NARLabs 國家實驗研究院

國家高速網路與計算中心

# Data Transfer Node 介紹

楊哲男

2018/04/24

# Data transfer: Overview

The key players

- Endpoints
- Network
- Transfer tool
- Transfer settings

- That’ s a lot of work

# Data transfer time

- You are getting speeds less than (for large datasets):
  - Within campus
    - 800 Mbps (Mega bits per second).
      - **100 GB** = ~820,000 Mb. Takes ~ 17 minutes
    - 3-4 Gbps if you have a 10G connection
      - **100 GB** = ~800 Gb. Takes ~ 4 minutes
  - Between campus and outside
    - Hard to tell because of things out of our control
    - 200 Mbps – 5000 Mbps (5 Gbps)
      - **100 GB** = ~820,000 Mb. Takes ~ 1 hour

# Science DMZ

其目的為如何讓科學資料能夠最佳化的在廣域網路上傳輸的設計方法。Science DMZ整合了三大重要關鍵因素，來達成快速傳遞資料之目的:

1. 適合於High-performance應用之專用網路，最好能與一般使用之網路區隔，避免防火牆，直接連接border router，減少其他網路設備發生錯誤時所帶來的影響，亦可減少除錯時之時間。
2. 擁有專屬的資料傳輸機器(群)，例如建置DTN(Data Transfer Node)並配合高效能的傳輸軟體及機器調教。
3. 良好的效能量測機制，可隨時量測點對點間之網路，以確保網路品質良好。

# Network Monitoring

- delay與packet loss影響網路效能
  - delay越長，若不做任何調整，效能將不如預期
  - 當loss發生，封包必須重傳，並且降速，會造成效能之惡性循環
  - 因為loss發生所需回復的時間，亦將造成RTT時間變長，因此效能又會降低
- 隨時監控網路delay、對外網路可用頻寬等資訊
- 長時間觀察網路品質及趨勢

# The Data Transfer Node (DTN)

- Dedicated, high-performance host for data transfer
- Typically PC-based Linux servers built with high-quality components and configured specifically for wide area data transfer
- Proper tools

# DTN subsystems

- 儲存
  - Local storage(RAID, SSD)
  - Networked storage- Distributed file system(Infiniband, SAN)
- 網路
  - 10G以上之網路或專用網路
- 主機
  - Server bus(PCI-e)
  - File system(ext4, xfs, btrfs)
- Tuning
  - bios、CPU、IRQ、儲存、網路、檔案系統、應用軟體

# NUMA Issues

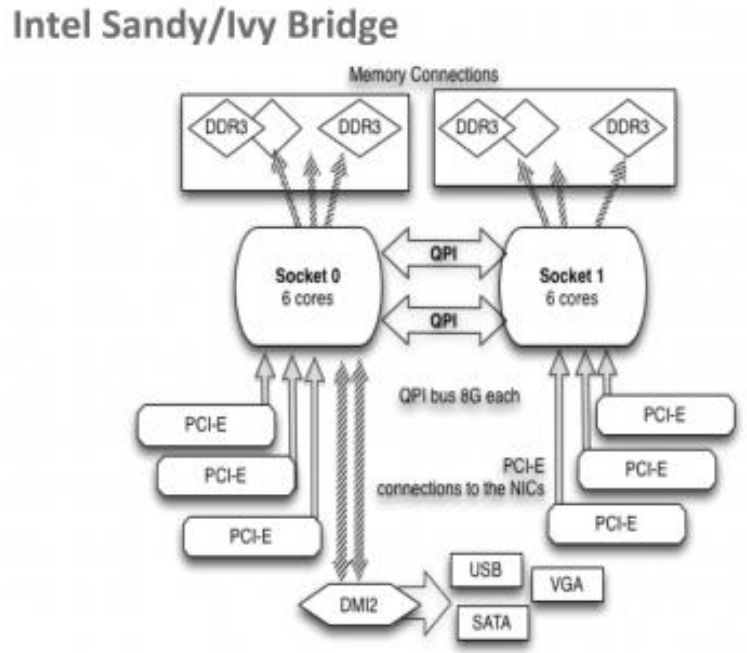
- Up to 2x performance difference if you use the wrong core.
- If you have a 2 CPU socket NUMA host, be sure to:
  - Turn off irqbalance
  - Figure out what socket your NIC is connected to:

    cat /sys/class/net/ethN/device/numa_node

  - Bind your program to the same CPU socket as the NIC:

    numactl -N 1 program_name
- Which cores belong to a NUMA socket?

