VERTIV™

Revertendo a Tendência de Crescimento nos Custos dos Downtimes em Data Centers

Frequência de Downtimes no Core e no Edge

Embora velocidade e eficiência do capital sejam necessidades no atual mercado altamente competitivo de data centers, essas metas devem ser colocadas no contexto da disponibilidade do data center.

Uma nova pesquisa do Ponemon Institute, Downtime em Data Centers de Core e de Edge: Uma Pesquisa sobre Frequência, Duração e Ações, revela que os 132 data centers de core (centrais) incluídos no estudo tiveram uma média de 2,4 paradas totais da instalação por ano e mais 10 eventos de downtime isolados em racks ou servidores específicos. Além disso, os 1.667 locais de edge incluídos no estudo tiveram uma média de 2,7 paradas completas não planejadas em um ano.

O que é particularmente alarmante sobre os achados dessa pesquisa é a duração das indisponibilidades ter aumentado em comparação com a última vez que o estudo foi realizado em 2016. A duração média de uma parada total em um data center de core aumentou para 138 minutos, um aumento de 8 minutos sobre o estudo anterior. Com as organizações dependendo cada vez mais de seus data centers e expandindo suas redes de edge, elas não estão apenas passando por uma alta frequência de indisponibilidades, mas estão também levando mais tempo para se recuperar dessas indisponibilidades.

Embora os participantes desse estudo estejam localizados nas Américas, os resultados do estudo são corroborados pela Pesquisa 2020 Global Data Center Survey do Uptime Institute. Essa pesquisa identificou que "indisponibilidades ocorrem com uma frequência preocupante, que as maiores indisponibilidades estão se tornando mais nocivas e mais caras e que, o que foi ganho com processos melhores e engenharia, foi parcialmente neutralizado pelos desafios de manter sistemas mais complexos."

Embora haja diversos desafios associados com o gerenciamento de data centers atualmente, incluindo a pressão para implementar capacidade com maior velocidade e com ótima relação custo-benefício, o verdadeiro desafio da disponibilidade é de tal importância que não pode ser relegado a uma prioridade mais baixa. Este estudo propõe estratégias que as organizações podem usar para minimizar sua exposição ao downtime, incluindo novos enfoques à redundância e escalabilidade de UPS, melhor monitoramento e acesso remoto, baterias de íon-lítio e estratégias para distribuição de energia de alta disponibilidade.

Avaliando as Ações que Afetam a Disponibilidade

Além de quantificar a frequência e a duração dos downtimes no core e no edge, o estudo do Ponemon explora também as ações organizacionais relacionadas a diversos fatores que podem afetar a disponibilidade do data center (Figura 1).

Nos dois tipos de instalação, a contenção de custos parece ser um dos principais motivos dos downtimes. Sessenta e nove por cento dos participantes disseram que o risco de downtime não planejado aumentou em seu data center de core como resultado de contenções de custos, enquanto 62% disseram o mesmo em relação às suas instalações de edge. Além disso, apenas a metade dos participantes disse que sua alta direção dava completo suporte aos seus esforços para evitar downtimes, tanto no core quanto no edge.

Nem as instalações de core, nem as instalações de edge estavam bem equipadas para se recuperar de uma indisponibilidade não planejada. Apenas 38% dos participantes acreditavam ter recursos suficientes no edge para retomar o funcionamento da instalação caso uma indisponibilidade não planejada ocorresse. Isso é de alguma forma esperado, já que essas instalações são muitas vezes remotas e sem a presença de pessoas. Mas foi surpreendente ver que apenas 43% dos participantes acreditavam ter esses recursos disponíveis nos data centers de core, provavelmente contribuindo para os tempos maiores de recuperação encontrados na pesquisa desse ano.

Figura 1: Comparação dos atributos do data center de core e de edge.

Finalmente, data centers de edge são mais propensos a utilizar melhores práticas do que data centers de core, embora em nenhum caso os percentuais sejam particularmente altos. Quarenta e seis por cento dos participantes disseram usar melhores práticas em seus data centers de core, comparado com 54% em suas instalações de edge.

