NUTANIX

O GUIA DEFINITIVO SOBRE

Computação do usuário final

Uma abordagem híbrida e multicloud para aplicações e desktops modernos

Índice

Autor
Brian Suhr tem mais de vinte anos de experiência com TI, em design, implementação e administração de infraestruturas corporativas. Já emprestou sua experiência em arquitetura e engenharia para diversos projetos baseados em nuvem, virtualização e data center, colaborando com equipes técnicas de alto desempenho em ambientes de alcance global. Como autor independente dos blogs DatacenterZombie e VirtualizeTips, Brian foca na criação de conteúdo sobre virtualização, automação, infraestrutura e divulgação de produtos e serviços que beneficiam a comunidade tecnológica. Siga Brian no Twitter: @bsuhr

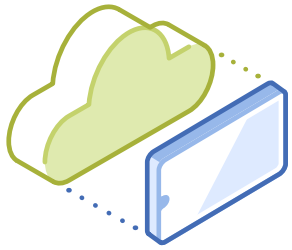
Sobre este eBook
O foco deste eBook é o design de infraestruturas para ambientes de VDI e Computação do Usuário Final (EUC). Seu conteúdo foi adaptado do capítulo sobre infraestrutura de "Architecting and Designing End User Computing Solutions" (disponível na Amazon).

Introdução

Após definir a estratégia e o fornecedor de software mais adequados para fornecer serviços e aplicações de EUC, as escolhas de implementação de arquitetura e infraestrutura são as próximas grandes decisões em projetos de virtualização de aplicações e desktops.

Escolher onde o plano de controle será executado e como será operado é cada vez mais importante, pois o número de opções cresce a cada dia. A infraestrutura é a base onde se constrói os serviços. Devemos poder contar com ela, assim como contamos com a eletricidade e a água quando apertamos um interruptor ou abrimos uma torneira. E também de forma parecida, a infraestrutura pode ser contratada como um serviço.

Sem um plano de controle e infraestrutura estáveis, altamente disponíveis e de alto desempenho, a TI pode enfrentar inúmeros desafios durante as fases de implementação e operação do seu projeto EUC. Embora a infraestrutura desempenhe um papel essencial, a TI não pode empenhar uma fatia significativa de seu tempo em implementação e manutenção, o que vem acontecendo com muita frequência. A infraestrutura certa deve simplesmente funcionar, liberando arquitetos e engenheiros para se concentrarem no fornecimento de serviços e aplicações de EUC.



Arquiteturas de Implementação da EUC

Identificar a arquitetura de implementação ideal depende dos requisitos da empresa, e essa decisão, por sua vez, influencia a escolha do plano de controle de EUC desta empresa. A seguir, temos os diferentes tipos de planos de controle e as opções de implementação para brokers de EUC. O plano de controle, como o nome sugere, lida com provisionamento, potência e intermediação, além de ser a interface primária para administradores. Geralmente também funciona como uma interface de API.

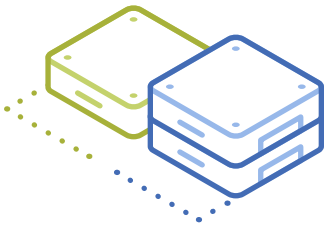
Os planos de controle de EUC tipicamente intermediam as conexões de um usuário com uma aplicação ou desktop. Para aplicações, isso é feito por meio da apresentação de aplicações (geralmente baseada em RDSH) e, para desktops, podem ser usados desktops compartilhados hospedados (HSD) ou uma sessão de infraestrutura de desktop virtual (VDI).

A seguir, temos as três principais alternativas e alguns exemplos de casos de uso.

Broker local tradicional

Com essa opção de broker de EUC, as empresas geralmente compram licenças por usuário e instalam em seu data center. Embora vários fornecedores e produtos se enquadrem nessa categoria, o Citrix Virtual Apps and Desktops (CVAD) e o VMware Horizon são, de longe, os mais usados.

Ao criar uma implementação de EUC no local, a responsabilidade pela arquitetura e pela implementação recai sobre a empresa ou sobre o parceiro, incluindo ter de definir todos os componentes do software intermediário a ser implantado, como servidores controladores, servidores de banco de dados, servidores de licenciamento, servidores ou dispositivos de segurança de borda e quaisquer outros serviços de suporte necessários. Como parte da implementação desses itens, você também define quais opções de alta disponibilidade (HA) estão disponíveis para cada serviço e seleciona uma que faça sentido para seus requisitos de design.



Geralmente, também há um aspecto de carga e escala para a maioria desses serviços. Você precisará dimensionar e arquitetar adequadamente quantas conexões cada serviço pode gerenciar para determinar o número adequado de servidores controladores necessários para, digamos, 5.000 conexões com HA. Em seguida, você pode determinar se implementará todos os servidores controladores para 5.000 conexões desde o início ou adicioná-los à medida que escalar sua implementação. Se você planeja escalar além desse número, você pode usar esses detalhes para continuar dimensionando os diferentes serviços e manter a conformidade com as práticas recomendadas e os máximos suportados.

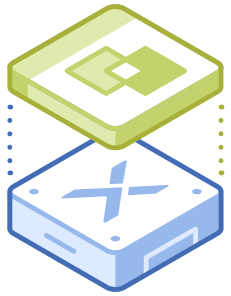
O próximo fator a ser considerado são as operações, nas quais é necessário entender como corrigir e atualizar todos os serviços no broker de EUC implementado. Além dos serviços dos quais já falamos (normalmente parte de todas as VMs de servidor implementadas), também há agentes e clientes que você precisa manter. Os agentes estão presentes nas imagens usadas para implementar os pools de servidores de aplicações e desktops aos quais os usuários se conectam. A forma como você atualiza os agentes depende de como você provisiona esses pools. Os clientes estão localizados nos endpoints utilizados pelos usuários para se conectarem a esses serviços, que variam muito dependendo do estilo do endpoint.

Apesar do nome, você pode implementar um broker local tradicional em uma nuvem pública. Geralmente faz mais sentido implementá-lo no local, daí o nome.

Cloud Broker

Os cloud brokers são oferecidos como serviço no modelo comum de software como serviço (SaaS) e geralmente são chamados de ofertas de desktop como serviço (DaaS). Vários fornecedores e produtos se enquadram nessa categoria. Os provedores de nuvem pública têm suas ofertas, no entanto, como a Citrix é líder absoluta no mercado de EUC, o Citrix Virtual Apps and Desktops Service (CVADS) é o lugar ideal para começar a avaliar os benefícios do DaaS. A maioria das ofertas oferece um conjunto de recursos semelhantes aos fornecidos por seus correspondentes locais tradicionais, mas isso não quer dizer que não tenham seus próprios benefícios e limitações.

Existem diversas ofertas diferentes de DaaS disponíveis de fornecedores de software, provedores de nuvem e provedores de serviços. Elas variam muito quanto ao que oferecem quando olhamos além dos grandes recursos, como VDI e apresentação de aplicações, e é por isso que é importante entender seus casos de uso e seus requisitos ao tomar uma decisão.



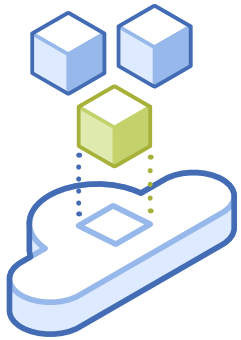
Os requisitos são seus melhores parâmetros; se você gosta de uma opção que não atende a alguns dos requisitos, você precisa decidir se pode viver sem eles até que essa lacuna seja preenchida, se um dia ela for.

Como um serviço que você assina, elas têm várias características interessantes. Primeiro, você normalmente paga por usuário e por mês, o que torna o custo fácil de calcular e monitorar. O custo também vem do orçamento de OpEx, o que para alguns é um benefício, já que são forçados a mudar para o modelo de consumo de nuvem. Você pode assinar acordos de longo prazo com essas ofertas de DaaS, que geralmente têm preços melhores a troco de um compromisso mais longo. Se você se inscrever, no entanto, você tem a opção de sair caso sua demanda diminua ou deixe de existir. Com a abordagem de licença tradicional, as empresas normalmente compram licenças e pagam pelo suporte eternamente.

Como é um serviço, as empresas podem consumir sob demanda. Todos os aspectos de design, implementação e dimensionamento da camada de intermediação são tratados para você. Além disso, você geralmente executa apps e desktops na nuvem pública junto à oferta de DaaS, para que não precise gerenciar a infraestrutura em que eles estão sendo executados, o que, por sua vez, o impede de monitorar e fazer o upgrade dessas camadas. Essas ofertas de DaaS podem fazer seu projeto decolar mais rapidamente e com menos esforço do lado operacional.

Embora o serviço de DaaS cubra as camadas de intermediação e infraestrutura, ainda há tarefas operacionais a serem executadas. Essas tarefas incluem criação de imagens e atualizações, instalações e atualizações de aplicações, dados do usuário, VPNs e assim por diante.

Outro benefício oferecido pela nuvem é a capacidade de usar data centers em todo o mundo para oferecer pools de aplicações ou desktops próximos a um grupo de usuários. Esse benefício nem sempre funciona. Por exemplo, se os usuários dependem muito de dados ou de uma aplicação específica em um data center distante, essa capacidade é nula.



