



北海道大学

北海道大学ハイパフォーマンス インターネットクラウドの概要

～ハードウェア調達から継続的なソフト力の強化へ～

北海道大学 情報基盤センター
システムデザイン研究部門 杉木章義

本日のアウトライン

システムのご紹介
ハードウェア調達から継続的進化へ
利用者から見たOCTOPUS
次期システムへ向けて

本日のアウトライン

システムのご紹介

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

次期システムへ向けて



北海道大学

北海道大学 情報基盤センター

大型計算機センターとメディア教育センターを統合して設立

- 1962年：大型計算機センター発足（全国共同利用施設）
- 1979年：情報処理教育センター発足（後の情報メディア教育総合センター）
- 2003年：両センターを統合、情報基盤センター発足

学内では、「研究センター」の位置付け

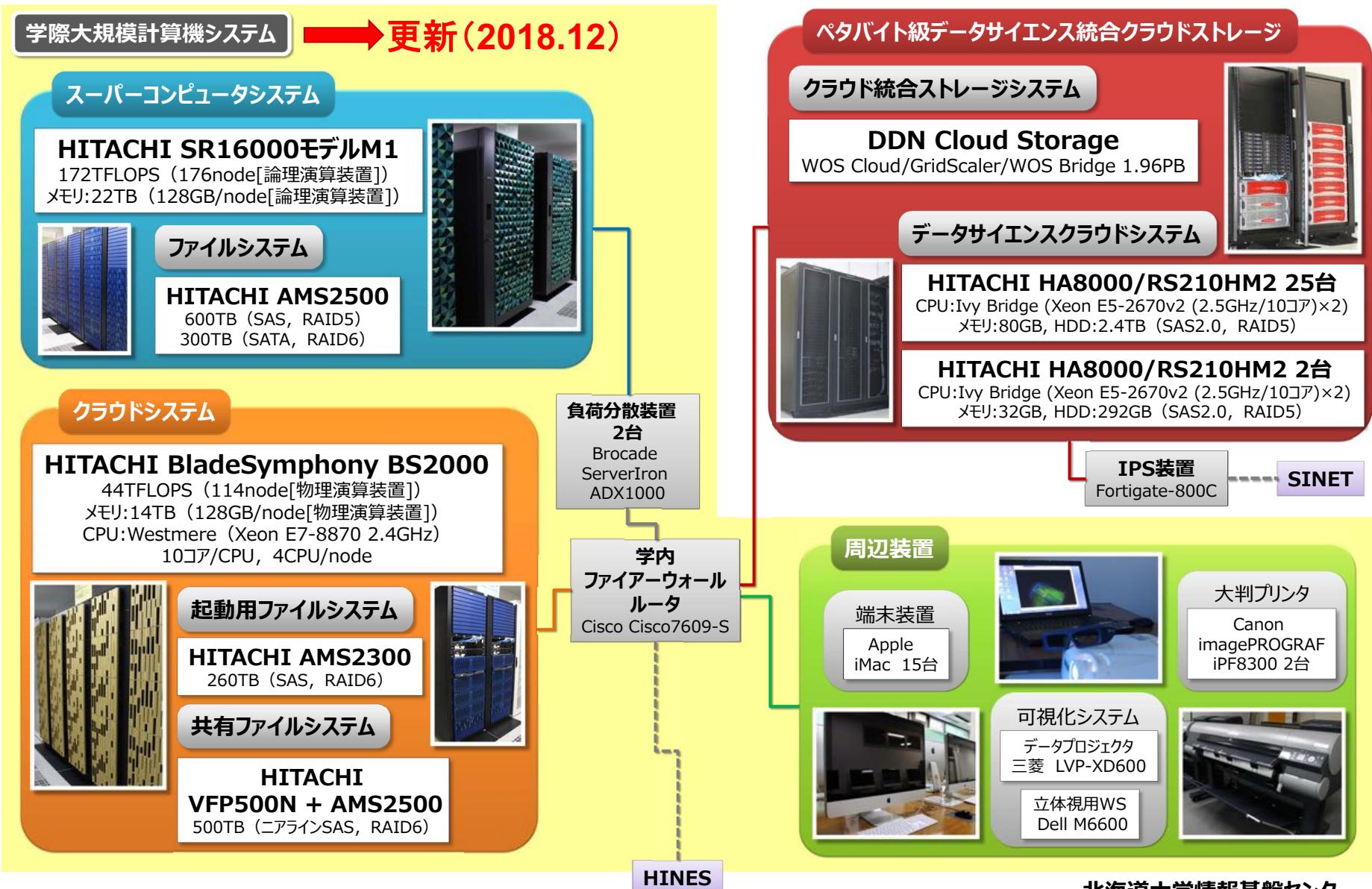
- JHPCN（学際大規模情報基盤共同利用・共同研究拠点）構成拠点
- HPCI（High Performance Computing Infra.）資源提供機関

スパコン・ネットワーク・クラウド・情報メディア教育・コンテンツ、
システムデザイン・サイバーセキュリティ
などに関する研究を実施



北海道大学

以前の学際大規模計算機システム(2018年11月まで)



学際大規模計算機システム(2018年12月運用開始)

(通称: 北海道大学ハイパフォーマンスインタークラウド)



学際大規模計算機システムパンフレットから引用

前システムに続き、スパコンとクラウドを併置したシステム

スーパコンピュータシステム:

- x86アーキテクチャへの変更
 - サブシステムA(Skylake)
 - サブシステムB(KNL)

インターネットクラウドシステム:

- ソフトウェアスタックを刷新
 - OpenStack, Nextcloud
- インターネットクラウド化
 - 遠隔サイト(阪大・東大・九大), 北見
 - SINET5への対応(100Gbps化, 各種VPNサービス)



北海道大学

写真: 学際大規模計算機システム

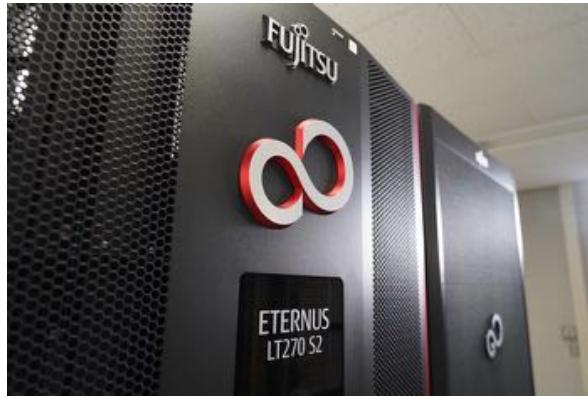
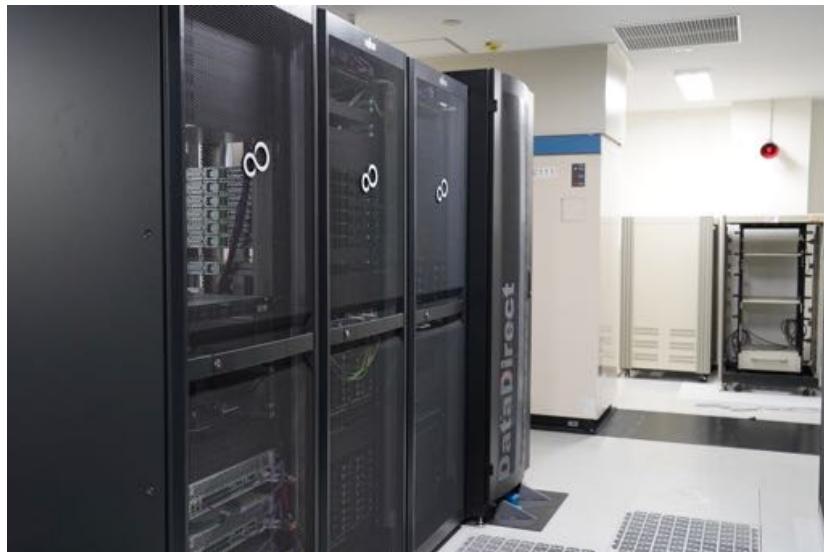


