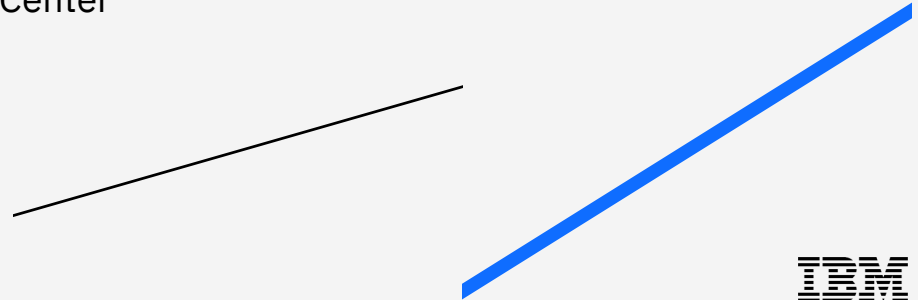


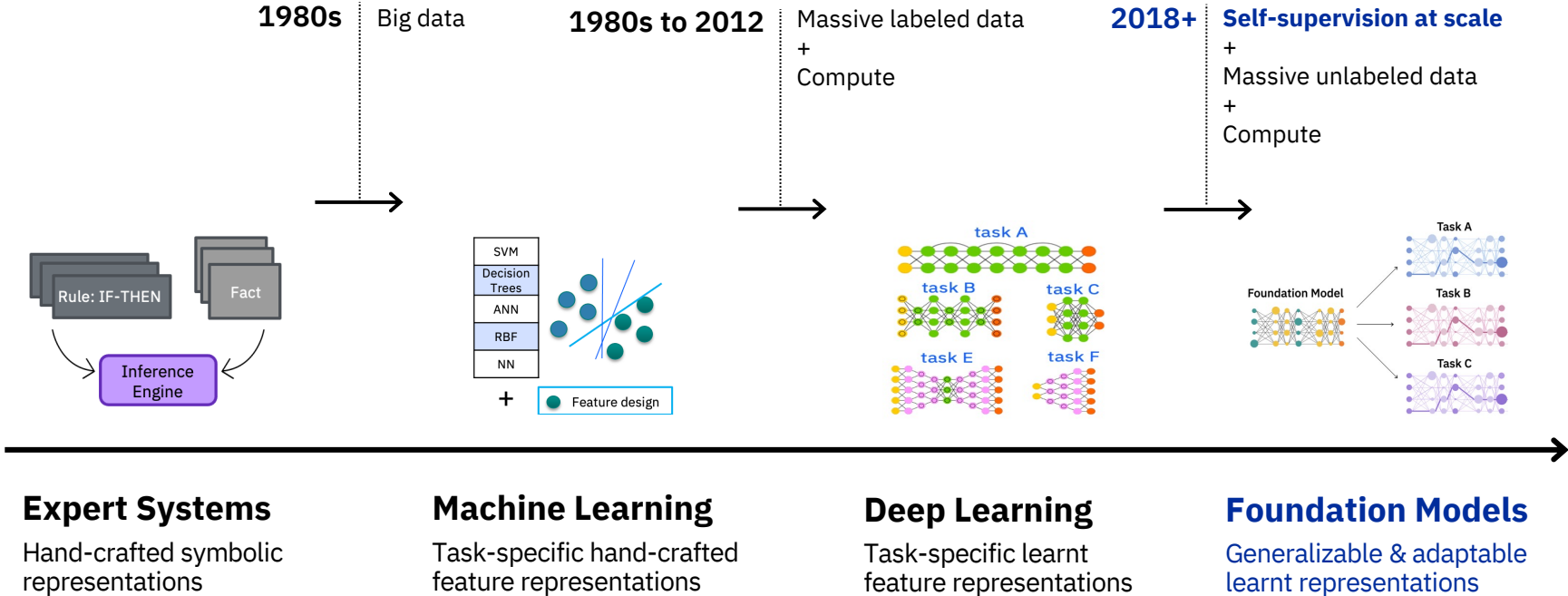
Challenges in AI Infrastructure for Enterprise Foundation Models

Jeffrey L. Burns, Ph.D.
Director, AI Compute and IBM Research AI Hardware Center
IBM Research

