

Oracle Maximum
Availability Architecture

Oracleホワイト・ペーパー
2012年10月

InfiniBandネットワークにおけるトラブル シューティングのガイドラインと方法

ORACLE®

概要	3
ネットワークの構成要素	4
開放型システム間相互接続 (OSI)	4
ネットワークを構成するハードウェア	5
ハードウェア・レベルでのデバッグ	5
ネットワークを構成するソフトウェア	6
InfiniBandサブネット・マネージャ	7
InfiniBandパーティション	7
ソフトウェア・レベルでのデバッグ	8
ヘルス・チェックのガイドライン	9
NTPサービスの重要性	9
ログのローテーションとアーカイブ	9
構成設計図	10
InfiniBandトポロジの保存	10
InfiniBand診断ツール	12
ホストに固有のコマンド	13
perfqueryカウンタの説明	16
InfiniBandスイッチに固有のコマンド	17
グローバルInfiniBandコマンド	21
ibdagnetのログによる分析	23
ネットワーク・パケットのキャプチャ	26
Wiresharkにおけるibdump pcapのスクリーンショット	27
ノード情報の問合せ	27
パフォーマンス・カウンタの問合せ	28
IPoIBでのIPv4 ICMP (ping)	29
SDPハンドシェイク	30
ARPリクエストの失敗	31
トラブルシューティングのシナリオ	32
ユースケース分析#1	32
ユースケース分析#2	32
ユースケース分析#3	33
ユースケース分析#4	33
ユースケース分析#5	34
ユースケース分析#6	35
ユースケース分析#7	35
ユースケース分析#8	36
ユースケース分析#9	36
ユースケース分析#10	36
ユースケース分析#11	37
ユースケース分析#12	38
ユースケース分析#13	38

概要

Oracle Maximum Availability Architecture (MAA) は、高可用性テクノロジーを実装するためのオラクルのベスト・プラクティス構想です。ネットワーク・テクノロジーは、最新のコンピューティング環境のインフラストラクチャ全体で、極めて重要なコンポーネントとなっています。オラクルのエンジニアド・システムは、10ギガビット・イーサネットや40ギガビットInfiniBand (IB) などの最先端の高速で効率的なネットワーキング・テクノロジーの1つに基づいて構築されています。このホワイト・ペーパーでは、最先端のネットワークのトラブルシューティングを行うための方法およびベスト・プラクティス・ガイドラインについて説明します。その中で、さまざまな構成要素と一般的な問題領域だけでなく、診断コマンドの簡単な説明とサンプル出力を紹介します。また、問題となる状況を含む一連の重要なシナリオおよび考えられる根本原因をトラブルシューティングするための系統的なアプローチを説明します。

ネットワークの構成要素

ネットワークのトラブルシューティング方法を説明する前に、詳細な調査に役立つ予備知識として、ネットワークを構成するコンポーネントについて解説します。大きく分けると、コンポーネントは、ソフトウェアとハードウェアの2つのカテゴリに分類されます。各カテゴリは、さらにサブカテゴリと個々の製品に分類できます。問題への対処と診断を試みる前に、これらの異なる機能領域をよく理解しておくことが不可欠です。最終的な目標は、問題の根本原因を突き止めて、問題を解決できるようにすることです。

複雑な環境では、特定領域に起因する問題が、別の領域においてより高いレベルの影響が現れるまで認識できないことがあります。ハードウェアのカテゴリとソフトウェアのカテゴリにあるすべてのレベルが互いに連携して動作することによって、要求された機能を提供しています。開放型システム間相互接続（OSI）モデルは、異なるレイヤーにまたがるハードウェア、ファームウェア、ソフトウェアのトラブルシューティングを、さまざまなプロトコルとメッセージを分析しながら行う場合に重要です。

開放型システム間相互接続（OSI）

以下の図に示したOSIモデルでは、下位にある3つのレイヤー（1～3）はハードウェアとオペレーティング・システムのネットワーク・ドライバ・スタックと密接な関係があり、上位にある4つのレイヤー（4～7）はアプリケーションとより密接な関係があります。この知識は、トラブルシューティングおよびさまざまなフェーズでの分析を実施する際に、極めて重要です。通常、下位レベルでの問題は上位レベルに継承されますが、その逆が起こることはありません。

OSIレイヤー	OSIモデル			例
7	ホスト・レイヤー	データ	アプリケーション	Webサーバー
6			プレゼンテーション	SSL
5			セッション	IPC
4	メディア・レイヤー	セグメント	トランスポート	UDP、TCP、SDP、RDS
3			ネットワーク	v4またはv6形式でのIPアドレス指定
2			データ・リンク	MAC、GUID
1		ビット	物理	ネットワーク・ハードウェア

本書では、最下層にある3つのレイヤー（1～3）について詳しく説明してから、他のレイヤーについて説明します。これらの3つのレイヤーは、ハードウェア、ファームウェア、オペレーティング・システムのネットワーク・スタック、およびもっとも重要なインターネット・プロトコル（IP）に対応するメディア・レイヤーを定義しています。それよりも上位にある4つのレイヤーで何かが発生しても一般化することが困難であり、通常、その領域はアプリケーション・モデルと使用モデルに極めて特異なものとなります。

ネットワークを構成するハードウェア

ネットワークによって、コンピューティング・プラットフォームは特定のプロトコルを介して情報を交換できます。つまり、ネットワークは、これらの独立したコンピューティング・プラットフォームを結び付けて、分散コンピューティング環境を形成します。この環境に参加する各エンド・ポイントには、マザーボード上のPCIe拡張スロットに取り付けるInfiniBand HCAまたはギガバイト・イーサネット・ポートなど、何らかのネットワーク・ハードウェアが必要です。このようなハードウェア・デバイス上にあるポートまたはソケットは、適切なケーブルを使用して外部と接続します。そのケーブルのもう一方の端は、スイッチに接続されています。さらに、他のホストもこのスイッチに接続します。このシナリオでは、物理レイヤーで一般的なスター型トポロジを作成します。まれではありますが、ホスト上のネットワーク・ポートを別のホストのネットワーク・インタフェース・ポートに直接接続して、ポイント・ツー・ポイント・トポロジを作成することもあります。

トラブルシューティングの観点から、これらの領域に問題が隠れている場合があるため、こうしたハードウェア・レイヤーについて理解することが重要です。ソフトウェアとは大きく異なり、一部のハードウェアは機械的に反応しやすいため、損傷によっては機能を損なう場合があります。このような問題がシステムに存在する場合、エンドユーザーがそれに直ちに気付くことはありませんが、結果としてアプリケーション・レイヤーでエラー状態を引き起こす可能性があります。コネクタの破損やピンの曲り、またはラック内で過熱状態になっている部分を見つけるために、ハードウェアを物理的に常時監視することはまずありません。コンピューティング・インフラストラクチャは、非常に複雑なハードウェアのセット上で動作する同じように複雑なソフトウェアから、一連のサービスを提供するように設計されています。

ハードウェア・レベルでのデバッグ

通常、トラブルシューティングのフェーズがこのレベルで始まることはありませんが、診断コマンドを実行しても問題が解決しない場合は、重要なチェックリスト項目になります。本書では後で詳しく説明しますが、ここでは、問題の原因を突き止めるため、またはハードウェアが期待どおりに機能していることを確認するために、ハードウェア・レイヤーで何ができるかについて簡単に説明します。ハードウェアの検証では、実際に設置場所の周辺に行き作業をすることが必要となる場合がありますが、それ以外の場合はセンサーおよびILOMプラグインを使用して検証を実施できます。

ケーブルは2つのエンド・ポイントを接続して最小単位のリンクを形成し、さらにエンド・ツー・エンド・ソフトウェアのアプリケーション・パスにはこのようなリンクが1つ以上含まれます。通常、銅ケーブルまたは光ファイバ・ケーブルの場合は、対応コネクタを使用してホストに接続します。RJ45イーサネット・ケーブルは、InfiniBand専用ケーブルと比較してもっとも簡単なケーブルの1つです。

2つのホスト間でpingが失敗すると、単純に問題の発生がレポートされます。オペレーティング・システムとアプリケーションのレベルですべての診断を実行することで、破損したラッチが原因でポートのソケットからイーサネット・ケーブルが脱落、クリンピング不良によるいずれかの配線が接触不良、またはケーブルを間違ったポートに接続といった原因を発見できます。すべてのケーブルは、業界標準規格に従って作成されています。イーサネット・ケーブルは物理的に丈夫ですが、InfiniBand銅ケーブルは許容できる最小曲げ半径が規格で厳密に規定されています。物理的な取り回しの最中にこれよりも小さな半径で曲げてしまうと、内部損傷が発生して結果的にインフラストラクチャ内でエラーを引き起こすこととなります。

設置されているハードウェア・システムに対しては、定期的に保守または更新作業を実施するようになっています。このような物理的な保守作業の間に、IB-HCAなどのネットワーク・インタフェース・カードが物理PCIeスロットから外れてしまうことがあります。これは故障ではありませんが、機能の中断を引き起こすこととなります。カードが拡張スロットから外れるという問題は非常に簡単に修正できますが、アップグレードまたは保守作業の間にプリント基板から小さな部品を脱落させてしまった場合はどうなるでしょうか。どのような部品が脱落するかにもよりますが、カードが完全に機能しなくなる、動作を継続できてもパフォーマンスへの何らかの影響や寿命の低下が起こる、または最悪の場合にデータが破損するといった問題が発生します。

検査する必要があるもう1つのハードウェア領域は、ケーブルを取り付ける実際のポート・ソケットです。InfiniBandケーブルは、ポートに取り付けまたは取り付けるポートを変更する場合に、適切に取り扱う必要があります。

部品の機械的な性質が原因で、不適切な取り扱いによって、ポート・ソケットまたはケーブル・コネクタが損傷する可能性があります。よくあるのは、これらのポートとコネクタの不適切な接続による接触不良で、これはエラーや機能の中断の原因となることもあります。この問題は、ネットワーク・スイッチのハードウェアや、ホストのネットワーク・インタフェースにも当てはまります。

各ネットワーク・インタフェース・デバイスには、オペレーティング・システム側のドライバとやり取りをする固有のファームウェアが搭載されており、これによって機能が適切に使用できるようになっています。また、ネットワーク・スイッチにも、固有のファームウェアを必要とするアクティブなハードウェア・コンポーネントがあります。本書の中では、環境内にファームウェアが存在することを理解して、正しいバージョンがインストールされていることを確認できるようになることが重要です。これは、トラブルシューティング・フェーズにおけるチェックリスト項目の1つに加える必要があります。オラクルのエンジニアド・システム内にあるネットワークでは、IB HCAと10GbE NICにファームウェアが搭載されています。オラクルのInfiniBandスイッチの内部には、InfiniScale-IVやBridge-Xなどのファームウェアを必要とするいくつかのコンポーネントがありますが、これらのファームウェアは、簡単な管理を目的として個別のファームウェアの代わりに総合的なファームウェア・バンドル・パッケージを使用して管理されます。また、一部のInfiniBand専用アクティブ光ケーブルでも、固有のファームウェアが必要になることがあります。これらのファームウェアは、通常、それぞれのコンポーネントの内部にあるPROMに書き込まれています。

ハードウェア障害は複雑で、根本原因分析の段階にたどり着くために、場合によっては専用の診断ツールと機器が必要になることがあります。本書では、対象範囲が限られているため、そのような複雑な領域については取り上げません。トラブルシューティングのフローがこの方向に向かい、ハードウェアが実際に設置されている現場で問題が解決できない場合は、ほぼ確実にハードウェアの交換が必要になります。障害が発生したハードウェアは、適切な設備を備えた技術研究所で詳細に診断と分析を実施できます。

診断とトラブルシューティングのフェーズで障害を特定するのは困難です。そこで重要なのは、トラブルシューティング・フェーズの適切な時点で、ハードウェア検証のチェックリストを実施することです。検証は、常に、事前に設定された予想結果に対して行われます。

ネットワークを構成するソフトウェア

目的のビジネス・サービスを提供するソフトウェアがあるのと同じように、インフラストラクチャのレベルには、ネットワーク・ソフトウェアがあります。このカテゴリの中で最初にあるのが、デバイス・ドライバです。デバイス・ドライバは、オペレーティング・システムで有効にする必要があるそれぞれのハードウェアに固有のもので、個別のハードウェアと、上位レイヤーにあるプロトコルおよびアプリケーションとの間には、デバイス・ドライバによってバインディングが作成されます。

たとえば例を挙げると、Linuxの'eth0'という名前のインタフェースは、オペレーティング・システムに適切なデバイス・ドライバをインストールしなければ表示されません。このようなデバイス・ドライバのリストは、Linuxコマンドの`lsmod`の出力で確認できます。InfiniBandの場合、このようなデバイス・ドライバは、*Open Fabrics Enterprise Distribution (OFED)* と呼ばれるソフトウェア・パッケージの一部分となっています。また、このパッケージには、さまざまな構成ユーティリティ、診断ツール、API、ライブラリ、テスト・ツール、およびサービス・バイナリが含まれています。OFEDは、カーネルとオペレーティング・システムのレガシー・ネットワーク・スタックに緊密に統合されています。TCP/IPドライバ、構成、分析ツールのような、いくつかの重要なコンポーネントがOFEDによって使用されます。