Implementações híbridas

Agora que já falamos sobre as diferentes arquiteturas para brokers de EUC, vejamos como estão sendo implementadas. Contrariando algumas previsões de que todos os workloads irão para a nuvem pública, é notório que a maioria das implementações ainda está no local e a maioria das empresas caminha rumo à arquitetura híbrida. Para os propósitos desta discussão, híbrido se refere simplesmente a uma combinação de nuvem pública e nuvem local. As porcentagens de cada uma variam de acordo com vários fatores. Quando falamos de híbrido no mundo da EUC, há formas de funcionar tanto para ofertas tradicionais quanto para DaaS. Primeiro, vejamos como cada uma das arquiteturas se encaixam em um mundo híbrido e depois mergulharemos em alguns exemplos de casos de uso.

A solução de DaaS é a abordagem mais comum para ofertas híbridas, e nem todos os fornecedores de DaaS oferecem a capacidade de implementação em uma arquitetura híbrida. O plano de controle geralmente permanece na nuvem para a versão de DaaS do híbrido e aqueles que o suportam contam com um método para controlar e gerenciar os pools de recursos privados, geralmente implementando uma VM local em seus clusters locais para atuar como um conector local. Através dessas VMs do conector de nuvem, o provedor de DaaS agora pode provisionar VMs, controlar seus estados de energia e intermediar conexões com elas. As VMs do conector são fáceis de configurar e você pode implementá-las em pares para oferecer alta disponibilidade. Essa é a arquitetura que o Citrix CVADS, o Horizon Cloud e o Nutanix Frame suportam.

No geral, a versão de DaaS do híbrido ainda é muito menos complexa de se projetar, implementar e operar do que a abordagem tradicional. Você ainda é o responsável por gerenciar a infraestrutura onde as máquinas virtuais locais são executadas, mas não precisará gerenciar a camada de intermediação.

Casos de uso

Vamos falar sobre os diferentes cenários e casos de uso adequados para o híbrido.Se usado corretamente, o design híbrido pode oferecer grande flexibilidade em termos de custo e recursos. A maioria dos casos de uso se enquadra em três cenários diferentes.

- **Recuperação de desastres (DR).** Esse cenário provavelmente se encaixa na maioria das implementações tradicionais locais que requerem continuidade de negócios (BC). Historicamente, a DR envolvia criar ou alugar um local secundário e usá-lo para implementar recursos de EUC para failover. A nuvem pública oferece uma alternativa muito atraente à abordagem legada, no sentido de que você pode reservar certa capacidade de nuvem e, depois, expandi-la para a capacidade total caso um desastre ocorra. Nessa abordagem, você paga para executar suas VMs de infraestrutura de broker e um pequeno pool de desktops constantemente com capacidade de armazenamento para os dados e perfis de usuário replicados. Então, se ocorrer um desastre, você pode expandir rapidamente esse pool de desktops para o tamanho que precisar e direcionar seus usuários, para que possam voltar a trabalhar. Quando tudo estiver corrigido, você tem como voltar ao tamanho estacionário para reduzir os custos.

- **Bursting.** Existem alguns casos de uso ou projetos que só precisam de recursos por um curto período, e manter espaço no local para eles pode não fazer muito sentido. Nesse caso, você ativa os recursos, normalmente em uma nuvem pública, e os desativa quando a demanda cai. Há diversos cenários onde podemos usar o bursting de forma eficaz. Os laboratórios estudantis estão entre os casos de uso mais comuns, pois têm necessidades sazonais e programadas de acordo com o calendário do aluno. Os alunos também podem ter necessidades de GPU que você não possui no local. Projetos especiais e trabalho temporário também são casos de uso muito comuns.

- **Híbrido de verdade.** Essa opção provavelmente reflete a maioria dos outros casos de uso no sentido de que não são necessidades temporárias, o que significa que você costuma executar alguns workloads no local e outros na nuvem e decide onde implementar com base em parâmetros que fazem sentido para sua empresa e seu design. Por exemplo, você pode ter um caso de uso em que precisa de GPUs e não as tem no local; você pode precisar implementar na região APAC em vez de fazer com que os usuários locais voltem para seus sites na América do Norte; ou você pode simplesmente estar sem capacidade em seu ambiente local.

Princípios Arquitetônicos

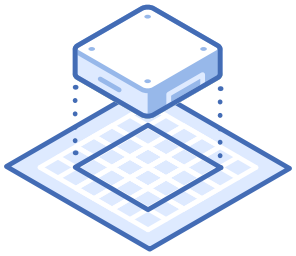
Existem vários fatores importantes que devem ser considerados no processo de design de infraestrutura da EUC. Ao utilizar esses fatores alinhados aos requisitos da empresa, é possível considerar melhor quais serão as alternativas de arquitetura. Os seguintes fatores devem ser considerados ao avaliar alternativas de arquitetura e opções de fornecedores em projetos de EUC:

- Ponto de entrada
- Escalabilidade
- Desempenho
- Monitoramento
- Capacidade

Ponto de entrada

O ponto de entrada ou partida para a infraestrutura muitas vezes pode ser a principal decisão de um projeto. É a quantidade de infraestrutura e o custo necessários para uma empresa iniciar a implementação de entrega e virtualização de aplicações ou desktops com base em diferentes tamanhos de ponto de partida.

Se o planejamento do projeto é atingir 10.000 usuários no fim da implementação total, com a fase de implementação inicial de 5.000 usuários, é menos provável que a empresa fique chocada com os custos iniciais. O raciocínio é que, dependendo do tipo de infraestrutura escolhida, o custo por usuário pode não fazer muito sentido até que você tenha implementado alguns milhares de usuários.



O outro lado é se uma empresa pretende implementar 10.000 usuários, mas quer começar apenas com 500 usuários e ir aumentando em um ritmo constante ao longo do cronograma do projeto. Eles analisarão melhor o custo da implementação inicial da infraestrutura nesse tamanho em vez de dar um primeiro passo maior. O custo por usuário nesse tamanho pode manter-se estável à medida que o ambiente aumenta, ou pode parecer realmente distorcido no início, devido a um gasto inicial maior com infraestrutura.

Embora o custo por usuário possa ser visto como vago e quase irrelevante como um fator para determinar seus custos de infraestrutura, você será questionado sobre isso ao tentar vender o projeto para a empresa ou justificar sua seleção de infraestrutura para a liderança. Se você escolher uma alternativa com custo inicial maior para o usuário, precisará estar preparado para explicar os detalhes. Analise as soluções que você acredita serem mais adequadas para o seu ambiente. Caso contrário, esteja preparado para tomar decisões sobre como os custos serão executados. Na Figura 1 temos um exemplo desses dois cenários.

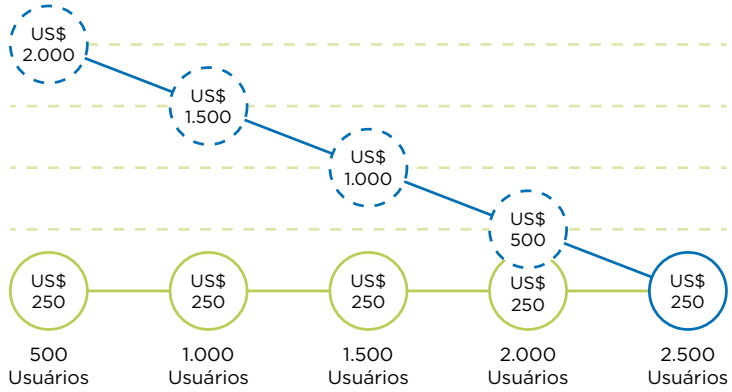


Figura 1:
Pontos de entrada por desktop

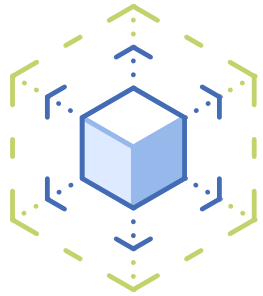
Escalabilidade

A escalabilidade da arquitetura é um fator importante ao avaliar a viabilidade de um projeto. Um arquiteto precisará entender as opções iniciais de tamanho para as diferentes alternativas; o que remete ao tópico do ponto de entrada que acabou de ser abordado. Essa alternativa permitirá iniciar o design em tamanho menor ou a empresa terá de comprar mais infraestrutura do que seria necessário para começar, por conta da dimensão inicial de um projeto—sem poder utilizar todo esse recurso até que o projeto alcance esse tamanho?

Além do tamanho mínimo com que é possível começar, é igualmente importante considerar o quanto se pode expandir. Se a ideia é começar com 500 e ainda ser capaz de escalar para 10.000 usuários, como a alternativa se comportará em cada um desses cenários? A empresa ficará satisfeita com os pontos altos ou baixos—ou ambos?

O tópico sobre escalabilidade não é só uma discussão sobre armazenamento. Isso também se aplica a computação, rede e possivelmente outras camadas dentro do design. Se forem feitos ajustes na configuração da camada de computação para conseguir uma densidade menor de VMs por servidor de host, como isso pode afetar diferentes opções de design ao dimensionar? Um exemplo: se o design inicial do host começasse com 128 GB de memória por host e a opção final fosse 256 GB ou mais, seria necessário garantir que os DIMMs de tamanho correto fossem usados para permitir que a configuração fosse dimensionada no futuro. Se as escolhas erradas forem feitas para cortar custos, isso afetará a densidade devido a restrições ou custará mais a longo prazo com DIMMs que não puderam ser reutilizados.

