Voice Microservices Private Edition Guide

ORS metrics and alerts

8/12/2022

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

Contents

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
orsnode_callevents Total number of received call events.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	Traffic
orsnode_ha_writes The number of HA writes to Redis.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	Traffic
orsnode_ha_reads The number of HA reads from Redis.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	Traffic
orsnode_interactions The number of active interactions.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	Traffic
orsnode_total_interactions The total number of interactions that have been created.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	Traffic
orsnode_cleared_interactions The total number of call interactions that have been cleared.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	

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

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

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

Metric and description	Metric details	Indicator of
Current Redis queue connection state.	**Type:** gauge **Label:** **Sample value:**	
orsnode_ixn_events Total number of interaction stream events received.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_rq_disconnect Number of times the ORS node disconnected from the RQ Service.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
service_version_info 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.	**Unit:** N/A **Type:** gauge **Label:** version **Sample value:** service_version_info{version="100.0.1000006"} 1	
orsnode_route_redirected Total number of EventRouteUsed events without a ReferenceID.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_balancer_stream_state The state of the voice balancer stream.	**Unit:** N/A **Type:** gauge **Label:** balancer_stream_type **Sample value:**	
orsnode_high_memory Indicates when the ORS node is using a lot of memory.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_urs_rlib_state Indicates a Tenant `rlib` request timeout.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_stuck_interactions The number of stuck interactions.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_urs_scxml_submit_requests The total number of URS `SCXMLSubmit` requests.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	

Metric and description	Metric details	Indicator of
orsnode_urs_scxml_cancel_requests The total number of URS `SCXMLQueueCancel` requests.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_urs_queue_submit_done_events Total number of URS `queue.submit.done` events.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_health_level Summarized health level of the ORS node: -1 – fail 0 – starting 1 – degraded 2 – pass	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_health_check_error Health check errors for the ORS node: 1 – has error 0 – no error	**Unit:** N/A **Type:** gauge **Label:** reason **Sample value:**	Errors
orsnode_running_applications The number of active sessions for each Designer application.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_failed_applications The number of failed sessions for each Designer application.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_total_applications The total number of sessions created for each Designer application.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_failed_scripts The number of scripts that failed to load in the Tenant Service configuration management environment.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_session_load_time_msec The time it takes for the strategy to be compiled and go through its initial states.	**Unit:** milliseconds **Type:** histogram **Label:** **Sample value:**	
orsnode_service_started Timestamp when the ORS node started.	**Unit:** N/A **Type:** gauge	

Metric and description	Metric details	Indicator of
	Label: started **Sample value:**	
orsnode_total_terminal_requests Total number of terminal requests (like Deliver, PlaceInQueue, StopProcessing for Digital and RequestClearCall, RequestRouteCall for Voice).	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_total_non_terminal_requests Total number of non-terminal requests to the Interaction Server (for Digital) or SIP Server (for Voice).	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_sip_post_errors Total number of errors encountered in POST requests to the SIP node.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	Errors
orsnode_pending_tlib_requests Total number of pending TLib requests.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_sips_rest_connections The number of active REST connections with SIP Cluster Service.	**Unit:** N/A **Type:** gauge **Label:** **Sample value:**	
orsnode_number_compiled_applications The number of compiled applications in the cache.	**Unit:** N/A **Type:** counter **Label:** **Sample value:**	
orsnode_cached_applications_size The sum of the sizes of the cached applications.	**Unit:** **Type:** gauge **Label:** **Sample value:**	
orsnode_tlib_latency_msec The TLib Rest API request latency, measured in (ms).	**Unit:** milliseconds **Type:** histogram **Label:** le **Sample value:**	Latency
orsnode_application_size The compiled size of the Designer application.	**Unit:** **Type:** gauge **Label:** **Sample value:**	
orsnode_application_microstep_count	**Unit:** N/A	

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

Alerts

The following alerts are defined for ORS.

