オープンクラウド基盤構築・運用のポイント

2012/11/20

NTTCommunications

本日の構成

- パブリッククラウド全体像（大野）
- ObjectStorage設計・運用（大野）
- Cloudstack設計・運用（鈴木）
- PaaS設計・運用（草間）

自己紹介

大野理望
NTTコミュニケーションズ クラウドサービス部

1996年 NTT入社
　GeneralMagicサービス、メールサービスのシステム設計・開発

2001年 NTTPCコミュニケーションズ
　ECサービス、社内システム、メールサービスのソフト開発

2009年 NTTコミュニケーションズ
　BizHostingBasicの企画・設計・開発

2011年 NTTコミュニケーションズ
　Cloud^{n} Computeの設計・開発

2012年 NTTコミュニケーションズ
　Cloud^{n} ObjectStorageの設計・開発

【パブリッククラウド全体像】

☑ オープンクラウドの全体設計を考える

☐ 日米共同開発の苦労

☐ 実際の構築作業のボトルネック

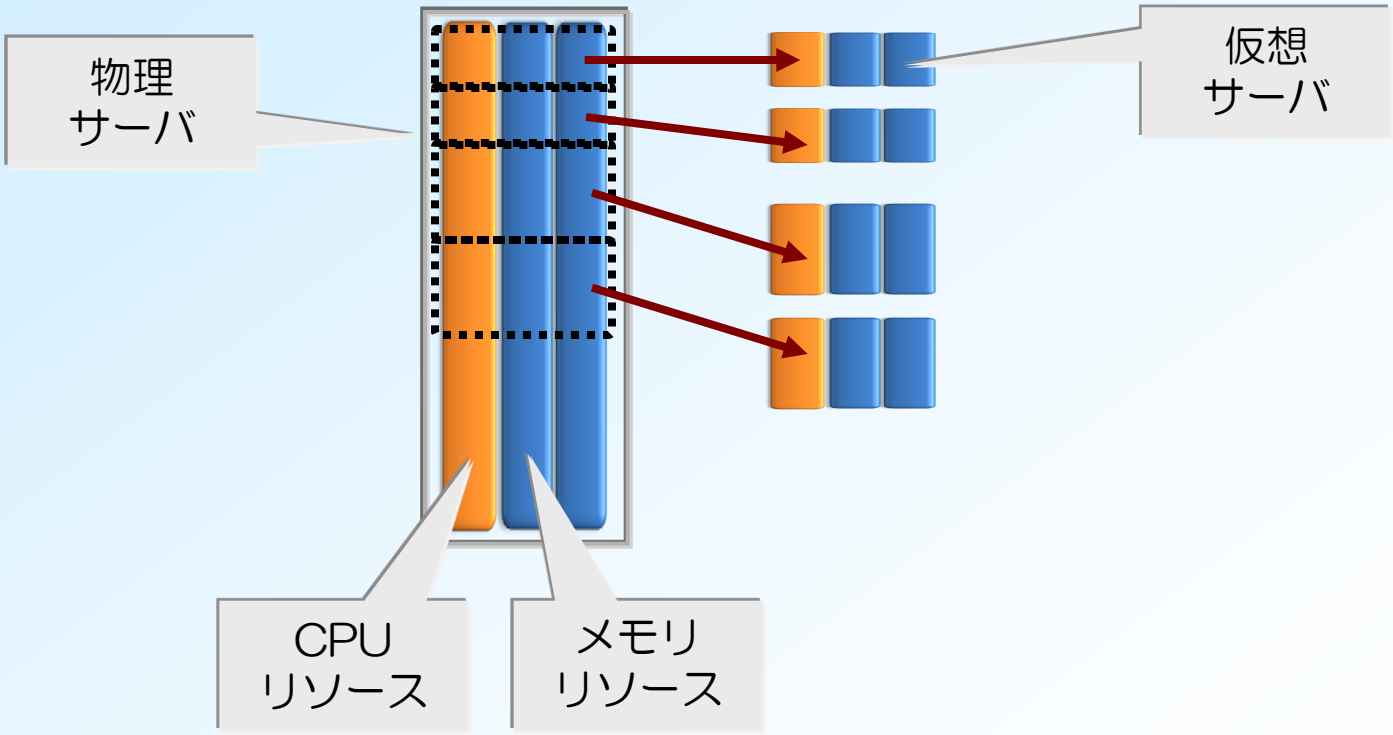
【ObjectStorage設計・運用】

☐ ObjectStorageサービス設計

☐ ObjectStorageの運用

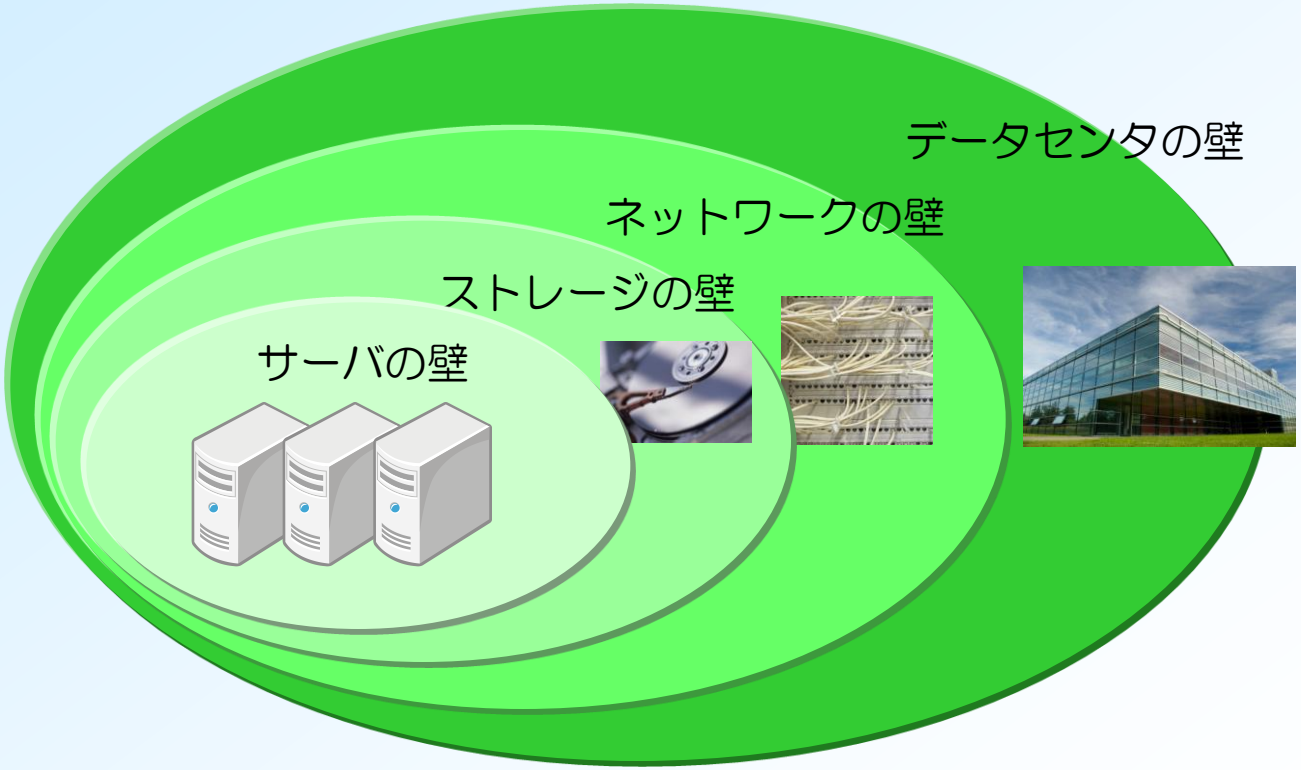
オープンクラウドの全体設計を考える（1）

- オープンクラウドはリソースの小売業。
- 物理サーバを論理的に輪切りにして切り売り。
- メモリにCPUを付けて売っているイメージ。



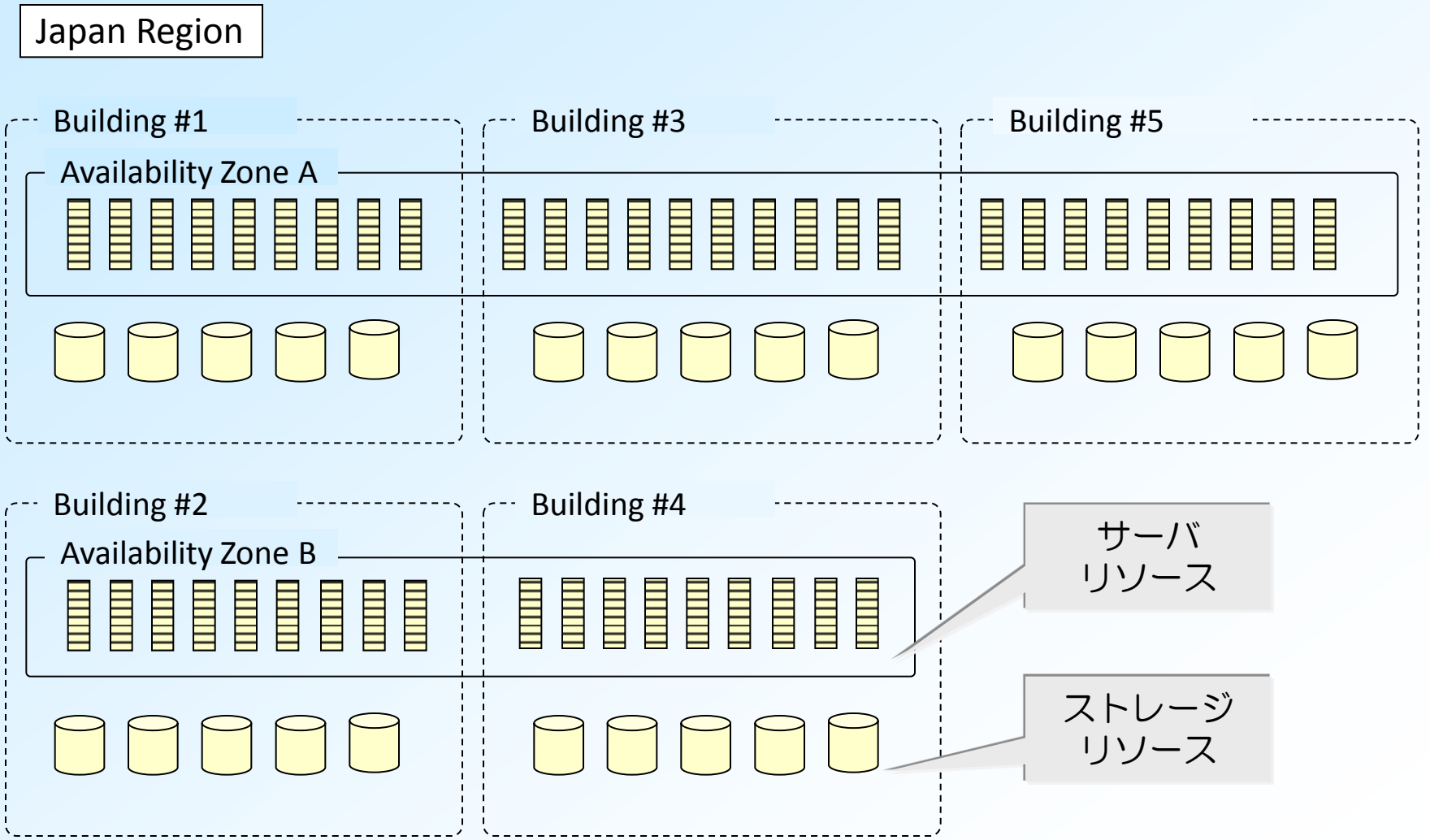
オープンクラウドの全体設計を考える（2）

- 共通リソースの在庫でお客様のバースト需要を叶える。
- 肝となる巨大な共有リソースを作り上げるにはいくつかの壁がある。



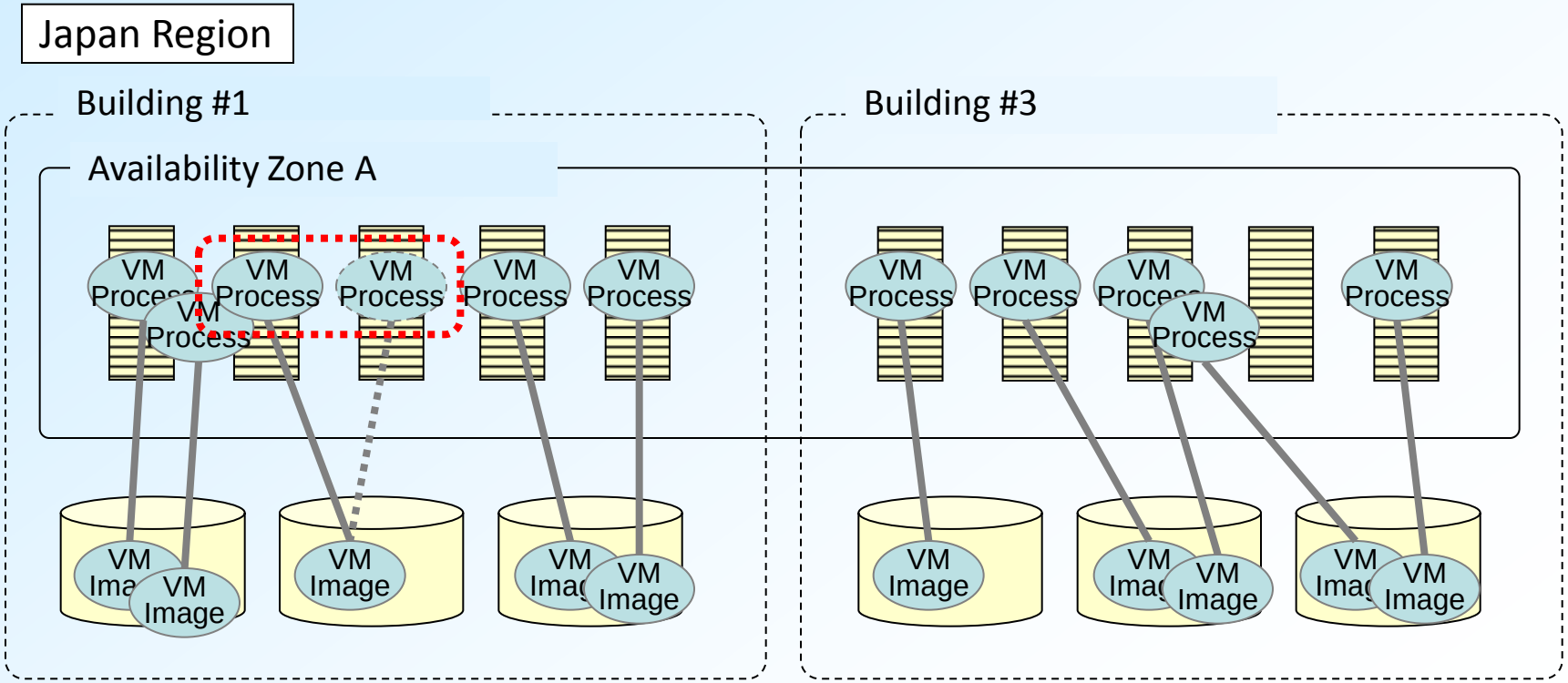
オープンクラウドの全体設計を考える（3）

- もう少し具体的なイメージで考える。



オープンクラウドの全体設計を考える（４）

- VMの管理・移動方法は主に３つ。
  - お客様の再起動タイミングで移動（運用中）
  - ライブマイグレーション（運用中）
  - ブロックマイグレーション（現状未対応）
- 結局ストレージもL2NWも無限にスケールさせるのは現実的でない。
- むしろ一定単位で分断されていてもお客様に見えない仕組みを考えるのが近道。



オープンクラウドの全体設計を考える（5）

現状はまだハードウェアの冗長化に依存している部分が多いが、スケールを考えるとソフトウェアによる冗長化に移行していきたい。

- サーバ
  - Cluster構成（従来）
    - → 別ノードでのVM起動（Compute）
    - → 複数ノードに複製（ObjectStorage）
    - → 別VMでプロセスを起動（PaaS）
- ブロックストレージ
  - 筐体ストレージ（現状）
    - → 分散ストレージ
- ネットワーク、DC
  - 機器冗長化、ロードバランサによる負荷分散（現状）
    - → DNSによる局分散、負荷分散（一部導入）

