Generative AI: Building and Scaling Adobe Firefly

Alexandru Costin, VP Firefly

March, 2024

Executive summary

Embracing Generative AI is essential for any business.

Adobe moved fast building on existing infrastructure.

Firefly – Adobe foundational models for creativity.

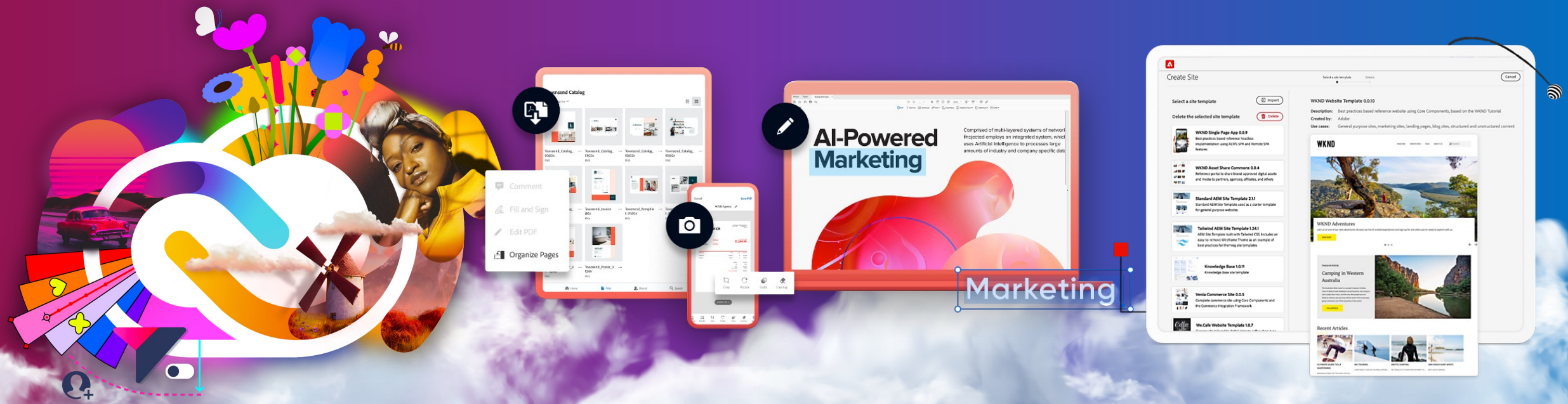
Powering Adobe apps at scale with Generative AI: Training, Data, and Inference

