

OpenAI



Microsoft Azure

Engineering A Cloud

Formative Assesment 1

Name	PRN
Yash D. Sonalekar	124B1F006
Vipul S. Pawar	124B1F010
Nishant Gawande	124B1F053

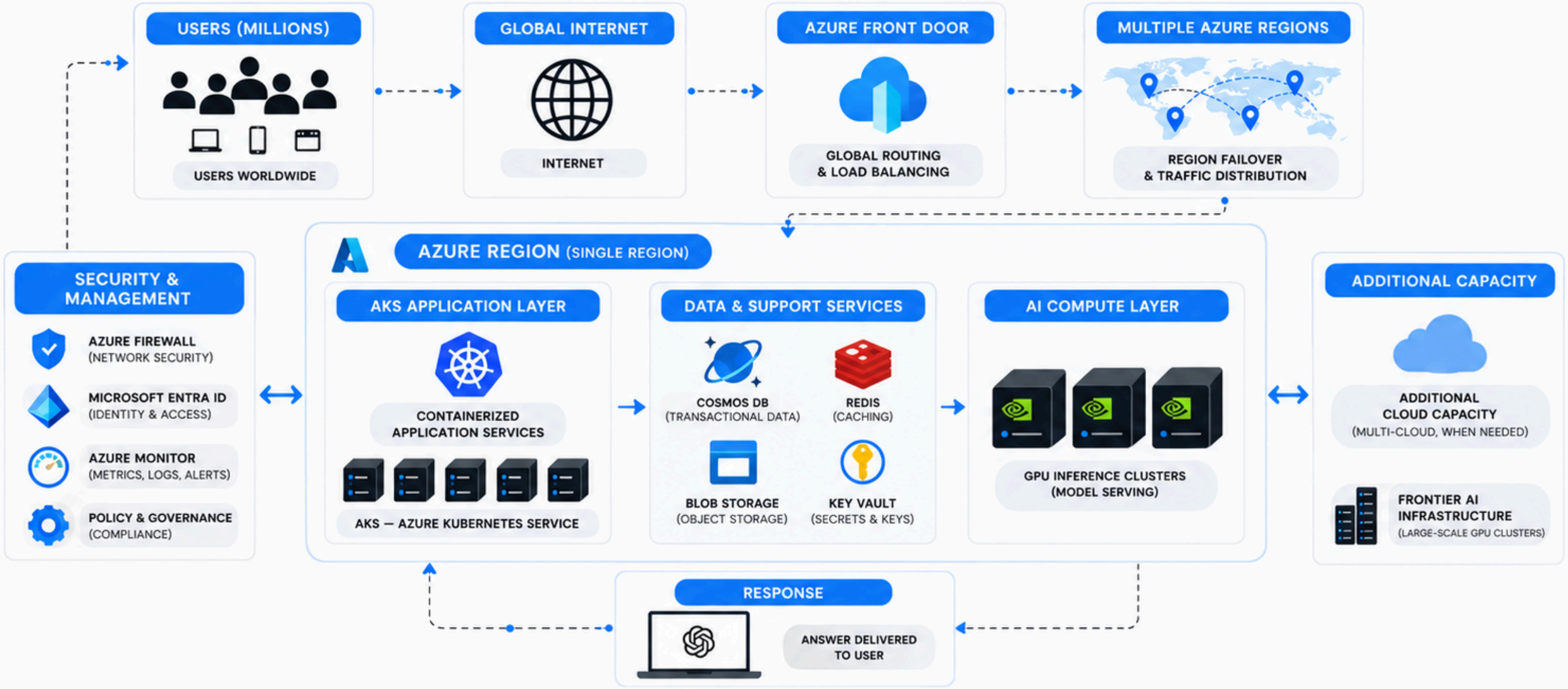




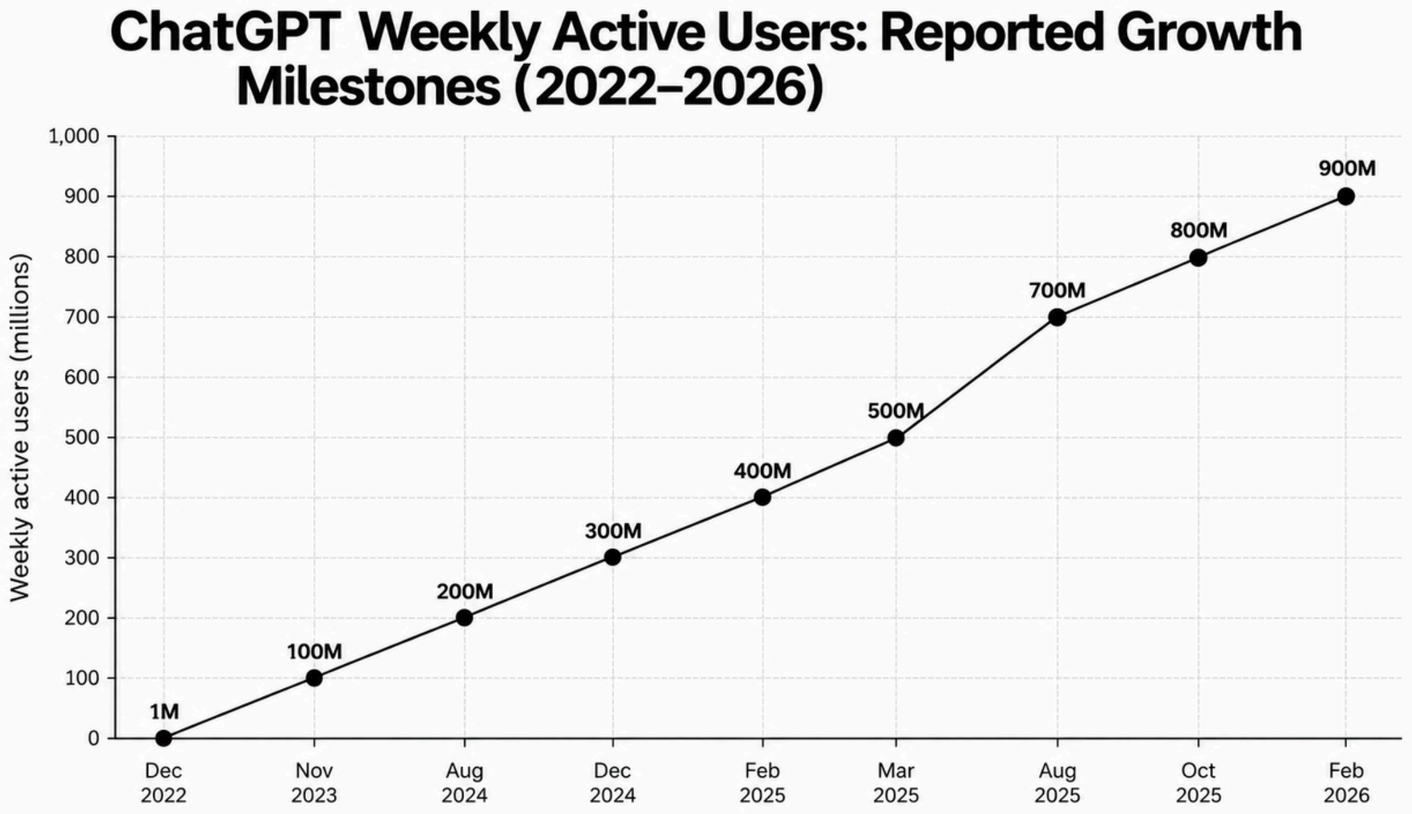
How ChatGPT Scales to Millions of Users Using Cloud Infrastructure?

OpenAI

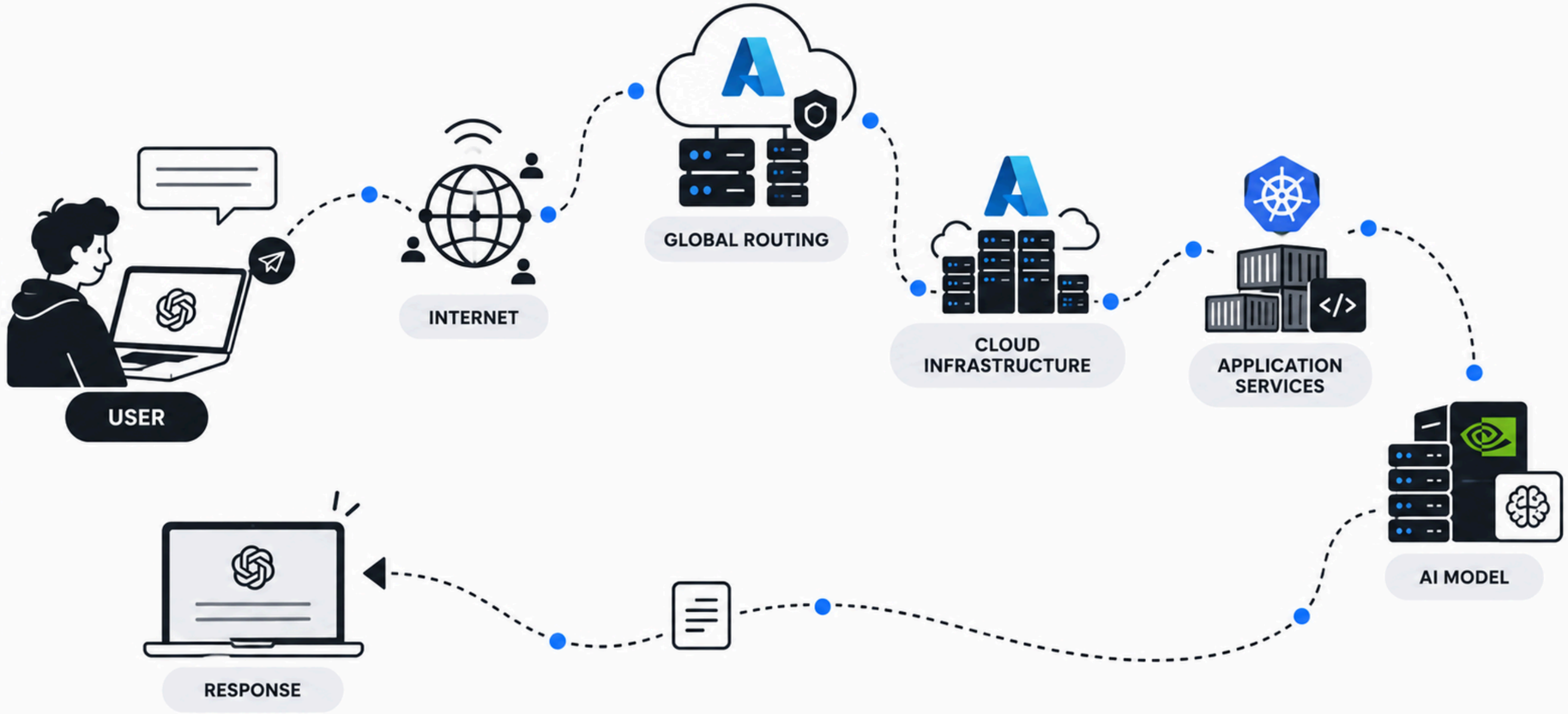
Microsoft Azure


USERS (MILLIONS)
USERS WORLDWIDE
GLOBAL INTERNET
INTERNET
AZURE FRONT DOOR
GLOBAL ROUTING
& LOAD BALANCING
MULTIPLE AZURE REGIONS
REGION FAILOVER
& TRAFFIC DISTRIBUTION
SECURITY &
MANAGEMENT
AZURE FIREWALL
(NETWORK SECURITY)
MICROSOFT ENTRA ID
(IDENTITY & ACCESS)
AZURE MONITOR
(METRICS, LOGS, ALERTS)
POLICY & GOVERNANCE
(COMPLIANCE)
AZURE REGION (SINGLE REGION)
AKS APPLICATION LAYER
CONTAINERIZED
APPLICATION SERVICES
AKS — AZURE KUBERNETES SERVICE
DATA & SUPPORT SERVICES
COSMOS DB
(TRANSACTIONAL DATA)
REDIS
(CACHING)
BLOB STORAGE
(OBJECT STORAGE)
KEY VAULT
(SECRETS & KEYS)
AI COMPUTE LAYER
GPU INFERENCE CLUSTERS
(MODEL SERVING)
ADDITIONAL CAPACITY
ADDITIONAL
CLOUD CAPACITY
(MULTI-CLOUD, WHEN NEEDED)
FRONTIER AI
INFRASTRUCTURE
(LARGE-SCALE GPU CLUSTERS)
RESPONSE
ANSWER DELIVERED
TO USER

ChatGPT reaches over 900 million weekly active users and processes roughly 2.5 billion prompts globally every day



