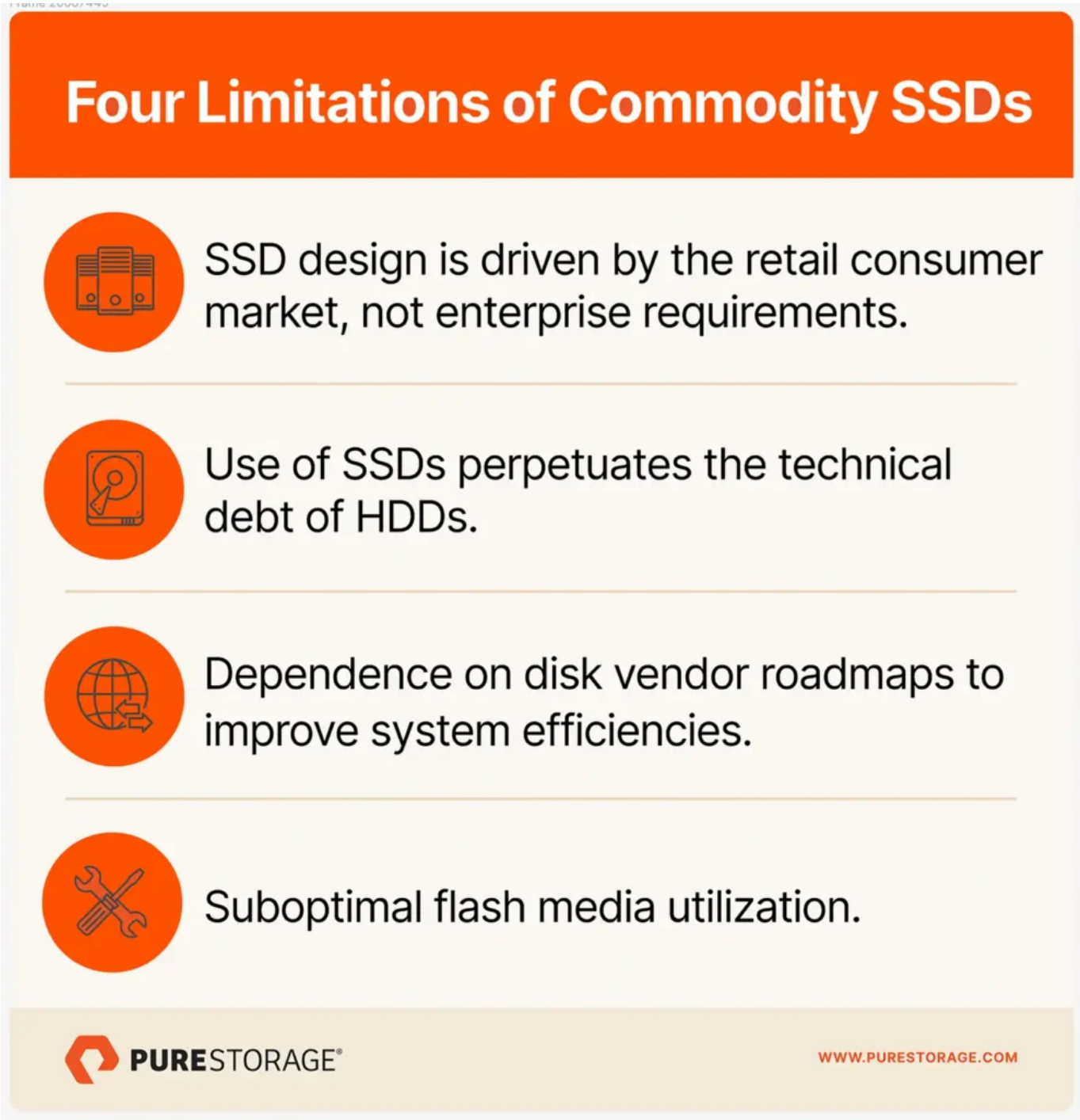
La trampa SSD: Cómo la dependencia del almacenamiento en las SSD puede afectarlo

A medida que los medios flash NAND de celda de nivel cuádruple (QLC) continúan expandiendo su prevalencia en los sistemas de almacenamiento, vemos una mayor rentabilidad de los SSD y, con eso, una caída en la implementación de los HDD tradicionales. Todas las mismas ventajas que impulsaron el desplazamiento histórico de los HDD para cargas de trabajo con mayor rendimiento ahora se aplican a todas las cargas de trabajo, impulsadas por los costos decrecientes de $/GB de los medios de QLC. Al mismo tiempo, la oportunidad de AI ha traído urgencia a las empresas que desean activar y monetizar sus datos de una manera operativamente eficiente.

