El punto óptimo de la informática empresarial moderna

Investigación realizada por:

Peter Rutten
Director de Investigación, Sistemas de infraestructura, Grupo de plataformas y tecnologías, Líder de investigación global de soluciones informáticas de alto rendimiento, IDC

Septiembre de 2021 | N.º de documento de IDC US48197921

Navegar por este documento técnico

Pulse los títulos o números de página para navegar a cada sección.

Opinión de IDC

El entorno de TI actual puede parecer un enigma. En el proceso de convertirse en una empresa digital y satisfacer las necesidades de clientes hiperexigentes, las empresas intentan alcanzar casi lo imposible.

- Los mercados pueden cambiar caprichosamente, provocando picos o caídas, y esta volatilidad no puede interpretarse como una excepción. La volatilidad es la línea base actual.
- Para cumplir las fluctuaciones de carga de trabajo de la demanda, los sistemas deben escalarse de forma impecable y dinámica sin necesidad de un crear centro de datos masivo, costoso y con un gran consumo de energía solo para los picos. La sostenibilidad ya no es solo un ardid de marketing.
- La complejidad de estos mercados tampoco puede analizarse y optimizarse con la capacidad y experiencia humana común. Gran parte de la información debe ser ahora artificial, operar en tiempo real y equilibrar innumerables variables mientras se obtienen ingentes cantidades de datos. La inteligencia artificial (IA) se incorporará cada vez más en todos los componentes y requiere prestaciones de hardware especialmente diseñadas.
- Dada la demanda de disponibilidad permanente y perpetua, las cargas de trabajo que dan soporte a la empresa digital no pueden ralentizarse ni obstaculizarse, y mucho menos reducirse completamente. En el mundo "siempre activo" actual, cualquier período de inactividad puede ser catastrófico.
- Con todos los componentes digitales y conectados para habilitar la empresa digital, los sistemas se han visto también expuestos a nuevos tipos de ataques que los ponen en peligro. Comunidades enteras de personas malintencionadas se han fusionado en un inframundo que libra una guerra permanente contra las empresas de todo el mundo, utilizando un amplio arsenal de herramientas y estrategias de ciberataque. Como resultado, ahora todo debe empezar por una seguridad integral e infalible.

Con este objetivo, para que cualquier plataforma de proceso de clase empresarial pueda funcionar como motor de la empresa digital, debe ser inequívocamente segura, fiable, escalable y sostenible, estar integrada en la nube como parte de un enfoque híbrido y haberse diseñado para incorporar IA. Este documento técnico profundizará en estas consideraciones desde el punto de vista de la infraestructura y el despliegue, y analizará cómo el nuevo procesador IBM Power10 y la nueva plataforma de clase empresarial IBM Power, la E1080, se ejecuta en ellas.

Información general de la situación

IDC considera que para que una empresa digital tenga éxito en el difícil entorno multidimensional actual, deben tenerse en cuenta las siguientes consideraciones críticas:

- **La seguridad como requisito imperativo**
- **El mandato de fiabilidad**
- **Escalabilidad y sostenibilidad**
- **La infraestructura de TI híbrida adecuada (nube híbrida y aplicaciones nativas en la nube)**
- **La importancia de la IA y dónde ejecutarla**

Las siguientes secciones profundizarán en cada una de estas consideraciones.

La seguridad como requisito imperativo

La seguridad se ha convertido en el requisito más importante de una empresa digital. Cuando IDC encuesta a las organizaciones sobre sus prioridades, la seguridad es invariablemente la primera o de las primeras de la lista. De hecho, cuando se les solicita, por ejemplo, que seleccionen los principales elementos de la infraestructura de IA que las empresas consideran que no son óptimos en las ofertas de servidores y almacenamiento de sus proveedores o distribuidores, la seguridad obtiene la puntuación más alta, ya que el 30 % de las empresas afirman no estar satisfechas con las funciones de seguridad.[1]

Este descontento también se manifiesta en el hecho de que muchas empresas no permiten que los dispositivos de almacenamiento que contienen los datos para sus cargas de trabajo de IA sean utilizados por otras cargas de trabajo. El motivo más aducido para ello (45 %) es la seguridad y la privacidad de los datos. Asimismo, la investigación de IDC ha descubierto que la seguridad es una gran preocupación en la infraestructura de nube pública como servicio, donde un 37 % de las empresas afirman que la seguridad es su mayor preocupación en esos despliegues.[2] Las empresas también están incorporando cada vez más la IA en sus cargas de trabajo de seguridad, más que en cualquier otra carga de trabajo, para que puedan predecir mejor las infracciones y actuar en consecuencia.

En la actualidad, la mayor parte de la atención y la inversión se destinan a la seguridad de las pilas de aplicación y de red. Sin embargo, un gran número de ataques se producen a bajo nivel y centrados en el hardware. A menudo se inician aprovechando las vulnerabilidades de los procesadores y/o el microcódigo. Estos ataques son sofisticados y difíciles de detectar.

Por ello, IDC está viendo cómo las empresas se interesan cada vez más en la "informática confidencial" para sus plataformas de negocio críticas. La informática confidencial permite el aislamiento de datos confidenciales en un subsistema de procesador designado y protegido (a veces denominado "enclave de procesador seguro") para su procesamiento.

[1] Fuente: IDC AI Infrastructure View 2021

[2] Fuente: IDC IaaSView 2020

En la actualidad, los datos suelen estar cifrados en reposo en el almacenamiento y en tránsito a través de la red, pero no mientras se utilizan en la memoria. La capacidad de proteger los datos y el código mientras están en la memoria es limitada en muchas plataformas informáticas. No obstante, las organizaciones que manejan datos confidenciales como, por ejemplo, información de identificación personal, datos financieros o información médica necesitan mitigar las amenazas dirigidas a la aplicación o los datos en la memoria del sistema.

En la informática confidencial, el contenido del subsistema, que puede estar cifrado a nivel de hardware, solo es accesible para el código autorizado dentro de un programa. Los contenidos no son accesibles para elementos externos, por ejemplo, otro código, otros sistemas u otros operadores. Las entidades no autorizadas no pueden ver ni manipular indebidamente los datos ni el proceso de ejecución de código autorizado. Una solución integral de informática confidencial protegerá los datos en uso y los datos en reposo; para ello, se utiliza el cifrado de contenido en almacenes de datos persistentes y memoria volátil o no volátil del sistema, ya sea en medios flash o de rotación.

