Universitat d'Alacant
Universidad de Alicante

Retos actuales en Computación de Alto Rendimiento (CAR)

Contenido

1. Introducción al escenario actual

¿Qué es la Computación de Alto Rendimiento (CAR) y por qué es relevante?

La *Computación de Alto Rendimiento* (High-Performance Computing, HPC) consiste en usar superordenadores y procesamiento en paralelo para resolver problemas muy complejos, imposibles de afrontar en equipos convencionales. Sus **tres ideas clave** son: **Rendimiento extremo** (múltiples CPUs/GPUs en paralelo para minimizar tiempos), **Escalabilidad** (mantener o mejorar rendimiento al añadir más recursos) y **Optimización** integral de hardware y software.

- *Rendimiento extremo:* miles de núcleos trabajan al unísono para lograr velocidades de cálculo gigantescas.
- *Escalabilidad:* el sistema sigue siendo eficiente al crecer en número de nodos o procesadores.
- *Optimización:* aprovechar al máximo la infraestructura con software y hardware especializados.

Relevancia actual de la CAR: La HPC sigue siendo vital porque permite afrontar necesidades emergentes:

- **Simulaciones científicas complejas:** clima global, dinámica molecular, física de partículas, etc. requieren cada vez más cómputo. Por ejemplo, modelos meteorológicos o de pandemia que antes eran inviables, hoy corren gracias a HPC.
- **Inteligencia Artificial masiva:** el auge de la IA generativa (entrenar modelos como GPT, DALL·E) exige infraestructuras de cómputo a escala planetaria. Entrenar un modelo gigante puede requerir miles de GPUs funcionando durante semanas. (Se estima que entrenar GPT-4 implicó decenas de miles de GPU por semanas, con un coste energético y económico enorme.
- **Importancia estratégica:** la CAR es una herramienta clave en defensa, salud, energías renovables y desarrollo tecnológico. Muchos avances médicos (ej. diseño de fármacos, análisis genómico), predicciones climáticas o innovaciones industriales dependen de supercomputadores.

De los superordenadores clásicos a la IA generativa: La HPC ha evolucionado constantemente:

- *Años 1960-1980:* superordenadores vectoriales (ej. **Cray-1**) para ciencia y uso militar.
- *Años 1990-2000:* arquitecturas masivamente paralelas (clústeres de cientos de nodos interconectados).
- *2010-presente:* sistemas híbridos heterogéneos (combina CPU, GPU, aceleradores como FPGAs, ASICs). Los supercomputadores actuales integran GPUs especializadas para IA (Tensor Cores, TPUs de Google, etc.).
- *Actualidad:* convergencia HPC + IA. Los modelos generativos (GPT, LLaMA, Google Gemini…) se entrenan en **HPC clusters** con miles de GPUs en paralelo. La frontera entre HPC tradicional y computación de IA se difumina: se habla ya

de *AI-HPC*. Ejemplo: Entrenar **GPT-3** requirió miles de GPU en paralelo durante semanas, y **GPT-4** habría multiplicado ese esfuerzo.

El papel de la CAR en la sociedad moderna: La HPC impacta transversalmente múltiples sectores. Por ejemplo:

Sector	Aplicaciones habilitadas por HPC
Sanidad	Simulación de proteínas, diseño de fármacos, análisis genómico
Clima y medioambiente	Predicción meteorológica, modelado climático global
Industria y energía	Optimización de procesos industriales, simulación de fluidos, energía nuclear
Finanzas	Análisis de riesgo, simulaciones Monte Carlo, detección de fraudes
IA y Big Data	Entrenamiento distribuido de modelos de IA, análisis masivo de datos

Un ejemplo reciente: durante la **pandemia COVID-19**, los supercomputadores aceleraron simulaciones de proteínas del virus y modelos de contagio a gran escala, apoyando el desarrollo de vacunas y políticas sanitarias.

Retos actuales de la CAR: A pesar de sus logros, la HPC enfrenta hoy varios desafíos cruciales:

- **Eficiencia energética:** Los superordenadores consumen cantidades enormes de electricidad. *(Ejemplo:* Frontier, primer sistema exaescala, consume ~20-21 MW constantes spectrum.ieee.org, lo equivalente al consumo de una ciudad pequeña.) Lograr la "exaescala" (10^18 operaciones/segundo) de forma sostenible es un reto abierto.
- **Escalabilidad del software:** Muchos programas no escalan bien a decenas de miles de núcleos. Se requiere rediseñar algoritmos pensando en paralelismo masivo y asincronía.
- **Gestión de datos masivos:** Los volúmenes de datos generados y procesados por HPC son gigantes. Mover datos se vuelve un cuello de botella; hacen falta sistemas de almacenamiento paralelos y E/S ultrarrápida (ej. sistemas de ficheros como Lustre, GPFS).
- **Heterogeneidad de hardware:** GPUs, TPUs, CPUs ARM, chips especializados… El software debe ser portátil a múltiples arquitecturas (usando estándares como OpenMP, MPI, CUDA, SYCL).
- **Formación y talento:** Escasez de profesionales con conocimientos en computación paralela, optimización de código HPC e IA a gran escala. Es vital fomentar programas educativos que integren HPC + IA desde etapas tempranas.
- **Accesibilidad e infraestructura:** No todos los países o centros académicos tienen acceso a supercomputadores. La nube HPC (AWS, Azure, Google TPU

pods) es una solución parcial, pero plantea desafíos de costo, ética y soberanía tecnológica.

Conclusión del bloque 1: La HPC no ha perdido vigencia; al contrario, está en un momento crítico de transformación. El auge de la IA, el Big Data, la necesidad de sostenibilidad y la convergencia de nuevas tecnologías están definiendo **nuevos horizontes** para la Computación de Alto Rendimiento. Los retos identificados orientarán los contenidos de las siguientes secciones.

Preguntas de reflexión (Introducción): **¿Cuál de los retos actuales de la CAR te parece más urgente y por qué?** ¿Crees que la sociedad en general es consciente de la importancia de la HPC en su vida cotidiana?

2. Escalabilidad de los sistemas HPC

Concepto de escalabilidad: En HPC, la *escalabilidad* mide la capacidad de un sistema para **mejorar su rendimiento al aumentar los recursos** (más procesadores, más nodos) sin perder eficiencia. Hay dos formas clásicas de evaluar la escalabilidad de una aplicación:

- **Escalabilidad fuerte (Strong Scaling):** Fijamos el tamaño del problema y aumentamos el número de procesadores, viendo cómo disminuye el tiempo de ejecución. Idealmente, duplicar procesadores debería **reducir a la mitad** el tiempo. En la práctica, hay sobrecarga de comunicación y sincronización, por lo que nunca se logra un speedup perfecto. *(Ejemplo:* Simulación meteorológica de 1 semana que tarda 10 horas con 128 procesadores: ¿podemos hacerla en 5 horas con 256? En teoría sí, en la práctica las comunicaciones impiden llegar exactamente a 5h.)
- **Escalabilidad débil (Weak Scaling):** Aumentamos en la misma proporción el tamaño del problema y el número de procesadores, comprobando si el **tiempo de ejecución se mantiene constante**. La meta es que más datos + más procesadores = mismo tiempo. En la realidad, suele haber degradación porque la comunicación entre nodos crece y añade demora. *(Ejemplo:* Simular 100 km² en 10 horas con 100 procesadores vs. 200 km² en ~10 horas con 200 procesadores.)

Problemas técnicos que afectan a la escalabilidad: Alcanzar una escalabilidad casi lineal es difícil por varios factores:

- *Latencia de comunicación:* Aunque usemos interconexiones ultra-rápidas (InfiniBand, NVLink), siempre hay un retardo mínimo en enviar mensajes entre nodos. En un clúster gigante, esa latencia acumulada limita el speedup.
- *Sincronización:* Muchas aplicaciones requieren sincronizar periódicamente todos sus procesos (p. ej. barreras MPI). Esto provoca que los procesos más rápidos deban esperar a los más lentos, generando ineficiencias (*efecto "straggler"*).
- *Cuellos de botella de E/S o memoria:* Si todos los nodos acceden al disco simultáneamente, la entrada/salida se vuelve un freno. De igual forma, en

arquitecturas CPU+GPU, las transferencias de datos entre memoria principal y GPU pueden saturar el rendimiento si no se gestionan bien.

- *Sobrecarga de gestión:* A medida que crece el número de nodos, aumenta la complejidad de orquestar tareas, asignar recursos y balancear cargas. Surgen problemas de *scheduling* (planificación de trabajos), mayor overhead del sistema operativo, fallos de componentes que hay que tolerar, etc.

