TMS Metrics

> ⚠ **Attention**
>
> NVIDIA Triton Management Service (TMS) will reach the end of life on July 31, 2024. The version 1.4.0 is the last release.

TMS provides a metrics endpoint from which Prometheus formatted runtime metrics can be retrieved.

The following Helm chart options can be used to affect how metrics are reported:

- `server.metrics.enabled` can be used to enable/disable metrics endpoint.
- `server.metrics.reportingWindow` can be used to configure the reporting window for metrics endpoint.
- `server.metrics.minimumVisibility` can be used to configure the which metrics are collected and reported.

By default, only high visibility are reported when metrics reporting is enabled.

Standard (high visibility) metrics are reported for each of the server's endpoints.

Metric Name	Description
tms_error_count	Number of errored requests during the reporting window.
tms_duration_avg_seconds	Average duration of successful requests during the reporting window.
tms_grpc_request_count	Number of gRPC requests made during the reporting window.

Additional metrics are available by adjusting the minimum-visibility Helm chart value. These metrics are self-describing as part of the Prometheus formatted output.