High Availability Cluster System using Linux-PC

Shiro Kusano[1,A)], Takuya Kudou[A)], Kazuro Furukawa[B)], Masanori Satoh[B)]
[A)] Mitsubishi Electric System & Service Co.,Ltd.
2-8-8 Umezono, Tsukuba, Ibaraki, 305-0045
[B)] High Energy Accelerator Research Organization (KEK)
1-1 Oho, Tsukuba, Ibaraki, 305-0801

Abstract

The KEK electron / positron linac provides its beams to four rings. As its operation time exceeds 7000 hours per year, recently, the stable operation is very important. The availability of the control system is essential. In order to achieve high availability, two kinds of cluster computer systems based on the Linux-PC technology was introduced and evaluated.

Linux-PCを用いた高可用性クラスタシステム

１．はじめに

KEK Linacでは、4つの異なるRingにビームを供給しており、年間の総運転時間は7000時間を越えている。2004年以降はKEKBへの入射が、連続入射となりBeam Mode Switchプログラムによるビームパラメータの切り替えは1日約300回にもなる。安定したビームを供給するためには、制御システムの高信頼性、長期安定運用は重要である。 KEK Linacでは、1993年にUnix計算機をベースとしたシステムに変更し、その後毎年変更、改善を行っている。安定運用のため、耐障害性、可用性が高いとされるRAIDディスクやクラスタサーバ（Compaq Tru64 Cluster）の導入を行った。

これまでのクラスタシステムには、専用計算機や専用のハードウェアが必要とされ、しかも高価であった。また運用管理、保守においても専門の知識、技術が必要とされる。一方、PCは近年高性能かつ低価格になり、またLinuxの普及によりLinux-PCを用いたクラスタシステムが注目されつつある。

KEK Linacでは、2003年にWebサーバ用、2004年にEPICS Channel Archiver用としてLinux-PCを用いたクラスタシステムの導入を行った。本稿では、Linux-PCクラスタの概要、導入、今後の改善点について報告する。

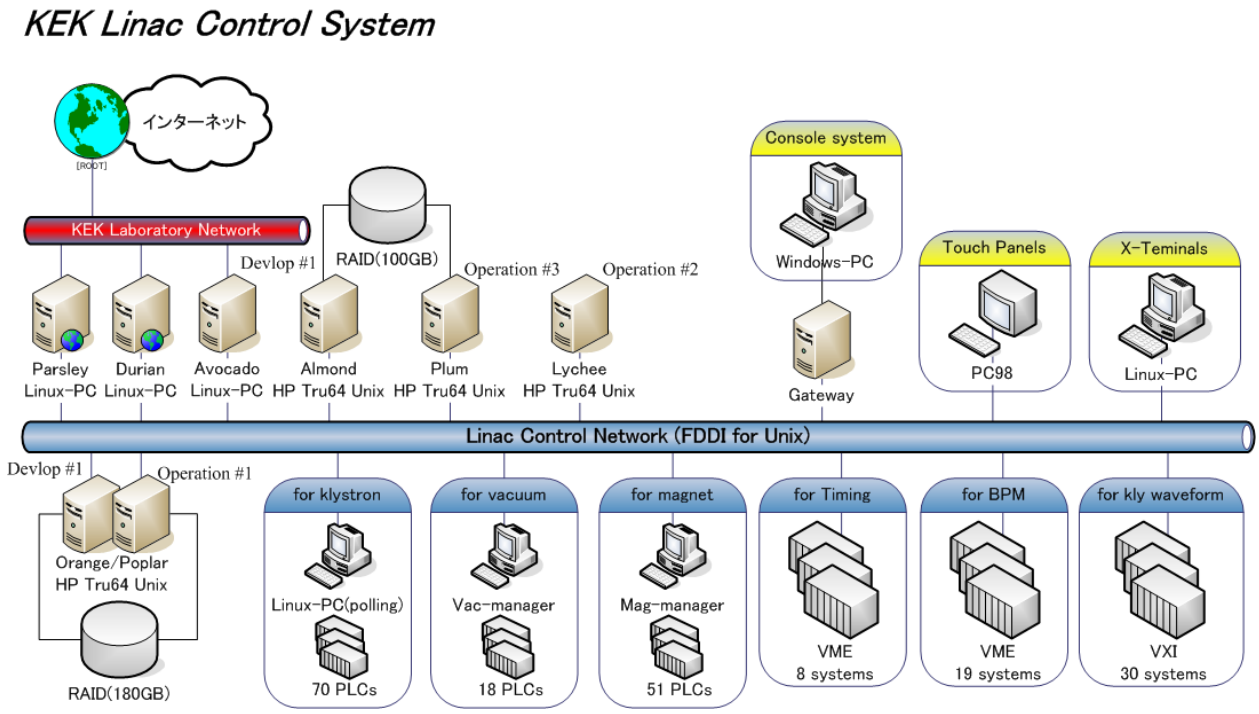

図1.KEK Linac Control System

[1] E-mail: skusa@post.kek.jp

２．KEK Linac 制御システムの構成

KEK Linac制御システムの構成を図1に示す。KEK Linacの制御システムは、Unix計算機（運転用3台、開発用2台）によるサーバ部と多様なフロントエンド（VME：約30台、PLC：約140台、CAMAC：約10台、等）による機器制御部とコンソールシステム、タッチパネル、X-Terminal（Linux-PC：約3台、Windows-PC：8台)によるオペレータインターフェース部の3階層の構成になっている。加速器の制御をするために、制御機器ごとにデバイスサーバが用意されており、クライアントとサーバ間はTCPやUDPを元にした独自開発のRPC（Remote Procedure Call）によって接続されている。