Changing the world through *personalized* digital experiences

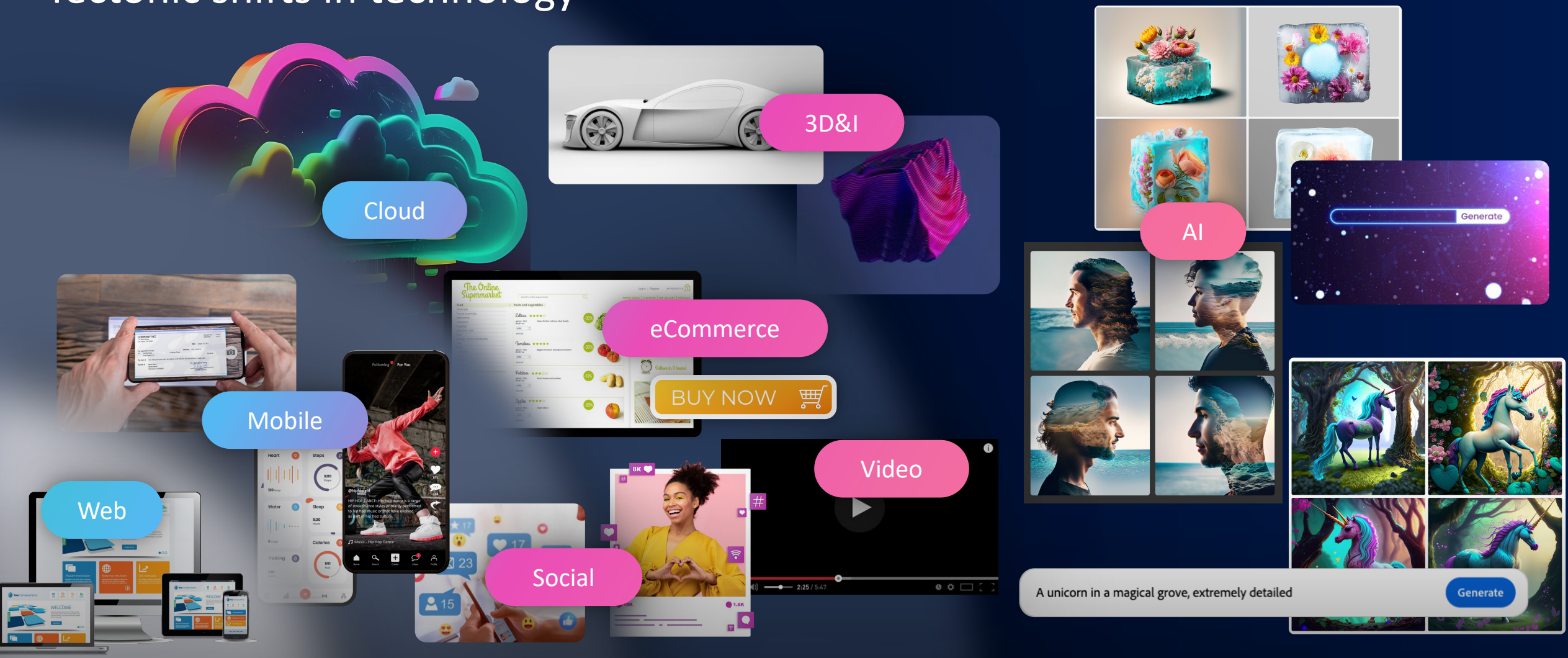
Unleashing creativity

Accelerating document productivity

Powering digital businesses

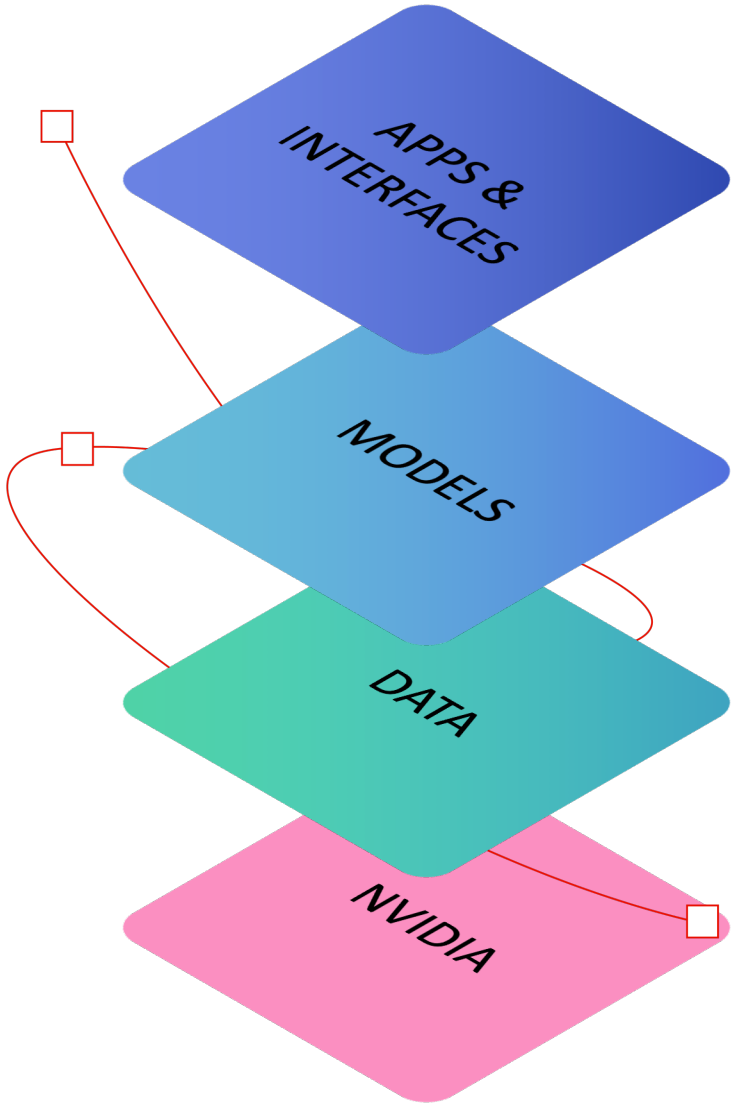


Tectonic shifts in technology



Internet era | Mobile era | Social era | AI era

Adobe’s unique approach to the AI stack



Transform industry-standard creative workflows; Integrate in existing commercial product workflows

Build models designed to be commercially safe and capitalize on Adobe Research

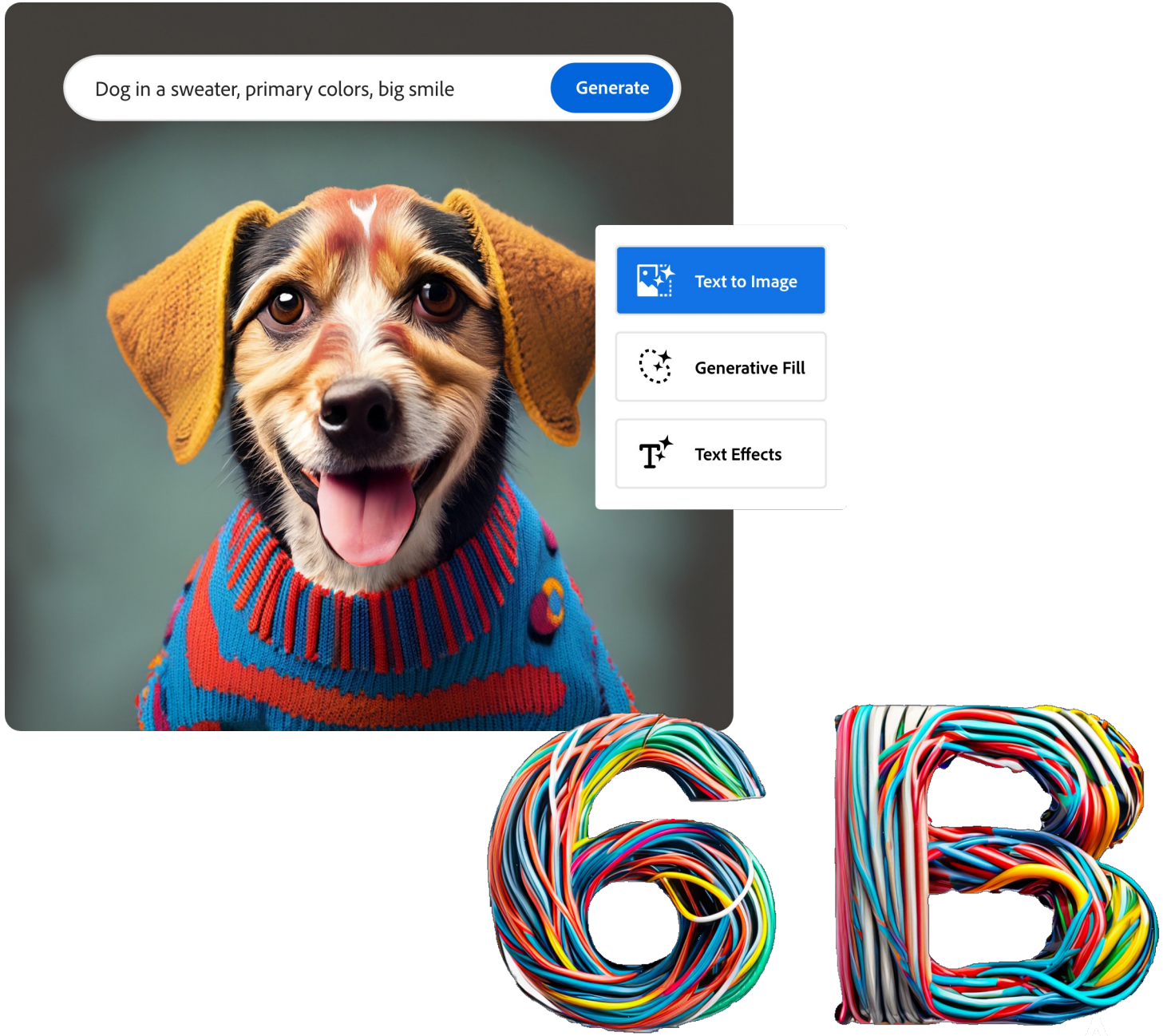
Harness Adobe data for individual and enterprise

