Best Splitting of DNNs for distributed deployment in edge-cloud continuum with Quality Requirements

Simone Nicosanti
simone.nicosanti.01@students.uniroma2.eu,snicosanti@tudelft.nl

Tor Vergata University of Rome, Delft University of Technology

August 17th, 2025

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Roadmap

Roadmap

Problem

- Number of quantized/not-quantized combinations can be very high ($2^{\#layers}$).
- Analysis of all possible combinations is infeasible.

Solution

- Consider only a subset of layer to be quantized.
- Those with higher number of FLOPS.

In this case, only 12 layers have been considered for quantization (most of which are *Conv* layers).

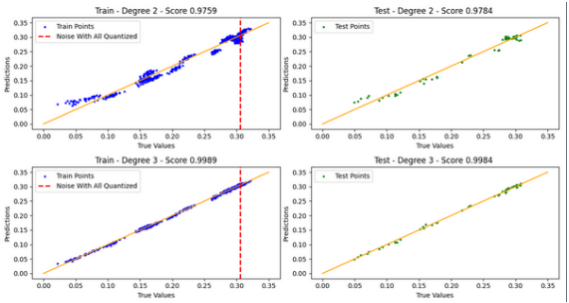

Figure 1: Noise Regressor Fit - Degree 2 and 3

Quantization noise defined as:
$\rho = \frac{1}{n} \sum_{i=1}^{n} mean(|o_i - o_{q,i}|)$

Test Context

Machines and Network

Machines GCP:

- Device
  - Machine Type: e2-standard-4
  - Docker `--cpu-affinity`: 1
- Edge
  - Machine Type: c3-standard-4
  - Docker `--cpu-affinity`: 1
- Cloud
  - Machine Type: n1-standard-4
  - GPU: nvidia-tesla-t4

Network Config:

- Bandwidth:
  - Device : Max Bandwidth: 5 MB/s
  - Edge : Max Bandwidth: 20 MB/s
  - Cloud : Max Bandwidth 100 MB/s
- Latencies:
  - Device $\leftrightarrow$ Edge: 5 ms
  - Device $\leftrightarrow$ Cloud: 55 ms
  - Edge $\leftrightarrow$ Cloud: 50 ms

	Device	**Edge**	**Cloud**
Computation Power [W]	2.9165	5.833	35
Transmission Power [W]	3.507	2.265	0.014

