RIBF 制御系における仮想化技術を用いたシステム設計と運用

SYSTEM DESIGN AND IMPLEMENTATION USING VIRTUALIZATION TECHNOLOGY FOR RIBF CONTROL SYSTEM

内山暁仁#, A), 込山美咲 A), 福西暢尚 A)
Akito Uchiyama #, A), Misaki Komiyama A), Nobuhisa Fukunishi A)
A) RIKEN Nishina Center

Abstract

In 2001, EPICS-based system has been introduced to RARF (RIKEN Accelerator Research Facility), which is now the injector part of RI Beam Factory (RIBF). The system consisted of HP-UX as the server system, and vxWorks systems as EPICS IOCs (Input/Output Controller). When we constructed the RIBF control system, we adopted a Linux system as both of the servers and IOCs for RIBF control system, because EPICS version R3.14 was available on Linux. To avoid system-wide failure, we constructed Linux clusters with a redundancy design based on a High-Availability (HA) system using open-source DRBD and Heartbeat in 2008. In order to use server resources more efficiently keeping the high availability of the RIBF control system, we have designed a new implementing system by using VMware vSphere as a virtualization environment and Network Attached Storage (NAS) as the shared storage. The system has successfully started its services such as DNS, HTTP, MySQL, EPICS CA (Channel Access) gateway and EPICS IOC. We report the design and the status of the virtualization system of RIBF control system.

1. はじめに

理研 RIBF 加速器制御システムは主に EPICS を用いた分散制御システムで構築されている。このシステムでは IOC は分散されてはいるが、共通部分のプログラムに関しては NFS 若しくは FTP サービスを用いた共有ディスクで管理されていた。これらのサーバに障害が発生すると IOC を含む全てのシステムに障害が発生するため、2008 年に HA クラスタシステムを構築する事で冗長化を行い、その結果サーバのハードウェアに障害が発生してもサービスを提供し続ける事が可能になった[1]。しかし我々が構築した HA クラスタは、実行系サーバと待機系サーバから成るアクティブ・スタンバイ構成であった為、待機系サーバであっても実行系サーバと同等のサーバリソースが必要であった。仮想化技術が進歩した事を背景に、更新後の新システムでは高可用性だけでなく、サーバリソースの運用効率の向上を実現させるため、仮想化技術を利用してシステムを構築する事とした。

2. 仮想化技術を利用した他施設の動向

2011 年フランス・グルノーブルで行われた制御の国際会議 ICALEPCS2011 では仮想化技術に関する多数の報告が行われた。これをもとに大型実験施設における仮想化技術を利用したシステムについての動向をまとめたのが表 1 である。[2-9]　仮想化ソフトウェアを検討した時、実績がある Xen, KVM, Hyper-V, VMware が候補に挙げられる。一方他施設の仮想化環境の動向をみると運用テストで Hyper-V や Xen を利用することはあっても、実運用の環境は KVM, VMware が採用されているということが分かった。

表 1: 大型実験施設の仮想化技術の動向

施設	仮想化ソフトウェア	用途
J-PARC Main Ring (日本)	KVM	・EPICS IOC
HLS (中国)	VMware vSphere 4	・開発環境, ・EPICS IOC
CERN (欧州)	VMware ESXi	・テスト環境
CLS (カナダ)	VMware ESX	・EPICS IOC, ・アプリケーションサーバ
SSRF (中国)	VMware vSphere 4	・EPICS IOC ・CA gateway
KSTAR (韓国)	VMware vSphere Hypervisor	・オペレータインターフェース
ANKA (ドイツ)	VMware ESXi	・TANGO
CERN /LHCB (欧州)	KVM	・アプリケーションサーバ

3. システム設計

3.1 仮想化ソフトウェア

RIBF 制御系における仮想化ソフトウェアの選択に際して以下を考慮した。

- サーバのハードウェアリソースを効率的に運用させる。

a-uchi@riken.jp

- 物理サーバの障害対策として HA 環境は必須だが、複雑なクラスタリングは運用面を考慮して避ける。
- システム構築コスト削減の為、現在運用している Linux 物理サーバから仮想イメージを作成、システムのディストリビューションや変更を加えずに対応させる。
- ハードウェア障害発生時サービスだけでなく OS も落とさずに対応したい。
- 仮想サーバを集約、管理可能なツールを利用。
- ベンダーからソフトウェアサポートが受けられる。

Hyper-V はサポートしているゲスト OS に現在利用している CentOS4 が含まれておらず、以上の事を考慮した結果、仮想化ソフトウェアは VMware vSphere 5.1 を採用した。