Partner with silicon providers to scale training

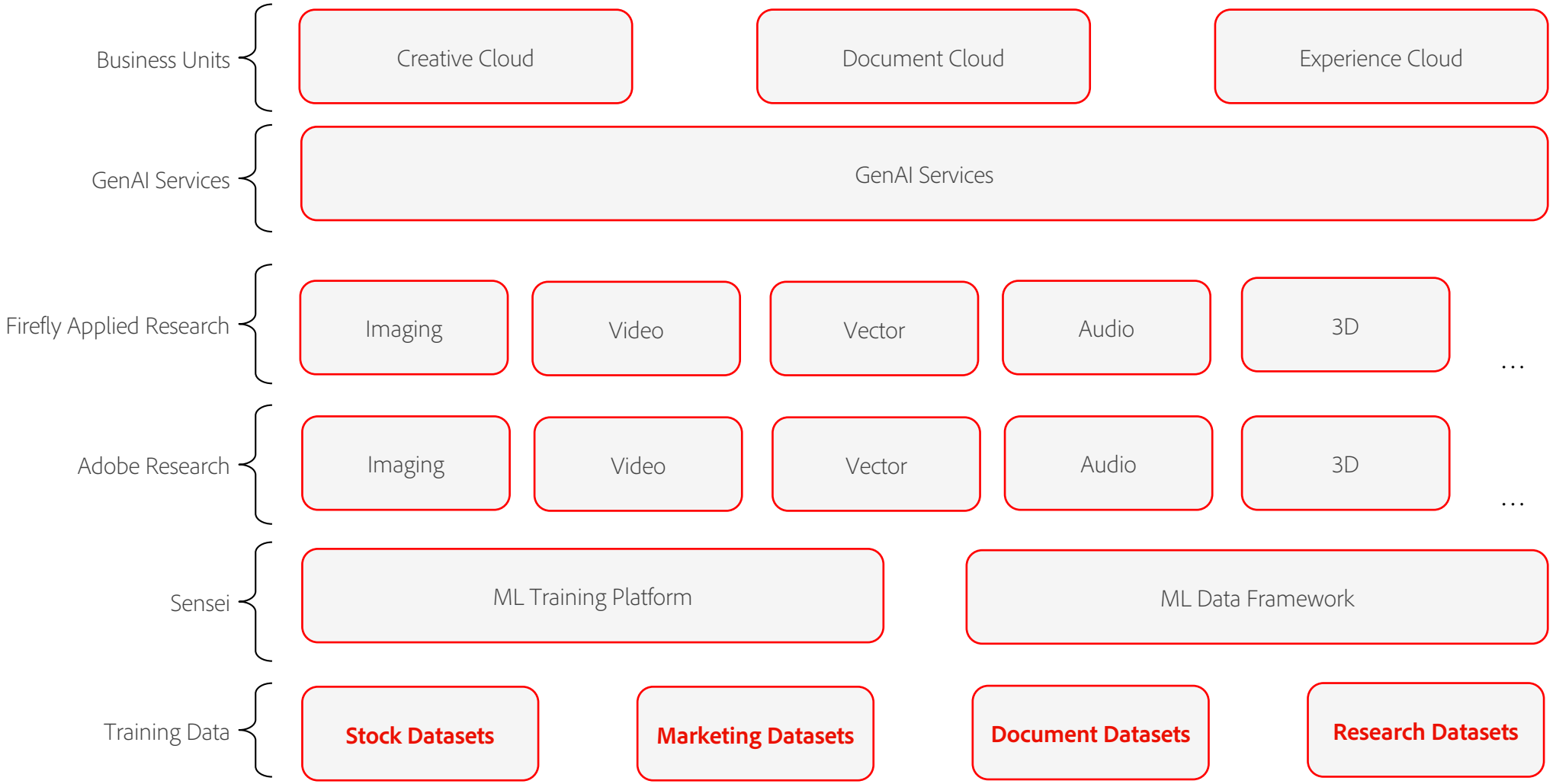
Adobe Firefly

Over 6B images generated with Adobe Firefly

- World's most advanced, popular AI image generation model, safe for commercial use; Ps Generative Fill most used Ps feature
- Natively Integrated directly into creative workflows
- Professional-quality content, trained on Adobe Stock assets and openly licensed, public domain content
- Content credentials auto-attached, new official "icon of transparency"



