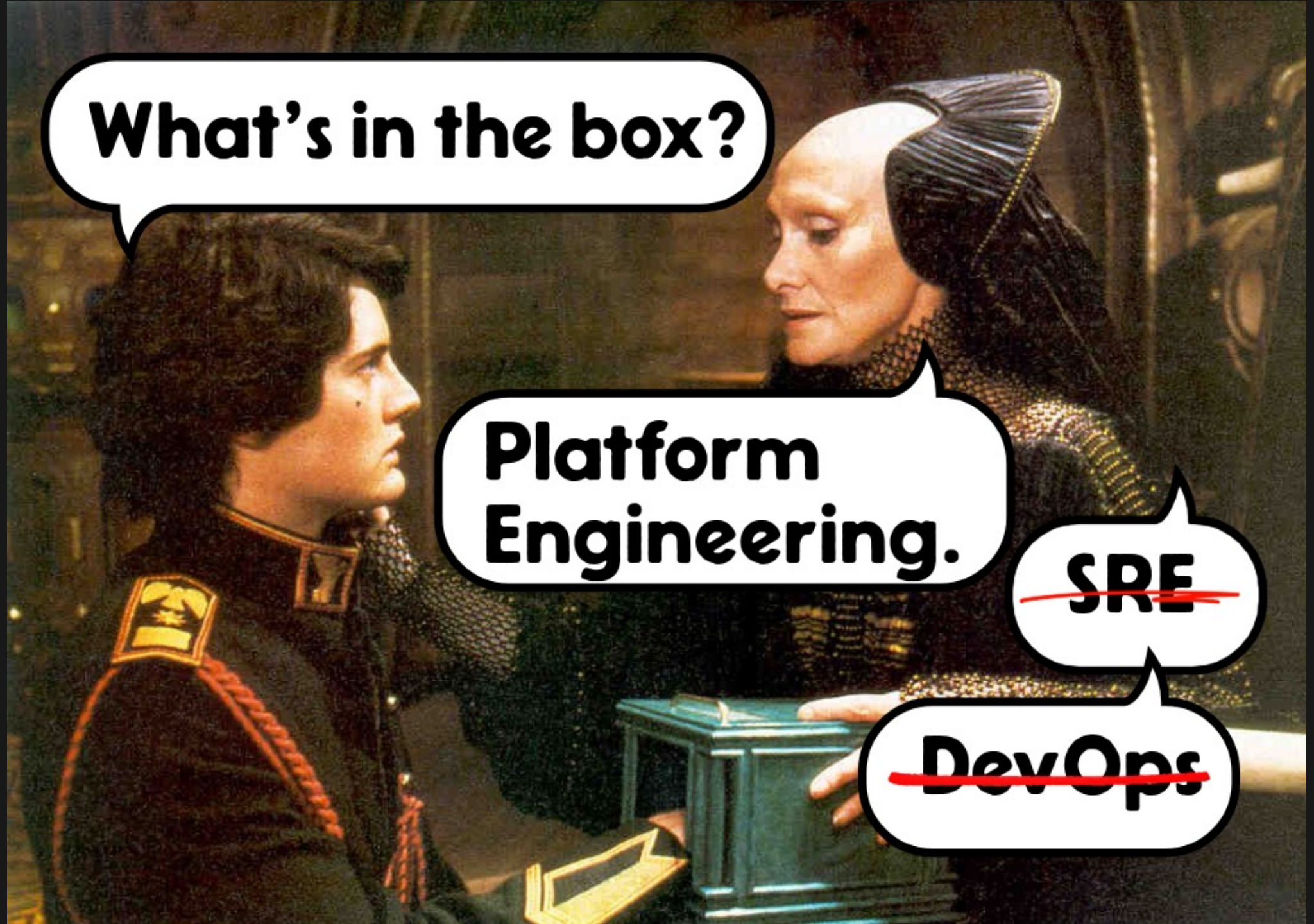
Platform Engineering
for Private Cloud

Coté – Civo Navigate London, September 30th, 2025


What's in the box?
Platform Engineering.
~~SRE~~
~~DevOps~~

50%

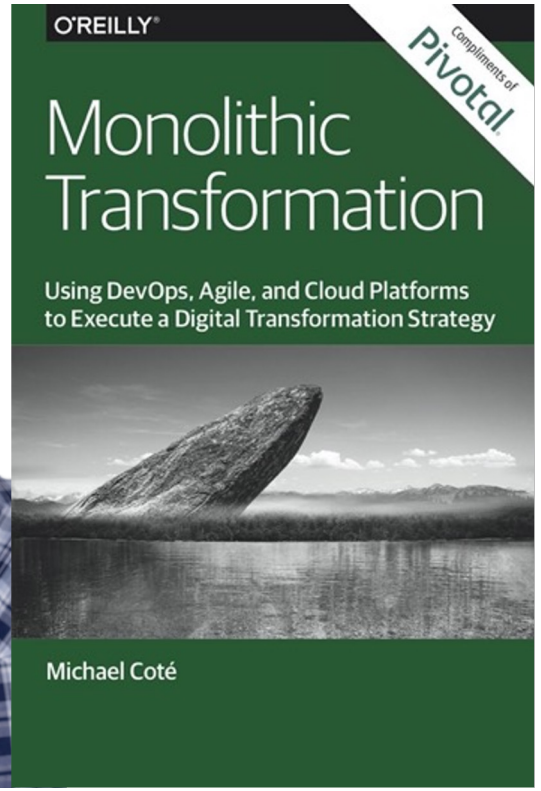
of enterprise apps
run on private cloud

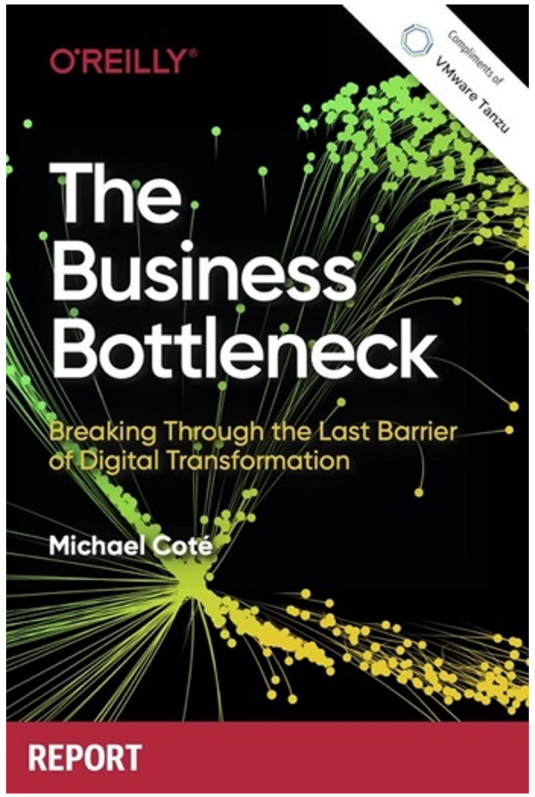
AI:

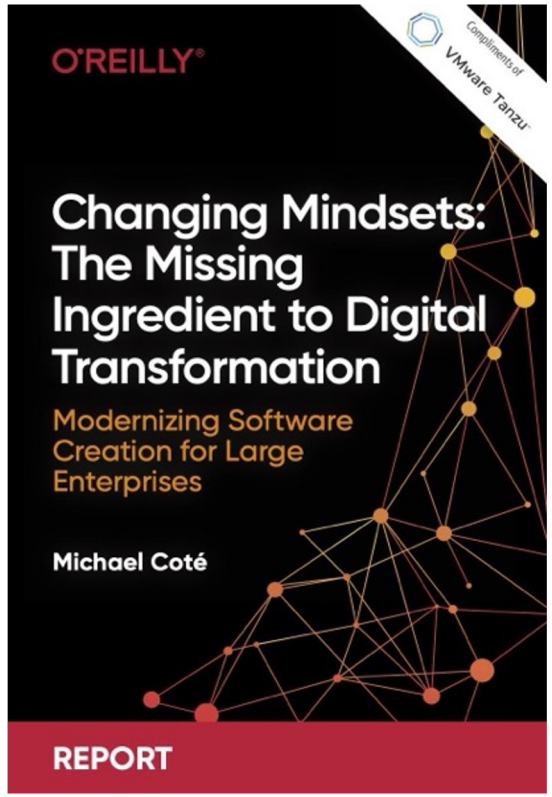
- Needs apps (maybe)
- Brings new platform customers

