Enabling Reliability Scalability Cluster Technology (RSCT) as clustering software for IBM PowerVM NovaLink

Nagendra Prasanna Dontamsetty
Santosh Chennuri
Sampath Sankoor

October 17, 2016

This article explains the Reliability Scalability Cluster Technology (RSCT) support to establish the formation of a management domain between IBM PowerVM NovaLink and its logical partitions (LPARs) automatically and to ensure that the communication between PowerVM NovaLink and the LPAR is possible over IPv6 link-local addresses using an OpenStack community driver on PowerVM NovaLink.

Introduction

IBM® PowerVM® NovaLink provides a PowerVM management interface running within a Linux® partition on the PowerVM system. This Linux partition acts as a host for the OpenStack compute drivers and other agents to provide an architecture for PowerVM, which is analogous to OpenStack with kernel-based virtual machine (KVM).

Similar to how Reliability Scalability Cluster Technology (RSCT) supports existing management consoles such as Hardware Management Console (HMC) and Integrated Virtualization Manager (IVM), it needs to support PowerVM NovaLink as a management console for logical partitions (LPARs) created using NovaLink (running inside the same system). This includes the automatic formation of a management domain between NovaLink and other LPARs and are needed to ensure that the communication between NovaLink and the LPARs is possible over IPv4 or IPv6 link-local addresses.

In RSCT, resource monitoring and control (RMC) plays a major role. RMC or management domain resource manager (MDRM) running on the LPARs continue to rely on the runtime abstraction services (RTAS) data to fetch NovaLink contact details and use them to form the management domain using IPv6 link-local address. In RSCT terminology, sometimes NovaLink is termed as management control point (MCP) and LPAR is termed as managed node.
What is RSCT

RSCT is a proven cluster software that can enable high availability and distributed management applications, such as Tivoli Storage Manager, IBM DB2®, PowerHA and so on. RSCT is a set of software components that together provide a comprehensive clustering environment for IBM AIX®, Linux, Solaris, and Microsoft® Windows®. RSCT is the infrastructure used by a variety of IBM products to provide clusters with improved system availability, reliability and scalability, and ease of use.

The following components are used in RSCT:

- Group services
- Topology services
- RMC subsystem
- Security services
- Cluster configuration manager

RSCT is used for building management clusters. A management domain is a set of nodes with resources that can be managed and monitored from one of the nodes, designated as MCP. All other nodes are considered to be managed nodes.

Management domain exploiter, such as HMC, IBM PowerVM NovaLink, IBM PowerLinux™ (big endian / little endian), IVM and so on.

RMC (which is a sub component of RSCT) is used for notifying changes to the cluster configuration. RMC is used to control operations, such as dynamic reconfiguration, Live Partition Mobility (LPM) and so on.

What is PowerVM NovaLink

Current distributed platforms have limitations in terms of centralized compute management, scalability, complex design, and so on, and this decreases agility and increases setup and maintenance cost. The new platform, PowerVM NovaLink, enables highly scalable modern cloud management and deployment for critical enterprise workloads using proven PowerVM solutions and OpenStack technology. PowerVM NovaLink can run on an Ubuntu Linux 15.10 (ppc64le) or later system.

NovaLink can be enabled on IBM POWER® processor-based managed systems (such as IBM POWER8® onwards), which provides management functionalities to support and create AIX and Linux partitions. Unlike HMC, NovaLink does not have a GUI.
Enabling Reliability Scalability Cluster Technology (RSCT) as clustering software for IBM PowerVM NovaLink

**Figure 1. Current PowerVM architecture**

**Figure 2. PowerVM NovaLink architecture**

**PowerVM NovaLink high-level architecture**

NovaLink is a partition running on IBM Power server, NovaLink runs on an Ubuntu 15.x or Ubuntu 16.x little endian partition on each system. It is designed for POWER8 processor-based servers
enabling reliability scalability cluster technology (RSCT) as clustering software for IBM PowerVM NovaLink only, and does not support backward compatibility. NovaLink services provide two primary components: Base-level virtualization interaction with IBM POWER Hypervisor™ and Virtual I/O Server (VIOS) as shown Figure 3. REST API is a virtualization interface with a K2-like schema that is heavily optimized and re-implemented wherever necessary. NovaLink can run services such as, OpenStack Nova compute, Neutron agents, and Ceilometer agents. VIOS instances and NovaLink partition can be sized during installation to minimize customer impact.