写真: スーパーコンピュータシステム



北海道大学

写真：インタークラウドシステム



磁気テープアーカイブ装置は
北見工業大学に設置



北海道大学

本日のアウトライン

システムのご紹介

- スーパーコンピュータシステム
- インタークラウドシステム

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

次期システムへ向けて



北海道大学



Intel Omni-Path

サブシステムA (Gran Chariot)

1,004ノード



ノード構成

- Intel Xeon Gold 6148 x2
(2.4GHz, 20コア)
- 3.07TFlops
- 384GB memory
- 240GB SSD

サブシステムB (Polaire)

288 ノード



ノード構成

- Intel Xeon Phi 7250
(1.4GHz, 68コア)
- 3.04TFlops
- 96GB memory
- 64GB SATA Flush

ソフトウェア: Linux OS, Fortran/C/C++ compiler, MPI library, Intel MKL, Applications

**総演算性能
3.96PFlops**



北海道大学

サブシステムA / B の概要

サブシステムA

ノードの構成

CPU	Intel Xeon Gold 6148 x 2個 (Skylake, 2.4GHz, 20コア)
メモリ	384GB
ストレージ	240GB SSD
OS	Cent OS

理論演算性能: 3.07FLOPS/ノード

×

1,004ノード

(ノード間ネットワーク: Intel Omni-Path)

サブシステムB

ノードの構成

CPU	Intel Xeon Phi 7250 x 1個 (KNL, 1.4GHz, 68コア)
メモリ	96GB (+16GB 高速メモリ)
ストレージ	64GB SATA Flush
OS	Cent OS

理論演算性能: 3.04FLOPS/ノード

×

288ノード

(ノード間ネットワーク: Intel Omni-Path)



新スパコンの概要: ソフトウェア環境

- Intel製の開発環境やライブラリを提供
 - C, C++, Fortranコンパイラ
 - VTune Amplifier, Trace Analyzer等のデバッガ・プロファイラ
 - MPI, MKL(数値計算), DAAL(ビッグデータ解析・機械学習)等のライブラリ
 - PythonやJavaの処理系を整備
 - 様々な分野のフリーソフトウェアを整備
 - 計算科学: OpenFOAM, PHASE, MEEP等
 - 機械学習(深層学習): Chainer, Tensorflow, Caffe
- ※サブシステムAでは有償ソフト(Gaussian, V-FaSTAR)も提供



新スパコンのサービスの概要

1. 利用者番号の取得 → 基本サービスを利用可能

- ・ スパコンストレージのhome領域100GBの利用
- ・ 試用・デバッグ用の共用ノードの利用
(4ノード程度以内, 経過時間1時間まで, 全ユーザが利用)
- ・ アプリケーションサーバの利用(民間ユーザは不可)

※他にも利用可能なサービスあり(スパコン関連のみ記載)

2. 付加サービスの追加(グループ利用可能)

- ・ 共用ノードの利用
- ・ 占有ノードの利用
- ・ 追加のスパコンストレージ領域(home領域, work領域)



付加サービスの詳細: 共用ノードと占有ノード

- **共用ノード利用: 演算時間(トークン)を申請**

- 演算時間(トークン)を消費してジョブを実行
(演算時間 = 利用ノード数 × 経過時間)
- 他のユーザと計算リソースを共用
- 多数のノードを利用するような大規模ジョブにも対応

- **占有ノード利用: ノードを申請**

- 申請したノードを占有してジョブを実行
(他のユーザのジョブを待つ必要なし)
- (どれだけジョブを実行しても)定額
※誤ったプログラムを実行しても問題なし



付加サービスの詳細: グループ利用

前提: グループのメンバーは利用者番号を保持
(例: A, B, C の3名)

1. Aが付加サービスを申請
2. AがBとCとグループのメンバーに追加
(利用者番号をメールアドレスを指定して追加)
3. BとCもAが申請した付加サービスを利用可能に
 - ・ 演算資源: ジョブスクリプト内で所定の方法で指定
 - ・ スパコンストレージ: work領域の所定の場所を利用可能

※各利用者が構成できる(代表となる)グループは原則1つ



新スパコンのサービス利用料金(負担金):一般コース

基本負担金 (スパコン・クラウド共通)	一般	¥ 12,960
	学生	¥ 2,160
サブシステムA 付加サービス	共用ノード 演算時間 (トークン)	3,000,000秒 (約35日) ¥ 24,000
		15,000,000秒 (約174日) ¥ 81,000
		100,000,000秒 (約3.1年) ¥ 405,000
		250,000,000秒 (約7.8年) ¥ 810,000
	占有ノード	1ノード (3TBのwork領域付き) ¥ 93,000
サブシステムB 付加サービス	共用ノード 演算時間 (トークン)	3,000,000秒 (約35日) ¥ 19,500
		15,000,000秒 (約174日) ¥ 66,000
		100,000,000秒 (約3.1年) ¥ 330,000
		250,000,000秒 (約7.8年) ¥ 660,000
	占有ノード	1ノード (3TBのwork領域付き) ¥ 78,000
スパコンスト レージ 付加サービス	home領域	1TB ¥ 20,000
	work領域	3TB ¥ 30,000

