

Internt Week 2011
InfiniBand

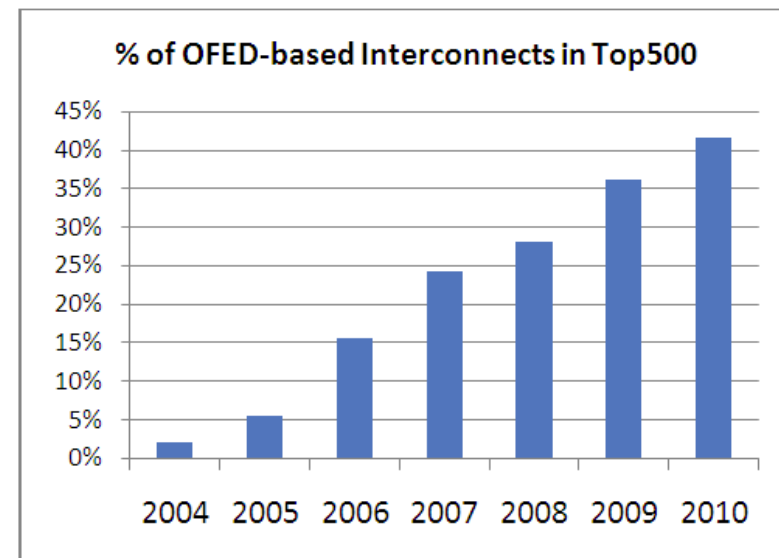
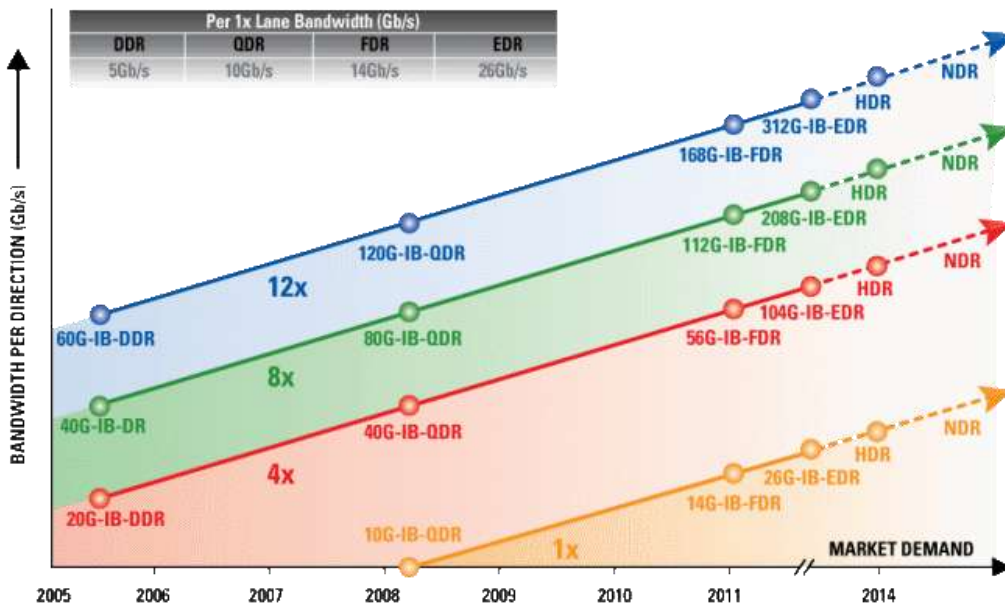
2011年12月

- InfiniBand基礎の理解
- EthernetでいうところのXXXは、InfiniBandではどうなっているの？
- データセンターでのInfiniBandの利用

- InfiniBand概要
- Ethernetとの機能比較
- データセンタへの適用

InfiniBand

- InfiniBand® Trade Association (IBTA)により、策定されたサーバ・ストレージなどインフラを接続するファブリック技術。
- 広帯域・低遅延・低消費電力・コストパフォーマンスのアドバンテージ
- スーパーコンピュータ/HPC、金融分野で普及
- ストレージ・バックエンド、データベースクラスタで実装



- データレート

- Single Data Rate (SDR), 10Gbps (実レートは 8Gbps)
 - Double Data Rate (DDR), 20Gbps (実レートは16Gbps)
 - Quad Data Rate (QDR), 40Gbps (実レートは32Gbps)
- エンコードに8B/10Bが使われるため、実効レートが80%
- 今後登場する規格ではエンコードが64B/66B
- Fourteen Data Rate (FDR), 56Gbps (実レートは54Gbps)
 - 2012年以降に予定されているEDR 100Gbps (実効レート96Gbps)

- InfiniBand デバイス

- ホスト・チャンネル・アダプタ (HCA)
 - ホスト側の接続デバイス
- ターゲット・チャンネル・アダプタ (TCA)
 - I/O ターゲット側のデバイス
- スイッチ

- DDR(20G)の場合:

- IPoIBで、ほぼワイヤーレート(16Gbps)

```
# iperf -c 7.7.7.2 -P 8
```

```
-----
```

```
Client connecting to 7.7.7.2, TCP port 5001
```

```
...
```

```
[SUM] 0.0-10.0 sec 18.2 GBytes 15.7 Gbits/sec
```

