Microsoft Azure

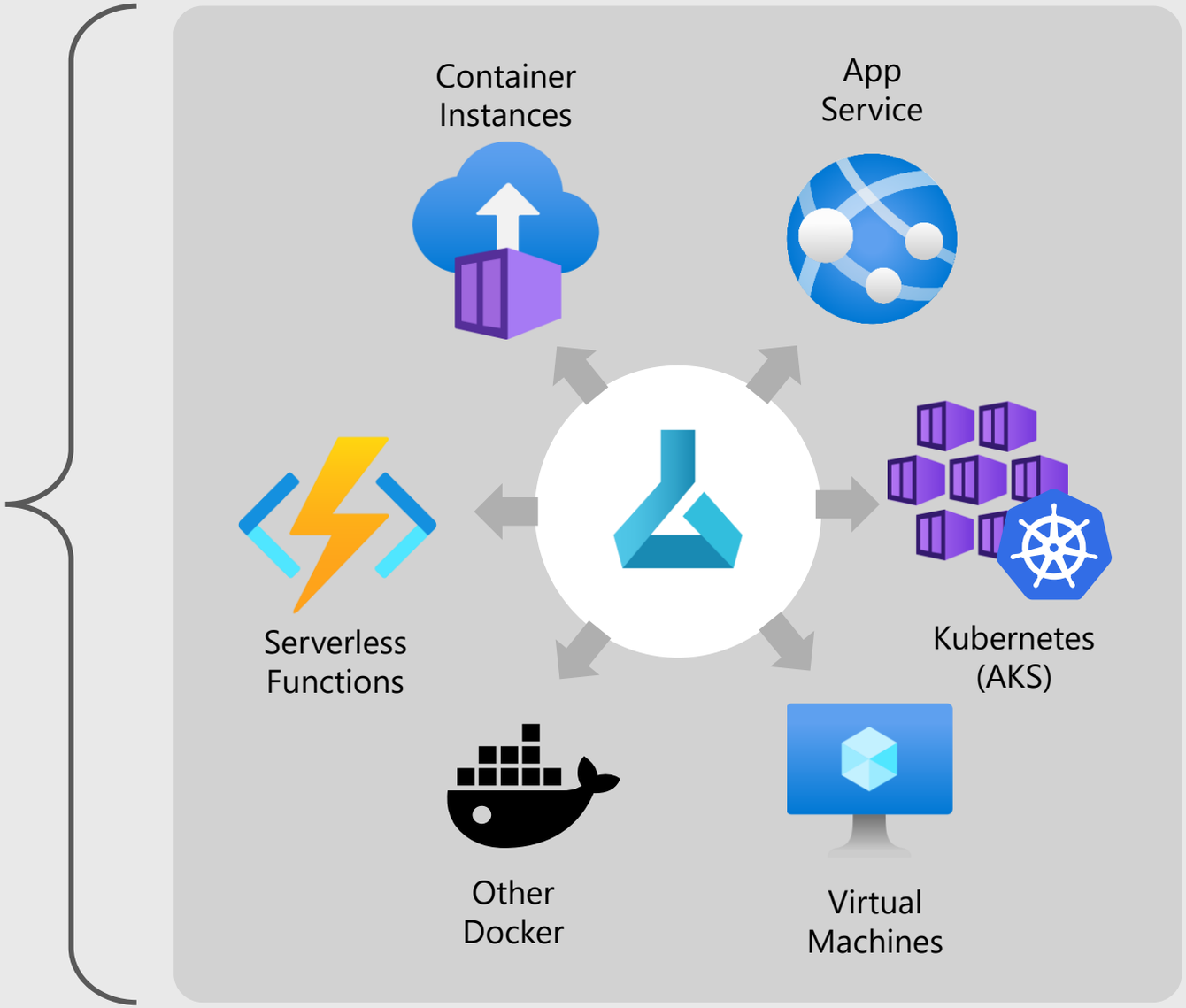
Securely Inferencing Models with APIs

Ben Coleman
Senior Cloud Architect & Engineer
@BenCodeGeek

Operationalizing a Model

- Client App
- End User
- System / Code
- Other

HTTP
(REST)



HTTP Endpoint Consideration

Access

Internal vs Public

Rate limiting

Content Filtering (WAF)