Las modernas infraestructuras de informática confidencial (especialmente las implementadas en entornos compartidos de arrendatario múltiple) utilizan coprocesadores discretos para descargar operaciones de procesador privilegiadas que puedan estar comprometidas por vulnerabilidades de ejecución de código de bajo nivel. Todavía no es un método común, pero para las cargas de trabajo básicas de la empresa es muy prometedor. Mientras tanto, las empresas están aplicando varias estrategias de seguridad simultáneamente, para el hardware y el software.

El Mandato de fiabilidad

Aunque las estrategias de seguridad son de vital importancia para proteger los datos, las aplicaciones y el hardware de los ataques, otro aspecto clave de la empresa digital es una fiabilidad absoluta del entorno de TI, incluida la infraestructura. La alta disponibilidad no es un concepto nuevo y las empresas pueden elegir plataformas de proceso con hasta un 99,999 % de disponibilidad y plataformas de almacenamiento con un 99,99999 %. No obstante, estas cifras solo se consiguen con el hardware, el software y las políticas adecuadas. IDC ha designado solamente a nueve plataformas de servidor de seis proveedores del mercado de servidores como de Nivel de Disponibilidad 4 (AL4)[3], que es el nivel más alto y representa una tolerancia a errores completa.

- La investigación de IDC[4] muestra que las tres causas principales de los períodos de inactividad de las aplicaciones son un error en la red (16,2 %), un error en los servidores (15,5 %) y un programa malicioso (10,3 %). Entre las causas más comunes de error de servidores están la sobrecarga en la memoria (DRAM) o las CPU y un error o un daño de memoria.
- Los volúmenes de transacciones están aumentando de forma espectacular y las empresas necesitan velocidades de transacción cada vez más rápidas para satisfacer las necesidades de sus clientes.
- Las cargas de trabajo de misión crítica y claves para el negocio están creciendo, y las funciones de soporte empresarial que antes podían ejecutarse en un nivel de baja disponibilidad — por ejemplo, mediante virtualización o una agrupación en clústeres — se consideran cada vez más importantes para el negocio.
- El coste de los períodos de inactividad aumenta a medida que las empresas dependen cada vez más de su infraestructura para las operaciones diarias. La investigación de IDC muestra que para el 20,7 % de las organizaciones, el coste del tiempo de inactividad es de 5000–10 000 dólares por hora; para el 18,4 %, es de 10 000–25 000 dólares por hora; para el 17 %, es de 25 000–100 000 dólares por hora; y para algunas empresas (1,4 %), es de 500 000 dólares por hora.
- El fin de las "horas laborable normales" y la existencia de aplicaciones de empresa que ahora deben estar disponibles para los clientes en todo momento han puesto una tremenda presión en la infraestructura que da soporte a dichas aplicaciones, por lo que el tiempo de inactividad planificado o no planificado que se permite es nulo o casi nulo.
- La tolerancia a interrupciones, retrasos, pérdida de datos y daño de datos es nula, tanto por parte de las empresas como de los consumidores, y cualquier infracción o error puede tener consecuencias catastróficas para la reputación de una organización.

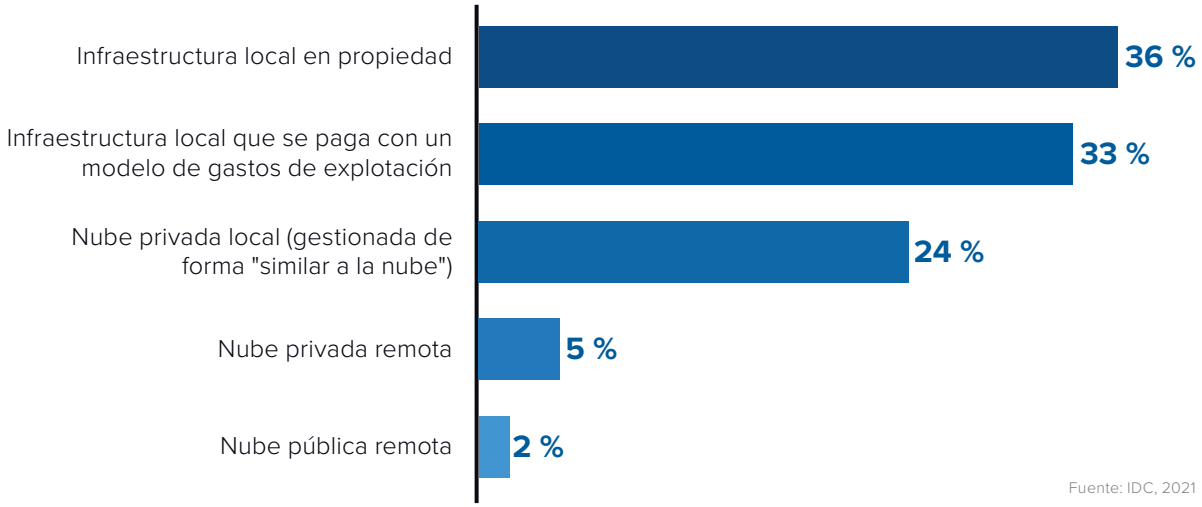
[3] Fuente: IDC Worldwide AL4 Server Market Shares, 2019: *Fault-Tolerant Systems Become Digital Transformation Platforms*

[4] Fuente: IDC *Server Storage Infrastructure Availability Survey,* 2018

- A medida que las empresas interactúan digitalmente con consumidores o ciudadanos y con otras empresas con mayor frecuencia y de formas muy diversas, la conformidad con la normativa nacional e internacional sobre disponibilidad de datos, seguridad y privacidad es de suma importancia.
- Aunque la disponibilidad y la seguridad en la nube pública han mejorado mucho, la verdadera tolerancia a errores sigue viéndose como una funcionalidad local o de nube híbrida, no como una funcionalidad de nube pública (véase la **Figura 1**).

