TPU COST SAVINGS CHEAT SHEET

How to Cut Your AI Costs Using Google's TPU Advantage

💡 THE BOTTOM LINE

Google's TPUs deliver **4-6x better cost efficiency** per unit of compute compared to NVIDIA GPUs. This is why Warren Buffett just invested $4.3 billion in Google, and why Apple, Anthropic, and Midjourney are all migrating to TPUs.

📊 THE PROOF: WHO'S ALREADY SAVING

COMPANY	WHAT THEY DID
Apple	Trained ALL Apple Intelligence models on 8,192 TPU v4 chips. Zero NVIDIA GPUs used for training.
Anthropic	Signed deal for up to 1 million TPUs worth tens of billions. Cited "strong price-performance and efficiency."
Midjourney	Migrated image generation infrastructure from GPUs to TPUs for inference cost savings.
CV Startup	Sold 128 H100s, switched to TPU v6e. Monthly bill dropped from $340K → $89K (74% savings).

💰 COST COMPARISON: TPU vs GPU

METRIC	NVIDIA GPU	GOOGLE TPU
Performance/Dollar	1x (baseline)	**4-6x better**
Energy Efficiency	Standard	**67% more efficient**
LLM Inference Cost	~$27/hr (8x H100)	**~$11/hr (8x TPU v5e)**

3 WAYS TO ACCESS TPU SAVINGS

1. USE GEMINI API (EASIEST) — Runs on TPUs automatically

When you call Google's Gemini API, you're already using TPU infrastructure.

- **Gemini 2.5 Flash:** $0.15/million input tokens, $0.60/million output
- **How:** Sign up at ai.google.dev → Get API key → Build

2. FREE TPU ACCESS FOR TESTING

- **Google Colab:** Free single TPU v5e for experiments
- **Kaggle Notebooks:** 8x TPU v5e chips, 20 hrs/month free
- **TPU Research Cloud:** Apply at sites.research.google/trc for free access

3. VERTEX AI (PRODUCTION SCALE)

Dr. Erin Jacques | Founder, ChatifyIT

Full TPU access for training & inference via Google Cloud Platform.
- **TPU v5e:** ~$2.70/chip/hour on-demand
- **Committed use:** Up to 57% discount with 3-year terms
- **$300 free credits:** New Google Cloud users get 90 days to experiment

✅ WHEN TPUs SAVE YOU THE MOST

BEST FOR TPUs ✓	STICK WITH GPUs
LLM inference at scale	Existing PyTorch codebases
AI agents & chatbots	Multi-cloud flexibility needs
Recommendation systems	Gaming/graphics workloads
New TensorFlow/JAX projects	Real-time low-latency needs
Fine-tuning LLMs	Small-batch experimentation

⚡ YOUR QUICK START ACTIONS

TODAY:

□ Get Gemini API key → ai.google.dev
□ Test in Google Colab with free TPU runtime
□ Calculate current AI spend vs. TPU alternative

THIS WEEK:

□ Run parallel pilot on TPU vs. current setup
□ Apply for TPU Research Cloud if qualifying

🎯 NOW THAT YOU'VE BUILT IT...

How will people discover it?

AI is changing how people search. Google isn't the only answer anymore.

Learn how to get YOUR product recommended by ChatGPT, Perplexity, Gemini & Claude.

ENROLL IN THE CHATGPT DISCOVERY METHOD

erinjacques.com/chatgpt_ai_discovery

SOURCES

- Nasdaq: "The Cost of AI Compute: Google's TPU Advantage" (4-6x cost efficiency)
- Bloomberg: Buffett acquires $4.9B stake in Alphabet (Nov 2025)
- Google Cloud Press: Anthropic TPU deal announcement (Oct 2025)
- Tom's Hardware: Apple trains AI on 8,192 TPU v4 chips (Jul 2024)
- Cloud TPU Pricing: cloud.google.com/tpu/pricing

Dr. Erin Jacques | Founder, ChatifyIT