学術情報基盤オープンフォーラム2018

ライフサイエンス研究の生産性を向上させるためのオンデマンドクラウド

二階堂 愛, PhD.

理化学研究所バイオインフォマティクス研究開発ユニット. ユニットリーダー

筑波大学. 教授 (協働大学院)

1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

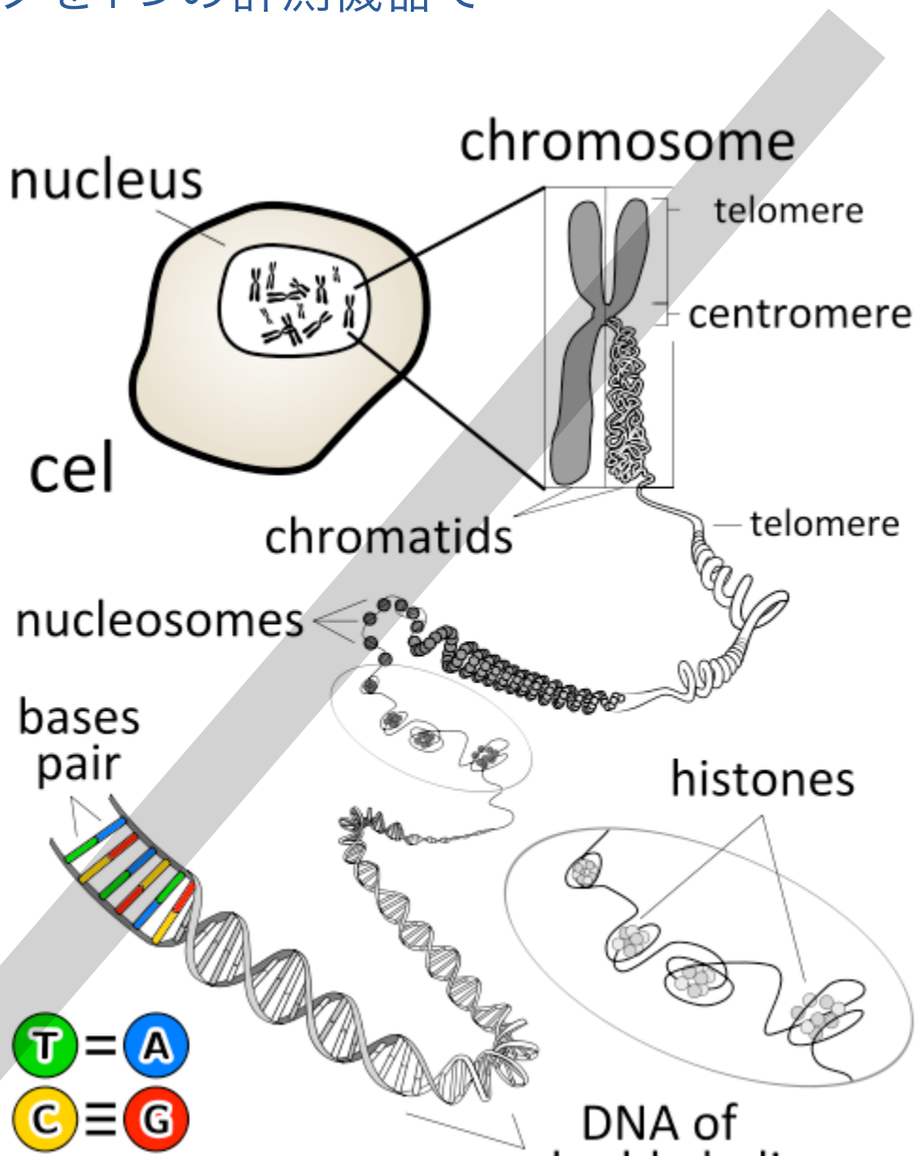
HPC on クラウド

オンデマンドハイブリッドクラウド

あらゆる現象を観測する多目的顕微鏡

セントラルドグマのすべてをステップを1つの計測機器で

One person:
30 trillion cells
400 type of cells
100 thousand mRNA
100 thousand protein
80 years old



4C/5C/HiC-Seq

ChIP-seq

ChIP-seq

DNA-seq
Exome

ChIP-seq

Methyl-Seq

RIP-seq

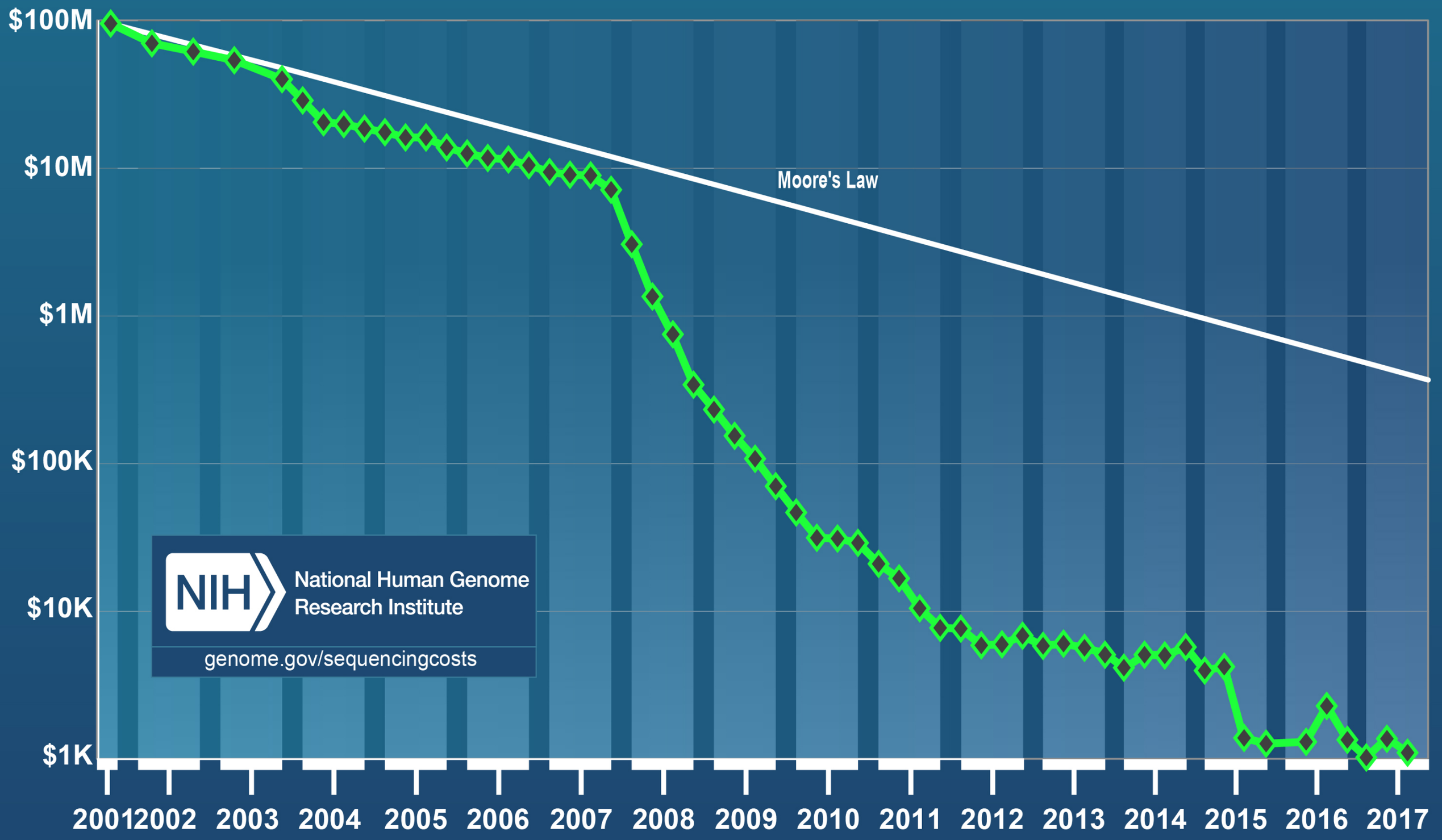
RNA-seq

Ribosomal Profiling

>100 *-Seq

http://www.chem-agilent.com/cimg/SystemWorkflow600B.gif

Cost per Genome



大型化と個人化に向かうゲノム科学

セントラルドグマのすべてをステップを1つの計測機器で

ゲノム科学の大型化 = ルーチンの解析を大規模に

Genomics england

10万人のヒトゲノムを収集する

10KP

1万の植物ゲノムを決定する

HUMAN CELL ATLAS

ヒトのすべての細胞種を同定する

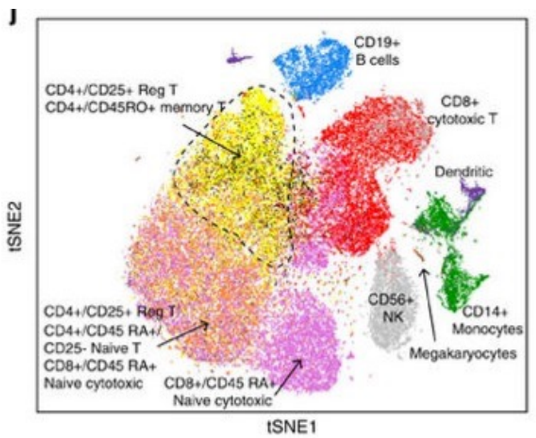
ゲノム科学の個人化 = だれでも大規模なデータが取れる

Illumina NextSeq

Chromium (10x genomics)

68K single-cells

peripheral blood mononuclear cells (PBMCs)



