

2023 AMD Solutions Day

探索HPC高性能叢集計算

凱穩電腦股份有限公司

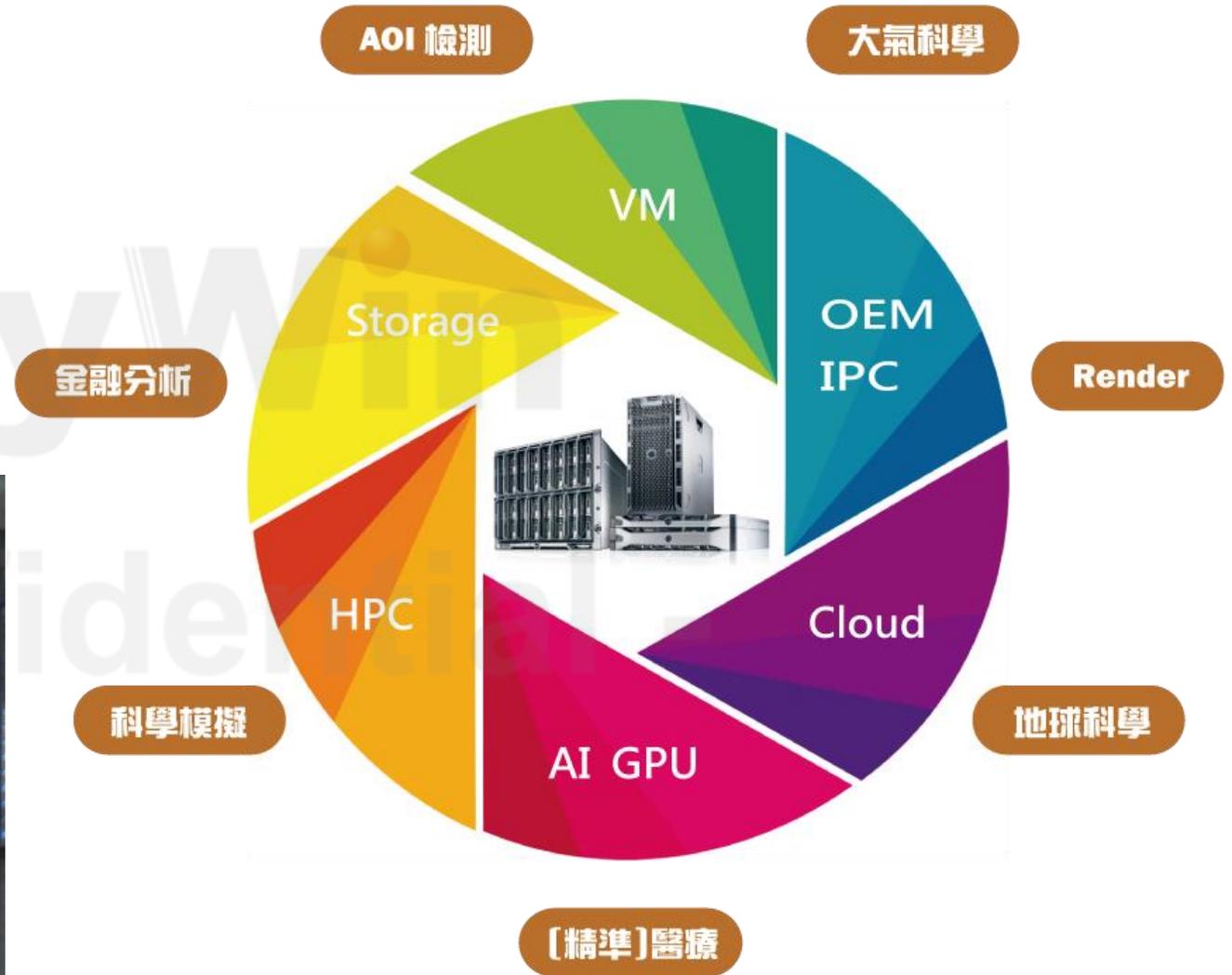
Key Win Computer Co., LTD.

Adam Chen

2023-06-20

about KeyWin 凱穩電腦

- 伺服器零件銷售
- 各白牌/品牌伺服器代理銷售
- 建置機房伺服器集群
- 佈建整合HPC / AI 系統



What is HPC? Why need HPC?

- HPC 是 High Performance Computing (高效能運算) 的縮寫。HPC 是一種計算科學和工程領域，旨在利用強大的計算能力來處理大量的數據和複雜的計算問題。
- HPC 的目標是提供超高速的計算能力，以解決需要大量計算、記憶和儲存資源的問題。這種計算能力通常來自於高效能的超級電腦或者由多台計算機組成的集群。
- HPC 可以應用於各種領域，包括科學研究、工程模擬、氣候預測、基因組學、藥物設計、金融模型等。它能夠加速數據處理和分析的速度，並提供更準確的模擬和預測結果。
- HPC 系統通常具備以下特點：
 1. **高性能處理器**：HPC 系統通常配備高效能的處理器，例如多核心的 CPU、GPU (圖形處理器) 或者其他特殊用途的加速器。
 2. **大容量記憶體**：HPC 系統需要具有足夠的記憶體容量，以處理大型數據集和複雜的計算任務。
 3. **高速互連網絡**：HPC 系統中的計算節點之間需要高速的互聯網絡連接，以實現高效的數據通信和協同計算。
 4. **平行計算**：HPC 系統可以同時運行多個計算任務，並使用並行計算的技術來分解和處理大型問題。
- 總之，HPC 是利用高效能計算資源解決大規模和複雜問題的領域，它在科學、工程和其他領域的研究和應用中扮演著重要角色。

What is HPC/HPCC

- **HPC is 高效能計算**

透過高效能處理器解決大量、複雜、高速的計算需求，其CPU工作負載幾乎全天**100%**

- **HPCC is HPC Cluster 高效能計算叢集**

HPC電腦的叢集，通常User希望每一台主機運算資源100%被利用，且7x24工作負載 **100%**

- **不同於HPC/HPCC 計算的電腦服務**

ex: PC, VM, 一般常態服務性主機...，提供非24小時100%負載的服務，包含
Web service, ERP系統...

Single Computer

15 years ago...



