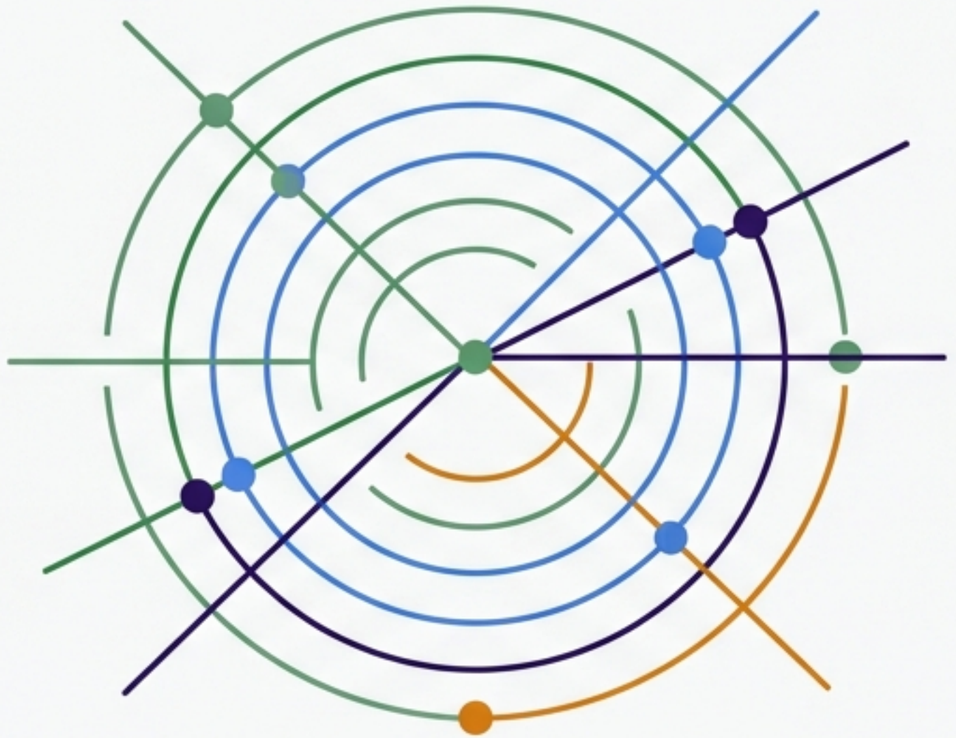
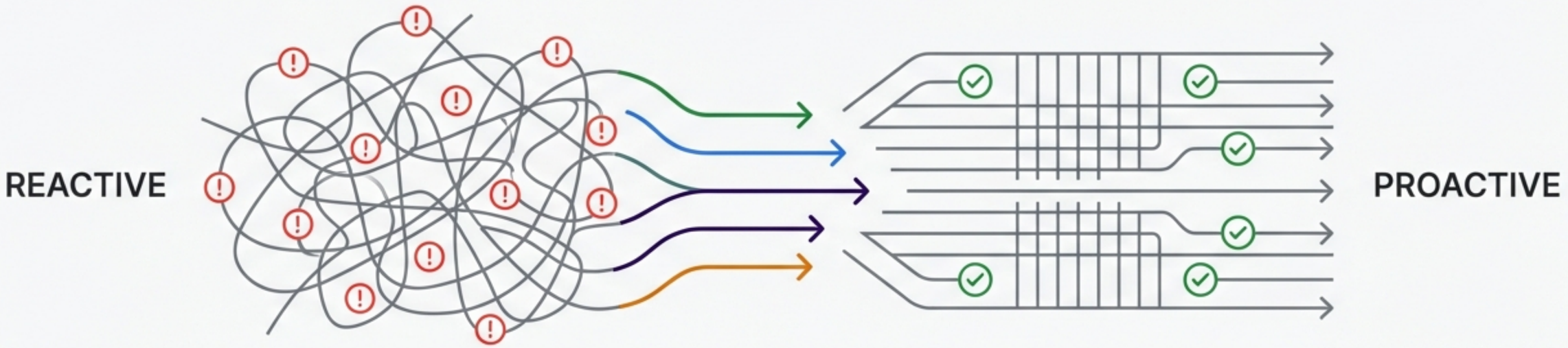
The Proactive Cloud

Mastering Workload Management on Google Cloud



Shift from Reactive Firefighting to Proactive Control



Managing a cloud environment at scale is a constant battle against inefficiency, unpredictability, and risk. The default state is reactive. Google Cloud provides a suite of intelligent, automated capabilities that fundamentally shifts this paradigm, empowering you to proactively manage your workloads.

