

# 塩基配列データベースを中心とした 生物学系研究の情報基盤について

国立遺伝学研究所  
DDBJセンター 助教  
小笠原 理

# 国立遺伝学研究所

HOME > 遺伝研について

## 遺伝研について



大学共同利用機関法人 情報・システム研究機構

## 国立遺伝学研究所



住所：〒411-8540 静岡県三島市谷田1111

TEL番号：055-981-6707（管理部）

FAX番号：055-981-6715

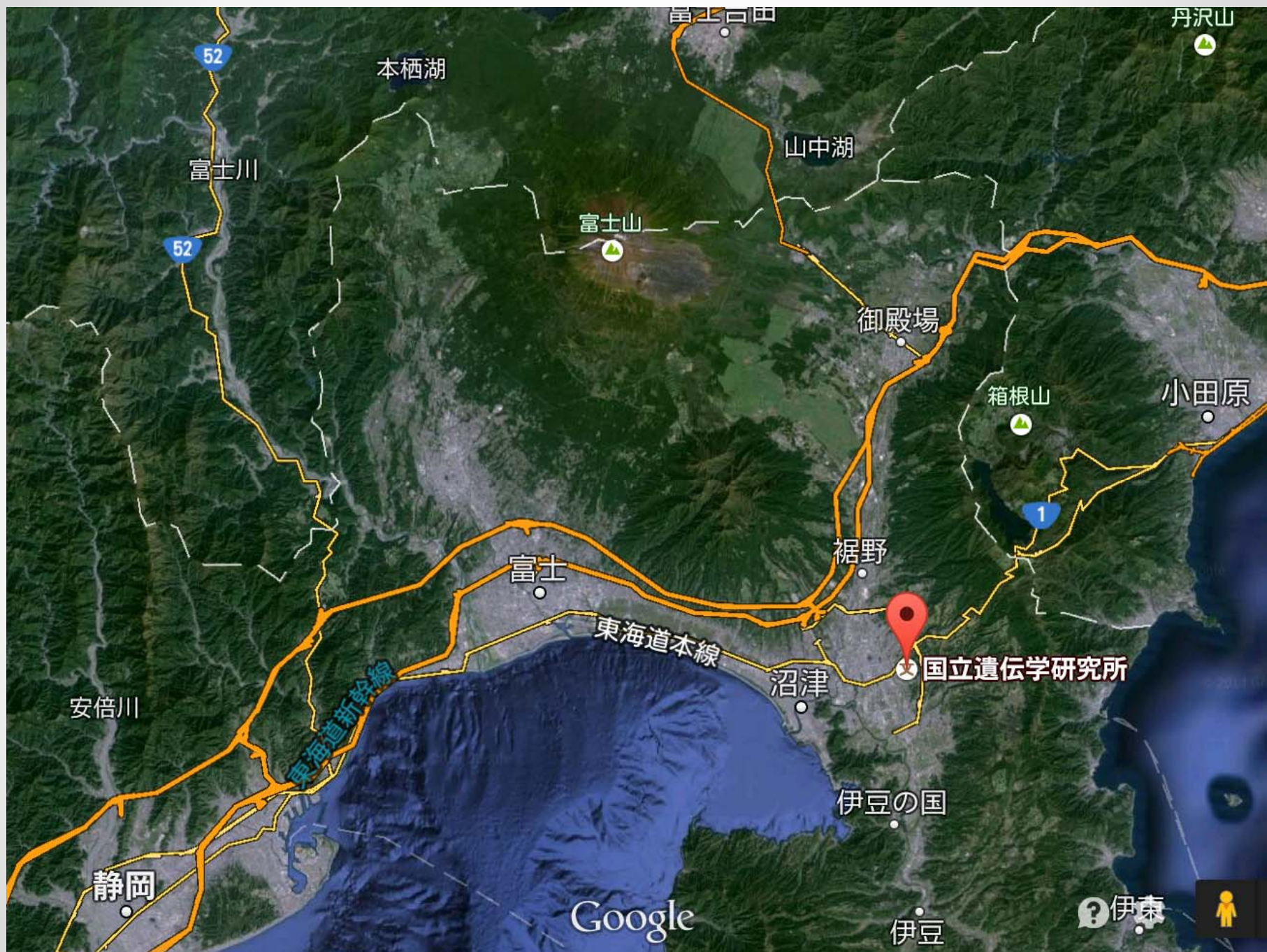
最寄駅：JR三島駅

三島駅～国立遺伝学研究所

距離 約4km，バスまたはタクシーで約20分

[アクセス方法や地図はこちら](#)





# 遺伝研スパコン：目的

1. 国際塩基配列データベース (International Nucleotide Sequence Database : INSD)の構築
2. 大学共同利用機関として、主に生物学医学系研究者への計算機資源の提供





# 遺伝研スパコン：構成概略

## 1. 分散メモリ型計算ノード (Thin)

- 554ノード 9672コア
- HP Proliant SL230 Gen8など
- XeonE5-2670, XeonE5-2680v2
- 64GB memory/node
- GPGPU搭載128ノード
- XeonPhi搭載32ノード

## 2. 共有メモリ型計算ノード (Medium)

- 10ノード 800コア
- HP Proliant DL980 G7
- 2TB memory/node

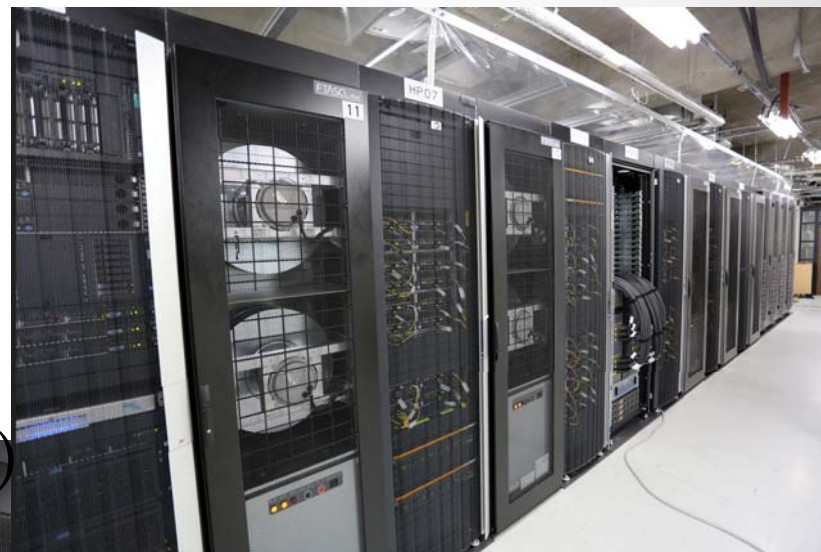
## 3. 共有メモリ型計算ノード (Fat)

