

Motivation

Disaggregated serving places KV-cache transfer directly on the TTFT critical path.

- Non-contiguous and diverse KV layouts
- Asymmetric prefill/decode parallelism
- Heterogeneous accelerator platforms
- Low CPU and synchronization overhead

Our Approach

Build a direct KV connector that maps framework-level metadata to efficient FlagCX P2P transfers.

- Framework-aware KV layout and block mapping
- Asymmetric prefill/decode parallelism
- Batched one-sided writes to decode memory
- Portable across accelerator platforms

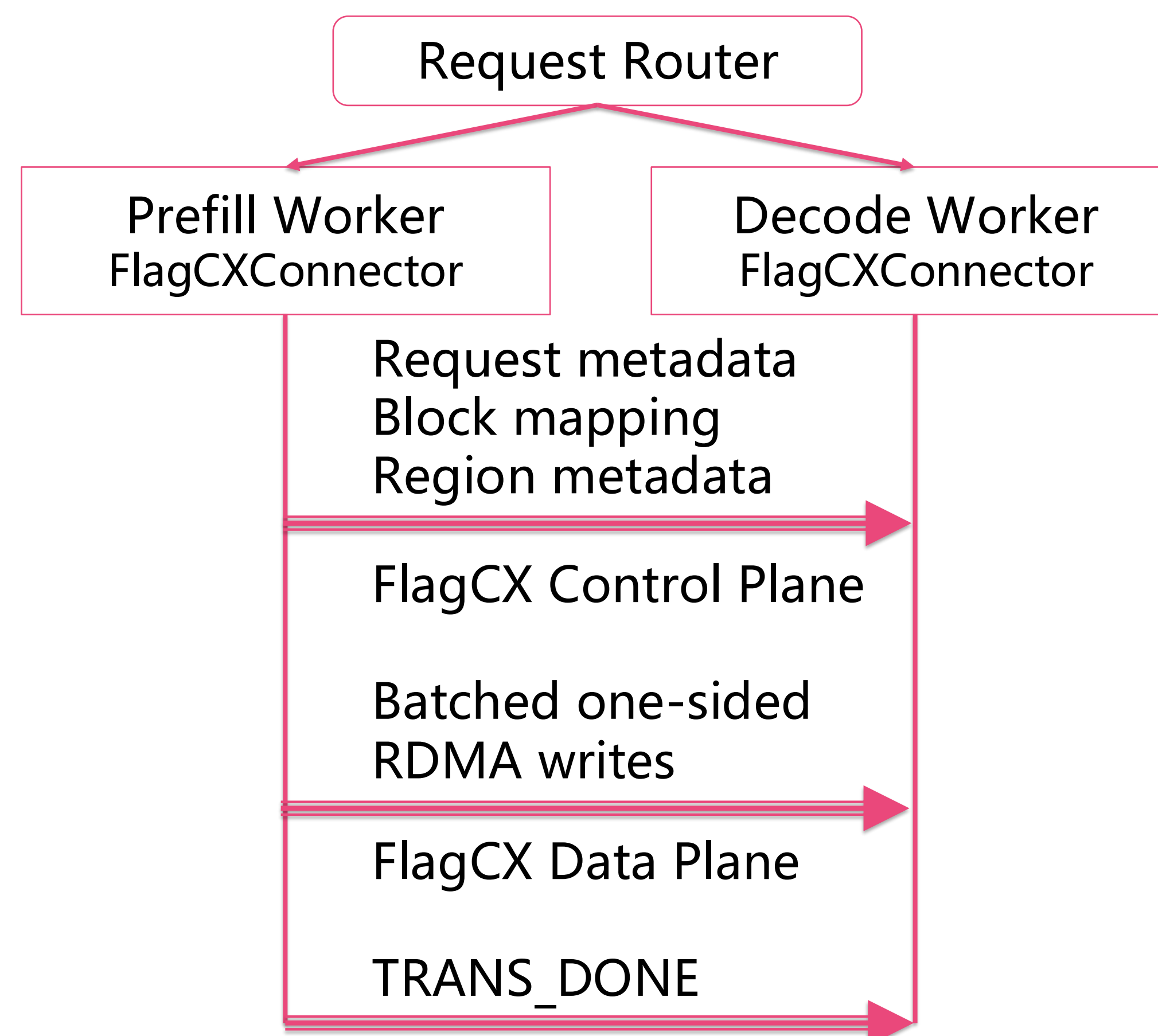
Design Goal

Provide a unified P2P engine for vLLM, powered by FlagCX' s portable abstraction layer across heterogeneous platforms.

