Vol.2018-HPC-165 No.8
2018/7/30

クラウドコールドストレージに対する大規模実験データ格納のケーススタディ

吉田 浩[†1]　合田 憲人[†1]　上田 郁夫[†2]　原 隆宣[†2]
小杉 城治[†3]　森田 英輔[†3]　中村 光志[†3]

概要：現在，主要なパブリッククラウドでは，アクセス頻度が比較的低いデータの保管を想定したコールドストレージサービスが提供されており，大量の科学研究データの長期保管に活用できないかという観点から考慮に値するものと言える．筆者らは，これまで，複数の商用パブリッククラウドが提供するコールドストレージサービスに対して，実際の科学研究データの格納を含む最大 1PiB の大量データ格納の試行と評価を実施してきた．本報告では，その最新のベンチマーク結果に加えて，具体的なケーススタディとして，研究で使われている観測データやシミュレーションデータをコールドストレージサービスに格納し解析アプリケーションやアーカイブ管理アプリケーションからアクセスする試行において得られた性能・運用性・コストの観点からの分析・評価結果を述べる．

キーワード：クラウド，コールドストレージ

Case Study Analyses of Storing Large Amount of Scientific Experimental Data in Cloud Cold Storage Services

HIROSHI YOSHIDA[†1]　KENTO AIDA[†1]
IKUO UEDA[†2]　TAKANORI HARA[†2]
GEORGE KOSUGI[†3]　EISUKE MORITA[†3]　KOHJI NAKAMURA[†3]

Abstract: Currently major cloud providers usually provide cold storage services which target to store data with relatively low access frequency for a long period as a part of their public IaaS offerings. The adoption of cold storage services should be considered in order to reduce the total cost of ownership and the labor of storage management of maintaining large amount of data for a long period. The authors have been conducting experiments on commercially available public cold storage services by storing large amount of data up to 1PiB as well as storing and accessing the actual scientific research data. In terms of performance, cost, and manageability, this report describes the case study analyses of the trials of storing actual scientific observation data and analysis data in the cold storage services and accessing them by the analysis applications and archive management applications, as well as the latest results of our benchmark tests.

Keywords: Cloud, Cold storage.

1. はじめに

現在，主要なパブリッククラウドの IaaS においては，オブジェクトストレージサービス（以下，特に断らない限り「オブジェクトストレージ」と呼ぶ）が提供されている．IaaS で提供される仮想マシン（以下 VM）の内蔵ディスクあるいは外付けディスクとして見えるブロックストレージサービスよりも容量単価（一般にクラウドでは 1GiB の容量を 1 か月保管する料金）が低く，さらにクラウド内外から REST API（http）によってアクセスすることが可能であるため，VM から独立したデータの保管や，データの共有・受渡しなどに利用される．オブジェクトストレージの利用が拡大するにつれ，標準的な性能や可用性などのサービスレベルを提供するサービス（「標準」,「ホット」など各社各様のサービス名で提供されているが，以下は「標準オブジェクトストレージ」に統一する）に加えて，アクセス頻度が比較的低いデータの長期保管を主な用途として，容量単価を相対的に低く設定したコールドストレージサービス（以下, 特に断らない限り「コールドストレージ」と呼ぶ）が提供されることが一般的となった[1].

大学・研究機関におけるコールドストレージの用途の一つとして，大量の科学研究データの長期保管が考えられ，データの保管に関わる TCO の低減や，ストレージシステムの運用管理の負担軽減といった効果が期待される．しかし一方で，コールドストレージには，性能や運用性の面で標準オブジェクトストレージとは異なる特有の仕様があり，

†1 情報・システム研究機構 国立情報学研究所
National Institute of Informatics
†2 高エネルギー加速器研究機構 素粒子原子核研究所
Institute of Particle and Nuclear Studies
†3 自然科学研究機構 国立天文台
National Astronomical Observatory of Japan

a) 本論文に記載されている社名，商品またはサービスの名称等は，各社の商標または登録商標です．

コスト面では優位であっても，現実の研究データ保管の運用に耐えられるかという懸念が生じることは否定できない．

この状況をふまえて，筆者らは，研究データをコールドストレージに保管するかどうかの判断，あるいはクラウドを含めた研究データ保管のストレージアーキテクチャ設計の一助となる実際的な情報を得る目的で，複数の商用パブリッククラウドで提供されるコールドストレージに関する実証実験を継続して進めている．実験には，基礎的なベンチマークテスト，最大 1PiB の大量データの格納，および実際の研究データの格納とアクセスなどが含まれている．前回の報告[2]では，ベンチマークテストやクラウドへのデータアップロード性能に関する実験結果を中心に述べたが，実際に研究者のデータ利用シーンに近い形でデータアクセスを行った場合の性能やコストがどうなるかという点については，十分に明らかにはなっていなかった．

本報告では，コールドストレージによる研究データ保管のケーススタディとして，高エネルギー物理学や天文学の実際の研究データの格納と，実際のデータ利用にできるだけ近い形の研究アプリケーションによるアクセスを試行し，性能・コスト・運用性の観点から検討を行った結果を述べる．合わせて，クラウドサービスの進化を考慮して，研究データアップロードや 1PiB のデータ格納に関する最近の実験結果を報告する．今回の実験で明らかになったことは，以下のとおりである．

- クラウドの進化は非常に速く，クラウドに対するデータのアップロードに関しては，1 年間で 2〜10 倍程度の性能向上が実現されている．
- 大量の研究データの格納・アクセスにおいて，総コストに影響するのは，主にコールドストレージのデータ保管コストとクラウド外へのデータ転送コストであり，コールドストレージのデータ読出しコストは，あまり影響しない．
- データアクセスに長時間（時間オーダの待ち時間）を要するタイプのコールドストレージは，コスト面からは有力な選択肢であり，適切なストレージ階層化を行うことによって，システム全体のコストパフォーマンスを最適化できる可能性がある．

2. 実験対象としたコールドストレージ

本実験は，次の 4 種類の商用パブリッククラウドのコールドストレージを対象とした．

- Amazon Web Services (AWS) Glacier [3]
  および S3 Infrequent Access [4]（以下 S3 IA）
- Google Cloud Platform (GCP) Coldline Storage [5]
  （以下 Coldline）
- Microsoft Azure block BLOB Cool（以下 Cool）[6]
- Oracle Cloud Archive Storage Service [7]

