J-PARC MR 制御での仮想マシンの応用

VIRTUAL MACHINES IN J-PARC MR CONTROL

上窪田紀彦#, A), 吉田奨 B), 本橋重信 B), 飯塚上夫 B), 根本弘幸 C),
高橋大輔 B), 山本昇 A), 山田秀衛 A), 佐藤健一 A)
Norihiko Kamikubota #, A), Susumu Yoshida B), Shigenobu Motohashi B), Takao Iitsuka B), Hiroyuki Nemoto C),
Daisuke Takahashi B), Noboru Yamamoto A), Shuei Yamada A), Kenichi C.Sato A)
A) J-PARC Center, KEK and JAEA
B) Kanto Information Service (KIS)
C) ACMOS Co. Ltd.

Abstract

In J-PARC MR, a scheme of virtual ioc (an EPICS I/O-controller which runs on a virtual machine) was introduced successfully in 2011. Following on this success, we intended to introduce more virtual machines even for server services. Scientific Linux 6 and KVM are selected as an OS and virtual environment for parent servers. Details of our virtual environment setups and experiences for the past two years are reported

1. はじめに

最近、計算機の仮想化技術が著しく発展・普及し、さまざまな機会で耳にするようになった。仮想環境を利用するには、かつては専門的な知識やそれなりの予算が前提であった。今や Windows には Virtual PC が、Linux では KVM が、仮想環境として標準で用意されている。仮想環境が身近になったのは、急速な 64 ビット PC の普及で、幅広いユーザが旧 OS（特に 32bit 時代）の backward-compatibility を必要としたためと思われる。

仮想化技術を導入すると、サーバ計算機の構成には構造的な変化が起こる。従来は１台の計算機には１種のサービスのみを受け持たせ、個々の計算機の管理を単純化しようと考えた。その結果多数の Low/Middle-spec の計算機が必要になるが、blade 型計算機システムの導入で計算機機体の扱いは効率的に行えた（図１左）。一方仮想環境では、１台の仮想マシンが１種のサーバを受け持つ（管理の単純さは従来と同じ）が、少数の High-spec な親計算機が複数の仮想マシンを持つ構成となる（図１右）。仮想環境では、負荷状態や保守都合に応じ仮想マシンを親サーバ間で柔軟に移動出来るため、少数の親サーバ機のハードウェアに集中投資して総コストを下げることが出来る。

J-PARC MR は、EPICS ベースの制御システムで運用している[1]。100 個以上の IOC (In Out Controller)が使用されているが、IOC の半数はこの分野で標準的な VME-bus 計算機である[2]。2010 年頃から VMWARE や XEN などを使用して仮想環境を加速器制御に応用する検討を行っていた。2011 年には仮想環境に KVM を導入し、仮想マシン上で IOC を動かす”Virtual IOC (vioc)”を開発した[3]。ソフトレコードやネットワークデバイスのみを扱う IOC（Real I./O 無し）の一部は、VME-bus 計算機から vioc に移行した。現在(2013 年)、約 30 台の vioc を運用している。

vioc の成功に引き続き、2012 年に J-PARC MR 制御に必要なサーバ機能の仮想マシンへの移動が試みられた[4]。本稿では、J-PARC MR の仮想環境を解説し、２年間の運用経験を報告する。

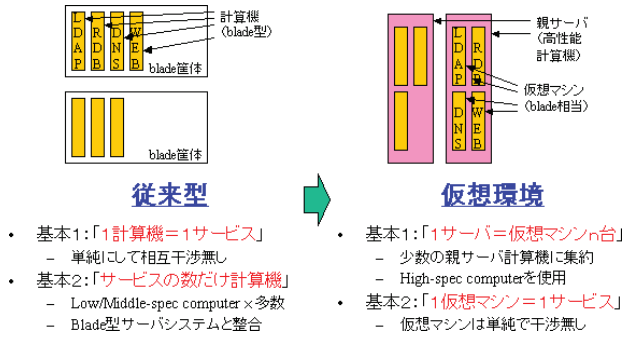

Figure 1: Server structures with/without virtualization.

2. J-PARC MR の仮想環境運用

2.1 仮想環境の概要

J-PARC MR では、主たる OS として Linux 系の Scientific Linux (SL)を採用している。運転が始まった 2008 年は SL4 であったが、2012 年夏季にサーバ系計算機を SL4 から SL6 に update した。当時、KVM で vioc の運用経験を積んでいたこと、SL6 に KVM が装備されたこと、などから、KVM を標準的に利用する仮想環境とした。

現在（2013 年）サーバ系と vioc 系の２系統の仮想環境を運用している。それぞれの系は、親サーバ計算機３台で構成する（ただし１台は両系で重複し総数５台）。通常時は３台で負荷が平均化するよう

norihiko.kamikubota@kek.jp

仮想マシンを分散しているが、１台が故障した際は残りの２台で運転を継続する。図２(a)にサーバ系、図２(b)に vioc 系の構成を示す。

親サーバ計算機は、20GB のメモリを積んだ比較的 High-spec な計算機(IBM Blade HS22, Xeon E5504 (4core 2cpu) 2GHz)を使用している。これに対し仮想マシンに割り当てる１台あたりの典型的なメモリは、サーバ系で 2-4GB、vioc 系で 512MB である。

図２(a)で示すように、サーバ系が実計算機から仮想環境に移行したことで、全体として８台の計算機を３台に集約したことになる。Archive engine は今後負荷が増えると思われるが、分割や親サーバ追加で対応する予定である。保守 server は特殊な仮想マシンで、SL6 移行以前の SL4.4 やサーバ系仮想マシンで使用する SL5.4 の開発環境を保持する（この事情は 2.3 節も参照されたい）。昔のアプリの動作検証などに重宝している。

サーバ系

親サーバ機(親) SL6.0, mem=20GB	仮想マシン(子) SL5.4, mem=2-4GB (typical spec for each v-machine)
jkjblade3a	・Archive engine ・CA-gateway
Jkjblade3b	・管理server (dhcp, tftp, ldap slave) ・App server (cron, zlog, cacti) ・RDB server (postgres, mysql) ・保守server (SL5.4環境)
jkjblade3c	・管理server (ldap master) ・保守server (旧OS:SL4.4環境)

Figure 2(a): Assignment for virtual machines (servers).