FIGURA 1

Infraestructura que aloja el nivel de disponibilidad máximo

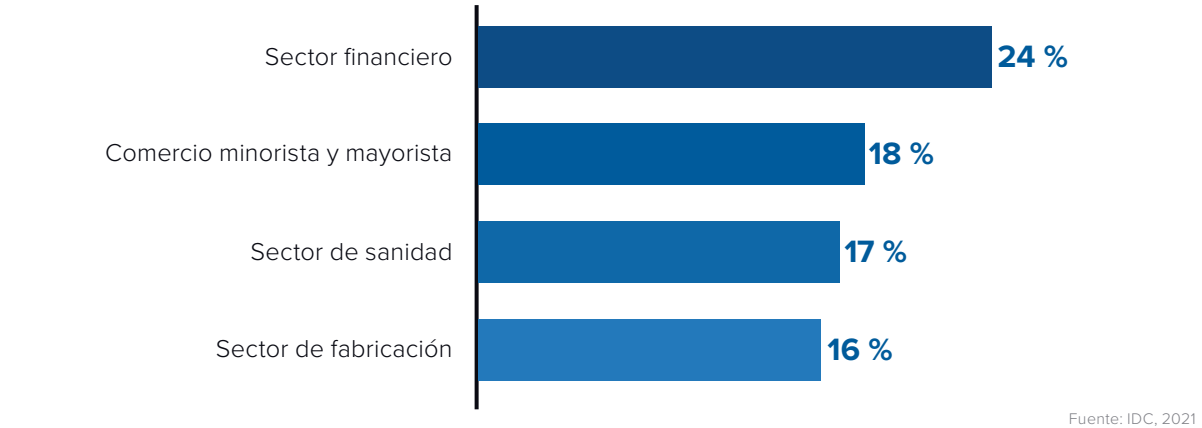


Fuente: IDC, 2021

Como resultado, el porcentaje de todos los sistemas que necesitan ser altamente disponibles está creciendo. En todos los sectores, más del 60 % de las empresas tienen entre el 21 y el 30 % de sus servidores en el nivel de disponibilidad máximo. La **Figura 2** muestra el porcentaje de sistemas que deben ser altamente disponibles en distintos sectores.

FIGURA 2

Porcentaje de sistemas que necesitan una alta disponibilidad, por sector



Fuente: IDC, 2021

Las actuales plataformas AL4 dominantes han dado pasos de gigante para convertirse en plataformas totalmente integradas en el centro de datos, que no solo participan en la transformación digital de una organización, sino que de hecho la dirigen. Estos sistemas procesan los datos más críticos y valiosos de muchas empresas, a menudo en mayor volumen que cualquier otro tipo de datos, y las empresas necesitan desbloquear estos datos y optimizarlos para convertirse en empresas digitales.

La necesidad de escalabilidad y sostenibilidad

Las empresas necesitan escalar cargas de trabajo que procesen cantidades de datos cada vez mayores en entornos de TI en constante expansión.

Al mismo tiempo, deben poder escalarse rápidamente en función de olas de demanda a veces imprevisibles, que ocasionalmente pueden mostrar fuertes picos. Todo esto significa centros de datos más grandes, más equipos, más renovación de equipos y más energía para ejecutar los equipos y al mismo tiempo refrigerarlos.

Las cargas de trabajo de IA son las cargas de trabajo con más crecimiento que consumen datos y determinan las inversiones de informática que realizan las empresas. En la actualidad, el 21 % de las organizaciones afirman estar invirtiendo en tecnologías de informática que habilitan el proceso paralelo necesario para el entrenamiento y la inferencia en redes de aprendizaje profundo de IA, y un 9 % adicional de empresas afirman tener previsto hacerlo en 2021. Además, el 46 % de las empresas está invirtiendo en tecnologías de aceleración de carga de trabajo, por ejemplo, GPU, FPGA y ASIC, y un 7 % adicional tiene previsto invertir en 2021.[5] Estas últimas sobre todo han provocado problemas en el centro de datos en cuanto a las necesidades energéticas y la refrigeración. El caso de uso más común para la aceleración es el de la inferencia del aprendizaje profundo de IA (llevando a producción un modelo de IA desarrollado con una red neuronal profunda [DNN]). Actualmente, el 38 % de las organizaciones utiliza la aceleración para la inferencia de IA, mientras que solo el 27 % utiliza la aceleración para el entrenamiento de una DNN.[6] La tendencia de que las inversiones de inferencia de IA empezaran a superar a las de entrenamiento de IA era una tendencia esperada. Asimismo, la IA no es la única carga de trabajo que está impulsando las inversiones en aceleración utilizando este tipo de coprocesadores. La analítica de datos, HPC, el modelado financiero, la ciberseguridad, la detección de fraudes y las operaciones financieras son otros ejemplos de cargas de trabajo que se ejecutan cada vez más en GPU, FPGA o ASIC, y la mayoría de las empresas ejecutan estas cargas de trabajo en el entorno local.

Sin embargo, un problema importante es que la mayoría de los centros de datos no están equipados para soportar varios bastidores de nodos de cálculo acelerados para entregar la potencia que requieren y disipar el calor que generan, que es mucho mayor que en los bastidores de servidores no acelerados. Según el Departamento de Energía de EE. UU. (2020), los centros de datos son uno de los tipos de edificación con un mayor uso de energía, ya que consumen entre 10 y 50 veces más energía por espacio de suelo que una edificación de oficina comercial típica. IDC ha descubierto que, como promedio, el 17,6 % del presupuesto operativo del centro de datos se dedica a electricidad, más que a cualquier otra partida de presupuesto. En Estados Unidos, los centros de datos representan el 2 % de la electricidad total utilizada en el sector comercial.

Sin embargo, al mismo tiempo, muchas organizaciones, especialmente en la industria tecnológica, están intentando mejorar su huella de carbono. Las empresas tecnológicas lideran la lista de empresas ecológicas de la Agencia de Protección Medioambiental (EPA), e IDC ha observado inversiones masivas en energías renovables en el sector tecnológico, así como inversiones en hardware y software de bajo consumo que ayudan a reducir el consumo de energía. IDC ha constatado que esto último ha ayudado a reducir el consumo de energía un 26 % como promedio.

En la actualidad, el 21 % de las organizaciones afirma estar invirtiendo en tecnologías que habilitan el proceso paralelo necesario para el entrenamiento y la inferencia en redes de aprendizaje profundo de IA.

[5] Fuente: IDC IT Infrastructure Plans for 2021 Survey, 2020

[6] Fuente: IDC IT Infrastructure for Compute Survey, 2021

Muchas empresas han seguido el ejemplo de los proveedores de servicios en la nube de aplicar un enfoque más sostenible a la TI, a saber, la reutilización y el reciclaje de sus equipos; el 33 % de los encuestados a una encuesta de IDC [7] afirmaron que esto tiene un papel clave para lograr una mayor sostenibilidad. La reutilización y el reciclaje de equipos pueden contribuir de forma significativa a la huella de carbono global de un centro de datos. Puede haber motivos para actualizar determinados componentes de un servidor, pero la cantidad de componentes nuevos imprescindibles entre dos generaciones de servidores no supera el número de componentes que podrían conservarse y reutilizarse.

