# HPC 簡介

國立臺灣師範大學物理學系 陳俊明

chunming@ntnu.edu.tw

# Outline

- 什麼是 HPC？
- HPC 的應用
- HPC Cluster 的基本架構
- HPC 實例
- HPC 規劃、趨勢與主流

# 什麼是 HPC？

# 什麼是 HPC？

**High Performance Computing(HPC)**

泛指可在較短的時間內計算一般電腦需長時間運算的複雜系統，以提高應用能力，以解決科學、工程或商業中的大問題。

- 計算能力(PFlop/s)
- 速度(時間)
- 能源的消耗(kW)
  - Energy Efficiency: GFlops/W
  - PUE(Power Usage Effectiveness) = (Total Usage) / (IT Usage)

# 平行計算的分類

- **多核心處理計算 (Multi-core computing)**：一個處理器內有多個處理單元 (processing units) 或稱核心 (core)
- **同時多執行緒 (Simultaneous multithreading)** 又稱超執行緒 (hyper-threading)。
- **對稱多處理 (Symmetric multiprocessing)**：電腦有多個處理器共用主記憶體。例如：2 way, 4 way, 8 way
- **分散式運算 (Distributed computing)**：位於不同地點的電腦透過網路相互連接傳遞訊息與通訊後，並協調它們的行為以達成共同目標而形成的系統。
- **大規模並行運算(Massively parallel computing)**：是由多個微處理器，局部存儲器及網路構成的節點 (node) 組成的並行計算系統；節點間以高速網路。大規模並行處理是一種異步的多指令及數據，因為它的程序有多個程序 (process)，它們分布在各個微處理器上，每個程序有自己獨立的地址空間，程序之間以訊息傳遞介面 (MPI) 進行相互溝通。
- **叢集運算 (Cluster computing)**：一組連接在一起工作的電腦。由於這些電腦協同工作，在許多方面它們可以被視為單個系統。與網格電腦不同，電腦叢集將每個節點設定為執行相同的任務，由軟體控制和排程。

# 平行計算的分類 - Multi-core computing

## AMD EPYC™ 7742

**Specifications**

| | | |
|---|---|---|
| **# of CPU Cores:** 64 | **# of Threads:** 128 | **Base Clock:** 2.25GHz |
| **Max Boost Clock:** Up to 3.4GHz | **Total L3 Cache:** 256MB | **Package:** SP3 |
| **Socket Count:** 1P/2P | **PCI Express® Version:** PCIe 4.0 x128 | **Default TDP / TDP:** 225W |

**Performance**

| | | | |
|---|---|---|---|
| # of Cores | 28 | # of Threads | 56 |
| Processor Base Frequency | 2.90 GHz | Max Turbo Frequency | 4.30 GHz |
| Cache | 38.5 MB | # of UPI Links | 6 |
| TDP | 250 W | | |

圖片來源：https://www.amd.com/en/products/cpu/amd-epyc-7742,
https://www.intel.com/content/www/us/en/products/processors/xeon/scalable/platinum-processors/platinum-8380hl.html

# 平行計算的分類

- **多核心處理計算 (Multi-core computing)**：一個處理器內有多個處理單元 (processing units) 或稱核心 (core)
- **同時多執行緒 (Simultaneous multithreading)** 又稱超執行緒 (hyper-threading)。
- **對稱多處理 (Symmetric multiprocessing)**：電腦有多個處理器共用主記憶體。例如：2 way, 4 way, 8 way
- **分散式運算 (Distributed computing)**：位於不同地點的電腦透過網路相互連接傳遞訊息與通訊後，並協調它們的行為以達成共同目標而形成的系統。
- **大規模並行運算(Massively parallel computing)**：是由多個微處理器，局部存儲器及網路構成的節點 (node) 組成的並行計算系統；節點間以高速網路。大規模並行處理是一種異步的多指令及數據，因為它的程序有多個程序 (process)，它們分布在各個微處理器上，每個程序有自己獨立的地址空間，程序之間以訊息傳遞介面 (MPI) 進行相互溝通。
- **叢集運算 (Cluster computing)**：一組連接在一起工作的電腦。由於這些電腦協同工作，在許多方面它們可以被視為單個系統。與網格電腦不同，電腦叢集將每個節點設定為執行相同的任務，由軟體控制和排程。

# 平行計算的分類 - Symmetric multiprocessing

NEW

**SYS-7089P-TR4T** X11

7U 1 Node | 8 CPUs | 96 DIMMs | 16 Drives

4x 10G | SAS Optional | SATA Native | Redundant

Learn more | Add to compare

**SYS-2049U-TR4** X11

2U 1 Node | 4 CPUs | 48 DIMMs | 24 Drives

4x 1G | nvm EXPRESS Optional | SAS Optional | SATA Native

Learn more | Add to compare

**SYS-2049P-TN8R** X11

2U 1 Node | 4 CPUs | 48 DIMMs | 10 Drives

1x 1G | nvm EXPRESS Native | SATA Native | Redundant

Learn more | Add to compare

圖片來源：https://www.supermicro.com/en/products/MP/

# 平行計算的分類

- **多核心處理計算 (Multi-core computing)**：一個處理器內有多個處理單元 (processing units) 或稱核心 (core)
- **同時多執行緒 (Simultaneous multithreading)** 又稱超執行緒 (hyper-threading)。
- **對稱多處理 (Symmetric multiprocessing)**：電腦有多個處理器共用主記憶體。例如：2 way, 4 way, 8 way
- **分散式運算 (Distributed computing)**：位於不同地點的電腦透過網路相互連接傳遞訊息與通訊後，並協調它們的行為以達成共同目標而形成的系統。
- **大規模並行運算(Massively parallel computing)**：是由多個微處理器，局部存儲器及網路構成的節點 (node) 組成的並行計算系統；節點間以高速網路。大規模並行處理是一種異步的多指令及數據，因為它的程序有多個程序 (process)，它們分布在各個微處理器上，每個程序有自己獨立的地址空間，程序之間以訊息傳遞介面 (MPI) 進行相互溝通。
- **叢集運算 (Cluster computing)**：一組連接在一起工作的電腦。由於這些電腦協同工作，在許多方面它們可以被視為單個系統。與網格電腦不同，電腦叢集將每個節點設定為執行相同的任務，由軟體控制和排程。

# 平行計算的分類 - Distributed computing

- 柏克萊開放式網路計算平台（英語：Berkeley Open Infrastructure for Network Computing，簡稱BOINC）是目前主流的分佈式計算平台之一，由加州大學柏克萊分校電腦學系發展出的分散式計算系統。
- 原本專為SETI@home專案而設計，目前納入的領域包括數學、醫學、天文學和氣象學等。BOINC匯集全球各地誌願者的電腦或行動裝置，提供運算能力給研究者。截至2024年7月，BOINC在全世界有約106,880台活躍的主機，提供約20.592 PFLOPS的運算能力。

圖片來源：https://boinc.berkeley.edu/, https://setiathome.berkeley.edu/

# 平行計算的分類

- **多核心處理計算 (Multi-core computing)**：一個處理器內有多個處理單元 (processing units) 或稱核心 (core)
- **同時多執行緒 (Simultaneous multithreading)** 又稱超執行緒 (hyper-threading)。
- **對稱多處理 (Symmetric multiprocessing)**：電腦有多個處理器共用主記憶體。例如：2 way, 4 way, 8 way
- **分散式運算 (Distributed computing)**：位於不同地點的電腦透過網路相互連接傳遞訊息與通訊後，並協調它們的行為以達成共同目標而形成的系統。
- **大規模並行運算(Massively parallel computing)**：是由多個微處理器，局部存儲器及網路構成的節點 (node) 組成的並行計算系統；節點間以高速網路。大規模並行處理是一種異步的多指令及數據，因為它的程序有多個程序 (process)，它們分布在各個微處理器上，每個程序有自己獨立的地址空間，程序之間以訊息傳遞介面 (MPI) 進行相互溝通。
- **叢集運算 (Cluster computing)**：一組連接在一起工作的電腦。由於這些電腦協同工作，在許多方面它們可以被視為單個系統。與網格電腦不同，電腦叢集將每個節點設定為執行相同的任務，由軟體控制和排程。

# 平行計算的分類 - Massively parallel computing

圖片來源：http://en.people.cn/n3/2017/0619/c90000-9230404.html

SUNWAY TAIHULIGHT - SUNWAY MPP, SUNWAY SW26010 260C 1.45GHZ, SUNWAY

| | |
|---|---|
| **Site:** | National Supercomputing Center in Wuxi |
| **Manufacturer:** | NRCPC |
| **Cores:** | 10,649,600 |
| **Processor:** | Sunway SW26010 260C 1.45GHz |
| **Interconnect:** | Sunway |
| **Installation Year:** | 2016 |
| **Performance** | |
| **Linpack Performance (Rmax)** | 93.01 PFlop/s |
| **Theoretical Peak (Rpeak)** | 125.44 PFlop/s |
| **Nmax** | 12,288,000 |
| **HPCG [TFlop/s]** | 480.848 |

# 平行計算的分類

- **多核心處理計算 (Multi-core computing)**：一個處理器內有多個處理單元 (processing units) 或稱核心 (core)
- **同時多執行緒 (Simultaneous multithreading)** 又稱超執行緒 (hyper-threading)。
- **對稱多處理 (Symmetric multiprocessing)**：電腦有多個處理器共用主記憶體。例如：2 way, 4 way, 8 way
- **分散式運算 (Distributed computing)**：位於不同地點的電腦透過網路相互連接傳遞訊息與通訊後，並協調它們的行為以達成共同目標而形成的系統。
- **大規模並行運算(Massively parallel computing)**：是由多個微處理器，局部存儲器及網路構成的節點 (node) 組成的並行計算系統；節點間以高速網路。大規模並行處理是一種異步的多指令及數據，因為它的程序有多個程序 (process)，它們分布在各個微處理器上，每個程序有自己獨立的地址空間，程序之間以訊息傳遞介面 (MPI) 進行相互溝通。
- **叢集運算 (Cluster computing)**：一組連接在一起工作的電腦。由於這些電腦協同工作，在許多方面它們可以被視為單個系統。與網格電腦不同，電腦叢集將每個節點設定為執行相同的任務，由軟體控制和排程。

# 平行計算的分類 - Cluster computing

圖片來源：https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/ ,

## SUMMIT - IBM POWER SYSTEM AC922, IBM POWER9 22C 3.07GHZ, NVIDIA VOLTA GV100, DUAL-RAIL MELLANOX EDR INFINIBAND

| | |
|---|---|
| **Site:** | DOE/SC/Oak Ridge National Laboratory |
| **System URL:** | http://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/ |
| **Manufacturer:** | IBM |
| **Cores:** | 2,414,592 |
| **Processor:** | IBM POWER9 22C 3.07GHz |
| **Interconnect:** | Dual-rail Mellanox EDR Infiniband |
| **Installation Year:** | 2018 |
| **Performance** | |
| **Linpack Performance (Rmax)** | 148.60 PFlop/s |
| **Theoretical Peak (Rpeak)** | 200.79 PFlop/s |
| **Nmax** | 16,473,600 |
| **HPCG [TFlop/s]** | 2,925.75 |

# HPC 加速

- 資源有不同需求，分別可用下列方式加速：
  - **Supercomputer(Computer Cluster，Massively Parallel Processor)**
  - GPU (Graphics Processing Unit) : OpenCL, CUDA, OpenACC
  - Cloud Services :
    - Cloud App : Google App Engine, AWS Lambda, Azure App Service
    - Cloud Container: Google Kubernetes Engine, Amazon EKS, Azure Kubernetes Service
    - Cloud Compute : Google Compute Engine, Amazon EC2, Azure Virtual Machine
- HPC 如何量測速度？以 HPL benchmark 程式，測速單位為 FLOPS (floating-point operations per second)

# HPC 加速 - Supercomputer

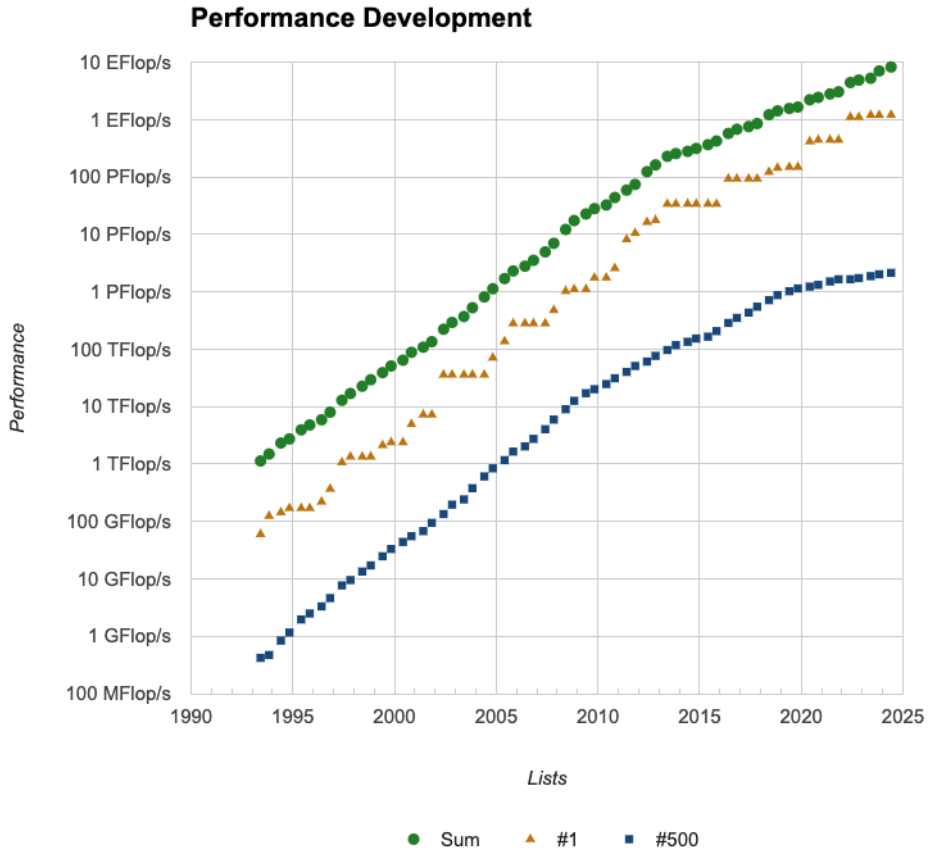
- 美國橡樹嶺國家實驗室 (Oak Ridge) 的 Frontier 在 TOP 500 目前排第一名。

