Analytical 1: Operator level
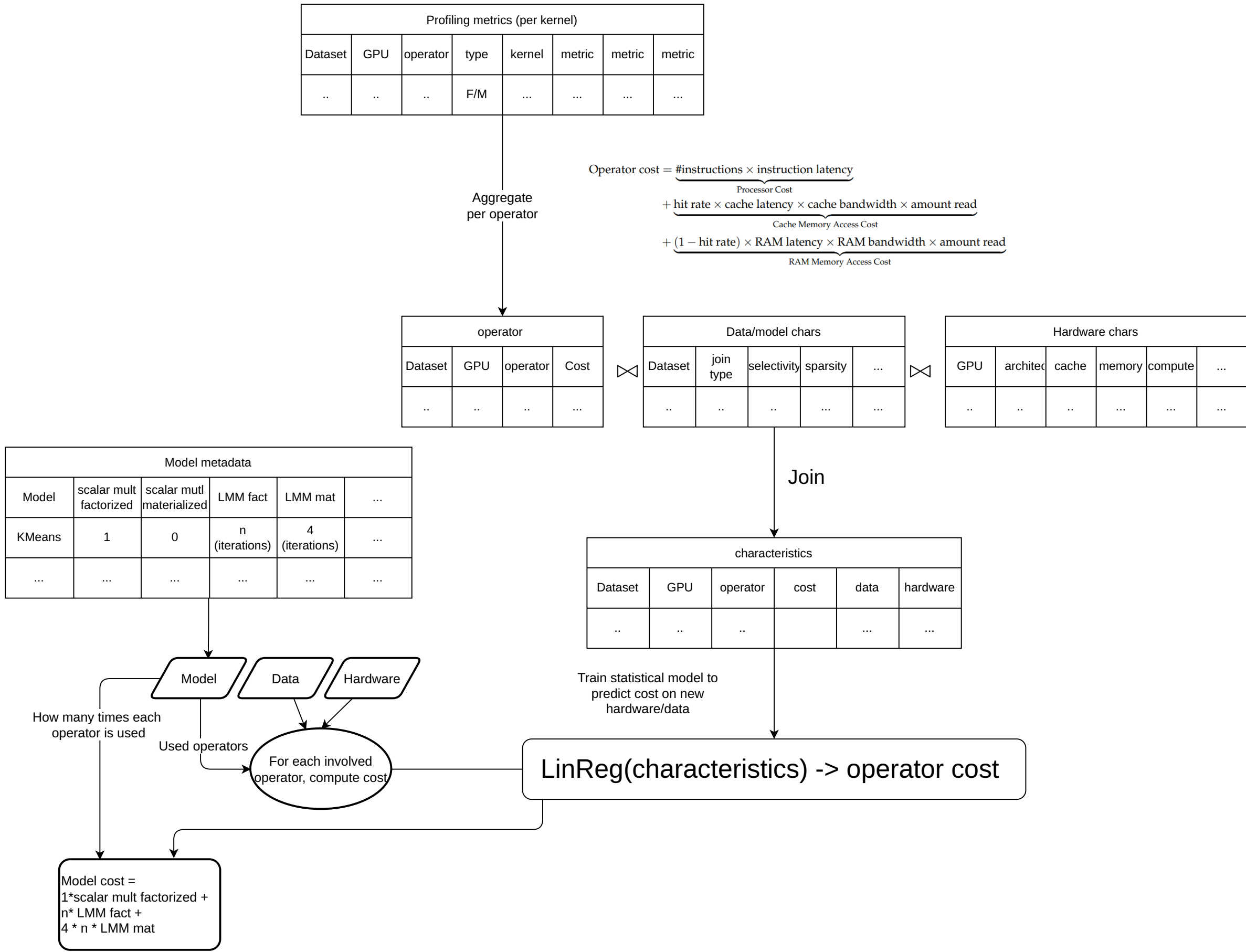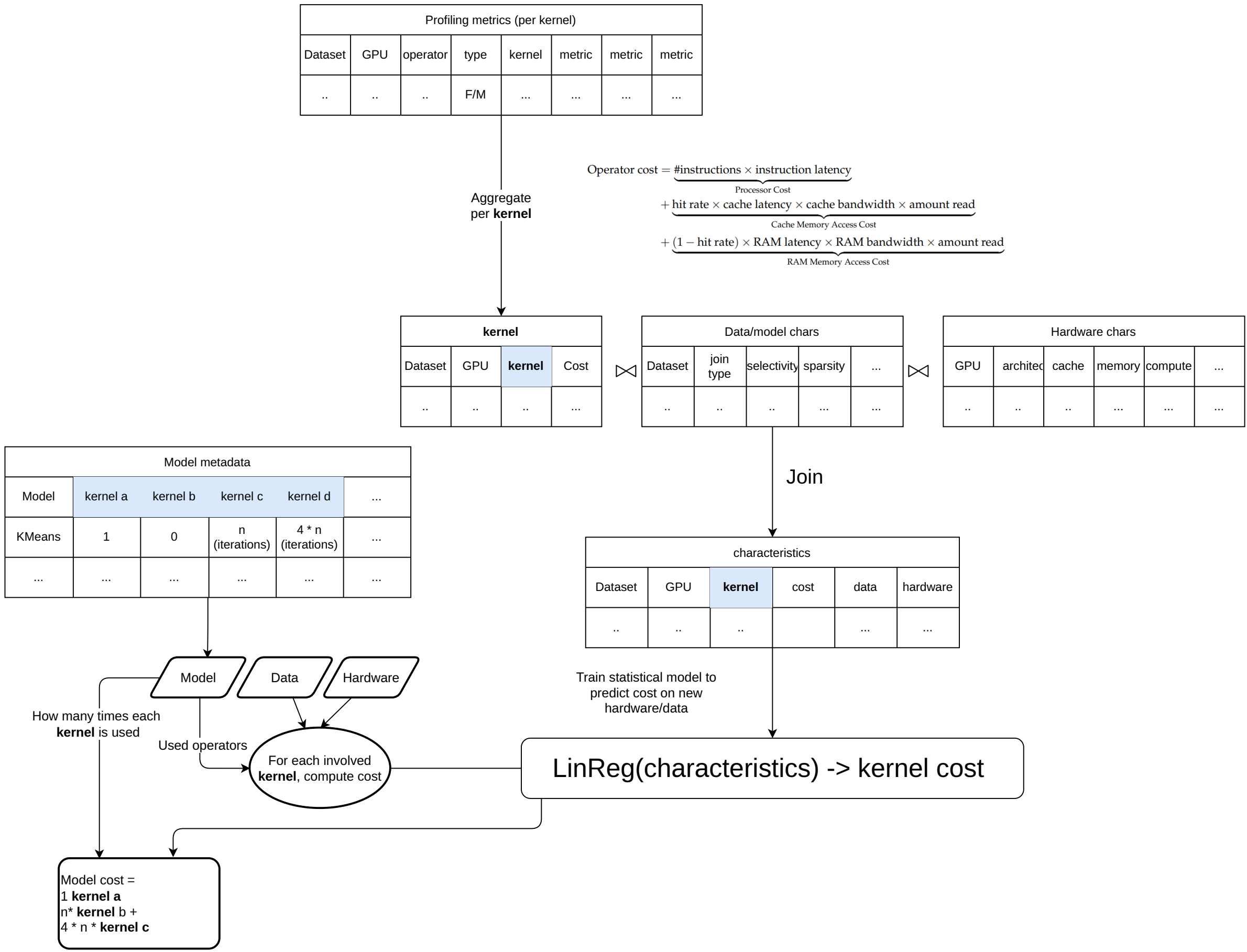
- Operator level aggregation easier to extend to new models
- Maybe less accurate

