NUTANIX

GUÍA DEFINITIVA SOBRE

END USER COMPUTING

Un enfoque híbrido y multinube para escritorios y aplicaciones virtuales

Índice

Autor
Brian Suhr tiene más de dos décadas de experiencia en TI en diseño, implementación y gestión de infraestructuras empresariales. Ha aplicado su experiencia en arquitectura e ingeniería en numerosos proyectos de virtualización, centros de datos y nube, trabajando con equipos técnicos de alto rendimiento en entornos de alcance global. Como autor independiente de los blogs DataCenterZombie y VirtualizeTips, Brian se enfoca en crear contenido relacionado con virtualización, automatización, infraestructura y difusión de productos y servicios que benefician a la comunidad tecnológica. Se le puede seguir en Twitter: @bsuhr

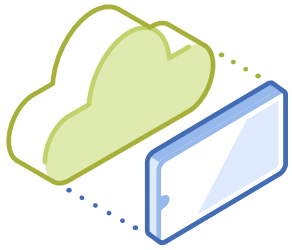
Acerca de este libro electrónico
Este libro electrónico se enfoca en el diseño de infraestructura para entornos VDI y de End-User Computing (EUC). El contenido de este libro electrónico se ha adaptado a partir del capítulo enfocado en infraestructura del libro Arquitectura y diseño de soluciones de End-User Computing (disponible en Amazon).

Introducción

Después de seleccionar la estrategia y el proveedor de software adecuados para ofrecer servicios y aplicaciones de EUC, la arquitectura de implementación y las opciones de infraestructura constituyen las siguientes decisiones importantes en los proyectos de virtualización de aplicaciones y escritorios.

La decisión sobre dónde se ejecutará y cómo se operará el plano de control es una decisión cada vez más importante, ya que existe una variedad de opciones cada vez mayor. La infraestructura es la base sobre la cual se construyen los servicios. Al igual que la energía eléctrica y el agua, es algo que se da por sentado, y deberían simplemente funcionar cuando abrimos el grifo o encendemos el interruptor. Se pueden alquilar como servicio, o alguien tiene que ocuparse de ellas.

Sin una infraestructura y un plano de control estables, de alta disponibilidad y alto rendimiento, el departamento de TI puede enfrentarse a una serie de desafíos durante las fases de implementación y operación de su proyecto de EUC. Si bien la infraestructura desempeña un papel esencial, el departamento de TI no debería dedicar una cantidad significativa de tiempo a implementarla y mantenerla, como suele ser el caso. La infraestructura adecuada debería funcionar, liberando a los arquitectos e ingenieros para que se concentren en proporcionar los servicios y aplicaciones de EUC.



Arquitecturas de **implementación de EUC**

La identificación de la arquitectura de implementación óptima depende de los requisitos de una empresa. Esta decisión, a su vez, influye en la elección del plano de control de EUC de una empresa. A continuación encontrará los diferentes tipos de planos de control y opciones de implementación para los agentes de EUC. El plano de control, como el nombre sugiere, se encarga del aprovisionamiento, alimentación e intermediación de datos y es la interfaz principal para los administradores. Por lo general, también funciona como una interfaz API.

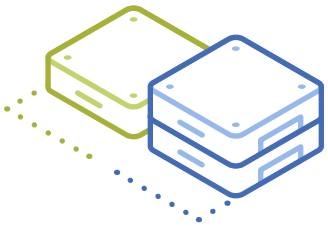
Los planos de control de EUC suelen intermediar una conexión de usuario para una aplicación o escritorio. Para las aplicaciones, intermedian a través de la presentación de la aplicación (generalmente basada en RDSH). Para los escritorios, pueden utilizar escritorios compartidos alojados (HSD) o una sesión de infraestructura de escritorio virtual (VDI).

A continuación se muestran las tres alternativas principales, junto con algunos ejemplos de casos de uso.

Agente on-premise tradicional

Con esta opción de agente de EUC, las empresas suelen comprar licencias por usuario e instalarlas en su centro de datos. Si bien varios proveedores y productos pertenecen a esta categoría, los más utilizados por mucho son Citrix Virtual Apps and Desktops (CVAD) y VMware Horizon.

Al construir una implementación de EUC local, la responsabilidad de la arquitectura y la implementación recae en la empresa o en el servicio subcontratado; esto incluye determinar todos los componentes del software intermediario que se implementará, como servidores de controladores, servidores de bases de datos, servidores de licencias, dispositivos o servidores de seguridad periféricos y cualquier otro servicio de soporte necesario. Como parte de la implementación de estos elementos, también debe determinar qué opciones de alta disponibilidad (HA) están disponibles para cada servicio y seleccionar una que tenga sentido para sus requisitos de diseño.



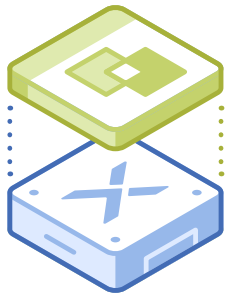
Por lo general, en la mayoría de estos servicios también hay un aspecto de carga y escalado. Deberá dimensionar y diseñar correctamente cuántas conexiones puede manejar cada servicio para determinar la cantidad adecuada de servidores de controlador que necesita, por ejemplo, para 5,000 conexiones con alta disponibilidad. A continuación, puede determinar si va a implementar todos los servidores del controlador para las 5,000 conexiones por adelantado o agregarlos a medida que escala la implementación. Si planea escalar más allá de este número, puede utilizar estos detalles para seguir escalando los diferentes servicios y cumplir con las mejores prácticas y los máximos permitidos.

La siguiente consideración son las operaciones, donde es necesario comprender cómo parchear y actualizar todos los servicios en el agente de EUC implementado. Aparte de los servicios de los que ya hemos hablado (generalmente contenidos en todas las máquinas virtuales de servidor implementadas), también debe mantener agentes y clientes. Los agentes se encuentran en las imágenes utilizadas para implementar los grupos de servidores de aplicaciones y escritorios a los que se conectan los usuarios. La forma en que se actualizan los agentes depende de cómo se aprovisionan estos grupos. Los clientes se encuentran en los endpoints que utilizan los usuarios para conectarse a estos servicios, que varían en gran medida según el estilo de endpoint.

A pesar del nombre, en realidad puede implementar un agente on-premise tradicional en una nube pública. Casi siempre tiene más sentido implementarlo on-premise, de ahí el nombre.

Agente de nube

Los agentes de nube se ofrecen como servicio en el modelo típico de software como servicio (SaaS) y se les conoce más comúnmente como productos de escritorio como servicio (DaaS). Varios proveedores y productos pertenecen a esta categoría. Los proveedores de nube pública tienen sus propios productos. Sin embargo, dado que Citrix es el líder claro en el mercado de EUC, el servicio Citrix Virtual Apps and Desktops (CVADS) es ideal para empezar a evaluar los beneficios de DaaS. La mayoría de los productos ofrecen un conjunto de características a la par con las proporcionadas por sus contrapartes on-premise tradicionales, pero eso no quiere decir que no tengan sus propios beneficios y limitaciones.



Hay muchos productos distintos de DaaS disponibles de proveedores de software, proveedores de nube y proveedores de servicios. Varían enormemente en lo que ofrecen cuando busca algo más allá de grandes características como infraestructura de escritorio virtual y presentación de aplicaciones, de modo que es importante comprender sus casos de uso y sus requisitos a la hora de tomar una decisión.

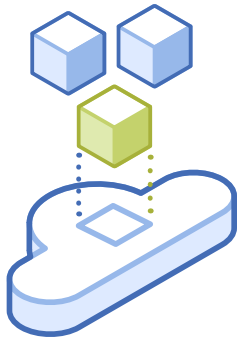
Los requisitos son el criterio con el cual guiarse. Si le gusta una opción que no cumple con algunos de los requisitos, debe determinar si puede vivir sin ellos hasta que se cierre esa brecha, si es que alguna vez lo hace.

En cuanto a servicios a los que se suscribe, estos tienen varias características interesantes. En primer lugar, normalmente paga por el servicio por usuario y por mes, lo cual hace que el costo sea fácil de calcular y rastrear. El costo también proviene del presupuesto OpEx, lo cual representa un beneficio para algunos, ya que se ven obligados a pasar al modelo de consumo en la nube. Puede firmar acuerdos a largo plazo con estos productos de DaaS que probablemente ofrezcan mejores precios a cambio de un compromiso prolongado; sin embargo, si se suscribe tiene la opción de retirarse si su demanda disminuye o desaparece. Con el enfoque tradicional de licencias, las empresas normalmente compran licencias y luego pagan por soporte perpetuo.

Dado que se trata de un servicio, las empresas pueden consumirlo bajo demanda. Usted gestiona todos los aspectos de diseño, implementación y escalado de la capa de intermediación. Además, normalmente ejecuta las aplicaciones y escritorios en la nube pública junto a la oferta de DaaS, de modo que tampoco tiene que gestionar la infraestructura en la que se ejecutan, lo cual a su vez hace que no tenga que monitorear y actualizar estas capas. Estas ofertas de DaaS pueden hacer que su proyecto despegue más rápido y representan menos esfuerzo desde el punto de vista operativo.

Si bien el servicio DaaS cubre las capas de intermediación e infraestructura, quedan tareas operativas por realizar. Estas funciones incluyen creación y actualizaciones de imágenes, instalaciones y actualizaciones de aplicaciones, datos de usuarios, VPN, etc.

Otro beneficio que ofrece la nube es la capacidad de utilizar potencialmente centros de datos en todo el mundo para ofrecer grupos de aplicaciones o escritorios cerca de un grupo de usuarios. Este beneficio no siempre funciona. Por ejemplo, si los usuarios dependen mucho de los datos o de una aplicación específica en un centro de datos que está lejos, esta capacidad es nula.



