インターネット環境におけるハイアベイラビリティの考え方

white paper

目次

hp invent

概要

ドット・コム系企業で、実行しているインターネット・アプリケーションが使用できなくなると、その時点で収益の道が絶たれることになります。こうした状況が発生した場合、これまでの顧客を別のサービス・プロバイダに取られてしまうことを考えれば、その影響は単にサービスの停止期間だけに限らず、深刻な事態にまで発展しかねません。今や、ドット・コム系企業やホスト側のサービス・プロバイダ各社にとって、ハイアベイラビリティ(HA)環境は任意の選択肢というより、必須の条件となりつつあります。

HA コンピューティング環境を設計するにあたり、その主体となる企業をはじめ、プラットフォーム・ベンダやサービス・プロバイダの各社が HA に関連する概念と用語について共通の認識を持つことが不可欠の条件となります。この資料では、アベイラビリティと障害に関する一般的な用語を定義することに加え、ハイアベイラビリティを実装するためのさまざまな手法、およびインターネットに特化した HP-UX 11i オペレーティング環境における実装方法を紹介します。

ドット・コム系企業向けのハイアベイラビリティ

最近ではドット・コム系企業をはじめ、それをホストするサービス・プロバイダの各社の間には、ハイアベイラビリティ(HA)環境を求める究極のニーズが広がりつつあります。ドット・コム系企業で、万が一インターネット・アプリケーションに障害が発生すると、ビジネスを続行するための手段が失われる結果になります。サービス停止の間は収入源が絶たれると同時に、これまでの顧客が他社に奪われ、ひいては企業イメージもダウンにまで至ることも考えられます。こうした意味で、HA 環境は単にドット・コム企業向けのテクノロジ戦略という次元を越え、ビジネス戦略の上でも不可欠の要素となりつつあります。

これまで、アプリケーションの停止の主たる原因は、ハードウェアの故障にありました。最近ではハードウェアの信頼性が向上するに伴い、アプリケーションの停止をもたらす最大の要因は、アプリケーション、オペレーティング・システム・ソフトウェア、人的エラーといった領域へ移行していく傾向にあります。したがってドット・コム企業では、SLO(Service Level Objectives:サービス・レベル・オブジェクティブ)構想の実現に向けて、社内のインフラや人的プロセス全体に適用できる HA ソリューションを導入していく必要があります。

アベイラビリティに関する用語

システムのアベイラビリティに寄せられる期待の度合は、具体的な条件によって大きく異なるものです。ここではまず、アベイラビリティに関するニーズとそのソリューションについて共通の認識が得られるように、一般的な用語の意味を定義しておきます。

アップタイム

アップタイムとは、ユーザーにアプリケーションへのアクセスが提供される時間のことです。一般には、1 週あたりの時間数で表します。たとえば 24×7、すなわち 1 日 24 時間、週 7 日間という言い方をします。アップタイムには一般に、保守作業の時間、すなわち保守作業が予定されているためにアプリケーションが一時的に使用できなくなる期間は含まれません。仮に、1 週間のうちアプリケーションが使用できなくなる期間を 6 時間とすると、24×6.75 のように表す方が現実的です。この間に、アプリケーションとシステムの全面的な保守作業を実施します。アップタイムの値を算定するときは、バックアップの所要時間や、バックアップ時にアプリケーションへのアクセスを停止させる必要があるかといった点を現実的に考慮する必要があります。

ダウンタイム

ダウンタイムは、アップタイムの逆の概念です。つまりユーザーがアプリケーションにアクセスできなくなる時間のことです。ダウンタイムには、計画ダウンタイムと計画外ダウンタイムの 2 種類があります。

- 計画ダウンタイムとは、データベースの再構成、ディスク空間の追加、バックアップの実行、新規/改訂アプリケーションやシステム・ソフトウェアのインストールなど、事前に予測できる保守作業を実施するためにあらかじめ確保しておいた時間のことです。
- 計画外ダウンタイムとは、ハードウェアやソフトウェアの障害といった予定外の問題が発生したために起きるダウンタイムのことです。

ダウンタイムの期間は、アプリケーションのシャットダウンからスタートアップまでに所要する時間のことで、場合によっては長期間に及ぶこともあります。

サービス停止

サービス停止とは、何らかの理由でユーザーがアプリケーションにアクセスできなくなる状況のことです。たとえば、停電、火災、洪水、地震といった環境要因から、システム・ハードウェア、システム、アプリケーション・ソフトウェアの障害までさまざまな原因が考えられます。ユーザーの最大の関心は、障害の原因が何であるかではなく、障害やサービス停止が発生したという事実に向けられます。ダウンタイムには事前に計画できる要素はあるものの、ユーザー側から見れば、いずれにせよアプリケーションにはアクセスできないという意味で、サービス停止には変わりありません。

アベイラビリティ（可用性）

アベイラビリティ（可用性）は通常、年間ベースで測定する概念で、ユーザーがアプリケーションを利用できる期間を比率で表したものです。アベイラビリティでは一般に、計画外ダウンタイムも計算に入れますが、計画ダウンタイムは考慮しないことがあります。したがって、アベイラビリティの条件が 99.86 %とある場合、年間を通じた計画外ダウンタイムが 12 時間まで認められることになります。アベイラビリティは次の式で求めることができます。

$$\frac{(8760\ \text{時間} - \text{年間あたりの計画外の時間数})}{8760}$$

8760 は、365 日×24 時間=8760 時間の値です。仮にアプリケーションが、1 日 24 時間、利用不可になった場合、8760 はアプリケーションをアクセス可能にすべき時間の年間あたりの数字になります。

