AZURE DAY ROME 2025

Un gateway per domarli tutti: dominare le architetture AI con Azure API Management

Massimo Bonanni

Microsoft

PORINI
A DGS COMPANY

BC Soft
consulenza sistemi informativi

KEY2

LOBRA

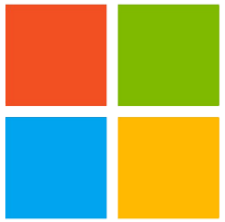
UNIKEY

Packt

JET BRAINS

NET CODE

Thanks to



Microsoft

PORINI
A DGS COMPANY

AZURE DAY

Introduction to Azure API Management

AZURE DAY

Introduction to Azure API Management

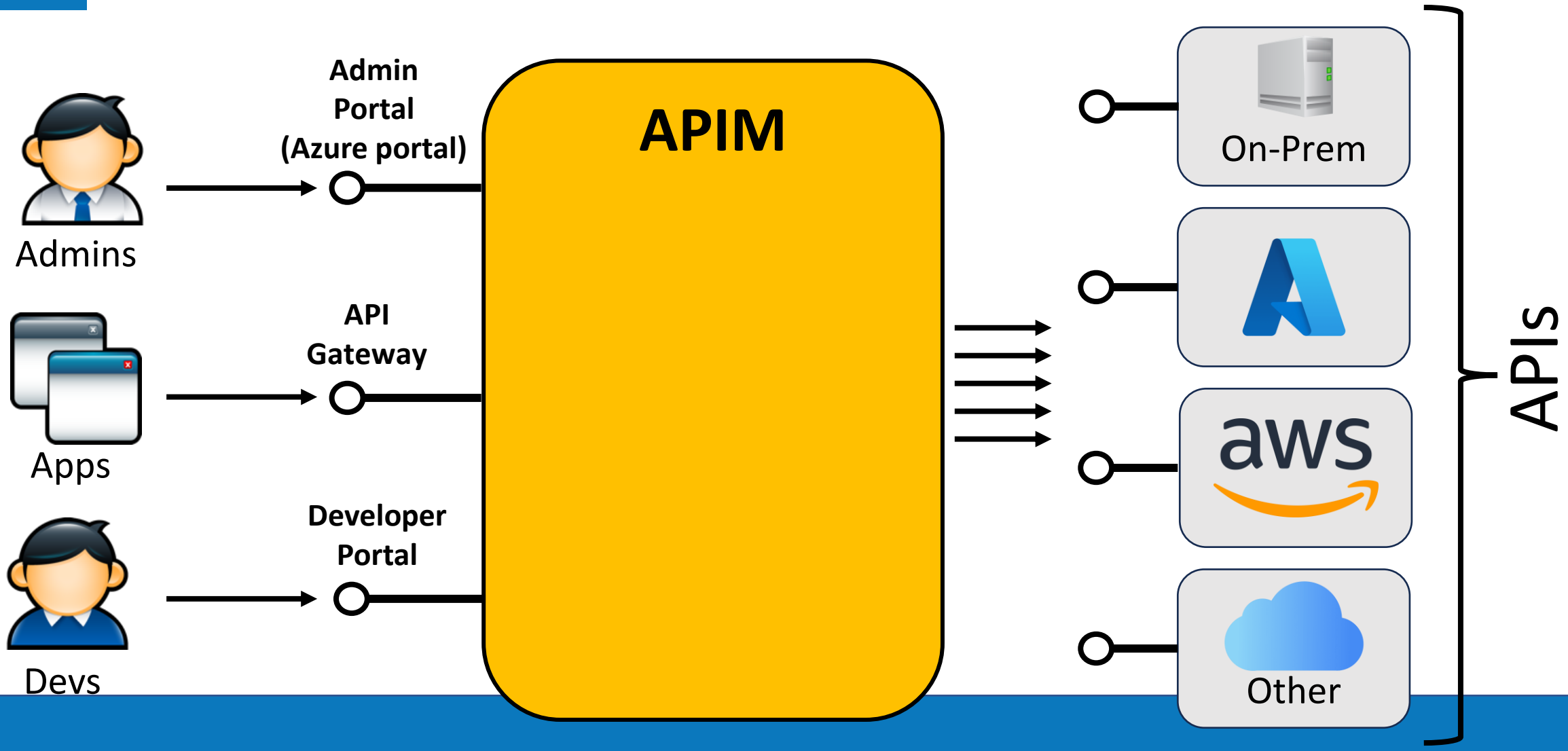
What is Azure API Management?

- A fully managed, hybrid, multi-cloud platform for managing APIs across all environments.
- Supports the complete API lifecycle: creation, publication, security, monitoring, and analytics.

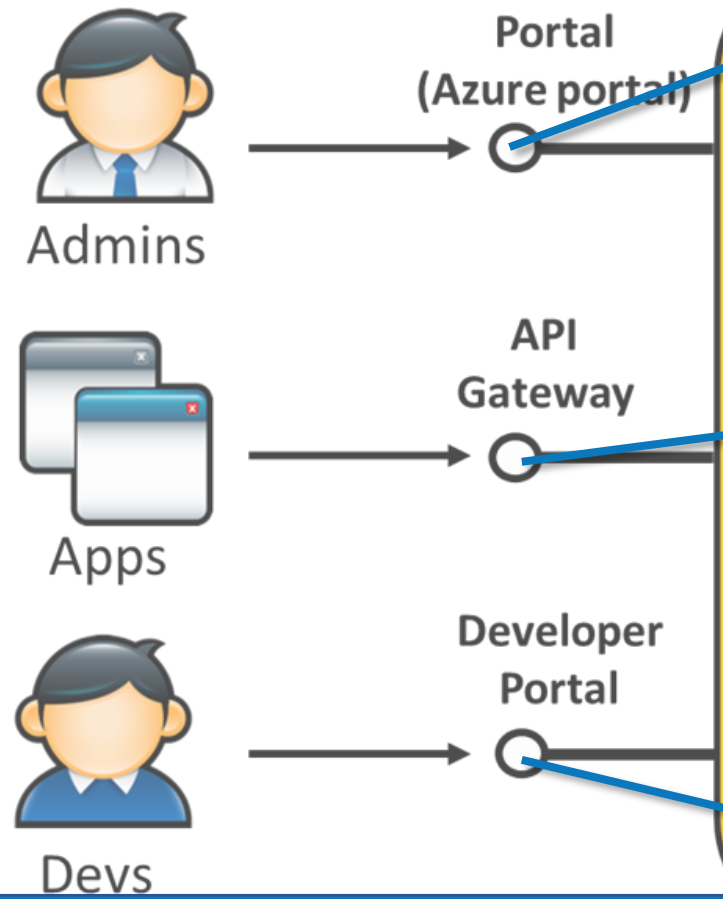
Why Use APIM?

- Abstracts backend complexities from API consumers.
- Enables secure and scalable API exposure.
- Facilitates API discovery and consumption by internal and external users.

Core Components of Azure APIM



Core Components of Azure APIM



- Provides full access to APIM capabilities via Azure Portal, CLI, PowerShell, REST API, and SDKs.
- Used for provisioning, configuration, and policy management.

- Acts as a facade to backend services.
- Handles request routing, security, throttling, caching, and observability.
- Supports both managed and self-hosted deployments.

- An automatically generated, customizable website for API documentation and testing.
- Allows developers to discover APIs, onboard, and manage subscriptions.

AZURE DAY

Key Concepts

APIs

- Define operations available to app developers.
- Map to backend services and operations.