Implementaciones híbridas

Ahora que hemos hablado sobre las diferentes arquitecturas para los agentes de EUC, veamos cómo se implementan. A pesar de algunas predicciones que apuntan a que todas las cargas de trabajo van a la nube pública, está claro que la mayoría de las implementaciones aún están on-premise y que la mayoría de las empresas está avanzando hacia una arquitectura híbrida. Para fines de esta explicación, el término "híbrido" se refiere simplemente a una combinación de on-premise y nube pública. Los porcentajes de ambas varían dependiendo de diversos factores. Cuando se trata de una estructura híbrida en el mundo de EUC, hay formas en las que funciona tanto para los sistemas tradicionales como para DaaS. Veamos primero cómo cada una de las arquitecturas puede encajar en un mundo híbrido. A continuación, profundizaremos en algunos ejemplos de casos de uso.

La solución DaaS es el enfoque más común para las soluciones híbridas, y no todos los proveedores de DaaS ofrecen la capacidad de implementar una arquitectura híbrida. El plano de control generalmente permanece en la nube para la versión híbrida de DaaS, y aquellos que lo soportan tienen un método para controlar y gestionar los grupos de recursos privados, generalmente mediante la implementación de una máquina virtual local en sus clústeres on-premise para que actúe como conector local. A través de estas máquinas virtuales conectoras en la nube, el proveedor de DaaS ahora puede aprovisionar máquinas virtuales, controlar sus estados de energía e intermediar conexiones con ellas. Las máquinas virtuales conectoras son fáciles de configurar y se puede implementar en pares para proporcionar HA. Esta es la arquitectura que soportan Citrix CVADS, Horizon Cloud y Nutanix Frame.

En general, la versión híbrida de DaaS sigue siendo mucho menos compleja de diseñar, implementar y operar que con el enfoque tradicional. Sigue siendo responsable de gestionar la infraestructura que está ejecutando las máquinas virtuales locales, pero no tiene que gestionar la capa de agente.

Casos de uso

A continuación, vamos a hablar acerca de los diferentes escenarios y casos prácticos que son una buena opción para un planteamiento híbrido. Si se utiliza correctamente, un diseño híbrido puede ofrecer una gran flexibilidad en términos de costos y características. La mayoría de los casos de uso se dividen en tres escenarios distintos.

- **Recuperación ante desastres (DR).** Este escenario probablemente se ajuste a la mayoría de las implementaciones locales tradicionales con una necesidad de continuidad del negocio (BC). Históricamente, la recuperación ante desastres ha implicado construir o alquilar un sitio secundario y utilizarlo para implementar recursos de EUC para conmutación por error. La nube pública ofrece una alternativa muy atractiva al enfoque tradicional, en el sentido de que se puede reservar parte de la capacidad de la nube y luego ampliarla al máximo en caso de desastre. En este enfoque, usted paga para ejecutar las máquinas virtuales de su infraestructura de intermediario y un pequeño grupo de escritorios constantemente con capacidad de almacenamiento para datos y perfiles de usuario replicados. Luego, en caso de que se produzca alguna eventualidad, puede expandir rápidamente ese grupo de escritorios al tamaño que necesite y avisar a sus usuarios para que empiecen a trabajar de nuevo. Cuando todo está hecho, vuelve a reducirse al tamaño de estado estacionario para volver a reducir los costos.

- **Expansión.** Hay algunos casos de uso o proyectos que solo necesitan recursos durante un corto tiempo, y puede que no tenga sentido mantener el espacio on-premise para ellos. En este caso, puede iniciar recursos (normalmente en una nube pública) y luego destruirlos cuando desaparece la necesidad. Hay diversos escenarios que podrían recurrir eficazmente a la expansión. Los laboratorios de estudiantes son un caso de uso común, porque las necesidades son estacionales y están programadas alrededor del calendario del estudiante. Los estudiantes también pueden tener necesidades de GPU que usted no tiene on-premise. Los proyectos especiales y el trabajo temporal también son casos de uso comunes de expansión.

- **Híbrida real.** Esta opción seguramente captura la mayoría de los otros casos de uso en el sentido de que no son necesidades temporales, lo cual significa que está ejecutando regularmente algunas cargas de trabajo on-premise y algunas en la nube y decide dónde implementar en función de los parámetros que tienen sentido para su empresa y su diseño. Por ejemplo, puede tener un caso de uso en el que necesite GPUs y no las tenga on-premise; es posible que deba implementar en la región Asia-Pacífico en lugar de que los usuarios locales regresen a sus sitios de Norteamérica; o puede que simplemente no tenga capacidad en su entorno local.

Principios de **arquitectura**

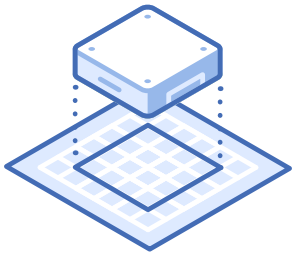
Hay varios factores importantes que deben tenerse en cuenta en el proceso de diseño de la infraestructura de EUC. Utilizando estos factores junto con los requisitos de la empresa, se puede apreciar mejor cuáles serán las alternativas de arquitectura. Hay que tener en cuenta los factores siguientes a la hora de evaluar alternativas de arquitectura y opciones de proveedores en proyectos de EUC:

- Punto de entrada
- Escalabilidad
- Rendimiento
- Monitoreo
- Capacidad

Punto de entrada

El punto de partida de la infraestructura a menudo puede ser una desición crucial en un proyecto. Esta es la cantidad de infraestructura y el costo que implicará para la empresa poner en marcha la implementación de entrega y virtualización de aplicaciones o escritorios en función de diferentes tamaños de punto de partida.

Si se planea que el proyecto llegue a 10,000 usuarios cuando se implemente por completo con la fase inicial de implementación de 5,000 usuarios, probablemente a la empresa le impacte menos los costos iniciales. El razonamiento es que dependiendo del tipo de infraestructura que se seleccione, es posible que el costo por usuario no empiece a tener sentido hasta que hayan implementado algunos miles de usuarios.



La otra cara de la moneda es cuando una empresa va a implementar 10,000 usuarios pero solo tiene la intención de empezar con 500 usuarios y escalar a un ritmo constante durante el proyecto. Van a analizar más de cerca el costo de la implementación de una infraestructura inicial de estas dimensiones en lugar de dar un primer paso más grande. El costo por usuario de este tamaño puede mantenerse constante a medida que escala el entorno, o puede parecer sesgado al principio, debido a un mayor gasto inicial en infraestructura.

Aunque el costo por usuario puede considerarse impreciso y casi irrelevante como un factor para determinar los costos de su infraestructura, se le preguntará sobre esto cuando intente vender el proyecto a la empresa o justificar a la directiva su selección de infraestructura. Si elige una alternativa que tenga un costo de usuario inicial más alto, debe estar preparado para explicar los detalles. Evalúe las soluciones que crea que serían más adecuadas para su entorno. De lo contrario, prepárese para definir la decisión sobre cómo se desarrollarán los costos. La figura 1 ilustra una muestra de estos dos escenarios.

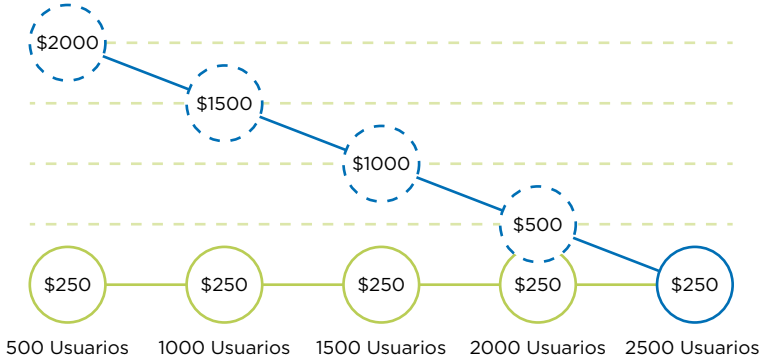


Figura 1:
Puntos de entrada por escritorio

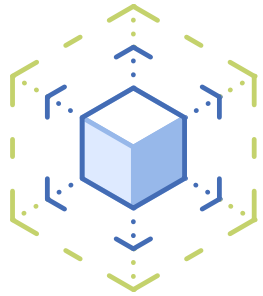
Escalabilidad

La escalabilidad de la arquitectura es un factor importante a la hora de evaluar la viabilidad del proyecto. Un arquitecto deberá comprender las opciones de tamaño inicial para las distintas alternativas; esto nos devuelve al tema del punto de entrada que acabamos de tratar. ¿La alternativa permitirá fácilmente que el diseño empiece con un tamaño más pequeño según sea necesario? ¿O la empresa necesitará comprar más infraestructura de la que sería necesaria para satisfacer el tamaño inicial de un proyecto y no podrá utilizar todos los recursos hasta que el proyecto avance?

Además de a qué tan pequeña escala puede empezar la alternativa, es igual de importante considerar hasta qué tamaño puede escalar. Si la intención es empezar con 500 y poder escalar hasta 10,000 usuarios, ¿cómo será la alternativa en ambos extremos de ese espectro? ¿Quedará satisfecha la empresa con los puntos bajos o altos? ¿O ambos?

