

Linux による基幹システム構築

Backbone System Configuration Using Linux

熊谷康成 *
Yasunari Kumagaya

加藤幸重 *
Yukishige Katou

由留部稔 *
Minoru Yurube

* ソリューションビジネス本部 システム事業部 インフラシステム統括部 第一システム部

基幹系システムは、24H365D 運用と厳しい稼働条件を求められることがしばしばである。これら企業における基幹系システムでは、従来高信頼システム化が重要視され Solaris^{注1)} OS 使用の UNIX^{注2)} サーバでの構築が通常行われていた。最近では、IA サーバや LinuxOS の完成度向上とコストダウン化の流れで、基幹系システムを LinuxOS 使用の IA サーバで構築する事例も登場している。今回は Linux カーネルで構築する高信頼システムの構築技術について解説する。

24/7 operation and stringent operational conditions are common requirements for backbone systems. Historically, company backbone systems have been configured on UNIX servers using Solaris OS because priority was placed on higher system reliability. In recent years, however, some backbone systems have been configured on IA servers using Linux OS following the increasing refinements of IA servers and Linux OS, and the trend toward cost reduction. The technique explained here is used for configuring higher reliability systems using Linux kernels.

1 まえがき

高信頼システムを構築する場合、可用性を高める事が必要となりプラットフォーム、OS、ミドルウェアと多角的に採用する上で検討する必要がある。どの要素が欠けても可用性を低める結果となってしまう。これら複数の要素からなる高信頼システムを構築する技術について Solaris OS 使用の UNIX サーバ(以降、Solaris 機)、Linux OS 使用の IA サーバ(以降、Linux 機)を取り上げ解説する。

2 背景とねらい

高信頼システムとはどのようなシステムかと問われて、連想するキーワードは何でしょうか。ハード・ソフト技術関連の用語を思い浮かべることでしょう。このような技術が基盤となり高信頼システムは構築されている

が、システムを導入するだけでは高信頼システムとして稼働させることはできない。運用まで考慮した企画、設計、構築、運用と一連の流れを踏まえ検討し採用していくことが必要である。本編ではこの一連の流れの中で設計、構築のフェーズを取り上げ、オールインワン富士通製品で構築するプラットフォーム別高信頼システムに視点を絞り解説していくこととする。

2.1 高信頼システム

2.1.1 Solaris 機での高信頼システム

Solaris 機で高信頼システムを構築するポイントとして、現状、ハード・ソフトの選択肢も複数存在し選択するのが難しいのが現状である。ここでは表 - 1 のハードウェア・ソフトウェアを前提にそれぞれの項について説明する。

注 1) Sun, Sun Microsystems, Sun ロゴ, Solaris およびすべての Solaris に関連する商標及びロゴは、米国およびその他の国における米国 Sun Microsystems, Inc.の商標または登録商標である。

注 2) UNIX は、米国およびその他の国におけるオープン・グループの登録商標である。

表 - 1 Solaris 機のハードウェア・ソフトウェア構成

区分	製品名	概要
ハードウェア	PRIMEPOWER	UNIX サーバ
	ETERNUS	ストレージシステム
OS	Solaris8	UNIX 系 OS
ソフトウェア	SafeDISK 注1)	ディスク装置のデータをコピーして持ち、可用性を高めるソフトウェア
	SafeLINK 注1)	複数の NIC (Network Interface Card) を使用して自システムが接続されるネットワーク伝送路を冗長化し、通信全体の高信頼化を実現するソフトウェア
	SafeCLUSTER 注1)	システムを二重化し、障害時に他方のサーバに業務を引き継ぎ業務継続を可能としたり、負荷分散を行えるソフトウェア
	SafeFILE 注1)	複数のディスク装置を一つのファイルシステムとして使用できる UNIX ファイルシステム

注 1) SafeDISK, SafeLINK, SafeCLUSTER, SafeFILE は富士通株式会社の商標である。

(1) 可用性の確保

サーバ単体の可用性を高める事がサーバをダウンさせない為の第一条件である。ハードの冗長化を行う上での主なポイントについて Solaris 機では以下のとおりである。

