Introducción a los DTNs:
Concepto y evaluación de rendimiento

Jorge Sasiaín, ***Eduardo Jacob****, Jasone Astorga y Juanjo Unzilla*

JJTT RedIRIS 2019 Sevilla, 28-30 de mayo

i2t Research Group

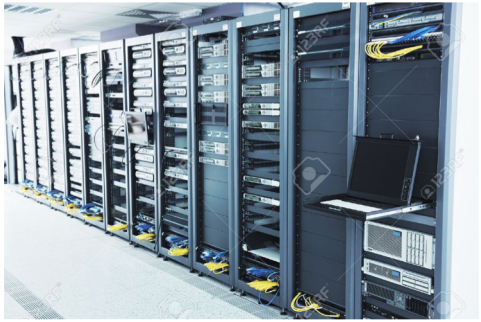
eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

¿Que es un DTN?

- Servidor Linux con componentes hardware de alta calidad
- Componentes hardware y software configurados (“tuneados”) con un objetivo principal
  - Maximizar el rendimiento en cuanto a lectura, escritura, y transmisión de datos a gran escala entre servidores situados en diferentes puntos de Internet



¿Que es un DTN?

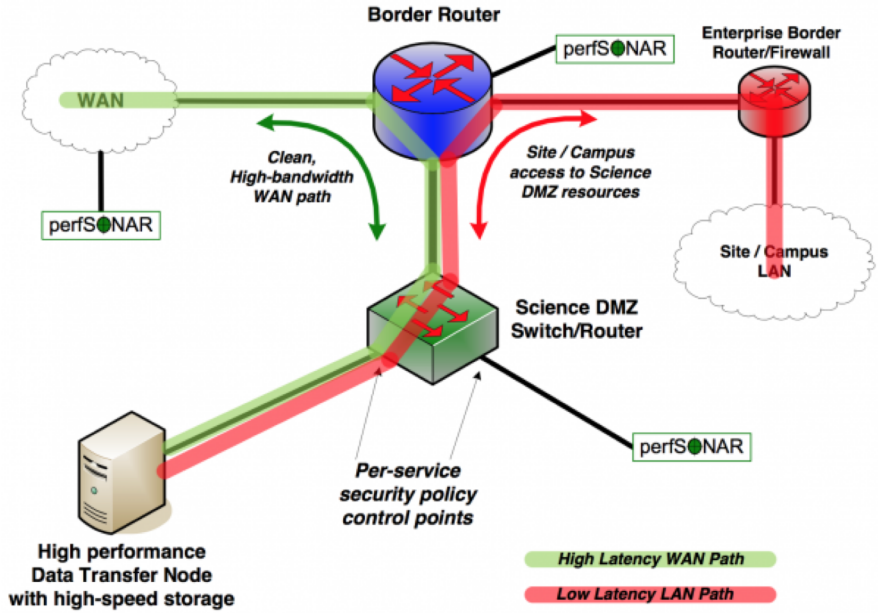
- Por consiguiente, para el óptimo funcionamiento de un DTN, se requiere
  - Servidor con acceso a un almacenamiento de alta velocidad
    - Disco local o conexión a una infraestructura de almacenamiento local
  - Interfaz o interfaces de red de alta velocidad
    - Típicamente entre 10 Gbps y 100 Gbps en función de cada caso
  - Infraestructura de red, tanto local como WAN, que soporte la capacidad en ancho de banda de los interfaces de red de los DTN

Origen y contexto de los DTNs

- Modelo de red Science DMZ (Esnet, ~2010)
  - El término “Science DMZ” proviene de las redes DMZ (demilitarized Zone)
  - Soporte para aplicaciones de alto rendimiento en las que la colaboración y compartición de grandes cantidades de datos científicos son críticas
  - Rendimiento necesario no soportado por redes de propósito general
    - Soporte de otros tipos de tráfico, procesamiento en firewalls, equipamiento de red con capacidad insuficiente
- Casos de uso fuera del ámbito científico
  - Supercomputación, Big Data, AI Training, Deep Learning, NFV, robótica
    - Punto en común: transferencia masiva de datos de forma rápida y eficiente

