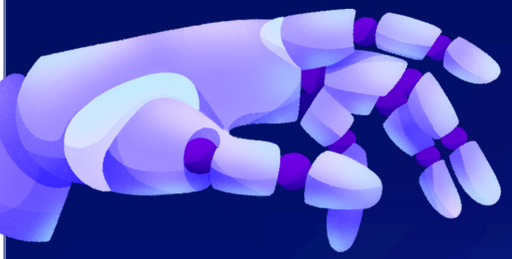
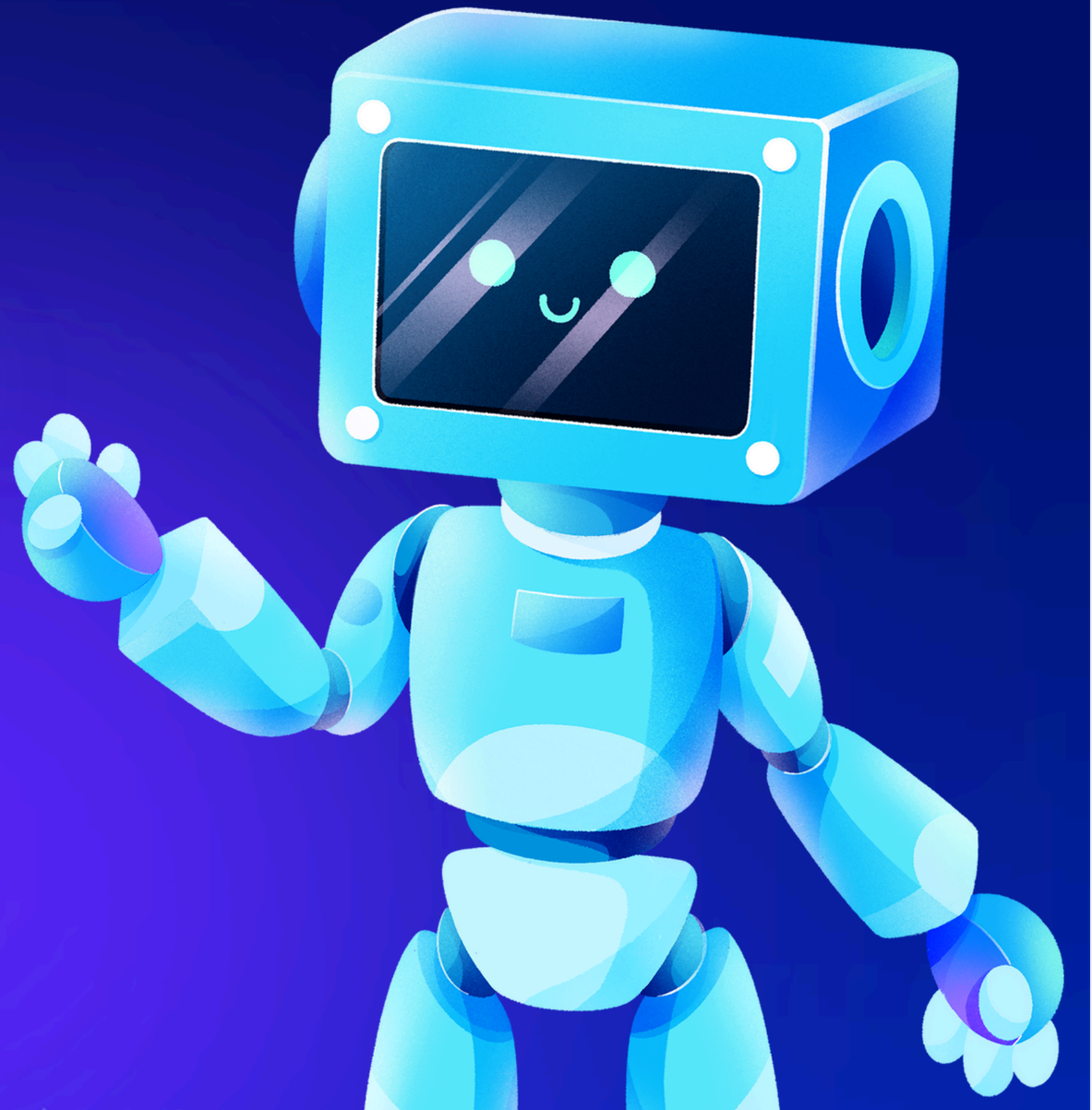


