檢視解決方案設計

BeeGFS on NetApp with E-Series Storage

NetApp
January 27, 2026

目錄

檢視解決方案設計

設計總覽

需要特定設備、纜線和組態來支援BeeGFS on NetApp解決方案、此解決方案將BeeGFS平行檔案系統與NetApp EF600儲存系統結合在一起。

深入瞭解：

- "硬體組態"
- "軟體組態"
- "設計驗證"
- "規模調整準則"
- "效能調校"

衍生架構的設計與效能差異：

- "高容量建置區塊"

硬體組態

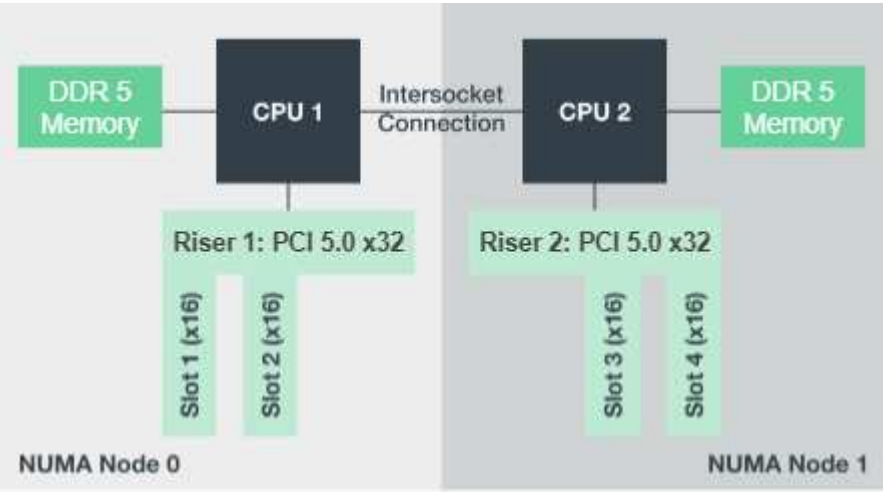
NetApp上BeeGFS的硬體組態包括檔案節點和網路纜線。

檔案節點組態

檔案節點有兩個CPU插槽、設定為獨立的NUMA區域、包括本機存取相同數量的PCIe插槽和記憶體。

InfiniBand介面卡必須安裝在適當的PCI擴充卡或插槽中、因此工作負載必須在可用的PCIe線道和記憶體通道之間取得平衡。您可以將個別BeeGFS服務的工作完全隔離到特定NUMA節點、藉此平衡工作負載。目標是從每個檔案節點取得類似的效能、就像是兩個獨立的單一插槽伺服器一樣。

下圖顯示檔案節點NUMA組態。



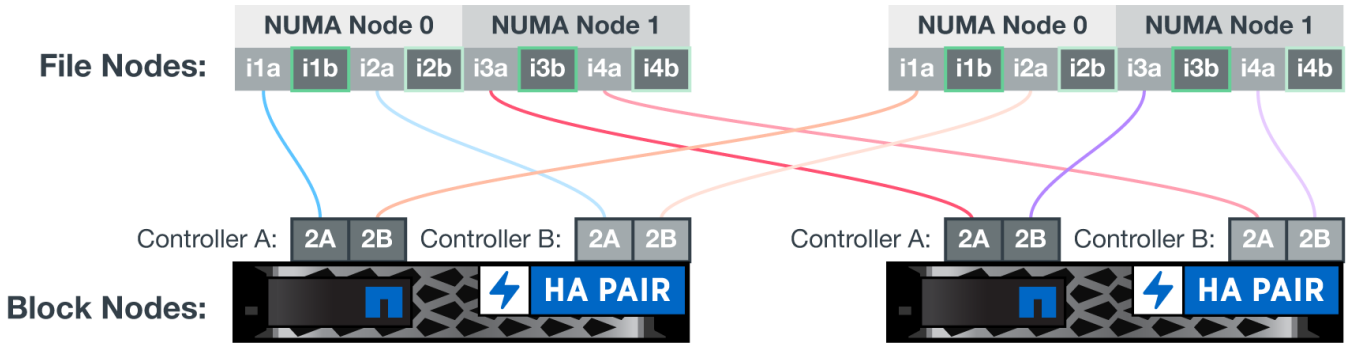
BeeGFS程序會固定在特定的NUMA區域、以確保所使用的介面位於相同的區域。此組態可避免透過插槽間連線進行遠端存取。插槽之間的連線有時稱為QPI或GMI2連結、即使是在現代化的處理器架構中、也可能是使用高速度網路（例如HDRInfiniBand）時的瓶頸。

網路纜線組態

在建置區塊中、每個檔案節點都會使用總共四個備援InfiniBand連線、連接至兩個區塊節點。此外、每個檔案節點都有四個與InfiniBand儲存網路的備援連線。

在下圖中、請注意：

- 所有以綠色顯示的檔案節點連接埠均用於連接至儲存架構；所有其他的檔案節點連接埠則是直接連接至區塊節點。
- 特定NUMA區域中的兩個InfiniBand連接埠會連接到同一個區塊節點的A和B控制器。
- NUMA節點0中的連接埠一律連線至第一個區塊節點。
- NUMA節點1中的連接埠會連線至第二個區塊節點。



當使用分離器纜線將儲存交換器連接至檔案節點時、一條纜線應分出並連接至淡綠色的連接埠。另一條纜線應分出並連接至暗綠色的連接埠。此外、對於具有備援交換器的儲存網路、淡綠色的連接埠應連接至一台交換器、而深綠色的連接埠則應連接至另一台交換器。

