

White paper

Cargas de trabalho de IA e o futuro da infraestrutura de TI

Encontrando o caminho certo para navegar em tempos incertos





Índice

Os dados sendo um novo propulsor econômico	3
A infraestrutura digital crítica que viabiliza a AI	4
A eficiência é primordial. Mas não será uma bala de prata	5
Entendendo o futuro	6
A chave para se adaptar? Alavancar a tecnologia e construir boas parcerias	11
Como será o futuro da infraestrutura de TI? É isto que os principais especialistas estão dizendo	12
Para onde vamos a partir de agora?	16

Os dados sendo um novo propulsor econômico

Comenta-se frequentemente que “os dados são o novo petróleo”. Mas assim como o petróleo, os dados precisam ser refinados.

Como o petróleo, os dados têm o potencial de transformar a economia global. Entretanto, da mesma forma como foram necessários avanços na tecnologia de refino para transformar o petróleo bruto no motor do crescimento econômico do século XX, a inteligência artificial (IA) é provavelmente a chave para fazer dos dados a base da economia do século XXI.

De acordo com a consultoria McKinsey & Company, a IA generativa sozinha poderia adicionar o equivalente a 4,4 trilhões de dólares por ano na economia global, incluindo até 100 bilhões de dólares no setor de telecomunicações, US\$ 130 bilhões no setor de mídia e US\$ 460 bilhões no setor de alta tecnologia¹

E o poder da IA não será medido apenas em dólares, euros, yuan ou yen. Ele também será medido pelos avanços que afetarão a qualidade da vida humana. Já estamos no limiar de

importantes conquistas, como a conversão de uma grande quantidade de dados médicos em cura para doenças como o câncer e até mesmo ajudando a desenvolver soluções inovadoras para combater as mudanças climáticas globais.

>US\$25 trilhões

É o potencial de aumento do PIB global como resultado da IA até 2030²

Entretanto, transformar a promessa da IA em realidade necessita de todo um ecossistema que dê suporte à tecnologia. A arquitetura do data center e a infraestrutura digital crítica que lhe dá suporte precisarão passar por uma importante transformação para atender às demandas das cargas de trabalho da IA. E esse ecossistema precisará ser projetado para fazer a diferença por muito tempo, e não para a obsolescência.



¹ McKinsey Digital; “The economic potential of generative AI: The next productivity frontier,” publicado em 14 de junho, 2023; <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>; acessado em 30 de maio, 2024

² ibid.

A infraestrutura digital crítica que viabiliza a IA

As tendências atuais sugerem que 19% do uso de energia nos data centers será associado à IA até 2028³. Então, o que isto significa para os data centers, especialmente no que se refere à arquitetura, ao uso de energia e ao resfriamento?

Embora possa não ser manchete de primeira página, o crescimento meteórico da IA já está desorganizando a arquitetura de TI e a infraestrutura digital crítica do data center. Mas os efeitos até agora são apenas a ponta do iceberg. Os aportes associados com a IA estão criando possíveis gargalos no tamanho dos chips, no peso dos racks e na alimentação e resfriamento do data center. O dispêndio com infraestrutura digital crítica para a IA generativa ultrapassará US\$ 18 bilhões em 2024, aumentando para mais de US\$ 48 bilhões até 2027.⁴



Quando primeiro surgiu a metáfora dos dados serem o novo petróleo, ela não era necessariamente precisa. Agora temos na IA uma ferramenta incrível, que pode prospectar dados e desvendar seu valor oculto. Mas isso também significa que você precisa estar preparado para as profundas mudanças na infraestrutura necessária para dar suporte à IA."



Stephen Liang

Diretor Executivo de Tecnologia (CTO) e
Vice-Presidente Executivo, Vertiv

O nível de investimento indica o quão profundamente o crescimento da IA generativa afetará a infraestrutura digital crítica. A revolução na IA levará, necessariamente, a uma revolução na forma como os data centers operam. Então, embora a maioria das pessoas esteja focada nos benefícios econômicos da IA, obter estes benefícios exigirá prever, compreender e resolver os desafios emergentes da infraestrutura.

Para aproveitar ao máximo o maior avanço na economia global desde o surgimento da era do petróleo, quem estiver confiando na IA para impulsionar seu crescimento precisará trabalhar com os parceiros certos e formular um plano para como lidar com esses desafios emergentes da infraestrutura.



³ Goldman Sachs; "AI is poised to drive 160% increase in data center power demand;" 14 de maio, 2024. <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>, acessado em 5 de agosto, 2024

⁴ Villars R. et al. "Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending;" IDC #US51539723, dezembro, 2023

A eficiência é primordial. Mas não será uma bala de prata.

A IA é baseada em uma série de coisas, mas um dos seus principais elementos são os microprocessadores. Os chips usados para treinar modelos de IA necessitam de uma quantidade significativa de energia e geram uma quantidade correspondente de calor.

As GPUs são a atual escolha de chip para executar as cargas de trabalho da IA e, usar GPUs ao invés de CPUs para executar cargas de trabalho de computação paralela é 100 vezes mais eficiente pela ótica do consumo de energia. Entretanto, embora as GPUs sejam mais eficientes, as necessidades de computação continuarão a crescer exponencialmente, superando – em muito – qualquer ganho de eficiência no chip.⁵ Então, embora as GPUs produzam, sem dúvida, muito mais computação com a mesma

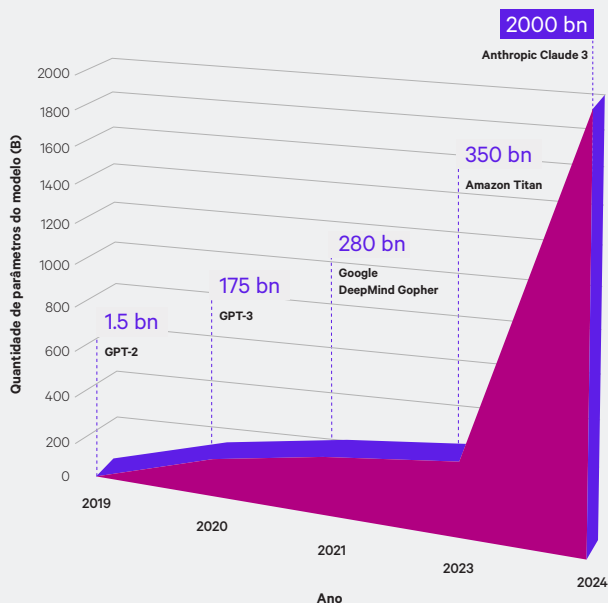
quantidade de energia usada, as demandas da IA significam que a quantidade de energia total usada ainda aumentará.⁶

Provavelmente, muito mais recursos financeiros serão dedicados ao desenvolvimento de GPUs mais eficientes. Com essa quantidade de investimento, seria imprudente não esperar progressos no que tange a eficiência.