表 1　実験の対象としたコールドストレージ
（2018 年 6 月 28 日現在）

	サービス名		価格[注1] ($/GiB･月)
標準オブジェクトストレージ	AWS S3		0.0250 [注2]
	GCP Regional Storage		0.0230
	Azure block BLOB Hot		0.0200 [注2]
	Oracle Cloud Storage		0.0255
コールドストレージ	標準オブジェクトストレージと別サービス	AWS Glacier [注3] • 復元処理要（料金により変化）[注4] • 読出し料割増 • 最低保持期間 90 日 • 少量データの切上げ格納 • 可用性未定義	0.0050
	標準オブジェクトストレージサービスのオプション	AWS S3 Infrequent Access (IA) • 読出し料割増 • 最低保持期間 30 日 • 少量データの切上げ格納 • S3 より低い可用性	0.0190
		GCP Coldline • 読出し料割増 • 最低保持期間 90 日	0.0100
		Azure block BLOB Cool • 読出し料／書出し料（従量）割増 • 最低保持期間 30 日	0.0150
	両者の中間	Oracle Archive Storage • 復元処理要(〜4 時間) • 読出し料割増 • 最低保持期間 90 日 • 長大オブジェクトの扱いに差異	0.0026

注 1: 図 1 に記載した地域（リージョン）における価格
注 2: 50TiB を超えると容量単価が逓減される．
注 3: ライフサイクル管理機能を利用することによって，
S3 あるいは S3 IA とほぼ同様の操作が可能
注 4: 復元処理時間（標準，バルク，高速）に応じて課金が異なる．

オブジェクトストレージ全体のサービス体系の観点からこれらのコールドストレージを分類すると，以下のように大別することができる．

i) 標準オブジェクトストレージとは別サービスとして提供されるもの．操作や API は別体系となる．
ii) 標準のオブジェクトストレージのオプションとして提供されるもの．操作や API は同一である．多くの場合，オブジェクトあるいはコンテナ（AWS，GCP ではバケット，Azure，Oracle，OpenStack ではコンテナと呼ぶが，AWS や GCP に関する記述を除いて，本報告ではコンテナに統一する）作成時にコールドストレージとすることを指定する．
iii) 両者の中間的なもの．

この分類に沿って，実験で利用したコールドストレージの特徴と保管料金（容量単価）を表 1 にまとめる．コールドストレージの保管料金は，標準オブジェクトストレージサービスに対して数分の一から一桁低い一方で，トレードオフとして，以下のような特徴がある．

表 2　実験で使用した研究データとアプリケーション

データ	データ量	アプリケーション	提供元	備考
高エネルギー物理学				
Belle II 実験の物理シミュレーションデータ	633GiB 1000 ファイル（ほぼ同一サイズ, 600〜700MiB）	解析支援ソフトウェア環境 BASF2 (Belle II Analysis Frame-work)経由のファイル読出し	高エネルギー加速器研究機構	
天文学				
ALMA 電波望遠鏡の観測／解析データ	58.5TiB 138 万ファイル（1MiB 以下〜100 GiB 以上に分布）	アーカイブシステム NGAS (Next Generation Archive System)	自然科学研究機構 国立天文台	他に Oracle DB 2.7TiB
野辺山望遠鏡観測データ	3.3TiB 29,075 ファイル（700B〜7.2GiB に分布）	アーカイブ用オブジェクトストレージシステム AOS (Adria Object Storage)		

- 標準オブジェクトストレージと比較して，アクセス性能が異なる，あるいは可用性が低いといった非機能要件上の差異があるものがある．
- 格納されたデータがオフラインとなっているかのように見えるものがある．すなわち，オブジェクトに即時にアクセスすることができず，アクセス前の復元処理（API 等で復元リクエストを発行し 1 時間から数時間待ってからアクセスを開始する）が必要となる．
- データの書込み・検索・復元・読出しなどのアクセス処理の料金が，標準オブジェクトストレージより高価に設定されていることがある．
- 少量データをある程度大きい単位（たとえば 128 バイト）に切り上げて課金することがある．
- データの最低保持期間が規定されていることがある．期間満了前にデータを削除した場合，最低保持期間分の保管料金が課金される．

これらの特徴が，実際の研究データの格納においてどのような影響を示すのかを調べるのが，本実験の重要な目的である．具体的には，復元処理時間がデータ読出し処理にどのように影響するのか，データアクセス料金の割増しがある場合，どのくらいのアクセス頻度であれば標準オブジェクトストレージより料金面の優位性が得られるのかといった問題を検討する．

3. 実験方法

3.1 実験内容

高エネルギー物理学分野および天文学分野において実際に研究に使われているデータをコールドストレージに格納し，さらに当該分野で使用しているアプリケーションまたはそのデータ入出力部を抽出したものを使ったアクセスを試行して，性能，コストパフォーマンス，運用性などを評価した．対象としたデータとアプリケーションの概要を表 2 に示す．

表 3　コールドストレージの一括処理用 CLI

ストレージ	CLI 名
AWS S3, S3 IA	AWS CLI，S3cmd
GCP Regional, Coldline	gsutil
Azure Hot, Cool	Azure CLI
Oracle Archive Storage	ftmcli，swift CLI（Swift 互換のため）

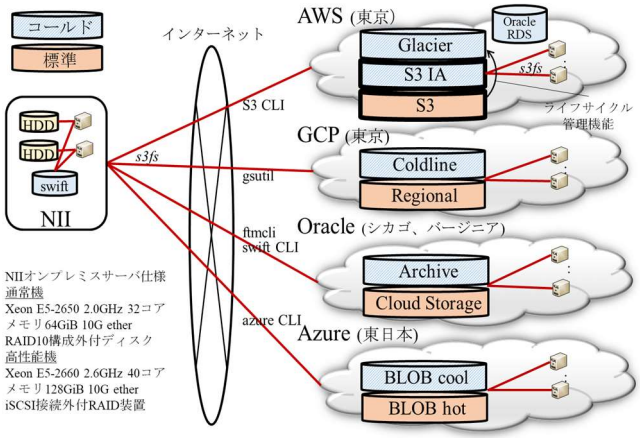

図 1　実験環境の全体像

3.2 実験の実施方針

コールドストレージにアクセスするためのインターフェースとしては，一般の利用者が最初に採用する方法に従うという方針から，主にクラウドサービスプロバイダが提供する GUI，CLI を使用した．特に，複数のファイルやオブジェクトを一括してアップロード, ダウンロード, コピー, 削除する場合の利便性が高い CLI を多用した．表 3 に，実験で使用した CLI を示す．CLI では，多くの場合，内部で並列処理を行う場合のスレッド数などを変更してチューニングを行うことが可能であるが，ここでは，一旦はデフォルト値を使用して実験を行い，何らかの不具合が生じた場合に変更するという方針で実験を進めた．

