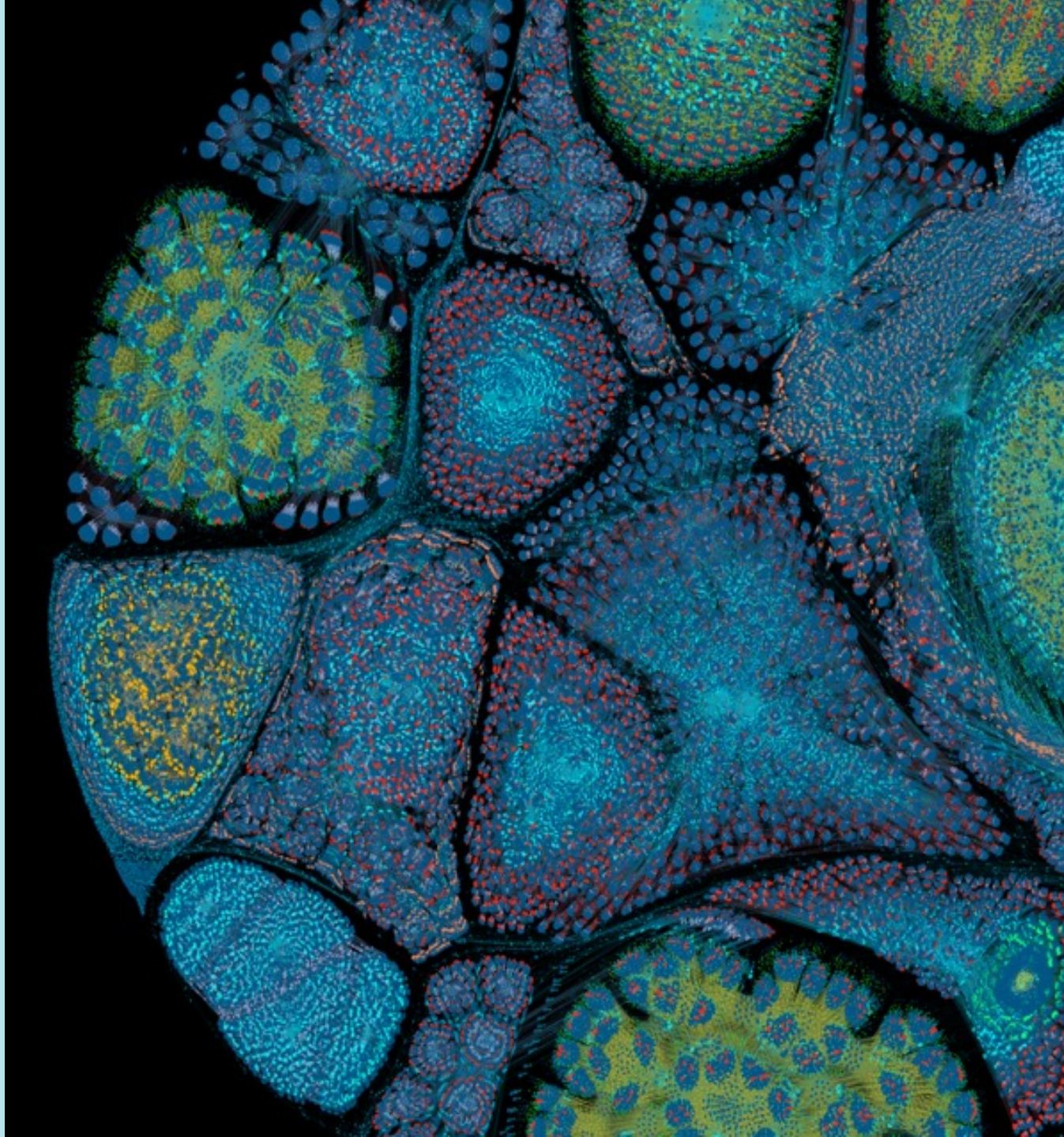
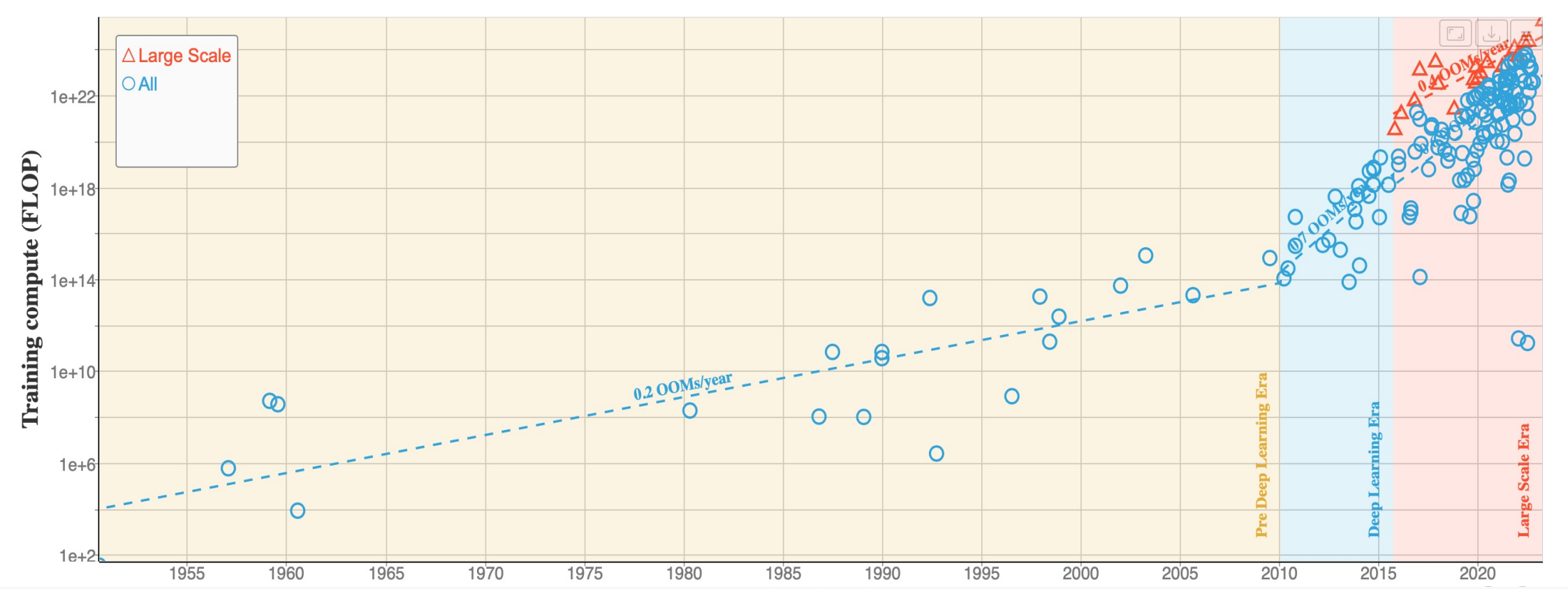
AI from Silicon to Solution