- 1ノード 768コア
- SGI Altix UV1000
- 10TB memory/node

## 4. 計算用ストレージ Lustre File system 7PB

## 5. 配列データアーカイブ用ストレージ (MAID) 5.5PB

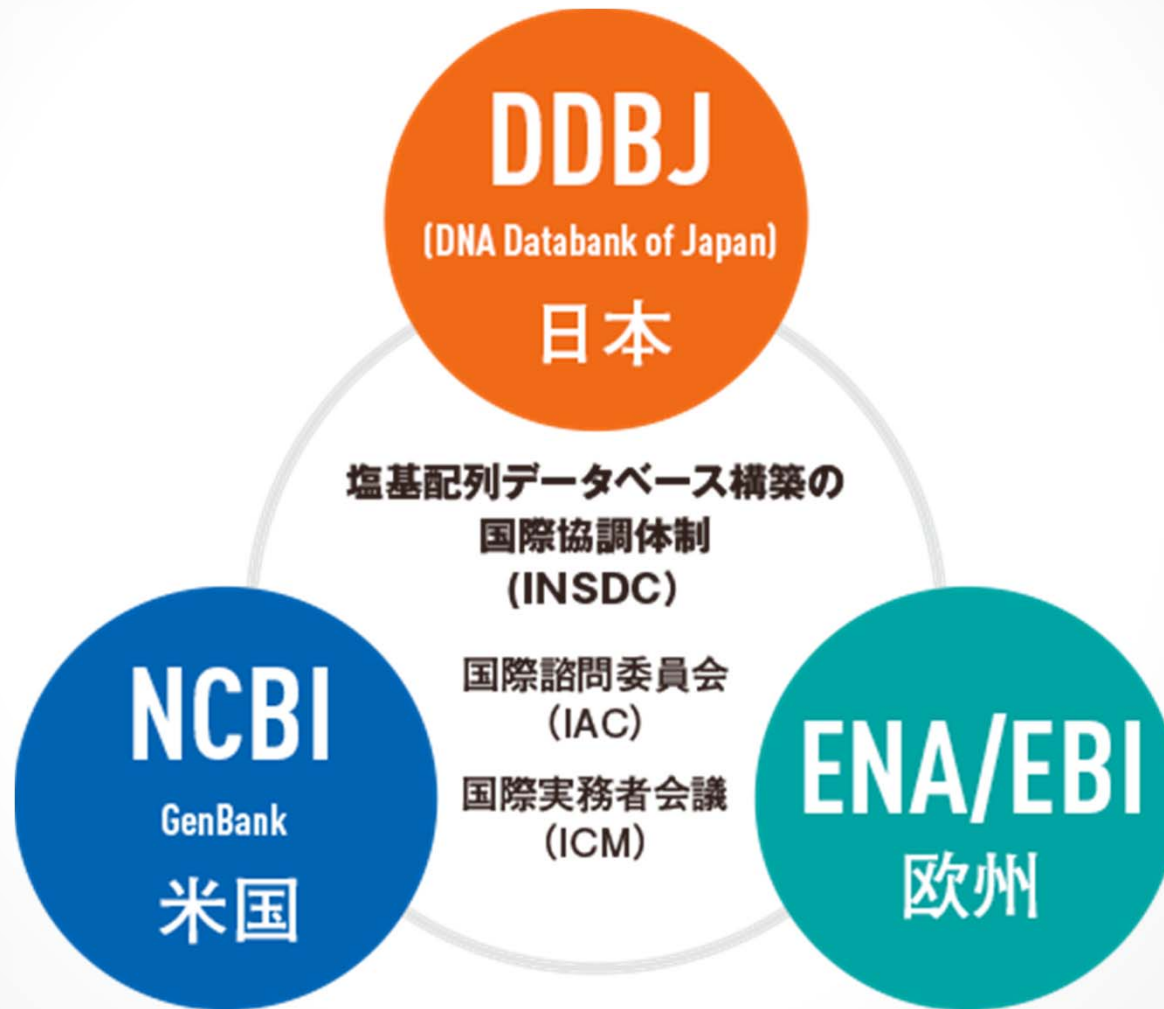
## 6. Infiniband QDR, FDR



# 国際塩基配列データベース(INSDB)と DDBJ (DNA Data Bank of Japan)の沿革(1/3) Traditional Data Archive

- 1977年 DNA塩基配列決定法 (Sanger法)
- 1979年 国際塩基配列データベース(INSDB)設立のための国際会議が開かれる
- 1981年 EMBL(欧) 1982年 GenBank (米) 設立
- **1987年 遺伝研にDDBJ設置。日米欧 3 極体制の確立**
- 1990 年 論文発表時にデータを登録することを著者に義務づける現在の運用形態が確立。

# 国際塩基配列データベースコンソーシアム (INSDC) = DDBJ (日) + NCBI(米) + EBI (欧)

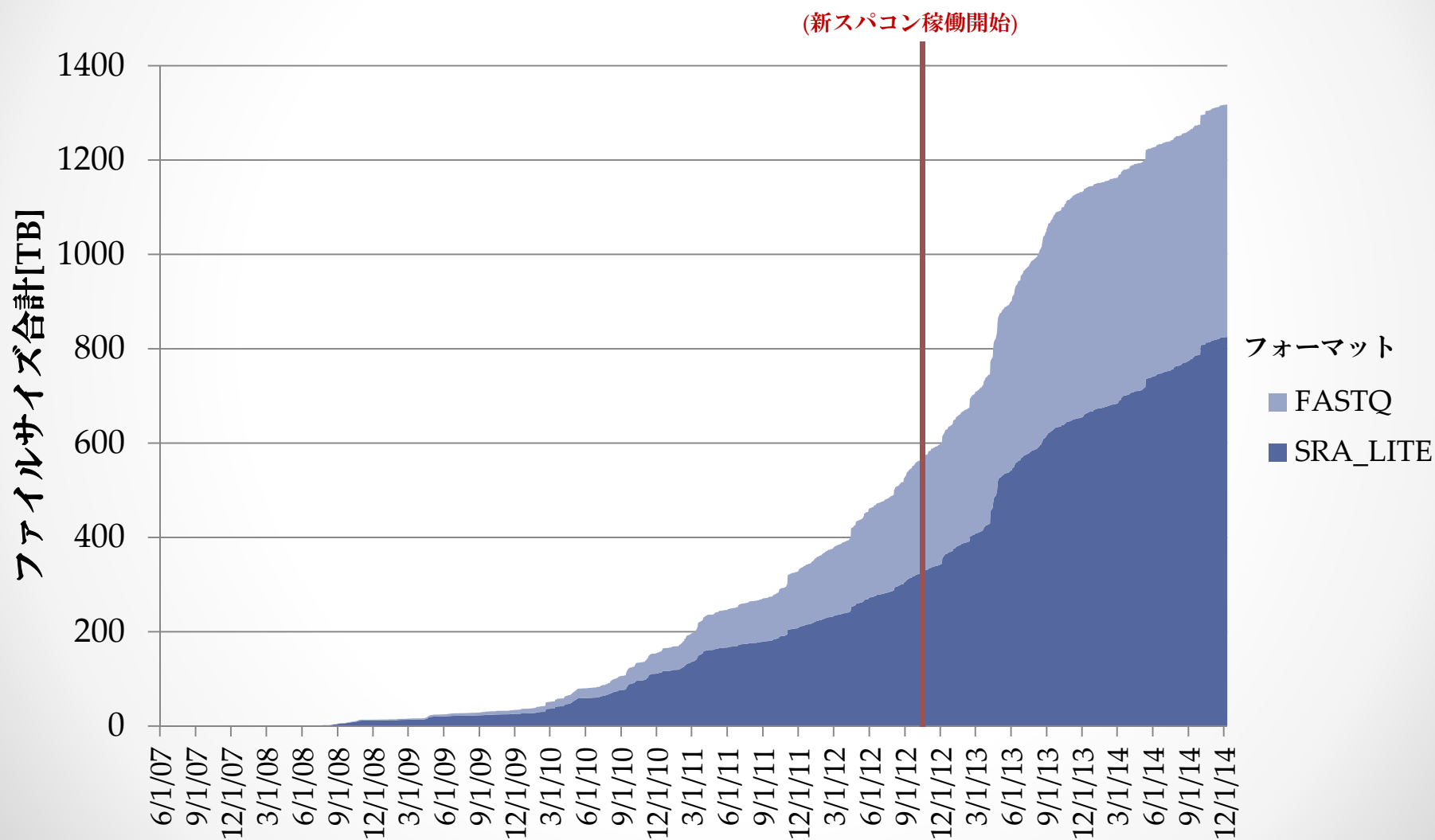


# 国際塩基配列データベース(INSD)と DDBJ (DNA Data Bank of Japan)の沿革(2/3) NGS Sequence Read Archive (Open Access Data)

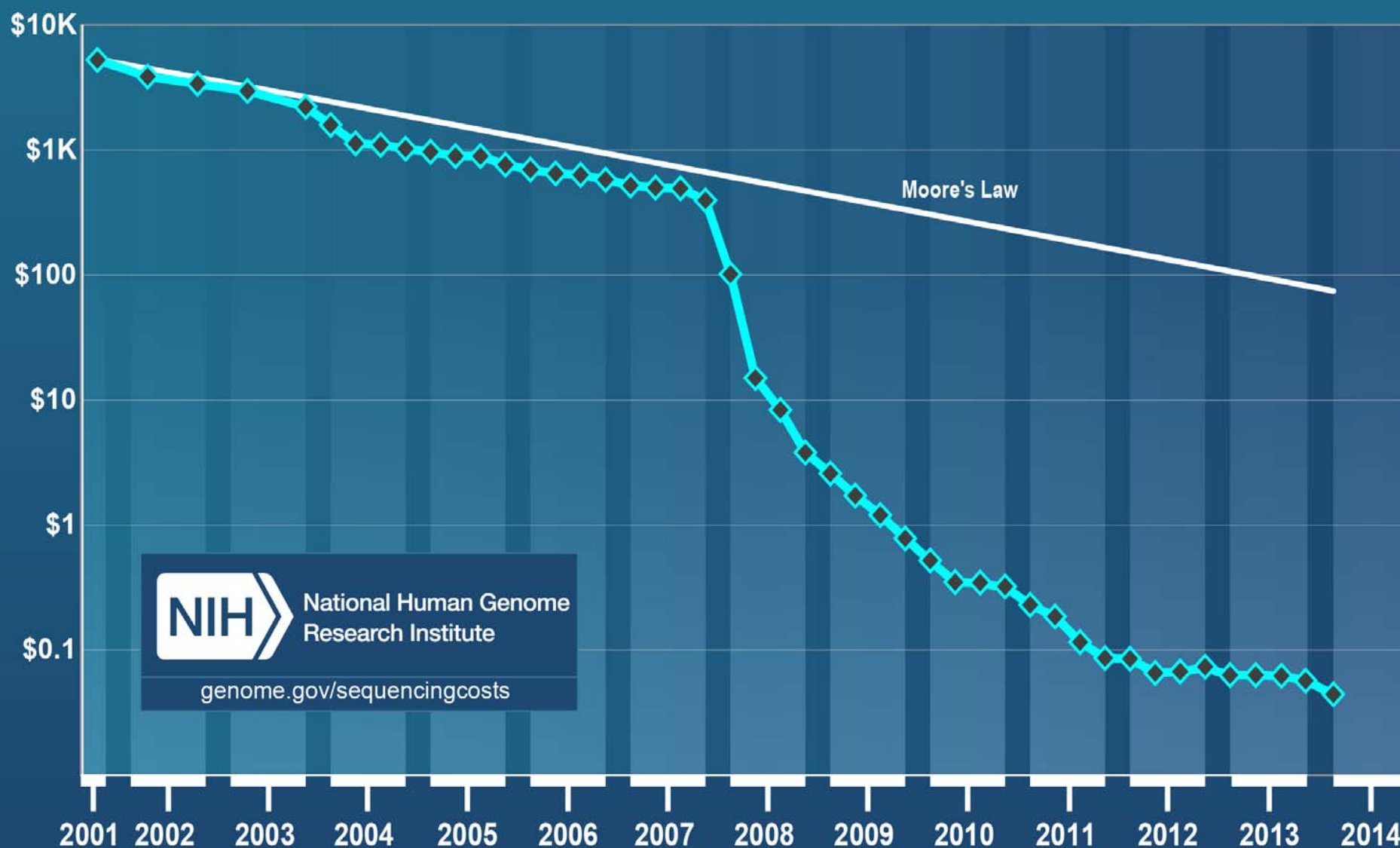
- 2001年 ヒトゲノムプロジェクトドラフト配列公開
- 2003年 ヒトゲノムプロジェクト終了
- 2005年 最初の次世代シーケンサー Roche454 GS FLX 販売開始
- 2009年 次世代シーケンサー(NGS)のアーカイブを日米欧3極で開始
- 2012年 3月 遺伝研スーパーコンピュータ更新 (第四期)



