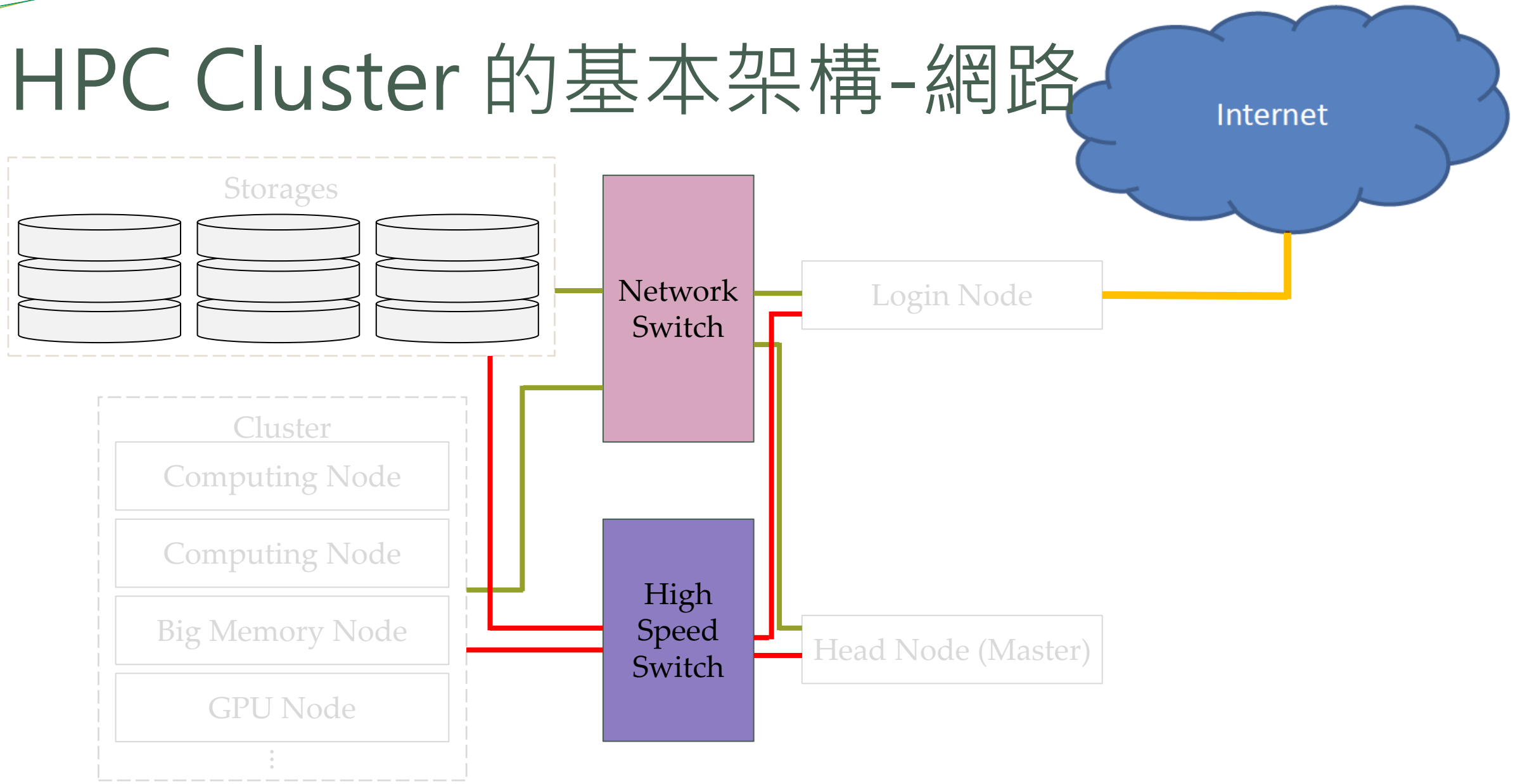
HPC 網路

國立臺灣師範大學物理學系 陳俊明

chunming@ntnu.edu.tw

HPC Cluster 的基本架構-網路

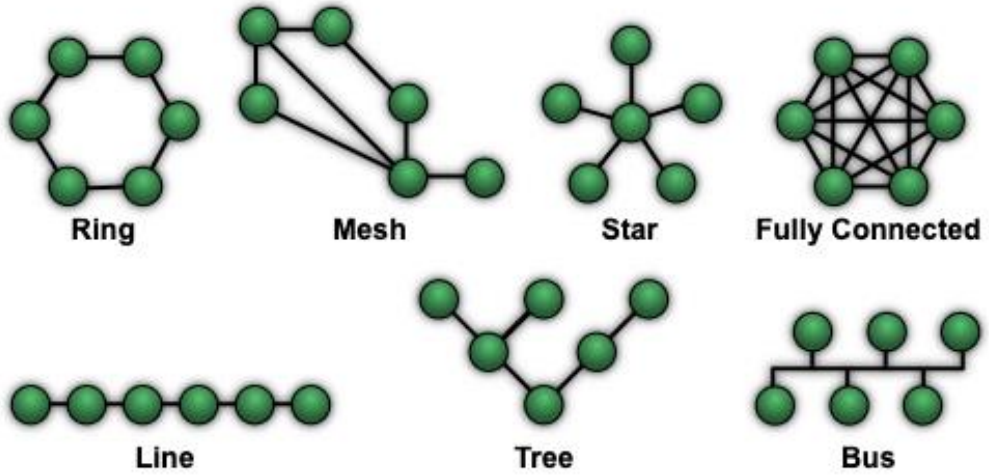


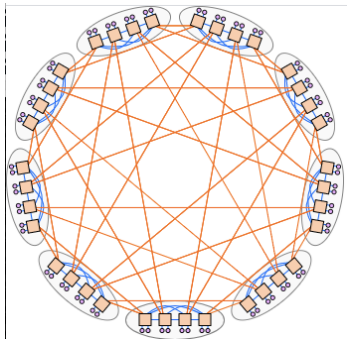
OSI七層模型
(Open System Interconnection Reference Model)

應用層 (Application)	應用軟體
表現層 (Presentation)	轉換網路封包格式
會談層 (Session)	確認網路端點之間的連線
傳輸層 (Transport)	定義網路傳輸技術
網路層 (Network)	定義網路IP
資料鏈結層 (Data Link)	定義軟體封裝
實體層 (Physical)	定義硬體實體

網路拓撲

- 點對點 (Point-to-point)
- 匯流排拓撲 (Bus)
- 星狀拓撲 (Star)
- 環狀拓撲 (Ring)
- 網狀拓撲 (Mesh)
- 樹狀拓撲 (Tree)
- 菊花鏈拓撲 (Daisy Chain)
- 蜻蜓拓撲 (Dragonfly)





上圖：Maksimderivative work: Malyszkz (talk) - NetworkTopologies https://commons.wikimedia.org/w/index.php?curid=15006915
下圖：cyberang3l / CC0 https://commons.wikimedia.org/wiki/File:Dragonfly-topology.svg

乙太網路 (Ethernet)

- 為現今最廣泛使用的區域網路類型。
  拓樸邏輯為匯流排型拓樸
