大規模な生物データ解析のためのUNIX入門

基礎生物学研究所
生物機能解析センター

大規模データ解析にUNIX が適している理由

- UNIX でしか使えないアプリケーションがある
  - 最新の研究用ソフト、並列化や巨大メモリに対応したソフト
- 巨大なテキストファイルの扱いが容易
  - テキストファイル処理のための豊富なコマンド群
- 多数の処理を一度に行なうバッチ処理に適している
  - シェルスクリプトを用いたコマンドの実行
- プログラムの開発環境として優れている
  - Perl, Ruby等のスクリプト言語、豊富な開発ユーティリティ
- 高性能なサーバ計算機では標準的なOS
  - サーバとしての高い安定性、データベースやサーバ管理などの標準的なフリーウェアの充実

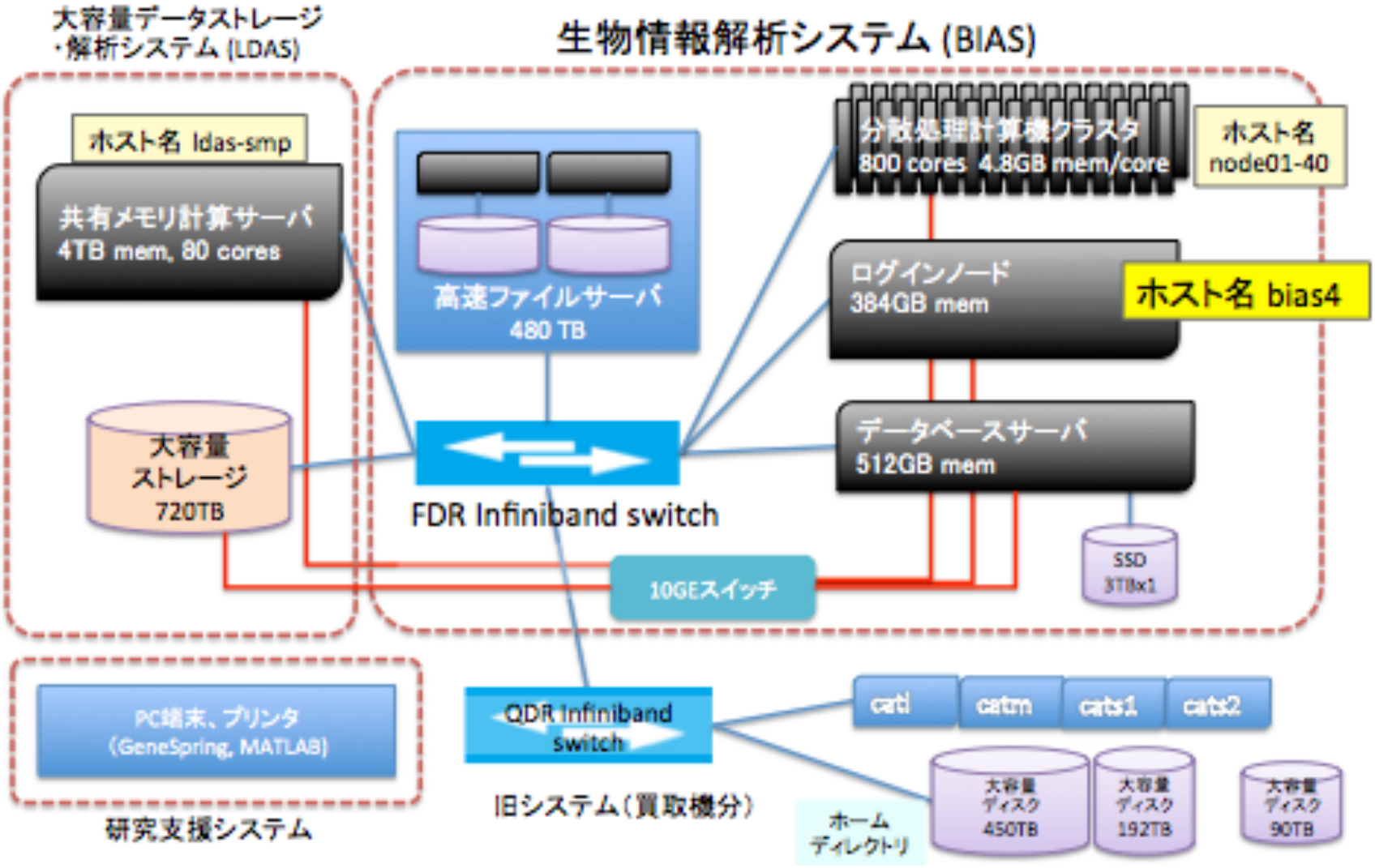
生物情報解析システムの紹介

Biological Information Analysis System

分散処理計算機クラスタ
SGI Rackable server C2112-4RP
Intel Xeon (2.8GHz) 20core/node
96GB/node Memory, 40node, 800core