Vioc系

親サーバ機(親) SL6.0, mem=20GB	仮想マシン(子) SL6.3, mem=512MB (typical spec for each vioc)	
jkjblade3a	・cont-group ３台	
Jkjblade3e	・mag-group ４台 ・mon-group ３台 ・rf-group １台	・inj-group ２台 ・accom-group １台
jkjblade3f	・cont-group ６台 ・mon-group １台	・sx-group ２台 ・個人試験･開発用 ８台

Figure 2(b): Assignment for virtual machines (viocs).

2.2 仮想環境の運用・管理

典型的な親サーバ機(jkjblade3b、jkjblade3e)の運用中の負荷状態（10 分間の CPU とメモリ）を図３に示す。負荷監視は、標準的な管理ツールを使っている。親サーバ機のうち１台(jkjblade3b)では、メモリ使用率がやや高くなっていることがわかる。

我々は全部で５台の親サーバ機を持っているが、ある仮想マシン（子）の親を変更したい場合、やはり標準的な管理ツールで作業する。図４は、ある仮想マシンを選択し、親サーバ機を変更しようとする作業の画面である。

Figure 3: Typical CPU and memory load status.

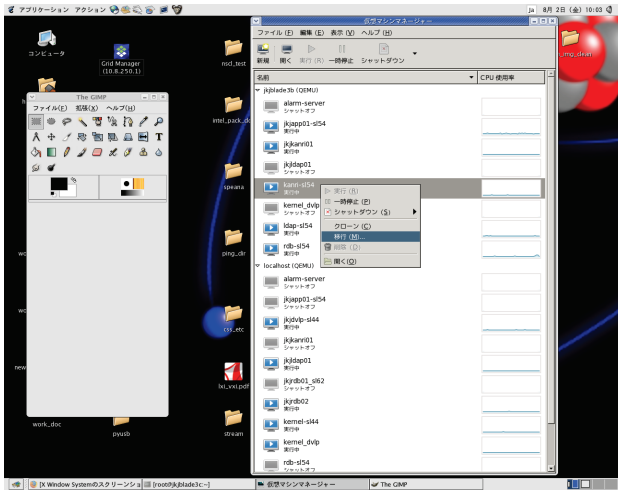

Figure 4: Management tool for virtual machines.

2.3 仮想環境の障害

サーバ系の仮想環境運用を開始した 2012 年は、仮想マシン（子）の OS に親サーバ機と同じ SL6.0 を使用していた。ところが、保守や故障でネットワークの一時停止が起こった後、仮想マシンの NFS が正常に復帰せず、read-only となる現象が見られた。この現象は、仮想でない SL6 計算機では起こらない。また、不調となった仮想マシンを復活させるには再起動が必要であり、サーバ系サービスにとっては大きな不具合であった。

この NFS 不調問題は、仮想マシンの NFS パラメータ調整などを試みたが効果が無く、原因ははっきりしていない。SL6 の NFS が v3 から v4 に改修された余波のように思えるが、情報収集を続ける。現時点では、仮想マシンの OS をあえて SL5.4 に downgrade し、問題を回避している。

2013 年 2 月 9 日、vioc 系の親サーバ機の 1 台(jkjblade3f)が突然停止する事件があった。その親サーバ機で稼動していた仮想マシンは全滅したため、別の親サーバ機でそれらの仮想マシンを手作業で再起動した。この作業には約 3 時間かかった。このように、親サーバ機が死んだ場合の対応は自動ではない。当面は手作業での対応手順の効率化検討を行うが、さらなる信頼度を期待するならば、自動で親を移動する仕組みなどが望まれる。

なお、この時の親サーバ停止の原因を追及すると、kernel config の soft_lockup 設定に行き当たった。仮想マシンが親から実 CPU を割り当ててもらえない場合、SL6.0 の初期設定では親がいきなり落ちる仕様になっている。親サーバ機全部で soft_lockup の見直し・再設定を行い、その後半年間この問題は発生していない。

4. まとめ

J-PARC MR では、計算機制御に仮想環境 KVM を導入した。2011 年から IOC 系仮想マシンを、2012 年からサーバ系仮想マシンの運用を始めた。

いくつかの障害を経験したが、おおむね運用は順調である。仮想の親サーバ機が停止した場合の復帰手順の確立が望まれる。

参考文献

[1] N.Kamikubota, et al., “J-PARC CONTROL TOWARD FUTURE RELIABLE OPERATION”, ICALEPCS2011, Grenoble, France, Oct. 2011, p.378-381
[2] 根本弘幸、他、”J-PARC MR 加速器制御システムにおける IOC 統合管理”、加速器学会（大阪）、Aug. 2012、p.745-747
[3] N.Kamikubota, et al., “VIRTUAL IO CONTROLLERS AT J-PARC MR USING XEN”, ICALEPCS2011, Grenoble, France, Oct. 2011, p.1165-1167
[4] 上窪田紀彦、他、”J-PARC MR 制御計算機の進展”、加速器学会（大阪）、Aug. 2012、p.741-744