

Introduction

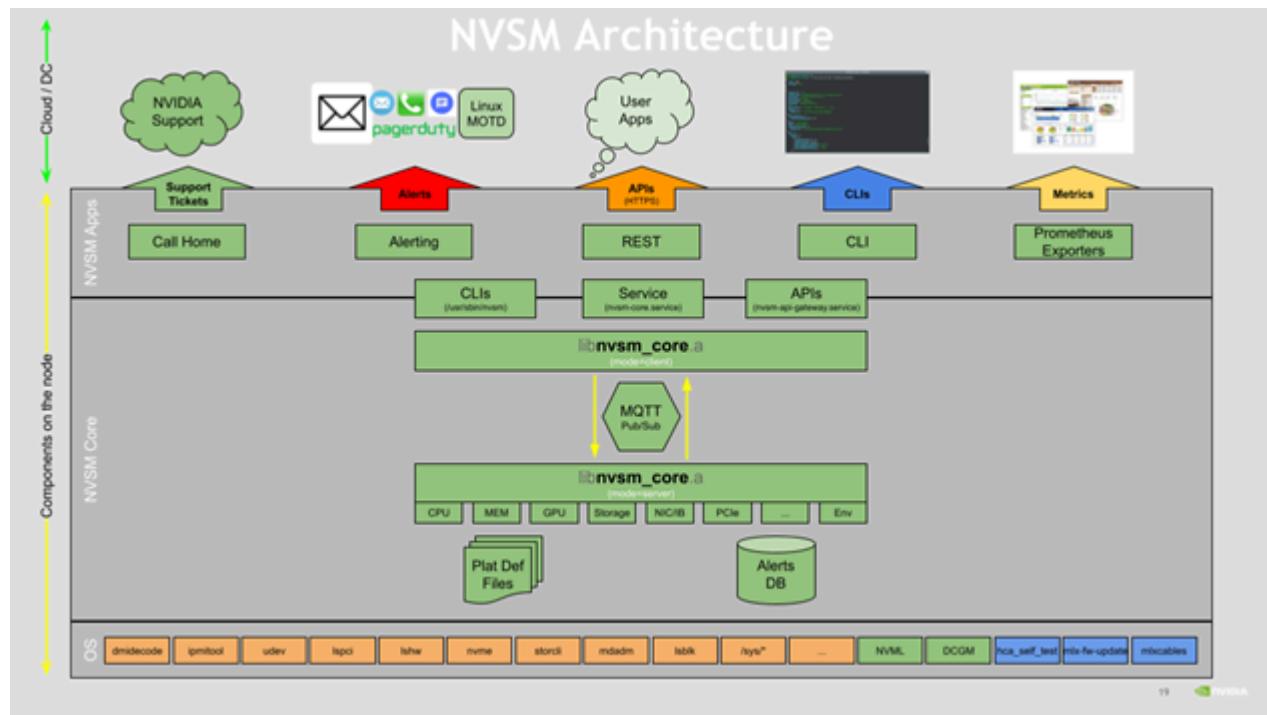
NVIDIA® System Management (NVSM) is an NVIDIA software stack for managing and monitoring NVIDIA-designed servers such as NVIDIA DGX™, CGX, and RTX servers.

- NVSM is an “always-on” health monitoring engine which catches issues proactively as opposed to other tools which need to be run post facto. By virtue of having deep knowledge of the underlying platform, NVSM has the optimal list of health checks to make as well as how frequently each check needs to happen.
- NVSM CLIs and APIs alleviate the need for users to
 - Have deep knowledge of tools such as ipmitool, dmidecode, lspci, storcli, mdadm, and lsblk.
 - Have deep knowledge of platform details such as the intended PCIe hierarchy, storage hierarchy, or error thresholds.
 - Manually correlate information from several tools; in many cases, the output of one tool needs to be manually parsed to know how to use the next tool. For example, BDF in SEL record vs BDF in lspci just to determine which device is faulty.
- NVSM catches issues which some customers might never notice. For example, some PCIe links might be running at lower link width/speed causing jobs to run slow. Without NVSM, customers might suspect something wrong with their jobs OR worse assume that DGX is simply that slow.
- NVSM provides
 - An on-demand health check suite which runs a battery of tests and reports deviations from expected results.
 - The ability to create a bundle of all relevant system logs required by NVIDIA support when reporting an issue.
 - A secure REST API interface removing the need for users/scripts to log-into the system. So it is easy to develop remote management SW applications using these APIs.
 - A Prometheus metrics exporter which can be enabled so an external Prometheus server can pull critical system metrics from the target DGX nodes.
- NVSM’s call-home feature, if enabled, creates a support ticket on behalf of the customer automatically in case of platform issues, even before the customer notices it.
- In addition, NVSM provides other notification mechanisms like email and PagerDuty.