CORS

Encryption

TLS / HTTPS

Client Certificate

Mutual TLS

Authentication

API Keys

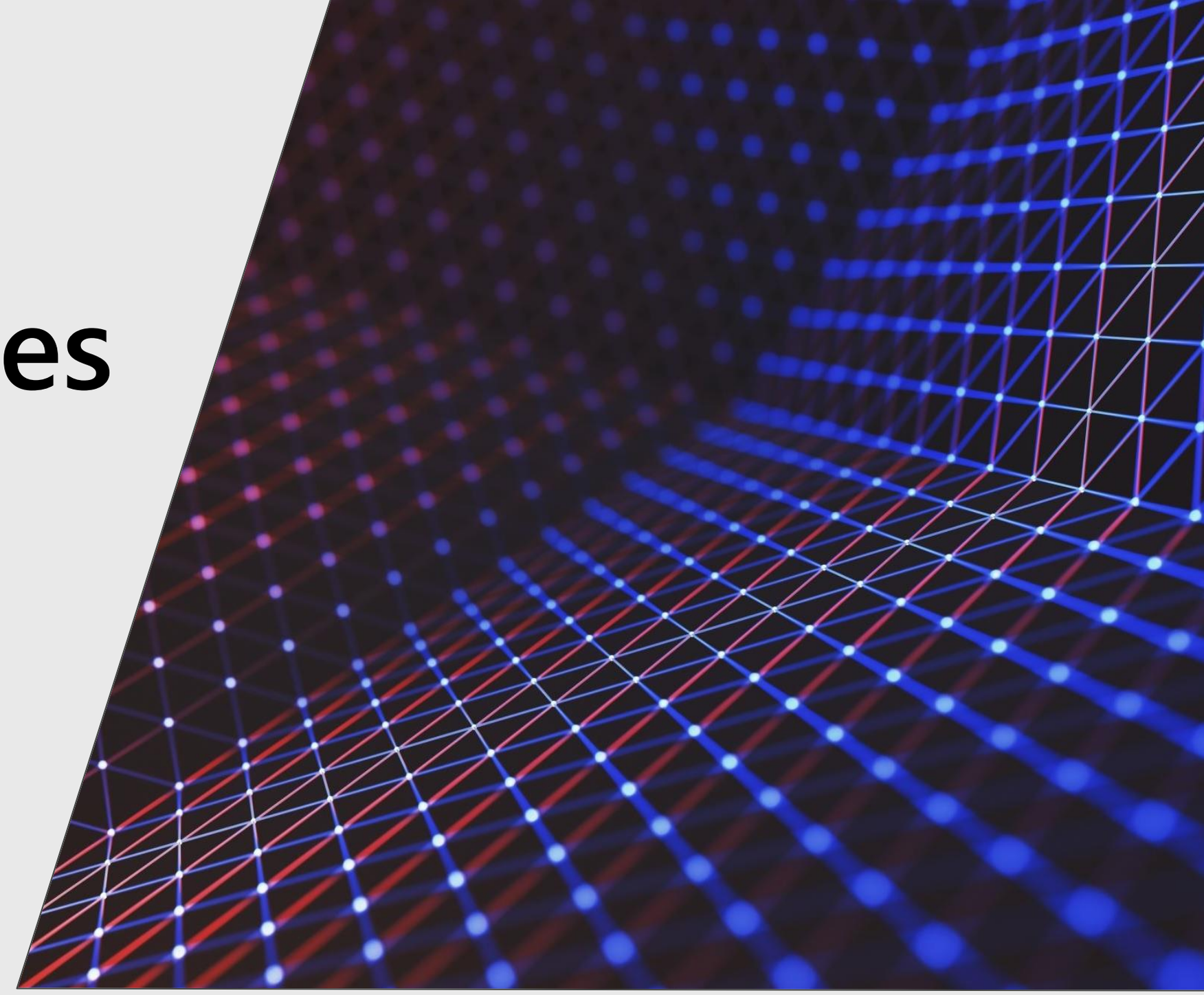
Token based (OAuth)

