「ビッグデータを取り巻く制約からの解放」

PCクラスタワークショップ in 大阪2023「ビッグデータとHPC」

2023年6月21日　於：富士通 Osaka Hub

Pacific Teck Japan 合同会社　Senior Engineer 森本 賢治

Pacific Teck ポートフォリオ

- ジョブスケジューラ
  - ALTAIR
- HPC向けコンテナ
  - PRO
  - ENTERPRISE
- ストレージソフトウエア
  - 並列ファイルシステム
    - BeeGFS®
    - BeeOND®
  - オブジェクトストレージ
    - DATACORE
    - cunoFS unleashing object storage
- S3互換クラウドストレージ
  - wasabi
- ルール指向高機能ストレージ構築ミドルウエア
  - iRODS®
- クラスター構築・管理ツール
  - Bright Computing
- 開発及びプロファイリングツール
  - Linaro
  - oneAPI
  - ALTAIR

随時拡大していきます

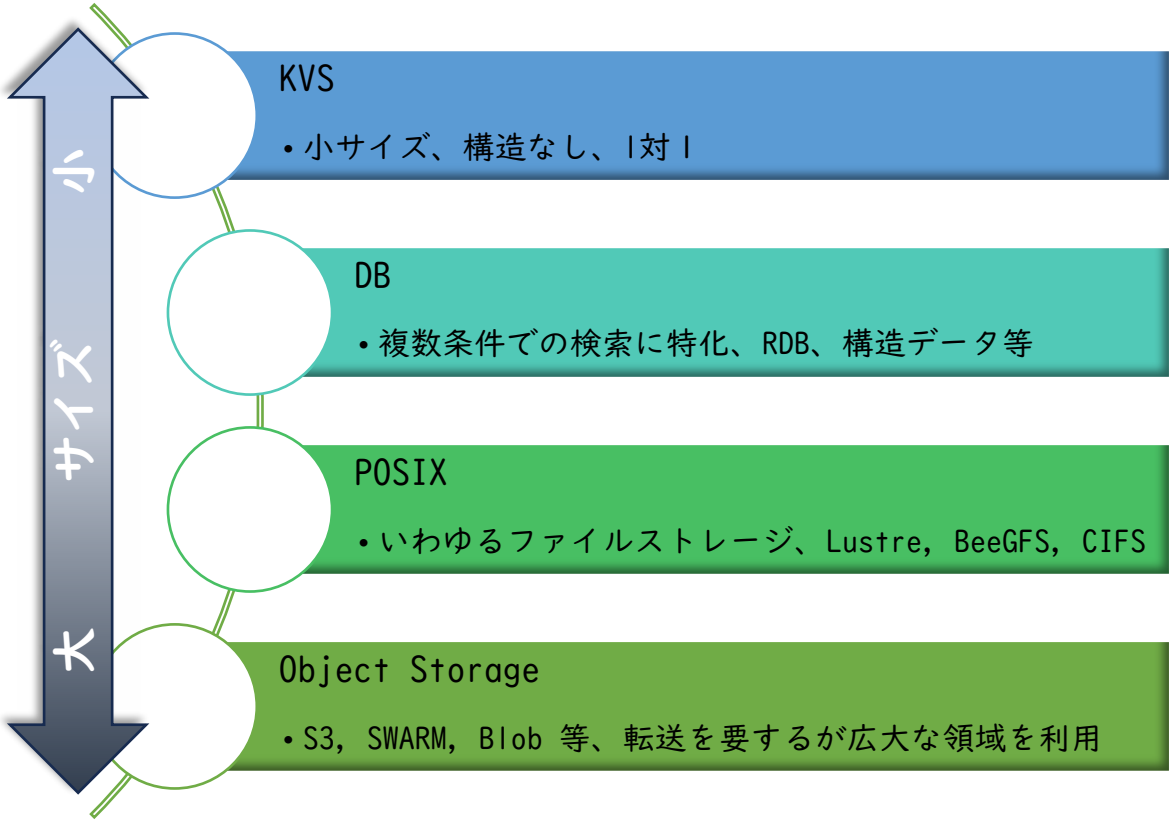
はじめに

ビッグデータとはもともとデータマイニング界隈で用いられていた単語。
一般的に用いられるようになったのは2010年頃と言われる。
定義は明確ではなく、容易にハンドリングできない量とサイズのデータ群という共通認識のみ。
当時は全世界のデータの合計はまだZBに届くか届かないかの様相。

2015年頃にその活用が始まったとされ、活用元年などと言われた

データを生成するデバイス数の増大、それを処理する計算能力の向上、
第3次から切れ目なく続く、第4次AIブーム。
データがデータを生み、益々複雑な構造が作られる状況に拍車がかかる。