An ecosystem view of building AI Products

Tim Santos
Director of Product, AI Cloud Solutions



WHY BOTHER?



"progress in machine learning (ML) is driven by three primary factors - algorithms, data, and compute"

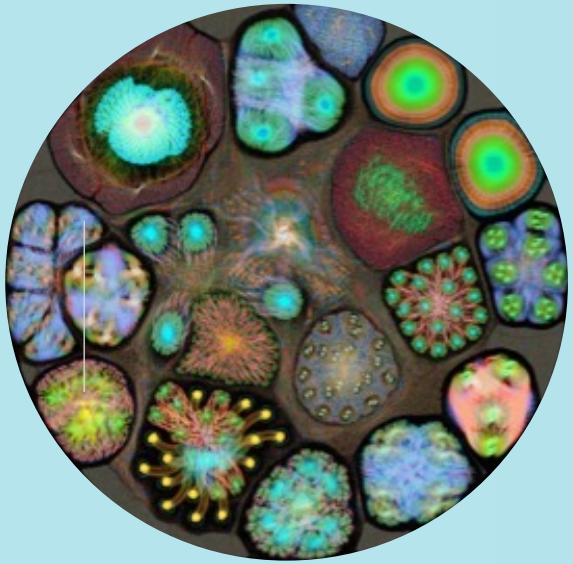
EPOCH

HOW WE ARE TACKLING IT

Hardware

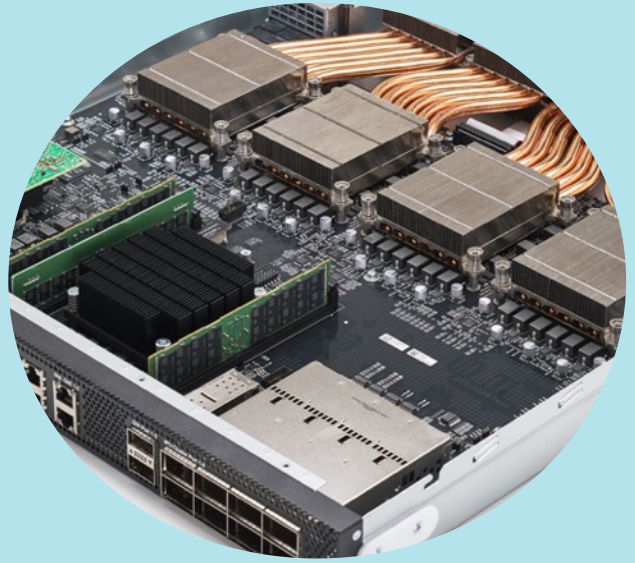
IPU processors designed for AI

Software



Poplar® software stack & development tools

Systems



IPU-M2000 and Server IPU-POD$_{64}$ scale-out

BOW IPU

Deep Trench Capacitor

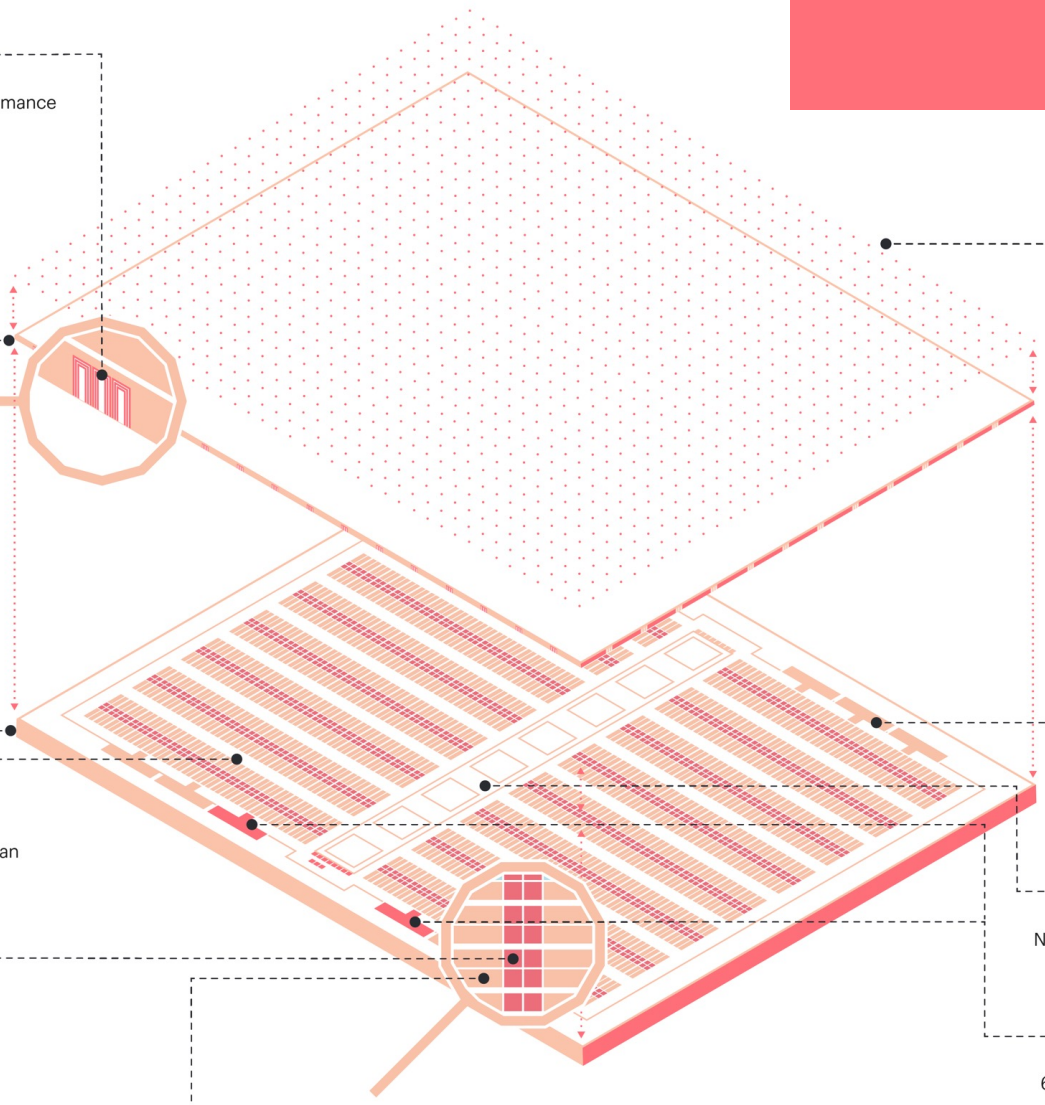
Efficient power delivery
Enables increase in operational performance