Currently, NVSM supports the following DGX systems:

- › DGX servers: Complete NVSM functionality described in this document.
- › DGX Station: Functionality is limited to using the CLI to check the health of the system and obtain diagnostic information.

The following is a high level diagram of the NVSM architecture:



! Note

“Always on” functionality is not supported on DGX Station.

Configurable “Always On” Features

NVSM contains the following features that you can configure using the NVSM CLI:

- › Health Monitor Alerts
- › Health Monitor Policies

Health Monitor Alerts

Alerts are events of significance that require attention. When a health monitor detects such an event in the subsystem that it monitors, it generates an alert to inform the user. The default behavior is to log the alerts in persistent storage as well as to send an E-mail notification to registered users. Refer to the section [Using the NVSM CLI](#) for details about configuring users for receiving alert E-mail notifications.

Each alert has a ‘state’. An active alert can be in a ‘critical’ or ‘warning’ state. Here, ‘critical’ implies an event that needs immediate action, and ‘warning’ implies an event that needs user attention. When the alerting condition is removed, the alert state changes to ‘cleared’. Details of how to view the generated alerts recorded in the database are available in the section [Using the NVSM CLI](#).

Health Monitor Alert List

The following table describes each alert ID:

Message and details	Alert ID	Component ID	Severity
Unsupported drive configuration. Affected component URI: {{ index .Params "Uri" }}	NV-DRIVE-01	Drive Slot <>	Warning
System entered degraded mode, drive in {{ index .Params "DriveSlot" }} is not supported.	NV-DRIVE-07	Drive Slot <>	Warning
Unsupported SED drive configuration.	NV-DRIVE-09	Volume label	Warning
Unsupported volume encryption configuration.	NV-DRIVE-10	Volume label	Critical
M.2 drive firmware version mismatch.	NV-DRIVE-11	Drive Slot <>	Warning
System entered degraded mode, volume {{ index .Params "ComponentName" }} is under rebuild.	NV-VOL-01	Volume name	Warning
System entered degraded mode, volume {{ index .Params "ComponentName" }} rebuild failed.	NV-VOL-02	Volume name	Critical

Message and details	Alert ID	Component ID	Severity
System entered degraded mode, volume {{ index .Params "ComponentName" }} is in a degraded state.	NV-VOL-03	Volume name	Critical
System entered degraded mode, volume {{ index .Params "ComponentName" }} is inactive or in a failed state.	NV-VOL-04	Volume name	Critical
Raid-0 Volume {{ index .Params "ComponentName" }} is misconfigured.	NV-VOL-05	Volume name	Warning
Raid-0 data volume for caching is not present.	NV-VOL-06		Informational
EFI partition missing on boot volume. Run 'blkid' to check the partition type.	NV-VOL-09	Volume name	Critical
Storage Volume {{ index .Params "ComponentName" }} utilization is nearing 90% of {{ index .Params "Capacity" }} bytes.	NV-VOL-10	Volume name	Critical
System entered degraded mode, {} is reporting an error. (Power supply module has failed.)	NV-PSU-01	<PSU#> where # is the PSU number.	Critical

Message and details	Alert ID	Component ID	Severity
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Operating temperature exceeds the thermal specifications of the component.)</p>	NV-PSU-02	<PSU#> where # is the PSU number.	Warning
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Input to the PSU is missing)</p>	NV-PSU-03	<PSU#> where # is the PSU number.	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Input voltage is out of range for the Power Supply Module)</p> <p>(Input voltage is out of range for the Power Supply Module)</p>	NV-PSU-04	<PSU#> where # is the PSU number.	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(PSU is missing)</p>	NV-PSU-05	<PSU#> where # is the PSU number.	Warning
<p>Failures in Power supply modules have been detected.</p> <p>(System is operating in degraded performance mode.)</p>	NV-PSU-06		Warning

Message and details	Alert ID	Component ID	Severity
<p>Failures in Power supply modules have been detected.</p> <p>(System is in power failed state)</p>	NV-PSU-07		Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Operating temperature exceeds the thermal specifications of the component.)</p>	NV-PDB-01	<PDB#> where # is the PDB number	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Fan speed reading has fallen below the expected speed setting.)</p>	NV-FAN-01	<p><FAN#_F> or <FAN#_R></p> <p>where # is the fan module number.</p> <p>F is for front fan.</p> <p>R is for rear fan.</p>	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Fan readings are inaccessible.)</p>	NV-FAN-02	<p><FAN#_F> or <FAN#_R></p> <p>where # is the fan module number.</p> <p>F is for front fan.</p> <p>R is for rear fan.</p>	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(An unrecoverable CPU Internal error has occurred.)</p>	NV-CPU-01	<p><CPU#></p> <p>where # is the CPU socket number (CPU0 or CPU1)</p>	Critical

