

BIND BEFORE YOU ALLOCATE — WHAT EACH DISTRIBUTED LAYER REMOVES BEFORE A GPU IS PAID FOR

illustrative rates, not measurements; the order matters: the gate runs before the cache, so a cached answer never skips consent

Bar = requests still travelling. Heavy box = the GPU bill.

STAGE	REQUESTS STILL TRAVELLING	WHAT IT REMOVES	SECURITY EFFECT
Incoming requests	<div><div></div></div> 1,000,000	every preview, render, caption and question	full payload 12 personal fields
1 · Device	<div><div></div></div> 700,000	previews and real-time 3D on the person's GPU removes 30% of what arrives	never sent
2 · Binding gate	<div><div></div></div> 630,000	unconsented, out-of-policy, unbound: denied removes 10% of what arrives	personal fields redacted, 12 to 3
3 · Edge cache	<div><div></div></div> 378,000	repeat transforms and answers from cache removes 40% of what arrives	served after the gate, never before
4 · Workers AI triage	<div><div></div></div> 113,400	small edge models settle easy cases removes 70% of what arrives	stays on the edge network
5 · GPU allocation	<div><div></div></div> 113,400	only the hard cases reach a GPU	redacted, signed payloads only

THE ILLUSTRATIVE RESULT

GPU calls: 1,000,000 to 113,400, about 89% fewer · personal-field exposures: 12,000,000 to 340,200, about 97% fewer

net saving = GPU cost avoided – (Workers + Workers AI + cache cost). Measure your own pass-through rate at each stage before sizing GPUs.