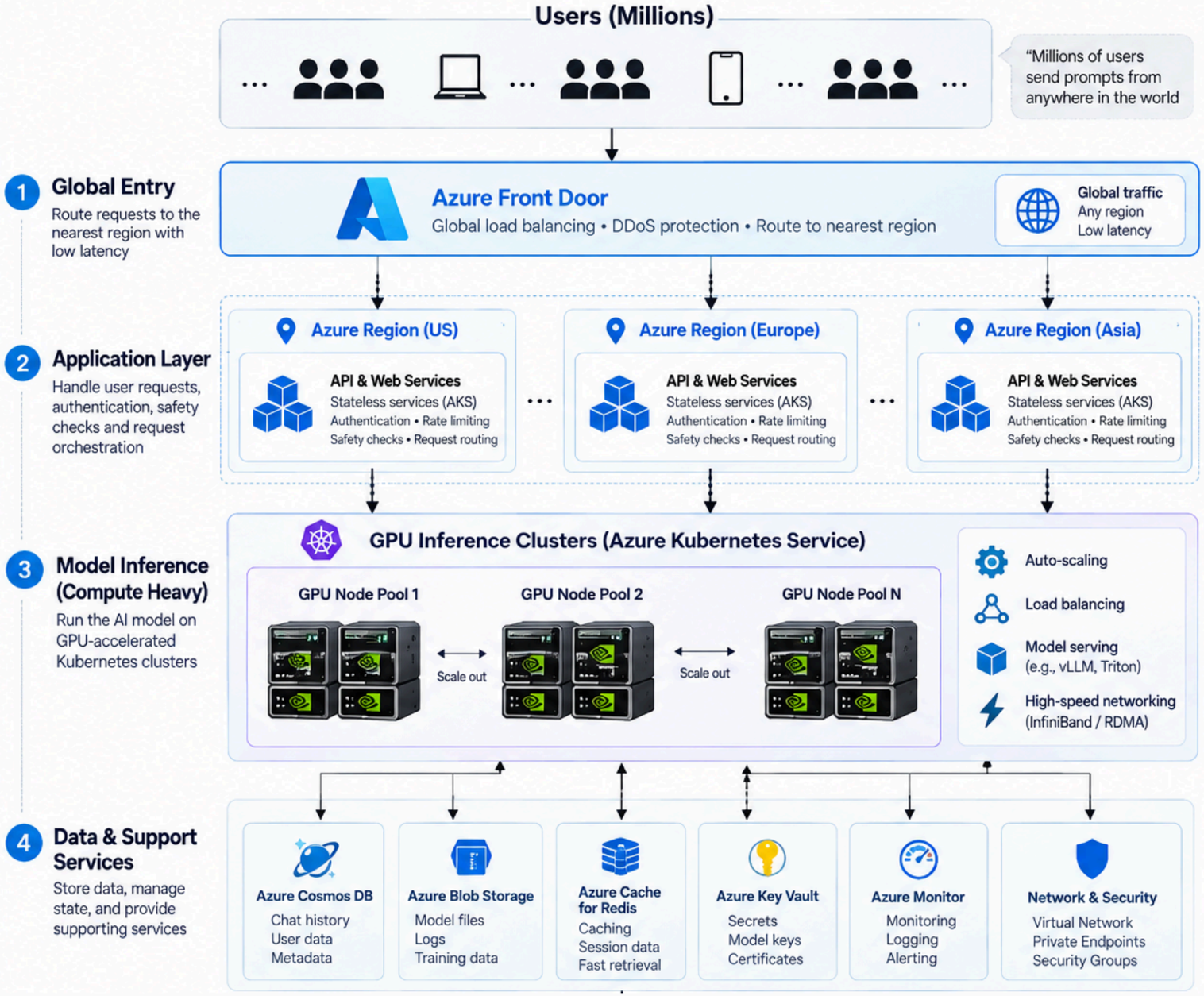
So, what actually happens behind a single request?

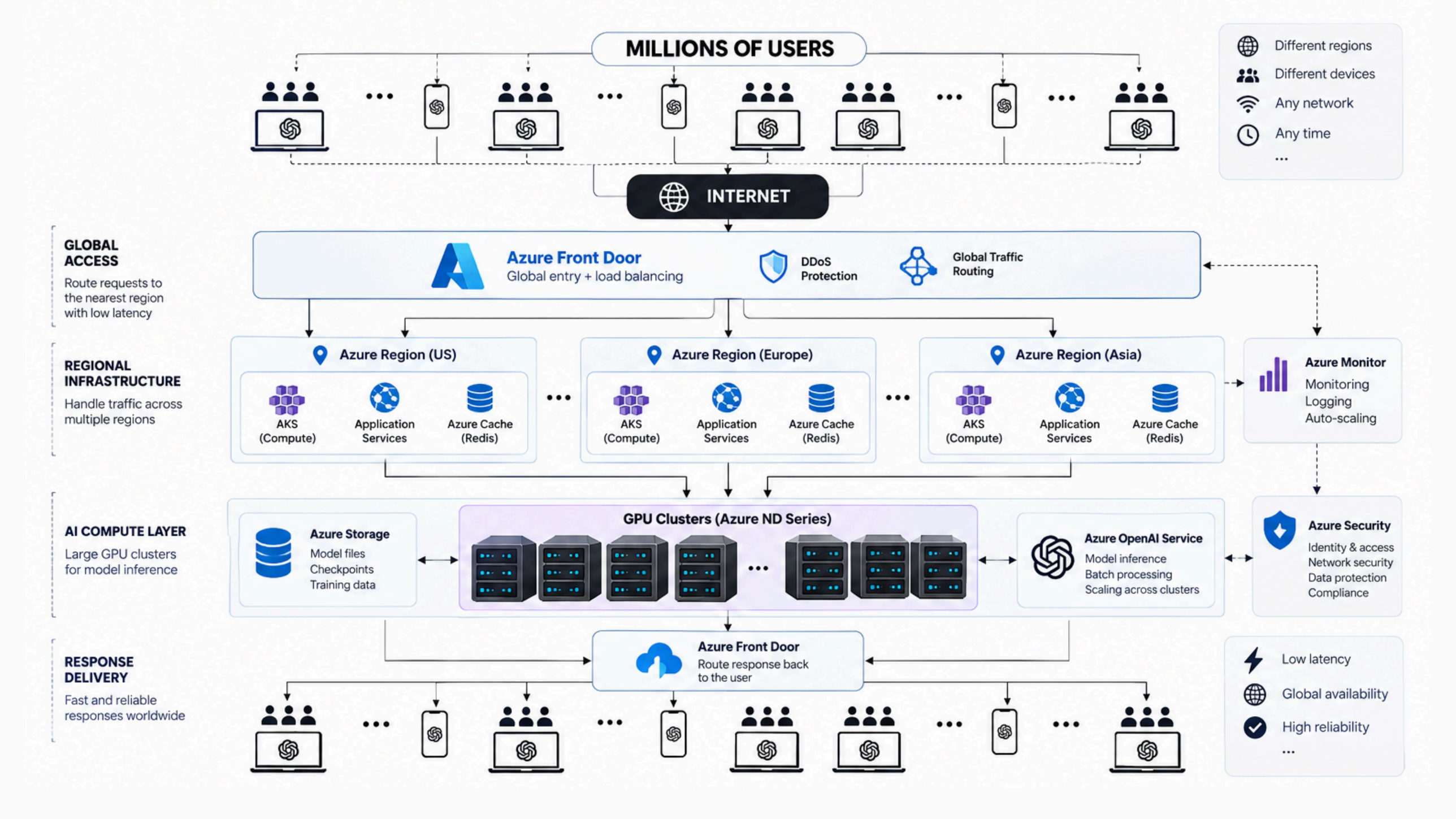


Before an answer appears on your screen, your message passes through multiple layers of networking, cloud infrastructure, application services and AI compute.

What Happens when millions of requests arrive together ?

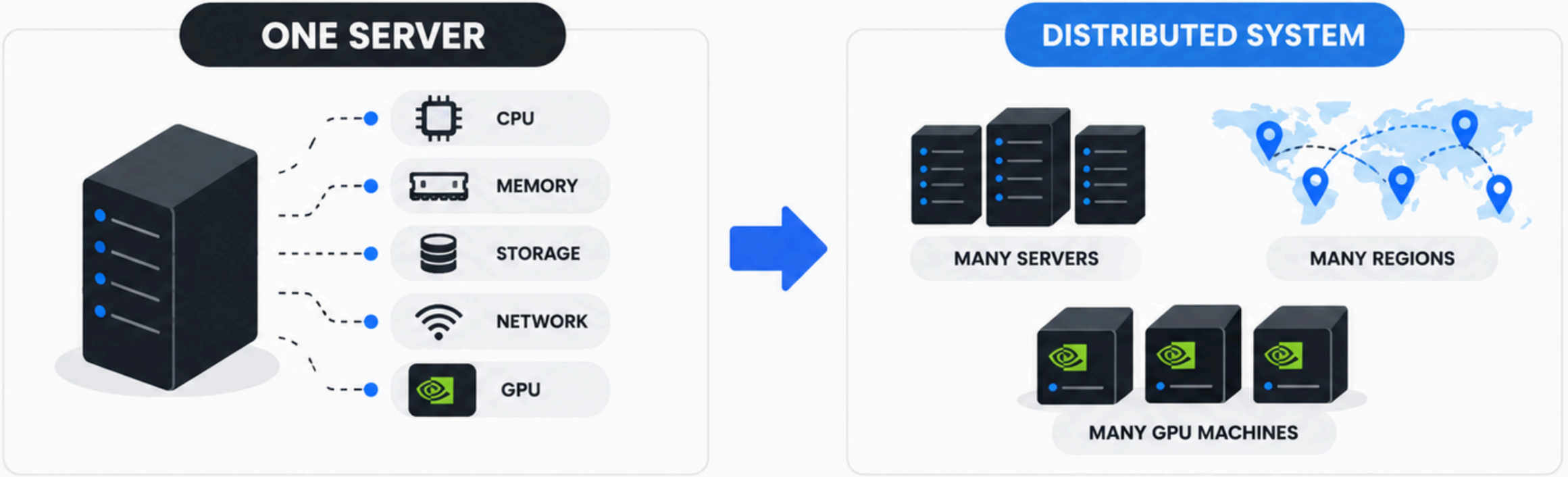
AI inference is compute intensive, so the system needs many different layers of infrastructure working together to handle the load.




MILLIONS OF USERS
Different regions
Different devices
Any network
Any time
...
INTERNET
GLOBAL ACCESS
Route requests to the nearest region with low latency
Azure Front Door
Global entry + load balancing
DDoS Protection
Global Traffic Routing
REGIONAL INFRASTRUCTURE
Handle traffic across multiple regions
Azure Region (US)
AKS (Compute)
Application Services
Azure Cache (Redis)
Azure Region (Europe)
AKS (Compute)
Application Services
Azure Cache (Redis)
Azure Region (Asia)
AKS (Compute)
Application Services
Azure Cache (Redis)
Azure Monitor
Monitoring
Logging
Auto-scaling
AI COMPUTE LAYER
Large GPU clusters for model inference
Azure Storage
Model files
Checkpoints
Training data
GPU Clusters (Azure ND Series)
Azure OpenAI Service
Model inference
Batch processing
Scaling across clusters
Azure Security
Identity & access
Network security
Data protection
Compliance
RESPONSE DELIVERY
Fast and reliable responses worldwide
Azure Front Door
Route response back to the user
Low latency
Global availability
High reliability
...

Why can't one server handle all of this?

