次世代グリーンデータセンターを構成するシステム：Mangrove

New System Architecture for Next-Generation Green Data Centers: Mangrove

● 三吉貴史　● 大江和一　● 田中　淳　● 山本　毅　● 山島弘之

あらまし

次世代グリーンデータセンターを構成する，サーバ，ストレージ，ネットワーク，ミドルウェア，ファシリティを融合した垂直統合による全体最適化システム“Mangrove”について紹介する。Mangroveは次世代グリーンデータセンターのコンセプトである資源プールと機能のミドルウェア化を実現するシステムアーキテクチャであり，Mangroveに基づくITプラットフォームにより，柔軟で効率的な資源利用，迅速な再構成，可用性・信頼性向上とその効果として低コスト，省電力を可能にする。Mangroveを構成する要素として，ハードウェア資源をプール化するサーバ・ストレージアーキテクチャ，ミドルウェアによる資源プール上でのストレージ機能の実現，スケーラブルなデータセンターネットワーク，低コスト・高集約な光インターコネクトによる高速インタフェース，VM（Virtual Machine）配置を最適化する運用管理技術の取組みについて取り上げ，それぞれのねらいと特徴について述べる。

Abstract

We propose new system architecture for Next-Generation Green Data Centers. It integrates servers, storage, networks, middleware and facilities into one consolidated system. This new system architecture, called Mangrove, realizes green data center concepts such as “resource pooling” and “offloading to middleware.” An IT platform based on Mangrove allows flexible and effective resource usage, agile configuration, high availability and high reliability. These features lead to reduced costs and power savings at data centers. Mangrove consists of several basic elements including server and storage architecture, which enables hardware resource pooling; storage functions on the resource pool as middleware; scalable data center networks; high-performance interfaces achieved by low-cost and high-density optical interconnects; and system management with virtual machine (VM) placement optimization. This paper describes the aims and features of the basic elements of Mangrove.

ま え が き

クラウドコンピューティングの浸透により，データセンターに求められる役割が大きく変化している。既存のエンタープライズ向けITシステムに加え，今後新しく出現してくるサービスに適応可能なコンピューティングプラットフォームが求められる。

富士通研究所では，次世代グリーンデータセンター（Green IDC）として，IT機器，運用管理，ファシリティの垂直統合によるコンピューティングプラットフォームの最適化に取り組んでいる。本稿ではGreen IDCのコンセプトである資源プール化と機能のミドルウェア化をもとにデータセンターの垂直統合を実現するアーキテクチャ“Mangrove”を構成する技術について紹介する。Mangroveに基づくITプラットフォームは，サーバ，ストレージ，ネットワークの基本構成要素とそれらを接続するインターコネクト，およびそれらを一体的に運用する運用管理から成る。さらに，Mangroveではファシリティとの一体的構築・運用により，データセンターの垂直統合を追求する。以下では，それぞれの構成要素のねらいと特徴について述べる。

サーバアーキテクチャ

Mangroveを構成するサーバは，資源プール化によるデータセンターの柔軟性を追求する。従来のデータセンターでは，サーバ筐体，ストレージ筐体といった装置を多数並べ，ソフトウェアによって資源プールを構成していた。[1] このため，資源の追加，削除が装置という枠で制約を受け，一度装置構成を決定すると変更が難しいという問題があった。そこで，Mangroveでは資源プールの粒度をもう一段細分化し，CPU，メモリ，HDDといったコンポーネントレベルでハードウェア資源プールを構成する。このアーキテクチャにより，資源利用率の向上と迅速な構成変更をメリットとすることができる。

コンポーネントレベルの資源プール化として，まず従来のサーバからマザーボード（M/B）とストレージドライブ（HDD，SSD：Hard Disk Drive, Solid State Drive）を分離し，**図-1**に示すようにM/Bプールとディスクプールを構築した。ディスクプールにより得られる効果として，第一

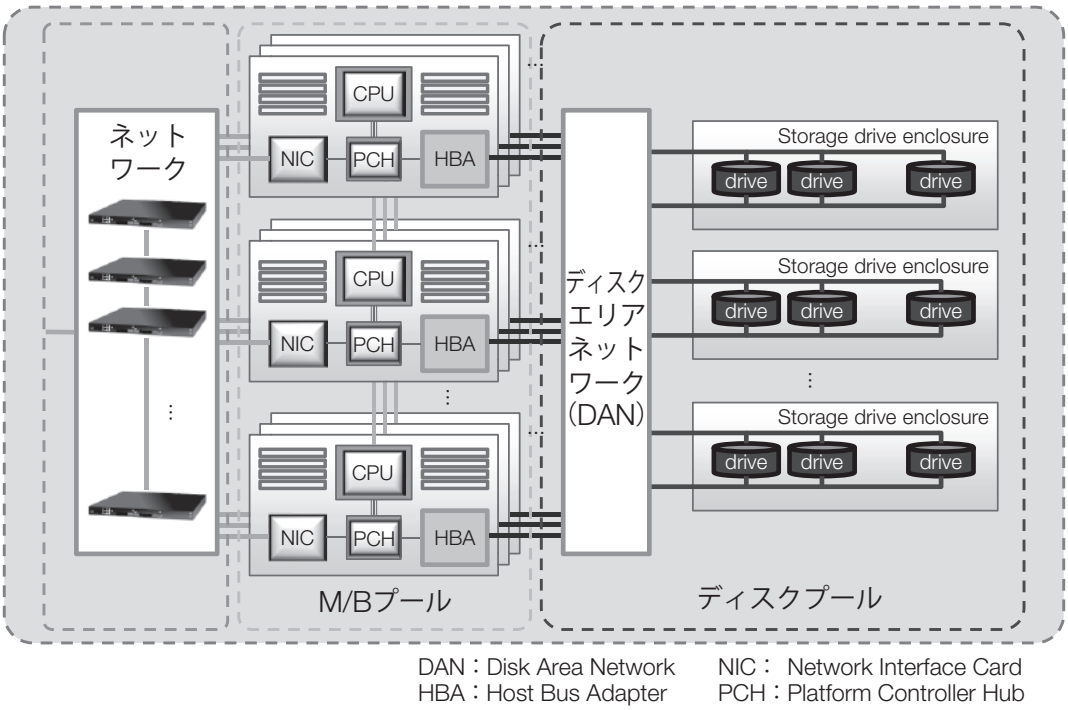


図-1　ディスクプール構成
Fig.1-Disk pool configuration.

に実装面の効果がある。高発熱のCPUと低発熱のHDD/SSDを分離することにより，冷却構造を最適化する。さらに，M/Bプールとディスクプールそれぞれに適した筐体，ラック実装により，資源プール化を前提としたデータセンターのレイアウト最適設計を行うことが可能になる。つぎに，柔軟なディスクプール化によってストレージの構成をシステム要件に応じて最適化することが可能になる。例えば，M/B1枚と4台のHDDを結合し通常の汎用サーバとすることや，M/B1枚に多数のディスクを接続し，ストレージサーバを提供することが考えられる。

このようなディスクプールを実現する上で重要なのが，M/Bプールとディスクプールの間を接続するインターコネクトである。ここでは，従来のファイバーチャネル[(2)]などによるSAN（Storage Area Network）と区別するため，ディスクを接続するインターコネクトの意味でDAN（Disk Area Network）と呼ぶ。DANは，従来のローカルディスクに相当する部分を接続するため高い性能と低コストの両立が求められる。DANはSANと異なり，以下のような特徴を持つ。

・トポロジ：1台のディスク（ターゲット）を複数のサーバ（ノード）で共有する必要がない。
・ルーティング：構成変更はまれであり，単純な回線スイッチでよい。

このように，トポロジ，ルーティングなどが簡単で済むため，低コストでの実装が容易となる。また，DANで構成されるディスクプールによって，物理ディスク資源を高速かつ柔軟に接続する。したがってシステム要件に応じてディスク構成を提供することが可能になり，ストレージドライブ資源の利用効率を向上させる。

ストレージシステムの構成

ストレージシステムの構築に関して，従来はキャパシティプランニングで事例ごとに必要なストレージ装置の種類や数を決めていた。Mangroveでは任意のCPUと複数の任意のストレージを自由に組み合わせることができるので，Mangroveの中から事例に合うストレージ装置を作り出すことができる。これは，事例ごとに異なる構成のストレージ装置を準備しなくてよいことを意味する。この目的を達成するためには以下の課題解決が必要である。

（1）M/B，ディスクプールから必要なリソースをオンデマンドで切り出し，システム構築を行える枠組みの構築
（2）上記（1）で切り出したシステム上にストレージ機能を提供するミドルウェア配備機能の構築

これらの課題に対し，（1）に関しては，ストレージ機能だけではなく汎用的にシステムを構築する枠組みとしてMangroveManagerを開発し，（2）に関しては，RAID機能を実現するAkashoubinを開発した。

MangroveManagerは以下のように動作する。

（1）サーバの構築

ユーザが指定したCPU，ストレージドライブの種類，数に応じて各プールからリソースを切り出し，DANを操作して接続を行う。

（2）OS/ミドルウェアのインストール

構築したサーバに対して，ユーザが指定したOS・ミドルウェアをネットワーク越しにインストールする。[(3)]

Akashoubinは以下から構成される（**図-2**）。

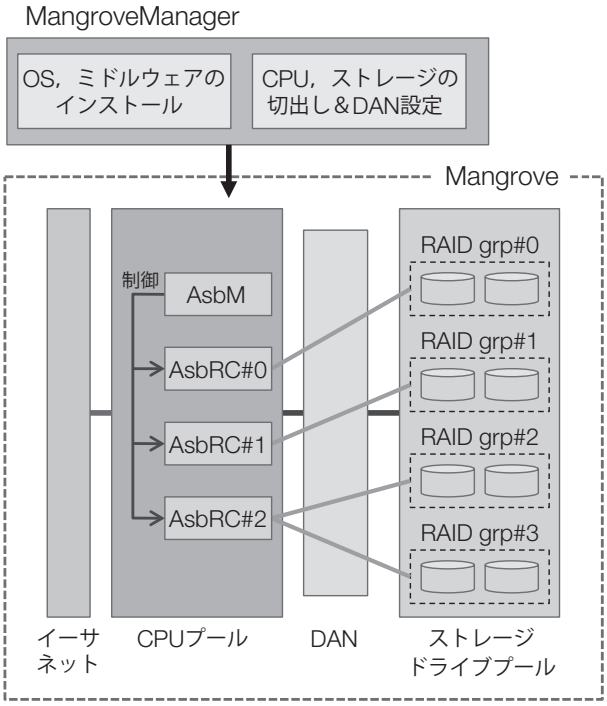


図-2　Akashoubinの構成
Fig.2-Configuration of Akashoubin.

・AsbM（Akashoubin Manager）
　複数のAsbRCの管理（生存監視など）
・AsbRC（Akashoubin RAID Controller）
　複数のRAIDグループが接続され，RAID機能を実現
・RAIDグループ
　ディスクプールから切り出された複数のストレージでRAIDを構成

Akashoubinでは，CPU/ストレージドライブ故障時に代替部品をプールから補給できるので運用の簡素化が期待できる。

データセンターネットワークの構築

本章ではデータセンターネットワークとしてサーバやストレージなどMangroveによって組み上げられたシステム間のLayer 2ネットワーク（従来のLAN，SANを指す）について述べる。

Green IDCの基本概念の一つに「垂直統合によるコストパフォーマンスの最適化」がある。これを体現し，次世代のクラウド基盤を提供するために，ネットワークとしては以下の点を重要課題ととらえる。

・スケーラブル（例2000サーバ程度）
・フラット（どのサーバとも一様な遅延，スループットで通信可能）
・仮想ネットワークのセキュリティと高信頼
・低コスト・省電力

今回，研究開発を進めているネットワークは大きく以下の三つの要素に分解される。これらを有機的に統合し，フラットでスケーラブルなデータセンターネットワークの実現を目指す（**図-3**）。

（1）高密度大容量スイッチ

高密度なMangroveを接続するためにはスイッチも10GE多ポート・高密度化を目指す。本スイッチの特徴としてルーティングプロトコルなどを処理するスイッチのローカルCPU機能をMangroveのCPUプールにオフロードする機能をサポートする。重いコントロールプレーン処理はCPUプールで，軽いマネジメント処理だけならスイッチのローカルCPUで実現するといった柔軟性を提供することで，処理に応じたコストを実現する。

（2）サーバサイドネットワーキング

ネットワーク機能が複雑化するとスイッチが高価格になるため，上記高密度大容量スイッチには基本的な転送機能・性能を求め，複雑なネットワーク機能はサーバ側で実現することで，システム全体の高機能・低コスト化を目指す。これをサーバサイドネットワーキングと呼ぶ。具体的には，スケーラブルな仮想ネットワーク分離機能[4]，ユーザ単位の細やかな帯域制御機能，冗長制御機能などをサーバのホストOSで実現する。

（3）ネットワークマネージャ

上記高密度大容量スイッチの管理と，サーバサイドネットワーキング機能を実現するため，リソー

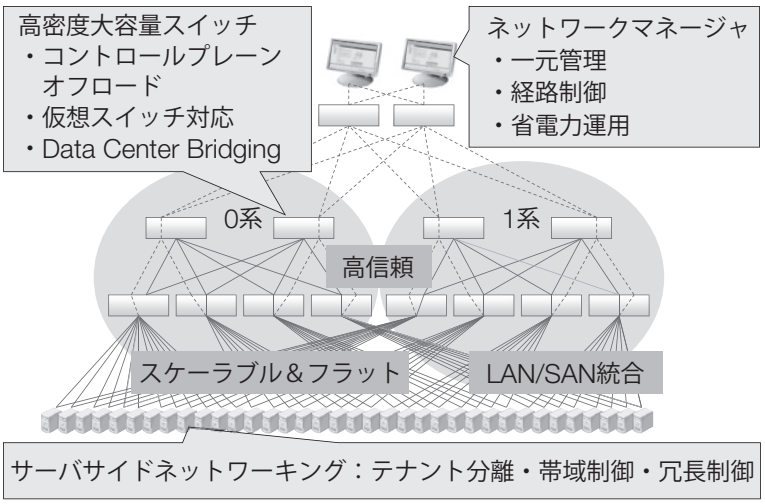


図-3　データセンターネットワーク
Fig.3-Data center network.

ス管理，トポロジ管理，経路制御，収容設計，可視化などLayer 2ネットワーク全体のマネジメントを行う。トラフィック監視を基にネットワークの省電力化と転送性能維持の両立を図る経路制御を行う。従来のイーサネットのような学習ベースではなく，ネットワークマネージャが経路制御を一括して行うことで，全体最適制御を実現する。

光インターコネクト

Green IDCではCPU，ストレージなどを機能別に集約してプール化し，高密度サーバの実現を目指している。高密度のリソースは多量の信号接続を要求し，高バンド幅の接続技術の重要性が増している。近年，高い伝送特性を持つ光技術の活用によるバンド幅の拡大が進んでいるが，非常に大容量の接続が要求されるGreen IDCでは，さらに一歩踏み込んだ光技術の活用方法を検討している。ここではそのための光インターコネクト技術（**図-4**）について述べる。

（1）サーバ間光インターコネクト

Green IDCの中核技術であるMangroveでは，DAN接続のため多数の筐体間I/Oが要求され，低コスト・小容積のインターコネクトの実現がかぎとなる。低価格なコンシューマ向け光技術をサーバに適用することでの抜本的なコストダウンをねらった。あまり高くない信頼度でサーバ適用の要求を満足させることが課題であり，冗長構成を検討している。コストパフォーマンスの高い10 Gbpsの伝送容量を持つ光モジュールを調査し，PC用途の光トランシーバを第一候補としてシステム検証（SAS信号の伝送に要求される伝送速度での誤り率の評価）を行い，サーバへの適用が可能であることを確認した。1モジュールあたりのポート数は冗長性と小容積のバランスを考慮して2 chとし，小規模のシステム試作での検証を目指している。

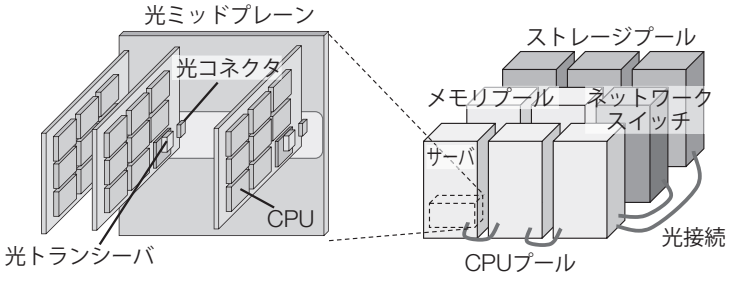
（2）サーバ内光インターコネクト

サーバの性能が向上した場合，サーバ間のみではなく，サーバ内（システムボード間）にも高バンド幅の信号接続が必要となる。また，入出力部，メモリの拡張など，バスを外部に引き出す機能も必要となってくる。これらの用途では，サーバ内の非常に限られた領域での高バンド幅接続となり，高速・高密度（当然低コスト）の光インターコネクト技術が要求される。富士通研究所では要素技術として，高速・高密度の光トランシーバ，高密度の光ミッドプレーン（約2000本の光ファイバをコンパクトに配置）の開発を行った。それらの技術を基に疑似サーバの試作を行い，10 Gbpsの光信号のミッドプレーン経由での伝送，PCI-E信号の光化の動作検証を行い，高バンド幅の装置内光インターコネクトの実現性を明らかにした。

VM配置最適化

この研究では，Mangroveが想定する利用シーンの中で必要となると考えられる，VM（Virtual Machine）配置先の管理者指定と完全自動化を両立する配置設計フレームワークの試作と評価を行っている。従来の一般的な制御に対する優位点として，つぎの3点がある。

（1）省電力や耐故障性向上など様々な最適化目的の独立した改善・追加・削除の実現のため，最適化目的別に個別プラグイン構造を採用した



（a）装置内光インターコネクト　（b）装置間光インターコネクト

図-4　光インターコネクトのサーバ適用
Fig.4-Optical interconnects' server applications.

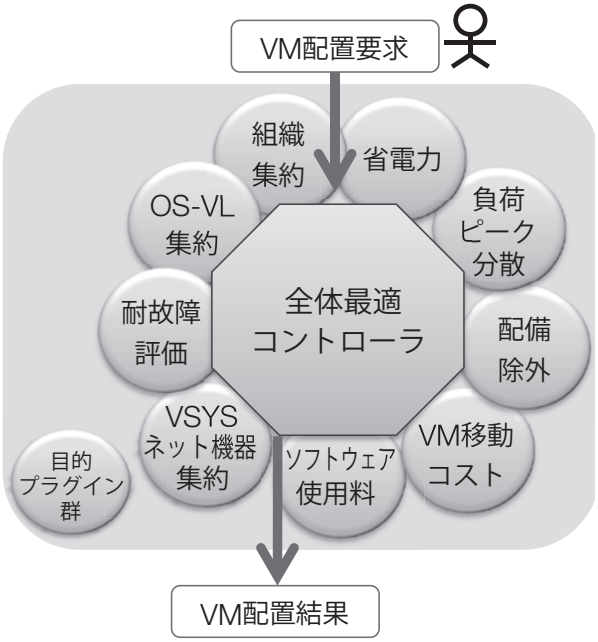


図-5　評価プラグインの概念図
Fig.5-Basic concept of evaluation plug-in.

（**図-5**）。

（2）観点の異なる個別最適化目標を統一的に処理するため，評価点合計値を最大化/最小化という共通フレームワークで処理する。

（3）任意のVM技術に対応可能とするため，VM配置設計とVM配置実行を分離実装する。

インフラ運用には様々な要件があり，従来，最適化目的となる対象として，消費電力，ソフトウェアライセンス料金，ネットワーク通信量，耐故障性などが知られている。ただし，環境や社会的要請により個々の要件の優先度も変化するため，特定の目的に向けた実装は直ちに陳腐化する恐れがある。また，定期保守・パッチ適用や機器交換などのインフラ停止操作への対応も同じ枠組みの中で対応することが望ましい。

富士通研究所では，これらの課題を総合的に扱い，要件の変化に柔軟に対処する方式として，個別プラグイン構造とVMの配置先物理サーバを基に得点を算出する評価得点方式を採用した。例えば，消費電力評価プラグインでは，VMを電源がON状態の物理サーバへ配置する場合は，電源がOFF状態の物理サーバへ配置する場合より評価得点を小さくする。このようにすると電源がONの物理サーバへVMを配置しようとする。

この方式では，最適化目的ごとに独立した個別の評価プラグインで評価値を算出し，その総合値を最適化する。評価値の恣意性を避け，実際的な評価とするために評価値は実際に発生するコストに関連付けることにした。これにより，評価軸の異なる多様な最適化を，コストの最小化と考えることができ，電力課金，ソフトウェアライセンス課金などは諸定数から客観的に定義できる。他方，冗長性やインフラキャパシティの消費に伴う潜在的コストについては，その数値化において恣意性が発生する余地がある。これは最適化目的のモデル化における問題であるが，評価プラグインの差替えによるモデルの変更という形で修正可能である。

つぎに直面する課題として，最適配置の計算量の問題がある。インフラ規模/能力がある程度大きくなり，場合の数が増大すると，計算量が爆発する。計算量削減の既知の方策としては，重複パターン排除，計算値キャッシュ，ヒューリスティック導入などが知られている。これらを検討した後に，最小となる評価得点を取り得るVMの配置先物理サーバが多数存在することに注目し，独自に開発した等価集合評価を導入した。

以上により，IaaSクラウドにおける典型的なVM配置要求に対して1秒未満での配置設計計算を完了することができ，事前に静的な配置を決定することなく，VM配置要求が与えられるごとに，リアルタイム・オンデマンドの配置設計を実行することが可能になった。

むすび

本稿では，次世代のデータセンターに向けたMangroveを構成する要素技術について述べた。ハードウェア資源をプール化するサーバアーキテクチャとストレージシステムにより，柔軟で構成変更可能なシステムを構築する方式について論じた。また，コストパフォーマンスを最適化するデータセンターネットワークを提案した。さらに，低コスト・高集約な光インターコネクトによる高速インタフェースや，VM配置を最適化する運用管理技術について紹介した。

今後は，これらの要素技術を一つの試作システムに統合し，コンセプト実証を進めていく。さらに，

メモリプール，オブジェクトストレージなど，新しい技術にも取り組んでいく。

参考文献

（1）吉田 浩ほか：サービス指向プラットフォーム. *FUJITSU*，Vol.61，No.3，p.283-290（2010）.

（2）T11 Home Page. http://www.t11.org/

（3）Cobbler Website. https://fedorahosted.org/cobbler/

（4）尾上浩一ほか：スケーラブルなクラウドネットワークを実現するホストベース論理分離技術．SACSIS，2011.

著者紹介

三吉貴史（みよし　たかし）
ITシステム研究所サーバテクノロジ研究部 所属
現在，サーバアーキテクチャ関連の研究に従事。

山本　毅（やまもと　つよし）
ITシステム研究所サーバネットワーキング研究部 所属
現在，サーバ用光インターコネクトの研究に従事。

大江和一（おおえ　かずいち）
ITシステム研究所システムミドルウェア研究部 所属
現在，ストレージ関連の研究に従事。

山島弘之（やましま　ひろゆき）
クラウドコンピューティング研究センター 所属
現在，クラウドデータセンターの運用管理の研究に従事。

田中　淳（たなか　じゅん）
ITシステム研究所サーバネットワーキング研究部 所属
現在，データセンターネットワークの研究に従事。