```
{
  "annotations": {
    "list": [
      {
        "builtIn": 1,
        "datasource": {
          "type": "grafana",
          "uid": "-- Grafana --"
        },
        "enable": true,
        "hide": true,
        "iconColor": "rgba(0, 211, 255, 1)",
        "name": "Annotations & Alerts",
        "target": {
          "limit": 100,
          "matchAny": false,
          "tags": [],
          "type": "dashboard"
        },
        "type": "dashboard"
      }
    ]
  },
  "description": "Swarms Cloud Monitoring",
  "editable": true,
  "fiscalYearStartMonth": 0,
```

```
"graphTooltip": 0,
"id": 29,
"links": [],
"liveNow": false,
"panels": [
 {
   "datasource": {
     "type": "prometheus",
     "uid": "prometheus"
   },
   "description": "End to end request latency measured in seconds.",
   "fieldConfig": {
     "defaults": {
       "color": {
         "mode": "palette-classic"
       },
       "custom": {
         "axisCenteredZero": false,
         "axisColorMode": "text",
         "axisLabel": "",
         "axisPlacement": "auto",
         "barAlignment": 0,
         "drawStyle": "line",
         "fillOpacity": 0,
         "gradientMode": "none",
         "hideFrom": {
```

```
      "legend": false,
      "tooltip": false,
      "viz": false
    },
    "lineInterpolation": "linear",
    "lineWidth": 1,
    "pointSize": 5,
    "scaleDistribution": {
      "type": "linear"
    },
    "showPoints": "auto",
    "spanNulls": false,
    "stacking": {
      "group": "A",
      "mode": "none"
    },
    "thresholdsStyle": {
      "mode": "off"
    }
  },
  "mappings": [],
  "thresholds": {
    "mode": "absolute",
    "steps": [
      {
        "color": "green",
```

```
          "value": null
        },
        {
          "color": "red",
          "value": 80
        }
      ]
    },
    "unit": "s"
  },
  "overrides": []
},
"gridPos": {
  "h": 8,
  "w": 12,
  "x": 0,
  "y": 0
},
"id": 9,
"options": {
  "legend": {
    "calcs": [],
    "displayMode": "list",
    "placement": "bottom",
    "showLegend": true
  },
```

```
        "tooltip": {
          "mode": "single",
          "sort": "none"
        }
      },
      "targets": [
        {
          "datasource": {
            "type": "prometheus",
            "uid": "prometheus"
          },
          "disableTextWrap": false,
          "editorMode": "builder",
          "expr": "histogram_quantile(0.99, sum by(le) (rate(vllm:e2e_request_latency_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
          "fullMetaSearch": false,
          "includeNullMetadata": false,
          "instant": false,
          "legendFormat": "P99",
          "range": true,
          "refId": "A",
          "useBackend": false
        },
        {
          "datasource": {
```

```
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "histogram_quantile(0.95, sum by(le) (rate(vllm:e2e_request_latency_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "P95",
      "range": true,
      "refId": "B",
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "histogram_quantile(0.9, sum by(le) (rate(vllm:e2e_request_latency_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])
```

```
))",
        "fullMetaSearch": false,
        "hide": false,
        "includeNullMetadata": false,
        "instant": false,
        "legendFormat": "P90",
        "range": true,
        "refId": "C",
        "useBackend": false
      },
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "disableTextWrap": false,
        "editorMode": "builder",
                                                    "expr": "histogram_quantile(0.5, sum by(le)
(rate(vllm:e2e_request_latency_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])
))",
        "fullMetaSearch": false,
        "hide": false,
        "includeNullMetadata": false,
        "instant": false,
        "legendFormat": "P50",
        "range": true,
```

```
        "refId": "D",
        "useBackend": false
      },
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "editorMode": "code",
