上海创智学院
Shanghai Innovation Institute

AI编程生态的历史与展望

History and Outlook of the Programming Ecosystem for AI

冯思远 | 上海创智学院

2025 年 10 月 25 日

SHANGHAI INNOVATION INSTITUTE

The Bridge between AI Model and Hardware

上海创智学院
Shanghai Innovation Institute

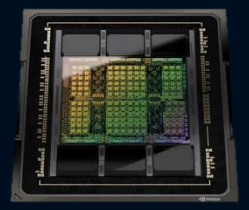
AI Models

Applications (Megatron, vLLM)

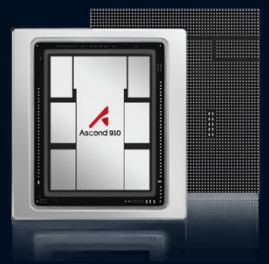
Frameworks (TensorFlow, PyTorch)

Kernels (CUTLASS, Triton, TileLang, PyPTO)

Programming Language (CUDA, Ascend C)



NVIDIA GPU



HUAWEI NPU

Mobile devices

Efficiency, Usability and Generality, a Trilemma

上海创智学院
Shanghai Innovation Institute

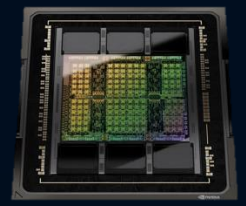
AI Models

Applications (Megatron, vLLM)

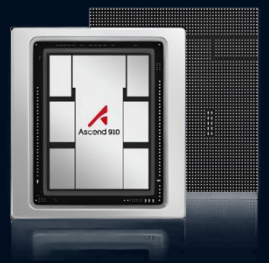
Frameworks (TensorFlow, PyTorch)

Kernels (CUTLASS, Triton, TileLang, PyPTO)

Programming Language (CUDA, Ascend C)