Science DMZ

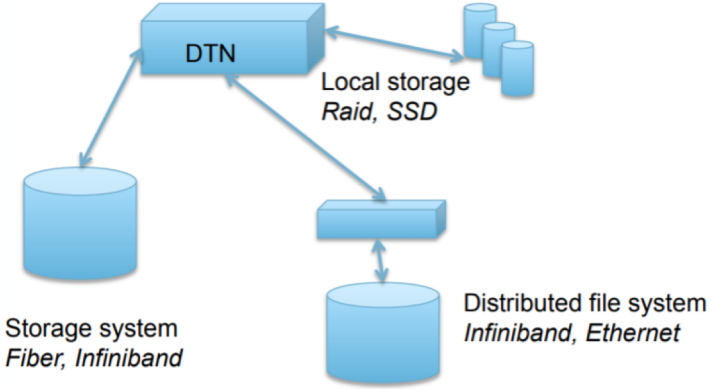
- Science DMZ separada de la red de propósito general
- Optimización independiente del resto de la red, sin comprometerse mutuamente
- El DTN se conecta a switch o router de alto rendimiento y se encarga de mover los datos científicos desde y hacia otras redes científicas



https://fasterdata.es.net

Hardware de un DTN

- Necesario el funcionamiento y desempeño óptimo de varios subsistemas
  - Sistema de almacenamiento
  - Sistema de red
  - Placa base y chasis
- Sistema de almacenamiento
  - Aspectos a considerar: disco, RAM, sistema de archivos, configuración RAID, canal de fibra al almacenamiento externo
  - Mejor opción: discos SSD NVMe
    - Hasta ~3 GB/s de lectura y ~2 GB/s de escritura por disco



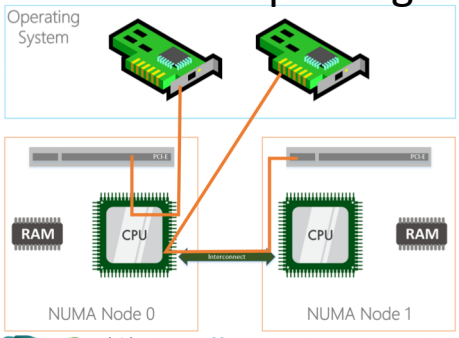
http://www.crc.nd.edu/~rich/OIN.10.2013/Science_DMZ/20131002-OIN-ScienceDMZ-3-DTN.pdf

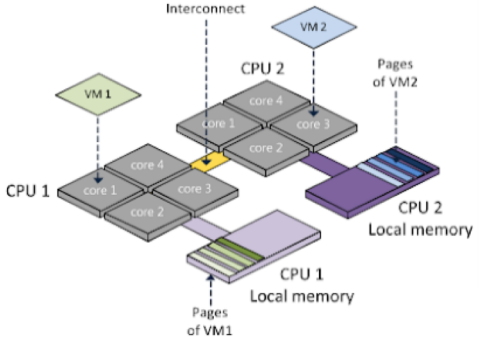
Hardware de un DTN

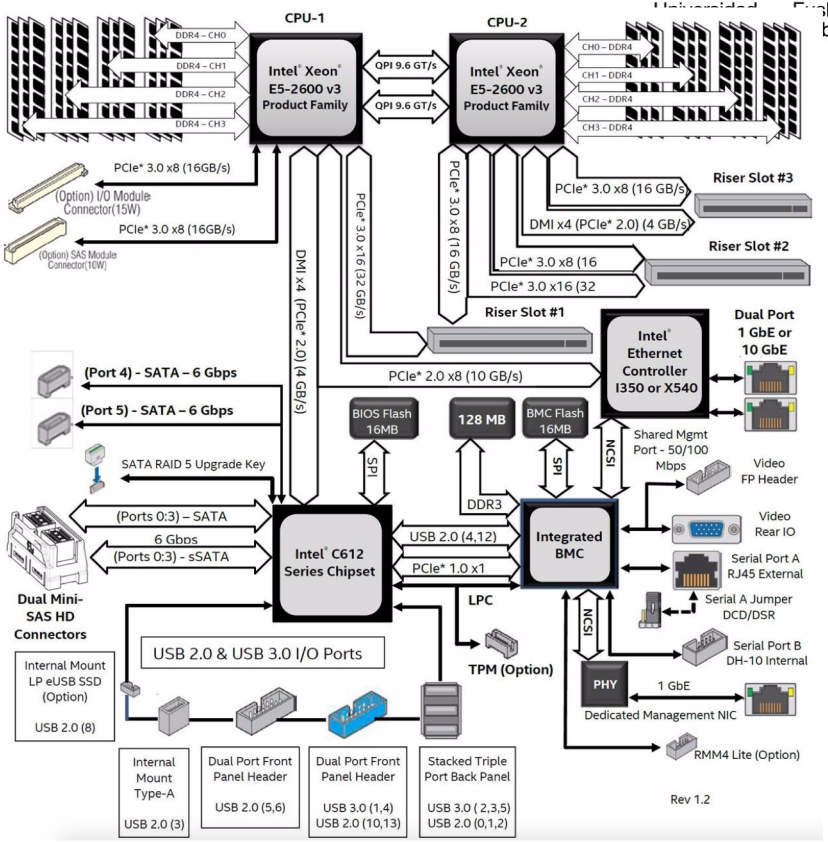
- Sistema de red
  - Interfaz/interfaces de red de alta calidad y velocidad (10G/40G/100G)
  - Soporte de varias optimizaciones
    - Coalescencia de interrupciones, MSI-X, TCP Offload Engine, TCP Segmentation Offload, UDP Fragmentation Offload, RDMA, …
- Placa base
  - Arquitectura de CPU, clock rate de CPU, PCI Express (generación PCIe de los slots y factor de forma), tipo y tamaño de la memoria RAM
  - Velocidad de buses QPI en arquitecturas NUMA
  - Refrigeración y fuente de alimentación adecuados