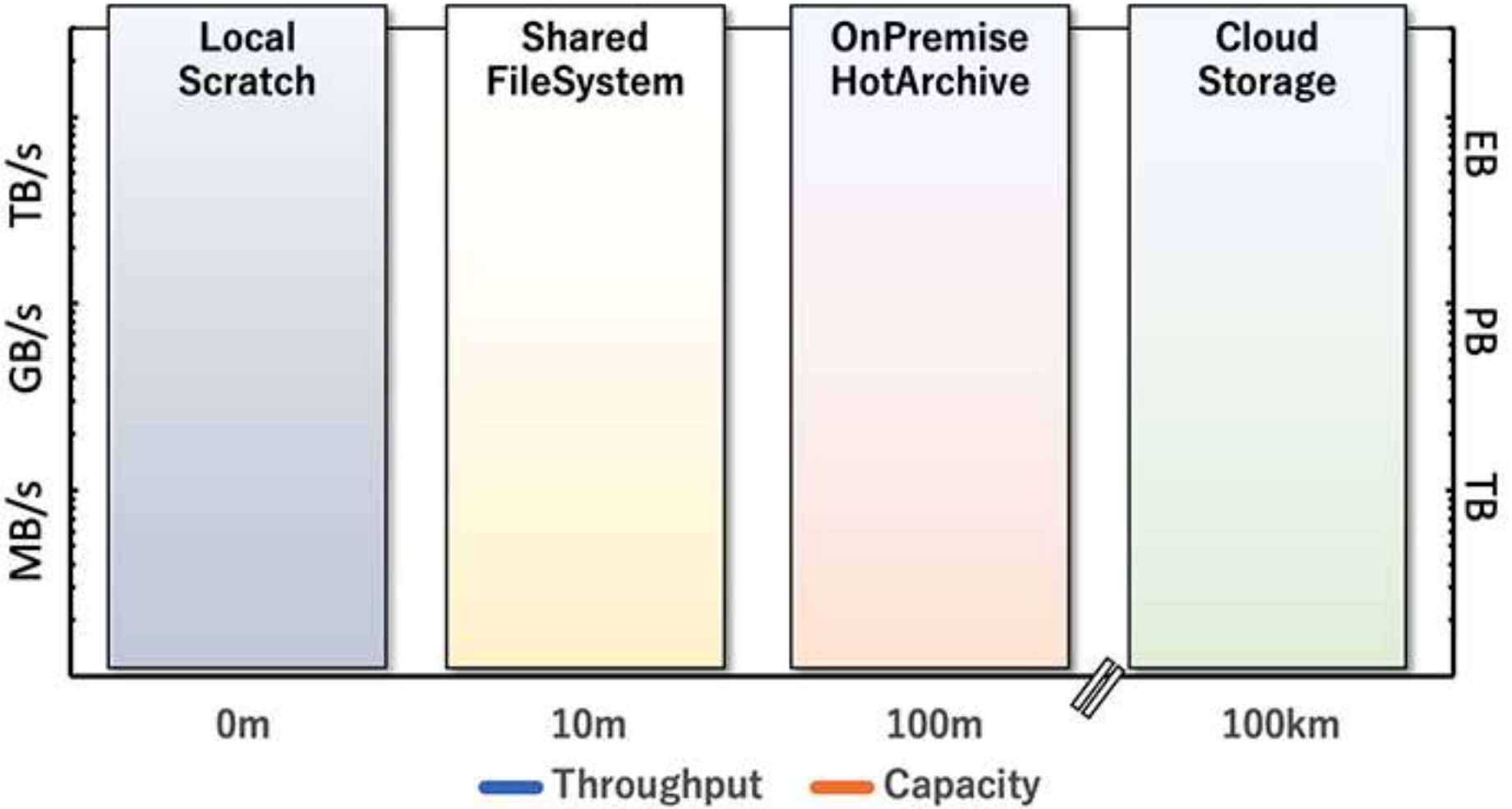
ストレージの現状について



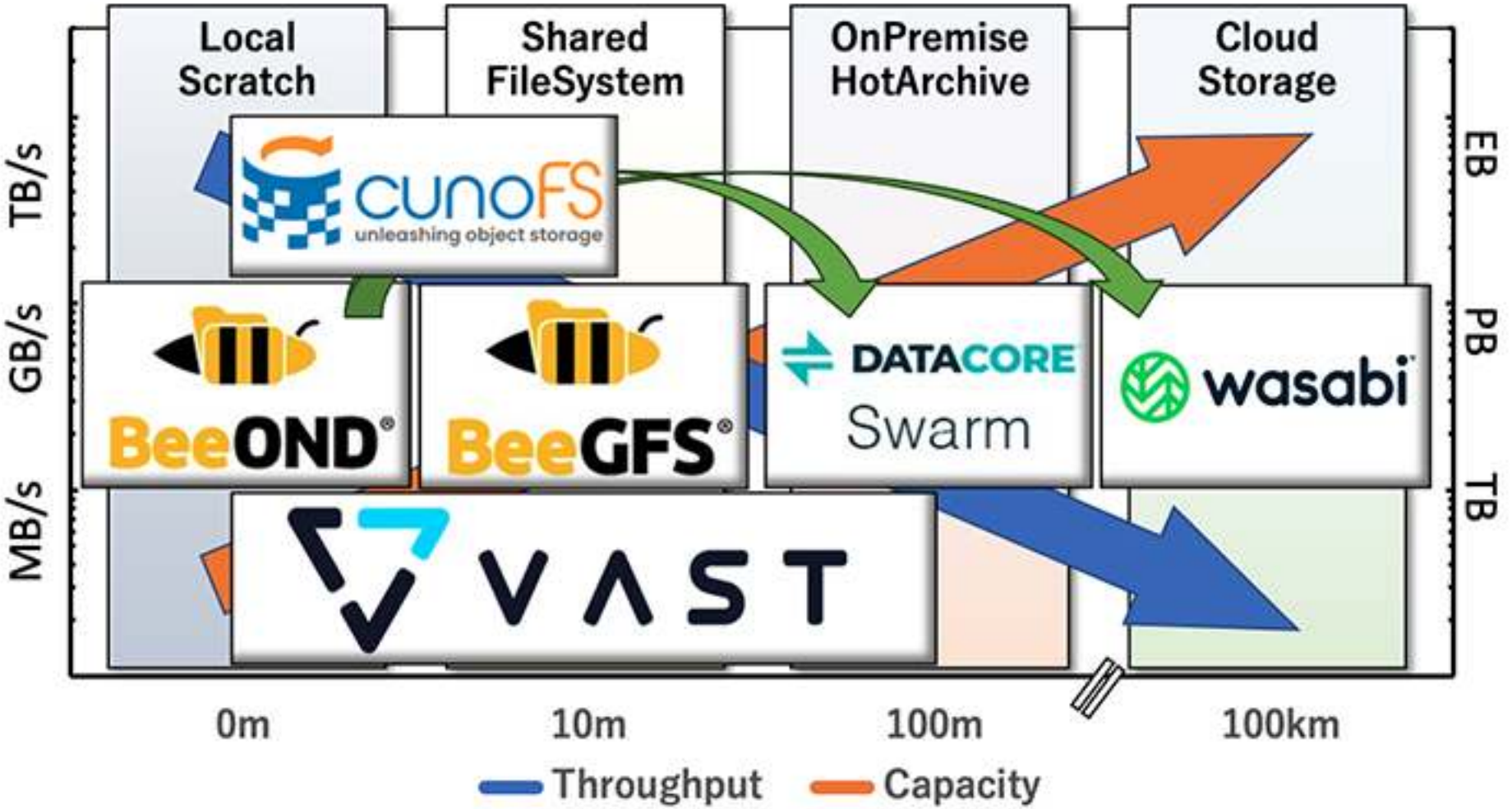
これら要素技術を組み合わせて、要件と予算に見合うストレージをカスタム

今後もこれで良いですか？

多彩なストレージポートフォリオ



多彩なストレージポートフォリオ



製品群ご紹介

VASTの構造と思想

CLIENTS: NFS, NFS+RDMA+GPUDirect, SMB, S3　同一データにマルチプロトコルでアクセス可能。さらに追加予定

PROTOCOL SERVERS プロトコル変換のみ。ステートレスでキャッシュ無し

NFS Multipathing

NVME FABRIC: COMMODITY ETHERNET OR INFINIBAND

NVMe ENCLOSURE
8x SCM
22x QLC FLASH

NVMe ENCLOSURE
8x SCM
22x QLC FLASH

NVMe ENCLOSURE
8x SCM
22x QLC FLASH

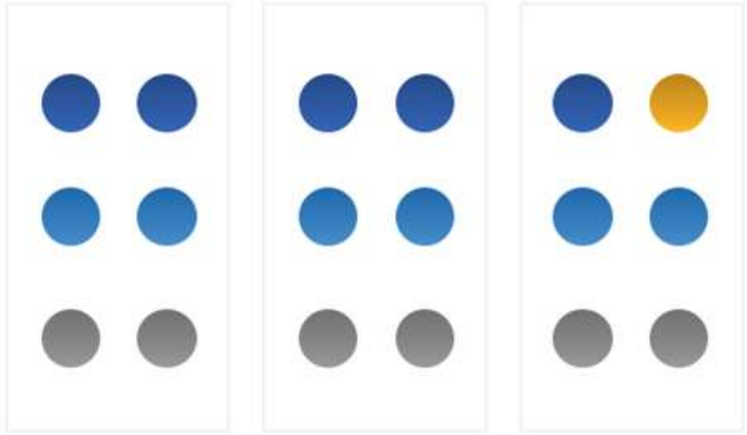
