Evergrid in High Performance Technical Computing

1. 概要

ソフトウェアとハードウェアがともに進歩してきたにもかかわらず、ハイパフォーマンステクニカルコンピューティング(HPTC)の世界では、依然としてアプリケーションダウンタイムがとても重要な事項、そして大きな重荷となっています。アプリケーションサービスの品質を管理し、システムの障害があってもクラスタ計算環境を継続する能力を手にすることは、HPTCシステムの進化のために非常に重要な項目です。本文書では今日の HPTC 分野でアプリケーションサービス品質を確保するための試みを概説し、Evergrid の製品が如何にそれらに対処していくのかを明らかにします。

本文書は以下の内容を含みます。
- 今日の HPTC 分野で行われている試み
- Evergrid 製品の概略
- アプリケーションの可用性確保
- アプリケーションサービス品質の管理
- 提供される機能
- Evergrid 社について
- まとめ
- 製品諸元

2. 今日の HPTC 分野で行われている試み

長年にわたり、HPTC 計算機インフラは、一連の根本的変化にさらされてきました。コンピューティング技術が進んだため、メインフレームはコモディティ製品を使ったクラスタに取って代わられました。今日の典型的な HPTC プロジェクトは、何千ものノードの上で何日もかかるほど巨大な並列計算が用いられています。

このクラスタベースのアプローチは利益をもたらす一方、計算資源の信頼性部分での基本的な弱さによって制限されてきました。今日の HPTC システムの多くでは、あるノードで障害が発生した場合、何のアウトプットも得られないまま、全ての計算の停止に至ります。これらの停止時間に伴うコストは、直接、潜在を問わず巨大です。停止期間は多大な計算時間の損失につながり、スタッフ時間を無駄し、プロジェクトは遅れ、そして、計算資源は無駄に浪費されることとなります。さらに重要なこととして、これらの停止期間は、障害が起こる前に解けたであろうモデルのサイズと細かさに制限を押しつけ、ひいてはお客様の研究対象そのものに対しても制限を強制するのです。

歴史的に、お客様がクラスタ環境において計算資源需要を迅速に予測する解決策はありませんでした。そのため、重要なアプリケーションは結局予定より遅れることになるか、全く失敗します。単により多くのハードウェアを加えること、そして、そのハードウェアを設置し、維持していくために必要な関連するスタッフを増員することは、内在する可用性の問題を解決しないばかりか、むしろしばしばそれを悪化させます。より多くの資源が加えられた結果、複雑さは増し、しかも信頼性は向上しないのです。あるノードでの停止期間は、依然としてすべての計算をダウンさせます。そのうえ、必要以上に準備されたサーバは、低い利用率、低い設定の柔軟性、過剰な床面積、過剰な電力、過剰な冷却能力、過剰なソフトウェアライセンス、および過剰人員からくる高いコストをもたらします。

これらの問題に対処するために、HPTC チームは、複雑なクラスタ計算機環境上で、より能率的にアプリケーションサービスの品質を管理する方法と、ハードウェア停止がアプリケーションに影響を及ぼさないことになる確実な手段を必要としています。

3. Evergrid ： 最適化と高可用性のためのソリューション

Evergrid は、クラスタ計算環境において、アプリケーションサービスの品質を管理するための基盤ソフトウェアを提供します。
Evergrid 社の Availability Management Suite を使えば、お客様は以下のようなことが実現できます。

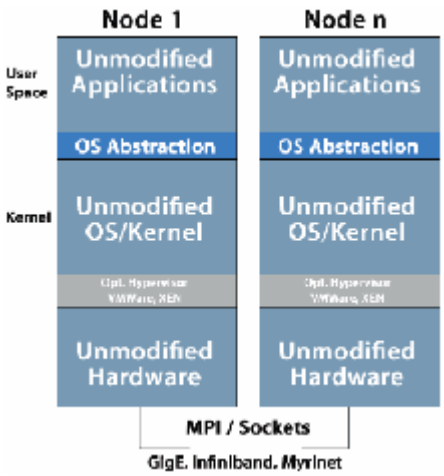
- アプリケーションの高可用性確保
  先進のオペレーティングシステム（OS）抽象化技術により、Evergrid はハードウェアまたはソフトウェアの障害に関わらず、アプリケーションの継続利用を確実にします。
- アプリケーションサービスの品質コントロール
  クラスタ化されたアプリケーション基盤を管理するための Evergrid の先進のコントロール機能で、お客様はリアルタイムに、停止時間とビジネス上の優先度の変化に対応することができます。