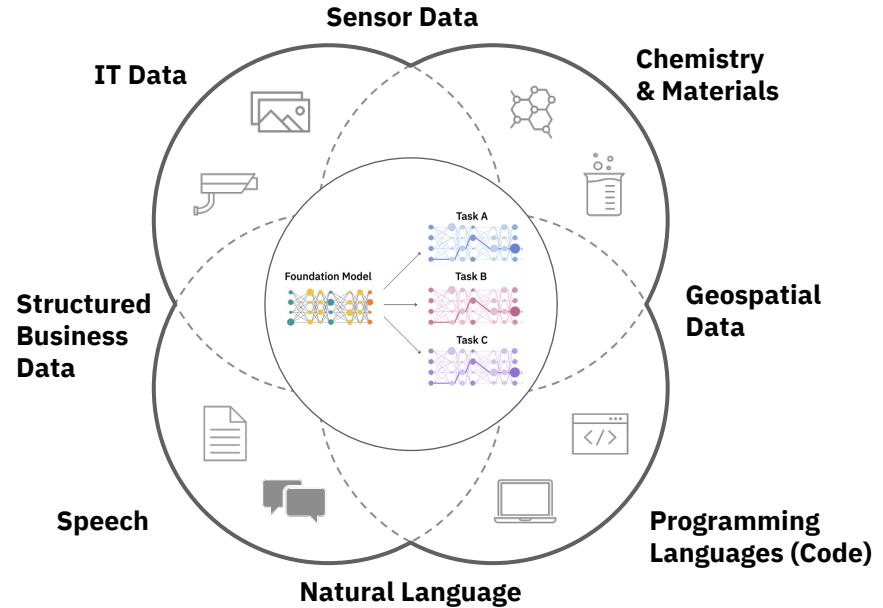
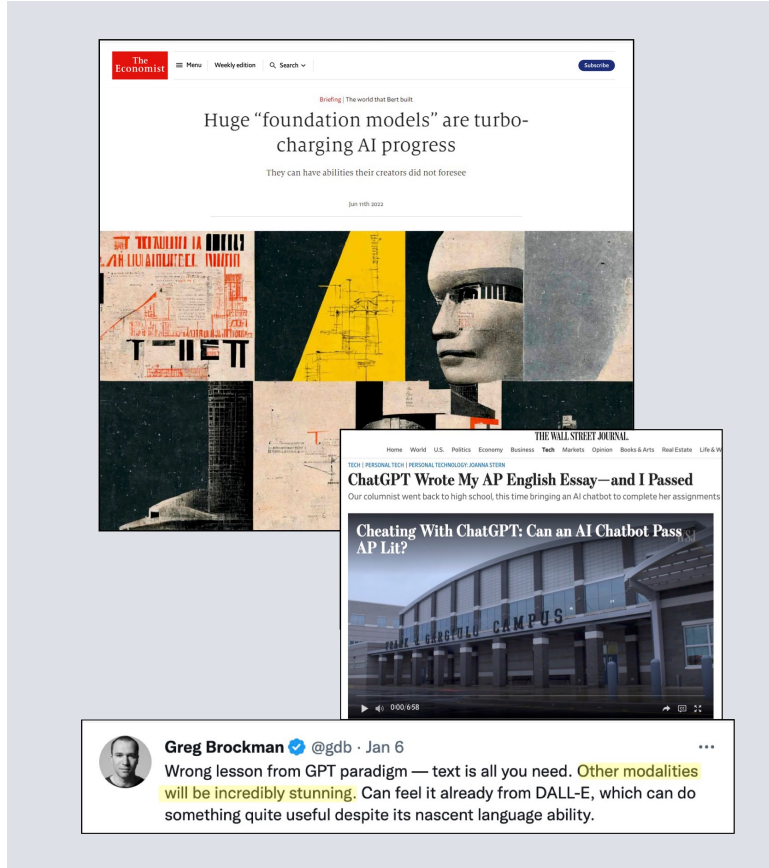
August 9, 2023



Foundation Models: An inflection point in generalizable and adaptable representations



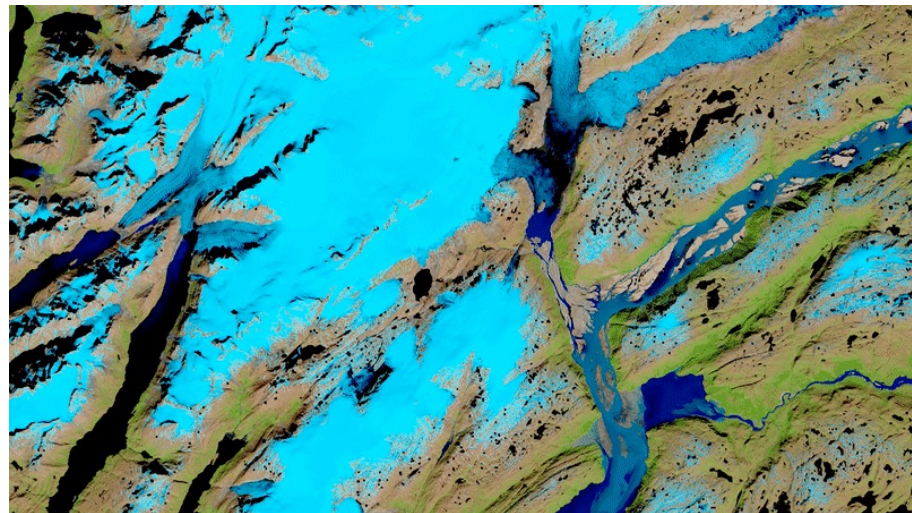
Incredible opportunities around enterprise applications