1) ハードの冗長

① MPU

MPU の選択は、サーバの処理能力の観点から語られることが多いが、高信頼システムの場合、MPU の選択基準としてマルチプロセッサが望ましい、最悪の場合、シングルプロセッサでは MPU の障害が即サーバダウンに繋がってしまいダウンタイムを長引かせる原因となってしまう。ただし、MPU としてマルチプロセッサを選択した場合、PRIMEPOWER 機は通常、MPU 障害時に動的に MPU の切り離しができない為、縮退(サーバの再起が必要)運転が必要である。

② メインメモリ

メインメモリも集積化が進み大容量化しているが、大容量のメモリだけでシステムを構築すると障害が発生した場合に障害メモリ(バンク単位)がシステムから切り離され、メモリ資源が大幅に減少し

サーバの能力が低下する。サーバ機のメモリバンク数にも依存するが、メモリ障害を考慮した実装を行うことが必要である。

③ ディスク

ディスクの冗長化は、データ保護の観点からまず検討に上がる項目である。単体のディスクでは障害が発生した場合、即サーバダウンに陥ると共に、ディスク交換後使用することが出来ず、システムの復元が必要である。このような事態を避ける為にも、ディスクの冗長化は検討に値する事項である。しかし、RAID の選択も複数の選択肢がありそれぞれ、一長一短で選択に迷うところなのも確かである。今回は SafeDISK で構築する RAID を取り上げる。SafeDISK は RAID1 のディスクミラーリングのみサポートしている RAID ソフトでディスクの単体障害時に性能劣化を起こすことなく業務を継続できる。しかし、片系運用時にもう一本のディスクに障害が発生した場合は、システムディスクでは、サーバダウンが発生し、データディスクでは業務停止が発生する。そこで、さらに可用性を高めるよう、同期を取って待機運転させるホットスペアのディスクを用意する機能が準備されている。ディスク障害時に自動的にディスクを補完し RAID 構成を自動構成し可用性を高める。ディスクの交換はサーバを停止せずに活性交換が可能であるが、ディスクの交換時、ソフトウェアで RAID 構成を実現しているソフトレイドの為、RAID 構成へのディスク組込み作業に SE が関与する必要がある。

④ ネットワーク

サーバが接続されるネットワークが一回線の場合、ネットワークカードの故障により即サーバは業務継続が出来ない状態に陥る。このようなネットワークカードの故障に対応する為、ネットワークの冗長化を検討する必要があるが発生するわけである。SafeLINK を導入する事により伝送経路を二重化し、耐障害性や可用性に優れた信頼性の高いネットワークを構築することが可能である。

また、接続先のネットワーク機器を冗長化することにより、さらに可用性の高いネットワークの構築が可能である。

⑤ 電源

商用電源の品質も向上し、事故等が起こらなければ長時間の停電等あまり聞かれないが、瞬停は地理的、気象条件により起こりえる。また、電源ユニット

ト、商用電源の障害は、即サーバのダウンに繋がり、しかも、正常なサーバ停止もできない。電源に関しても冗長化することにより電源ユニットの冗長化及び、電源ケーブル自体も冗長化される。無停電電源装置 UPS によるサーバの保護も含め検討し導入が必要である。

2) ストレージ

データの信頼性やデータ量、サーバ間でのデータ共有、障害時のダウンリカバリ時間を考えた場合、サーバは搭載できるディスク量に限りがある。このような場合にストレージの導入を行う必要が出てくる。ストレージシステム ETERNUS は複数サーバからの共有や多彩な RAID 構成、ディスク内でのバックアップ等の多彩な機能を有し、異なった OS 間での共用も可能である。クラスタリングソフトウェアとの連携にも対応しデータを保全する。また、ディスクの障害に備えホットスペア機能も搭載されている。