  - Ethernet：10BASE
  - Fast Ethernet：100BASE
  - Gigabit Ethernet：1000BASE
  - 10Gb Ethernet：10GBASE
  - 40Gb Ethernet：40GBASE
  - 100Gb Ethernet：100GBASE
  - 200Gb Ethernet：200GBASE
- 準備向 Terabit Ethernet (TbE) 邁進

Ethernet的OSI模型

應用層 (Application)	應用層	HTTP, FTP, SMTP, POP3, NFS, SSH
表現層 (Presentation)		
會談層 (Session)		
傳輸層 (Transport)	傳輸層	TCP, UDP
網路層 (Network)	網路層	IP, ICMP
資料鏈結層 (Data Link)	資料鏈結層	Ethernet MAC
實體層 (Physical)		Ethernet Pysical

Ethernet的硬體

- 網路卡：1 / 10 / 25 / 40 /100 / 200 Gbps
- 網路線
  - Cat 5e, Cat 6, Cat 6a, Cat 7, Cat 8
  - SFP, SFP+, QSFP, QSFP+, XFP
  - CX4
- 網路設備
  - (router)
  - switch
  - (bridge)
  - (hub)
  - (repeater)

網路設備額外資訊：https://notfalse.net/66/repeater-hub-bridge-switch

InfiniBand

InfiniBand為一種通訊傳輸標準，具有低網路延遲及非常高的網路傳輸帶寬，原生支援遠端記憶體直接存取(Remote Direct Memory Access, RDMA)