4. アプリケーションの高可用性確保

Evergrid 社は、クラスタ計算環境のための高性能 OS 抽象化技術に基づく高可用性を提供します。Evergrid 社製品により、お客様は、システムに障害が発生している場合にさえ、ビジネス優先度、サービスレベルアグリーメント（SLA）と計算ニーズを満足させることを確実とすることができます。

Evergrid 社のソリューションでは、以下のような機能を用いて高可用性を実現します。

- 洗練された OS 抽象化技術により、多層の、複数ノードにわたるアプリケーションリカバリ
- アプリケーションに対して手を加えることなく、いかなるアプリケーションの組み合わせ、システムの障害であっても、迅速なリカバリ
- ポリシーベースの、プリエンプティブ・スケジューリングにより、管理者は、率先してボトルネックと停止期間を除くことができます。
- 多層、並列、分散アプリケーションの信頼性と回復性を強化します。



業界初にして唯一の分散並列アプリ向け可用性向上技術

- チェックポイント
  メモリ、ファイル IO、ネットワークの状態を取得し、保存します。
- 首尾一貫性
  ノード間の状態や、個々のノードの状態を追跡します。
- 状態を保持したリカバリ
  どんな障害からもアプリケーションを迅速に復旧します。

アプリケーション透過性

- アプリケーション、OS、ハードウェアに一切修正を加える必要はありません。
- カーネルに組み込まれるソフトウェアではありません。ユーザレベルで動作します。

5%以下のパフォーマンス・オーバーヘッド

- アプリケーションに対しても OS に対してもノンブロッキング動作
- 障害からの迅速な復旧

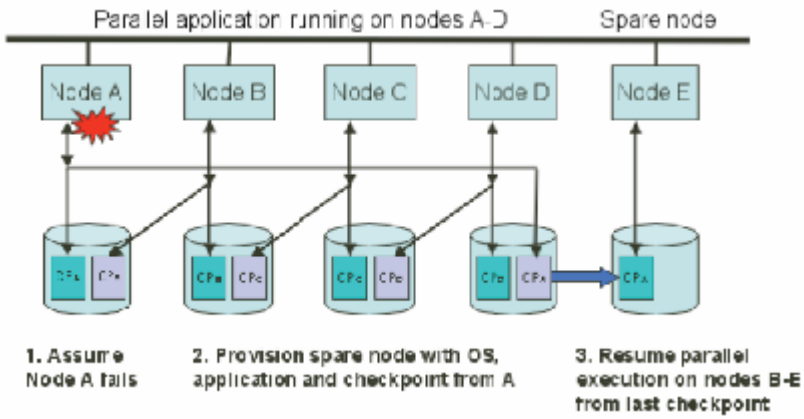
千ノードにおよぶ販売実績

洗練された、効率的なチェックポイント

Evergrid 社のソリューションでは分散チェックポイント技術を使用し、それにより分散環境においてノード、ネットワークとファイル I/O の状態を常に取得し続けます。このチェックポイント技術は、アプリケーションに透過的に機能し、さらに複数のノード間、層間で機能します。従って、ハードウェアまたはソフトウェアに障害が発生した場合、多層アプリケーションや分散並列アプリケーションでさえ、首尾一貫した状態に迅速に復帰するのです。Evergrid 社だけが、分散環境において、このレベルのアプリケーション高可用性を実現することができるのです。

過去には、チェックポイントを取得するために専用のコードを組み込む必要がありました。Evergrid 社のソフトウェアは、アプリケーションを修正することなく、アプリケーション、IO、アプリケーション間の通信を止めずに、どんなアプリケーションであっても、その状態を取得することができます。

