GateBleed – A Timing-Only Membership Inference Attack, MoE-Routing Inference, and a Stealthy, Generic Magnifier Via Hardware Power Gating in AI Accelerators

NC STATE UNIVERSITY intel

Department of Electrical & Computer Engineering

Joshua Kalyanapu, Farshad Dizani, Darsh Asher, Rosario Cammarota, Aydin Aysu, Samira M. Ajorpaz

Artifacts Evaluated — Functional

Results Reproduced

Artifacts Available

GateBleed Discovery & Mechanism

Intel's Advanced Matrix Extensions (AMX) use power gating to save energy. We discovered that these transitions create five distinct latency states when matrix units wake from idle. This reuse-distance–dependent delay exposes timing fingerprints that can leak private model information.

Leak **stored data**:
- Spectre
- Meltdown
- Cache Telepathy
- MDS

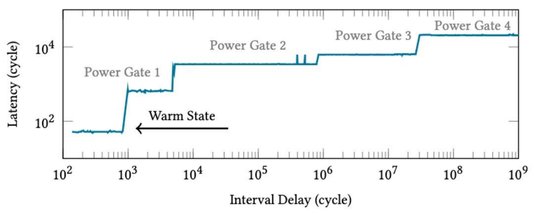
GateBleed is the first side channel to leak data **privacy** through power optimizations via timing

AI Privacy Inference Attacks

Infer information not stored on the target machine as bytes
training-set membership or expert routing choice.

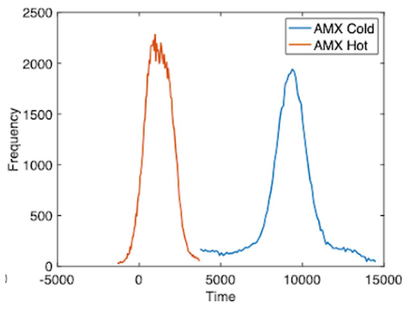
Core Idea: GateBleed Root Cause

Undocumented AMX power gating creates a measurable timing source that survives across privilege levels.



Performance States of TMUL due to Power Gating:

Each AMX power gate introduces dis latency (up to 20 000 cycles) visible e within SGX enclaves.

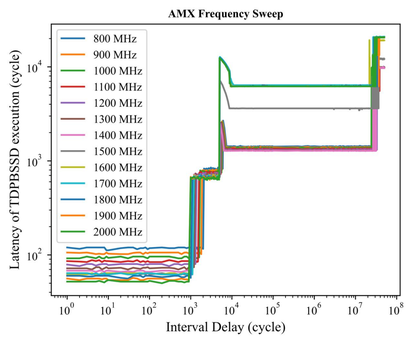


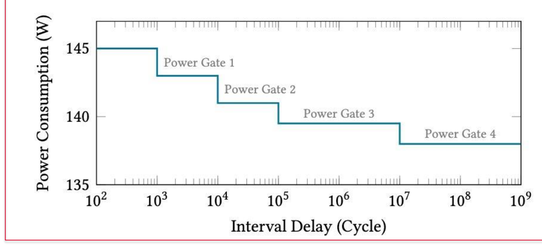
Conditional AMX Timing Difference:

AMX invocation conditional on an AI secret, the timing of the code becomes correlated with the secret.
Conditional AMX execution exists in the wild
•Mixtral - input-dependent expert routing
•BranchyNet/MSDNet - Early exiting

Frequency Independence of GateBleed Leakage:

Even with fixed CPU frequencies and Turbo Boost disabled, AMX exhibits the same five latency stages. This proves that GateBleed originates from **hardware power gating**, not from DVFS or throttling effects, broadening the attack's applicability.



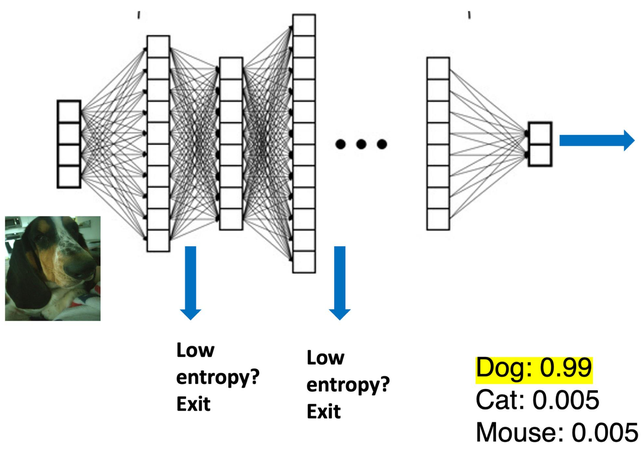


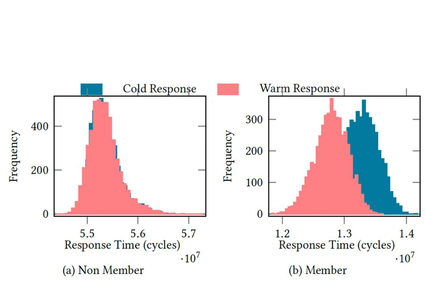
AMX Power Consumption across Power States:
Power gating transitions produce step-wise drops in wattage that align with latency stages.