Coté

https://newsletter.cote.io/ | cote@broadcom.com







SOFTWARE DEFINED Interviews

SOFTWARE DEFINED TALK

RedMonk

DELL

451 Research S&P Global Market Intelligence

Pivotal

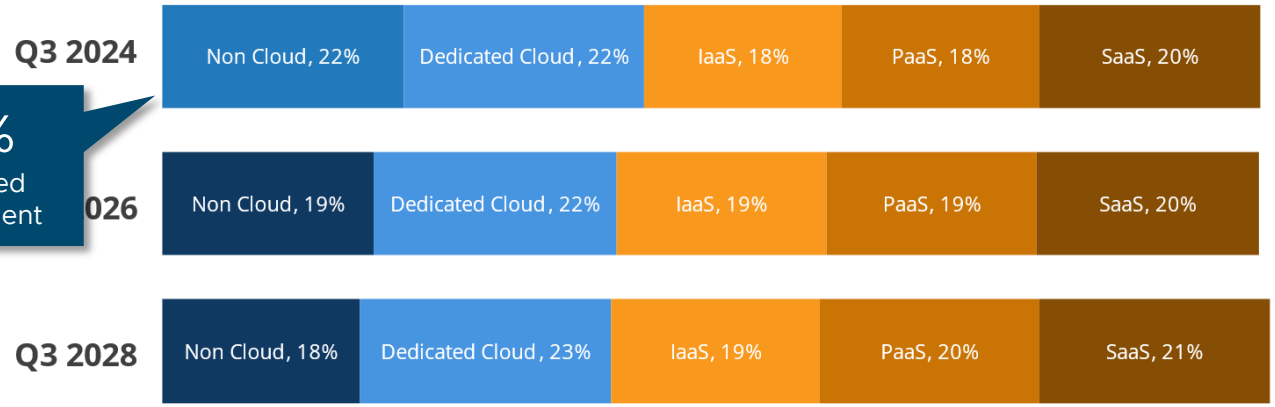
vmware

Tanzu by Broadcom

Where are the apps?

Cloud buyers increased their level of cross-platform interoperability in 2024. They also relied more upon providers for the management of their IT environments

Where are all your organizations applications deployed?



44% dedicated environment

Level of Interoperability Between Clouds

Public Cloud	**41%**	Hosted Dedicated Cloud
On Prem Dedicated Cloud	**39%**	Public Cloud
Hosted Dedicated Cloud	**38%**	On Prem Dedicated Cloud
Public Cloud	**38%**	Public Cloud

HOW THE EDGE IS MANAGED

IAAS EDGE

- 49% of IaaS deployments are considered as remote of edge deployments

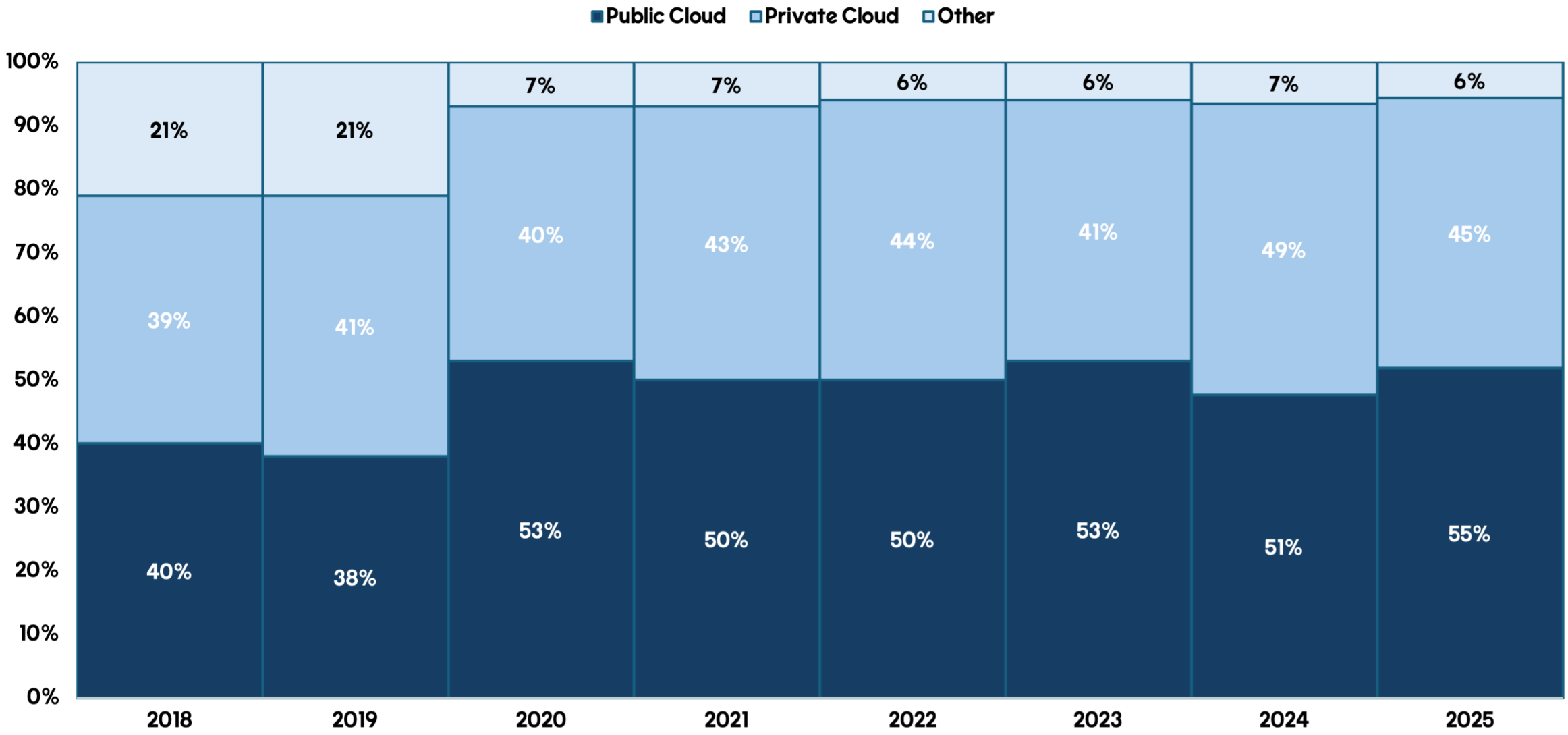
DEDICATED EDGE

- 48% of dedicated cloud environments are classified as remote or edge deployments.

N = 1,724, QC6 Source: QDV1 3Q24 Cloud Pulse Survey, October 2024, IDC

Source: IDC, IDC Cloud Pulse, 3Q24: Executive Summary – Vendor Perception, doc #US51134624, December 2024. n=1,724 IT decision makers, developers and LOB cloud influencers and decision makers; 35% tech companies, 25% 1k to 4.9k staff, 25% 5k+ staff.

Workload placement over the years (Flexera)



"Other" is "Non-cloud" in 2018 and 2019, "Additional workloads in public cloud in 12 months" in remaining years.