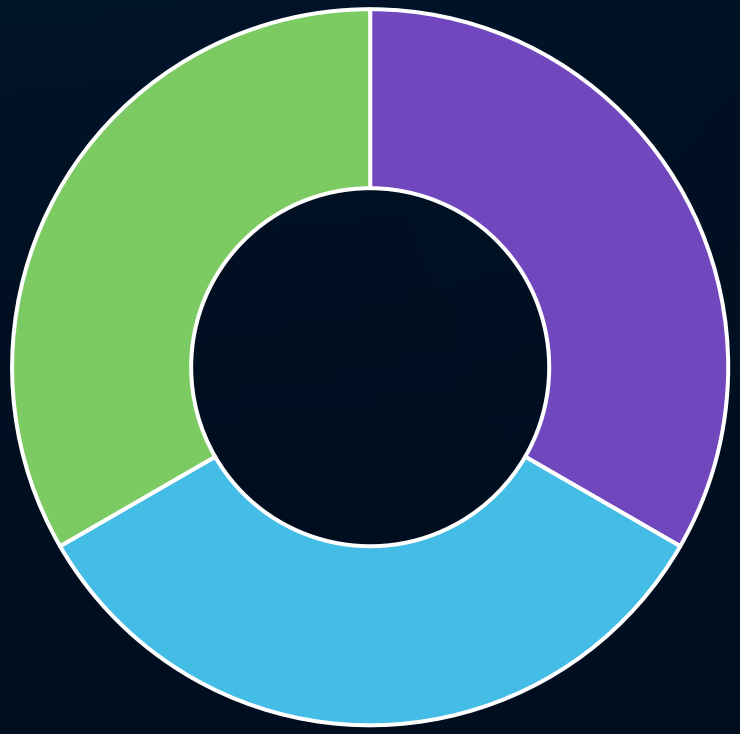


NVIDIA GPU



HUAWEI NPU

Mobile devices



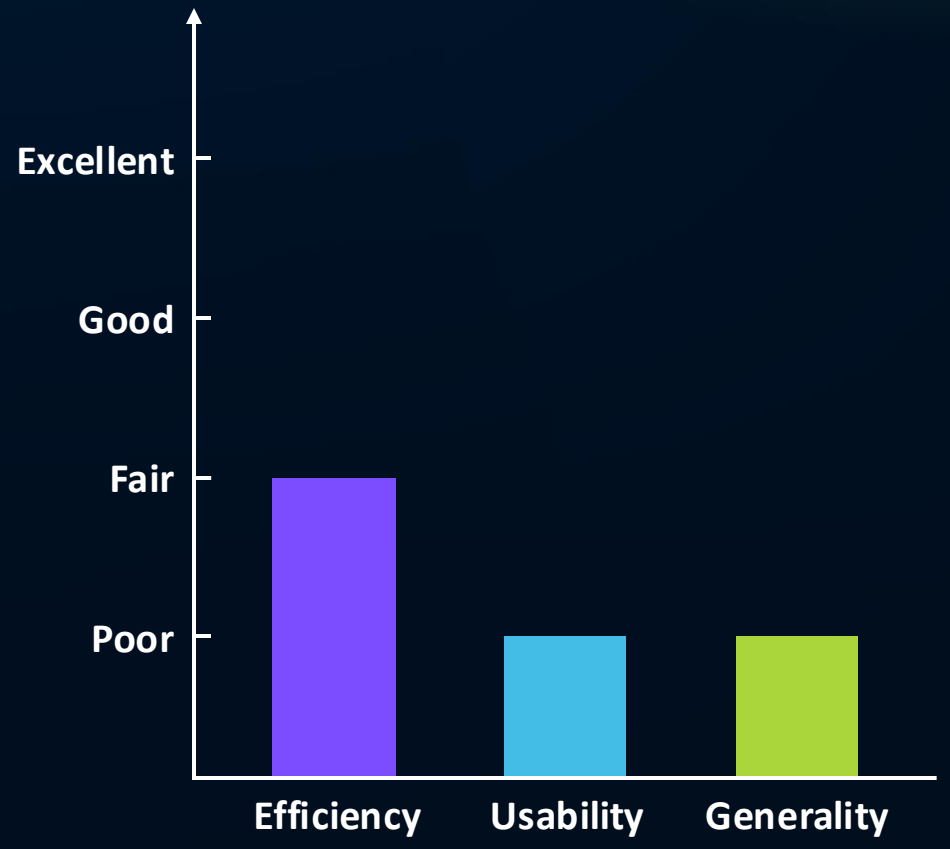
First Try: Train Network with Native CUDA

上海创智学院
Shanghai Innovation Institute

The breakthrough model AlexNet (2012) is a classic example of this era. It was trained using hand-written, native CUDA code.

This direct-to-hardware approach was the only way to achieve the performance needed for a breakthrough, but required heroic effort and deep GPU expertise.

Development was slow and complex, requiring researchers to write low-level C++ code to define even basic operations like convolutions. Besides, AlexNet's architecture and the specific NVIDIA GPUs it ran on, making it difficult to adapt or reuse for other models.



The Classic Era: From Raw Power to Usability

上海创智学院
Shanghai Innovation Institute

The early days were a clear trade-off.

Native CUDA gave us performance but was difficult to use and not portable.

TensorFlow and **PyTorch** solved this, creating frameworks with excellent Usability and Generality.

However, this came at a cost: the "eager mode" execution in PyTorch was too slow for production.

The Turning Point: LLMs Demand Extreme Efficiency

上海创智学院
Shanghai Innovation Institute

LLMs changed the game. Their massive scale made the inefficiency of classic frameworks unacceptable.

This forced a new trade-off across the entire stack—from libraries like **FlashAttention** to training and serving frameworks like **Megatron/ MindSpeed** and **vLLM**.

This incredible efficiency comes from highly-specialized, hand-tuned code.

However, the frameworks and libraries are specialized to LLM cases, with poor **Generality.**

The Path Forward: Model-Hardware Co-Design

上海创智学院
Shanghai Innovation Institute

The industry is moving away from sacrificing one goal for another. The future is co-design, where hardware and programming models are developed together.

This has led to new programming languages that seek a better balance in the trilemma.

