Platform LSF
Version 9 Release 1.1

Release Notes for Dynamic Cluster

GI13-3417-01

Platform LSF
Version 9 Release 1.1

Release Notes for Dynamic Cluster

GI13-3417-01

Note
Before using this information and the product it supports, read the information in "Notices" on page 11.

First edition

This edition applies to version 9, release 1 of IBM Platform Dynamic Cluster (product number 5725G82) and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Release Notes for IBM Platform Dynamic Cluster

About IBM Platform Dynamic Cluster 9.1.1

IBM Platform Dynamic Cluster (Dynamic Cluster) is an add-on product for Platform LSF (LSF) that can turn a static LSF cluster into a dynamic compute environment capable of optimizing the characteristics of resources based on workload demand by changing operating systems, running reliability-critical workload in virtual machines and performance-critical workload on physical machines. It has the following benefits:

- Job mobility
  - Optimize resources by defragmenting them. For example, if there are two hypervisors in a cluster, each with 1 GB of memory free, you cannot run a 2 GB job unless you defragment the hypervisors.
- Dynamic resource provisioning
  - Create new virtual machines to meet job demand.
  - There is no need to keep dedicated legacy physical machines.

System requirements

This section describes the system requirements for installing and running Dynamic Cluster. For detailed installation steps for Dynamic Cluster, see the *Setup* section of the *Using IBM Platform Dynamic Cluster* guide.

Supported hypervisor operating systems are as follows:

- RHEL 6.3 KVM with the following patches:
  - `kernel-2.6.32-279.14.1.el6`
  - `libvirt-0.9.10-21.el6_3.5`
  - `qemu-kvm-0.12.1.2-2.295.el6_3.2`
- VMware 5.x

Supported virtual machine guest operating systems are as follows:

- RHEL 4.x, 5.x, and 6.x (64-bit).
- Windows 7 (32-bit)
- Windows 2008 (64-bit)

Compatibility notes

To use Dynamic Cluster 9.1.1, you must complete a fresh installation of LSF 9.1.1. Dynamic Cluster can be enabled for some or all of the hosts in an existing cluster. The Platform Cluster Manager Advanced Edition 3.2.0.2 package must be installed to manage the virtual machines.

Host roles and requirements

Hosts in the Dynamic Cluster system have three roles:

- Platform LSF (LSF) master host: A few LSF master binaries have been enhanced to support Dynamic Cluster functionality. In Dynamic Cluster 9.1.1, the only supported master host platform is linux2.6-glibc2.3-x86_64.

- IBM Platform Cluster Manager Advanced Edition (Platform Cluster Manager Advanced Edition) manager host: This is the host with the Platform Cluster Manager Advanced Edition manager package installed. The supported platform is linux2.6-glibc2.3-x86_64. Minimum hardware requirements are as follows:
  - CPU: Single- or dual-core 2.0 GHz.
  - RAM: 4 GB.
  - Local disk space: 50 GB
  - Operating system: RHEL 5.x and 6.x.
- Dynamic Cluster host: The Dynamic Cluster host is a special LSF compute host that can run a hypervisor operating system (KVM or VMware). The Platform Cluster Manager Advanced Edition agent package is installed on the KVM hypervisor, allowing virtual machines to be created when required to run a workload. LSF binaries are installed in the virtual machines.

What is in Dynamic Cluster

Functionality

Workload-driven provisioning

Dynamic Cluster supports workload-driven virtual machine provisioning in the LSF cluster. Without Dynamic Cluster, LSF finds suitable resources and schedules jobs, but the resource attributes are fixed, and some jobs may be pending while resources that do not match job requirements are idle. With Dynamic Cluster, idle resources that do not match job requirements are repurposed, so that LSF can schedule the pending jobs. Dynamic Cluster can provision the machine type that is most appropriate for the workload. Jobs are contained in virtual machines (VMs) for greater flexibility. VM memory and CPU allocations can be modified when powering them on.

