yasp

Coding Agents for Compiler Construction: Beyond the AI Assistant Paradigm

PyTorch CONFERENCE EUROPE 2026

7-8 April 2026

The Challenge

up to 80%

of AI budgets allocated to infrastructure, software & related salaries

- **Vendor Lock-In**

  NVIDIA dependency limits hardware choice and inflates costs

- **Manual Optimization**

  Expert-dependent, time-consuming tuning of every model

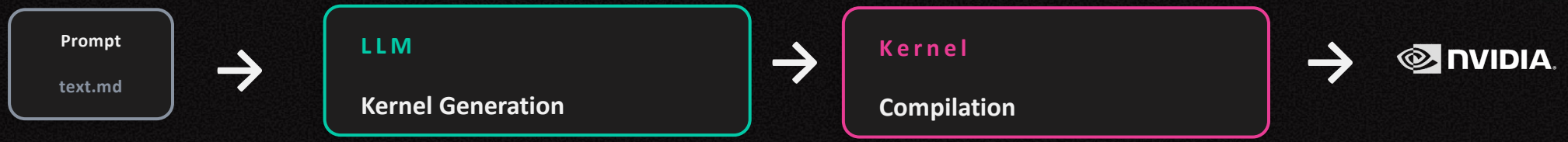
- **Rising Costs**

  60%+ of AI budgets consumed by compute

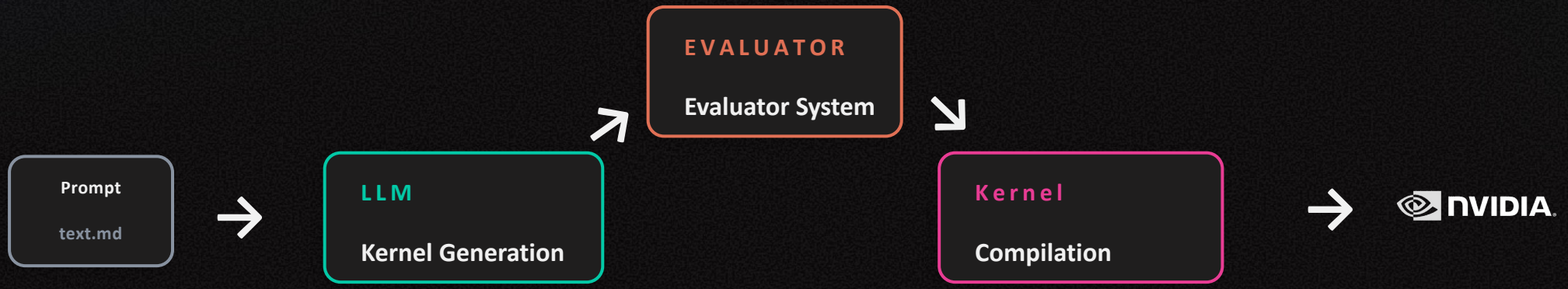
- **Fragmented Deployment**

  Cloud-to-edge complexity slows scaling of AI applications

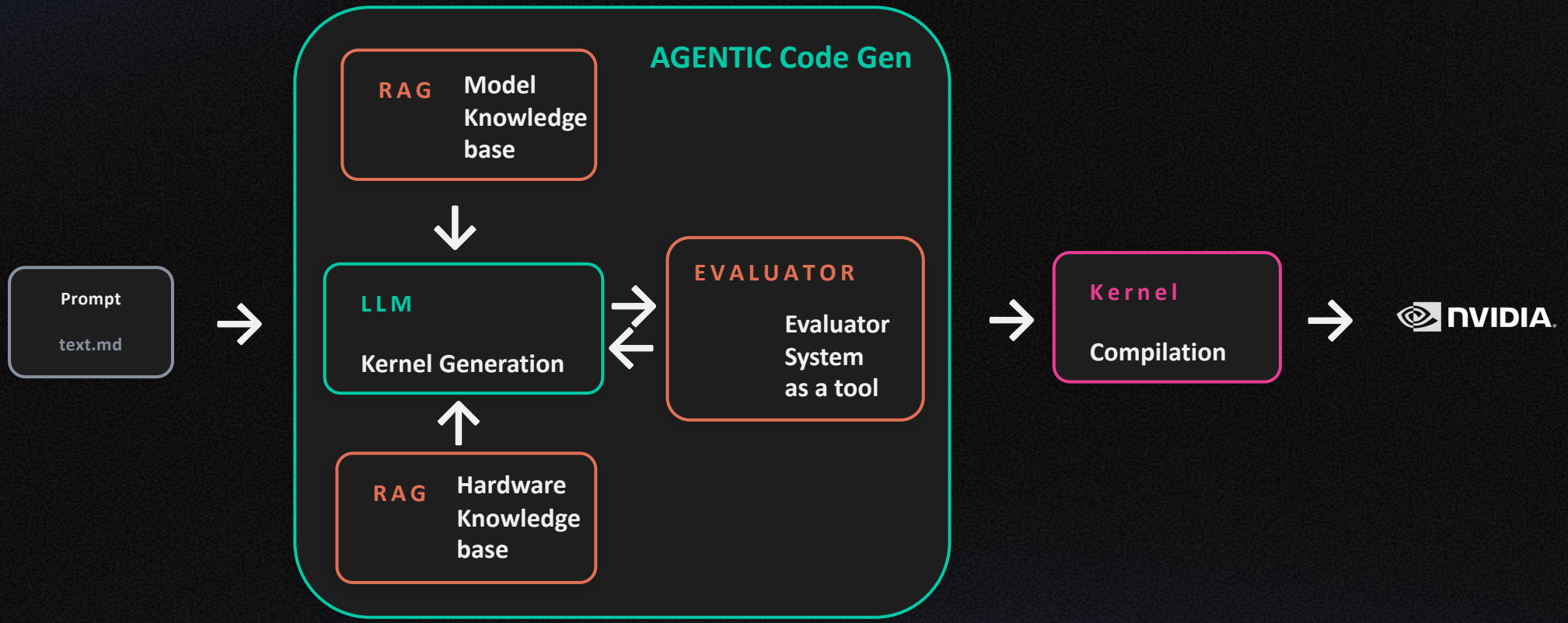
Using LLMs for GPU Kernel Generation



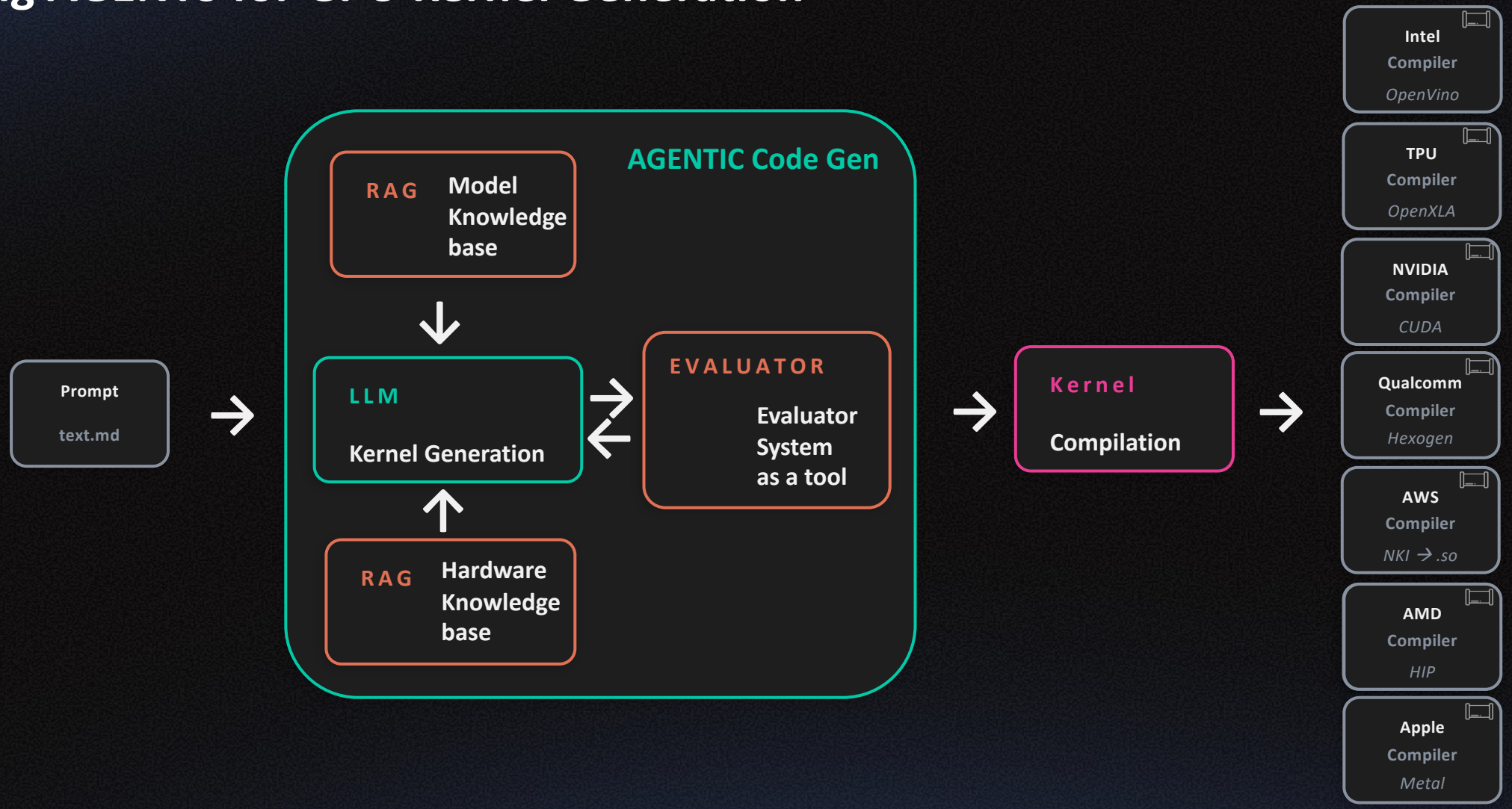