Entretanto, se a tendência atual continuar, os modelos que tornam a IA generativa melhor também continuarão a se tornar maiores. A concorrência entre os principais players da indústria, como a OpenAI, Anthropic, Google e a Meta está sendo impulsionada pelo uso de modelos ainda maiores, resultando que os modelos têm duplicado sua necessidade de computação a cada seis meses.⁷

Além do crescimento exponencial⁸

O passado irá prever o futuro no crescimento dos LLMs?



Os resultados inacreditáveis associados com a IA certamente justificam o aumento cada vez maior da entrada de energia, mas permanece sendo um fato que avanços nos recursos da IA parecem também necessitar de modelos ainda maiores.

A demanda por precisão está, em parte, por trás do crescimento contínuo destes modelos, mas o crescimento vai além da precisão simples. Novas funcionalidades provavelmente demandarão modelos ainda maiores para realizar tarefas como a realização de trabalho em diversas etapas ao invés de apenas responder a prompts ou determinar qual resposta algoritmicamente disponível melhor atende às necessidades do usuário.

Embora seja verdadeiro que modelos menores estejam mostrando resultados promissores, a demanda dos usuários por maior precisão e funcionalidade significa que esses modelos menores provavelmente continuarão a ser um nicho.

Ao mesmo tempo, a base de usuários de IA está crescendo rapidamente. Consumidores individuais, empresas de todos os tipos e até mesmo organizações governamentais estão impulsionando uma demanda crescente por IA.

⁵ Newmark; "2023 U.S. Data Center Market Overview & Market Clusters," janeiro, 2024; <https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview-market-clusters>, acessado em 10 de julho, 2024

⁶ ibid.

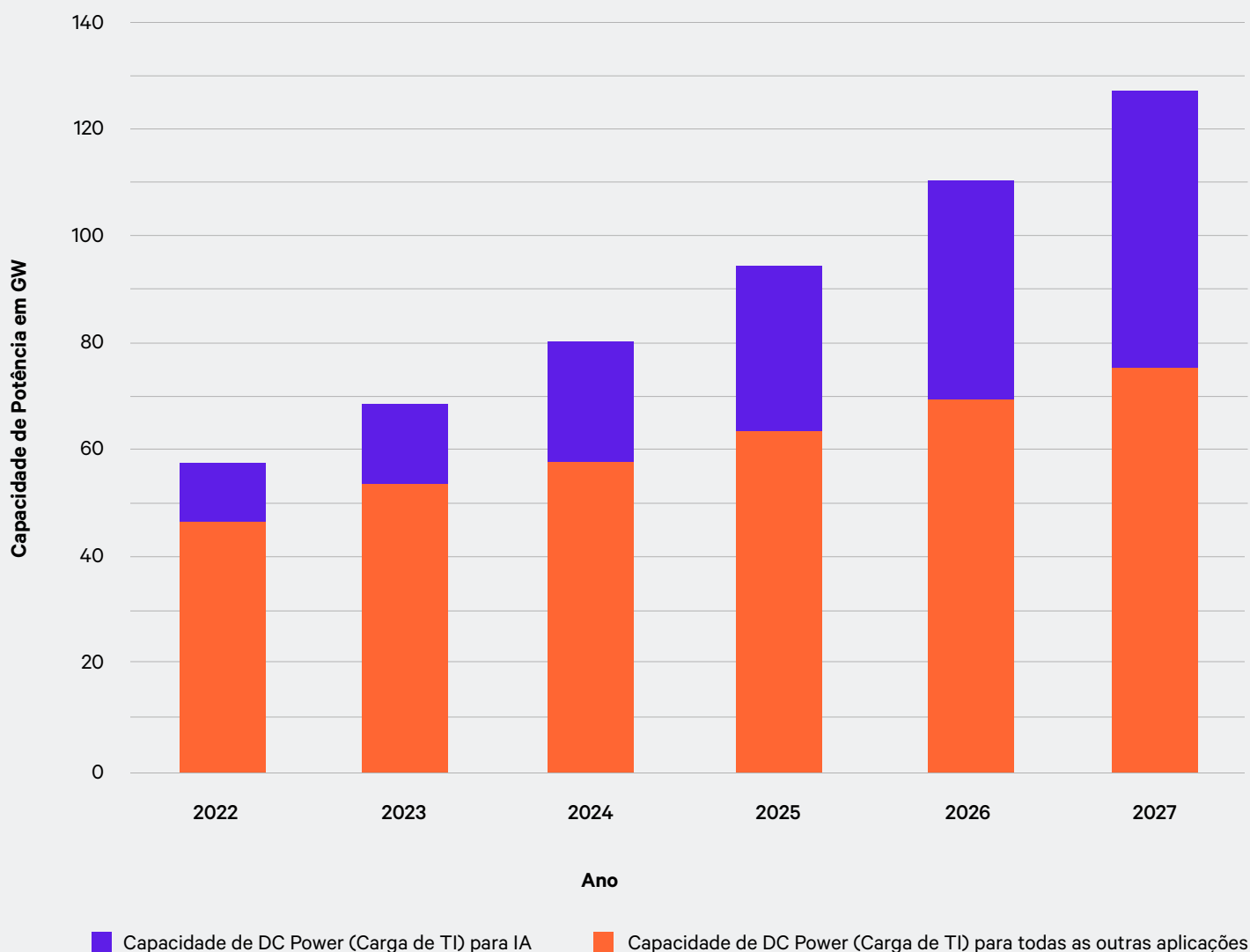
⁷ Epoch AI; "Notable AI Models," 19 de junho, 2024; <https://epochai.org/data/notable-ai-models>, acessado 10 de julho, 2024

⁸ Adaptado de The Rise and Rise of A.I. Large Language Models (LLMs) & seus bots associados como ChatGPT, da McCandless, D. Evans T. e Barton P, 2024 (<https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/>). Em domínio público.

Entendendo o Futuro

De acordo com o paradoxo de Jevons, um aumento na eficiência de um recurso irá gerar um aumento no consumo ao invés de uma redução. Dados os diversos fatores que parecem estar impulsionando tanto a demanda por IA quanto o uso de modelos maiores, é difícil ver como os ganhos de eficiência nos chips ou nos modelos poderá compensar completamente as iniciativas que demandam uma grande quantidade de computação no espaço da IA. Em outras palavras, o futuro pode ser mais eficiente, mas a capacidade das cargas de TI continuará a aumentar nesta década apesar dos ganhos de eficiência.

