aws

COMMUNITY DAY

AOTEAROA

K8s & AI Agents

Dipin Thomas

aws
COMMUNITY DAY
AOTEAROA

Intro Video / Demo

aws COMMUNITY DAY AOTEAROA

Agenda

- Building Blocks
  - MCP
  - LangChain & LangGraph
  - Slack demo architecture
- Human in-loop k8s agent - kagent
  - Importance of System Prompts
- Q & A

A Bit About Me

Lead Engineer at Xero

xero

Cloud & DevOps enthusiast who's worked in both MNCs and start-ups — and published a few courses for on KodeKloud

My global Journey

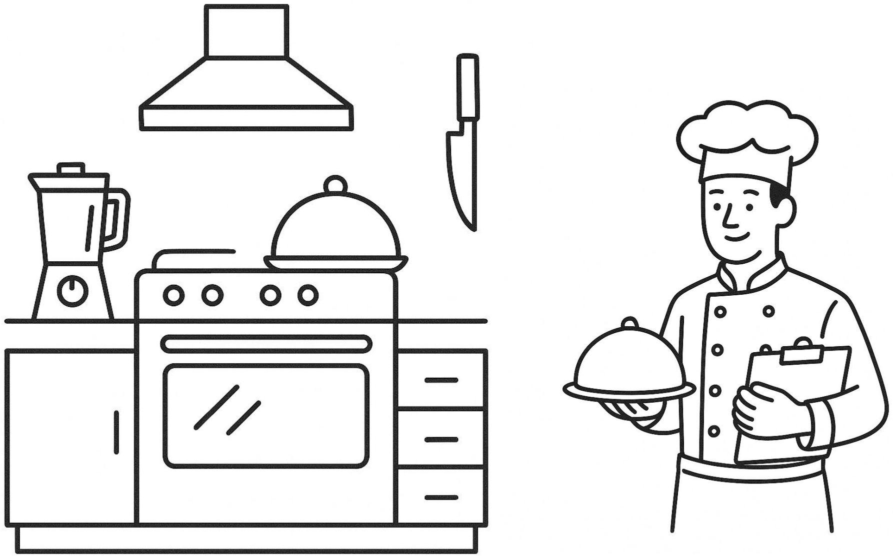
- India
- Malaysia
- Switzerland
- Singapore
- USA
- New Zealand

Building Blocks

MCP
(Model Context Protocol)

aws
COMMUNITY DAY
AOTEAROA

MCP Server



- **list_k8s_resources**
- **list_api_versions**
- **get_k8s_events**
- **get_pod_logs**
    - cluster_name 
    - pod_name 
    - namespace 

LangChain / LangGraph

LangChain

LangChain

Advantages

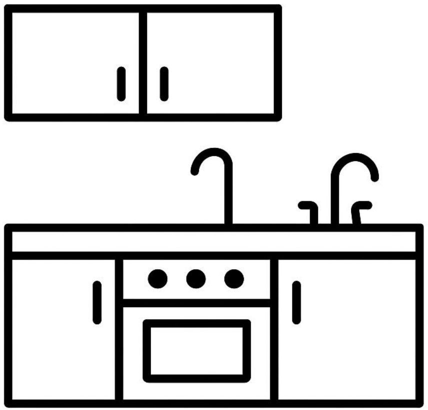
- Standardized recipe
- Chef follows fixed steps
- Consistent, scalable

Trade offs

- Less flexible

LangChain Workflow : Linear Sequential Chain

LangChain



FRAMEWORK

LangChain

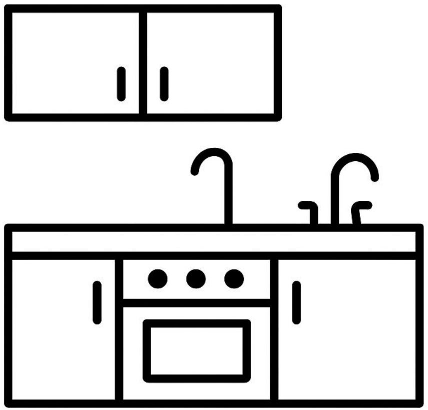
```python
from openai import OpenAI

# Initialize OpenAI client
client = OpenAI()

# Direct API call
response = client.chat.completions.create(
    model="gpt-4.1",
    messages=[{"role": "user", "content": "Hello, OpenAI!"}]
)

print(response.choices[0].message.content)
```

LangChain

```python
from openai import OpenAI

# Initialize OpenAI client
client = OpenAI()

# Direct API call

print(response.choices[0].message.content)

```

```
(.venv) admin@Admins-MacBook-Pro langraph_3node_agent % python3 ppt_test.py
Hello! How can I help you today? 😊
```

LangChain



FRAMEWORK

LangChain

```python
from langchain.chat_models import init_chat_model

# Initialize the model (choose OpenAI GPT-4.1 here)
model = init_chat_model("gpt-4.1", model_provider="openai")

# Send a simple prompt
response = model.invoke("Hello, LangChain!")

# Print the response
print(response)
```