- 10Gb IB：SDR (Retired)
- 20Gb IB：DDR (Retired)
- 40Gb IB：QDR
- 56Gb IB：FDR
- 100Gb IB：EDR
- 200Gb IB：HDR

InfiniBand Bandwidth

- IB bandwidth = IB width x Single lane speed
- IB width: 1,4,8,12X
- Single lane speed

Notation	Year	Signal rate Gb/s
SDR	2002	2.5
DDR	2005	5
QDR	2008	10
FDR	2011	14
EDR	2013	25
HDR	2017	50

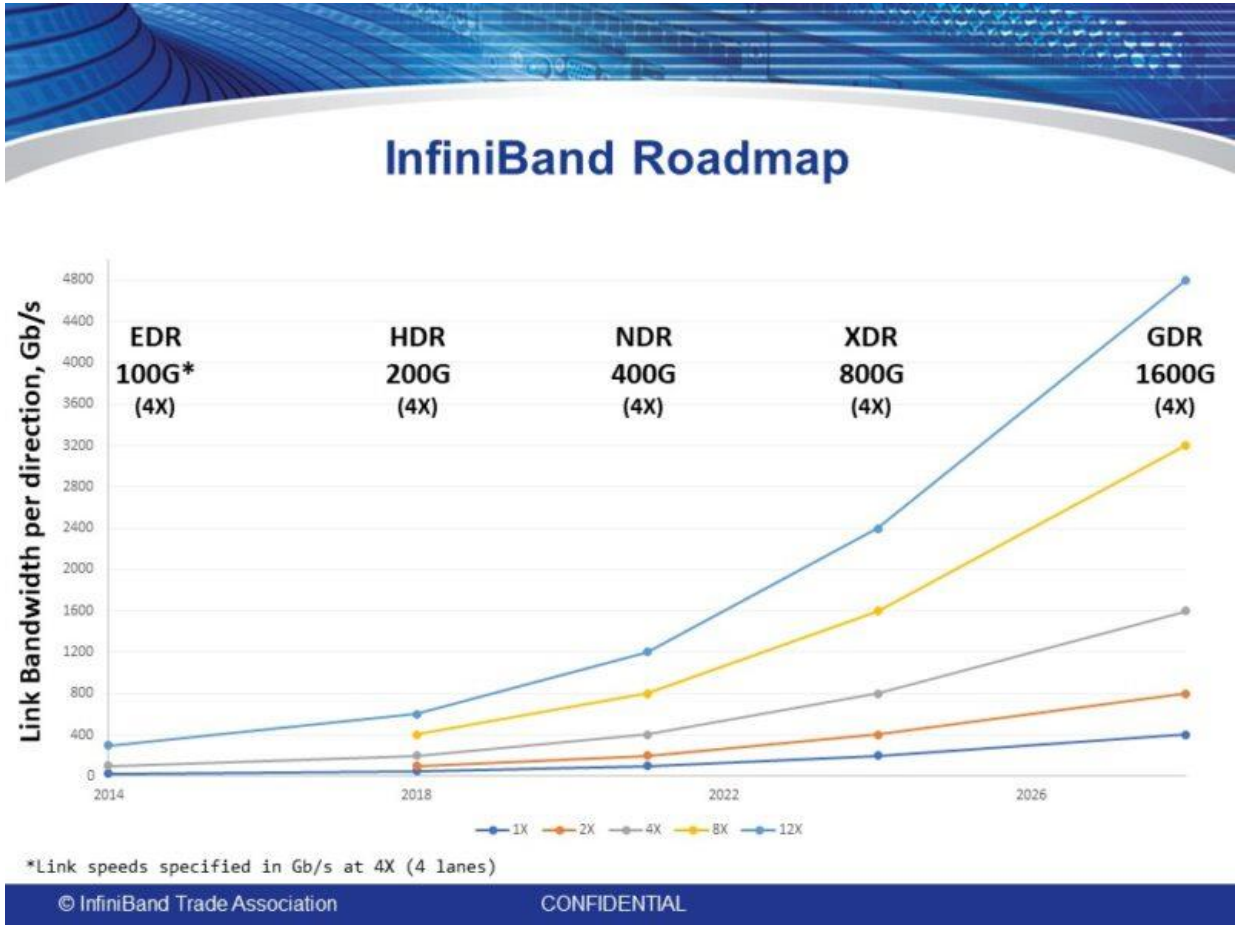
EDR 4X 4 x 25 Gb/s = 100Gb/s

Subnet Managment

SM (Subnet Manager) 用於建立及管理 Infiniband 網路

- 可用 opensm 軟體管理或用有管理功能的交換器
- InfiniBand 網路特色
  - 隨插即用
  - 集中式管理
  - 1 個 SM 可以同時管理48,000 個 IB 端點

