# IMPROVEMENT OF COMPUTER SYSTEMS FOR J-PARC MR CONTROL

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)], Hidetoshi Nakagawa[A)]
[A)] J-PARC Center, KEK and JAEA, 2-4 Shirakata Shirane, Tokai-mura, Ibaraki, Japan, 319-1195
[B)] Kanto Information Service (KIS), 8-21 Bunkyo, Tsuchiura, Ibaraki, Japan, 300-0045
[C)] ACMOS Co. Ltd., 2713-7 Muramatsu, Tokai-mura, Naka-gun, Ibaraki, 319-1112

*Abstract*

In J-PARC MR, major components of computer systems for MR controls were introduced in 2008. Since them, they have been improved year be year. In this report, we focused on three components; sever CPU, control network, and disk system. Reviews of 5-year operation and future perspectives will be given.

# J-PARC MR制御計算機システムの進展

#

## １．はじめに

J-PARC MR加速器は、2008年5月にビーム運転を開始した。その後着実に加速器性能を向上しつつ、実験施設（HadronおよびNeutrino）へのビーム供給を続けてきた。2011年3月の東日本大震災も乗り越え、現在(2012年)は本格的な実験施設向けビーム供給を行っている[1]。

J-PARC加速器制御システムはEPICSで整備された[2]が、本報告ではMR制御システムの計算機インフラ部分に注目する。MR向け制御計算機資源には、計算サーバ、制御ネットワーク、ディスクシステム、などがあるが、いずれもビーム運転開始直前の2007年頃に集中して導入された[3]。これらの計算機資源が過去5年間どのように使用されまた増強されたか、経験を総括して報告する。また、次の5年を見据えた進展の方向について議論する。

## ２．計算サーバ（CPU）

### 2.1 概要

主要な計算サーバ(CPU)として、MR制御用にはIBM社Blade serverを採用している。2007年にHS20型Blade 5枚で運用を開始した。現在（2012年）は、総数29枚（HS20型5枚、HS21型13枚、HS22型11枚）にまで増加した。なお、IOC（Input Output Controller）のCPUについては[4]を参照されたい。

表１に、現在（2012年）とMRビーム運転開始時（2008年）に運転に使用されたBladeの型・役割・台数を示す（予備・特殊用途のものは除いた）。３種類の型それぞれで各Bladeは全く同等の性能・環境を持たせ、局所的に負荷の高いBladeがあれば、アプリやサーバを別のBladeに移動させて負荷を平滑化できる。このような体制で、運転・開発業務の需要を満たしてきた。Bladeの運転時CPU負荷について、2010年時点の報告がある[5]。

表１　現在(2012)と運転開始時(2008)の主要Blade

| | 型 x台数 | 役割 |
|---|---|---|
| **2012** | HS20 x2 | サーバ用(web, login)<br>2008から存命、HS21に移行予定 |
| | HS21 x8<br>HS22 x3 | SL4.4/SL6.2 アプリ用 |
| | HS22 x6 | SL6.2 仮想マシンの親<br>各種サーバ、vioc [6]、など |
| | HS21 x3<br>HS22 x2 | SL4/SL5 RCS simulation用 |
| **2008** | HS20 x4 | サーバ用 (web, RDB, ldap, dhcp,..) |
| | HS21 x9 | SL4/SL3 アプリ用 |
| | HS21 x2 | SL4 RCS simulation用 |

HS20 = Pentium D Xeon64-LV 2.8GHz / 1GB-memory
HS21 = DualCore Xeon 5110 1.6GHz / 2GB-memory
HS22 = 4Core2Cpu Xeon E5504 2GHz / 20GB-memory

### 2.2 進展（5年間の運用で出てきた新しい方向）

仮想OS環境を利用して1台の計算機に多数の仮想マシンを搭載し、計算機資源を有効利用することが出来る。我々のサーバ計算機の主力OSであるSL6ではKVM（仮想OS環境）が標準装備されている。2008年、主たるサーバ（web, RDB, ldap, dhcp, ..）は1 Bladeで1機能づつ稼動していた。現在（2012年）、サーバ機能のほとんどは親仮想マシン（数台）で稼動する仮想マシン（多数台）で動作している。また、仮想マシン上で動作するEPICS仮想IOC (vioc)がMR加速器の運転に投入されている[4,6]。

MR加速器の大強度化に伴い、ビームの振る舞いやロスのシミュレーション計算が重要になってきた[7]。特殊な仕様（巨大メモリ/48GBや多重Core/64coreなど）が必要なため汎用仕様のBladeになじまず、2010年からシミュレーション専用の計算機が導入されている。現在(2012年)は、2機種で計10台まで台数が増えた。

