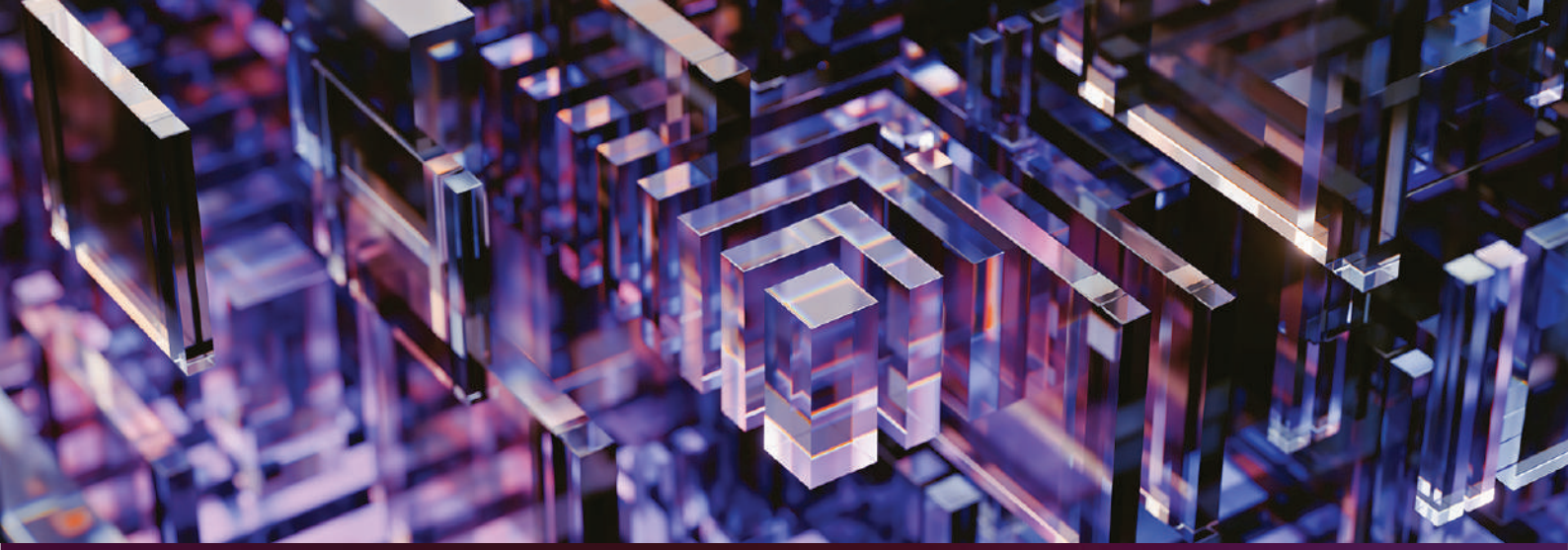


Proporcionando Infraestructura de Alto Rendimiento para Inteligencia Artificial Generativa y Empresarial

Soluciones de Lenovo y NVIDIA® para Mejorar la Productividad, Innovación y Tiempo de Lanzamiento al Mercado

Resumen Ejecutivo

En muchos sectores, la Inteligencia Artificial (IA) y la Inteligencia Artificial Generativa (GenAI) pueden acelerar la innovación y mejorar la posición competitiva de una empresa, la calidad de productos/servicios, operaciones y el compromiso con el cliente. Sin embargo, existen numerosos desafíos de implementación al desplegar casos de uso en la vida real.

Lenovo ayuda a las empresas a superar estos obstáculos proporcionando un conjunto integral de servicios con las mejores prácticas y una infraestructura optimizada para la IA, que incluye servidores, almacenamiento, estaciones de trabajo, dispositivos móviles y software, desde el edge hasta el centro de datos y la nube. Por ejemplo, los servidores Lenovo ThinkSystem y ThinkEdge, impulsados por unidades de procesamiento gráfico (GPUs) de NVIDIA y software, pueden acelerar la jornada de IA y GenAI del cliente ofreciendo:

- Resultados más rápidos para workloads de entrenamiento e inferencia en texto, video, imagen y otras modalidades de datos.
- Más flexibilidad para personalizar y optimizar diversas workloads de IA, GenAI y otras relacionadas, desde el edge hasta el centro de datos y la nube.
- Mejor eficiencia energética y menor costo total de propiedad (TCO).
- Diversas ofertas y servicios inmersivos complementarios para facilitar la transformación digital de la empresa con IA y GenAI, incluido el acceso al Centro de Excelencia Lenovo AI Discover (AIDiscover@lenovo.com).

Introducción

Las soluciones de IA y GenAI están creciendo rápidamente en el ámbito empresarial, brindando muchos beneficios en diversas industrias. Las empresas están implementando IA y GenAI para acelerar la innovación y mejorar su posición competitiva, la calidad de productos/servicios, operaciones y el compromiso con el cliente.

Aunque la promesa y el valor económico de la IA son inmensos, también lo son los desafíos de implementación, dadas las grandes cantidades de datos y la necesidad de almacenar, analizar y proteger de manera efectiva todos sus datos valiosos a lo largo de su ciclo de vida. Con GenAI, estos problemas se vuelven aún más agudos.

Este documento técnico discute cómo Lenovo y NVIDIA se asocian con sus respectivas tecnologías únicas para proporcionar la arquitectura óptima que ofrece IA y GenAI para las empresas. Basándose en la participación de Lenovo y NVIDIA con clientes y socios, este documento ofrece orientación valiosa para seleccionar configuraciones optimizadas para el rendimiento en varios casos de uso de IA y GenAI en diversas industrias para obtener una ventaja competitiva. Lenovo continúa invirtiendo[1] en asociaciones de IA, incluida NVIDIA, para acelerar la implementación de la IA para empresas en todo el mundo y ayudar a los clientes a iniciar su viaje de IA y GenAI hoy mismo.

IA y GenAI Generan Valor Empresarial en Diversas Industrias

La IA, que incluye Aprendizaje Automático (ML) y Aprendizaje Profundo (DL), está creciendo rápidamente y transformando una amplia gama de industrias y aplicaciones. Se espera que el mercado global de IA alcance los US $241.80 mil millones en 2023 y se proyecta que crecerá a una tasa de crecimiento anual compuesta (CAGR) del 17.30% de 2023 a 2030, lo que resultará en un volumen de mercado de US $738.80 mil millones para 2030[2].

GenAI, que incluye Modelos de Lenguaje Grandes (LLMs), es un nuevo tipo de DL poderoso que puede crear nuevo contenido, como texto, imágenes, audio y video. Lo hace aprendiendo patrones a partir de datos existentes y luego utilizando este conocimiento para generar salidas nuevas y únicas. GenAI puede producir contenido altamente realista y complejo que imita la creatividad humana. Está creciendo incluso más rápido que la IA[3] (más del 58%) y se está convirtiendo en una herramienta valiosa para muchas industrias, como servicios financieros, salud, manufactura, retail, telecomunicaciones/medios etc.

La Figura 1 representa varios casos de uso destacados de GenAI resumidos de un estudio reciente de Deloitte[4] que agregan un valor significativo a través de diversas industrias extrayendo ideas profundas y prácticas en muchas modalidades de datos (Texto, Audio, Imagen, Video, Código y Artefactos 3D/Especializados). Los colores sólidos y ricos en el gráfico de torta representan las modalidades de datos típicamente prominentes para cada caso de uso. El color blanco en el gráfico de torta es para modalidades de datos que generalmente no son significativas.