Ley de Amdahl: límite teórico del paralelismo: La *Ley de Amdahl* formula el techo de aceleración (*speedup*) posible al paralelizar una tarea. Si una fracción `f` del código es paralelizable y el resto `(1-f)` es secuencial, entonces con p procesadores el speedup máximo `S(p)` viene dado por:

$$S(p) = \frac{1}{(1-f)+\frac{f}{p}}.$$

- **Interpretación:** Por muchos procesadores que tengamos, cualquier porción secuencial `(1-f)` limita el speedup. Con **infinitos** procesadores, el speedup tiende a `1/(1-f)` como máximo. Por ejemplo, si un programa tiene $f = 0.99$ (99% paralelo, 1% secuencial), el speedup máximo teórico sería ~100×, aunque pongamos miles de CPUs. Esto significa que más allá de cierto punto, añadir recursos *no* mejora el rendimiento.
- En la práctica, las aplicaciones reales suelen tener entre un 1% y 10% de código secuencial, lo que impone un límite significativo incluso en sistemas exaescala.
- **Ley de Gustafson:** (complementaria a Amdahl) plantea un escenario más optimista: si aumentamos el tamaño del problema junto con los procesadores (weak scaling), la fracción paralelizable f puede crecer, aliviando el límite. En otras palabras, podemos aprovechar más CPUs haciendo problemas proporcionalmente más grandes, obteniendo speedups mayores de lo que predice Amdahl para un problema fijo.

Ejemplo representativo – Supercomputador Frontier (EE.UU.): Frontier es el primer superordenador de nivel *exaescala* (más de 1 exaFLOP) y actualmente uno de los más potentes del mundo (Oak Ridge National Lab, EUA, 2022). Sus características ilustran avances en escalabilidad:

- +**1 exaflop** de rendimiento sostenido Linpack (10^18 operaciones/segundo).
- **Arquitectura:** ~9,400 nodos HPE Cray EX interconectados, cada nodo con 1 CPU AMD EPYC de 64 núcleos y 4 GPUs AMD Instinct MI250X, sumando más de 8.7 millones de núcleos en totalspectrum.ieee.org.
- **Red de interconexión Slingshot** de HPE, de baja latencia, para minimizar costos de comunicación entre miles de nodos.
- **Consumo energético:** ~21 MW en operaciónspectrum.ieee.org – muestra el desafío energético de la exaescala, pero también su eficiencia (Frontier ocupa el #1 del ranking Green500 de supercomputadores más eficientes energéticamentespectrum.ieee.orgspectrum.ieee.org).
- **Refrigeración:** por agua directa a los componentes, dada la altísima densidad de calor, más eficiente que aire.

- **Escalabilidad de Frontier:** Está diseñado para mantener >50% de eficiencia paralela incluso en trabajos exaescala completos. Logra esto mediante algoritmos optimizados para GPUs que reducen las necesidades de sincronización global. Frontier muestra que, combinando *ingeniería de hardware*, *software especializado* y *arquitectura heterogénea*, es posible acercarse a la escalabilidad casi ideal a gran escala

Conclusión del bloque 2: La escalabilidad es un reto estructural en HPC. Aunque el hardware siga mejorando (más núcleos, redes más rápidas), **el software y los algoritmos** deben evolucionar para aprovecharlo. Las limitaciones físicas (latencia, sincronizaciones, partes secuenciales) marcan un techo práctico. Proyectos punteros como Frontier indican que con la co-innovación en hardware y software podemos empujar ese techo, pero entender leyes como la de Amdahl nos recuerda que siempre habrá un límite que vencer o esquivar (por ejemplo, replanteando el problema para aprovechar la Ley de Gustafson).

Preguntas de reflexión (Escalabilidad): **¿Qué estrategias** (técnicas o algorítmicas) **se te ocurren para mitigar los efectos de la Ley de Amdahl en un programa paralelo?** En tu experiencia (ej., trabajos en equipo), ¿has visto situaciones donde "más recursos" dejaron de mejorar el resultado? ¿Qué analogía cotidiana harías con los problemas de escalabilidad en HPC?

3. Sostenibilidad y eficiencia energética

El desafío energético en HPC: Los centros de datos y supercomputadores consumen enormes cantidades de energía eléctrica, gran parte de la cual se disipa en forma de calor. Mejorar la eficiencia energética es un objetivo crítico por motivos tanto económicos (coste de electricidad) como ambientales (huella de carbono). En esta sección exploramos métricas, retos y soluciones para lograr una HPC sostenible.

¿Qué es el PUE (Power Usage Effectiveness)? El *PUE* (Eficiencia en el Uso de Energía) es la métrica estándar para evaluar la eficiencia energética de un centro de datos. Se define como el cociente entre la **energía total** consumida por el centro de datos (incluyendo climatización, iluminación, etc.) y la **energía consumida por el equipamiento IT** (servidores, almacenamiento, networking):

$$PUE = \frac{\text{Energía total centro datos}}{\text{Energía usada por equipos IT}}.$$

- Un **PUE ideal** sería 1.0, indicando que **toda** la energía va a computación pura (nada a extras). En la práctica nunca es 1; la diferencia (PUE – 1) refleja la sobrecarga de infraestructura.
- Ejemplos: En 2010 el PUE medio de la industria era ~1.7 (por cada 1 kW en computadoras, 0.7 kW adicionales a enfriamiento y otros. Las compañías líderes han bajado ese valor: Google, por ejemplo, reporta PUE ~1.1 en sus centros de datos más eficientes, acercándose bastante al ideal. Esto se logra con diseños innovadores de cooling, distribución eléctrica optimizada, uso de IA para gestión térmica, etc.

Retos de refrigeración y consumo eléctrico: La **refrigeración** suele representar ~40% del consumo energético de un centro de datos típico. A medida que aumenta la **densidad de potencia** (equipos más potentes por metro cuadrado, especialmente para cargas de IA), los retos térmicos se agravan. Las soluciones tradicionales (aire acondicionado y ventilación) están llegando a su límite. Nuevas tecnologías en despliegue para enfrentar este reto incluyen:

- **Refrigeración por inmersión:** Sumergir los servidores en un líquido dieléctrico especial que absorbe el calor más eficientemente que el aire. Esto puede reducir el consumo de energía para enfriamiento hasta un 30-50% y permite empaquetar hardware más densamente. Grandes operadores experimentan con ello, aunque con desafíos en mantenimiento y fiabilidad a largo plazo (ej. corrosión, acceso físico al hardware).
- **Refrigeración líquida directa:** Circular líquido (agua o refrigerantes) directamente por bloques fríos acoplados a CPUs/GPUs. Absorbe el calor en la fuente y lo lleva fuera del rack. Mejora la eficiencia térmica y disminuye la necesidad de climatizar toda la sala. Requiere infraestructura de tuberías y controles, pero cada vez es más común en supercomputadores actuales.
- **"Free cooling":** Aprovechar el aire exterior frío (en climas templados o fríos) para enfriar el centro de datos, reduciendo el uso de chillers mecánicos. Por ejemplo, en Escandinavia y otros lugares se instalan data centers que casi todo el año usan aire exterior o agua de mar/lagos para refrigeración.

Casos reales – Mejores prácticas:

- *Meta en Dinamarca:* Meta (Facebook) construyó centros de datos en Odense, Dinamarca, aprovechando el clima frío y 100% alimentados con energía eólica local. Usan enfriamiento por aire exterior gran parte del año y sistemas altamente eficientes. Además, reutilizan el calor residual: el calor de los servidores se captura para alimentar la red de calefacción distrital, proporcionando agua caliente a miles de hogaresramboll.comramboll.com. Es un ejemplo de simbiosis energética donde un data center se integra sosteniblemente en su entorno.
- *IA de Google para refrigeración:* Google aplicó algoritmos de *machine learning* (DeepMind) para optimizar en tiempo real la climatización de sus data centers. El sistema de IA ajusta dinámicamente ventiladores, enfriadores y flujos de aire mejor que la configuración manual, logrando reducciones de hasta un 30% en la energía de refrigeración. Este enfoque de "centro de datos autónomo" demuestra cómo la IA puede ayudar a hacer más verde a la propia IA.

Iniciativa innovadora – Project Natick (Microsoft):*Project Natick* es un proyecto experimental de Microsoft que explora centros de datos **submarinos** como alternativa eficiente. En 2018 se sumergió un contenedor sellado con servidores (cuya apariencia recuerda a un cilindro metálico) en las Islas Orcadas (Escocia), a ~35 metros bajo el mar. Estuvo 2 años operando bajo el agua. ¿Resultados? Al recuperarlo en 2020, se encontró que tuvo *1/8 de la tasa de fallos* de un centro de datos terrestre y se enfrió naturalmente con el agua fría del mar. Esto sugiere que los océanos podrían proporcionar refrigeración gratuita y un ambiente estable. Además, los módulos submarinos pueden ser alimentados por energías renovables marinas (olas, mareas).

Aunque Natick es experimental, abrió la puerta a diseños fuera de la caja para data centers del futuro.

*figura – Recuperación de un centro de datos submarino de **Project Natick** (Microsoft) tras dos años de operación en el océano (Orcadas, Escocia). La refrigeración natural del agua de mar y la atmósfera controlada permitieron una alta eficiencia y fiabilidad, mostrando el potencial de enfoques no convencionales en la búsqueda de HPC sostenible.*

Impacto ecológico del entrenamiento de modelos de IA:

Un aspecto crítico de la HPC actual es el **coste ambiental de entrenar grandes modelos** de IA. Modelos como GPT-3 o GPT-4 requieren recursos computacionales inmensos, lo que se traduce en consumo eléctrico abrumador y también en uso de agua (para enfriar los servidores).

Un estudio estimó que entrenar GPT-3 (175 mil millones de parámetros) pudo emitir del orden de **500 toneladas de CO_2** y consumir **700.000 litros de agua** en los data centers durante su entrenamiento. GPT-4, con más parámetros, habría multiplicado esas cifras. Estos números equivalen a las emisiones de varios coches a lo largo de su vida útil, solo por entrenar un modelo de IA.