システムで提供されるアベイラビリティのレベルは個々の条件によって異なります。標準的（基本的）な高信頼システムは、ハードウェアの基本的な信頼性に応じて、アベイラビリティは 95～98%の範囲で変動します。またハイアベイラビリティ・システムでは 99%以上の可用性が保証されます。さらに継続的なアベイラビリティ・システムでは 99.999%となり、計画外ダウンタイムは年間あたり数分以内という水準が達成されます。

アベイラビリティは、購入する大型コンピューティング・システムの仕様を記述する提案要求（RFP）で必須の条件とされる場合があります。一般的な基準として 99.86%（年間 12 時間）～99.8%（年間 18 時間）程度のアベイラビリティが要求されます。

フォールト・トレラント

フォールト・トレラントと称するシステムは、同時並行で動作する複数のハードウェア・コンポーネントが搭載され、演算処理と I/O 動作すべてを複製します。フォールト・トレラント・システムでは、シングル・システム中のハードウェア・コンポーネントを冗長構造で組み込むことにより、ハードウェアの障害からシステムを保護します。なおフォールト・トレラントとは、システムが決して故障しないという意味ではありません。つまりシステムやアプリケーション・ソフトウェアに障害が起きれば、アプリケーションは容易に使用不可になり得るということです。通常フォールト・トレラント・システムの価格は、非フォールト・トレラント・システムの約 10 倍ですが、ハイアベイラビリティ・ソリューションとはみなされません。

SPOF（単一障害点）

SPOF（単一障害点）とは、仮に特定の部分に障害が発生した場合に、システムやアプリケーションの停止にまで発展してしまう重要なコンポーネントのことを指します。SPOF の典型的な例として、次の要素が考えられます。

- コンピュータ・システム（SPU）
- ディスク
- ホスト・アダプタ/ケーブル
- ネットワーク
- 電源

ハイアベイラビリティの手法を利用することにより、SPOF 要因の大部分を取り除くことができます。たとえば、一次 SPU に障害が生じた場合にも相互のアプリケーションの動作が保証されるように、クラスタ・ソフトウェアを使用して、複数の SPU をリンクすることができます。データの保護には、ディスク・ミラーやディスク・アレイといったテクノロジを利用できます。また、UPS（無停電電源装置）や冗長構造の電源を採用することにより、電源障害による影響を未然に防ぐことができます。さらにホスト・アダプタの障害に対しては、オペレーティング・システムの機能とホスト・アダプタの冗長化によって対応することができます。

フェールオーバ

フェールオーバとは、アプリケーションやデータを含め、システムの動作を故障したサイトからバックアップ・サイトへ切り換えることを示します。フェールオーバに要する時間は、わずか数分で済むことから、アプリケーションの復旧時間も含めて、数時間以上に及ぶこともあります。

ハイアベイラビリティ(HA)

SPOF の要因を取り除くことは、コンポーネントの冗長化と、迅速な復旧と即時フェールオーバといった対策によって、ハードウェアにハイアベイラビリティの機能を持たせることです。冗長構造を適用できるのは、特定のコンポーネントやバスで単一障害点の起きる部分に限定されます。たとえば、デュアル構造のミラー・ディスクの双方で障害が起きた場合は、保護されません。HA が適用されるのは単にハードウェアの領域だけではありません。アプリケーションの可用性が可能な限り保証されるように、管理と運用、さらにアプリケーションの領域でも相応の変更が必要になります。ハイアベイラビリティ・システムの主な目標は、従来の標準的なシステム(非 HA システム)を上回る高度なアベイラビリティをフォールト・トレラント・システムよりも安価な価格で実現することです。

連続アベイラビリティ

次世代の HA システムの目標は、継続的なアベイラビリティを実現することです。ハードウェアとソフトウェアのいずれにも障害が起きる可能性がありますが、継続的なアベイラビリティの目的はあくまでも、ユーザーの業務を障害から隔離すると同時に、障害からの復旧時間をわずか数分以内にとどめることにあります。

HA を実装すべきかどうかを判断するにあたって、ハードウェアおよびソフトウェアのコストと、ダウンタイムの発生に伴うコストを比較する必要があります。ダウンタイム時のコストには、たとえば次のようなことが考えられます。

- 実際の収益損失
- 永久的な顧客の喪失
- アイドル状態、または有効利用されていないリソースのコスト
- 作業遅れに伴うコスト
- 罰金
- データの復元に要する時間のコスト

投資回収(ROI)に伴う期間は、上記のダウンタイム・コストを考慮して算定する必要があります。典型的な組織の場合、投資の回収には 6～12 か月の期間がかかります。

またダウンタイムには、データの永久喪失、翌日までの作業延期、顧客の不満といった無形の犠牲もかかわってきます。

災害復旧(Disaster Recovery)

火災、洪水、地震といった自然災害の影響から、短期間で復旧することを災害復旧(Disaster Recovery)といいます。こうした災害が発生すると、結果としてシステムが物理的に破壊されたり、データの喪失、通信の切断、作業空間の喪失といった事態が生じます。復旧にかかる時間はわずか数分間で済む場合や、数日、数週間に及ぶことまであります。また復旧に要する時間は、システムへのアクセス、データやアプリケーションのロード、通信の復元の時間に直接依存する場合があります。災害復旧の為の冗長性は一般に、地理的に離れたリモート・サイトで重複したシステムを構成することによって提供されます。

災害復旧に関連する 2 つの重要な問題として、データの複製とデータの鮮度という条件があります。データを別のサイトで複製することは、リンクの距離的な条件と転送速度の影響を受けます。複製の方法が低速であるほど、災害が起きたときに、失われるデータの量は多くなります。

災害復旧に関連したソリューションとサービスを求めるニーズは最近、急速な勢いで高まっています。災害が起きた後のダウンタイムに伴うコストは膨大な額に及ぶほか、システムやアプリケーションに対するアクセスの復旧に向けたニーズは高まる方向です。災害が組織に及ぼす影響度は、災害復旧ソリューションの実装に必要なコストとともに算定する必要があります。

