Vol.2013-HPC-141 No.9
2013/9/30

HPCクラスタシステム上で動作する仮想マシンを用いたHadoopクラスタの構築

土屋 雅稔[1,a)]

概要：近年の HPC システムでは，1 つのノードで処理できる計算量とメモリ量の上限を超えるために，できるだけ高速なインターコネクトで多数のノードを疎結合したクラスタシステムを採用することが一般的である．同時に，HPC アプリケーションも，このような HPC クラスタシステムで実行することを前提として，複数のノードにまたがる並列化が行われてきた．近年，Hadoop に代表されるようなデータ集約的なアプリケーションが注目を集めている．これらのアプリケーションは，従来型 HPC アプリケーションとは異なり，HPC クラスタシステム向けに最適化されていないため，HPC クラスタシステムで処理するためには注意が必要である．
本稿では，従来型の HPC アプリケーションに限らず様々なアプリケーションを実行できる HPC クラスタシステムを目指して，HPC クラスタシステムを IaaS (Infrastructure-as-a-Service) 基盤として用いることを試みる．その例題として，HPC クラスタシステム上で動作する仮想マシンを用いて Hadoop クラスタを構成した結果について述べる．

Development of a Hadoop Cluster Using Virtual Machines Running on an HPC Cluster

Masatoshi Tsuchiya[1,a)]

Abstract: A modern HPC system is generally designed as a loose coupling cluster system of many computing nodes, in order to cope the limitaion of computing power and memory. Such system cannot process data-intensive applications as well as traditional HPC applications. This paper presents our ongoing development of a Hadoop cluster using virtual machines running on an HPC cluster system.

1. はじめに

HPC システムでは，1 つのノードで処理できる計算量とメモリ量の上限を超えるために，できるだけ高速なインターコネクトで多数のノードを結合したクラスタシステムを採用することが一般的である．同時に，HPC アプリケーションも，このような HPC クラスタシステムで実行することを前提として，複数のノードにまたがる並列化が行われてきた．

このような従来型 HPC クラスタシステムには，システム規模および実行環境についての柔軟性の欠如という問題点が存在する．HPC クラスタシステムの処理能力は，基本的には，ノード数，ノード毎の計算処理能力，ノード間のインターコネクト性能，ファイル入出力性能などによって決まる．特に，ノード数は重要な要素であるが，従来型 HPC クラスタシステムでは，必要な計算処理の規模に応じてノードを追加するなどの対応は困難である．また，大規模な HPC クラスタシステムは，計画から導入までに長時間を必要とするため，研究課題に応じて機動的に必要な計算資源を確保することも難しい．また，計算集約的な従来型 HPC アプリケーションに加えて，Hadoop に代表されるようなデータ集約的な HPC アプリケーションが注目を集めてきている．これらのアプリケーションにおいては，

[1] 豊橋技術科学大学情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology
[a)] tsuchiya@imc.tut.ac.jp

表 1　ログインノードおよび演算ノードの構成

ハードウェア	CPU	Intel Xeon E5–2680 (2.70GHz, 20MB キャッシュ, 8 コア) × 2 ソケット
	メインメモリ	64GB (ECC, DDR3–1600)
	補助記憶装置	HDD (10krpm, SAS 接続, 300GB) × 2 台 (RAID1 構成)
ソフトウェア	OS	Red Hat Enterprise Linux Server 6.2
	カーネル	Linux 2.6.32-220.4.2.el6.x86_64
	仮想マシンハイパーバイザ	kvm (上記 OS およびカーネルに同梱)

表 2　ストレージシステムの構成

ソフトウェア	並列ファイルシステム	Lustre 1.8.8
ハードウェア	MDS	HP ProLiant DL380P Gen8 (Intel Xeon E5–2650 ×1, メモリ 320GB) × 2 台
	OSS	HP ProLiant DL380P Gen8 (Intel Xeon E5–2670 ×1, メモリ 320GB) × 2 台
	MDT / ODT	DataDirect NETWORKS SFA10K–M (240TB)