Powerful Computer
(AMD Opteron Server
– 2Core x 8CPU)
16Core

Today 2023



Personal Computer
(AMD Ryzen™ Threadripper™ PRO)
64Core

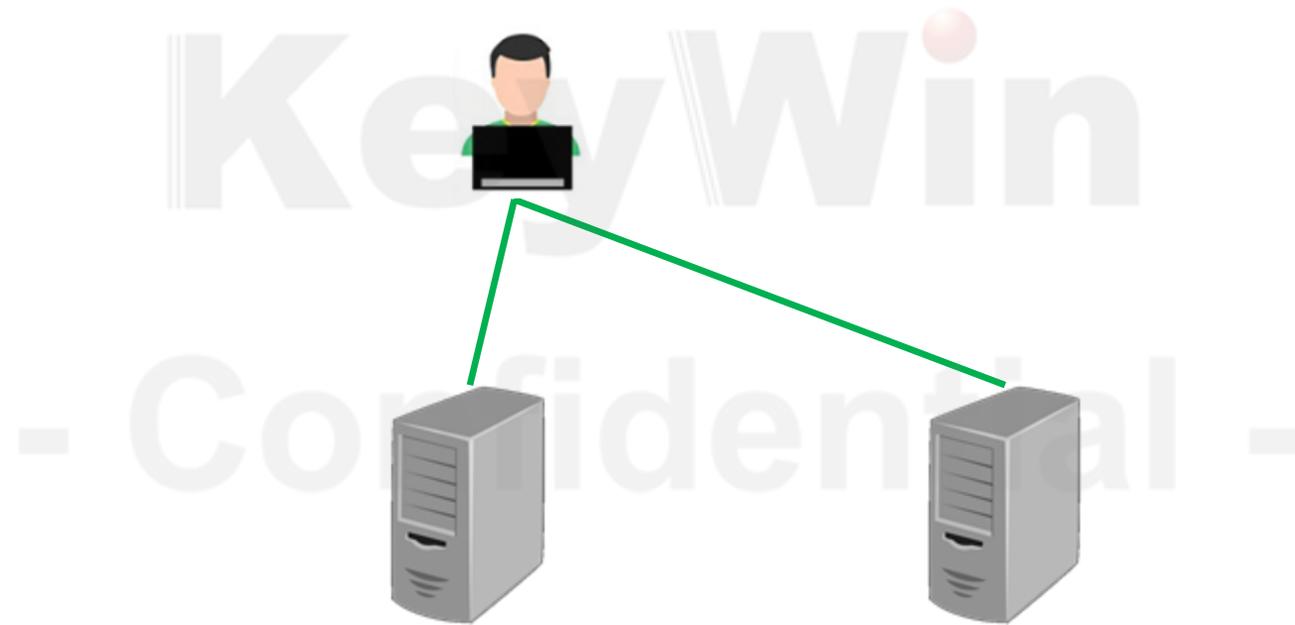
Today 2023



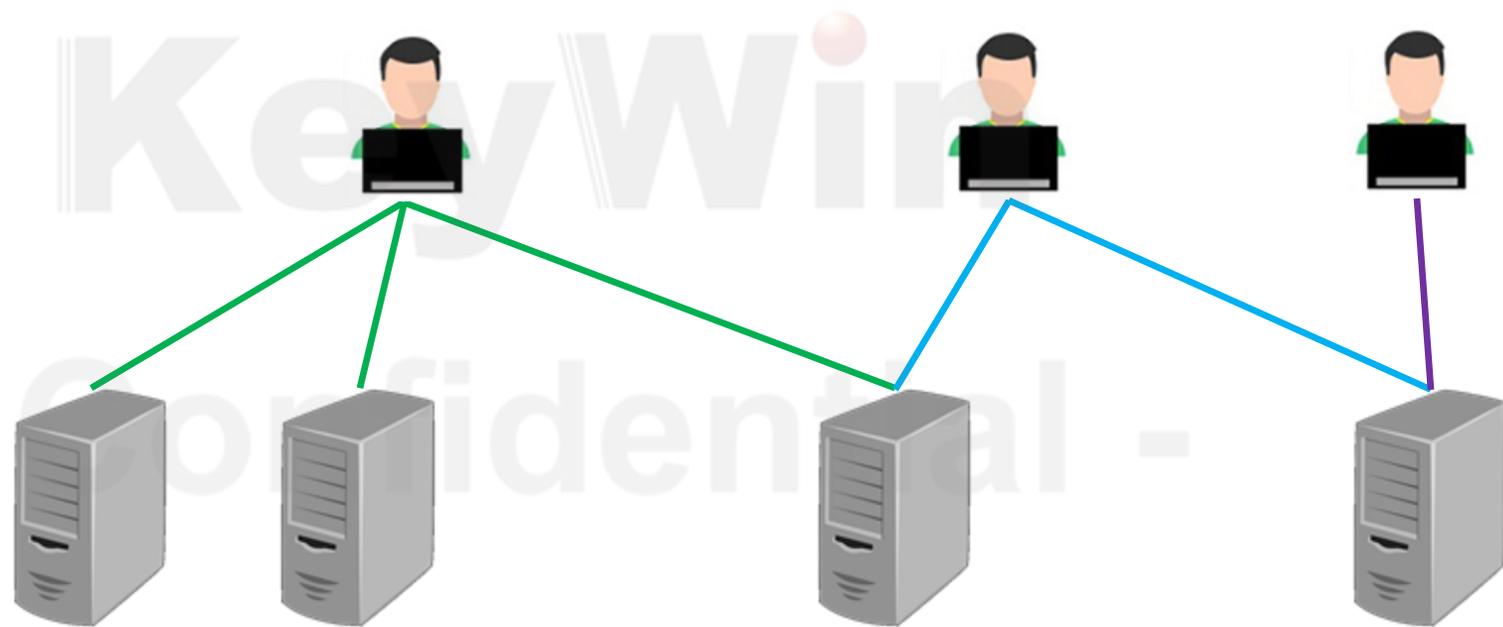
2U Server
(AMD 9654 96Core * 2)
192Core

KeyWin
- Confidential -