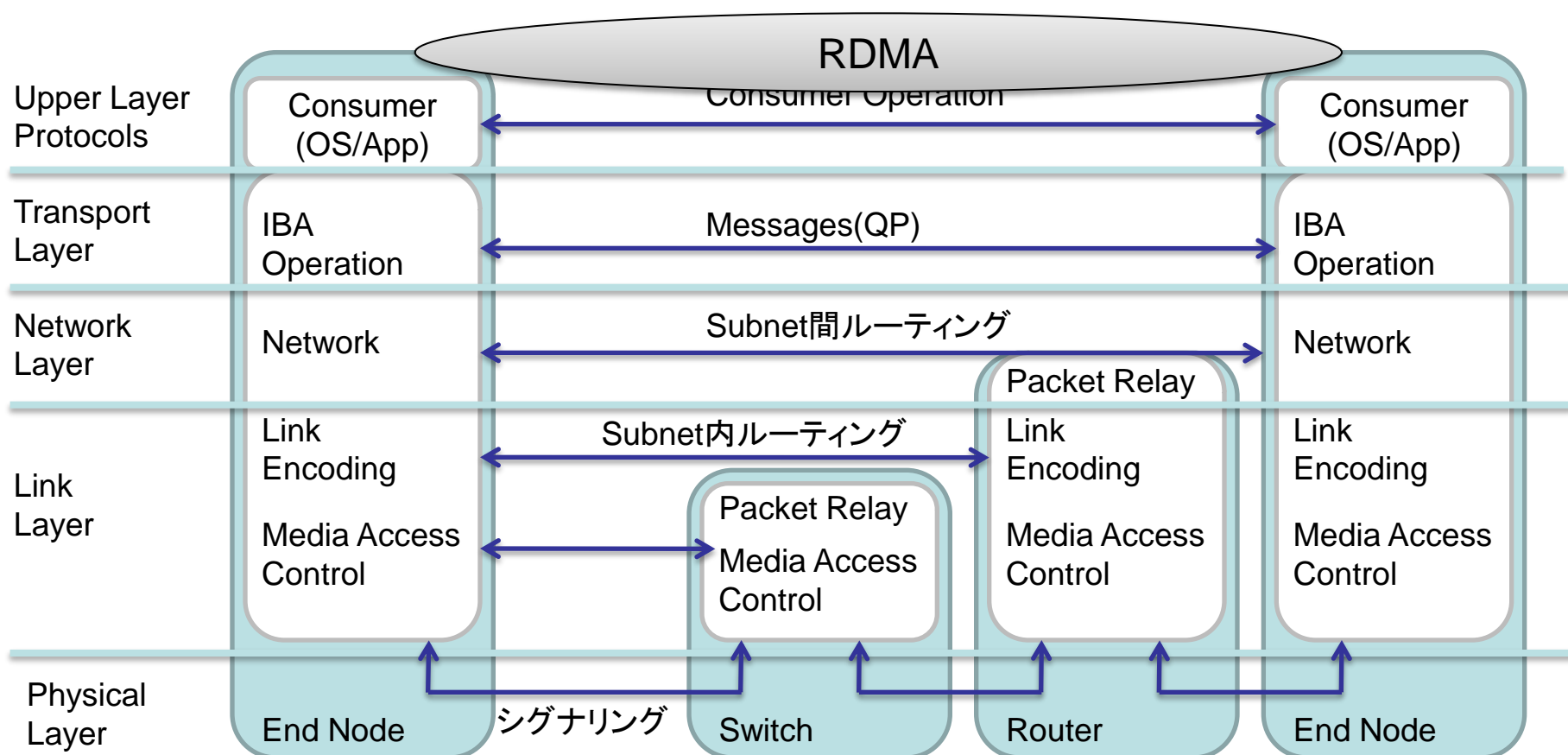
- QDR(40G)の場合では、26-27Gbps

- PCIe Gen2の帯域が上限

- FDR(56G)には、PCIe Gen3が必要

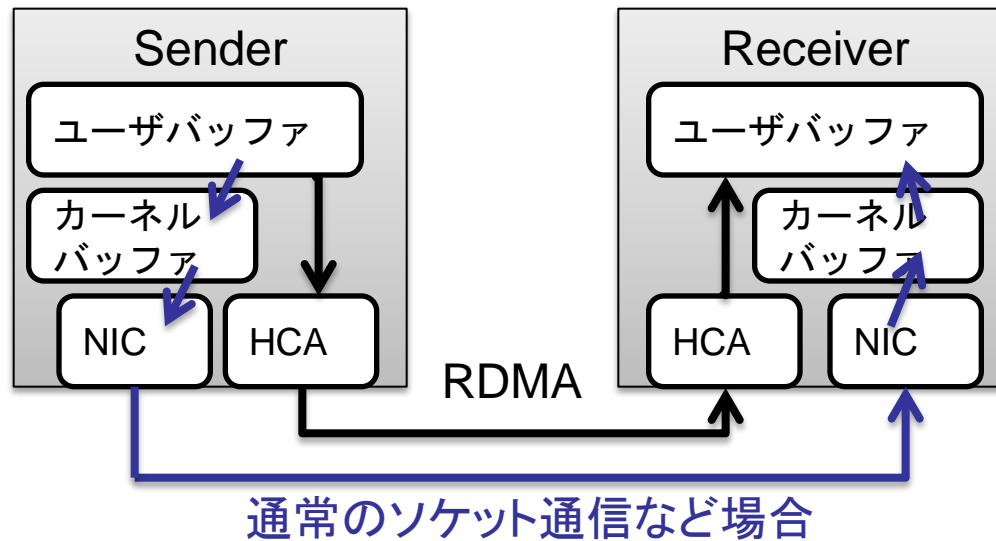
レイヤー・アーキテクチャ

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



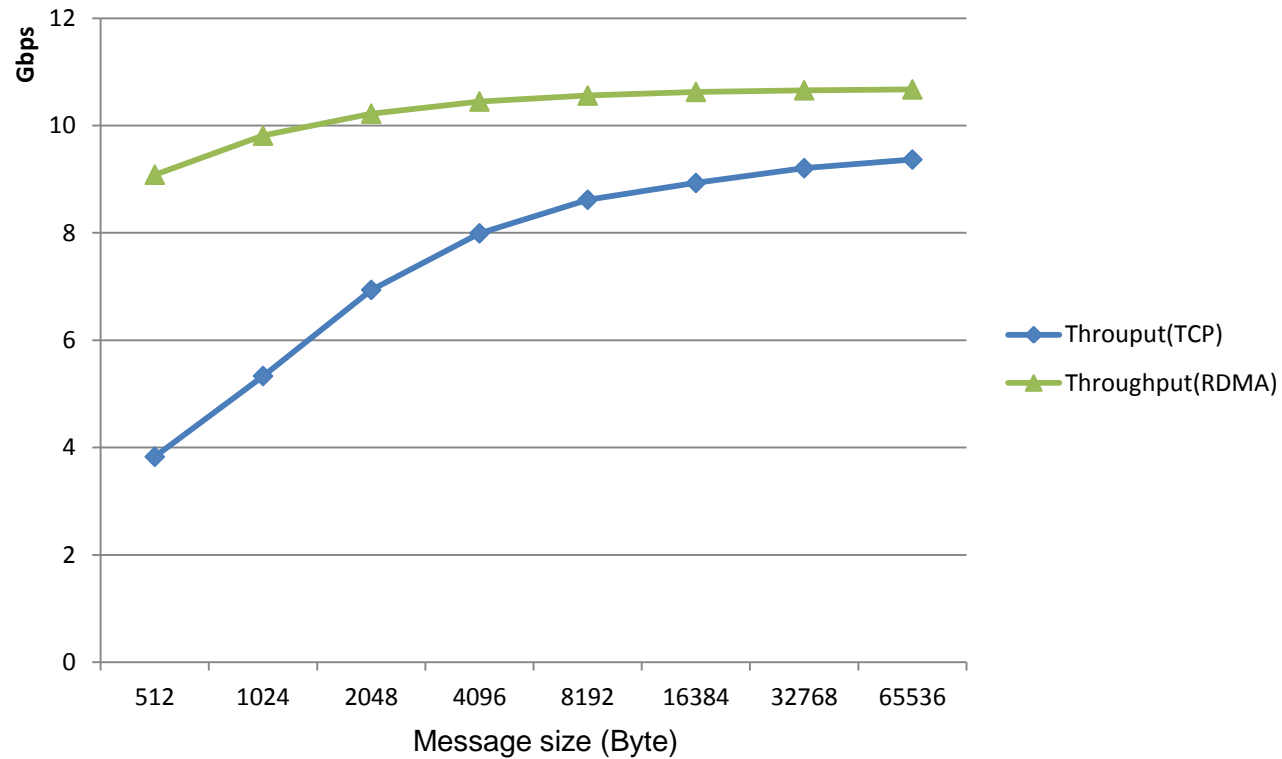
RDMA技術とは？

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



