De DMA a UltraEthernet

Eduardo Collado
edu@tecnocratica.net
https://www.tecnocratica.net

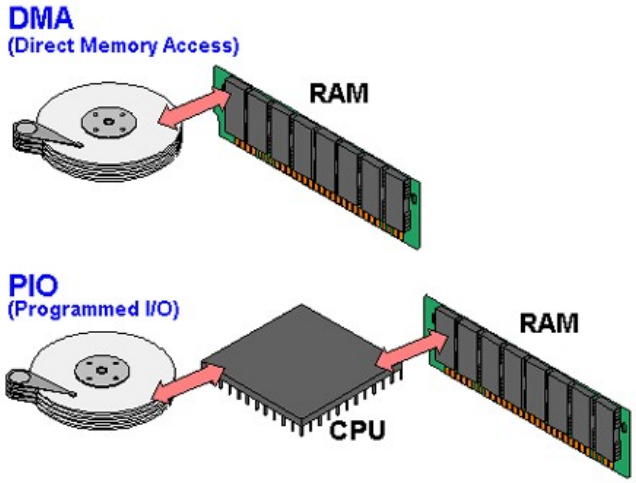
ES.NOG 31

TECNOCRÁTICA

DMA

DMA

- Se permite a ciertos componentes de hw acceso a memoria sin que la CPU tenga que intervenir.



- Funcionamiento:
  - El dispositivo que quiere acceder pide permiso al controlador DMA.
  - El controlador verifica que el canal esté libre y lo asigna al solicitante.
  - El controlador gestiona transferencia entre memoria y dispositivos enviando direcciones de memoria y controles de lectura/escritura al bus de memoria.
  - Para terminar el controlador DMA envía señal de interrupción a la CPU indicando la finalización.

DMA

- Ventajas:
  - Mayor eficiencia al liberar la CPU.
  - Menor latencia.

¿Y si la memoria no está en el mismo host?

⇓

RDMA

RDMA

Principio de Funcionamiento de RDMA

- Las NIC ordinarias requieren múltiples copias de datos, tanto enviados como recibidos, lo que consume significativamente los ciclos de cómputo de la CPU.
- RDMA (Acceso Directo a Memoria Remota) ofrece una solución para mejorar el rendimiento de la red y la eficiencia de la CPU mediante la eliminación de copias de datos innecesarias.

Proceso de Trabajo de una NIC Ordinaria

- El emisor de datos copia los datos del espacio de usuario al buffer del socket en el espacio del kernel y agrega encabezados de paquetes.
- La transmisión de datos implica múltiples copias entre buffers y finaliza con la recepción y procesamiento de datos en el sistema receptor.
- Este proceso implica un uso intensivo de la CPU y restricciones de velocidad de red debido a la gestión de datos entre el kernel y la memoria de la aplicación.

Desafíos de las NIC Ordinarias

- La intensa tarea de copia de datos ejerce presión sobre los ciclos de cómputo de la CPU, limitando el rendimiento general.
- Se requieren recursos adicionales de CPU y mejoras de software para manejar el aumento en la comunicación de la aplicación y la gestión de datos.

¿Qué es RDMA?

- RDMA permite la transferencia directa de datos entre memorias de sistemas remotos sin la intervención del sistema operativo, liberando así ciclos de CPU.
- Elimina las copias de datos innecesarias y reduce significativamente la latencia de los mensajes entre sistemas.

Proceso de trabajo de RDMA (I)

- Las aplicaciones pueden realizar solicitudes de lectura o escritura directamente desde el espacio del usuario a la NIC, sin replicación de datos.
- Los datos se transmiten directamente a la memoria del sistema remoto, utilizando **direcciones virtuales** y **claves de memoria** para la asignación de datos, lo que permite un procesamiento eficiente y rápido.