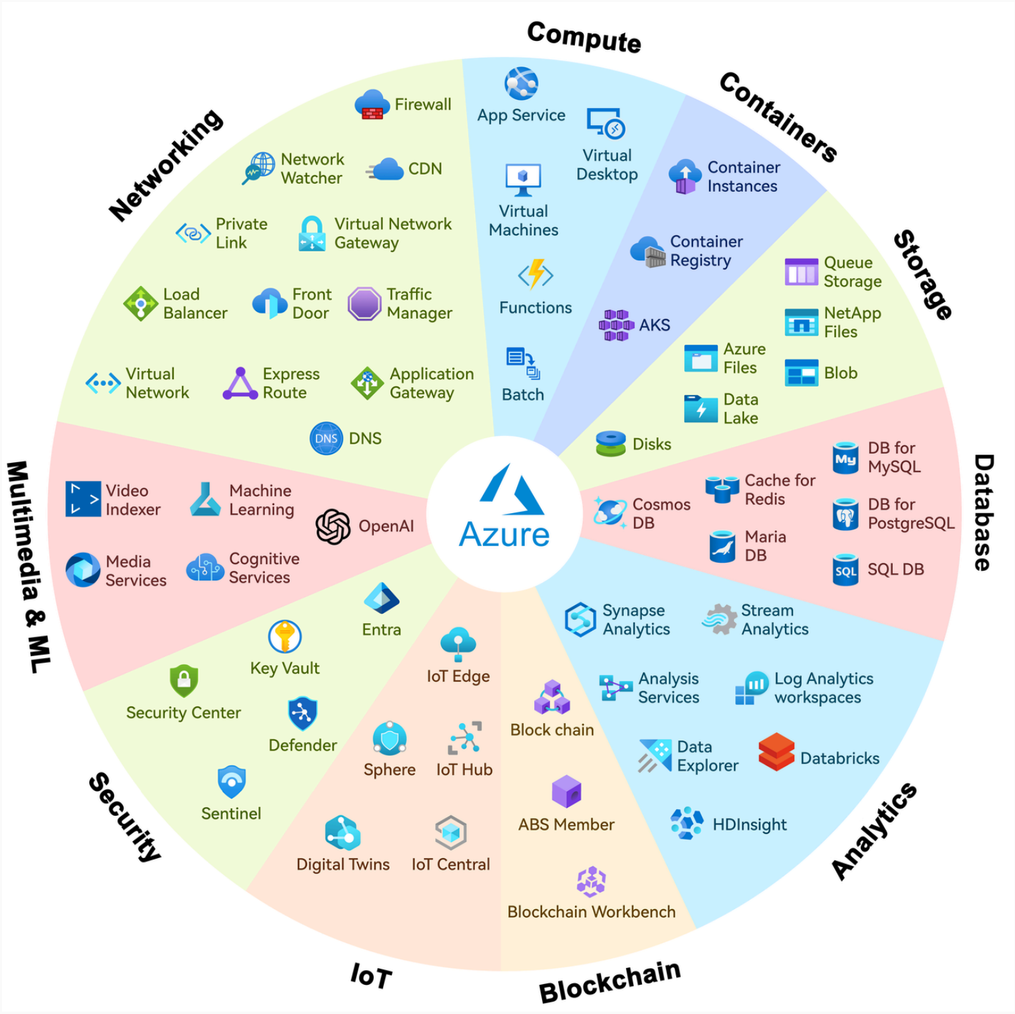
A single machine has limited CPU, memory, storage and network capacity, so the workload has to be distributed across many machines.



Scale comes from distributing the workload, not from endlessly growing one machine.

The System Doesn’t scale as a single giant server.

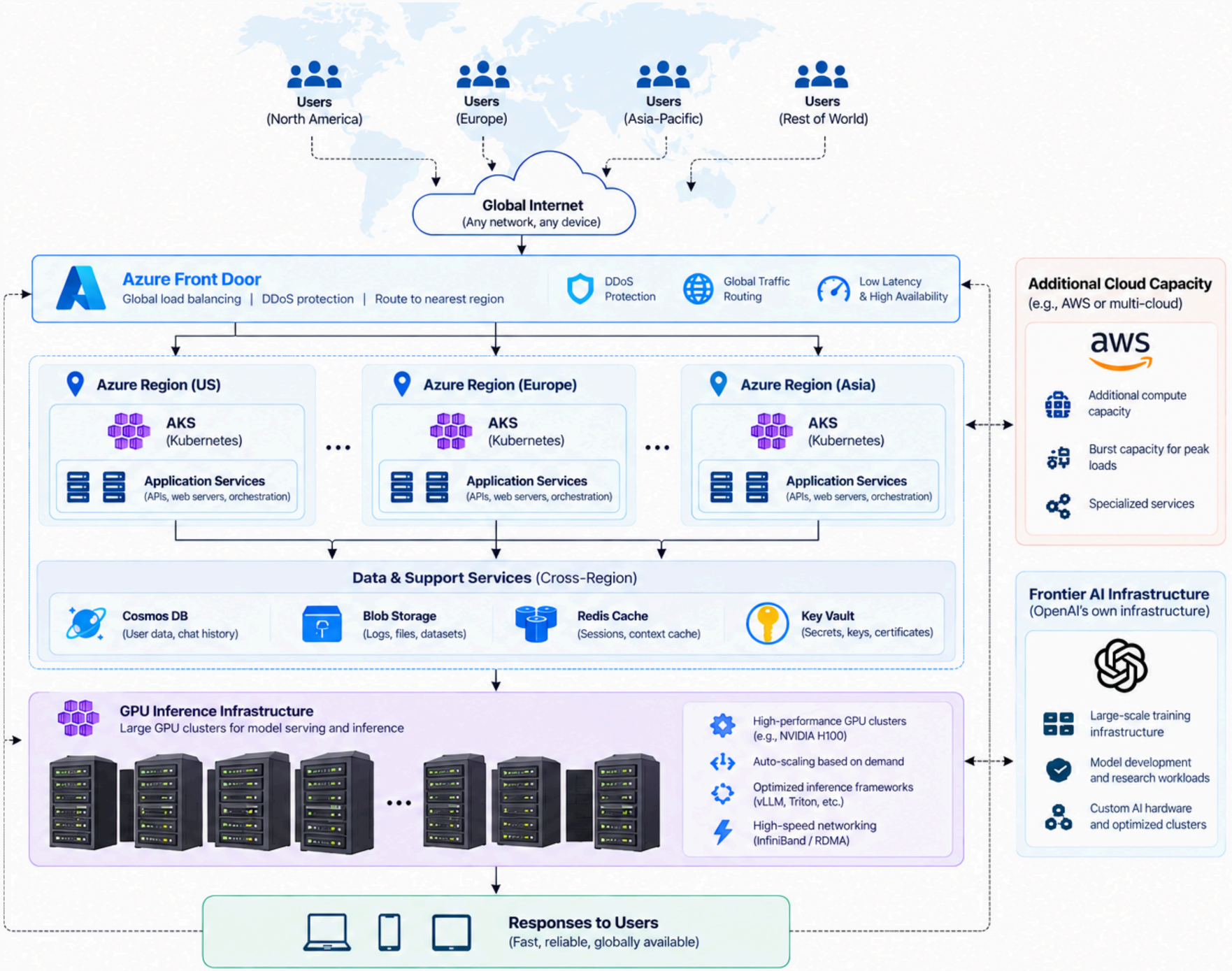
The workload is distributed across regions, application servers, cloud services and GPU clusters.



Category	Azure Service	Icon	What it does	Used for in ChatGPT-like system
Global Entry & Networking	Azure Front Door		Global load balancing and content delivery	Routes users to the nearest region, protects from DDoS, improves performance
	Azure DNS		Domain name system	Resolves chat.openai.com to the right endpoints
	Azure Virtual Network		Private networking	Secure communication between services
	Azure Load Balancer		Distributes traffic within a region	Balances traffic across application servers
Compute & Application Layer	Azure Kubernetes Service (AKS)		Container orchestration	Runs and scales application services (APIs, web servers, workers)
	Azure App Service		Host web applications	Hosts auxiliary services, dashboards, internal tools
	Azure Container Instances		Run containers without managing servers	Used for smaller or short-lived workloads
	Azure Functions		Serverless compute	Event-driven tasks (e.g., webhooks, background jobs)
AI Compute & Model Inference	Azure OpenAI Service		Access to OpenAI models (e.g., GPT-4)	Model inference, embeddings, completions
	Azure Virtual Machines (GPU)		GPU-accelerated compute	Runs large model inference workers (e.g., NVIDIA GPUs)
	Azure Batch		Large-scale job scheduling	Batch processing, offline inference, data processing
Storage & Data	Azure Cosmos DB		Globally distributed NoSQL database	Store chat history, user data, metadata (low-latency, global)
	Azure Blob Storage		Object storage	Store files, logs, model artifacts, datasets
	Azure Cache for Redis		In-memory caching	Cache frequently accessed data (sessions, user context)
	Azure Data Lake		Analytics storage	Store large datasets for training, analytics, and logs
Security & Identity	Azure Key Vault		Secrets and key management	Store API keys, certificates, encryption keys
	Microsoft Defender		Threat protection	Detect and respond to security threats
	Azure Sentinel		Security monitoring (SIEM)	Centralized logging and threat detection
	Azure Firewall		Network security	Control inbound/outbound traffic, protect internal services
Monitoring & Analytics	Azure Monitor		Monitoring and telemetry	Collect metrics, logs, and traces from all services
	Log Analytics Workspace		Log storage and analysis	Analyze logs, troubleshoct issues, detect anomalies
	Application Insights		Application performance monitoring	Track performance, latency, and failures
	Azure Data Explorer		Fast data analytics	Analyze large volumes of telemetry and usage data
Additional / Other Services	Azure Traffic Manager		DNS-based traffic routing	Route traffic based on policies (e.g., failover, performance)
	Azure Virtual Desktop		Remote desktop infrastructure	Internal access for development and operations
	Azure DevOps / Entra ID		CI/CD and identity management	Developer workflows and access control

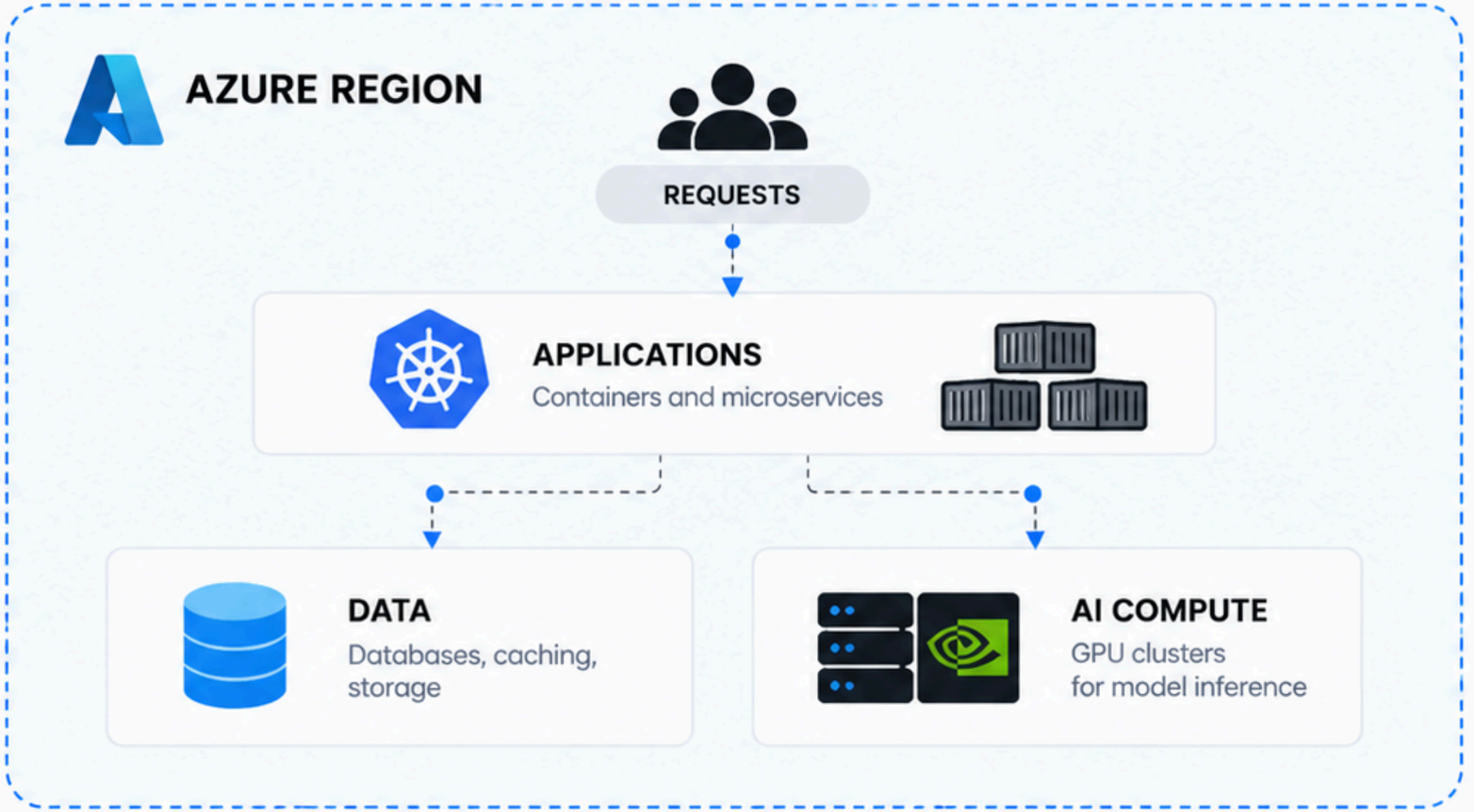
So, how do all these pieces work together ?

