オープン系エンタープライズシステム構築

Development of Enterprise System on an Open Platform

石 崎 達 也，滝 沢 和 明

要　約　全日本空輸株式会社の基幹系システムである，国内線旅客システムがオープンプラットフォーム環境にて 2013 年 2 月に稼働した．旧システムはメインフレームの高い可用性と信頼性に支えられ，30 年以上稼働し続けてきた．オープン系技術の台頭から久しく，IT コストの抑制からオープンプラットフォームへの移行が情報系システムを中心に進められてきたが，基幹系システムのオープンプラットフォーム移行には大きなリスクを伴う．

本稿は，今回の国内線旅客システムの移行において，メインフレームと同等の高いサービスレベルをいかに実現したかについて記載する．

Abstract　The Domestic passenger system is a mission-critical system of All Nippon Airways Co., Ltd., and is running on open platform environment since February 2013. The previous system supported by high availability and reliability of the mainframe has continued to operate for more than 30 years. For a long time from the rise of open-system technology, the migration to an open platform has been carried on mainly in information systems due to IT cost reduction, as open platform migration of mission-critical systems is very risky.

This paper describes about how high service levels equivalent to that of the mainframe systems were realized in the open platform environment as a result of Domestic Passengers System migration.

1. は じ め に

全日本空輸株式会社（以下，ANA）の国内線旅客システム“able-D”は，社会公共性の高い，ミッションクリティカルなシステムで，1978 年より米国 Unisys 社製メインフレームで稼働してきた．当該システムはメインフレームの堅牢性に加え，拡張トランザクション処理アーキテクチャ XTPA（eXtended Transaction Processing Architecture）により，複数台のメインフレームを疎結合接続させ，高信頼性と高可用性および高処理性能を提供してきた．従来のメインフレームからオープンプラットフォームへの移行にあたり，大規模な国内旅客系業務アプリケーションの再開発に加え，メインフレームと同等のサービスレベル（高可用性・高信頼性・高処理性能）を実現するインフラ基盤（以下，ANACore）の構築が必要であった．

本稿では，インフラプロダクトセット（OpenXTPA）の机上検討から本番稼働までの開発計画とシステム構成定義上の主な考慮点について記載する．

2.　システム構成の前提条件

システム構成の主な前提条件を本章に記載する．

2. 1　システム化要件

本番環境は以下のシステム構成要件を満たすシステム構成とした．

1） サーバ処理性能：xxx 件/秒
2） レスポンスタイム：空港設置端末 n 秒以内，一般端末 m 秒以内
3） サービス提供率：99.xxx％（サービス提供時間中（計画停止を除く））
4） 復旧時間：n 分以内（障害検知から暫定復旧までの最大時間）

なお，上記の正確な値は諸般の事情により非公開とする．

2.2 システム構成方針

2.2.1 信頼性

1） システム全体

システムを構成する機器（サーバ，ストレージ，ネットワーク機器，FC スイッチ等）は，システム障害につながる SPOF（Single Point Of Failure）となる部位が発生しないよう二重化以上の多重化を基本とする．多重度や多重化方式（Active-Active または Active-Standby）については各機器の特性や保守性を考慮して決定するものとする．

障害が発生した場合のフェイルオーバー処理は，障害復旧時間に完了する構成を前提とする．

外部ストレージについては，通常，外部ストレージ筐体内において基本部位（電源，ディスク，コントローラ等）は全て多重化されているが，本システムにおいては業務データベースを保存するストレージ筐体についても，筐体二重化構成を実装し，信頼性を更に高める構成とする．

2） ハードウェア単体

ハードウェア単体（サーバ，ネットワーク機器，ディスクストレージ等）の構成について，信頼性を高めるために，以下の方針で冗長化などの対応を行う．

ⅰ）ハードディスクは RAID1（もしくは RAID1+0），RAID5[*1] により冗長構成を基本とする．サーバ OS，業務データ格納エリアなど本番業務性能に影響を与える箇所は RAID1 を採用し，外部ストレージ副ボリューム（バックアップ用）など業務に直接影響を与えない箇所は RAID5 による構成とする

ⅱ）電源，ネットワーク（NIC），ファイバチャネル（FC）は二重化を基本とする

ⅲ）メモリ単体での冗長構成（オンラインスペア，ミラーリングメモリ，メモリ RAID 等）の考慮は行わない

2.2.2 可用性

1） システム全体

本システムの可用性（サービス提供率）目標値は 99.xxx％となっているが，システム化要件に挙げられている通り，アプリケーションプログラムを含めたソフトウェア定期リリースなどの計画停止時間は対象に含まない．また，性能縮退が発生している場合も，サービスが何らかの形で継続されている場合はサービス提供状態と見なす．システム全面停止時のみサービスが提供できない状態と定義する．

サービスを提供するには，システムを構成するすべてのハードウェア（サーバ，ネットワーク機器，ストレージ等）とソフトウェア（OS，ミドルウェア，アプリケーション）が少なくとも 1 セット以上正常に稼働していることが必要となる．

システム構成定義にあたる可用性の考え方としては，ハードウェアレベルでの可用性と多重度（冗長化）を対象としてMTBF（Mean Time Between Failure：平均故障間隔）とMTTR（Mean Time To Repair：平均修理時間）から算出される稼働率（図1）をベースとして問題がないかどうかを確認する．ANACoreシステムを構成するサーバ，ネットワークなど機能別単位は並列型モデルとし，システム全体では機能別の構成が全て直列型に接続されているモデルとして評価する．

