Primeros pasos con IA

Cómo Lenovo e Intel están potenciando aplicaciones prácticas para IA hoy.

intel. XEON PLATINUM

Lenovo ThinkSystem SR650 V3 servidores construidos con la 5ª generación de los procesadores escalables Intel Xeon, diseñados para IA.

Smarter technology for all

Lenovo

Sumario

El rápido crecimiento de la IA

La inteligencia artificial (IA) ha hecho avances tremendos desde sus días pioneros en la década de 1950. Los primeros y rigurosos enfoques de diseño para el aprendizaje estadístico y el análisis predictivo realizados en computadoras dieron paso a las primeras instancias de aprendizaje automático en la década de 1980, cuando los algoritmos fueron enseñados a reconocer relaciones y construir modelos de sistemas complejos.

La llegada de grandes redes neuronales en los años 2000 pavimentó el camino para expansiones masivas de capacidad computacional e introdujo la capacidad de gestionar y analizar grandes cantidades de datos complejos y patrones abstractos.

Desde una perspectiva empresarial, el potencial para obtener insights, reducir cargas de trabajo y acelerar la productividad parece casi ilimitado, y las empresas están investigando activamente maneras de poner la IA a trabajar.

80%

de los CIOs de hoy tienen la tarea de investigar y evaluar posibles implementaciones de IA en sus soluciones tecnológicas..[1]

Habilitando la IA en todos lados

Abordar iniciativas de IA puede ser desalentador. Históricamente, la IA ha estado solo en el dominio de los motores de búsqueda, instituciones financieras y la investigación científica. Además del costo de adquirir el hardware necesario, en muchos casos los centros de datos existentes no soportan el poder adicional y los requisitos de enfriamiento, lo que necesita más inversiones de capital y tiempo.

La buena noticia es que la introducción de modelos basados en IA entrenados en datos públicos ha reducido las barreras para que las organizaciones implementen soluciones de IA.

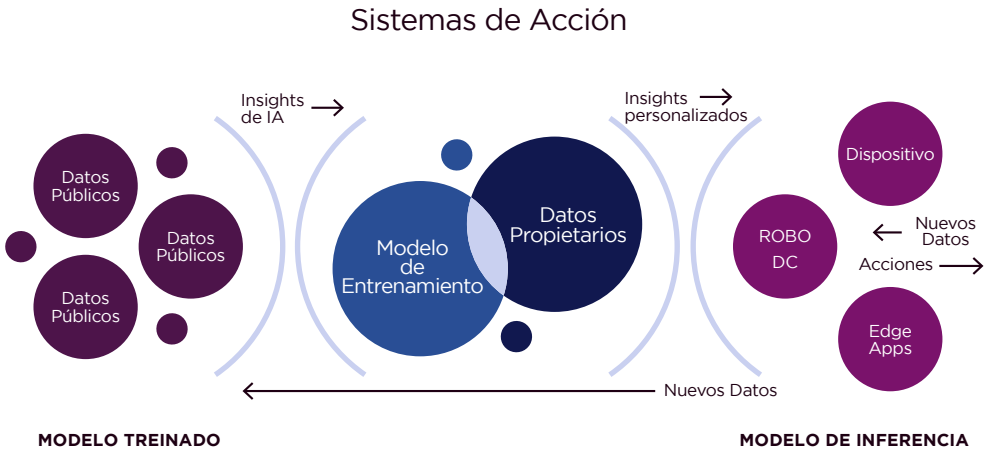
Lenovo e Intel están utilizando su larga asociación para trabajar, entregando soluciones que permiten a las empresas aprovechar todo el gran trabajo que se está haciendo de manera rápida y práctica en maneras que entregan resultados medibles.

Vamos a comenzar con algunos conceptos básicos

En los términos más simples, la IA se define como cualquier sistema de automatización que simula la inteligencia humana aprendiendo en el trabajo. La IA se implementa en dos fases:

1. **Entrenamiento o desarrollo de modelo:** Esta es la fase donde los científicos de datos desarrollan y optimizan modelos fundamentales con un conjunto de datos seleccionado.

2. **Inferencia:** Aplicar nuevos datos a un modelo entrenado para derivar nueva información y acelerar la automatización.