El tema de la escalabilidad no es solo una cuestión de almacenamiento. También se aplica al procesamiento, a las redes y posiblemente a otras capas dentro del diseño. Si se realizan ajustes en la configuración de la capa de procesamiento para lograr una densidad menor de máquinas virtuales por servidor host, ¿cómo podría afectar esto a diferentes opciones de diseño a la hora de escalar? Un ejemplo sería si el diseño de host inicial empieza con 128 GB de memoria por host y la opción final es de 256 GB o más, será necesario asegurarse de que se utilicen DIMMs del tamaño adecuado para permitir que la configuración escale en el futuro. Si se toman decisiones incorrectas para ahorrar costos, las restricciones tendrán un efecto en la densidad, o acabará costando más a largo plazo debido a los DIMMs que no se pudieron reutilizar.

El arquitecto debería enfocarse en cómo la solución podrá empezar a escala reducida, además de poder escalar hasta el punto más grande. Pero tampoco se pueden ignorar todos los puntos intermedios, porque dependiendo de cómo se escale la implementación, podría haber muchos puntos de escalado entre el inicio y el final. Es ideal buscar una solución que permita al diseño escalar fácilmente en grupos de usuarios que encajen en el proyecto, pero sin superar los períodos previstos ni las capacidades de implementación. El tamaño de escalado ideal para un proyecto puede ser en incrementos de 500 a 1,000 usuarios. Pero si la alternativa de arquitectura elegida escala a una dimensión superior, debe entender cómo afecta esto a los costos y a la implementación.



Rendimiento

El rendimiento de EUC medido por la experiencia del usuario final siempre es una de las consideraciones principales. La arquitectura seleccionada debe poder cumplir con los requisitos en cualquier fase del proyecto. Este puede ser un camino difícil de recorrer con algunas alternativas. Si se escala una solución para cumplir con los requisitos mínimos del usuario inicial, puede terminar sacrificando el rendimiento si no se puede escalar de forma lineal. Los arquitectos no quieren hacer concesiones en la arquitectura para alcanzar este pequeño punto de partida que pueda afectar las opciones generales de rendimiento máximo de una solución. Si le dedica tiempo al principio a tomar la decisión correcta, puede evitarse problemas más adelante.

Un diseño de solución EUC normalmente tendrá muchos requisitos de rendimiento distintos. Seleccione una alternativa de arquitectura que sea lo suficientemente flexible como para cumplir con todos los requisitos de rendimiento dentro de una sola opción. Ya sea que el diseño proporcione varios tipos de servicios de EUC o solo se enfoque en la virtualización de aplicaciones y escritorios, hay que tener en cuenta múltiples necesidades de rendimiento. Comprender cómo cada alternativa podrá o no cumplir con los requisitos individuales de rendimiento afectará en gran medida a su proceso de evaluación y diseño.

Capacidad

El argumento sobre la capacidad es similar al de rendimiento. Hay una serie de requisitos de capacidad distintos dentro de los diseños de EUC que deberán proporcionarse. La solución requerirá la ejecución de máquinas virtuales de servidor, máquinas virtuales de escritorio, aplicaciones, perfiles de usuario y datos de usuario para este tipo de arquitectura. Cada capa dentro del diseño puede tener requisitos de capacidad muy distintos. Algunas utilizan grandes cantidades de datos que normalmente se deduplican bien. Otros aspectos como los perfiles de usuario y los datos consisten en cantidades más pequeñas de datos por usuario, pero multiplicadas por miles de usuarios al final constituyen una cantidad considerable.

En los últimos años, un problema habitual ha sido la adquisición de demasiada o muy poca capacidad al intentar alcanzar los niveles de rendimiento requeridos. Examine atentamente las alternativas de arquitectura durante la fase de diseño para ver cómo podrán ofrecer la capacidad requerida, mientras que se asegura de que se cumplan los requisitos mínimos de rendimiento. La alternativa no debería proporcionar 2-3x o más capacidad para cumplir con los requisitos de rendimiento de almacenamiento, ni agregar rendimiento adicional significativo para cumplir con los requisitos de capacidad.

La solución ideal es aquella que permite la suficiente flexibilidad para escalar el rendimiento y la capacidad a tasas similares, de modo que ninguna salga del rango de la otra.

En el pasado, este tema ha causado muchos problemas. Muchas empresas han tenido dificultades de planificación de rendimiento y capacidad al escalar la capacidad más rápido que el rendimiento. El hecho de que la solución tenga 5TB de espacio libre no significa que pueda escalar a otros 500 usuarios. Este escenario puede hacer que el rendimiento se vea muy afectado. Los administradores y directores de TI que no tengan una comprensión sólida de cómo se escala la solución pueden caer en esta trampa.

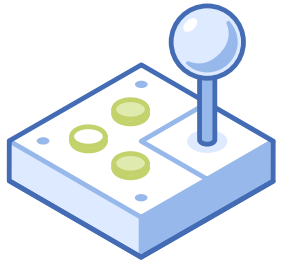
Monitoreo

El monitoreo es muy importante, y a menudo se pasa por alto. Cuando se trata de monitorear la infraestructura en un entorno EUC, los administradores generalmente se enfocan en el aspecto del rendimiento. Necesitan la habilidad de entender qué es normal y cuándo hay un problema activo.

El uso del monitoreo debería ser simple y a la vez proporcionar una gran cantidad de información detallada. Este no es el caso con muchos fabricantes, así que hay que observar detenidamente cuál es la experiencia de monitoreo con cada alternativa.

Otro requisito es la capacidad de proporcionar monitoreo de rendimiento a nivel de máquina virtual. Desafortunadamente, la mayoría de los proveedores de infraestructura todavía no pueden ofrecer este nivel de visibilidad en el entorno de virtualización. El mejor tipo de monitoreo de rendimiento de infraestructura debería ofrecer a los administradores la capacidad de examinar rápidamente la capa de almacenamiento y determinar si el problema de rendimiento del almacenamiento es global o si está aislado a un host, grupo de máquinas virtuales o a una sola máquina virtual.

Al gestionar el rendimiento del almacenamiento a nivel de máquina virtual, se puede utilizar un enfoque similar a la gestión del rendimiento de CPU y memoria de una máquina virtual a nivel de host. Los administradores necesitan saber si una máquina virtual está utilizando rendimiento adicional temporalmente, o si se trata de un consumidor habitual de más rendimiento de almacenamiento que los usuarios habituales. Esto les permitirá comprender cuándo hay un pico y cuándo buscar más a fondo para identificar el problema.



Bloques de **construcción**

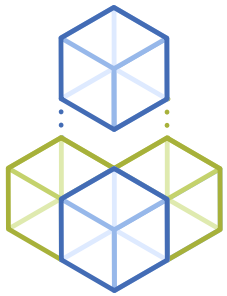
Un bloque de construcción es un conjunto predefinido de infraestructura que se asigna a una cantidad específica de recursos o a un número de usuarios. Este es uno de los mejores enfoques para abordar el diseño de infraestructura con End-User Computing.

Utilizando este enfoque, se puede desarrollar una arquitectura que ofrezca un modelo predecible de costo, rendimiento y escalado de capacidad. Al determinar el tamaño de los bloques de construcción, elija qué incrementos necesita para escalar los usuarios y cómo la selección de la infraestructura puede adaptarse a las opciones. Por ejemplo, es posible que quiera escalar los usuarios en incrementos de 50 a 100 usuarios, pero la elección de infraestructura no escala bien en incrementos tan pequeños. Esto puede obligar al diseño a escalar en incrementos más grandes de 500 o 1000 usuarios. Si la opción de infraestructura escala en bloques grandes, puede elegir escalar para que se integre con eso o simplemente aceptar el hecho de que los costos de infraestructura no escalarán de la misma manera que los bloques de implementación del usuario. Esto simplemente significa que la empresa compraría infraestructura en bloques de 1,000 usuarios y solo se implementaría en grupos de 50 a 100 usuarios.

Hace que los costos de los escritorios virtuales o las sesiones de usuario parezcan caros al comprar un bloque grande para implementar una cantidad menor de usuarios. Esto se equilibra si la empresa implementa todos los usuarios planificados.

Las arquitecturas al estilo de bloques de construcción son útiles en cualquier proyecto de diseño, pero las implementaciones de EUC siempre tienen partes comunes de usuarios y casos de uso que tienen características similares y se implementan en grupos. Siguiendo con el ejemplo de un tamaño de bloque de 100 usuarios, si comprende los requisitos de recursos de 100 usuarios, podrá asegurarse de que el bloque de infraestructura pueda proporcionar todo lo que necesitan esos usuarios.

Si cada usuario requiere 15 IOPS en estado estable y 30 GB de capacidad de almacenamiento, junto con 2 GB de memoria y 200 MHZ de CPU, el arquitecto sabe que los componentes básicos deben proporcionar 1,500 IOPS, 3 TB de capacidad, 200 GB de memoria y 20 GHZ de CPU. El arquitecto puede diseñar los bloques de construcción para que contengan recursos adicionales, pero ninguno de ellos puede estar por debajo de esos valores. También conviene



evitar el desperdicio que representa el incluir demasiados complementos adicionales en cada bloque que no se puedan utilizar.

Con este enfoque y granularidad en el diseño, ahora se puede escalar el entorno en grupos de 100 usuarios. Esto permite un enfoque lento y constante y proporciona valores predecibles que las empresas pueden planificar para la implementación, rendimiento, capacidad y costos. Si las empresas quieren escalar más rápido y en grandes cantidades, solo tienen que desplegar múltiples bloques de construcción a la vez.

Por último, el enfoque de bloques de construcción ha demostrado ser atractivo ya que a la mayoría de las implementaciones de clientes les gusta empezar con despliegues más pequeños y escalar a partir de ahí. El modelo de "comenzar poco a poco y pagar por uso" les permite invertir cantidades más pequeñas de capital por adelantado y adquirir experiencia a medida que crece la implementación. La sección siguiente trata los distintos tipos de arquitecturas de infraestructura disponibles actualmente y cómo cada una de ellas admite o no el enfoque de bloques de construcción.