コンポーネントの障害に関する用語

まず障害に関する専門用語を正しく理解し、適切に利用することがハイアベイラビリティや HA ソリューションの概念を把握する上で第一の条件となります。

次の一般的に使用されている用語は、ときに誤って認識されることがあります。いずれの用語もコンポーネントの障害を表すもので、システムの障害を指すものではありません。組織の立場からすると、システム・アベイラビリティの仕様で捉える傾向がありますが、ここでコンポーネントの障害とシステムの障害の間でそれぞれの統計データの関係を理解しておくことが重要です。

MTBF（平均故障間隔：Mean Time Between Failures）

MTBF (平均故障間隔: Mean Time Between Failures) とは、ユニット全体のパフォーマンスに関する過去のデータに基づいて、予測される今後のパフォーマンスを予測するものです。ハードウェアが新しければ、過去のデータに基づいて MTBF を予測することはできません。とは言え、いずれかの数式モデルを使用して、特定のコンポーネントに予測される MTBF を計算することは可能です。こうした計算の精度は、ベンダ、アセンブリの品質管理、コンポーネントの品質などによって大きく異なるものです。
MTBF は次の数式で計算します。

$$\frac{\text{実際の動作時間の合計}}{\text{障害の合計回数}}$$

分子は、実際の動作時間を表す値で、該当するコンポーネントを有効にした日の時間数で表します。

MTTR（平均修復時間：Mean Time To Repair）

MTTR (平均修復時間: Mean Time To Repair) は、実際の統計データに基づいて、コンポーネントの修復にかかる時間の平均値を表したものです。MTTR は、次の方法で計算することができます。たとえば、MTTR をコンポーネントの交換に必要なオンサイトの時間として適用すると、顧客はユニットが故障して交換するまでの時間が含まれるものと解釈します。これには恐らく、ハードウェアが使用不能になる状態、レスポンス・タイム、移動時間、そして実際のオンサイトにおける修復時間も含まれていることが考えられます。

AFR（年間故障率：Annualized Failure Rate）

AFR (年間故障率: Annualized Failure Rate) とは、信頼性を計るもう 1 つの基準です。ユニット時間の単位で表す MTBF や MTTR とは異なり、AFR (Annualized Failure Rate) は常に比率で表します。AFR は、次の数式で計算します。

$$\frac{\text{合計故障回数} \times 8760 \times 100}{\text{連続経過時間の合計}}$$

連続経過時間の合計には、コンポーネントの動作時間とダウンタイムの両方が含まれます。仮に AFR が 200%とある場合、年間あたりの故障回数は 2 回、また 50%であれば、2 年あたりに 1 回故障するということになります。

システムの故障率

システムの故障率は、特殊な方法で計算します。この数式では、同じコンポーネントについては乗算、別のコンポーネントについては加算を使用します。システムの故障率は次の数式で計算します。

$S = \Sigma$（特定のコンポーネント数 × このコンポーネントの故障率）

システムの故障率には、コンピュータ、ディスク、I/O カードなど個々のコンポーネントすべてを対象とした故障率、すなわち MTBF が含まれます。これゆえ、同じ種類の複数のコンポーネントを使用すると、特定のシステムの故障率に影響を与えることになります。

例えば、最近のディスクの MTBF は一般に 500,000 時間とも言われています。すなわち 57 年間に 1 回の頻度で故障するということです。しかし、ディスクがこれほど長期間にわたって使われることは現実にはありません。MTBF は、実際の故障に基づいて計算する値です。また、MTBF は単にディスク・メカニズム自体に適用される概念です。電源、コントローラ、ファンといった要素も考慮に入れると、MTBF は 200,000 時間、すなわち 22 年以上になります。

1 つのシステムに 200 台ものディスクが使われていたら、いったいどうなるのでしょう。この場合の故障率は乗算で計算するため、200 台のディスクの MTBF は 200,000 時間/200 で、すなわち 1000 時間となります。AFR は約 900%、つまり年間 9 回の故障が発生することになります。

以上の数式から、大規模なシステムでは故障は頻繁に起きる可能性があることがわかります。アプリケーションやユーザーに影響を及ぼす可能性のある故障の頻度や復旧までの時間を減らす対策として、HA ソリューションを導入する方法があります。

アベイラビリティ

アベイラビリティとは、ユーザーがシステム（アプリケーションも含むことが多い）にアクセスし、利用できる合計時間の比率のことを言います。アベイラビリティは、次の数式で表します。

$$\frac{\text{合計経過時間} - \Sigma\text{動作不能時間}}{\text{合計経過時間}} \times 100$$

アベイラビリティに関する格言

- システム・アベイラビリティは本来、単一のコンポーネント（ことの是非を問わず）に基づいて評価するものではない。
- システム全体から可用性(HA)を測定しない限り、結局は最も信頼性の低いコンポーネントの値となってしまう。
- ハードウェアの数が多いほど、いずれかのコンポーネントに障害が起きる可能性が高くなる。

ハイアベイラビリティの実装方法

ハイアベイラビリティは、保護するコンポーネントの種類によってその実装方法が異なります。総合的な HA ソリューションとは、これらの手法をすべて統合したアプローチのことを言います。こうした手法にはそれぞれハードウェアやソフトウェア、その両者を伴うことがあると同時に、アプリケーション、システム管理、動作手順といった領域の変更に関連することもあります。HP-UX 11i オペレーティング環境と特定の HP 製品を組み合わせることにより、幅広い高度な HA 機能を提供しています。