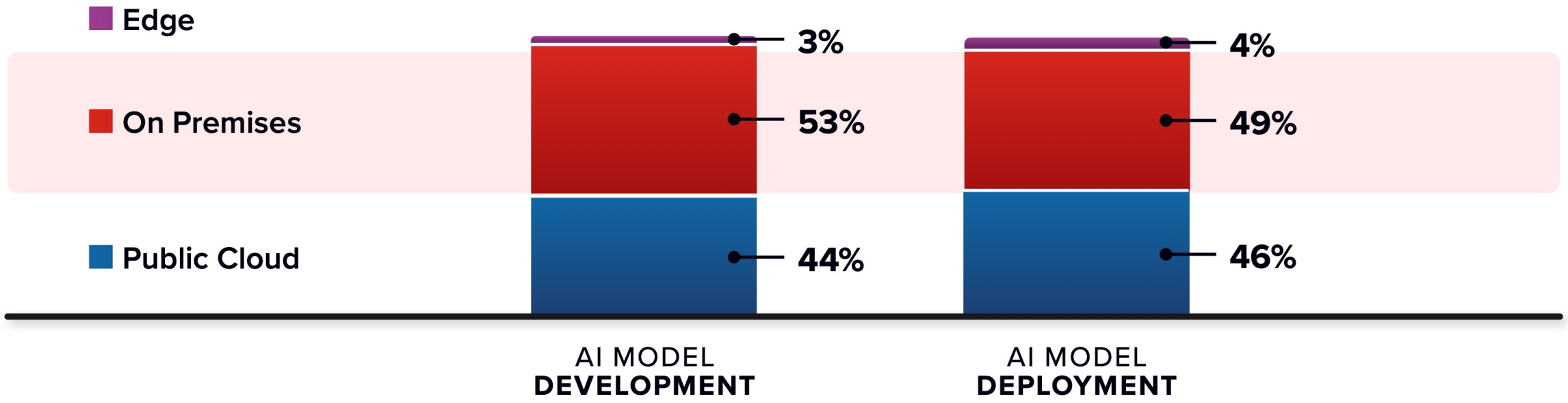
Sources: Flexera State of Cloud Report, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025.

FIGURE 5

Deployment Location for the Development and Deployment of AI Models

Where does your organization primarily develop and deploy AI models?

(Percentage of respondents)



Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey*, July 2024.
For an accessible version of the data in this figure, see Figure 5 Supplemental Data in Appendix 1.

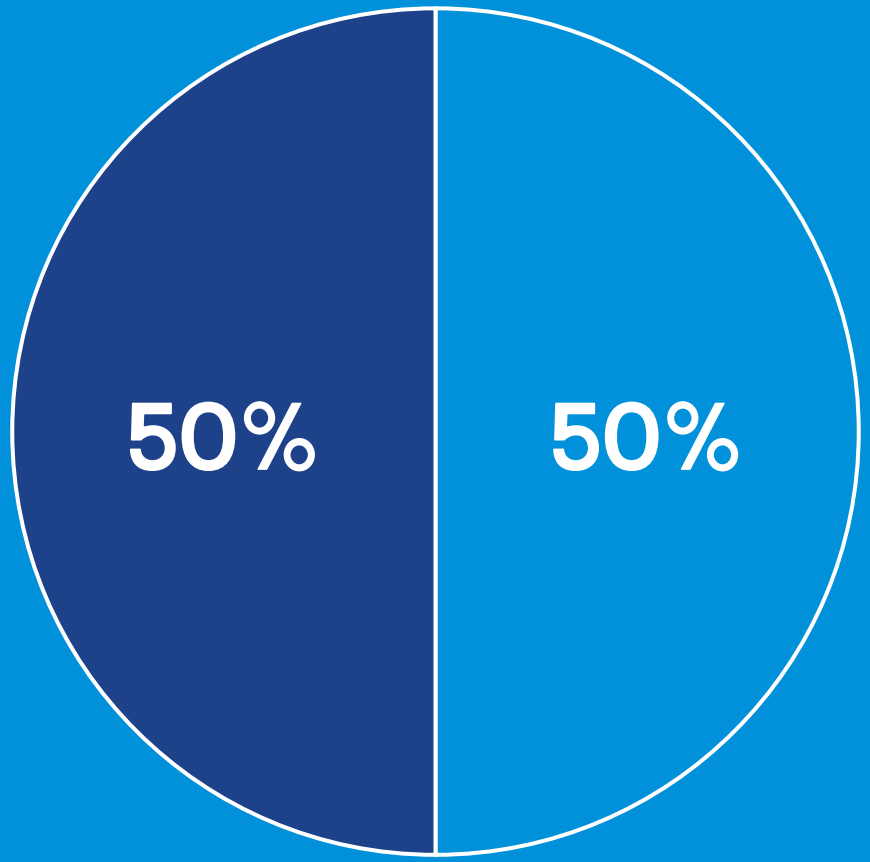
Source: IDC White Paper, sponsored by Broadcom, On-Premises AI Infrastructure Balances Innovation and Security, doc #US52747024 December, 2024. Conducted July, 2024, n=411.

🇺🇸

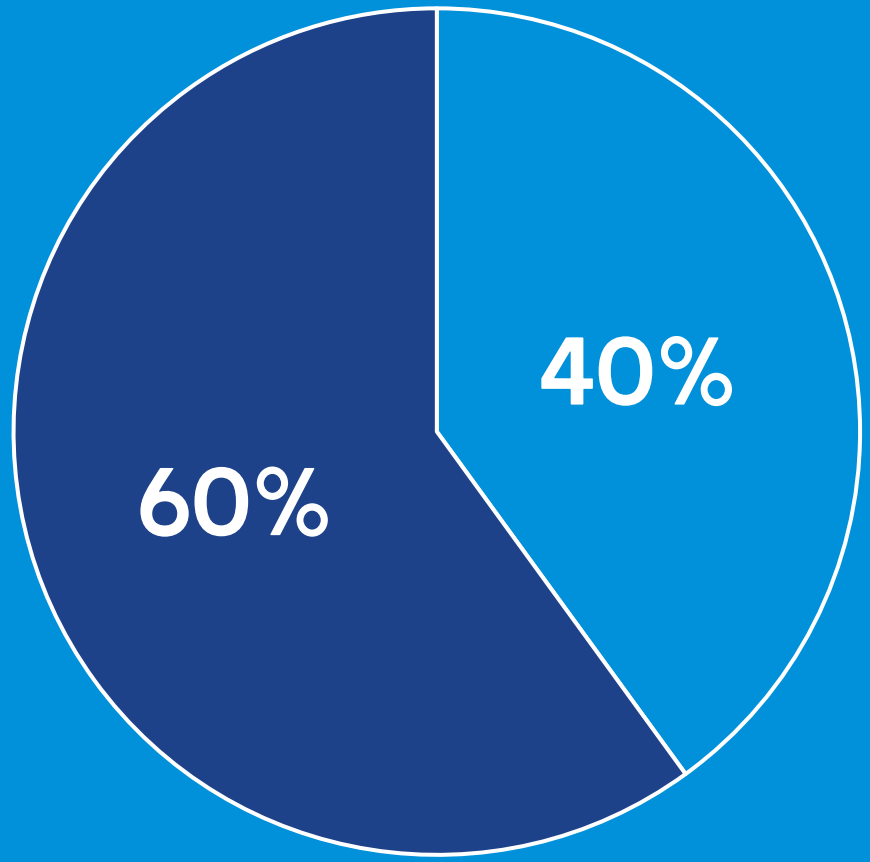
🇪🇺

Source: AI generated Trump Cloud; "The Cloud Man," found on Whoopah, does not appear to ben actual Rene Magritte painting.