Hipervisores

El hipervisor es una capa importante en el diseño de su infraestructura. Es directamente responsable de una parte saludable del rendimiento, disponibilidad, resiliencia y capacidad de gestión de su solución.

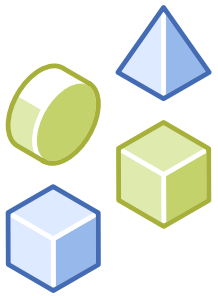
Panorama del hipervisor

En el entorno de virtualización actual, existe una amplia oferta de opciones de hipervisor listas para la empresa. La lista se reduce cuando nos enfocamos en hipervisores que habitualmente tienen casos de uso de EUC y VDI implementados. Se reduce a los siguientes:

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

Motivos para considerar el cambio

Hay varias razones por las cuales las empresas se planteen cambiar su hipervisor, que van desde la simplificación de la arquitectura y las operaciones de la capa del hipervisor hasta el aumento de la seguridad, la reducción de la dependencia de un solo proveedor y la eliminación de costos.



Criterios de evaluación

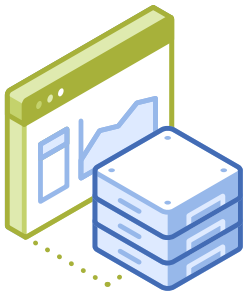
Los hipervisores son complejos elementos de software compuestos por cientos de características posibles. Las características básicas incluyen programación de recursos, alta disponibilidad, redes virtuales y compatibilidad con múltiples sistemas operativos. Actualmente estas son características habituales, que están fácilmente disponibles en todas las alternativas de hipervisores descritas anteriormente.

Más allá de esto, las empresas deben enfocarse en el valor que le ofrece al negocio cada alternativa y en cómo podría cambiar o mejorar las operaciones.

Una buena forma de empezar es observar la fase de diseño de las diferentes alternativas de hipervisor. Utilice los requisitos de diseño de un proyecto reciente o uno inminente como solución para la cual estaría diseñando. Debe desarrollar la comprensión de la cantidad necesaria de esfuerzos para diseñar una solución para el proyecto seleccionado. Observe si esta fase consume días o semanas de trabajo para diseñar la capa de hipervisor de la solución debido a la complejidad de las opciones de diseño. Idealmente, el hipervisor debería ofrecer toda la funcionalidad requerida y simplificar la fase de diseño, reduciéndola a unas pocas opciones claras y simples.

A continuación, asegúrese de comprender cómo es el esfuerzo de implementación. Basándose en el diseño propuesto seleccionado en su evaluación, identifique el esfuerzo necesario para implementar la solución y en qué aspectos difiere (si es que lo hace) de una implementación básica. La fase de implementación, al igual que la fase de diseño, idealmente debería requerir una aportación mínima del ingeniero y estar altamente automatizada. Implementar un nuevo clúster no es algo que deba requerir días o semanas de esfuerzo.

Por último, darle un vistazo a las operaciones del hipervisor ayudará a entender cualquier diferencia con respecto a sus esfuerzos actuales. Históricamente, parchear y actualizar un hipervisor ha sido uno de los mayores esfuerzos operativos. ¿Alguna de las alternativas de hipervisor ofrece alguna ventaja para simplificar este proceso en términos de esfuerzo y fiabilidad de las actualizaciones? Aparte de las actualizaciones, ¿cómo es la gestión diaria de las máquinas virtuales? ¿Se puede llevar a cabo desde una sola interfaz sencilla?



Alternativas de **infraestructura**

Actualmente existen tres alternativas de arquitectura principal para la virtualización de aplicaciones y equipos de escritorio, o soluciones EUC en general. Las alternativas son construir la propia (BYO), la infraestructura convergente (CI) y la infraestructura hiperconvergente (HCI).

Construir la propia

Como sugiere el nombre, en la alternativa de infraestructura BYO el arquitecto o el equipo elige de forma independiente los productos que más les gustan o que consideran los mejores. Esta alternativa genera un aumento significativo en el período inicial de planificación e investigación, ya que el equipo debe evaluar cada producto por separado y valorar cómo pueden o no trabajar juntos.

Esta alternativa también ofrece la capacidad de seleccionar y seguir una arquitectura de referencia que un proveedor ha publicado para el tipo de solución que se está construyendo. Estas arquitecturas de referencia suelen ser publicadas por un único proveedor y se enfocan en su producto. Estas arquitecturas de referencia personalizadas pueden ahorrar tiempo y reducir algunos riesgos, pero no siempre aplican a sus requisitos de diseño, casos de uso y entorno.

Como mínimo, una alternativa BYO para un diseño basado en EUC contendrá recursos de procesamiento y almacenamiento. Es posible que pueda utilizar la conectividad de red existente, por lo que puede que no sea un componente de esta alternativa. La figura 2 ilustra un ejemplo simple de las partes de una alternativa BYO. Al tener flexibilidad en el escalado, los costos son bastante predecibles; la única excepción sería en el lado del almacenamiento. Según el tamaño máximo de su diseño y de la opción de almacenamiento que elija, es posible que necesite varios dispositivos o cabinas de almacenamiento. A medida que escala el almacenamiento y necesita agregar una nueva matriz o dispositivo, el costo aumentará en esos puntos.

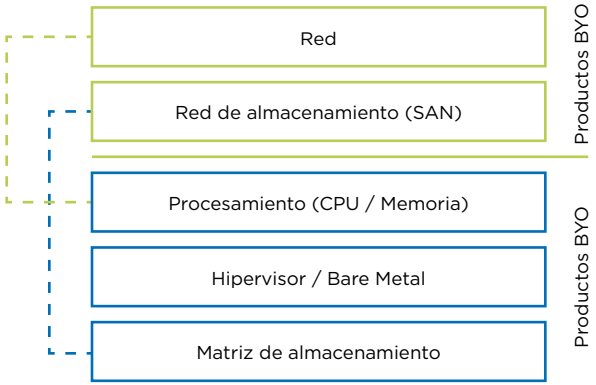
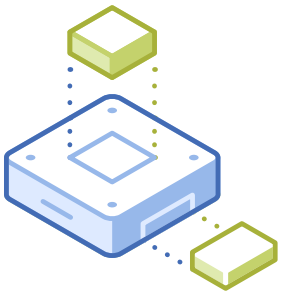


Figura 2:
Construir su propia infraestructura (BYO)

Siempre que esté reuniendo una serie de productos del mismo proveedor o de varios proveedores sin experiencia previa, existe un riesgo adicional. Habrá cierto nivel de incertidumbre sobre el rendimiento y la fiabilidad de la solución hasta el momento de adquirir la infraestructura real e implementarla en la arquitectura.

Si puede aceptar las incógnitas y el riesgo adicional, la alternativa BYO maximiza la flexibilidad. La capacidad de tomar casi cualquier decisión de proveedor y producto que sean capaces de trabajar juntos le permite permanecer con los proveedores existentes con los que ha tenido buenos resultados, y a la vez pasar a nuevos proveedores en otras áreas.

La alternativa BYO es capaz de escalar independientemente los recursos de procesamiento y almacenamiento. El único límite para el método de escalado o el tamaño máximo sería una restricción de la elección del producto individual. Debido a que los productos se adquieren por separado, no hay cantidades mínimas o fijas para el escalado de los productos. Esto permite flexibilidad al plantearse el enfoque de bloques de construcción descrito anteriormente.



Infraestructura convergente

La alternativa de infraestructura convergente (CI) es una arquitectura que se lanzó al mercado alrededor del 2010. Por lo general, las soluciones de infraestructura convergente ofrecen los mismos productos que podrían seleccionarse como parte de la alternativa BYO y los agrupan en una solución productizada. Esto significa que un proveedor de CI incluirá procesamiento, almacenamiento y redes en su oferta. Normalmente, la mayoría de las ofertas de CI suelen contener productos de varios proveedores y se incluyen como parte de una única oferta, o bien un proveedor puede ofrecer todas las capas de una oferta de CI de su propia línea de productos. La figura 3 ilustra un ejemplo simple de una alternativa de infraestructura convergente.

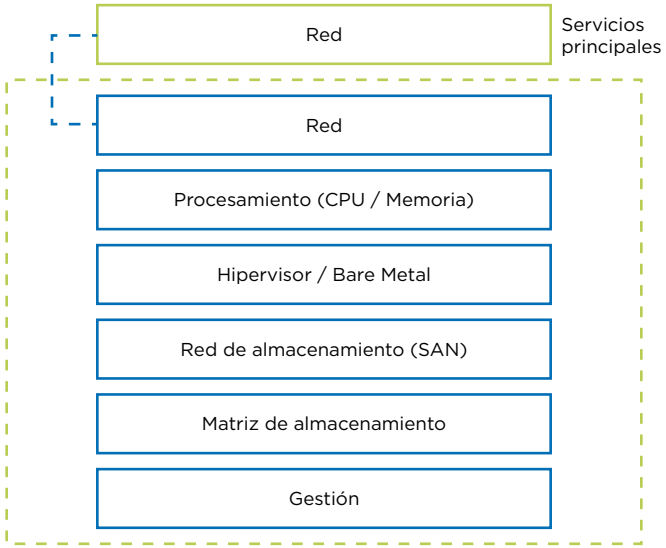
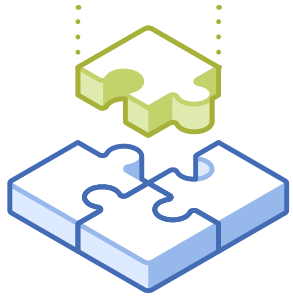


Figura 3:
Infraestructura convergente

Una solución de infraestructura convergente le permitirá comprar productos conocidos que han sido empaquetados en una única solución. Esto se puede considerar como una arquitectura de referencia que se puede comprar como un producto. Según el producto de CI que se evalúe, este puede ofrecer o no convergencia adicional, como lo haría si comprara los productos por separado en una alternativa BYO.