(注意)全て年額で、年度内の利用に限る。民間企業等利用コースは別料金。



本日のアウトライン

システムのご紹介

- スーパーコンピュータシステム
- インタークラウドシステム

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

次期システムへ向けて



北海道大学

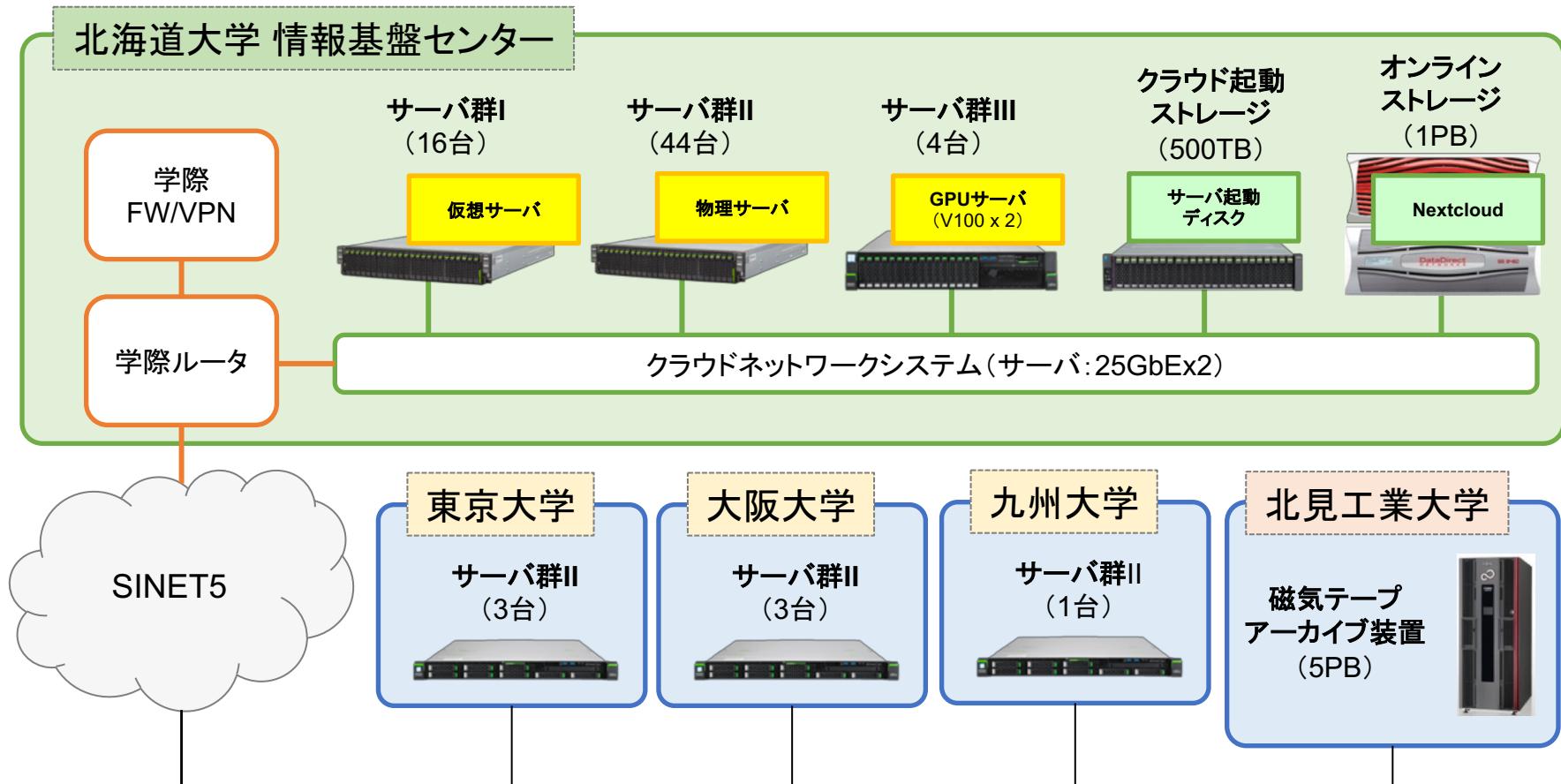
新クラウドシステムの概要

- **研究目的でのクラウド活用を重視**
 - ベアメタル(物理)サーバを中心に提供(高性能クラウド)
- **ソフトウェアスタックを前システムから刷新**
 - OpenStack(ベアメタル用Ironic含む), Nextcloud
- **インタークラウドを活用した研究のサポート**
 - 北大・東大・阪大・九大に配置した計算機 + SINET L2VPN
 - SINETプロジェクトと連携(計画)
 - オンデマンドクラウド構築, 広域データ収集基盤, GakuNin RDM



インタークラウドシステム概要(ハードウェア編)

北大情報基盤センターに加えて、東大・阪大・九大・北見工大に機器を設置



インタークラウドシステム概要(ソフトウェア編)

サーバ基盤



openstack.

OpenStack (Mirantis Cloud Platform)

- 仮想マシンをIaaSで提供
- ベアメタルもIronicで統一的に管理(一部手動あり)



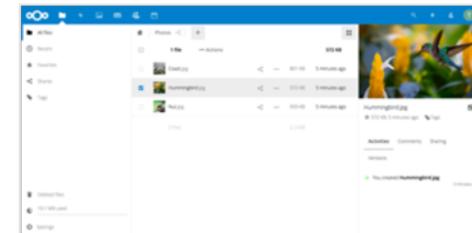
ストレージ基盤



Nextcloud

Nextcloud

- ownCloudから派生したWebDAVストレージ
- Dropboxなどオンラインストレージと同等の使い勝手



前システムのCloudStackから変更

ownCloudから変更



北海道大学

提供サービス

サーバサービス

仮想サーバ

仮想マシンを2コア以上
1コア単位で提供(**柔軟性を重視**)

物理サーバ

物理マシン(40コア)を
1台単位で提供(**性能を重視**)

GPUサーバ

Tesla V100搭載サー
バを1台単位で提供

※最大4台

インターネットサービス

※原則、HPCI-JHPCNのみ

インターネットパッケージ(3拠点または4拠点)

北大・東大・阪大・九大に設置した物理サーバをSINET L2VPNで接続して提供

ストレージサービス

クラウドストレージ(Nextcloud)

大学に設置したストレージによるファイル共有サービス(**一般100GB/学生10GB無料**)



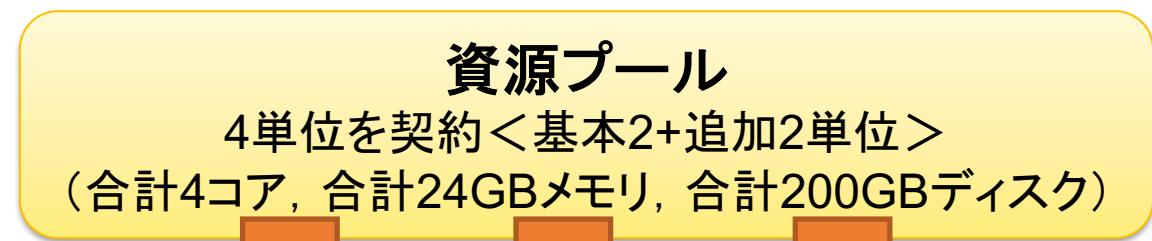
北海道大学

新サービス(1): 仮想サーバ

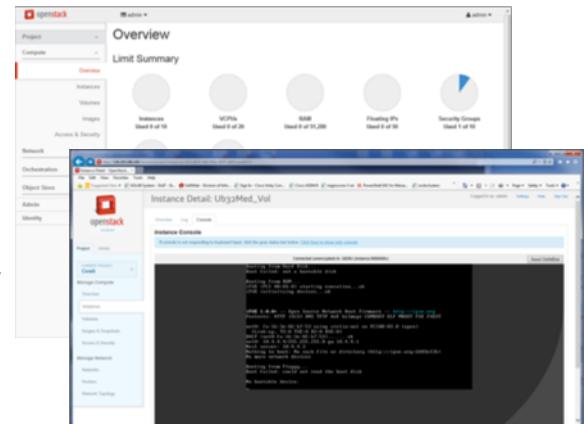
資源プール方式で提供

- 合計コア数, 合計メモリ容量, 合計ストレージ容量を指定
- 複数の仮想サーバに合計リソースを資源プールから再配分
- 本学では, 1コア/6GB/50GBを1単位として提供(最低2単位以上)