分散並列アプリケーションチェックポイント取得: Evergrid 社のソリューションは、並列プロセス間通信の状態の同期のための、初めて真に中央集権化されていない手段です。この革新はコストのかかるバリア命令の必要を無くすだけでなくて、本当に非同期なチェックポイント取得を可能とします。もうチェックポイント取得を妨げるいかなるアプリケーション命令も存在しないのです。
非同期並列チェックポイント取得: Evergrid 社は、最適化され低遅延並列チェックポイントシステムを使います。それにより、アプリケーションへの影響を最小限にしたまま、チェックポイントデータの格納を同時に行うことができます。
インクリメンタル・チェックポイント: 最後のチェックポイント取得から変更のあったアプリケーションの状態だけを保存することで、パフォーマンスを下げるディスク書き込みを最小にします。



可用性とデータ保全性を確実に

ファイル I/O 動作のロールバック機能により、Evergrid 社のソリューションは、システム障害発生時もデータ保全性を確実にします。Evergrid 製品には、全てのファイルシステムコールをインターセプトし、ファイルシステムをロールバックするモジュールを含んでいます。リスタートプロトコルの一部として、最後のチェックポイントが起こった時から、ノードが変わったりファイルシステム階層に変化があった場合でも、ファイルシステム上で起こった全ての変化をロールバックします。その結果、障害が発生したときにアプリケーションプロセスがどこで稼動していても、Evergrid 社の製品はアプリケーションデータの完全性を確実にします。

5. アプリケーションサービス品質の管理

Evergrid はお客様に、クラスタ計算環境の完全な制御に入れ、資源利用率を最適化し、基盤コストを下げ、より負担になるサービスレベルアグリーメントを提供する、洗練された能力を提供します。Evergrid の技術は、ユーザ、アプリケーションと、しソース管理、全リソース、ビジネス優先度に基づいたノード間アプリケーション実行スケジュールを統合します。さらに、Evergrid で、お客様は、自動的にスケジュール優先度を確実にし、リソース不足になるアプリケーションをなくすことで、計算資源を最適化できます。

Evergrid 製品を用いることで、お客様は以下のようなことが可能となります。
- 要求に応じ、サーバを起動前、起動、電源停止の状態にできます。
- 要求に応じ、動的に、OS、バーチャルマシン、アプリケーション、サーバを準備します。
- アプリケーションや資源優先度をリアルタイムに変更するプリエンプティブ・スケジューリング

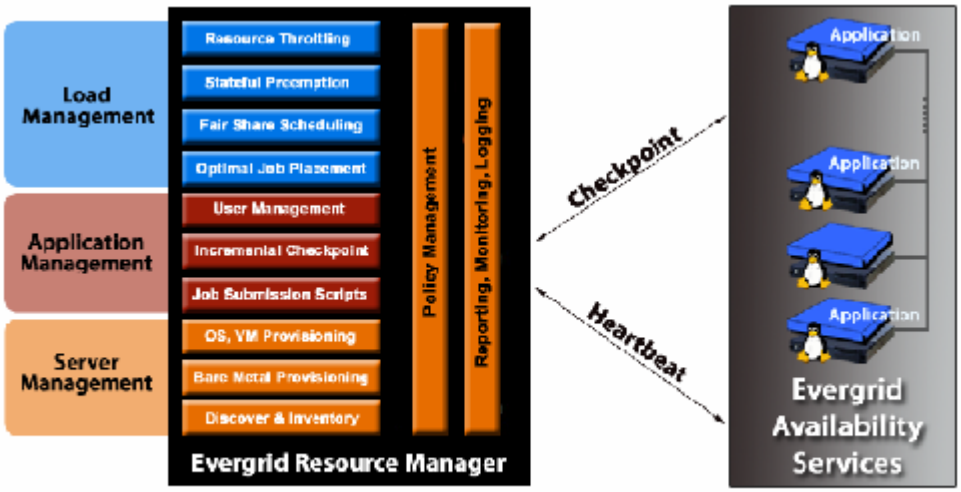
- 監視と課金システムのために、使用したすべての資源に対しての完全なアカウンティング情報を取得できます。

以下に、Evergrid 製品により提供される包括的でプロアクティブな管理機能に関する詳細を示します。

包括的な管理

リソースマネージャーは、クラスタ計算基盤の中であらゆるレイヤ上に包括的な管理を提供します。管理者に、特定ノードの電源投入から、負荷分散やスケジューリングまでの能力を提供します。Evergrid 製品は、これらの機能を提供します。