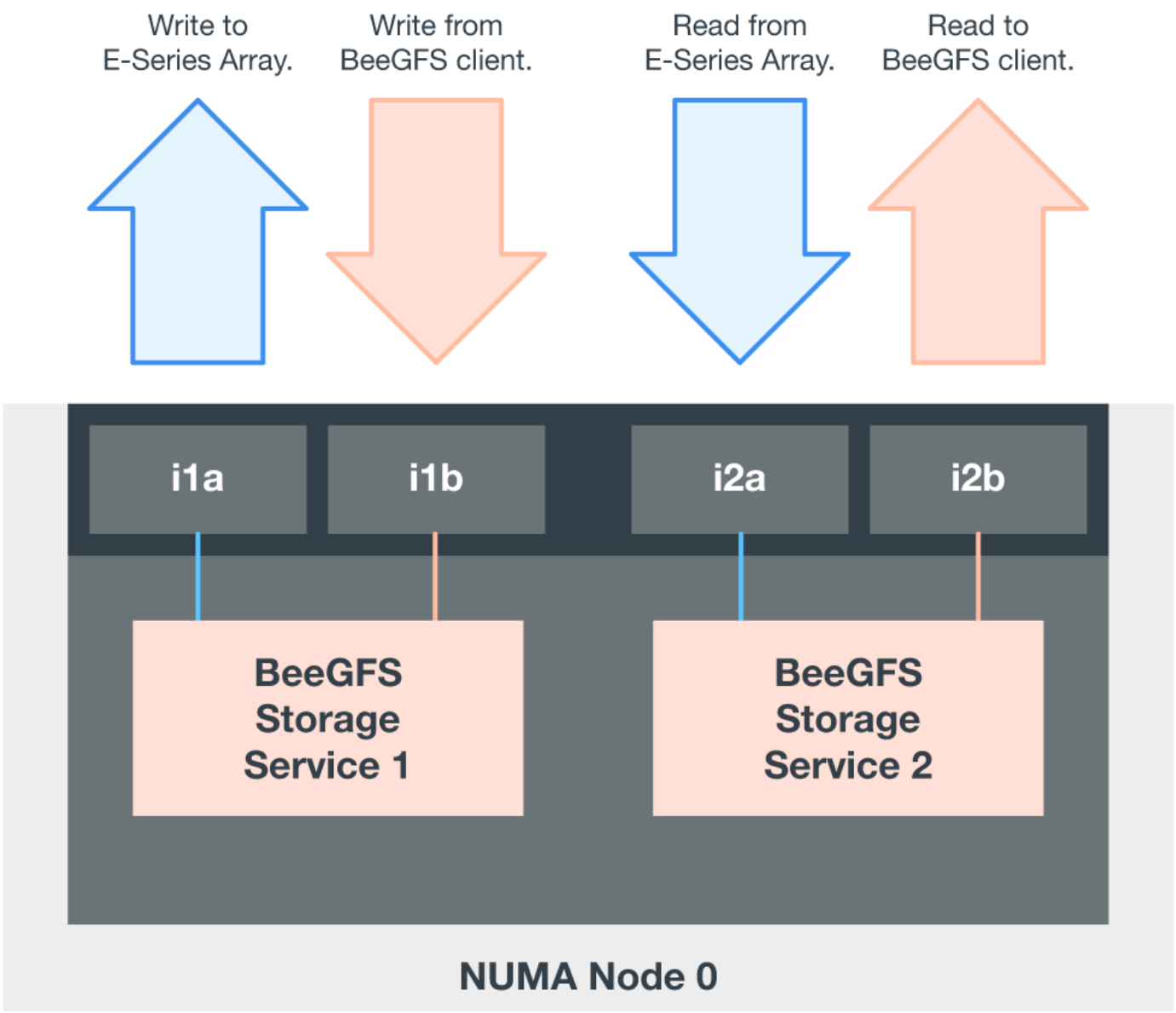
圖中所示的佈線組態可讓每個BeeGFS服務：

- 無論執行BeeGFS服務的檔案節點為何、都可在相同的NUMA區域中執行。
- 無論故障發生在何處、都要有次要的最佳路徑可通往前端儲存網路和後端區塊節點。
- 如果區塊節點中的檔案節點或控制器需要維護、請將效能影響降至最低。

利用頻寬的纜線

若要充分運用PCIe雙向頻寬、請確定每個InfiniBand介面卡上的一個連接埠連接至儲存架構、另一個連接埠則連接至區塊節點。

下圖顯示用於充分運用PCIe雙向頻寬的纜線設計。



對於每個BeeGFS服務、請使用相同的介面卡、將用戶端流量所使用的慣用連接埠、與服務磁碟區的主要擁有者區塊節點控制器路徑連線。如需詳細資訊、請參閱 "軟體組態"。

軟體組態

NetApp上BeeGFS的軟體組態包括BeeGFS網路元件、EF600區塊節點、BeeGFS檔案節點、資源群組和BeeGFS服務。

BeeGFS網路組態

BeeGFS網路組態包含下列元件。

- *浮動IP*浮動IP是一種虛擬IP位址、可動態路由傳送至同一個網路中的任何伺服器。多部伺服器可以擁有相同的浮動IP位址、但在任何指定時間、只能在一部伺服器上啟用。

  每個BeeGFS伺服器服務都有自己的IP位址、可視BeeGFS伺服器服務的執行位置而在檔案節點之間移動。此浮動IP組態可讓每個服務獨立容錯移轉至其他檔案節點。用戶端只需知道特定BeeGFS服務的IP位址、就

不需要知道目前執行該服務的檔案節點。

- * BeeGFS伺服器多重主頁組態*為了提高解決方案的密度、每個檔案節點都有多個儲存介面、其中IP設定在同一個IP子網路中。

  需要額外的組態、以確保此組態能與Linux網路堆疊正常運作、因為在預設情況下、如果某個介面的IP位在同一子網路中、則可在不同的介面上回應對該介面的要求。除了其他缺點、這種預設行為也使得無法正確建立或維護RDMA連線。

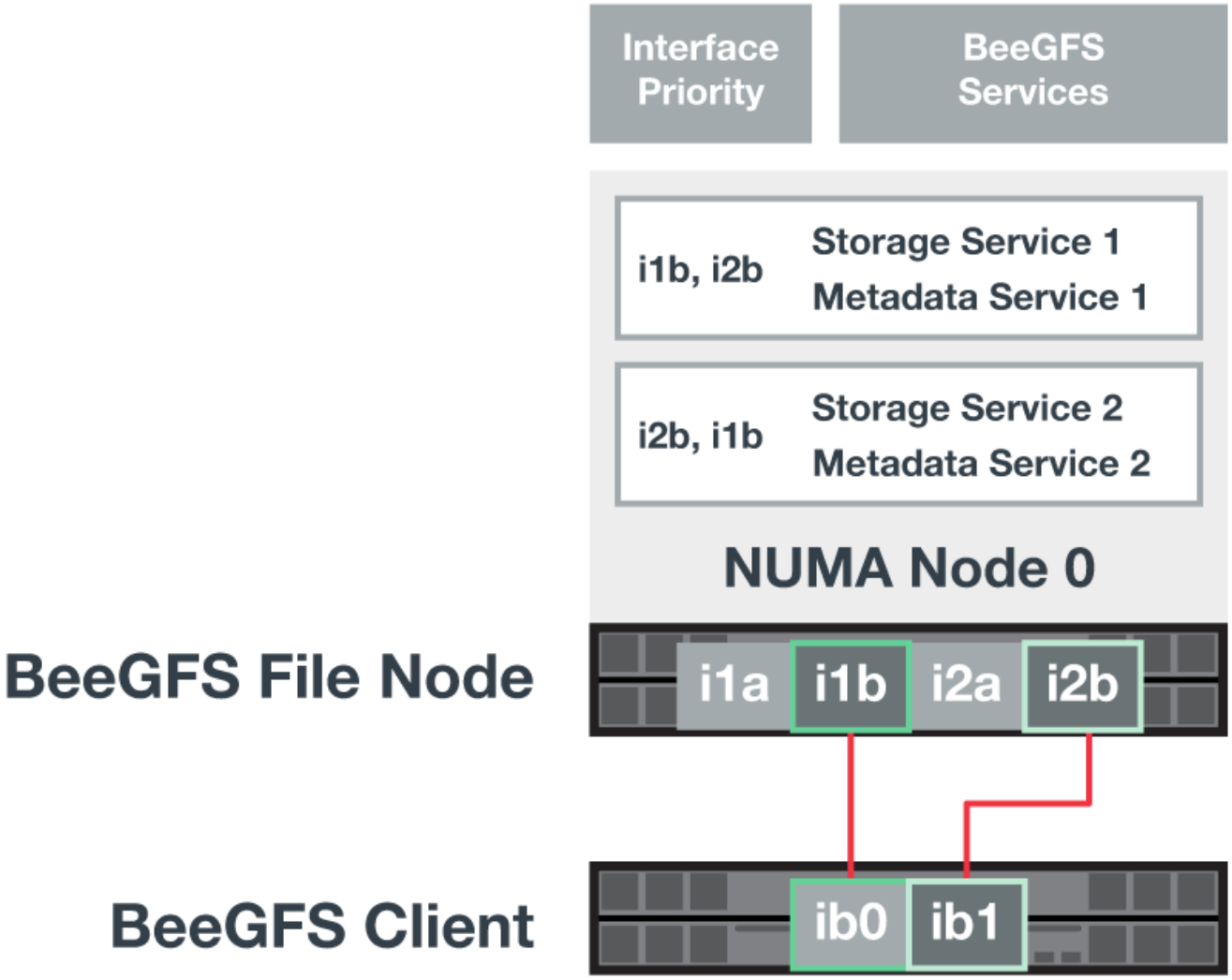
  Ansible型部署可處理反向路徑（RP）和位址解析傳輸協定（Arp）行為的強化、同時確保啟動和停止浮動IP；動態建立對應的IP路由和規則、讓多重主目錄網路組態正常運作。