HMAC

B[illegible]ut[illegible]s[illegible]rd

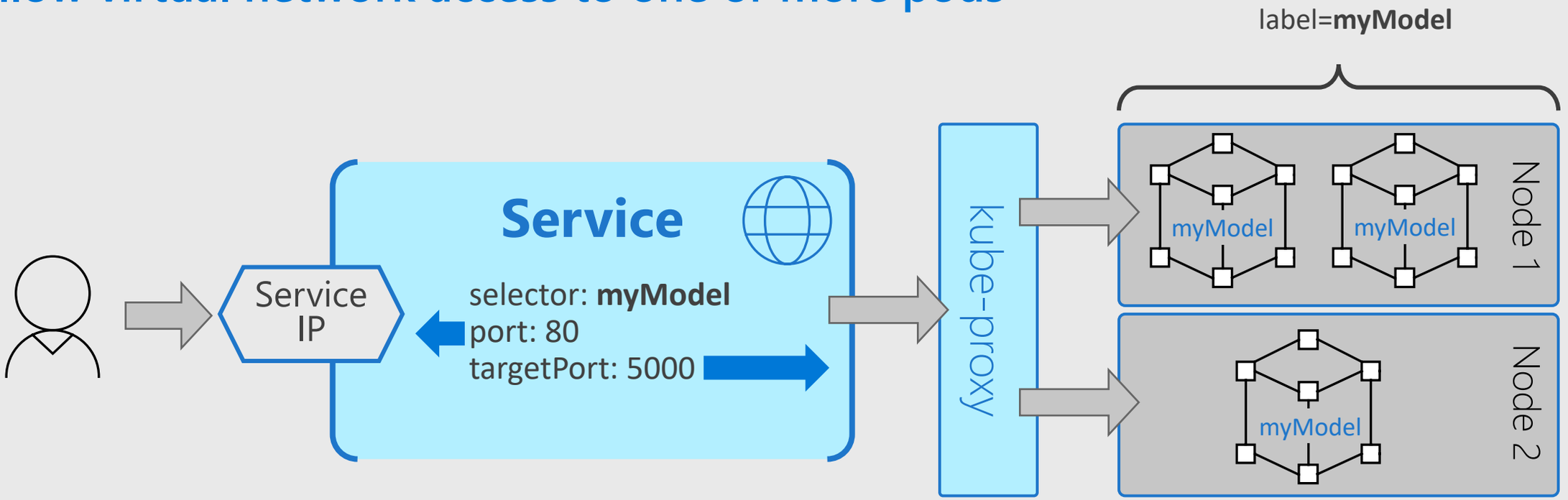
DON'T DO THIS

Kubernetes



Kubernetes Services – Illustrated

Allow virtual network access to one or more pods



EXTERNAL

LoadBalancer

Uses cloud provider to present an external load-balanced IP

INTERNAL

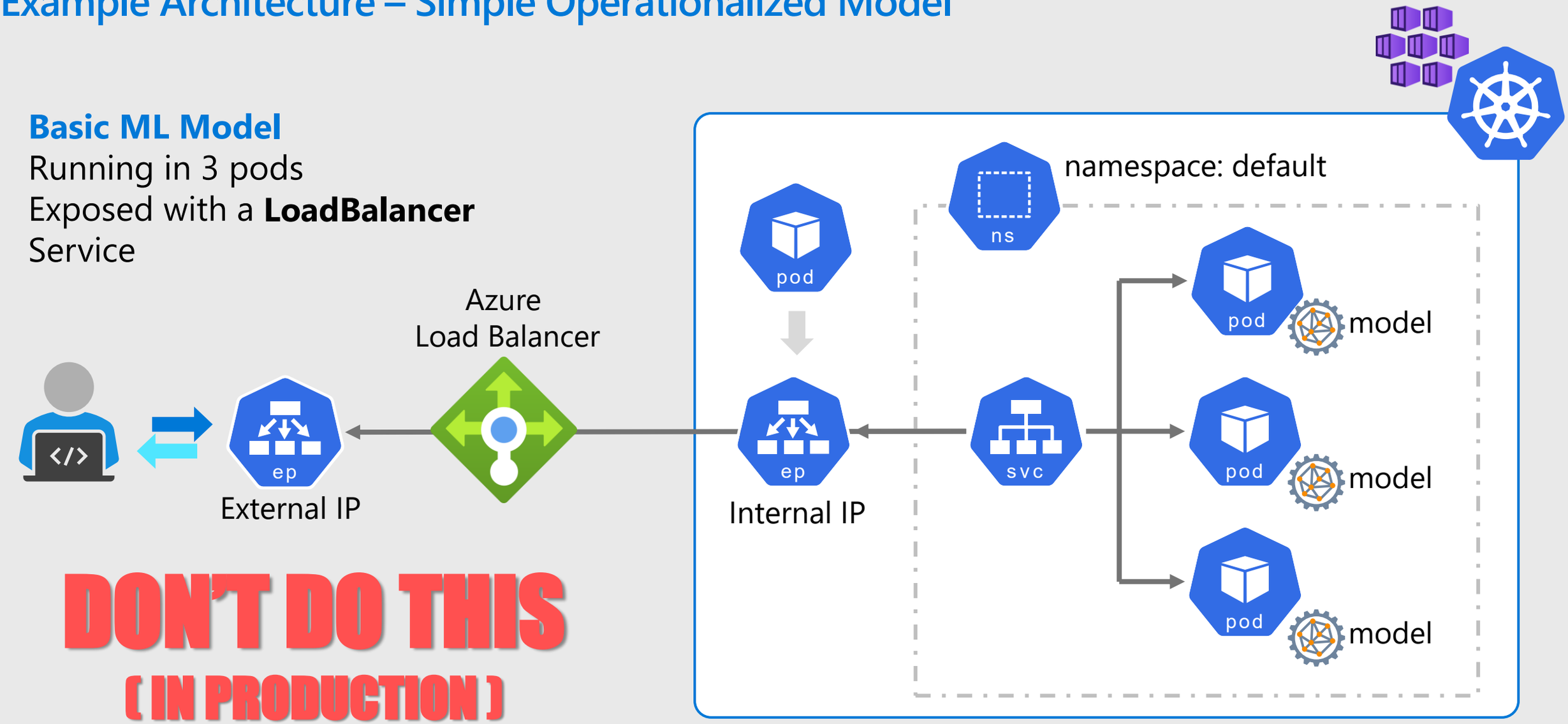
ClusterIP

Internal virtual IP, only accessible by other pods/services

Letting People Call Your Model in Kubernetes

Example Architecture – Simple Operationalized Model

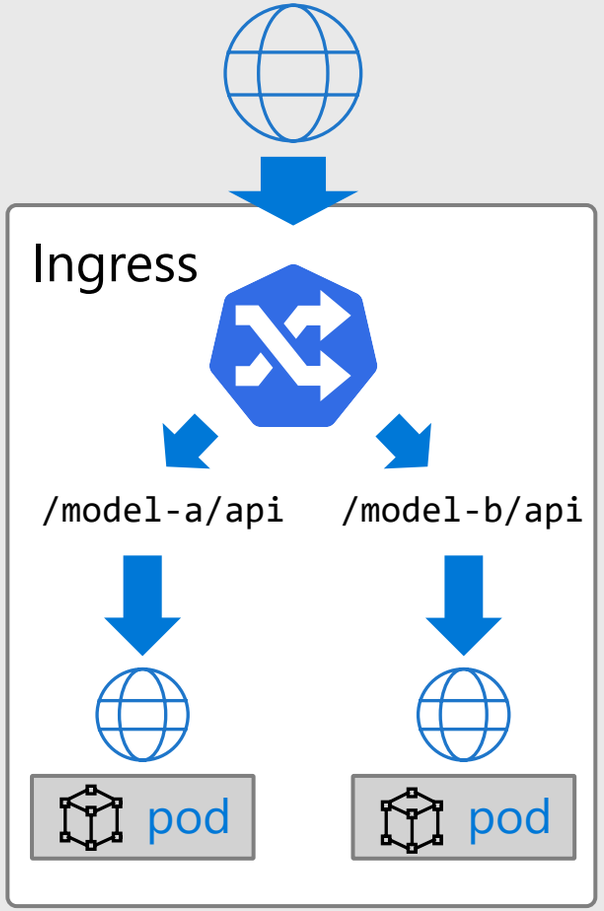
Basic ML Model
Running in 3 pods
Exposed with a **LoadBalancer** Service



DON'T DO THIS
(IN PRODUCTION)

Kubernertes Ingress – Reverse Proxy

External access for HTTP and HTTPS traffic



An *Ingress* allows you to **route** HTTP/HTTPS traffic to **services** based on URL and/or domain host name

Ingress Controller has a **public IP** and *LoadBalancer* service, it routes traffic to internal *ClusterIP* services

Various implementations of controllers are available

Use an Ingress when you want to route HTTP(S) traffic into your workloads and pods

NGINX

traefik

Kong

CONTOUR

HAPROXY

... many others ...