Wafer-On-Wafer

Advanced silicon 3D stacking technology

Closely coupled power delivery die

Higher operating frequency and enhanced overall performance



Solder Bumps

IPU-Links™

10x IPU-Links,
320GB/s chip to chip bandwidth

IPU-Exchange™

11 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

PCIe

PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Tiles™

1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

IPU-Core™

1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

65.4TB/s memory bandwidth per IPU

THE BOW IPU
WORLD'S FIRST 3D WAFER-ON-WAFER PROCESSOR

3D silicon wafer stacked processor

350 TeraFLOPS AI compute

Optimized silicon power delivery

0.9 GigaByte In-Processor-Memory @ 65TB/s

1,472 independent processor cores

8,832 independent parallel programs

11.4 TB/s internal IPU-Exchange bandwidth

10x IPU-Links™ delivering 320GB/s

GRAPHCORE

BOW-2000 IPU MACHINE

4 x Bow 3D Wafer-on-Wafer IPUs

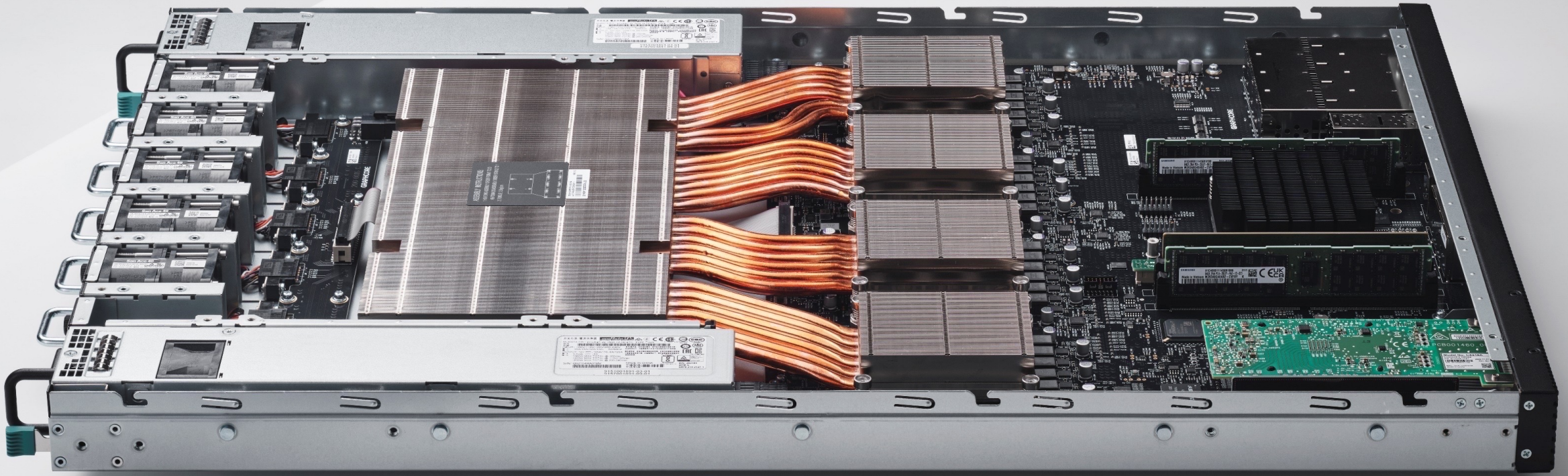
1.4 PetaFLOPS AI Compute

3.6 GB In-Processor-Memory @ 260TB/s

Up to 256 GB IPU Streaming Memory

2.8 Tbps IPU-Fabric™

Same 1U blade form factor



BOW: 3RD GENERATION IPU SYSTEMS

BOW POD16

4x Bow-2000
5.6 PetaFLOPS
1 CPU server

BOW POD64

16x Bow-2000
22.4 PetaFLOPS
1-4 CPU server(s)

BOW POD256

64x Bow-2000
89.6 PetaFLOPS
4-16 CPU server(s)