GLOBAL NAMESPACE

QLC NANDとSCMを用い、
大容量、高速、長寿命
高可用性を同時に実現
150/146のErasureCode

データ圧縮と重複排除のハイブリッド技術

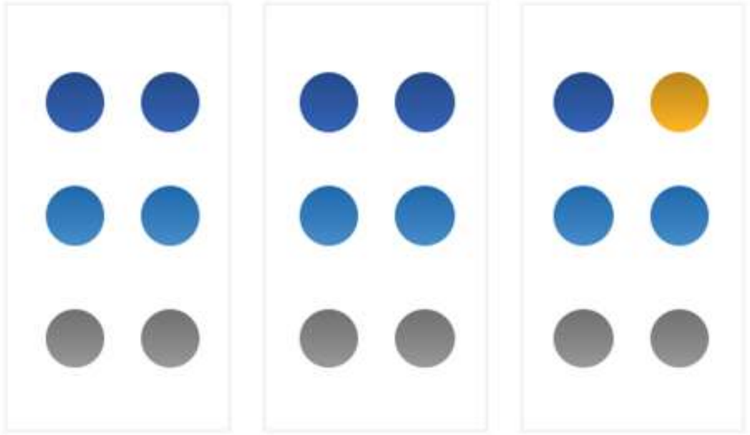
SIMILARITY IS GAME-CHANGING

COMPRESSION



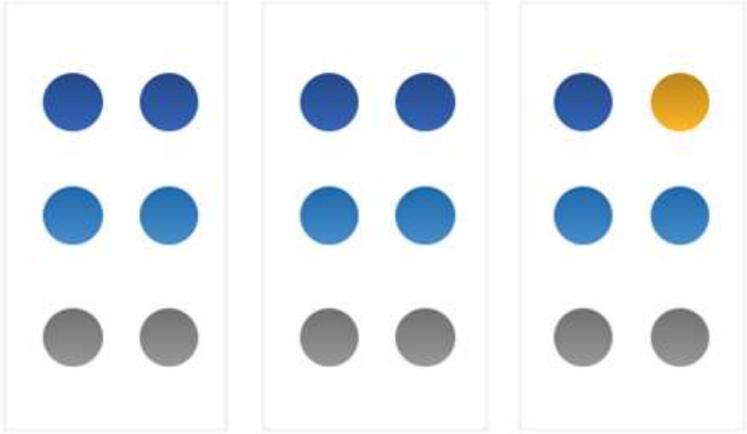
Fine Grained, But Local

DEDUPLICATION



Global, But Coarse

VAST DATA SIMILARITY REDUCTION



Global & Fine Grained

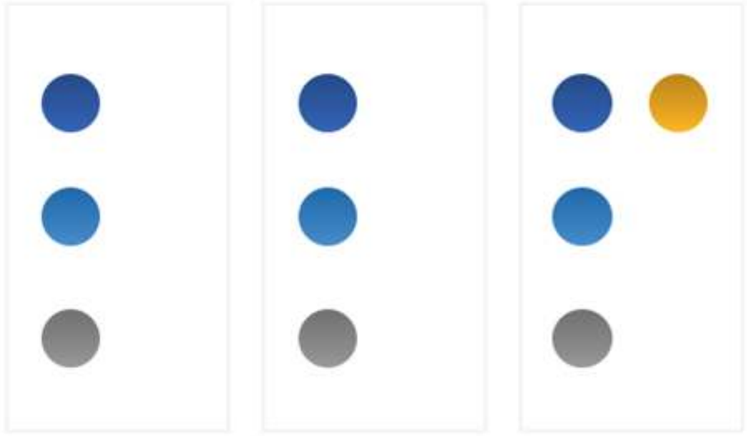
EXAMPLE SAVINGS FROM SIMILARITY

3:1	Pre-Reduced Backups	**3:1**	Pre-Compressed Log Files	**2:1**	Life Science Data	**3:1**	HPC Data	**8:1**	Uncompressed Time-series Data

