GUÍA DEFINITIVA SOBRE

VDI en la infraestructura hiperconvergente

NUTANIX™

Brian Suhr [autor] tiene más de dos décadas de experiencia en TI en el diseño, implementación y administración de infraestructura empresarial. Ha aportado su experiencia en arquitectura e ingeniería en varios proyectos de virtualización, centros de datos, así como basados en la nube, al tiempo que trabaja con equipos técnicos de alto rendimiento en entornos de alcance global. Como autor independiente de los blogs DataCenterZombie y VirtualizeTips, Brian se centra en crear contenido que versa sobre la virtualización, la automatización, la infraestructura y la promoción de los productos y servicios que beneficien a la comunidad tecnológica. Sigue a Brian en Twitter: @bsuhr

Sachin Chheda [editor] es el director de soluciones y marketing vertical de Nutanix. Durante mucho tiempo ha trabajado con tecnología de la información, en puestos de ingeniería, gestión y marketing en las empresas más innovadoras del sector, desarrollando y comercializando productos y soluciones que impulsan a algunas de las empresas más grandes y con visión de futuro. Sigue a Sachin en Twitter: @StorSC

Lanzamiento oficial “1.0”

ÍNDICE

SOBRE ESTE LIBRO

Este libro trata sobre el diseño de infraestructura para VDI y entornos 'End User Computing' (EUC). El contenido de este libro ha sido extraído del capítulo sobre infraestructura del libro "Arquitectura y diseño de soluciones End User Computing", que se publicará próximamente.

INTRODUCCIÓN

Después de escoger la estrategia correcta y el proveedor de software para la entrega de servicios y aplicaciones EUC, la elección de infraestructura constituye la siguiente gran decisión para los proyectos de aplicaciones y virtualización de escritorios (VDI).

La infraestructura de procesamiento y almacenamiento es la base sobre la cual construir los servicios. Es algo parecido a la electricidad y el agua, contamos con ellos, y deberían funcionar cuando abrimos el grifo o presionamos el interruptor.

Sin una infraestructura estable, altamente disponible y de alto rendimiento, que sea subyacente al diseño, el departamento de TI se enfrentará a todo tipo de obstáculos durante la implementación y las fases operativas de su proyecto EUC. Esto refuerza la realidad de que la infraestructura es muy importante, pero que gastar una gran parte del tiempo de los equipos de TI en ella es perder el tiempo. Los arquitectos e ingenieros deben centrarse en proporcionar los servicios de EUC y aplicaciones, en lugar de gestionar la fontanería.

Hay varios factores importantes que deben considerarse en el proceso de diseño de infraestructura EUC. Al tener en cuenta estos factores junto con los requisitos de la organización, se está mejor capacitado para decidir cuáles serán las opciones de arquitectura. Deben tenerse en cuenta los siguientes factores al evaluar alternativas de arquitectura y opciones de proveedores en los proyectos EUC:

- Punto de partida
- Escalabilidad
- Rendimiento óptimo
- Supervisión
- Capacidad

PUNTO DE PARTIDA

El inicio o punto de partida para la infraestructura de un proyecto es, a menudo, una decisión crucial. Se trata de cuánta infraestructura e inversión requerirá la organización para iniciar la implementación de la virtualización y la entrega de aplicaciones / escritorios, en función de los diferentes tamaños de dicho punto de partida.

Si se planea que el proyecto llegue a 10.000 usuarios cuando esté completamente implementado, con una fase inicial de 5.000, la organización probablemente se escandalice menos por los costes iniciales. El razonamiento es que, dependiendo del tipo de infraestructura que se seleccione, no tiene sentido pensar en el coste por usuario hasta que se hayan implementado unos cuantos miles de ellos.

La otra cara de esto es que, si una organización quiere llegar a 10.000 usuarios, pero pretende comenzar con solo 500 y luego ir escalando a un ritmo constante a lo largo del proyecto, analizará más detenidamente el coste de la implementación inicial de la infraestructura a este tamaño antes de tomar una decisión mayor. El coste por usuario puede mantenerse estable a medida que el entorno escala, o puede verse realmente sesgado al principio, debido a un mayor gasto inicial en infraestructura.

Si bien el coste por usuario puede considerarse vago y casi irrelevante como factor para determinar los gastos de su infraestructura, es algo sobre lo que se le preguntará al intentar que su negocio acepte el proyecto o al justificar ante sus jefes su elección de infraestructura. Si elige una opción con un coste inicial más alto por usuario, debe estar preparado para explicar los detalles. Evalúe las soluciones que considere que serían más adecuadas para su entorno. De lo contrario, esté preparado para definir cuál será la decisión sobre cómo los costes se desarrollarán. Una muestra de estos dos escenarios se puede ver en el Gráfico 1.

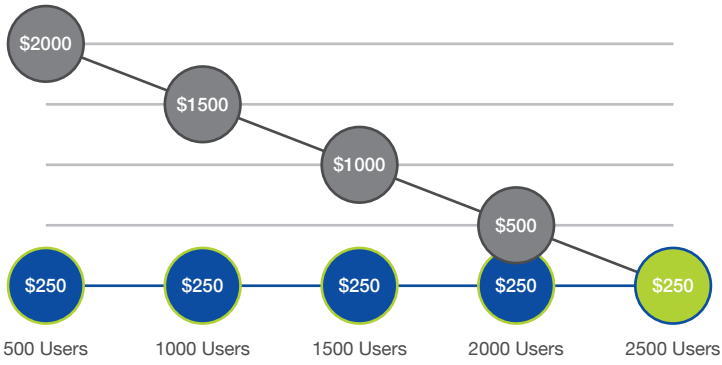

Gráfico 1
Puntos de partida por escritorio

ESCALABILIDAD

La escalabilidad de la arquitectura es un factor importante en la evaluación de la viabilidad del proyecto. Un arquitecto deberá comprender las opciones iniciales de tamaño según cada alternativa, lo que nos devuelve a la ya mencionada cuestión del punto de partida. ¿La opción escogida facilitará que el diseño comience en un tamaño tan pequeño como sea necesario? ¿O la organización tendrá que comprar más infraestructura de la que se necesita en un principio, manteniendo algunos recursos sin utilizar hasta que el proyecto alcance ese tamaño?