Desarrollo del modelo de entrenamiento

El entrenamiento de la IA se logra a través de un proceso llamado Machine Learning (ML), en el cual un modelo es entrenado con base en parámetros específicos que definen la tarea (por ejemplo, color, formas y bordes) y usa técnicas como agrupamiento, regresión y redes neuronales para extraer esos elementos de enormes cantidades de datos para desarrollar predicciones. A partir de ahí, el modelo continúa consumiendo y analizando datos mientras mejora sus capacidades de toma de decisiones futuras.

Los conjuntos de datos usados para el entrenamiento fundamental han crecido a una escala que requiere grandes cantidades de poder computacional especializado, funcionando en paralelo en miles de procesadores, lo que históricamente ha estado bajo el dominio exclusivo de organizaciones académicas, financieras y gubernamentales.

La cantidad de poder computacional necesario para entrenar los mayores modelos de IA es de

3 a 10 meses.[2]

Introduciendo la inferencia de IA

La inferencia de IA involucra la utilización de modelos entrenados existentes y su aplicación a nuevos conjuntos de datos propietarios para tareas específicas de aplicación. Los resultados e insights son entonces adaptados para nuevas aplicaciones que están diseñadas para ofrecer experiencias más precisas y relevantes.

La inferencia expande el aprendizaje ya realizado, por lo tanto, las demandas de procesamiento para generar predicciones e insights son significativamente menores que aquellas necesarias durante el entrenamiento inicial.

Con las demandas reducidas de procesamiento, se están abriendo puertas para que empresas y organizaciones de todos los tamaños aprovechen el poder de la IA para una amplia gama de aplicaciones.

Vea cómo Lenovo e Intel están acelerando la Industria 4.0 con inspecciones visuales asistidas por IA y análisis de datos más rápidos.
Descubre más.

El número de empresas usando IA ha crecido

300% en **5 años.**[3]

Las organizaciones no necesitan desarrollar modelos de entrenamiento fundamentales, esto acelera dramáticamente el desarrollo y les ayuda a mover hacia aplicaciones reales más rápidamente. Además, como el enfoque solo necesita de nueva información que se aplique al modelo, las demandas de datos y procesamiento pueden extenderse más allá del centro de datos. Esto significa que la inferencia puede ocurrir donde se recolectan los datos, incluso en el borde.

Esto es importante porque permite funciones críticas en tiempo real sin penalizaciones de latencia al enviar datos de ida y vuelta hasta la nube, como en los sistemas autónomos encontrados en vehículos autodirigidos o en fábricas automatizadas.

El modelo de inferencia de IA entrenado funciona solo con los datos necesarios para tomar decisiones, lo que acelera el proceso de decisión y reduce la necesidad de mover grandes cantidades de datos a través de las redes.

Por ejemplo, en un ambiente de manufactura, los servidores de borde que ejecutan modelos de inferencia de IA pueden usar visión computacional (basada en modelos entrenados existentes) para identificar defectos, tomar decisiones y tomar las medidas apropiadas (usando datos locales propietarios) para resolver el defecto manteniendo la línea de producción.

Desbloquee insights en sus datos más rápidamente

A medida que las aplicaciones de IA evolucionan, la tecnología que las soporta también está evolucionando para adaptarse a estas nuevas expectativas, acelerando y posibilitando el despliegue de la IA en cada etapa, desde el edge hasta la nube. Lenovo e Intel han unido fuerzas para entregar soluciones construidas específicamente para aplicaciones de inferencia de IA.

La más reciente generación de servidores ThinkSystem, como el ThinkSystem SR650 V3, están construidos sobre los procesadores escalables Intel® Xeon® de 5ª generación, diseñados para IA. La aceleración integrada ofrece un aumento de rendimiento para tareas de inferencia de IA y reduce los requisitos de energía y enfriamiento, lo que significa que los servidores ThinkSystem SR650 V3 pueden ser implementados en centros de datos existentes en lugar de construir nuevos.

Hasta
2.7x más rendimiento de IA
en cualquier otro CPU[4] con procesadores escalables Intel® Xeon® de 5ª generación con Intel® AI Engines.