O arquiteto deve se concentrar em como a solução poderá começar pequena e ser capaz de escalar até o ponto máximo. Mas também não podemos ignorar todos os pontos intermediários, pois dependendo de como alguém dimensiona a implementação, pode haver muitos pontos de escala entre o início e o fim. O ideal é procurar uma solução que permita ao projeto escalar facilmente em grupos de usuários de acordo com o projeto, sem ultrapassar o cronograma e os recursos da implementação. O tamanho de escala ideal para um projeto pode estar em incrementos de 500 a 1.000 usuários. Mas se a alternativa de arquitetura escolhida pode escalar ainda mais, entenda como isso afeta os custos e a implementação.



Desempenho

O desempenho da EUC avaliado pela experiência do usuário final é sempre uma consideração importante. A arquitetura escolhida deve ser capaz de atender aos requisitos em qualquer fase do projeto. Esse pode ser um caminho difícil de se percorrer com algumas alternativas. Se alguém reduzir uma solução para atender aos requisitos mínimos de usuário inicial, pode acabar sacrificando o desempenho se não conseguir escalar de forma linear. Os arquitetos não querem fazer concessões na arquitetura para alcançar esse pequeno ponto de entrada que pode afetar as opções gerais de desempenho máximo de uma solução. Se você investir tempo tomando a decisão certa no início, poderá evitar problemas futuros.

Um projeto de solução de EUC geralmente é composto por vários requisitos diferentes de desempenho. Escolha uma opção de arquitetura flexível o suficiente para atender a todos os requisitos de desempenho em uma única opção. Se o design fornecerá vários tipos de serviços de EUC ou focará apenas em virtualização de apps e desktops, devemos considerar várias necessidades de desempenho. Entender como cada opção será ou não capaz de atender aos requisitos de desempenho individuais afetará bastante seu processo de design e avaliação.

Capacidade

A discussão sobre capacidade é semelhante à de desempenho. Existem vários requisitos diferentes de capacidade nos projetos de EUC que precisarão ser oferecidos. A solução exigirá a execução de VMs de servidor, VMs de desktop, aplicações, perfis de usuário e dados do usuário para esse tipo de arquitetura. Cada camada dentro do projeto pode ter requisitos de capacidade bem diferentes. Alguns usam grandes quantidades de dados que normalmente desduplicam bem. Outras, como perfis de usuário e dados, são compostas por quantidades menores de dados por usuário, mas, multiplicadas por milhares de usuários, acabam representando uma grande quantidade no final.

Um grande problema nos últimos anos foi a compra de muita ou pouca capacidade, enquanto tentava-se atingir os níveis de desempenho exigidos. Observe atentamente as opções de arquitetura durante a fase de projeto e veja se elas serão capazes de oferecer a capacidade necessária e, ao mesmo tempo, garantir que os requisitos mínimos de desempenho também sejam atendidos. A solução não deve oferecer capacidade de 2, 3 vezes ou mais para atender aos requisitos de desempenho de armazenamento ou adicionar um desempenho extra significativo para atender aos requisitos de capacidade.

A solução ideal é aquela que oferece flexibilidade suficiente para dimensionar o desempenho e a capacidade em taxas semelhantes, de modo que nenhum fique muito fora do ritmo do outro.

No passado, esse assunto já causou muitos problemas. Muitas empresas se depararam com problemas de planejamento de desempenho e capacidade ao dimensionar a capacidade mais rápido que o desempenho. Só porque a solução possui 5 TB de espaço livre não significa que ela seja capaz de ser dimensionada para outros 500 usuários. Esse cenário pode afetar muito o desempenho. Administradores e líderes de TI sem um conhecimento sólido de como a solução é dimensionada podem acabar caindo nessa armadilha.

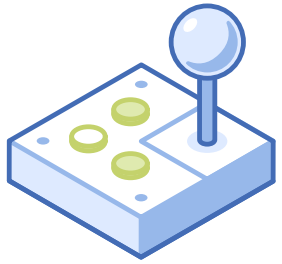
Monitoramento

O monitoramento é muito importante, mas costuma ser negligenciado. Quando se trata de monitorar a infraestrutura em um ambiente de EUC, os administradores geralmente focam no desempenho. Eles precisam entender o que é normal e quando há um problema ocorrendo.

O uso do monitoramento deve ser simples, ao mesmo tempo que oferece uma riqueza de informações detalhadas. Esse não é o caso de muitos fabricantes, portanto, devemos observar atentamente a experiência de monitoramento de cada solução.

Outro requisito é a capacidade de oferecer monitoramento do desempenho a nível de máquina virtual. Infelizmente, a maioria dos fornecedores de infraestrutura ainda oferecem esse nível de visibilidade do ambiente de virtualização. O melhor monitoramento de desempenho de infraestrutura da categoria deve oferecer aos administradores a capacidade de analisar rapidamente a camada de armazenamento e determinar se o problema de desempenho do armazenamento é global ou se é limitado a um host, grupo de VMs ou uma única VM.

Ao gerenciar o desempenho do armazenamento no nível da VM, é possível usar uma abordagem semelhante para gerenciar o desempenho da CPU e da memória de uma VM no nível do host. Os administradores precisam saber se uma VM está usando temporariamente desempenho extra ou se é um consumidor regular de desempenho extra de armazenamento em comparação aos usuários padrão. Isso permitirá entender quando há um pico e quando investigar mais para identificar o problema.



Blocos de construção

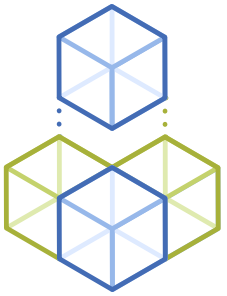
Um bloco de construção é um conjunto pré-definido de infraestrutura alocado a uma quantidade específica de recursos ou número de usuários. Essa abordagem é uma das melhores formas de abordar o design da infraestrutura com a computação de usuário final.

Com essa abordagem, é possível desenvolver uma arquitetura que ofereça um modelo previsível de dimensionamento de capacidade, desempenho e custos. Ao determinar o tamanho do bloco de construção, escolha quais incrementos você precisa para dimensionar os usuários e como a infraestrutura pode adaptar-se às opções. Por exemplo, pode ser que você queira escalar usuários em incrementos de 50 a 100, mas a infraestrutura escolhida não escala bem com pequenos incrementos como este. Isso pode forçar o projeto a ser dimensionado em incrementos maiores, de 500 ou 1.000 usuários. Se a infraestrutura for dimensionada em blocos grandes, você pode optar por escalar para combinar com isso ou simplesmente aceitar o fato de que os custos de infraestrutura não serão dimensionados da mesma forma que os blocos de implementação do usuário. Isso quer dizer apenas que a empresa compraria infraestrutura em blocos de 1.000 usuários e implementaria apenas em grupos de 50 a 100 usuários.

Isso faz com que os custos dos desktops virtuais ou das sessões do usuário pareçam caros, ao comprar o bloco maior para implementar uma quantidade menor de usuários. Isso se equilibra caso a empresa implemente todos os usuários planejados.

As arquiteturas do estilo bloco de construção são úteis para qualquer projeto de design, mas as implementações de EUC sempre têm partes comuns de usuários e casos de uso que possuem características semelhantes e são implementados em grupos. Seguindo no exemplo de bloco com tamanho de 100 usuários, ao entender os requisitos de recursos para 100 usuários, podemos garantir que o bloco de infraestrutura seja capaz de fornecer tudo o que esses usuários precisam.

Se cada usuário precisa de 15 IOPS em estado estacionário e 30 GB de capacidade de armazenamento, junto com 2 GB de memória e 200 MHZ de CPU, então o arquiteto sabe que os blocos de construção devem fornecer 1.500 IOPS, 3 TB de capacidade, 200 GB de memória e 20 GHZ de CPU. O arquiteto pode projetar os blocos de construção para conter



recursos adicionais, mas nenhum deles poderá estar abaixo desses valores. Também queremos evitar o desperdício de incluir em cada bloco muito mais do que podemos utilizar.

Com essa abordagem e granularidade no design, agora é possível dimensionar o ambiente em grupos de 100 usuários. Isso permite uma abordagem lenta e constante e fornece valores previsíveis que as empresas podem usar para planejar a implementantação, o desempenho, a capacidade e os custos. Se as empresas quiserem escalar mais rapidamente e em maiores quantidades, basta elas colocarem vários blocos de construção de uma só vez.

Por fim, a abordagem de blocos de construção provou ser interessante, pois a maioria dos clientes gosta de começar com implementações pequenas e aumentar a partir daí. O modelo "comece pequeno e pague à medida que crescer" permite que as empresas invistam uma menor quantia de capital no início e ganhem experiência à medida que suas implementações crescem. Na próxima seção, falaremos sobre os diferentes tipos de arquiteturas de infraestrutura disponíveis atualmente e como cada uma delas suporta ou não a abordagem de blocos de construção.