A IA é o motor de propulsão do crescimento da capacidade das cargas de TI⁹

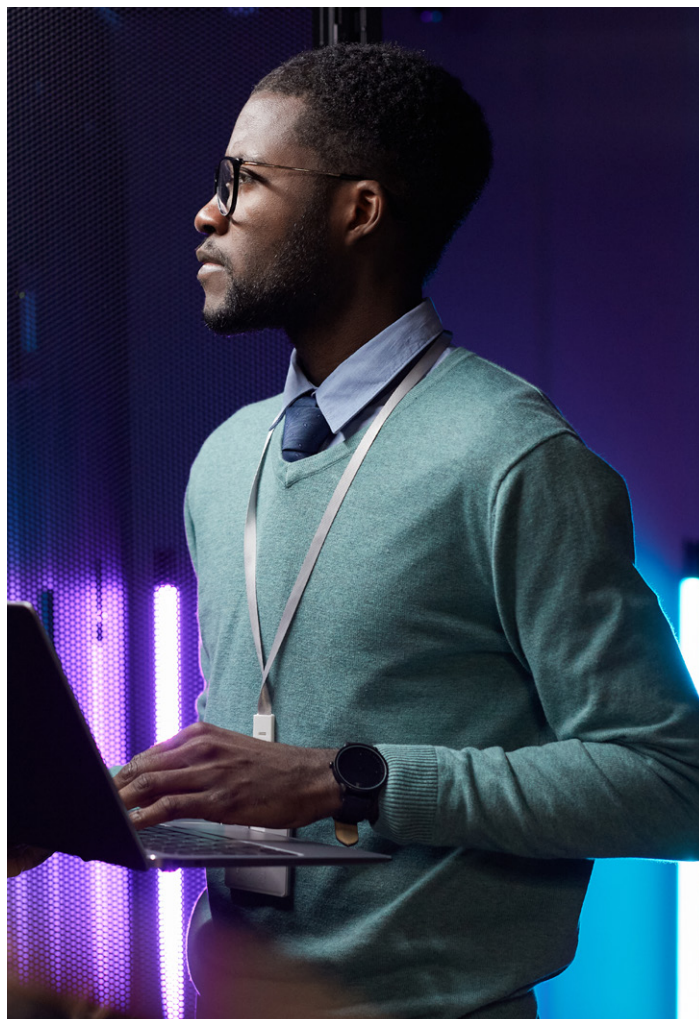


⁹ Galabov, V and Sukumaran M; "Cloud and Data Center Market Snapshot – Novembro de 2023;" Omdia; Novembro, 2023

O jogador de baseball norte-americano e filósofo amador Yogi Berra disse alguma vez, “É difícil fazer previsões, especialmente sobre o futuro.” Embora isso possa parecer especialmente adequado com tantas incógnitas conhecidas no desenvolvimento da infraestrutura para IA, com base nas tendências atuais e nos prováveis avanços tecnológicos, é possível antecipar os maiores desafios para a implementação e o uso da IA, especialmente no contexto dos data centers. Então, quais são os elementos mais importantes a serem buscados se você estiver interessando em tornar as suas operações preparadas para o futuro?

Especificamente, as tendências abaixo devem estar no radar de todos:

- **Mudanças na capacidade de saída:** Precisamos estar preparados para um mundo onde a potência de saída aumente. As cargas dos racks, PDUs, UPSs, painéis elétricos e CDUs terão mais capacidade e terão um efeito dominó correspondente em toda a infraestrutura.
- **Mudanças na escala total:** Sem dúvida, o efeito dominó mais óbvio será as instalações também tornarem-se maiores. Os blocos de 3 MW de hoje serão os blocos de 20 MW de amanhã.
- **A busca pela energia sempre ligada:** Maior significa também mais energia e mais calor. E mais calor significa que o resfriamento líquido e o híbrido são o futuro. Como as GPUs do futuro precisam ser continuamente resfriadas, a energia sempre ligada é um imperativo.
- **A IA significa perfis de carga não convencionais:** A IA está associada com cargas de potência que podem variar de uma ociosidade de 10% a uma sobrecarga de 150% em um piscar de olhos.
- **Pressão para modernizar e atualizar:** Operadores de data centers precisarão determinar a melhor forma de fazer o retrofit em suas operações existentes para antever as demandas listadas acima.



Acredito que a arquitetura passará por mudanças importantes e uma parte crucial do sucesso será antecipar estas mudanças. Por exemplo, é importante se perguntar: “Como fazemos o planejamento para racks de 250 kilowatts ou maiores? A forma como é feito hoje é inviável, portanto, o momento para pensar sobre as mudanças é agora.”



Gregory Ratcliff
Chief Innovation Officer, Vertiv

Estão vindo grandes mudanças na arquitetura, na alimentação de energia e no resfriamento

Um resumo dos maiores avanços desta década

As atuais tendências apontam para grandes mudanças ocorrendo na infraestrutura de TI em geral. Mas, quais mudanças específicas veremos e quanto elas irão acontecer?

Determinar as respostas para essas questões começa no nível mais básico. À medida que a arquitetura das GPUs e a arquitetura das filas evoluem, veremos uma evolução

correspondente tanto na alimentação de energia como no resfriamento.

Embora ninguém possa estar completamente certo sobre o futuro da arquitetura de TI, olhar para os principais motivos da mudança é a melhor forma de determinar como esse futuro poderá ser.

	2024	2025	2027**	2029+**
	Os atuais designs das GPUs de IA são principalmente baseados em resfriamento a ar e resfriamento híbrido a líquido e a ar para uma maior eficiência operacional. ¹⁰	Os próximos designs das GPUs serão principalmente baseados no resfriamento líquido ao chip de uma fase e no resfriamento a ar para cargas térmicas residuais. ¹¹	A próxima geração (não anunciada, e sim prevista) de designs de GPUs para IA será principalmente baseada em maior fluxo e menor temperatura para uma performance melhor. ¹²	As gerações futuras (trajetória prevista) dos designs de GPUs para IA serão principalmente baseadas no resfriamento líquido ao chip ou alguma forma de imersão confinada. ¹³
Arquitetura das GPUs	Até 32 núcleos de GPU NVIDIA Hopper H100 (ou similar) por rack, configurados em clusters/pods na fila de até 240 núcleos de GPU. ¹⁴	Até 72 núcleos de GPU NVIDIA Blackwell B200 (ou similar) por rack, configurados em clusters/pods na fila de até 720 núcleos de GPU. ¹⁵	Núcleos de GPU de nova geração (ciclos de lançamento de 12 meses) em configurações similares de cluster/pod na fila.	
Arquitetura dos Racks¹⁶	Até 50 kW por rack com distribuição de energia N+1 e 400/230 VCA. Resfriados usando trocadores de calor de porta traseira de líquido para ar.	Até 140 kW por rack com distribuição de energia N+1 e 1400/230 VCA ou 400 VCC. Resfriados usando placas frias de líquido para o chip em uma fase (80% da carga térmica), complementadas por trocadores de calor de porta traseira de líquido para ar (20% da carga líquida para ar (20% da carga térmica).	Até 300 kW por rack com distribuição de energia N+1 e 400/230 VCA ou 400 VCC. Resfriados usando placas frias de líquido para o chip em uma fase complementadas por trocadores de calor de porta traseira de líquido para ar.*	Densidades de até 1 MW por rack. Até 2031, deve ser considerado no design 480/277 VCA, 600/347 VCA ou 800 VCC. Distribuição de energia N+1 e 400/230 VA ou 400 VCC. Resfriamento por placas frias (100% da carga térmica) deve ser considerado no design.