Table 1: Power Consumption per Server

Prediction Accuracy

Device Only

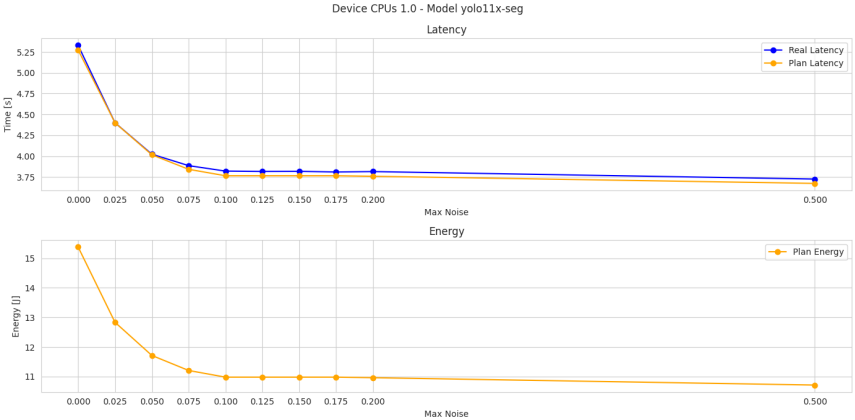

Figure 2: Device - Prediction VS Real Values

Prediction Accuracy

Device & Edge

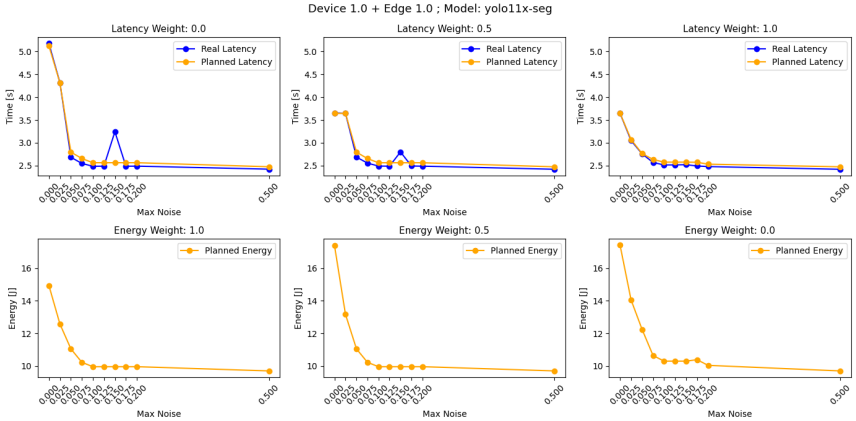

Figure 3: Device + Edge - Prediction VS Real Values

Prediction Accuracy

Device & Edge & Cloud

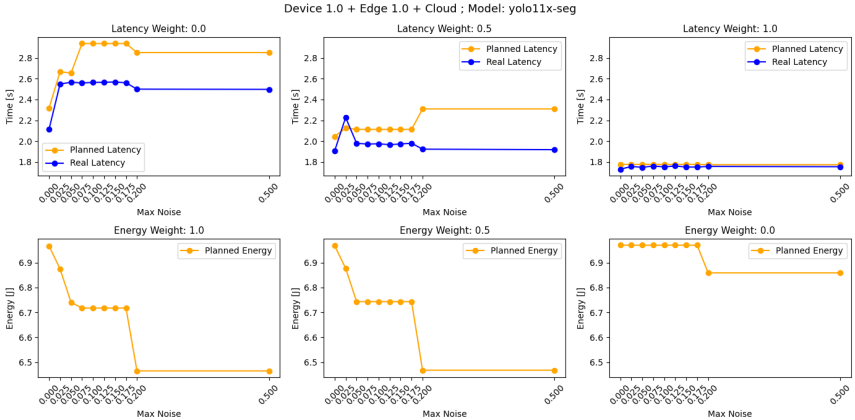

Figure 4: Device + Edge + Cloud - Prediction VS Real Values

Baseline Comparison

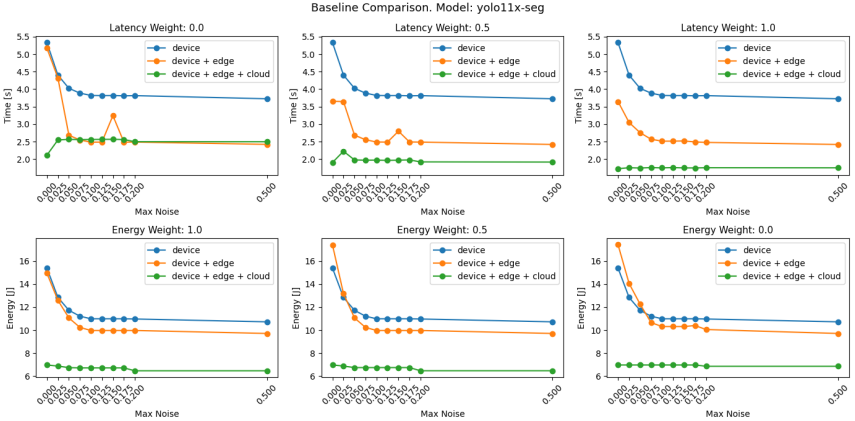

Figure 5: Baseline Comparison

Assigned Layers

Device & Edge

