

## サーバ クラスタリングの概要

### クラスタとは

クラスタとは、複数のコンピュータを接続して、1 台のコンピュータでは処理できない問題を、適切かつ効率的に解決できるように設定したものです。クラスタを使用すると、ハイ アベイラビリティなどのビジネス ニーズのほか、気象パターンの追跡や台風予測などの科学的なニーズに応えることができます。クラスタを構成する個々のコンピュータは「ノード」と呼ばれます。あらゆるクラスタは、ハイ アベイラビリティ、スケーラビリティ、管理性、負荷分散の 4 つのタイプに分類できます。最新のクラスタ設計のなかには、2 つのタイプ、あるいは 3 つのタイプを包含したもののさえあります。

### クラスタの種類

#### ハイ アベイラビリティ

ハイ アベイラビリティ クラスタは、通常は 2～3 台のサーバで構成されます。クラスタを構成するサーバは、互いの状態を監視し、何らかの障害が発生すると適切な措置を実行します。このタイプのクラスタは、1 つのラック内に 2 台のサーバを設置（1 方をプライマリ、他方をセカンダリとする）するだけの構成から、世界中に展開するマルチノード コンチネンタル クラスタ構成まで、システム規模に柔軟性があります。プライマリ ノードに障害が発生すると、セカンダリ ノードが処理を引き継ぎ、ユーザにはほとんど影響を与えません。コンチネンタル クラスタでは、深刻な災害が起きた場合でも、アプリケーションのハイ アベイラビリティを維持できます。サーバ障害の原因として最も一般的なものには、ディスクや I/O システムの障害、電源障害、ソフトウェア障害、自然災害、人為的ミスなどがあります。

#### スケーラビリティ

ビジネスが成長するに従い、ミッションクリティカルなアプリケーションに必要とされる処理能力や帯域幅もますます大きくなります。この問題をコスト効率良く解決する方法は、既存の IT インフラストラクチャにサーバを追加し、処理の負荷を複数のサーバ間で分散することです。どのような種類の処理であっても、複数に分割できる処理であれば、このタイプのクラスタで効果的に対処できる可能性があります。クラスタ化によってスケーラビリティを実現している一般的なアプリケーションには、E メールサーバや Web サーバのほかにも、地震解析や画像処理のような科学技術計算があります。小規模なシステムをクラスタ化する代わりに、大規模な対称型マルチプロセッシング サーバを採用することも、このようなビジネス ニーズに対する一般的なソリューションです。

スケーラビリティ クラスタの目的は、機能性と容量の 2 つに分類されます。大容量を目的とするクラスタは、独立した複数のコンピュータが集中管理され、個々のジョブが 1 つまたは少数のノードに割り当てられて処理されるシステムと考えることができます。このタイプのクラスタを設計する場合は、効率的で信頼性の高いノード間相互接続と、包括的な管理基盤を実現することが最大の課題となります。このタイプのクラスタは、ノード間の相互運用をあまり必要としないため、「疎結合」クラスタと呼ばれることもあります。

高機能性を実現するように設計されたクラスタでは、複数のノードが協調して 1 つのタスクを処理する必要があります。高機能計算を必要とするアプリケーションには、ゲノム配列解析、計算流体力学、経済予測などがあります。このタイプのクラスタでは、あるプロセスの処理結果が別のプロセスへの入力となります。高機能クラスタは、一般に大容量クラスタよりも規模が大きく、ノード数が 1000 を超える場合もあります。ただし、高機能クラスタを大規模に展開するためには、いくつかの課題があります。たとえば、並列処理の効率性向上、システム規模を拡大する際の安定性の確保、システムで実行可能なアプリケーションセットを拡張するための並列プログラミング技術の開発などです。このタイプのクラスタは、ノード間の相互動作が連続的に必要とされるため、「密結合」クラスタと呼ばれています。