【パブリッククラウド全体像】

- ☑ オープンクラウドの全体設計を考える
- ☑ 日米共同開発の苦労
- ☐ 実際の構築作業のボトルネック

【ObjectStorage設計・運用】

- ☐ ObjectStorageサービス設計
- ☐ ObjectStorageの運用

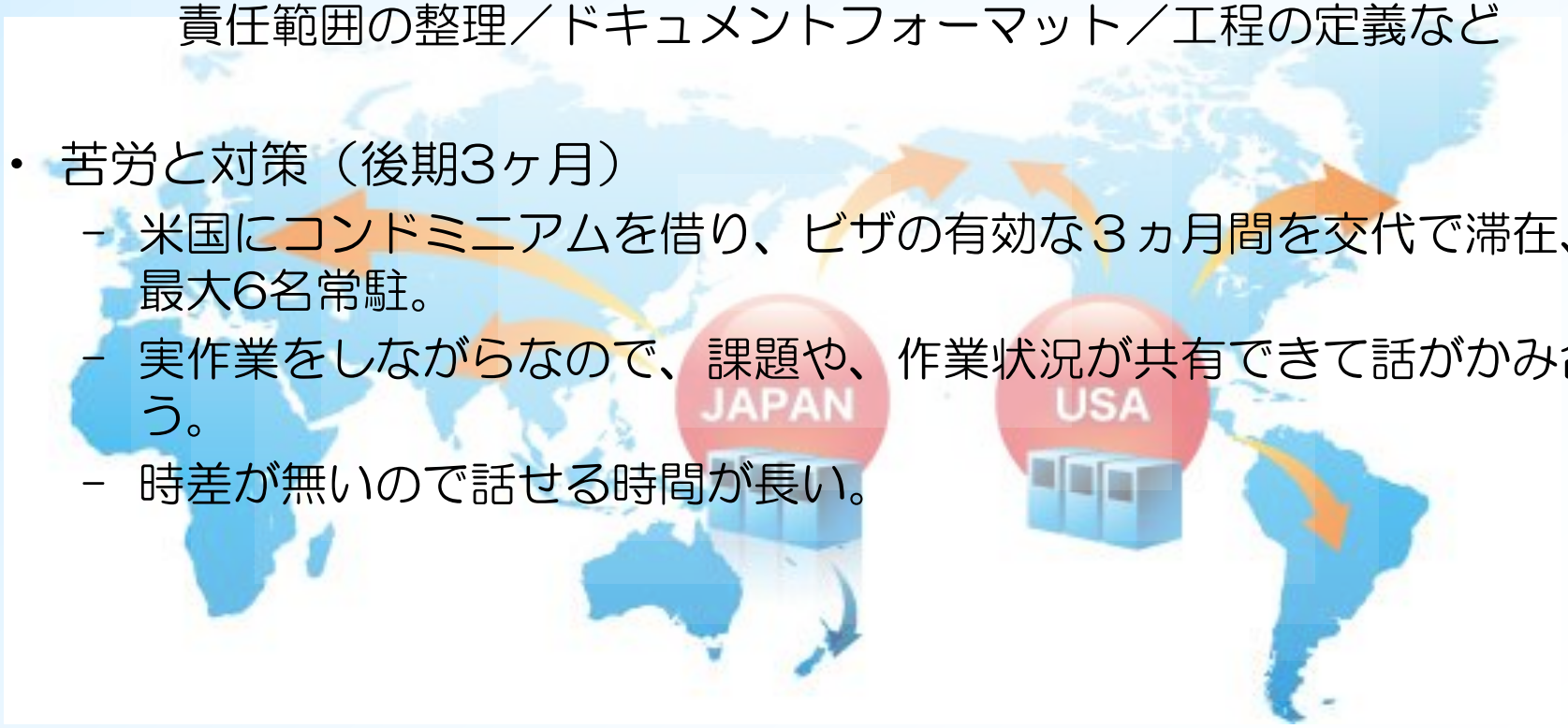
日米共同開発の苦労（１）

- VERIOとの共同開発による世界共通クラウド基盤開発
- 苦労と対策（当初3ヶ月）
  - クラウド開発者と海外業務経験者は少なからず一致しない。
  - 役割分担が異なるのが意思決定のギャップに。
  - 時差の関係で打ち合わせ時間が少ない。

日米共同開発の苦労（２）

- 苦労と対策（中期3ヶ月）
  - 日米のサービス／システムに求める品質の違い。
  - 設計の常識／運用の常識／お客様の常識の違い。
  - PMPが世界共通の実践資格だったと実感。
    責任範囲の整理／ドキュメントフォーマット／工程の定義など

- 苦労と対策（後期3ヶ月）
  - 米国にコンドミニアムを借り、ビザの有効な３ヵ月間を交代で滞在、最大6名常駐。
  - 実作業をしながらなので、課題や、作業状況が共有できて話がかみ合う。
  - 時差が無いので話せる時間が長い。



【パブリッククラウド全体像】

- ☑ オープンクラウドの全体設計を考える
- ☑ 日米共同開発の苦労
- ☑ 実際の構築作業のボトルネック

【ObjectStorage設計・運用】

- ☐ ObjectStorageサービス設計
- ☐ ObjectStorageの運用

実際の構築作業のボトルネック（１）

クラウドだからこそ事業者にとっては低レイヤの悩みが多い。サーバエンジニアが日頃あまり認識しない世界。

- フロアスペースの確保
- 電力の確保
- UPSの確保
- 空調の確保

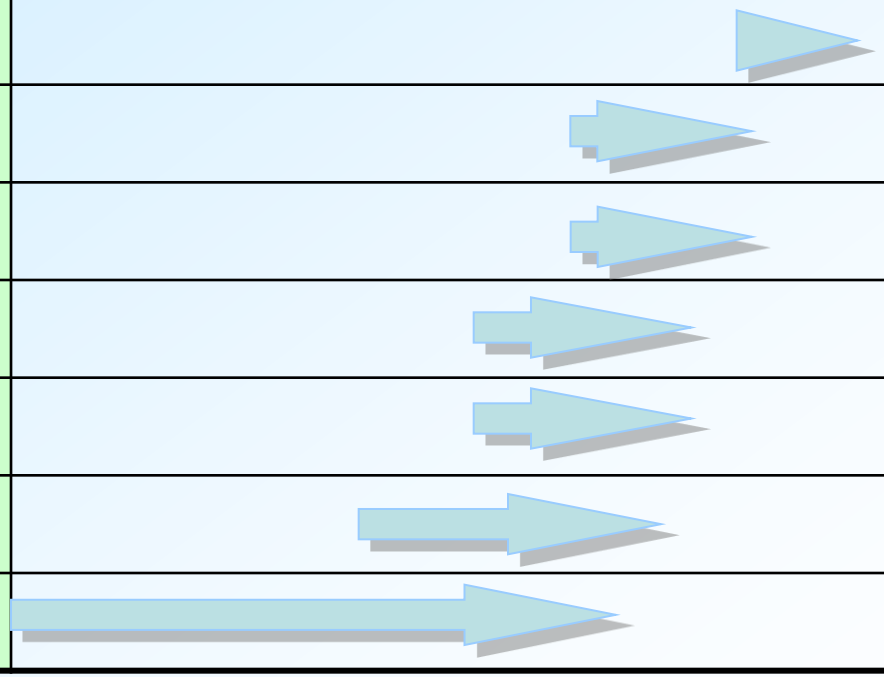
100ラック単位のファシリティの調達は、クラウドのスピードとはかけ離れて半年、１年単位の増設期間が必要。

通信系局舎としてのストックがある中で多少の泳ぎ幅があるのはNTTコムの強み。

実際の構築作業のボトルネック（2）

サーバの調達期間は早くネットワーク機器の調達期間が長い。
サーバ ＜ NW機器 ＜ 回線 ＜ DC・ラック・電源。

		調達期間
設備	サーバ	
	ストレージ	
	スイッチ	
	ロードバランサ	
	ルータ	
回線		
DC、ラック電源		



【パブリッククラウド全体像】

- ☑ オープンクラウドの全体設計を考える
- ☑ 日米共同開発の苦労
- ☑ 実際の構築作業のボトルネック

【ObjectStorage設計・運用】

- ☑ ObjectStorageサービス設計
- ☐ ObjectStorageの運用

ObjectStorageサービス設計（１）

Cloudn ObjectStorageの特徴

１）Amazon互換API

Amazon互換APIを提供

２）データ堅牢性

２拠点に分散し、３データ複製によりハードウェア故障に対するデータ堅牢性としての99.999999999%（11nine）。

３）トラフィック課金フリー

トラフィック課金無料

４）低価格

月額7.4円/GB～

ObjectStorageサービス設計（２）

- ソフトウェアによるマルチDC構成
  ソフトウェアで複数拠点にオブジェクトの複製を生成

- NWとLB
  複数拠点にて両系ActiveのAPIサーバを運用。
  ロードバランサのGSLB機能を組み合わせて両拠点Active－Activeの状態で運用。