Está surgiendo una mayor concienciación sobre la oportunidad de reutilización para reducir la huella medioambiental, e IDC ha previsto que, para 2025, el 90 % de las empresas del G2000 exigirán materiales reutilizables en las cadenas de suministro de hardware de TI, objetivos de neutralidad de carbono para las instalaciones de los proveedores y un menor uso de energía como requisitos previos para hacer negocios.[8] Estas medidas también permiten reducir los costes de las empresas, ya sea por el menor uso de energía o por la reducción de las inversiones de hardware.

La infraestructura de TI híbrida adecuada

Cambio hacia la nube híbrida

En la actualidad, el 54 % de las aplicaciones de las organizaciones siguen desplegadas en el entorno local.[9] IDC no ve que este porcentaje vaya a disminuir de forma significativa; las empresas afirman que en dos años esperan seguir ejecutando el 52 % de sus aplicaciones en el entorno local. De esas aplicaciones locales, el 56 % se ejecuta como una nube privada, una cifra que se espera que aumente al 60 % en dos años. En cuanto a si la nube privada cumple sus objetivos, el 61 % de las organizaciones afirma que no solo los cumple, sino que supera sus expectativas.

Muchas de estas aplicaciones, especialmente las aplicaciones empresariales críticas, tienen complejas interdependencias. De media, las empresas afirman que el 49 % de sus aplicaciones empresariales tienen algunas dependencias y el 27 % tienen interdependencias complejas. Hoy en día, solo el 18 % de todas las aplicaciones se consideran "nativas en la nube", lo que significa que son microservicios modulares y desagregados que representan suites de servicios desplegables independientemente. Por su parte, el 32 % de las aplicaciones siguen siendo monolíticas. Sin embargo, esto cambiará muy rápidamente. Las empresas afirman que en dos años, solo el 21% de las aplicaciones empresariales críticas serán monolíticas, mientras que el 44 % serán nativas en la nube.

Al mismo tiempo, las empresas esperan utilizar en el futuro diferentes despliegues de nube local y remota, lo que se denomina de nube "híbrida", que IDC considera un caso de ejemplo en rápido crecimiento. La **Figura 3** muestra que la combinación de nubes más común hoy en día es tener varias nubes entre las que migrar cargas de trabajo y datos. Para el caso de ejemplo de nube privada/nube pública, aproximadamente el 40 % de las organizaciones afirman que estos dos despliegues interoperan en sus organizaciones, es decir, que sirven como una nube híbrida más o menos integrada.

Obsérvese que para la parte local de una nube híbrida, las empresas desean mayoritariamente (84 %) cambiar de un modelo capex a otro opex. Actualmente, el 42 % de los presupuestos de TI de las empresas se financian con un método opex; hace tres años, esta cifra era del 36 %.

Obsérvese que para la parte local de una nube híbrida, las empresas desean mayoritariamente (84 %) cambiar de un modelo capex a otro opex.

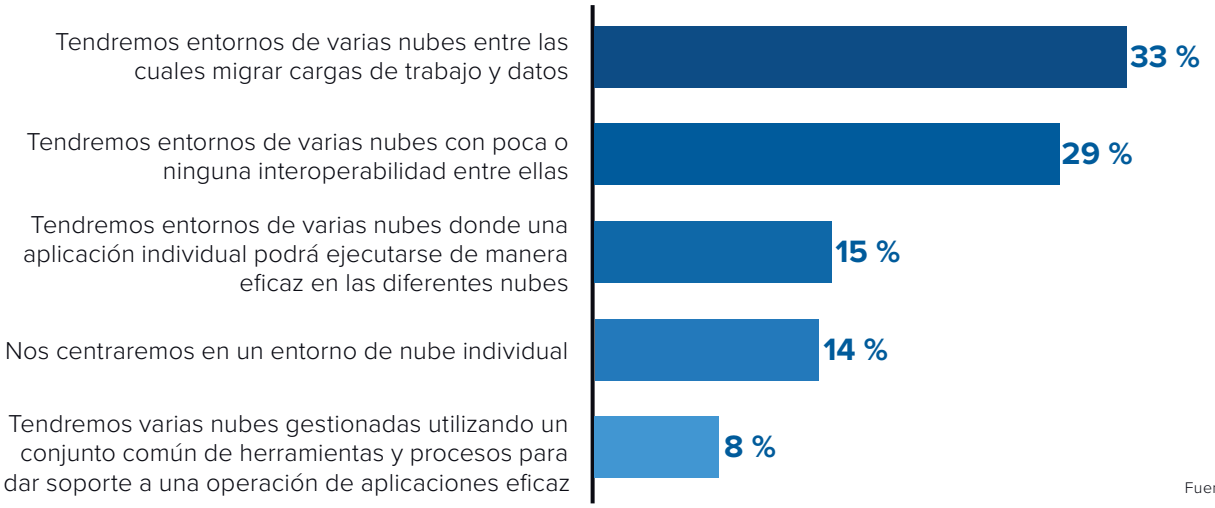
[7] Fuente: IDC 2021 Datacenter Operational Survey

[8] Fuente: IDC Worldwide Future of Digital Infrastructure 2021 Predictions

[9] Fuente: IDC 1Q21 Cloud Pulse Survey, May 2021

FIGURA 3

Uso de entornos de nube locales y remotos



Fuente: IDC, 2021

A medida que la nube híbrida es más prevalente, la repatriación desde una nube pública a una nube privada es también muy común: el 66 % de las empresas afirman que cambian aplicaciones a su nube privada o a entornos sin nube por varias razones, siendo las tres primeras el rendimiento, la seguridad y la disponibilidad (véase la **Figura 4**).

FIGURA 4

Razones para cambiar aplicaciones de IaaS a la nube privada o a un entorno sin nube

Fuente: IDC, 2021

Nube híbrida y aplicaciones nativas en la nube

Una nube híbrida correctamente diseñada es una plataforma ideal para desarrollar y ejecutar aplicaciones nativas en la nube, algo que cada vez más empresas consideran una funcionalidad clave para su transformación digital. IDC ha descubierto que la mayoría de las empresas consideran que la implementación de varias prestaciones es de "importante" a "extremadamente importante" para cumplir sus necesidades empresariales, ya que invierten en una estrategia de nube adecuada para desarrollar y ejecutar aplicaciones nativas en la nube. Estas capacidades son:

- Mayor rendimiento, disponibilidad, portabilidad y dirección de las aplicaciones
- Mejora de la integración de datos, la orquestación, la observabilidad, la gestión de API y AIOps en los entornos de nube
- Ciclos de desarrollo y tiempos de comercialización más rápidos con CI/CD (desarrollo y despliegue continuos) y automatización
- Políticas integrales de seguridad, gestión de riesgos, estrategias de recuperación tras el desastre y conformidad con la normativa
- Un modelo opex en lugar de capex, que incluye prestaciones de reembolso
- Optimización de la productividad, eficiencia y habilidades del personal

Las empresas que deseen aumentar su inversión en una nube híbrida deben "colocar marcas" al lado de estos puntos para asegurarse el poder conseguir el ROI que tienen previsto.