Figure 3. PowerVM NovaLink platform view
Figure 4. The NovaLink partition

NovaLink consists of the following three major components (as shown in Figure 4):

- NovaLink core – Communicates with virtualization resources
- NovaLink API – Provides an API that allows access to the POWER Hypervisor and VIOS instances
- OpenStack services – Runs the distributed Nova compute, Neutron agents, and Ceilometer agents for services such as IBM PowerVC and IBM Cloud Manager with OpenStack

NovaLink communicates with Power Hypervisor using the virtual management channel (VMC). Power Hypervisor supports only one virtual management channel connection. NovaLink communicates with all VIOS instance through RMC and the communication are over an internal, secure RMC network with the trunk adapter in the NovaLink partition. This ensures there is no interpartition communication.

For installation of ubuntu 15.x/16.x of the NovaLink, refer to Installing PowerVM NovaLink.

Software and hardware prerequisites

You need the following software to configure, set up, and install PowerVM NovaLink:
• POWER8 - 840 level firmware version SV840_050_050 or later
• OS – Ubuntu 15.10 or later
• RSCT: 3.2.1.2 or later

LPARs of the following operating systems can be configured for PowerVM NovaLink.

• AIX version 7.1.4 or later
• AIX version 7.2.0 or later
• Ubuntu 15.10 or later
• RHEL 7.1 or later
• SUSE 12 or later

You need to perform the following steps to start using NovaLink with RSCT support:

1. Set up the system for the NovaLink mode.
2. Create a partition with NovaLink and load the images on this partition.
3. Configure NovaLink host to use the new internal RMC communication.
4. Update the RSCT packages within the VM.
   - These packages include rsct.basic, rsct.core, rsct.core.utils, and any optional RSCT package that can be installed.
   - RSCT must be version 3.2.1.2 or later to contain the enhanced RMC support.

**Configuration of VMs**

This section provides a few examples of creating VMs, and assigning SCSI adaptors and virtual adaptors using PowerVM NovaLink.

To list the VIOS, run the following command:

```
root@r8r3m2Nova:~# pvmctl vios list
```

```
Virtual I/O Servers
+-------+----+---------+----------+------+-----+-----+--------------+
| Name   | ID |  State  | Ref Code | Mem  | CPU | Ent |
+-------+----+---------+----------+------+-----+-----+--------------+
| vios1  | 2  | running |          | 4096 |  2  | 1.0 |
+-------+----+---------+----------+------+-----+-----+--------------+
```

To create an LPAR, run the following command:

```
root@r8r3m2Nova:~# pvmctl LogicalPartition create --name r8r3m31 --mem 2048 --min-mem 512 --max-mem 2048 --proc-type dedicated --proc 1 --type AIX/Linux --sharing-mode "keep idle procs"
```

Notice that the LPAR shows the state as **not activated** after the creation of the LPAR.
Enabling Reliability Scalability Cluster Technology (RSCT) as clustering software for IBM PowerVM NovaLink

To activate the LPAR, run the following command:

```
root@r8r3m2Nova:~# pvmctl lpar power-on -i id=12
```

Powering on partition r8r3m31, this may take a few minutes.

Partition r8r3m31 power-on successful.

After activating LPAR, notice that the state is displayed as running.

Creating adapters

NovaLink communicates with all VIOS instances through RMC. Communications are over an internal, secure RMC network with the trunk adapter in the NovaLink host. This ensures that there is no interpartition communication. Using NovaLink, you can create any number of virtual Ethernet adapters.

To create an Ethernet adapter, run the following command:

```
root@r8r3m2Nova:~# pvmctl vea create -p id=12 --pvid 1 --vswitch ETHERNET0
```

To list all the Ethernet adapters created already, run the following command:

```
root@r8r3m2Nova:~# pvmctl vea list
```

To create a virtual adapter, run the following command:

```
root@r8r3m2Nova:~# pvmctl vea create -p id=10 --pvid 4094 --vswitch MGMTSWITCH
```

To list virtual adapters created already, run the following command:
Enabling Reliability Scalability Cluster Technology (RSCT) as clustering software for IBM PowerVM NovaLink

You need to make the following changes in RSCT to support PowerVM NovaLink:

**Figure 5. Management domain formation between MCPs and its LPARs**
• Establish a management domain formation between PowerVM NovaLink and its LPARs automatically
• Establish the communication between PowerVM NovaLink and LPAR over IPv6 link-local addresses

RTAS data

On LPARs managed by a NovaLink, the RTAS data is expected to contain an IPv6 link-local address listed under HMCAddIPv6s as the contact points for NovaLink. MDRM must use these supplied IPv6 link-local addresses to form the management domain between NovaLink and itself.