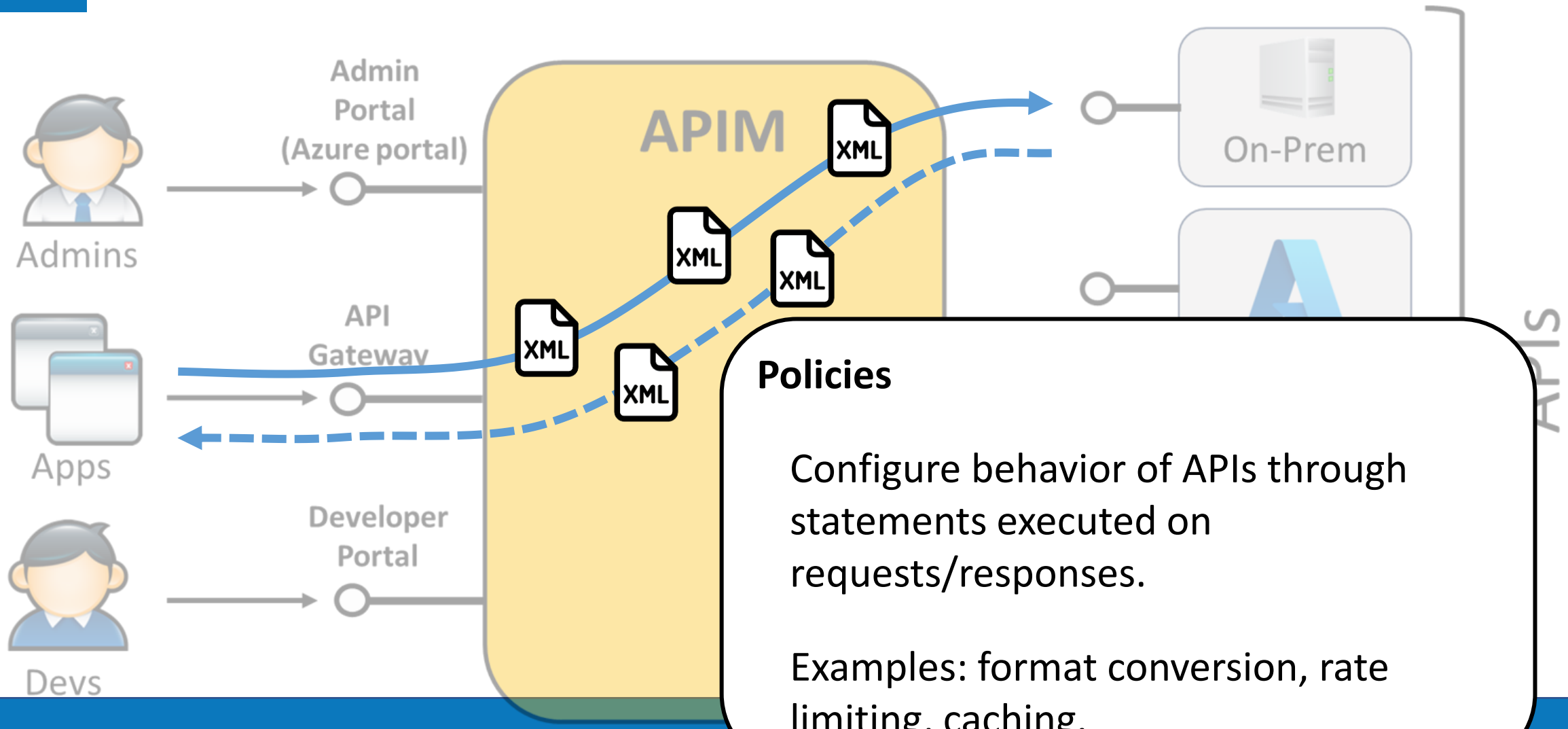
Products

- Group one or more APIs for publication.
- Can be open or require subscriptions.

Users and Groups

- Manage access and visibility to APIs.
- Support custom groups and integration with Microsoft Entra ID.

Policies

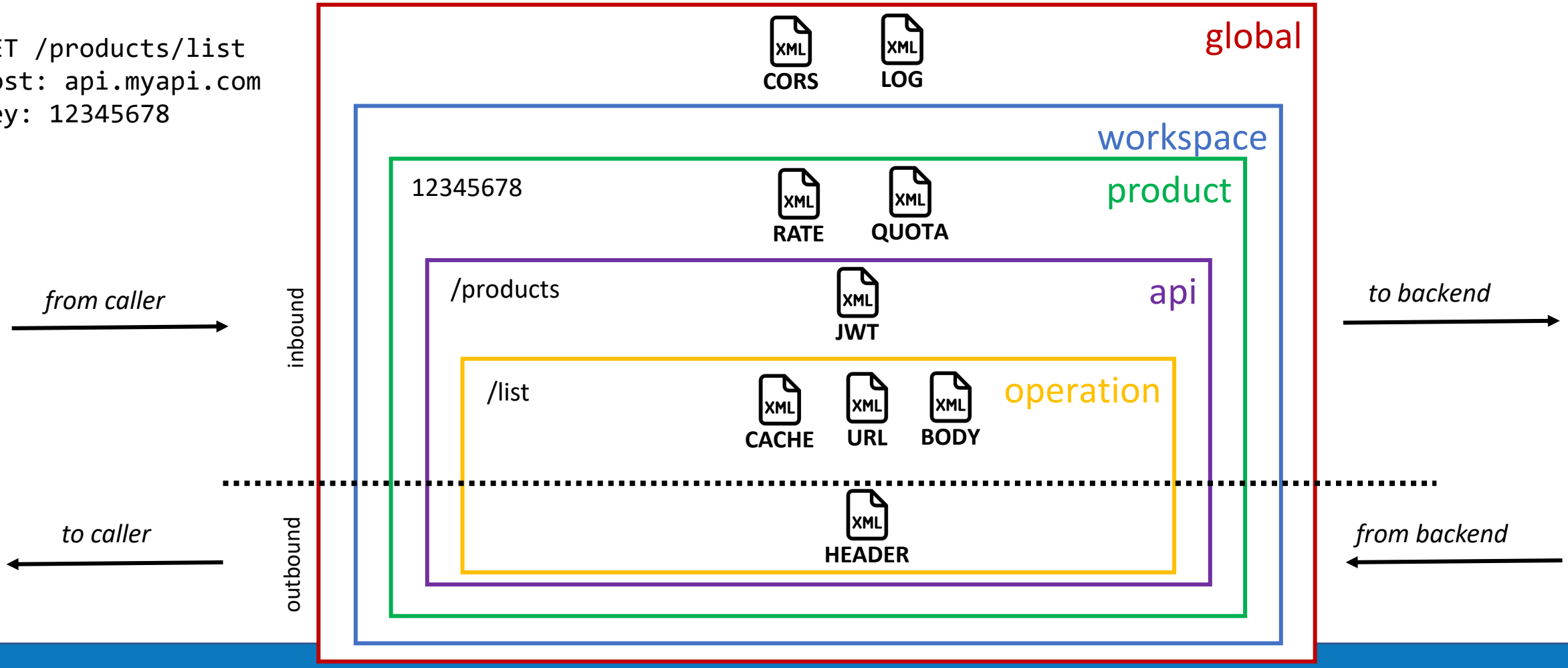


Policies

Configure behavior of APIs through statements executed on requests/responses.