- 可用性に対する考え方の違い
  Computeサービス（VMサービス）の場合には、APIサーバ障害でも影響範囲はVMの新規作成・変更に留まりVM自体は稼働しつづけるが、ObjectStorageではサービス自体が利用できないため2拠点分散、拠点内冗長構成を活用して可用性を確保。

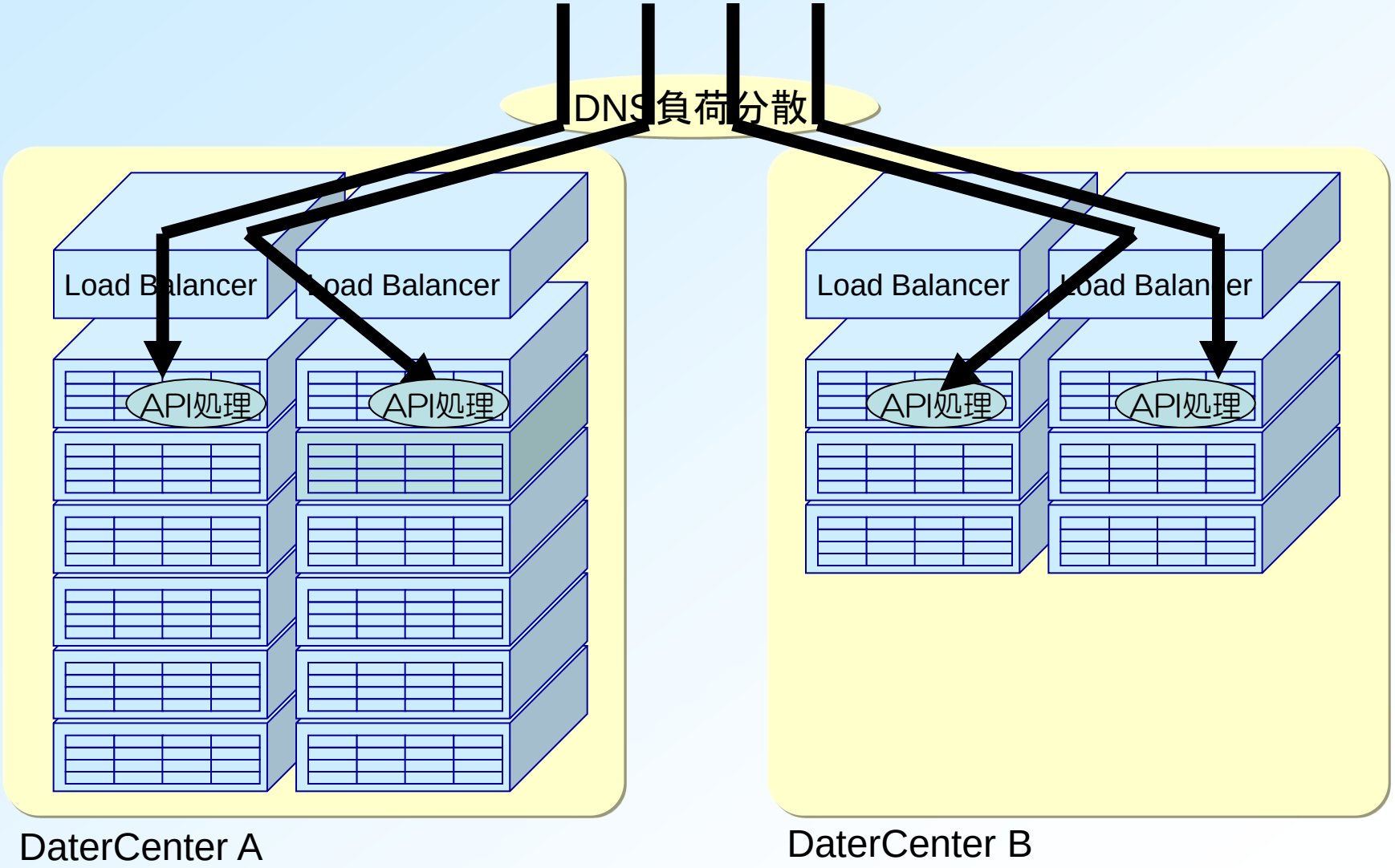
ObjectStorageサービス設計（３）

- ノードが多いためちょっとした作業が大変
  Kickstart、Puppet等の活用