データ解析や計算機科学に詳しくない実験科学者がデータ解析に参加

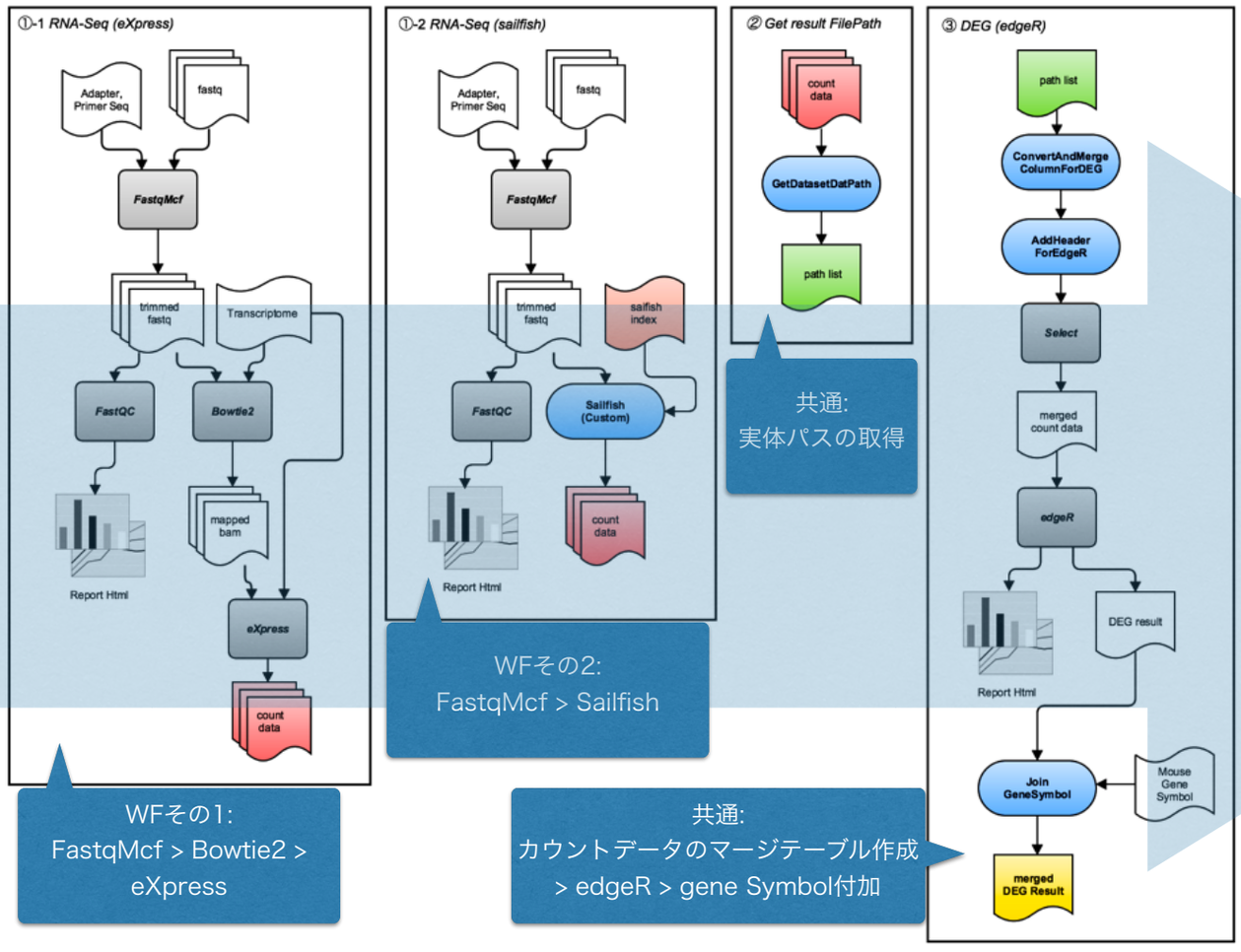
DNAシーケンスのデータ解析の例

たくさんのプログラムとデータベースの組み合わせ

DNA配列 (0.4GB-600TB)

```
@HWUSI-EAS747_0001:1:1:12:1307#0/1
ATATTCATTAGGATCTCATCTTTTGTCACATGACTCTCTAAGTGCTGACT
+HWUSI-EAS747_0001:1:1:12:1307#0/1
ggggggdgggff`fffffffffffffffffffffffffaffffffffff
@HWUSI-EAS747_0001:1:1:12:471#0/1
GCCAACTCTCCTGTGCTGCTCAGGTGAGGAGCAGGGCCCATTCTCCTGCG
+HWUSI-EAS747_0001:1:1:12:471#0/1
ggggfgcggWggagegggggcgdg_eg_Pg_ggTcagUe\Nebbfea_^b
@HWUSI-EAS747_0001:1:1:12:1047#0/1
ACTGGGCACTGACTTTCTAAAGTGCTTTTGGGAGGACATGTTCAGACCAT
+HWUSI-EAS747_0001:1:1:12:1047#0/1
gggbdfgfggcbgggfffZbfffffffffffaffffaffeffZfZdffd
@HWUSI-EAS747_0001:1:1:13:210#0/1
GGAGCTCAGAAACGTGGCTGTGTCTCCTTGGTTACCAGCTCCTGGTGCTG
+HWUSI-EAS747_0001:1:1:13:210#0/1
e`^ee^e^e``\c_[f`bfebOfffffffT`b`eLf``ffeeBBBBBBBB
@HWUSI-EAS747_0001:1:1:13:1770#0/1
ATTCTAGTTTCATTCGAGTACTTTCTGAGGGGAAAGGGACTGTATAGGAG
+HWUSI-EAS747_0001:1:1:13:1770#0/1
fgfgggcegggagggf^ZfdfffeffdfZ`cTceed^^^feceTecX\Mc
@HWUSI-EAS747_0001:1:1:13:615#0/1
GAACAATATGCACAATAGCTACATATCTGTGGAACAAGAACAAAACACAG
+HWUSI-EAS747_0001:1:1:13:615#0/1
gcggcgggggfeffYfffffffffefffff]effeRe``feVMafeeaa
@HWUSI-EAS747_0001:1:1:13:1493#0/1
CAATGCTGCCTCCTGGTGGATCTGGACTTTCCAAGTTCATCATTTTAAAT
+HWUSI-EAS747_0001:1:1:13:1493#0/1
gfgggggggggggggghgfeggggegggggggefgdgggcgcgegfg`gg
@HWUSI-EAS747_0001:1:1:13:1038#0/1
CAGAAAGTCGCAGTAGGGAAATTTCTATTTCAAGGCCAGCCTGGGCAACA
+HWUSI-EAS747_0001:1:1:13:1038#0/1
e`e^_Xa]ecY_a__f^ZJbQ\bbbfefefcffbQb\b\ZffdY`b_ReZ
@HWUSI-EAS747_0001:1:1:13:1147#0/1
AATGAAAAACACATTCGTTGGAAACGGGATTTGTAGAACAGTGTATATCA
+HWUSI-EAS747_0001:1:1:13:1147#0/1
cgfggfgg^gfgfgggggggfgffggggcggfggggWcggfgggggcggg
@HWUSI-EAS747_0001:1:1:13:1024#0/1
AGAATTCCAGTGAAGCTGACGATGGAAAATGGCGGTATCTTGAATTATTC
```

大規模な並列計算・パイプライン処理



行列 (MBからGB程度)

サンプル

遺伝子/変異

これらのワークフローを数千から数万サンプルを扱う

データ解析の再現性とライフサイエンス

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 543 > Issue 7645 > Editorial > Article

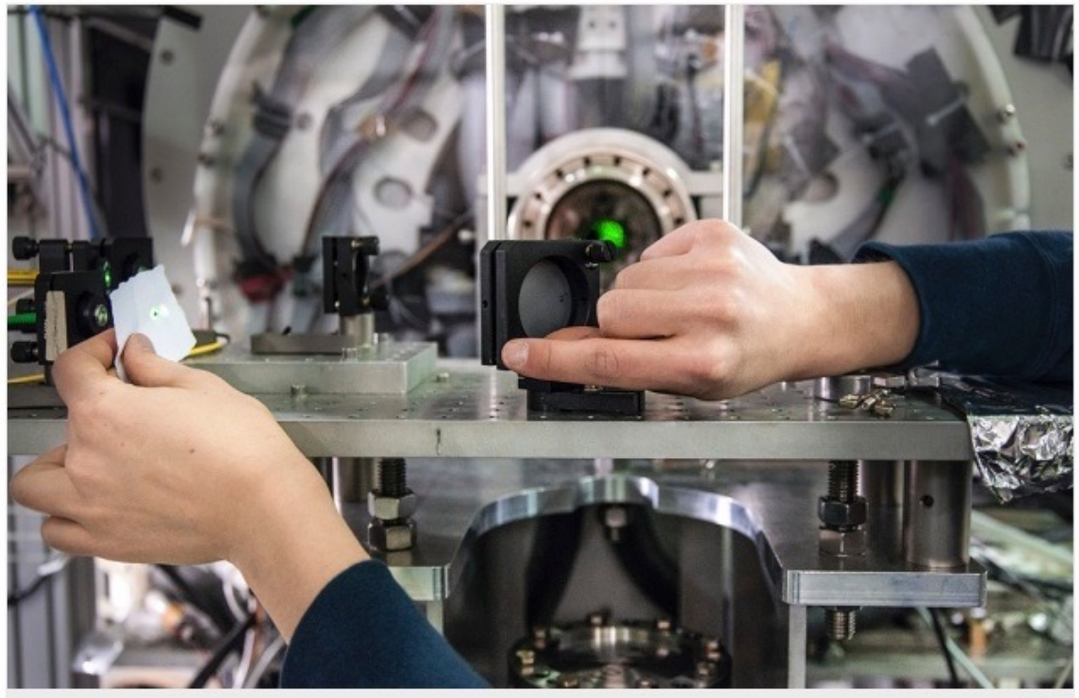
NATURE | EDITORIAL

Announcement: Transparency upgrade for Nature journals

The Nature journals continue journey towards greater rigour.

15 March 2017

PDF | Rights & Permissions



CERN/SPL

Laser physics is being targeted for better reporting of experiments.

nature biotechnology

Home | Current issue | News & comment | Research | Archive | Authors & referees | About the journ

home ▸ advance online publication ▸ abstract

NATURE BIOTECHNOLOGY | RESEARCH | ANALYSIS

Reproducibility of computational workflows is automated using continuous analysis

Brett K Beaulieu-Jones & Casey S Greene

Affiliations | Contributions | Corresponding author

Nature Biotechnology (2017) | doi:10.1038/nbt.3780

Received 10 June 2016 | Accepted 22 December 2016 | Published online 13 March 2017

Full text | PDF | Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract · Accession codes · References · Author information · Supplementary information

Replication, validation and extension of experiments are crucial for scientific progress. Computational experiments are scriptable and should be easy to reproduce. However, computational analyses are designed and run in a specific computing environment, which may be difficult or impossible to match using written instructions. We report the development of continuous analysis, a workflow that enables reproducible computational analyses. Continuous analysis combines Docker, a container technology akin to virtual machines, with continuous integration, a software development technique, to automatically rerun a computational analysis whenever updates or improvements are made to source code or data. This enables researchers to reproduce results without contacting the study authors. Continuous analysis allows reviewers, editors or readers to verify reproducibility without manually downloading and rerunning code and can provide an audit trail for analyses of data that cannot be shared.

