

Scality 大規模ストレージ 運用における 課題とポイント

スキャリティ・ジャパン株式会社
October 26th 2018

会社概要

Scality会社概要

オブジェクトストレージのソフトウェアベンダー

Scality, Inc

- 創業: 2009年 (フランス・パリ)
- 製品出荷: 2010年7月
- サンフランシスコ (HQ), パリ (エンジニアリング)、ワシントンDC、東京
- 非構造化データの基盤として幅広く採用
- 主な市場: クラウドサービスプロバイダ、通信事業者、メディア系企業、大手製造業、政府系機関など

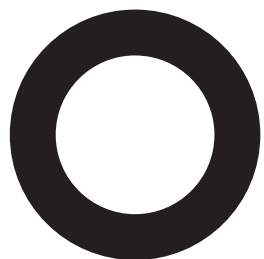
スキャリティ・ジャパン株式会社

- 設立: 2015年3月
- 本社所在地: 東京都千代田区大手町
- 代表: 江尾 浩昌
- 24時間の日本語保守受付体制
- 主な取引先: 通信事業者様、クラウドサービスプロバイダ様、インターネットサービス事業者様、流通事業者様、製造業様



100% SOFTWARE SOLUTION

FOR ANY STANDARD x86 SERVER PLATFORM



RING

PRIVATE CLOUD

- マルチプロトコル (object & file)
- 無制限にパフォーマンスをスケールアウト
- 堅牢なデータ保護メカニズム
- データの自動自己修復機能
- システムの無停止運用



MULTI-CLOUD DATA CONTROLLER

TO ACCESS AND MANAGE DATA ACROSS CLOUDS



ZENKO

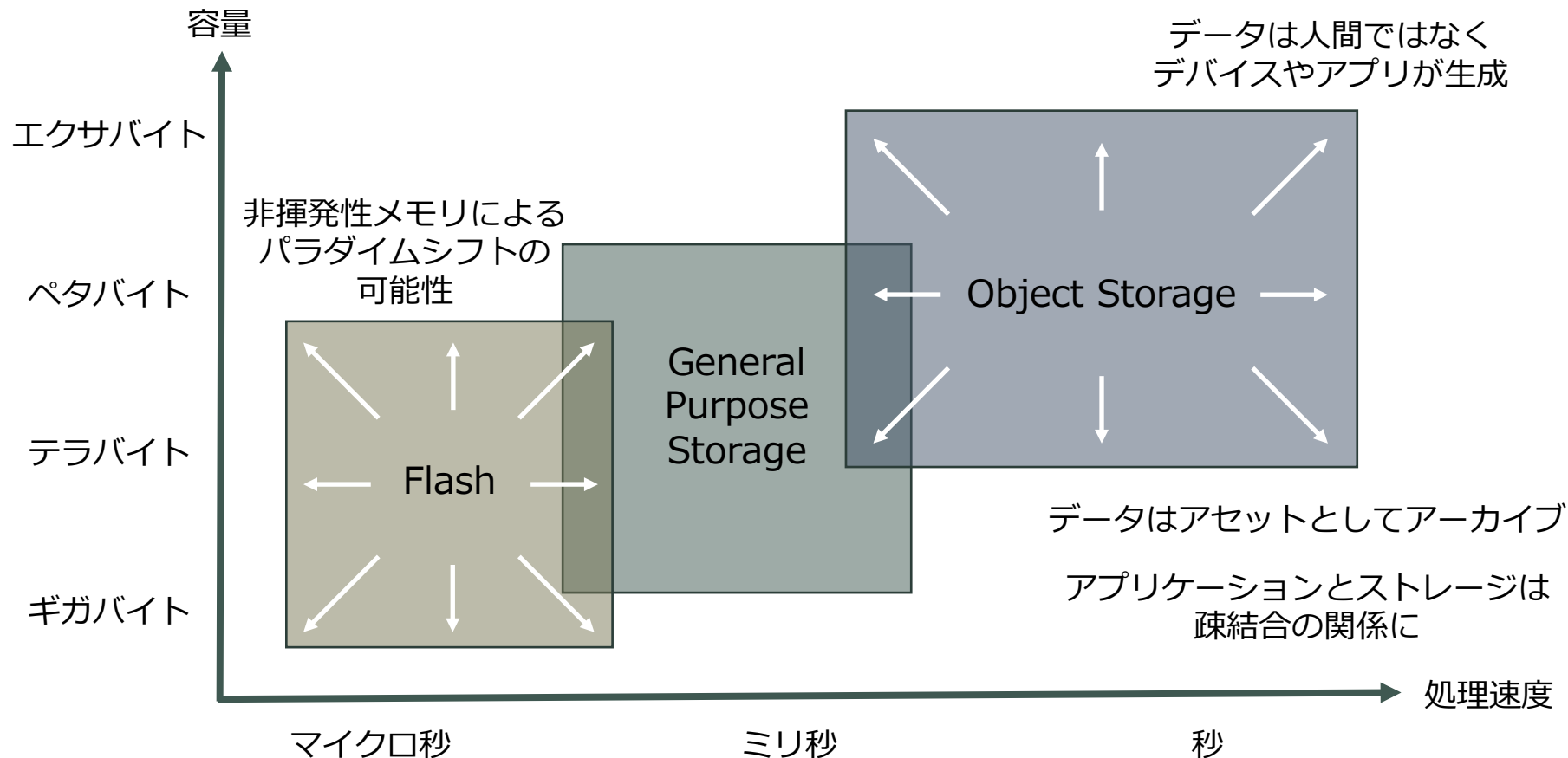
PUBLIC CLOUDS

- マルチクラウドデータ管理ソリューション
- 全てのクラウドに単一APIでアクセス
- データは各クラウドのネイティブフォーマットで保存
- クラウド間での横断検索