Además de desde qué tamaño puede comenzar nuestra opción, es igualmente importante considerar hasta qué tamaño puede llegar. Si se quiere comenzar en 500 y llegar a escalar a 10.000 usuarios, ¿cómo será el desempeño de nuestra opción en ambos extremos del espectro? ¿La organización estará contenta con ello en el punto más bajo? ¿Y en el más alto? ¿O en ambos?

El tema de la escalabilidad no es solo una cuestión de almacenamiento. También se aplica al procesamiento, las redes y muy probablemente a otras capas del diseño. Si se realizan ajustes en la configuración de la capa de procesamiento para lograr una menor densidad de máquinas virtuales por servidor host, ¿cómo podría afectar esto a las diferentes opciones de diseño al escalar? Por ejemplo, si el diseño inicial del host comienza con 128 GB de memoria por host y la opción final es de 256 GB o más, será necesario asegurarse de disponer de DIMM del tamaño adecuado para permitir que la configuración se escale en el futuro. Si se toman las decisiones incorrectas al principio para ahorrar costes, esto afectará a la densidad debido a las restricciones o costará más a largo plazo, ya que quedarán DIMM que no se pudieron reutilizar.

El arquitecto debe centrarse en cómo hacer que la solución se ponga en marcha a pequeña escala, y más adelante en hacerla alcanzar su máximo potencial. Pero tampoco se pueden ignorar todos los puntos intermedios, porque dependiendo de cómo se escale la implementación, podría haber muchas fases de escalado entre el inicio y el final. Lo ideal es buscar algo que permita que el diseño se escale fácilmente en sectores de archivo de cuentas de usuario identificadas por el proyecto, sin sobrepasar el plazo o las capacidades de implementación. El tamaño ideal del escalado de un sector para un proyecto puede encontrarse en incrementos de 100-200 usuarios. Pero si la opción de arquitectura elegida es mayor que esta, debe entender la manera en la que esto afecta sus costes e implementación.

RENDIMIENTO

El rendimiento de EUC medido por la experiencia del usuario final siempre se analiza cuidadosamente. La arquitectura seleccionada debe poder cumplir con lo requerido en cualquier fase del proyecto. Puede ser un proceso complicado, pero hay algunas alternativas. Si se reduce la escala de una solución para cumplir con los requisitos mínimos del usuario inicial, al no poder escalar linealmente, puede estar sacrificando rendimiento. Los arquitectos no quieren comprometer el diseño para adaptarse a este pequeño punto de partida porque puede afectar a las opciones de máximo rendimiento de la solución. Si invierte un poco de tiempo al principio para tomar la decisión correcta, puede evitar problemas más adelante.

El diseño de una solución de EUC generalmente presentará muchos requisitos de rendimiento diferentes. Seleccione una alternativa de arquitectura que sea lo suficientemente flexible como para cumplir con todos ellos dentro de una sola opción, bien porque el diseño proporcione varios tipos de servicios EUC o bien porque se centre solo en la virtualización de aplicaciones y escritorios. Se deben tener en cuenta sus múltiples necesidades. Comprender cómo cada alternativa podrá o no cumplir con los requisitos de rendimiento individuales afectará en gran medida su proceso de evaluación y diseño.

CAPACIDAD

El tema de la capacidad es similar al del rendimiento. Hay varios requisitos de capacidad diferentes dentro de los diseños EUC que deberán cumplirse. La solución requerirá la ejecución de máquinas virtuales de servidor, máquinas virtuales de escritorio, aplicaciones, perfiles de usuario y datos de usuario adecuados para esta arquitectura. Cada capa del diseño puede presentar requisitos de capacidad muy diferentes. Algunos usan grandes cantidades de datos que generalmente se deduplican bien. Otras partes, como los perfiles de usuario y los datos, consisten en pequeñas cantidades de datos comprimibles por usuario pero, multiplicadas por miles de usuarios, al final resultan ser enormes.

En el pasado, un problema que se producía a menudo era que se compraba demasiada o muy poca capacidad para intentar alcanzar los niveles de rendimiento requeridos. Estudie atentamente las opciones de arquitectura durante la fase de diseño para ver cómo podrán proporcionar la capacidad requerida, al tiempo que garantiza que también se cumplan los requisitos mínimos de rendimiento. La opción no debe proporcionar más de 2-3 veces la capacidad para cumplir con los requisitos de rendimiento de almacenamiento, ni añadir un rendimiento adicional significativo para cumplir con los requisitos de capacidad. La solución ideal es aquella que permita suficiente flexibilidad para escalar el rendimiento y la capacidad de forma más o menos paralela, de modo que ninguno de ellos se aleje demasiado del ritmo del otro.

En el pasado, este tema ha causado muchos debates y problemas. Muchas organizaciones se han encontrado con dificultades de planificación del rendimiento y de la capacidad al escalar la capacidad más rápido que el rendimiento. El hecho de que la solución tenga 5 TB de espacio libre no significa que se pueda escalar en otros 500 usuarios. Este escenario puede hacer que el rendimiento se resienta bastante. Este problema puede afectar a administradores y gerentes de TI que no tengan conocimientos sólidos sobre cómo se escala la solución.

SUPERVISIÓN

Aunque a menudo se pase por alto, la supervisión es muy importante. Cuando se trata de supervisar la infraestructura en un entorno EUC, los administradores generalmente se centran en el aspecto del rendimiento. Necesitan la capacidad de comprender lo que es normal y cuándo hay un problema activo.

Supervisar debería ser una tárea fácil pero al mismo tiempo proporcionar una gran cantidad de información detallada. Este no es el caso para muchos fabricantes, por lo que se debe observar de cerca cuál es la experiencia de supervisión con cada alternativa.

Otro requisito es la capacidad de proporcionar supervisión del rendimiento a nivel de máquina virtual. Desafortunadamente, la mayoría de los proveedores de infraestructuras todavía no pueden ofrecer este nivel de visibilidad en el entorno de virtualización. La capacidad de observar rápidamente la capa de almacenamiento y determinar si el problema del rendimiento del almacenamiento está a escala global o si está aislado en un host, en un grupo de máquinas virtuales o en una sola máquina virtual ya no es una opción.

Si se administra el rendimiento del almacenamiento en el nivel de VM, se puede utilizar un enfoque similar para administrar el rendimiento de la CPU y la memoria de una VM en el nivel de host. Los administradores deben saber si una máquina virtual está utilizando temporalmente un rendimiento adicional o si es una consumidora habitual de más rendimiento de almacenamiento que los usuarios típicos. Esto permitirá la comprensión de cuándo hay un pico y cuándo es necesario seguir indagando para identificar el problema.