  (https://www.top500.org/system/180047/) - AMD EPYC 64C
- 日本的理化學研究所 (RIKEN) 與富士通 (Fujitsu) 共同合作打造的 Fugaku (富岳) 在 TOP 500 目前排行第四，曾在 2020 年 6 月 至 2022 年 5 月期間，排行第一 。

  (https://www.top500.org/system/179807/) - ARM A64FX
- Top 500
  - 世界前 500 名 HPC 效能排名
  - https://www.top500.org/
- Green 500
  - 世界前500名HPC的節能排名
  - https://www.top500.org/green500/

圖片來源： https://www.olcf.ornl.gov/frontier/, https://www.fujitsu.com/global/about/innovation/fugaku/

# Top500的效能趨勢(JUNE 2024)



# Top5 (JUNE 2024)

| Rank | System | Cores | Rmax (PFlop/s) | Rpeak (PFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 8,699,904 | 1,206.00 | 1,714.81 | 22,786 |
| 2 | **Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel<br>DOE/SC/Argonne National Laboratory<br>United States | 9,264,128 | 1,012.00 | 1,980.01 | 38,698 |
| 3 | **Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure<br>Microsoft Azure<br>United States | 2,073,600 | 561.20 | 846.84 | |
| 4 | **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>Japan | 7,630,848 | 442.01 | 537.21 | 29,899 |
| 5 | **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>EuroHPC/CSC<br>Finland | 2,752,704 | 379.70 | 531.51 | 7,107 |
| 38 | **TAIPEI-1** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Nvidia<br>NVIDIA Corporation<br>Taiwan | 40,960 | 22.30 | 34.53 | |

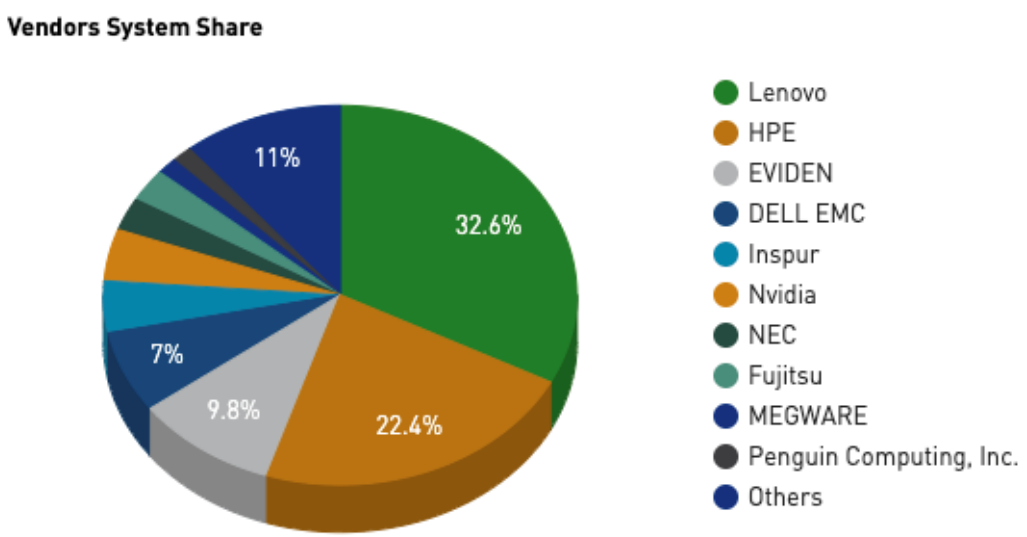
https://www.top500.org/lists/top500/2024/06/

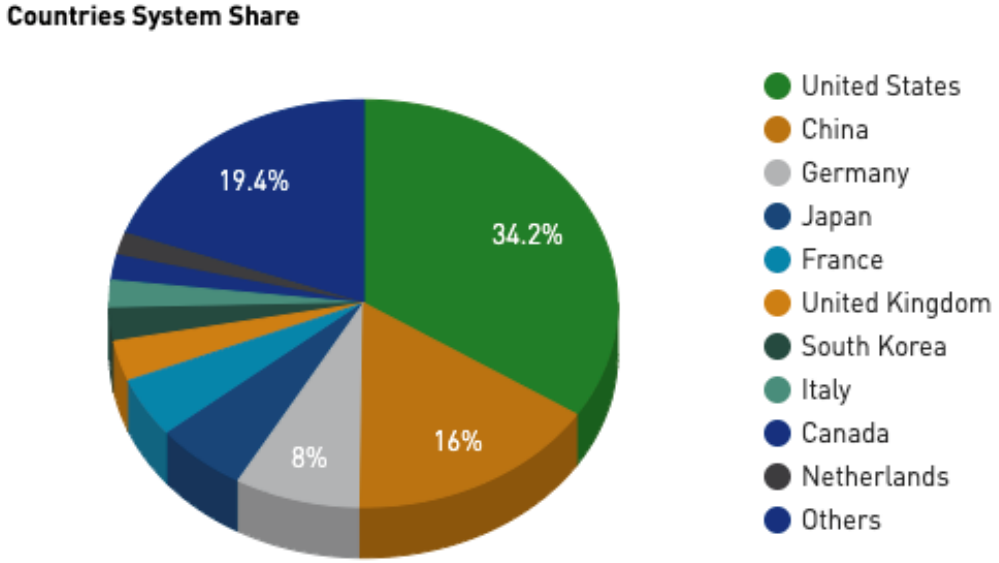
# Green 500 – Top5 (JUNE 2024)

| Rank | TOP500 Rank | System | Cores | Rmax (PFlop/s) | Power (kW) | Energy Efficiency (GFlops/watts) |
|---|---|---|---|---|---|---|
| 1 | 189 | **JEDI** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany | 19,584 | 4.50 | 67 | 72.733 |
| 2 | 128 | **Isambard-AI phase 1** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom | 34,272 | 7.42 | 117 | 68.835 |
| 3 | 55 | **Helios GPU** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cyfronet Poland | 89,760 | 19.14 | 317 | 66.948 |
| 4 | 328 | **Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States | 8,288 | 2.88 | 44 | 65.396 |
| 5 | 71 | **preAlps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland | 81,600 | 15.47 | 240 | 64.381 |

https://www.top500.org/lists/green500/2024/06/

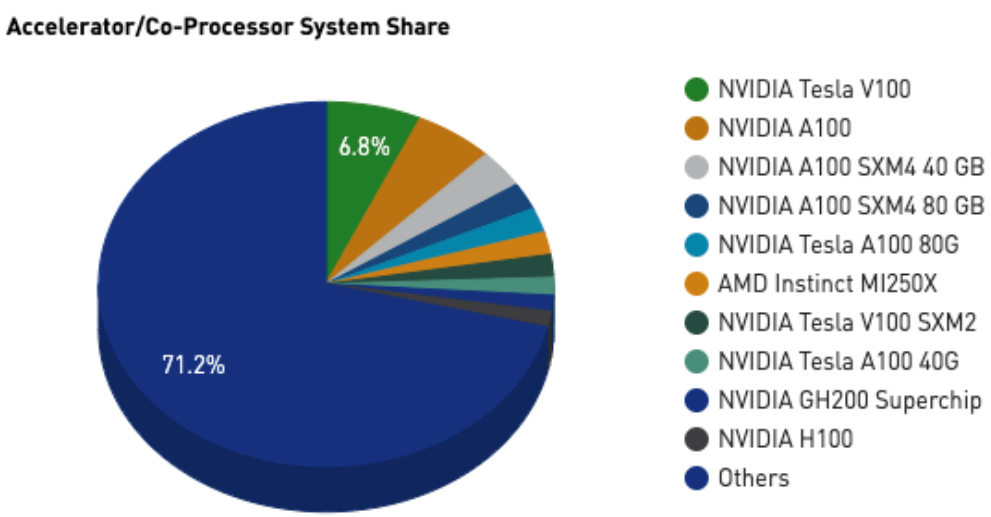
# Top500統計資料

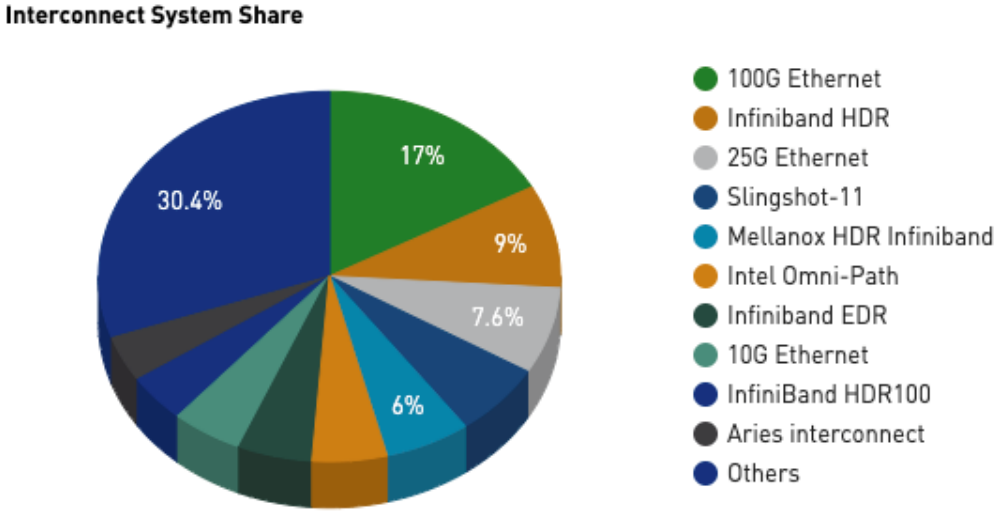



資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# Top500統計資料



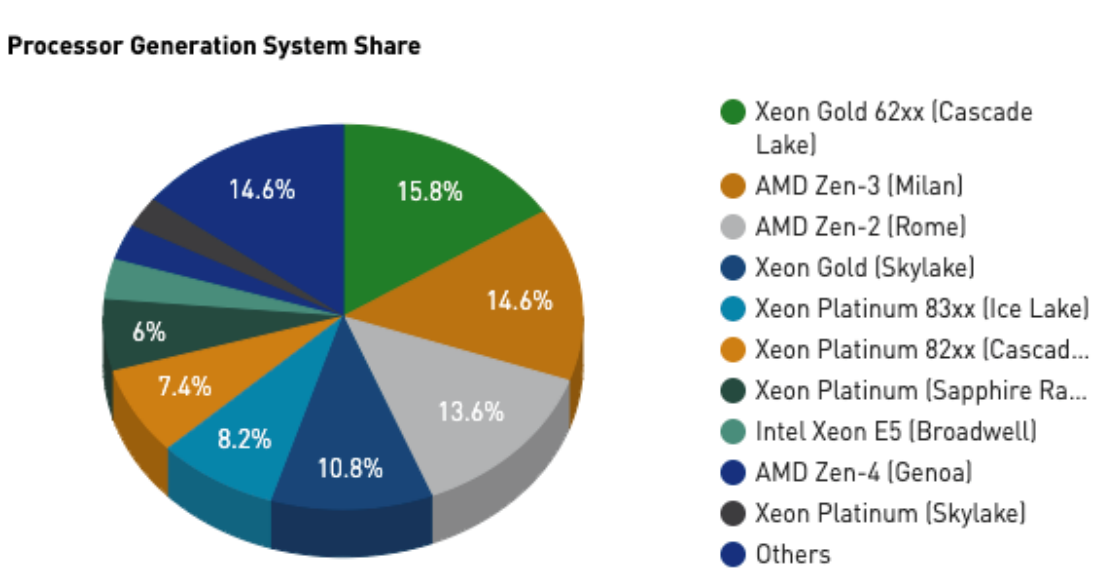



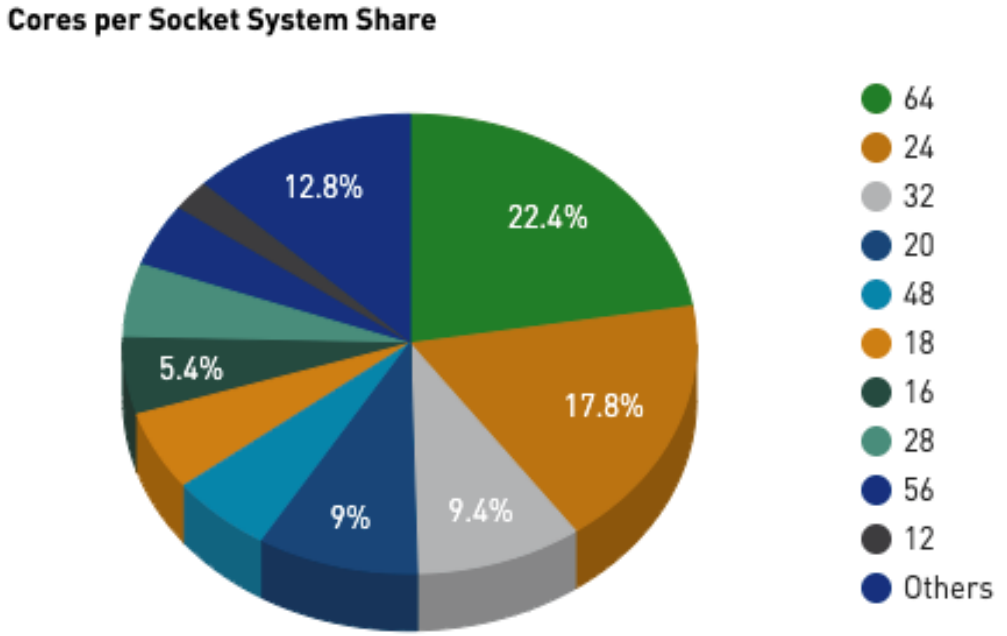

Top500中：
有加速器：38%
無加速器：62%

Top10中：
有加速器：60%
無加速器：40%

資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# Top500統計資料




2018年6月
Intel CPU佔總比例 95%
AMD CPU佔總比例 0

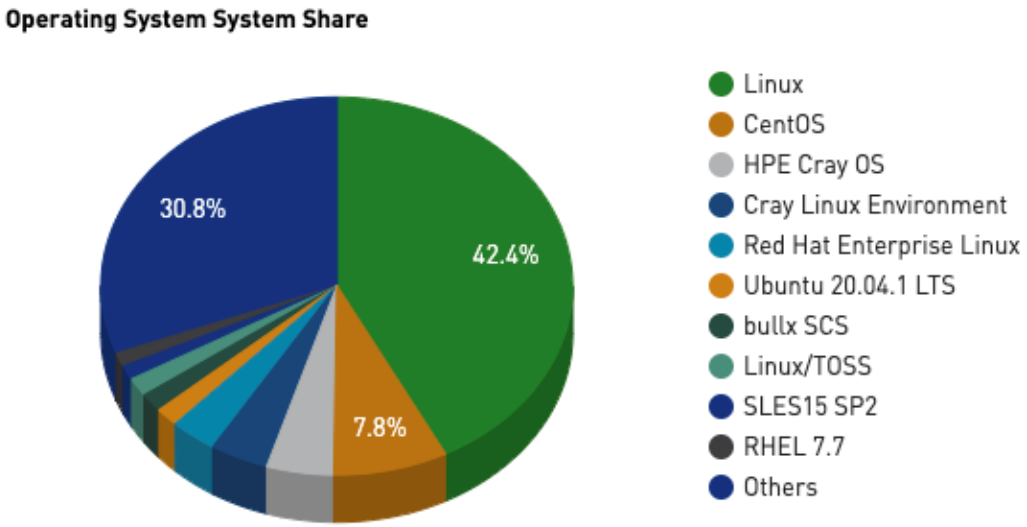
2022年6月
Intel CPU佔總比例 87.9%
AMD CPU佔總比例 18.6%