このカテゴリの中で2番目にあるのが、ネットワーク・スイッチを制御するソフトウェアです。ここでの説明は一部分がファームウェアと重複しますが、スイッチの制御の中身を具体的に理解しておくことは重要です。ファームウェアとは、固定化されたブラック・ボックス・バージョンのソフトウェアです。イーサネットとInfiniBandのどちらでも、スイッチはそれぞれのファームウェアによって動作し、管理と構成に必要なユーザー・インタフェースを公開します。また、InfiniBandスイッチのソフトウェアには、ファブリック・レベルの管理機能を提供できるカスタマイズされたバージョンのOFEDが含まれています。InfiniBandスイッチでもっとも重要なコンポーネントの1つがサブネット・マネージャ (*Subnet Manager*) であり、OpenSMまたはOpen Source Subnet Managerとしても知られています。サブネット・マネージャの詳細については、次の項で説明します。スイッチに専用ハードウェアが搭載されている場合には、関連するソフトウェア・コンポーネントも組み込まれることになります。たとえば、InfiniBandゲートウェイ・スイッチには、仮想Ethernet over InfiniBand機能を提供するBridge-Xデバイスが搭載されています。

そのため、ソフトウェアまたはいわゆるファームウェアにも、**Bridge-X Manager (BXM)** が含まれています。

InfiniBandサブネット・マネージャ

サブネット・マネージャは、InfiniBandネットワークの中でもっとも重要なソフトウェアの1つです。オラクルのエンジニアド・システムでは、サブネット・マネージャは、互換性の理由からファームウェア・リビジョンが最新になっている一連のInfiniBandスイッチ上で動作します。サブネット・マネージャのもっとも目立つ機能の1つが、ファブリック内のすべてのエンティティにローカル識別子を割り当てて、それらの識別子に対するルーティング・テーブルを作成する機能です。注意しなければならないのは、このテーブルは、OSIモデルのレイヤー3にあるルーティング・テーブルではなく、レイヤー2にあることです。信頼性の低いデータグラム・サービスを介して管理パケットを交換するために、仮想レーン15 (VL15) が使用されます。ファブリックの初期検出とアクティブ化が完了した後は、サブネット・マネージャが、必要に応じて定期的にスキャンを実行して情報を更新します。高可用性と冗長性を確保するために、サブネット・マネージャの複数のインスタンスを作成することもできます。ただし、マスター・ロールの所有権を得られるのは1つのインスタンスだけなので、他のすべてのインスタンスはスタンバイ状態となります。サブネット・マネージャの構成ファイルでは、コマンドライン・インタフェースを使用していくつかのユーザー構成オプションを指定できます。スイッチをスパインに指定すると、リーフ・スイッチと比較してより高い優先順位が設定されます。制御ハンドオーバーは、すべてのInfiniBandスイッチでtrueに設定されます。2台以上のスイッチに同じ優先順位が割り当てられている場合は、マスター・ロールの優先権は、GUIDが小さい方のスイッチに与えられます。

また、サブネット・マネージャは、構成ファイルで設定されているロギング・レベル3に基づいて、定期的にイベントをログに記録します。一部のメッセージは、見ればすぐ分かりますが、その他は分かりにくいものとなっています。showsmllogコマンドを使用すると、サブネット・マネージャのログを表示できます。表1に、エラー・コードの一部を示します。

表1: サブネット・マネージャのエラー・コード

サブネット・マネージャのエラー・コード	説明
ERR 1B11	結合要求でパラメータが欠落していたことが原因で、新しいマルチキャスト・グループの作成に失敗したことを示しています。これらのメッセージは、サブネットの初期化とアクティブ化のフェーズでよく表示されます。再試行によって作成に成功した場合、これらのイベントはログに記録されません。
ERR 1B10	要求元がFullメンバーになることを希望していないことが理由で、失敗したことを示しています。
ERR 5411とERR 3113	メッセージ転送がタイムアウトになったことを示しています。この2つのエラー・メッセージは単一のイベントを表しており、通常はペアで表示されます。このタイプのメッセージは、SM-SM同期に使用されます。たとえば、スタンバイSMIは、マスターSMが有効であることをこのメッセージによって確認します。
ERR 3315	SMAが応答しないために、リモート側のノードが不明です。
ERR 3809	受信したトラップの送信元となっている物理ポートを発見できませんでした。
ERR 7502	LIDでポート・オブジェクトを特定できなかったことを示しています。
ERR 7503	キャッシュ情報が古いために、LIDが範囲外となっています。

InfiniBandパーティション

InfiniBandパーティションは、サブネット・マネージャによって厳密に管理されます。このパーティションは、ファブリック内にあるノードの論理グループを定義します。マルチホームのコンピューティング・プラットフォームを考えると、InfiniBandパーティションを理解できます。イーサネット環境に例えると、InfiniBandパーティションはVLANと非常に良く似ています。InfiniBand HCAを複数のパーティションに追加することによって、異なるネットワーク・パスを介して通信する機能を有効にできます。すべての構成をマスター・サブネット・マネージャで実行し、最終的にコミットした構成をすべてのスタンバイ・スイッチに伝播できます。最初にパーティション・キーを定義してから、次にHCAポートのGUIDをメンバーとして追加する必要があります。メンバーシップには、LimitedとFullの2種類があります。

表2： InfiniBandパーティションのメンバーシップ

InfiniBandパーティションのメンバーシップ	役割と説明
Limited	同じパーティション内のFullメンバーと通信できますが、別のLimitedメンバーと通信することはできません。
Full	同じパーティションのすべてのメンバーと通信ができます。
Both	Oracle VM Server (OVS) ノード用の特別なメンバーシップで、そのリソースでホストするゲスト仮想マシンにLimitedとFullメンバーシップを割り当てることができます。

ソフトウェア・レベルでのデバッグ

ほとんどの場合、診断とトラブルシューティングのフェーズは、ソフトウェア・レベルで始まります。問題の原因がユーザー・エラーや構成の構文エラーのように単純な場合もありますが、実際の原因が他の部分にあつて影響が他の領域にまで及ぶためにデバッグが困難になる場合もあります。状況によっては、これらの問題は特定の領域に収まらず、相互運用性の課題に対応することが必要となります。このようなトラブルシューティングのシナリオについては、本書の全体を通してさらに詳しく説明していきます。

さまざまなコマンドとトラブルシューティングのシナリオを評価する前に、InfiniBandファブリックを構築するための主要なソフトウェア機能のいくつかについて理解しておくことが重要です。この後にある各項では、サブネット・マネージャとパーティションについて確認していきます。

ここで重要なのは、ソフトウェアのバージョンが正しいことを確認して、ネットワーク・インフラストラクチャ内にあるコンポーネントの適切な互換性と相互運用性を確保することです。

ヘルス・チェックのガイドライン

非常に大きく分けると、2種類の診断コマンドがあります。一部のコマンドはブール型の結果を返しますが、その他のコマンドは本質的に非常に主観的です。前者のカテゴリにあるのは、簡単で分かりやすいコマンドです。これらのコマンドの出力は、合格または不合格という形で結果を提示するので、複雑な分析は必要ありません。たとえば、`sminfo`はマスター・サブネット・マネージャについての情報を返します。状況によっては、この出力で十分かもしれませんが、返されたマスター・サブネット・マネージャについての情報が想定した範囲内にあるかどうかをさらに分析することが必要になる場合もあります。

この後にある各項では、構成設計図について説明します。簡単に言えば、トラブルシューティング・フェーズになったときにその時点での状態を照合確認できるように、ネットワーク構成の状態を事前に定義しておく必要があるということです。

問題が出てきた場合に備えて、使用されているテクノロジーに対して利用可能なさまざまな診断ツールやコマンドを熟知し精通していることが重要です。このテクノロジーは、レイヤー1~2にはInfiniBandや10GbE、レイヤー3~4にはTCP/IPなどの環境内のさまざまなテクノロジーを組み合わせたものになります。ハードウェア・コンポーネントにも同様に固有のコマンド・セットがあります。このような知識があれば、知識を分類して適用することで、トラブルシューティングを系統的なアプローチによって実施できます。

NTPサービスの重要性

時刻同期サービスであるNTPは、最初から設定しておくことが非常に重要です。クラスタ化されたコンピューティング環境には常にいくつものコンポーネントがあるため、すべてのコンポーネント間で、共通のソースを使用してシステム・クロックを同期する必要があります。トラブルシューティングの実施中は、個々のコンポーネントでログに記録されたイベントを分析して相互に関連付けます。多くの場合、これらのイベント・ログはミリ秒単位で記録されるので、順序とシーケンスは極めて重要な役割を果たします。

ログのローテーションとアーカイブ

すべてのホストとスイッチは、定期的にログ・ファイルのローテーションを実行するメカニズムを実装しています。コンピューティング・ホストには過去のアーカイブがいくつもありますが、NM2スイッチは、ディスク領域に制限があるため最新のアーカイブを1つだけ保持しています。トラブルシューティング・フェーズでNM2スイッチの外部にログ・ファイルをコピーしておくと、役立つことがあります。このコピーは、セキュア・コピー（`scp`）コマンドを使用して定期的に行うことができます。または、外部マシンから`cronjob`を使用して、同じようにコピーを実行することもできます。

構成設計図

設計図は固有であり、それぞれのインストールに一意のものであります。想定している構成とスナップショットを照合して検証するために、このような構成設計図を保持しておくことが重要です。トラブルシューティングで使用されるいくつかのコマンドが非常に主観的な出力を生成するので、このような出力を事前に定義された既知の状態と比較して検証する必要があります。工場出荷時のデフォルト構成を使用している場合は、このような設計図がすでに利用できるようになっています。ただし、ほとんどの場合、標準の出荷時構成との間に何らかの差異があります。そのもっとも一般的な例が、IPアドレス、ホスト名、外部接続です。

ここでの推奨事項およびベスト・プラクティスは、最初のインストールを実施した直後に、良好であることが確認されている構成の状態を保存しておくことです。これらの情報は、後でトラブルシューティングが必要となった場合に非常に役立ちます。