*Fator de forma do rack provavelmente mudará em tamanho e configuração a partir de 2027 e provavelmente não corresponderá às versões da EIA e OPC usadas hoje.

¹⁰ NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; acessado em 12 de agosto, 2024

¹¹ NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; acessado em 12 de agosto, 2024

¹² Trueman, C. "Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;" Data Center Dynamics; Janeiro 8, 2024; <https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/>; acessado em 14 de agosto, 2024

¹³ ibid.

¹⁴ NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; acessado em 12 de agosto, 2024

¹⁵ NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; acessado em 12 de agosto, 2024

¹⁶ Wang, S. "2024 Trends to Watch: Data Center Physical Infrastructure;" Omdia; Dezembro 22, 2023; <https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure>; acessado em 12 de agosto, 2024



	2024	2025	2027**	2029+**
Arquitetura da Fila ¹⁷	Recomenda-se barramento blindado rígido suspenso de 400 A com caixas de derivação, proporcionando a máxima flexibilidade para substituição rápida e atualização da GPU. Circuitos dedicados cabeados no campo a partir da PDU podem ser considerados para fins de custos. Unidades de distribuição de resfriamento de 600 kW localizadas na fila e unidades perimetrais de tratamento de ar de sala de computadores	Recomenda-se barramento blindado rígido suspenso de 2000 A com caixas de derivação, proporcionando a máxima flexibilidade para troca rápida e atualização da GPU. Circuitos dedicados cabeados no campo a partir da PDU não são recomendados devido ao custo de instalação e a bitola inflexível dos fios. Unidades de distribuição de líquido de 1350 kW localizadas na fila e unidades perimetrais de tratamento de ar de sala de computadores.	Recomenda-se barramento blindado rígido suspenso de 4000 A com caixas de derivação, proporcionando a máxima flexibilidade para troca rápida e atualização da GPU. Circuitos dedicados cabeados no campo a partir da PDU não são recomendados. Unidades de distribuição de líquido de 3000 kW localizadas na fila e unidades perimetrais de tratamento de ar de sala de computadores.	Recomenda-se barramento blindado rígido suspenso de 4000 A com caixas de derivação, proporcionando a máxima flexibilidade para troca rápida e atualização da GPU. Circuitos dedicados cabeados no campo a partir da PDU não são recomendados. Unidades de distribuição de líquido de diversos MW.
Gerenciamento de Energia	Alimentação da rede elétrica no mínimo Tier II ou Tier III apoiada por geração de energia de standby N+1 com Sistemas de Armazenamento de Energia em Baterias (BESS) no site ou similar devem ser considerados para interação com a rede, arbitragem e resiliência.		Alimentação da rede elétrica no mínimo Tier II ou Tier III apoiada por geração de energia de standby N+1 com Recursos de Energia Distribuídos (DER), Sistemas de Armazenamento de Energia em Baterias (BESS) no site ou células de combustível devem ser considerados para interação com a rede totalmente dinâmica, arbitragem e resiliência. Monitoramento de energia integrado de alta velocidade na caixa de derivação (barramento blindado) com controle térmico inteligente é exibido (proprietário da Vertiv).	
Carga de Trabalho da GPU	Até 160% Estamos considerando que as implementações iniciais de GPU sejam >20% da carga de trabalho total para empresas até 2025 e aumentando para 100% da carga total até 2027.	Até 160%	Até 160%	Até 160%
Ciclo Térmico ¹⁸	Expansão direta (DX) e água gelada (CW).		Circuito de água gelada de alta temperatura.	Circuito de água gelada.
Reuso do Calor	Uso do Ciclo Rankine Orgânico (ORC), Captura de Ar Direta (DAC), e chillers de absorção.			
	Mais de 25% de reuso de calor.	Mais de 25% de reuso de calor.	Mais de 25% de reuso de calor.	Mais de 80% de reuso de calor.
Controles	Capturar a telemetria operando no data lake e implementar manutenção baseada na condição (CBM).		Implementar gêmeos digitais em todo o data center e alavancar informações contínuas de telemetria para a otimização da operação usando IA.	Auto-otimização e operação robótica autônoma.

**Os detalhes, estimativas e projeções técnicas e de gerenciamento aqui citadas são baseadas em estudos da Vertiv, expertise interno e informações atualmente disponíveis de parceiros da Vertiv na indústria. Previsões analisadas são baseadas no design e nas tecnologias de infraestrutura disponíveis na data de publicação desde white paper. As projeções são apenas para fins ilustrativos.

A trajetória ascendente da densidade dos racks desencadeada pela onda atual de IA não será um mero ruído histórico. Para atender à crescente demanda por IA, a arquitetura dos racks aumentará de densidades dos 30 kW superiores para 300-600 kW no curto prazo e possivelmente para 1 MW e mais até 2030.

Logicamente, um certo nível de incertezas permanecerá ao redor desse número com base não apenas nos avanços tecnológicos, mas também sobre a demanda por IA. Entretanto, é razoável esperar que o futuro seja baseado na crescente densidade dos racks exigidas pela IA e pelos avanços tecnológicos na infraestrutura crítica, necessárias para atender a essas demandas.



As coisas estão mudando hoje muito mais rapidamente do que em qualquer momento nos últimos 30 anos. Para administrar estas mudanças, precisamos trabalhar juntos, como setor, para estabelecer as melhores práticas e designs de referência. É essencial para estarmos preparados para as mudanças que já estão acontecendo."

Peter Panfil
Vice-Presidente, Global Power, Vertiv

¹⁷ ibid.

¹⁸ Wang, S. "AI-driven Data Center Cooling Systems and Technologies;" Omdia; Agosto 2, 2024; <https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies>; acessado em 15 de agosto, 2024

Embora possa haver incertezas sobre o quão rapidamente a arquitetura dos racks evoluirá, especialmente no longo prazo, há poucas dúvidas de que as mudanças na arquitetura demandarão novas abordagens ao resfriamento. Os atuais designs das GPUs de IA são, em sua maioria, baseados em resfriamento a ar.