NEURAL PROCESSING UNITS:THE FUTURE OF AI

Kamdon Bird, Chris Darnell, Jaylin Mendoza, Mark Timpson



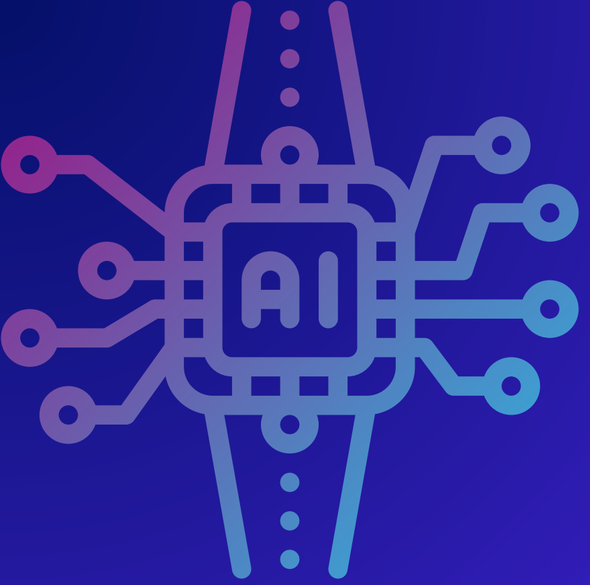
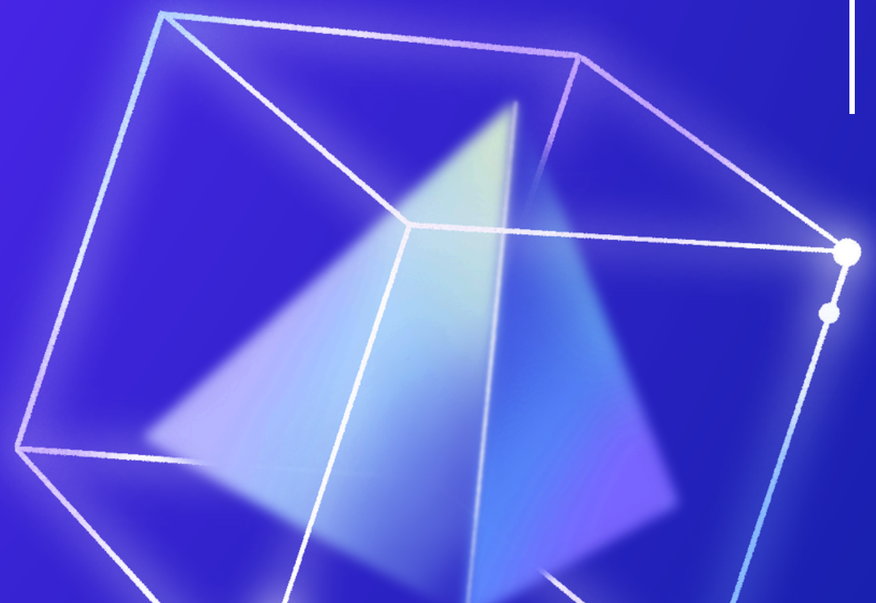
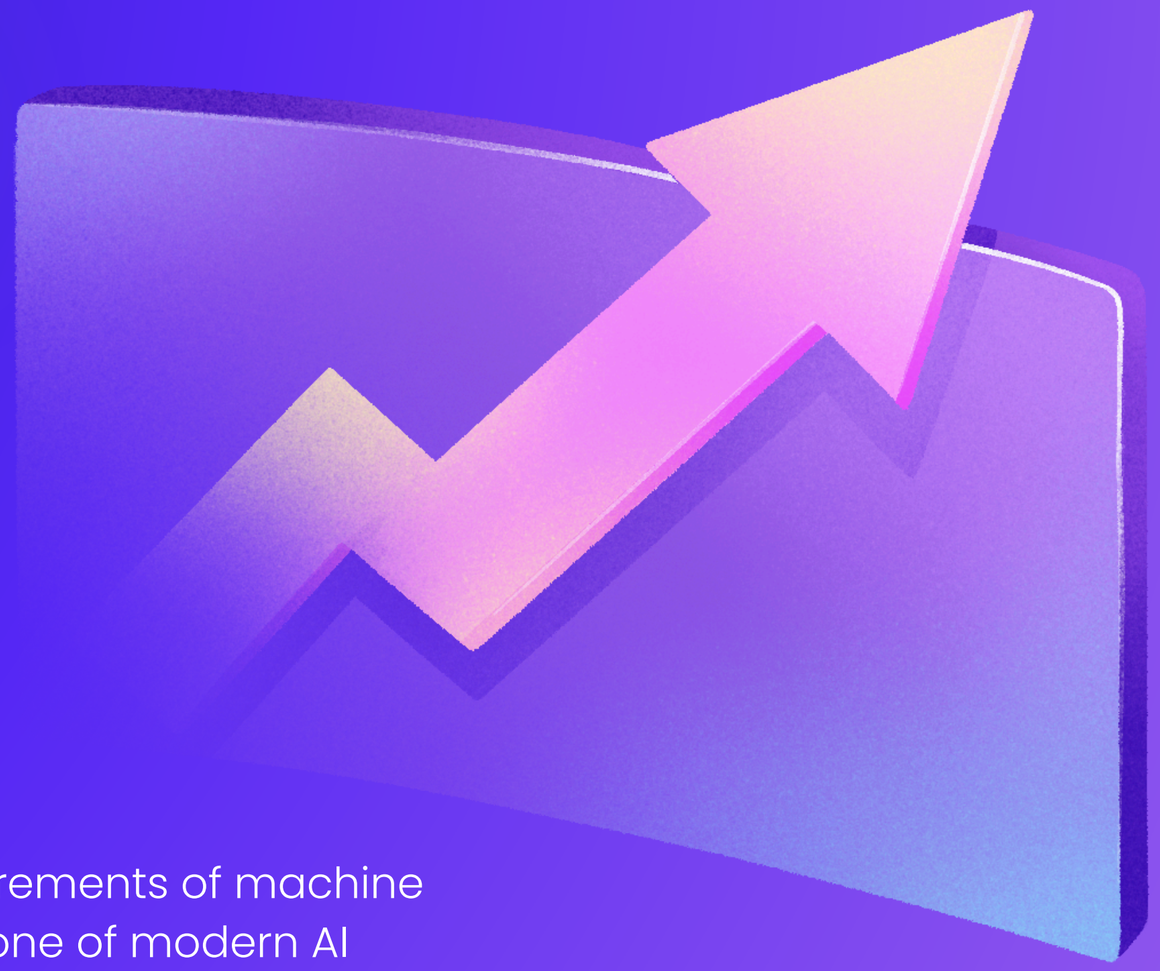