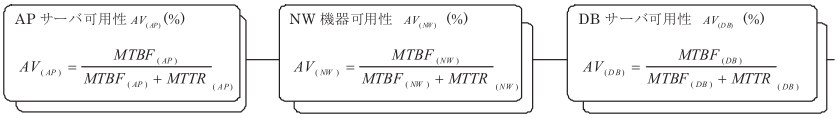


図1　可用性の考え方

2）　ハードウェア単体

ハードウェアは，各ベンダーが提供するエンタープライズクラスで可用性が高く実績のある機器を選択する．

2.2.3　拡張性

各機器における拡張性は，その用途・部位によって1)スケールアウト方式，2)スケールアップ方式のいずれか，または3)両方式の併用の3種類で実現する．それぞれの方式で以下のように拡張性を確保する．

1）　スケールアップ方式：対象となる機器の構成要素（CPU，メモリ，ディスク）は当初搭載量と同等もしくはそれ以上の余裕を持って増設可能であること

2）　スケールアウト方式：対象となる機器（サーバ等）の増設に対応可能な構成とすること

3.　開発計画

開発計画は大きく，Step1（事前検証），Step2（実質開発）の二フェーズから構成され，Step2の実開発に先立ち，Step1は全体計画の精度向上とリスク検証を目的に実施した．図2に開発計画全体のマイルストーンを示す．以降インフラ基盤関連の計画について記載する．

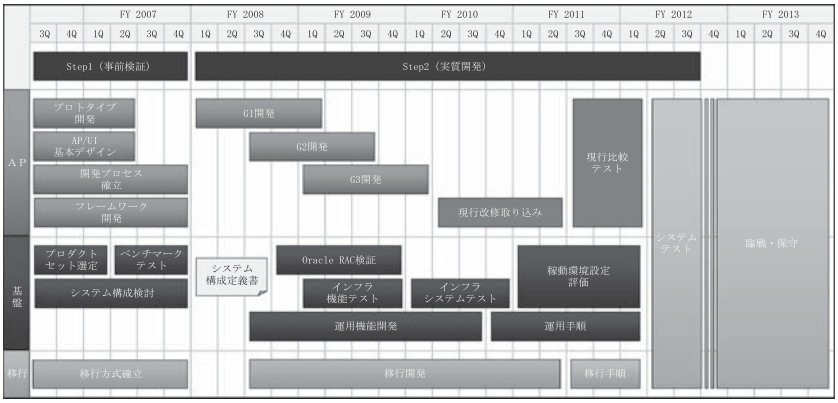


図2　開発計画

オープンプラットフォームへの移行にあたり，日本ユニシスグループおよびベンダーコンソーシアム（日本ヒューレット・パッカード株式会社，日本オラクル株式会社，株式会社日立製作所との協力体制）により，オープン系プロダクトの机上比較検討から稼働および稼働後の運用保守に至るまで，図2に示す入念な検証プロセスを経てメインフレームと同等のサービスレベルを実現することとした．

インフラ基盤関連の主要タスクとその作業概要を以下に示す．

1） プロダクトセット選定とベンチマークテスト

プロダクトセット選定にあたり，エンタープライズシステム[*2]要件およびミッションクリティカル要件を明確にし，以下の観点で調査・検討・評価を実施した．選定結果は4章に記載する．

ⅰ）機能性

RASIS[*3]の観点を中心にプラットフォーム基盤（OS）およびミドルウェアの検討を実施するとともに，ベンチマークテストを通じて可用性・信頼性検証を実施した．

ⅱ）市場性

各プロダクトについての動向，シェア等に関して調査を実施した．

ⅲ）性能（ベンチマークテスト）

主要機能のプロトタイプ開発を実施し，アプリケーション稼働状態で，現行システムのトランザクションミックスモデルを用いてベンチマーク計測を実施した．ベンチマーク計測は，プロダクトセット候補の複数のプラットフォーム基盤について実施した．

2） システム構成検討

プロダクトセット選定とベンチマークテスト結果およびシステム化要件より，機器構成算定[*4]を実施し，システム構成を検討した（システム構成定義書）．

3） Oracle RAC（Real Application Cluster）検証

able-Dのサービスレベル維持を目標とした性能・可用性・信頼性の評価テストを，プロダクトセット選定およびシステム構成検討で選定した構成（HP Superdome，Oracle RAC）にて実施することにより，プロダクト系不具合を事前に検出し，プロダクトリスクを排除した．

4） 運用機能開発

インフラ基盤上に以下の九つの運用機能を開発した．

ⅰ）監視運用機能：システムに異常が発生したことを迅速に検知する機能

ⅱ）障害時運用機能：システム障害が発生した際，迅速にリカバリ処理を実施する機能

ⅲ）オンライン入力制御機能：オンライン入力の制御機能

ⅳ）リリース運用機能：アプリケーションを含めたソフトウェアのリリース機能

ⅴ）キャパシティ運用機能：システムリソース状況を逐次・定期的に管理できる機能

ⅵ）オペレーション運用機能：システム操作の一元化や省力化を図る機能

ⅶ）ログ運用機能：出力されたログを一元管理する機能

ⅷ）バックアップ運用機能：データのバックアップ/リストアを簡易に行う機能

ⅸ）セキュリティ運用機能：規定されたガイドライン準拠のセキュリティ対策機能

5）インフラ機能テスト

運用機能プログラム，機器設定・ソフトウェア設定が連携し，各運用機能が設計通りに正しく動作することを確認した．

6）インフラシステムテスト

システムテスト開始前のインフラ基盤完成を目的に，設計開発したインフラ基盤が設計通り正しく動作することを確認した．基盤運用テスト，障害リカバリテスト，基盤系システム接続テスト，FTP 接続テストの各テストテーマについて実施した．