# E-mail: norihiko.kamikubota@kek.jp

### 2.3 総括

最初にBlade型サーバ計算機を選択し、需要と予算に応じて臨機応変に枚数を増やした戦略は良かった。今後は初期モデルのHS20型を退役させ、OSは初期のSL4から最新のSL6へ移行を進めていく。一方、シミュレーション用計算機の比重が増しているが、機種統一するなど管理を楽にする方向を模索していきたい。

## 3．制御ネットワーク

### 3.1 概要

J-PARC加速器制御用ネットワークは、先行するLINAC向けの整備が2004年頃から始まり、Extreme社のBlack-DiamondとSummit-switchの構成が採用された。MR向けは、既存部を拡張する形で2006年から整備を始め、2008年夏にはHadron棟・Neutrino棟を含むMR加速器全エリアをカバーするに至った。

2008年夏、MR各建屋（5ヶ所=D1,D2,D3, HD, NU）に各1台の24port switch（計5台）、中央制御棟に2台の48port switch、を配備した。LINACと共通の冗長コア(Black-Diamond、中央制御棟)は、リング構造により1ヶ所の故障がシステムを止めないよう設計されている。冗長コアから各switchはスター形接続（各switch2系統の配線、計1Gbpsの容量）である。

### 3.2 5年間の障害事例

2008年以降の5年の運用で、印象的な障害を記す。

(1) Neutrino switch障害：Neutrinoへは2009年からビームを供給している。NU棟のswitchが2010年3月, 2010年5月, 2011年8月, 2012年4月と、3年の運用で計4回故障した。いずれもfirmwareが暴走するなどの症状である。このswitchは末端部なので、故障するとNU棟全域が全滅する。2012年4月の故障では加速器運転を2時間35分停止させ（Neutrino実験再開まではもっと長時間かかった）、深刻な影響を及ぼした。繰り返し故障するのはNU棟特有の問題があると思われるが、原因は判明せず、対応は電源系統の確認など対症療法に留まる。

(2) MR server用switch障害：MR制御用の計算サーバ（Blade server）やディスクシステムのLAN配線を、当初トラフィックの閉じ込めを期待して1 switchに集約した。2011年10月、このswitchが故障し、MR制御システムは約5時間壊滅状態となった。このswitchは2010年に導入されたが、故障の数時間前に何回かrebootを起こした以外の事前兆候はなかった。

(3) 想定外の異常トラフィック：2012年2月、保守作業のミスで光メディコンで光回線をloopさせ、broadcast増殖によりMR制御ネットワークが麻痺した。このケースでは、光回線のloop先が別建屋（しかも放射線管理区域）で、発見までに4時間かかった。2012年5月、LINACトンネル内のネットカメラが故障し、大量のパケットを吐き出してLINAC制御を停め、（図１）その後加速器運転再開まで約5時間かかった。放射線による経年損傷と見られる。2012年夏のネットワーク機器更新では、各switchでloopや故障による大量trafficを検出してportを遮断する機能を組み込んだ。

(4) 冗長を無効にする温度障害：（3.5節参照）

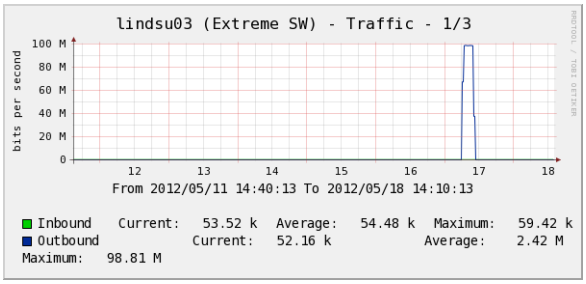



図１　2012年5月の故障カメラのbroadcast

### 3.3 運用時のネットワークトラフィック量

MR制御ネットワークのトラフィックは、2010年時点ではMR各電源棟（D1,D2,D3）で20-60Mbpsと報告されている[5]。2012年6月、ビーム運転中のNeutrino向け(FX)とHadron向け(SX)の両運転モードでのトラフィック量を計測した（表２）。また、同じ時期、FX運転時のディスクシステムへのデータ転送量はread/writeで55Mbps/60Mbpsであった。

表２　2012年6月 加速器運転時トラフィック(bps)

| 建屋 | FX (Neutrino向け) User-run 160kW | SX (Hadron向け) User-run 3.5kW |
|---|---|---|
| **D1** | -> CER 140 M<br><- CER 5 M | -> CER 70 M<br><- CER 5 M |
| **D2** | -> CER 90 M<br><- CER 6 M | -> CER 160 M<br><- CER 5 M |
| **D3** | -> CER 95 M<br><- CER 22 M | -> CER 45 M<br><- CER 15 M |
| **HD** | -> CER 3 M<br><- CER 0 M | -> CER 4 M<br><- CER 4 M |
| **NU** | -> CER 17-25 M<br><- CER 5-7 M | -> CER no data<br><- CER no data |
| **CER03** | -> out 230 M<br><- in 460M | -> out 230 M<br><- in 400M |

