

JDSF FCT部会 勉強会

Infinibandプロトコルアナライザ

2004年5月12日



I/Oアクセス解析ソリューション部
堀部 勝義

Infiniband プロトコルアナライザ

- 内 容

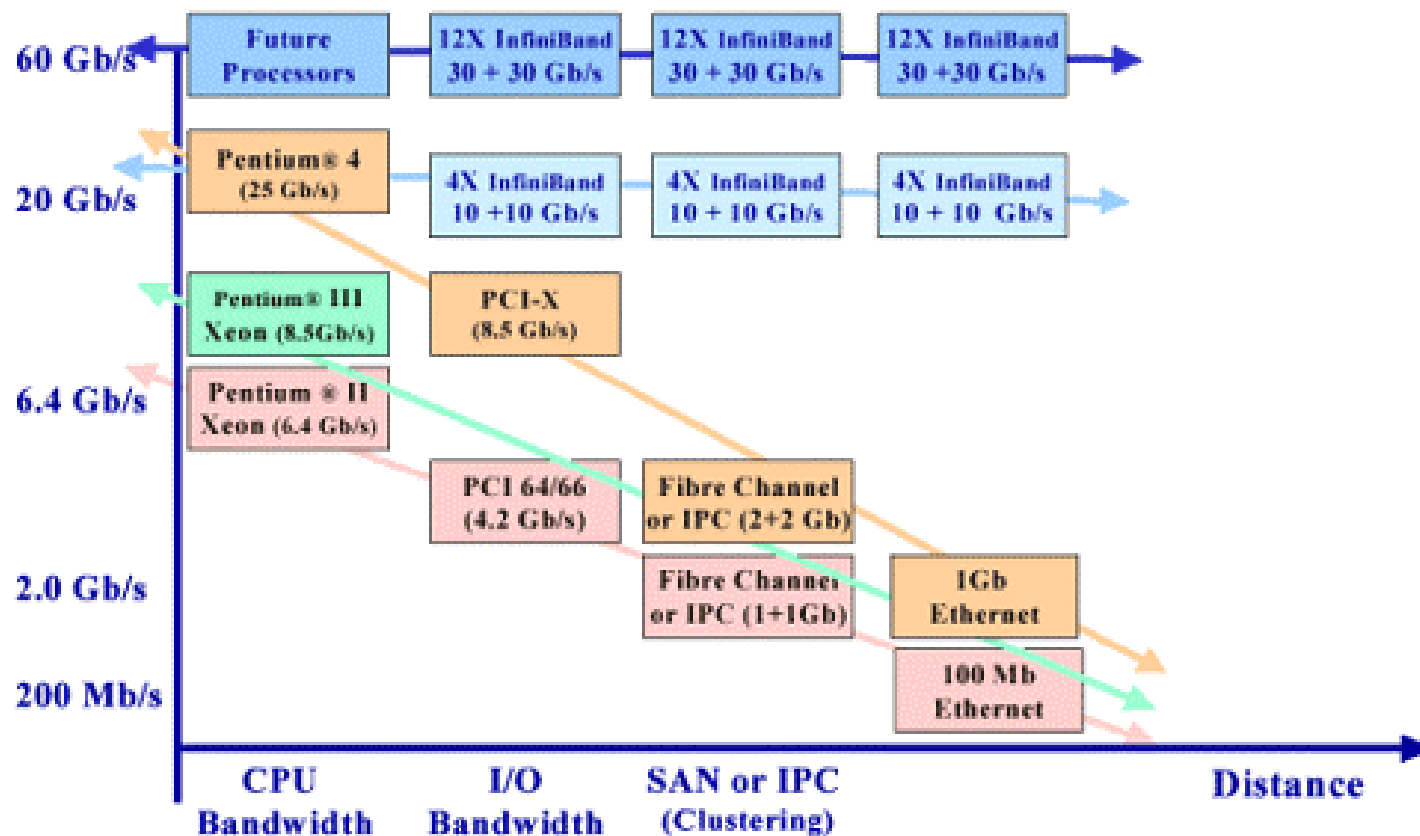
- Infinibandの概要

- プロトコルのロードマップ
 - Infinibandの特徴
 - Infinibandの技術概要

- Infinibandプロトコルアナライザについて

Infinibandの概要

Infinibandロードマップ



Infinibandの概要

Infinibandの特徴

- 階層化されたプロトコル
- マルチ・レイヤーの接続性
- パケットベースのコミュニケーション
- マルチキャスト機能
- パケットおよびエンド・ノードのフォルト・トレランス
- サブネット・マネージメント機能
- リンクスピード - 1x, 4x, 12x
- 2.5 ~30 Gbit / sec.の転送レート
- PCB、銅、ファイバーのフィジカル・リンク
- リモートDMA をサポート

Infinibandの概要

- **Channel-Based Architecture** – InfiniBandアーキテクチャはチャンネルベースの I/O モデルにより、ファブリックノード間で信頼性の高い接続が可能
- **Message-Passing Structure** – InfiniBandアーキテクチャ・プロトコルは データ転送に効果的な message-passingストラクチャを採用
- **Natural Redundancy** - InfiniBandファブリックでは、ノードはlink redundancyをファブリックにアタッチできるので、1つのパスが失敗しても、トラフィックは最終的なエンドポイントに再ルート可能。
- **Quality of Service (QoS)** – リンクレイヤはInfinibandのQoS 特性を可能にします。InfiniBand は 15の独立したレベル (VL0-14)とひとつのマネージメントパス (VL15) でデバイス特定の優先順位を構成することが可能。これによりファブリックにおけるI/Oオペレーションに優先権がアサインされクリティカルな通信をすることが可能。