## 管理性

より多数のサーバを限られた資源で管理するために、多くの IT 企業がクラスタリング技術を利用するようになってきています。クラスタ管理コンソールを使用すると、多数のシステムを一箇所で集中的に監視および設定できます。管理者は幅広い種類のアプリケーションに対し、問題点や障害の発生を一目で把握でき、適切な処置を取ることができます。これは、多数のサーバを抱える企業にとって、時間とコストの大幅な節約となります。さらに、ソフトウェアのバージョン管理と更新が容易になるという利点もあります。先進的なクラスタ コンソールでは、たった 1 つのコマンドにより、多数のノードにソフトウェア パッチを配布できます。

## 負荷分散

このタイプのクラスタは、各ノードが同一のサービスを提供するように構成されます。Web サーバは、最も典型的な負荷分散アプリケーションです。マスター スイッチまたは管理ノードに要求が送られると、負荷分散ソフトウェアにより、クラスタ化された各ノードにこの処理が分配されます。処理の分配方法には、ラウンド ロビン、ランダム、さまざまな加重アルゴリズムなど、多くの種類があります。負荷分散に使用される技術は、ハイ アベイラビリティ クラスタに使用される技術とよく似ています。実際、多くのハイ アベイラビリティ クラスタ実装には、負荷分散機能も組み込まれています。

## サーバの種類

### ワークステーション

ワークステーションは、多くのユーザが使用しているデスクトップ PC によく似ています。ワークステーションと通常のデスクトップ コンピュータとの主な違いは、堅牢な I/O サブシステムと、高性能な画像処理機能にあります。ワークステーションのメーカーが対象としているのは、エンジニア、科学者、およびデジタル コンテンツ制作者などの市場です。また、ほとんどのワークステーションには、ラインレートのパフォーマンスを提供する高速ネットワーク接続が実装されています。また、クロックレートの高い 1 基または 2 基のプロセッサが搭載されています。

### タワー型サーバ

タワー型サーバには、ラックを必要としないデスクトップまたはデスクサイドに配備するサーバ構成と、I/O 接続の追加を必要とするサーバ構成の 2 種類があります。一般にタワー型サーバは、1 ～ 4 基のプロセッサを搭載し、拡張された I/O 接続と大容量を備えています。

### ラックマウント タイプのサーバ

ラックマウント タイプのサーバは、最も普及しているサーバ構成です。このようなラックマウント タイプのサーバは、標準 19 インチ ラックでの占有ラック ユニット (RU) 数で分類します。その種類は、わずか 1 RU しか使用しないものから、ラック全体を占有するものまであります。多数のサーバベンダーは、高性能の 1 RU デュアル プロセッサ x86 ベースシステムを、クラスタアプリケーション用に特別に設計して提供しています。

### ブレード サーバ

ブレード サーバは、省電力かつ高性能なプロセッサと高速バックプレーンの発達により、現在大きな注目を集めています。ブレード サーバは、シャーシ (エンクロージャ) とサーバブレードの 2 つの部分で構成されます。シャーシのサイズは 4 ～ 6 RU で、電源、ネットワーク接続、およびバックプレーン インターフェイス接続をブレードに提供します。ブレードは、電源と I/O 接続以外のサーバ機能を 1 枚の回路基盤に実装したもので、シャーシに挿入し、シャーシのバックプレーンに接続して使用します。ブレード サーバはラックマウント タイプのサーバと比べ、はるかに高い CPU 密度を提供します。さらにブレード サーバは、管理性と保守性の面でも利点があります。たとえばブレードの交換やアップグレードは、使用中のブレードと新しいブレードをホットスワップするだけで済みます。

## スイッチの種類

### 固定構成

固定構成スイッチのポートは数と種類が固定されており、ポートの変更が必要になったときには機器自体を交換する必要があります。固定構成スイッチのなかには、冗長ファンや冗長電源などの基本的なアベイラビリティ機能を備えているものもあります。