RAID によるデータ保護

一般に、データは組織にとって不可欠の要素であり、コンピュータ・システムに保存されたデータは厳重に保護する必要があります。データを保護するための最も簡単な方法は、頻繁にバックアップを取得することです。しかし、バックアップを実行しても、ディスク・ドライブが故障する時点までに生成されたデータが現実には取得できていないことも想定されます。

データは、いくつかのレベルで保護することができます。いずれの場合も RAID のレベルとして定義します。RAID はハードウェアの手法として見られがちですが、実際にはソフトウェア、またはハードウェアのいずれかを利用して実装されます。なお、すべての RAID レベルでデータが保護されるわけではないこと、またデータ保護が要件に合わなければ、HA も実現されない、という点に注意する必要があります。

最も一般的に利用されている RAID のレベルは次のとおりです。

- **RAID 0** — データは保護されず、ディスク・グループを対象としてインターリーブ（ストライプ）されます。RAID 0 は HA ソリューションではありません。
- **RAID 1** — 各種のディスク・メカニズムによってミラー化を行うことにより、データを保護します。
- **RAID 0/1** — ブロック・ストライプ/ミラーリングを組み合わせた手法です。
- **RAID 3** — パリティ生成（XOR）によりデータを保護します。ディスク・グループ全体を通じてバイト・ストライプを使用します。
- **RAID 5** — パリティ生成（XOR）によりデータを保護します。ディスク・グループ全体を通じてブロック・ストライプを使用します。

RAID 1 では、各ミラー・コピーに 100%の追加ディスク容量が必要になります。したがって、最大の欠点はコストということになります。読み取り時のパフォーマンスは通常、要求の待ち行列が最も短いミラー・コピーに I/O 要求を入れれば、改善されます。書き込み時のパフォーマンスは、RAID 1 では複数の書き込が同時に行われるため、一般には低下します。

RAID 5 では、5 種類のディスク・メカニズムで構成された各 RAID グループごとにパリティ用として 25%の追加ディスク容量が必要になります（これは一般的な実装方法です）。したがって、RAID 1 よりもコスト効果のあるソリューションとなります。RAID 5 による読み取り時のパフォーマンスは一般に改善されますが、RAID 1 に比べ効果は大きくありません。パリティが生成されるため、RAID 5 による書き込みパフォーマンスは、I/O サイズが小さい場合（64 KB 以下の場合）に大きく低下することがあります。

MirrorDisk/UX

HP-UX 11i には、RAID 0 と RAID 1 の 2 つのレベルがソフトウェア上で実装されています。MirrorDisk/UX は RAID 1 ソリューションを提供するソフトウェア製品で、これを使用することにより、選択された論理ボリューム（LV）について 2 つまたは 3 つのミラー・コピーが提供されます。いずれのミラー・コピーに障害が起きた場合は、読み取りは正常に動作しているコピーを使用して継続されます。3 つのミラー・コピーにより、同じデータを含む 2 つのディスク・メカニズムに障害が発生した場合にも、データが保護されます。また分割機能を併用することにより、同じシステム、または別のシステム上にあるデータのスナップショット・コピーをバックアップしたり、アクセスすることができます。

ディスク・アレイと JBOD

MirrorDisk/UX は、スタンドアロン・ディスク（JBOD："just a bunch of disks" -「ディスクの集合」の意）にも、ディスク・アレイにも利用できます。ディスク・アレイでは、ハードウェア上で各種の RAID を実装することにより、RAID 1 の複数の書き込みや RAID 3 と RAID 5 のパリティ生成に伴うシステム・オーバヘッドを軽減しています。ディスク・アレイは、ソフトウェア・ミラーリングを使用した JBOD よりも優れたコスト効果を発揮することがあります。

ディスクのリンクには SCSI インタフェースが幅広く利用されています。SCSI は、現在では大型のコンピュータ・システムをクラスタ化する場合に何台ものディスクを接続できるという点で、本来の設計をはるかに越えて進化しています。ディスク・アレイでは、アレイ全体で 10 個、20 個、あるいはそれ以上のディスク・メカニズムが組み込まれていても、SCSI のターゲット・アドレスは 1～2 個しか使用しないことから、JBOD に比べ、占有する SCSI ターゲット・アドレスは少なくて済みます。これは、1 つのアレイが、ターゲット・アドレスのサブアドレスである SCSI LUN を介してアクセスする RAID グループに分類されているからです。システム・ディスク全体の容量は、SCSI ホスト・アダプタの数によって制限されるため、ディスクの容量は、ディスク・アレイよりはるかに大きくなります。

最近のディスク・アレイには、複数のコントローラ、電源、ファンが搭載されており、いずれも各コンポーネントの SPOF 要因が排除されています。ディスク・アレイ全体が故障する可能性はわずかながら残されており、これは、ミラーJBOD がディスク・アレイよりも有利な点です。

論理ボリューム・マネージャ

LVM（論理ボリューム・マネージャ）は、HA ソリューションには分類されませんが、HP-UX 11i に標準装備されているため、PV（物理ボリューム）のグループを通じてデータをストライプ化することにより、オプションとして RAID 0 を実装することができます。

ディスク・ホスト・アダプタの保護

HP-UX 11i の標準で提供される LVM（論理ボリューム・マネージャ）の機能のことを PVLink と言います。PVLink では、同じ物理ディスク・デバイスについて複数のパスが確保されます。この機能は、複数のコントローラを搭載したディスク・アレイには最も優れた効果があります。各ディスク・アレイ・コントローラは、それぞれ別のホスト・アダプタに接続されます。一次ホスト・アダプタ、またはケーブルに障害が生じた場合、LVM は自動的に冗長ホスト・アダプタに切り換えられます。したがって、ホスト・アダプタとケーブルの SPOF 要因は解消されます。