CER=中央制御棟計算帰室、CER03=計算機室のswitch

FX運転時のトラフィックの流れを図２に示す。MR各電源棟から中央への転送量はそれぞれ~100Mbpsに達し、合流する計算機室switch (CER03)では200-500Mbpsのデータが流れている。

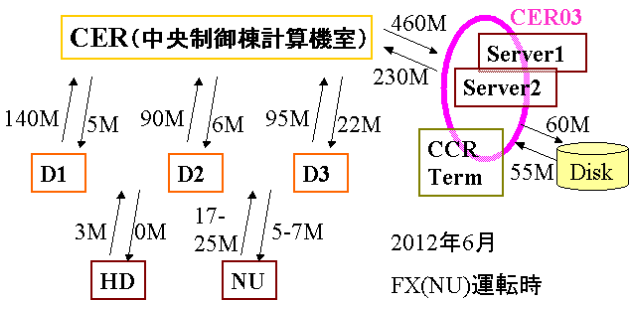



図２　FX運転時のトラフィックの流れ

### 3.4 Switch機種更新と故障を想定した工夫

制御switchは信頼性の高い機器のはずであるが、MRに限っても何度も故障している。冗長機能でシステムは停止しなかったが、3.2節で紹介した以外にも2例のswitch故障がある。

LINACの機器は2011年から順次機種更新されていることから、MRの機器も2011年後半～2012年前半に最新機種への更新を進めた。MRでは各建屋switch 1台から2台として故障時の回避ルートを確保するとともに、中央制御棟との回線を1Gbpsから10Gbpsに増強した（HD, NUを除く）。MR建屋（D1,D2,D3）で各2台（計6台）の48port switch、HD,NU建屋で各2台（計4台）の24port switch、中央制御棟に3台の48port switch、を配備した。一連の更新の最後となった2012年7月の更新・増強の内容を図３に示す。

これら更新・増強と平行し、MRで運転に使用するswitchは常に2台ずつpairを組ませるよう工夫した。例えば計算機室のサーバ用switch pairでは、一方は奇数番portのみ（他方は偶数portのみ）を使用し、奇数番のBladeのみ（他方は偶数番のBladeのみ）を接続する。仮に1台のswitchが停止しても残った他方で加速器運転を継続できる。また、ラック上下のLAN配線の移動だけで仮復旧できる（図４参照）。

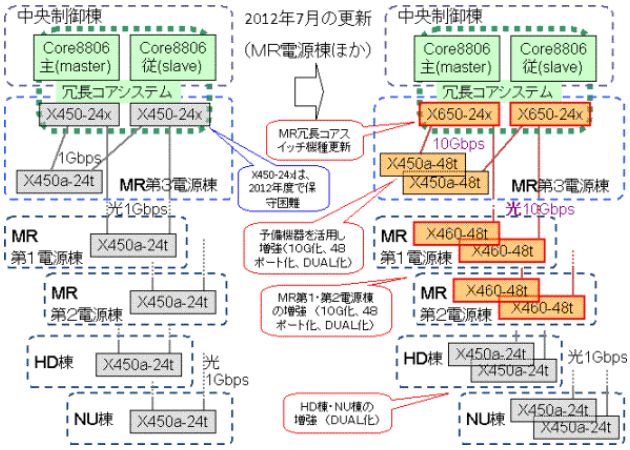
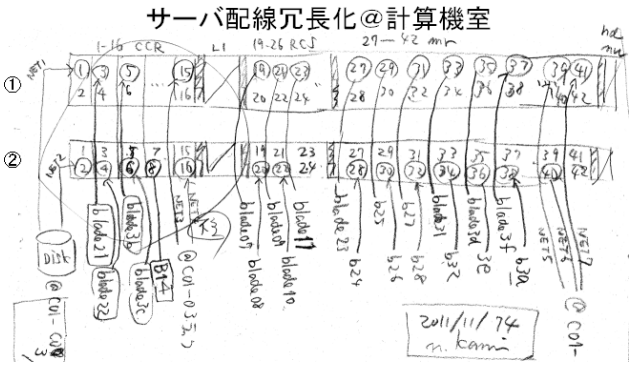



図３　2012年7月のMR制御ネット機器更新・増強




- Blade奇数番は①switch、偶数番は②switchに配線
- 一方が死んでも他方で片肺運転が可能
  - 死亡時はラック内の配線変更のみで仮復旧できる

図４　Switch pairでのサーバ配線工夫による冗長化

### 3.5 空調によるネットワーク障害

3.4節で述べた徹底的な冗長化の工夫で、1ヶ所の故障では加速器運転が止まらない、また故障しても復旧しやすい環境が実現した、と考えていた。ところが2011年11月、計算機室の空調整備（2式のうち1式）が停止して室温が上がり、冗長を組んだ複数のswitchが揃って同時に落ちるという事件が起こった。建設から数年の時間がたって計算機室の熱源が増え、1式のみの空調では冷却能力不足となっていた。3式目の空調を追加することを検討している。