OpenStackのダッシュボード(管理画面)を提供



資源プールから
仮想マシンを作成



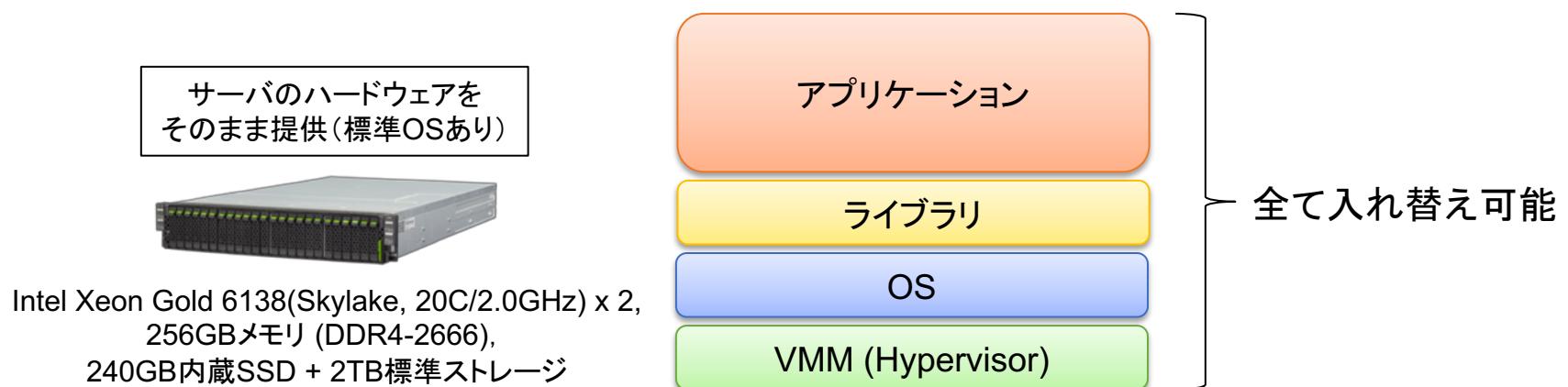
仮想サーバの利点:

- 研究予算に応じて, さまざまなサイズのマシンを作成できる
- サーバの使い捨てが用意(短時間の作成と破棄の繰り返し)

新サービス(2)：物理サーバ

物理サーバを1台単位でまるごと提供

- OS, ライブラリ, アプリケーションの全てのソフトウェアを入れ替え可能(管理者権限あり)
- OpenStackダッシュボード上で仮想サーバと同じ画面で管理
- 仮想サーバと比較して, 20コア以上は物理サーバがお得



物理サーバの利点:

- ・ 高性能, 仮想化のオーバーヘッドなし
- ・ 旧XLサーバよりも低廉な負担金, 豊富な台数



新サービス(3) : GPUサーバ

Volta世代のGPUを物理サーバで提供

- Tesla V100 PCIe/16GB × 2

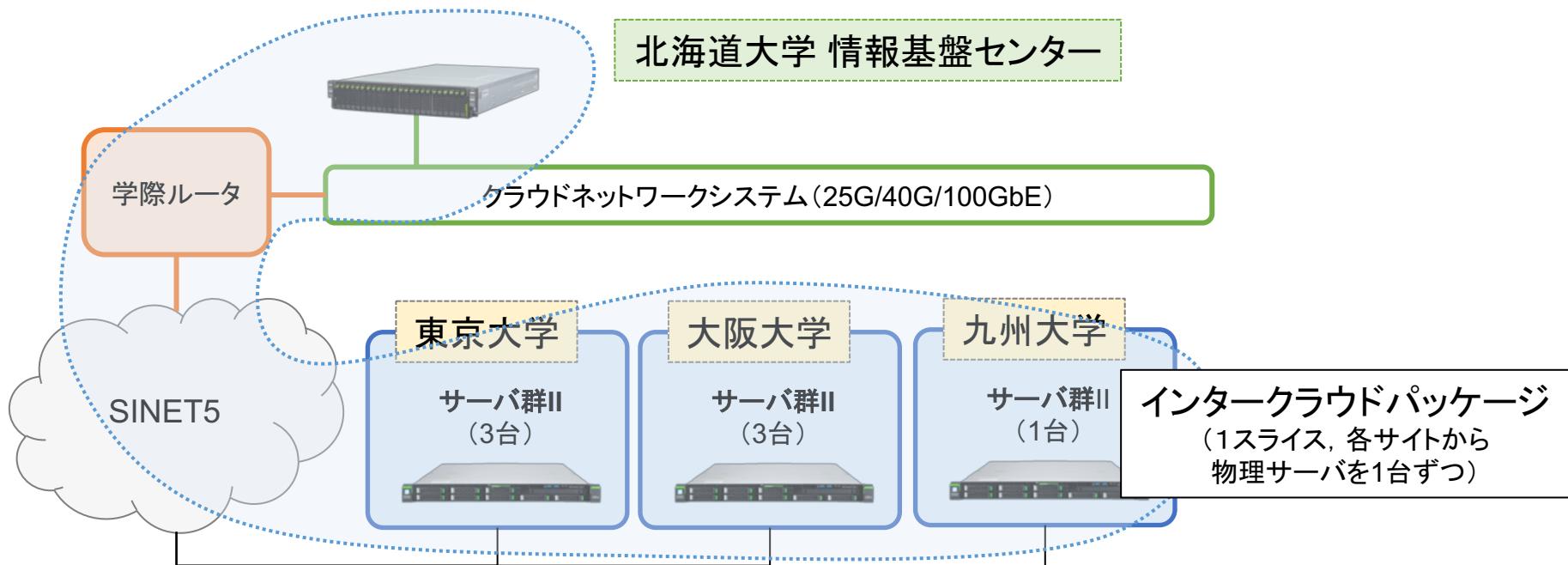
利用負担金と性能のバランスを最適化

- パイロット的な研究開発を想定
- 本格計算は、最近整備の進むGPUスパコンへ



複数サイトの物理サーバをスライスとして一括提供

- 北大の遠隔サイトの位置付けのため、面倒な交渉は一切不要
- 構成拠点：東京大学(3台)，大阪大学(3台)，九州大学(1台)
- SINET5 L2VPNによる仮想プライベート接続



新サービス(5)：クラウドストレージ

Nextcloudベースのクラウドストレージ

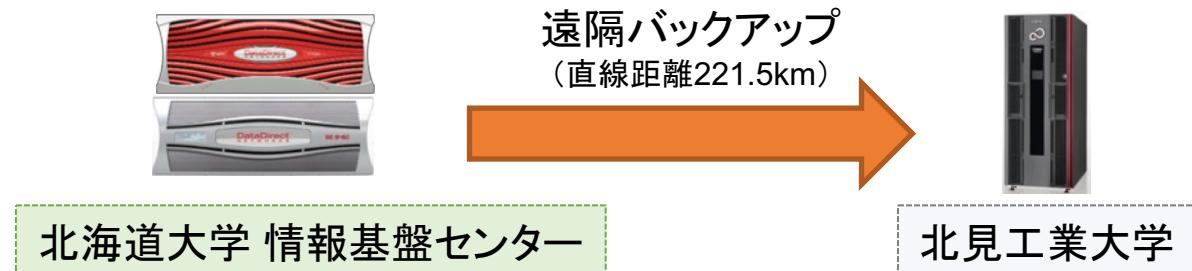
- Web, PC, スマートフォンからのストレージ操作に対応
- Dropboxライクな使い勝手