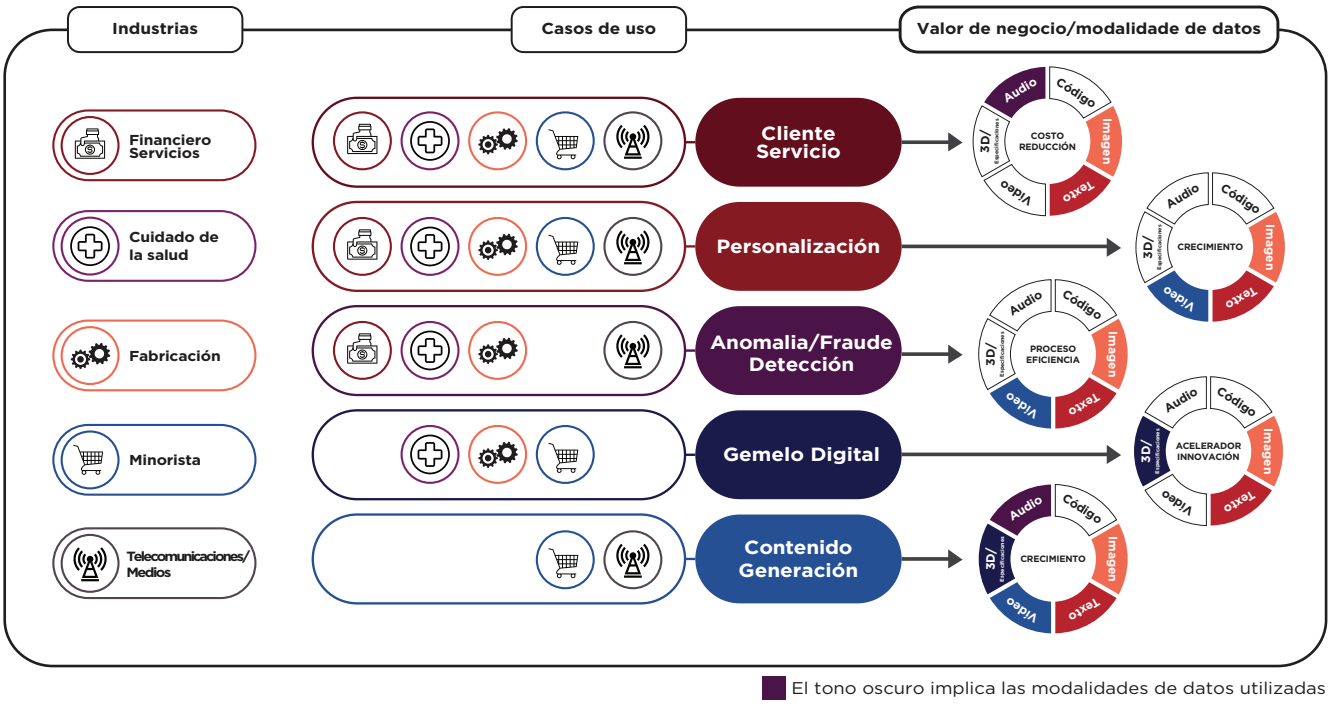


Figura 1: Casos de Uso de Alto Valor de GenAI en Diversas Industrias y Modalidades de Datos[5]

Servicios Financieros: Bancos y compañías de seguros están incorporando GenAI a sus procesos intensivos en datos para mejorar:

- **Atención al Cliente:** Interfaz de avatar digital impulsada por GenAI con opciones de texto, audio e imágenes que mejoran el soporte al cliente las 24 horas, los 7 días de la semana, responden a consultas y ayudan con tareas financieras para mejorar la eficiencia del proceso y el compromiso del cliente.
- **Personalización:** Entregar materiales de marketing en conformidad con las regulaciones, promociones de productos y participación en ventas con texto, imágenes y videos personalizados en diferentes geografías para impulsar el crecimiento y adquirir nuevos clientes.
- **Detección de Fraudes:** Identificar transacciones fraudulentas en tiempo real analizando patrones y anomalías en datos reales y sintéticos en varias modalidades, ayudando a mejorar procesos y prevenir pérdidas financieras.

Cuidado de la salud: Pagadores, proveedores, organizaciones farmacéuticas y de biotecnología están incorporando GenAI para:

- Servicio al Cliente: Acelerar la autorización previa para pacientes y generar respuestas a preguntas sobre el proceso de reclamación, cobertura de seguros y otros detalles del plan para mejorar el proceso. Habilitar el monitoreo continuo y proacti-vo las 24 horas, los 7 días de la semana, de pacien-tes con dispositivos IoT y análisis impulsados por IA para vigilar los signos vitales del paciente, alertar a los proveedores de salud sobre desviaciones de valores típicos y tomar medidas correctivas.
- **Personalización:** Descubrir y adaptar tratamientos y medicamentos a pacientes individuales según su composición genética y antecedentes médicos para mejorar la eficacia del cuidado y fomentar el crecimiento del negocio y la ventaja competitiva.
- **Detección de Fraude/Anomalía:** Identificar reclamaciones fraudulentas en tiempo real, analizando patrones y anomalías en datos de varias modalidades, ayudando a mejorar procesos y prevenir pérdidas financieras. Analizar imágenes médicas como radiografías, resonancias magnéticas y tomografías computarizadas para asistir en la detección temprana de enfermedades y anomalías.
- **Gemelo Digital:** Construir réplicas digitales centradas en el paciente de extremo a extremo para analizar datos del paciente, incluidos registros médicos/imagenes y síntomas, para mejorar el diagnóstico de enfermedades y recomendar planes de tratamiento. Identificar posibles candidatos a medicamentos, predecir su eficacia y optimizar estructuras moleculares.

Prever brotes de enfermedades, readmisiones de pacientes y la utilización de recursos de atención médica, ayudando a hospitales y clínicas a asignar recursos de manera eficiente. Todo esto impulsa más innovación en todo el ecosistema de salud.

Manufactura: Fabricantes automotrices, aeroespaciales y de semiconductores están incorporando GenAI para:

- **Mantenimiento:** Analizar datos de sensores de máquinas en varias modalidades para predecir cuándo podrían fallar. Esto ayuda a realizar servicios preventivos y evitar tiempos de inactividad costosos y interrupciones.
- **Personalización:** Permitir la personalización masiva mediante el análisis de datos y la adaptación eficiente de procesos de fabricación para producir productos personalizados. Impulsa un mayor atractivo para el cliente y el crecimiento del negocio.
- **Detección de Anomalías:** Sistemas de visión computarizada impulsados por IA pueden inspeccionar de manera rápida y precisa productos en busca de defectos, reduciendo el número de artículos defectuosos en la línea de producción para impulsar la eficiencia del proceso.
- **Gemelo Digital:** Construir una réplica digital de extremo a extremo de todo el ciclo de vida del producto, desde el desarrollo hasta la fabricación y el servicio. Ayuda a generar y evaluar nuevos diseños de productos, optimizándolos para rendimiento, costo y fabricabilidad. Optimizar los procesos de fabricación mediante el análisis de datos de sensores y líneas de producción para mejorar la eficiencia y la calidad del producto. Proporciona información en tiempo real sobre la cadena de suministro y la operación del cliente, ayudando a los fabricantes a rastrear materias primas, monitorear el progreso de la producción, responder a interrupciones, rastrear el uso real del cliente de sus productos y garantizar un mantenimiento oportuno. Todo esto mejora drásticamente la innovación de productos y procesos y la calidad.

Minorista: Empresas orientadas al consumidor, como grandes minoristas, pequeños comerciantes y minoristas especializados, están utilizando GenAI para:

- **Servicio al Cliente:** Opciones de interfaz de avatar digital alimentadas por GenAI con texto, audio e imágenes mejoran el soporte al cliente las 24 horas, los 7 días de la semana, responden a consultas y ayudan con recomendaciones de productos para mejorar la eficiencia del proceso y el compromiso empático con el cliente, construyendo lealtad de marca y equidad. También libera recursos humanos costosos para abordar problemas más complejos del cliente.
- **Personalización:** Entregar materiales de marketing, promociones de productos y participación en ventas con texto, imágenes y videos personalizados en diferentes geografías para impulsar el crecimiento y adquirir nuevos clientes. Generar recomendaciones más específicas y dirigidas en muchas modalidades que los motores de búsqueda para hacer que la compra sea más personalizada y conveniente.
- **Gemelo Digital:** Crear salas de exhibición virtuales, demostraciones de productos y planogramas, personalizar la experiencia del cliente, pronosticar la demanda, simular el diseño y las operaciones de una tienda e identificar áreas de mejora. Todo esto puede ayudar a los minoristas a innovar más.
- **Generación de Contenido:** Crear descripciones de productos, imágenes, videos y más de manera más rápida y consistente que las herramientas y procesos tradicionales. Para hacer crecer el negocio, personalizar este contenido por geografía, idioma, matices culturales y regulaciones locales.

Telco/Media: Con GenAI, estas empresas están acelerando la transformación digital con mejoras en:

- **Servicio al Cliente:** Opciones de interfaz de avatar digital y asistencia virtual alimentadas por GenAI con texto, audio e imágenes mejoran el soporte al cliente las 24 horas, los 7 días de la semana, responden a consultas y ayudan con recomendaciones de servicios para mejorar la eficiencia del proceso y el compromiso empático del cliente, construyendo lealtad y conteniendo costos de cambio. También libera recursos humanos costosos para abordar problemas más complejos del cliente. Optimizar el rendimiento de la red y reducir la congestión, mejorando la experiencia del cliente.
- **Personalización:** Organizar y gestionar tipos de archivos complejos, analizar contenido antes de la traducción para optimizar la localización e integrar otras herramientas de idiomas en el flujo de trabajo para aumentar las conversiones y el compromiso para construir lealtad. El reconocimiento de voz ayuda a transcribir contenido de video y audio a texto y traducir contenido hablado a otros idiomas para hacer crecer el negocio.
- **Detección de Fraude:** Utilizar datos reales y sintéticos para mejorar la eficiencia del proceso y detectar actividades fraudulentas en redes de telecomunicaciones, como el cambio de SIM y el acceso no autorizado.
- **Generación de Contenido:** Imitar el estilo de los materiales de marketing de la empresa y generar nuevas y diversas versiones de contenido de manera rápida y bajo demanda adaptadas a diferentes audiencias. Mejorar la calidad del lenguaje de materiales de marketing con formulación, gramática, estilo de la empresa y adherencia a los valores de la empresa. Crear rápidamente numerosas versiones de contenido en varios estilos para identificar la mejor opción para hacer crecer el negocio.

