All the cloud native news that's fit to reconcile

The Daily Reconcile.

COUNTDOWN
2 days
to the Conf42 Chaos CFP

VOL. I · NO. 321 TUESDAY, NOVEMBER 17, 2026 DAILY EDITION · KUBERNETES & PLATFORM ENGINEERING

INSIDE TODAY

◆ LLM RELEASE
Alibaba's Qwen4-Max: $1.50 in, $4.50 out, 1M context, no open weights P.2

◆ KUBERNETES AT SCALE
AWS pushes Kubernetes clusters toward 100,000 nodes and 300 pods a second P.2

◆ INGRESS
Traefik 3.7.18 fixes a forged client-certificate bypass in its NGINX provider P.2

KUBERNETES NETWORKING

Kubernetes will remove kube-proxy's IPVS mode in 1.40

Deprecated since 1.35, the ipvs backend should give way to nftables, GA since 1.33. SIG Network says every ipvs user should be planning the move.

COMPILED FROM KUBERNETES BLOG, NOVEMBER 16

Kubernetes SIG Network plans to remove kube-proxy's ipvs mode in 1.40, Dan Winship of Red Hat wrote on the Kubernetes blog on November 16. It was deprecated in 1.35; KEP-5495 has the timeline.

The backend is complex, not 100% feature complete and still leans on iptables for some Service features; the IPVS API has no batch updates, and kube-proxy cannot use most IPVS schedulers usefully. Red Hat also plans to remove the iptables API from RHEL.

On 1.36 or later, with kernel 5.13+ nodes, move to nftables: on kubeadm, set mode: nftables in the kube-proxy ConfigMap and restart its pods. If you fall back to iptables, reboot each node.

Full story → kubernetes.io/blog/2026/11/16/ipvs-removal

KUBE-PROXY BY KUBERNETES VERSION

33 NFTABLES GA, 1.33

35 IPVS DEPRECATED, 1.35

36 NFTABLES ADVISED FROM 1.36

40 IPVS REMOVAL PLANNED, 1.40

Values are Kubernetes 1.x minor versions. nftables is advised on nodes with kernel 5.13 or later. Source: Kubernetes Blog, KEP-5495.

AI INFERENCE

vLLM 0.34 sizes the KV cache from free memory and drops Model Runner V1

MiniMax-M3 on H200 gets 672k cache tokens instead of 433k, but some H200 models lose 6–18% in per-token decode time.

COMPILED FROM VLLM RELEASE NOTES, GITHUB, NOVEMBER 16

vLLM 0.34.0 (November 16) has 1,061 commits from 412 contributors. Its extensible KV cache, now on by default, is sized from memory free after warmup: MiniMax-M3 on H200 grows from 433k to 672k tokens.

TPOT rises 6–18% on some H200 models at concurrency 128; opt out with --no-enable-extensible-kv-cache. New models include Mistral Large 4 (1T total, 52B active).

Before upgrading: Model Runner V2 is the only GPU runner, and VLLM_USE_V2_MODEL_RUNNER=0 fails at startup. The default --block-size is 64, not 16.

Full story → github.com/vllm-project/vllm/releases/tag/v0.34.0

CONTAINER ISOLATION

AWS Nitro Sandboxes give each EKS pod its own microVM, at no extra charge

COMPILED FROM INFOQ, NOVEMBER 16

Set runtimeClassName to isolated-sandbox or confidential-sandbox and each pod runs in its own Firecracker microVM, in all commercial regions. Sandbox vCPUs are capped at five times the node's. GPUs are not supported, and AWS publishes no overhead figures.

Full story → www.infoq.com/news/2026/11/aws-nitro-sandboxes-eks

DEVSECOPS

GitLab 19.5.6 fixes 37 flaws, one exposing other namespaces' Kubernetes Secrets

COMPILED FROM DAILY CYBERSECURITY, NOVEMBER 17

CVE-2026-97187 (7.6) in Workspaces let privileged users read Secrets across namespaces. Self-managed installs should upgrade; GitLab.com is patched.