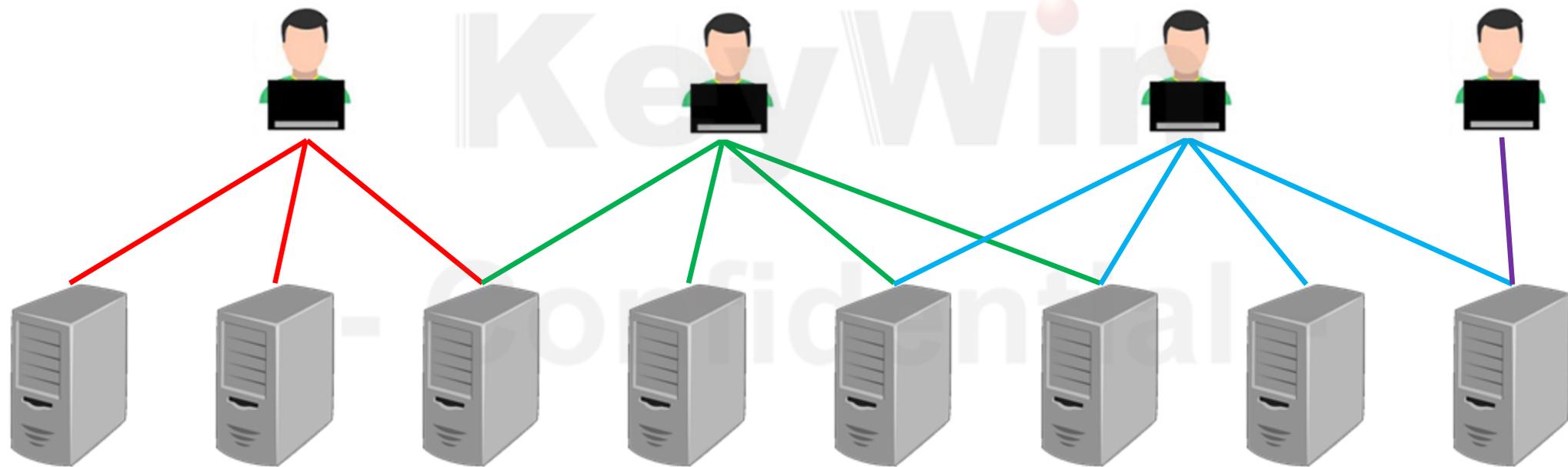
HPC使用情境 - A lot of Computers/Servers



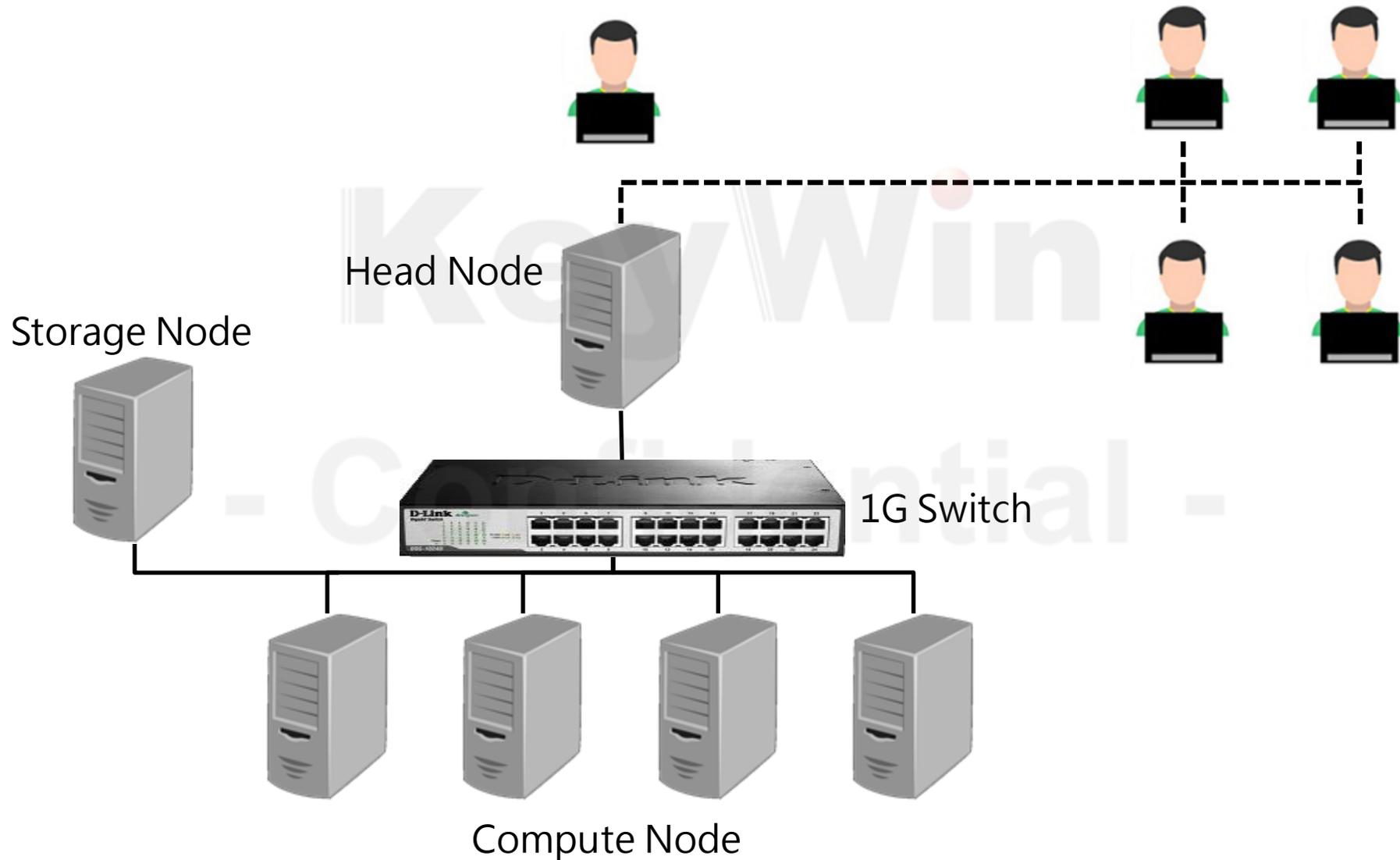
HPC使用情境 - A lot of Computers/Servers



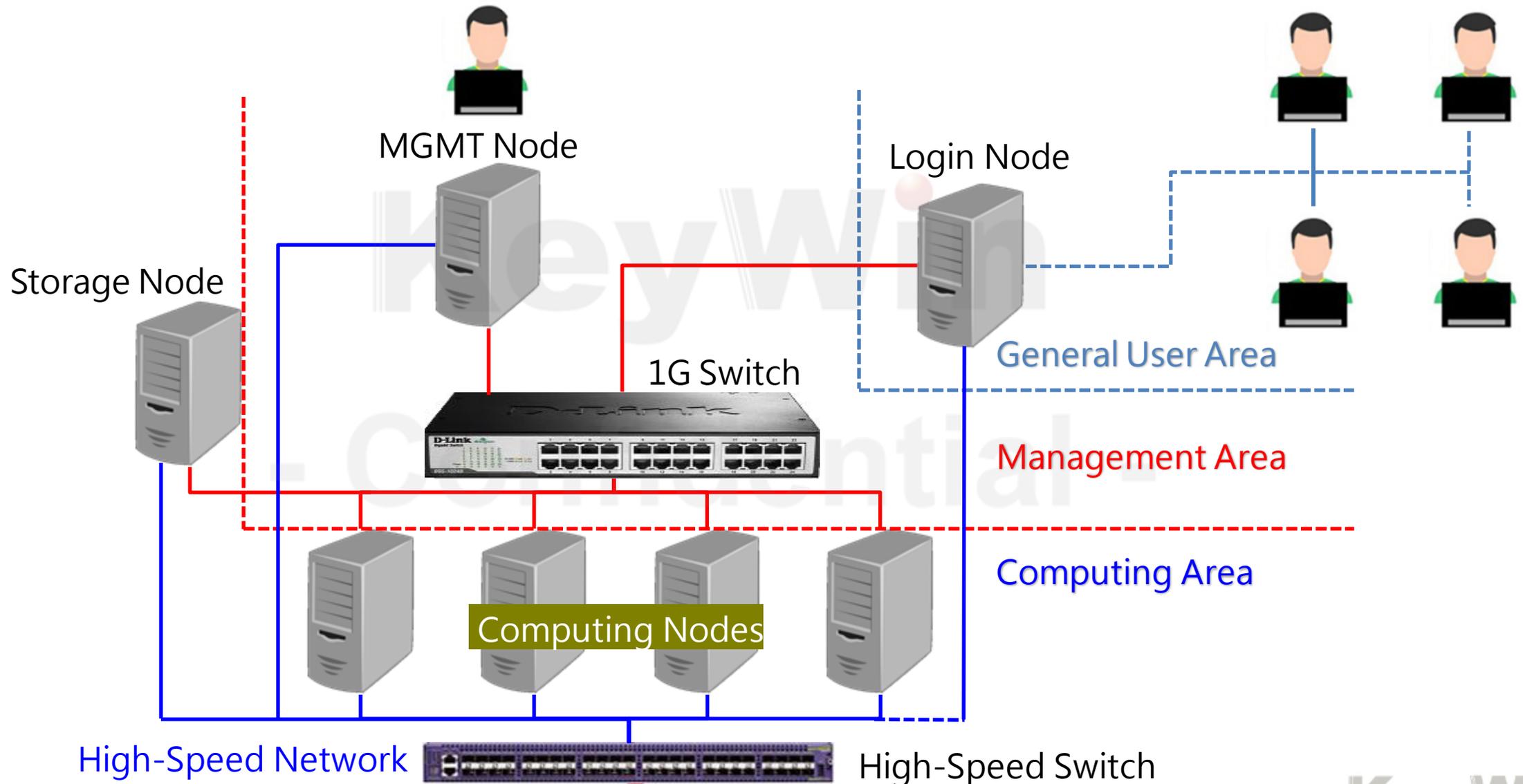
HPC使用情境 - A lot of Computers/Servers



HPC使用情境 - Managed Cluster

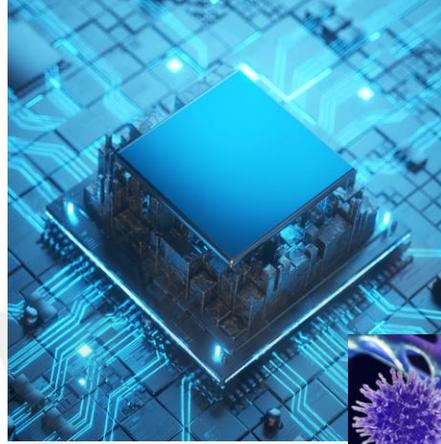


HPC使用情境 – HPC 標準架構



