

京プリポストクラウドサービスの紹介

RIKEN R-CCS

Mar. 28, 2019

導入の背景

プリポスト処理のための計算リソースの不足

- 「京」ではプリポスト処理用の物理サーバ4台が導入されているが、8万ノードに達する計算ノード規模に比べると極めて小さく、以前よりプリポストのためのリソース拡充の要望があった。

ソフトウェア利用環境に関する課題

- 「京」のプリポスト処理ノードは、Intelアーキテクチャを採用しているため、計算ノード (SPARCアーキテクチャ) で動作しないようなソフトウェアが簡単に実行できることが期待されていることが多く、オープンソースソフトウェアや、商用ソフトウェアがその代表である。
- 商用ソフトウェアの動作環境に関しては、産業利用課題のユーザから環境整備の要望があった。
- ソフトウェアによっては、インストールにroot権限が必要であったり、共有ライブラリ等の関連ソフトウェアの依存関係の解決が困難であったり、一つのオペレーティングシステムの上に多種多様な要望を満たそうとすることで生じる、従来型の共有環境に起因する問題があった。

京のプリポスト処理に関する機能強化の一環として、「京」コンピュータ環境に、IaaSタイプのプライベートクラウドを導入し、2018年度10月に実験サービスとして開始しました。

想定される適用例

データ処理

- メッシュ生成
- 計算結果アーカイブのための圧縮 (gzip, bzip2, pigz)
- 他サイトへのファイル転送 (rsync)
- マルチコア・大容量メモリ（最大96コア, 最大メモリ量320GiB）向けのシングルノード処理
- SSDを利用した高スループットI/O処理
- GPGPUを利用したデータ処理

可視化

- リモート可視化

その他

- オープンソースソフトウェアの活用
- 最新Linuxディストリビューションの利用
- Windowsの利用（OSイメージとライセンスはユーザが用意することが前提）
- ISVソフトウェアの利用（商用ソフトに必要な費用はユーザが負担することが前提）
- CLI/REST APIによるVMの自動化
- バッチ処理に向かないタスク
- 割り当て資源消費後の、期末の追加計算を行うための小規模な計算リソース

基本的には、従来の京ブリポスト処理ノードでの利用形態の延長で考えて頂いて問題ありません。

京プリポストクラウドサービスの特徴

VM=仮想マシン

x86アーキテクチャ

x86アーキテクチャを採用することで、手間のかかるポータリング作業を行うことなく、潤沢なオープンソースソフトウェアのエコシステムをすぐに利用することが可能です。結果、研究における time-to-result を短縮することが可能です。

仮想化

ユーザ利用環境を向上させるためにOpenStackによるサーバ仮想化を採用しています。サーバ仮想化の利点は、一般的なサーバと同様に高い自由度をセキユアな状態でユーザに提供することができます。例えば、各ユーザはroot権限によるコマンド実行が可能です。

オペレーティングシステム

様々なタイプのオペレーティングシステム(CentOS, Ubuntu)が仮想マシン上で利用可能です。Linuxだけでなく、Windowsや他のオペレーティングシステムもVM上で起動可能です。（なお、その場合ライセンスとイメージはユーザで用意する必要があります。）