Proceso de trabajo de RDMA (II)

- Solicitud de RDMA desde la aplicación.
  - La aplicación realiza una solicitud de lectura o escritura de RDMA.
  - La solicitud se envía desde el espacio del usuario a la NIC local sin replicar datos.
  - La NIC lee el contenido del búfer y lo transmite a través de la red a la NIC remota.

Proceso de trabajo de RDMA (III)

- La NIC remota recibe y procesa la información RDMA.
  - La información RDMA transmitida a través de la red contiene la dirección virtual de destino, la clave de memoria y los datos en sí.
  - La NIC remota reconoce la clave de memoria y escribe los datos directamente en la memoria caché de la aplicación.
  - La dirección de memoria virtual remota utilizada para la operación está contenida en la información de RDMA.

Proceso de trabajo de RDMA (IV)

- Finalización de la solicitud en espacio de usuario o kernel.
  - La finalización de la solicitud puede manejarse completamente en el espacio del usuario mediante el sondeo de la alineación de finalización del nivel de usuario.
  - Si la aplicación duerme hasta que se completa la solicitud, la finalización puede gestionarse a través de la memoria del kernel.

Beneficios de Implementar RDMA

- **Menor latencia**: RDMA reduce la latencia de la transferencia de datos.
- **Menor uso de CPU**: Al eliminar las copias de datos, RDMA libera recursos de la CPU para otras tareas.
- **Mayor ancho de banda**: RDMA permite alcanzar anchos de banda de red más altos al eliminar las copias de datos innecesarias, al reducir la sobrecarga del kernel. Además de pueden utilizar mecanismos de aceleración como RoCE o Infiniband.

Dónde puedo encontrar RDMA

- **Infiniband**: Soporta RDMA de forma nativa. Solución de NVIDIA.
- **RoCE**: RDMA over Converged Ethernet. Disponible en repositorios de distribuciones GNU/Linux.
  - En Debian 12: rdma-core/stable
- **IWARP**: Internet Wide Area RDMA Protocol. Hace posible el uso de RDMA en redes WAN.
- **Slingshot**: Expone características optimizadas de RDMA y HPC al software utilizando Libfabric. Solución de HP-Cray.
- **Omni-Path**: Solución de Intel.

RoCE

Qué es RoCE

- Es una tecnología que permite la comunicación de RDMA sobre redes ethernet.
  - RoCE significa RDMA over Converged Ethernet.
  - Permite la transferencia directa de datos entre memorias de sistemas a través de Ethernet.
  - Evita la sobrecarga del CPU y reduce la latencia.

Funcionamiento de RoCE

- Utiliza el acceso directo a memoria remota (RDMA).
- Facilita transferencias de datos de alta velocidad sin cargar el procesador central.
- Mejora significativamente la eficiencia operativa.

Aplicaciones de RoCE

- Ideal para centros de datos, almacenamiento de alto rendimiento y computación en la nube.
- Beneficia entornos que requieren altas tasas de transferencia de datos y baja latencia.

Ventajas de RoCE

- Se beneficia de la infraestructura Ethernet existente.
- Fácil integración y adopción sin hardware especializado (se aprovecha la infraestructura de red existente.
- Baja latencia.
- Alta eficiencia.

Versiones de RoCE

- **RoCE v1**: Opera en redes sin pérdidas con Priority Flow Control (PFC).
  - Ethertype 0x8915.
- **RoCE v2**: Funciona sobre redes IP, utiliza UDP para encapsular paquetes RDMA.
  - UDP puerto 4791. IPv4 o IPv6.
  - El mecanismo de control de congestión con ECN. Los ACKs con CNP.

La Importancia de RoCE

- Mejora la eficiencia de la red y reduce la carga en procesadores.
- Disminuye la latencia de la red, mejorando el rendimiento de las aplicaciones.

Diseño de Redes para Soporte RoCE

- Selección de hardware adecuado.
- Configuración de red optimizada para baja latencia.
- Implementación de prácticas para coexistencia «armoniosa» con otras aplicaciones.