A large-scale AI service combines global routing, regional application services, data systems and GPU infrastructure.



What happens inside a single region?

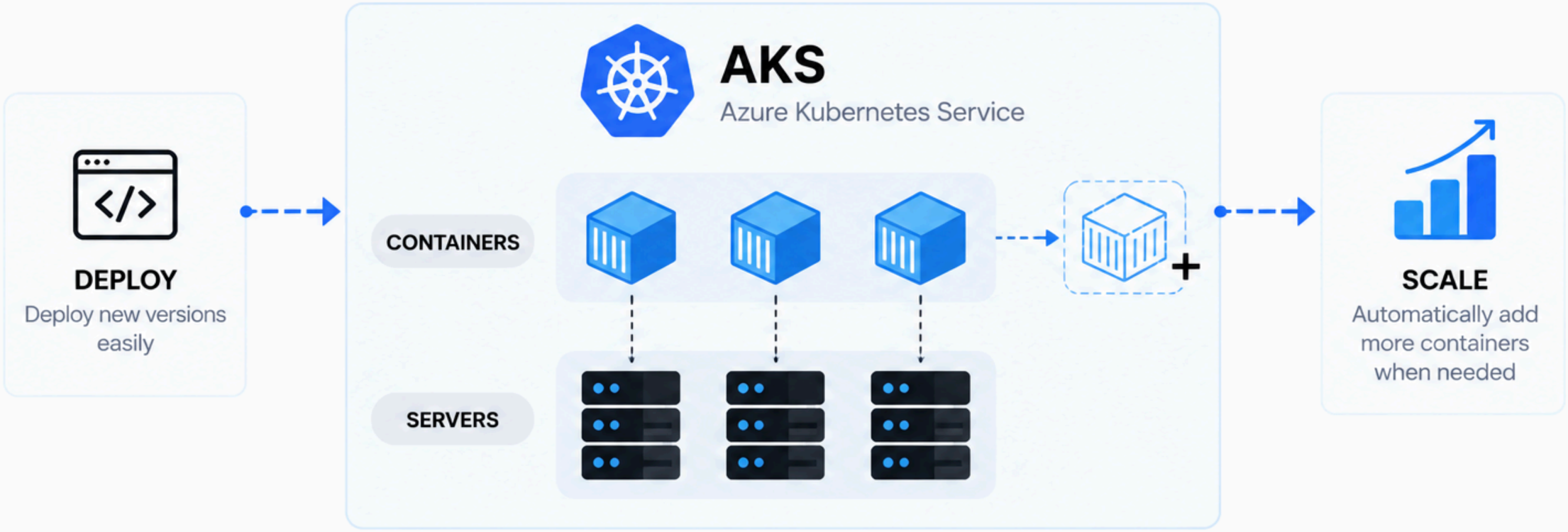
A region can handle requests, supporting data services and AI compute independently, so the system does not depend on one location.



Each region can handle its own share of the workload.

Who manages all those application servers?

Azure Kubernetes Service can run containerized applications across many machines and automatically manage their deployment and scaling.

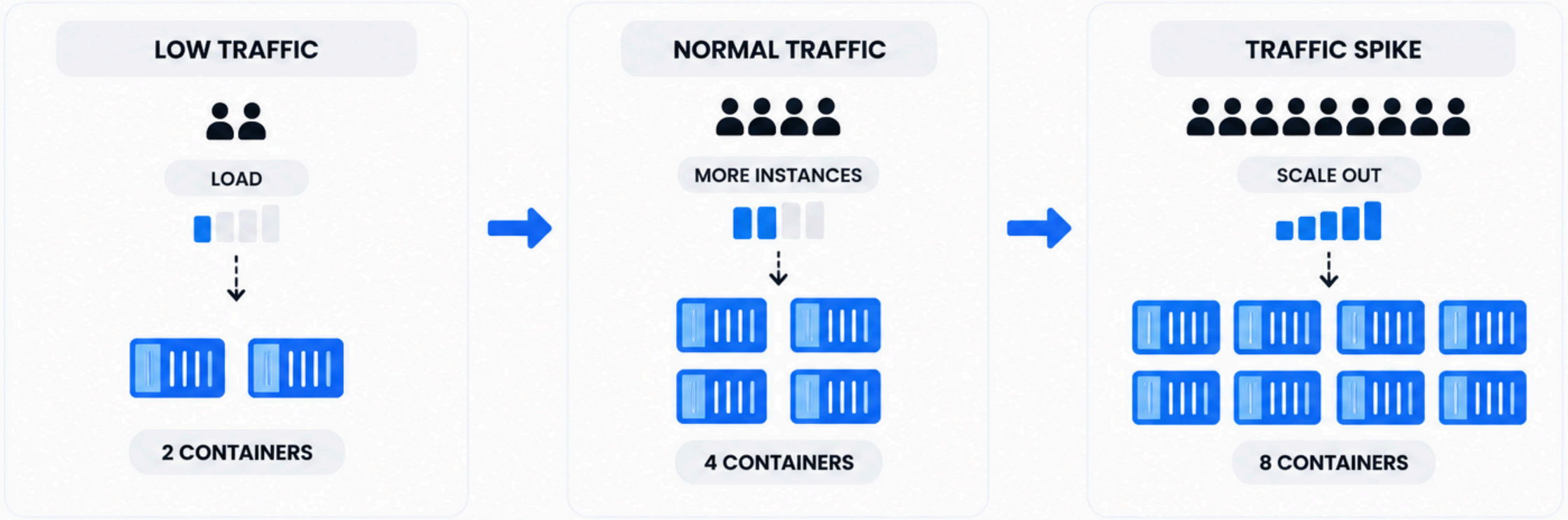


Instead of managing every server manually, Kubernetes manages the application workloads for us.

What happens when traffic suddenly increases?

Instead of making one server infinitely powerful, the system can add more application instances to share the workload.

AKS
Azure Kubernetes Service

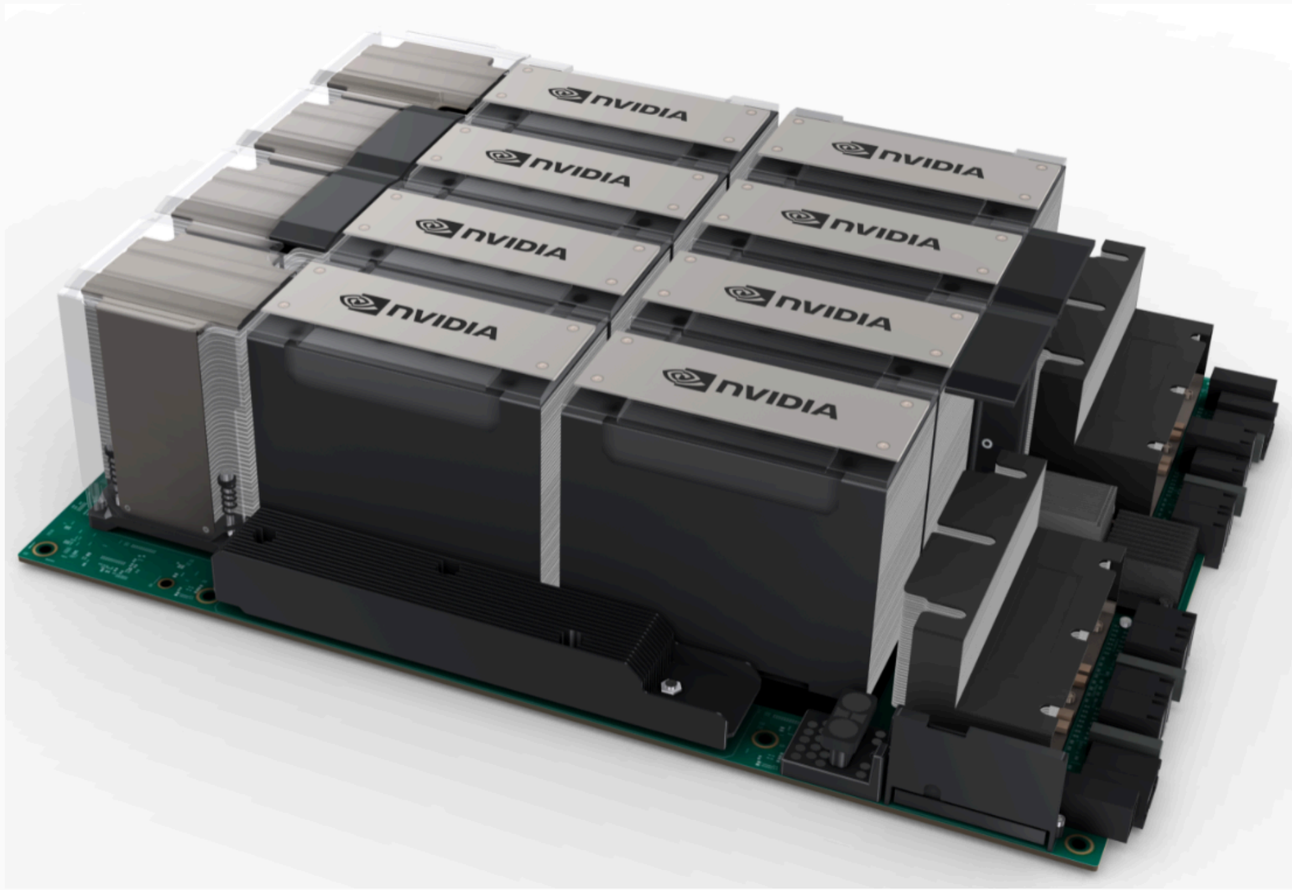


When demand grows, the system adds more capacity instead of relying on one giant server.

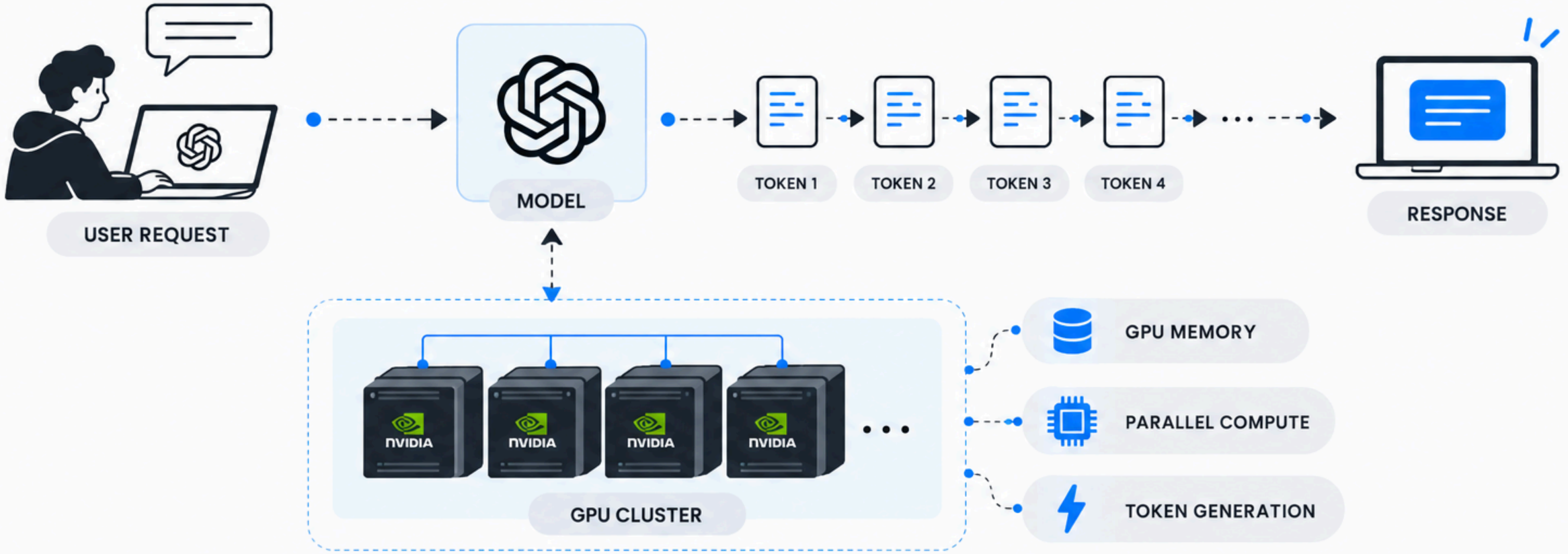
But can we scale everything the same way?

No. Web applications, databases and AI inference have very different computing requirements.

AI inference needs specialized GPU infrastructure.