TABLE OF CONTENTS

- Introduction 01
- What is a Neural Network 02
- How NPUs Work 03
- Applications of NPUs 04
- Benefits of NPUs 05
- Current Challenges 06
- Key Takeaways 07



INTRODUCTION



A neural processing unit (NPU), also known as an AI accelerator, is a specialized microprocessor designed to perform neural network and machine learning computations faster than other processors.

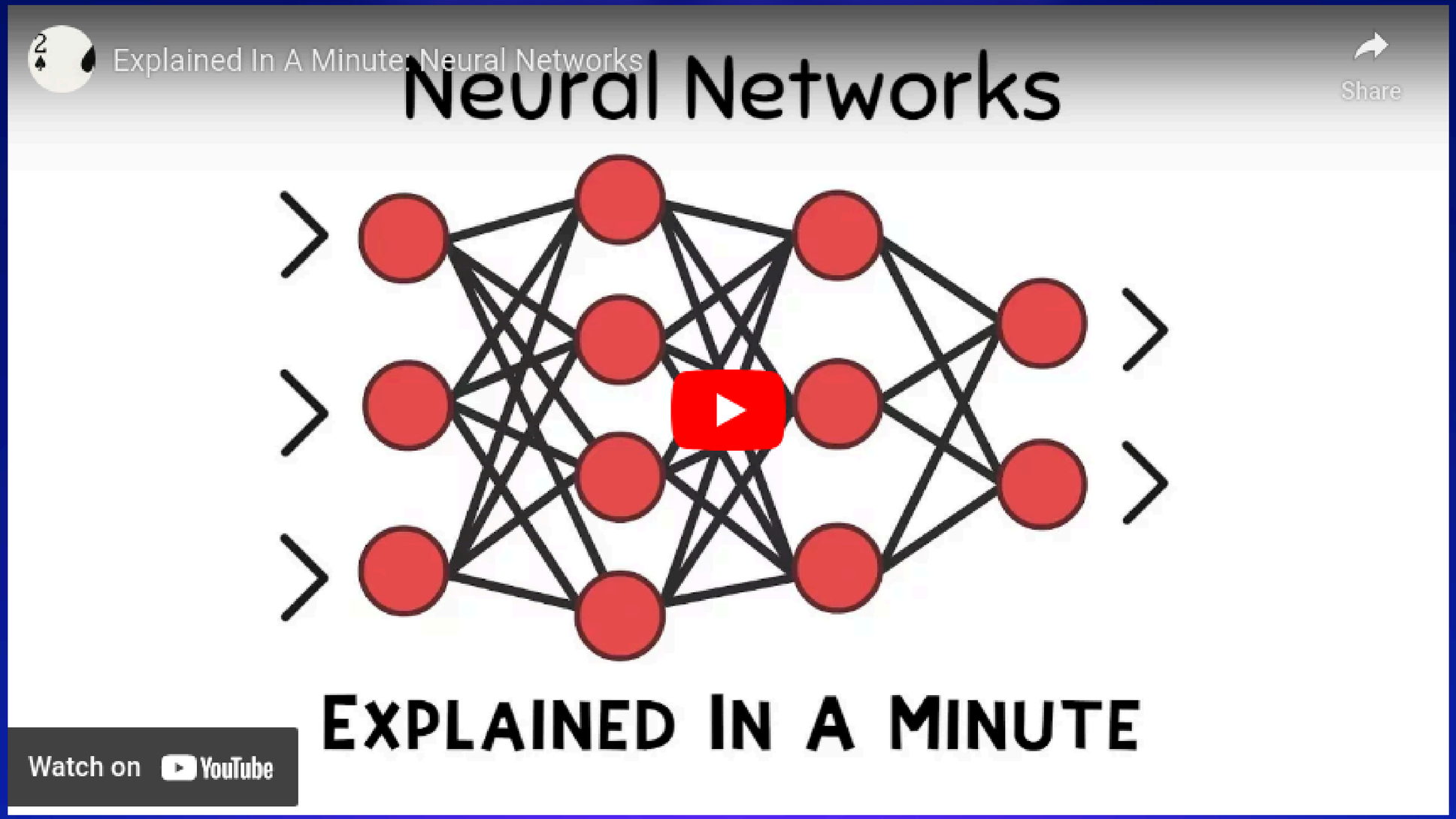
Artificial Intelligence

By focusing on the specific computational requirements of machine learning models, NPUs have become a cornerstone of modern AI development.

NEURAL NETWORKS

A neural network is a model for machine learning. It simulates a human brain, with artificial neurons interconnected by weighted connections. The system can be trained on inputs and outputs to strengthen or weaken the connections between neurons, which allows computer-generated predictions without explicit instructions.

NEURAL NETWORKS



HOW NPUS WORK (OVERVIEW)

Specialized Design

Unlike traditional processors such as Central Processing Units (CPUs) or Graphics Processing Units (GPUs), NPUs are tailored for the unique demands of artificial intelligence (AI) workloads. They are optimized for executing tasks like image recognition, natural language processing, and autonomous decision-making with exceptional speed and efficiency.

Key Features

An NPU uses parallel processing, low-precision arithmetic, high-bandwidth memory, and hardware acceleration to perform AI tasks quickly and efficiently. These attributes are crucial for the real-time processing required in AI applications.

APPLICATIONS | CONSUMER DEVICES

SMARTPHONES

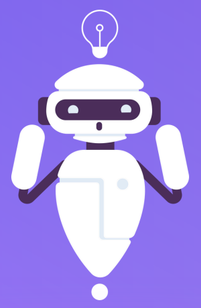
AI-powered photo processing, voice recognition.

SMARTWATCHES

Health tracking, voice assistants.