Structured to Innovate: AI Superhighway



The story of Firefly

March 21
Adobe Firefly & Sensei GenAI

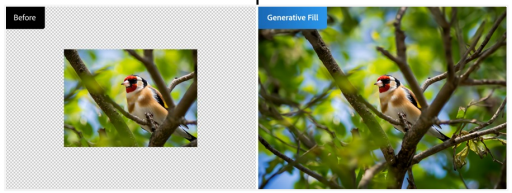
May 23
Photoshop with Generative Fill

June 8
Generative AI for Enterprises
(Firefly & Express for Enterprise; Sensei GenAI services)

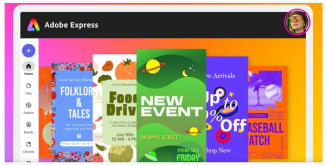
June 13
Illustrator with Generative Recolor

July 12
Firefly Expands Globally supporting text prompts in 100+ languages

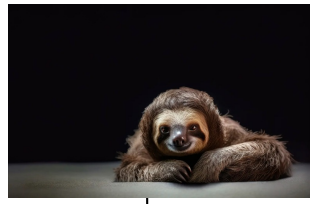
July 27
Photoshop with Generative Expand



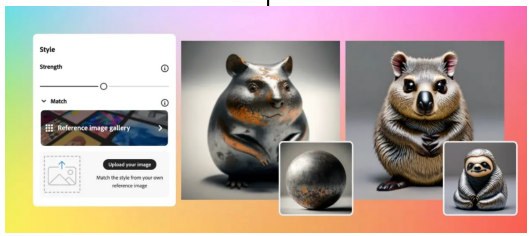
Aug 16
Major Adobe Express Update with Firefly Beta Capabilities