Si bien el retorno de inversión (ROI) de GenAI puede ser sustancial para la empresa, implementar el aprendizaje profundo (DL) y la infraestructura de tecnología de la información (TI) de alto rendimiento asociada puede ser complejo y costoso. Existen numerosos desafíos de implementación.

Para discutir su caso de uso específico, le invitamos a ponerse en contacto con el Laboratorio de Descubrimiento de IA de Lenovo enviando un correo electrónico a AIDiscover@lenovo.com.

Desafíos de Implementación de IA y GenAI

La implementación de flujos de trabajo de DL y GenAI en producción generalmente pasa por cuatro etapas (Figura 2)[6]:

1. **Gestión de Datos** para preparar los datos necesarios para construir el modelo DL y **GenAI.**
2. **Aprendizaje del Modelo (Entrenamiento)** para definir, seleccionar y entrenar el modelo DL y **GenAI.**
3. **Verificación del Modelo (Entrenamiento)** para asegurar que el modelo cumpla con requisitos específicos de funcionalidad y rendimiento.
4. **Implementación del Modelo (Inferencia)** para integrar el modelo entrenado en la infraestructura de TI y ejecutar, mantener y actualizar el modelo según sea necesario.

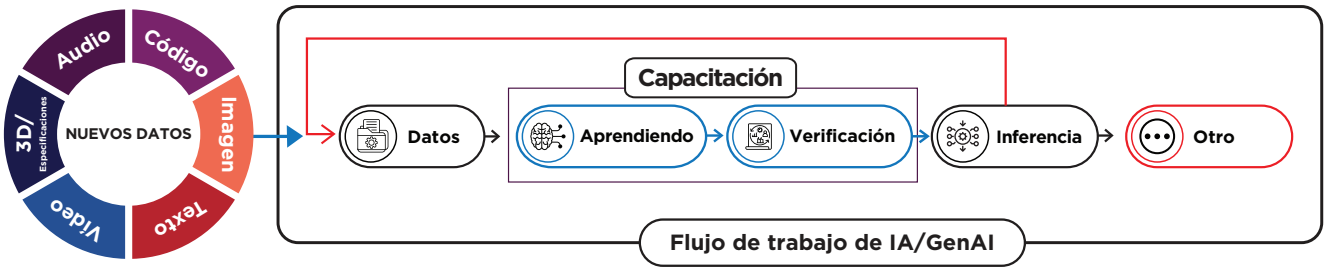


Figura 2: Fases Clave en el Flujo de Trabajo de DL y GenAI

Estas etapas tienen pasos más pequeños (Figura 2) que pueden ejecutarse en paralelo y con retroalimentación. Además, existen otras consideraciones éticas, legales, de confianza y seguridad. Todo esto hace que la implementación de GenAI sea muy desafiante. La Figura 3 representa estos desafíos de Datos, Procesos, Negocios, Infraestructura y Otros:

Figura 3: Desafíos de Implementación de GenAI

Datos

Calidad y Cantidad de Datos

- Disponibilidad de Datos: Los modelos GenAI a menudo requieren grandes cantidades de datos de alta calidad, lo cual puede ser desafiante de recopilar y depurar.
- Diversidad de Datos: Asegurar que los datos de entrenamiento representen una amplia gama de escenarios y demografías puede ser complejo.

Sesgo y Equidad

- Sesgo en los Datos: Los modelos GenAI pueden heredar sesgos en los datos de entrenamiento, lo que resulta en salidas sesgadas o injustas.
- Equidad: Asegurar la equidad en el contenido generado, especialmente en dominios sensibles como finanzas y salud, es un desafío significativo.

Interpretabilidad y Explicabilidad

- Modelos de Caja Negra: Muchos modelos GenAI son como "cajas negras", lo que hace difícil entender sus procesos de toma de decisiones. Esto puede ser problemático para aplicaciones donde la transparencia es crucial.

Procesos

Gestión del Cambio

- Cultura Organizacional: Implementar GenAI puede requerir cambios significativos en la cultura, procesos y flujos de trabajo de una organización, lo que puede implicar superar la resistencia organizativa.

Colaboración Humano-AI

- Entrenamiento y Monitoreo: Las empresas deben establecer procesos de colaboración entre operadores humanos y sistemas GenAI, incluido el monitoreo continuo y el mantenimiento.

Aceptación y Confianza del Usuario

- Escepticismo del Usuario: Los usuarios pueden ser escépticos respecto al contenido generado por IA, lo que afecta las tasas de adopción.
- Construcción de Confianza: Construir confianza en el contenido generado por IA es crucial para la aceptación del usuario.

Negocios

Evaluación del ROI

- Medición del Impacto: Evaluar el retorno de inversión (ROI) de la implementación de GenAI puede ser desafiante, especialmente al cuantificar el valor generado por las soluciones de IA.

Habilidades y Talento

- Escasez de Talento: Puede haber escasez de expertos en IA y científicos de datos con las habilidades necesarias para implementar y mantener sistemas GenAI de manera efectiva.

Infraestructura

Recursos Computacionales

- Infraestructura de Alto Rendimiento: Entrenar e implementar modelos GenAI a gran escala requiere recursos computacionales sustanciales, lo que conlleva a costos de infraestructura elevados.
- Escalabilidad: Asegurar que la infraestructura pueda expandirse para manejar aumentos en las demandas computacionales a medida que evolucionan los modelos GenAI es un desafío continuo.
- Eficiencia Energética: La infraestructura debe ser eficiente en términos de energía. El análisis ha demostrado que entrenar un LLM, un modelo GenAI con 200 mil millones de parámetros, produce aproximadamente 75,000 kg de emisiones de CO_2, en comparación con solo 900 kg de emisiones de CO_2 por un vuelo de Nueva York a San Francisco[7].

Integración con Sistemas Existentes

- Sistemas Heredados: Integrar GenAI con la infraestructura de TI existente y los sistemas heredados puede ser complejo y requerir esfuerzo sustancial.

Entrenamiento y Ajuste del Modelo

- Tiempo de Entrenamiento: Entrenar modelos GenAI complejos puede llevar tiempo, retrasando la implementación de soluciones de IA.
- Ajuste de Hiperparámetros: Afinar modelos para tareas específicas y optimizar su rendimiento puede requerir un esfuerzo significativo.

Otros

Preocupaciones Éticas

- Uso Malintencionado: Existen preocupaciones sobre el mal uso del GenAI para generar contenido falso, deep fakes u otros fines maliciosos.
- Privacidad: Generar contenido altamente personalizado puede plantear preocupaciones sobre la privacidad, requiriendo medidas sólidas de protección de datos.
- Alucinaciones: Son salidas del modelo que son o sin sentido o completamente falsas.

Cumplimiento Regulatorio

- Privacidad de Datos: Cumplir con regulaciones de privacidad de datos, como el GDPR o el HIPAA, puede ser complejo al manejar datos generados por el usuario.
- Regulaciones de Contenido: Algunas industrias, como la farmacéutica, bancaria y financiera, tienen regulaciones estrictas que rigen el contenido que producen y comparten.

Seguridad

- Vulnerabilidades: Los modelos GenAI pueden ser vulnerables a ataques adversos, comprometiendo potencialmente su confiabilidad y seguridad.
- Propiedad Intelectual: Los modelos GenAI y los procesos utilizados para construirlos suelen ser los "tesoros" de una organización y deben ser protegidos.

Lenovo espera que la IA se desarrolle y utilice consistentemente con sus valores fundamentales. El Comité de IA Responsable de Lenovo garantiza que todas las soluciones, incluidas las de los socios Innovadores de IA, cumplan con los requisitos que protegen a los usuarios finales y aseguran que el uso de la IA sea justo, ético y responsable, enfocándose en:

- Diversidad e Inclusión
- Privacidad y Seguridad
- Responsabilidad y Confianza
- Explicabilidad
- Transparencia
- Impacto Ambiental y Social

Lenovo y NVIDIA están trabajando con un ecosistema amplio y en crecimiento de socios y clientes para desarrollar mejores prácticas y soluciones que ayuden a las empresas a

superar estos desafíos de implementación. Lenovo también ha creado una arquitectura de referencia[8] para GenAI basada en GPU y software de NVIDIA.

