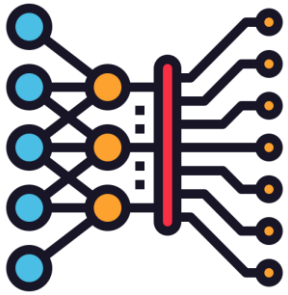
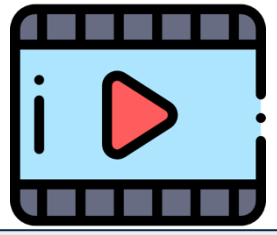
Enabling Energy Efficiency for Distributed Machine Learning

Tawhid Bhuiyan and Tanvir Ahmed Khan

Large Machine Learning Models Consume a Lot of Energy



Large Machine Learning Model



6s 480p Video
110Wh

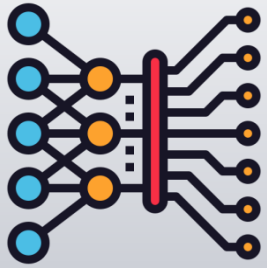
thin, medium-well steak
220Wh

Essential to model **end-to-end energy consumption** for optimization

Source: https://archive.is/l3Klo

What Does Modeling End-To-End Energy Mean?

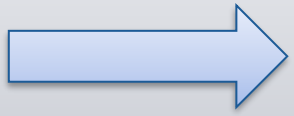
Problem Definition



ML Workload

System Setup

Inputs



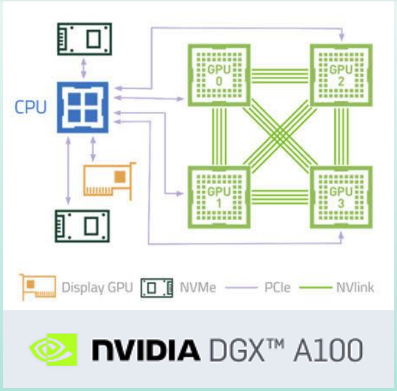
Energy Breakdown

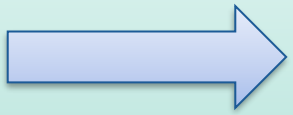
Time Breakdown

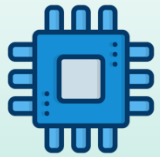
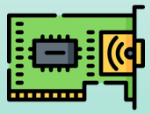
Outputs

Example

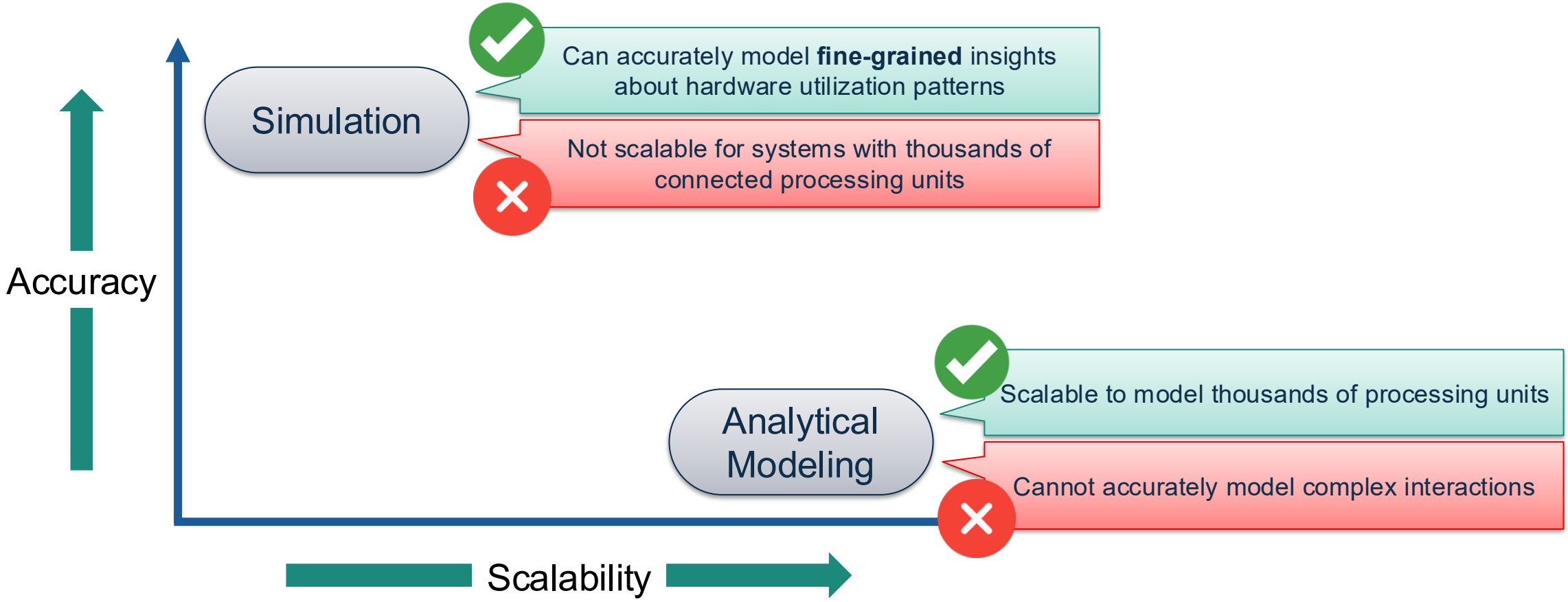
Llama 3



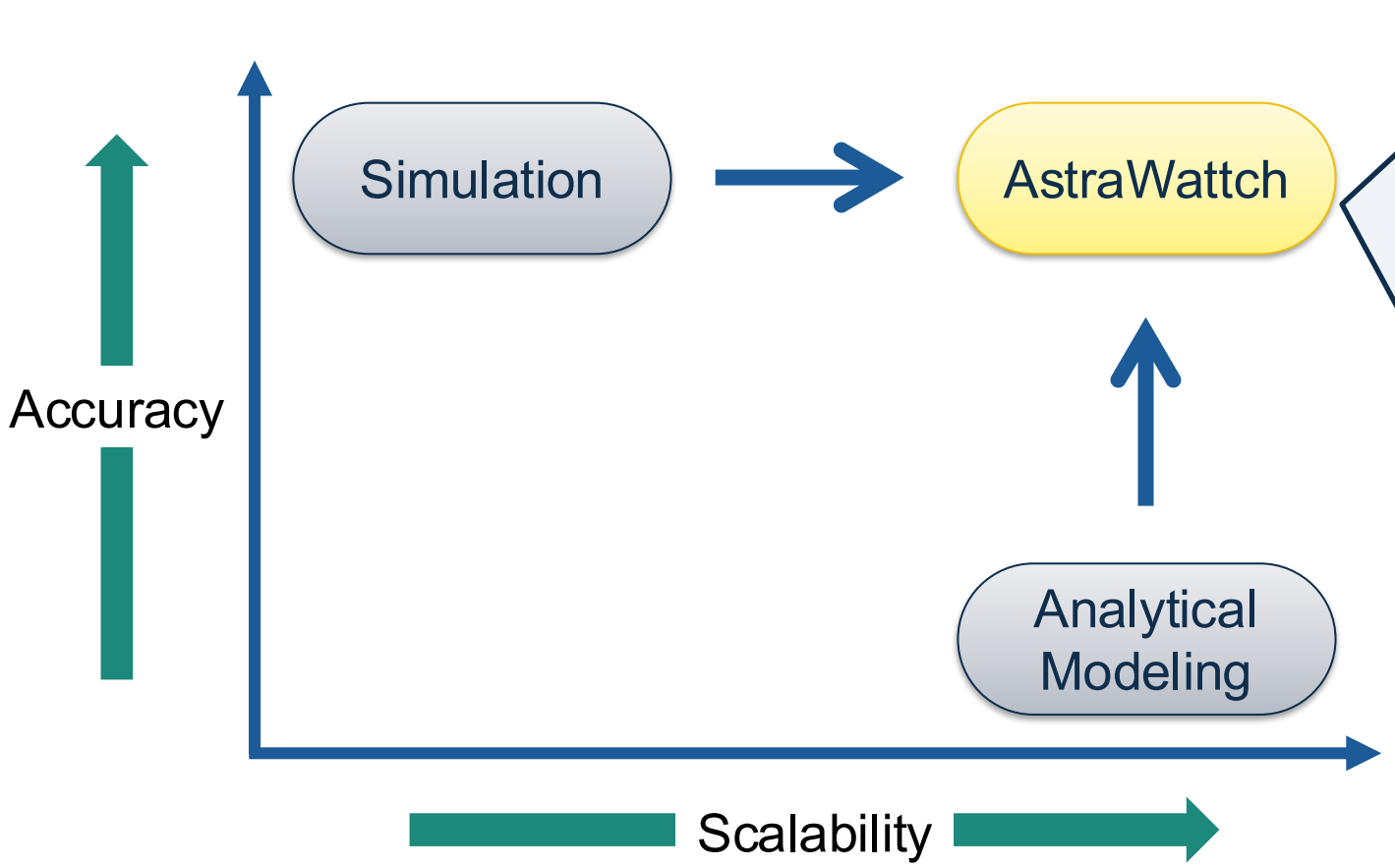


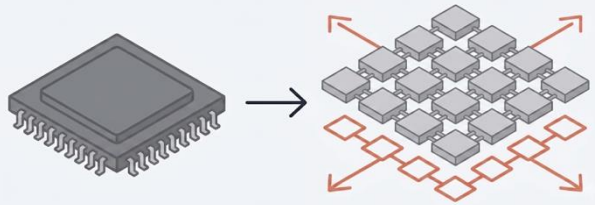
Compute		120J	6s
Memory		20J	3s
Network		15J	2s