# 次世代シーケンサー(NGS) データアーカイブのファイルサイズ合計



## Cost per Raw Megabase of DNA Sequence



# 国際塩基配列データベース(INSD)と DDBJ (DNA Data Bank of Japan)の沿革(3/3) NGS Sequence Data (Controlled Access Data)

- 2009年 NGSのデータアーカイブを日米欧3極で開始
- 2012年 3月 遺伝研スーパーコンピュータ更新（第四期）
- 2013年 個人ゲノム情報のアーカイブの運用を日米欧3極で開始。

日本では遺伝研DDBJが個人ゲノム情報のアーカイブ  
Japanese Genotype-phenotype Archive (JGA) をJST傘下の  
NBDC と協同で開始



[ホーム](#)[データの利用](#)[データの提供](#)[ガイドライン](#)[NBDCヒトデータ審査委員会](#)

## データの利用

NBDCヒトデータベースにおいて提供しているヒトに関するデータの利用手続きは以下の通りです。

### オープンデータ（利用可能な研究データ一覧のアクセス制限欄に”オープン”と表記。）

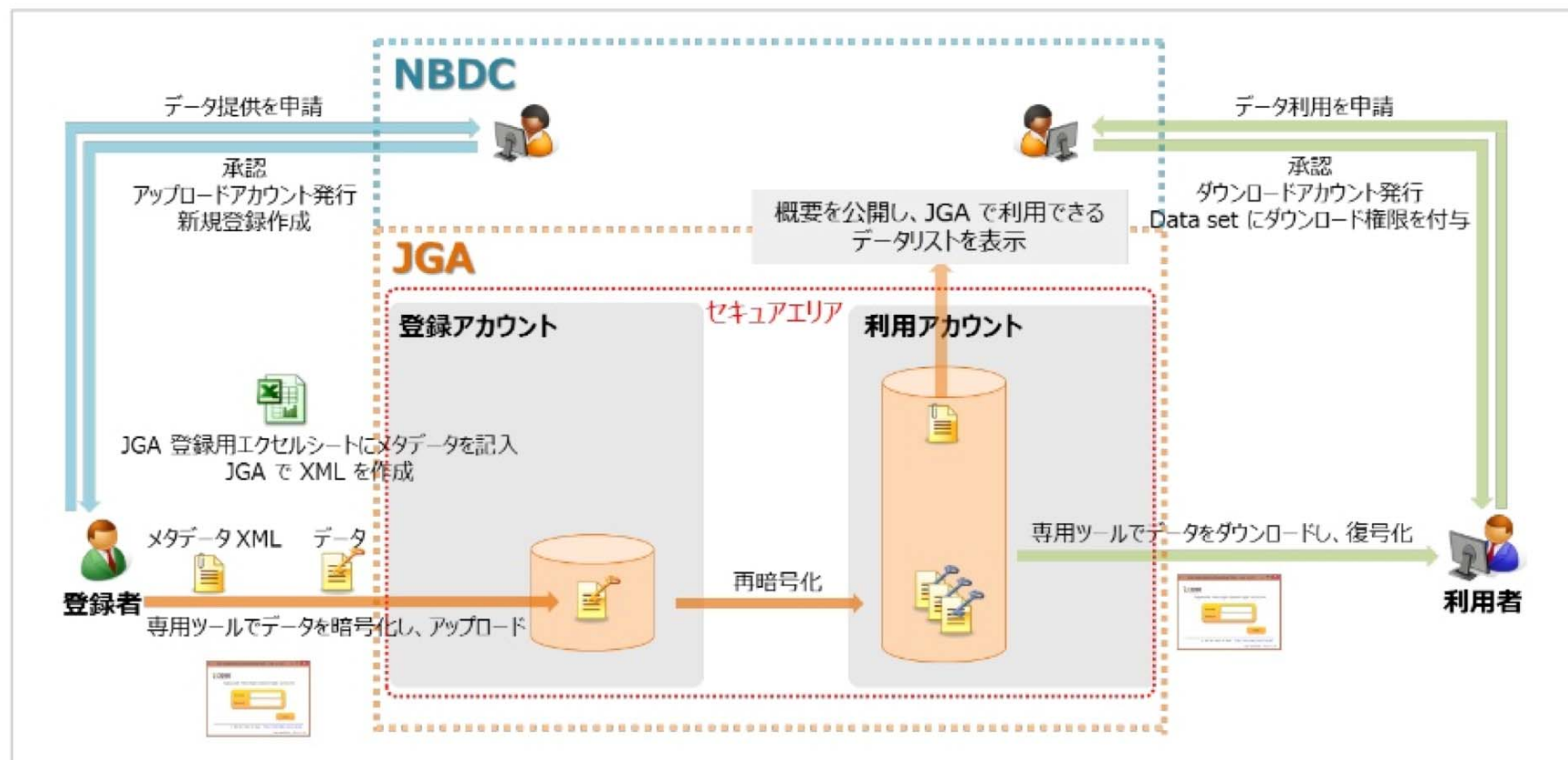
1. [NBDCヒトデータ共有ガイドライン](#)を熟読して下さい。
2. [利用可能な研究データ一覧](#)から各研究内容を確認（Research IDをクリック）のうえ、自由にデータをダウンロードして下さい。

### 制限公開データ（利用可能な研究データ一覧のアクセス制限欄に”制限”と表記。データの利用には利用申請が必要。）

利用申請方法は以下の通りです。

1. [NBDCヒトデータ共有ガイドライン](#)を熟読して下さい。
2. [利用可能な研究データ一覧](#)から各研究内容を確認（Research IDをクリック）のうえ、内容を確認後、利用したいデータセットのデータIDを控えて下さい。データ利用申請時に必要になります。
3. データ利用申請のために必要な以下の情報を準備して下さい。
  - 研究代表者および研究分担者全員の氏名、所属情報 ※一回の申請で同一所属機関の研究分担者を30名まで登録可能です。
  - 研究代表者によって利用申請される研究内容に関連した論文の情報

