APPENDIX A

InfiniBand の概念

この章では、InfiniBand の概念について説明します。内容は次のとおりです。

InfiniBand の概要

InfiniBand は CPU の利用率を向上させ、遅延を削減し、データセンターの管理を容易にする、高速で高密度のシリアル相互接続です。InfiniBand という用語は、ハードウェア、通信、および管理のインフラストラクチャ全体を意味します。このテクノロジーを利用すると、次の通信速度を向上させることができます。

- CPU
- サーバ内のデバイス
- ネットワーク全体にわたって配置されたサブシステム

InfiniBand では、高速ハードウェア、専用プロトコル、および Remote Data Memory Access（RDMA）技術を組み合わせて、CPU 利用率の向上と遅延の削減が図られています。InfiniBand アーキテクチャの動作は、Subnet Manager で管理されます。

ここでは、次の内容を説明します。

InfiniBand のコンポーネント

次の 1 つまたは複数のコンポーネントを使用することにより、サーバ ネットワークを最大限に活用できます。

- InfiniBand スイッチ — InfiniBand ネットワーク上で、InfiniBand 対応デバイス同士のトラフィックの受け渡しを行います。
- Host Channel Adapter（HCA; ホスト チャネル アダプタ）（ホストに搭載）— Network Interface Card（NIC; ネットワーク インターフェイス カード）の InfiniBand 版として機能し、ホストを InfiniBand ネットワークに接続します。
- イーサネット ゲートウェイ — イーサネットと InfiniBand ネットワークを接続します。

- ファイバ チャネル ゲートウェイ — ファイバ チャネルと InfiniBand ネットワークを接続します。

プロトコル

InfiniBand では新しいプロトコル セットが必要になります。サーバ スイッチには、必要なすべてのプロトコル ドライバが含まれています。プロトコルは、次のとおりです。

- IPoIB（p.A-2）
- SDP（p.A-2）
- SRP（p.A-2）
- uDAPL（p.A-2）

IPoIB

IP over InfiniBand（IPoIB）リンク ドライバは、InfiniBand ファブリック上で標準の IP カプセル化を行います。IPoIB では、IP over InfiniBand テクノロジーを透過的に使用できます。このテクノロジーは、イーサネットに IP を通す方法と同様の技術です。

IPoIB ドライバを使用して、アドレスの解決とマルチキャスト メンバーシップの管理ができます。

SDP

Socket Direct Protocol（SDP）は、InfiniBand ネットワーク上で使用される透過的なプロトコルです。これにより、ソケットベースのアプリケーションは InfiniBand ネットワーク上で RDMA のパフォーマンスを利用できます。SDP を使用すると、プロセス コンテキスト内で実行されるソフトウェアのサイズが小さくなります。SDP がサポートするゼロコピー機能を使用すると、データベースは作業待ち時間が減少し、アプリケーション サーバは応答待ち時間が減少し、CPU は他の処理に割ける時間が増えるので、データベース、アプリケーション サーバ、および CPU の動作効率を向上させることができます。

SRP

SCSI RDMA Protocol（SRP）は、RDMA 対応ネットワークを介して SCSI コマンドを実行する、InfiniBand の上位レイヤのストレージ プロトコルです。InfiniBand ホストは、このプロトコルを使用してファイバ チャネル ストレージ デバイスと通信します。このプロトコルを使用することで、InfiniBand ホストはストレージが直接接続されている場合と同様に、SCSI コマンドをネイティブ状態で送信できます。

SRP プロトコルには、コンシューマ ペア間の通信手段である RDMA 通信サービスが採用されています。具体的には、制御情報の転送にはメッセージを、データの転送には RDMA の機能を使用します。

SRP プロトコルは、InfiniBand システムにファイバ チャネル ゲートウェイが搭載されている場合にのみ使用されます。

uDAPL

user Direct Access Programming Library（uDAPL）は、InfiniBand とファブリックをネイティブでサポートする標準のユーザ モード API です。uDAPL は、名前からアドレスへの変換、接続の確立、およびデータの確実な転送を行います。接続の管理を行い、データを遅延の少ない方法で転送し、完了させることが uDAL の主な役割です。