Challenges in Modeling End-To-End Energy



AstraWattch: Profile Once and Model Everywhere



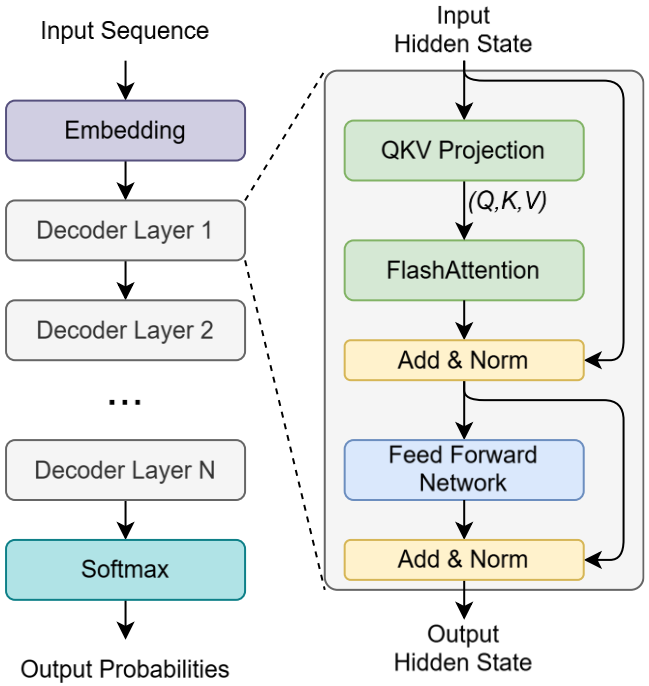


AstraWattch achieves high accuracy and scalability by:

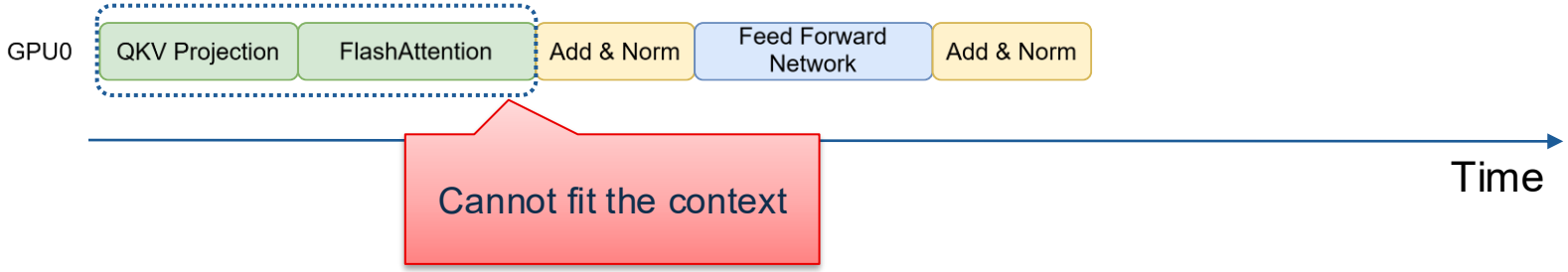
- **Profiling Once:** Drawing fine-grained insights from simulation or real silicon.
- **Modeling Everywhere:** Using the insights to drive fast analytical models.

