IA: Uso e Impacto en Data Centers

Alma Esparza
Consultor IoT | **Grupo Salinas**

ICT SUMMIT
CONFERENCIA & EXHIBICION
MÉXICO 2025

BICSI CALA

IA - Uso e Impacto en Data Centers

Dra. Alma Laura Esparza Maldonado

Agenda

- Demanda creciente de IA
- Demanda energética
- Consumo Energético de IA: Data Centers.
- Gestión térmica de IA en Data Centers.
- Aplicaciones de IA en Data Centers
- Conclusión

Demanda creciente de cómputo de IA

- Entre 2012 y 2018, el cómputo requerido para entrenar modelos de IA se duplicó cada 3.4 meses (OpenAI), creciendo más de 300,000 veces en total.

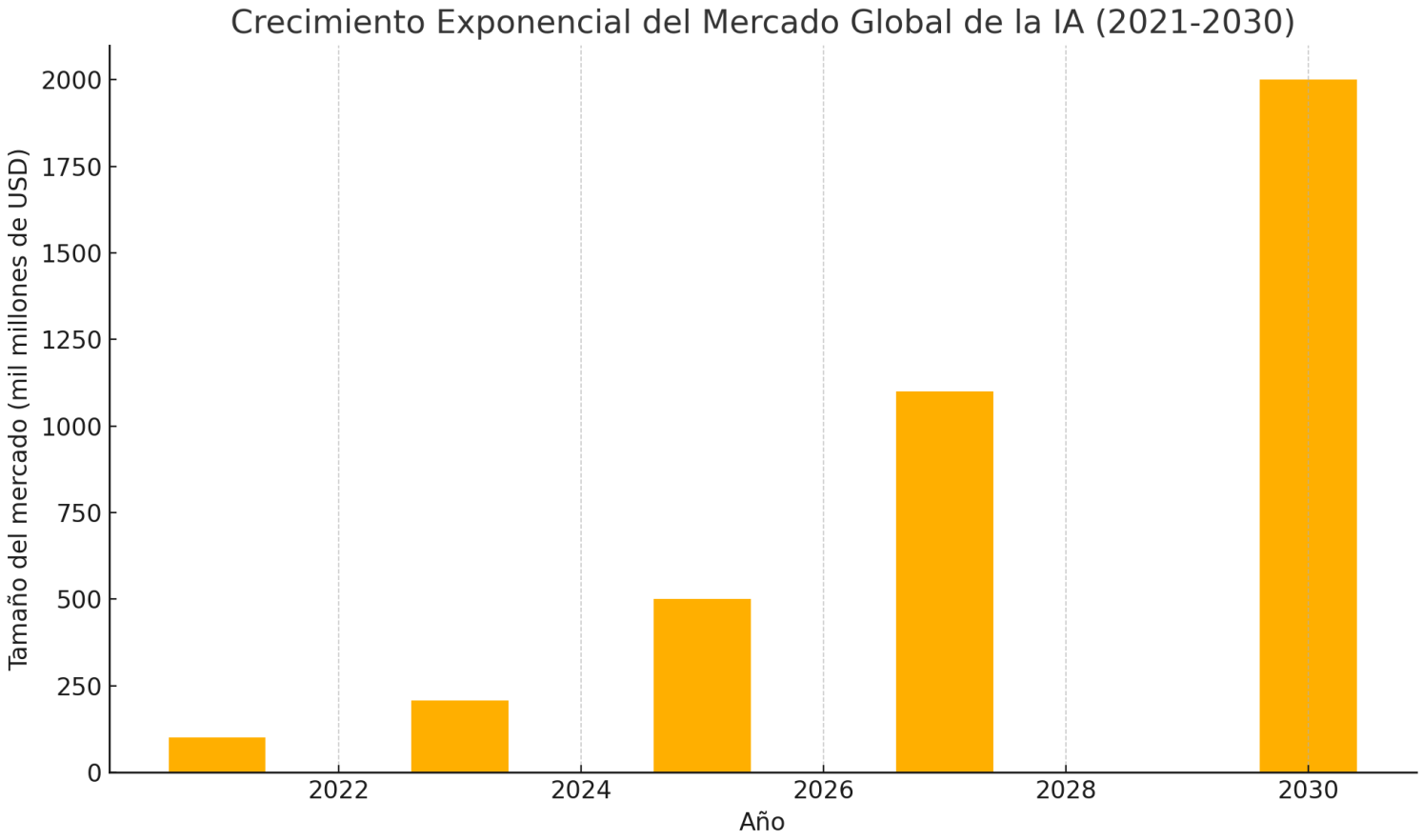
- GPT-3: 175 mil millones de parámetros, mientras que GPT-4 y Gemini lo superan ampliamente aproximadamente 1 billón.