La canción de cisne de los HDD ha comenzado, pero ¿es realista en este momento programar la estela?

En esta entrega de dos partes de nuestra serie de blogs “Más allá de lo esperado”, analizaremos por qué el uso de SSD estándar (COTS) de productos básicos, propuesto por algunas nuevas empresas de almacenamiento definido por software (SDS) basadas en NVMe, es una alternativa subóptima a las matrices basadas íntegramente en tecnología flash (AFA) que utilizan dispositivos de almacenamiento flash diseñados específicamente para reemplazar los HDD. En la Parte 2, veremos una comparación más directa entre los SSD COTS y los módulos Pure Storage® DirectFlash®.



Las inquietudes de sustentabilidad acelerarán la desaparición de los HDD

En comparación con los sistemas de almacenamiento en HDD heredados, el almacenamiento basado en flash ofrece beneficios abrumadores para todas las cargas de trabajo: rendimiento superior, mayor confiabilidad, mayor densidad de almacenamiento, menor consumo de energía y menor sobrecarga operativa. De 2012 a 2019, las AFA han aumentado en popularidad y ahora impulsan aproximadamente el 80 % o más de todos los envíos de almacenamiento para entornos de aplicaciones de alto rendimiento.

Básicamente, los HDD se dejaron en el polvo magnético.

¿Por qué? A medida que más empresas priorizan la sustentabilidad como un criterio clave para las nuevas compras de AFA, las métricas como la eficiencia energética (TB/vatio) y la densidad de almacenamiento (TB/U) se vuelven fundamentales para evaluar el costo de los nuevos sistemas. La mayor densidad del SSD (en relación con los HDD) y el rendimiento significativamente mayor son factores clave que contribuyen a una mejor eficiencia y, en última instancia, a un costo más bajo, ya que se necesitan muchos menos dispositivos de medios (así como una infraestructura de soporte mucho menor para aquellos dispositivos como controladores, gabinetes, ventiladores, fuentes de alimentación, cables, conmutadores, etc.) para construir un sistema que cumpla con los requisitos de rendimiento y capacidad determinados.

En Pure//Accelerate® 2023, ponemos en juego que 2028 sería el último año en que se venderían nuevos sistemas de almacenamiento construidos en torno a los HDD para uso empresarial. Lo que se perdió en medio del hulabaloo generado por esa declaración fue esto: La predicción agresiva de Pure Storage se basó en empresas que usaban *nuestros* dispositivos flash, no SSD estándar (COTS). Somos alcistas en flash, pero incluso no creemos que los SSD COTS vayan a eliminar los HDD esta década.

Hay algo muy diferente e innovador sobre cómo implementamos flash en nuestros sistemas de almacenamiento empresarial en Pure Storage. Sabemos esto porque los módulos DirectFlash (DFM) de Pure Storage han demostrado ser dos a cinco veces mejores en cada métrica significativa que los basados en SSD COTS. Además, son 10 veces mejores que los discos duros en cada métrica, excepto uno: el costo de adquisición sin procesar.

Limitaciones de los proveedores que aprovechan las SSD COTS

Cuatro cosas funcionan en comparación con los proveedores que usan SSD COTS, incluso cuando están basadas en NVMe. Echemos un vistazo a cada uno de ellos:

1. El diseño de SSD está impulsado por el volumen de los mercados minoristas de consumo en lugar de los requisitos empresariales.
2. El uso de SSD COTS perpetúa la deuda técnica de los HDD.
3. Crea dependencia de las hojas de ruta de los proveedores de discos para mejorar la eficiencia del sistema.
4. Ocasiona una utilización subóptima de medios flash.