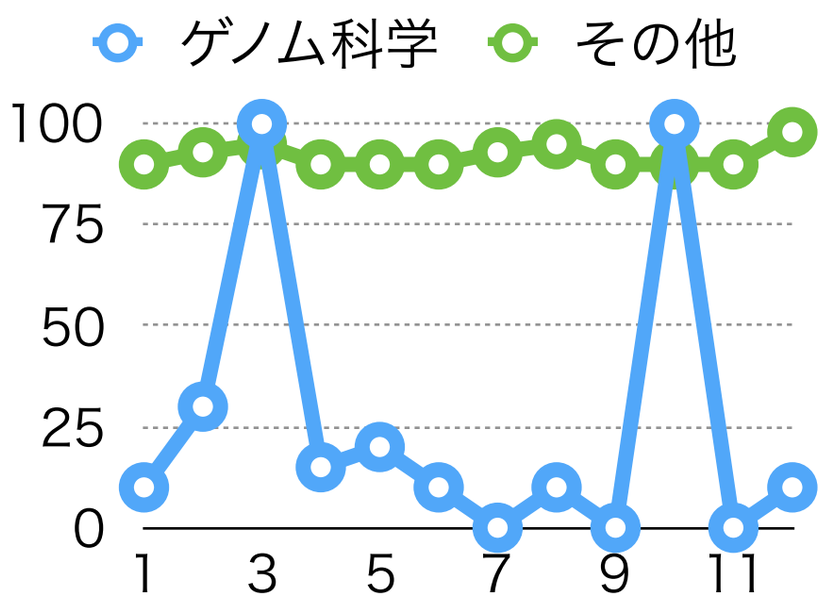
「計算」の高速化から「研究」の高速化へ

問題点1: 複雑な環境構築がデータ解析のボトルネック

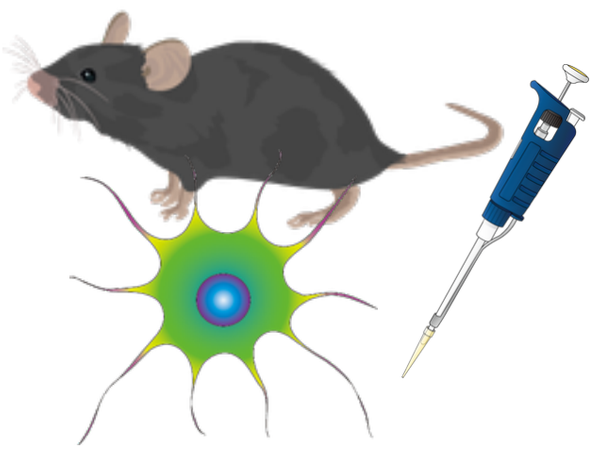
・これまでのHPCとライフサイエンス

・現在のHPCとゲノム科学

問題点2: スポットでデータが出力される



問題点3: エンジニアリングと科学



計算機利用のリテラシー格差が存在

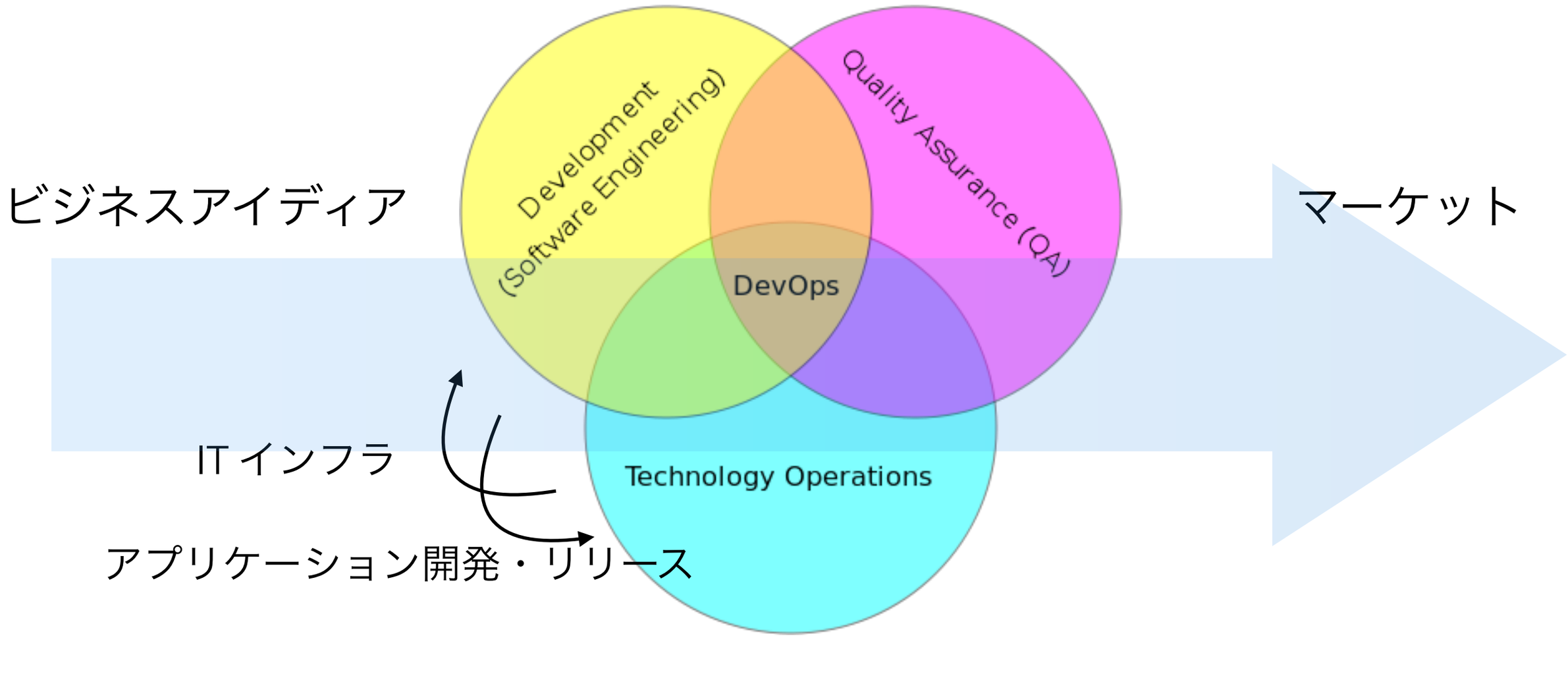
ライフサイエンスを取り巻くデータ解析環境構築の問題点

- ライフサイエンス研究に集中したいが、データ解析環境を構築することは手間がかかる
    - 計算機やソフトウェアの調達や管理、保守
    - オープンソースツールやDBのアップデートが速い
- データ解析の再現性担保できなくなる
- 生命情報科学者・計算生命科学者が計算機利用支援に借り出される
    - 計算導入、環境設定、計算機利用指導、簡単な作図、などのサポートに忙殺され、自身の研究に専念できない

DevOps = Development + Operations

ITインフラとアプリケーション開発の一体化

ビジネスアイディアを素早くマーケットに出すための
ITに関する思想とその技術

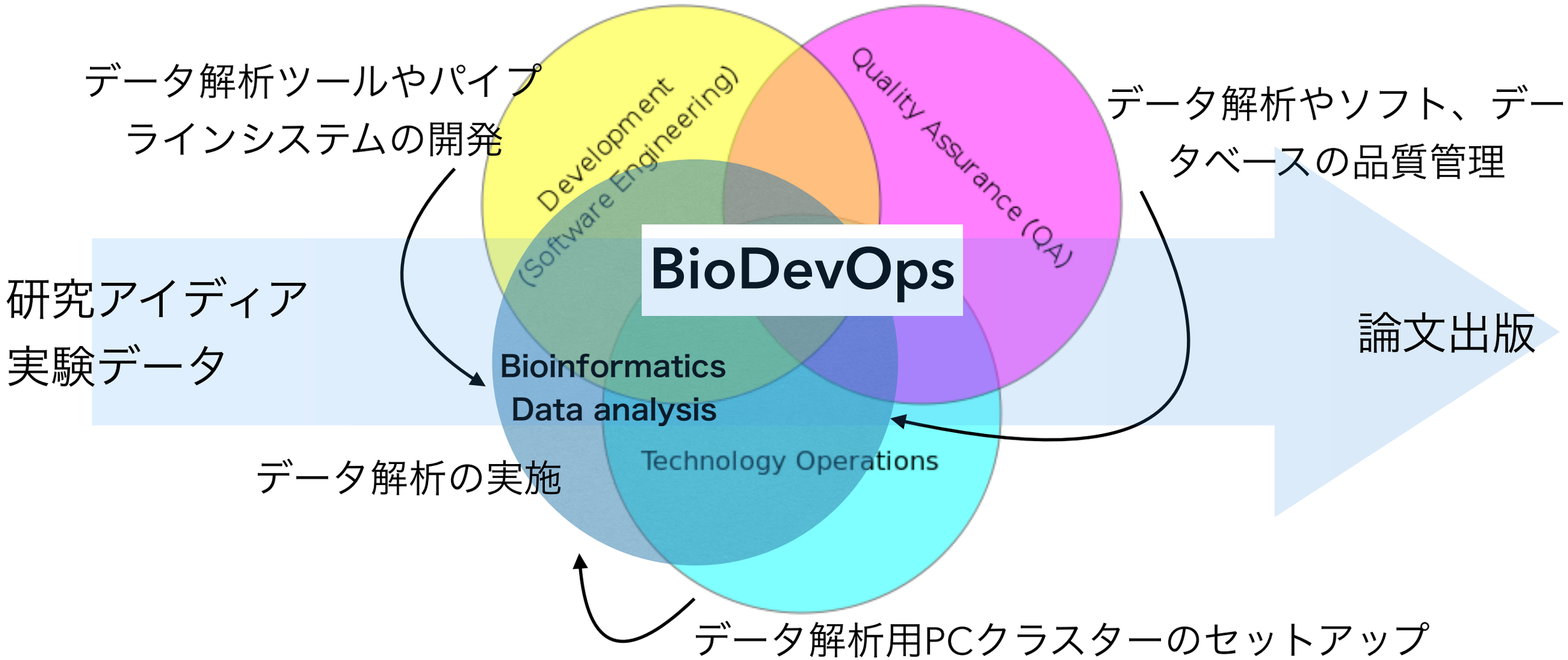


http://ja.wikipedia.org/wiki/DevOps. modified

BioDevOps = Bioinformatics + Development + Operations

バイオインフォマティクス解析とITインフラとアプリケーション開発の一体化

研究アイディアを素早く論文として出すための考えかたと技術



BioDevOps =
クラウドコンピューティング、Infrastructure as Code, パイプライン管理で実現する!

1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

クラウドの利用

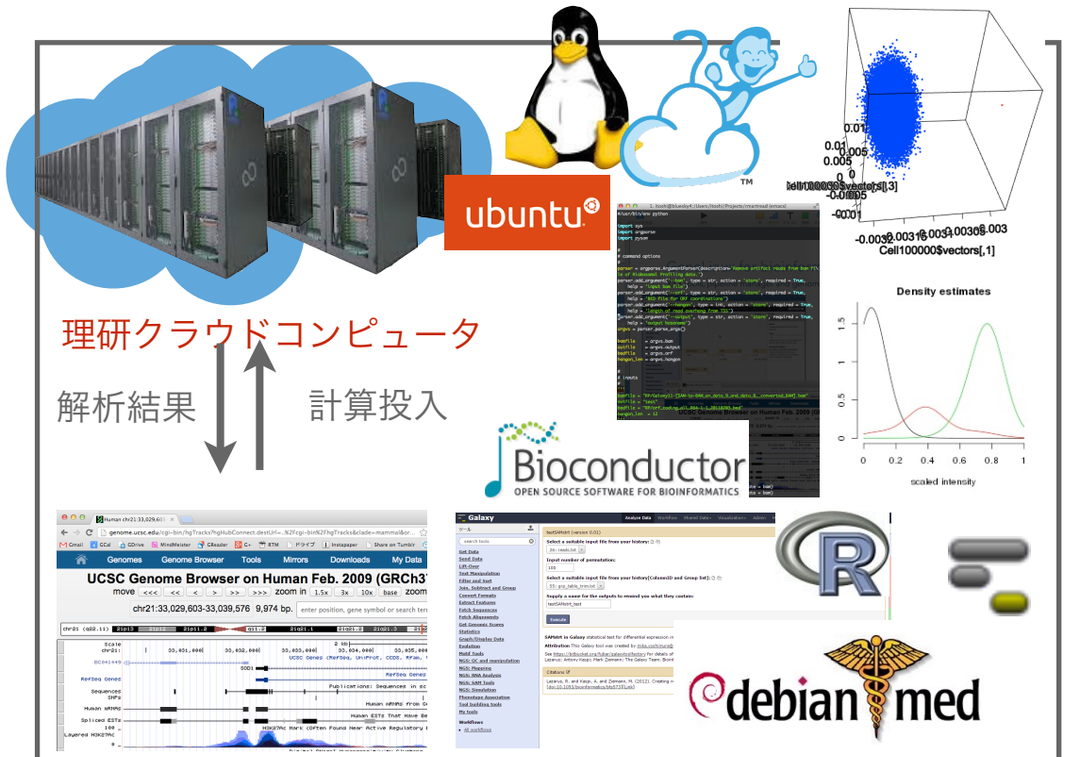
- Private Cloud
- Public Cloud
- Hybrid Cloud

