2013 年 11 月 26-27 日 遺伝研

新学術 「ゲノム支援」
平成 25 年度 情報解析講習会
Linux サーバの構築

東京工業大学 大学院生命理工学研究科
森 宙史
Hiroshi Mori

ゲノム研究の情報解析で行う計算は、メモリを大量に使用したり、データサイズが大きかったりなど、一般的な Web サーバで行う計算とは異なった特徴を持っている。それらの計算をクラウド上で行う試みや、研究インフラとしてのスーパーコンピュータの整備が進行中であるが、他のユーザとのマシンリソースの取り合いで待たされること無く、最新の解析ソフトウェアを気軽に試して、迅速に解析結果を得たいのであれば、研究室で解析用のサーバを持つ意義は大きい。この 1 時間の講義では、ゲノム研究の情報解析を行うためのサーバを構築する上で特に重要になる、以下の 4 項目について解説する。

[1] 目的に応じたハードウェアのスペックの目安
[2] OS の選択
[3] Linux ディストリビューションの選択
[4] 実際の Linux OS のインストール&初期設定

[1] 目的に応じたハードウェアのスペックの目安

ゲノム研究の情報解析で用いられる計算手法は多岐にわたっているが、ここではゲノム研究と特に関連の深い計算手法に焦点を絞り、それらの計算を行うために必要なサーバのハードウェアのスペックの大体の目安を解説する。

(1) サーバの物理的なタイプ

・タワー型 簡単に言うと通常のデスクトップ型 PC の大きいもの。1 台のサーバに後から様々な部品を搭載できるように拡張性を重視して設計されている。

・ラック型 19 インチのラックに数ユニットごとにサーバや無停電電源装置、

外付けディスク等を搭載し、縦にどんどんと積み上げていく形式　ラックは通常横 60cm 奥行き 105cm 縦 200cm または縦 82cm と大きいが、何台もタワー型サーバを導入する予定があるのなら、ラック型にした方が省スペースになる。

・ブレード型　電源やファン等をサーバとは切り離すことによって、ラック型よりもさらに高密度に筐体にサーバを搭載できるようにしたもの。

・その他

ラック型やブレード型は何台もサーバを運用することが前提となっているので、最初はタワー型を選んだ方が無難。

(2) サーバの機種の違い

サーバの機種（チップセットまたはマザーボードの種類）がハードウェアの拡張性の大部分を決めるので、ある程度拡張性が高いものを選ぶ。ある程度というのは、経験的にサーバの部品（ディスクやメモリ等）は導入後約 3-4 年過ぎたあたりからちらほらと壊れだし、無料での部品交換の保証期間もそれぐらいで切れる。保証期間が切れた後は、例えばエンジニアを派遣してもらって 2TB のハードディスク 1 本を交換すると、10 万円ぐらいかかることもある。したがって、サーバの寿命は大体 4-5 年ぐらいと考えるべき。購入後それくらいの期間は研究室の情報解析をそのサーバで行えることが期待される。

以下の 4 項目はサーバの機種選びで特に気をつけるべきもの。

・ディスクの種類と最大搭載数
・メモリの種類と最大搭載数
・CPU の種類と最大搭載数
・電源の容量

デスクトップ PC を自作したことがある人もいるかもしれないが、サーバ関連のハードウェアの知識に自信が無いのであれば、サーバは自作しないほうが無難。

(3) 目的別に必要なサーバの大まかなハードウェアスペック

Web サーバ

1. 論文の Supplements を公開したり、研究プロジェクトの Web ページを公開したりする

大量のアクセスが見込まれたりする場合を除けば、メモリ数 GB、1CPU 数コア、ディスク数百 GB で十分な場合が多い。新型シーケンサーのリードデー

タ等の巨大なファイルをプロジェクト内で交換出来るように、ftp サーバや WebDAV サーバとしても利用するのであれば、ディスクはもっと積んだ方が良い。

2. 研究で得られたデータを何らかの DB 管理システムに格納して Web で公開し、検索可能にする

DB のデータの種類や数、アクセス数に依存するが、メモリ数十 GB（16GB-32GB ぐらい）、1CPU 数コア、ディスク数 TB あれば、大抵は十分。

解析用サーバ

3. 原核・真核・メタゲノム・Transcriptome のアセンブル

特に重要なハードウェア: メモリ・CPU

原核ゲノムの場合、数十 GB メモリ(50GB あれば余裕)、1CPU 数コア。Transcriptome アセンブルの場合、今の HiSeq のリード 1 億 pair ぐらいなら大体 128GB あれば十分。1-2CPU 数コア-数十コア。例えば、よく使われているソフトウェアである Trinity の場合は、A basic recommendation is to have 1G of RAM per 1M pairs of Illumina reads. (http://trinityrnaseq.sourceforge.net/) と目安を設定している。