- **Credit-based flow control** –クレジット・ベースのフロー制御管理アプローチにより、IBAネットワーク上の各受信ノードは、パケットロスなしに効率的に転送可能な最大データ量の値を送信デバイスに転送する。クレジット・データは、IBAファブリックに沿って専用のリンク経由で転送。送信デバイスは受信デバイスからプライマリ通信バッファ経由でデータ転送が可能であることを知るまで、パケットは送信しない。
- **CRC Check** – 2つのCRCによりエンド to エンドのインテグリティをチェック; 16-bit variant CRC値は各データ・フィールドに指定されて、各IBAファブリック・ホップで再計算される。32ビットinvariant CRC値は、各IBAホップポイント間で変わらない静的データを保護するように設計されている。
- **Subnet Management** – InfiniBandのネットワークレイヤは サブネット間のパケットルーティングを提供。ルートされた各パケットは、ソースとデスティネーション・ノードのためにグローバル・ルート・ヘッダ (GRH)と128ビットIPv6アドレスがある。またネットワークレイヤは、全てのサブネットに沿って各デバイスのために標準64ビットのユニークなグローバルな識別子を埋めます。これらの一貫した識別値のハンドリングにより、IBAネットワークは、複数の論理的サブネット経由のデータ転送が可能。

Infiniband プロトコルアナライザ

内 容

- Infinibandの概要
 - プロトコルのロードマップ
 - Infinibandの特徴
 - Infinibandの技術概要
- **Infinibandプロトコルアナライザについて**
 - IB Tracer 1X アナライザ
 - IB Tracer 4X アナライザ
 - IB Trainer 4X エキササイザ
 - Infiniband SPEC1.1のサポート
 - Lane Reversalについて
 - Script Verification Engineについて
- 東陽テクニカ ストレージアプリケーション用
プロトコル解析評価ツール製品ライン