7）システムテスト

本番稼働に向けた最終工程として，業務サイクルテスト，パフォーマンステスト，ロングランテスト，運用テストなどを実施し稼働品質を確保した．パフォーマンステストについては，ピーク時の性能目標を達成し，システムリソース的にも良好な結果であった．また，ロングランテストでの安定性の確認も問題ない状況であった．

4. プロダクトセット選定結果

機能性，市場動向，性能（ベンチマークテスト結果）について比較検討した結果，本システムの中核となるプロダクトについては以下の通りとした．

1）データベースは Oracle RAC を採用

2）アプリケーションサーバは Oracle WebLogic Server を採用

3）プラットフォーム OS は DB サーバについては HP-UX，AP サーバについては Linux を採用

また，プラットフォーム OS の比較検討サマリを表 1 および以下に示す．

ⅰ）Linux または Unix いずれもエンタープライズシステムの基本機能/性能は提供可能

ⅱ）マルチプラットフォームでの動作はオープンソースである Linux が有利

ⅲ）大規模構成システムが必要な場合は Unix が優れている

ⅳ）保守性を重視する場合，自己完結可能なメインフレームおよび Linux が有利

ⅴ）DB サーバの OracleRAC クラスタ実績としては Unix（HP-UX）が多数である

ⅵ）WebLogic Server 稼働環境の JavaVM の対応が Linux，Solaris および Windows のみである

表 1　機能比較結果サマリ

プラットホーム	性能	可用性 信頼性	保守性	ﾏﾙﾁﾌﾟﾗｯﾄﾌｫｰﾑ適用度	拡張性	機密性	DBｸﾗｽﾀ 実績
Unix（HP-UX）	○	◎	○	△	◎	○	◎
Unix（Solaris）	○	◎	○	△	◎	○	○
Linux	○	○	○	◎	○	○	△
Windows	－	○	○	○	○	△	△

採用理由とともに，プロダクトセット一覧を表 2 に示す．

表2　プロダクトセット一覧

分類	利用ミドルウェア	採用理由
OS	Linux(RedHat)（Red Hat） APサーバ、バッチ実行サーバ、 ジョブ管理サーバ、ログ管理サーバ Unix(HP-UX)（日本HP） DBサーバ、バックアップサーバ、 CAMPGW、運用管理サーバ Unix(Solaris)（日本オラクル） CATGW Windows（マイクロソフト） ブレード管理サーバ、運用管理端末	Linux, Unix(HP-UX, Solaris)およびWindowsを候補として挙げ、比較検討を行った。また、ベンチマークテストにより、DBサーバOSとして、Linux, Unix(HP-UX, Solaris)を検証したが、ハードウェア性能とほぼ比例した結果となったため、OS依存による大きな性能差はないとういう結果となった。 1）DBサーバ：OracleRACでのクラスタ実績の多さから、Unix(HP-UX)を採用 2）APサーバ：スケールアウトによる拡張時のプラットホーム選択肢の多さ(マルチプラットホーム適合性)の観点より、Linuxを採用 ※その他のサーバOSについては省略
業務アプリケーション	AirCore（米国Unisys）	AirCore(Airline Core Systems Solutions) オープンエアラインパッケージ
アプリケーションサーバ	WebLogic Server（日本オラクル）	WebLogic, WebSphere, Oracle Application Server, JBOSSを候補として比較を行い、以下の理由からWebLogicを採用した。 1）性能面、信頼性に優れている 2）機能が豊富であり、容易に利用可能である 3）AirCoreにおいて実績がある（推奨） 4）実績/サポート力が高い
データベース	Oracle with RAC（日本オラクル）	Oracle, DB2, HiRDB, PostgreSQL, MySQLを候補に挙げ、比較を行った。米国Unisys社製メインフレームで稼働するable-Dで利用されているXTPAに基づく、無停止システム相当を現在実現しているのは、唯一Oracle RAC(Real Application Cluster)機能であり、日本ユニシスグループでの導入/稼働実績面からも、Oracleを採用した。
ログ管理	SenSage（米国Sensage）	基本機能、カスタマイズ性、性能、システム構成、可用性、導入実績より、他の製品と比較検討を実施し、SenSageを採用した。
クラスタウェア	CLUSTERPRO (Linux/RedHat)（NEC） HP Serviceguard (Unix/HP-UX)（日本HP）	Linux：導入実績より、CLUSTERPROを採用 HP-UX：Oracle稼働構成としての実績より、HP ServiceGuardを採用
システム監視	JP1/IM, SSO, NNM（日立）	日本ユニシスグループでの導入/サポート実績より、JP1/IM, SSO, NNMを採用
キャパシティ管理	JP1/PFM（日立）	日本ユニシスグループでの導入/サポート実績より、JP1/PFMを採用
ジョブ管理	JP1/AJS（日立）	日本ユニシスグループでの導入/サポート実績より、JP1/AJSを採用
バックアップ	HP DataProtector (ファイルバックアップ)（日本HP） HP Ignite/UX (OSバックアップ)（日本HP） JP1/SC/DPM (ブレードサーババックアップ)（日立）	HP-UX：ANA環境にて数多く導入されていることより、HP DataProtector, HP Ignite/UXを採用 Linux：サーバは全てブレード型となることから、JP1/SC/DPMを採用