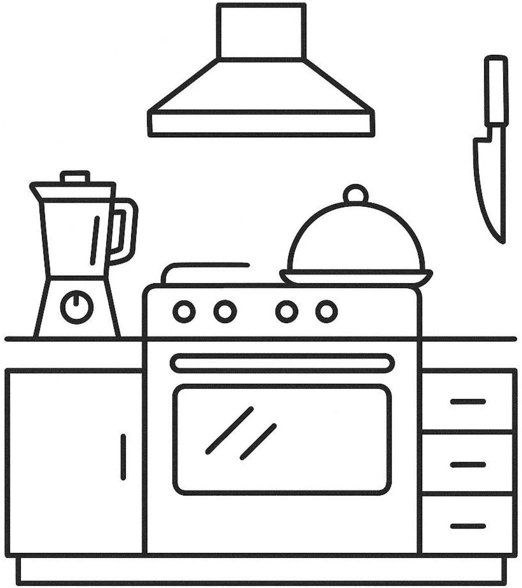
LangChain

```python
from langchain.chat_models import init_chat_model
```

```
(.venv) admin@Admins-MacBook-Pro langraph_3node_agent % python3 ppt_test.py
content="Hello! Just to clarify—I'm not LangChain itself, but I'm familiar with LangChain and can help answer questions about
it, or use it as a topic of conversation. How can I help you today?" al_kwargs={'refusal': None} response_metadata={'t
4, 'completion_tokens_details': {'accepted_predic
tion_tokens': 0, 'audio_tokens': 0, 'reasoning_tokens': 0, 'rejected_prediction_tokens': 0}, 'prompt_tokens_details': {'audio_
tokens': 0, 'cached_tokens': 0}}, 'model_name': 'gpt-4.1-2025-04-14', 'system_fingerprint': 'fp_daf5fcc80a', 'id': 'chatcmpl-C
GPA3fh77eCez3W1vQKKrhYbV9RNl', 'service_tier': 'default', 'finish_reason': 'stop', 'logprobs': None} id='run--84513d51-3c5d-4e
e7-a57c-c11adb470c93-0' usage_metadata={'input_tokens': 12, 'output_tokens': 42, 'total_tokens': 54, 'input_token_details': {'
audio': 0, 'cache_read': 0}, 'output_token_details': {'audio': 0, 'reasoning': 0}}
```

```python
print(response)
```

