RIBF 制御系におけるサーバ・インフラストラクチャの現状

CURRENT STATUS OF SERVER AND SYSTEM INFRASTRUCTURE FOR RIBF CONTROL SYSTEM

内山暁仁#, A), 込山美咲 A))
Akito Uchiyama #, A), Misaki Komiyama A)
A) RIKEN Nishina Center

Abstract

The RIBF control system is mainly constructed with a distributed control system based on EPICS (Experimental Physics and Industrial Control System). In 2008, a high-availability system utilizing open source software was implemented to prevent stopping unexpectedly the important services such as NFS which is used for EPICS IOC (Input / Output Controller). Since 2013, to realize operational efficiency of server resources as well as the high availability, we have provided virtualization environment that adopted NetApp FAS 2240 for shared storage and VMware vSphere 5.1 for server virtualization platform. On the other hand, system migration in the virtualization environment has been processing since 2018, because the useful life of physical servers has passed. In the newly system, we constructed the high-availability system by adopting NetApp FAS 2620-based shared disk and VMware vSphere 6.5 as virtualization platform. For the early detection of system failure, a management tool, which is monitoring of network traffic at the protocol level, and an alive-monitoring tools for EPICS-based system are also introduced.

1. はじめに

加速器制御系において重要なサービスが停止する障害が発生した場合、加速器オペレーションを続ける事ができず、様々なコスト増大に直結してしまうケースがある。したがって理研 RIBF 制御システムでは NFS, FTP, DNS, LDAP, PostgreSQL といったサービスの予期せぬ停止を防ぐ目的で、2008 年よりオープンソースによって高可用性システム（フェールオーバークラスタ）を構築し運用してきた[1]。 一方で構築されたシステム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ため、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。

2013 年に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、仮想化技術を導入した[2]。構築されたシステムは NAS (Network Attached Storage) 共有ストレージに NetApp FAS2240、サーバ仮想化のプラットフォームに VMware vSphere 5.1[3]を採用、物理サーバ3台に約40台の仮想マシンを運用し、リレーショナルデータベースを除いたサービスの集約をする事に成功した。 結果として上記仮想マシン上に実装されたサービスについて、運用開始より 2018 年 7 月現在までほぼ 100%の可用性を実現できている。

2018 年より上記物理サーバの耐用年数が経過した事を背景に、新しいシステムへの置き換えを進めている。また高可用性システムの構築だけでなく、EPICS IOC (Input/Output Controller)を含めた RIBF 制御系で利用されているサービスについての死活監視やネットワークの帯域使用状況についても可視化を行い、解析を行う事でさらなる障害対策の強化を行っている。

a-uchi@riken.jp

2. サーバシステム設計

2.1 サーバ仮想化ソフトウェア

以前のシステムにおいて VMware vSphere でクラスタリングを構築する事によって、サーバのハードウェアリソースを効率的に運用する事に成功した。 また VMware vSphere の vMotion[4]機能を用いる事でマザーボードやメモリといったハードウェア交換作業を伴う障害発生時に OS を落とさずに修理対応する事が可能になった。VMware vSphere を採用した結果、高い可用性と運用効率を実現できたという事から、新たなシステムにおいても以前のシステム設計を踏襲し、主に以下の点に基づいてシステム構築した。

- 共有ディスクには NAS の利用
- 物理サーバ3台をクラスタリング
- 仮想化ソフトウェアは現状の最新バージョンである VMware vSphere 6.5 の採用

システム概要を Fig. 1 に示す。

2.2 ネットワーク共有ストレージ

本システムにおける共有ストレージは可用性の実績と性能を考慮した結果、NetApp FAS2620 の OEM 製品である富士通 ETERNUS NR1000 F2620[5]を採用した。共有ストレージのサービス提供には、1 Gbps イーサネットベース上を用いた NFS を採用していることから、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る。