全理研に向けた計算生命科学の支援体制の構築

生命機能科学研究センターと情報システム部の連携

1. 計算生命科学の大規模計算環境を研究者へ提供・教育実施

スペース・管理費などを集約化。約1億円の節約。研究に集中

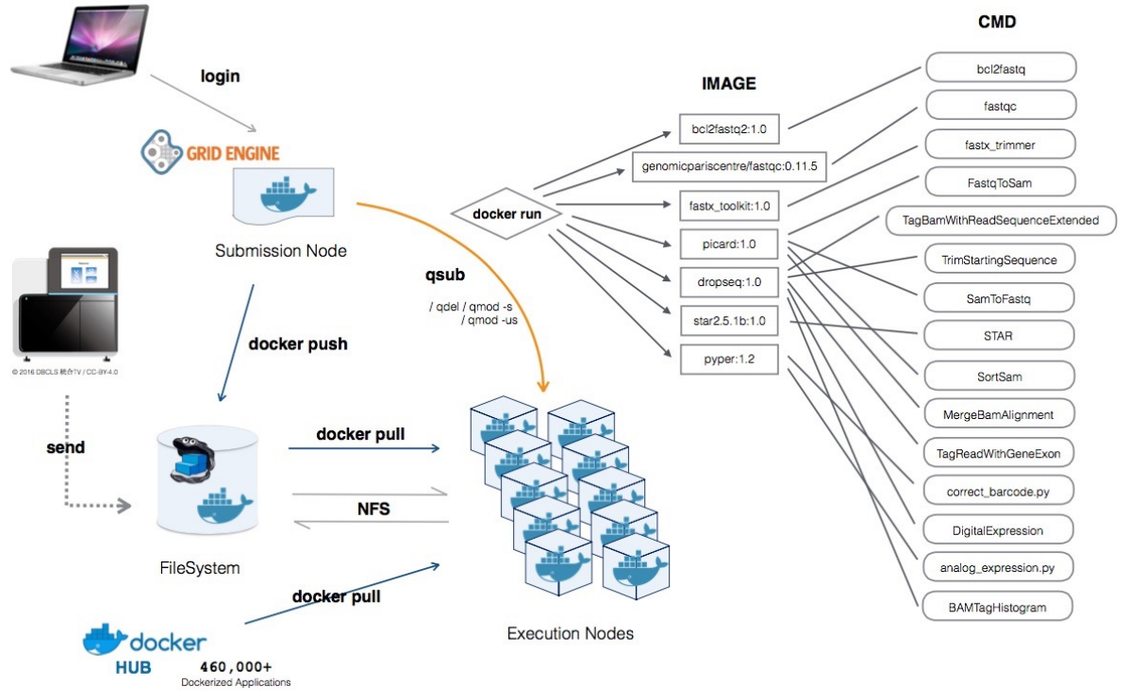


1600程度の解析ソフトウェアを収録

8 CPU cores (2GHz), 64 GB RAM and 2 TB strage

2. 必要なときに必要なだけスパコンをパブリッククラウドから

納期3-6ヶ月を15分に。研究速度の圧倒的な向上



理研内にあるデータに対して自在に計算が可能な仕組み

(国立情報学研究所/マイクロソフトとの共同研究)

Bioinformatics Analysis Environment as Code

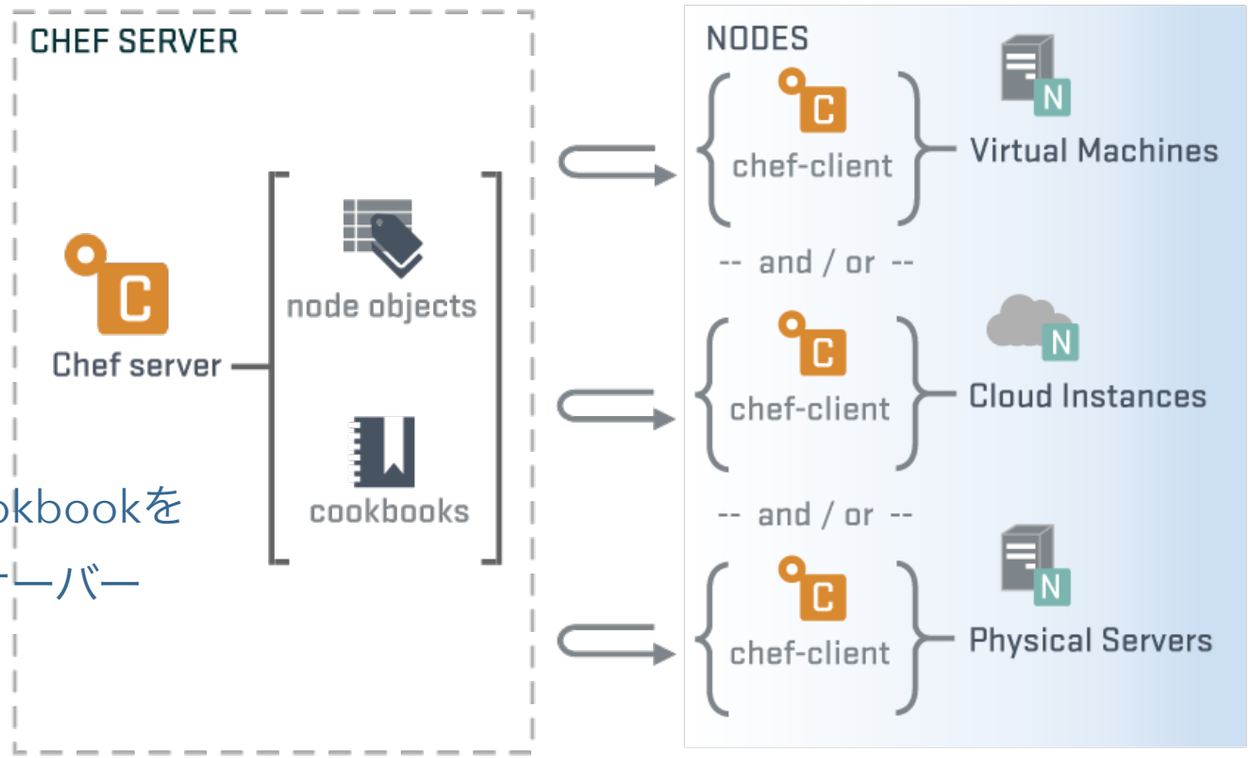
Dev, Ops, Analysis の疎結合化

0. vegrantでマシンを管理

仮想計算機構築

chef cookbookを
実行しプログラムを
自動セットアップ

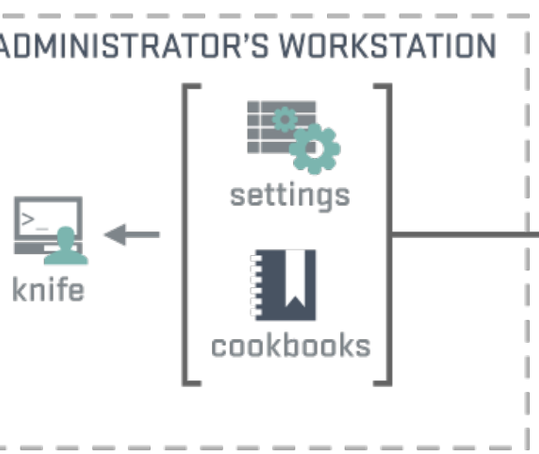
User



3. chef cookbookを
配信するサーバー

ZABBIX

Zabbixでマシンの状況を監視



2. chef cookbookを
配信サーバーへ提供

Jenkinsで継続的インテグレーション

1. chef cookbookを実装し
コード共有レポジトリで管理

http://www.getchef.com/chef/

Chef recipe and Integration Test

Example: Installing NCBI BLAST by chef

branch: **master** **blast-cookbook** / recipes / **default.rb**

manabuishii 7 days ago init commit

1 contributor

file | 10 lines (10 sloc) | 0.253 kb | Open | Edit | Raw | Blame | History | Delete

```
case node.platform
when 'debian','ubuntu'
  package "ncbi-blast+"
when 'centos'
  #
  execute "install blast" do
    command "yum -y install ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.2.29+-1.x86_64.rpm"
    action :run
  end
end
```

debian, Ubuntuの場合は“ncbi-blast+”というパッケージをインストール

CentOSの場合はNCBIからRPMパッケージを取ってきてインストール

Chef recipe and Integration Test

Example: Installing NCBI BLAST by chef

branch: **master** **blast-cookbook** / test / integration / default / bats / **blast_installed.bats**

manabuishii 7 days ago init commit

1 contributor

file | 5 lines (5 sloc) | 0.102 kb Open | Edit | Raw | Blame | History | Delete

```
#!/usr/bin/env bats
@test "blastp  is found in PATH" {
    run which blastp
    [ "$status" -eq 0 ]
}
```

blastpを実行できたらテスト成功

ハードウェア構成をプログラムする

Azure Resource Managerテンプレート

```
- {
    apiVersion: "2016-03-30",
    type: "Microsoft.Compute/virtualMachines",
    name: "[variables('vmNameMaster')]",
    location: "[resourceGroup().location]",
    tags: "[variables('tagValues')]",
  - dependsOn: [
        "[variables('storageAccountName')]",
        "[variables('nicNameMaster')]",
        "[concat('Microsoft.Compute/virtualMachines/', parameters('vmNameNFSServer'))]"
    ],
  - properties: {
    - hardwareProfile: {
        vmSize: "[parameters('vmSizeLeader')]"
      },
    - osProfile: {
        computerName: "[variables('vmNameMaster')]",
        adminUsername: "[parameters('adminUserName')]",
      - linuxConfiguration: {
          disablePasswordAuthentication: "true",
        - ssh: {
          - publicKeys: [
            - {
                path: "[variables('sshKeyPath')]",
                keyData: "[parameters('sshKeyData')]"
              }
            ]
          }
        }
      },
    - storageProfile: {
      - imageReference: {
```

