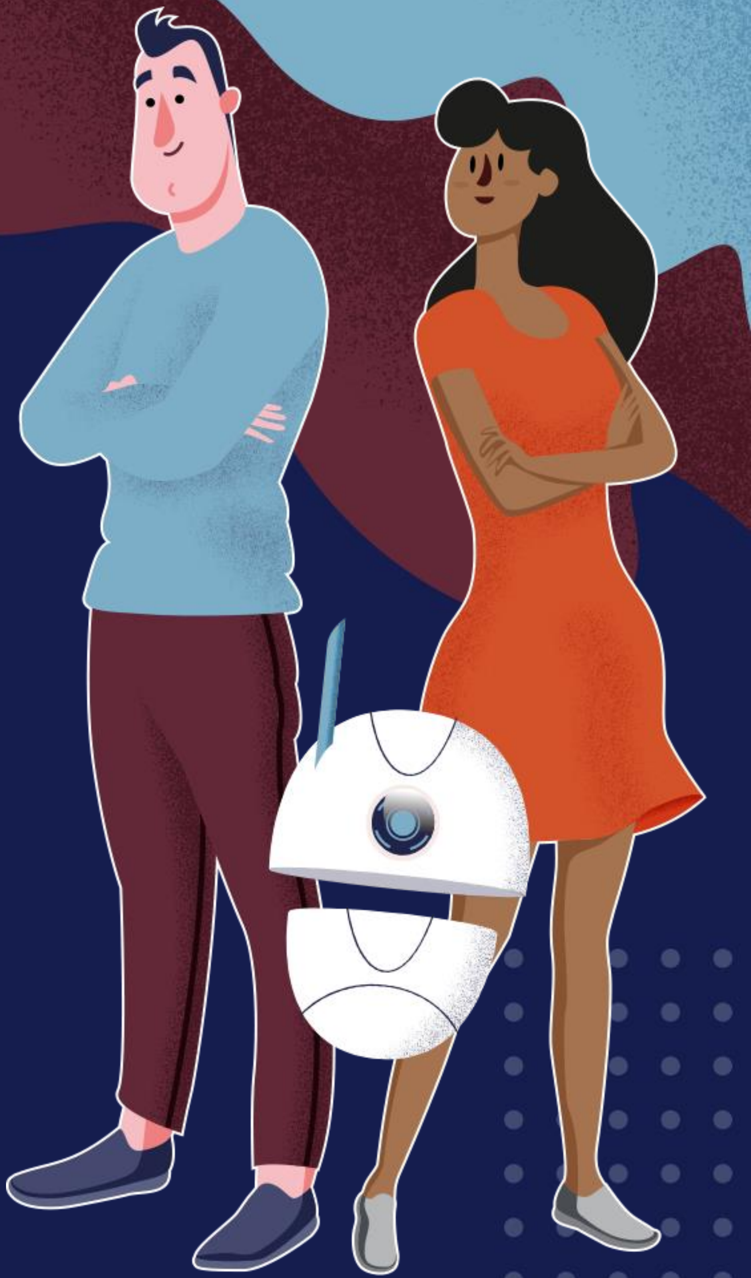
global AI bootcamp

March 4th 2023, Torino - Italy

Making AI optimization part of every developer's toolkit

Diego Fiori
CTO of Nebuly

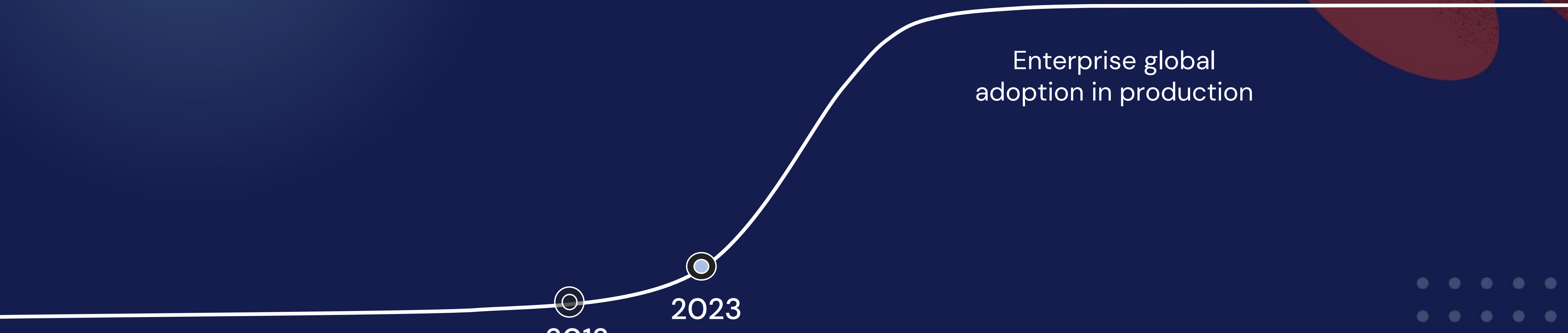


global AI community

Deltatre Innovation Lab

#GlobalAIBootcamp