高速ファイルサーバ
DDN SFA7700
Lustre file system : 480TB

生物情報解析システムの構成



biasへのログインとログアウト

ターミナルを開いて

```
$ ssh ユーザ名@bias4.nibb.ac.jp
The authenticity of host 'nibb.ac.jp (133.48.32.2)' can't be established.
RSA key fingerprint is 22:43:98:7c:b0:9a:20:f4:e4:71:95:64:44:60:a3:2c.

Are you sure you want to continue connecting (yes/no)?  yes

username@nibb.ac.jp's password: パスワード
[username@bias4-login ~]$
```

bias4にログインできているかはここで確認

```
[username@bias4-login ~]$ exit （ログアウト）
```

生物情報解析システムWiki

http://www.nibb.ac.jp/cproom/wiki2/

生物情報解析システム
メインページ
コミュニティ・ポータル
最近の出来事
最近の更新
おまかせ表示
ヘルプ
ツール
リンク元
関連ページの更新状況
特別ページ
印刷用バージョン
この版への固定リンク
ページ情報

メインページ　議論

生物情報解析システム

基礎生物学研究所　生物機能解析センター　情報管理解析室

目次 [非表示]
1 概要
2 障害・障害復旧情報
3 停止予定
4 システム構成
5 利用手続き
6 使い方
7 キュー利用状況

概要

- 生物情報解析システム Version4 は、2014年1月から運用中です。
  - **「共有メモリ型計算サーバ」メインメモリ 4TB**　（メモリが多く必要な解析に）
  - **「分散処理用計算機クラスタ」800コア**　（CPUを沢山使って分散処理したい解析に）
  - 次世代シーケンサデータ解析向け計算サーバとして、メインメモリ 512GB、1TBを持つ計算機も提供されています。
- 基礎生物学研究所内の方はもちろん、所外の研究者の方にもお使いいただけます。ユーザーアカウントの取得方法
- 利用資格などは内規をご覧ください。　生物情報解析システム:利用内規
- 全てのアプリケーション、データベース/ホームディレクトリを含めた全てのディスク領域は、**全計算機から同様に**使うことができます。
- システムをお使いいただくには「**UNIXコマンドの知識**」があることが前提になります。デスクトップ画面はありません。

障害・障害復旧情報

- 現在大容量ストレージ上にある「~/save」領域が故障のため使えません。鋭意修復作業中です。ご迷惑をおかけして申し訳ありません（2017/04/26~）

停止予定

- 現在停止予定日はありません

システム構成

- マシン構成
- 分子生物学データベース
- ソフトウェア

生物情報解析システムで利用可能なソフトウェア

次世代シーケンサ解析

- マッピング
  - Bowtie, BWA, SOAP
- RNA-Seq解析
  - TopHat, Cufflinks, Trinity
- アセンブラ
  - Velvet, ABySS, AllPaths-LG
- ユーティリティ
  - samtools, bamtools, BEDtools, cutadapt SRA toolkit

その他のツール

- ホモロジー検索
  - BLAST, FASTA
- 遺伝子予測
  - GeneMark, GenScan, Augustus
- ゲノムアライメント
  - lastz, MUMmer, BLAT
- マルチプルアライメント
  - ClustalW, Muscle, MAFFT
- 系統樹解析
  - Phylip, PhyML, MrBayes
- モチーフ解析
  - InterProScan, HMMER, MEME
- データベース検索
  - DBGET
- 統合配列解析
  - EMBOSS