The Challenge of Cost Efficiency

The Challenge of Reliability

The Challenge of Security

The Challenge of Compliance

The Challenge: Spiraling Costs & Wasted Resources

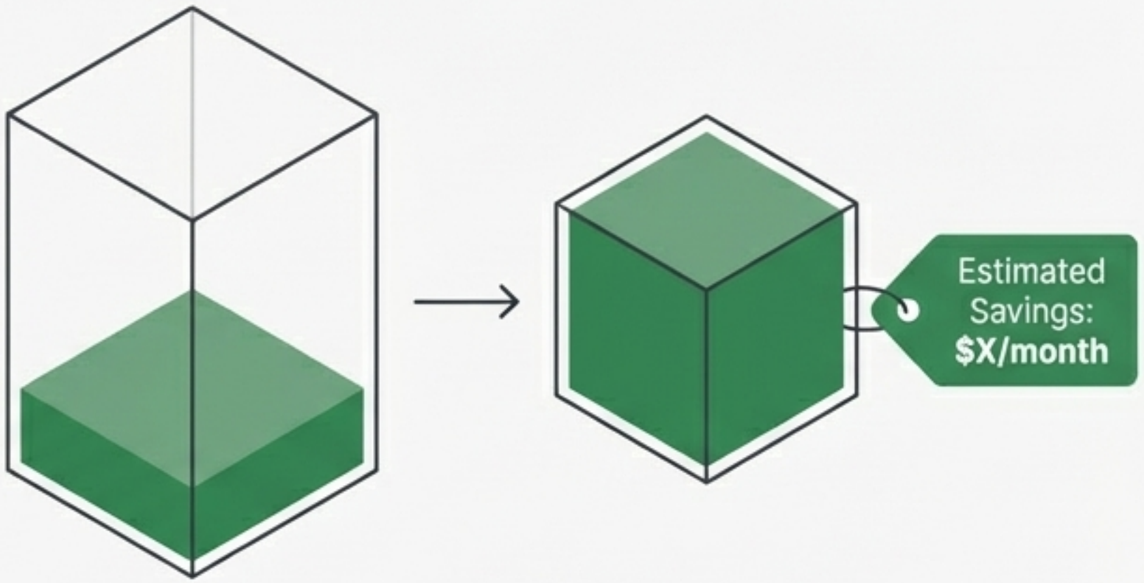
Overprovisioning is the default, leading to significant and often hidden cloud waste. Workloads with fluctuating demand or those that shrink over time frequently run on unnecessarily expensive infrastructure.

- Persistent underutilization of CPU and memory is common.
- Manually tracking and rightsizing hundreds or thousands of VMs is not scalable.
- The business impact is direct: paying for capacity you don't use erodes margins and inflates operational costs.

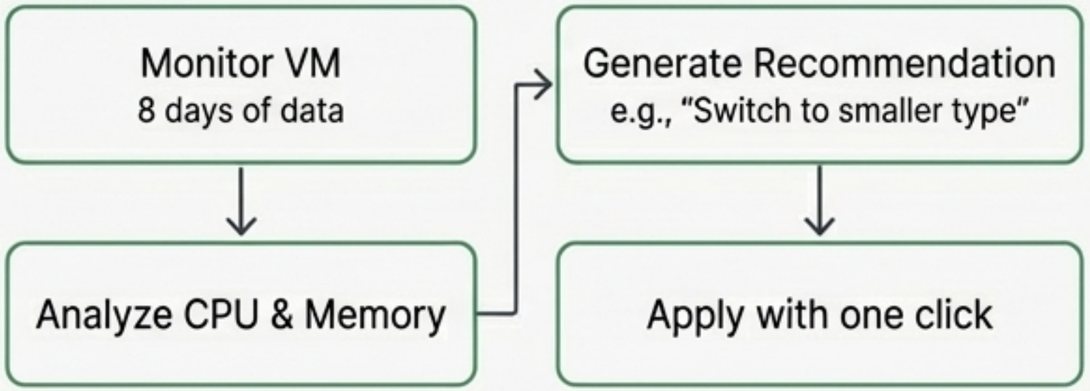
The GCP Capability: Intelligent Rightsizing Recommendations

What

Compute Engine automatically generates machine type recommendations to help you optimize & resource recommendations to help you optimize the resource utilization of your VM instances. These recommendations are based on system metrics gathered by the Cloud Monitoring service.