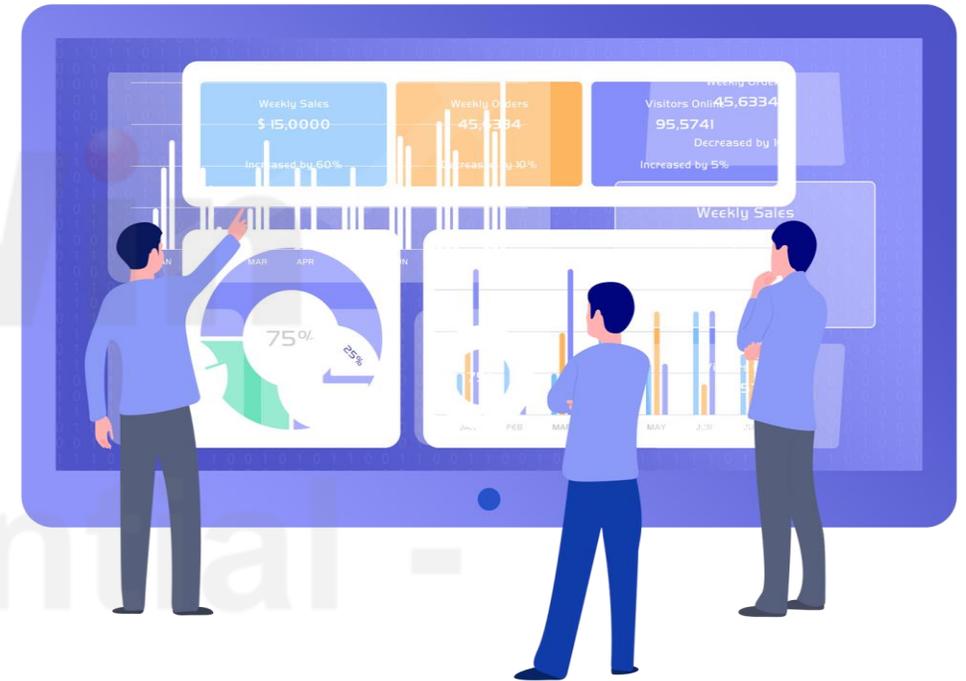
Why HPC? For Who?

- Covid-19
- 氣候變遷
- 物理/化學
- 商業/金融分析
- 晶片戰爭 (EDA)
- 醫療研究

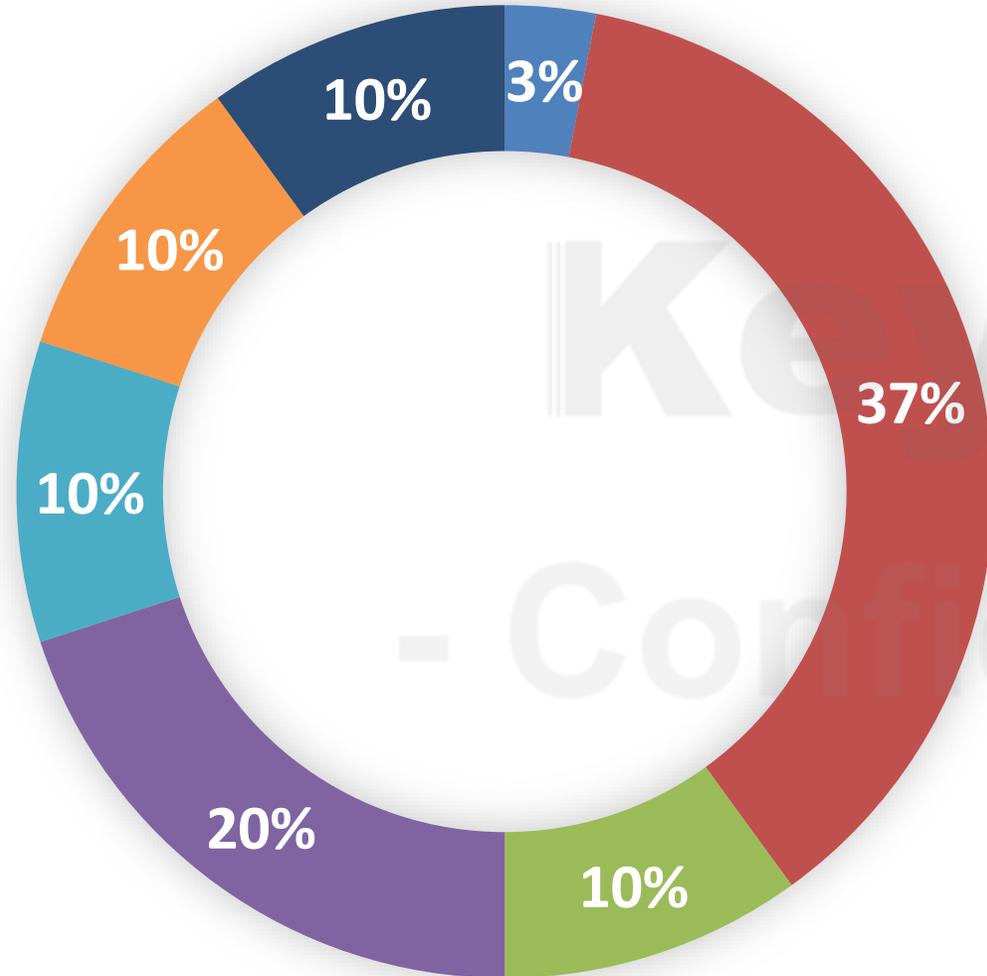


HPC 系統考量

- 應用?
- 有哪些計算軟體?
- 多少人使用?
- 有無管理者?
- CPU? GPU?
- 儲存空間?
- 設備使用環境? 空間/電力/空調散熱
- 使用年限?
- 可接受down time多久? 服務等級 5x8 or 7x24?
- 預算?



HPC Budget?



