Análisis

Los modernos centros de datos están compuestos por decenas de miles de servidores y realizan el procesamiento para muchas aplicaciones de negocios en Internet. Los centros de datos están utilizando cada vez más la virtualización para simplificar la gestión y aprovechar mejor los recursos del servidor. Con la llegada de la computación en la nube a gran escala, los usuarios pueden obtener recursos informáticos bajo demanda con modelos de precios flexibles.

Antecedentes y motivación

Las aplicaciones de Internet y de negocios se están trasladando cada vez más a grandes centros de datos que albergan clústeres masivos de servidores y almacenamiento. Los centros de datos actuales pueden albergar decenas o cientos de miles de servidores, y ya se están implementando estrategias para centros de datos que utilizan más de un millón de servidores.

Memoria Trascendente

La memoria trascendente es un tipo de memoria que el kernel de Linux no puede enumerar, rastrear ni direccionar directamente, pero que ayuda a utilizar de manera más eficiente la memoria por parte de un solo kernel o a equilibrar la carga de memoria entre múltiples kernels en un entorno virtualizado. La implementación de tmem se divide en dos partes: la parte frontal (frontend) y la parte trasera (backend).

Existen dos frontends de tmem en el kernel de Linux que abarcan dos tipos principales de memoria del kernel: páginas anónimas y páginas respaldadas por archivos.

- Cleancache: Cleancache se utiliza para almacenar páginas respaldadas por archivos en un disco. El kernel puede optar por recuperar estas páginas en momentos de presión de memoria. Estas páginas se eliminan de la memoria.
- Frontswap: Frontswap se utiliza para almacenar páginas de intercambio (swap). El subsistema de intercambio de Linux almacena páginas anónimas en un dispositivo de intercambio cuando necesita eliminarlas.
- Auto-reducción de Frontswap
  Cuando el kernel intercambia una página, asume que la página irá al disco y puede permanecer allí durante mucho tiempo, incluso si no se vuelve a utilizar, ya que el kernel supone que el espacio en disco es menos costoso y abundante.

Las políticas de recuperación de memoria varían según el estado de memoria definido en el servidor ESX. Hay cuatro estados de memoria que determinan las técnicas de recuperación utilizadas:

Alto (High): En este estado, más del 6% de la memoria total del hipervisor está libre. Solo se utiliza el intercambio de páginas (page sharing) en este estado.

Blando (Soft): La memoria libre se encuentra entre el 6% y el 4%. El intercambio de páginas está activo y el controlador de balón (balloon driver) comienza a recuperar memoria inactiva.

Duro (Hard): La memoria libre se encuentra entre el 4% y el 2%. En este punto, el hipervisor comienza a recuperar memoria de forma agresiva a través del intercambio (swapping). También se activa la compresión de páginas. En la recuperación de páginas, si una página es comprimible o compartible, se comprime o comparte, de lo contrario, se intercambia.

Bajo (Low): Cuando la memoria libre está por debajo del 1%, se encuentra en este estado. Además de la recuperación de memoria, ESX bloquea cualquier asignación de memoria nueva por parte de cualquier invitado (guest).

Gestión de recursos en la nube

La gestión de recursos es una técnica esencial para utilizar eficientemente el hardware subyacente de la nube. El papel del administrador de recursos es gestionar la asignación de recursos físicos a las máquinas virtuales desplegadas en un clúster de nodos en la nube. Los diferentes sistemas de gestión de recursos pueden tener diferentes objetivos según las necesidades. En el caso de una nube privada, como en una institución educativa, el objetivo más común podría ser maximizar el rendimiento de las máquinas virtuales al tiempo que se minimizan los costos operativos de la infraestructura de la nube.

Conclusión

La gestión eficiente de recursos es crucial para optimizar el rendimiento y minimizar los costos operativos en entornos de computación en la nube. Los administradores de recursos desempeñan un papel fundamental al asignar los recursos físicos a las máquinas virtuales, asegurándose de utilizar de manera óptima el hardware subyacente. Los diferentes sistemas de gestión de recursos pueden tener objetivos específicos según las necesidades, ya sea maximizar el rendimiento, reducir los costos o cumplir con acuerdos de nivel de servicio. Independientemente del objetivo, la gestión eficiente de recursos es fundamental para garantizar un uso eficiente de los recursos y una operación rentable de la infraestructura en la nube.