Entretanto, o tremendo crescimento no uso de energia significará que várias formas de resfriamento híbrido e líquido farão parte do futuro. O resfriamento líquido é significativamente mais eficiente do que o resfriamento a ar, e esta eficiência é extremamente importante para o desenvolvimento da infraestrutura necessária para IA. Ao final da década, veremos os data centers recorrendo principalmente ao resfriamento líquido ao chip, imersão independente e resfriamento a ar para as cargas térmicas residuais.



O aumento do resfriamento líquido deverá suscitar algumas dúvidas para quem opera data centers. Por exemplo, “Como ter certeza de que o circuito de líquido está estável? Como ter certeza de que ele é redundante? Como coloco um rack no sistema? Como removo um rack do sistema? Com as mudanças que estão vindo, todas essas são questões que as pessoas precisam avaliar.”



Steve Madara
Vice-Presidente,
Global Cooling, Vertiv



O resfriamento não é a única preocupação. Os data centers do futuro também precisarão ser flexíveis o suficiente para lidar com “cargas de pulso”, ou cargas periódicas com alta corrente de inrush em um período de tempo muito curto. Os clusters de treinamento da IA são conhecidos por exibir picos de curta duração na corrente elétrica durante determinados ciclos de computação, portanto, serão obrigatórias provisões para sobrecarga.

Porém, os data centers provavelmente têm um problema maior do que picos de corrente no que se refere à energia. Somente nos Estados Unidos, o consumo total de eletricidade deve crescer quase que inteiramente com base na demanda dos data centers. De acordo com dados da Goldman Sachs, demanda dos data centers por energia terá uma taxa de crescimento anual composta (CAGR) de 15% até 2030, com os data centers representando 8% do total da demanda americana por energia até o final da década, partindo dos atuais 3%.



Uma de nossas funções hoje é tornar mais disponível a energia que existe e garantir que ela seja usada com mais eficiência no data center. Você precisa ter certeza de que usa a energia que tem de forma eficaz.”



Stephen Liang
Diretor Executivo de
Tecnologia (CTO) e
Vice-Presidente
Executivo, Vertiv



Embora seja difícil apontar exatamente todos os marcos tecnológicos que ocorrerão nesta década e depois, as linhas gerais do futuro estão se tornando mais claras. Sabemos que a densidade dos racks aumentará. Sabemos que os métodos de resfriamento e de gerenciamento precisarão evoluir. Sabemos que a alimentação de energia e o gerenciamento da carga estão rapidamente se tornando desafios importantes.

Agora precisamos aplicar o que sabemos para abastecer com informações o desenvolvimento dos produtos e das soluções que resolverão as necessidades tanto dos data centers de hoje quanto as necessidades dos data centers do futuro.

Por exemplo, precisamos planejar para maiores densidades do rack através do desenvolvimento de unidades de distribuição de energia para racks de alta densidade, para gerenciar a distribuição de energia de forma mais eficaz nos data centers que rapidamente evoluem. Também devemos revisitar as tecnologias existentes de resfriamento líquido e nos perguntar como podemos estar preparados para um mundo onde o resfriamento das placas frias de líquido ao chip de duas fases é uma necessidade. E precisamos olhar para os perfis reais das cargas de IA e desenvolver formas de otimizar a performance com base nos perfis que estamos vendo agora e nos perfis que prevemos para o futuro à medida que a IA avance.

A chave para se adaptar?

Alavancar a tecnologia e construir boas parcerias

Realizar todas estas coisas não é necessariamente fácil. Embora preparar-se para as mudanças sendo trazidas pela IA seja essencial, é igualmente importante reconhecer que qualquer tentativa de tornar as suas operações prontas para o futuro não pode ser feita no vácuo.

O CEO da NVIDIA, Jensen Huang, resumiu a situação de forma excelente quando disse, “A AI não é uma questão de chip. É uma questão de reinventar a computação. Você não pode resolver essa nova forma de computação apenas projetando um chip. Todos os aspectos do computador mudaram profundamente.”

Como damos suporte e alavancamos a IA também está mudando. Já estamos vendo a ascensão da fábrica de IA, um hub centralizado que permite que especialistas, cientistas de dados e engenheiros colaborem no desenvolvimento, implementação e escalada de soluções com base na IA.

À medida que a IA cresça, o modelo da fábrica também o fará. Funcionando de forma similar às fábricas tradicionais, mas criando inteligência e novos insights implementáveis ao invés de produtos físicos, o modelo precisa de uma infraestrutura complexa e um design holístico para funcionar.

Na prática, isso quer dizer que nenhum player sozinho tem todas as respostas. Portanto, prosperar em um mundo dominado pela IA significará ir além de resolver um problema com energia, resfriamento ou chips. Ao invés disso, os operadores inteligentes verão como os problemas que enfrentam se encaixam no ecossistema geral.

É por isso que estar pronto para o futuro exigirá desenvolver relacionamentos fortes com parceiros qualificados que não estão apenas assistindo as mudanças trazidas pela IA, mas estão impulsionando estas mudanças.

Mesmo para as empresas liderando o espaço de IA, reunir expertise dos players mais experientes é uma parte importante para maximizar os benefícios da IA. Integrar as perspectivas de todos, de desenvolvedores de chips a gestores de energia, é fundamental para fazer previsões inteligentes sobre as demandas da IA e ter a capacidade de atender a essas demandas.

Na prática, o que isto significa? Dado que elementos como chips inovadores terão um efeito enorme na arquitetura e nas operações futuras dos data centers, a Vertiv tem trabalhado em estreita colaboração com os principais fabricantes de chips, atuando como uma consultora sobre a [tecnologia dos chips que evolui rapidamente](#) e como arquitetura para a [infraestrutura necessária para abrigar esses chips](#). Como parte de vários

programas globais envolvendo inovadores tecnológicos que oferecem soluções fomentadas ou movidas por tecnologia, podemos ajudar a garantir que a infraestrutura digital crítica e as GPUs funcionem em conjunto para entregar uma performance eficaz.

E, ao oferecer nossa expertise para redes que também incluem os principais fornecedores de software, provedores de serviços de nuvem, provedores de soluções e integradores de sistemas, estamos fortalecendo o trabalho que realizamos hoje e nos preparando para as mudanças do futuro através da [integração do portfólio completo da Vertiv de soluções para energia e para resfriamento com as mais novas tecnologias](#) de todo o setor.

No mercado atual, um amplo banco de expertise técnica incomparável é muito importante, mas se você realmente quiser tornar suas operações a prova de futuro, essa expertise precisa trabalhar em conjunto com as outras partes do ecossistema de IA em desenvolvimento.



