Monitor volume latency

EDA workloads

NetApp
February 02, 2026

Table of Contents

Monitor volume latency

Monitor volume latency

Using latency analysis you can proactively monitor volume performance by tracking read and write latency metrics across your FSx for ONTAP file systems. Configure customizable thresholds for warning and critical events to identify potential performance bottlenecks before they impact your EDA workloads.

Overview

Latency analysis collects and monitors CloudWatch metrics for volume read and write operations. When both latency and IOPS thresholds are breached for all data points within a specified time range, the system generates alerts that appear in the latency events table. This enables you to:

- Identify volumes experiencing performance degradation.
- Distinguish between warning-level and critical-level performance issues.
- Track latency trends over time to optimize storage configurations.
- Take proactive action before latency impacts workload performance.

Before you begin

To use latency analysis, you must have AWS credentials configured in Workload Factory. The feature requires access to CloudWatch metrics for all FSx for ONTAP volumes associated with your AWS credentials.

If you haven't configured AWS credentials, see Add AWS credentials.

Configure latency thresholds

You can configure thresholds for both warning and critical events. Each event type includes separate thresholds for read and write operations. The system evaluates these thresholds continuously and generates alerts when conditions are met.

> ⓘ You must set critical event thresholds higher than warning event thresholds to ensure proper alert escalation. If not, you cannot save your configuration.

About this task

For an alert to trigger, both the latency threshold and the IOPS threshold must be breached for all data points within the specified time range. This dual-condition logic helps reduce false positives by ensuring that high latency is sustained under significant load.

Steps

1. Log in using one of the console experiences.
2. Select the menu ☰ and then select **EDA**.
3. From the EDA menu, select **Latency**.
4. In the EDA latency configuration page, configure the following thresholds:

    - **Warning events**
        - **Read latency threshold**: Enter the latency threshold in milliseconds. Default: 6 ms.
        - **Read IOPS threshold**: Enter the IOPS threshold in operations per second. Default: 100 ops/sec.
        - **Read time range**: Enter the time range in minutes (5-20). Default: 10 minutes.
        - **Write latency threshold**: Enter the latency threshold in milliseconds. Default: 8 ms.
        - **Write IOPS threshold**: Enter the IOPS threshold in operations per second. Default: 100 ops/sec.
        - **Write time range**: Enter the time range in minutes (5-20). Default: 10 minutes.
    - **Critical events**
        - **Read latency threshold**: Enter the latency threshold in milliseconds. Default: 12 ms.
        - **Read IOPS threshold**: Enter the IOPS threshold in operations per second. Default: 100 ops/sec.
        - **Read time range**: Enter the time range in minutes (5-20). Default: 10 minutes.
        - **Write latency threshold**: Enter the latency threshold in milliseconds. Default: 15 ms.
        - **Write IOPS threshold**: Enter the IOPS threshold in operations per second. Default: 100 ops/sec.
        - **Write time range**: Enter the time range in minutes (5-20). Default: 10 minutes.

5. Select **Apply**.

Result

Workload Factory begins collecting latency metrics for all FSx for ONTAP volumes associated with your AWS credentials. Metrics are collected at least every 20 minutes. The latency events table displays any volumes that breach your configured thresholds.

Understanding alerts

The latency analysis feature uses CloudWatch alarms to monitor volume performance. Understanding how alerts are triggered helps you configure appropriate thresholds and interpret the results.

Metrics collected

The system collects the following CloudWatch metrics for each volume:

- **Read latency threshold**: Calculated as 1000 * m2/(m1+0.000001) where m1 = DataReadOperations and m2 = DataReadOperationTime
- **Write latency threshold**: Calculated as 1000 * m2/(m1+0.000001) where m1 = DataWriteOperations and m2 = DataWriteOperationTime

Alert trigger conditions

An alert is triggered when all of the following conditions are met:

- The latency threshold is exceeded for the operation type (read or write).
- The IOPS threshold is exceeded for the operation type.
- Both conditions persist for all data points within the configured time range.

For example, with default warning thresholds, a read alert triggers only if read latency exceeds 6 ms AND read IOPS exceeds 100 ops/sec for all data points within a 10-minute period.

Event severity

- **Warning events**: Indicate elevated latency that might need attention.
- **Critical events**: Indicate severe latency that requires immediate investigation.

View latency events

The latency events table displays all warning and critical events detected within the last 72 hours. Use this table to monitor volume performance and identify volumes that require optimization.

Additional information

- Only the latest breach for each volume appears in the table. If a volume experiences multiple breaches, only the most recent event is displayed.
- Events are automatically removed after 72 hours.
- The table displays a maximum of 200 events. Older events are removed as new events are added.

Steps

1. In the **Latency** tab, view the latency events table.
2. Review the information for each event including:
    - **Severity**: Indicates whether the event is Critical or Warning.
    - **Volume name**: The name of the affected volume.
    - **Volume ID**: The ID of the affected volume.
    - **File system**: The FSx for ONTAP file system containing the volume.
    - **Time detected**: When the breach was detected
    - **Median latency**: The median latency value during the breach period.
3. To sort the table, select any column header. By default, critical events appear first sorted by time, followed by warning events sorted by time.
4. To dismiss one or more events, next to each event select **Dismiss**.
5. To add columns to the table, select the column icon, choose the columns, and select **Apply**.

Manage latency configuration

After the initial configuration, you can edit your thresholds.

Steps

1. In the **Latency** page, select **Edit**.
2. Modify any of the threshold values as needed.

    > ⓘ Ensure that critical thresholds remain higher than warning thresholds. The system displays an error if you configure critical thresholds lower than warning thresholds.

3. Select **Apply** to save your changes.

Best practices

Consider these recommendations when configuring and using latency analysis:

- **Set realistic thresholds**: Configure thresholds based on your workload requirements. Default values provide a starting point but might need adjustment for your specific environment.
- **Start with warning thresholds**: Use warning events to establish baseline performance expectations before fine-tuning critical thresholds.
- **Consider time ranges carefully**: Shorter time ranges (5-10 minutes) detect issues faster but might generate more alerts. Longer time ranges (15-20 minutes) reduce false positives but might delay detection.
- **Monitor trends**: Regularly review the latency events table to identify patterns or recurring issues that might indicate underlying configuration problems.
- **Coordinate IOPS and latency thresholds**: The dual-condition logic means both must be exceeded. Setting very high IOPS thresholds might prevent alerts even when latency is problematic.
- **Review dismissed events**: Periodically review why events were dismissed to identify opportunities for threshold adjustment or infrastructure improvements.

Copyright information

Trademark information