CLAUDE ARCHITECT IN 30 DAYS · DAY 7

Prompt caching: pay less for the same tokens.

Often the cheapest win in a Claude system.

Shanker Dhand

shankerdhand.com

THE PATTERN

Most requests repeat themselves.

Same system prompt, same tools, same documents. Only the question changes.

THE SAVING

A cache hit on Claude Opus 5.5 or Sonnet 5.5 costs

0.05×

the normal input price

Source: Anthropic pricing page, Nov 2026. Most other Claude models: 0.1×.

HOW IT MATCHES

It’s a prefix match.

Change one character early in the request and everything after it misses the cache.

THE ORDER

Stable parts first.

Tools

↓

System prompt

↓

Messages: the part that changes

SILENT CACHE KILLERS

What breaks it without errors.

- A timestamp in the system prompt
- Tools listed in a different order
- JSON keys that shuffle

THE COST

Writing the cache costs a little more.

1.25× the input price for a 5-minute cache. It pays for itself after one hit.

CHECK IT

Confirm it’s working.

cache_read_input_tokens should be above zero on repeat requests. Zero means something changed.

TRY THIS TODAY

Find your largest repeated prefix.

Move it to the front of the request, mark it for caching, then check the usage numbers.

Day 8: how Claude calls your functions

Follow →