- BeeGFS 用戶端多軌組態 * _Multi-rail 是指應用程式使用多個不同網路連線（或「 rail 」）來提高效能的能力。

  BeeGFS 實作多軌支援、可在單一 IPoIB 子網路中使用多個 IB 介面。此功能可在 RDMA NIC 之間啟用動態負載平衡等功能、以最佳化網路資源的使用。它也與 NVIDIA GPUDirect 儲存設備（ GDS ）整合、可提供更高的系統頻寬、並減少用戶端 CPU 的延遲和使用率。

  本文件提供單一 IPoIB 子網路組態的說明。支援雙 IPoIB 子網路組態、但並未提供與單一子網路組態相同的優勢。

  下圖顯示多個BeeGFS用戶端介面之間的流量平衡。



由於BeeGFS中的每個檔案通常會跨越多個儲存服務進行等量分佈、因此多重軌道組態可讓用戶端達到比單一InfiniBand連接埠更高的處理量。例如、下列程式碼範例顯示通用的檔案分段組態、可讓用戶端在兩個介面之間平衡流量：

+

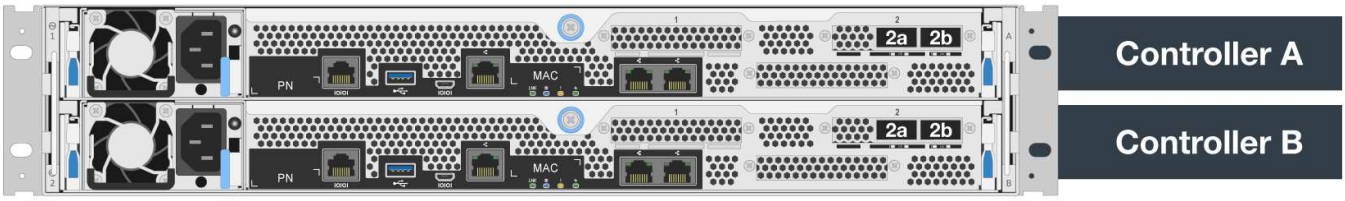
```
root@beegfs01:/mnt/beegfs# beegfs-ctl --getentryinfo myfile
Entry type: file
EntryID: 11D-624759A9-65
Metadata node: meta_01_tgt_0101 [ID: 101]
Stripe pattern details:
+ Type: RAID0
+ Chunksize: 1M
+ Number of storage targets: desired: 4; actual: 4
+ Storage targets:
  + 101 @ stor_01_tgt_0101 [ID: 101]
  + 102 @ stor_01_tgt_0101 [ID: 101]
  + 201 @ stor_02_tgt_0201 [ID: 201]
  + 202 @ stor_02_tgt_0201 [ID: 201]
```

EF600區塊節點組態

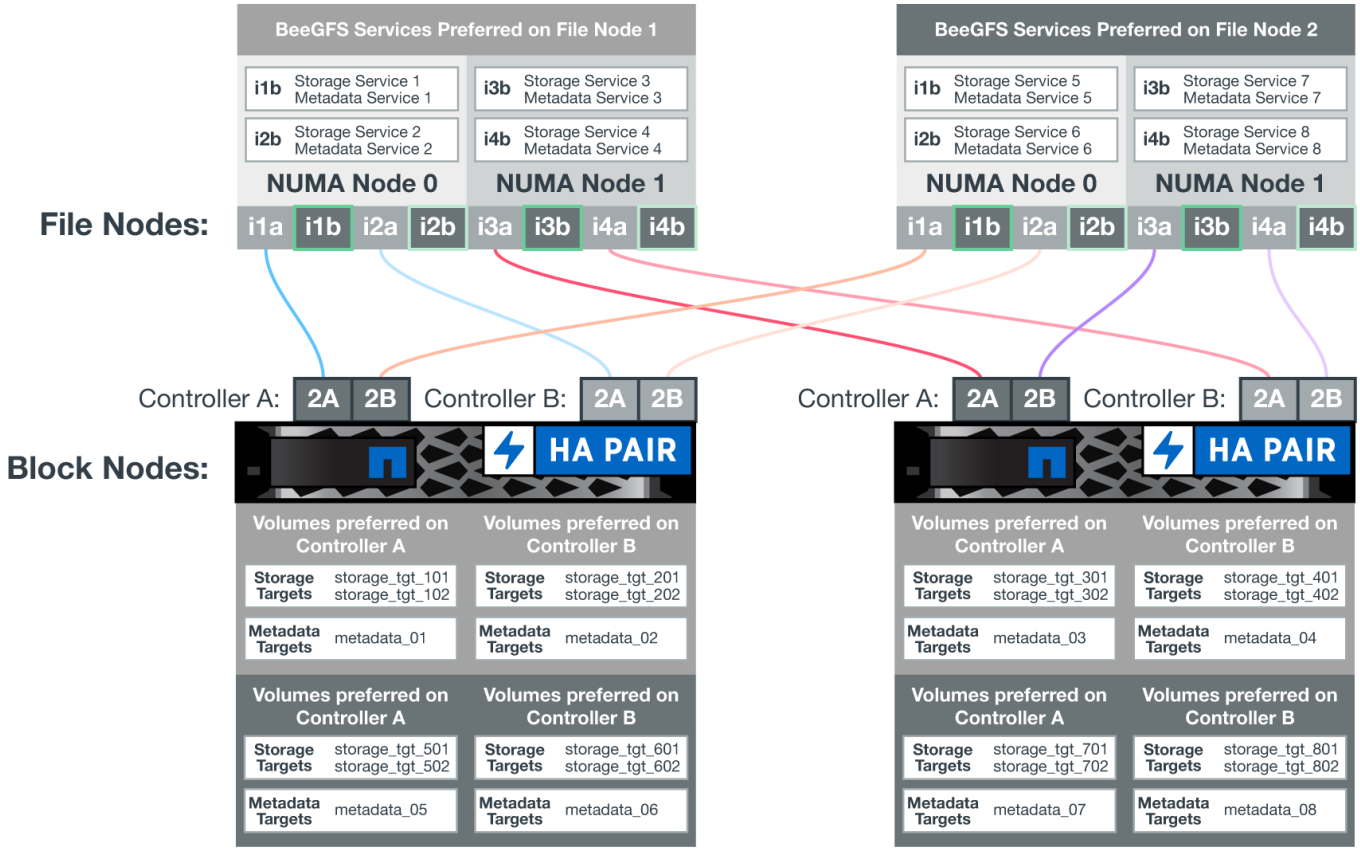
區塊節點由兩個主動/主動式RAID控制器組成、可共用存取同一組磁碟機。一般而言、每個控制器擁有系統上設定的一半磁碟區、但可視需要接管其他控制器。