データ圧縮と重複排除のハイブリッド技術

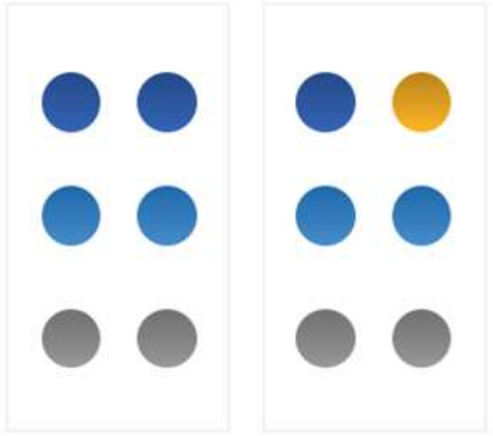
SIMILARITY IS GAME-CHANGING

COMPRESSION



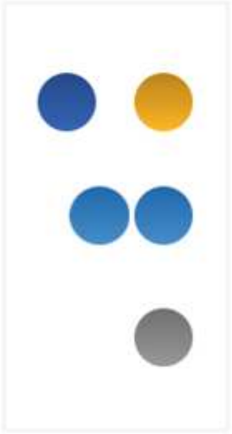
Fine Grained, But Local

DEDUPLICATION



Global, But Coarse

VAST DATA SIMILARITY REDUCTION



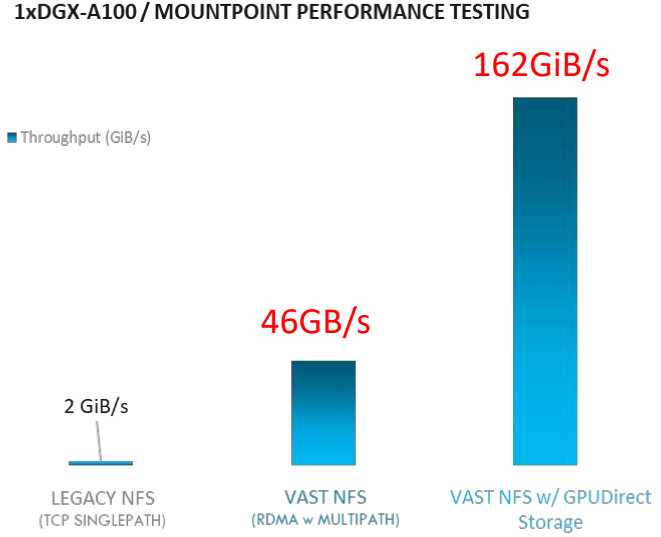
Global & Fine Grained

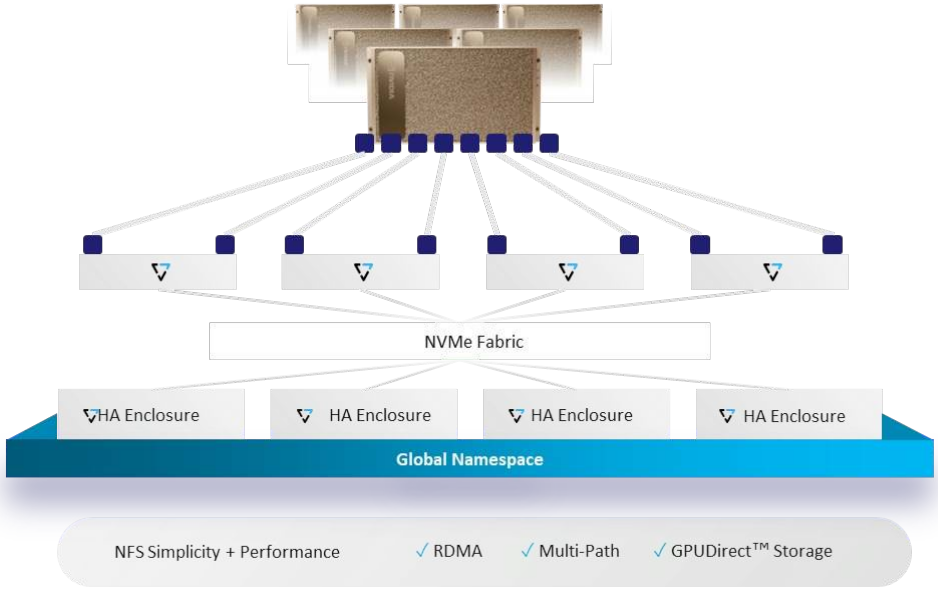
EXAMPLE SAVINGS FROM SIMILARITY

3:1	Pre-Reduced Backups	**3:1**	Pre-Compressed Log Files	**2:1**	Life Science Data	**3:1**	HPC Data	**8:1**	Uncompressed Time-series Data

VASTのスループット

AI PERFORMANCE + NAS SIMPLICITY





VASTに標準搭載の機能

NFS v3.0 & v4.1

w/ RDMA, GPUDirect, byte-range locks

Encryption-At-Rest

FIPS-grade security

Similarity-Based Data Reduction

Unprecedented storage efficiency

SMB 2.1 (resilient SMB)

Stateful file services; stateless servers.

No-Overhead Snapshots

Space and performance efficient

Locally-Decodable Data Protection

Unrivaled resilience, only 2.5% overhead

S3-Compatible API, w HTTPS

Stateful file services from containers.

Replication to Cloud/S3

Snapshot to an S3 endpoint of choice

Automation Plugins

Kubernetes CSI, Manilla Driver

Multi-Protocol Namespace

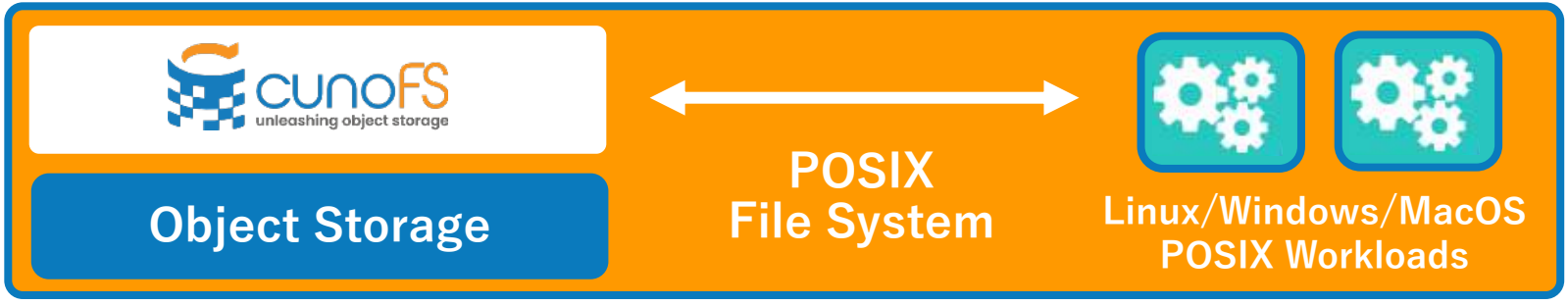
Easily traverse between file & object

Enterprise Authentication

Support for LDAP, AD, NIS

Directory Quotas

CUNO FS



CUNOは多彩なオブジェクトストレージをPOSIXファイルシステムとして
マウントしているかのように使うことができるツール。
S3等の広大な容量を持つオブジェクトストレージから、ステージングする
ことなくファイルの直接操作を実現。
同様なツールとは一線を画す、高いパフォーマンスと使い勝手。