3.3 実験の全体構成

4 種のパブリッククラウドからなる実験環境を構築した．日本国内にデータセンタ（リージョン）を持つクラウドでは，それを利用した．ストレージに対するアクセス元としては，各クラウドが提供する同一リージョンの計算環境内の VM と，国立情報学研究所（以下 NII）のオンプレミスの物理サーバを用意し，必要なアプリケーションやデータを配置した．環境の全体像を図 1 に示す．

本実験では，NII のサーバとクラウド間の通信はインターネット経由で行った．通信速度の目安として，NII のオンプレミス環境のサーバおよびクラウドの同一リージョンの

表 4　実験環境における ping 応答時間
（パケットサイズ：1024 バイト，単位：ミリ秒）

発行元 ＼ 発行先[注1]	NII	AWS	GCP	Oracle
NII	0.20	3.0 ～7.0 に分布	2.8	150.7[注2]
AWS EC2		0.24～0.91 に分布		

注 1: Azure のエンドポイントはセキュリティ上 ping を通さない設定となっていると見られ，測定対象からはずした．

注 2: Oracle は 2017 年 6 月測定．他は 2018 年 6 月測定

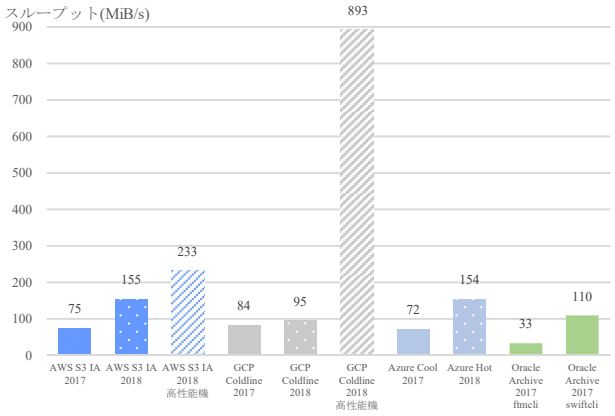

図 2　高エネルギー物理学データのアップロード性能
(オブジェクト数 1000，データ量 633GiB)

VM からクラウドストレージの API エンドポイントに ping を発行したときの応答時間を表 4 に示す．NII からの通信よりもクラウド内からの通信のほうが速く，さらに，Oracle は米国データセンタへの通信のため，時間が長くなる．

実際に研究アプリケーションを使ったアクセス試行は，もっぱら AWS 上で実施した．理由は以下のとおりである．

- 標準オブジェクトストレージおよびコールドストレージは，REST API でアクセスするのが本来の仕様である．しかし，今回の実験対象も含めて，既存の研究アプリケーションの多くは，POSIX 互換ファイルシステムの API でファイルにアクセスするように作られている．ここでは，アプリケーションを修正せずに実験を進められるように，コンテナをファイルシステムとして FUSE（Filesystem in USErspace）[8]でマウントし，個々のオブジェクトをファイルとしてアクセスする機能を提供するソフトウェアを利用した．AWS では OSS である s3fs [9]が著名であるが，他クラウドでは，このようなソフトウェアが得られないものがある．
- 測定項目の一つに課金額があるが，AWS では，詳細な課金レポートを取得する機能が提供されている．

4. 実験結果と考察

図 1 の環境の構築は 2017 年 1 月から開始し，実際の実験は 2017 年 2 月から 2018 年 6 月にかけて実施した．

4.1 ケーススタディ(1): 高エネルギー物理学データの格納とアクセス

(1) コールドストレージに対するデータアップロード

表 2 で述べた Belle II 実験の物理シミュレーションデータ 1000 ファイル，633GiB を，1 ファイルを 1 オブジェクトに対応させて，表 3 の CLI が提供する一括アップロード機能によって NII のオンプレミス環境からアップロードした．スループットの測定結果を図 2 に示す．なお，これまでの実験[2]から，コールドストレージが標準オブジェクトストレージのオプションであるタイプの AWS（S3 IA），GCP，Azure においては，両者の性能は同等であることがわかっているため，測定の一部は，早期削除課金がなく時間的な制約の少ない標準オブジェクトストレージで実施した．

2017 年と 2018 年に同じ測定を行ったクラウドに関しては，すべて 1 年間で性能の向上が見られた．これは，クラウドプロバイダのクラウドサービス基盤の性能改善および CLI の内部処理の改善が行われたためと考えられる．なお，NII オンプレミス環境のサーバとして，特に明記していないものは図 1 に記した通常機を利用したが，2018 年の再測定で性能向上が見られた一部のサービスについては，高性能機による測定も行い，結果としてさらなる著しい性能向上が認められたものがある．

(2) コールドストレージに対するデータアクセス

解析アプリケーションのデータ読出し部分を抽出したプログラムを AWS の VM（インスタンス）上および NII のオンプレミス環境のサーバ上で動作させ，(1)でアップロードした 1,000 オブジェクトを格納したバケットを s3fs でマウントしたものに対する読出し性能（経過時間）および課金額を測定した．このプログラムは，Belle II 実験で実際に使用している解析支援ソフトウェア環境 BASF2 のライブラリ経由でファイルを読み込むものである．

コールドストレージとして AWS S3 IA および Glacier，比較のための標準オブジェクトストレージとして AWS S3 について測定を行った．なお，Glacier へのデータ格納には，S3 のライフサイクル管理機能を利用した．これは，S3 のバケットに，一定期間経過後に Glacier に移行するポリシーを設定することによって，Glacier への書出しが自動的に行われるものである．Glacier に移行されたオブジェクトにアクセスするには，まずアクセス対象の全オブジェクトに対して復元リクエストを発行し，すべての完了を待ってからアプリケーションプログラムを実行する必要があるが，これは，クラウドプロバイダ提供の CLI（s3api コマンド）と OSS の CLI（S3cmd）を併用し，手作業によって実施した．

表 5 に得られた性能値（1,000 ファイル処理の経過時間）とアクセスおよび復元に対する課金額を示す．合わせて，それぞれのストレージサービスにおけるデータ保管料金も示した．

表 5　高エネルギー物理学データの読出し性能と課金額
(オブジェクト数：1,000，データ量：633GiB)