PVLink 機能は、各ミラーが個別のホスト・アダプタに接続されるため、ミラーJBOD では必要なくなります。また PVLink では現在のところ、同じ物理ディスク・デバイスに対する各ホスト・アダプタ上で複数の並列 I/O アクセスはサポートされていません。

データの複製（Data Replication）

DR（Data Replication: データ複製）も、データを保護するための手段となります。DR には通常、データを別のシステムに保存するために、転送処理が必要になります。なお、HA という観点から、データ複製は主要サイトから地理的に離れた別のサイトで保存されます。万が一、災害が主要サイトで起きた場合でも、重要なデータは DR サイトで利用できるようになります。

データは、次の方法で複製することができます。
- データベース複製ソフトウェアを使用する
- 複製用のカスタム・アプリケーションを使用する
- トランザクション処理モニタによる複製
- リモート・データ・ミラーリング

データ複製に伴う最大の問題は、データ変更の頻度に応じて、最適のパフォーマンスを維持する目的で各サイト間に設定するリンクにコストが必要になる点です。このコストは、組織が災害時に失ってもかまわないと認識するデータの量と勘案して算定する必要があります。例えば、低速リンクは DDR サイトにおけるデータまで 4 時間を経過している可能性があります。仮に災害が起きると、転送を待機しているデータが失われます。また、テープにデータをコピーしたり、人手を使ってサイト間を物理的に運ぶような組織でのリスクはさらに高くなります。この結果、DR サイトには、24 時間も経過した旧データが送られることになります。

このような状況は、多大なコストをかけて高速リンクを設定している組織にとっては、大きな負担となっています。ただし、組織にとってのコストは、重要なデータが失われた場合の犠牲に比べれば、はるかに少なくて済むわけです。

システムの保護

クラスタ化によって、追加コストを比較的低く抑えたまま標準的な市販（COTS：commercial, off-theshelf）のコンピュータ・システムを保護することができます。クラスタ化には、実装するためのソフトウェアとハードウェアが必要になるほか、同じアプリケーション・セットに対応できるコンピュータ・システムをグループ化する必要があります。アプリケーションに属するデータには、該当するアプリケーションに対応するクラスタ中の全システムからアクセスできなければなりません。

SCSI バスの場合、このデータ共有化には、HP 9000 サーバのシステムでサポートされるマルチ・イニシエータを接続する必要があります。マルチ・イニシエータを接続する場合、最大 7 つのアドレスのうち最も優先順位が高いアドレスを割り当てられるのは、バス上のホスト・アダプタだけであるため、ホスト・アダプタの SCSI ターゲット・アドレスを変更する機能が必要になります。また、バスのコントロールを別のイニシエータに受け渡すためにドライバの特殊サポートも必要になります。

図 1 は、システムのクラスタ化によってディスクのグループをどのように共有するかを示したものです。バス上の各ホスト・アダプタには高い優先順位と別の SCSI ターゲット・アドレスを割り当てる必要があるほか、特殊な V 型ケーブルを使用して、バス上の複数のホストを接続しなければなりません。

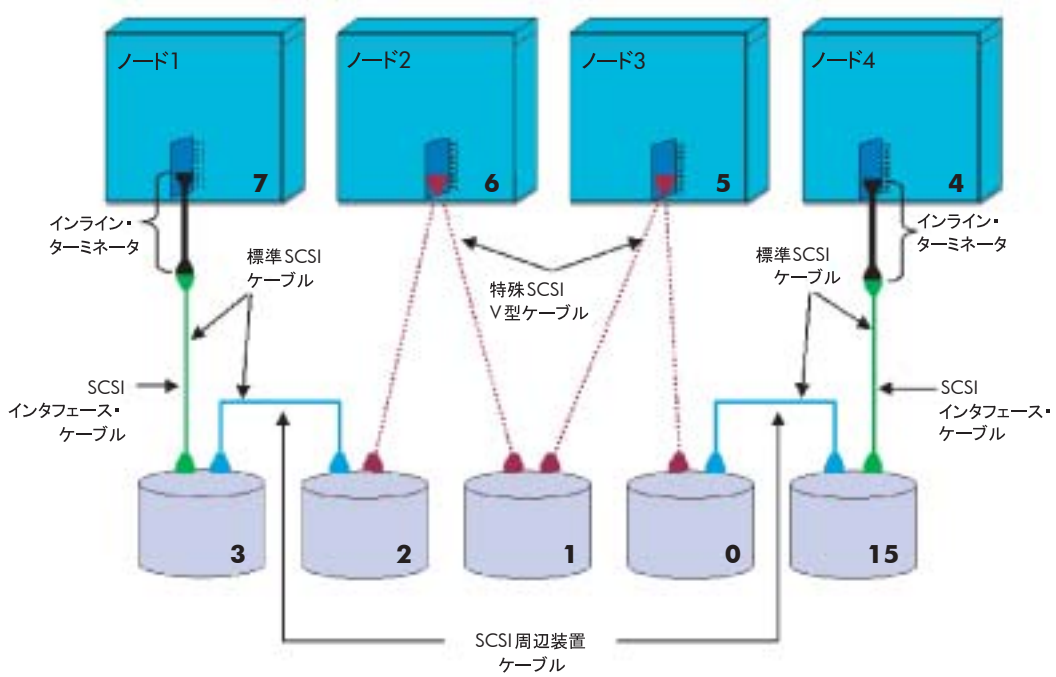

図 1：HA クラスタにおける SCSI マルチ・イニシエータ・バス

ソフトウェアのクラスタ化により、モニタとフェールオーバの機能が提供されます。HP-UX 用として、次のクラスタ製品が提供されています。

- MC/ServiceGuard
- ServiceGuard for RAC

MC/ServiceGuard

MC/ServiceGuard は、最大 16 システムで構成されたクラスタ中で複数のアプリケーションをサポートする第二世代の HA 製品です。クラスタ中の各システムがすべてアクティブなピアとなり、各システムがそれぞれ 1 つ以上のアプリケーションに対応します。

