Collect data and monitor workload performance

Active IQ Unified Manager

NetApp
October 07, 2026

This PDF was generated from https://docs.netapp.com/us-en/active-iq-unified-manager/performance-checker/concept_collect_data_and_monitor_workload_performance.html on October 07, 2026. Always check docs.netapp.com for the latest.

Table of Contents

Collect data and monitor workload performance . . . 1
- Collect data and monitor workload performance in Active IQ Unified Manager . . . 1
- Monitor workload types with Active IQ Unified Manager . . . 1
- Understand workload performance measurement values in Active IQ Unified Manager . . . 2
- Understand how Active IQ Unified Manager calculates the expected performance range . . . 4
  - How the latency forecast is formed . . . 4
- Use the latency forecast for performance analysis in Active IQ Unified Manager . . . 5
- Use workload latency to identify performance issues in Active IQ Unified Manager . . . 6
- Assess how cluster operations can affect workload latency in Active IQ Unified Manager . . . 7
- Monitor MetroCluster performance with Active IQ Unified Manager . . . 7

Collect data and monitor workload performance

Collect data and monitor workload performance in Active IQ Unified Manager

Unified Manager collects and analyzes workload activity every 5 minutes to identify performance events, and it detects configuration changes every 15 minutes. It retains a maximum of 30 days of 5-minute historical performance and event data, and it uses this data to forecast the expected latency range for all monitored workloads.

Unified Manager must collect a minimum of 3 days of workload activity before it can begin its analysis and before the latency forecast for I/O response time can be displayed on the Workload Analysis page and in the Event details page. While this activity is being collected, the latency forecast does not display all changes occurring from workload activity. After collecting 3 days of activity, Unified Manager adjusts the latency forecast every 24 hours at 12:00 a.m., to reflect workload activity changes and establish a more accurate dynamic performance threshold.

During the first 4 days that Unified Manager is monitoring a workload, if more than 24 hours have passed since the last data collection, the latency charts will not display the latency forecast for that workload. Events detected prior to the last collection are still available.

Daylight savings time (DST) changes the system time, which alters the latency forecast of performance statistics for monitored workloads. Unified Manager immediately begins to correct the latency forecast, which takes approximately 15 days to complete. During this time you can continue to use Unified Manager, but, since Unified Manager uses the latency forecast to detect dynamic events, some events might not be accurate. Events detected prior to the time change are not affected.

Monitor workload types with Active IQ Unified Manager

You can use Unified Manager to monitor the performance of two types of workloads: user-defined and system-defined.

- ***User-defined workloads***

  The I/O throughput from applications to the cluster. These are processes involved in read and write requests. A volume, LUN, NFS share, SMB/CIFS share, and a workload is a user-defined workload.

  Unified Manager only monitors the workload activity on the cluster. It does not monitor the applications, the clients, or the paths between the applications and the cluster.

  If one or more of the following is true for a workload, it cannot be monitored by Unified Manager:

  - It is a data protection (DP) copy in read-only mode. (DP volumes are monitored for user-generated traffic.)
  - It is an offline data clone.
  - It is a mirrored volume in a MetroCluster configuration.
- ***System-defined workloads***

The internal processes involved with storage efficiency, data replication, and system health, including:

