DELL EMC

White paper

# PCIe Card Cooling with Dell EMC PowerEdge Servers

Updates based on servers with iDRAC9


## Abstract

An overview of thermal management of PCIe adapter cards in Dell EMC PowerEdge Servers and potential airflow customization options for third-party cards



December 2019

# Revisions

| Date | Description |
|---|---|
| 11/20/2019 | New white paper that consolidates all PCIe cooling related topics, including airflow customization options for third-party cards. Supported under iDRAC9 (latest available release recommended). Some features may have iDRAC license dependency. |

# Acknowledgments

This paper was produced by the following:

Authors: Hasnain Shabbir, Jon Brown

Support: Server Thermal Controls Team

# Table of contents

# Executive summary

The adoption of PCIe adapters, especially of customer-specific appliance solutions based on PowerEdge servers, is increasing. With this increase, it is important to highlight how PowerEdge servers address thermal management of PCIe adapters, especially of third-party cards. This paper also provides details on airflow customization options available for such cases. The following figure is a high-level workflow of how PCIe adapters are thermally managed—from systems management perspective—in PowerEdge Servers. The details of various sections of this flowchart are discussed in this white paper.

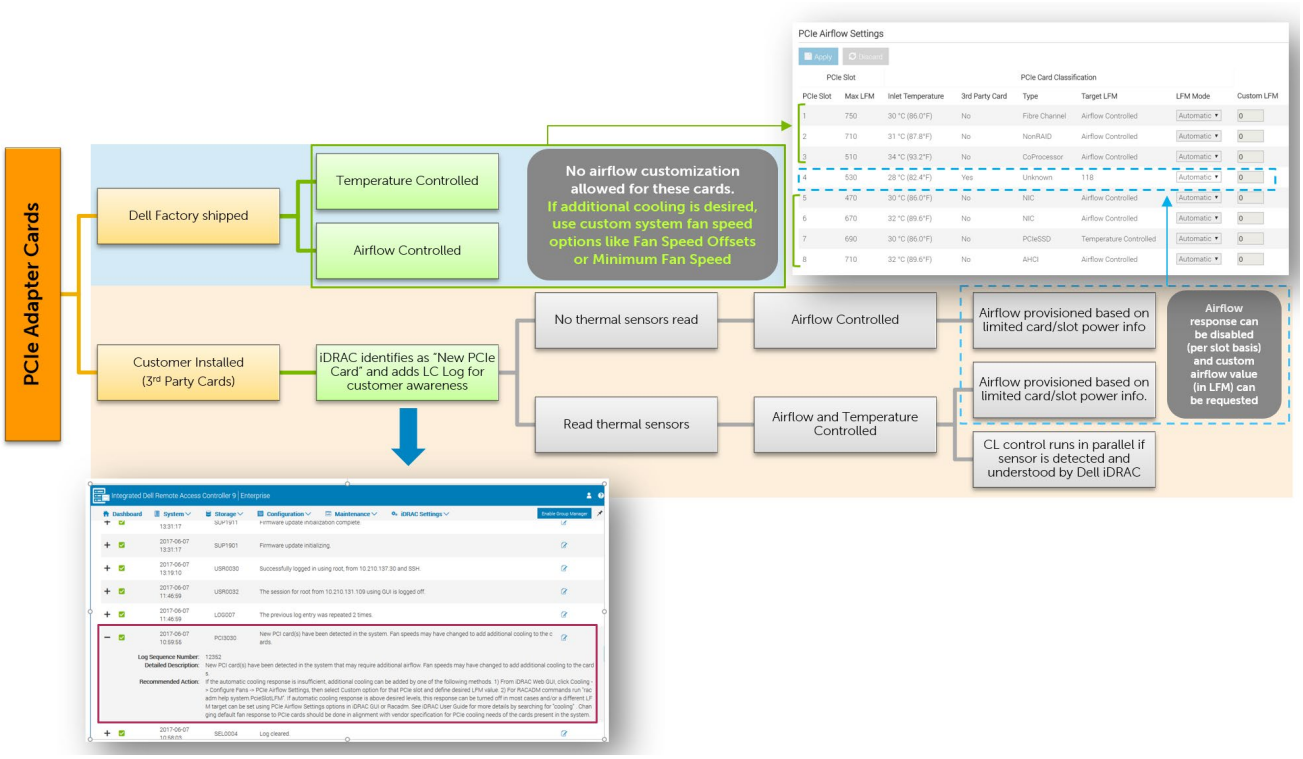



Figure 1 Third-party card airflow customization options

# 1 Introduction

PowerEdge servers have the cooling capacity to support a broad array of PCIe adapters. These servers use customized fans, airflow shrouding, and optimized system topologies to maximize airflow that is provided to PCIe slots. The design also ensures the temperature of the air that is delivered to the PCIe cards is at or below industry requirements.

Embedded within Dell EMC's industry-leading iDRAC systems management features, an intelligent thermal controls algorithm is used to reduce fan power consumption and acoustics without compromising component reliability or performance. This algorithm uses several temperature monitoring sensors and knowledge of the installed system hardware configuration to achieve higher efficiency.

