MARCIN CHRAPEK, MARCIN COPIK, ETIENNE METAZ, TORSTEN HOEFLER, ETH ZURICH

Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs

LLM productivity frontier and their increased security needs

LLMs create a new frontier of productivity and are widely adopted

Harvard Business School

DIGITAL DATA DESIGN INSTITUTE AT HARVARD

Navigating the Jagged Technological Frontier

*"[…] (In a consultancy field study) ChatGPT-4 significantly increased performance, **boosting speed by over 25%,** human-rated performance by over 40%, and task completion by over 12%[…]"*

QuantumBlack AI by McKinsey

The state of AI: How organizations are rewiring to capture value

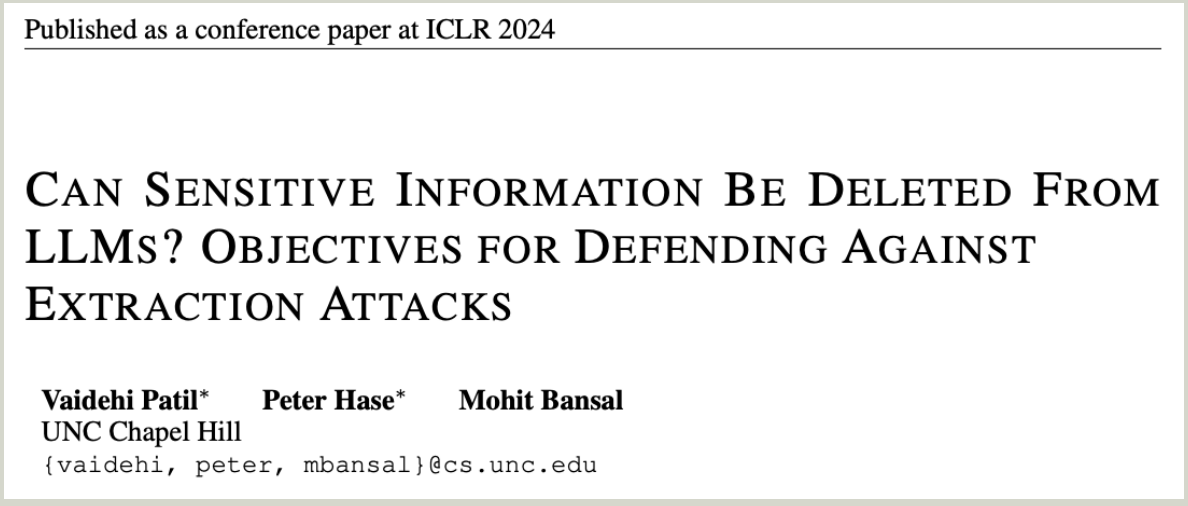
March 12, 2025 | Survey

*"[…] latest survey, **78 percent of respondents say their organizations** use AI in at least one business function (71% of which is gen AI) […]"*

LLMs and their increased security needs

LLMs create a new frontier of productivity and are widely adopted

However, they operate on confidential user on unprecedented scale



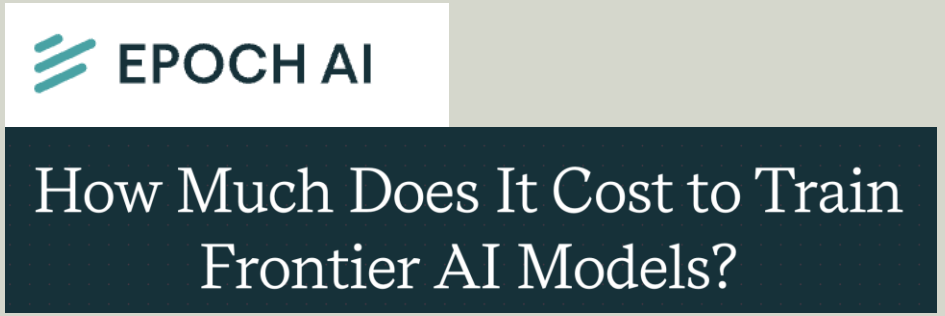
Published as a conference paper at ICLR 2024

CAN SENSITIVE INFORMATION BE DELETED FROM LLMS? OBJECTIVES FOR DEFENDING AGAINST EXTRACTION ATTACKS

Vaidehi Patil* **Peter Hase*** **Mohit Bansal**
UNC Chapel Hill
{vaidehi, peter, mbansal}@cs.unc.edu

LLMs and their increased security needs

*"[…] OpenAI's GPT-4 used an estimated **$78 million worth of compute to train, while Google's Gemini Ultra cost $191 million** […]"*



*"[…] the largest models will cost **over a billion dollars** by 2027 […]"*

If trained or fine-tuned on expensive datasets, models are also expensive.