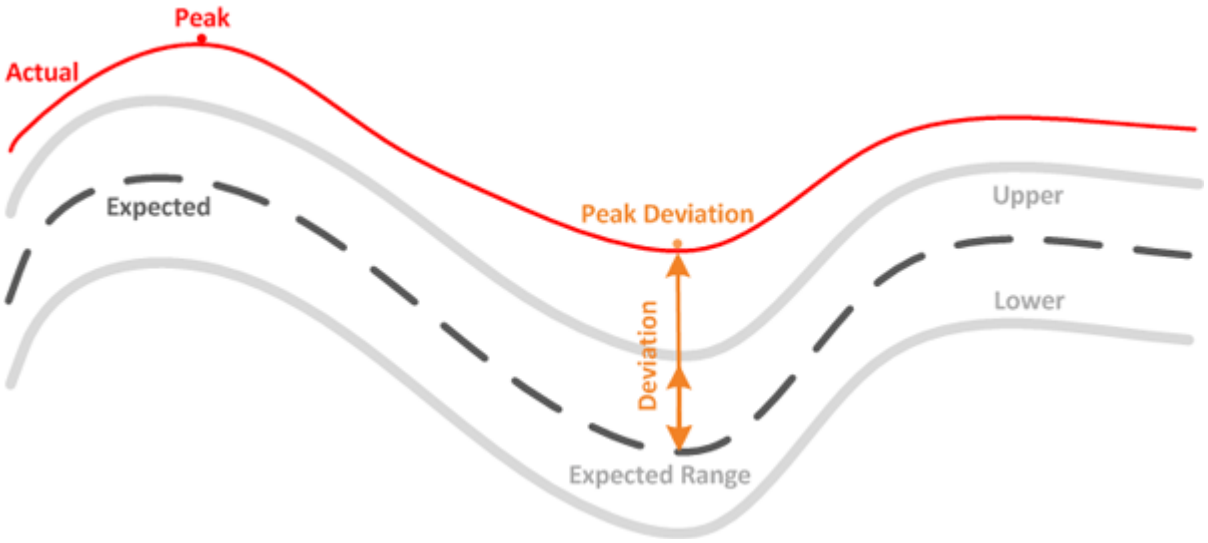
- Storage efficiency, such as deduplication
- Disk health, which includes RAID reconstruct, disk scrubbing, and so on
- Data replication, such as SnapMirror copies
- Management activities
- File system health, which includes various WAFL activities
- File system scanners, such as WAFL scan
- Copy offload, such as offloaded storage efficiency operations from VMware hosts
- System health, such as volume moves, data compression, and so on
- Unmonitored volumes

Performance data for system-defined workloads is displayed in the GUI only when the cluster component used by these workloads is in contention. For example, you cannot search for the name of a system-defined workload to view its performance data in the GUI.

Understand workload performance measurement values in Active IQ Unified Manager

Unified Manager measures the performance of workloads on a cluster based on historical and expected statistical values, which form the latency forecast of values for the workloads. It compares the actual workload statistical values to the latency forecast to determine when workload performance is too high or too low. A workload that is not performing as expected triggers a dynamic performance event to notify you.

In the following illustration, the actual value, in red, represents the actual performance statistics in the time frame. The actual value has crossed the performance threshold, which is the upper bounds of the latency forecast. The peak is the highest actual value in the time frame. The deviation measures the change between the expected values (the forecast) and the actual values, while the peak deviation indicates the largest change between the expected values and the actual values.



The following table lists the workload performance measurement values.

Measurement	Description
Activity	The percentage of the QoS limit used by the workloads in the policy group. If Unified Manager detects a change to a policy group, such as adding or removing a volume or changing the QoS limit, the actual and expected values might exceed 100% of the set limit. If a value exceeds 100% of the set limit it is displayed as >100%. If a value is less than 1% of the set limit it is displayed as <1%.
Actual	The measured performance value at a specific time for a given workload.
Deviation	The change between the expected values and the actual values. It is the ratio of the actual value minus the expected value to the upper value of the expected range minus the expected value. A negative deviation value indicates that workload performance is lower than expected, while a positive deviation value indicates that workload performance is higher than expected.
Expected	The expected values are based on the analysis of historical performance data for a given workload. Unified Manager analyzes these statistical values to determine the expected range (latency forecast) of values.
Latency Forecast (Expected Range)	The latency forecast is a prediction of what the upper and lower performance values are expected to be at a specific time. For the workload latency, the upper values form the performance threshold. When the actual value crosses the performance threshold, Unified Manager triggers a dynamic performance event.
Peak	The maximum value measured over a period of time.
Peak Deviation	The maximum deviation value measured over a period of time.
Queue Depth	The number of pending I/O requests that are waiting at the interconnect component.

Measurement	Description
Utilization	For the network processing, data processing, and aggregate components, the percentage of busy time to complete workload operations over a period of time. For example, the percentage of time for the network processing or data processing components to process an I/O request or for an aggregate to fulfill a read or write request.
Write Throughput	The amount of write throughput, in Megabytes per second (MB/s), from workloads on a local cluster to the partner cluster in a MetroCluster configuration.

Understand how Active IQ Unified Manager calculates the expected performance range

The latency forecast is a prediction of what the upper and lower performance values are expected to be at a specific time. For the workload latency, the upper values form the performance threshold. When the actual value crosses the performance threshold, Unified Manager triggers a dynamic performance event.

For example, during regular business hours between 9:00 a.m. and 5:00 p.m., most employees might check their email between 9:00 a.m. and 10:30 a.m. The increased demand on the email servers means an increase in workload activity on the back-end storage during this time. Employees might notice slow response time from their email clients.