Este elevado consumo está llevando a medidas como: empresas que exploran alimentar sus centros de datos **con energía nuclear** o solar dedicada, optimizar el código de entrenamiento para usar menos operaciones, o incluso impuestos/regulaciones verdes para incentivar la eficiencia. La comunidad de IA también investiga técnicas como *Green AI* (modelos más pequeños o entrenamientos más cortos) y compensaciones de carbono.

(Nota: Microsoft ha llegado a **considerar reactivar plantas de energía nuclear para abastecer la demanda de sus clústeres de IA** y compañías como OpenAI y NVIDIA trabajan en hardware y métodos más eficientes energéticamente por necesidad de sostenibilidad.)

Conclusión del bloque 3: La sostenibilidad se ha convertido en una preocupación central para la HPC. Cada mejora de rendimiento debe equilibrarse con su coste energético. Métricas como PUE nos ayudan a medir avances, y casos de éxito (data centers en climas fríos, IA optimizando IA, datacenters submarinos) muestran que **sí es posible reducir drásticamente la huella** de la computación avanzada. Sin embargo, el crecimiento exponencial de la demanda (especialmente por la IA) plantea la pregunta de si el modelo actual de aumento continuo de capacidad es sostenible. La innovación en **tecnologías de refrigeración**, **fuentes de energía renovable** y **diseños eficientes** será fundamental para mantener el progreso de la HPC sin comprometer el planeta.

Preguntas de reflexión (Sostenibilidad):Si tuvieras que proponer una medida para hacer más "verde" un centro de supercomputación, ¿cuál sería? Piensa en **coste vs. beneficio**: ¿Crees que iniciativas radicales como centros de datos submarinos o usar IA para control ambiental se adoptarán ampliamente, o son soluciones demasiado específicas? ¿Deberían los grandes modelos de IA estar sujetos a una "etiqueta energética" pública (al estilo electrodomésticos) para visibilizar su impacto?

4. Procesamiento masivo de datos (Big Data)

¿Qué es Big Data en el contexto HPC?*Big Data* se refiere al conjunto de tecnologías y arquitecturas para almacenar, procesar y analizar volúmenes enormes de datos que superan las capacidades de sistemas tradicionales. HPC y Big Data están íntimamente ligados: la CAR proporciona la potencia bruta para manejar Big Data, mientras que Big Data aporta problemas y datasets que empujan al límite a la CAR. Las **4 V** clásicas caracterizan el Big Data:

- **Volumen:** Cantidades masivas de datos generados continuamente. Ejemplo: **Google procesa más de 100 PB de datos al día** en sus centros (logs de búsqueda, videos de YouTube, etc.). La escala de almacenamiento se mide en petabytes o exabytes.
- **Variedad:** Los datos vienen en muchos formatos: *estructurados* (bases relacionales), *semiestructurados* (JSON, XML, CSV) y *no estructurados* (texto libre, imágenes, audio, video). Un entorno Big Data típico combina bases SQL, documentos, registros de sensores IoT, redes sociales, etc. con necesidad de integrarlos para análisis.
- **Velocidad:** Tanto la velocidad de generación de datos (streams en tiempo real de sensores, transacciones, clics web) como la velocidad necesaria de procesamiento (analítica *casi en tiempo real*). Ejemplo: sistemas antifraude bancario o predicción del tiempo necesitan procesar flujos de datos en segundos o menos.
- **Veracidad:** Calidad e incertidumbre de los datos. Datos masivos a menudo contienen ruido, errores, duplicados o sesgos. Confiar en Big Data exige

técnicas de limpieza, validación y cruzamiento de fuentes para garantizar que las conclusiones sean fiables.

(Nota: A veces se añaden más "V": Valor, Variabilidad, Visualización… pero las cuatro anteriores son las esenciales.)

Sistemas de almacenamiento y acceso distribuido: Para manejar Big Data, las organizaciones usan sistemas distribuidos que reparten datos y cómputo entre muchos servidores. En HPC/BigData existen soluciones emblemáticas desarrolladas por pioneros como Google, que luego inspiraron herramientas open source:

- **Colossus (Google):** Es el sistema de archivos distribuido de Google (sucesor evolucionado de GFS). Divide los archivos enormes en bloques y los almacena redundados en múltiples máquinas. Permite que miles de nodos lean/escriban datos simultáneamente con tolerancia a fallos (si un nodo cae, los bloques están replicados en otros). Colossus es base de servicios planetarios como YouTube, Gmail, Google Drive, etc., manejando **exabytes** de información[blog.quastor.orgblog.quastor.org](blog.quastor.orgblog.quastor.org). *(Inspiró a Hadoop HDFS en la comunidad open source).*
- **Borg (Google):** Es el sistema interno de gestión de clúster de Google. Borg toma miles de trabajos (consultas, microservicios, procesos MapReduce) y los distribuye eficientemente en los recursos disponibles del centro de datos. Automatiza la orquestación de contenedores, asignación de CPU/RAM, reintentos en caso de fallo… Fue la base conceptual de Kubernetes (Google liberó su experiencia en Borg hacia la comunidad). En Big Data, Borg permite a Google correr *concurrentemente* infinidad de tareas de procesamiento de datos optimizando el uso de su inmensa infraestructura.
- **Hadoop & Spark (open-source):** Siguiendo el modelo de Google MapReduce, Apache Hadoop ofreció almacenamiento distribuido (HDFS) y cómputo paralelo escalable en clusters commodity. Spark, de segunda generación, mejoró el rendimiento manteniendo datos en memoria y ofreciendo un modelo de programación más flexible (algo así como "HPC para Big Data"). Hoy son estándares en análisis de grandes datos en muchas empresas (sobre clústeres on-premise o en la nube).

En resumen, estos sistemas permiten que **ni los datos ni el procesamiento estén centralizados**, sino repartidos en decenas o miles de máquinas, logrando escalabilidad horizontal (crece la capacidad agregando más nodos).

Problemas técnicos en Big Data HPC:

- **Data locality (localidad de datos):** "Mover computación es más barato que mover datos". En Big Data, idealmente uno envía la *tarea* a los nodos donde ya están los datos, evitando transferir grandes ficheros por la red. Si no se respeta, las copias de datos en red se vuelven cuellos de botella, con alta latencia y congestión. Los frameworks Big Data modernos intentan maximizar esta localidad.
- **Coste de mover datos:** A veces no hay más remedio que mover datos (por ejemplo, agregarlos de muchos nodos hacia uno para un resultado final).

Transferir datos a través de la red implica: **latencia** (tiempo de ida/vuelta), consumo de **ancho de banda** (la red es recurso compartido) y potenciales riesgos de seguridad. Por ello surgen arquitecturas de *Edge Computing* (procesar en la fuente), *caching inteligente* y *replicación selectiva* para minimizar tráfico. Un dicho en HPC/BigData es que "los discos duros nunca se subieron a un avión, pero las cintas sí" – ilustrando que a veces es más rápido enviar físicamente medios de almacenamiento que mover petabytes por internet.

- **Integridad y calidad de datos:** A diferencia de cálculos puramente numéricos de HPC tradicional, en Big Data la heterogeneidad de fuentes puede introducir datos corruptos o incoherentes. La HPC tiene que lidiar con limpieza en escala: ¿cómo filtrar outliers en 100 PB? ¿cómo asegurar consistencia entre datos distribuidos? Son retos de software que impactan el rendimiento (un mal join de datos puede costar horas de cómputo desperdiciado).
- **Dimensionamiento y coste:** Procesar Big Data en HPC requiere planificar cuidadosamente cuántos nodos y cuánta memoria se necesita para evitar saturación. Un dataset puede no caber en RAM ni disco local, obligando a E/S remota continua (lo cual es lento). La nube ofrece elasticidad, pero alquilar cientos de núcleos/hora tiene un coste significativo, así que la eficiencia económica se vuelve también un parámetro técnico.

Caso real – Google: Google es un referente en Big Data a escala HPC. Sus cifras ilustran el estado del arte:

- Procesa **100+ PB diarios** (indexación web, análisis de uso, traducciones, etc.).
- Emplea internamente sistemas como Colossus y Borg para soportar esta escala. Su pipeline MapReduce original ha evolucionado a sistemas como **Flume** (procesamiento de stream) y **Beam** (modelo unificado batch/stream).
- Aloja servicios de escala planetaria:
  - **YouTube:** cientos de horas de vídeo subidas por minuto, con HPC transcodificando vídeos, analizando contenido y recomendando en tiempo real.
  - **Búsquedas de Google:***mil millones de búsquedas diarias*, donde índices gigantes (~100 petabytes) son recorridos en milisegundos para dar resultados relevantes.
  - **Google Translate:** entrenado en HPC con datos multilingües masivos, y que en producción requiere HPC para inferencia rápida en docenas de idiomas simultáneamente.
- Su modelo es altamente distribuido y tolerante a fallos: ningún servidor individual es crítico, todo está replicado; si falla un rack, los procesos se reinician en otro automáticamente (Borg se encarga); el sistema está diseñado para escalar casi linealmente añadiendo máquinas.