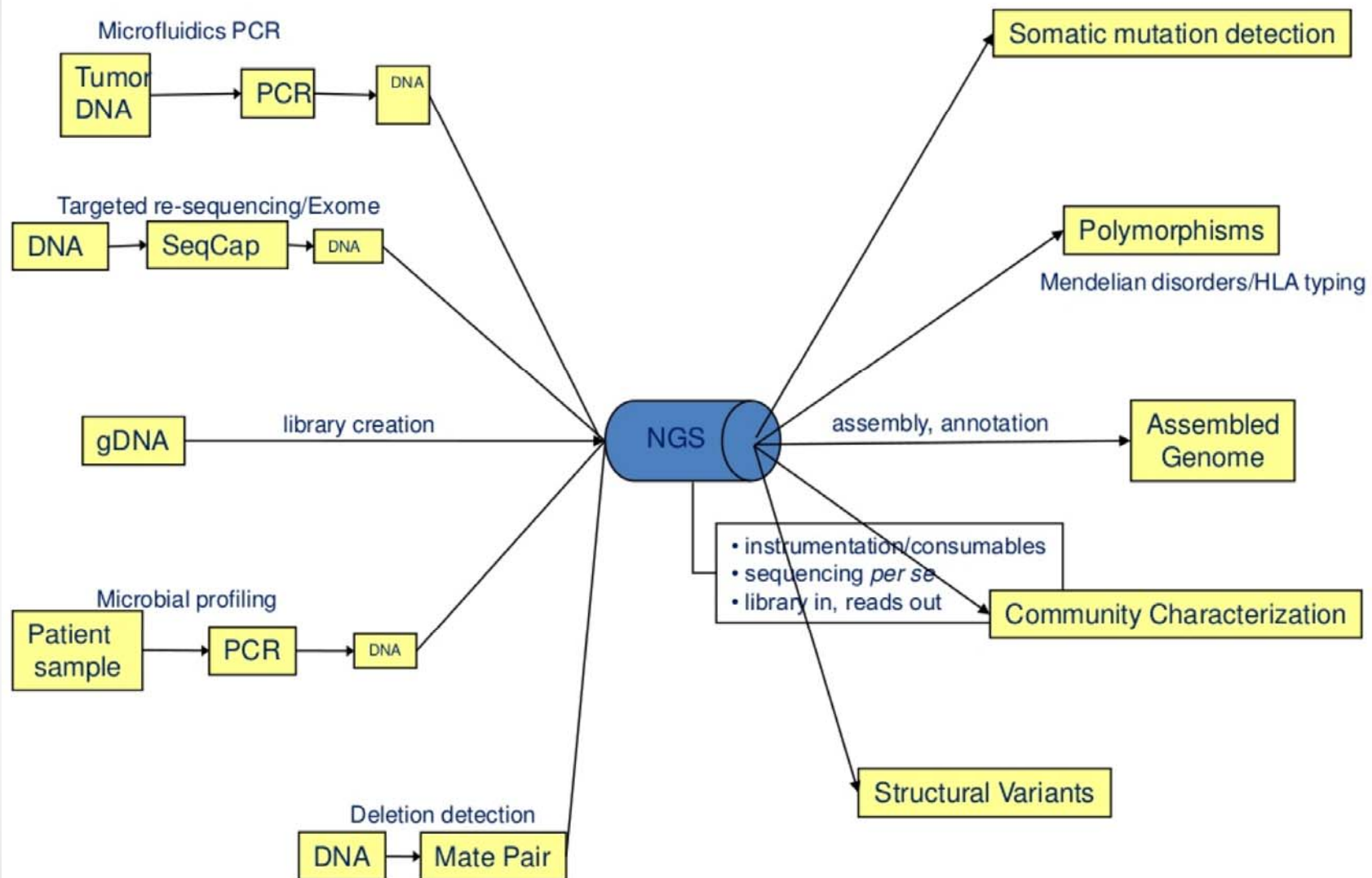
# DDBJ と NBDC の役割分担



- ✓ 利用制限が必要な個人ゲノムデータを受入・保存・提供
- ✓ 匿名化されたメタデータのみ受付

- ✓ NBDC ヒトデータ共有ガイドラインに従い、データの登録と利用を審査

# FUTURE APPLICATIONS OF NGS



## Clinical Applications of Next Generation Sequencing

SlideShare : <http://www.slideshare.net/lampweb/clinical-applications-of-next-generation-sequencing>

# 国際塩基配列データベース(INSDB)

## データの種類

| データベース                                      | 件数     | サイズ    |
|---------------------------------------------|--------|--------|
| <b>DNA塩基配列データ (Traditional)</b>             |        |        |
| DDBJ/ENA/GenBank                            | 3 億件   | 10TB   |
| The Trace Archive                           | 21.7億件 | 100TB  |
| <b>DNA塩基配列データ (次世代シーケンサー)</b>               |        |        |
| DRA/ERA/SRA                                 | 120万件  | 1500TB |
| JGA/EGA/dbGaP Personal Genome Data Archives | 26万人   | 500TB  |
| <b>遺伝子発現データ</b>                             |        |        |
| DOR/Array Express/GEO                       | 125万件  | 20TB   |



# 1. DDBJが受け付けるデータの登録

1. ログイン、ユーザー情報の入力・編集・管理
2. BioSample, BioProjectのデータ入力 (メタデータ)
3. 配列生データのアップロード(DRA, Trace Archive)
4. (アノテーションの計算)
5. Traditionalデータの登録

# 2. 査定業務 (TSUNAMI)、公開管理 => DDBJ分のDaily update作成 (FFmaker)

3. NCBI, EBIのDailyデータとのマージ
4. リリース作成プログラム
5. 公開系 (全文検索システム等)

| Rel.(YYYY/MM)       | DDBJ       | JPO       | KIPO    | EMBL       | EPO        | GenBank     | USPTO     | TOTAL       |
|---------------------|------------|-----------|---------|------------|------------|-------------|-----------|-------------|
| Rel.99<br>(2014/12) | 31,919,197 |           |         | 26,350,404 |            | 120,556,014 |           | 178,825,615 |
|                     | 22,404,242 | 9,105,839 | 409,116 | 15,995,591 | 10,354,813 | 111,997,616 | 8,558,398 |             |
| Rel.98<br>(2014/9)  | 31,801,247 |           |         | 22,751,293 |            | 119,838,741 |           | 174,391,281 |
|                     | 22,359,492 | 9,072,486 | 369,269 | 12,818,994 | 9,932,299  | 111,280,343 | 8,558,398 |             |
| Rel.97<br>(2014/6)  | 31,082,977 |           |         | 22,554,995 |            | 118,764,352 |           | 172,402,324 |
|                     | 21,869,332 | 8,966,715 | 246,930 | 12,754,223 | 9,800,772  | 110,273,226 | 8,491,126 |             |

# DDBJの活動

- 国際塩基配列データベース(INSDB)の構築
- 計算機資源の提供

次世代シーケンサーの登場により、データ量が飛躍的に増大



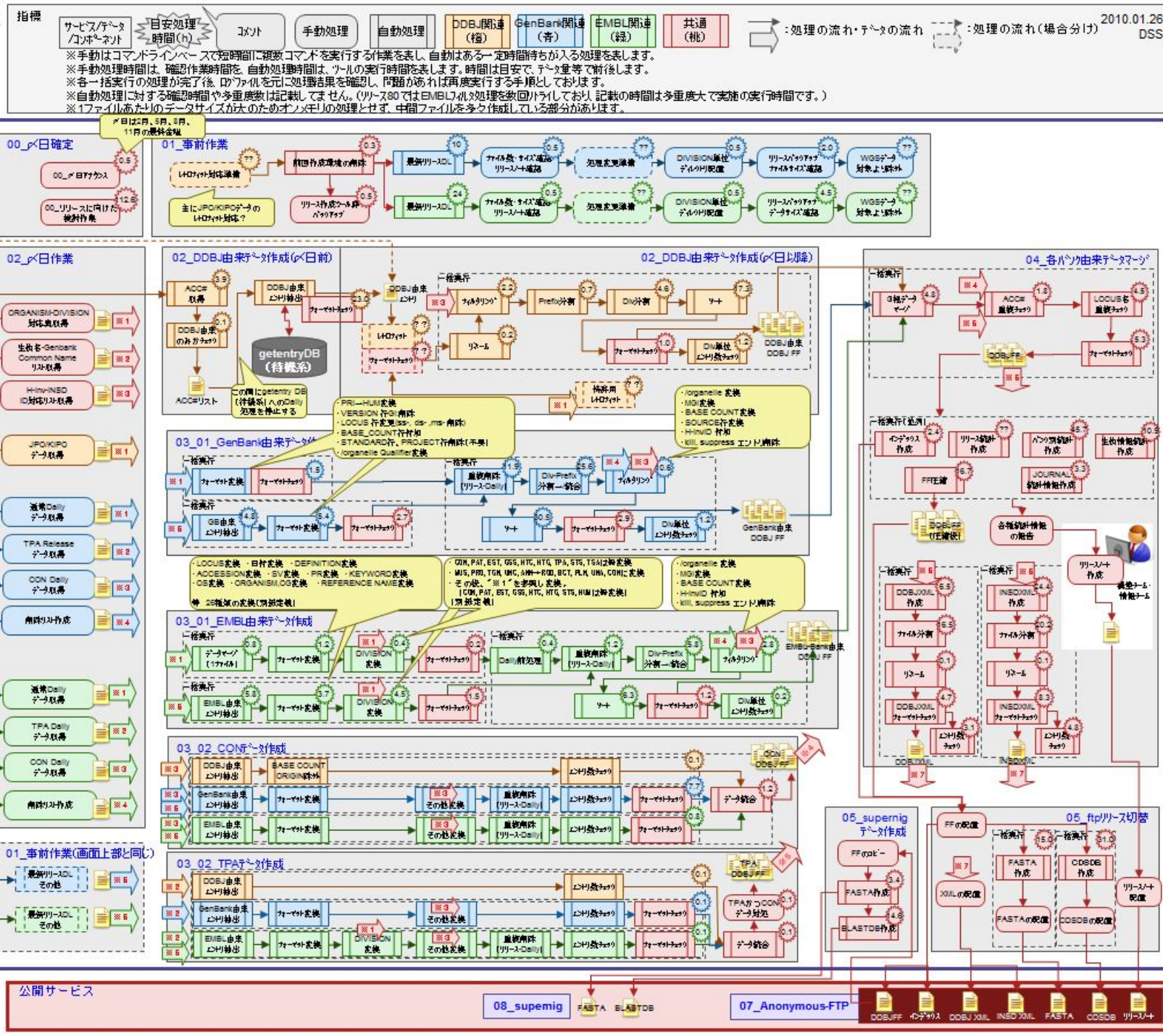
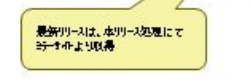
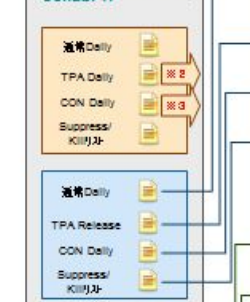
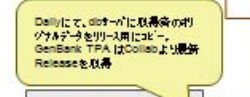
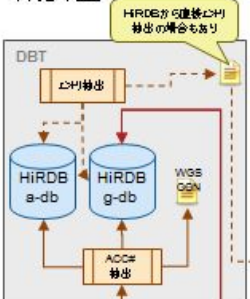
- データ量の増大に対応するため、まず業務用ソフトウェアを全面刷新。
- スーパーコンピュータをCOTS(commercial off-the-shelf)型に移行。
- ディスクサイズ約20倍・計算処理性能約14倍とすることに成功。