As a feature of Dell EMC's thermal control architecture, your specific hardware configuration determines the baseline fan speed. The fan speed is set to cool components that do not have temperature monitoring. Fan speeds may never go below this level unless the inlet ambient temperature or system configuration changes. When inlet ambient temperature or hardware changes occur, the thermal control uses open-loop control algorithms to determine when to increase fan speeds to cool components whose temperatures are not monitored. The PCIe card responses that are described in this white paper can increase fan speeds above the baseline.

PCIe adapter vendors develop cards with a general 55° C inlet card temperature limit, or 45° C for a few adapters, and define the airflow cooling requirement for the card in terms of linear feet per minute (LFM). The vendors also ensure that the card cooling boundary conditions are within server vendor expectations and aligned with the PCIe cooling specifications of the server vendor. Recent updates to PCI-SIG (industry standard for PCIe devices) allow a card to specify a cooling tier, called *MaxTherm level*. Each level represents the LFM requirement for the card as a function of inlet air temperature to the PCIe card. When the vendors adopt this updated specification and report MaxTherm levels, Dell EMC's thermal controls algorithm can read that information from the third-party PCIe cards and provide the required LFM.

This document provides detailed guidelines on how PowerEdge servers manage airflow to third-party PCIe cards and airflow customization options where applicable.

# 2 PCIe cooling strategy in PowerEdge servers

PowerEdge servers identify the device and provide customized cooling for each Dell EMC-designed or qualified PCIe adapter card. Many customers choose to use their own cards and those cards are also known as third-party cards. Dell EMC has designed the logic to detect these third-party PCIe adapters and provide a default cooling response that is based on an estimate of the cooling requirements for the card.

At a high level, PowerEdge servers handle PCIe card cooling depending on whether the cards are Dell EMC-qualified or sourced from a third party.

## 2.1 PCIe airflow settings in iDRAC9

Given the increasing power consumption and heat advanced PCIe adapters produce, PowerEdge engineers developed an innovative solution for PCIe adapter cooling that gives users a slot-by-slot view of airflow delivered, and the capability to customize the delivery of airflow for their custom PCIe cards. This new solution is called PCIe Airflow Settings.

PCIe Airflow Setting in iDRAC represents the PCIe cooling setup, slot-by-slot, for the server. See the following figure for a visual representation of the slot's maximum LFM capability.