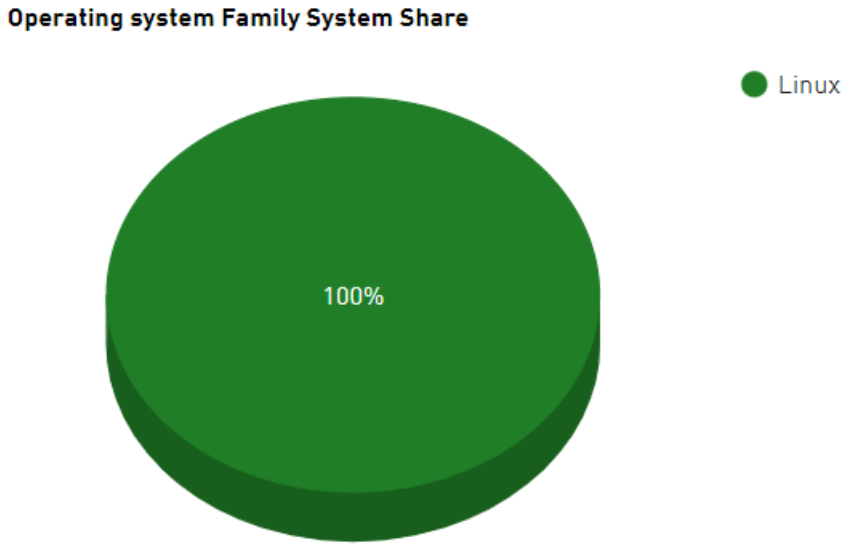
2023年6月
Intel CPU佔總比例 72%
AMD CPU佔總比例 24.2%

2024年6月
Intel CPU佔總比例 54.2%
AMD CPU佔總比例 31.2%

資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# Top500統計資料

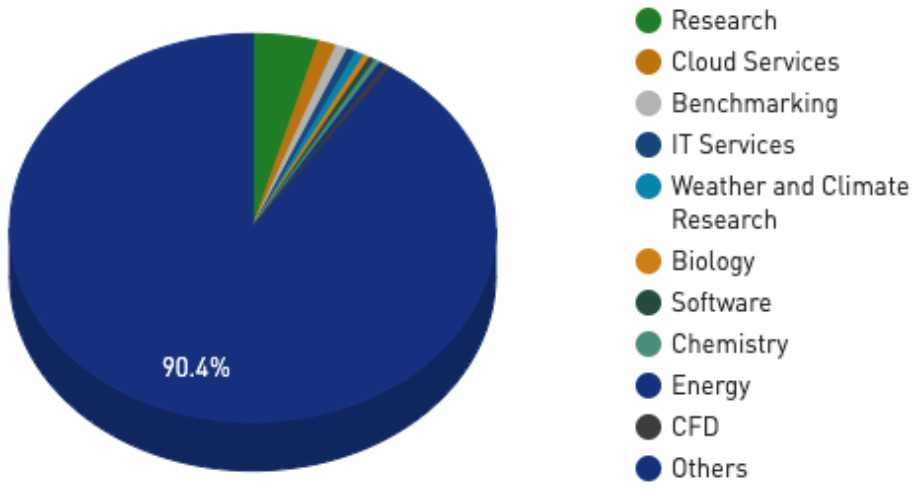



資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# Top500統計資料

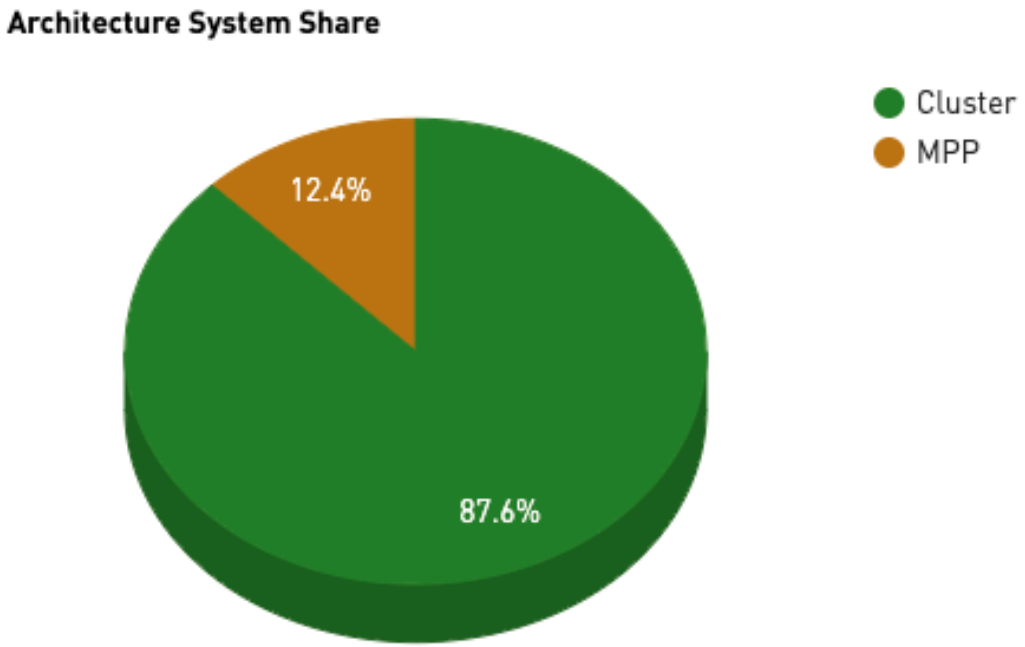
**Application Area System Share**




| | Application Area | Count | System Share (%) | Rmax (GFlops) | Rpeak (GFlops) | Cores |
|---|---|---|---|---|---|---|
| 1 | Research | 22 | 4.4 | 829,683,774 | 1,128,267,179 | 16,661,168 |
| 2 | Cloud Services | 6 | 1.2 | 681,460,000 | 1,021,717,542 | 3,093,408 |
| 3 | Benchmarking | 4 | 0.8 | 75,276,960 | 93,905,001 | 716,896 |
| 4 | IT Services | 3 | 0.6 | 50,360,000 | 70,080,460 | 459,296 |
| 5 | Weather and Climate Research | 3 | 0.6 | 23,497,260 | 38,398,266 | 420,032 |
| 6 | Biology | 2 | 0.4 | 13,591,000 | 24,475,970 | 116,352 |
| 7 | Software | 2 | 0.4 | 34,340,000 | 49,444,360 | 324,384 |
| 8 | Chemistry | 2 | 0.4 | 13,591,000 | 24,475,970 | 116,352 |
| 9 | Energy | 2 | 0.4 | 14,649,000 | 22,613,060 | 161,664 |
| 10 | CFD | 2 | 0.4 | 13,591,000 | 24,475,970 | 116,352 |
| 11 | Aerospace | 1 | 0.2 | 5,951,550 | 7,107,149 | 241,108 |
| 12 | Biochemistry | 1 | 0.2 | 4,504,000 | 5,131,780 | 19,584 |
| 13 | Economics | 1 | 0.2 | 9,087,000 | 19,344,190 | 96,768 |
| 14 | Finance | 1 | 0.2 | 9,087,000 | 19,344,190 | 96,768 |
| 15 | Medicine | 1 | 0.2 | 9,087,000 | 19,344,190 | 96,768 |
| 16 | Not Specified | 1 | 0.2 | 79,230,000 | 112,997,790 | 888,832 |
| 17 | Environment | 1 | 0.2 | 4,504,000 | 5,131,780 | 19,584 |
| 18 | Information Service | 1 | 0.2 | 63,460,000 | 79,215,000 | 555,520 |
| 19 | Geophysics | 1 | 0.2 | 9,087,000 | 19,344,190 | 96,768 |

資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# Top500統計資料




資料計算依2024年6月份Top500列表，圖片來源為Top500網站

# HPC 加速

- 資源有不同需求，分別可用下列方式加速：
  - Supercomputer(Computer Cluster，Massively Parallel Processor)
  - **GPU (Graphics Processing Unit) : OpenCL, CUDA, OpenACC**
  - Cloud Services :
    - Cloud App : Google App Engine, AWS Lambda, Azure App Service
    - Cloud Container: Google Kubernetes Engine, Amazon EKS, Azure Kubernetes Service
    - Cloud Compute : Google Compute Engine, Amazon EC2, Azure Virtual Machine
- HPC 如何量測速度？以 HPL benchmark 程式，測速單位為 FLOPS (floating-point operations per second)

# HPC 加速 - GPU

- 目前單顯卡 CUDA 核心數高達18432核，記憶體可高達 188 GB 頻寬可達 7.8 TB/sec，算力FP32 134 TFLOPs FP64 68 TFLOPs
- 用 OpenCL 、 CUDA 、 OpenACC 進行加速

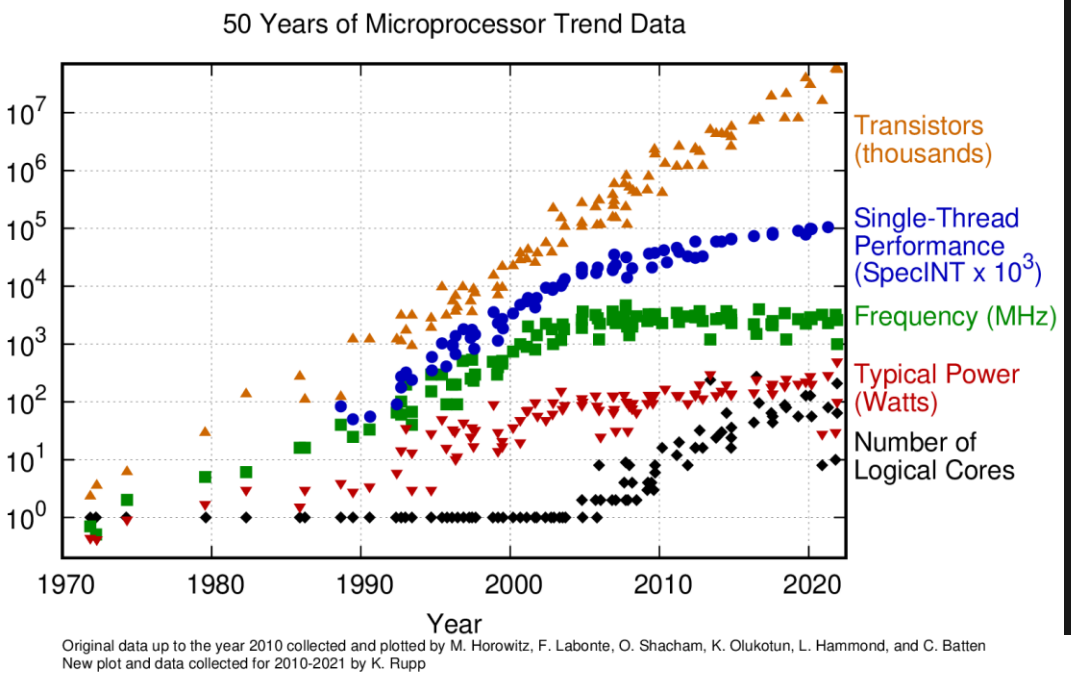
OpenCL™

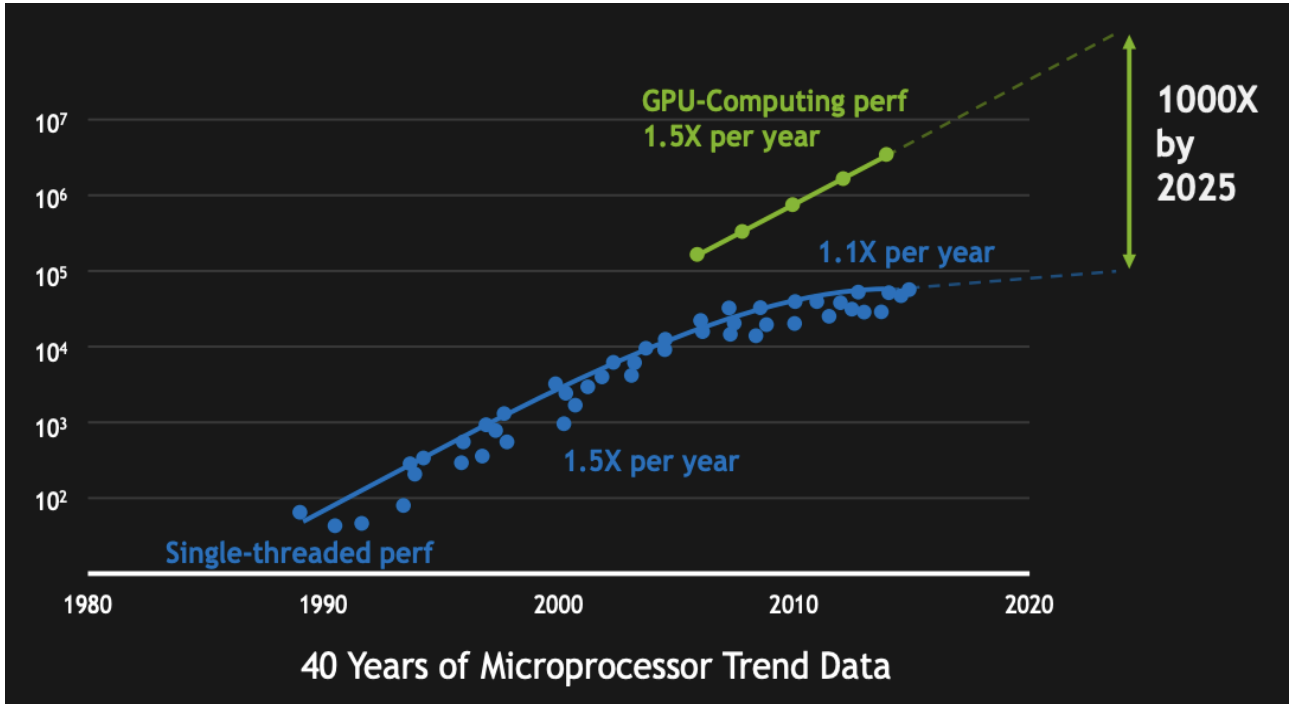
NVIDIA CUDA

OpenACC
More Science, Less Programming

圖片來源：https://opencl.org/，https://en.wikipedia.org/wiki/CUDA，https://www.openacc.org/，https://www.nvidia.com/en-us/data-center/dgx-systems/，https://www.amd.com/en/graphics/servers-radeon-instinct-mi-powered-servers