■ 管理主機

■ 運算主機 30-50%

■ 網路系統 5-15%

■ 儲存系統 5-25%

■ 管理監控軟體 0-10%

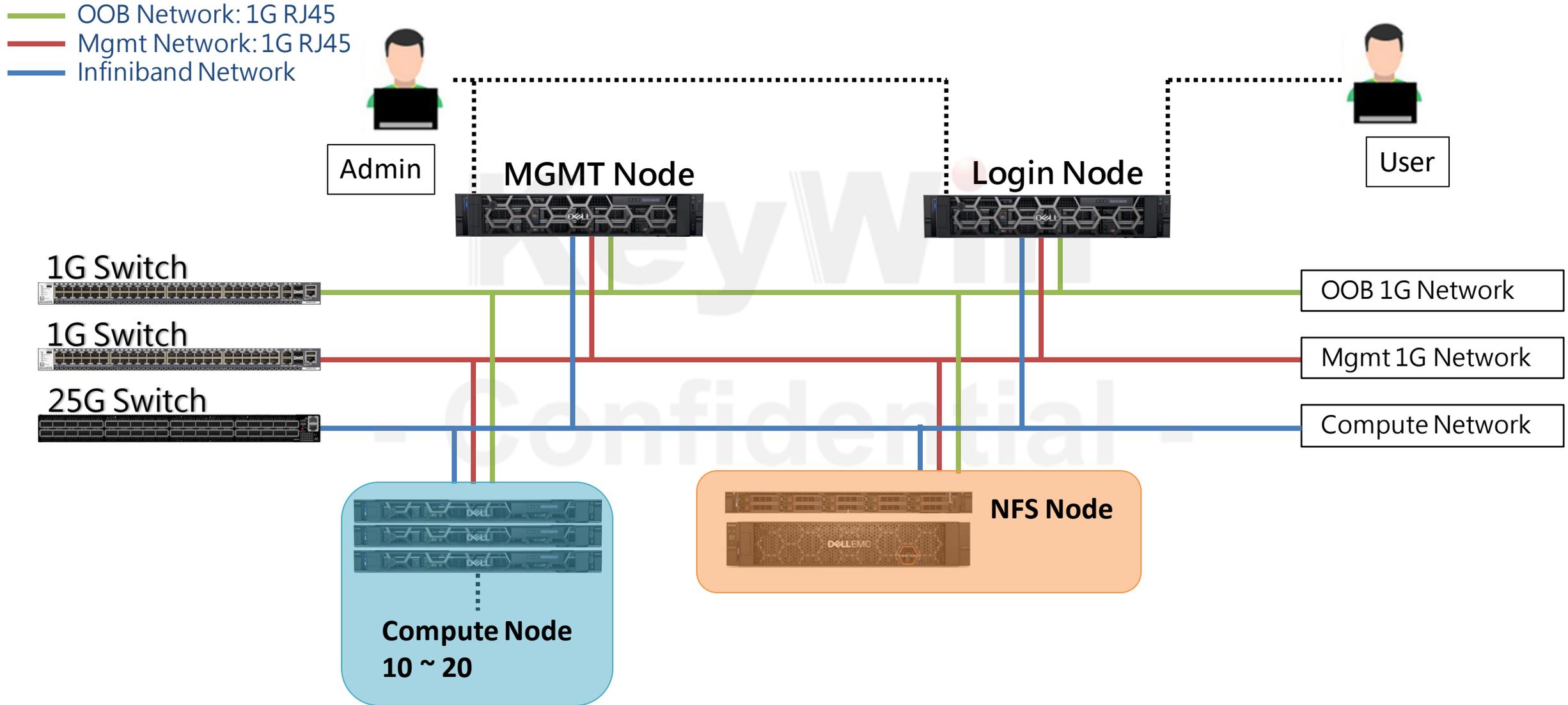
■ 機電整合 5-15%

■ 建置服務 0-15%

■ 維運成本(電力,人力,保固)- 不計入

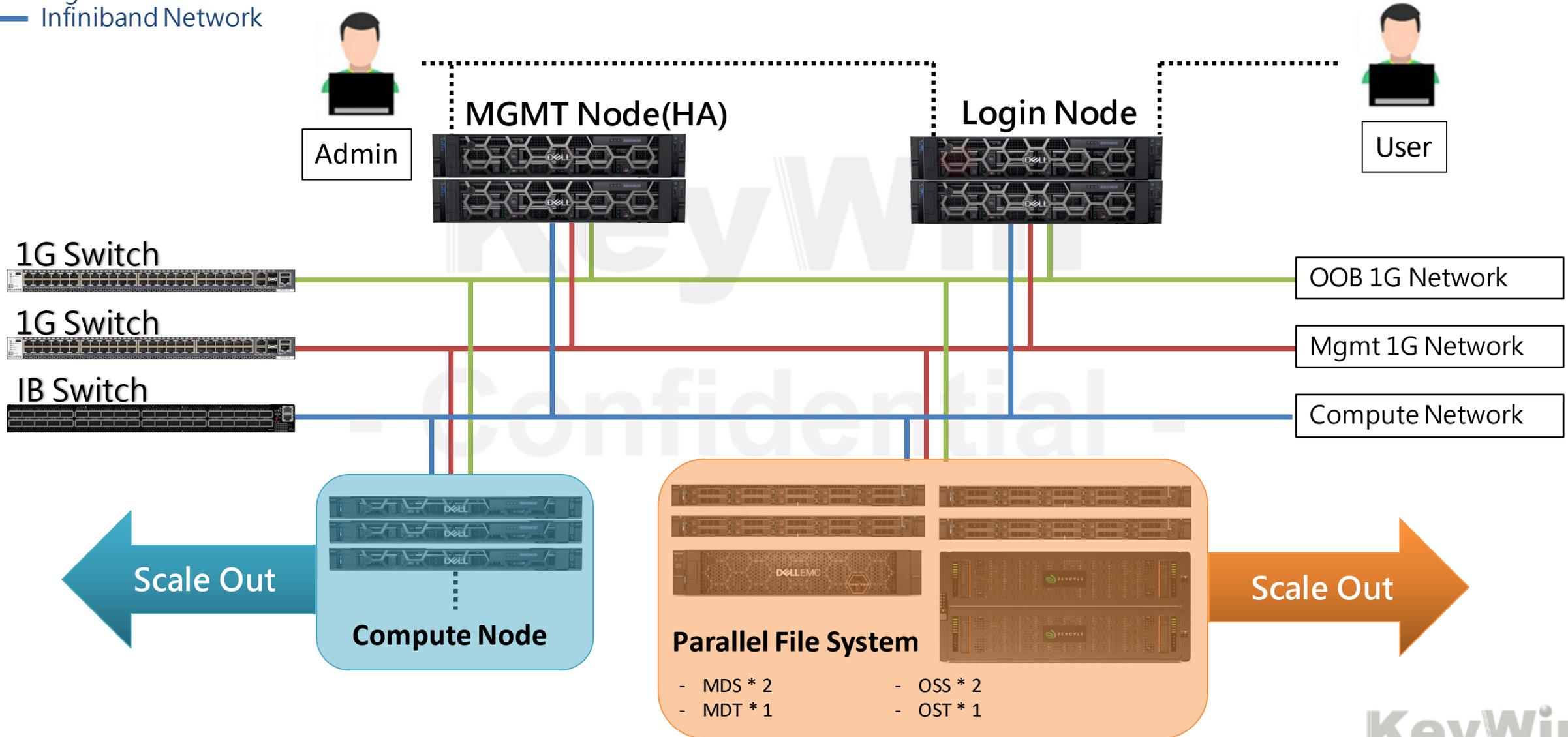
■ 計算軟體 - 不計入

小型HPCC基本架構圖



中大型HPC網路架構圖

- OOB Network: 1G RJ45
- Mgmt Network: 1G RJ45
- Infiniband Network



HPC節點角色

- **MGMT Node 管理節點**

通常指的是管理節點 (Management Node) 。這是一個專門用於執行一些關於叢集管理的任務的節點，包括但不限於系統配置、性能監控、故障診斷和恢復、軟體部署和更新，以及安全管理等等。管理節點通常運行工作排程器，如 Slurm 或 PBS，這些排程器負責在叢集中分配計算資源，排列並管理使用者的計算任務。

- **Login Node 登入節點**

也被稱為 "Access Node" 或者 "Front-end Node"，是使用者進行遠程登錄，提交計算任務，編譯程式碼，以及管理資料等工作的節點。

- **Compute Node 計算節點**

是叢集中的一個或多個節點，專門用於執行使用者提交的計算任務或者程式。在一些需要大量計算資源的情況下，多個 Compute Node 可能需要通過網路協同工作，一起完成一個計算任務。

- **Parallel File System 平行檔案系統**

高效能計算 (HPC) 中的平行檔案系統， ex: Lustre、BeeGFS、WekaFS、GPFS... 是一種能在多節點同時讀寫數據的檔案系統，主要用於提高資料存取效率和速度。它將檔案切分為多個部分，分佈在多個硬碟或儲存節點上，實現大量同時讀寫操作。此外，平行檔案系統也提供數據冗餘和錯誤檢測功能，以增強數據的可靠性和容錯性，並允許多個節點共享同一份數據資料。

- **OoB 1G Switch 硬體網路**

Out-Of-Band (OOB) 1G Switch 主要負責管理和監控系統硬體的健康狀態和性能。它提供一條獨立的通道來監控硬體狀態、警報、以及其他硬體相關的資訊，如溫度、風扇轉速、電源狀態等，並用於執行硬體配置和管理，如韌體更新、遠端開機等。此外，這個通道也可以提供遠端診斷和故障修復的能力，以提高系統的可用性和穩定性。

- **Mgmt 1G Switch 系統網路**

在高效能計算 (HPC) 環境中，MGMT (Management) 1G Switch 主要用於支援網路管理活動。它連接所有節點 (MGMT Node)，並提供網路通道來執行各種管理任務，如配置管理、工作排程、性能監控、安全管理、以及軟體和系統更新等。這個管理網路通常與用於高性能計算和資料傳輸的主要網路分開，以防止管理流量影響計算和資料傳輸的性能。

- **Compute Network (IB Switch) 高速網路**

在高效能計算 (HPC) 環境中，InfiniBand (IB) Switch 是核心網路組件，主要負責在各個節點 (如管理節點、登錄節點、計算節點) 之間進行高速、低延遲的資料傳輸。透過優化的路由和交換能力，IB Switch 能有效地將大量並行的資訊流導向正確的目的地，進而支援 HPC 的大規模並行運算和高效的資料共享。

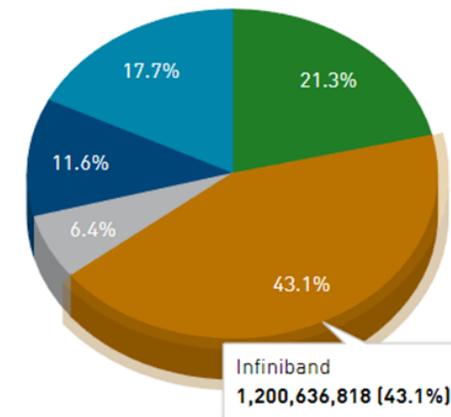
HOW to choose HPC Server?