ユーザー数も増大(現在約1700名)。  
生命系に特化したスパコンとして広く使われている。

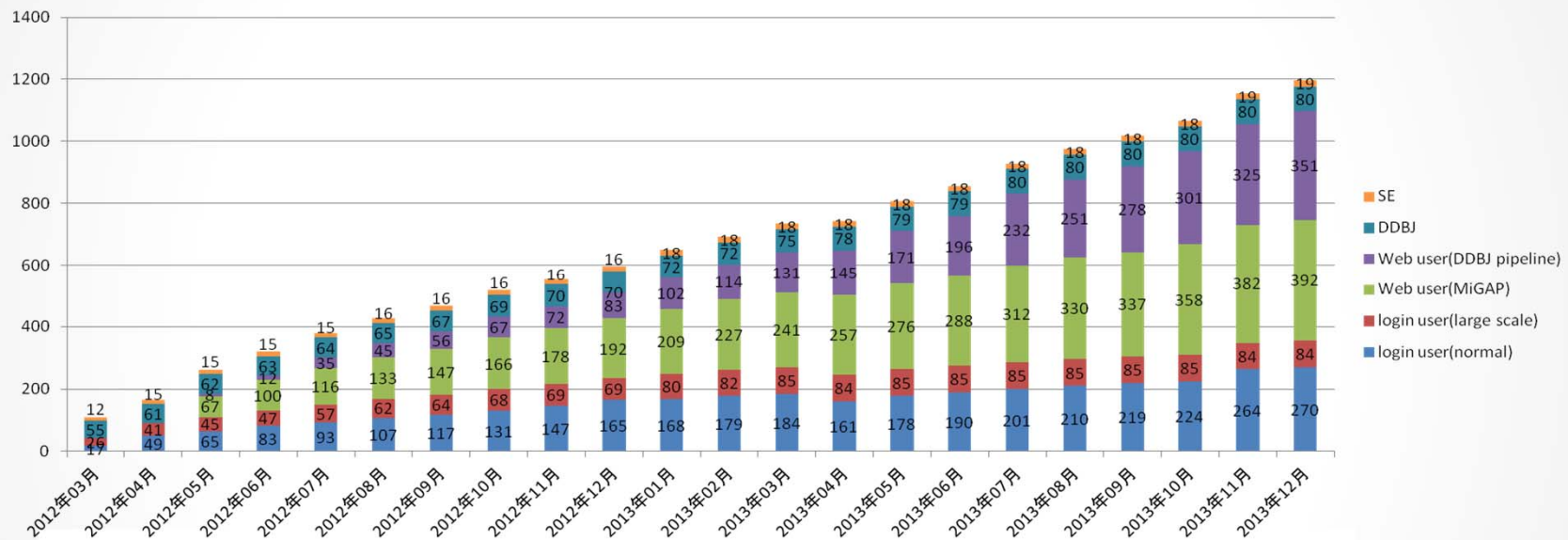


## 02\_定期リリース処理概要図



2010.01.26  
DSS

# ユーザー数の推移

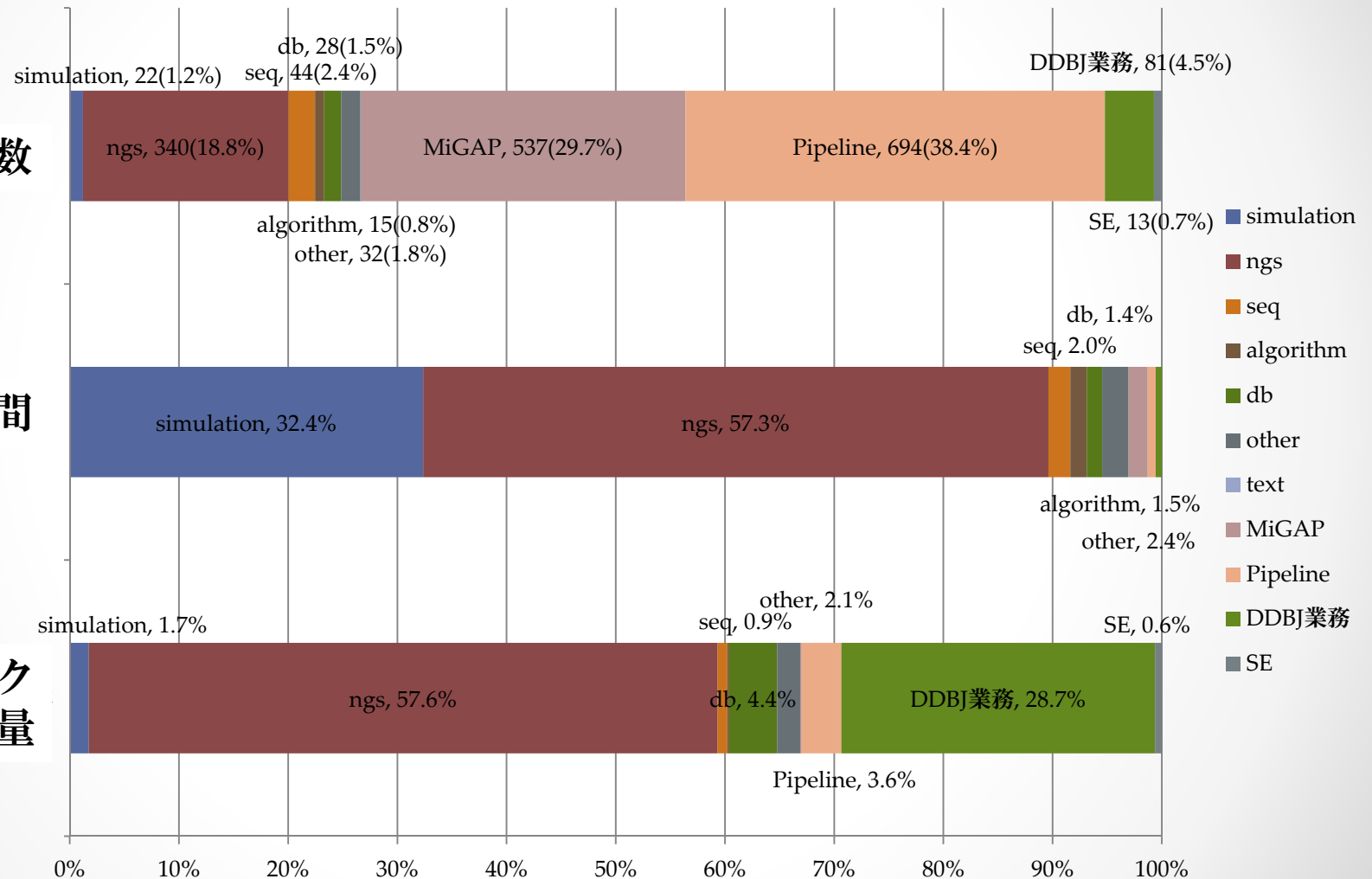


| ユーザ分類            | 外部ユーザ数 | 遺伝研所属ユーザ数 | 総ユーザ数 |
|------------------|--------|-----------|-------|
| 一般研究ユーザ          | 349    | 43        | 392   |
| 一般研究ユーザ - 大規模 -  | 76     | 29        | 105   |
| Webサービスユーザ       | 534    | 12        | 546   |
| DDBJ pipelineユーザ | 687    | 33        | 720   |
| 業務ユーザ            | 0      | 84        | 84    |
| スパコン管理者          | 0      | 13        | 13    |
| 全ユーザ合計数          | 1646   | 214       | 1860  |