檔案節點上的多重路徑軟體可決定每個磁碟區的作用中最佳化路徑、並在纜線、介面卡或控制器故障時自動移至替代路徑。

下圖顯示EF600區塊節點中的控制器配置。



為了簡化共享磁碟HA解決方案、磁碟區會對應至兩個檔案節點、以便視需要彼此接管。下圖顯示如何設定BeeGFS服務和慣用磁碟區擁有權以達到最大效能的範例。每個BeeGFS服務左側的介面會指出用戶端和其他服務用來與其聯絡的偏好介面。



在前一個範例中、用戶端和伺服器服務偏好使用介面i1b與儲存服務1通訊。儲存服務1使用介面i1a做為首選路徑、以便在第一個區塊節點的控制器A上與其磁碟區（儲存設備_tgt_101、102）進行通訊。此組態可利用InfiniBand介面卡可用的全雙向PCIe頻寬、並從雙埠的HDRInfiniBand介面卡獲得比PCIe 4.0更好的效能。

BeeGFS檔案節點組態

BeeGFS檔案節點已設定為高可用度（HA）叢集、以便在多個檔案節點之間進行BeeGFS服務的容錯移轉。

HA叢集設計是以兩個廣泛使用的Linux HA專案為基礎：叢集成員資格的電暈器同步、以及叢集資源管理的起搏器。如需更多資訊、請參閱 "適用於高可用度附加元件的Red Hat訓練"。

NetApp撰寫並擴充數個開放式叢集架構（OCF）資源代理程式、讓叢集能夠智慧地啟動及監控BeeGFS資源。

BeeGFS HA叢集

一般而言、當您啟動BeeGFS服務（無論是否有HA）時、必須有幾個資源：

- 可連線服務的IP位址、通常由Network Manager設定。
- 作為BeeGFS儲存資料目標的基礎檔案系統。

  這些通常是在/etc/stab'中定義的、並由systemd掛載。

- 負責在其他資源準備就緒時啟動BeeGFS的系統服務。

  如果沒有其他軟體、這些資源只會在單一檔案節點上啟動。因此、如果檔案節點離線、則無法存取BeeGFS檔案系統的一部分。

由於多個節點可以啟動每個BeeGFS服務、因此心臟起搏器必須確保每個服務和相依資源一次只能在一個節點上執行。例如、如果兩個節點嘗試啟動相同的BeeGFS服務、則如果兩個節點都嘗試寫入基礎目標上的相同檔案、就會有資料毀損的風險。為了避免這種情況、心臟起搏器必須仰賴電暈器同步、才能在所有節點之間可靠地保持整體叢集的狀態同步、並建立仲裁。

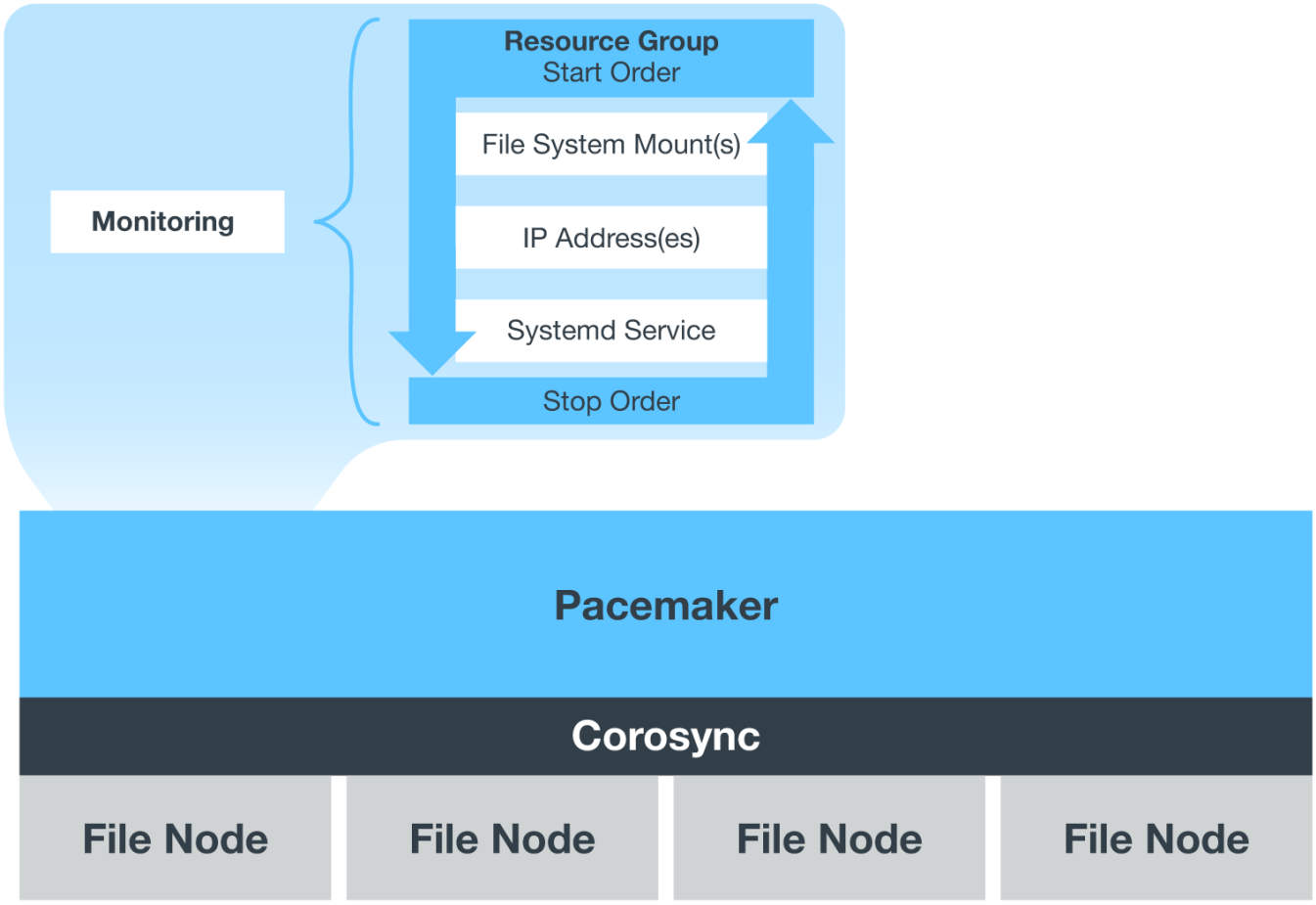
如果叢集發生故障、心臟起搏器會在另一個節點上反應並重新啟動BeeGFS資源。在某些情況下、心臟起搏器可能無法與原始故障節點通訊、以確認資源已停止。若要在重新啟動BeeGFS資源之前驗證節點是否已關閉、請先移除電源、使心臟起搏器從故障節點上關閉。