IBTracer × 4 プロトコルアナライザ

Infiniband 4xのプロトコル解析



- Infiniband Spec1.1をサポート
- InfiniBand 1X(2.5Gbps), 4X(10Gbps)転送速度のバストラフィックを記録

IBTracer × 1 プロトコルアナライザ

Infiniband 1xのプロトコル解析



- Infiniband Spec1.0をサポート
- InfiniBand 1X(2.5Gbps), 転送速度のバストラフィックを記録

IBTracer Infinibandプロトコルアナライザ

IBTracerの特徴

- ドリル・ダウン表示: MAD (マネージメント・データグラム)、エラー、ペイロード、パケット
- マルチレーンリンク解析: late-to-lane スキューの記録と表示
- Lane-Reversal 対応: レーンの反転をトリガ、記録、表示対応
- SPEC1.1の新しいIMAD タイプをサポート
- Script Verification Engine: ユーザ定義可能なスクリプトにより自動解析が可能
- トレーニング・シーケンスを表示
- スキップ・オーダードセット、フラグ・バイオレーション を表示

IB Tracer Infinibandプロトコルアナライザ

トレース画面

The screenshot displays the VndrApplSample.ibt application window. The main interface shows a trace entry for VendorGetResp() with the following details:

| MAD | ReqLID | RespLID | Vendor | VendorGetResp() | ClassPortInfo | Vendor Data | BaseVersion |
|-----|--------|---------|--------|-----------------|---------------|---------------|-------------|
| 0 | 0x0002 | 0x0004 | Class | (Response) | (0x0000) | ClassPortInfo | 0xAB |

Below this, a tooltip for RedirectGID is visible, containing the text: "The GID a requester shall use as the destination GID in the GRH of messages used to access redirected class services. If redirection is not being performed, this shall be set to zero." The value for RedirectGID is 0000000000.

Other fields shown include:

| ClassVersion | ClassSpec_CpblMsk | Common_CpblMask | RespTimeValue | RedirectGID |
|--------------|-------------------|-----------------|---------------|-------------|
| | | | | 0000000000 |

Trap information:

| TrapGID | TrapTC | TrapSL | TrapFL | TrapLID | TrapP_Key | TrapHL | TrapQP | TrapQ_Key |
|--|--------|--------|---------|---------|-----------|--------|----------|------------|
| 0xACDB0104-0415-0000-B700-00D004150000 | 0x00 | 0x0 | 0x00000 | 0x3316 | 0x0000 | 0xB7 | 0x0000D0 | 0x00003101 |

Operation details:

| Operation | SEND | ReqLID | RespLID | SrcQP | DestQP | Bytes Transferred |
|-----------|------|--------|---------|----------|----------|-------------------|
| 0 | UD | 0x0002 | 0x0004 | 0x000055 | 0x000000 | 256 |

Packet details (Packet 0, Rx):

| Packet | Rx | LRH | DLID | SLID | BTH | UD | SEND Only | SE | M | Pad | P_Key | TVer | DestQP |
|--------|----|-----|--------|--------|-----|-------|-----------|----|---|--------|-------|----------|--------|
| 0 | | | 0x0004 | 0x0002 | 011 | 00100 | 1 | 1 | 0 | 0x0321 | 0x1 | 0x000000 | |

Additional fields:

| A | PSN | DETH | Q_Key |
|---|-----|------|------------|
| 1 | 5 | | 0x000009AC |

Data section:

| Offset | Data |
|--------|---|
| 0: | 01090181 FFFF0101 34000000 00000000 0001DF5F 00000000 ABCD0000 00000000 |
| 8: | 0000000A 00000000 00000000 00000000 00000000 00000000 00000000 00000000 |

IB Trainer × 4エキサイサイザ



- システムやデバイス・レベルの検証のために再現可能な任意の InfiniBand 4xパケットを生成
- Invariant CRC / Variant CRC、パケットヘッダ、パケットフレーミング、ランニングディスパリティ、8B/10Bシンボルなどのエラーをトラフィックストリームにインジェクション可能
- InfiniBandコンポーネントの機能テストのため、リンク・トレーニング、アイドルデータの送信、スキップオーダードセット、リンクレベルのフロー制御が可能