El diseño de SSD está impulsado por el volumen de los mercados minoristas de consumo en lugar de los requisitos empresariales

En el mundo de los medios de almacenamiento, la rentabilidad de la fabricación se correlaciona directamente con los volúmenes de producción. Las SSD de consumo (PC y móviles), no empresariales, conforman el mercado de volumen. De hecho, las SSD empresariales representan aproximadamente solo el 15 % del mercado general de SSD. Siguiendo el volumen, las innovaciones clave para la tecnología COTS SSD están impulsadas por el mercado de consumo, que valora el bajo costo y las capacidades más bajas, no los requisitos empresariales. Por el contrario, las empresas valoran el rendimiento, la resistencia, la confiabilidad y la integridad de los datos, todas características que requieren una ingeniería profunda más allá de las SSD principales del consumidor. Esta dicotomía continúa invadiendo el ciclo de innovación de los SSD COTS y su aplicación en los mercados empresariales.

El uso de SSD COTS perpetúa la deuda técnica de los HDD

Los SSD se diseñaron para que pudieran insertarse fácilmente en sistemas de almacenamiento que habían sido diseñados para HDD. Específicamente, las SSD COTS están diseñadas en gran medida para adaptarse a gabinetes de unidades de factor de forma pequeño (SFF) de 2,5”. La mayoría de las AFA empresariales se desarrollaron específicamente para usar estas SSD COTS. Las AFA empresariales que utilizan dispositivos que tienen que adaptarse a huellas de SFF de 2,5” están limitadas por su capacidad para aumentar la densidad, ofrecer vatios/TB más bajos y usar medios flash de manera óptima.

¿Cuáles son esas limitaciones? Las SSD COTS incluyen un controlador interno, DRAM para almacenamiento en caché para ayudar a mejorar el rendimiento y los medios flash en sí. Arquitectónicamente, una SSD COTS necesita 1GB de DRAM por cada 1TB de capacidad flash, principalmente para impulsar la capa de traducción flash (FTL). Crear una SSD de mayor capacidad significa que debe encontrar el espacio dentro del dispositivo para el controlador, la DRAM y los medios flash, así como enfriar el dispositivo. Se está volviendo cada vez más difícil poder hacer eso y mantenerse dentro de los límites del empaque de 2,5” mientras se impulsa el aumento de la densidad.

Primero, veamos el controlador y los requisitos de firmware resultantes. Las SSD en sí tienen controladores integrados que manejan el direccionamiento lógico de

bloques, la administración del espacio libre y las tareas de mantenimiento en segundo plano, así como el FTL. También incluyen controladores flash NAND integrados que manejan el acceso directo a la memoria a la NAND. Todos estos controladores tienen firmware, y cada sistema de almacenamiento empresarial puede tener fácilmente cientos de SSD, por lo que hay mucho firmware. El firmware es una de las dos causas principales de falla en las AFA (las fallas de dispositivos son las más comunes). Este laberinto de complejidad de firmware plantea riesgos de confiabilidad.

Si alguna vez se preguntó por qué la mayoría de los proveedores de sistemas no recomiendan actualizar el firmware en línea a nivel de SSD, se debe al riesgo asociado con eso dada la complejidad del firmware. El resultado neto de esto es que es raro que una empresa incluso intente una actualización del firmware del disco durante la vida útil de un sistema, lo que relega eficazmente el sistema a usar firmware del disco más antiguo y menos eficiente que se envió inicialmente con el sistema. Sería bueno poder actualizar el firmware sin interrupciones durante la vida útil de un sistema para mejorar el rendimiento, la utilización de la capacidad, la resistencia, la eficiencia energética y otras métricas, pero en los sistemas que usan SSD COTS, eso rara vez sucede a menos que se vea forzado por algún tipo de problema de integridad de datos.

A continuación, veamos DRAM. La DRAM es mucho más costosa que los medios flash NAND, consume mucha potencia en relación con los medios flash en sí y ocupa espacio en el dispositivo que no contribuye directamente a su capacidad utilizable. La cantidad de DRAM necesaria en SSD de mayor capacidad es un factor limitante cuando necesita encajar en un factor de forma de 2,5”. Y en sistemas con cientos de SSD, eso es mucha DRAM, todo lo cual consume mucha más energía por GB que los medios flash NAND en sí. La DRAM también está sujeta a fallas que afectan la confiabilidad del dispositivo. Las SSD más grandes que requieren más DRAM aumentan las preocupaciones de confiabilidad de la DRAM.

Si bien surgió un nuevo enfoque de empaque de disco denominado factor de forma estándar empresarial y de centro de datos (EDSFF) que permite una capacidad de almacenamiento ligeramente más densa, sigue estando limitado por los mismos factores molestos de “empaque HDD” y actualmente se enfrenta a superar los 30,72 TB de tamaño mientras mantiene la confiabilidad.

