DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI

ISSN: 2007-6363

Publicación Semestral Pädi Vol.11 No. Especial 2 (2023) 55-60

Diseño e implementación de una red de almacenamiento distribuido basada en Ceph
Design and implementation of a Ceph based distributed storage network

D. Higuera-Balderrama[a,*], D. Morejón-López[b], R. Canosa-Reyes[b], R. Alonso-Labrada[c], F. Marín-Hernández[c]

[a] *Tecnológico Nacional de México, Instituto Tecnológico de La Paz, La Paz, Baja California Sur, México.*
[b] *ITX Consulting & Services, Ensenada, Baja California, México.*
[c] *Empresa de Telecomunicaciones de Cuba S.A, Cienfuegos, Cuba.*

Resumen

Los sistemas de almacenamiento de datos constituyen un asunto al que se le debe prestar especial importancia en los locales de servidores conocidos como "Sites" de las organizaciones. Esto se debe a que la pérdida de información por rotura de dispositivos de almacenamiento aislados como discos, arreglo de discos, servidores, etc., constituye el problema más sensible hoy en día en las organizaciones. Por otra parte, las soluciones tecnológicas de redes de área de almacenamiento (SAN) existentes son por lo general privativas, de elevado precio en el mercado, y carentes de mecanismos de integración con otras soluciones SAN de distintos fabricantes. Es por ello que en este trabajo se estudiaron dos soluciones de almacenamiento distribuido de código abierto para implementar en un futuro próximo en el Instituto Tecnológico de La Paz (ITLP), se seleccionó la solución nombrada Ceph, se diseñó una red de almacenamiento utilizando servidores con prestaciones de hardware heterogéneas, y fue implementada en fase de pruebas midiendo el impacto de su funcionamiento.

Palabras Clave: SAN, NAS, Ceph, Código abierto.

Abstract

Data storage systems are a special topic in server sites of organizations. This is because the lost of information due to isolated storage devices failure such as disks, disk arrays, servers, etc., is the most sensitive issue in organizations. On the other hand, Storage Area Network (SAN) technologies are private solutions, with a high market price, and without integration mechanisms with other SAN solutions. That's why in this paper we studied two open sourced and distributed storage systems to be applied in the the future in Instituto Tecnológico de La Paz (ITLP), Ceph system was selected, it was designed a storage network using servers with heterogeneous hardware features, and it was implemented in testing mode, measuring the impact of its performance.

Keywords: SAN, NAS, Ceph, Open source.

1. Introducción

Los sistemas de almacenamiento de datos constituyen un asunto al que se le debe prestar especial importancia en los locales de servidores de las organizaciones o centros de datos para nubes públicas y privadas.

En estos locales y en los centros de datos existen plataformas de virtualización que ejecutan máquinas virtuales (VM). Estas VM son las que ejecutan las aplicaciones. A su vez, las VM guardan su información persistente en discos duros virtuales (vDisks). Los vDisks son ficheros que se crean usualmente en dispositivos de almacenamiento remoto, o sea, fuera del servidor donde se ejecuta la VM. De esta manera el vDisk es un elemento que permanece fuera y debe ser almacenado en dispositivos de almacenamiento compartido que sean fiables, ya que si estos se averían se pueden perder los vDisks de varias VM. La figura 1 muestra un esquema de la topología descrita.

Hay que tener en cuenta que la topología mostrada en la figura 1 no incluye el sistema de salvas de máquinas virtuales que debe tener en adición toda organización. No obstante, la pérdida de un disco virtual por rotura de dispositivos de almacenamiento aislados como discos, arreglo de discos, servidores, etc., constituye un grave problema que pudiera compensarse con la restauración de la salva de la VM, pero perdiendo seguramente alguna información y sobre todo tiempo de disponibilidad del servicio que se brinda. Esto puede causar molestias a los usuarios y pérdidas económicas notables

*Autor para la correspondencia: david.hb@lapaz.tecnm.mx
Correo electrónico: david.hb@lapaz.tecnm.mx (David Higuera Balderrama), denis.morejon.lopez@gmail.com (Denis Morejón López), rewer@itxconsult.com (Rewer Miguel Canosa Reyes), redeis.alonso@nauta.cu (Redeis Alonso Labrada), marinescfg@gmail.com (Freddy Marín Hernández).