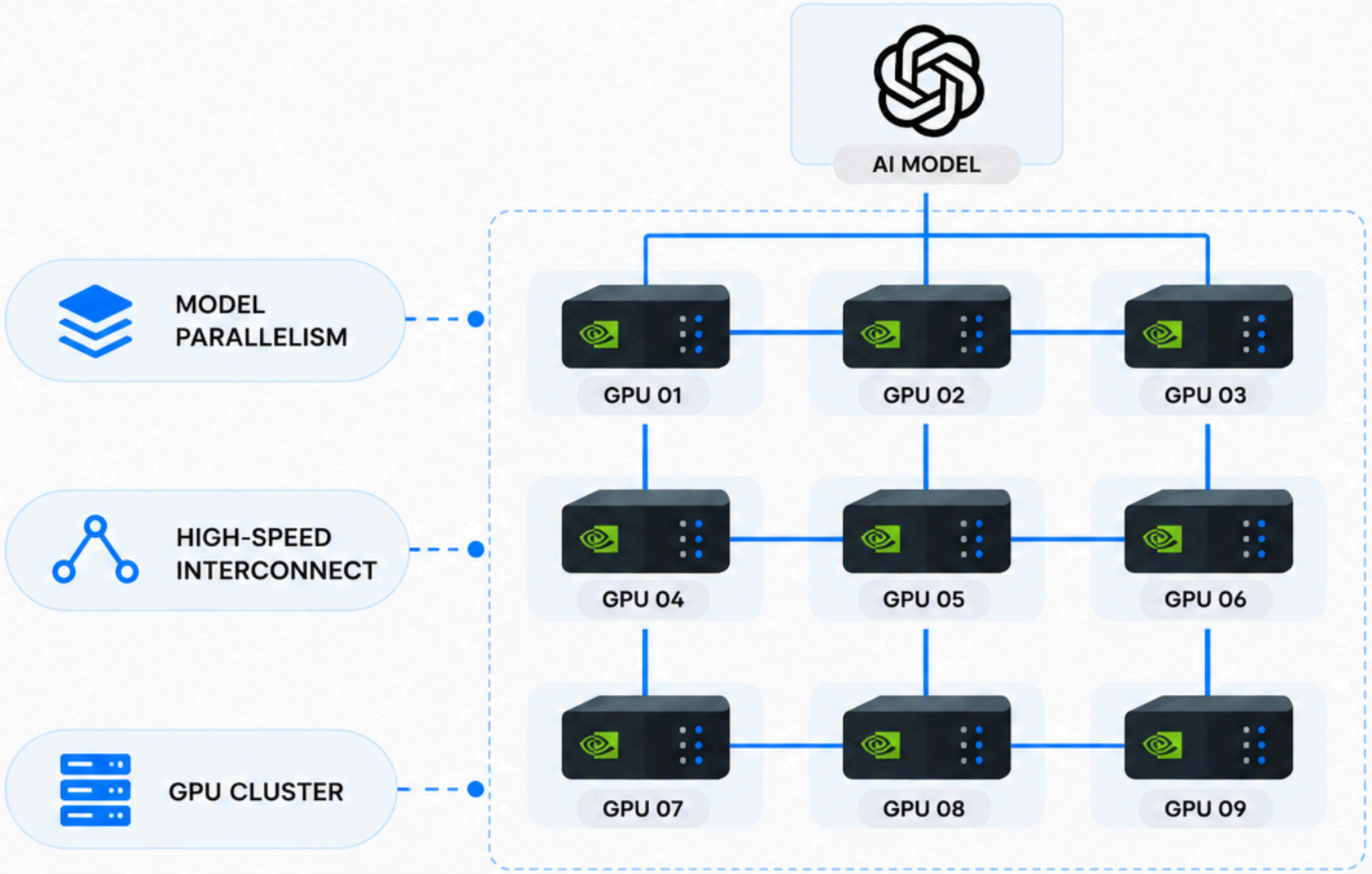


Why does one AI request need so much compute?



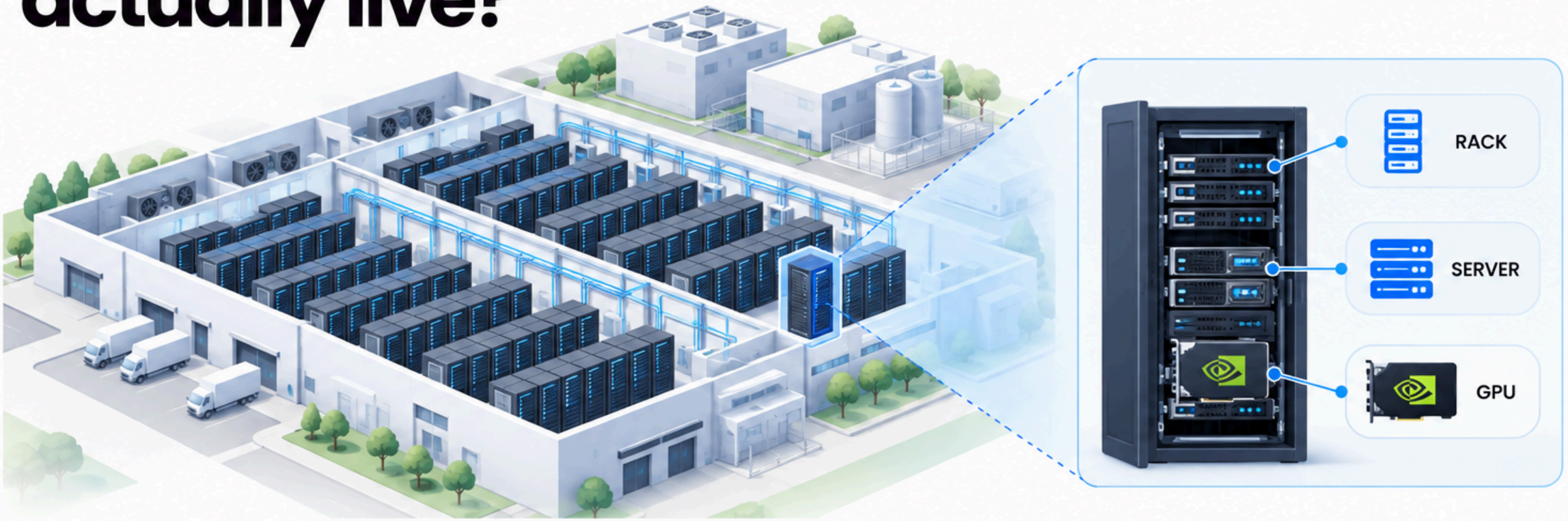
One response is not one calculation.
It is a sequence of computations repeated across the model.

What if the model is too large for one GPU?



More GPUs allow the workload to be distributed instead of forcing everything onto one machine.

Where do all these GPUs actually live?



GPU RACKS

NETWORKING

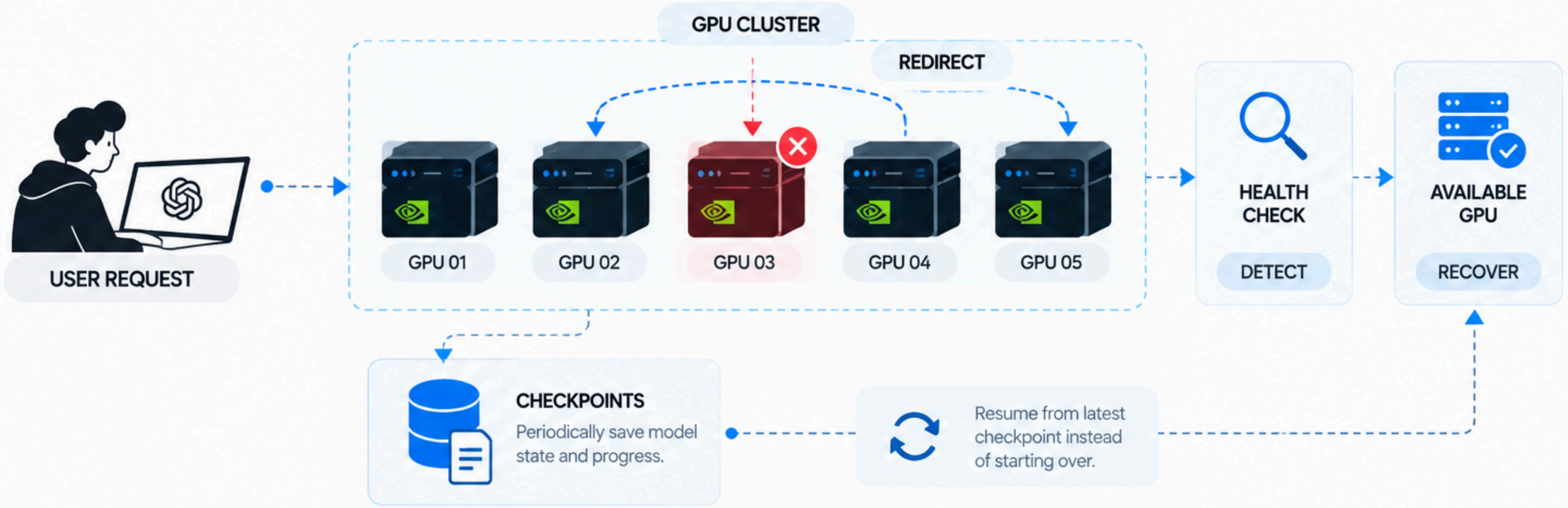
POWER

COOLING

Cloud scale ultimately depends on **physical infrastructure.**

What happens when a GPU or server fails?

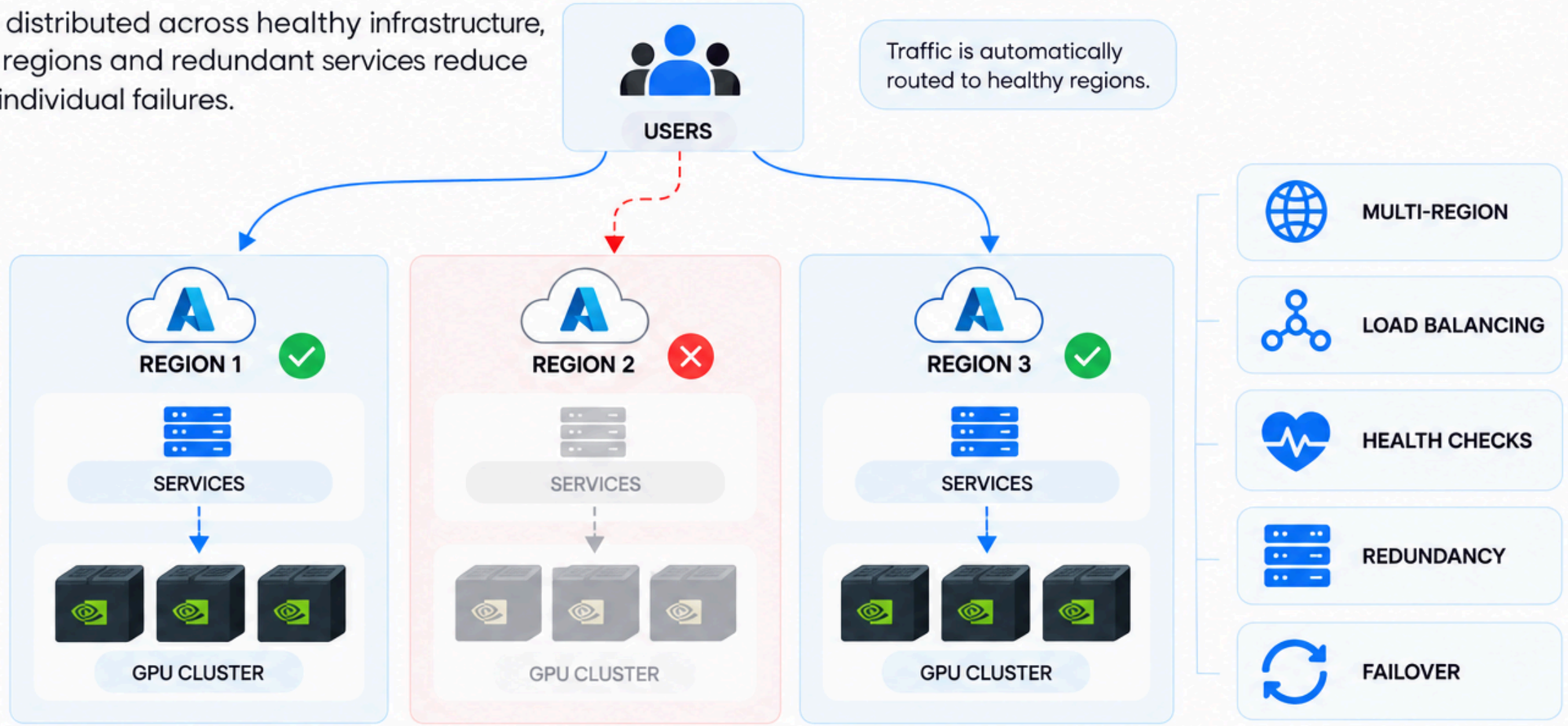
At large scale, hardware failures are expected, so the architecture has to recover without taking the entire service down.