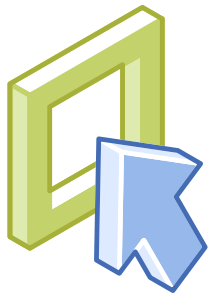
Por lo general, la mayoría de los proveedores y productos de CI ofrecerán la posibilidad de comprar todas las partes de la infraestructura en un solo SKU de producto. El proveedor de CI debe poder ofrecer soporte con una sola llamada para toda la solución de CI, lo cual significa que el proveedor de CI puede ofrecer soporte para todos los productos dentro de la solución. Este es un beneficio adicional, ya que permite a los clientes eliminar la necesidad de trabajar con múltiples proveedores en el proceso de resolución de problemas.

Con la mayoría de las ofertas de CI, hay un número limitado de productos ofrecidos dentro de la solución. Esto permite al proveedor de CI realizar una prueba previa y validar todas las piezas para garantizar que funcionen correctamente juntas, eliminando así gran parte del riesgo de la alternativa BYO.

Incluso después de varios años de venta de productos de CI en el mercado, los proveedores de CI han hecho poco para simplificar la gestión de estos productos. Con ofertas de CI que incluyen los mismos productos que las alternativas BYO, normalmente se gestionan ambas alternativas de una forma dispersa similar. Esta alternativa puede incluir la adquisición y/o algunos de los productos, pero normalmente no incluye la gestión operativa diaria de la solución.

Un producto de infraestructura convergente debería ser capaz de escalar los recursos que contiene de forma totalmente independiente. Esto significaría que solo puede agregar capacidad de procesamiento, aunque puede haber incrementos mínimos para el escalado. El otro recurso que se escalaría en un entorno de CI es el almacenamiento, y esto dependerá en gran medida del tipo de solución de almacenamiento seleccionada como parte del producto de CI. Un producto de infraestructura convergente tendrá un tamaño máximo, lo cual significa que tendrá un límite en el número de servidores que puede admitir y un límite de almacenamiento basado en la cabina de almacenamiento incluida.

Los límites de escalado de un producto de CI suelen ser bastante grandes, pero en algún momento del escalado los recursos dentro del producto de CI alcanzarán sus máximos. Llegado a este punto, para continuar escalando el diseño será necesario adquirir un producto de CI adicional. Esto causará grandes picos en los costos de infraestructura en diferentes puntos del proceso de escalado según el tamaño máximo de su diseño.



Infraestructura hiperconvergente

La arquitectura hiperconvergente se introdujo en el mercado aproximadamente un año después de la CI. Las verdaderas arquitecturas hiperconvergentes se consiguen mediante la convergencia de los recursos de procesamiento, recursos de almacenamiento y la capa de gestión en un solo producto. Es posible implementar una solución hiperconvergente en un modelo de solo software (SWO) o en un modelo de dispositivo que incluya hardware diseñado específicamente.

Al incluir un elemento de hardware como parte del producto, el proveedor ahora puede incluir la gestión de la infraestructura junto con los demás recursos que convergen en el producto. La figura 4 ilustra un ejemplo simple de una alternativa de infraestructura hiperconvergente.

La arquitectura de solo software puede proporcionar una gran flexibilidad en la capa de hardware, lo cual permite elegir la plataforma en la cual implementar. Al considerar opciones de SWO, deberían proporcionar una experiencia similar a la de un dispositivo. Esto se diferencia de las ofertas con una lista de compatibilidad de hardware débilmente acoplada que solo incorpora pruebas mínimas.

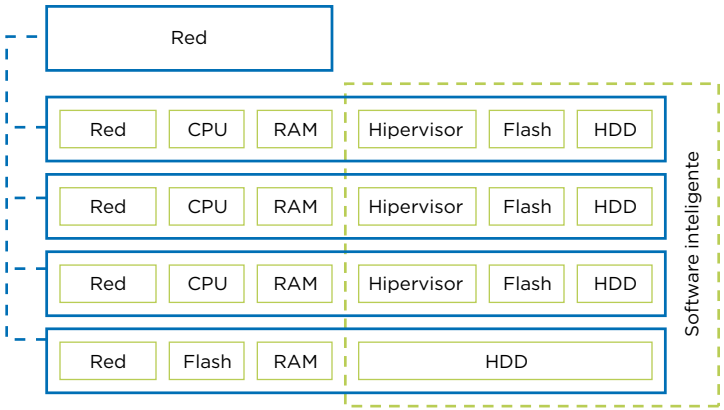
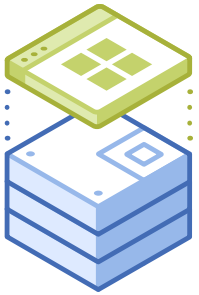


Figura 4:
Infraestructura hiperconvergente



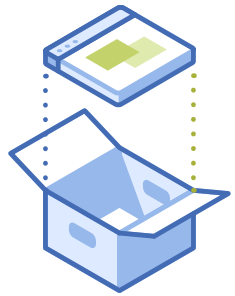
Instalación simple: los productos HCI líderes deben instalar nodos en cuestión de minutos u horas, no días o semanas, utilizando un proceso altamente automatizado.

Fácil escalabilidad: el producto debe ser fácil de escalar y desescalar. La adición de nuevos nodos al entorno debe realizarse de manera fácil y rápida a través de la interfaz de gestión.

Gestión moderna: una interfaz de gestión moderna debe enfocarse en la máquina virtual (VM) como punto de gestión. Un administrador debe poder ver el rendimiento de las máquinas virtuales, la cantidad de recursos que consume cada máquina virtual y si hay eventos o errores, además de ser capaz de extraer fácilmente informes basados en máquinas virtuales.

Extensibilidad: debe poder integrar fácilmente la infraestructura con otras partes de la solución y controlarla mediante programación. Esto requiere que el producto de HCI ofrezca una API y posiblemente otro método, como cmdlets de PowerShell. Con una API, podrá automatizar la comunicación y el control entre productos para reducir todavía más el esfuerzo y aumentar la precisión del entorno.

El rendimiento se ha excluido intencionalmente de la lista de beneficios de HCI porque todo el mundo espera que una solución moderna híbrida o basada en flash funcione bien. El objetivo de HCI es crear una capa de infraestructura que sea simple y eficiente. Evita que los equipos tengan que dedicar tiempo a hacer trabajos repetitivos, y proporciona valor adicional al negocio a nivel de automatización o aplicación.



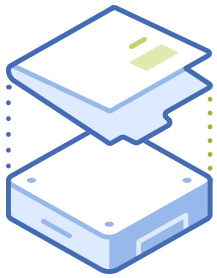
Requisitos de **almacenamiento**

Hay una serie de requisitos de recursos de almacenamiento diferentes que existen con cualquier diseño EUC. Deberá tener en cuenta las máquinas virtuales basadas en servidor, los datos de usuario y la infraestructura de escritorio virtual (VDI). Los requisitos de almacenamiento de VDI serán los más exigentes dentro del entorno, y también son los que hacen que la mayoría de los proyectos de VDI fracasen o sufran una mala experiencia.

Por esta razón, la sección de este libro electrónico dedicada al almacenamiento se enfoca en las necesidades del servicio VDI de la solución. Las necesidades de cada escritorio virtual a menudo pueden parecer pequeñas e insignificantes, pero cuando se combinan en grupos grandes al escalar el almacenamiento, los requisitos de rendimiento pueden sobrecargar fácilmente el almacenamiento que no fue diseñado correctamente para satisfacer estas necesidades.

Si cada escritorio virtual tiene un promedio de 15 IOPS y se esperan 2,000 usuarios simultáneos, eso equivale a 30,000 IOPS. Es un número bastante elevado, y podría abrumar la cabina de almacenamiento promedio. Pero no es conveniente diseñar la solución de almacenamiento para satisfacer el promedio de I/O del entorno. El diseño debe tener en cuenta los picos, incluyendo los inicios de escritorio y los eventos de inicio de sesión de los usuarios.

Las cargas de trabajo de escritorio virtual tienen muchos picos de I/O, lo cual las hace muy diferentes de otros tipos de cargas de trabajo dentro del centro de datos empresarial promedio. Por ejemplo, abrir una aplicación como Outlook por primera vez en una sesión puede generar hasta 1,000 IOPS para esa sesión de usuario. Eso supera con creces las 15 IOPS mencionadas anteriormente. En la figura 5 se muestra un ejemplo del impacto de IOP de distintas aplicaciones.