        "expr": "rate(vllm:e2e_request_latency_seconds_sum{model_name=\"$model_name\"}[$__rate_interval])\n/\nrate(vllm:e2e_request_latency_seconds_count{model_name=\"$model_name\"}[$__rate_interval])",
        "hide": false,
        "instant": false,
        "legendFormat": "Average",
        "range": true,
        "refId": "E"
      }
    ],
    "title": "E2E Request Latency",
    "type": "timeseries"
  },
  {
    "datasource": {
      "type": "prometheus",
```

```
  "uid": "prometheus"
},
"description": "Number of tokens processed per second",
"fieldConfig": {
  "defaults": {
    "color": {
      "mode": "palette-classic"
    },
    "custom": {
      "axisCenteredZero": false,
      "axisColorMode": "text",
      "axisLabel": "",
      "axisPlacement": "auto",
      "barAlignment": 0,
      "drawStyle": "line",
      "fillOpacity": 0,
      "gradientMode": "none",
      "hideFrom": {
        "legend": false,
        "tooltip": false,
        "viz": false
      },
      "lineInterpolation": "linear",
      "lineWidth": 1,
      "pointSize": 5,
      "scaleDistribution": {
```

```
      "type": "linear"
    },
    "showPoints": "auto",
    "spanNulls": false,
    "stacking": {
      "group": "A",
      "mode": "none"
    },
    "thresholdsStyle": {
      "mode": "off"
    }
  },
  "mappings": [],
  "thresholds": {
    "mode": "absolute",
    "steps": [
      {
        "color": "green",
        "value": null
      },
      {
        "color": "red",
        "value": 80
      }
    ]
  }
```

```
  },
  "overrides": []
},
"gridPos": {
  "h": 8,
  "w": 12,
  "x": 12,
  "y": 0
},
"id": 8,
"options": {
  "legend": {
    "calcs": [],
    "displayMode": "list",
    "placement": "bottom",
    "showLegend": true
  },
  "tooltip": {
    "mode": "single",
    "sort": "none"
  }
},
"targets": [
  {
    "datasource": {
      "type": "prometheus",
```

```
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "rate(vllm:prompt_tokens_total{model_name=\"$model_name\"}[$__rate_interval])",
      "fullMetaSearch": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "Prompt Tokens/Sec",
      "range": true,
      "refId": "A",
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
                                                                                "expr":
"rate(vllm:generation_tokens_total{model_name=\"$model_name\"}[$__rate_interval])",
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
```

```
      "legendFormat": "Generation Tokens/Sec",
      "range": true,
      "refId": "B",
      "useBackend": false
    }
  ],
  "title": "Token Throughput",
  "type": "timeseries"
},
{
  "datasource": {
    "type": "prometheus",
    "uid": "prometheus"
  },
  "description": "Inter token latency in seconds.",
  "fieldConfig": {
    "defaults": {
      "color": {
        "mode": "palette-classic"
      },
      "custom": {
        "axisCenteredZero": false,
        "axisColorMode": "text",
        "axisLabel": "",
        "axisPlacement": "auto",
        "barAlignment": 0,
```

```
  "drawStyle": "line",
  "fillOpacity": 0,
  "gradientMode": "none",
  "hideFrom": {
    "legend": false,
    "tooltip": false,
    "viz": false
  },
  "lineInterpolation": "linear",
  "lineWidth": 1,
  "pointSize": 5,
  "scaleDistribution": {
    "type": "linear"
  },
  "showPoints": "auto",
  "spanNulls": false,
  "stacking": {
    "group": "A",
    "mode": "none"
  },
  "thresholdsStyle": {
    "mode": "off"
  }
},
"mappings": [],
"thresholds": {
```

```
      "mode": "absolute",
      "steps": [
        {
          "color": "green",
          "value": null
        },
        {
          "color": "red",
          "value": 80
        }
      ]
    },
    "unit": "s"
  },
  "overrides": []
},
"gridPos": {
  "h": 8,
  "w": 12,
  "x": 0,
  "y": 8
},
"id": 10,
"options": {
  "legend": {
    "calcs": [],
```

```
        "displayMode": "list",
        "placement": "bottom",
        "showLegend": true
      },
      "tooltip": {
        "mode": "single",