Arquitecturas de x86

- Intel® Server Board S2600WT
  - Aspecto crítico.
- Las arquitecturas son asimétricas.
  - Arquitecturas NUMA: Non Uniform Memory Access.
  - NIC y Buses
  - CPU pinning

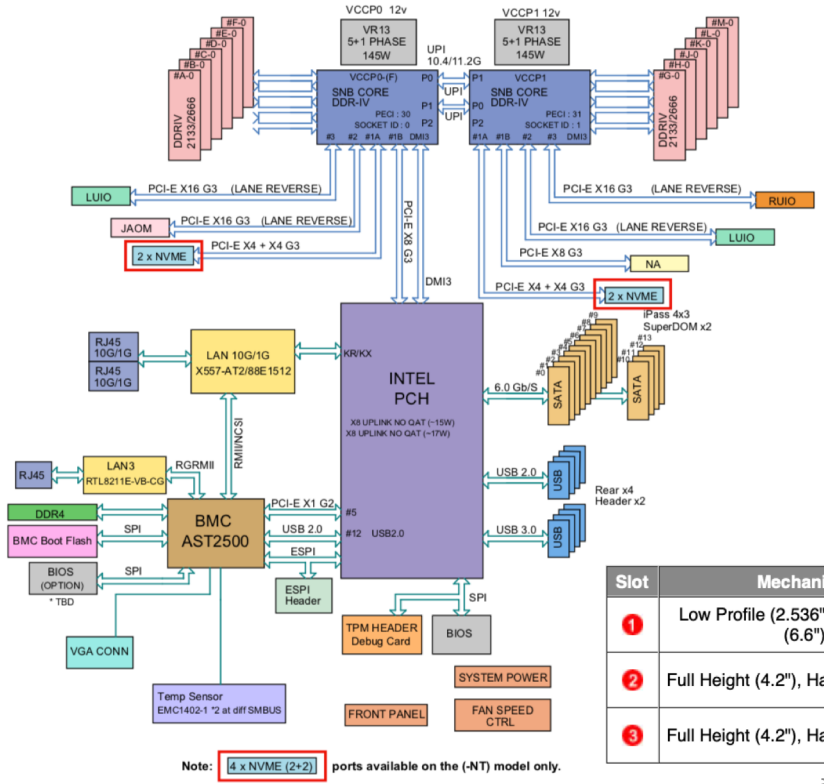






(1) http://www.darrylvanderpeijl.com/windows-server-2016-networking-optimizing-network-settings/
(2) http://openstack-in-production.blogspot.com/2015/08/numa-and-cpu-pinning-in-high-throughput.html

Un ejemplo de asignación de slots.