従来型 HPC アプリケーションに比べて必要とする実行環境構成が大きく異なるため，クラスタシステムの利用者から見ると，かなり大幅に実行環境を変更したいという要望が発生する．しかし，従来型 HPC クラスタシステムは，システム管理者によって厳密に管理されていることが一般的であり，そのような利用者の要望にはなかなか応えられていない．

これらの問題に対応するため，パブリッククラウド上に仮想クラスタシステムを構築し，HPC アプリケーションを実行するという手法が注目されている．He ら [1] は，Amazon EC2, GoGrid Cloud, IBM Cloud の 3 社のパブリッククラウド上に仮想クラスタシステムを構築した結果について報告している．Roloff ら [2] は，Amazon EC2, Microsoft Azure, Rackspace の 3 社のパブリッククラウド上に仮想クラスタシステムを構築した結果について報告している．これらの手法では，必要な計算処理に応じてクラスタシステムの規模を柔軟に拡大することができるという利点がある．しかし，パブリッククラウド内部で用いられているインターコネクトの詳細については非公開のため厳密な比較は難しいが，Infiniband などの高速なインターコネクト技術に基づいて構築された従来型 HPC クラスタシステムと比較すると，パブリッククラウド上で構築された仮想クラスタシステム内のノード間通信速度は非常に低速であり，ノード間の通信速度が重要な HPC アプリケーションでは性能の低下が大きい．Mehrotra ら [3] は，Amazon によって提供されている Cluster Compute Instance を用いて仮想クラスタを構築した結果について報告している．Cluster Compute Instance は，通常の Amazon EC2 の仮想マシンとは異なり，仮想クラスタシステム向けに特に調整されているが，やはりインターコネクトのオーバーヘッドが大きく，一般的な HPC アプリケーションにおいては性能劣化が大きいと報告している．

本稿では，従来型の HPC アプリケーションに限らず様々なアプリケーションを実行できる HPC クラスタシステムを目指して，HPC クラスタシステムを IaaS (Infrastructure-as-a-Service) 基盤として用いることを試みる．仮想マシンハイパーバイザとして Linux カーネルと kvm[4] を用いると，通常のアプリケーションを物理マシン上で動かしつつ，同時に，kvm 上で仮想マシンを動作させることが可能である．本学では，このような kvm の特徴に注目して，従来型の HPC アプリケーションについては物理マシン上で稼働させ，同時に，IaaS 環境を kvm 上で稼働させるというハイブリッドなクラスタシステムを実装した．本稿では，そのようなクラスタシステム上において，Hadoop クラスタを構築し，ベンチマーク試験を行った結果について報告する．

2. システム構成

本システムは，2 ノードのログインノードと 28 ノードの演算ノード (日立製 HA8000–tc/HT210) を，4x FDR InfiniBand と GbE を用いて相互接続した，非常に一般的な構成の Intel64 アーキテクチャクラスタシステムである．ログインノードおよび演算ノードの構成を表 **1** に，ストレージシステムの構成を表 **2** に示す．汎用 GPU などの演算加速装置は搭載していない[*1]．ノードあたりの理論演算性能は 172.8GFLOPS，システム全体の演算性能は，Linpack ベンチマークで測定して，$R_{\max}$ 値が 9.656TFLOPS, R_{peak} 値が 10.368 TFLOPS だった[*2]．

IaaS 環境を実現するには，大別すると，仮想マシンハイパーバイザ，仮想ディスクを格納するストレージ，および仮想マシン管理ミドルウェアという 3 つの構成要素が必要である．本システムの各ノードは，表 1 に示す通り，一般的な Linux システムを OS として用いているため，仮想マシンハイパーバイザとしては kvm を利用することができる．kvm では，仮想マシンの仮想ディスクは QCOW2 形式 [5] を用いて，通常のファイルとして格納することができるため，ストレージシステム (表 2) を仮想ディスクを格納するために用いることができる．なお，本システムで

[*1] 正確には，ログインノード 2 ノードと演算ノード 2 ノードに，NVIDIA Tesla K20X を 1 基/ノードずつ搭載しているが，全演算ノードに搭載していない状態である．
[*2] アクセラレータは利用せずに測定した．

