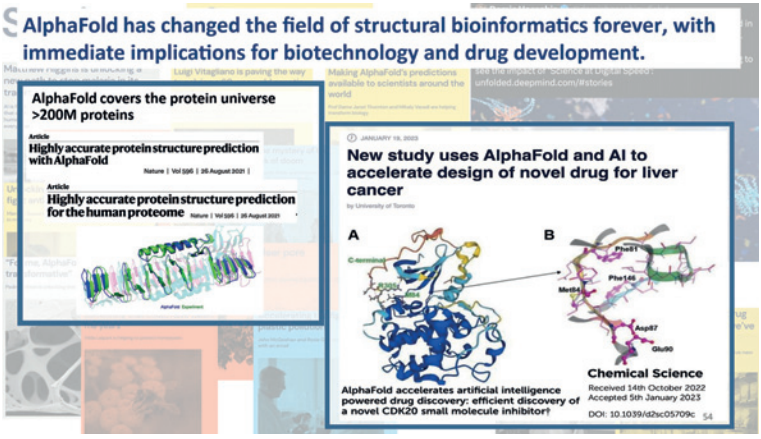
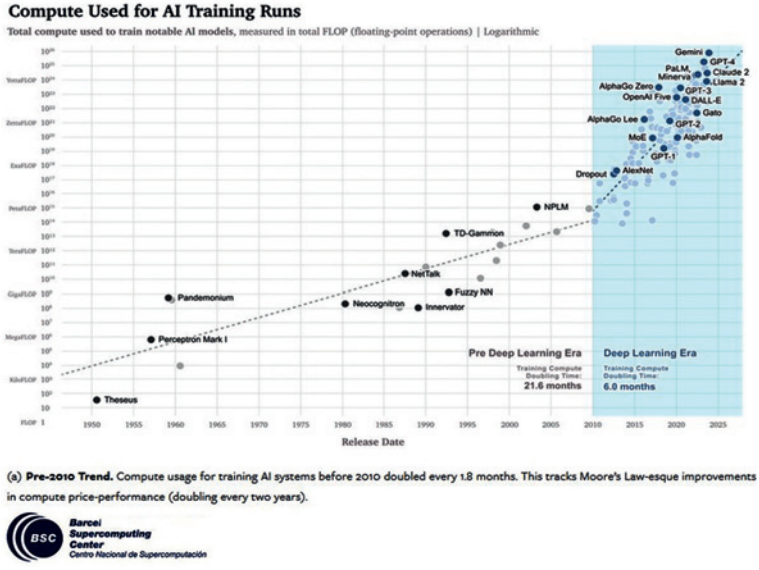


Figura 8. Plegamiento de las proteínas a partir de la secuencia de aminoácidos.

(LLM), que introducen el concepto de *inteligencia artificial generativa,* que es la que permite crear/generar respuestas muy inteligentes a preguntas hechas por los usuarios. Con ChatGPT se abría un nuevo mundo en el campo de la IA y su uso por la sociedad. Juntos, la supercomputación y la IA están produciendo resultados impresionantes, tales como predecir como será la estructura 3D de una proteína a partir de la secuencia de aminoácidos que la produce (figura 8). Este problema era de Premio Nobel y tanto es así que a los investigadores se les ha concedido el Premio Nobel de Química de 2024.

7. ¿Puede haber inteligencia artificial sin la ayuda de los supercomputadores?

La respuesta es sí. Existen muchas variantes de la IA y la mayor parte de ellas no necesitan de la supercomputación. Sin



(a) Pre-2010 Trend. Compute usage for training AI systems before 2010 doubled every 1.8 months. This tracks Moore's Law-esque improvements in compute price-performance (doubling every two years).

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Figura 9. Cálculos necesitados para entrenar los *large language models* (LLM).

embargo, la IA generativa, que es la que está de moda, necesita muchos datos y mucha computación. Las redes hay que entrenarlas con datos, de forma que al final del entrenamiento, la red ha adquirido una serie de valores o parámetros que caracterizan al modelo LLM de la red. Dependiendo del número de parámetros que queramos obtener (indicativo en principio de la calidad del sistema) y del conjunto de valores de entrada, el número de operaciones que hay que realizar es astronómico. Una aproximación es decir que las operaciones necesarias serán seis multiplicado por el número de parámetros de la red y por el número de *tokens* utilizados para entrenar la red.

En la figura 9, de enero de 2025, se muestran el número de operaciones necesarias para entrenar LLM populares, pertenecientes a diferentes compañías. En la parte alta de la tabla está

el modelo Gemini 2 de Google que necesitó casi 10^{26} operaciones para obtener los parámetros. Este es un número crítico que ha sido definido por el Gobierno americano como el valor límite a partir del cual el modelo debe ponerse a disposición de la Administración americana por motivos de la seguridad nacional. En la figura 9 se puede ver cómo ChatGPT-4 necesitó un poco más de dos por 10^{25} operaciones. Otra información que se puede obtener de la figura 9 es que hasta el año 2010, la necesidad de cálculos de los modelos se multiplicaba por dos cada veintiún meses y, desde el año 2010, se multiplican por dos cada seis meses. Esta escala de crecimiento no se puede mantener por mucho tiempo.

Profundizaremos un poco más en este tema. Se estima que entrenar GPT-3, que contiene 175 000 millones de parámetros, necesitó de 3 por 10^{23} operaciones, y consumió alrededor de 1300 MWh de electricidad. Las estimaciones, más bien conservadoras, para GPT-5, estiman que su entrenamiento podría haber consumido entre 50 000 y 100 000 MWh; es decir, entre 40 y 80 veces más. Para expresarlo en valores que entendemos mejor, podríamos decir que 100 000 MWh es el consumo anual de electricidad de más de 8000 hogares y equivalente, también, a la energía que generarían 3000 turbinas eólicas funcionando todo el día.

En cuanto al agua requerida, podemos decir que entrenar GPT-3 en los centros de datos de Microsoft pudo haber consumido 700 000 litros de agua, equivalente para producir 370 coches BMW o 320 Teslas.

Como hemos comentado anteriormente, ChatGPT-3 necesitó 3,14 por 10^{23} flops (314 zettaflops) operaciones. Las estimaciones para ChatGPT-4 fueron de 2,15 por 10^{25}; es decir, setenta veces más que GPT-3. Una estimación para ChatGPT-5 puede estar entre 1 a 5 por 10^{26} que es un número increíble. Como analo-

gía, si cada flops fuera un grano de arena 10^{26} sería mayor que la arena de todas las playas y desiertos del planeta, estimadas en menos de 10^{19} y que todas las estrellas del universo observable estimadas en 10^{24}.

El número de operaciones para entrenar Gemini-3 es parecido al de ChatGPT-5. Un dato adicional es que al precio actual del *cloud computing* 10^{26} operaciones podrían costar unos pocos cientos de millones de dólares.

¿Qué significan 10^{26} operaciones cuando se ejecutan en supercomputadores o *data centers*? Supongamos que el acelerador es el H100 de Nvidia. Este chip es capaz de realizar 2 peta operaciones por segundo, utilizando como operandos números codificados en formatos de 16 bits tales como el FP16 o el BF16. Haciendo una división sencilla nos da que un solo chip tardaría 1585 años en realizar esas 10^{26} operaciones. Debido a que vamos a usar muchos chips en paralelo, podemos decir que, si utilizamos 25 000 chips, tardaremos 23 días y que si utilizamos 50 000 tardaríamos 11,5 días. A título de anécdota, el Marenostrum 5 tiene 4960 chips del tipo Hopper 100, por lo que el tiempo necesario sería de 125 días más o menos.

8. ¿Podemos comparar estos valores con los de otras industrias tecnológicas como la minería de bitcoins, los videojuegos o los centros de datos tradicionales?

Tal como hemos comentado, el entrenamiento de un GPT-4 avanzado puede consumir entre 0,1-0,5 TWh (similar al consumo de una ciudad de entre 50 000 y 100 000 habitantes en un año). Es muy difícil calcular cuánto se consume en inferencias, pero la energía necesaria está creciendo día a día y podemos valorarla entre 1 y 2 TWh/año.

La minería de bitcoin a nivel mundial puede estar consumiendo alrededor de 150 TWh/año, equivalente a 20 gigavatios, que es el consumo global de países como Polonia, o también a 2/3 del consumo de España.

En cuanto a los videojuegos, la industria a nivel mundial consume entre 230 y 350 TWh/año, incluyendo consolas, PC y centros de datos para jugar en la nube. Este consumo es equivalente al de Italia. Tengamos en cuenta que hay millones de usuarios.

En los centros de datos (nube más IA), el gasto de energía se calcula entre 400 y 450 TWh/año e incluye búsquedas en la web, *streaming* como Netflix y YouTube, correo, redes sociales; es decir, la infraestructura digital mundial. Solo el *streaming* de vídeo más Netflix y YouTube, consumen entre 80 y 100 TWh/año, casi tanto como los bitcoin. El gasto de los centros de datos (entre 250 y 300 TWh), equivale al consumo de España más el de los Países Bajos.

El gasto de todas las *foundries* del mundo podría estimarse entre 150 y 250 TWh, valor que va creciendo año a año. Comentaremos el caso particular de la *foundry* TSMC más adelante

Como resumen podemos decir que el gasto en centros de datos gana por mucho en el consumo total. Si los datos son válidos, podemos decir que su consumo es alrededor de 150 veces más que el entrenamiento de todos los modelos grandes de IA combinados. Los videojuegos son grandes consumidores, aunque su uso es de millones de jugadores.

Toda la IA generativa es alrededor del 1-2 % del consumo total de los centros de datos, aunque el aumento en el consumo de energía de los centros de datos podría duplicarse para el año 2030, impulsado por el uso de la IA.

9. ¿Cómo está influyendo la inteligencia artificial en las investigaciones que se realizan en el BSC?

Como hemos dicho, la IA y la supercomputación se dan la mano para hacer que los investigadores puedan soñar. La IA está cambiando la forma de hacer investigación en cualquier centro de supercomputación. De hecho, podríamos decir ahora que no existen ya centros de supercomputación clásicos, sino que lo que existen son centros de inteligencia artificial y que los mejores, tal vez, son los que albergan los supercomputadores más rápidos del mundo.

Desde hace años, el BSC adquirió máquinas orientadas a la ejecución de programas de IA. Y, cada vez más, la potencia de las máquinas dedicadas a la IA en el BSC ha ido aumentando, de forma que una de las cuatro partes del MN5 (figura 10), es una máquina basada en aceleradores de Nvidia de tipo Hopper, H-100, que eran los más rápidos del mundo en el año 2023. Desde entonces, Nvidia diseñó el Blackwell que es la generación siguiente a los H-100, y recientemente el Rubin.

Tal como se ve en la figura 10, hay una partición del MN5 que alcanza una velocidad pico de 260 petaflops. Contiene 4096 aceleradores Hopper de Nvidia y, en noviembre de 2023, esta máquina se clasificó como la octava del mundo y, en noviembre de 2025, está en la posición decimocuarta. Conviene resaltar aquí que le empresa Meta tiene una máquina, que no compite por el Top-500, que contiene 100 000 aceleradores; es decir, más de veinte veces más rápida que la del BSC, y que hubiera sido el número uno del mundo si pasara la prueba del HPL, para posicionarse en la lista del Top-500.