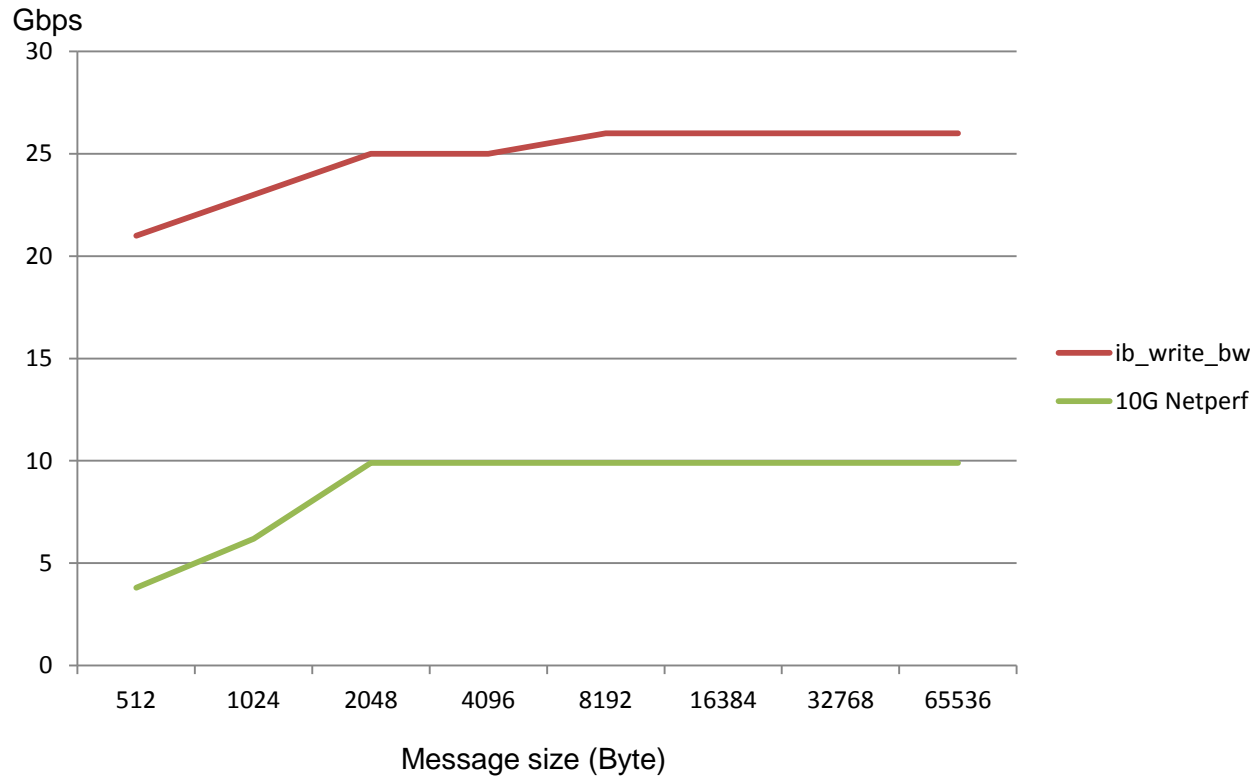
TCP vs RDMA

少し古いサーバでのNetperf(TCP) vs ib_write_bw(RDMA)



TCP vs RDMA

10GE Netperf(TCP) vs QDR ib_write_bw(RDMA)



- RoCE(ロッキー)

- Ethernet上でRDMAを実装

- Soft HCA

- Software(ドライバ)でRDMAを実装

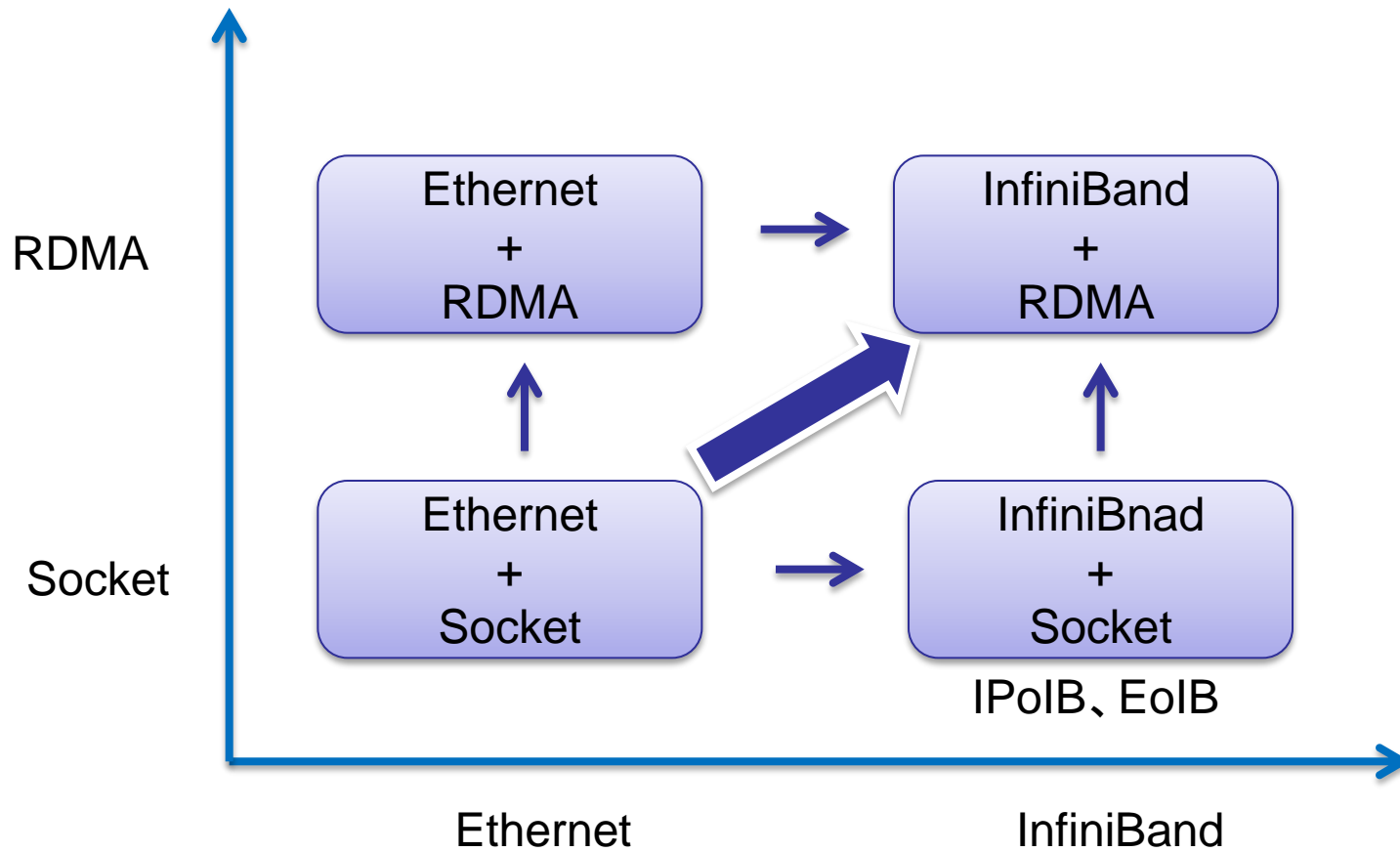
- IntelやBroadcomの標準NICでも動作可能(CNAでも動作可)