How

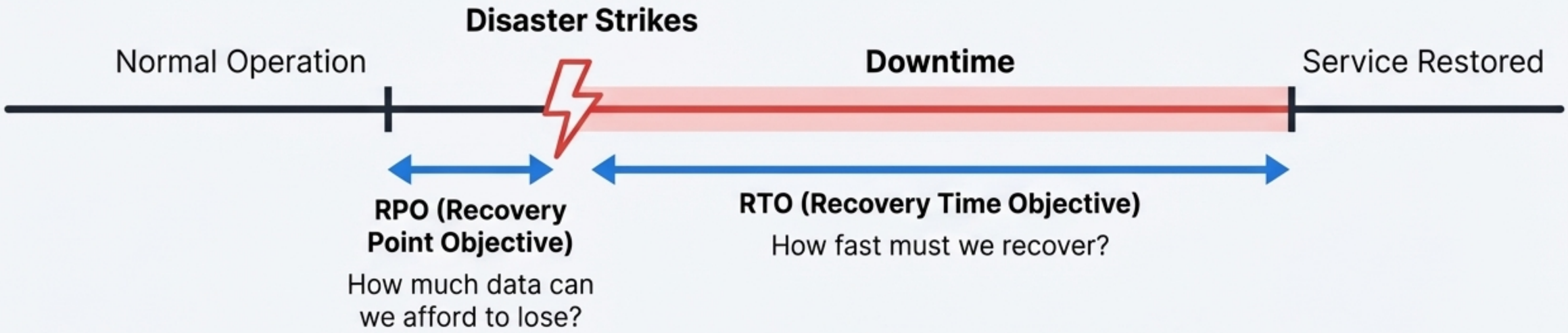


Why (The Benefits)

- **Eliminate Waste**: Stop paying for unused vCPU and memory.
- **Improve Performance**: Identify and resolve resource bottlenecks for constrained VMs.
- **Automated Intelligence**: Frees up engineering time from manual analysis. Cost difference estimations are provided for each recommendation.

The Challenge: The Constant Threat of Downtime

Application failures, infrastructure issues, and regional outages can cripple services, damage user trust, and cause direct revenue loss.



Recovery Time Objective (RTO)

The maximum acceptable length of time your application can be offline.

Recovery Point Objective (RPO)

The maximum acceptable length of time during which data might be lost.

The Goal: Achieving aggressive RTO and RPO targets requires proactive, automated systems that can react to failures faster than humans can.

The GCP Capability: Proactive Autohealing for High Availability

What

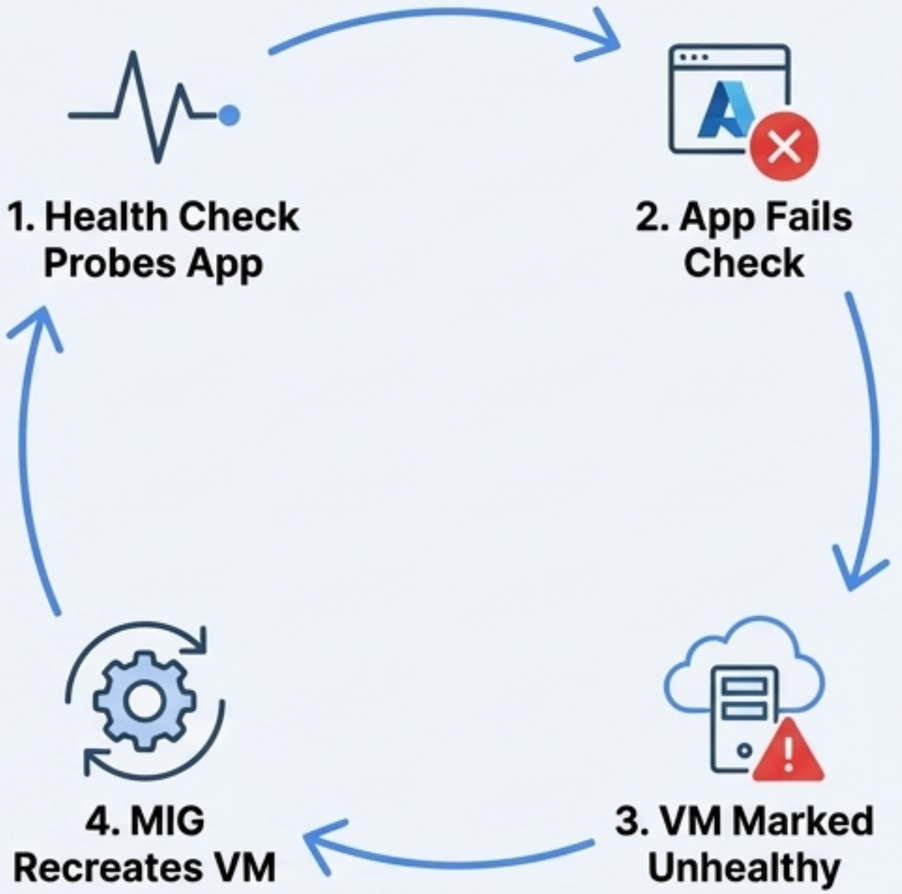
Autohealing is a feature of Managed Instance Groups (MIGs) that automatically detects and recreates unhealthy VMs based on application-level health checks.