Towards AI global enterprise adoption in production



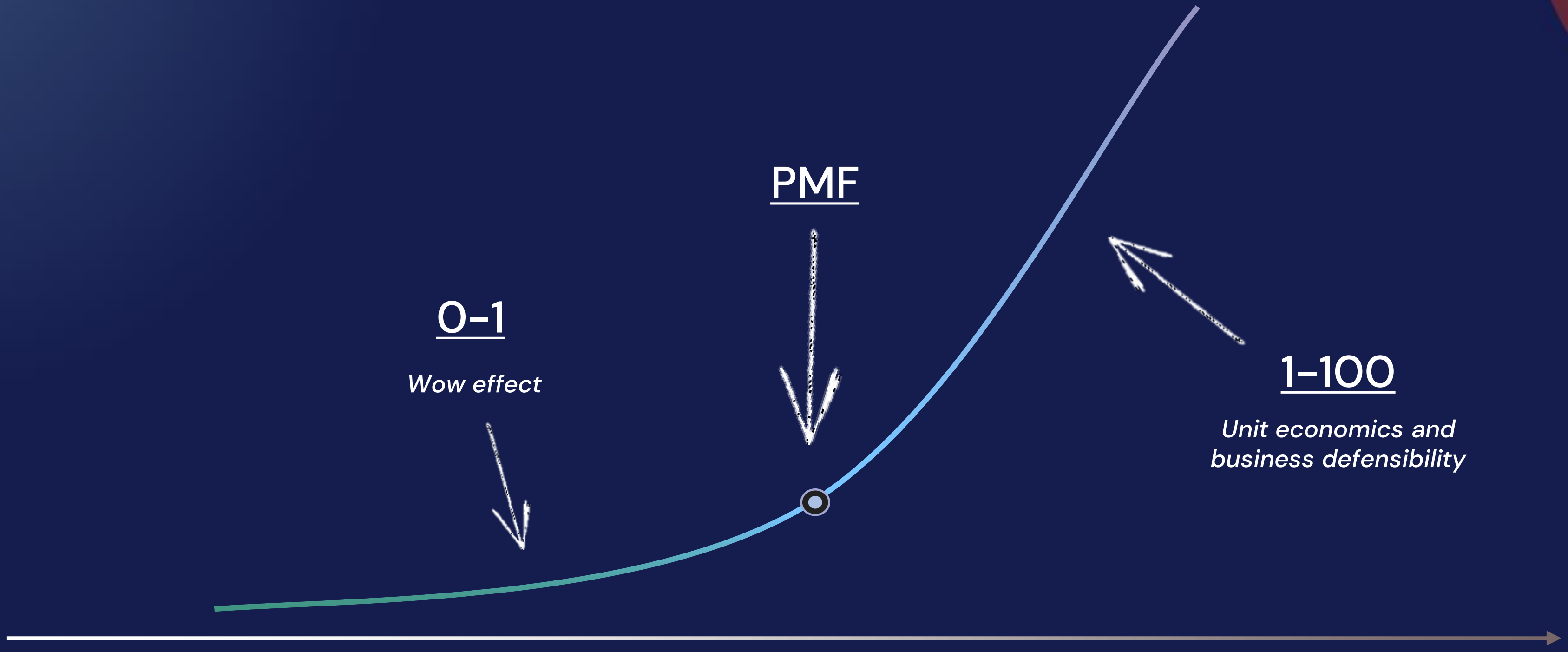
Enterprise global adoption in production

AI confined in research labs

2012 AlexNet

2023

Many new AI companies will emerge. At first what matters is to demonstrate the value of large models capabilities, but what’s next?