- IBのHCA(Host Channel Adaptor)では通常HWでRDMA処理

- Mellanox社の10GE NICではHWでRDMA処理

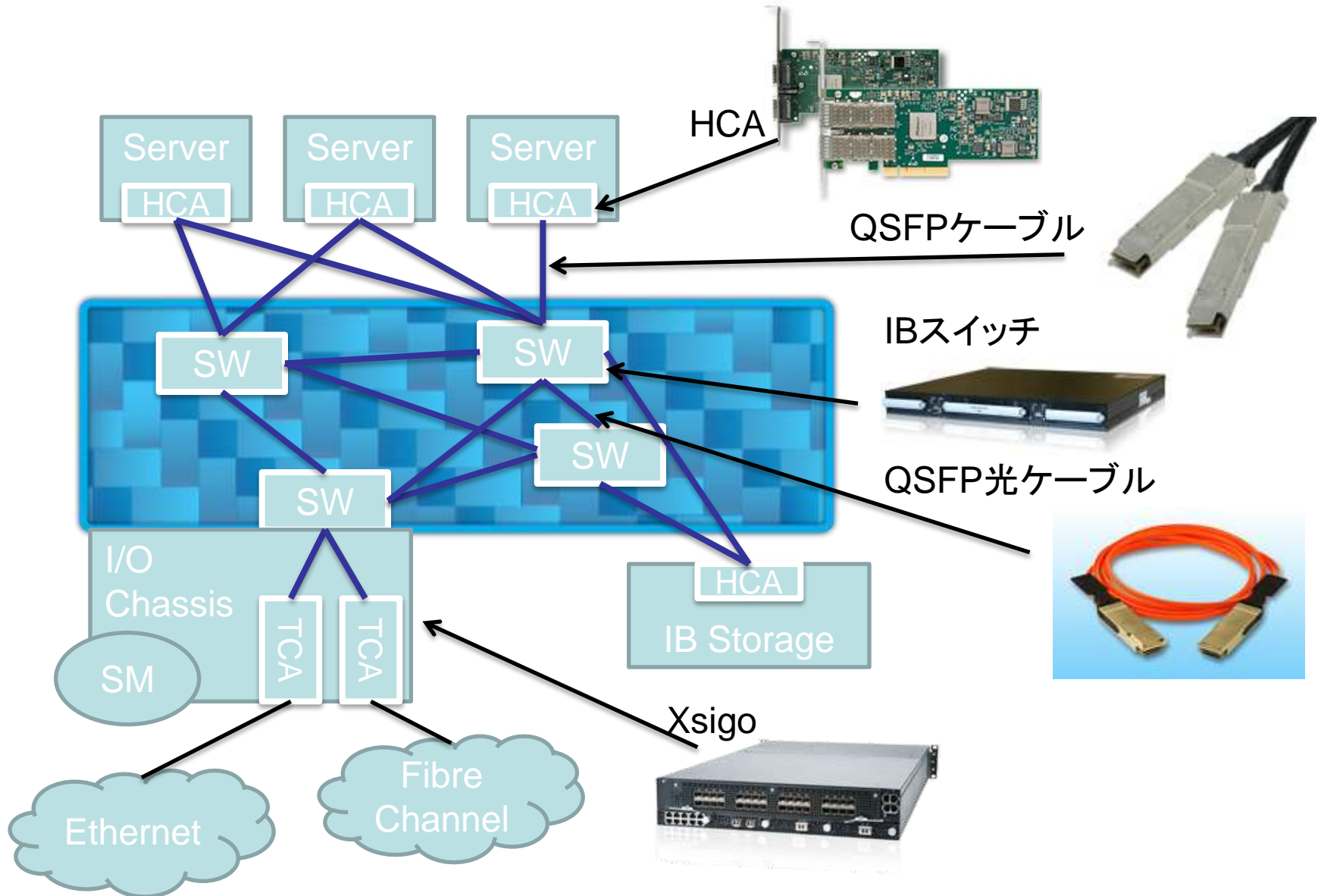
- OFED1.5(2010年)より実装

InfiniBandとRDMA



- InfiniBand概要
- Ethernetとの機能比較
- データセンタへの適用

InfiniBand ファブリック

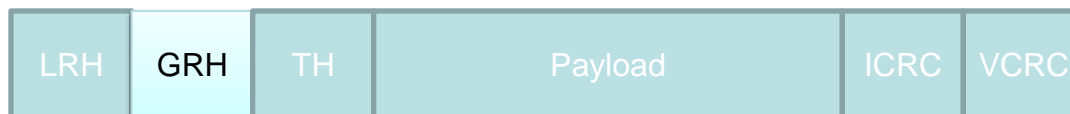


IBパケット・フォーマット



- 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 Packet:



- 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

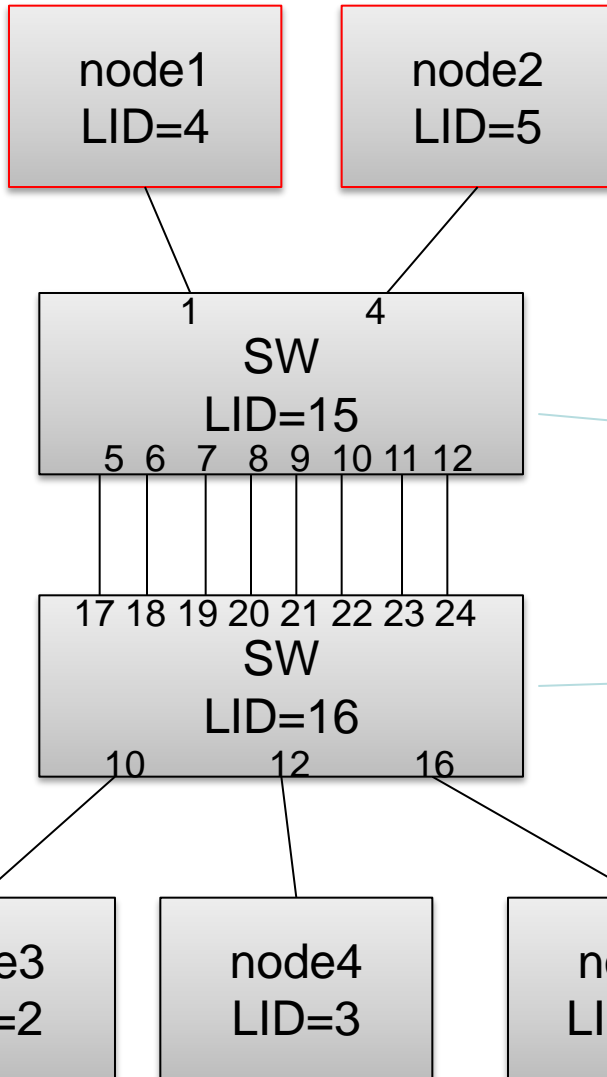
- すべてのデバイスおよびポートは、*Globally Unique Identifiers (GUID)* というグローバルでユニークなIDが割り付けられる
 - 64-bit アドレス
 - 例 GUID: 0x0013970102000157
 - GUIDの前半部分は、ハードウェアベンダーの情報となる
- 各ポートは、*Local Identifier (LID)*がダイナミックにアサインされる
 - 16-bit長
 - Unicast LID 0x0001-0xBFFF = 48K アドレス
 - Multicast LID 0xC000-0xFFFFE = 16K アドレス
 -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

Ethernetスイッチ	InfiniBnadスイッチ
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
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
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
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



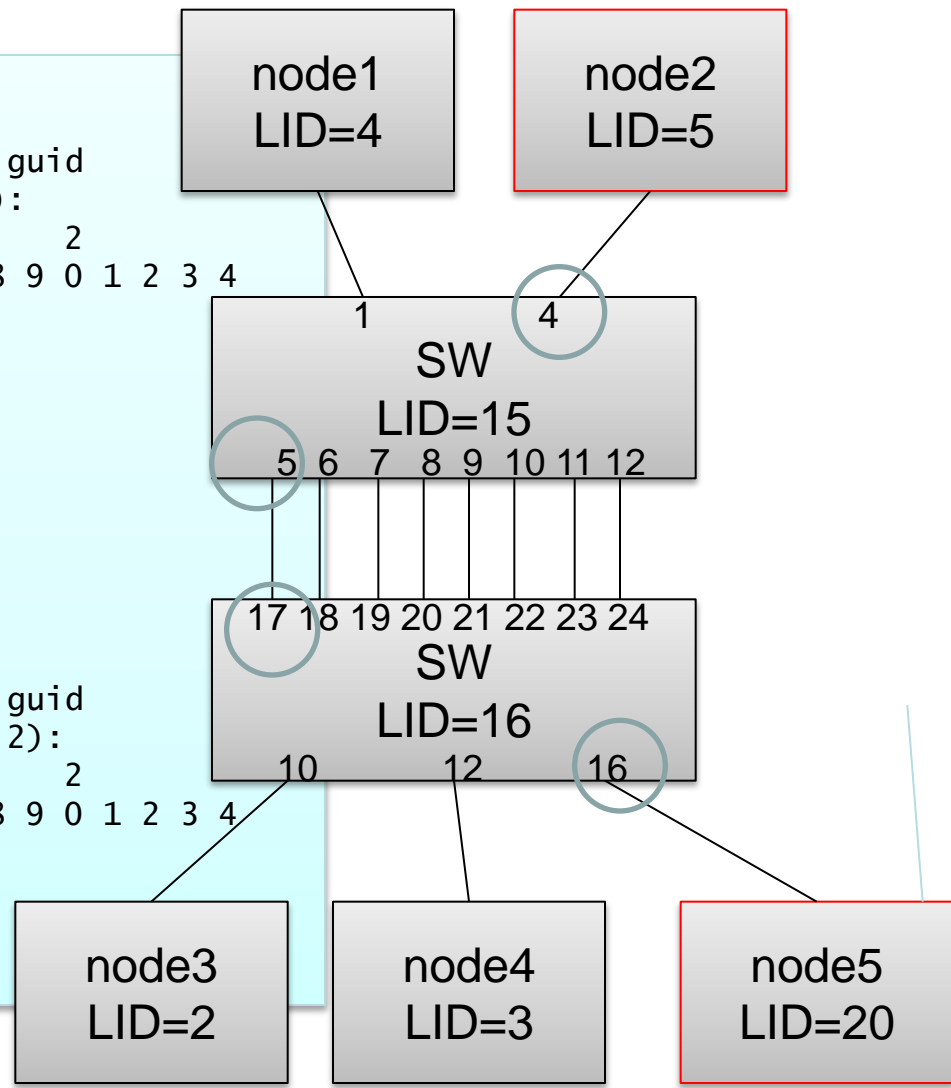
```
~# ibroute 15 -n
Unicast lids [0x0-0x14] of
switch Lid 15..:
  Lid  Out  Destination
      Port  Info
-----
0x0002 006
0x0003 005
0x0004 001
0x0005 004
0x000f 000
0x0010 008
0x0011 013
0x0014 007
8 valid lids d
```

```
~# 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
```