Full story → securityonline.info/gitlab-patch-release-19-5-6

MCP SECURITY

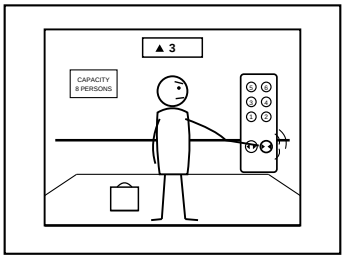
Kubeshark's MCP server let one injected argument take over a cluster

COMPILED FROM GITHUB ADVISORY DATABASE, NOVEMBER 16

CVE-2026-129236 (7.2): with --allow-destructive, a pod_regex value became a Helm override pulling attacker images. v53.4.0 fixed it in August.

Full story → github.com/advisories/GHSA-8vf2-hxg5-crqh

CARTOON OF THE DAY



Drawing: Claude

"I know it isn't wired to anything. It still helps."

WHY IT'S FUNNY The door-close button in many lifts is said to do little or nothing in normal use, yet people keep pressing it while they wait. He knows it is a placebo and presses it anyway, because pressing something feels better than standing still.

DATES TO WATCH
Events and calls for papers

Nov 17 EVENT
OpenInfra Day North America, Chicago

Nov 18 EVENT
gRPConf India, Bengaluru

Nov 19 CFP
Chaos Engineering (Conf42, online, Dec 3): talks due

NEWSROOM

PRESS RELEASE · FROM GITHUB RELEASES AND NOTIFICATIONS

Krateo runs many composition controllers in one process

November 17, 2026 — On November 16, core-provider added a pool mode for composition controllers, oasgen-provider stopped leaking ClusterRoles and observability gained alerts for vendor 5xx errors.

KRATEO PLATFORMOPS

GitHub releases, github.com/krateo-platformops

core-provider 2.13.105–2.13.107 · Nov 16

- Pool mode: one process runs many isolated members.
- Follower definitions publish schemas in their own namespace.

github.com/krateo-platformops/core-provider/releases

installer 0.3.668–0.3.671 · Nov 16

- Pins Kyverno 3.9.2; Keycloak routed in path mode.

github.com/krateo-platformops/installer/releases

oasgen-provider 0.74.0–0.77.0 · Nov 16

- Settable request timeout; no leaked ClusterRoles.
- Five transient existence refusals now requeue.

github.com/krateo-platformops/oasgen-provider/releases

unstructured-runtime 1.22.0–1.24.0 · Nov 16

- Unknown create outcomes confirmed before a retry.

github.com/krateo-platformops/unstructured-runtime/releases

portal, frontend 1.8.129–1.8.130, 1.8.48 · Nov 16

- Callers see only blueprint controls they may use.

github.com/krateo-platformops/portal/releases

observability 3.1.95–3.1.102 · Nov 16

- Vendor 5xx alerts; rollup.ttlDays minimum now 31.

github.com/krateo-platformops/observability/releases

Latest elsewhere: snowplow 1.12.70 (Nov 11; 1.12.71–1.12.73 are pre-releases), authn 0.52.0 (Nov 14), provider-runtime 1.20.0 (Nov 14), alert-provider 0.2.85 (Nov 15).

KRATEO AGENTIKO

Merged into main this week

Krateo Agentiko notifications could not be read this morning.

NEWS IN BRIEF

LLM RELEASE

Alibaba's Qwen4-Max: $1.50 in, $4.50 out, 1M context, no open weights

The mixture-of-experts model has 2.4T total and 71B active parameters, and runs on the QwenCloud API, Qwen Studio and Qoder; cached input costs $0.075. Artificial Analysis scores it 43, behind Kimi K3.1 Pro at 46. Only smaller Qwen4 models will get open weights, with no date given.