CUNOの使い勝手

URI-based access

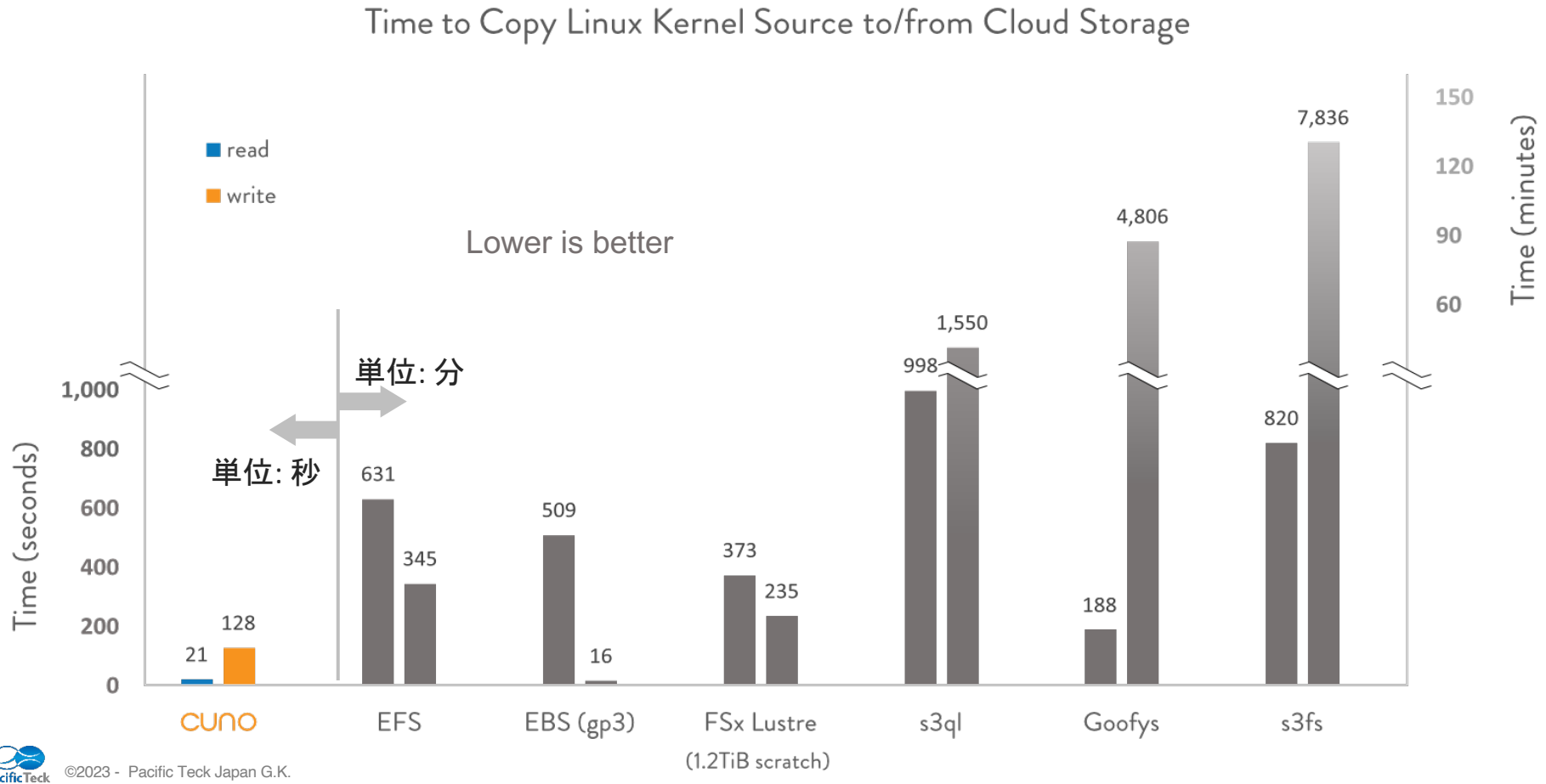
```
~ cuno
(cuno) ~ cd s3://bkt
(cuno) s3://bkt tar xfz az://pg/bkt/archive.tar.gz
(cuno) s3://bkt chmod a+x script.sh
(cuno) s3://bkt grep -R test > tests.txt
```

Unmodified binaries