大容量

- 基本サービス：**一般100GB/学生10GB無料**, 追加：1TBにつき月額500円

高い信頼性・安全性

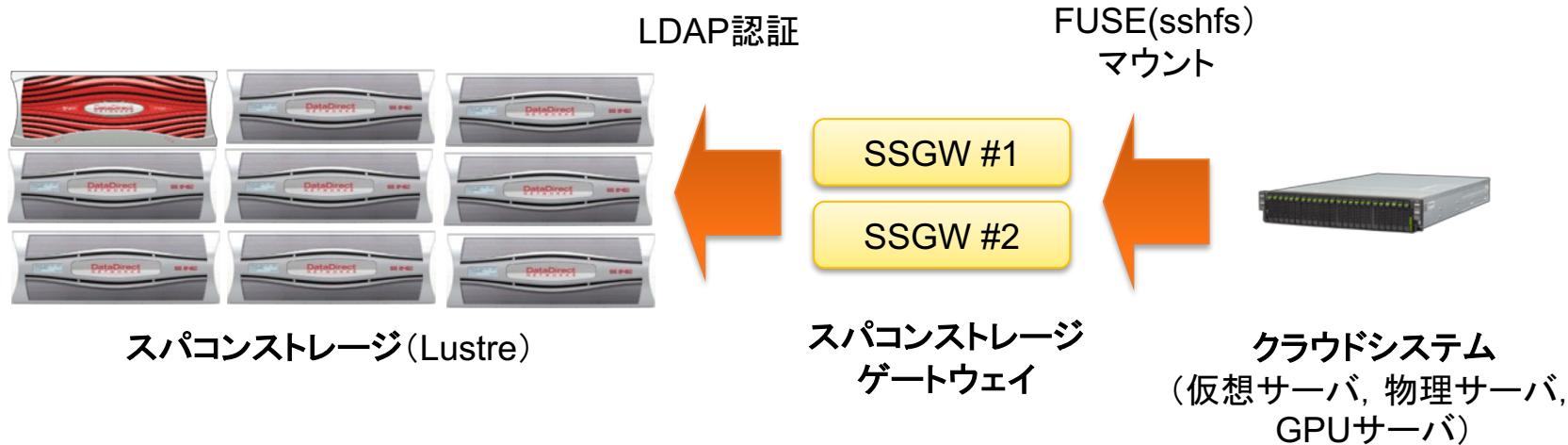
- データは北大内のストレージに保存
- 北見工大のテープアーカイブ装置に自動遠隔バックアップ



スパコン・クラウド連携

スパコンストレージゲートウェイ(2台)を新規に導入

- クラウド側の仮想・物理・GPUサーバから、スパコンのストレージを認証付きでマウント、高速にアクセス可能
- FUSE(sshfs)を使用
- (活用例) クラウドによるプレポスト処理、スパコン計算結果の解析



負担金表(1/2)

基本サービス:

提供資源	性能	負担金
基本サービス(一般)	基本スペコン+クラウドストレージ(100GB)	年額 12,960円
基本サービス(学生)	基本スペコン+クラウドストレージ(10GB)	年額 2,160円

サーバサービス:

提供資源	性能	負担金
(1)仮想サーバ	基本	2コア/12GBメモリ/100GBディスク
	追加	1コア/6GBメモリ/50GBディスク 追加につき
(2)物理サーバ		月額 14,000円
(3)GPUサーバ	上記に加えて、Tesla V100 × 2	月額 20,000円
追加ストレージ	1TBあたり	月額 500円



負担金表(2/2)

インタークラウドサービス:

提供資源	性能	負担金
(4) インタークラウド パッケージ	3拠点(北大/東大/阪大)	月額 42,000円
	4拠点(北大/東大/阪大+九大)	月額 56,000円

ストレージサービス:

提供資源	性能	負担金
(5) クラウドストレージ	1TB 追加につき (基本サービスに一般100GB/学生10GB含む)	月額 500円

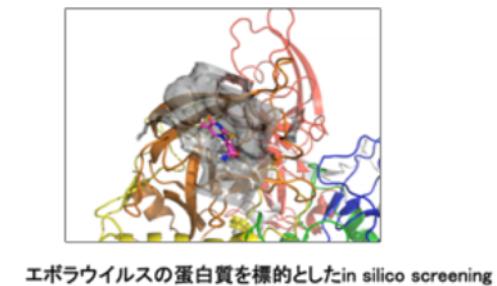
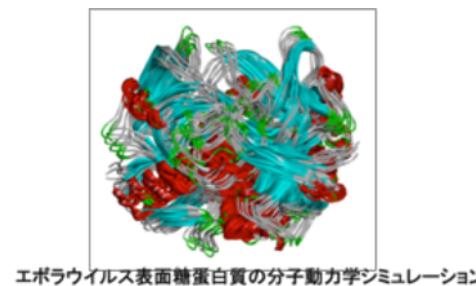
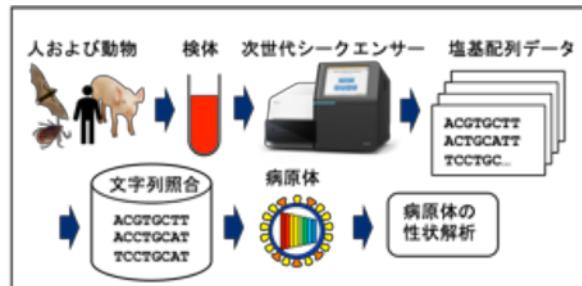


北海道大学

人獣共通感染症リサーチセンター

大量の遺伝子塩基配列データと既知の病原体塩基配列と照合

- 新興感染症の出現は人類の脅威
 - 人, 野生動物, 家畜, 節足動物の検体に含まれる遺伝子
- ## 感染症の流行および進化ダイナミクスの大規模シミュレーション
- 感染症の流行は複雑で予測が困難, 効果的な介入が行えず
 - 感染症流行制圧

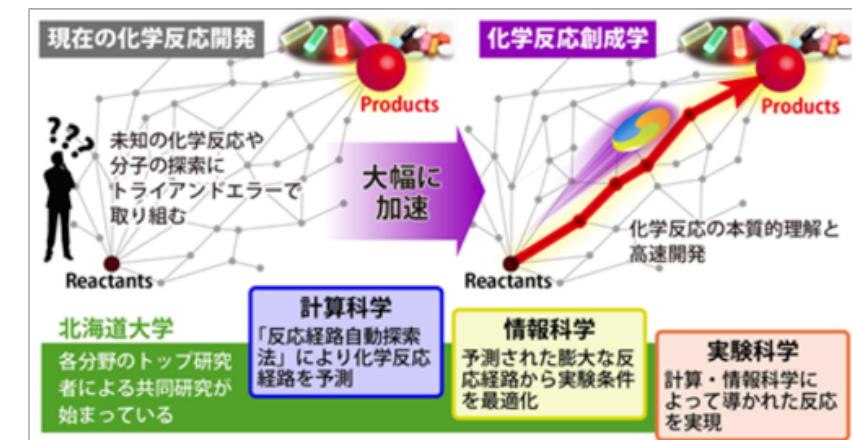


化学反応創成研究拠点(WPI-ICReDD)