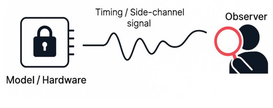
Attacks on Model Privacy via AMX Timing

GateBleed lets an attacker infer **training membership**, **expert routing**, or **early-exit decisions** from timing alone—no logits, confidences, or model access required.

Membership Interference





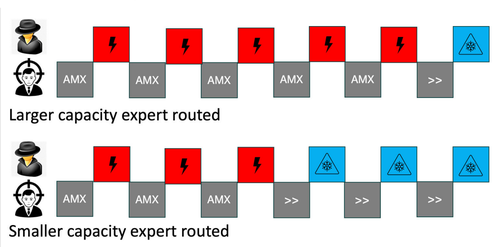


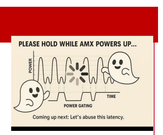
Entropy - measure of how "sure" model is of choice

Real Examples: Early Exits BranchyNet, MSDNet

Attack Results

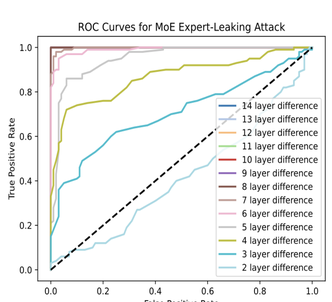
GateBleed exploits conditional AMX execution in AI workloads.



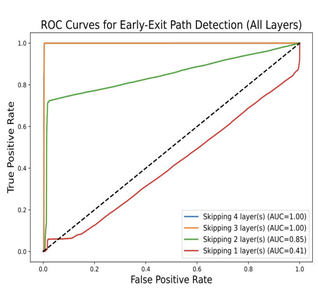


- *Top:* Transformer membership inference (training vs non-member).
- *Bottom:* HMoE expert inference (routing differences). AMX "hot" vs "cold" transitions expose per-input secret-dependent timing.

Routing & Early Exit Leakage



ROC curves show 100% detection for expert routing leakage in Mixtral-like MoE.

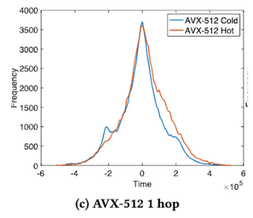


ROC curves for early-exit CNN (1–4 layer skips).

Remote Leakage, Noise Resilience, and Stealth

Beyond local timing, GateBleed amplifies signals enough to enable **remote Spectre-style leaks** over real networks and to **evade all known microarchitectural detectors.**

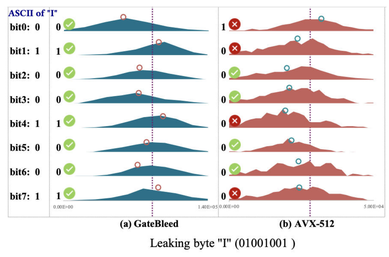
Remote Spectre Demonstration



(c) AVX-512 1 hop

We wanted to see if AI power gating can make microarchitectural attacks realistic on production network for the first time

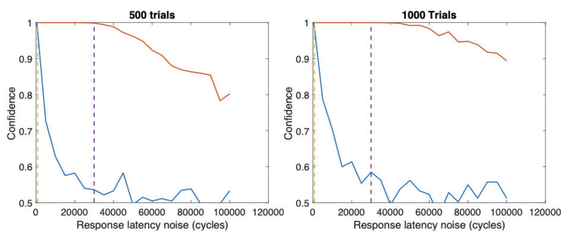
Remote Spectre Attack Net-Spectre leaks 0.000001 bps on a real production network that's about 3 months to leak a single byte!! Malware detectors flag such high repetition attacks before they leak.



Leaking byte "I" (01001001)

GateBleed clearly separates bits; AVX-512 fails under identical noise. GateBleed transmits 8-bit secrets over real traffic (0.07 bps), 70 000× faster than NetSpectre.

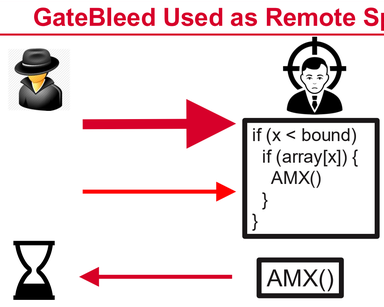
Noise Resilience & Timer Coarsening



GateBleed (orange) sustains > 99 % confidence under 30 000-cycle noise; AVX-512 fails early. AMX's 20 000-cycle gap acts as a built-in timing magnifier that defeats 10 µs timer coarsening."