LangChain



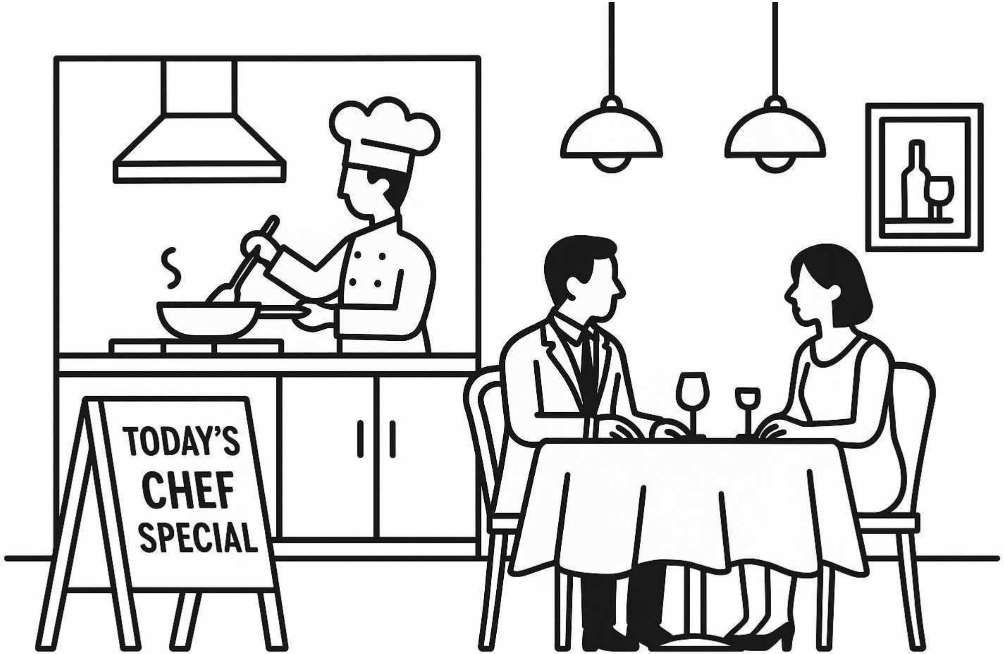
LangChain

Chef = LLM

LangGraph

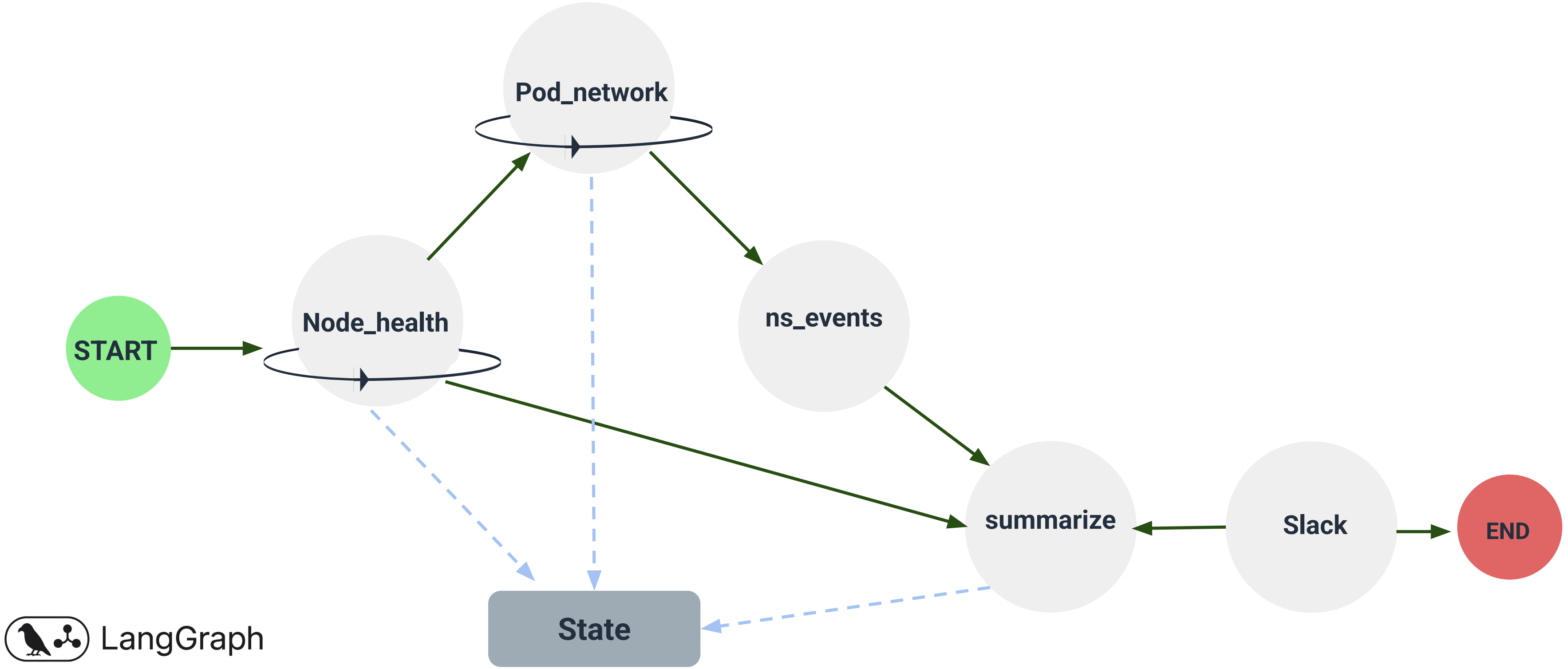
aws
COMMUNITY DAY
AOTEAROA

LangGraph

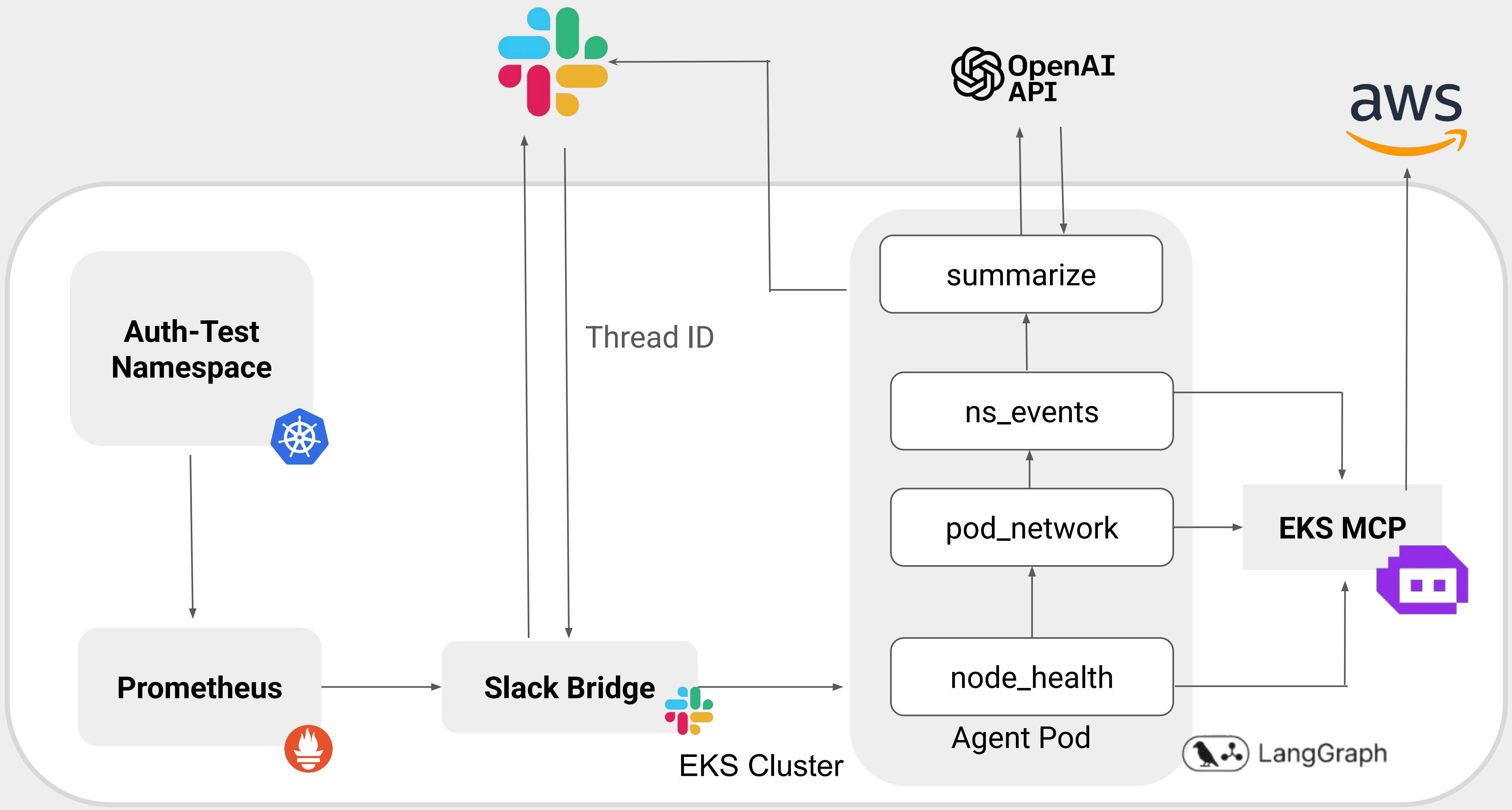


LangGraph

LangGraph


Pod_network
Node_health
START
ns_events
summarize
Slack
END
State
LangGraph

Architecture of Slack Demo


OpenAI
API
aws
Auth-Test
Namespace
Thread ID
summarize
ns_events
pod_network
EKS MCP
Prometheus
Slack Bridge
node_health
Agent Pod
EKS Cluster
LangGraph

```python
def build_graph():
    g = StateGraph(AgentState)

    # Initial diagnostic nodes
    g.add_node("node_health", partial(node_health_checker, eks_client=eks_mcp))
    g.add_node("pod_network", partial(pod_network_checker, eks_client=eks_mcp))
    g.add_node("ns_events", partial(namespace_events_checker, eks_client=eks_mcp))
    g.add_node("summarize", summarizer)
    g.add_node("slack", slack_poster)

    # Initial diagnostic flow
    g.add_edge(START, "node_health")
    g.add_edge("node_health", "pod_network")
    g.add_edge("pod_network", "ns_events")
    g.add_edge("ns_events", "summarize")
    g.add_edge("summarize", "slack")
    g.add_edge("slack", END)

    return g.compile()
```

```python
prompt = (
    "You are a Kubernetes expert analyzing diagnostic data. Create a concise, human-readable summary for DevOps teams.\n\n"
    "IMPORTANT: Base your analysis ONLY on the actual diagnostic data provided. Do not make assumptions.\n\n"
    "ANALYSIS GUIDELINES:\n"
    "- If node health data is available, compare node capacity vs namespace quotas vs actual usage\n"
    "- If pod logs are available, look for application-level errors or startup issues\n"
    "- Prioritize the most critical issues that need immediate attention\n\n"
    "Format your response as:\n"