Triton offers near-CUDA performance with a much more usable, Python-like syntax. While **TileLang** expose more low-level information to developers, achieve a better balance between **efficiency, usability and generality. PyPTO** leverage **MPMD** style to utilize NPU hardware with large single kernel.

Lessons Learned

上海创智学院
Shanghai Innovation Institute

01

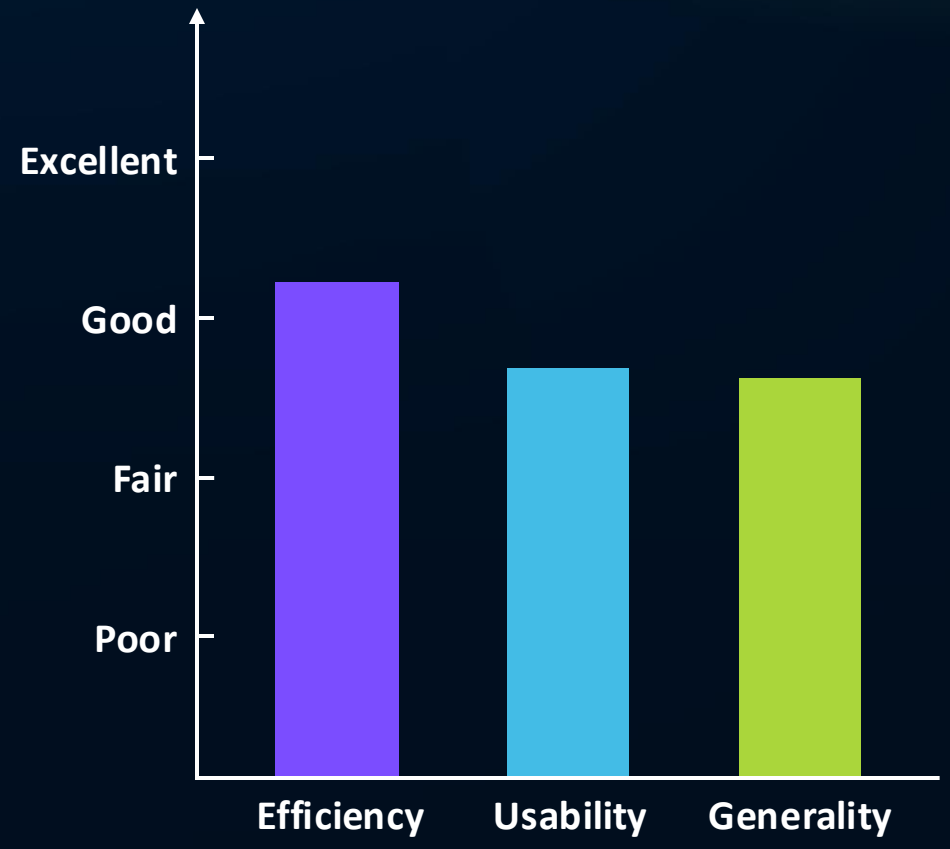
The Trilemma is Real

The tension between Efficiency, Generality, and Usability drives innovation. The ideal balance shifts with the needs of the era.

02

Specialization is Unavoidable

There is no "one size fits all" solution. The ecosystem will continue to have specialized tools for different tasks like training, inference, and research.



Future Opportunities

上海创智学院
Shanghai Innovation Institute

01

Deeper Co-Design, from Chip to Superpod

The concept of co-design will extend from a single chip to the entire AI supercomputer. Future programming languages will be built with awareness of the whole system, including the high-speed network fabric. This allows for optimizing not just computation, but also communication at the scale of an entire data center.

02

AI-Driven Programming

The ultimate solution to the usability challenge is to use AI itself. Future AI agents will automatically generate and optimize the complex kernels we write by hand today, freeing developers to innovate at the model level

上海创智学院
Shanghai Innovation Institute

AI编程生态的历史与展望

History and Outlook of the Programming Ecosystem for AI

冯思远 | 上海创智学院

2025 年 10 月 25 日

SHANGHAI INNOVATION INSTITUTE