大規模クラスタ構築における DCAST の有効性

Effectiveness of DCAST for setting up large-scale cluster

中尾　昌広
Masahiro NAKAO

Abstract: Recently, the demand of large-scale calculation is increasing in the technology calculation field. Then, although PC cluster is one of the ideas which replied it, the work and update which take numerous time are mentioned as a fault in the case of construction. DCAST(Doshisha Cluster Auto Setup Tool) was developed as a tool for canceling the fault. This report describes the large-scale cluster construction which used DCAST.

1　はじめに

科学技術分野における計算には大規模計算資源を必要とし，かつ高速なものが多く存在する．近年，このような計算を行うインフラとして PC クラスタが注目されている．PC クラスタは，大規模計算のために特別に設計されたハードウェアではなく，複数台の PC によって構成される並列計算機である．全て市販品で構築できるため，PC クラスタは専用計算機に比べて以下の利点がある．

- ノード数やネットワークなどといったシステム構成の自由度が高い
- 最新技術をすぐに取り込むことができる
- 特殊なハードウェアを使用しないために安価である
- 他の高性能計算機に比べて安価である

しかし PC クラスタの問題点として，クラスタ構築やシステムアップのためにかかる作業時間の多さが挙げられる．そこで同志社大学工学部知的システムデザイン研究室のクラスタグループ (以下クラスタグループ) では短時間でクラスタが構築でき，簡単にシステムのアップデートができる DCAST（Doshisha Cluster Auto Setup Tool）を開発した．

本報告では大規模クラスタにおける DCAST の有効性について述べる．

2　PC クラスタシステムの構成

PC クラスタは Fig. 1 のように，マスター (Master) と呼ばれるクラスタシステム全体を管理するためのマシンと，スレーブ (Slave) と呼ばれる計算を行うマシンの 2 種類から構成される．

スレーブの内，ハードディスクのあるスレーブは ”ディスクフルノード (diskfull node) ”，ハードディスクがないノードは ”ディスクレスノード (diskless node) ”と呼ばれる．ディスクレスノードにおいて，起動に必要なファイルは NFS(Network File System)[1]などを利用して，マスターノード上のファイルをネットワーク経由で参照する．

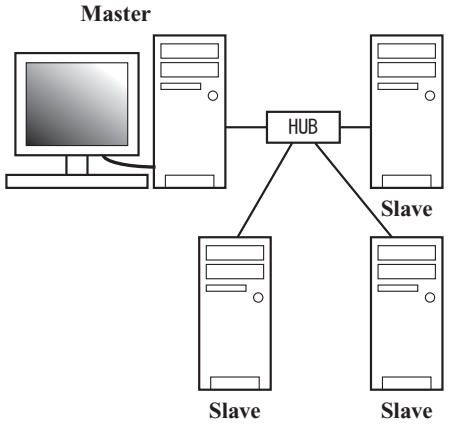

Fig. 1　クラスタシステムの概念図

2.1　PC クラスタの構築手順

次に一般的な PC クラスタの構築手順について述べる．PC クラスタの構築手順はディスクフルクラスタとディスクレスクラスタで異なる．以下にそれぞれの構築の手順を示す．

- ディスクフルクラスタの構築手順

  1. マスターに OS をインストール
  2. カーネルの再構築
  3. 必要なソフトウェアのインストール
  4. rsh，NFS，NIS サーバの設定
  5. スレーブに OS をインストール
  6. カーネルの再構築
  7. 必要なソフトウェアのインストール

[1]分散ファイルシステムを実現するソフトウエア．ネットワークで相互接続された別のコンピュータのファイルを自分のコンピュータにマウントできる

8. rsh(remote shell)[2]，NFS，NIS(Network Information Service) [3]クライアントの設定
9. NFS を用いてのマスターへのマウント

- ディスクレスクラスタの構築手順

1. マスターに OS をインストール
2. カーネルの再構築
3. 必要なソフトウェアのインストール
4. rsh，bootp[4]，tftp[5]，NFS サーバの設定
5. スレーブ用のディレクトリの作成
6. スレーブ用の設定ファイルを作成
7. スレーブ用カーネルの作製
8. スレーブ用のブートフロッピーの作製
9. ディスクレスマシンの起動