1クリック or 1 コマンドで仮想的なHPCをクラウドに構築

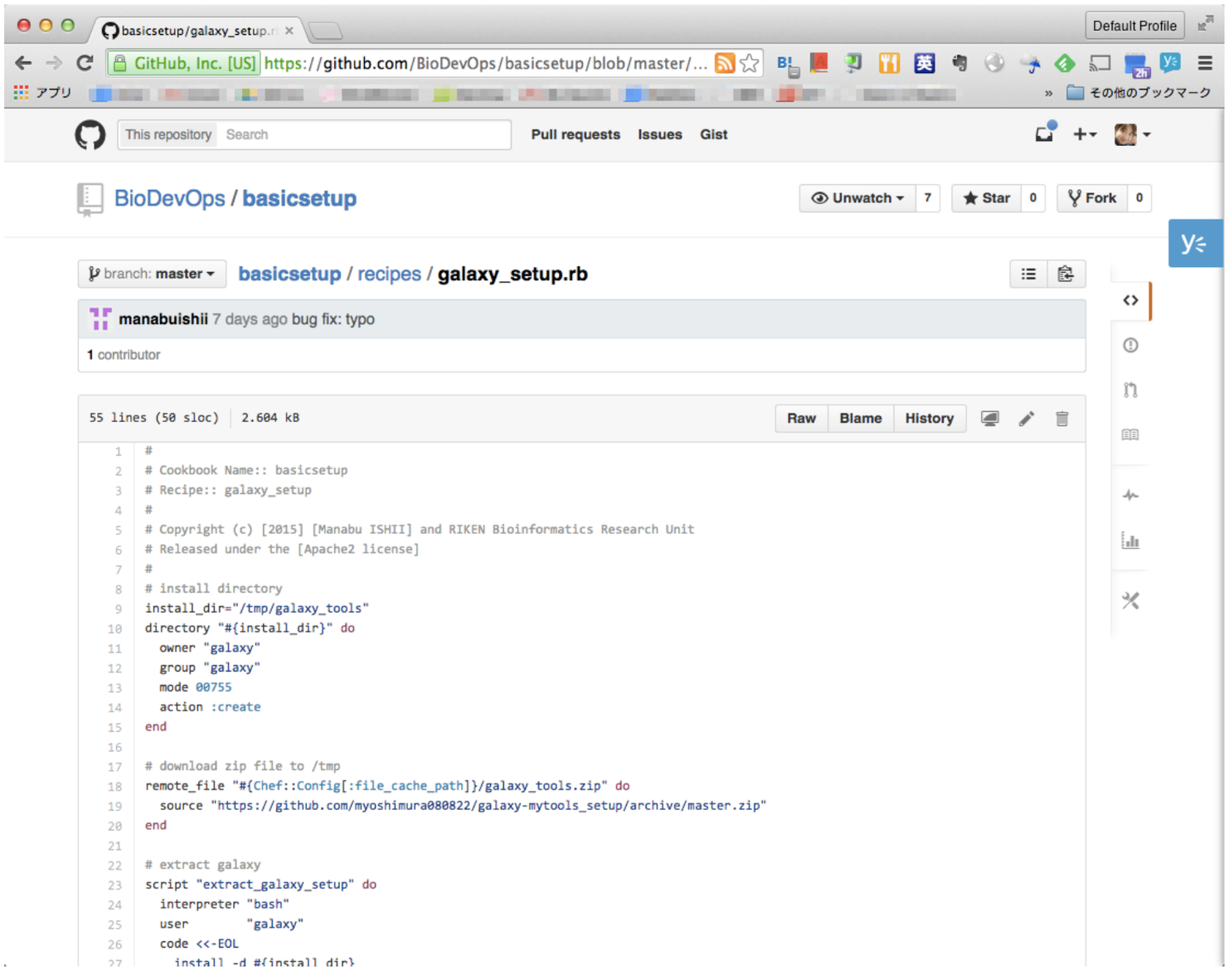
Deploy to Azure

OR

```
azure group deployment create -g AZURERESOURCEGROUP6 \
  -n Deploy1 -f azuredeploy.json -e local.parameters.json -v
```

計算機とその設定をソースコードのように管理する

ソーシャルコードレポジトリ: GitHub



誰でもアップロード

誰でもダウンロード

バージョン管理

修正反映依頼 (pull request)