En definitiva, Google muestra cómo **HPC + Big Data** permiten crear aplicaciones de impacto global, apoyadas en infraestructuras capaces de manejar información y cálculos a una escala impensable hace pocas décadas.

Conclusión del bloque 4: El Big Data plantea retos únicos a la HPC: no se trata solo de calcular rápido, sino de **gestionar datos masivos eficientemente**. Las arquitecturas distribuidas, tanto propietarias (Google Colossus/Borg) como abiertas (Hadoop/Spark),

son la respuesta para alcanzar escalabilidad. Sin embargo, manejar los 4V sigue requiriendo innovaciones constantes en almacenamiento, redes y software. El procesamiento masivo de datos es ya rutina en las grandes tecnológicas, pero el resto del mundo busca democratizar ese poder (apoyándose en la nube HPC, nuevas plataformas, etc.). La capacidad de obtener *valor* de Big Data a través de HPC se ha convertido en ventaja competitiva en ciencia y negocios.

Preguntas de reflexión (Big Data): **¿Cuál de las "4 V" del Big Data te parece más difícil de abordar con HPC y por qué?** Por ejemplo, ¿es más complejo lidiar con la *velocidad* (streaming en tiempo real) o con la *variedad* (integrar datos dispares)? En tu opinión, ¿qué es más crítico: invertir en hardware más potente o en mejores algoritmos de procesamiento de datos?

5. Respuesta de las grandes tecnológicas

Las compañías tecnológicas líderes (**Google, Meta (Facebook), Amazon, Microsoft**) están a la vanguardia de la HPC, impulsando avances en IA, infraestructuras masivas y sostenibilidad. A continuación, veremos las iniciativas destacadas de cada una en cuanto a supercomputación y procesamiento de datos.

5.1 Google: TPUs, sostenibilidad y data centers inteligentes

- **TPUs (Tensor Processing Units):** Google diseñó sus propios aceleradores de IA llamados TPUs, optimizados para tareas de *machine learning*. Estas ASICs (circuitos integrados específicos) aceleran operaciones de redes neuronales con gran eficiencia. Van por la 5ª generación (TPU v5e), usada en Google Cloud, que ofrece rendimiento exaflop para IA con menor consumo energético. Con TPUs, Google entrena modelos como Google Translate o AlphaGo sin depender de GPUs de terceros, logrando un stack vertical integrado de hardware-software.
- **Compromiso de sostenibilidad 24x7:** Google anunció la meta de operar todos sus centros de datos con energía **libre de carbono 24/7 para 2030**. Esto va más allá de ser 100% renovable en balance anual; implica que *cada hora* del día la electricidad provenga directamente de fuentes eólicas, solares, hidro, etc. Para lograrlo, invierten en nuevas plantas renovables, baterías y otras tecnologías de almacenamiento, además de mejorar la eficiencia. Google ya logró reducir drásticamente su PUE global (≈1.1), y utiliza IA para minimizar uso de agua y electricidad (como vimos, DeepMind AI Cooling).
- **Gestión inteligente de data centers:** Google aplica algoritmos de IA y analítica avanzada para micro-optimizar sus operaciones: desde predecir fallos de discos para reemplazarlos antes de que fallen, hasta ajustar la distribución de cargas de trabajo entre regiones según la disponibilidad de energía renovable en ese momento. Estas optimizaciones "invisibles" han dado grandes frutos, como reducir hasta un 30% la energía de enfriamiento, o ajustar voltajes de CPUs en vivo para ahorrar energía cuando la carga es baja.

En conjunto, Google lidera integrando **innovación hardware (TPUs)** con **estrategias agresivas de sostenibilidad** e **inteligencia operativa** para mantener su posición puntera en HPC verde. Un resultado concreto: en 2023 Google alcanzó un promedio del

64% de energía libre de carbono en tiempo real en sus data centers (y 100% renovable en compensación anual)techtarget.comtechtarget.com, camino al 24x7 para 2030.

5.2 Meta: AI SuperCluster y energía renovable

- **AI Research SuperCluster (RSC):** Meta construyó en 2022 uno de los supercomputadores de IA más potentes del mundo, orientado a investigación en IA. El RSC está compuesto inicialmente por **760 sistemas NVIDIA DGX-A100**(más de 6000 GPUs A100 interconectadas) y escalará a **16,000 GPUs** en su fase finaldatacenterfrontier.com. Su objetivo es entrenar modelos de visión por computador y NLP de próxima generación (por ejemplo, modelos capaces de entender cientos de idiomas y realizar traducción automática en tiempo real para el metaverso de Meta). Meta reporta que RSC será posiblemente el **AI supercomputer más rápido** una vez completo, capaz de alcanzar ~5 exaFLOPs en precisión reducida para AI. Esto ilustra cómo las empresas montan infraestructuras HPC dedicadas específicamente a IA.

*Figura – **Meta AI Research SuperCluster (RSC)**: Vista de racks de GPU dentro del nuevo supercomputador de Meta para IA (cada rack contiene 2 sistemas DGX A100 con 16 GPUs en total). RSC cuenta con un diseño de red y almacenamiento de última generación para mantener las miles de GPUs alimentadas con datos de entrenamiento sin cuellos de botella*

- **Energía 100% renovable:** Meta alcanzó en 2020 que el 100% de la energía de sus operaciones globales provenga de fuentes renovables (eólica y solar

principalmente. Ha invertido en proyectos equivalentes a 12 GW de capacidad renovable. Sus data centers, como vimos con Odense (Dinamarca), usan energía limpia y buscan innovación como reutilizar calor. Meta incluso está explorando opciones de comprar energía nuclear modular para asegurar suministro limpio a futuro (pues sus necesidades de IA crecen y la solar/eólica son intermitentes).

- **Innovación energética:** Además de eólica/solar, Meta investiga otras fuentes: geotermia (usar calor del subsuelo para climatización) y celdas de combustible de hidrógeno para backup en vez de generadores diésel. Todo con la meta de **cero emisiones netas** en su huella para 2030 y mantener la sostenibilidad aún a medida que crece la demanda de computación para el *metaverso* y IA.

En resumen, Meta combina **inversión en HPC de IA** (RSC) con un fuerte **compromiso en renovables**. Su escala global (Facebook, Instagram, WhatsApp) obliga a pensar en eficiencia a cada paso. El hecho de que su supercomputador esté *alimentado enteramente con energía limpia* marca un estándar para la industria.

5.3 Amazon Web Services (AWS): democratización del HPC en la nube

- **HPC en la nube para todos:** AWS (Amazon) ha convertido la HPC en un servicio disponible bajo demanda. Ofrece instancias de cómputo intensivo (p. ej. instancias con 96 núcleos, GPU Tesla, incluso instancias con múltiples GPUs por nodo) y servicios como **AWS ParallelCluster** para desplegar clústeres HPC virtuales. Esto permite que startups, universidades pequeñas o cualquier entidad con una tarjeta de crédito accedan a recursos HPC sin comprar hardware propio. Simulaciones, modelado y entrenamiento de IA a gran escala ahora se pueden hacer "rentando" horas de un supercomputador virtual en AWS . Es un cambio de paradigma: *HPC as a Service*.
- **Servicios especializados:** AWS no solo da máquinas virtuales, también herramientas para facilitar su uso: *Elastic Fabric Adapter* (adaptador de red de baja latencia) que conecta instancias HPC como si fueran un Infiniband, almacenamiento paralelo FSx for Lustre para datos compartidos de alta velocidad, y optimizaciones para ejecutar herramientas científicas populares. Todo administrado desde la consola AWS. Así, un investigador puede lanzar en minutos 1000 núcleos con MPI configurado y un sistema de archivos paralelo, algo que antes requería una infraestructura física costosa.
- **Sostenibilidad AWS:** Amazon (como empresa global) se comprometió a cero emisiones netas de carbono para 2040 y a alimentar sus operaciones con 100% renovables para 2025. AWS, al ser la mayor plataforma cloud, está invirtiendo en enormes parques solares y eólicos. Además, argumenta que su nube es más eficiente que centros de datos empresariales tradicionales (por economías de escala y alta optimización). Un informe indicó que la nube de AWS puede ser 3,6 veces más eficiente que un centro de datos mediano típicotechtarget.comtechtarget.com. En HPC esto significa que mover cargas a la nube podría en teoría reducir la huella (siempre que AWS efectivamente use energía verde para esas cargas, lo cual persiguen).

En síntesis, AWS está *democratizando* la HPC: lo que antes solo unos pocos laboratorios podían costear, ahora está a disposición bajo demanda. Esto abre oportunidades, pero también crea dependencia de un proveedor centralizado (un

punto a reflexionar). Su apuesta en sostenibilidad busca que ese crecimiento no incremente la contaminación global.