How

- A health check is configured to probe a specific port or path on your application.
- If the application on a VM fails to respond correctly for a configured number of consecutive checks (the unhealthy threshold), the MIG marks the VM as `UNHEALTHY`.
- The MIG automatically initiates a repair process, recreating the unhealthy VM to restore service.

Why (The Benefits)

- **Increases Application Availability**: Automatically recovers from application crashes, freezes, or corruption.
- **Reduces Manual Intervention**: Eliminates the need for operators to manually detect and restart failed instances.
- **Helps Achieve Low RTO**: The automated repair process is significantly faster than manual response, helping to meet aggressive recovery time objectives.



Configuring an Effective Autohealing Policy

A well-configured health check is crucial. An overly aggressive check can mistake busy instances for failed ones, reducing availability. A conservative check ensures repairs only happen when necessary.

Key Parameters for a Resilient Health Check

Initial Delay: The time a MIG waits after a VM is created before starting health checks. Allows the application to initialize properly.

Unhealthy Threshold: Number of consecutive failures before a VM is marked unhealthy. Recommended: 3 or more to protect against transient issues.

Timeout: How long to wait for a response. Recommended: Generous, at least 5x the expected response time to account for busy instances.

Check Interval: Frequency of probes. A balance between fast detection and not overloading the system.

Health Check Policy

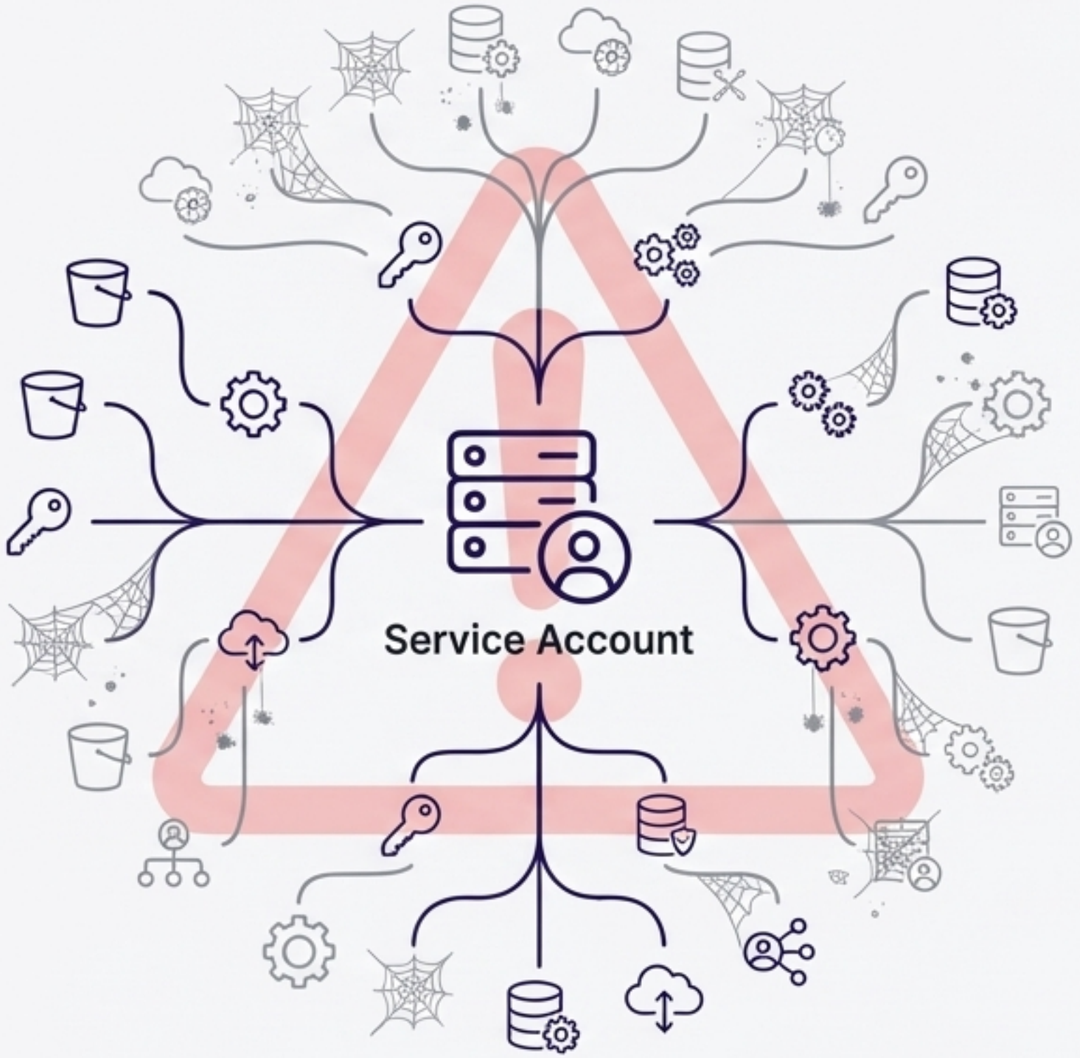
Parameter	Value	Annotation
initialDelaySec	300	The time a MIG waits after a VM is created before starting heaith checks. Allows the application to initialize properly.
unhealthyThreshold	3	Number of consecutive failures before a VM is marked unhealthy. Recommended: 3 or more to protect against transient issues.
timeoutSec	10	How long to wait for a response. Recommended: Generous, at least 5x the expected response time to account for busy instances.
checkIntervalSec	30	Frequency of probes. A balance between fast detection and not overloading the system.