Adoption of LLM capabilities within a company

In the “1-100” phase, compute will be the primary cost driver of companies using large models and constrain the business models they can choose.

Cristobal Valenzuela
@c_valenzuelab

runway

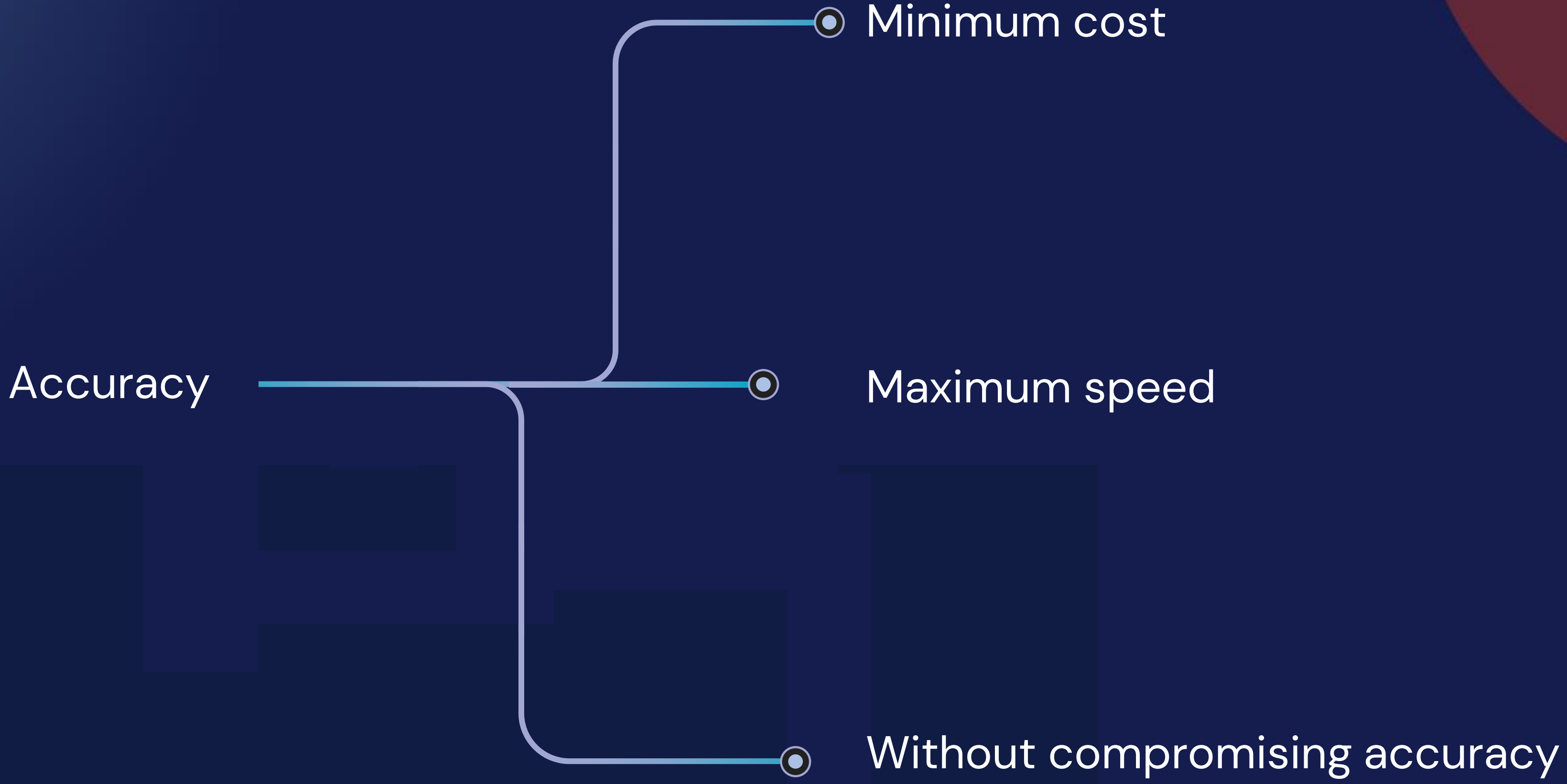
“If language models become a commodity, and I think we’re starting to see that already, then it becomes a function of cost. Whoever can run the cheapest and most effective model will win.”