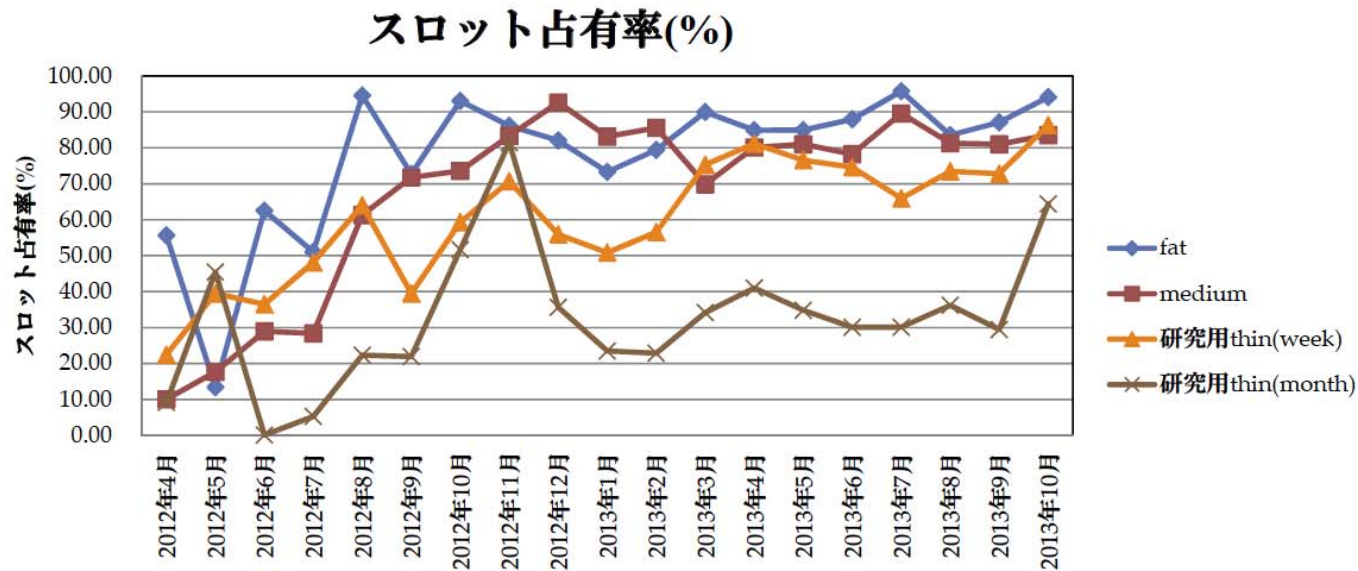


# ユーザーの利用目的とリソース使用量

ユーザー数



# 計算ノードの種類と稼働率



Thin node (64GB memory 554nodes)



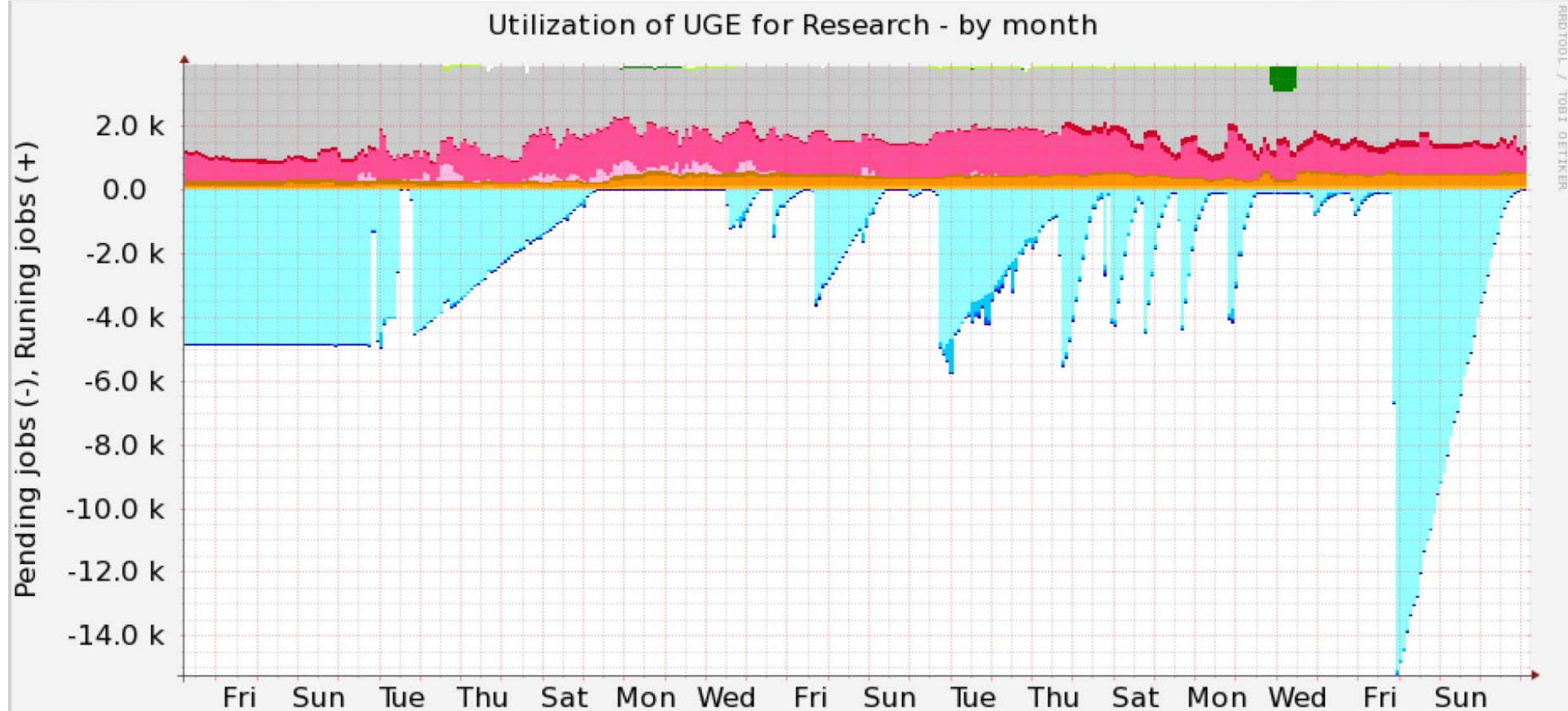
Medium node (2TB memory, 10nodes)



Fat node (10TB memory, 1node)



# 実行中のジョブ数（赤）と 待機中のジョブ数（青）



## ディスク利用状況

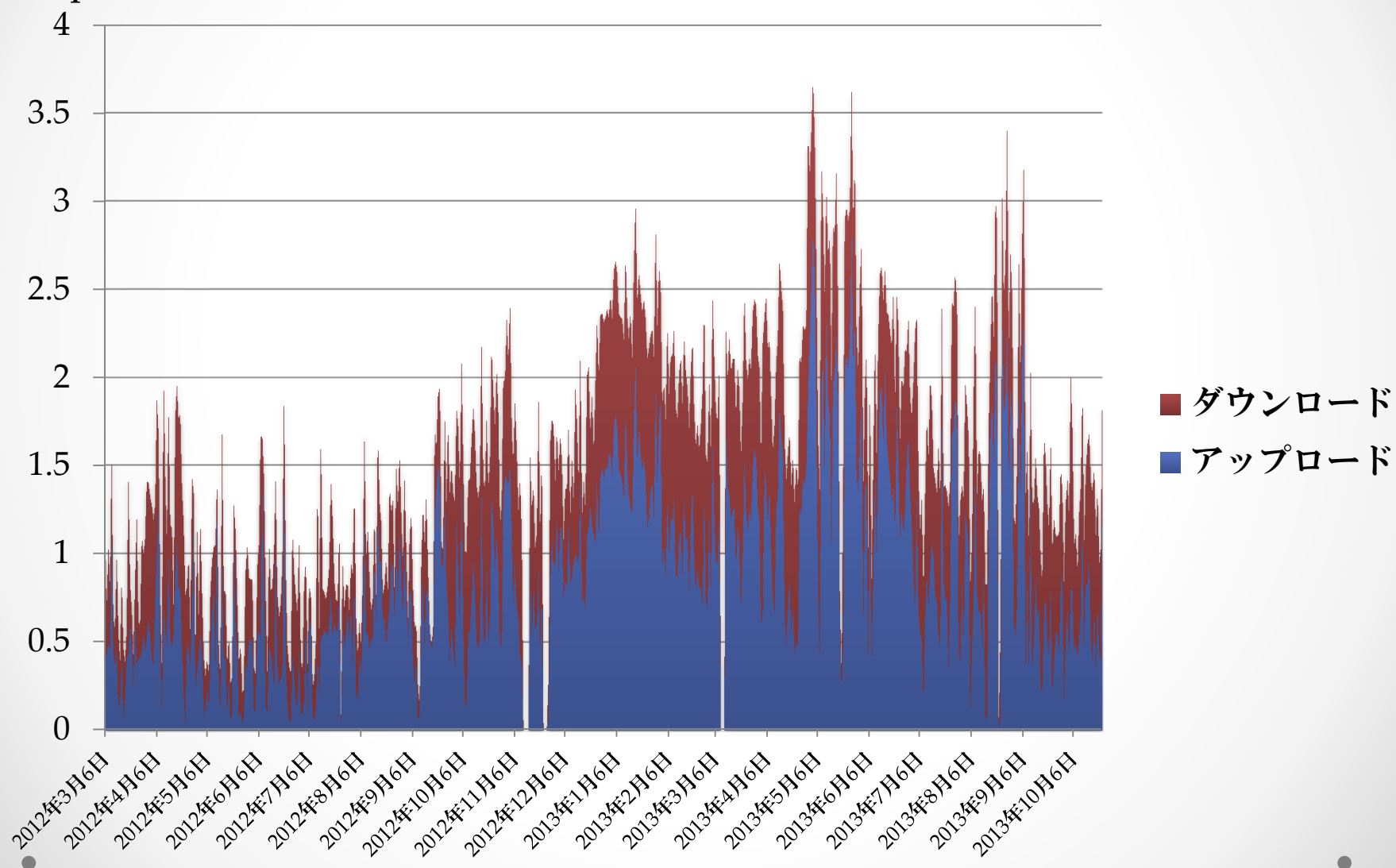
lustreのsize,file項目は、現在のディスク使用量（全ユーザ合計）／ディスク容量で表現しています。  
quotaのsize項目は、申請ディスク使用量（全ユーザ合計）／ディスク容量で表現しています。  
sizeの単位は、Tbyteです。