なぜオブジェクトストレージ？

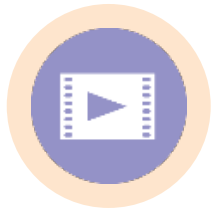
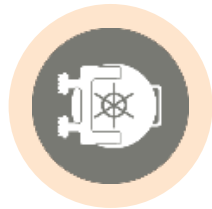
“One fits all” ストレージの終焉

パフォーマンス重視と容量重視の2極化



Scalityのユースケース

Scality RINGの用途

				
コンテンツ配信	アクティブ アーカイブ	アナリティクス	バックアップ	クラウド サービス基盤
• 動画配信基盤 • ビデオアーカイブ • 音楽配信基盤	• 実験、観測データ • 監視カメラ • HPC • メディカルイメージ	• ログ蓄積 • ゲノムデータ • IoT	• バックアップ統合	• クラウドサービス • メールシステム • ハイブリッドクラウド
S3 APIを使用したアプリケーション				

dailymotion

EUROSPORT

DMM.com Labo



hhmi



RENAULT



NATIXIS



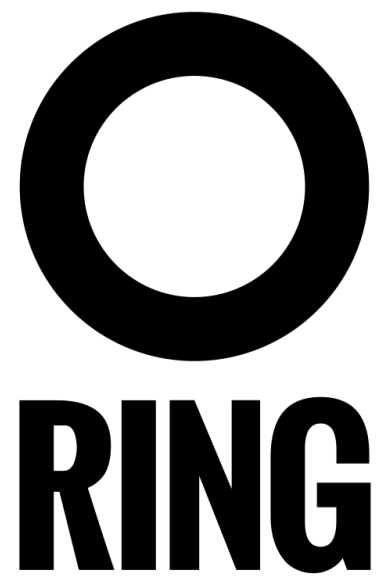
COMCAST

telenet

@nifty

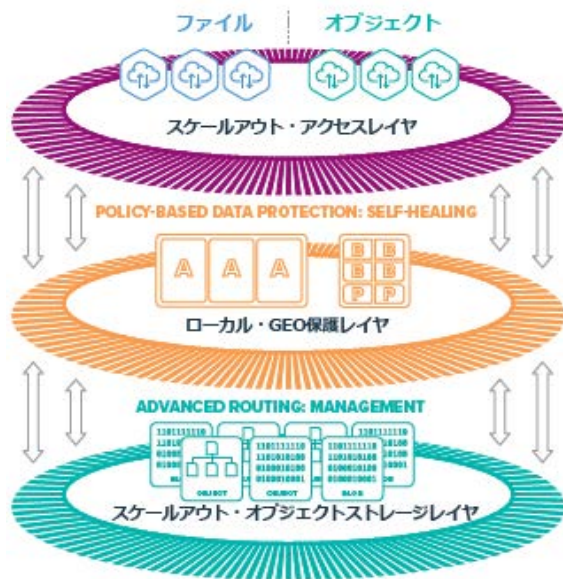
KDDI





Scality RING概要

従来型ストレージアプライアンスの課題を解決するSDS製品



専用のアプライアンス型
ストレージ装置に代えて、
汎用のx86サーバを利用

1. ソフトウェアデファインドストレージ

- ハードウェアを一切選ばないため、各種コストの削減が可能
- 汎用x86サーバを利用することで、費用対効果に優れた大容量ストレージが構成できます。

2. マルチプロトコルにネイティブ対応

- オブジェクトアクセスにもファイルアクセス (NFS, SMB)にも対応し、ゲートウェイ製品を必要とすることなく、様々なアプリケーションに対応可能

3. 高い信頼性と無停止運用

- サーバ間でデータを分散保護し筐体障害にも対応
- 複数のレプリカ(ミラー)保護やイレイジャーコーディングによる4つ以上のパリティ保護
- サーバの増設や交換も無停止で可能なため、EOSLによるデータ移行から解放されます
- 7年以上無停止の海外通信事業者様の事例があります