In each of these domains there is ample unlabeled data available in enterprises, which can be used to train custom foundation models, potentially opening the doors for solving business problems that were previously considered intractable.

Geospatial Foundation Models

IBM and NASA have teamed up to apply **foundation model AI technology** to leverage earth science data for **geospatial intelligence**.



This work with NASA is part of an effort across IBM Research to pioneer **applications of foundation models beyond language**.

<https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model>



Pre-trained on sufficient datasets in partnership with content-rich institutions (e.g. NASA)



Leverage **self-supervised learning** (i.e., masking imagery or timeseries)

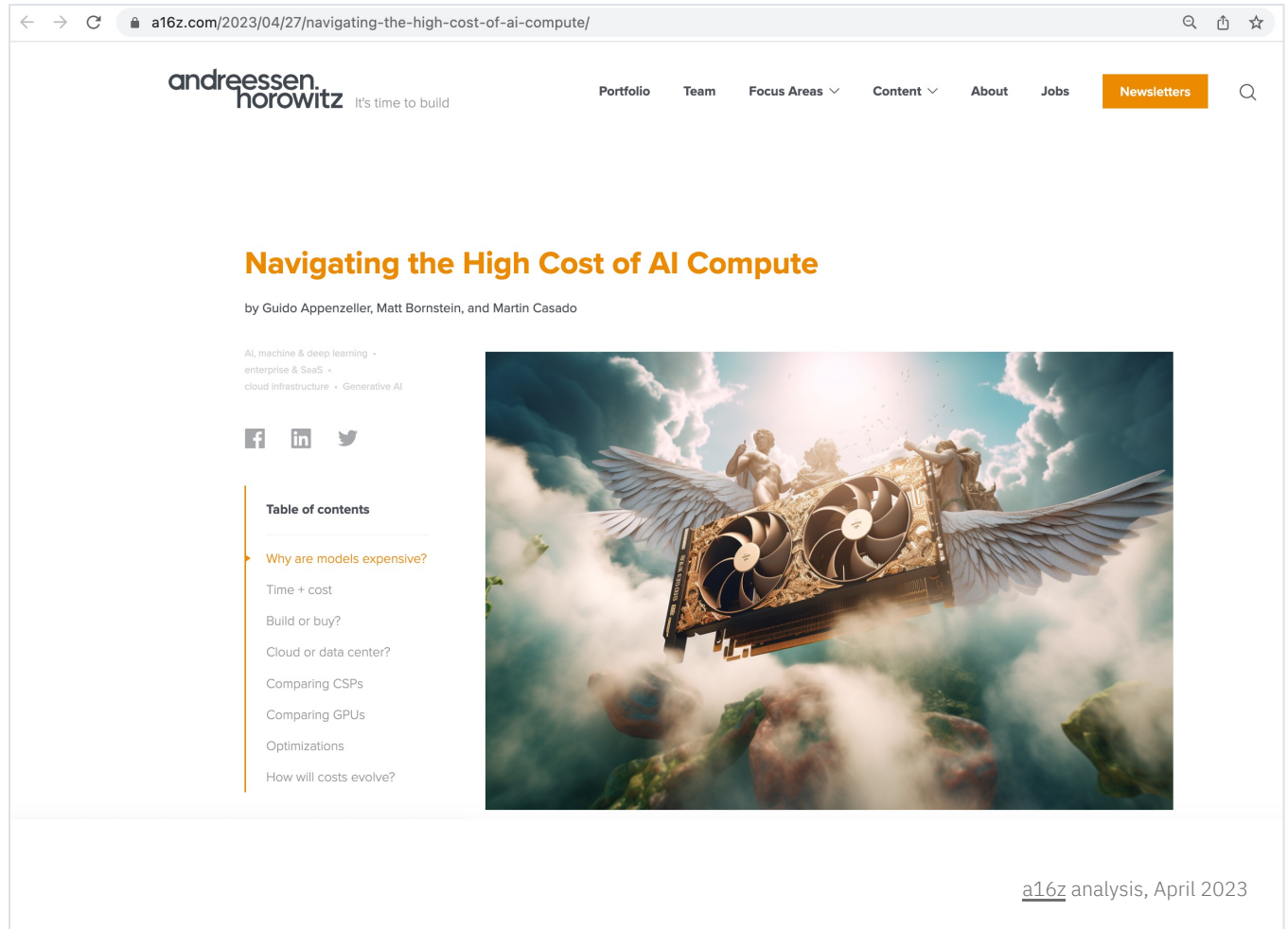


Able to effectively complete **multiple downstream tasks** while meeting accuracy baselines (e.g., flood mapping, land cover classification, outage prediction)

Note: while transformer architecture is most prevalent in foundation models, definition not restricted by model architecture

The flip side

“So, we think it’s fair to say that, right now, access to compute resources — at the lowest total cost — has become a determining factor for the success of AI companies.”



The screenshot shows a web browser displaying an article from a16z.com. The URL is a16z.com/2023/04/27/navigating-the-high-cost-of-ai-compute/. The page features the Andreessen Horowitz logo and navigation links for Portfolio, Team, Focus Areas, Content, About, and