IOT(INTERNET OF THINGS) DEVICES

Real-time decision-making.



AUTONOMOUS VEHICLES

incoming data from multiple cameras, LiDAR, and other sensors for tracking multiple objects in real-time

BENEFITS

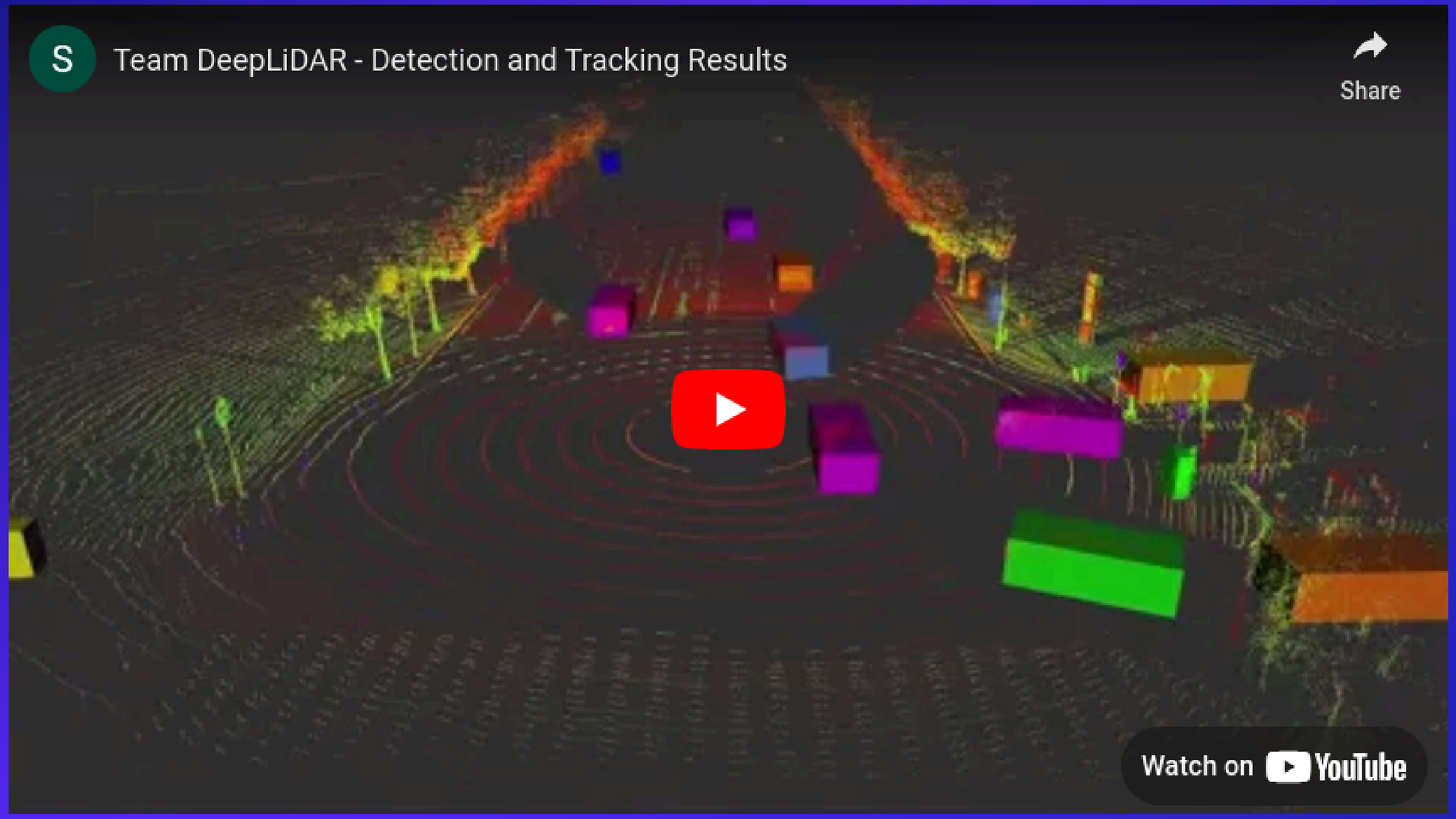
01

- For Consumers:
- Improved energy efficiency.
- Enhanced user experience with tools like:
- Improved graphics.
- Advanced note-taking tools (e.g., OneNote that summarizes and creates to-do lists).
- Live captions for real-time audio transcription and translation.