CHORDについて

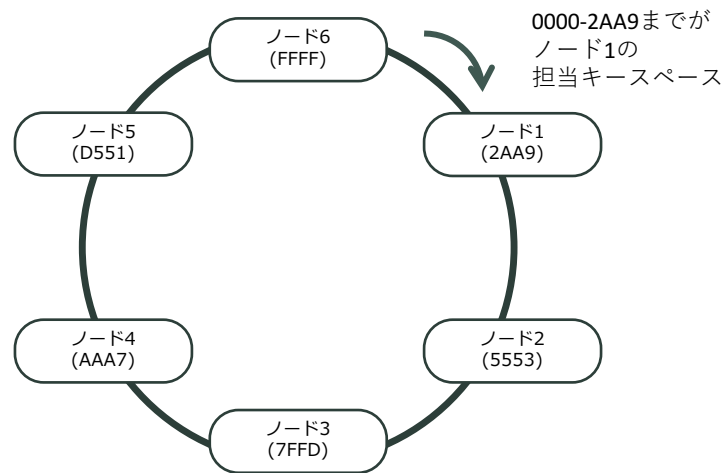
“CHORD” の背景

- オブジェクトストレージは（ディレクトリのような階層構造ではなく）フラットな空間に（何億という）大量のオブジェクトを配置するための仕組みが必要とされる。
- パフォーマンスや容量の観点から、一つのコントローラがこれらを管理するというのは難しく、分散型のストレージという形態が適している。
- CHORDとは分散されたノードにどのようにデータを配置するのかを規定する分散ハッシュテーブルのアルゴリズムの一つ

“CHORD” の特徴

- データはキーとバリューのペアで管理される
- キーはデータを読み出す時に必要となる鍵でハッシュで生成される
バリューはデータの中身
- CHORDではキーは円環状に配置されノードは自分のIDまでのキー
スペースを受け持つ

ノード	キー始点	キー終点	ノードID
ノード1	0000	2AA9	2AA9
ノード2	2AAA	5553	5553
ノード3	5554	7FFD	7FFD
ノード4	7FFE	AAA7	AAA7
ノード5	AAA8	D551	D551
ノード6	D552	FFFF	FFFF