    "🚨 **ALERT SUMMARY**\n"
    "- Issue: [brief description]\n"
    "- Affected: [cluster/namespace/pod details]\n\n"
    "🔍 **KEY FINDINGS**\n"
    "- [bullet points based on ACTUAL investigation results from diagnostic data]\n"
    "- [include node capacity vs quota limits if available]\n"
    "- [include pod status and logs if available]\n\n"
    "⚡ **RECOMMENDED ACTIONS**\n"
    "- [prioritized action items based on findings]\n\n"
    "📊 **TECHNICAL DETAILS**\n"
    "- [specific numbers and facts from the diagnostic data]\n"
    "- [node specifications if available]\n"
    "- [resource utilization metrics]\n\n"
    f"Alert Context: {json.dumps(alert_info, indent=2)}\n\nDiagnostic Data:\n{findings_text}"
)
```

Thread

k8s-agent APP Yesterday at 14:42

🚨 **ALERT: HPAIsScalingLimited**

Cluster: ai-agents-cluster
Namespace: `auth-test`
Status: FIRING

Reason:

> HPA auth-deployment-hpa in namespace auth-test has hit min/max replica limits.

1 reply

k8s-agent APP Yesterday at 14:43

🚨 **ALERT SUMMARY**
- Issue: ResourceQuota violation preventing pod creation in the `auth-test` namespace.
- Affected: Deployment `auth-deployment` in namespace `auth-test`.

🔍 **KEY FINDINGS**
- The Horizontal Pod Autoscaler (HPA) for `auth-deployment` has reached its minimum and maximum replica limits, impacting scaling.
- The `auth-test` namespace has a memo
- Multiple attempts to create new pods w
- Events indicate successful scaling down

blocked due to exceeding the memory quota, sp

⚡ **RECOMMENDED ACTIONS**
1. Review and potentially increase the memory quota for the `auth-test` namespace to accommodate the required pod scaling.
2. Investigate the current memory usage to identify any optimizations or reductions that could free up resources.

kload and resource availability.

