O ponto ideal da computação corporativa moderna

Pesquisa por:

Peter Rutten
Diretor de pesquisa, Grupo de sistemas de infraestrutura, Plataformas e tecnologias, Líder de Pesquisa global de soluções de computação intensiva de desempenho, IDC

Setembro de 2021 | IDC Doc. #US48197921

Navegando neste boletim informativo

Clique nos títulos ou nos números de página para navegar para cada seção.

Opinião da IDC

O cenário de TI de hoje pode parecer um enigma. No caminho para se tornar uma empresa digital e satisfazer as necessidades de clientes extremamente exigentes, as empresas se encontram tentando alcançar o quase impossível.

- Os mercados podem mudar por capricho, causando picos ou depressões, e essa volatilidade não pode ser interpretada como uma exceção. Volatilidade é a referência de hoje.
- Para atender ao fluxo e refluxo da carga de trabalho, os sistemas devem ser dimensionados sem falhas e dinamicamente, sem exigir uma expansão de um datacenter enorme e cara, que consome energia apenas nos picos. Sustentabilidade já não é apenas um truque de marketing.
- A complexidade desses mercados também não pode mais ser analisada e alavancada com experiência e inteligência humanas comuns. Grande parte da inteligência agora deve ser artificial, operando em tempo real e fazendo malabarismos com inúmeras variáveis enquanto obtém grandes quantidades de dados. A inteligência artificial (IA) vai infundir cada vez mais tudo em todos os lugares, e ela requer recursos de hardware específicos.
- Dada a demanda por disponibilidade permanente e perpétua, as cargas de trabalho que sustentam a empresa digital não podem ser reduzidas ou impedidas, muito menos totalmente. No mundo de hoje, qualquer tempo de inatividade pode ser catastrófico.
- Com tudo digital e conectado para permitir a empresa digital, tudo também foi exposto e pode ser comprometido por novos tipos de ataques. Comunidades inteiras de pessoas mal-intencionadas se aglutinaram em um submundo que trava uma guerra permanente contra empresas em todo o mundo usando um vasto arsenal de ferramentas e estratégias de ataque cibernético. Como resultado, tudo agora deve começar com segurança abrangente e estanque.

Com isso em mente, para que qualquer plataforma de computação de classe corporativa seja capaz de funcionar como o motor da empresa digital, ela deve ser inequivocamente segura, confiável, escalável, sustentável, integrativa com a nuvem como parte de uma abordagem híbrida e construída para IA. Este Boletim Informativo aprofundará essas considerações a partir de uma perspectiva de infraestrutura e implementação e analisará como o novo processador IBM Power10 e a nova plataforma de classe corporativa IBM Power, o E1080, são executados neles.

Visão geral

A IDC acredita que para uma empresa digital ter sucesso no ambiente multifacetado e desafiador de hoje, as seguintes considerações são críticas:

- **Segurança como um requisito imperativo**
- **O mandato de confiabilidade**
- **Escalabilidade e sustentabilidade**
- **A infraestrutura de TI híbrida certa (nuvem híbrida e aplicações nativas da nuvem)**
- **A importância da IA e onde executá-la**

As próximas seções se aprofundarão em cada uma dessas considerações.

Segurança como um requisito imperativo

A segurança se tornou o requisito mais importante de uma empresa digital. Quando a IDC pesquisa organizações sobre suas prioridades, a segurança está invariavelmente no topo da lista ou próximo a ela. De fato, quando solicitados, por exemplo, a selecionar os principais itens de infraestrutura de IA que as empresas consideram não serem os melhores nas ofertas de seus fornecedores ou provedores de servidor e armazenamento, a pontuação de segurança é mais alta, com 30% dizendo que estão insatisfeitos com os recursos de segurança.[1]

Esse descontentamento também é demonstrado pelo fato de que muitas empresas não permitem que os dispositivos de armazenamento que contêm os dados de suas cargas de trabalho de IA sejam usados por outras cargas de trabalho. A razão mais frequentemente apresentada para isso (45%) é a segurança e privacidade de dados. Além disso, a pesquisa do IDC descobriu que a segurança é uma das principais preocupações na infraestrutura de nuvem pública como serviço (IaaS), com 37% das empresas afirmando que a segurança é sua maior preocupação nessas implementações.[2] As empresas também estão infundindo cada vez mais suas cargas de trabalho de segurança com IA, mais do que qualquer outra carga de trabalho, para torná-las mais capazes de prever e agir em violações.

Atualmente, a maior parte da atenção e do investimento está voltada para a segurança das aplicações e das pilhas de rede. Um grande número de ataques, no entanto, são de baixo nível e centrados no hardware. Eles geralmente são iniciados aproveitando as vulnerabilidades nos processadores e/ou microcódigo. Esses ataques são sofisticados e difíceis de detectar.

Portanto, a IDC está vendo as empresas se tornarem cada vez mais interessadas em “computação confidencial” para suas plataformas críticas de negócios. A computação confidencial permite o isolamento de dados sensíveis para um subsistema de processador designado e protegido (às vezes chamado de um "enclave de processador seguro") para processamento. Hoje, os dados geralmente são criptografados em repouso no armazenamento e em trânsito pela rede, mas não durante o uso na memória.

[1] Fonte: IDC AI Infrastructure View 2021

[2] Fonte: IDC IaaSView 2020

A capacidade de proteger dados e código enquanto estão na memória é limitada em muitas plataformas de computação. Ainda assim, as organizações que lidam com dados confidenciais, como informações de identificação pessoal (PII), dados financeiros ou informações de saúde, precisam mitigar ameaças que visam a aplicação ou os dados na memória do sistema.

Na computação confidencial, o conteúdo do subsistema, que pode ser criptografado no nível do hardware, é acessível apenas para código autorizado dentro de um programa. O conteúdo é inacessível para qualquer coisa fora dele, incluindo outro código, outros sistemas ou outros operadores. Entidades não autorizadas não podem visualizar ou adulterar os dados ou o processo de execução de código autorizado. Uma solução de computação confidencial abrangente protegerá os dados em uso e em repouso. Isso pode ser habilitado pela criptografia de conteúdo em memória de sistema volátil ou não volátil e armazenamentos de dados persistentes, em mídia flash ou rotacional.

As infraestruturas de computação confidenciais modernas, especialmente aquelas implementadas em ambientes diversos locatários compartilhados, fazem uso de coprocessadores discretos para descarregar operações de processador privilegiadas que podem ser comprometidas por vulnerabilidades de execução de código de baixo nível. Essa ainda não é uma abordagem comum, mas para cargas de trabalho corporativas centrais, ela é uma promessa significativa. Nesse ínterim, as empresas estão aproveitando várias estratégias de segurança simultaneamente, com hardware e software.