During the lunch hour between 12:00 p.m. and 1:00 p.m. and at the end of the work day after 5:00 p.m., most employees are likely away from their computers. The demand on the email servers typically decreases, also decreasing the demand on back-end storage. Alternatively, there could be scheduled workload operations, such as storage backups or virus scanning, that start after 5:00 p.m. and increase activity on the back-end storage.

Over several days, the increase and decrease in workload activity determines the expected range (latency forecast) of activity, with upper and lower boundaries for a workload. When the actual workload activity for an object is outside the upper or lower boundaries, and remains outside the boundaries for a period of time, this might indicate that the object is being overused or underused.

How the latency forecast is formed

Unified Manager must collect a minimum of 3 days of workload activity before it can begin its analysis and before the latency forecast for I/O response time can be displayed in the GUI. The minimum required data collection does not account for all changes occurring from workload activity. After collecting the first 3 days of activity, Unified Manager adjusts the latency forecast every 24 hours at 12:00 a.m. to reflect workload activity changes and establish a more accurate dynamic performance threshold.

Daylight savings time (DST) changes the system time, which alters the latency forecast of performance statistics for monitored workloads. Unified Manager immediately begins to correct the latency forecast, which takes approximately 15 days to complete. During this time you can continue to use Unified Manager, but, since Unified Manager uses the latency forecast to detect dynamic events, some events might not be accurate. Events detected prior to the time change are not affected.

Use the latency forecast for performance analysis in Active IQ Unified Manager

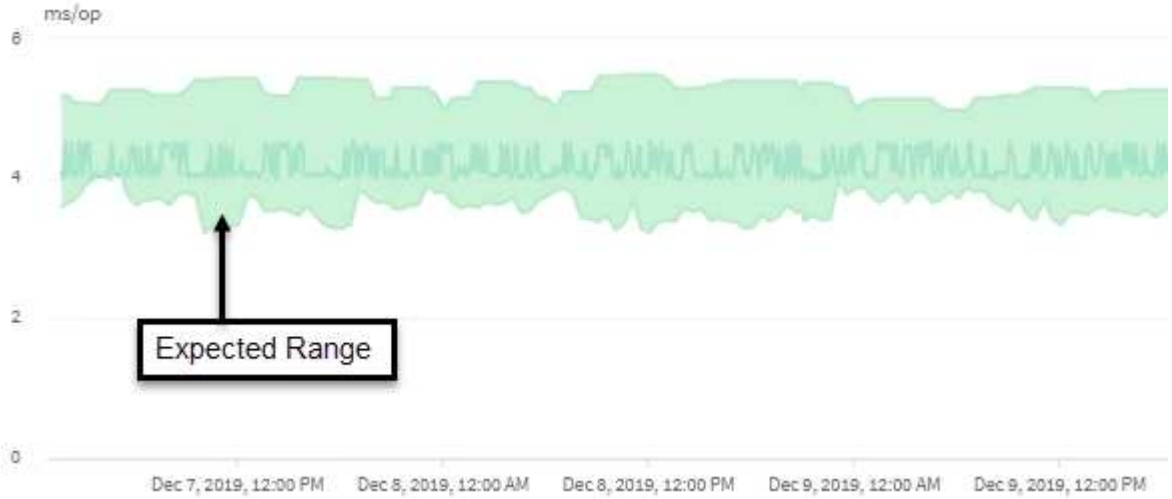
Unified Manager uses the latency forecast to represent the typical I/O latency (response time) activity for your monitored workloads. It alerts you when the actual latency for a workload is above the upper bounds of the latency forecast, which triggers a dynamic performance event, so that you can analyze the performance issue and take corrective action for resolving it.

The latency forecast sets the performance baseline for the workload. Over time, Unified Manager learns from past performance measurements to forecast the expected performance and activity levels for the workload. The upper boundary of the expected range establishes the dynamic performance threshold. Unified Manager uses the baseline to determine when the actual latency is above or below a threshold, or outside the bounds of their expected range. The comparison between the actual values and the expected values creates a performance profile for the workload.

When the actual latency for a workload exceeds the dynamic performance threshold, due to contention on a cluster component, the latency is high and the workload performs more slowly than expected. The performance of other workloads that share the same cluster components might also be slower than expected.

Unified Manager analyzes the threshold crossing event and determines whether the activity is a performance event. If the high workload activity remains consistent for a long period of time, such as several hours, Unified Manager considers the activity to be normal and dynamically adjusts the latency forecast to form the new dynamic performance threshold.

Some workloads might have consistently low activity, where the latency forecast for latency does not have a high rate of change over time. To minimize the number of events during analysis of performance events, Unified Manager triggers an event only for low-activity volumes whose operations and latencies are much higher than expected.