# HPC 加速 - GPU




圖片來源：https://github.com/karlrupp/microprocessor-trend-data . JENSEN HUANG, FOUNDER & CEO | GTC CHINA 2017

# HPC 加速

- 資源有不同需求，分別可用下列方式加速：
  - Supercomputer(Computer Cluster，Massively Parallel Processor)
  - GPU (Graphics Processing Unit) : OpenCL, CUDA, OpenACC
  - **Cloud Services :**
    - **Cloud App : Google App Engine, AWS Lambda, Azure App Service**
    - **Cloud Container: Google Kubernetes Engine, Amazon EKS, Azure Kubernetes Service**
    - **Cloud Compute : Google Compute Engine, Amazon EC2, Azure Virtual Machine**
- HPC 如何量測速度？以 HPL benchmark 程式，測速單位為 FLOPS (floating-point operations per second)

# HPC 加速 - Cloud App 服務

- 不必擔心基礎架構的管理工作。開發人員完全不需要管理伺服器及部署設定，因此可以專心建構出色的應用程式，沒有任何管理作業負擔。

Google App Engine

AWS Lambda

Azure App Service

# HPC 加速 - Cloud Container 服務

- 提供預建部署範本的容器化解決方案，具有可攜性、簡化授權程序及合併帳單等特色。這類應用程式不僅是容器映像檔，也是提高開發人員生產力而建構的開放原始碼商業應用程式。

Google Kubernetes Engine (GKE)

AWS EKS

Azure Kubernetes Service (AKS)

# HPC 加速 - Cloud Compute 服務

- 虛擬機器可在主機系統間即時遷移，而且不用重新啟動，即便主機系統在進行必要維護時，應用程式也能照常運作。

Google Compute Engine (GCE)

Amazon Elastic Compute Cloud (EC2)

Azure Virtual Machine

# HPC 的應用

# HPC 的應用

特色

- 巨量資料分析需求
- 高效能計算需求
- 需要花很長時間處理的需求
- 大量平行分散計算需求
- 大量批次處理需求（不要人工介入）

# HPC 的應用

**應用層面太廣泛，不勝枚舉...**

# HPC 的應用-天氣預報

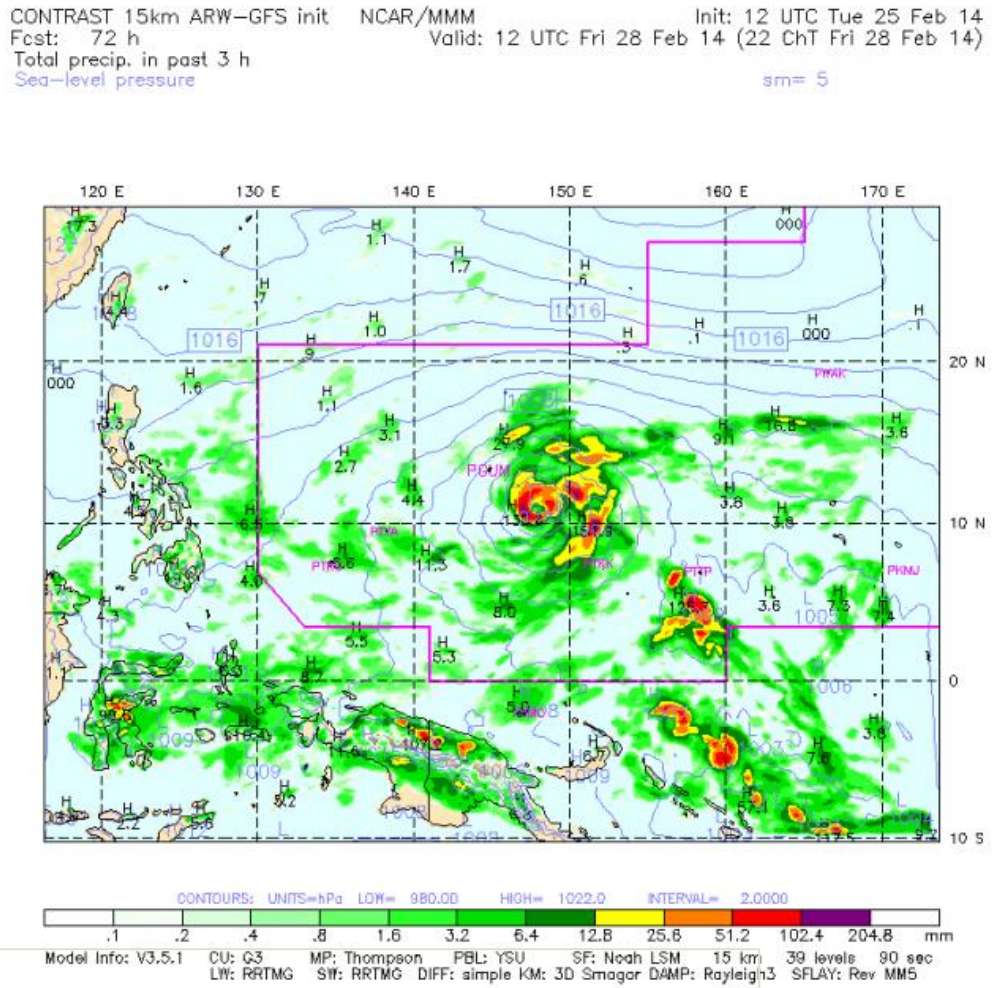
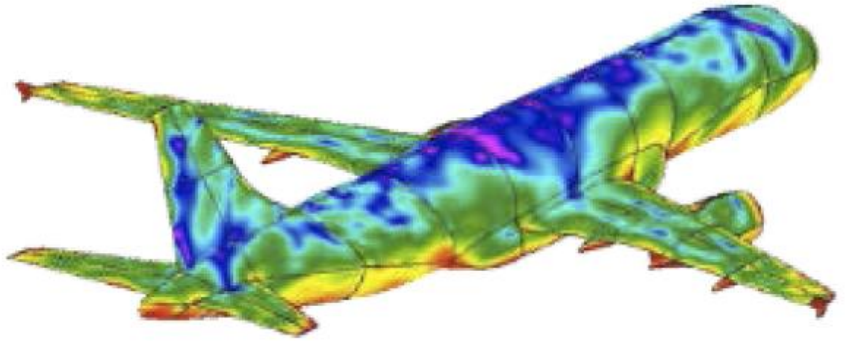



**Fig**: WRF domain (15-km) configured for the forecasting support of CONTRAST. 72-hr forecast of SLP (contour interval= 2 mb) and 3-hourly precipitation (mm, shaded, scale at bottom) shown (1200 UTC 25 Feb 2014 initialization). Purple boundary marks the Oakland flight information region (FIR).

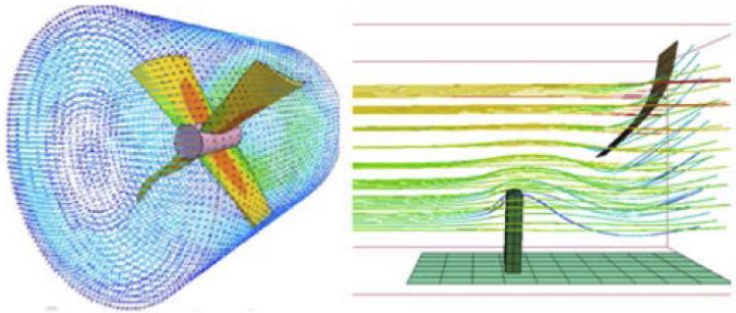
Source:
https://www.mmm.ucar.edu/forecast_support

# HPC 的應用-電腦輔助工程(CAE)

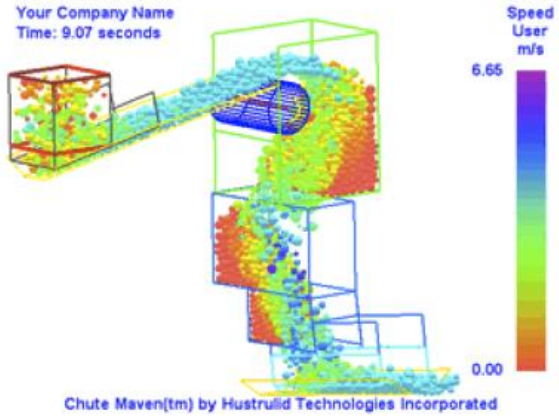
圖片來源:
http://www.hpcadvisorycouncil.com/pdf/LS-DYNA%20_analysis.pdf