Hipervisores

O hipervisor é uma camada importante no design da sua infraestrutura. Ele é diretamente responsável por uma boa parte do desempenho, da disponibilidade, da resiliência e da capacidade de gerenciamento da sua solução.

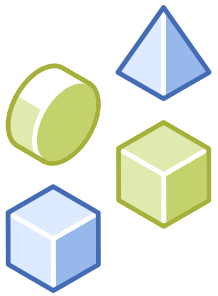
Cenário de hipervisor

No ambiente de virtualização atual, há uma lista sólida com opções de hipervisores prontos para empresas. Essa lista se estreita quando analisamos os hipervisores que rotineiramente têm casos de uso de EUC e VDI implementados, que são os seguintes:

- Citrix Hypervisor (XenServer)
- Nutanix AHV
- VMware vSphere (ESXi)
- Microsoft Hyper-V

Questões que devem ser consideradas

Há vários motivos para as empresas considerarem mudar seu hipervisor, desde a simplificação da arquitetura e das operações da camada do hipervisor, até o reforço da segurança, a redução do aprisionamento tecnológico e a economia de custos.



Critério de avaliação

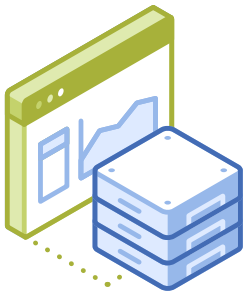
Os hipervisores são peças complexas de software compostas por centenas de recursos possíveis. Recursos básicos incluem agendamento de recursos, alta disponibilidade, redes virtuais e compatibilidade com vários sistemas operacionais. Hoje, esses são recursos prontamente disponíveis em todas as opções de hipervisor citadas acima.

Além disso, as empresas devem focar no valor que cada uma das soluções oferecem aos seus negócios e como isso pode mudar ou melhorar suas operações.

Um bom ponto de partida é examinar a fase de design para as diferentes opções de hipervisor disponíveis. Use os requisitos de design para um projeto recente ou futuro, como a solução para a qual você estaria projetando. Desenvolva um entendimento sobre o esforço necessário para criar uma solução para o projeto selecionado. Observe se essa fase requer dias ou semanas de esforço para projetar a camada de hipervisor da solução devido à complexidade das opções de design. O ideal é que o hipervisor ofereça todos os recursos e funcionalidades necessários, além de simplificar a fase de design, reduzindo a algumas poucas opções claras e simples.

Em seguida, trate de entender como é o esforço de implementação. Com base no projeto proposto escolhido na sua avaliação, identifique o esforço necessário para implementar a solução e como ele difere, se necessário, de uma implementação básica. Idealmente, assim como na fase de design, a fase de implementação deve exigir o mínimo de intervenção do engenheiro e ser altamente automatizada. Implementar um novo cluster não é algo que deve levar dias ou semanas.

E, por fim, examinar melhor o lado operacional de um hipervisor ajudará a entender qualquer diferença em relação aos seus esforços atuais. Historicamente, a aplicação de correções e upgrades em hipervisores tem sido um dos maiores desafios operacionais. Alguma opção de hipervisor oferece algum tipo de vantagem ao simplificar esses processos tanto em temos de esforço quanto de confiabilidade dos upgrades? Além dos upgrades, como é o gerenciamento diário de máquinas virtuais? Isso pode ser feito através de uma simples interface única?



Infraestrutura

Opções de infraestrutura

Atualmente, existem três opções principais de arquitetura para virtualização de aplicações e desktops, ou soluções de EUC. As alternativas são construir sua própria infraestrutura (BYO), infraestrutura convergente (CI) e infraestrutura hiperconvergente (HCI).

Construa sua própria infraestrutura

A opção BYO é exatamente o que seu nome implica, o arquiteto ou a equipe escolhem independentemente seus produtos preferidos ou que consideram ser os top de linha. Essa opção resulta em um aumento significativo no período inicial de pesquisa e planejamento, pois a equipe precisa avaliar cada produto separadamente e de que modo eles podem ou não operar juntos.

Essa opção também permite selecionar e seguir uma arquitetura de referência publicada por um fornecedor para o tipo de solução que está sendo criada. Essas arquiteturas de referência geralmente são publicadas por um único fornecedor e o foco é o seu produto. Essas arquiteturas de referência do tipo faça você mesmo (DIY) podem economizar tempo e reduzir alguns riscos, mas nem sempre estão alinhadas com requisitos de design, casos de uso e ambiente.

No mínimo, uma alternativa BYO para um projeto baseado em EUC conterá recursos de computação e armazenamento. Você pode usar a conectividade da rede existente, portanto, ela pode nem ser um componente dessa opção. A figura 2 é um exemplo simples das partes de uma opção de BYO. Com flexibilidade no dimensionamento, os custos são bastante previsíveis; a única exceção seria quanto ao armazenamento. Dependendo do tamanho máximo do seu design e da escolha do armazenamento, você pode precisar de vários dispositivos ou matrizes de armazenamento. Conforme você dimensiona o armazenamento e precisa adicionar um novo dispositivo ou matriz, o custo aumenta nesses pontos.

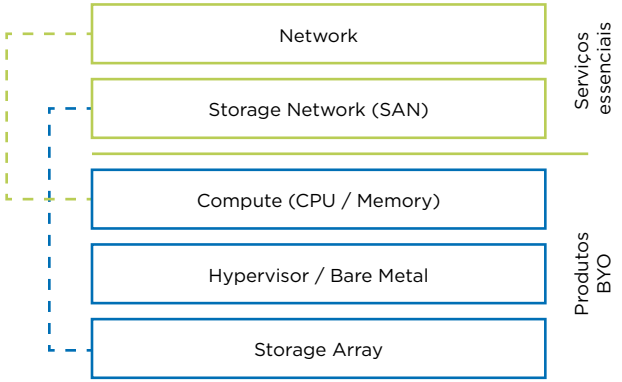
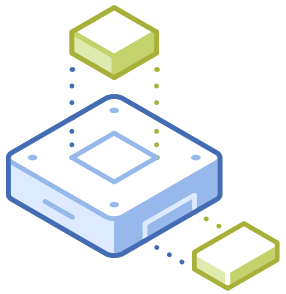


Figura 2:
Traga sua própria (BYO) infraestrutura

Sempre que montamos diversos produtos do mesmo fornecedor ou de vários fornecedores diferentes sem experiência prévia, há certos riscos envolvidos. Haverá um nível de incerteza sobre o desempenho e a confiabilidade da solução até que a infraestrutura seja, de fato, comprada e implementada na arquitetura.

Se você estiver disposto a aceitar as incertezas e o risco adicional, o BYO maximiza a flexibilidade, de fato. Como você pode tomar praticamente qualquer decisão sobre fornecedor e produtos que funcionam juntos, isso permite seguir com os fornecedores de sempre, enquanto migra para novos fornecedores em outras áreas.

O BYO é capaz de dimensionar os recursos de computação e armazenamento de forma independente. O único limite para o método de dimensionamento ou o tamanho máximo seria uma restrição na escolha do produto individual. Como os produtos são adquiridos separadamente, não há valores mínimos ou definidos para o dimensionamento desses produtos. Isso permite flexibilidade na abordagem de blocos de construção citada anteriormente.



Infraestrutura convergente

A opção de infraestrutura convergente (CI) é uma arquitetura que foi lançada no mercado por volta de 2010. As ofertas de infraestrutura convergente geralmente oferecem os mesmos produtos que podem ser selecionados como parte da BYO e os agrupam em uma solução produzida. Isso significa que um fornecedor de CI incluirá computação, armazenamento e rede em sua oferta. Geralmente, a maioria das ofertas de CI contém produtos de vários fornecedores e pode ser incluída como parte de uma única oferta, ou um fornecedor pode oferecer todas as camadas de uma oferta de CI a partir de sua própria linha de produtos. A figura 3 ilustra um exemplo simples de uma opção de infraestrutura convergente.

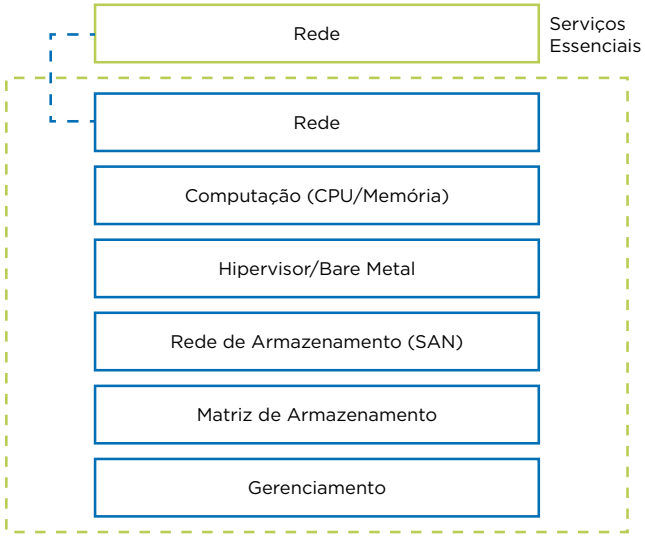
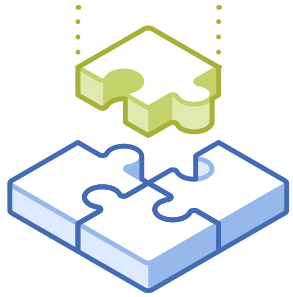


Figura 3:
Infraestrutura convergente

Uma oferta de infraestrutura convergente permitirá que você compre produtos que já conhece e foram agrupados em uma única solução. Podemos considerar essa oferta uma arquitetura de referência, que pode ser comprada como um produto. Dependendo do produto de CI avaliado, ele pode ou não oferecer alguma convergência adicional em relação à alternativa de comprar os produtos separadamente em uma opção de BYO.