Pero, ¿no se desarrolló NVMe específicamente para dispositivos flash? Sí, pero se diseñó principalmente para acelerar la latencia y aumentar el ancho de banda en los sistemas flash. En esas métricas, claramente supera a SAS, pero tiene un impacto mucho menor en la resistencia, confiabilidad, utilización de la capacidad o

eficiencia energética cuando se usa con SSD COTS que siguen siendo administradas por controladores individuales que trabajan de forma aislada.

Dependencia de los mapas de ruta de los proveedores de discos para mejorar las eficiencias

Una arquitectura de dispositivo de almacenamiento basada en las SSD COTS significa que el proveedor del sistema de almacenamiento depende de las hojas de ruta de SSD de los proveedores de discos para mejorar la eficiencia energética y la densidad de almacenamiento. Al principio, esto puede sonar bien: los proveedores de medios pueden enfocarse en hacer las “mejores” SSD y los proveedores de dispositivos de almacenamiento que usan las SSD COTS en sus sistemas pueden enfocarse en hacer los mejores sistemas, ¿verdad? Echemos un vistazo más de cerca.

En la actualidad, los SSD de 15,36 TB están disponibles en volumen, los SSD de 30,72 TB están disponibles (pero aún no en volumen, por lo que los precios siguen siendo relativamente altos) y los SSD de 61,44 TB están en la lista de precios de al menos un proveedor de discos (aunque la disponibilidad parece limitada y los precios son altos). A nivel del mercado, surgen preguntas sobre si los SSD de 30,72 TB lograrán la caída de precios asociada con los envíos de alto volumen. Y hay aún más preguntas sobre si los dispositivos de 61,44 TB alguna vez lograrán esa caída de precios. No hay duda de que los dispositivos flash de mayor capacidad marcan una gran diferencia en la eficiencia energética de una plataforma de almacenamiento, ya que reducen el consumo de energía y espacio en rack por terabyte, así como también abren la oportunidad de comprar menos dispositivos para alcanzar un objetivo de rendimiento y capacidad a nivel del sistema. En teoría, esto también mejora la confiabilidad, ya que un sistema con menos dispositivos también requiere una infraestructura de soporte mucho menor (controladores, gabinetes, ventiladores, fuentes de alimentación, cables, interruptores, etc.).

Sin embargo, crear una SSD de mayor capacidad no es solo una cuestión simple de instalar más medios flash en el dispositivo. Dependiendo del factor de forma, debe colocar el controlador, DRAM (1 GB por TB) y medios flash en el paquete. Suponiendo que un proveedor pueda lograrlo, existen preocupaciones empresariales válidas sobre SSD de más de 15,36 TB de tamaño, específicamente la utilización de la capacidad y los tiempos de reconstrucción del disco.

Independientemente de las afirmaciones de los proveedores de discos sobre la utilización de la capacidad a nivel del dispositivo, la mayoría de los proveedores de

sistemas de almacenamiento no recomiendan que llene las SSD en un sistema de almacenamiento más del 60 % al 70 % de su capacidad. (Sin embargo, es aún peor con los HDD, ya que los proveedores de sistemas sugieren no llenar esos dispositivos más del 50 % al 60 %). Esta incapacidad de utilizar completamente toda la capacidad de esos SSD significa que tiene que comprar más para cumplir con los objetivos de rendimiento y capacidad.

Para aumentar la utilización de la capacidad de los medios flash en cada dispositivo, los proveedores pueden agregar una gran capa de caché. Esta caché actúa como un búfer de escritura de mayor rendimiento y su gran tamaño le permite realizar una coalescencia de escritura para aumentar la resistencia y la utilización de la capacidad del flash de respaldo mejor que los sistemas convencionales. Pero este enfoque tiene sus riesgos. La caché generalmente consiste en un almacenamiento de rendimiento extremadamente alto (es decir, flash Optane o SLC NAND) que es significativamente más costoso sobre una base de $/GB y consume mucha más potencia que el flash QLC NAND. El rendimiento de escritura de esa caché debe ser lo suficientemente alto para evitar que el sistema golpee un acantilado de escritura, ya que simultáneamente ingiere nuevas escrituras, presta servicios a solicitudes de metadatos, maneja cualquier tráfico de lectura y administra cómo se eliminan los datos a medios flash de menor costo. ¿Qué es exactamente ese medio de alto rendimiento, cuánto cuesta y dónde se experimenta el acantilado de escritura con diferentes cargas de trabajo?