CLOUD SERVICE

Paperspace

Notebook based development environment

Jupyter notebooks ready to run on IPU

N America focus initially

For Model Experimentation and Development

G-CORE LABS

G-Core Labs

On-Demand IPU Cloud Infrastructure

Scalable, Flexible, Convenient

Europe focus initially

For Deployment and Productisation

GET UP & RUNNING ON IPUS RIGHT AWAY

AI COMPUTE ON DEMAND

IMPROVE INNOVATION WITH FASTER & LOWER COST TO TRAIN FOR:
- NLP, CV & GRAPH ML MODELS

EXTENSIVE RANGE OF ML MODELS READY TO RUN ON IPU

IPU-POD AND BOW POD PLATFORMS AVAILABLE FROM POD4 THORUGH TO LARGER SYSTEMS

STANDARD ML FRAMEWORK SUPPORT

Develop models using standard high-level frameworks or port existing models

PyTorch

TensorFlow

PyTorch Lightning

HUGGING FACE

PyG

Keras

PaddlePaddle

HALO

Existing models on alternative platforms

Easy port of high-level framework models

POPLAR®

IPU-Processor Platforms

MODEL GARDEN COVERAGE

COMPUTER VISION

IMAGE CLASSIFICATION

- ResNet50 v1.5
- EfficientNet-B0
- EfficientNet-B4
- ResNeXt-101
- MobileNet v2
- MobileNet v3
- ViT
- DINO
- SWIN
- MAE

OBJECT DETECTION

- YOLO v3
- YOLO v4
- Faster RCNN
- EfficientDet

OBJECT SEGMENTATION

- Unet (Industrial)
- Unet (Medical)

GNN

- TGN NEW
- MPNN-GIN NEW
- GPS++ NEW
- Distr. KGE NEW
- Cluster-GCN NEW
- SchNet NEW
- NBFNet NEW

AI FOR SIMULATION

- DeepMD
- DeepDriveMD
- ETO
- CosmoFlow
- ABC Covid-19

REINFORCEMENT

- RL
  Reinforcement Learning

PROBABILISTIC

- MCMC
- VAE

NLP

- Dolly NEW
- BERT
- Group | Packing
- GPT-2
- GPT-J NEW
- RoBERTa
- Deberta
- BART
- T5
- Hubert
- DistilBERT

SPEECH

STT (ASR)

