Azure Compute - Detailed Reference

Azure Compute — Detailed Reference (Topics 21-28)

Topic 21: Virtual Machines (VMs)

The most fundamental Azure compute service. Infrastructure-as-a-Service (IaaS). You get a VM, you manage everything above the hypervisor.

VM Sizing — Families

Family	Optimized For	Example Sizes	Use Case
B	Burstable, low cost	B1s, B2s, B2ms	Dev/test, low-traffic web, small workloads
D	General purpose	D2s_v5, D4s_v5, D8s_v5	Web servers, app servers, small databases
E	Memory optimized	E2s_v5, E4s_v5, E8s_v5	In-memory databases, cache, analytics
F	Compute optimized	F2s_v2, F4s_v2, F8s_v2	Batch processing, gaming, high CPU
L	Storage optimized	L8s_v3, L16s_v3	Big data, databases, high disk throughput
N	GPU	NC6s_v3, ND40rs_v2	ML training, rendering, CUDA workloads
H	High performance computing	HB120rs_v3, HC44rs	HPC, fluid dynamics, simulation
M	Memory extreme	M128s, M416ms_v2	SAP HANA, large in-memory databases

v5 suffix = latest generation (better performance per dollar). Always prefer latest generation when available.

VM Naming Convention Decoded

Example: D4s_v5

Position	Meaning
D	Family (general purpose)
4	vCPU count
s	Premium storage supported (always choose this)
v5	Generation/version

Other suffixes:

Suffix	Meaning
s	Premium storage + burst capable
d	Local temp disk included
i	Isolated (dedicated host)
r	RDMA (remote direct memory access for HPC)
m	Memory amplified (more RAM per vCPU)
p	Premium storage only, no temp disk

VM Creation — What You Configure

Setting	Options	What to Choose
VM name	Follow naming convention	e.g., vm-app-prod-eastus-001
Region	Azure region	Where your users/resources are
Availability options	None, Availability Set, Availability Zone, VMSS	Based on HA needs
Image	Windows Server, Ubuntu, RHEL, custom image	OS for your workload
Size	VM family/size	Based on CPU/RAM/disk requirements
Authentication	SSH key (Linux), Password (Windows)	SSH key preferred
OS disk type	Premium SSD, Standard SSD, Standard HDD	Premium SSD for production
OS disk size	30GB-4TB depending on type	127GB default is common
Data disks	Add additional disks	For databases, app data, logs
Public IP	None, Basic, Standard	Standard for production
NSG	Create new or existing	Use existing if you have one
Boot diagnostics	Enable/disable	Enable for troubleshooting
Auto-shutdown	Enable with schedule	Good for dev/test cost savings

VM Disk Types

Disk Type	IOPS	Throughput	Max Size	Cost (1TB/month)	Use Case
Standard HDD	~500	~60 MB/s	32TB	~$20	Backup, non-critical
Standard SSD	~6,000	~300 MB/s	32TB	~$50	Web servers, dev/test
Premium SSD	~7,500	~250 MB/s	32TB	~$125	Production workloads
Premium SSD v2	~80,000	~1,200 MB/s	64TB	~$80+	High-performance databases
Ultra Disk	~160,000	~2,000 MB/s	64TB	~$140+	Mission-critical databases

Premium SSD Performance Tiers

Tier	IOPS	Throughput	Additional Cost
P1 (4GB)	120	25 MB/s	Base
P10 (128GB)	500	100 MB/s	Base
P30 (1TB)	5,000	200 MB/s	Base
P50 (4TB)	7,500	250 MB/s	Base
+ 15K tier	Up to 15,000	Up to 500 MB/s	Extra

Tier	IOPS	Throughput	Additional Cost
+ 20K tier	Up to 20,000	Up to 667 MB/s	Extra

Ultra Disk — Sub-line CRUD

- You can adjust IOPS and throughput on Ultra Disk without detaching the disk
- IOPS can be changed every 5 minutes
- Throughput can be changed every 5 minutes
- Great for databases where you need to scale performance on demand

Disk Caching

Cache Type	When to Use
None	Write-heavy workloads (database log files)
ReadOnly	Read-heavy workloads (database data files)
ReadWrite	OS disk (default) — dangerous for databases without application-aware writes

Availability Options

Option	What it Does	SLA	Cost
Single VM (premium disk)	No redundancy. If host fails, VM goes down.	99.9% (disk only)	Base
Availability Set	2+ VMs across fault domains and update domains	99.95%	No extra (you pay for VMs)
Availability Zone	VMs in different physical data centers in same region	99.99%	No extra (you pay for VMs)
VM Scale Sets	Auto-scaling group of identical VMs	99.95%+	No extra (you pay for VMs)

Availability Sets — Domains

Domain Type	What it Means	Count
Fault Domain	Different physical rack, power, cooling, network. If one rack dies, others survive.	Max 3 (configurable 2-5 in some regions)
Update Domain	Different maintenance schedule. Azure patches one UD at a time. VMs in other UDs stay up.	Max 20 (default 5)

When you put 2 VMs in an Availability Set, Azure automatically places them in different fault domains and update domains.