種別	コールド				標準
サービス	S3 IA		Glacier		S3
出力先	EC2	NII	EC2	EC2	EC2
復元オプション[注1]	-	-	標準	バルク	-
読出し時間(分)	501	717	476	489	493
復元時間(分)[注2]	-	-	203	338	-
リクエスト課金($)	0.083	0.296	0.029	0.029	0.031
読出しデータ課金($)	6.982	28.357 [注4]	-	-	-
復元課金($)[注3]	-	-	8.365	2.324	-
アクセス課金総額($)	7.065	28.653 [注4]	8.394	2.353	0.031
クラウド外転送課金($)	-	396.878 [注4]	-	-	-
月額データ保管料金($)	12.0		3.2		15.8

注 1: 復元後の保持期間は 1 日とした.
注 2: 1,000 オブジェクトの復元完了確認に約 6 分要するため
最大 6 分の誤差を含んでいる可能性がある.
注 3: 復元リクエスト課金+復元データ量課金+復元スペース課金
注 4: 他の場合の 4 倍のデータ読出し・転送に相当する課金額

アクセス性能に関しては,

- 同じクラウド内の VM からのアクセスに関しては, S3 IA，Glacier，S3 とも，ほぼ同等である.
- Glacier からの復元時間は, クラウドプロバイダが提示している公称値の下限に近い値となっている.

一方，課金額に関しては,

- コールドストレージの場合，リクエスト課金は，読出しデータ従量課金や復元課金に対して，相対的に無視できる程度に小さい.
- オンプレミスからのインターネット経由のアクセスでは，クラウド外データ転送課金が支配的である．なお，本測定では，この場合の読出しデータ課金と転送課金が他の場合の 4 倍のデータ量に相当する値となっている．s3fs のバッファリング処理などが関係しているとも推測されるが，十分な解析はできていない.

今回使用したデータとアプリケーション（クラウド内アクセス）に関して，表 5 のアクセス課金総額と月額データ保管料金を比較検討すると，以下のような考察ができる.

- S3 と S3 IA を比較すると，アクセス頻度が 1.9 か月に 1 回以下であれば S3 IA が有利（月あたりアクセス回数を a とし，7.065a+12.0（S3 IA の月総コスト）= 0.031a+15.8（S3 の月総コスト）から a=0.54, 1/a=1.9）.
- S3 IA と Glacier を比較すると，標準復元では月 6.6 回以上のアクセスで S3 IA が有利（上記同様 7.065a+12.0 =8.394a+3.2 から a=6.6）．しかし，この場合もっとも有利なのは，アクセス課金の不要な標準 S3 である.

表 6　大量データの書出し性能

ストレージ	サーバ	データ量 (TiB)	総時間 (時間)[注5]	CLI 多重度	スループット (MiB/s)[注1]	スループット 標準偏差
AWS S3 IA (2018)	VM [注2]	528	945.9	4	162.6	3.1
	NII [注3]	508	858.5	4	172.4	6.8
	合計	1,036	1804.4	8	167.2	-
Azure BLOB Hot (2018)	VM-1 [注4]	472	861.4	4	159.5	7.0
	VM-2 [注4]	263	546.0	4	140.5	9.8
	NII [注3]	294	822.0	4	137.8	28.9
	合計	1,029	2029.4	12	147.7	-
Oracle Archive [注5]	(VM)	718	3832.9	12	54.5	-

注 1: スループットは 1 コマンドあたり　注 2: m4.16xlarge 使用
注 3: オンプレミス環境・高性能機使用（図 1）
注 4: D5V2 使用　注 5: 2017 年に測定した参考値

表 7　大量オブジェクト群からの部分読出しの性能

ストレージ	読出し方法	オブジェクト母数	総所要時間(分)
AWS S3	BASF2 ライブラリ経由で 1,000 オブジェクト読出し	228,000	501
		1,000	493
Azure BLOB Hot	ダウンロード CLI により 1,000 オブジェクト読出し	106,835	198
		1,000	184

表 8　1PiB オブジェクトの一覧取得性能

ストレージ	総オブジェクト数	所要時間(秒)
AWS S3	1,677,000	860
Azure BLOB Hot	1,665,000	1712

- Glacier のバルク復元を使う場合は, アクセス頻度によらず，常に Glacier が有利となる.

ただし，Glacier では時間オーダの復元時間が必要となる．すなわち，Glacier を選択するかどうかの基準は，アクセス課金の額よりは，時間オーダの復元処理を許容できるかどうかにかかっていると言える.

(3) 大量データの書出しと操作

(1)と同じ 1,000 ファイル, 633GiB のデータを繰り返しアップロードすることによって，クラウドに 1PiB のデータを格納する実験を行った．書出しを加速するために，別々のコンテナにオブジェクトを書き出すアップロード CLI を 1 サーバあたり 4 多重で並行に動作させた．これを 2〜3 台のサーバで同時に実施したが，1 台は NII のオンプレミス環境の高性能サーバとし，残りのサーバは，ストレージと同じクラウド内の VM とした.

得られた経過時間およびスループットを表 6 に示す.

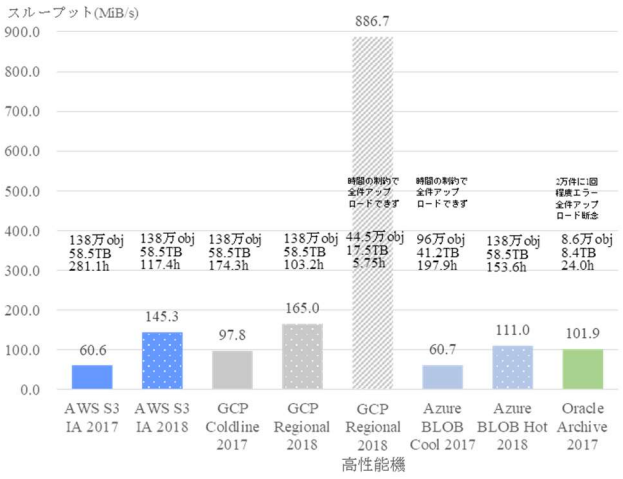

図 3　天文学(ALMA 望遠鏡) データのアップロード性能

2018 年に測定したクラウドでは，サーバ 1 台（コマンド 4 多重）で 500MiB/秒以上のスループットが得られている．また，オンプレミスのサーバとクラウド内の VM からの書出し性能には大きな差はない．従って，この条件下では，アップロード性能はクラウドストレージ内部の処理性能によってほぼ決まっていると推測される．なお，実際の実験実施時間は，総時間を CLI 多重度で除算した時間に近い．2018 年の実験では 200 時間前後（8～9 日）となっており，この程度の時間で 1PiB のデータをクラウドにアップロードできるということを示している．