LLMs and their increased security needs

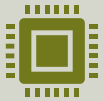
LLMs create a new frontier of productivity and are widely adopted

ANTHROP\C

"[…] At Anthropic, we […] ***ensure that our users' trust is warranted****—and in fact, to ensure that their* ***trust is cryptographically guaranteed****.[…]"*

ed scale

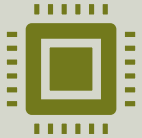
so expensive.



For compute, efficiency, and scalability, they need clouds but from security perspective this is challenging

Main research questions

LLMs create a new frontier of productivity and are widely adopted



For practical deployments:

How can we securely deploy LLMs?

How viable is that?

What are the overheads and limitations?

What can we learn for other workloads?

For compute, efficiency, and scalability, we need clouds but from security perspective this is challenging

Potential solutions for securing LLMs

Machine Learning methods

– signature embedding
– passport authentication
– backdoor authentication
– model watermarking

Require retraining

No confidentiality of user prompts

No measurable security properties

Homomorphic encryption

– conducting operations on encrypted data
– strong security guarantees

Works only in well structured examples, e.g., HEAR [1]

Doesn't generalize and takes minutes to conduct simple RESNET and MNIST inference

Trusted Execution Environments

– Secure, isolated, and verifiable hardware environment ensuring confidentiality and integrity

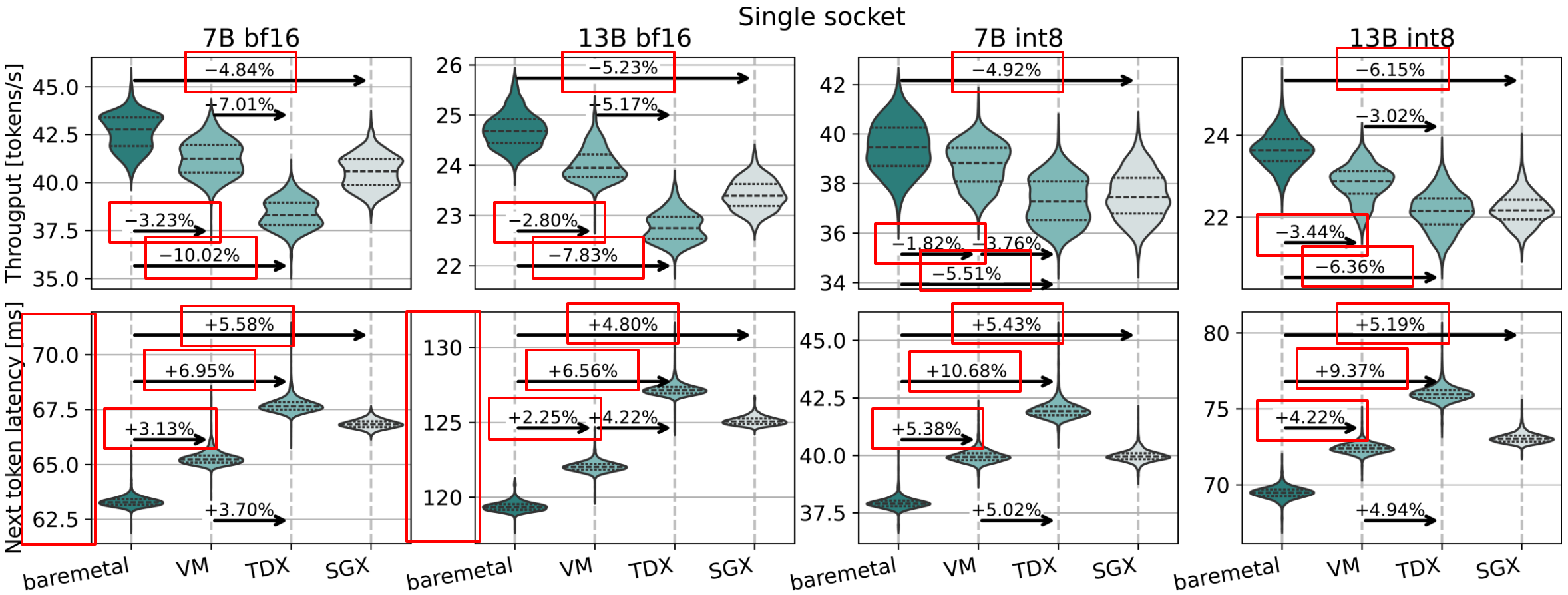
Good ratio between provided security and performance

How viable are LLMs in CPU (SGX, TDX) and GPU (NVIDIA H100) TEEs (costs and limitations)?

