白皮書
2009 年 10 月

利用 IBM DB2 pureScale 實現應用程式的透明擴充

Information Management

DB2 software

內容

簡介

在經濟復甦的過程中，對於核心業務資料的立即存取始終是企業賴以生存，甚至是獲得成功的關鍵因素。隨著越來越多的金錢以各種方式流入經濟體系，企業必須靈活多變、具備高可用度與可適應性的基礎設施，才能把握恢復成長的契機。

大多數分散式軟體公司在行銷時，都會採用將可用度等級與「近似*大型主機*」或「5 – 9s」可用度等詞彙相關聯的手段。這些關鍵措辭都是要試著表達業界視為高可用度「黃金級」標準：DB2® for z/OS® 所訂下的連續可用度目標。

可用度	每年停機時間
99.999%	5 分鐘
99.99%	50 分鐘
99.9%	8 小時 20 分鐘
99%	3 天 11 小時 18 分鐘
95%	18 天 6 小時
90%	34 天 17 小時 17 分鐘
85%	54 天 18 小時

如今，可用度的定義不僅止於不受故障元件影響和恢復正常交易流程。如果您的服務層級協議(SLA)規定預期的查詢回應時間應該是在數秒鐘內，而伺服器卻過了好一會兒才傳回查詢，那這就是可用度的問題。為了確保可用度，您的系統不僅需要提供交易服務，也必須在您 SLA 中定義的時間內提供服務。

例如，如果景氣循環的季節性波動導致擴充方面的可用度問題，則真正具備可用度的架構就必須在*不變更應用程式*的前提下，以*透明*的方式新增資源，才能滿足不斷變化的效能需求。*透明*這個詞是關鍵：新增設備時，不應該讓應用程式成為叢集式的應用程式(應用程式知道哪些資料位於哪個節點上，以避免節點間出現爭用的情況)。企業無法負擔建立這些複雜的應用程式所需要的資金，因而無法進行適當的擴充。為什麼？首先，最明顯的是：叢集感知應用程式必須隨著您的資料與程式分佈量改變時而改變。叢集感知應用程式不僅需要隨著叢集發展而變更程式碼：而且這些程式碼還需要經過測試、通過品質保證(Q/A)流程、進行部署、認證等。這可能會讓整個企業花費數星期的時間努力協調，而且免不了會耗盡基礎設施中本來應該有更好用途的資源。

其他用於分散式平台(非大型主機)中交易式水平擴充資料庫產品的過時架構會為擴充帶來不必要的妨礙(例如增加的額外開支)，而導致違反 SLA。

IBM DB2 pureScale 技術(以下簡稱 DB2 pureScale)可透過結合**高可用度**與真正**透明的應用程式擴充**系統，滿足您對連續可用度當下與未來的業務需求。IBM® Power™ Systems 伺服器與 IBM 儲存設備解決方案的整合利用，進一步支援 DB2 pureScale 架構實現這種高價值解決方案。

截至目前為止，*近似大型主機*仍是一種行銷口號。但現在 DB2 pureScale 將是第一個讓分散式平台可以使用真正透明擴充架構的技術。本文將為您介紹 DB2 pureScale 技術的外觀、由來，以及如何在高可用度及透明應用程式擴充方面提供獨特的優勢。

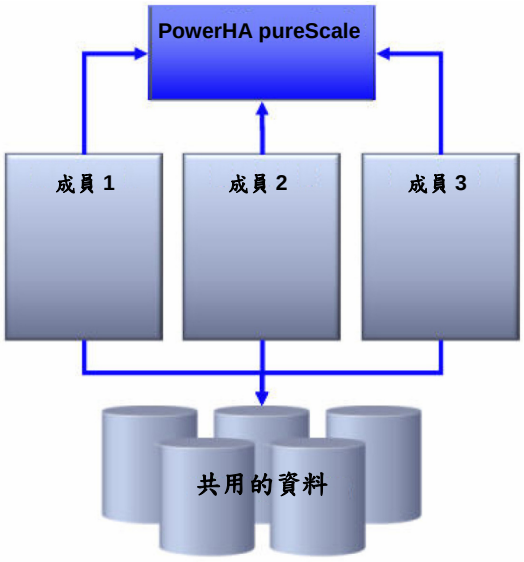
DB2 pureScale 的外觀？

DB2 pureScale 是一種全新的 DB2 選用功能，可讓您在一組「主動/主動」(active-active)組態的伺服器中水平擴充您的資料庫，以提供高水準的可用度及可擴充性。在這種組態中，在每部主機(或伺服器)中執行的 DB2 副本，可以同時讀取和寫入相同的資料。