3) クラスタ化

サーバ単体の可用性を高めることが、高信頼システムの必須条件であるが、サーバ単体の可用性だけでは業務を引き継ぐサーバが無いなどの理由から、高信頼システムとして限界がある。そこで、さらに信頼性を高める方法として、クラスタシステム等の検討、導入が必要になってくる。クラスタシステムとは複数のサーバを 1 台のサーバとして扱い、どれかのサーバがダウンした場合に、残りのサーバがダウンしたサーバの処理を引き継いで処理継続を可能としたり、特定のサーバに負荷が過大にかかる場合に、他のサーバに負荷を分散させ、安定した性能を実現する技術である。可用性を求め、しかも拡張性を求める場合に選択する。クラスタシステムの SafeCLUSTER は運用形態も多彩で、スタンバイ型、スケラブル型、複合型に分類される。例えば、アプリケーションサーバやデータベースサーバは複数台の内、固定しない 1 台を障害や負荷分散に備えて待機させるスタンバイ型の N : 1 移動待機型にした場合、処理能力の増大時にサーバを増設することにより必要な能力が確保できる。この方式ではフェールオーバーが発生した場合も待機サーバに業務が引き継がれ処理能力が保証される。ただし、複数のサーバで待機サーバを共有する為、障害サーバの復旧までの間待機サーバが存在しない運用時間が発生する。これらの点については、先に述べたサーバの可

用性を高める個々の冗長化や、保守の 24H365D の導入で対応する必要がある。

4) 監視システム

導入後の運用では障害の検知・対応に監視システムの導入は有効で、障害の検知をリアルタイムで監視し、フェールオーバーに至る以前に適切な対応が可能である。ただし、監視システムだけでは運用できず、監視システムを使った障害報知・障害対応策等を考慮したシステムとして構築しないと旨く運用することが出来ないため、運用設計を十分行い構築、導入することが肝要である。

2.1.2 Linux 機で構築する高信頼システム

Linux 機で高信頼システムを構築するポイントとして、ハードとして I/A サーバ及びサーバ機上で稼動する Linux カーネル (RedHat^{注3)}) と高信頼システムを実現するソフトウェア群が整えられて来おり Solaris 機と同等の信頼性が確保されつつある。本項では、表 - 2 に示すハードウェア・ソフトウェアを前提にそれぞれの項目の信頼性を高めるための主なポイントについて述

表 - 2 Linux 機のハードウェア・ソフトウェア構成

区分	製品名	概要
ハードウェア	PRIMERGY (32 ビット)	IA サーバ
	ETERNUS	ストレージシステム
OS	RedHat Enterprise Linux ES V2.1	Linux 系 OS
ソフトウェア	PrimeCLUSTER GDS	ディスク装置のデータをコピーして持ち、可用性を高めるソフトウェア
	PrimeCLUSTER GLS	複数の NIC を使用して、自システムが接続されるネットワーク伝送路を冗長化し、通信全体の高信頼化を実現するソフトウェア
	PrimeCLUSTER HA Server	HA (切替え) クラスタ機能、ポリシー管理機能、共用ファイルシステム機能、ネットワーク多重化機能をセットにした、切替え型クラスタシステムの基盤ソフトウェア
	PrimeCLUSTER GFS	オンラインサイズ拡張、高速リカバリ、連続ブロック割り当て機能等を備えるファイルシステム

注 3) Red Hat は米国その他の国で Red Hat Inc. の登録商標又は商標である。

べる。

(1) 可用性の確保

1) ハードの冗長

サーバ単体の可用性を高める事がサーバをダウンさせない為の第一条件である。ハードの冗長化を行う上での主なポイントについて Linux 機では以下のようなになる。

① MPU

詳細は Solaris 機と同一の為割愛し、異なる点について説明する。MPU は 1CPU を論理的に分割するハイパースレッディング機能を備え、CPU の効率的な使用にも対応している。ただし、本機能を使用した場合は、論理分割した CPU 単位でのライセンス購入が必要となり、本サーバで稼働させるソフトウェア費用を増大させる要因ともなるので、導入には注意が必要である。また、本機能はアプリケーション、ソフトウェアで個々にサポートしているか確認が必要である。

② メインメモリ

詳細は Solaris 機と同一の為割愛し、異なる点について説明する。PRIMERGY ではメモリエラーによるサーバダウンを回避するためスเปアメモリを搭載、割付け可能である。障害が発生したメモリをシステムから切り離すには縮退(サーバの再起動)が必要であるが、スเปアメモリを搭載することにより縮退を回避できる。ただし、スเปアメモリは、通常のサーバ稼働時にはメモリとして認識及び使用できないので、通常の業務で使用するメモリ量を考慮し、割付可能か判断する必要がある。