インターネット

VMからインターネットにアクセス可能です。インターネットから最新のオープンソースソフトウェアを容易にダウンロード、インストール、アップデートすることができます。

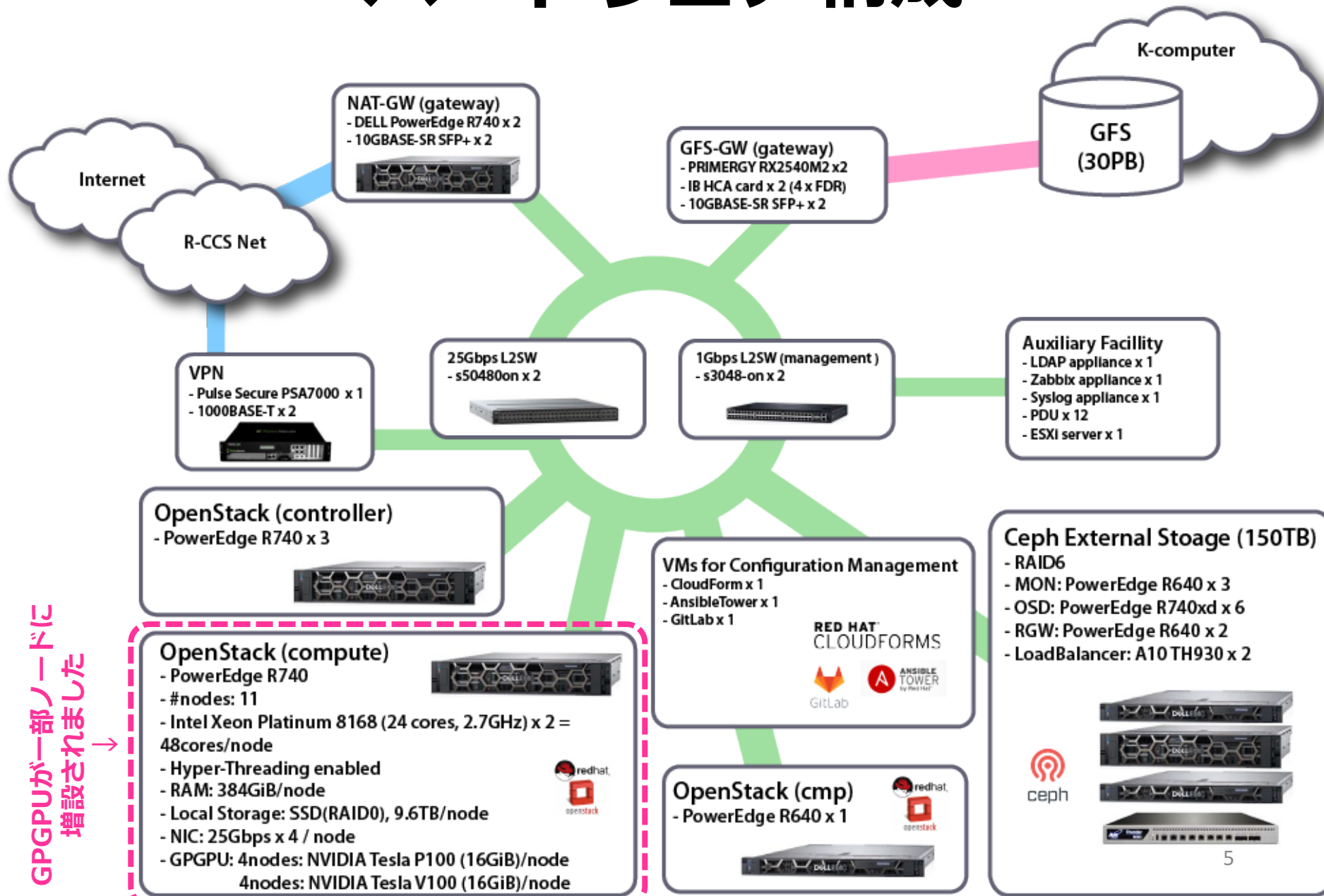
ストレージ

3種類以上のストレージが利用できます。ゲストOSのインストール領域および一時的なデータ処理の領域にSSDを用いることができます。また、VMのイメージのバックアップ先としてCephによる外部ストレージを用意しています。さらに、仮想マシンから京のGFSストレージをマウントして、利用可能です。

CLI/REST API

VMを制御するための、Pythonパッケージによるcommand line interface(CLI) が提供されています。このインターフェースを利用することで、クラウドシステム外部から利用するアプリケーション開発が可能です。

ハードウェア構成



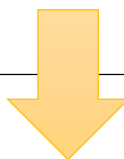
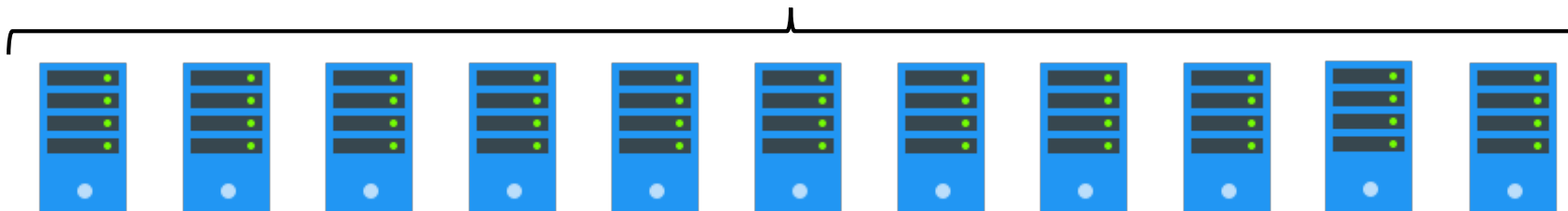
GPGPU増設の概要

- 11ノード(compute)の内, 各8ノードにGPGPUが1基搭載されています.
- 8ノードの構成内訳は以下になります.
 - 各4ノードにNVIDIA Tesla P100 (16GiB) x 1
 - 各4ノードにNVIDIA Tesla V100 (16GiB) x 1
- 1つのGPGPUは1つのインスタンスで専有する利用形態となります.
 - システム全体で最大8インスタンスが同時生成可能
 - すでに8インスタンス生成されている場合は, エラーになります.
 - 1つの物理GPGPUを複数インスタンスで共有すること (例: VDI用途) はサポートしません.
- GPGPU増設に伴う変更点
 - GPGPU専用のAvailability Zone(AZ)が提供されます.
 - AZがデフォルトの場合, GPGPU非搭載ノードでインスタンスが生成されます.
 - GPGPU専用のフレーバが提供されます.

Availability Zone(AZ)の変更

GPGPU増設前

nova ← Availability zone

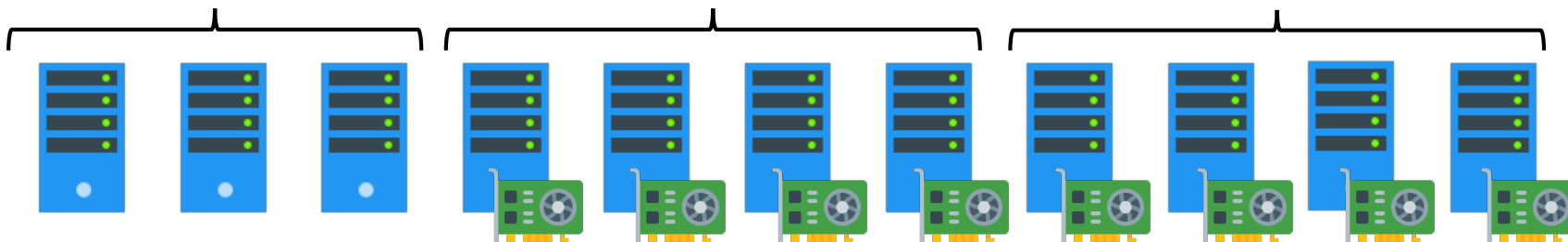


GPGPU増設後(2019/3/11以降)

