Dell EMCインフラストラクチャへの仮想化HPCクラスターの実装

ピサ大学CTO、Maurizio Davini氏

ピサ大学、HPCシステム管理者、Fabio Pratelli氏

概要

最新のハイ パフォーマンス コンピューティングでは依然として物理リソースを使用しなければなりませんが、仮想化は、インフラストラクチャの柔軟性という重要なアドオンのメリットを活用して、HPCにコスト効率と拡張性に優れた信頼性の高いIT環境を提供することが実証されています。ピサ大学での取り組みにおいて、私たちは、仮想化を使用してVMware vSphere®環境で仮想HPCクラスターの導入を迅速化する方法を示しました。

2020年5月

目次

Published in the USA 5/20.

ピサ大学のVMware環境

ピサ大学には、本番環境に複数のVMware vSphereクラスターがあります。この仮想化されたITインフラストラクチャは、200 Gb/sのデータ センター インターコネクト（DCI）によって接続されている3つのデータ センターに分散されています。

VMwareクラスターは、最新のハードウェア テクノロジーを使用して、エンタープライズ サービスからVDI（仮想デスクトップ インフラストラクチャ）までのリソースをホストします。インフラストラクチャの主要なコンポーネントは次のとおりです。

- Dell EMC PowerEdge R730XDおよびR740XDサーバー
- インテル® Xeon®プロセッサー
- 100Gb/s Mellanox Connectx-5 Ethernetカード
- インテル® Optane™ P4800X
- インテルSSD DC P4600 NVMeストレージ
- Dell EMC Isilonハイ パフォーマンスNAS
- Dell Networking Z9100スイッチ
- VMware vSAN™ハイパーコンバージド ストレージ

HPC仮想クラスターを選ぶ理由

仮想化の基本的な要素となるのは仮想マシン（VM）です。これは、リソース構成が基盤となるハードウェアのリソース構成とは異なる可能性がある環境でオペレーティング システムとそのアプリケーションの実行をサポートするソフトウェアを抽象化したものです。

HPCにおけるVMのメリットの分かりやすい説明は、VMwareのホワイト ペーパー『Virtualizing High-Performance Computing (HPC) Environments: Reference Architecture』に記載されています。このホワイト ペーパーでは、VMのメリットを次の用語でまとめています。

仮想マシンのメリット：

- **異種混在環境：**VMを使用することにより、異なるリソース構成、オペレーティング システム、HPCソフトウェアを同じ物理ハードウェア上で柔軟に混在させることができます。さらにセルフプロビジョニング モデルを使用すると、IT部門は各ユーザーの要件に応じて、研究者、科学者、エンジニアに対する課題解決までの時間を短縮して、さまざまな環境を提供できます。
- **制御性と調査の再現性の向上：**インフラストラクチャおよびHPCの管理者は、ロール ベースのアクセス権に基づいて動的なサイズ変更、一時停止、スナップショットの作成、バックアップ、他の仮想環境への複製、または単にVMのワイプと再導入を行うことができます。構成とファイルは各VM内にカプセル化されるため、コンプライアンスなどの調査目的でVMをアーカイブして再実行することができます。
- **リソースの優先順位付けとバランシングの向上：**VMのコンピューティング リソースは、個別に、またはプール内で優先順位を付けることができます。ロード バランシングのために、実行中のVMとそのカプセル化されたワークロードをクラスター全体にわたって移行することもできます。この移行により、ベア メタル アプローチと比較してクラスター全体の効率性が向上します。
- **障害分離：**分離されたVM環境でジョブを実行することにより、各ジョブは別のVMで実行されているジョブによって引き起こされる潜在的な障害から保護されます。

次のメリットも提供されます。

セキュリティ

- セキュリティ ルールとポリシーを環境、ワークフロー、VM、物理サーバー、オペレーターに基づいて定義および適用できます。以下が含まれます。
- ユーザー権限によって制御され、監査レポート用にログに記録されるアクション。たとえば、root権限は必要に応じて特定のVMに基づいてのみ付与されるため、他のHPCワークフローに対する侵害を防ぐことができます。
- 機密データを他のHPC環境、ワークフロー、または同じ基盤となるハードウェアで稼動しているユーザーと共有できない分離されたワークフロー。