O mandato de confiabilidade

Embora as estratégias de segurança sejam extremamente importantes na proteção de dados, aplicações e hardware contra ataques, outro aspecto crucial da empresa digital é a confiabilidade absoluta do ambiente de TI, incluindo a infraestrutura. Alta disponibilidade dificilmente é um conceito novo, e as empresas podem escolher plataformas de computação com disponibilidade de até 99,999% e plataformas de armazenamento com 99,99999%. Mas esses números apenas são alcançados com o hardware, software e políticas corretos. IDC designou apenas nove plataformas de servidor de seis fornecedores no mercado de servidores como Availability Level 4 (AL4)[3], que é o nível mais alto e representa tolerância total a falhas.

- A pesquisa da IDC[4] mostra que as três principais causas de tempo de inatividade de aplicações são falha na rede (16,2%), falha nos servidores (15,5%) e malware (10,3%). Entre as causas mais comuns de falha do servidor estão sobrecarga de memória (DRAM) ou CPUs e falha ou corrupção de memória.
- Os volumes de transações estão aumentando drasticamente e as empresas precisam de velocidades de transação cada vez mais rápidas para satisfazer seus clientes.
- As cargas de trabalho essenciais para a missão e os negócios estão crescendo, e as funções de suporte aos negócios que antes podiam ser executadas em um nível de baixa disponibilidade — por exemplo, por meio de virtualização ou armazenamento em cluster — são cada vez mais considerados críticos para os negócios.
- O custo do tempo de inatividade está aumentando à medida que as empresas se tornam cada vez mais dependentes de sua infraestrutura para as operações diárias. A pesquisa da IDC mostra que para 20,7% das organizações, o custo do tempo de inatividade é de US$ 5.000 a 10.000 por hora; para 18,4% delas, é de US$ 10.000 a US$ 25.000 por hora; para 17% delas, é de US$ 25.000 a US$ 100.000 por hora e para algumas empresas (1,4%), é de US$ 500.000 por hora.
- O fim do "horário comercial normal", com as aplicações das empresas agora obrigados a estar disponíveis aos clientes o tempo todo, colocou uma pressão tremenda sobre a infraestrutura que dá suporte a essas aplicações, permitindo pouco ou nenhum tempo de inatividade programado ou não programado.
- A tolerância quanto a interrupções, atrasos, perda de dados e corrupção de dados é zero - de empresas e consumidores - e quaisquer violações ou erros podem ter consequências catastróficas para a reputação de uma organização.

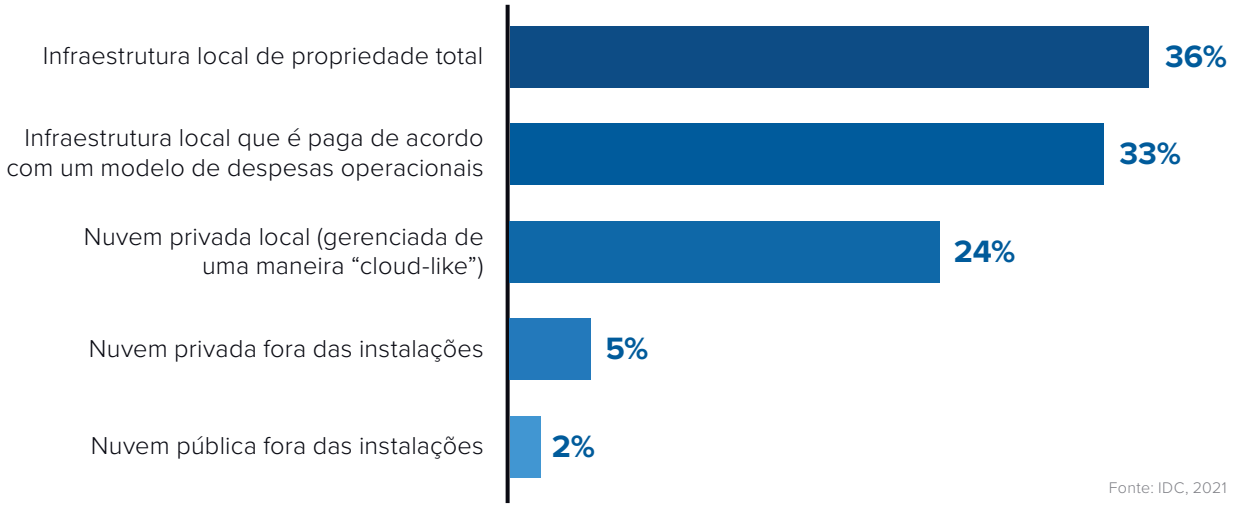
[3] Fonte: IDC Worldwide AL4 Server Market Shares, 2019: *Fault-Tolerant Systems Become Digital Transformation Platforms*

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

- À medida que as empresas se envolvem digitalmente com consumidores ou cidadãos e outras empresas com mais frequência e de muitas maneiras mais diversas, a conformidade com os regulamentos nacionais e internacionais sobre disponibilidade, segurança e privacidade de dados é de suma importância.
- Mesmo que a disponibilidade e a segurança na nuvem pública tenham melhorado muito, a verdadeira tolerância a falhas continua a ser vista como um recurso de nuvem híbrida ou local, não como um recurso de nuvem pública (consulte a **Figura 1**).

FIGURA 1

Infraestrutura que hospeda o nível de maior disponibilidade

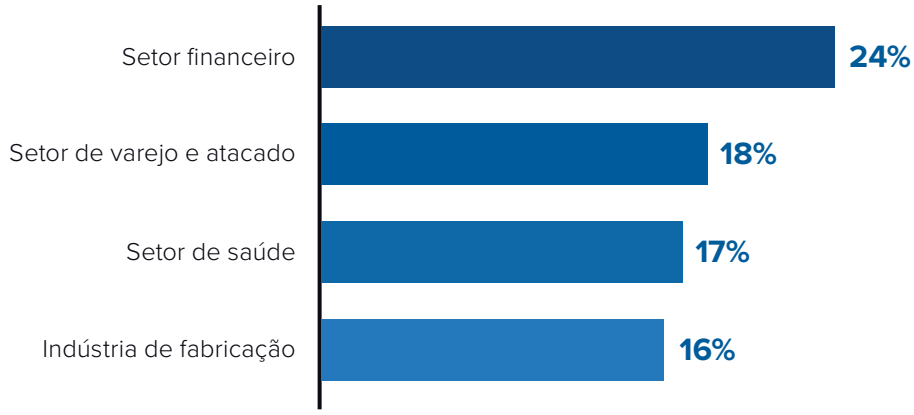


Fonte: IDC, 2021

Como resultado, a porcentagem de todos os sistemas que precisam estar altamente disponíveis está crescendo. Em todos os setores, mais de 60% das empresas têm de 21% a 30% de todos os seus servidores no nível de disponibilidade mais alto. A **Figura 2** mostra a porcentagem de sistemas que precisam estar altamente disponíveis em vários setores.