        "sort": "none"
      }
    },
    "targets": [
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "disableTextWrap": false,
        "editorMode": "builder",
        "expr": "histogram_quantile(0.99, sum by(le) (rate(vllm:time_per_output_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
        "fullMetaSearch": false,
        "includeNullMetadata": false,
        "instant": false,
        "legendFormat": "P99",
        "range": true,
        "refId": "A",
```

```
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "histogram_quantile(0.95, sum by(le) (rate(vllm:time_per_output_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "P95",
      "range": true,
      "refId": "B",
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
```

```
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "histogram_quantile(0.9, sum by(le) (rate(vllm:time_per_output_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "P90",
      "range": true,
      "refId": "C",
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
      "expr": "histogram_quantile(0.5, sum by(le) (rate(vllm:time_per_output_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))",
      "fullMetaSearch": false,
      "hide": false,
```

```
        "includeNullMetadata": false,
        "instant": false,
        "legendFormat": "P50",
        "range": true,
        "refId": "D",
        "useBackend": false
      },
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "editorMode": "code",
        "expr": "rate(vllm:time_per_output_token_seconds_sum{model_name=\"$model_name\"}[$__rate_interval])\n/\nrate(vllm:time_per_output_token_seconds_count{model_name=\"$model_name\"}[$__rate_interval])",
        "hide": false,
        "instant": false,
        "legendFormat": "Mean",
        "range": true,
        "refId": "E"
      }
    ],
    "title": "Time Per Output Token Latency",
    "type": "timeseries"
```

```
},
{
  "datasource": {
    "type": "prometheus",
    "uid": "prometheus"
  },
  "description": "Number of requests in RUNNING, WAITING, and SWAPPED state",
  "fieldConfig": {
    "defaults": {
      "color": {
        "mode": "palette-classic"
      },
      "custom": {
        "axisCenteredZero": false,
        "axisColorMode": "text",
        "axisLabel": "",
        "axisPlacement": "auto",
        "barAlignment": 0,
        "drawStyle": "line",
        "fillOpacity": 0,
        "gradientMode": "none",
        "hideFrom": {
          "legend": false,
          "tooltip": false,
          "viz": false
        },
```

```
  "lineInterpolation": "linear",
  "lineWidth": 1,
  "pointSize": 5,
  "scaleDistribution": {
    "type": "linear"
  },
  "showPoints": "auto",
  "spanNulls": false,
  "stacking": {
    "group": "A",
    "mode": "none"
  },
  "thresholdsStyle": {
    "mode": "off"
  }
},
"mappings": [],
"thresholds": {
  "mode": "absolute",
  "steps": [
    {
      "color": "green",
      "value": null
    },
    {
      "color": "red",
```

```
          "value": 80
        }
       ]
     },
     "unit": "none"
   },
   "overrides": []
 },
 "gridPos": {
   "h": 8,
   "w": 12,
   "x": 12,
   "y": 8
 },
 "id": 3,
 "options": {
   "legend": {
     "calcs": [],
     "displayMode": "list",
     "placement": "bottom",
     "showLegend": true
   },
   "tooltip": {
     "mode": "single",
     "sort": "none"
   }
```

```
},
"targets": [
  {
    "datasource": {
      "type": "prometheus",
      "uid": "prometheus"
    },
    "disableTextWrap": false,
    "editorMode": "builder",
    "expr": "vllm:num_requests_running{model_name=\"$model_name\"}",
    "fullMetaSearch": false,
    "includeNullMetadata": true,
    "instant": false,
    "legendFormat": "Num Running",
    "range": true,
    "refId": "A",
    "useBackend": false
  },
  {
    "datasource": {
      "type": "prometheus",
      "uid": "prometheus"
    },
    "disableTextWrap": false,
    "editorMode": "builder",
    "expr": "vllm:num_requests_swapped{model_name=\"$model_name\"}",
```

```
  "fullMetaSearch": false,
