SuperKEKB における制御用サーバー計算機の現状

PRESENT STATUS OF SERVER INFRASTRUCTURE FOR THE SuperKEKB CONTROL SYSTEM

佐々木信哉[#, A)], 杉村仁志 [A)], 中村達郎 [A)], 中村卓也 [B)], 吉井兼治 [B)], 廣瀬雅哉 [C)]
Shinya Sasaki [#, A)], Hitoshi Sugimura[A)], Tatsuro Nakamura[A)], Takuya Nakamura[B)], Kenji Yoshii[B)], Masaya Hirose[C)]
[A)] High Energy Accelerator Research Organization (KEK)
[B)] Mitsubishi Electric System & Service Co., Ltd.
[C)] Kanto Information Service Co., Ltd.

Abstract

The server infrastructure for the SuperKEKB control system has been migrated from blade servers to rack-mount servers. The new server infrastructure provides virtualized environments with KVM that allows to run applications in the isolated environment. The configuration of the physical servers and virtual machines is managed with Ansible to automate the deployment and prevent the system construction from depending on a particular person. The server system is monitored with iDRAC and Zabbix. iDRAC, IPMI device integrated in DELL servers, provides physical server management platform, and shows server status and log data. The Zabbix server monitors resource of the servers and virtual machines and alerts administrators to system issues.

1. はじめに

SuperKEKB では DNS や DHCP、LDAP、NTP などの制御システムの基幹となるサービスの実行、および加速器制御用アプリケーションの開発や実行のためにブレードサーバーを運用してきた[1]。ブレードサーバーは SuperKEKB の前身である KEKB の頃から利用しており、その保守が終了する前に新しいシステムへ移行する必要があった。そのため、2020 年からラックマウントサーバーの導入とサービスの移行を進めてきた。

ブレードサーバーを利用したシステムでは仮想化技術を利用せず、物理サーバー上で直接アプリケーションを実行してきた。しかし、用途の異なる複数のアプリケーションが同一の実行環境で動作しており、管理が煩雑化していた。ラックマウントサーバーを利用した新しいシステムでは KVM[2]による仮想環境を利用することで、用途ごとに実行環境が分離され、管理を容易にすることができた。また、サーバーや仮想マシンの環境構築作業が属人化することを防ぐことや、環境構築の再現性や再利用性を高めることを目的に、IT 自動化ツールである Ansible[3]の導入も行った。サーバーや仮想マシンの監視には iDRAC や Zabbix[4]を利用して、利用状況の確認やアラートのメール通知を行っている。

本稿では導入したサーバーの詳細およびその運用状況に関して報告する。

2. システム構成

2.1 全体構成

Figure 1 に SuperKEKB における制御用サーバー計算機システムの全体構成図を示す。計算機システムは 2 台のコアサーバー、6 台の VM ホストサーバー、3 台のオフィスサーバー、3 台のネットワーク共有ストレージによって構成されている。

SuperKEKB 加速器制御ネットワークは KEK オフィスネットワーク（機構内ネットワーク）とファイアウォールによって分離されている。ほとんどの機器は SuperKEKB 加速器制御ネットワークに接続しているが、オフィスサーバーのみ KEK オフィスネットワークに接続している。制御ネットワークに接続する機器は 10 GBASE-T で接続し、オフィスネットワークに接続する機器は 1000 BASE-T で接続している。

以下からは、まずシステムを構成するサーバー計算機およびストレージシステムの詳細を説明する。その後 KVM による仮想環境の運用状況に関して説明する。

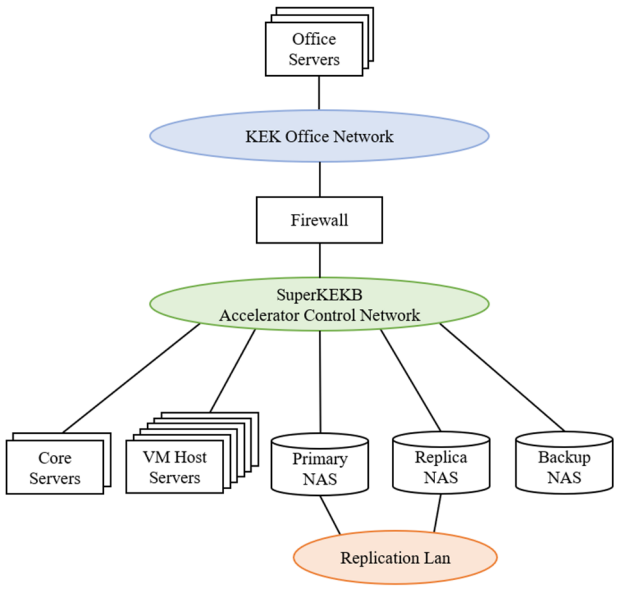

Figure 1: Schematic view of the SuperKEKB server infrastructure.