FIGURA 2

Porcentagem de sistemas que precisam estar altamente disponíveis, por setor



Fonte: IDC, 2021

As plataformas AL4 dominantes de hoje deram grandes passos para se tornarem plataformas totalmente integradas ao data center,que não apenas participam da transformação digital de uma organização, mas também a impulsionam. Esses sistemas processam os dados mais importantes e valiosos de muitas empresas, muitas vezes em volumes maiores do que quaisquer outros tipos de dados, e as empresas precisam desbloquear esses dados e aproveitá-los para se tornarem empresas digitais.

A necessidade de escalabilidade e sustentabilidade

As empresas precisam dimensionar cargas de trabalho que processam quantidades cada vez maiores de dados em ambientes de TI em expansão consistente.

Ao mesmo tempo, elas precisam ser capazes de aumentar ou diminuir rapidamente com base em ondas de demanda às vezes imprevisíveis que podem ocasionalmente assumir a forma de picos severos. Tudo isso significa data centers maiores, mais equipamentos, mais renovação de equipamentos e mais energia para operar o equipamento e, ao mesmo tempo, resfriá-lo.

As cargas de trabalho de IA são a parte de crescimento mais rápido das cargas de trabalho que consomem dados e conduzem os investimentos em computação que as empresas fazem. Atualmente, 21% das organizações dizem que estão investindo em tecnologias de computação que permitem o processamento paralelo necessário para treinamento e inferência em redes de aprendizado profunda de IA, e outros 9% das empresas afirmam que planejam fazer isso em 2021. Além disso, 46% das empresas estão investindo em tecnologias de aceleração de carga de trabalho, como GPUs, FPGAs e ASICs e um plano adicional de 7% para investir em 2021.[5] Este último, especialmente, levou a problemas de data center no que diz respeito aos requisitos de energia e resfriamento. O caso de uso mais comum para aceleração é a inferência de aprendizado profundo de IA (colocando em produção um modelo de IA que foi desenvolvido com uma rede neural profunda [DNN]). Atualmente, 38% das organizações usam aceleração para inferência de IA, enquanto apenas 27% usam aceleração para o treinamento de um DNN.[6] Essa tendência, de que os investimentos em computação de inferência de IA começam a exceder o treinamento de IA, era esperada. Além disso, a IA não é a única carga de trabalho que está gerando investimentos em aceleração usando tais coprocessadores. Análise de dados, HPC, modelagem financeira, cibersegurança e detecção de fraude e negociação financeira são exemplos adicionais de cargas de trabalho que estão cada vez mais sendo executadas em GPUs, FPGAs ou ASICs, e a maioria das empresas executa essas cargas de trabalho no local.

Um grande problema, no entanto, é que a maioria dos data centers não está equipada para sustentar vários racks de nós de computação acelerados em termos de fornecer a potência necessária e dissipar o calor que geram, que é muito maior do que com racks de servidores não acelerados. De acordo com o Departamento de Energia dos EUA (2020), os data centers são um dos tipos de edifícios com maior consumo de energia, consumindo de 10 a 50 vezes a energia por área de um edifício comercial típico. A IDC descobriu que, em média, 17,6% do orçamento operacional do data center é gasto com eletricidade, mais do que qualquer outro item do orçamento. Nos Estados Unidos, os data centers respondem por 2% do uso total de eletricidade no setor comercial.

Ao mesmo tempo, porém, muitas organizações, especialmente na indústria de tecnologia, estão tentando melhorar sua pegada de carbono. As empresas de tecnologia lideram a lista de empresas verdes da Agência de Proteção Ambiental (EPA), e a IDC tem visto grandes investimentos em energia renovável no setor de tecnologia, bem como investimentos em hardware e software mais amigáveis à energia que ajudam a reduzir o consumo de energia. A IDC descobriu que este último ajudou a reduzir o consumo de energia em 26%, em média.

Atualmente, 21% das organizações afirmam que estão investindo em tecnologias de computação que permitem o processamento paralelo necessário para treinamento e inferência em redes de aprendizado profundo de IA.

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

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

Muitas empresas seguiram a sugestão de provedores de serviços de nuvem para uma abordagem mais sustentável de sua TI, ou seja, reutilizando e reciclando seus equipamentos; 33% dos entrevistados em uma pesquisa da IDC [7] disseram acreditar que isso desempenha um papel para alcançar maior sustentabilidade. A reutilização e reciclagem de equipamentos podem, de fato, contribuir significativamente para a pegada de carbono geral de um data center. Pode haver motivos para atualizar certos componentes de um servidor, mas a quantidade de novos componentes obrigatórios entre duas gerações de servidores não excede o número de componentes que poderiam simplesmente permanecer no local e ser reutilizados.

Mais consciência está surgindo em torno desta oportunidade de reutilização para reduzir a área de cobertura ambiental, e a IDC previu que até 2025, 90% das empresas do G2000 exigirão materiais reutilizáveis em cadeias de suprimentos de hardware de TI, metas de neutralidade de carbono para instalações de fornecedores e menor uso de energia como pré-requisitos para fazer negócios.[8] Essas medidas também ajudam a reduzir custos para as empresas, seja pelo menor uso de energia ou investimentos reduzidos em hardware.

A infraestrutura de TI híbrida certa

Gravitando em direção à nuvem híbrida

Hoje, 54% das aplicações das organizações ainda são implementadas no local.[9] A IDC não vê essa porcentagem diminuindo significativamente; as empresas dizem que, em dois anos, ainda esperam executar 52% de suas aplicações no local. Dessas aplicações locais, 56% são executadas como nuvem privada, um número que deve aumentar para 60% em dois anos. Quanto à questão de saber se a nuvem privada cumpre seus objetivos, 61% das organizações dizem que ela não apenas atende, mas excede suas expectativas.

Muitas dessas aplicações, especialmente as aplicações críticas de trabalho, têm interdependências complexas. Em média, as empresas dizem que 49% de suas aplicações de negócios têm algumas dependências e 27% têm interdependências complexas. Hoje, apenas 18% de todas as aplicações são consideradas “nativas da nuvem " o que significa que são microsserviços modulares desagregados que representam conjuntos de serviços implementáveis independentemente. Em contraste, 32% das aplicações continuam monolíticas. No entanto, isso mudará muito rapidamente. As empresas dizem que, em dois anos, apenas 21% das aplicações críticas de negócios serão monolíticas, enquanto 44% serão nativas da nuvem.

Ao mesmo tempo, as empresas esperam aproveitar diferentes implementações de nuvem locais e externas, o que costuma ser chamado de nuvem “híbrida”, que a IDC vê como um cenário de crescimento rápido. A **Figura 3** mostra que a combinação de nuvem mais comum hoje é ter várias nuvens para migrar cargas de trabalho e dados entre elas. Para o cenário de nuvem privada/nuvem pública, cerca de 40% das organizações dizem que essas duas implementações interoperam em suas organizações, em outras palavras, servindo como uma nuvem híbrida mais ou menos integrada.