③ ディスク

ディスクの冗長化は、Solaris 機と異なりハードウェアで RAID 構成を実現するハードレイドとなる。RAID 構成の選択も多彩で、RAID0/1/5/0+1 から選択できる。購入時の標準構成では RAID5 となっており、購入時点で既に冗長化構成となっている。しかし、RAID5 の場合は、ディスク障害が発生すると、パリティチェックで性能が劣化してしまうので、障害時の性能劣化の回避及び、可用性を高める施策として、ホットスペアのディスクを搭載する機能が準備されている。ディスク障害時に自動的にディスクを補完し RAID 構成を自動構成し可用性を高める。ディスク交換はサーバを停止することなく活性交換が可能で、障害時のディスク交換及び、RAID へのディスク組込み等の作業は SE を

介さずハードの保守要員だけで作業が可能である。また、サーバのシステム退避・復元用にバックアップ装置を内蔵出来るが、ディスクベイ及び SCSI ケーブルを共有する為、ディスク 2 台分のリソースを占有する。採用する上では、データ量及びディスク本数を考慮する必要がある。

④ ネットワーク

ネットワークの伝送経路を二重化するソフトウェアは PrimeCLUSTER GLS を使用する。以降、伝送経路を二重化する機能についての詳細は Solaris 機と同一の為割愛。

⑤ 電源

電源の冗長化に関する機能の詳細は Solaris 機と同一の為割愛。

2) ストレージ

ストレージの機能に関する詳細は Solaris 機と同一の為割愛。

3) クラスタ化

サーバが PRIMERGY でしかも OS が RedHat ではクラスタシステム構築に PrimeCLUSTER のみ選択可能である。クラスタソフトの各機能についての詳細は Solaris 機と同一の為割愛。

4) 監視システム

障害の検知・対応に監視システムを導入することにより障害の検知をリアルタイムで監視し、フェールオーバーに至る以前に適切な対応が可能であるが、監視システムを構築しない場合でも、サポート部門へハード障害情報を通知し、障害箇所の特定やハードの交換要員の派遣要求を自動的に行うリモート通報機能が用意されている。本サービスを利用するには、サーバにリモートサービスボード(RSB)の搭載が必須で、保守契約の無償オプションとなっている。サポート部署への通知方法も選択可能で、PtoP 及び、SMTP サーバ、Internet 経由での選択も可能である。あくまでもハードに限定したサービスなので、リソース枯渇や、ソフトウェアが原因となるサーバ障害には対応できないので、これらへの対応が、個別に必要となる。

2.2 Linux 機で構築する場合の注意点

ここまで読み進んで頂けた方は、Linux 機でも Solaris 機と同等の高信頼システムが簡単に構築出来るのではないかと思われたのではないのでしょうか。しかし、筆者らのこれまでの Linux 機の構築経験を振り返

ると各フェーズでいろいろな取捨選択を行いシステム構築まで漕ぎつけたと言うのが本音である。本節では、Linux 機での高信頼システムを構築する場面場面で遭遇した事象及び、対応について述べることにする。

(1) システム構成

今回構築した高信頼システムのシステム概要を図 - 1 に示す。IA サーバは PRIMERGY RX300 を使用した。

(2) 冗長化の取捨選択

サーバを選択するにあたり、必要なオプションカードの設計を行った所、冗長化に必要なカード数を満たす搭載ができない為、機種種のグレードアップで対応するか、一部の冗長化を諦めるかの選択を迫られた。冗長化の取捨選択結果を表 - 3 に示す。

結論として、機種種のグレードアップは、冗長化の条件を満たせるが、大幅な費用アップとなり断念し、業務 LAN の伝送経路を単線とした場合は、ネットワークカードや経路の障害で、即業務停止に陥る為、機種は変更せずにクラスタ機の系間通信をオンボード LAN ポートで単線とする案を選択した。

(3) OS バージョンの選定

今回、OS は、Red Hat Enterprise Linux ES V2.1 を選択した。選定理由としてまず第一に、DiskDump 機能が実装されていることがある。既に V3.0 が発売されているが、NetDump しか対応していない為、耐障害用に専用の NetDump サーバ（共有可）を用意し、しかも、NetDump サーバ用ネットワークをサーバ間に専用を用意する必要がある。サーバの