InfiniBand Roadmap

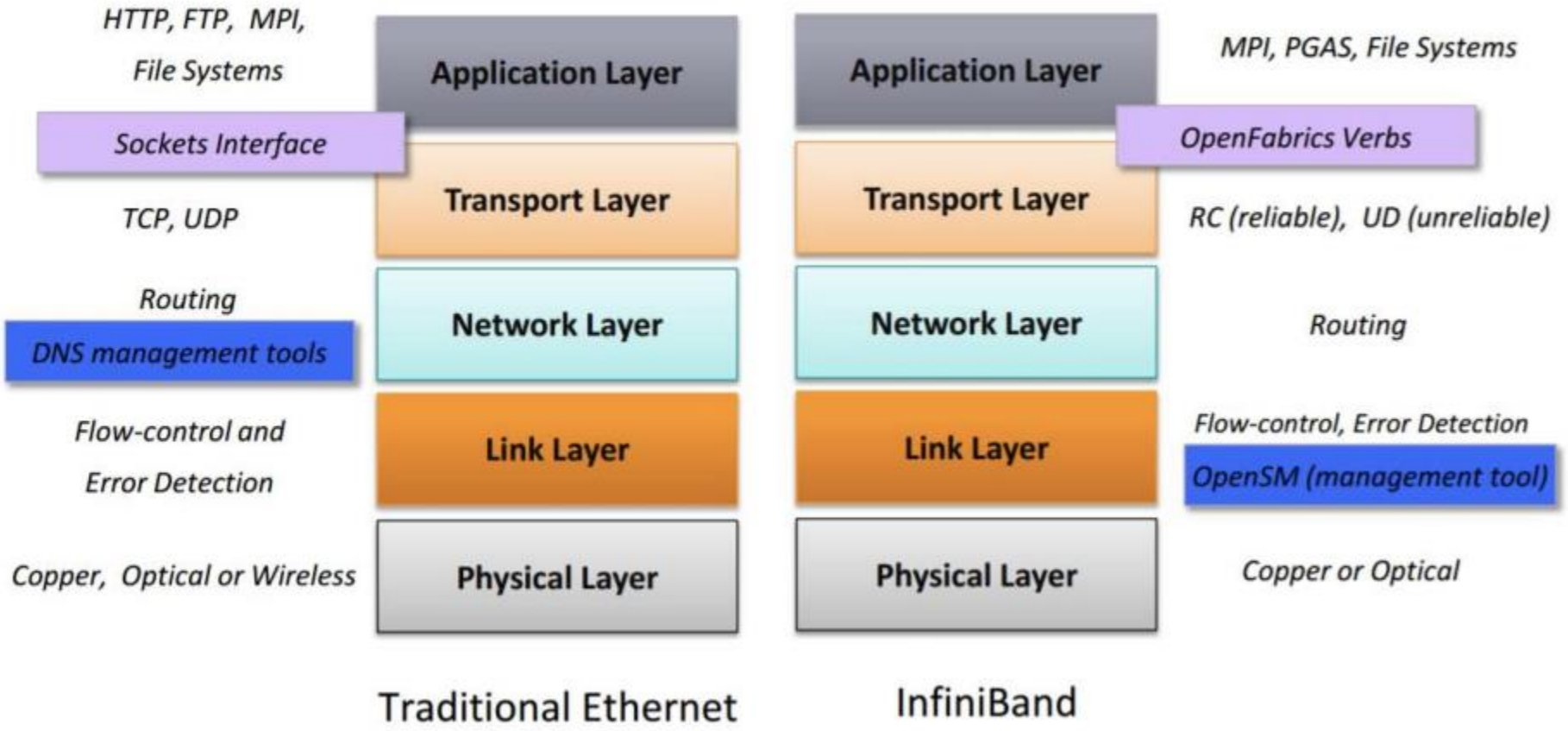


資料來源：https://www.infinibandta.org/infiniband-roadmap/

InfiniBand的OSI模型

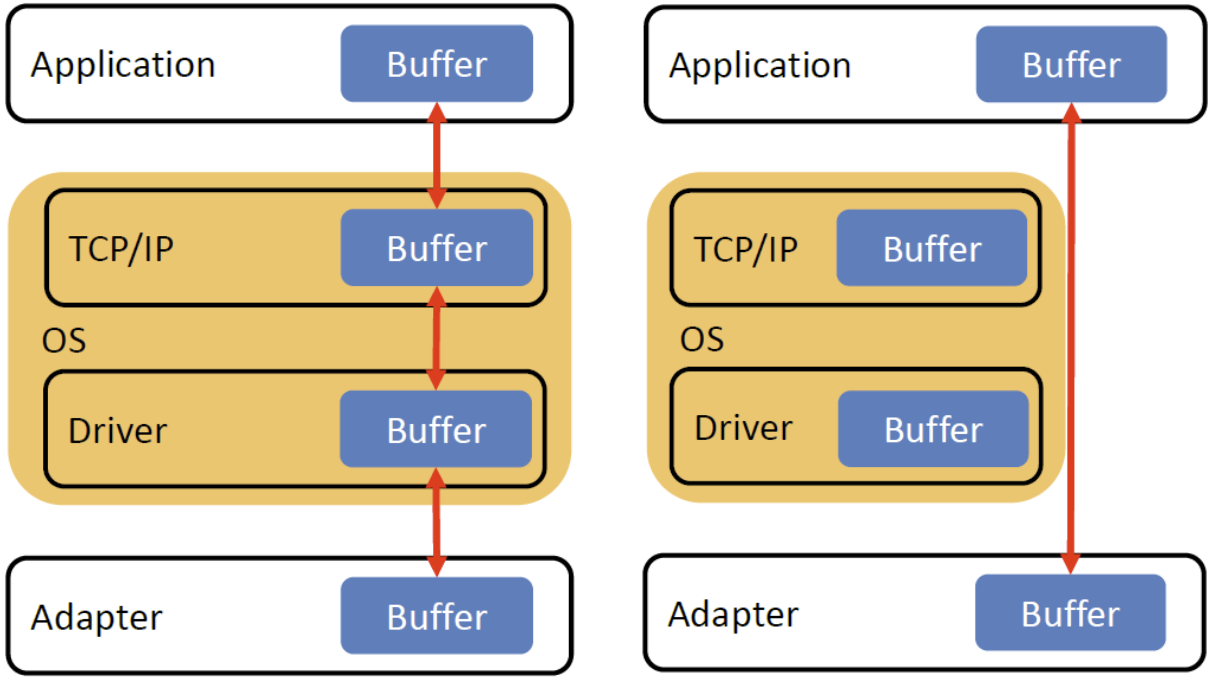
應用層 (Application)	應用層	NFS MPI
表現層 (Presentation)		
會談層 (Session)		
傳輸層 (Transport)	傳輸層	OFA
網路層 (Network)		
資料鏈結層 (Data Link)	資料鏈結層	InfiniBand
實體層 (Physical)		