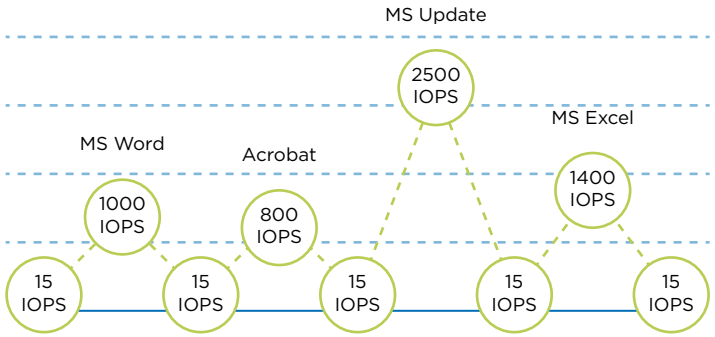
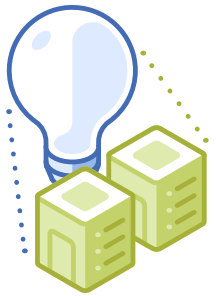


Figura 5:
VDI IOPS

Otros elementos de implementación y operativos, como la aplicación de parches y las actualizaciones del entorno, también pueden crear enormes picos en las IOPS y afectar al rendimiento si no se tienen en cuenta y se planifican en consecuencia. Si se implementan otros 50 escritorios virtuales, esa acción puede crear un pico significativo de I/O. Por estos motivos, la arquitectura de almacenamiento debe estar diseñada para adaptarse a las IOPS máximas de las operaciones de mantenimiento.

Hay varias formas de diseñar soluciones VDI con presentación de imágenes compartidas o clones completos, y cada una puede tener diferentes efectos en los requisitos de almacenamiento en términos de capacidad y rendimiento. Dado que los clones completos consumen capacidad y almacenamiento adicionales, la deduplicación será importante. Los clones completos también se deben parchar de forma independiente, lo cual aumentará la I/O durante esas operaciones.

El enfoque de imagen compartida que ofrece Citrix con MCS o PVS, y VMware con clones vinculados, presenta diferentes desafíos de I/O. Por naturaleza, estos enfoques de imagen compartida requieren menos capacidad de almacenamiento, ya que la imagen principal se comparte y cada escritorio virtual consume una cantidad menor de espacio para sus datos únicos. La imagen compartida tiene requisitos de rendimiento diferentes a los de la máquina virtual típica. Esta imagen ahora la utilizan cientos o miles de escritorios virtuales y debe poder generar grandes cantidades de IOPS. Si la imagen compartida es un cuello de botella, todos los escritorios virtuales que la utilicen se verán afectados negativamente y la experiencia del usuario será negativa.



Teniendo en cuenta estas consideraciones para picos y diferentes tipos de arquitecturas de virtualización de aplicaciones/escritorios, hay que seleccionar y diseñar una solución de almacenamiento que sea capaz de satisfacer los requisitos máximos de arranque, inicio de sesión y estado estable del entorno. Para comprender los requisitos de almacenamiento del diseño, se debe realizar una evaluación de escritorio en el entorno de PC físico existente. Esta evaluación de escritorio recopilará los detalles reales de rendimiento y capacidad de la base de usuarios para poder aplicarlos a los cálculos de diseño.

Una última reflexión sobre los requisitos de almacenamiento relacionados con la virtualización de aplicaciones/escritorio es que además de ser muy impredecibles en cuanto a I/O, las cargas de trabajo de escritorio también son muy pesadas. A diferencia de muchas cargas de trabajo de servidor que en su mayoría leen datos y se los entregan a los usuarios, los escritorios generalmente dedican más tiempo a escribir en el disco. En la cabina de almacenamiento, la escritura es más intensiva que la lectura. Una carga de trabajo normal de servidor podría ser del 80% de lecturas y del 20% de escrituras, mientras que la carga de trabajo de escritorio virtual de estado estacionario podría ser lo contrario. Al evaluar sus opciones de almacenamiento, asegúrese de prestar mucha atención a cómo la solución de almacenamiento almacena en búfer y asigna escrituras, en comparación con la promesa de que hace un "trabajo excelente" al almacenar en caché bloques de lectura común.

Tipos de **almacenamiento**

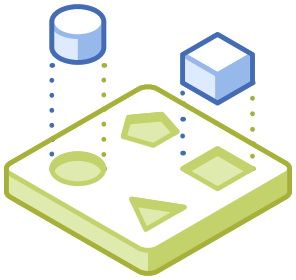
Existen varios tipos diferentes de almacenamiento. Las principales alternativas de almacenamiento disponibles actualmente son los sistemas de almacenamiento tradicionales por niveles, los sistemas flash híbridos y los sistemas todo flash. Cada alternativa adopta un enfoque diferente para proporcionar rendimiento y capacidad a las cargas de trabajo. Dentro de cada alternativa, los proveedores adoptan diferentes enfoques para crear sus productos. A continuación encontrará una breve explicación de cada una.

Arquitecturas tradicionales por niveles

Estos son los sistemas empresariales tradicionales que se han utilizado para cargas de trabajo basadas en servidor durante los últimos 10 a 20 años. Por lo general, son arquitecturas dobles basadas en controladores. En la última década, se han modificado para permitir incluir en la arquitectura múltiples niveles de rendimiento y capacidad de discos. Se proporcionan diferentes capas de discos para tratar de atender las demandas de capacidad y rendimiento de las diversas cargas de trabajo. Hay dos opciones en este enfoque: puede diseñar para el rendimiento creando grupos específicos de discos de alto rendimiento para una carga de trabajo, pero esto puede ser muy costoso y limitante. La otra opción es intentar aprovechar la nivelación que se agregó a esta arquitectura para pedir a la matriz que promueva o degrade los bloques de datos basados en la demanda. El problema con esta nivelación automática es que a menudo se necesita demasiado tiempo para tomar esas decisiones para las cargas de trabajo de VDI.

Todo flash

Los sistemas de almacenamiento todo flash están compuestos por un almacenamiento basado en flash. Hay muchos tipos diferentes de flash que se pueden utilizar dentro de estos sistemas de almacenamiento. Los sistemas todo flash modernos han sido diseñados para aprovechar las características del almacenamiento flash, lo cual significa que el sistema operativo y el sistema de archivos se han diseñado pensando en flash. Algunos productos han adoptado un diseño tradicional y simplemente han reemplazado los discos giratorios por todo flash. Aunque esta opción sigue siendo más rápida que la anterior, el producto final no fue diseñado para este fin.

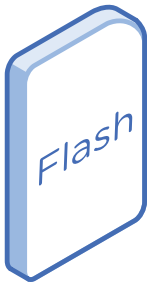


Los sistemas de almacenamiento todo flash son muy rápidos, con solo una capa de rendimiento en el producto. Para garantizar que el sistema también pueda proporcionar la capacidad necesaria para el diseño a un precio accesible, debe buscar sistemas que ofrezcan eliminación de datos duplicados y compresión. Aunque casi todos los sistemas modernos todo flash son más fáciles de gestionar que sus homólogos tradicionales, no siempre ofrecen la misma facilidad de gestión y gestión por máquina virtual que muchas de las ofertas de flash híbrida.

Flash híbrida

Los sistemas de almacenamiento híbrido son arquitecturas modernas que se diseñaron para usar de manera eficiente una combinación de unidades flash y discos giratorios. Los proveedores han adoptado diferentes enfoques arquitectónicos sobre cómo utilizan la capacidad y el rendimiento en sus sistemas, pero los resultados finales son similares. Todos pueden ofrecer un rendimiento impresionante con una cantidad menor de flash, a la vez que ofrecen una gran cantidad de capacidad mediante el almacenamiento de datos en grandes discos giratorios en cabina. Las alternativas ideales de arquitectura de almacenamiento híbrido utilizan inteligencia integrada para jerarquizar automáticamente los datos en unidades flash y de disco en base a la demanda, lo cual elimina la necesidad de ajustes manuales y posibles problemas de rendimiento.

Las arquitecturas que mejor se adaptan a un diseño VDI moderno son las arquitecturas de almacenamiento híbridas y todo flash. Estas arquitecturas son capaces de proporcionar el rendimiento requerido para los entornos VDI y, por lo general, también ofrecen las experiencias de gestión modernas descritas anteriormente. Las cargas de trabajo de VDI son muy impredecibles por naturaleza, y si su solución de almacenamiento debe esperar para tomar decisiones de almacenamiento o promover bloques a un nivel de almacenamiento en caché, la demanda de rendimiento habrá desaparecido mucho antes de que esto ocurra y la experiencia se habrá visto afectada negativamente.



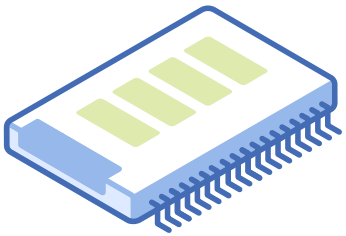
GPUs

Una unidad de procesamiento de gráficos (GPU) es una parte importante de una experiencia de escritorio moderna. Los equipos de escritorio y laptops físicos siempre han incluido una GPU, pero no era un aspecto muy importante a menos que fuese para jugar videojuegos o trabajar mucho con gráficos de gama alta. Aunque esto sigue aplicando, los sistemas operativos modernos y las nuevas cargas de trabajo, como el aprendizaje automático (ML) e inteligencia artificial (AI), también están aprovechando las GPUs para mejorar su rendimiento.

En el espacio de VDI, las GPUs no suelen ser necesarias y, por lo general, se evalúan según el caso de uso para ver si se requiere el rendimiento adicional en comparación con el equilibrio entre los costos adicionales y el valor percibido para los casos de uso con requisitos más bajos. Utilizar Windows 10 o ver streaming de video son ejemplos de casos de uso con requisitos más bajos que pueden beneficiarse de la GPU, pero pueden no ofrecer valor suficiente en comparación con la oferta de una experiencia de usuario aceptable sin GPU a un costo total más bajo. Sin embargo, al igual que con la fabricación de CPU, las GPUs están ganando rendimiento a un costo menor. El valor derivado de las GPUs debe ser parte de una evaluación general al diseñar un entorno EUC.