アーキテクチャの構成要素

InfiniBand アーキテクチャは、次の基本要素で構成されています。

RDMA

InfiniBand は RDMA テクノロジーを採用しています。RDMA を使用すると、1 台のコンピュータが別のコンピュータのメモリに情報を直接書き込むことができます。RDMA により、ユーザ空間のアプリケーションがハードウェアに直接アクセスできるので、ゼロコピーでデータを移動できます。

ハードウェアとソフトウェアの連係により、ユーザ空間のアプリケーションは、カーネルの介入や不要なデータ コピーを行わずに、遠隔システムのメモリの読み取りと書き込みができます。InfiniBand の高速ネットワーク ハードウェアにかかるメッセージング負荷の大半はアプリケーションによるものなので、この機能により、I/O 動作ごとの CPU 利用率が減少し、システム リソースをより効率的に使用できます。

キュー ペア

Queue Pair（QP）は、InfiniBand アーキテクチャの主要要素の 1 つです。InfiniBand の通信は、ポート間ではなく、QP 間で行われます。

QP は、送信ワーク キューと受信ワーク キューで構成された、アドレス指定が可能なエンティティです。チャネル アダプタ ハードウェアは、送信キューに対するアクセスを多重化し、受信キュー上のメッセージを逆多重化することで通信の調停を行います。

（注）　HCA の機能の定義には、Verb が使用されます。Verb を直接使用するユーザは、「Verb コンシューマ」と呼ばれます。

コンシューマは、ワーク キューに一連の命令をキューイングして、チャネル アダプタに命令を実行させることができます。ワーク キューには、送信ワーク キュー（アウトバウンド）と受信ワーク キュー（インバウンド）の 2 種類があります。QP は、この 2 種類のワーク キューを組み合わせて作成します。

ローカルの QP とリモートの QP がリンクされると接続が確立します。アプリケーションは QP を共有しません。QP を設定するとアプリケーション レベルで管理が可能なので、システム コールのオーバーヘッドが発生しません。

送信ワーク キューと受信ワーク キューには次の特徴があります。

- 必ずペアで作成されます。
- 常にペアのまま維持されます。
- ペアは QP と呼ばれます。
- キューはペア番号で識別され、チャネル アダプタ内にあります。

QP には次の特徴があります。

- メモリ領域をバッファとして使用します（QP の数はメモリのサイズによって制限されます）。
- キー（Q_Key）を使用してパケットの正当性を確認します（各着信パケットのキーが一致する必要があります）。

- （ある場合）パーティション キーで QP がアクセスできるファブリックの範囲を指定します。

QP は、QoS（Quality Of Service）、システム保護、エラーの検出と応答、および使用可能なサービスを定義するメカニズムです。

サービスのタイプごとに個別に QP を設定します。次に示すサービス タイプは、それぞれサービスのレベルとエラーの回復特性が異なります。

- 信頼性のある接続
- 信頼性のない接続
- 信頼性のあるデータグラム
- 信頼性のないデータグラム

ファブリック接続が検出されると、QP と保護ドメインが確立されます。QP ごとにサービスのタイプと QoS が定義され、ファブリックは QoS を最大限に生かした信頼性のあるセキュアな動作をします。システムのハードウェア リソースとソフトウェア リソースには影響しません。

Subnet Manager の概要

Subnet Manager は、ファブリックの動作の設定とメンテナンスを行います。Subnet Manager は複数存在できますが、マスターになるのは 1 つだけです。Subnet Manager には、InfiniBand ファブリックの設定と起動に必要なすべての情報が一元的に格納されています。

マスター Subnet Manager は次のことを行います。

- ファブリック トポロジの検出
- エンド ノードの検出
- 次のパラメータなどによるスイッチとエンド ノードの設定
  - Local Identifier（LID）
  - Global Unique Identifier（GUID）
  - パーティション キー（P_Key）
- スイッチの転送テーブルの設定
- Subnet Management Agents（SMA）からのトラップの受信
- サブネットのスイープ（トポロジの変化の検出、ノードの追加と削除の変更管理）