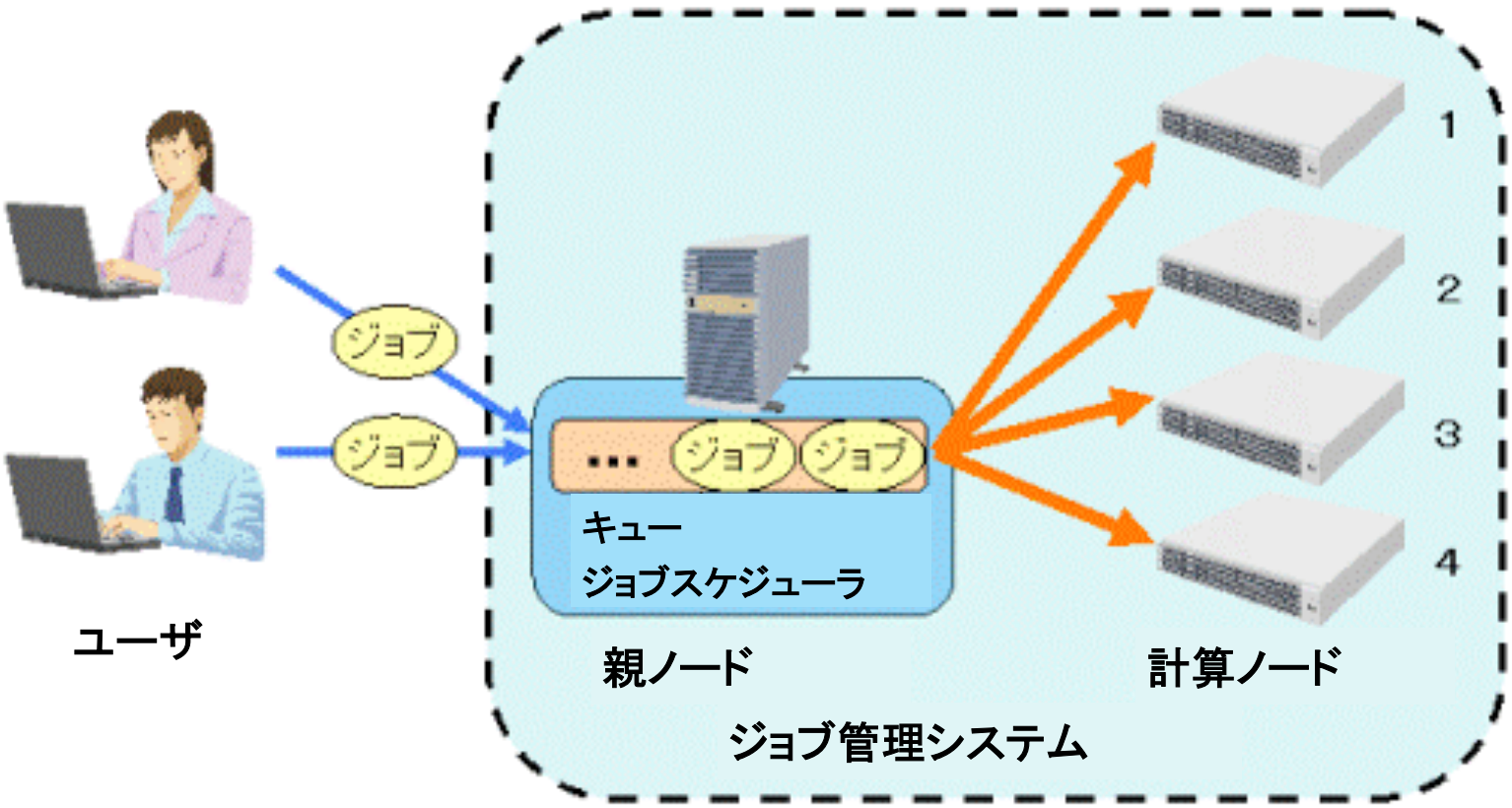
分子生物学データベース

データベース名	アクセス名	説明	フォーマット	更新型
GenBank	genbank	核酸塩基配列	DBGET	定期
GenBank-upd	genbank-upd	核酸塩基配列	DBGET	日々
EMBL	embl	核酸塩基配列	DBGET	定期
EMBL-upd	embl-upd	核酸塩基配列	DBGET	日々
RefSeq	refseq	refnuc + refpep	DBGET	日々
RefSeq Nuc.	refnuc	核酸塩基配列	DBGET	日々
RefSeq Pep.	refpep	タンパク質アミノ酸配列	DBGET	日々
RefSeq Protein	refseq_protein	タンパク質アミノ酸配列	FASTA, BLAST	日々
RefSeq Genomic	refseq_genomic	ゲノム配列	FASTA, BLAST	日々
RefSeq RNA	refseq_rna	RNA塩基配列	FASTA, BLAST	日々
EST_human EST_mouse EST_others	est_human est_mouse est_others	核酸塩基配列	FASTA, BLAST	定期
NCBI nr-nt	nt	非冗長核酸塩基配列	FASTA, BLAST	定期
gss	gss	核酸塩基配列	FASTA, BLAST	定期
HTGS	htgs	核酸塩基配列	FASTA, BLAST	定期
dbsts	dbsts	核酸塩基配列	FASTA, BLAST	定期
patnt	patnt	核酸塩基配列	FASTA, BLAST	定期
env_nt	env_nt	核酸塩基配列	FASTA, BLAST	定期
pdbnt	pdbnt	核酸塩基配列	FASTA, BLAST	定期
NCBI nr-aa	nr	非冗長アミノ酸配列	FASTA, BLAST	定期
UniProt	uniprot	TrEMBL + Swissprot	DBGET, FASTA, BLAST	日々
TrEMBL	trembl	タンパク質アミノ酸配列	DBGET, FASTA, BLAST	日々
Swissprot	swissprot	タンパク質アミノ酸配列	DBGET, FASTA, BLAST	日々
pataa	pataa	タンパク質アミノ酸配列	FASTA, BLAST	定期
env_nr	env_nr	タンパク質アミノ酸配列	FASTA, BLAST	定期
pdbaa	pdbaa	タンパク質アミノ酸配列	FASTA, BLAST	定期
PDB	pdb	タンパク質立体構造	DBGET	定期

- ホモロジー検索用データベース(BLAST/FASTA)
