Modeling and Analysis of Inference Latency on USB Edge TPUs

Haopeng Gao (Santa Clara University),

Hyunjong Choi (San Diego State University),

Yidi Wang (Santa Clara University)

Real-Time Autonomous and
Intelligent Systems Lab
raiselab.scu@gmail.com

www.scu.edu/engineering

Challenges

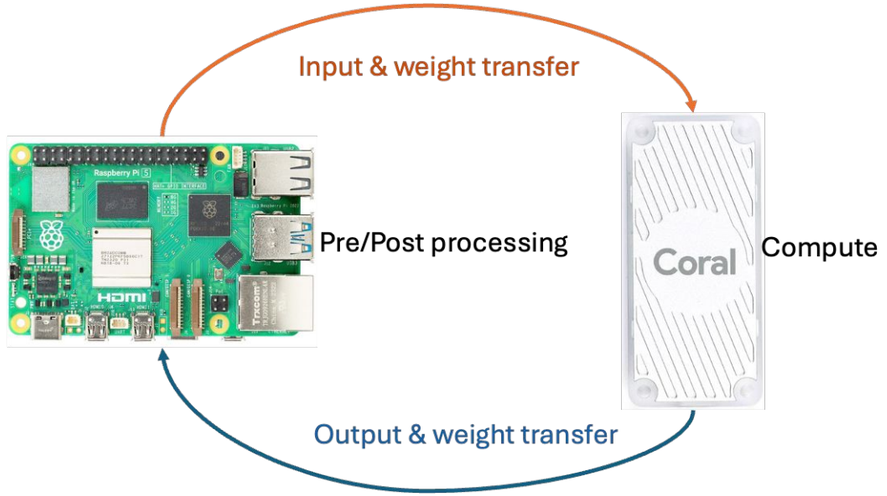
Limited SRAM

The TPU's usable on-chip SRAM is small (< 8 MB). It must hold model weights and intermediate activations.

USB as Bottleneck

When the model weights exceed the on-chip SRAM, the runtime streams weights from the host over USB during computation. USB bandwidth may also be affected by other devices on the bus.

SRAM < 8 MB



Limitations of Prior Work

Allocating

[1].Changhun Han, Hoon Sung Chwa, Kilho Lee, and Sangeun Oh. Spet: Transparent SRAM allocation and model partitioning for real-time DNN tasks on Edge TPU.

—Focuses on on-chip SRAM allocation and model partitioning on a single Edge TPU, and does not model host–device USB transfer delays or USB contention.

Profiling

[1].Binqi Sun, Bohua Zou, Yigong Hu, Tomasz Kloda, Ling Wang, Tarek Abdelzaher, and Marco Caccamo. Sapar: A surrogate-assisted DNN partitioner for efficient inferences on Edge TPU pipelines.

[2].Bohua Zou, Binqi Sun, Yigong Hu, Tomasz Kloda, Marco Caccamo, and Tarek Abdelzaher. A performance prediction-based DNN partitioner for Edge TPU pipelining.

—Rely on profiling based latency evaluation for specific partition candidates, without an explicit analytical model of USB transfer or weight streaming behavior.

Our Contribution

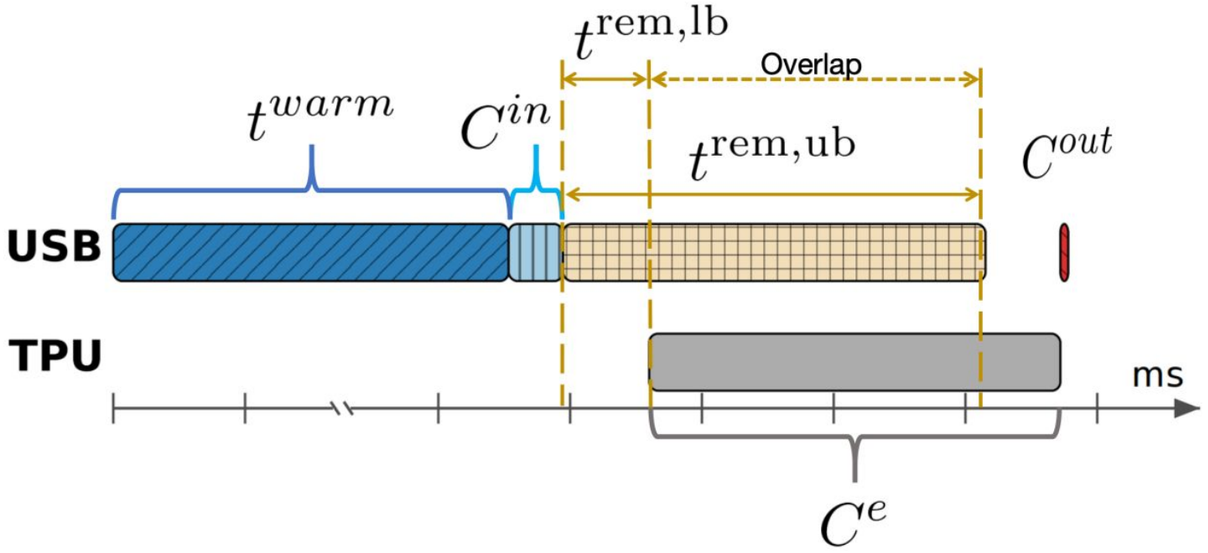
We present a **modeling-based latency framework** for USB-connected Edge TPUs that **separates compute and transfer components and captures their overlap**, enabling principled reasoning beyond profiling-based or SRAM-allocation–centric approaches.