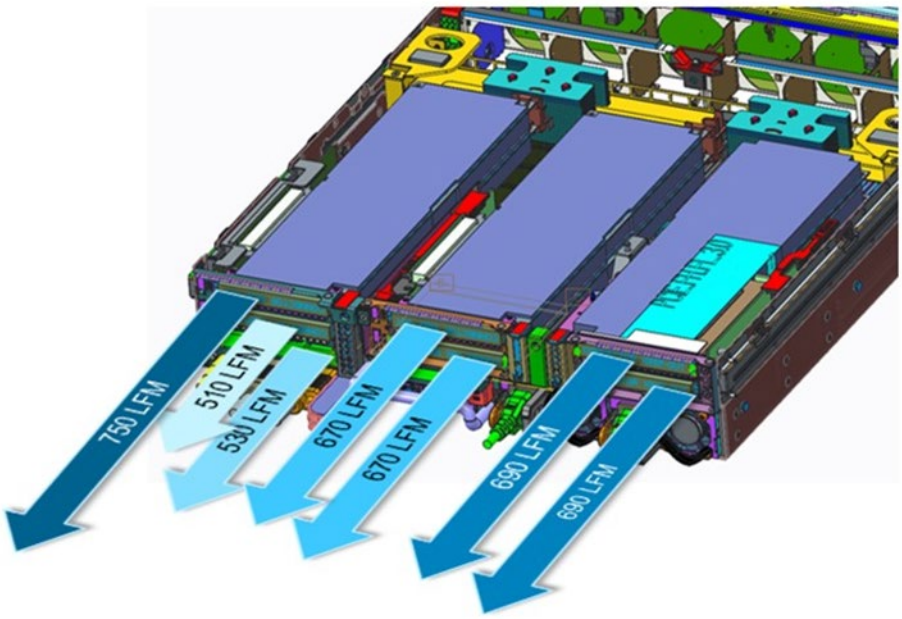
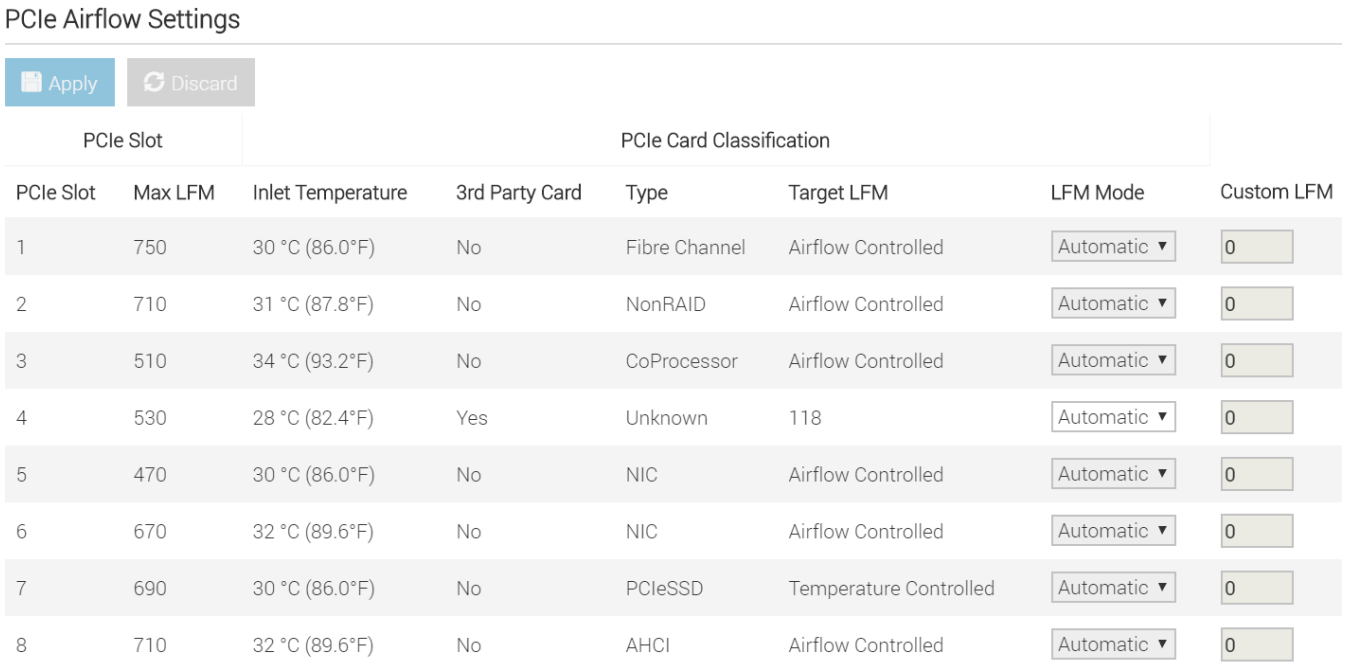



Figure 2 Visual representation of the slot-by-slot max LFM capability for each PCIe slot

See next figure for an example snapshot from the iDRAC UI and how it represents PCIe cooling status.