La importancia de la IA y dónde ejecutarla

IDC prevé que el mercado mundial de plataformas de servidores de inteligencia artificial crecerá hasta los 27 000 millones de dólares en 2025.[10]

Este crecimiento se verá impulsado por la creciente adopción de tecnologías conversacionales, procesamiento del lenguaje natural (NLP), análisis de imágenes y vídeos, aprendizaje profundo, aprendizaje automático (ML), generación de hipótesis y análisis predictivo. Como resultado, las plataformas de servidores de IA constituirán el 21 % del total del mercado mundial de servidores en 2025.

En una sección anterior, hemos analizado la creciente necesidad de coprocesadores para poder ejecutar cargas de trabajo de entrenamiento de IA e inferencia de IA. Dado que la nube privada local es el escenario de despliegue preferido para la IA, y los entornos locales sin nube el segundo más común, esto se traduce directamente en importantes inversiones para las empresas en términos de GPU, FPGA y ASIC añadidos. Para el entrenamiento de IA, estas inversiones son más o menos inevitables; el entrenamiento de un algoritmo DNN simplemente no se puede realizar en un procesador de host. No obstante, para la inferencia de IA, hay muchos modelos de IA que funcionarán muy bien en un procesador de host avanzado o en un procesador de host con un procesador de IA de especialidad integrado. Estos casos de ejemplo tienen una clara ventaja en el coste para las empresas, ya que añadir unas pocas GPU a un servidor puede duplicar rápidamente el precio del paquete total.

Dado que la nube privada local es el escenario de despliegue preferido para la IA, y los entornos locales sin nube el segundo más común, esto se traduce directamente en importantes inversiones.

Desde luego, esto invita a preguntarse por qué las empresas continúan ejecutando sus aplicaciones de IA en el entorno local en primer lugar. ¿Por qué no las ejecutan en la nube, por ejemplo, para evitar capex por completo? Ciertamente, una parte del entrenamiento de IA se realiza en nubes públicas en las plataformas de IA de los proveedores y, una vez desarrollados, estos modelos a veces permanecen en la nube como cargas de trabajo de producción.

[10] IDC Worldwide AI Server Forecast, 2021–2025, julio de 202

El factor más importante que determina la nube frente al entorno local son los datos, y las siguientes preguntas constituyen el motivo subyacente:

¿Cuáles son los datos necesarios para desarrollar el modelo?
Si se trata de datos de aplicaciones básicas de la empresa como, por ejemplo, los datos transaccionales, es preferible permanecer en la plataforma transaccional, incluso por motivos de latencia.

¿Qué grado de confidencialidad tienen los datos?
Si los datos son confidenciales, lo que significa que deben protegerse rigurosamente, no se recomienda moverlos a la nube, ni para el entrenamiento ni para la inferencia.

¿Cuál es la infraestructura normativa de los datos?
Algunos datos no pueden moverse por ley a una nube pública; por regla general, ocurre muy a menudo con los datos empresariales básicos. Las empresas están sujetas a todo tipo de regulaciones, desde las normativas nacionales de protección de datos hasta el RGPD, pasando por regulaciones industriales como la HIPPA, normativas de ISO o la Ley de Protección al Consumidor de California.

¿Qué se puede hacer y qué no con los datos para que sigan cumpliendo la normativa?
Una vez que los datos empiezan a moverse de un lado a otro, es más difícil asegurarse de que la organización siga cumpliendo la normativa.

¿Qué volumen de datos hay?
Cuantos más datos se requieran para el entrenamiento, o cuantos más datos utilice la inferencia del modelo de IA, especialmente si la inferencia es en tiempo casi real, más difícil será hacerlo en la nube.

¿Qué grado de integración tienen las aplicaciones que utilizan los datos?
La plataforma que ejecuta las transacciones seguramente tendrá varias aplicaciones profundamente integradas con la base de datos para ejecutar la analítica y otras funciones, lo que dificulta mover los datos a la nube.

¿Cuánto cuesta el almacenamiento de los datos?
El almacenamiento en grandes volúmenes en la nube puede superar rápidamente el gasto de capital que sería necesario para el almacenamiento en el entorno local.

En conjunto, estas consideraciones pueden llevar a muchas organizaciones a permanecer en el entorno local con las cargas de trabajo de inferencia y entrenamiento de IA. Pueden seguir entrenando en un entorno aparte en el centro de datos detrás de sus cortafuegos, pero luego trasladar el modelo entrenado de nuevo a la plataforma que ejecuta las aplicaciones empresariales básicas para la inferencia. Si la plataforma permite una inferencia sólida, las empresas pueden utilizar la IA en datos básicos que habrían sido inaccesibles en el pasado.

IBM Power10 e IBM Power E1080

Para transformarse con éxito en un negocio digital, las empresas necesitan plataformas de proceso que puedan absorber cualquier tipo de volatilidad del mercado, que sean seguras, sin riesgos, que se escalen sin problemas a la vez que reducen la huella física y de carbono de las empresas, que proporcionen el máximo nivel de resiliencia, y que puedan ejecutar IA en tiempo real en un gran número de transacciones, todo como parte de una nube híbrida eficaz. El nuevo procesador Power10 de IBM y la plataforma IBM Power E1080 de clase empresarial basada en Power10 ofrecen una amplia gama de innovaciones que cumplen estos requisitos con metodologías nuevas e interesantes.

El nuevo procesador Power10

La nueva arquitectura y el nuevo procesador IBM Power10 de IBM incluye importantes nuevas tecnologías que ayudarán a las empresas con cargas de trabajo que requieren muchos procesos, memoria y ancho de banda, incluidas las nuevas tecnologías de inferencia rápida de IA en el chip sin hardware adicional, basadas en un acelerador matemático de matriz (MMA) incorporado diseñado específicamente.

Desde el punto de vista de la seguridad, Power10 implementa el cifrado de memoria sin degradación del rendimiento (a diferencia del cifrado de memoria basado en software); proporciona seguridad de contenedor cooptimizada por hardware/software para el aislamiento del contenedor; e incluye características de seguridad para anticipar la funcionalidad inminente de computación cuántica para romper las claves de cifrado tradicionales.