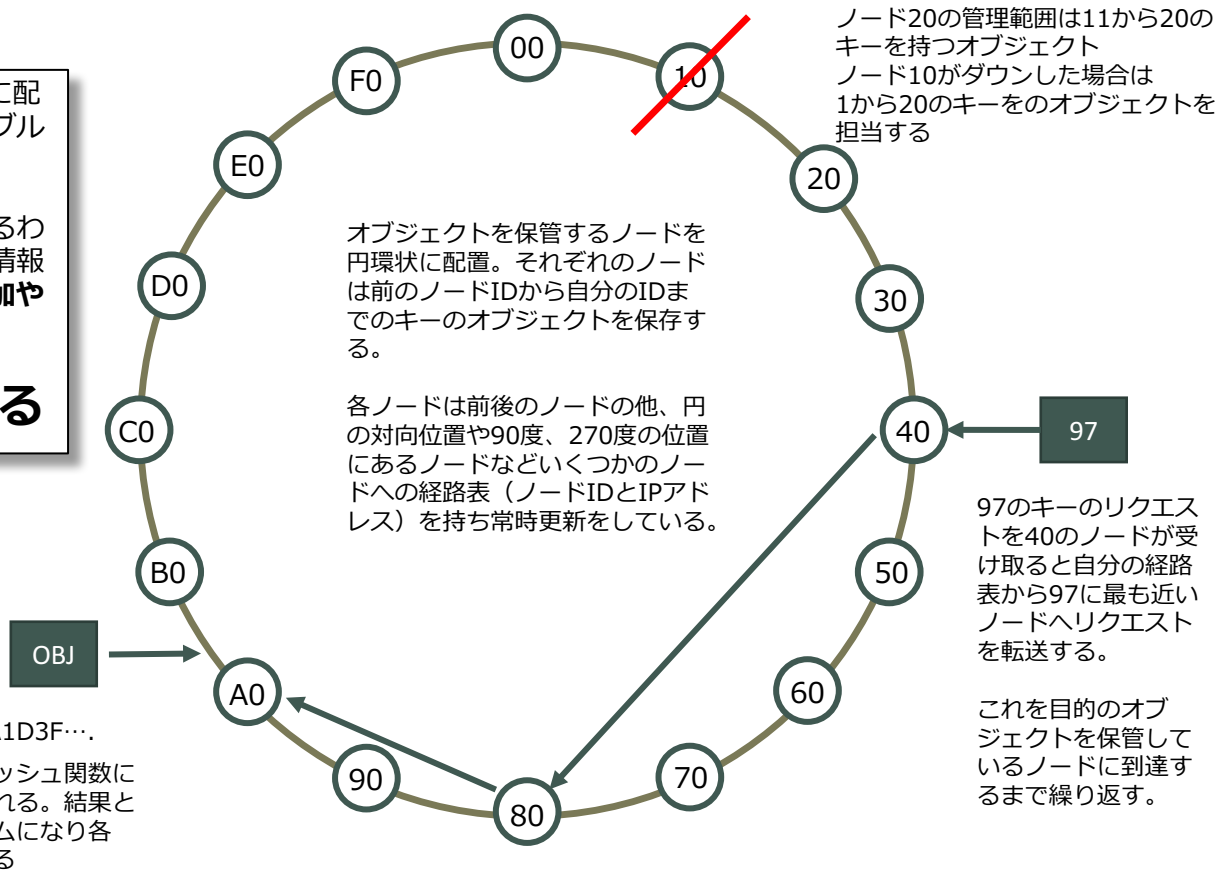


“CHORD” の手続き

CHORDとは、大量のオブジェクトを効率的に配置しアクセスを可能とする分散ハッシュテーブルの方式。

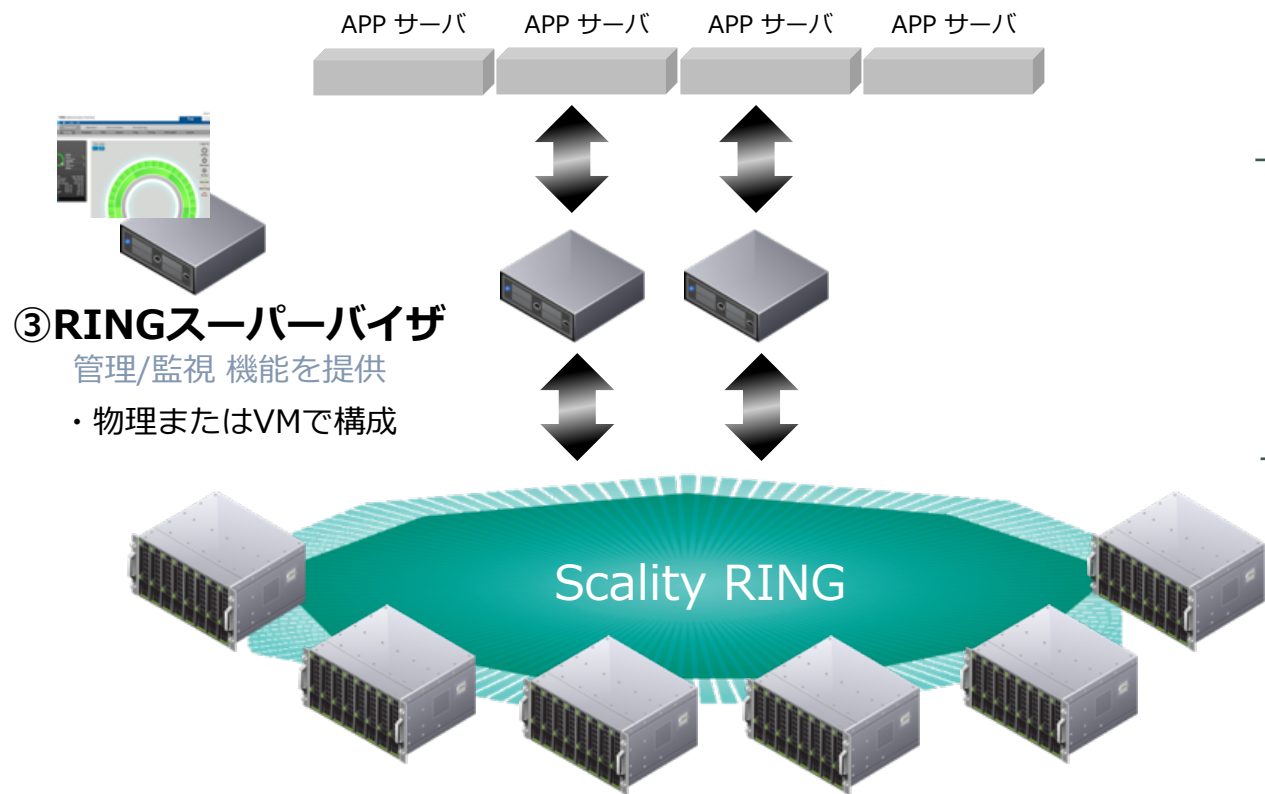
全てのノードとキー組み合わせを管理しているわけではなく、データに到達できる必要十分な情報のみをメンテナンスしているため**ノードの追加や削除に対して柔軟性が高い**。

→**耐障害性や拡張性に優れる**



RINGの実装

Scality RINGの構成



アプリケーション層

②RINGコネクタサーバ

アクセス処理および
書き込み先の演算を担当

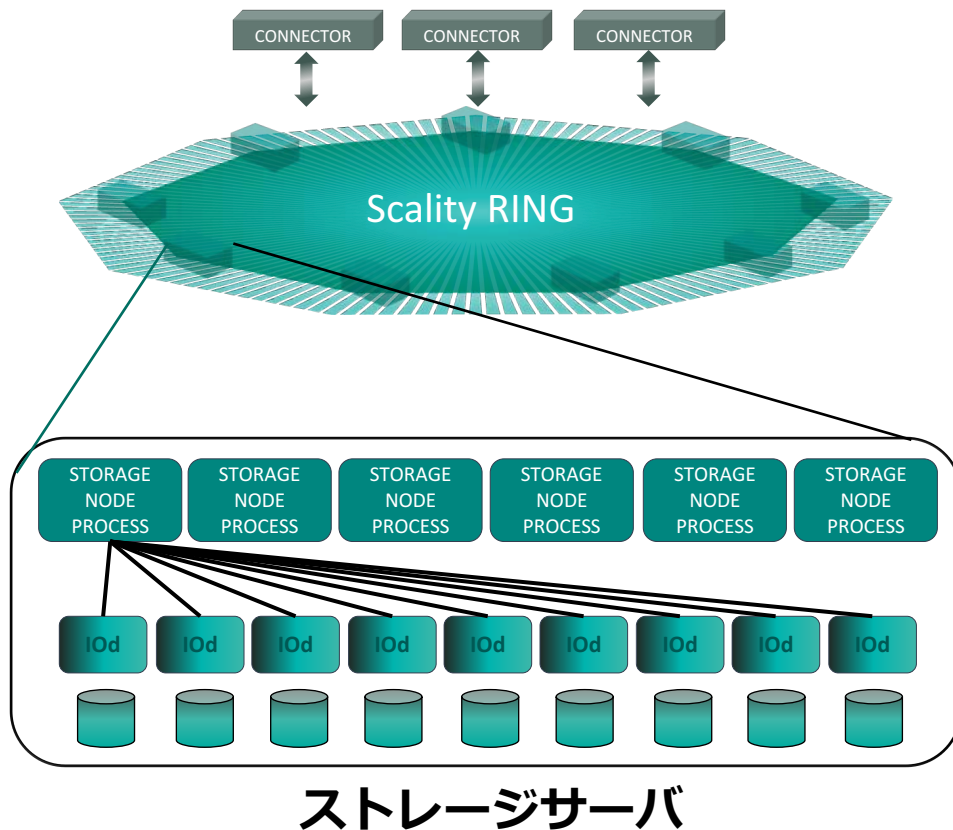
- 性能要件に応じて適切な機種と
数量や配置先を設計
(①への導入も要件次第では可能)