NVIDIA es el fabricante de tarjetas gráficas diseñadas para virtualización de escritorio más conocido y con mejor rendimiento. Las tarjetas gráficas de AMD funcionan solo en ciertos casos de uso y no ofrecen las mismas optimizaciones que las tarjetas NVIDIA.

Las tarjetas NVIDIA Grid permiten la virtualización de GPU, lo cual permite que varios usuarios compartan una única tarjeta gráfica. La virtualización de GPU no solo admite mayores densidades de usuario, sino que también ofrece un rendimiento nativo mientras accede a un escritorio virtual. Las GPUs NVIDIA también tienen un motor para la codificación H.264 que descarga procesos de la CPU, lo cual aumenta todavía más la densidad de usuarios en su hardware. Las tarjetas NVIDIA Grid suelen tener múltiples GPUs, lo cual mejora el escalado.



GPU dedicada

Con acceso directo a la GPU, puede crear una máquina virtual con una GPU dedicada. Esta configuración proporciona una experiencia de usuario comparable al uso de un cliente pesado con una tarjeta gráfica de gama alta. Sin embargo, asignar un núcleo de GPU a una sola máquina virtual, ya sea un escritorio compartido alojado (SBC) o un escritorio privado alojado (VDI), limita la escalabilidad.

GPU compartida

La tecnología Grid permite que varios escritorios virtuales compartan una GPU, mientras que ofrece la misma experiencia de usuario que las GPUs nativas. Este uso compartido se conoce comúnmente como vGPU y es una función del software Grid y del hipervisor. Una tarjeta NVIDIA Grid M10, por ejemplo, tiene cuatro núcleos de GPU físicos que pueden alojar hasta dieciséis usuarios por núcleo, lo cual resulta en 64 usuarios, cada uno con un escritorio habilitado para vGPU, por tarjeta M10.

La GPU procesa directamente los comandos gráficos de las máquinas virtuales, lo cual significa que los usuarios obtienen gráficos de gama alta sin una penalización de rendimiento debido a la interferencia del hipervisor. vGPU es más escalable que el acceso directo, ya que asignamos perfiles vGPU a nuestros usuarios, y por lo tanto conseguimos más usuarios en la misma tarjeta.

Los perfiles vGPU ofrecen memoria gráfica dedicada a través de vGPU Manager, que asigna la memoria configurada para cada escritorio. Un paquete de instalación de vSphere (VIB) instala vGPU Manager en el hipervisor. Con AHV, un administrador de paquetes RPM realiza esta tarea. Cada instancia de VDI tiene recursos preestablecidos basados en las necesidades de las aplicaciones.

Licencias Grid

Algo exclusivo de NVIDIA es que para utilizar la funcionalidad vGPU, se requiere una licencia. Hay diferentes niveles de licencias para aplicaciones virtuales que se utilizan para soluciones basadas en RDSH. El nivel de usuario avanzado y de diseñador para la mayoría de los casos de uso de VDI cubre aplicaciones de alto nivel, como aplicaciones de Adobe, ingeniería y aplicaciones CAD. Las licencias de Grid están disponibles en opciones de usuario nominal o concurrente.

Servicios de **archivos**

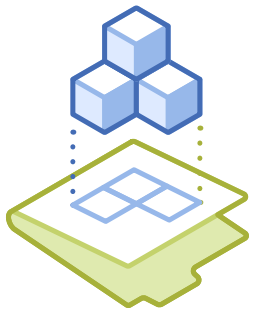
Los servicios de archivos se utilizan ampliamente y constituyen una capa importante en las arquitecturas de soluciones VDI. Tradicionalmente, los servicios de archivos se han proporcionado mediante un dispositivo NAS o un dispositivo basado en un servidor de Windows. Esto se presenta como un recurso compartido de SMB que consumen distintos servicios y dispositivos de sistema operativo invitado para el acceso compartido o privado.

Dentro de un diseño VDI, generalmente hay varios requisitos diferentes para los servicios de archivos. La gestión y la captura de perfiles de usuario son importantes para proporcionar una experiencia de usuario persistente. La mayoría de las soluciones de gestión de perfiles almacena los datos en recursos compartidos de SMB. También es muy común redirigir las carpetas como parte del perfil del usuario a un recurso compartido de SMB. Por último, el usuario ha creado datos, como documentos, medios e imágenes. Se almacenan en carpetas privadas o compartidas en recursos compartidos SMB.

Servicios de archivos de Nutanix

Nutanix ofrece Nutanix Files como característica nativa en la plataforma de nube de Nutanix. Files es una plataforma integrada de servicio de archivos escalable que se gestiona desde Prism junto con todas las demás funciones de Nutanix. Esto permite implementaciones sencillas con un solo clic y actualizaciones sin interrupciones. Esta simplificación permite a los equipos de VDI administrar más de la solución completa cuando lo deseen.

Files proporciona una arquitectura altamente escalable que permite agregar capacidad adicional a las máquinas virtuales de servicios de archivos con un solo clic, permitiendo capacidad adicional o conexiones de usuario. Las instancias de Files pueden ejecutarse en el mismo clúster de su servidor o máquinas virtuales VDI o en un clúster específico, si lo desea. Para permitir flexibilidad en la conectividad, los archivos son compatibles con las conexiones SMB 2.1, SMB 3.0 y NFS V3 y V4.



Dimensionamiento de **procesamiento**

Existen diferentes líneas de pensamiento sobre el dimensionamiento de la capa de procesamiento del diseño. La primera es el enfoque de escalado vertical, que utiliza menos hosts grandes para proporcionar recursos, mientras que el enfoque de escalado horizontal utiliza más hosts pequeños para proporcionar recursos. El método preferido se encuentra entre los dos enfoques, utilizando 2 hosts de socket y haciéndolos lo más densos posible sin violar las relaciones de consolidación establecidas como parte del diseño. Este libro electrónico se enfoca en ayudar a dimensionar los recursos de procesamiento para la carga de trabajo de VDI.

Hay tres cálculos principales en los que uno se enfocará al dimensionar los recursos de procesamiento en el diseño. Son la cantidad de memoria física en cada host, la frecuencia y cantidad de núcleos de CPU y su ratio de CPU.

Memoria física

En primer lugar, nunca se debe comprometer demasiado la memoria en un diseño de VDI. Romper esta regla solo sirve para ocasionar problemas de rendimiento en el entorno.

Cálculo de frecuencia de CPU

El cálculo de frecuencia de CPU depende en gran medida de los detalles recopilados en la evaluación de escritorio anterior. Los informes de la evaluación proporcionarán la cantidad de CPU que han utilizado las sesiones de usuario en promedio y durante los picos. Se utilizarán estos detalles junto con los detalles de la memoria de la evaluación para realizar los cálculos.

Límites de utilización del host

Un par de recomendaciones más de host y clúster de virtualización son: no superar nunca el 80% de utilización del host y dimensionar siempre su clúster para N+1. La utilización del host del 80% no es solo para implementaciones de virtualización de aplicaciones/escritorios, sino que es una recomendación que se aplica a cualquier carga de trabajo que se ejecute en un hipervisor. Si está ejecutando sus hosts más allá de la marca del 80%, tiene muy poco espacio para los picos y es posible que tampoco tenga suficiente margen de error de recursos para considerar una falla del host, según el tamaño de su clúster.

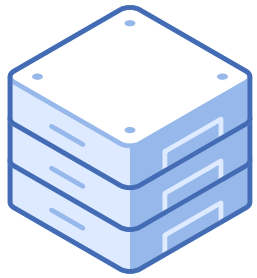
Dimensione siempre su clúster para N+1

El segundo elemento para calcular N+1 en el tamaño de su clúster es asegurarse de que haya suficientes recursos en su clúster para considerar una falla del host, para garantizar que todas las máquinas virtuales puedan seguir funcionando y que las que fallan se reinicien sin problemas. Una falla única del host es el nivel más común de resiliencia; hay un pequeño conjunto de clientes que requieren N+2 para dar considerar requisitos de SLA más altos.

Ratios de CPU

El elemento final del tema del dimensionamiento de procesamiento es el ratio de CPU, que se enfoca en el número de CPU virtuales y CPU físicas (vCPU:pCPU). Este ratio es muy importante, porque si se eleva demasiado, llegará a un punto en el que surgirá un problema de programación de CPU, lo cual afectará drásticamente el rendimiento y la experiencia del usuario. Cuando ocurre un problema de programación de CPU en los hosts de vSphere, aumenta la cantidad de tiempo de preparación de la CPU y esto le permite saber que el programador tiene problemas para programar todas las vCPU en las pCPU. Esto significa que la vCPU tendrá que esperar, aunque esté lista. El ratio de CPU es muy diferente para los diversos tipos de cargas de trabajo que se virtualizan en los clústeres VMware. Por lo general, las cargas de trabajo de servidor y de bases de datos tienen un ratio mucho menor, mientras que las cargas de trabajo de VDI pueden tener un ratio mayor.

El uso de vCPU no es un cálculo lineal, lo que significa que se puede crear un host que tenga un mayor ratio de consolidación si todas las máquinas virtuales tienen una sola vCPU. Si muchas máquinas virtuales tienen dos o más vCPU, esto afecta los cálculos. No es tan fácil como dividir entre dos para tener en cuenta el doble de vCPU. La figura 6 representa un rango que se ha demostrado que funciona con implementaciones reales de clientes. Los fabricantes que realizan pruebas sintéticas pueden mostrar ratios más altos. Hay que tener cuidado con estos ratios, ya que no siempre aplican a los diseños del mundo real.