InfiniBandトポロジの保存

ここでは、InfiniBandファブリックについての重要情報を保存する方法を、基本的なシナリオを使用して説明します。

iblinkinfo.pl - ファブリック内にあるすべてのリンクのリンク情報を表示します。

```
[root@scab01db01 ~]# iblinkinfo.pl -R
Switch 0x0021283a866aa0a0 SUN DCS 36P QDR scab01sw-ib3:
 18 1[ ] == ( 4X 2.5 Gbps Down/ Polling) ==> [ ] "" ( )
 18 2[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 41 1[ ] "scab01cel101 C 192.168.40.29 HCA-1" ( )
 18 3[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 42 2[ ] "scab01cel104 C 192.168.75.14 HCA-1" ( )
 18 4[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 46 2[ ] "scab01cel103 C 192.168.40.31 HCA-1" ( )
 18 5[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 32 2[ ] "scab01cel106 C 192.168.40.34 HCA-1" ( )
 18 6[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 36 2[ ] "scab01cel105 C 192.168.40.33 HCA-1" ( )
 18 7[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 2 2[ ] "scab01db01 S 192.168.40.21 HCA-1" ( )
 18 8[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 25 2[ ] "scab01cel107 C 192.168.40.35 HCA-1" ( )
 18 9[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 33 2[ ] "scab01db03 S 192.168.40.23 HCA-1" ( )
 18 10[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 52 2[ ] "scab01db02 S 192.168.40.22 HCA-1" ( )
 18 11[ ] == ( 4X 2.5 Gbps Down/ Polling) ==> [ ] "" ( )
 18 12[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 31 2[ ] "scab01db04 S 192.168.40.24 HCA-1" ( )
 18 13[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 14[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 14[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 13[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 15[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 16[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 16[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 15[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 17[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 18[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 18[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 17[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 19[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 19 1[ ] "scab01cel113 C 192.168.40.41 HCA-1" ( )
 18 20[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 21 1[ ] "scab01cel114 C 192.168.40.42 HCA-1" ( )
 18 21[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 29 1[ ] "scab01cel111 C 192.168.75.21 HCA-1" ( )
 18 22[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 26 1[ ] "scab01cel112 C 192.168.40.40 HCA-1" ( )
 18 23[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 43 1[ ] "scab01cel109 C 192.168.40.37 HCA-1" ( )
 18 24[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 23 1[ ] "scab01cel110 C 192.168.40.38 HCA-1" ( )
 18 25[ ] == ( 4X 2.5 Gbps Down/ Disabled) ==> [ ] "" ( )
 18 26[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 39 1[ ] "scab01cel108 C 192.168.40.36 HCA-1" ( )
 18 27[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 11 1[ ] "scab01db06 S 192.168.40.26 HCA-1" ( )
 18 28[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 10 1[ ] "scab01db07 S 192.168.40.27 HCA-1" ( )
 18 29[ ] == ( 4X 2.5 Gbps Down/ Polling) ==> [ ] "" ( )
 18 30[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 48 1[ ] "scab01db05 S 192.168.40.25 HCA-1" ( )
 18 31[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 1 31[ ] "SUN DCS 36P QDR scab01sw-ib2" ( )
 18 32[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 38 2[ ] "scab01cel102 C 192.168.40.30 HCA-1" ( )
 18 33[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 5 1[ ] "MT25408 ConnectX Mellanox Technologies" ( )
 18 34[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 16 2[ ] "MT25408 ConnectX Mellanox Technologies" ( )
 18 35[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 7 2[ ] "MT25408 ConnectX Mellanox Technologies" ( )
 18 36[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 14 1[ ] "MT25408 ConnectX Mellanox Technologies" ( )
Switch 0x002128469727a0a0 SUN DCS 36P QDR scab01sw-ib2:
 1 1[ ] == ( 4X 2.5 Gbps Down/ Polling) ==> [ ] "" ( )
 1 2[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 45 2[ ] "scab01cel101 C 192.168.40.29 HCA-1" ( )
 1 3[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 9 1[ ] "scab01cel104 C 192.168.75.14 HCA-1" ( )
 1 4[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 35 1[ ] "scab01cel103 C 192.168.40.31 HCA-1" ( )
 1 5[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 13 1[ ] "scab01cel106 C 192.168.40.34 HCA-1" ( )
 1 6[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 4 1[ ] "scab01cel105 C 192.168.40.33 HCA-1" ( )
 1 7[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 8 1[ ] "scab01db01 S 192.168.40.21 HCA-1" ( )
```

```
1 8[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 34 1[ ] "scab01cel107 C 192.168.40.35 HCA-1" ( )
1 9[ ] == ( 4X 10.0 Gbps Active/ LinkUp) ==> 28 1[ ] "scab01db03 S 192.168.40.23 HCA-1" ( )
```

```

1    10[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 51  1[ ] "scab01db02 S 192.168.40.22 HCA-1" ( )
1    11[ ] == ( 4X 2.5 Gbps Down/ Polling)==>  [ ] "" ( )
1    12[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==>  3  1[ ] "scab01db04 S 192.168.40.24 HCA-1" ( )
1    13[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 14[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    14[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 13[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    15[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 16[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    16[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 15[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    17[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 18[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    18[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 17[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    19[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 20 2[ ] "scab01cel13 C 192.168.40.41 HCA-1" ( )
1    20[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 22 2[ ] "scab01cel14 C 192.168.40.42 HCA-1" ( )
1    21[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 30 2[ ] "scab01cel11 C 192.168.75.21 HCA-1" ( )
1    22[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 27 2[ ] "scab01cel12 C 192.168.40.40 HCA-1" ( )
1    23[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 44 2[ ] "scab01cel09 C 192.168.40.37 HCA-1" ( )
1    24[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 24 2[ ] "scab01cel10 C 192.168.40.38 HCA-1" ( )
1    25[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 37 2[ ] "scab01db08 S 192.168.40.28 HCA-1" ( )
1    26[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 40 2[ ] "scab01cel08 C 192.168.40.36 HCA-1" ( )
1    27[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 17 2[ ] "scab01db06 S 192.168.40.26 HCA-1" ( )
1    28[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 12 2[ ] "scab01db07 S 192.168.40.27 HCA-1" ( )
1    29[ ] == ( 4X 2.5 Gbps Down/ Polling)==>  [ ] "" ( )
1    30[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==>  6  2[ ] "scab01db05 S 192.168.40.25 HCA-1" ( )
1    31[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 18 31[ ] "SUN DCS 36P QDR scab01sw-ib3" ( )
1    32[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 15 1[ ] "scab01cel02 C 192.168.40.30 HCA-1" ( )
1    33[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 49 1[ ] "MT25408 ConnectX Mellanox Technologies" ( )
1    34[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 50 1[ ] "MT25408 ConnectX Mellanox Technologies" ( )
1    35[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 47 2[ ] "MT25408 ConnectX Mellanox Technologies" ( )
1    36[ ] == ( 4X 10.0 Gbps Active/ LinkUp)==> 53 2[ ] "MT25408 ConnectX Mellanox Technologies" ( )

```

ibdiagnet -ls 10 -lw 4x -vlr

ここで非常に重要なのは、コマンドの実行後に/tmp/ibdiagnet.*を保存することです。

ibnetdiscover - InfiniBandトポロジを検出します。

```

vendid=0x2c9
devid=0x673c
sysimgguid=0x2c903000a740f
caguid=0x2c903000a740c
Ca      2 "H-0002c903000a740c"          # "scac01cel14 C 192.168.40.92 HCA-1"
[2](2c903000a740e)      "S-002128468d20a0a0"[20]      # lid 113 lmc 0 "SUN DCS 36P
QDR scac01sw-ib2 10.133.40.117" lid 14 4xQDR
[1](2c903000a740d)      "S-002128468d57a0a0"[20]      # lid 112 lmc 0 "SUN DCS 36P
QDR scac01sw-ib3 10.133.40.118" lid 33 4xQDR

vendid=0x2c9
devid=0x673c
sysimgguid=0x2c903000a77b3
caguid=0x2c903000a77b0
Ca      2 "H-0002c903000a77b0"          # "scac01cel13 C 192.168.48.253 HCA-1"
[2](2c903000a77b2)      "S-002128468d20a0a0"[19]      # lid 58 lmc 0 "SUN DCS 36P
QDR scac01sw-ib2 10.133.40.117" lid 14 4xQDR
[1](2c903000a77b1)      "S-002128468d57a0a0"[19]      # lid 57 lmc 0 "SUN DCS 36P
QDR scac01sw-ib3 10.133.40.118" lid 33 4xQDR

vendid=0x2c9
devid=0x673c
sysimgguid=0x2c903000a73bb
caguid=0x2c903000a73b8
Ca      2 "H-0002c903000a73b8"          # "scab01db04 S 192.168.10.103 HCA-1"
[1](2c903000a73b9)      "S-002128468d20a0a0"[12]      # lid 17 lmc 0 "SUN DCS 36P

```

```
QDR scac01sw-ib2 10.133.40.117" lid 14 4xQDR
[2](2c903000a73ba) "S-002128468d57a0a0"[12] # lid 35 lmc 0 "SUN DCS 36P
QDR scac01sw-ib3 10.133.40.118" lid 33 4xQDR
```

ibswitches - トポロジ内のInfiniBandスイッチ・ノードを表示します。

```
[root@scab01db01 ~]# ibswitches
Switch : 0x0021283a866aa0a0 ports 36 "SUN DCS 36P QDR scab01sw-ib3.us.oracle.com" enhanced
port 0 lid 18 lmc 0
Switch : 0x002128469727a0a0 ports 36 "SUN DCS 36P QDR scab01sw-ib2.us.oracle.com" enhanced
port 0 lid 1 lmc 0
```

ibhosts - InfiniBandファブリック内のホストとそのチャネル・アダプタを表示します。

```
[root@scab01db01 ~]# ibhosts
Ca      : 0x0021280001cef69a ports 2 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0021280001cef60a ports 2 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0021280001cef68e ports 2 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0021280001cef626 ports 2 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x00212800013e6982 ports 2 "scab01cel02 C 192.168.40.30 HCA-1"
Ca      : 0x00212800013e69ee ports 2 "scab01db05 S 192.168.40.25 HCA-1"
Ca      : 0x00212800013e695e ports 2 "scab01db07 S 192.168.40.27 HCA-1"
Ca      : 0x00212800013e699a ports 2 "scab01db06 S 192.168.40.26 HCA-1"
Ca      : 0x00212800013e6986 ports 2 "scab01cel08 C 192.168.40.36 HCA-1"
Ca      : 0x00212800013e69a6 ports 2 "scab01db08 S 192.168.40.28 HCA-1"
Ca      : 0x00212800013e6966 ports 2 "scab01cel10 C 192.168.40.38 HCA-1"
Ca      : 0x00212800013e6a06 ports 2 "scab01cel09 C 192.168.40.37 HCA-1"
Ca      : 0x00212800013e696e ports 2 "scab01cel12 C 192.168.40.40 HCA-1"
Ca      : 0x00212800013e697e ports 2 "scab01cel11 C 192.168.75.21 HCA-1"
Ca      : 0x00212800013e698e ports 2 "scab01cel14 C 192.168.40.42 HCA-1"
Ca      : 0x00212800013e6962 ports 2 "scab01cel13 C 192.168.40.41 HCA-1"
Ca      : 0x00212800013e69ae ports 2 "scab01db04 S 192.168.40.24 HCA-1"
Ca      : 0x00212800013e69f2 ports 2 "scab01db02 S 192.168.40.22 HCA-1"
Ca      : 0x00212800013e69ca ports 2 "scab01db03 S 192.168.40.23 HCA-1"
Ca      : 0x00212800013e69fe ports 2 "scab01cel07 C 192.168.40.35 HCA-1"
Ca      : 0x00212800013e6a02 ports 2 "scab01cel05 C 192.168.40.33 HCA-1"
Ca      : 0x00212800013e697a ports 2 "scab01cel06 C 192.168.40.34 HCA-1"
Ca      : 0x00212800013e69d6 ports 2 "scab01cel03 C 192.168.40.31 HCA-1"
Ca      : 0x00212800013e695a ports 2 "scab01cel04 C 192.168.75.14 HCA-1"
Ca      : 0x00212800013e6a0e ports 2 "scab01cel01 C 192.168.40.29 HCA-1"
Ca      : 0x00212800013e6a16 ports 2 "scab01db01 S 192.168.40.21 HCA-1"
```

InfiniBand診断ツール

利用可能なInfiniBandコマンドは、さまざまなレベルに分類できます。これによって、後の項で説明するトラブルシューティングのアプローチが簡素化されます。一般に、一部のコマンドはシステム固有のものであり、その他はグローバル・コマンドとなっています。また、システム固有のコマンドは、システムやオペレーティング・システムなどの種類によっても分類できます。

ホストに固有のコマンド

パケット・フロー全体で見た場合、InfiniBandリンクの終点となるエンティティがホストです。通常、このホストには、デュアル・ポートのIB HCAを搭載しています。Linuxでは、次のコマンドを使用して、このエンド・ポイントのヘルス・チェックを実行できます。

lspci - システム内にあるすべてのPCIバスとそれらに接続されているすべてのデバイスに関する情報を表示するユーティリティです。

```
[root@scae01cn01 ~]# lspci | grep -i InfiniBand
19:00.0 InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR /
10GigE] (rev b0)

[root@scae01cn01 ~]# lspci -vvv -s 19:00.0
19:00.0 InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR /
10GigE]
Subsystem: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR /
10GigE]
Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr+ Stepping- SERR+
FastB2B-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort-
>SERR- <PERR-
Latency: 0, Cache Line Size: 256 bytes
Interrupt: pin A routed to IRQ 32
Region 0: Memory at df600000 (64-bit, non-prefetchable) [size=1M]
Region 2: Memory at de800000 (64-bit, prefetchable) [size=8M]
Capabilities: [40] Power Management version 3
Flags: PMEClk- DSI- D1- D2- AuxCurrent=0mA PME(D0-,D1-,D2-,D3hot-,D3cold-)
Status: D0 PME-Enable- DSel=0 DScale=0 PME-
Capabilities: [48] Vital Product Data
Capabilities: [9c] MSI-X: Enable+ Mask- TabSize=256
Vector table: BAR=0 offset=0007c000
PBA: BAR=0 offset=0007d000
Capabilities: [60] Express Endpoint IRQ 0
Device: Supported: MaxPayload 256 bytes, PhantFunc 0, ExtTag-
Device: Latency L0s <64ns, L1 unlimited
Device: AtnBtn- AtnInd- PwrInd-
Device: Errors: Correctable+ Non-Fatal+ Fatal+ Unsupported-
Device: RlxdOrd- ExtTag- PhantFunc- AuxPwr- NoSnoop-
Device: MaxPayload 128 bytes, MaxReadReq 512 bytes
Link: Supported Speed unknown, Width x8, ASPM L0s, Port 8
Link: Latency L0s unlimited, L1 unlimited
Link: ASPM Disabled RCB 64 bytes CommClk- ExtSynch-
Link: Speed unknown, Width x8
Capabilities: [100] Unknown (14)
```

ibv_devices - システムに取り付けられているRDMA対応デバイスを一覧表示します。

```
[root@scae01cn01 ~]# ibv_devices
device          node GUID
-----
mlx4_0          0021280001a0a384
```