システムに障害が起きると、アプリケーションは事前に構成されたクラスタ中の別のシステムに自動的に移行されます。このアプリケーションに対応する IP（インターネット・プロトコル）アドレスも対応するノードに移行されるため、クラスタでは同じ IP 名と IP アドレスに接続することができます。この方法を"move the service point"（サービス・ポイント移行）パラダイムと言います（図 2 を参照してください）。対応するシステムはリブートされないため、パフォーマンスの理由上、止むを得ない場合を除き、該当するアプリケーションのサービスを停止する必要はありません。

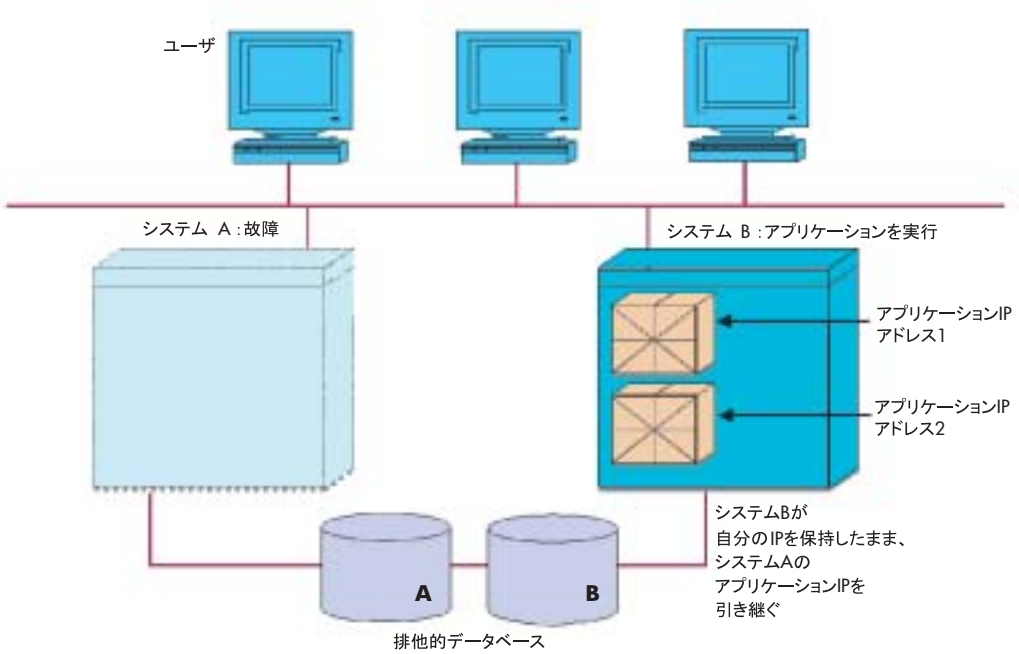

図 2：HA パラダイム： サービス・ポイントの移行（アクティブ/アクティブ）

複数のシステムから成るクラスタでは、故障したシステムを引き継ぐ正常なシステムも含めて複数のシステムに障害が発生することもあります。各ネットワーク・アダプタでは複数の IP アドレスに対応できるため、別のアプリケーションを停止させなくても、そのままフェールオーバを実行することができます。

MC/ServiceGuard ではクラスタ中の各ノードの動作状態だけに限らず、ネットワーク・ホスト・アダプタの状態や、冗長アダプタへのスイッチ（利用可能な場合）の状況についても監視できます。これについては、次に詳しく説明します。

アプリケーション自体の動作状態も監視できます。アプリケーションに障害が起きた場合は、アプリケーションはクラスタ中の別のシステムで再起動、または移行することができます。

MC/ServiceGuard では、アプリケーションの属するディスクへの排他的アクセスが保証されるため、同じディスク・バスに接続された別のシステムからの不当なアクセスは禁止されます。

ServiceGuard for RAC

ServiceGuard for RAC は、2 つのシステムで構成されるクラスタ中の特定のアプリケーション、すなわち Oracle 9i RAC をサポートする HA 製品です。各システムがアクティブ・ピアとして実行され、ユーザー向けの Oracle アプリケーションに対応します。あるシステムに障害が起きると、別のシステムがそのまま Oracle アプリケーションを続行します。故障したシステムに接続されていたユーザーは、正常に動作しているシステムに再接続されます。この方法を"multiple service point"（マルチ・サービス・ポイント）パラダイムと言います（図 3 を参照してください）。正常に動作しているシステムは、すでに実行状態になっているため、リブートの必要がないことに加え、アプリケーションを再起動する必要もありません。データベースは、故障したシステムが使用していたログに従って、自動的に復元されます。