Profiling metrics (per kernel)

Dataset	GPU	operator	type	kernel	metric	metric	metric
..	..	..	F/M	...	...	...	...

Aggregate per operator

$$\text{Operator cost} = \underbrace{\#\text{instructions} \times \text{instruction latency}}_{\text{Processor Cost}}$$
$$+ \underbrace{\text{hit rate} \times \text{cache latency} \times \text{cache bandwidth} \times \text{amount read}}_{\text{Cache Memory Access Cost}}$$
$$+ \underbrace{(1 - \text{hit rate}) \times \text{RAM latency} \times \text{RAM bandwidth} \times \text{amount read}}_{\text{RAM Memory Access Cost}}$$

operator

Dataset	GPU	operator	Cost
..	..	..	...

⋈

Data/model chars

Dataset	join type	selectivity	sparsity	...
..	..	..	...	...

⋈

Hardware chars

GPU	archited	cache	memory	compute	...
..	..	..	...	...	...

Join

characteristics

Dataset	GPU	operator	cost	data	hardware
..	..	..		...	...

Train statistical model to predict cost on new hardware/data

Model metadata

Model	scalar mult factorized	scalar mutl materialized	LMM fact	LMM mat	...
KMeans	1	0	n (iterations)	4 (iterations)	...
...	...	...	...	...	...

How many times each operator is used

Used operators

Model | Data | Hardware

For each involved operator, compute cost

LinReg(characteristics) -> operator cost

Model cost =
1*scalar mult factorized +
n* LMM fact +
4 * n * LMM mat

Analytical 1: Operator level

Profiling metrics (per kernel)

Dataset	GPU	operator	type	kernel	metric	metric	metric
..	..	..	F/M	...	...	...	...

Aggregate per **kernel**

$$\text{Operator cost} = \underbrace{\#\text{instructions} \times \text{instruction latency}}_{\text{Processor Cost}}$$
$$+ \underbrace{\text{hit rate} \times \text{cache latency} \times \text{cache bandwidth} \times \text{amount read}}_{\text{Cache Memory Access Cost}}$$
$$+ \underbrace{(1 - \text{hit rate}) \times \text{RAM latency} \times \text{RAM bandwidth} \times \text{amount read}}_{\text{RAM Memory Access Cost}}$$

kernel

Dataset	GPU	**kernel**	Cost
..	..	..	...

⋈

Data/model chars

Dataset	join type	selectivity	sparsity	...
..	..	..	...	...

⋈

Hardware chars

GPU	archited	cache	memory	compute	...
..	..	..	...	...	...

Join

characteristics

Dataset	GPU	**kernel**	cost	data	hardware
..	..	..		...	...

Train statistical model to predict cost on new hardware/data

Model metadata

Model	kernel a	kernel b	kernel c	kernel d	...
KMeans	1	0	n (iterations)	4 * n (iterations)	...
...	...	...	...	...	...

How many times each **kernel** is used

Used operators

Model | Data | Hardware

For each involved **kernel**, compute cost

LinReg(characteristics) -> kernel cost

Model cost =
1 **kernel a**
n* **kernel** b +
4 * n * **kernel c**

Hybrid

Use neural network to compute weight importance of kernel cost in relation to the model/data/hardware characteristics

Profiling metrics (per kernel)

Dataset	GPU	operator	type	kernel	metric	metric	metric
..	..	..	F/M	...	...	...	...

Kernel cost embedding

characteristics

model	data	hardware
...	...	...

Characteristic embedding

fully connected layer

output layer

Factorize/ materialize