ibv_devinfo - システム内にあるRDMA対応デバイスを問い合わせます。

```
[root@scae01cn01 ~]# ibv_devinfo
hca_id: mlx4_0
  transport: InfiniBand (0)
  fw_ver: 2.7.8130
  node_guid: 0021:2800:01a0:a384
  sys_image_guid: 0021:2800:01a0:a387
  vendor_id: 0x02c9
  vendor_part_id: 26428
  hw_ver: 0xB0
  board_id: SUN0170000009
  phys_port_cnt: 2
    port: 1
      state: PORT_ACTIVE (4)
      max_mtu: 2048 (4)
      active_mtu: 2048 (4)
      sm_lid: 61
      port_lid: 53
      port_lmc: 0x00
      link_layer: IB
    port: 2
      state: PORT_ACTIVE (4)
      max_mtu: 2048 (4)
      active_mtu: 2048 (4)
      sm_lid: 61
      port_lid: 54
      port_lmc: 0x00
      link_layer: IB
```

mstflint - Mellanox InfiniBand HCAとイーサネットNICカードのFW（ファームウェア）書き込みとフラッシュ・メモリ操作を実行するためのツールです。

```
[root@scae01cn01 ~]# mstflint -d mlx4_0 q
Image type: ConnectX
FW Version: 2.7.8130
Device ID: 26428
Chip Revision: B0
Description: Node Port1 Port2 Sys image
GUIDs: 0021280001a0a384 0021280001a0a385 0021280001a0a386 0021280001a0a387
MACs: 002128a0a384 002128a0a385
Board ID: (SUN0170000009)
VSD:
PSID: SUN0170000009
```

ibstat - システムに取り付けられているInfiniBandデバイスの基本ステータスを問い合わせます。

```
[root@scae01cn01 ~]# ibstat
CA 'mlx4_0'
  CA type: MT26428
  Number of ports: 2
  Firmware version: 2.7.8130
  Hardware version: b0
  Node GUID: 0x0021280001a0a384
  System image GUID: 0x0021280001a0a387
  Port 1:
    State: Active
    Physical state: LinkUp
    Rate: 40
    Base lid: 53
    LMC: 0
    SM lid: 61
    Capability mask: 0x02510868
    Port GUID: 0x0021280001a0a385
    Link layer: IB
  Port 2:
    State: Active
    Physical state: LinkUp
    Rate: 40
    Base lid: 54
    LMC: 0
    SM lid: 61
    Capability mask: 0x02510868
    Port GUID: 0x0021280001a0a386
    Link layer: IB
```

ibstatus

```
[root@scae01cn01 ~]# ibstatus
Infiniband device 'mlx4_0' port 1 status:
  default gid:    fe80:0000:0000:0000:0021:2800:01a0:a385
  base lid:      0x35
  sm lid:        0x3d
  state:         4: ACTIVE
  phys state:    5: LinkUp
  rate:          40 Gb/sec (4X QDR)
  link_layer:    IB

Infiniband device 'mlx4_0' port 2 status:
  default gid:    fe80:0000:0000:0000:0021:2800:01a0:a386
  base lid:      0x36
  sm lid:        0x3d
  state:         4: ACTIVE
  phys state:    5: LinkUp
  rate:          40 Gb/sec (4X QDR)
  link_layer:    IB
```

perfquery - InfiniBandポートのパフォーマンス・カウンタを問い合わせます。

```
[root@scae01cn01 ~]# perfquery
# Port counters:Lid 53 port 1
PortSelect:.....1
CounterSelect:.....0x0400
SymbolErrors:.....0
LinkRecovers:.....0
LinkDowned:.....0
RcvErrors:.....0
RcvRemotePhysErrors:.....0
RcvSwRelayErrors:.....0
XmtDiscards:.....0
XmtConstraintErrors:.....0
RcvConstraintErrors:.....0
CounterSelect2:.....0x00
LinkIntegrityErrors:.....0
ExcBufOverrunErrors:.....0
VL15Dropped:.....0
XmtData:.....1915835586
RcvData:.....4294967295
XmtPkts:.....27742207
RcvPkts:.....106026118
```

perfqueryカウンタの説明

表3 : perfqueryカウンタ

perfqueryカウンタ	説明
SymbolErrorCounter	1つまたは複数の物理レーンで検出された、リンクのマイナー・エラーの総数です。
LinkErrorRecovery	ポートがトレーニング状態になっているマシンが、リンク・エラーのリカバリ処理を正常に完了した回数の合計です。
LinkDownedCounter	ポートがトレーニング状態になっているマシンが、リンク・エラーのリカバリ処理に失敗してリンクが停止した回数の合計です。
PortRcvErrors	ポートで受信したパケットに含まれているエラーの総数です。 <ul style="list-style-type: none"> ローカルの物理エラー 不正なデータまたはリンクのパケット・エラー バッファ・オーバーランが原因で破棄されたパケット
PortRcvRemotePhysicalErrors	ポートで受信した、EBPデリミタ付きパケットの総数です。
PortRcvSwitchRelayErrors	ポートで受信したパケットのうち、スイッチ・リレーによって転送できなかったことが原因で破棄されたパケットの総数です。
PortXmitDiscards	ポートの停止または輻輳が原因でポートによって破棄された、アウトバウンド・パケットの総数です。 <ul style="list-style-type: none"> 出力ポートがアクティブ状態になっていません。 パケット長が近接するMTUを超過しました。 スイッチのライフタイム制限を超過しました。

	<ul style="list-style-type: none"> • スイッチのHOQライフタイム制限を超過しました。 <p>また、このようなパケットには、VLが停止状態の間に破棄されたパケットも含まれます。</p>
PortXmitConstraintErrors	<p>次の理由で、スイッチの物理ポートから送信されなかったパケットの総数です。</p> <ul style="list-style-type: none"> • FilterRawOutboundがtrueで、パケットがrawになっています。 • PartitionEnforcementOutboundがtrueで、パケットがパーティション・キーのチェックまたはIPバージョンのチェックに失敗しました。
PortRcvConstraintErrors	<p>スイッチの物理ポートで受信したパケットのうち、次の理由によって破棄されたパケットの総数です。</p> <ul style="list-style-type: none"> • FilterRawInboundがtrueで、パケットがrawになっています。 • PartitionEnforcementInboundがtrueで、パケットがパーティション・キーのチェックまたはIPバージョンのチェックに失敗しました。
LocalLinkIntegrityErrors	ローカル物理エラーの数が、LocalPhyErrorsによって指定されたしきい値を超えた回数です。
ExcessiveBufferOverRunErrors	連続してOverrunErrorsのフロー制御更新期間が発生し、そのたびに1回以上のオーバーラン・エラーが発生した回数です。
VL15Dropped	受信したVL15パケットが、ポートのリソース制限（例：バッファ不足）が原因で破棄された回数です。
PortXmitData	ポートからすべてのVL上で送信されたデータ・オクテットの総数です（4で除算した値）。この数には、パケット・デリミタの開始からVCRCまでの（デリミタとVCRCは除く）すべてのオクテットを含みます。また、エラーのあるパケットが含まれている可能性があります。
PortRcvData	ポートにおいてすべてのVL上で受信されたデータ・オクテットの総数です（4で除算した値）。この数には、パケット・デリミタの開始からVCRCまでの（デリミタとVCRCは除く）すべてのオクテットを含みます。また、エラーのあるパケットが含まれている可能性があります。
PortXmitPkts	ポートからすべてのVL上で送信されたパケットの総数です。この数には、エラーのあるパケットが含まれている可能性があります。
PortRcvPkts	ポートにおいてすべてのVLから受信されたパケットの総数です。ただし、エラーのあるパケットは含みますが、リンク・パケットは除外しています。
PortXmitWait	PortSelectによって選択されたポートに送信するデータがあるにも関わらず、クレジットの不足またはアービトレーションの欠如のいずれかが原因でその全ティック期間中にデータが送信されなかったティックの数です。

InfiniBandスイッチに固有のコマンド

env_test - Oracle InfiniBandスイッチの環境をテストする自己診断機能です。さまざまな内部コンポーネントのテストを実行して、レポートを出力します。

```
[root@scae01sw-ib02 ~]# env_test
Environment test started:
Starting Environment Daemon test:
Environment daemon running
Environment Daemon test returned OK
Starting Voltage test:
Voltage ECB OK
Measured 3.3V Main = 3.27 V
Measured 3.3V Standby = 3.37 V
Measured 12V = 11.90 V
Measured 5V = 5.02 V
```

```

Measured VBAT = 3.04 V
Measured 1.0V = 1.01 V
Measured I4 1.2V = 1.22 V
Measured 2.5V = 2.50 V
Measured V1P2 DIG = 1.17 V
Measured V1P2 ANG = 1.17 V
Measured 1.2V BridgeX = 1.22 V
Measured 1.8V = 1.78 V
Measured 1.2V Standby = 1.19 V
Voltage test returned OK
Starting PSU test:
PSU 0 present OK
WARNING PSU 1 present AC Loss
PSU test returned 1 faults
Starting Temperature test:
Back temperature 28
Front temperature 25
SP temperature 40
Switch temperature 53, maxtemperature 55
Bridge-0 temperature 42, maxtemperature 44
Bridge-1 temperature 44, maxtemperature 45
Temperature test returned OK
Starting FAN test:
Fan 0 not present
Fan 1 running at rpm 12426
Fan 2 running at rpm 12317
Fan 3 running at rpm 12317
Fan 4 not present
FAN test returned OK
Starting Connector test:
Connector test returned OK
Starting Onboard ibdevice test:
Switch OK
Bridge-0 OK
Bridge-1 OK
All Internal ibdevices OK
Onboard ibdevice test returned OK
Environment test FAILED

```

Listlinkup - コネクタ・ラベルを表示して、論理ポート情報およびリンク状態を切り替えます。

```

[root@scae01sw-ib02 ~]# listlinkup
Connector 0A Present <-> Switch Port 20 up (Enabled)
Connector 1A Present <-> Switch Port 22 up (Enabled)
Connector 2A Not present
Connector 3A Not present
Connector 4A Not present
Connector 5A Not present
Connector 6A Present <-> Switch Port 35 up (Enabled)
Connector 7A Present <-> Switch Port 33 up (Enabled)
Connector 8A Present <-> Switch Port 31 up (Enabled)
Connector 9A Present <-> Switch Port 14 down (Enabled)
Connector 10A Present <-> Switch Port 16 up (Enabled)
Connector 11A Present <-> Switch Port 12 up (Enabled)
Connector 12A Present <-> Switch Port 18 down (Enabled)

```

```
Connector 13A Present <-> Switch Port 9 up (Enabled)
Connector 14A Present <-> Switch Port 7 up (Enabled)
Connector 15A Present <-> Switch Port 5 up (Enabled)
Connector 0A-ETH Present
  Bridge-0 Port 0A-ETH-1 (Bridge-0-2) up (Enabled)
  Bridge-0 Port 0A-ETH-2 (Bridge-0-2) up (Enabled)
  Bridge-0 Port 0A-ETH-3 (Bridge-0-1) down (Enabled)
  Bridge-0 Port 0A-ETH-4 (Bridge-0-1) down (Enabled)
Connector 1A-ETH Present
  Bridge-1 Port 1A-ETH-1 (Bridge-1-2) up (Enabled)
  Bridge-1 Port 1A-ETH-2 (Bridge-1-2) up (Enabled)
  Bridge-1 Port 1A-ETH-3 (Bridge-1-1) down (Enabled)
  Bridge-1 Port 1A-ETH-4 (Bridge-1-1) down (Enabled)
Connector 0B Present <-> Switch Port 19 up (Enabled)
Connector 1B Present <-> Switch Port 21 up (Enabled)
Connector 2B Not present
Connector 3B Not present
Connector 4B Not present
Connector 5B Not present
Connector 6B Present <-> Switch Port 36 up (Enabled)
Connector 7B Present <-> Switch Port 34 up (Enabled)
Connector 8B Present <-> Switch Port 32 up (Enabled)
Connector 9B Present <-> Switch Port 13 up (Enabled)
Connector 10B Present <-> Switch Port 15 up (Enabled)
Connector 11B Present <-> Switch Port 17 up (Enabled)
Connector 12B Present <-> Switch Port 11 up (Enabled)
Connector 13B Present <-> Switch Port 10 up (Enabled)
Connector 14B Present <-> Switch Port 8 up (Enabled)
Connector 15B Present <-> Switch Port 6 up (Enabled)
Connector 0B-FC Not present
Connector 1B-FC Not present
```