nova

gpu-p (Tesla P100用)

gpu-v (Tesla V100用)



novaについては、従来通り
1vCPU~96vCPUまでの複数の
フレーバーが選択できます。

上記以外のAZとしてcmpが別途あるが、ユーザ向けに開放していないので省略します。

新旧プリポストシステム比較

	プリポスト処理ノード	京プリポストクラウド
CPU	Intel Xeon X7560 (Nehalem-EX) (8cores/2.26GHz/24MB) x 8 (/node)	Intel Xeon Platinum 8168 (Skylake) (24cores/2.7Ghz/33MB) x 2 (/node)
#nodes	2 (front nodes) + 2 (batch nodes)	11 (compute nodes)
Total #cores	128 cores (batch nodes)	528 cores (1056 vCPUs, Hyper-Threading enabled)
RAM	0.5TiB/node or 1TiB/node (The batch nodes have memory devices in different sizes.)	384GiB/node
Storage	GFS(30PB)	SSD(9.6TB/node)+Ceph(150TB)+GFS(30PB)
OS	RHEL 6.5	HostOS: RHEL 7.4 GuestOS: CentOS7, Ubuntu 18.04LTS (VMごとにゲストOSをリストから選択できる)
GPGPU	なし	NVIDIA Tesla P100 x 4基/system NVIDIA Tesla V100 x 4基/system ※一部の計算ノードに増設
備考	バッチ処理システム(SLURM)がインストールされている。ユーザはバッチジョブマネージャーを通して、ジョブをサブミットできる。	仮想マシンを管理するための各種インターフェース (Web/CLI/REST API) を提供。インターフェースを通して、ユーザは必要に応じて仮想マシンを確保できる。

クラウドコンピューティング

1. サーバ仮想化

サーバ仮想化は、物理リソースを仮想化された独立した環境(VM)に分割、再構成し、それぞれのVMを異なるユーザに割り当てることができます。それぞれのVMには異なるオペレーティングシステムをインストールすることができます。