節 點	型號/規格	說 明
Compute Node	<ul style="list-style-type: none"> - CPU: AMD/Intel x 2 - Memory: 384GB (16GB x 24) - OS Drive: SSD x 2 (RAID 1) - NIC: HDR Infiniband Single Port NIC x 1 - LAN: 1G Dual Port - PSU x 2 	<ol style="list-style-type: none"> 1. 設備以2 CPU Socket為主，其他4/8 CPU Socket設備通常是有特殊用途。CPU 盡量選擇2.4GHz 以上高時脈為主 2. 一般計算節點記憶體大小之選取，以節點總記憶體與CPU核心數之比值考量(Total Memory Size/Total CPU Cores)，一般比值介於2~4GB/Core。 3. 記憶體通道插滿，ex: AMD EPYC 9004 support 12 Channel 4. 少數計算節點可以用大記憶體空間，例如1TB以上總記憶體空間，作為特殊需要大記憶空間之計算使用，但前提是有足夠經費的話。 5. 少數計算節點可以安裝GPU，作為特殊需要GPU之計算使用，但前提是有足夠經費的話。 6. OS：CentOS/Red Hat/Rocky Linux/Almalinux皆可，以作業系統穩定考量。 7. 使用高密度伺服器或高瓦度CPU應注意環境散熱規劃。
MGMT/Login Node	<ul style="list-style-type: none"> - CPU: AMD/Intel x 2 - Memory: 256GB (16GB x 16) 3200MT/s - OS Drive: SSD x 2(RAID 1)或x3 x4 x5..(RAID 6) - NIC: HDR Single Port NIC x 1 - LAN: 1G Dual Port - PSU x 2 	<ol style="list-style-type: none"> 1. CPU應與Compute Node同一世代。 2. MGMT Node的OS Disk可以考慮組成Raid 6，因為此節重要性較高。 3. OS：CentOS/Red Hat/Rocky Linux/Almalinux皆可，以作業系統穩定考量。

HOW to choose HPC Network?

產品	型號/規格	說明
高速網路系統 IB Switch	Infiniband HDR Switch QM8700 - HDR 200G x 40Port - PSU * 2	- 中小型規模之HPC可採用具備IB網路管理功能之機種
管理網路系統 1G Switch	節點管理網路, 硬體管理網路 - 1G x 48Port - PSU * 2	