Preliminary results: AstraWattch enables accurate energy modeling in minutes

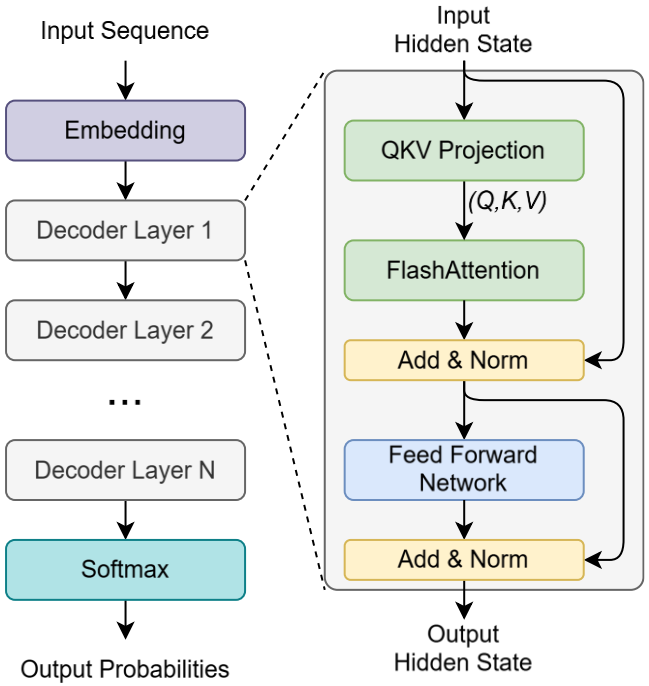
Profiling Machine Learning



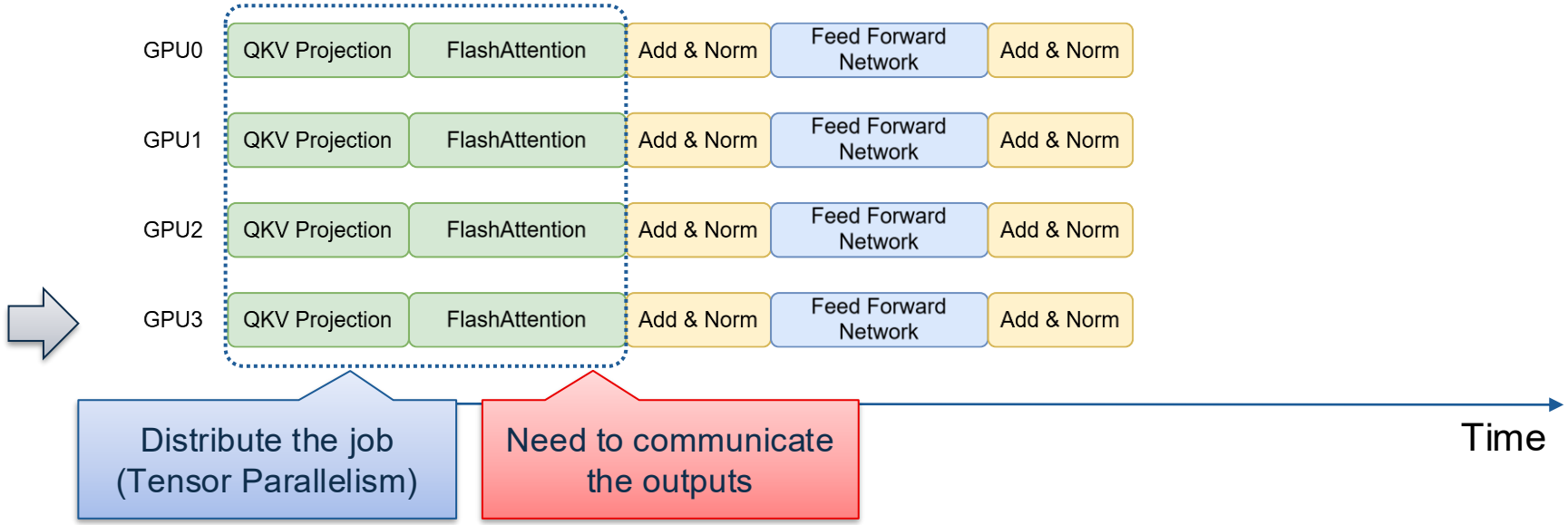
LLM Architecture



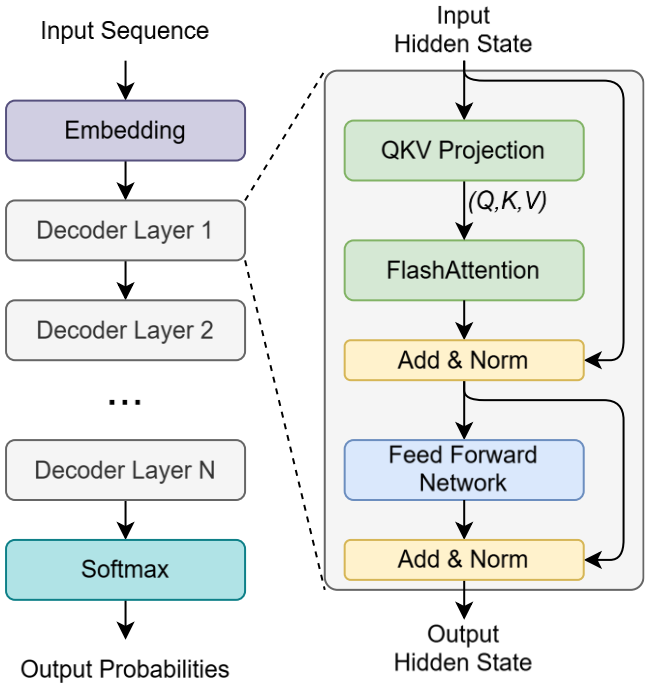
Profiling Distributed Machine Learning



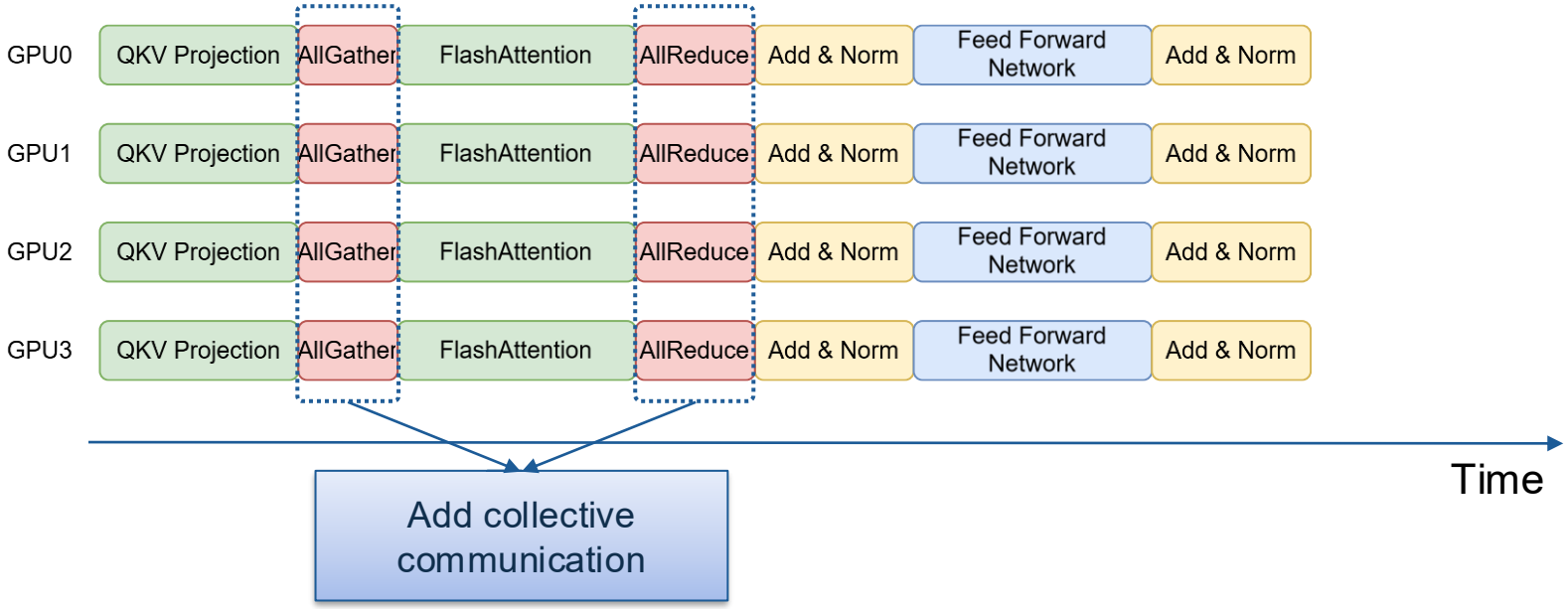
LLM Architecture



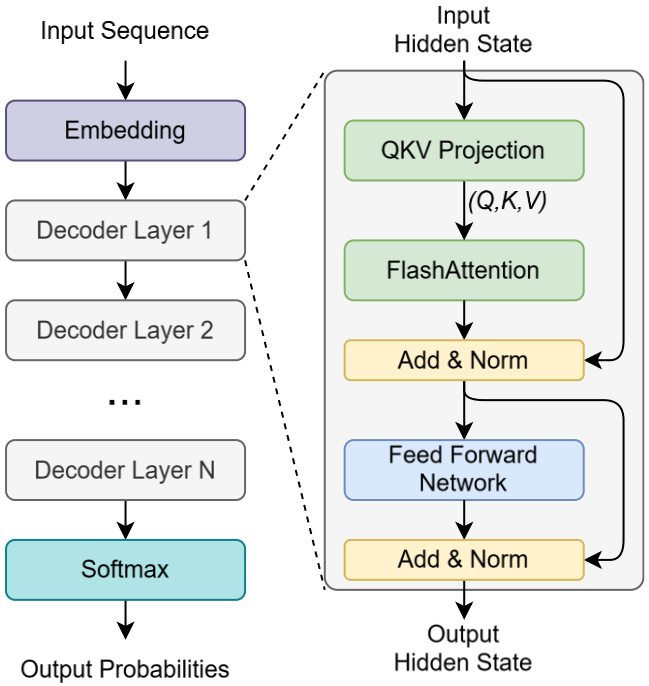
Profiling Distributed Machine Learning