The Challenge: Managing Permissions at Scale is a Security Risk

In large, dynamic environments, identities and service accounts often accumulate permissions they no longer need, creating a massive, unnecessary attack surface.

Key Problems (The “Permission Creep” Effect)

- **Over-privileged Identities:** Principals (users or service accounts) are granted broad permissions “just in case” and they are rarely revoked.
- **Stale Permissions:** A permission granted for a one-time task remains active indefinitely.
- **Lack of Visibility:** It’s nearly impossible to manually track which permissions have been used and which are dormant.
- **Increased Blast Radius:** A compromised, over-privileged account can cause catastrophic damage.



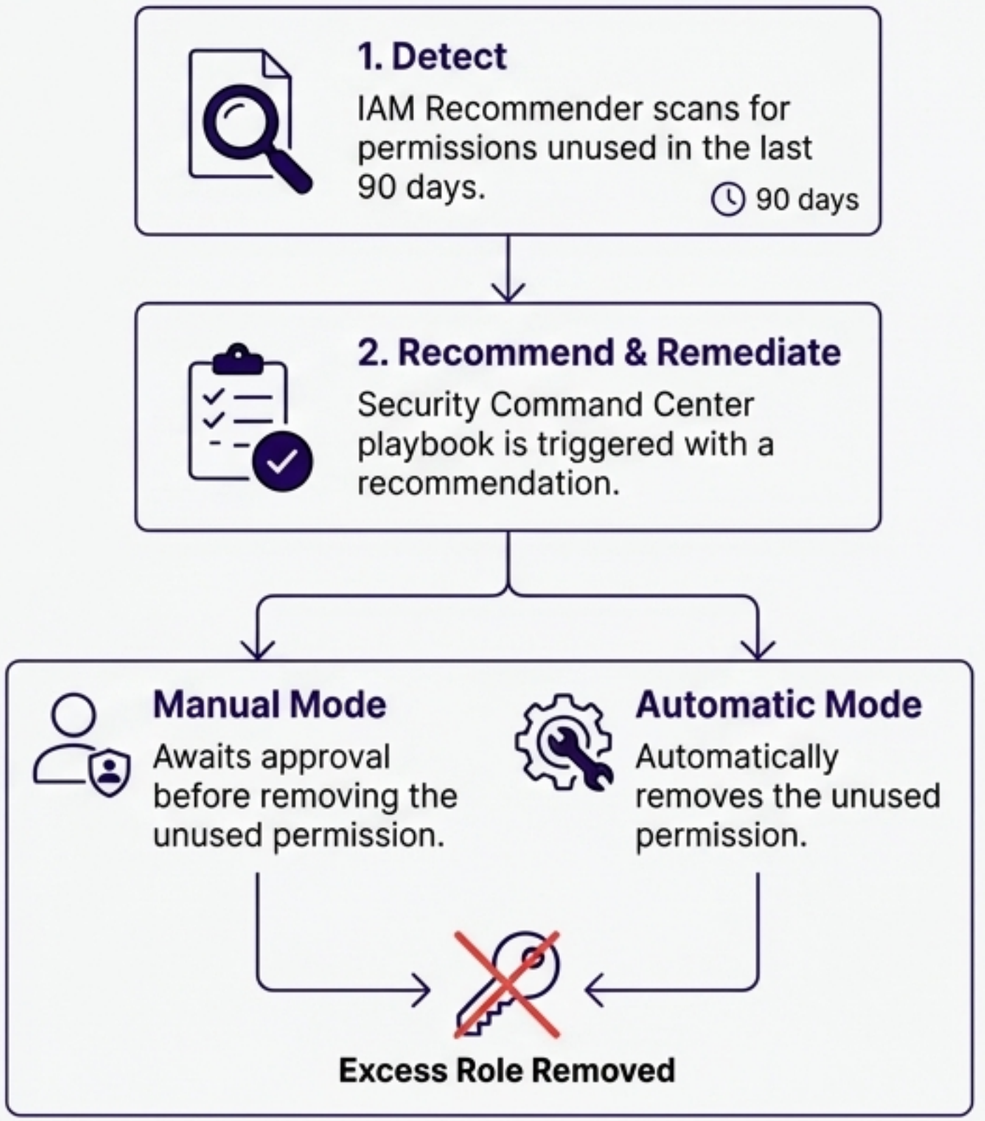
The GCP Capability: Automated IAM Remediation