Arquitectura de Alto Nivel de las Soluciones de Lenovo Impulsadas por NVIDIA

Lenovo simplifica la implementación de IA y GenAI con una infraestructura optimizada, lista para implementar (hardware, software y servicios), experiencia comprobada y soluciones prevalidadas de ISVs y socios diseñadas para cualquier tamaño o escala. En la base de esta arquitectura de alto nivel (Figura 4) se encuentran sistemas líderes expertamente diseñados de alto rendimiento y almacenamiento para IA y GenAI, desde estaciones de trabajo hasta el edge, pasando por el centro de datos hasta la nube.

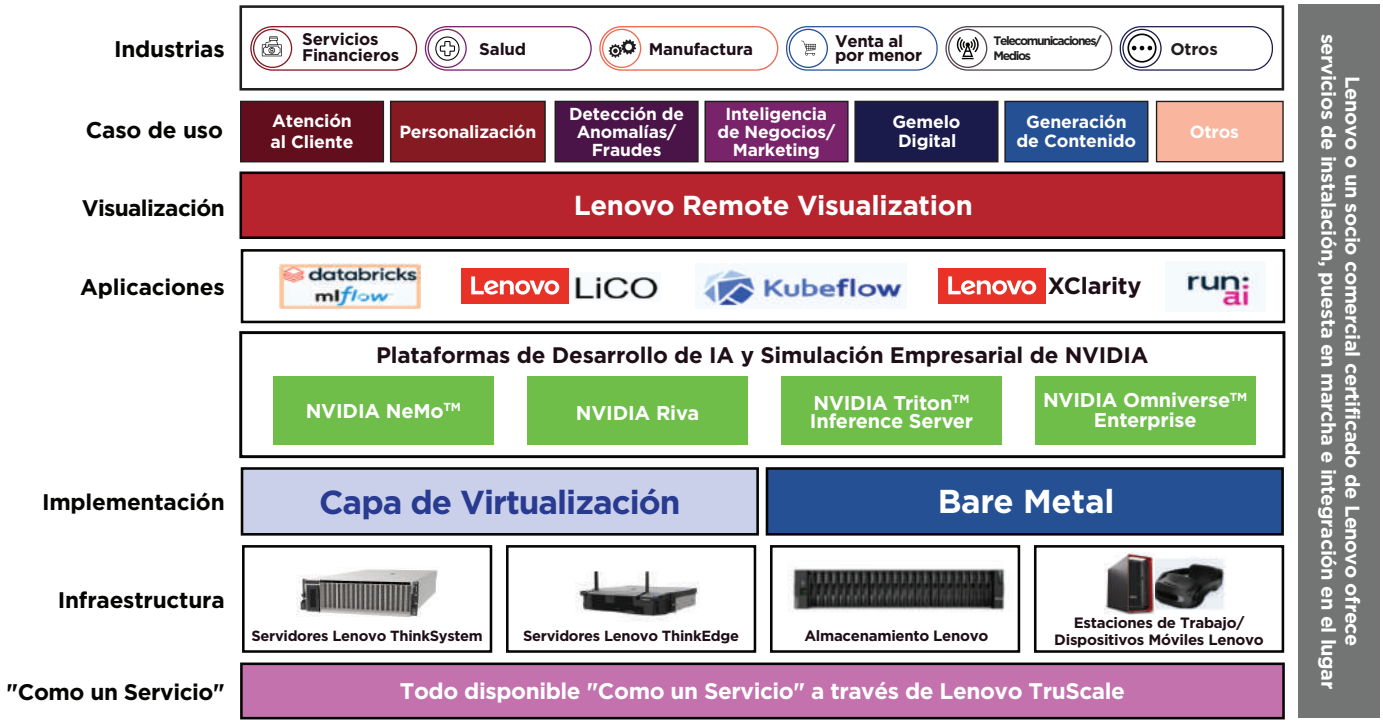


Figura 4: Arquitectura de Alto Nivel de IA y GenAI

Algunos componentes (no abordados anteriormente) de esta arquitectura de alto nivel, comenzando en la capa de infraestructura, incluyen:

- **Servidores Lenovo ThinkSystem Optimizados para Rendimiento:** Servidores altamente confiables, escalables y de alto rendimiento para acelerar significativamente la IA y GenAI. Este portafolio de servidores de Lenovo incluye el Lenovo ThinkSystem SR675 V3 rico en GPU. Aprovechando las tecnologías de enfriamiento líquido de Lenovo, algunos sistemas van desde la refrigeración directa por agua para CPUs y GPUs hasta sistemas mejorados con líquido donde el líquido aumenta el enfriamiento estándar por aire.
- **Servidores Lenovo ThinkEdge:** Ofrecen plataformas específicas y seguras adecuadas para aplicaciones intensivas en cómputo y sensibles a la latencia, como el Lenovo ThinkEdge SE455 V3 desplegado fuera de los centros de datos tradicionales.
- **Almacenamiento Lenovo:** Los JBODs de almacenamiento directamente conectados y las unidades de expansión proporcionan almacenamiento flexible, rentable y de alta capacidad, ideal para entornos con restricciones de espacio y clientes sensibles a los costos. Las matrices Lenovo ThinkSystem DE Series All-Flash están diseñadas para un rendimiento extremo con hasta 2.0 millones de IOPS y latencia inferior a 100 microsegundos, e incluyen funciones de disponibilidad y seguridad probadas en empresas líderes en la industria.
- **Estaciones de Trabajo Lenovo:** Las estaciones de trabajo de la serie ThinkStation P con GPU de NVIDIA ofrecen un rendimiento potente y están certificadas por ISV, son eficientes en energía y altamente versátiles.
- **Opciones de Implementación:** Proporciona la autonomía para adaptar el enfoque de implementación, eligiendo entre una configuración robusta de metal desnudo o una implementación virtual versátil.
- **AI Enterprise de NVIDIA:** Como una pila completa de software de IA, AI Enterprise de NVIDIA (con componentes clave como NVIDIA NeMoTM, NVIDIA Riva y NVIDIA TritonTM) acelera los flujos de trabajo de IA y simplifica el desarrollo e implementación de IA en producción para una amplia gama de casos de uso, desde visión por computadora hasta GenAI, incluidos los LLM.
- Omniverse Enterprise de NVIDIA: Es una plataforma de software nativa OpenUSD para conectar tuberías 3D complejas y desarrollar aplicaciones para la digitalización industrial. Unifica fácilmente sus herramientas y datos 3D para romper los silos de información, minimizar la tediosa preparación de datos y potenciar la colaboración entre equipos empresariales. Aproveche las herramientas de desarrollo fáciles de usar para construir aplicaciones 3D avanzadas en tiempo real que le permitan visualizar y simular productos, activos e instalaciones en plena fidelidad

de diseño. Implemente la plataforma en su entorno preferido, ya sea en estaciones de trabajo móviles profesionales NVIDIA RTXTM, estaciones de trabajo y servidores certificados por NVIDIA, o NVIDIA OVXTM.

- **Aplicación:** Componentes principales en esta capa incluyen:
  - **Databricks MLflowTM:** Proporciona una plataforma unificada para gestionar el ciclo de vida del aprendizaje automático, desde el seguimiento de experimentos y el registro de modelos hasta la implementación y el monitoreo del modelo.
  - **Lenovo XClarity:** Es una familia de software que simplifica y automatiza la implementación y gestión de la infraestructura de Lenovo, permitiendo que los clientes se centren en sus proyectos de alto valor.
  - **Lenovo Intelligent Computing Orchestration (LiCO):** Reduce la complejidad de utilizar un clúster HPC masivo y simplifica el despliegue, operación y aceleración de aplicaciones.
  - **Run:ai:** Es un programador que gestiona tareas en lotes utilizando múltiples colas en la parte superior de Kubernetes®, permitiendo que los administradores del sistema definan diferentes reglas, políticas y requisitos para cada cola según las prioridades comerciales.
- **Lenovo Remote Visualization:** Proporciona acceso confiable y seguro a aplicaciones intensivas en gráficos en cualquier momento y lugar. En lugar de emitir nuevas estaciones de trabajo costosas para todo el personal de diseño, la TI puede implementar computadoras personales empresariales o de consumo menos costosas. Además, los departamentos de TI pueden mantener la seguridad y reducir los costos utilizando la visualización remota alojada en un centro de datos interno o desde la nube. La visualización remota realiza operaciones gráficas intensivas en un servidor gráfico de alta gama y genera una versión de píxeles en 2D que los usuarios pueden recibir rápidamente. Además, la renderización del lado del servidor acelera considerablemente el proceso de utilizar gráficos en sesiones remotas.
- **Lenovo o Servicios Certificados por Socios:** Lenovo y su ecosistema global de socios altamente especializados en software y servicios de IA pueden ofrecer toda o partes de la pila integrada de Lenovo representada en la Figura 4. También pueden proporcionar servicios de instalación y puesta en marcha en el lugar para integrar esto en el entorno de trabajo del cliente, incluida la instalación de aplicaciones de IA y GenAI en diversas industrias.
- **"Como un Servicio":** Suscríbase a la innovación que crece con usted con Lenovo TruScale, que proporciona servicios de entrega, gestión y actualización de extremo a extremo, lo que significa que sus equipos de TI no tienen que levantar un dedo al implementar nuevos dispositivos y escalar su infraestructura de TI.