- Inversión proyectada: de $100 mil millones (2021) a $300 mil millones (2026) (Statista)

Demanda creciente de cómputo de IA

Modelo	Año	Parámetros Estimados	Características Relevantes	Referencia
GPT-2	2019	1.5 mil millones	Primer salto significativo en lenguaje natural	GPT-2: Better Language Models
GPT-3	2020	175 mil millones	Fundacional para tareas multitarea	Language Models are Few-Shot Learners
PaLM (Google)	2022	540 mil millones	Multilingüe, de código cerrado	PaLM: Scaling Language Modeling
Claude	2023	No revelado	Foco en alineación y seguridad	Anthropic FAQ: Claude 2 release
LLaMA 2	2023	7B a 65B	Código abierto, eficiente	Meta AI: Introducing LLaMA 2
Mixtral	2023	MoE con 12.9B activos	Modelo mixto eficiente, libre	Mistral AI blog: Mixtral of Experts
GPT-4	2024	~1 billón (estimado)	Arquitectura MoE, multimodal, no confirmados	SemiAnalysis y The Decoder, estiman entre 500B y 1T usando MoE (Mixture of Experts)
Gemini 1.5	2024	Comparable a GPT-4+	Ventana de contexto de hasta 1 millón de tokens	Google DeepMind blog (2024): Gemini 1.5 models

Demanda creciente en el mercado de IA



Demanda Energética: Entrenamiento

- Según OpenAI y Google Research, entrenar un modelo como GPT-3 (175 mil millones de parámetros) puede consumir entre 1.2 y 3 GWh (gigavatios-hora) en total.

- Eso implica que cada parámetro en GPT-3 requiere aproximadamente entre:
  - $\frac{1.2\ GWh}{175*10^9} = 6.9 \times 10^{-6}\ Wh \rightarrow Mínimo\ estimado$
  - $\frac{3\ GWh}{175*10^9} = 1.7 \times 10^{-5}\ Wh \rightarrow Máximo\ estimado$

- Cada parámetro consume entre ~7 a ~17 μWh (microvatios-hora) durante su entrenamiento.

Bicsi CALA

Demanda Energética: Inferencia

- Inferencia con GPT-3 para generar una respuesta puede consumir entre 0.001 y 0.01 kWh.
- El costo energético por token puede ir desde 0.1 a 1.0 mWh/token, dependiendo del modelo, longitud de respuesta y hardware (según HuggingFace y DeepMind).

Plataforma / Modelo	Consultas por Minuto (estimado)	Consumo Mínimo (kWh/min)	Consumo Máximo (kWh/min)	Referencias
OpenAI / ChatGPT (GPT-3.5/4)	~185,000	185	1850	Basado en tráfico estimado de 1.6B visitas/mes (SimilarWeb, 2024)
Google Gemini (Bard)	50,000 - 150,000	50-150	500 – 1500	Estimado por uso global de Gemini (Google DeepMind)
Plataforma Empresarial Privada	100 - 10,000	0.1 – 10	1 – 100	Varía según arquitectura y carga de usuarios
Otros LLMs (Claude, LLaMA, etc.)	Variable	Variable	Variable	Depende del despliegue local o en nube

Bicsi CALA

Consumo Energético de IA: Data Centers

- En 2024, los Data Centers consumieron aproximadamente 415 TWh, representando el 1.5% del consumo eléctrico global.
- Para 2030, se proyecta un consumo de 945 TWh, más del doble del valor actual.
- China y Estados Unidos concentran cerca del 80% del crecimiento en esta demanda energética.

Gestión térmica de IA: Data Centers

- En 2024, consumo total: 415 TWh → Refrigeración (30-50%): 125 – 208 TWh

- En 2030, proyección total: 945 TWh → Refrigeración (30-50%): 284 – 472 TWh

Gestión térmica de IA: Data Centers

Tecnología	Ventajas	Referencias
Liquid Cooling (Direct-to-Chip)	Alta eficiencia térmica, ideal para GPUs/TPUs de alto TDP.	Intel, ASHRAE, NVIDIA
Inmersión líquida (1 o 2 fases)	Reduce hasta 95% el consumo de refrigeración. Silencioso.	Microsoft, Schneider Electric
Rear Door Heat Exchangers	Compatible con racks actuales. Ahorra energía sin rediseño.	IBM, Vertiv
Agua caliente directa	Permite reutilizar calor. Bajo impacto ambiental.	Lenovo, RISE Institute
Evaporación indirecta	Alta eficiencia en climas secos. Bajo costo energético.	Google, Meta (Facebook)