一組共用 DB2 資料的一或多部 DB2 伺服器稱為*資料共用群組*。屬於資料共用群組的 DB2 伺服器是該群組的成員。資料共用群組的所有成員會共用相同的資料庫。目前一個資料共用群組的成員數量上限為 128。

除了 DB2 成員以外，PowerHA pureScale™元件還會提供集中式鎖定管理和資料分頁的集中式通用快取(也稱為群組緩衝集區)。

資料共用群組中的每個成員，都可以透過一個非常有效的 InfiniBand™網路直接和 PowerHA pureScale 元件互動(如下所示)，這表示每個成員都有到集中鎖定與快取設備的點對點連線。



DB2 pureScale 的由來？

您所聽到或看到*大型主機等級可用度*的描述，都是指 DB2 for z/OS 所立下的可用度黃金級標準。事實上，世界上還沒有任何一種資料庫解決方案，在可用度方面可媲美 DB2 for z/OS 的 System z®伺服器。

DB2 for z/OS 資料共用實施背後的技術，讓利用這項技術的伺服器能夠持續提供符合 SLA 規範的服務，因為具有集中式鎖定和通用快取一致性(global cache coherency)的串聯設備(Coupling Facility)能夠從故障快速復原。事實上，DB2 for z/OS 可提供真正 5-9s 級的可用度，並因為其無縫線性擴充工作負載的能力而聞名。

當您看到或聽到 DB2 for z/OS 時，您可能會想到廣泛的可擴充性和極高的可用度。這個名聲可不是行銷噱頭，而是源自這些系統在資料庫工作負載方面連續領先業界相當長的時間而聞名。或許最能證明 DB2 for z/OS 強大功能的人，是 Oracle 共同創辦人兼 CEO Larry Ellison 的評論[1]：

我嘲笑過許多其他的資料庫，實際上是除了 DB2 大型主機版本以外的所有資料庫。DB2 是一流的技術。

DB2 for z/OS 到底有什麼獨到之處，可以讓 Ellison 先生讚譽有加？DB2 for z/OS 的使用者非常熟悉其資料共用的「秘訣」：*串聯設備(Coupling Facility)*。串聯設備賦予 DB2 for z/OS 線性水平擴充的能力、提供集中式的設備來管理鎖定、扮演供修改分頁使用的通用共用緩衝集區(有助於可擴充性與可復原性的操作)等。

DB2 pureScale 的技術直接承襲自 DB2 for z/OS 串聯設備，也因此累積許多優勢，讓 DB2 for z/OS 擁有可用度及可擴充性「黃金級」標準的稱號。何以如此？因為 DB2 pureScale 所附的 powerHA pureScale 元件，可提供相同的集中式鎖定和真正的通用共用緩衝集區架構。

其他廠商已經實施共用磁碟架構的資料庫，其中最特別的是 Oracle Real Application Clusters (Oracle RAC)。不過，當這些廠商開發 Oracle RAC 時，分散式平台的技術還不允許有效存取*集中式*共用快取。因此，Oracle RAC 的設計是嘗試模擬在 DB2 for z/OS 中發現的技術，這也是 Oracle RAC 的分散式鎖定管理技術和*分散式*快取架構的起源。當 Oracle RAC 架構導入其水平擴充共用磁碟架構時，失去了 DB2 for z/OS 解決方案的簡潔價值。另一方面，DB2 for z/OS 和 DB2 pureScale 都提供相同的集中式資源管理，可以解決這些負責的可擴充性及可用度問題，我們將於本文稍後說明這點。

最基本的問題是市場上只有一種架構提供真正透明應用程式可擴充性和超高可用度。隨著在分散式平台上使用現代的硬體互連，以及透過 InfiniBand 提供深入利用無中斷遠端直接記憶體存取(RDMA)的能力，現在終於可以利用同樣可在 DB2 for z/OS 中找到的集中式鎖定和緩衝快取演算法。DB2 pureScale 是 IBM 系列產品的演進，將此業界公認的技術延伸到分散式平台。

DB2 pureScale 實現透明的應用程式可擴充性

在水平擴充的資料庫環境中，真正節省成本的關鍵，在於提供真正的透明的應用程式擴充。透明擴充表示資料庫引擎可以為 OLPT 應用程式提供更高的傳輸量和更快的反應時間，而不需要*局部性(locality)*的資料。

局部性的資料表示應用程式需要的資料位於該應用程式連線的伺服器，而且節點之間少有爭用相同資料分頁的情況。在水平擴充的架構中，如果擁有透過網路傳送大量訊息的基礎設施共用叢集中的資料，局部性的資料就顯得非常重要。