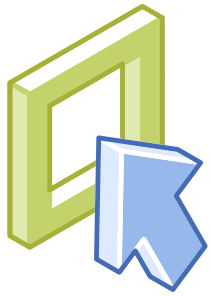
Geralmente, a maioria dos fornecedores e produtos de CI permitirá a compra de todas as partes da infraestrutura em um único SKU de produto. O fornecedor de CI deve oferecer suporte via chamada única para toda a solução de CI, o que significa que ele é capaz de suportar todos os produtos da solução. Esse é um benefício adicional, pois permite aos clientes eliminar a necessidade de lidar com vários fornecedores no processo de solução de problemas.

A maioria das ofertas de CI oferecem um número limitado de produtos dentro da solução. Isso permite que o fornecedor de CI faça um pré-teste e valide todas as partes e peças para garantir que funcionem corretamente juntas, eliminando grande parte do risco da opção de BYO.

Mesmo após vários anos no mercado, pouco foi feito pelos fornecedores de CI para simplificar o gerenciamento desses produtos. Com ofertas de CI que incluem os mesmos produtos que as opções de BYO, normalmente é possível gerenciar ambas de maneira semelhante e dispersa. Essa opção pode convergir a compra e/ou alguns dos produtos, mas geralmente não converge o gerenciamento operacional diário da solução.

Um produto de infraestrutura convergente deve ser capaz de dimensionar os recursos inclusos sem que um dependa do outro. Isso significaria que você pode simplesmente adicionar computação, embora possa haver incrementos mínimos para dimensionamento. O outro recurso que seria dimensionado em uma oferta de CI é o armazenamento, e dependerá muito do tipo de solução de armazenamento escolhido como parte da oferta de CI. Um produto de infraestrutura convergente terá um tamanho máximo, o que significa que haverá um limite no número de servidores que ela poderá suportar e um limite de armazenamento baseado na matriz de armazenamento inclusa.

Os limites de dimensionamento de uma oferta de CI geralmente são bastante grandes. No entanto, em algum momento, à medida que os recursos dentro do produto de CI escalam, eles atingirão os limites máximos. Para continuar dimensionando o design a partir deste ponto, será necessário comprar um produto de CI adicional. Isso causará grandes picos de custos de infraestrutura em diferentes pontos do processo de dimensionamento, dependendo do tamanho máximo do seu projeto.



Infraestrutura hiperconvergente

A arquitetura hiperconvergente chegou ao mercado aproximadamente um ano após a CI. As arquiteturas hiperconvergentes de verdade são obtidas convergindo os recursos de computação, os recursos de armazenamento e a camada de gerenciamento em um único produto.É possível implementar uma solução hiperconvergente em um modelo somente de software (SWO) ou em um modelo de equipamento que inclua hardware específico.

Ao incluir um dispositivo de hardware como parte do produto, o fornecedor agora pode incluir o gerenciamento da infraestrutura junto com os outros recursos que estão sendo convergidos no produto. A Figura 4 ilustra um exemplo simples de uma infraestrutura hiperconvergente.

A arquitetura somente de software pode oferecer bastante flexibilidade na camada de hardware, permitindo que você escolha a plataforma em que deseja implementar. Ao considerar as opções de SWO, elas ainda devem oferecer uma experiência semelhante a de um dispositivo. Isso é diferente das ofertas com uma lista de compatibilidade de hardware fracamente acoplada, que incorpora apenas testes mínimos.

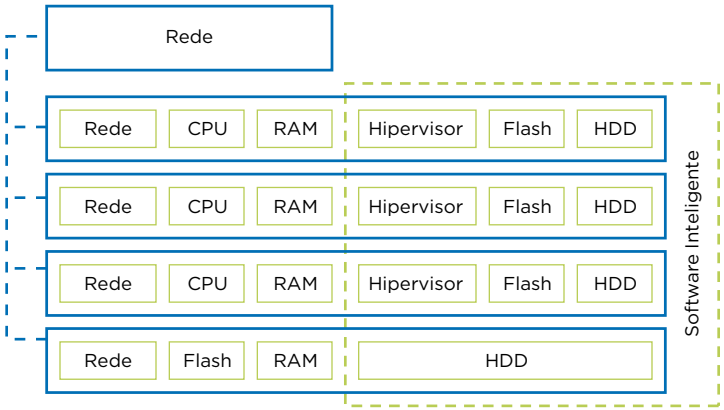
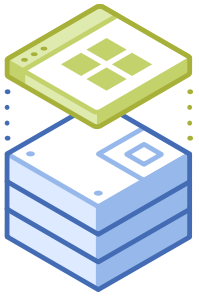


Figura 4:
infraestrutura hiperconvergente



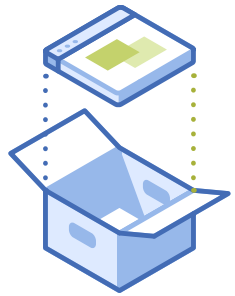
Instalação simples – Os principais produtos de HCI devem instalar nós em questão de minutos e horas, não em dias ou semanas, utilizando processos altamente automatizados.

Escalabilidade fácil – O produto deve ser fácil de escalar ou reduzir. A inserção de novos nós no ambiente deve ocorrer de forma fácil e rápida através da interface de gerenciamento.

Gerenciamento moderno – Uma interface de gerenciamento moderna deve focar na máquina virtual (VM) como ponto de gerenciamento. O administrador deve ser capaz de ver o desempenho das VMs, a quantidade de recursos que cada VM está consumindo, se há alguma falha ou erro, além de poder exportar relatórios facilmente com base nas VMs.

Extensibilidade – Você deve ser capaz de integrar a infraestrutura com outras partes da solução de forma fácil e controlá-la por meio de programação. Isso exige que o produto de HCI ofereça uma API e, possivelmente, outro método, como os cmdlets da PowerShell. Com uma API, você poderá automatizar a comunicação e o controle dos produtos para reduzir ainda mais o esforço e aumentar a precisão do ambiente.

O desempenho foi intencionalmente deixado de fora da lista de benefícios da HCI, pois todos esperam que uma solução híbrida moderna ou baseada em flash tenha um bom desempenho. O objetivo da HCI é criar uma camada de infraestrutura simples e eficiente. Ela permite que as equipes parem de perder tempo com tarefas repetitivas e permite que elas gerem mais valor aos negócios a nível de automação ou aplicação.



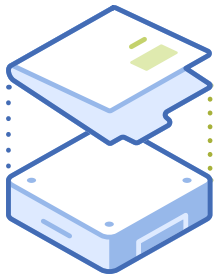
Requisitos de armazenamento

Há inúmeros requisitos de recursos de armazenamento diferentes em todo projeto de EUC. Eles devem levar em conta as VMs baseadas em servidor, os dados do usuário e a infraestrutura de desktop virtual (VDI). Os requisitos de armazenamento de VDI são os que mais exigem do ambiente, e também os que fazem com que os projetos de VDI falhem ou forneçam uma experiência ruim.

Por esse motivo, a parte relacionada a armazenamento neste eBook foca nas necessidades do serviço de VDI da solução. As necessidades de cada desktop virtual muitas vezes podem parecer pequenas e insignificantes, mas quando você as combina em grandes grupos conforme escala o armazenamento, as demandas de desempenho podem facilmente sobrecarregar o armazenamento que não foi projetado adequadamente para atender a essas necessidades.

Se a média de cada desktop virtual for 15 IOPS e são esperados 2.000 usuários simultâneos, isso equivale a 30.000 IOPS. Esse número é muito grande e pode sobrecarregar uma matriz de armazenamento padrão. Mas não se pode simplesmente projetar a solução de armazenamento para atender à E/S média do ambiente; o design deve levar em conta os picos, incluindo inicializações de desktop e eventos de login do usuário.

Os workloads de desktops virtuais apresentam muitos picos de E/S, o que os difere muito de outros workloads dentro do data center corporativo tradicional. Por exemplo, abrir uma aplicação como o Outlook pela primeira vez em uma sessão pode gerar mais de 1.000 IOPS para essa sessão de usuário. É muito além da média de 15 IOPS citada anteriormente. Um exemplo de impacto de aplicação IOP diferente é mostrado na Figura 5.