2.2 PC クラスタの問題点

PC クラスタは市販のハードウェアを使用して構成できるため，高い汎用性・安価などの利点がある．その一方で PC クラスタにはいくつかの欠点も存在する．以下にその問題点を列挙する．

- 構築時の労力
  ディスクフルクラスタの場合，PC クラスタを構成する全てのマシンに OS をインストールしなければならない．またインストール後，PC クラスタを構築するためのすべての設定をスレーブノードで行う必要がある．
- ソフトウェアのインストール
  PC クラスタに新しいソフトウェアをインストールを行う場合，またシステムのアップデートを行う場合，PC クラスタを構成する全てのマシンにおいてその作業が必要となる．
- ノードの追加作業
  PC クラスタのノードを増やす場合は，既存の全てのノードにおいて追加するノードに対する通信を許可する設定などを書き加える必要がある．

つまり，ディスクフルクラスタ，ディスクレスクラスタともにノード数に比例して作業量が増加する．またこれらの作業を手動で行う場合，スレーブノードの数が多くなると記述ミスが起こる可能性も高くなる．

[2]リモートシステム上で指定したコマンドを実行するためのソフトウェア
[3]パスワードやホスト名など，ネットワークの利用に必要な情報を一括管理するサービス
[4]自ら OS を持たない端末がサーバに接続するためのプロトコル
[5]ファイル転送プロトコル．これを用いてサーバ上のカーネルをダウンロードする

3 DCAST とは

3.1 DCAST の特徴

DCAST はクラスタグループが開発した PC クラスタの構築･管理を容易に行うためのツールである[6]．DCAST ではマスター・スレーブの両方で必要な以下の設定を自動で行ってくれる．

- ノード固有のホスト名・ネットワークの設定
- マスターにインストールしたソフトウェアのコピー
- 並列計算を行うための暗号化なしで通信を行う設定

DCAST の特徴を以下に示す．

- PC クラスタ構築の高速化
  PC クラスタ構築にはマスター/スレーブともにさまざまな作業が必要となるが，マスターの設定のみでスレーブに設定を反映させ，クラスタとして動作させることができる．
- PC クラスタ構築の簡易化
  PC クラスタの構築を行うには，ネットワークや Linux などの知識が必要となるが，特定の書式に従った設定ファイルを用意するだけで，特別な知識を持たなくても PC クラスタが構築できる．
- システム全体のアップグレード作業の軽減
  Linux カーネルのアップデートや，新たなソフトウェアの導入などにその変更を反映させることができる．
- 設定の一元管理
  ゲートウェイ，ネットマスク，ホスト名などのネットワーク設定をマスターで一元管理を行うことにより，設定の変更を容易に行うことができる．

3.2 DCAST を用いた PC クラスタ構築の手順

DCAST を用いた PC クラスタ構築の手順を以下に示す．

1. マスターに OS をインストールする
2. マスター，スレーブのカーネルを再構築する
3. 必要なソフトウェアのインストール，設定
   PC クラスタとして動作するためのソフトウェアをインストールする．
4. DCAST のダウンロード，slave.lst の設定
   DCAST をダウンロードし，展開したファイルにある slave.lst にホスト名やネットワークの設定などを記述する．Fig. 2 に slave.lst の記述例を示し，以下それぞれの項目について説明する．

[6]アーカイブは http://mikilab.doshisha.ac.jp/dia/research/cluster/dcast/download/list.cgi にある

```
#Enter PARTITION size.
FPRT /dev/hda1 128 boot *
SPRT /dev/hda2 512 swap
TPRT /dev/hda3 -   /
4PRT /dev/hda4 -   etc
HOSTNAME rna-
NISDOMAIN cluster
LOCALETHCARD eth0
#   NETWORK    NETMASK    BROADCAST
NET 10.0.0.0 255.0.0.0 10.255.255.255
#MASTER Master's name Master's IP
MASTER rna-101 10.0.2.101
SMASTER cambria 10.0.2.1
GATEWAY 10.0.2.1
#slave's name slave's IP slave's MACaddress
1   192.168.0.1 000000000000
102 10.0.1.102 0040C7975A9E
103 10.0.1.103 0040C79757DE
104 10.0.1.104 0040C79761E1
```