ここでは、次の内容を説明します。

- Subnet Manager Agent（p.A-4）
- Subnet Manager のホットスタンバイ（p.A-5）

Subnet Manager Agent

Subnet Manager Agent（SMA）は、Subnet Manager の一部です。SMA は各ノードに存在し、Subnet Manager からのパケットを処理します。

Subnet Manager がマスターに選出されると、Subnet Agent を含む Subnet Manager のすべてのコンポーネントが暗黙的にマスターに選出されます。Subnet Manager がマスターでなくなると、そのすべてのコンポーネントはクライアントのメッセージに応答することを中止します。

Subnet Manager のホットスタンバイ

マスターとスレーブの Subnet Manager の同期を取ることで、フェールオーバーが発生した場合にマスターの情報をスレーブに引き継がせることができます。ホットスタンバイの Subnet Manager を設定する場合は、「Subnet Manager のデータベース同期のイネーブル化」（p.8-14）を参照してください。

異なるシャーシ上で稼働している Subnet Manager 間のデータベースの同期には、ホットスタンバイ / データベース同期機能が使用されます。

データベースは、マスター Subnet Manager の不揮発性メモリに維持されます。データベースの同期は、次の 2 つのステージで行われます。

- コールド同期 — マスター Subnet Manager は、スタンバイ Subnet Manager と同期セッションを開始する準備ができると、このステージを開始します。このステージでは、マスター Subnet Manager が、未同期のテーブルをスタンバイ Subnet Manager にコピーします。
- トランザクション同期 — コールド同期が正常に終了すると、このステージに入ります。このステージでは、マスターがデータベースのすべてのアップデート トランザクション要求を処理し、スタンバイの Subnet Manager に複製します。

スタンバイ Subnet Manager は、次のいずれかの事象が発生するとマスターになることができます。

- 現在マスターになっている Subnet Manager が稼働するノードのクラッシュ
- （リンク障害などによる）サブネットの分割
- （メンテナンス目的などによる）マスターのグレースフル シャットダウン

障害が発生すると、次のことが行われます。

- スタンバイ Subnet Manager が新たにマスターになります。
- 新規マスターは、サブネット検出フェーズで取得した情報を元にデータベースを再構築します。
- 既存の LID 割り当てを可能な範囲で保持します。
- 全ポートのリセット、マルチキャスト グループへの再参加、サービスの再通知、イベント転送の再要求、および接続の再確立を行います。
- SlaveToMaster イベント トラップを生成します。外部の管理アプリケーションは、このトラップを受信して必要な処理を開始します。

Subnet Manager のルーティング

InfiniBand ルーティングには、2 つの異なる概念があります。

- スイッチ内のルーティング（スイッチ チップ間のホップ）
- 全スイッチのルーティング（ノード間のホップ）

（注）　この処理は、スイッチ要素間ルーティングとも呼ばれます。

スイッチ内部ルーティングでは、トラフィックを最高のパフォーマンスで通過させ、スイッチ内で輻輳が生じる恐れが最小になるように設定できます。

次に、ルーティングの処理を示します。

ステップ 1　Subnet Manager は、ネットワーク内のすべての InfiniBand スイッチ チップを検出します。

ステップ 2　Subnet Manager は、シャーシごとに、すべての内部スイッチ チップを 1 つのスイッチ要素にグループ化します。

ステップ 3　Subnet Manager は、すべての InfiniBand スイッチをスイッチ要素にグループ化するまでこの処理を続けます。

ステップ 4　すべてのスイッチ チップをグループ化すると、「最小コンテンション、最短パス、ロード バランシング アルゴリズム」（p.A-6）に記載されたルーティング アルゴリズムに従い、スイッチ要素間をルーティングします。