Failure is not an exception at this scale. It has to be part of the design.

How does the system keep running when something goes wrong?

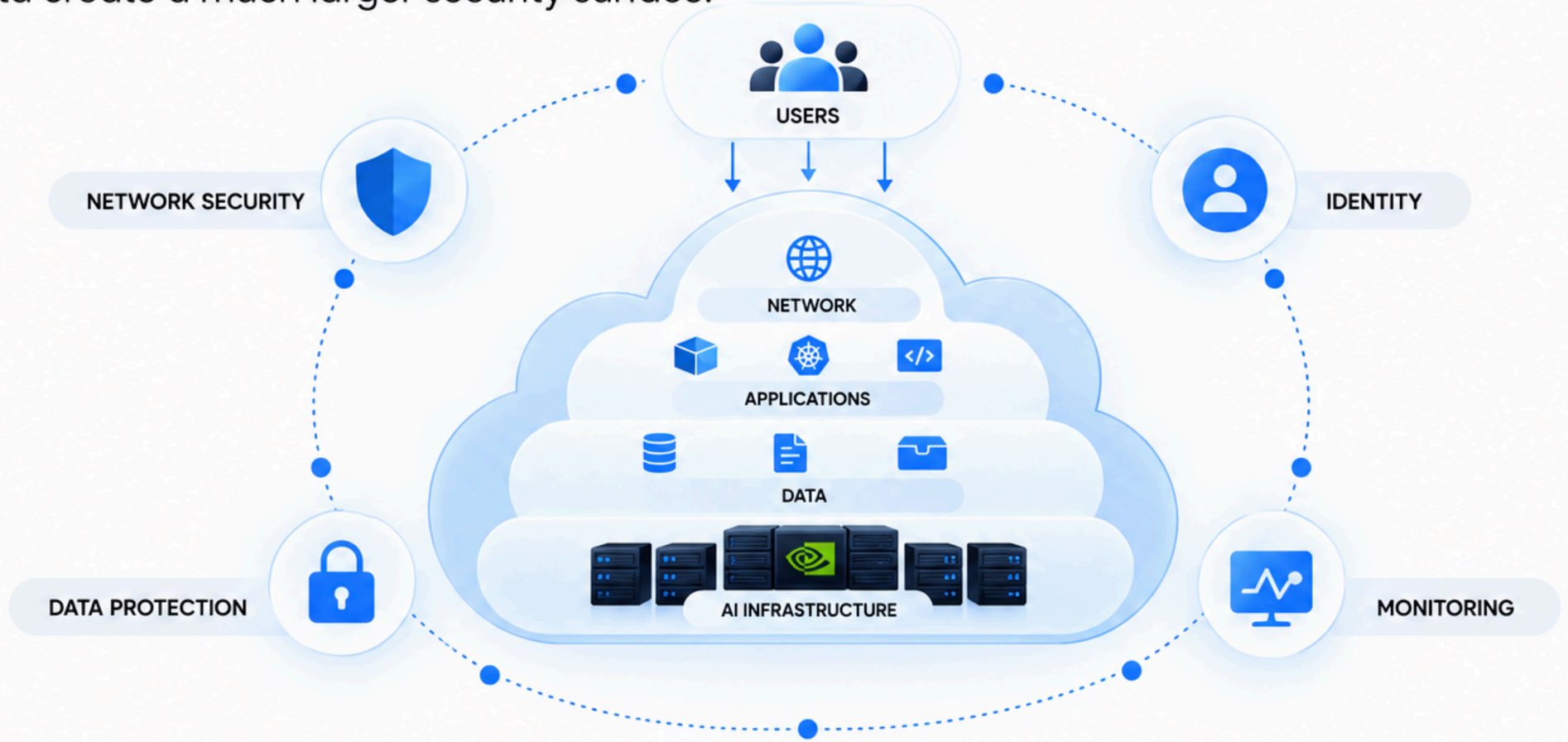
Traffic can be distributed across healthy infrastructure, while multiple regions and redundant services reduce the impact of individual failures.



The goal is not to prevent every failure. It is to prevent one failure from becoming an outage.

What happens when the system gets this big?

Millions of users, thousands of services and huge amounts of data create a much larger security surface.

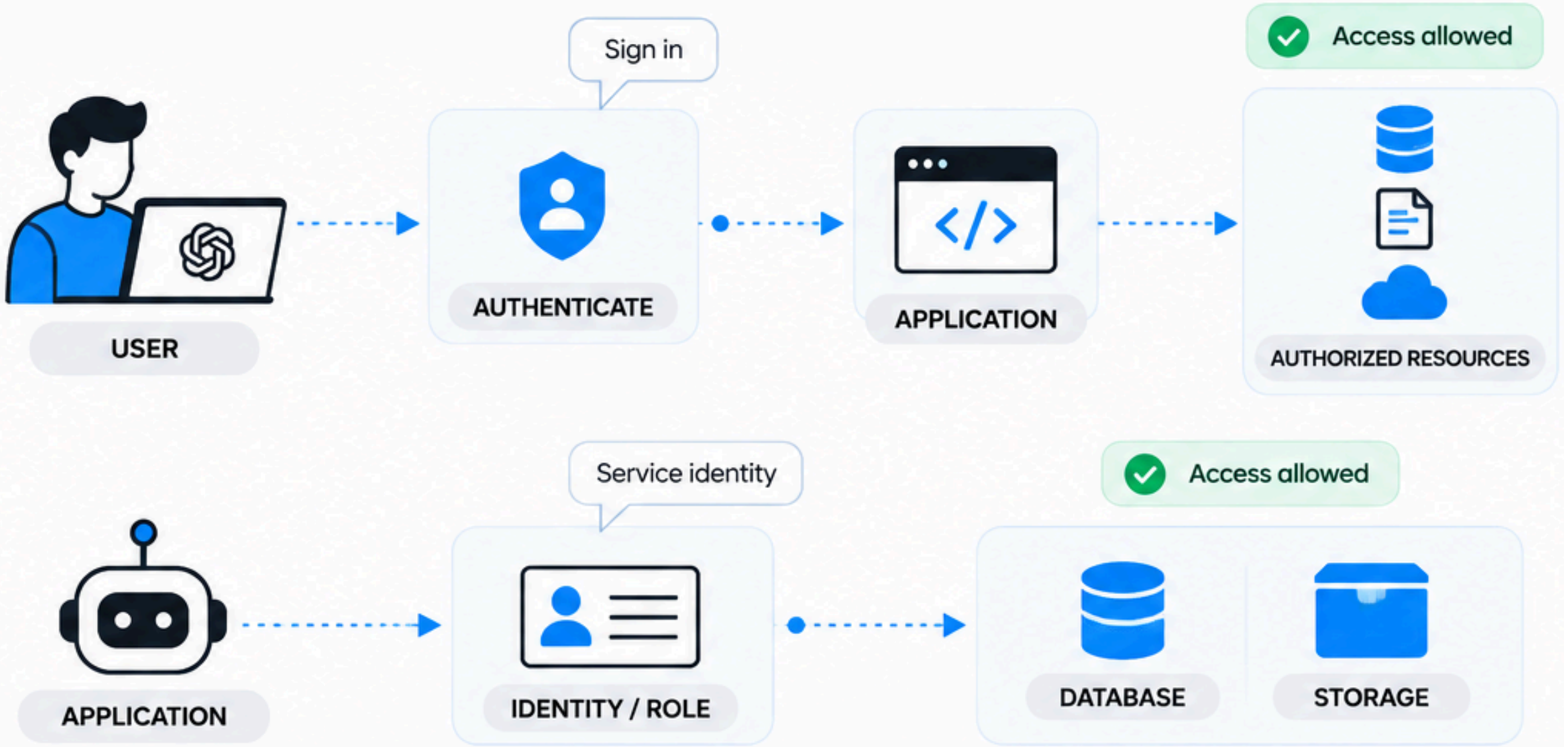


Security has to be built into every layer of the system.

How does the system know who is allowed to access what?

Identity and access controls ensure that users, applications and administrators only get the permissions they need.