5.　システム構成定義上の考慮点

システム構成定義において，検討を要した以下の考慮点について本章に記載する．

1）　外部ストレージによる筐体間ミラー構成

2）　AirCore QueueManager サーバアーキテクチャ変更

3）　Oracle RAC インターコネクト用ネットワーク機器変更

5.1　外部ストレージによる筐体間ミラー構成

米国 Unisys 社製メインフレームで稼働する able-D では，MHFS（Multi Host File Sharing）機能と UD（Unit Duplex）機能により，2 台のホストが同時稼働するシステムにおいて，複数の外部ストレージに同一データを保存する筐体間ミラーリング構成を実現している．この構成により，1 台の外部ストレージ全体が障害になった場合でも残存する外部ストレージを利用して業務処理が継続可能な構成となっている．

外部ストレージは RAID[*1] 構成によりディスクが冗長化されていること，電源やコントローラなど主要コンポーネントが多重化構成となっており，筐体障害の可能性は限りなく少ないものであるため，筐体障害の発生を考慮する必要はないという考え方が一般的である．

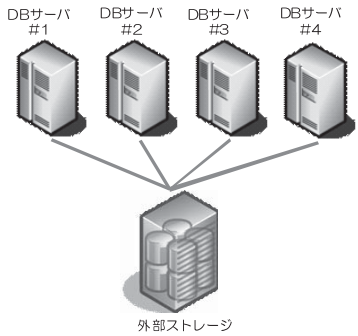
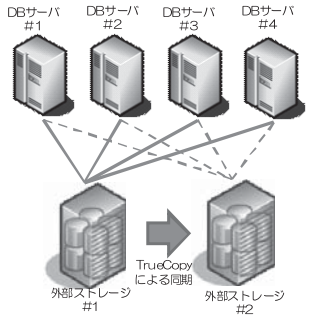
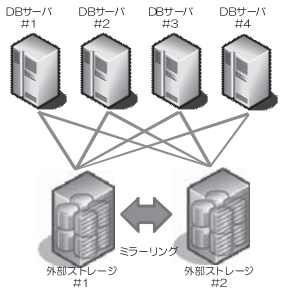
しかしながら，本システムでは外部ストレージにおいてもバックプレーンなど一部コンポーネントは多重化されておらず，障害発生時に筐体障害となるため，外部ストレージ 1 筐体構成の場合は SPOF を含む構成となることから，外部ストレージによる筐体間ミラー構成を検討

した．

オープン系システム環境においては，大規模オンライントランザクションシステムでの複数の外部ストレージ筐体によるミラーリング構成の実例がなかった．本システムで想定していたHP Superdome 4ノード構成ではOS（HP-UX 11i v3）が3ノード以上のミラーリング構成をサポートしていなかったことから，構成検討当初は，1)外部ストレージ単体構成と，2)外部ストレージ機能TrueCopy[*5]を使用した待機系（Standby）外部ストレージ構成の2案で検討せざるを得ない状況であった．

2008年9月にリリースされたOSバージョンにおいて，ディスク管理機能であるLVM（Logical Volume Manager）が2.1となり，ミラーリング機能SLVM（Shared LVM）が最大16ノード環境をサポートするようになったため，新たに3)SLVMによる筐体間ミラー構成を加えて検討を実施した．

表3　外部ストレージ構成案

	構成イメージ	特　徴
案1	外部ストレージ単体構成 	DBサーバが外部ストレージ1台と接続する構成 筐体障害時にはシステム全面停止となり，データ破損の場合は外部保存されたバックアップテープなどからのリカバリが必要になる オープンシステムで最も一般的な構成であり，導入実績多数
案2	TrueCopyによる待機系外部ストレージ構成 	DBサーバが本番系（Active）と待機系（Standby）の外部ストレージ2台と接続．通常稼働時は本番系の外部ストレージのみを使用する構成 外部ストレージが持つ筐体間コピー機能（TrueCopy）により，本番系外部ストレージに書き込まれたデータ内容は待機系外部ストレージ側にほぼリアルタイムで反映される 本番系外部ストレージに筐体障害が発生した場合，システム全面停止となるが，手動切り替えにより待機系外部ストレージを使用し復旧する オープンシステムでの導入実績は少ない
案3	SLVMによる筐体間ミラー構成 	DBサーバ4台と外部ストレージ2台がすべて接続される構成であり，外部ストレージ2台に対してすべてのデータは同時に書き込みを行う構成 外部ストレージ筐体障害（およびFC経路二重障害など筐体障害に準ずる障害を含む）においてもオンラインサービスを継続可能 4ノードでのSLVMによるミラー構成は2008年9月版でサポートされたLVM2.1の新機能のため，国内を初めとして世界的にも当該システム規模での導入実績がない状態

各案の構成イメージと特徴は表3の通りである．案1，案2ともに，外部ストレージに対する障害発生時または計画停止時にシステム全面停止となることから，案3を選択した．

ただし稼働実績がなかったため，稼働検証および各種障害系テストを実施することで本番稼働に必要となる挙動確認およびリカバリ手順の確立を行い，本番稼働に耐えうる構成であることをあらゆる方面から検証し最終構成に採用した．本構成は，オープンシステムにおいては世界初となる外部ストレージ筐体間ミラー構成であり，本番稼働後も安定的に稼働している．

5. 2　AirCore QueueManager サーバアーキテクチャ変更

AirCore は，業務単位（予約，発券，搭乗など）にモジュール化された機能が独立，または連携して処理を行うアーキテクチャとなっている．モジュール間で送受信されるメッセージはモジュールの独立性を維持したまま連携するため，非同期メッセージ連携を行う QueueManager サーバ（以下，QM サーバ）を経由してやりとりをする構成としている．メッセージ基盤には WebSphereMQ を使用している．AirCore における実装概要は図3の通りである．