また待機系は実装せずに、デュアルコントローラを持ったアクティブ・アクティブ構成のため、サービ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る。実際の運用としてはゲスト OS のイメージファイルの提供と EPICS プログラム・ユーザディレクトリの提供と言った接続先の用途に応じコントローラを分けている。

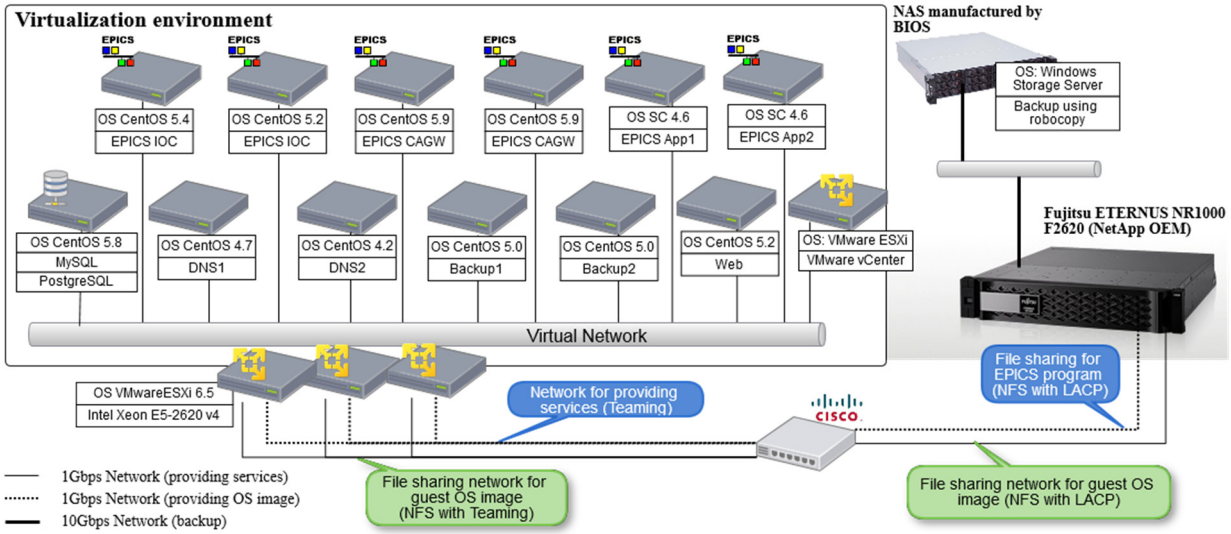

Figure 1: System diagram for newly virtualization environment. The network is constructed by redundancy method, which are teaming and LACP. 10Gbps network is installed for backup purpose.

サービス提供用ネットワークにおいても冗長化を行っている。 本システムで採用された共有ストレージは管理用ポートとは別にコントローラ毎に4ポートのイーサネットポートを備えている（1 Gbps×2ポート、10 Gbps×2ポート）。 本システムではその内2ポートをリンクアグリケーション（LACP）で構成する事により冗長性を持った論理的な1ポートとして扱っている。

2.3 物理サーバ仕様

物理サーバの仕様を Table 1 に示す。 従来のシステムでは rsync でのバックアップタスク中にゲスト OS の挙動が体感的に遅くなる現象が起こっていた。 これを解決するために新しいシステムでは内蔵 SATA ベースの SSD に VMware vSphere の Swap を格納（ホストスワップキャッシュ）するように設計して、実際のゲスト OS のファイル I/O をより高速に実現させた。 従来システムと新システムを CrystalDiskMark [6]でベンチマークの計測をし、ファイル I/O 性能を比較したデータを Fig. 2 に示す。 物理サーバのスペックが異なるため単純な比較はできないが、従来サーバシステム（Intel Xeon CPU E5-2630×2ソケット, 96GB メモリ, SAS RAID1, 1 Gbps Ethernet）と比べ、本システム上に構成されたゲスト OS（Microsoft Windows 10）は、シーケンシャルリードで約2倍、シーケンシャルライトで約3倍高い性能を持つ事を確認した。またランダムアクセスに関しても高い性能を示している。