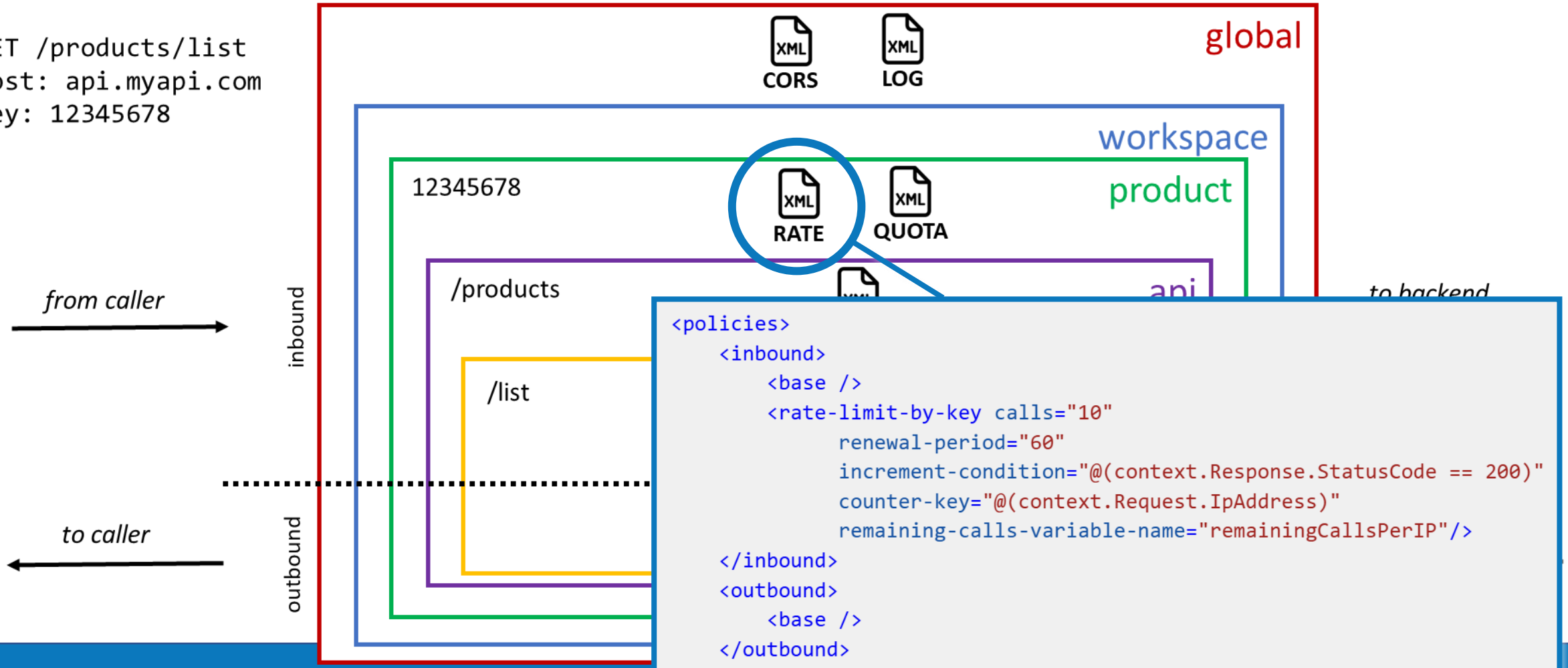
Examples: format conversion, rate limiting, caching.

Policy scopes



Policy scopes

```
GET /products/list
Host: api.myapi.com
Key: 12345678
```




```xml
<policies>
    <inbound>
        <base />
        <rate-limit-by-key calls="10"
              renewal-period="60"
              increment-condition="@(context.Response.StatusCode == 200)"
              counter-key="@(context.Request.IpAddress)"
              remaining-calls-variable-name="remainingCallsPerIP"/>
    </inbound>
    <outbound>
        <base />
    </outbound>
</policies>
```

AZURE DAY

APIM & AI

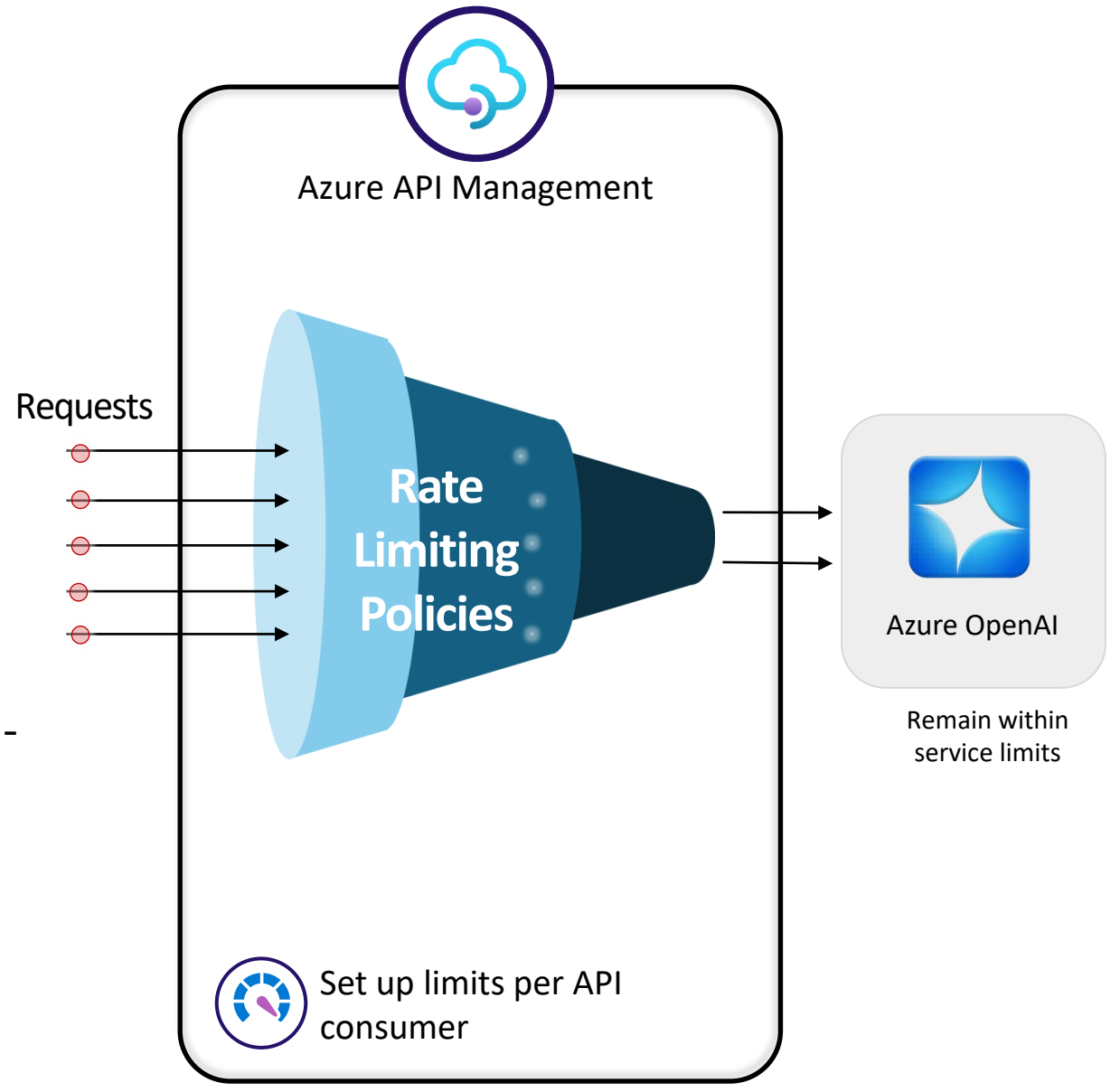
AZURE DAY

Challenges in managing Generative AI APIs

- How is token usage tracked across multiple applications?
- How is the API key securely distributed across multiple applications?
- How is load distributed across multiple endpoints?
- Can you ensure that the committed capacity in PTUs is exhausted first?

Token Limit Policy

- **Rate Limiting**: You can specify a maximum number of tokens per minute (TPM) for each API consumer.
- **Quota Management**: Set limits over specific periods (hourly, daily, weekly, monthly, yearly) to control long-term usage.
- **Flexible Counters**: Limits can be configured based on Subscription Key, IP Address, Custom-defined keys
- **Pre-calculation of Tokens**: The policy can estimate the number of prompt tokens before sending the request to minimize unnecessary API calls.