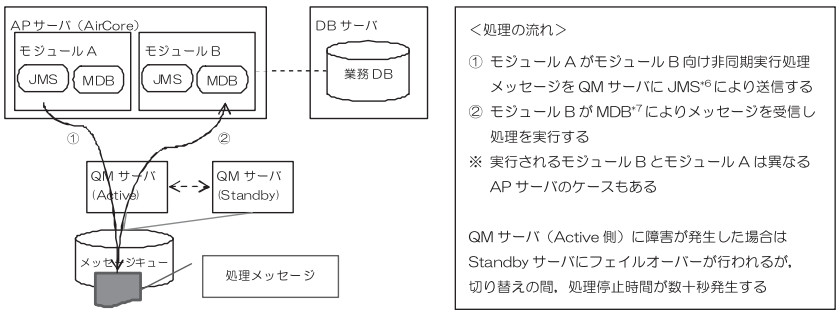


図3　AirCore での実装概要

ここで検証テストなどの結果から，全トランザクションのうちおよそ10%の処理が複数モジュール間をまたいだ処理になると想定されたため，最大処理件数からこの非同期連携アーキテクチャ上では20-30件/秒程度の処理が行われることになり，WebSphereMQ が稼働する Queue Manager サーバ（以下，QM サーバとする）の安定稼働を実現することは業務処理全体の可用性をあげるためには必須条件であった．WebSphereMQ の製品仕様から QM サーバの高可用構成は Active-Standby 以外の選択肢がなく，QM サーバ障害時には切り替わりに数十秒の処理停止（＝業務停止）が発生することが問題視され，更なる高可用構成を検討することとした．

開発元である米国 Unisys 社を含めた検討を実施した結果，代替案として図4と図5の二案から選択することとした．

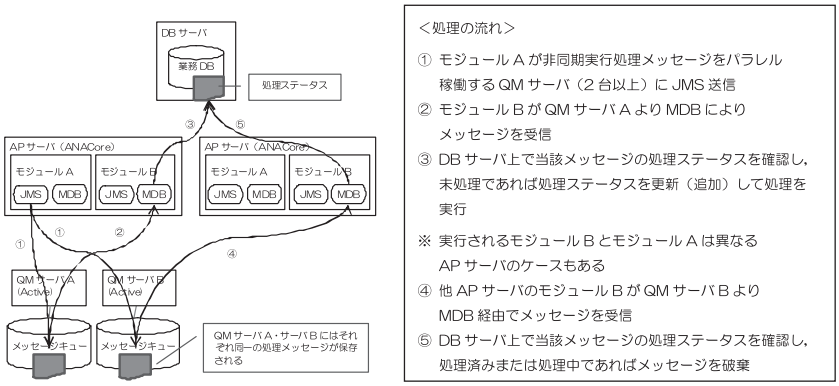


図 4　代替案 1（QM サーバ Active-Active 構成）

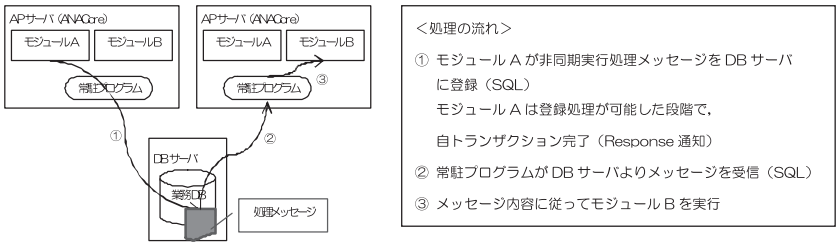


図 5　代替案 2（Oracle 常駐プロセス構成）

両案のメリット・デメリットを比較した結果，ANACore では代替案 2（図 5）を採用することとした．両案の比較結果を表 4 に示す．

	代替案1（QMサーバ Active-Active構成）	評価	代替案2（Oracle常駐プロセス構成）	評価
概要構成	処理経路が常駐プロセス構成と比較すると複雑	△	処理経路はQMサーバActive-Active構成と比較すると簡素 APサーバ上に常駐プロセスを稼働させる必要あり	○
可用性	QMサーバはパラレル構成のため、1台以上稼働していれば業務サービス提供は可能	◎	AP,DBサーバの冗長化レベルと同様であるため、可用性は高い	◎
機能性	内部スケジューラ・端末/ROP配信機能ともに実装可能	◎	内部スケジューラ・端末/ROP配信機能ともに実装可能	◎
性能面	DBサーバからのステータス情報取得、既処理済みメッセージの受け捨て処理が入るが、取扱情報サイズは小さいため性能への影響は軽微と予想	○	常駐プロセスがMDBと同じ役割を代替し、定期的にOracleへの問い合わせおよび（select処理）が追加となる処理ロジックの追加はない。性能への劣化はないと想定	◎
保守運用面	処理経路が比較的複雑になるため、障害発生時の対応手順として様々なパターンへの考慮が必要	○	構成は簡素となるため、障害発生時の対応手順は比較的パターンが絞られるただし、常駐プログラム本体のロジック変更は頻度は少ないと予想されるが 十分な考慮が必要	○
システム構成全体への影響	[各QM/バッチサーバ] Active-Activeになるため、クラスタウェア（CLUSTERPRO）が不要 [SANARENA] 各QM/バッチサーバ用に独立したキューデータ領域を保持	○	[QM/バッチサーバ] QM機能（WebSphereMQ、CLUSTERPRO）がなくなり、バッチ専用サーバに変更 [APサーバ] MDB相当の常駐プログラムが追加 [SANARENA] WebSphereMQ用のキューデータ領域が不要	○
開発・改修ポイント	JMS送信機能：複数のQueueManagerに対するメッセージ送信 ステータスをDBに書き込む処理 MDB受信機能：複数のQueueManagerからのメッセージ受信 ステータスをDBに書き込む処理 処理済みメッセージの受け捨て処理 データベース：ステータス保存テーブル追加 →　開発・改修ボリューム：大	△	メッセージ送信機能：データベースに書き込む機能（SQL）を追加 常駐プロセス　：MDB相当の機能を実装 データベース　：メッセージ格納用テーブル追加 （特定テーブルへの集中アクセスとなるため、 テーブル設計、物理配置、バッファプール独立 など設計時の考慮が必要） →　開発・改修ボリューム：中	○
マスタスケジュールへの影響	修正部分を加味しても機能テスト開始（2009/2Q）までの提供は可能であるので、マスタスケジュールへの大きな影響はなし	◎	同左	◎
NULでの実績	他ユーザでの販売・予約系システムにて実績あり 一般的な技術アーキテクチャを使用した構成であり、性能面でも問題がないと判断する	◎	他ユーザでの販売・予約系システムにて実績あり 一般的な技術アーキテクチャを使用した構成であり、性能面でも問題がないと判断する	◎
総合		△		◎