5.4 Microsoft: refrigeración autónoma y sostenibilidad global

- **Refrigeración autónoma y experimental:** Microsoft ha innovado en cooling para sus centros de datos de Azure. Ha implementado **refrigeración líquida por inmersión** en algunas de sus instalaciones: sumergen servidores en bañeras con un líquido dieléctrico especial que hierve a baja temperatura llevando el calor fuera eficientemente. Es el primer gran proveedor cloud en usar inmersión a escala comercial, lo que puede reducir el consumo de agua y energía significativamente (porque no necesitan torres de enfriamiento evaporativas). También han probado enfriamiento "on chip" con fluidos especiales corriendo dentro de bloques en contacto directo con CPU/GPUs. Estas medidas permiten mayor densidad de computación por rack y menos requerimientos de aire acondicionado.
- **Carbono negativo 2030:** Microsoft fue una de las primeras en prometer que para 2030 no solo será carbono neutral, sino *carbono negativo* (eliminará más CO_2 del que emite). Además, se comprometió a eliminar para 2050 todo el carbono emitido por la empresa en su historia. También anunció objetivo "water positive" para 2030 (devolver más agua de la que consume) – relevante porque los data centers consumen mucha agua en refrigeración evaporativa. Para ello, invierten en proyectos de reforestación, captura de carbono, restauración de cuencas de agua, etc., y rediseñan sus data centers para usar **cero agua** en enfriamiento en climas secos (por ejemplo, usando más aire y menos evaporación).
- **Innovación en centros de datos:** Microsoft, al igual que otros, también explora energías alternativas para sus data centers. Ha probado celdas de combustible de hidrógeno para sustituir generadores de respaldo, y evalúa pequeñas plantas nucleares modulares para futuros centros de gran tamaño. Continúa con el proyecto **Natick** (ya visto) para posibles despliegues en zonas costeras. Todo orientado a que los futuros centros de datos sean *hiper-eficientes*: 100% renovables, mínimos residuos de calor (y los que haya, reutilizados), mínima huella hídrica, etc.

En resumen, Microsoft combina **investigación en cooling revolucionario** con compromisos ambientales ambiciosos. Como proveedor de Azure (servicios cloud y HPC en la nube), tiene la presión de escalar sin aumentar emisiones. Sus avances en inmersión y metas carbono-negativo marcan pauta para la industria.

Conclusión del bloque 5: Las grandes empresas tecnológicas están redefiniendo los límites de la HPC:

- Google con hardware a medida (TPUs) y operaciones ultra-eficientes 24x7 libres de carbono.
- Meta con uno de los mayores supercomputadores de IA y alimentado por renovables.
- AWS llevando HPC a la nube para que cualquiera la use, mientras se expande con renovables a escala masiva.

- Microsoft apostando por *tecnología verde* (inmersión, energía alternativa) y ambiciones de eliminar su huella histórica.

Estas compañías concentran mucho poder computacional, pero también están presentando soluciones pioneras a los retos de escalabilidad y sostenibilidad. Surge, no obstante, un debate: **la dependencia global de la HPC de unas pocas empresas** – lo exploraremos en la dimensión ética y geopolítica.

Preguntas de reflexión (Grandes Tecnológicas): **¿Te parece positivo que el avance de la HPC esté liderado por empresas privadas?** Por un lado, innovan rápidamente (TPUs, RSC, etc.), por otro, concentran poder. ¿Qué riesgos ves en depender de servicios de nube como única vía de acceso a HPC? ¿Crees que iniciativas como los compromisos de carbono cero de estas empresas son suficientes o hacen falta regulaciones externas?

6. Dimensión ética, social y política de la CAR

La Computación de Alto Rendimiento no es solo tecnología; implica **poder**, **equidad**, **soberanía digital** y **ética**. ¿Quién controla el cómputo masivo? ¿Cómo se reparte el acceso? ¿Qué riesgos acarrea la concentración de tanta capacidad en pocas manos? Y ¿cómo están respondiendo las sociedades y gobiernos para regular esto? Veamos estos puntos.

¿Quién tiene el poder computacional hoy? La capacidad de entrenar una IA puntera o de simular algo complejo como el clima global está limitada a quienes poseen HPC a gran escala:

- **Corporaciones tecnológicas:** Google (con sus TPU pods), Microsoft/OpenAI (Azure supercomputing clusters para GPT-4), Meta (RSC), Amazon, NVIDIA, etc. Todas invierten miles de millones en HPC porque saben que *más cómputo = ventaja competitiva* en IA.
- **Gobiernos y ejércitos:** Supercomputadoras nacionales en EE.UU., China, la UE, Japón, etc., a menudo con fines estratégicos (desde diseñar armas nucleares hasta predicción de desastres naturales).
- **Consorcios científicos y grandes universidades:** CERN, proyectos como el telescopio Event Horizon o secuenciadores genómicos globales usan HPC, aunque muchos dependen de apoyos gubernamentales o alianzas con la industria por el costo.

Preguntas que emergen: Si solo un puñado de actores tienen "superpoder computacional", **¿quién decide qué problemas se investigan con él?** Por ejemplo, si una empresa prioriza modelos de publicidad en vez de, digamos, simulaciones de cura del cáncer, ¿qué implica? **¿Qué sucede si la infraestructura pública (universidades, laboratorios estatales) queda rezagada frente a la privada?** Podríamos ver una dependencia de investigadores académicos hacia recursos de empresas para realizar ciencia (ya está ocurriendo en IA). **¿Y los países en desarrollo?** Hoy pocos de ellos tienen supercomputadores; dependen de comprar

servicios a otros. Esto puede aumentar la brecha global en investigación y desarrollo.

Riesgos de concentración, dependencia y monopolios: El acceso desigual al HPC puede reforzar monopolios tecnológicos y relaciones de dependencia:

- *Soberanía digital:* Países sin sus propios supercomputadores o sin capacidad de fabricar chips avanzados quedan sujetos a proveedores externos. Ej., si X país depende de la nube de Amazon para su e-gobierno o su meteorología, su autonomía se ve limitada.
- *Investigación sesgada:* Si la mayor parte del HPC está en manos privadas, las prioridades de investigación podrían inclinarse a intereses comerciales (ej., modelos de consumidor) más que al bien público, a menos que haya políticas correctoras.
- *Mercado laboral y económico:* Quien controla HPC puede automatizar y optimizar procesos masivamente, ganando ventaja económica. Empresas con HPC dominante pueden ofrecer servicios que ninguna startup sin acceso a HPC podría igualar, consolidando su posición.

Ejemplo actual: OpenAI (con Microsoft) tiene acceso prácticamente exclusivo a un supercluster de IA entrenando GPT-4, GPT-5, etc. Una pequeña startup o universidad no puede replicar ese entrenamiento, generando una barrera de entrada enorme en el campo de los LLM (Large Language Models). Esto tiende a un **oligopolio**: solo gigantes pueden costear entrenar los modelos más avanzados, que luego otros solo pueden *usar*, no crear.

Marco normativo emergente: Los gobiernos comienzan a reaccionar regulando el uso de la IA y, por extensión, del poder computacional:

- ♡ **RGPD (UE) – Reglamento General de Protección de Datos:** Aunque enfocado a datos personales, indirectamente afecta HPC e IA. Prohíbe usar datos personales sin consentimiento y exige minimización de datos. Esto limita, por ejemplo, entrenar modelos con ciertos datos sensibles en Europa. Obliga a desarrollar HPC con **privacidad** en mente (anonimización, federated learning, etc.).
- **AI Act (UE) – Ley de Inteligencia Artificial:** En negociación (a final de 2024), propone clasificar sistemas de IA por riesgo y poner requisitos. Para *IA de alto riesgo* (p.ej. sistemas médicos, de transporte, o IA general tipo GPT), exigirá *auditorías, trazabilidad de datos, documentación del proceso de entrenamiento y supervisión humana*. Esto implicará que quienes entrenan grandes modelos deben registrar qué datos usan, cómo los evalúan, etc. Si se aprueba, influirá en cómo se desarrollan modelos en HPC dentro de la UE. También discute la posibilidad de requerir permisos para entrenar ciertos modelos fundacionales.
- us**EE.UU.:** No tiene (a 2025) una ley federal integral de IA, pero sí movimientos: control de exportaciones (prohibir venta de GPUs avanzadas a ciertos países por seguridad nacional), marcos éticos voluntarios, y enormes inversiones en HPC a través de la NSF, DOE, etc. (El CHIPS Act destina fondos

para fomentar fabricación de chips y HPC domesticamente). El énfasis es mantener la competitividad y no quedarse atrás de China.

- CN**China:** Implementó regulaciones estrictas sobre algoritmos (algoritmos recomendadores deben registrarse), y censura sobre outputs de modelos generativos. A la vez, invierte fuertemente en su HPC e IA estatal (Ej: supercomputadores secretos posiblemente ya exaescala, chips nacionales como Huawei Ascend) y limita salida de datos al extranjero. Su enfoque es la *soberanía tecnológica total*.
- **Geopolítica de la computación:** Estas leyes y políticas muestran que HPC e IA son consideradas **estratégicas**. Tener o no tener HPC influye en la posición de un país en el mundo. Se está configurando una carrera similar a la nuclear en el siglo XX, pero con supercomputadores e IA.

Accountability y transparencia en grandes modelos: Un reto ético serio es la falta de trazabilidad en los modelos de IA entrenados con HPC. Surgen cuestiones:

- *¿Cómo se entrenó exactamente el modelo?* Muchas veces no se documenta bien qué datos exactos se usaron ni qué procesos siguió (por secretismo o falta de metodología), lo que dificulta replicabilidad científica y responsabilidad.
- *¿Quién es responsable si el modelo causa daño?* Si un modelo de IA toma decisiones (concede un préstamo, conduce un coche) y falla, ¿la culpa es del desarrollador, del operador que le proveyó los datos, o del modelo "en sí"? En HPC entrenamos modelos tan complejos que ni sus creadores entienden totalmente su funcionamiento interno.
- *Caja negra:* Muchos modelos enormes (ej. GPT-4) son cajas negras incluso para sus creadores. Esto desafía principios de **auditoría** y **explicabilidad**. Desde lo legal, se podrían requerir "*model cards*" o documentación detallada, pero mientras los modelos sean opacos, hay un vacío de accountability.