Run the following command to display the RTAS data:

```
root@r8r3m39:~# /usr/sbin/rsct/bin/getRTAS
Number of RTAS slots in use: 1
9:HmcStat=1;HscName=1*82B422A*10D71CT;HscHostName=r8r3m2Nova;HscIPAddr=fe80::8426:6ff:fefe:5481RMCKey=6825ba91d99d6dfd;RMCKeyLength=8;HscAddIPs=192.168.128.1,9.3.207.138;HMCAddIPv6s=fe80::848f:d4ff:fe84:8172,2002:903:15f:180:848f:d4ff:fe84:8172;
```

At present, to support RSCT in the management domain and peer domain clusters, it is required that the nodes are able to communicate over either IPv4 addresses or IPv6 global addresses. However, to support RSCT in PowerVM NovaLink environments, it is necessary for RSCT to support using IPv6 link-local addresses for communication.

RSCT components such as Phoenix Reliable Messaging (PRM) and RMC API are used in forming and maintaining the management domain that needs to be modified to support IPv6 link-local addresses.

Management domain

After the management domain is formed, RMC/MDRM running on NovaLink needs to continue monitoring the LPARs for healthiness and aliveness.

Refer to the following examples for the formation of a management domain having communication with different switches. If one of the switches break, RSCT would take care of the communication between NovaLink and its LPARs using other switch over IPv4 or IPv6 addresses.

Example 1: When switch (0) breaks, the communication of the interfaces between MCP (NovaLink) and the managed node (LPAR) is through switch(1) interfaces.

1. Management domain status from MCP (NovaLink) to managed node (LPAR):

```
root@r8r3m2Nova:~# /usr/sbin/rsct/bin/rmcdomainstatus -s ctrmc -a IP
Management Domain Status: Managed Nodes
  I A  0x8e6a3aa4e7d0c8dd  0008  fe80::4a3:c9ff:fea2:2079 (C)
  10.40.1.216 (C)
  fe80::e067:fbff:fe1d:82c6 (C)
  fe80::b0ac:93ff:fe42:6069 (C)
  2002:903:15f:180:e067:fbff:fe1d:82c6 (C)
```

2. Management domain status from managed node (LPAR) to MCP (NovaLink):
Example 2: When switch (1) breaks, the communication of the interfaces between MCP and the managed node is through switch (0) interfaces.

1. Management domain status from MCP (NovaLink) to managed node (LPAR):

   root@r8r3m2Nova:~# /usr/sbin/rsct/bin/rmcdomainstatus -s ctrmc -a IP

   Management Domain Status: Managed Nodes
   I A 6xaaa6a3aa4e7d0cddd 0008 fe80::e067:fbff:fe1d:82c6 (C)
   10.40.1.216 (C)  
   fe80::4a3:c9ff:fea2:2079 (C)  
   fe80::c0ac:93ff:fe42:6069 (C)  
   2002:903:15f:180:e067:fbff:fe1d:82c6 (C)

2. Management domain status from managed node (LPAR) to MCP (NovaLink):

   root@r8r3m31_lpar:~# /usr/sbin/rsct/bin/rmcdomainstatus -s ctrmc -a IP

   Management Domain Status: Management Control Points
   I A 0x5dea90a1b9cccc29 0001 fe80::8426:6ff:fefe:5481 (C)
   9.3.207.138 (C)  
   fe80::84ff:d4ff:fe84:8172 (C)  
   fe80::848f:d4ff:fe84:8172 (C)  
   2002:903:15f:180:848f:d4ff:fe84:8172 (C)

IPv6 link-local addresses and scope ID

IPv6 link-local addresses are always associated with a scope ID. Some of the C APIs (such as getnameinfo and so on) support scope ID. However, this support is limited only to a few platforms (AIX is not one of them).

Programming using link-local addresses would involve sin6_scope_id to be filled in when calling the bind function. However, no need of supplying it if in6addr_any is used for filling sin6_addr, for example:

```c
serveraddr.sin6_addr = in6addr_any;
```

Scope ID for remote addresses

However, platforms such as AIX does not provide support for the NI_NUMERICSCOPE flag in the getnameinfo function. So, a new library interface is proposed to find a local scope ID for a given remote IPv6 link-local address.

