AYI NEDJIMI Consultants

GPU Side-Channel sur LLM Inference 2026 : T Attacks

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3266 mots •

Le KV-cache des LLM produit des timings observables qui leak le prompt. 2026 sur Nvidia H100, A100, B200. Defenses constant-time.

À RETENIR

A retenir — GPU Side-Channel sur LLM Inference

KV-cache timing : un attaquant co-localise sur la meme GPU detecte les c hits via timing TTFT (Time-To-First-Token).

Prompt-Cache leak (OpenAI 2024) : la mise en cache prefix prompts entre utilisateurs leak des informations sur les prompts d'autres tenants.

NVIDIA MIG isole les SM mais pas les caches L2, HBM bandwidth ni le DRA controller. Cross-tenant attaques observees

> Defenses 2026 : *constant-time inference*, *MIG strict*, *cache flush per-requ*
> *scheduling randomization*.
>
> Cas reel : Yan et al. (2025) reconstruisent 47% du prompt d'un tenant voisi
> timing GPU sur instance multi-tenant H100.

Les **gpu side channel llm** sont la frontiere R&D la moins exploree mais la plus inqu
securite LLM 2026. Le constat est mecanique : un LLM moderne (Llama 4, GPT-5,
s'execute sur GPU en partage de cache, en partage de bande passante HBM, en p
DRAM controller, et avec des optimisations de performance (KV-cache, prompt ca
speculative decoding) qui rendent l'inference time non-constante en fonction du p
ce qui n'est pas constant-time est potentiellement observable par un attaquant co
Cet article presente les attaques de timing sur LLM, le code Python pour les mesu
reproductibilite sur H100/B200, et les defenses (MIG strict, cache flush, randomisa
les fournisseurs SaaS LLM multi-tenant, les side-channels GPU representent un ri
reputational et legal majeur, qu'aucune defense unique ne couvre completement -
approche defense-in-depth combinant 7+ controles atteint un niveau acceptable
protection.

1. Genese et etat de l'art

Les attaques side-channel sur GPU remontent a Jiang et al. (2016) *A Complete Ke*
Timing Attack on a GPU contre AES sur Nvidia Tesla. Pour les NN, Hua et al. (2018

Réponse sous 24h

Devis gratuit →