Resource Requests vs. Limits in Kubernetes

Overview

In Kubernetes, resource requests and limits are crucial for managing container resources effectively.

- **Requests**: Minimum resources guaranteed for a container (used by the scheduler).
- **Limits**: Maximum resources a container can use (prevents overconsumption).

Impact

- **Requests** ensure containers have the minimum required resources, impacting performance.
- **Limits** prevent containers from consuming excessive resources, protecting cluster performance and stability.

Best Practices

1. **Set Requests and Limits Based on Actual Needs**: Analyze usage patterns to determine appropriate values.
2. **Monitor and Adjust Regularly**: Use metrics to refine resource settings.
3. **Avoid Extreme Values**: Prevent inefficient resource usage by avoiding very high or very low settings.

Conclusion

Properly balancing requests and limits optimizes performance and resource utilization in Kubernetes clusters.