InfiniBand 與乙太網路比較



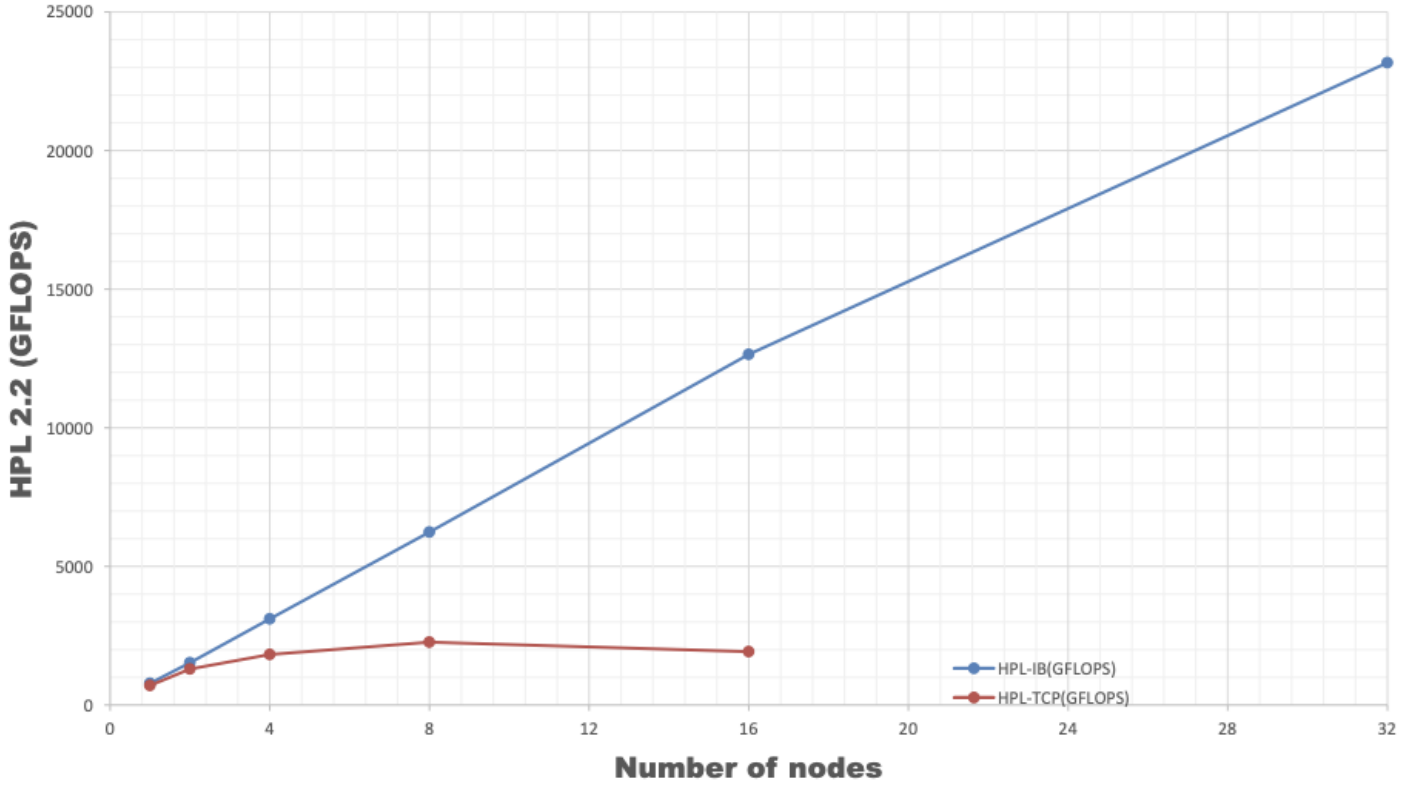
RDMA優點

- 跳過系統軟體層直接處理網路流量
  - 使用者端的應用程式直接跳過kernel進行資料交換
  - 核心端的程式直接跳過系統驅動程式
- 從某機器直接將資料抄寫到另一台機器不用複製到各個暫存區 (buffer)
- 增加傳輸速度降低延遲及 CPU 使用時間