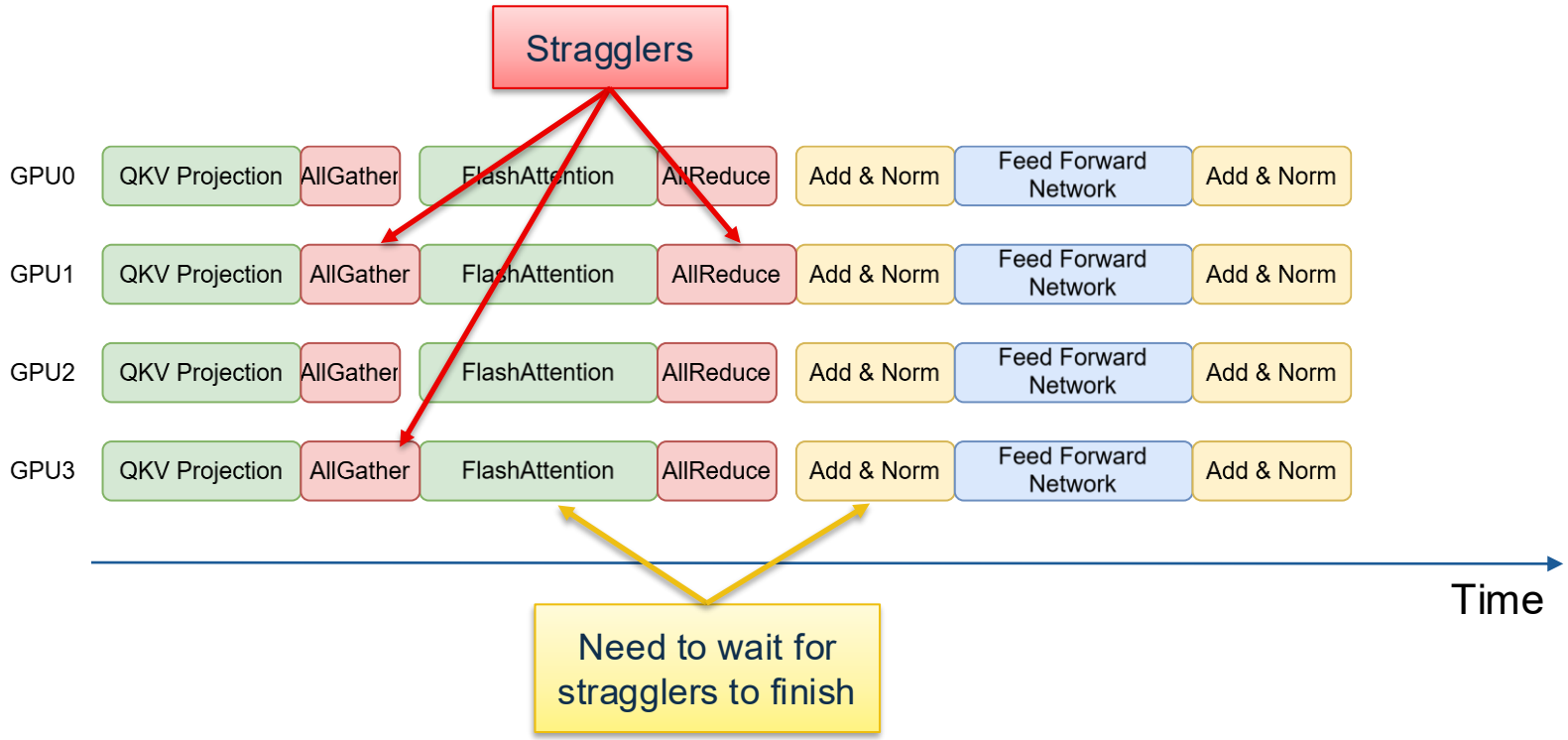
LLM Architecture



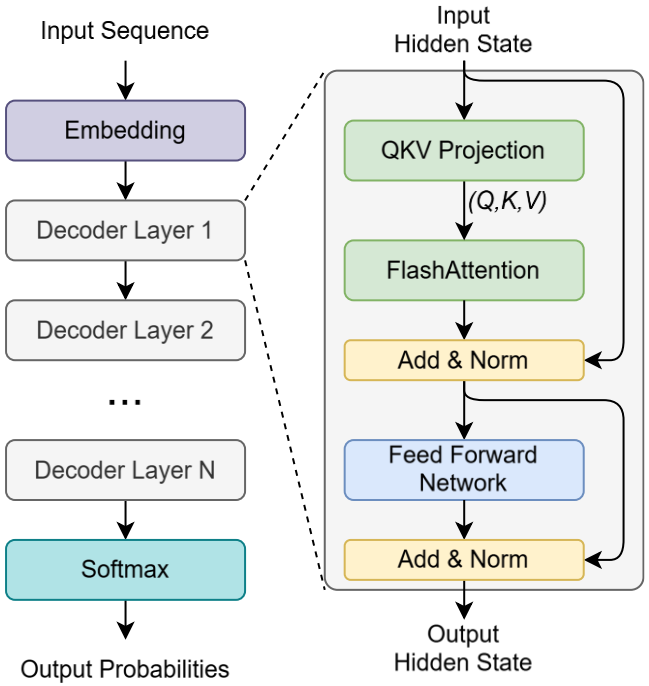
Profiling Insights



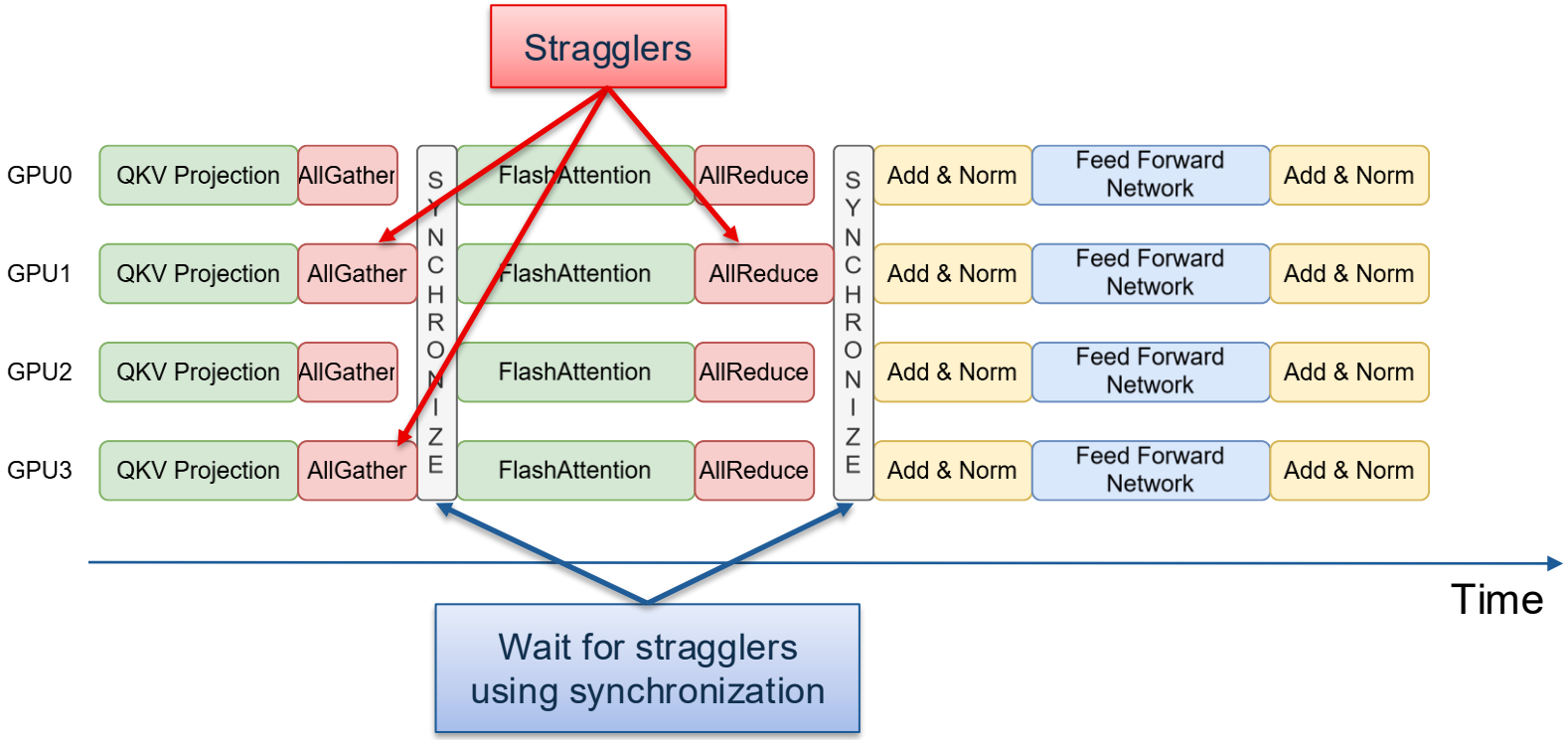
LLM Architecture



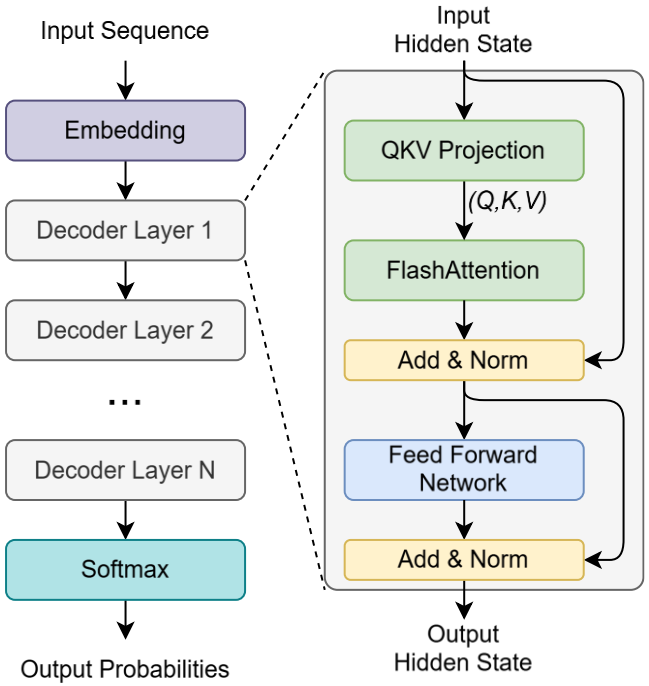
Profiling Insights



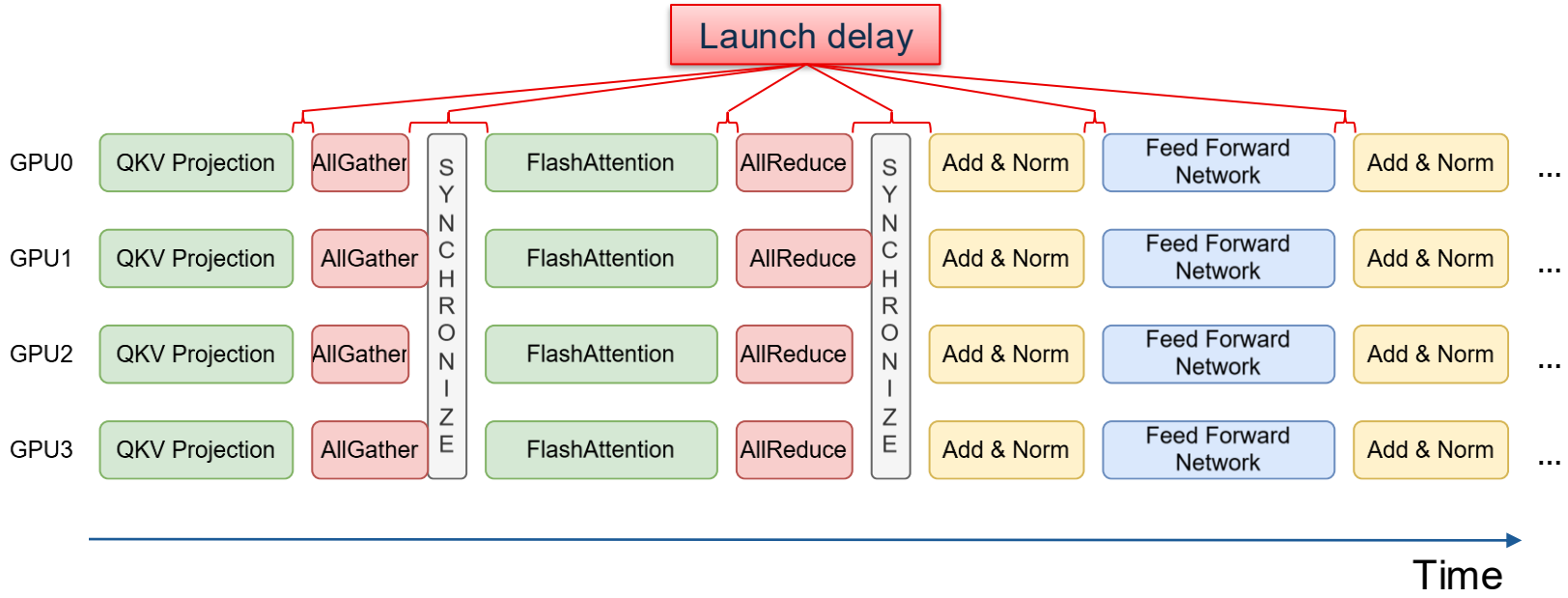
LLM Architecture



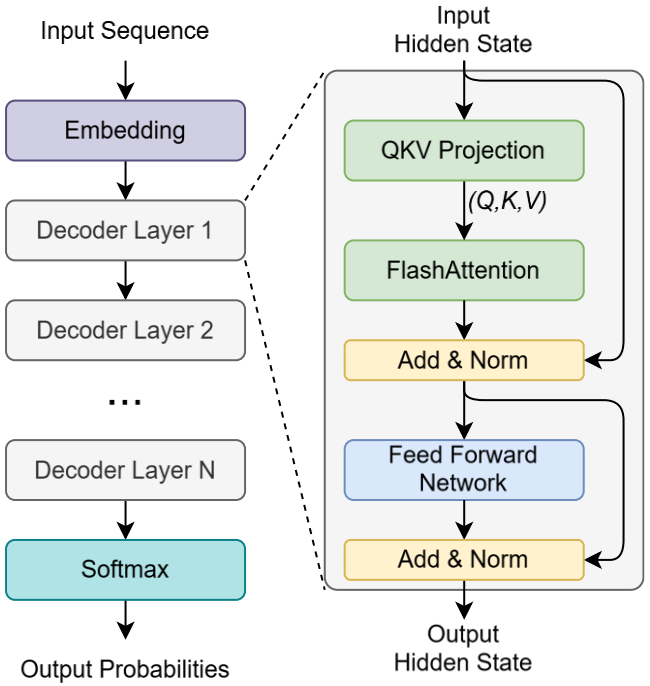
Profiling Insights



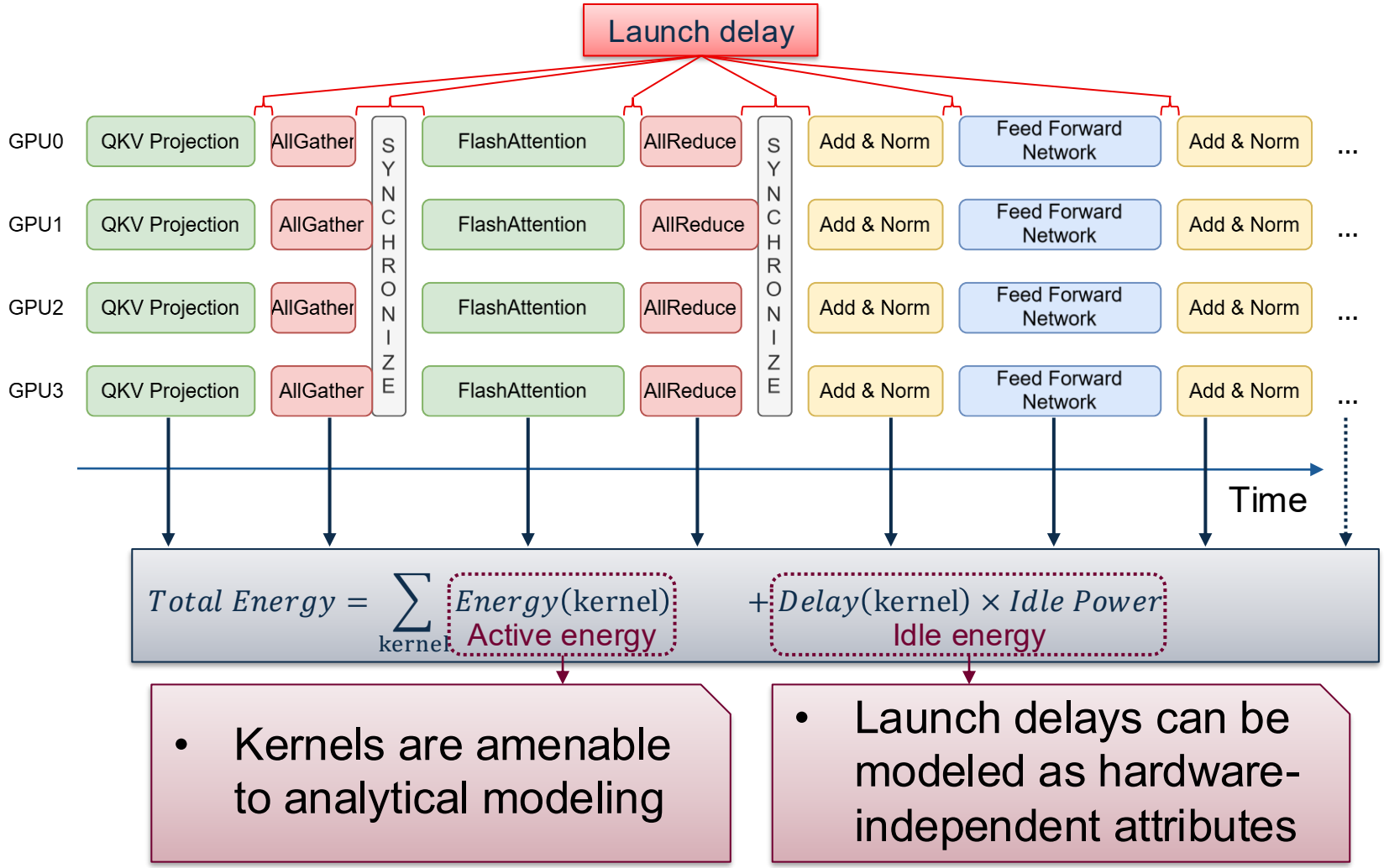
