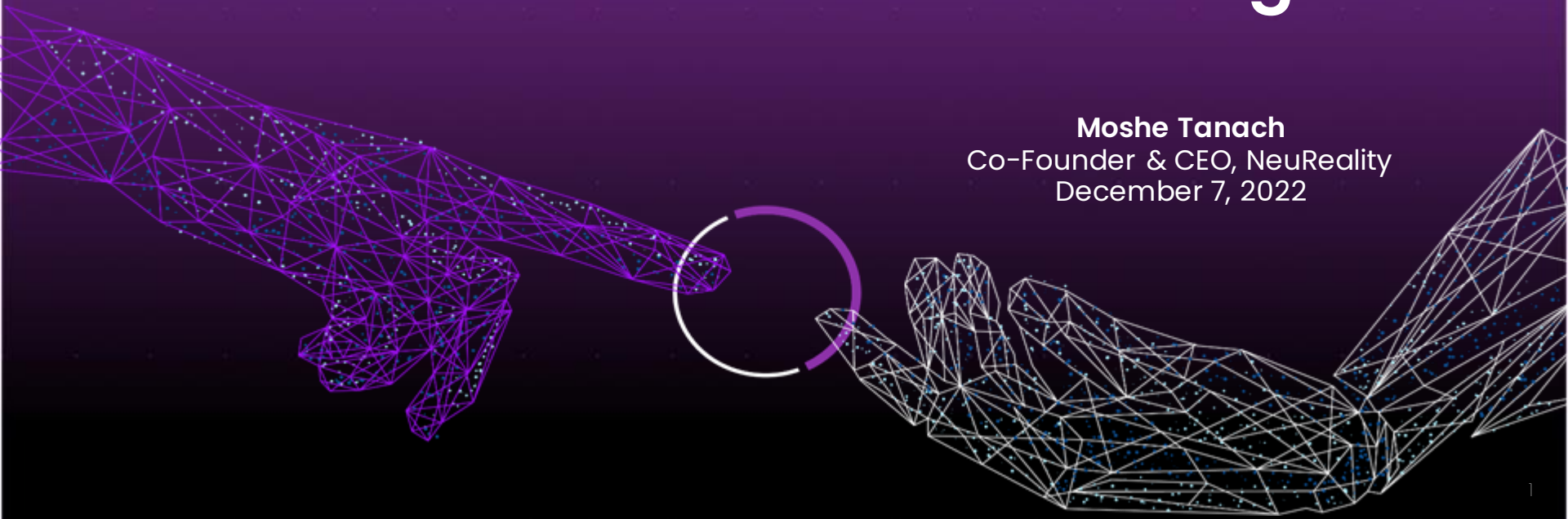
NEUREALITY

Deploying any Inference Use Case with Network Attached Processing Units



Moshe Tanach
Co-Founder & CEO, NeuReality
December 7, 2022

AI Inference

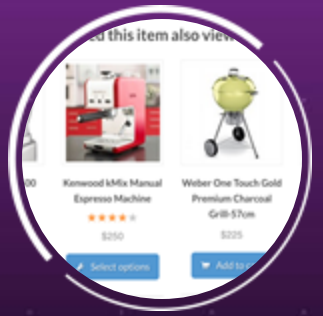
AI is Evolving with Deep-Learning

Computer Vision

Smart city, retail, industrial, Automotive

Natural Language Processing

FSI, Customer service, Digital assistance



Recommendation Engines

eCommerce, Social networks, Online media

Enormous Opportunity to Scale Inference…

Financial Services AI Use Cases

Credit Decisions

Fraud Detection

Process Automation

Risk Management

Algorithmic Trading

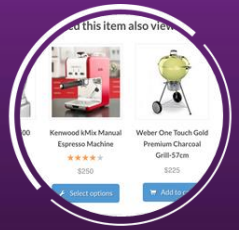
Customer Service & Call Centers

Challenges to Scale AI Inference Usage

..but many AI possibilities can't be realized and deployed due to the major complexity and cost barriers of today's infrastructure..