En general, los *data centers* tienen centenares de miles de aceleradores y pueden llegar a consumir hasta 500 megavatios; es una locura. En ellos, la velocidad no se mide por los flops

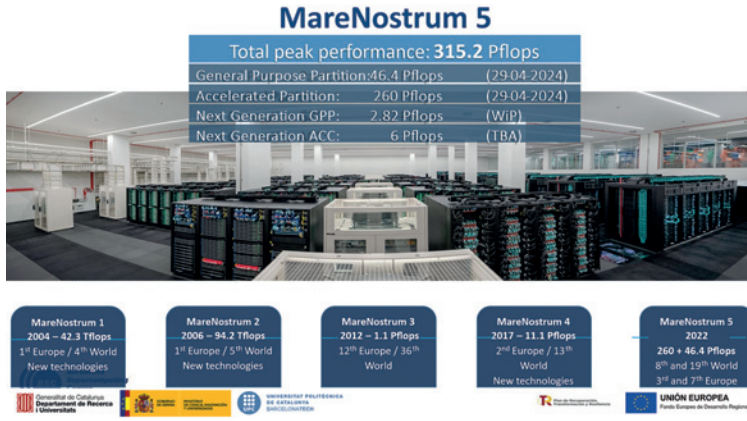


Figura 10. Componentes del MareNostrum 5.

realizando operaciones con números de 64 bits como en el Top-500. En ellos se hacen más operaciones con números que necesitan menos bits para ejecutar los LLM. Hoy en día, podemos decir que es normal alcanzar unas velocidades entre 1-2 exaflops en un formato de números de 16 bits denominado BF16, cada 20 megavatios. Así, una instalación, de las que hay varias, con 200 megavatios puede alcanzar una velocidad de entre 10-20 exaflops con datos formateados en BF16.

En todos los departamentos del BSC hay investigadores que investigan en IA o usan sistemas basados en la IA. Y cada vez más, el número va creciendo. En la actualidad, son más de trescientas personas (figura 11). Lo más importante es que los investigadores dominan desde el *software* para desarrollar, por ejemplo, un Large Language Model (LLM), hasta el diseño de chips orientados a una ejecución eficiente de las aplicaciones de inteligencia artificial.

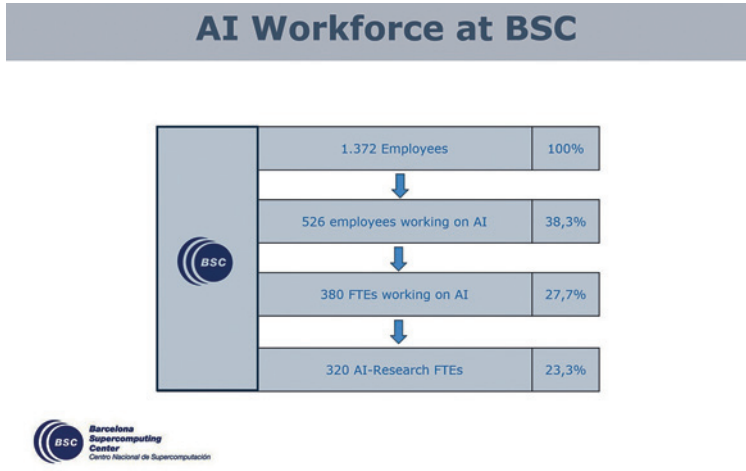


Figura 11. Personal del BSC relacionado con la inteligencia artificial en diciembre de 2025.

No es el objetivo de este escrito describir los muchos y buenos trabajos que hacen los investigadores del BSC en este campo. En la figura 12 hay un resumen de los temas en los que trabajan los diferentes departamentos.

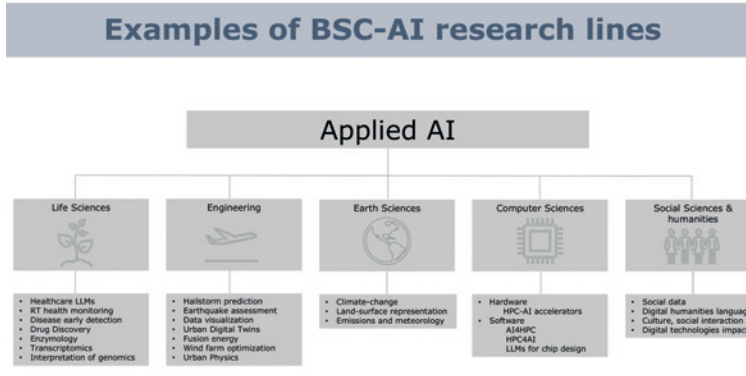


Figura 12. Departamentos del BSC y temas de la inteligencia artificial donde investigan.

En algunas aplicaciones se ha sustituido el uso de las técnicas tradicionales de resolver problemas —por ejemplo, uso de la teoría de la dinámica de fluidos— por el entrenamiento y posterior uso de tecnologías del aprendizaje automático *(surrogate computing)*. Un ejemplo concreto, desarrollado por investigadores del Departamento de Ingeniería, es la simulación de la aerodinámica de un avión en vuelo. Este problema tarda muchísimo tiempo de simulación usando las técnicas tradicionales y el uso de los sistemas basados en la IA hacen que la simulación se realice en un tiempo *razonable*. Hay algunos ejemplos de los gemelos digitales en los que el uso de la IA ha hecho mejorar los diseños y tiempos de ejecución. Un ejemplo claro es el proyecto Destination Earth, que es un gemelo digital de la tierra. Todos los departamentos tienen actividades de los modelos LLM fundacionales. Por ejemplo, el Departamento de Ciencias de la Vida está desarrollando los modelos Alia de procesado de lenguaje natural para preservar las cuatro lenguas del Estado español, y el Departamento de Ciencia de los Computadores un modelo funcional de texto e imágenes médicas. También realizamos investigación en procesado de imágenes, diseño de arquitecturas *hardware* y actividades orientadas a potenciar los aspectos éticos y de credibilidad de la IA.

Debido a la importancia de la IA en los centros de supercomputación, hemos creado en el BSC un Instituto de inteligencia artificial (figura 12), que es una estructura horizontal a los cuatro departamentos de investigación. Desde el Instituto queremos coordinar las investigaciones realizadas en los departamentos, así como dar visibilidad al BSC.

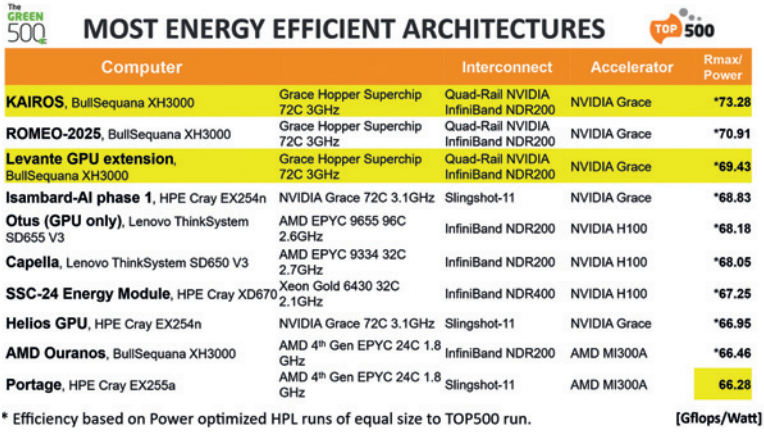
10. ¿Consumen mucha energía los supercomputadores?

Los supercomputadores consumen mucha energía por el hecho de que contienen hasta millones de procesadores y ace-

leradores de alta velocidad realizando operaciones, así como grandes memorias y discos donde se leen y escriben los datos del programa y una red de interconexión ultrarrápida que permite intercambiar datos a los procesadores.

En la figura 2 de los Top-10 supercomputadores, hay información del gasto energético en la última columna. Algunos supercomputadores llegan a consumir más de 30 megavatios, siendo el de Argonne (la actual número 3 del mundo) el que más consume con 38,7 megavatios. A ese consumo hay que sumarle la energía necesaria para refrigerar la máquina y que puede tener un valor superior al 20 % del consumo. Si esto es así, el consumo del supercomputador Aurora instalado en Argonne será superior a 45 megavatios. A nivel económico, ese consumo costaría en España alrededor de 50 millones de euros por año. Por otra parte, significa que 22 máquinas como ella necesitarían toda la energía de una central nuclear como la de Vandellós, que produce 1 gigavatio.

La lista Green-500 clasifica a los 500 supercomputadores más rápidos del mundo; es decir, los que están en la lista Top-500, en función de la energía que consumen al ejecutar el test HPL, que es el usado para clasificarlos en la lista de los Top-500. Los valores se expresan como gigaflops por vatio; es decir, cuantos gigaflops ejecuta la máquina con un vatio. La primera lista se estableció en junio de 2013 y el computador primero de la lista era uno instalado en Cineca (Bolonia, Italia), construido con procesadores Intel y aceleradores Kepler de Nvidia. El valor de su eficiencia fue de 3,3 gigaflops por vatio. En la última lista de noviembre de 2025 (figura 13) el número uno en la lista fue el supercomputador KAIROS instalado en Jüelich (Alemania). Alcanzó el valor de 73,28 gigaflops por vatio. A esa mejora energética de casi 25 veces en 11 años, ha contribuido, de manera fundamental, la reducción en el tamaño de los transistores,



MOST ENERGY EFFICIENT ARCHITECTURES

Computer		Interconnect	Accelerator	Rmax/ Power
KAIROS, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*73.28
ROMEO-2025, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*70.91
Levante GPU extension, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*69.43
Isambard-AI phase 1, HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA Grace	*68.83
Otus (GPU only), Lenovo ThinkSystem SD655 V3	AMD EPYC 9655 96C 2.6GHz	InfiniBand NDR200	NVIDIA H100	*68.18
Capella, Lenovo ThinkSystem SD650 V3	AMD EPYC 9334 32C 2.7GHz	InfiniBand NDR200	NVIDIA H100	*68.05
SSC-24 Energy Module, HPE Cray XD670	Xeon Gold 6430 32C 2.1GHz	InfiniBand NDR400	NVIDIA H100	*67.25
Helios GPU, HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA Grace	*66.95
AMD Ouranos, BullSequana XH3000	AMD 4th Gen EPYC 24C 1.8 GHz	InfiniBand NDR200	AMD MI300A	*66.46
Portage, HPE Cray EX255a	AMD 4th Gen EPYC 24C 1.8 GHz	Slingshot-11	AMD MI300A	66.28

* Efficiency based on Power optimized HPL runs of equal size to TOP500 run. [Gflops/Watt]

Figura 13. Lista de los computadores que consumen menos energía por operación en noviembre de 2025.

junto con el uso de nuevos algoritmos que reducen, significativamente, el movimiento de datos entre procesadores, fuente de una parte muy importante del consumo energético de los supercomputadores.