Selección de Hardware Compatible con RoCE

- Importancia de elegir adaptadores de red y conmutadores compatibles con RoCE.
  - Adaptadores: **Mellanox** (Series ConnectX), **Broadcom** (Serie NetXtreme).
  - Switches: **Arista**, **Cisco Nexus**.
- Capacidades necesarias: Priorización de tráfico y control de flujo.

Configuración de la Red para Tráfico de Baja Latencia

- Implementación de Quality of Service (QoS).
- Uso de Priority Flow Control (PFC) para prevenir pérdida de paquetes.

Ejemplo: Configuración en Cisco Nexus (NX-OS)

- Habilitar Priority Flow Control (PFC)

```
configure terminal
interface ethernet 1/25
priority-flow-control mode on
```

- Configuración políticas QoS

```
class-map type qos match-any roce
match cos 3
policy-map type qos roce-policy
class roce
set qos-group 4
```

- Aplicar política QoS en interfaces

```
interface ethernet 1/25
service-policy type qos input roce-policy
```

- Ajustar tamaño del buffer

Data Center Bridging (DCB)

- **Definición**: Se trata de un conjunto de extensiones a la Ethernet tradicional que permite un transporte más confiable. Data Center Bridging (DCB) esencial para RDMA over Converged Ethernet.
- **Objetivo**: Garantiza el manejo eficiente del tráfico de datos de alta velocidad y la fiabilidad requerida por RDMA.
- **Necesidad**: RDMA realiza transferencias directas de memoria a memoria sobre Ethernet, reduciendo latencia y uso de CPU.
- **Desafío**: RDMA sobre Ethernet requiere de una red capaz de manejar tráfico congestionado y de alto rendimiento sin pérdida de paquetes.

Tecnologías que Componen DCB

- **Priority Flow Control (PFC)**: Evita la pérdida de paquetes mediante la pausa selectiva de flujos de datos. Prioridades en flujos y pausas por flujos por congestión.
- **Enhanced Transmission Selection (ETS)**: Asigna el ancho de banda entre flujos de tráfico, priorizando según necesidad.
- **Quantized Congestion Notification (QCN)**: Mitiga la congestión informando a los emisores sobre la necesidad de ajustar las tasas de envío.
- **DCB Capability Exchange (DCBX)**: Permite la configuración automática entre dispositivos para soporte uniforme de DCB.

Implementación de DCB para RoCE

- Importancia de DCB en redes convergentes.
- Herramientas para control de flujo y asignación de prioridades.

Planificación de la Topología de la Red para RoCE

- Diseño de topología para minimizar latencia (ej. malla, spine-leaf).
- Optimización para eficiencia de transferencia de datos.

Seguridad en Redes RoCE

- Implicaciones de seguridad del acceso directo a memoria.
- Medidas de seguridad necesarias: Segmentación de red, políticas de acceso.

UltraEthernet

Misión de UltraEthernet

- *«Ofrecer una arquitectura de pila de comunicaciones completa, abierta, interoperable y de alto rendimiento basada en Ethernet para satisfacer las crecientes demandas de red de IA y HPC a escala».*

¿Por qué UltraEthernet?

- UltraEthernet se presenta como una evolución de RoCE diseñada para satisfacer las necesidades de AI y HPC.
  - Mejoras en escalabilidad.
  - Manejo de la congestión.
  - Rendimiento y seguridad.

Mejoras de UltraEthernet respecto a RoCE

- Optimización para AI y HPC.
- Escalabilidad.
- Manejo avanzado de la congestión.
- Multipath y rociado de paquetes.
- Transporte UltraEthernet.
- Seguridad integrada y compatibilidad.
- Herramientas de desarrollo.