Table 1: Specification of Newly Physical Server for Virtualization Environment

CPU	Intel Xeon E5-2620 v4 ×2 socket
Logical processor	32 threads
Memory	128 GB
Storage	SAS 280 GB RAID1 (OS) SATA SSD 230 GB (SWAP)
Ethernet	1 Gbps×4 port 10 Gbps×4 port

Sequential Access	Read (MB/sec)	Write (MB/sec)
New Server	**117.1**	**116.7**
Previous Server	**56.1**	**32.32**
Random Access, 8queue, 8thread	Read (MB/sec)	Write (MB/sec)
New Server	**81.88**	**111.9**
Previous Server	**10.87**	**32.75**
Random Access, 32queue, 1thread	Read (MB/sec)	Write (MB/sec)
New Server	**51.63**	**108.1**
Previous Server	**11.27**	**32.28**
Random Access, 1queue, 1thread	Read (MB/sec)	Write (MB/sec)
New Server	**11.36**	**9.134**
Previous Server	**11.3**	**8.087**

Figure 2: Comparison of file I / O performance on guest OS in previous server and new server.

3. ネットワーク監視

3.1 RIBF 制御系ネットワーク

RIBF 制御系では、中心に Cisco 製コアスイッチ（Catalyst 4506）を実装したスター型でネットワークが構成されている。 スター型ネットワークの特徴として、スイッチの増設といった拡張を簡便に行う仕組みを実現する事ができる一方で、通信経路が一部の機器に集中し負荷が増大する可能性があり、それが通信障害になるケースもある。 したがって、ネットワークトラフィックやサーバ死活監視が重要になってくる。

3.2 PRTG Network Monitor

現在 RIBF 制御系では、ネットワーク監視ツールとして PAESSLER 社の PRTG[7]を採用している。 RIBF 制御系では従来 Nagios[8]でネットワーク監視を行っていたが PRTG の利点は SNMP を利用した Cisco スイッチのトラフィック監視が簡便に行えるだけでなく、様々なプロトコル監視を標準でサポートしており、導入の閾値が低い事である。本システムでは主に以下の用途で PRTG を利用している。

- 仮想化用物理サーバの状態（温度、ファン、ディスク等）
- HTTP や FTP といった重要サービスの死活監視
- ネットワークスイッチの状態（CPU,メモリ、ファン等）
- ネットワークスイッチのポート毎における帯域監視
- NAS の状態（ディスク使用状況、温度、ファン等）
- NetFlow（3.4 にて後述）

一方で Zabbix も標準で Cisco を含む様々なプロトコル監視がサポートされており実装例[9]も報告されているが、PRTG は後述する NetFlow の監視も簡便に行え、かつ一元管理が可能であるという利点がある。

PRTG を実装する事によって、ポート単位のネットワークトラフィック、NetApp 製 NAS の使用状況、そして仮想マシンの負荷や温度等を可視化し一元管理する事が可能になった。 また PC ブラウザだけでなく iOS 専用の PRTG クライアントが提供されているため、iPhone を利用してネットワークの状況を確認する事が可能である (Figure 3 参照)。

3.3 EPICS 死活監視システム

2015 年ぐらいより EPICS IOC、CA (Channel Access) プロトコルを監視する目的として、EPICS 死活監視システムを開発、導入している [10]。 本システムの特徴は EPICS PV 管理システム[11]と連携する事によって、監視対象の EPICS PV, IOC, ネットワークデバイスといった情報を自動で取得し、監視を開始するという点にある。 したがって、従来の監視システムの様に、テンプレートから監視用ファイルを個別に設定、もしくは編集といった作業をする必要がなくなる。 サポートする機能は次の通りである。