Hasta
14x más rendimiento de inferencia en tiempo real para detección de objetos comparado con los procesadores Intel® Xeon® de 3ª generación.[5]

Además, Lenovo ofrece un portafolio líder de soluciones de edge AI, como el **ThinkEdge SE350 V2** y **ThinkEdge SE360 V2**, usando procesadores Intel® Xeon® D para proporcionar insights en tiempo real. Los recursos de computación mejorados y los diseños flexibles de implementación soportan múltiples tipos de cargas de trabajo de IA con rendimiento avanzado y diseños eficientes. Con la IA en el borde, las organizaciones pueden capitalizar en información dinámica en tiempo real y entregar automatización, remediación e insights donde son más accionables — directamente en el frente.

Aproveche la expertise en transformación de IA

Diseñar e implementar modelos de inferencia de IA que entreguen insights confiables y accionables requiere un conjunto muy específico de habilidades y extrema atención a los detalles.

El **AI Discover Center of Excellence de Lenovo** reúne expertos de Lenovo y de Intel para ayudar a sus desarrolladores a crear y acelerar la entrega de aplicaciones de IA y modelos de inferencia.

Nuestros expertos en IA conducen una amplia serie de talleres para proporcionar evaluaciones de negocios integrales, evaluaciones de TI y blueprints de diseño documentados.

Ingenieros técnicos, socios y científicos de datos optimizan sus códigos de IA usando frameworks de código abierto para funcionar en servidores ThinkSystem con hardware y software Intel.

Podemos ayudarte a aprovechar el conjunto completo de recursos de Intel, como la herramienta OpenVINO™ y la oneAPI Deep Neural Network Library (oneDNN) para simplificar la implementación de inferencia de aprendizaje profundo para cientos de modelos preentrenados.

Lenovo también ofrece una amplia serie de talleres de Servicios Profesionales de Lenovo para acelerar su viaje de transformación de IA.

intel. XEON PLATINUM

Lenovo

El kit de herramientas OpenVINO™

Las barreras para la adopción de IA generalmente incluyen la necesidad de modelos grandes, optimizados, diversificados, una amplia gama de arquitecturas de XPU (a menudo implementadas juntas), y un vasto ecossistema de frameworks de software API para elegir. Implementar IA puede ser un proceso difícil y que consume tiempo, involucrando muchas elecciones de proveedores.

Con todas estas complejidades, los conceptos comprobados a menudo no llegan a la producción, creando un “cementerio de POCs.”

Esas barreras necesitan ser derribadas para crear oportunidades, y es eso lo que OpenVINO™ hace al ofrecer un kit de herramientas de código abierto que soporta una amplia gama de arquitecturas de XPU y frameworks de software de IA.

intel. XEON PLATINUM

OpenVINO™

Beneficios:

1. Amplia accesibilidad para múltiples arquitecturas de XPU a través de un modelo de código abierto.

2. Una solución de inferencia de IA accesible y eficiente que reduce los costos de adopción y aplicación de la tecnología de IA desde la nube hasta PCs locales.

3. Una arquitectura abierta que permite la colaboración a través de todo el ecosistema, desde científicos de datos creando modelos hasta desarrolladores aplicando frameworks de aprendizaje profundo en una variedad de mercados verticales, aprovechando funciones de IA múltiples como procesamiento de lenguaje natural, sistemas de recomendación y IA generativa.

Emparejado con la **Plataforma Edge de Intel**, soluciones nativas de edge pueden ser construidas para acelerar iniciativas de edge IA con recursos de optimización de modelo, entrenamiento y desarrollo de aplicaciones.

Las empresas también pueden embarcar y gestionar de forma segura una flota de nodos de edge, aprovechando los componentes más adecuados y costo-efectivos, ya sea en ambientes nuevos o ya existentes, en asociación con nuestro ecosistema inigualable para un costo total de propiedad más bajo.

Vea cómo Lenovo e Intel están agilizando la adopción de IA con OpenVINO™. **Descubre más.**

Acelere su viaje con soluciones comprobadas de implementación

Cuando llegue la hora de implementar su solución de inferencia de IA, el programa Lenovo AI Innovators simplifica el proceso con soluciones comprobadas usando software ISV de mejor categoría en infraestructura optimizada para IA de Lenovo y de Intel.