Jobs. A prominent orange 'Newsletters' button is also visible. The article title is 'Navigating the High Cost of AI Compute' by Guido Appenzeller, Matt Bornstein, and Martin Casado. The article is categorized under AI, machine & deep learning, enterprise & SaaS, cloud infrastructure, and Generative AI. Social media sharing icons for Facebook, LinkedIn, and Twitter are present. A 'Table of contents' section lists the following topics: 'Why are models expensive?' (highlighted), 'Time + cost', 'Build or buy?', 'Cloud or data center?', 'Comparing CSPs', 'Comparing GPUs', 'Optimizations', and 'How will costs evolve?'. The main image is a conceptual illustration of a GPU with large white wings, appearing to fly through a sky filled with clouds and a bright sun. Below the GPU, there are several large, colorful, abstract shapes resembling data points or clouds.

andressen horowitz It's time to build

Portfolio Team Focus Areas Content About Jobs Newsletters

Navigating the High Cost of AI Compute


by Guido Appenzeller, Matt Bornstein, and Martin Casado

AI, machine & deep learning · enterprise & SaaS · cloud infrastructure · Generative AI

Facebook LinkedIn Twitter

Table of contents

- Why are models expensive?
- Time + cost
- Build or buy?
- Cloud or data center?
- Comparing CSPs
- Comparing GPUs
- Optimizations
- How will costs evolve?



a16z analysis, April 2023

Optimizing the infrastructure for Foundation Models

Across the whole AI workflow

Data preparation



e.g., remove hate and profanity, deduplicate, etc.