|                |                     |                             |                |                           |                            |
|----------------|---------------------|-----------------------------|----------------|---------------------------|----------------------------|
| lustre1 size : | 946.87 / 1,024.90   | <div><div></div></div> 92%  | lustre1 file : | 193,347,536 / 731,840,512 | <div><div></div></div> 26% |
| quota size :   | 2,074.99 / 1,024.90 | <div><div></div></div> 202% |                |                           |                            |
| lustre2 size : | 944.13 / 1,024.90   | <div><div></div></div> 92%  | lustre2 file : | 630,671,061 / 731,840,512 | <div><div></div></div> 86% |
| quota size :   | 1,015.14 / 1,024.90 | <div><div></div></div> 99%  |                |                           |                            |
| lustre3 size : | 349.16 / 1,787.70   | <div><div></div></div> 19%  | lustre3 file : | 52,174,115 / 731,840,512  | <div><div></div></div> 7%  |
| quota size :   | 834.46 / 1,787.70   | <div><div></div></div> 46%  |                |                           |                            |
| lustre4 size : | 232.80 / 1,787.70   | <div><div></div></div> 13%  | lustre4 file : | 13,093,785 / 731,840,512  | <div><div></div></div> 1%  |
| quota size :   | 817.70 / 1,787.70   | <div><div></div></div> 45%  |                |                           |                            |
| lustre5 size : | 239.37 / 1,787.70   | <div><div></div></div> 13%  | lustre5 file : | 25,122,321 / 731,840,512  | <div><div></div></div> 3%  |
| quota size :   | 597.22 / 1,787.70   | <div><div></div></div> 33%  |                |                           |                            |



# ネットワークトラフィック(一日平均)

[Gbps]



# 遺伝研スパコン ユーザー会

## 【日程】

### - 三島開催 -

日時： 7月22日（火） 13:30～16:30

会場： 国立遺伝学研究所 本館2F 大会議室  
静岡県三島市谷田1111

<http://www.nig.ac.jp/>

### - 東京開催 -

日時： 7月31日（木） 13:30～16:30

会場： 国立情報学研究所（学術総合センター） 12F 1208会議室  
東京都千代田区一ツ橋2-1-2

<http://www.nii.ac.jp/>

## 【プログラム】

遺伝研スパコンの紹介 (1時間)

1. 遺伝研スーパーコンピュータの概要
2. INSDの構築(DDBJ登録業務、DRA, JGAの紹介)
3. 検索・解析サービス(DDBJ pipelineとMiGAPの紹介)
4. 計算機資源の貸出に関する話題
  - ・ NGS関係の活用法, GPUなどの使い方
  - ・ 共同研究について

意見交換 (1時間)

- ・ 各種サービス、DDBJ登録運用への意見
- ・ 次期スパコンに望むこと
- ・ 課金を含めたシステムの効率化
- ・ 運用方法についての意見
- ・ 使い方の質疑応答 など

## 【ご要望一覧】

- ・ [スパコンユーザ会意見・要望一覧](#)

## 【議事録】

- ・ [三島開催議事録](#)
- ・ [東京開催議事録](#)

<http://sc.ddbj.nig.ac.jp/index.php/component/content/article/10-ja-category/ja-important-notice/214-nig-supercomputer-meeting-result>

|                   |                                                                                                                                                                                                                                                                                                                                           |
|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 東京工業大学            | 2. 複数processor/core を利用するjob の優先順位向上<br>上記の1にも関連していますが、8 cpu で実行できるのに、8 cpuで実行せずに4 cpu で実行した方が、早くjobが投入されて先に終わるという事が見られます。スパコンの利用法としては、各ユーザがjobを動かす実時間を減らし（ピーク性能を高く出し）なるべくユーザに、機会を与える方がよいかと思っています。                                                                                                                                        |
| 東京工業大学            | 3. 導入ソフトウェアのバージョンの時系列的な明示について<br>スパコンに大量のバイオインフォ系ソフトが導入され、メンテナンスされています。バイオ系のユーザが、気楽に使える機会を増やすため、また、日本で開発されたソフトウェアをみんなが利用できるように普及機械を作るため、現状は継続して頂くのが良いかと思います。<br># 本格的なユーザは、自分のディレクトリの下に、自分のバイナリを入れてあるかとは思いますが。一方、バージョンアップが頻繁に起きると、それにより結果が変わる事があるかと思っています。そのようなことから起きる問題を避けるため、たとえば、最新のバージョン番号だけでなく、特定の日（期間）のバージョン番号一覧を掲載しておくのは、いかがでしょうか？ |
| 東京工業大学            | 4. 開発用プロセスと実行用プロセスの分けについて<br>スパコンのJOBには、大きく分けて、ソフト開発用のプロセスとソフトを利用して、解析をするプロセスがあるかと思っています。（どのように区別するかは別問題として）スパコンの目的にもよりますが、世界に向けて日本発のソフトウェアを増やすため、開発用のプロセスを解析用のプロセスに優先するような仕組みが導入できないでしょうか？                                                                                                                                               |
| 東京工業大学            | 5. 有償jobの導入について<br>無料で利用できることは、非常に大きなメリットではある一方、失礼ながら、「このjobは、今やらなくても・・・」というjobのせいで、至急走って欲しいjob が進まないような事があるように思います。例えば、有償のjob queueを導入して、そこは有償の代わりに無償のjob に優先して実行してくれる、という形を導入いただくことはできませんでしょうか？                                                                                                                                         |
| 東京工業大学            | 6. 動的なqsub のリソースの変化<br>実行プロセスによって、殆どがsingle process だけど、部分的に複数プロセスで走ることがあると思います（ゲノムアセンブリのプログラムなど）。この時、qsubでリソースを確保すると、最大プロセス数とメモリに合わせて確保しないといけなため、実際には使われない部分もリソースを確保しておかなければなりません。この様なケースの時、qsubが動的にリソースを管理するような機構はありませんでしょうか？あったら、導入して頂けませんか？                                                                                           |
| DBCLS             | データ解析再現性を担保するため、ソフトウェアのversion管理が必要である。解析環境をVMのイメージファイルとして用意し共有すれば良いのでは？イメージを共有する事で環境依存の問題が減り、計算の選択肢が増える。                                                                                                                                                                                                                                 |
| 大阪大学蛋白質研究所        | NIGスパコンでもユーザが公開Webサーバを立てられれば良いと思う。                                                                                                                                                                                                                                                                                                        |
| 遺伝研・植物遺伝研究室       | NIGスパコンへの要望としてジョブとディスクが逼迫している状況の改善を望んでいる。また、NIGユーザ以外とのデータ授受が出来る仕組みがあれば利便性が高いと考える。仮想化および特権ユーザの貸与サービスも実施して欲しい。                                                                                                                                                                                                                              |
| 東京工業大学 学術国際情報センター | HPCとBigDataの融合について進めている。HPCI共用ストレージがあり、DDBJ DBをこのストレージにおいては如何か。それによってHPCI参加サイトからDDBJ DBに直接アクセス可能となる。まずはその前段階としてGSIGとDDBJで連携しては如何か。                                                                                                                                                                                                        |

スパコンユーザー会のご意見 その他多数 ...