LLM Architecture



Profiling Insights For Energy

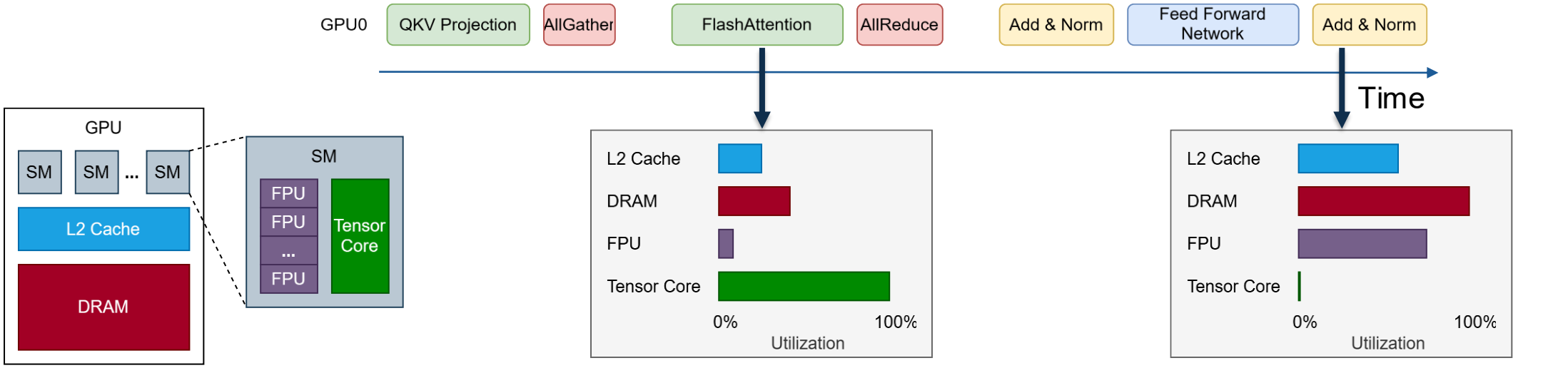


LLM Architecture



- Kernels are amenable to analytical modeling
- Launch delays can be modeled as hardware-independent attributes

Profiling Kernels



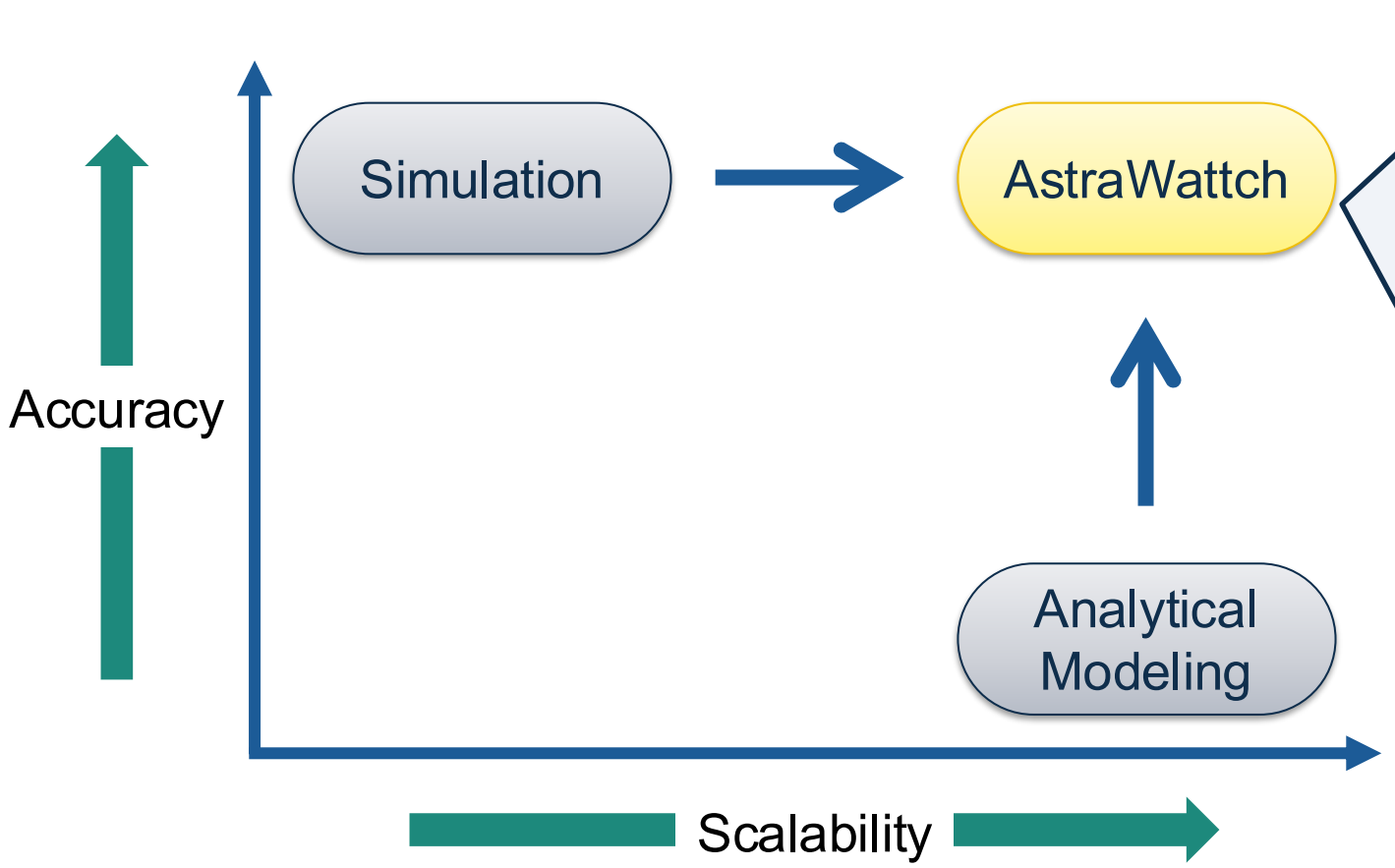
Major GPU Components

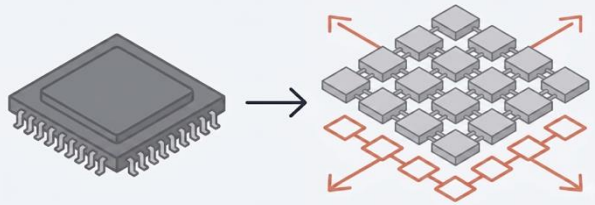
Large matrix multiplication makes use of the tensor core