圖片來源:
http://cihanfea.blogspot.tw/

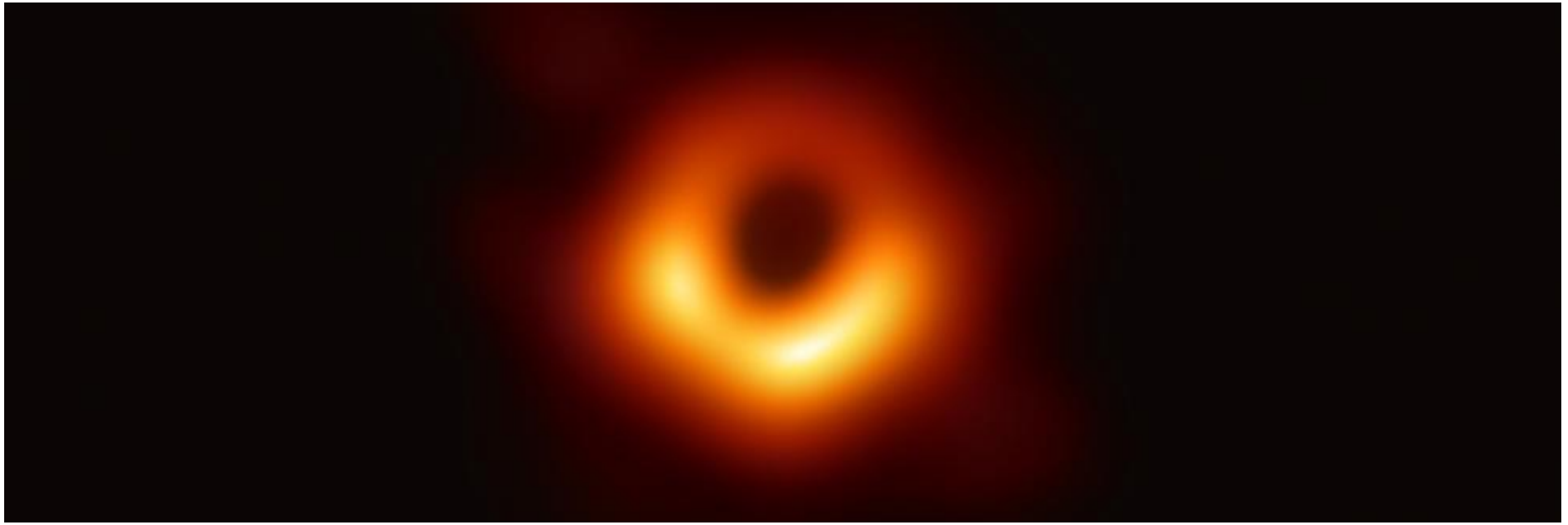


圖片來源:
http://www.lstc.com/applications/new_multiphysics




圖片來源:
https://www.chutemaven.com/images/movie8.gif

# HPC 的應用 - 天文模擬

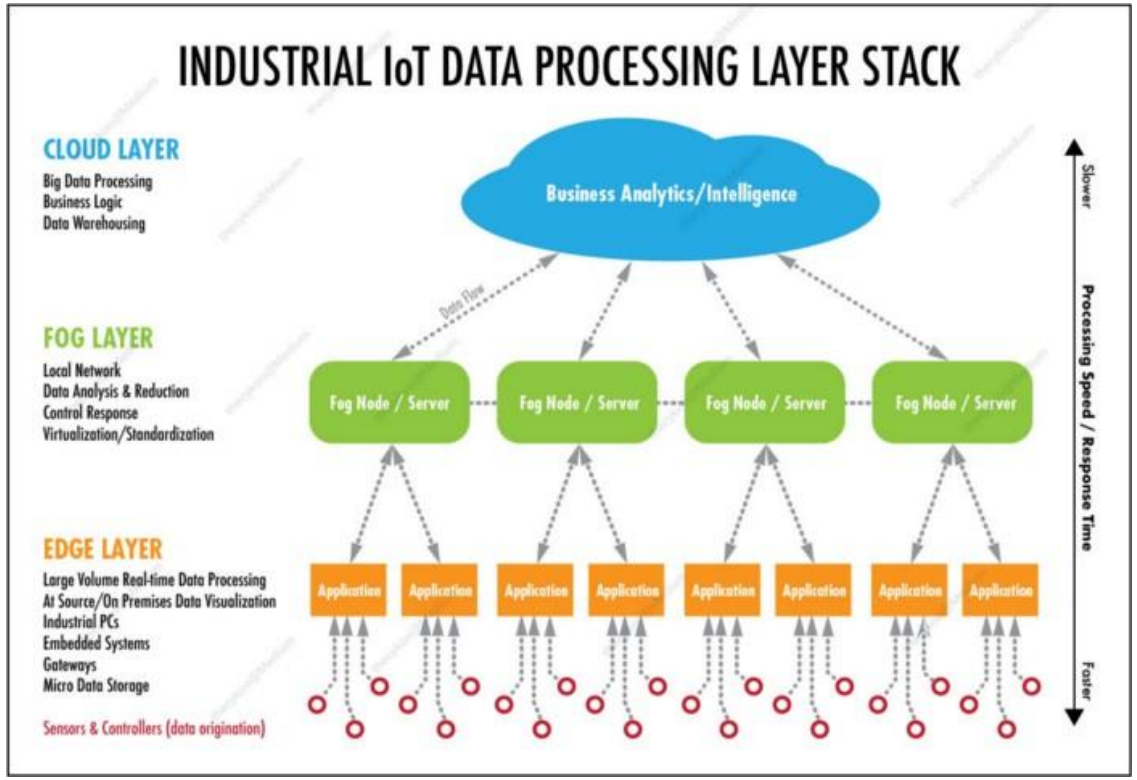


**2019/4/10**

圖片來源：https://eventhorizontelescope.org/

# HPC 的應用-邊緣運算(Edge Computing)

- *邊緣計算是一種分散式運算的架構，將應用程式、數據資料與服務的運算，由網絡中心節點，移往網絡邏輯上的邊緣節點來處理*



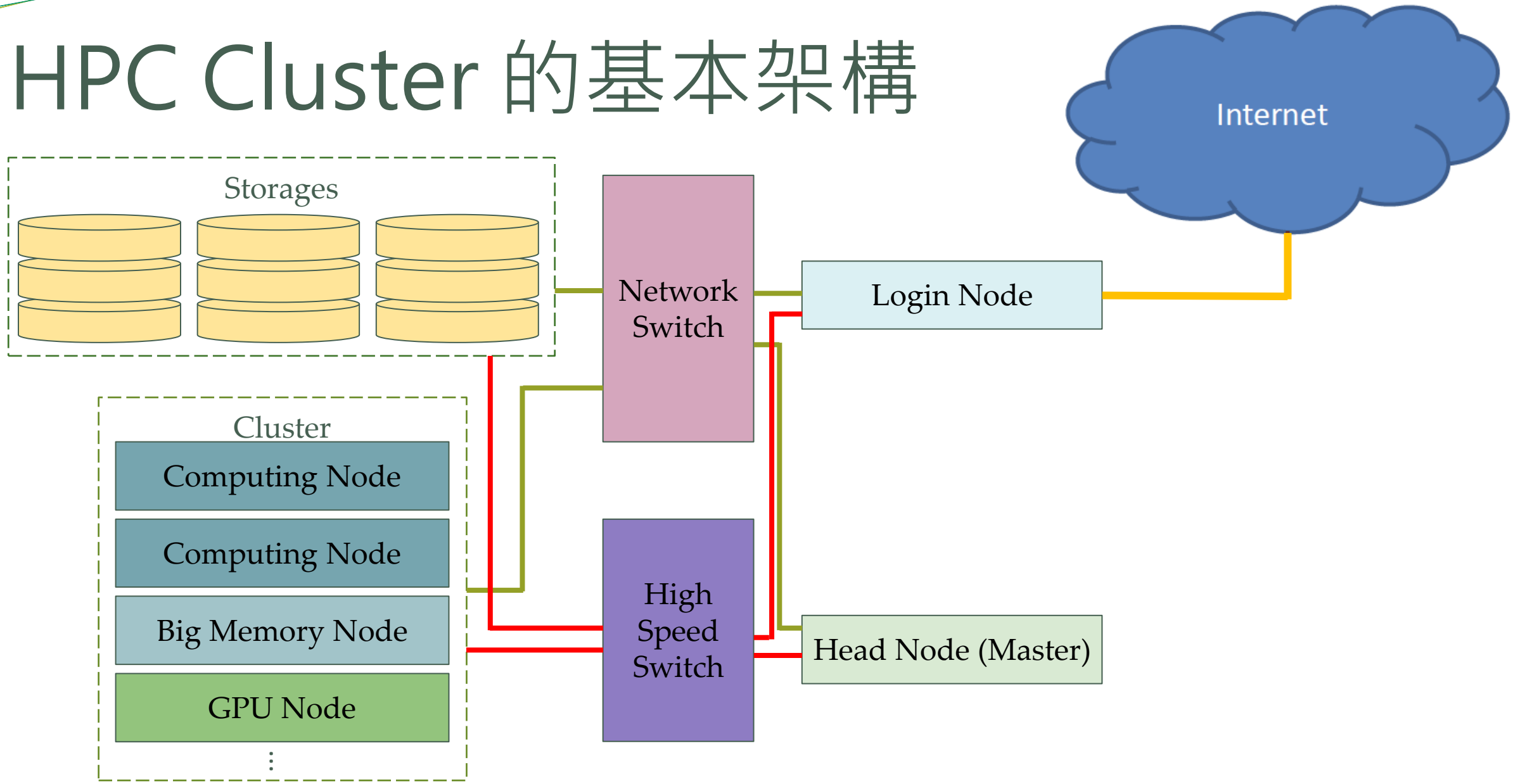

邊緣計算優點：

- 低延遲: 數據由近場產生，能快速回應
- 獨立性: 在沒有網絡連接下，系統亦能運作
- 合規性: 無需傳送用戶資料，保護個人數據
- 簡化數據: 終端先處理部份數據，數據簡化後才向雲服務器傳輸
- 安全性: 數據傳輸減少，減少網絡安全風險

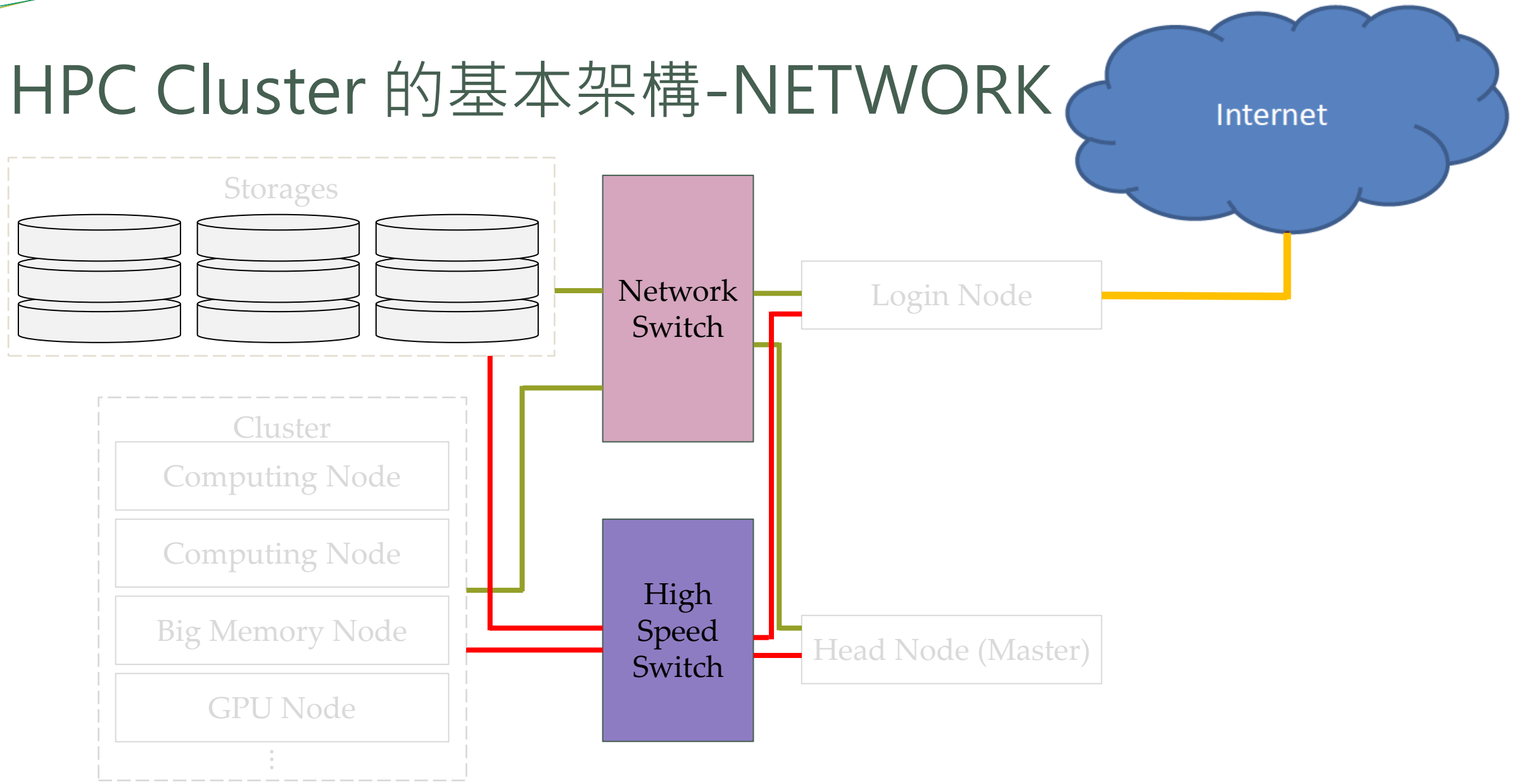
資料來源: IT+Digital@互聯網

# HPC Cluster 的基本架構

# HPC Cluster 的基本架構



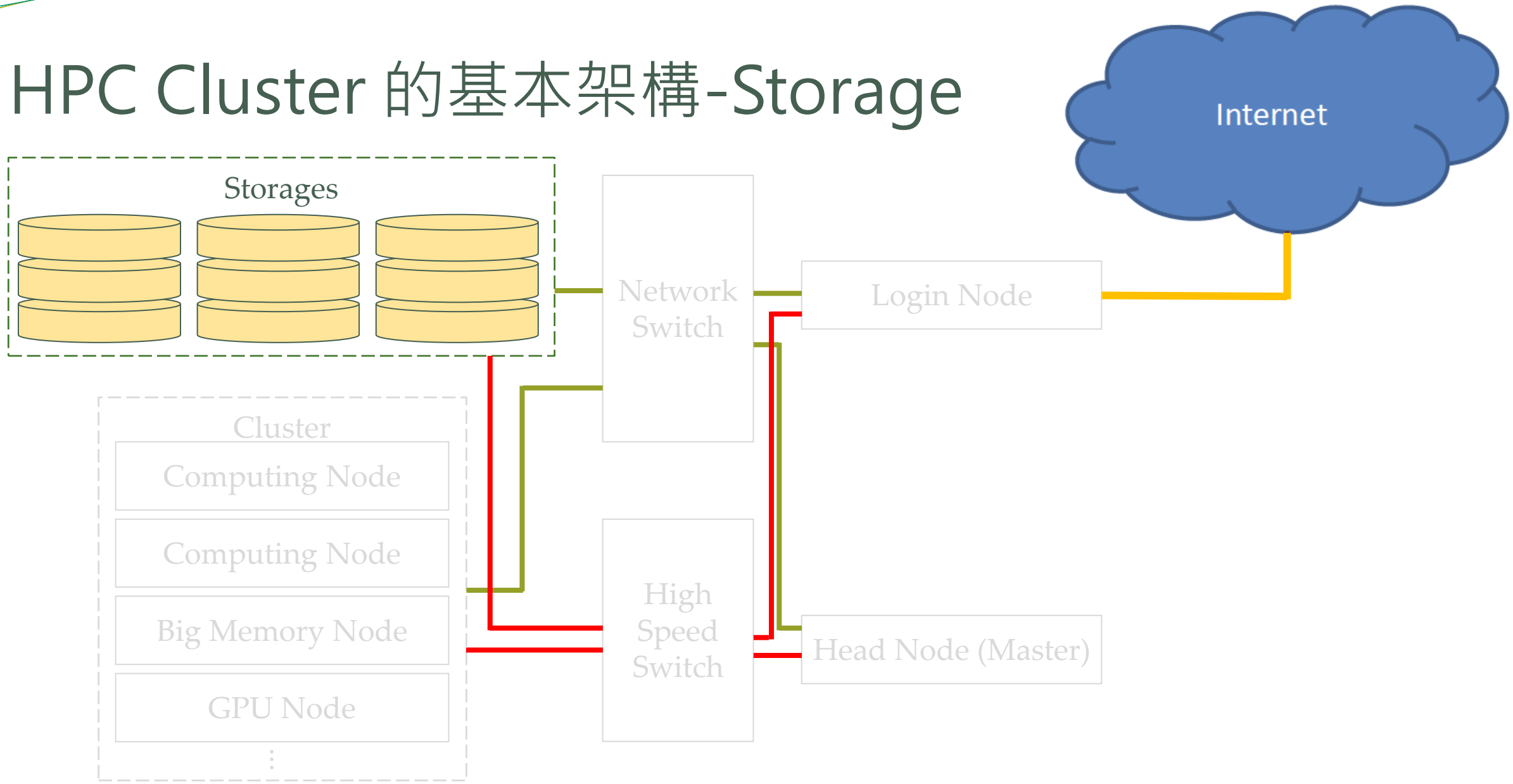
# HPC Cluster 的基本架構-NETWORK



# HPC Cluster的基本架構-Network

- Infiniband
  - FDR 56Gbs (Mellanox ConnectX-3, ConnectX-4)
  - EDR 100Gbs (Mellanox ConnectX-4, ConnectX-5)
  - HDR 200Gbs (Mellanox ConnectX-6)
- Fiber
  - 8Gbs / 16Gbs / 32Gbs
- Ethernet
  - Mellanox 10/25/40/50/56/100/200GbE
- Intel Omni-Path
  - 100Gbs