マルチキャスト・フォワーディング・テーブル

```

vp780p:~# ibroute -M 15
Multicast mlids [0xc000-0xc3ff] of switch Lid 15 guid
0x0013970101000175 (Xsigo Core: INI Ver: 1.0.0.2):
      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 guid
0x0013970102000175 (Xsigo Leaf 1: INI Ver: 1.0.0.2):
      0          1          2
Ports: 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
    
```



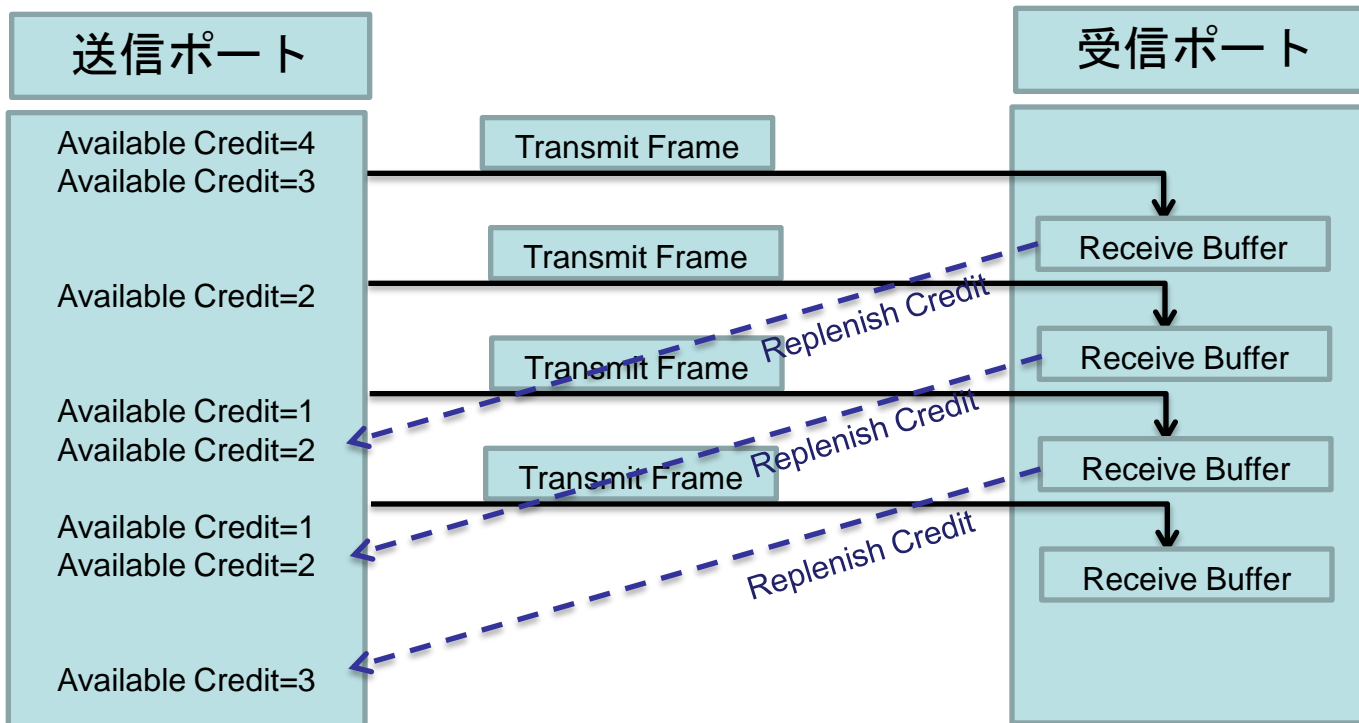
- Ethernetで使われるスパニングツリーのような仕組みは不要
 - SMで最短経路が計算され、SWに渡される
 - 障害時には、TrapがSMへ送られて、即時に経路情報が再計算
- LAG、LACPなどの設定不要
 - トポロジーは、SMで管理されており、設定不要
 - 同一HopのMultipathは、ランダムに割り当て

- トランスポートサービス



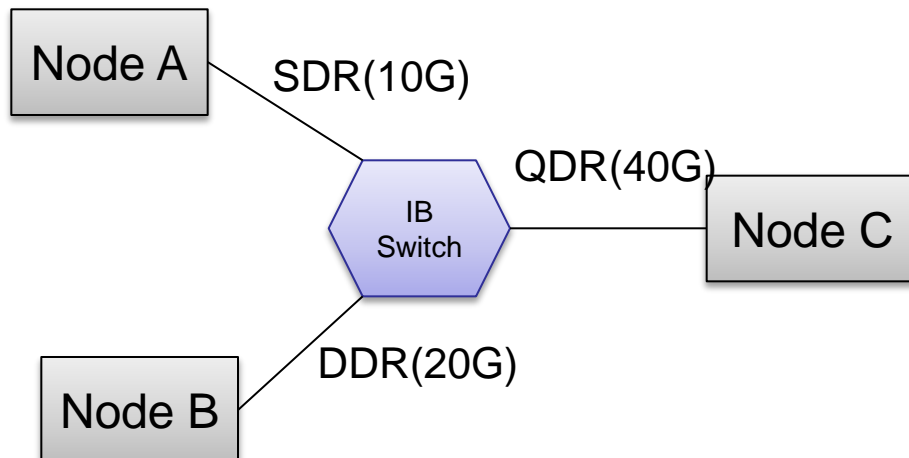
- Reliable Connection(RC)
 - Unreliable Connection(UC)
 - Reliable Datagram (RD)
 - Unreliable Datagram (UD)
- Reliableサービスでは、データは保証される。=ロス・レス

- 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

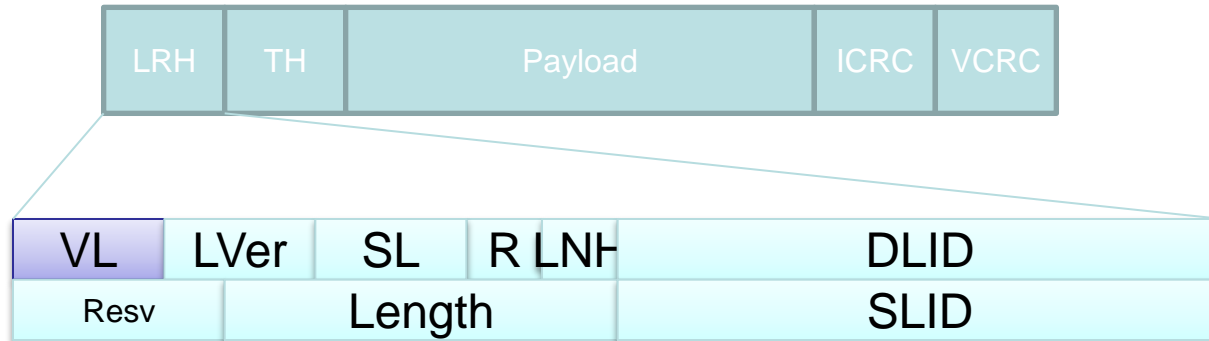


・ Inter Packet Delay (IPD)