仮想化されるリソース

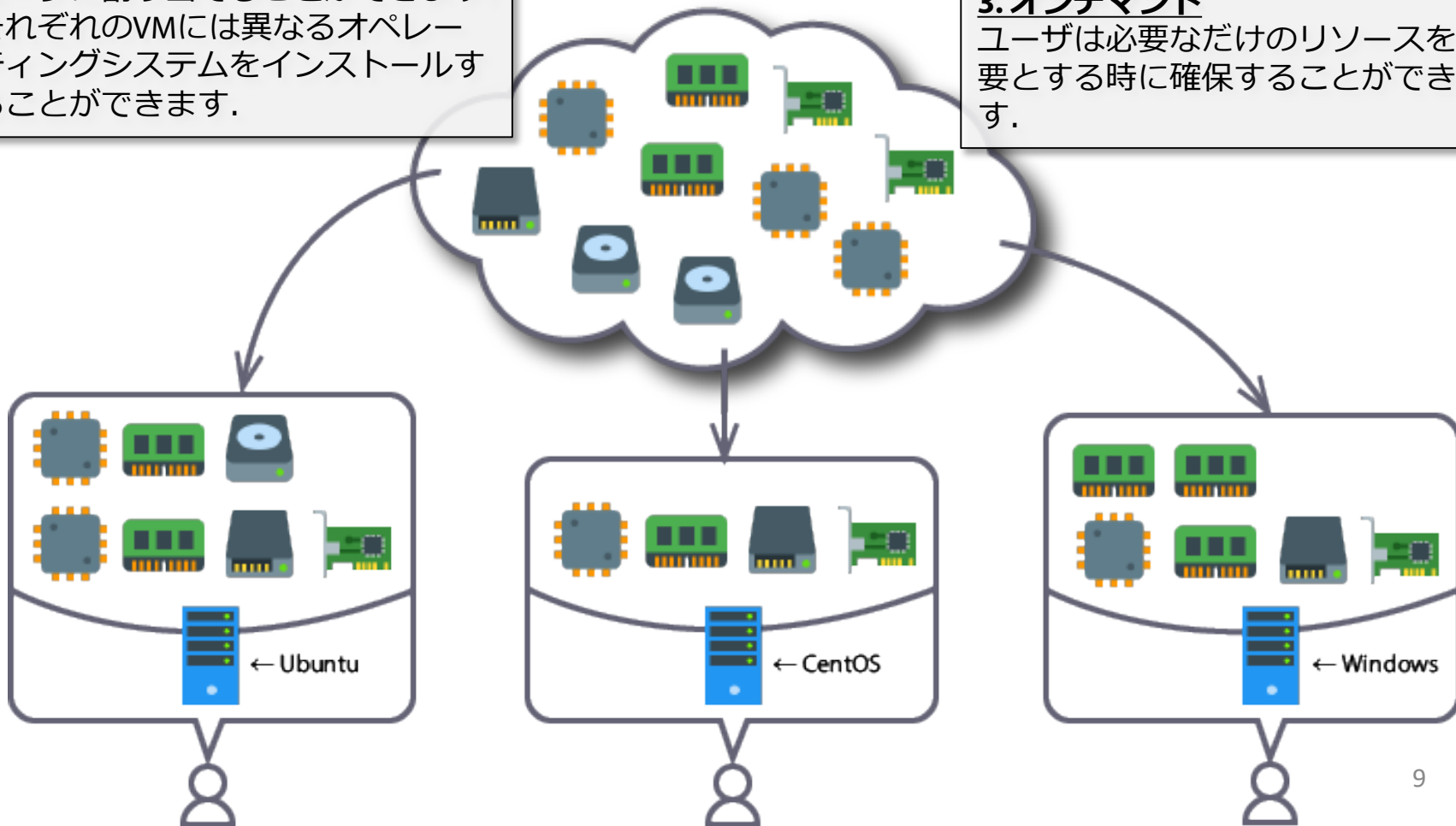
- CPU/GPGPU
- メモリ
- ストレージ
- ネットワーク

2. マルチテナンシー

OpenStackは完全に独立したVMを提供することができます。

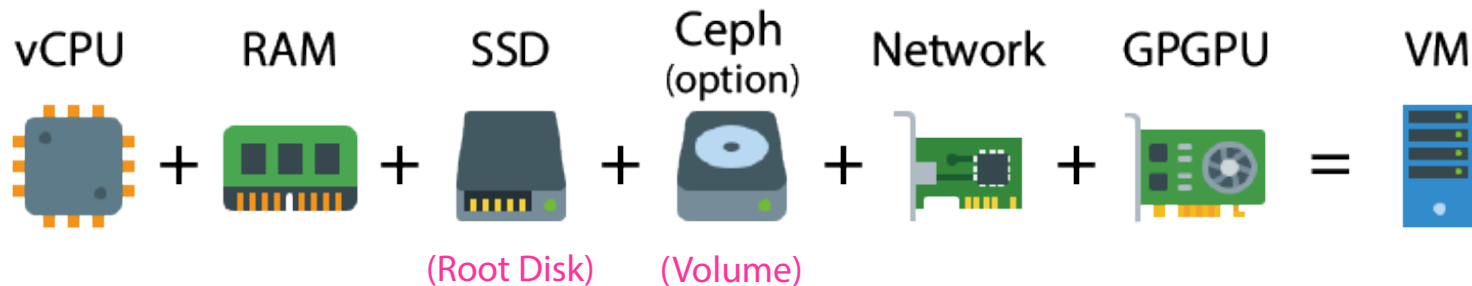
3. オンデマンド

ユーザは必要なだけのリソースを必要とする時に確保することができます。



ソフトウェア定義リソース

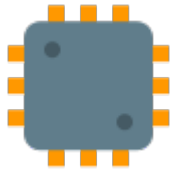
- サーバ仮想化では、物理リソースを複数のVMで共有するために抽象化（仮想化）します。
- 仮想化されたリソース (vCPU, RAM, SSD, Ceph, Network) のことを、ここではまとめてソフトウェア定義リソースと呼びます。
- VMにログインすると、ソフトウェア定義リソースは一般的な（専有の）物理リソースと同じように見えます。



- 京プリポストクラウド(OpenStack)では、“flavor”と呼ばれるソフトウェア定義リソースのセットメニューのようなものを提供します。
- ユーザは、VM生成時に、必要なリソースサイズが定義された“Flavor”をリストから選択して、VMを起動させることができます。（ただし、クォータを超えることはできません。）
- “flavor”では、vCPU（仮想CPU）、RAM（メモリサイズ）、SSD（ローカルディスク）のリソースサイズ、GPGPUの搭載の有無が選択できます。
- それ以外に、Ceph（分散ストレージ）による外部ストレージを用意しており、その一部領域を払い出すことで追加ボリュームを作成できます。また、その領域にVMイメージのバックアップを保存することができます。
- また、プロジェクトごとに内部ネットワークを提供します。（現在、ネットワークに関するカスタマイズは利用できません。）

フレーバー (Flavor)

- 現在, 以下のflavorに基づくリソースが利用できます.



VM (instance) Type									
vCPUs		RAM [GiB]							
		4	8	16	32	64	128	256	320
	1	A1							
	2		A2						
	6			A3					
	12				A4				
	24				C1	A5	B1	B2	B4
	48				C2	C4	A6	B3	B5
	96				C3	C5	C6	A7	A8

仮想マシン (VM) タイプ
 A1-8: 標準タイプ
 B1-5: メモリ消費タイプ
 C1-6: 計算コア消費タイプ



Root (ephemeral) Disk Size (SSD)	
tiny	16GiB
small	128GiB
medium	512GiB
large	2TiB
huge	8TiB

vCPU+RAM size Root Disk size

Example: A5.medium

24vCPUs 64GiB 512GiB (SSD)

仮想マシンの寿命

- 京プリポストクラウドは商用クラウドのように無限のリソースがあるわけでないので、より多くのユーザに利用する機会を与えるために、Flavorのタイプに応じて古いVMを一定の時間で強制終了するポリシーを導入しています。
- VM削除の3日前に、VMを作成したユーザの登録メールアドレス宛に通知メールが送られます。
- このポリシーは、より多くのリソースを消費するVMに対して、短い寿命（最短1週間）が設定されています。一方で、消費リソースが小さいVMについては、長い期間利用できるように設定されています。

VM Maximum Duration (tiny, small, medium)									
		RAM [GiB]							
		4	8	16	32	64	128	256	320
vCP Us	1	inf							
	2		inf						
	6			inf					
	12				inf				
	24				4w	4w	2w	2w	1w
	48				4w	2w	2w	1w	1w
	96				2w	2w	1w	1w	1w