In this example, the latency for a volume has a latency forecast, in gray, of 3.5 milliseconds per operation (ms/op) at its lowest and 5.5 ms/op at its highest. If the actual latency, in blue, suddenly increases to 10 ms/op, due to an intermittent spike in network traffic or contention on a cluster component, it is then above the latency forecast and has exceeded the dynamic performance threshold.

When network traffic has decreased, or the cluster component is no longer in contention, the latency returns within the latency forecast. If the latency remains at or above 10 ms/op for a long period of time, you might

need to take corrective action to resolve the event.

Use workload latency to identify performance issues in Active IQ Unified Manager

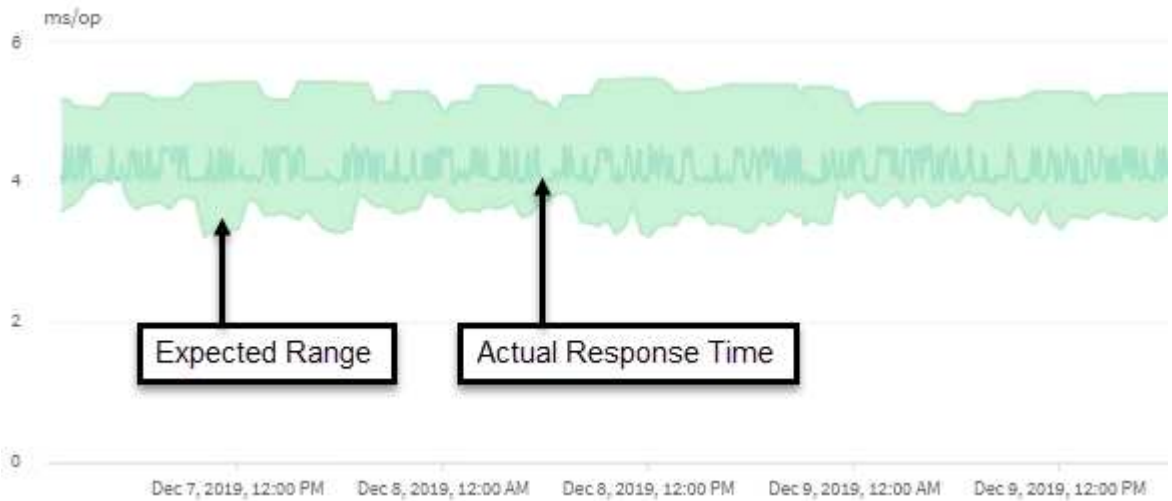
The workload latency (response time) is the time it takes for a volume on a cluster to respond to I/O requests from client applications. Unified Manager uses the latency to detect and alert you to performance events.

A high latency means that requests from applications to a volume on a cluster are taking longer than usual. The cause of the high latency could be on the cluster itself, due to contention on one or more cluster components. High latency could also be caused by issues outside of the cluster, such as network bottlenecks, issues with the client hosting the applications, or issues with the applications themselves.

Unified Manager only monitors the workload activity on the cluster. It does not monitor the applications, the clients, or the paths between the applications and the cluster.

Operations on the cluster, such as making backups or running deduplication, that increase their demand of cluster components shared by other workloads can also contribute to high latency. If the actual latency exceeds the dynamic performance threshold of the expected range (latency forecast), Unified Manager analyzes the event to determine whether it is a performance event that you might need to resolve. The latency is measured in milliseconds per operation (ms/op).

On the Latency Total chart in the Workload Analysis page, you can view an analysis of the latency statistics to see how the activity of individual processes, such as read and write requests, compares to the overall latency statistics. The comparison helps you determine which operations have the highest activity or whether specific operations have abnormal activity that is impacting the latency for a volume. When analyzing performance events, you can use the latency statistics to determine whether an event was caused by an issue on the cluster. You can also identify the specific workload activities or cluster components that are involved in the event.



This example shows the Latency chart . The actual response time (latency) activity is a blue line and the latency forecast (expected range) is green.