PCIe Airflow Settings

Apply Discard

| PCIe Slot | | PCIe Card Classification | | | | | |
|---|---|---|---|---|---|---|---|
| PCIe Slot | Max LFM | Inlet Temperature | 3rd Party Card | Type | Target LFM | LFM Mode | Custom LFM |
| 1 | 750 | 30 °C (86.0°F) | No | Fibre Channel | Airflow Controlled | Automatic | 0 |
| 2 | 710 | 31 °C (87.8°F) | No | NonRAID | Airflow Controlled | Automatic | 0 |
| 3 | 510 | 34 °C (93.2°F) | No | CoProcessor | Airflow Controlled | Automatic | 0 |
| 4 | 530 | 28 °C (82.4°F) | Yes | Unknown | 118 | Automatic | 0 |
| 5 | 470 | 30 °C (86.0°F) | No | NIC | Airflow Controlled | Automatic | 0 |
| 6 | 670 | 32 °C (89.6°F) | No | NIC | Airflow Controlled | Automatic | 0 |
| 7 | 690 | 30 °C (86.0°F) | No | PCIeSSD | Temperature Controlled | Automatic | 0 |
| 8 | 710 | 32 °C (89.6°F) | No | AHCI | Airflow Controlled | Automatic | 0 |

Figure 3 iDRAC web UI snapshot for PCIe Airflow Settings

The UI illustrated in the above figure enables the user to observe:

- The maximum LFM capability of each slot within the server when the max LFM is at full fan speeds
- How each slot is being managed—by airflow control, temperature control, or target LFM value
- The minimum LFM delivered to any third-party card slots allows users to decide whether to keep this LFM, or to customize its settings at a higher or lower value based on the card specifications
- PCIe inlet temperature distribution across the slots

This UI overview of the system PCIe cooling plan provides key information such as max LFM capability. With this information, the user can reconfigure their controller layout to achieve reduced system fan power consumption, ensuring, for example, that the card with the highest LFM requirement is installed in the slot with the highest LFM capacity. This feature is only available for third-party cards. For Dell EMC cards, LFM value is not displayed since Dell EMC's thermal algorithm automatically determines the optimal cooling requirements for Dell EMC cards.

The default system fan response for third-party PCIe cards can only be turned off one slot at a time to prevent cooling issues.

**NOTE: It is recommended that you do not turn off the default system fan response unless you have a good understanding of the PCIe adapter cooling requirements in your system.**

## 2.2 Dell EMC-qualified cards and optimal cooling response

PowerEdge servers identify Dell EMC-qualified cards through proprietary communications. All Dell EMC-designed cards and many EMC-qualified cards have temperature monitoring capabilities that are used to provide optimal cooling through closed-loop control. If system level thermal controls identify cards without these capabilities, it prescribes a predetermined fan speed response based on the airflow needs of the card.

Cards that are cooled based on temperature monitoring are called *temperature controlled,* while those adapter cards that require specific airflow only are called *airflow controlled*. These terms are used in iDRAC UI associated with PCIe airflow and are discussed later in this white paper.

Dell EMC-qualified cards have been thoroughly validated with PowerEdge servers and require no additional customization from a cooling perspective. For such cards, no slot-level customization options are available to change the airflow to them. A general fan speed increase can be used to achieve this goal by using custom fan speed options such as fan speed offset or minimum fan speed (MFS).

Note that there is no single PCIe adapter card temperature monitoring standard. General networking cards follow NC-SI standard, FGPA cards use PLDM specifications, and GPU card vendors like Intel, Nvidia, and AMD have their own proprietary methods defined. PowerEdge servers must have prior understanding of the temperature monitoring method used by the technology for them to enable closed-loop control. Hence the idea of using a Dell EMC-qualified or -branded card is important because Dell EMC uses prior knowledge from these cards to optimally regulate cooling to the cards.