- Discovery
  発見段階で、Evergrid 社のソリューションは、利用できるすべてのリソースを見つけ出し、OS のイメージをダウンロードし、すべてのノードの属性を検出します。それにより、各々のシステムが効果的に管理されることができます。
- 初期状態でのプロビジョニング
  Evergrid 社のソリューションはベースハードウェアレベルでの管理を提供しますので、基盤の中、あるいは計算資源の各々の MAC アドレスを必要であるときまで、使用せずに残しておくことができます。そして動的に、起動、起動前、電源停止のそれぞれの状態に変更することができます。
- 仮想マシンのプロビジョニング
  Evergrid 社の製品は、XenSource と VMware のような既存の仮想化ソリューションと統合し、既存の仮想マシンのプロビジョニングをコントロールできます。
- OS のプロビジョニング
  Evergrid 製品で、お客様はオペレーティングシステム全体のプロビジョニングを管理することができます。これらの能力で、お客様は、プラットフォーム全体に統一されたコントロールを得る事ができ、ある特定の計算を、それにあったプラットフォームタイプ上に割り当てることを可能とします。
- アプリケーション管理
  Evergrid ソリューションで、アプリケーションは、重要なアプリケーションのための可用性を確実にし、またはリソース利用を最適化することにより、様々なプライオリティを与えるか、または都合のよい処理スケジュールかシステムを割り当てることができます。さらに、特定のユーザかグループに、より高いかより低い処理能力を割り当てることができます。
- 負荷の管理
  Evergrid の包括的で、包括的で中央集権化された管理能力は、管理者に対して OS、仮想計算機、アプリケーション、およびサーバをオンデマンドで、動的にプロビジョニングすることで個々の負荷に応じた完全な制御のための能力を与えます。



End-to-End optimization using Policy Based Automation

先を見越す管理

Evergrid ソリューションで、お客様は非常に洗練された方法で、クラスタ計算環境を管理することができます。Evergrid を使うとお客様はアプリケーションを、ノードをまたがって移動させることができます。それによりお客様は今以上に迅速に技術要件やビジネスの優先度の変更に対応することができるようになります。Evergrid の製品では以下のようなことが可能になります。

- リアルタイムでアプリケーションとリソース優先度を変える、先を見越すプリエンプティブ・スケジューリング
  例えば、低い優先度のアプリケーションから横取りし、より利用資源の少ない状況に追いやり、高い優先度のアプリケーションに多くのリソースを割り当てることができます。
- 複数のクラスタやシステム間での負荷分散
  これにより、最適化された可用性と利用率を確実にします。
- 重み付けされた割当
  これによりアプリケーションは、利用可能なメモリ、処理速度など、システム固有の測定基準に基づきリソースを割り当てることができます。
- フェアシェア・スケジューリングアルゴリズム
  速度、スケーラビリティ、およびリソース利用を改善します。
- リソース優先順位づけ
  優先されたアプリケーションが最適のサービスレベルを保証されます。
- 動的なパーティショニング
  ビジネスか技術的な要求に従って、クラスタあるいはサーバの特定のグループを区分して、プールすることができます。
- ソフトウェア・プロビジョニング
  必要な操作環境が必要に応じていつも利用可能であることを確実にします。

6. 提供される機能

上述の可用性や管理機能に加えて、Evergrid は HPTC 環境に特化した機能を有しています。インテリジェント・オートメーション、柔軟性の高い実装、最小のオーバーヘッドなどです。

インテリジェント・オートメーション

Evergrid のソリューションは、介入なしでアプリケーション障害かインフラストラクチャ障害を自己検知し、人手の介入なしで瞬時のアプリケーション移行することにより、管理インフラストラクチャを拡張し、自動化します。 結局、Evergrid ソリューションは、インフラストラクチャを、自己修復し、自己最適化し、自己に最適化して、自己回復する

インフラストラクチャとするのです。

フレキシブルな統合
Evergrid は展開における最大の柔軟性を管理者に与える機能を提供します:

- アプリケーション及び OS 透過性
  Evergrid は、アプリケーションバイナリやオペレーティングシステムカーネルに対して、なんら修正を必要としません。
- MPI と InfiniBand のサポート
  現在の実装では、MPICH の ADI 層で MPI 操作をインターセプトします。
- 移動可能性
  アプリケーションに含まれるすべての位置依存の状態を、Evergrid Availability Services を用いて、新しいノードに移動できるように、OS 層で抽象化します。