11. ¿Qué otros *benchmarks* hay para clasificar a los supercomputadores?

Hasta ahora hemos comentado los *benchmarks* asociados al Top-500 y al Green-500. Existe otro denominado *High Performance Conjugent Gradient* (HPCG), donde el objetivo es calcular el gradiente conjugado de una matriz muy grande (comparada con el tamaño de las memorias cachés de los chips) y dispersa (el número de elementos de la matriz diferentes a cero son muy pocos). Este problema tiene la característica de que la operación básica es en matriz por vector, por lo que la reutilización

HPCG TOP 10, NOVEMBER 2025

Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	DOE/SC/LLNL **USA**	**El Capitan**, HPE Cray 255a, AMD 4th Gen EPYC 24C 1.8 GHz, AMD Instinct MI300A, Slingshot-11	11,039,616	1809	1	17.4	**0.6%**
2	RIKEN Center for Computational Science **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D, Fujitsu	7,630,848	442	7	16.0	**3.0%**
3	DOE/SC/ORNL **USA**	**Frontier**, HPE Cray Ex235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot-11	9,066,176	1353	2	14.1	**0.7%**
4	DOE/SC/ANL **USA**	**Aurora**, HPE Cray EX, Intel Max 9470 52C, 2.4 GHz, Intel GPU MAX, Slingshot-11	9,264,128	1012	3	5.6	**0.3%**
5	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,752,704	380	9	4.6	**0.9%**
6	Softbank **Japan**	**CHIE-4**, NVIDIA DGX B200, Xeon Platinum 8570 56C 2.1 GHz, B200 SXM 180GB, Infiniband NDR400	662,256	135	17	3.76	**2.5%**
7	CSCS **Switzerland**	**Alps**, HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	2,121,600	435	8	3.67	**0.6%**
8	EuroHPC/CINECA **Italy**	**Leonardo**, BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 40 GB, Quad-rail NVIDIA HDR100 Infiniband	1,824,768	241	10	3.1	**1.0%**
9	ExxonMobil **USA**	**Discovery 6**, HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	822,528	164	15	2.6	**1.2%**
10	AIST **Japan**	**ABCI 3.0**, HPE Cray XD670, Xeon Platinum 8558 48C 2.1GHz, NVIDIA H200 SXM5 141 GB, Infiniband NDR200	479,232	145	16	2.4	**1.3%**

Figura 14. Lista de los supercomputadores mas rápidos del mundo ejecutando el test HPCG en noviembre de 2025.

de los datos es prácticamente nula. Además, las matrices son dispersas, con la mayoría de los elementos con valor cero. Esto hace que el acceso a estos elementos en memoria se haga de manera mucho más lenta que en el caso de que las matrices fueran densas, como es el caso del test HPL, usado para crear la tabla de los TOP-500. Por lo tanto, es un problema cuya eficiencia está limitada por la relación entre el ancho de banda de las memorias externas y la capacidad de cálculo de los procesadores dentro del chip. Cuanto menor sea el cociente bytes/flops, peor será la eficiencia en la ejecución del *benchmark* HPCG. Los resultados son bastante desalentadores tal como se indican en la tabla de los 10 computadores con mejores eficiencias.

El número uno en esta nueva lista (figura 14) es de nuevo el computador Capitán, instalado en el Lawrence Livermore National Laboratory (número 1 en la lista del Top-500), que obtiene una eficiencia del 0,60 %. Vemos que apenas hay supercomputadores que pasen el valor del 2,6 %, con la excepción del

supercomputador Fugaku instalado en Riken, Japón, que tiene una eficiencia del 3,0 %.

Los *benchmarks* utilizados para clasificar los supercomputadores deberían representar las aplicaciones reales que ejecutan. Durante muchos años, las aplicaciones que ejecutaban venían del campo de la ingeniería y se caracterizaban, entre otras cosas, porque sus variables se codificaban como números reales con 64 y/o 32 bits. Así, el test HPL del Linpack y el HPCG usan datos codificados en 64 bits. Las aplicaciones provenientes de la inteligencia artificial usan datos codificados de manera variable, con un número menor de bits. Si los datos son más pequeños, se pueden utilizar más unidades funcionales en el mismo espacio del chip e ir más rápido. Y este es el caso real que aparece, por ejemplo, en los LLM.

Para reflejar este nuevo tipo de aplicaciones se ha propuesto el test HPL-MxP, que es un HPL con precisión mixta. La idea es sustituir algunos de los cálculos del HPL que se realizan con 64 bits, como la descomposición LU y hacerlos con formatos de

HPL-MxP Top 10 for November 2025

Rank	Site	Computer	Cores	HPL Rmax (Eflop/s)	TOP500 Rank	HPL-MxP (Eflop/s)	Speedup
1	DOE/SC/LLNL **USA**	**El Capitan**, HPE Cray 255a, AMD 4th Gen EPYC 24C 1.8 GHz, AMD Instinct MI300A, Slingshot-11	11,340,000	1.809	1	16.7	9.6
2	DOE/SC/ANL **USA**	**Aurora**, HPE Cray EX, Intel Max 9470 52C, 2.4 GHz, Intel GPU MAX, Slingshot-11	8,159,232	1.012	3	11.6	11.5
3	DOE/SC/ORNL **USA**	**Frontier**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	8,560,640	1.353	2	11.4	8.4
4	EuroHPC/FZJ **Germany**	**JUPITER Booster**, BullSequana XH3000, NVIDIA Grace 72C 3GHz, GH200, NVIDIA Infiniband NDR200	4,801,344	1.0	4	6.25	6.3
5	Softbank **Japan**	**CHIE-4**, NVIDIA DGX B200, Xeon Platinum 8570 56C 2.1 GHz, B200 SXM 180GB, Infiniband NDR400	662,256	0.135	17	3.3	24.4
6	AIST **Japan**	**ABCI 3.0**, HPE Cray XD670, Xeon Platinum 48C 2.1GHz, NVIDIA H200 SXM5 141 GB, Infiniband NDR200	479,232	0.145	16	2.36_8 1.2_{16}	16.3_8 8.3_{16}
7	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,752,704	0.380	9	2.35	6.2
8	RIKEN Center for Computational Science, **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D	7,630,848	0.442	7	2.0	4.5
9	EuroHPC/CINECA **Italy**	**Leonardo**, BullSequana XH2000, Xeon Pl. 32C 2.6GHz, NVIDIA A100 SXM4 40 GB, QR NVIDIA HDR100 IB	1,824,768	0.241	10	1.8	7.6
10	KAIST **Saudi Arabia**	**Shaheen III**, HOE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200, Slingshot-11	574,464	0.123	18	0.7	5.7

Figura 15. Lista de los supercomputadores más rápidos del mundo ejecutando un test relacionado con la inteligencia artificial.

datos más pequeños tales como los FP16 o bf16, que usaba 16 bits. En la figura 15 se puede ver que el supercomputador más rápido ejecutando esta prueba es de nuevo el Capitán. La velocidad que obtiene es de 16,7 exaflops por segundo. Tengamos en cuenta que las operaciones se hacen con formatos más pequeños y por eso superan la velocidad de 1,74 exaflops que obtiene al ejecutar el HPL clásico. En la última columna se puede ver la relación, para cada computador, al ejecutar la prueba HPL (que da lugar a la lista Top-500) y el test HPL-MxP.

Esta evaluación de las velocidades a las que ejecutan los supercomputadores las aplicaciones de la IA hará posible que, mucho antes que si solo se utilizara la prueba HPL, habrá supercomputadores que batirán la barrera de los zettaflops, es decir, 10^{21} operaciones por segundo.

12 ¿Pero el uso agresivo de la inteligencia artificial ha aumentado/aumentará enormemente el consumo de energía?

La respuesta es afirmativa si seguimos aumentando al ritmo actual, por ejemplo, el número de LLM en el mundo, el número de parámetros asociados a esos modelos, el número y el tipo de datos para entrenarlos, así como el número y el tipo de preguntas que les hacemos. Asociado a estas hipótesis está el anuncio de finales de enero de 2025 de Donald Trump de cofinanciar un proyecto con un presupuesto de 500 000 millones de dólares. A él contribuirían por el momento, el Gobierno americano, OpenAI, Oracle, Microsoft, Nvidia, un fondo soberano de Abu Dhabi y Softbank. El proyecto se llama Stargate

y el dinero se dedicará a la construcción de enormes centros de datos y de centrales nucleares de 100 megavatios cada una, para alimentar a los centros de datos. Según los americanos, el proyecto Stargate es similar al proyecto Manhattan en cuanto a dinero y objetivos: en ambos es dominar el mundo. La instalación se está construyendo en Abilene (Texas), ocupando una extensión aproximada de tres millones y medio de metros cuadrados, trescientas cincuenta hectáreas, equivalente a todo el Central Park de Nueva York que tiene 34 hectáreas. Pero para ver un poco cómo está el tema de la construcción de máquinas con miles y miles de aceleradores, diremos que no había pasado ni una semana del anuncio del proyecto Stargate, cuando Mark Zuckerberg, fundador y presidente de Meta (antigua Facebook), anunció que su compañía dará servicio en el año 2025 a más de 1000 millones de personas y que para ello necesita construir nuevos servidores. La máquina tendrá más de 1,3 millones de GPU, llegará a consumir hasta 2 gigavatios (equivalente a dos centrales nucleares o al 7 % de toda la energía que consumimos en España). El precio de la inversión será de 60 000 millones de dólares.

ChatGPT utilizó, para entrenar la red, 10 000 chips Ampere de Nvidia, equivalentes en potencia a 3000 Hopper, chip que siguió al Ampere, y de los que hay 4480 que contiene el MareNostrum 5, o a 700 de los recién anunciados Blackwell, que sigue al Hopper. Para fabricar una máquina que permitiera obtener el 10 % de la capacidad de cálculo de Google, se necesitaría una inversión inicial de 48 000 millones de dólares, con 1000 millones adicionales anuales para mantener las peticiones de los usuarios.

ChatGPT es un gran modelo con muchos parámetros de entrada. Pero hay propuestas para hacer modelos mucho más grandes. Tengamos en cuenta que entrenar un modelo de

1 billón europeo de parámetros (1 trillón de parámetros para los americanos) puede necesitar más de 10^{26} operaciones con números en coma flotante de 16 bits (formato bf16, propuesto por Google). Para decirlo de otra manera, necesitaría 204 días enteros del supercomputador Aurora de Argonne, que es desde noviembre de 2023, el segundo supercomputador más rápido del mundo. Si se tuviera que alquilar ese servicio a Amazon, Google o Microsoft, el coste sería de 1000 millones de euros, aproximadamente.

Y los supercomputadores, tal como hemos comentado anteriormente, consumen mucha energía. Si se sigue cumpliendo la evolución en el número de modelos que se van a desarrollar, así como de los parámetros a utilizar y el tipo de inferencias que esos modelos han de servir, al ritmo de los últimos años, que ha sido el de duplicar la potencia de cálculo necesitada cada seis meses, vamos a necesitar una cantidad de energía preocupante.