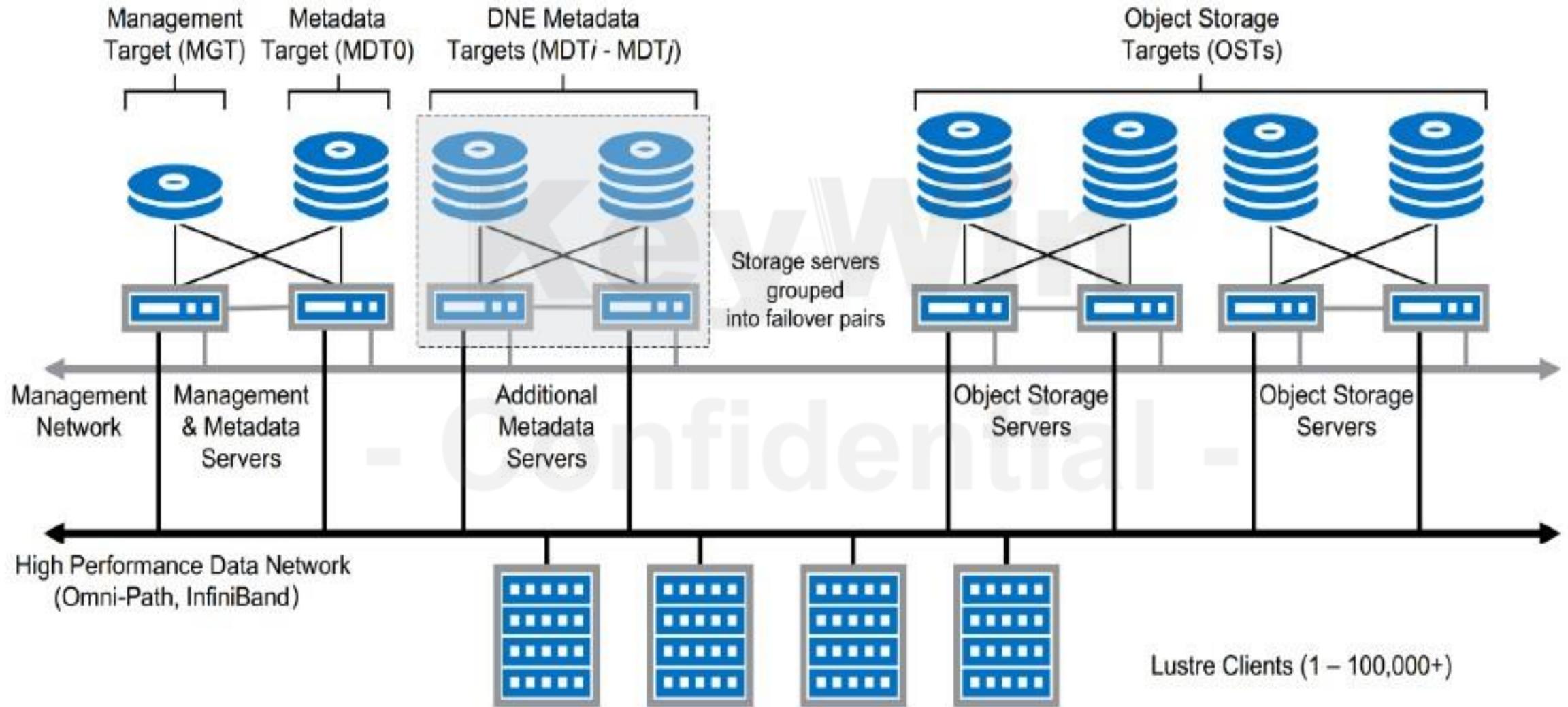
Interconnect Family Performance Share



QM8700系列比較表

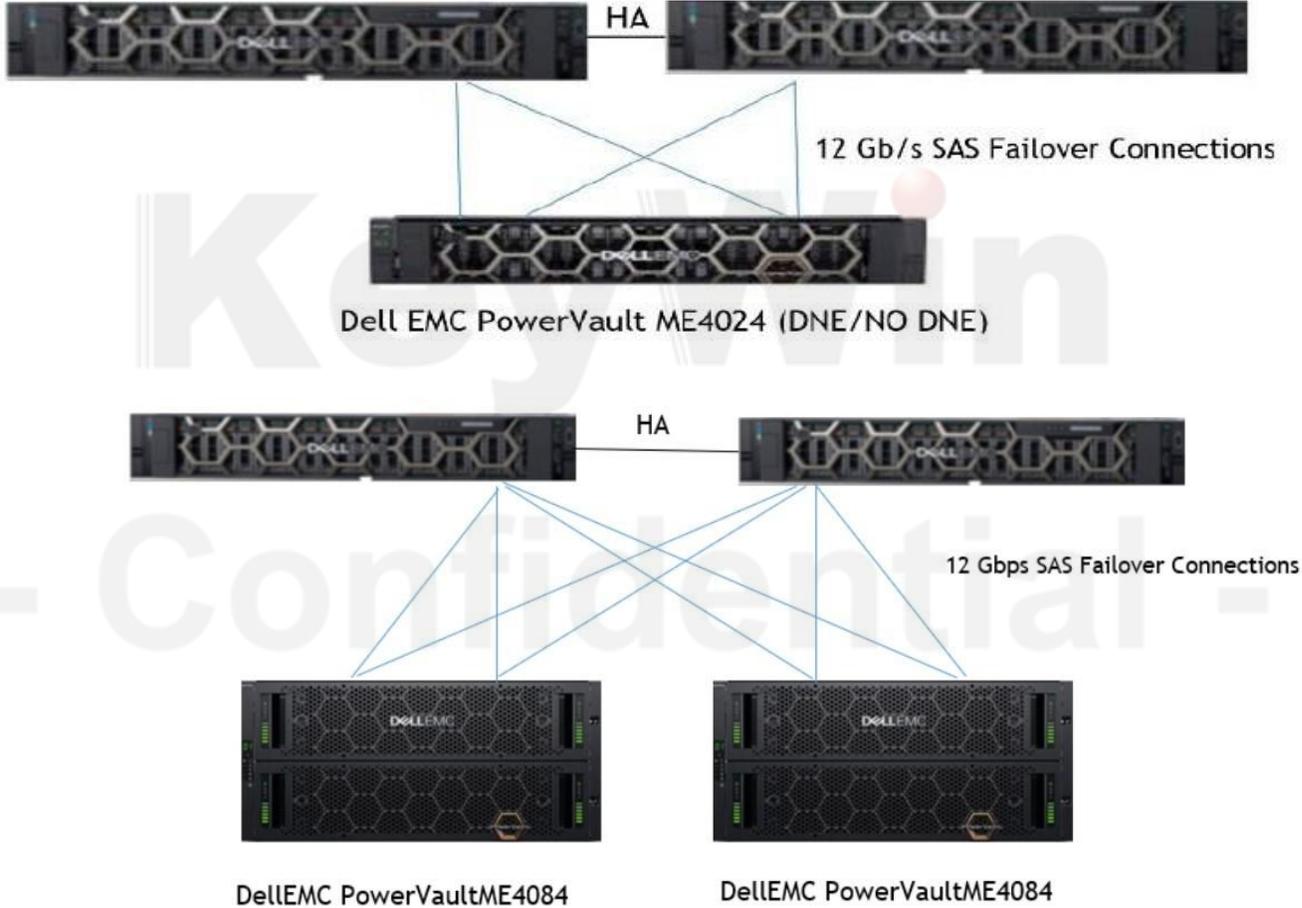
	連結速度	連接埠	高度	交換器容量	冷卻系統	骨幹模組	分葉模組	介面	PSU 數量	管理	子網路管理軟體
QM8700	每秒200Gb	40	1U	每秒 16Tb	氣冷	20	20	QSFP56	2	頻內/頻外	+
QM8790	每秒200Gb	40	1U	每秒 16Tb	氣冷	20	20	QSFP56	2	頻內	-

How to Choose HPC Storage? (HDD?)



* From <https://www.lustre.org>

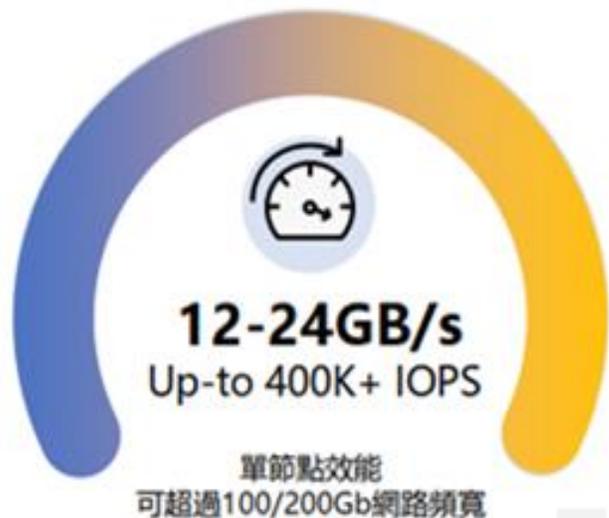
HPC 平行檔案系統 HA 連接架構



How to Choose HPC Storage? (NVMe SSD?)

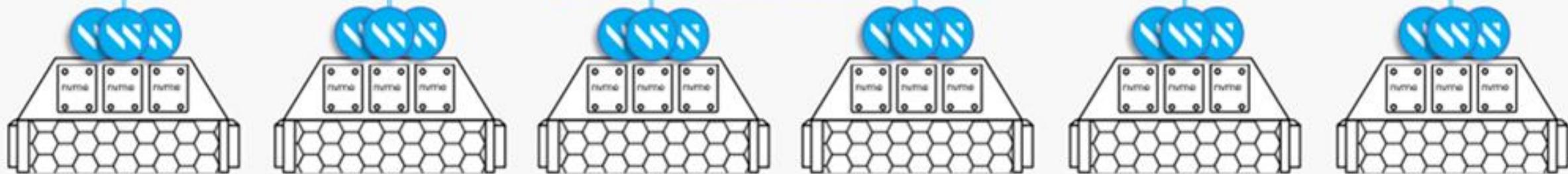
現代平行檔案管理系統

專為 NVMe 存儲與高速網路打造



- Weka POSIX
- Weka service
- Weka backend

乙太網路或無限帶寬



How to manage HPC



排程軟體

PBS pro/OpenPBS/Torque、SLURM、LFS



編譯器

GNU Compiler、Intel Compiler(Intel OneAPI)、AMD AOCC Compiler、PGI Compiler(NVIDIA HPC SDK)
Support C, C++, Fortran



Library

MPI : Intel MPI(Intel OneAPI)、OpenMPI、MPICH...

Math : Intel MKL(Intel OneAPI)、Lapack、ScaLapack、FFTW...

IO : HDF5、netCDF4、PnetCDF、PIO...



環境變數管理模組

Environment Module、Lmod

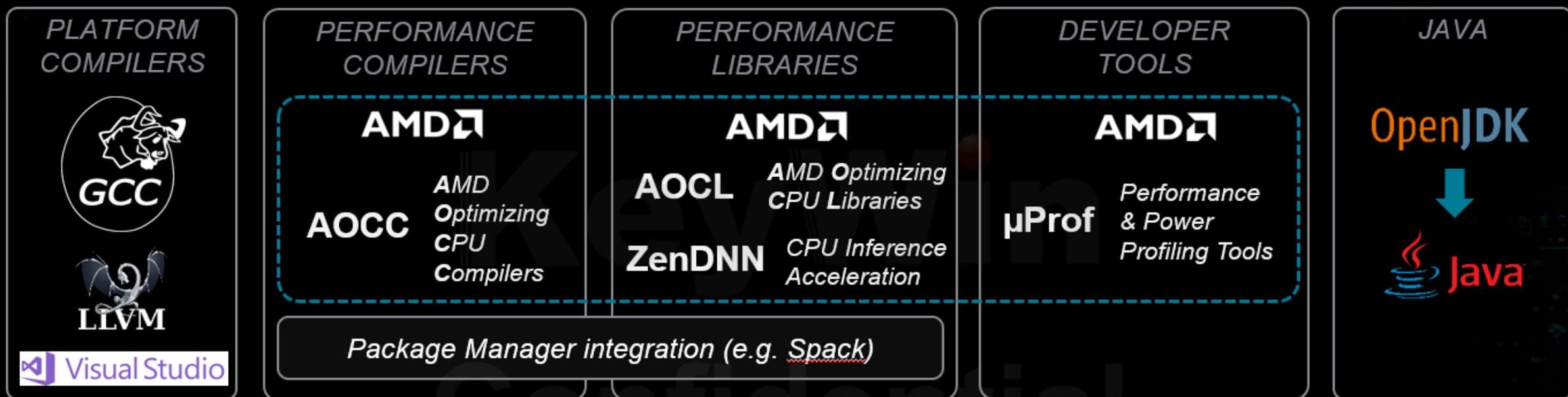


硬體/系統監控

Dell Openmanage Enterprise、HPE OneView、ASUS Control Center、
LibreNMS、Ganglia、BCM

TOOLCHAIN ECOSYSTEM

AMD EPYC™ Software Development Environment



Use AMD tools for best performance and code efficiency on EPYC CPUs

Compilers → Focus on delivering the best out-of-the-box code generation for C, C++, Fortran, Java.