Ingress Shopping List

- ✓ SSL/TLS Termination
- ✓ Rate Limiting
- ✓ JWT Validation
- ✓ WAF

Traffic Routing
Distribution
Load Balancing
Observability
Service Discovery

Cert Manager

Optional Addon – Automate issuing of TLS certificates

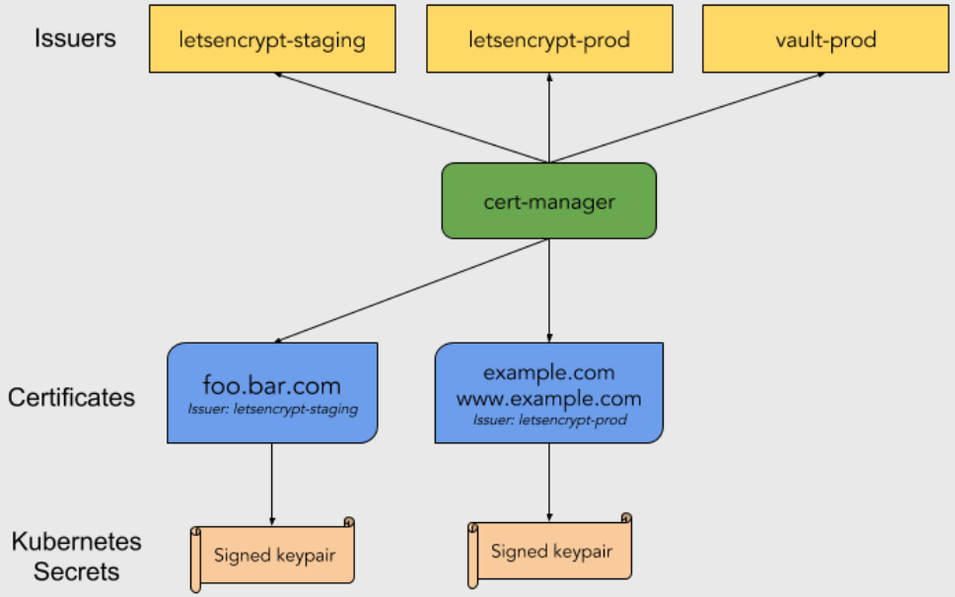
Ensures **certificates are valid** and up to date