shinya.sasaki@kek.jp

2.2 サーバー計算機

サーバー計算機は用途によって大きく以下の 3 つに分類される。

- コアサーバー
- VM ホストサーバー
- オフィスサーバー

それぞれのサーバーの仕様を Table 1 および Table 2、Table 3 に示す。

Table 1: Specification of Core Server

機種	Dell PowerEdge R6515
CPU	AMD EPYC 7262 3.2GHz 8 コア
Memory	16 GB
Storage	SATA HDD 2TB×2 RAID1
OS	CentOS 7.9

Table 2: Specification of VM Host Server

機種	Dell PowerEdge R6515
CPU	AMD EPYC 7302P 3.0GHz 16 コア
Memory	64 GB
Storage	SATA HDD 2TB×2 RAID1
OS	CentOS 7.9

Table 3: Specification of Office Server

機種	Dell PowerEdge R340
CPU	Intel Xeon E-2224 3.4GHz 4 コア
Memory	8 GB
Storage	SATA HDD 1TB×2 RAID1
OS	CentOS 7.9

コアサーバーは加速器制御システムの基幹となるサービスや、VM ホストサーバーの管理を行うサービスを実行するサーバーである。2 台の内 1 台は DNS や DHCP、LDAP、NTP などの基幹となるサービスの実行に利用している。もう 1 台は VM ホストサーバーの管理のために使用する Ansible や Cockpit[5]などのサービスの実行に利用している。

VM ホストサーバーは、KVM を実行して仮想環境を提供するためのサーバーであり、主に加速器制御用アプリケーションを実行するために利用する。加速器制御用アプリケーションは物理サーバー上で直接実行せず、KVM の仮想マシン上で実行する。そのため、全ての VM ホストサーバーの環境を統一することができる。6 台ある VM ホストサーバーの内 1 台は開発用として運用している。

オフィスサーバーは加速器制御ネットワーク内の情報に KEK オフィスネットワーク内からアクセスするためのサービスを実行することが主要な役割である。例えば、加速器制御ネットワーク内の EPICS PV をオフィスネットワークから読み出すための CA Gateway や、加速器制御ネットワーク内で運用している Web サービスにオフィスネットワークからアクセスするためのポートフォワードのサービスなどが実行されている。

2.3 ストレージシステム

ストレージシステムはプライマリ機、レプリカ機、バックアップ機の 3 台の NAS から構成される。プライマリ機とレプリカ機は 10 GBASE-T の専用ネットワークで接続している。プライマリ機とレプリカ機の仕様を Table 4 に、バックアップ機の仕様を Table 5 に示す。

Table 4: Specification of Primary NAS and Replica NAS

機種	Dell EMC NX3240
CPU	Intel Xeon Bronze 3204 1.90 GHz 6 コア
Memory	16 GB
OS Storage	SAS HDD 600GB×2 RAID1 (10K-RPM, 2.5 inch)
Data Storage	NLSAS HDD 8TB×6 RAID6 (7.2K-RPM, 3.5 inch)
OS	Windows Server 2019 Standard Edition

Table 5: Specification of Backup NAS

機種	QNAP TS-1886-XU
CPU	Intel Xeon D-1622 2.60 GHz 2 コア
Memory	16 GB
Data Storage	NLSAS HDD 16TB×12 RAID6 （12 台中 2 台がホットスペア）
OS	QTS

導入当初の予定では、プライマリ機に配置した仮想マシンのイメージファイルを VM ホストサーバーから iSCSI でマウントして利用する予定だった。そして、プライマリ機のデータに変更があった場合は即座にレプリカ機にデータが同期されるように、記憶域レプリカ[6]によってレプリケーションしようとしていた。しかし、iSCSI でマウントしているデータを記憶域レプリカによってレプリケーションしていると、VM ホストサーバーとプライマリ機の iSCSI 接続が切れてしまう症状が発生していた。仮想マシンのイメージファイルを iSCSI でマウントしている場合、その接続が切れると仮想マシンも正常に動作しなくなってしまう。そのため、仮想マシンを安定に動作させることを優先して、仮想マシンのイメージファイルは VM ホストサーバーのローカルディスク上に置くことにした。

現在プライマリ機は、KVM の仮想マシンから大きなサイズのデータを高速に読み書きして永続化したい場合に

iSCSI 接続でマウントして利用している。例として、リレーショナルデータベースのデータが挙げられる。

レプリカ機は上述した iSCSI の接続切れの問題があったため、利用方法を再検討して整備を進めている状況である。現在は 2 時間に 1 回など定期的にプライマリ機のスナップショットデータをレプリカ機にバックアップするような使い方を検討している。

