AKS

https://learn.microsoft.com/en-us/azure/aks/use-system-pools?tabs=azure-powershell#system-and-user-node-pools

*In Azure Kubernetes Service (AKS), what happens when you scale up a node pool? Specifically, how does this scaling process affect the underlying virtual machines (VMs) and the cluster's capacity to handle workloads?*

When you scale up (Horizontal Scaling) a node in a node pool within Azure Kubernetes Service(AKS), it creates the new Virtual Machine in the underlying Azure Infrastructure. This newly created VM is then added to the node pool, effectively increasing the pool's capacity to run more pods or handle increased workload.

Here's how it works:

1. **Scaling Decision:** You or the AKS autoscaler decide to scale up a node pool, either manually or based on workload demands.
2. **VM Creation:** Azure automatically provisions a new VM that matches the configuration of the other nodes in the node pool.
3. **Node Addition:** Once the VM is created, it is joined to the Kubernetes cluster as a new node. This node then becomes available to schedule additional pods.
4. **Workload Distribution:** The Kubernetes scheduler can now distribute workloads across the expanded set of nodes, improving the cluster's ability to handle increased demand.

In summary, scaling up a node pool in AKS directly results in the creation of additional VMs, each of which functions as a node within the Kubernetes cluster, thereby expanding the cluster's capacity.

*How many Node Pools are in AKS? How do they work and where should I use them?*

Within AKS, there are two types of node pools: System Node Pool

User Node Pool

**System Node Pool:** This is Primarily to run all the required Services OR components (CoreDNS, metrics server, and kube-proxy) for Kubernetes Cluster Itself.

* Namespaces, Deployments, Replica Sets and Pods …etc.

These services and the components are necessary for the k8s cluster to run properly, they should be in System Node Pool.

**User Node Pool:** This Primarily to run or host your application pods (actual application).

* This is where our actual application pods are run. Especially for production.

Note:

If you want, you can run your actual application pods within System Node Pool but it’s for a non-production-based application and to Cost Optimizing only.

Overview:

\*\*It’s important to understand for every AKS cluster includes at least one System Node Pool must contain a minimum of two nodes. This ensures redundancy and HA of System services and components.

|  |
| --- |
| **Important**  If you run a single system node pool for your AKS cluster in a production environment, we recommend you use at least three nodes for the node pool. |

**System and user node pools**

In a system node pool, AKS automatically adds a label called **`kubernetes.azure.com/mode: system`** to the nodes. This label makes AKS prefer placing system pods on these nodes. While you can still run your application pods on these nodes, it's recommended to keep system pods separate from application pods. This separation helps protect critical system pods from being accidentally affected by any issues or misconfigurations in your application pods.

System node pools have the following restrictions:

* System node pools must support at least 30 pods.
* System pools osType must be Linux.
* User node pools osType may be Linux or Windows.
* System pools must contain at least two nodes.
* User node pools may contain zero or more nodes.
* System node pools require a VM SKU of at least 4 vCPUs and 4GB memory.
* ‘[B series VMs](https://learn.microsoft.com/en-us/azure/virtual-machines/sizes-b-series-burstable)’ are not supported for system node pools.
* A minimum of three nodes of 8 vCPUs or two nodes of at least 16 vCPUs is recommended (for example, Standard\_DS4\_v2), especially for large clusters (Multiple CoreDNS Pod replicas, 3-4+ add-ons, etc.).
* Spot node pools require user node pools.
* Adding another system node pool or changing which node pool is a system node pool *does not* automatically move system pods. System pods can continue to run on the same node pool, even if you change it to a user node pool. If you delete or scale down a node pool running system pods that were previously a system node pool, those system pods are redeployed with preferred sc.

\*\*It’s important to understand for every AKS cluster includes at least one System Node Pool must contain a minimum of two nodes. This ensures redundancy and HA of System services and components.

|  |
| --- |
| **Important**  **‘CriticalAddonsOnly=true:NoSchedule’**, which will prevent application pods from being scheduled on System node pool. |

System Node Pool Configuration Requirements

In **Azure Kubernetes Service (AKS)**, the **System Node Pool** runs critical system pods like CoreDNS, kube-proxy, and metrics-server. The recommended **CPU and memory** for VMs in the system node pool depend on the **cluster size** and workload.

**Recommended VM Sizes for the System Node Pool**

| **Celuster Size** | **Recommended VM Size** | **vCPUs** | **Memory (GiB)** |
| --- | --- | --- | --- |
| Small (≤10 nodes) | Standard\_D2s\_v3 | 2 vCPUs | 8 GiB |
| Medium (10–100 nodes) | Standard\_D4s\_v3 | 4 vCPUs | 16 GiB |
| Large (100–500 nodes) | Standard\_D8s\_v3 | 8 vCPUs | 32 GiB |
| Very Large (500–5000 nodes) | Standard\_D16s\_v3 | 16 vCPUs | 64 GiB |

**Best Practices for System Node Pools**

1. **Dedicated Node Pool**: Use a **separate node pool** for system workloads.
2. **VM Size Considerations**:
   * At least **2 vCPUs and 8 GiB RAM** to avoid performance bottlenecks.
   * Larger clusters (100+ nodes) benefit from **4–8 vCPUs**.
3. **Taints and Labels**: Apply taints (CriticalAddonsOnly=true:NoSchedule) to prevent user workloads from running on system nodes.
4. **Scaling**: Use **at least 2 nodes** for redundancy.

**For 2 VMs in the System Node Pool:**

Simply multiply the above CPU and memory by 2.

Would you like help configuring your system node pool in Terraform? 😊

**Update existing cluster system and user node pools**

You can change modes for both system and user node pools. You can change a system node pool to a user pool only if another system node pool already exists on the AKS cluster.

Important Things to Consider:

| **Requirement** | **Details** |
| --- | --- |
| **Max Nodes per Cluster** | 5,000 Nodes |
| **Max Nodes per Node Pool** | 1,000 Nodes (**Recommended: 900 Nodes**) |
| **Max Pods per Node** | Up to **250 Pods** (depends on VM size and pod density) |
| **Scaling Beyond 5,000 Nodes** | Deploy multiple AKS clusters and use external load balancers or federation strategies. |

 If each pod has **only one container**, you can run **250 containers per node**.

 If each pod has **two containers**, you can run **125 pods (250 containers)**, and so on.

**Enable Azure Spot Instances**  
Azure Spot Instance offers unused Azure capacity at Low price versus pay as you go prices. Workloads should be tolerant to infrastructure loss as Azure may recall capacity for pay as you go workloads.

Only suitable for User Nodes.