- RNN-T
- Conformer
- Wav2Vec2 NEW
- Whisper NEW

TTS

- DeepVoice3
- FastSpeech2
- FastPitch

RECOMMENDER

- Autoencoder
- DIN
- DIEN

MULTIMODAL

- LXMERT
- CLIP
- Stable Diffusion NEW
- Mini DALL-E
- Frozen In Time NEW

OTHER

- Sales Forecast
- Neural Image Fields

PyG

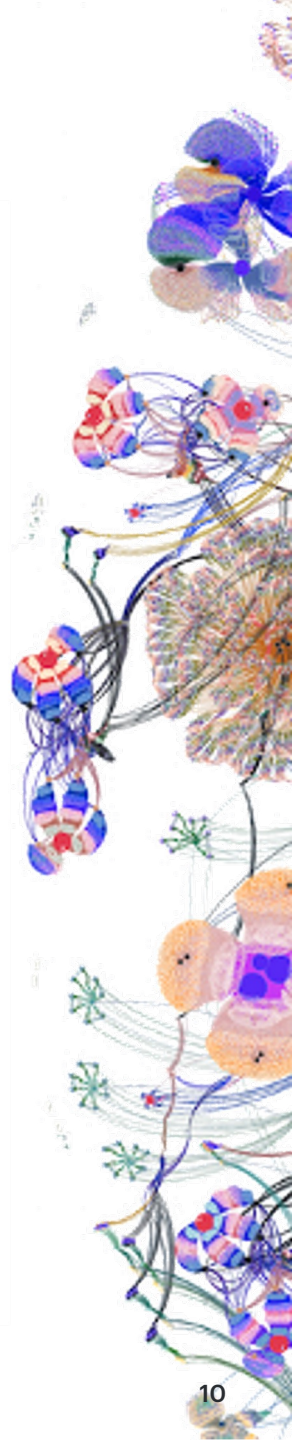
PyTorch

TensorFlow

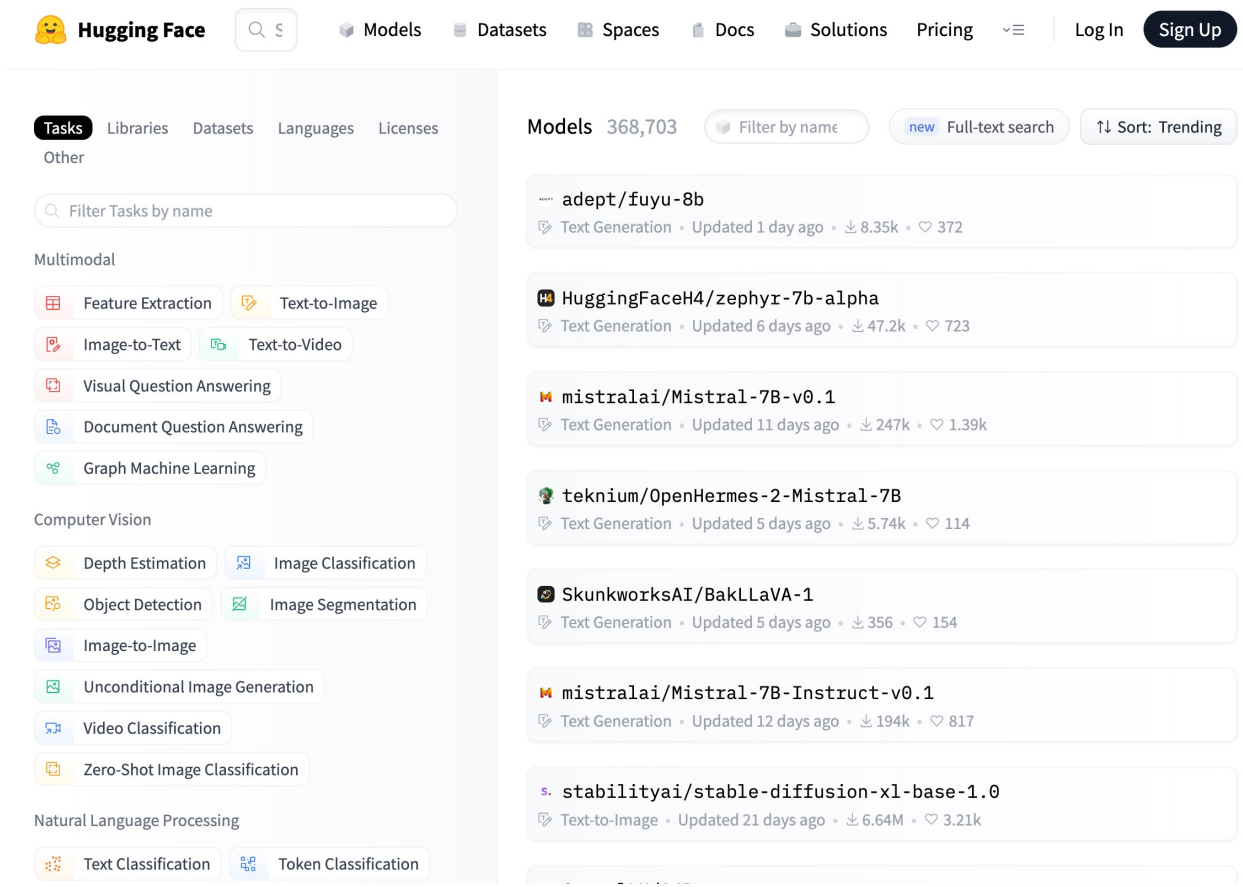
Hugging Face

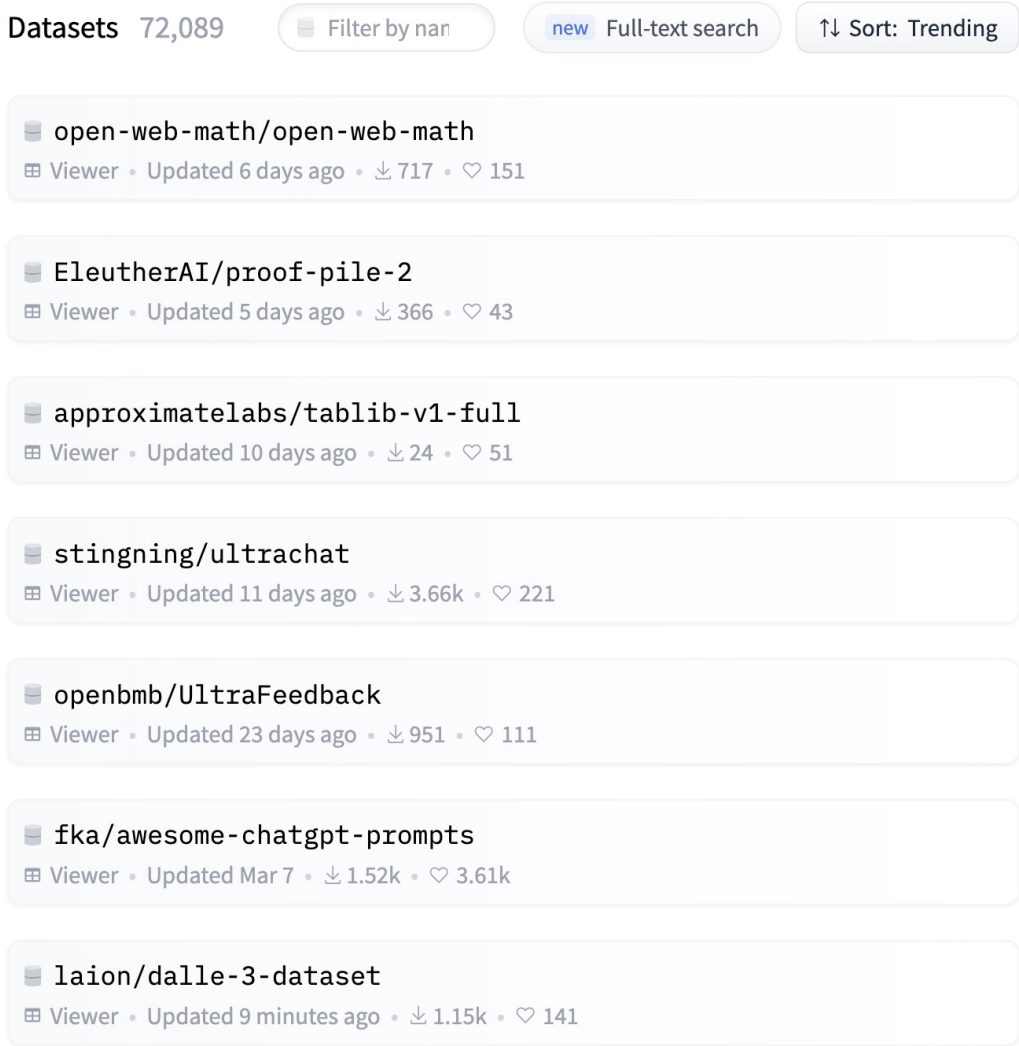
Keras

PaddlePaddle



MODEL HUB


Hugging Face
Models
Datasets
Spaces
Docs
Solutions
Pricing
Log In
Sign Up
Tasks
Libraries
Datasets
Languages
Licenses
Other
Filter Tasks by name
Multimodal
Feature Extraction
Text-to-Image
Image-to-Text
Text-to-Video
Visual Question Answering
Document Question Answering
Graph Machine Learning
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Image-to-Image
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Natural Language Processing
Text Classification
Token Classification
Models 368,703
Filter by name
new Full-text search
Sort: Trending
adept/fuyu-8b
Text Generation • Updated 1 day ago • 8.35k • 372
HuggingFaceH4/zephyr-7b-alpha
Text Generation • Updated 6 days ago • 47.2k • 723
mistralai/Mistral-7B-v0.1
Text Generation • Updated 11 days ago • 247k • 1.39k
teknium/OpenHermes-2-Mistral-7B
Text Generation • Updated 5 days ago • 5.74k • 114
SkunkworksAI/BakLLaVA-1
Text Generation • Updated 5 days ago • 356 • 154
mistralai/Mistral-7B-Instruct-v0.1
Text Generation • Updated 12 days ago • 194k • 817
stabilityai/stable-diffusion-xl-base-1.0
Text-to-Image • Updated 21 days ago • 6.64M • 3.21k