計算科学・情報科学・実験科学の三分野融合により、化学反応を複雑なネットワークとして理解し、制御することを目指す

- 反応経路自動探索により、化学反応経路ネットワークを算出
- 情報科学で、実験的に検討する情報を抽出し、実験条件を絞り込み
- 実験科学のデータを、情報科学を通じて、計算科学へフィードバック
化学反応の高度なデザインと迅速開発の実現を目指す

新スパコンは、反応経路ネットワークの算出等における計算の劇的な加速に関与



SINET各種サービスとの連携

(1) 学認クラウドオンデマンド構築サービス

- 2018年10月1日事業化

- 仮想サーバ・物理サーバの上でアプリケーションを自動構築

(2) SINETデータ収集基盤 実証実験 公募

- SINET L2VPNを通じてモバイル網とプライベート接続

(3) オープンサイエンス管理基盤 GakuNin RDM

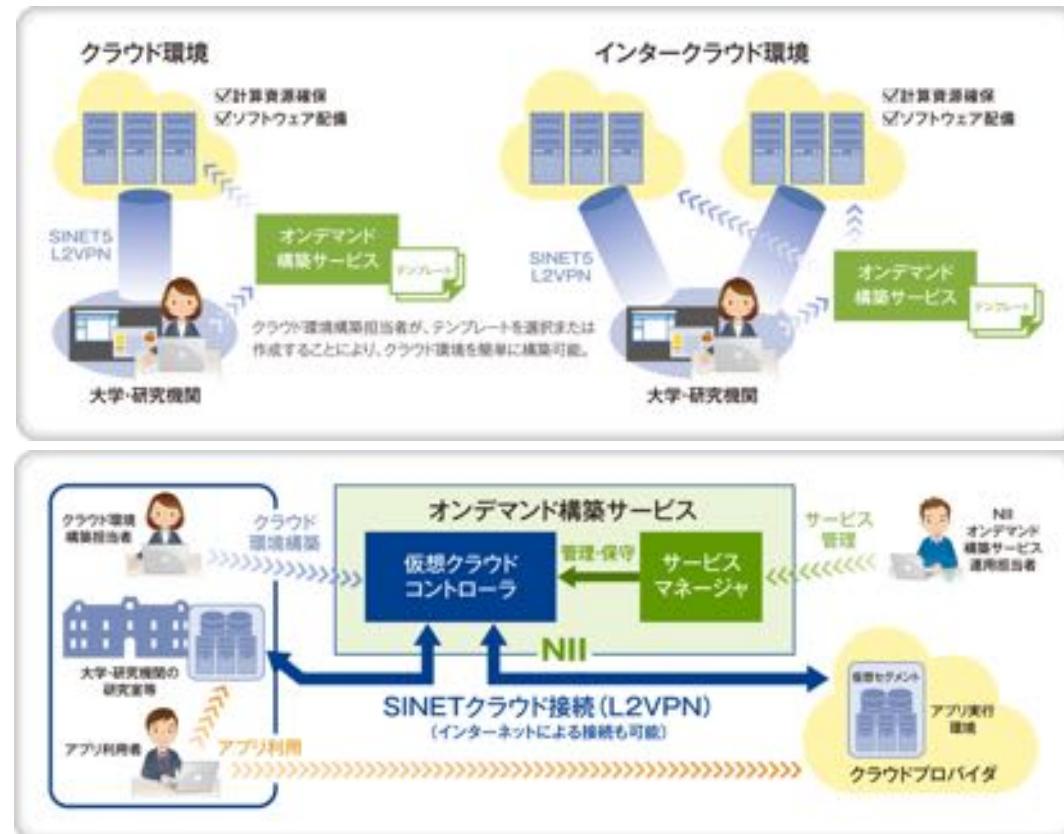
- 研究データのマネージメント

- Nextcloudのストレージと連携



学認クラウドオンデマンド構築サービス

大学クラウドとしてオンデマンド構築のための資源を提供



北大としては、調達時に手薄い
アプリケーション支援の部分で
オンデマンド構築サービスを活用

北大の標準機能では、仮想サーバ・物理サーバを、箱として貸し出すのみ

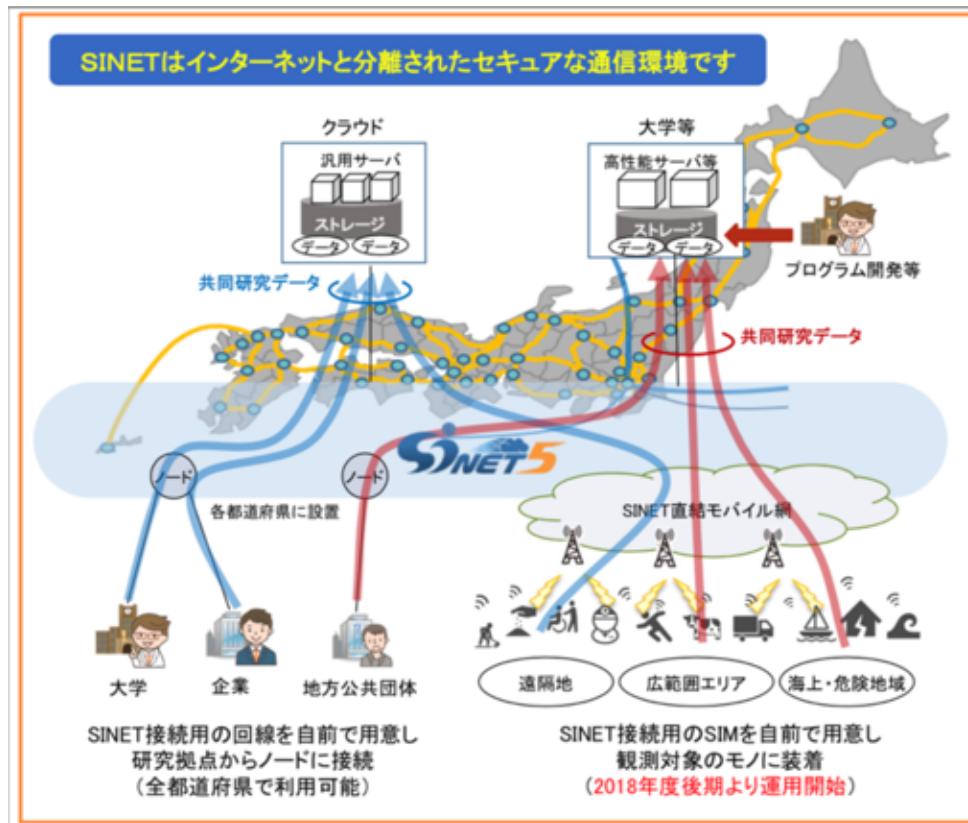
L2VPN/L2ODの活用もある程度想定
(SINET～クラウド間のバイパス線、
VLAN IDをNII連携用に確保)

引用: 学認クラウド(国立情報学研究所),
学認クラウドオンデマンド構築サービス
<https://cloud.gakunin.jp/ocs/>



SINETデータ収集基盤 実証実験 公募(1/2)

