



This PDF is generated from authoritative online content, and is provided for convenience only. This PDF cannot be used for legal purposes. For authoritative understanding of what is and is not supported, always use the online content. To copy code samples, always use the online content.

# Voice Microservices Private Edition Guide

ORS metrics and alerts

12/18/2025

---

## Contents

- [1 Metrics](#)
- [2 Alerts](#)

---

Find the metrics ORS exposes and the alerts defined for ORS.

Service	CRD or annotations?	Port	Endpoint/Selector	Metrics update interval
ORS	Supports both CRD and annotations	11200	http://:11200/metrics	30 seconds

See details about:

- ORS metrics
- ORS alerts

## Metrics

You can query Prometheus directly to see all the metrics that the Voice Orchestration Service exposes. The following metrics are likely to be particularly useful. Genesys does not commit to maintain other currently available Orchestration Service metrics not documented on this page.

Metric and description	Metric details	Indicator of
<b>orsnode_callevents</b> Total number of received call events.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_ha_writes</b> The number of HA writes to Redis.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_ha_reads</b> The number of HA reads from Redis.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_interactions</b> The number of active interactions.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_total_interactions</b> The total number of interactions that	<b>Unit:</b> N/A <b>Type:</b> counter	Traffic

Metric and description	Metric details	Indicator of
have been created.	<b>Label:</b> <b>Sample value:</b>	
<b>orsnode_cleared_interactions</b> The total number of call interactions that have been cleared.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_strategies</b> The number of strategies that are running.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_total_strategies</b> The total number of strategies that have been created.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_load_errors</b> The total number of strategy load errors.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_fetch_errors</b> The total number of errors encountered when a strategy tried to fetch data from a Designer Application Server (DAS).	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_config_errors</b> The total number of strategy configuration errors.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_invoke_errors</b> The total number of strategy invoke errors.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_treatments</b> The total number of strategy treatments.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_failed_treatments</b> The total number of failed strategy treatments.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_userdata_updates</b> The total number of times that a strategy	<b>Unit:</b> N/A <b>Type:</b> counter	Traffic

Metric and description	Metric details	Indicator of
updated user data.	<b>Label:</b> <b>Sample value:</b>	
<b>orsnode_scxml_transitions</b> The total number of SCXML transitions.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_scxml_events</b> The total number of SCXML events.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Traffic
<b>orsnode_scxml_error_events</b> The total number of SCXML error.* events.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_http_fetch_requests</b> The total number of HTTP fetch requests.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_http_fetch_duration</b> The HTTP fetch time, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	Latency
<b>orsnode_http_fetch_errors</b> The total number of HTTP fetch errors.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_http_fetch_error_status</b> Status of the HTTP fetch error.	<b>Unit:</b> <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_urs_rlib_latency_msec</b> The Universal Routing Server (URS) rlib latency, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	Latency
<b>orsnode_urs_rlib_errors</b> The total number of URS rlib errors.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_urs_rlib_requests</b> The total number of URS rlib requests.	<b>Unit:</b> N/A <b>Type:</b> counter	

Metric and description	Metric details	Indicator of
	<b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_rlib_events</b> The total number of URS rlib events.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_rlib_timeouts</b> The total number of URS rlib timeouts.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_redis_state</b> Current Redis connection state.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> redis_cluster_name <b>Sample value:</b>	
<b>orsnode_redis_disconnect</b> The number of times that the ORS node disconnected from Redis.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_sdr_messages_sent</b> The number of SDR messages that have been sent.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_rq_latency_msec</b> Redis queue latency, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> le, service <b>Sample value:</b>	Latency
<b>orsnode_routing_latency_msec</b> Routing latency, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	Latency
<b>orsnode_rstream_latency_msec</b> Redis stream latency, measured in (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> le, node <b>Sample value:</b>	Latency
<b>orsnode_digital_latency_msec</b> Digital stream latency, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	Latency
<b>orsnode_sip_health_check</b> ORS health check.	<b>Unit:</b> N/A <b>Type:</b> gauge	

Metric and description	Metric details	Indicator of
	<b>Label:</b> node <b>Sample value:</b>	
<b>orsnode_ixn_health_check</b> Interaction health check.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_rq_state</b> Current Redis queue connection state.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_ixn_events</b> Total number of interaction stream events received.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_rq_disconnect</b> Number of times the ORS node disconnected from the RQ Service.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>service_version_info</b> Displays the version of Voice Orchestration Service that is currently running. In the case of this metric, the labels provide the important information. The metric value is always 1 and does not provide any information.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> version <b>Sample value:</b> service_version_info{version="100.0.1000006"} 1	
<b>orsnode_route_redirected</b> Total number of EventRouteUsed events without a ReferenceID.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_balancer_stream_state</b> The state of the voice balancer stream.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> balancer_stream_type <b>Sample value:</b>	
<b>orsnode_high_memory</b> Indicates when the ORS node is using a lot of memory.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_rlib_state</b> Indicates a Tenant rlib request timeout.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_stuck_interactions</b>	<b>Unit:</b> N/A	