３．クラスタシステムの概要

クラスタシステムは、複数台の計算機を相互に接続し、ユーザーや他の計算機に対して全体で１つのシステムとして運用する技術である。クラスタを方式別に分類すると、HPC（High Performance Computing）クラスタとHA（High Availability）クラスタに分類される。HPCクラスタは、複数台の計算機を束ねて高速な並列計算を目的とした環境を提供するクラスタで、大規模な数値計算に向いている。HAクラスタは、計算機を複数台使用し冗長化することにより、システムの停止時間を最小限に抑え、可用性を向上させる。さらにHAクラスタ方式は、フェイルオーバー型と負荷分散型に分けることができる。以下に、本稿の対象となるHAクラスタについて簡単に説明する。

3.1 フェイルオーバー型クラスタ

フェイルオーバー型クラスタには、データミラー型、共有ディスク型及び遠隔クラスタ型がある。標準的な共有ディスク型クラスタの構成を図2に示す。障害が発生した計算機上にあるサービスを他の計算機に引き継ぐ（フェイルオーバー）ことにより高可用性を実現する。サービスのフェイルオーバーを実現させるためには、計算機間でデータを引き継ぐ必要あるが、引き継ぐデータを計算機間で共有するディスクに置く方式を共有ディスク型、計算機間でデータをコピーする方式をデータミラー型という。共有ディスク型はデータの同期を取る必要がないため、大量のデータを取り扱うシステムに向いている。その反面、共有ディスクを構成するために複数のチャンネル接続に対応したSCSIディスク、Fiber Channelディスクを必要とし高価なハードウェア構成となる。一方、データミラー型は計算機間でデータのコピーを行うことでデータの整合性を保つ。一定間隔でデータの同期を行うが、そのつど計算機間で通信が発生するため大量のデータを取り扱うシステムには不向きである。しかし、共有ディスクを持たないため、比較的低価格なハードウェア構成で実現することが可能である。

3.2 負荷分散型クラスタ

負荷分散型クラスタには、ロードバランス型、並列分散型に分類され、主に負荷分散を目的としたHAクラスタである。必要に応じて計算機へのアクセスを制御することにより、システム全体で可用性や応答性を向上させている。

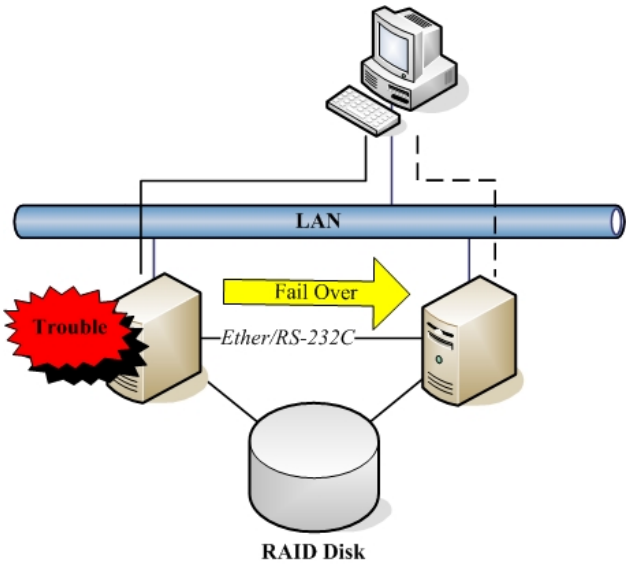

図2．標準的なHAクラスタシステム

４．Linux-PCクラスタシステム

4.1 所外Web用クラスタシステム

世界的にネットワーク環境は発達し、日常生活においても計算機の利用は必要不可欠になっている。その反面、ネットワークによる犯罪も多くなってきている。KEKでもセキュリティ対策として、2003年に所内ネットワークにDMZ（DeMilitarized Zone）が導入された。KEK Linacでは、Web上に加速器の状態や運転情報を提供[1]しているが、DMZが導入されたため外部からの運転情報などの閲覧が困難となった。その対応策として所外Webサーバ用にLinux-PCを用いることになった。運転情報など多くのファイルを扱うことやできるだけサービスを停止しないことを考慮して、Linux-PCを用いたクラスタの導入を行うことになった。システムの構成を表1に示す。