Microsoft Entra ID | Azure Key Vault

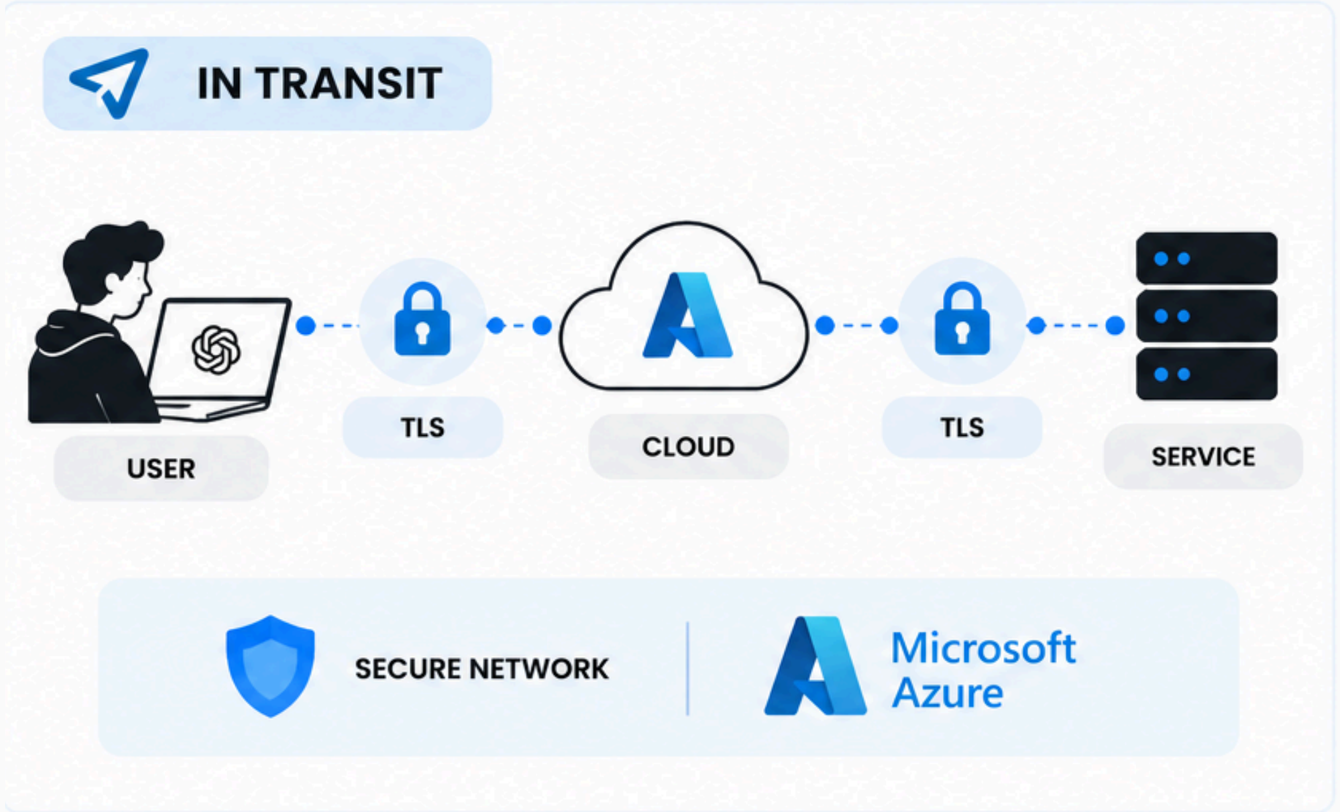


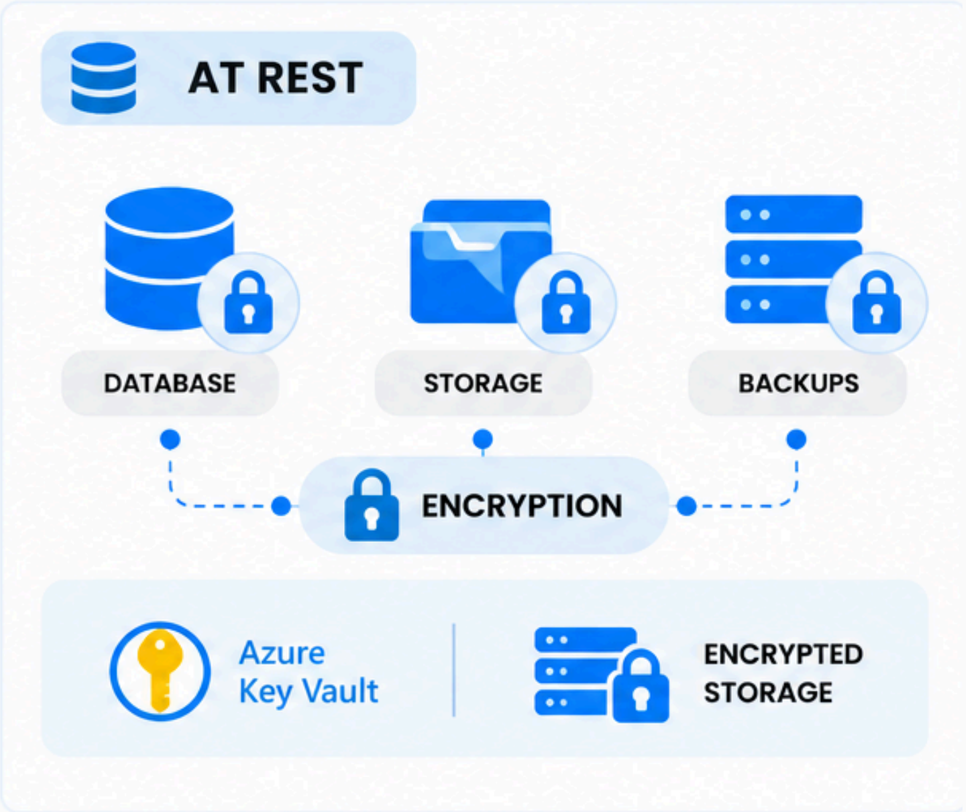
- AUTHENTICATE
- AUTHORIZE
- LEAST PRIVILEGE
- MANAGE SECRETS

Being inside the system does not automatically mean you can access everything.

What happens to the data while it is moving and while it is stored?

Sensitive data can be protected both in transit across networks and at rest inside storage and databases.

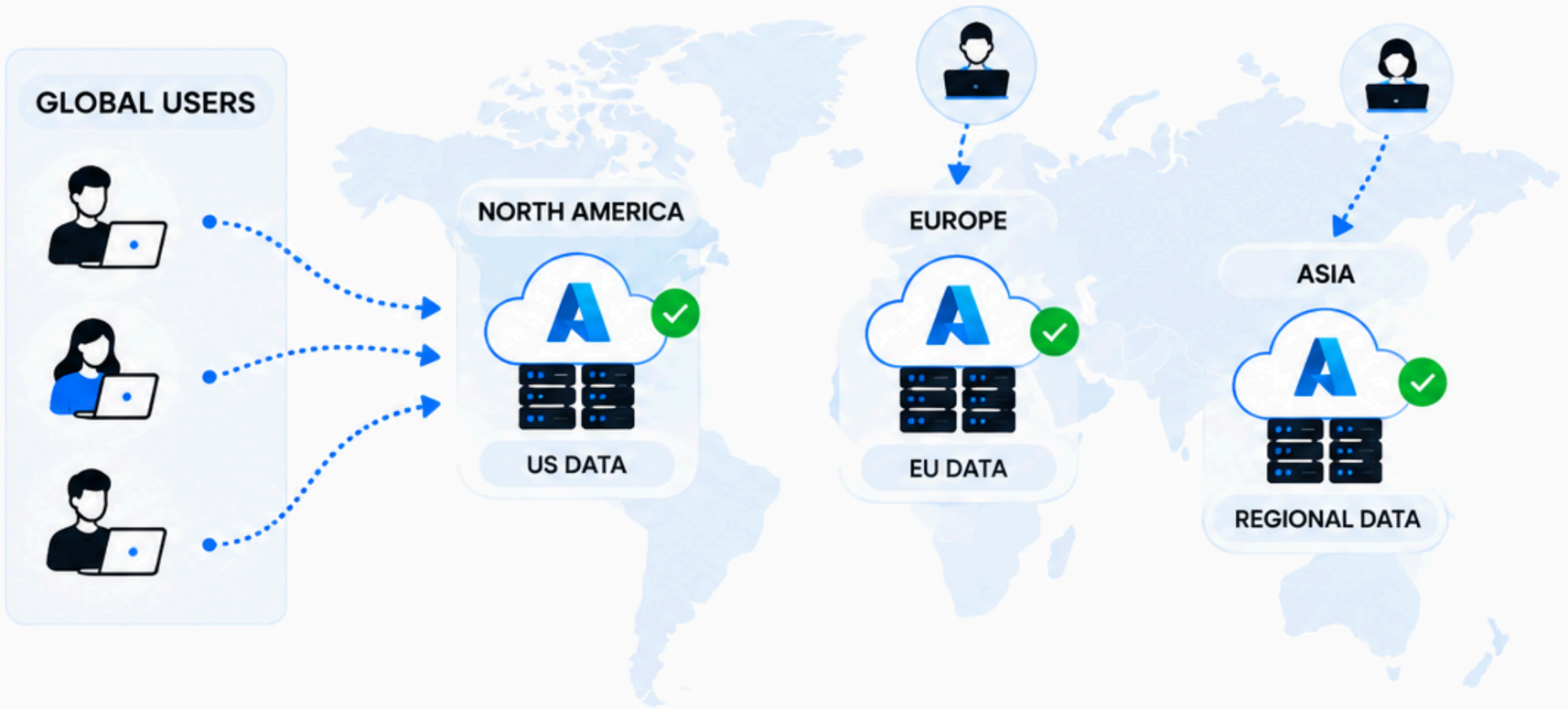




The data needs protection wherever it exists.

Does every user's data have to travel everywhere?

For some workloads and customers, data may need to remain within specific geographic regions because of regulatory or organizational requirements.



DATA RESIDENCY

REGIONAL PROCESSING

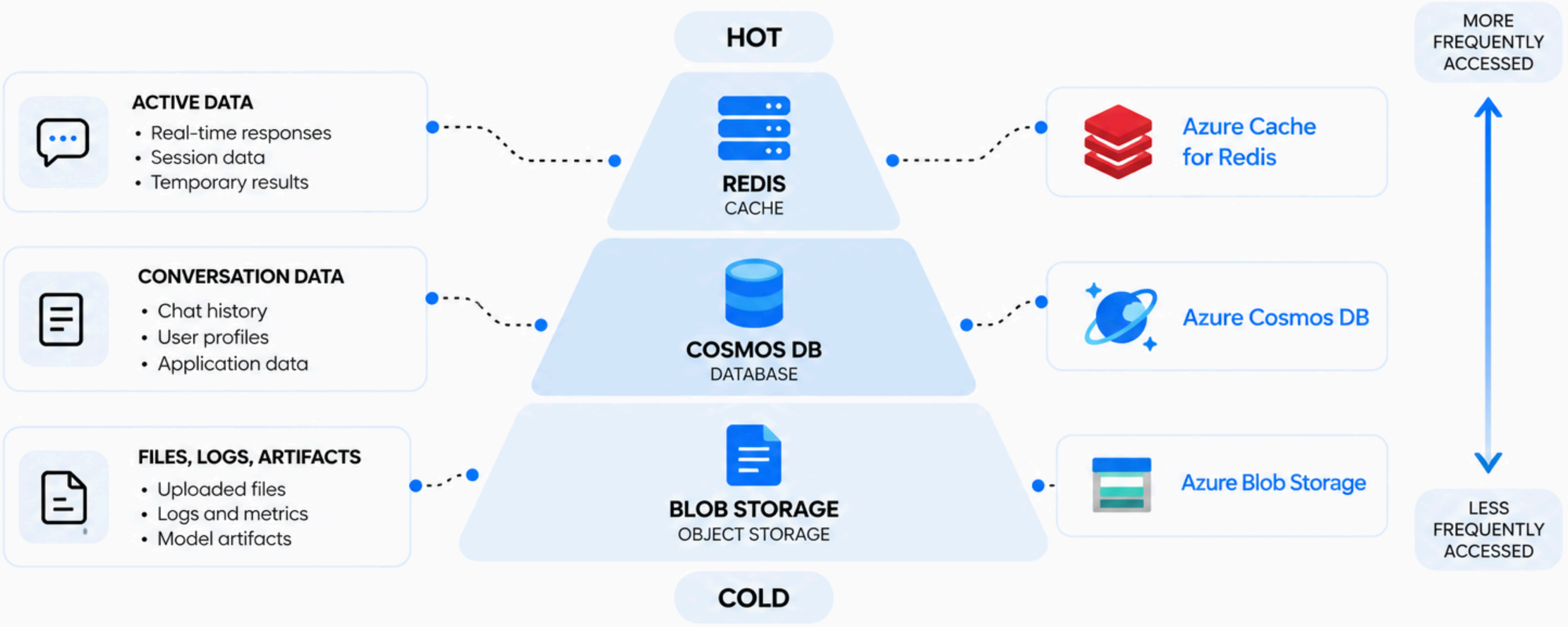
COMPLIANCE

DATA
STAYS WITHIN
REQUIRED REGION

Global infrastructure does not mean every piece of data has to be stored everywhere.

Where does all this data actually go?

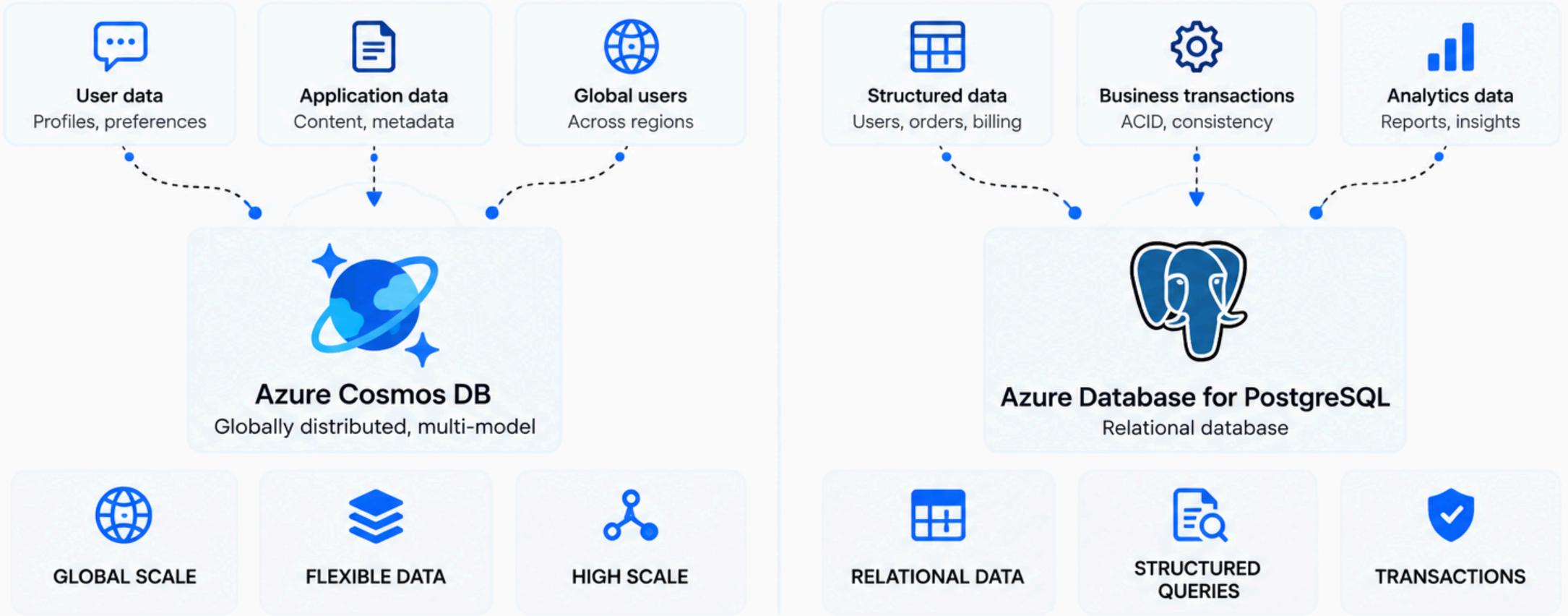
Different types of data need different storage systems depending on how frequently they are accessed and how they are used.



The right storage system depends on the workload.

Why not just use one database for everything?