傳統模式　　RDMA模式

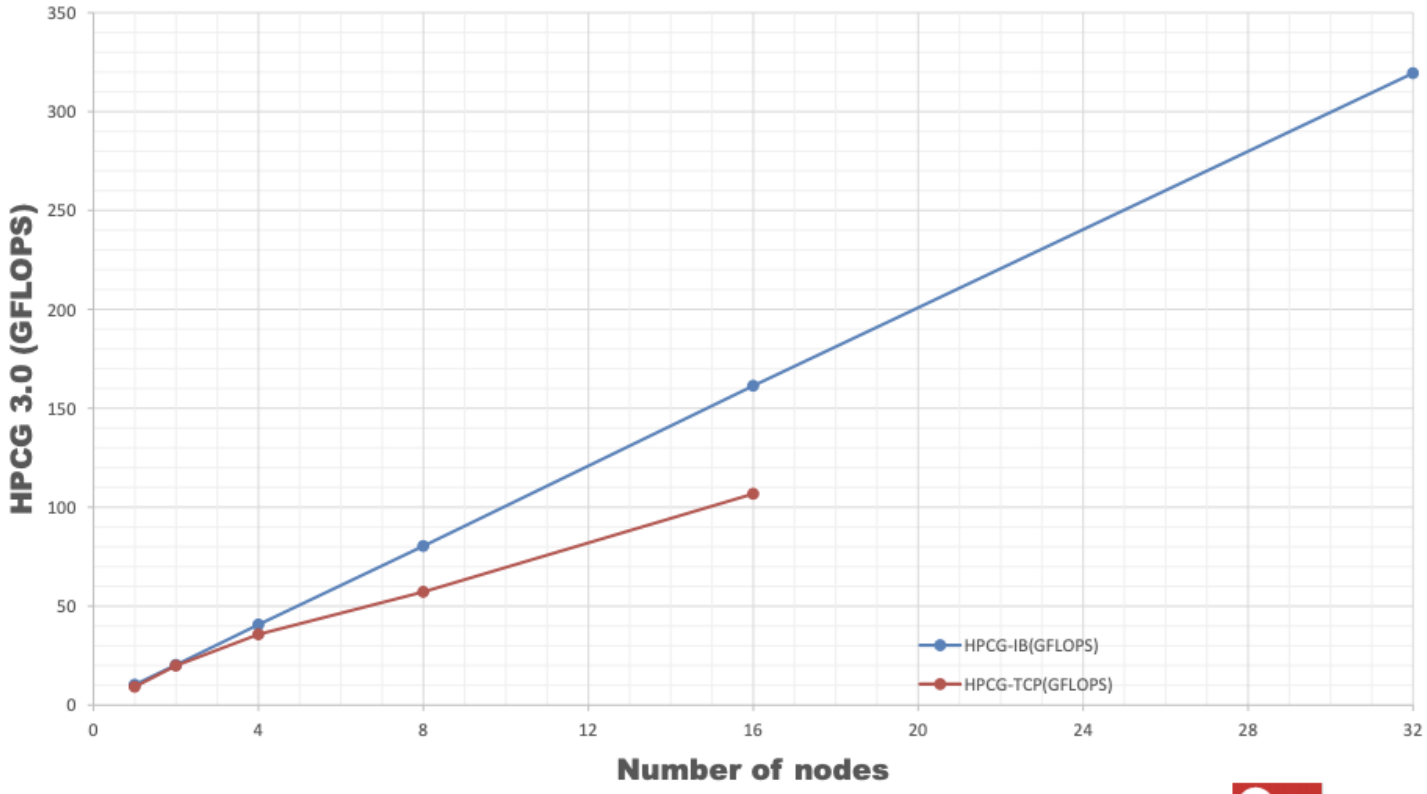
RDMA vs TCP - HPL 2.2



Support by Dr. James Chen, National Supercomputing Center, Singapore

National Supercomputing Centre

RDMA vs TCP - HPCG 3.0



Support by Dr. James Chen, National Supercomputing Center, Singapore

National Supercomputing Centre

InfiniBand的優點

- 高吞吐量 (Higher throughput) ：200Gb/s 已經有完整的配套，乙太網路緊跟在後
- 低延遲 (Lower latency)：HDR 網卡點對點低於 600ns，降低作業系統額外負擔(overhead) 使得資料傳遞更有效率
- 強化擴展性 (Enhanced scalability)：網路容易橫向擴展，只要不夠接加上交換器免額外設定，網路就連通。
- 較高 CPU 使用效率 (Higher CPU efficiency)：資料搬移的工作不在 CPU 身上，所以 CPU 可以做更多事情
- 較好投報率 (Better ROI)：高吞吐量及CPU使用效率降低使用成本

InfiniBand 網路硬體

- 網路卡：HCA (Host Channel Adapter) card ConnectX-6 VPI, ConnectX-5 VPI
- 網路線及接頭：DAC (Direct Attach Copper) Cables, AOCs (Active Optical Cables), Optical Transceivers
- 網路設備：Switch, Gateway, Router
  - Mellanox(Nvidia Networking)
  - Intel Omni-path

InfiniBand 網路設定

- 提供 OpenSM 服務（SM Server only，選其中一台擔任）

```
# systemctl start opensm
# systemctl enable opensm
```

- OFED (OpenFabrics Enterprise Distribution) 每個運算節點

```
# systemctl start rdma
# systemctl enable rdma
```

OFED 可由 Mellanox / 設備原廠官網下載安裝或是用系統安裝 yum groupinstall "Infiniband Support"