SuperServer 1029P-WTR

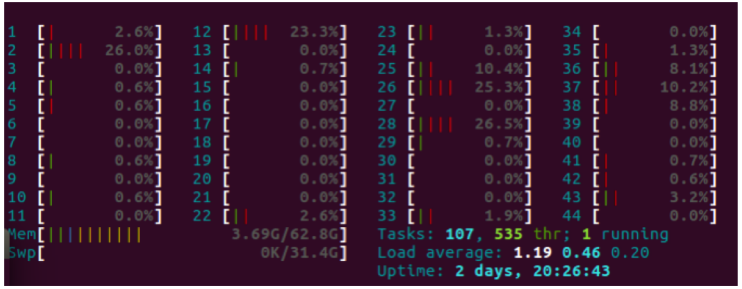
Slot	Mechanical	Electrical	Images / Illustration
1	Low Profile (2.536"), Half Length (6.6")	PCI-E 3.0 x8 (CPU2)	 SYS-1029P-WTR, SYS-1029P-WTRT, SYS-6019P-WTR
2	Full Height (4.2"), Half Length (6.6")	PCI-E 3.0 x16 (CPU2)	
3	Full Height (4.2"), Half Length (6.6")	PCI-E 3.0 x16 (CPU1)	

Tuning de un DTN

- La configuración por defecto no permite lograr el máximo rendimiento
- Necesario “tunear” varios componentes
  - BIOS, drivers, SO, sistema de red, sistema de archivos, memoria, I/O, …

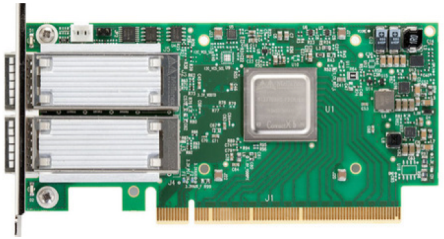
Tuning de un DTN

- Ejemplos de tuning
  - BIOS: asegurar que se trabaja en máximo rendimiento; deshabilitar Hyper-Threading y escalado de frecuencia, y habilitar Turbo Boost
  - Asignación de core óptimo a interrupciones (IRQs) en arquitecturas NUMA
  - Sistema de archivos: aumentar readahead y reducir/deshabilitar journaling
  - Memoria virtual: acelerar escritura de datos de memoria a disco



Tuning del sistema de red

- Objetivo: maximizar el rendimiento en la transmisión de datos a otro DTN a través de la red
  - No solo entra en juego el propio DTN; también ancho de banda, latencia, y características de la red y de nodos intermedios
  - Evitar DTN factor limitante: tuning de tarjeta(s) de red y protocolos TCP/UDP
- NIC
  - Configuración de buffers de transmisión y recepción (descriptores TX/RX)
    - Indican la longitud de la cola de tx/rx de la NIC
    - Compromiso entre throughput y latencia
    - En general, buffers grandes en DTNs



Tuning del sistema de red

- TCP
  - El rendimiento y throughput de TCP se ve afectado por múltiples factores
    - Latencia, tamaño de ventana, pérdida de paquetes, ...
  - Muy importante
    - Optimización de la capacidad de los buffers utilizados en la conexión
    - Selección de un algoritmo de control de congestión óptimo
    - Ajuste de la transmisión de la NIC a capacidad de la red y del receptor
  - Objetivos
    - Mayor acercamiento posible al valor Bandwidth-Delay Product ideal
    - Evitar degradación de throughput por eventos aleatorios (pérdida de paquetes, recepción fuera de orden…) no debidos a situación de congestión

Tuning del sistema de red

- UDP
  - Throughput UDP frecuentemente limitado por CPU
    - Aliviar la necesidad de procesamiento a la hora de generar y transmitir paquetes UDP a altas velocidades
      - Uso de “Jumbo Frames” (MTU = 9000)
      - Selección de core óptimo para interrupciones de la NIC en arquitecturas NUMA
  - Capacidad de buffers suficiente para la transmisión y recepción

Pruebas realizadas

- Pruebas realizadas
  - Lectura de disco
  - Escritura en disco
  - Rendimiento TCP
  - Rendimiento UDP
- Algunas herramientas utilizadas
  - dd
  - iperf3
  - htop
  - netem (insertar delay, pérdidas, duplicados en la red)

- Escenario de pruebas
  - Procesador Intel® Xeon® Processor E5-2699 v4 (2,2 GHz)
  - 2 nodos NUMA; 22 cores/nodo
  - NIC: Intel® Ethernet Controller X540-AT2 (10 GbE)
  - 64 Gb RAM
  - Discos SSD NVMe: Intel® DC P3700 SSDPE2MD400G4
  - Switch Dell EMC S4048-ON (10 Gbps) interconectando dos equipos similares.
  - Netem se ejecutaba en otro.