Oct 10
Adobe Image 2

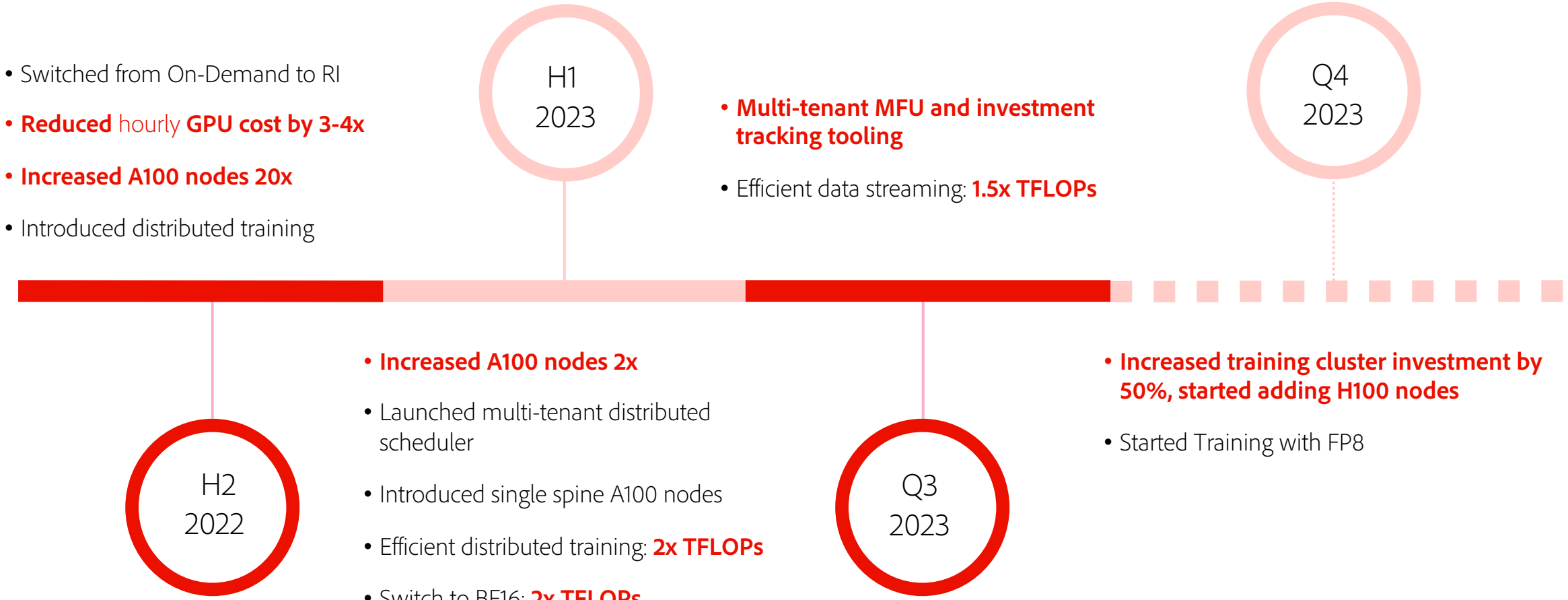


Oct 10
Generative Match
Text to Vector
Text to Template

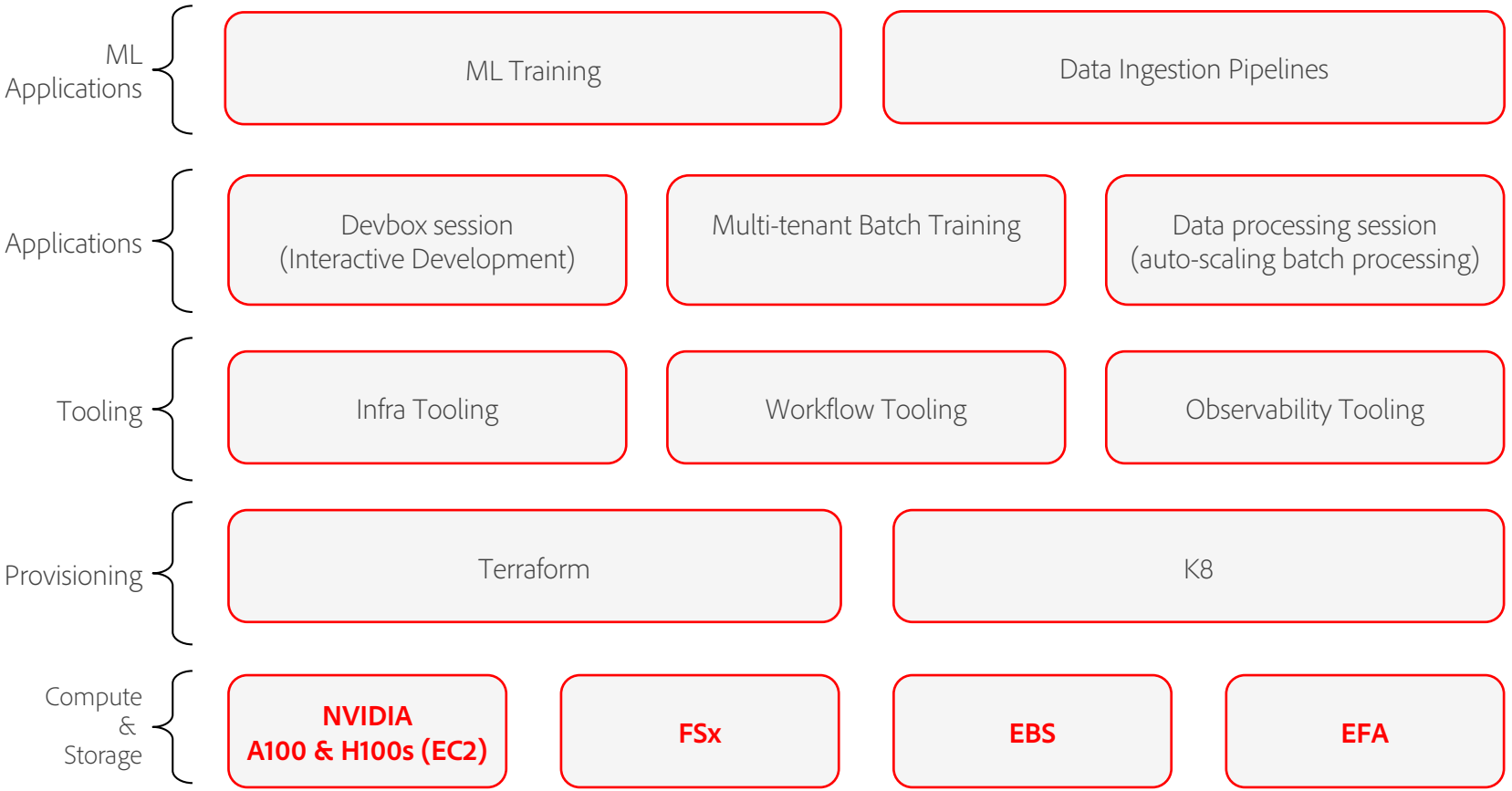


2023

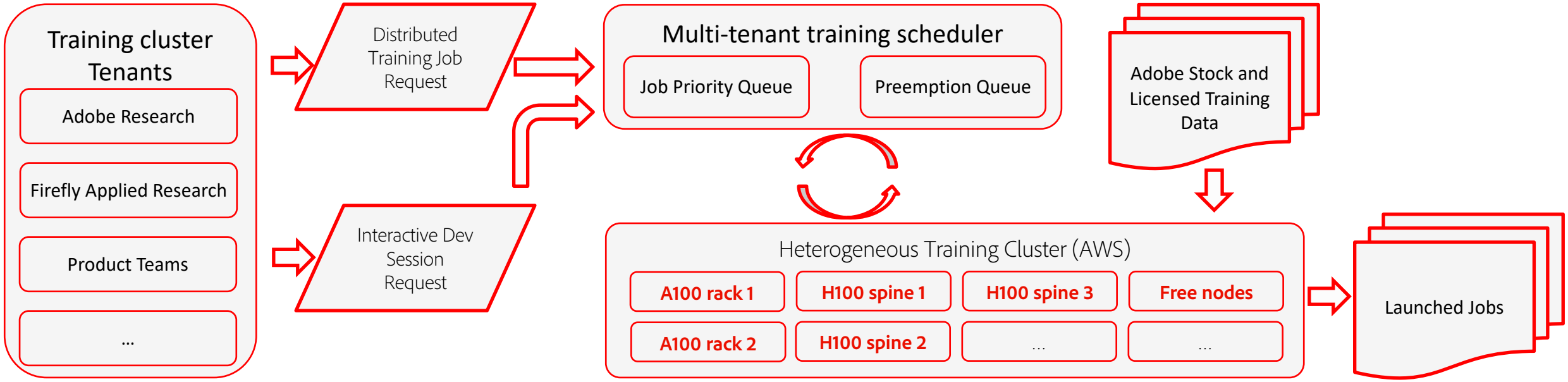
ML Training Investments



Sensei Training



Training

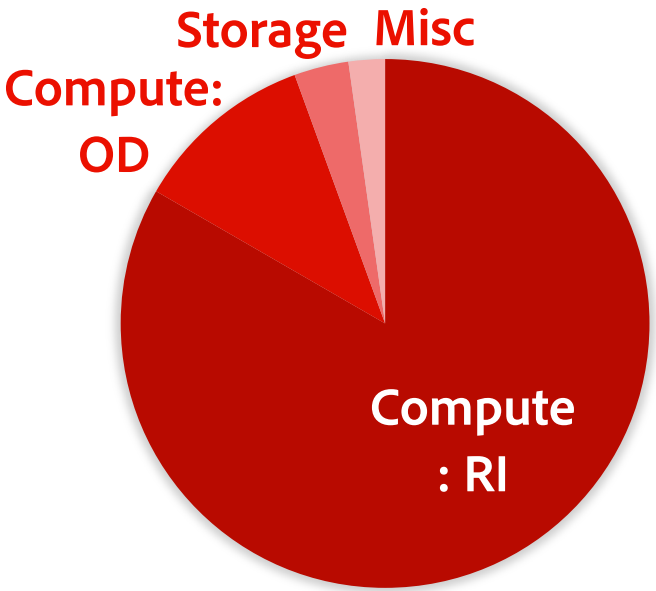


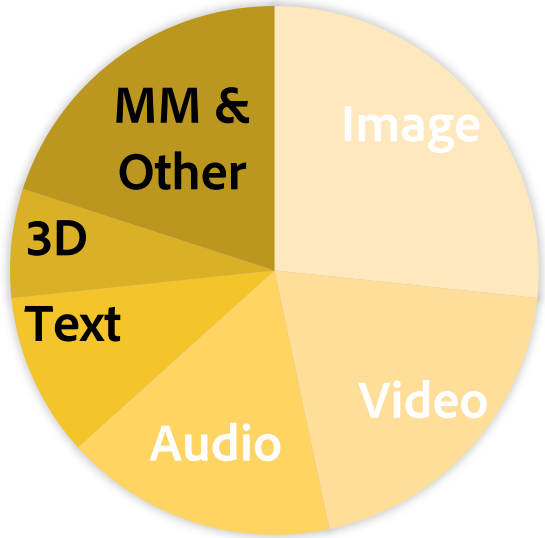
- Hundreds of distributed training runs in parallel
- Team and project quotas