Availability Zones

- Physically separate data centers within a region
- Each zone has independent power, cooling, networking
- Zones 1, 2, 3 (varies by region)
- VM in Zone 1 + VM in Zone 2 = survive an entire data center failure
- Requires zone-redundant services (Standard Load Balancer, zone-redundant storage)

VMSS — Virtual Machine Scale Sets

Feature	What it Does
Auto-scale	Add/remove VMs based on CPU, memory, or custom metrics
Uniform orchestration	All VMs identical from same image
Flexible orchestration	VMs can have different sizes (newer model)
Scale-in policy	Which VM to delete when scaling in (newest, oldest, default)
Overprovisioning	Azure creates extra VMs during scale-out, deletes failures. Improves reliability.
Automatic repairs	If a VM becomes unhealthy, auto-replace it

Autoscale Rules Example

Rule	Condition	Action
Scale out	CPU > 75% for 5 minutes	Add 1 VM
Scale out	CPU > 90% for 2 minutes	Add 2 VMs
Scale in	CPU < 25% for 10 minutes	Remove 1 VM
Schedule	Weekdays 8:00 AM	Set min count to 3
Schedule	Weekdays 8:00 PM	Set min count to 1

Custom VM Images

- Create a VM, install software, configure settings
- Generalize the VM (sysprep for Windows, waagent -deprovision for Linux)
- Capture as a managed image or Shared Image Gallery image
- Deploy new VMs from this image — all pre-configured

Shared Image Gallery

- Share images across subscriptions, regions, and tenants
- Supports versioning (v1.0, v1.1, v2.0)
- Roll back if needed
- RBAC for access control

VM Extensions

- Small scripts that run on VM after deployment or on-demand
- Examples: Custom Script Extension (run a script), RunCommand (execute commands), Azure Monitor Agent, Dependency Agent, Backup extension

---

Topic 22: Azure App Service

Fully managed PaaS for hosting web applications, REST APIs, and mobile backends.

What App Service eliminates

- No OS management, patching, or updates
- No web server configuration (IIS, Nginx)
- Built-in auto-scale

- Built-in high availability
- Built-in CI/CD (deployment slots, local Git, GitHub Actions)

App Service Plan Tiers

Tier	Use Case	Cost/month
Free (F1)	Dev/test, 60 min/day CPU	Free
Shared (D1)	Dev/test, shared infrastructure	~$10
Basic (B1)	Small production, no auto-scale	~$13
Basic (B2/B3)	Small production, more resources	~$26/$52
Standard (S1)	Production, auto-scale, deployment slots	~$70
Standard (S2/S3)	Production, more resources	~$140/$280
Premium (P1v3)	High performance production	~$140
Premium (P2v3/P3v3)	High performance	~$280/$560

Key differences between tiers

Feature	Free/Shared	Basic	Standard	Premium
Auto-scale	No	No	Yes	Yes
Deployment slots	No	No	5	20
Custom domains	No (shared)	Yes	Yes	Yes
SSL	Shared	Free managed cert	Free managed cert	Free managed cert
Private VNet	No	No	Yes (via integration)	Yes
Backup	No	No	Yes	Yes
Staging	No	No	Yes (slots)	Yes (slots)
Zone redundancy	No	No	No	Yes

Deployment Slots

- Deploy your app to a staging slot without affecting production
- Test the deployment in staging
- Swap staging and production slots — zero downtime
- If something goes wrong, swap back instantly
- Available on Standard tier and above

How slots work:

Production slot: myapp.azurewebsites.net -> v1.0 (live) Staging slot: myapp-staging.azurewebsites.net -> v2.0 (testing)

After swap: Production slot: myapp.azurewebsites.net -> v2.0 (live) Staging slot: myapp-staging.azurewebsites.net -> v1.0 (old)

App Service VNet Integration

- Connect your App Service to a VNet
- App can access resources in the VNet (databases, internal APIs)
- Two types:
  - Regional VNet integration: app integrates with a VNet in the same region. Delegated subnet required.