許多開放原始碼的屏障代理程式可讓心臟起搏器使用電力分配單元（PDU）或伺服器基板管理控制器（BMC）搭配API（例如RedfISH）來隔離節點。

當BeeGFS在HA叢集中執行時、所有BeeGFS服務和基礎資源都是由資源群組中的心臟起搏器管理。每個BeeGFS服務及其所依賴的資源都會設定成資源群組、以確保資源以正確的順序啟動和停止、並配置在同一個節點上。

對於每個BeeGFS資源群組、心臟起搏器都會執行自訂BeeGFS監控資源、負責偵測故障情況、並在特定節點上無法存取BeeGFS服務時、以智慧方式觸發容錯移轉。

下圖顯示由心臟起搏器控制的BeeGFS服務和相依性。



為了在同一個節點上啟動多個相同類型的BeeGFS服務、心臟起搏器已設定為使用多重模式組態方法來啟動BeeGFS服務。如需詳細資訊、請參閱 "多重模式的BeeGFS文件"。

由於BeeGFS服務必須能夠在多個節點上啟動、因此每項服務的組態檔（通常位於「/etc/beegfs」）會儲存在其中一個E系列磁碟區上、作為該服務的BeeGFS目標。如此一來、可能需要執行服務的所有節點都能存取特定BeeGFS服務的組態和資料。

```
# tree stor_01_tgt_0101/ -L 2
stor_01_tgt_0101/
├── data
│   ├── benchmark
│   ├── buddymir
│   ├── chunks
│   ├── format.conf
│   ├── lock.pid
│   ├── nodeID
│   ├── nodeNumID
│   ├── originalNodeID
│   ├── targetID
│   └── targetNumID
└── storage_config
    ├── beegfs-storage.conf
    ├── connInterfacesFile.conf
    └── connNetFilterFile.conf
```

設計驗證

NetApp解決方案BeeGFS的第二代設計已使用三種建置區塊組態設定檔進行驗證。

組態設定檔包括下列項目：

- 單一基礎建置區塊、包括BeeGFS管理、中繼資料和儲存服務。
- BeeGFS中繼資料加上儲存建置區塊。
- BeeGFS純儲存建置區塊。

建置區塊連接至兩台 NVIDIA Quantum InfiniBand （ MQM8700 ） 交換器。十個BeeGFS用戶端也連接到InfiniBand交換器、用來執行綜合基準測試公用程式。

下圖顯示用於驗證NetApp解決方案BeeGFS的BeeGFS組態。

BeeGFS檔案分段

平行檔案系統的一項優點是能夠跨越多個儲存目標、將個別檔案等量磁碟區、這可能代表相同或不同基礎儲存系統上的磁碟區。

在BeeGFS中、您可以根據每個目錄和每個檔案來設定分段、以控制用於每個檔案的目標數量、並控制用於每個檔案分段的chunksize（或區塊大小）。此組態可讓檔案系統支援不同類型的工作負載和I/O設定檔、而不需要重新設定或重新啟動服務。您可以使用「beegfs-CTL」命令列工具或使用分段API的應用程式來套用等量磁碟區設定。如需詳細資訊、請參閱的BeeGFS文件 "分段" 和 "分段API"。

為了達到最佳效能、在整個測試過程中都會調整等量磁碟區模式、並記錄每項測試所使用的參數。

IOR頻寬測試：多個用戶端

IOR頻寬測試使用OpenMPI來執行綜合I/O產生器工具IOR的平行工作（可從以下網站取得 "HPC GitHub"）跨所有10個用戶端節點、移至一或多個BeeGFS建置區塊。除非另有說明：

- 所有測試均使用直接I/O、傳輸大小為1MiB。
- BeeGFS檔案分段設定為1MB chunksize、每個檔案一個目標。

下列參數用於IOR、區段數經過調整、可將一個建置區塊的Aggregate檔案大小維持在5TiB、三個建置區塊的區段數維持在40TiB。

```
mpirun --allow-run-as-root --mca btl tcp -np 48 -map-by node -hostfile
10xnodes ior -b 1024k --posix.odirect -e -t 1024k -s 54613 -z -C -F -E -k
```

一個**BeeGFS**基礎（管理、中繼資料和儲存）建置區塊

下圖顯示單一BeeGFS基礎（管理、中繼資料和儲存）建置區塊的IOR測試結果。

BeeGFS中繼資料**+**儲存建置區塊

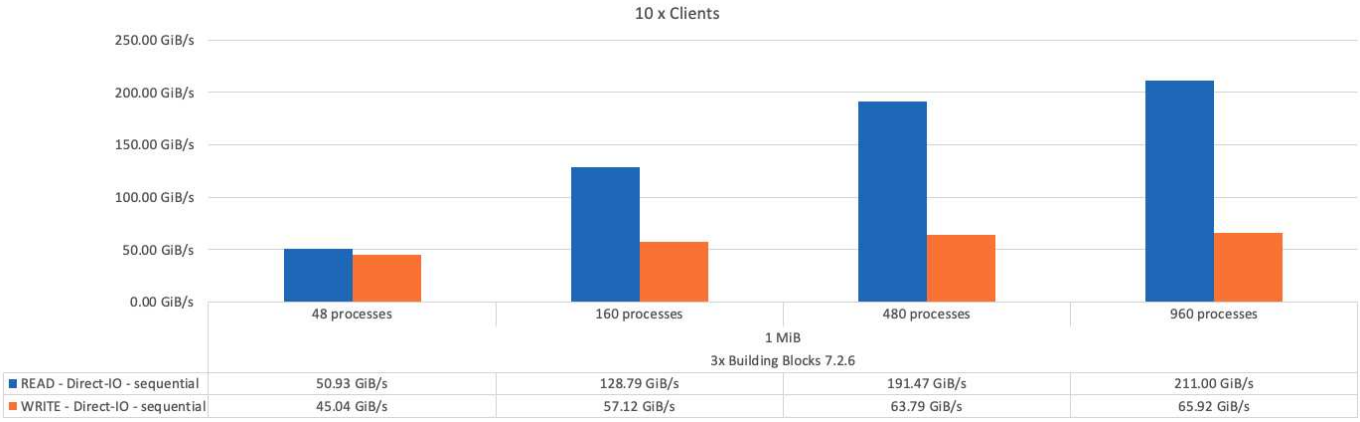
下圖顯示單一BeeGFS中繼資料+儲存建置區塊的IOR測試結果。

BeeGFS純儲存建置區塊

下圖顯示單一BeeGFS純儲存建置區塊的IOR測試結果。

三個**BeeGFS**建置區塊

下圖顯示使用三個BeeGFS建置區塊的IOR測試結果。



如預期、基礎建置區塊與後續中繼資料+儲存建置區塊之間的效能差異可忽略不計。比較中繼資料+儲存建置區塊與純儲存建置區塊、可看出讀取效能略有提升、因為使用額外的磁碟機做為儲存目標。不過、寫入效能並無顯著差異。若要達到更高的效能、您可以將多個建置區塊一起新增、以線性方式擴充效能。