Datasets 72,089
Filter by nan
new Full-text search
Sort: Trending
open-web-math/open-web-math
Viewer • Updated 6 days ago • 717 • 151
EleutherAI/proof-pile-2
Viewer • Updated 5 days ago • 366 • 43
approximatelabs/tablib-v1-full
Viewer • Updated 10 days ago • 24 • 51
stingning/ultrachat
Viewer • Updated 11 days ago • 3.66k • 221
openbmb/UltraFeedback
Viewer • Updated 23 days ago • 951 • 111
fka/awesome-chatgpt-prompts
Viewer • Updated Mar 7 • 1.52k • 3.61k
laion/dalle-3-dataset
Viewer • Updated 9 minutes ago • 1.15k • 141

SELF- SERVICE NOTEBOOKS

Aug 01, 2023

LLAMA 2: RUN META'S OPEN SOURCE LARGE LANGUAGE MODEL FOR FREE ON IPUS

Written By:
Tim Santos and Arsalan Uddin

Chatbot Open Source LLM
Llama 2 - Inference
Run on Gradient

May 31, 2023

OPENASSISTANT FINE-TUNED PYTHIA-12B: OPEN-SOURCE CHATGPT ALTERNATIVE

Written By:
Steve Barlow

Chatbot using OpenAssistant
Pythia 12B - Inference
Run on Gradient

Apr 26, 2023

DOLLY 2.0 - OPEN SOURCE LANGUAGE MODEL WITH CHATGPT-LIKE INTERACTIVITY

Written By:
Alex McKinney

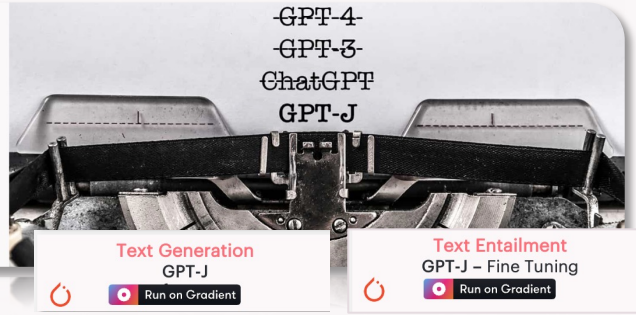
Dolly

Instruction Tuned LLM
Dolly 2.0 - Inference
Run on Gradient

Mar 24, 2023

FINE-TUNE GPT-J: A COST-EFFECTIVE GPT-4 ALTERNATIVE FOR MANY NLP TASKS

Written By:
Sofia Liguori



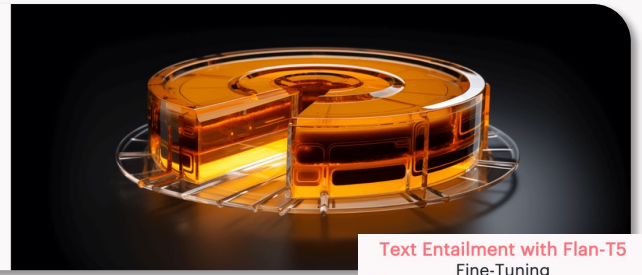
Text Generation
GPT-J
Run on Gradient

Text Entailment
GPT-J – Fine Tuning
Run on Gradient

Jul 27, 2023

FINE-TUNING FLAN-T5 XXL - THE POWERFUL AND EFFICIENT LLM

Written By:
Manuele Sigona



Text Entailment with Flan-T5
Fine-Tuning
Run on Gradient

May 30, 2023

FLAN-T5: SWEET RESULTS WITH THE SMALLER, MORE EFFICIENT LLM

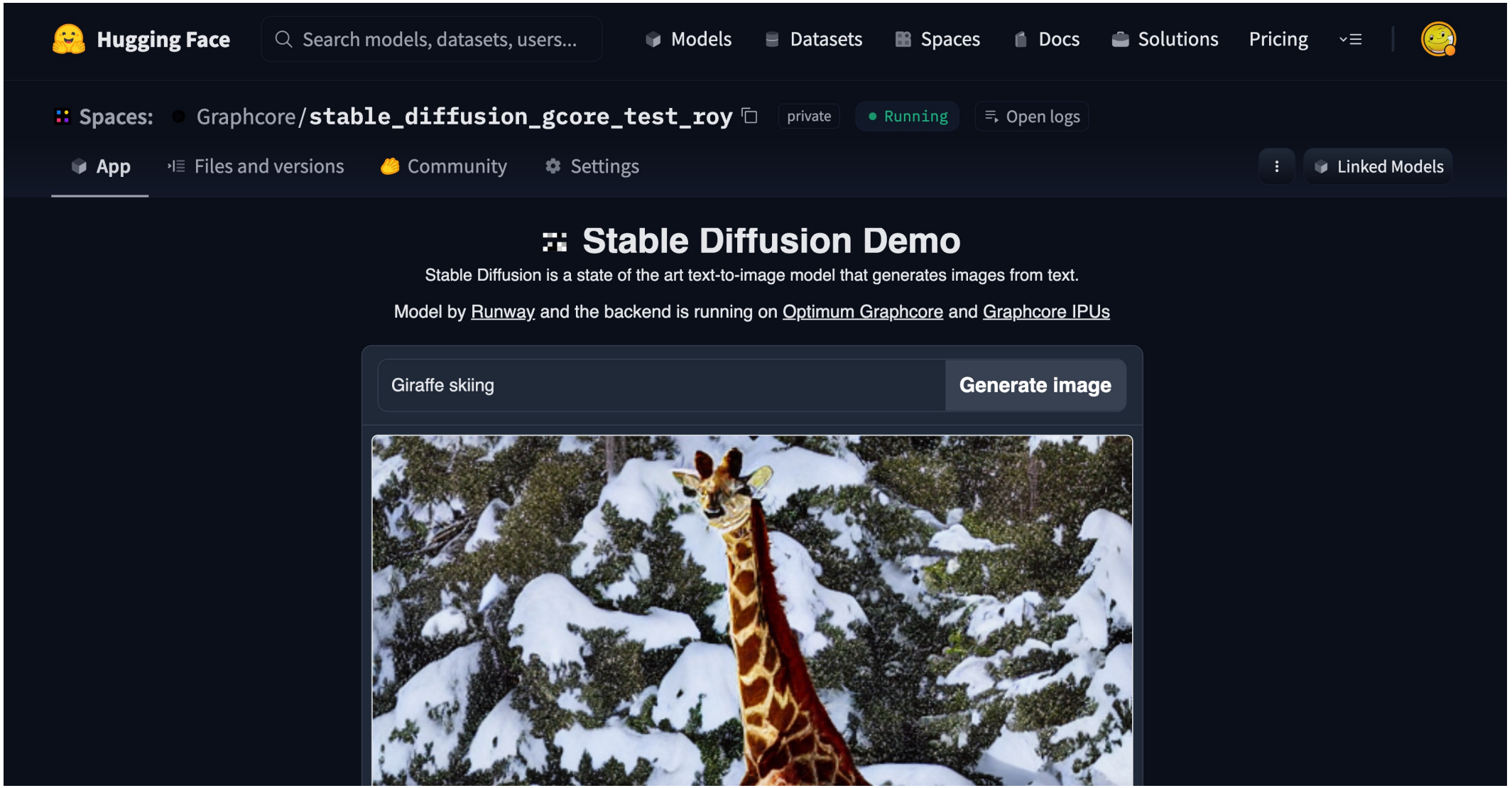
Written By:
Harry Mellor

Flan-T5 is all you need
Inference
Run on Gradient

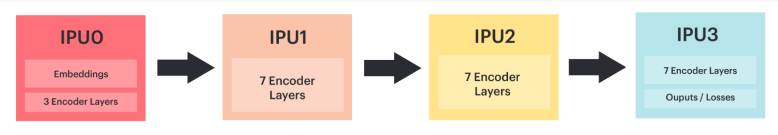
Text Generation with Flan-T5
Inference
Run on Gradient

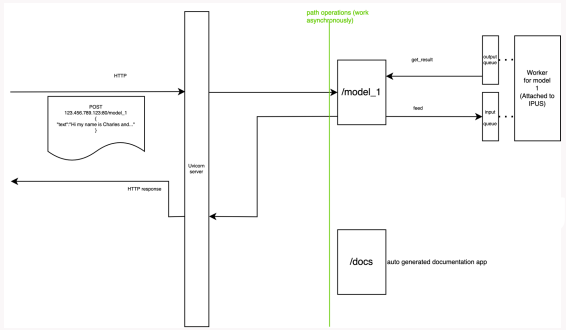
ipu.dev/notebooks

DEMOS AND APPLICATIONS



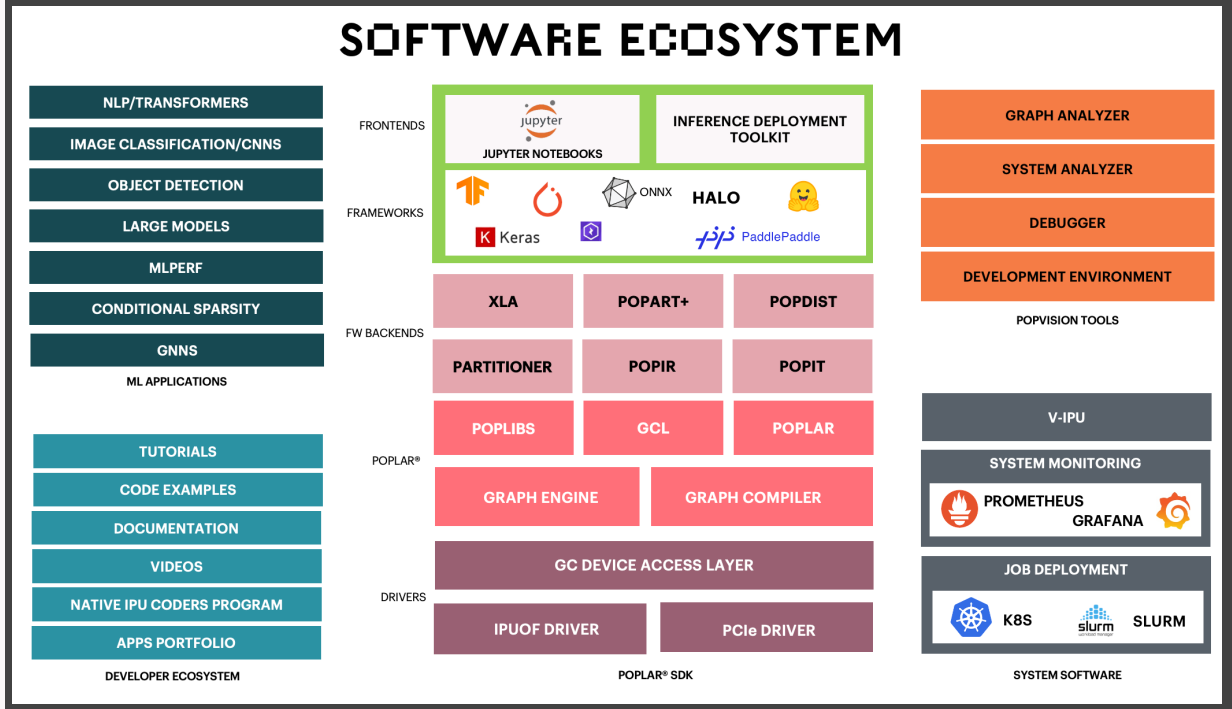
PRODUCTISATION





SOME LEARNINGS

IT TAKES A VILLAGE...