VM Maximum Duration (large, huge)									
		RAM [GiB]							
		4	8	16	32	64	128	256	320
vCP Us	1	1w							
	2		1w						
	6			1w					
	12				1w				
	24				1w	1w	1w	1w	1w
	48				1w	1w	1w	1w	1w
	96				1w	1w	1w	1w	1w

TIPS

- VMのスナップショットを保存するためのストレージ(Ceph)が用意されています。
- スナップショットのファイルサイズはVMのroot diskのサイズに依存します。
- すべてのシミュレーションデータを保存できる容量はCephにありません。
- ストレージリソースを節約するために、tiny, small, mediumのroot diskを利用することをお勧めします。
- また、VMが消失することを防ぐために、VMのバックアップ（スナップショット）を定期的にとられることをお勧めします。

仮想マシンの寿命

1w: 1週間

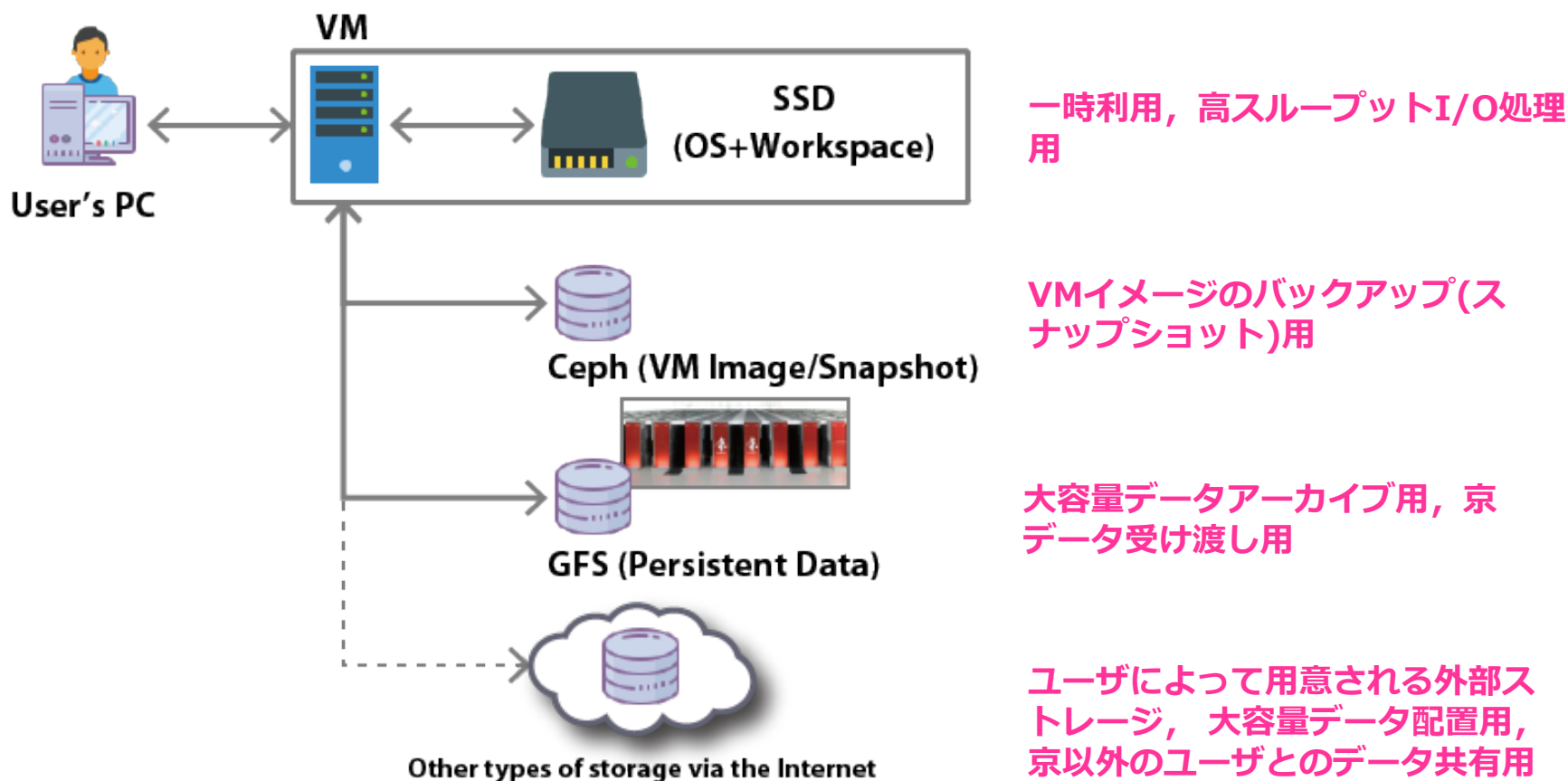
2w: 2週間

4w: 4週間

inf: アカウント失効日もしくは年度末まで¹²

ストレージ（概要）

- 京プリポストクラウドサービスでは用途別に複数のタイプのストレージを提供しています。
- 現在，以下の図で示される4種類のストレージの利用が想定されます。



ストレージ（補足）

- **SSD (ephemeral, root disk)**

- 計算ノードには、RAID0のディスク構成のローカルディスクがインストールされていて、OpenStackによって仮想ストレージリソース(root disk)としてVMに割り当てられます。
- 仮想マシンからはブロックデバイスとして見えます。
- この領域はOSのインストール、ユーザデータの一時的な保存の使用を想定しています。
- なお、OpenStackの用語として、“root disk”と“volume”は異なるので混同しないように注意してください。
- **VMを終了するとき、root diskも削除されます。**