Using AGENTS for GPU Kernel Generation



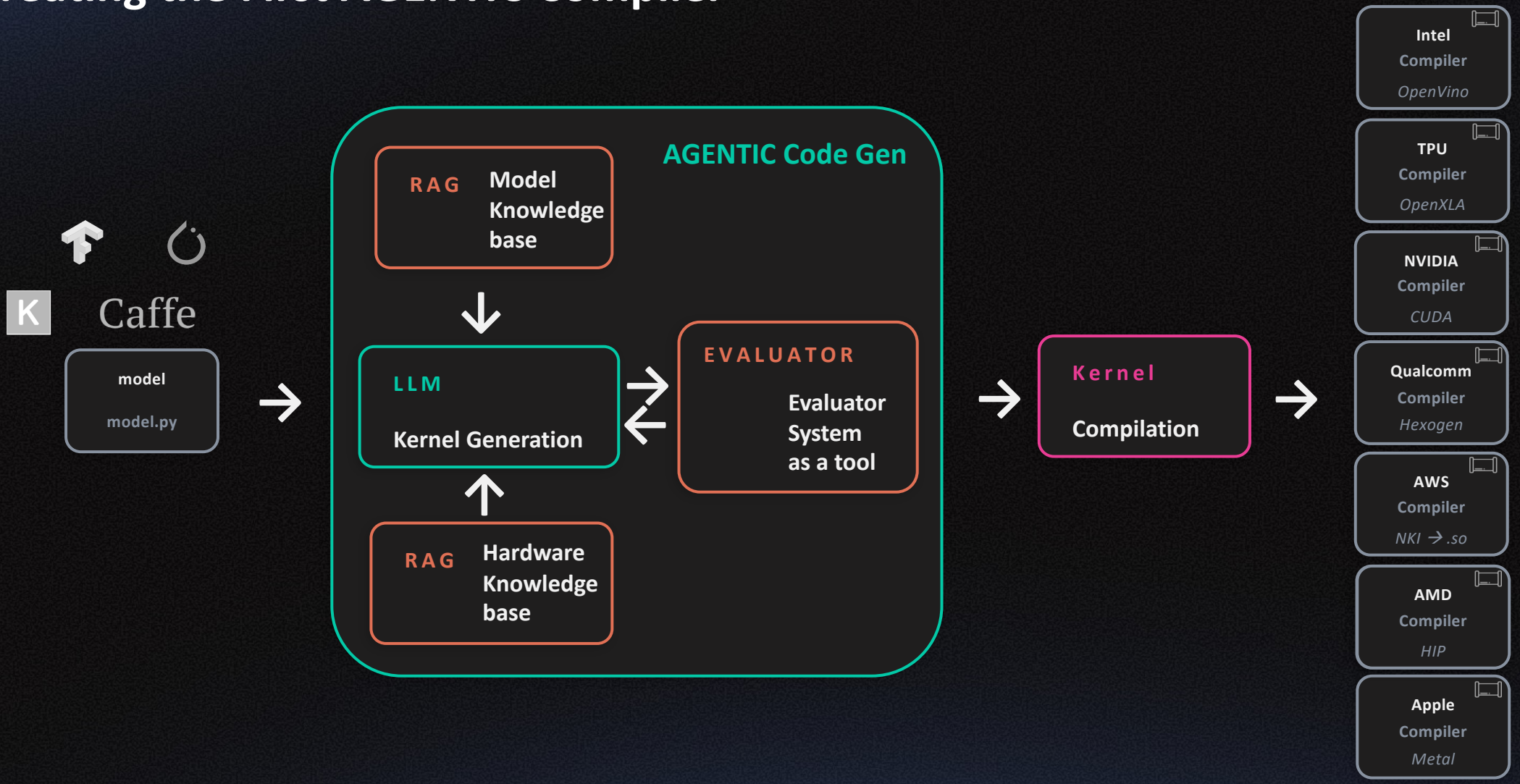
Using AGENTS for GPU Kernel Generation



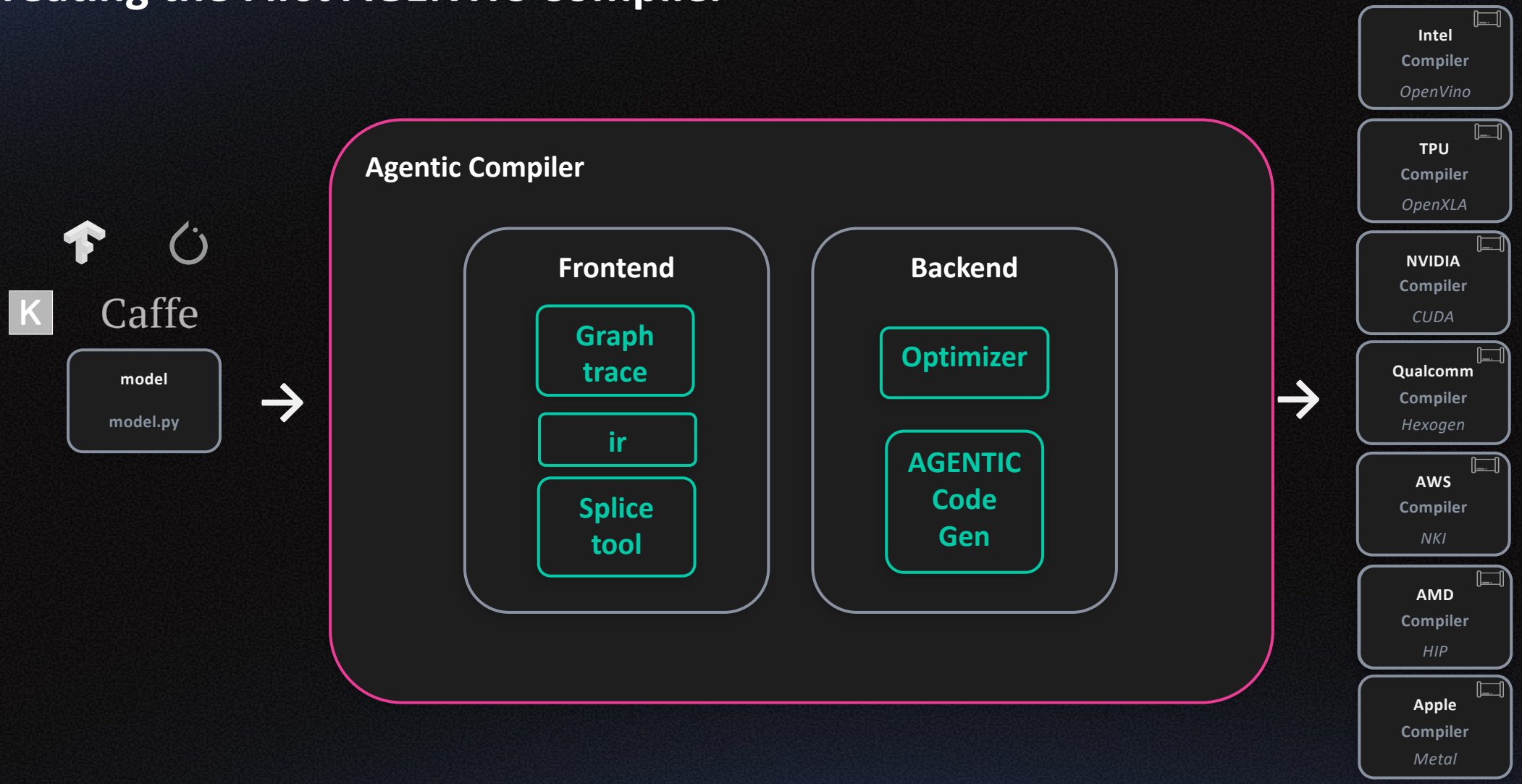
Using AGENTS for GPU Kernel Generation



Creating the First AGENTIC Compiler



Creating the First AGENTIC Compiler



We optimized IBM‘s Granite model

NVIDIA H200

Granite

6.2×

faster inference compared to torch.compile
(IBM granite-4, -620% latency)

How yasp.compile Achieved a 6.25x Speedup on IBM Granite's Mamba Layer with a Single Algebraic Insight

From Micro-Kernels to Macro-Impact: How yasp.compile Sped Up IBM Granite 4.0 by up to 3x End-to-End Over torch.compile

Up to 6.2× faster inference achievable | Almost no accuracy loss (99.9%) | Works with existing PyTorch workflows


yasp