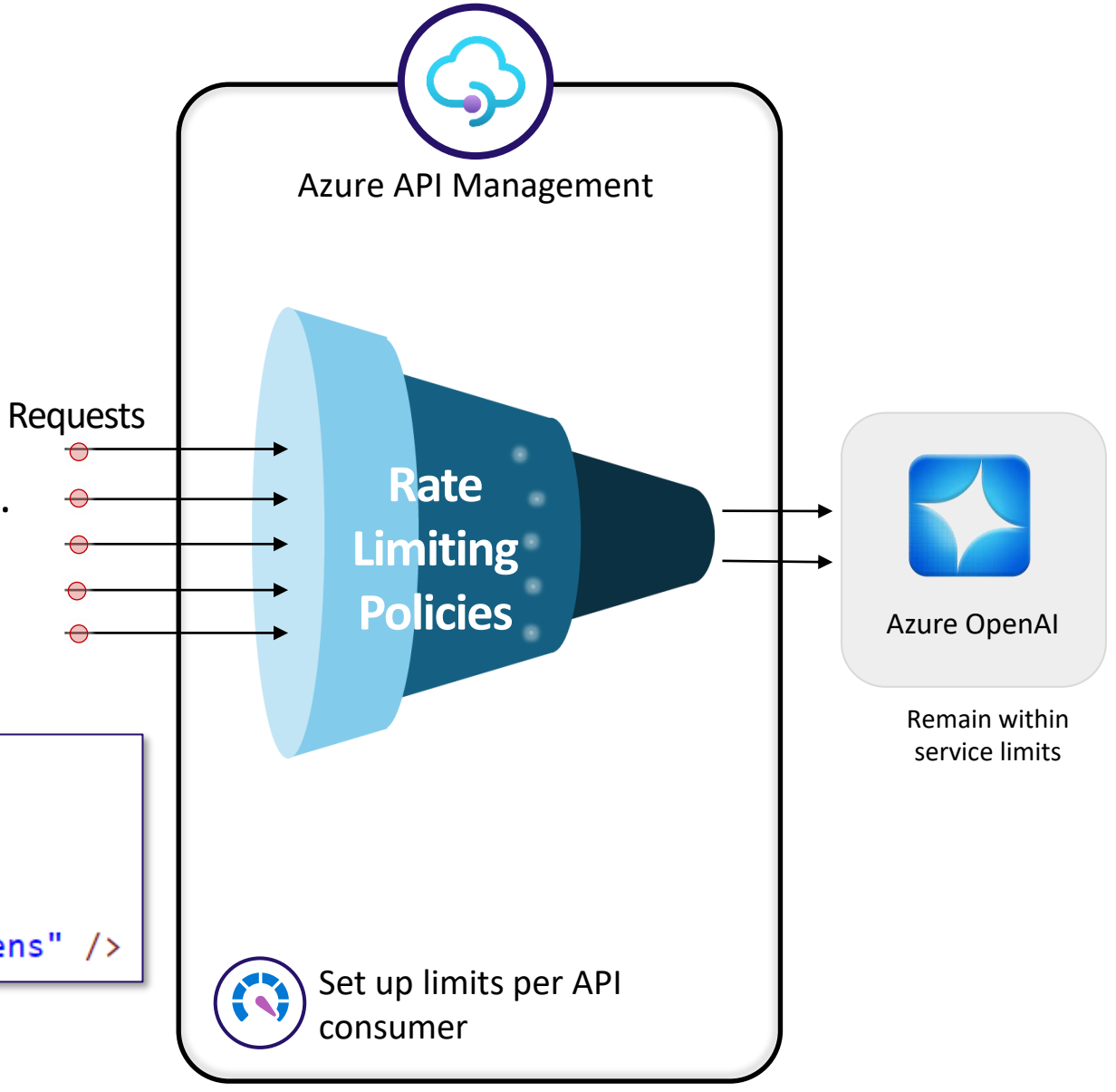


Token Limit Policy

Benefits:

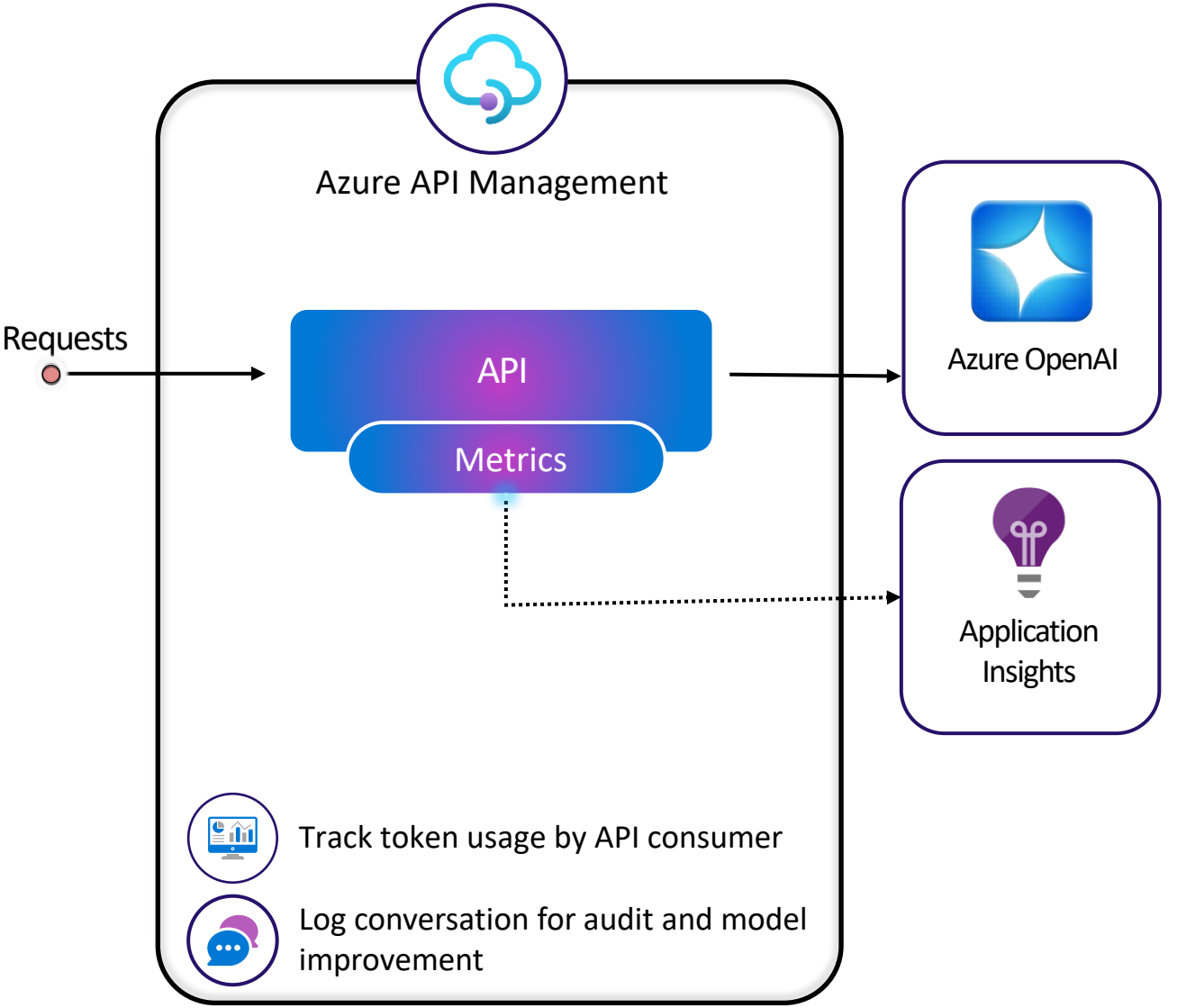
- Prevents single apps from monopolizing token quotas.
- Ensures fair usage across multiple applications.
- Helps track and limit token usage effectively.

```
<azure-openai-token-limit
    counter-key="@(context.Request.IpAddress)"
    tokens-per-minute="5000"
    estimate-prompt-tokens="false"
    remaining-tokens-variable-name="remainingTokens" />
```