  "hide": false,
  "includeNullMetadata": true,
  "instant": false,
  "legendFormat": "Num Swapped",
  "range": true,
  "refId": "B",
  "useBackend": false
},
{
  "datasource": {
    "type": "prometheus",
    "uid": "prometheus"
  },
  "disableTextWrap": false,
  "editorMode": "builder",
  "expr": "vllm:num_requests_waiting{model_name=\"$model_name\"}",
  "fullMetaSearch": false,
  "hide": false,
  "includeNullMetadata": true,
  "instant": false,
  "legendFormat": "Num Waiting",
  "range": true,
  "refId": "C",
  "useBackend": false
}
```

```
  ],
  "title": "Scheduler State",
  "type": "timeseries"
},
{
  "datasource": {
    "type": "prometheus",
    "uid": "prometheus"
  },
  "description": "P50, P90, P95, and P99 TTFT latency in seconds.",
  "fieldConfig": {
    "defaults": {
      "color": {
        "mode": "palette-classic"
      },
      "custom": {
        "axisCenteredZero": false,
        "axisColorMode": "text",
        "axisLabel": "",
        "axisPlacement": "auto",
        "barAlignment": 0,
        "drawStyle": "line",
        "fillOpacity": 0,
        "gradientMode": "none",
        "hideFrom": {
          "legend": false,
```

```
      "tooltip": false,
      "viz": false
    },
    "lineInterpolation": "linear",
    "lineWidth": 1,
    "pointSize": 5,
    "scaleDistribution": {
      "type": "linear"
    },
    "showPoints": "auto",
    "spanNulls": false,
    "stacking": {
      "group": "A",
      "mode": "none"
    },
    "thresholdsStyle": {
      "mode": "off"
    }
  },
  "mappings": [],
  "thresholds": {
    "mode": "absolute",
    "steps": [
      {
        "color": "green",
        "value": null
```

```
        },
        {
          "color": "red",
          "value": 80
        }
      ]
    },
    "unit": "s"
  },
  "overrides": []
},
"gridPos": {
  "h": 8,
  "w": 12,
  "x": 0,
  "y": 16
},
"id": 5,
"options": {
  "legend": {
    "calcs": [],
    "displayMode": "list",
    "placement": "bottom",
    "showLegend": true
  },
  "tooltip": {
```

```
        "mode": "single",
        "sort": "none"
      }
    },
    "targets": [
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "disableTextWrap": false,
        "editorMode": "builder",
                                                "expr": "histogram_quantile(0.99, sum by(le)
(rate(vllm:time_to_first_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))"
,
        "fullMetaSearch": false,
        "hide": false,
        "includeNullMetadata": false,
        "instant": false,
        "legendFormat": "P99",
        "range": true,
        "refId": "A",
        "useBackend": false
      },
      {
        "datasource": {
```

```
            "type": "prometheus",
            "uid": "prometheus"
          },
          "disableTextWrap": false,
          "editorMode": "builder",
                                                "expr": "histogram_quantile(0.95, sum by(le)
(rate(vllm:time_to_first_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))"
,
          "fullMetaSearch": false,
          "includeNullMetadata": false,
          "instant": false,
          "legendFormat": "P95",
          "range": true,
          "refId": "B",
          "useBackend": false
        },
        {
          "datasource": {
            "type": "prometheus",
            "uid": "prometheus"
          },
          "disableTextWrap": false,
          "editorMode": "builder",
                                                "expr": "histogram_quantile(0.9, sum by(le)
(rate(vllm:time_to_first_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))"
,
```

```
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "P90",
      "range": true,
      "refId": "C",
      "useBackend": false
    },
    {
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "disableTextWrap": false,
      "editorMode": "builder",
                                                "expr": "histogram_quantile(0.5, sum by(le)
(rate(vllm:time_to_first_token_seconds_bucket{model_name=\"$model_name\"}[$__rate_interval])))"