Observe que, para a parte local de uma nuvem híbrida, a esmagadora maioria das empresas (84%) deseja passar de um modelo de capex para um modelo de opex. Atualmente, 42% dos orçamentos de TI das empresas são financiados com uma abordagem opex; há três anos, esse número era de 36%.

Observe que, para a parte local de uma nuvem híbrida, a esmagadora maioria das empresas (84%) deseja passar de um modelo de capex para um modelo de opex.

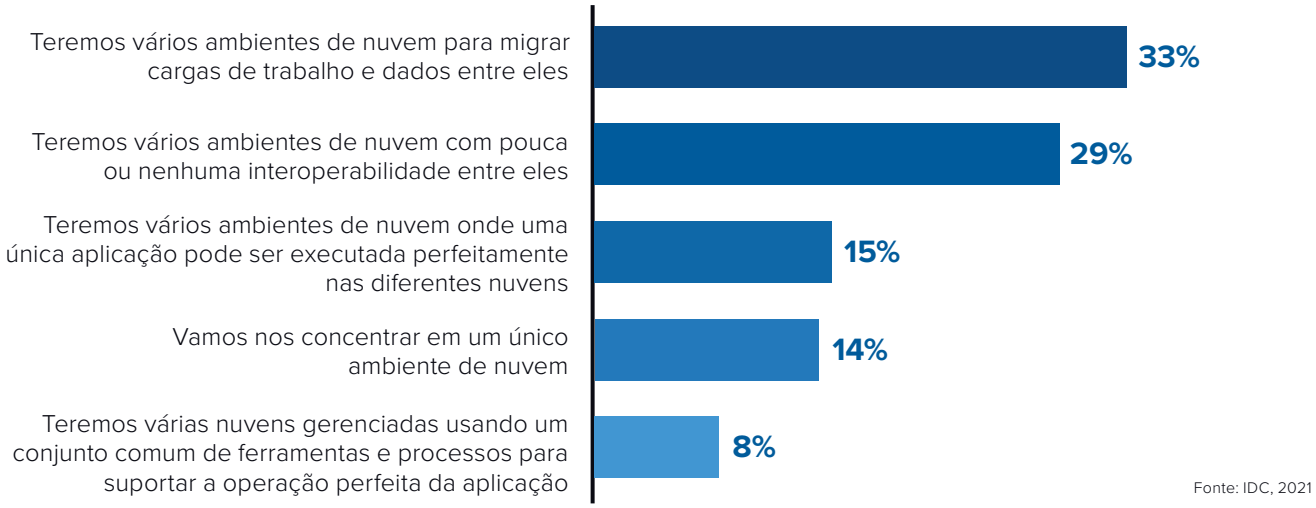
[7] Fonte: IDC 2021 Datacenter Operational Survey

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

[9] Fonte: IDC 1Q21 Cloud Pulse Survey, Maio de 2021

FIGURA 3

Uso de ambientes de nuvem no local e fora do local

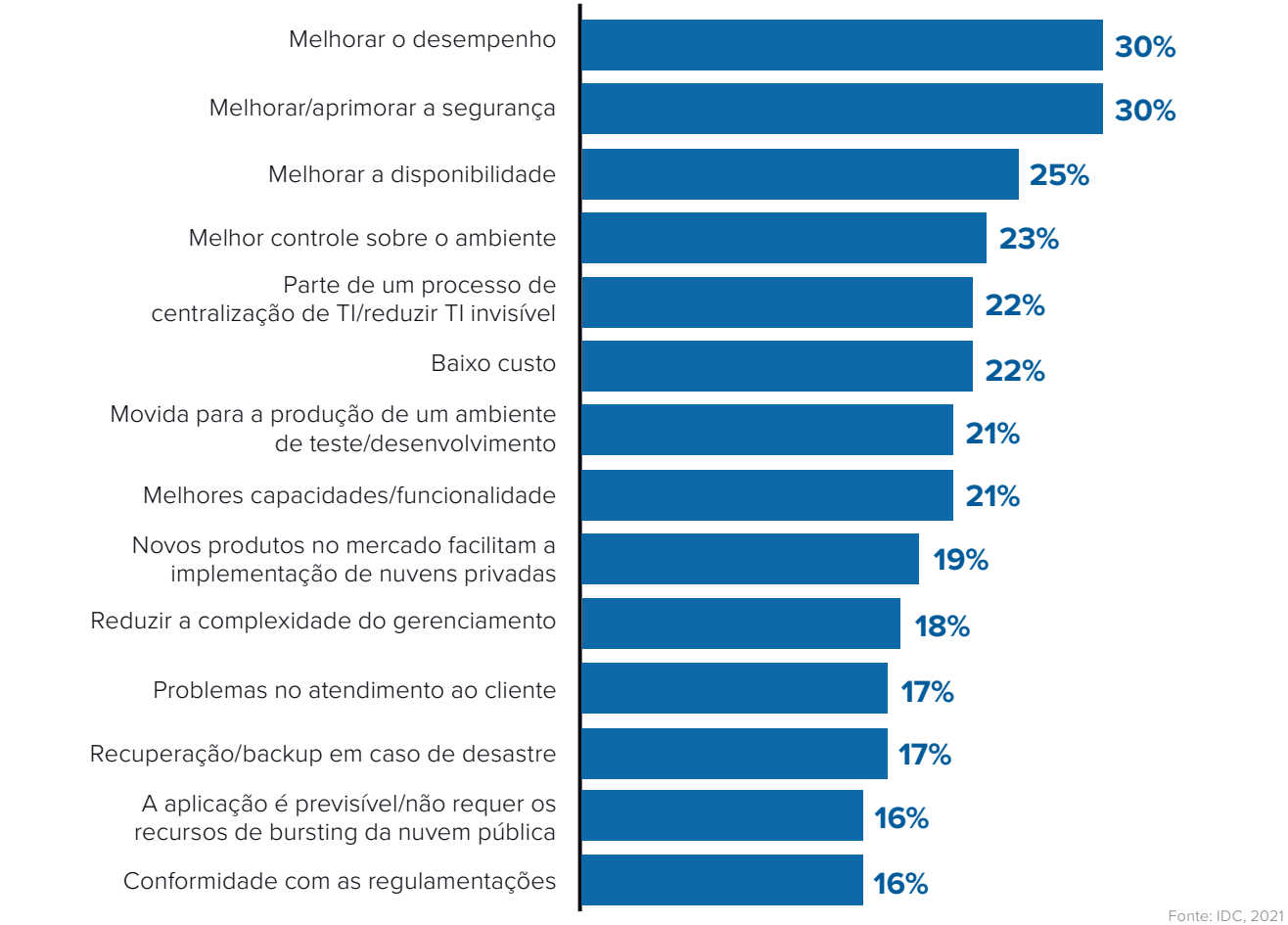


Fonte: IDC, 2021

À medida que a nuvem híbrida se torna mais prevalente, a repatriação de uma nuvem pública para uma nuvem privada também é muito comum: 66% das empresas afirmam que movem aplicações para sua nuvem privada ou ambientes fora da nuvem por vários motivos, com desempenho, segurança e disponibilidade sendo os três principais (ver **Figura 4**).