Distributed training and model validation



Long-running job on massive infrastructure

Model adaptation

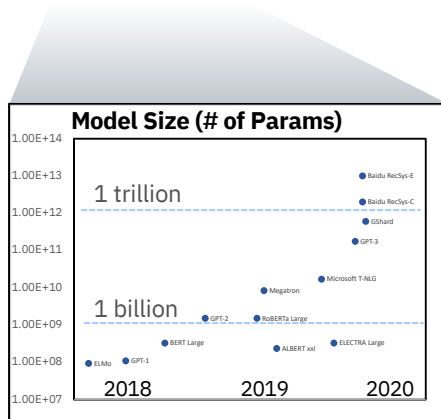


Model tuning with custom data set for downstream tasks

Inference



May have sensitivity to latency/throughput, always cost-sensitive

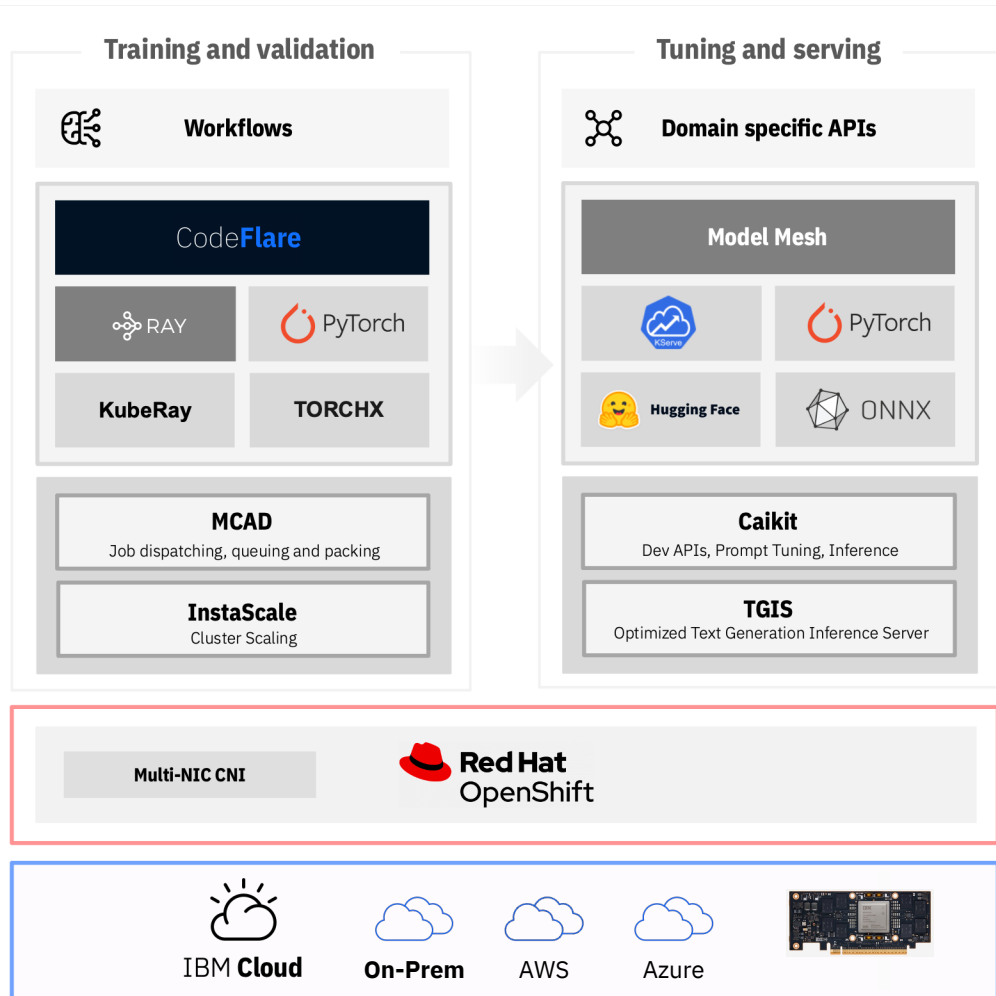


Building the FM technology stack

Middleware that simplifies end-to-end AI workflow and optimizes use of underlying infrastructure

Platform that deliver portability and abstracts infrastructure complexity

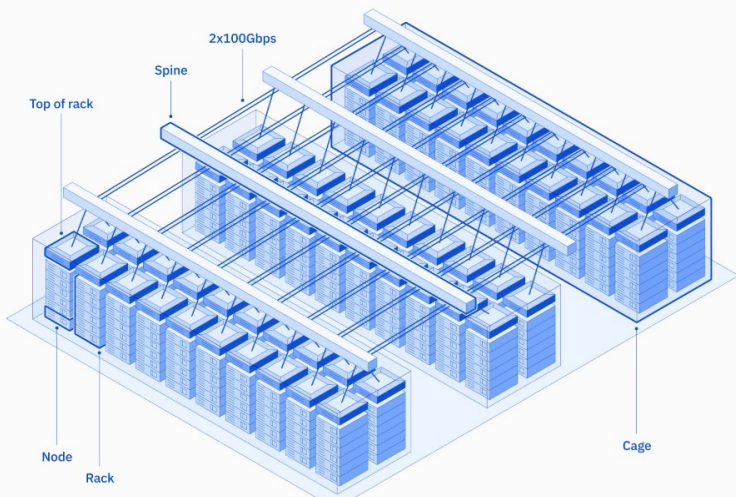
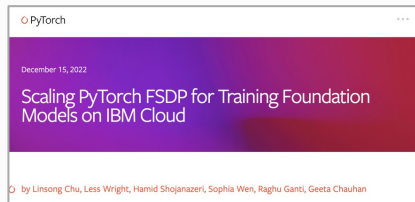
World-class infrastructure for training, tuning and serving foundation models (on-prem and in the cloud)