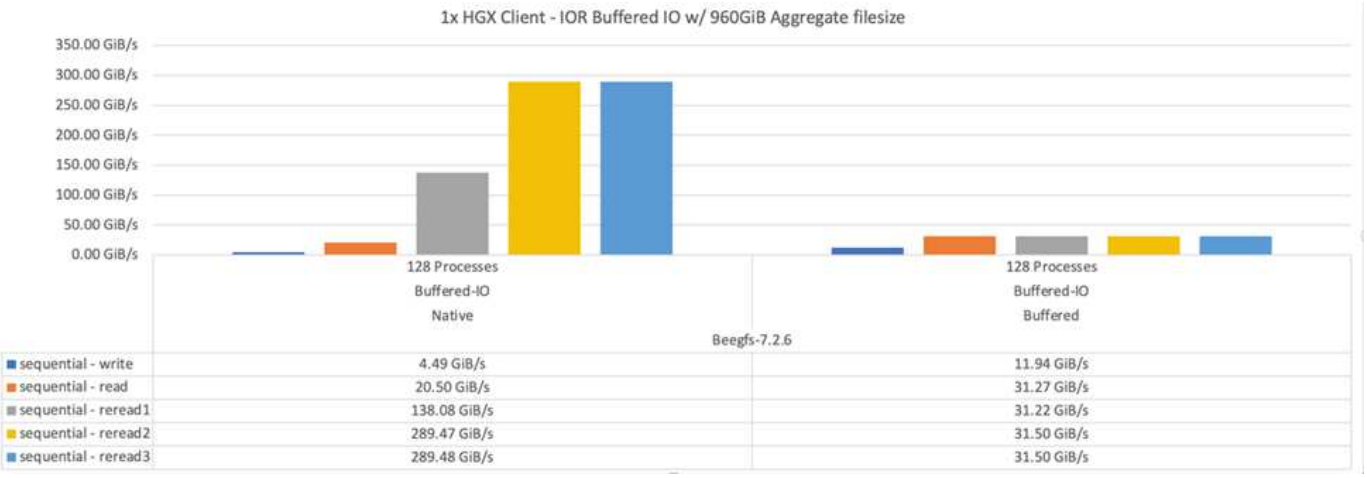
IOR頻寬測試：單一用戶端

IOR頻寬測試使用OpenMPI、使用單一高效能GPU伺服器執行多個IOR程序、以探索單一用戶端所能達到的效能。

此測試也會比較BeeGFS在用戶端設定為使用Linux核心分頁快取（「tuneFileCacheType = Native」）時的重新讀取行為和效能、以及預設的「緩衝」設定。

原生快取模式會使用用戶端上的Linux核心分頁快取、讓重新讀取作業從本機記憶體產生、而非透過網路重新傳輸。

下圖顯示使用三個BeeGFS建置區塊和單一用戶端的IOR測試結果。



這些測試的BeeGFS分段設定為1MB chunksize、每個檔案有八個目標。

雖然使用預設的緩衝模式時、寫入和初始讀取效能較高、但對於重讀相同資料多次的工作負載、原生快取模式可大幅提升效能。這項改善的重新讀取效能對於深度學習等工作負載來說非常重要、因為深度學習會在許多時期重讀相同的資料集多次。

中繼資料效能測試

中繼資料效能測試使用MDTest工具（包含在IOR中）來測量BeeGFS的中繼資料效能。測試使用OpenMPI在所有十個用戶端節點上執行平行工作。

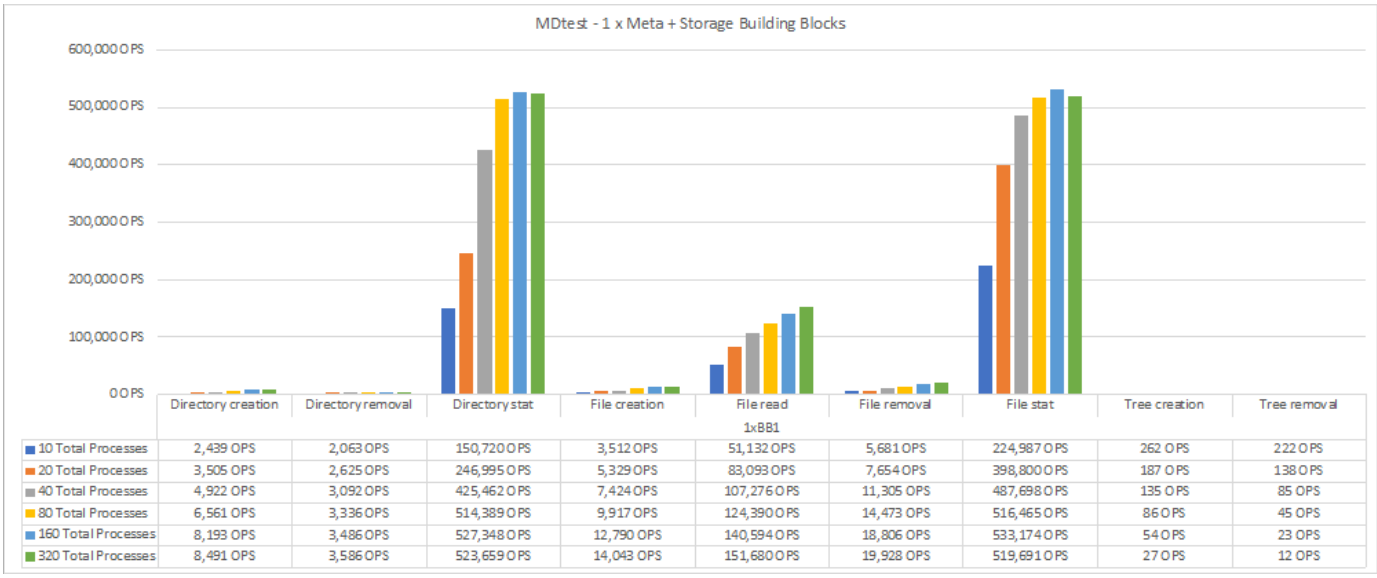
下列參數用於執行基準測試、其處理程序總數從10個增加到320個、步驟2個、檔案大小為4K。

```
mpirun -h 10xnodes –map-by node np $processes mdtest -e 4k -w 4k -i 3 -I
16 -z 3 -b 8 -u
```