FIGURA 4

Razões para mover aplicações de IaaS para nuvem privada ou fora da nuvem



Fonte: IDC, 2021

Nuvem híbrida e aplicações nativas da nuvem

Uma nuvem híbrida que foi projetada corretamente é uma plataforma ideal para desenvolver e executar aplicações nativas da nuvem, que cada vez mais empresas consideram um recurso importante para sua transformação digital. A IDC descobriu que a maioria das empresas considera a implementação de vários recursos de "importante" a "extremamente importante" para atender às suas necessidades de negócios, pois investem em uma estratégia de nuvem adequada para desenvolver e executar aplicações nativas da nuvem. Estas capacidades incluem:

- Melhor desempenho, disponibilidade, portabilidade e gerenciamento de aplicações
- Melhor integração de dados, orquestração, observabilidade, gerenciamento de API e AIOps em ambientes de nuvem
- Ciclos de desenvolvimento mais rápidos e tempo de comercialização com CI/CD (desenvolvimento e implementação contínuas) e automação
- Políticas de segurança abrangentes, gerenciamento de riscos, estratégias de recuperação de desastres e conformidade regulatória
- Um modelo opex em vez de capex, incluindo recursos de estorno
- Produtividade otimizada da equipe, eficiência e conjuntos de habilidades

As empresas que desejam aumentar seus investimentos em uma nuvem híbrida precisam "colocar marcas de verificação" ao lado desses itens para garantir que alcancem o ROI que estão antecipando.

A importância da IA e onde executá-la

A IDC espera que o mercado mundial de plataformas de servidores de inteligência artificial cresça para US$ 27 bilhões em 2025.[10]

Esse crescimento será impulsionado pela crescente adoção de tecnologias conversacionais, processamento de linguagem natural (PNL), análise de imagem e vídeo, aprendizado profundo, aprendizado de máquina (ML), geração de hipóteses e análise preditiva. Como resultado, as plataformas de servidores de IA constituirão 21% do mercado mundial de servidores em 2025.

Em uma seção anterior, discutimos a necessidade crescente de coprocessadores para executar cargas de trabalho de treinamento e inferência de IA. Dado que a nuvem privada local é o cenário de implementação preferencial para IA e fora da nuvem local é o segundo mais comum, isso se traduz diretamente em investimentos significativos para as empresas em termos de GPUs, FPGAs e ASICs incluídos. Para o treinamento de IA, esses investimentos são mais ou menos inevitáveis, treinar um algoritmo DNN simplesmente não pode ser feito em um processador host. Para inferência de IA, entretanto, existem muitos modelos de IA que funcionarão muito bem em um processador host avançado ou em um processador host com um processador de IA especializado integrado. Esses cenários têm uma vantagem de custo distinta para as empresas, visto que incluir algumas GPUs em um servidor pode rapidamente dobrar o preço do pacote total.

Isso, é claro, levanta a questão de porque as empresas continuam a executar suas aplicações de IA no local em primeiro lugar. Por que não executá-las na nuvem, por exemplo e evitar o capex por completo? Certamente, parte do treinamento de IA é feito em nuvens públicas nas plataformas de IA dos fornecedores e, uma vez desenvolvidos, esses modelos às vezes permanecem na nuvem como cargas de trabalho de produção.

Visto que a nuvem privada local é o cenário de implementação preferencial para IA e fora da nuvem local é o segundo mais comum, isso se traduz diretamente em investimentos significativos.

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

O fator mais importante que determina a nuvem versus local são os dados, e as seguintes questões são o motivo subjacente para isso:

Quais são os dados necessários para desenvolver o modelo?

Se forem dados de aplicações corporativas centrais, como dados transacionais, é preferível permanecer na plataforma transacional, inclusive por motivos de latência.

Qual é o grau de confidencialidade desses dados?

Se os dados forem confidenciais, o que significa que devem ser protegidos vigorosamente, será indesejável movê-los para a nuvem, seja para treinamento ou inferência.

Qual é o modelo regulamentar em torno dos dados?

Alguns dados não podem ser legalmente movidos para uma nuvem pública, na maioria das vezes, dados corporativos principais. As empresas estão sujeitas a tudo, desde regulamentações locais de proteção de dados ao GDPR e regulamentações do setor, como HIPAA, às regulamentações ISO e à Lei de proteção ao consumidor da Califórnia.

O que pode e não pode ser feito com os dados para manter a conformidade?

Uma vez que os dados começam a ser movidos, torna-se difícil garantir que a organização retenha sua compliance.

Qual é o volume dos dados?

Quanto mais dados são necessários para o treinamento ou quanto mais dados o modelo de IA estiver inferindo, especialmente se essa inferência é em tempo quase real, mais difícil se torna fazer isso na nuvem.

Qual é o grau de integração das aplicações que alavancam os dados?

A plataforma que executa as transações provavelmente terá várias aplicações profundamente integradas ao banco de dados para realizar análises e outras funções, dificultando sua movimentação para a nuvem.

Qual é o custo do armazenamento dos dados?

O armazenamento em grandes volumes na nuvem pode exceder rapidamente quaisquer despesas de capital que seriam necessárias para o armazenamento local.

Em conjunto, essas considerações levam muitas organizações a continuarem com instalações locais para seu treinamento de IA e para inferir cargas de trabalho. Eles ainda podem treinar em um ambiente de computação separado no data center por trás de seu firewall, mas, depois disso, mover o modelo treinado de volta para a plataforma que executa as principais aplicações corporativas para inferência. Se a plataforma permitir inferência robusta, isso permitirá que as empresas usem IA em dados essenciais que estavam fora dos limites no passado.

IBM Power10 e IBM Power E1080

Para se transformar com sucesso em direção à corporatividade digital, as empresas precisam de plataformas de computação que possam absorver qualquer tipo de volatilidade do mercado, que sejam seguras sem concessões, que possam ser dimensionadas sem esforço, ao mesmo tempo que reduzem as pegadas físicas e de carbono das empresas, que forneçam os mais altos níveis de resiliência e possam executar IA em tempo real em um grande número de transações, tudo como parte de uma nuvem híbrida perfeita. O novo processador Power10 da IBM e a plataforma de classe corporativa IBM Power E1080 baseada no Power10 oferecem uma grande variedade de inovações que atendem a esses requisitos de maneiras novas e interessantes.