5%以下のパフォーマンス・オーバーヘッド

Evergrid のオペレーティングシステム抽象化技術は高能率的な方法でこれらの物凄い利益をもたらします。 通常 20～40%ものプロセッサパワーを必要とする伝統的な仮想化技術とは対照的に、Evergrid のオペレーティングシステム抽象化能力は非常に最適化されており、処理オーバーヘッドとして 5%未満しか消費しません。

実例で示すプリエンプティブ・スケジューリングのメリット

これから、Evergrid リソースマネージャが、お客様に、あるシナリオ中でどう重要な利益を提供するかに関する例を示します。100 ノードクラスのクラスタを保有するお客様は、比較的低い優先度のアプリケーションを少し実行しています。突然、優先度の高い、90 の利用可能なノードを必要とするアプリケーションが実行されます。以下に、どう解決するかの 2 つのシナリオを示します。1つ目はリソースマネージャ製品を使わない場合、2 番目はリソースマネージャ製品を使う場合です。

シナリオ 1:
リソースマネージャ製品がなければ、お客様は実行中の既存アプリケーションを停止するか(これは、計算時間の無駄、遅延、そして低い利用率につながります)、あるいは、資源が利用可能になるまで高優先度のジョブを保留にしておく(これは、非常にビジネス上重要なプロジェクトでの長い遅延、あるいはおそらく仕事の窮乏につながります)かの選択に直面しています。

シナリオ 2:
リソースマネージャ製品を使う場合、お客様は、プリエンプティブ・スケジューリング機能により、動的に高優先度アプリケーションに 90 ノードを割り当て、残っている利用可能な 10 ノードへの実行中の低優先度アプリケーションを再割り当てします。低優先度アプリケーションが先取りされた後に、それらは再スケジュールされ、高優先度アプリケーションが終了したときに、一旦停止した時点から再開されます。その結果、リソースマネージャ製品は、計算時間が全く無駄にならず、すべてのリソースが完全に利用され、そして、高い優先度アプリケーションは最適の性能を得ることを確実にします。

7. まとめ

まとめると、Evergrid はアプリケーションサービス品質を高め、アプリケーションの高可用性をもたらします。Evergrid はビジネスのプライオリティに基づき、アプリケーション間でノードをプールし、共有することが可能です。そのため、お客様は、リソースの利用率を極限まで高めることが可能となるのです。Evergrid を使うことで、お客様は、

- アプリケーションやリソースのプライオリティを、リアルタイムに変更できます。
- 重要なアプリケーションを優先的に実行するために、プリエンプティブ･スケジューリングが利用できます。
- システム負荷を、非破壊的にかつ即時に移動できます。
- 持っているリソースをフルに投入することで、過剰なプロビジョニングを防ぎます。
- ダウンタイムを増やすことなく、サービス可用性要求に取り組むことができます。

8. Evergrid 社について

Evergrid 社は、次世代のデータセンターにアプリケーション品質管理のインフラを提供し、並列または分散アプリケーションを、ハイパフォーマンスクラスタグリッド上で、ほぼ100%の稼動をもたらします。Evergrid 社のフォルトトレラント OS 抽象化ソフトウェアにより、ダウンタイムをなくし、自動的なチェックポイント取得、マイグレーション、アプリケーションの回復を行います。さらに千台クラスのノードでもスケールし、しかも 5%未満のオーバーヘッドしかないのです。
Evergrid 社の経営チームは、IBM、Amdahl、VERITAS、Motorola、Tandem Computers そしてバージニア工科大学で、広範囲のマネージメントと技術経験を積んでいます。Evergrid 社は私企業です。詳細は www.Evergrid.com を参照ください。

9. 製品諸元

実行環境	
他社製品との連携	チェックポイント/レジュームに対しての、他社スケジューラ製品との連携可能
オペレーティングシステム	
制御ノード	Red Hat Linux RHEL v.4.x SUSE 9.x、10.x
アプリケーションノード	Red Hat Linux RHEL v.4.x SUSE 9.x、10.x
インターコネクト	
ライブラリ	MPI、MPICH、MVAPICH などの並列プログラミングライブラリをサポート
物理的なインターコネクト	Infiniband Gigabit Ethernet Myrinet 2000　　(Myrinet は 07 年 2Q 予定)