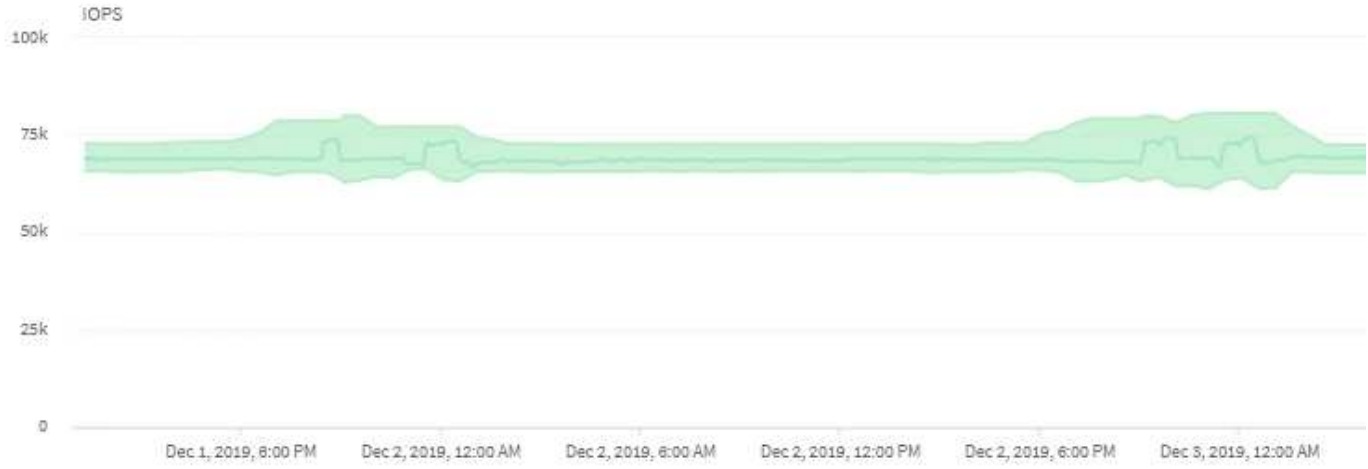
There can be gaps in the blue line if Unified Manager was unable to gather data. This can occur because the cluster or volume was unreachable, Unified Manager was turned off during that

time, or the collection was taking longer than the 5 minute collection period.

Assess how cluster operations can affect workload latency in Active IQ Unified Manager

Operations (IOPS) represent the activity of all user-defined and system-defined workloads on a cluster. The IOPS statistics help you determine whether cluster processes, such as making backups or running deduplication, are impacting workload latency (response time) or might have caused, or contributed to, a performance event.

When analyzing performance events, you can use the IOPS statistics to determine whether a performance event was caused by an issue on the cluster. You can identify the specific workload activities that might have been the main contributors to the performance event. IOPS are measured in operations per second (ops/sec).



This example shows the IOPS chart. The actual operations statistics is a blue line and the IOPS forecast of operations statistics is green.

In some cases where a cluster is overloaded, Unified Manager might display the message `Data collection is taking too long on Cluster` *`cluster_name`*. This means that not enough statistics have been collected for Unified Manager to analyze. You need to reduce the resources the cluster is using so that statistics can be collected.

Monitor MetroCluster performance with Active IQ Unified Manager

Unified Manager enables you to monitor the write throughput between clusters in a MetroCluster configuration to identify workloads with a high amount of write throughput.

If these high-performing workloads are causing other volumes on the local cluster to have high I/O response times, Unified Manager triggers performance events to notify you.

Unified Manager treats the clusters in a MetroCluster configuration as individual clusters. It does not distinguish between clusters that are partners or correlate the write throughput from each cluster.

When a local cluster in a MetroCluster configuration mirrors its data to its partner cluster, the data is written to NVRAM and then transferred over the interswitch links (ISLs) to the remote aggregates. Unified Manager analyzes the NVRAM to identify the workloads whose high write throughput is overutilizing the NVRAM, placing the NVRAM in contention.

Workloads whose deviation in response time has exceeded the performance threshold are called *victims* and workloads whose deviation in write throughput to the NVRAM is higher than usual, causing the contention, are called *bullies*. Because only the write requests are mirrored to the partner cluster, Unified Manager does not analyze read throughput.

You can view the throughput of any of the clusters in a MetroCluster configuration by analyzing the workloads of the corresponding LUNs and volumes from the following screens. You can filter the results by the cluster. From the left navigation pane:

- **Storage > Clusters > Performance: All Clusters** view. See
- **Storage > Volumes > Performance: All Volumes** view.
- **Storage > LUNs > Performance: All LUNs** view.
- **Workload Analysis > All workloads**

Related information

Performance event analysis and notification

Performance event analysis for a MetroCluster configuration

Roles of workloads involved in a performance event

Identifying victim workloads involved in a performance event

Identifying bully workloads involved in a performance event

Identifying shark workloads involved in a performance event

Copyright information

Copyright © 2026 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.