Un bloque de construcción es un conjunto predefinido de infraestructura que se asigna a una cantidad específica de recursos o número de usuarios. Este enfoque es una de las mejores formas de abordar el diseño de infraestructura con la End User Computing.

Utilizando este enfoque, se puede desarrollar una arquitectura que ofrezca un modelo predictivo de escalado de costes, rendimiento y capacidad. Al determinar el tamaño del bloque de construcción, elija qué incrementos necesita para escalar los usuarios y cómo la selección de infraestructura puede ajustarse a estas opciones. Por ejemplo, se puede querer escalar usuarios en incrementos de 50 a 100 usuarios, pero la elección de infraestructura no se escala bien en un incremento tan pequeño. Esto puede obligar al diseño a escalar en incrementos mayores de 500 o 1.000 usuarios. Si la infraestructura elegida se escala en bloques grandes, se puede elegir escalar para encajar con este dato o simplemente aceptar el hecho de que los costes de infraestructura no se escalarán de la misma como los bloques de implementación del usuario. Esto significa que la organización compraría infraestructura en bloques de 1.000 usuarios y solo se implementaría en grupos de 50 a 100.

Al adquirir un bloque mayor para implementar una cantidad menor de usuarios, el precio de los escritorios virtuales o las sesiones de usuario parece dispararse. Esto se equilibrará cuando la organización implemente todos los usuarios planificados.

Las arquitecturas de estilo de bloques de construcción son útiles en cualquier proyecto de diseño, pero las implementaciones de EUC siempre tienen conjuntos comunes de usuarios, casos de uso con características similares y que se implementan en grupos. Para continuar con el ejemplo de un tamaño de bloque de 100 usuarios, comprendiendo los requisitos de recursos de esos 100 usuarios se puede garantizar que el bloque de infraestructura pueda proporcionar todo lo que esos usuarios requieren.

Si cada usuario requiere 15 IOPS en estado estable y 30GB de capacidad de almacenamiento, junto con 2GB de memoria y 200MHZ de CPU, el arquitecto sabe que los bloques de construcción deben proporcionar 1.500 IOPS, 3TB de capacidad, 200GB de memoria y 20GHZ de CPU. El arquitecto puede diseñar los bloques de construcción para contener recursos adicionales, pero ninguno de ellos puede estar por debajo de esos valores. Además, hay que evitar el desperdicio que supondría incluir demasiados recursos en un bloque y que luego no se puedan utilizar.

Con este enfoque y con un diseño granular, ahora se puede escalar el entorno en grupos más pequeños de 50-100 usuarios. Esto permite un enfoque lento y constante, con valores predecibles que permiten a las organizaciones planificar su implementación, rendimiento, capacidad y costes. Si las organizaciones quieren escalar más rápido y en grandes cantidades, solo tienen que agregar varios 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 comenzar con implementaciones más pequeñas y escalar desde allí. El modelo de ‘empezar pequeño y pagar por uso’ les permite invertir pequeñas cantidades de capital por adelantado y ganar experiencia a medida que crece la implementación. La siguiente sección cubre los diferentes tipos de arquitecturas de infraestructura disponibles en la actualidad y cómo cada una de ellas es o no compatible con el enfoque de bloques de construcción.

Actualmente hay tres opciones principales de arquitectura para la virtualización de aplicaciones / escritorios, o en general, soluciones EUC. Las opciones son Build Your Own (BYO, “Construya la suya propia”), Infraestructura convergente (CI) e Infraestructura hiperconvergente (HCI).

BUILD YOUR OWN

La opción de infraestructura BYO es exactamente lo que su nombre indica: el arquitecto o equipo eligen de forma independiente los productos que les gustan o creen que son los mejores. Esta opción da como resultado un aumento significativo en el período inicial de planificación e investigación, ya que el equipo debe evaluar cada producto por separado y cómo pueden o no trabajar en conjunto.

Esta opción también permite 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 generalmente las publica un solo proveedor y se centran en su producto. Estas arquitecturas de “hágalo-usted-mismo” (DIY) pueden ahorrar tiempo y reducir algunos riesgos, pero no siempre son aplicables a sus requisitos de diseño, casos de uso y entorno.

Como mínimo, una opción BYO para un diseño basado en EUC contendrá recursos de procesamiento y de almacenamiento. Es posible que pueda utilizar la conectividad de red existente, por lo que puede no ser un componente de esta alternativa. El Gráfico 2 ilustra con un sencillo ejemplo las partes de una opción BYO. Con flexibilidad en el escalado, los costes son bastante predecibles; la única excepción sería en el apartado del almacenamiento. Dependiendo del tamaño máximo de su diseño y la elección de almacenamiento realizada, puede requerir múltiples cabinas de almacenamiento o dispositivos. A medida que escale el almacenamiento y necesite añadir una nueva cabina o dispositivo, el coste aumentará.

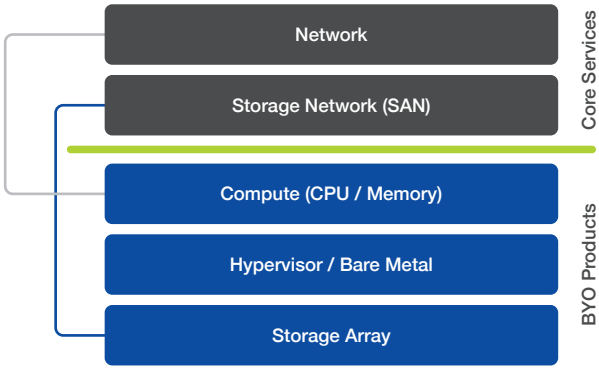

Gráfico 2
Infraestructura Bring Your Own (BYO)

Montar sin experiencia previa una serie de productos del mismo proveedor o de múltiples proveedores implica un riesgo. No tendrá seguridad acerca del rendimiento y confiabilidad de la solución hasta el momento de adquirir e implementar la infraestructura.

Si puede aceptar las incógnitas y el riesgo adicional, la opción BYO maximiza la flexibilidad. Dado que puede tomar casi cualquier decisión en cuanto a proveedor y producto que sea capaz de trabajar en conjunto, esto le permite permanecer con los proveedores con los que ha tenido una buena relación, mientras se muda a nuevos proveedores en otras áreas.