Fig. 2　slave.lst の記述例

- "FPRT〜4PRT"
  パーティションの設定を表す．順番に，デバイス，サイズ (単位は MB)，ディレクトリ (swap のみスワップパーテションを示す)，ブートデバイス指定を表す．*はどれか一つのみ記入し，swap の行には記入しないようにする．
- "HOSTNAME"
  スレーブのホスト名の接頭辞を記入する．上記の場合，スレーブのホスト名は，rna-102，rna-103... となる．記入しなかった場合は，ホスト名は 102，103... となる．
- "NISDOMAIN"
  NIS を用いた際のドメイン名を記入する．
- "LOCALETHCARD"
  DCAST を実行するマスター (PC クラスタ内部向き) の NIC を指定する．
- "NET"
  ネットワーク，ネットマスク，ブロードキャストの IP をそれぞれ指定する．
- "MASTER"
  DCAST サーバのホスト名，IP アドレスを順に指定する．
- "SMASTER"
  構築する PC クラスタのマスターノードのホスト名，IP アドレスを順に指定する．作成されたたスレーブはこのノードにマウントする仕様になっている．
- "GATEWAY"
  ゲートウェイとなるノードのホスト名と IP を順に記述する．
- "slave's name"
  slave となるマシンのマシン名と IP，NIC の MAC アドレスを指定する．

  編集した slave.lst は/bin/tftpboot 以下に置く．

5. スクリプトの起動
   DCAST のアーカイブの中にある makecluster というスクリプトを実行する．このスクリプトがマスターとスレーブのネットワーク設定などのファイルの書き換えと，ディスクレスクラスタが起動するのに必要なファイル群の作製を自動で行う．
6. grub フロッピーの用意
   スレーブは起動後まずネットワークを経由してカーネルを入手する．そのためブートローダである grub をインストールしたフロッピーディスクを作成する．

4　DCAST を用いた大規模クラスタ構築

Cambria システムは，本研究室が所有する 256 + 1 ノード[7]の大規模クラスタである．この Cambria システムを DCAST を用いて構築した．その過程について報告する．

4.1　Cambria システムの構成

Cambria システムの構成については Fig. 3 のように，dna，rna，prot，amin の各スレーブノード 64 台がマスターの Cambria システムに接続されている．256 台

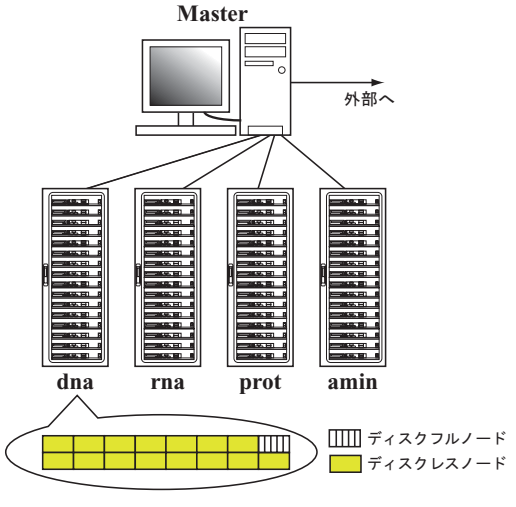

Fig. 3　Cambria システムの構成

のスレーブは，16 台に 1 台がハードディスクを備えており (このノードを bootmaster と呼ぶ)，残り 15 台はディスクレスノードである．つまり dna-101〜164 なら

[7]ディスクレス 240 台，ディスクフル 16 台，マスターホード 1 台

dna-101，dna-117，dna-133，dna-149がハードディスクを持っている．ハードディスクを備えたノードは，マスターの/homeをマウントするだけでよいが，ディスクレスノードはハードディスクを備えたノードにもマウントする必要がある．