Message and details	Alert ID	Component ID	Severity
<p>System entered degraded mode, {} is reporting an error.</p> <p>(CPU Thermtrip has occurred, processor socket temperature exceeded the thermal specifications of the component.)</p>	NV-CPU-02	<p><CPU#></p> <p>where # is the CPU socket number (CPU0 or CPU1)</p>	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Processor socket temperature exceeded the thermal specifications of the component.)</p>	NV-CPU-03		Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Processor socket temperature exceeded the thermal specifications of the component.)</p>	NV-CPU-04		Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Uncorrectable error is reported).</p>	NV-DIMM-01	<p><CPU#_DIMM_@\$></p> <p>where # = (1, 2)</p> <p>@ = (A, B, C, D, E, F)</p> <p>\$ = (1, 2)</p>	Critical

Message and details	Alert ID	Component ID	Severity
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Correctable errors reported exceeds the configured threshold.)</p>	NV-DIMM-02	<CPU#_DIMM_@\$> where # = (1, 2) @ = (A, B, C, D, E, F) \$ = (1, 2)	Warning
<p>System entered degraded mode, {} is reporting an error.</p> <p>(Unrecoverable error is observed on the DIMM, specific details of the error are unavailable.)</p>	NV-DIMM-03	<CPU#_DIMM_@\$> where # = (1, 2) @ = (A, B, C, D, E, F) \$ = (1, 2)	Critical
<p>System entered degraded mode, {} is reporting an error.</p> <p>(DIMM presence is not expected in this slot, please verify the DIMM details.)</p>	NV-DIMM-04		
<p>System entered degraded mode, GPU is reporting an error</p> <p>(Critical error has been reported by the GPU.)</p>	NV-GPU-01		Critical

Message and details	Alert ID	Component ID	Severity
<p>GPU{} power Limits are not configured correctly</p> <p>(Expected limits (Power: 200000W, Clock: 1597MHz), Actual limits (Power: 200000W, Clock: 1163MHz).)</p>	NV-GPU-02		Critical
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Link speed degradation observed between {BDF1, BDF2}, expected link speed is {} actual link speed is {})</p>	NV-PCI-01		Warning
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Link width degradation observed between {BDF1, BDF2},, expected link width is {} actual link width is {})</p>	NV-PCI-02		Warning
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Correctable errors reported on {BDF}).</p>	NV-PCI-03		Warning

Message and details	Alert ID	Component ID	Severity
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(UnCorrectable errors reported on {BDF})</p>	NV-PCI-04		Critical
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Device is missing on {BDF})</p>	NV-PCI-05		Critical
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Device Error Reporting is disabled on {BDF} for {})</p>	NV-PCI-06		Critical
<p>System entered degraded mode, {ID} is reporting an error.</p> <p>(Device is disabled on {BDF})</p>	NV-PCI-07		Critical
<p>System entered degraded mode, controller {{ index .Params "ComponentName" }} is reporting an error.</p>	NV-CONTROLLER-01	Controller name	Warning
<p>System entered degraded mode, Storage controller {{ index .Params "ComponentName" }} is reporting a PHY error.</p>	NV-CONTROLLER-02	Controller name	Warning

Message and details	Alert ID	Component ID	Severity
System entered degraded mode, controller {{ index .Params "ComponentName" }} is set at lower than expected speed.	NV-CONTROLLER-03	Controller name	Warning
System entered degraded mode, controller {{ index .Params "ComponentName" }} is reporting an error.	NV-CONTROLLER-04	Controller name	Warning
System entered degraded mode, controller {{ index .Params "ComponentName" }} is reporting an error.	NV-CONTROLLER-05	Controller name	Critical
System entered degraded mode, controller {{ index .Params "ComponentName" }} is reporting an error.	NV-CONTROLLER-06	Controller name	Critical
LEDStatus for controller {{ index .Params "ComponentName" }} needs to be cleared.	NV-CONTROLLER-07	Controller name	Critical
Link error on {}. (Network Link is down)	NV-NET-01		Warning

Message and details	Alert ID	Component ID	Severity
<p>Network traffic errors observed on {}.</p> <p>(Rx collision rate of {}, has crossed threshold value of {} on {}network port.)</p>	NV-NET-02		Warning
<p>Network traffic errors observed on {}.</p> <p>(Tx collision rate of {}, has crossed threshold value of {} on {}network port.)</p>	NV-NET-03		Warning
<p>Network traffic errors observed on {}.</p> <p>(CRC error rate of {}, has crossed threshold value of {} on {}network port.)</p>	NV-NET-04		Critical
<p>{} is reporting an error.</p> <p>({}Network port is disabled.)</p>	NV-NET-05		Critical
<p>Ethernet interface error on port {}.</p> <p>({}Ethernet health check failing with Online NVRAM test failure.)</p>	NV-ETH-01		Critical