中繼資料效能是先以一到兩個中繼資料+儲存建置區塊來測量、藉由新增額外的建置區塊來顯示效能如何擴充。

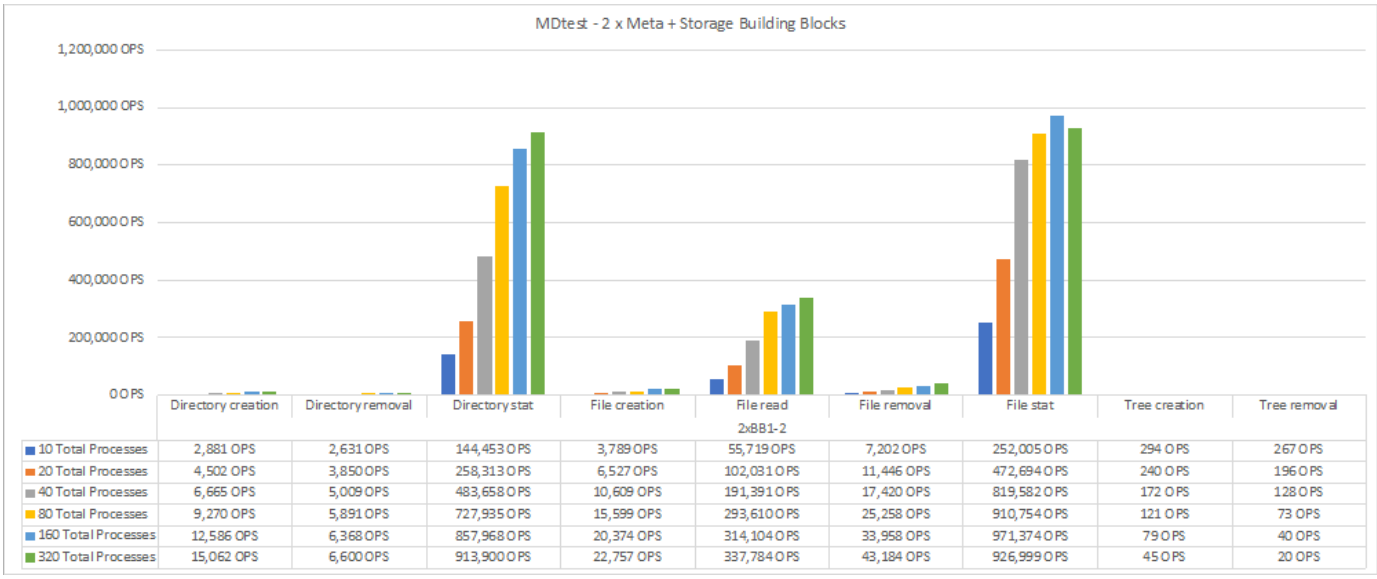
一個**BeeGFS**中繼資料**+**儲存建置區塊

下圖顯示含有一個BeeGFS中繼資料+儲存建置區塊的MDTest結果。



兩個**BeeGFS**中繼資料**+**儲存建置區塊

下圖顯示含有兩個BeeGFS中繼資料+儲存建置區塊的MDTest結果。



	Directory creation	Directory removal	Directory stat	File creation	File read	File removal	File stat	Tree creation	Tree removal
					2xBB1-2				
10 Total Processes	2,881 OPS	2,631 OPS	144,453 OPS	3,789 OPS	55,719 OPS	7,202 OPS	252,005 OPS	294 OPS	267 OPS
20 Total Processes	4,502 OPS	3,850 OPS	258,313 OPS	6,527 OPS	102,031 OPS	11,446 OPS	472,694 OPS	240 OPS	196 OPS
40 Total Processes	6,665 OPS	5,009 OPS	483,658 OPS	10,609 OPS	191,391 OPS	17,420 OPS	819,582 OPS	172 OPS	128 OPS
80 Total Processes	9,270 OPS	5,891 OPS	727,935 OPS	15,599 OPS	293,610 OPS	25,258 OPS	910,754 OPS	121 OPS	73 OPS
160 Total Processes	12,586 OPS	6,368 OPS	857,968 OPS	20,374 OPS	314,104 OPS	33,958 OPS	971,374 OPS	79 OPS	40 OPS
320 Total Processes	15,062 OPS	6,600 OPS	913,900 OPS	22,757 OPS	337,784 OPS	43,184 OPS	926,999 OPS	45 OPS	20 OPS

功能驗證

在驗證此架構時、NetApp執行了數項功能測試、包括：

- 停用交換器連接埠、使單一用戶端InfiniBand連接埠故障。
- 停用交換器連接埠、使單一伺服器InfiniBand連接埠故障。
- 使用BMC觸發立即關閉伺服器電源。
- 將節點正常置於待命狀態、並將故障切換服務移轉至其他節點。
- 正常地將節點重新連線、並將服務容錯回復至原始節點。
- 使用PDU關閉其中一個InfiniBand交換器。所有測試都是在壓力測試進行期間執行、並在BeeGFS用戶端上設定「SysSessionChecksEnabled:假」參數。未發現I/O錯誤或中斷。

有已知問題（請參閱 "Changelog"）當BeeGFS用戶端/伺服器RDMA連線意外中斷時、可能是因為主要介面遺失（如「connInterfacesFile」中所定義）、或是BeeGFS伺服器故障；作用中用戶端I/O在恢復前最多可掛斷10分鐘。若BeeGFS節點在規劃維護時正常放置在待命或使用TCP、則不會發生此問題。

NVIDIA DGX SuperPOD 和 BasePOD 驗證

NetApp已使用類似的BeeGFS檔案系統（由三個建置區塊組成、並套用中繼資料加上儲存組態設定檔）、驗證NVIDIAs DGX A100 SupermPOD的儲存解決方案。此NVA所描述的解決方案、需要測試資格、測試20部DGX A100 GPU伺服器、執行各種儲存設備、機器學習和深度學習基準測試。以 NVIDIA DGX A100 SuperPOD 所建立的驗證為基礎、 NetApp 上的 BeeGFS 解決方案已獲得 DGX SuperPOD H100 、 H200 及 B200 系統的核准。這項延伸是根據 NVIDIA DGX A100 所驗證的先前基準測試和系統需求而定。

如需詳細資訊、請參閱 "NVIDIA DGX超級POD與NetApp合作" 和 "NVIDIA DGX基礎POD"。

規模調整準則

BeeGFS解決方案包含根據驗證測試來調整效能和容量規模的建議。

建置區塊架構的目標、是透過新增多個建置區塊來建立易於調整規模的解決方案、以符合特定BeeGFS系統的需求。根據以下準則、您可以預估BeeGFS建置區塊的數量和類型、以符合您環境的需求。

請記住、這些預估是最佳的效能表現。綜合基準測試應用程式是以實際應用程式可能無法使用的方式來撰寫及使用、以最佳化基礎檔案系統的使用。

效能規模調整

下表提供建議的效能規模調整。

組態設定檔	**1MiB**讀取	**1MiB**寫入
中繼資料+儲存設備	62GiBps	21GiBps
僅儲存設備	64GiBps	21GiBps

中繼資料容量規模預估是根據「經驗法則」、在BeeGFS中、500 GB的容量足以容納約1.5億個檔案。（如需詳細資訊、請參閱BeeGFS文件 "系統需求"）

使用存取控制清單等功能、以及每個目錄的目錄和檔案數量、也會影響中繼資料空間的使用速度。儲存容量預估會考慮可用磁碟機容量、以及RAID 6和XFS負荷。

中繼資料+儲存建置區塊的容量規模

下表提供中繼資料與儲存建置區塊的建議容量規模調整。

磁碟機大小（**2+2 RAID 1**）中繼資料**Volume**群組	中繼資料容量（檔案數）	磁碟機大小（**8+2 RAID 6**）儲存**Volume**群組	儲存容量（檔案內容）
1.92TB	1,938,577,200	1.92TB	51.77TB
3.84 TB	3,880,388,400	3.84 TB	103.55TB
7.68TB	8、125、278、000	7.68TB	216.74 TB
15.3TB	17、269、854000	15.3TB	460.60TB