  cat /sys/devices/system/node/node0/cpulist



# DTN Storage

- 儲存種類
  - Local Storage(RAID, SSD): ex: NVME SSD
  - External Storage: Distributed file system(ex:Lustre檔案系統), lustre client mount
- 連接方式
  - DTN透過Ethernet、IB、omni-path連接後端DFS
  - NFS mount

# Data Transfer Tools

- anonymous: anyone can access the data.
  ex: FTP HTTP(wget)
- simple password: most sites no longer allow this method since the password can be easily captured.
  ex: FTP HTTP(wget)
- password encrypted: control channel is encrypted, but data is unencrypted.
  ex: bbcp, bbftp, GridFTP, FDT
- everything encrypted: both control and data channels are encrypted.
  ex: scp, sftp, rsync over ssh, GridFTP, HTTPS-based web server

# Secure Copy (SCP)

- **Secure Copy (SCP)**
  - Widely used for file transfers
  - Uses SSH for authentication and data transfer (TCP port 22)
  - Unix-based systems
  - Windows: WinSCP
- 若需使用scp或是rsync等傳輸軟體，可更新OpenSSH版本(例如hpn-ssh)

```
[sun1@tp-server1 ~]$ scp sun1@chi-server1:/home/100GB /home/sun1/worker/
100GB                                         0%  386MB   9.1MB/s 3:03:14 ETA
```

# rsync

- **rsync (rsync over SSH)**
  - Sync files and directories between two endpoints
  - Good for running backups
  - Unix-based systems
  - Windows: CwRsync

# File Transfer Protocol (FTP)

- **‘Secure’ File Transfer Protocol (‘S’ FTP)**
  - Widely used for file transfers
  - SFTP is more secure. Use it if available
  - Unix-based systems (including Mac OS X): Should have it by default
  - Windows: FileZilla (https://filezilla-project.org)

# These tools are okay, but not always

- Great compatibility. Widely available.
- Small datasets. Quick transfers. (< 15 mins)

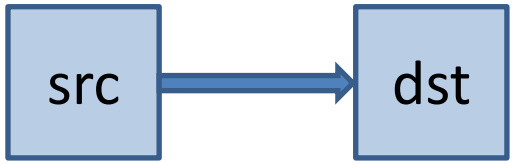
- Large bulk data transfers.
- Transfers on unreliable connections and hosts.

# Data Transfer Tools

- Parallelism is key
  - It is much easier to achieve a given performance level with four
  - parallel connections than one connection
  - Several tools offer parallel transfers
- Latency interaction is critical
  - Wide area data transfers have much higher latency than LAN transfers
  - Many tools and protocols assume a LAN
  - Examples: SCP/SFTP

# Single vs multi stream

- Single stream
  - scp
  - ftp
  - rsync




Better utilization of link

**Faster transfer speed**

- Multi stream(建議使用)
  - GridFTP
  - BBCP
  - FDT
  - Nuttcp
  - Aspera

# GridFTP

- GridFTP from ANL has features needed to fill the network pipe
  - Buffer Tuning
  - Parallel Streams
- Supports multiple authentication options
  - Anonymous
  - **ssh**
  - X509

```
[sun1@tp-server1 ~]$ globus-url-copy -vb  -p 10 -sync  \
sshftp://sun1@chi-server1/home/100GB   file:///dev/null
Source: sshftp://sun1@chi-server1/home/
Dest:   file:///dev/
 100GB  ->  null

 25728909312 bytes      766.78 MB/sec avg      995.20 MB/sec inst
```

# BBCP & FDT

- BBCP
  - comparable performance to Globus
  - Mac OS X, Linux-based systems. SSH-based access control
  - Both endpoints need it installed, but easier to install and configure