02

- For Businesses:
- Efficient systems for critical operations.
- Cost reduction through better task execution.
- Dedicated processing power for crucial business applications.
- Competitive advantage in adopting AI-powered systems.

TRACKING MULTIPLE OBJECTS



CURRENT CHALLENGES

01

02

03

- Limitations
  - Not suitable for general computing
  - Limited blockchain use (cyptocurrency)
  - Mainly focused on AI and machine learning
- Integration
  - Requires significant changes to existing systems
  - Businesses need AI expertise for optimal performance
- High Costs
  - Custom-built, requiring substantial research and development
  - Expensive for smaller companies to adopt and maintain

TAKEAWAYS

- What NPUs Are
  - NPUs are specialized processors designed for AI tasks like neural network computation and machine learning.
  - They use parallel processing, low-precision arithmetic, and high-bandwidth memory for efficient AI computation.
- Benefits of NPUs
  - High performance, low latency, and efficient AI processing.
  - Power AI-driven features in devices like Copilot+ PCs, improving productivity, graphics, and cloud management.
- Applications
  - Real-time processing in autonomous vehicles and medical devices.
  - On-device AI tools, improving efficiency and user experience.
- Challenges
  - NPUs are expensive and require significant investment and expertise for integration.
  - They are limited to specific tasks, making them unsuitable for general-purpose computing.
- Conclusion
  - While NPUs excel in AI applications, they face integration and cost challenges, limiting their broader use in general computing.

THANK YOU!

RESOURCES

Patterson, D. A., & Hennessy, J. L. (2017). Computer organization and design: The hardware/software interface (5th ed., p. 502). Morgan Kaufmann.

Schneider, J., & Smalley, I. (n.d.). What is a neural processing unit (NPU)? IBM. Retrieved December 3, 2024, from https://www.ibm.com/think/topics/neural-processing-unit

Malatino, K. (2023, September 27). The future of AI is wide, deep, and large. Rensselaer Polytechnic Institute. Retrieved December 3, 2024, from https://news.rpi.edu/content/2023/09/27/future-ai-wide-deep-and-large

Chambers, P. (2024, July 2). Ai Golden Child: How AI PCs Will Change the Game for Enterprises. The Fast Mode. https://www.thefastmode.com/expert-opinion/36153-ai-golden-child-how-ai-pcs-will-change-the-game-for-enterprises.

Miller, J. (2024, October 16). Neural Processing Units (NPUs): The dawn of AI pcs. CIO. https://www.cio.com/article/3566907/neural-processing-units-npus-the-dawn-of-ai-pcs.html.

Nzube, S. (2024, July 2). NPUs: How They Are Changing the Computing Game with the New Copilot+ PCs. Auslogics Blog | Tech Wisdom Blog. https://www.auslogics.com/en/articles/npus-how-they-are-changing-the-computing-game-with-the-new-copilot-pcs/.

RESOURCES

Sheikh, R. A., Borkar, P. G., Sable, D. G., & Sugandhi, M. Y. (2024, 11 11). The Role of Neural Processing Units in Enhancing Real-Time Processing and Decision-Making in Autonomous Systems. International Journal of Humanities Social Science and Management (IJHSSM), 6(6), 213-217. https://ijhssm.org/issue_dcp/The%20Role%20of%20Neural%20Processing%20Units%20in%20Enhancing%20Real%20Time%20Processing%20and%20Decision%20Making%20in%20Autonomous%20Systems..pdf

Morrison, R. (2023, February 8). IBM launches Vela, its new cloud-native AI supercomputer. Tech Monitor https://www.techmonitor.ai/digital-economy/ai-and-automation/ibm-vela-cloud-native-ai-supercomputer?cf-view

Schneider, J., & Smalley, I. (2024, October 10). NPU vs GPU: Understanding the differences. IBM. https://www.ibm.com/think/topics/npu-vs-gpu

Clarke, H. (2024, May 22). AI hardware showdown: CPU vs GPU vs NPU. Harrison Clarke. https://www.harrisonclarke.com/blog/ai-hardware-showdown-cpu-vs-gpu-vs-npu