さまざまな検討や対処で、ネットワーク障害の頻度を下げる努力を尽くしてきたつもりであった。つくづく現実は想定を超えることを思い知らされた事例であった。

## ４．ディスクシステム

### 4.1 概要

MR制御用のディスクシステムとして、2008年3月にIBM N3600 (NetApp)を導入した。それまでのDell server (PE2650)の250GBから9TBになっただけでなく、拡張性・信頼性も格段に向上した。ディスクシステムが提供するのは、(a) 運転アプリ領域(/jk)、(b) User領域(/home)、(c) データ領域(/jkdataなど)、である。MR運転開始時(2008年)と現在(2012年)の容量を表３に示すが、データ領域が増設されているのが分かる。このディスクシステムをmountしているNFS client数は、MRサーバ 約40台、MR IOC 約150台、端末 約30台、その他 30台程度、と推定している。

表３　現在(2012)と運転開始時(2008)のDisk

| | アプリ・ホーム領域 | データ領域 |
|---|---|---|
| **2012** | /jk 1TB<br>/home 1TB | /jkdata 11TB<br>/jkdata2 8TB<br>/jkdata3 7TB |
| **2008** | /jk 1TB<br>/home 1TB | /jkdata 7TB |

### 4.2 運用（データ量と障害事例）

MR加速器では、加速器機器のdata archive systemが整備されている[8]。Archiveの中で、各種波形データ（DCCT, BPM, SX-Spill, など）を記録するcadumpデータがディスクを消費する。2012年、ビーム運転中のNeutrino向け(FX)とHadron向け(SX)の両運転モードでディスク消費率を計測した。FXでは50GB/day（2012年5月、5月1ヶ月で1.5TB）、一方SXでは 110GB/day（2012年2月）であった。

SX運転モードで加速器運転中の2012年2月、ディスクシステムは深刻な障害に陥った。この時、ディスク容量残量が逼迫したためビーム運転中に大量の不要データ削除を行ったが、通常20-30%のdisk controller CPU負荷が100％となって多数の計算機（NFS client）が反応しなくなった。ディスクシステム負荷が高くなりすぎたと考えられている。

4.3 将来の方向

MR運転の各種波形データをきちんとArchiveすれば、おおむね~2TB/月の量のディスク消費となる。これほど巨大なデータすべてを記録する意味があるのか、高機能のN3600に巨大データファイルを置く整合の悪さ、などの議論がある。

1年以上古いデータはあまりアクセスされないと考えられる。2012年4月、相対的に安価・大容量の新ディスクシステム（IBM X3630）を導入し、新しくデータ領域15TBを提供した。MR運転時は波形データをN3600に記録し、古くなったデータは運転休止時にX3630に移動する、というシナリオを策定した。15TBではせいぜい1年分なので、毎年継続してX3630相当のディスクを増やしていくことになる。さらに良いシナリオが無いか、今後も検討を続ける。

## 参考文献


[1] 佐藤洋一 他、”J-PARC Main Ring における大強度運転”、第９回加速器学会(this meeting)
小関忠、”J-PARC MR の運転状況”、加速器学会誌 2012 年 9 巻 1 号 p.30-40 (2012)

[2] N.Kamikubota, et al., “J-PARC Control toward Future Reliable Operation”, Proceedings of the ICALEPCS 2011, Grenoble, France, Oct. 2011, p.378-381
(EPICS homepage) http://www.aps.anl.gov/epics/

[3] 上窪田紀彦 他、”J-PARC 計算機制御システムのインフラ整備”、第 4 回加速器学会、和光、2007 年 8 月、p.378-380

[4] 根本弘幸 他、”J-PARC MR 加速器制御における IOC 統合管理”、 第９回加速器学会(this meeting)

[5] 高木誠 他、”J-PARC MR 制御システムの CPU 負荷とネットワークトラフィック”、第７回加速器学会、姫路、2010 年 8 月、p.693-695

[6] N.Kamikubota, et al., “Virtual IO Controllers at J-PARC MR using Xen”, Proceedings of the ICALEPCS 2011, Grenoble, France, Oct.2011 p.1165--1167

[7] 佐藤洋一 他、”J-PARCMain Ring 速い取出し運転のビームシミュレーション：実測ビームロスとの比較ベンチマーク、高繰り返しによるビーム増強の要点”、第８回加速器学会、つくば、2011 年 8 月、p.199-203

[8] N.Kamikubota, et al, “Data archive system for J-PARC main ring”, IPAC10, Kyoto, Japan, May 2010, p.2680-2682