G-CORE LABS
G-Core Labs

On-Demand
IPU Cloud Infrastructure

Scalable, Flexible,
Convenient

Paperspace

Notebook based
development environment

Jupyter notebooks ready to
run on IPU

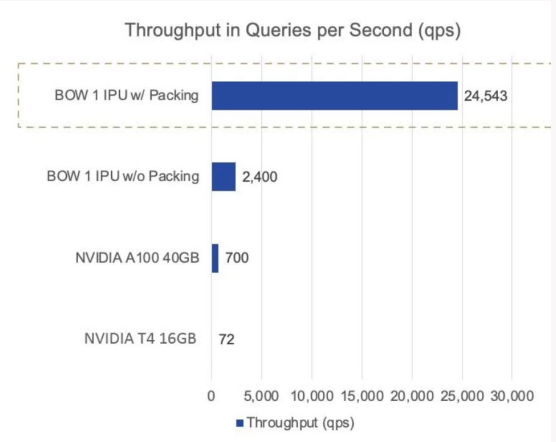
GRAPHCORE + pienso

AI-for-all
AI now accessible by non-technical subject matter experts

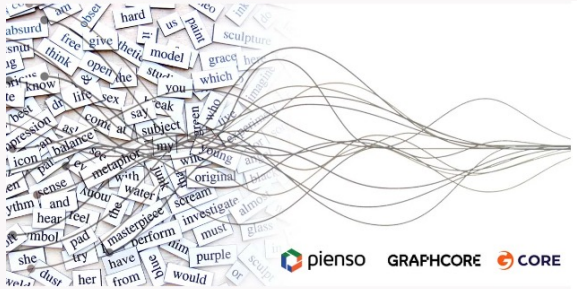
Performance
IPUs enable Pienso with upto **35x** speed-up over leading GPUs

MLOps Solution
Fully integrated MLOps platform with Training & Inference

CogX
BEST INNOVATION IN NLP : 2023
pienso + GRAPHCORE



Pienso is a low/no-code deep learning platform that allows non-technical users to analyse high-volume text data to drive insights – without seeing a single line of code.



PIENSO OFFERS EFFICIENT LLM ACCESS FOR BUSINESS, POWERED BY CLOUD IPUS

TO A HAMMER...

The Hardware Lottery

Sara Hooker

Google Research, Brain Team
shooker@google.com

Abstract

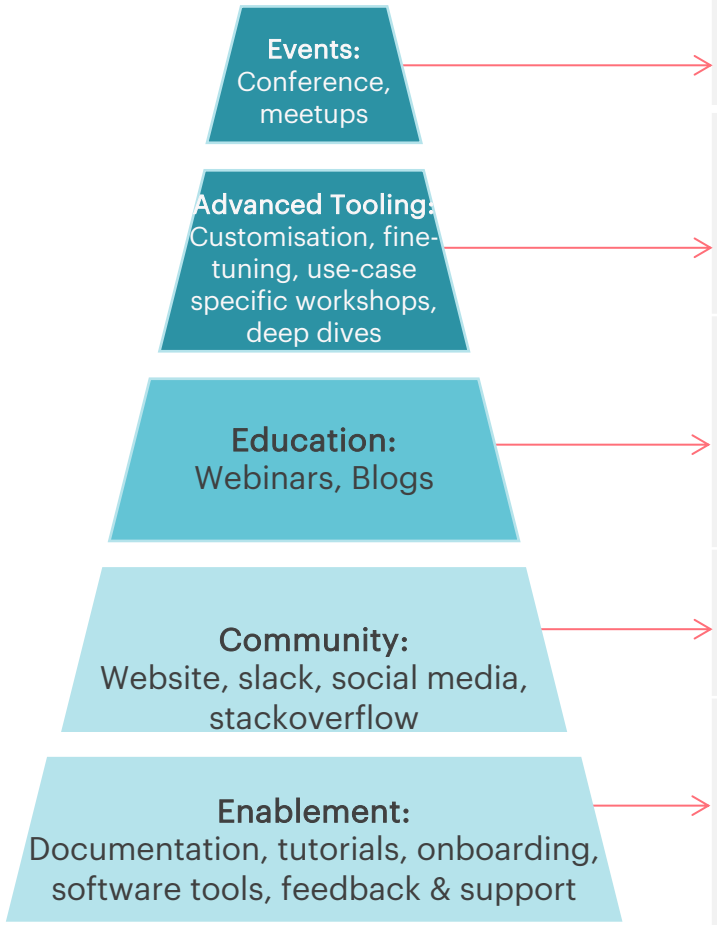
Hardware, systems and algorithms research communities have historically had different incentive structures and fluctuating motivation to engage with each other explicitly. This historical treatment is odd given that hardware and software have frequently determined which research ideas succeed (and fail). This essay introduces the term hardware lottery to describe when a research idea wins because it is suited to the available software and hardware and *not* because the idea is superior to alternative research directions. Examples from early computer science history illustrate how hardware lotteries can delay research progress by casting successful ideas as failures. These lessons are particularly salient given the advent of domain specialized hardware which make it increasingly costly to stray off of the beaten path of research ideas. This essay posits that the gains from progress in computing are likely to become even more uneven, with certain research directions moving into the fast-lane while progress on others is further obstructed.

PILLARS OF TECH/DEV ADOPTION

Code

Content

Community



Create personal connections and reach developers to foster engagement

Empower developers beyond basics, improve the product with new features

The art of the possible, exhibit thought leadership beyond the product

Collaboration and resources beyond the developer portal

Accelerate developer adoption by improving **ease of use** and **ease of access** to IPUs, and knowledge of core IPU principles

Q&A

Connect:

 @internetoftim

 linkedin.com/in/internetoftim

 https://internetoftim.xyz