Dadas as mudanças sem precedentes que os data centers estão começando a enfrentar com a ascensão da IA, eu pensei: “Acredito que precisamos ir atrás do pessoal dos chips, especificamente do pessoal dos chips disruptores. Precisamos aprender o que eles irão fazer e construir que irá provocar uma disrupção em nosso ecossistema.”

Greg Stover
Diretor Global,
Desenvolvimento de
Alta Tecnologia, Vertiv



Como será o futuro da infraestrutura de TI?

É isso que os principais especialistas estão dizendo

A necessidade de dar suporte à IA já está conduzindo grandes mudanças na infraestrutura dos data centers. Entretanto, os maiores avanços ainda estão por vir. Durante o restante da década, a IA será a força motriz de uma série de mudanças, com novas arquiteturas das GPUs, dos racks e das filas, todas contribuindo para mudanças significativas na infraestrutura de TI e na infraestrutura digital crítica. O que será preciso para aproveitar as mudanças e as oportunidades apresentadas pela ascensão da IA?

Alimentar o futuro significa enxergar além do óbvio



Peter Panfil

Vice-Presidente, Global Power, Vertiv

O Peter lidera o desenvolvimento estratégico de clientes para os negócios de Power da Vertiv. Ele é especializado em resolver os desafios dos clientes com as mais novas tecnologias de energia e de controle entregando os níveis de disponibilidade, escalabilidade e eficiência para atender as necessidades diversas dos clientes e de sustentabilidade. Ele tem um profundo conhecimento em soluções de energia para dar suporte às necessidades únicas das aplicações de IA e é um defensor da abordagem “Traga sua Própria Energia” (“Bring your Own Power”) para solucionar os problemas da dependência de concessionárias de energia elétrica. Com mais de 30 anos no espaço de infraestrutura crítica, Peter ocupou posições executivas, incluindo VP de Engenharia e VP/GM de AC Power antes de suas responsabilidades atuais. Ele é um apresentador e palestrante frequente nas feiras, conferências e nos meios de comunicação, atendendo aos segmentos de TI, facilities e engenharia e é um autor publicado através de sua contribuição para o livro “Greener Data Vol. 2”, de 2024.



Por muito tempo vivemos felizes no mundo dos data centers com blocos de 3 MW diretamente acoplados a um gerador de 3 MW com um barramento de 4000 A. Estes dias estão acabando. Com a ascensão da IA, agora é comum começar a falar em blocos de 10 MW e até de 20 MW.

Então, no que se refere ao uso de energia nos data centers, não é segredo que estamos vendo um aumento massivo em tamanho e em escala. A conclusão é: Os operadores precisarão descobrir de onde essa energia virá.

Mas da mesma forma importante, a IA também resultará em cargas passando por picos associados ao treinamento da IA generativa e a demanda por alimentação de energia contínua associada com a necessidade de resfriamento. Esses últimos dois fatores são necessidades relativamente recentes e serão um desafio cada vez maior em toda a indústria nos próximos anos.

Então, qual o segredo para resolver esses desafios relativos à energia? Creio que uma grande parte será antecipar quais serão as mudanças mais específicas da energia. Você precisa saber para onde os fabricantes de chips estão indo e precisará saber o que os profissionais da indústria de TI acham.

Quando se trata de soluções de energia, os inovadores já estão fazendo perguntas como: ‘Posso simplificar e implementar facilmente e com um custo baixo?’ e ‘Posso ter certeza de que essa solução é confiável?’ Na minha mente, as respostas são definitivamente ‘Sim’. Entretanto, é importante compreender que chegar até o ‘sim’ envolverá trabalhar com as pessoas certas; pessoas que entendem tanto dos chips mais recentes como das soluções de energias mais recentes.”

Use a resiliência para proteger seus ativos



Stephen Liang

Diretor Executivo de Tecnologia (CTO) e
Vice-Presidente Executivo, Vertiv

Stephen Liang alinha as estratégias tecnológicas e os recursos da Vertiv para priorizar a voz do cliente na pesquisa e desenvolvimento de produtos. Ele possui três décadas de experiência em soluções para alimentação de energia, operações de fabricação e P&D. Com uma graduação e mestrado em engenharia mecânica pelo Massachusetts Institute of Technology (MIT), ele combina tanto conhecimento técnico como uma longa experiência em sua visão sobre como a tecnologia está mudando a infraestrutura de TI.

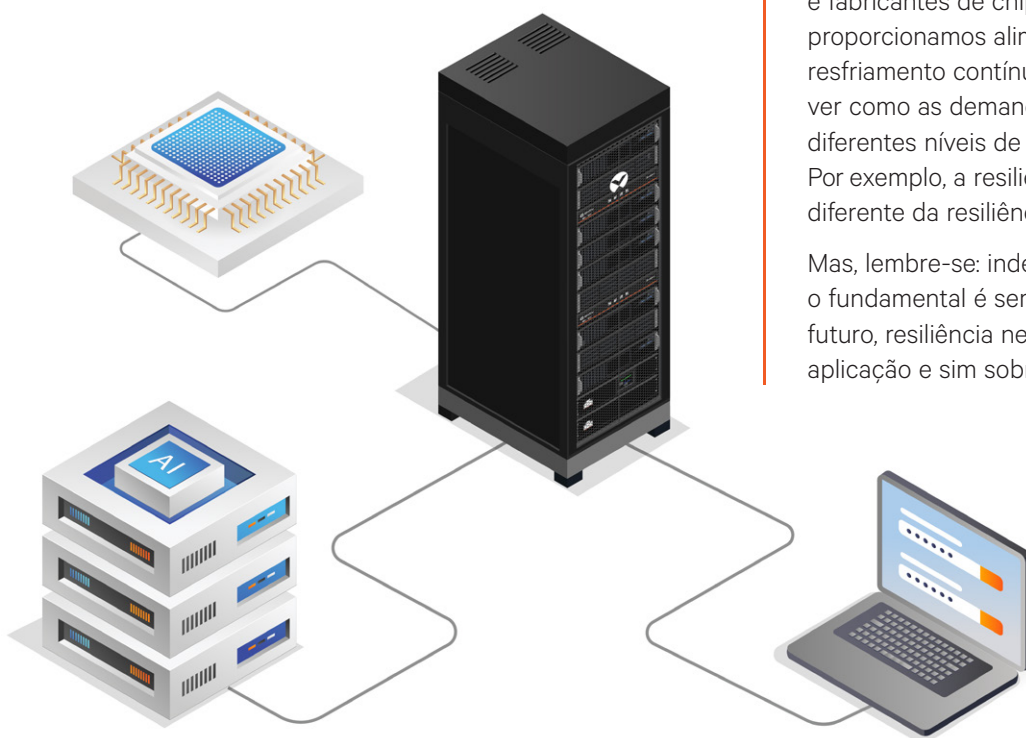


Tenho uma ampla experiência com infraestrutura de TI e posso dizer sem nenhuma dúvida que a mudança tecnológica no setor está ocorrendo mais rápido do que nunca. Uma das razões é porque esperamos que a tecnologia faça mais do que nunca e estamos pedindo à infraestrutura como um todo que faça o mesmo.