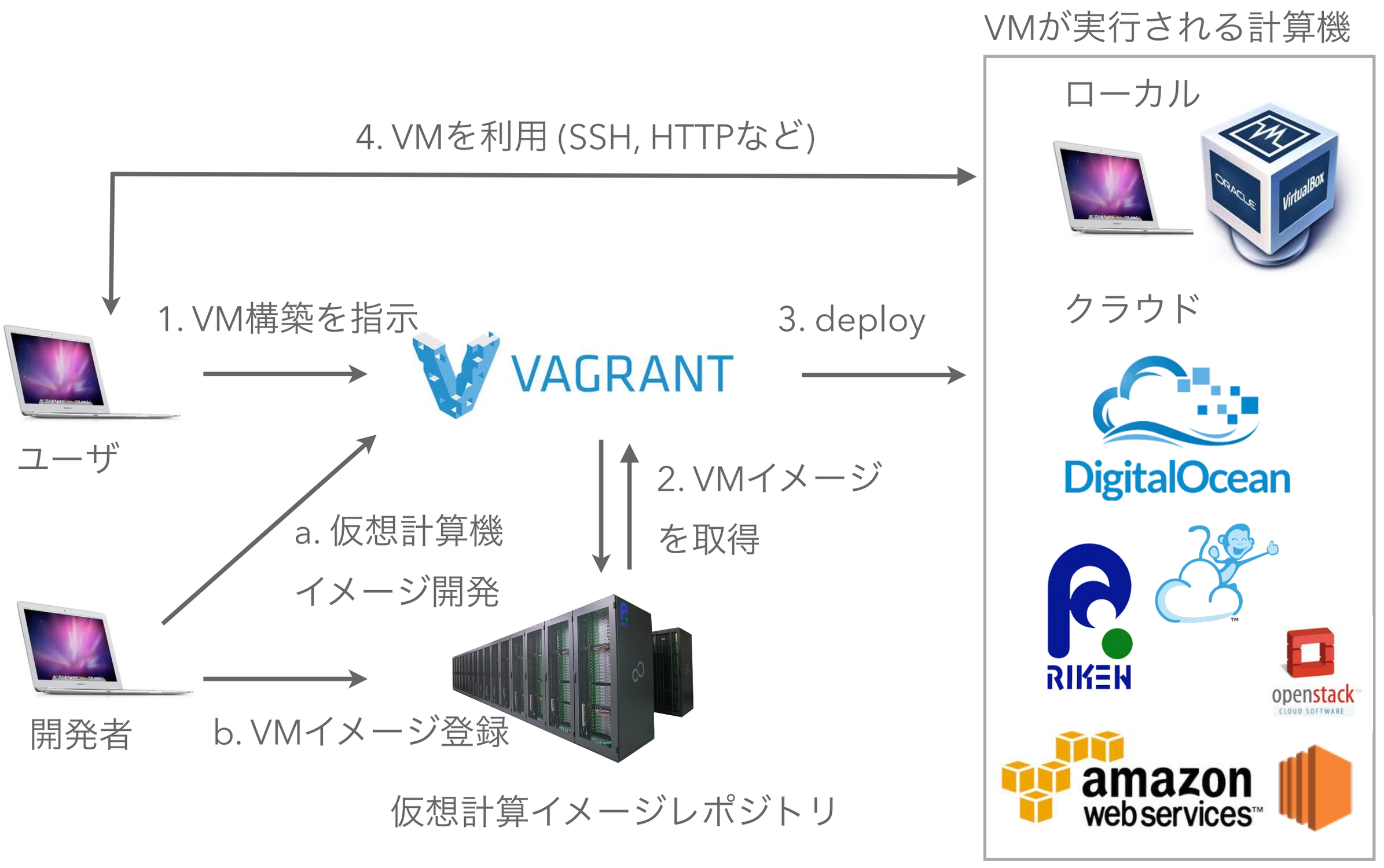
ユーザーやレポジトリをフォロー

Dev, Opsの疎結合に貢献

http://github.com/biodevops/

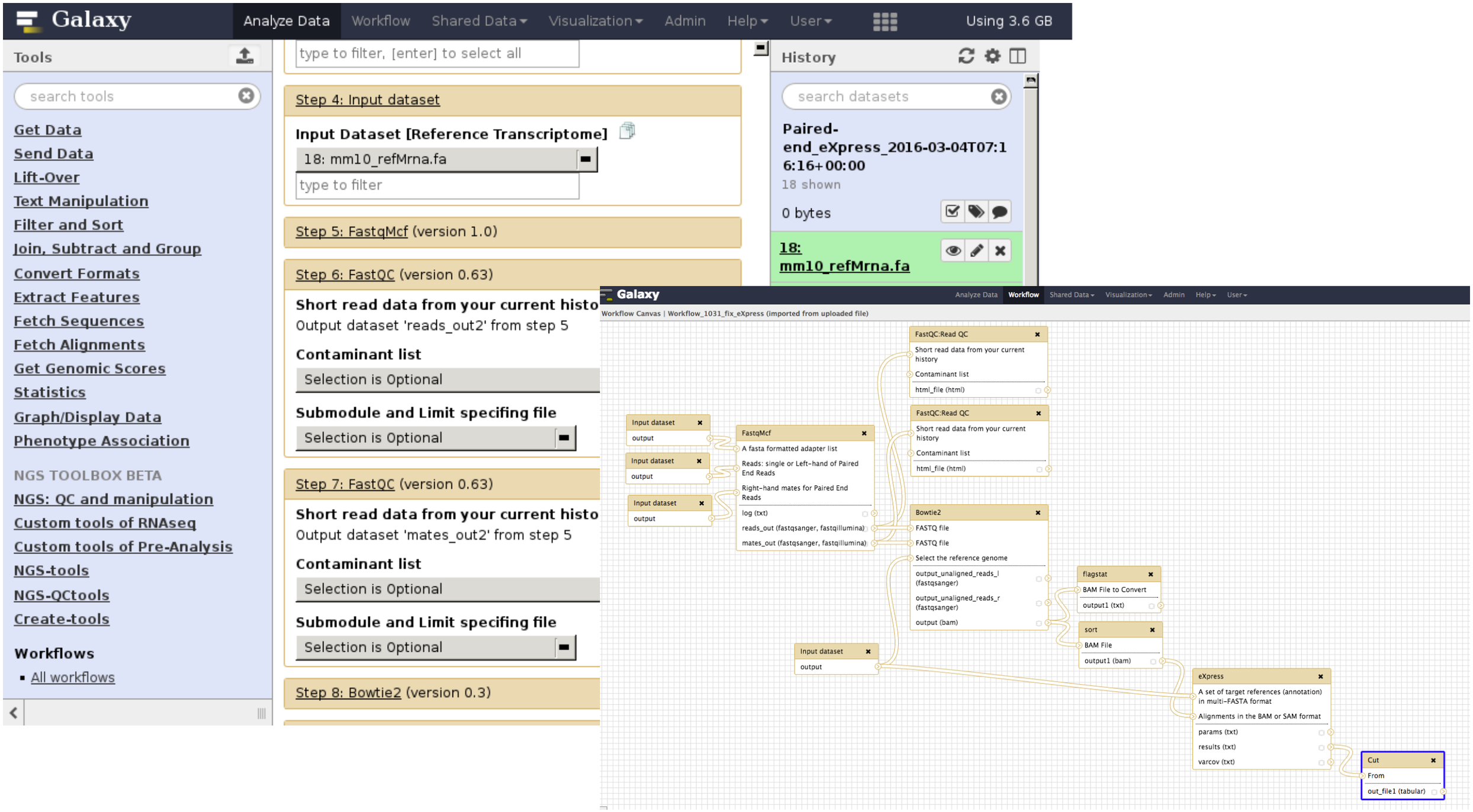
deploy: ソフトウェア環境を利用可能なように配置する

Vagrant: どのようなクラウドコンピュータでも簡単に環境をインストールできる



ワークフロー管理システムGalaxy

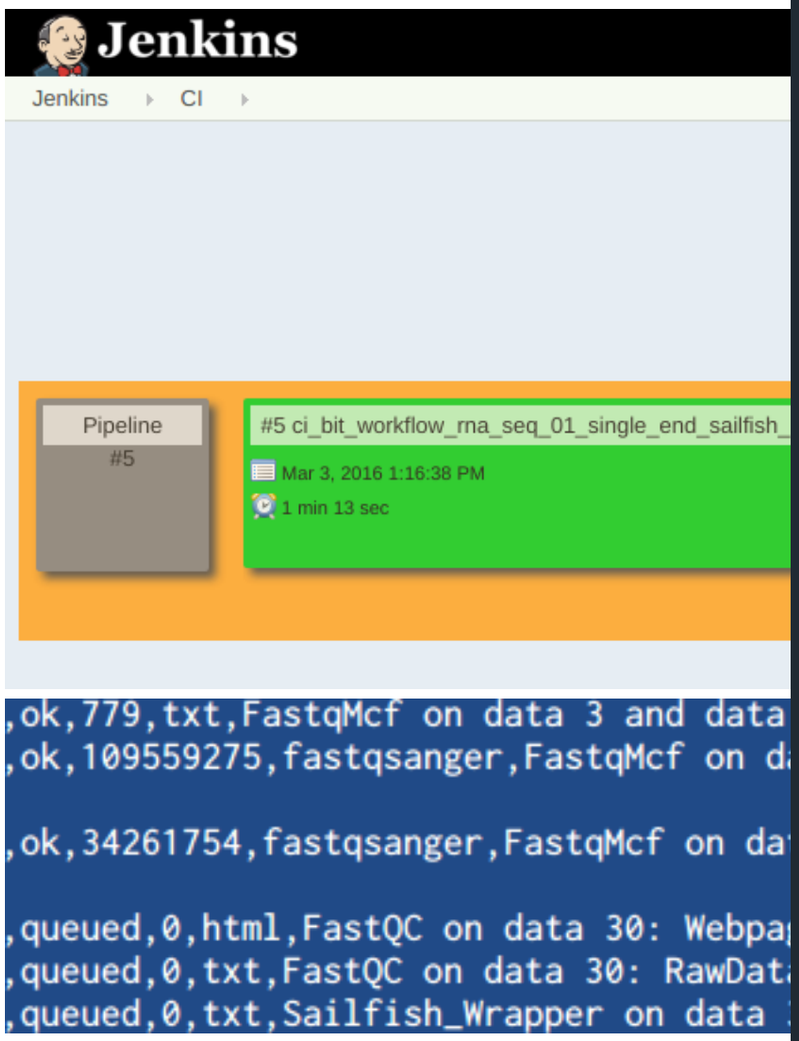
オープンソースのデータ解析ワークフロー実行・管理システム



ワークフローの再配布が可能。様々なツールのレポジトリもあり簡単に利用できる

継続的インテグレ

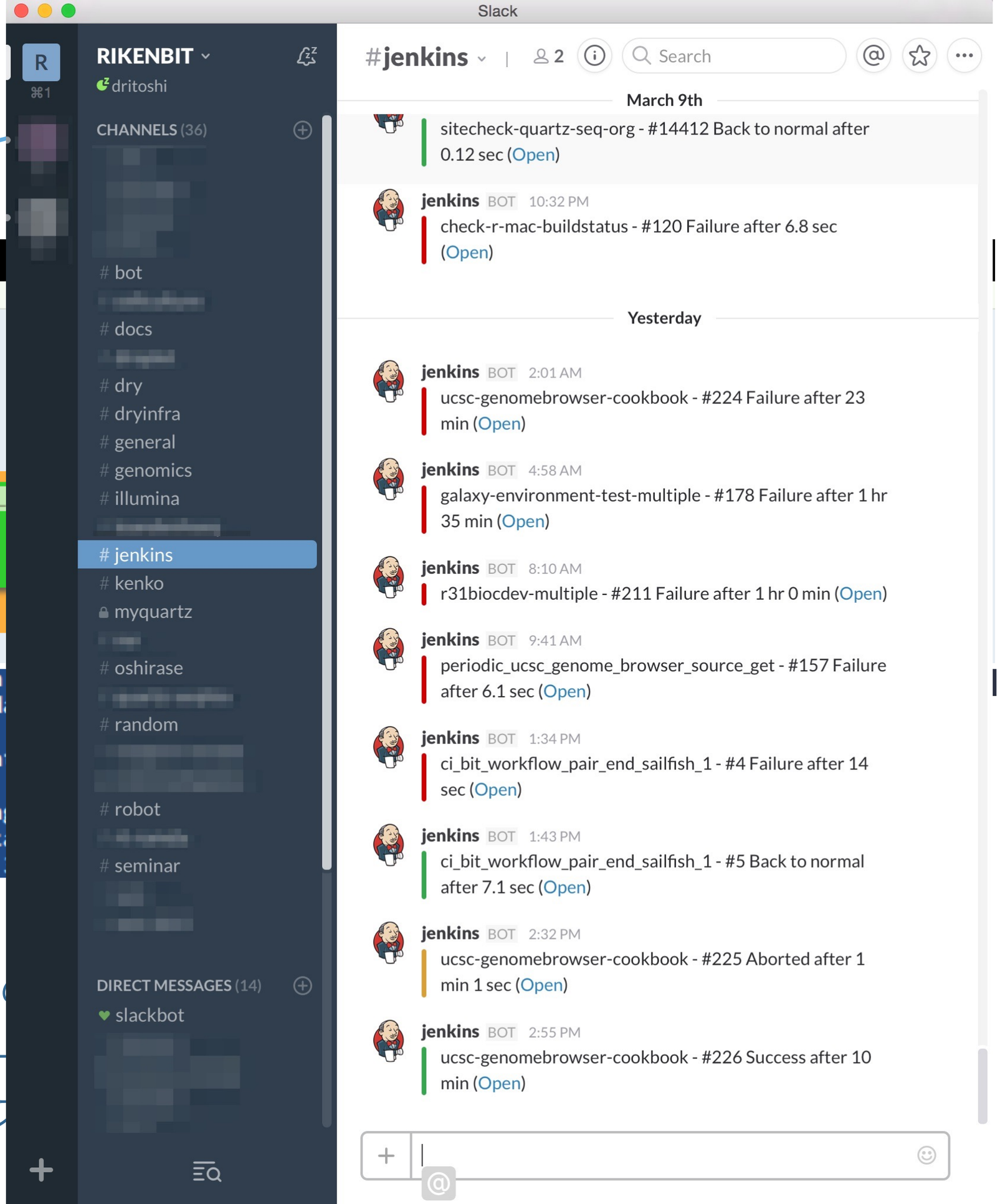
ソフトウェアだけでなく



WebUIから

ワークフロー

自動的にテス



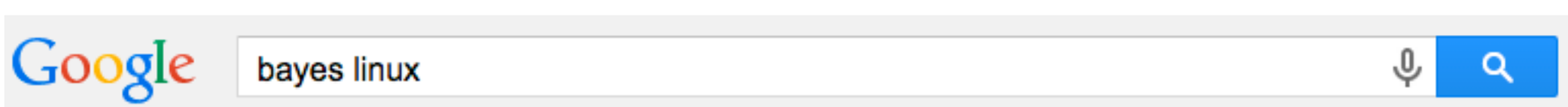
Bayes Linux: Bioinformatics Analysis Environment

Virtual machine with NGS Data analysis tools and pipelines

- >125 tools on Galaxy
- 900 R/Bioconductor Packages
- 600 command line tools (DebianMed)

OPEN SCIENCE AWARD

2015 ソフトウェア部門3位



1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

クラウドの利用

- Private Cloud
- Public Cloud
- Hybrid Cloud

ライフサイエンスデータ解析環境: Bayes Linux on RIKEN Cloud System

2015.4より運用開始： 次世代DNAシーケンサーのデータ解析を自動で行うクラウドシステム

シーケンス拠点

シーエンスデータ

2016年オープンサイエンスアワード受賞

※理研クラウドのハードウェアは情報システムが管理

ソフトウェア開発・提供
解析環境の自動構築

理研クラウドコンピュータ

バイオインフォマティクス
研究開発ユニット

解析結果

計算投入

ユーザ開発
ツールの提供

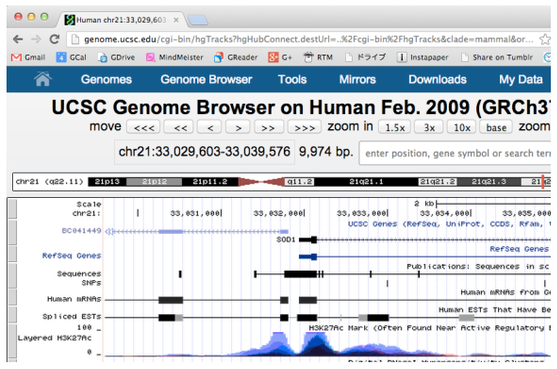
コンサルテーション
チュートリアル

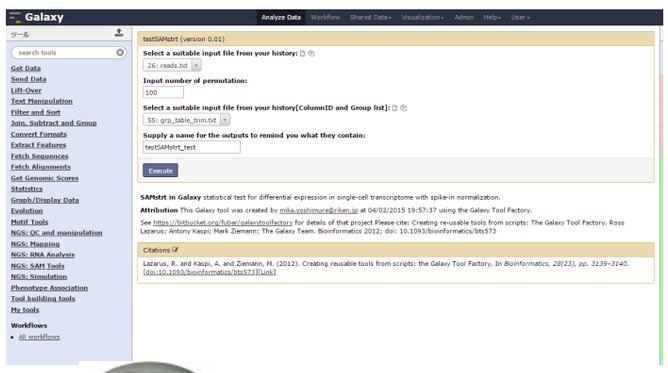
試料

解析実行・可視化

SSH, HTTPS

ユーザー
(理研内外)