La escalabilidad con Power10 se amplía a nuevos niveles con varias innovaciones de ancho de banda. IBM ha mejorado la tecnología de conectividad POWER AXON y ha añadido la Interfaz de memoria abierta (OMI), ambas funcionando a 32 GT/s. La interfaz de Power10 AXON conecta hasta 16 sockets a un gran sistema escalable. La OMI se comunica con la memoria DRAM DDR4 a través de 16 puertos DDR por socket, lo que proporciona un ancho de banda de hasta 409 GB/s por socket. Estas dos interfaces pueden utilizarse para proporcionar soluciones de cálculo muy flexibles e incluso componibles.

Es el primer procesador de 7 nanómetros de IBM, que afirma que su eficiencia es 3 veces superior a la de IBM Power9 en términos de potencia de proceso (número de usuarios, número de transacciones) y energía.[11] Con la atención constante de IBM a la nube híbrida, esto se traduce directamente en una menor huella en el centro de datos y una reducción significativa de la energía. Hay 15 núcleos de procesador en el chip, y Power10 incluirá PCI Gen5, que está empezando a emerger en la industria.

IBM Power E1080

IBM Power E1080 es la primera plataforma de clase empresarial de IBM creada con el procesador Power10. El sistema puede escalarse a hasta 16 procesadores y está claramente dedicado a las principales consideraciones de TI para las organizaciones que deben cumplir las demandas de la empresa digital.

Seguridad

Para que la seguridad sea persistente y no tenga penalizaciones, IBM ha incorporado el cifrado en el procesador Power10. Esto permite cifrar los datos sin comprometer el rendimiento del sistema. Además, el sistema ha sido equipado con características de seguridad adicionales para protegerlo contra los ataques de programación orientados al retorno,

[11] El rendimiento 3X se basa en el análisis de ingeniería previo al silicio de los entornos Integer, Enterprise y Floating Point en una oferta de servidor de socket dual POWER10 con 2 módulos de 30 núcleos frente a la oferta de servidor de socket dual POWER9 con 2 módulos de 12 núcleos; ambos módulos tienen el mismo nivel de energía.

una técnica mediante la cual un atacante puede ejecutar código malicioso en presencia de defensas de seguridad. Power E1080 ofrece protección avanzada de datos con cifrado de memoria transparente, el tipo de seguridad a nivel de hardware para los datos en uso en el que se basa la informática confidencial, y cuenta con cuatro veces más aceleradores de cifrado criptográfico que su predecesor. Las particiones en la plataforma han mejorado el aislamiento, y el sistema está protegido de futuras amenazas basadas en computación cuántica con criptografía postcuántica (PQC) y cifrado totalmente homomórfico, una tecnología donde las entradas en el sistema no necesitan descifrarse, lo que significa que una parte no fiable puede ejecutarlas sin revelarlas.

Resiliencia

IDC considera que la familia de servidores Power de clase empresarial tiene AL4, es decir, que es totalmente tolerante a errores y, por lo tanto, ofrece un 99,999% o más de disponibilidad. Con Power10, IBM Power E1080 va un paso más allá que su predecesor al ofrecer un ancho de banda muy alto y fiabilidad, disponibilidad y capacidad de servicio (RAS) de memoria con la nueva interfaz de memoria abierta. El procesador puede detectar errores leves, aislarlos y recuperarse automáticamente sin interrupciones o sin depender del sistema operativo para gestionar los fallos y reparar automáticamente los errores recuperables. El sistema también cuenta con prestaciones mejoradas de reparación simultánea como, por ejemplo, los cables subminiatura push-on (SMP) entre nodos para reducir el tiempo de inactividad de la aplicación.

Escalabilidad y sostenibilidad

En términos de escalabilidad y sostenibilidad, IBM Power E1080 se beneficia enormemente del hecho de que la familia de servidores Power está excepcionalmente bien integrada desde el procesador hasta el firmware, pasando por el SO y el hardware, ya que son todos componentes de IBM. La eficiencia del software y el contenedor OpenShift de la plataforma es excepcional, según IBM. Como resultado, la plataforma con el nuevo procesador Power10 consigue un 50 % más de rendimiento en la misma huella física y de energía si se compara con Power E980.[12] Esto también se traduce en un 33 % menos de consumo de energía para la misma carga de trabajo, afirma IBM.[13] La mayor eficiencia permite a las empresas reducir significativamente su huella de carbono y consolidar potencialmente las cargas de trabajo, lo que permite un ahorro de costes de hardware y software.

Cloud híbrido

Power E1080 da soporte a tres entornos operativos (AIX, IBM i y Linux) en la misma plataforma, y está diseñado para permitir la adopción de nube híbrida de las empresas para los tres entornos operativos. Desde luego, AIX es el sistema operativo Unix totalmente modernizado de IBM que sigue siendo la plataforma preferida para la plataforma Power de escalado y clase empresarial. IBM i es el entorno operativo de IBM que integra la base de datos y otro software de empresa en el sistema operativo, lo que simplifica enormemente la gestión de la plataforma; para muchas medianas empresas, IBM i es el núcleo de sus operaciones. AIX e IBM i dan soporte absoluto al código abierto, admiten los modernos lenguajes preferidos por los desarrolladores y pueden operarse totalmente como una nube híbrida. Al igual que las generaciones anteriores, Power E1080 también puede ejecutarse total o parcialmente en Linux con las mismas características de seguridad, disponibilidad y escalabilidad, lo que representa una gran oportunidad para las empresas que deseen mover sus cargas de trabajo transaccionales y analíticas a una plataforma totalmente de código abierto.

Los siguientes componentes de software de IBM Power desempeñan un papel importante al permitir a las empresas optimizar su plataforma Power de nivel empresarial con AIX, IBM i y Linux, para garantizar una modernización de la carga de trabajo segura, basada en la nube y altamente disponible:

IBM PowerVM

Las cargas de trabajo del servidor IBM Power están virtualizadas, son móviles y están totalmente habilitadas para la nube con PowerVM, que se ha mejorado recientemente con nuevas funciones, entre las que se incluyen la Compresión y el Cifrado de datos de Movilidad de particiones activas (LPM), lo que significa que cuando una partición activa se migra desde un servidor Power a otro, lo que se produce con un tiempo de inactividad nulo, los datos se cifran y se comprimen automáticamente. Se trata de una importante función de seguridad y rendimiento.

IBM PowerVC

PowerVC es una herramienta de virtualización de gestión basada en OpenStack, que simplifica la gestión de los recursos virtuales en los entornos Power. El software ha sido mejorado recientemente con varias características nuevas, incluida una funcionalidad de exportación/importación para compartir imágenes de VM entre centros de datos.