- Gateway-required VNet integration: older method, uses VPN gateway. Avoid for new deployments.

App Service Networking

Feature	What it Does
VNet integration	App can access resources in VNet (outbound)
Private Endpoints	App is accessible from VNet via private IP (inbound)
Access restrictions	IP-based allow/deny rules for inbound traffic
Hybrid Connections	Connect to on-prem resources via relay (no VNet needed)

Supported Runtimes

Language	Versions
.NET	6, 8, 9
Java	11, 17, 21
Python	3.8-3.12
Node.js	16-20
PHP	7.4-8.2
Ruby	2.7+

App Service for Containers

- Deploy containerized apps without managing Kubernetes
- Pull from Azure Container Registry, Docker Hub, or custom registries
- Linux containers only
- Custom container support on Standard tier and above

Topic 23: Azure Container Apps

Serverless container platform. Run containers without managing VMs or Kubernetes clusters.

Why Container Apps instead of AKS or App Service

Feature	App Service	Container Apps	AKS
Container support	Yes (limited)	Yes (full)	Yes (full)
Kubernetes	No	Hidden (managed)	Yes (you manage)
Auto-scale	Yes	Yes (KEDA-based, event-driven)	Yes (manual or KEDA)
Serverless	No (always running)	Yes (scale to zero)	No
Ingress	Built-in	Built-in (external/internal)	You configure
Dapr	No	Built-in	Manual install
Cost	Per plan	Per vCPU-second + memory	Per node
Complexity	Low	Low	High

Key Features

Feature	What it Does
Scale to zero	No traffic = no containers running = no cost
KEDA autoscaling	Scale based on HTTP traffic, queue length, CPU, custom metrics
Revisions	Each deployment creates a new revision. Active/inactive revisions. Blue/green deployments.
Ingress	Built-in HTTP ingress (external or internal). TLS termination.
Dapr	Built-in microservice building blocks (service discovery, pub/sub, state management)
Secrets	Store secrets in Container Apps environment or reference Key Vault
VNet integration	Connect to internal resources via VNet

Container Apps Environment

- A logical grouping of Container Apps
- Shares a VNet and internal ingress
- All apps in the same environment can communicate internally
- Internal traffic is free

Cost Model

- Pay per vCPU-second and memory-GB-second
- Free grant: 180,000 vCPU-s + 360,000 GiB-s per month
- Active replicas: full price
- Idle replicas (scale to zero traffic but min=1): reduced price

---

Topic 24: Azure Kubernetes Service (AKS)

Managed Kubernetes cluster. You manage the applications, Microsoft manages the control plane.

AKS Architecture

Component	Managed By	What it Is
Control plane	Microsoft (free)	API server, etcd, scheduler — the Kubernetes brain
Node pools	You (paid)	Worker VMs that run your containers
System node pool	You (paid)	Runs system pods (CoreDNS, metrics-server). Min 1 node.
User node pool	You (paid)	Runs your application pods. Can have multiple.

Node Pool Types

Type	When to Use
Standard (VMSS)	Always-on workloads, predictable traffic
Spot	Fault-tolerant, batch processing, cost savings (up to 80% cheaper, can be evicted)
Virtual nodes	Serverless — burst to ACI for quick scale. No VM management.

AKS Networking

Model	How it Works	When to Use
kubenet	Pods get IPs from a separate CIDR, NAT to node IP	Simple, small clusters, IP-constrained VNets
Azure CNI	Pods get IPs from VNet subnet directly	Advanced networking, network policies, pod-to-pod connectivity

AKS Add-ons

Add-on	What it Does
Azure Monitor	Container insights, logs, metrics
Azure Policy	Enforce policies on Kubernetes resources
Ingress controller	NGINX or Application Gateway Ingress Controller (AGIC)
Azure Key Vault provider	Mount secrets from Key Vault as volumes
Open Service Mesh	Service mesh (istio-based) for mTLS, traffic management
Dapr	Microservice building blocks

AKS Security

Feature	What it Does
Entra ID integration	Use Azure AD for Kubernetes RBAC. Users authenticate with Azure AD.
Azure RBAC	Manage Kubernetes permissions via Azure RBAC (unified model)
Private cluster	API server has no public IP. Access only via private endpoint.
Network policy	Control pod-to-pod traffic (Calico or Azure network policies)
Pod identity	Pods authenticate to Azure services using managed identity (Azure AD Workload Identity)
Secrets Store CSI	Mount Azure Key Vault secrets as Kubernetes secrets

AKS Upgrades

- Microsoft releases new Kubernetes versions regularly
- You must upgrade within N-2 (cannot be more than 2 versions behind)
- Upgrade process: cordon + drain nodes, upgrade, uncordon
- Use planned maintenance windows to control when upgrades happen
- Auto-upgrade: set channel (stable, rapid, node-image only)

Topic 25: Azure Container Instances (ACI)

Run a container without provisioning VMs or adopting higher-level services. Fastest way to run a container in Azure.