La opción BYO es capaz de escalar los recursos de procesamiento y de almacenamiento de forma independiente. El único límite para el método de escalado o el tamaño máximo sería una restricción de la elección individual del producto. Dado que estos productos se compran por separado, no hay mínimos o cantidades establecidas para escalarlos. Esto permite flexibilidad al intentar contabilizar el enfoque de bloques de construcción mencionado anteriormente.

INFRAESTRUCTURA CONVERGENTE

La infraestructura convergente (CI) es una arquitectura que se lanzó al mercado alrededor de 2010. Las ofertas de infraestructura convergente generalmente ofrecen los mismos productos que forman parte de la opción BYO, agrupándolos en una solución productiva. Esto significa que un proveedor de CI incluirá en su oferta ordenadores, almacenamiento y redes. Por lo general, la mayoría de las ofertas de CI contendrán productos de múltiples proveedores y se incluirán como parte de una sola oferta; otra opción es que el proveedor ofrezca todas las capas de una oferta de CI desde su propia línea de productos. El Gráfico 3 ilustra con un sencillo ejemplo una opción de infraestructura convergente.

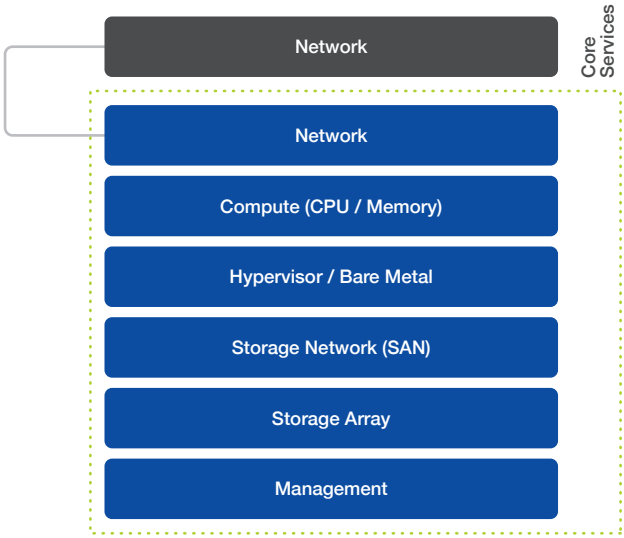

Gráfico 3
Infraestructura convergente

Una oferta de infraestructura convergente le permitirá comprar productos que le resulten familiares, reunidos en una única solución. Esto puede pensarse como una arquitectura de referencia adquirible como producto. Dependiendo del producto de CI que se evalúe, el producto puede ofrecer o no más convergencia que si hubiera comprado los productos por separado en una opción BYO.

Por lo general, la mayoría de los proveedores y productos de CI le ofrecerán la posibilidad de adquirir todas las partes de infraestructura en un único SKU de producto. El proveedor de CI debe poder ofrecer soporte general para toda la solución de CI, lo que significa que puede proporcionar soporte a todos los productos dentro de la solución. Esto supone una ventaja adicional, ya que permite a los clientes eliminar la necesidad de trabajar con múltiples proveedores en el proceso de solución de problemas.

En la mayoría de las ofertas de CI se ofrece una cantidad limitada de productos. Esto permite que el proveedor de CI realice una prueba previa y valide todos los componentes para garantizar que funcionen correctamente juntos, eliminando gran parte del riesgo existente en la opción BYO.

Incluso tras varios años de venta de productos de CI, los proveedores no han trabajado mucho para simplificar la administración de dichos productos. Con las ofertas de CI que incluyen los mismos productos que las opciones BYO, normalmente se administrarán ambas alternativas de manera similar y dispersa. Esta alternativa puede combinar el producto adquirido con algunos de los demás productos, pero generalmente no lo hace con la gestión operativa diaria de la solución.

Un elemento de la infraestructura convergente debería poder escalar los recursos en ella de manera independiente. Esto significaría poder añadir únicamente procesamiento, aunque pueda haber incrementos mínimos en los que poder escalar. El otro recurso que se escalaría en una oferta de CI es el almacenamiento, que dependerá en gran medida del tipo de solución seleccionada para el mismo como parte de dicha oferta. Un elemento de infraestructura convergente tendrá un tamaño máximo, lo que significa que tendrá un límite en la cantidad de servidores que puede admitir, así como en la capacidad de almacenamiento, basado en las cabinas existentes.

Los límites de escalado de una oferta de CI suelen ser bastante amplios, pero, a medida que se escala, se alcanzará el máximo. Para continuar a partir de este punto, será necesario comprar un producto CI adicional, lo que provocará grandes picos en los costes de infraestructura en diferentes puntos del proceso de escalado.

INFRAESTRUCTURA HIPERCONVERGENTE

La arquitectura hiperconvergente se introdujo en el mercado aproximadamente un año después de la CI. Las auténticas arquitecturas hiperconvergentes se logran mediante la convergencia de los recursos de procesamiento, los recursos de almacenamiento y la capa de administración en un solo producto unificado. Es posible implementar una solución hiperconvergente en un método de arquitectura BYO o de referencia, pero para ser realmente hiperconvergente, el producto debe incluir el dispositivo de hardware.

Esto permite al proveedor incluir la administración de la infraestructura junto con los otros recursos que están convergiendo en el producto. En el Gráfico 4 se ilustra con un sencillo ejemplo una alternativa de infraestructura hiperconvergente.

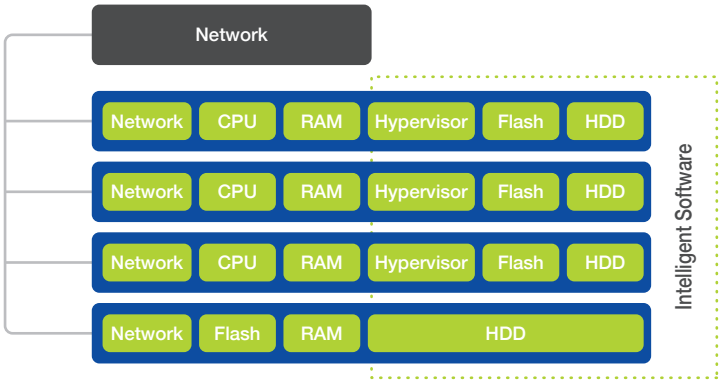

Gráfico 4
Infraestructura hiperconvergente

Un producto verdaderamente hiperconvergente ofrece una serie de ventajas que otras arquitecturas de referencia no pueden ofrecer:

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

Escalabilidad fácil: el producto debe ser fácil de escalar hacia arriba o hacia abajo. La adición de nuevos nodos al entorno debe suceder de forma fácil y rápida a través de la interfaz de administración.