Historial del manuscrito: recibido el 30/03/2023, última versión-revisada recibida el 06/06/2023, aceptado el 13/06/2023 publicado el 11/09/2023. **DOI:** https://doi.org/10.29057/icbi.v11iEspecial2.10839

en muchos casos. Esta problemática es afrontada por todos los centros de datos en mayor o menor medida, dependiendo de los recursos monetarios y tecnológicos de los que dispongan y que empleen de manera acertada.

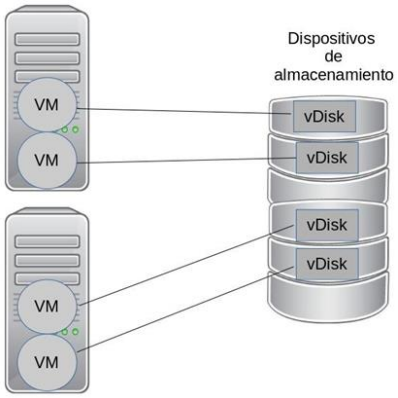

Figura 1: Esquema de uso típico de los dispositivos de almacenamiento en centros de datos. Se muestran las máquinas virtuales VM alojadas en los servidores, mientras que los discos virtuales (VDisks) están alojados en un dispositivo de almacenamiento el cuál comparten ambos servidores.

La figura 2 presenta un escenario donde los dispositivos de almacenamiento de los fabricantes B y D no tienen formas, por regla general, para llegar a un acuerdo y presentar un único almacenamiento a los servidores de los fabricantes A y C que ejecutan las VM, teniendo que presentarse el almacenamiento de forma independiente, sin réplicas de información entre ellos, etc. De esa manera no se puede lograr la abstracción del almacenamiento ante los servidores que lo utilizan, y disminuye la flexibilidad y seguridad en el uso de este valioso recurso. Supongamos que una organización proveedora de servicios de hosting, pudiera haber invertido en la implementación de sus servicios utilizando servidores y dispositivos de almacenamiento SAN de un fabricante determinado. Al cabo de un tiempo quiere expandirse en equipamiento, pero no puede hacerlo con el mismo proveedor por cualquier motivo. En ese caso la interoperabilidad del equipamiento actual con el de los futuros proveedores es crucial para llevar a término su objetivo a largo plazo, como se muestra en la figura 2.

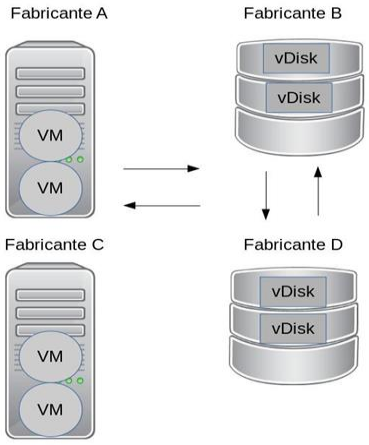

Figura 2: Necesidad de interoperabilidad en dispositivos de almacenamiento de distintos fabricantes.

Por otra parte, los dispositivos o sistemas de almacenamiento tradicionales precisan de discos duros del mismo tamaño para conformar arreglos que implementen la replicación de datos entre ellos. Además, una vez armados los arreglos es difícil expandirlos con discos nuevos sin tener que salvar la información, destruir el arreglo, y construir uno nuevo con el disco adicional.

Ese es el escenario que existe en varias organizaciones como el Instituto Tecnológico de La Paz (ITLP). Donde el centro de cómputo cuenta con un cluster de la plataforma de virtualización Proxmox VE (Ahmed, 2017) para ejecutar las VM, pero los vDisks de dichas máquinas se almacenaban en los discos duros locales en cada servidor nodo del cluster Proxmox VE. De esta forma si se averían los discos de un servidor se pierden los vDisks de las VM que ejecuta, y sólo se podrán restaurar utilizando el sistema de salvas, pero con pérdida parcial de la información procesada en el día y la demora en el proceso de restauración.

