Scale-upai server



Overview

GitHub - Nels2/SCALE-API-Helper-MCP: A FastAPI-based tool server for exploring and interacting with the SCALE Computing HyperCore (HC3) API, featuring secure Bearer Auth, session lifecycle control, and full MCP support for OpenWebUI, Claude Desktop, and LLM-powered. GitHub - Nels2/SCALE-API-Helper-MCP: A FastAPI-based tool server for exploring and interacting with the SCALE Computing HyperCore (HC3) API, featuring secure Bearer Auth, session lifecycle control, and full MCP support for OpenWebUI, Claude Desktop, and LLM-powered. Now in its *fifth* year, ScaleUp:AI is one of the first global conference series bringing together top AI builders, executives, luminaries, and innovators. Powered by *Insight* *Partners* and our extensive network of founders, investors, academics, experts, and enterprise leaders, our events across. FastAPI is a high-performance, easy-to-use Python web framework, which makes it a popular way to serve machine learning models. In this blog post, we'll scale up a FastAPI model serving application from one CPU to a 100+ CPU cluster, yielding a 60x improvement in the number of requests served per. Today we will learn how to do auto-scaling with Kubernetes. We will use a FastAPI application as an example. The Foundation: An ASGI Server (Don't Use uvicorn Directly) This is the most common mistake for beginners. When you run you're using Uvicorn's development. Scale-up servers and solutions with the software, services, and support to help you scale your business and turn your data into insights.

© 2026 QK FIBER SYSTEMS - All rights reserved

Article Content

How to Scale Up Your FastAPI Application Using Ray Serve

In this blog post, we''ll scale up a FastAPI model serving application from one CPU to a 100+ CPU cluster, yielding a 60x improvement in the number of requests served per second. All of

Hands On Auto Scaling FastAPI with Kubernetes

Today we will learn how to do auto-scaling with Kubernetes. We will use a FastAPI application as an example. Overview Architecture. In this project,

From Prototype to Production: A Practical Guide to

You need a production-grade ASGI (Asynchronous Server Gateway Interface) server. Uvicorn itself is great, but you should run it via a process

Scaling Up vs. Scaling Out: Choosing the Right Strategy

When designing a server network infrastructure, the choice between scaling up and scaling out should be driven by the specific needs of your

Compute Scale-up Servers | HPE

Modernize and transform core banking with HPE Compute Scale-up servers. Bank on mission-critical availability, security, and low latency. Scale-up servers and

Running and Scaling a FastAPI ML Inference Server

Running and Scaling a FastAPI ML Inference Server with Kubernetes A guide on scaling your model''s inference capabilities.

Vertical Scaling vs Horizontal Server Scaling Solutions and Tips

Server scaling solutions help businesses efficiently expand resources to handle increased workloads. This article will cover different scaling strategies, their benefits, & how to implement them. 3 key

Hewlett Packard Enterprise (HPE) Launches Compute Scale-Up

On May 11, HPE announced the availability of the HPE Compute Scale-up Server 3250, a purpose-built server designed for large, in-memory databases and critical business workloads.

AMD, NVIDIA, OpenAI & Others Form An Optical Scale-up Consortium

AMD, Broadcom, Meta, Microsoft, NVIDIA and OpenAI jointly announced today the formation of the Optical Compute Interconnect (OCI) Multi-Source Agreement (MSA) optical scale-up

© 2026 QK FIBER SYSTEMS - All rights reserved

Code execution with MCP: building more efficient AI

Code execution with MCP improves context efficiency With code execution environments becoming more common for agents, a solution is to

Designed for AI Reasoning Performance & Efficiency

NVIDIA Grace CPU The NVIDIA Grace CPU is a breakthrough processor designed for modern data center workloads. It provides outstanding performance and

NVLink & NVLink Switch: Fastest HPC Data Center

The Need for a Faster Scale-Up Interconnect Reaching the highest performance for the latest AI models requires seamless, high-throughput GPU-to-GPU

Power AI at Scale with NVIDIA MGX™ AI Servers

NVIDIA MGX is a modular server architecture built to power AI, HPC, and cloud-scale workloads. With flexible support for multiple generations of CPUs and

HPE Compute Scale-up Server 3200

Scale up your computing power with the HPE Compute Scale-up Server 3200. It delivers high-performance solutions for intensive workloads, security and reliability.

What is the HPE Scale Up Server?

Conclusion: The HPE Scale Up Server is a powerful and effective solution for businesses looking to expand their IT infrastructure to handle data

Compute Scale-up Servers | HPE

Mission critical HPE Compute Scale-up Servers power critical applications, accelerate analytics and tackle AI workloads and in-memory databases holistically.

Scale Up vs Scale Out: What''s the Difference?

Everything you need to know about when to scale up vs. scale out, and how each can impact cost, performance, and resource use.

HPE Compute Scale-up Servers

HPE Compute Scale-up Servers are designed for growth, speed, and reliability. Run demanding workloads efficiently with HPE Compute Scale-up Servers today.

Related Video Reference

This video was associated with the source search result. Verify technical details against current product documentation and project requirements.

© 2026 QK FIBER SYSTEMS - All rights reserved

Contact Us

Continue your research online:

Website: https://qualitykitchens.co.za
Technical inquiry: https://qualitykitchens.co.za/project-desk.html
African project coordination: Umhlanga Ridge, Durban, KwaZulu-Natal, South Africa

This document is for preliminary research. Verify specifications against current standards, product documentation and project requirements.

© 2026 QK FIBER SYSTEMS - All rights reserved