Restrict resource usage with VMs

Users of HPC applications cannot always predict the memory or CPU usage of a job. Without Dynamic Cluster, a job might unexpectedly use more resources than it asked for and interfere with other workloads running on the execution host. When Dynamic Cluster jobs run within virtual machine containers, one physical host can run many jobs, and each job is isolated in its environment. For example, a job that runs out of memory and fails will not interfere with other jobs running on the same host.

Live migration of VMs

VMs (and jobs running on them) may be migrated from one hypervisor host to another. VMs might be migrated for a number of reasons:

- Shut down a hypervisor for maintenance purposes or to save power
- Balance CPU and memory usage of hypervisors to improve VM performance
- Free up high priority high priority resources without killing jobs
- Save or restore VMs

The system releases all resources normally used by the job from the hypervisor host, then migrates the job to the destination host without any detectable delay. During this time, the job remains in a RUN state.

VM job checkpoint and restart

Checkpointing enables Dynamic Cluster users to pause and save the current state of memory and disk of a VM running a job to a separate set of files. The checkpoint files allow users to restart the VM job on the same physical server or a different physical server running the same hypervisor so that it continues processing from the point at which the checkpoint files were written.

When a user submits a VM job with a checkpoint, Dynamic Cluster saves the current state of the VM ("checkpoints") at the initial specified time, and repeats the process again when the job reaches the specified time interval. Dynamic Cluster only checkpoints the VM job when the job is running, so if the job state changes during the checkpointing period (for example, the job is finished, killed, or suspended), Dynamic Cluster does not checkpoint the VM job. Dynamic Cluster only keeps one checkpoint file for each job. If Dynamic Cluster checkpoints a VM job multiple times, the newest checkpoint file always overwrites the last checkpoint file.

Dynamic Cluster automatically restarts a checkpointed VM job only when the job status becomes `UNKNOWN` and **`lim`** reports that the VM is `unavail`. When restarting the VM job, Dynamic Cluster restores the VM from the last checkpoint. If there is no checkpoint for the VM job yet, LSF kills the job and requeues it, where it is rerun from the beginning with the same job ID. To use checkpointing, LSF must be able to rerun the job, either by submitting it to a rerunnable queue or by using the **`bsub -r`** option.

It is possible for the VM to be down but the physical execution host to be available. Therefore, it is possible for the VM to reschedule on the same execution host. To reduce the chance of repeating the failure, Dynamic Cluster places the original execution host at the end of the candidate host list, so that Dynamic Cluster attempts to reschedule the job on other execution hosts first.

If Dynamic Cluster fails to create a checkpoint, the VM job continues to run and Dynamic Cluster attempts to create another checkpoint at the next scheduled checkpoint time. The last successful checkpoint is always kept regardless of subsequent checkpoint failures.

If the VM could not be restored from the last checkpoint, the VM job cannot be restarted, so the job status remains UNKNOWN. If this occurs, LSF will continue attempting to trigger a restoring provision action.

New or changed commands

bdc

Provides a set of sub-commands to monitor Dynamic Cluster:

- **`vm`**: Displays information about VMs.
- **`host`**: Displays information about physical machines and hypervisors.
- **`tmpl`**: Displays information about Dynamic Cluster machine templates.
- **`action`**: Displays information about active provisioning requests and their status.
- **`hist`**: Displays historical information about provisioning requests.
- **`param`**: Displays information about Dynamic Cluster configuration parameters.

bjobs and bhist

These commands now display job information specific to Dynamic Cluster, including job virtual machine containers, job provisioning requests, and preemption or restore actions and status.

bsub

This command now includes the following job submission options specific to Dynamic Cluster:

- **-dc_tmpl** *template_name*: Specifies the name of one or more Dynamic Cluster templates that the job can use.
- **-dc_mtype vm**: Specifies the machine type for the Dynamic Cluster job. vm indicates a virtual machine.
- **-dc_vmaction** *action_name*: Specifies the action taken on the VM if the job is preempted:
  - **-dc_vmaction savevm**: Save the VM, which allows this job to continue later on.
  - **-dc_vmaction livemigvm**: Live migrate the VM (and the jobs running on them) from one hypervisor host to another.
  - **-dc_vmaction requeuejob**: Kill the VM job and resubmit it to the queue.
- **-dc_chkpntvm "init=***initial_minutes interval_minutes***"**: Enables VM job checkpointing by specifying an initial checkpoint time and recurring checkpoint interval.