Version - スイッチのファームウェア・バージョンとシリアル番号を表示します。

```
[root@scae01sw-ib02 ~]# version
SUN DCS gw version: 2.0.6-1
Build time: Jan 17 2012 14:29:13
FPGA version: 0x33
SP board info:
Manufacturing Date: 2010.05.05
Serial Number: "NCD4J0943"
Hardware Revision: 0x0006
Firmware Revision: 0x0102
BIOS version: NOW1R112
BIOS date: 04/24/2009
```

smpartition list active - 構成されているパーティションの情報を表示します。

```
[root@scae01sw-ib02 ~]# smpartition list active
# Sun DCS IB partition config file
# This file is generated, do not edit
#! version_number :333
Default=0x7fff, ipoib :
ALL_CAS=full,
ALL_SWITCHES=full,
SELF=full;
SUN_DCS=0x0001, ipoib :
ALL_SWITCHES=full;
  = 0x8001, ipoib:
0x0021280001a0a5c6=both,
0x0021280001a0a5c5=both,
0x0021280001a0a36e=both,
0x0021280001a0a36d=both,
0x00212800013ea434=full,
0x00212800013ea433=full,
0x00212800013ea3ec=full,
0x00212800013ea3eb=full,
0x0021280001a0a30e=both,
0x0021280001a0a30d=both,
0x0021280001a0a392=both,
0x0021280001a0a391=both,
0x0002c903000a7776=both,
0x0002c903000a7775=both;
  = 0x8002, ipoib:
0x0021280001a0a5c6=both,
0x0021280001a0a5c5=both,
0x0021280001a0a36e=both,
0x0021280001a0a36d=both,
0x00212800013ea434=full,
0x00212800013ea433=full,
0x00212800013ea3ec=full,
0x00212800013ea3eb=full,
0x0021280001a0a30e=both,
0x0021280001a0a30d=both,
0x0021280001a0a392=both,
0x0021280001a0a391=both,
0x0002c903000a7776=both,
0x0002c903000a7775=both;
```

setsmpriority list - サブネット・マネージャの優先順位、ハンドオーバー状態、サブネット接頭辞、M_Key値を表示します。

```
[root@scae01sw-ib02 ~]# setsmpriority list
Current SM settings:
smpriority 5
controlled_handover TRUE
subnet_prefix 0xfe80000000000000
M_Key None
```

Showsmlog - サブネット・マネージャのログを表示します。

```
-----
OpenSM 3.2.6_20120116 - Oracle patch 10.5.2

Jul 13 04:47:22 649004 [B7EFB8D0] 0x80 -> OpenSM 3.2.6_20120116 - Oracle patch 10.5.2
Entering DISCOVERING state

Jul 13 04:47:22 684996 [B7EFB8D0] 0x80 -> Entering DISCOVERING state
Using default GUID 0x2128547f22c0a0
Entering STANDBY state

Jul 13 04:47:22 695994 [B7EFB8D0] 0x02 -> osm_vendor_bind: Binding to port 0x2128547f22c0a0,
class 129 version 1
Jul 13 04:47:22 722987 [B7EFB8D0] 0x02 -> osm_vendor_bind: Binding to port 0x2128547f22c0a0,
class 3 version 2
Jul 13 04:47:22 738984 [B7EFB8D0] 0x02 -> osm_console_init: Console listening on port 10000
Jul 13 04:47:22 965932 [B66F7B90] 0x80 -> Entering STANDBY state
Jul 14 00:00:47 841754 [B6EF8B90] 0x01 -> osm_vendor_send: ERR 5430: Send p_madw = 0x8244b48
of size 256 TID 0x455c540004639e failed -5
Jul 14 00:00:47 841754 [B6EF8B90] 0x01 -> vl15_send_mad: ERR 3E03: MAD send failed
(IB_UNKNOWN_ERROR)
Jul 20 19:26:18 802645 [B6EF8B90] 0x01 -> osm_vendor_send: ERR 5430: Send p_madw = 0x824cff8
of size 256 TID 0x455c540018e943 failed -5
Jul 20 19:26:18 803645 [B6EF8B90] 0x01 -> vl15_send_mad: ERR 3E03: MAD send failed
(IB_UNKNOWN_ERROR)
Jul 23 05:57:46 940382 [B6EF8B90] 0x01 -> osm_vendor_send: ERR 5430: Send p_madw = 0x824cff8
of size 256 TID 0x455c54001ef85f failed -5
Jul 23 05:57:46 941382 [B6EF8B90] 0x01 -> vl15_send_mad: ERR 3E03: MAD send failed
(IB_UNKNOWN_ERROR)
OpenSM: Got signal 15 - exiting...
Exiting SM

Jul 27 18:08:45 154750 [B7EFB8D0] 0x80 -> Exiting SM
```

Getmaster - サブネット・マネージャのマスター・ロールに関する情報を表示します。

```
[root@scae01sw-ib02 ~]# getmaster
Local SM enabled and running
20120727 18:09:10 Master SubnetManager on sm lid 65 sm guid 0x212856d0a2c0a0 : SUN IB QDR GW
switch scae01sw-ib04 10.133.42.177 leaf:3
```

グローバルInfiniBandコマンド

グローバルInfiniBandコマンドとは、ネットワークに参加している任意のシステムから実行できるコマンドのことです。通常、これらのコマンドは、一般的なソフトウェア・スタックAPIによって提供されます。たとえば、OFEDソフトウェア・パッケージは、Linuxベースのコンピュータ・ノードおよびSun DCS QDRスイッチ向けのツール・セットおよびコマンド・セットを提供します。また、Solarisベースのホストにも、使用方法と出力に関しては完全に同一となっている類似のコマンド・セットが提供されています。

ibdiagnet - `ibdiagnet`は、指示されたルート・パケットを使用してファブリックをスキャンし、その接続とデバイスに関して得られる情報をすべて抽出します。

複数のスイッチを使用して、さまざまな種類の情報を問合せできます。通常、サマリーはターミナル自体に出力されますが、詳細なログは/tmpディレクトリに生成されます。

```
[root@scae01cn01 ~]# ibdiagnet -ls 10 -lw 4x
Loading IBDIAGNET from: /usr/lib64/ibdiagnet1.5.4
-W- Topology file is not specified.
    Reports regarding cluster links will use direct routes.
Loading IBDM from: /usr/lib64/ibdml.5.4
-W- A few ports of local device are up.
    Since port-num was not specified (-p option), port 1 of device 1 will be
    used as the local port.
-I- Discovering ... 65 nodes (8 Switches & 57 CA-s) discovered.

-I-----
-I- Bad Guids/LIDs Info
-I-----
-I- No bad Guids were found

-I-----
-I- Links With Logical State = INIT
-I-----
-I- No bad Links (with logical state = INIT) were found

-I-----
-I- General Device Info
-I-----

-I-----
-I- PM Counters Info
-I-----
-I- No illegal PM counters values were found

-I-----
-I- Links With links width != 4x (as set by -lw option)
-I-----
-I- No unmatched Links (with width != 4x) were found

-I-----
-I- Links With links speed != 10 (as set by -ls option)
-I-----
-I- No unmatched Links (with speed != 10) were found

-I-----
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I-----
-I-    PKey:0x7fff Hosts:113 full:113 limited:0

-I-----
-I- IPoIB Subnets Check
-I-----
-I- Subnet: IPv4 PKey:0x7fff QKey:0x00000b1b MTU:2048Byte rate:10Gbps SL:0x00
-W- Suboptimal rate for group. Lowest member rate:40Gbps > group-rate:10Gbps
```

```

-I-----
-I- Bad Links Info
-I- No bad link were found
-I-----

-I- Stages Status Report:
  STAGE                               Errors  Warnings
Bad GUIDs/LIDs Check                 0       0
Link State Active Check               0       0
General Devices Info Report           0       0
Performance Counters Report           0       0
Specific Link Width Check              0       0
Specific Link Speed Check              0       0
Partitions Check                       0       0
IPoIB Subnets Check                  0       1

Please see /tmp/ibdiagnet.log for complete log
-----

-I- Done. Run time was 10 seconds.
[root@scae01cn01 ~]# ls -l /tmp/ibdiagnet*
-rw-r--r-- 1 root root 317258 Apr  6 17:02 ibdiagnet.db
-rw-r--r-- 1 root root  63784 Apr  6 17:02 ibdiagnet.lst
-rw-r--r-- 1 root root   7768 Apr  6 17:02 ibdiagnet.pkey
-rw-r--r-- 1 root root  33752 Apr  6 17:02 ibdiagnet.fdfs
-rw-r--r-- 1 root root   398 Apr  6 17:02 ibdiagnet.sm
-rw-r--r-- 1 root root   6161 Apr  6 17:02 ibdiagnet.mcfdfs
-rw-r--r-- 1 root root 716912 Apr  6 17:02 ibdiagnet.slv1
-rw-r--r-- 1 root root 353078 Apr  6 17:03 ibdiagnet.psl
-rw-r--r-- 1 root root   1931 Apr  6 17:03 ibdiagnet.log

```

ibdagnetのログによる分析

/tmp/ibdiagnet.sm - ネットワーク内で動作しているopensmのすべてのインスタンスが含まれています。

```

[root@scae01cn01 ~]# cat /tmp/ibdiagnet.sm
ibdiagnet fabric SM report

SM - master
  Port=13 lid=0x003d guid=0x0021286cc8aca0a0 dev=48438 priority:14

SM - standby
  Port=6 lid=0x0001 guid=0x002128547f22c0a0 dev=48438 priority:5
  Port=21 lid=0x0040 guid=0x00212856d162c0a0 dev=48438 priority:5
  Port=20 lid=0x0041 guid=0x00212856d0a2c0a0 dev=48438 priority:5
  Port=22 lid=0x003f guid=0x002128548042c0a0 dev=48438 priority:5

/tmp/ibdiagnet.pkey - 使用中のすべてのパーティションおよび関連付けられているホストとそのメンバーシップのリストです。

[root@scae01cn01 ~]# cat /tmp/ibdiagnet.pkey
GROUP PKey:0x7fff Hosts:113
Full scac01cel08/U/P1 lid=0x006e guid=0x0002c903000a7c41 dev=26428
Full scac01cel08/U/P2 lid=0x006f guid=0x0002c903000a7c42 dev=26428
Full scac01cel14/U/P1 lid=0x0017 guid=0x0002c903000a740d dev=26428
Full scac01cel14/U/P2 lid=0x0003 guid=0x0002c903000a740e dev=26428

```

```
Full scae01cn14/U/P1 lid=0x002f guid=0x0002c903000a7b6d dev=26428
```

```

Full scae01cn14/U/P2 lid=0x0037 guid=0x0002c903000a7b6e dev=26428
Full MT25408/P1 lid=0x006d guid=0x0002c903000a7429 dev=26428
Full MT25408/P2 lid=0x0007 guid=0x0002c903000a742a dev=26428
<.. snip ..>
Full scac01db05/U/P1 lid=0x0018 guid=0x0002c903000a7711 dev=26428
Full scae01cn02/U/P1 lid=0x0030 guid=0x0021280001a0a35d dev=26428
Full scae01cn02/U/P2 lid=0x0031 guid=0x0021280001a0a35e dev=26428
-----

```

/tmp/ibdiagnet.fdfs - ファブリック・スイッチのユニキャスト・フォワーディング・テーブルのダンプです。

このファイルには、多数のエントリが含まれている可能性があります。その一部は、"UNREACHABLE"となっていることがあります。これは、以前にスイッチ側で把握していたこれらのエントリのLIDが、すでに利用できなくなっていることを意味しています。

最後の列は、サブネット・マネージャのルーティング・アルゴリズムによって計算された最適パスであれば、キーワードが'yes'になるはずですが。

ibdiagpath - 2つのエンド・ポイント間でパスをトレースして、パスに沿ってトラバースするノートとポートに関する情報を提供します。

```

[root@scae01cn06 ~]# ibdiagpath -l 65,1
Loading IBDIAGPATH from: /usr/lib64/ibdiagpath1.5.4
-W- Topology file is not specified.
    Reports regarding cluster links will use direct routes.
Loading IBDM from: /usr/lib64/ibdml.5.4
-W- A few ports of local device are up.
    Since port-num was not specified (-p option), port 1 of device 1 will be
    used as the local port.

-I-----
-I- Traversing the path from local to source
-I-----
-I- From: lid=0x0072 guid=0x0021280001a0a3fd dev=26428 scae01cn06/U/P1
-I- To:   lid=0x0001 guid=0x002128547f22c0a0 dev=48438 Port=9

-I- From: lid=0x0001 guid=0x002128547f22c0a0 dev=48438 Port=33
-I- To:   lid=0x000f guid=0x002128468d27a0a0 dev=48438 Port=25

-I- From: lid=0x000f guid=0x002128468d27a0a0 dev=48438 Port=24
-I- To:   lid=0x0041 guid=0x00212856d0a2c0a0 dev=48438 Port=35

-I-----
-I- Traversing the path from source to destination
-I-----
-I- From: lid=0x0041 guid=0x00212856d0a2c0a0 dev=48438 Port=20
-I- To:   lid=0x003d guid=0x0021286cc8aca0a0 dev=48438 Port=9

-I- From: lid=0x003d guid=0x0021286cc8aca0a0 dev=48438 Port=32
-I- To:   lid=0x0001 guid=0x002128547f22c0a0 dev=48438 Port=21

-I-----
-I- PM Counters Info
-I-----
-I- No illegal PM counters values were found