Tightly coupled to *Ingress*, e.g. host rules

Renew certificates before expiry

Uses ACME issuers, i.e. **Let's Encrypt**

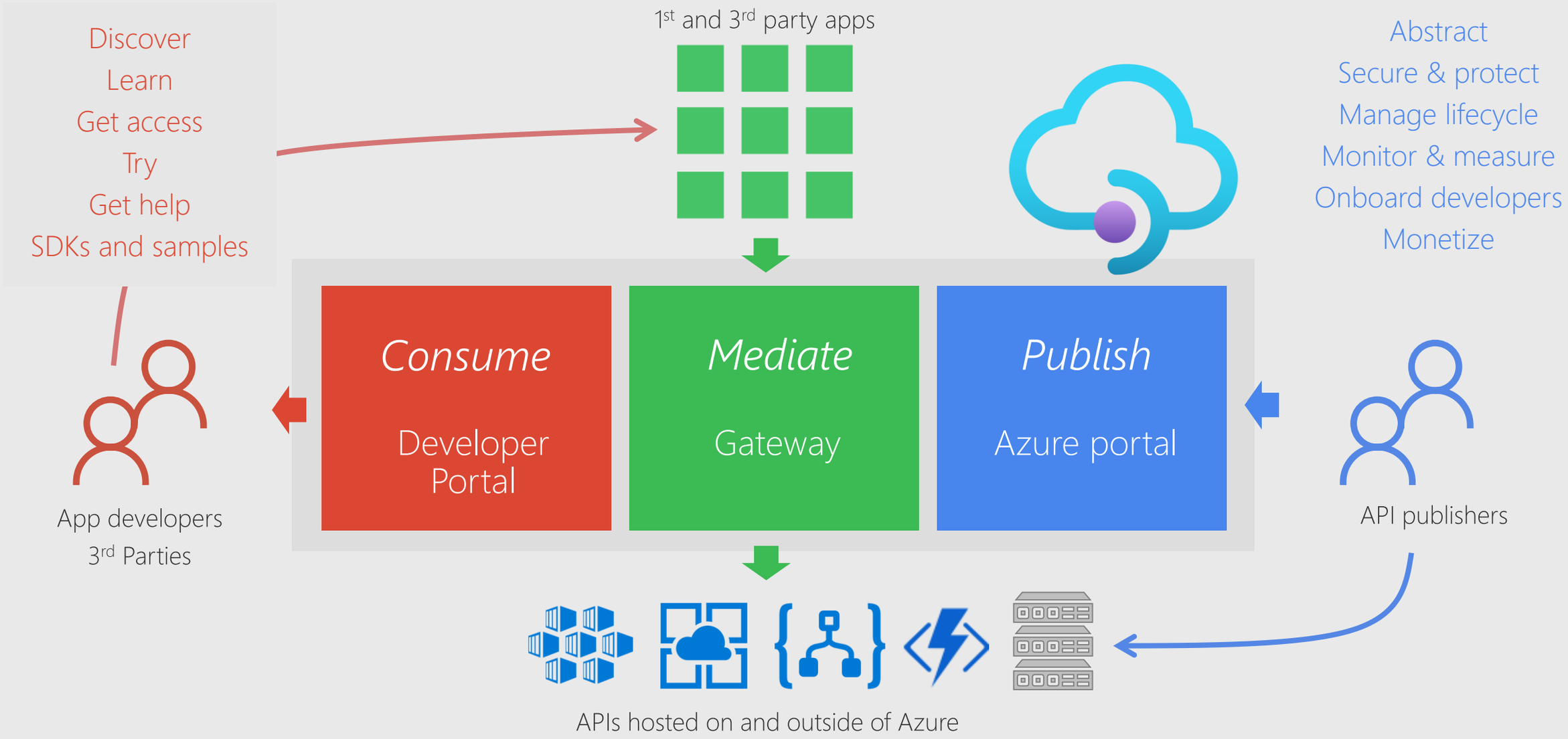
github.com/jetstack/cert-manager



Issue TLS certs for HTTPS access to services & Ingress

Other Azure Services

API Management



API Management - Data Plane Security

CONSUMERS
CLIENTS

API Keys

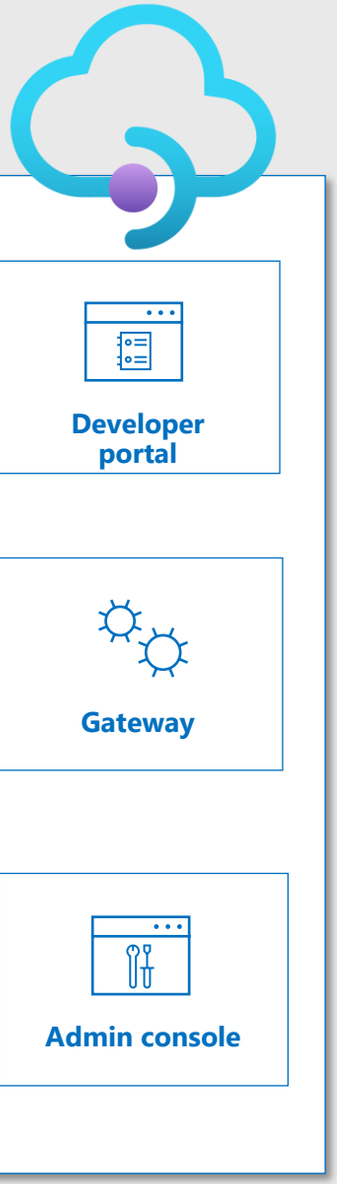
OAuth 2 & OpenID Connect

Client certificate

IP filter

External authorizer (custom)

Rate limits and quotas



API Keys

Mutual certificate (mTLS)

OAuth 2 & OpenID Connect

HTTP Basic

IP filter

Network security (VNET)

YOUR APIS
YOUR MODELS

Keys vs. OAuth 2.0 (JWT – JSON Web Token)

	KEY	OAuth JWT
Credential type	Bearer	Bearer
Granularity	All or nothing	Fine grained control
Sensitivity	Is a secret	Doesn't contain secrets
Validation	Known	Signature
Expiration	External, ad hoc	Built-in, pre-defined
Subject	Developer or an app	End user or an app