O novo processador Power10

A nova arquitetura e processador IBM Power10 apresenta novas tecnologias importantes que ajudarão as empresas com cargas de trabalho que exigem computação, memória e largura de banda, incluindo novas tecnologias para inferência de IA rápida no chip sem hardware adicional, com base em um acelerador de matriz matemática (MMA), intencionalmente integrado.

De uma perspectiva de segurança, o Power10 implementa criptografia de memória sem degradação de desempenho (em oposição à criptografia de memória baseada em software), fornece segurança de contêiner otimizada de hardware/software para isolamento de contêiner e inclui recursos de segurança para impedir a capacidade iminente da computação quântica de quebrar as chaves de criptografia tradicionais.

A escalabilidade com o Power10 é levada a novos níveis com várias inovações de largura da banda. A IBM aprimorou a tecnologia de conectividade POWER AXON e incluiu o Open Memory Interface (OMI), ambas executando a 32 GT/s. A interface Power10 AXON conecta até 16 soquetes em um sistema grande e escalável. A OMI se comunica com a memória DDR4 DRAM por meio de 16 portas DDR por soquete, fornecendo largura de banda de até 409 GB/s por soquete. Essas duas interfaces podem ser usadas para fornecer soluções de computação muito flexíveis e até compossíveis.

Este é o primeiro processador de sete nanômetros da IBM, e a IBM afirma um ganho de eficiência de três vezes mais em comparação com o IBM Power9 em termos de poder de computação (número de usuários, número de transações) e energia.[11] Com o foco contínuo da IBM na nuvem híbrida, isso se traduz diretamente em uma área de cobertura menor no data center e redução significativa de energia. Existem 15 núcleos de processador no chip, e o Power10 contará com PCI Gen5, que está começando a surgir na indústria.

O IBM Power E1080

O IBM Power E1080 é a primeira plataforma de classe corporativa da IBM construída com o processador Power10. O sistema pode ser dimensionado para até 16 processadores e é focado distintamente nas principais considerações de TI para organizações que devem atender às demandas da empresa digital.

Segurança

Para tornar a segurança persistente e livre de penalidades, a IBM incorporou criptografia ao processador Power10. Isso permite que os dados sejam criptografados sem comprometer o desempenho do sistema. Além disso, o sistema foi equipado com recursos de segurança adicionais para proteger contra ataques de programação orientados para

[11] O desempenho três vezes maior é baseado na análise de engenharia pré-silício de ambientes Integer, Enterprise e Floating Point em um servidor de soquete duplo POWER10 com módulos 2x30-core vs oferta de servidor de soquete duplo POWER9 com módulos 2x12-core; ambos os módulos têm o mesmo nível de energia.

retorno, uma técnica na qual um invasor pode executar código malicioso na presença de defesas de segurança. O Power E1080 oferece proteção avançada de dados com criptografia de memória transparente, o tipo de segurança em nível de hardware para dados em uso em que se baseia a computação confidencial, e apresenta quatro vezes mais aceleradores de criptografia criptográficos do que seu predecessor. As partições na plataforma melhoraram o isolamento e o sistema está protegido contra ameaças futuras baseadas em quantum com criptografia pós-quântica (PQC), bem como criptografia totalmente homomórfica (FHE), uma tecnologia em que as entradas no sistema não precisam ser decriptografadas, o que significa que podem ser executadas por uma parte não confiável sem revelar essas entradas.

Resiliência

A IDC considera a família de servidores Power de classe corporativa como tendo AL4, em outras palavras, totalmente tolerante a falhas e, portanto, fornecendo 99,999% ou mais de disponibilidade. Com o Power10, o IBM Power E1080 vai um passo além do seu predecessor ao fornecer largura de banda muito alta e confiabilidade, disponibilidade e capacidade de manutenção (RAS) com a nova interface de memória aberta. O processador pode detectar, isolar e se recuperar automaticamente de erros de software sem interrupção ou sem depender do sistema operacional para gerenciar falhas e erros recuperáveis de autocorreção. O sistema também apresenta recursos aprimorados de reparo simultâneo, como cabos subminiatura push-on (SMP) entre os nós para reduzir o tempo de inatividade da aplicação.

Escalabilidade e sustentabilidade

Em termos de escalabilidade e sustentabilidade, o IBM Power E1080 se beneficia enormemente do fato de que a família de servidores Power é excepcionalmente bem integrada, do processador ao firmware, do sistema operacional ao hardware, já que todos são componentes IBM. O software e a eficiência do OpenShift contêiner da plataforma é excepcional, de acordo com a IBM. Como resultado, a plataforma com o novo processador Power10 atinge 50% mais desempenho no mesmo espaço e consumo de energia em comparação com o Power E980.[12] Isso também se traduz em 33% menos consumo de energia para a mesma carga de trabalho, afirma a IBM.[13] A maior eficiência ajuda as empresas a reduzir significativamente sua pegada de carbono e potencialmente consolidar cargas de trabalho, economizando em custos de hardware e software.

Cloud Híbrida

O Power E1080 oferece suporte a três ambientes operacionais, AIX, IBM i e Linux, na mesma plataforma e é projetado para oferecer suporte à adoção de nuvem híbrida pelas empresas para todos os três ambientes operacionais. O AIX é, obviamente, o sistema operacional Unix totalmente modernizado da IBM que continua a ser a plataforma preferida para a plataforma Power de escalabilidade vertical de classe corporativa. IBM i é o ambiente operacional da IBM que integra o banco de dados e outros softwares corporativos ao sistema operacional, simplificando muito o gerenciamento da plataforma; para muitas empresas de médio porte, o IBM i é o coração de suas operações. O AIX e o IBM i são extremamente amigáveis ao software livre, suportam linguagens de desenvolvedor modernas e preferenciais e são totalmente operados como uma nuvem híbrida. Assim como nas gerações anteriores, o Power E1080 também pode ser executado total ou parcialmente no Linux com os mesmos recursos de segurança, disponibilidade e escalabilidade, representando uma oportunidade para as empresas moverem suas cargas de trabalho transacionais e analíticas para uma plataforma totalmente de software livre.

Os seguintes componentes de software IBM Power desempenham um papel importante para permitir que as empresas aproveitem sua plataforma Power de classificação corporativa com AIX, IBM i e Linux para uma modernização de carga de trabalho segura, altamente disponível e baseada em nuvem:

IBM PowerVM

As cargas de trabalho do servidor IBM Power são virtualizadas, móveis e totalmente habilitadas para nuvem com o PowerVM, que foi recentemente aprimorado com vários novos recursos, incluindo compactação e criptografia de dados Live Partition Mobility (LPM), o que significa que quando uma partição ativa é migrada de um Power servidor para outro, o que ocorre com tempo de inatividade zero, os dados são criptografados e compactados automaticamente, um recurso importante de segurança e desempenho.