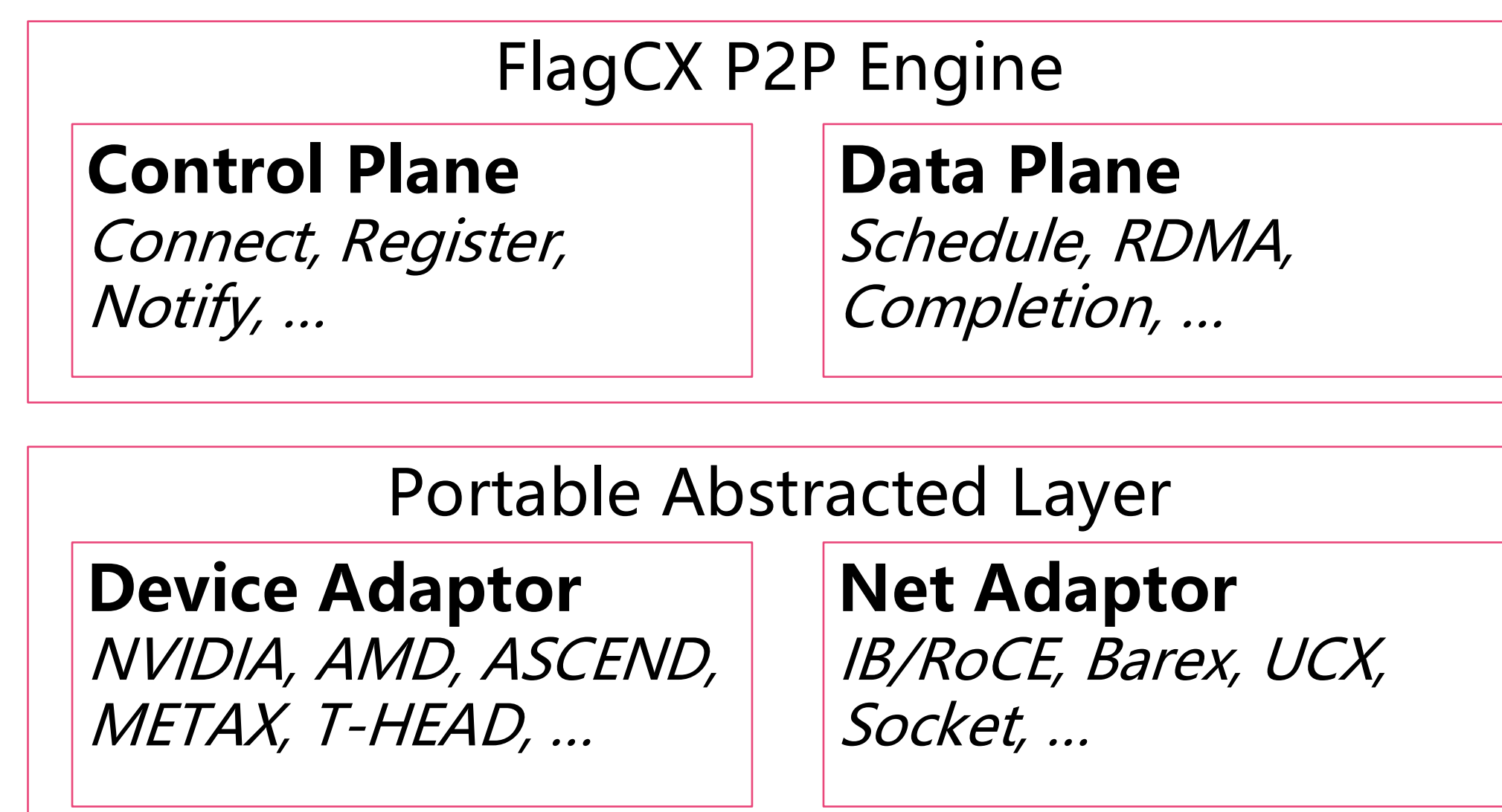
Core Capabilities

- **13** accelerator platforms
- **3** network software stacks
- **2** prefill-decode disaggregation modes