En varias jurisdicciones se discute obligar a que los modelos tengan algún tipo de "certificación", y que las decisiones automatizadas tengan posibilidad de explicarse a un humano (ej. el RGPD tiene el derecho a explicación en decisiones algorítmicas). Pero estamos corriendo para poner reglas *después* de haber soltado modelos muy potentes.

Dilema ético: ¿Proteger los datos o proteger el conocimiento? Aquí aparece una tensión interesante:

- Por un lado, proteger al máximo los datos personales (privacidad) es crucial – leyes como RGPD lo imponen.
- Por otro lado, entrenar IA útiles (en salud, ciencia) requiere acceder a muchos datos, a veces sensibles. Si la regulación bloquea demasiado el uso de datos, podríamos estar frenando avances importantes (*proteges la privacidad, pero tal vez a costa de no descubrir una cura, por ejemplo*).
- Si priorizamos acceso abierto a datos para "proteger el conocimiento" y permitir más IA, arriesgamos privacidad, posibles abusos y discriminación.

Ejemplo: Modelos de IA médica que podrían diagnosticar enfermedades raras necesitan entrenar con historiales clínicos de miles de pacientes. Si las leyes

impiden compartir esos datos por privacidad, el modelo puede no desarrollarse. ¿Qué prima, la privacidad de unos datos o el avance médico que salvaría vidas? No es trivial. Incluso anonimizados los datos, siempre hay riesgo. Este dilema requiere equilibrios: técnicas como *federated learning* (los datos nunca salen de hospitales, se entrena de forma distribuida) buscan mitigar el conflicto. Pero sigue la pregunta de fondo: **¿hasta dónde restringir datos por privacidad sin ahogar la innovación?** Y a la inversa, **¿cómo impulsar innovación sin atropellar derechos individuales?**

Conclusión del bloque 6: El poder de la HPC conlleva responsabilidad. Estamos viendo cómo la **concentración de recursos computacionales** puede crear desigualdades y tensiones de poder (tecnológicas, económicas y entre naciones). La regulación intenta ponerse al día para asegurar que la IA y HPC se desarrollen de forma **ética y transparente**. Sin embargo, es un terreno complejo: la tecnología avanza rápido, globalmente, mientras las leyes son locales y lentas. Como futuros profesionales, tendremos que navegar estos dilemas, balanceando innovación con responsabilidad social.

Preguntas de reflexión (Ética y sociedad): **¿Quién debería decidir los usos permitidos de la HPC y la IA?** (¿Gobiernos, empresas, una entidad internacional?). **¿Debería el acceso a supercomputación considerarse un bien público?** Piensa en la analogía con el acceso a internet o a la electricidad. Y en el plano personal: si trabajas en un equipo que desarrolla una IA muy poderosa, ¿qué medidas tomarías para asegurar que se use de forma justa y transparente?

7. Geopolítica y soberanía tecnológica

La HPC se ha convertido en un **activo estratégico nacional**. Tener los mejores supercomputadores y la capacidad de procesar grandes datos es visto como crucial para la competitividad económica y la seguridad nacional. Por ello, existe una verdadera **carrera global** por el liderazgo en HPC, que involucra a EE.UU., China, Europa, Japón, entre otros.

Supercomputación como activo estratégico: ¿Por qué le importa a un país tener supercomputadores punteros?

- *Autonomía tecnológica:* Poder procesar datos sensibles (defensa, inteligencia) en infraestructura propia, sin depender de terceros. Por ejemplo, simulaciones nucleares o de criptografía requieren HPC que un país no querrá ejecutar en la nube de otro país.
- *Ventaja en ciencia e industria:* HPC acelera descubrimientos (nuevos materiales, fármacos, energías). Un país con HPC de primer nivel puede liderar en innovación.
- *Defensa y seguridad:* HPC se usa en criptoanálisis, en diseño de armamento avanzado, en inteligencia artificial militar. Es una pieza más del poderío militar moderno (como tener satélites o drones).

- *IA general:* Entrenar IA de propósito general (como GPTs) requiere HPC inmenso. Quien lo posea podrá desarrollar IAs avanzadas de forma independiente.

Ejemplos de énfasis nacional: En 2022 EE.UU. declaró prioridad estratégica lograr computadores exaescala antes que sus rivales – y lo hizo con Frontier. La Unión Europea creó la iniciativa *EuroHPC* (detallada abajo) para no quedarse atrás en HPC. Países como España se posicionan: el Barcelona Supercomputing Center será un nodo clave para IA europea. Esto muestra que HPC está en agendas gubernamentales.

Competencia global por el liderazgo en HPC: Una tabla resumen de la "carrera" HPC:

País/Región	Estrategia & Enfoque	Superordenador emblemático
US **EE.UU.**	Defensa, IA, clima, biomedicina	*Frontier* (Oak Ridge) – 1ª exaescala
CN **China**	Seguridad nacional, IA gubernamental	*Sunway TaihuLight* (antiguo #1) y sistemas exa escala **no públicos**
JP **Japón**	Investigación climática, salud, industria	*Fugaku* (RIKEN) – Top 3 mundial
EU **Europa**	Ciencia abierta, IA ética, colaboración UE	*LUMI* (Finlandia), *Leonardo* (Italia), *MareNostrum5* (España)

(Nota: En Nov 2023, Frontier (EE.UU.) era #1; China se rumorea tiene 2 exa-computadores no listados oficialmente; Japón Fugaku #2; Europa tiene varios pre-exa en Top10).

Cada bloque geopolítico busca controlar toda la cadena de valor: desde la **fabricación de chips** (por eso tensiones con Taiwán/TSMC), el desarrollo de **software HPC** propio, el acceso a **energía** para alimentarlos, una **regulación favorable** y retener el **talento humano** (científicos e ingenieros). Vemos por ejemplo a China invirtiendo en sus equivalentes de NVIDIA para no depender de occidente, y a EE.UU. bloqueando exportaciones de GPUs A100/H100 a China para ralentizar su avance. Europa, que no fabrica los chips más punteros, invierte en colaboración pan-europea para atajar su dependencia.

Proyectos cooperativos – EuroHPC: Reconociendo que ningún país europeo individual podía competir con EE.UU./China solo, la UE lanzó *EuroHPC JU* (Joint Undertaking) para aunar recursos:

- Construir **supercomputadoras de clase mundial en suelo europeo** (por ejemplo LUMI en Finlandia, 151 PFlops, 100% renovable; Leonardo en Italia, 240 PFlops; MareNostrum 5 en España en progreso).
- Compartir el acceso a estos sistemas entre países miembros para fines de investigación pública y también pymes innovadoras.

- Financiar desarrollo de tecnologías HPC europeas (chips específicos, software, algoritmos) y aplicaciones en ciencia (desde medicina personalizada hasta cambio climático). La meta es tener al menos **dos supercomputadores exaescala** europeos para 2025-2026.
- Esto es una estrategia de *soberanía digital compartida*: Europa unida puede tener HPC competitivo (ej. LUMI estuvo #3 global un tiempo) sin depender totalmente de nubes de EE.UU.

Acceso desigual a la infraestructura HPC: A nivel global, la potencia computacional está concentrada en países ricos. Consecuencias:

- Países en desarrollo o menos adelantados en tecnología *dependen* de infraestructura extranjera (p. ej. usar supercomputadores vía convenios o la nube comercial). Esto puede ser costoso y limitante.
- Muchas universidades pequeñas o centros de investigación carecen de HPC propio; investigadores con ideas brillantes en esos lugares pueden no disponer de los medios para probarlas, a menos que colaboren con quienes sí tienen (generalmente en países G7 o grandes empresas).
- El *coste* del HPC (no solo comprar la máquina, sino pagar energía y contratar expertos para operarla) impide que se difunda ampliamente. Esto refuerza la brecha tecnológica: los "ricos en computación" se vuelven más ricos en capacidad de innovación, mientras otros se quedan mirando.
- Incluso dentro de países desarrollados, puede haber brecha: institutos top vs. universidades pequeñas.

Consecuencia: riesgo de *exclusión científica* de regiones enteras. Si África, gran parte de Latinoamérica, etc., no tienen acceso pleno a HPC, podrían quedar fuera de la próxima ola de descubrimientos en IA, clima, genómica, etc., profundizando desigualdades globales.

Por eso hay llamados a considerar la HPC (especialmente para investigación que beneficie a la humanidad, como clima o salud) como una infraestructura global compartida, donde consorcios internacionales ofrezcan tiempo de supercómputo a proyectos merecedores independientemente del país. Iniciativas como **Latin American HPC collaboration** o **African Supercomputing Centre** empiezan a emerger, aunque aún incipientes.