[12] Información proporcionada por IBM. Basada en los resultados publicados de rPerf para Power E980/12 núcleos en comparación con las mediciones internas de rPerf de IBM (utilizando la misma metodología) para Power E1080/15 núcleos.

[13] Power9 (12c) es 5081 rPerf @ 16.520 vatios (0,31 rPerf/W), Power10 (15c) es 7998 rPerf @ 17.320 vatios (0,46 rPerf/W) 0,46 / 0,31 = 1,48 Más rPerf/W

IBM PowerSC

PowerSC es el portfolio de seguridad de la plataforma, que simplifica la gestión de la seguridad y la conformidad, e incluye automatización de conformidad, detección de intrusiones de programas maliciosos, gestión de parches, etc. Se ha mejorado con varias características o incluso nuevas ofertas, incluida la habilitación de la autenticación multifactorial (MFA), otra importante función de seguridad. En general, la seguridad en IBM Power con AIX se consigue con una solución integral que incluye el procesador, el firmware, el hipervisor y las innumerables características de seguridad del propio sistema operativo para proteger los datos en todos los niveles.

IBM PowerHA y VM Recovery Manager HA y DR

PowerHA es una tecnología de alta disponibilidad que ayuda a proporcionar disponibilidad casi continúa a las aplicaciones y a mejorar la fiabilidad de los servicios. Es un colaborador clave de IBM Enterprise Power, está designado por IDC como tolerante a fallos (AL4) y se ha mejorado con varias características como, por ejemplo, las métricas de migración tras error mejoradas y la verificación entre distintos clúster (por ejemplo, para comparar un clúster de desarrollo con un clúster de prueba). VM Recovery Manager (VMRM) es una solución HA/DR simplificada, basada en la réplica y el reinicio de VM independientes del sistema operativo, e incluye agentes de supervisión de aplicaciones para DB2, Oracle y SAP HANA.

Cloud Management Console

Cloud Management Console (CMC) proporciona una visión completa sobre el rendimiento, el inventario y el registro de la infraestructura Power local y remota. CMC se aloja en IBM Cloud, lo que libera a las empresas de tener que mantener un software para supervisar su infraestructura, y permite simplificar la dirección de despliegues de nube híbrida y la supervisión y gestión de su infraestructura.

Enterprise Cloud Edition 2.0

Enterprise Cloud Edition reúne todos los componentes clave de una infraestructura de gestión de nube simplificada además de PowerVM, por ejemplo, PowerSC, MFA, PowerVC, CMC, VMRM y Aspera. Permite el despliegue y la gestión rápidos de una nube privada; la gestión de seguridad y conformidad simplificada; alta disponibilidad simplificada; y transferencias aceleradas de archivos de gran tamaño a través de las nubes. Enterprise Cloud 2.0 puede adquirirse con AIX 7.2 incorporado.

Red Hat Ansible Automation Platform

Red Hat Ansible Automation Platform permite escalar y asegurar la automatización de varios aspectos de las operaciones de TI, incluidos el suministro de recursos, la gestión del ciclo de vida de las aplicaciones y las operaciones de red. Está formado por Ansible Engine, Ansible Tower y Ansible Hosted Services. Los demás productos de la cartera de Red Hat pueden integrarse utilizando Red Hat Ansible Automation Platform. Red Hat Ansible Automation Platform permite la coherencia en el centro de datos al proporcionar métodos programáticos para desplegar, gestionar y proteger los recursos de la infraestructura.

Red Hat OpenShift

Red Hat OpenShift es una plataforma de nivel empresarial, certificada por Kubernetes (una orquestación de contenedores) para generar, desplegar y gestionar aplicaciones contenerizadas. Red Hat OpenShift puede consumirse como un servicio totalmente gestionado en diferentes proveedores de nube, o el cliente puede gestionarlo utilizando Red Hat OpenShift Container Platform o Red Hat OpenShift Kubernetes Engine. Puede desplegarse en el entorno local en servidores bare metal, en las plataformas de virtualización (Red Hat Virtualización, VMware o Red Hat OpenStack) o en los principales proveedores de nube como IBM Cloud, AWS, Google o Azure. Asimismo, Red Hat Advanced Cluster Management for Kubernetes puede utilizarse para gestionar varios clústeres y aplicaciones de Red Hat OpenShift desde una única consola, con políticas de seguridad incorporadas, lo que habilita a los clientes en la nube híbrida abierta. Red Hat OpenShift es compatible con IBM Power, IBM Z y las plataformas basadas en x86, y puede utilizarse con AIX, IBM i y Linux.

IBM Cloud Paks

IBM Cloud Paks son productos de software cada vez más populares preempaquetados en contenedores y altamente integrados en varios servicios de OpenShift para garantizar un despliegue rápido y sencillo en OpenShift. Los productos IBM Cloud Paks ofrecen herramientas de desarrollador, datos y servicios de inteligencia artificial, así como un software middleware de código abierto. Se ejecutan en la plataforma de nube Red Hat OpenShift.

Algunos Cloud Paks que son particularmente relevantes para IBM Power son:

- ***Cloud Pak for Data:*** ayuda a los clientes con información útil ampliada de los datos y las prestaciones de IA
- ***Cloud Pak for Integration:*** consiste en herramientas de integración de datos, servicios de aplicación y servicios en la nube, para ayudar a integrar aplicaciones, datos, servicios en la nube y API
- ***Cloud Pak for Watson AIOps:*** ofrece visibilidad, gobierno y automatización de las distintas nubes, a partir del uso común de despliegues en varias nubes

Inteligencia Artificial

IBM afirma que Power E1080 acelera la inferencia de IA en un orden de magnitud en comparación con su predecesor. No requiere ningún hardware especializado como, por ejemplo, un coprocesador (GPU, FPGA o ASIC). En su lugar, la inferencia tiene lugar en un acelerador matemático de matriz (MMA). Cada núcleo del chip Power10 tiene un MMA incorporado para realizar correctamente las operaciones matemáticas de matriz. Estas operaciones se han optimizado en una amplia gama de tipos de datos para distintas precisiones, que son importantes para el aprendizaje profundo, desde la precisión doble y la precisión simple hasta dos tipos de precisión media, incluidos Bfloat-16, Int-16, Int-8 e Int-4. El rendimiento de inferencia de IA se ha incorporado en cada capa del procesador. La memoria caché L2 se ha cuadruplicado: las unidades de almacén de carga y los SIMD se han duplicado. Esto significa que una carga de trabajo transaccional que tenga componentes de IA incorporados puede ejecutar las transacciones y la inferencia de IA en el mismo procesador Power10 sin necesidad de un coprocesador.