- 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。

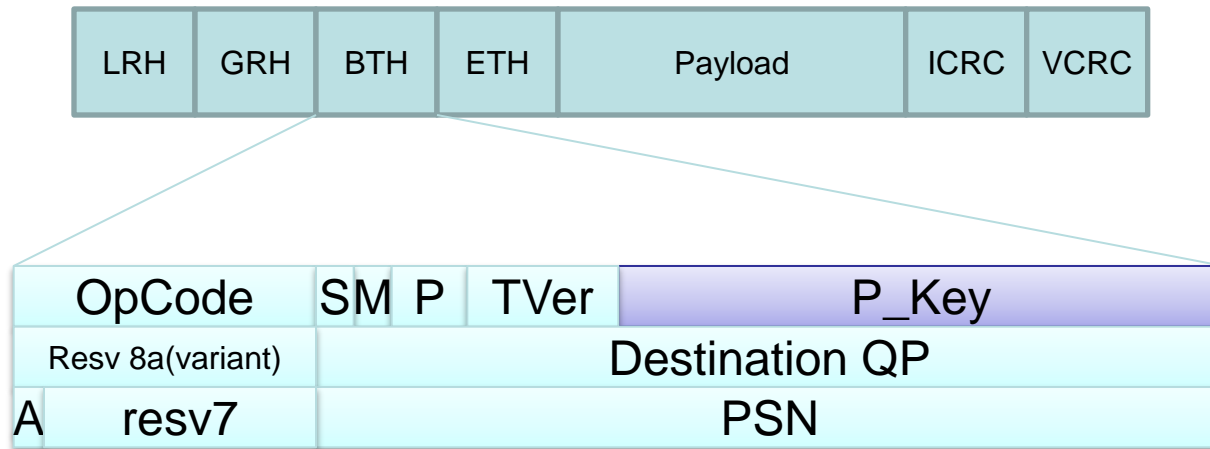


IPD	rate	Comment
0	100%	Suited for matched links
1	50%	
2	33%	Suited for 30 Gbps to 10 Gbps conversion
3	25%	Suited for 10 Gbps to 2.5 Gbps conversion
11	8%	Suited for a 30 Gbps to 2.5 Gbps conversion



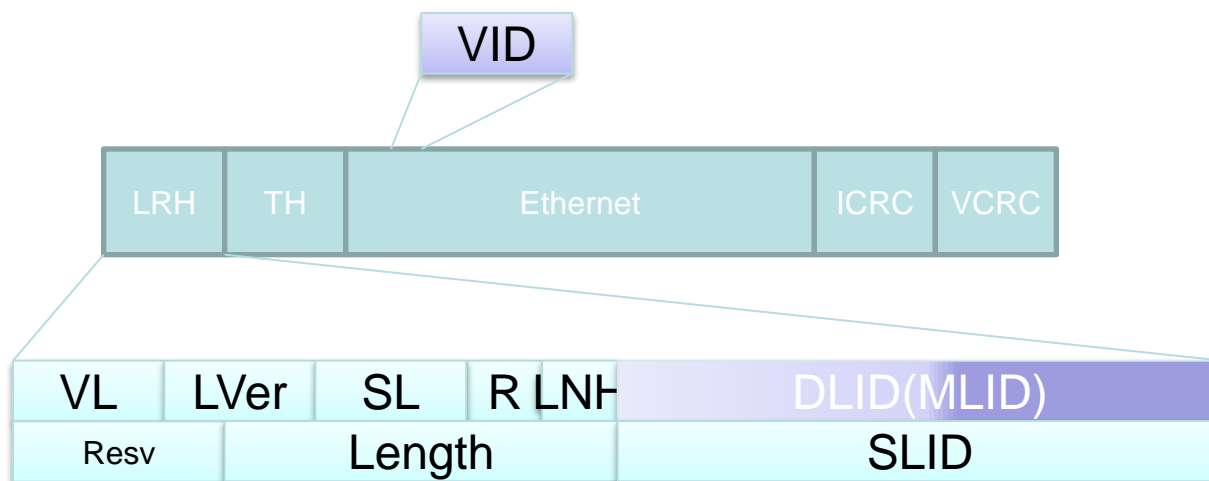
• バーチャル・レーン(4ビット)

- 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)



・ Partition Key(P_Key): 16ビット

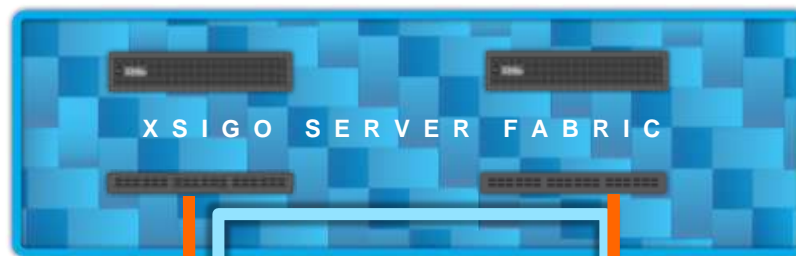
- 最上位ビットはメンバーシップ・タイプに使用。InfiniBandファブリック内を分割。P_Keyがマッチしたパケットのみを受信。P_KeyテーブルはPartition Managerによりコントロール。