Conclusión del bloque 7: La HPC se ha politizado: es parte de la *realpolitik* del siglo XXI. Quien lidere en HPC e IA tendrá ventajas económicas y geopolíticas enormes. Vemos carreras entre potencias, alianzas (EuroHPC) para no quedarse atrás, y al mismo tiempo riesgos de exclusión de quienes no puedan subirse a este tren. El panorama sugiere que, así como existe la diplomacia científica, veremos también *diplomacia computacional* – acuerdos para compartir recursos, o disputas por restricciones de tecnología. La soberanía tecnológica (poder controlar tu propio destino digital) está ligada a tener acceso soberano a HPC.

Preguntas de reflexión (Geopolítica): **¿Debería la supercomputación de primer nivel considerarse una infraestructura global a la que todos los países deban tener acceso básico, similar a internet?** ¿Qué riesgos conlleva que empresas

privadas tengan hoy más capacidad computacional que muchos países (ej. Google > la mayoría de gobiernos)? **¿Cómo podría la cooperación internacional ayudar a democratizar el HPC sin que un solo país acapare el control?** (Piensa en modelos ONU, o consorcios tipo CERN pero para computación).

8. Casos de uso actuales con alto impacto

Para aterrizar todo lo anterior, veamos dos casos actuales donde la HPC está teniendo un **alto impacto** y cambiando paradigmas: **(1) IA generativa** y **(2) Ciudades inteligentes**. Son ejemplos de cómo la HPC transforma tanto la tecnología como la vida cotidiana.

8.1 IA generativa (entrenamiento masivo de modelos)

¿Qué es la IA generativa? Es la rama de la IA donde los modelos aprenden la distribución de los datos de entrenamiento tan bien que pueden *generar contenido nuevo similar*: texto, imágenes, audio, código, etc. Los ejemplos más conocidos:

- Modelos de texto: GPT-3, GPT-4 de OpenAI; **Google Gemini** (en desarrollo); Meta LLaMA. Generan ensayos, chats, traducciones...
- Imágenes: DALL·E, **Midjourney**, **Stable Diffusion** – crean imágenes a partir de descripciones.
- Audio/Música: **Suno AI**, **MusicGen** – sintetizan voces o música original.
- Código: GitHub Copilot (basado en OpenAI Codex) – autocompleta y genera código fuente.

Estos modelos han explotado en 2022-2023 y todos tienen en común necesitar **inmensos recursos HPC** para entrenarse.

¿Cómo se entrena un modelo generativo grande? (ejemplo: GPT-3):

1. **Recolección de datos:** Se juntan terabytes de textos (libros digitalizados, Wikipedia, páginas web) o imágenes (raspado web), etc., según el tipo de modelo.
2. **Preprocesamiento:** Se limpian los datos (quitar ruido, formato uniforme) y se convierten a la entrada esperada (por ejemplo, tokenizar texto a números).
3. **Entrenamiento iterativo:** Aquí entra la HPC en acción:
    - Se configura un supercomputador con **miles de GPUs trabajando en paralelo**. Por ejemplo, GPT-3 se entrenó con alrededor de *10,000 GPU* (NVIDIA V100) durante varias semanas.
    - Las GPUs calculan gradientes y actualizan los parámetros del modelo (que en GPT-3 son 175 mil millones). Este proceso se repite *miles de millones de veces* (iteraciones) sobre fragmentos de texto, ajustando pesos para que el modelo aprenda.
    - Se usa paralelismo de datos (cada GPU entrena con porciones distintas) y paralelismo de modelo (el modelo mismo se parte entre GPUs, porque no cabe completo en una sola).

- En total, entrenar GPT-3 costó en torno a **4-5 millones de dólares** en cómputo y consumió una enorme cantidad de energíacnbc.comcnbc.com. GPT-4 fue bastante más.
- Además del pre-entrenamiento "bruto", suelen añadirse etapas de *fine-tuning* (ajuste fino con datos específicos) e incluso *RLHF* (aprendizaje por refuerzo con feedback humano) que también requieren HPC.

4. **Generación/Inferencia:** Una vez entrenado, usar el modelo también requiere recursos, pero mucho menores que entrenarlo. Aún así, servir ChatGPT por ejemplo implica cientos de GPUs operando en clusters para responder a millones de usuarios simultáneos.

Sin HPC, nada de esto sería posible. De hecho, se dice que la reciente "revolución ChatGPT" llegó en 2023 no solo por avances algorítmicos, sino porque *por fin* se contaba con la **capacidad computacional** y datos suficientes para hacer entrenamientos de modelos gigantes que antes eran impensables.

¿Qué hace posible la CAR aquí?:

- Procesar en paralelo cantidades astronómicas de operaciones (los últimos modelos requieren del orden de 10^23 operaciones de cálculo para entrenarse por completo).
- Entrenar modelos con **cientos de miles de millones de parámetros** (cada parámetro es un número ajustable que la HPC va optimizando). Sin memoria y computación masiva, no se podrían manejar modelos tan grandes.
- Repetir ciclos de entrenamiento una y otra vez hasta que el modelo converge (lo que puede tardar días o semanas continuos). La HPC actual permite entrenar en, digamos, 2 semanas lo que con menos recursos tardaría años.
- Probar técnicas avanzadas como *ensembles*, *búsqueda de hiperparámetros* u otros que multiplican aún más la carga, pero dan mejor resultado final.

Aplicaciones concretas de IA generativa (impacto):

Sector	Aplicaciones con IA generativa
Educación	Tutores virtuales personalizados; generación de contenidos educativos adaptados (ej. preguntas, resúmenes).
Salud	Síntesis de informes clínicos; análisis de textos médicos; asistentes que orientan diagnósticos a partir de síntomas descritos.
Creatividad	Herramientas para diseñadores gráficos (generar bocetos, imágenes conceptuales), escritores (ideas de trama, borradores), músicos (melodías o arreglos sugeridos).
Programación	Asistentes tipo Copilot que sugieren código, documentan funciones o incluso generan módulos enteros a partir de descripciones.
Empresas	Chatbots de atención al cliente avanzados (entienden consultas complejas, responden coherentemente), generación de informes de negocio automatizados, análisis de sentimientos en redes sociales a escala.

(Ejemplos: Un diseñador puede pedir a DALL·E "muéstrame variaciones de logo" ahorrando horas de bocetaje; un estudiante puede practicar conversación en otro idioma con ChatGPT como tutor; un médico rural podría consultar a una IA (con las

debidas precauciones) para orientación diagnóstica cuando no tiene especialista cerca.)

En todos estos casos, la HPC ha sido el *catalizador*: entrenar el modelo generativo subyacente en un supercomputador, para luego democratizar su uso en aplicaciones. Esto está cambiando cómo **creamos contenido y software**, con enormes implicaciones económicas y sociales.

8.2 Ciudades inteligentes y simulación urbana

¿Qué son las ciudades inteligentes? Son urbes que integran tecnología, datos masivos e inteligencia para mejorar la calidad de vida, optimizar servicios públicos y hacer sostenible la ciudad. Ejemplos de ámbitos: tráfico, transporte público, gestión de energía, seguridad ciudadana, recogida de basura eficiente, respuesta a emergencias, participación ciudadana digital, etc..

La HPC entra aquí para permitir crear **gemelos digitales urbanos**: réplicas virtuales de la ciudad en las que se pueden simular miles de escenarios "what if" rápidamente:

- Simular el tráfico vehicular en tiempo real, detectando congestiones y replanificando semáforos o rutas (*traffic management*).
- Probar cómo una nueva línea de metro o la peatonalización de calles afectaría al flujo de personas y coches antes de implementarla.
- Evaluar la calidad del aire o contaminación sonora en distintos sectores según cambios en industrias o clima.
- Planificar **evacuaciones masivas** o respuestas a desastres (terremoto, inundación) en el gemelo digital para ver cuellos de botella y mejorar planes de emergencia.
- Gestionar redes eléctricas inteligentes: simular la demanda y oferta en cada barrio, integrar energías renovables, detectar posibles apagones con antelación.

Para todo ello se requiere procesar **gran cantidad de datos urbanos** (sensores IoT, cámaras de tráfico, datos meteorológicos, telecomunicaciones móviles, etc.) en tiempo real o cuasi-real. Un sistema de ciudad inteligente avanzado podría ingerir *millones de eventos por minuto*. La HPC (ya sea un supercomputador municipal o recursos cloud distribuidos) permite hacer análisis y simulaciones veloces con esos datos.

Ejemplos reales:

- *Virtual Singapore:* Singapur ha desarrollado un gemelo digital 3D completo de la ciudad-estado. Integra datos BIM de edificios, tráfico, clima, etc., para permitir simulaciones urbanas. Empresas y agencias pueden probar en Virtual Singapore desarrollos urbanos antes de construirlos.
- *Barcelona:* El Barcelona Supercomputing Center (BSC) colabora con la ciudad para simular tráfico y consumo energético. Usan MareNostrum para modelar escenarios como reorganización de autobuses o el impacto de aislantes en edificios sobre la demanda eléctrica.

- *Smart Dublin:* En Dublín, Irlanda, hay plataforma abierta de datos urbanos y simulaciones HPC para cosas como optimizar rutas de camiones de basura, monitorear uso de bicicletas compartidas y predecir inundaciones en ciertos puntos con modelos HPC combinando lluvia + marea.