Gestión moderna: una interfaz de gestión moderna debe centrarse en la máquina virtual (VM) como punto de gestión. Un administrador debe ser capaz de comprender cómo funcionan las máquinas virtuales, la cantidad de recursos que consume cada máquina virtual, si hay eventos o errores, y proporcionar la capacidad 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 HCI ofrezca una API, además de otros métodos posibles como los comandos de PowerShell. Con una API, usted podrá automatizar la comunicación y el control entre productos para reducir aún más el esfuerzo y aumentar la precisión del entorno.

El rendimiento se ha omitido intencionalmente de la lista de ventajas de HCI porque todo el mundo espera que una solución híbrida o flash moderna funcione bien. La hiperconvergencia debería crear una capa de infraestructura sencilla y eficiente, que permitiera a los equipos dejar de pasar el tiempo dándole a los botones posibilitándoles proporcionar un valor adicional a la empresa a nivel de automatización o aplicación.

Los distintos diseños de EUC necesitan diferentes recursos de almacenamiento. El diseño deberá tener en cuenta las máquinas virtuales basadas en servidor, los datos de usuario y la infraestructura de escritorio virtual (VDI, también conocida como máquinas virtuales de usuario). Los requisitos de almacenamiento asociados serán los más exigentes dentro del entorno y también los que provoquen que la mayoría de los proyectos fallen o generen una mala experiencia.

Por esta razón, el apartado sobre almacenamiento de este eBook se centra en las necesidades del servicio VDI de la solución. A menudo, las necesidades de cada escritorio virtual pueden parecer pequeñas y de poca importancia, pero cuando se combinan en grandes grupos a medida que escalamos el almacenamiento, las demandas de rendimiento pueden fácilmente sobrepasarlo si no fue diseñado para satisfacer estas necesidades adecuadamente.

Si cada escritorio virtual tiene 15 IOPS de promedio a una latencia razonable y se esperan 2.000 usuarios concurrentes, eso equivale a 30.000 IOPS. Ese número es bastante grande y podría sobrepasar la cabina de almacenamiento promedio. Pero no se puede simplemente diseñar una solución de almacenamiento para cumplir con la media de entradas/salidas del entorno; el diseño debe tener en cuenta los picos, incluidos los arranques de escritorio y los eventos de inicio de sesión del usuario.

La carga de trabajo de un escritorio virtual es muy diferente de otros tipos de cargas de trabajo que se ejecutan dentro de un centro de datos empresarial medio, ya que presentan muchos picos de entrada/salida. Por ejemplo, abrir una aplicación como Outlook por primera vez en una sesión puede generar más de 1.000 IOPS para esa sesión de usuario. Eso está más allá del promedio de 15 IOPS del que hablamos anteriormente. En el Gráfico 5 se muestra un ejemplo del impacto de la IOP de diferentes aplicaciones.

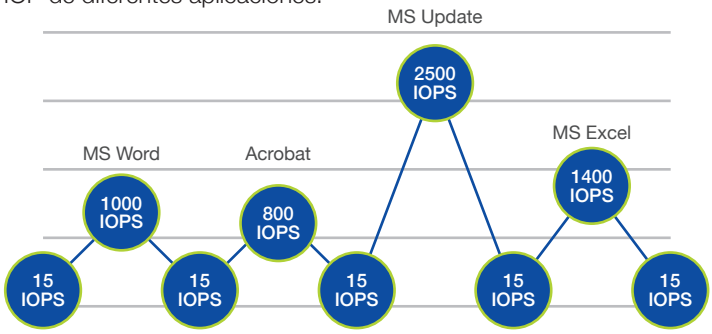

Gráfico 5
IOPS en VDI

Otros elementos operativos y de implementación, como parches y actualizaciones del entorno, también pueden crear picos tremendos en IOPS y afectarán el rendimiento si no se tienen en cuenta y se planifican en consecuencia. Si se implementan otros 50 escritorios virtuales, esa acción puede causar un aumento significativo en entradas/salidas. Por estos motivos, será necesario tener en cuenta las operaciones de mantenimiento en la arquitectura de almacenamiento para picos en IOPS.

Hay varias formas de diseñar soluciones VDI, con clones completos o imágenes compartidas, 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 deben parchearse de forma independiente, lo que aumentará las entradas/salidas durante esas operaciones.

El enfoque de imagen compartida que ofrece Citrix con MCS o PVS y VMware con clones vinculados presenta diferentes desafíos de entrada/salida. En esencia, estos enfoques de imágenes compartidas requieren menos capacidad de almacenamiento, ya que la imagen principal se comparte y cada escritorio virtual solo consume una cantidad menor de espacio para sus datos únicos. La imagen compartida presenta diferentes requisitos de rendimiento respecto a la VM típica. Esta imagen ahora es utilizada por cientos o miles de escritorios virtuales y debe poder generar grandes cantidades de IOPS para manejar situaciones como tormentas de arranque. Si la imagen compartida es un cuello de botella, todos los escritorios virtuales que la usen se verán afectados negativamente y la experiencia del usuario será mala.

Teniendo en cuenta estas consideraciones para los picos y los diferentes tipos de arquitecturas de virtualización de aplicaciones / escritorios, se debe seleccionar y diseñar una solución de almacenamiento que sea capaz de cumplir con las demandas máximas 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 físico del PC. Ésta reunirá los detalles reales de rendimiento y capacidad de la base de usuarios para que pueda aplicarlos a los cálculos de diseño.

Una última reflexión sobre los requisitos de almacenamiento relacionados con la virtualización de aplicaciones / escritorios es que las cargas de trabajo de escritorio, aparte de ser muy imprevisibles desde el punto de vista de las entradas/salidas, también son muy pesadas. A diferencia de muchas cargas de trabajo de servidores que en su mayoría leen datos y se los presentan a los usuarios, los escritorios suelen pasar más tiempo escribiendo en el disco. Las escrituras son más intensivas para la cabina de almacenamiento que las lecturas. Una carga de trabajo típica del servidor podría ser 80% de lectura y 20% de escritura, mientras que la carga de trabajo de escritorio virtual de estado estable podría ser al contrario. Al evaluar sus opciones de almacenamiento, asegúrese de prestar mucha atención a cómo la solución de almacenamiento amortigua y confirma las escrituras, en lugar de centrarse en la promesa de que el almacenamiento hará un "trabajo excelente" para almacenar en caché los bloques comúnmente leídos para manejar múltiples arranques.