Emit token metric policy

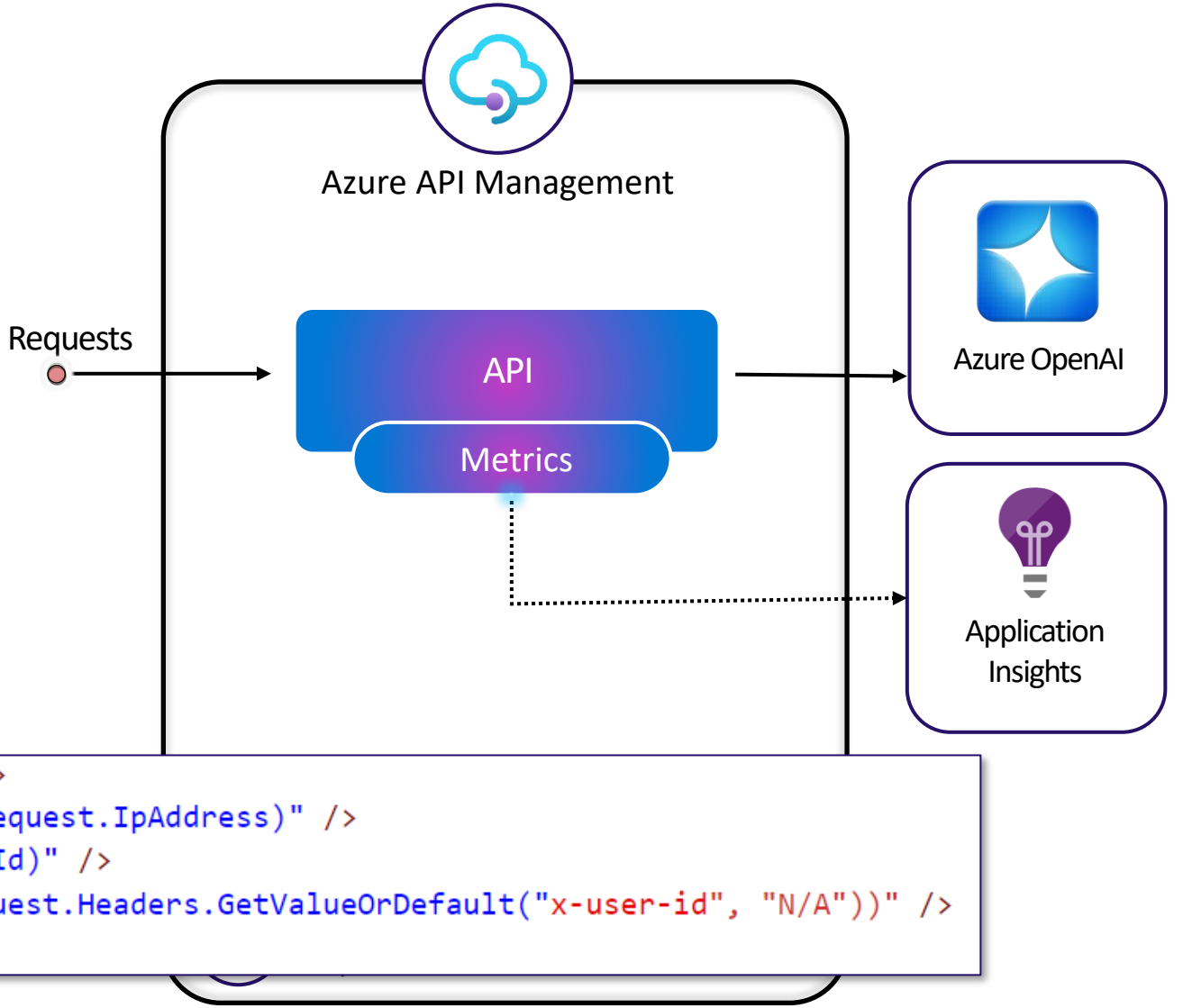
- **Metrics Collection:** Tracks token usage, including prompt, completion, and total token counts.
- **Integration with Application Insights:** Sends detailed metrics to Application Insights for visualization and analysis.
- **Custom Dimensions:** Allows you to segment and analyze data by dimensions such as: Client IP Address, API ID, User ID
- **Chargeback and Monitoring:** Facilitates chargeback calculations, usage monitoring, and capacity planning.



Emit token metric policy

Benefits:

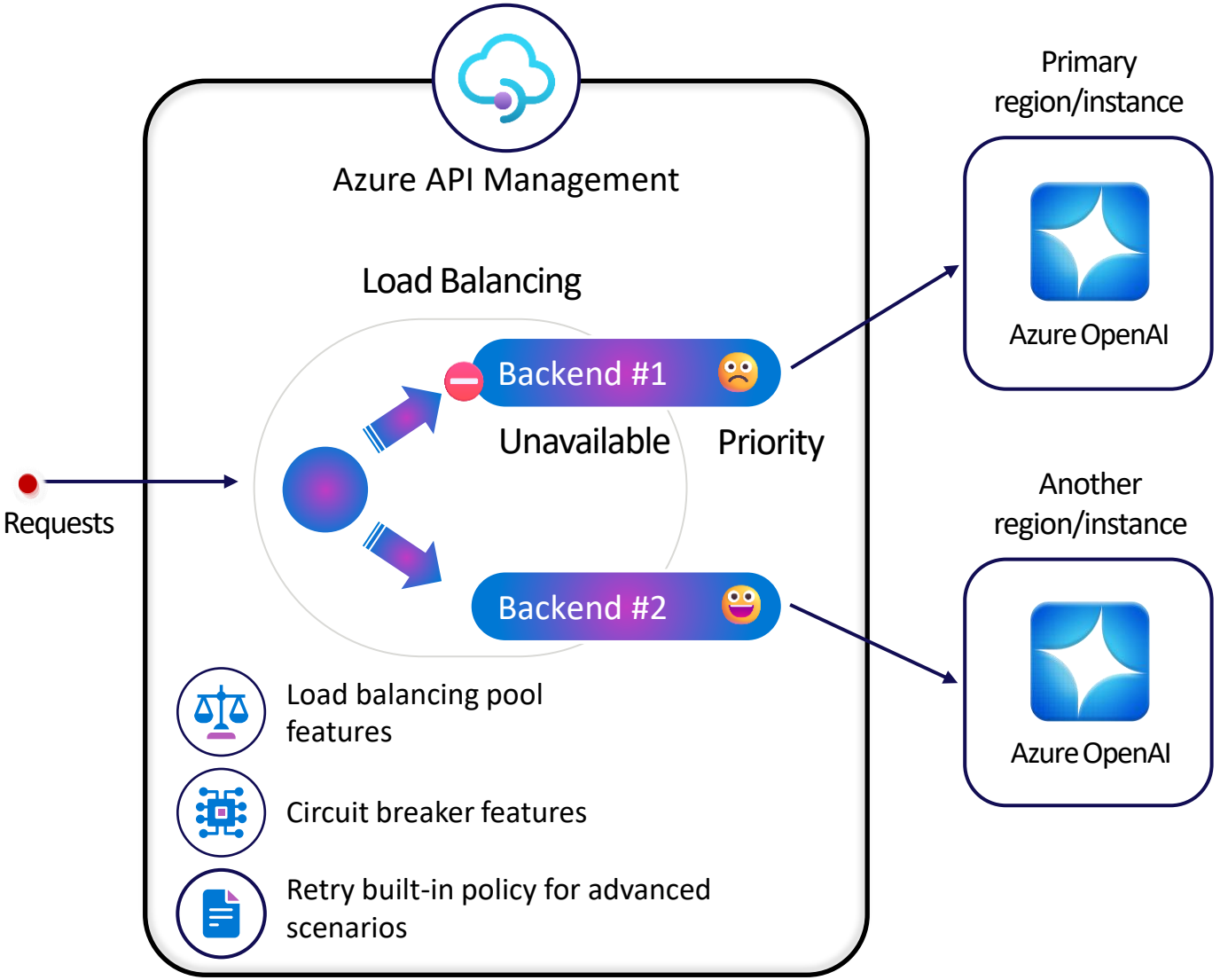
- Helps understand token consumption across different apps or users.
- Enables more precise allocation and optimization of AI resources.
- Assists in billing and chargeback by providing accurate usage data.




```
<azure-openai-emit-token-metric namespace="openai">
    <dimension name="Client IP" value="@(context.Request.IpAddress)" />
    <dimension name="API ID" value="@(context.Api.Id)" />
    <dimension name="User ID" value="@(context.Request.Headers.GetValueOrDefault("x-user-id", "N/A"))" />
</azure-openai-emit-token-metric>
```