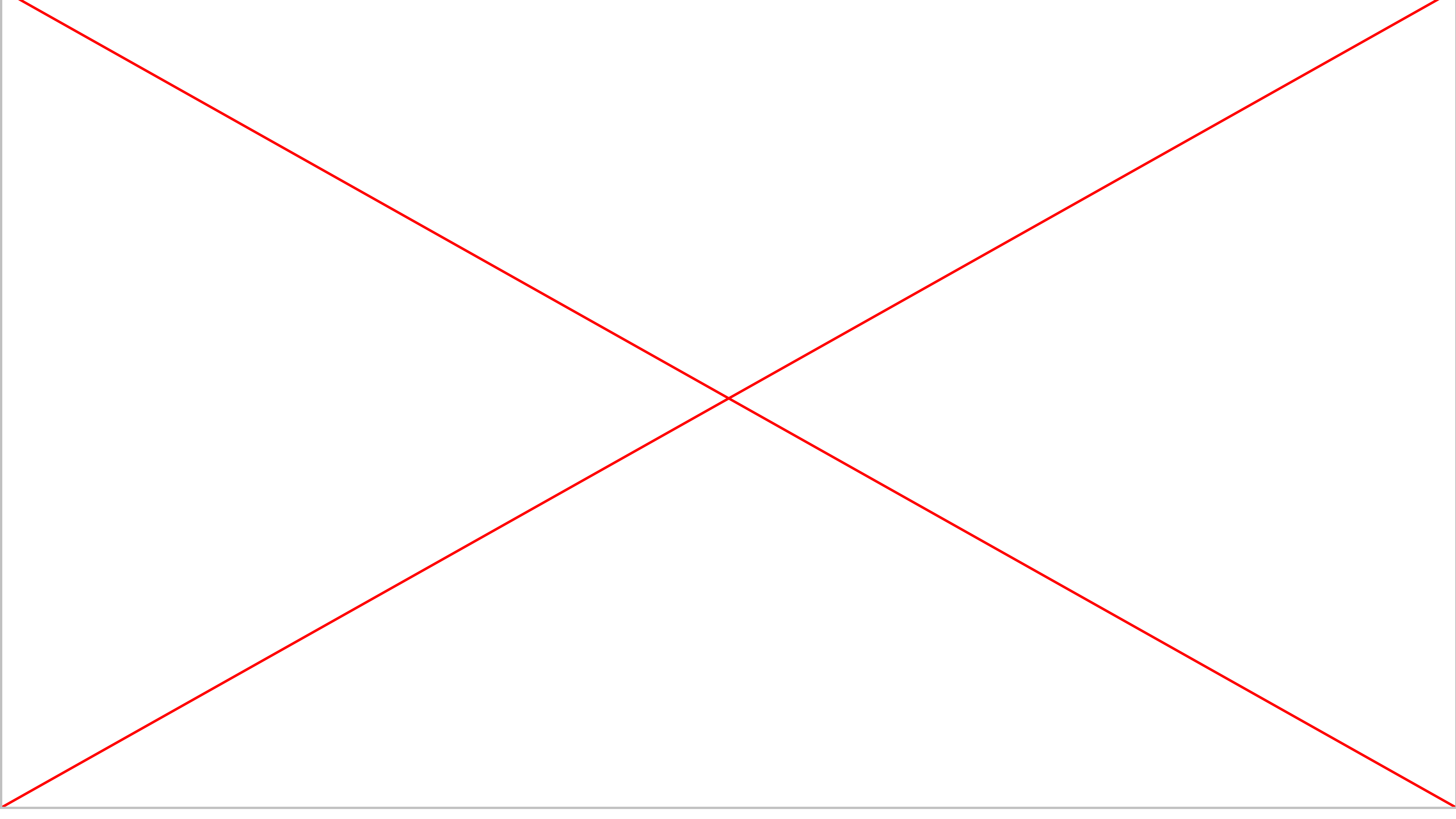
📊 **TECHNICAL DETAILS**
- Current memory usage in `auth-test`: 384Mi
- Memory quota limit in `auth-test`: 500Mi

nts.

- Multiple errors logged for pod creation attempts due to exceeding the quota, indicating a critical need for resource management.

kagent

kagent Video



kagent

- An open-source framework for running AI agents in Kubernetes.
- GitOps way of deploy agents and mcp servers
- Default integration with widely used tools like
    - kubernetes
    - prometheus
    - kgateway etc

```yaml
apiVersion: kagent.dev/v1alpha2
kind: Agent
metadata:
  name: eks-agent
  namespace: kagent

spec:
  declarative:
    a2aConfig:
      skills:
      - name: EKS Monitoring
        description: Monitor and analyze AWS EKS cluster performance and health
      - name: EKS Troubleshooting
        description: Diagnose and troubleshoot EKS cluster issues
      - name: CloudWatch Analysis
        description: Analyze CloudWatch logs and metrics for EKS clusters

    systemMessage: |
      You are EKSAssist, an AI Agent specialized in AWS EKS management,
      monitoring, troubleshooting, and CloudWatch analysis.
      - EKS Expert Knowledge
      - CloudWatch Mastery
      - Security-First Mindset
      - Cost Optimization
      - Clear Communication
```

```yaml
  tools:
  - name: mcp-eks
    type: McpServer
    toolNames:
    - get_eks_insights
    - list_k8s_resources
    - get_pod_logs
  - name: mcp-cloudwatch
    type: McpServer
    toolNames:
    - get_metric_data
    - analyze_log_group
    - get_active_alarms

description: An AWS EKS Expert AI Agent for monitoring,
  troubleshooting, and CloudWatch analysis.
type: Declarative
```

Scan to connect with me!

https://shorturl.at/HR83P

Go to
pigeonhole.at

Enter passcode

AWSCDNZ25

Scan here!