Por consiguiente, el problema de esta investigación se plantea como sigue.

1.1. Problemática

Los dispositivos de almacenamiento para la plataforma de virtualización Proxmox VE del Instituto Tecnológico de La Paz, no garantizan la interoperabilidad entre ellos, ni mecanismos de abstracción para presentar mayor capacidad de almacenamiento, ni la replicación de datos necesaria para mantener la información en estado de alta disponibilidad

1.2. Hipótesis

La búsqueda e implementación de un sistema de código abierto para almacenamiento distribuido, que permita aglutinar a varios servidores con discos duros de distintas capacidades, en un sólo elemento de almacenamiento virtual con alta disponibilidad, permitirá resolver las dificultades de flexibilidad y seguridad en el almacenamiento de discos virtuales que presenta el ITLP.

1.3. Objetivo General

Implementar un sistema de almacenamiento distribuido y de código abierto que agrupe servidores y dispositivos de almacenamiento heterogéneos en un sólo espacio de almacenamiento virtual y de alta disponibilidad.

2. Desarrollo

En el presente trabajo se estudiaron dos sistemas de almacenamiento distribuido de código abierto que eran integrables en la plataforma Proxmox VE (Ahmed, 2017) existente en el ITLP.

Es importante destacar que son dos aspectos diferentes el hecho de ser un almacenamiento compartido y otro el de ser un almacenamiento distribuido y se explicarán las diferencias.

Un almacenamiento compartido es aquel que radica fuera de la plataforma de virtualización como un ente independiente y capaz de ser accedido por todos los nodos que ejecutan las máquinas virtuales. Es accedido a través de un protocolo de acceso a datos remotos como pueden ser iSCSI o NFS (Network File System). La ventaja del almacenamiento compartido es que las VM que se ejecutan en los virtualizadores se pueden migrar de un servidor físico a otro sin que sea perceptible por los usuarios de las aplicaciones que

radican en dichas VM. A esto se le denomina migración en caliente o live migration (Ahmed, 2017) (en idioma inglés).

En la figura 1 se observa cómo el dispositivo de almacenamiento es externo y a la vez compartido por los servidores que ejecutan las VM. De manera que la migración de una VM hacia otro servidor ocurre solamente a nivel del contexto de la VM en memoria y CPU, porque el vDisk asociado a dicha VM permanece en el dispositivo de almacenamiento compartido. Es por eso que ocurre muy rápido. Esta condición es un aspecto clave y ya casi elemental en la implementación de entornos de virtualización.

Pero, ¿Qué ocurre si se avería el dispositivo de almacenamiento compartido? Ciertamente es un problema que afrontan muchas organizaciones, aun las que poseen dispositivos de fabricantes reconocidos a nivel mundial. El dispositivo de almacenamiento por lo regular permite conformar un arreglo de discos, como puede ser RAID5 o RAID6, etc. en el cual se garantiza la replicación de datos entre los discos de manera que puedan averiarse uno o dos discos al mismo tiempo (En dependencia del tipo de arreglo que se prepare) y aun así el arreglo permita disponer de la información asociada al disco averiado y permanece la capacidad de almacenamiento intacta. Pero pocos ofrecen alternativas cuando se avería el dispositivo en su totalidad, o son muy caras de implementar.

Aquí llega el concepto de almacenamiento distribuido, se trata de que el almacenamiento compartido no radique en un sólo equipo, sino que existan varios equipos en una red de almacenamiento, donde sean capaces de organizar y replicar datos entre ellos y sean presentados al cliente de este recurso (Los servidores de virtualización) como un elemento único donde almacenar los vDisks.

En el esquema de la figura 2 se muestra a la derecha dos dispositivos de almacenamiento de distintos fabricantes que necesitan ejercer esta réplica y coordinación de datos para que sean presentados a los servidores de la plataforma de virtualización que están a la izquierda.

