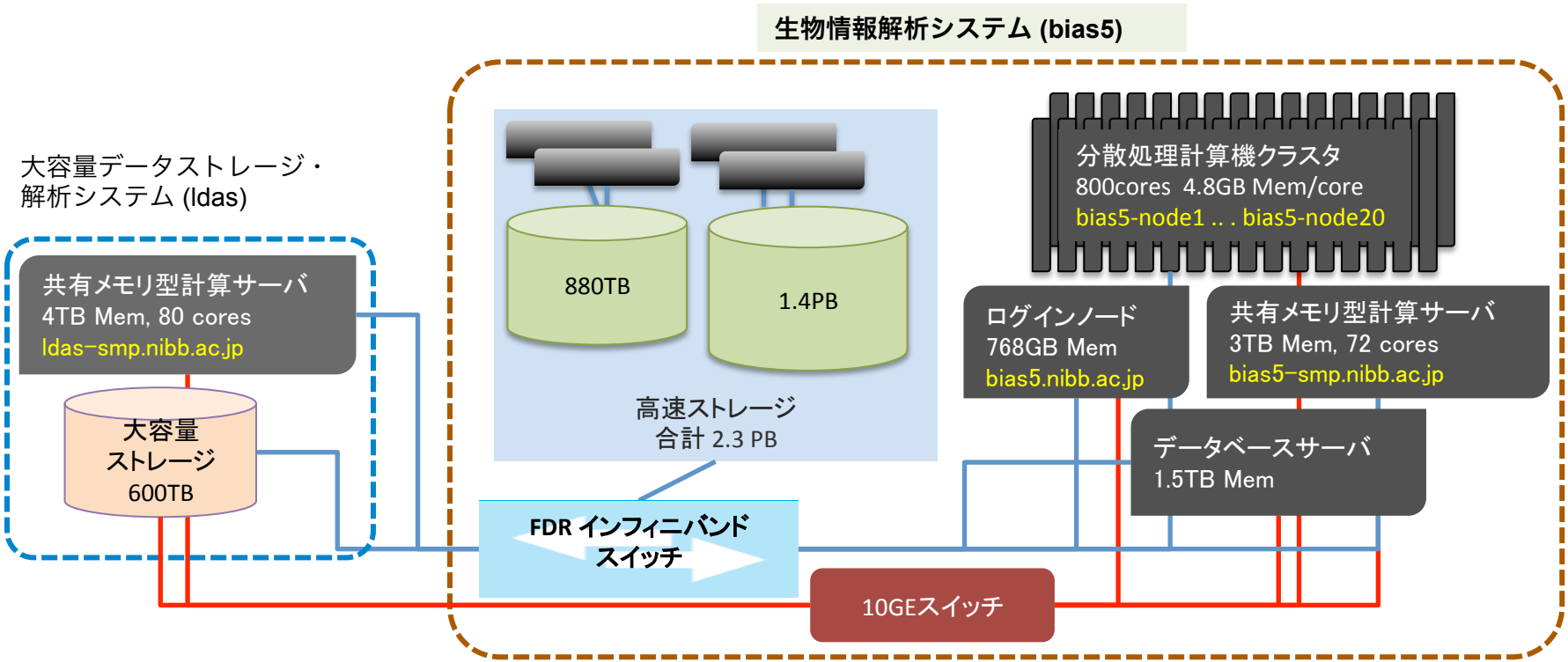
生物情報解析システム
新システム(BIAS5)の概要

情報管理解析室
内山　郁夫

新・生物情報解析システムの構成



ログインノード

ホスト名　bias5

HP ProLiant DL360 Gen10

CPU: Intel Xeon Gold 6138 (2.1 GHz) 32 cores

Memory: 768 GB

主な用途：
ジョブの実行
プログラムの作成
ログインして利用するが、プログラムはqsub 経由で実行

分散処理用計算機クラスタ

ホスト名 bias5-node01
– bias5-node20

HPE Apollo r2800

CPU: Intel Xeon Gold 6138 (2.0GHz) 40 cores/node
Memory: 192GB/node (4.8GB/core)
Total: 20 nodes, 800 cores

主な用途:
大規模な並列計算
直接ログインせず qsub 経由で利用

キュー:
small, medium, large

共有メモリ計算サーバ

ホスト名　bias5-smp

HP ProLiant DL560 Gen10

CPU: Intel Xeon Gold 6140 (2.3 GHz) 72 cores
Memory: 3TB

主な用途:
大きなメモリを使う計算
直接ログインせずqsub 経由で利用

キュー:
smps, smpm, smpl

高速ストレージシステム

DDN SFA7700X

実効容量 A) 880TB + B) 1.52PB
並列分散ファイルシステム
GPFS

主な用途:
ホームディレクトリ(クオータ制限あり)
スクラッチ領域(クオータ制限なし)
共通データベース
プロジェクト領域(要相談)

Type A

Type B

大容量データ解析システム(LDAS)
共有メモリ計算サーバ

ホスト名　ldas-smp

HP ProLiant DL980 G7

CPU: Intel Xeon (2.4 GHz) 80 cores
Memory: 4TB　(50GB/core)