Essas atitudes estão aparecendo no design dos data centers de edge. Pela ótica da disponibilidade, estamos vendo uma maior redundância ser usada no edge. Enquanto os data centers de core podem estar trocando para N+1, o edge é percebido como sendo a linha de frente da disponibilidade e muitas vezes é implementado como 2N.

Lidando com as Causas Raiz

As principais causas de downtimes não planejados identificadas pelos participantes na pesquisa do Ponemon incluem ciberataques, falha em equipamentos de TI, erro humano, falha na bateria do UPS e falha do equipamento UPS. Quando considerando essas causas raiz, é importante pensar nos resultados da Pesquisa 2020 Global Data Center Survey do Uptime Institute, que relatou que três em cada quatro participantes afirmaram que seus mais recentes eventos de downtime eram evitáveis.

Por exemplo, muitas falhas de equipamentos de TI poderiam ser evitadas através do monitoramento e troca antes da falha? A mesma pergunta pode ser feita sobre as falhas na bateria do UPS. Sistemas de monitoramento de baterias, quando implementados adequadamente, podem identificar prováveis falhas na bateria antes que elas ocorram.

Claramente, as contenções de custos sendo impostas sobre os responsáveis pela disponibilidade das instalações e o correspondente uso limitado das melhores práticas estão tendo influência na frequência relativamente alta dos eventos de downtime revelados pelo estudo do Ponemon.

Conforme o Uptime Institute observa em sua pesquisa 2020 Global Data Center Survey: “não está claro se operadores estão realmente aprendendo com seus problemas de processo ou culpando seus gestores. Também é possível que os gestores culpem os operadores – ou todos podem estar culpando os investidores por não investir o suficiente. De qualquer forma, os achados apontam para uma oportunidade clara: com mais investimentos em gerenciamento, processos e treinamento, a frequência das indisponibilidades certamente teria uma redução significativa.”

Eventos de downtime representam uma situação de crise. O foco sempre é em colocar o data center para funcionar o mais rápido possível. Mas, muitas vezes, parece que a recuperação não é seguida por planejamento e investimentos suficientes para fortalecer a infraestrutura crítica do data center de forma que a probabilidade de eventos futuros seja reduzida.

Estratégias para Reduzir a Frequência e a Duração das Indisponibilidades do Data Center

2020 foi um ano desafiador para o gerenciamento de data centers. Várias organizações tiveram uma demanda para aumento de capacidade devido à pandemia global, ao mesmo tempo em precisaram implementar novos protocolos e trabalhar com orçamentos reduzidos. Porém, esses fatores não podem ser aceitos como desculpas para o aumento de downtimes. A disponibilidade dos serviços é mais importante do que nunca.

A situação atual também criou oportunidades para fortalecer a infraestrutura contra falhas futuras. Estamos vendo mais organizações planejando upgrades significativos na infraestrutura conforme preparam suas organizações para capitalizar sobre a recuperação econômica. As estratégias a seguir podem ajudar a garantir que esses upgrades entreguem a maior disponibilidade possível: redundância da infraestrutura, monitoramento da infraestrutura e gerenciamento remoto de TI, escalabilidade do UPS, baterias de íon-lítio e design de distribuição de energia.

Redundância da Infraestrutura

Avaliar redundância e as oportunidades para fortalecimento do sistema é um investimento que pode oferecer um retorno positivo ao reduzir a frequência dos eventos de downtime. O desafio é alcançar o nível certo de redundância de UPS da forma mais simples e mais eficiente possível. As necessidades de redundância devem ser consideradas sob a ótica da necessidade dos Acordos de Nível de Serviço (SLA). Pode haver a necessidade de aumentar a resiliência para 2N em alguns casos ou a oportunidade de reduzir para N em outros. A análise e o fortalecimento no nível do sistema podem também reduzir a vulnerabilidade com relação aos downtimes a partir de eventos relacionados com o UPS.