- Over-quota scheduling capability
- Usage policy, heterogeneous policy governance

ML Training Analytics

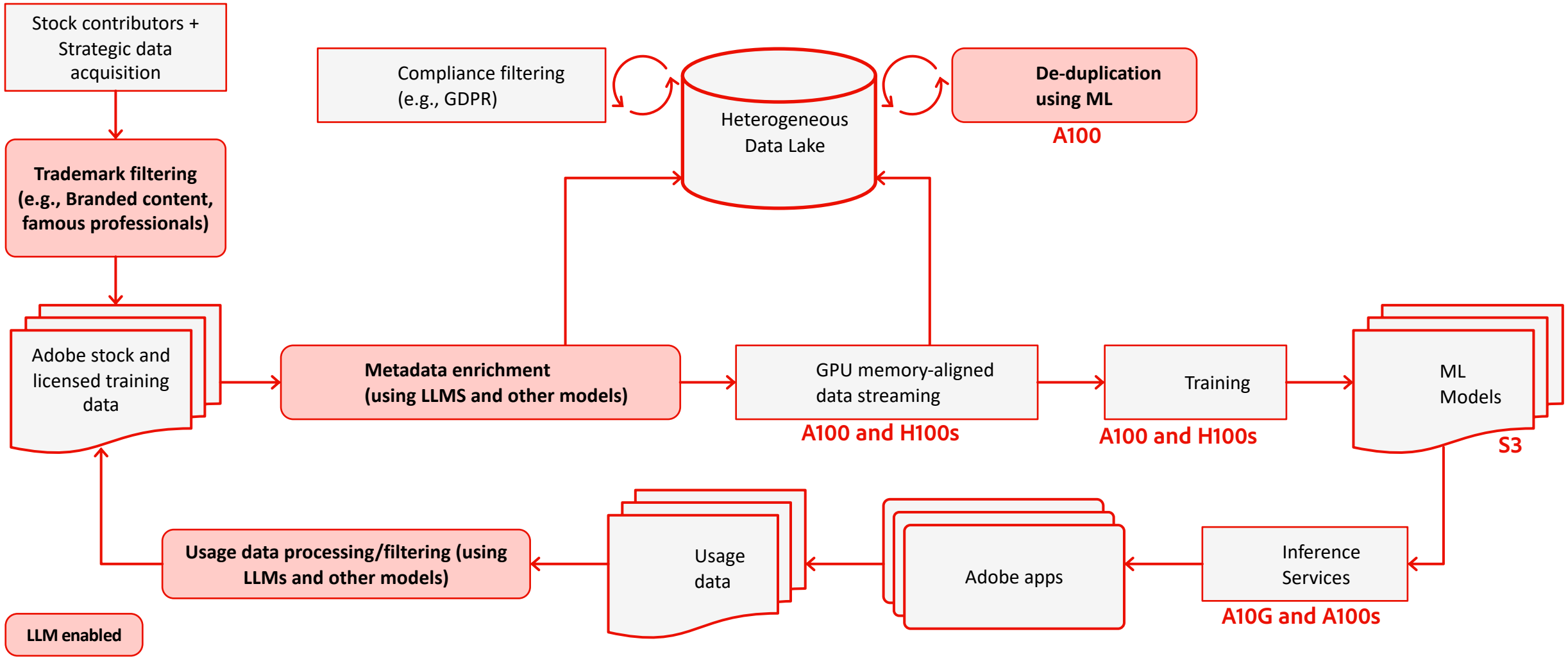
- MFU monitoring
- Usage and cost per project
- Strategy <> Investment alignment
- Early warning for unoptimized code

Project	Utilization	SM Activity	SM Occupancy	TFLOPs
Experiment #100	81%	74%	22%	103
Experiment #101	96%	51%	36%	85
Experiment #102	79%	27%	30%	42
Experiment #103	77%	35%	12%	39
…	…	…	…	…