①RINGストレージサーバ

データを分散格納する領域

- 物理サーバ6台以上
- 容量要件や性能要件に応じて適
切な機種の性能と数量をサイジ
ング

ストレージサーバで稼働しているプロセスについて



- **ストレージノードプロセス (Storage Node Process)**

- 他の物理サーバ内の "Storage Node" との間でP2P通信を行う
- 担当する範囲のキーのオブジェクトを書き込み/読み込みするために筐体内のIOdと通信を行う
- 物理サーバ1台につき標準で6つの "Storage Node Process" が動作

- **ディスク管理デーモン (IOd)**

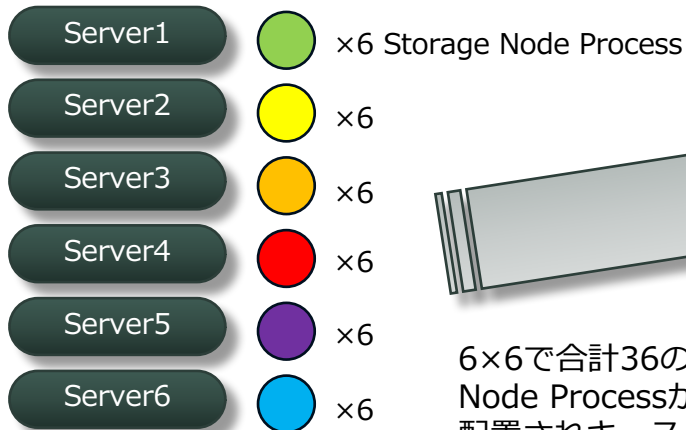
- Disk内のオブジェクトを管理するデーモン
- マウントした "Disk" 1つにつき、一つの "IOd" が起動

ストレージノードプロセスへのアクセス

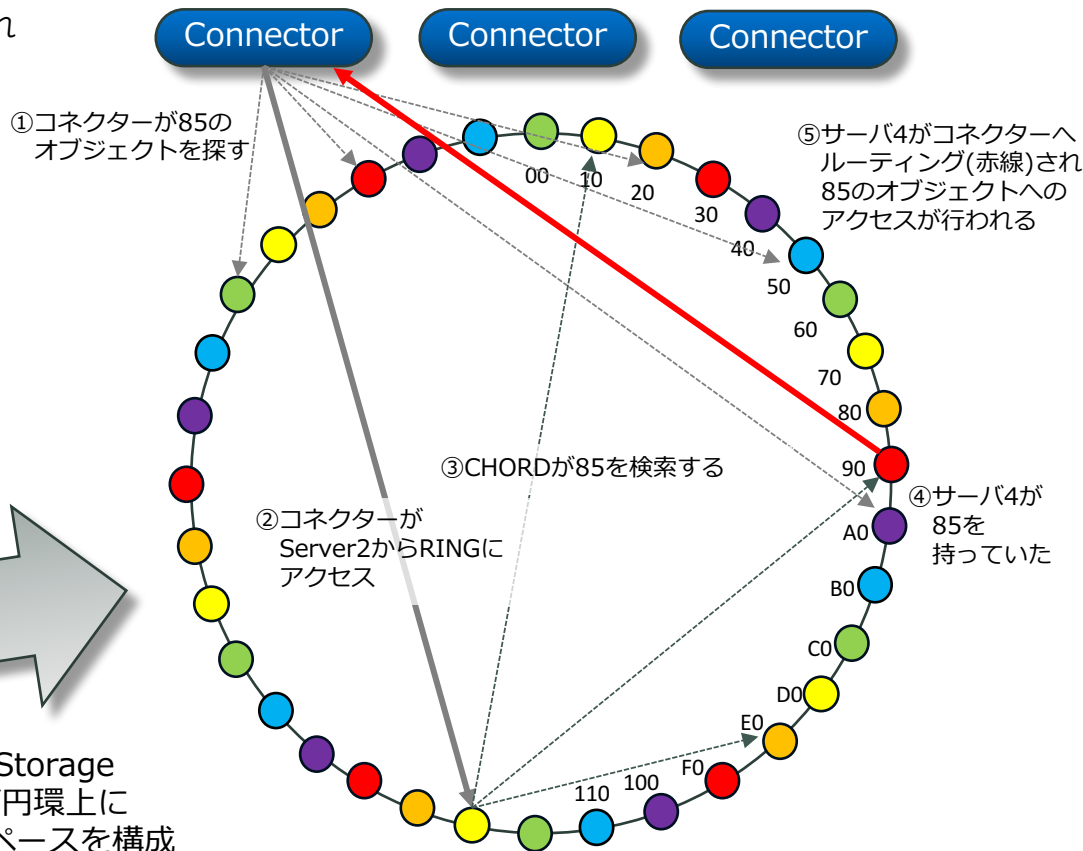
各コネクタはブートストラップリストと呼ばれるエントリーポイントのリストを持っている