Latency Decomposition with Sufficient TPU SRAM

- C^{in}: time to transfer model inputs from host to device before execution.
- C^{out}: time to transfer model outputs from device to host after execution completes.
- C^{e}: the intrinsic cumulative device (TPU) computation time.
- t^{tot}: time to transfer total weights of the model.

➡ The total latency is a simple summation of transfer and computation.

Latency Decomposition with Insufficient TPU SRAM



Observation: the runtime weight streaming t^{rem} via USB I/O overlaps with the computation C^{e} due to insufficient SRAM.

➡ **The total latency is not a simple summation of all the components.**

Experimental Setup

Hardware

- Raspberry Pi 5 (4× Cortex-A76 @ 2.4 GHz, 8 GB RAM)
- Google Coral USB Edge TPU (USB 3.0)

Software

- Runtime library: libedgetpu1-std
- Tracing tool: Linux usbmon

Model Benchmarks

- ResNet-50, ResNet-101, DenseNet-201, InceptionV3, and Xception for comprehensive evaluation.

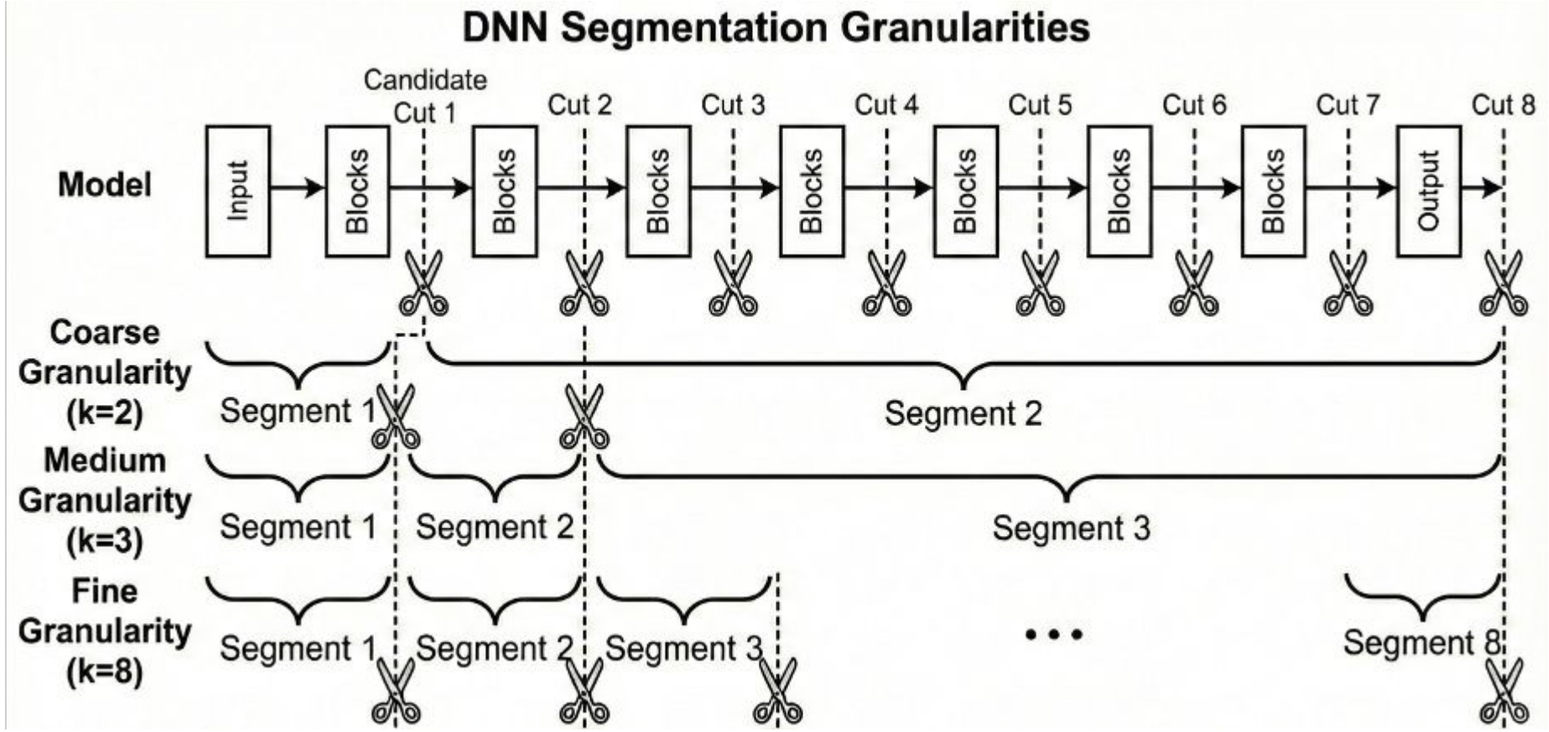
Model Partition

Each model is partitioned into 8 candidate segments. The number of active segments k is varied systematically from 2 to 8. Segments are chosen based on

(1) sizes that fit within TPU SRAM,

(2) natural structural boundaries compatible with the TPU.

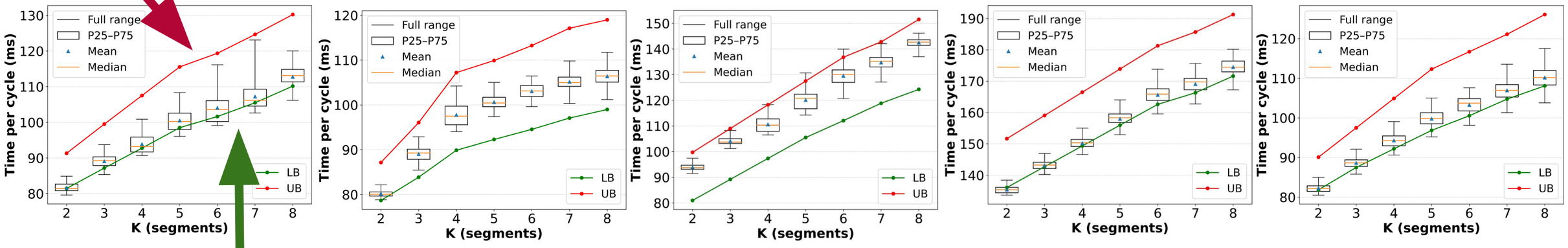
	ResNet-50	
Cut	Position	Size
1	conv4_block1_add	2.98
2	conv4_block4_add	3.31