Siguiendo este ritmo, se calcula que, en el año 2028, la IA consumiría alrededor de 300 teravatios hora, equivalente a más de 30 gigavatios. Como comparación diremos que España consume un total de 30 gigavatios y que una central nuclear produce alrededor de un gigavatio. Para los Estados Unidos supondría casi un 0,5 % de su gasto energético. Microsoft está consumiendo casi 2 gigavatios.

Sin embargo, también a finales de enero de 2025, cuatro días después del anuncio de Trump, apareció una noticia que puede cambiar todo lo que hemos comentado hasta ahora. Se trata de que una compañía china, denominada DeepSeek, anunció un modelo R1 que necesitó mucho menos tiempo de cálculo y menos datos de entrenamiento que los grandes modelos de Google, Amazon, Meta, y que obtuvo mejores resultados que los modelos más complicados.

Hay un refrán que dice: «El hambre agudiza el ingenio». Y está admitido a nivel científico que ese refrán cobra sentido porque cuando el cuerpo nota que necesita comer, dispara mecanismos orientados a buscar comida. Y por eso, entre otras cosas, es tan difícil perder peso cuando la comida sobra. China fue *castigada* por los americanos para que no tuviera acceso a fabricar chips que usaran las últimas tecnologías en el tamaño de los transistores (ahora está entre 2 y 3 nanómetros) y para que no pudiera comprar los chips más potentes, tales como el Hopper y el Blackwell de la empresa Nvidia. Estas restricciones agudizaron el ingenio de muchos investigadores chinos de forma que DeepSeek es un gran ejemplo. En diciembre de 2025, esta restricción ha sido suprimida.

El modelo fue comparado con el último modelo de OpenAI y con el Claude de Anthropic y los resultados le fueron favorables en *benchmarks* como MATH-500. El modelo tiene una herramienta especializada en razonamiento en matemáticas, química, programación y razonamiento lógico. El coste de desarrollar el modelo fue de unos 6 millones de dólares y empleó menos de 3 millones de horas de GPU. Meta gastó 60 millones en entrenar el modelo Llama 3 y OpenAI 100 millones en su último modelo. Hemos de decir que, anticipándose al embargo americano, la empresa compró varios miles de GPU de tipo Ampere, que fueron las anteriores a Hopper y que son la mitad de rápidas. Eso quiere decir que con las Hopper hubiera tardado la mitad de las horas y gastada menos energía. El modelo no maneja vídeos por el momento y tiene anuladas las respuestas comprometedoras para China. Por otro lado, ha desarrollado minimodelos para aplicaciones específicas.

13. ¿Podemos decir algo acerca del consumo de electricidad y agua de los *data centers* y en los centros de supercomputación?

Cuando hablamos de supercomputadores y de *data centers,* mencionamos el número y el tipo de procesadores y de aceleradores, la memoria central, en discos y cintas, y la red de interconexión. Y ya vemos que en la lista del Top-500, aparece, junto a cada supercomputador, su consumo energético. Sin embargo, no se habla de la importancia del agua para generar la energía utilizada en los *data centers,* así como para refrigerarlos. Hablaremos más delante de los consumos de agua y electricidad en las *foundries,* tales como la TSMC.

Es bastante normal que los *data centers* y los supercomputadores usen, actualmente, agua para refrigerar los circuitos que los componen. Al trabajar, los procesadores, los aceleradores, las memorias, la red de interconexión y los sistemas de almacenamiento producen calor que hay que quitarles para su correcto funcionamiento. Hay que enfriar los *racks* o bastidores y existen diferentes técnicas que dependen del calor a disipar, y del coste que se quiera aceptar.

Cuando tenemos *racks* con poca capacidad de cálculo, la refrigeración por aire suele ser la utilizada, ya que es la más económica. Por otro lado, hay la previsión de tener dentro de poco, *racks* que podrán consumir 1 megavatio de electricidad. En esta situación, la densidad de cálculo será enorme, así como el calor generado y ya se está pensando en refrigerar las placas, introduciéndolas en un líquido no conductor. El precio de diseño será más costoso, pero no se necesitará agua para refrigerar las placas por lo que el precio inicial se podrá ir recuperando con el tiempo.

En la actualidad existen muchos sistemas que usan el agua para refrigerarlos. Y hay varias posibilidades, desde las que utilizan un circuito cerrado, por lo que no necesitan gran cantidad de agua pero sí más electricidad, hasta las que evaporan el agua por lo que necesitan una entrada continua de agua al sistema. En estos últimos sistemas las máquinas necesitan una cantidad de agua proporcional a la energía que consume la máquina para funcionar, o lo que es igual, proporcional al número y los tipos de chips que componen el *data center* o supercomputador.

Esta cantidad de agua que vuelve a la naturaleza en forma de vapor puede ser significativa y puede afectar enormemente al ecosistema donde está ubicada la instalación. De hecho, hay empresas como Meta, Google y Microsoft que prometen retornar, en el año 2030, al entorno una cantidad de agua superior a la que consumen con proyectos de ayuda al medio ambiente.

¿Cuánta agua potable representa la consumida por estas supermáquinas en relación con el consumo a nivel mundial? Es muy difícil calcular ambas cantidades y, dependiendo de las fuentes, los datos varían muchísimo: una de las informaciones indica que el agua dulce consumida por todo el mundo es aproximadamente de 10 000 a 15 000 millones de metros cúbicos por día; es decir, 5000 billones (europeos) de litros al año. Por usos, diremos que la agricultura utiliza el 70 %, la industria el 20 % y el consumo doméstico el 10 %.

Aunque el consumo de agua en los *data centers* está aumentando, se estima que su valor, admitiendo que en diferentes fuentes los valores discrepan bastante, en el año 2025, está situado entre el 0,015 % y el 0,03 % del consumo mundial. Otras fuentes indican que la cantidad de agua consumida en los *data centers* es de 560 millones de metros cúbicos al año, cifra que se podría duplicar en el año 2030, si seguimos el actual ritmo de

construcción de *data centers*. EE. UU. consume entre 250 y 280 millones de metros cúbicos al año, China 140 y Europa 84.

Al igual que en los supercomputadores se usa una métrica, el Green-500, que da el número de gigaflops que pueden calcular utilizando un vatio de energía eléctrica (recordemos que el valor más alto en los supercomputadores en la lista de noviembre de 2025 es de 73,28), en el caso de gasto de agua existe el concepto denominado *Water Usage Effectiveness* (WUE) o efectividad en el uso del agua. De hecho, mide la cantidad de litros de agua utilizada por kWh de energía consumida. Lo importante es que el valor sea lo más bajo posible. Se calcula que, en promedio, el valor del WUE es de 1,8 litros de agua por cada kilovatio hora de energía consumida. Así pues, con este valor, un computador como el Marenostrum 5 que necesita 12 megavatios para su funcionamiento, consumiría alrededor de 21 000 litros de agua al día, o casi 8 millones de litros de agua al año.

En cuanto a empresas, se estima que Google consumió 24 000 millones de litros (solo el centro que tiene ubicado en Council Bluffs consumió 3 700 millones de litros), y que Microsoft en el año 2022 consumió 6,4 millones de metros cúbicos. Además, hemos de resaltar que, en el año 2023, el promedio de su WUE fue de 0,30 litros por kW/hora y su objetivo es que el consumo de agua sea prácticamente nulo en el 2030. AWS está muy activo en reducir el valor de la WUE en sus centros de datos consiguiendo en el 2023 un valor de 0,18 L/kWh.

El valor de 1,8 para el WUE es el promedio, aunque existen instalaciones con un valor de 0,1 litros, utilizando circuitos de agua cerrados.

Para producir electricidad, también se necesita agua. Y la cantidad de agua depende del mecanismo utilizado para producir la electricidad. Y dentro del agua necesaria aquí considerare-

mos el agua consumida y no la que se utiliza y vuelve al medio ambiente.

Daremos algunos valores típicos de los litros de agua que se necesitan para producir un kilovatio hora (L/kWh). La generación termoeléctrica (carbón/gas/nuclear) da un promedio de 1,8 litros por kWh. La hidroeléctrica consume mucha agua por la evaporación del embalse, dando un promedio de 68 L/kWh, dependiendo del lugar y de la temporada del año. Las energías eólicas y solar no necesitan apenas agua con valores inferiores a 0,1 L/kWh. En el otro extremo están las termoeléctricas (que no se usan ya) con un valor de 57 L/kWh.

14. Uniendo partes: ¿cuánta agua necesita un *data center* en total?

El agua que se necesita globalmente es la suma del agua utilizada para producir la energía eléctrica más el agua utilizada en la refrigeración.

Supongamos un *data center* que usa 20 megavatios de electricidad, que es equivalente a 175,2 GWh/año. El agua consumida para refrigerar el centro dependerá del método usado y podría ir desde los 27 000 metros cúbicos al año (WUE igual a 0,15), hasta los 242 000 metros cúbicos con un WUE igual a 1,8. En cuanto al agua necesaria para producir los 20 GWh/año irían desde valores de 17 500 metros cúbicos al año si se utiliza energía eólica o solar (0,1 L/kWh), 1 millón de metros cúbicos al año si la energía se produce en un entorno medio termoeléctrico (1,25 L/kWh), hasta casi 3 millones de metros cúbicos por año si toda la energía se produjera en un embalse. Tomando los valores mínimos y máximos y sumándolos nos sale que el agua dulce utilizada para generar la electricidad y para refrigerar el

data center está entre los 44 500 metros cúbicos hasta los 3242 metros cúbicos de agua al año. Suponiendo que una persona consume un promedio de 100 litros al día, el agua del *data center* equivale al consumo de entre 1200 y 90 000 personas en un año.

Como se ve, la cantidad de agua total necesaria para que estos centros de datos funcionen no es despreciable.

Como ejemplo, podemos comentar la instalación del *data center* de MareNostrum 5, pensado y diseñado a finales de 2018. Este nuevo *data center* se plantea con una capacidad máxima de 20 megavatios, con 5 transformadores de 4 MW cada uno, imponiendo una redundancia de N+1 en los transformadores. Decidimos refrigeración por agua, DLC, por las siglas en inglés de Direct Liquid Cooling, de forma que un mínimo de un 85 % del calor generado se disipa por DLC, un 10 % por aire en puerta trasera (RDHX, Rear Door Heat Exchanger, en inglés) y el restante 5 % en ambiente, en refrigeración en aire.

El MN5 dispone de los siguientes elementos:

— 16 torres de evaporación en redundancia de N+2, para una capacidad total del 17 MW, con un volumen máximo de 60 m^3, y un flujo máximo de 1500 m^3/h, con impulsión a 28 ºC, y trabajando con delta T de 10º. Este circuito alimenta a los intercambiadores para DLC y a los *chillers* para puertas traseras y *crash*. En este circuito tenemos tratamiento de agua para conductividad y legionela, entre otras.