- 10.0.13.1:4554
- 10.0.13.2:4554
- 10.0.13.3:4554
- 10.0.13.4:4554
- 10.0.13.5:4554
- 10.0.13.6:4554

RINGストレージサーバ



6x6で合計36のStorage Node Processが円環上に配置されキースペースを構成



RINGのメンテナンスタスク

RINGのメンテナンス・タスク

- サーバの追加などキースペースを変更した時やノードの停止や再起動などでキースペースが自動的に変化した時のタスク
 - **Balance** - オブジェクトを「正しい」ノードへ移動
 - **Proxy** - 「正しい」ノードにまだ移動されていないデータへの操作を中継
- **データの自動修復タスク**
 - **Rebuild** - レプリカの存在をチェックし必要に応じて他サーバにコピー
 - **Repair** - ディスク障害で失われたオブジェクトを他サーバより復元
- **データ削除タスク**
 - **Purge** - オブジェクトの物理削除
 - **Relocation** - ディスクのデフラグ処理

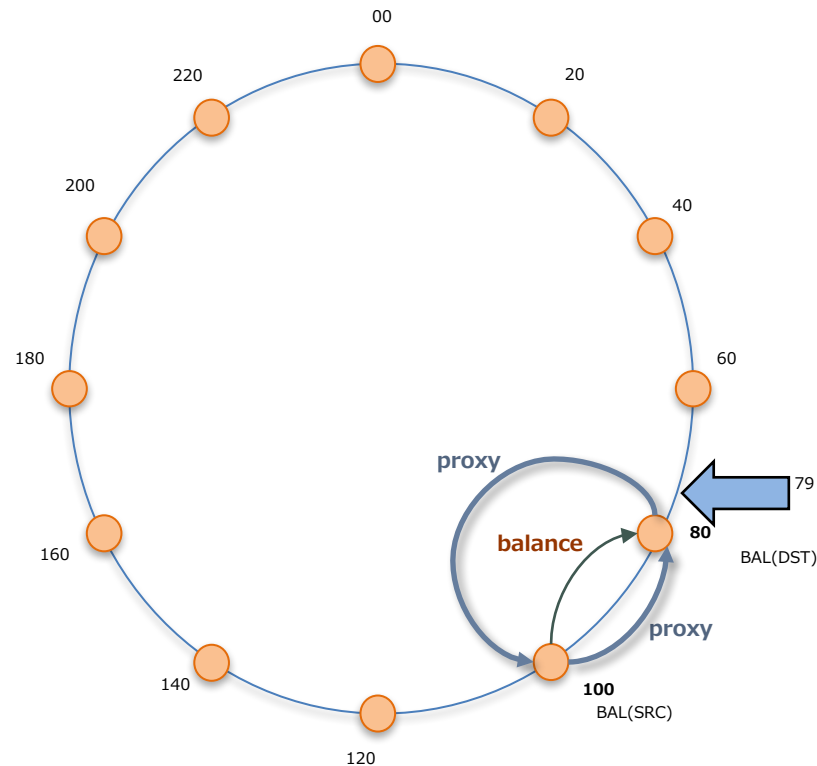
RINGを維持するためのタスク – キースペース変更

- **Balance**

- 各ノードに割り当てられたキー範囲と実際に保持しているのキーの整合性をとるタスク。ノードの追加や一時的に外れていたノードの復旧、キー割当の変更などにより、ノードが範囲外のキーを持つ場合がある。これらのキーを実際に保持すべきノードへ移動するタスク。

- **Proxy**

- Balance 中は、コネクタがあるキーのオブジェクトを保持すべきノードにリクエストをしても、データがまだ移動されていないことがある。この時、実際にデータを持っているノードにリクエストを中継する処理。