表1．所外Webサーバクラスタシステムの構成

	メーカー	理由
PC	DELL Precision 340(Pentium4, 2.8GHz)	安価で一般的なPCを選択
RAIDディスク	IAI AE-5600EX3-5400	複数Channelを持つIDE RAIDディスク
クラスタソフト	Redhat Advanced Server 2.1(kernel 2.4.9-e.43)	

所外Web用クラスタは2003年夏に導入し約2年間、KEK Linacの運転状況などの情報を提供している。

4.2 EPICS Channel Archiver用クラスタシステム

KEK Linacでは、機器の状態を監視するためにWebなどによる履歴表示ツールを提供している[3]。履歴表示ツールの1つであるEPICS Channel Archiverは、BPM, Vacuum, Klystronなどの加速器装置の情報を約1秒周期で履歴を蓄積している。現在のEPICS Channel Archiver用システムはLinux-PCを用い、収集したデータはローカルのディスクに保存している。データファイルは日付毎に作成され、ファイルサイズはKlystronで50MB、BPMで300MB、Vacuumで40MBとなっている。これまでも1年に2、3回Linux-PCが原因不明で停止することがあり、安定した運用、サービスの提供ができていないことがあった。この対策として、Linux-PCを用いたクラスタの導入を検討している。システムの構成を表2に示す。

表2．EPICS Channel Archiver用クラスタの構成

	メーカー	理由
PC	DELL PowerEdge 600SC(Pentiuam4, 2.8GHz)	安価で一般的なPC
RAIDディスク	IAI SNX-R740000D-2300XS	複数のChannelを持ち、信頼性の高いSCSIディスクを選択
クラスタソフト	RoseHA[2]	設定、管理用ツールが提供されており、設定管理が容易（図3（

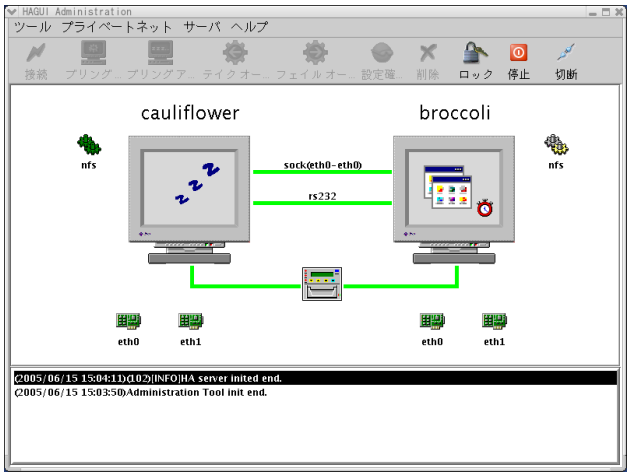

図3.RoseHA管理ツール

クラスタソフトでRoseHAを選んだ理由は、評価の際運用が容易と判断。また、Redhatがヴァージョン更新の際にソフトウェアの構成も変更したため、使用が困難となっていた。

[2] http://www.roseha.com

５．今後の課題

5.1 所外Webサーバ用クラスタ

・ フェイルオーバーの失敗

システム導入後、1年に1、2回原因不明で計算機が停止することがあった。計算機が停止するとバックアップの計算機にサービスが移行してサービスを継続するのがクラスタシステムの特徴であるが、フェイルオーバーが正しく動作しないことがあった。この問題は、クラスタシステムにKEK Linac独自のスクリプトを追記しており、その記述が正しくなかったことが原因であった。スクリプトを改修することで、フェイルオーバーの問題は解決した。また、さまざまな障害を想定した試験によってフェイルオーバーが正しく行われることを確認した。

5.2 EPICS Channel Archiver用クラスタ

・ EPICS Channel Archiverのサービス化

今回使用したクラスタソフトであるRoseHAでは、サービスの監視としてNFS, Oracle DBなどをサポートしている。しかし、ユーザーが定義したサービスは監視していないため、EPICS Channel Archiverのサービスがバックアップの計算機にフェイルオーバーができない。この問題に関しては、現在検討中である。特にファイルサーバーに使用するNFSのフェイルオーバーと動機する機構を整えるつもりである。

４．まとめ

本稿では、クラスタシステムの概要とKEK LinacにおけるLinux-PCクラスタの導入、今後の課題について述べた。専用のハードウェアでなくても、安価なPCでクラスタシステムを構成することができた。クラスタシステムの導入により、障害時においてもサービスの継続が可能となり、システムの停止時間を抑えることができる。また、計算機の保守も容易となった。Linux-PCクラスタは、加速器制御に使用するにあたって十分信頼でき、長期安定運用も可能と思われる。

参考文献

[1] S.Kusano, et al., “KEKB LinacとRingの運転ログブックシステム”, Proceedings of the 28th Linear Accelerator Meeting in Japan, Tokai, Aug, 2003, p449-451.
[2] 工藤拓弥、他、本会議で報告