Where the workloads live, rough estimates

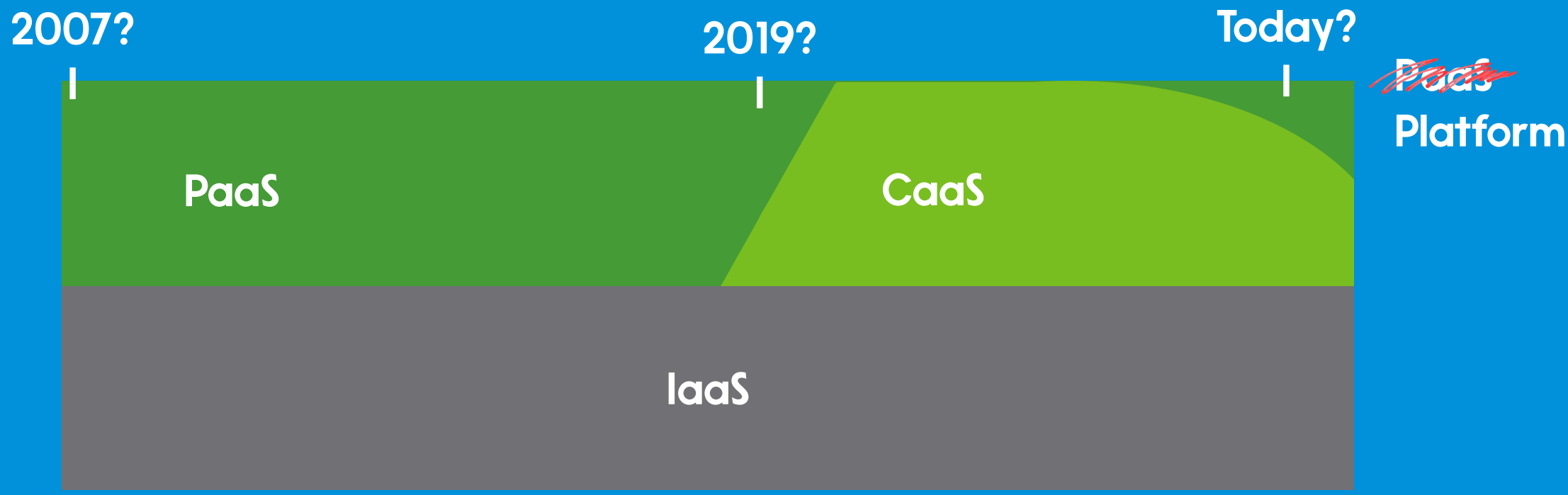


?



Source: me!

What is a platform?

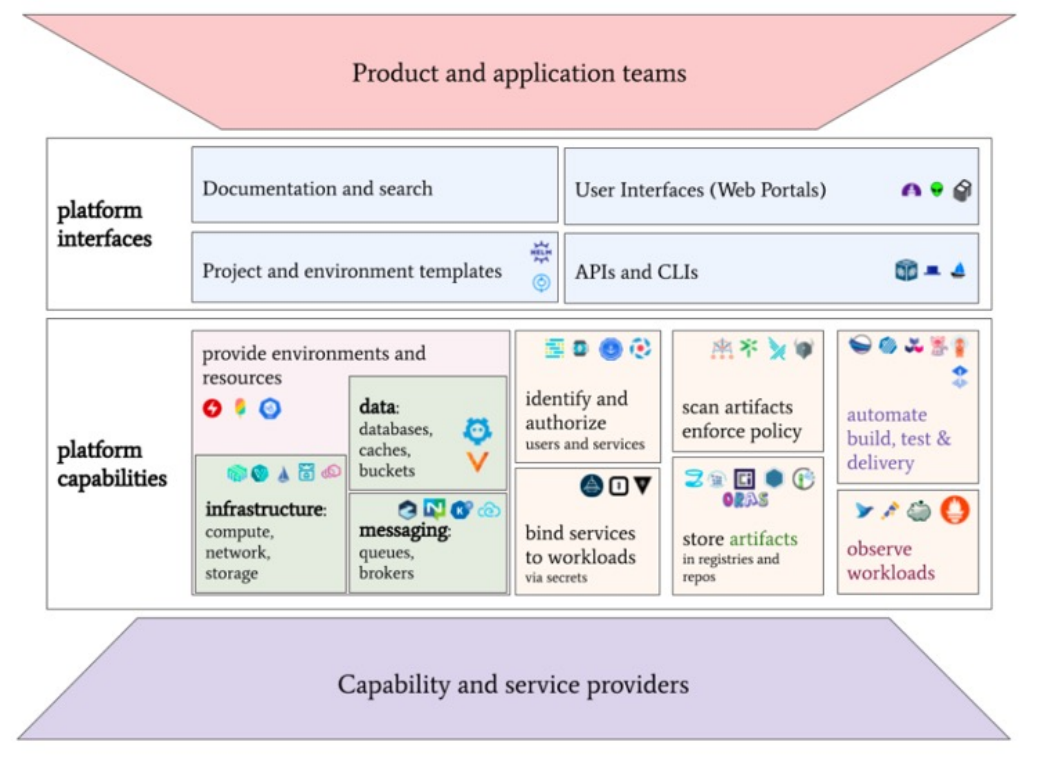

2007?
2019?
Today?
PaaS
Platform
PaaS
CaaS
IaaS

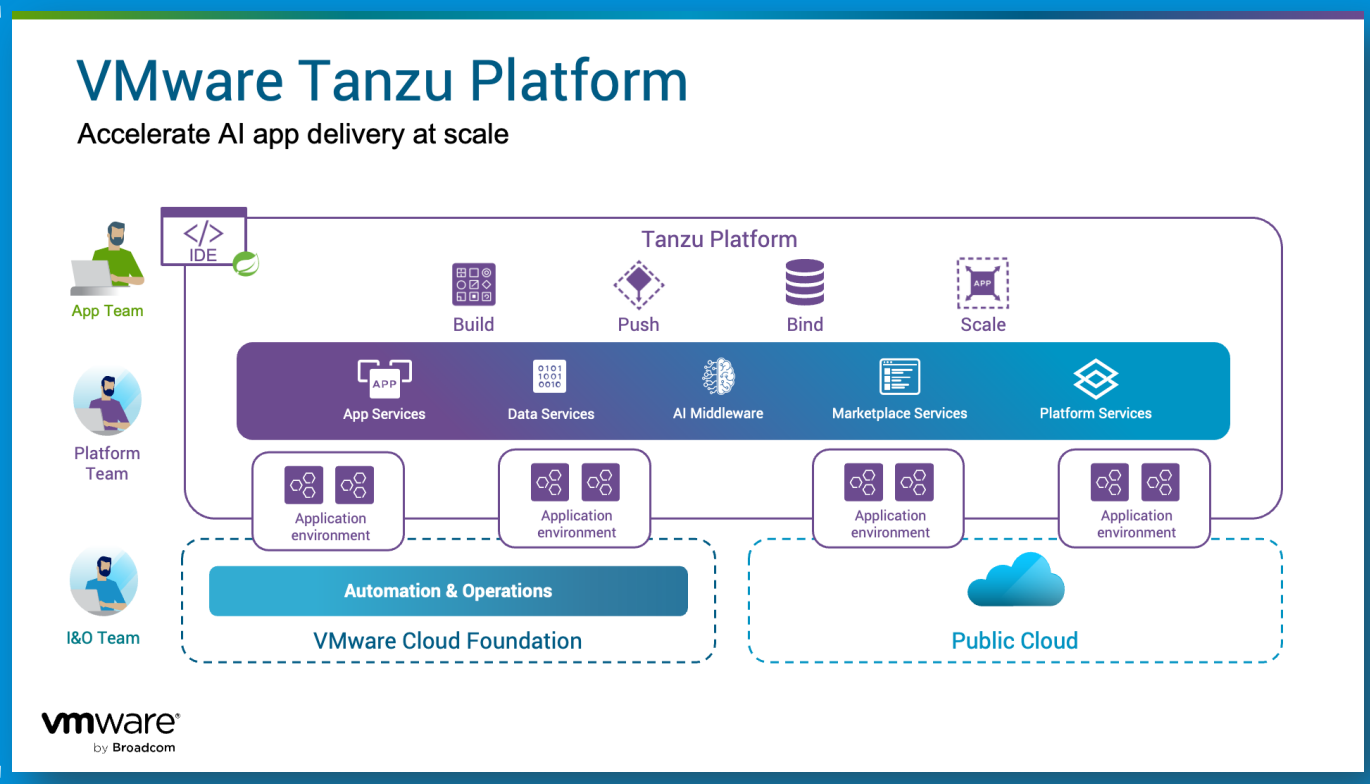
> A *digital platform* is a foundation of self-service APIs, tools, services, knowledge and support which are arranged as a compelling internal product.
>
> [SO THAT] Autonomous delivery teams can make use of the platform to deliver product features at a higher pace, with reduced co-ordination.