TIPOS DE ALMACENAMIENTO

Hay varios tipos diferentes de almacenamiento. Las principales opciones disponibles en la actualidad son cabinas de almacenamiento en capas tradicionales, cabinas flash híbridas y cabinas totalmente flash. Cada opción adopta un enfoque diferente para proporcionar rendimiento y capacidad a las cargas de trabajo y, para cada una de ellas, los proveedores adoptan diferentes enfoques al crear sus ofertas. A continuación se incluye una breve explicación de cada una.

Arquitecturas tradicionales por capas: estas son las cabinas empresariales tradicionales que se han utilizado para cargas de trabajo basadas en servidor durante los últimos 10-20 años. Por lo general, son arquitecturas basadas en dos controladores. En la última década, se han modificado para permitir que se incluyan en esta arquitectura múltiples capas de discos de rendimiento y capacidad. Se proporcionan diferentes capas de discos para probar y atender las demandas de capacidad y rendimiento de las cargas de trabajo dispersas. Hay dos opciones en este enfoque: para una determinada carga de trabajo puede enfocar el diseño en el rendimiento, creando grupos dedicados de discos de alto rendimiento, pero esto puede ser muy costoso y limitante. La otra opción es tratar de aprovechar las capas que se añadieron a esta arquitectura para exigirle a la cabina que promueva o degrade bloques de datos según demanda. El problema con esta distribución automática de capas es que a menudo se invierte demasiado tiempo en tomar esas decisiones para las cargas de trabajo de VDI.

All-Flash: las cabinas de almacenamiento All-flash se componen completamente de almacenamiento basado en flash. Hay muchos tipos diferentes de flash que pueden ser utilizados en ellas. La cabina all-flash moderna se diseñó para aprovechar las características del almacenamiento flash, lo que significa que el sistema operativo y el sistema de archivos se diseñaron para ello. Algunos productos han tomado un diseño de cabina tradicional y se han limitado a reemplazar los discos por flash. Si bien obtenemos mayor rapidez que con la opción anterior, el producto final no fue diseñado para este propósito.

Las cabinas de almacenamiento all-flash son muy rápidas, con solo un nivel de rendimiento. Para garantizar que la cabina proporcione la capacidad requerida para el diseño a un precio asequible, debe buscar cabinas que ofrezcan deduplicación y compresión. Si bien casi todas la cabinas all-flash actuales son más fáciles de administrar que sus homólogas heredadas, no siempre ofrecen la misma facilidad de gestión y gestión por VM que ofrecen muchas cabinas de flash híbrido.

Flash híbrido: las cabinas de almacenamiento híbrido son arquitecturas modernas diseñadas para usar de manera eficiente una combinación de unidades flash y discos giratorios. Los proveedores han adoptado diferentes enfoques de arquitectura sobre cómo usar la capacidad y el rendimiento en sus cabinas, aunque los resultados finales son similares. Todas ofrecen un rendimiento impresionante con una menor cantidad de flash, a la vez que proporcionan una gran cantidad de capacidad almacenando datos en discos giratorios grandes en la cabina. Las opciones ideales de arquitectura de almacenamiento híbrido utilizan inteligencia integrada para agrupar datos automáticamente mediante unidades flash y de disco según demanda, eliminando la necesidad de ajustes manuales y posibles dificultades de rendimiento.

Las arquitecturas que mejor se ajustan a un diseño de VDI moderno son arquitecturas de almacenamiento híbrido y all-flash. Estas arquitecturas son capaces de proporcionar el rendimiento requerido para los entornos VDI y, por lo general, también ofrecen las experiencias de administración modernas de las que hemos hablado anteriormente. Las cargas de trabajo de VDI son muy impredecibles por naturaleza y, si su solución de almacenamiento debe esperar para tomar decisiones o promover bloques a un nivel de almacenamiento en caché, la demanda de rendimiento habrá desaparecido antes de que suceda y la experiencia se verá afectada negativamente.

Existen varios métodos distintos para calcular la capa de procesamiento del diseño. El primero se basa en la ampliación y utiliza menos hosts grandes para proporcionar recursos, mientras que el método de disminución emplea más hosts pequeños para proporcionar recursos. El método ideal se encuentra en algún punto intermedio, en un enfoque que utiliza 2 hosts de socket y los hace lo más densos posible sin violar las relaciones de consolidación establecidas como parte del diseño. Este libro le ayuda a calcular los recursos informáticos para la carga de trabajo de VDI.

Hay tres cálculos principales en los que centrarse al calcular los recursos informáticos en diseño: la cantidad de memoria física en cada host, la cantidad de velocidad de reloj de la CPU y la cantidad de núcleos de CPU y la ratio de CPU para ellos. En primer lugar, nunca se debe comprometer demasiado la memoria en un diseño VDI. Incumplir esta regla tiene muy pocas ventajas y solo provocará problemas de rendimiento en el entorno.

El cálculo de la velocidad del reloj de la 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 las sesiones de usuario usaron de media y en picos. Se usarán esos detalles junto con los de la memoria de la evaluación para hacer los cálculos.

Hay dos recomendaciones cruciales sobre el clúster de host y la virtualización. En primer lugar, el uso del host no debe exceder nunca el 80%. En segundo lugar, siempre hay que dimensionar el clúster para N+1. El 80% de la utilización del host no es solo para implementaciones de virtualización de aplicaciones / escritorios, sino que la recomendación 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 la sobrecarga de recursos suficiente para dar cuenta de un fallo del host, dependiendo del tamaño de su clúster. El segundo elemento de cálculo para el N + 1 en el tamaño de su clúster es asegurarse de que haya suficientes recursos en éste como para dar cuenta de un solo error de host, garantizando así que todas las máquinas virtuales puedan seguir ejecutándose y las que fallen se reiniciarán sin problemas. Un único fallo del host es el nivel más común de resistencia; hay una pequeña cantidad de clientes que requieren N + 2 para cubrir requisitos de SLA más exigentes.