Responsible AI data lifecycle



Key insight: Data size in generative AI workflow

Training data size

5PB

raw data

2PB

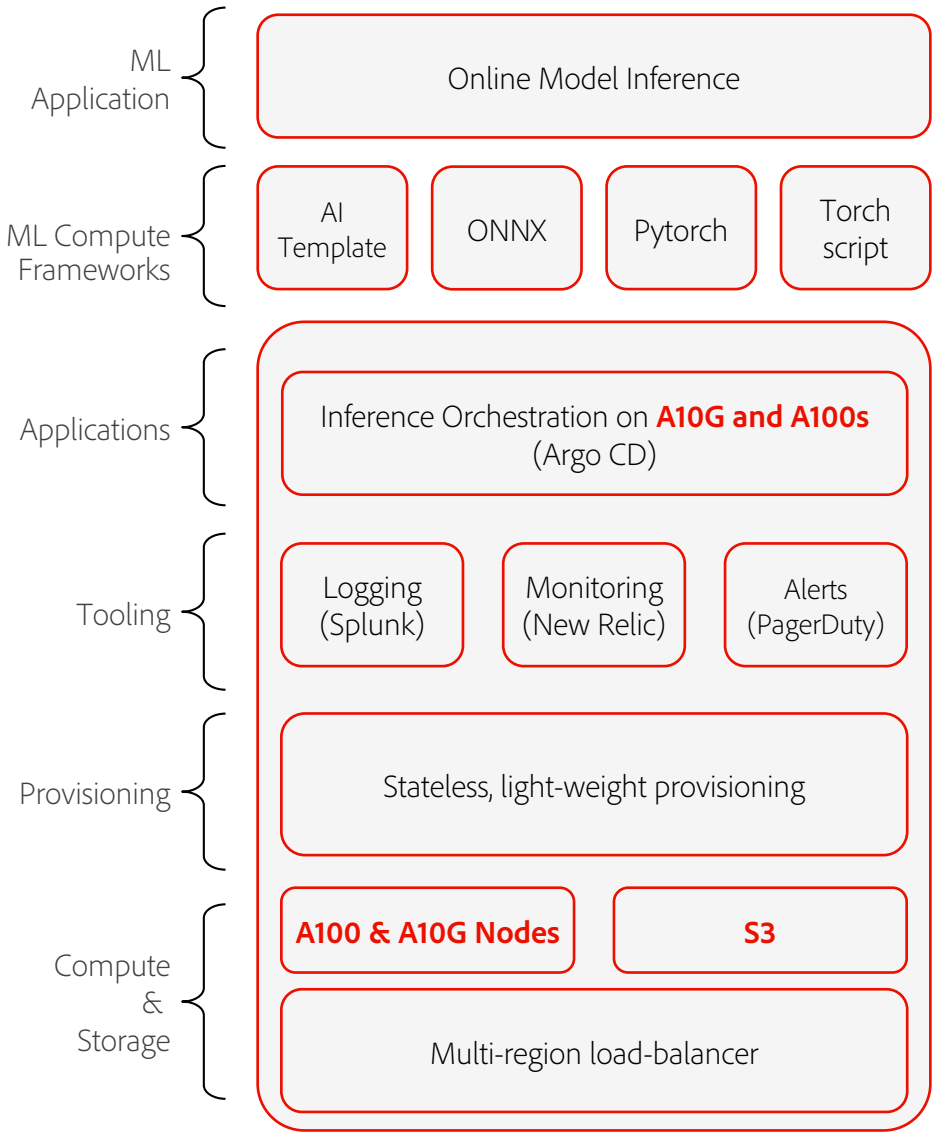
embeddings

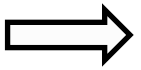
Fine-tuning data size

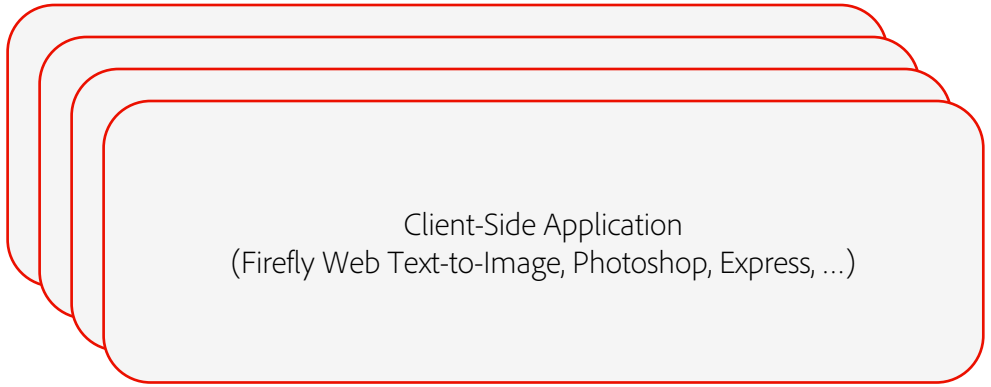
0.1-3 PB

depending on model/usecase

Firefly Inference





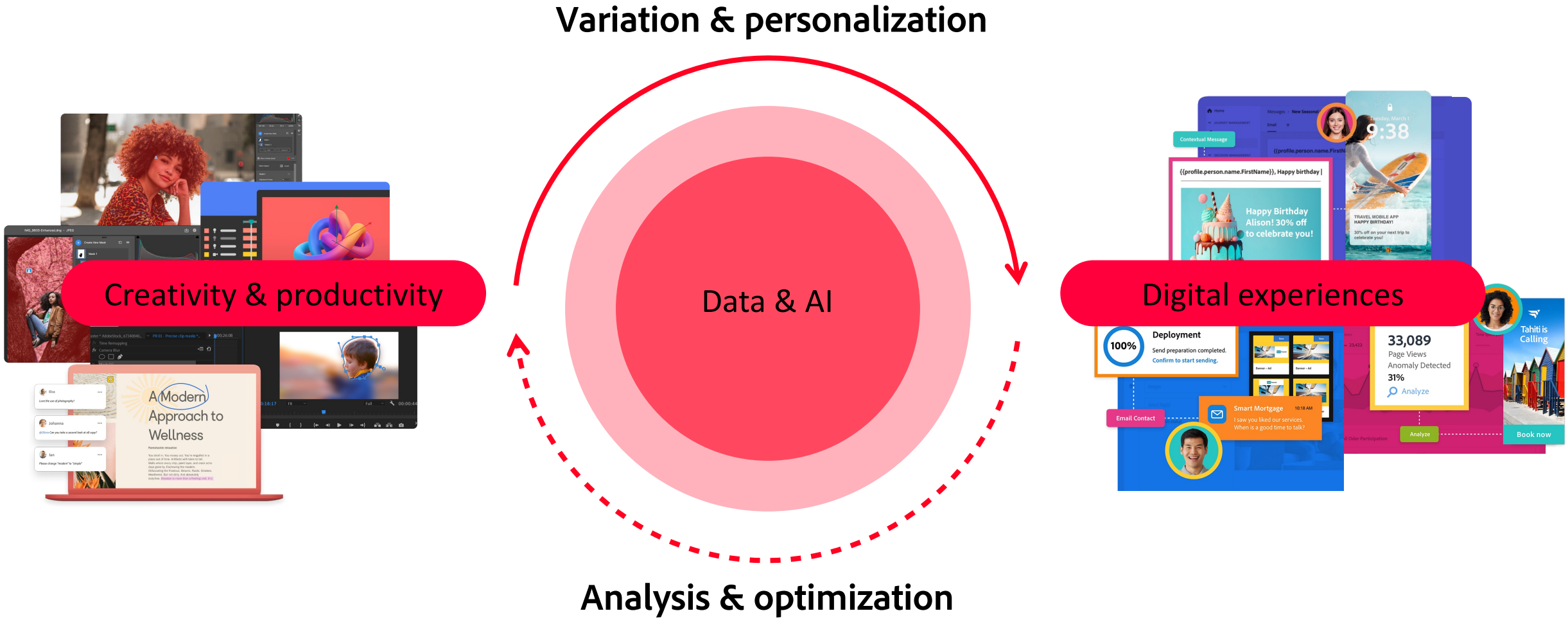


- Stateless provisioning for inference instances.
- It's a TFLOPs game. 1 image – 10-15 seconds.
- Dynamic scaling is not great when silicon is scarce. Move to RI.
- Optimizing models for inference on lower-cost compute (G5 and P4).

Insights on Inference

- **Multiple, specialized models working together in a DAG**
  - Large foundational models for generation
  - Smaller models for specific business logic
- **Initially inference on A100s**
  - Mistake! (A100s are scarce). We're back now.
- **Quickly switched to A10G inference (slowly rolling out A100s again into the mix)**
  - Worked with partners at AWS and NVIDIA

The digital experience flywheel