Evan Bottcher, March, 2018

What is a *platform*?

Centralized, standardized stack for building, running, and managing in-house apps.





Sources: "CNCF Platforms White Paper," March 2023; VMware Tanzu.

Developer productivity & all the ops -illities

DEVELOP → OPERATE → OPTIMIZE

App Teams

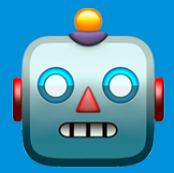
- Bring their own framework
- Simple command to push to production
- Frictionless data services & content
- Plug into any CI/CD
- Self-service access

Platform Teams

- Autoscaling & load balancing
- Flexibility to run on any Cloud
- Ability to curate services
- AI & platform quota management
- Zero downtime deployments

Platform Teams

- Rapidly respond to CVEs
- Lifecycle automation & patching
- Credential rotation
- Automated compliance
- Built-in Observability



Adding AI…?

A platform treats AI like any other service, adding AI middleware & focusing on new models & frameworks

DEVELOP → OPERATE → OPTIMIZE

App Teams

- Bring their own framework
- Simple command to push to production
- Frictionless data services & content
- Plug into any CI/CD
- AI-Ready dev framework
- AI Self-service model access
- AI Integrate with enterprise data

Platform/AI Teams

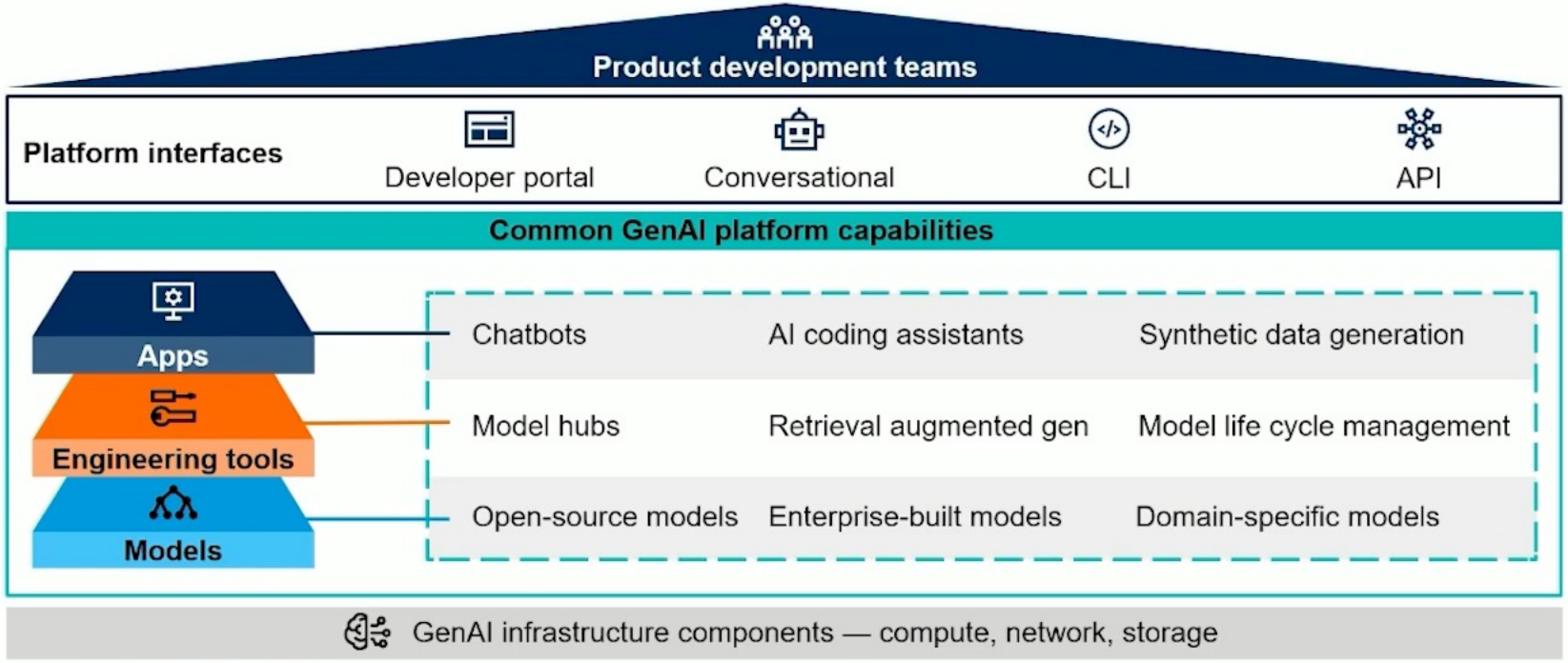
- Autoscaling & load balancing
- Flexibility to run on any Cloud
- Ability to curate services
- AI & platform quota management
- Zero downtime deployments
- AI Model running & updating
- AI Control costs & policy with guardrails

App, Platform/AI Teams

- Rapidly respond to CVEs
- Lifecycle automation & patching
- Credential rotation
- Automated compliance
- Built-in Observability
- AI Integrated model observability
- AI Continuous model curation