— 6 intercambiadores, en redundancia N+2 (que ahora con la puesta en marcha de una ampliación para AI, pasaremos a redundancia N+1), con una capacidad de refrigeración de 13,5 MW, con un volumen de agua de 26 m^3, y un flujo máximo de 1170 m^3/h. Esta agua, en circuito cerrado, se trata para garantizar pH, conductividad, etc. Estos inter-

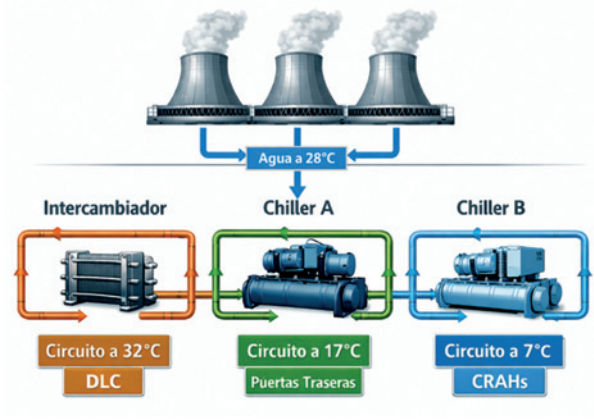


Figura 16. Sistema de refrigeración del MareNostrum 5.

cambiadores proporcionan agua a 32 °C o más, en función de la operativa necesaria.

— 5 *chillers,* de los cuales 4 pueden operar para generar 17 ºC y 3 para generar 7 ºC. En operativa normal, 2 *chillers* operan a 17 ºC y 1 opera a 7 ºC, definiendo una redundancia superior a N+1 en cada caso. El circuito a 17 ºC, para las puertas traseras, tiene un volumen de 12 m^3 y un flujo de 302 m^3/h, mientras que el circuito de 7 ºC son 8 m^3 y 151 m^3/h, respectivamente.

En el mes de marzo de 2025, y con la máquina instalada en carga normal, que varía en cada momento en función de los trabajos que envían los investigadores europeos y españoles, el *data center* de MN5 tuvo un consumo IT de 5318 MWh y un consumo de agua de 7683 m^3, dando un WUE de 1,44 m^3/MWh. Se debe considerar que el sistema de refrigeración de MN5 utiliza parte de refrigeración para otro *data center* de pequeña capacidad y consumo, refrigerado solo en puerta trasera, pero por simplicidad no se deduce del consumo presentado.

15. ¿Es el tema de diseño de chips para la IA estratégico?

Cuando hemos comentado las necesidades de cálculo de los LLM con elevado número de parámetros, hemos visto que el número de operaciones necesarias para entrenar la red superan las 10^{25}. Por otra parte, también hemos comentado al describir el test HPL-MxP que hay algoritmos, entre ellos los de la IA, que pueden ser resueltos con representaciones más pequeñas de los números.

Otro aspecto para comentar es que cada vez más los supercomputadores están ejecutando aplicaciones basadas en la IA. Y este es el caso para los servidores, *laptops,* teléfonos, terminales en la red y futuros coches autónomos. Toda esta situación lleva al hecho de que el diseño de chips que ejecuten eficientemente las aplicaciones de la IA es fundamental ya hoy, y lo será los próximos años (figura 17).

Ya hemos comentado la enorme capacidad de cálculo que necesitó OpenAI para entrenar el modelo ChatGPT y para usarlo posteriormente. Y estos modelos se van a complementar/completar, con muchos más datos (textos, audios, vídeos, señales de sensores, resultados de las simulaciones de los supercomputadores). Y van a entrenarse modelos en todos los centros de supercomputación y en otros muchos laboratorios y empresas. La necesidad de cálculo va a crecer muchísimo. Además, estarán los servidores de las empresas y la enorme cantidad de puntos donde se hará *edge computing* que requerirá un enorme número de procesadores especializados en diferentes aplicaciones de la inteligencia artificial.

Europa ha lanzado la iniciativa IA Factories para facilitar el acceso a *start-ups* y pymes a la IA con apoyo de centros como BSC. El objetivo de las IA Factories es favorecer el desarrollo de la IA en Europa, dando acceso a empresas emergentes (*start-ups* y *spin-off* surgidas de entornos científicos) y también a pymes

Figura 17. La inteligencia artificial está cambiando el tipo de procesadores y aceleradores que construimos.

para que puedan tener recursos de supercomputación gratuitos y a servicios de alto nivel. Las pequeñas y medianas empresas, así como las administraciones, poseen datos y necesitan computadores potentes para que, utilizando las técnicas de la inteligencia artificial, puedan obtener resultados competitivos en los múltiples sectores a los que se dedicarán. También se va a decidir el montaje de 5 gigafactorías, que serán entre diez y cien veces más potentes que las IA Factories, con el objetivo de ser usadas por cualquier empresa, grupos de investigación y administraciones públicas europeas. Son dos grandes iniciativas que son necesarias, pero que no son suficientes.

El desarrollo de nuevas redes neuronales es una necesidad en los centros de supercomputación. Por ejemplo, los centros de supercomputación de Argonne, Riken y el BSC, junto con otros muchos centros, vamos a entrenar y usar modelos con más de un billón (europeo) de parámetros con el objetivo de utilizar estas redes para mejorar nuestras investigaciones en campos

tales como la energía, el cambio climático y la medicina personalizada. Indudablemente, necesitamos reorientar el *hardware* que han de tener los futuros supercomputadores de forma que ejecuten, de manera eficiente, las aplicaciones derivadas de la inteligencia artificial. A esa iniciativa se la denomina Trillion Parameters Consortium (TPC).

¿Y qué influencia tiene todo este tema de la explosión de la inteligencia artificial en el diseño y la fabricación de chips? Muchísima. Para empezar, recordemos que una *foundry* (fábrica) de chips de las últimas tecnologías con transistores de entre 2 y 3 nanómetros cuesta entre 20 000 y 30 000 millones de dólares. Solamente la máquina de litografía, fabricada por la empresa holandesa ASML, cuesta entre 300 y 400 millones de euros, y una *foundry* contiene varias de ellas. Sam Altman, fundador y director de la empresa OpenAI, empezó liderando, a finales de 2024, una cruzada con el objetivo de construir *foundries,* máquinas basadas en la energía atómica, y las mejores ideas en arquitectura de computadores y técnicas de diseño, para construir los millones y millones de chips avanzados orientados a ejecutar aplicaciones derivadas de la inteligencia artificial.

No solamente quieren ser los más aventajados en el diseño del *software,* sino que quieren competir en el mundo del *hardware*. Para ello, empleó tiempo intentado crear una bolsa de dinero de entre 3 y 8 billones (europeos) de dólares. Tengamos en cuenta que esa cantidad es entre 2 y 6 veces el PIB español. En comparación, la deuda corporativa de Estados Unidos en el año 2023 fue de 1,44 billones europeos y que la capitalización de Microsoft y Apple, de manera conjunta, es de 6 billones de dólares. Solo Microsoft ha aumentado su valor bursátil a 3 billones debido al efecto positivo de haber invertido fuertemente en OpenAI. La continuación de esta iniciativa ha sido el proyecto Stargate, anunciado por Trump y comentado anteriormente.

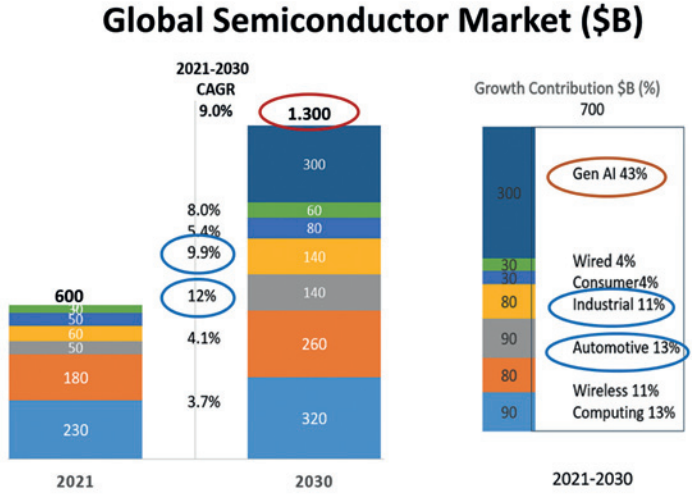


Figura 18. Evolución de los valores de los chips producidos en el 2021 y previsiones para el año 2030.

En el mercado de semiconductores, la aparición de la IA, y sobre todo de la IA generativa, está haciendo que se necesiten más y más chips, y para algunas aplicaciones, chips más y más potentes.

En la figura 18 se ven los valores de los chips que se consumieron por aplicaciones del año 2021, y las predicciones para el año 2030. En esos diez años se van a doblar el valor de los productos que se venderán, llegando a 1,3 billones europeos de dólares (valor muy cercano al PIB de España). En la columna de la derecha se puede apreciar la predicción para el crecimiento de los chips que se usan en la IA generativa que es de un 43 %, o lo que es lo mismo, de 300 000 millones de dólares.

Este hecho ha motivado la reorientación en el diseño de chips de altas prestaciones en las empresas clásicas como Intel, AMD y ARM. Ha ocasionado una verdadera revolución en el diseño de chips.

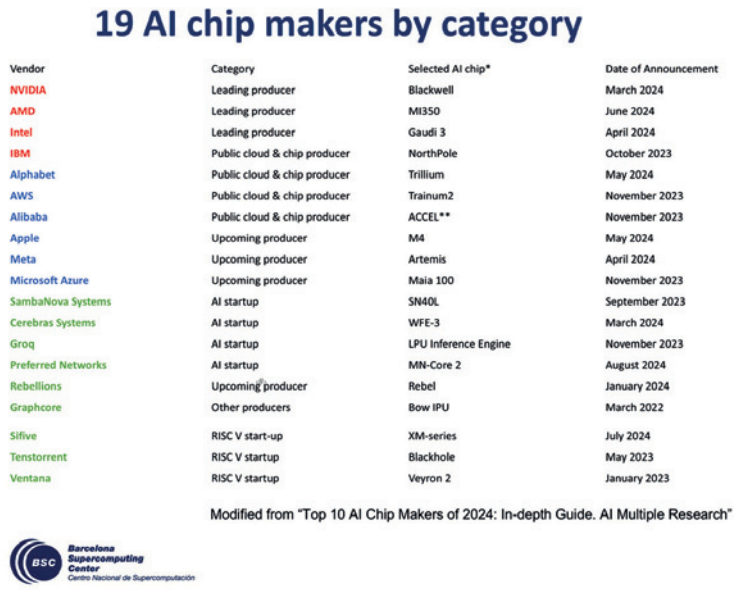


Figura 19. Empresas diseñadoras de chips.

Además de las empresas tradicionales, las proveedoras de servicios, como Alphabet, Amazon, etc., han considerado oportuno desarrollar sus propios chips para no depender tanto de Nvidia, tales como las TPU de Google, para orientar sus diseños a ejecutar de manera más eficiente sus aplicaciones y para obtener precios más bajos. Y, por último, han aparecido muchas otras empresas con el objetivo de disputar a Nvidia parte de la enorme tarta que posee.