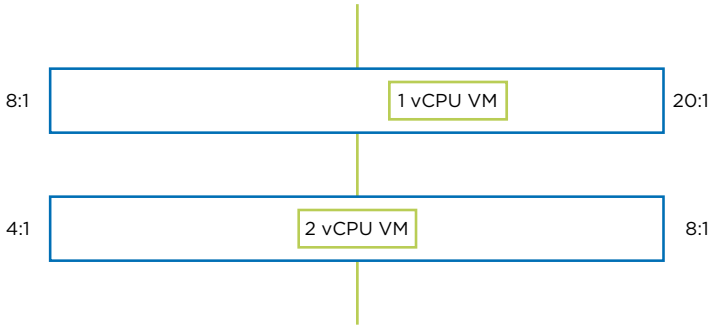
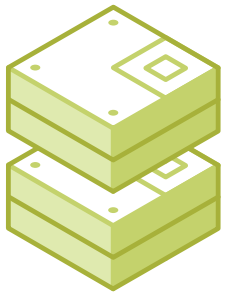


Figura 6:
La consolidación de VDI depende en gran medida del ratio de vCPU con el que se configurarán sus escritorios virtuales. El gráfico representa un rango que por experiencia ha demostrado ser seguro.

El rango de trabajo que funciona normalmente para escritorios virtuales vCPU únicos es de entre 8:1 y 20:1. Este es un rango muy amplio. El lugar preciso dentro de este rango depende de diferentes decisiones. Una sería lo grandes que son los hosts, el número de máquinas virtuales por host y el nivel de comodidad del cliente con ese número. Un ejemplo sería un host de doble socket con CPU duales de 18 núcleos. Esto podría acomodar más de 700 máquinas virtuales por el lado alto, lo cual le permite tener la cantidad adecuada de memoria y suficiente frecuencia disponible. Normalmente, tener tantas máquinas virtuales en un solo host asustaría a la mayoría de los clientes. En consecuencia, existen dos decisiones a realizar en este escenario. La primera es elegir una densidad más baja de lo que se limita artificialmente. Si se elige el extremo inferior del ratio, tendría 288 máquinas virtuales netas en el mismo host. La segunda opción sería elegir una CPU con menos núcleos, pero elegir un ratio intermedio. Si se elige una CPU de 12 núcleos y se utiliza un ratio de 12:1, se generarían 288 máquinas virtuales. Esta decisión suele ser una combinación de comentarios de los clientes, recomendaciones de la arquitectura y precios de la infraestructura. Se pueden obtener ahorros de costos significativos al elegir diferentes configuraciones físicas de CPU.

Los cálculos para un escritorio virtual de vCPU dual son similares, excepto que ahora se maneja el doble de la cantidad de vCPUs. El rango para operar aquí es entre 4:1 y 8:1. Algunos proveedores prometen más, pero estas recomendaciones se basan en implementaciones reales de los clientes. Se deben usar los mismos puntos de decisión que en el ejemplo anterior, solo que con un rango diferente de ratio de CPU.

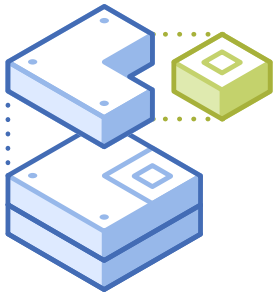


Otro aspecto a tener en cuenta es que si selecciona un ratio de CPU en medio de esos rangos, tendrá libertad para escalar verticalmente la densidad de consolidación en caso de que el entorno siga funcionando dentro de las tolerancias. Hay que tener en cuenta que no se pueden configurar estos ratios de CPU en ninguna otra herramienta actual. Estos son atributos que se deben declarar en el diseño, y que se convierten en puntos de datos a tener en cuenta en la gestión y el escalado del entorno. Al igual que con la memoria y la frecuencia, el ratio de CPU debe tenerse en cuenta al decidir agregar más máquinas virtuales a un clúster y cuándo agregar otro host a un clúster para proporcionar más recursos.

Se puede administrar el ratio de CPU a través de cálculos manuales mediante la recopilación de datos. Algunos administradores utilizan una secuencia de comandos de PowerShell, que recopilará datos y presentará el ratio como resultado de la secuencia de comandos. Con una secuencia de comandos, podría ejecutarse como un trabajo programado diariamente para asegurarse de que no se esté violando el ratio y no esté en peligro en ninguno de los clústeres.

La RAM o frecuencia de bus de memoria también está asociada con el dimensionamiento del procesamiento. La regla general al dimensionar la memoria es apuntar a la densidad más alta con los presupuestos de velocidad de bus más rápidos. El desafío al que se enfrenta constantemente la memoria es que una memoria más lenta puede dar lugar a ciclos de CPU inactivos en espera de que se completen las transacciones de lectura/escritura en la RAM.

La incorporación de GPUs en sus clústeres y diseño de VDI generalmente también afectará a la densidad de usuarios por host. Esto está directamente relacionado con el número de tarjetas GPU y el número de GPUs por tarjeta que se pueden colocar en su host deseado y luego con el perfil de vGPU seleccionado para sus usuarios. Ejemplo simple de un host que puede aceptar dos tarjetas GPU, cada una con una GPU. A continuación, se elige un perfil de vGPU que permite 16 usuarios por tarjeta, lo cual significa que solo 32 usuarios que necesitan GPU se ajustarían a ese tipo de host. Si hay CPU y memoria disponible en el host, todavía puede ejecutar máquinas virtuales que no tengan GPU. Identificar el perfil correcto de vGPU es importante para un dimensionamiento eficiente.

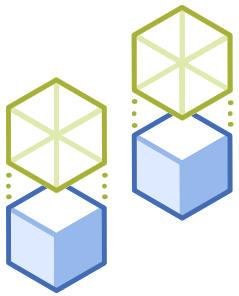


Diseño de clústeres de **virtualización**

Existen varias razones para crear diferentes clústeres de virtualización en un diseño EUC. Por lo general, la decisión de tener diferentes clústeres se basará en diferentes cargas de trabajo y tamaños de clústeres. En este libro electrónico no se dedica mucho tiempo a este tema, pero aquí tiene algunas recomendaciones que se basan en los temas tratados en otras partes del libro más amplio y en internet.

En primer lugar, al crear un diseño VDI de más de unos cientos de usuarios, es esencial separar la infraestructura de gestión de virtualización de la carga de trabajo de VDI. Esto significa que todos los servidores de gestión, agentes de VDI, servidores de archivos, servidores de gestión de aplicaciones y cualquier otra función que no sean escritorios virtuales deben ejecutarse en un clúster diferente.

Si el clúster de gestión necesita ser solo uno dedicado al diseño de EUC, dependerá del tamaño del entorno. Si el diseño es más pequeño, se pueden ejecutar máquinas virtuales de gestión en un clúster de virtualización de servidores existente. Es posible escalar estos clústeres de escritorio virtual hasta alcanzar un tamaño entre 16 y 32 hosts. Este rango permite crear un grupo de recursos más grande para que lo utilicen las máquinas virtuales, y también empuja a la mayoría de los clientes a adoptar un clúster más grande de lo habitual. Las actualizaciones recientes de hipervisor permiten clústeres de hasta 64 hosts, pero tomará algo de tiempo hasta que muchos arquitectos y clientes se sientan cómodos con ese tamaño. Si el entorno es lo suficientemente grande como para que el número de hosts supere estos rangos, se necesitará más de un clúster de VDI.



Otra razón por la cual se diseñaría para múltiples clústeres de virtualización, además del tamaño del entorno, sería para diferentes cargas de trabajo. Existen diferentes cargas de trabajo dentro de los clústeres de VDI. Si hay una cantidad significativa de escritorios virtuales de 1 vCPU y 2 vCPU, se debería diseñar un clúster separado para cada uno. La figura 7 ilustra un enfoque de diseño de múltiples clústeres. Esto permite gestionar el ratio de CPU de manera diferente en cada clúster, lo cual permite un diseño más fácil de gestionar. Si se quisiera mezclar las diferentes configuraciones de CPU, habría un nuevo ratio combinado a calcular, lo cual confundiría las cosas.

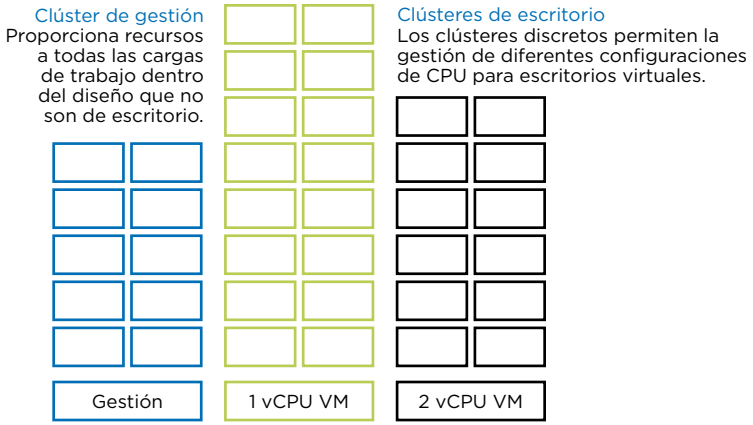


Figura 7:
Clústeres de gestión y escritorio

El uso de GPUs puede ser otro motivo para considerar un clúster específico para usuarios con GPU. Puede combinar usuarios con GPU y usuarios sin GPU en el mismo clúster, pero esto puede hacer que las operaciones y la programación de estos usuarios sean un poco más complejas. Si tiene suficientes usuarios con GPU para cumplir con los requisitos mínimos para un clúster pequeño, generalmente vale la pena hacerlo.

Ofrezca una experiencia de usuario excelente dondequiera que implemente

Explore las soluciones de Nutanix para End-User Computing en **nutanix.com/euc**

NUTANIX