Jon Turow
@jturow

Madrona aws

“A new generation of tooling and infrastructure for deployment optimization, training, and infrastructure, is helping builders to operate more efficiently to unblock new business models.”

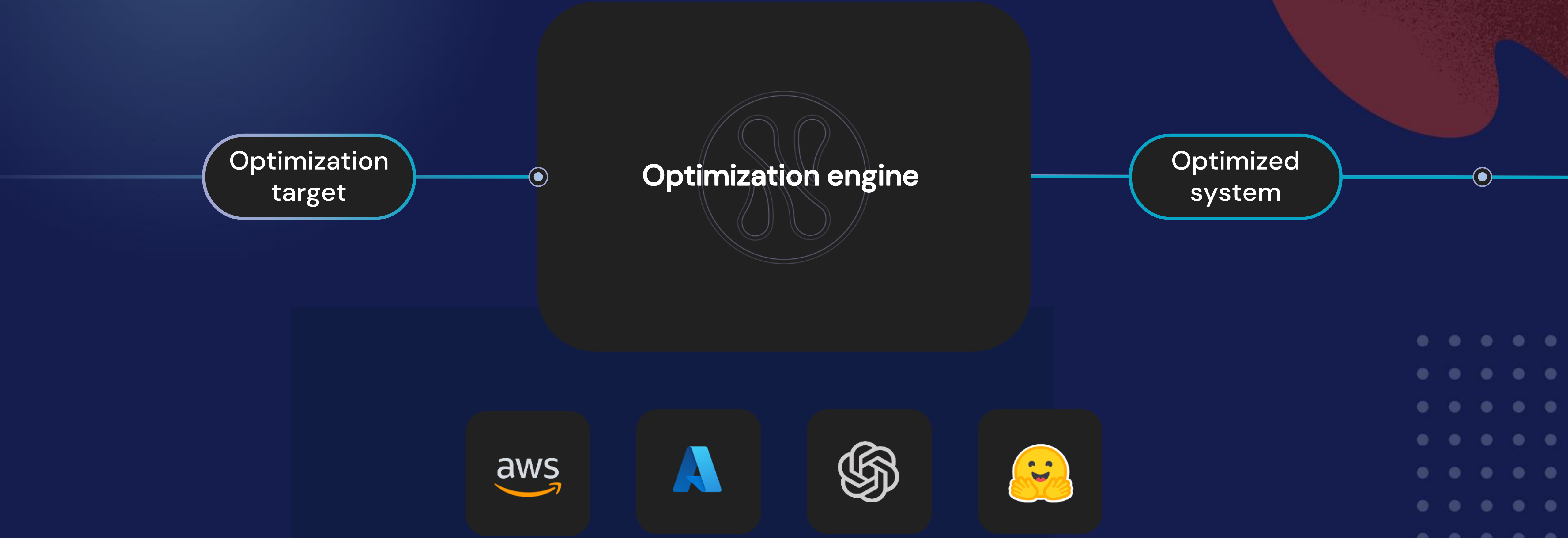
Using AI in production requires a shift in focus from accuracy to performance