En el núcleo de esta arquitectura de alto nivel se encuentran los servidores de Lenovo con software y GPUs de NVIDIA que brindan un rendimiento excelente para IA y GenAI.

Software y GPUs de alto valor de NVIDIA para IA y GenAI

El software de alto valor de NVIDIA representado en esta arquitectura de alto nivel incluye:

- **NVIDIA AI Enterprise** es una plataforma de software de IA nativa de la nube, segura y de alto rendimiento con seguridad, estabilidad, capacidad de gestión y soporte de nivel empresarial para la creación e implementación de modelos de IA. Acelera los flujos de trabajo de IA y simplifica el desarrollo e implementación de IA en producción, cubriendo una variedad de casos de uso desde visión por computadora hasta IA y GenAI. NVIDIA AI Enterprise incluye:
  - **NVIDIA NeMo™ (Neural Models)** es un marco integral nativo de la nube para construir, personalizar e implementar modelos de IA y GenAI. Viene con un conjunto completo de herramientas y recursos, que incluyen:
    - Una biblioteca de modelos preentrenados para diversas tareas, como generación de texto, traducción, reconocimiento de voz y generación de imágenes.
    - Un conjunto de herramientas para personalizar y entrenar modelos.
    - Una plataforma basada en la nube para implementar y gestionar modelos a escala.
    - NeMo Guardrails ayuda a las empresas a mantener aplicaciones construidas en LLM alineadas con sus requisitos de seguridad.
  - **NVIDIA Riva** es un SDK de IA acelerado por GPU para construir e implementar pipelines de IA conversacional totalmente personalizables y en tiempo real para:
    - Reconocimiento automático de voz (ASR).
    - Avatares digitales de IA conversacional.
    - Sistemas interactivos de respuesta por voz (IVR).
    - Traducción neuronal de máquina (NMT).
    - Texto a voz (TTS).
    - Asistentes de voz.
  - **NVIDIA Triton™ Inference Server** Inference Server es un software de código abierto que estandariza la implementación y ejecución de modelos de inteligencia artificial en cada workloads. Triton acelera y optimiza la implementación y ejecución de modelos de inteligencia artificial en la nube, centro de datos y dispositivos periféricos.
- **NVIDIA Omniverse™** es una plataforma potente en las GPU de NVIDIA que facilita la integración de tecnologías de realidad aumentada (AR) y realidad virtual (VR) en empresas. Proporciona un entorno colaborativo donde los equipos pueden crear, simular y visualizar mundos virtuales y mejorar varios aspectos de sus flujos de trabajo.

NVIDIA ofrece varias GPU de alto rendimiento para ayudar a los clientes a implementar workloads de inteligencia artificial, GenAI y otras.

Aquí hay algunas GPU asequibles basadas en la arquitectura NVIDIA Ada Lovelace que son adecuadas para estas workloads:

- **La NVIDIA L40S (2U)** ofrece aceleración de extremo a extremo para la próxima generación de aplicaciones habilitadas para inteligencia artificial, desde el entrenamiento e inferencia de modelos GenAI hasta gráficos 3D y aceleración de medios. Las potentes capacidades de inferencia del L40S, combinadas con el trazado de rayos acelerado por NVIDIA RTX y motores de codificación y decodificación dedicados, aceleran el audio, el habla, la inteligencia artificial en 2D, video y GenAI en 3D.
- **La NVIDIA L40 (2U)** ofrece gráficos neurales revolucionarios, virtualización, cómputo y capacidades de inteligencia artificial para workload aceleradas por GPU en centros de datos.
- **La NVIDIA L4 (1U)** es un acelerador universal, rentable y eficiente en energía diseñado para satisfacer las necesidades de inteligencia artificial en video, cómputo visual, gráficos, virtualización y numerosas aplicaciones, incluidos juegos en la nube, simulación y ciencia de datos. Ofrece alto rendimiento y baja latencia en cada servidor, desde el edge hasta el centro de datos y la nube.

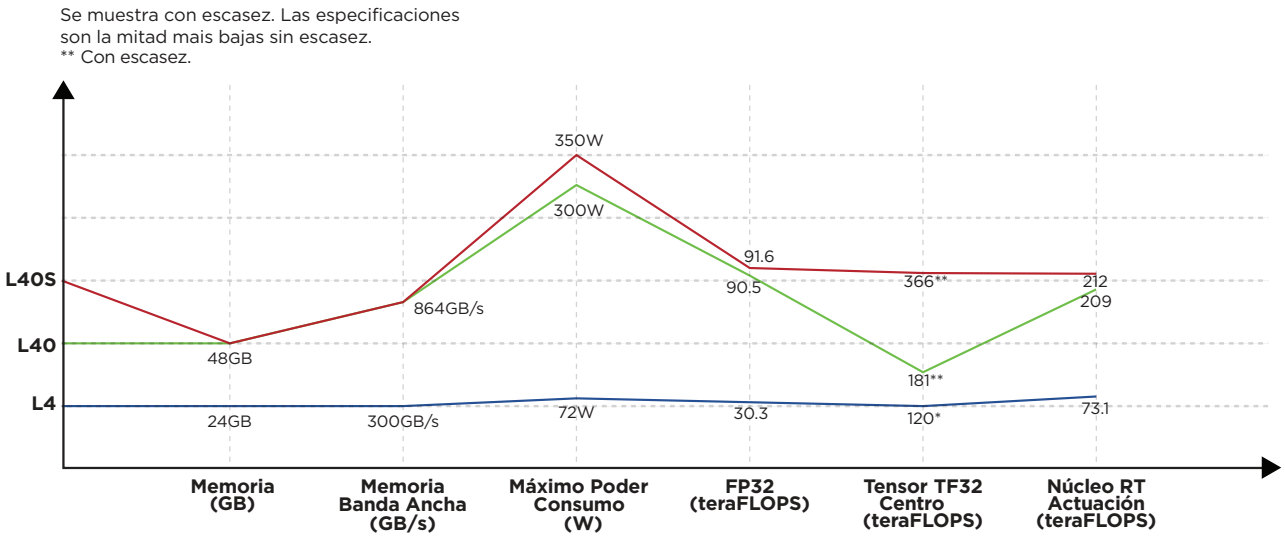


Figura 5: Características Comparativas de la GPU

La Figura 5 representa las características clave de estas tres GPU. La Tabla 1 muestra un mapa sugerido de afinidad de workload (Lo Mejor, Mejor y Bueno) por GPU, aunque depen-dería de los requisitos específicos del cliente.

Las celdas en blanco en la Tabla 1 significan que la GPU correspondiente es excesiva o insuficiente para esa workload en particular. Por ejemplo, para escritorio virtual (VDI) y aceleración en el edge, la L4 es la única GPU sugerida. La L40 y la L40S son excesivas, más caras y ocupan más ranuras ya que ambas son de 2U. Para entrenamiento de DL y cómputo de alto rendimiento (HPC), la L40S es la única GPU sugerida debido a su rendimiento significativamente mejor en el núcleo Tensor TF32. La L40 es la mejor para renderizado con su excelente rendimiento del núcleo RT y mayor asequibilidad que la L40S.

Portafolio de GPU de NVIDIA y Afinidad de Workloads

GPU	Entrenamiento de DL	Inferencia DL	HPC/AI	Renderizado	Estación de Trabajo Virtual.	Escritorio Virtual (VDI)	Video de IA	Aceleración en el Edge Lejano
L40S	◎	◎	◎	◎	◎		◎	
L40				●	●			
L4		○		○	●	●	●	●

● Lo Mejor ◎ Mejor ○ Bueno

Tabla 1: Portafolio de GPU de NVIDIA para Afinidad de Workloads

Las Tablas 2 y 3 muestran las GPU sugeridas para entrenamiento de IA y GenAI (solo L40S) e inferencia de workloads, incluidos LLM.

