
Top picks

Kubernetes In Anger

Lobsters firehose

• samof76

Kubernetes is a powerful container orchestration tool that automates the deployment and scaling of applications. However, misconfigurations can cause it to fail spectacularly, leading to downtime and lost revenue. Understanding how to diagnose and resolve these critical errors is essential for maintaining reliable cloud infrastructure.

<https://samof76.space/kubernetes-in-anger.html>

Multi-Stream LLMs: new paper on parallelizing/separating prompts, thinking, I/O

Hacker News top

• atomicthumbs

A new paper proposes multi-stream Large Language Models that split prompts into separate thinking, input, and output streams to run in parallel. This architecture aims to drastically reduce latency and improve efficiency by allowing different parts of the request to be processed simultaneously. For developers, this could mean faster response times and lower costs, making advanced AI more practical for real-time applications.

<https://arxiv.org/abs/2605.12460>

Announcing etcd 3.7.0-beta.0

Kubernetes blog

etcd v3.7.0-beta.0 introduces RangeStream, a feature that streams large resultsets in chunks to reduce latency and memory usage. This beta also marks the end-of-life for version 3.4 and includes significant security and reliability improvements. Developers are encouraged to test the new binaries and report any issues found.

<https://kubernetes.io/blog/2026/05/20/etcd-370-beta>

Also worth a glance

Erasing Existentials

Lobsters firehose

<https://wolfgirl.dev/blog/2026-05-20-erasing-existentials>

Show HN: Rmux – A programmable terminal multiplexer with a Playwright-style SDK

Hacker News top

<https://github.com/helvecsec/rmux>

Welcome to the Raspberry Pi Podcast

Raspberry Pi

<https://www.raspberrypi.com/news/welcome-to-the-raspberry-pi-podcast>

Show HN: KVBoost – chunk-level KV cache reuse for HuggingFace, 5–48x faster TTFT

Hacker News top

<https://pythongiant.github.io/KVBoost>

CODA: Rewriting Transformer Blocks as GEMM-Epilogue Programs

Hacker News top

<https://arxiv.org/abs/2605.19269>

Twelve Ways to Be Wrong About AI-Assisted Coding

Lobsters firehose

<https://third-bit.com/2026/05/20/twelve-ways-to-be-wrong>

Saying goodbye to asm.js

Lobsters firehose

<https://spidermonkey.dev/blog/2026/05/20/saying-goodbye-to-asmjs.html>

Bliki: Vibe Coding

Martin Fowler

<https://martinfowler.com/bliki/VibeCoding.html>

Qwen3.7-Max: The Agent Frontier

HN Best

<https://qwen.ai/blog?id=qwen3.7>

[\$] BPF support in GCC 16 and beyond

LWN

<https://lwn.net/Articles/1071973>

How fast is 10 tokens per second really?

Simon Willison

<https://simonwillison.net/2026/May/20/tokens-per-second>