[1] "HEAR: Homomorphically Encrypted Allreduce," M. Chrapek, M. Khalilov, and T. Hoefler @ SC23, Best Paper and Best Reproducibility Awards

Single socket CPU TEE (Llama2)

Overheads consistent <10%.

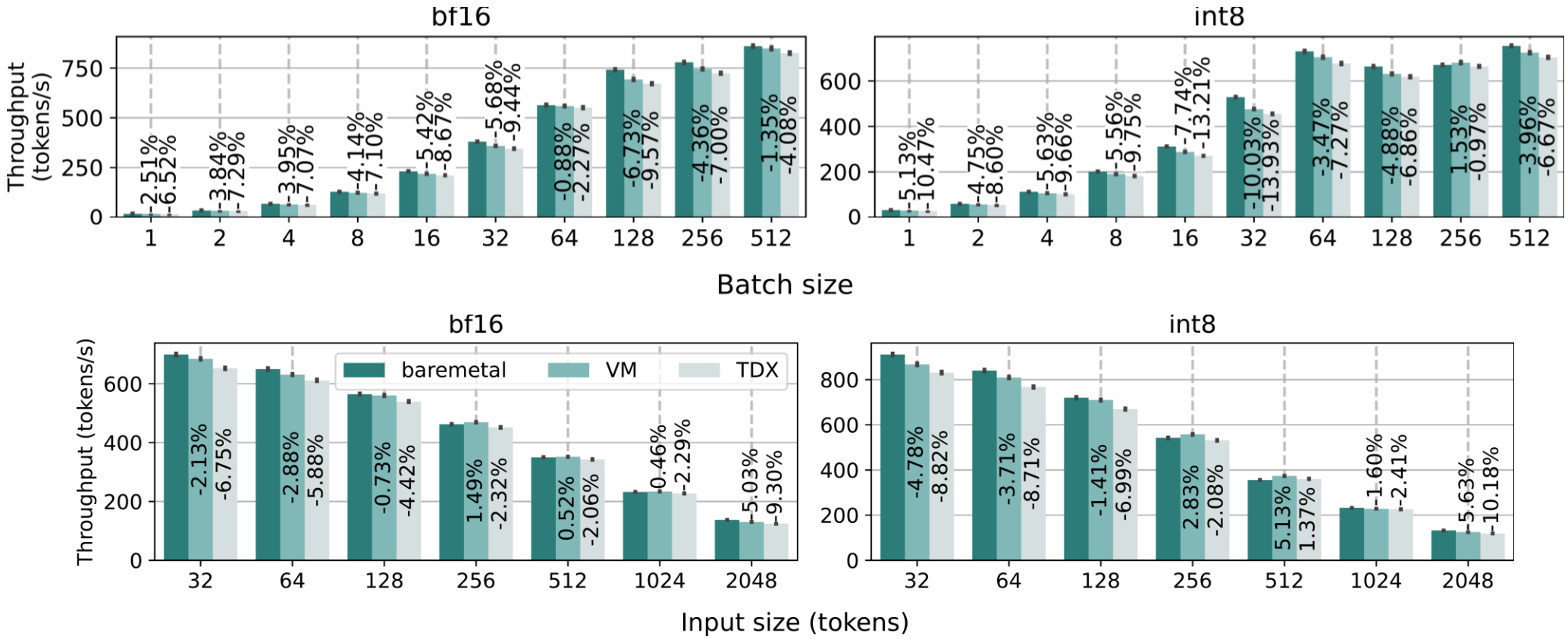


Latency considerably below the human reading speed of 200ms/word (typical golden standard)

SGX typically performs better than TDX but worse than VM.
TDX introduces a 1-5% virtualization tax on top of its security overheads.