このようにして作成されたコンテナの一つの中の 1,000 オブジェクトに対して(1)と同じ読出しを同じクラウド内の VM から行い，その性能を比較した．ただし，Azure の場合は，s3fs 相当のファイル API アクセスの手段が存在しないため，CLI によって 1,000 オブジェクトをダウンロードすることで代替した．結果を表 7 に示す．読出し性能はオブジェクトの総数には大きな影響を受けず，大量データの格納に対して，アクセスのスケーラビリティが実現されていることがわかる．

合わせて，CLI による 1PiB の全オブジェクトの一覧取得を実施し，その経過時間を測定した．結果を表 8 に示す．対話的な処理には適用困難な時間を要しており，大量データ格納システムの運用には，オブジェクト名を階層化して表示するなど，何らかの考慮が必要である．

4.2 ケーススタディ(2):
天文学データ（ALMA 望遠鏡データ）の格納とアクセス

(1) コールドストレージに対するデータのアップロード

表 2 に示した ALMA 電波望遠鏡の観測／解析データ約 138 万ファイル，58.5TiB を，NII のオンプレミス環境のサーバから，1 ファイルを 1 オブジェクトとして，コールドストレージ（早期削除課金を避けるために，性能が同等である標準オブジェクトストレージを使った場合もある）に対して，表 3 の CLI を使用してアップロードした．図 3 に，その経過時間およびスループットを示す．

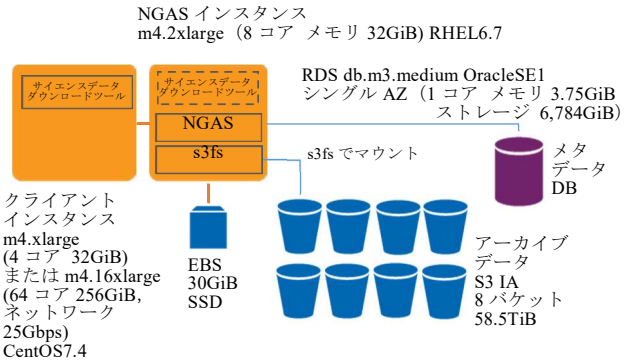

図 4　AWS 上に移植した NGAS の構成

高エネルギー物理学データの場合と同様に，2018 年に再測定を行ったクラウドについては，性能向上が認められる．高性能機を使用した場合に著しい性能向上が得られる場合があることも同様である．ただし，高エネルギー物理学データの場合と比較すると，平均オブジェクト長が小さい（高エネルギー物理学 648MiB に対して 44.4MiB）ために，スループットの値自体は小さくなっている場合が多い．

(2) アプリケーションによるデータの読出し

観測／解析データを管理しアクセスするアーカイブアプリケーション NGAS（Next Generation Archive System）は，複数のホスト，そのホスト配下のアーカイブデータを格納したディレクトリ群，およびメタデータ管理用の Oracle データベースによって構成される．本実験では，2 個のホストを AWS EC2 の同一インスタンス上に配置し，(1)でアップロードしたデータを格納したバケット 8 個を s3fs でファイルシステムとしてマウントした．一方，データベースは AWS のサービスである RDS for Oracle Database を使用して，構築および運用の負担軽減を図った．これらの方法によって，NGAS ソフトウェア自体を修正することなく，AWS 上で動作させることができた．さらに，NGAS から研究データを取り出すサイエンスデータダウンロードツールを，クライアントインスタンスあるいは NGAS ノードインスタンス上で動作させて測定を実施した．この構成を図 4 に示す．

ダウンロードツールでは，アーカイブデータ中の特定のファイル（オブジェクト）のグループに付された id を指定して，関連する全ファイルとそのメタデータをダウンロードする．いろいろなダウンロードデータ量に対するクラウド上の NGAS システムのスループット測定結果と，同じダウンロードを国立天文台のオンプレミスの NGAS システムで実行した場合のスループット測定結果を図 5 に示す．

全般に，クラウド上の NGAS は，オンプレミスの NGAS に対して性能が 1/2 程度まで低下している．原因として，S3 とオンプレミスのストレージシステムのスループット差，s3fs のオーバヘッド，さらに，実験で使用した RDS インスタンスの能力が低いことが考えられる．ただし，クラ

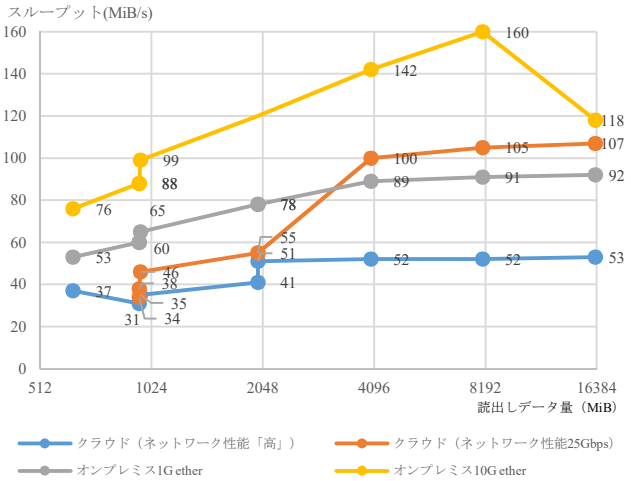

図 5　NGAS アーカイブデータのダウンロード性能

表 9　S3 IA 内の NGAS データに対するアクセスコスト

データ量(MiB)	628	949	1,985	4,012	8,047	16,253
リクエスト課金($)	0.0002	0.0004	0.0003	0.0005	0.0010	0.0179
データ量課金($)	0.0067	0.0100	0.0201	0.0399	0.0797	0.1604
読出し課金総計($)	0.0068	0,0103	00204	0.0404	0.0807	0.1783
インターネット経由転送課金($)[注1]	0.0859	0.1298	0.2714	0.5485	1.1002	2.2221
SINET 接続サービス経由転送課金($)[注1]	0.0258	0.0389	0.0814	0.1645	0.3301	0.6666

注 1: カタログ記載の料金からの算出値

イアントのネットワーク性能を上げればスループットが向上する傾向は共通であり，適切なサイジングを行えば，クラウド上においても実用的な性能が得られると予想される．

さらに，いくつかのダウンロードに対するアクセスコスト（課金額）を表 9 に示す．なお，この測定はダウンロードツールを NGAS インスタンス内に配置して実施したが，ストレージアクセスに関する課金額は同じである．