Portafolio de Entrenamiento de GPU de NVIDIA						
GPU	NLP/LLM				Imagen/ Video Gen AI	Recsys
	Up to 5B	6B to 65B	66B to 175B	>175B		
L40S	◎	◎	○		◎	○

● Lo Mejor ◎ Mejor ○ Bueno

Tabla 2: Portafolio de Entrenamiento de GPU de NVIDIA

Portafolio de Inferencia de GPU de NVIDIA								
GPU	NLP/LLM				Imagen/ Video Gen AI	Recsys	Visión por Computadora	Video de AI
	Up to 5B	6B to 65B	66B to 175B	>175B				
L40S	◎	◎	○		●		◎	◎
L4	○				○		●	●

● Lo Mejor ◎ Mejor ○ Bueno

Tabla 3: Portafolio de Inferencia de GPU de NVIDIA

Con base en estas GPU de NVIDIA y software, Lenovo proporciona a los clientes en muchas industrias soluciones validadas y optimizadas en rendimiento con la elección y flexibilidad de personalizar según casos de uso específicos, workloads, presupuestos y otros requisitos.

Lenovo Ofrece la Arquitectura Óptima con NVIDIA para IA y GenAI

Las Figuras 6 y 7 muestran un mapa de alto nivel de los-servidores Lenovo ThinkSystem con GPU específicas de NVIDIA. Estos sistemas están diseñados y desarrollados desde cero para cumplir y superar los rigurosos requisitos de rendimiento de aplicaciones y flujos de trabajo de inteligencia artificial y GenAI de la industria.

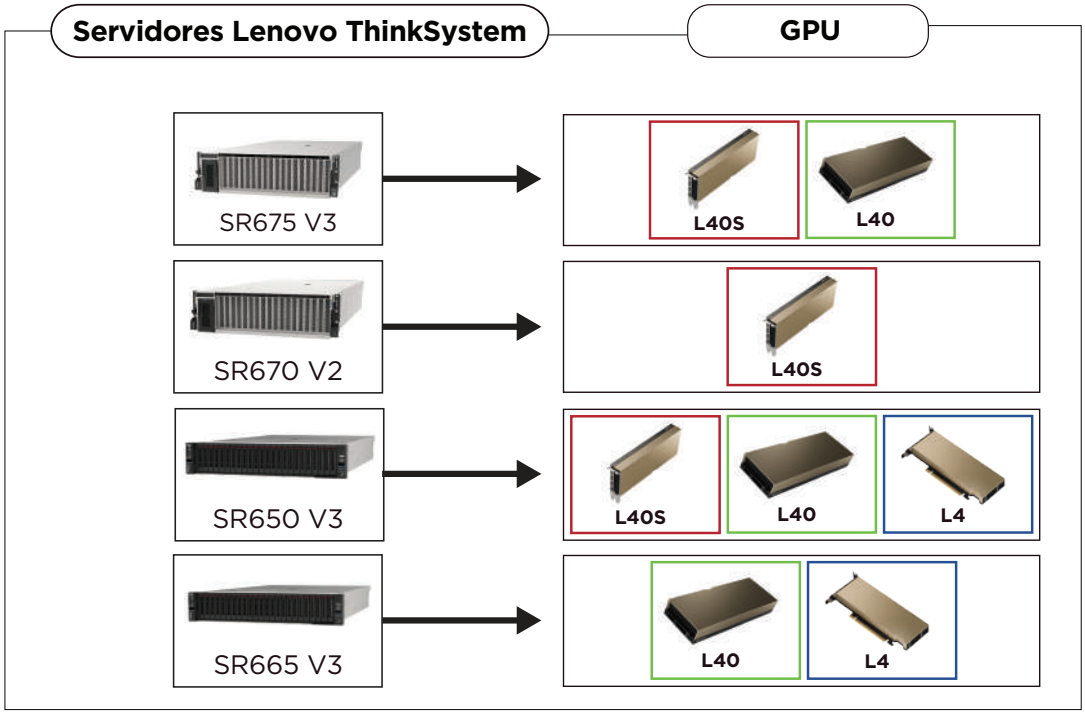


Figura 6: Servidores Lenovo ThinkSystem con GPU Relevantes Compatibles para IA y GenAI

- Lenovo ThinkSystem SR675 V3 Server y Lenovo ThinkSystem SR670 V2 Server son servidores versátiles de bastidor de 3U ricos en GPU que admiten ocho GPU de doble ancho, incluidas las GPU Tensor Core L40S, con NVLink y refrigeración híbrida líquida-aire de Lenovo Neptune. Estos servidores ofrecen un rendimiento óptimo en muchas industrias para workloads de IA, GenAI, HPC y gráficos.
- Lenovo ThinkSystem SR665 V3 Server ofrece el rendimiento definitivo de servidor de dos sockets en un factor de forma de 2U. Es ideal para workloads densas que utilizan el procesamiento de GPU y unidades NVMe de alto rendimiento.
- Lenovo ThinkSystem SR650 V3 Server es un servidor versátil de bastidor de 2U con 2 sockets ideales para confiabilidad, gestión y seguridad líderes en la industria, maximizando el rendimiento y la flexibilidad para el crecimiento futuro. Puede manejar varias workloads empresariales, como bases de datos, virtualización, computación en la nube, transmisión de medios etc.

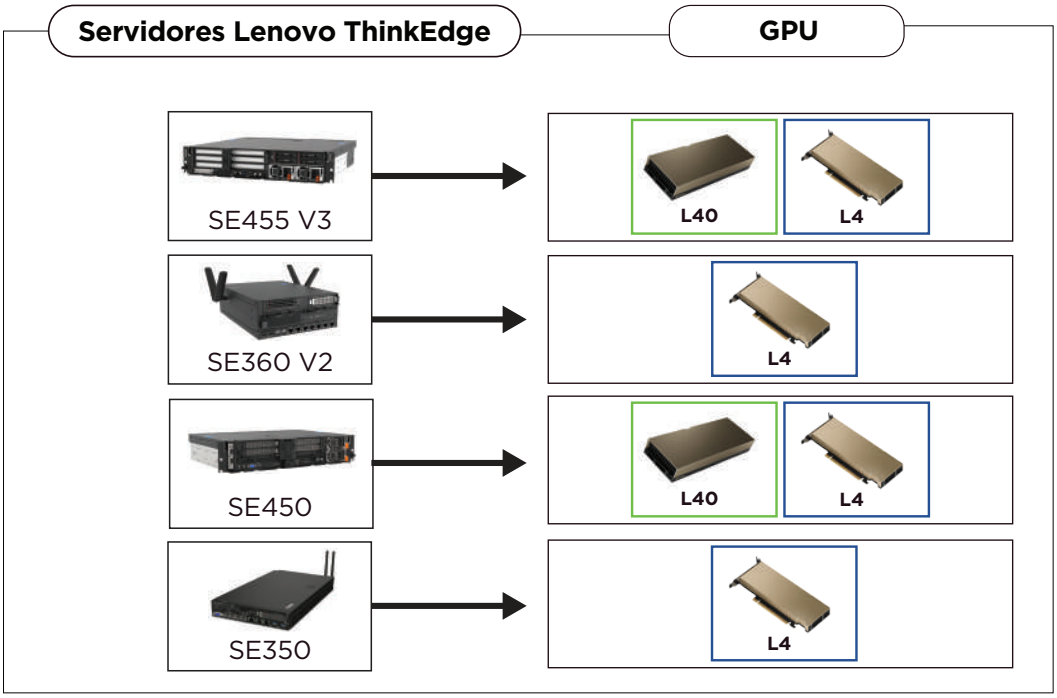


Figura 7: Servidores Lenovo ThinkEdge con GPU Relevantes Compatibles para IA y GenAI

- ThinkEdge SE455 V3 Edge Server es para soluciones AI y específicas de telecomunicaciones y admite estrategias emergentes de consolidación de workloads en el edge, con un gran recuento de núcleos en un espacio más pequeño.
- Lenovo ThinkEdge SE450 Edge Server es un servidor de un solo zócalo con una altura de 2U y un estuche de profundidad corta que puede ir casi a cualquier lugar, operar silenciosamente en un amplio rango de temperaturas y tolerar polvo y vibraciones.
- Lenovo ThinkEdge SE360 V2 Edge Server y Lenovo ThinkEdge SE350 V2 Edge Server tienen la mitad del ancho y son significativamente más cortos que un servidor tradicional, ideales para su implementación en espacios reducidos. Ofrecen mayor potencia de procesamiento, almacenamiento y red más cerca de la fuente de generación de datos para workloads en tiempo real como AR/VR, vigilancia, IA, etc.