表 3　Hadoop クラスタのソフトウェア構成

ソフトウェア	バージョン
OS	CentOS 6.2 (64bit)
Hadoop	CDH (Cloudera's Distribution Including Apache Hadoop) 4.1.2
Java	Oracle Java SE 7u11

表 4　Hadoop クラスタの仮想マシン資源割り当て

	CPU コア数	メモリ容量	ディスク容量
マスタノード	8 コア	60GB	100GB
スレーブノード	1 コア	3GB	100GB

は，仮想マシンの配置に関する制約が生じることを避けるために，各ノードの補助記憶装置には仮想ディスクを格納しない方針を採る．多数の仮想マシンを生成 · 管理するミドルウェアとして，本システムでは，CloudStack[6] を用いる．CloudStack は，事前に用意した仮想マシンの雛形 (テンプレート) に基づいて，多数の仮想マシンを容易に配備する機能を有する．本稿では，この機能を用いて，Hadoop クラスタを配備しベンチマーク試験を行った．以上のように，一般的な構成の HPC クラスタシステムは，仮想マシン管理ミドルウェアを加えると，IaaS 環境として容易に利用できるようになる．なお，従来型の HPC アプリケーションについては，仮想化による影響を最低限に抑えるため，ジョブスケジューラ[*3]を経由して，仮想化していない演算ノードの OS 上で直接に実行する．

3. 実験

実験に用いた Hadoop クラスタのソフトウェア構成を，表 **3** に示す．Hadoop のディストリビューションとして広く普及している CDH (Cloudera's Distribution Including Apache Hadoop) を利用した典型的な構成である．Hadoop クラスタの性能測定には，CDH に標準で含まれている Teragen / Terasort プログラムを利用した．Hadoop クラスタを構成する仮想マシンに対する資源割り当てを，表 **4** に示す．Hadoop クラスタは，割り当てられた部分演算を担当するスレーブノードと，タスク全体の管理 · 部分演算の割り当てなどを行うマスタノードという，2 種類のノードからなる．マスタノードが制約条件となることを避けるため，マスタノードには余裕を持たせた資源割り当てを行い，さらに，独立した物理ノードを割り当てた．

最初に，物理ノードを 1 ノードだけ用いて，その物理ノードで実行するスレーブノード数 (VM 数) を変化させた場合の結果を，表 **5** に示す．表 5 より，VM 数を増やすことによって，データサイズが増えても処理時間がほとんど増えないこと，言い換えれば，VM 数を増やすことによって処理性能が上がっていることが分かる．表 5 と同じ実験において，物理ノードのネットワーク入出力量を測定し，得ら

[*3] 本システムでは Torque を使用している．

表 5　単一演算ノードにおいて VM 数を変化させた場合の処理時間

VM 数	データサイズ	Teragen 処理時間 (秒)	Terasort 処理時間 (秒)
1	16GB	222	2343
4	32GB	219	2448
8	64GB	279	1540
16	128GB	454	2316