What: The IAM Recommender, integrated with Security Command Center, identifies excess permissions and provides a playbook to automatically and safely remove them.

How: This process is detailed in the diagram to the right.

Why (The Benefits):

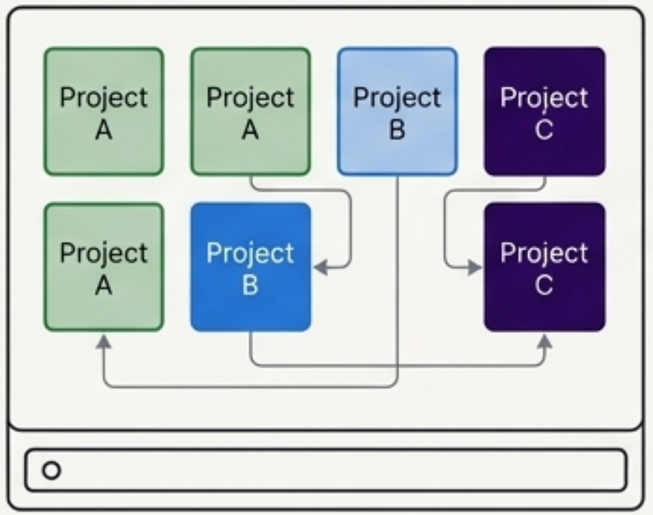
- **Enforce Least Privilege:** Systematically reduces the attack surface by removing unnecessary permissions.
- **Continuous Security:** Turns a manual, periodic audit into an automated, continuous process.
- **Safe Remediation:** Focuses only on permissions that are verifiably unused over a long period, minimizing risk to active applications.



The Challenge: Meeting Strict Isolation & Licensing Requirements

Certain workloads require more than logical separation. They demand physical hardware isolation for compliance, or need to run on dedicated servers to satisfy third-party software license terms.

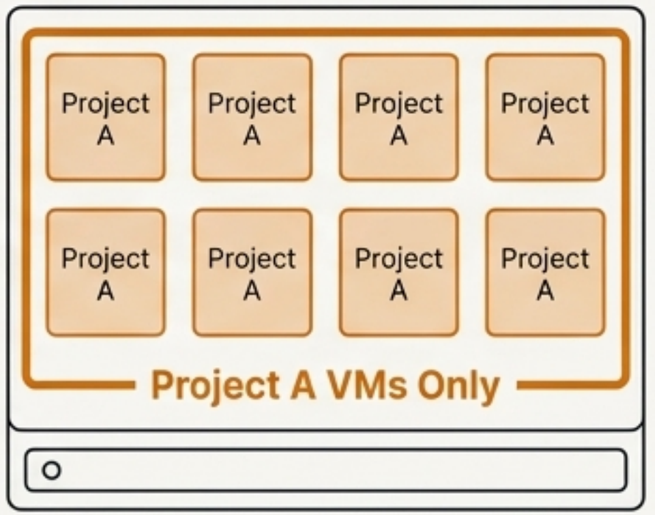
Multi-Tenant Host



Mixed Workloads on Shared Hardware

Exclusive Access, Physically Isolated

Dedicated Host Requirement



Common Scenarios

Compliance

Regulations in industries like finance or healthcare may mandate that workloads run on physically dedicated hardware.

Bring Your Own License (BYOL)

Enterprise software licenses are often sold per-processor or per-core, requiring proof of dedicated physical server use.

Performance Predictability

Eliminating the ‘noisy neighbor’ effect by having exclusive access to a physical host.