耐障害性と冗長性

- HPC VMは、従来のHPC環境では利用できない耐障害性、動的リカバリー、その他の機能を提供します。具体的には、HPC VMは次のことを可能にします。
- 運用上のHPCワークフローや保守性に影響を与えないハードウェア メンテナンス
- サーバーの障害が発生した後のクラスター内の別の物理サーバーでの自動再起動
- 特定のホストのリソースがいっぱいになった場合の別の物理ホストへのライブ マイグレーション

ピサ大学の仮想HPC

この取り組みは、
Dell EMC PowerEdgeサーバー、インテル® Omni-Path InfiniBandネットワーク、並列ファイル システム（BeeGFS）で作成された従来のHPCベア メタル環境に仮想HPCリソースを統合するという試みから生まれました。

これを行うことで、柔軟なオンデマンドのHPCインフラストラクチャをユーザーに提供するか、VMware Sphereインフラストラクチャの機能を100%活用することができました。

VMware vSphereインフラストラクチャでは、次の2つのアプローチを使用することにしました。

1. ベア メタル クラスターでの本番環境で使用するOpenHPC/BeeGFSベースの同じHPCソフトウェア インフラストラクチャ環境
2. HPCおよびAI向けのKubernetes/Slurmベースの革新的なアプローチ

最初のアプローチはVMwareおよびOpenHPCのvHPCスクリプトをベースとしています。2つ目はデル・テクノロジーズのOmniaプロジェクトをベースとしています。

OpenHPCを選ぶ理由

OpenHPCはLinux Foundationの共同プロジェクトであり、その使命は、オープンソースのHPCソフトウェア コンポーネントとベスト プラクティスのリファレンス コレクションを提供して、最新のHPCメソッドとツールの導入、発展、使用に対するハードルを下げることです。

このプロジェクト自体は次のように説明されています。

「*OpenHPC*は、プロビジョニング ツール、リソース管理、*I/O*クライアント、開発ツール、さまざまな科学ライブラリーなど、*HPC*（ハイ パフォーマンス コンピューティング）*Linux*クラスターの導入および管理に必要な多くの一般的な要素を集約したいという願望から始まった、共同でのコミュニティーの取り組みです。*OpenHPC*によって提供されるパッケージは、*HPC*コミュニティーに再利用可能な構成要素を提供することを目的として、*HPC*統合を念頭に置いて事前に構築されています。」

ピサ大学のHPCコンピューティング施設のソフトウェア環境は、主にOpenHPCをベースとしています。バッチ システムとしてSlurmを使用したWarewulfステートレス プロビジョニングを使用しています。ローカル ノード ディスクは、BeeGFSを内部並列ファイル システムとして導入するために使用されます。

VMware vSphereでHPCクラスターをマッピングする方法：vHPC

私たちは、VMwareから提供されるvHPC Toolkit pythonスクリプトを使用することにしました。スクリプトは次の場所からダウンロードできます。

- github：https://github.com/vmware/vhpc-toolkit.git
- Flingsリポジトリー：https://flings.vmware.com/virtualized-high-performance-computing-toolkit

vHPCツールキットは、VMware vSphere APIを活用することで、HPCの特殊な構成（コンピューティングやRDMAインターコネクトにGPUおよびFPGAハードウェア アクセラレーターを活用するなど）のライフサイクル管理を容易にする方法です。