3	conv4_block5_out	1.13
4	conv4_block6_add	1.13
5	conv5_block1_add	5.86
6	conv5_block2_add	4.33
7	conv5_block3_add	4.33
8	output	2.10



Evaluation Results

- 100 runs per configuration
- Compared against predicted lower bound (LB) and upper bound (UB)

Upper bound



Lower bound

- 98.65% samples within upper bound
- 83.6% within all bounds
- 0.85% mean violation magnitude

Our model effectively constraints most of the latency distribution within the predicted lower and upper bounds.

Future Work

Multiple Concurrent Workloads

Extend the model to efficiently manage and support multiple concurrent workloads.

Multi-Device USB Contention

Develop advanced modeling techniques to account for and mitigate contention arising from multiple devices sharing USB bandwidth.

Real-Time Schedulability Analysis

Integrate the model with real-time schedulability analysis to provide robust guarantees on task completion times and system reliability.

Optimization Decisions

Utilize the integrated model to inform and optimize key decisions such as segmentation granularity, batching strategies, and optimal device allocation.

Thank you

Modeling and Analysis of Inference Latency on USB Edge TPUs

Haopeng Gao (Santa Clara University)

Hyunjong Choi (San Diego State University)

Yidi Wang (Santa Clara University)

Our work is available at: https://github.com/raiselab-scu/tpu-usbmon-analysis