The GCP Capability: Sole-Tenant Nodes for Dedicated Hardware

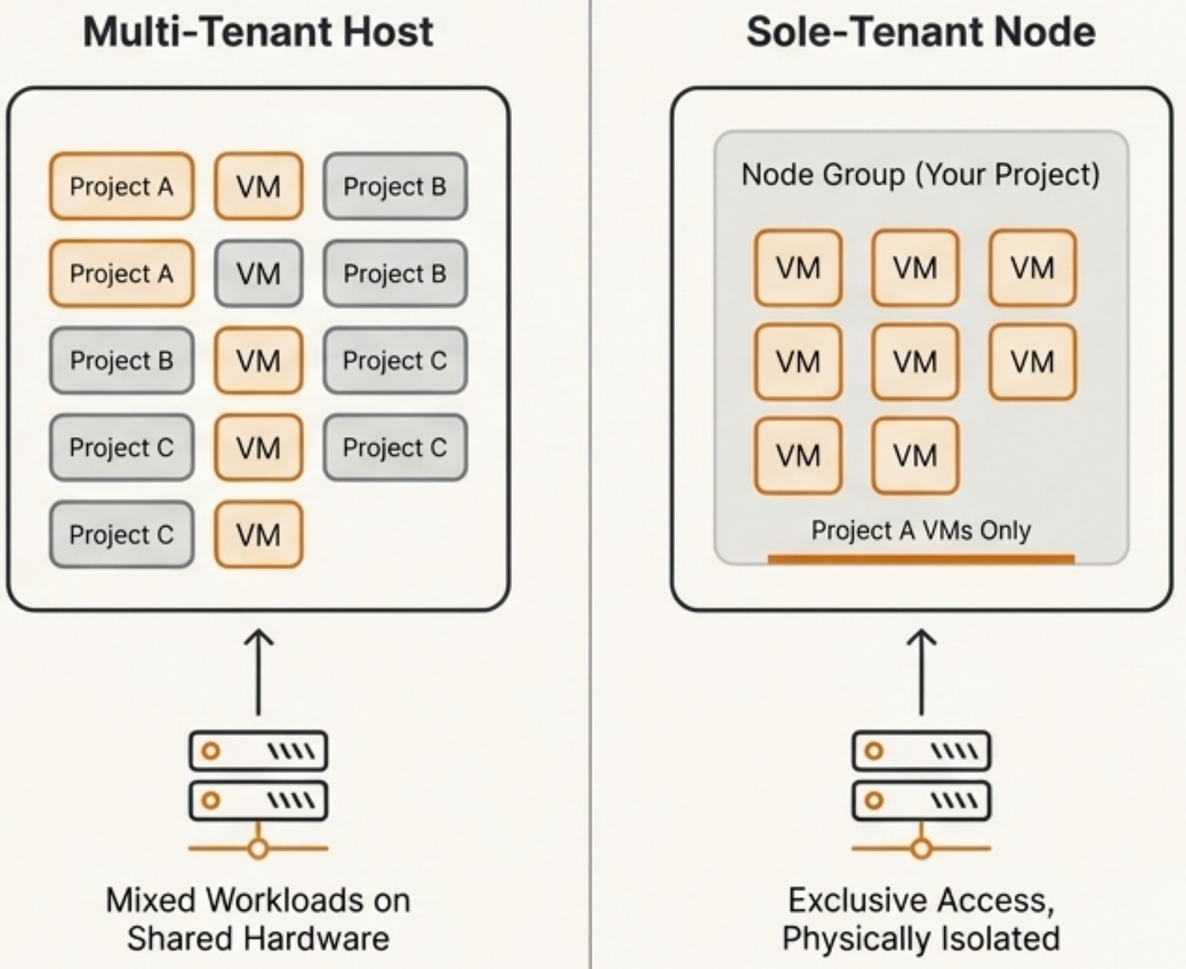
What:
A sole-tenant node is a physical Compute Engine server dedicated exclusively to hosting your project's VMs. It provides hardware isolation, ensuring your VMs are physically separated from those of other projects.

How it Works:

- You reserve a physical server (a 'node') of a specific type (e.g., `n2-node-80-640`).
- This node is provisioned within a 'node group' in a specific zone.
- You then place your VMs onto this node with full control and visibility.