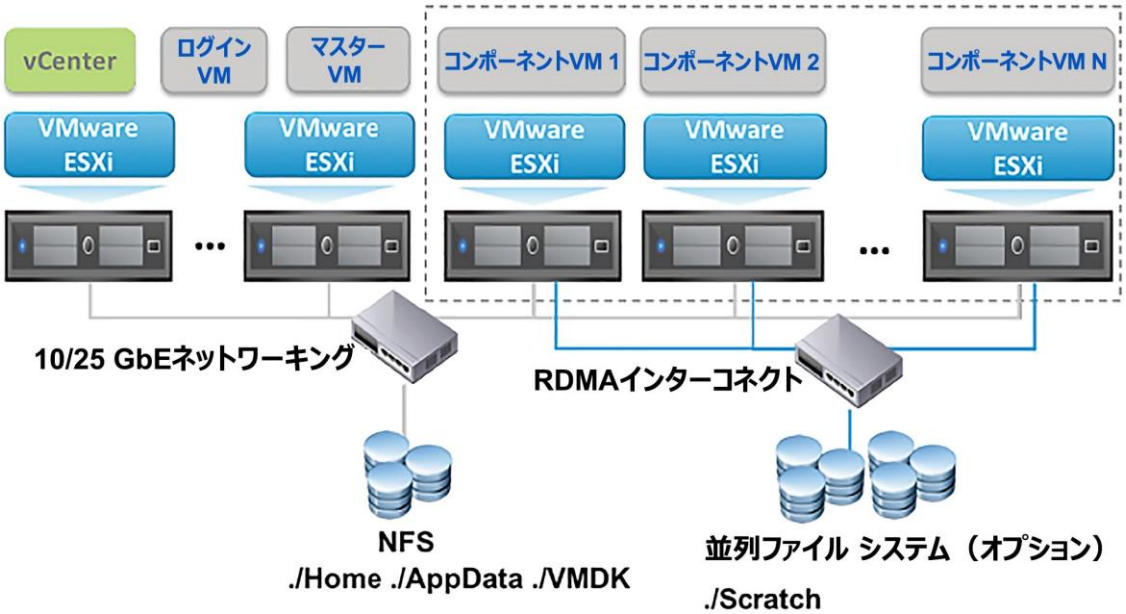
また、VMのクローン作成、レイテンシー感度の設定、vCPU、メモリーのサイズ設定など、このようなハイパフォーマンス環境の作成に関連するいくつかの一般的なvSphereタスクをvSphere管理者が実行する上で役立つ機能も含まれています。

主な特長：

- GPGPU、FPGA、RDMAインターコネクトなどのDirectPath I/OモードでのPCIeデバイスの構成
- NVIDIA vGPUの構成
- RDMA SR-IOV（シングル ルートI/O仮想化）の構成
- PVRDMA（準仮想化RDMA）の構成
- クラスター構成ファイルを使用して、仮想HPCクラスターを簡単に作成および破壊できるようにする
- VMのクローン作成、vCPUの構成、メモリー、予約、共有、レイテンシー感度、分散仮想スイッチ/標準仮想スイッチ、ネットワーク アダプター、ネットワーク構成などの一般的なvSphereタスクを実行する

仮想HPC vSphereインフラストラクチャ

基本的な考え方はVMware HPCのベスト プラクティスに基づいています。



6.2 基本的な仮想化HPC（vHPC）アーキテクチャ

Dell EMC PowerEdgeサーバー、インテル Optane、インテルNVMeドライブで構築された3ノードのvSphere/vSAN 6.7U3クラスターを導入しました。接続は、Dell EMC Networking Z9100スイッチの100-Gb/s Mellanoxカードによって提供されます。