(3) Glacier を利用する場合の推定

(2)の測定では，アーカイブデータを S3 IA に格納したが，さらに容量単価の低い Glacier に格納することも考慮に値する選択肢である．しかし，Glacier にデータを格納した場合は，オブジェクト（ファイル）の読出し前に復元リクエストを発行する必要があり，NGAS ソフトウェアの修正を要する．本実験では，ダウンロードツールでダウンロードしたアーカイブデータを別の S3 バケットに格納し，ライフサイクル管理機能によって Glacier に自動移行させたものに対して CLI を使用して手作業で復元処理を実施した結果から，復元性能と復元課金を推定した．結果を表 10 に示す．復元時間は，高エネルギー物理学データの測定結果とほぼ同等である．すなわち，今回の測定の範囲では，復元時間は，データ量や同時に発行する復元リクエスト数には

表 10　Glacier 上の NGAS データの復元時間と課金額

復元データ量(MiB)	949	8,047	16,253	16,253
復元オブジェクト数	350	348	291	291
復元方法[注1]	標準	標準	標準	バルク
復元時間(分)	200	200	200	327
復元リクエスト課金($)	0.0200	0.0199	0.0166	0.0080
復元データ量課金($)	0.0102	0.0864	0.1745	0.0436
復元スペース課金($)	0.0005	0.0041	0.0082	0.0082
読出し課金($)[注2]	0.0001	0.0004	0.0066	0.0066
課金額総計($)	0.0308	0.1107	0.2060	0.0665

注 1: 復元後の保持期間は 1 日とした．

注 2: S3 IA のリクエスト課金額実測値から，カタログ記載の料金の差異を考慮して推定（$0.000001/リクエストに対して $0.00000037/リクエスト）．

依存しないことがわかる．

NGAS で Glacier を利用するためには，おおよそ以下のような修正が必要である．なお，Glacier へのデータ移行は S3 のライフサイクル管理機能によって行われているとする．

i) メタデータデータベースから，取り出すファイルに対応するバケット名とオブジェクト名を得る．
ii) S3 の API を使って当該オブジェクトの格納状態の情報を取得する．Glacier に移行されていなければ（あるいはたまたま前回の復元結果が残っていれば），そのまま S3 オブジェクトとしてアクセスを続行する．
iii) Glacier に移行されている場合は，当該オブジェクトに対する復元要求を API 経由で発行する．
iv) 復元処理の完了を適切な時間間隔（最初は 1 時間ごと，3 時間を過ぎたら 5 分ごと，など）でポーリングする．
v) 復元完了後に S3 オブジェクトとしてアクセスする．

なお，復元要求の発行と完了待ちを複数オブジェクトに対して順次行うと膨大な時間を要するため，1 回のデータのダウンロード要求で取り出すべきオブジェクト群を先に特定し，それらに対して一挙に復元要求を発行して，すべてが完了するまでポーリングを繰り返すという考慮が必要である．また，復元処理によって時間オーダの遅延が生じるため，ダウンロード処理全体に対するタイムアウト時間の見直しも必要である．

(4) データ保管コストおよびシステム維持コスト

AWS の課金情報および(2)で測定したアーカイブデータダウンロードにおける課金情報に基づいて，S3 IA, Glacier, S3 のそれぞれにデータを格納した場合の保管コストと，図 4 の AWS 上の NGAS の構成から NGAS システム維持コストを算出したものを表 11 に示す．本実験では，RDS インスタンスはテスト用の小規模構成（1 コア，メモリ 3.75GiB）としたが，表 11 では，より現実的な規模（4 コア，メモリ 15GiB）を想定して算出を行った．

表 11　NGAS のシステム維持コストとデータ保管コスト

項目	仕様・量	月額($)
NGAS ノード インスタンス	m4.2xlarge 1 台 (8 コア/メモリ 32GiB/RHEL)	371.5
NGAS ノードストレージ	30GiB SSD	12.0
Oracle RDS インスタンス	db.m3.xlarge 1 台 (4 コア/メモリ 15GiB)	756.0
Oracle RDS ストレージ	6,784GiB 注 1	936.2
NGAS ファイルチェック	350,000 リクエスト/時注 2	288.0
データ保管	S3 IA　1TiB	19.46
	Glacier　1TiB	5.12
	参考: S3 1TiB	25.60

注 1: DB 作成時のデータが残っており，削減の余地がある．
注 2: S3 IA を想定（Glacier は方式自体の検討が必要）

表 12　年間総運用コストの試算例

項目		値
アーカイブデータ総量		500TiB
年間ダウンロード量注 1		250TiB
アーカイブデータの平均オブジェクト長		40MiB
年間総運用コスト	全データを S3 IA に格納	$158,668
	全データを Glacier 注 2 に格納	$74,191

注 1:データ転送は SINET 接続サービスを利用するものとする．
注 2:Glacier 復元速度は「標準」とする（測定結果では 200 分待ち）

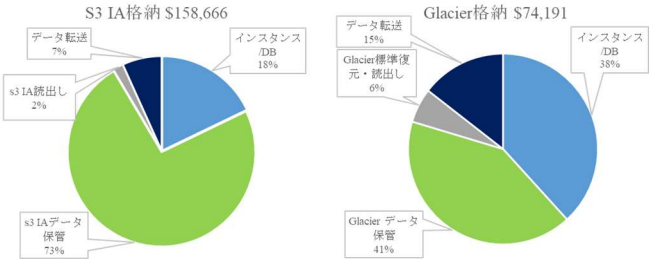

図 6　年間総運用コストの試算例における内訳

なお，NGAS は，ファイルデータの逸失や破壊を検出するために，定期的にファイルのチェックを行っているが，クラウドストレージでは，そのアクセスリクエストに対して課金される．リクエストの数は時間によって変動しているが，実測結果から 1 時間あたり約 350,000 リクエストが発行されると推定して，費用を算入した．

(5)　いろいろな構成例における総コスト試算

これまで述べてきた値を用いた年間総運用コストの試算例を表 12 に，その内訳を図 6 に示す．

表 12 では，すべてのアーカイブデータをクラウドに置く場合を想定しているが，より現実的な構成として，すでにオンプレミス環境に存在する NGAS システムとクラウド