IBM PowerVC

O PowerVC é a ferramenta de gerenciamento de virtualização construída no OpenStack, simplificando o gerenciamento de recursos virtuais em ambientes Power. O software foi recentemente aprimorado com vários novos recursos, incluindo um recurso de exportação/importação para compartilhar imagens de VM entre data centers.

[12] Informações fornecidas pela IBM. Com base nos resultados de rPerf publicados para o núcleo Power E980/12 em comparação com as medições internas de rPerf da IBM (usando a mesma metodologia) para o núcleo Power E1080/15.

[13] Power9 (12c) é 5081 rPerf @ 16.520 Watts (0,31 rPerf/Watt), Power10 (15c) é 7998 rPerf @ 17.320 Watts (0,46 rPerf/Watt) 0,46/0,31 = 1,48 Mais rPerf/Watt

IBM PowerSC

PowerSC é o portfólio de segurança da plataforma, simplificando o gerenciamento de segurança e conformidade, apresentando automação de conformidade, detecção de intrusão de malware, gerenciamento de correção e muito mais. Ele foi aprimorado com vários recursos ou até mesmo novas ofertas, incluindo a ativação de autenticação com múltiplos fatores (MFA), outro recurso de segurança importante. Em geral, a segurança no IBM Power with AIX é obtida com uma solução abrangente que inclui o processador, firmware, hipervisor e os incontáveis recursos de segurança do próprio sistema operacional para proteger os dados em todos os níveis.

IBM PowerHA e VM Recovery Manager HA e DR

PowerHA é uma tecnologia de alta disponibilidade que ajuda a fornecer disponibilidade quase contínua de aplicações e melhora a confiabilidade do serviço. É um contribuidor chave para o IBM Enterprise Power ser caracterizado como tolerante a falhas (AL4) pela IDC e foi aprimorado com vários recursos, como métricas de failover aprimoradas e verificação de cluster cruzado (por exemplo, para comparar um desenvolvimento com um cluster de teste). VM Recovery Manager (VMRM) é uma solução HA/DR simplificada com base na replicação e reinicialização da VM que independe do sistema operacional e inclui agentes de monitoramento de aplicações, como DB2, Oracle e SAP HANA.

Cloud Management Console

O Cloud Management Console (CMC) fornece uma visão completa do desempenho, inventário e registro de infraestrutura de energia local e externa. A configuração de gerenciamento de comunicação é hospedada na nuvem IBM, liberando assim as empresas de ter que manter software para monitorar sua infraestrutura e ajudando a simplificar o gerenciamento de implementações de nuvem híbrida e a simplificar o monitoramento e gerenciamento de sua infraestrutura.

Enterprise Cloud Edition 2.0

A Enterprise Cloud Edition reúne todos os componentes principais de uma infraestrutura de gerenciamento de nuvem simplificada em cima do PowerVM, incluindo PowerSC, MFA, PowerVC, CMC, VMRM e Aspera. Ela permite rápida implementação e gerenciamento de uma nuvem privada; gerenciamento simplificado de segurança e conformidade; alta disponibilidade simplificada e transferências aceleradas de arquivos grandes entre nuvens. A Enterprise Cloud 2.0 pode ser adquirida com o AIX 7.2 integrado.

Plataforma de automação Red Hat Ansible

A Plataforma de automação Red Hat Ansible permite a automação escalonável e segura de vários aspectos das operações de TI da empresa, incluindo provisionamento de recursos, gerenciamento do ciclo de vida de aplicações e operações de rede. Ela consiste em Ansible Engine, Ansible Tower e Ansible Hosted Services. Todos os outros produtos do portfólio da Red Hat podem ser integrados usando a Plataforma de automação Red Hat Ansible, que permite consistência no data center, fornecendo métodos programáticos para implementar, gerenciar e proteger os recursos de infraestrutura.

Red Hat OpenShift

O Red Hat OpenShift é uma plataforma certificada Kubernetes de classe corporativa (uma orquestração de contêineres) para construir, implementar e gerenciar aplicações em contêineres. O Red Hat OpenShift pode ser consumido como um serviço totalmente gerenciado em diferentes provedores de nuvem ou gerenciado pelo cliente usando o Red Hat OpenShift Container Platform ou Red Hat OpenShift Kubernetes Engine. Ele pode ser implementado localmente em servidores bare metal, plataformas de virtualização (Red Hat Virtualization, VMware ou Red Hat OpenStack) ou grandes provedores de nuvem, como IBM Cloud, AWS, Google ou Azure. Além disso, o Red Hat Advanced Cluster Management para Kubernetes pode ser usado para gerenciar vários clusters e aplicações Red Hat OpenShift a partir de um único console, com políticas de segurança integradas, permitindo aos clientes uma nuvem híbrida aberta. O Red Hat OpenShift é compatível com IBM Power, IBM Z e plataformas baseadas em x86 e pode ser usado com AIX, IBM i e Linux.

IBM Cloud Paks

Os IBM Cloud Paks são produtos de software cada vez mais populares pré-embalados em contêineres e altamente integrados em vários serviços OpenShift para implementação rápida e fácil no OpenShift. Os IBM Cloud Paks oferecem ferramentas de desenvolvedor, dados e serviços de inteligência artificial e software de middleware de software

livre. Eles são executados na plataforma de nuvem Red Hat OpenShift. Alguns Cloud Paks que são particularmente relevantes para o IBM Power são:

- ***Cloud Pak for Data:*** ajuda os clientes a expandir insights de dados e recursos de IA
- ***Cloud Pak for Integration:*** consiste em ferramentas de integração para dados, serviços de aplicações e serviços em nuvem para ajudar a integrar aplicações, dados, serviços em nuvem e APIs
- ***Cloud Pak for Watson AIOps:*** oferece visibilidade, governança e automação multinuvem, dado o uso comum de implementações desse tipo

Inteligência Artificial

A IBM afirma que o Power E1080 acelera o desempenho de inferência de IA em uma ordem de magnitude em comparação com seu antecessor. Isso não requer nenhum hardware especializado, como um coprocessador (GPU, FPGA ou ASIC). Em vez disso, a inferência ocorre em um acelerador de matriz matemática (MMA). Cada núcleo do chip Power10 tem um MMA integrado para realizar operações de matrizes matemáticas de maneira eficiente. Essas operações foram otimizadas em uma ampla variedade de tipos de dados para várias precisões, que são importantes para o aprendizado profundo, de precisão dupla e precisão simples a dois tipos de meia precisão, incluindo Bfloat-16, bem como Int-16, Int-8 e Int-4. O desempenho de inferência de IA foi infundido em todas as camadas do processador. O cache L2 foi quadruplicado: as unidades de armazenamento de carga e o SIMD dobraram. Isso significa que uma carga de trabalho transacional que possui componentes de IA incorporados pode executar as transações e a inferência de IA no mesmo processador Power10 sem exigir um coprocessador.