SINET L2VPNを通じてモバイル端末(SIMカード)と 北大クラウドをプライベート接続



利用者はSIMカードを購入
モバイルデータ通信量を負担

北大は大学クラウドを提供

SINETがモバイルと
大学クラウドの間で
高速な閉域網を提供

引用: 学術情報ネットワークSINET5,
2018年度SINET広域データ収集基盤実証実験
<https://www.sinet.ad.jp/wadci>



北海道大学

オープンサイエンス管理基盤 GakuNin RDM

GakuNin RDMのバックエンドストレージとして クラウドストレージ(Nextcloud)を提供

- 各利用者の所属機関がGakuNin RDMに参加
- 学際システムの利用者番号・パスワードでNextcloudを指定



前提:
利用者の所属機関が
GakuNin RDMに参加

RCOSによるサービスの概要

・複数リポジトリ・各種開発プロジェクトやデータリポジトリとも連携
・専門用語や研究プロセスの用語と連携
・既存のデータベースを統合
・研究者による発見のプロセスをサポート

National Institute of Informatics

バックエンドとして
Nextcloudを指定
(利用者番号, パスワード)

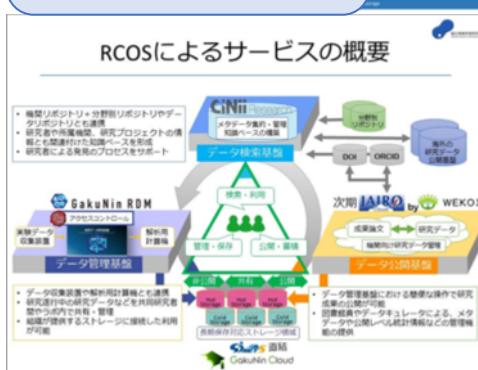


一般100GB, 学生10GBを標準
提供(1TB単位で追加可能)



北海道大学 情報基盤センター

GakuNin RDM
(NIIクラウドサービス版)



引用: 国立情報学研究所オープンサイエンス基盤
研究センター, 管理基盤(GakuNin RDM)
<https://rcos.nii.ac.jp/service/rdm/>



北海道大学

本日のアウトライン

システムのご紹介

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

次期システムへ向けて

新システム導入後の経過(3ヶ月)

成功した点(**strong points** ☺)

- 安定的に運用されている
- OpenStack APIによる自動連携は便利
 - クラウドにはクラウドの良さがあることを再確認

失敗した点(**weak points** ☹)

- 同一世代の成熟した仕様
- 明らかにリソースが足りない
- 供給者側と需要者側の期待のミスマッチ



新システム導入までの経緯

不落→再調達

保守的な路線に変更

- 同一仕様のまま、リソース量を削減
 - 特にクラウド

同一世代で成熟したハードウェアを導入

- 大きな遅れなどもなく、納入自体はスムーズ
- 水冷の一部部品、北海道胆振東部地震の影響により途中、2週間程度の遅れあり(稼働開始には余裕あり)



北海道大学

5年のシステム運用期間(2023年12月まで)

調達サイクルは長い(5年以上)

- **LTE(Long Term Evolution)**
 - 同一世代のまま進化させる
- 主にソフトウェア面
 - サーバの箱貸しではなく、アプリケーションの直接支援が必要
 - コンテナ, Kubernetes?
- パブリッククラウドは調達ルールが変わらないと難しい
 - 現状の製品調達に乗せるのは難易度が高い
 - 単価契約も本当に使いたいのはサーバか？



北海道大学

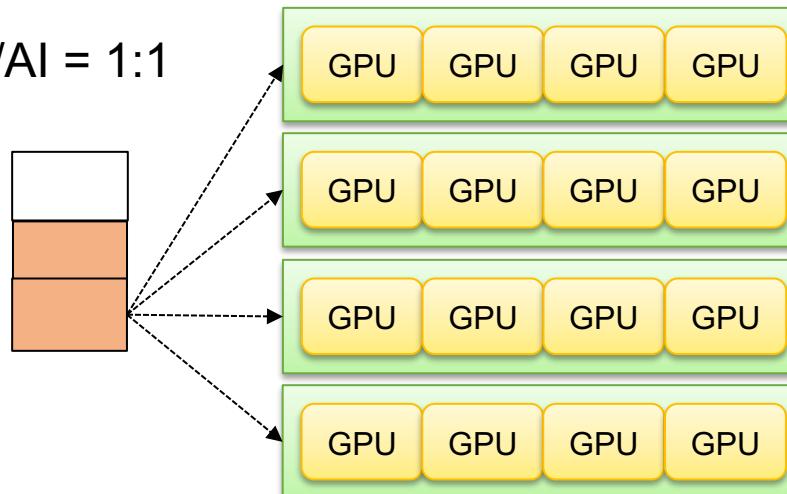
人工知能対応先進的計算機システム(2019年12月-)

学内利用者対象のアクセラレータ搭載クラスタ

- 倍精度演算180TFLOPS以上(6ノード以上)
- オールフラッシュストレージ(70TB以上)
- スパコン仕様(ジョブスケジューラ, ノード間接続)

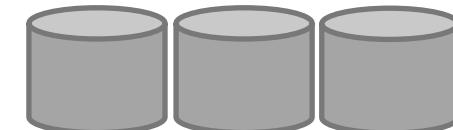
Workload

HPC:ML/AI = 1:1



4 GPUs x 6 nodes
(or 8 GPUs x 3 nodes)

All flush storage (70TB+)



ノード占有 or 時間課金

...



北海道大学

新しい利用者を見据えて

AI/ML研究者はジョブスケジューラが最適解か？

- 研究室と同じように占有して使いたい
 - 年度末の研究室のクラスタが混み合う時期に使いたい
- 理想：フロントエンドはJupyter(Hub), バックエンドは速い
 - いつでも使って、研究室のJupyterより速い環境(ログインノード化)
- GPUの共有利用に課題(デバイス分割)
 - GoogleのColabのようには使えない

コンテナ・オーケストレーションの経験を積んでいなかったことを反省

- Kubernetesでジョブスケジューラの代替は可能か？
- セキュリティ対策、安定的に継続運用していくことは可能か？



北海道大学

クラウドの利用状況(運用開始から3ヶ月)

	1月の利用状況(カッコ内は2月速報値)
仮想サーバ	421/600コア(→利用率 約70%)
物理サーバ(インタークラウド パッケージを含む)	12/22台(→利用率 約87%)
GPUサーバ	3台/4台(→利用率 75%)
移行用サーバ(プロジェクト)	281コア(92台)
移行用サーバ(ホスティング)	146コア(73台)
事務用クラウド	760コア

運用期間中にリソースが不足することは予想していた

- 運用開始2-3年目を想定していたが…

足りなくなったらどうするか？

- 追加調達？パブリッククラウド利用？



運用開始後(蓋を開けてみて)

当初は研究利用を想定して設計

- OpenStack, ベアメタル, SINET5連携, ...