La inferencia en el chip también significa que todas las características de seguridad del procesador y el sistema están disponibles para proteger los datos en los que se aplica la inferencia. Asimismo, la plataforma es compatible con Open Neural Network Exchange (ONNX). ONNX es un ecosistema de IA de código abierto de empresas tecnológicas y organizaciones de investigación que trabajan para establecer estándares abiertos para representar algoritmos y herramientas de IA con el fin de promover la innovación y la colaboración en el sector de la IA. Las empresas con IBM Power E1080 pueden incorporar sin cambios los modelos ONNX en la plataforma y ejecutarlos, aprovechando las características RAS de la plataforma durante la inferencia.

Desafíos y oportunidades

Para las empresas

Las plataformas de clase empresarial que ejecutan las cargas de trabajo analíticas y transaccionales básicas de una organización tienden a tratarse como silos en el centro de datos, aunque estén diseñadas y creadas con características y tecnologías integrales para evitarlo. El personal de TI que tiene una amplia experiencia con el sistema, pero que desconfía de la exposición de los datos, la integración de la plataforma con la nube, la ejecución de código abierto en la plataforma y la ejecución de modelos de IA en datos en tiempo real a menudo "protege" estas plataformas de las nuevas tecnologías. Para las empresas, el reto es superar esta cultura de la indecisión lo antes posible. Es absolutamente fundamental que las plataformas de clase empresarial sean apreciadas como los sistemas abiertos que son; esto permitirá a las empresas optimizarlas completamente como plataformas de transformación digital que crean nuevas oportunidades de ingresos. Al mismo tiempo, estas plataformas ofrecen la oportunidad de empezar a gestionar seriamente los problemas de sostenibilidad, para reducir la huella de carbono de la organización. La ejecución de la IA en una plataforma de empresa sin necesidad de costosos coprocesadores con un gran consumo de energía es un requisito importante, ya que se está incorporando cada vez más la funcionalidad de IA en las aplicaciones básicas.

Para IBM

Con la nueva plataforma Power E1080, IBM continúa fomentando en las empresas el código abierto, la nube híbrida, la IA y la sostenibilidad en una plataforma altamente segura, eficiente y fiable. IBM suele afrontar los retos de innovación con nuevas e interesantes tecnologías que, en algunos casos, son rompedoras y se adelantan a la competencia, como ocurre, por ejemplo, con MMA en el nuevo procesador Power10. La innovación no es el mayor reto de IBM. El verdadero reto de IBM es cambiar una parte de la mentalidad de sus clientes de forma que pasen de tratar su plataforma empresarial como un sistema en silos, o quizás un sistema cuidadosamente abierto, a tratarla como una plataforma totalmente integrada con el resto del centro de datos y con la nube, que aproveche plenamente todas sus prestaciones para innovar y generar más ingresos con los datos básicos que residen en la plataforma. IBM necesita seguir animando a sus clientes a ser atrevidos y creativos con su plataforma empresarial a través de estudios de formación, incentivos y ROI.

Conclusión

Las empresas modernas necesitan plataformas de proceso que puedan manejar la extrema volatilidad del mercado, proporcionar una seguridad férrea, escalarse sin esfuerzo y de forma sostenible, permitir la máxima resiliencia, ejecutar IA en tiempo real y operar como una nube híbrida. El nuevo procesador Power10 de IBM y la plataforma de clase empresarial IBM Power E1080 basada en Power10 cumplen estos requisitos sin problema. La nueva generación de procesadores Power de IBM no es un simple paso más y se aventura en un importante territorio de cara al futuro.

El procesador habilita la tecnología de informática confidencial con cifrado basado en hardware que protege los datos en movimiento. El ancho de banda en Power10 se ha aumentado considerablemente para habilitar una potente escalabilidad de 16 sockets. La resiliencia se ve reforzada con la capacidad de detectar errores leves, aislarlos y recuperarse automáticamente sin interrupciones o sin depender del sistema operativo. El MMA en el chip permite la inferencia de IA en tiempo real sin necesidad de un coprocesador. Entre las soluciones de Red Hat y el software de IBM Cloud, la capacidad de operar plenamente como una nube híbrida es un hecho. Con el chip Power10 como motor de la nueva plataforma Power E1080, IBM sigue llevando a la informática empresarial a un punto óptimo donde se une lo mejor de todos los mundos: código abierto, potencia de procesamiento, nube híbrida, IA, seguridad, escalabilidad, sostenibilidad y fiabilidad en una única plataforma.

Acerca del analista

Peter Rutten

Director de investigación, Sistemas de infraestructura, Grupo de plataformas y tecnologías,
Líder de investigación global de soluciones informáticas de alto rendimiento, IDC

Peter Rutten es Director de investigación en Worldwide Infrastructure Practice de IDC, que cubre la investigación en plataformas informáticas. El Sr. Rutten es el líder de investigación global de IDC sobre soluciones informáticas y casos de uso con un rendimiento intensivo. Incluye investigación sobre Inteligencia Artificial (IA), Modelado y Simulación (M&S), y la infraestructura de Big Data y analítica (BDA), y soluciones asociadas. Su cobertura de la informática con un rendimiento intensivo incluye sistemas, plataformas y tecnologías de infraestructura informática heterogénea, en memoria, acelerada, de gama alta y supercomputación. Incluye plataformas de proceso con GPU, FPGA, ASIC y otros aceleradores que se despliegan tanto en la nube como en el entorno local. También incluye la investigación sobre plataformas x86 de misión crítica, mainframes y sistemas basados en RISC, así como sus entornos operativos (Linux, z/OS, Unix). El Sr. Rutten también examina las tecnologías y plataformas emergentes, como la computación cuántica, la computación neuromórfica y tecnologías potencialmente disruptivas para los mercados de infraestructura avanzada. Como parte de su trabajo, el Sr. Rutten realiza un análisis cuantitativo (dimensionamiento y previsión del mercado) y cualitativo (basado en la investigación primaria) , así como un dimensionamiento personalizado del mercado para los clientes de IDC.

Más información sobre Peter Rutten

IDC Custom Solutions

Esta publicación ha sido elaborada por IDC Custom Solutions. Como principal proveedor mundial de inteligencia de mercado, servicios de asesoría y eventos para los mercados de tecnologías de la información, telecomunicaciones y tecnología del consumidor, el grupo de soluciones personalizadas de IDC ayuda a sus clientes a planificar, comercializar, vender y tener éxito en el mercado global. Creamos programas accionables de inteligencia de mercado y marketing de contenidos influyentes que ofrecen resultados medibles.

IDC @idc @idc idc.com