Cloud storage behaves like regular paths

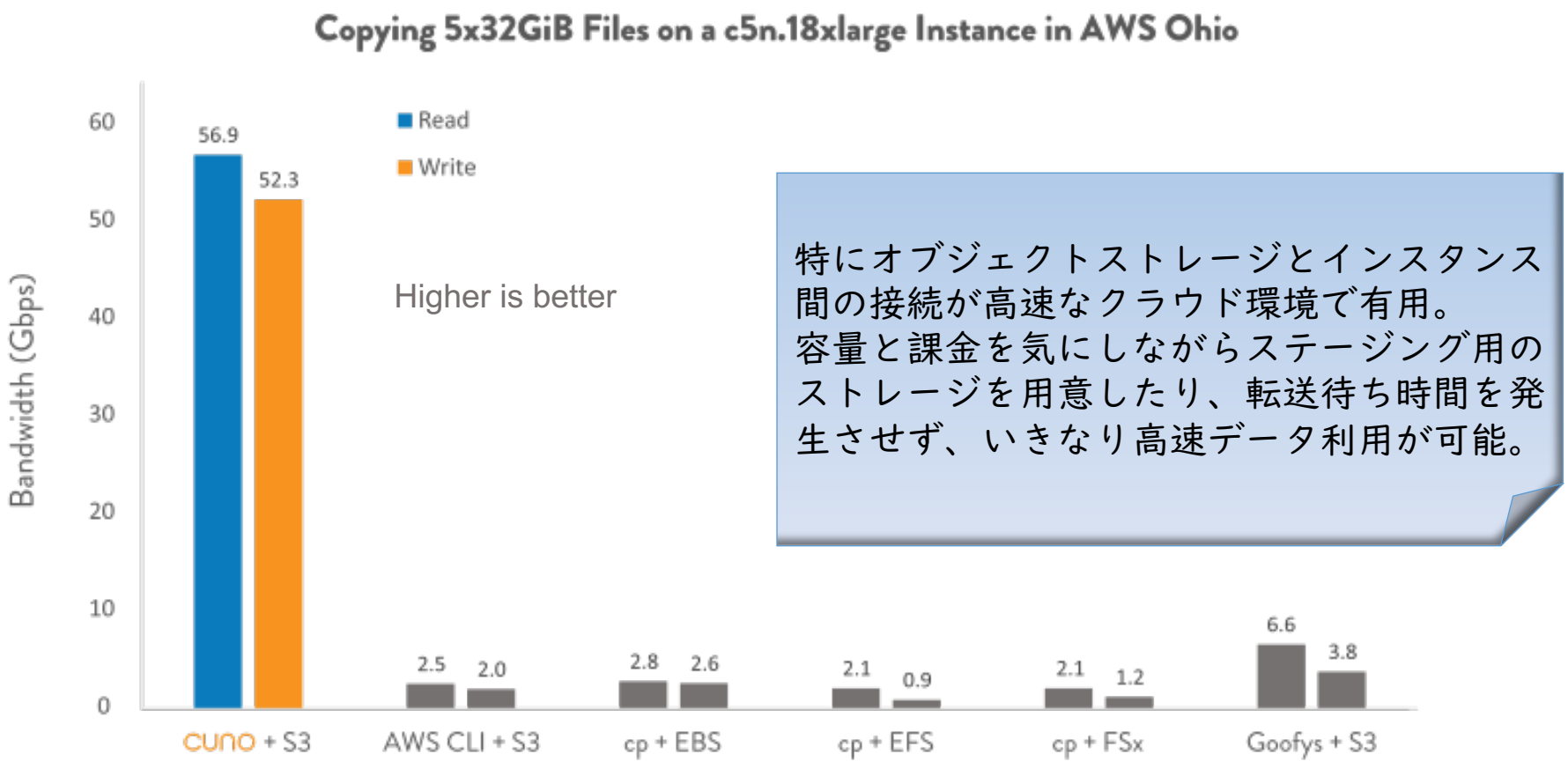
Path-based access

```
~ cuno
(cuno) ~ cd /cuno/s3/bkt
(cuno) /cuno/s3/bkt tar xfz /cuno/az/pg/bkt/archive.tar.gz
(cuno) /cuno/s3/bkt chmod a+x script.sh
(cuno) /cuno/s3/bkt grep -R test > tests.txt
```