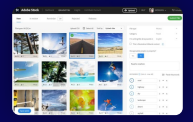


Students Consumers Communicators Creative Professionals Developers SMBs Enterprises

Exciting time for innovation at Adobe

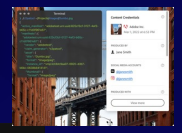
Strong Foundation

Adobe Stock



Neural Filters

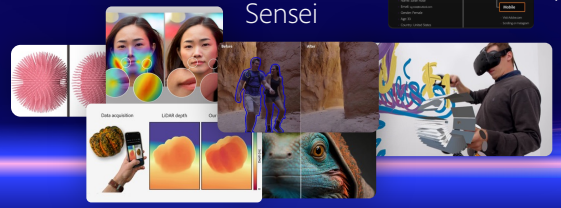
Content Authenticity



Partnerships

Real time customer profile

Sensei

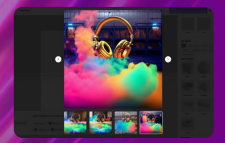


Cloud Ecosystem

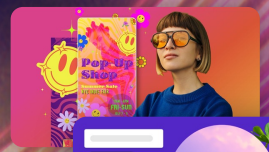
Accelerating Pipeline

AI Assistant

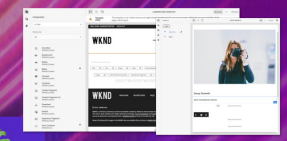
Text to 3D



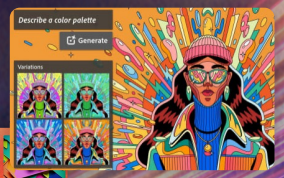
Express text effects



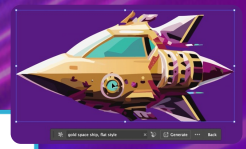
Adobe Experience Manager



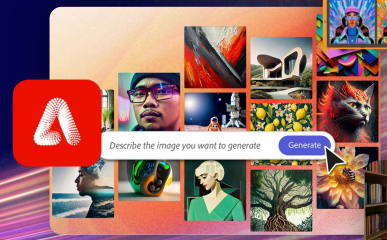
Illustrator recolor



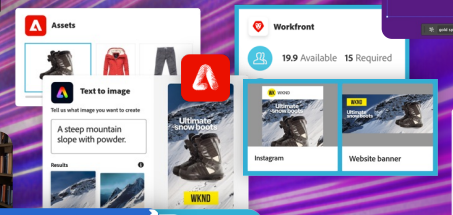
Text to vector



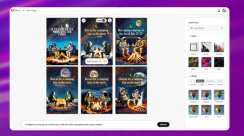
Firefly app



GenStudio



Text to template



Photoshop gen-fill, gen-expand

Productivity and creativity unleashed

Development and design cycles in real time

Talent, insights, and scale at speed