Nebuly AI

The platform for AI optimization

Automatically optimize your AI system

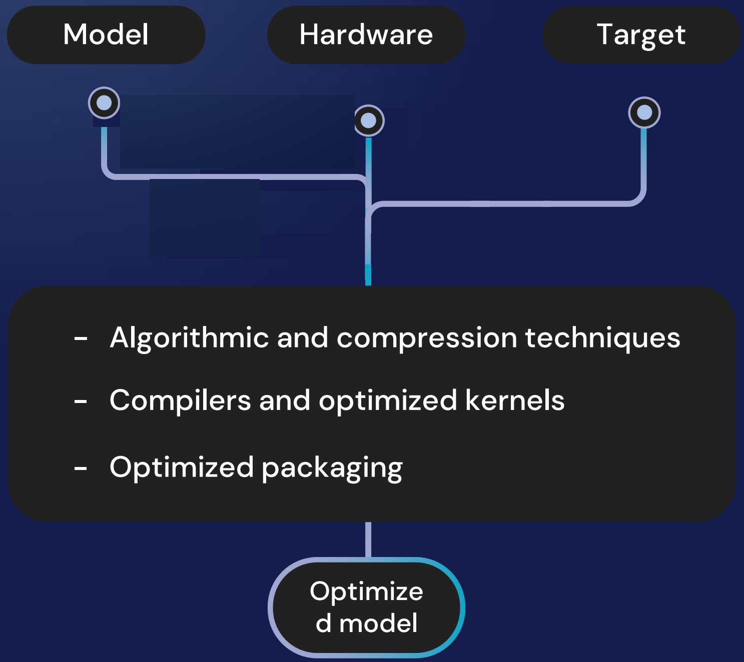


Nebuly AI

We build on strong open-source roots

Inference - Speedster

Automatically apply the best set of SOTA optimization techniques to achieve the maximum inference speed-up on your hardware.



- Algorithmic and compression techniques
- Compilers and optimized kernels
- Optimized packaging

Integrations

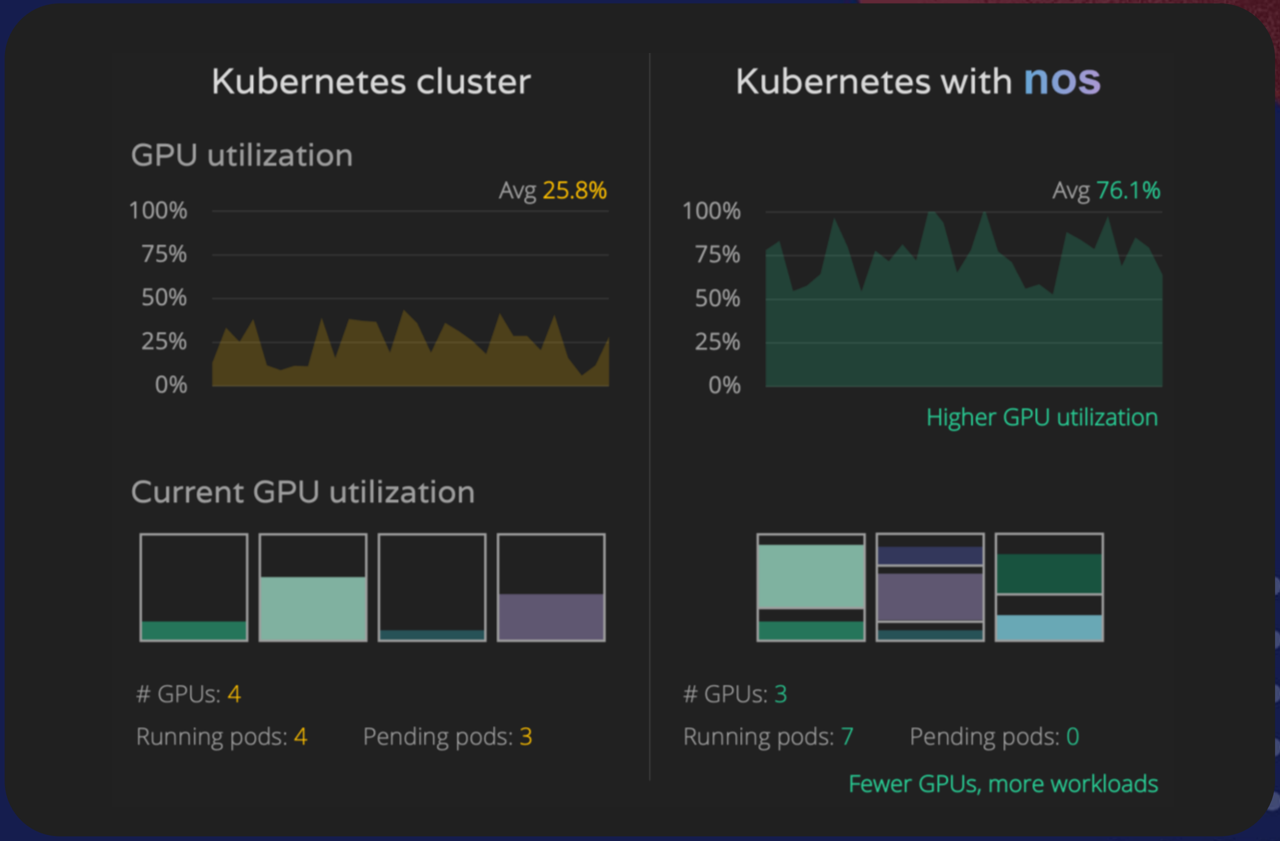
PyTorch, Hugging Face, ONNX, TensorFlow, Docker, NVIDIA Triton

```
import torch
import torchvision.models as models
from speedster import optimize_model, save_model

model = models.resnet50()
input_data = [((torch.randn(1, 3, 256, 256), ), torch.tensor([0])) for _ in range(100)]

# Run Speedster optimization
optimized_model = optimize_model(
  model,
  input_data=input_data,
  optimization_time="constrained",
  metric_drop_ths=0.05)
```