# HPC Cluster 的基本架構-Storage



# HPC Cluster 的基本架構-Storage

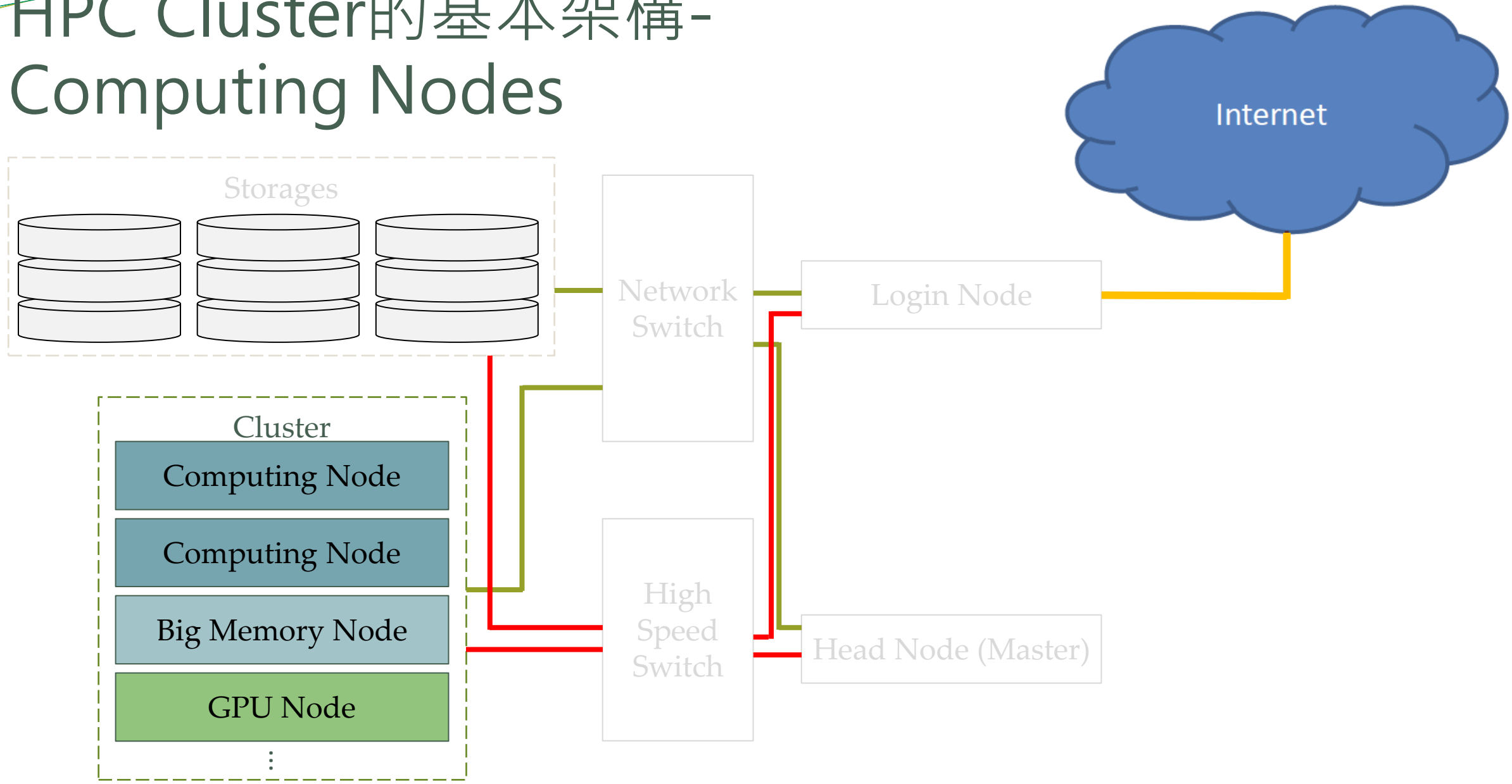
Lustre

- Open Source
  - Lustre
  - Glusterfs
  - BeeGFS
  - OrangeFS
  - Ceph
  - NFS

IBM Spectrum Scale

- Enterprise
  - IBM GPFS
  - DELL Isilon OneFS

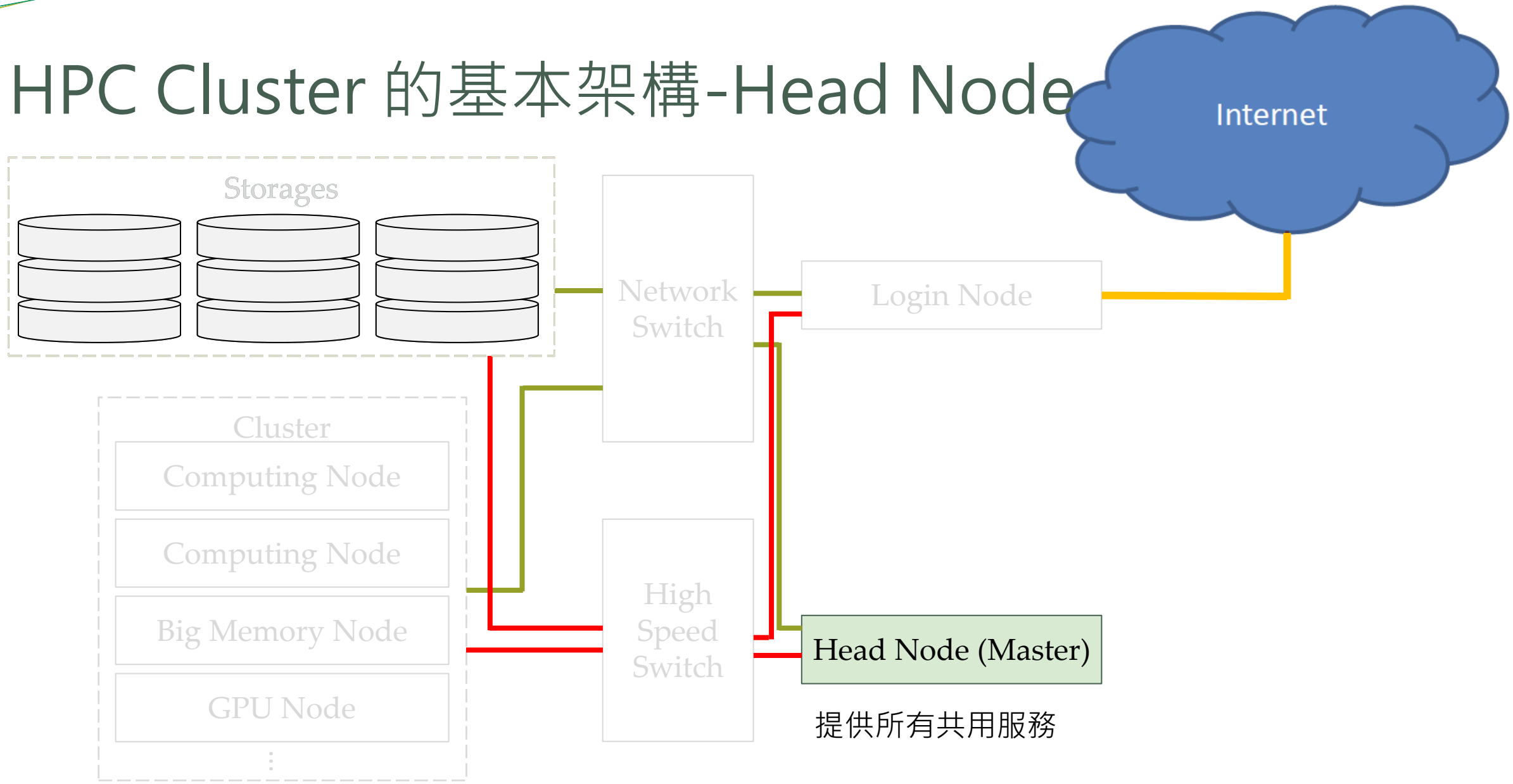
# HPC Cluster的基本架構-Computing Nodes



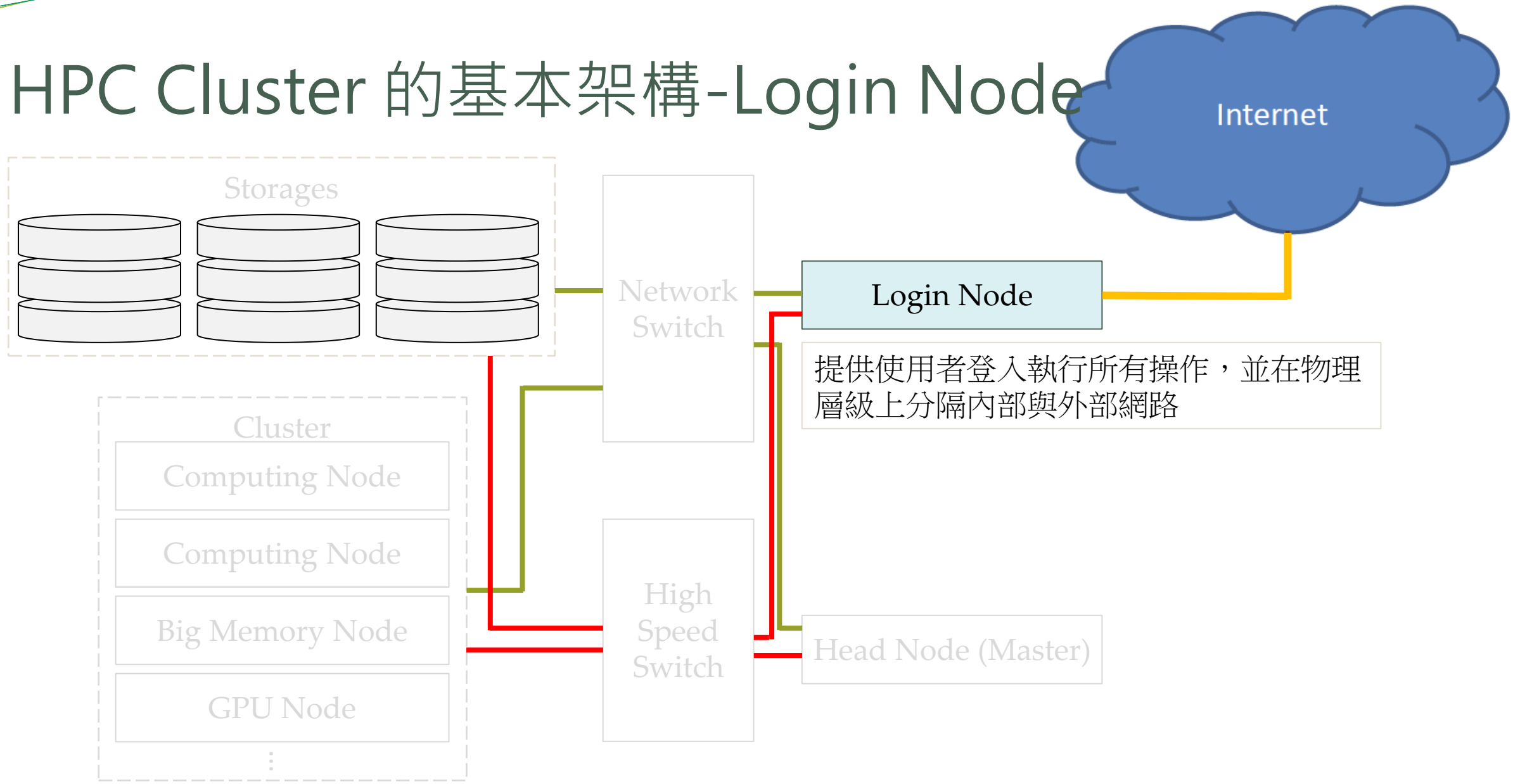
# HPC Cluster的基本架構-Computing Nodes

| 主要組件 | 規格 |
| --- | --- |
| CPU | Intel Xeon / AMD / IBM Power 9 / ARM |
| 記憶體 | 8GB/16GB/32GB/64GB/128GB ECC |
| 硬碟 | SAS HDD / SAS SSD / SATA SSD / NVMe |
| 網路卡 | InfiniBand / Fiber / 超高速Ethernet |
| 管理模組 | IPMI / iLO (HPE) / iDrac (DELL) / IMM (IBM, Lenovo) |

# HPC Cluster 的基本架構-Head Node



# HPC Cluster 的基本架構-Login Node



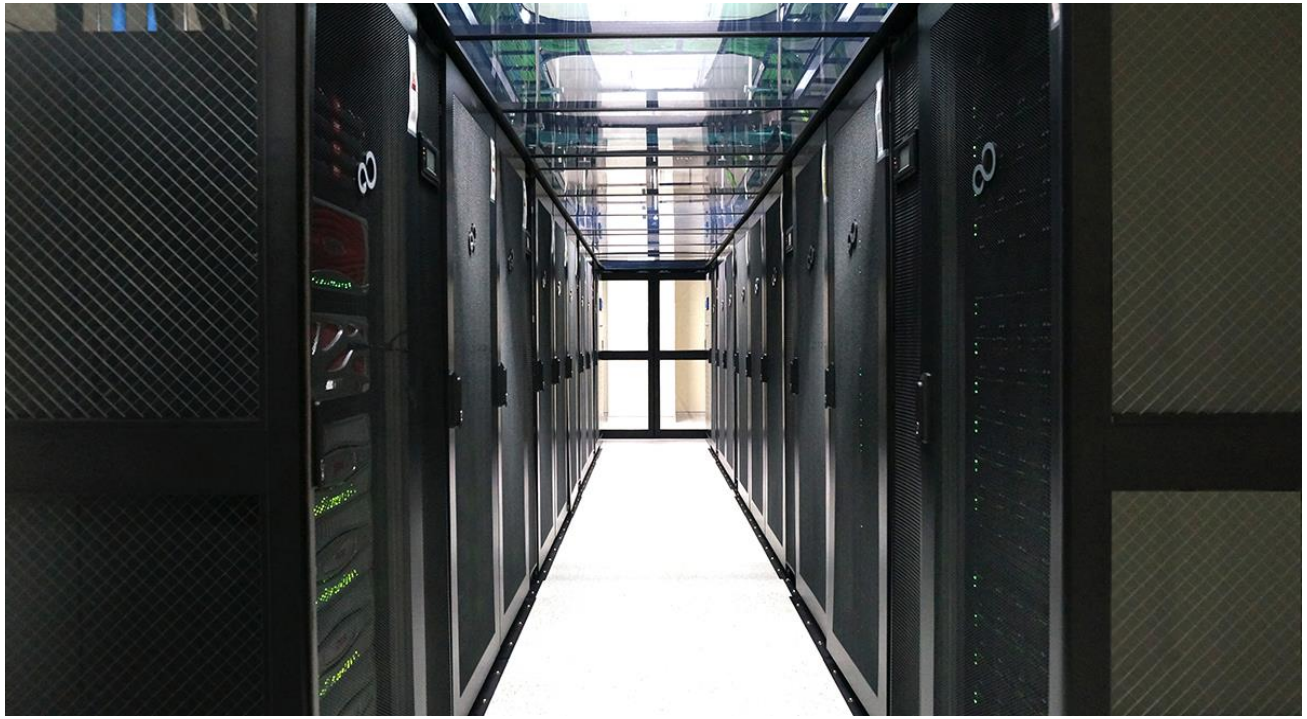
# HPC Cluster的基本架構-服務

| 功能 | 軟體元件 |
| --- | --- |
| 網路磁碟 | NFS / DFS / CIFS |
| 帳號 | NIS / LDAP |
| 工作排程 | Torque / OpenPBS / LSF |
| 校時 | NTP |