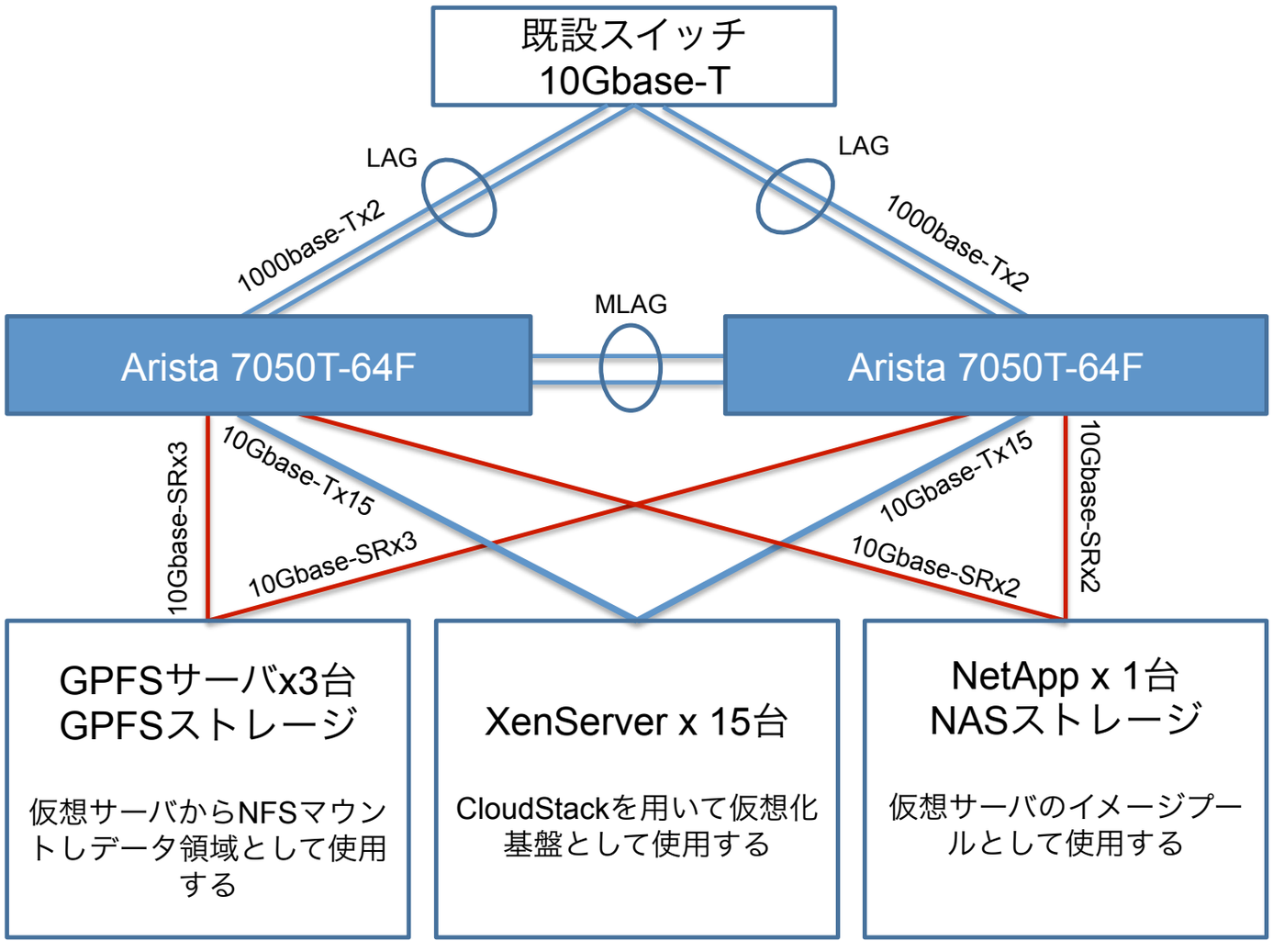
解析パイプライン
データ可視化

理研全所の15ラボを支援。さらにのべ7大学に提供中

計算機購入と運用人件費、電気代・スペース利用料、調達作業コストなどを大きく削減

理研クラウドシステム (ビックデータ基盤)

理研内へ計算資源を提供する実験サービスを開始 (2015.4)



IBM DCS3700
+NDSサーバ 総容量:544TB

HP DL360p Gen8 x 12
CPU: Xeon E5-2670 2.6GHz(8コア) x 2
MEM: 512GB(DDR3-1333 32GBx16)
HDD: 2.5インチSAS 146GB 15Krpm x 3
NIC: 10GbE-Tx2、1000Base-Tx4

DELL PowerEdge R620 x 3
CPU: Xeon E5-2650v2 2.6GHz(8コア) x 2
MEM: 512GB(DDR3-1333 32GBx16)
HDD: 2.5インチSAS 146GB 15Krpm x 4
NIC: 10GbE-Tx2、1000Base-Tx4

NetApp FAS2220 5.2TB 10G NIC

- ハードウェア
  - 1,170 1GHz vCPUs
  - 10 TB RAM
  - 734 TB Strage
- ソフトウェア
  - CloudStack
  - Xen
- 導入・運用コストはオンプレミスと変わりがない

1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

クラウドの利用

- Private Cloud
- Public Cloud
- Hybrid Cloud

データ解析用スパコンをクラウド上に自動構築

1コマンド/クリックで、欲しいときに、欲しいだけ、自分専用スパコンを。マイクロソフトとの共同研究

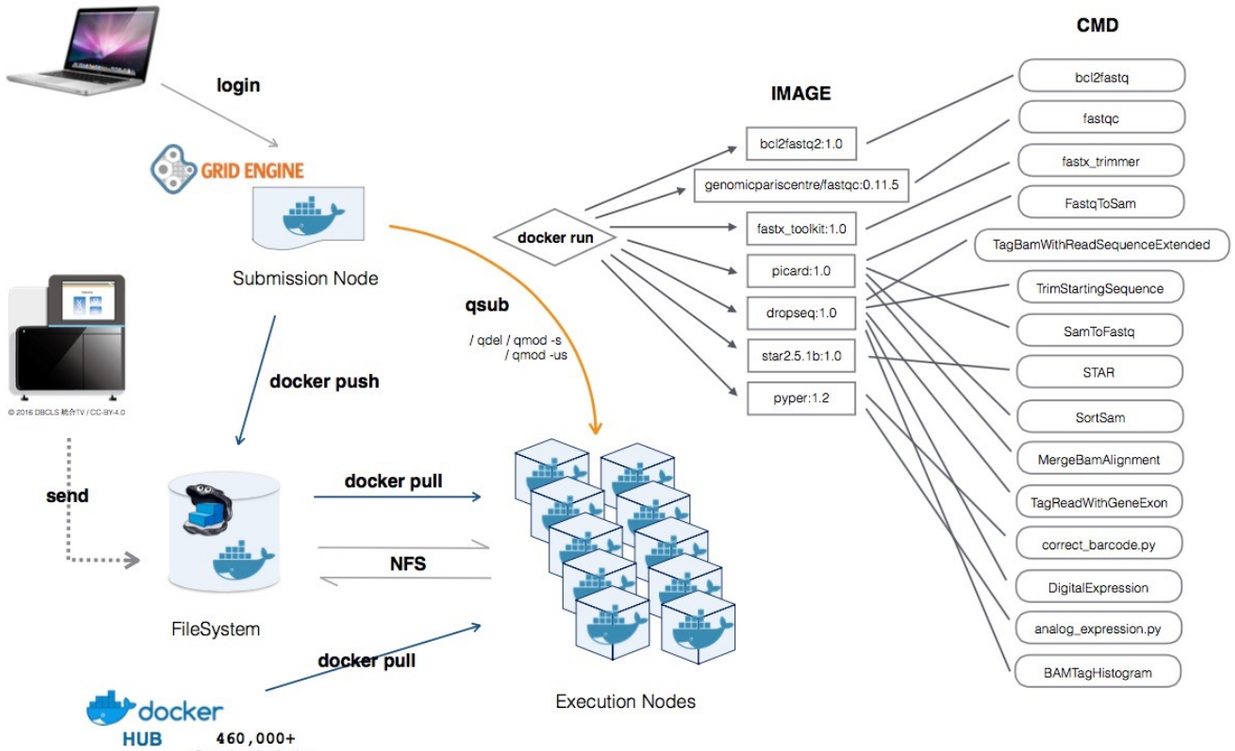
Web上のボタンをクリック/1コマンドで計算機が手に

Deploy to Azure

OR

```
azure group deployment create -g AZURERESOURCEGROUP6 \
  -n Deploy1 -f azuredeploy.json -e local.parameters.json -v
```

仮想計算機とクラウドを利用し、スパコンを自動構築し、計算を投入



Cost

	Cost
MacBookPro	$3000
On-premise	$50000 ~
Cloud	**$200/run**

Execution Time

	Execution Time
MacBookPro	Not finished
On-premise	half a day
Cloud	half a day

Reproducibility

	Reproducibility
MacBookPro	No - if manually
On-premise	Hard - procedure by
Cloud	**YES**

Procurement Time

	Procurement Time
MacBookPro	1 day ~ 2weeks
On-premise	half a year
Cloud	**15 min**

Microsoft

日経ビジネスオンライン

This compare is done by Not Galaxy Pipeline. This compare is illumine NextSeq500 1 run , almost 2000 single-cell RNA-seq.And compare only computational anaylisys.

https://github.com/manabuishii/NGS5th/

1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

クラウドの利用

- Private Cloud
- Public Cloud
- Hybrid Cloud

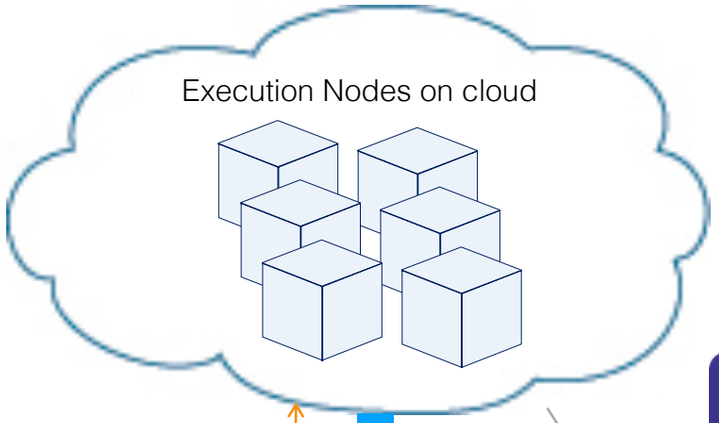
未解決な問題

- ローカルにある巨大なデータをクラウドに移動する必要がある
  - データ転送料金 vs. ストレージ費用
- 普段の計算機を自然に拡張できないか?
  - オンデマンドハイブリッドクラウドの実現へ

オンデマンドハイブリッドクラウド

計算ノードを欲しいときに欲しい量だけパブリッククラウドから調達

- NIIがVirtual Cloud Provider (VCP)を提供
- 理研のPCクラスターのノードをAWSから調達
- ジョブキューイングシステムから自然に計算