依賴局部性以實現有效擴充的水平擴充架構，需要開發人員建立複雜的交易應用程式，以讓其應用程式成為*叢集感知(cluster aware)*。叢集感知應用程式的開發與部署更為複雜且成本更高，而且叢集變更時也必須改寫應用程式。部分的廠商可能會宣稱其架構可以執行任何應用程式，而不需要進行修改，不過，如果廠商在設計時沒有利用某種形式的叢集感知，這些架構就無法***擴充***任何應用程式。

透明應用程式的擴充表示應用程式不需要具備叢集感知的特性才能利用水平擴充的架構。DB2 pureScale 是分散式平台上唯一的，其效率源自現代網路、硬體架構，以及 pureScale 集中式鎖定和快取的利用。

為了減少叢集中各節點之間通訊，以提供鎖定管理和通用快取服務，DB2 pureScale 使用 powerHA pureScale 叢集加速設備(Cluster Acceleration Facility) (以下簡稱為 *CF*)和 RDMA 技術，提供透明應用程式可擴充性。

RDMA 允許叢集中的每個成員直接存取 CF 中的記憶體，而且 CF 也可以直接存取每個成員的記憶體。例如，假設叢集中的某個成員(成員 1)想要讀取不在其本機緩衝集區中的資料分頁。DB2 會指派一個代理程式(或執行緒)來執行這項交易，接著代理程式會使用 RDMA 直接寫入 CF 的記憶體，指出它需要讀取的某個特定分頁。如果成員 1 想要讀取的分頁已經位於 CF 的通用集中式緩衝集區，CF 會將該分頁直接推入成員 1 的記憶體，而不是讓該成員中的代理程式執行讀取的 I/O 作業，以從磁碟讀取該分頁。使用 RDMA 允許成員 1 的代理程式只對遠端伺服器呼叫一個 memcopy (記憶體副本)，而不需要進行成本高昂的流程間通訊呼叫、處理器中斷、IP 堆疊呼叫等。簡而言之，pureScale 允許成員的代理程式在實際目標為遠端機器的記憶體位址時，執行看起來像是本機記憶體複製的作業。

這些輕量型遠端記憶體呼叫，以及集中式緩衝集區和鎖定管理設備，表示應用程式不需要連線到已經具有資料的成員。不論叢集大小，其中的任何成員都可以從通用緩衝集區有效接收資料分頁。大多數的 RDMA 呼叫都非常快速，這使得進行呼叫的 DB2 流程不需要在等候 CF 回應的同時讓出 CPU 資源，而且不需要重新排程工作即可完成工作。例如，為了要通知 CF 某列即將更新(因此需要一個 X 鎖)，某個成員的代理程式會透過將鎖定資訊直接寫入 CF 記憶體的方式執行 Set Lock State (SLS)要求。CF 會確認叢集中是否有其他成員鎖定此列 X，然後會直接寫入要求成員的記憶體以授予鎖定。這個 SLS 只需要不到 15 微秒即可完成整個過程，因此代理程式不需要讓出 CPU 資源。代理程式可以繼續以高效率運作，而不需要像其他水平擴充架構等候 IP 中斷(避免不必要的內文切換)。如果是針對特定操作(例如長時間執行批次交易)，DB2 代理程式就有必要讓出 CPU 的資源，而且 DB2 會自動決定是否要動態讓出 CPU 資源。

DB2 pureScale 針對整個叢集成員內建的負載平衡，是另一項重要的 DB2 可擴充性功能。即使不是叢集感知應用程式，也可以 DB2 pureScale 內建的利用負載平衡。DB2 for z/OS 資料共用客戶如今使用相同的客戶端磁碟機，可以和 DB2 pureScale 搭配使用以進行叢集負載平衡。

DB2 pureScale 實現高可用度

這種橫向擴充的架構不只提升了系統交易容量。在元件故障時，還可以繼續提供交易處理能力，因而提高了系統的可用度。

與可以在分散式平台上使用的其他產品相比，DB2 pureScale 將高可用度提升到全新的層次。DB2 pureScale 允許完整存取所有不需要復原的資料分頁，並且隨時都可以知道有哪些的特定分頁需要復原，而不需要執行任何 I/O 操作。這是透過集中式 CF 獨特功能實現的另一個重要的創新。