3.2 共有ストレージ

前システムではアクティブ・スタンバイ構成の HA-Linux クラスタで共有ストレージを提供していた。本システムにおける共有ストレージは高性能でかつ実績[10]があり、可用性、信頼性、保守性を備えた NAS (Network Attached Storage)である NetApp FAS2240 を採用した。我々の NAS 使用方法では NFS で共有ストレージを提供させているため、ファイバーチャネルベースの共有ストレージに比べ導入コストを抑える事ができた。また、デュアルコントローラのアクティブ・アクティブ構成で HA システムを構築した為、リソースの提供先に応じてコントローラを使い分ける事が可能であり、待機系は実装されない。したがってサーバリソースの無駄がなくなる。RIBF 制御システムにおいて、この共有ストレージは EPICS プログラムや各サーバで共有されるユーザ領域と VMware イメージファイルを格納する目的で利用される。

3.3 ネットワーク

FAS2240 では物理ネットワークポートはコントローラにつき管理用とは別に 4 ポート備えている。ネットワークトラフィック量を考慮し、仮想 OS のイメージファイル提供用ネットワークポートとその他のファイル提供用のネットワークポートを分ける事とした。その際ネットワーク帯域の拡張と冗長性確保を目的として、2 ポートを束ねて論理的な 1 つのポートとして扱う技術である、リンクアグリケーション(LACP)で構成した。これによって、本システムでは冗長性を持つ 2 つの論理ポートとして利用可能である。

3.4 サーバハードウェア

現時点(2013/7)旧システムからの移行に伴う仮想化すべき物理サーバ台数は 10 台程度であった。さらに今後 20~30 台程度の仮想サーバの増加に対応させるべくハードウェアの選定を行い、その結果 HP ProLiant DL360p Gen8 を採用、物理サーバ 3 台のクラスタリングで構成をした。物理サーバの仕様を表 2 に示す。

3.5 サーバ移行

VMware vCenter Converter Standalone を利用する事で現在動作している物理サーバから直接仮想サーバ用イメージファイルを作成する事が可能である。これにより簡便にシステム移行作業が完了した。また、システム寿命の延命対策で一部古いディストリビューションを仮想環境でそのまま運用している。なお、予めテンプレートのイメージファイルを作成しておく事でファイルコピーをする感覚で新規ホストを追加する事が可能であるので、同じサービスを提供するシステム(EPICS IOC 等)を構築する手間は省ける。

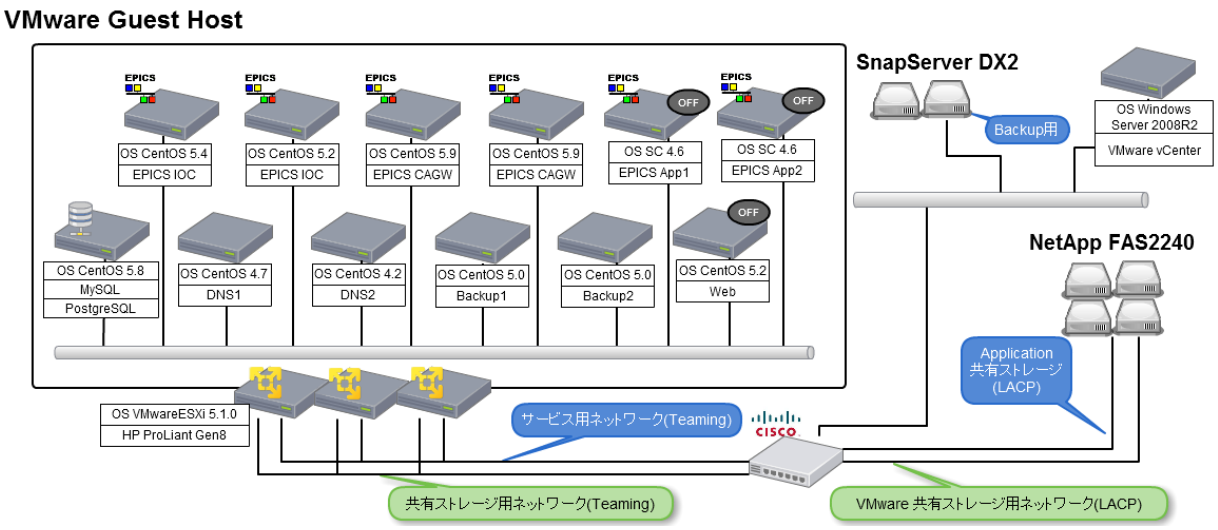

図 1：システムの全体概要図

表 2: 物理サーバ 1 台におけるハードウェア仕様

CPU	Intel Xeon E5-2630 2.30GHz × 2 ソケット
コア数	12 コア (Hyper threading で 24 スレッド利用)
メモリ	32G byte
HDD	146G byte (RAID 1)
ネットワークアダプタ	Gigabit Ethernet×4 100Base-T Ethernet ×1

3.6 サーバ管理

Windows Server 2008R 上にインストールされた VMware vCenter 上で仮想ホストの増減や仮想ネットワークの接続といった管理を統合的に行う。