Elementwise operation becomes memory bound

- Coarse-grained component utilization can accurately model energy

AstraWattch: Profile Once and Model Everywhere



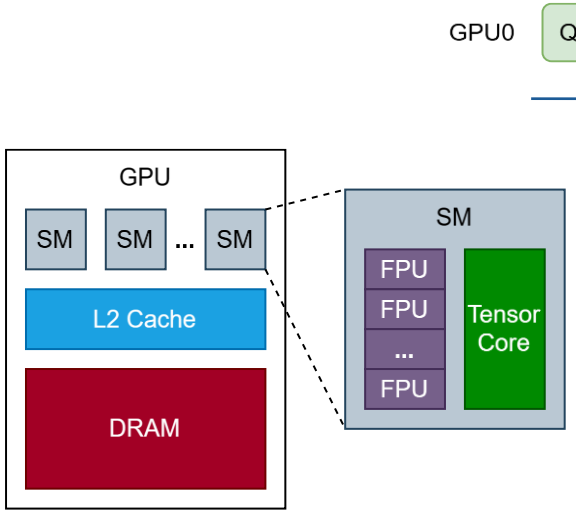


AstraWattch achieves high accuracy and scalability by:

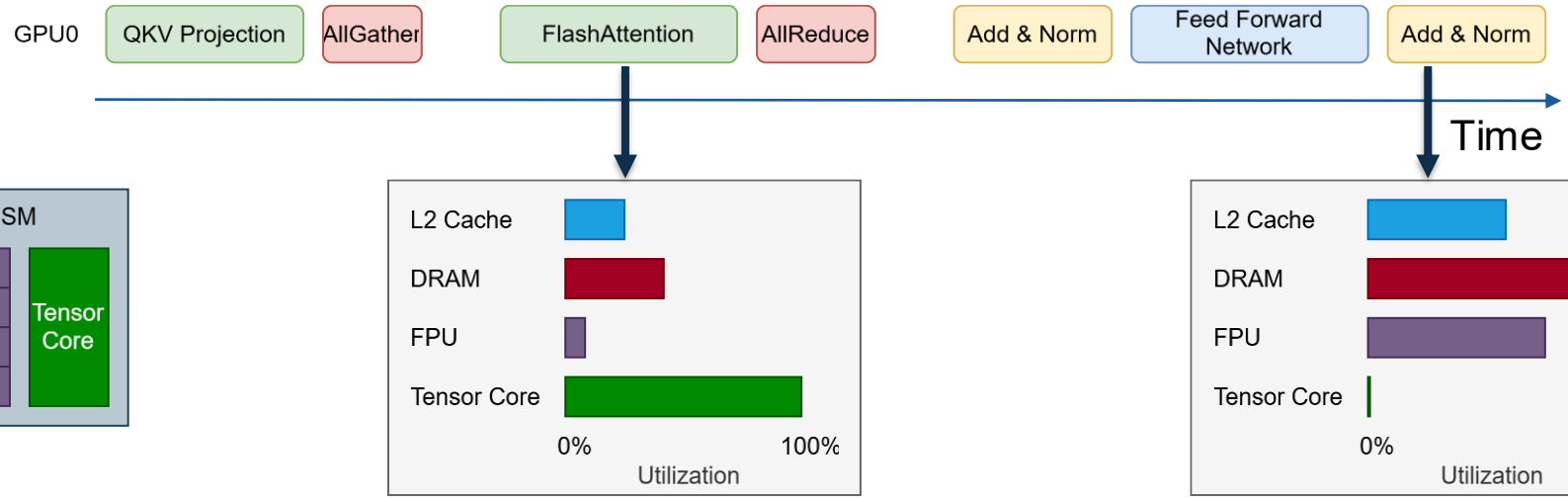
- **Profiling Once:** Drawing fine-grained insights from simulation or real silicon.
- **Modeling Everywhere:** Using the insights to drive fast analytical models.

Preliminary results: AstraWattch enables accurate energy modeling in minutes

Modeling Active Energy



Major GPU Components



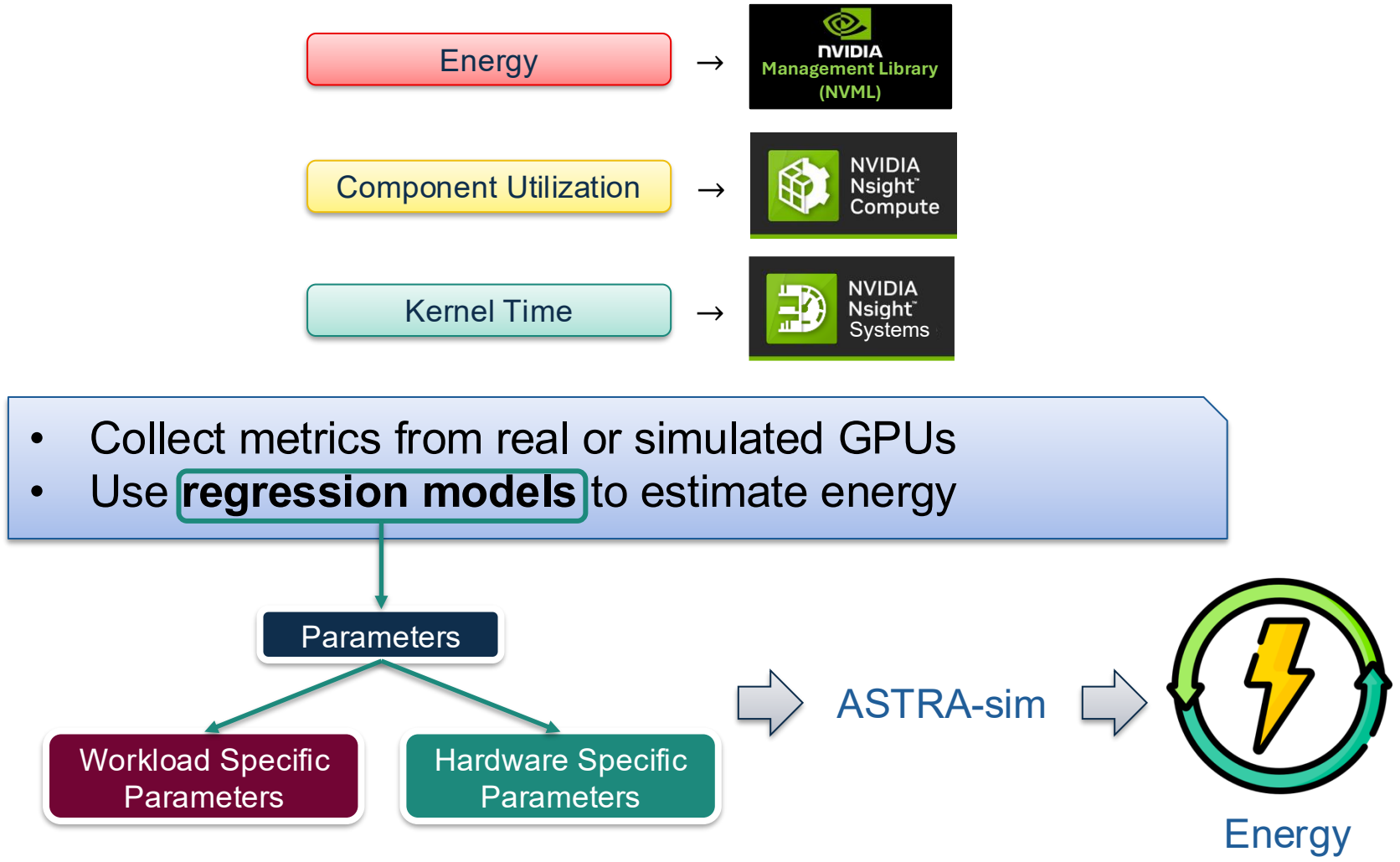
Large matrix multiplication makes use of the tensor core

Elementwise operation becomes memory bound

$Energy(\text{kernel})$ = Sum of component energies for the kernel

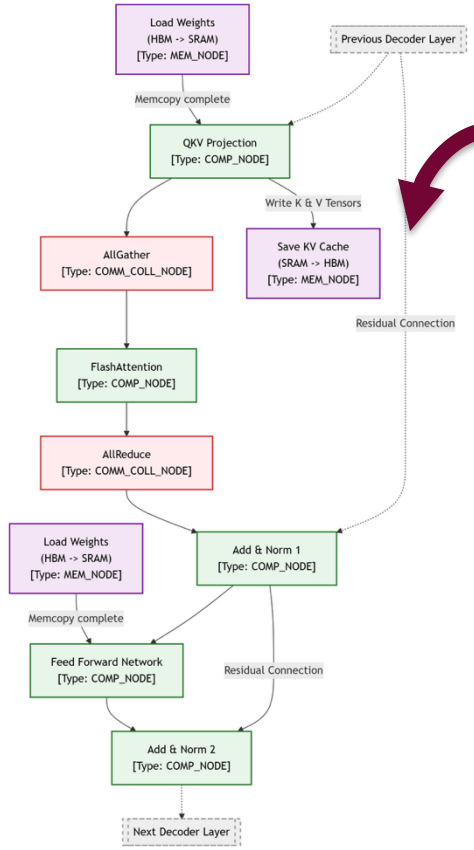
- We can build simple **analytical models for each kernel**
- One analytical model covers all invocations of a kernel