メタゲノム・真核ゲノムは、ゲノムサイズやどれくらい hetero なのか等によって変わるが、数百 GB から数 TB 必要なこともよくある。2CPU 以上 数十コア。

4. ゲノムやメタゲノムの遺伝子アノテーションや比較解析のための、BLAST などでの配列相同性検索

特に重要なハードウェア: CPU

クエリ配列の数が数千本以上になる場合がほとんどなので、高速化のために CPU のコア数は多い方が良い。注意が必要なのは、BLAST では、使うコア数に応じて必要なメモリ量も増加する。現在の nr 相手の数千本の配列の BLASTX や BLASTP なら、1 プロセスごとに 5GB ほどメモリを消費する場合が多い。日常的に数千本以上の配列を nr 等の大きな DB 相手に配列相同性検索をするなら、十数コアは欲しい。その場合にはメモリも 100GB ほど必要。

5. Resequencing, RNA-Seq, ChIP-Seq 等の解析のための、Bowtie2 などのマッピングツールでの Reference ゲノムへのマッピング

特に重要なハードウェア: ディスク

リードのゲノムへのマッピングは、高速でありメモリ使用量も少ない。しかし、マッピング結果の SAM や BAM ファイルは数十〜数百 GB になることもよ

くあり、sort したりするたびに新しいファイルが出来るので、頻繁にこれらのデータを解析するなら、ディスクは最低でも数十 TB は用意しておくべき。

［2］OS の選択

Windows

・バイオインフォマティクスの様々なソフトウェアが動かない

・Cygwin や Linux の Virtual OS をインストールする手もあるが、本末転倒であり、メモリも多く消費してしまう

・マウスクリックでサーバの設定が可能であるが、Linux とは設定方法が大きく異なる。

Mac OSX Server

・最近は Mac OS でも動作するバイオインフォマティクスのソフトウェアが増えてきている。

・Linux コマンドが、標準でインストールされている、ターミナルというソフトウェアで実行可能

・マウスクリックでサーバの様々な設定が可能であるが、Linux とは異なる独自の設定方法の場合が多い。Linux と同じと思って設定ファイルを直接編集しても、設定がうまく変更できない場合もある。

・Mac OSX Server を利用しているコミュニティが小さいため、少し踏み込んだ設定をしたり、障害の原因を探ろうとした場合に、Web 上に情報が少なくて困ることが多い

Linux

・バイオインフォマティクスの様々なソフトウェアは Linux OS 専用のものが多い

・最近は多少はマウスクリックでサーバの設定も出来るようになってきたが、細かい設定はまだまだコマンドラインでの設定ファイルの編集が必要

・様々な種類の Linux ディストリビューションが存在（例: CentOS, Fedora, Ubuntu, openSUSE 等）

・スーパーコンピュータのほとんどの OS は Linux

［3］Linux ディストリビューションの選択

大きく分けて、Debian 系、Red Hat 系、Slackware 系、その他の 4 種類に分けられる。ソフトウェアをインストールする際の方法や、システムのディレクトリ階層などが異なる。今回は、サーバ構築の際に広く使われているオーソドックスなディストリビューションである、RedHat 系の CentOS を使用したサーバ構築について紹介する。

［4］実際の Linux OS のインストール＆初期設定

ここからは主にスライドで説明する。ただし、ネットワークの設定については、重要であるが複雑なため、補足資料として以下に記述しておく。

3. ネットワークの設計

(1). グローバル IP アドレスをサーバに設定して、サーバを外部のネットワークに直につなげる

(2). プライベート IP アドレスをサーバに設定して、サーバを外部のネットワークからは隔離する

グローバル IP アドレスをサーバに設定する場合には、所属する組織のネットワーク管理者等から、グローバル IP アドレスを取得する必要がある。大抵の場合、その際にサーバのホスト名（例: http://huga.hoge.ac.jp/の huga)をその組織の DNS サーバにグローバル IP アドレスと共に登録することになるので、サーバのホスト名も考えておく必要がある。組織の内と外のネットワークを隔てる Gateway に、ファイアーウォールが設置されている場合は、グローバル IP アドレスの申請の際にどの port を空けるのか申請する必要がある場合が多い。
一般的には、http (80 番)、https (443 番)、ssh (22 番)の port を空けることが多い。