— The Decoder, SiliconANGLE, November 16
the-decoder.com/alibabas-qwen4-max-takes-on-opus-5-5-gpt-6-and-kimi-k3-1-pro-at-much-lower-prices

KUBERNETES AT SCALE

AWS pushes Kubernetes clusters toward 100,000 nodes and 300 pods a second

AWS's Mike Stefaniak says the fixes shipped in Kubernetes 1.37; in a million-pod test, a watch-cache lock wait fell from over 0.5 s to 2 ms. containerd 2.2's parallel unpack cut a 10 GB image's pull from 2 minutes to under 45 seconds. The figures are AWS's own.

— Cloud Native Now, November 13
cloudnativenow.com/features/kubernetes-is-strengthening-its-position-as-the-ai-platform-of-choice

KUBECTL TIP

```
kubectl edit svc/docker-registry
```

Edit the service named docker-registry *(kubernetes.io)*

INGRESS

Traefik 3.7.18 fixes a forged client-certificate bypass in its NGINX provider

GHSA-ccgh-9g52-6vwg (CVSS 7.0) lets an unauthenticated client send forged Ssl-Client-* headers on shared hosts in v3.7.15 to 3.7.17. Version 2 is not affected. The latest releases are 3.7.19 and 2.11.63, from November 13.

— GitHub advisory, November 12
github.com/traefik/traefik/security/advisories/GHSA-ccgh-9g52-6vwg

INFERENCE

Hugging Face routes LLM inference through OpenRouter, with no markup

Inference Providers now uses OpenRouter's router for language models; :fastest and :cheapest suffixes pick a policy, and failed requests retry on another provider. You can no longer bring your own provider key for these models. PRO users get $2 a month in credits.

— Hugging Face blog, November 16
huggingface.co/blog/openrouter

DEVELOPER TOOLS

AWS Emulator mimics more than 200 services locally, under CC0

The Rust tool is generated from over 400 Smithy API models; point AWS_ENDPOINT_URL at it. It does not evaluate IAM or resource policies, so AWS says not to use it for security testing. A third-party count finds CloudFormation passing under a fifth of tests.

— InfoQ, November 16
www.infoq.com/news/2026/11/aws-emulator-agents

THREAT INTELLIGENCE

Google finds an attacker tool that turns patches into exploits unattended

GTIG found exploit_agent on an exposed staging server: it diffs patched code, drafts exploits and runs them. Google names no actor and gives no CVE or victim count. It expects low-skill attacks with little to no human interaction in about three months.

— VentureBeat, BleepingComputer, November 16
venturebeat.com/security/google-says-attackers-built-an-ai-framework-to-turn-software-patches-into-exploits-without-human-intervention

SUPPLY CHAIN

Optimizely's CDN served injected malware to visitors of 178 organisations' sites

For about 13 hours on November 16, the code swapped wallet addresses and skimmed cards.

— BleepingComputer, November 16
www.bleepingcomputer.com/news/security/optimizely-supply-chain-attack-delivers-malware-to-cnn-fox-news-bbc-salesforce-sites

HELM TIP

```
helm get notes my-release
```

Show a release's chart notes *(helm.sh)*

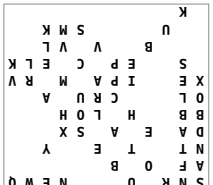
PUZZLE · WORD SEARCH

Ten words from today's two pages are hidden in the grid.

FIND EMULATOR · GITLAB · IPVS · KUBEPROXY · KUBESHARK · KVCACHE · NFTABLES · QWEN · SANDBOX · VLLM

C	H	P	K	G	I	T	L	A	B	V	S
S	N	R	E	U	T	D	A	N	E	W	Q
A	F	T	O	O	B	C	N	W	I	U	M
N	T	U	S	T	H	E	I	T	Y	D	A
D	A	I	E	W	A	U	S	X	I	E	G
B	B	F	Y	H	A	L	O	H	I	B	U
O	L	E	G	T	C	R	U	R	A	G	L
X	E	W	R	I	P	A	S	M	D	R	V
N	S	I	E	E	P	N	C	A	E	L	K
C	R	A	B	F	R	V	T	V	L	M	P
B	H	U	I	L	S	D	S	M	K	G	I
V	K	G	W	H	Y	Y	W	H	R	O	S

Words run in every direction, also backwards.



SOLUTION