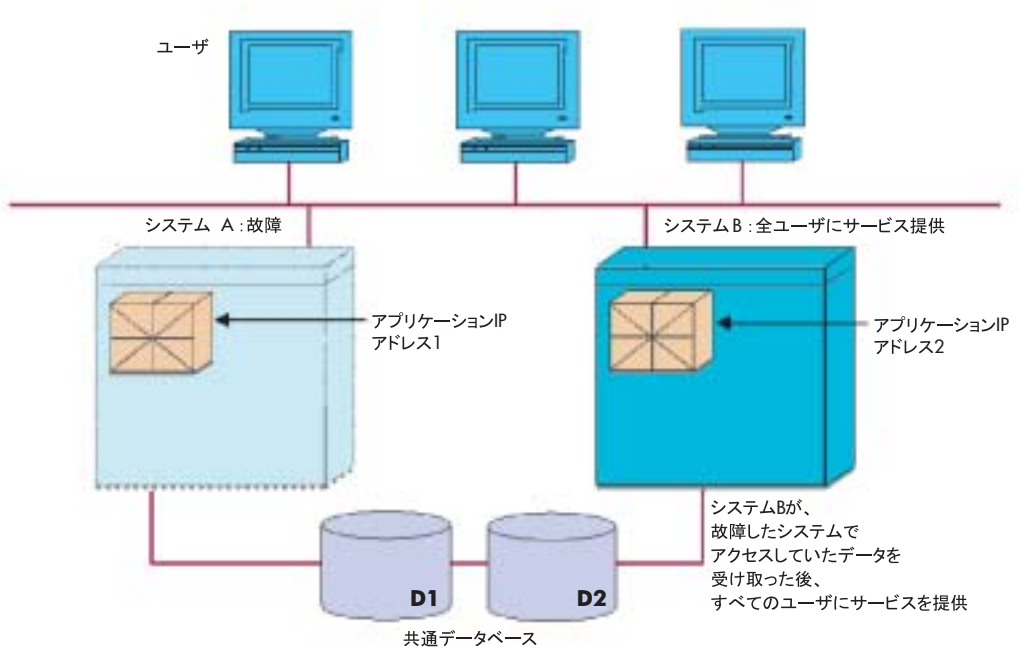

図 3：HA パラダイム： マルチ・サービス・ポイント

ServiceGuard for RAC には、MC/ServiceGuard の機能も組み込まれています。1 人のユーザーを別の IP アドレスではなく同じ IP アドレスに接続する方が望ましい場合は、故障したシステムの IP アドレスを正常なシステムに割り当てることにより、「サービス・ポイント移行」パラダイムを実現できます。ServiceGuard for RAC では、ネットワーク・ホスト・アダプタが故障した場合に備えて、冗長ネットワーク・インタフェースを使用することもできます。最後に、OPS の最上段にあるアプリケーション、または OPS からは独立したアプリケーションは、MC/ServiceGuard と同様に、パッケージとしてまとめることにより、クラスタ中の別のシステムに移行することができます。たとえば、OPS と連動するトランザクション処理モニタ（TPM）は、クラスタ中の別のシステムにフェールオーバできます。

クラスタ中の全システムで同じアプリケーションが実行されるため、ServiceGuard for RAC により、同じデータ・ディスクを対象とした読み書きアクセスを共有することができます。分散ロック・マネージャを通じて、システム間の書き込み処理が調整されるため、特定のシステムが同時に同じデータを対象に変更を加えることはできません。

ネットワーク・リンクの保護

HA ソリューションを実装すると、ネットワーク・リンクが無視されることがあります。
MC/ServiceGuard、または ServiceGuard for RAC のいずれかと、HP-UX 11i を組み合わせることにより、ハイアベイラビリティ対応のネットワーク・リンクを設定するための各機能が提供されます。

ネットワークをハイアベイラビリティ対応に移行するには、まず冗長構造のリンクを設定する必要があります。Ethernet、FDDI、および Token Ring の LAN リンクは、簡単にハイアベイラビリティに対応することができます。ハイアベイラビリティには、同じ種類の 2 つ目のホスト・アダプタをインストールし、これを一次ホスト・アダプタと同じ物理サブネットに接続する必要があります。物理媒体は、デュアル構造のメディアを実行し、2 つのメディアとハブ、またはブリッジを接続することにより、物理媒体を保護することができます。ハブやブリッジの SPOF 要因を除外するには、デュアル構造のハブ、またはブリッジを使用した構成を採用することができます。

MC/ServiceGuard と ServiceGuard for RAC では、各ネットワーク・リンクを定期的にテストします。リンクに障害が起きた場合に、スタンバイ・ホスト・アダプタが構成されていれば、スタンバイ・ホスト・アダプタが、故障したリンクを引き継ぎます。ローカル LAN フェールオーバと呼ばれるこのプロセスは、フェールオーバとは異なり、ユーザーへのネットワーク接続が切断されないため、優先的に使用されています。

現在、OSI（オープン・システム・インターコネクト）はクラスタのハートビート・ネットワークとしては使われていませんが、MC/ServiceGuard、または ServiceGuard for RAC によって冗長 OSI リンクのフェールオーバが可能になります。OSI ネットワークのユーザーは、ネットワークの障害から保護されます。

SNA（システム・ネットワーク体系）ネットワーク・リンクをハイアベイラビリティに対応させるのは、さらに難しい問題で、リンク・ソフトウェア自体の連携が不可欠になります。HP-UX SNA リンク製品では、冗長リンクがサポートされます。HP-UX SNA リンクに関するマニュアルでは、各リンクを MC/ServiceGuard クラスタに統合する方法について説明してあります。SNA リンクでは、メインフレームと HP 9000 サーバエンドの両方で特定の構成が必要になります。したがって、冗長 SNA リンクは、アイドル状態、または重要でないアプリケーションに対応した構成になっていなければなりません。

X.25 をハイアベイラビリティ対応に移行するには、システムに障害が起きると、X.25 の物理回路を別のシステムに切り換えることができる特殊なハードウェアが必要になります。

電源の保護

電源部分の SPOF 要因を排除するには、次のステップが必要になります。
- 電源の分離
- 無停電電源装置の構成
- 独立したラッキング

電源の分離

まず、次の重要なハードウェアごとに、電源サブパネルからは独立した主電源回線を確保する必要があります。
- クラスタ中の各システム
- ミラー・ディスク・セットの両側
- 複数の電源をサポートするディスク・アレイに対応した電源入力
- ネットワーク・コンポーネントの冗長化