多ファイルコピーでの性能評価



大ファイルアクセス時の性能比較



特にオブジェクトストレージとインスタンス間の接続が高速なクラウド環境で有用。
容量と課金を気にしながらステージング用のストレージを用意したり、転送待ち時間を発生させず、いきなり高速データ利用が可能。

BeeGFSの新しい機能拡張

BeeGFS HIVE

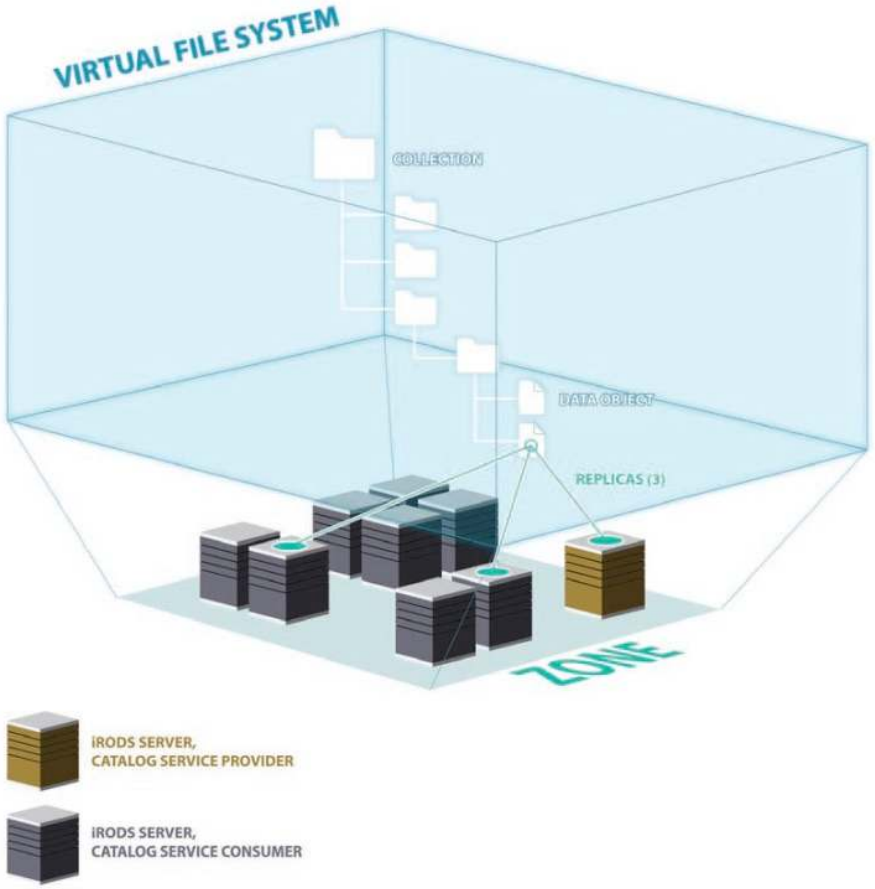
- BeeGFSに保存されているデータのインデックスサービス。OSのシステムコールを使うことなく、
  膨大な数のファイルの検索等を、クエリで超高速に行うことが可能。
- OSSのGUFI（Grand Unified File Index)をBeeGFSに取り込んだ実装。
- 最新版 7.3.3 で適用可能。
  https://doc.beegfs.io/latest/hive/hive_index.html