Models, batch & input sizes

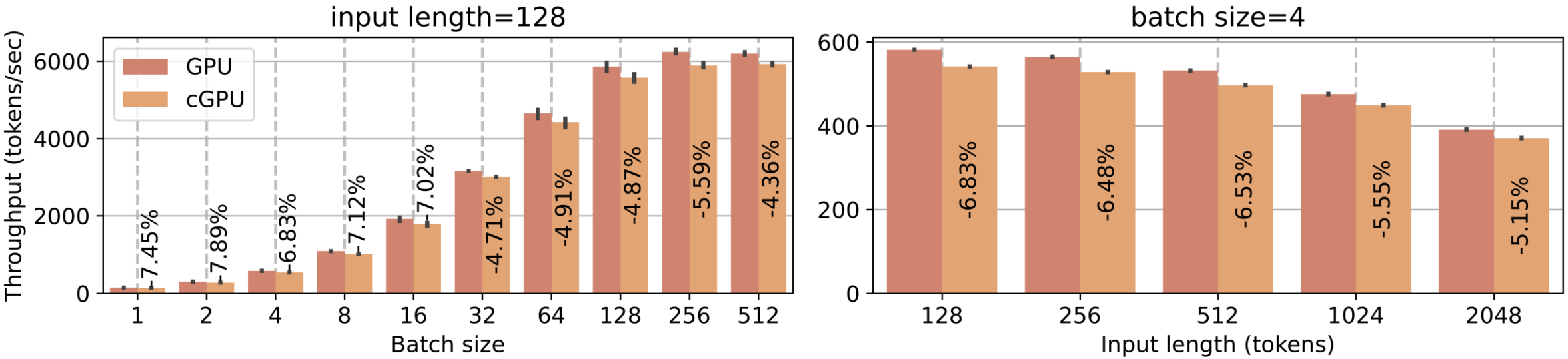
We found similar overheads (0.1-13.1%) for Llama3 8B, GPT-J 6B, Falcon 7B, Baichuan2 7B, and Qwen 7B.



These overheads get minimized with larger batch/input sizes as the workload gets more compute bound.

Confidential GPUs (cGPUs)

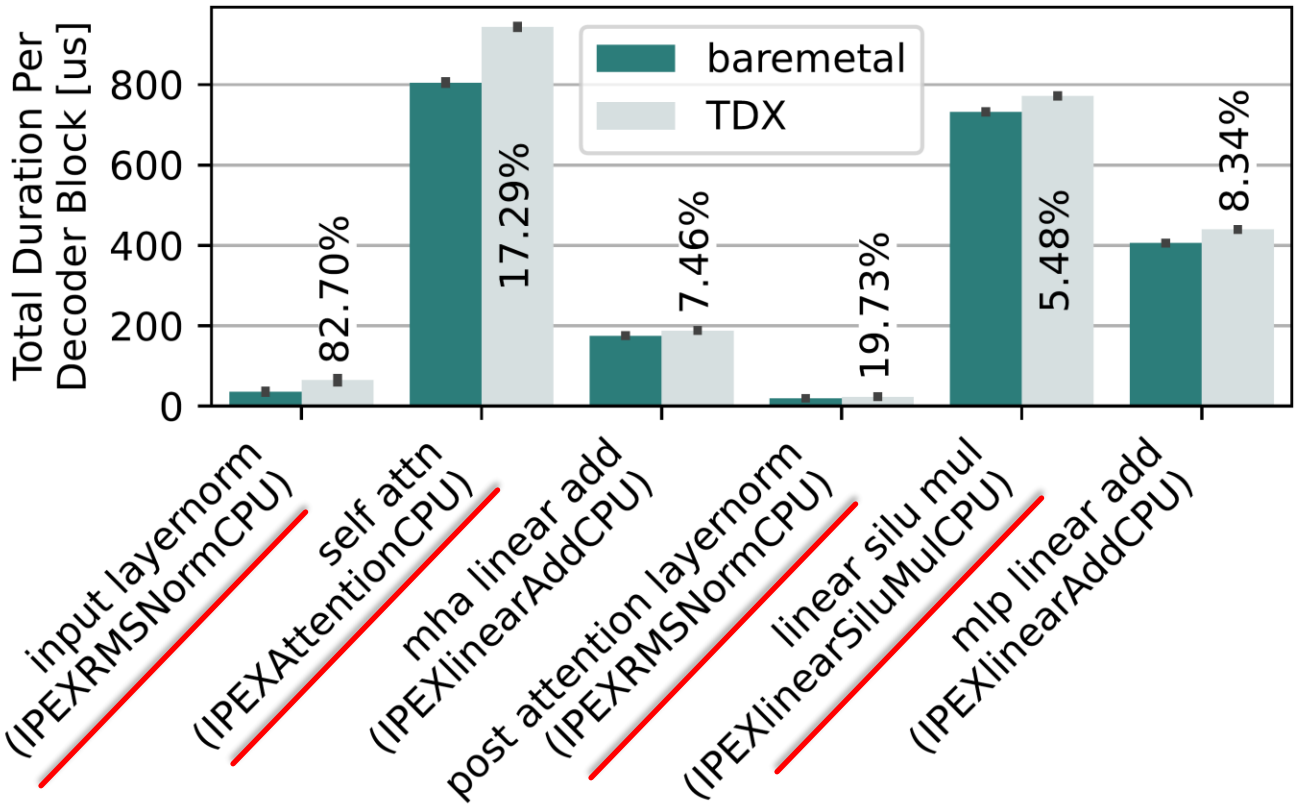
We conducted similar experiments on H100s (no baremetal as we rented an instance)