Bicsi CALA

Aplicaciones de IA en Data Centers

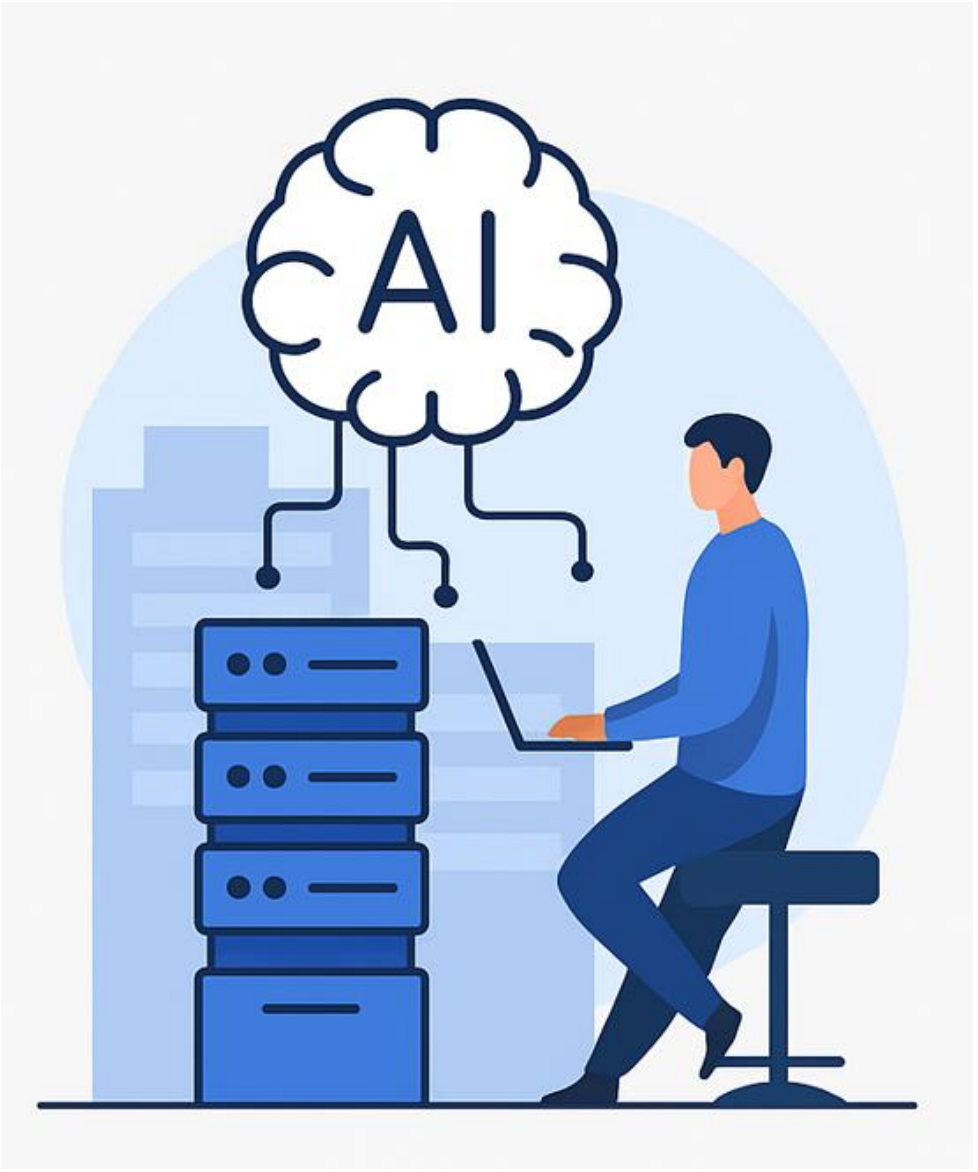
- Optimización del Consumo Energético
  - Google usa IA de DeepMind para reducir hasta 40% la energía en refrigeración.
  - Reprogramación de cargas de trabajo según disponibilidad de energía limpia.

- Mantenimiento Predictivo
  - Microsoft reduce 25% el downtime con modelos de predicción de fallas.
  - IBM Watson predice fallos en servidores mediante análisis de sensores.

- Automatización y Monitoreo Inteligente
  - IA detecta anomalías, automatiza reinicios y balancea cargas.
  - Herramientas como AIOps y Watson optimizan la operación 24/7.

Conclusiones

- La IA ha crecido de manera exponencial aumentando significativamente los requerimientos de energía
- Los Data Centers ha duplicado su demanda energética y requieren mejorar su gestión térmica
- La IA optimiza el consumo energético y minimiza las fallas mediante aplicaciones avanzadas.



Referencias

- [1] OpenAI, “AI and Compute,” 2018.
- [2] Statista, “AI Market Size Worldwide,” 2021–2023.
- [3] Scientific American, “AI Will Drive Doubling of Data Center Energy Demand by 2030,” 2024.
- [4] Google DeepMind, “Reducing Cooling with AI,” 2020.
- [5] Microsoft Azure, “Predictive Maintenance in Data Centers,” 2023.
- [6] Uptime Institute, “Global Data Center Survey,” 2023.
- [7] IBM, “Watson AIOps Overview,” 2022.
- [8] HuggingFace & DeepMind, “Energy Cost per Token for Inference,” 2023.
- [9] ASHRAE, “Thermal Guidelines for Data Processing Environments,” 2023.
- [10] Schneider Electric, “Liquid Cooling for AI Systems,” 2024.