- EPICS IOC への ICMP を利用した死活監視
- EPICS IOC へのポートチェックによる EPICS CA の死活監視
- EPICS IOC への caMonitor を利用した EPICS CA の死活監視

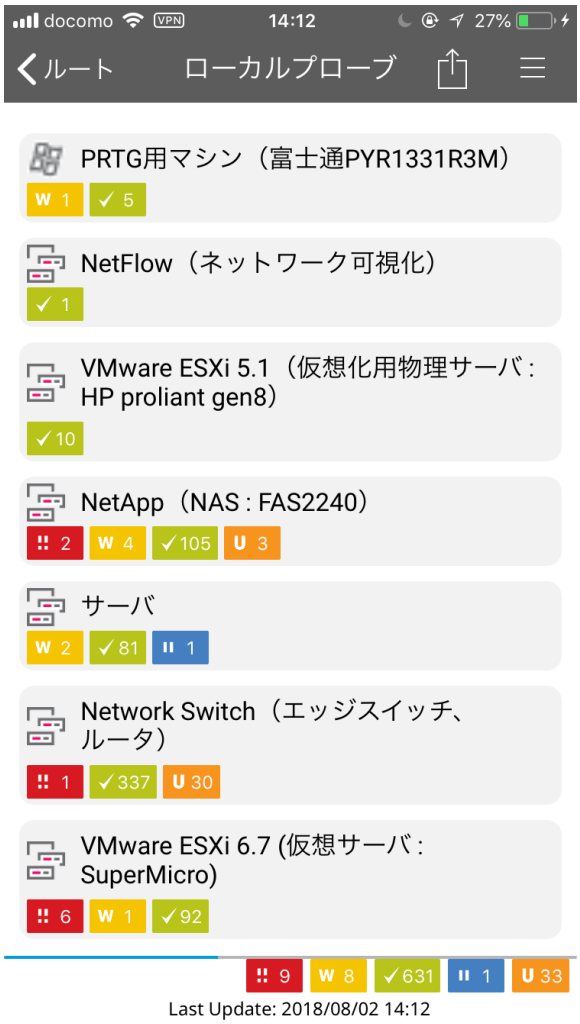

Figure 3: Screenshot of PRTG client displayed for RIBF control system network on iPhone6. The application is installed through App Store.

- ネットワークデバイスに対する ICMP を用いた死活監視
- ネットワークデバイスに対するポートチェックによるサービスの死活監視

EPICS 死活監視システムのユーザインターフェースは Web アプリケーションによって実現されている。このユーザインターフェース上で CA プロトコルとネットワークデバイスの現在の死活監視状況と過去ステータスがログとして確認する事ができる。

3.4 NetFlow

RIBF 制御系ネットワークでは、コアスイッチを経由するトラフィックを NetFlow で収集、分析も行っている。NetFlow は Cisco によって開発された、スイッチを通過するトラフィック情報を収集、分析するためのプロトコルである [12]。 従来ネットワーク監視には SNMP が広く用いられている。 しかし SNMP は SNMP マネージャがネットワーク機器である SNMP エージェントに対して、ポーリングアクセスを用いてデータ収集をするプロトコルであり、プロトコル毎の帯域等、詳細なトラフィック情報の収集を

する事は困難であると考えられる。 一方 NetFlow は予め設定された NetFlow コレクタに対して全てのトラフィック情報を送信するため、例えば送信元 IP アドレス、宛先 IP アドレス、そしてアプリケーションレベルのプロトコルといった詳細な情報を取得する事が可能である。

本システムにおいてNetFlowコレクタは、他のRIBF制御系ネットワーク監視系と同様、3.2 にある PRTG を利用している。 他社製品は NetFlow コレクタとしてのみ動作するのに比べ、PRTG を採用する利点は NetFlow コレクタだけでなく他の SNMP を利用したトラフィック監視や ICMP（いわゆる ping）を利用した死活監視としても動作するので、一元的に管理することが可能な事である。