## スタックابل

スタックابل スイッチは、固定スイッチの一種です。スタックابل スイッチは、スイッチの高速バックプレーンを他のスイッチにまで拡張することで、スイッチの性能を高めることができます。この機能には、スイッチの高速アップリンク ポートをお互いのスイッチに接続する場合と同様の効果があります。

## モジュラ型

モジュラ型スイッチは、最も用途の広いスイッチです。モジュラ型スイッチには、電源、高速バックプレーン、およびラインカードやスーパーバイザ エンジンを追加するためのスロットがあります。スイッチがサポートするすべてのレイヤ 2～4 の処理は、スーパーバイザ エンジンがインテリジェントに制御します。ライン カードによってスイッチのインターフェイスが提供され、さまざまな速度、ポート数、種類のインターフェイスが選択できます。モジュラ型スイッチは、新たなライン カードやスーパーバイザ エンジンを追加するだけで、簡単にアップグレードできます。

## 相互接続の種類

### イーサネット

イーサネットは世界で最も普及率が高く、最も広く展開されているネットワーク技術です。イーサネットはもともと、近くにある機器同士を接続する技術として開発されました。それがやがて、クラスタ相互接続、LAN、WAN などの幅広い種類のアプリケーションに対応する、優れた相互接続技術へと成長しました。イーサネットには、銅線ケーブルだけでなく、光ファイバケーブルを使用することもできます。

イーサネットを使用すると、Quality of Service (QoS; サービス品質) やセキュリティなど、他のクラスタ相互接続では実現できない高度な機能をもったネットワークを実現できます。QoS とは、それぞれのデータに指定された遅延要件を監視する機能です。たとえば、ノード間通信には高いプライオリティを割り当て、外部からの要求や管理トラフィックといった緊急性の低いトラフィックよりも優先されるように設定できます。

### InfiniBand

InfiniBand は、インターネット インフラストラクチャでの I/O 接続をサポートするために設計された新技術で、サーバ向け I/O 相互接続の次世代標準として設計されました。InfiniBand は、従来のコンピュータ内部でのバス接続の拡張を目的としており、コンピュータの内部と外部のどちらの相互接続にも利用できます。サーバ同士の通信やサーバとストレージ間の通信だけでなく、プロセッサとメモリ間の通信も、同一の I/O アーキテクチャによって処理できます。InfiniBand のリンク速度は、標準データレートである 2.5 Gbps の倍数として表現されます。最も一般的な InfiniBand は 4x ですが、このほかに 1x や 12x があります。

### Myrinet

Myrinet は高性能の packets 通信およびスイッチング技術で、遅延の少ないクラスタ相互接続が必要なアプリケーションに使用されます。表 1 に、クラスタ相互接続の各タイプの概要をまとめます。

表 1 クラスタ相互接続

インターフェイス タイプ	帯域幅	ケーブル タイプ	ポート単価 *
ギガビット イーサネット	全二重 1 Gbps	光ファイバまたは銅線	US\$500
10 ギガビット イーサネット	全二重 10 Gbps	光ファイバ	US\$3000
InfiniBand	全二重 10 Gbps (2.5 Gbps、30 Gbps)	銅線	US\$1500
Myrinet	全二重 2 Gbps	光ファイバ	US\$975

\* ポート単価には、スイッチ ポートとホスト アダプタも含まれます。価格は参考のためであり、実際にはさまざまな価格があります。

## ネットワーク パフォーマンス

クラスタの相互接続のパフォーマンスに最も大きく影響するのは、帯域幅と遅延です。帯域幅とは、1 回線で管理可能な情報量です。ほとんどのネットワークでは、帯域幅は 1 秒当たりのビット数 (bps) で表現されます。遅延とは、データが送信されてから宛先に到達するまでにかかる時間であり、通常はマイクロ秒単位で計測されます。よくある誤解は、帯域幅と遅延が互いに関連しているというものです。遅延を少なくしても、ネットワークが仕様以上の帯域幅を提供することはありません。