  一元監視による状態の把握

- ログのトレース、ノードでのちょっとしたコマンド確認などはそれでも手間で、現時点では簡単なスクリプトを作成して対応

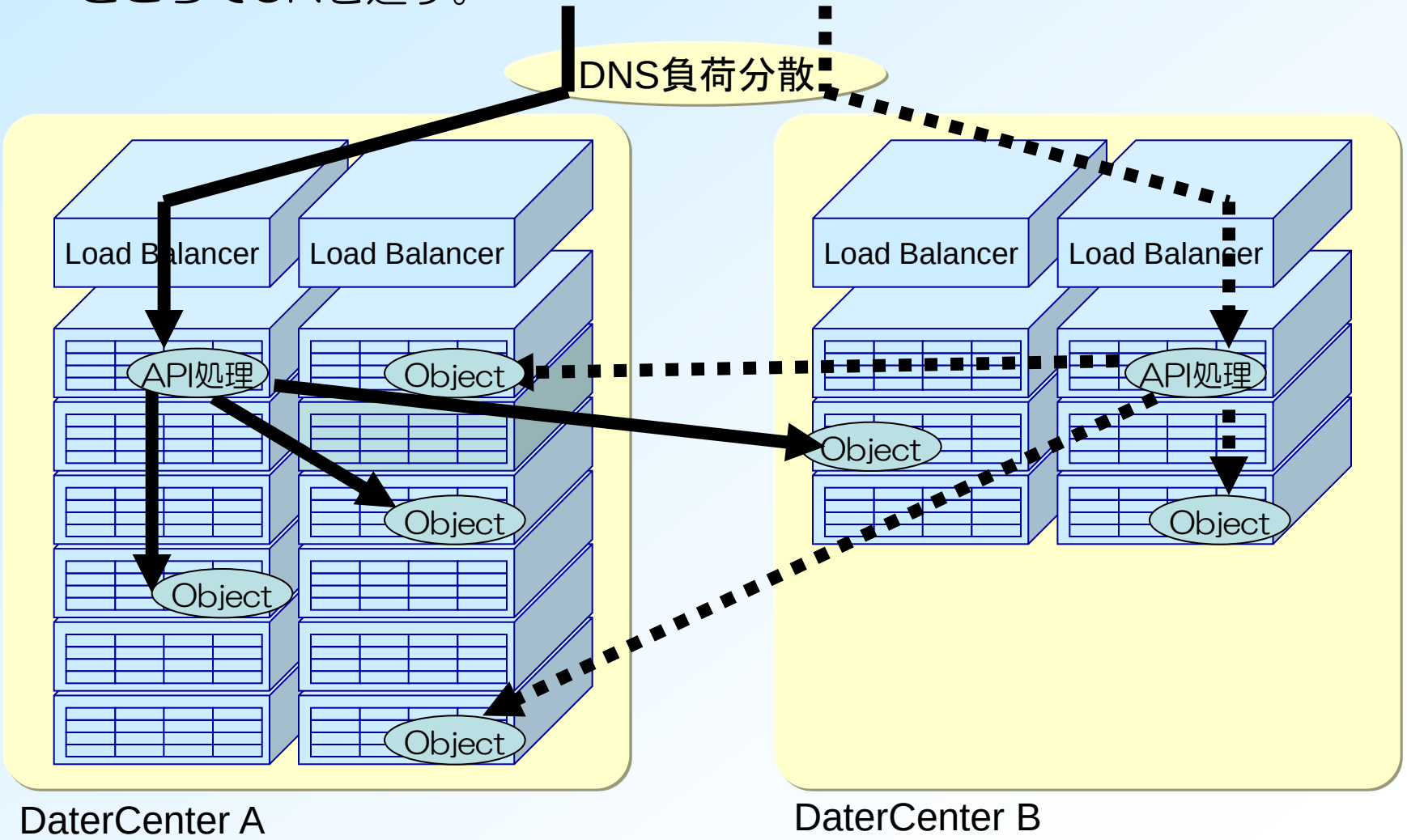
ObjectStorageサービス設計（４）

- DNSとLoadBalancerによる分散



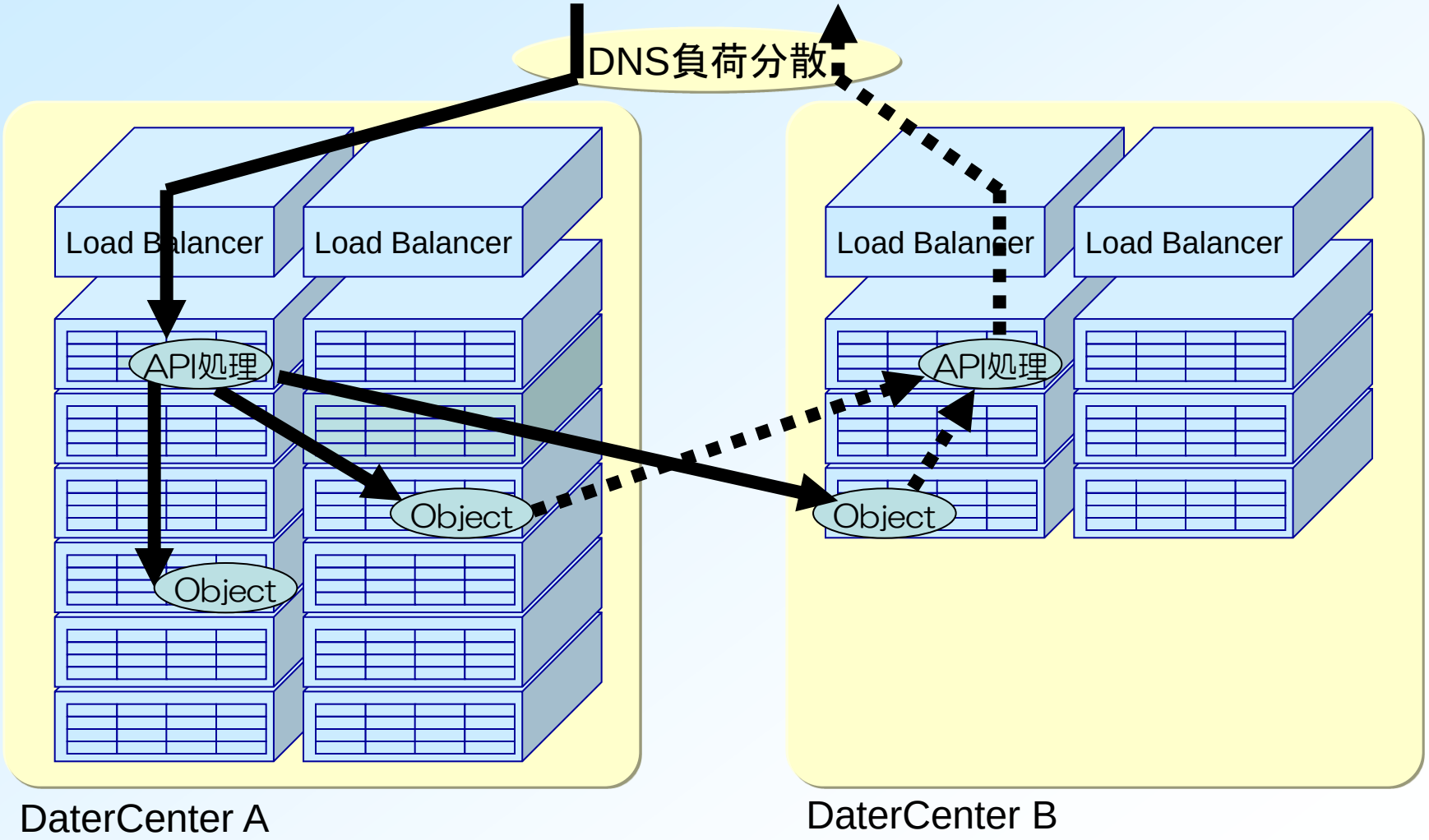
ObjectStorageサービス設計（５）

- DNSとLBによる分散によりAPI処理は負荷分散される