Infiniband SPEC1.1のサポート

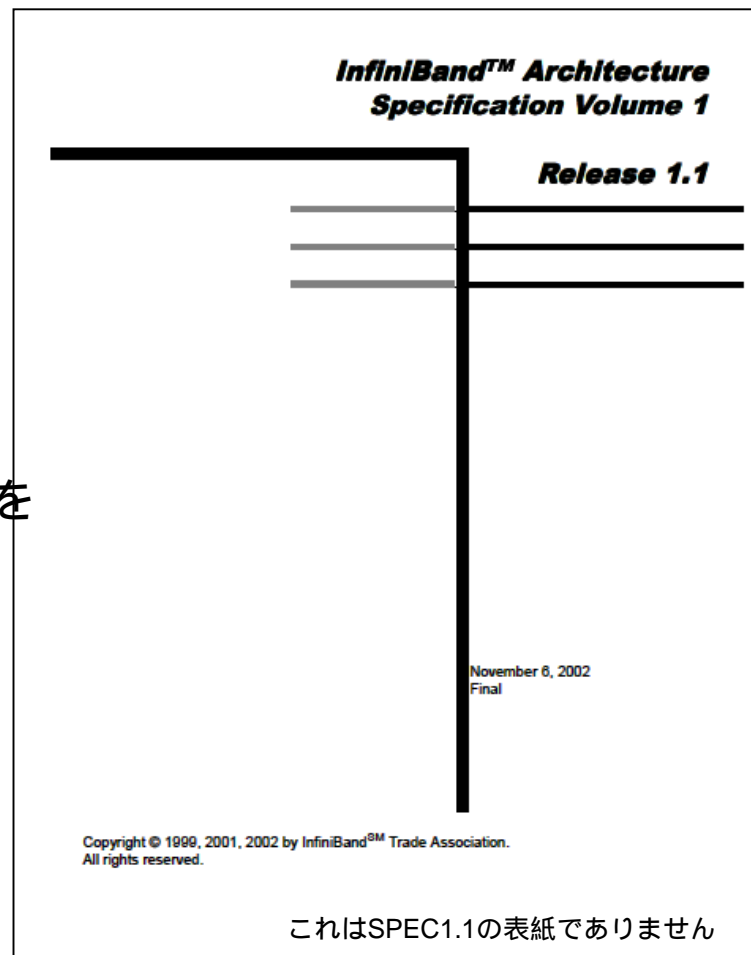
MAD (Management Datagram)のアーキテクチャ変更を含む
InfiniBand SPEC1.1をサポート

以下の新しいIMADパケットを
デコード可能

- Subnet Management Class
- Subnet Administration Class
- Communication Management Class

以下の新しいトランザクションタイプを
デコード可能

- Reliable Multi-Packet Transaction Protocol (RMPP) for SADM



SPEC1.1 : Subnet Management Class (1)

- サブネット・マネージメントクラスは、サブネット内の構成、モニタリング、ノードのクエリーを提供する。
- マスター・サブネット・マネジャーは、IBサブネットを初期化し構成する重要なエレメントであり、サブネットの初期化プロセスの一部として選択される。

SPEC1.1 : Subnet Management Class (2)

CATC IBTracer(TM) InfiniBand Protocol Analyzer(spec. 1.1) - [C:\Progr...