Known Issues and Limitations

Live migration not triggered for a VM already live migrated once

Symptom

When a VM is already live migrated once, that VM may not be live migrated again. As a result, a high priority job may be pending in favor of a lower priority job.

Cause

This issue occurs because the low priority job keeps its original hypervisor allocation information and the live migration is canceled when its target host is matched to the original hypervisor allocation. A live migration is not triggered when the hypervisor allocated the first time is selected as a target hypervisor for a lower priority job.

Action

To work around this issue, reconfigure the LSF cluster by running the **badmin reconfig** command. After the reconfiguration is complete, the low priority job only keeps the current hypervisor allocation.

Reference

213378

"Failed to define domain" error when running provisioning actions on KVM hypervisors

Symptom

Provisioning actions on KVM hypervisors may fail with the following error message:

```
Register VM failed due to: Failed to define domain.
```

This error message is seen in the output of the **bdc hist** command.

Cause

This error occurs because KVM hypervisors that are under stress fail to register new VMs.

Action

To work around this issue, reduce stress on the KVM hypervisor by increasing the VM TTL (time to live) to an appropriately large value.

Reference

210295

Action failed with an RFITooManyOperationsException error

Symptom

An action or request failed with a `com.platform.rfi.manager.exceptions.RFITooManyOperationsException` error, which shows that the action or request was not accepted because there are too many errors in the system.

Cause

This error occurs because there is another root problem with the cluster. This root problem repeats because LSF continuously retries the provisioning action, causing the errors.

Action

To resolve this issue, use **bdc hist -n 0 -a -l** (ignoring the `com.platform.rfi.manager.exceptions.RFITooManyOperationsException` errors) to look for the root cause of the error, then fix the root cause.

Reference

212286

PVMO removes unavailable VMware hypervisors from the inventory

Symptom

When a VMware running on a VMware hypervisor crashes and the hypervisor is unavailable, PVMO will permanently remove the hypervisor from the PVMO inventory.

During this time, **bhosts** and **lshosts** shows the status of the hypervisor as dc_unknown.

Note that if the hypervisor is ok when the VM crashes, PVMO does not remove the hypervisor from the inventory.

Cause

PVMO permanently removes the hypervisor from the inventory due to technical issues with VMware.

Action

To work around this issue, you must manually add the hypervisor back to the inventory using the VMware vSphere Client.

Reference

211502

VMware live migration fails on memory-intensive programs

Symptom

When attempting a live migration on a VMware host that is running a memory-intensive program, the live migration may fail. Checking the live migration action shows the following error:

```
A general system error occurred: The migration was canceled because the amount of changing memory for the VM was greater than the available network bandwidth, meaning the migration was not making forward progress. Please attempt the migration again when the VM is not as busy or more network bandwidth is available.
```

Cause

This issue occurs because the program is using too much memory for the network to handle the live migration data.

Action

To work around this issue, kill the memory-intensive program on the VMware host before performing the live migration.

Reference

203450

VMware live migration does not support the migrationMaxTimeOut parameter

Symptom

If you specified a value for the **migrationMaxTimeOut** parameter, the host will remain timed out beyond this value if you perform a live migration on VMware hosts.

Cause

This is an issue with VMware because VMware does not let you change the down time of hosts.

Action

To work around this issue, ignore the **migrationMaxTimeOut** parameter when performing a VMware live migration.

Reference

203463

ActiveMQ has a java.io.EOFException error

Symptom

The PVMO log reports a `java.io.EOFException` error in ActiveMQ.

For example,

```
...
Oct 24 13:13:54 2012 ERROR [ActiveMQ Connection Executor: tcp://ondemand1/172.17.8.113:61616]
                     VMOManager - javax.jms.JMSException: java.io.EOFException
...
```

Cause

This is an issue with ActiveMQ that may occur in an unstable network environment. However, Platform Cluster Manager Advanced Edition recovers when the network is back up.