Alert	Severity	Description	Based on	Threshold
Number of running strategies is too high	Warning	Too many active sessions. Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached. • Check the number of voice, digital, and callback calls in the system.	orsnode_strategies	More than 400 strategies running in 5 consecutive seconds.
Number of running strategies is	Critical	Too many active sessions.	orsnode_strategies	More than 600 strategies running

Alert	Severity	Description	Based on	Threshold
critical		Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached. • Check the number of voice, digital, and callback calls in the system.		in 5 consecutive seconds.
Redis disconnected for 5 minutes	Warning	Actions: • If the alarm is triggered for multiple services, make sure there are no issues with Redis, and then restart Redis. • If alarm is triggered only for pod {{ $labels.pod }}, check if there is an issue with the pod.	redis_state	Redis is not available for the pod {{ $labels.pod }} for 5 minutes.
Redis disconnected for 10 minutes	Critical	Actions: • If the alarm is triggered for multiple services, make sure there are no issues with Redis, and then restart Redis. • If the alarm is triggered only for pod {{ $labels.pod }}, check if there is an issue with	redis_state	Redis is not available for the pod {{ $labels.pod }} for 10 minutes.

Alert	Severity	Description	Based on	Threshold
		the pod.		
Pod status Failed	Warning	Pod {{ $labels.pod }} failed. Actions: • One of the containers in the pod has entered a Failed state. Check the Kibana logs for the reason.	kube_pod_status_phase	Pod {{ $labels.pod }} is in Failed state.
Pod in Unknown state	Warning	Pod {{ $labels.pod }} is in Unknown state. Actions: • If the alarm is triggered for multiple services, make sure there are no issues with the Kubernetes cluster. • If the alarm is triggered only for pod {{ $labels.pod }}, check whether the image is correct and if the container is starting up.	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. Actions: • If the alarm is triggered for multiple services, make sure the Kubernetes nodes where the pod is	kube_pod_status_phase	Pod {{ $labels.pod }} is in Pending state for 5 minutes.

Alert	Severity	Description	Based on	Threshold
		running are alive in the cluster. • If the alarm is triggered only for the pod {{ $labels.pod }}, check the health of the pod.		
Pod Not ready for 10 minutes	Critical	Pod {{ $labels.pod }} in NotReady state. Actions: • If this alarm is triggered, check whether the CPU is available for the pods. • Check whether the port of the pod is running and serving the request.	kube_pod_status_ready	Pod {{ $labels.pod }} in NotReady state for 10 minutes.
Container restored repeatedly	Critical	Actions: • One of the containers in the pod has entered a failed state. Check the Kibana logs for the reason.	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 }}. Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of	container_memory_working_set_bytes, kube_pod_container_resource_requests_memory_bytes	Container {{ $labels.container }} memory usage exceeded 65% for 5 minutes.

Alert	Severity	Description	Based on	Threshold
		pods has been reached. • Check Grafana for abnormal load. • Collect the service logs; raise an investigation ticket.		
Pod memory greater than 80%	Critical	Critical memory usage for pod {{ $labels.pod }}. Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached. • Check Grafana for abnormal load. • Restart the service. • Collect the service logs; raise an investigation ticket.	container_memory_working_set_bytes, kube_pod_container_resource_requests_memory_bytes	Container {{ $labels.container }} memory usage exceeded 80% for 5 minutes.
Pod CPU greater than 65%	Warning	High CPU load for pod {{ $labels.pod }}. Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been	container_cpu_usage_seconds_total, container_spec_cpu_period	Container {{ $labels.container }} CPU usage exceeded 65% for 5 minutes.

Alert	Severity	Description	Based on	Threshold
		reached. • Check Grafana for abnormal load. • Collect the service logs; raise an investigation ticket.		
Pod CPU greater than 80%	Critical	Critical CPU load for pod {{ $labels.pod }}. Actions: • Check whether the horizontal pod autoscaler has triggered and if the maximum number of pods has been reached. • Check Grafana for abnormal load. • Restart the service. • Collect the service logs; raise an investigation ticket.	container_cpu_usage_seconds_total, container_spec_cpu_period	Container {{ $labels.container }} CPU usage exceeded 80% for 5 minutes.