La proporción de lecturas a escrituras que los medios de alto rendimiento pueden ofrecer es un factor de éxito crítico para manejar con éxito las cargas de trabajo de alto rendimiento, y la proporción de caché a los medios flash variará con la intensidad de escritura de la carga de trabajo. Hay impactos en el rendimiento por no obtener la relación correcta y los impactos en los costos del gran tamaño de caché.

Ahora veamos el problema del tiempo de reconstrucción. La mayoría de los proveedores de sistemas de almacenamiento implementan alguna forma de protección de datos en línea en el disco, como RAID o codificación de borrado (EC) para proteger los datos de fallas individuales de SSD. A raíz de las fallas, las empresas están preocupadas por una segunda falla en el disco que provocaría la falta de disponibilidad o pérdida de datos antes de que se complete la reconstrucción del primer disco defectuoso y, por esa razón, están muy preocupadas por los tiempos de reconstrucción del disco. Cuando una SSD falla y debe reemplazarse, todos los datos de ese dispositivo deben reconstruirse. Las SSD COTS son mucho más rápidas que las HDD, por lo que las empresas se sintieron más cómodas con la implementación de SSD más grandes. Los

proveedores de sistemas de almacenamiento suelen citar un tiempo de reconstrucción de 8 a 12 horas para un SSD de 15,36 TB, mientras que la estimación sería de 25 a 30 horas para un HDD de ese tamaño. Pero esto supone que no sucede nada más en el sistema. En la práctica real, la tasa de reconstrucción para un HDD utilizando un enfoque EC de 20 bandas que continúa prestando servicio a I/O normal es de aproximadamente 1TB por día, lo que significa que la reconstrucción de un HDD de 24TB podría tomar más de tres semanas. Las SSD pueden reconstruirse mucho más rápido, pero para dispositivos de mayor capacidad, seguimos hablando de muchos días.

Debido a las preocupaciones sobre los tiempos de reconstrucción, las empresas tienden a querer implementar dispositivos de capacidades más pequeñas en sistemas que tienen alguna sensibilidad de disponibilidad de datos. Esto ha afectado en particular la capacidad de los proveedores de discos para vender dispositivos de 30,72 TB. De hecho, he visto que los clientes implementan dispositivos de 30,72 TB en los sistemas y luego deciden pasar a dispositivos de 15,72 TB o 7,68 TB porque no pudieron utilizar suficiente capacidad mientras cumplían con sus requisitos de rendimiento para hacerlos económicamente atractivos.1 También he visto a los proveedores ofertar SSD de 30,72 TB para ganar un acuerdo sobre el costo, pero luego mover al cliente a SSD de menor capacidad en la instalación por motivos similares.

Por lo tanto, con las SSD COTS, un tamaño de dispositivo más pequeño elimina las preocupaciones sobre los tiempos de reconstrucción, pero conduce a un sistema más costoso y potencialmente menos confiable. Los tamaños de dispositivos más grandes conducen a un sistema con mayor eficiencia energética y espacial, pero plantean inquietudes sobre la facilidad de uso. De hecho, por razones de rendimiento a escala y utilización de capacidad, es posible que los proveedores de discos ni siquiera estén dispuestos a producir SSD de capacidad extremadamente grande (más de 75 TB) que se necesitarán para que flash reemplace los HDD en las cargas de trabajo empresariales. Esto se debe a que incluso si los proveedores pueden fabricar dispositivos de mayor capacidad, aún hay una cuestión de si las empresas realmente los usarán en los sistemas de producción. Y si los clientes no los compran con mucha frecuencia, los proveedores no pueden obtener la fabricación de alto volumen necesaria para bajar los precios. Eso hace que los clientes empresariales sean aún menos propensos a comprarlos.

Eso lleva a lo que nuestros hermanos en el estanque podrían referirse como “batear un wicket pegajoso” para los proveedores de discos y las empresas, dependiendo de los SSD COTS.

Utilización subóptima de medios flash

¿Qué significa diseñar un sistema para usar medios flash de manera óptima? Significa dispensar equipaje de HDD, optimizar el diseño del sistema en torno a un protocolo que se creó específicamente para flash (NVMe) y mapear y administrar todos los medios flash en un sistema, tanto de forma global como directa.