Libraries → Support common kernels for core math, solvers and FFT.

Profiling tools → Enable developers to access the full capabilities of EPYC CPUs.

All tools available at <https://www.amd.com/en/developer/zen-software-studio.html>

LibreNMS Dashboard

The dashboard provides a comprehensive overview of network health and performance. It includes several key sections:

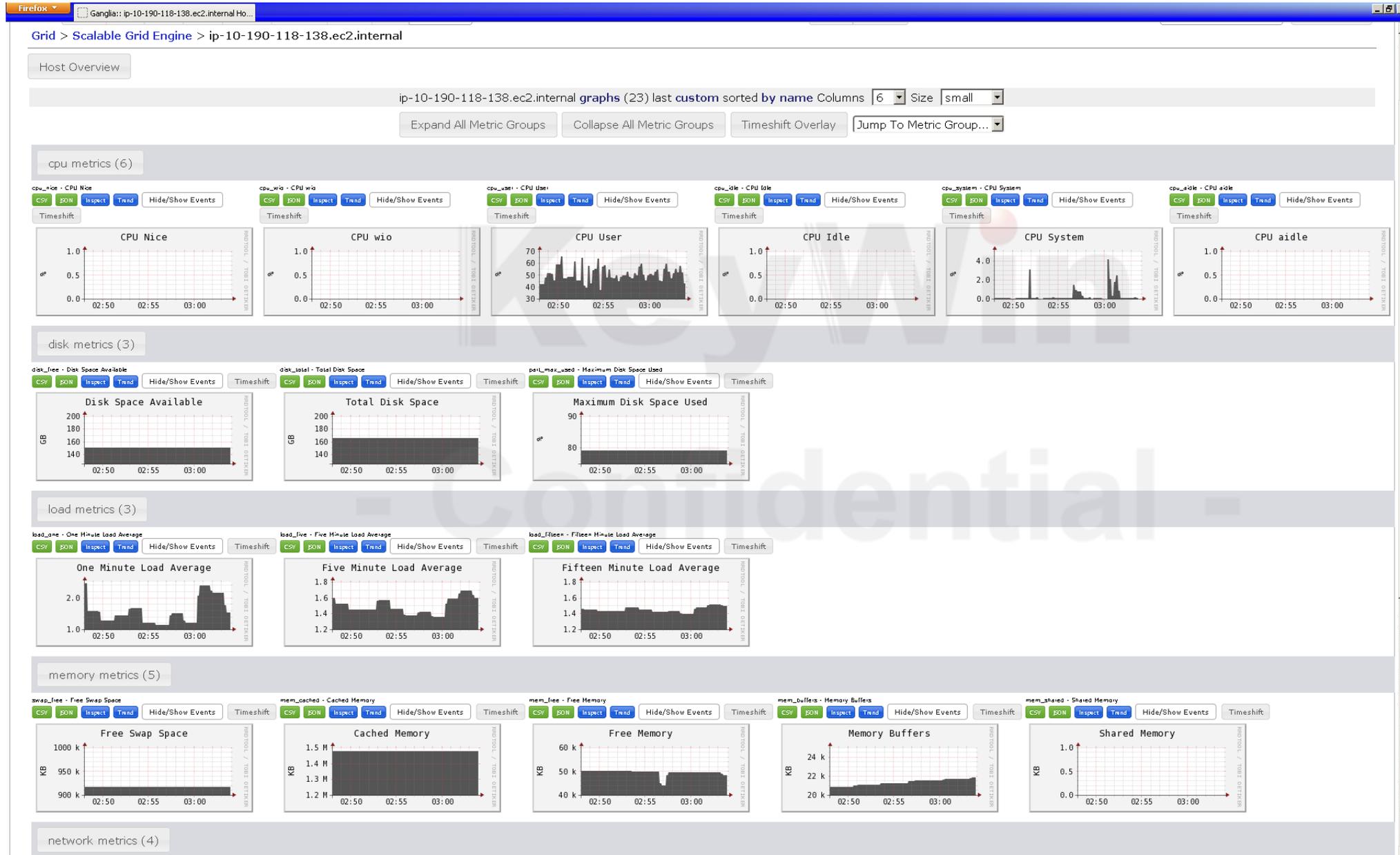
- Availability Map:** Shows the status of hosts and services. Hosts: 378 up, 2 warn, 0 down. Services: 14 up, 0 warn, 0 down.
- Device-summary-horiz:** A summary table of device and port status.
- Worldmap:** A geographical map showing the location of network devices.
- Top CPU, Top Memory, Top-interfaces:** Performance charts for the most active devices and interfaces.
- Alerts:** A table of system alerts with search and filter capabilities.
- Eventlog:** A detailed log of network events and messages.

	Total	Up	Down	Ignored	Disabled
Devices	381	380	0	1	0
Ports	16309	6382	6323	177	109
Services	14	14	0	0	0

Status	Rule	Hostname	Timestamp	Severity	Acknowledge	Procedure
No results found!						

Datetime	Hostname	Type	Message	User
2017-11-08 19:32:59	nbms-idf-704-wlan	discovery	CDP discovery of CN23D323XT (172.16.247.72) failed - Could not connect to CN23D323XT, please check the snmp details and snmp reachability	
2017-11-08 19:31:58	lse-cafe-lan-2	36	ifHighSpeed: 10 -> 1000	
2017-11-08 19:31:58	lse-cafe-lan-2	36	ifSpeed: 100000000 -> 1000000000	
2017-11-08 19:31:30	ss-idf-wlan	16	ifOperStatus: down -> up	
2017-11-08 19:31:23	ec-noc-lan-2-top	25	ifHighSpeed: 10 -> 1000	
2017-11-08 19:31:23	ec-noc-lan-2-top	25	ifSpeed: 100000000 -> 1000000000	
2017-11-08 19:31:18	ngc-411-lan1-faculty	42	ifHighSpeed: 1000 -> 10	
2017-11-08 19:31:18	ngc-411-lan1-faculty	42	ifSpeed: 1000000000 -> 100000000	
2017-11-08 19:31:17	kre-mdf-lan-2-bottom	19	ifHighSpeed: 100 -> 1000	
2017-11-08 19:31:17	kre-mdf-lan-2-bottom	19	ifSpeed: 1000000000 -> 1000000000	
2017-11-08 19:31:17	ngc-504-lan1-student	23	ifHighSpeed: 100 -> 1000	

Ganglia



How to Benchmark/Test HPC

HPL

HPL (High Performance Linpack) 是衡量高效能計算系統性能的標準工具，用於解決密集型線性代數問題。通過測試系統在執行浮點運算（特別是矩陣運算）時的性能，HPL提供了一種衡量和比較HPC系統性能的方法。

HPCG

HPCG (High Performance Conjugate Gradients) 是一種測試高效能計算系統性能的基準測試工具，用於模擬現實世界中的計算負載。與傳統的Linpack基準測試相比，HPCG更加能反映當前實際應用程序中的計算和數據訪問模式。

Memtest86

Memtest86是一款檢測電腦記憶體錯誤的自由及開源的軟體。它會對系統的隨機存取記憶體 (RAM) 進行嚴格的測試，檢查並識別可能的故障或不良部分，以確保記憶體的穩定性和正確性。

IOR

IOR (Interleaved or Random) 是一種常用於測試HPC系統中並行文件系統性能的工具。它可進行連續或隨機的讀寫測試，以衡量儲存系統在不同工作負載下的性能，從而幫助用戶評估和優化他們的儲存基礎設施。

Thank You

www.keywin-computer.com
HPC solution provider

專注於客戶需求
從零件、主機到解決方案

Server Hardware Supplier

KeyWin