Optimización para AI y HPC

- La **latencia** es crítica en AI y HPC, donde las operaciones deben realizarse rápidamente para procesar grandes volúmenes de datos en tiempo real o en tiempos de ejecución aceptables.
- Se desarrollan **protocolos** de red que reduzcan la sobrecarga en las comunicaciones. Se busca rapidez.
- Se trabaja en la **compresión** de la información para reducir las transferencias y mejorar la eficiencia del ancho de banda.

Escalabilidad

- **Diseño de red modular**: Arquitecturas de red que permiten la expansión fácil y eficiente del sistema, permitiendo añadir más nodos al clúster sin degradar el rendimiento.
- **Soporte para clústeres de gran escala**: Mecanismos y protocolos de red diseñados para mantener un alto rendimiento incluso cuando el sistema se escala a millones de nodos, como es común en aplicaciones de HPC y centros de datos de AI.

Manejo avanzado de la congestión

- Las aplicaciones de AI y HPC suelen transferir mucha información y puede haber congestión.
- UltraEthernet introduce mecanismos de control de congestión dinámicos que utilizan:
  - ECN.
  - Control de flujo basado en prioridades (PFC).

Multipath y rociado de paquetes

- **Multipath**: permite la distribución de paquetes de datos a través de múltiples rutas de red.
- Maximiza:
  - Utilización.
  - Redundancia.
- **Rociado de paquetes**: Distribuye uniformemente la carga de datos a través de múltiples rutas.

Transporte UltraEthernet

- Diseñado para ser altamente eficiente y adaptable a las necesidades específicas de AI y HPC:
  - Optimización de la **inicialización de la conexión**.
  - Gestión de la **ventana de congestión**.
  - Manejo de **errores**.

Seguridad integrada y compatibilidad

- UltraEthernet incorpora características de seguridad directamente en la capa de red para asegurar que los datos críticos permanezcan seguros y protegidos en todo momento, incluyendo:
  - **Cifrado** de datos en tránsito.
  - **Autenticación** de nodos.
  - Protección contra ataques de denegación de servicio (**DoS**).

Herramientas de desarrollo

- Proporcionar interfaces de programación de aplicaciones (APIs) y herramientas que faciliten a los desarrolladores la creación, prueba, y despliegue de aplicaciones de AI y HPC sobre UltraEthernet, reduciendo la complejidad y acelerando el ciclo de innovación.

Bibliografía

Recursos utilizados

- https://blog.ipspace.net/2023/10/ultra-ethernet.html
- https://ultraethernet.org/wp-content/uploads/sites/20/2023/10/23.07.12-UEC-1.0-Overview-FINAL-WITH-LOGO.pdf
- https://ultraethernet.org/wp-content/uploads/sites/20/2023/09/23.08.10-UEC-Overview-Presentation-FINAL-2.pdf
- https://www.diva-portal.org/smash/get/diva2:1844498/FULLTEXT01.pdf
- https://www.nextplatform.com/2022/08/15/hpe-slingshot-makes-the-gpus-do-control-plane-compute/
- https://www.youtube.com/watch?v=z4GAr5cud7g
- https://enterprise-support.nvidia.com/s/article/understanding-rocev2-congestion-management
- https://enterprise-support.nvidia.com/s/article/rocev2-cnp-packet-format-example
- https://enterprise-support.nvidia.com/s/article/download-wireshark-with-rocev2-support
- https://www.fibermall.com/es/blog/rdma-key-technology-for-arithmetic-networks.htm
- https://www.nextplatform.com/2023/07/20/ethernet-consortium-shoots-for-1-million-node-clusters-that-beat-infiniband/
- https://www.theregister.com/2023/07/20/ultra_ethernet_consortium_ai_hpc/
- https://www.anandtech.com/show/18965/ultra-ethernet-consortium-to-adapt-ethernet-for-ai-and-hpc-needs
- https://ultraethernet.org/
- https://www.roceinitiative.org/wp-content/uploads/2016/11/SoftRoCE_Paper_FINAL.pdf

Muchas Gracias