- オブジェクトのPUT処理として３複製されるが、２複製が完了したところでOKを返す。



ObjectStorageサービス設計（６）

- オブジェクトのGET処理については２複製を確認し、最新のオブジェクトを返す。
- 本処理により結果整合性ではなく、即時整合性を保証。



【パブリッククラウド全体像】

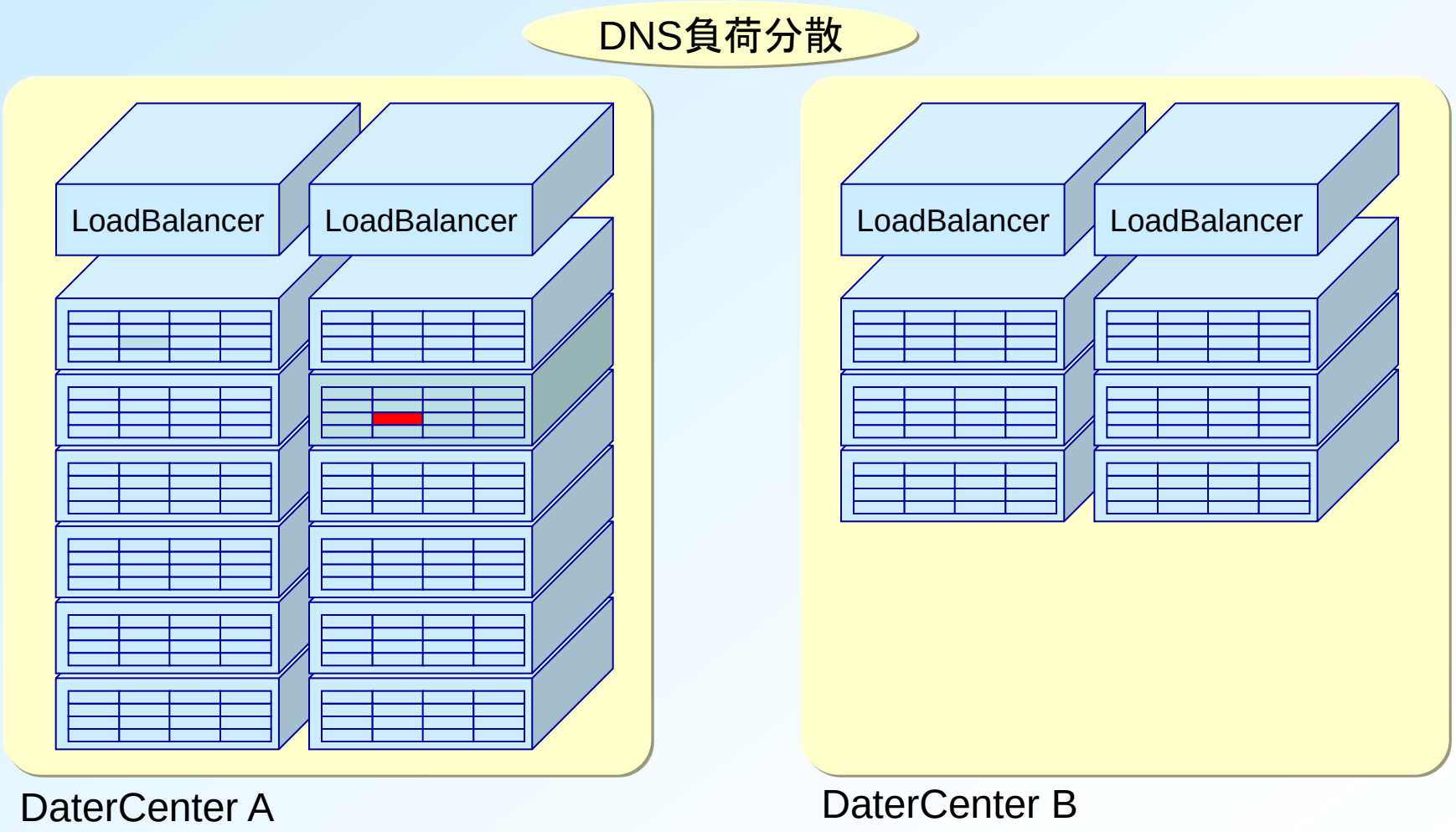
- ☑ オープンクラウドの全体設計を考える
- ☑ 日米共同開発の苦労
- ☑ 実際の構築作業のボトルネック