Infrastructure - Nos

Automatically maximize GPUs utilization in a Kubernetes cluster via real-time dynamic partitioning and elastic quotas.

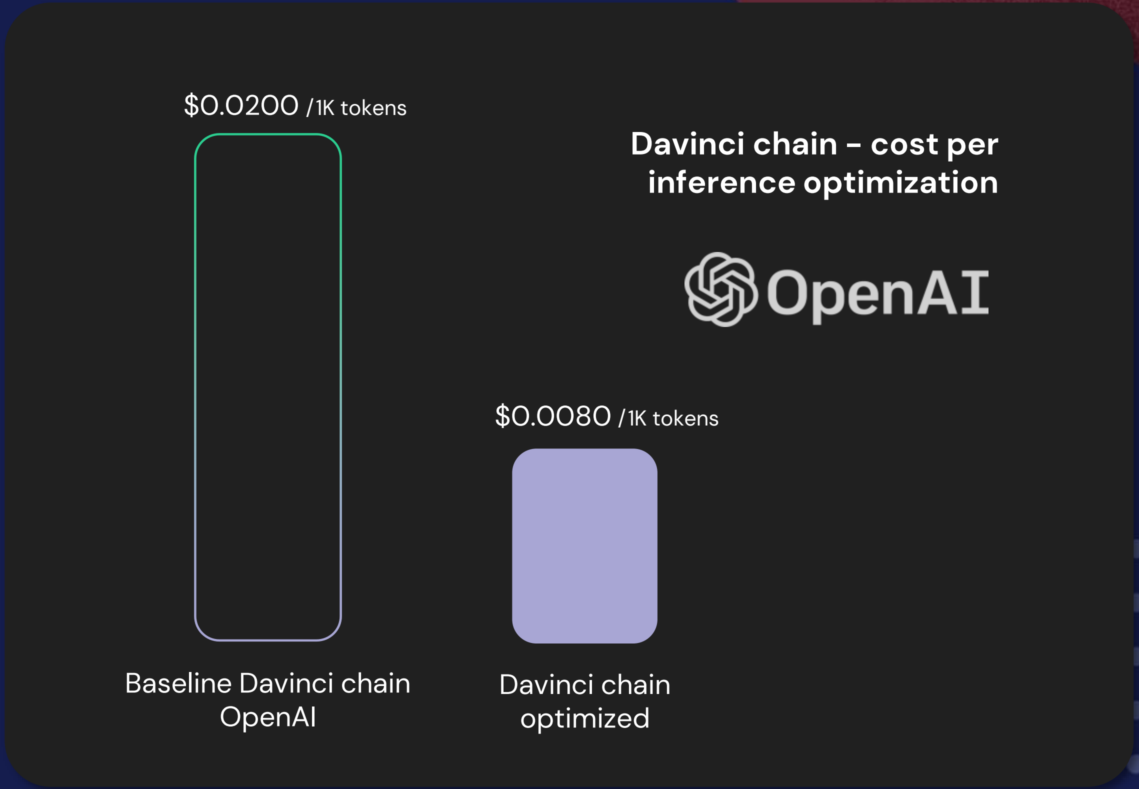
- Accelerate time-to-value by minimizing pending AI workloads
- Slash cloud spending / boost on prem infrastructure ROI



APIs - GPT Optimizer 🚧 COMING SOON 🚧

Turn any third-party API into your own API. Regain control of inference costs.

- Slash ballooning APIs inference costs
- Leverage your company data to create your own personalized APIs
- Continuously optimize the performances of your APIs
- Fully integrated with LangChain, GPT index etc



Davinci chain – cost per inference optimization

A global community of leading AI companies

5000+

GitHub stars.

2000+

Members of our community.

500+

Logos that interacted with our products.

50+

Countries.

Run your ML application as usual, just optimized