El último elemento a tener en cuenta en relación con el cálculo del procesamiento es la ratio de CPU, que se centra en la proporción entre CPU virtuales y CPU físicas (vCPU: pCPU). Esta relación es muy importante porque si esta proporción es muy elevada llegará a un punto donde surgirá un problema de programación de la CPU, lo que afectará drásticamente el rendimiento y la experiencia del usuario. Cuando ocurre un problema de programación de la CPU en los hosts vSphere, la cantidad de tiempo de preparación de la CPU aumenta, lo que le permite saber que el planificador está teniendo problemas para programar todas las vCPU en las pCPU. Esto significa que la vCPU tendrá que esperar, aunque esté lista. La ratio de CPU es muy diferente para los diversos tipos de cargas de trabajo que se virtualizan en clústeres de VMware. Por lo general, las cargas de trabajo del servidor y la base de datos tienen una proporción mucho menor, mientras que las cargas de trabajo VDI pueden tener una proporción más alta.

El uso de vCPU no es un cálculo lineal, lo que significa que se puede construir un host que tenga una ratio de consolidación más alta si todas las máquinas virtuales disponen solo de una vCPU. Cuando muchas máquinas virtuales tienen 2 o más vCPU, esto afectará los cálculos. No es tan fácil como dividir entre 2 para dar cuenta del doble de vCPU. El Gráfico 6 representa una ratio que ha demostrado funcionar con implementaciones reales de clientes. Los fabricantes que hacen pruebas teóricas pueden mostrar proporciones más altas. Se debe tener cuidado con estas pruebas, ya que no siempre son reproducibles en los diseños del mundo real.

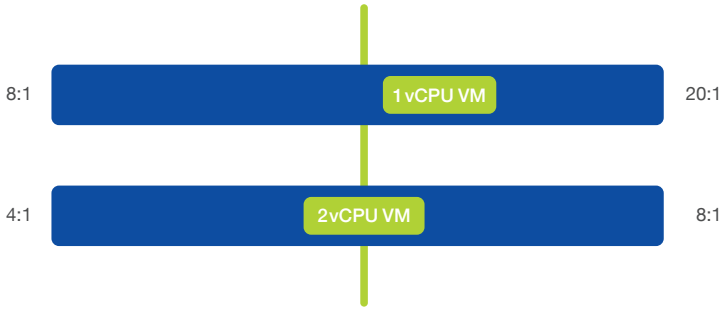

Gráfico 6
La consolidación de VDI depende en gran medida de la proporción de vCPU con la que se configurarán sus escritorios virtuales. Esta tabla representa el rango que, por experiencia, se sabe que es seguro.

La ratio de trabajo normal para escritorios virtuales de vCPU individuales es entre 8: 1 y 20: 1. Este es un rango amplio, por lo que la cantidad exacta dependerá de diferentes alternativas, como el tamaño de los hosts, la cantidad de máquinas virtuales por host o 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 tirando por lo alto, siempre que tenga la cantidad correcta de memoria y suficiente velocidad de reloj disponible. Por lo general, tener tantas máquinas virtuales en un solo host asustaría a la mayoría de los clientes. En este escenario, hay que tomar dos decisiones importantes. En primer caso, hay que elegir una densidad más baja limitada artificialmente.

Si se elige la ratio menor, se vincularían 288 VM en el mismo host. La segunda opción sería elegir CPU con menos núcleos, pero elegir una ratio en alguna cantidad intermedia. Si elige 12 CPU centrales y usa una proporción de 12: 1, se vincularían 288 máquinas virtuales. Esta decisión suele ser una combinación de feedback del cliente, recomendaciones de arquitectos y precios de infraestructura. Puede haber importantes ahorros de costes al elegir diferentes configuraciones físicas de CPU

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

Otra cosa a tener en cuenta es que, si selecciona una ratio de CPU en la zona media, se podrá escalar la densidad de consolidación hacia arriba si el entorno continúa funcionando dentro de las tolerancias. Una cosa a tener en cuenta es que hoy en día no hay lugar para configurar estas relaciones de CPU en ninguna otra herramienta. Estos son atributos que deben declararse en el diseño y convertirse en puntos de datos que se tendrá que tener en cuenta en la gestión y escala del entorno. Al igual que la memoria y la velocidad del reloj, la ratio de CPU debe calcularse si se decide añadir más máquinas virtuales a un clúster, así como cuándo agregar otro host a un clúster para proporcionar más recursos.

Se puede administrar la ratio de CPU mediante cálculos manuales recopilando datos. Algunos administradores usan un script de PowerShell que recopila datos y presenta esta ratio. Con un script, podría ejecutarse diariamente como un trabajo programado para garantizar que no se está alterando la ratio ni poniendo en peligro a ninguno de los grupos.

La RAM o frecuencia del bus de memoria también está asociada con el tamaño de la capacidad de procesamiento. La regla general al dimensionar la memoria es apuntar a la densidad más alta con los presupuestos de velocidad de bus más rápidos que lo permitan. El desafío que a menudo presenta la memoria es que, si es más lenta, puede dar como resultado ciclos inactivos de la CPU que esperan que se completen las transacciones de lectura / escritura en la RAM.

Hay varias razones para crear diferentes clústeres de virtualización en un diseño EUC. La decisión de tener diferentes grupos suele diferentes cargas de trabajo y tamaños de clúster. En este eBook no se le dedica mucho tiempo a este tema, pero sí que se incluyen algunas recomendaciones basadas en los temas que se tratan en la versión ampliada del libro, disponible online.

En primer lugar, cuando se crea un diseño VDI de cientos de usuarios, es esencial separar la infraestructura de administración de virtualización de la carga de trabajo de VDI. Esto significa que todos los servidores de administración, intermediarios de VDI, servidores de archivos, servidores de administración de aplicaciones y cualesquiera funciones que no sean escritorios virtuales deberían ejecutarse en un clúster diferente. Que el clúster de gestión sea uno solo dedicado al diseño de EUC dependerá de lo grande que sea el entorno. Si el diseño es más pequeño, se pueden ejecutar máquinas virtuales de administración en un clúster de virtualización de servidor existente.