【ObjectStorage設計・運用】

- ☑ ObjectStorageサービス設計
- ☑ ObjectStorageの運用

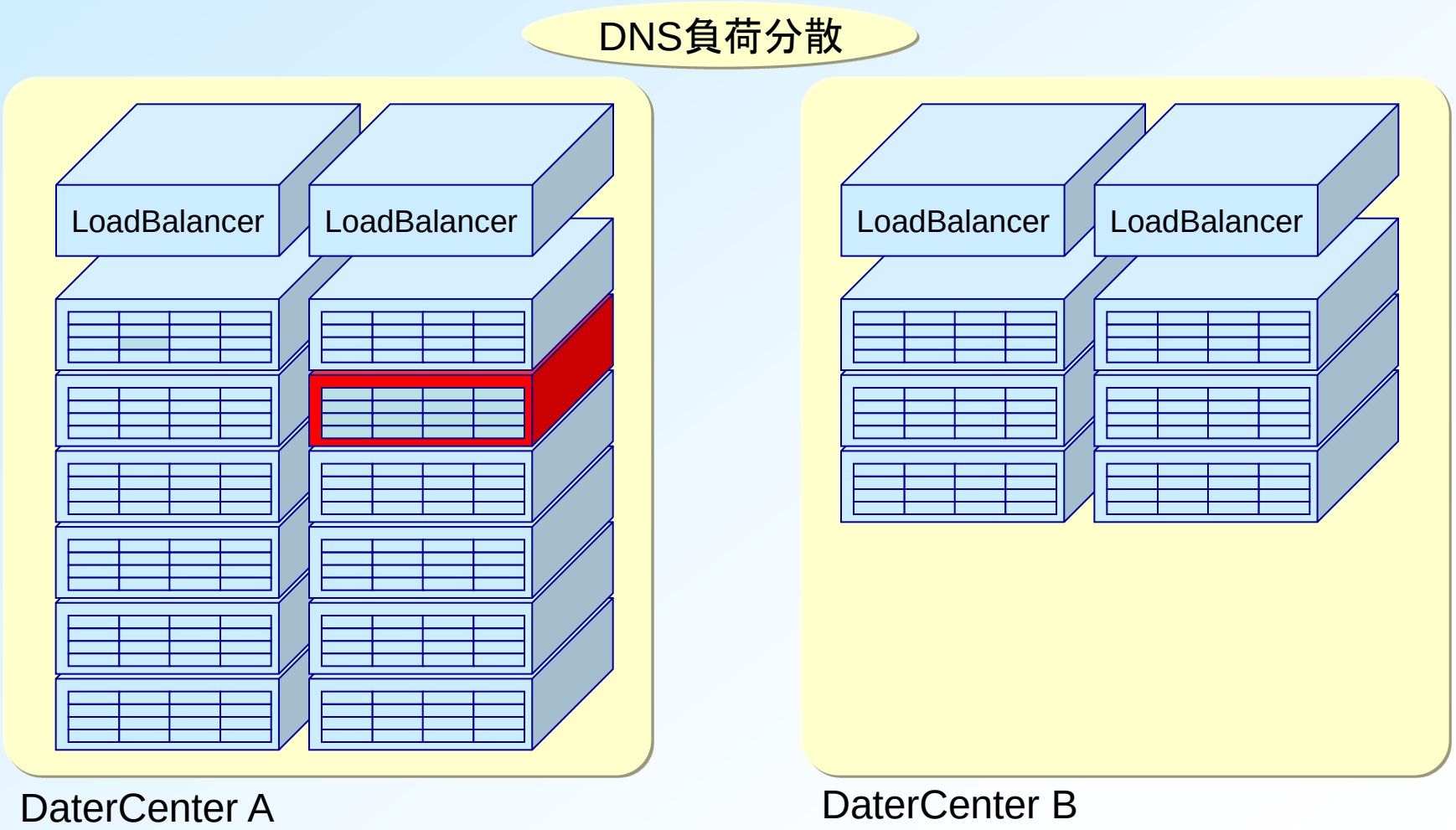
ObjectStorageの運用（１）

１）故障パターン１・・・HDD故障
RAIDにて修復（２TBの修復：数時間）



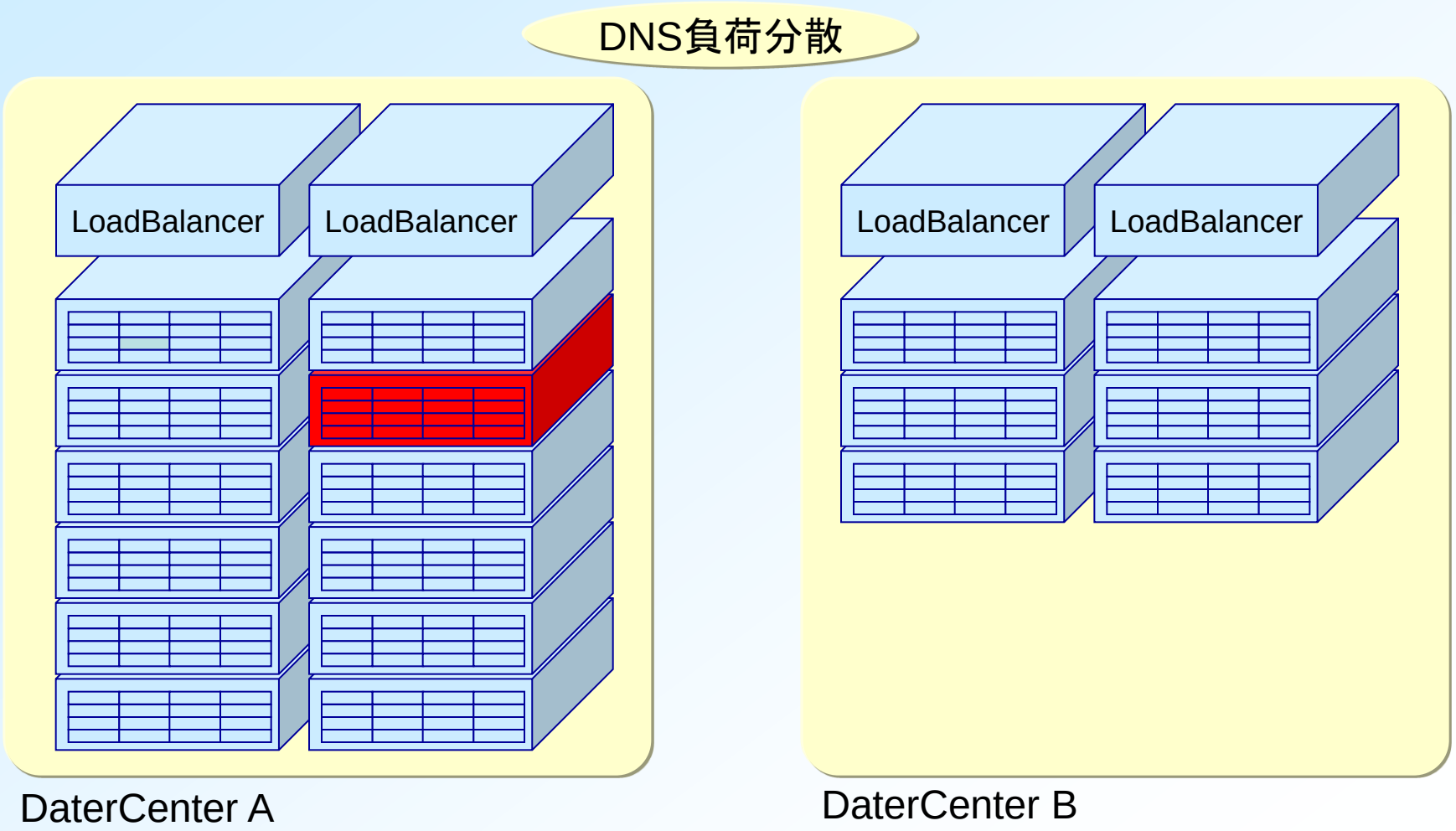
ObjectStorageの運用（2）

2）故障パターン2・・・ノード故障（データ有）
ノード間同期修復（差分データ：数時間）



ObjectStorageの運用（３）

３）故障パターン３・・・ノード故障（データ無）
ノード間同期修復（数＋TB：数日～1週間）



ObjectStorageの運用（４）

- HDD故障に関しては定期的に起きるものなので、淡々と交換
  → 将来的には品質改善や予防交換などもう少しノウハウをためたい

- ノード故障に関しては確率は低いが一旦起きると手動対応と時間が必要
  → 故障対応の自動化やデータの引き継ぎなど今後工夫

- ソフトウェアの品質管理／バージョンアップ
  → 実はここも重要。バージョンアップポリシーなど議論中

・・・ Cloudstack設計・運用について鈴木がご紹介します