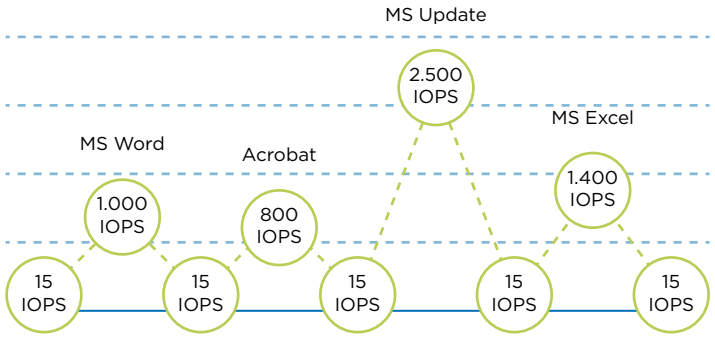
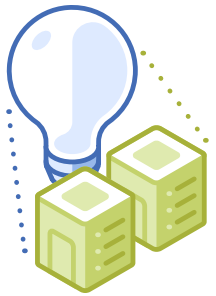


Figura 5:
IOPS de VDI

Outros itens operacionais e de implementação, como correções e atualizações do ambiente, também podem criar enormes picos nos IOPS e afetar o desempenho se não forem considerados e planejados adequadamente. Se implantarmos mais 50 desktops virtuais, essa ação pode criar um pico significativo de E/S. Por esses motivos, a arquitetura de armazenamento deve ser projetada para suportar picos de IOPS das operações de manutenção.

Há várias formas de arquitetar soluções de VDI com clones completos ou apresentação de imagens compartilhadas e cada uma pode gerar efeitos diferentes nos requisitos de armazenamento em termos de capacidade e desempenho. Como os clones completos consomem capacidade e armazenamento extras, a desduplicação será importante. Os clones completos também devem ser corrigidos de forma independente, o que aumentará a E/S durante essas operações.

A abordagem com imagem compartilhada oferecida pela Citrix através do MCS ou PVS, e a VMware com clones vinculados, apresenta diferentes desafios de E/S. Por natureza, essas abordagens de imagem compartilhada exigem menos capacidade de armazenamento, pois a imagem controladora é compartilhada e cada desktop virtual consome apenas uma pequena quantidade de espaço para seus dados exclusivos. A imagem compartilhada possui requisitos de desempenho diferentes de uma VM padrão. Essa imagem agora é usada por centenas ou milhares de desktops virtuais e precisa gerar grandes quantidades de IOPS. Se a imagem compartilhada for um gargalo, todos os desktops virtuais que a utilizam serão afetados negativamente e a experiência do usuário será ruim.



Levando em conta essas considerações para picos e diferentes tipos de arquiteturas de virtualização de apps/desktops, é preciso selecionar e projetar uma solução de armazenamento que seja capaz de atender às demandas de pico de inicialização, login e estado estacionário do ambiente. Para entender os requisitos de armazenamento do projeto, devemos realizar uma avaliação de desktop no ambiente físico de PC existente. Essa avaliação de desktop reunirá as informações sobre desempenho e capacidade da base de usuários para que seja possível aplicá-las aos cálculos do projeto.

Uma última reflexão sobre os requisitos de armazenamento relacionados à virtualização de apps/desktops é que, além de serem muito imprevisíveis em termos de E/S, os workloads de desktop também são muito pesados para gravação. Ao contrário de muitos workloads de servidor, que geralmente estão lendo dados e os servindo aos usuários, os desktops costumam gastar mais tempo gravando em disco. Na matriz de armazenamento, as gravações são mais intensas que as leituras. Um workload padrão de servidor pode ser 80% de leituras e 20% de gravação, enquanto o workload de desktop virtual de estado estacionário pode ser o oposto. Ao avaliar suas opções de armazenamento, preste muita atenção em como a solução de armazenamento armazena em buffer e aloca as gravações, em vez de prometer que ela faz um “ótimo trabalho” armazenando em cache blocos de leitura comum.

Tipos de armazenamento

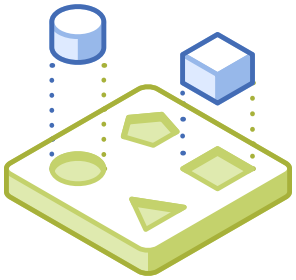
Existem vários tipos diferentes de armazenamento. As principais opções disponíveis atualmente são matrizes de armazenamento em camadas, matrizes de flash híbridas e matrizes all-flash. Cada uma dessas opções adota uma abordagem diferente para oferecer desempenho e capacidade aos workloads. Em cada uma delas, os fornecedores adotam abordagens distintas na criação de suas ofertas, portanto, preparamos uma breve explicação sobre elas, que você encontrará a seguir.

Arquiteturas tradicionais de camadas

São as matrizes corporativas tradicionais, que têm sido usadas para workloads baseados em servidor nos últimos 10 a 20 anos. Normalmente, são arquiteturas duplas baseadas em controladores. Na última década, elas foram modificadas para permitir a inclusão de vários níveis de discos de capacidade e desempenho na arquitetura. São fornecidos diferentes níveis de discos para testar e atender às demandas de capacidade e desempenho de workloads distribuídos. Há duas opções nessa abordagem: você pode projetar visando desempenho, criando pools dedicados com discos de alto desempenho para um workload, o que pode se tornar muito caro e restritivo. Outra opção é tentar aproveitar o armazenamento em camadas que foi adicionado a essa arquitetura para solicitar que a matriz promova ou rebaixe blocos de dados com base na demanda. O problema com essa divisão automática em camadas é que geralmente leva muito tempo para se tomar essas decisões sobre workloads de VDI.

All-Flash

As matrizes de armazenamento all-flash são totalmente compostas por armazenamento baseado em flash. Existem muitos tipos de flash que podem ser usados nessas matrizes de armazenamento. As matrizes de armazenamento all-flash modernas foram projetadas para aproveitar as características do armazenamento flash, o que significa que o sistema operacional e o sistema de arquivos foram projetados com o flash em mente. Alguns produtos adotaram um design de matriz tradicional e simplesmente substituíram os discos rígidos por all-flash. Embora seja mais rápido que a opção mais antiga, o produto final não foi projetado para essa finalidade.

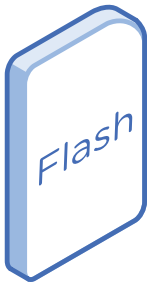


As matrizes de armazenamento all-flash são muito rápidas, com apenas um nível de desempenho no produto. Para garantir que a matriz também possa fornecer a capacidade necessária para o projeto a um preço acessível, você deve procurar matrizes que ofereçam desduplicação e compactação. Embora quase todas as matrizes de armazenamento all-flash modernas sejam mais fáceis de gerenciar do que seus equivalentes legados, elas nem sempre oferecem a mesma facilidade no gerenciamento ou o gerenciamento por VM que várias ofertas de flash híbrido fornecem.

Flash híbrido

As matrizes de armazenamento híbridas são arquiteturas modernas projetadas para usar uma combinação de pendrives e discos rígidos de forma eficiente. Os fornecedores adotaram diferentes abordagens de arquitetura em relação a como usam a capacidade e o desempenho em suas matrizes, mas os resultados finais são semelhantes. Todas são capazes de oferecer um desempenho impressionante com uma quantidade menor de flash e ainda assim fornecer uma grande quantidade de capacidade armazenando dados em grandes discos rígidos na matriz. As alternativas ideais de arquitetura de armazenamento híbrido utilizam inteligência integrada para classificar automaticamente os dados em unidades flash e de disco com base na demanda, eliminando a necessidade de ajuste manual e possíveis falhas de desempenho.

As arquiteturas mais adequadas para projetos de VDI modernos são as de armazenamento all-flash e híbridas. Essas arquiteturas são capazes de fornecer o desempenho necessário para ambientes de VDI e, geralmente, também oferecem as experiências de gerenciamento modernas citadas anteriormente. Os workloads de VDI são muito imprevisíveis por natureza e se sua solução de armazenamento precisa esperar para tomar decisões de armazenagem ou promover blocos para uma camada de cache, a demanda por desempenho irá desaparecer muito antes e a experiência será afetada negativamente.



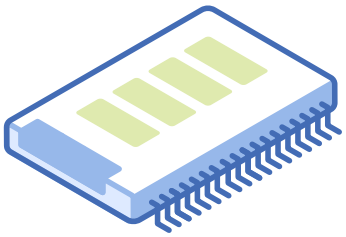
GPUs

Uma unidade de processamento gráfico (GPU) é uma parte importante de uma experiência de desktop moderna. Desktops e laptops físicos sempre tiveram uma GPU inclusa, mas elas não eram importantes, a menos que você quisesse jogar jogos ou realizar trabalhos com uso intenso de gráficos. Embora isso ainda se aplique, sistemas operacionais modernos e novos workloads, como machine learning (ML) e inteligência artificial (IA), também estão se valendo das GPUs para melhorar seu desempenho.

No espaço de VDI, as GPUs geralmente não são necessárias e costumam ser avaliadas com base em casos de uso para verificar se o desempenho adicional faz sentido em comparação ao balanceamento de custos adicionais e ao valor percebido para casos de uso com requisitos mais baixos. O Windows 10 ou streamings de vídeos são exemplos de casos de uso com requisitos mais baixos que podem aproveitar a GPU, mas podem não oferecer valor suficiente ao comparar com a oferta de uma experiência de usuário aceitável sem GPUs a um custo menor. No entanto, assim como na fabricação de CPU, as GPUs estão ganhando desempenho a um custo menor e o valor derivado das GPUs precisa ser parte integrante da avaliação geral ao projetar um ambiente de EUC.