A inferência no chip também significa que todos os recursos de segurança do processador e do sistema estão disponíveis para proteger os dados que estão sendo inferidos. Além disso, a plataforma é compatível com Open Neural Network Exchange– (ONNX-). ONNX é um ecossistema de IA de software livre de empresas de tecnologia e organizações de pesquisa que trabalham para estabelecer padrões abertos para representar algoritmos e ferramentas de IA a fim de promover inovação e colaboração no setor de IA. As empresas com IBM Power E1080 podem trazer modelos ONNX para a plataforma inalterados e executá-los, aproveitando os recursos RAS da plataforma durante a inferência.

Desafios/oportunidades

Para empresas

As plataformas de classe corporativa que executam as principais cargas de trabalho transacionais e analíticas de uma organização tendem a ser tratadas como silos no data center, mesmo que sejam projetadas e construídas com recursos e tecnologias abrangentes para evitar isso. Essas plataformas costumam ser "protegidas" de novas tecnologias pela equipe de TI que tem profundo conhecimento do sistema, mas que tem medo de expor os dados, integrar a plataforma com a nuvem, executar software livre na plataforma e executar modelos de IA em dados em tempo real. Para as empresas, o desafio é romper com essa cultura de hesitação o mais rápido possível. É absolutamente crítico permitir que as plataformas de classe corporativa sejam apreciadas como os sistemas abertos que são. Isso permitirá que as empresas as aproveitem totalmente como plataformas de transformação digital que geram novas oportunidades de receita. Ao mesmo tempo, essas plataformas oferecem uma oportunidade para começar a abordar seriamente as questões de sustentabilidade, reduzindo a pegada de carbono da organização. Executar a IA em uma plataforma corporativa sem a necessidade de coprocessadores caros e que consomem energia é um requisito importante, pois cada vez mais aplicações centrais estão sendo infundidas com a funcionalidade de IA.

Para a IBM

Com a nova plataforma Power E1080, a IBM continua a conduzir os negócios em direção à abertura, nuvem híbrida, IA e sustentabilidade em uma plataforma altamente segura, confiável e de alto desempenho. A IBM tende a enfrentar os desafios da inovação com novas tecnologias interessantes que, em alguns casos, são inovadoras e estão à frente do pacote: por exemplo, o MMA no novo processador Power10. A inovação não é o maior desafio da IBM. O verdadeiro desafio para a IBM é mudar uma parte da mentalidade de seus clientes de tratar sua plataforma corporativa como um sistema em silos, ou talvez um sistema cuidadosamente aberto, para uma plataforma agressivamente integrada com o restante do data center e com a nuvem, aproveitando totalmente todos os seus recursos para fazer coisas novas e gerar mais receita com os dados principais que residem na plataforma. A IBM precisa continuar incentivando seus clientes a serem ousados e criativos com sua plataforma corporativa por meio de educação, incentivos e estudos de ROI.

Conclusão

As empresas modernas precisam de plataformas de computação que possam lidar com a extrema volatilidade do mercado, fornecer segurança inflexível, ser dimensionadas sem esforço e de forma sustentável, fornecer resiliência máxima, executar IA em tempo real e operar como uma nuvem híbrida. O novo processador Power10 da IBM e a plataforma de classe corporativa IBM Power E1080 baseada no Power10 atendem a esses requisitos de frente. O processador Power de nova geração da IBM é tudo menos um passo incremental e se aventura em um importante território voltado para o futuro.

O processador ativa tecnologia de computação confidencial para criptografia baseada em hardware que protege os dados em uso. A largura de banda no Power10 foi grandemente aumentada para ativar uma poderosa escalabilidade de 16 soquetes. A resiliência é ainda mais aprimorada com a capacidade de detectar, isolar e se recuperar automaticamente de erros de software sem uma interrupção ou sem depender do sistema operacional. O MMA no chip possibilita a inferência de IA em tempo real sem a necessidade de um coprocessador. E entre as soluções Red Hat e o software IBM Cloud, a capacidade de operar totalmente como uma nuvem híbrida é um dado adquirido. Com o chip Power10 como mecanismo da nova plataforma Power E1080, a IBM continua a impulsionar a computação corporativa em direção a um ponto ideal onde o melhor de todos os mundos se reúne: abertura, poder de computação, nuvem híbrida, IA, segurança, escalabilidade, sustentabilidade e confiabilidade em uma única plataforma.

Sobre o analista

Peter Rutten

Diretor de pesquisa, Grupo de sistemas de infraestrutura, plataformas e tecnologias,
Líder de Pesquisa global de soluções de computação intensiva de desempenho, IDC

Peter Rutten é Diretor de Pesquisa da prática de infraestrutura mundial da IDC, cobrindo pesquisas em plataformas de computação. O Sr. Rutten é o líder de pesquisa global da IDC em soluções e casos de uso de computação de alto desempenho. Isso inclui pesquisas em Inteligência Artificial (AI), Modelagem e Simulação (M&S) e Infraestrutura de Big Data e Analytics (BDA) e pilhas de soluções associadas. Sua cobertura de computação intensiva de desempenho inclui sistemas, plataformas e tecnologias de supercomputação, high-end, acelerada, em memória e de infraestrutura de computação heterogênea. Inclui plataformas de computação com GPUs, FPGAs, ASICs e outros aceleradores que são implementados na nuvem e também no local. Também inclui pesquisas sobre plataformas x86 de missão crítica, mainframes e sistemas baseados em RISC, bem como seus ambientes operacionais (Linux, z/OS, Unix). O Sr. Rutten também examina tecnologias e plataformas emergentes, como computação quântica, computação neuromórfica e tecnologias que são potencialmente prejudiciais para mercados de infraestrutura maduros. Como parte de sua função, o Sr. Rutten realiza análises quantitativas (dimensionamento de mercado e previsão) e qualitativas (com base em pesquisas primárias), bem como dimensionamento de mercado personalizado para os clientes da IDC.

Mais sobre Peter Rutten

IDC Custom Solutions

Esta publicação foi produzida por Soluções Customizadas da IDC. Como o principal fornecedor global de inteligência de mercado, serviços de consultoria e eventos para os mercados de tecnologia da informação, telecomunicações e tecnologia de consumo, o grupo de Soluções Personalizadas da IDC ajuda os clientes a planejar, comercializar, vender e ter sucesso no mercado global. Criamos inteligência de mercado acionável e programas de marketing de conteúdo influentes que geram resultados mensuráveis.

IDC @idc @idc idc.com

© 2021 IDC Research, Inc. Os materiais da IDC são licenciados para uso externo, e de forma alguma o uso ou publicação de pesquisas da IDC indica o endosso pela IDC dos produtos ou estratégias do patrocinador ou licenciado.

Política de privacidade | CCPA