ほとんどのネットワークでは、帯域幅は簡単に拡大できます。1 回線の帯域幅が 1 Gbps の場合は、2 回線にすれば帯域幅は 2 Gbps になります。これは単純化した例ではありますが、回線を増やせばそれだけ帯域幅も拡大します。イーサネット標準の IEEE 802.3ad では、複数のイーサネット接続を使って帯域幅を拡大するための仕様を定義しています。

一方、遅延の改善はそれほど単純ではありません。1 Gbps のネットワーク上で 1 Gb のデータブロックを転送する場合、データが宛先に到達するまでの時間は 1 秒 (帯域幅) + x (遅延) です。ネットワーク速度を上げて 1000 Gbps にすれば、この時間は 0.001 秒 + x (遅延) となります。多くのネットワーク アーキテクチャでは、大きなデータブロックをパケットと呼ばれる小さなブロックに分割することで、高速な転送を実現しています。しかし、すべてのパケットにはネットワーク遅延が追加されます。ネットワーク遅延は、ネットワーク上で転送されるあらゆるデータの転送時間を増加させる要素です。

ネットワーク遅延は、データがさまざまなネットワーク機器を経由して転送される過程で生じます。データパス上にある各機器の処理速度および数が、遅延を決定する最大の要因となります。すべてのイーサネット スイッチの遅延は同じだという考えも、一般的な誤解です。イーサネット スイッチのメーカーによっては、遅延をミリ秒 (1/1000 秒) 単位で測定していますし、シスコシステムズではスイッチの遅延をマイクロ秒 (1/100 万秒) 単位で測定しています。スイッチに組み込まれているチップセットの品質と性能によって、スイッチの遅延に関するパフォーマンスは大きく異なります。

クラスタの設計、およびクラスタ上で動作するアプリケーションの種類によっては、帯域幅や遅延がパフォーマンスの拡張性に大きく影響します。以下に、典型的な科学技術計算処理の例と、これらの処理における相互接続の問題について説明します。

## 大容量環境

### レンダリング ファーム

写真のように精巧なキャラクターが本物のような 3 次元イメージとして登場する、驚くべき品質のアニメーション映画が制作されています。映画は、何千ものフレームによって作成されます。1 つのフレームは 1 つの画像であり、これらを連続して表示すると、人間の目には一続きの動きのように見えます。個々のフレームは互いに完全に独立しているため、複数のコンピュータを使用している場合でも、互いにほとんど情報のやり取りを行わずに各フレームをレンダリングできます。

ほとんどのレンダリング ファームは、多数の小型高速サーバを高速かつ信頼性の高い方式で相互接続することで構成されています。クラスタ内の 1 つのノードはディレクタの役割を担い、各フレームのレンダリングに必要な情報を他のノードに配布します。各ノードは 1 つのフレーム処理が完了すると、その情報をディレクタ ノードまたはストレージ デバイスに送り返してから、引き続き次のフレームのレンダリングを開始します。1 フレーム当たりの処理時間は、その複雑さに応じて数マイクロ秒から数時間程度までさまざまです。

レンダリング ファームの相互接続設計には、ディレクタ ノードと実際に処理を行うワーキング ノードとの接続に高い帯域幅が必要になります。遅延やノード間の帯域幅はそれほど重要ではありません。ディレクタ ノードは、ノンブロッキングのワイヤスピードでネットワークに接続する必要があります。ワーキング ノードはオーバーサブスクライブであっても問題ありません。帯域幅要件を決定するには、次の数式を使用します。

フレームの水平解像度 × 垂直解像度 × ピクセル当たりのビット数 (bpp) = 転送ビット数

平均的なレンダリング時間が 1 秒未満であれば、ネットワークは全ノードからのフレーム転送を同時に処理する必要があります。レンダリング時間がこれより長くかかる場合には、レンダリング時間に比例して帯域幅要件も低くなります。次の例を参考にしてください。