Em instalações maiores, arquiteturas de reserva estão sendo implementadas cada vez mais para reduzir os custos de capital e aumentar a eficiência dos sistemas UPS. Estas arquiteturas caem em duas principais categorias: reserva de bloco e reserva distribuída. Configurações com reserva de bloco implementam uma Chave Estática de Transferência (STS) e simplificam o gerenciamento da carga. Elas são em geral recomendadas quando os SLAs exigem alimentação para ambas as linhas. Arquiteturas de reserva distribuídas cada vez mais não

implementam uma STS e demandam uma maior atenção para o gerenciamento da carga de forma que não excedam os níveis de redundância. Elas podem ser usadas onde os SLAs exigem alimentação apenas para uma linha.

Tecnologias mais novas de UPS, como as empregadas pelo Vertiv™ Liebert® Trinergy™ Cube, usam redundância interna para eliminar a complexidade do design de sistemas UPS multimodulares. O UPS Liebert Trinergy Cube possibilita que as empresas modernizando seus data centers reduzam as despesas de capital e operacionais enquanto melhoram a disponibilidade. Ao usar uma configuração N+1 interna, esse UPS pode trocar a redundância do nível do sistema para o nível do módulo. Ao integrar diversas linhas de alimentação dentro do sistema, ele também proporciona melhor escalabilidade para arquiteturas 2N ou de reserva de alta disponibilidade.

Monitoramento de Infraestrutura e Gerenciamento Remoto de TI

Da telemedicina ao comércio eletrônico e ao home office, a pandemia acelerou a velocidade da transformação digital. O monitoramento da infraestrutura do data center e o gerenciamento remoto de TI são outros exemplos disso. Essas tecnologias não apenas estão ajudando as organizações a se adaptar às situações onde o acesso às instalações críticas é limitado devido as restrições da pandemia, mas também são ferramentas críticas para responder mais rapidamente a indisponibilidades e proteger contra a falha de equipamentos críticos.

Ao monitorar os sistemas de infraestrutura em tempo real, as organizações podem muitas vezes identificar precocemente sinais de alertas de falhas iminentes e tomar as medidas corretivas antes que a falha ocorra. Esses sistemas também coletam os dados necessários para aproveitar as análises preditivas e a transição para uma estratégia de manutenção proativa. Conjugar dados em tempo real com estratégias de serviço e de manutenção que correlacionem a manutenção com o tempo médio entre falhas (MTBF) permite a realização de serviços mais efetivos e mais eficazes nos equipamentos. Esses recursos são particularmente valiosos na medida em que proporcionam visibilidade em locais de edge remotos e simplificam o gerenciamento de múltiplos locais de edge.

Além disso, sistemas de monitoramento e gerenciamento de infraestrutura podem dar suporte a relatórios regulares sobre o estado do data center para garantir que servidores e outros equipamentos estejam operando em condições que não contribuirão para falhas. Eles também possibilitam modelagem, para garantir que uma nova capacidade tenha o suporte de alimentação e ambiental antes que ela seja implementada.

Figura 2: O Liebert® Trinergy™ Cube apresenta redundância interna e escalabilidade tridimensional.

Sistemas de gerenciamento remoto de TI, como consoles seriais e KVMs, reduzem a necessidade de interação física com os sistemas de TI ao mesmo tempo em que simplificam o gerenciamento, a identificação e resolução de problemas e a recuperação. Aproximadamente 80% das falhas dos equipamentos de TI são relacionadas ao software ou ao firmware. Nesses casos, engenheiros usando ferramentas de acesso remoto podem geralmente resolver a situação rápida e remotamente para minimizar a duração dos eventos de downtime.

Escalabilidade do UPS

A capacidade do UPS pode ser uma limitação na capacidade do data center e, quando eventos como a pandemia criam uma demanda inesperada que excede a capacidade do UPS, ela pode levar diretamente ao downtime.

Atualmente, há uma solução que permite que as organizações minimizem seu investimento de capital enquanto mantêm a flexibilidade para escalar o sistema UPS rapidamente. O anteriormente citado UPS Liebert Trinergy Cube apresenta um design modular, escalável a quente (funcionando) que permite a adição de nova capacidade sem desligar a unidade.