- **Ceph (volume)**

- VMイメージのバックアップ（スナップショット）を保存するための外部ストレージです。
- このストレージは、分散ストレージ用クラスタノード間でデータを3重に冗長配置して、可用性を高めるフォールトトレラントの機能を持っています。
- 大規模なディスク容量が必要な場合、SSDの代わりにゲストOSのインストール領域（ブロックデバイス）として利用することができます。
- シミュレーションによって生じる大量の入出力データのすべてを保存するための容量はありませんのでご注意ください。

- **京グローバルファイルストレージ (GFS)**

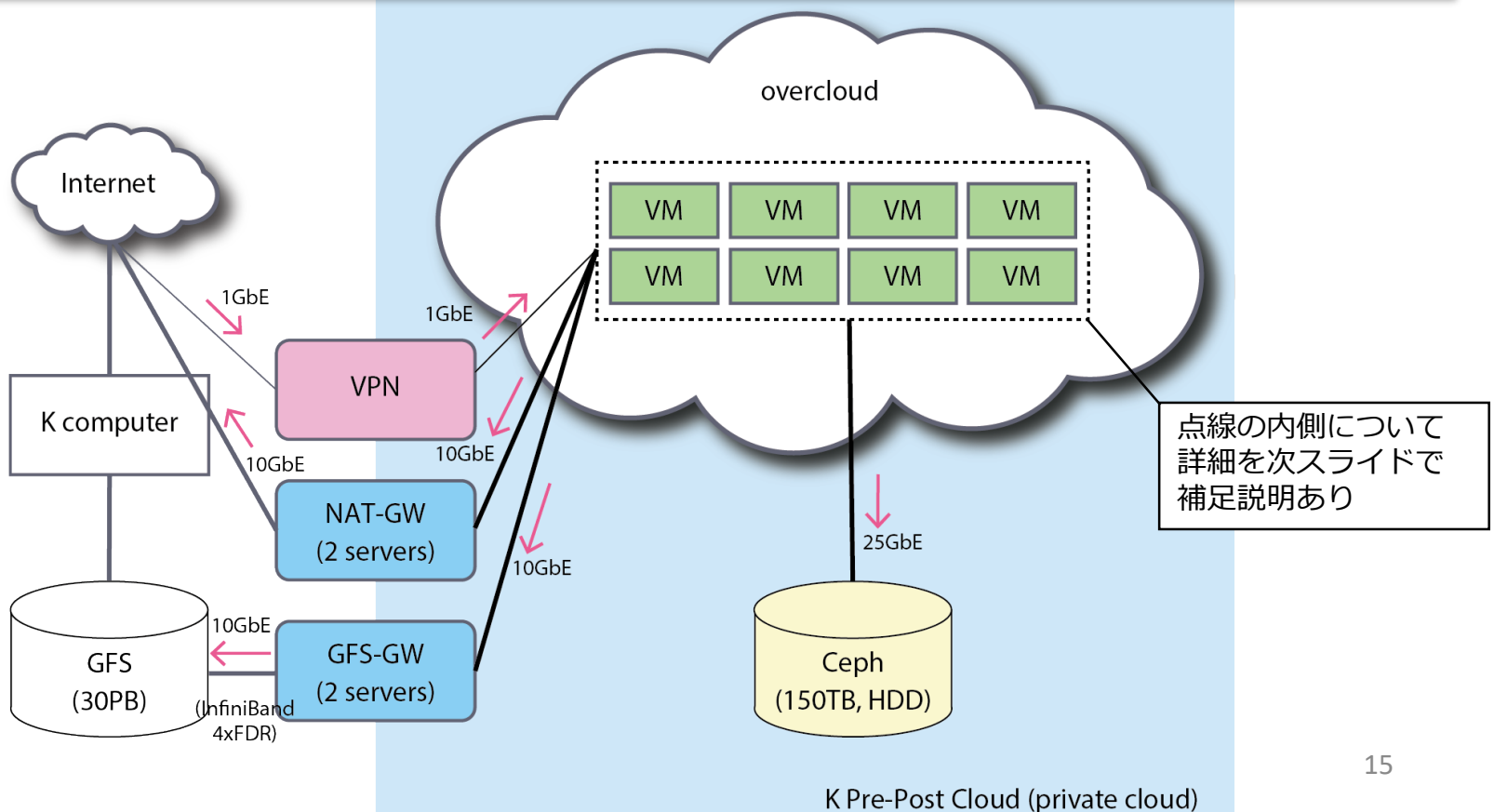
- VMからSFTPもしくはSSHFSを用いて、「京」のグローバルファイルストレージの領域をマウントし、アクセスすることができます。

- **インターネット上で利用できる各種ストレージ**

- VMからインターネット経由でアクセスできる外部ストレージ（例えば、商用クラウドストレージ）がユーザによって用意できる場合、利用することができます。

ネットワーク（概要）

- VMとCephストレージは25GbEによって構成される内部ネットワークによって接続されます。
- VMには、プライベートIPアドレス(10.9.0.0/16)がVMに割り当てられます。
- VPN接続を通して、ユーザは内部ネットワークにアクセスすることができます。
- VMは、10GbEで接続されたNAT-GW（ゲートウェイサーバ）を通して、インターネットにアクセスすることができます。
- また、VMは、GFS-GW（ゲートウェイサーバ）を通して京のグローバルファイルストレージにアクセスすることができます。