File Setup Record Report Search View Window Help

TRG REC STOP LNK JS

Pkt Tra MAD SA SDP SRP

サブネット・マネージメントのデコード画面

| MAD | ReqLID | RespLID | S M P | SubnSet() | PortInfo | | |
|------------|------------|------------|---------------------|---------------|--------------------|------------|----------|
| 10 | 0x0002 | 0x0004 | Directed Route | (Request) | (Port # 0) | | |
| M_Key | DrSLID | DrDLID | S M P Data | Initial Path | Return Path | Time Delta | |
| (0x0)(0x0) | 0xFFFF | 0xFFFF | PortInfo Port (# 0) | (5) | 0 | 2.336 μs | |
| MAD | ReqLID | RespLID | S M P | Status | D | HopPointer | HopCount |
| 11 | 0x0002 | 0x0004 | Directed Route | 0x0000 | 0 | 1 | 1 |
| BaseVer | MgmtClass | ClassVer | R | Method | TrailD | AttrID | AttrM |
| 0x01 | Subn | 0x01 | 0 | Set() | 0x3400000000000000 | VendorDiag | 0x2 |
| M_Key | DrSLID | DrDLID | S M P Data | Initial Path | Return Path | Time Delta | |
| (0x0)(0x0) | 0xFFFF | 0xFFFF | VendorDiag | (5) | 0 | 2.336 μs | |
| MAD | ReqLID | RespLID | S M P | SubnGetResp() | NodeInfo | | |
| 12 | 0x0002 | 0x0004 | LID Routed | (Response) | (0x0000) | | |
| M_Key | S M P Data | Time Delta | | | | | |
| (0x0)(0x0) | NodeInfo | 2.336 μs | | | | | |

Ready Search: Fwd

SPEC1.1 : Subnet Administration Class (1)

- サブネット・アドミニストレーション(SA)クラスは、サブネットの運営に必要な情報の格納場所を提供する。
- サブネット・トポロジとパーティション情報、データ・パス、イベント、サービスレベルの通知を含む。

SPEC1.1 : Subnet Administration Class (2)

View Fields for SADM Transaction #0

AdmGetTableResp() : 2 segment(s), 384 bytes

Subnet Administration Multi-Packet Response Transaction, contains bunch of PathRecord attributes

Subnet Administration transaction records

| SubnAdm Data | PathRID | RESERVED | DGID |
|---------------|------------|----------|--|
| PathRecord(1) | 0x00010000 | 0 | 0xFE800000-0000-0000-00D0-3C00010003B5 |
| PathRecord(2) | 0x00040000 | 0 | 0xFE800000-0000-0000-0002-C900000FA251 |
| PathRecord(3) | 0x00060000 | 0 | 0xFE800000-0000-0000-0002-C900000FA253 |
| PathRecord(4) | 0x00010000 | 0 | 0xFE800000-0000-0000-00D0-3C00010003B5 |
| PathRecord(5) | 0x00040000 | 0 | 0xFE800000-0000-0000-0002-C900000FA251 |
| PathRecord(6) | 0x00060000 | 0 | 0xFE800000-0000-0000-0002-C900000FA253 |

Save As... Find... Layout... Previous Next Close

SPEC1.1 : Communication Management Class (1)

- コミュニケーション・マネージメント(CM)クラスは、IB Reliable Connection、Unreliable Connection、Reliable Datagram 転送サービスのタイプのためにチャネルの確立、維持、開放に使用されるメカニズムを含む。

The screenshot shows a network analysis tool interface. On the left, a tree view under 'All reports' shows 'Pkt Packets' expanded to 'Services and OpCodes', with 'RC Reliable Connection' selected. Below this are 'UC Unreliable Connection', 'RD Reliable Datagram', and 'UD Unreliable Datagram'. Further down are 'Virtual Lanes', 'Source LID', 'Destination LID', 'Tra Operations', 'MAD MADs', and 'Errors'. On the right, a table titled 'OpCodes' lists various operations and their total counts, all of which are currently zero.

| OpCodes | Total |
|---------------------------|-------|
| SEND First | 0 |
| SEND Middle | 0 |
| SEND Last | 0 |
| SEND Last w/ ImmDt | 0 |
| SEND Only | 0 |
| SEND Only w/ ImmDt | 0 |
| RDMA WRITE First | 0 |
| RDMA WRITE Middle | 0 |
| RDMA WRITE Last | 0 |
| RDMA WRITE Last w/ ImmDt | 0 |
| RDMA WRITE Only | 0 |
| RDMA WRITE Only w/ ImmDt | 0 |
| RDMA READ Request | 0 |
| RDMA READ response First | 0 |
| RDMA READ response Middle | 0 |
| RDMA READ response Last | 0 |
| RDMA READ response Only | 0 |
| Acknowledge | 0 |
| ATOMIC Acknowledge | 0 |
| ATOMIC CmpSwap | 0 |
| ATOMIC FetchAdd | 0 |
| | 0 |