Computer Vision

Natural Language Processing



Recommendation Engines

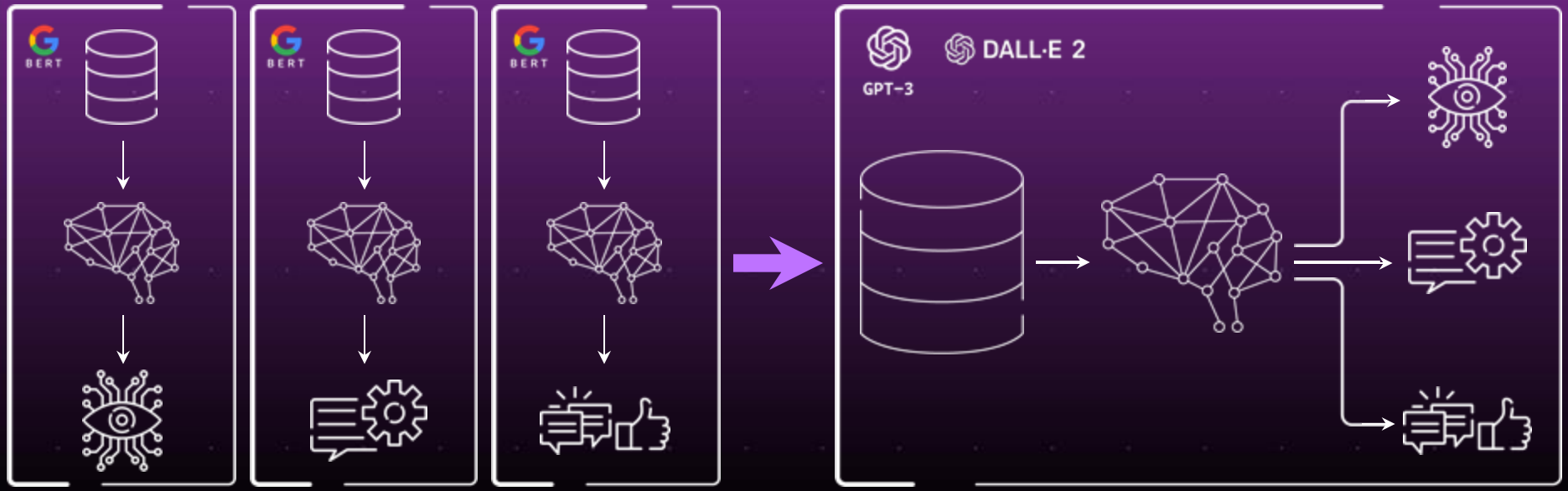
✗ Existing AI solutions, built for Training are not optimized for inference

✗ CPU-centricity introduces System bottlenecks, and high cost and power consumption

✗ Deployment complexities of AI set high barriers to entry

NEUREALITY

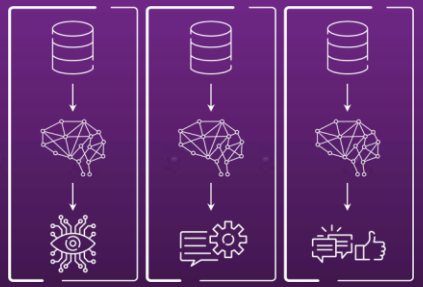
NLP for Customer support and Call Centers



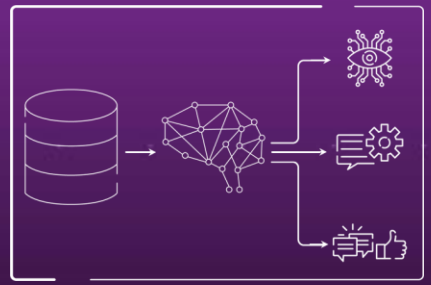
Task-Specific

Foundation Models

NLP for customer support and call centers



Task-Specific



Foundation Models

System Complexities

(beyond deep learning processing)