2.1. Plataforma Proxmox VE existente

En el ITLP existe un cluster de servidores Proxmox VE (Ahmed, 2017) para ejecutar las VM dentro de los discos físicos de los propios servidores. En la figura 3 se exponen a modo de ejemplo sólo dos servidores Proxmox VE con sus respectivos discos locales utilizados para almacenar las VM.

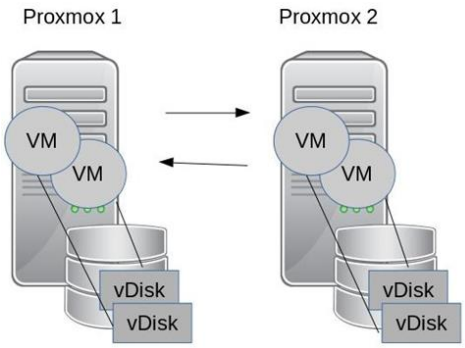

Figura 3: Cluster Proxmox con discos duros locales.

La desventaja fundamental de esta topología radicaba en que si fallan los discos de un Proxmox VE habrá que restaurar toda la VM desde el sistema de salvas que implementa Proxmox VE y no está reflejado en la figura porque está fuera del objeto de estudio. Eso implica pérdida de tiempo de disponibilidad y pérdida parcial de datos. Además, no se logran realizar migraciones de VM en caliente para efectuar mantenimientos de hardware a los servidores. Este mantenimiento hay que realizarlo en horarios nocturnos o de la madrugada en dependencia de la utilización de sus VM por parte de los clientes. Es necesario recurrir a un sistema de almacenamiento compartido, pero que a su vez tenga un carácter distribuido para conservar los datos almacenados en alta disponibilidad. Es así que se estudiaron y valoraron las características técnicas de los siguientes sistemas de almacenamiento distribuido:

- GlusterFS (Gluster File System)
- Ceph (Ceph storage system)

2.2. GlusterFS

El sistema GlusterFS (Gluster, 2012) es de código abierto. La plataforma Proxmox VE puede utilizarlo, y además presenta un almacenamiento compartido y a su vez distribuido. Puede instalarse sobre una distribución Linux como Debian, incorporando al sistema operativo un repositorio desde donde se instala el GlusterFS server.

El sistema presenta a los Proxmox VE un único almacenamiento, y se encarga de distribuir los ficheros almacenados entre los distintos servidores donde es instalado de forma homogénea, es decir, con GlusterFS (Gluster, 2012) la forma de replicación consiste en repartir los datos en N partes iguales.

Por ejemplo, si dispone de dos servidores con cuatro discos duros cada uno, se deben organizar los cuatro discos con un arreglo interno tipo RAID5, pero el volumen resultante debe ser del mismo tamaño en ambos servidores porque luego el sistema distribuye los datos que se escriben en un servidor hacia el otro para lograr la redundancia. El inconveniente que tiene esta manera simétrica de distribuir los datos es que siempre se requiere el doble de los recursos de almacenamiento, y esa forma de distribución no es eficiente. Además, se necesitan discos duros de igual tamaño para la distribución simétrica de los datos.

2.3. Sistema de almacenamiento Ceph

El sistema de almacenamiento Ceph (Singh, 2015), (Fisk, 2019) es de código abierto, integrable en Proxmox VE como almacenamiento compartido y además distribuido. Tiene al igual que GlusterFS varios puntos de acceso a la información, de manera que Proxmox VE pueda acceder a un nodo cuando otro falle.

En cambio, el mecanismo de replicación y construcción del almacenamiento virtual es más flexible ya que permite utilizar discos duros de distintos tamaños y en distintos servidores. En GlusterFS sólo se permiten un número par de servidores debido a la distribución simétrica de los datos, pero aquí en Ceph se pueden instalar cualquier número de servidores.

El sistema Ceph (Singh, 2015), (Fisk, 2019) permite configurar en él la distribución física de dónde se encuentran los servidores, permite insertar varios centros de datos, locales, bastidores, etc. e incluso el número de réplicas de los datos que se requieran, para así tenerlos en cuenta cuando se ejecute dicha replicación, y tratar de tener copias de los datos en servidores, bastidores, y locales distintos para conservar mejor

la información, Ceph cuenta con algoritmos para efectuar la réplica en las distintas localidades (Sage *et al.*, 2007).