Algunos resultados y conclusiones

- Lectura de disco
  - Tamaño de bloque (`block size`) rendimiento similar y óptimo : entre ~16 kB y ~16 MB
  - `readahead` con rendimiento similar y óptimo: entre 8192 y 262144 bloques
  - En condiciones óptimas se alcanza 2630 MB/s para tamaño de fichero 32 GiB
    - Velocidad de lectura según especificación disco SSD: 2700 MB/s
  - Velocidad limitada por SSD (utilización CPU no alcanza 100%; máx. ~75%)
  - La velocidad de lectura decrece al aumentar el tamaño del fichero, especialmente con valores de `readahead` superiores a 262144 bloques

Algunos resultados y conclusiones

- Escritura en disco
  - Se utiliza tamaño de bloque 256 kB y tamaño de fichero 32 GiB; escritura en modo `fdatasync` (se garantiza que todos los datos son persistidos a disco)
  - El mejor throughput se consigue al disminuir `vm.dirty_background_ratio` (frecuencia de bloqueo del proceso de escritura) y al aumentar `vm.dirty_ratio` (inicio de escritura de memoria a disco)
    - Se alcanza 1020 MB/s, siendo especificaciones SSD 1080 MB/s
  - Influencia de journaling (utilizado sistema de archivos EXT4)
    - Modos writeback y ordered alcanzan el mismo throughput (1020 MB/s)
    - Con modo journal decrece significativamente (~360 MB/s)

Algunos resultados y conclusiones

- TCP
  - El impacto del utilizar tamaños de buffer no suficientemente grandes para el RTT dado es muy considerable
    - Limita directamente el tamaño de la ventana de congestión
    - La reducción del tamaño de buffer a la mitad puede implicar la disminución a la mitad del throughput logrado
    - Cuando el cuello de botella se da únicamente en uno de los extremos, la principal limitación se encuentra en el lado del receptor

Algunos resultados y conclusiones

- TCP
  - Comparación(1) entre algoritmos de control de congestion (no basados en retardo): HTCP, Cubic (default de Linux), y BBR
    - BBR ofrece un throughput de un orden de magnitud superior a HTCP Cubic en condiciones de 0.1% de pérdida de paquetes
      - BBR: 5.68 Gbps ; HTCP: 205 Mbps ; Cubic: 378 Mbps
    - Cubic se ha comportado ligeramente mejor que HTCP, pero ambos colapsan a partir de una tasa de pérdidas de 0.01%
      - BBR: 8.29 Gbps ; HTCP: 1.67 Gbps ; Cubic: 2.79 Gbps

(1) https://intronetworks.cs.luc.edu/current/html/newtcps.html

Algunos resultados y conclusiones

- UDP
  - Estudio de diferentes valores de MTU
    - Con “Jumbo Frames” se obtenido un throughput más de 3 veces superior al obtenido con tramas de MTU = 1500
  - Asignación de los procesos UDP cliente y servidor a diferentes cores
    - Cuando el core asignado a los procesos de transmisión y recepción pertenece al nodo NUMA en el que se encuentra el slot de la NIC, se obtiene una mejora del throughput de alrededor del 10% respecto al caso opuesto

Conclusiones

- El estudio es un poco limitado porque no hemos podido desplegar el experimento sobre una red de 10Gbs en operación que no sea de ámbito local.
- Para velocidades de 10Gbs con SSD NVMe y CPUs de 2,2 GHz (no el tope de gama) es posible saturar el enlace.
- Hay una configuración que tiene que recoger las características de la red.

Agradecimientos

- Este trabajo ha sido financiado en parte por:

  El Ministerio de Ciencia, Innovación y Universidades a través del proyecto TEC2016-76795-C6-5-R, de título GESTION FLEXIBLE DE SERVICIOS 5G ORIENTADA A SOPORTAR SITUACIONES CRITICAS URBANAS

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Introducción a los DTNs:
Concepto y evaluación de rendimiento

*Jorge Sasiaín, **Eduardo Jacob**, Jasone Astorga y Juanjo Unzilla*

JJTT RedIRIS 2019 - Sevilla, 28-30 de mayo

i2t Research Group