## 2.3 Third-party cards and their thermal management

The automatic system cooling response for third-party cards provisions airflow based on common industry PCIe requirements, regulating the inlet air to the card to a maximum of 55°C. The algorithm also approximates airflow in linear foot per minute (LFM) for the card based on card power delivery expectations for the slot (not actual card power consumption) and sets fan speeds to meet that LFM expectation. Since the airflow delivery is based on limited information from the third-party card, it is possible that this estimated airflow delivery may result in overcooling or undercooling of the card. Therefore, Dell EMC provides airflow customization for third-party PCIe adapters installed in PowerEdge platforms.

### 2.3.1 How do I know that the system has identified a card as a third-party card?

If a card is identified as a third-party card upon installation, the following message is recorded in the LC log:

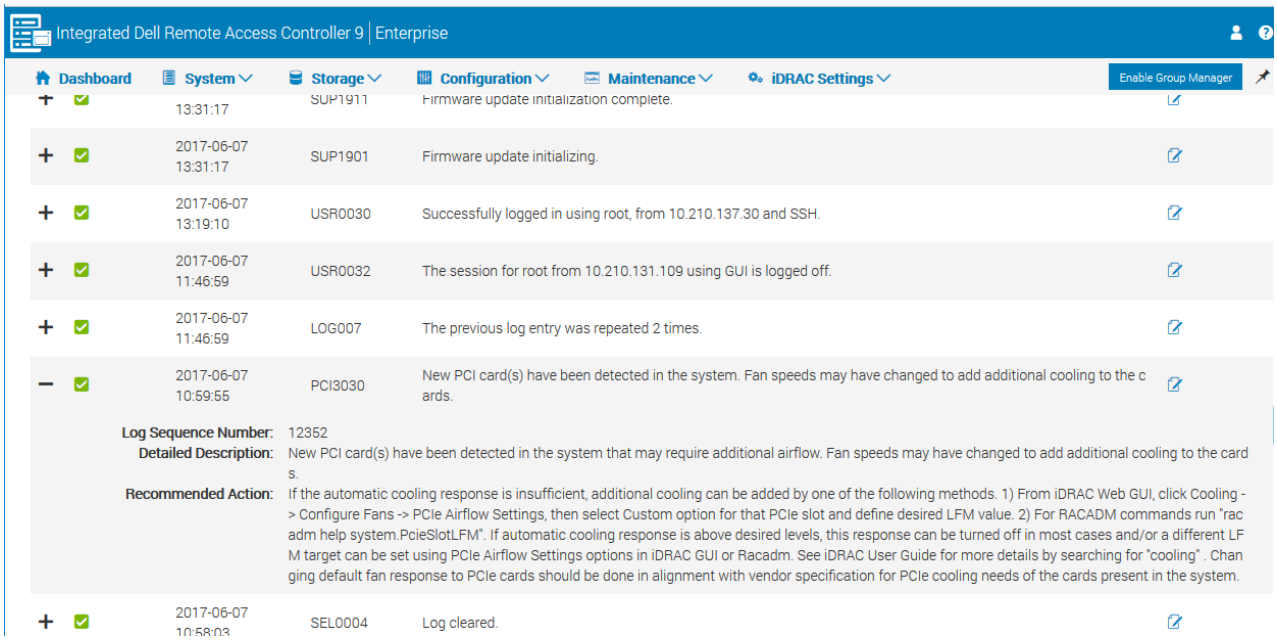



Figure 4 LC log message in iDRAC web UI if the system has at least one third-party PCIe adapter card

If the LC log does not record this message upon card installation, the server is not treating the card as a third-party card and no airflow customization can be done for the cards.