Por otra parte, permite balancear la carga de escritura y lectura de información del cliente, en este caso Proxmox VE.

Por ejemplo, con Ceph se pueden tener 3 servidores con la distribución Linux Debian instalada. Luego se instala el sistema Ceph en los tres servidores. Pudiera existir un servidor A de propósito general con dos discos de 2TB, un servidor B con tres discos de 1TB, y un servidor C con un arreglo de discos tipo RAID5 de 4TB de capacidad al combinar las capacidades de 3 discos duros. Luego, al combinar todas las capacidades y teniendo en cuenta la redundancia deseada (Que es configurable dinámicamente) el sistema presenta a la red de Proxmox VE un almacenamiento único virtual que pudiera ser, para el ejemplo, de 9TB, como se muestra en la figura 4.

Este almacenamiento será ideal para los vDisks de las VM. Aunque también puede emplearse para el sistema de salvas de la plataforma Proxmox VE.

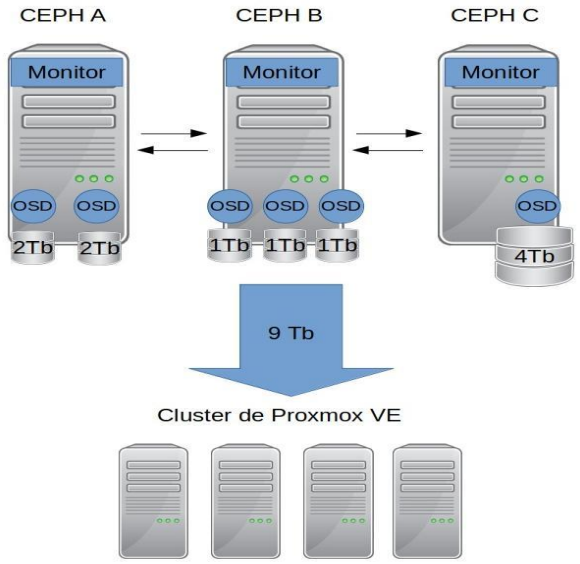

Figura 4: Sistema de almacenamiento Ceph con Cluster Proxmox VE.

El sistema de almacenamiento Ceph posee los siguientes componentes:

- Object Storage Device (OSD)
- Monitor

Los OSD (Object Storage Device) son los procesos que corren en un servidor y que se encargan de la lectura y escritura de fragmentos de información en discos. Por lo general se selecciona un OSD por disco. Los monitores son aquellos procesos encargados de chequear el estado de los OSD e interactuar con los clientes (En este caso los Proxmox VE). En este ejemplo, al cluster Proxmox VE se le agrega por la interfaz web el almacenamiento Ceph equivalente a 9TB aproximadamente, y las direcciones IP de los 3 monitores, de esta manera, si alguno falla se puede acceder a la información a través de otro monitor.

2.4. Implementación en fase de pruebas del sistema de almacenamiento Ceph

Una vez analizadas las características técnicas de los sistemas de almacenamiento GlusterFS y Ceph, como se muestra en la Tabla 1, se tomó la decisión de implementar el sistema de almacenamiento Ceph para realizar el experimento descrito en la siguiente sección. La selección de Ceph se basó principalmente en la infraestructura con la que se cuenta en el centro de cómputo del Instituto Tecnológico de La Paz, dicha infraestructura consta de 5 servidores y cada uno de ellos tienen diferentes capacidades de almacenamiento,

Tabla 1: Comparación entre el sistema de almacenamiento GlusterFS y el sistema de almacenamiento Ceph.