Metric and description	Metric details	Indicator of
The number of stuck interactions.	<b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_scxml_submit_requests</b> The total number of URS SCXMLSubmit requests.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_scxml_cancel_requests</b> The total number of URS SCXMLQueueCancel requests.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_urs_queue_submit_done_events</b> Total number of URS queue.submit.done events.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_health_level</b> Summarized health level of the ORS node:  -1 – fail 0 – starting 1 – degraded 2 – pass	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_health_check_error</b> Health check errors for the ORS node:  1 – has error 0 – no error	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> reason <b>Sample value:</b>	Errors
<b>orsnode_running_applications</b> The number of active sessions for each Designer application.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_failed_applications</b> The number of failed sessions for each Designer application.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_total_applications</b> The total number of sessions created for each Designer application.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_failed_scripts</b> The number of scripts that failed to load in the Tenant Service configuration	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b>	



Metric and description	Metric details	Indicator of
management environment.	<b>Sample value:</b>	
<b>orsnode_session_load_time_msec</b> The time it takes for the strategy to be compiled and go through its initial states.	<b>Unit:</b> milliseconds <b>Type:</b> histogram <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_service_started</b> Timestamp when the ORS node started.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> started <b>Sample value:</b>	
<b>orsnode_total_terminal_requests</b> Total number of terminal requests (like Deliver, PlaceInQueue, StopProcessing for Digital and RequestClearCall, RequestRouteCall for Voice).	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_total_non_terminal_requests</b> Total number of non-terminal requests to the Interaction Server (for Digital) or SIP Server (for Voice).	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_sip_post_errors</b> Total number of errors encountered in POST requests to the SIP node.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	Errors
<b>orsnode_pending_tlib_requests</b> Total number of pending TLib requests.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_sips_rest_connections</b> The number of active REST connections with SIP Cluster Service.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_number_compiled_applications</b> The number of compiled applications in the cache.	<b>Unit:</b> N/A <b>Type:</b> counter <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_cached_applications_size</b> The sum of the sizes of the cached applications.	<b>Unit:</b> <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_tlib_latency_msec</b> The TLib Rest API request latency,	<b>Unit:</b> milliseconds <b>Type:</b> histogram	Latency

Metric and description	Metric details	Indicator of
measured in (ms).	<b>Label:</b> le <b>Sample value:</b>	
<b>orsnode_application_size</b> The compiled size of the Designer application.	<b>Unit:</b> <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_application_microstep_count</b> The number of microsteps while executing the Designer application.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_application_run_time_msec</b> The length of time the Designer application was running, measured in milliseconds (ms).	<b>Unit:</b> milliseconds <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_application_compiled_date</b> The date on which the Designer application was compiled.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	
<b>orsnode_application_last_invoked_date</b> The date when the Designer application was last invoked.	<b>Unit:</b> N/A <b>Type:</b> gauge <b>Label:</b> <b>Sample value:</b>	

## Alerts

The following alerts are defined for ORS.

Alert	Severity	Description	Based on	Threshold
Number of running strategies is too high	Warning	<p>Too many active sessions.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> </ul>	orsnode_strategies	More than 400 strategies running in 5 consecutive seconds.

Alert	Severity	Description	Based on	Threshold
		<ul style="list-style-type: none"> <li>Check the number of voice, digital, and callback calls in the system.</li> </ul>		
Number of running strategies is critical	Critical	<p>Too many active sessions.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> <li>Check the number of voice, digital, and callback calls in the system.</li> </ul>	orsnode_strategies	More than 600 strategies running in 5 consecutive seconds.
Redis disconnected for 5 minutes	Warning	<p>Actions:</p> <ul style="list-style-type: none"> <li>If the alarm is triggered for multiple services, make sure there are no issues with Redis, and then restart Redis.</li> <li>If alarm is triggered only for pod {{ \$labels.pod }}, check if there is an issue with the pod.</li> </ul>	redis_state	Redis is not available for the pod {{ \$labels.pod }} for 5 minutes.
Redis disconnected for 10 minutes	Critical	<p>Actions:</p> <ul style="list-style-type: none"> <li>If the alarm is triggered for multiple services, make</li> </ul>	redis_state	Redis is not available for the pod {{ \$labels.pod }} for 10 minutes.