Lenovo e Intel construyen, prueban y validan soluciones de inferencia de IA con un ecosistema de socios de Innovadores en IA para asegurar implementaciones fluidas y óptimas que te mantengan dentro del cronograma y presupuesto.

- Solución de gestión remota de **Nybl**
- Solución de inspección visual asistida por IA de **byteLAKE**
- Soluciones de visión computacional, mantenimiento predictivo y detección de anomalías de **Guise AI**
- Solución de Análisis de Filas y Multitudes de **WaitTime**
- Solución de **Sunlight.io** que acelera la transformación digital de restaurantes y drive-thrus
- Solución de inteligencia industrial de **Smartia** que conecta y transforma datos en insights accionables

Continuamos monitoreando, evaluando y construyendo relaciones con socios ISV a medida que sus soluciones evolucionan.

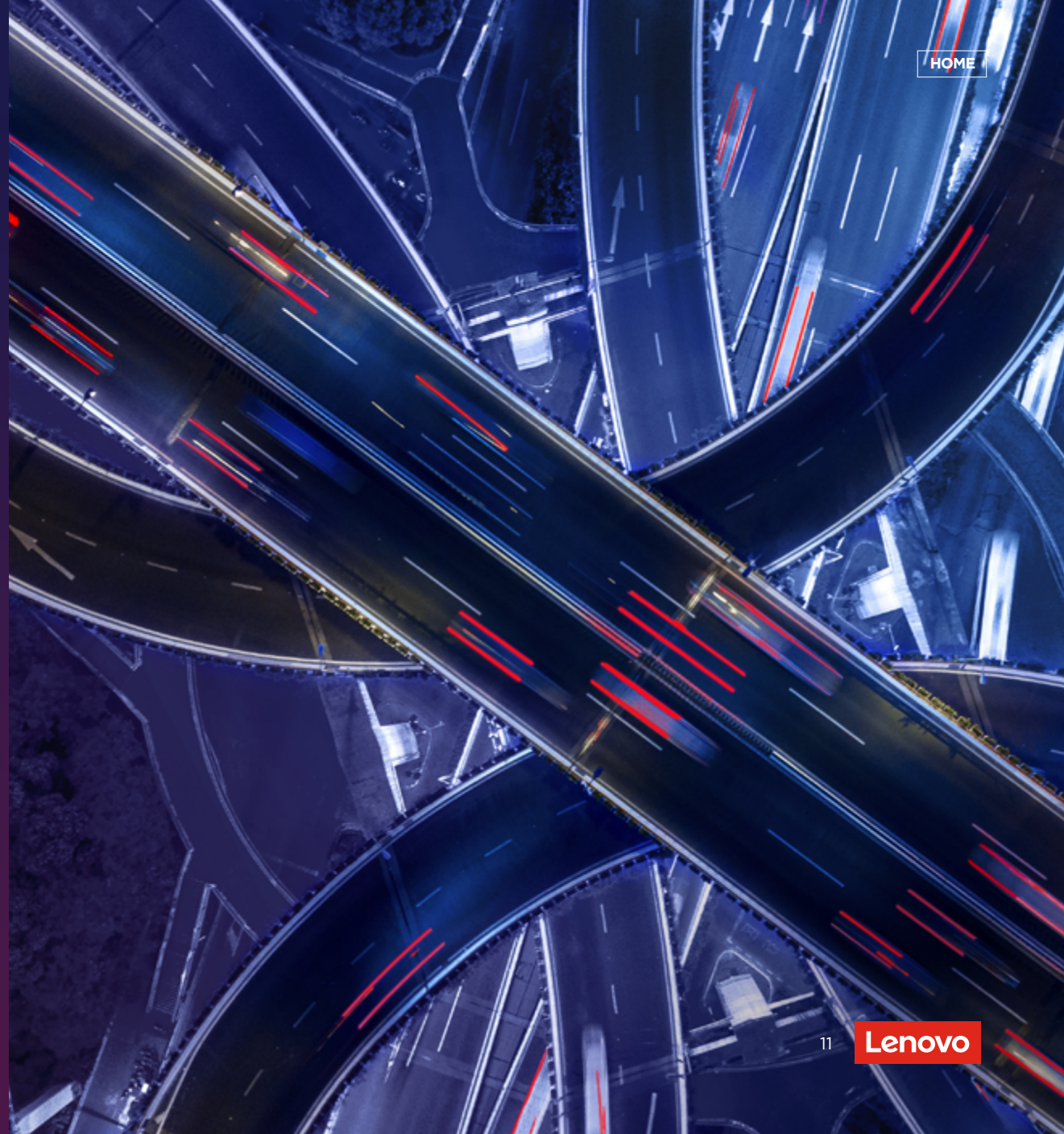
Estudio de caso:
IA está transformando las experiencias de los espectadores

