我的英特爾MPI學習筆記

綜述

MPI 這個東西整體來看還是相當依賴具體實現和機器架構的。NAMD（一個分子模擬軟體）文件中有下面可能的話：

> 對於工作站叢集和其他具有**特殊高效能網路的**大規模平行機器，NAMD 使用系統提供的 MPI 庫（有少數例外）並使用 mpirun 等標準系統工具來啟動作業。由於 MPI 庫在版本之間通常不相容，因此您可能需要重新編譯 NAMD 及其底層 Charm++ 程式庫才能並行使用這些機器。

大意是說，MPI在版本之間以及廠商之間存在著更大的兼容性差異，所以很多時候為了能夠在高速網路互聯的大型集群系統之上正常運行的數值模式軟體，必須採用合適的編譯器重新編譯。

文檔細節

兩個行程調度器

這部分翻譯了，文檔裡面說得很清楚。很多老的PBS腳本裡面都預設使用MPD作為進程調度器，但是這種用法在新的編譯器下即將被廢棄。不過看起來老負載也不是一般的怎麼升級編譯器的樣子，畢竟大多數模式程式碼的編譯器相容性都有點慘。

九頭蛇

發佈時間:

2015-03-06

分類:

筆記 (https://blog.finaltheory.me/note/index.html)

標籤:

MPI (https://blog.finaltheory.me/tag/mpi.html)[1]

目錄

我的英特爾MPI學習筆記

1. 綜述
2. 文檔細節
   2.1. 兩個進程調度器
      2.1.1. 九頭蛇
      2.1.2. MPD
      2.1.3. 筆記
   2.2. 參數不相容
3. 命令列參數
   3.1. 關閉架構選擇…
   3.2. 架構選擇
   3.3. 調整進度分配

Hydra 是一個簡化的、可擴展的流程管理器。Hydra 將檢查已知的資源管理器，以確定進程可以在哪裡運行，並使用每個主機上的代理程式在目標之間分配進程。這些代理將用於進程啟動、清理、I/O 轉發、訊號轉發和其他任務。

您可以使用 來啟動 Hydra `mpiexec.hydra` 。請參閱可擴充進程管理系統 (Hydra) 指令主題，以了解英特爾® MPI 函式庫參考手冊中選項的詳細清單。
也可以透過**直接呼叫適當的 mpiexec 檔案來**選擇進程管理器：
`mpiexec.hydra` 對於 Hydra 或 `mpiexec` MPD。

MPD

MPD 代表多用途守護程式。這是用於啟動必須在所有節點上執行的平行作業的英特爾® MPI 庫進程管理系統。MPD 收集有關係統和硬體的信息，並相互通信以交換所需資訊。例如，需要 MPD 環才能在 MPD 流程管理器下正確固定。

筆記

從英特爾® MPI 庫 5.0 版本開始，多用途守護程序 (MPD)**已被棄用**。轉換為使用可擴展的流程管理系統 (Hydra) 來啟動平行作業。

參數不相容

大意是說兩個進程調度器採用不同的參數，並且會默默地忽略對方的一些特有參數：

在作業管理器下執行時，mpirun 指令會忽略該 `-r | --rsh` 選項 if `Hydra*` 用作底層行程管理器。在這種情況下， `Hydra*` 將使用相應的引導伺服器。明確使用引導程式特定選項或對應的環境變數來覆寫自動偵測到的引導程式伺服器。

4. 參數調節
5. 評論

如果您選擇作為活動進程管理器，則出於相容性原因，mpirun 命令**會靜默忽略**MPD 特定選項。 Hydra* 下表提供了靜默忽略和不支援的選項的清單 MPD* 。 Hydra* 如果使用流程管理器，請避免這些不支援的選項。

命令列參數

總結匯總的命令列參數如下：

```
 mpiexec.hydra -n 96 -hostfile ~/mpi_hosts -perhost 12 -genv
I_MPI_FABRICS shm:dapl ./test
```

關閉架構選擇時的Fallback功能

意思是說，預設選擇MPI自動架構，並且按照一個架構清單自上而下嘗試。其中如果一個運行成功，就不會報錯，但有可能會導致實際運行時的方式不是你所期望的。為了關閉該特性，可以設置 export I_MPI_FALLBACK=0 ，或指定 I_MPI_FABRICS 環境變數。

預設情況下，如果未設定 I_MPI_FABRICS，將啟用回退。如果設定了 I_MPI_FABRICS，則回退將被停用。

架構選擇

```
 I_MPI_FABRICS
```

選擇要使用的特定網路結構。

這是非常重要的一部分，指定程式運行時的架構。一般的架構描述是兩個縮寫，用號分割，如： 。 前面表示節點內部 `shm:dapl` 的連接架構，晚期表示節點內部的連接架構。 例如， 對於我們常用的叢集架構，節點之間是普通伺服器實際上多核心共享記憶體的SMP架構；節點之間採用高速Infiniband網路互聯，屬於 `DAPL-capable network fabrics` ，所以有了上述的架構。具體的可用架構清單請參閱文件說明。

調整流程分配

```
 -perhost <# of processes >, -ppn <# of processes >, or -grr
<# of processes>
```

使用此選項可使用循環調度在群組中的每個主機上放置指定數量的連續 MPI 進程。有關更多詳細信息，請參閱 I_MPI_PERHOST 環境變數。

首先這裡指出了MPI在調度進程的時候所採用的演算法，簡單來說就是各個節點輪流取得一個行程的意思。這個參數用來設定每個節點運行多少個行程。

參數調整

Intel MPI 實現了一個自動調優 MPI 參數的功能，雖然實際測試的效果一般，不過總算聊勝於無。MPI User Guide 中的「Tuning with mpitune Utility」章節描述了這個工具的最常用。這個東西主要有兩種工作模式：針對叢集架構的參數調優，以及針對某些特定應用的參數

調優。對於之前，顧名思義就是在特定的叢集上面跑起一些自帶的應用（當然也可以指定自己的），然後看哪個配置速度快，就會產生一份對應的設定檔。 復活節，跑一個使用者自訂的設定檔

(http://fortawesome.github.io/Font-Awesome)設計。
(https://github.com/FinalTheory/pelican-theme)
(http://fortawesome.github.io/Font-Awesome)