This log is informational only, and this message does not mean that there is a problem with the PCIe card. The intent of this message is to make the user aware that a third-party card is detected in the system and that Dell EMC has provisioned cooling to the card. However, if they determine that the system response is inadequate, over provisioned, or not required at all, the user may customize cooling settings

## 2.3.2 Manually customizing PCIe card cooling

The cooling requirements of a card may be different from the default system fan response, and if the default response is deemed insufficient there are additional options available to customize fan speeds. The iDRAC UI and RACADM interfaces both enable customization. Usually, the speed of all system fans changes for any one PCIe card airflow customization, and therefore it can result in additional airflow to other adapters as well. The goal is to dial in the minimum LFM that is being requested. The maximum LFM per slot is visible in all iDRAC interfaces and should be used as a guide to how much additional LFM can be added to a third-party card. Note that Dell EMC only allows third-party PCIe adapter default cooling responses to be disabled or customized on a slot-by-slot basis. There is no universal disable flag or customization option that allows configuration for multiple third-party cards.

### 2.3.2.1 Using the iDRAC UI

Within the iDRAC UI, customization of PCIe cooling can be accomplished for third-party cards only on a slot-by-slot basis. Again, referring to Figure 3, the UI enables the user to observe:

- The maximum LFM capability of each slot within the server when the max LFM is at full fan speeds
- How each slot is being managed in terms of airflow and temperature, from controls perspective.

The minimum LFM delivered to any third-party card slots allows users to decide on whether to keep this LFM, or to customize it to a higher or lower value based on the card specifications.

To change the LFM, select the **Custom** option under **LFM Mode**, and then type in a new LFM value. System fan speeds change to meet that minimum LFM requirement.

**NOTE: An increase in the system fan speed to meet this new LFM requirement also increases LFM to other cards, though the max LFM capability does not change for the configuration. The max LFM value changes only if the system configuration changes. For example, where hard disk drives in the front bay are uninstalled or added, or where a system with a different configuration within the same chassis has different limits.**

If a general increase is required in the cooling settings of all PCIe adapter cards, use Fan Speed Offset or Min Fan Speed options.

For example, a card that has its own cooling fan may not need additional cooling or may require less than the automatic response provides. If so, you can disable the automatic response for that slot to reduce fan power consumption and noise levels in the server. Disabling the response for that slot does not necessarily mean that fan speeds drops or that the slot does not get any system airflow at all. It means that the thermal algorithm ignores the airflow requirement of that slot, but incidental airflow from system fans running to meet the cooling needs of other devices or PCIe cards result in additional airflow to that slot.

#### 2.3.2.2 Using IPMI

There is no customer-facing support for configuring the third-party PCIe fan response through IPMI.

#### 2.3.2.3 Using RACADM commands

Use the following RACADM command to check the status of third-party PCI fan responses per slot. Replace the last digit which is PCIe slot number with the applicable slot number for your system.

```
racadm get system.pcieslotlfm.1
[Key=system.Embedded.1#PCIeSlotLFM.1]
#3rdPartyCard=Yes
#CardType=P2PBridge
CustomLFM=0
LFMMode=Automatic
#MaxLFM=700
#SlotState=Defined
#TargetLFM=118
```

The **LFMMode** property indicates how the fan speed for the slot is being controlled. This value can be **Automatic** (controlled by iDRAC), **Custom** (user-configured), or **Disabled.** The **TargetLFM** property shows the LFM that is provided to the slot. The **MaxLFM** property indicates the highest LFM value that can be provided to the slot.

To set a custom third-party PCIe fan response:

1) Enable the custom LFM mode.

```
racadm set system.pcieslotlfm.1.lfmmode custom
```

2) Set the desired LFM. The allowable range is 0-**MaxLFM**. **LFMMode** must be set to **Custom** before configuring the **CustomLFM**.

```
racadm set system.pcieslotlfm.1.customlfm 50
```

To disable the third-party PCI fan response:

```
racadm set system.pcieslotlfm.1.lfmmode disabled
```

To enable the automatic response if customized or disabled by above commands:

```
racadm set system.pcieslotlfm.1.lfmmode automatic
```