ネットワーク (詳細)

- 標準構成では、各プロジェクトに、仮想スイッチ1つと内部ネットワークが払い出されます。
- 外部ネットワーク(external)は、すべてのプロジェクトで共有されます。
- 内部ネットワーク(*-internal)は、当該プロジェクトで生成したすべてのVMで共有されます。

Internet

local PC

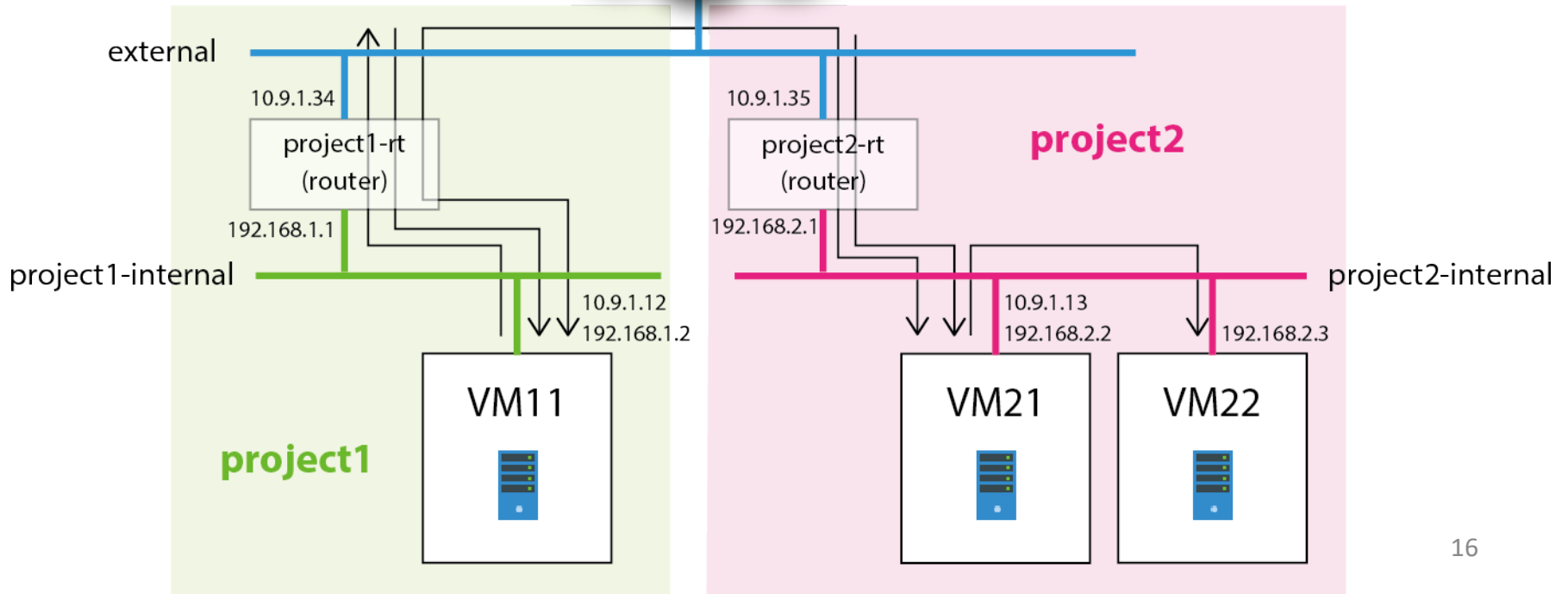
VPN connection

TIPS

- 京プリポストクラウドは、“security group”と呼ばれる簡易的なファイアウォール（パケットフィルタ）機能が提供されます。
- ユーザは、“security group”を用いることで、VMの TCP/UDP/ICMPプロトコルの ingress/egressに対するアクセス制御を行うことができます。

private IP: 192.168.0.0/16

public IP (external): 10.9.0.0/16



ダッシュボード (Horizon)