ステップ 5　続いて、各 InfiniBand スイッチの内部ネットワークを、スイッチ要素ごとの最善のアルゴリズムに基づいてルーティングします。

ここでは、次の内容を説明します。

複数パス

Subnet Manager を使用すると、サブネット単位に Local Identifier Mask Control（LMC）を指定できます。LMC のデフォルト値は 0 です。デフォルトでは、Local Identifier（LID）が 1 つだけ各ホスト ポートに割り当てられます。

LMC の値が割り当てられると、Subnet Manager は同じホスト ポートに関連付けられた LID ごとに異なるパスをルーティングします。これらのパスによる結果は、適用されるルーティング アルゴリズムに基づきます。

Subnet Manager のルーティング用語の概要

ルーティングに Subnet Manager が使用する多様なアルゴリズムについて説明する前に、次の用語について理解することが重要です。

- 「トレランス」は、特定のパスが、すでに選択されているパスよりも距離的に優れているかどうか決定する際に使用されます。次のように、最短パスの計算にトレランスを使用できます。
  - トレランスが 0 に設定されると、エンドポートまでのパスが異なるペアは、パスのホップ数が同じであれば距離が等しいとされます。
  - トレランスが 1 に設定されると、エンドポートまでのパスが異なるペアは、パスのホップ数の差が 1 以下であれば距離が等しいとされます。
- 「コンテンション」は、同じホスト ポートに関連付けられている別の LID のルーティングにパスがすでに使用されていた場合に、そのパス上のすべてのスイッチ ポートに対して宣言されます。

最小コンテンション、最短パス、ロード バランシング アルゴリズム

デフォルトでは、スイッチ要素間のルーティング、および各スイッチ要素内の InfiniBand 内部スイッチ チップ間ルーティングには、最小コンテンション、最短パス、およびロード バランシング アルゴリズムが使用されます。

次に、アルゴリズムでの計算手順を示します。

ステップ 1　各ホスト ポートの最短パスを計算します。

ステップ 2　（最短パス + トレランス）距離内にある使用可能なすべてのパスについて、コンテンションを計算します。

a. 最小のコンテンションのパスを選択します。

b. 2 つのパスのコンテンションが同じ場合は、短い距離の方のパスを選択します。

c. 2 つのパスのコンテンションが同じで距離も同じ場合は、ポートの使用状況カウントの値を参照して、両パス間のロード バランスが図られます。使用状況カウントは、その特定のポートを使用するように設定された LID の数を示しています。

確定的ソースベース ルーティング アルゴリズム

確定的ソースベース ルーティング アルゴリズムは、要件をより厳格に定義する必要のある一部のハイパフォーマンス コンピューティング環境で使用されます。管理者は、ネットワークを通過するために任意のポートおよび LID が選択するルートを厳密に定義できます。

現在、このルーティング アルゴリズムは、Cisco SFS 7008（96 ポート スイッチ）の内部ルーティングに限定してサポートされています。詳細については、Cisco SFS 7008 のハードウェア ガイドを参照するか、Cisco TAC にお問い合わせください。

ルーティングの最適化設定

ルーティングの最適化には、次の手順を推奨します。

- スイッチ要素間の等しいパスを作成します。
- 最初に検出されるパスを決定します。

可能であれば、InfiniBand スイッチの要素のあらゆる組み合わせ間のパスが同じ距離（同じホップ数）になるようにスイッチ要素を接続することを推奨します。このようにすれば、デフォルトのトレランス（0）を使用する最適パスが得られます。パスの長さが異なる場合は、トレランスの値を決定する必要があります。

Subnet Manager のルーティング アルゴリズムでは、最初に検出した最善のパスを選択します。同じ特性を持つ複数のパスを使用できる場合は、それらのパスのうち、最初に検出したパスを選択します。スイッチ要素間のケーブル配線を設定して、強制的に特定のパスを優先させることができます。ネットワーク要件に応じて、優先パスを特定のスイッチ要素に集中させたり、複数のスイッチ要素に分散させて耐障害性を向上させることができます。