Criterio	GlusterFS	Ceph
Puntos de acceso al almacenamiento	El almacenamiento puede ser accedido a través de todas las direcciones IP que poseen los nodos involucrados.	El almacenamiento puede ser accedido a través de las direcciones IP de varios nodos definidos como monitores.
Réplica de datos	Los datos están replicados entre varios nodos.	Los datos están replicados entre varios nodos.
Protocolo de acceso soportado por Proxmox VE	El protocolo GlusterFS es soportado por Proxmox para asociar el almacenamiento a su cluster.	El protocolo Ceph es soportado por Proxmox para asociar el almacenamiento a su cluster.
Eficiencia en el almacenamiento	Distribuye los datos de forma simétrica. No permite almacenar datos en discos con diferentes tamaños, por lo que siempre se utilizan solo la mitad de las capacidades.	Distribuye los datos teniendo en cuenta las desigualdades en los tamaños de los discos, y el nivel de replicación deseado. Por esto aprovecha mejor las capacidades de almacenamiento disponibles.
Flexibilidad	La unidad de almacenamiento se centra en el nodo, por lo que hay que lidiar con otros mecanismos para organizar los discos internos.	La unidad de almacenamiento se centra en el disco (OSD). Por lo que se pueden organizar las réplicas entre los discos de todos los nodos y de diferentes tamaños.
Tiene en cuenta la ubicación física de los nodos	No	Si

Por lo tanto, la implementación de GlusterFS era inviable, ya que como se expuso en secciones anteriores se requiere de

un número par de servidores los cuales deben tener la misma capacidad de almacenamiento.

A continuación, se describen las pruebas que se realizaron al sistema de almacenamiento Ceph. En primer lugar, el objeto de estudio del presente trabajo es medir el rendimiento del sistema de almacenamiento seleccionado y para esto se realizaron diversas operaciones de escritura y se midió el ancho de banda utilizado para cada una de estas operaciones, el experimento se aborda con mayor detalle en la sección 2.6.

Otras pruebas que se realizaron fueron verificar el funcionamiento correcto de las migraciones en caliente y de la alta disponibilidad. Se comprobaron las migraciones de VM en caliente satisfactoriamente, estas pruebas de migración se llevaron a cabo manualmente dando la instrucción para que la VM migrará o se moviera hacia otro nodo del cluster de Proxmox VE. Luego se configuraron las VM en modo alta disponibilidad (HA). En este modo las VM configuradas son capaces de migrarse automáticamente hacia otro Proxmox VE cuando el servidor físico se apaga por algún motivo. Este mecanismo fue comprobado también satisfactoriamente retirando el cable de red de uno de los servidores Proxmox VE con VM en modo HA, y verificando cómo las VM se trasladaban a los 3 minutos hacia otro servidor Proxmox VE.

2.5. Experimento para comprobar el rendimiento del Ceph en las operaciones de entrada/salida.

Se realizó un experimento para comprobar el rendimiento de una máquina virtual cuando utiliza el sistema de almacenamiento Ceph con relación a cuando utiliza un almacenamiento local. Específicamente se midió el ancho de banda de las operaciones de entrada/salida en el disco virtual (vDisk).

Se creó una VM con sistema operativo GNU/Linux, distribución Ubuntu. Se programó un script en bash (Sage *et al.*, 2007) nombrado test_disk_io.sh.

```
#!/bin/bash
datafile="file.dat"
tmpfile="/tmp/tmpfile"
logfile="test_disk_io.log"
megas=500 #500Mb
    dd oflag=nocache bs=1M count=$megas if=/dev/zero
of=$datafile 2>$tmpfile
timestamp=`date +%F_%R`
disk_rate=`cat $tmpfile | awk '/copied/{print $10,$11}'`
echo $timestamp $disk_rate >> $logfile
rm $datafile
rm $tmpfile
```

Este script crea en el disco un fichero de 500Mb y luego lo borra. Después se colocó el script en una tarea programada (dentro de la VM) que se ejecutaba cada 1h en horario de la madrugada. Y se registraron los datos del ancho de banda de escritura en disco de las operaciones en cada hora en el fichero: test_disk_io.log. Los resultados de este experimento se analizan en la siguiente sección.

3. Resultados

A continuación, se analizarán los resultados del experimento descrito en la sección 2.6. La figura 5 presenta los resultados del experimento realizado, en dónde se compara el ancho de banda en MB/s (Eje Y) cuando la VM tiene el vDisk local (Curva superior), y cuando tiene el vDisk sobre el almacenamiento Ceph (Curva inferior).