大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

オンデマンドクラウド構築サービス

構築・管理ソフトウェアVCP

login

qsub

SINET5

NFS

GRID ENGINE

RIKEN 情報システム部

Submission Node

FileSystem

qsub

NFS

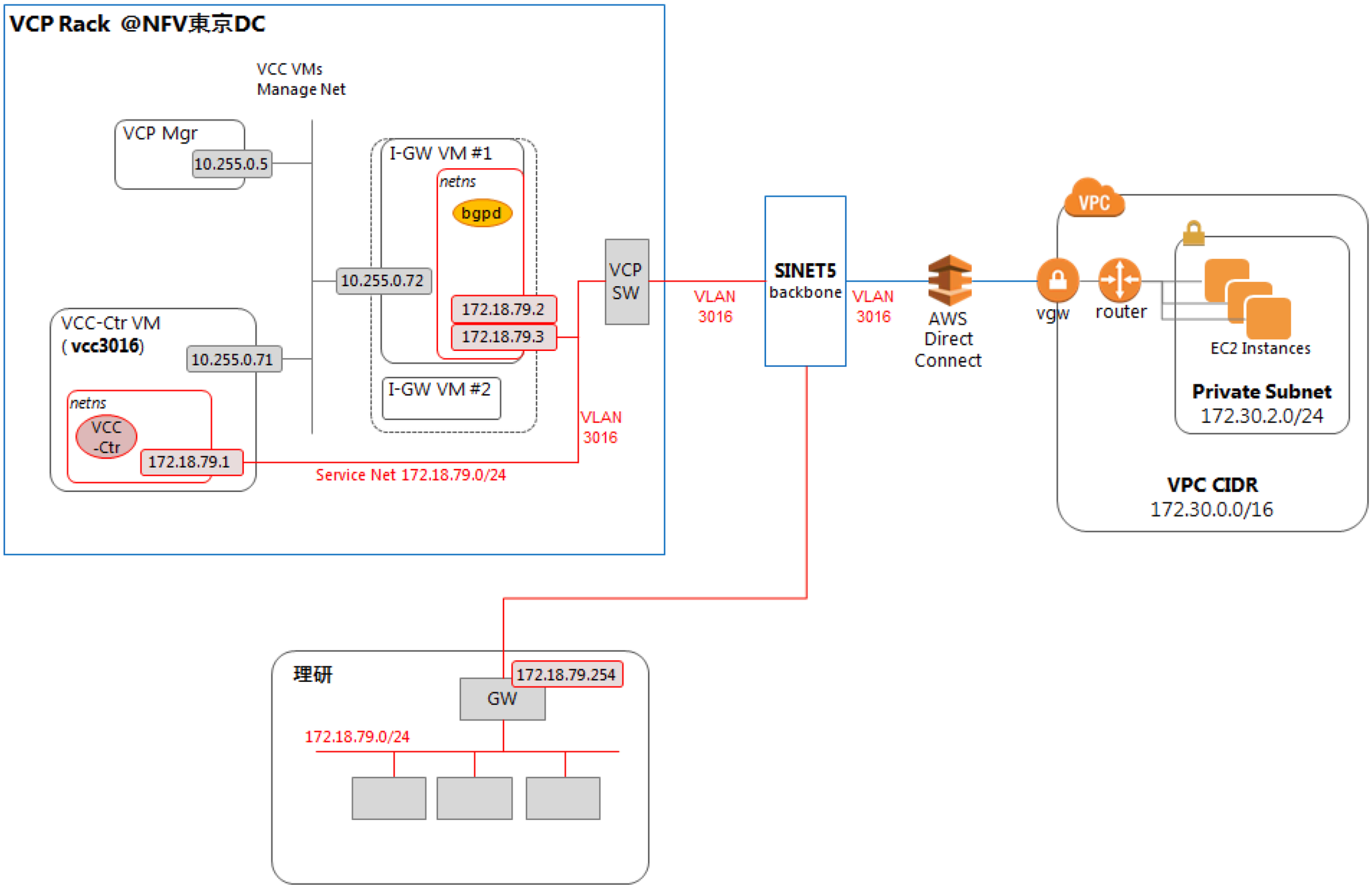
send

Execution Nodes on-premiss

数100万円～数1000万円

OHCのネットワーク構成

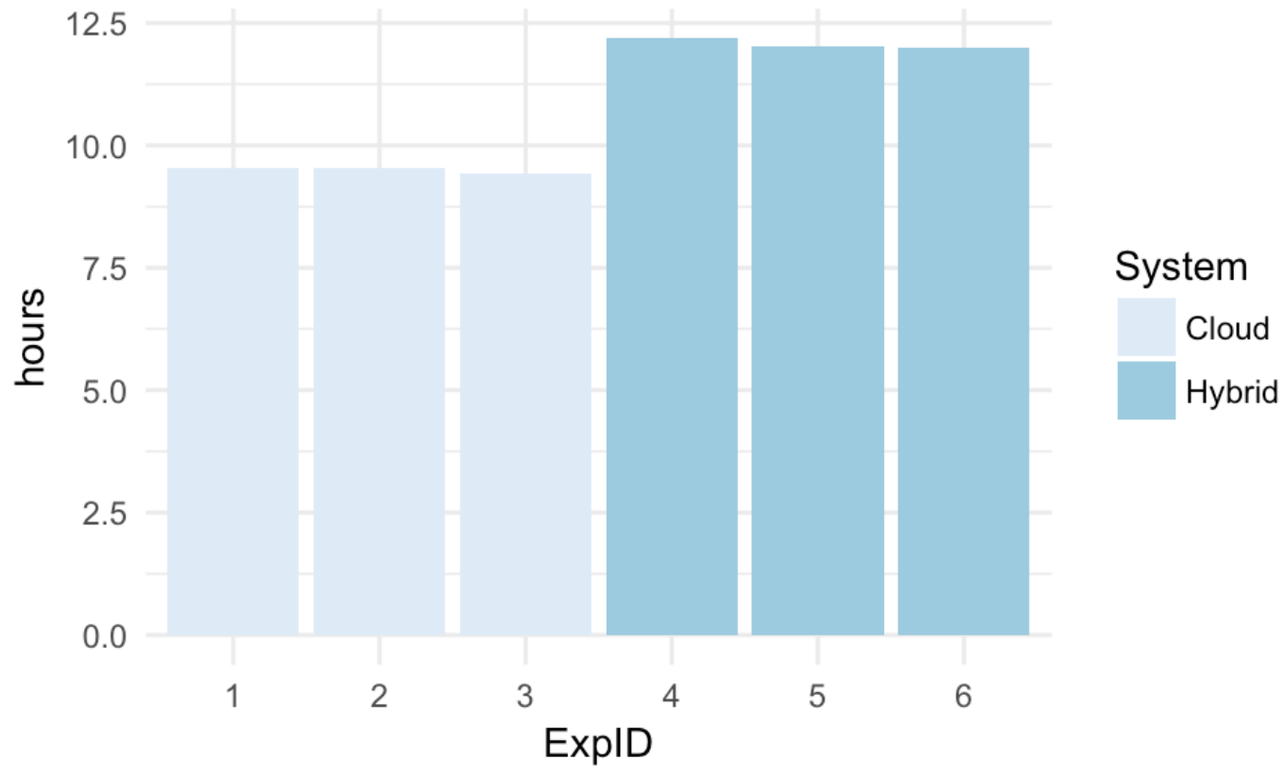
計算ノードを欲しいときに欲しい量だけパブリッククラウドから調達

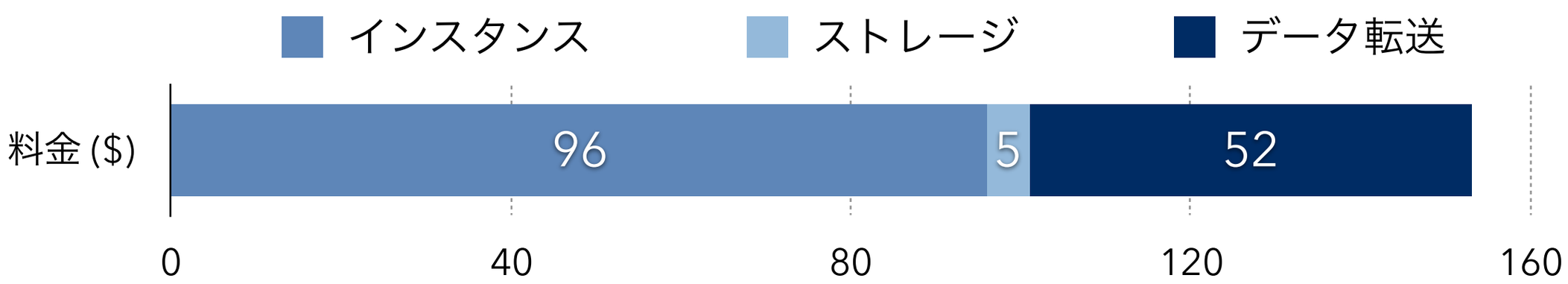


計算速度とコスト: Hybrid Cloud vs. Public Cloud

実験条件: AWS, m4.10xlarge x 4, EBS 20GB, 593GB転送, 2000サンプル

総計算料金: $153.02





展望: ローカルにあるヘッドノードのチープ化

課題

- **通信速度**
- **信頼性/堅牢性**

1. ライフサイエンスを取り巻く計算機事情

大規模・複雑な解析を実施できる計算環境
実験生物学者が計算しなければならない
データ解析環境構築と解析の再現性の向上

2. クラウド技術によるライフサイエンス研究の生産性向上

DevOps技術

クラウドの利用

- Private Cloud
- Public Cloud
- Hybrid Cloud

HPC OPS研究会

第2回 HPC OPS 研究会のお知らせ

第2回 HPC OPS研究会

概要

HPC OPS (えいちぴーしー おぷす) 研究会は、自然科学の研究成果を最大化するための科学計算環境やその構築技術についての研究会です。計算環境構築の時間やコストを低下させ、本来の研究活動に多くの時間を割けられるよう科学計算環境の開発・運用のノウハウを共有します。また、そのような計算環境そのものを研究開発したり、提供する研究者や技術者との交流を目指します。

具体的な技術としては、コンテナ型仮想計算やクラウドでのHigh Performance Computing、DevOps による科学計算環境の自動構築、データ解析ワークフローエンジンの実装や利用、最適なオンプレミスPCクラスタの運用構築などについて議論します。産学官などの垣根を越えて、クラウドやDevOps, HPCに関わる技術者や科学者などからの参加を広く募集します。

- 日時：2018年7月2日、13時30分-18時
- 場所：日本マイクロソフト (品川)、31階セミナールーム（C+D)
- 参加申し込み： こちらの参加申し込みフォームからお申し込みください。
  - 6月29日金曜日 13:00 まで
  - 席が埋まりましたらその前に締め切らせていただくことがあります
- 主催: 理化学研究所. 協賛: 日本マイクロソフト
- 問い合わせ： support-bayes at riken dot jp

講演者 (敬称略)

- 政谷好伸、国立情報学研究所 「NIIでの計算機環境の運用及び、Literate Computing(for reproducible infrastructure)について」
- 白石友一、国立がん研究センター 「Extraction Transformation Load (ETL)アプローチに基づくがんゲノム解析パイプラインの開発」
- 海津一成、理化学研究所 「TBA」
- 佐藤仁、産業技術総合研究所 「TBA」
- 近藤宇智朗、GMOペパボ株式会社 「TBA」

7/2 13:30

日本マイクロソフト (品川)

参加受付中

Twitter: #hpcopsjp

https://bit.riken.jp/

謝辞

理研バイオインフォ研究開発ユニット

- 石井学
- 松嶋明宏
- 芳村美佳
- 團野宏樹

理研情報システム部

- 黒川原佳
- 加茂聡
- Cho Wukui

国立情報研究所

- 合田憲人
- 竹房あつ子
- 丹生智也
- 政谷好伸

日本マイクロソフト

- 林勝典
- 久木田弦
- 中田寿穂

敬称略