https://shorturl.at/HR83P

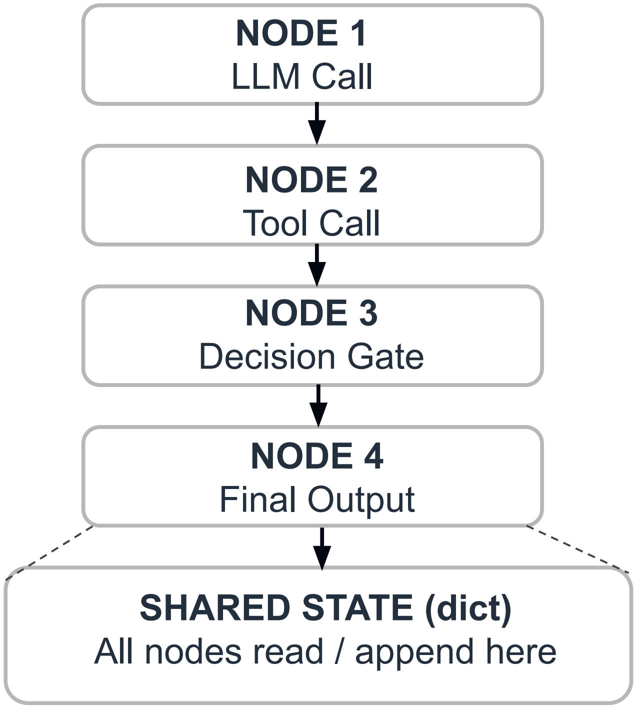
aws COMMUNITY DAY AOTEAROA

aws
COMMUNITY DAY
AOTEAROA

LangGraph

LangGraph

WORKFLOW (DAG)



Q&A

aws COMMUNITY DAY AOTEAROA

K8s & AI Agents

- Dipin Thomas

LangGraph

Fine Dining (LangGraph)

Chef = LLM

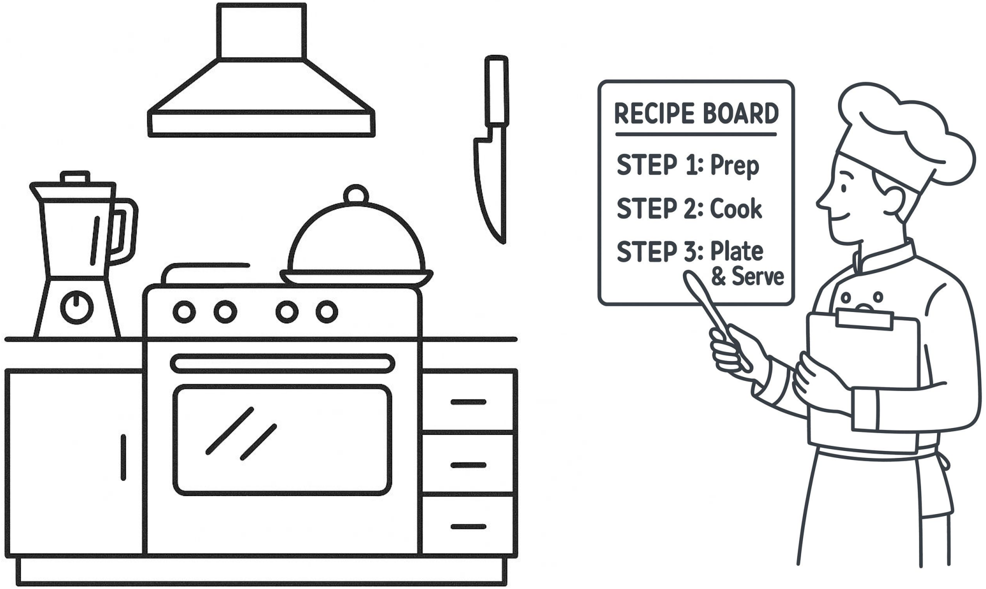
LangGraph

- Building Blocks
    - MCP
    - LangChain & LangGraph
    - Slack demo architecture