# HPC 實例

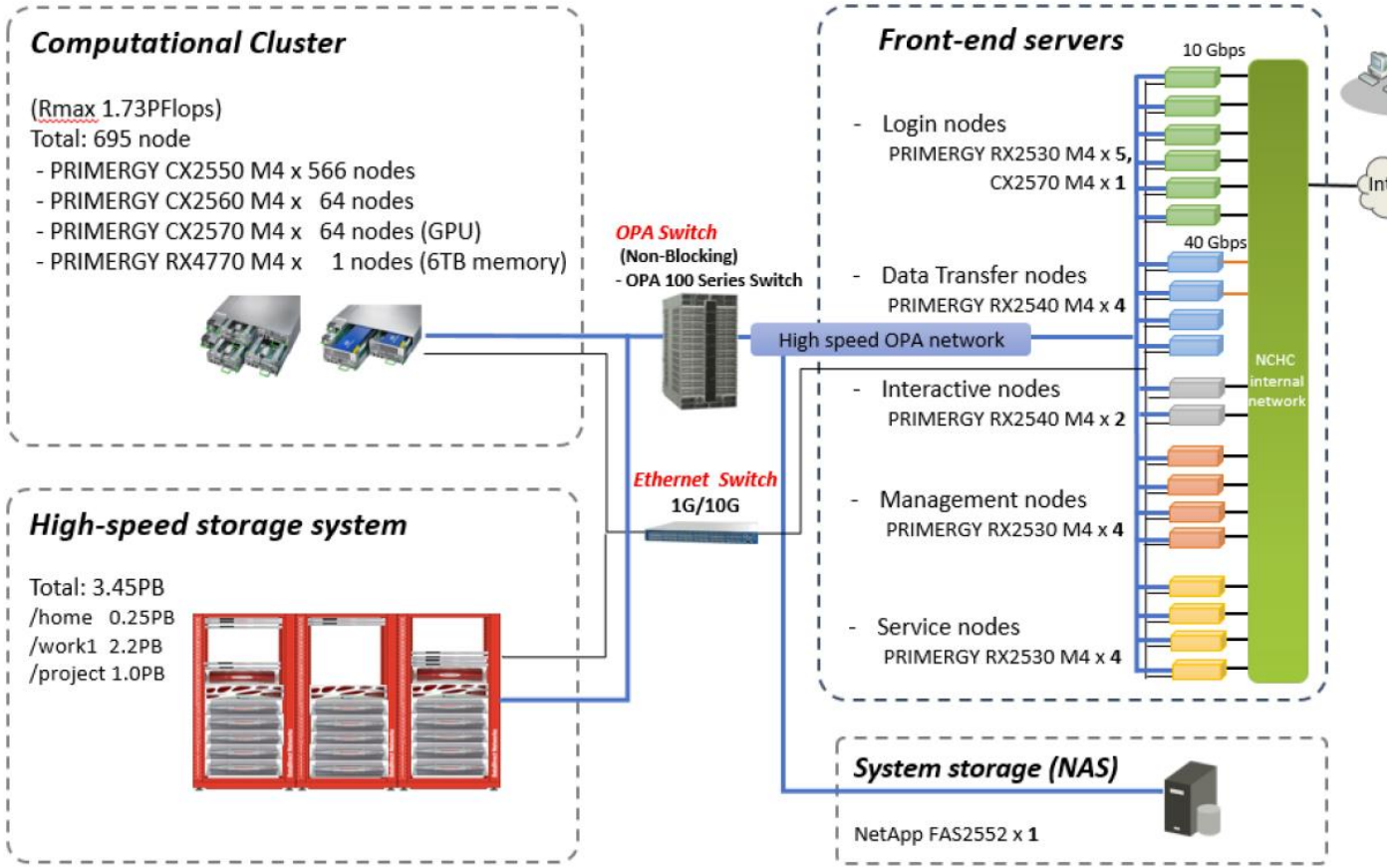
# HPC 實例 - 台灣杉一號 TAIWANIA 1

- 計算節點：Intel Xeon Gold 6148 2.4GHz CPU x 2 & 192 / 384 GB RAM
- 網路介面：Intel Omni-Path HFI
- 作業系統：RHEL 7.3 x86_64
- 編譯器：Intel Parallel Studio XE 2018、PGI Compilers & Tools 2017
- 排程軟體：Altair PBS Professional



資料來源：https://www.nchc.org.tw/Page?itemid=2&mid=4

# HPC 實例 - 台灣杉一號 TAIWANIA 1

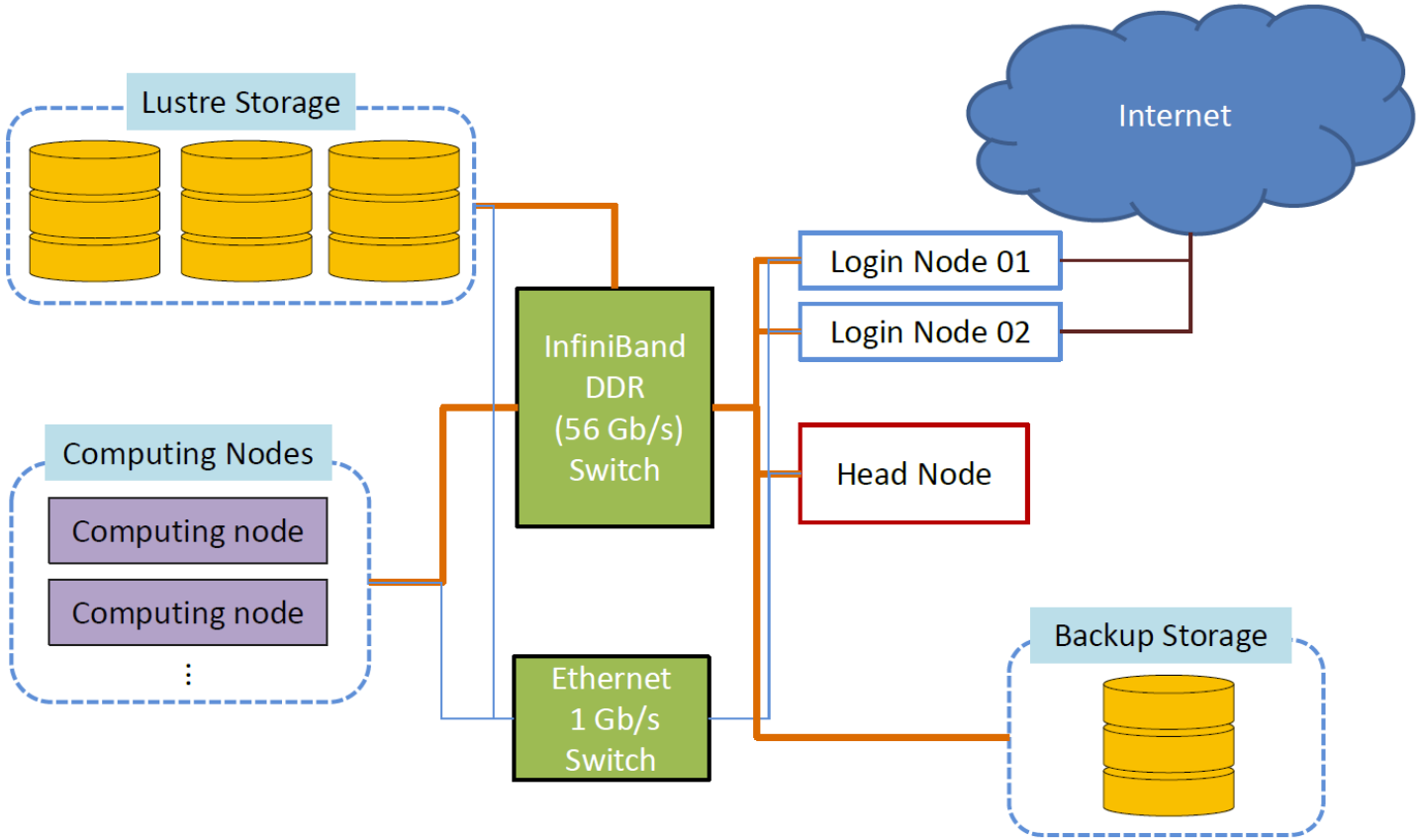



資料來源：台灣杉一號操作手冊

# HPC 實例 - 臺師大理學院HPC

- 計算節點：52 computing nodes
  - 44 CPU nodes (1056 CPU cores)
  - 2 GPU nodes(48 CPU cores, 4 GPU)
  - 6 Big Memory nodes (192 CPU cores)
- 硬碟空間：100TB Lustre filesystem
- 網路連線：Infiniband FDR (56 Gb/s)
- 作業系統：Linux (CentOS 7 x86_64)
- 排程軟體：Torque
- 編譯程式：GCC, Intel, PGI, Python 2.x/3.x

# HPC 實例 - 臺師大理學院HPC



# HPC 趨勢、主流與規劃

# HPC 趨勢與主流

- 高密度主機 : 1U 512 cores CPU
- GPU 加速 : many-core 8192 cores GPU
- 使用容器 : Container provides different environments
- 資料處理單元 : Offloading network functions from the CPU to the network
- 浸沒式冷卻系統 : Energy saving around 30% ~ 40%
- 雲端服務 : Pay-as-you-go

# HPC 趨勢與主流 - 高密度主機

- 2U 可安裝 4 節點，每個節點可安裝 2 CPUs，目前技術可高達 512 核心 (cores)，單一節點可支援 2TB 記憶體
- Cray 1U 可安裝 8 CPUs，最高可達 512 核心

**GIGABYTE H262-ZL0**
**AMD DP 2U 4 Nodes Server Gen4 NVMe Liquid Cooling Solution**

資料來源：GIGABYTE高密度伺服器

# HPC 趨勢與主流 - GPU 加速

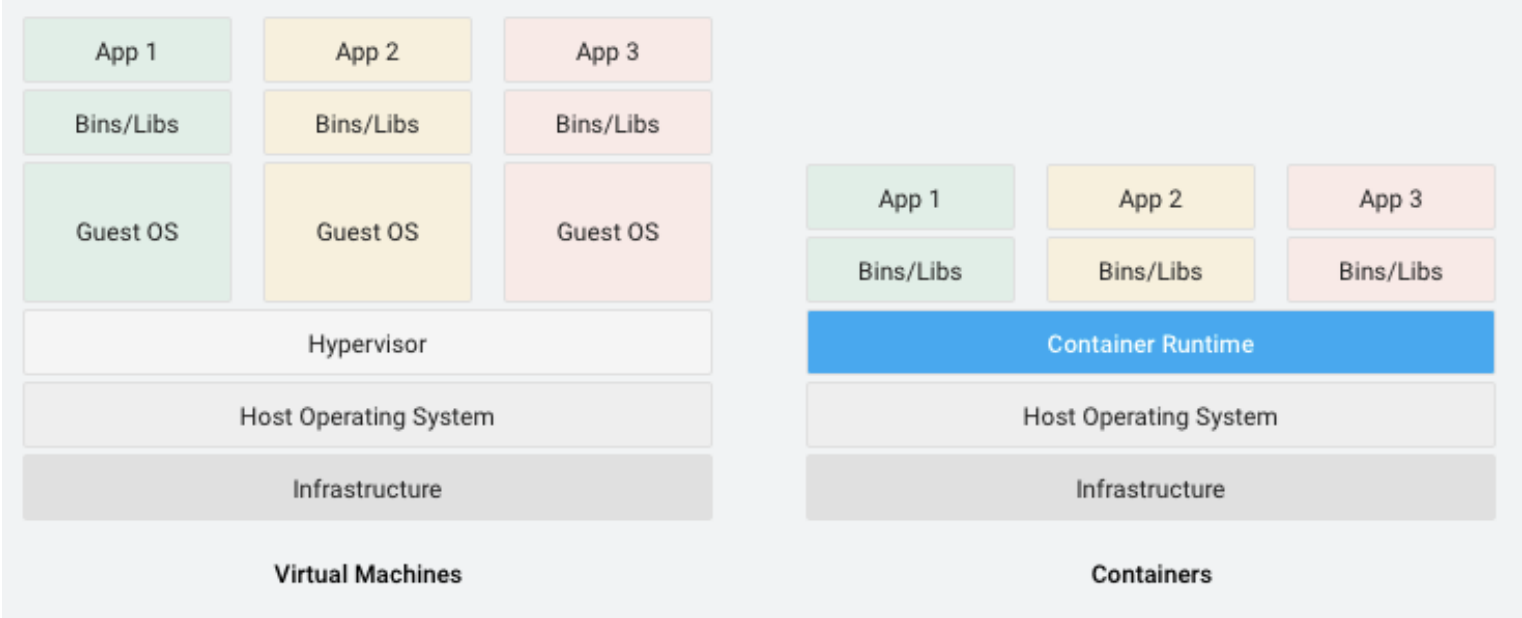
- 程式內如果有大量的 Loop 且運算邊界交換次數很少就很適合用 GPU 加速
- 透過 NVIDIA HPC SDK與 OpenACC 程式開發門檻降低
- 在 NVIDIA 提供 Unified Memory 技術，解決 GPU 記憶體太小的詬病