実際には研究支援(グレーゾーン)利用が多数

- 旧システムでは何でも受け入れた
- サーバのクラウド利用はある程度は浸透
 - そこそこお金のある部局がやってくる



北海道大学

運用をめぐる緊張関係

OpenStackでクリティカルなサービスを動かすのは難しい

- 学内のかなり重要なサーバが稼働している
- 今回のクラウド導入コストの大半は、移行作業
 - スパコンは止められるが、クラウドは止められない

OpenStack Ironic(ベアメタル)が全然成熟していない

- 特にコンソール周り(ネットワーク、ストレージ)

OpenStackの全ての機能が使えない

- ブロックストレージはストレージH/W支援が必要



本日のアウトライン

システムのご紹介

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

次期システムへ向けて



北海道大学

OCTOPUSの良い点

ヘテロな計算機が共通のOCTOPUSポイントで使える

- 北大クラウドが当初目標にしていたH/W構成に近い
- システムソフトウェア研究としては、いろいろなハードウェアが使えるのは大変有難い

Condaが使えるのは案外便利

- 利用者ごとに機械学習環境を切り替えるには、Docker/Singularityでなければいけないと思っていた



北海道大学

OCTOPUSの良くない点

ジョブがどの計算機も混んでいる印象

- 非常に好評であることの裏返しだが…

3つの利用者向けポータル

- Web利用申請システム
- 大型計算機システムポータル
- 利用者管理Webシステム
 - 特にホームページからたどった場合の入り口が分かりにくい…



北海道大学

本日のアウトライン

システムのご紹介

ハードウェア調達から継続的進化へ

利用者から見たOCTOPUS

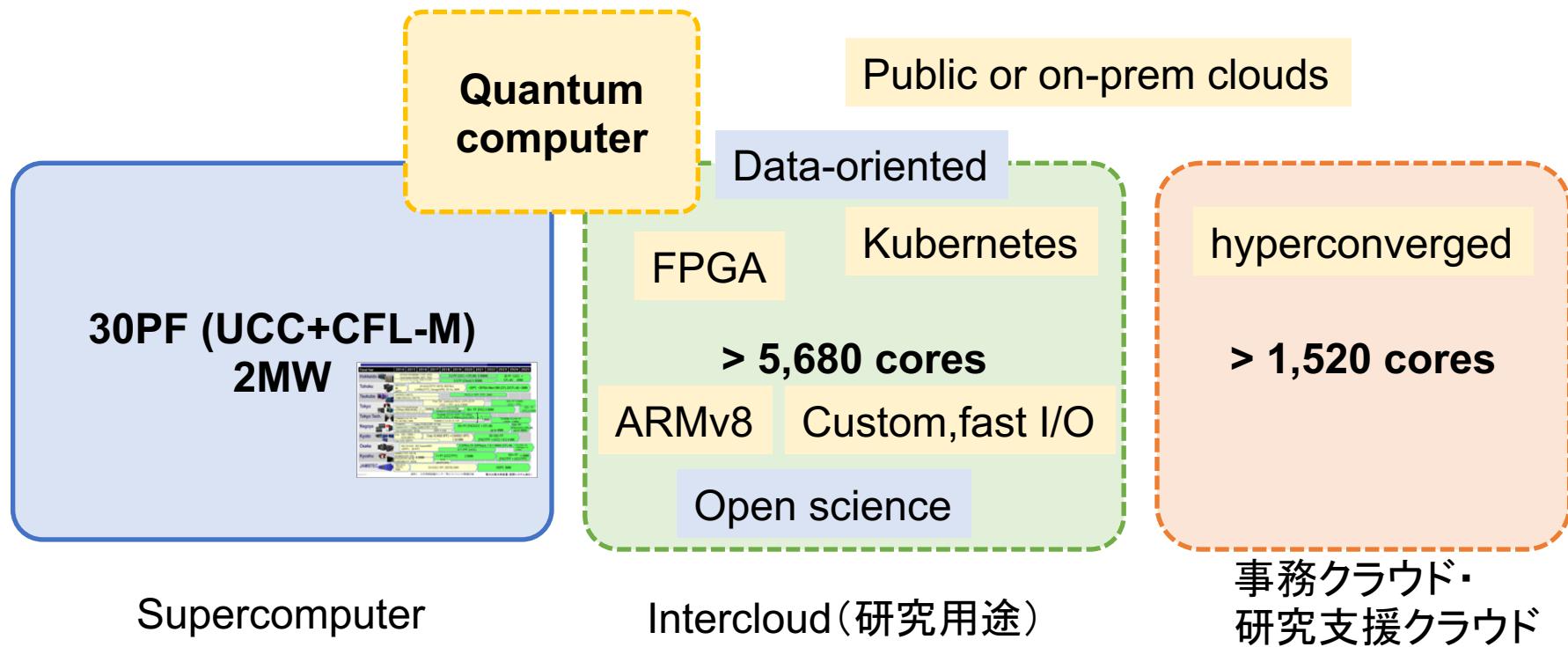
次期システムへ向けて



北海道大学

次期システムの予想(2023年12月)

次期システムでスパコン・クラウドは統合されると考えていた
クラウドはクラウドとしての道を追求
学内サーバの受け皿も必要



今後のクラウド設計(< 3-5年)

ベンダ提供のベアメタルソリューション

- OpenStackの未来は？

Kubernetes

- 単一クラスタ、ネームスペース切り替えによる運用は厳しい
パブリッククラウド(マシン以外または短時間利用のみ)

(+ 量子コンピュータ)

(+ GPU以外のアクセラレータ)



北海道大学

調達を取り巻く状況

スパコンの作り方はコモディティ化

- x86(またはGPU) + OPAまたはIB + Lustre
- ハードウェアに違いはない(導入時期のみ)
 - 差別化要因は何か？

一方、クラウドはまだまだ高い

- パッケージ化がほとんどされていない
 - 特にソフトウェアライセンス、作業費用



北海道大学

調達したハードウェアのソフト力の強化へ

ソフト力の強化

– ソフトパワー(soft power)

- 国際政治用語

- 軍事力・経済力などの強制力によらず、その国の文化、価値観、魅力等によって支持を得ること(Wikipedia「ソフトパワー」要約)

- 北風と太陽

– 新システムの設計や実装はよい

- アプリケーション支援の強化
- 利用負担金のモデル
- 運用や利用者支援などのソフト面の強化



パネルに対する意見

計算機センターが廃業するかどうかは分からぬ

- 我々にはコントロールできない力が突然働く(予測不能)

研究者個人としては、廃業することを想定して備えることが大事ではないか？

- これまでの自身の反省を強く込めて
- 大学を取り巻く状況を考えると、業務をしている場合ではない
- 業務は自身の研究・教育に関係するもののみに留めるべき

廃業しないための中途半端な努力は、組織を捻じ曲げて、かえって崩壊を早める恐れがある



引用：大阪大学 Cyber HPC Symposium 2019のポスター



北海道大学

パネルに対する意見(2)

極めて現実的な今後の大学の計算機センターのシナリオ

– シナリオ1: 純粹な研究活動を推進

- 研究・教育活動に大きく舵を切る(業務は捨てる)
- 選択肢1:オープンサイエンス化
- 選択肢2:クローズドサイエンスを強化
- 業務はパブリッククラウドに

– シナリオ2: ノウハウセンター化

- 業務方面に舵を切る
- 計算機リソースはパブリッククラウド,
構築・運用ノウハウを蓄積

– シナリオ3: 現状維持(do...while)

- どちらにも舵を切らない