- Human in-loop k8s agent - kagent

- Q & A

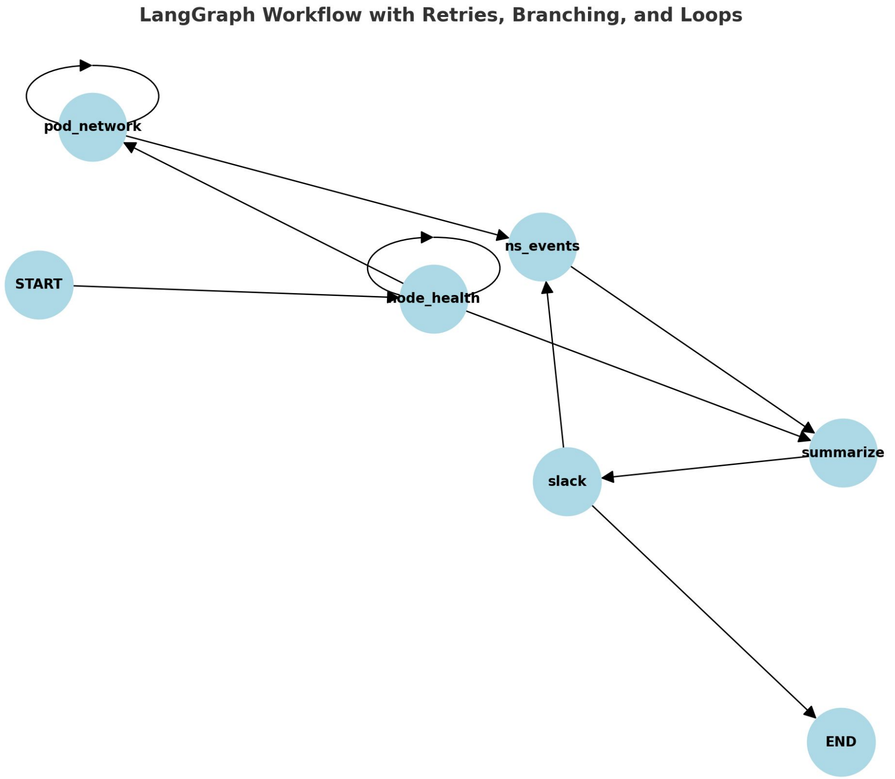
LangChain



LangGraph

Chef = LLM

LangGraph



LangGraph

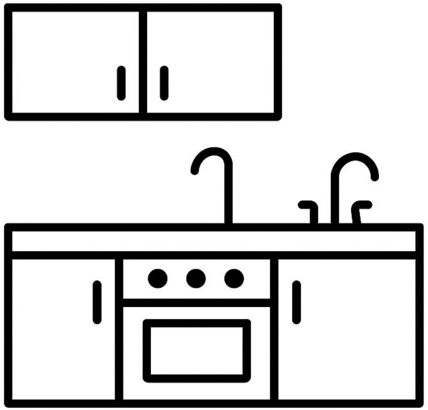
Your Guide Today

Lead Engineer at Xero

Cloud & DevOps enthusiast who's worked in both MNCs and start-ups — and published a few courses for on KodeKloud

My global Journey

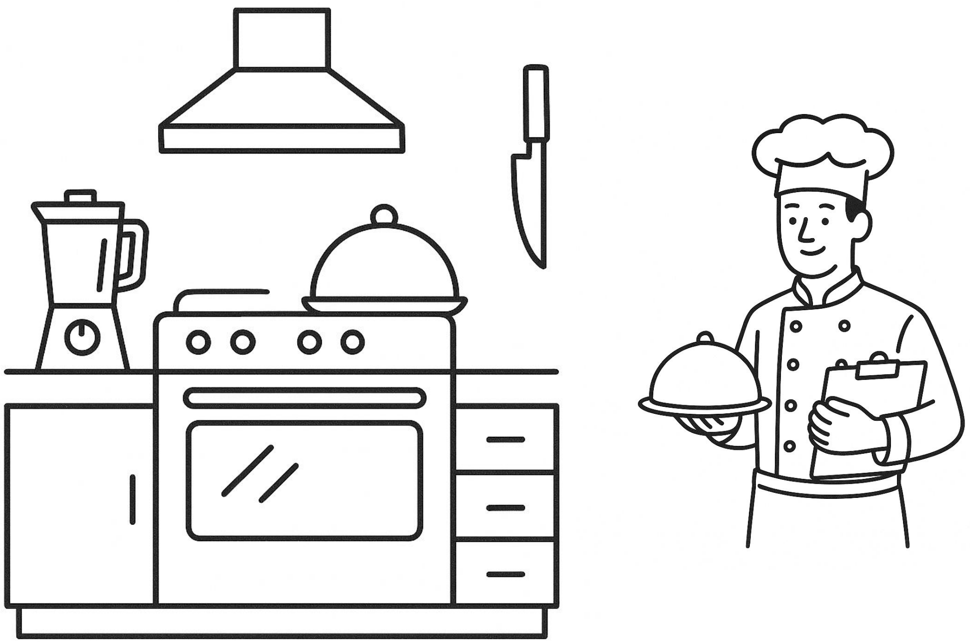
- 🇮🇳 India
- 🇲🇾 Malaysia
- 🇨🇭 Switzerland
- 🇸🇬 Singapore
- 🇺🇸 USA
- 🇳🇿 New Zealand



FRAMEWORK

LangChain

```python
import getpass
import os

if not os.environ.get("OPENAI_API_KEY"):
  os.environ["OPENAI_API_KEY"] = getpass.getpass("Enter API key for OpenAI: ")

from langchain.chat_models import init_chat_model

model = init_chat_model("gpt-4.1", model_provider="openai")
```



AWS EKS MCP

- list_k8s_resources
- list_api_versions
- get_k8s_events
- get_pod_logs
  - cluster_name
  - pod_name
  - namespace

LangGraph

WORKFLOW (DAG)