Otra manera de ver el mismo tema es ir a la figura 20. En ella se ve que Nvidia posee alrededor del 90 % de los chips que se usan para la IA en general y para la generativa en particular. Por otra parte, se indica que la empresa europea ASML es la única en el mundo que puede proporcionar a las *foundries,* máquinas

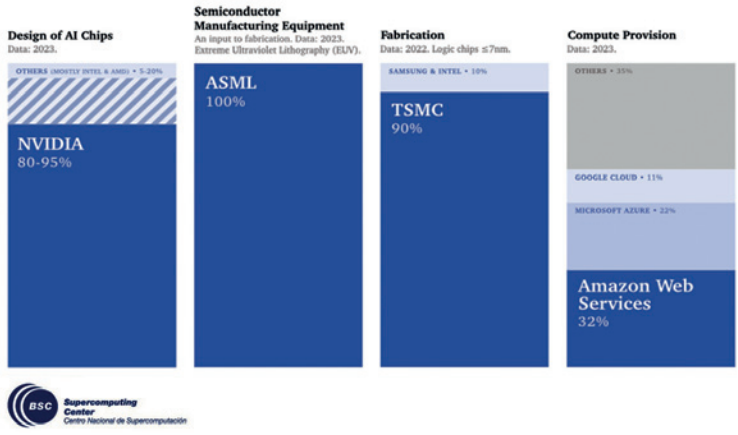


Figura 20. Algunos datos sobre chips, *foundries* y empresas usuarias.

para la litografía para hacer chips, que en la actualidad tienen un tamaño de transistores menor de dos nanómetros. ASML es un cuello de botella en el diseño de chips súper rápidos. En cuanto a las *foundries* (empresas que fabrican chips), se ve que la taiwanesa TSMC fabrica el 90 % de los chips de más alta tecnología y velocidad.

Si vamos a datos más concretos, diremos que TSMC en Taiwán fabrica alrededor de 3 millones de obleas, de diámetro de 300 milímetros, por mes y 0,25 millones de obleas de diámetro de 200 milímetros por mes. Las obleas de 300 milímetros se usan con tecnologías de 7, 5 y 3 nanómetros para diseñar chips de altas prestaciones como los empleados en inteligencia artificial y supercomputadores.

En promedio, la estimación de chips por oblea es la siguiente: en obleas de 300 milímetros, el área útil es de 70 000 milímetros cuadrados que, con chips de 10 centímetros cuadrados (como una CPU moderna), da 700 chips por oblea de los que de

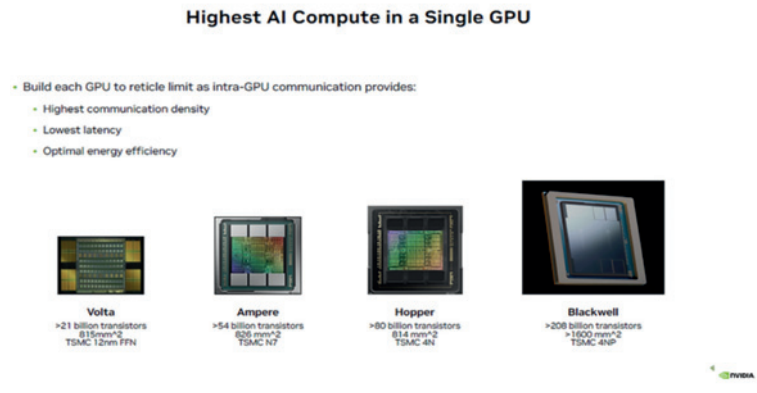


Figura 21. Familia de aceleradores de Nvidia.

500 a 600 son utilizables; en obleas de 200 milímetros, el área útil es de 31 000 milímetros cuadrados y producen 310 chips por oblea. Con tamaño de obleas de 150 milímetros, la superficie útil es de 17 000 milímetros cuadrados y se obtiene 170 chips por oblea.

Como ejemplo, si queremos contabilizar la producción diaria de chips para obleas de 300 milímetros y chips de 100 milímetros cuadrados, tendríamos 2000 obleas por día por 600 chips por oblea, darían una producción de 1,2 millones de chips por día.

La batalla por producir los chips más potentes para ejecutar las aplicaciones más demandantes de la IA es el problema geopolítico más importante que tenemos en la actualidad a nivel técnico. En la figura 21 se ve la evolución de los últimos chips de Nvidia, que es la empresa más potente hoy. En los últimos siete años, Nvidia ha sido capaz de pasar de producir el chip acelerador Volta al nuevo denominado Blackwell.

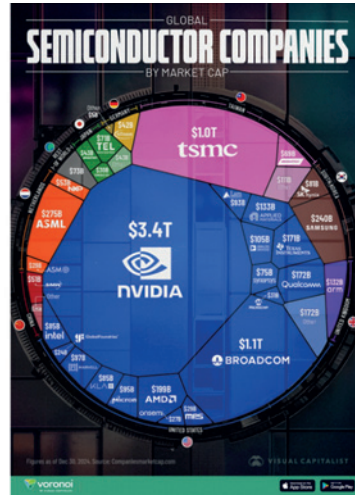


Figura 22. Valorización de empresas fabricantes de chips en la primavera de 2025.

El Volta contenía 21 000 millones de transistores con una tecnología de 12 nanómetros y el Blackwell contiene más de 200 000 millones de transistores con una tecnología de 4 nanómetros y ocupando el doble de espacio. En el intermedio, Nvidia ha producido los chips denominados Ampere y Hopper en tecnologías de 7 y 4 nanómetros. Lo que obtiene esta empresa es increíble.

Antes hemos dado información sobre el negocio que significa el diseño de chips y cómo evolucionará hasta el año 2030. En esta figura 22, de la primavera de 2025, vemos el valor de las empresas dedicadas al diseño de chips.

Sorprendente es que Nvidia fuera el número uno con un valor de 3,4 billones europeos de dólares. Era la segunda empresa mundial con el mayor valor empresarial después de Apple. Vemos que la segunda es Broadcom con un valor de

1,1 billones europeos de dólares y la tercera TSMC con el valor de 1 billón de dólares. Como contrapartida, Intel solo tiene un valor de 85 000 millones de dólares. En septiembre de 2025, Nvidia pasó a ser la empresa más valorada superando los 5 billones europeos de dólares.

16. Consumos de agua y electricidad en las *foundries*

El tema es similar al descrito para los supercomputadores y *data centers*. La diferencia es que las *foundries* necesitan enormes cantidades de agua muy pura para limpiar las obleas donde se fabricarán los chips. Esta agua ha de ser de una pureza extrema para evitar que en la oblea haya cualquier partícula que pudiera dañar la grabación de transistores.

Hemos comentado que la empresa TSMC de Taiwán es la que produce los chips más avanzados del mundo. Vamos a dar datos para chips de tecnología de 3 nanómetros. Aunque no hay datos comprobados, se estima que el consumo de agua de la fábrica de chips de TSMC en Taiwán es de 150 000 metros cúbicos de agua al día, equivalente a casi 50 000 millones de litros para limpiar las obleas. Recicla casi el 90 % del agua que utiliza para limpiar las obleas.

Tengamos en cuenta que la limpieza de cada oblea, con transistores de 5 nanómetros, de 33 centímetros de diámetro, necesita entre 8000 y 10 000 litros de agua, de los que 5500 son de agua ultrapura (UPW, Ultra Pure Water). Esta cantidad de agua equivale al 10 % del consumo industrial para todo Taiwán.

En cuanto al consumo eléctrico, el promedio para la generación de electricidad puede ser de alrededor de 2 litros de agua por kilovatio producido. Se estima que la fábrica necesita 25,5 TWh/año equivalente a casi 3 gigavatios; es decir, tres centrales

nucleares. Es un poco más del 6 % del consumo nacional. Gran parte de esa energía la necesitan los escáneres de Extrem Ultra Violet (EUV). Una fábrica de chips avanzada puede tener más de cuarenta máquinas de este tipo, de forma que cada una consume alrededor de 1,5 megavatios. Es decir, que solo las máquinas que realizan la litografía de los chips pueden consumir 50 megavatios. Y ese número se hace más grande cada vez que utilizamos transistores más pequeños. ¿Cuánta agua se necesita para generar 3 gigavatios de electricidad? Depende de qué se utilice para generar la electricidad. Pero podemos estimar que el agua necesaria será alrededor de 45 millones de metros cúbicos al año. Sumando ambas cantidades, nos da un valor de más de 100 millones de metros cúbicos al año.

El total de agua utilizada serviría para abastecer una ciudad de casi 3 millones de habitantes. Todo lo escrito relacionado con el consumo de agua, diseño de chips y desarrollo de la inteligencia artificial nos permite acuñar frases únicas tales como: «Sin transistores no hay chips», «Sin chips no hay inteligencia artificial», «Sin agua no hay electricidad», «Entrenar un modelo requiere un río de agua», en definitiva, «Sin agua no hay paraíso».

Sumando el agua que necesita TSMC en Taiwán (TSMC tiene otras fábricas fuera de Taiwán), vemos que TSMC gasta alrededor de 250 000 metros cúbicos de agua al día. Por analogía, un *data center* que consumiera 1 gigavatio necesitaría entre 15 000 y 25 000 metros cúbicos al día. Es decir, TSMC consume el agua equivalente a 10-15 centros de datos de un gigavatio, o el equivalente al agua necesaria para los habitantes de Barcelona.

Lógicamente, la cantidad de agua directa (limpiar obleas, fundamentalmente) e indirecta (para producir electricidad) varían en función de la tecnología. Así, para una tecnología de 2 nanómetros se estima que para limpiar cada oblea *(wafer)*

se debe aumentar la cantidad de agua a casi 30 000 litros. Se calcula que si TSMC produce dentro de dos años 150 000 obleas al mes necesitará un valor de 4000 millones de metros cúbicos de agua al mes y casi 50 000 millones de metros cúbicos al año. En cuanto a la electricidad, se estima un 25 % mayor que en tecnología de 3 nanómetros, por lo que necesitaría casi 4 kilovatios hora. Para el número de obleas indicado anteriormente, la energía sería de 6 TWh/año, que requerirían, para su producción, 8 millones de metros cúbicos por año. Sumando los dos datos, da un valor de casi 60 millones de metros cúbicos, que tendríamos que sumar a los casi 100 calculados antes y que corresponden a la suma total de la producción de chips con tecnologías de 7, 5 y 3 nanómetros.

A nivel global, los *data centers* de todo el mundo, consumieron en el 2024 un valor de 400 TWh, aproximadamente 45 gigavatios, equivalente a la electricidad producida por 45 centrales nucleares. Hay tres centros (China Telecom, Range International en China y Switch en Las Vegas que consumen más de 500 megavatios cada uno). Empresas como Alibaba, Google, Microsoft y AWS tienen centros que consumen más de 250 megavatios. EE. UU. y China poseen el 70 % de los *data centers* más grandes. Esto equivale al 1,5 % de la energía total consumida a nivel mundial

17. ¿Cómo es la situación europea en el tema de la IA, diseño de chips y supercomputadores?

