VIGIL GUARD



# Scaling Profiles 1.8

How Vigil Guard Enterprise 1.8 scales in production environments

## PREDICTABLE SCALING FOR DEPLOYMENTS

Vigil Guard Enterprise 1.8 scales production environments using ready-made scaling profiles. A profile is selected and validated against host resources during installation or an approved configuration change, then persisted.

As a result, the deployment runs reproducibly across restarts, upgrades, and maintenance operations, and the runtime configuration stays consistent with what was approved at install time. Choosing a higher profile raises both the host class and the throughput of the heaviest analysis stages.

## KEY BENEFITS

- Four ready-made scaling profiles matched to the host class
- Profile selected and validated against host resources at install
- A higher profile increases parallelism of the heaviest analysis stages
- Reproducible deployments across restarts, upgrades, and maintenance
- Runtime configuration consistent with the approved installation
- Tuned concurrency limits and overload protection
- Upgrades preserve the selected profile
- Signals when the environment needs a larger host class

**4**
scaling profiles

**20-52**
CPU cores (class range)

**30-256 GB**
RAM (class range)

**1 host**
single-node Linux deployment

On-premise deployment • Air-gapped • Single Linux server

## SCALING PROFILES

Every deployment uses one of four scaling profiles, matched to the host class. A profile defines the minimum server requirements and the intended production use.

| PROFILE | DEPLOYMENT CLASS | CPU (MIN.) | RAM (MIN.) | DISK (RECOMMENDED) | INTENDED USE |
|---|---|---|---|---|---|
| **prod-32-balanced** | Entry | 20 cores | 30 GB | 150 GB+ NVMe SSD | Low traffic, internal environments, or controlled pilots. |
| **prod-64-balanced** | Baseline (recommended) | 28 cores | 64 GB | 150-200 GB NVMe SSD | Default choice for most production deployments. |
| **prod-128-balanced** | Growth | 36 cores | 128 GB | 300 GB+ NVMe SSD | Higher sustained concurrency and larger analytics load. |
| **prod-256-balanced** | Large single-node | 52 cores | 256 GB | 500 GB+ NVMe SSD | Largest single-node deployments with substantial headroom. |

**RECOMMENDED STARTING POINT**

For most new production deployments the recommended profile is **prod-64-balanced** on a dedicated Linux server. It provides safe headroom for the data layer, analytics, and model inference while leaving room to grow. The entry profile prod-32-balanced is the smallest compatible shape, not the default choice for enterprise deployments.

## WHAT A HIGHER PROFILE CHANGES

Scaling in 1.8 is more than a bigger server. Real processing capacity grows with the profile, and the system stays stable under load.

### Greater analysis parallelism

A higher profile increases the throughput of the heaviest detection and analysis stages, improving resilience to peak traffic and unique queries.

### Tuned concurrency limits

Parallel processing limits are matched to the profile to keep system behavior predictable.

### Overload protection

Traffic admission and work budgets limit degradation under load and protect the environment.

### Headroom for data and analytics

A larger host class increases memory, disk, and cache headroom for the event store and analytics workloads.

## PLATFORM LAYERS AND SCALING

The platform runs as a single, cohesive stack on a Linux server. Each layer affects host sizing differently.

| LAYER | ROLE | IMPACT ON HOST CLASS |
|---|---|---|
| **Reverse proxy / TLS gateway** | Traffic entry and TLS termination | Lightweight, constant entry path. |
| **Data and analytics core** | Message bus, cache service, analytics database, vector store, metrics collector | Sets the floor for memory, disk, and I/O. |
| **Application and API** | Request handling and management console | Moderate memory, user-facing latency sensitivity. |
| **Processing and analysis** | Parallel detection layers | Main CPU consumer during analysis. |
| **ML model services** | Detection model inference | Scaled with the profile; the main performance bottleneck. |

## VALIDATION AND REPRODUCIBILITY

Each profile is validated against available host resources: CPU core count, RAM, and free disk space. If the server does not meet a profile's requirements, the profile is not applied. The approved configuration is persisted, so subsequent restarts of the environment are reproducible and the running state does not drift from what was approved.

## UPGRADES AND PROFILE CHANGES

An upgrade preserves the selected scaling profile and re-fits the runtime configuration to the current host resources and the new platform version. This keeps an upgrade from changing the characteristics of a running environment in an uncontrolled way.

Profile changes are made through the provided administrative tooling and confirmed by a deployment verification procedure. This keeps the configuration aligned with the installation process and preserves reproducibility.

> **PERFORMANCE AND ITS DRIVERS**
>
> Each deployment's performance depends on traffic characteristics, the share of unique queries, data retention configuration, analytics load, host parameters, and the customer's environment. Throughput and latency figures should therefore not be treated as a universal guarantee. For larger deployment classes, final performance should be confirmed in the customer's target environment.

## WHEN TO MOVE TO A LARGER PROFILE

The following signals indicate it is worth planning a larger host class and a higher scaling profile:

- Memory usage frequently approaches the configured limits of analytics and model services.
- Analytics load, event volume, or data retention requirements are growing.
- Analysis latency or response time rises under normal business traffic.
- Monitoring reports partial capacity loss or host resource warnings.
- Disk usage approaches the configured thresholds.

When more capacity is needed, move to the next larger host class and its matching scaling profile rather than making ad hoc configuration changes.

## DEPLOYMENT OPTIONS

- **On-premise**: a single dedicated Linux server, full control over your data.
- **Air-gapped**: full functionality without internet access.
- **No GPU required**: model inference runs on CPU.

# Scaling that fits your deployment

Four ready-made scaling profiles let you match Vigil Guard Enterprise to your host class and load, while keeping deployments reproducible and predictable.

**RECOMMENDED STANDARD: PROD-64-BALANCED**

A dedicated Linux server, 28 CPU cores, 64 GB RAM, 150-200 GB NVMe SSD minimum. The most balanced starting point for production deployments: safe headroom for analytics, model inference, and growth.

## VIGIL GUARD ENTERPRISE 1.8

Four ready-made scaling profiles. Validated against host resources. Reproducible deployments.

On-premise, air-gapped, single Linux server, no GPU required.

### Deployment Teams

Reproducible, predictable environment startups

### Operations / SRE

Profile selection and controlled upgrades

### Architecture / Infrastructure

Clear host requirements and capacity planning

**vigilguard.ai** | contact@vigilguard.ai