Integrated Rule Oriented Data System

- RENCI(Renaissance Computing Institute)
  米国の複数大学のコンソーシアムにより
  20年間開発されているミドルウエア
- 修正BSDライセンスのオープンソース
- 組織間連携を含む大規模サイトでの
  導入事例が多数。
- 複数のサーバーの管理下にあるファイル
  ストレージを単一ネームスペースで統合
  した仮想ファイルシステムを構成
- ユーザー定義のメタデータ付与により
  ポリシーベースのデータ移動をはじめ
  さまざまな自動処理を実現

iRODS の基本概念 Zone, Catalog, Resource

- 複数のサーバーの管理下にある、ファイルシステムやオブジェクトストレージ(リソース)を持ち寄って構成される、単一名前空間の仮想ファイルシステムを Zone と呼ぶ。
- ファイルやリソースに付与されるメタデータや配置情報等の総称をカタログデータと呼びRDBで管理。
- リソース間でレプリケーション等を実現する仮想リソースを設定し、それらを組み合わせて Zone を構成。



Catalog Data

ファイル

ファイル名＋拡張子
所有者
ファイルサイズ
最終アクセス日時

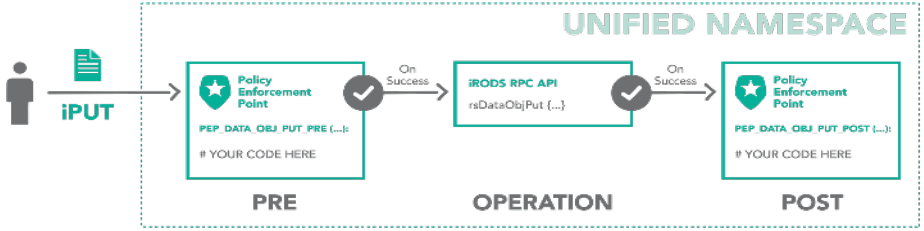
ファイルフォーマット
想定用途
チェックサム
画像解像度
前処理済みか未然か
正常データかエラーデータか
生成元情報　etc…

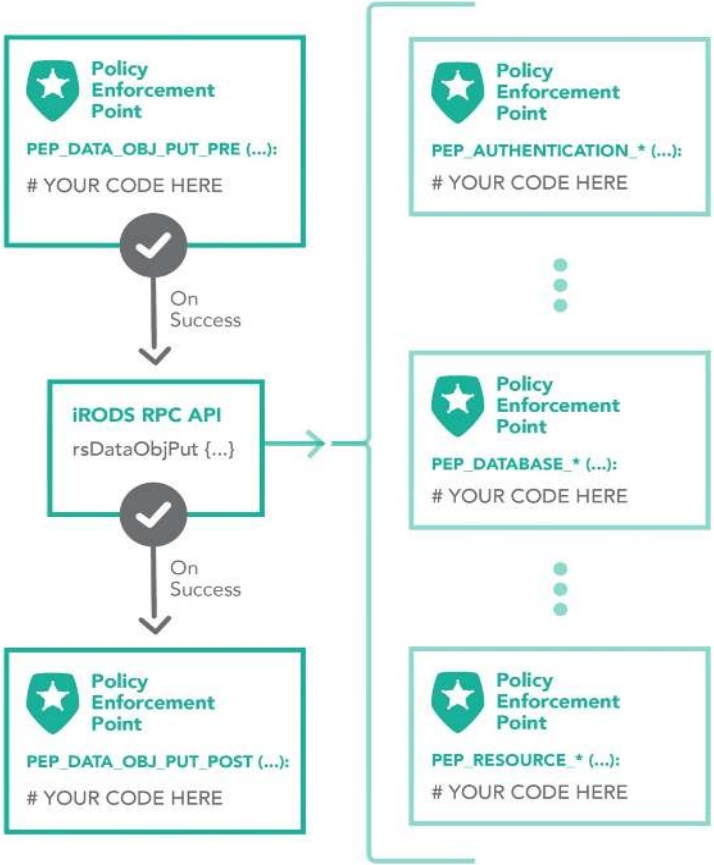
ファイルに付随するメタデータ
POSIXファイルシステムでは
これらが仕様として用意されるが、
ユーザーにとっては全く不十分。

ファイルに関するメタデータは多種多彩
iRODSでは、これらをカタログデータ
として自在に付与でき、RDBで管理。
ユーザーはAPIやコマンドからクエリを
発行し、検索に利用できるだけでなく
iRODSによるデータ処理のトリガーや
タグとしても利活用される。

PEP –Policy Enforcement Point-

- iRODSは処理の途中に独自のコードを差し込むことができる。様々なシチュエーションで処理をフックしデータの処理や移動を記述。配布されているティアリングプラグインも、これで実現したサンプル。
- この再実装可能なポイントをPEPと呼び、全体では数百実装されている。
- ルール記述はPEPに対してイベントドリブンコードを入れたり、ルールエンジンのコードを呼び出すことで行う。
- 独自文法の言語があるが、昨今はプラグインで利用可能なPythonが使われることが多い。APIについては、JavaやC++も対応する。





まとめ

- データの量・サイズ・複雑性・生成速度、全ての属性が拡大傾向にある現在、既存手法の改善と共に、新しいアーキテクチャの開発も続いており、引き続き動向に注視したい。
- 一方で、機能付加により、利便性・パフォーマンス共に大幅な改善を得られる場合があり、要検討。

お問合せはコチラ

Pacific Teck
HPC and Machine Learning Experts

Thank you!