#### 2.3.2.4 Using WSMAN commands

Use the following RACADM command to check the status of third-party PCI fan responses per slot. Replace the PCIe slot number (in bold text) with the applicable slot number for your system.

```
winrm g
cimv2/root/dcim/DCIM_SystemEnumeration?InstanceID=System.Embedded.1#pcieslotlfm.1#1
fmmode -u:<username> -p:<password> -r:https://<idrac_ip>/wsman -SkipCNcheck -
SkipCAcheck -encoding:utf-8 -a:basic

DCIM_SystemEnumeration
    AttributeDisplayName = LFM Mode
    AttributeName = LFMMode
```

```
    CurrentValue = Automatic
    DefaultValue = Automatic
    Dependency = <Dep><AttrLev Op="OR"><ROIf Op="NOT"
Name="PCIeSlotLFM.n#3rdPartyCard">Yes</ROIf></AttrLev></Dep>
    DisplayOrder = 7
    FQDD = System.Embedded.1
    GroupDisplayName = PCIeSlotLFM
    GroupID = pcieslotlfm.1
    InstanceID = System.Embedded.1#pcieslotlfm.1#lfmmode
    IsReadOnly = false
    PendingValue = null
    PossibleValues = Automatic, Disabled, Custom
```

To set a custom third-party PCI response:

**Note: This feature may be tied to an iDRAC license upgrade.**

1) Enable the custom LFM mode:

```
winrm i SetAttribute
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_Comput
erSystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Na
me=DCIM:SystemManagementService -u:<username> -p:<password> -
r:https://<idrac_ip>/wsman -SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";AttributeName="PCIeSlotLFM.1#LFMMode";AttributeValu
e="Custom"}

winrm i CreateTargetedConfigJob
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_Comput
erSystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Na
me=DCIM:SystemManagementService -u:<username> -p:<password -
r:https://<idrac_ip>/wsman -SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";ScheduledStartTime="TIME_NOW"}
```

2) Set the desired LFM. The allowable range is 0-**MaxLFM**. **LFMMode** must be set to **Custom** before configuring the **CustomLFM**.

```
winrm i SetAttribute
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_Comput
erSystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Na
me=DCIM:SystemManagementService -u:<username> -p:<password> -
r:https://<idrac_ip>/wsman -SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";AttributeName="PCIeSlotLFM.1#CustomLFM";AttributeVa
lue="50"}

winrm i CreateTargetedConfigJob
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_Comput
erSystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Na
me=DCIM:SystemManagementService -u:<username> -p:<password -
r:https://<idrac_ip>/wsman -SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";ScheduledStartTime="TIME_NOW"}
```

To disable the third-party PCI response:

```
winrm i SetAttribute
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_ComputerS
ystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Name=DCI
M:SystemManagementService -u:<username> -p:<password> -r:https://<idrac_ip>/wsman -
SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";AttributeName="PCIeSlotLFM.1#LFMMode";AttributeValue="
Disabled"}
```

```
winrm i CreateTargetedConfigJob
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_ComputerS
ystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Name=DCI
M:SystemManagementService -u:<username> -p:<password -r:https://<idrac_ip>/wsman -
SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";ScheduledStartTime="TIME_NOW"}
```

To enable the automatic response if customized or disabled by above commands:

```
winrm i SetAttribute
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_Com
puterSystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementServ
ice+Name=DCIM:SystemManagementService -u:<username> -p:<password> -
r:https://<idrac_ip>/wsman -SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";AttributeName="PCIeSlotLFM.1#LFMMode";AttributeV
alue="Automatic"}
```

```
winrm i CreateTargetedConfigJob
cimv2/root/dcim/DCIM_SystemManagementService?SystemCreationClassName=DCIM_ComputerS
ystem+SystemName=srv:system+CreationClassName=DCIM_SystemManagementService+Name=DCI
M:SystemManagementService -u:<username> -p:<password -r:https://<idrac_ip>/wsman -
SkipCNcheck -SkipCAcheck -encoding:utf-8 -a:basic
@{Target="System.Embedded.1";ScheduledStartTime="TIME_NOW"}
```

# 3 References

See one of the following references for more details on manual cooling options:

- White Paper: Custom Cooling Options
- iDRAC9 User Guide