またマスターには5枚のNICが備えられており，内1枚が外部へ，残り4枚がdna，rna，prot，aminのそれぞれ64台ずつにつながっている．これは通信負荷を分散するためである．

4.2 Cambriaシステム構築作業の手順

Cambriaシステムの再構築の手順を以下に示す．

1. Grubフロッピーの作成
   dsh[8]を用いて，ディスクフルノード用Grubフロッピー16枚とディスクレスノード用Grubフロッピー240枚を作製した．
2. DCASTサーバへのOSと必要なアプリケーションのインストール
   DCASTサーバとはDCASTをダウンロードし，makeclusterのスクリプトを実行させるノードのことである[9]．このノードにOSとしてDebian GNU/Linux3.0，DCASTに必要なソフトウェア(rsh，tftpなど)と，PCクラスタの運用・管理，そして研究に必要なソフトウェア(mpichなど)をインストールした．
3. bootmasterの作製
   まずDCASTサーバでmakeclusterを実行させて新しいbootmasterを作製した．1台のbootmasterに対し15台のディスクレスノードが/home以外にもマウントするためである．
4. bootmasterの設定変更
   DCASTは一度の実行で各スレーブのマウント先は一つしか指定できない．しかしCambriaシステムは通信負荷を分散するために/home以下のマウント先がdna，rna，prot，aminによって異なる．このためマウント場所の設定変更を行った．
5. ディスクレスノードの作製
   それぞれのbootmasterのslave.lstを編集し，各ディスクレスノード用のファイル群(15台)をmakeclusterを実行させて作製した．
6. 全てのノードの必要なファイルの書き換え
   それぞれのbootmasterでmakeclusterを実行したので，このままでは各bootmasterとそこで作製したディスクレスノード15台間でしか通信が行えない．また，アカウントとそのパスワードの組が新しく作製したDCASTサーバのままなので，構築前に保存しておいた/etc/passwdで上書きする[10]．

最後に，以前とIPアドレスを変えたため，マスターに通信やマウントを許可するファイルの書き換えを行わなければならなかった．以上の操作を行い，Cambriaシステムが構築できた．

4.3 Cambria構築から得た今後の展望

Cambriaシステムはネットワーク構成がかなり特殊であり，Cambriaに対応できるようにDCASTを変更したことが原因で予想外のバグが発生した．特に時間がかかった箇所はslave.lstにホスト名・IP・MACアドレスの組を記述することであった．しかしこれらのファイルは一度作製し保存しておくと，もう一度Cambriaを再構築するときは以前よりもはるかに短時間で作製できる．

また，このことを教訓としてDCASTは色々な構成のPCクラスタにも対応していく必要があることを認識できた．

5 まとめ

DCASTを用いて大規模クラスタを構築したことにより，PCクラスタの欠点である構築・システムのアップデートの際に生じる労力が大分減少されたことを確認できた．しかし，マウント先が一つというネットワーク構成のクラスタにしか対応していないため，様々なネットワーク構成のクラスタシステムに対応させるための拡張が必要である．

参考文献

1) 児玉憲造，PCクラスタ自動構築ツールの開発，第50回月例発表会

2) 児玉憲造，大規模クラスタにおける簡易セットアップ・管理ツールの提案，同志社大学理工学研究報告VOL.43,No.4

3) 児玉憲造，DCASTホームページ，
   http://mikilab.doshisha.ac.jp/dia/research/\\cluster/dcast/index.html

4) 谷村勇輔，Diskless Cluster構築入門，
   http://mikilab.doshisha.ac.jp/~tanisuke/pcc/\\archives/seminer_dc.pdf

5) 斉藤宏樹，クラスタゼミ・第2回配布資料，
   http://mikilab.doshisha.ac.jp/dia/seminar/2002/pdf/cluster02.pdf

[8] dsh(distributed shell)とは，複数のシステムに同時に実行させることのできる分散シェルである

[9] データ消失のおそれがあるのでマスターをDCASTサーバとせず，amin-149をDCASTサーバとした．なお，amin-149はハードディスクを持つbootmasterである

[10] dna-101～dna-132までは外部利用者用ノードであり，その他のノードとは/etc/passwdの内容が違う