Vimos um enorme investimento nos valiosos chips de computação e sistemas de servidores de última geração, muitos dos quais relacionados com a IA. Esses investimentos significam que temos uma maior responsabilidade maior do que nunca em manter esses sistemas funcionando. Para tirar o máximo proveito desses ativos, nossos clientes precisarão que eles funcionem 24x7. Portanto, acho que uma das palavras mais importantes a ser observada no futuro é resiliência. Obviamente, ter uma infraestrutura de TI resiliente sempre foi uma prioridade. Mas no futuro, a resiliência será mais importante e mais complexa do que nunca.

O que quero dizer com isso? Na Vertiv, temos um longo histórico na formação de parcerias com desenvolvedores e fabricantes de chips para ganhar informações enquanto proporcionamos alimentação de energia contínua e resfriamento contínuo – faz parte de nosso DNA. Podemos ver como as demandas da IA muitas vezes requerem diferentes níveis de resiliência para diferentes funções. Por exemplo, a resiliência necessária para treinamento é diferente da resiliência necessária para inferência.

Mas, lembre-se: independentemente do seu objetivo final, o fundamental é sempre proteger o seu investimento. No futuro, resiliência nem sempre será sobre a proteção da aplicação e sim sobre a proteção do valor do ativo.”



Aprenda a aproveitar o que você tem



Martin T. Olsen

Vice-Presidente Sênior, Estratégia Global para Produtos, Vertiv

Martin Olsen é responsável pela estratégia e planejamento global para produtos. Seu trabalho inclui a condução do roadmap para a interoperabilidade dos produtos para entregar soluções completas, escopo do investimento, avaliação e priorização ao longo das unidades de negócios da Vertiv mundialmente. Com experiência em várias posições de liderança em engenharia, estratégia e planejamento, ele é um especialista em infraestrutura digital crítica no segmento de data centers. Martin possui várias patentes dos EUA relacionadas ao gerenciamento térmico e tem o título de Bacharel em Ciências em Marketing e Análise de Negócios pela Universidade de Syddanmark, Kolding, Dinamarca e completou o Programa de Gestão Geral na Harvard Business School.



Vivemos em um mundo onde a ascensão da IA significa que a densidade dos racks pode chegar a 1 MW por rack no final da década. Para quem já tem importantes investimentos em infraestrutura, é necessário pensar sobre como podem aproveitar ao máximo esses investimentos enquanto se preparam para uma nova escala de demandas no futuro.

Em outras palavras, em um ambiente que está mudando rapidamente, como você aproveita o que já tem sem arriscar se tornar obsoleto?

Muitos operadores de data center precisarão levar em consideração fazer o retrofit de seus ativos. Mas à medida que a densidade dos racks continuar a crescer com uma taxa acelerada, apenas decidir sobre um retrofit não será suficiente. Ao invés, você precisará levar em consideração para quais densidades você está fazendo o retrofit e quais tecnologias de resfriamento serão necessárias como parte do seu retrofit.

Implementar e usar IA nos negócios é muito transformador e os clientes estão extremamente conscientes da frase “Faça a Disrupção ou Sofra a Disrupção” de John Chamber, antigo CEO e Conselheiro Emérito da Cisco Systems. Eles querem fazer as coisas de uma forma que os diferencie de seus concorrentes. Em relação ao retrofit, deve-se tomar uma decisão apenas após considerar todas as opções que existem e analisar como estas opções podem trabalhar em conjunto de forma inteligente.

Isso pode ser uma tarefa difícil. Afinal, como você sabe quais são as suas opções? Qual é a melhor maneira de ver ao mesmo tempo os detalhes e o panorama geral? Na minha mente, o que fará o operador do data center estar verdadeiramente preparado para o futuro será encontrar parceiros que não apenas tenham um amplo conhecimento sobre as tecnologias emergentes, mas também saiba como estas tecnologias podem trabalhar em conjunto para atender a demandas futuras. Sem este parceiro, acho que os operadores de data centers correm o risco de ver o futuro passar por eles.”

Onde procurar os produtos mais inovadores do futuro



Greg Ratcliff
Diretor Executivo de Inovação, Vertiv

Greg Ratcliff participa da indústria há mais de 30 anos com ênfase em inovações orientadas por dados. Ele liderou o grupo global de analytics avançada para a equipe de serviços, especializando-se em dados em tempo real e conectando os cerca de 1 bilhão de produtos Vertiv em operação à nuvem da Vertiv. O Greg constrói relacionamentos com parceiros tecnológicos, universidades e as equipes de engenharia da Vertiv para garantir que a empresa mantenha o foco no futuro e esteja pronta para dar suporte às mudanças na indústria. Sua formação acadêmica inclui graduações em Matemática Aplicada, Engenharia da Computação e Engenharia Elétrica pela Ohio State University, um MBA pela Universidade de Phoenix e um treinamento especializado centralizado em gerenciamento ágil de projetos para projetos de IoT e Big Data.



Gosto de pensar sobre P&D em três contextos: Agora, médio prazo e longo prazo. O ‘agora’ é onde você encontra os produtos que pode usar hoje. Estes são essenciais para manter a infraestrutura de TI atualizada.

Mas, qualquer empresa querendo acompanhar o ritmo das mudanças rápidas que já estão acontecendo não pode se dar ao luxo de ignorar o ‘médio’ e o ‘longo’ prazo. As tecnologias essenciais para a infraestrutura de TI serão significativamente diferentes das que usamos hoje, e é por isso que ficar de olho no pipeline de desenvolvimento é um ponto essencial para estar preparado para o futuro.

Na minha realidade, uma forma de nos prepararmos para o futuro é ter relacionamentos ativos com dezenas de universidades. Alguns desses relacionamentos são formais, como o The Center for Energy- Smart Electronic Systems (ES2). O ES2 é um Centro de Pesquisas de Ciências Colaborativo da National Science Foundation/ Universidades/Indústria (I/UCRC) estabelecido para desenvolver ferramentas para maximizar a eficiência energética para a operação de sistemas eletrônicos, incluindo data centers, ‘do chip a torre de resfriamento’. Outros relacionamentos com a academia são menos formalizados e muito valiosos.