vHPCツールキットは、VMware vSphereインフラストラクチャ内のUbuntu 18.04 VMにインストールされています。

vHPCツールキットとOpenHPC

使用したワークフローは次のとおりです。

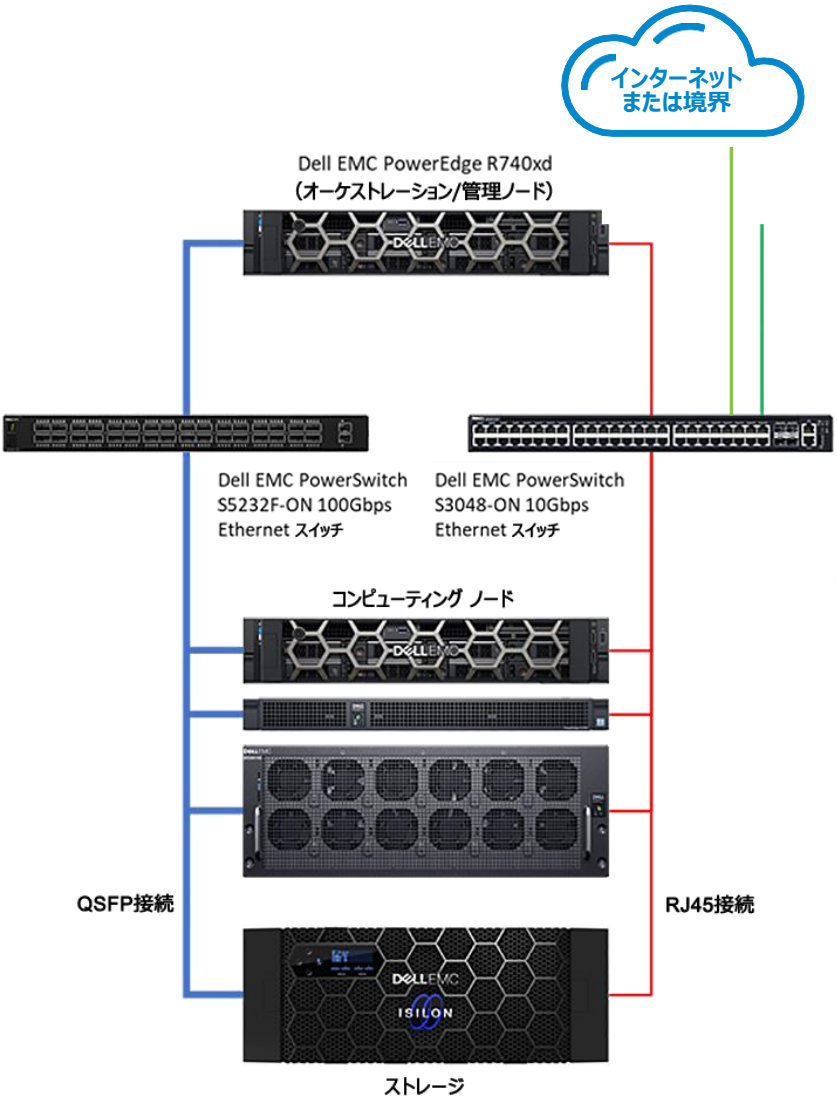
1. OpenHPCコンピューティング ノードのVMテンプレートをセットアップする
2. OpenHPCヘッド ノードのVMテンプレートをセットアップする
3. vSphere Distributed Switchをセットアップする
4. ヘッド ノードVMのクローンを作成し、インストールをパーソナライズする
5. vHPCスクリプトを使用してVMware HPCインフラストラクチャを導入する
6. OpenHPCヘッド ノードの構成を完了する：
   a. プロビジョニング
   b. Slurm構成
   c. BeeGFS
3. ノードを起動する

数分で仮想OpenHPCクラスターが稼働しました。

OMNIAプロジェクト

Omnia（ラテン語の意味：すべてまたはあらゆるもの）は、RPMベースのLinuxイメージを備えたDell EMC PowerEdgeサーバーを機能するSlurm/Kubernetesクラスターに変換するための導入ツールです。以下の開発者によって開発されました。

- Lucas A. Wilson（デル・テクノロジーズ）
- John Lockman（デル・テクノロジーズ）



Omniaは、サーバーのインベントリーにSlurmまたはKubernetesをインストールおよび構成するためのAnsibleプレイブックのコレクションであり、追加のソフトウェア パッケージやサービスも含まれています。

Omniaはgithubからダウンロードできます。

https://github.com/dellhpc/omnia

vHPCとOMNIA

vHPCスクリプトを使用して、実際のインフラストラクチャと同様のVMとネットワークを設計して導入しました。

行ったワークフローは次のとおりです。

1. Centos 7.8 VMテンプレートを導入する
2. vHPCスクリプトを使用してVMネットワークを設計し、Centos 7.8 VMを導入する
3. マスター ノードをセットアップする（Ansibleをインストールしてrootアカウントの SSHキーをセットアップする）
4. Omniaプレイブックを実行する

その後短時間で環境を使用する準備ができました。

まとめ

VMware vSphere環境で仮想HPCクラスターを正常に導入するために用いた方法をいくつか説明しました。私たちが使用したアプローチは依然として人の手による制御が必要であり、完全に自動化されているわけではありません。

オートメーションと利用可能な機能を向上させるため継続的に取り組んでいますが、現在では重要なメリットを実感しています。VMwareは、エンタープライズ ワークロード、仮想デスクトップ インフラストラクチャ、リモート ワークステーション、スマート ワーキング、科学コンピューティング、HPCのサポートなど、多くのワークロードにインフラストラクチャを使用できる柔軟性に優れた可能性を提供し、同時にライセンス コストを最適化する非常に柔軟な方法で同じインフラストラクチャでこれを実現します。

参考資料

VMware、『Virtualizing High-Performance Computing (HPC) Environments: Reference Architecture』（2018年9月）。

VMware、『Virtualized High Performance Computing Toolkit』

Omniaプロジェクト

OpenHPC

詳細については、hpcatdell.comをご覧ください。