Why (The Benefits):

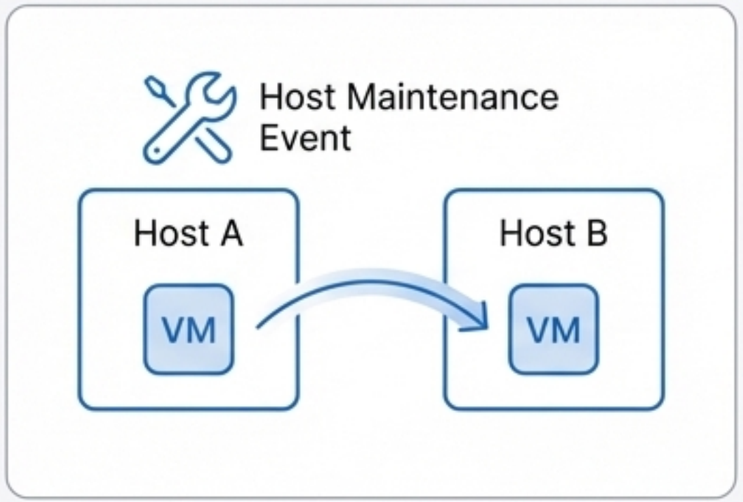
- **Security & Compliance**: Achieves physical isolation to meet strict regulatory requirements.
- **Optimize Licensing Costs**: Enables BYOL scenarios by providing visibility into the underlying physical cores and processors.
- **Performance Consistency**: Guarantees exclusive access to host resources, eliminating performance variability.



Advanced Control over Your Dedicated Infrastructure

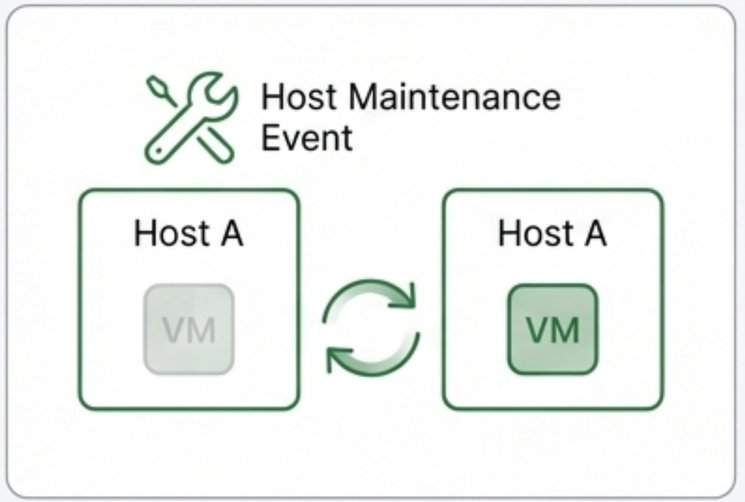
Sole-tenancy offers granular control over VM placement and behavior during host maintenance events, which is critical for managing licensed software.

Key Configuration: Host Maintenance Policy



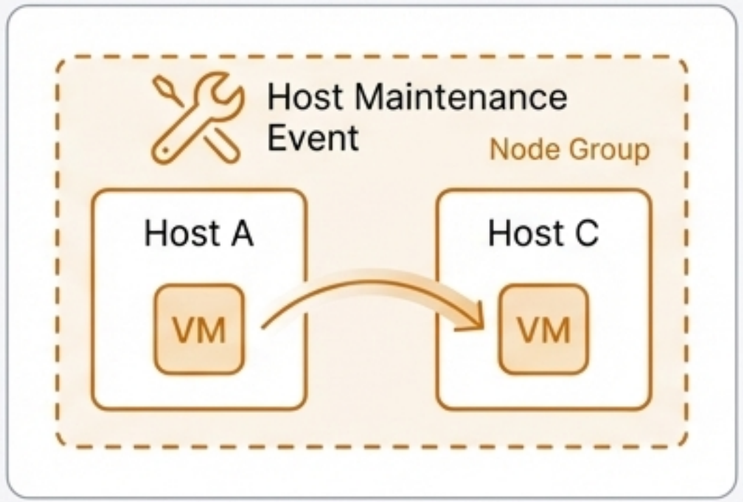
Default

VMs are live-migrated to another sole-tenant node in the group. Ideal for general workloads.



Restart in place

VMs are stopped and restarted on the *same physical server*. Essential for licenses tied to a specific machine.



Migrate within node group

VMs live-migrate only within a pre-defined pool of physical servers, containing licensing costs.

From Reactive Firefighting to Proactive Control

Challenge:
Spiraling Costs

GCP Capability:
Intelligent Rightsizing

Challenge:
Threat of Downtime

GCP Capability:
Proactive Autohealing

Challenge: Security &
Permission Creep

GCP Capability: Automated
IAM Remediation

Challenge:
Compliance & Licensing

GCP Capability:
Sole-Tenant Nodes

Google Cloud's intelligent workload management capabilities empower you to build more cost-effective, resilient, secure, and compliant applications by design.