A NVIDIA é a fabricante mais conhecida e com melhor desempenho de placas gráficas projetadas para virtualização de desktop. As placas gráficas da AMD funcionam apenas em determinados casos de uso e não oferecem as mesmas otimizações que as placas da NVIDIA.

As placas NVIDIA Grid permitem a virtualização de GPU, onde vários usuários podem compartilhar uma única placa gráfica. A virtualização de GPU não só suporta densidades mais altas de usuário, como também oferece desempenho nativo ao acessar um desktop virtual. As GPUs NVIDIA também contam com um mecanismo para codificação H.264 que descarrega processos da CPU, aumentando ainda mais a densidade do usuário em seu hardware. As placas NVIDIA Grid geralmente contam com várias GPUs, o que melhora o dimensionamento.



GPU dedicada

Com a passagem de GPU, você pode criar uma VM com uma GPU dedicada. Essa configuração oferece uma experiência de usuário comparável ao uso de um cliente fat com uma placa gráfica de ponta. No entanto, ao atribuir um núcleo de GPU a uma única VM, um desktop compartilhado hospedado (SBC) ou um desktop privado hospedado (VDI), limitam a escalabilidade.

GPU compartilhada

A tecnologia Grid permite que vários desktops virtuais compartilhem uma GPU, oferecendo a mesma experiência de usuário que as GPUs nativas. Esse compartilhamento é comumente conhecido como vGPU e é uma função do hipervisor e software Grid. Uma placa NVIDIA Grid M10, por exemplo, tem quatro núcleos físicos de GPU capazes de hospedar até dezesseis usuários por núcleo, resultando em 64 usuários, cada um com um desktop habilitado para vGPU, via placa M10.

A GPU processa os comandos gráficos da VM diretamente, o que significa que os usuários aproveitam gráficos de ponta sem perder desempenho devido à interferência do hipervisor. A vGPU é mais escalável do que a passagem, pois atribuímos perfis de vGPU aos nossos usuários e, assim, colocamos mais usuários na mesma placa.

Os perfis de vGPU fornecem memória gráfica dedicada através do vGPU Manager, que atribui a memória configurada para cada desktop. Um pacote de instalação do vSphere (VIB) instala o vGPU Manager no hipervisor. Com o AHV, um gerenciador de pacotes RPM executa essa tarefa. Cada instância de VDI conta com recursos predefinidos baseados nas necessidades das aplicações.

Licenças de Grid

Algo exclusivo da NVIDIA é que, para utilizar a funcionalidade vGPU, precisamos da licença. Existem diferentes níveis de licenciamento para apps virtuais usadas para soluções baseadas em RDSH. O nível de usuário avançado e designer para a maioria dos casos de uso de VDI abrangem aplicações avançadas, como aplicações da Adobe, engenharia e aplicações CAD. As licenças do Grid estão disponíveis nas opções de usuário nominal ou simultâneo.

Serviços de arquivos

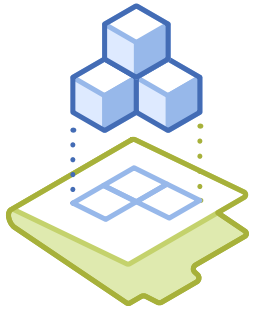
Os serviços de arquivos são amplamente utilizados e representam uma camada importante em arquiteturas de soluções de VDI. Tradicionalmente, os serviços de arquivo são fornecidos por um dispositivo NAS ou um dispositivo baseado em servidor Windows. Isso é apresentado como um compartilhamento SMB que diferentes serviços e dispositivos do sistema operacional convidado consomem para acesso compartilhado ou privado.

Geralmente, existem vários requisitos diferentes para serviços de arquivos em um projeto de VDI. O gerenciamento e a captura de perfis de usuário são importantes para proporcionar uma experiência de usuário consistente e a maioria das soluções de gerenciamento de perfis armazena os dados em compartilhamentos SMB. Também é bastante comum redirecionar pastas como parte do perfil do usuário para um compartilhamento SMB. E, por fim, os dados criados pelo usuário, como documentos, mídia e imagens, são armazenados em pastas privadas ou compartilhadas via SMB.

Serviços do Nutanix Files

A Nutanix oferece o Nutanix Files como recurso nativo na plataforma de nuvem da Nutanix. O Files é uma plataforma de serviço de arquivos escalável integrada que é gerenciada através do Prism, juntamente com todas as outras funções da Nutanix. Isso permite implementações simples com apenas 1 clique e upgrades sem interrupções. Essa simplificação possibilita que as equipes de VDI gerenciem mais da solução completa quando desejado.

O Files oferece uma arquitetura altamente escalável que permite adicionar mais capacidade às VMs de serviços de arquivos com apenas um clique, viabilizando conexões de usuário ou capacidade extras. As instâncias do Files podem ser executadas no mesmo cluster e suas VMs de VDI ou servidor em um cluster dedicado, se desejado. Para melhorar a flexibilidade na conectividade, o Files oferece suporte para conexões SMB 2.1, SMB 3.0, bem como NFS V3 e V4.



Dimensionamento de processamento

Existem diferentes correntes de pensamento sobre o dimensionamento da camada de computação do projeto. A primeira é a abordagem de escalabilidade vertical, que utiliza menos hosts grandes para fornecer recursos, enquanto a abordagem de escalabilidade horizontal utiliza mais hosts pequenos para fornecer recursos. O método preferido está entre as duas abordagens, que utiliza 2 hosts de soquete e os torna o mais densos possível, sem violar as taxas de consolidação definidas pelo projeto. O foco deste eBook é ajudar a dimensionar os recursos de computação para o workload de VDI.

Há três cálculos principais nos quais precisamos focar ao escalar os recursos de computação no projeto. São eles: a quantidade de memória física em cada host, a velocidade de clock da CPU e o número de núcleos de CPU com a taxa de CPU para eles.

Memória física

Em primeiro lugar, nunca devemos comprometer demais a memória em um projeto de VDI. Violar essa regra tem pouco valor e só levará a problemas de desempenho no ambiente.

Cálculo da frequência de clock da CPU

O cálculo da frequência de clock da CPU depende muito das informações coletadas na avaliação anterior do desktop. Os relatórios da avaliação fornecerão a quantidade média e o pico de utilização da CPU pelas sessões do usuário. Essas informações serão usadas juntamente com as informações sobre a memória derivadas da avaliação para fazer os cálculos.

Limites de utilização do host

Algumas outras recomendações sobre cluster de virtualização e host são nunca exceder 80% da utilização do host e sempre dimensionar seu cluster para N+1. A utilização de 80% do host não é apenas para implementações de virtualização de apps/desktops, é uma recomendação que se aplica a qualquer workload executado em um hipervisor. Se você estiver executando seus hosts acima de 80%, você deixa pouco espaço para picos e também pode acabar com sobra de recursos insuficiente para suportar uma falha de host, dependendo do tamanho do cluster.

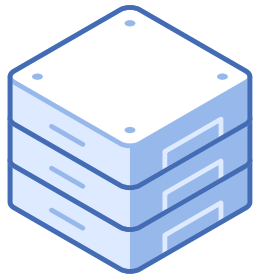
Sempre dimensione seu cluster para N+1

O segundo elemento para calcular N+1 no dimensionamento do seu cluster é garantir que haja recursos suficientes no cluster para suportar uma falha única de host, garantindo que todas as VMs possam continuar em execução e que as que falharam sejam reiniciadas sem problemas. Uma falha única de host é o nível mais comum de resiliência. Poucos clientes exigem N+2 para cumprir com mais exigências de SLA.

Taxas da CPU

O elemento final na questão de dimensionamento da computação é a taxa de CPU, que foca no número de CPUs virtuais para CPUs físicas (vCPU:pCPU). Essa taxa é muito importante porque, se aumentar muito, chegará a um ponto em que surgirão problemas de agendamento de CPU, afetando drasticamente o desempenho e a experiência do usuário. Quando ocorre um problema de agendamento de CPU em hosts do vSphere, o tempo de preparação da CPU aumenta, apontando que o agendador está com problemas para fazer o agendamento de todas as vCPUs nas pCPUs. Isso significa que a vCPU terá que esperar, mesmo que esteja pronta. A taxa de CPU varia muito de acordo com os diferentes tipos de workloads virtualizados em clusters VMware. Geralmente, os workloads de servidor e banco de dados contam com uma taxa muito menor, enquanto os workloads de VDI podem apresentar uma taxa maior.

O uso de vCPUs não é um cálculo linear, o que significa que é possível criar um host com taxa de consolidação mais alta caso todas as VMs tenham uma única vCPU. Quando muitas VMs possuem duas ou mais vCPUs, isso afeta os cálculos. Não é tão fácil quanto dividir por dois para ter o dobro de vCPUs. A figura 6 ilustra uma faixa que comprovadamente funciona com implementações reais de clientes. Os fabricantes que realizam testes sintéticos podem apresentar taxas mais altas. Deve-se ter cuidado com isso, pois nem sempre se aplicam a projetos reais.