- **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**

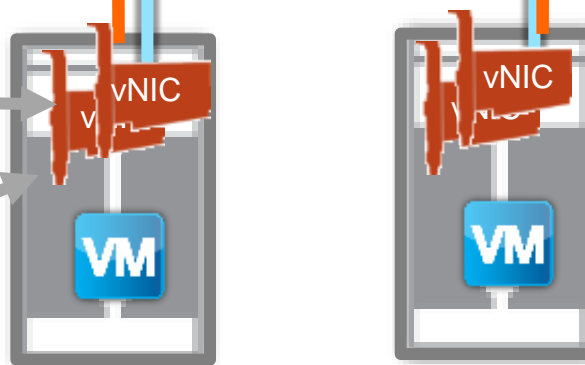
Xsigo Server Fabric

InfiniBand上でEthernet フ
レームをトンネリング



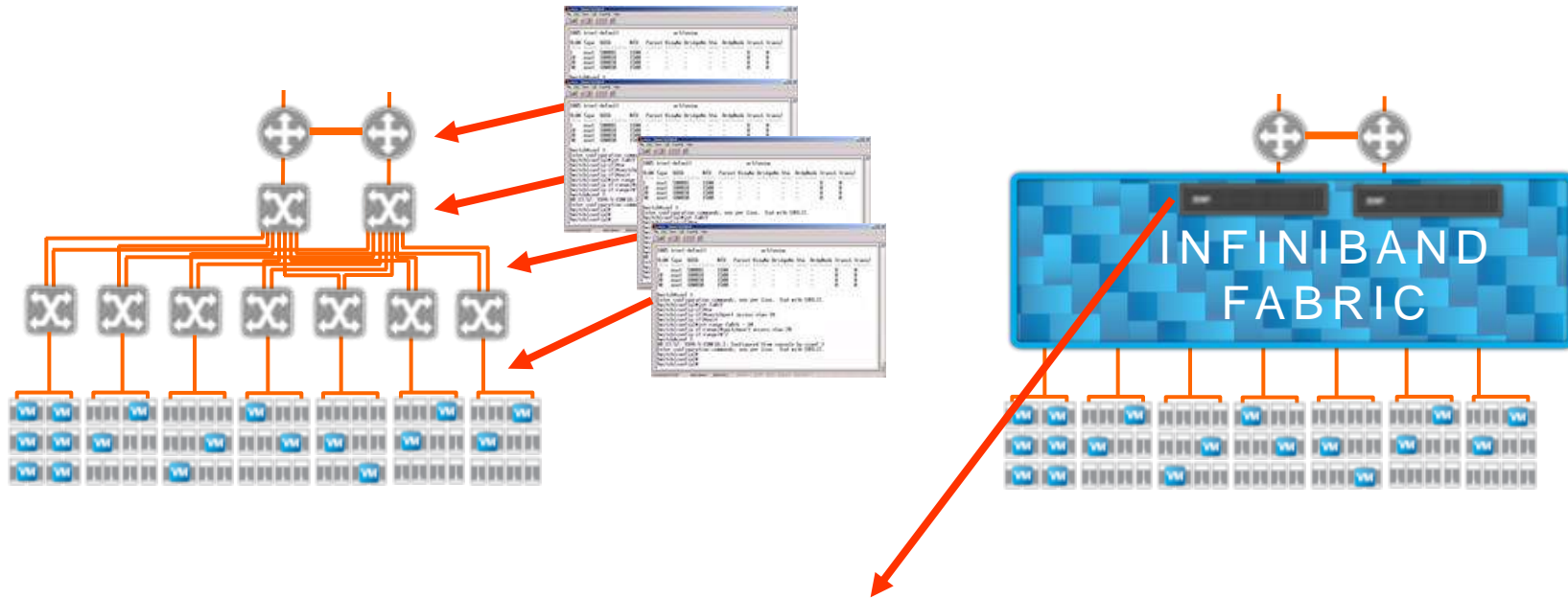
サーバ上では標準
のEthernet NICと
して見える

複数作成可能



サーバにはIB
HCAと、Xsigoド
ライバをインス
トール

Diagnostic



```
admin@vp780p[xsigo] show diagnostics ib-path hpesx40.shibuya.xsigo.com 10
```

name	lid/port	width/speed	SymErrs	LnkRcov	LnkDwnd	RxErrs	RxSwErr	TxDisc	TxPkts	RxPkts	XmtWait
hpesx40 InfiniHostEx HCA	1/1	4X/2.5 Gbps	0	0	0	0	0	0	473958	476090	0
Xsigo Core Switch	15/2	4X/2.5 Gbps	0	0	0	0	0	0	476092	473961	0
Xsigo Core Switch	15/5	4X/5.0 Gbps	0	0	0	0	1	0	741245	515937	0
Xsigo Leaf 1 Switch	16/17	4X/5.0 Gbps	0	0	0	0	0	0	515939	741249	0
Xsigo Leaf 1 Switch	16/10	4X/5.0 Gbps	0	0	0	0	0	0	701435	701423	0
slot=10 vn10x1gcard	6/1	4X/5.0 Gbps	0	0	0	0	0	0	513412	209259	0

```
admin@vp780p[xsigo]
```

- InfiniBand概要
- Ethernetとの機能比較
- データセンタへの適用

- IPoIB

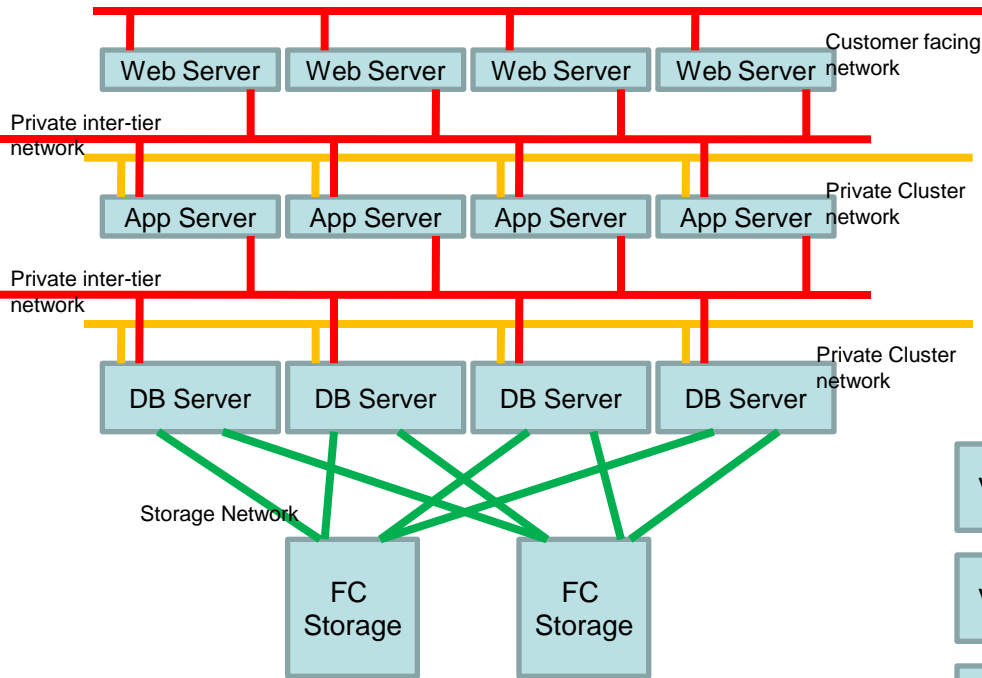
- OFEDで標準で提供
- Bonding (=NIC Teaming)サポート
- IPoIBルータは・・・1GEくらいまで？

- EoIB、FCoIB

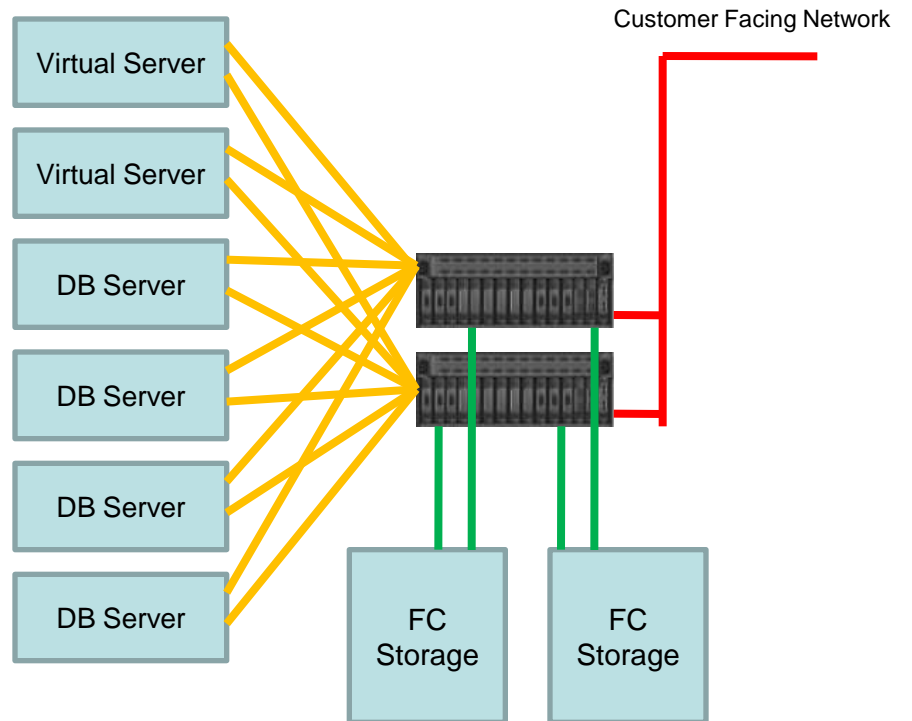
- サーバとIO Chassis間をInfiniBandでトンネリング
- GE、10GE、4G FC、8G FCへの接続可能