Lenovo también ofrece valor adicional y varios servicios y soluciones complementarios para ayudar a los clientes en su viaje de IA y GenAI con:

- La Arquitectura de Referencia de Lenovo para GenAI basada en LLM: Lenovo creó recientemente esta arquitectura de referencia para ayudar a los clientes en su viaje de IA y GenAI.
- Lenovo Innovador en Eficiencia Energética de Enfriamiento: A medida que las frecuencias de los procesadores y el número de núcleos aumentan, y las GPUs se vuelven más potentes para proporcionar el mejor rendimiento, es crucial enfriar estos sistemas de manera eficiente para evitar problemas de sobrecalentamiento que causan apagones, rendimiento más lento y posibles pérdidas de datos. Durante más de una década, Lenovo ha liderado en tecnología de energía y enfriamiento de centros de datos y cuenta con varias soluciones innovadoras y únicas con disipadores de calor especializados o líquido para aire (L2A) y ventiladores de alta velocidad con baja impedancia. Si el enfriamiento por aire no es viable, los clientes pueden utilizar otras tecnologías de enfriamiento líquido en el portafolio Lenovo Neptune.

- Aceleración en el Descubrimiento y Adopción de IA: Muchas empresas enfrentan desafíos de implementación debido a limitaciones de recursos y complejidades de infraestructura, lo que interrumpe el lanzamiento de iniciativas de IA y GenAI. El programa Lenovo AI Innovators incluye un ecosistema de socios de software líderes que colaboran con Lenovo para proporcionar a los clientes soluciones de IA y GenAI personalizadas, probadas y listas para implementar en sus casos de uso.
- Laboratorio de Descubrimiento de IA: Trabaje con expertos en IA de Lenovo y NVIDIA para obtener el máximo valor, reduciendo los riesgos del proyecto. Lenovo ha estado a la vanguardia de la IA durante casi una década. Benefíciese del Laboratorio de Descubrimiento de IA de Lenovo, talleres de evaluación de IA y un comité de IA que impulsa la adopción de IA para clientes en todos los continentes.

El Laboratorio de Descubrimiento de IA puede ofrecer los siguientes servicios:

  - Acceso a científicos de datos, arquitectos de soluciones e ingenieros de rendimiento de GPU.
  - Ayuda en la identificación y entrega de soluciones de IA que cumplan o superen los KPI establecidos por su empresa. Identificará y entregará soluciones de IA que generen retorno de inversión, no solo proyectos de IA.
  - Enfoque en casos de uso en manufactura, venta al por menor, salud y finanzas, pero también se han llevado a cabo muchos proyectos en varias otras industrias.
  - Ayuda en la determinación de la estrategia de IA y adaptación a nuevas tecnologías GenAI.
  - Entrega de casos de uso de implementación de visión computacional, como se ha hecho en muchos proyectos, desde NASCAR hasta Island Conservation, hasta la detección de defectos de fabricación.
  - Entrega de soluciones GenAI para implementaciones locales que preservan la privacidad y la seguridad, como se ha hecho con muchos LLMs de código abierto.

- **Práctica de Servicios Profesionales de IA de Lenovo:** Ofreciendo una variedad de servicios, soluciones y plataformas, la Práctica de Servicios Profesionales de IA de Lenovo ayuda a empresas de todos los tamaños a navegar por el panorama de la IA, encontrar las soluciones correctas y poner la IA en funcionamiento en sus organizaciones de manera rápida, efectiva y a escala. Ayuda a llevar la IA del concepto a la realidad, desde la elaboración de mapas de IA hasta la implementación de plataformas y proporcionando transparencia en el uso de tecnología con el Lenovo TruScale Hub.
- **Soluciones Complementares Inovadoras:** Lenovo está entregando muchas tecnologías de vanguardia en estaciones de trabajo, laptops, tablets, dispositivos móviles, AR/VR (ThinkReality) y computación en la nube (TruScale), que satisfacen las necesidades de integración, flexibilidad y experiencia inmersiva de clientes en varias industrias.
- **Del Edge al Data Center hasta una Plataforma de Nube:** El modelo de computación de IA y GenAI es híbrido, con el entrenamiento realizado en el centro de datos con servidores ThinkSystem e inferencia realizada en el edge con servidores ThinkEdge (Figura 8).

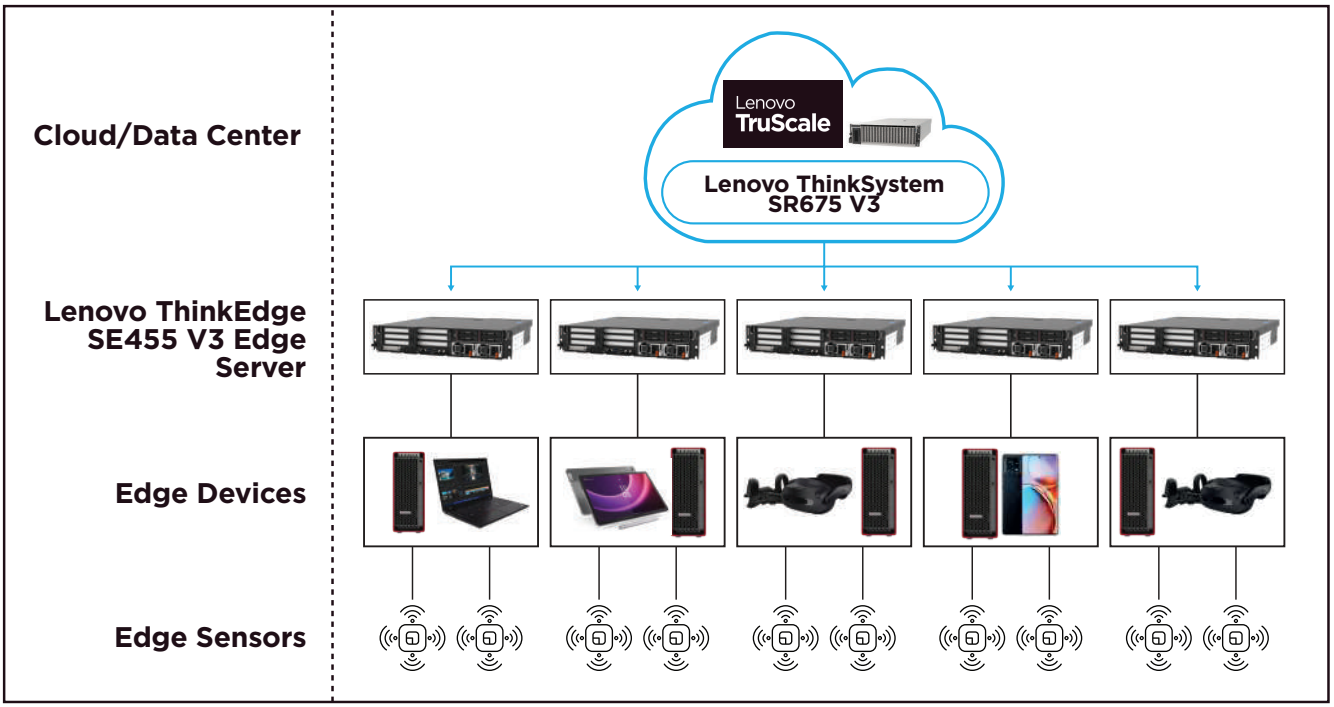


Figura 8: Soluciones Lenovo del Edge al Data Center para IA y GenAI

Casos de Uso del Edge al Data Center Específicos por Industria

Aquí hay algunos ejemplos específicos por industria:

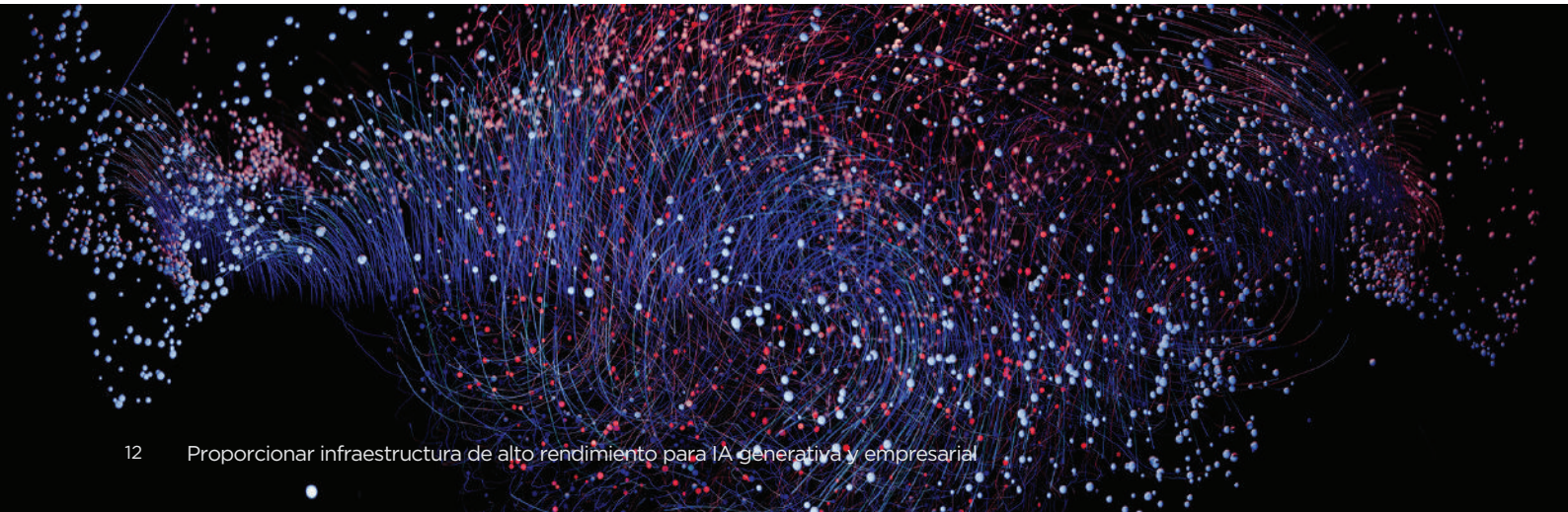
- **Empresas de Servicios Financieros** pueden emplear los servidores Lenovo ThinkEdge para la detección de fraudes en tiempo real. Estos servidores ahora cuentan con la capacidad de llevar a cabo la autenticación biométrica (inferencia) utilizando modelos entrenados en el centro de datos, los cuales se actualizan periódicamente con datos reales y sintéticos para mejorar la precisión.
- **Los Proveedores de Salud** pueden utilizar los servidores Lenovo ThinkEdge para monitorear los signos vitales de los pacientes y otros datos de salud en tiempo real desde sus dispositivos vestibles. Al analizar estos datos (inferencia) mediante los servidores ThinkEdge en sus consultorios, los proveedores pueden identificar posibles problemas de salud de manera temprana y ofrecer recomendaciones de salud personalizadas a los pacientes. Estos proveedores pueden compartir estos datos con organizaciones de salud afiliadas, las cuales pueden utilizar dichos datos de pacientes anonimizados para construir modelos de entrenamiento de IA más efectivos y realizar predicciones más precisas en el futuro.
- **Los Minoristas** pueden emplear GenAI para crear recomendaciones personalizadas para los clientes en sus tiendas. El minorista entrena un modelo GenAI con datos de ventas para aprender qué productos es probable que los clientes compren juntos. Un servidor ThinkEdge en una tienda específica puede ejecutar este modelo para generar (inferencia) recomendaciones personalizadas cuando un cliente entra en la tienda.
- **Los Fabricantes** pueden inspeccionar productos mediante análisis de imagen (inferencia) en los servidores Lenovo ThinkEdge para detectar defectos en la línea de ensamblaje, reduciendo el desperdicio y mejorando la calidad, el diseño y la fabricabilidad del producto. Estas percepciones influyen en los nuevos diseños de productos analizados en servidores Lenovo ThinkSystem de alto rendimiento en el centro de datos.
- **Empresas de Telecomunicaciones y Medios** pueden personalizar la experiencia televisiva para sus clientes. Al entrenar un modelo de IA con datos del cliente en un servidor Lenovo ThinkSystem, una empresa puede aprender qué tipos de programas y películas es probable que gusten a los clientes. Luego, este modelo puede ser implementado en dispositivos de edge, como decodificadores, para generar recomendaciones personalizadas.

Inicie su Jornada con Lenovo y NVIDIA en su Viaje de IA y GenAI

A medida que las empresas incorporan la IA y GenAI como parte de sus procesos comerciales centrales, no pueden permitirse problemas de rendimiento, demoras o tiempos de inactividad. Por lo tanto, el soporte debe ser proactivo, realizado por especialistas técnicos que trabajen estrechamente con el cliente y comprendan profundamente su entorno.

Como parte de su contrato con Lenovo, las empresas pueden recibir un gerente de cuenta técnico dedicado o un administrador de sistemas como su único punto de contacto. Ya sea en el lugar, trabajando de forma remota o una combinación de ambos, los profesionales de soporte pueden identificar y resolver rápidamente cualquier problema, asegurando que el entorno de IA funcione de manera óptima las 24 horas del día, los 7 días de la semana.

Sin embargo, Lenovo va mucho más allá del soporte técnico especializado. El servicio integral de Lenovo para IA y GenAI incluye consultas iniciales, talleres, análisis y configuración del entorno correcto, pasando por la evaluación continua de enfriamiento y servicios de monitoreo/mantenimiento hasta la facturación y administración. Estos servicios integrales pueden ayudar a los clientes a maximizar el retorno de la inversión en sus iniciativas de IA.



La Ventaja Lenovo y NVIDIA

A medida que la IA, especialmente la GenAI, se convierte en parte integral de los procesos comerciales centrales de una empresa, debe superar varios desafíos de implementación. Lenovo y NVIDIA ayudan a las empresas a maximizar el retorno de la inversión en sus iniciativas de IA, acelerar el tiempo de valor y impulsar la innovación y productividad al ofrecer:

- **Sistemas Optimizados para Rendimiento:** Los servidores ThinkSystem y ThinkEdge alimentados por GPUs y software NVIDIA ofrecen un excelente rendimiento para workloads exigentes de entrenamiento e inferencia en modalidades de datos de texto, video e imagen para casos de uso en diversas industrias.
- **Servicios y Software de Alto Valor:** El programa Lenovo AI Innovators incluye un ecosistema líder en software y socios de servicios para proporcionar a los clientes soluciones personalizadas, probadas y listas para implementar de IA y GenAI desde consultas iniciales, talleres, análisis hasta la configuración del entorno correcto.
- **Liderazgo en Eficiencia Energética:** Lenovo lidera en tecnología de energía y enfriamiento de centros de datos y ofrece varias soluciones innovadoras y únicas de enfriamiento por aire y líquido, incluyendo las tecnologías de enfriamiento líquido Neptune™.
- **Soporte a Nivel Empresarial:** Los sistemas se prueban, validan y optimizan para rendimiento, capacidad de gestión, seguridad y escalabilidad. Lenovo, o un socio de negocios certificado, proporciona instalación en el lugar, puesta en marcha, integración y monitoreo y resolución proactiva de cualquier problema operativo.
- **Un Portafolio Completo de Soluciones:** Con Lenovo, los clientes pueden implementar soluciones de IA de extremo a extremo utilizando un amplio portafolio de dispositivos móviles inteligentes, estaciones de trabajo hasta servidores ThinkEdge y los servidores ThinkSystem más escalables. Estos sistemas vienen con una amplia gama de almacenamiento, software y servicios integrales que brindan un excelente rendimiento, confiabilidad y seguridad para el entorno de TI del cliente, desde el edge hasta el centro de datos y la nube.
- **Un Sólido Plan con Innovación Continua:** NVIDIA continúa liderando el mercado de GPUs al proporcionar consistentemente un portafolio de GPUs y software de alto rendimiento para acelerar las workloads de entrenamiento e inferencia más exigentes de la GenAI, reduciendo el TCO. De manera similar, Lenovo entrega servidores de centro de datos y edge que integran rápidamente estas GPUs NVIDIA con otras tecnologías de vanguardia en la computación en la nube (TruScale) y AR/VR (ThinkReality), que abordan las futuras necesidades de rendimiento, accesibilidad, eficiencia energética y experiencia inmersiva para empresas y sus clientes.

Maximice el Retorno de la Inversión en su Iniciativa de IA

Por favor, póngase en contacto con su representante de Lenovo o envíe un correo electrónico a AIDiscover@lenovo.com para programar una consulta inicial con un Experto en IA de Lenovo o solicitar un taller personalizado de IA.

[1] Lenovo incrementa los ingresos de su infraestructura de IA a más de US$2 mil millones y lleva la IA a los datos con el portafolio más completo de la industria - Lenovo StoryHub

[2] "Inteligencia Artificial - En todo el mundo", Statista

[3] Se espera que el mercado de software de IA generativa supere los US$36 mil millones en ingresos acumulados para 2028, con una tasa de crecimiento anual compuesta del 58% entre 2023 y 2028 | S&P Global Market Intelligence (spglobal.com)

[4 e 5] Instituto de IA de Deloitte, "Dossier de IA Generativa: Una selección de casos de uso de alto impacto en seis industrias principales", 2023.

[6] Desafíos en la Implementación del Aprendizaje Automático: Un Estudio de Casos, 2011.09926v2.pdf (arxiv.org), Jan 2021

[7] Cómo funcionan los Transformers - Curso de NLP de Hugging Face

[8] https://lenovopress.lenovo.com/lp1798-reference-architecture-for-generative-ai-based-on-large-language-models#authors.