このステップでは単に、電源ケーブル、または回路ブレーカの障害が保護されるだけであることは言うまでもありません。また重要な冗長コンポーネントごとに個別のサブパネルを使用すれば、保護機能をさらに強化できます。

無停電電源装置

次のステップでは、UPS（無停電電源装置）を使用します。UPS は、重要なデータを保護できるように、安全なシャットダウンを保証するのに十分な電源が供給されるように、適切なサイズに設定する必要があります。大型の UPS であれば、長期間にわたってシステムの動作を続行するための電源が確保されます。しかし、UPS はあくまでバッテリであるため、一定期間を超えてシステムに電源供給し続けることはできません。長期間に及ぶ電源障害から保護する対策として、ディーゼル・ジェネレータを使用することもできます。

ディーゼル・ジェネレータ自体には、災害から保護する能力はありません。災害が発生した場合に備えて、地理的に離れた場所にホットスタンバイ装置を確保しておく必要があります。こうした災害復旧サイトを設置するにあたり、さらに次の課題についても対応する必要が出てきます。
- 通信リンク
- データの複製
- 当該の災害によって、ユーザー自身が影響を受けることがあるか
- コスト
- コンピュータ・システムの冗長化

独立したラッキング

キャビネットは、電源に直接プラグを接続する部分であるため、重要な冗長コンポーネントがすべて同じキャビネットに収納されている限り、電源障害からの保護の点では効果はありません。たとえば、ミラー・ディスク・セットの各サイドを、それぞれ別のキャビネットに収納することができます。

デュアル電源をサポートするディスク・アレイを単独のキャビネットに収納する場合は、特殊なプランニングが必要になります。まずアレイを収納するキャビネットから 1 本、電源を確保します。2 本目の電源は、隣のキャビネットから供給します。これには、サイド・パネルを取り除くか、サイド・パネルに穴を開ける必要があります。

アプリケーションの保護

HA 環境でアプリケーションにアクセスする場合、その設計については慎重に検討する必要があります。HA アプリケーションを設計するにあたり、サービス停止やフェールオーバからユーザーを保護することが第一の目標となります。たとえば、クライアント/サーバのアプリケーションは、接続が切断された場合にも、クライアントからサーバに再接続できるような形式で設計する必要があります。
"move the service point"(サービス・ポイント移行)パラダイムでは、クライアントが、同じ IP 名、または IP アドレスに再接続されます。また"multiple service point"(マルチ・サービス・ポイント)の場合は、クライアント・セットで、IP 名、または IP アドレスから選択する必要があります。ユーザーではなく、クライアント・ソフトウェアで再接続に対応できなければなりません。

アプリケーションを保護する上で、次の点についても対応する必要があります。
- エラー処理の自動化
- TPM(トランザクション処理モニタ)の使用
- 再起動可能なトランザクションの設計
- システム固有の情報の回避

運用と管理

HA システム・クラスタの管理は、単一システムの場合よりも、さらに複雑な作業を伴います。一貫した形式で管理するマルチ・システムが存在すること以外に、システムやアプリケーションを管理するため新しいコマンドを学習しなければなりません。ユーザー側のエラーが原因となって、サービス停止が起きる率が増えているため、ユーザーや管理者が起こすエラーの可能性を低減するための対策が必要になります。

HA クラスタの管理作業に利用できる次のようなソフトウェア製品が提供されています。
- ServiceGuard Manager(Service Guard 製品をお持ちの場合、無償で提供)
- OpenView Operations

ServiceGuard Manager は、OpenView Network Node Manager や OpenView OperationsCenter と連動する OpenView アプリケーションです。これにより、HA クラスタ、クラスタ・ノード、およびアプリケーション・パッケージの各ステータスを監視する機能が提供されます。Network Node Manager は、クラスタ中のアプリケーション IP アドレスや、LAN ホスト・アダプタの各 IP アドレスを表示できるように、機能が強化されています。ノードとクラスタのサブマップにより、現在実行されているアプリケーション・パッケージの関係が表示されます。

OpenView OperationsCenter 自体は、管理上のミスを低減したり、ネットワーク接続されたコンピュータ・システム・グループで起きるエラーについてオペレータが対応できる機能を強化したりする上で役に立ちます。ClusterView と OpenView OperationsCenter を統合することにより、クラスタのエラー・メッセージ・フィルタ、詳細手順、その他特定のエラー状況から回復するためのオペレータの対応手順が組み込まれます。さらに問題の発生が管理者に自動的に通知され、原因究明の作業を始めることができます。

クラスタ全体を完全なエンティティとして管理することが重要になります。たとえば、クラスタ中の複数のシステム上で、あるアプリケーションを並列的に実行する場合、パスワード・ファイルには、各システムで一貫性のあるエントリが保証されなければなりません。OpenView AdminCenter では、管理者がクラスタ全体を一貫して管理できます。この結果、バックアップ・システムへのアプリケーションのフェールオーバができなくなったり、ダウンタイムが長期化したりする事態に至る前に、エラー発生の可能性を低減することができます。

結論

ハイアベイラビリティ・コンピューティング環境では、組織のアプリケーション環境で必要とされる機能が提供されます。HA 保護ソリューションはいずれも、アベイラビリティに関するニーズに基づいて評価する必要があります。まず、この段階を経てから、重要なアプリケーションへのユーザー・アクセスを保証するための適切なソリューションを実装します。

お問い合わせはカスタマー・インフォメーションセンターへ

03-5304-6660　月～金 9:00～19:00　土 10:00～18:00(日、祝祭日、年末年始および 5/1 を除く)

HP-UX 製品に関する情報はhttp://www.hp.com/jp/hpux

日本ヒューレット・パッカード株式会社
〒140-8641 東京都品川区東品川 2-2-24 天王洲セントラルタワー

invent

PDFHS03-021-01