AI-optimized infrastructure

Training: Vela

Cloud-native design for large-scale distributed model training



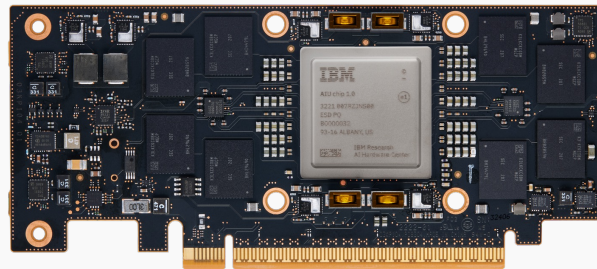
<https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster>

Inference: IBM AIU

Designed for energy-efficient AI compute at reduced precision

Precision Sparsity	BERT-base (F1%)	Wav2vec2.0 (WER %)	ViT (Accuracy %)
FP32	88.69	4.20	84.12
INT8	88.35 (-0.34)	3.85 (+0.35)	82.47 (+0.35)
INT8+50%Sp	87.70 (-0.99)	4.21 (-0.01)	84.03 (-0.09)
INT4	87.86 (-0.83)	4.53 (-0.33)	83.49 (-0.63)
INT4+50%Sp	87.07 (-1.62)	4.65 (-0.45)	83.60 (-0.52)

N. Wang et al, NeurIPS 2022

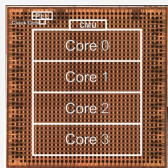


<https://research.ibm.com/blog/ibm-artificial-intelligence-unit-aiu>

IBM Research AIU background

Gen-3 AI Core Prototype

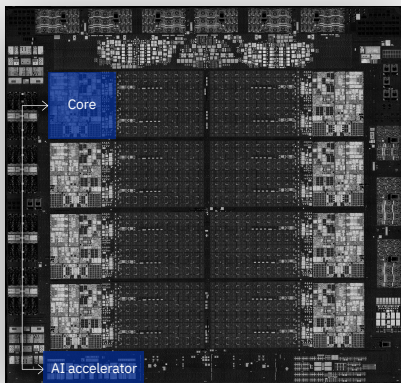
2019



IBM z16 Telum Chip

2022 GA

1 Gen-3 AI Core



zAIU

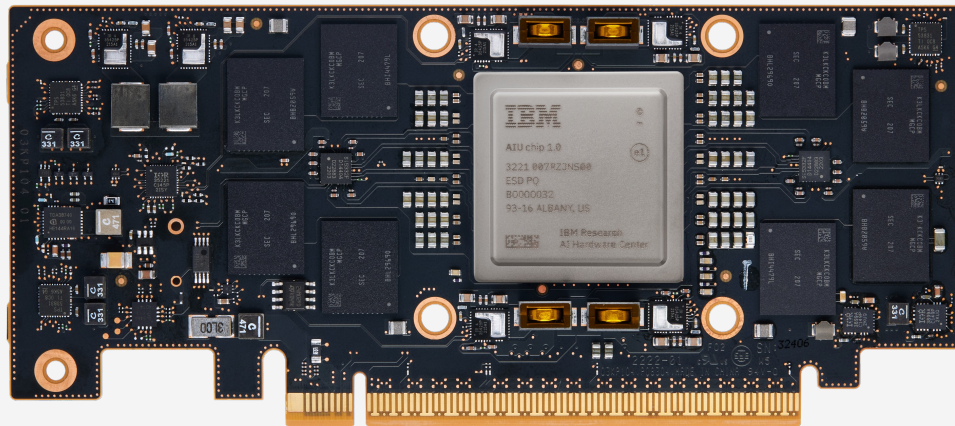
zAIU overview:

- One Gen-3 AI core, integrated in the z16 processor chip
- Off-loads AI tasks from the 8 CPU cores
- Optimized for in-transaction AI inferencing
- Seamless integration into z software stack

AIU (Artificial Intelligence Unit)

2022

32 Gen-3 AI Cores



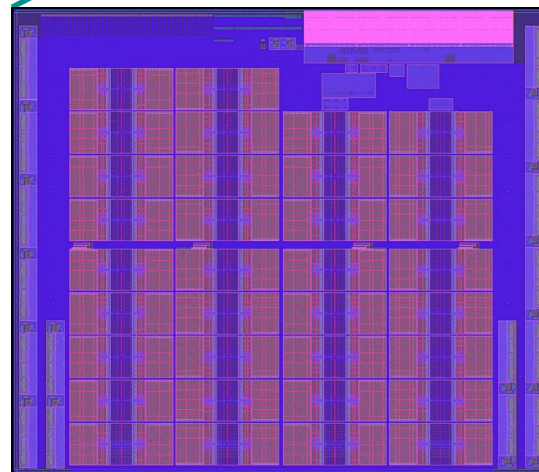
AIU overview:

- Complete AI accelerator, plugs into a standard PCIe slot
- 32 Gen-3 AI cores
- Optimized for AI inferencing, supports all operations for fine-tuning and training as well
- Designed to ease cloud integration, enabled in Red Hat stack
- Support for all common neural network types

IBM Artificial Intelligence Unit (AIU)

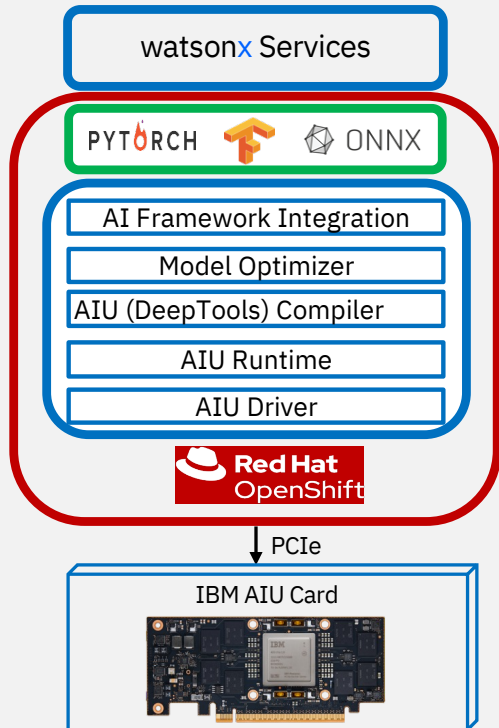
SoC implements IBM's leadership innovations in **low-precision** AI arithmetic and algorithms

- Chip architecture optimized for **enterprise AI** workloads, including foundation models
- Enabled in the **Red Hat** and **Foundation Models** software stacks
- Supports multi-precision inference (and some training) **FP16, FP8, INT8, INT4, INT2**
- Implemented in leading edge **5nm** technology

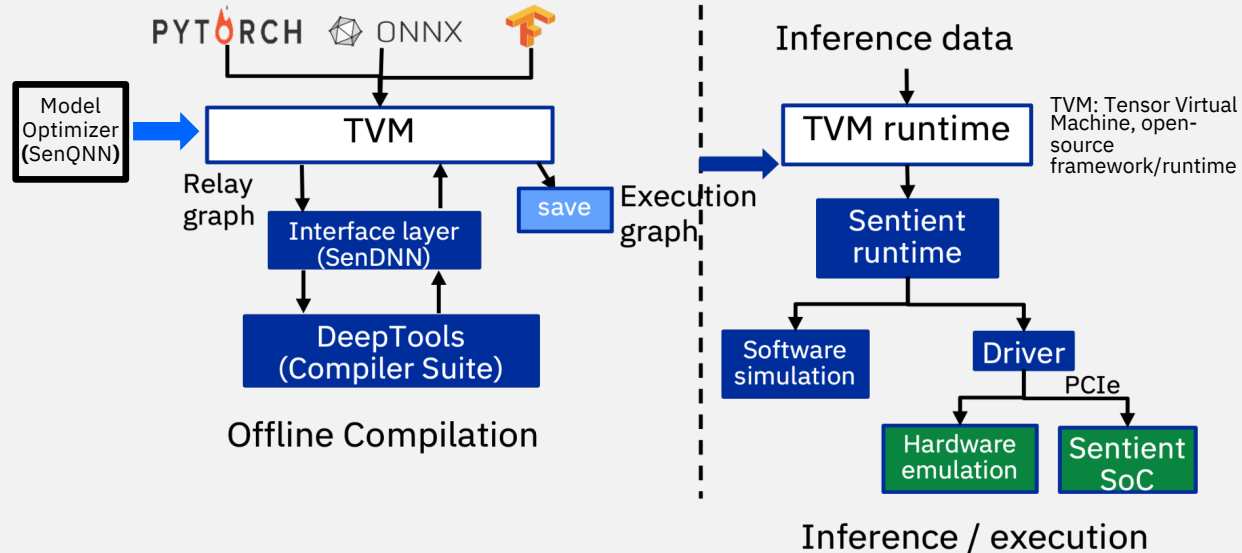


IBM AIU inference stack integrated with watsonx

User's view: watsonx services (only)



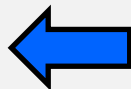
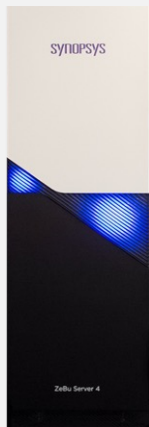
Internal software architecture components