Lenovo y WaitTime presentaron una solución innovadora para locales de eventos, transformando la experiencia de los espectadores de la Fórmula 1® con tecnología de punta. Al combinar 18 cámaras estratégicamente instaladas en el autódromo Circuit of The Americas (COTA), con la tecnología patentada de IA de WaitTime en servidores Lenovo ThinkEdge, alimentados por procesadores Intel® Xeon®, los operadores del COTA pueden monitorear meticulosamente las multitudes y filas de personas.

"Esta plataforma de análisis de datos en tiempo real proporciona insights valiosos, permitiendo que los operadores comprendan dinámicamente cómo las multitudes están creciendo y cambiando," dijo Zachary Klima, fundador y CEO de WaitTime.

"Estas informaciones instantáneas nos empoderan para hacer ajustes en tiempo real en las operaciones y estrategias de ingresos, garantizando una experiencia óptima y continua para los espectadores, al mismo tiempo maximizando la eficiencia y los ingresos para el evento."

Puede leer más sobre la solución **aquí.**



Obtenga la flexibilidad para escalar sin problemas

Implementar inferencia de IA requiere mucho menos gastos iniciales comparado con la construcción y entrenamiento de modelos fundamentales desde cero, pero aún hay costos que deben ser considerados para hardware, software y servicios.

Lenovo TruScale ofrece la flexibilidad de un modelo de pago escalable conforme al uso para sus iniciativas de inferencia de IA — proporcionando acceso a expertise que acelera sus iniciativas.

El modelo OpEx reduce la inversión inicial y escala con las necesidades de negocios en cambio, permitiendo que lleve proyectos desde el concepto hasta la implementación y más allá.

intel. XEON PLATINUM

Implementación más rápida

Al reemplazar los requisitos de aprobación de CapEx y cambiar a un modelo OpEx, TruScale puede aumentar la flexibilidad y acelerar los tiempos de adquisición e implementación.

Opciones escalables

Elija entre un contrato fijo o consumo medido para satisfacer las necesidades de su organización.

Expertise y servicios integrados de IA

Aproveche los servicios especializados de Lenovo para cerrar brechas de habilidades y recursos y asegurar el éxito de la implementación. Además, los gerentes de éxito del cliente dedicados de Lenovo pueden ayudar a facilitar y coordinar con los recursos de Lenovo.

Esta flexibilidad no solo hace más fácil para una gama más amplia de organizaciones aprovechar la IA, sino que también anticipa el futuro de la tecnología y elimina el riesgo de obsolescencia conforme la tecnología evoluciona.

IA con un ojo en la sostenibilidad

El aumento de la capacidad computacional necesaria para entrenar y operar modelos de IA resulta en mayor consumo de energía y generación de calor, lo cual sigue siendo una fuente de preocupación global.

A medida que la IA se integra más en los aspectos del día a día, el aumento en el poder computacional necesario solo tiende a crecer.

Por ejemplo, una búsqueda típica en Google consume **menos de 0.3 watt-horas (Wh)** por solicitud. Añadir un gran modelo de lenguaje que reaccione a la solicitud puede aumentar ese consumo a algo entre **7Wh y 9Wh** por pedido. Considerando el volumen actual de búsquedas de Google, si cada búsqueda incluyera un componente de IA, Google solo podría consumir cerca de **30 terawatt-horas (TWh)** por año, aproximadamente el equivalente al consumo del país de Irlanda.[7]

Entrenar un solo modelo de IA puede producir

626,000

libras de CO_2 equivalente.[6]

Lenovo e Intel estão comprometidos com soluções de inferência de IA sustentáveis, eficientes em termos energéticos e ambientalmente responsáveis.