主な用途:
大きなメモリを使う計算
直接ログインせずqsub 経由で利用

キュー:
smps, smpm, smpl

大容量データ解析システム(LDAS)
大容量ファイルサーバ

DELL PowerVault MD3660f

DELL PowerEdge R620

物理容量 720TB

実効容量 596TB

並列分散ファイルシステム
GlusterFS

主な用途:

頻繁に利用しないデータの格納
（現在サービス休止中）

生物情報解析システムで利用可能なバイオ関連ソフトウェア（一部）

次世代シーケンサ解析

- マッピング
  - Bowtie, BWA, SOAP
- RNA-Seq解析
  - TopHat, Cufflinks, HISAT2, StringTie, 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, Diamond
- 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

module ツールによって複数のバージョンが管理されている

利用可能なバイオ関連データベース

項番	データベース	概要	フォーマット	更新型
1	GenBank/GenBank-upd	核酸塩基配列	フラット, DBGET	定期/日々
2	EMBL/EMBL-upd	核酸塩基配列	フラット, DBGET	定期/日々
3	RefSeq/RefSeq-upd	核酸塩基配列	フラット, DBGET, FASTA, BLAST	定期/日々
4	EST_human/EST_mouse/EST_others	核酸塩基配列	FASTA, BLAST	定期
5	NCBI nr-nt	非冗長核酸塩基配列	FASTA, BLAST	定期
6	gss	核酸塩基配列	FASTA, BLAST	定期
7	HTGS	核酸塩基配列	FASTA, BLAST	定期
8	dbsts	核酸塩基配列	FASTA, BLAST	定期
9	patnt	核酸塩基配列	FASTA, BLAST	定期
10	env_nt	核酸塩基配列	FASTA, BLAST	定期
11	pdbnt	核酸塩基配列	FASTA, BLAST	定期
12	NCBI nr-aa	非冗長アミノ酸配列	FASTA, BLAST, DIAMOND	定期
13	RefSeq-protein	タンパク質アミノ酸配列	フラット, DBGET, FASTA, BLAST, DIAMOND	定期
14	UniProt(TrEMBL. Swissprot)	タンパク質アミノ酸配列	フラット, DBGET, FASTA, BLAST, DIAMOND	日々
15	pataa	タンパク質アミノ酸配列	FASTA, BLAST	定期
16	env_nr	タンパク質アミノ酸配列	FASTA, BLAST	定期
17	pdbaa	タンパク質アミノ酸配列	FASTA, BLAST	定期
18	PDB	タンパク質立体構造	FASTA, BLAST	定期
19	kegg	遺伝子/ゲノム統合データベース	フラット, DBGET, FASTA, BLAST, DIAMOND	定期

ジョブ管理システムPBS

- 多数の処理要求（ジョブ）を受け付けて管理し、CPUやメモリなどの計算資源を適切に割り当てて順次実行させる仕組み。
- 今回のシステムは PBS (Portable Batch System)が導入されている。
- 基本的にすべての計算は、ジョブ管理システムを通して実行する。
- 計算の規模や種類によって複数のキューが用意されている。

キューの設定

	分散処理用計算機クラスタ(bias5-node01～node20)			共有メモリ計算サーバ(bias5-smp)		
キュー名	small(default)	medium	large	smps	smpm	smpl
ジョブの特徴	短時間・並列多	中規模	長時間	中メモリ	大メモリ	最大メモリ
ジョブ実行サーバ	bias5-node01～ bias5-node20	bias5-node01～ bias5-node20	bias5-node01～ bias5-node20	bias5-smp (ldas-smp)	bias5-smp (ldas-smp)	bias5-smp (ldas-smp)
最大実行時間	6時間	72時間	no limit	no limit	no limit	no limit
ジョブで利用できる最大メモリ	96GB	96GB	96GB	500GB	1TB	3TB
キューで利用できるジョブ数	no limit	no limit	no limit	6	3	1
キューで利用できるCPU数	580	200	20	48	48	36
ユーザー人当たりのCPU数(サーバ全体)	480					
ユーザー人当たりのCPU数(キューあたり)	400	150	10	no limit	no limit	no limit
デフォルトCPU数	1	1	1	1	1	1
デフォルトメモリサイズ	3GB	3GB	3GB	250GB	500GB	1500GB

システム利用上の注意

- まずログインノード bias5.nibb.ac.jp にログインする。
- ログインノードではプログラム作成などCPUを消費しない処理のみを行い、ジョブの実行は必ずジョブ管理システムを通して行う。
- 分散処理クラスタ、および共有メモリサーバを使うには、それぞれの専用のキューにジョブをサブミットする。
- ホームディレクトリはクオータ制限（デフォルト3TB）をかける予定。
- 一時的に大量のディスクを使う場合はクオータ制限がかかっていないスクラッチ領域（/scratch/ユーザ名） を使う。ただし、スクラッチ領域のファイルは1ヶ月後に消去される。
- バックアップやアーカイブとして長期的に保存したいファイルはSave領域に保存する（現在は利用不可）。
- 利用法に関する詳しい情報は、専用ページ http://www.nibb.ac.jp/cproom/wiki で確認する。