Alert	Severity	Description	Based on	Threshold
		<p>sure there are no issues with Redis, and then restart Redis.</p> <ul style="list-style-type: none"> <li>If the alarm is triggered only for pod {{ \$labels.pod }}, check if there is an issue with the pod.</li> </ul>		
Pod status Failed	Warning	<p>Pod {{ \$labels.pod }} failed.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>One of the containers in the pod has entered a Failed state. Check the Kibana logs for the reason.</li> </ul>	kube_pod_status_phase	Pod {{ \$labels.pod }} is in Failed state.
Pod in Unknown state	Warning	<p>Pod {{ \$labels.pod }} is in Unknown state.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>If the alarm is triggered for multiple services, make sure there are no issues with the Kubernetes cluster.</li> <li>If the alarm is triggered only for pod {{ \$labels.pod }}, check whether the image is correct and if the container is starting up.</li> </ul>	kube_pod_status_phase	Pod {{ \$labels.pod }} is in Unknown state for 5 minutes.
Pod in Pending state	Warning	Pod {{ \$labels.pod }} is in Pending state.	kube_pod_status_phase	Pod {{ \$labels.pod }} is in Pending state for 5 minutes.

Alert	Severity	Description	Based on	Threshold
		<p>Actions:</p> <ul style="list-style-type: none"> <li>If the alarm is triggered for multiple services, make sure the Kubernetes nodes where the pod is running are alive in the cluster.</li> <li>If the alarm is triggered only for the pod {{ \$labels.pod }}, check the health of the pod.</li> </ul>		minutes.
Pod Not ready for 10 minutes	Critical	<p>Pod {{ \$labels.pod }} in NotReady state.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>If this alarm is triggered, check whether the CPU is available for the pods.</li> <li>Check whether the port of the pod is running and serving the request.</li> </ul>	kube_pod_status_ready	Pod {{ \$labels.pod }} in NotReady state for 10 minutes.
Container restored repeatedly	Critical	<p>Actions:</p> <ul style="list-style-type: none"> <li>One of the containers in the pod has entered a failed state. Check the Kibana logs for the reason.</li> </ul>	kube_pod_container_status_restarts_total	Container {{ \$labels.container }} was restarted 5 or more times within 15 minutes.
Pod memory greater than 65%	Warning	High memory usage for pod {{ \$labels.pod }}.	container_memory_working_set_bytes kube_pod_container_resource_requests_memory_bytes	Container {{ \$labels.container }} memory usage greater than 65% of kube_pod_container_resource_requests_memory_bytes

Alert	Severity	Description	Based on	Threshold
		<p>Actions:</p> <ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> <li>Check Grafana for abnormal load.</li> <li>Collect the service logs; raise an investigation ticket.</li> </ul>		exceeded 65% for 5 minutes.
Pod memory greater than 80%	Critical	<p>Critical memory usage for pod {{ \$labels.pod }}.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> <li>Check Grafana for abnormal load.</li> <li>Restart the service.</li> <li>Collect the service logs; raise an investigation ticket.</li> </ul>	<p>container_memory_working_set_bytes</p> <p>kube_pod_container_resource_requests_memory_bytes</p>	<p>Container {{ \$labels.container }} memory usage exceeded 80% for 5 minutes.</p>
Pod CPU greater than 65%	Warning	<p>High CPU load for pod {{ \$labels.pod }}.</p> <p>Actions:</p>	<p>container_cpu_usage_seconds_total</p> <p>container_spec_cpu_period</p>	<p>Container {{ \$labels.container }} CPU usage exceeded 65% for 5 minutes.</p>

Alert	Severity	Description	Based on	Threshold
		<ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> <li>Check Grafana for abnormal load.</li> <li>Collect the service logs; raise an investigation ticket.</li> </ul>		
Pod CPU greater than 80%	Critical	<p>Critical CPU load for pod {{ \$labels.pod }}.</p> <p>Actions:</p> <ul style="list-style-type: none"> <li>Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached.</li> <li>Check Grafana for abnormal load.</li> <li>Restart the service.</li> <li>Collect the service logs; raise an investigation ticket.</li> </ul>	<p>container_cpu_usage_seconds_total,</p> <p>container_spec_cpu_period</p>	<p>Container {{ \$labels.container }}</p> <p>if CPU usage exceeded 80% for 5 minutes.</p>