,
      "fullMetaSearch": false,
      "hide": false,
      "includeNullMetadata": false,
      "instant": false,
      "legendFormat": "P50",
      "range": true,
      "refId": "D",
```

```
        "useBackend": false
      },
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "editorMode": "code",
                                                                                "expr":
"rate(vllm:time_to_first_token_seconds_sum{model_name=\"$model_name\"}[$__rate_interval])\n/\n
rate(vllm:time_to_first_token_seconds_count{model_name=\"$model_name\"}[$__rate_interval])",
        "hide": false,
        "instant": false,
        "legendFormat": "Average",
        "range": true,
        "refId": "E"
      }
    ],
    "title": "Time To First Token Latency",
    "type": "timeseries"
  },
  {
    "datasource": {
      "type": "prometheus",
      "uid": "prometheus"
    },
```

```
"description": "Percentage of used cache blocks by vLLM.",
"fieldConfig": {
  "defaults": {
    "color": {
      "mode": "palette-classic"
    },
    "custom": {
      "axisCenteredZero": false,
      "axisColorMode": "text",
      "axisLabel": "",
      "axisPlacement": "auto",
      "barAlignment": 0,
      "drawStyle": "line",
      "fillOpacity": 0,
      "gradientMode": "none",
      "hideFrom": {
        "legend": false,
        "tooltip": false,
        "viz": false
      },
      "lineInterpolation": "linear",
      "lineWidth": 1,
      "pointSize": 5,
      "scaleDistribution": {
        "type": "linear"
      },
```

```
      "showPoints": "auto",
      "spanNulls": false,
      "stacking": {
        "group": "A",
        "mode": "none"
      },
      "thresholdsStyle": {
        "mode": "off"
      }
    },
    "mappings": [],
    "thresholds": {
      "mode": "absolute",
      "steps": [
        {
          "color": "green",
          "value": null
        },
        {
          "color": "red",
          "value": 80
        }
      ]
    },
    "unit": "percentunit"
  },
```

```
  "overrides": []
},
"gridPos": {
  "h": 8,
  "w": 12,
  "x": 12,
  "y": 16
},
"id": 4,
"options": {
  "legend": {
    "calcs": [],
    "displayMode": "list",
    "placement": "bottom",
    "showLegend": true
  },
  "tooltip": {
    "mode": "single",
    "sort": "none"
  }
},
"targets": [
  {
    "datasource": {
      "type": "prometheus",
      "uid": "prometheus"
```

```
        },
        "editorMode": "code",
        "expr": "vllm:gpu_cache_usage_perc{model_name=\"$model_name\"}",
        "instant": false,
        "legendFormat": "GPU Cache Usage",
        "range": true,
        "refId": "A"
      },
      {
        "datasource": {
          "type": "prometheus",
          "uid": "prometheus"
        },
        "editorMode": "code",
        "expr": "vllm:cpu_cache_usage_perc{model_name=\"$model_name\"}",
        "hide": false,
        "instant": false,
        "legendFormat": "CPU Cache Usage",
        "range": true,
        "refId": "B"
      }
    ],
    "title": "Cache Utilization",
    "type": "timeseries"
  }
],
```

```
"refresh": "",
"schemaVersion": 37,
"style": "dark",
"tags": [],
"templating": {
  "list": [
    {
      "current": {
        "selected": false,
        "text": "vllm",
        "value": "vllm"
      },
      "datasource": {
        "type": "prometheus",
        "uid": "prometheus"
      },
      "definition": "label_values(model_name)",
      "hide": 0,
      "includeAll": false,
      "label": "model_name",
      "multi": false,
      "name": "model_name",
      "options": [],
      "query": {
        "query": "label_values(model_name)",
        "refId": "StandardVariableQuery"
```

```
        },
        "refresh": 1,
        "regex": "",
        "skipUrlSync": false,
        "sort": 0,
        "type": "query"
      }
    ]
  },
  "time": {
    "from": "now-5m",
    "to": "now"
  },
  "timepicker": {},
  "timezone": "",
  "title": "vLLM",
  "uid": "b281712d-8bff-41ef-9f3f-71ad43c05e9b",
  "version": 2,
  "weekStart": ""
}
```