Calibrating the Model Parameters and Using Everywhere



ASTRA-sim's Modeling of Distributed Machine Learning

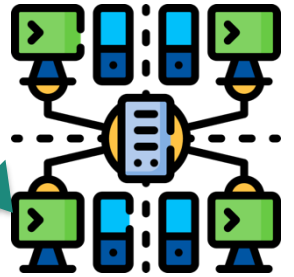
Encode



Chakra Trace
(Execution Graph)

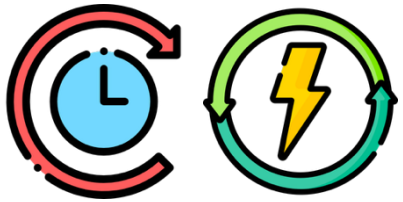
Workload Specific Parameters

Hardware Specific Parameters



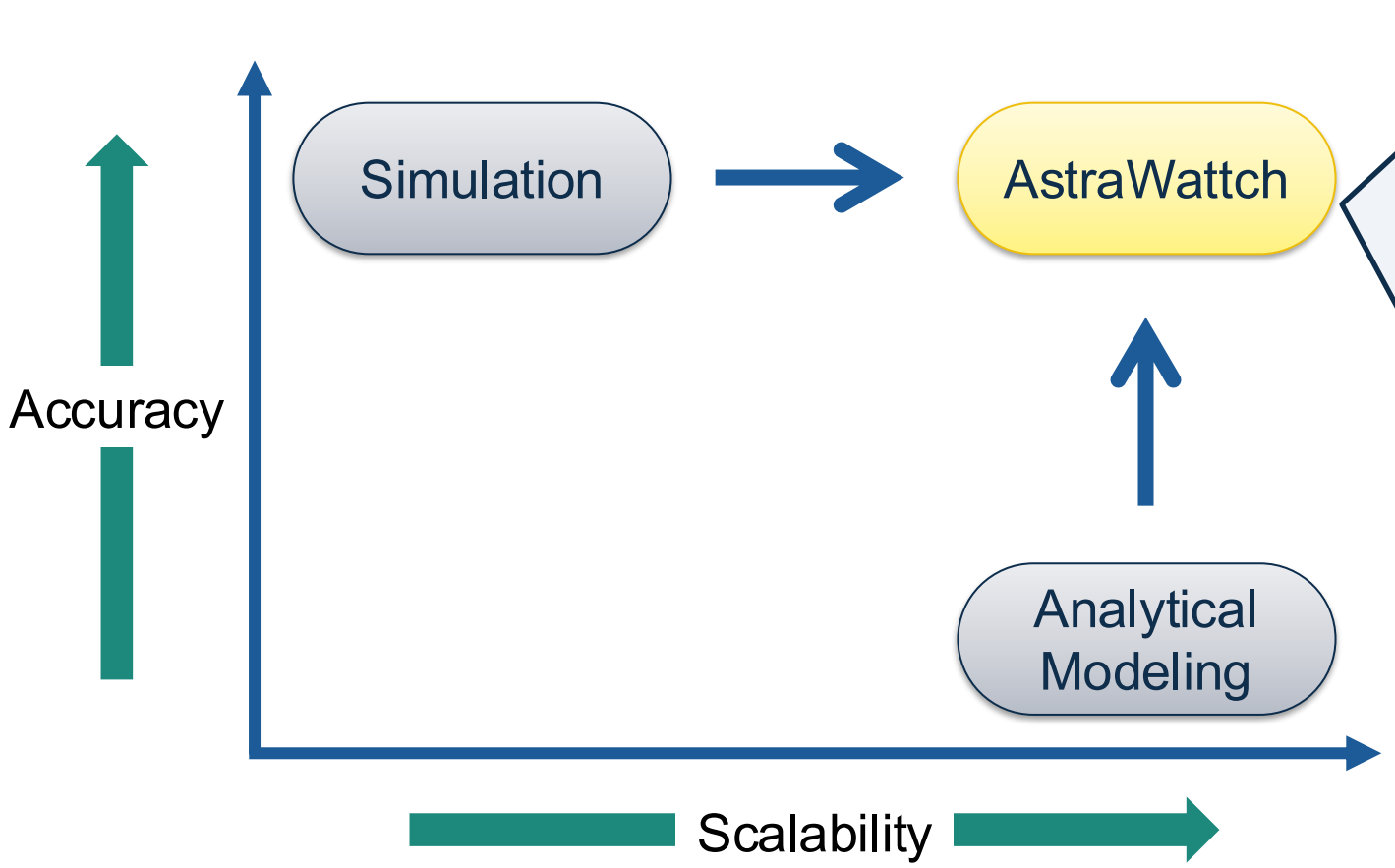
System Setup

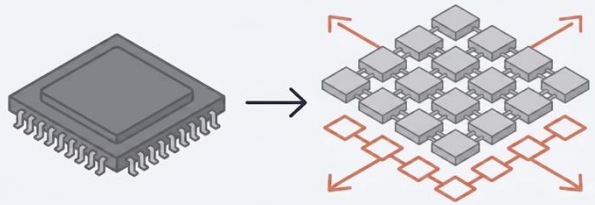
ASTRA-sim



Time Energy

AstraWattch: Profile Once and Model Everywhere





AstraWattch achieves high accuracy and scalability by:

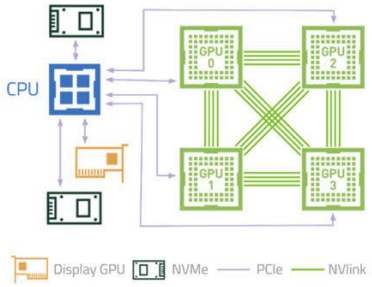
- **Profiling Once:** Drawing fine-grained insights from simulation or real silicon.
- **Modeling Everywhere:** Using the insights to drive fast analytical models.

Preliminary Results: AstraWattch enables accurate energy modeling in minutes

Preliminary Results: AstraWattch's Modeling Accuracy

BERT

Training



NVIDIA DGX V100

NVIDIA Management Library (NVML)

NVIDIA Nsight Compute

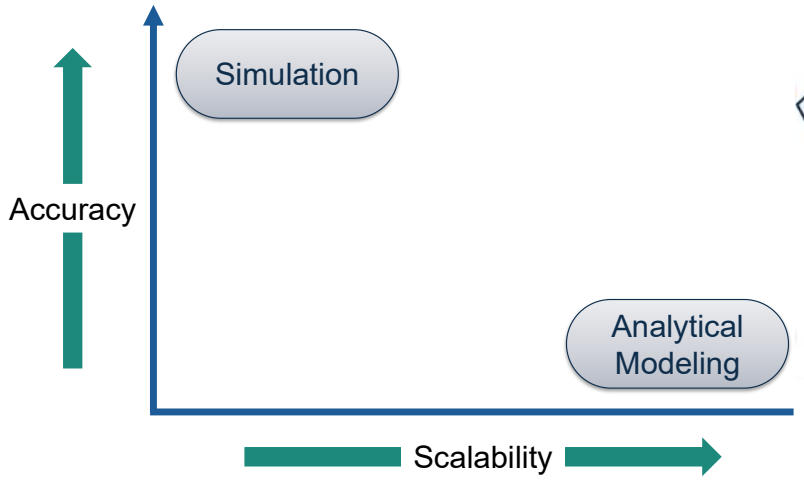
NVIDIA Nsight Systems

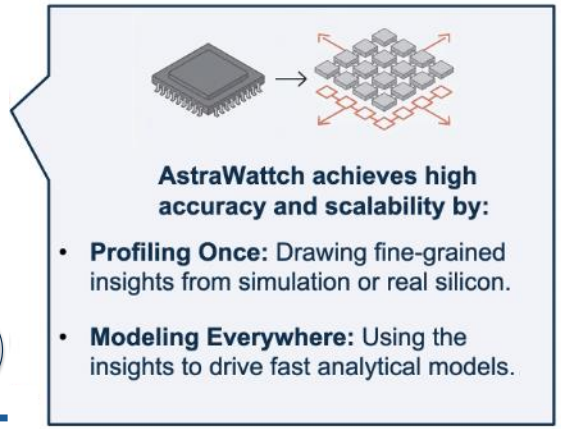
10 Minutes

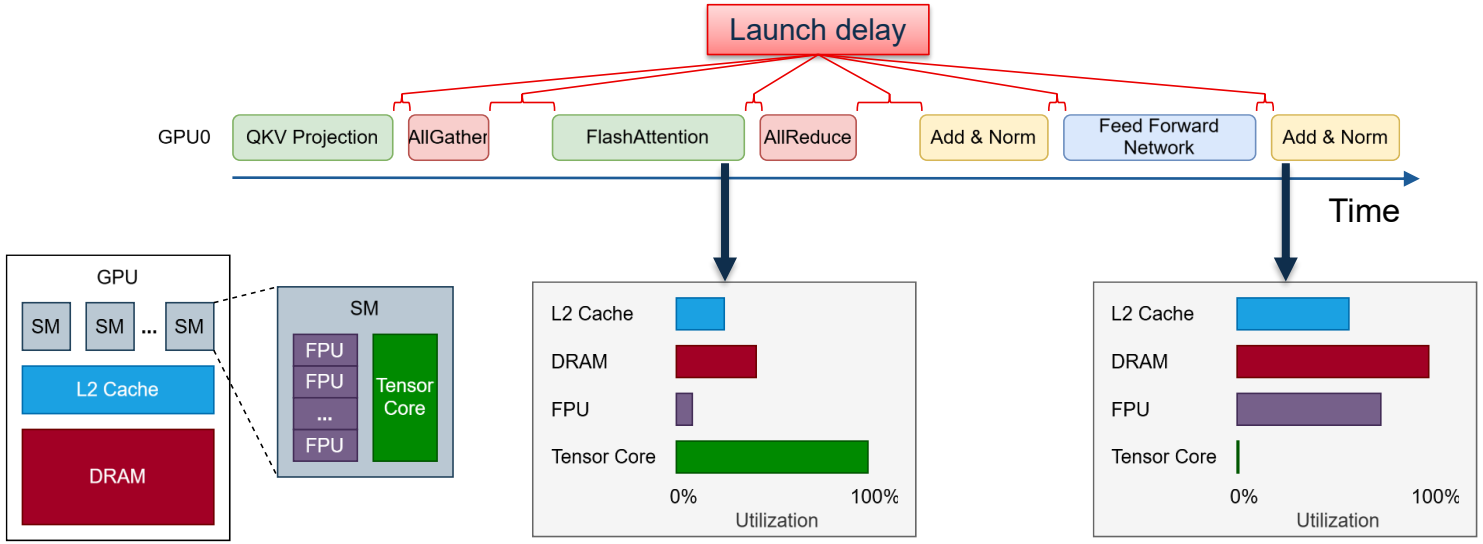
Frequency (MHz)	Time Modeling Error (%)	Energy Modeling Error (%)
735	9.9	11.8
930	4.6	2.2
1132	2.2	2
1335	2.1	4
1530	7.4	0.9