- 三島開催 -

1) 大田達郎 DBCLS

「大量NGSデータの並列処理と共用スパコンにおける環境構築の今後について」

2) 辰本将司 国立遺伝学研究所 生命情報研究センター 比較ゲノム解析研究室

「NGS解析における遺伝研スパコンの活用事例について」

3) 徳田直子 名古屋大学 大学院工学研究科 計算理工学専攻 計算生物物理グループ(笹井研究室)

「Hi-Cデータに基づくゲノム動力学シミュレーション」

4) 笠原浩太 大阪大学蛋白質研究所

「共用スパコンを用いた分子動力学シミュレーション研究」

5) 長崎英樹 国立遺伝学研究所 生命情報研究センター 大量遺伝研究室

「NIGスパコンでのNGSデータ処理 - 個人的にスパコン操作で悩ましかったこと」

6) 大柳一 国立遺伝学研究所 系統生物研究センター 植物遺伝研究室

「植物大規模オミックス解析に必須なNIGスパコンの明治大&遺伝研での活用事例」

7) 三浦信一 東京工業大学 情報理工学研究科

- 東京開催 -

1) 安井康夫 京都大学農学研究科

「どうしても避けて通れぬNGS ～ドライ初心者による遺伝研スパコン利用例～」

2) 笠原雅弘 東京大学大学院新領域創成科学研究科 情報生命科学専攻

「遺伝研スパコンの『ここが困った!』事例集と今後への提言」

3) 二階堂愛 理化学研究所情報基盤センター

「DevOpsとcloudで達成する再現性のあるスーパーコンピューティング」

4) 亀田倫史 産総研・CBRC

「分子動力学を活用した創薬研究」

5) 森田史 東京工業大学大学院生命理工学研究科

「メタゲノム解析におけるスーパーコンピュータの利用例」

6) 近藤伸二 国立極地研究所新領域融合研究センター

「DDBJスーパーコンピュータを使用した大規模極限環境メタゲノム解析及び ハイブリッドマウス全トランスクリプトーム解析」

7) 山本希 東京工業大学大学院

8) 柚木克之 東京大学大学院理学研究科

「遺伝研スパコンシステムによるシステム生物学研究の高速化・大型化」



# スパコンユーザーからの 要望まとめ

1. スパコン上で動作するソフトウェアの開発支援
2. データ解析をする研究者に便利な環境の提供
3. 個人ゲノム等制限アクセスデータの解析を可能とする環境の提供
4. データベースをミラーして必要なデータを揃えておくこと
5. スパコン間の連携

検索

[> STAFF](#)

[> ENGLISH](#)

[HOME](#)

[最新情報  
TOPICS](#)

[機構案内  
GUIDE](#)

[情報公開  
INFORMATION](#)

[リンク  
LINK](#)

[アクセス  
ACCESS](#)



[▶ 国立極地研究所](#)

[▶ 国立情報学研究所](#)

[▶ 統計数理研究所](#)

[▶ 国立遺伝学研究所](#)

[▶ 新領域融合研究センター](#)

[▶ ライフサイエンス統合データベースセンター](#)

[新着情報](#) WHAT'S NEW

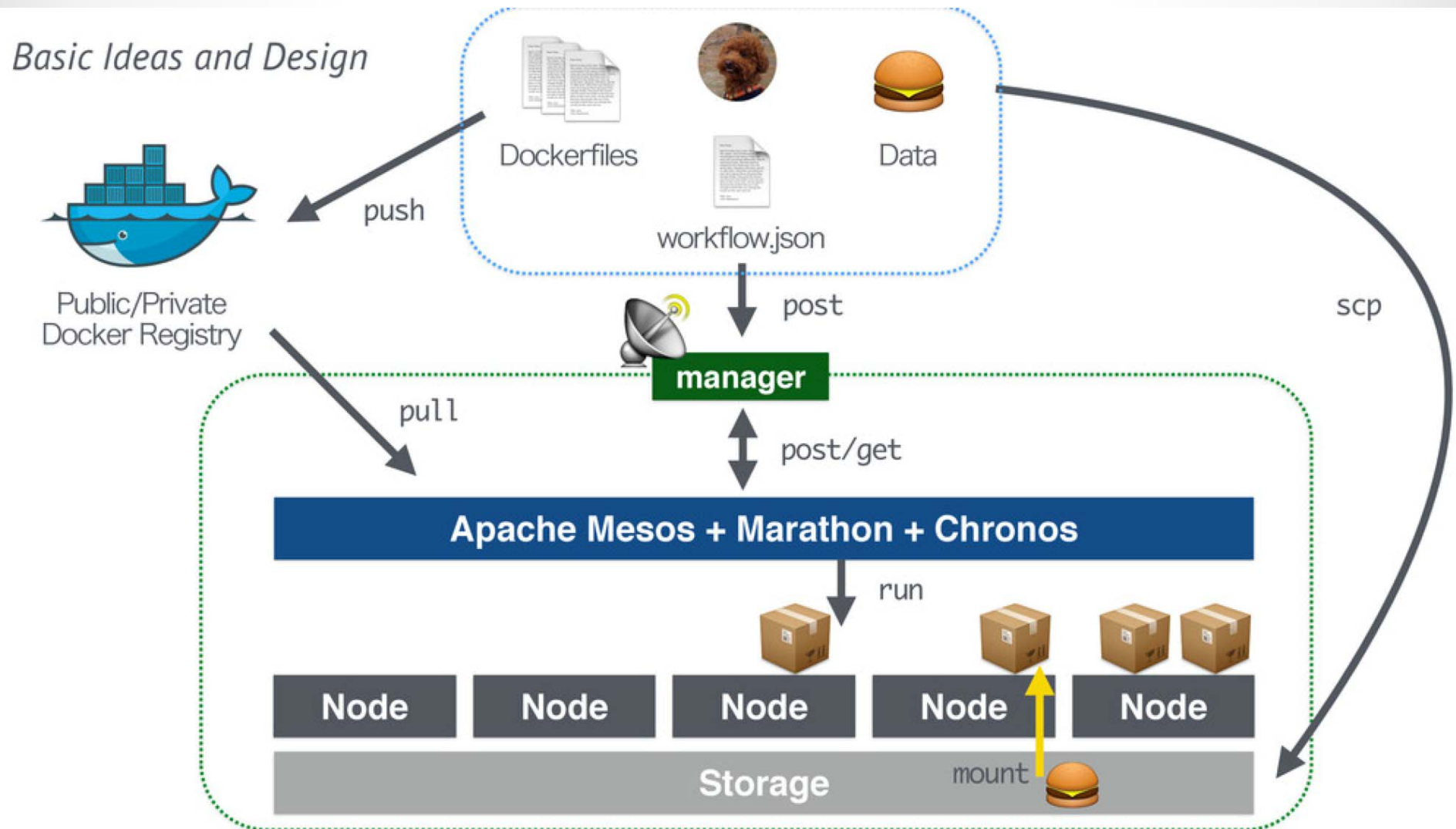
[一覧へ](#)

# RaaS: Reproducibility as a Service by docker container on NIG supercomputer

by [Tazro Inutano Ohta](#)

Published January 22, 2015

<https://speakerdeck.com/inutano/raas-reproducibility-as-a-service-by-docker-container-on-nig-supercomputer>



# RaaS: Reproducibility as a Service by docker container on NIG supercomputer

by [Tazro Inutano Ohta](#)

Published January 22, 2015

<https://speakerdeck.com/inutano/raas-reproducibility-as-a-service-by-docker-container-on-nig-supercomputer>

```
{
  "user": "inutano",
  "project": "exome",
  "volume": "/data",
  "workflow": [
    {
      "image": "awk",
      "args": ["1", "10"],
      "cpu": 1,
      "memory": "500M"
    },
    {
      "image": "bwa/bwa",
      "args": ["-v"],
      "cpu": 8,
      "memory": "8G"
    },
    {
      "image": "gatk/gatk",
      "cpu": "2",
      "memory": "4G"
    }
  ]
}
```

- JSON format configuration file
- Describe a workflow contains multiple steps
- 1 container for 1 app
- Include directory to be mounted on containers



# RaaS: Reproducibility as a Service by docker container on NIG supercomputer

by [Tazro Inutano Ohta](#)

Published January 22, 2015

<https://speakerdeck.com/inutano/raas-reproducibility-as-a-service-by-docker-container-on-nig-supercomputer>



Benefit

- Cost-Efficiency
  - ~5mins to prepare data analysis environments
- Scalability
  - Parallel execution without env dependencies
- Reproducibility
  - Manuscript = Raw data + workflow.json + Containers

# クラウド利用への期待 (まとめに替えて)

1. 解析用の計算ノードが不足（要求は多い）
  - o 多人数のユーザーがGrid Engineを用いて様々な大きさの計算を実行している状態。
2. 生物学系の研究者がWebサイトを立てて解析結果を他の研究者と共有する環境が要る。
3. データアーカイブ領域：数十PBのデータに対するアクセス
4. 個人ゲノムデータのような制限アクセスデータのアーカイブと計算環境