バックアップ機は VM ホストサーバーのローカルディスク上に置くことになった仮想マシンのイメージデータや仮想マシン内の永続化データのバックアップに利用している。

各 NAS ではスナップショットの取得を実施しており、その頻度はプライマリ機およびレプリカ機では 2 時間に 1 回、バックアップ機では 1 日に 1 回である。

2.4 KVM による仮想環境

仮想マシンによる仮想環境の実現のために、VM ホストサーバーでは KVM を実行している。KVM は Linux に組み込まれたオープンソースの仮想化テクノロジーである。2022 年 10 月現在、開発用を除く 5 台の VM ホストサーバーにおいて 17 個の仮想マシンを稼働している。

iSCSI の接続切れの問題があったため、仮想マシンのイメージファイルは VM ホストサーバーのローカルストレージ上に配置して管理している。サーバー間でイメージファイルを共有していないため、仮想マシンのライブマイグレーションは実施できない。サーバー間で仮想マシンを移動したい場合には、対象の仮想マシンを停止してから、イメージファイルを移動先のサーバーにコピーして起動する必要がある。仮想マシンのイメージファイルは仮想マシンが停止する長期メンテナンス中にバックアップを行っている。

仮想マシンの管理や起動、終了処理には Linux サーバーの管理用 Web インターフェイスである Cockpit を利用している。Figure 2 に Cockpit の操作画面を示す。Cockpit では複数のサーバーを統一的に管理することができ、KVM の仮想マシンも Cockpit から管理することが出来る。個別のサーバーにログインすることなく全体の状況を把握し、起動や終了ができるため、効率的な管理に役立っている。

3. サーバーの構成管理

各サーバーの構成管理のために Ansible を利用している。Ansible はサーバーの構成管理やソフトウェアのデプロイメントを自動化する IT 自動化ツールである。従来はスクリプトや手動で行っていた操作を Ansible によって効率的に実行できるようになった。

Ansible ではプレイブックと呼ばれる YAML 形式のファイルに、実行したい処理の流れを記載する。プレイブックから呼び出される「パッケージのインストール」や「サービスの起動」などの処理はモジュールと呼ばれる単位で管理される。モジュールは Ansible の開発チームが保守しているものや、コミュニティによって保守されているものなど多数あり、プレイブックを作成するだけで様々な自動化を実現できる。複数の処理を目的ごとにロールという単位でまとめることも可能である。ロールを利用することでプレイブックの管理を容易にし、その再利用性を高めることができる。

例えば、新しく導入した全てのサーバー計算機と仮想マシンにはシステム監視のために Zabbix エージェントをインストールしている。この Zabbix エージェントのインストール処理を「zabbix-agent」というロールによって管理することで、インストールに必要となるパッケージのインストールや設定ファイルの変更などの複数の処理をまとめて管理することが出来る。

Ansible のプレイブックはグループ内で運用している GitLab[7]で管理している。プレイブックを変更する際は GitLab のイシューやマージリクエストを発行し、課題や変更内容をグループ内で共有している。これにより、サーバーの構成管理が個人に依存しないようにしている。

4. 監視システム

システムの監視には iDRAC と Zabbix を利用している。

iDRAC は Dell のサーバー製品に搭載されている IPMI デバイスである。iDRAC ではサーバーの状態やログの確認および BIOS 設定などのサーバー管理を行うことができる。また、iDRAC はサーバーの障害や状態を通知するアラート機能も備えている。SuperKEKB では各サーバー計算機において iDRAC のアラートをメールで通知するほか、SNMP トラップで取得したアラートをグループ内で利用しているチャットツール Mattermost[8]にも通知するようにしている。Figure 3 に iDRAC から Mattermost に通知されたアラートを示す。

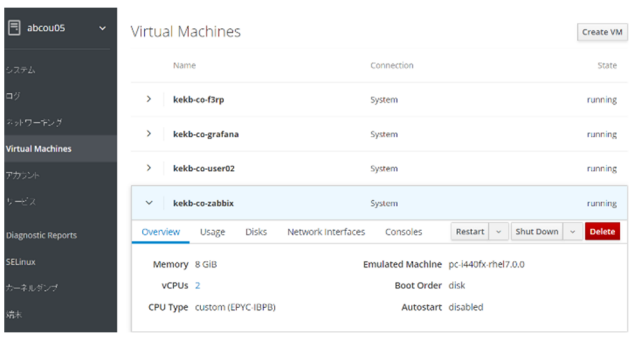

Figure 2: Screenshot of Cockpit Web UI. It allows to manage virtual machines from Web browser.