SPEC1.1: Communication Management Class (2)

CATC IBTracer(TM) InfiniBand Protocol Analyzer(spec 1.1)

File Setup Record Report Search View Window Help

REC STOP

Pkt Tra MAD SA SDP SRP

C:\Program Files\CAT...\AllDecodedMADs (spec 1.1).ibt

コミュニケーション・マネージメントクラスのデコード画面

| MAD | ReqLID | RespLID | Communication | Status | BaseVer | MgmtClass | ClassVer | R |
|-----------------------|--------|-----------------------|---------------------|-----------------|---------------------|-----------------|------------|---|
| 23 | 0x0002 | 0x0004 | Management | 0x0000 | 0x01 | ComMgt | 0x01 | 0 |
| Method | | TraID | AttrID | AttrM | CMP Data | | RequestID | |
| Send() | | 0x3400000000000000 | ServiceIDResReq | 0x0 | ServiceIDResReq | | 0xABCD0000 | |
| P_Key | | ServiceID | Private | Data | Time Delta | | | |
| 0x00000000 | | IBTA: 00000A 00000000 | Data | 216 bytes | 2.336 μs | | | |
| MAD | ReqLID | RespLID | Communication | Status | BaseVer | MgmtClass | ClassVer | R |
| 24 | 0x0002 | 0x0004 | Management | 0x0000 | 0x01 | ComMgt | 0x01 | 0 |
| Method | | TraID | AttrID | AttrM | CMP Data | | | |
| Send() | | 0x3400000000000000 | ServiceIDResReqResp | 0x0 | ServiceIDResReqResp | | | |
| RequestID | | Status | AddInfoLen | !!! WARNING !!! | QPN | !!! WARNING !!! | | |
| 0xABCD0000 | | ClavVer not supported | 18 | 0xAA04 - not 0! | Not Valid | 0xA - not 0! | | |
| ServiceID | | Q_Key | Hghst CM ClaVer | Private | Data | Time Delta | | |
| IBTA: 000000 00000011 | | Not valid | 0x00 | Data | 136 bytes | 2.336 μs | | |

Ready

Search: Fwd

SPEC1.1 : Reliable Multi-Packet Transaction Protocol RMPP (1)

- InfiniBandアーキテクチャでは、マネージメントクラスがシングルMADより大きい量のデータを確実に転送しなければならない状況がある。
- これを円滑に行うために、SPEC1.1ではシングルの論理トランザクションで大量のデータ(最高 2^{32} パケット)を転送可能なRMPPを定めた。

SPEC1.1: Reliable Multi-Packet Transaction Protocol (RMPP) for SADM (2)

CATC IBTracer(TM) InfiniBand Protocol Analyzer(spec. 1.1) - [C:\Program Files\CAT...SADM Transactions (spec. 1.1)...]

File Setup Record Report Search View Window Help

RMPPトラフィックの全てのポートの特定の管理データを表示

| MAD | ReqLID | RespLID | Subnet | SubnAdmGetTable() | PortInfoRecord | RMPP |
|-----|--------|---------|----------------|-------------------|----------------|--------|
| 0 | 0x0004 | 0x0002 | Administration | (Request) | (0x0000) | Header |

| RMPPType | ComponentMask | SubnAdm Data | Time Delta |
|-------------|---------------------|----------------|------------|
| Not an RMPP | 0x00000000 00000010 | PortInfoRecord | 2.336 μs |

| SADM | ReqLID | RespLID | AdmGetTableResp() | SubnAdm Data | LID | PortN |
|------|--------|---------|-------------------------|-------------------|--------|-------|
| 0 | 0x0002 | 0x0004 | 5 segment(s), 960 bytes | PortInfoRecord(1) | 0xAABB | 1 |