3.7 バックアップ

以前のバックアップシステムは指定したシステムのディレクトリごとに rsync でデータをネットワークコピーしていた。今回のシステムでは仮想環境で構築された仮想イメージはファイルとして全て共有ディスクである NetApp FAS2240 上に格納されている。よって FAS2240 上に格納されている仮想イメージを全てもう一つの NAS である SnapServer DX2 にコピーする事とした。本システムにおけるバックアップの考え方としては、たとえ火災が起こってもバックアップからシステム復旧できる事である。よって、メインの共有ディスクである FAS2240 を仁科記念棟の計算機室、バックアップの共有ディスクをリニアック棟の計算機室、とそれぞれ別の建物に設置する事とした。

4. 可用性

メンテナンスやハードウェア障害といった計画的な物理サーバのシャットダウン時においても OS を継続して運用する為に、vMotion[11]という機能が用意されている。これはライブマイグレーションと呼ばれており、実行中の仮想マシンを物理サーバ間で移行する技術である。RIBF では 3 台のクラスタリング構成の為、vMotion 時は残り 2 台のうちサーバリソースに余裕がある物理サーバへ仮想マシンを移行させ、サーバのパフォーマンス低下を防ぐ。

5. 仮想サーバ上の現行サービス

5.1 全体概要

全体のシステム構成を図 1 に示す。現行で移行が完了したサービスは DNS (Primary/Secondary), EPICS IOC, EPICS CA Gateway, バックアップ(rsync), PostgreSQL, MySQL である。データアーカイバで利用される PostgreSQL は頻繁なディスクアクセスが発生する為、仮想環境で運用する事は難しいと考えている。一方、CSS (Control System Studio) Alarm と各種 Web アプリケーションで利用するデータベースにおいては、それほど頻繁なディスクアクセスは発生しないので、仮想環境での利用でも十分可能であると考えている。よって、システムは仮想環境で運用した。また GUI 用のアプリケーションである MEDM/EDM やアラーム情報を管理する Alarm Handler 等を走らせる EPICS アプリケーション用サーバと運転ログを表示させる用途で利用されている Web サーバにおける移行準備は既に整っており、今夏メンテナンス時期に移行予定である。

5.2 EPICS IOC

RIBF 制御系において EPICS IOC は重要なサービスの一つである。そこで仮想環境で運用している EPICS IOC の状況について一例を報告する。

RIBF 制御系において EPICS で制御されていない、スタンドアロンな制御システムのデータを統合する目的で MyDAQ2[12]がアーカイブシステムとして採用されている[13]。我々は EPICS IOC と MySQL データベースをインターフェースするプログラムを開発し、MyDAQ2 にデータを格納さえすれば EPICS CA からデータを読み出せる機能を追加した。よってスタンドアロンな制御システムのデータを EPICS にて統合する事が可能になった。システムの概要を以下に示す。実際のクライアントからのデータ読み出しは EPICS IOC への負荷を減らす為に CA gateway 経由で行われている。

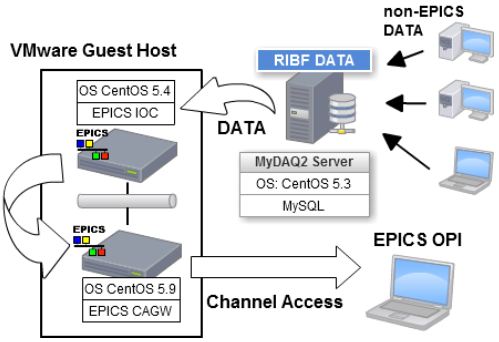

図 2：仮想環境での EPICS IOC 運用例

6. まとめ

VMware vSphere を用いた仮想環境を RIBF 制御システムに導入する事で可用性だけでなく、効率的にサーバリソースを利用する事が可能になり、また統合的なサーバ運用やバックアップも可能になった。近い将来 vMotion 時の EPICS の挙動について、例えば CA クライアントの値変化の遅延がどれだけ起こるのか、試験的に調査したいと考えている。

参考文献

[1] A. Uchiyama, et al., Proc. 6 Annu Meet. Particle Accelerator Society of Japan, Himeji, P.1092-1095 (2010)
[2] N. Kamikubota, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.1165-P.1167

[3] G. Liu, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.323-P.325
[4] O. Khalid, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.622-P.625
[5] E. Matias, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.454-P.457
[6] L. R. Shen, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.1138-P.1140
[7] Sangil Lee, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.694-P.697
[8] T. Spangenberg, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.749-P.751
[9] E. Bonaccorsi, et al., Proceedings of ICALEPCS2011, Grenoble, France, P.1157-P.1160
[10] N. Kamikubota, et al., Proc. 9 Annu Meet. Particle Accelerator Society of Japan, Osaka, P.741-744 (2012)
[11] http://www.vmware.com/
[12] T. Hirono, et al., Proceedings of the4th Annual Meeting of Particle Accelerator Society of Japan and the 32nd Linear Accelerator Meeting in Japan, Wako, P.154-156 (2007)
[13] M. Komiyama, et al., Proc. 8 Annu Meet. Particle Accelerator Society of Japan, Tsukuba, P.225-229 (2011)