パーティションの概要

パーティションとは、相互に通信することが許可された複数の InfiniBand ノードの集合です。パーティションにより、次のことが実現します。

- セキュリティの向上
- 大規模クラスタを隔離された小規模のサブクラスタに分割
- InfiniBand ノードを特定の VLAN にマッピング

（注）　db-sync がイネーブルの場合は、マスター Subnet Manager が稼働しているシャーシでのみパーティションの設定を変更できます。詳細については、「データベース同期の設定」（p.8-14）を参照してください。

次の内容について説明します。

パーティションの仕組み

パーティションとは、相互に通信することが許可された複数の InfiniBand ノードの集合です。システム管理者は、ノードを複数のパーティションに所属させることで、必要に応じてパーティション同士をオーバーラップさせて定義できます。通常のデータ パケットには、16 ビットの P_Key（パーティション キー）が含まれており、このキーによってパーティションが一意に定義されます。Subnet Manager は、各ノードのチャネル アダプタをノードの P_Key のセットで設定します。パケットがノードに到着すると、チャネル アダプタは、Subnet Manager の設定を基準にしてパケットの P_Key が有効なことを確認します。無効な P_Key を持ったパケットは廃棄されます。P_Key を確認することで、サーバがパーティションの外側の別なサーバと通信することを防止しています。

InfiniBand パーティションは、イーサネット VLAN やファイバ チャネル ゾーンなど、従来の I/O ネットワーキング テクノロジーにおけるハードウェア実行型のセキュリティ機能に相当します。

パーティション メンバー

メンバーのいないパーティションは、システムには意味がありません。ポートはパーティションに追加されると、そのパーティションのメンバーになります。ポートを複数のパーティションに所属させることで、必要に応じてパーティション同士をオーバーラップさせて定義できます。

ポート メンバーをパーティションに追加するときには、そのポートをフル メンバーシップにするのか、限定メンバーシップにするのかを決定する必要があります。

メンバーシップ タイプ

パーティションは複数のメンバーから構成され、1 つのパーティションにはタイプの異なるメンバーが存在できます。パーティションのメンバーシップを使用すると、パーティション外部との通信だけでなく同じグループの構成メンバー同士の通信を定義できるので、さらに細かい制御が可能になります。

パーティションのメンバーシップには、フル メンバーシップと限定メンバーシップの 2 つのタイプがあります。フル メンバーシップのパーティション メンバーは、フル メンバーと限定メンバーを含めた他のすべてのパーティション メンバーと通信できます。限定メンバーシップのパーティション メンバーは、他の限定メンバーシップのパーティション メンバーとは通信できません。ただし、限定パーティション メンバーは、フル メンバーと通信できます。

デフォルト パーティションについて

Subnet Manager は、デフォルト パーティションを自動的に設定します。これは常に p_key が ff:ff です。

接続されているすべてのポートは、デフォルト パーティションにより制御されます。また、デフォルトでは、すべてがデフォルト パーティションのフル メンバーです。p_key のデフォルトは、すべてのパーティションの設定を管理する制御メカニズムであることから、変更したり削除したりできません。