AstraWattch enables accurate energy modeling in minutes by generalizing analytical models to different frequencies

Enabling Energy Efficiency for Distributed Machine Learning







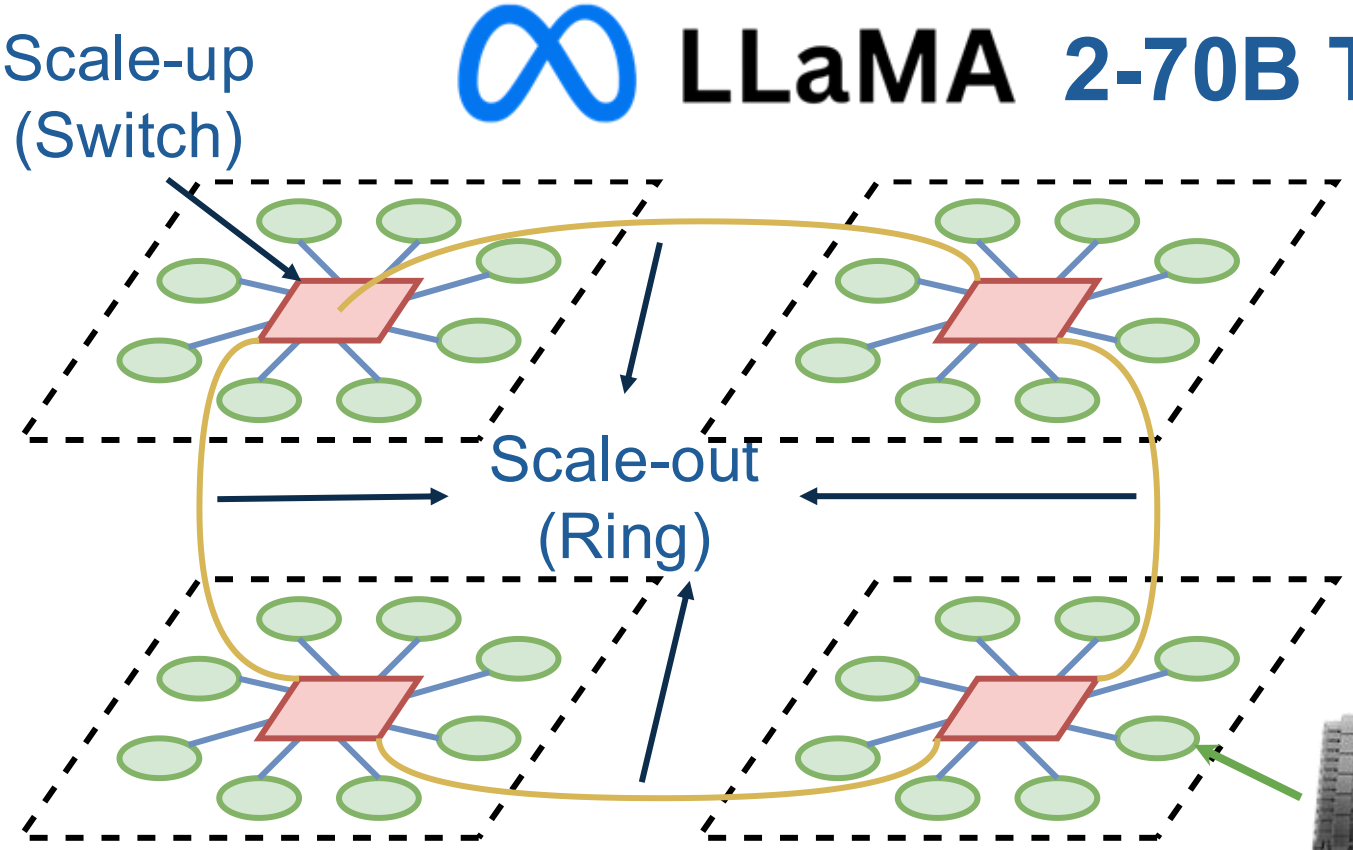
- Component utilization and delay information → **chakra trace**
- Component power → **simulation configuration**

Tawhid Bhuiyan
Email: mb5332@columbia.edu
Website: mthbhuiyan.github.io

Preliminary results: AstraWattch enables accurate energy modeling in minutes

Case Study: Marginal Efficiency Analysis

Machine Learning Model and System Configuration



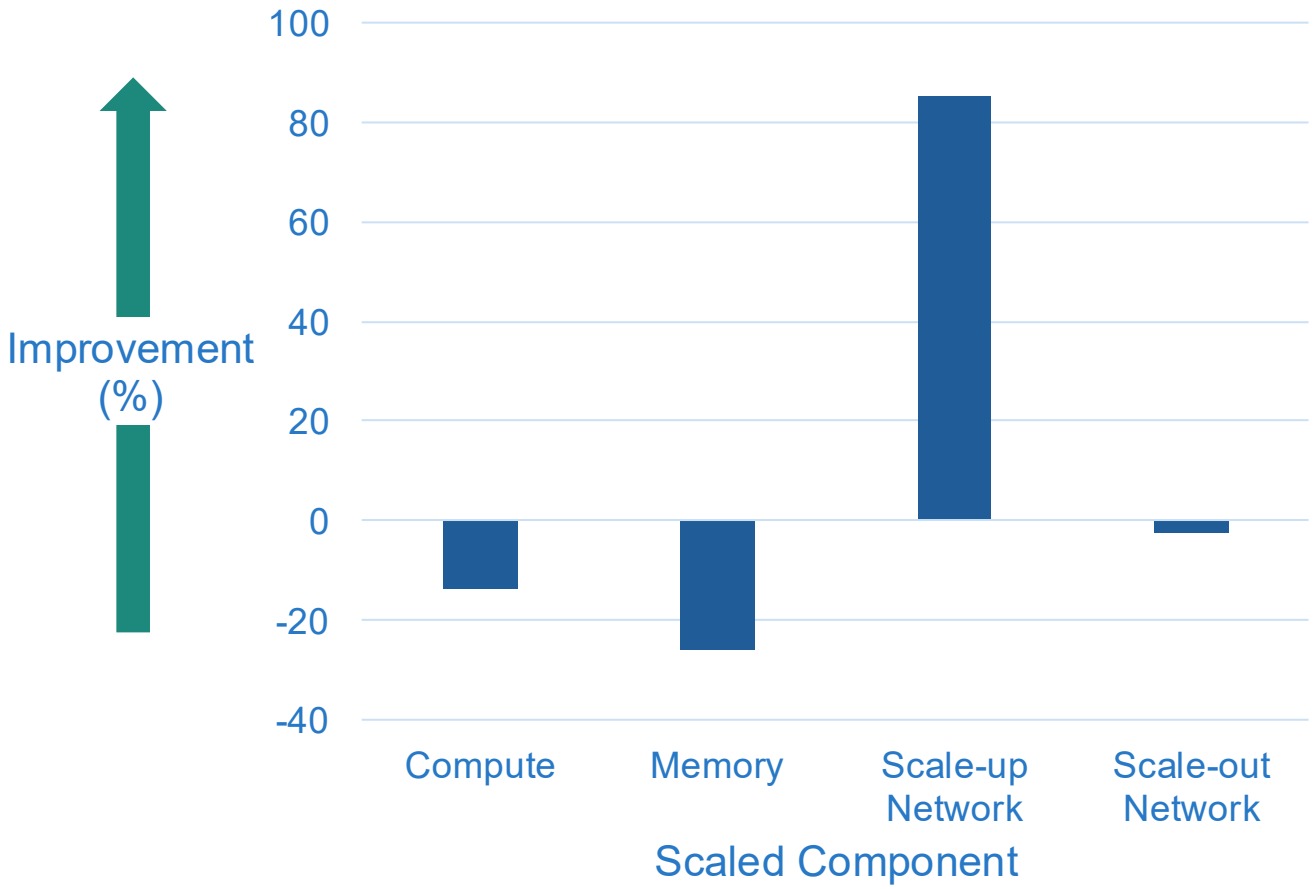
LLaMA 2-70B Training

Active Power = 700 W
Idle Power = 80 W
Scale-up Network:
Energy Per Bit = 2pJ
Bandwidth = 400GB/s
Scale-out Network:
Energy Per Bit = 25pJ
Bandwidth = 50GB/s

Switch (8) - Ring (4) Topology

NVIDIA H100

Case Study: Marginal Efficiency Analysis



- Scale-out network voltage–frequency scaling delivers efficiency gain
- Due to
  - Low direct network energy
  - Poor compute-communication overlap
  - High idle energy impact