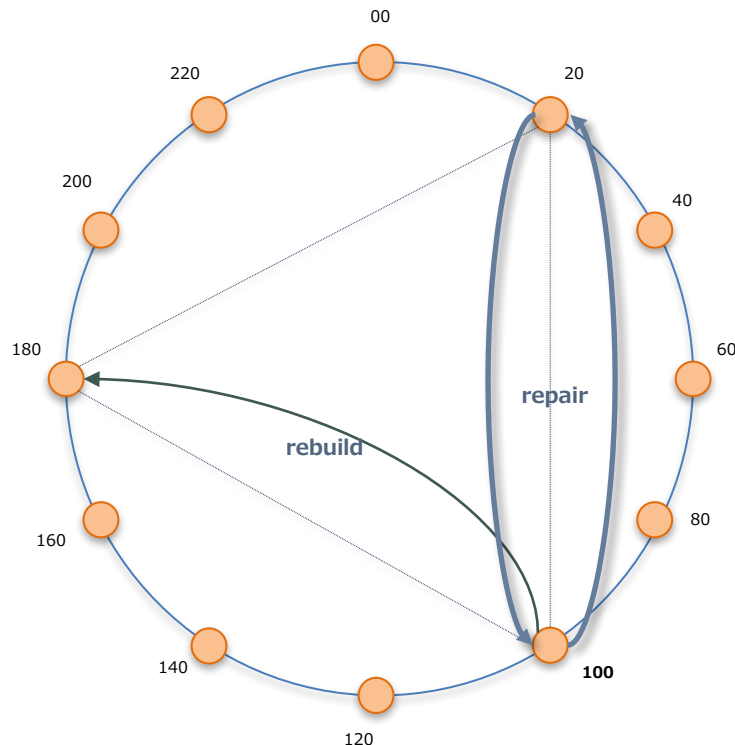
RINGを維持するためのタスク – データの自動修復

- **Rebuild**

- ノードが、自身の管理するキーの次のレプリカが対象ノードに存在し、かつ最新バージョンであることをチェックする。
- 存在しないか旧バージョンの場合、自身のレプリカを対象ノードにコピーする。
- 存在しバージョンが自身のもつレプリカより大きい場合、自身のレプリカを更新する。

- **Repair (Disk repair)**

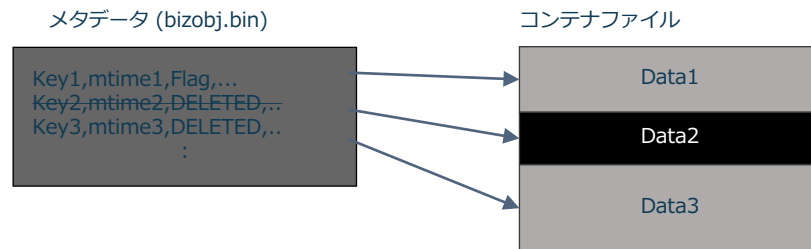
- ディスクが故障した場合、ノードプロセスがメモリ上の情報から失われたオブジェクトの一覧を作成、それを元に前のレプリカからデータを取得する。
- ノードプロセスはIOデモン経由でディスク故障を認識、Repair タスクを実行する。
- Rebuild より高速だが、ノードプロセスを一度停止すると使用できない。



RINGを維持するためのタスク – データ削除

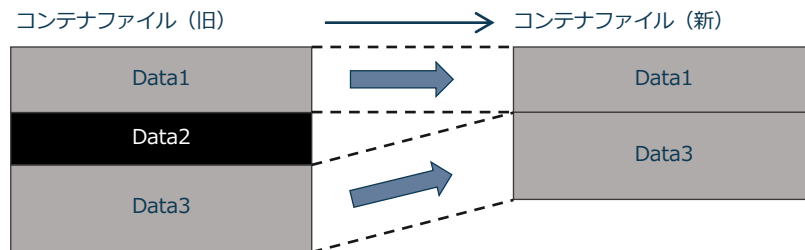
• Purge

- オブジェクトが削除されると、IOデーモンが管理するメタデータのレコードに DELETE フラグが付与される（論理削除）。
- Purge は物理削除を行い、メタデータのレコードを削除する。
- 論理削除の際にはオブジェクトの mtime が更新される。Purge までの保持期間はこの時刻を元に決定される。



• Relocation

- RING では複数のオブジェクトが1つのコンテナファイルに格納されている。Purge によって発生したコンテナファイルのフラグメンテーションを解消し、ファイルシステム上の空き容量を増加させるため処理が Relocation である。
- Relocation は IOデーモンが行う。
- Relocation は大きな I/O 負荷をかけるため、多数のディスクに同時に実行しないよう各ディスクでの実行時間を調整する。



RINGバックグラウンドタスク負荷調整について

- **スループットで制御（relocation以外）**
In/OutをMbps単位で帯域制御が可能
- **同時に実行するタスク数で制御（relocation以外）**
ノード毎にバックグラウンドで動作するタスク数の調整が可能
- **キースペースの割り当てで制御（サーバ増設時のbalance時）**
増設するサーバに割り当てるキースペースを調節する事によって当該サーバへ移動するデータ量を調節
- **時間で制御（relocation）**
開始(再開)時間と中断時間が指定可能
(例) 22:00開始で翌6:00中断

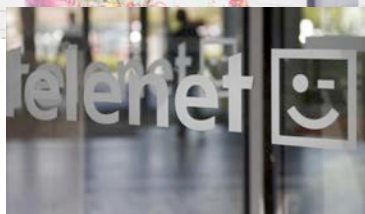
大規模ストレージのポイント

Scality 運用面でポイントとなる事項

1. クリティカルなシステムを支えるストレージとして
 - ✓ 通常運用/障害対応/構成変更のトレードオフ
2. サポート体制
 - ✓ 10名の体制の中でサポートの部隊に5名
 - ✓ アプリケーションやサービスに関するノウハウ