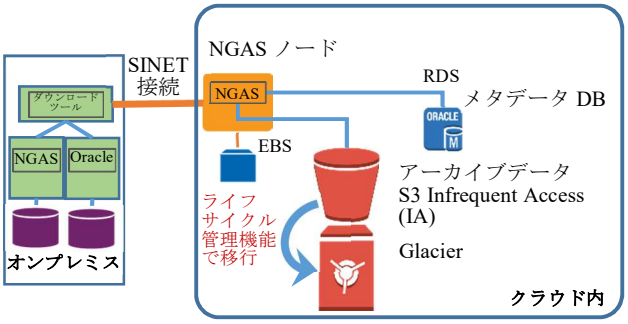

図 7　ハイブリッド・階層ストレージ構成

表 13　ハイブリッド構成の年間総運用コストの試算例

項目	値
アーカイブデータ総量	1,000TiB
うち　オンプレミスのデータ量	500TiB
S3 IA のデータ量	400TiB
Glacier のデータ量	100TiB
年間ダウンロード量注 1	250TiB
うち S3 IA からのダウンロード量（全体の 20%）	50TiB
Glacier 注 2 からのダウンロード量（全体の 10%）	25TiB
アーカイブデータの平均オブジェクト長	40MiB
クラウドの年間総運用コスト	$128,895

注 1:データ転送は SINET 接続サービスを利用するものとする
注 2:Glacier 復元速度は「標準」とする（測定結果では 200 分待ち）

上の NGAS システムによるハイブリッド構成が考えられる．一定の保管期間を経過したアーカイブデータをクラウドに移行する運用を行うことによって，オンプレミス環境に対する投資額を一定とし，また，アクセス頻度の高い新しいデータに対するクラウドのアクセスコストとデータ転送コストを削減することが可能となる．NGAS は，もともと複数ホストによる構成をサポートしており，このような運用の実現は，比較的容易と考えられる．

一方，Glacier は，保管コストは低いものの 3 時間を超える復元時間が必要であり，すべてのアーカイブデータを Glacier に配置すると，データダウンロードのサービスレベルが大幅に低下する．そこで，S3 IA に対してライフサイクル管理機能を適用し，一定期間経過後のアーカイブデータを Glacier に自動移行することによって，大半のデータダウンロードに対するサービスレベルを維持することが可能となる．この二点を考慮した構成の概略を図 7 に，年間総運用コストの試算例を表 13 に示す．

4.3　ケーススタディ(3):
天文学データ（野辺山望遠鏡データ）の格納とアクセス

AOS（Adria Object Storage）は，野辺山望遠鏡データのアーカイブのバックエンドシステムである．本実験は，アー

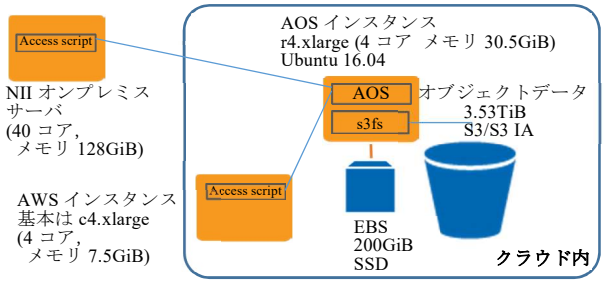

図 8　AWS 上に移植した AOS の構成

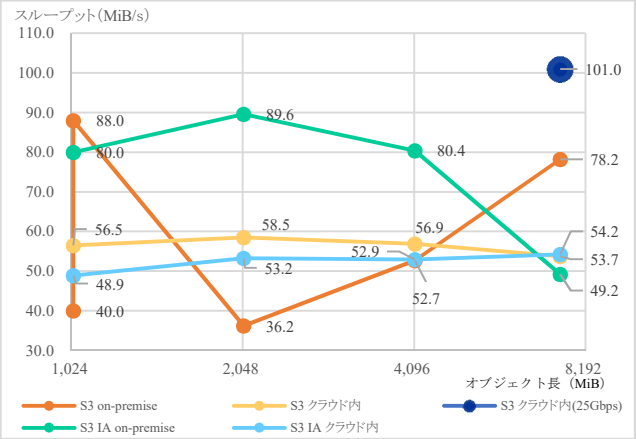

図 9　AOS のスループット実測値

カイブデータを S3 および S3 IA に格納し，AOS ソフトウェアを AWS 上で動作させることによって，クラウド上の運用における性能やコストに関する感触を得ることを目的として実施した．AOS 自体はオブジェクトストレージシステムであるが，現時点では， POSIX ファイルシステムを対象ストレージとして実装されているので，今回は無修正のまま実験できるように，S3/S3 IA のバケットを s3fs で FUSE マウントしてアクセスするようにした．

図 8 に構成を示す．ここで，アーカイブデータを S3 と S3 IA に格納し，同一リージョン内のインスタンスあるいは NII のオンプレミス環境のサーバ（インターネット経由）から種々のサイズのオブジェクトを取り出すスループットを測定した結果を図 9 に示す．クラウド内からのアクセスでは，S3，S3 IA ともにスループットはほぼ同等である．一方，オンプレミスからのアクセスではスループットが変動し，同じオブジェクトに対して異なる値が測定されることもあるが，上限値に関しては比較的近い値が得られている．

次に，S3 および S3 IA から AOS 経由でアーカイブデータを取り出した場合の課金額を実測した．結果の一例を表 14 に示す．表には，今回実験で使用した 3.5TiB のデータを S3 および S3 IA に 1 か月保管した場合の課金額，S3 から S3 IA への自動移行に要した課金額も併記した．コールドストレージである S3 IA では，保管課金は下がるが，読出し課金が上がる．他のオブジェクト長の測定値と合わせて，

表 14　AOS のオブジェクト読出しに対する課金額

データ量 1,032 MiB のオブジェクト読出しの実測値				
サービス	S3 IA		S3	
出力先	EC2	NII	EC2	NII
読出しリクエスト課金($)	0.00011	0.00011	0.00004	0.00004
読出しデータ量課金($)	0.01008	0.01008	-	-
インターネット経由転送課金($)	-	0.14756	-	0.14769
アクセスコスト総計($)	0.01019	0.15775	0.00004	0.14773
28,979 オブジェクト 3.5TiB の自動移行および保管料金				
月額保管課金($)[注 1]	60.6		81.3	
S3 から S3 IA への自動移行料金($)	$0.2898			

注 1: カタログ記載の料金からの算出値

本例のコールドストレージの適用に関しては，約 2.2TiB/月以下の読出し量であれば，S3 から S3 IA に移行することによって総費用が下がるという結果が得られた．

5.　まとめと今後の取組み

本報告では，複数の商用パブリッククラウドで提供されるコールドストレージについて，ケーススタディとして，実際の研究データと研究アプリケーションを使ったいくつかの実験結果を述べた．