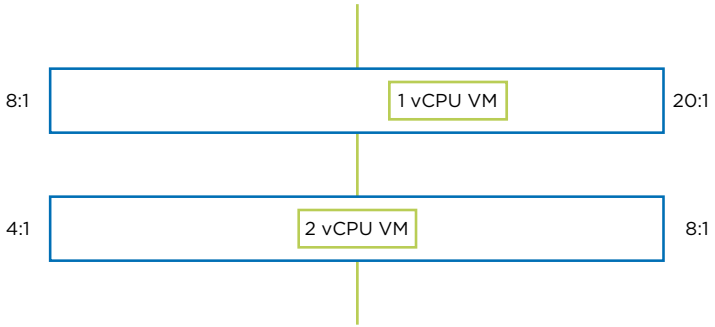
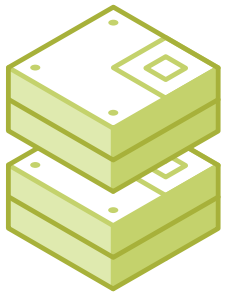


Figura 6:
A consolidação da VDI é fortemente orientada pela taxa de vCPU com a qual seus desktops virtuais serão configurados. O gráfico representa uma faixa que a prática provou ser segura.

A faixa de trabalho para operar normalmente com desktops virtuais de vCPU única é entre 8:1 e 20:1. Esta é uma faixa ampla. O ponto exato dentro desta faixa depende de diferentes decisões. Uma delas seria o tamanho dos hosts, o número de VMs por host e o nível de conforto do cliente com esse número. Um exemplo seria um host de soquete duplo com CPUs duplas de 18 núcleos. Essa configuração pode acomodar mais de 700 VMs na melhor das hipóteses, desde que você tenha a quantidade certa de memória e frequência de clock suficiente disponíveis. Geralmente, ter muitas VMs em um único host assusta a maioria dos clientes. Consequentemente, há duas escolhas a serem feitas neste cenário. A primeira é optar por uma densidade mais baixa que se limita artificialmente. Se a taxa mais baixa for escolhida, ela renderá 288 VMs no mesmo host. A segunda opção seria escolher CPUs com menos núcleos, mas com uma taxa no meio do caminho. Se escolhermos CPUs de 12 núcleos e usarmos uma taxa de 12:1, isso resultaria em 288 VMs. Normalmente, essa decisão é uma combinação de feedback do cliente, recomendações do arquiteto e preços da infraestrutura. A escolha de diferentes configurações físicas de CPU pode gerar uma economia significativa.

Os cálculos para um desktop virtual de vCPU dupla são semelhantes, exceto pelo fato de que agora estamos lidando com o dobro da quantidade de vCPUs. A faixa para operar aqui é entre 4:1 e 8:1. Alguns fornecedores prometem mais, mas essas recomendações são orientadas por implementações reais de clientes. Os mesmos pontos de decisão que o exemplo anterior devem ser usados, mas com uma faixa de taxa de CPU diferente.

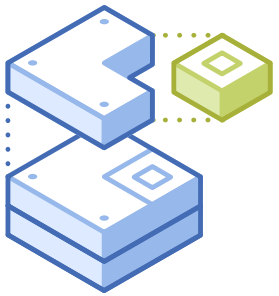


Outra coisa que você deve ter em mente é que, se você selecionar uma taxa de CPU no meio dessas faixas, ela fornecerá a liberdade de dimensionar a densidade de consolidação verticalmente, caso o ambiente continue funcionando dentro das tolerâncias. Um fator que deve ser observado é que não há lugar para configurar essas taxas de CPU como configuração em qualquer outra ferramenta atualmente. Esses são atributos que devem ser declarados no projeto e se tornar pontos de dados que precisarão ser considerados no gerenciamento e no dimensionamento do ambiente. Assim como a memória e a frequência de clock, a taxa de CPU precisa ser calculada dentro da decisão de adicionar mais VMs a um cluster e quando adicionar outro host a um cluster para fornecer mais recursos.

É possível gerenciar a taxa da CPU através de cálculos manuais, coletando dados. Alguns administradores usam um script do PowerShell que coleta dados e apresenta a taxa como resultado do script. Com um script, ele pode rodar como um trabalho programado diariamente, para garantir que ninguém esteja violando a taxa ou corra perigo em nenhum dos clusters.

A frequência do barramento de memória ou RAM também está associada ao dimensionamento de computação. A regra geral ao dimensionar a memória é buscar a maior densidade com os orçamentos de velocidade de barramento mais rápidos. O desafio frequentemente enfrentado com a memória é que a memória mais lenta pode resultar em ciclos de CPU ociosos aguardando a conclusão das transações de leitura/gravação na RAM.

A incorporação de GPUs aos clusters e ao projeto de VDI geralmente também afetará a densidade de usuários por host. Isso está diretamente relacionado ao número de placas GPU e ao número de GPUs por placa que podem ser colocadas no host desejado e, portanto, com o perfil de vGPU selecionado para seus usuários. Exemplo simples de um host que pode aceitar duas placas GPU, cada uma com uma GPU. Então, um perfil de vGPU que permite 16 usuários por placa é escolhido, o que significa que apenas 32 usuários que precisam de GPU caberiam nesse tipo de host. Se houver CPU e memória disponíveis no host, ele ainda poderá executar VMs sem GPU. Identificar o perfil correto da vGPU é importante para um dimensionamento eficiente.

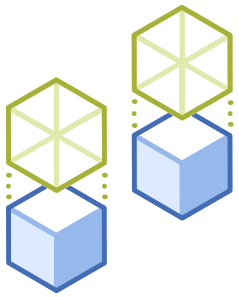


Design do cluster de **Virtualização**

Há várias razões para se criar diferentes clusters de virtualização em um projeto de EUC. A decisão de se ter diferentes clusters geralmente está relacionada a diferentes workloads e tamanho do cluster. Não perderemos muito tempo com esse assunto neste eBook, mas aqui estão algumas recomendações com base nos assuntos abordados em outras partes do livro mais completo, online.

Em primeiro lugar, ao criar um projeto de VDI com mais que algumas centenas de usuários, é essencial separar a infraestrutura de gerenciamento de virtualização do workload de VDI. Isso significa que todos os servidores de gerenciamento, agentes de VDI, servidores de arquivos, servidores de gerenciamento de aplicações e quaisquer outras funções que não sejam desktops virtuais devem ser executados em um cluster diferente.

Se o cluster de gerenciamento precisa ser dedicado apenas ao projeto de EUC, isso dependerá do tamanho do ambiente. Se o projeto for menor, é possível executar VMs de gerenciamento em um cluster de virtualização de servidor existente. É possível escalar esses clusters de desktop virtual para alcançar um tamanho entre 16 e 32 hosts. Essa faixa permite que seja criado um pool de recursos maior para uso das VMs e também leva a maioria dos clientes a adotar um cluster maior do que seus tamanhos padrão. Atualizações recentes de hipervisores permitem clusters de até 64 hosts, mas levará algum tempo até que muitos arquitetos e clientes se sintam confortáveis com esse tamanho. Se o ambiente for grande o suficiente para que o número de hosts exceda essas faixas, haverá a necessidade de mais de um cluster VDI.



Outro motivo para se fazer um projeto para vários clusters de virtualização, além do tamanho do ambiente, seria para diferentes workloads. Existem diferentes workloads nos clusters VDI. Se houver uma quantidade significativa de desktops virtuais de 1 vCPU e 2 vCPUs, deve ser projetado um cluster separado para cada um. A figura 7 ilustra uma abordagem de projeto com vários clusters. Isso permite o gerenciamento da taxa de CPU de forma diferente em cada cluster, permitindo um projeto mais fácil de gerenciar. Se combinássemos as diferentes configurações de CPU, haveria uma nova taxa combinada que precisaria ser calculada, o que só confunde as coisas.

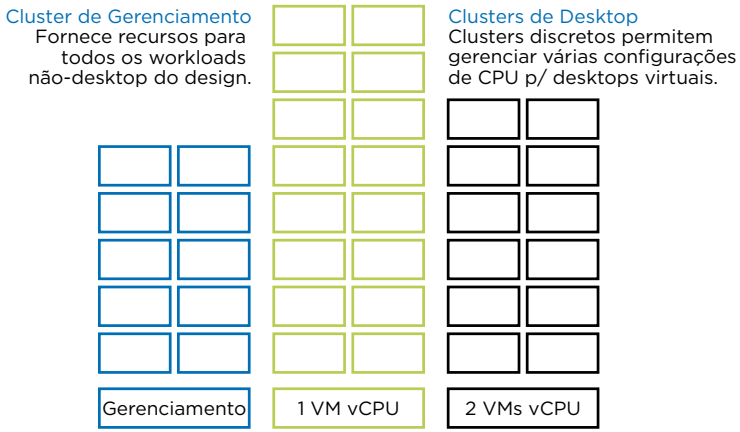


Figura 7:
Clusters de gerenciamento e desktop

O uso de GPUs pode ser outro motivo para considerar um cluster dedicado para usuários de GPU. Você pode combinar usuários de GPU e não GPU no mesmo cluster, mas isso pode tornar as operações e a programação desses usuários um pouco mais complexas. Se você tiver usuários de GPU suficientes para atender aos requisitos mínimos de um pequeno cluster, geralmente vale a pena fazê-lo.

Forneça uma excelente experiência de usuário final onde quer que você implemente

Conheça as soluções da Nutanix para computação do usuário final em **nutanix.com/euc**

NUTANIX