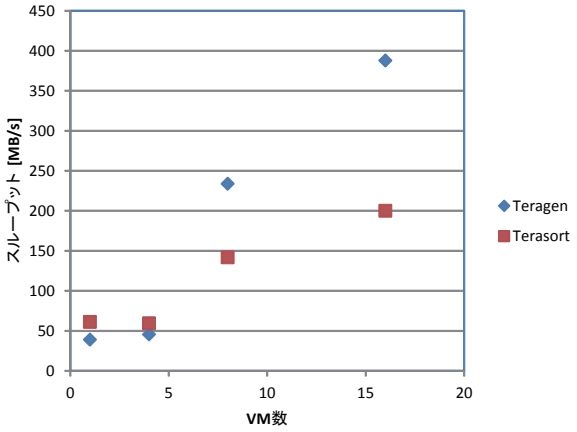

図 1　単一演算ノードにおいて VM 数を変化させた場合のスループット

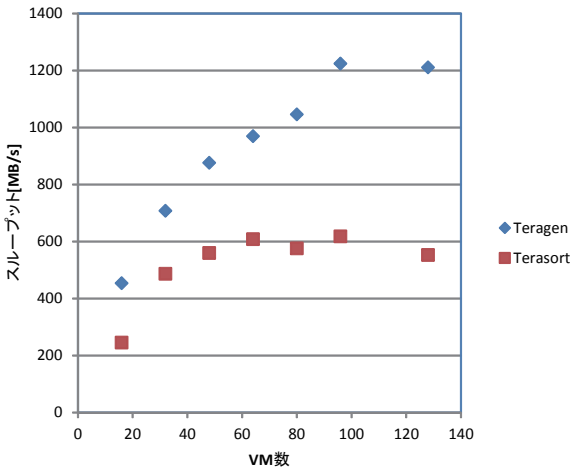

図 2　VM 数を変化させた場合

れた入出力量を処理時間で割った値を，平均スループットとする．この平均スループットと VM 数の関係を，図 **1** に示す．図 1 より，物理ノードに実装されている物理コア数まで VM を増やしても，スケールしていることが分かる．

次に，物理ノードあたり 16 個の仮想マシンを割り当て，物理ノード数を増やすことによって仮想マシンを増やした場合の結果を，図 **2** に示す．Teragen / Terasort ともに，VM 数が少ない区間では VM 数と性能が比例している．しかし，Teragen についてはスループットが 1200MB/s 程度に達した時点で，Terasort についてはスループットが 600MB/s 程度に達した時点で，性能が向上しなくなっている．この要因については，現在引き続き検討中である．

4. おわりに

本稿では，従来型の HPC アプリケーションに限らず様々なアプリケーションを実行できる HPC クラスタシステムを目指して，HPC クラスタシステムを IaaS (Infrastructure-as-a-Service) 基盤として用いることを試みた．仮想マシンハイパーバイザとして kvm を利用，仮想ディスクを HPC 用並列ファイルシステムに格納し，仮想マシン管理ミドルウェア (CloudStack) と組み合わせて，IaaS 基盤とした．このシステムでは，従来型の HPC アプリケーションについては物理マシン上で稼働させると同時に，各種の仮想マシンを実行することが可能である．このクラスタシステムを IaaS 基盤として用いて Hadoop クラスタを構築し，ベンチマーク試験を行った結果について報告した．

今後の検討課題としては，Hadoop クラスタの性能上限を決定している要因を明らかにする必要がある．また，従来型 HPC アプリケーションを管理するジョブスケジューラからは，どの演算ノードにおいて仮想マシンが実行中かの情報が見えていないため，現状では適切な負荷分散ができていない点も問題である．

参考文献

[1] He, Q., Zhou, S., Kobler, B., Duffy, D. and McGlynn, T.: Case study for running HPC applications in public clouds, *High Performance Distributed Computing, 2010. Proceedings. 19th IEEE International Symposium on*, pp. 395–401 (2010).

[2] Roloff, E., Diener, M., Carissimi, A. and Navaux, P. O. A.: High Performance Computing in the cloud: Deployment, performance and cost efficiency, *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, pp. 371–378 (2012).

[3] Mehrotra, P., Djomehri, J., Heistand, S., Hood, R., Jin, H., Lazanoff, A., Saini, S. and Biswas, R.: Performance evaluation of Amazon EC2 for NASA HPC applications, *Proceedings of the 3rd workshop on Scientific Cloud Computing Date*, ScienceCloud '12, pp. 41–50 (2012).

[4] Kivity, A., Kamay, Y. and Laor, D.: kvm: the Linux Virtual Machine Monitor, *Proceedings of the Linux Symposium, Volume One*, pp. 225–230 (online), available from ⟨`http://www.kernel.org/doc/ols/2007/ols2007v1-pages-225-230.pdf`⟩ (2007).

[5] McLoughlin, M.: The QCOW2 Image Format, (online), available from ⟨`https://people.gnome.org/~markmc/qcow-image-format.html`⟩ (accessed 2013–09–01).

[6] Childers, C.: CloudStack, Apache CloudStack Project Management Committee (online), available from ⟨`http://cloudstack.apache.org/`⟩ (accessed 2013–09–01).