Los procesadores escalables Intel® Xeon® de 5ª generación son los más sostenibles jamás ofrecidos por Intel para centros de datos, entregando hasta 10 veces más desempeño por watt con aceleradores dirigidos para cargas de trabajo específicas.[8] Y pueden ser implantados en centros de datos existentes sin requisitos adicionales de energía o enfriamiento.

En el centro de datos, la tecnología de medición TruScale puede ayudarle a monitorear consumo de energía, utilización y temperatura para gestionar el uso y costos de energía de manera más eficiente. Además, nuestro software Runtime Energy Aware (EAR) y el xClarity Energy Manager ayudan a optimizar el desempeño con un nivel bajo de consumo de energía, optimizando estados de energía, apagando componentes no usados y direccionando cargas de trabajo hacia los recursos más eficientes.

Optimizar su centro de datos con Lenovo TruScale puede ayudar a reducir emisiones de CO_2 hasta un 20%.[9]

Búsqueda Google <0.3Wh

Búsqueda de Google con tecnología de IA 7–9Wh

Todas las búsquedas de Google IA 30TWh por año

Un enfoque más inteligente para la inferencia de IA en cualquier lugar

La inferencia de IA tiene un tremendo potencial para acelerar el crecimiento de los negocios, reducir cargas de trabajo y optimizar la eficiencia para empresas en todas las industrias.

No importa dónde se encuentre en su viaje para implementar soluciones de IA en su organización, Lenovo e Intel están listos para ayudar con soluciones hechas a medida, expertise líder en el sector y socios de la mejor clase.

Visite la **página de la Alianza Intel AI** para aprender más.

Fuentes

1. 1Foundry, "State of the CIO Survey 2024"
2. Accenture, "Technology Vision 2023", marzo de 2023
3. Tidio, "Los 10 datos esenciales de estadísticas de IA que necesita saber para 2023", octubre de 2023
4. Basado en ganancias de desempeño de 119% a 269% con las extensiones Intel® Advanced Matrix Extensions (Intel® AMX) para inferencia en GPU-T, LLAMA-2 128, DL RM, DiscBERT, BERT-Large, y ResNet50v1.5 comparado a AMD EPYC 9654 y 9754. Vea A201, A202, A208-A211 en intel.com/processors/claims. Los procesadores escalables de la 5ª generación Intel Xeon Scalable presentan resultados que pueden variar.
5. Vea A20 en intel.com/processors/claims. Los procesadores escalables de la 5ª generación Intel Xeon Scalable proporcionan hasta 1.4x (BFI6) y 1.3x (INT8) vs. 4th Gen y hasta 1.4x (BFI6) y 6.7x (INT8) vs. 3rd Gen Intel® Xeon® processors. Resultados pueden variar.
6. Universidad de Massachusetts, "Energy and Policy Considerations for Deep Learning in NLP", junio de 2019
7. Be Davis, "The growing energy footprint of artificial intelligence", octubre de 2023
8. Basado en el desempeño por watt que va de 1.46x a 10.6x con aceleradores incorporados en una gama de cargas de trabajo de IA, banco de datos y redes, vea el Sitio A9-A15, D2, D5, D25, N6 en intel.com/processors/claims: Los procesadores escalables de la 5ª generación Intel Xeon Scalable presentan resultados que pueden variar.
9. TruScale mide de manera precisa solo el desempeño y la capacidad, permitiendo que infraestructuras gestionadas sean diseñadas, implementadas y afinadas no solo para desempeño y capacidad, sino también para emisiones de CO_2. El monitoreo continuo del sistema usando Lenovo xClarity Power Monitor y sistemas de desempeño permite optimizar el consumo de energía por la infraestructura. Las emisiones de CO_2 son calculadas basadas en la huella de carbono localizada de la fuente de energía utilizada.

intel. XEON PLATINUM

Lenovo ThinkSystem SR650 V3 servers construidos sobre procesadores escalables de la 5ª generación Intel® Xeon® diseñados para IA.