```



```
-I-----  
-I- Path Partitions Report  
-I-----  
-I- Source Port=35 lid=0x0041 guid=0x00212856d0a2c0a0 dev=48438 Port 35  
    PKeys:Not-Enforced  
-I- Destination lid=0x0001 guid=0x002128547f22c0a0 dev=48438 PKeys:Not-Enforced  
  
-E- No shared PKeys found on Path! Nodes can not communicate!  
    Aborting route tracing.
```

ネットワーク・パケットのキャプチャ

ネットワークのトラブルシューティングと診断のフェーズでは、分析のためにパケットのキャプチャが必要になることがあります。これらのパケット・キャプチャによる情報は、問題の根本原因を突き止めたり、または他の調査結果を裏付けたりするのに、非常に役立ちます。メッセージのシーケンス、ハンドシェイク・フロー、待機時間、パケット形式、欠落したメッセージなど、さまざまなものを表示できます。

Linuxでパケットをキャプチャする場合は、複数のオプションがあるtcpdumpを使用できます。多くの場合、tcpdumpは、キャプチャしたパケットを保存して事後分析するのに役立ちます。キャプチャされるパケット・フローは非常に高速となる場合があるため、TMPFSファイル・システムで複数の増分ファイルに保存することを推奨します。ファイルを別の場所に保存した後は、TMPFSの格納場所をクリーンアップしてください。Solarisでは、同じ操作をsnoopで実行できます。どちらのツールもキャプチャ形式は同じで、pcapと呼ばれます。

また、InfiniBandでは、ibdumpと呼ばれる類似のツールもあります。このツールは、Mellanoxによって提供されており、同社の以下のWebサイトからダウンロードできます。

http://www.mellanox.com/content/pages.php?pg=products_dyn&product_family=110&menu_section=34

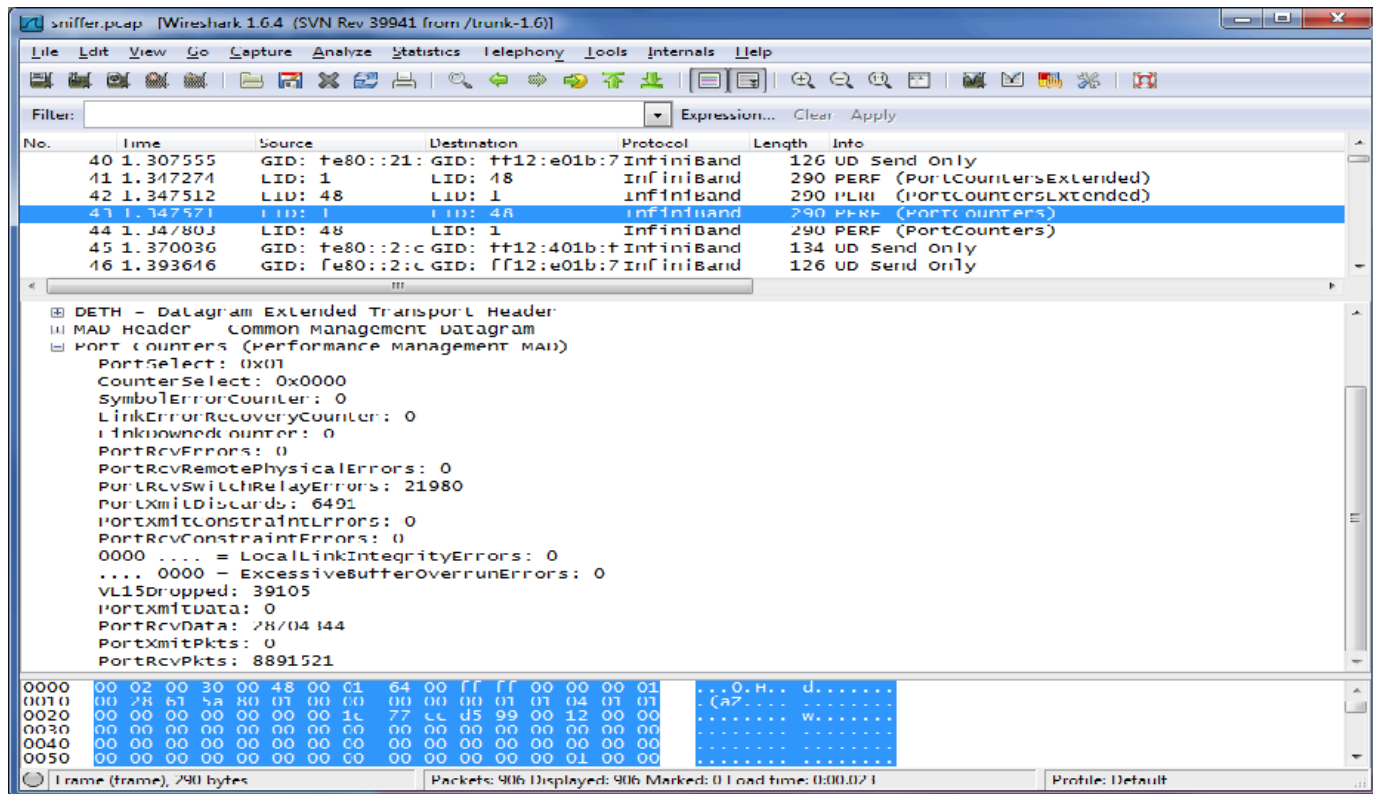
このツールは比較的新しいため、現時点では、tcpdumpやsnoopに相当する機能のすべてを使用できるわけではありません。ただし、出力形式はpcapと同じです。また、ibdumpによるキャプチャは、インタフェース上でInfiniBandメッセージングのデバッグを実行する場合に、非常に役立つ情報を提供します。

すべてのpcapファイルは、Wiresharkと呼ばれるグラフィカル・ツールで再生できます。もっと以前には、このツールはetherealと呼ばれていました。Wiresharkは、次のWebサイトからダウンロードできます。

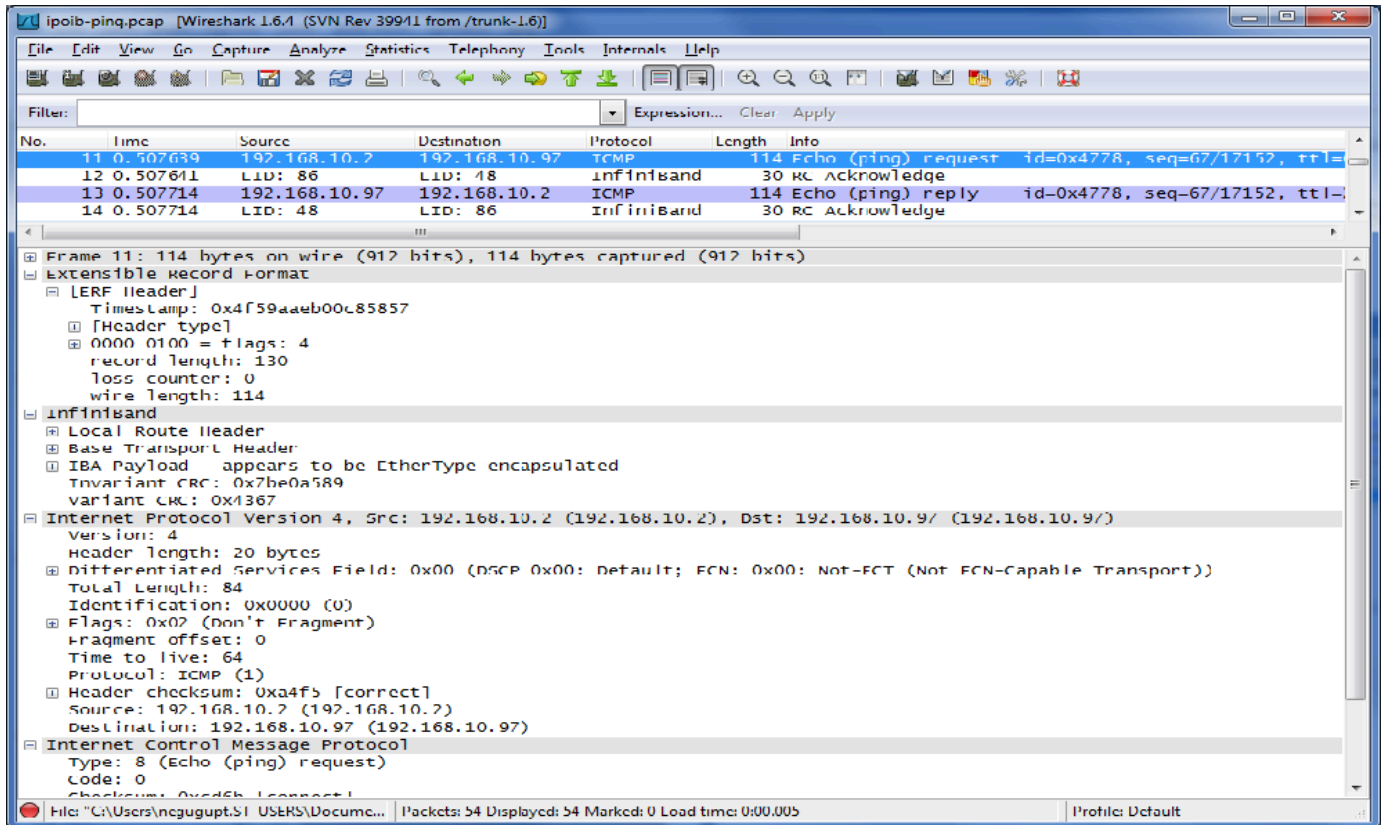
<http://www.wireshark.org/download.html>

このインタフェースは非常に強力で、各種のフィルタを適用することで分析が簡単にできます。また、さまざまなプロトコルを解析して表示する機能があり、ハンドシェイクとそれらのフロー・シーケンスを把握するのに役立ちます。

パフォーマンス・カウンタの間合せ



IPoIBでのIPv4 ICMP (ping)



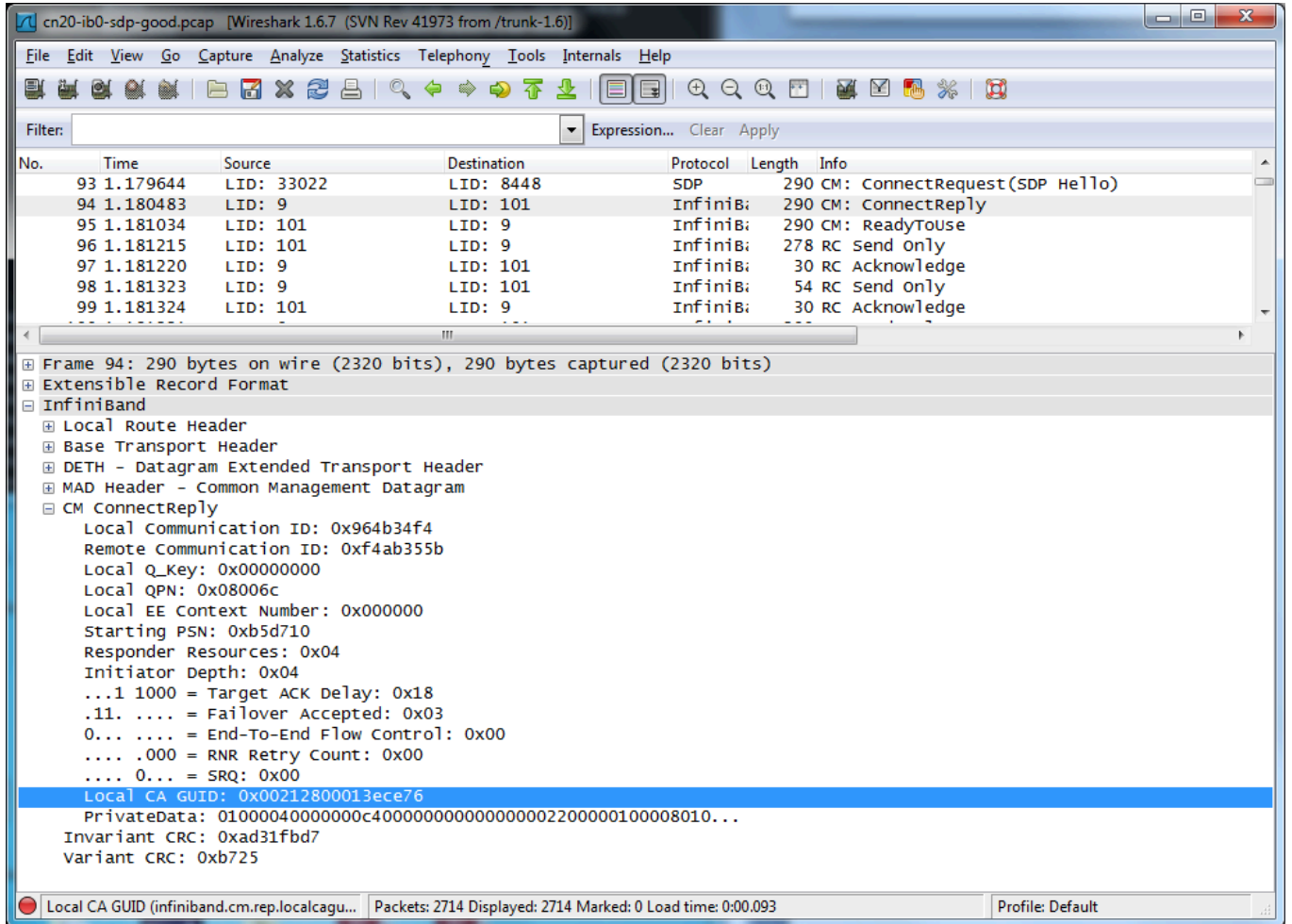
The image shows a Wireshark capture of IPoIB ping traffic. The packet list pane displays four packets:

No.	Time	Source	Destination	Protocol	Length	Info
11	0.507639	192.168.10.2	192.168.10.97	ICMP	114	Echo (ping) request id=0x4778, seq=67/17152, ttl=...
12	0.507611	LID: 86	LID: 48	InfiniBand	30	RC Acknowledge
13	0.507714	192.168.10.97	192.168.10.2	ICMP	114	Echo (ping) reply id=0x4778, seq=67/17152, ttl=...
14	0.507714	LID: 48	LID: 86	InfiniBand	30	RC Acknowledge

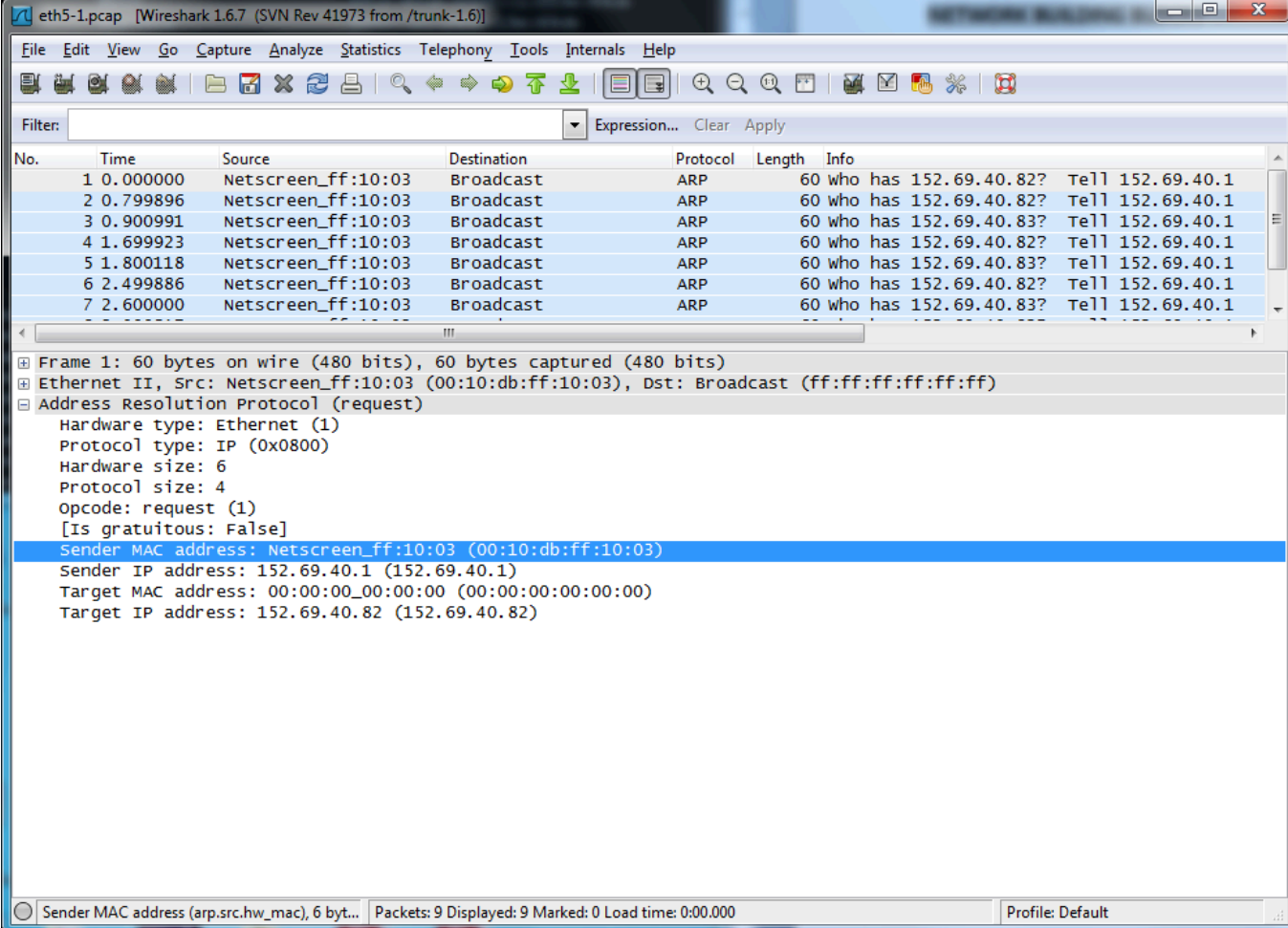
The packet details pane for packet 11 shows the following structure:

- Frame 11: 114 bytes on wire (912 bits), 114 bytes captured (912 bits)
- Extensible record format
 - [ERF Header]
 - Timestamp: 0x4f59aaeb00c85857
 - [Header type]
 - 0000 0100 = flags: 4
 - record length: 130
 - loss counter: 0
 - wire length: 114
- InfiniBand
 - Local Route Header
 - Base Transport Header
 - IBA Payload appears to be EtherType encapsulated
 - Invariant CRC: 0x7be0a589
 - variant CRC: 0x4367
 - Internet Protocol Version 4, Src: 192.168.10.2 (192.168.10.2), Dst: 192.168.10.97 (192.168.10.97)
 - Version: 4
 - Header length: 20 bytes
 - Differentiated Services Field: 0x00 (DSCP 0x00: Default; ECN: 0x00: Not-ECT (Not ECN-Capable Transport))
 - Total Length: 84
 - Identification: 0x0000 (0)
 - Flags: 0x02 (Don't Fragment)
 - Fragment offset: 0
 - Time to live: 64
 - Protocol: ICMP (1)
 - Header checksum: 0xa4f5 [correct]
 - Source: 192.168.10.2 (192.168.10.2)
 - Destination: 192.168.10.97 (192.168.10.97)
 - Internet Control Message Protocol
 - Type: 8 (Echo (ping) request)
 - Code: 0
 - Checksum: 0x2d6b [correct]

SDPハンドシェイク



ARPリクエストの失敗



The image shows a Wireshark capture of an ARP request packet. The packet list pane shows seven frames, all of which are ARP requests from source MAC address Netscreen_ff:10:03 to destination Broadcast. The selected frame (No. 1) is expanded in the packet details pane, showing the following information:

- Frame 1: 60 bytes on wire (480 bits), 60 bytes captured (480 bits)
- Ethernet II, Src: Netscreen_ff:10:03 (00:10:db:ff:10:03), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
- Address Resolution Protocol (request)
 - Hardware type: Ethernet (1)
 - Protocol type: IP (0x0800)
 - Hardware size: 6
 - Protocol size: 4
 - Opcode: request (1)
 - [Is gratuitous: False]
 - Sender MAC address: Netscreen_ff:10:03 (00:10:db:ff:10:03)
 - Sender IP address: 152.69.40.1 (152.69.40.1)
 - Target MAC address: 00:00:00_00:00:00 (00:00:00:00:00:00)
 - Target IP address: 152.69.40.82 (152.69.40.82)

The status bar at the bottom indicates: Sender MAC address (arp.src.hw_mac), 6 byt... Packets: 9 Displayed: 9 Marked: 0 Load time: 0:00.000 Profile: Default

トラブルシューティングのシナリオ

ユースケース分析#1

定期的なデータ収集プロセスにより、`perfquery`の出力でシンボル・エラーまたはリンク・リカバリが示されています。InfiniBandリンクが4X QDRになっていません。

補足説明

`perfquery`のシンボル・エラー・カウンタは、特定のリンクで電氣的レベルに問題があることを示します。このようなエラーは、特定のポートに限定されます。これらのカウンタは、この特定のリンクとそれに関連付けられているハードウェア・コンポーネントの問題だけを示します。

`perfquery`でのリンク・リカバリは、指定された電氣的レベルでリンクが安定しないためにオート・ネゴシエーションが再試行されていることを示しています。

InfiniBandリンクの幅が1Xまたは速度がDDRやSDRと表示されている場合は、ハードウェア・レイヤーに何らかの問題があるために、指定されたパラメータでネゴシエーションができないことを示しています。

トラブルシューティングのガイドライン

ポート・カウンタをリセットして、観察期間中にエラーの症状が再発するかを観察します。リンクの両側にあるポートを一度に1つずつリセットして、ステータスを再び観察します。

物理的に、リンクの両端を一度に1つずつ再配置して、ステータスを再び観察します。可能であればスイッチ側の予備ポートを使用して、エラー状態が解消するかを確認します。

一時的に別のケーブルを使用して、ケーブル自体に障害がなかったかを確認します。

これらのタスクを実行した後、問題が解決した場合は、ハードウェアにリカバリ不可能な障害がないと結論付けられます。つまり、エラー状態は、コネクタの不適切な接続またはリンクのネゴシエーションが原因で発生した可能性があります。ただし、問題が解決しない場合は、実行する各手順からの終了ステータスに基づいて、どのコンポーネントに障害があるかを結論付けられます。

ユースケース分析#2

InfiniBandファブリックにマスター・サブネット・マネージャのインスタンスがありません。LIDの値が0になっています。

補足説明

サブネット・マネージャは、InfiniBandネットワークを機能させる上で最も重要なソフトウェアです。サブネット・マネージャはNM2スイッチ上で動作し、同じInfiniBandサブネット内で複数のインスタンスを実行する場合は、その中で選択された1つのインスタンスがマスターになり、他のインスタンスはスタンバイになります。OFEDツールの`sminfo`またはNM2スイッチのコマンド`getmaster`を使用すると、現在のマスター・サブネット・マネージャを問い合わせることができます。サブネット・マネージャのすべてのインスタンスを問い合わせる場合は、OFEDツールの`ibdiagnet`を使用して`/tmp/ibdiagnet.sm`を生成できます。

トラブルシューティングのガイドライン

opensmインスタンスを実行しているすべてのNM2スイッチで、ログ・ファイルを検査します (/var/log/messagesと/var/log/opensm.log)。

各NM2スイッチが、/etc/hostsファイルによってlocalhostを127.0.0.1に解決できることを確認します。パーティションの構成が正しく、構文に問題があるかどうかを確認します。

/conf/configvalidファイルを検査して、数値'1'を含む行が1つあることを確認します。

ibdiagnetの出力を検査して、ファブリック内に重複したGUIDやエラーになっているリンクがないことを確認します。

各NM2スイッチのシステム状態を検査して、ファイル・システムに空き領域があること、およびメモリの空き容量が十分な状態で実行していることを確認します。サブネット・マネージャのインスタンスを実行しているホストがないことを確認します。

ユースケース分析#3

IPoIBネットワーク内にあるいくつかのノードへのpingが失敗します。

補足説明

InfiniBandファブリック内にある有効なIPoIBサブネットで一連のホストが構成されている場合、それらのホストは互いに通信することができます。

トラブルシューティングのガイドライン

IPoIBサブネットの構成に問題がないこと、つまりすべてのノードのサブネット・マスクが一致していることを確認します。すべてのノードが同じInfiniBandパーティション内にあることを確認します。

トラブルシューティングの対象になっているノードが、パーティションのメンバーシップに基づいて互いに通信できることを確認します。

ネットワーク内にマスター・サブネット・マネージャがあることを確認します。

サブネット・マネージャのログ・ファイル/var/log/opensm.logを検査して、明らかな問題点の有無を確認します。

ノードとスイッチが、ibhostsコマンドとibswitchesコマンドの出力に表示されることを確認します。MTUコードが0x84を超えるInfiniBandマルチキャスト・グループがあるかどうかを確認します。

分析対象となっているホスト間にあるパスが、正しいことを確認します。

ホスト上にあるボンディング・ドライバのステータスを検査して、アクティブ・インタフェースがあることを確認します。

ホストのログ・ファイルを検査して、IPoIBインタフェースが送信タイムアウトを報告しているかどうかを確認します。

ユースケース分析#4

NM2 InfiniBandスイッチに、ログインできません。

補足説明

NM2スイッチには、管理アクセス用のギガビット・イーサネット・ポートがあります。必要であれば、USB変換器を経由してシリアル接続でスイッチにアクセスすることもできます。

トラブルシューティングのガイドライン

NM2スイッチに対してpingテストを実行します。このテストが成功した場合は、IP接続がOKになります。スイッチのブラウザ・インタフェース経由で、スイッチにアクセスできるかどうかを確認します。

同じInfiniBandファブリックで実行したibswitchesの出力に、スイッチが表示されているかどうかを確認します。USBシリアル・ポートを使用してスイッチに接続して、ログインを試行します。