Esse sistema também redefine os limites da escalabilidade. Ele é escalável até 12,8 megawatts (MW) através de seu exclusivo design modular tridimensional. Verticalmente, as gavetas empilhadas em cada núcleo podem ser extraídas individualmente para manutenção enquanto o UPS continua a proteger a carga. Horizontalmente, o sistema pode ser escalado até 1,6 MW acrescentando-se quatro núcleos individuais de 400 quilowatts (kW) (como opção, um quinto núcleo de 400 kW de redundância). E ortogonalmente, até oito unidades de UPS Liebert Trinergy Cube de 1,6 MW podem operar em paralelo para dar suporte à carga de 12,8 MW.

Baterias de Íon-Lítio

Baterias tradicionais de chumbo-ácido são muitas vezes consideradas o elo fraco na cadeia de energia do data center, portanto, não é surpreendente que as baterias sejam a principal causa de downtime. Com strings e strings de baterias necessários para dar suporte às modernas instalações, pode sentir-se que uma falha está para acontecer a qualquer momento. Essas baterias tendem a ser de alta manutenção, pesadas e com necessidade frequente de trocas. Os avanços no monitoramento, gerenciamento e manutenção ajudaram a aliviar um pouco esses problemas, mas nem todos os data centers têm a vantagem de ter esses recursos.

Baterias de íon-lítio surgiram como uma alternativa viável para as baterias de chumbo-ácido e deveriam ser consideradas por operadores de data centers buscando limitar seus riscos de downtime. As baterias de íon-lítio têm uma vida útil consideravelmente maior do que as de chumbo-ácido, necessitando de menos manutenção e serviços. Algumas baterias de íon-lítio também provaram ter reduzido as necessidades de refrigeração, resultando em custos operacionais menores. Talvez o mais importante seja que, quando usadas com um sistema UPS, essas baterias usam um sistema de gerenciamento de baterias integrado para melhorar a operação e reduzir o risco de falhas no sistema e downtimes não planejados.

As baterias de íon-lítio têm, de fato, um custo inicial maior, mas a sua maior vida útil resulta em um menor custo total sobre a vida da bateria, mesmo sem levar em conta os custos de downtime.

Para as organizações que não estão em uma posição de fazer a transição para as baterias de íon-lítio, implementar uma solução de monitoramento de baterias para as baterias de chumbo-ácido proporciona a visibilidade no desempenho da bateria necessária para minimizar ou eliminar indisponibilidades devidas a falhas nas baterias.

Design da Distribuição de Energia

Há diversas opções para gerenciar a distribuição da alimentação de energia no data center, desde o uso de grandes unidades de distribuição centralizada até unidades de distribuição menores.

Na Vertiv, analisamos como os vários designs de sistemas de distribuição afetam as indisponibilidades do data center. Alguns operadores preferem uma mentalidade de “pequena falha” e implementaram unidades de STS no rack ao invés de uma STS maior centralizada. A Vertiv considera a maior STS como um possível ponto único de falha e fortaleceu a arquitetura da STS para incluir fontes de alimentação redundantes, lógica de transferência triplamente redundante e algoritmos de controle avançados, como a Transferência Otimizada, para limitar a inrush devida a magnetização dos transformadores da PDU. Isso resultou em um MTBF com uma ordem de magnitude maior do que o sistema UPS.

Investindo no Seu Futuro

Fazer as mudanças necessárias para minimizar os eventos de downtime requer uma mudança de uma abordagem reativa para uma abordagem proativa, na qual as infraestruturas críticas e as práticas para dar suporte a elas são avaliadas e investimentos são feitos para endereçar as causas raiz. Em vários casos, isso incluirá a substituição de equipamentos antigos por novos sistemas e a implementação de sistemas de gerenciamento e monitoramento remotos. Embora o investimento necessário possa ser percebido como significativo, ele deve ser colocado em perspectiva ao considerar-se os custos de downtime que a organização tem todos os anos.

Vertiv.com | **Vertiv América Latina,** 550 W. Cypress Creek Rd. Suíte 200, Fort Lauderdale, FL 33309, Estados Unidos da América

PT-SL-70705 (R01/21)