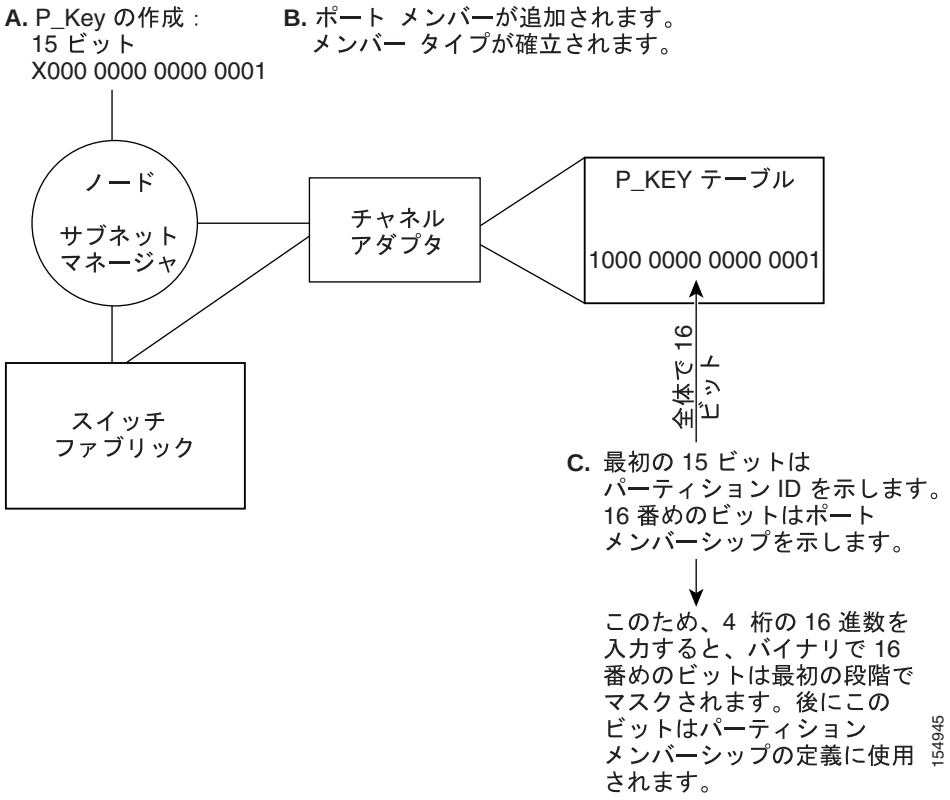
P_Key 値の選択

許容される P_Key 値については、表 A-2（p.A-11）を参照してください。

作成時点の p_key 値（図 A-1 を参照）は、厳密には 15 ビットの数値です。p_key が作成されてからポートのメンバーシップ タイプが確立されると、全体で 16 ビットの値になります。最上位ビット（MSB）はメンバーシップのタイプを示します（0 = 限定メンバー、1 = フル メンバー）。

p_key 値を割り当てる際は、16 進数の 4 桁の数値を選択する必要があります。ただし、16 番めのビットは用途が決まっているため、最上位桁に使用できる数は限られています。16 ビット数の最上位ビットだけが異なる P_Key を 2 つ作成しないようにしてください。システムはそれらを同じ P_Key として認識します。たとえば、0 #:# # は 8#:# # と同じ P_Key になります。

図 A-1　　パーティション キー



次のセクションで、P_Key 値の選択について説明します。

- 16 進数から 2 進数への変換（p.A-10）
- 有効な P_Key 値の例（p.A-11）

16 進数から 2 進数への変換

P_Key の作成に役立つように、表 A-1 を示します。パーティション p_key の作成時は、2 進数の 16 ビットに相当する 16 進数を入力してください。たとえば、80:00（16 進数）は、1000000000000000（2 進数）に相当します。デフォルト パーティション（変更できません）は、7f:ff です。

表 A-1　　2 進数変換

16 進数	2 進数
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000

表 A-1　2 進数変換（続き）

16 進数	2 進数
9	1001
A	1010
B	1011
C	1100
D	1101
E	1110
F	1111

有効な P_Key 値の例

P_Key の値は、任意の値を選択するか、表 A-2 から選択します。

表 A-2　有効な P_Key 値

00:01	00:11
00:02	00:12
00:03	00:13
00:04	00:14
00:05	00:15
00:06	00:16
00:07	00:17
00:08	00:18
00:09	00:19
00:10	00:20

P_Key の保存方法の概要

パーティション情報は、マスター Subnet Manager が保存します。db-sync がイネーブルになっていると、マスター Subnet Manager は、スタンバイ Subnet Manager との間で P_key 情報の同期を取ります（スタンバイ Subnet Manager は現行では 1 つのみ許可されます）。スタンバイ Subnet Manager は、同期が取れるとマスターからの情報を保持します。

1 台の InfiniBand スイッチだけを設定した場合、そのスイッチが自動的にマスターになり、パーティション設定を自スイッチ上に恒久的に保存します。詳細については、「Subnet Manager のデータベース同期のイネーブル化」（p.8-14）を参照してください。