4. バックアップタスク

日々のシステム運用でバックアップは重要なタスクである。 従来のシステムにおいて、各ゲスト OS のユーザディレクトリ、設定ファイル、そして EPICS アプリケーションは rsync でネットワーク経由にてバックアップされている。 同様に NetApp FAS2240 に格納された各ゲスト OS のイメージファイルも rsync でバックアップしている。 しかし RIBF 制御系ネットワークは従来 1 Gbps ベースの帯域を持っているが、バックアップ時にその帯域を全て使用している事が NetFlow のデータから判明した (Figure 4 参照)。 上記を解決するため新たなシステムにおいてはバックアップタスクの手法について変更を行った。

新たなシステムでは物理サーバ、ETERNUS NR1000 F2620、そしてバックアップファイル格納用 NAS に全て 10 Gbps イーサネットポートを備えている事から、バックアップタスク専用のインターナルな 10 Gbps ネットワークを実装してバックアップタスクが制御ネットワークに影響を及ぼさない様に対応した。

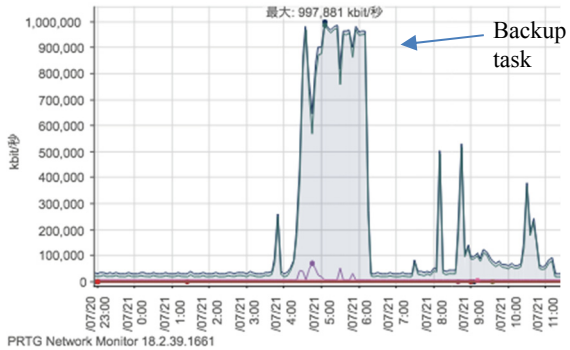

Figure 4: Network traffic measured by NetFlow. The backup task occupies about 1Gbps network bandwidth.

5. まとめ

RIBF 制御システムで 2013 年に導入した仮想環境は現在まで高い可用性と運用効率を実現できている。2018 年に新たなシステムへのシステムリプレースした時、以前のシステム設計を踏襲しただけでなく、ファイル I/O とバックアップタスクの改善も行っている。 また障害の迅速な検知を行うためネットワークやデバイス監視の強化を行っている。

参考文献

[1] A. Uchiyama *et al.,* Proceedings of the 7th Annual Meeting of Particle Accelerator Society of Japan, Himeji, Aug 4-6, 2010, pp. 1092-1095.
[2] A. Uchiyama *et al.,* Proceedings of the 10th Annual Meeting of Particle Accelerator Society of Japan, Nagoya, Aug 3-5, 2013, pp. 1109-1112.
[3] https://www.vmware.com/jp/products/vsphere.html
[4] https://www.vmware.com/files/jp/pdf/vmotion_datasheet.pdf
[5] http://www.fujitsu.com/jp/products/computing/storage/disk/nas/nr1000f-entry/
[6] https://crystalmark.info/en/software/crystaldiskmark/
[7] https://www.paessler.com/jp/prtg
[8] D. E. R. Quock *et al.,* Proc. PCaPAC08, Ljubljana, Slovenia (2008), p.19.
[9] T. Sugimoto *et al.,* Proceedings of the 13th Annual Meeting of Particle Accelerator Society of Japan, Chiba, Aug 8-10, 2016, pp. 652-655.
[10] A. Uchiyama *et al.,* Proceedings of the 13th Annual Meeting of Particle Accelerator Society of Japan, Chiba, Aug 8-10, 2016, pp. 664-667.
[11] A. Uchiyama *et al.,* Proc. ICALEPCS2015, Melbourne, Australia (2015), pp. 769-771.
[12] Claise, Benoit. Cisco systems netflow services export version 9. No. RFC 3954. 2004.; http://www.rfc-editor.org/info/rfc3954