Es posible escalar estos clústeres de escritorios virtuales para alcanzar un tamaño de entre 16 y 32 hosts. Este rango permite que se cree un grupo de recursos mayor para que las VM los usen, y también empuja a la mayoría de los clientes a adoptar un clúster que sea más grande que sus tamaños típicos. Las actualizaciones recientes del hipervisor permiten clústeres de hasta 64 hosts, pero a muchos arquitectos y clientes les llevará tiempo sentirse cómodos con ese tamaño. Si el entorno es lo suficientemente grande como para que los recuentos de host excedan estos rangos, sería necesario más de un clúster VDI.

Otra razón por la que se diseñaría para clústeres de virtualización múltiple, además del tamaño del entorno, serían las diferentes cargas de trabajo que existen en los clústeres de VDI. Si hay una cantidad significativa de escritorios virtuales de 1 vCPU y 2 vCPU, se debe diseñar un clúster separado para cada uno. El Gráfico 7 ilustra un enfoque de diseño multi-clúster. Esto permite administrar la ratio de CPU de manera diferente en cada clúster, lo que proporciona un diseño más fácil de administrar. Si se tuvieran que combinar las diferentes configuraciones de CPU, sería necesario calcular una nueva ratio combinada, y esto conllevaría cierta confusión añadida.

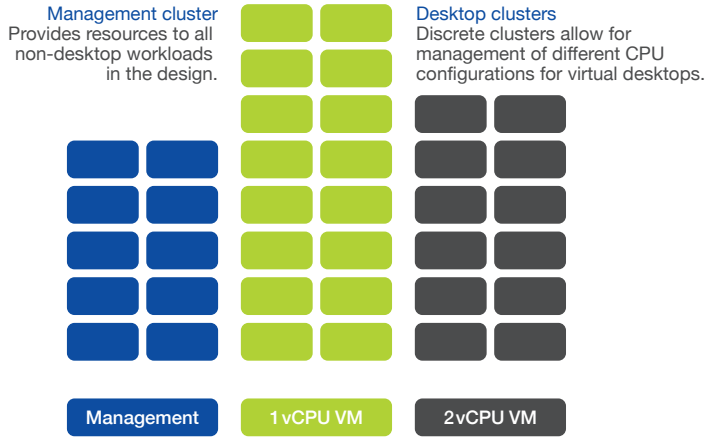

Gráfico 7
Gestión y clústeres de escritorio

Si ha llegado hasta aquí, esperamos que al menos le intriguen las posibilidades de virtualización de escritorios y aplicaciones. Si usted y su organización están listos para aprender cómo implementar con éxito la virtualización de escritorio y aplicaciones, desde Nutanix nos encantaría ayudarles. La infraestructura invisible de Nutanix puede simplificar enormemente el viaje a través de su galardonada arquitectura a escala web: la mejor plataforma VDI.

ES HORA DE EMPEZAR

Probablemente no le sorprenda que Nutanix haya dedicado mucho tiempo a pensar cómo descubrir las mejores formas de garantizar con éxito las implementaciones de virtualización de aplicaciones y escritorios.

Comprender su entorno actual:

El proceso comienza con una comprensión completa de su entorno actual de usuario final, lo que incluye:

- Métricas del usuario final: recopilar perfiles del usuario final y factores relacionados, como las aplicaciones utilizadas, los dispositivos de acceso del usuario final, la ubicación y la conectividad.

- Servicios de red y métricas específicas de infraestructura: recopilar información adecuada sobre diferentes servicios de usuario final, como servicios de archivos, autenticación y control de acceso, y firewalls / equilibrio de carga. También recopile métricas de rendimiento, latencia, productividad, etc.

- Asignarlo todo a los propietarios: la responsabilidad es un factor clave para el éxito.

CÁLCULO DEL NUEVO ENTORNO:

Con toda la información anterior, puede calcular con precisión su nuevo entorno. Nutanix Sizer simplifica esta tarea, pero debe tener en cuenta los aspectos siguientes:

- Siempre tenga en cuenta la alta disponibilidad para servidores y escritorios clave.

- Una infraestructura adicional y / o clústeres adicionales pueden ser necesarios según estas consideraciones:
  - Empresariales: acuerdos a nivel de servicio, licencias, seguridad, presupuesto, políticas.
  - Planificación de la transición: siga las mejores prácticas y directrices de Nutanix y del sector para las migraciones P2V y tenga cuidado con la creación de la imagen maestra que se utilizará para crear otros escritorios. Si está migrando una implementación existente, le recomendamos utilizar un partner Nutanix o herramientas nativas siempre que sea posible.

Naturalmente, Nutanix Global Services puede ayudarlo con cualquiera de estos pasos para encaminarlo hacia un mayor éxito de infraestructura. A través de nuestra organización de Servicios Globales, Nutanix ofrece la única solución del sector para eliminar el riesgo de un cálculo incorrecto de la infraestructura para proyectos de virtualización de escritorios.

Con el programa VDI Assurance, Nutanix garantiza que sus escritorios virtuales siempre obtengan los recursos de procesamiento (CPU virtual y memoria) y de almacenamiento (rendimiento y capacidad) necesarios para cumplir con las expectativas de VDI del usuario final. Solo tiene que determinar el tipo y la cantidad de usuarios de VDI en su entorno y transferir el riesgo de dimensionar los requisitos de infraestructura a Nutanix con VDI Assurance.

Si quiere saber más sobre infraestructura invisible para aplicaciones empresariales, póngase en contacto con nosotros en info@nutanix.com, síganos en Twitter y envíenos un DM a @nutanix, o envíenos una solicitud en www.nutanix.com/demo para configurar su propia sesión informativa personalizada y una demostración para comprobar cómo las soluciones validadas y certificadas de Nutanix pueden ayudar a su organización a aprovechar al máximo sus aplicaciones empresariales.

Manténgase en contacto con los expertos y clientes de Nutanix en la comunidad en línea de Nutanix Next (next.nutanix.com).

Nutanix ofrece una infraestructura invisible para el procesamiento empresarial de próxima generación, llevando la TI al siguiente nivel y centrándose en las aplicaciones y servicios que impulsan sus negocios. La Xtreme Computing Platform, basada en software, de la compañía unifica de manera nativa el procesamiento, la virtualización y el almacenamiento en una solución única que simplifica el centro de datos. Con Nutanix, los clientes se benefician de un rendimiento predecible, escalabilidad lineal y un consumo de infraestructura similar al de la nube. Obtenga más información en www.nutanix.com o síganos en Twitter: @nutanix.

NUTANIX™

T.855.NUTANIX (855.688.2649)
info@nutanix.com | www.nutanix.com | @nutanix