- ✓ Complementing Model processing
- ✓ Model Loading and Switching
- ✓ Quality of Service (QM/SCH/LB)
- ✓ Data Movement
- ✓ Data Processing (pre/post)

+

- ✓ Model Parallelism
- ✓ Intra-Model Data Movement
- ✓ Intermediate Data Processing

NEUREALITY

NeuReality Architecture

NeuReality AI-as-a-Service

The first system solution optimized for inference

NeuReality is enabling AI everywhere by offering a holistic solution for inference deployment that lowers cost, complexity & power consumption with a revolutionary new AI-centric architecture, SDK, and APIs

Architecture

Optimizing end-to-end experience & efficiency for Inference-as-a-Service

Hardware

New generation of Network Addressable Processing Units

Software

Multi-layer SDK & Runtime for easy deployment of AI-pipelines

APIs

Simplified UX for development

From Data Centers to Near-Edge

Data Center

On-Premise

Near-Edge

Edge

NEUREALITY

AI-Centric Architecture
Maximize the Power of DLAs with NAPU's

- ✓ Purpose-built for AI Inference
- ✓ Ultra Scalability
- ✓ Lowest Cost and Power consumption
- ✓ Zero System Bottlenecks

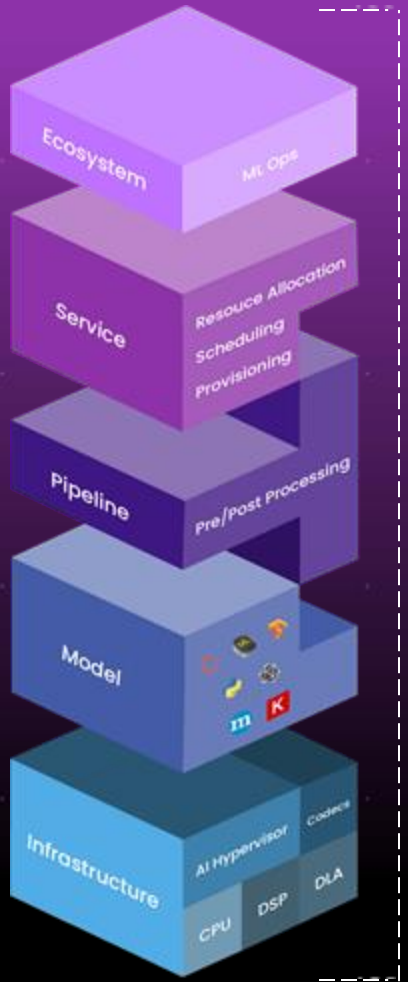
NEUREALITY

NeuReality Software

NeuReality Software

NeuReality provides three unique levels of value add to inference that don't exist today with software tools and runtime that:

- **Model:** NeuReality provides holistic AI inference model execution which enhances the inference system to handle any trained model
- **Media Processing:** NeuReality provides full AI pipeline offload tools and runtime for processing the media
- **Interface:** NeuReality provides server interface application that connect to any environment to provide inference service