表 4　QM サーバ代替案比較表

2. 2. 1 項で述べた通り，業務トランザクションに対して SPOF をすべて排除するというシステム化方針を満たすために，AirCore パッケージが採用しているアーキテクチャを崩すこととなったが，結果として数十秒の業務停止が発生するという可能性を排除した構成とすることができた．

5. 3　Oracle RAC インターコネクト用ネットワーク機器変更

Oracle RAC は米国 Unisys 社製メインフレームで稼働する able-D で利用されている XTPA に基づく，無停止システム相当を実現できる唯一のデータベース・ソフトウェアとして，プロダクトセット選定において採用した（4 章に記載）．

Oracle RAC による高可用性アーキテクチャは，複数の DB サーバが外部ディスク上にあるデータベースを共有してアクセスする構成（シェアードエブリシング）をとっている．そのため，データベース上のデータが一貫性を持って参照・更新されるために，DB サーバ間でブロック単位でのデータ転送やロック・ロック解除処理などを行う専用のプライベート LAN（以下，インターコネクト LAN）を使用している．インターコネクト LAN 上で送受信されるデータはキャッシュ・フュージョンと呼ばれ，ノード数が多くなるほど通信対象が増加することからデータ量も比例して増加する（図 6）．AirCore ベンチマークテストでは最大数百 Mbps になることがあったが，本システムではアプリケーションパーティショニング[*8] 機能を実装することにより，最大でも 100Mbps 以下のトラフィックを実現している．

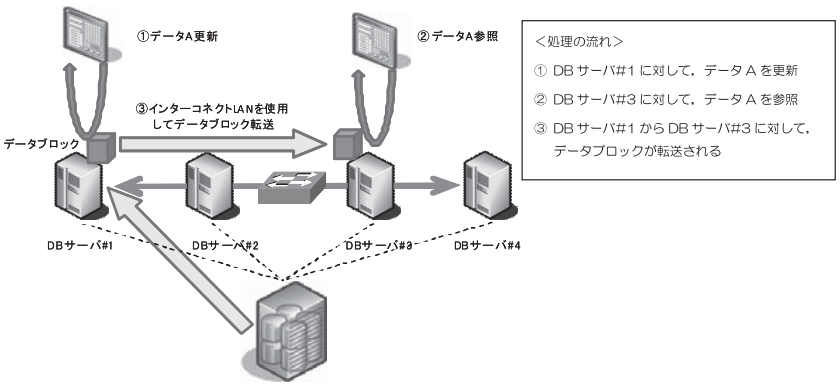


図 6　Oracle RAC キャッシュフュージョン転送

本構成において，データベース起動時のウォームアップ処理（Keep バッファキャッシュへのデータロード[*9]）などを実行した場合に，Oracle 上で gc_cr_multi_block_request（キャッシュ・フュージョンの再送処理）イベントの頻発により処理遅延となる事象が発生した．

原因を調査したところ，インターコネクト LAN 用ネットワーク機器 Cisco Catalyst2960（1000Mbps 対応）の処理能力に対して実トラフィックは数 10Mbps であるものの，当該機器においてパケットドロップが発生していることが確認された．

Oracle のアーキテクチャ上，インターコネクト LAN にはデータサイズの小さいパケットが大量に送られることと，パケット転送に UDP（User Datagram Protocol）プロトコルが使用されていることから，ネットワーク機器のパケット処理能力数を超えた場合に，当該事象が発

生していた．

UDP プロトコルは送達確認などをしない無手順方式のデータ転送で，信頼性・順序性・データ完全性を保障しない．このため，インターコネクト LAN の通信経路上（OS，NIC，ネットワーク機器）でパケットドロップが発生した場合，Oracle RAC による再送処理がなされる結果となり，処理遅延を引き起こす原因となっていた．

そのため使用機種（Cisco Catalyst 2960）およびその上位機種（Catalyst 3750 / Catalyst 4948）を使用し，DB サーバ稼働台数を増加させた場合の DB サーバ間の実効 NW 転送レート（サイズ，パケット数）推移を確認したところ，図 7 に示す通りの結果を得た．

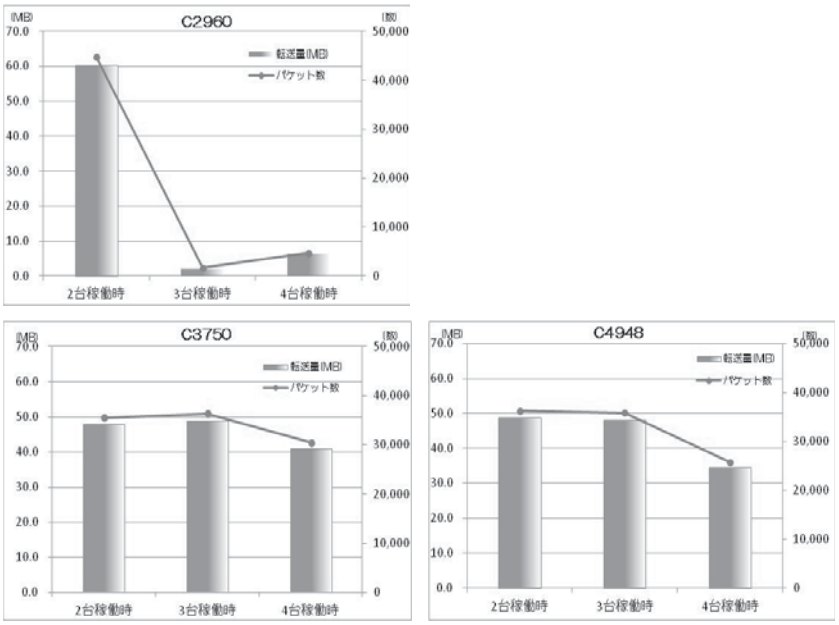


図 7　実効 NW 転送レート推移比較

テスト結果より，Catalyst 2960 では DB サーバ 3 台以上の構成において実効 NW 転送レートが極端に低下したが，Catalyst 3750/4948 といった上位機種では同様な事象は観測されなかった．この結果から 1000Mbps という同一処理能力を持った機器においてもより単位時間のパケット処理能力が高い機器を利用することにより，Oracle の処理遅延を解消することが確認できた．

本事象は，機器選定段階においてネットワーク機器上でのパケットロストの有無確認を実施しなかったことが原因特定を遅らせた要因である．また，ネットワーク機器のカタログスペックでは，データ処理レート（Mbps）は公表されているものの，単位時間のパケット処理能力は公表されていない．このことからも，機器選定段階での検証作業において，ネットワーク機器の設定（OS および NIC の設定を含む）およびトラフィックレートの確認のみならず，パケット処理能力の確認が非常に重要となる．

6. システム構成概要

システムテスト（業務サイクルテスト，パフォーマンステスト，ロングランテスト，運用テストなど）を完了し，システム構成を最終確定した，本番稼働時のシステム構成概要を図 8 および表 5 に示す.

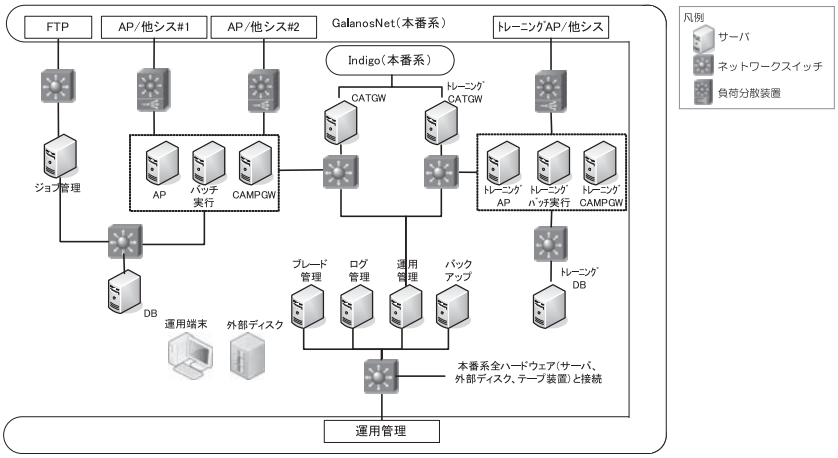


図 8　システム構成概要図

表 5　システム構成一覧

サブシステム	機種	冗長化方式	スペック	用途
DBサーバ	HP Superdome	Oracle Real Application Cluster、 HP ServiceguardによるParallel構成	1台あたり CPU：Itanium2(1.66GHz) × 12 メモリ：64GB	業務データベース管理を行う
APサーバ	Unisys rE6000 （ブレードサーバ）	NEC CLUSTERPRO Single Server Safe によるParallel構成	1台あたり CPU：Xeon(2.93GHz) × 2 メモリ：8GB	オンライン業務アプリを実行する
バッチ実行サーバ		NEC CLUSTERPRO Xを使用 Active-StandbyとParallelのハイブリッド構成	1台あたり CPU：Xeon(2.93GHz) × 2 メモリ：16GB	バッチ業務アプリを実行する
ジョブ管理サーバ		NEC CLUSTERPRO Xを使用 Active-Standby構成	1台あたり CPU：Xeon(2.93GHz) × 2 メモリ：8GB	ジョブの実行制御、FTP送受信を行う
ログ管理サーバ		ログ管理プロダクト固有機能による冗長化 N+1によるActive-Standby NEC CLUSTERPRO Single Server Safeによる Parallel構成	1台あたり CPU：Xeon(2.93GHz) × 2 メモリ：16GB	各サーバのログを収集・保持する
トレーニングAPサーバ		NEC CLUSTERPRO Single Server Safe によるParallel構成		オンライン業務アプリを実行する（トレーニング環境）
トレーニングバッチ実行サーバ		NEC CLUSTERPRO Xを使用 Active-StandbyとParallelのハイブリッド構成		バッチ業務アプリを実行する（トレーニング環境）
ブレード管理サーバ	Unisys rE5000	Parallel構成	1台あたり CPU：Xeon(2.93GHz) × 2 メモリ：4GB	ブレードサーバ（rE6000）を管理する
CAMPGW	HP rx2660	HP ServiceguardによるParallel構成	1台あたり CPU：Itanium2(1.66GHz) × 2 メモリ：8GB	ASW/CAPと接続を行う
トレーニングCAMPGW				CAPと接続を行う（トレーニング環境）
バックアップサーバ	HP rx3600	HP ServiceguardによるActive-Standby構成	1台あたり CPU：Itanium2(1.66GHz) × 2 メモリ：8GB	バックアップ・テープライブラリ管理を行う
運用管理サーバ		HP ServiceguardによるActive-Standby およびActive-Activeハイブリッド構成		監視、障害検知、障害通知等を行う
トレーニングDBサーバ		HP ServiceguardによるActive-Standby構成	1台あたり CPU：Itanium2(1.66GHz) × 2 メモリ：32GB	業務データベース管理を行う（トレーニング環境）

7. お わ り に

現行のメインフレームからオープンプラットフォームへの移行にあたり，本稿記載の計画に沿って進めてきた．途中幾多の問題に直面したが，全日本空輸株式会社様，ANA システムズ株式会社様を始めとして，日本ヒューレット・パッカード株式会社様，日本オラクル株式会社様，株式会社日立製作所様，多くの方々のご協力により，無事本番稼働を迎えることができた．ご協力頂いた全ての皆様に，誌面を借りて深く感謝申し上げたい．

今回稼働を遂げたオープン系システムは，2006 年に技術検討を開始したシステムである．2013 年 10 月現在の技術動向として，クラウドコンピューティングによる IT 資産の保有から利用への転換，それを支える要素技術として，仮想化技術，OSS（Open Source Software）

の成熟と利用など，技術革新が進んでいる．このような状況下，新たなインフラプロダクトセット検討の機会があれば参画し，優れたインフラ基盤の構築に貢献したい．

＊1　RAID（Redundant Arrays of Inexpensive Disks）：複数台のハードディスクを組み合わせて，仮想的な1台のハードディスクとして運用し冗長性を向上させる技術であり，主に信頼性・可用性の向上を目的として用いられる実装形態のこと．ハードディスクの構成によって，RAID0，RAID1，RAID5などの組み合わせが存在する．RAID0：複数台のハードディスクにデータを分散して読み書きし高速化した構成であり，ストライピングと呼ばれる．RAID1：複数台のハードディスクに同時に同じ内容を書き込む構成であり，ミラーリングと呼ばれる．RAID1（1+0）：RAID1とRAID0の組み合わせ構成．RAID5：複数のハードディスクでデータに誤り訂正符号データを加えて分散して読み書きする構成．

＊2　エンタープライズシステム：大企業の基幹系システムを指す．企業の業務システムにおいて，システムの価格・機能・性能に加え，サービスを停止させない信頼性，大量の処理を行った時の安定性，ベンダーのサポート体制の充実などが求められるシステムのこと．

＊3　RASIS：「信頼性（Reliability）」ハードウェアやソフトウェアによるエラーがないように保つ．MTBF（平均故障間隔）が長い．「可用性（Availability）」エラーが発生した場合に短時間で回復できる．MTTR（平均復旧時間）が短い．「保守性（Serviceability）」エラー原因を短時間に究明でき修復できる．エラー診断機能や故障装置の切り離し，予備装置への切り換えなどができる．「保全性（Integrity）」ハードウェアやソフトウェアのエラーによって，正常動作している機能を侵害しないように防止できる．また，矛盾なく復旧できる．「機密性（Security）」意図的な操作によって情報漏えいや改ざん・破壊が起きないよう保護できる機能を備えていること．

＊4　機器構成算定：必要なハードウェアのリソース（サーバ台数・CPU数・メモリ数）を算出すること．構成算定は特定のハードウェア製品を仮決めして実施するが，採用製品を決定することまでは目的としていない．採用製品の決定は必要リソース構成を組めるかということの他に，価格・ベンダーサポートなど，他の事由を勘案して決める必要がある．仮決めした製品とは別の製品を採用する場合は，構成算定結果と製品性能比を勘案して実際に調達するハードウェア構成を決定する．

＊5　TrueCopy：日立製ストレージシステムにおいて，ホスト/サーバを経由せず，ストレージ筐体間で，直接データコピーを実施する機能

＊6　JMS（Java Message Service）：Java標準APIで提供されているメッセージ送信機能

＊7　MDB（Message Driven Bean）：EJBで定義されたメッセージ駆動型処理機能

＊8　アプリケーションパーティショニング：データベースに保存されたデータを，特定のデータは特定DBサーバからアクセスされるようにアプリケーションロジックで制御すること．本システムではフレームワーク部において実装されており，便名・端末IDなどの情報からDBサーバを特定してアクセスするようにしている．本機能の実装により，キャッシュ・フュージョントラフィックの抑制を実現している．

＊9　Keepバッファキャッシュへのデータロード：アクセス頻度の高いオブジェクト（テーブル，インデックス）を対象として，バッファキャッシュのKeepバッファプールに事前にデータをロードしておくことで，データベース処理の高速化を図っている．

執筆者紹介　石　崎　達　也（Tatsuya Ishizaki）

1983年日本ユニシス(株)入社．プロダクト主管部にて主にネットワーク系プロダクトの開発・保守を担当．2001年より公共システム部門にてエアライン系のシステムサービスに従事し，現在に至る．

滝　沢　和　明（Kazuaki Takizawa）
1994年日本ユニシス(株)入社．官公庁関連システムの提案・開発・保守作業を担当し，2006年よりエアライン関連システムの基盤設計・開発を担当．現在はエアラインサービス二部に所属．