Support for link-local addresses in RSCT (in management domain) needs the following change in different components.
Resource management control (RMC)

As of now, RMC daemon, if found to be running on an LPAR (AIX/PowerLinux) will start the RTAS look-up thread (poll_RTAS function). This thread performs the normal RTAS look up to notify MDRM about any changes in MCP configurations.

Similarly, if it is found to be running on an MCP, it starts a getcimstatus loop to initiate the necessary managed node domain cleanup push the Central Electronics Complex (CEC) key into RTAS (for every new CEC), and build an internal partition cache table.

- During the getcimstatus loop, MCP needs to identify whether a traditional HMC is connected to NovaLink. If connected, it needs to start an RTAS thread to allow it to be managed by the HMC.
- Continue to drive the getcimstatus loop to allow NovaLink to be the MCP for the LPARs created so far and for the ones to be created.
- While adding a node or the IP configuration about the managed node to PRM (using PrmDRCCAddNode/PrmDRCAddIP), if the managed node is found to be using the link-address, RMC will discover local scope IDs associated with the remote link-local address and make it available to PRM.

RMC API

RMC API provides session interfaces for opening a session with the remote RMC subsystem. The RMC API can contact the remote RMC subsystem to start a session.

IPv6 link-local addresses can be specified along with a scope ID in the format `<ip address>%<scope>`. The scope can be a numeric index or an interface name. However, at times when there is no scope ID mentioned for ‘name’, the sessions APIs will be able to discover the applicable scope ID using the library function.

Management domain resource manager (MDRM)

Existing design

1. As a part of the initiation of the public key exchange protocol, MDRM running on the LPAR, collects all the IP addresses of MCP/HMC through RTAS fields such as HscIPAddr, HscAddIPs and HMCAddIPv6s.
2. An RMC session is opened for each IP address collected in step (1). If successful, the local adapter's IP address through which this session was made possible is found.
3. After all the local IP addresses (mentioned in step 2) are collected, the information will be sent to MCP/HMC.
4. MDRM running on MCP/HMC now gathered the LPARs IP addresses (mentioned in step 3) and each LPAR uses this IP address to start an RMC session and identify the local adapter's IP address used for that.
5. MDRM running on HMC/MCP sends the local adapter's IP address gathered in step (4) to the LPAR by calling the p0nePublicKeyExchange class.

Changes proposed in MDRM to make work with IPv6 Link Local addresses
• No additional changes are expected in the overall design for exchanging the IP address between MCP and the managed node.
• Starting the public key exchange protocol would begin with opening an RMCAPI session with RMC running on the HMC with a supplied link-local address in the RTAS data. Because the scope ID is not mentioned in the RTAS data, MDRM would discover or harvest the scope ID before initiating the `mc_timed_start_session` session.
• MDRM has to avoid reading RTAS data of the NovaLink (in case NovaLink is managed by a traditional HMC) so that it can avoid creating a circular dependency.

Phoenix Reliable Messaging (PRM)

RMC initializes PRM with a bound socket. This socket is bound with `in6addr_any`, and therefore, no changes are needed in the PRM to receive messages from the remote nodes using the link-local addresses. However, the PRMAPI changes are needed in PRM while sending messages using IPv6 link-local addresses.

PRM API changes

When a new node needs to be added to the management domain, RMC uses a PRM interface called `PrmDRCAddNode`, which needs to be changed.

If any of the remote IP addresses supplied as a part of `pIPAddr` is a valid IPv6 link-local address, RMC discovers all the applicable scope IDs for this remote address and supplies them as part of `scope_ids`. For a non IPv6 link-local address, `scope_ids` is set to null and `no_of_scope_ids` set to zero.

Conclusion

You can use PowerVM NovaLink (similar to HMC) to configure, manage, monitor, and communicate with LPARs and interfaces. However, by enabling RSCT to PowerVM NovaLink, you can establish management domain formation between PowerVM NovaLink and its LPARs automatically over IPv6 link-local addresses.

References

• Introducing PowerVM NovaLink
• PowerVC: Simplified NovaLink RMC Connections
• Welcome to High Performance Computing (HPC) Central: RSCT

© Copyright IBM Corporation 2016
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)