Dominar la IA es dominar el mundo. Una parte importante para tener ese dominio es dominar el diseño y la fabricación de chips de alta velocidad. Sin chips de alta velocidad, no hay IA generativa avanzada. La batalla por el liderazgo de la IA hace

que EE. UU. dedique enormes recursos al tema, para evitar que China avance y los alcance. Además, impone trabas en la exportación de los chips más avanzados a China para dificultar el desarrollo de modelos de IA generativa. Europa está en medio y hemos de movernos rápidamente, ya que estamos perdiendo un tren que no volverá a pasar. España debe contribuir dentro de Europa al desarrollo y al uso de la IA y de los supercomputadores.

En Europa, la situación en estos temas no es muy saludable. Diseñamos y fabricamos chips de no muy altas prestaciones, tales como los que se utilizan en los coches actuales. Sin embargo, ni diseñamos ni fabricamos chips de altas prestaciones. No existe una política coordinada, por ejemplo, ni para la creación de talento, ni para repartir las *foundries* entre los países.

Europa (y España en particular) deben invertir en la educación y en la formación de la próxima generación de ingenieros e investigadores que puedan allanar el camino para un ecosistema de semiconductores vibrante. Hoy, Europa tiene algunos de los componentes, como Europractice, pero necesita invertir más y habilitar estas opciones para las pymes y la industria. Recientemente, se han hecho anuncios para expandir el apoyo europeo al desarrollo de semiconductores, pero los mecanismos no están bien definidos. Desafortunadamente, Europa no ha abordado el desafío de la fuerza laboral. Europa debe establecer programas similares para habilitar la fuerza laboral de semiconductores. La Unión Europea puede apoyar los esfuerzos de Francia (CEA), Alemania (HM, Fraunhofer), Italia (Universidad de Bolonia), España (BSC) y Suiza (ETHZ) para poner en marcha los programas.

La ingeniería moderna y las ciencias no se pueden hacer sin supercomputación. La IA ha hecho más fuerte esta tendencia. La supercomputación actual permite simulaciones muy realistas. Ejemplos claros son la aeronáutica o la descarbonización del transporte y la industria.

En el tema de las *foundries,* las diferentes naciones miran este tema a nivel muy egoísta. Por ejemplo, Alemania quiere llevarse todas las *foundries* avanzadas que se puedan construir en Europa en un futuro próximo. Existen programas parciales de R+D a veces desconectados entre ellos. La política y la burocracia no dejan avanzar a Europa en estos temas.

Un producto de todo ello lo constituye el caso de los supercomputadores. Europa posee supercomputadores muy rápidos, como el MareNostrum 5 que inauguramos el 21 de diciembre de 2023 en el BSC. Más en concreto, en la lista del Top-500 de noviembre de 2025, Europa poseía 5 de los 10 supercomputadores más potentes. Pues bien, ni este ni ningún supercomputador europeo tiene procesadores y aceleradores diseñados y/o construidos en Europa. Creo que esta situación no debe continuar. Desde siempre, el BSC ha estado detrás de evangelizar y de motivar a Europa de que no podemos permitirnos esta dependencia tecnológica. Por ejemplo, los futuros coches autónomos basarán su autonomía, entre otras cosas, en la existencia de chips de muy alta velocidad. Europa ni los diseña ni los fabrica, por lo que vamos a ser totalmente vulnerables si no tenemos acceso a ellos. Y de igual forma, si no fabricamos los superchips que necesita la inteligencia artificial, así como su *software* asociado, no podremos influir en que la inteligencia artificial contemple los principios éticos europeos al máximo nivel. Sin la tecnología, Europa solo puede intentar ser «el cuarto árbitro de partidos en los que sabe que no tiene equipo para jugar». Poco a poco, desde el BSC hemos influido en la Unión Europea para cambiar esta situación.

Reducir la dependencia tecnológica de Europa frente a las compañías norteamericanas y asiáticas, que dominan este mercado, es uno de los objetivos estratégicos de la Comisión Europea, que a través de la iniciativa Chip Act europea y otras, busca una buena posición europea en el diseño y en la produc-

ción de todo tipo de chips. Europa dedicará 43 000 millones de euros a esta actividad que intentará potenciar la creación de fábricas con tecnologías avanzadas, y el diseño de los chips que necesitamos. En cuanto a casos concretos, TSMC va a instalar en Alemania una fábrica en Dresden, orientada a la producción de chips para coches, entre otras aplicaciones. Es importante que empresas como Infineon, NXP y Bosch hayan invertido cada una de ellas un 10 % en el proyecto y, sobre todo, que hayan decidido fabricar los futuros chips para coches en ella. Por otra parte, se intenta que Intel construya otra fábrica con la última tecnología, también en Alemania, y que mejore las condiciones de la que tiene en Irlanda. Francia quiere tener una fábrica de Global Foundries con tecnología lo más avanzada posible

Los chips dominan el mundo, por lo que, si Europa confía en tecnología extranjera para nuestras infraestructuras más trascendentales, puede ser un problema desde un punto de vista geopolítico. Y no se trata solo de tener tecnología europea, sino también de garantizar que esta sea competitiva en el mundo. Europa tiene que garantizar su soberanía tecnológica y, para ello, debe fabricar sus propios procesadores basados en *hardware* abierto. Esto, sin ninguna duda, va a mejorar nuestra competitividad y rendimiento económico, pero, además, nos va a permitir crear aplicaciones que se ajusten a los principios éticos, legales, socioeconómicos y culturales europeos.

Y ahora se dan unas condiciones únicas para intentar reducir esa dependencia tecnológica. Por una parte, estamos acercándonos al límite del tamaño mínimo de los transistores de silicio, de forma que todo el mundo podrá tener acceso a la tecnología más avanzada. Ahora entramos en una cierta estabilización tecnológica, solo que algunas empresas como Apple y Nvidia tienen prioridad, debido a sus grandes volúmenes de consumo. Por ejemplo, se sabe que Nvidia y Apple han obtenido de TSMC la prioridad

para que, en el primer año y medio, sean casi los únicos clientes de chips que usen los tamaños de 3 y 2 nanómetros. Este hecho les da una gran ventaja competitiva para fabricar productos que otros no podrán tener. En segundo lugar, y más importante, es que existe una iniciativa mundial, denominada RISC V, que define un juego común de instrucciones del lenguaje máquina de los procesadores, de forma que nadie es propietario, y todo el mundo puede fabricar procesadores que usen estas instrucciones. Este hecho está cambiando el ecosistema en el diseño de procesadores. Hasta ahora, las empresas como Intel, Nvidia, ARM o IBM son las propietarias de los juegos de instrucciones del lenguaje máquina de sus procesadores, y esto, prácticamente, hacía imposible el desarrollo de procesadores con juegos de instrucciones nuevos, ya que no hay *software* desarrollado para ellos. Era misión imposible. Europa tuvo en los procesadores ARM su única oportunidad hasta que dicha empresa fue adquirida por Softbank[8] (Japón).

El RISC V es equivalente para el desarrollo del *hardware* a lo que Linux fue para el desarrollo de *software*. Linux democratizó el *software* haciendo que todo el mundo pudiera hacer programas que se puedan ejecutar en cualquier plataforma *hardware*. RISC V está haciendo lo mismo en *hardware*. Multitud de empresas y centros de investigación están desarrollando *hardware* y *software* basado en RISC V y Linux. Se abre una posibilidad jamás imaginada, que va a permitir la existencia de un mundo abierto en el que la colaboración sea lo más natural. En pocos años, la inmensa mayoría de los procesadores serán RISC V. Con estas dos condiciones anteriores, aparecerán innumerables nuevas ideas que podrán ser llevadas a la práctica abriendo un campo de innovación nunca imaginado. La idea RISC V permite

8 <https://www.softbank.jp/en/corp/aboutus/message/>.

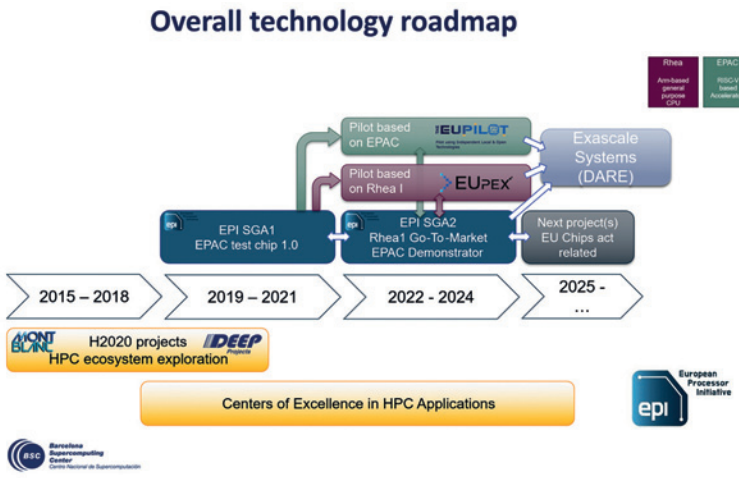


Figura 23. Proyectos europeos relacionados con el RISC V.

diseñar procesadores orientados a las aplicaciones, y esta será una ventaja adicional. La tercera razón es que Europa de manera global, y algunos países como España a nivel particular, han decidido dedicar importantes recursos para intentar obtener esta autonomía tecnológica. La Unión Europea y España estamos ante un tren que debemos tomar; será la última vez que pase.

Por todo ello, Europa ha financiado algunos proyectos de procesadores y aceleradores basados en RISC V (figura 23). Empezó con el proyecto European Processor Initiative (EPI), donde se desarrollaron tres sistemas de procesadores y aceleradores que fueron integrados y enviados para realizar el chip físico. A partir de ahí, se han financiado otros proyectos como el EU-Pilot y el E-Processor. Todos estos proyectos han sido liderados por el BSC. En la actualidad, estamos esperando comenzar el proyecto DARE. Está coordinado por el BSC y tiene un presupuesto de 260 millones de euros, de los que la mitad están financiados por la Unión Europea.

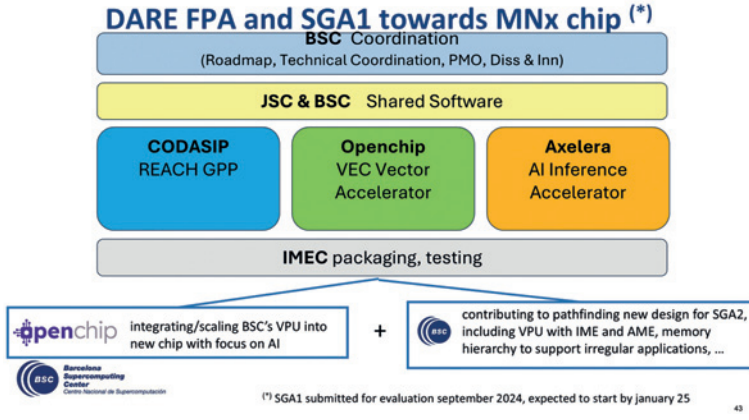


Figura 24. Proyecto europeo DARE.

En dicho proyecto se van a desarrollar tres chips de muy alta velocidad. Uno de ellos será desarrollado por la empresa española Openchip que ha recibido una fuerte financiación del PERTE chip. Y la idea dentro del BSC es que ese chip sea la antesala de la siguiente generación, que será utilizada en el futuro MareNostrum 6.

Figura 25. Inauguración del MareNostrum 5.