每當成員將分頁讀取到其緩衝集區時，CF 都會感知到並持續對其進行追蹤。只要成員希望更新分頁中的資料列(row)，CF 也會感知到該事件。每當應用程式確認交易時，該成員會將修改的分頁直接寫入 CF 的記憶體。這個程序讓叢集中其他任何想要讀取此已變更分頁的成員可以直接從 CF 取得更新。更重要的是，從復原的觀點來看，如果任何成員故障，CF 會有一個故障成員正在進行更新的分頁清單，以及故障的成員承諾要更新但尚未寫入磁碟的分頁。

*任何*關聯式資料庫管理系統(RDBMS)的復原程序，首先都要重做已經確認的任何交易，才能確保這些交易在磁碟中的分頁會是磁碟中最新的分頁(這個流程也稱為 *Redo Recovery*)。此外，任何的資料庫伺服器也必須復原所有進行中的(in-flight)交易，也就是在對已寫入至磁碟但在故障前尚未確認的資料進行清除(這個流程也稱為 *Undo Recovery*)。

在共用的磁碟叢集中，應先確保叢集中沒有其他的節點從磁碟讀取或更新尚未復原的任何分頁(復原這些分頁之後，才能在這些列執行任何新的交易)。這是 CF 的特出之處：因為 CF 知道故障節點正在更新哪些分頁，而且 CF 的集中式緩衝集區中已經有透過該節點確認的已修改分頁，所以 DB2 pureScale 不需要在判斷必須復原哪些分頁時封鎖其他成員繼續交易。其他的架構則需要瞭解哪些交易要花許多時間處理，才能判斷哪些分頁因為其發佈的鎖定資訊而必須加以復原(稍後將於該主題做進一步的說明)。

從較高的層面來看，可以很容易說明這種在 DB2 pureScale 環境中的復原過程。每個成員都有處於閒置的處理流程，但都已準備好處理故障事件。某個成員故障時，就會啟動其中一個復原流程，因為這些流程已經存在，所以作業系統不需要浪費寶貴的系統時間來建立流程、為其配置記憶體等。此復原流程會立即開始將修改的分頁從 CF 預先擷取至其所屬的本機緩衝集區。絕大多數的復原流程都不需要執行 I/O 操作，因為需要復原的分頁已經位於 CF 的集中式緩衝集區。除此之外，此分頁預先擷取會使用輕量型 RDMA，以在 CF 和復原成員間進行迅速有效的傳輸。在這段時間內，所有其他成員上所有其他的應用程式將繼續處理要求。如果這些應用程式需要從任何不需要復原的分頁中獲得任何資料，這些應用程式就可以繼續執行其交易。同樣地，他們可以繼續讀取磁碟中的分頁，因為 CF 已經確實知道磁碟中有哪些分頁是乾淨的，有哪些分頁需要復原。接著，為了重新顯示必要的交易，復原流程會讀取故障成員的交易日誌檔，以重做或復原故障成員所進行的更新。

就一般的交易工作負載而言，從成員故障到資料頁在另一成員復原而能執行下一筆交易，通常只需要 20 秒或更短的時間。請注意，這也包含故障偵測的時間，部分的廠商提到的復原時間可能都不含這段時間。即使***成員故障***之後，資料庫中***所有其他的分頁還是隨時都可供使用***。

除此之外，像是在 PowerHA pureScale 叢集加速系統中的元件都是多重元件。DB2 pureScale 允許雙工的 CF 功能，如此鎖定及共用快取資訊會儲存在兩個不同的位置，以因應主要 CF 故障的情況。

摘要

藉由利用現代的硬體架構，DB2 pureScale 能夠將先前只能在 DB2 for z/OS 中使用的集中式鎖定與快取功能導入分散式平台使用。利用這種硬體和網路可以進行更高等級的同步存取並可大幅降低耗損，從而提供更高水準的可用度。此外，集中式鎖定和分頁快取讓 DB2 pureScale 隨時知曉在成員故障時該復原哪些分頁。因此在遇到故障時，所有不需要復原的資料仍然可以繼續供其他應用程式使用，而且系統已經知道故障節點正在進行更新的分頁，復原的速度會更快。

對於需要高可用度以及水平橫向擴充能力的應用系統而言，DB2 pureScale 提供一個可以滿足這些需求、而且承襲自己受市場肯定的技術的客製化解決方案。

進一步的資訊

如需要瞭解 IBM DB2 降低管理資料成本的方式，請聯絡您的 IBM 代表，或請造訪 ibm.com/db2

IBM Canada Ltd.
8200 Warden Avenue
Markham, ON, Canada L6G 1C7

於台灣列印
2009 年 10 月

附註

[1] http://www.eweek.com/c/a/Database/In-Larrys-Own-Words/2/

IMW14253-USEN-01