

# Serve your next hundred customers on the GPUs you already have

PROTOPIA

Stained Glass Transform (SGT)  
the inference privacy layer for  
Sovereign Token Factories.

## The easiest supply you will ever add: idle cycles on GPUs you already run

Your anchor tenant takes the big carve-out. The capacity outside it is what you want to sell by the token, but a growing share of it leaves as **dedicated capacity**, because "I'll give you an isolated endpoint or your own box" is the only answer a sensitive customer has ever had to "who can see my data on your platform?" Built in, SGT fills those idle cycles both ways: more of each customer's work on dedicated endpoints, and the rest sold by the token.

## What dedicated capacity costs any token factory

### Supply

Every customer on dedicated capacity takes at least a whole node. The nodes run out quickly while the GPUs sit idle, and the next customer hears "no capacity."

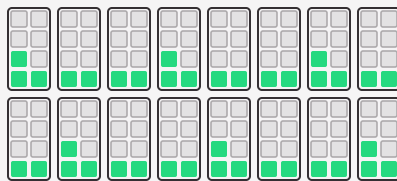
### Revenue

Idle GPUs leave money on the table. Dedicated endpoints bill by the hour whether requests arrive or not. **Every idle cycle on a dedicated GPU is a token you could have sold and never will.**

### Trust

Dedicated does not mean protected. Your stack still handles plaintext in transit, at rest and in compute: logs, traces, metadata, host and GPU memory. Reserved GPUs still leave the most sensitive workflows at home.

### SAME 128 GPUS, TWO WAYS TO SELL THEM



**DEDICATED BOXES**  
16 customers, one 8-GPU box each. About 38 GPUs of real work; **about 90 booked and idle.**



**SERVERLESS POOL + SGT**  
The same 16 customers' load in about 40 GPUs. **About 88 free to sell by the token to the next 30-plus customers.**

Illustrative: 8-GPU boxes at 30% average load; substitute your own fill rates. A pool needs headroom for coincident peaks.

## Build SGT in for every endpoint you sell, dedicated and serverless

### Expedite security reviews for highly regulated customers.

On dedicated and serverless endpoints alike, your stack receives protected representations, not plaintext, which customers can inspect.

### Unlock serverless for more customer workflows.

Any request can land on any GPU, so idle time becomes sellable time. Bill the SGT-enabled endpoint by the token like any catalog item and **monetize every cycle.**

### Add SGT early and the share you must dedicate stays small as you grow.

Customers who still want dedicated endpoints bring the workflows they kept at home onto them, because their plaintext is never on your stack.

## How Protopia AI Stained Glass Transform works

**Created per target model.** The model is not retrained or modified; the transform is what is trained. The unchanged model reads its output directly on vLLM or NIM. Transforms exist today for NVIDIA Nemotron 3 and Llama 3.3 70B among others.

**Runs where the data owner wants it, with no code changes.** As a container in the customer's environment, or as the entry point of your token factory with an API the customer can call to see the transform's output; the client still talks to an OpenAI-compatible endpoint.

**Costs the request a few percent, keeps the answers.** The transform runs once per generation call. Output quality lands within a few points of the raw model, some benchmarks at parity.

**Complements encryption, masking and redaction.** Mask or redact what policy requires; the model still has to reason over what remains, and encryption ends where the AI factory begins. SGT protects that context in transit, at rest and in compute, with no key and no inversion step. Compatible with, but not dependent on, Confidential Computing.

## Next step: stand up your own

Protopia provides the container and a transform for an evaluation period. Launch the model you already serve, route requests through the transform, inspect what reaches the endpoint, run your own benchmarks. Then price the SGT-enabled endpoint onto your catalog.

[protopia.ai](https://protopia.ai)

© 2026 Protopia. All rights reserved.