The Missing Meter — Worked Example

Computing the served token from public data

Companion to Appendix A.9. Every number here is public and dated; the calculation reproduces from the sources cited. This document shows the served token (svt) is computable today, from a single public benchmark page, with arithmetic anyone can check.

The claim being demonstrated

The same model, at the same prices, inverts in cost-effectiveness depending only on the workload's service-level objective. Raw dollars-per-million-tokens cannot see this inversion. The served token can. That is the entire argument for why the unit is needed, reduced to arithmetic.

The unit

One **served token (svt)** is one output token delivered within its stated latency SLO and above its stated quality floor. The cost of useful work over a window is:

$$\mathbf{C = S / W}$$

where **S** is total spend attributable to the workload (all tokens billed, including failures and retries) and **W** is compliant work delivered, in millions of served tokens. A token emitted late, or wrong, contributes to S but not to W.

The public inputs

All figures from Artificial Analysis, provider benchmarks for Llama 3.3 Instruct 70B, accessed 12 July 2026. The same open-weight model is served by fifteen providers. Two are enough to demonstrate the inversion:

Provider class	Price ($/Mtok)	Output speed (tok/s)	Artifact
Cheapest	0.12	15.2	fp8-quantized
Fastest	0.64	329.6	bf16

Public price spread across all fifteen providers: 8.6x. Public speed spread: 21.7x. Same weights.

Buyer 1 — Interactive voice agent

SLO: time-to-first-token $\leq$ 1.0 s, sustained output $\geq$ 100 tok/s (below which generated speech falls behind playback). **Quality floor:** the fp8 artifact passes the buyer's acceptance test (declared in advance).

- **Cheapest provider:** delivers 15.2 tok/s. This is below the 100 tok/s floor, so *no output is compliant*. W ≈ 0. Therefore **C = S / 0** $\rightarrow \infty$. The cheapest token in the market is, for this buyer, the most expensive placement at any price.
- **Fastest provider:** delivers 329.6 tok/s at 0.95 s TTFT. Compliant. **C = $0.64 per million served tokens.**

For the voice buyer, paying 5.3x more per raw token buys the only compliant output that exists. The served-token cost of the "cheap" option is not 0.12; it is infinite.

Buyer 2 — Overnight batch summarization

Same model, same prompts. **SLO:** complete within 12 hours; no per-token latency floor.

- **Cheapest provider:** every provider now meets the deadline, so W equals billed output everywhere and **C = $0.12 per million served tokens** — equal to raw price.
- **Fastest provider:** also compliant, at **C = $0.64** — paying 5.3x for speed the deadline cannot use.

For the batch buyer, the cheapest provider wins by 5.3x — the exact placement that was worthless for voice.

The inversion

	Cheapest ($0.12/Mtok, 15.2 tok/s)	Fastest ($0.64/Mtok, 329.6 tok/s)
Voice SLO ($\geq$100 tok/s)	∞ per Msvt (zero compliant)	$0.64 per Msvt — winner
Batch SLO (12h)	$0.12 per Msvt — winner	$0.64 per Msvt

Same weights, same prices, opposite rankings. Both facts are invisible to the two metrics the industry prices with today:

- **$/GPU-hour** cannot see the workload at all.
- **Raw $/Mtok** ranks the two placements identically for both buyers — and is wrong for one of them every time.

The gap between the two correct answers — a factor of 5.3 on identical work — is the routing margin. It is computed here from nothing but a public webpage and division.

One detail the tuple forces into the open

The cheapest artifact is fp8-quantized; the fastest is bf16. These are not guaranteed to be the same quality. Under a rigorous served-token measurement the two buyers must *declare and test* their quality floor before the comparison is valid — a step current market convention never requires, and the reason the served token carries a quality leg at all.

What this example is and is not

This is a p50 illustration of the method, using median public figures. A fully conforming served-token record requires the declared percentile (typically p99), the six-field conditioning tuple, and provenance labels on every figure — specified in Appendix A of the paper. The point is narrower and complete: the served token is not a theoretical construct awaiting infrastructure. It is computable today, and the computation changes the answer.

The Missing Meter themissingmeter.org Data: Artificial Analysis, accessed 2026-07-12. Prices and latencies are volatile; re-pull before citing.