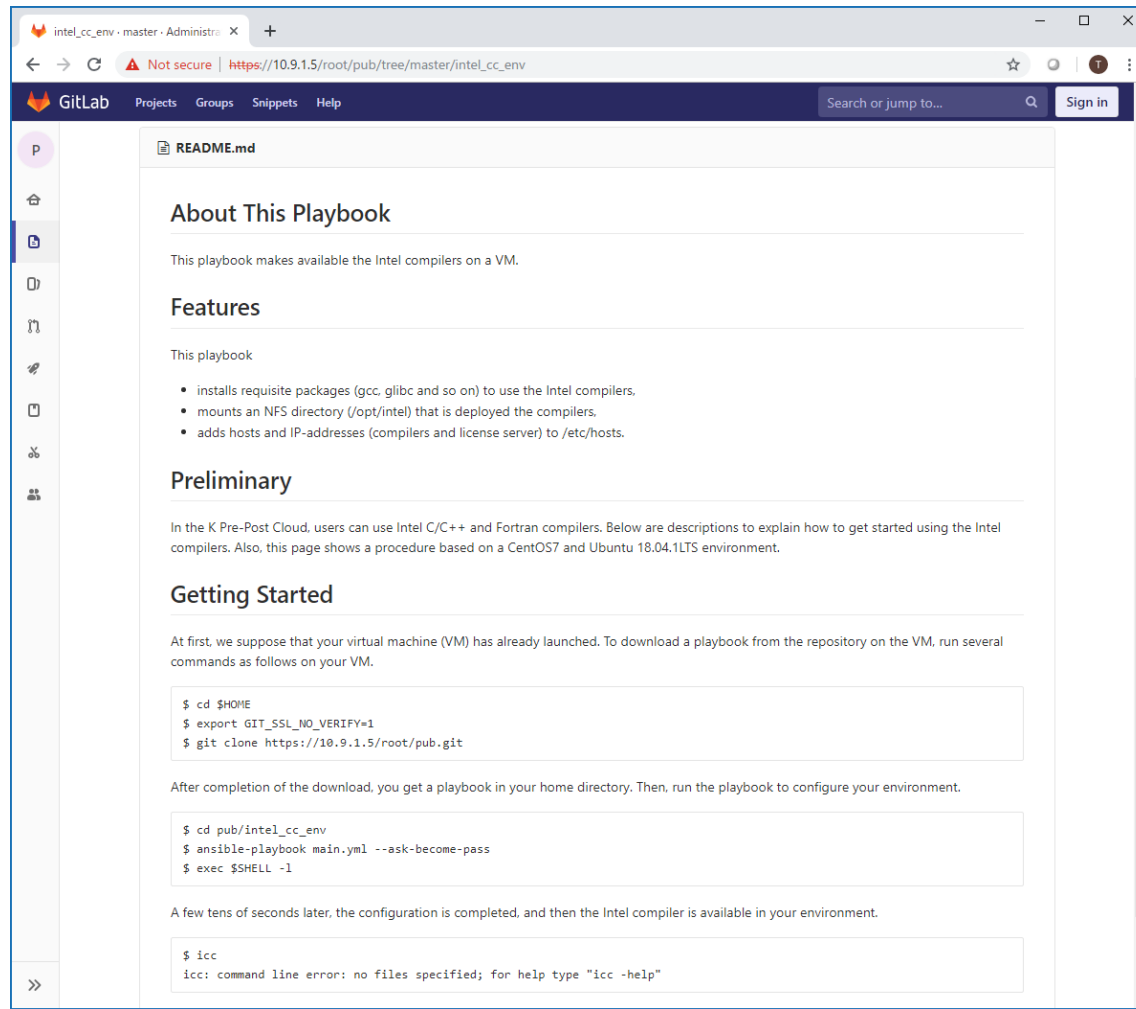
- ここまで説明したソフトウェア定義リソースに対する操作は、Horizonと呼ばれるWebインターフェース（ダッシュボード）を利用することで達成できます。
- VM作成後、ユーザはターミナルを用いて、VMにSSH鍵認証でログインして、あとは一般的なサーバと同様に利用を開始します。
- ユーザには、VMのroot権限が与えられるので、基本的にテクニカルな制限はありません。

The screenshot displays the OpenStack Horizon web interface. The main page shows the 'Instances' section with a table of existing VMs. A 'Launch Instance' dialog is open, showing the 'Flavor' selection screen. The dialog includes a sidebar with configuration options like Source, Networks, and Key Pair. The main area shows a table of available flavors with columns for Name, VCPUS, RAM, Total Disk, and Public access.

Name	VCPUS	RAM	Total Disk	Public
A1.tiny	1	4 GB	16 GB	Yes
A2.small	2	8 GB	64 GB	Yes
A3.medium	6	16 GB	128 GB	Yes
A4.medium	12	32 GB	256 GB	Yes
A5.large	24	64 GB	1 TB	Yes
A6.large	48	128 GB	2 TB	Yes
A7.huge	96	256 GB	4 TB	Yes

Intelコンパイラの提供

- Intel compiler (Linux版parallel_studio_xe_2019)が利用可能です.
- 詳細は, VPN接続した上で, 以下サイト(GitLab)をご参照ください.
- https://10.9.1.5/root/pub/tree/master/intel_cc_env



利用申請

- 京プリポストクラウドサービスは無料で利用できます。
- 利用するには申請時点で、京のユーザである必要があります。
 - VPNの証明書、アカウント名は、メールで通知されます。
 - パスワードは、郵送等で通知されます。
 - 申請に基づいて、京のグループID（課題）単位に、京プリポストクラウドのプロジェクト（内部ネットワーク）が作成されます。
 - 本サービスのユーザIDとして、京のユーザIDの末尾に'c'を付加したものを、別途発行します。
 - VPNアカウント、証明書も同時に発行します。

• 申請方法

- 下記Webサイトの“Application”から利用申請ができます。
 - <https://www.r-ccs.riken.jp/ungi/prpstcloud/>
- Webサイトでは、利用のための役立つ情報を提供していますのでご参照ください。

• 問い合わせ窓口

- 利用に関する一般的な質問については、以下のアドレス（京サポートデスク）にメールでお問い合わせください。
 - r-ccs-k-desk@riken.jp

免責事項

- 京プリポストクラウドは、京コンピュータ環境におけるプリポスト処理機能強化のための、実験サービス用プラットフォームです。
- スパコン環境に導入されるプリポストサーバの技術的な知見や運用に関するノウハウを得るための、実験サービスを目的としている。
- サービスの品質や手続きが「京」などの正式にサービスしているものに比べると見劣りする可能性はありますがご了承ください。
- 利用規約に関しては以下URL先に示されるWebサイトをご参照ください。
 - <https://www.r-ccs.riken.jp/ungi/prpstcloud/>
- 利用規約にはいろいろ厳しいことが書かれているかもしれませんが、要望にお応えできることは対応していきたいと考えています。