フレーム解像度 1024 × 768

色深度 = 32

ビット数 = 25 Mb

ノード数 = 24

必要な帯域幅 = 603 Mbps

## 大容量アプリケーション

### 計算流体力学

計算流体力学や負荷の高いアプリケーションの処理を想定したクラスタでは、1つの大きな問題を細分化し、その各部分を処理するためのノードを多数用意します。あるノードの出力が別のノードへの入力となるため、この設計では遅延がスケーラビリティの障害要素となります。クラスタでの処理はディレクタ ノードによって管理されますが、多くの処理は、前の処理からの出力が転送されるまで開始できません。ほとんどの場合、この相互接続間では小さいサイズのデータ パケットだけが送受信されます。あるノードが、他のノードから送られる処理結果を待機する間は、CPU サイクルが無駄に消費されることとなります。このようなタイプのアプリケーションに対して相互接続方式を選択する場合は、遅延が最も重要な課題です。

### クラスタのスケーリング

ほとんどのクラスタ実装で制限となるのは、相互接続です。クラスタを設計する上での最終目標は、線形スケーリングです。つまり、1つのノードからなるシステムが1つの計算を完了するまでに60分かかるのであれば、2つのノードからなるクラスタによる処理時間は30分となることです。ただし、これは実現可能な計算というよりは理論上の目標です。ノードをクラスタに追加すると、一定量のオーバーヘッドも追加されます。大規模なクラスタ実装では、スーパー ノードまたは管理ノードと呼ばれる、ノード間通信を管理するための専用マシンが必要です。これらのノードは、システムの処理能力には関与しませんが、クラスタのオペレーションには重要な役割を果たします。

クラスタのパフォーマンスをどれだけ拡張できるかは、クラスタ化したアプリケーションのノード間通信の特性に依存します。前に説明した4つのタイプのクラスタのどれについても、スケーリングに関してそれぞれの課題があります。パフォーマンスの向上を目的としたハイ アベイラビリティ クラスタの管理者は、「計画的および計画外のダウンタイムをどのようにすれば減らすことができるか」、「特定の場所、地域、または国全体で発生する災害からアプリケーションをどのように保護するか」といった課題を解決しなければなりません。ハイ アベイラビリティ クラスタを拡張する場合は、クラスタ相互接続の距離およびパフォーマンス要件が厳しくなります。必要な距離が長くなれば、光ファイバケーブルの導入も必須となります。

容量の拡大を目的としたスケーラビリティ クラスタの管理者であれば、「アプリケーションの処理時間をどうしたら短縮できるか」、「ノードの追加時に線形スケーリングを維持するにはどうすればよいのか」といった問題に直面します。スケーラビリティ クラスタのパフォーマンスを向上させるには、複数のノードをワイヤ スピードで接続できるスイッチング ファブリックが必要になります。最初にクラスタを設計する時点では、遅延だけが問題になります。既存のクラスタにノードを追加しても、ネットワークの遅延は増加しません。ただし、クラスタにレイヤを追加すると遅延も増加します。これは、各パケットが目的のノードに辿り着くまでに通過しなければならないスイッチが増えるためです。

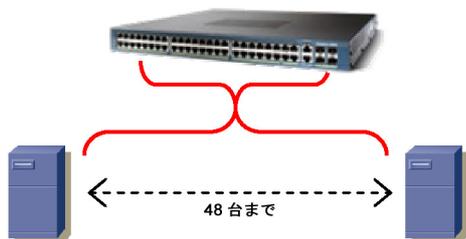
多数のサーバを管理するように設計されたクラスタには、「もっと離れた場所から、より多くのシステム数を管理するにはどうしたらよいか」といった問題が伴います。ここでは、ハイ アベイラビリティ クラスタと同様、相互接続距離が重要な要素になります。最後のクラスタ タイプである負荷分散型クラスタは、線形スケーリングを簡単に実現しやすいといえます。負荷分散アプリケーションでの相互接続方式における主な要素は、帯域幅と複数のノードをサポートできる能力です。