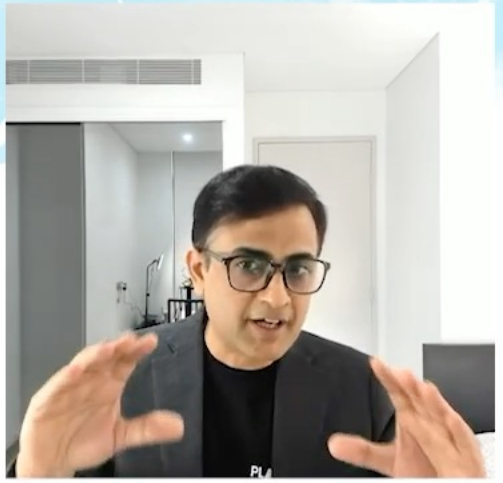
PLATFORM CON 25

Platform Teams Provide Internal Platform Services to Support Common GenAI Needs



[a]TRiSM — AI trust, risk and security management
Source: Gartner

Gartner



Manjunath Bhat

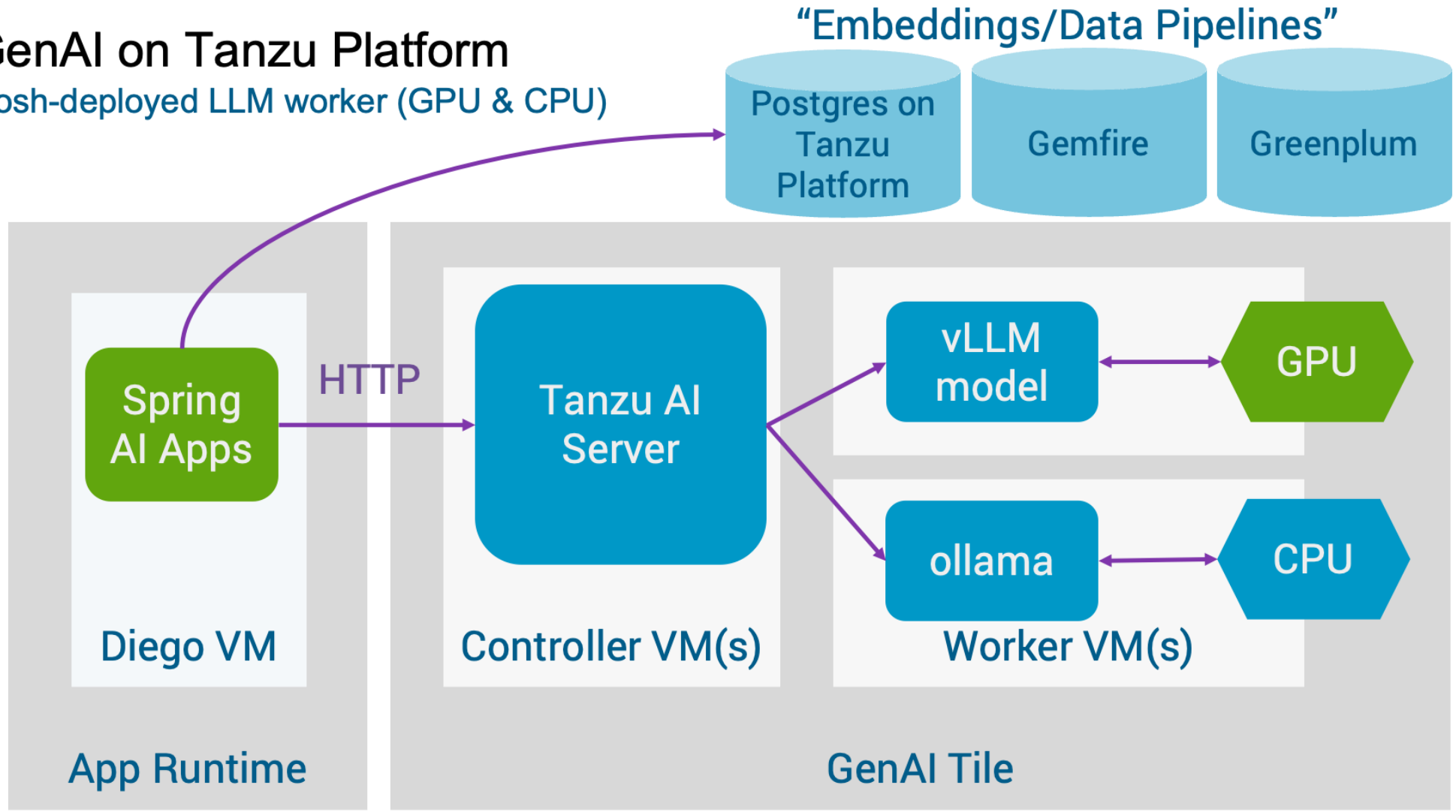
Research VP, Gartner

IMPACT

Source: "How platform teams can help scale generative AI application delivery," Manjunath Bhat, Gartner, PlatformCon 2025, June, 2025 For another stack, see also "Can Platform Engineering Accelerate AI Adoption?", Jennifer Riggins, The New Stack, August, 2025.

GenAI on Tanzu Platform

Bosh-deployed LLM worker (GPU & CPU)



Source: "Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform," Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.

How do you run a platform?

(in private cloud)

“We are building this platform not for us, we are building it for Mercedes-Benz developers.”

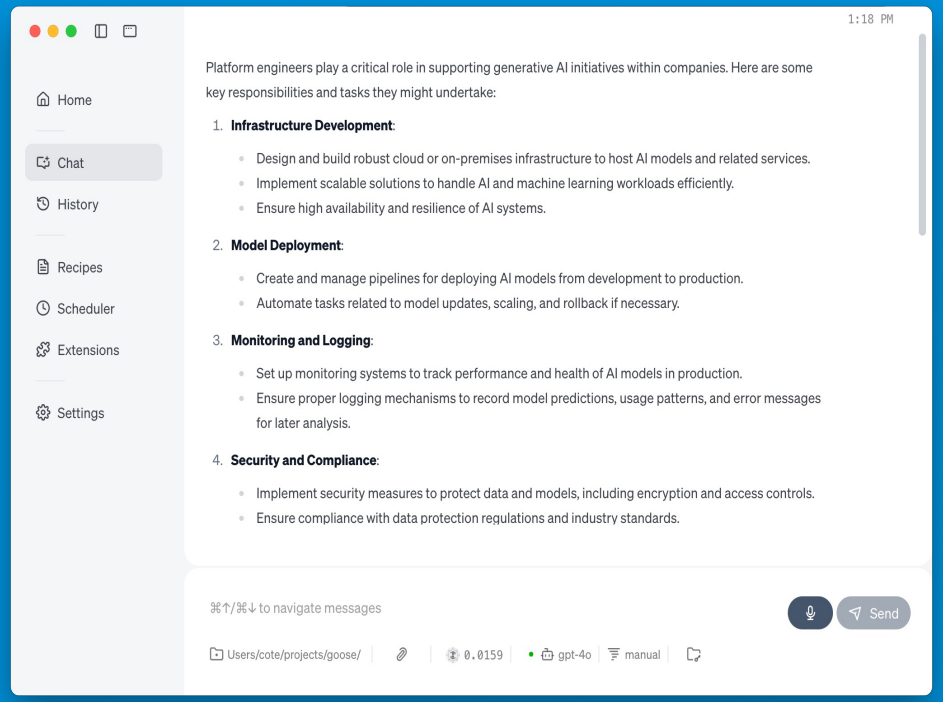
Thomas Müller, Mercedes-Benz

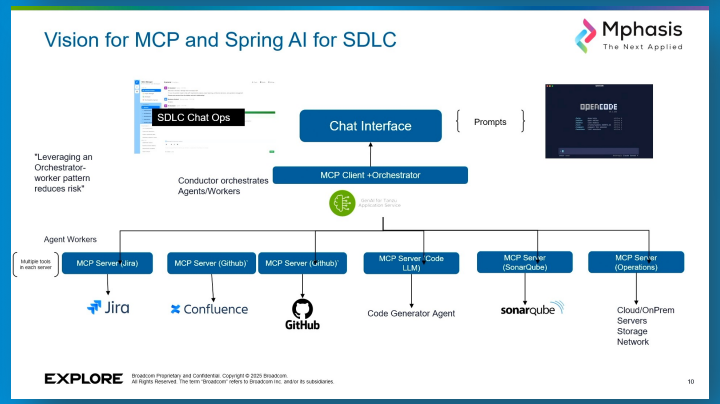
Find the Developer Toil, Confusion, Blockers

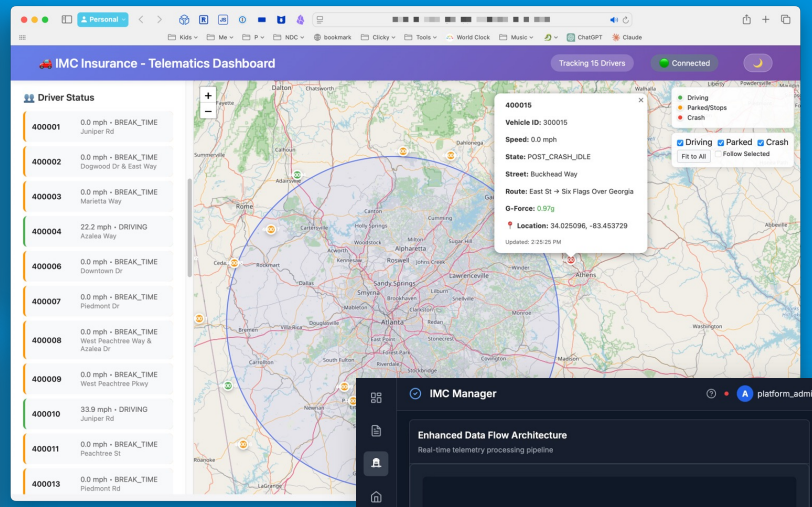
Find the Developer Toil, Confusion, Blockers