性能面では，クラウド自体が非常に速く進化しており，1 年間で大幅な性能向上が実現されていることが明らかとなった．また，性能に加えて，実際のユースケースにおける課金額を測定することによって，クラウド上のデータ保管と利用に関する総コストの推定に利用できる情報を得ることができた．コールドストレージでは，データ保管課金に注意が向きがちであるが，その他に，アクセスのためのリクエストと読出しデータ量に対する課金，さらにクラウド外へのデータ転送課金が発生する．アクセス課金が総コストに及ぼす影響はあまり大きくないが，クラウド外へのデータ転送課金は総コストに大きく影響する場合があり，クラウドとオンプレミス環境を通したシステム構成を考える際には，考慮が必要であることがわかった．合わせて，Glacier のようなデータアクセスに時間オーダの待ち時間を要するタイプのコールドストレージは，コスト面からは有力な選択肢であり，適切なストレージ階層化を行うことによって，システム全体のコストパフォーマンスを最適化できる可能性があることがわかった．

今後の取組みとしては，個々のケーススタディにおいて，より最適なアクセス方法や管理方法を検討し，性能・コスト・運用性の改善の可能性を探ることに加えて，以下のような展開が考えられる．

(1)　SINET クラウド接続サービスの活用

NII が提供する「SINET クラウド接続サービス」[10]は，

クラウドプロバイダのデータセンタを SINET に直結するものであり，実験対象のクラウドでは，AWS および Azure が対応している（2018 年 6 月現在）．クラウド外へのデータ転送課金がアクセスのコストに影響するケースでは，本サービスを利用することによって，データ転送性能向上やセキュリティ向上に加えて，データ転送課金低減が期待できる．クラウド内の VM に対する SINET 経由のアクセスは，これまでの使用実績も多く，NGAS や AOS のように VM 経由でストレージのデータにアクセスする場合に適用できる．一方，クラウドストレージに直接 SINET 経由でアクセスする場合は，VM 経由とは異なる接続方法が必要であり，利用例も少ないため，試行を通じた情報収集とノウハウ蓄積を進めたい．

(2) コールドストレージの特性を考慮したアプリケーションや運用の最適化方針および実装方法の蓄積

今回の実験では，研究データの 1 ファイルをそのまま 1 オブジェクトとして格納した．しかし，Glacier のように復元処理が必要なコールドストレージでは，アクセスする全オブジェクトに対して事前に復元リクエストを発行して完了を待つ必要があり，アクセスするオブジェクト数が多い場合は，リクエスト発行数が増えるとともに，処理が複雑化する．従って，1 回のアクセス単位を少数のオブジェクトにまとめてコールドストレージに格納するといった対策を考える必要がある．

一方，アプリケーションを無修正で動作させるために，本実験では s3fs を多用した．しかし，そのオーバヘッドや連続運用時の安定性の問題に加え，意図しないバッファリングなどによるストレージの読出し課金やデータ転送課金が発生する可能性も否定できない．アプリケーションのファイル API の使用をオブジェクトストレージ API に変更することによって，性能向上に加えて，課金額の予測可能性を高める効果も期待できる．

(3) クラウドコールドストレージ活用の支援情報・ツールの継続的蓄積

情報蓄積としては，ベンチマークや実証実験の結果の蓄積と共有を進めたい．特に，今回明らかになったように，1 年間でクラウドストレージの性能が大幅に向上するといったクラウドサービスの進化の速さを考えると，定点観測としての測定を継続し，データを蓄積してゆくことが重要と考える．

ツールの蓄積としては，オンデマンドクラウド構築サービス[11]において，コールドストレージ利用環境構築（オブジェクトストレージアクセス・管理環境や，コールドストレージのデータ復元機能などを含む）を自動化するテンプレートを提供・拡充してゆくことが考えられる．

はじめに述べたように，研究データの長期保管におけるコールドストレージ活用の検討材料となる知識を，実践を通じて獲得することを目指して実験を継続してきた．研究データの長期保管において，クラウドコールドストレージサービスは一つの素材であり，最終的には，オンプレミスとクラウドの双方を組み合わせるアーキテクチャと運用によって費用対効果を最適化してゆくことが必要であると考える．今後の実証実験を，そのためのベストプラクティスの蓄積と共有につなげてゆきたい．

謝辞 実験を進めるにあたりデータやソフトウェアの提供と，ご指導・ご助言をいただいた各研究機関の皆様，結果の分析やチューニング方法の検討に関してご協力いただいた各プロバイダの皆様，実験に必要なクラウドサービスの調達をご支援いただいた国立情報学研究所の関係者の皆様に，謹んで感謝の意を表する．なお，本研究で使用したクラウド資源の一部は，平成 29 年度国立情報学研究所クラウド利活用実証実験において提供された．

参考文献

[1] ストレージ ネットワーキング・インダストリ・アソシエーション（SNIA）日本支部．コールドストレージの最新動向．情報処理, 2017, vol.58, no. 12, p. 1107-1113.
[2] 吉田浩, et al.. 情報処理学会研究報告ハイパフォーマンスコンピューティング（HPC）, 2017, vol. 2017-HPC-160, no. 25, p.1‐8.
[3] “Amazon Glacier Developer Guide API Version 2012-06-01”. http://docs.aws.amazon.com/amazonglacier/latest/dev/glacier-dg.pdf (参照 2018-06-26).
[4] “Working with Amazon S3 Objects - Storage Classes”. https://docs.aws.amazon.com/AmazonS3/latest/dev/storage-class-intro.html (参照 2018-06-29).
[5] “ストレージクラス”. https://cloud.google.com/storage/docs/storage-classes (参照 2018-06-29).
[6] “Azure Blob Storage: ホット，クール，およびアーカイブストレージ層”. https://docs.microsoft.com/ja-jp/azure/storage/storage-blob-storage-tiers (参照 2018-06-29).
[7] “Oracle Cloud Infrastructure Object Storage Classic”. http://docs.oracle.com/en/cloud/iaas/storage-cloud/index.html (参照 2018-06-29).
[8] “libfuse”. https://github.com/libfuse/libfuse (参照 2018-06-29).
[9] “s3fs-fuse”. https://github.com/s3fs-fuse/s3fs-fuse (参照 2018-06-27).
[10] “クラウド接続”. https://www.sinet.ad.jp/connect_service/service/cloud_connection (参照 2018-06-27).
[11] 竹房あつ子, et al.. インタークラウド環境構築システムの開発．信学技報, 2017, vol. 117, no. 153, CPSY. 2017-17, p.7-12.