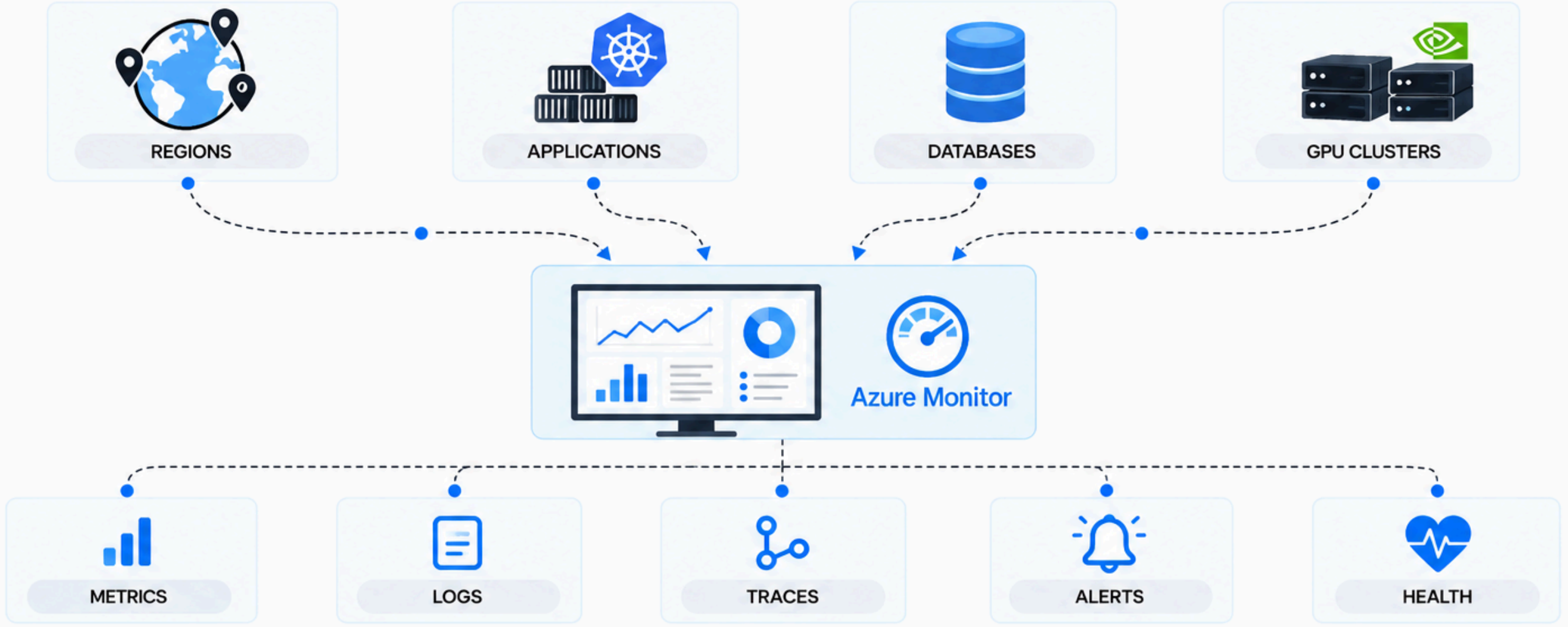
Different workloads need different ways of storing and accessing data.



The database should match the workload, not the other way around.

How do you know what is happening across thousands of services?

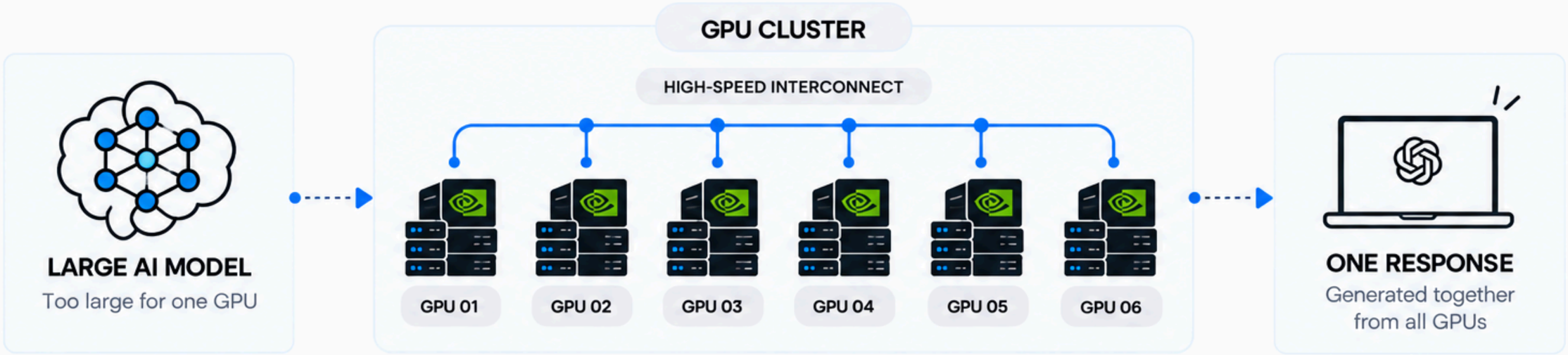
At global scale, the system needs continuous monitoring to detect failures, bottlenecks and unusual traffic before they become major problems.



You cannot operate a system at this scale without being able to see inside it.

What if the model is too large for one GPU?

A large AI model can be distributed across multiple GPUs so that the workload and model parameters can be processed together.



MODEL PARALLELISM
Split the model across multiple GPUs

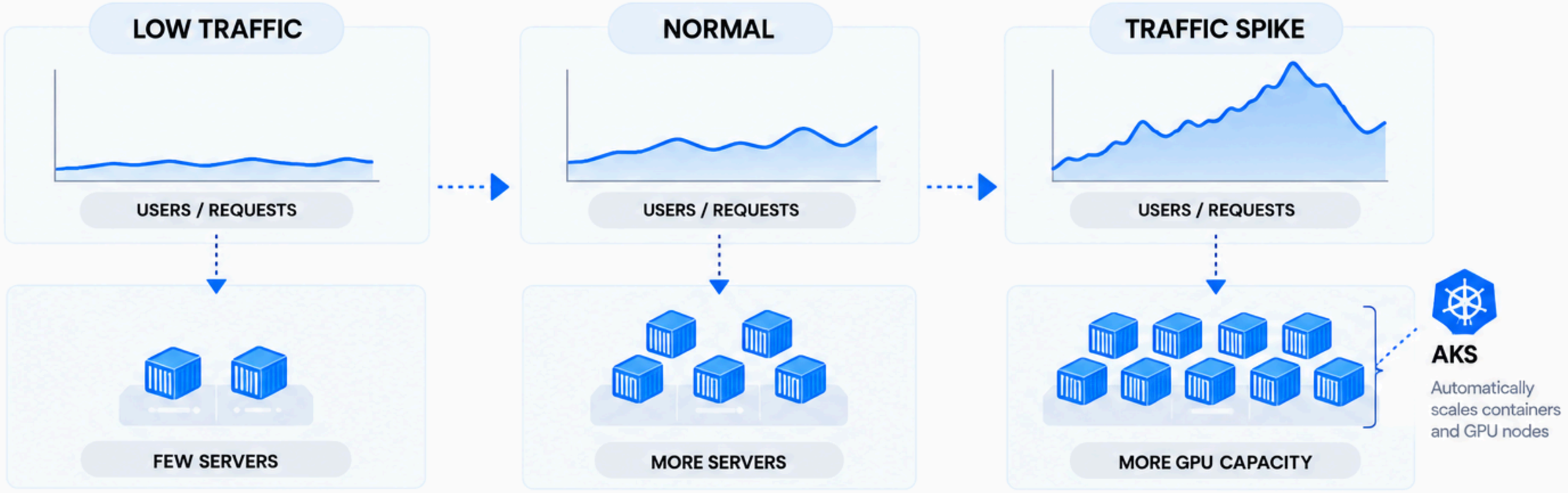
HIGH-SPEED INTERCONNECT
GPUs work together

GPU CLUSTER
Scales to very large models

The model does not have to fit on one machine. The computation can be distributed across many GPUs.

What happens when demand suddenly increases?

The system needs to add capacity when traffic increases and reduce unused capacity when demand falls.



LOAD
Monitor incoming traffic

HEALTH
Check service health

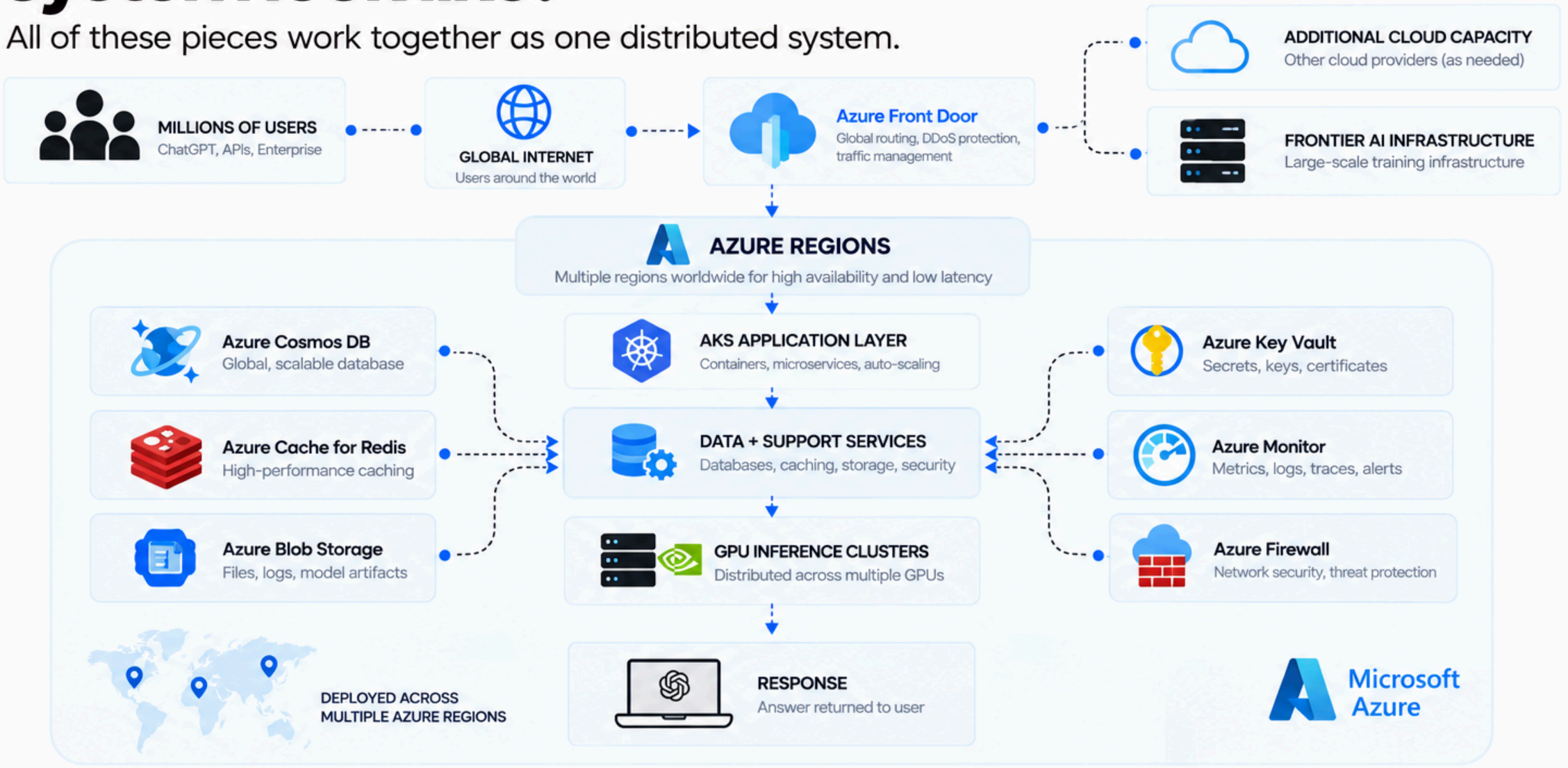
SCALE OUT
Add more containers/GPUs

CAPACITY
Adjust to demand

Scaling means adding capacity before demand overwhelms the system.

So, what does the complete system look like?

All of these pieces work together as one distributed system.



There is no single server doing all the work. The system is a collection of distributed services working together.

So, how does ChatGPT scale to millions of users?



TRAFFIC
Distributed across regions

NOT BY USING ONE GIANT SERVER.

↓

BY DISTRIBUTING THE WORK.



APPLICATIONS
Scaled with containers



AI COMPUTE
Distributed across GPU clusters



DATA
Stored across specialized services

Millions of users require a system that can scale, distribute, protect and recover.

Thank You Very Much