Telenet

ベルギー最大の通信複合企業 キャリアの他にMSP事業とケーブル事業を運営
Scality社の最初のお客様(2010年7月)



- 数百万ユーザー向けのe-mail基盤用ストレージとしてScality RINGを導入
- 同サービスは一般ユーザに加えて法人も利用するため、信頼性が高くスケールアウト可能なストレージ基盤を構築することが課題
- Scality RING導入による効果・ハイライト
 - 高い可用性: 2010年7月のScality RING導入以来、ストレージのダウンタイム無し。100%の可用性を継続中
 - ハードウェア・フリー: 当初導入したRING用のサーバを、システム停止やデータ移行などを行うことなく新世代のサーバにリプレイス済み
 - スケールアウト: 240億オブジェクトをストレッチ構成のRINGに格納

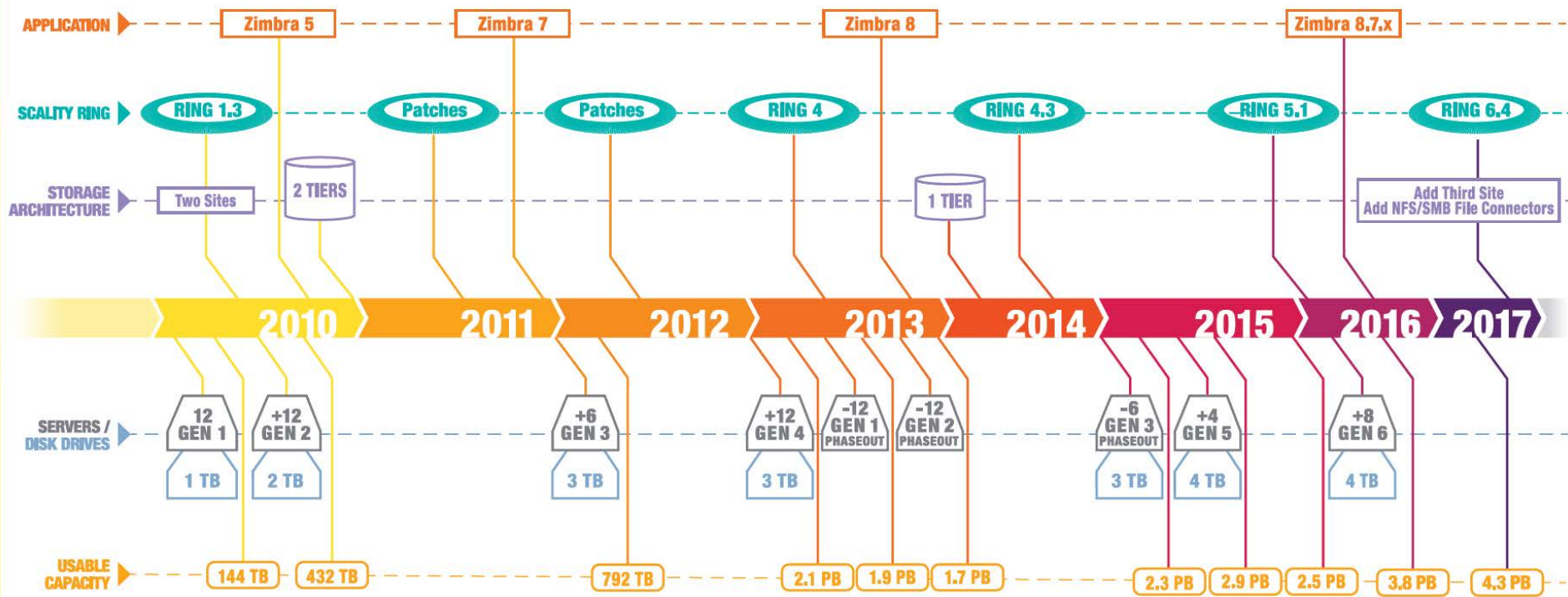
Telenetは、7年間無停止で、無駄なくストレージを運用

GREW 30+X
in Capacity

EVOLVED
Architecture, Application, Infrastructure

GREW 150%
to Over Two Million Happy Users

ALL WITH 100% UPTIME
in Over Six Years with Scalability



Scality SDS Day Tokyo 2018

Scality SDS Day x French Tech Tokyo 2018



名称 : Scality SDS Day x French Tech Tokyo 2018

日時 : 11月22日(木) 10:00 – 19:00 (終了予定)

場所 : 東京ミッドタウン日比谷 6F BASE-Q

主催 : スキャリティ・ジャパン株式会社

共催 : 在日フランス大使館 貿易投資庁 - ビジネスフランス

後援 : 在日フランス商工会議所、マージェンシー

URL : <https://www.scality.com/scality-sds-day-2018-tokyo/>



スキャリティ・ジャパン株式会社
東京都千代田区大手町1-6-1 大手町ビル4F
Email: sales.japan@scality.com | Telephone: 03-4405-5400