Si lo hacemos posible, aunque el reto es muy grande, habremos contribuido a que Europa sea un poco más autónoma en el desarrollo de esos chips que ejecutan las aplicaciones avanzadas de la inteligencia artificial y que han conquistado el mundo.

18. A modo de despedida

Comentar sobre los temas anteriores implica aceptar que cada día se debería reescribir parte del documento. Las noticias vuelan. Estamos en un momento cámbrico de investigaciones reales, propuestas de futuras investigaciones y competencia entre empresas y países. Está muy aceptado decir que los países que gobiernen el cómputo, los datos y la inteligencia artificial dominarán el mundo. Y que los países que no tengan esos instrumentos serán esclavos de los anteriores.

Por ejemplo, hay una batalla encarnizada en producir y lanzar al mercado los mejores modelos de lenguaje (LLM). Cuando terminé de escribir la versión anterior el 3 de enero de 2026, parecía ser que en el top estaban ChatGPT-4o, de OpenAI, Gemini 2.0 Ultra, de Google, y Claude 3.5 Sonnet. Entre los de código abierto, Llama 3.1 405B, de Meta, lideraba por su escalabilidad y rendimiento general. Otros le seguían la estela como el chino Deep Seek. Pero la primera posición va cambiando cada muy poco tiempo. Hay una pelea atroz para conquistar el mercado.

Hoy, 3 de abril de 2026, existen otros modelos compitiendo. OpenAI tiene el GPT-5 y el GPT-5X, Anthropic el Claude Sonnet 4.X y el Claude Sonnet Opus 4.X, Google el Gemini 3, Meta el LLaMA-4, Deep-see el V3, xAI el Grok, Alibaba el Quen y Microsoft el Phi. Se intenta llegar a una inteligencia cercana a la humana extendiendo el número de parámetros de los mode-

los y los datos usados para entrenarlos. Parece ser que dentro de muy poco se producirá una gran burbuja tecnológica en estos temas debido a su incapacidad para llegar a ese tipo de inteligencia. Los investigadores más brillantes en el campo indican que es necesario que los modelos aprendan de la vida real tal como hacemos los humanos. Por los ojos de cualquier niño de cuatro años ha pasado más información que toda la usada para entrenar los modelos más avanzados y el niño, con un cerebro que consume unos pocos vatios y ocupa un pequeño volumen (comparados con los monstruosos centros de datos), es capaz de reconocer y generar conocimiento que el modelo no puede.

Hay otra batalla para construir los *data centers* más grandes del mundo. En esos centros de datos se concentran los modelos sobre los que millones de usuarios hacen preguntas, hacen lo que se denominan *inferencias*. Y es ahí donde está el negocio. Se habla de construir centros de datos que contengan más de 1 millón de chips aceleradores de la IA (tales como el Blackwell de Nvidia). Tengamos en cuenta que los supercomputadores más rápidos del mundo pueden tener alrededor de 30 000 aceleradores, por lo que son 30 veces más lentos que los *data centers* más grandes. Estos *data centers* pueden necesitar varios centenares de megavatios para alimentar a sus chips y refrigerarlos. Estamos ahora hablando ya de centros que van a requerir más de un gigavatio de potencia eléctrica.

En enero de 2025, Mark Zuckerberg anunció que Meta planeaba terminar 2025 con al menos 1,3 millones de GPU en servicio. El centro de datos Stargate de OpenAI planea usar más de 450 000 GPU Nvidia Blackwell GB200, y el Colossus 2 de xAI, una expansión de Colossus, está construido para acomodar a más de 550 000 GPU.

El principal de ellos es el centro de datos de 5 gigavatios planeado por Meta en Louisiana, llamado Hyperion, anun-

ciado en junio de 2025. El CEO de Meta, Mark Zuckerberg, dijo que Hyperion «cubrirá una parte significativa de la huella de Manhattan» y que la primera fase, una versión de 2-GW, se completará para 2030. Los centros de datos de IA modernos a menudo utilizan sistemas a escala de *rack,* como el Nvidia GB200 NVL72, que ocupan un solo *rack*. Cada *rack* contiene 72 GPU, 36 CPU y hasta 13,4 *terabytes* de memoria GPU. Además del rendimiento de la computación en bruto, los *racks* Nvidia GB200 NVL72 también requieren grandes cantidades de memoria. Un *rack* Nvidia GB200 NVL72 puede incluir hasta 13,4 *terabytes* de memoria de alto ancho de banda, lo que implica que un campus de centro de datos a escala de Hyperion requerirá al menos varias docenas de petabytes. La inmensa demanda ha hecho que los precios de la memoria se disparen: el precio de la DRAM, específicamente DDR5, ha aumentado un 172 % en 2025.

Los bastidores miden más de 2,2 metros de altura y pesan más de una tonelada y media, obligando a los centros de datos de IA a usar hormigón más grueso con más refuerzo para soportar la carga. Un solo *rack* GB200 puede utilizar hasta 120 kilovatios. Si Hyperion cumple con sus objetivos de 5 gigavatios, el campus del centro de datos podría incluir más de 41 000 *racks,* para un total de más de 3 millones de GPU. Es probable que el número final de GPU utilizadas por Hyperion sea menor porque las futuras GPU serán más grandes, más capaces y usarán más energía. Según ConstructConnect, el gasto en centros de datos se aproximó a los 27 000 millones de dólares de Estados Unidos, hasta julio de 2025 y, según los últimos datos, se acercará a los 60 000 millones hasta el final del año. El proyecto Hyperion de Meta es una gran parte del pastel, con un coste superior a los 10 000 millones de dólares.

Al igual que existe una batalla entre compañías desarrollando LLM, la mayoría de ellas americanas, existe una batalla para el desarrollo de chips que contengan el tamaño más pequeño de los transistores. Estamos en 3 nanómetros y hay prototipos de transistores de menos de 2 nanómetros. Los americanos prohibieron a Nvidia vender a China sus chips más avanzados como el Blackwell. Con ello frenaban el avance chino en desarrollar modelos avanzados de LLM, aunque Deep Seek demostró lo contrario. Hace dos semanas, los americanos quitaron ese veto a Nvidia. Las cosas pasan sin ninguna explicación razonable. Por otra parte, los americanos prohibieron a la empresa europea ASML vender máquinas de litografía de la última generación para que China no pudiera producir chips con el tamaño de transistores muy pequeños, al igual que lo hace la empresa TSMC de Taiwán, de muy altas prestaciones que compitieran con los americanos de AMD y Nvidia. Sin embargo, hace menos de una semana a principios de diciembre de 2025, se anunció que China será capaz de tener esa tecnología de ASML en menos de cinco años. Increíble, pero cierto. China va a conquistar el mercado de los chips y de los LLM comerciales en muy poco tiempo.

Nvidia se caracteriza por sacar al mercado nuevos productos cada dos años. En el texto anterior, nos habíamos quedado en el acelerador Blackwell y la CPU Grace (figura 21). Pues bien, ya tiene en el mercado la continuación de ambos que se denomina Vera Rubin.

En la figura 26 se puede ver el conjunto de seis chips con los que Nvidia puede construir supercomputadores dedicados a la inteligencia artificial de amplia gama. Describiremos brevemente el acelerador Rubin. Contiene 336 000 millones de transistores, es decir, 1,6 veces más que el Blackwell, y puede realizar 50 petaflops de operaciones con números en formato NVFP4

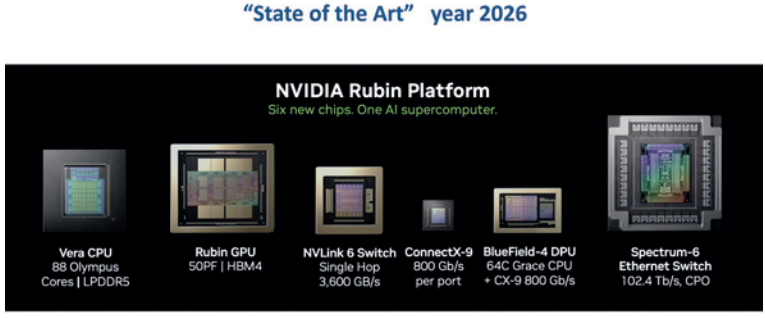


Figura 26. Conjunto de seis chips de Nvidia entre los que se encuentran el Vera y el Rubin.

entrenando modelos, por lo que es 5 veces más rápido que el Blackwell.

Al igual que Nvidia, las empresas que usan gran número de chips aceleradores desarrollan sus propios chips cada vez más y más potentes. En el caso de Google, después de la TPU V6 (Trillium) ha desarrollado la TPU V7 (Ironwood) que es más de 3 veces más rápida que la Trillium. En el caso de Amazon, después de la Trainium 3 ha desarrollado la Trainium 4. Sus intereses consisten en poseer chips más baratos que los proporcionados por Nvidia y capaces de ejecutar sus aplicaciones con una eficiencia igual o mejor.

En este documento hemos hablado del proyecto Trillion of Parameters Consortium (TPC), como alianza entre algunos centros de supercomputación del mundo para desarrollar modelos fundacionales que ayudarán a hacer una mejor investigación en los centros de supercomputación. Pues bien, esa idea de centros de supercomputación a nivel mundial (hay más de ochenta *partners*), ha sido retomada y amplificada por el Gobierno americano para lanzar la iniciativa denominada Génesis.

Figura 27. Proyecto Génesis.

El objetivo de Génesis es utilizar los supercomputadores (los 17 centros de supercomputación del Department of Energy [DOE] se coordinarán), la inteligencia artificial y las empresas y centros de investigación americanos para producir la mejor ciencia e ingeniería del mundo. Quieren que América sea líder indiscutible de la generación de esas ideas disruptivas en campos como la fusión, la computación cuántica, energía, seguridad nacional, medicina personalizada, cambio climático, nuevos materiales... Quieren que el resto del mundo adopte y compre sus tecnologías y pretenden que el resto del mundo seamos incapaces de competir contra ellos. Y visto esto... *¿Quo vadis,* Europa?

Barcelona, 3 de abril de 2026

El presente discurso
fue leído el 5 de mayo de 2026,
157 años después de que Santiago Ramón y Cajal
se incorporara a las aulas
de la Universidad de Zaragoza

LECCIONESCAJAL

1 Las nuevas neurotecnologías y su impacto en la ciencia, medicina y sociedad // **Rafael Yuste**

2 La oncología en el siglo XXI: de las terapias personalizadas a la inmunoterapia // **Mariano Barbacid**

3 De Cajal, los embriones y el cáncer // **Ángela Nieto Toledano**

4 Aplicación de la nanotecnología al desarrollo de nuevas vacunas y terapias personalizadas // **María José Alonso**

5 Del diagnóstico microscópico a la patología digital, molecular y algorítmica // **Santiago Ramón y Cajal Agüeras**

6 Pasado, presente y futuro de las vacunas: El reto de vivir en sociedad // **Carlos Martín Montañés**

7 Supercomputación, inteligencia artificial, chips y autonomía europea // **Mateo Valero**

MAGNA

Vicerrectorado de
Cultura y Patrimonio
Universidad Zaragoza