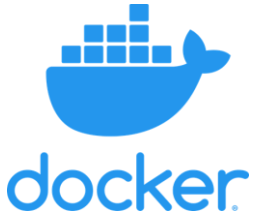
| | HGX H100 | | | |
|---|---|---|---|---|
| | H100 PCIe | 4-GPU | 8-GPU | 256-GPU |
| GPUs | 1 個 NVIDIA H100 PCIe | HGX H100 4-GPU | HGX H100 8-GPU | 32 個 HGX H100 8-GPU |
| 尺寸規格 | PCIe | 4 個 NVIDIA H100 SXM | 8 個 NVIDIA H100 SXM | 32 個各自擁有 8 個採用 NVLink Switch 系統連結的 NVIDIA H100 SXM 節點 |
| 高效能運算與人工智慧運算 (FP64/TF32/FP16/FP8/INT8) | 48TF/800TF/1.6PF/3.2PF/3.2 POPS | 240TF/4PF/8PF/16PF/16POPS | 480TF/8PF/16PF/32PF/32POPS | 15PF/256PF/512PF/1EF/1EOPS |
| 記憶體 | 每個 GPU 各 80 GB | 最高 320 GB | 最高 640 GB | 最高 20 TB |
| NVLink | 第四代 | 第四代 | 第四代 | 第四代 |
| NVSwitch | N/A | N/A | 第三代 | 第三代 |
| NVLink Switch | N/A | N/A | N/A | 第一代 |
| NVSwich GPU 至 GPU 頻寬 | N/A | N/A | 每秒 900 GB | 每秒 900 GB |
| 彙總頻寬大小總計 | 每秒 900 GB | 每秒 3.6 TB | 每秒 7.2 TB | 每秒 57.6 TB |

資料來源: https://www.nvidia.com/zh-tw/data-center/hgx/

# HPC 趨勢與主流 - 使用容器

- 效能接近使用實體機器 (bare-metal)
- 方便部署不同版本軟體及環境
- 顯示卡驅動程式直接透通
- 封裝好的容器可以到任何環境執行
- 容器與系統互相隔離 (isolation)
- 容器：docker、singularity、nvidia-docker

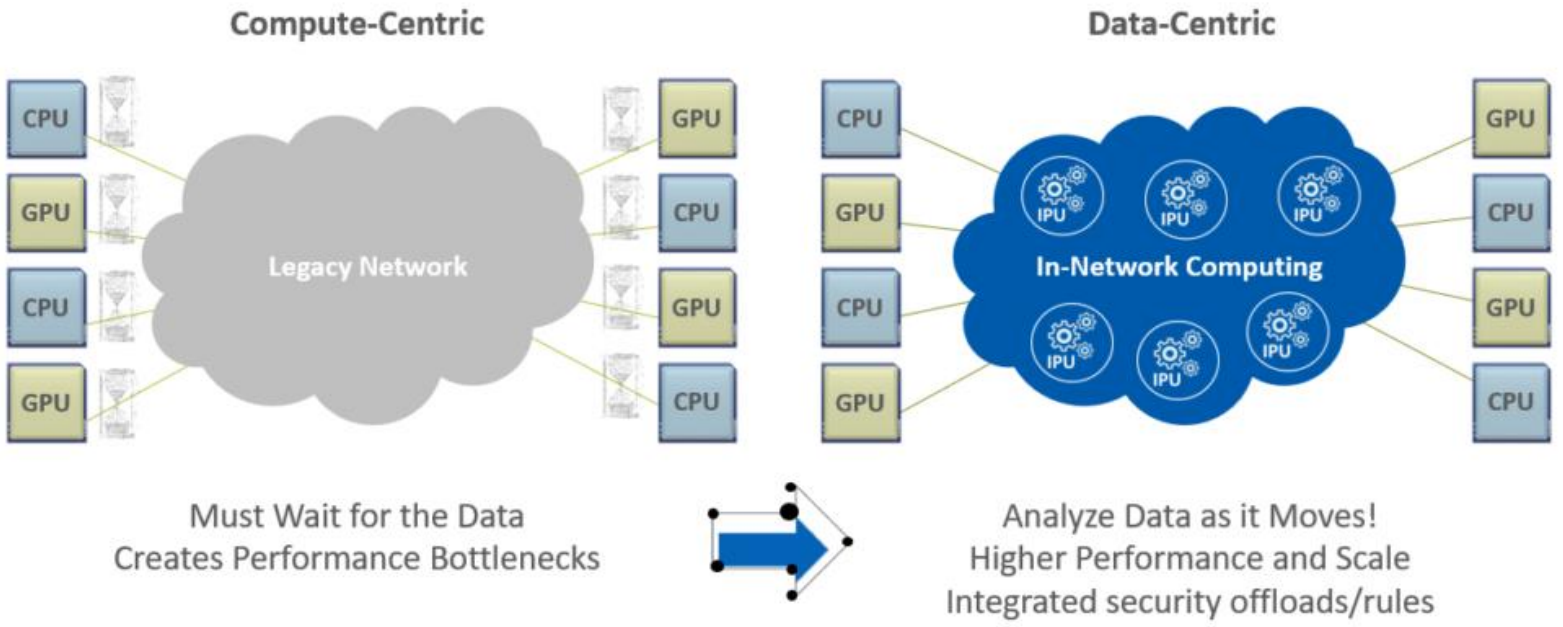


參考：https://cloud.google.com/containers, https://www.docker.com/, https://sylabs.io/docs/, https://github.com/NVIDIA/nvidia-docker

# HPC 趨勢與主流 - 資料處理單元 (DPU)

- 傳統資料搬移都是 CPU 的工作，移動資料會讓 CPU 閒置
- 未來移動資料不用透過 CPU，且在移動資料時還能進行運算
- Mellanox HDR 200G switch 和 HDR 網卡支援 in-networking computing (application offload capability) 和 in-network memory

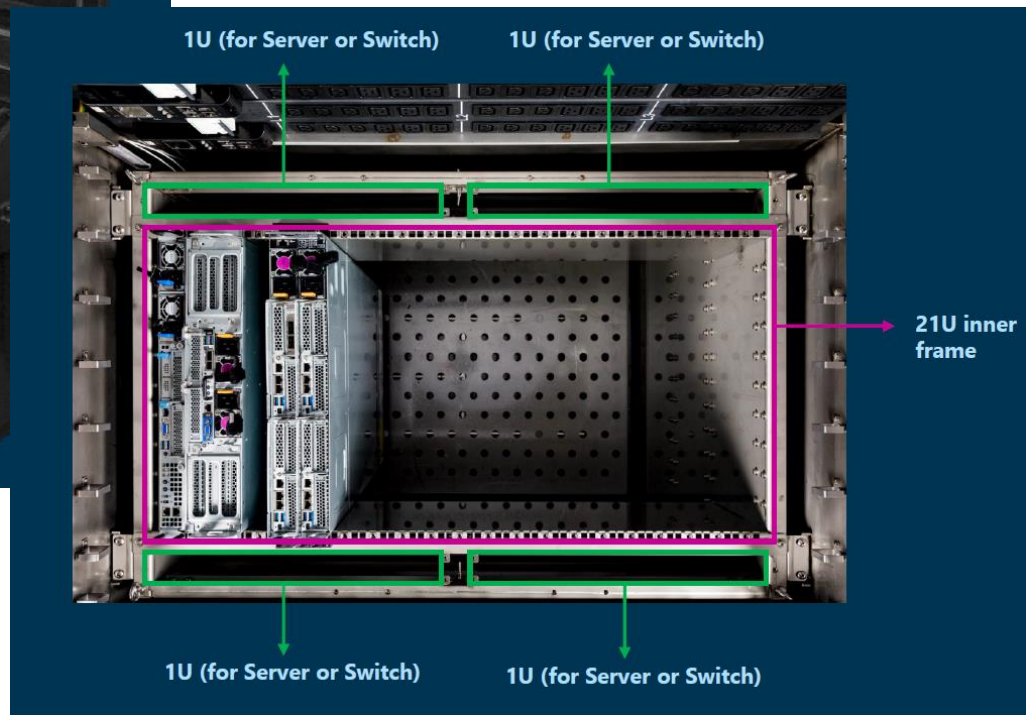



資料來源：https://blog.mellanox.com/2020/03/welcome-to-the-dpu-enabled-data-revolution-era/, https://www.mellanox.com/files/doc-2020/wp-in-network-computing-next-generation-hdr-200g-ib.pdf

# HPC 趨勢與主流 – 浸沒式冷卻系統

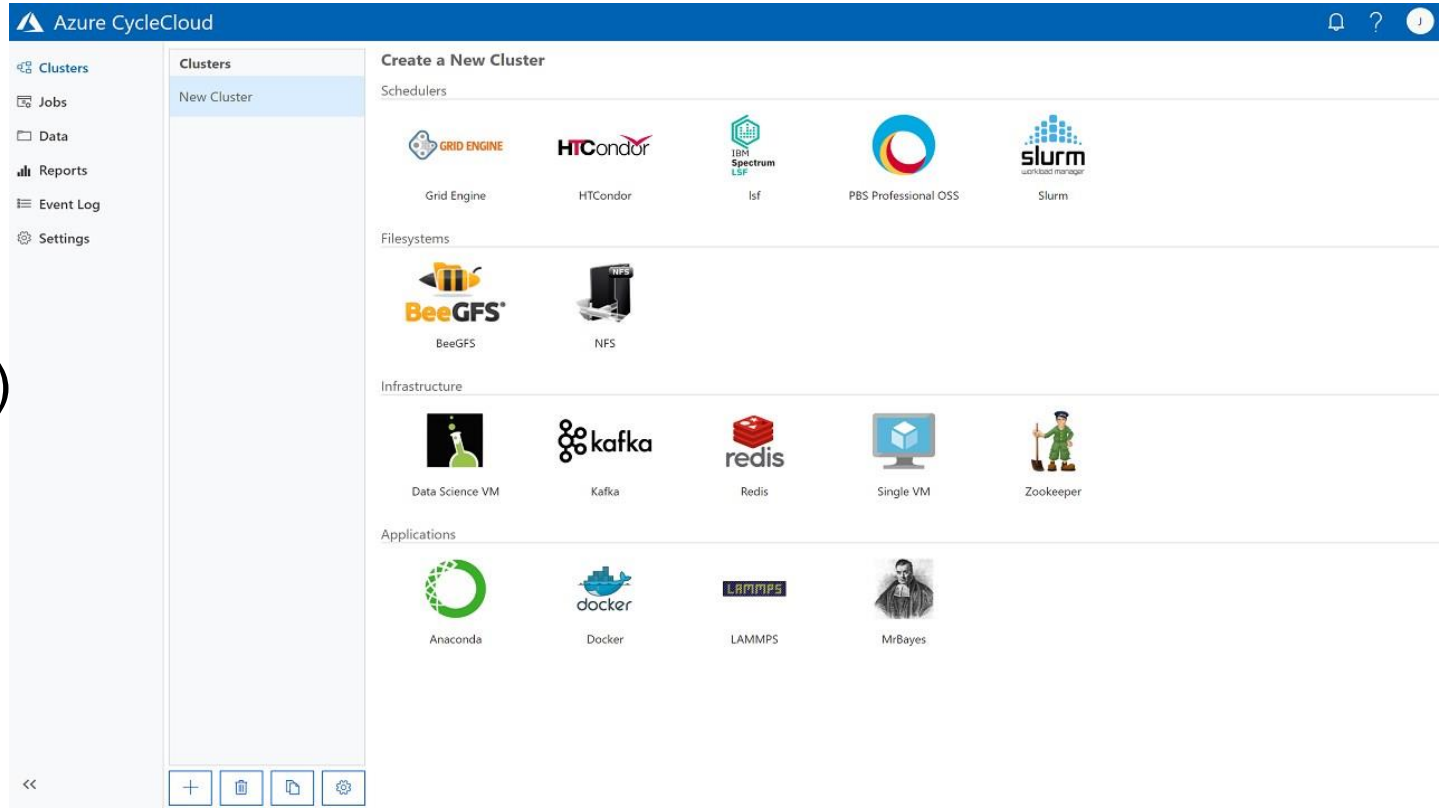
- 傳統用氣流帶走熱效率不彰
- 浸沒式冷卻(immersion cooling)，是透過將伺服器直接浸泡在不導電的液體中，將零組件產生的熱能傳導給流體不需要其他主動式的冷卻零件譬如散熱鰭片、導熱銅管或風扇等等，溫度上升的液體透過循環冷卻方式再回流繼續吸收熱能，有助於提升資料中心的能源效率。
- 可節省能源耗損可達 30~40%
- 節省機櫃空間

技嘉 25U Ready Tank



# HPC 趨勢與主流 - 雲端服務

- 以 Azure CycleCloud 為例：
  - 免建置 HPC Cluster
  - 排程軟體隨選 (Grid Engine, HTCondor, lsf, PBS Pro, Slurm)
  - 平行檔案系統隨選 (BeeGFS, GlusterFS, NFS, NetApp, lustre)
  - 動態增長、縮減 compute nodes
- 快速部署
- 可滿足臨時需求
- 可用最新的 CPU 或 GPU
- 免地端機房電力管理
- 有技術人員隨時支援




資料來源：https://azure.microsoft.com/en-us/features/azure-cyclecloud/, https://docs.microsoft.com/en-us/azure/cyclecloud/tutorials/create-cluster?view=cyclecloud-7

# 規劃 HPC Cluster

- 經費預算??
- 程式特性及用途，攸關 CPU 或 GPU 計算能力及主記憶體需求
- 網路交換能力
- 儲存設備速度、容量及資料備份保護（TSM）
- 作業系統及軟體（編譯器、公用函式庫、應用軟體）
- 機房空間及樓板承重（1000 kg/$m^2$）
- 機房電力、空調能源消耗（KW/BTU）
- 資訊安全（firewall, patch)