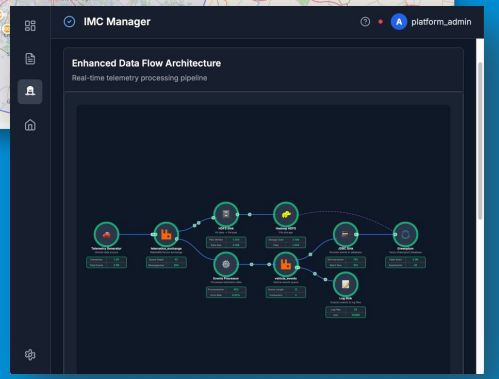
- What are we making?
- We have a strong vision for our product, and we're doing important work together every day to fulfill that vision.
- I have the context I need to confidently make changes while I'm working.
- I am proud of the work I have delivered so far for our product.
- I am learning things that I look forward to applying to future products.
- My workstation seems to disappear out from under me while I'm working.
- It's easy to get my workstation into the state I need to develop our product.
- What aspect of our workstation setup is painful?
- It's easy to run our software on my workstation while I'm developing it.
- I can boot our software up into the state I need with minimal effort.
- What aspect of running our software locally is painful? What could we do to make it less painful?
- It's easy to run our test suites and to author new ones.
- Tests are a stable, reliable, seamless part of my workflow.
- Test failures give me the feedback I need on the code I am writing.
- What aspect of production support is painful?
- We collaborate well with the teams whose software we integrate with.
- When necessary, it is within my power to request timely changes from other teams.
- I have the resources I need to test and code confidently against other teams' integration points.
- What aspect of integrating with other teams is painful?
- I'm rarely impacted by breaking changes from other tracks of work.
- We almost always catch broken tests and code before they're merged in.
- What aspect of committing changes is painful?
- Our release process (CI/CD) from source control to our story acceptance environment is fully automated.
- If the release process (CI/CD) fails, I'm confident something is truly wrong, and I know I'll be able to track down the problem.
- What aspect of our release process (CI/CD) is painful?
- Our team releases new versions of our software as often as the business needs us to.
- We are meeting our service-level agreements with a minimum of unplanned work.
- When something is wrong in production, we reproduce and solve the problem in a lower environment.

Sources: "Developer Toil: The Hidden Tech Debt," Susie Forbath, Tyson McNulty, and Coté, August, 2022. See also Michael Galloway's interview questions for platform product managers.









👨‍💼 Chat (for normals)…

- Your own ChatGPT.
- Customer service.
- Better search.
- Chat—as-UI

👩‍💻 Programming…

- New & old code.
- SDLC juicing.
- Trad'l "data science."
- Making pptx?

💾 AI in apps…

- Sales assistants.
- Sloppy integration.
- Science-ing.
- ???

Note: audio, pictures, video are omitted.

Sources: Tanzu customers; AI at Goldman, *FT*, September 14th, 2025; "Leverage Generative AI to Streamline the Software Development Lifecycle," Banu Parasuraman, Andrew Berenato, Explore 2025, August, 2025.

Run and customize your PaaS, don't build it

350 apps supported by 7 platform engineers

300 apps supported by 8 platform engineers

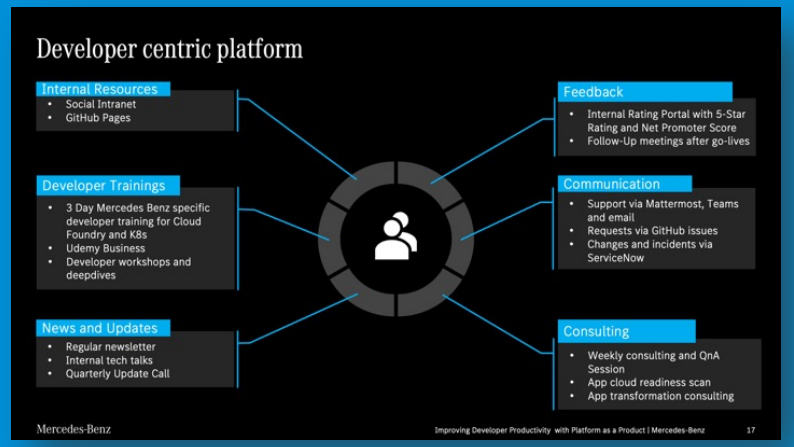
1,200 developers supported by 6 platform engineers

2,500 developers supported by 5 platform engineers

45 app teams supported by 1 ops team

Sources: Kroger, GAIC, Mercedes-Benz, two anonymized large enterprises. See also ESG ROI study.

Platform marketing



Sources: ING, 2023;BT Canvas team; MB.io; Duke Energy; Allstate; "Take DevOps to 11 and Sprinkle Cloud on it with Rainbows and Unicorns," Matt Curry, s1p 2017. "Improve Developer Productivity with Platform as a Product," VMware Explore, Nov. 2022; Kessel Run Wikipedia page (circa Feb 2025); platform marketing whitepaper, Coté, 2025.

“What have you done for me lately?”

Tales of ROI, or,

Metrics == Money

Picture: Know Your Meme.

Metrics for the LINE OF BUSINESS

Indicators

Speed

Velocity is a vector comprised of speed and direction.

We bring a raw speed of advantage to the LOBs and also enable them to rapidly and reliably respond to changes in direction in the service of the business based on user feedback loops.

MEASUREMENTS

- Time to value (cycle time)
- Frequence of customer feedback
- Time between bug identification and fix
- Time from feedback to deployment of change
- Customer satisfaction (NPS)
- Business satisfaction

Stability

Reality is a complex landscape of changing priorities, emergent bugs, evolving architectures, and staffing changes.

We help the LOB achieve resiliency and low volatility as they deliver customer value in the face of this complex reality.

MEASUREMENTS

- Volatility (std dev in velocity / mean velocity)
- # of defects generated per developer - year
- % of software launches / upgrades delayed due to defects
- Employee satisfaction (ENPS)

Scalability

LOBs need to scale across two dimensions:

People - LOBs strive to attract developers and ramp productivity linearly with personnel.

Apps - LOBs need to rapidly scale their applications and their complexity to handle demand.

MEASUREMENTS

- # of products in development
- # of products measuring business success
- Investment ratios: spend developing software vs operating and systems
- Disruption caused by doubling workload
- Ability to attract and retain talent (# of internal referrals)

Security

To move rapidly the team needs to feel secure in making code changes aggressively. Automated test coverage provides this safety net.

To rapidly search for customer value LOBs must adopt a learning culture that fosters psychological safety necessary to fail and learn from failure.

MEASUREMENTS

- % teams using CI
- % teams doing TDD
- Time from commit to deployment

Savings

Teams must reduce risk and waste through small batch delivery and fast consumer feedback.

This drives significant savings as use of the product grows and is key to maintaining their trust and enabling them to go fast, forever.

MEASUREMENTS

- Fraction of developer time spend writing code and delivering value
- Product:dev ratio
- Business satisfaction
- # of go/no-go decisions based on business success

Sources: Tanzu customer interviews and consulting, ~2015 to 2025; "Common platform team mistakes and how to avoid them," Camille Fournier, PlatformCon 2025, July, 2025.