冗長化の観点から、これ以上のオプションカードを追加できないことと、サーバの調達費用まで膨らむことを考慮し V3.0 の導入は断念した。

(4) ミドルウェアのバージョンの選定

OS が決定され、次に問題として浮上したのは、ミドルウェアの各バージョンと、Linux カーネルのサポート版数との整合性をとり、調達できるバージョンに絞り込むことであった。相互の関連については、表 - 4 に纏めている。結論として、今回のシステムでは、監視関連のミドルウェアの購入が出来ず、監視を断念することになった。ただし、先に監視システムとして記載しているとおりで、ハードウェアの監視については、リモート通報機能を採用することにした。

表 - 3 冗長化取捨選択一覧

区分	種別	DBサーバ	Web・連携サーバ
オプション	業務 LAN	1	1
	ファイバチャネルカード	2	2
	RAID カード	1	1
	リモートサービスボード	1	1
	オプション搭載最大数	5	5
オンボード	LAN1 (業務 LAN)	1	1
	LAN2 (系間通信)	1	1

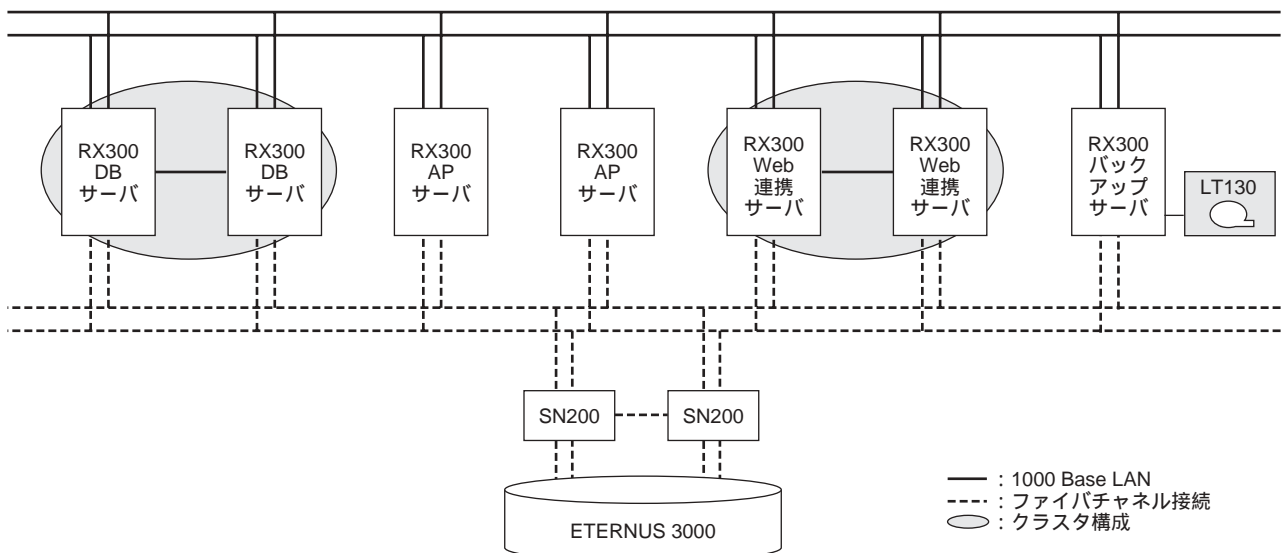


図 1 システム構成図
(Fig.1-System configuration)

表 - 4 カーネル版数とミドルウェアの選定

OS	カーネル版数	ミドルウェア	選定バージョン
Red Hat Enterprise Linux ES V2.1	2.4.9-e.12	Interstage Application Server Enterprise Edition	V5.0L20
		PrimeCLUSTER HA Server	V4.1A20
		Symfoware Server Enterprise Edition	V6.0
		Systemwalker 群	
		Systemwalker Centric Manager Enterprise Edition	
		Systemwalker Service Quality Coordinator Enterprise Edition	
		SystemWalker/CentricMGR-A EE	

3 むすび

現在，IA チップの圧倒的な性能向上を考えると，今後ますます Linux 機での高信頼システムの需要が高まると考えられる．構築事例もさらに増加し，技術の蓄積と，より良い構築への技術の反映と頑張っ

ばいけない．その為にも，Linux 機を安定的に構築，稼働させる為，富士通のミドルウェア群の Linux カーネルへの追従のスピードアップ，製品ラインナップの平準化や市場で稼働しているミドルウェアの継続サポートまで考慮した製品の市場投入を期待してやまない．