Azure Front Door

Scalable and secure entry point for fast delivery of your global applications

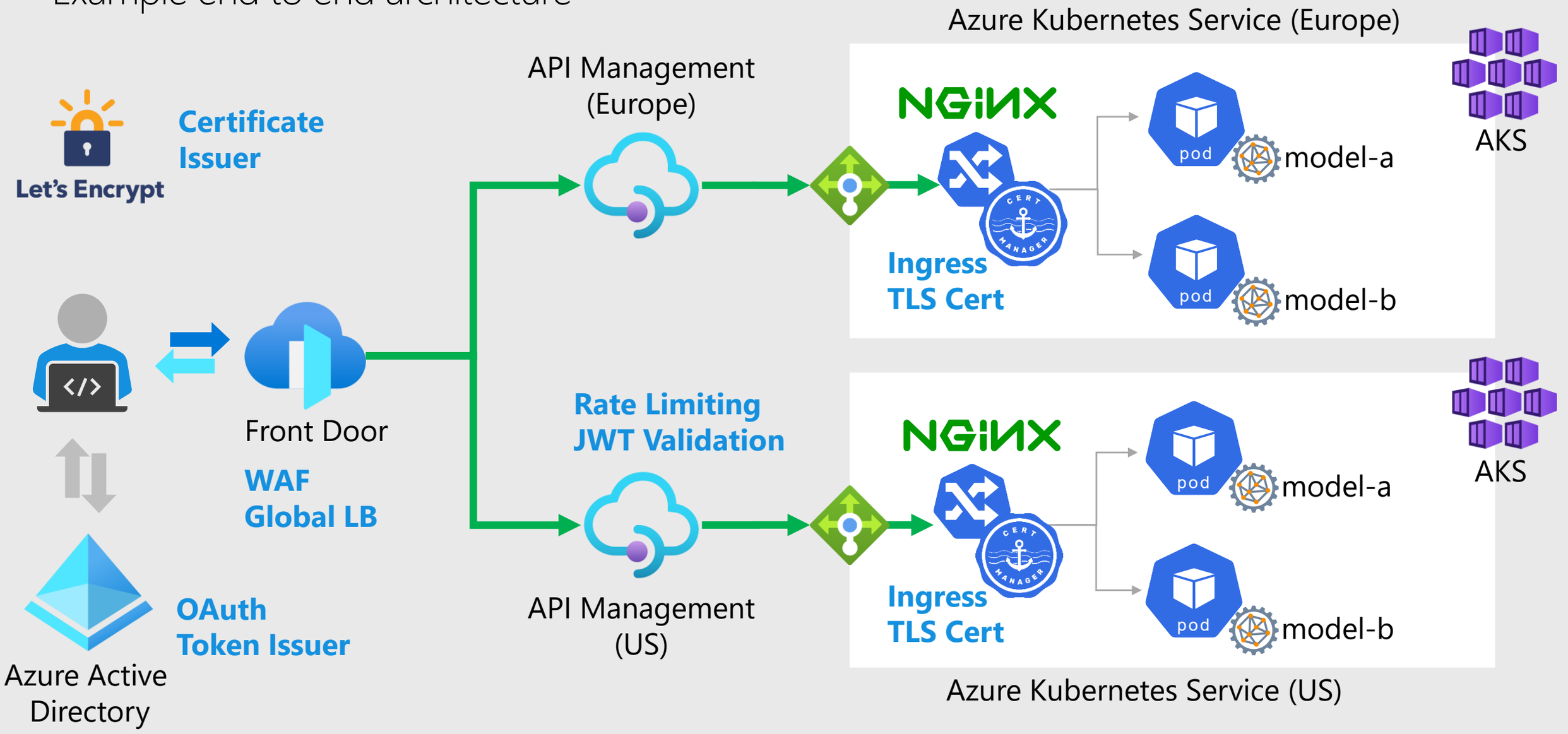
- Accelerate application performance
- Smart health probes
- URL-based routing
- TLS termination
- URL redirection
- URL rewrite
- Rules engine
- WAF (Web Application Firewall)

Protect your API from attacks

Stop network and application layer attacks at the edge with Web Application Firewall and Azure DDoS Protection. Harden your service using Microsoft managed rule sets and author your own rules for custom protection of your app.

Putting It All Together – Global Deployment

Example end to end architecture



Summary

- Don’t expose your APIs without considering security
- The range of options is wide, but something is better than nothing :)
- Don’t re-invent the wheel, let Azure Services do the heavy lifting
- Securing Kubernetes is complex, but there’s tools to help

Q&A

@BenCodeGeek

Microsoft Azure