それでもまだスイッチにアクセスできない場合は、スイッチの電源を入れ直す必要があります。

USBシリアル・ポート経由でログインできた場合は、次に、ILOM CLI経由でSSHサービスの再起動を試行します。

IPのpingでスイッチに到達できなかった場合は、イーサネットへのリンクがアクティブになっているかどうかを確認します。これには、ethtoolコマンドを使用します。

スイッチが回復した後は、ファイル・システムのステータスとログ・ファイルを検査して、明らかな問題点がないことを確認します。

ユースケース分析#5

EoIB VNICを動作状態にできません。

補足説明

この状況はNM2GWに当てはまりますが、トラブルシューティング中にNM2GWとNM2-36Pの相互作用についても分析する必要があります。VNICがUP状態にならない理由は数多くあるため、診断を段階的に進める必要があります。

トラブルシューティングのガイドライン

作成したVNICの現在のGUIDとMACアドレスが正しいことを確認します。NM2GWにある10GbEアップリンクのステータスを確認します。

/conf/configvalidファイルを検査して、数値'1'を含む行が1つあることを確認します。

InfiniBandファブリックに、マスター・サブネット・マネージャが存在していることを確認します。

showgwconfigコマンドを実行して、Data SLが1およびControl SLが2になっていることを確認します。必要に応じて、正しいPKEYとVLANが使用されていることを確認します。

Bridge-XのGUIDが、適切なInfiniBandパーティションに追加されていることも確認します。該当するホスト上のOFEDバージョンが正しいことを確認します。

ドライバがホストに適切にインストールされて動作していることを確認します。/etc/infiniband/openib.confにあるVNICパラメータが、'yes'になっている必要があります。ismodコマンドによって、mlx4_vnicドライバがロードされる必要があります。

ホストがSolarisを実行している場合は、最初にdladmコマンドを使用して、デバイス・インスタンスがすでに構成されていることを確認します。

可能であれば、BXNを再起動してVNICが機能するかどうかを確認します。

NM2GWスイッチで、メモリやディスク領域などのシステム・リソースが不足していないかどうかを確認します。

ユースケース分析#6

VNICが外部のイーサネット・ネットワークと通信できません。

補足説明

場合によっては、VNIC間で通信ができて、VNICと外部ネットワークの間で通信ができないことがあります。この状況では、外部ネットワークのコンポーネントと構成を分析する必要があります。

トラブルシューティングのガイドライン

10GbEアップリンクのステータスを検査して、アップリンクがアクティブであることを確認します。

アップリンクが外部LANに正しく接続されていることを確認します。

VNICホストが目的のネットワークに到達するのに、ルーターが必要かどうかを確認します。VNICからルーターのIPに到達できる必要があります。

VNICによるボンディングを使用している場合は、外部ネットワーク・スイッチへの両方のパスが正しく構成されていることを確認します。

アクティブ/スタンバイ・ボンディングでは、テストによって、両方のインタフェースが機能していないのか、または問題が片方のインタフェースだけなのかを確認できます。

VLANを使用している場合は、外部スイッチ側も検査して、スイッチがVLAN向けに正しく構成されていることを確認します。

テストの対象となるホストとスイッチが数ホップ離れている場合は、別のホストでホップ数がより少ない通信のテストを実行して、問題が解消するかを確認します。

ユースケース分析#7

SDP over InfiniBandを使用して接続を確立することができません。

補足説明

InfiniBandドライバでSDPサポートが有効になっている場合、対応しているアプリケーションでは、TCPの代わりに、ソケット・ベースの接続用のSDPを使用できます。

トラブルシューティングのガイドライン

同じホスト・ペアにあるアプリケーションで、TCPが使用できるかどうかを確認します。TCPが使用できる場合は、SDPに固有の問題ということになります。別のクライアントが、このサーバーにSDPで接続できるかどうかを確認します。

また、クライアントが、別のサーバーにSDPで接続できるかどうかを確認します。

これらの結果に基づいて、サーバーとクライアントのどちら側に問題があるかを特定します。ibdumpユーティリティを使用してパケットのキャプチャを行い、オフライン分析を実施します。

異常が発見されなかった場合は、可能であれば、問題があるノードを再起動します。チケットのファイルへの保存が、必要になることがあります。

ユースケース分析#8

EoIBボンディングで、10GbEアップリンクに障害が発生してもフェイルオーバーが実行されません。

補足説明

EoIB VNICは、NM2GWの10GbEアップリンクに依存しています。Linuxボンディングを使用して1+1冗長パスを設定している場合は、アップリンクに障害が発生したときにVNICのフェイルオーバーが実行される必要があります。

トラブルシューティングのガイドライン

eport_state_enforceオプションが、1に設定されているかどうかを確認します。これは必須です。OFEDとNM2GWのバージョンに、互換性があることを確認します。

ネットワークのアーキテクチャによっては、Linuxボンディングでarp_ip_targetオプションの使用が必要になることがあります。

ユースケース分析#9

rds-pingで、最初の応答までの待機時間が非常に長くなっています。

補足説明

rds-pingは基本的な接続テスト・ツールで、特に、2つのExadataデータベース・ノードまたはセル・ノードの間にあるRDS接続を検証することを目的としています。したがって、他の種類のノードまたはストレージ・アプライアンスでは機能しません。ファブリック全体の何らかの問題によって、rds-pingが通信チャネルを確立するのに必要な時間が長くなることもあり、そのために最初の応答までの待機時間がそれ以降の応答よりも長くなる可能性があります。最初の応答までの待機時間が通常よりも長い場合は、ファブリックに何らかの問題があります。

トラブルシューティングのガイドライン

ibdiagnet -ls 10 -lw 4xを使用して、ファブリックが正常であることを確認します。

InfiniBandホストおよびスイッチの数を確認します。この数は、ファブリック設計図にある数と一致している必要があります。pingコマンドを使用して、すべてのIPoIBターゲットが正常であることを確認します。

ファブリック内に、応答しないInfiniBandノードがあるかどうかを確認します。大抵は、このようなノードが、InfiniBand問合せの遅延を引き起こします。

ユースケース分析#10

NFSマウント・ポイントが、ハングまたは停止します。

補足説明

Exalogicマシンには、NFS上のファイル・システムとして機能するSun ZFS Storage Appliance (ZFSSA) が組み込まれています。多くの場合、Exadataマシンは、バックアップおよび他の目的で外部のZFSSAも使用します。通常、これらのNFSマウントは、IPoIBチャネルを通じて実行されます。ファイル・システムの動作中にマウント・ポイントがハングまたは停止する場合は、IPoIBまたはNFSサーバー/クライアント自体のいずれかに何らかの問題があります。

トラブルシューティングのガイドライン

`ibdiagnet -ls 10 -lw 4x`を使用して、ファブリックが正常であることを確認します。

InfiniBandホストおよびスイッチの数を確認します。この数は、ファブリック設計図にある数と一致している必要があります。

`ping`コマンドを使用して、すべてのIPoIBストレージ・ターゲットが正常であることを確認します。通常、コンピュータ・ノードとストレージ間でのpingが成功した場合は、IPoIBが正常であることを表しているため、問題は別の場所にあります。

マウントが、特定のノードまたはすべてのNFSクライアントでハングするかどうかを確認します。これは、NFSクライアントとNFSサーバーのどちらに問題があるかを特定するのに役立ちます。

NFS v4を使用している場合は、ドメイン・マッピングに使用されるNISやLDAPのような、関連する重要なサービスを検査します。

また、ZFSSAソフトウェアのバージョンをチェックして、必要であれば、最新リリースのバージョンにアップグレードして既知の問題を回避することを推奨します。

ユースケース分析#11

パーティションに基づくIPoIBパスを通じて、NFSサービスに到達することができません。

補足説明

InfiniBandパーティションは、同じ物理媒体上で仮想化されたネットワーク・パスを提供します。接続を成功させるには、一連の構成を行う必要があります。パーティション・ベースのネットワークではなく、デフォルトのネットワークが機能している場合は、この状態の分析が必要になることがあります。

トラブルシューティングのガイドライン

`smpartition list active`コマンドを使用して、InfiniBandスイッチ上にパーティションが作成されているかどうかを確認します。また、`/tmp`にある`ibdiagnet`のログ・ファイルを使用して、同じ情報を問い合わせることもできます。

ストレージのGUIDが、Fullメンバーとしてパーティションに追加されたかどうかを確認します。ストレージは、その設計上、Limitedメンバーになることができません。

パーティション構成が、InfiniBandスイッチにコミットされたかどうかを確認します。コミットは、マスター・サブネット・マネージャ上で実行する必要があります。必要であれば、確認のためにもう一度コミットできます。

パーティション化されたインタフェースをIPMPで構成している場合は、アクティブとスタンバイのスイッチングが有効になっているかどうかを試すことができます。スイッチングが有効になっていれば、両方のインタフェースではなく、片方のインタフェースのみに問題があります。

同じパーティション化されたインタフェースが、同じパーティションの別のノード・メンバーと通信できることを確認します。また、この確認は、NFSクライアントでも実施する必要があります。

また、IPoIBサブネットも検査して、レイヤー3の構成にエラーがないことを確認します。

ユースケース分析#12

コンピュータ・ノード上のInfiniBandインタフェースがアクティブになりません。

補足説明

InfiniBandインタフェースには、物理状態と論理状態があります。物理状態は、ファームウェアを使用して電氣的なレベルで制御できます。論理状態は、デバイス・ドライバおよびデバイス・ドライバとサブネット・マネージャの相互作用によって制御されます。

トラブルシューティングのガイドライン

ibstatコマンドを使用して、ローカルのInfiniBandデバイスが正常であることを確認します。物理状態が停止またはポーリングになっている場合は、ハードウェア・レベルに何らかの問題があることを示しています。

InfiniBandスイッチ上の該当するポートが、無効または非アクティブになっているかどうかを確認します。これには、listlinkupコマンドを使用します。

ケーブルを検査して、ケーブルの両端が正しく接続されていることを確認します。必要であれば、予備ケーブルでテストするか、または別のスイッチ・ポートに差し替えます。

物理状態が停止からポーリングに変化すると、InfiniBandインタフェースはアクティブになります。

論理状態は、'初期化中'になることがあります。このような状態が長時間続く場合は、サブネット・マネージャとの通信に問題があります。

サブネット・マネージャを検査して、サブネット・マネージャがアクティブであること、およびエンド・ポイントにLIDが割り当てられていることを確認します。

ユースケース分析#13

netstatコマンドの出力で、EoIBインタフェースに破棄パケットがあることが示されています。

補足説明

EoIBインタフェースには、標準のネットワーク・インタフェースと同じように、統計カウンタがあります。ネットワークの構成に誤りがあると、破棄される受信パケットが増加する原因となることがあります。このようなシナリオには、複数の理由が存在することもあります。

トラブルシューティングのガイドライン

netstatコマンドの出力にあるゼロではないカウンタを確認して監視し、何らかの相関関係を見つけ出します。

また、VLANに関連するエラー・カウンタが増加しているかも確認します。エラー・カウンタが増加している場合は、VLANの構成に誤りがあります。tcpdumpを使用してパケットのキャプチャを行い、Wiresharkで分析します。

複数のポートが同じイーサネット・スイッチに接続されているマルチホーム環境では、ルーティング・エラーが原因で、受信パケットが破棄されることがあります。



InfiniBand ネットワークにおけるトラブル
シューティングのガイドラインと方法

2012年8月

著者: Neeraj Gupta

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

海外からのお問い合わせ窓口:
電話: +1.650.506.7000
ファクシミリ: +1.650.506.7200
www.oracle.com



Oracle is committed to developing practices and products that help protect the environment

Copyright © 2011, Oracle and/or its affiliates. All rights reserved. 本文書は情報提供のみを目的として提供されており、ここに記載される内容は予告なく変更されることがあります。本文書は一切間違いがないことを保証するものではなく、さらに、口述による明示または法律による黙示を問わず、特定の目的に対する商品性もしくは適合性についての黙示的な保証を含み、いかなる他の保証や条件も提供するものではありません。オラクル社は本文書に関するいかなる法的責任も明確に否認し、本文書によって直接的または間接的に確立される契約義務はないものとします。本文書はオラクル社の書面による許可を前もって得ることなく、いかなる目的のためにも、電子または印刷を含むいかなる形式や手段によっても再作成または送信することはできません。

OracleおよびJavaはOracleおよびその子会社、関連会社の登録商標です。その他の名称はそれぞれの会社の商標です。

AMD、Opteron、AMDロゴおよびAMD Opteronロゴは、Advanced Micro Devicesの商標または登録商標です。IntelおよびIntel XeonはIntel Corporationの商標または登録商標です。すべてのSPARC商標はライセンスに基づいて使用されるSPARC International, Inc.の商標または登録商標です。UNIXはX/Open Company, Ltd.によってライセンス提供された登録商標です。

1010

Hardware and Software, Engineered to Work Together