| !!! WARNING !!! | S M P Data | M_Key | GIDPrefix | LID |
|-----------------|---------------------|--------------------|-----------------------|--------|
| 0x2 - not 0! | PortInfo Port (# 1) | 0x00000000ACDB0104 | 0x0415-0000-B700-00D0 | 0x0415 |

| MasterSMLID | CapabilityMask | DiagCode | IndexForward | VendorDiagCode | StandardDiag |
|-------------|----------------|------------|--------------|----------------|--------------|
| 0x0000 | 0x000000D0 | (3 fields) | Not set | 0x331 | Unknown code |

| M_KeyLeasePeriod | LocalPortNum | LnkWidthEnabled | LnkWidthSupported | LnkWidthActive |
|------------------|--------------|-----------------|-------------------|----------------|
| Infinite | 183 | No Change (NOP) | 0 : reserved | 208 : reserved |

| LnkSpeedSupported | PortState | PortPhysState | LnkDwnDefState | M_KeyProtBits | !!! WARNING !!! |
|-------------------|-----------------|---------------|----------------|---------------|-----------------|
| 0 : reserved | No Change (NOP) | No Change | No Change | 0 | 0x6 - not 0! |

| LMC | LnkSpeedActive | LnkSpeedEnabled | NeighborMTU | MasterSMSL | VLCap | InitType | VLHighLimit |
|-----|----------------|-----------------|--------------|------------|--------------|----------|-------------|
| 0x1 | 0 : reserved | 2.5Gbps | 0 : reserved | 0 | 0 : reserved | 0x0 | 0 |

| VLArbitrHighCap | VLArbitrLowCap | InitTypeReply | MTUCap | VLStallCount | HOQLife | OperVLs |
|-----------------|----------------|---------------|--------------|--------------|---------|--------------|
| 0 | 0 | 0x0 | 0 : reserved | 4 | 0 | 8 : reserved |

Ready Search: Fwd

Lane Reversalについて

- The IBTracer 4X アナライザは、IBポートの Lane Reversalを サポート
- IBA 4Xリンクで接続するとき起こりうる、正しくないピンアサインを補正をオンチップ・ロジックで行うことを目的とした InfiniBandの特長である
- Lane Reversalを実装したIB 製品は、反転したレーン番号の順序を使用して4x linksの接続をするためPWB (printed wiring board) レイアウトを割り当てる
- IB Tracer 4X は、自動的に反転したレーンコンディションを検出するためにトレーニングシーケンス(TS1 or TS2)の間に送信されたレーン番号シンボルを使用する
- IB Tracer は、レーンの物理構成に関係なく論理的にIBトラフィックの表示、トリガリング、記録が可能

Script Verification Engineについて

- InfiniBand Trade Association でプラグテスト用にデザインされたコンプライアンス・テストスクリプトをサポート
- ユーザ独自のテストスクリプトの生成可能
- トレースデータからテストスクリプトの生成可能
- Lane-Reversalモードなどの拡張機能もサポート
- プラグテストにおいて
 - Infiniband TAの複雑なリンクレイヤテスト(c1v07-053)を自動化
 - リンクパケットが正しいオペコードを使用しているか (リンクの初期化の間、アーム、アクティブステート)
 - リンクパケット間の平均インターバル時間
 - パケットレングスフィールドの正確な計算

東陽テクニカ ストレージアプリケーション用 プロトコル解析評価ツール製品ライン

Infiniband アナライザ・ジェネレータ

Serial Attached SCSI アナライザ・ジェネレータ

Serial ATA アナライザ・ジェネレータ

ATA/ATAPI アナライザ&テスター・ジェネレータ

Ultra320SCSI アナライザ&テスタ

iSCSI / Fibre Channel アナライザ&テスタ

PCI Express アナライザ&ジェネレータ

I²C アナライザ&ジェネレータ