- オリジナル形式のフラットファイル (DBGET)
- KEGG データベース
- 次世代シーケンサデータ解析用データベース (Illumina iGenomes)
- モチーフ解析用データベース(InterproScan)

ジョブ管理システム

ユーザが要求した計算処理（ジョブ）を、計算ノードに順に割り振って実行させる



bias4では、SunGridEngine というジョブ管理システムが導入されており、
解析のための計算は、必ずジョブ管理システムを通して実行する。

本コースの狙い

基生研のサーバ計算機にログインして、UNIXで大規模なデータ解析を行うための基礎を習得する

- **UNIX基本コマンド**　UNIXでファイル操作などを行う基本的なコマンドや、複数のコマンドを組み合わせて実行するパイプなどの技法を習得し、UNIX上のコマンドベースの操作に慣れる。
- **次世代シーケンサ解析コマンド**　基本的なNGSデータ解析コマンドを題材として、UNIXでデータ解析を行う感覚をつかむ。
- **テキスト処理**　ちょっとしたデータ解析や、解析の前処理や後処理を行うのに便利なUNIXのテキスト処理コマンドを習得する。
- **シェルスクリプト**　一連のコマンドを一度に実行するためのスクリプトの基本的な書き方を習得する。
- **ジョブ管理システムの使い方**　クラスター計算機上で並列にジョブを実行する方法を習得する。

コーススケジュール

- **6月22日（木）**
- 13:00-13:30　はじめに [内山]
- 13:30-16:30　UNIX基本コマンド　[三輪]
- 16:30-17:00　（休憩・復習タイム）
- 17:00-18:00　エディタとシェルスクリプト　[中村]

- **6月23日（金）**
- 08:30-09:00　（開場・復習タイム）
- 09:00-10:30　NGS基本ツール　[内山]
- 10:30-11:30　SunGridEngine使用方法　[西出]
- 11:30-12:00　UNIXによるテキストファイル処理（前編）　[中村]
- 12:00-13:00　（昼休憩）
- 13:00-13:30　UNIXによるテキストファイル処理（後編）　[中村]
- 13:30-15:00　シェルスクリプト２　[西出]
- 15:00-17:00　実践演習

講師

- 講師
  - 内山郁夫　情報管理解析室・助教
  - 三輪朋樹　技術課長
  - 西出浩世　情報管理解析室・技術職員
  - 中村貴宣　情報管理解析室・技術職員

- 事務局・フロアスタッフ
  - 尾納隆大　生物機能情報分析室・技術職員

コースページ

https://github.com/nibb-unix/gitc201706-unix/wiki