Message and details	Alert ID	Component ID	Severity
<p>Ethernet interface configuration error on {}.</p> <p>(MAC address is missing on the Ethernet interface of {}.)</p>	NV-ETH-02		Critical
<p>IB driver error.</p> <p>(HCA self test reports IB driver initialization failure.)</p>	NV-IB-01		Critical
<p>Counter errors on IB port {}</p> <p>({}HCA self test on IB port reports counter error.)</p>	NV-IB-02		Critical
<p>Configuration error on IB port {}.</p> <p>(GUID is missing on {}HCA.)</p>	NV-IB-03		Critical
<p>System entered degraded mode, {} is reporting a fatal error</p> <p>(Critical error has been reported by the NVSwitch Id {} with SXID error {})</p>	NV-NVSWITCH-01		Critical

Message and details	Alert ID	Component ID	Severity
<p>System entered degraded mode, {} is reporting a non fatal error</p> <p>(Critical error has been reported by the NVSwitch Id {} with SXID error {})</p>	NV-NVSWITCH-02		Warning

Recommended Actions

(A)

1. Run 'sudo nvsm dump health'
2. Open a case with NVIDIA Enterprise Support at this address
<https://nvid.nvidia.com/dashboard/>
3. Attach this notification and the nvsyinfo log file from /tmp/nvsm-health-<hostname>-<timestamp>.tar.xz

(B)

1. Check the physical link connection
2. Open a case with NVIDIA Enterprise Support at <https://nvid.nvidia.com/dashboard/>

(C)

1. Check OFED installation troubleshooting
2. Open a case with NVIDIA Enterprise Support at this address
<https://nvid.nvidia.com/dashboard/>

(D)

1. Check the status of the Subnet Manager
2. Open a case with NVIDIA Enterprise Support at this address
<https://nvid.nvidia.com/dashboard/>

Health Monitor Policies

Users can tune certain aspects of health monitor behavior using health monitor policies. This includes details such as email related configuration for alert notification, selectively disabling devices to be monitored, etc. Details of the supported policies and how to configure them using the CLI are provided in the section [Using the NVSM CLI](#).

Verifying the Installation

Before using NVSM, you can verify the installation to make sure all the services are present.

Verifying NVSM Services with systemctl

NVSM is part of the DGX OS image and is launched by systemd when DGX boots. The following are the services running under NVSM:

- › nvsm-api-gateway.service
- › nvsm-core.service
- › nvsm-mqtt.service
- › nvsm-notifier.service
- › nvsm.service

You can verify if each NVSM service is up and running using the `systemctl` command. For example, the following command verifies the core service:

```
$ sudo systemctl status nvsm-core
```

You can view all the NVSM services and their status with the following command:

```
$ sudo systemctl status -all nvsm*
```

Verifying NVSM Services with nvsm status

The `nvsm status` command displays the NVSM services and their status, example output:

```
$ sudo nvsm status
  SERVICE          ENABLED   ACTIVE   SUB      DESCRIPTION
=====
nvsm-mqtt.service    enabled    active  running      MQTT broker for NVSM API
for signaling within NVSM API components
nvsm-core.service     enabled    active  running      NVSM Core Service for
System Management
nvsm-api-gateway.service  enabled    active  running      NVSM API Server to
provide DGX System Management APIs
nvsm-notifier.service  enabled    active  running      NVSM Notifier service.

Overall
=====
Overall Health: Healthy
Overall Status: Active
```

If you run `nvsm status` while NVSM is starting, the output resembles the following example:

```
$ sudo nvsm status
SERVICE           ENABLED   ACTIVE    SUB          DESCRIPTION
=====
nvsm-mqtt.service     enabled   active    running    MQTT broker for NVSM API for
signalizing within NVSM API components
nvsm-core.service      enabled   activating start-post  NVSM Core Service for System
Management
nvsm-api-gateway.service  enabled   inactive   dead       NVSM API Server to provide
DGX System Management APIs
nvsm-notifier.service   enabled   inactive   dead       NVSM Notifier service.

Overall
=====
Overall Health: Transient
Overall Status: Starting

Recommendations:
=====
0. NVSM is starting, this state should be transient, please try again later
1. nvsm-core.service is activating. If it stay in this state, please run "journalctl -f
fu nvsm-core.service" for more details
```

Note

The `nvsm` CLI command works only if all NVSM services are up and running.

If any sub service fails or stuck in starting, run the following command to get additional information:

```
sudo systemctl status <service-name>
```

For example:

```
sudo systemctl status nvsm-core.service
```