All wrapped in a single, easy to use UI/UX

NEUREALITY

NeuReality Hardware

Range of NeuReality AI-Centric Offerings

NR1-P
AI-centric Prototype

The first AI-centric server with AMD FPGA

NR1-M
CPU centric Server

Augment OEM systems with a PCIe module

NR1-S
AI centric Server

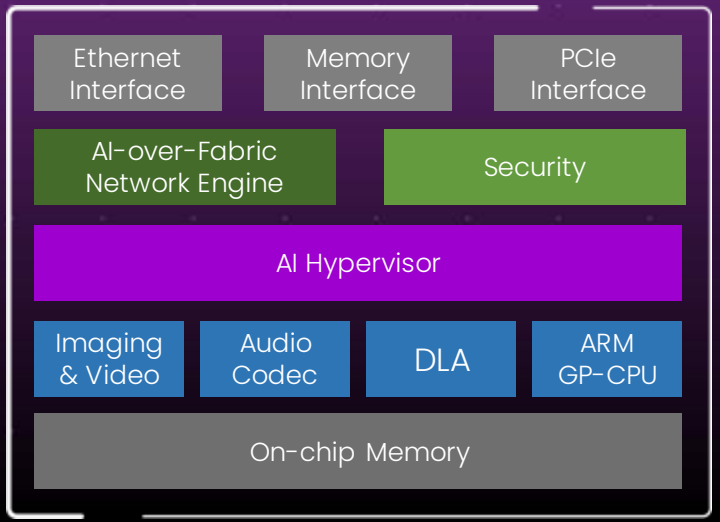
Next generation CPU-less server with highest density of NAPU+DLA technology

AMD IBM Lenovo SUPERMICRO

NEUREALITY

NR1 - Network Addressable Processing Unit

First AI-centric Inference Server-on-a-Chip



NEUREALITY

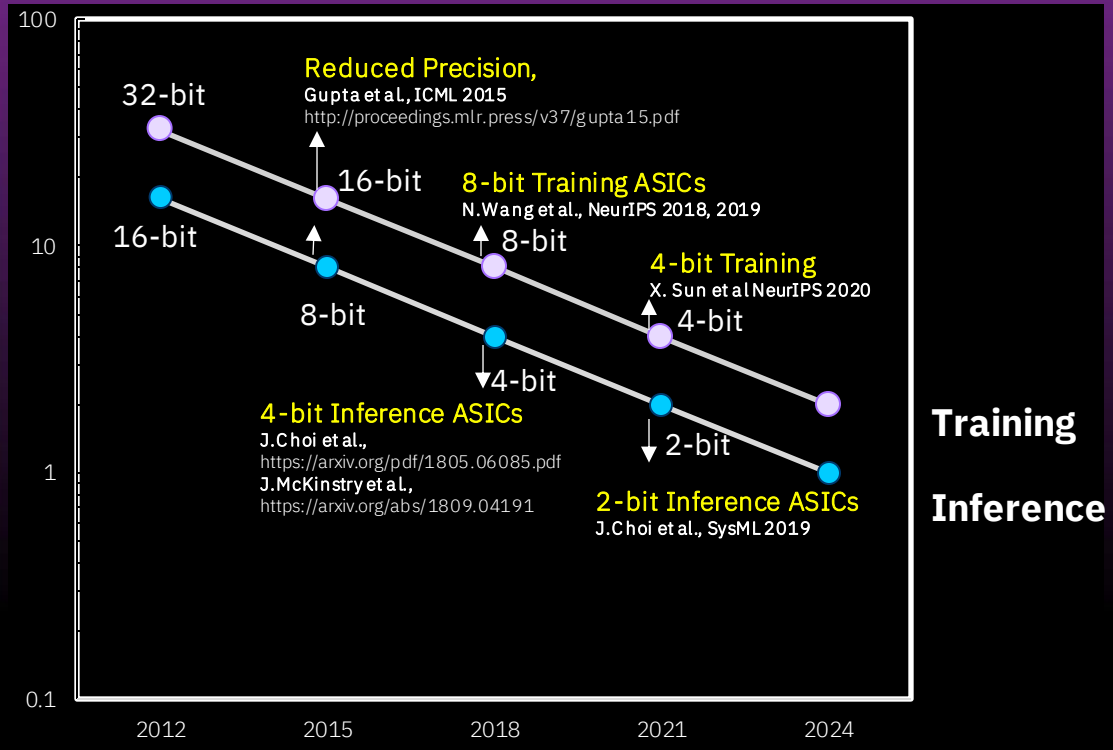
Dr. Mukesh Khare

Vice President, Hybrid Cloud

IBM

Driving reduced precision *with iso accuracy*

IBM Research leadership in reduced precision arithmetic for AI



Extending performance by 2.5X / year

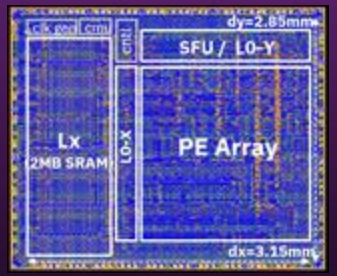
Approximate computing principles applied to **Digital AI Cores** with reduced precision