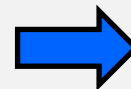
Key challenge: develop the entire AIU software stack in parallel with developing the SoC and PCIe card

IBM AIU emulation overview

- **Emulation systems have been essential for:**
 - Hardware verification: Uncover functional/performance bugs
 - **Software development:** Provide platform for chip internal/external software development



IBM AIU



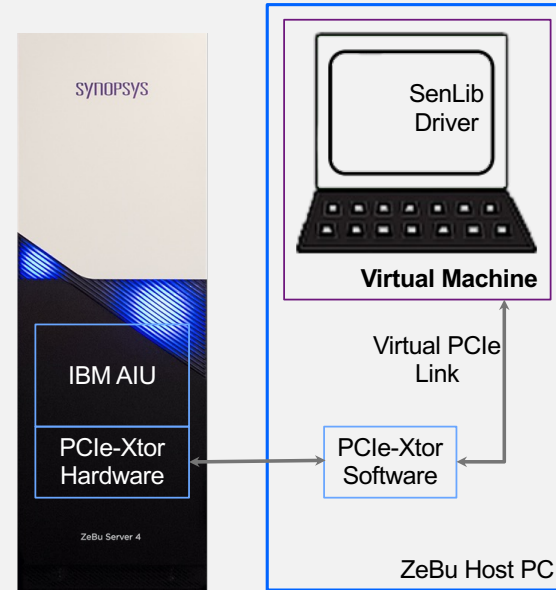
Synopsys HAPS

- Synopsys ZeBu**
- 96 Xilinx VU440 FPGAs
 - Hardware verification
 - Compiler / hardware co-development

- 4-8 Xilinx VU440 FPGAs
- Device driver development

Full AIU computational emulation

- Objective: high-fidelity model of **all computational elements – cores and interconnect** – of the SoC
- Model build:
 - ZeBu system from Synopsys
 - 96 Xilinx VU440 FPGAs
 - Very high fill rate, ~90% LUT utilization
 - 24h model build time (RTL to bitfiles)
 - 1 – 1.5 MHz operating frequency; limited by memory interface
- Impact highlights:
 - Found several high impact hardware bugs
 - Rare, hard to hit scenarios, practically impossible to find in simulation
 - Vital for compiler development
 - Complete cycle-accurate processing of 1 image: **1 min on ZeBu vs. 9 hours in simulation**



Example	
Number of different NNs exercised	14
Tests run (32 images/features per run)	100,000
Image/feature inferences completed	3.2 million
Total emulation run time	7000 hours
Equivalent SoC run time	7 hours

AIU nest emulation

Why a second emulation platform?

- **Develop device driver stack for AIU:** require SoC-like hardware fidelity (e.g., host-PCIe interface)

Platform and model details:

- HAPS system from Synopsys
 - SoC faithful nest + 1 AI core (vs 32 AI cores)
 - Running at MHz speed
- Includes PCIe Gen5 PHY daughter card from Synopsys
- Includes DDR4 DIMMs
- Uniquely suited for AIU driver development
 - Faithfully realizes the host-PCIe interface of the SoC



Network	HAPS runtime (sec/image or sec/feature)	ZeBu runtime (sec/image or sec/feature)
ResNet50	1.46	10.02
MobileNetV1	0.59	3.37
InceptionV4	4.35	43.76
BERT-large (seq=384)	67	292

Modeling and emulation impact

- Multiple software and FPGA-based methods have been essential to IBM's full-stack AIU and AI system development
- Our SoC design process leverages multiple levels of simulation for architecture development, logic and chip design, and design verification
- Our software stack development, accelerator software integration development, and compiler / hardware co-optimization leveraged FPGA-based emulation systems
 - Full-chip emulation via ZeBu for full-chip performance & accuracy analyses of AI models on multi-core models, compiler optimizations, architectural modifications and power estimation
 - Detailed SoC nest emulation via HAPS for device driver development, low-level software stack development, and evaluation of multi-chip configurations
- These methods enabled us to develop a full system, end-to-end hardware and software stack for Foundation Model inference **in parallel to** SoC and PCIe card development

Foundation Models are an inflection point for enterprise AI

- FMs enable a **proliferation of task-specific models**, but with **large** and **escalating** compute demands
 - Inference, fine-tuning, and distributed training systems differing in requirements
 - Full-system innovation is required
- Our approach emphasizes:
 - Cloud-native architectures
 - Ease-of-use for developers and clients
 - Hybrid cloud consumption
 - AI accelerator design and technology innovations