26. ネットワークの設定

(1) ifcfg-eth0 の編集

CentOS の場合、GUI でネットワークの設定も可能であるが、一般的にはテキストファイルを編集することで、ネットワークの設定を行う。

cd /etc/sysconfig/network-scripts/

ネットワークインターフェイスが何個搭載されているかに依存するが、少なく

とも ifcfg-eth0 というファイルはあるはず。これが、サーバに設置されている 1 つめのネットワークインターフェイスの設定ファイルである。
中身は、
less ifcfg-eth0
DEVICE=eth0　#ネットワークインターフェイス名
HWADDR=いろいろ　#MAC アドレス
TYPE=Ethernet　#ネットワークのタイプ
UUID=いろいろ　#NetworkManager が割り振っているネットワークインターフェイスの ID
ONBOOT=no　#サーバ起動時にこのネットワークも起動するか否か
NM_CONTROLLED=yes　#NetworkManager での設定を反映するか否か
BOOTPROTO=dhcp　#IP アドレスの設定方法

となっているかと思うが、
このファイルを、赤字のように修正する。
vi ifcfg-eth0
DEVICE=eth0
HWADDR=いろいろ
TYPE=Ethernet
UUID=いろいろ
ONBOOT=yes
NM_CONTROLLED=no
BOOTPROTO=none
IPADDR="設定したい IP アドレス"
BROADCAST="設定したいブロードキャストアドレス"
NETMASK="設定したいサブネットマスク"
NETWORK="設定したい所属ネットワークのアドレス"
GATEWAY="設定したいデフォルトゲートウェイのアドレス"
それぞれの項目の詳細は、長くなるのでここでは説明しません。

(2) resolv.conf の編集
さらに、所属組織の DNS サーバの情報がある場合には、

/etc/resolv.conf　に記述する。

vi /etc/resolv.conf

nameserver 設定したい DNS サーバのグローバル IP アドレス

なお、サーバがプライベート IP アドレスを用いている場合は、上記に加えて /etc/hosts　に IP アドレスと結びつけたいホスト名を記述する。

(3) SELinux の停止

セキュリティ設定が厳しすぎて何かとサーバの運用上不都合な、セキュリティプログラムである SELinux を以下のコマンドで停止させる。

setenforce 0

さらに、以下のファイルを編集し、サーバ起動時に SELinux を自動起動しないように設定する。

vi /etc/selinux/config

ファイルの中身は、

SELINUX=enforcing

SELINUXTYPE=targeted

となっているので、

SELINUX=disabled

SELINUXTYPE=targeted

にする。

(4) NetworkManager の停止

GUI でネットワーク設定が出来る NetworkManager というプログラムが CentOS 6 には標準でインストールされているが、このプログラムは、ファイルを直接編集してネットワーク設定を変更した場合にその設定変更を強制的に書き換えてしまうため、テキストファイルを編集してネットワークの設定変更をしたい場合には、NetworkManager をサーバ起動時から停止させる必要がある。コマンドは以下。

service NetworkManager stop

chkconfig NetworkManager off

(5) 遠隔からの root でのログインの禁止

不正侵入を防ぐため、遠隔からの root でのログインを禁止する。

vi /etc/ssh/sshd_config

#PermitRootLogin yes

を

PermitRootLogin no

にする。

その後、以下のコマンドで ssh を再起動。

/etc/init.d/sshd restart

あとは、iptables の設定を変更してファイアーウォールを構築し、セキュリティを高めれば、初期設定は完了。

参考文献

・中井悦司　プロのための Linux システム構築・運用技術　技術評論社 2011
Linux サーバをインストールし管理したことがある中級者向けの本。

・中島能和　Linux サーバーセキュリティ徹底入門　翔泳社 2013
Linux サーバを運用する際のセキュリティ設定方法について書かれた本。