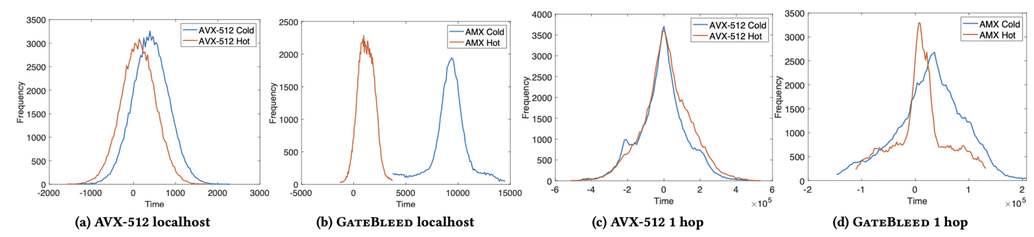
GateBleed Used as Remote Spectre Channel

1. Attacker mistrains branch predictor with in-bounds requests
2. Attacker performs out-of-bounds request
3. Response time leaks arbitrary out-of-bounds value



GateBleed enables out-of-bounds value leakage via AMX speculation.
A mispredicted branch executes AMX operations speculatively, producing measurable timing differences at the network layer.

Stealth and Detection Evasion

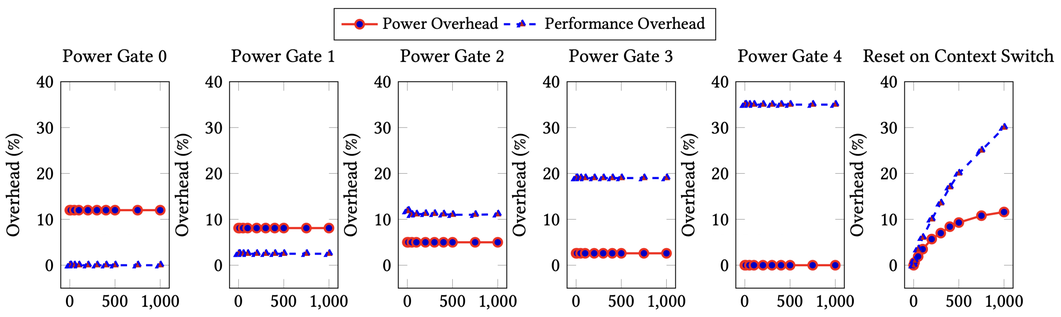


(a) AVX-512 localhost (b) GateBleed localhost (c) AVX-512 1 hop (d) GateBleed 1 hop

GateBleed produces no cache/TLB footprint—undetectable by EVAX, RHMD, or PerSpectron (< 10 % accuracy). Single-instruction latency, zero microarchitectural signature.

Remote Leakage & Stealth Resilience

GateBleed transmits the ASCII byte 'I' across a real network with measurable latency separation, while AVX-512 collapses under identical conditions. Demonstrates 0.07 bps exfiltration—70 000× faster than NetSpectre.



GateBleed's large latency gap enables **remote information leakage** even under severe network jitter and timer coarsening.
Unlike prior frequency-based channels, it amplifies timing differences enough to survive realistic internet noise, proving **first practical accelerator-level remote timing attack** feasibility.

Evading Detection with Power Gating

(i) Low repetition rates required for GateBleed to succeed
(ii) Reset phase of GateBleed has no anomalous instructions
(Passive vs. Active Reset due to automatic cool down of AMX)
(iii) Few AMX performance counters

Attack / Gadget	EVAX [8]	PerSpectron [78]	RHMD [59]
GateBleed	**10%**	**9%**	**6%**
Microscope [103]	80%	78%	63%
Flush+Flush [45]	99%	87%	72%
Binoculars [137]	98%	97%	85%
NetSpectre [97]	97%	95%	94%
Hacky Racers [128]	100%	98%	90%

Mitigations

Method	Power Overhead	Effect
Keep AMX always "warm" (microcode fix)	+12%	Stops leak
Compiler-inserted dummy AMX ops	+8–12%	Partial
Disable AMX on context switch	+2–12%	Practical trade-off

Responsible disclosure to **Intel (2023–2024)** led to mitigation in **Lenovo UEFI v3.20 (June 2024)**.

Takeaways

1. **First attack** exploiting **hardware accelerator power gating** for AI privacy leakage.
2. **Breaks hardware isolation** (VMs, OS, SGX).
3. **Bypasses existing privacy defenses** (instruction padding, confidence masking).
4. Opens new research frontier on **accelerator-level security**.

Attack Target	Accuracy	Precision	FPR
MoE Expert Routing	100%	1.0	0%
Early-Exit CNN	99.7%	0.99	0.54%
Transformer MIA	81%	0.89	16%

GateBleed achieves up to 99.7 % accuracy on CNNs and 100 % on MoE routing — all via timing

All results in this paper arereproducible within 40 minutes following the instructions at https://github.com/jkalya/gatebleed, and have been verified through the MICRO artifact evaluation.