

## Turning Idle Consumer GPUs into Useful Inference Capacity

### Problem

Consumer GPUs are expensive, energy-hungry assets that sit idle most of the day, and are replaced on short cycles. Meanwhile, demand for inference keeps rising.

- Newzoo estimates **936M PC players** in 2025<sup>1</sup>.
- Steam has **132M monthly active users**<sup>2</sup>(order-of-magnitude proxy for “gaming PCs online”).

### Hypothesis

If even a small fraction of consumer-owned GPUs could be pooled, they could provide meaningful inference capacity during idle hours, extending the useful life of hardware and reducing “wasted” capital.

### Key observation: lots of idle time exists

Playtime is not 24/7. Newzoo reports PC/console playtime has fallen materially from the pandemic peak (down ~**26% from 2021–2023**)<sup>3</sup>.

That implies large windows where GPUs are underutilized.

### Back-of-envelope capacity sizing

Define

- **N** = number of eligible consumer GPUs participating
- **h/day** = average idle hours contributed
- **tps** = tokens/sec per GPU for a target model class
- **U** = utilization efficiency (network + scheduling losses)

Then

- **tokens/day** =  $N \times h/day \times 3600 \times tps \times U$

Example (illustrative):

- **N** = **100,000** participating GPUs (<< Steam MAU)
- **h/day** = **8** hours contributed
- **tps** = **1 tokens/sec** (depends heavily on model / quantization)
- **U** = **0.5**

$$\begin{aligned}\text{Tokens / day} &\approx 1e5 \times 8 \times 3600 \times 1 \times 0.5 \\ &\approx 1e5 \times 14400 \\ &\approx 1.44e10 \text{ tokens/day}\end{aligned}$$

Steam’s GPU distribution show modern-ish GPUs are widely deployed (e.g. RTX 4060 laptop GPU and RTX 3060 laptop GPU are the top entries)<sup>4</sup>. This suggests a nontrivial base of machines that can run quantized 7B-8B class models locally.

### Constraints

- Reliability: churn, NAT, sleep states, throttling
- Security: untrusted workers, data privacy, model theft
- Verification: ensuring work was not faked
- Latency: many inference workloads are latency-sensitive
- Economics: electricity cost vs. revenue, wear-and-tear, incentives

---

1 “Global Games Market to Hit \$189 Billion in 2025 as Growth Shifts to Console.”

2 Backlinko, “Steam Usage and Catalog Stats for 2025.”

3 “Playtime Has Decreased since the Start of 2021 across the PC and Console Market.”

4 “Steam Hardware & Software Survey.”

## Turning Idle Consumer GPUs into Useful Inference Capacity

Petals<sup>5</sup> demonstrates that large-scale collaborative inference is technically feasible on consumer GPUs by sharding model layers across many independent machines. The system achieves competitive real-world throughput on heterogeneous, globally distributed hardware and can outperform RAM- or SSD-based offloading for very large models.

PETALS, however, assumes cooperative participants and does not address the economic incentives required to sustain such a system in an adversarial setting, leaving incentive mechanisms to future work. In contrast, our work focuses on the incentive layer: we design a trust-unnecessary mechanism in which honest computation is the dominant economic strategy, making protocol compliance cheaper than deviation even in the presence of self-interested or malicious actors.

---

5 Borzunov et al., “Petals.”