Backend Load Balancer and Circuit Breaker

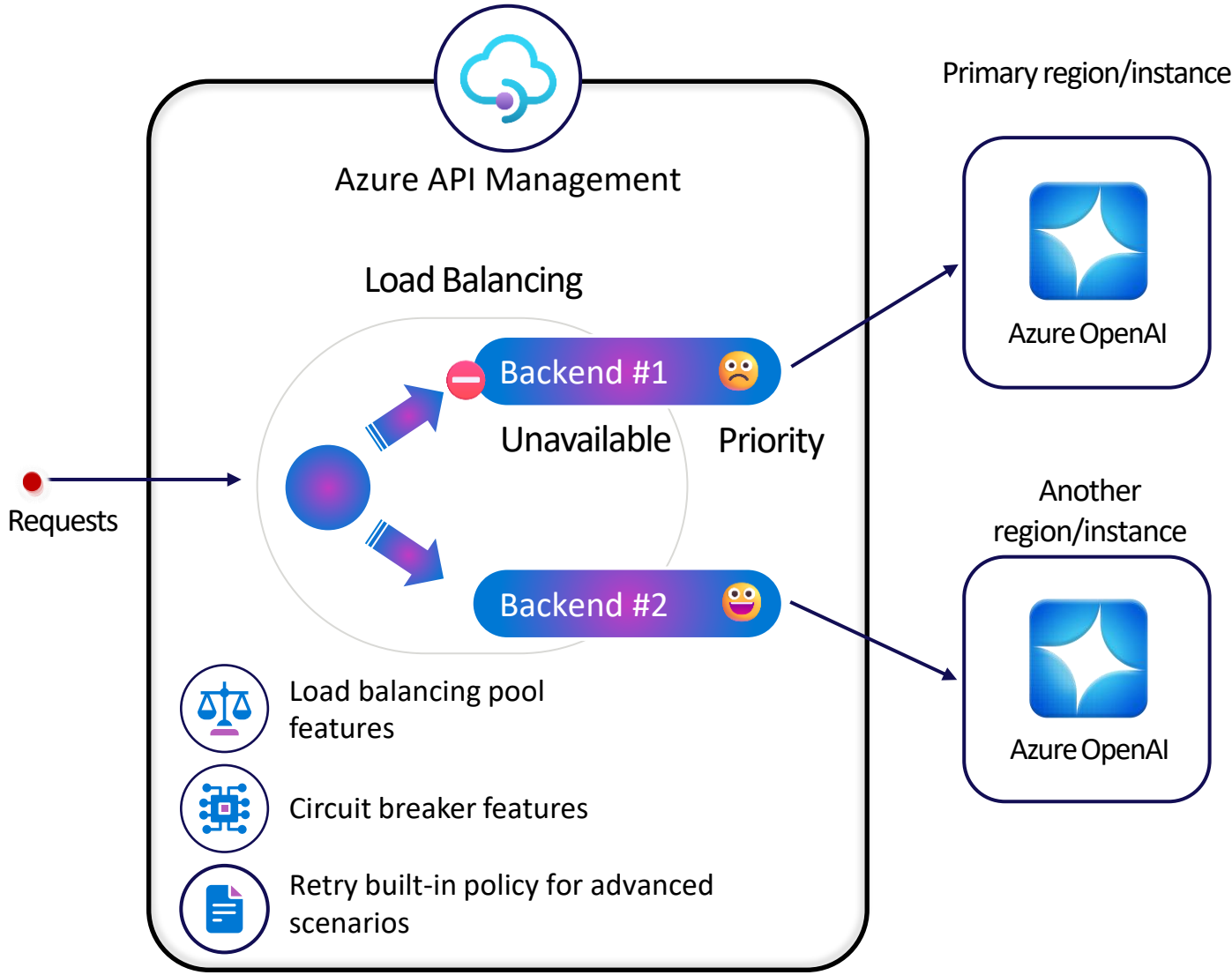
- **Load Balancer:**
  - Distributes requests across multiple Azure OpenAI Service endpoints.
  - Supports different balancing strategies like: Round-Robin, Weighted, Priority-Based.
  - Ensures that Provisioned Throughput Units (PTUs) are utilized before falling back to pay-as-you-go instances.
- **Circuit Breaker:**
  - Monitors the health of backend services.
  - Stops forwarding requests to unresponsive or degraded endpoints.
  - Uses dynamic trip duration based on the "Retry-After" header from the backend.
  - Ensures smooth recovery by resuming traffic only when the endpoint is healthy.



Backend Load Balancer and Circuit Breaker

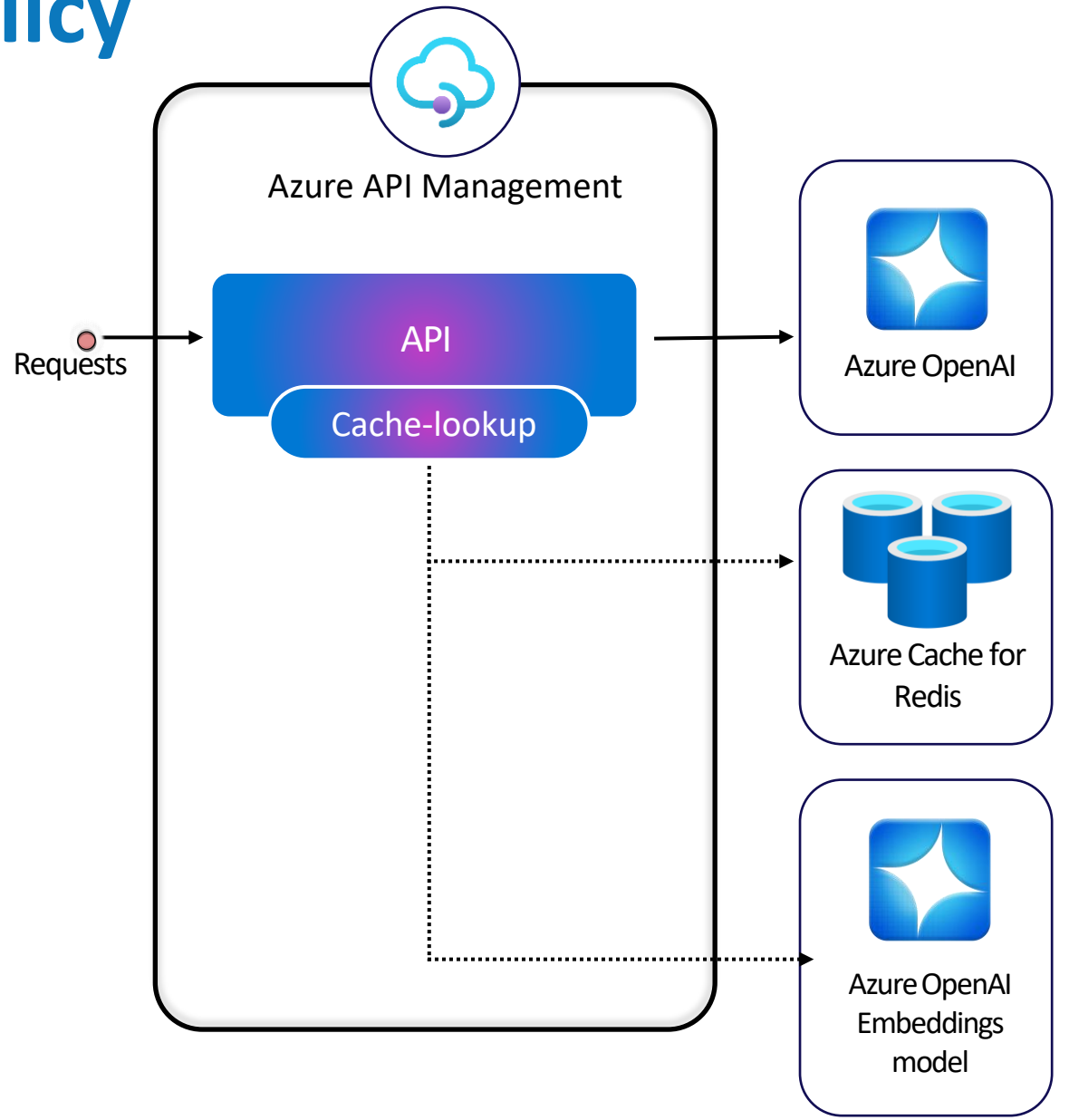
Benefits:

- High Availability: Minimizes downtime by dynamically rerouting traffic.
- Performance Optimization: Efficiently balances load, reducing the risk of overloading any single endpoint.
- Resilient Architecture: Automatically handles failures and maintains service continuity.



Semantic caching policy

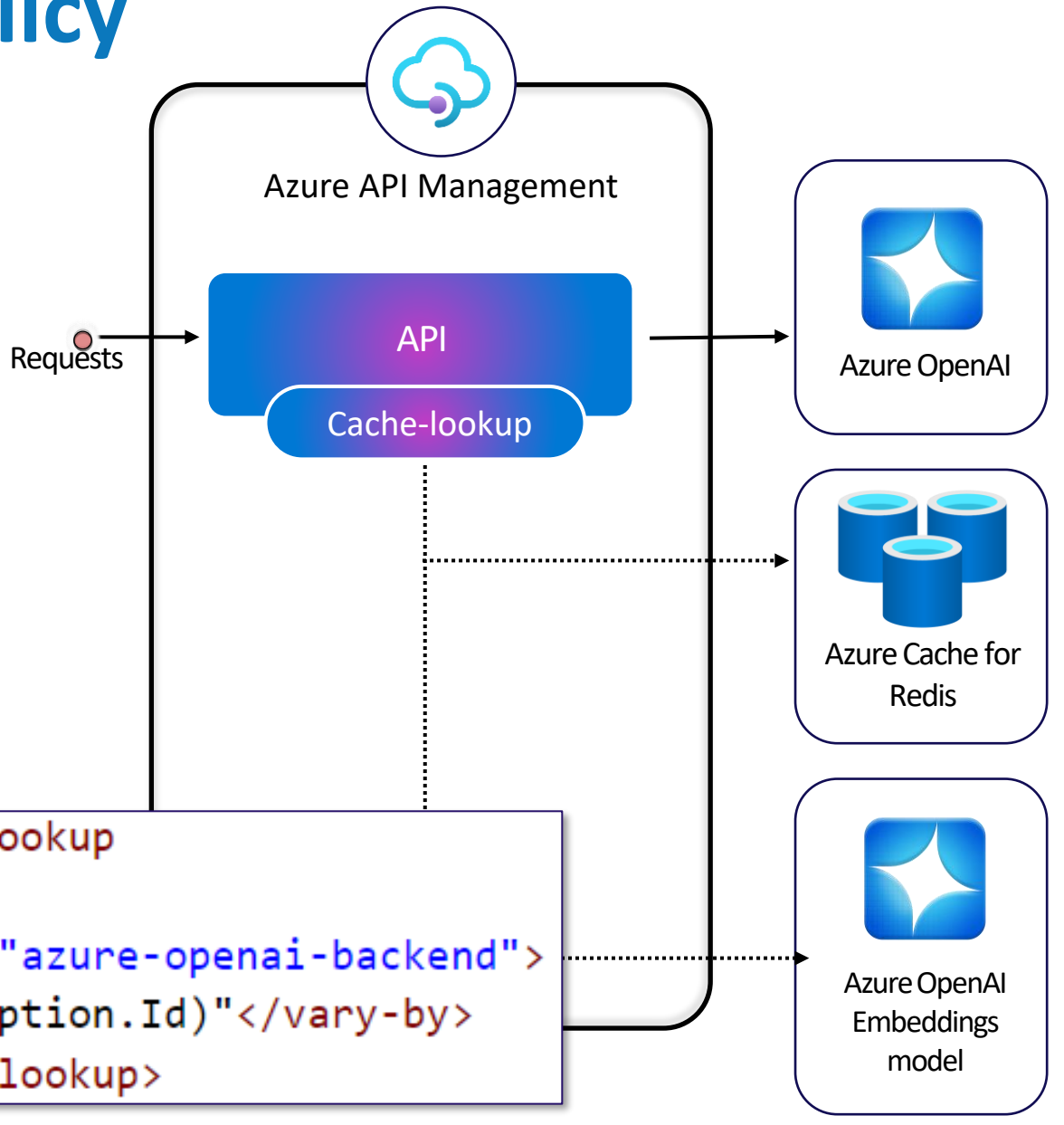
- **Response Caching:**
  - Stores responses to frequently used or similar prompts, reducing the number of requests to the backend.
  - Uses semantic similarity to determine when cached responses can be reused.
- **Integration with Redis:**
  - Uses Azure Redis Enterprise or Azure Managed Redis for storing and retrieving cached completions.
  - Supports any external cache compatible with RediSearch.
- **Semantic Matching:**
  - Uses the Azure OpenAI Service Embeddings API to calculate similarity between new and cached prompts.
  - Retrieves cached responses when similarity is within a defined threshold.



Semantic caching policy

Benefits:

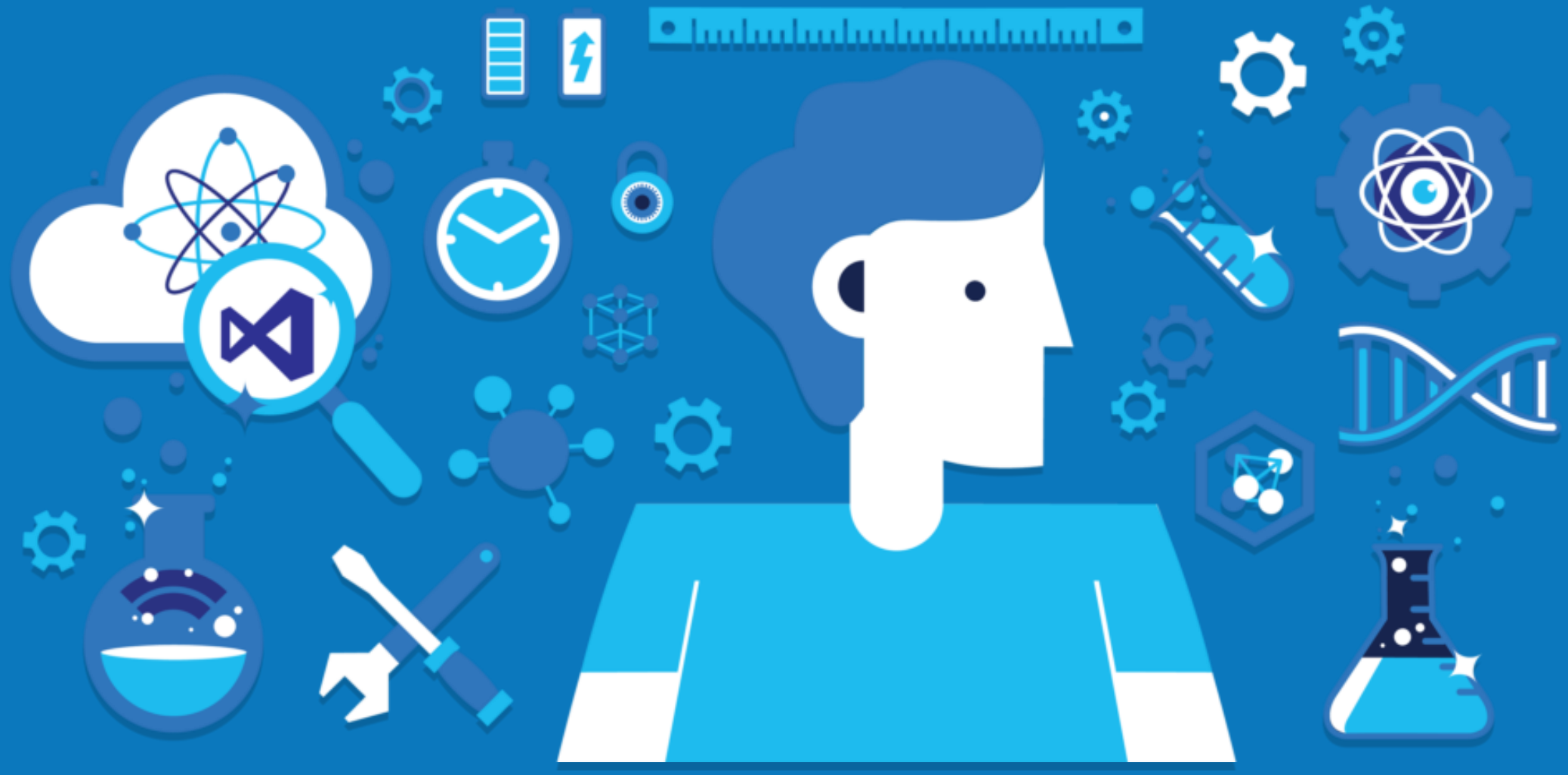
- Reduced Token Consumption: Minimizes API calls by reusing cached responses.
- Improved Performance: Reduces latency, especially for repetitive or similar prompt completions.
- Cost Efficiency: Decreases the overall usage of OpenAI tokens.




```
<azure-openai-semantic-cache-lookup
        score-threshold="0.05"
        embeddings-backend-id="azure-openai-backend">
    <vary-by>@(context.Subscription.Id)"</vary-by>
</azure-openai-semantic-cache-lookup>
```

AZURE DAY

DEMO: APIM & Azure OpenAI



AZURE DAY

Azure API Management policy toolkit

An open-source toolkit comprising C# libraries and developer tools designed to simplify the creation, testing, and management of Azure API Management policies

- Replaces Razor/XML editing: Instead of authoring policy XML in Razor with embedded C#, developers can now write policies entirely in C#, improving readability and maintainability
- Faster feedback loops: Policies compile locally, enabling early syntax validation and generation of XML equivalents—no more deploying to live APIM instances for each tweak

```csharp
[Document]
public class ApiOperationPolicy : IDocument
{
    public void Inbound(IInboundContext context)
    {
        context.Base();
        context.SetHeader("X-Hello", "World");
    }
}
```

```xml
<policies>
    <inbound>
        <base/>
        <set-header name="X-Hello" exists-action="override">
            <value>World</value>
        </set-header>
    </inbound>
</policies>
```

[Introducing Azure API Management Policy Toolkit | Microsoft Community Hub]

315% ROI + AI-Readiness: The Business Case for Azure API Management

ROI & Productivity

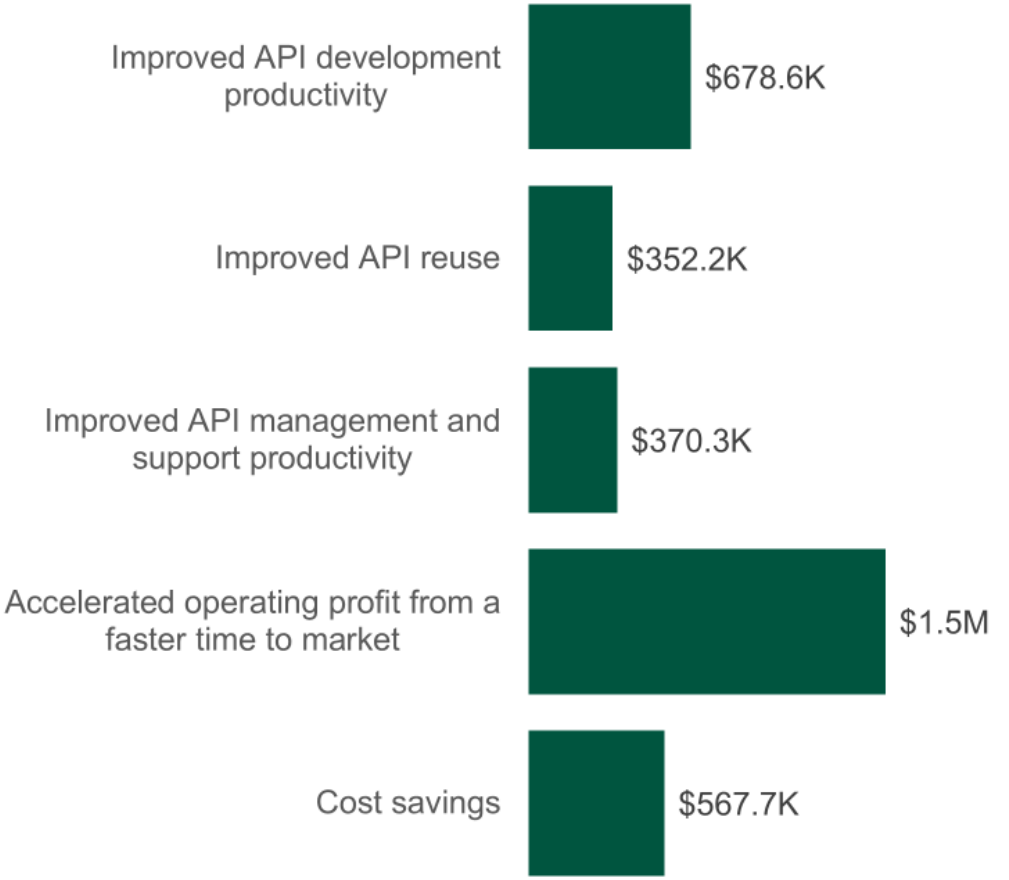
- 315% ROI over 3 years
- 50% faster time-to-market
- 80% productivity boost

AI Governance Gateway

- Centralized visibility & control
- Rate limiting, auditing, cost management

Security & Integration

- Unified secure gateway
- Threat detection & SIEM integration
- Deep Azure ecosystem synergy



Forrester Study Finds 315% ROI with Azure API Management and a Path to AI Readiness | Microsoft Community Hub

AZURE DAY

Vote my session

Massimo Bonanni

Senior Technical Trainer @ Microsoft

massimo.bonanni@microsoft.com

References

- Implement API Management - Training | Microsoft Learn
- AI gateway capabilities in Azure API Management | Microsoft Learn
- Introducing Azure API Management Policy Toolkit | Microsoft Community Hub
- Forrester Study Finds 315% ROI with Azure API Management and a Path to AI Readiness | Microsoft Community Hub
- Hello from AI Gateway workshop | AI Gateway workshop
- microsoft/AzureOpenAI-with-APIM: Deploy APIM. Auto-configure it to work with your Azure Open AI.