### クラスタ構築のためのアーキテクチャ

クラスタの規模は、アプリケーション、スペース、予算、利用可能な技術といったさまざまな要因に応じて、2つのノードを直接接続する形態から、何千ものノードで構成されたキャンパス クラスタまで、多岐にわたります。

## 単一層構造

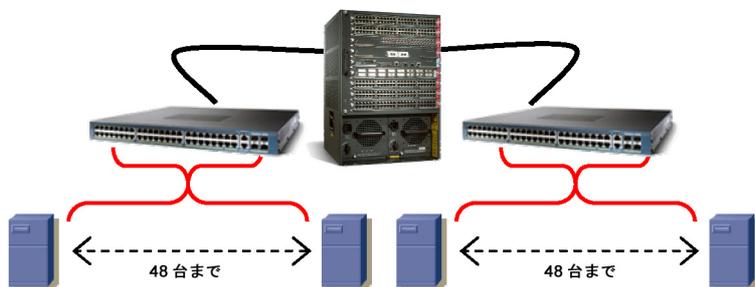
図1 単一層構造のネットワーク設計



単一層クラスタは、単一レベルのスイッチング ファブリックが特徴です。図1に示す48ポートの固定構成スイッチでは、ノンブロッキング ラインレート対応のギガビット イーサネット ポートに最大48のノードを接続できます。このトポロジーでは、非常に高速な、低遅延のクラスタ ノード間通信を実現できます。単一層クラスタは、スイッチのポート数を増やすことで拡張できます。固定構成スイッチまたはスタックブル スイッチで実現できるのは、4～48ポートまでです。それ以上の数のノードをクラスタ化するには、モジュラ型スイッチを使ったほうがよいでしょう。モジュラ型スイッチは数百ポートにまで拡張でき、構成次第では、どのポート間でもラインレートのパフォーマンスを提供することが可能です。多くのクラスタでは、ノード間通信のための専用のネットワークが構築されます。また、クラスタとLANの両方のトラフィックを伝送する統合ネットワークを使ったクラスタ アプリケーションの場合は、高速アップリンクを使用してネットワーク コアに接続します。

## 2層構造

図2 2層構造のネットワーク設計



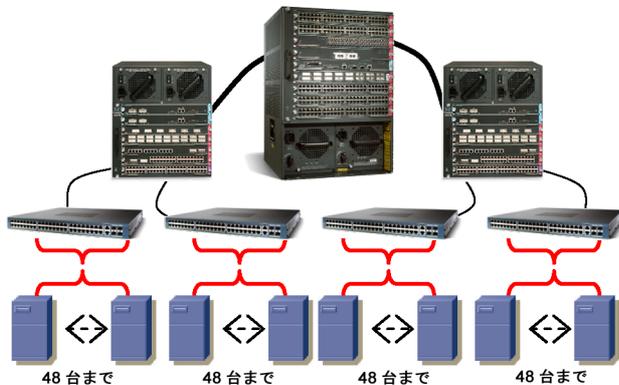
このタイプの相互接続設計(図2)では、より大きなスケーラビリティを実現でき、場合によっては1台のモジュラ型スイッチを使用するよりもコストを抑えることができます。ただし、一方のサーバグループ内のノードは、他方のサーバグループ内のノードとの接続に対して、ノンブロッキング ラインレートのパフォーマンスを提供できません。ラインレート パフォーマンスを実現するには、上りの帯域幅と下りの帯域幅を一致させます。イーサネット環境では、高速アップリンクにより、多層クラスタ構成でのラインレート パフォーマンスを実現できます。モジュラ型スイッチを使用する構成であれば、1台のスイッチの全ポートでラインレート パフォーマンスを提供することも可能です。レイヤ1スイッチからのアップリンクは、いったんアグリゲーション スイッチに接続されてから、ネットワーク コアに接続されます。この構成が実現可能なソリューションとなるかどうかは、アプリケーション、およびネットワーク上を流れるトラフィックの種類に依存します。このトポロジーを決定する基準は、ラインレート パフォーマンスが必要かどうかという点です。