When to use ACI

Scenario	Choose
Run a one-off task (batch job, script)	ACI
Quick prototyping	ACI
Burst from AKS (virtual nodes)	ACI (via AKS)
Long-running production app	Container Apps or AKS

ACI Properties

Property	Options
OS	Linux, Windows
CPU	1-4 cores
Memory	1-16 GB
GPUs	Available (preview)
Restart policy	Always, OnFailure, Never
VNet integration	Yes (deploy into subnet)
Container groups	Multiple containers in one group (share network, storage)

Cost: Per second, per CPU/memory. No upfront commitment.

Topic 26: Azure VMware Solution (AVS)

Run VMware workloads natively on Azure. Your VMware VMs, on Azure infrastructure, managed together.

Why AVS

- Organization has 100s/1000s of VMware VMs
- Does not want to re-architect everything for cloud
- Wants to move to Azure while keeping VMware tooling (vCenter, vMotion, HCX)
- Gradual modernization path: VMware on Azure then eventually PaaS

AVS Components

Component	What it Is
Private Cloud	An isolated VMware environment on Azure. Contains clusters.
Cluster	Minimum 3 hosts. Each host is a dedicated physical server.
Host	Bare-metal server. SKUs: AV36 (36 cores, 576GB RAM), AV36P (enhanced), AV52 (52 cores).
vCenter Server	Manage VMs using familiar vSphere client
NSX-T	Network virtualization and segmentation within AVS
HCX	Migration tool — move VMs from on-prem vSphere to AVS with minimal downtime

AVS Networking

Connection	How
AVS to Azure native VNet	ExpressRoute (managed, auto-provisioned)
AVS to On-prem	ExpressRoute Global Reach or HCX
AVS to Internet	Managed SNAT or Azure Public IP
Within AVS	NSX-T segments (micro-segmentation)

AVS Cost: ~$8,000-12,000/month per host (3 host minimum). Not cheap — enterprise use only.

Topic 27: Azure Functions

Serverless compute for event-driven workloads. Write a function, it runs when triggered. Pay only for execution time.

Trigger Types

Trigger	When Function Runs
HTTP	When an HTTP request arrives
Timer	On a schedule (cron expression)
Blob	When a blob is created/updated in Storage
Queue	When a message arrives in Storage Queue
Service Bus	When a message arrives in Service Bus
Event Grid	When an event is published
Event Hub	When events arrive in Event Hub
Cosmos DB	When a document changes in Cosmos DB
SignalR	When a SignalR message arrives

Hosting Plans

Plan	How it Works	When to Use
Consumption	Auto-scale, scale to zero, pay per execution	Event-driven, intermittent traffic
Premium	Pre-warmed instances, no cold start, VNet integration	Production, low-latency requirements
App Service (Dedicated)	Runs on your App Service plan	Already have an App Service plan, predictable costs

Cold Start

- Consumption plan: if no executions for a few minutes, function is unloaded
- Next invocation takes 1-5 seconds to start (cold start)
- Premium plan: pre-warmed instances eliminate cold start
- For latency-sensitive apps, use Premium plan

Cost (Consumption)

- First 1 million executions free per month
- ~$0.000016 per execution after
- ~$0.000016/GB-s for memory + execution time

Topic 28: Azure Batch

Run large-scale parallel and HPC (high-performance computing) workloads. Manage hundreds or thousands of VMs for batch processing.

When to use Batch

- Rendering 3D images across 100 VMs simultaneously
- Financial risk modeling across thousands of scenarios

- Video transcoding
- Scientific simulations

How it works

1. Create a Batch account
2. Define a pool of VMs (type, count, auto-scale)
3. Submit a job with one or more tasks
4. Batch schedules tasks across VMs in the pool
5. When tasks complete, results are stored in Storage
6. Pool can auto-scale down when idle