調整中繼資料加上儲存建置區塊規模時、您可以使用較小的磁碟機來進行中繼資料磁碟區群組、而非儲存磁碟區群組、藉此降低成本。

專為儲存設備建置區塊調整容量

下表針對純儲存建置區塊提供經驗法則容量規模調整。

磁碟機大小（**10+2 RAID 6**）儲存**Volume**群組	儲存容量（檔案內容）
1.92TB	59.89TB
3.84 TB	119.80TB
7.68TB	251.89TB
15.3TB	538.55TB

除非啟用全域檔案鎖定、否則在基礎（第一）建置區塊中納入管理服務的效能和容量負荷最小。

效能調校

BeeGFS解決方案包含根據驗證測試進行效能調校的建議。

儘管BeeGFS提供合理的開箱即用效能、但NetApp已開發出一套建議的調校參數、以最大化效能。這些參數會考量基礎E系列區塊節點的功能、以及在共享磁碟HA架構中執行BeeGFS所需的任何特殊需求。

檔案節點的效能調校

您可以設定的可用調校參數包括：

1. *檔案節點的UEFI/BIOS中的系統設定。*若要發揮最大效能、建議您在做為檔案節點的伺服器機型上設定系統設定。您可以使用系統設定程式（UEFI/BIOS）或底板管理控制器（BMC）提供的Redfish API來設定檔案節點時的系統設定。

   系統設定會因您用來做為檔案節點的伺服器機型而有所不同。這些設定必須根據使用中的伺服器機型手動設定。若要了解如何配置已驗證的 Lenovo SR665 V3 檔案節點的系統設置，請參閱"調整檔案節點系統設定以獲得效能" 。

2. *必要組態參數的預設設定。*必要的組態參數會影響BeeGFS服務的設定方式、以及E系列磁碟區（區塊裝置）如何由心臟起搏器設定格式及掛載。這些必要的組態參數包括：

   - BeeGFS服務組態參數

     您可以視需要覆寫組態參數的預設設定。如需可針對特定工作負載或使用案例進行調整的參數，請參閱 "BeeGFS服務組態參數"。

   - Volume格式化和掛載參數會設定為建議的預設值、而且只能針對進階使用案例進行調整。預設值會執行下列動作：

     - 根據目標類型（例如管理、中繼資料或儲存設備）、以及基礎磁碟區的RAID組態和區段大小、最佳化初始磁碟區格式。
     - 調整心臟起搏器如何掛載每個Volume、以確保變更立即排清至E系列區塊節點。如此可在檔案節點發生故障且正在進行作用中寫入時、避免資料遺失。

       如需可針對特定工作負載或使用案例進行調整的參數，請參閱 "Volume格式化與掛載組態參數"。

3. * 安裝在檔案節點上的 Linux 作業系統中的系統設定。 *當您在的步驟 4 中建立 Ansible 庫存時、您可以覆寫預設的 Linux 作業系統設定 "建立可Ansible庫存"。

   預設設定是用來驗證NetApp解決方案上的BeeGFS、但您可以變更這些設定、以因應您的特定工作負載或使用案例進行調整。您可以變更的一些Linux作業系統設定範例包括：

   - E系列區塊裝置上的I/O佇列。

     您可以在作為BeeGFS目標的E系列區塊裝置上設定I/O佇列、以便：

     - 根據裝置類型（NVMe、HDD等）調整排程演算法。
     - 增加未處理要求的數量。

- 調整要求大小。
- 最佳化預先讀取行為。

◦ 虛擬記憶體設定：

您可以調整虛擬記憶體設定、以獲得最佳的持續串流效能。

◦ CPU設定：

您可以調整CPU頻率調節器和其他CPU組態、以獲得最大效能。

◦ 讀取要求大小。

您可以增加 NVIDIA HCA 的讀取要求大小上限。

區塊節點的效能調校

根據套用至特定BeeGFS建置區塊的組態設定檔、區塊節點上設定的Volume群組會稍微變更。例如、使用24個磁碟機EF600區塊節點：

- 對於單一基礎建置區塊、包括BeeGFS管理、中繼資料和儲存服務：
  - 1個2+2個RAID 10 Volume群組、用於BeeGFS管理和中繼資料服務
  - 2個8+2個RAID 6 Volume群組用於BeeGFS儲存服務
- 若為BeeGFS中繼資料+儲存建置區塊：
  - 1個2+2個RAID 10 Volume群組、用於BeeGFS中繼資料服務
  - 2個8+2個RAID 6 Volume群組用於BeeGFS儲存服務
- 僅適用於BeeGFS儲存設備建置區塊：
  - 2個10+2個RAID 6 Volume群組用於BeeGFS儲存服務

> 由於BeeGFS需要的管理與中繼資料儲存空間比儲存空間大幅減少、因此有一個選項是針對RAID 10 Volume群組使用較小的磁碟機。較小的磁碟機應安裝在最外側的磁碟機插槽中。如需詳細資訊、請參閱 "部署指示"。

這些都是由Ansible型部署所設定、以及其他一些一般建議的設定、以最佳化效能/行為、包括：

- 將全域快取區塊大小調整為32KiB、並將需求型快取排清調整為80%。
- 停用自動負載平衡（確保控制器磁碟區指派維持原定狀態）。
- 啟用讀取快取和停用預先讀取快取。
- 啟用含鏡射的寫入快取、並需要電池備份、以便在區塊節點控制器故障時、快取仍會持續存在。
- 指定磁碟機指派給磁碟區群組的順序、在可用磁碟機通道之間平衡I/O。

大容量建置區塊

標準BeeGFS解決方案設計以高效能工作負載為設計考量。尋求高容量使用案例的客戶應

觀察此處概述的設計與效能特性差異。

硬體與軟體組態

高容量建置區塊的硬體和軟體組態為標準配置、但EF600控制器應更換為EF300控制器、並可選擇連接1到7個IOM擴充支架、每個儲存陣列各有60個磁碟機、每個建置區塊總計2至14個擴充托盤。

部署高容量建置區塊設計的客戶、可能只會使用由BeeGFS管理、中繼資料及每個節點儲存服務所組成的基礎建置區塊樣式組態。為了節省成本、大容量儲存節點應在EF300控制器機箱的NVMe磁碟上配置中繼資料磁碟區、並應將儲存磁碟區配置至擴充托盤中的NL-SAS磁碟機。

[]

規模調整準則

這些規模調整準則假設大容量建置區塊在基礎EF300機箱中設定一個2+2 NVMe SSD Volume群組作為中繼資料、並在每個IOM擴充匣中設定6x 8+2 NL-SAS Volume群組作為儲存設備。

磁碟機大小（容量**HDD**）	每個寬板的容量（**1**個紙匣）	每個寬帶容量（**2**個磁碟匣）	每個寬帶容量（**3**個磁碟匣）	每個寬帶容量（**4**個磁碟匣）
4TB	439TB	878 TB	1317 TB	1756 TB
8 TB	878 TB	1756 TB	2634 TB	3512 TB
10 TB	1097 TB	2195 TB	3292 TB	4390 TB
12 TB	1317 TB	2634 TB	3951 TB	5268TB
16 TB	1756 TB	3512 TB	5268TB	7024 TB
18 TB	1975 TB	3951 TB	5927 TB	7902 TB