Action

To work around the issue, restart the VMOManager service:

```
egosh service stop VMOManager
egosh service start VMOManager
```

Reference

205476

Do not use short names for hypervisors

Symptom

If you installed the Platform Cluster Manager Advanced Edition agent on a hypervisor host with a short host name, the host status is `unavail` in LSF.

Cause

This is a known issue with Platform Cluster Manager Advanced Edition.

Action

To work around this issue, make sure that you always use long host names (that is, a fully-qualified host name) for hypervisor hosts.

Reference

206306

A VM shutdown operation failed because the timeout expired

Symptom

A provision request to shut down a VM failed. When viewing information on the provision request, you get the following error: `Request failed: Shutdown VM failed due to: Shutdown VM failed because timeout expired`.

For example, if the provision request ID is 123:

```
bdc hist -l -p 123
Provision request <123> for Job <321>
Sun Nov  4 22:43:03: Requested on 1 Hosts <hostA>; Power off 1 Machine; Power on 1 Machine
                     with Template <rhel62vm> Processors <2> Memory <2048 MB>
Sun Nov  4 22:43:04: Requested Power off 1 Machine <bbfbb1c5-44e3-4ee1-a39b-b60fb3358c22>
Sun Nov  4 22:58:24: Failed Power off 1 Machine <bbfbb1c5-44e3-4ee1-a39b-b60fb3358c22>
Sun Nov  4 22:58:24: Request failed: Shutdown VM failed due to: Shutdown VM failed because
                     timeout expired: 900 sec.
```

Cause

This issue occurs if the kernel crashed on the VM.

Action

Because this has no impact on the functionality or performance of Dynamic Cluster, you can ignore this issue.

Reference

206338

Live migration fails if libvirtd is restarted on the target hypervisor

Symptom

A VM live migration may fail if the **`libvirtd`** daemon is restarted on the destination hypervisor, and the error message is "cannot send data: Broken pipe". From the Platform Cluster Manager Advanced Edition Portal, the VM remains in `Migrating` status, while **`libvirtd`** shows the VM as undefined in the source hypervisor and paused in the target hypervisor.

Cause

This is a problem with **libvirt** because **libvirt** cannot recover from a restart during a live migration.

Action

To work around this issue, delete the VM that is affected by the live migration failure.

Reference

192701

RHEL 5.6 guest OS may fail to boot on a RHEL 6.2 hypervisor due to kernel panic

Symptom

A RHEL 5.6 guest OS may fail to boot up on a RHEL 6.2 KVM hypervisor due to kernel panic. The kernal panic may occur with the following error message:

```
Kernel panic - not syncing: IO-APIC + timer doesn't work! Boot with
apic=debug and send a report. Then try booting with the 'noapic' option
```

Action

To work around this issue, set the **noapic** boot option for the RHEL Linux kernel for the RHEL 5.6 guest OS.

Reference

198302

A socket operation failed after restarting the LIM

Symptom

When restarting the LIM (using **lsadmin limrestart** or **lsadmin reconfig**), the restart fails with the following error message:

```
initSock(): chanServSocketExt_(). A socket operation has failed on the
configured UDP port <port_number> on host <host_name>. Reason: <Address
already in use>. Fatal error. Either change the port number in lsf.conf
(LSF_LIM_PORT) or terminate the other process that is bound to the port.
```

Cause

This issue occurs when the **LSF_LIM_UDP_QUEUE** parameter is enabled in `lsf.conf`, which causes a conflict in the UDP port.

Action

To work around this issue, disable the LIM UDP port by specifying `LSF_LIM_UDP_QUEUE=n` in `lsf.conf`, then reconfigure the LIM by running **lsadmin reconfig**.

Reference

199585

Notices

This information was developed for products and services offered in the U.S.A.

IBM® may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web

sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
Intellectual Property Law
Mail Station P300
2455 South Road,
Poughkeepsie, NY 12601-5400
USA

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application

programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

LSF®, Platform, and Platform Computing are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Printed in USA

GI13-3417-01