Metrics for the IT

Indicators

Speed

IT can efficiently upgrade, patch, and manage the platform.

They rapidly onboard new application teams and provide the necessary services to quickly unblock teams and enable them to deliver consumer value.

MEASUREMENTS

- # prod/dev deploys per month
- # platform upgrades per month
- Platform upgrade speed
- # of new apps onboarded/month
- Team distribution of skills

Stability

Our customers entrust us with their production workloads and their developer productivity.

We must provide adequate SLOs to meet their needs and earn their trust by ensuring compatibility and uptime across platform upgrades.

MEASUREMENTS

- Minutes of prod outage per year
- Minutes of dev outage per year
- Mean time to recovery
- Mean time between failures
- # of upgrade-related failures

Scalability

IT needs to provide an "at-scale" service on-demand at the whim of the business.

They need to explore all options with minimal friction as they grapple with the mix of workloads on-premise and in the cloud.

MEASUREMENTS

- Queries per second
- # of AIs per foundation
- # of SIs per foundation
- # of foundations
- # of teams using the platform
- Does increasing workload on existing

Security

Security is a paramount concern for our customers. We earn their trust by providing a platform that is secure by default.

We solve for security and reduce security-related friction and toil in order to enable our customers to go fast, forever.

MEASUREMENTS

- Time between identifying and patching a CVE
- Cost in person-hours or dollars of leaked credential
- Fraction of operator time spent on security configuration
- # of disruptions/suspensions due to security concerns

Savings

IT must meet the needs of thousands of developers within tight budgetary constraints.

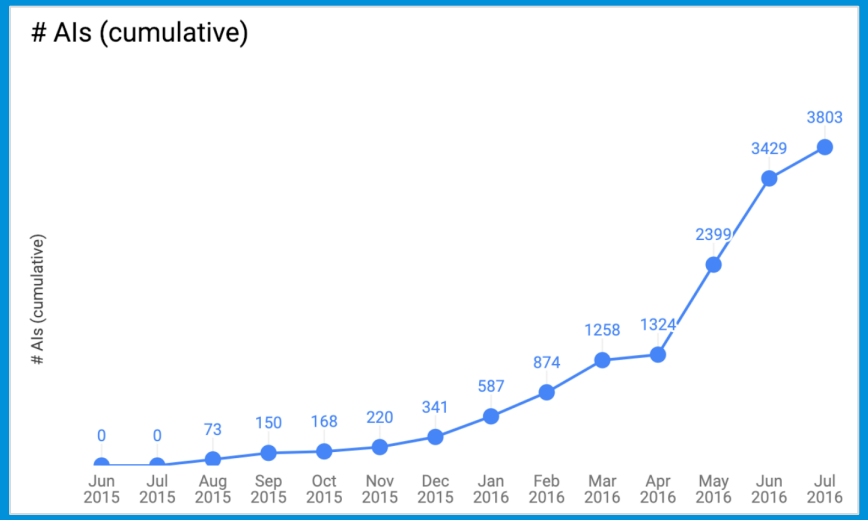
We provide a platform that simultaneously reduces complexity and sprawl and improves the ops:dev ratio.

MEASUREMENTS

- Operator:developer ratio
- # of apps per operator
- # of foundations per operator
- Degree of automation for provisioning, build, test, change approval governance, deployment, perf

Sources: Tanzu customer interviews and consulting, ~2015 to 2025.

Scaling Phase – Pairing & Seeding to build trust & training



1. Create platform marketing program.
2. Find two to five more apps.
3. Pair & seed from first dev & platform team to new teams.
4. ”Shift Left” - build golden paths for governance, security, etc.
5. Add more infrastructure staff with pairing & seeding.
6. Do this for three months.
7. Repeat, growing number of apps as pairing & seeding allows.

Sources: “From 0 to 1000 Apps: The First Year of Cloud Foundry at The Home Depot,” Anthony McCulley, The Home Depot, Aug 2016; “Cloud Native at The Home Depot, with Tony McCulley,” Pivotal Conversations #45; USAF presentations and write-ups; ”Driving Business Agility Without Large-Scale Transformation Programs,” Venkatesh Arunachalam, Sep 2021; The Home Depot 2022[?]Q4 earnings call; *The Business Bottleneck*, Coté.

Running a platform: Platform Engineering Maturity Model

Aspect		Provisional	Operational	Scalable	Optimizing
Investment	*How are staff and funds allocated to platform capabilities?*	Voluntary or temporary	Dedicated team	As product	Enabled ecosystem
Adoption	*Why and how do users discover and use internal platforms and platform capabilities?*	Erratic	Extrinsic push	Intrinsic pull	Participatory
Interfaces	*How do users interact with and consume platform capabilities?*	Custom processes	Standard tooling	Self-service solutions	Integrated services
Operations	*How are platforms and their capabilities planned, prioritized, developed and maintained?*	By request	Centrally tracked	Centrally enabled	Managed services
Measurement	*What is the process for gathering and incorporating feedback and learning?*	Ad hoc	Consistent collection	Insights	Quantitative and qualitative

Source: "Platform Engineering Maturity Model," CNCF Platforms Working Group, October, 2023.

Getting ready for a long journey

Questions, topics, & analysis

1. Expect 3 to 5 years: "Small organizations might swim through level three, while large enterprise organizations with lots of heritage and legacy, might be here for years."
2. Do you have and/or need microservices and 12 factor apps? How many new apps do you have?
3. How will you modernize your existing apps? Would lift-and-shift or leave alone be better?
4. How will you standardize & centralize infrastructure, [dev|ops|security|compliance] practices?
5. How can you make continuous learning stick?
6. How will you change the organization and team definition, structures, and incentives?
7. How will you get The Business to work weekly with the developers?
8. Costs do not decrease until you decommission old infrastructure and legacy apps.

Get an AI App Up and Running in Under 2 Weeks

What is it:

Developed by innovators of Cloud Foundry, the AI Starter Kit provides a seamless path to your first AI application on Tanzu Platform.

Who is it for:

Platform engineers, application developers and data teams can utilize AI Starter Kit to develop patterns for scaling AI apps.

Why it's needed:

- Take AI out of the shadows and support internal innovation with minimal friction
- De-risk AI exploration and prepare for scaled AI app rollout
- Ensure your models are ready for primetime with performance, accuracy and ROI

TryTanzu.ai

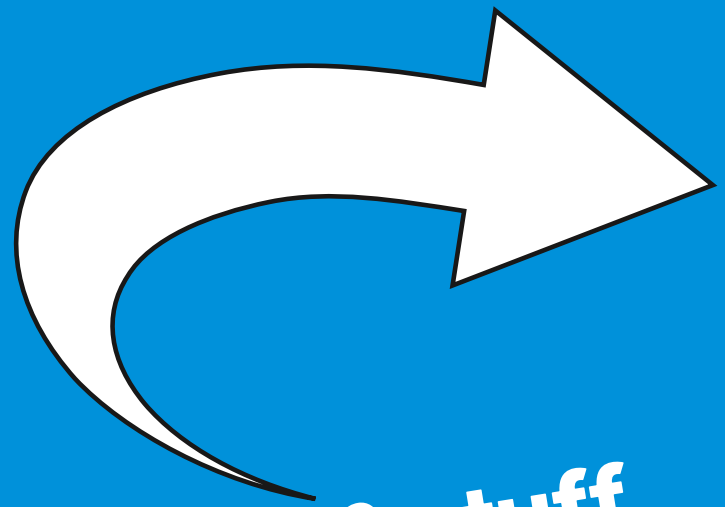
Expected results:

- Frictionless AI Prototyping
- Forecast AI needs more accurately
- Identify optimal delivery strategies
- Optimize production performance
- Reduce GPU dependency

Thanks!

📨 https://newsletter.cote.io/

🏢 cote@broadcom.com



Slides & stuff