NODE 1
LLM Call

↓

NODE 2
Tool Call

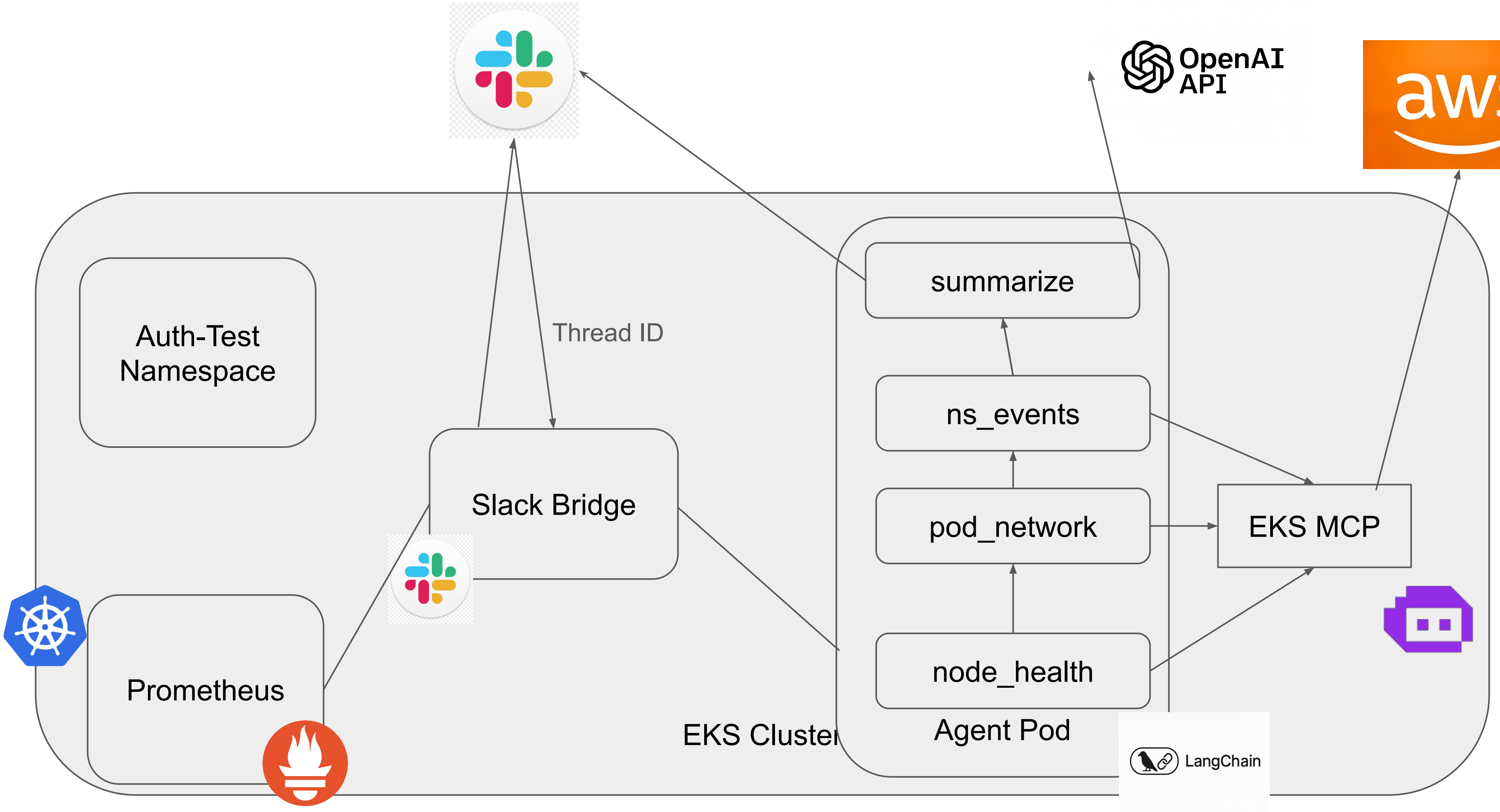
↓

NODE 3
Decision Gate

↓

NODE 4
Final Output

SHARED STATE (dict)
All nodes read/
append here



OpenAI
API
aws
Auth-Test
Namespace
Thread ID
summarize
ns_events
pod_network
node_health
EKS MCP
Slack Bridge
Prometheus
EKS Cluster
Agent Pod
LangChain

Thread

k8s-agent APP Yesterday at 14:42

🚨 ALERT: HPAIsScalingLimited

Cluster: ai-agents-cluster
Namespace: `auth-test`
Status: FIRING

Reason:

> HPA auth-deployment-hpa in namespace auth-test has hit min/max replica limits.

1 reply

k8s-agent APP Yesterday at 14:43

🚨 **ALERT SUMMARY**
- Issue: ResourceQuota violation preventing pod creation in the `auth-test` namespace.
- Affected: Deployment `auth-deployment` in namespace `auth-test`.

🔍 **KEY FINDINGS**
- The Horizontal Pod Autoscaler (HPA) for `auth-deployment` has reached its minimum and maximum replica limits, impacting scaling.
- The `auth-test` namespace has a memory quota of 500Mi, with 384Mi currently in use.
- Multiple attempts to create new pods were blocked due to exceeding the memory quota, specifically for requests of 128Mi each.
- Events indicate successful scaling down of replicas but also highlight repeated failures to create new pods due to quota constraints.

⚡ **RECOMMENDED ACTIONS**
1. Review and potentially increase the memory quota for the `auth-test` namespace to accommodate the required pod scaling.
2. Investigate the current memory usage to identify any optimizations or reductions that could free up resources.
3. Monitor the HPA settings to ensure they align with the expected workload and resource availability.

📊 **TECHNICAL DETAILS**
- Current memory usage in `auth-test`: 384Mi
- Memory quota limit in `auth-test`: 500Mi
- Number of replicas scaled down from 20 to 6 due to resource constraints.
- Multiple errors logged for pod creation attempts due to exceeding the quota, indicating a critical need for resource management.

LangChain Workflow: Linear Sequential Chain