El rol de la HPC aquí es ser el "*cerebro analítico*" de la ciudad:

- Procesa datos en streaming (ej. detecta en video en vivo si hay un accidente en X calle y alerta al centro de control).
- Ejecuta **algoritmos de predicción** complejos (modelos de tránsito, predicción de contaminación) continuamente.
- Permite a los gestores públicos ensayar políticas en la computadora y basar decisiones en evidencia cuantitativa (urbanismo basado en simulaciones en vez de intuición).

Beneficios esperados de aplicar HPC en ciudades:

Área urbana	Mejoras gracias a simulación HPC
Movilidad	Menos atascos, tiempos de viaje reducidos, rutas de transporte público optimizadas según demanda.
Medio ambiente	Reducción de emisiones (p. ej. optimizando semáforos para menos paradas), mejor gestión del agua (evitando inundaciones), monitoreo de calidad del aire en tiempo real.
Urbanismo	Planificación basada en evidencia: se prueban planes urbanos en el gemelo digital para ver impacto en densidad, flujos, etc., antes de invertir millones en obras.
Seguridad	Respuesta más rápida a emergencias (simulaciones dicen dónde posicionar recursos, o cómo evacuar de forma óptima). Policía y bomberos usan análisis de datos para patrullaje predictivo o prevención de incendios.
Energía	Optimización del consumo en edificios (smart grids deciden cómo redistribuir electricidad, cortar picos, almacenar excedentes renovables).

En definitiva, HPC habilita las "centrales cerebrales" de las *smart cities*. Muchas de estas simulaciones son imposibles sin computación seria: por ejemplo, simular microscopicamente el movimiento de cada vehículo y persona en una ciudad grande es una tarea HPC (podría involucrar millones de agentes, resolviendo en pasos de tiempo pequeños).

Conclusión del bloque 8: Tanto la IA generativa como las simulaciones urbanas demuestran cómo la HPC está saliendo del ámbito puramente científico para **transformar actividades humanas cotidianas**: desde cómo creamos arte hasta cómo gestionamos nuestras ciudades. Estos casos nos recuerdan que detrás de innovaciones muy visibles (ChatGPT respondiendo, o una ciudad con tráfico fluido) hay un poder de cómputo enorme operando tras bambalinas. A su vez, plantean reflexiones: ¿estamos preparados para las implicaciones de estas tecnologías (en empleo, privacidad, seguridad)? ¿Cómo aseguramos que sus beneficios (mejoras en vida cotidiana) lleguen a todos y no solo a unos pocos entornos privilegiados?

Preguntas de reflexión (Casos de uso): **¿Cuál de estos avances te parece más impactante en la vida diaria: la IA generativa o las ciudades inteligentes?** ¿Ves algún riesgo en su adopción masiva (por ejemplo, dependencia de la IA para tareas creativas, o vigilancia excesiva en ciudades hiperconectadas)? Y pensando en el futuro próximo, **¿qué otro gran problema o sector** (salud, cambio climático,

9. Conclusión y preparación para la reflexión posterior

Una mirada al futuro cercano: La sesión nos mostró que la Computación de Alto Rendimiento no es solo un tema técnico; es *estratégico, social y ambiental*. Los avances en HPC están redefiniendo industrias, acelerando la ciencia y alimentando una revolución basada en datos… **pero** también están consumiendo recursos a un ritmo sin precedentes. Enfrentamos un momento de inflexión donde hay que equilibrar las increíbles oportunidades de la HPC con los desafíos que conlleva.

Cerrando, más que enfocarnos en la tecnología per se, debemos mirarnos a **nosotros mismos** como futuros profesionales de IA/HPC y ciudadanos globales:

¿Qué retos nos interpelan más como futuros profesionales?

Como dijo Einstein: *"No podemos resolver problemas con el mismo tipo de pensamiento que usamos cuando los creamos."* Hoy tenemos varios retos que nos llaman a pensar diferente:

- **Eficiencia energética:** ¿Podremos diseñar sistemas que sigan creciendo en potencia sin crecer en consumo? La ley de Moore tal vez murió, pero la necesidad de ser *verdes* no.
- **Equidad en acceso:** Si la HPC se concentra, ¿quién se beneficia de los avances? ¿Ampliará la brecha digital y económica? Nos toca idear formas de democratizarla (ej. plataformas abiertas, colaboraciones).
- **Algoritmos responsables:** Cada vez más decisiones críticas las toman sistemas computacionales (desde préstamos bancarios hasta diagnósticos médicos). Nosotros seremos quienes diseñen esos algoritmos: ¿cómo aseguramos que sean **transparentes, justos y auditables**?
- **Impacto ambiental:** La industria tech pasó de ignorar su huella a tratar de mitigarlo. Como ingenieros, ¿estamos dispuestos a repensar el ciclo de vida completo de la tecnología (fabricación de equipos, energía, residuos electrónicos)? Quizá debamos incluir "criterios verdes" en cada proyecto, por diseño, no como añadido después.

¿Es sostenible el modelo actual de crecimiento computacional?

Hoy duplicamos capacidad de cómputo en meses en algunas áreas (ej. número de parámetros de IA). Pero:

- Los centros de datos ya consumen 1–2% de la electricidad mundial, y subiendo.
- Entrenar un solo modelo de IA grande puede emitir tanto CO_2 como varios automóviles en toda su vida útil.

- Muchos países no pueden proveer la infraestructura energética ni digital para seguir ese ritmo de crecimiento.

☞**Reflexión clave:***Si el costo energético y ambiental sigue aumentando así, ¿podremos seguir creciendo sin comprometer el planeta?* Quizá la respuesta esté en un cambio de paradigma: **computación cuántica** (promete calcular más con menos energía en ciertos problemas), **hardware neuromórfico**, o simplemente **frenar la carrera** y optimizar mejor lo que tenemos. Es un debate abierto en la comunidad: buscar **soberanía computacional sostenible**.

¿Qué papel queremos jugar en esta transformación?

Esta pregunta es personal y profesional a la vez. Tenemos en nuestras manos un poder (conocimiento en IA/HPC) que poca gente tiene, y como dice el dicho, con gran poder viene gran responsabilidad. Podemos elegir:

- Ser simplemente usuarios/consumidores de las tecnologías cada vez más potentes que otros creen, **o**
- Liderar un cambio para que la computación sea más **justa, accesible y sostenible**.

Como estudiantes y pronto ingenieros/científicos, podemos orientar nuestras carreras a:

- Innovar técnicamente *y* al mismo tiempo influir en políticas o prácticas para abrir el acceso a HPC (por ejemplo, contribuir a proyectos open-source de HPC, o impulsar clusters compartidos en nuestra universidad).
- Abogar por la **transparencia**: si trabajamos en big tech, empujar internamente por documentar mejor los modelos, por externalizar más conocimiento a la comunidad científica.
- Mantener siempre presente la ética: preguntarnos en cada proyecto de HPC/IA los potenciales impactos negativos y cómo mitigarlos desde el diseño.

En resumen, el futuro de la HPC/IA no está escrito; lo escribiremos nosotros. Podemos ser agentes de un cambio positivo para que la computación de alto rendimiento beneficie al mayor número de personas, minimizando daños.

Preguntas clave para debate en clase: (para consolidar y discutir colectivamente)

- *¿Debería el acceso al poder computacional avanzado considerarse un derecho público (similar al acceso a internet)?* ¿O es algo que pertenece al ámbito privado de quien pueda pagarlo?
- *¿Cómo evitamos que el desarrollo de la IA dependa de unas pocas empresas o países?* ¿Hacen falta iniciativas tipo "ONU de la supercomputación"?
- *¿Qué límites éticos habría que imponer al uso de HPC?* (Ej: ¿Debería prohibirse entrenar ciertos modelos muy riesgosos? ¿Regular cuánta energía puede usar un entrenamiento, o cuántos datos personales?)
- *Legal y éticamente, ¿estamos preparados para un mundo con IA potenciadas por HPC casi ilimitado?* (Por ej., si llega IA general superinteligente, ¿tenemos

marcos para controlarla? Si un país desarrolla una IA militar muy superior con su HPC, ¿cómo responde el mundo?)

- *En un mundo con HPC ubicua, ¿cómo garantizar que la privacidad y los derechos de individuos y comunidades sigan respetados?* (¿Necesitaremos "pausas" o moratorias al estilo que algunos piden para la IA?)

Actividad breve de cierre (preparación para la práctica de la semana 13):

Para terminar, les proponemos una **reflexión personal escrita (≤150 palabras)** sobre tres preguntas, que usaremos de base en la próxima sesión práctica:

1. **¿Qué reto de los comentados hoy consideras más urgente abordar y por qué?** (Puede ser técnico, ético, ambiental, etc.)
2. **¿Qué idea o dato te ha impactado más de la sesión?** (Algo que no sabías o no habías dimensionado antes).
3. **¿Qué compromiso podrías asumir tú como futuro profesional de IA/HPC para contribuir a un modelo de computación más responsable y sostenible?** (Por pequeño que sea: ej. divulgar, especializarte en optimización energética, evitar ciertos usos, etc.)

Por favor, escriban sus respuestas y tráiganlas la semana próxima en que tengamos clase. Serán el punto de partida para una actividad de diseño en la que imaginaremos soluciones concretas alineadas con un modelo de computación de alto rendimiento **ético y sostenible**. ¡Gracias por vuestra atención y participación!