- 網路設定

```
# vi /etc/sysconfig/network-scripts/ifcfg-ib0
TYPE=InfiniBand
NAME=ib0
DEVICE=ib0
ONBOOT=yes
IPADDR="192.168.2.100“
NETMASK="255.255.255.0"
```

```
# vi /etc/sysconfig/network-scripts/ifcfg-enp0s3
TYPE=Ethernet
NAME=enp0s3
DEVICE=enp0s3
ONBOOT=yes
IPADDR="192.168.2.100“
NETMASK="255.255.255.0"
```

InfiniBand 原廠 OFED 驅動程式安裝

- Linux InfiniBand Drivers 下載
- 安裝程序隨著不同版本會有所差異

```
# tar zxvf MLNX_OFED_LINUX-23.07-0.5.0.0-rhel8.8-x86_64.tgz
# cd MLNX_OFED_LINUX-5.7-1.0.2.0-rhel7.9-x86_64
# ./mlnxofedinstall -v --enable-affinity --enable-mlnx_tune --all --force
# systemctl start openibd
# systemctl enable openibd
```

- 安裝好後手動改 ifcfg-ib0

```
# vi /etc/sysconfig/network-scripts/ifcfg-ib0
```

設定 limits.conf 檔案

- RDMA 會讓使用者對應記憶體到設備端，所以建議把 memlock 設定成不限制。至於堆疊 (stack) 大小也建議設定為不限制，這和 RDMA 沒關。

```
# vi /etc/security/limits.conf
*               hard    memlock         unlimited
*               soft    memlock         unlimited
*               hard    stack           unlimited
*               soft    stack           unlimited
```

檢查 IB 網卡狀態

```
# ibstat
CA 'mlx4_0’
        CA type: MT4099
        Number of ports: 2
        Firmware version: 2.35.5100
        Hardware version: 0
        Node GUID: 0x50014850002aa240
        System image GUID: 0x50014850002aa243
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 56
                Base lid: 44
                LMC: 0
                SM lid: 9
                Capability mask: 0x02594868
                Port GUID: 0x50014850002aa241
                Link layer: InfiniBand
```

確認程式是否支援IB

- 將程式編譯好後用 ldd 指令確認是否使用到 IB 的 shared library

```
# ldd YOUR_BINARY
        linux-vdso.so.1 =>  (0x00007ffee718a000)
        libpthread.so.0 => /lib64/libpthread.so.0 (0x00002b0167519000)
        libm.so.6 => /lib64/libm.so.6 (0x00002b0167735000)
        libdl.so.2 => /lib64/libdl.so.2 (0x00002b0167a37000)
        libmpi_usempif08.so.0 => /opt/openmpi-1.8.8-intel17/lib/libmpi_usempif08.so.0 (0x00002b0167c3b000)
        libmpi_usempi_ignore_tkr.so.0 => /opt/openmpi-1.8.8-intel17/lib/libmpi_usempi_ignore_tkr.so.0 (0x00002b0167e7b000)
        libmpi_mpifh.so.2 => /opt/openmpi-1.8.8-intel17/lib/libmpi_mpifh.so.2 (0x00002b0168085000)
        libmpi.so.1 => /opt/openmpi-1.8.8-intel17/lib/libmpi.so.1 (0x00002b01682ee000)
        libc.so.6 => /lib64/libc.so.6 (0x00002b016889c000)
        libgcc_s.so.1 => /lib64/libgcc_s.so.1 (0x00002b0168c5d000)
        /lib64/ld-linux-x86-64.so.2 (0x00002b01672f7000)
        librdmacm.so.1 => /lib64/librdmacm.so.1 (0x00002b0168e73000)
        libibverbs.so.1 => /lib64/libibverbs.so.1 (0x00002b016908a000)
        libopen-rte.so.7 => /opt/openmpi-1.8.8-intel17/lib/libopen-rte.so.7 (0x00002b016929d000)
        libtorque.so.2 => /usr/local/lib/libtorque.so.2 (0x00002b01695c2000)
 …(omit)
```

排除網路狀態

- 確認硬體
  - 網路線連接是否正確
  - 網卡燈號是否正常
- 由指令確認環境
  - ping
  - traceroute
  - nslookup
  - arping
  - route
  - ibstat, iblinkinfo, ibnetdiscover, ibswitches, ibhosts, ibstatus
  - smpquery