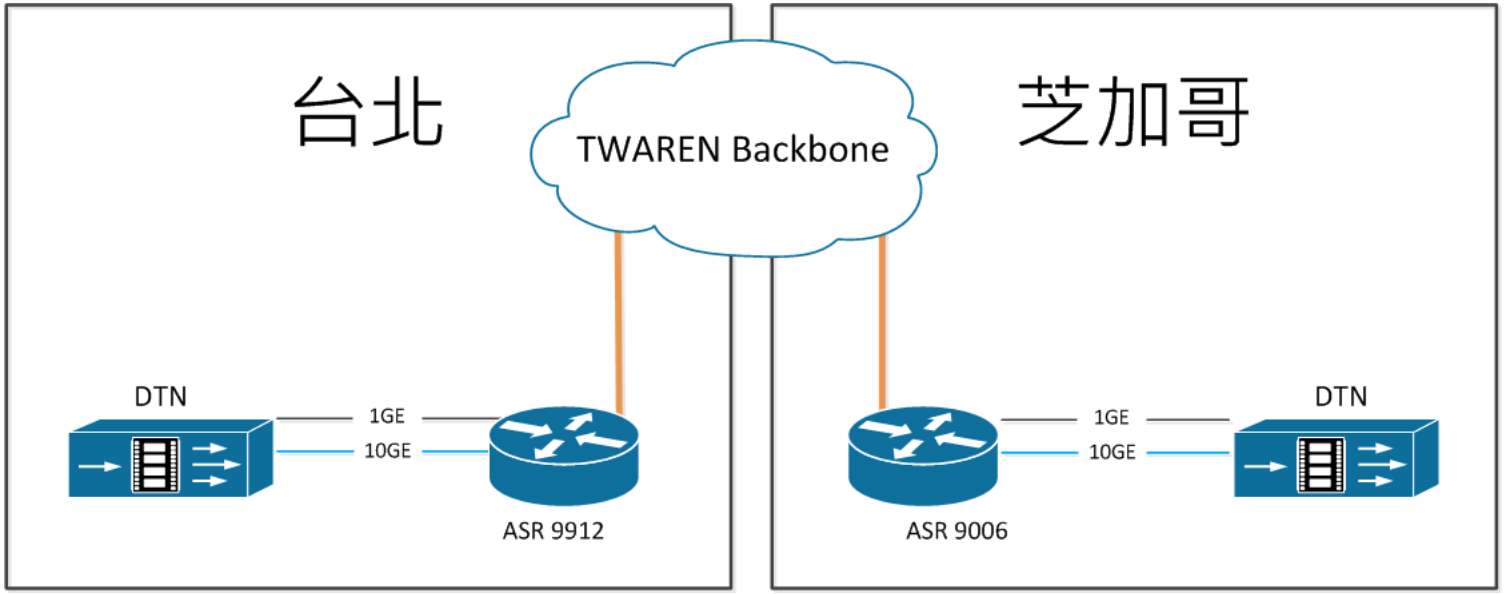
"$bbcp -V -s 16 /local/path/largefile.tar remotesystem:/remote/path/largefile.tar"

- Fast Data Transfer (FDT)
  - Java-based tool from Caltech & CERN
  - Can theoretically run in any Operating System, including Windows
  - Need server-side running in server mode

"$java -jar fdt.jar -c <remote_address> -d destinationDir ./local.data"

# Demo

- 架構




- 從芝加哥節點傳輸100GB檔案至台北節點

/home/sun1/100GB to /dev/null (disk to ram)

/home/sun1/100GB to /home/sun1/worker/ (disk to disk)

# Summary

- 為了得到較佳之傳輸效能，我們可以：
  - 單一solution無法解決所有問題，有時調整完也不見得會得到效果
  - Large bulk transfers: 建置DTN並透過DTN傳輸
  - 確定可用頻寬是足夠的，或可建立傳輸專用頻寬
  - 確定TCP之相關參數已調整
  - 為了得到較好之傳輸效能，可規劃檔案傳輸DMZ架構
  - 可使用多重(個)串流之方法
  - 考量磁碟I/O效能，可使用磁碟陣列或是GPFS、Lustre等平行分散檔案系統
  - 需注意網路品質，特別是RTT與packet loss
  - 長期監控效能，了解網路品質是否有異常(即時今天網路沒問題，也無法保證明日網路是好的)