I/O統合

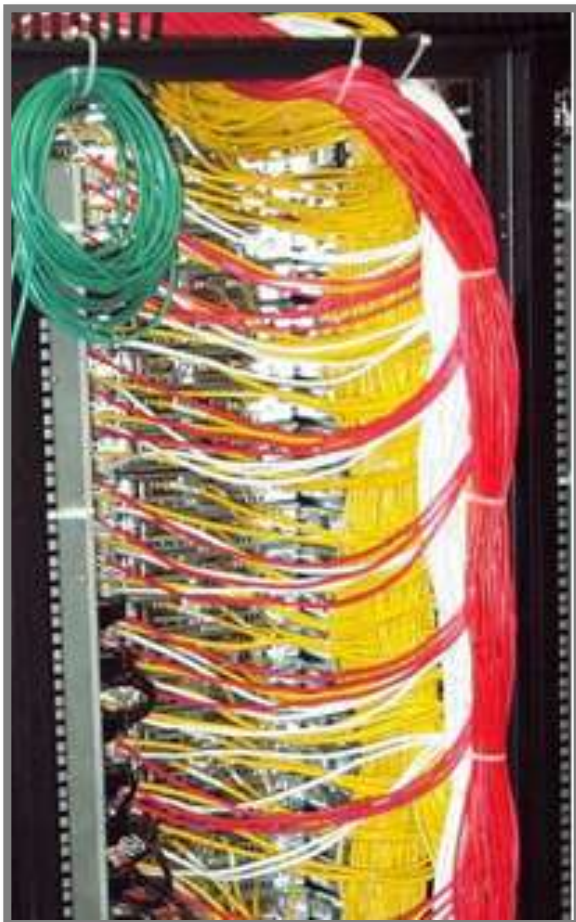


3 tier application stack



ケーブル数の削減

統合前



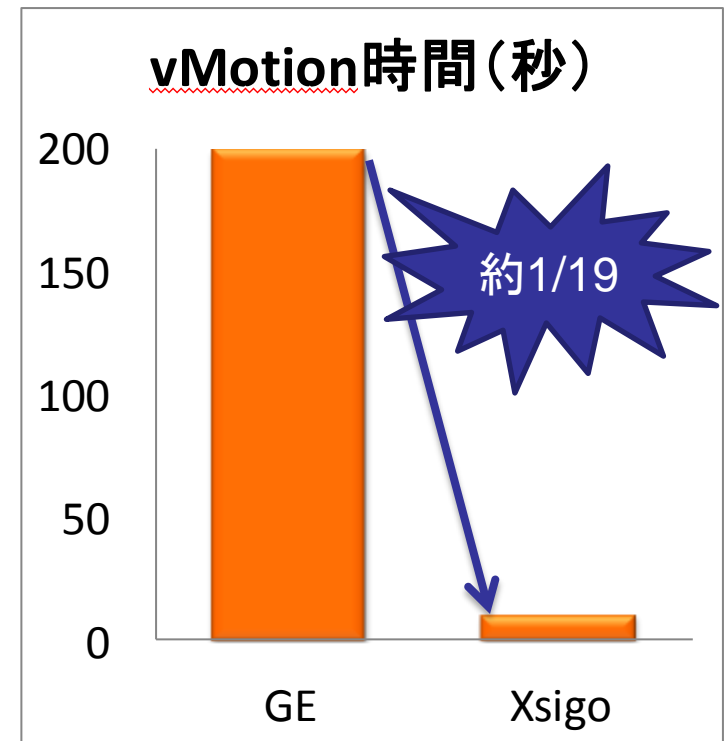
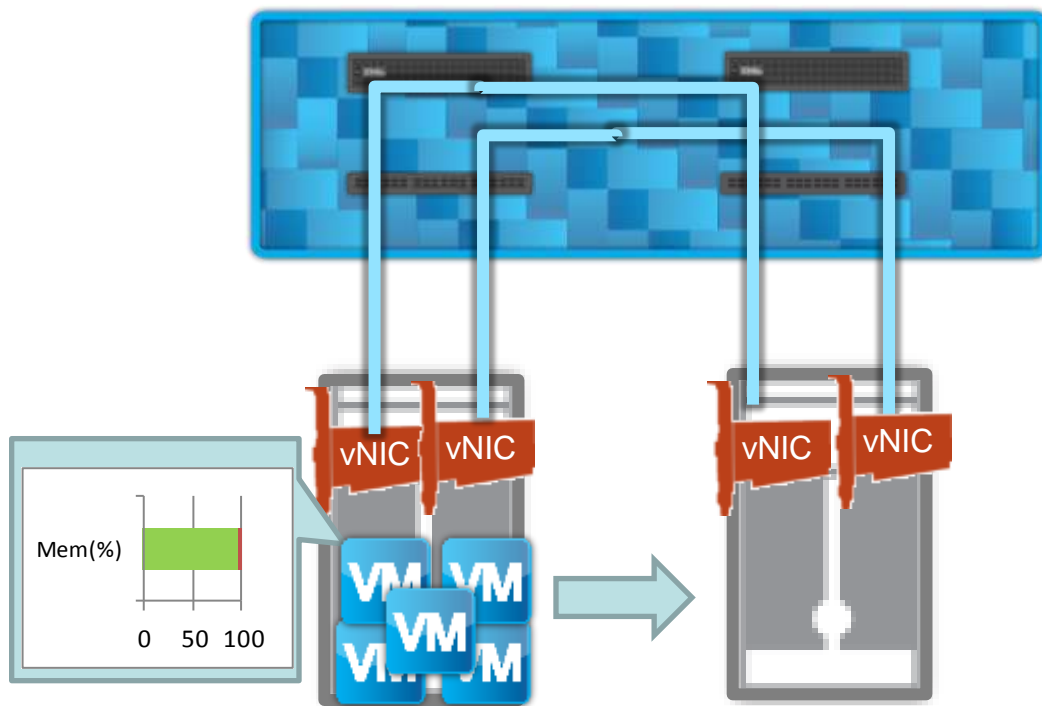
統合後



ライブ・マイグレーション

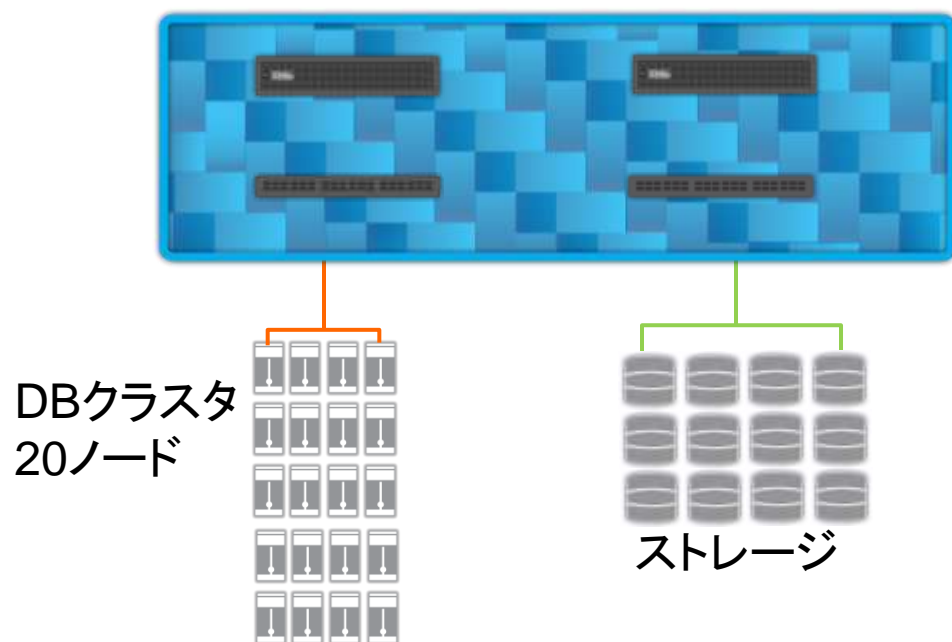
- 5台のVMへメモリ負荷をかけた状態で、ライブマイグレーションに要する時間を測定
- 1GE接続に比べて、XSFでは19倍高速に

※ESX5より、複数vmknicsでのvMotionが可能



DBクエリ時間の短縮

- 20ノードのDBクラスタをXsigoへIB接続
- 監査クエリに対して、1GEで50秒→ Xsigoで3秒へ短縮
- クエリ処理速度は平均でも2倍以上向上



- InfiniBandは、下位レイヤと上位のRDMAレイヤで構成されている
- 下位レイヤは、非常にシンプルかつ高速
- データセンターでは、I/O統合が可能