Analog AI Cores, which could potentially offer another **100x in energy-efficiency**

IBM Research AI Hardware Journey: *Path to Product*

2018

First Gen Reduced Precision Core



14nm Technology

2019

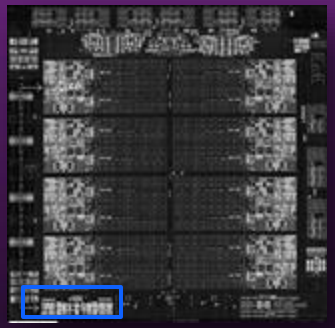
AI HW Center Launch

"IBM Invests $2 Billion in New York Research Hub for AI"
Bloomberg

Full stack approach

2021

AI Core in IBM z16 **Telum** processor



7nm Product

On-chip AI accelerator enables real-time data insights for applications such as **fraud detection**

IBM's New Telum Chip Reboots the Mainframe ›Big Blue's z16 computer—and the cache-savvy design at its core—gives new relevance to the platform

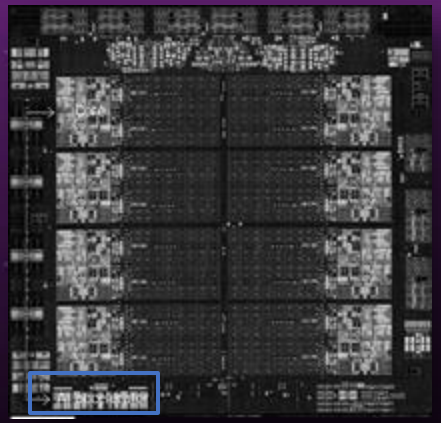
IEEE Spectrum

April 29, 2022

IBM digital accelerator supports diverse use cases *with flexible integration style*

IBM Telum processor

AI Core integrated in processor



7nm Technology

Single AI core

300B inferences per day with ***1 ms*** latency

IBM Artificial Intelligence Unit (AIU)

Flexible system integration as PCIe card, 5nm Technology, Designed for any Cloud

NeuReality NR1M

Stand-alone AI inference system


NEUREALITY
We Make AI
EASY

NeuReality AI-Centric Solutions

NeuReality takes care of everything needed
for a streamlined deployment of AI at scale

NR1 NAPU

NR1-M PCI Module

NR1-S Server

NeuReality Makes AI Easy

Hardware and software solutions that optimize AI inference usage and
Make setup easy for both inexperienced and sophisticated users

✓ EASY TO DEVELOP

- Helps implement complex customization and optimization of models
- Serves customers with various levels of experience

✓ EASY TO INTEGRATE

- APIs can easily be integrated into existing IT infrastructures
- AI developers can continue working within their favorite environment

✓ EASY TO DEPLOY & RUN

- Takes care of deploying and provisioning compiled models
- Sets up clients to run inference tasks without special IT expertise

✓ EASY TO SCALE

- Enables scaling and optimizes inference capacity based on usage
- Reduces cost of adoption for AI services

With NeuReality, the Future for AI-based Financial Services can now be Realized

Credit Decisions

Fraud Detection

Process Automation

Risk Management

Algorithmic Trading

Customer Service & Call centers

Call to Action

- Adopting a new AI-centric architecture provides a huge advantage for handling AI use cases from small task-specific models to huge foundation-models

  - Visit our booth #215 at The AI Summit New York to learn more about our solution for AI inference
  - Join our Early Adopter Program to evaluate financial services AI applications
  - Visit www.neureality.ai

NEUREALITY

We Make AI EASY