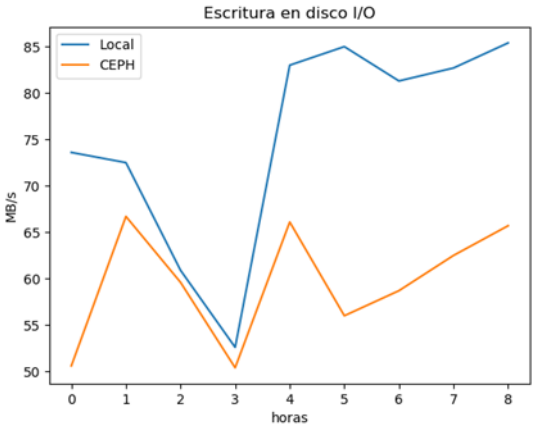

Figura 5: Resultados de escritura en disco.

Como se puede observar, existe una caída del rendimiento en escritura en disco expresada en MB/s. La diferencia promedio entre las dos operaciones de escritura fue de 15.63 MB/s. Esta caída en el rendimiento que se presenta con el sistema de almacenamiento Ceph, se debe principalmente a que las operaciones de escritura se realizan s través de la red y a su vez la información es replicada en distintos servidores para tener así alta disponibilidad en la información, a diferencia que cuando los discos virtuales vDisk se encuentran localmente en cada servidor las operaciones de escritura se hacen a través del bus de datos de cada servidor el cuál es de una velocidad mayor que la transferencia de datos por medio de la red a pesar que el tráfico en esta última sea mínimo.

A pesar del resultado, esa diferencia no afectó de manera significativa las operaciones de escritura en las 10 VM cuyos vDisks fueron movidos del almacenamiento local al almacenamiento Ceph, motivados por las múltiples ventajas de poseer un almacenamiento compartido y distribuido, que permite obtener alta disponibilidad tanto a nivel de VM como a nivel de almacenamiento.

4. Conclusiones

- Se estudiaron dos sistemas de almacenamiento distribuidos, de los cuales se descartó la implementación de GlusterFS debido a que los servidores de experimentación tenían todos diferentes capacidades de almacenamiento, por lo que, gracias a su flexibilidad, resultó ser Ceph el más conveniente para aplicar en la plataforma Proxmox VE del centro de cómputo del ITLP.

- El sistema Ceph, una vez instalado, permite que uno de sus servidores deje de funcionar y aun así las VM que se ejecutan en los Proxmox VE siguen realizando operaciones de lectura/escritura.

- Las VM escriben datos en el sistema Ceph con menor ancho de banda que como lo hacen cuando escriben en discos virtuales locales. Pero aun así es más conveniente utilizar Ceph porque permite migración en caliente de VM, y alta disponibilidad tanto en la plataforma de virtualización como en el almacenamiento.

- Es importante destacar que en el experimento se realizó la escritura de un archivo de 500 MB y se obtuvieron los resultados presentados, sin embargo, esta diferencia de menor ancho de banda al momento de realizar operaciones de escritura en el sistema Ceph sería prácticamente imperceptible para el usuario cuando se trabajara con archivos pequeños.

Referencias

Ahmed, W. (2017) Mastering Proxmox third Edition, Ed. Packt Publishing, Birmingham – MUMBAI.

Fisk, N. (2019) Mastering Ceph: Infrastructure storage solutions with the latest Ceph release, 2nd Edition, Ed. Packt Publishing.

Gluster (2012) Gluster File System Developers Gluster File System 3.3.0 Administration Guide. Using Gluster File System Edition 1.

Robbins A. (2021) Bash Pocket Reference: Help for Power Users and Sys Admins 2nd Edition, Ed. O'Reilly.

Sage A. Weil, Scott A. Brandt, Carlos Maltzahn (2007) CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data, IEEE Xplore.

Singh, K. (2015) Learning CEPH, Ed. Packt Publishing, Birmingham UK.