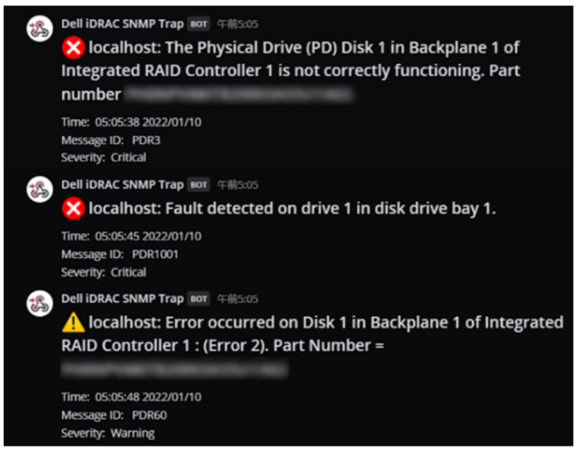

Figure 3: Screenshot of iDRAC alert messages on Mattermost.

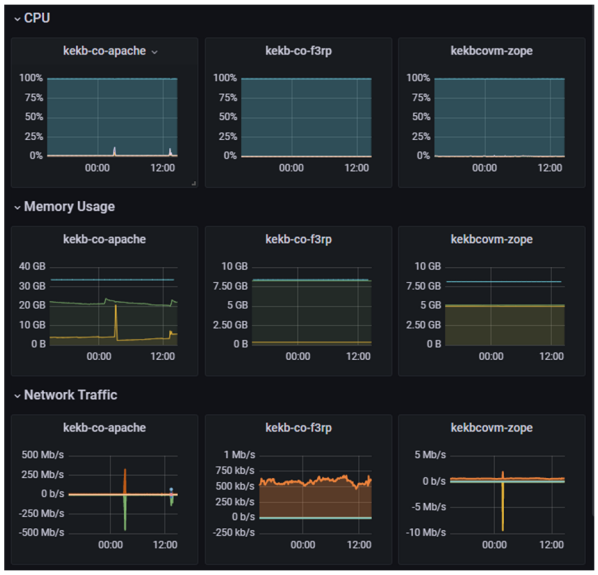

Figure 4: Grafana dashboard for resource monitoring of virtual machines.

サーバー計算機や仮想マシンの CPU 使用率などのリソース監視には Zabbix を利用している[9]。各サーバー計算機や仮想マシンには Zabbix agent をインストールしており、監視対象のデータが定期的に収集される。障害を検知した際はメールでアラートを通知するようにしている。収集したデータは Grafana[10]上で可視化し、利用状況を理解しやすいようにしている。Figure 4 に仮想マシンのリソース状況を表示する Grafana のダッシュボードを示す。

5. まとめ

SuperKEKB の制御用サーバー計算機としてこれまで利用してきたブレードサーバーから移行するために、ラックマウントサーバーの導入とサービスの移行を進めてきた。加速器制御用アプリケーションは KVM による仮想環境で実行することで、効率よく管理することが出来ている。また、サーバーの構成管理に Ansible と GitLab を利用する事で効率的にサーバーを構築し、個人への依存も小さくすることが出来た。システム監視には iDRAC や Zabbix を利用して、障害発生時にはアラートが通知されるようにしている。

一方で、レプリケーションを行っているデータを iSCSI でマウントすると iSCSI の接続が切断される症状がストレージシステムで発生している。そのため、当初の予定とは異なる構成でシステムを運用している。今後は iSCSI 接続が切断される症状の調査を継続し、適切なシステム構成の検討を進める。

参考文献

[1] H. Sugimura *et al.*, "Control system for SuperKEKB accelerator", Proceedings of the 16th Annual Meeting of Particle Accelerator Society of Japan, Kyoto, Japan, Jul. 31 - Aug. 3, 2019, pp. 144-148; https://www.pasj.jp/web_publish/pasj2019/proceedings/PDF/THOI/THOI06.pdf
[2] http://www.linux-kvm.org
[3] https://www.ansible.com
[4] https://www.zabbix.com
[5] https://cockpit-project.org
[6] https://docs.microsoft.com/ja-jp/windows-server/storage/storage-replica/storage-replica-overview
[7] https://about.gitlab.com
[8] https://mattermost.com
[9] S. Sasaki *et al.*, "Monitoring system with Zabbix at SuperKEKB", Proceedings of the 16th Annual Meeting of Particle Accelerator Society of Japan, Kyoto, Japan, Jul. 31 - Aug. 3, 2019, pp. 596-599; http://www.pasj.jp/web_publish/pasj2019/proceedings/PDF/THPH/THPH005.pdf
[10] https://grafana.com