Mas esses relacionamentos com universidades não significam que vivemos em uma torre de marfim. Na verdade, os clientes usuários finais da Vertiv são muitas vezes uma parte importante desses relacionamentos. Gosto de pensar nisso como sendo uma banquetta de três pés que sustenta a inovação prática. A inovação pode vir de um laboratório de uma universidade, mas a Vertiv e nossos clientes têm um papel fundamental para tornar essas inovações em um produto pronto para atender às demandas de um mercado mais amplo.

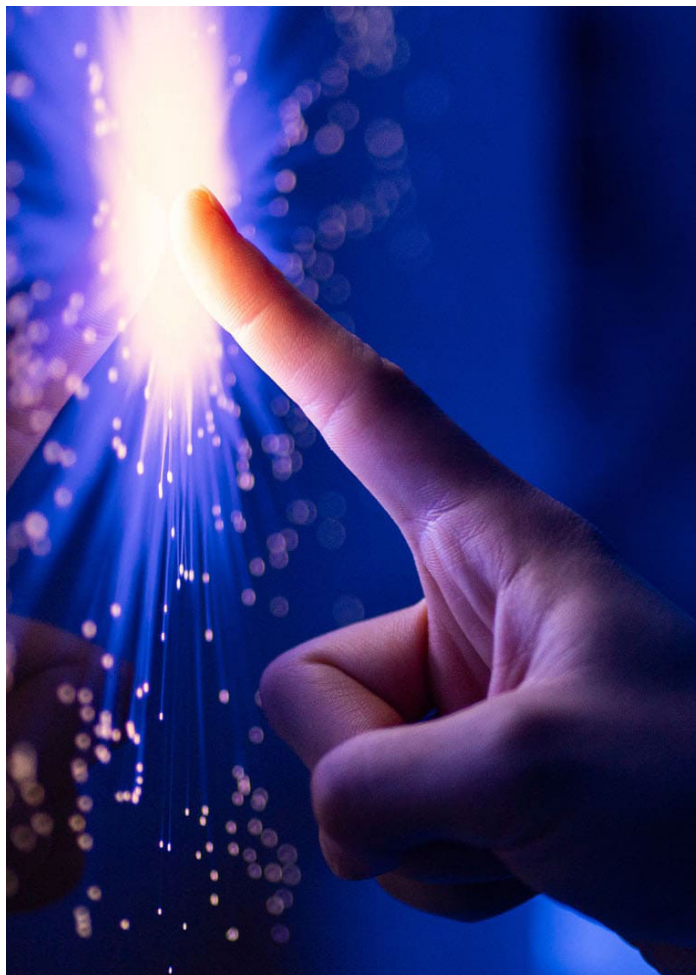
E as inovações que se tornam produtos reais serão uma parte essencial da evolução da infraestrutura digital crítica. Se você quer estar pronto para o futuro, faça uma parceria com uma empresa que tenha laços profundos com líderes tecnológicos e inovadores. Pois no ambiente atual você pode estar focado no ‘agora’, mas você também precisa manter um olho no ‘médio’ e ‘longo’ prazo.”

Para onde vamos a partir de agora?

A ascensão da IA tornou esta época verdadeiramente emocionante no mundo da computação. Mas com a emoção e as possibilidades surgem as perguntas e as incertezas. Muitas entidades com data centers estão lutando para adicionar capacidade de IA em suas operações existentes. Ao mesmo tempo, aqueles que estão planejando data centers novos se perguntam exatamente quanta capacidade para IA eles devem incluir em seus projetos. E como ficará a arquitetura dos racks? E as necessidades de alimentação de energia e resfriamento?

Responder a esse tipo de questão será difícil, mas uma forma de fazer com que seja mais fácil é garantir que esteja projetando a sua infraestrutura digital crítica com líderes no desenvolvimento da IA, e não com seguidores. Pois em um mundo onde a mudança é tanto rápida como certa, a abordagem errada poderia ser o investimento em uma solução que se torna obsoleta logo após ter sido implementada. Ao ser inteligente e intencional, você pode evitar fazer projetos para a obsolescência e, ao invés, criar algo mais sustentável, confiável e duradouro.

Ter sucesso na revolução da IA significa escolher suas soluções com inteligência, mas principalmente significa ser ainda mais inteligente em relação a quem fornece as suas soluções. Você não precisa apenas da tecnologia certa, mas precisa também de um parceiro com uma ampla compreensão da cadeia de energia e da cadeia térmica, de ponta a ponta; e com especialistas que possam aconselhar sobre como alimentar as necessidades da sua IA tanto hoje quanto amanhã.



Seu parceiro ideal é um parceiro com um amplo portfólio de tecnologias de alimentação e resfriamento, com uma grande rede de serviços e uma presença global. Por que? Porque você precisa de um parceiro que possa capacitá-lo a ser transformador, e para ser transformador nesse momento do jogo com a IA, você precisa de velocidade e de escala”



Martin Olsen

Vice-Presidente Sênior, Estratégia
Global para Produtos, Vertiv



Referências

Adaptado de The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT, por McCandless, D. Evans T. e Barton P, 2024 (<https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt>). De domínio público

Epoch AI; "Notable AI Models;" 19 de junho, 2024; <https://epochai.org/data/notable-ai-models>, acessado em 10 de julho, 2024

Galabov, V e Sukumaran M; "Cloud and Data Center Market Snapshot – Novembro 2023;" Omdia; Novembro 2023

Goldman Sachs; "Generational growth: AI, data centers and the coming US power demand surge;" publicado em 28 de abril, 2024; <https://www.goldmansachs.com/intelligence/pages/gs-research/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf>; acessado em 30 de maio, 2024

Goldman Sachs; "AI is poised to drive 160% increase in data center power demand;" 14 de maio, 2024. <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>, acessado em 5 de agosto, 2024

McKinsey Digital; "The economic potential of generative AI: The next productivity frontier;" publicado em 14 de junho, 2023; <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>; acessado em 30 de maio, 2024

Newmark; "2023 U.S. Data Center Market Overview & Market Clusters;" Janeiro 2024; <https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview-market-clusters>, acesso em 10 de julho, 2024

NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid>; acessado em 12 de agosto, 2024

NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" <https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid>; acessado em 12 de agosto, 2024

Trueman, C. "Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;" Data Center Dynamics; 8 de janeiro, 2024; <https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/>; acessado em 14 de agosto, 2024

Villars R. et al. "Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending;" IDC #US51539723, Dezembro, 2023

Wang, S. "2024 Trends to Watch: Data Center Physical Infrastructure;" Omdia; 22 de dezembro, 2023; <https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure>; acessado em 12 de agosto, 2024

Wang, S. "AI-driven Data Center Cooling Systems and Technologies;" Omdia; 2 de agosto, 2024; <https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies>; acessado em 15 de agosto, 2024