すべてのノード間でラインレート パフォーマンスを実現可能な、Fat Tree 設計と呼ばれるスケーラブルなトポロジーがあります。Fat Tree 設計では原則的に、ツリー内の各レベルが備える回線数および帯域幅が、それぞれの1つ前のレベルよりも少なくなります。下りのエンド ノードへの接続の帯域幅は、相互接続に使用する上りパスへの接続の帯域幅と同じです。

このネットワーク設計のもう1つの利点は、ケーブル管理が容易なことです。各サーバ ノードとのネットワーク接続は、1つのラック内に集約されます。モジュラ型スイッチを使った大規模なアプリケーションでは、距離がかなり離れている場合であっても、個々のサーバをモジュラ型スイッチに接続しなければなりません。

### 3 層構造

図3 3層構造のネットワーク設計



一般に3層構造の設計（図3）では、24ポートまたは48ポートの小型スイッチを使って各ノードを接続します。これらのレベル1スイッチからのアップリンクは、高速アップリンク機能によってアグリゲーションスイッチと接続されます。アグリゲーションスイッチは、クラスタ化されたコアネットワークスイッチとアップリンク接続します。この設計は、ノード間でラインレートパフォーマンスを必要としない、大規模なクラスタに最適です。図3で示しているのは最小構成での接続例です。さらに接続数を増やし、アベイラビリティやパフォーマンスを高めることも可能です。

多層クラスタ設計では、高性能イーサネットスイッチのレイヤ3ルーティング機能が大きな役割を果たします。レイヤ3ルーティングを使用すると、ネットワーク上で伝送される集約トラフィックの量が減少するため、ネットワークパフォーマンスが向上します。VLANを使用するとサーバグループまたはクラスタ全体を分割し、VLANを不要なネットワークトラフィックから分離できます。

### 結論

複数のコンピュータの相互接続により、問題解決と情報の共有を可能にすることは、シスコの20年以上にもわたる強みです。何年もの間、コンピュータクラスタは主に、高価な独自仕様のハードウェアで構成されたハイエンド実装に使用されてきました。しかし近年では、低価格で高性能なx86ベースサーバが導入され、クラスタ対応のソフトウェアおよびOSも改良が続けているため、クラスタの人気も高まっています。

ゼロからクラスタを設計する場合、最も重要な決定事項の1つとなるのが、相互接続のタイプです。相互接続の性能が不十分だと、情報処理の完了を待機する間、ノードがアイドル状態になってしまうため、クラスタのパフォーマンスが制限されます。逆に必要以上に高性能の相互接続を選択すると、クラスタ構築が高コストで複雑になってしまうため、その分のコストを処理能力の増強にかけの方が妥当ということにもなります。ネットワークトラフィックとアプリケーションの依存性を正確に分類した上で、適切な相互接続技術を選択するようにしてください。アプリケーションに適した正しい相互接続技術を選択することが重要です。

©2005 Cisco Systems, Inc. All rights reserved.

Cisco、Cisco Systems、および Cisco ロゴは米国およびその他の国における Cisco Systems, Inc. の商標または登録商標です。  
この文書で説明した商品、サービスはすべて、それぞれの所有者の商標、サービスマーク、登録商標、登録サービスマークです。  
この資料に記載された仕様は予告なく変更する場合があります。



シスコシステムズ株式会社

URL: <http://www.cisco.com/jp/>

問合せ URL: <http://www.cisco.com/jp/go/contactcenter/>

〒 107-0052 東京都港区赤坂 2-14-27 国際新赤坂ビル東館

TEL: 03-6670-2992

電話でのお問合せは、以下の時間帯で受付けております。

平日 10:00 ~ 12:00 および 13:00 ~ 17:00

お問合せ先