System Architecture



FlagCX Architecture

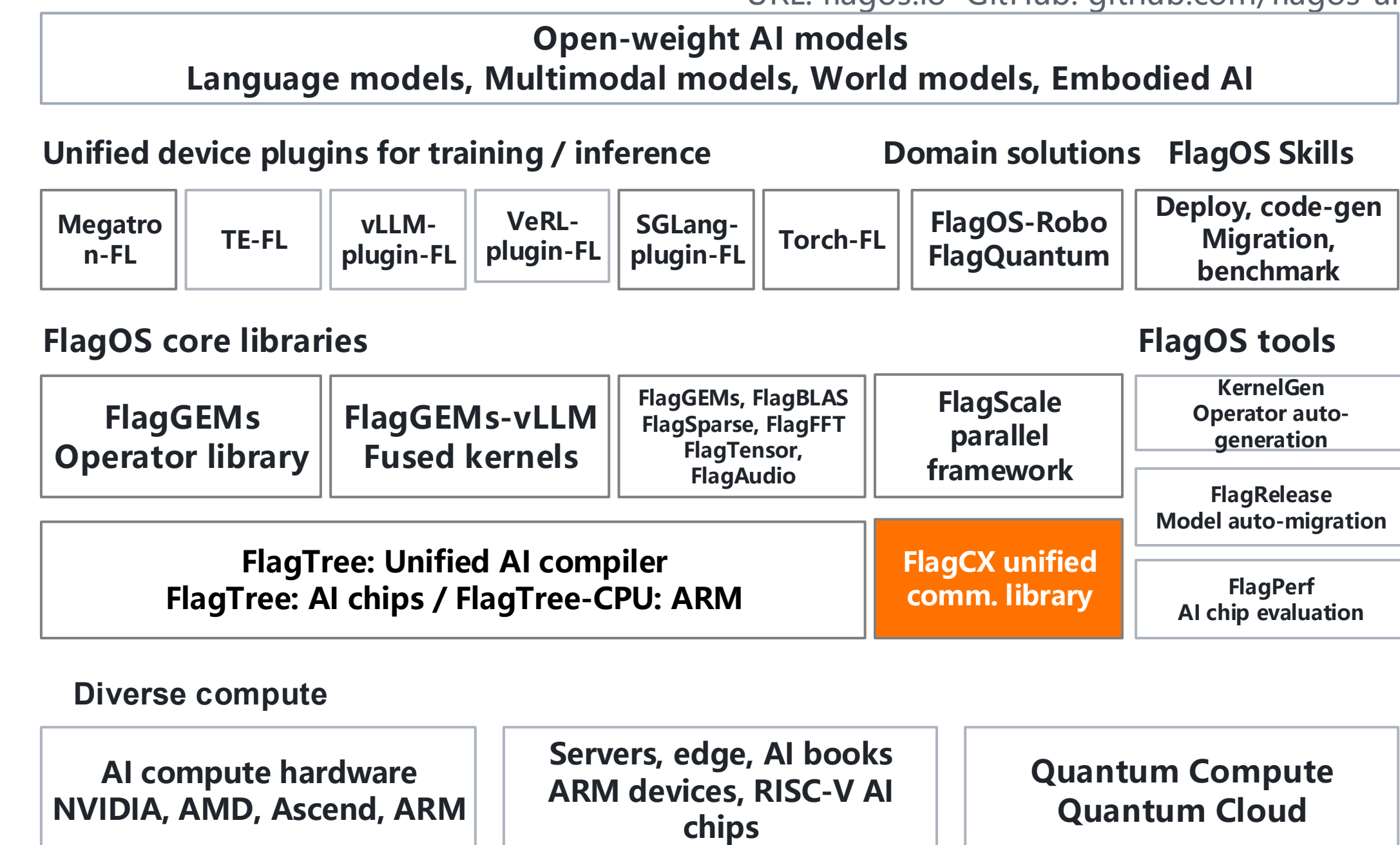


FlagOS: Open-sourced System SW

for Multi-Vendor AI chipsets

20+ Vendors 30+ Chips

URL: flagos.io GitHub: github.com/flagos-ai



Main Results

Model	Platform	Config	TTFT speedup
Qwen3.6-35B	Nvidia	1P1D, TP8, 16K/1K	1.02
	Metax	1P1D, TP8, 16K/1K	1.01
	Hygon	1P1D, TP8, 16K/1K	1.04
GLM5.2	T-Head	1P1D, TP16, 16K/1K	1.00
	Metax	1P1D, TP16, 16K/1K	1.01

FlagCX matches Mooncake' s performance (**1.00×–1.04×**) across multiple models and accelerator platforms.