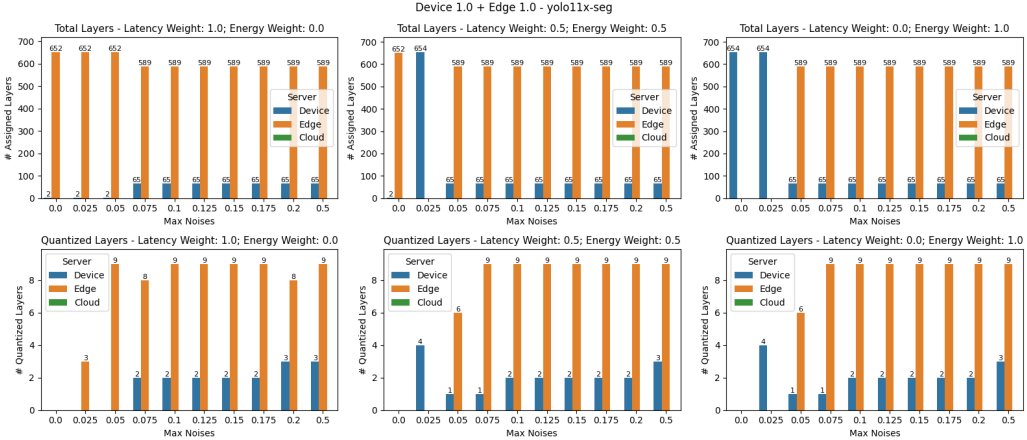

Figure 6: Device + Edge - Assigned Nodes

Assigned Layers

Device & Edge & Cloud

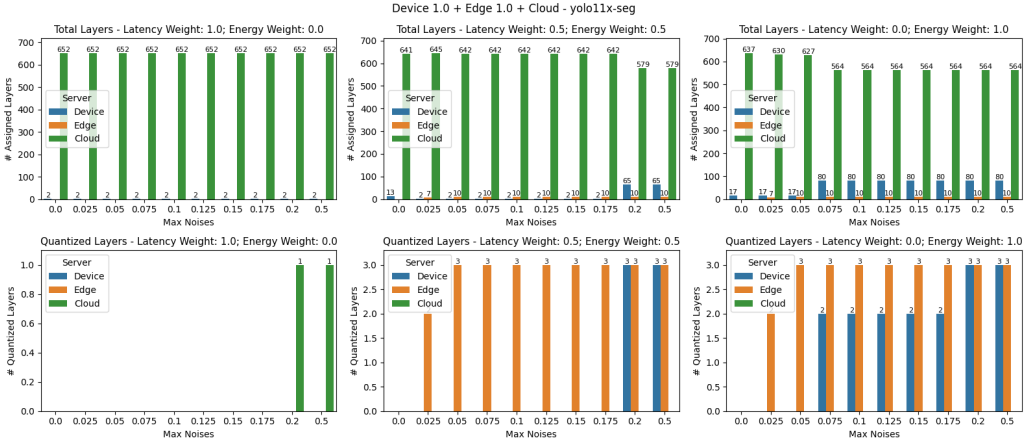

Figure 7: Device + Edge + Cloud - Assigned Nodes

Components Number

Device & Edge & Cloud

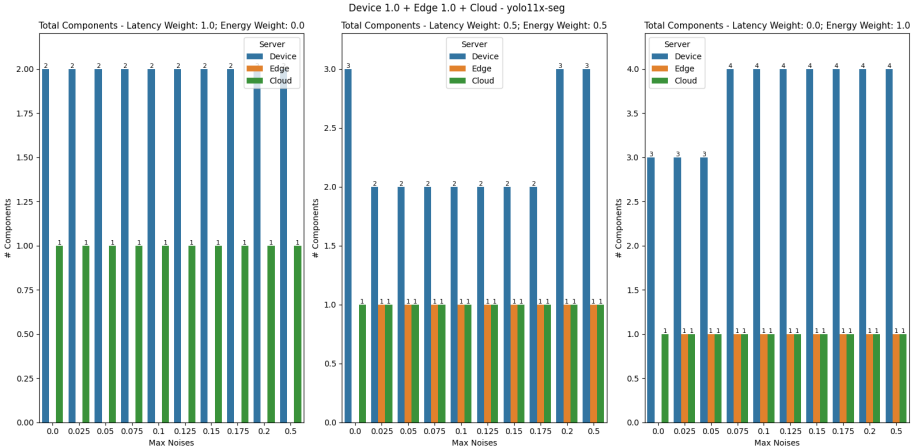

Figure 8: Device + Edge + Cloud - Assigned Components

Example of Parallel Graph Configuration:

- Max Noise: 0.5
- Latency Weight: 0.0
- Energy Weight: 1.0
- Device + Edge + Cloud

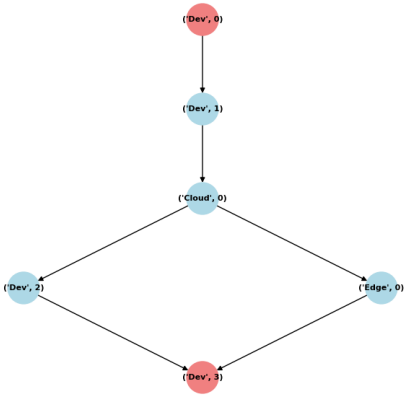

Figure 9: Parallel Execution Graph