Las SSD empresariales emplean un pequeño controlador en cada disco que administra los medios solo en ese dispositivo, coordinando la I/O lectura y escritura con la administración del espacio libre y lidiando con los errores I/O. Básicamente, esta fue la forma en que se construyeron los HDD y los SSD COTS se construyeron de la misma manera. Al tomar decisiones sobre la utilización de medios, el controlador de disco no tiene visibilidad de lo que sucede en el sistema. Al optimizar la administración de medios dentro de un solo disco, termina administrándola de manera subóptima desde el punto de vista de los sistemas. ¿El resultado? Una mayor amplificación de la escritura, un problema debido a las preocupaciones de resistencia a flash y una recolección de basura menos eficiente que tiene un impacto en la consistencia del rendimiento, particularmente a medida que se llenan los SSD. Para intentar compensar esto, flash está "sobreaprovisionado" (generalmente entre un 15 % y un 20 %) en el SSD COTS para ayudar a mejorar el rendimiento y aumentar la resistencia. Por supuesto, este sobreaprovisionamiento aumenta el costo y elimina la "capacidad utilizable" que una unidad puede proporcionar realmente (ya que la capacidad sobreaprovisionada ocupa espacio, pero solo es visible para el controlador de ese disco, no para el sistema).

Todavía hay sistemas de almacenamiento en el mercado que básicamente son versiones actualizadas de sistemas diseñados en las décadas de 1990 y 2000 para HDD que ahora ejecutan SSD. También hay sistemas más nuevos, algunos de los cuales fueron diseñados en torno a NVMe, que tienen menos equipaje de HDD y son más eficientes. Sin embargo, el "equipaje de HDD" que aún está allí con estos sistemas NVMe es que siguen usando dispositivos construidos alrededor del diseño original de HDD con un controlador interno, DRAM y medios que deben encajar dentro de un factor de forma de HDD. Y acceder a los medios de disco a través de controladores de disco internos que no tienen una visión global de lo que está haciendo el sistema presenta desafíos al intentar aprovechar al máximo los medios flash en muchas métricas: consistencia del rendimiento, resistencia, confiabilidad, consumo de energía, densidad y utilización de la capacidad.

En las SSD, el controlador interno asigna y presenta los medios al FTL, que luego los presenta a un controlador de almacenamiento que los presenta a servidores

que ejecutan varias aplicaciones. Un SSD de 15TB que funciona con una tasa de utilización de capacidad del 70 % presenta un poco menos de 11TB de capacidad utilizable. Eso es mejor que un HDD con una tasa de utilización de capacidad del 60 % que solo presentaría 9,2 TB de capacidad utilizable desde un dispositivo de 15TB. Pero, ¿cómo se compara esto con un sistema que usa dispositivos de almacenamiento flash (no SSD COTS) que no tienen herencia de consumidores y se diseñaron específicamente para impulsar el rendimiento y la eficiencia en entornos empresariales? Eso es exactamente lo que hemos hecho en Pure Storage con nuestros módulos DirectFlash (DFM), y responderemos esta pregunta en la Parte 2 de este blog.

En resumen, invertir en un sistema de almacenamiento empresarial construido en torno a las SSD COTS obliga a los clientes a aceptar una infraestructura ineficiente que los afecte negativamente en términos de consistencia del rendimiento, resistencia de los medios, confiabilidad del dispositivo, consumo de energía, densidad de almacenamiento, utilización de la capacidad y, en última instancia, costo del sistema. Pero como los puntos de comparación para muchos de ellos son sistemas basados íntegramente en HDD, los sistemas basados en SSD COTS parecen atractivos.

Preparación para la Parte 2

Hemos visto las consecuencias del uso de los SSD COTS. Para tantas cargas de trabajo empresariales, son mucho mejores que los HDD. Pero, ¿qué pasaría si un proveedor de almacenamiento pudiera construir dispositivos de almacenamiento flash que fueran de dos a cinco veces mejores que los SSD COTS en cuanto a consistencia de rendimiento, resistencia, confiabilidad, consumo de energía, densidad y utilización de capacidad, y un costo menor en $/GB? Analizaremos eso en la Parte 2.

1Pagar el 30 % de un SSD de 30TB que no puede usar es más costoso que pagar el 30 % de un SSD de 7,68 TB que no puede usar, un factor que desalienta la compra de SSD COTS de mayor capacidad.

Escrito por: