ORACLE

なぜ、InfiniBandなのか？
オラクル製品に見るInfiniBand

日本オラクル株式会社
システム事業統括
プロダクト・マネジメント・オフィス
シニア・セールスコンサルタント
谷 茂俊

#odddtky

日本オラクル、今年最大の技術トレーニング・イベント
Oracle DBA & Developer Day 2013

以下の事項は、弊社の一般的な製品の方向性に関する概要を説明するものです。また、情報提供を唯一の目的とするものであり、いかなる契約にも組み込むことはできません。以下の事項は、マテリアルやコード、機能を提供することをコミットメント（確約）するものではないため、購買決定を行う際の判断材料になさらないで下さい。オラクル製品に関して記載されている機能の開発、リリースおよび時期については、弊社の裁量により決定されます。

Program Agenda

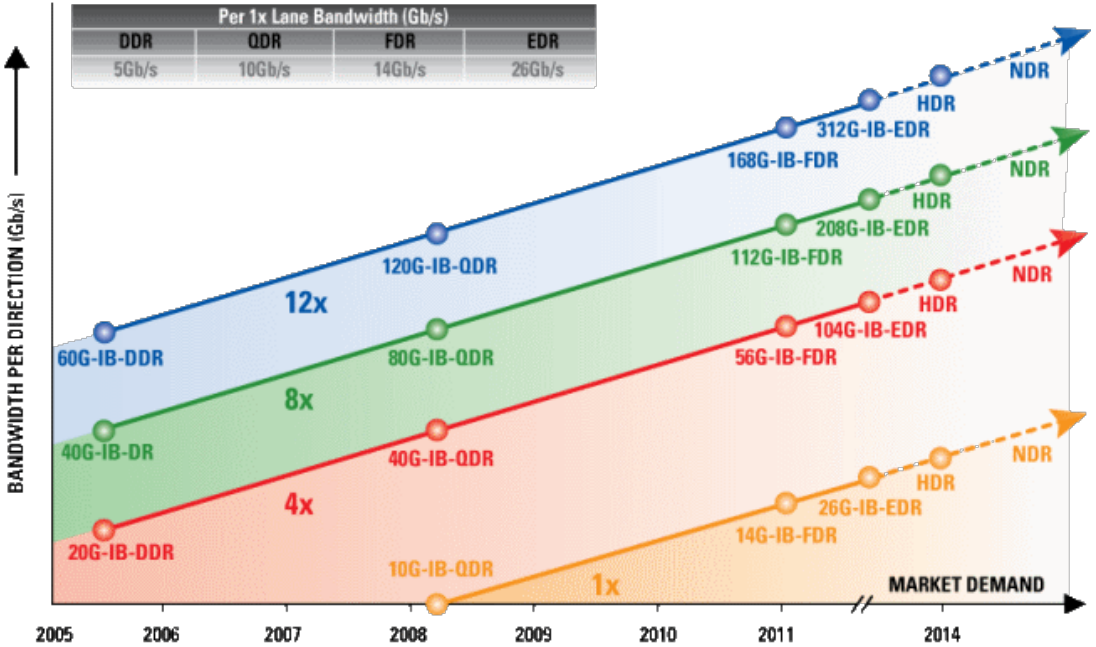
- InfiniBand概要
- リンクレイヤー
- データセンターへの適用
- お客様事例

InfiniBand概要

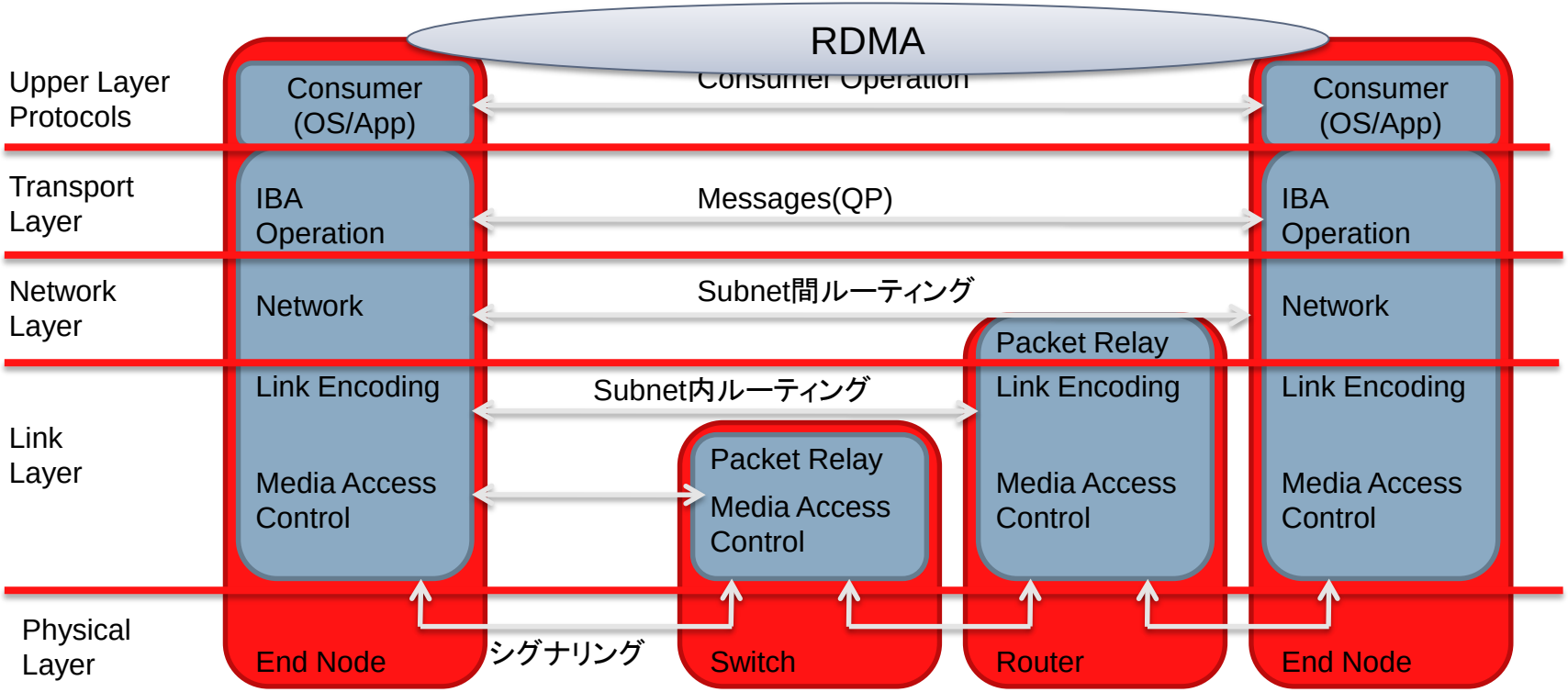
InfiniBand (IB) の特徴と業界動向

エンタープライズでの利用も

- Microsoft Windows Server 2012 が RDMA を標準対応
  - SMB Direct：RDMAを使ったファイルサービス
- ストレージバックエンド接続
- EDR（100G）ロードマップ
- HPCでは、TOP500 のうち 205がInfiniBandのシステム
- SSD over InfiniBand

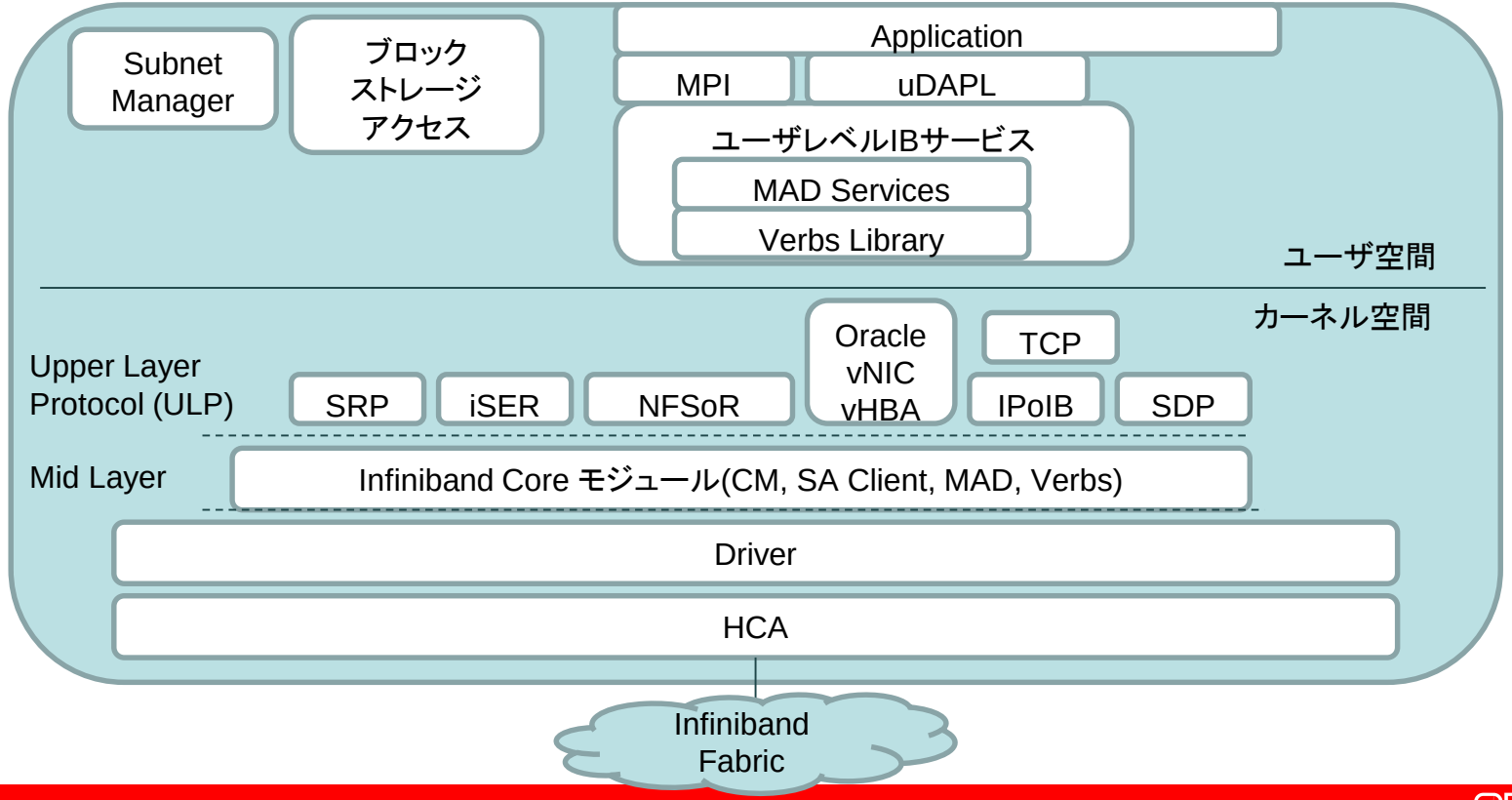


レイヤー・アーキテクチャ

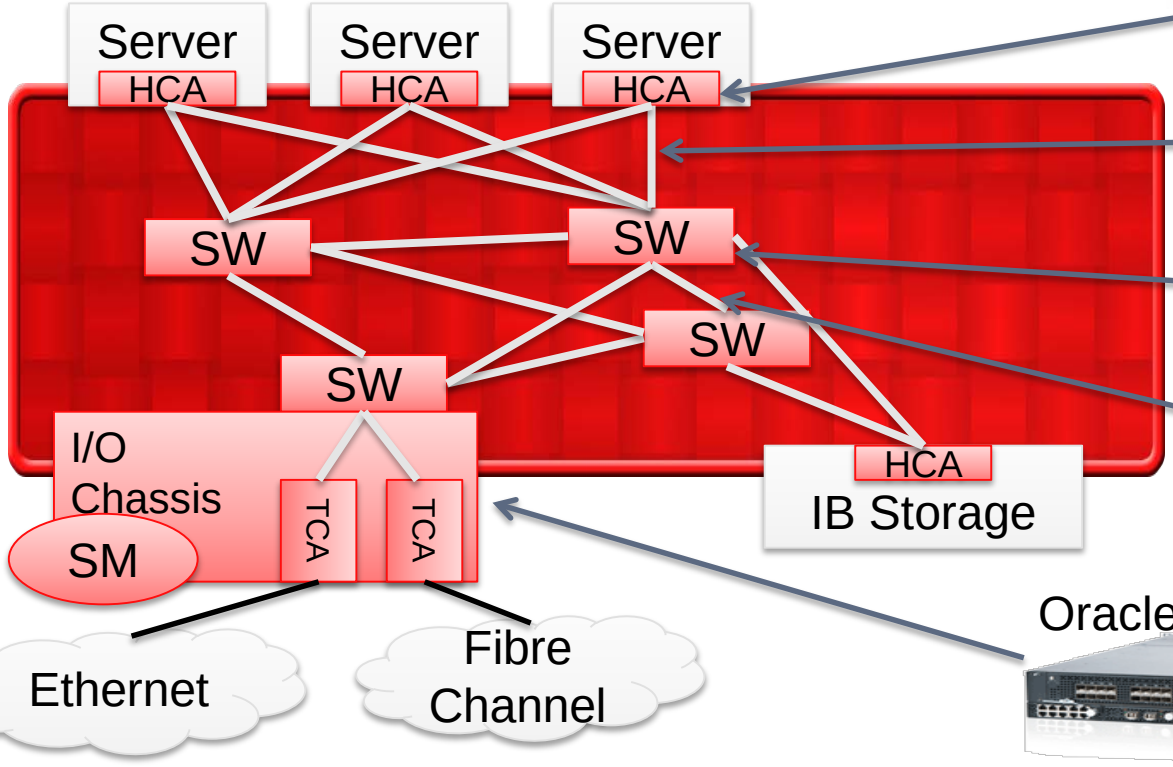


- 詳細は、*"InfiniBand Architecture Specification"*をご参照ください。

InfiniBandソフトウェア・アーキテクチャ



InfiniBand ファブリック



HCA

QSFP
カッパーケーブル

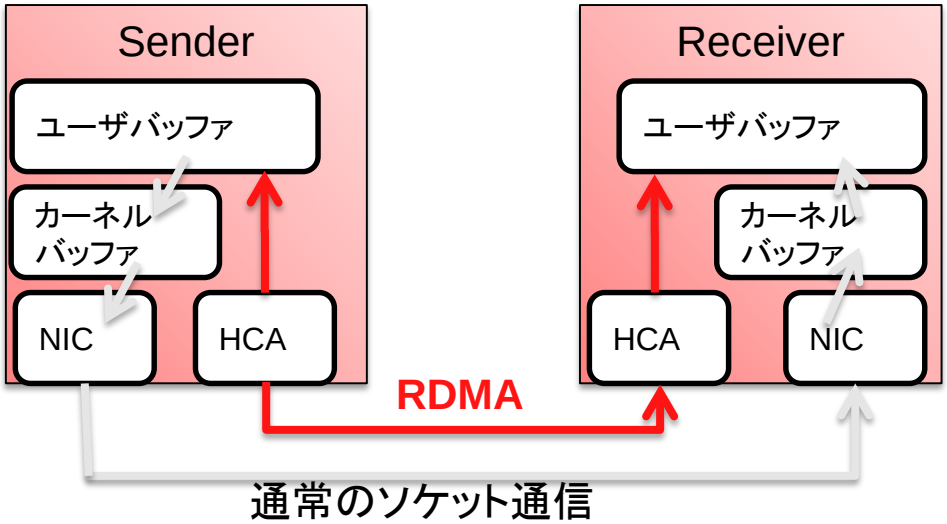
IBスイッチ

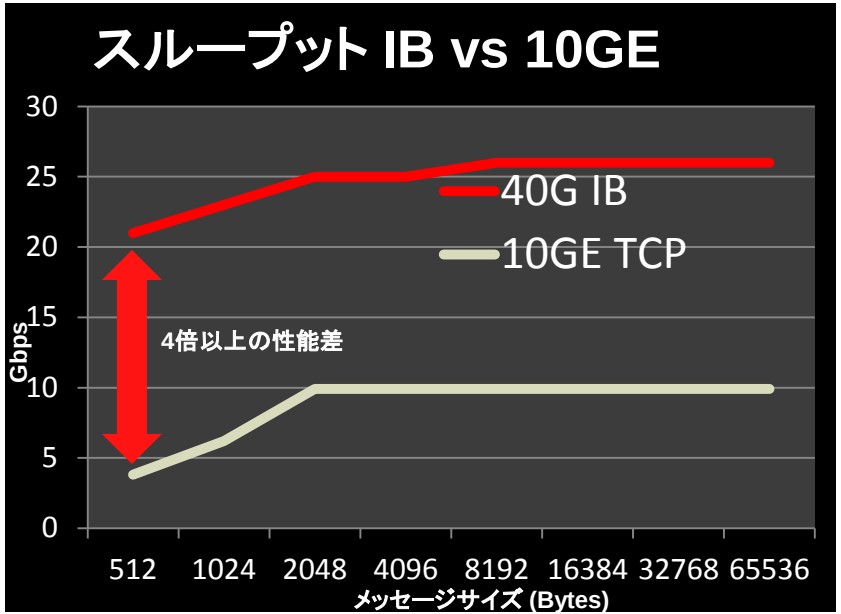
QSFP光ケーブル

Oracle Fabric Interconnect

InfiniBand (IB) とRDMA

- Remote Direct Memory Access(RDMA)
  - アプリケーションはリモートノードと直接通信
  - システムバス、CPUの負荷も低く抑えられる



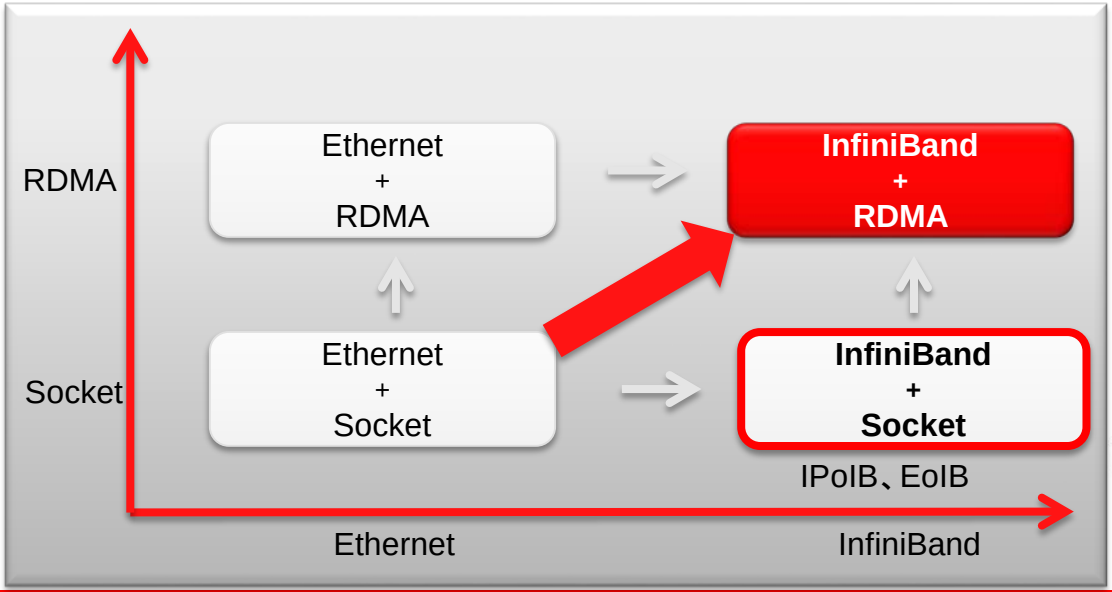


RDMAソリューション Over Ethernet

- RoCE(ロッキー)
  - RDMA over Converged Ethernet
  - Soft HCA
    - SoftwareでRDMAを実装
    - IntelやBroadcomの標準NICでも動作可能(CNAでも動作可)
    - IBのHCA(Host Channel Adaptor)では通常HWでRDMA処理
    - Mellanox社の10/40 GENICではHWでRDMA処理
  - OFED1.5(2010年)より実装
- iWARP(アイワープ)
  - Internet Wide Area RDMA Protocol
  - RDMA over TCP
    - TCP Offload Engine(TOE)
    - Chelsio社の10/40GE NICではHWでRDMA処理

OracleがリードするInfiniBand 技術

- エンジニアド・システム（RDMAにより最適化）
  - データベース層の最適化：**Exadata**
  - アプリケーション層の最適化：**Exalogic**
  - Unix環境の最適化: **SuperCluster**



- 仮想化
  - **OVCA**
  - **Oracle Virtual Networking**

リンクレイヤー

IBパケット・フォーマット

- Local Routing Header(LRH): 宛先LID、送信元LID、サービスレベル(VL)を指定。
- Transport Header(TH): 宛先QP番号、パケットシーケンス、オペレーションコード(Opcode)を指定など。
- Invariant CRC(ICRC): ファブリック内で不変のCRC(GRH以外を対象)
- Variant CRC(VCRC): GRHも含め対象
- 最大パケット長は2Kバイト。(オプションで最大4Kバイトまで拡張)

- Global Routing Header(GRH): 異なるサブネット間でのルーティング。RouterはVCRCを再計算。

ノードアドレス≒MACアドレス

- すべてのデバイスおよびポートは、*Globally Unique Identifiers (GUID)*というグローバルでユニークなIDが割り付けられる
  - 64-bit アドレス
    - 例 GUID: *0x0013970102000157*
  - GUIDの前半部分は、ハードウェアベンダーの情報となる
- 各ポートは、*Local Identifier (LID)*がダイナミックにアサインされる
  - 16-bit長
    - Unicast LID 0x0001-0xBFFF = 48K アドレス
    - Multicast LID 0xC000-0xFFFE = 16K アドレス
    - Permissive LID 0xFFFF ディスカバリプロセスで使用

ノードアドレスの割り当て

- サブネットマネージャ（SM）から動的にアサインされる
- エンドポイントのアドレスとして使用される
  - スイッチもHCAもLIDを持つ
- 通常は、サブネット内で1から順にアサインされる
  - 100程度のノード数では、容易に把握できる
- 通常リブート後にはSMがキャッシュしているリブート前のアドレスをアサイン

サブネット・マネージャ (SM)

- Infinibandファブリックは最低ひとつのSM
- 網内に複数のSMが存在する場合
  - ひとつのSMが「Master」、ほかのSMは「Standby」となる
- ファブリック内のどこでもSMを配置できる
  - ノード、スイッチ、スペシャルデバイスなど、SMの場所はどこでもよい
- SM とSMA (サブネット・マネージャ・エージェント)
- すべてのIBデバイスはサブネット・マネージャ・エージェントを持つ

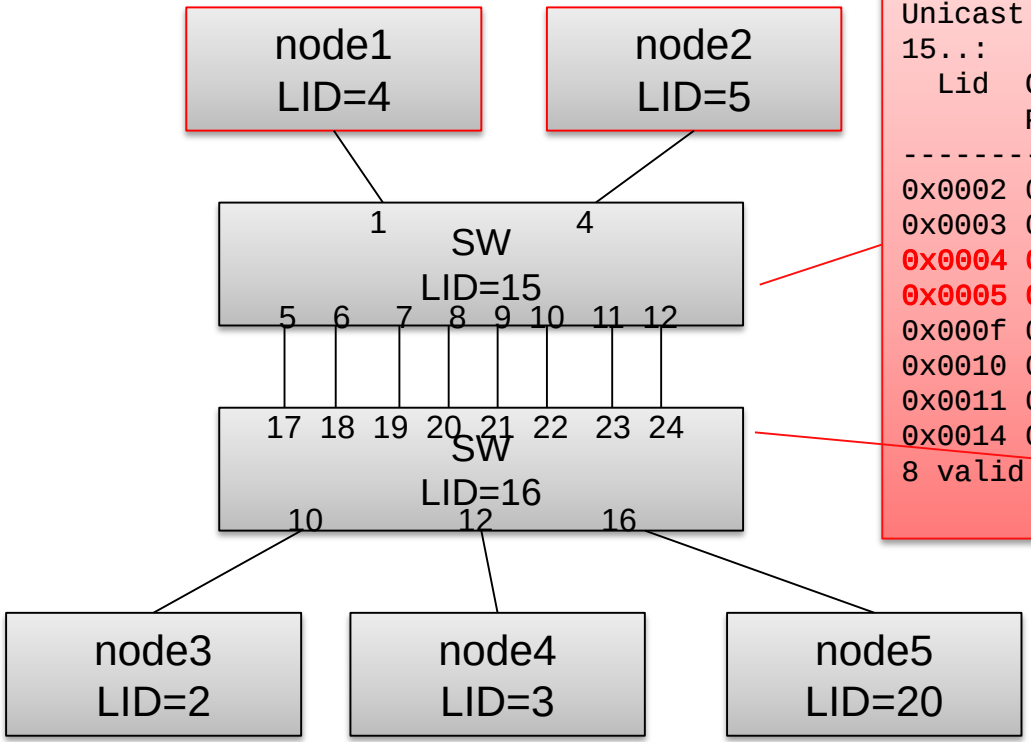
サブネット・マネージャ (SM)続き

- SMはマネージメント・データグラム・パケット（MAD）をSMAへ送信する
- SMAは、ローカルのステータスの変更を通知する際に、TrapをSMへ送信する
- SM がサブネット・トポロジーを管理
  - NodeInfo
  - portInfo
  - switchInfo
  - GUIDInfo
  - ForwardingTable、
  - LinkInfoなど
- サブネットのトポロジーとPathInfoを作成

パケット・フォワーディング

	Ethernetスイッチ	InfiniBnadスイッチ
テーブル管理	パケットが到着した時点で、送信元MACアドレスをフォワーディング・テーブルに登録	ノードが起動した時点で、SMから全スイッチにLIDを登録。 SMではMin-Hopアルゴリズムで最短経路が計算。
Unknown パケット	Unknownユニキャストはフラッディング	SMからLIDを取得するので、Unknownユニキャストは無
マルチキャスト	Multicastは、IGMP-Snoopingでは必要なポートのみ転送	MLIDで必要なポートのみ転送。SMにより管理
ブロードキャスト	ブロードキャストはフラッディング	FFFFはPLIDで予約のため、無
転送方式	ストア・アンド・フォワーディングの場合、パケット長に応じて転送遅延増大	カットスルー方式

ユニキャスト・フォワーディング・テーブル

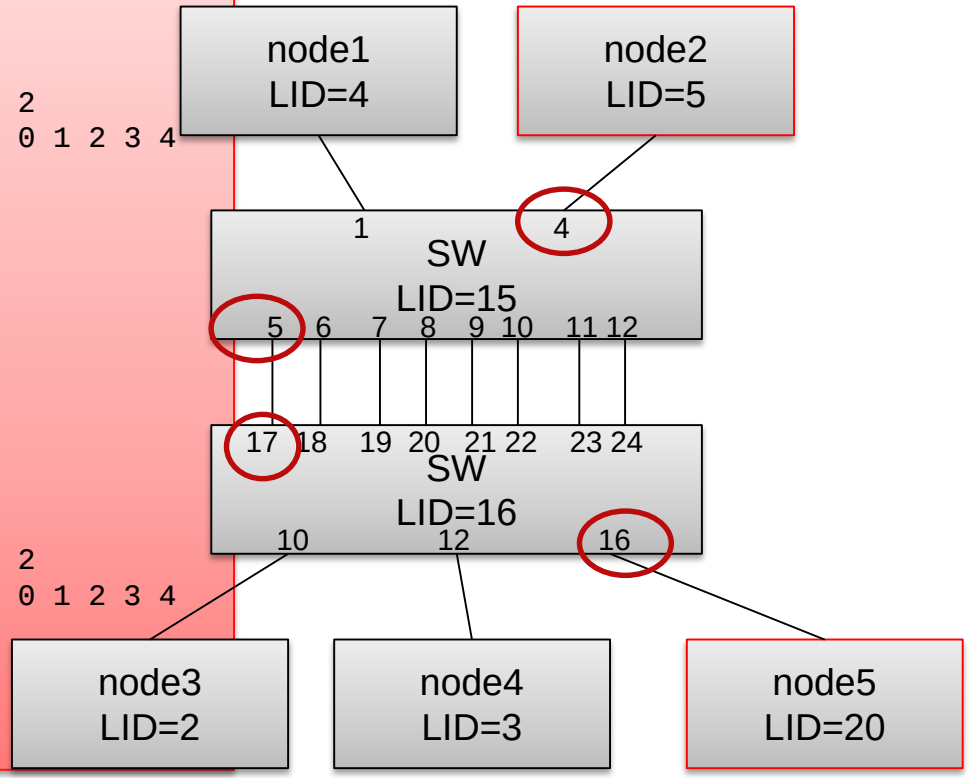



```
~# ibroute 15 -n
Unicast lids [0x0-0x14] of switch Lid
15..:
  Lid  Out   Destination
       Port     Info
------------------------
0x0002 006
0x0003 005
0x0004 001
0x0005 004
0x000f 000
0x0010 008
0x0011 013
0x0014 007
8 valid lids
```

```
~# ibroute 16 -n
Unicast lids [0x0-0x14] of switch Lid
16:
  Lid  Out   Destination
       Port     Info
--------------------------
0x0002 012
0x0003 010
0x0004 017
0x0005 018
0x000f 019
0x0010 000
0x0011 019
0x0014 016
8 valid lids dumped
```

ユニキャスト・フォワーディング・テーブル

```
vp780p:~# ibroute -M 15
Multicast mlids [0xc000-0xc3ff] of switch Lid 15 …:
           0                   1                   2
     Ports: 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4
 MLid
0xc000              x x
0xc001              x x
0xc002              x
0xc003              x
0xc004              x
0xc005              x
0xc006              x
7 valid mlids dumped
vp780p:~# ibroute -M 16
Multicast mlids [0xc000-0xc3ff] of switch Lid 16…:
           0                   1                   2
     Ports: 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4
 MLid
0xc000                                      x x
0xc001                                      x x
2 valid mlids dumped
```

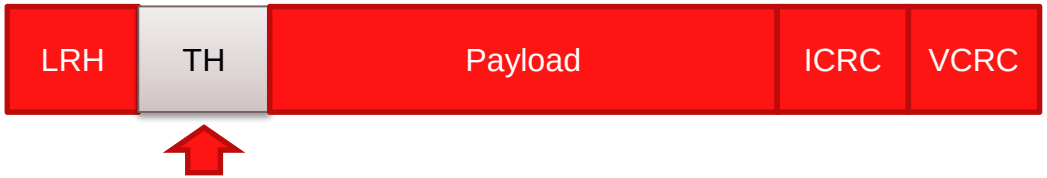


冗長化

- Ethernetで使われるスパニングツリーのような仕組みは不要
  - SMで最短経路が計算され、SWに渡される
  - 障害時には、TrapがSMへ送られて、即時に経路情報が再計算
- LAG、LACPなどの設定不要
  - トポロジーは、SMで管理されており、設定不要
  - 同一HopのMultipathは、ランダムに割り当て

フロー制御

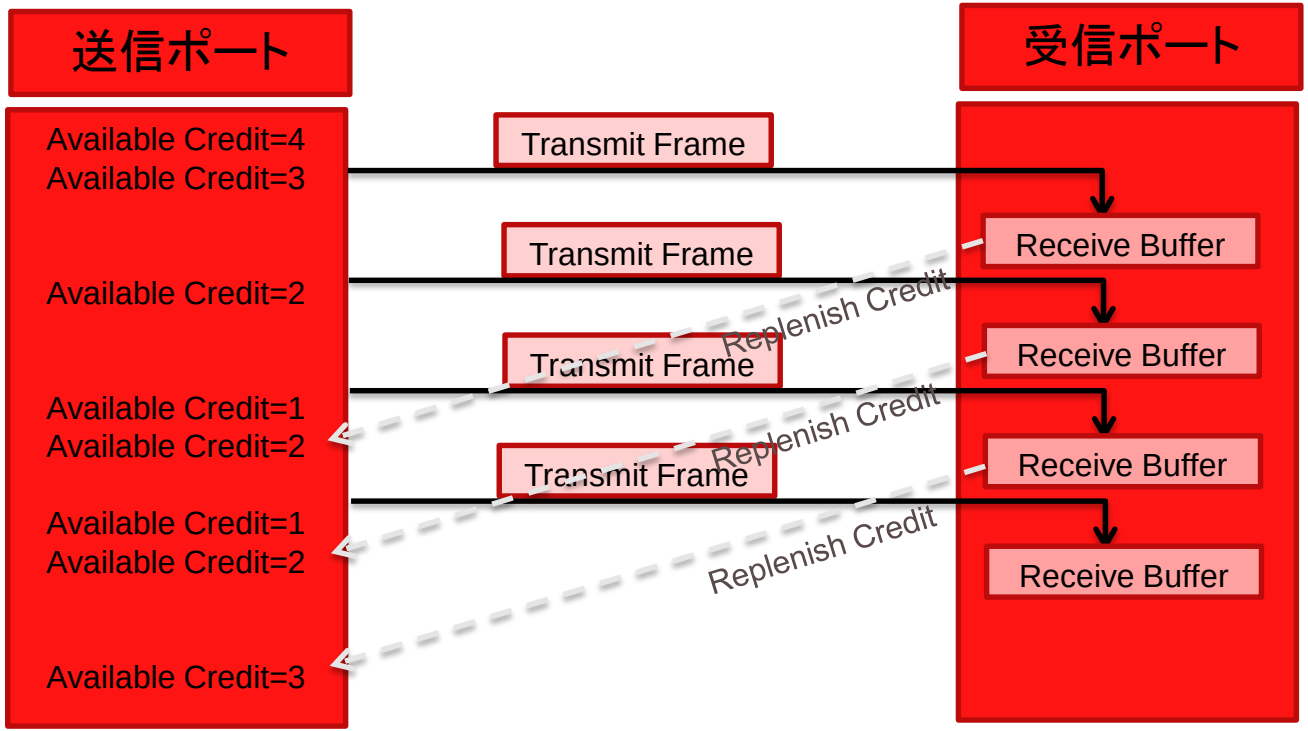
- トランスポートサービス



  - Reliable Connection(RC)
  - Unreliable Connection(UC)
  - Reliable Datagram (RD)
  - Unreliable Datagram (UD)
- Relieble サービスでは、データは保証される。=ロス・レス

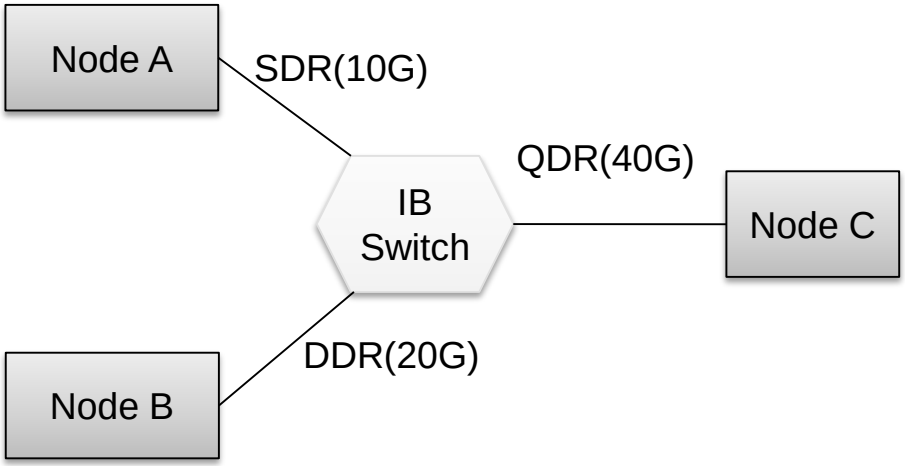
フロー制御 続き

- クレジットベースのフロー制御
  - FCで使われている信頼性の高いフロー制御
  - Ethernetでは Pause Frame による制御



フロー制御 続き

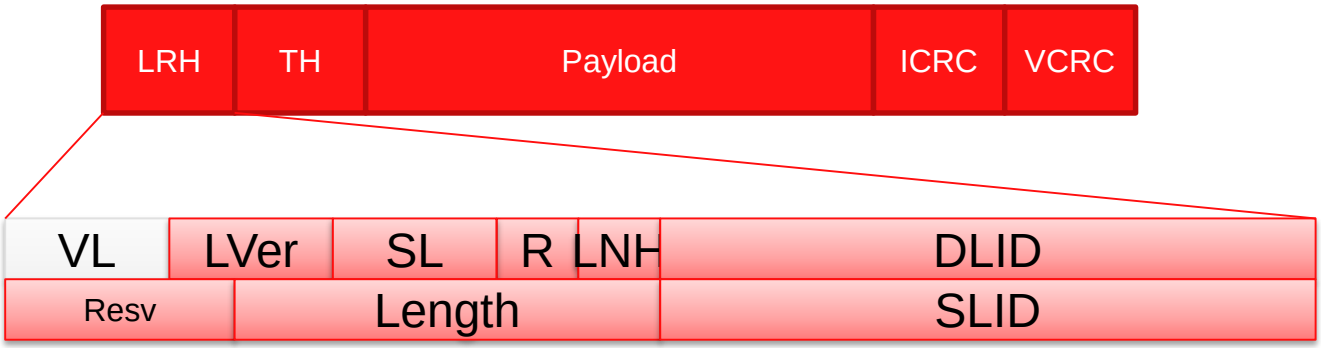
- Inter Packet Delay (IPD)
  - 異なるリンクスピードとの通信においては、送信パケット間に適当なDelayが指定され、バッファのオーバフローを防ぎます。



IPD	rate	Comment
0	100%	Suited for matched links
1	50%	
2	33%	Suited for 30 Gbps to 10 Gbps conversion
3	25%	Suited for 10 Gbps to 2.5 Gbps conversion
11	8%	Suited for a 30 Gbps to 2.5 Gbps conversion

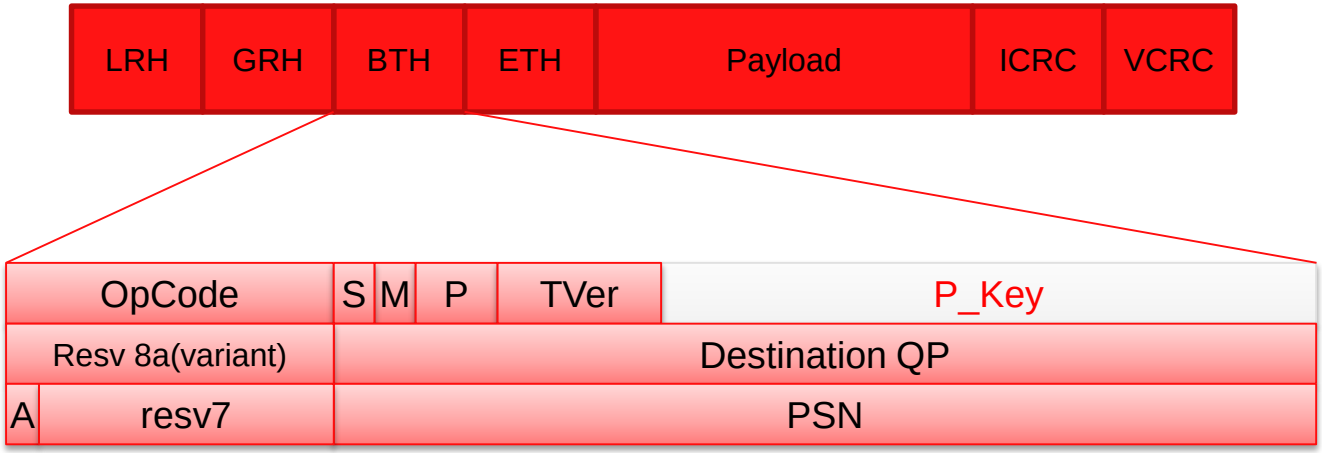
QoS、CoS

- バーチャル・レーン（4ビット）
  - 物理リンク毎に、最大16バーチャル・レーン（VL)を持ち、キューイングが行われ、Head-of-Line Blockingを防ぐ。各VLでそれぞれのバッファスペースを持つ。VL15はSMP専用となる。
  - ≒802.1p（3ビット）



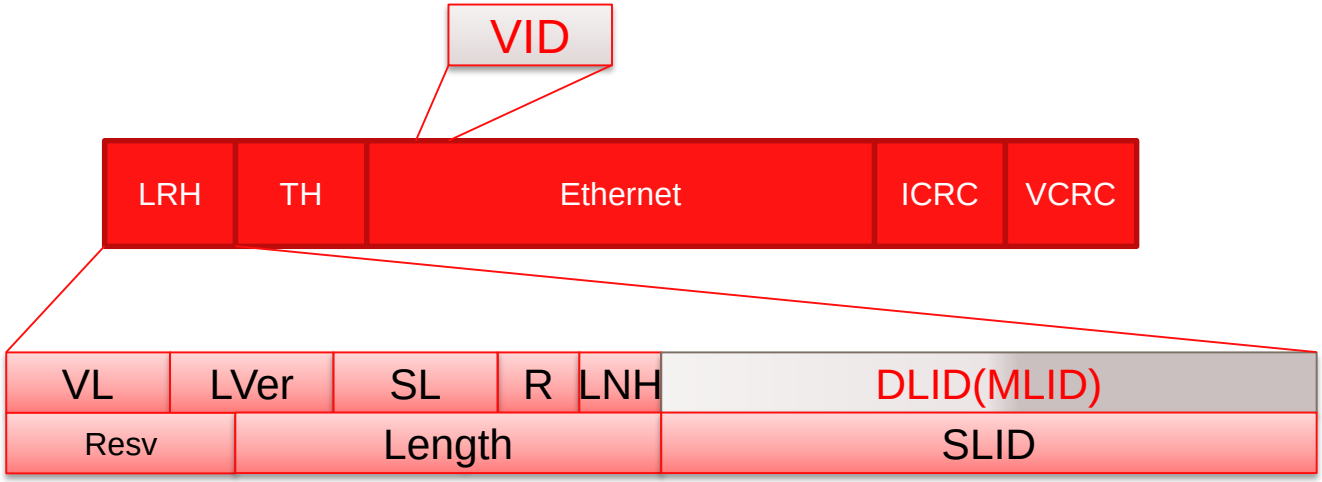
VLAN的なもの～パーティショニング

- Partition Key(P_Key): 16ビット
  - 最上位ビットはメンバーシップ・タイプに使用。InfiniBandファブリック内を分割。P_Keyがマッチしたパケットのみを受信。P_KeyテーブルはPartition Managerによりコントロール。



Oracle SDN

- Multicast LID(MLID): 16K アドレス
  - MLID単位でEthernetブロードキャストドメインを分割
- EthernetのVLAN ID(4K VLAN)も併用可能



データセンターへの適用

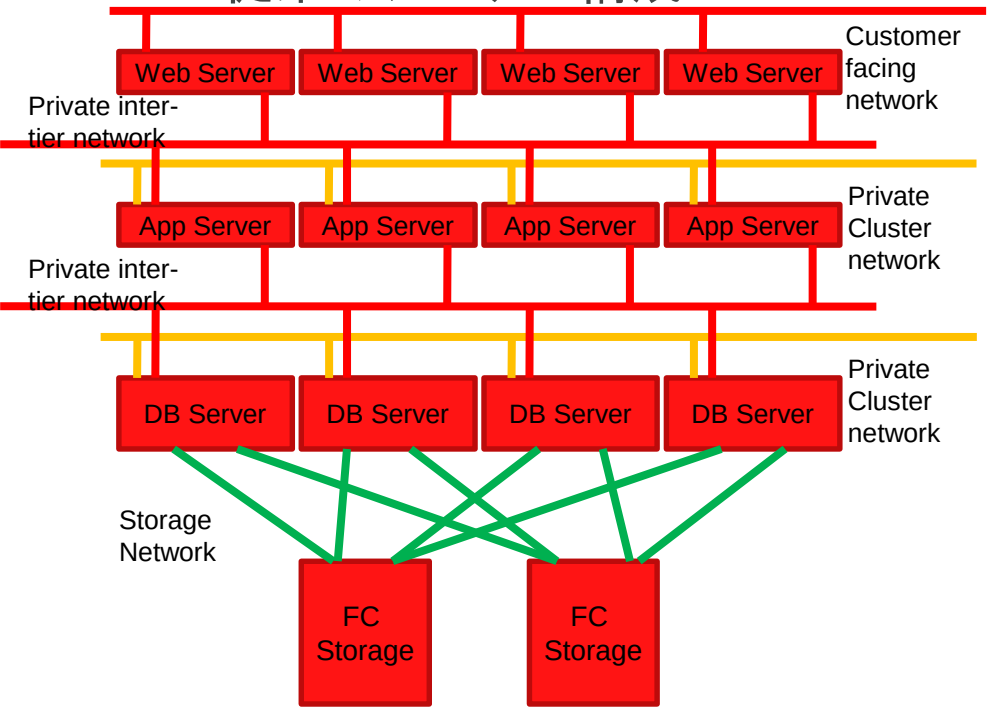
既存インフラとの混在

- Exadata、Exalogic → RDMA + IPoIB + EoIB
- IPoIB
  - OFEDで標準で提供
  - Bonding (=NIC Teaming)サポート
- EoIB、FCoIB
  - サーバとIO Chassis間をInfiniBandでトンネリング
  - GE、10GE、8G FCへの接続可能

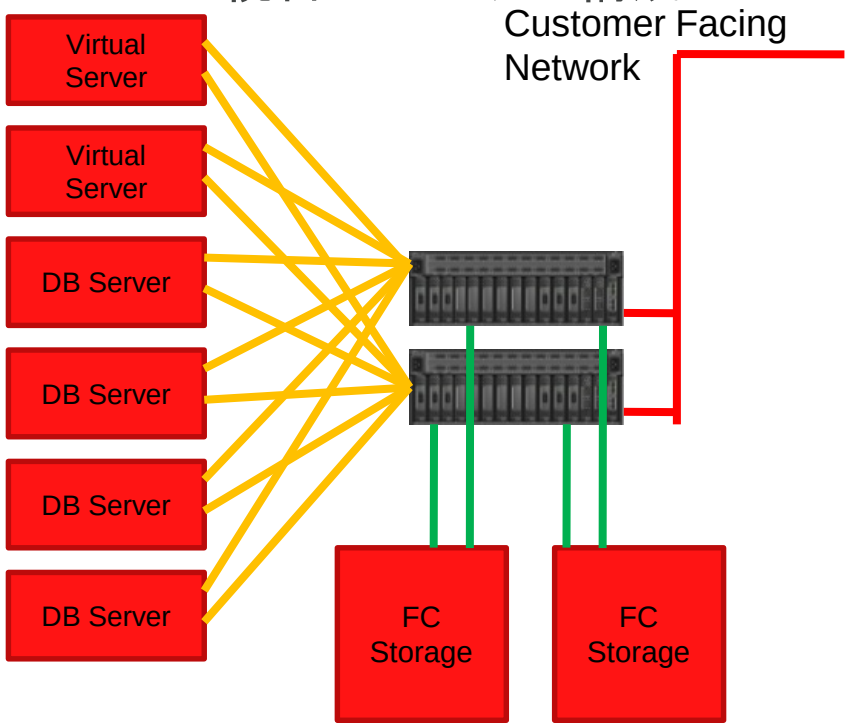
I/O統合

3 tier application stack

従来のシステム構成

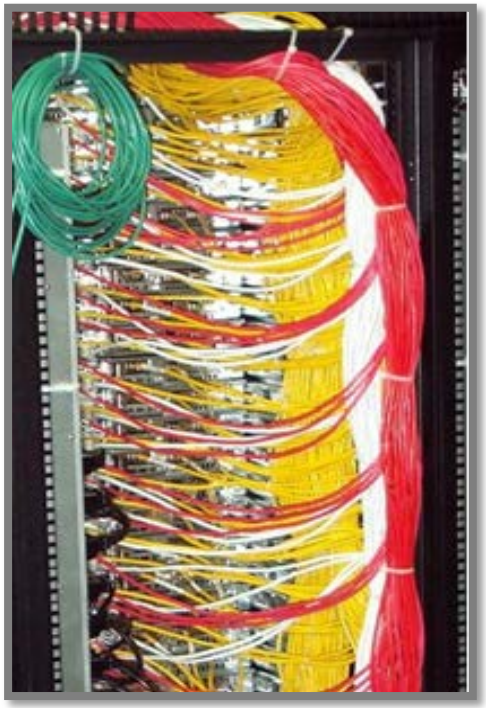


I/O統合のシステム構成

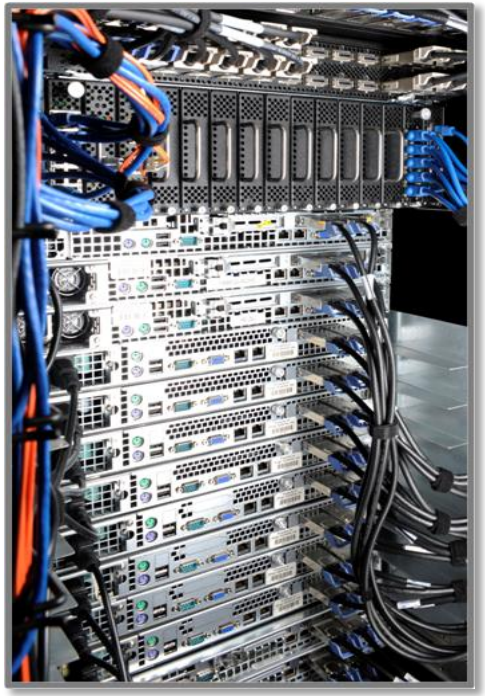


ケーブル数の削減

統合前

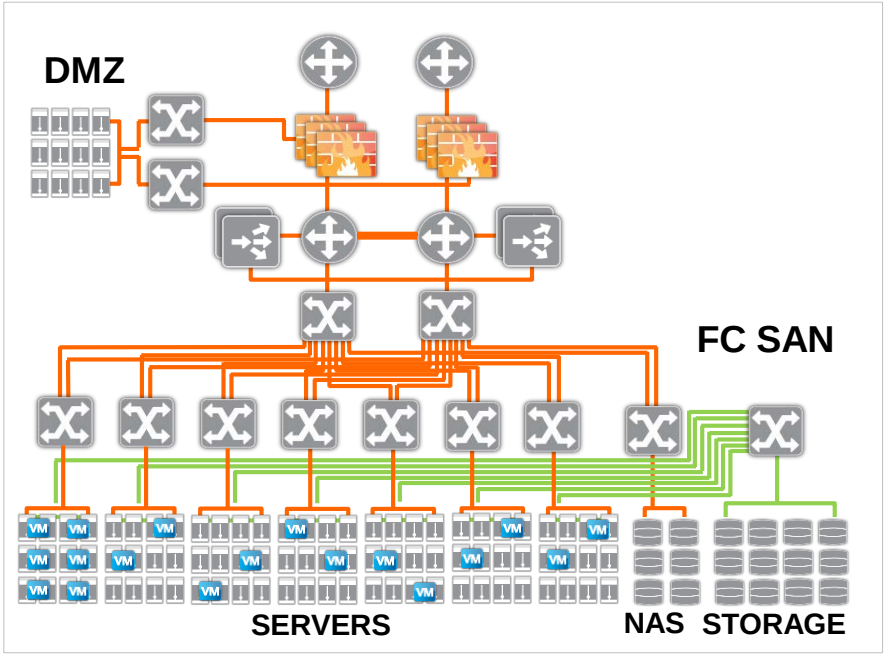


統合後

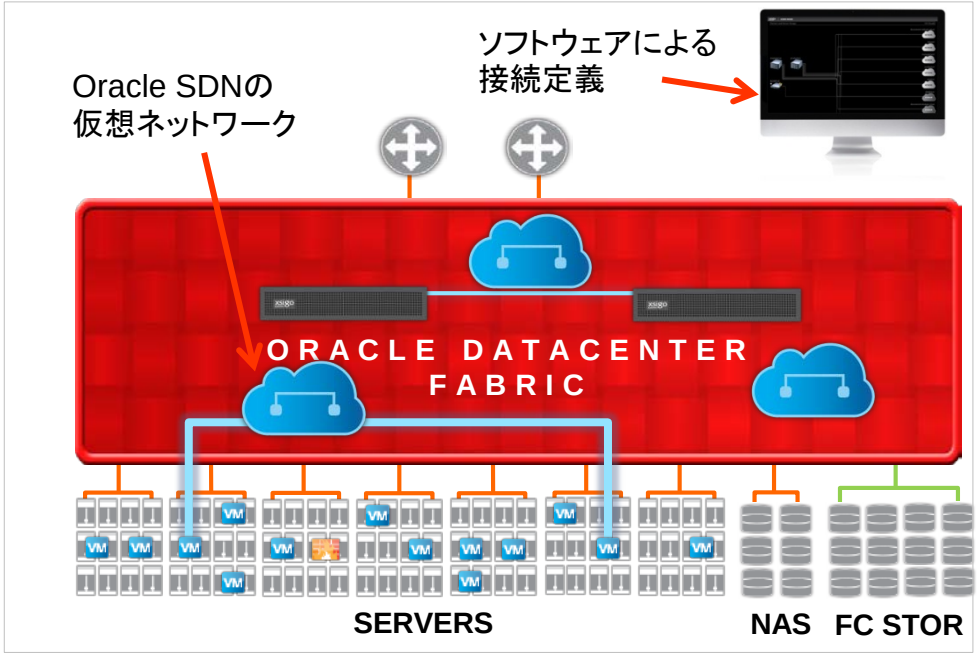


Oracle仮想ネットワーキングによる仮想化基盤

従来の共通基盤



オラクルデータセンターファブリック

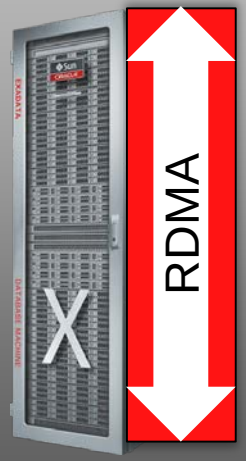


Oracle製品でのInfiniBandの活用

Workload Centric

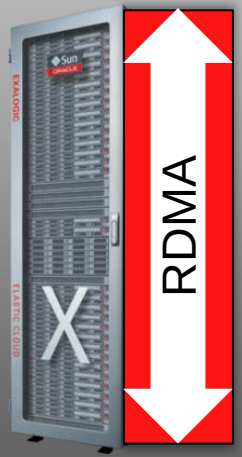
General Purpose

オラクルDB統合基盤



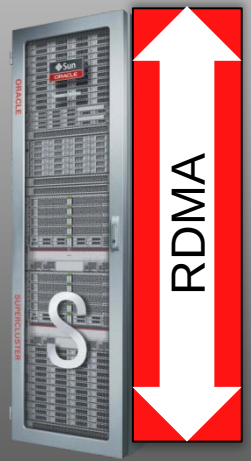
Exadata

Java アプリケーション 実行基盤



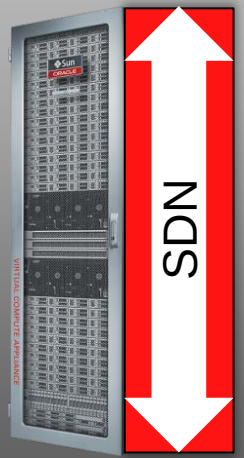
Exalogic

アプリケーション& DB統合基盤



SuperCluster T5-8

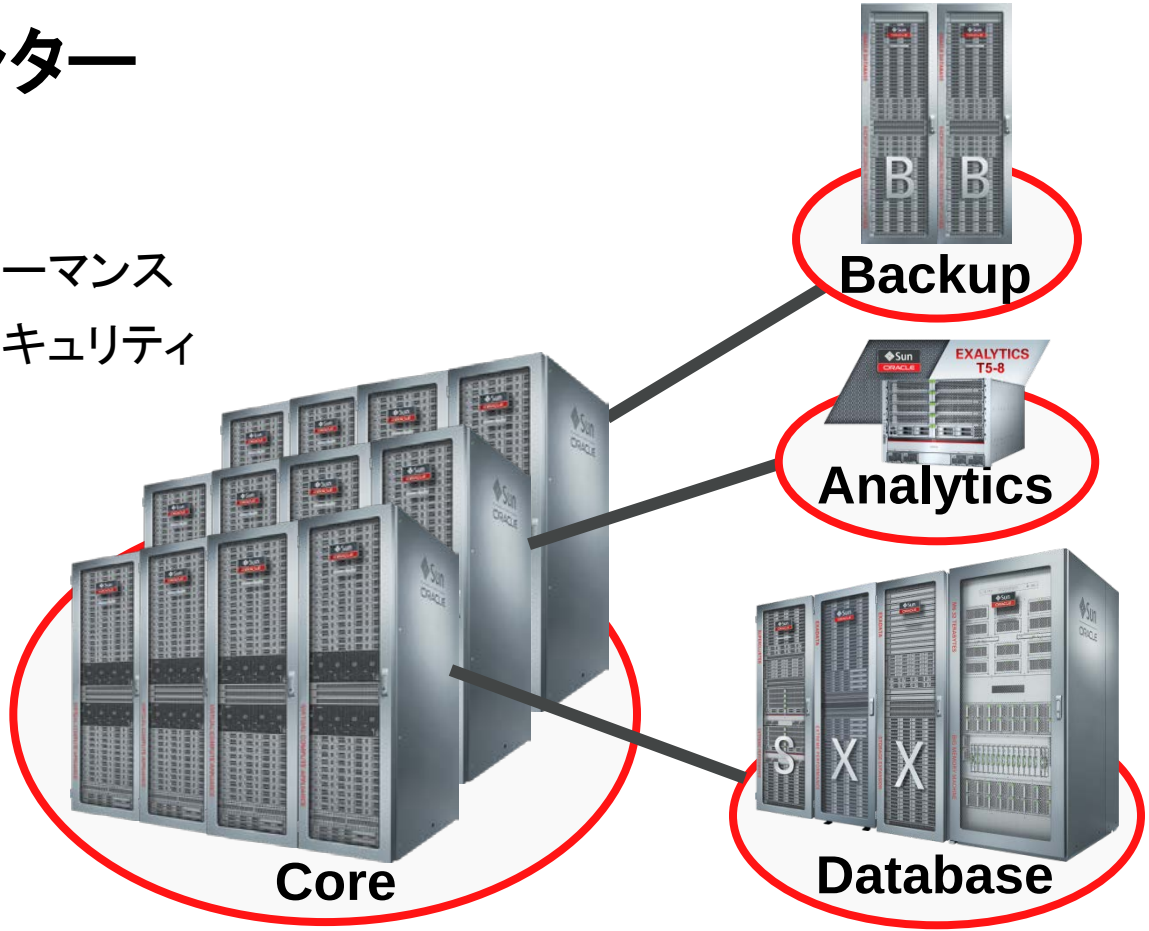
オープン系 アプリケーション 仮想化統合基盤



Oracle Virtual Compute Appliance(OVCA)

今後のデータセンター

- 目的毎の専用マシン
  - ベスト・コスト/パフォーマンス
  - より高い信頼性とセキュリティ
  - 迅速なと導入
- コア
  - Intel サーバ
  - 仮想化Linux
  - Ethernet接続

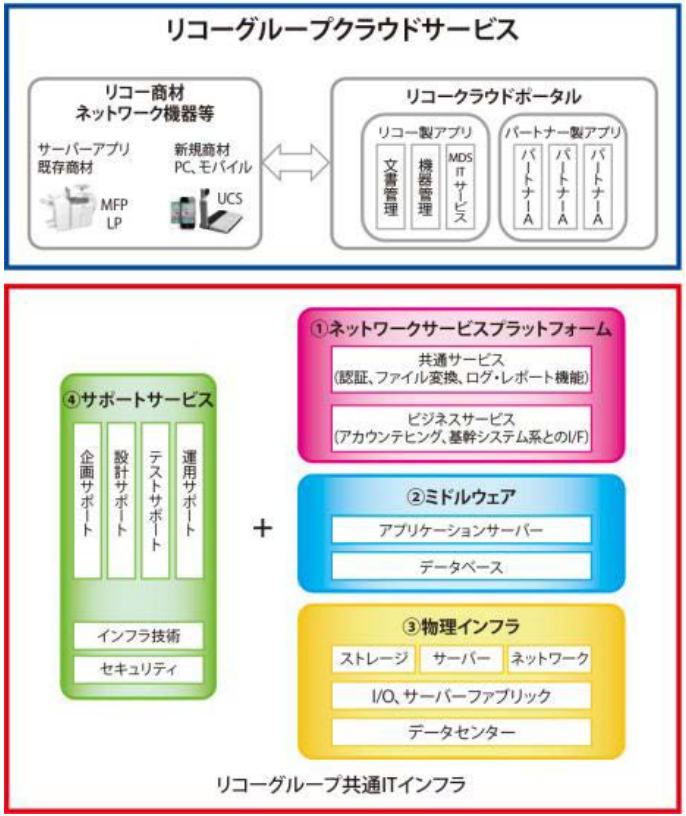


お客様事例

リコー：シンプルで柔軟なグループ共通インフラ

Solutions

- 【背景】SaaSビジネスの展開を可能にするサービスのグループ共通インフラ
  - 【課題】将来のサービス要件は、基盤構築時点では予測困難
- Oracle Virtual Networkingにより、I/Oを仮想化
- 需要に応じた迅速な展開が可能に
- 機器コスト、運用コストを削減



まとめ

- InfiniBandは、下位レイヤと上位のRDMAレイヤで構成
- 下位レイヤは、非常にシンプルかつ高速
- RDMAにより、アプリケーションを高速化
- 既存アプリケーションの移行において、InfiniBandで統合

Hardware and Software

ORACLE

Engineered to Work Together


ORACLE®