cGPUs achieve similar performance overheads to CPUs (4-8%).

Increased batch and input sizes make the workload more compute bound amortizing overheads.

Are there any other factors that impact the overhead?



We traced inference and parsed the results to understand how much time is spent in each layer.

99.9% of the time was spent in decoder blocks.

Layernorms introduce largest overhead but have large noise and contribute only 3% of total block time.

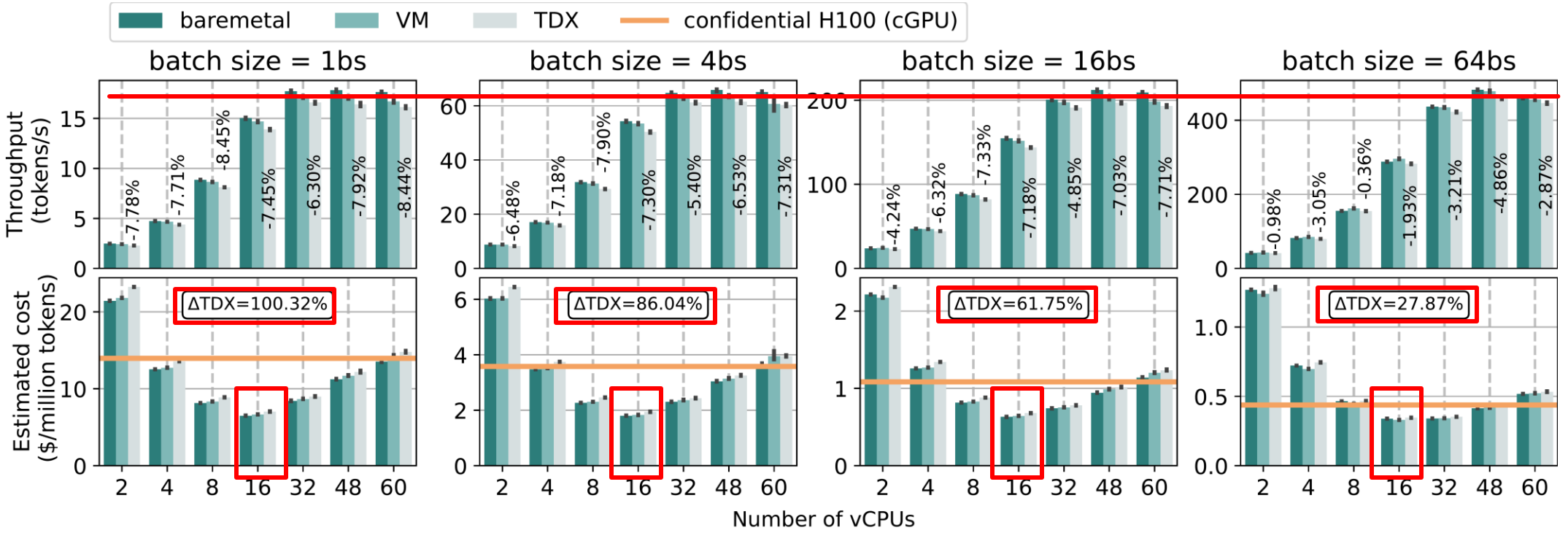
SiLU and attention contribute most to overall cost. Expected as they have a large data movement [3].

These are impacted by arithmetic intensity. Apart from batch and input size scaling, using AMX and frameworks minimizing data transfers is critical.

[3] Ivanov, A., Dryden, N., Ben-Nun, T., Li, S. and Hoefler, T., 2021. Data movement is all you need: A case study on optimizing transformers.

Cost efficiency – CPU TEEs vs cGPUs

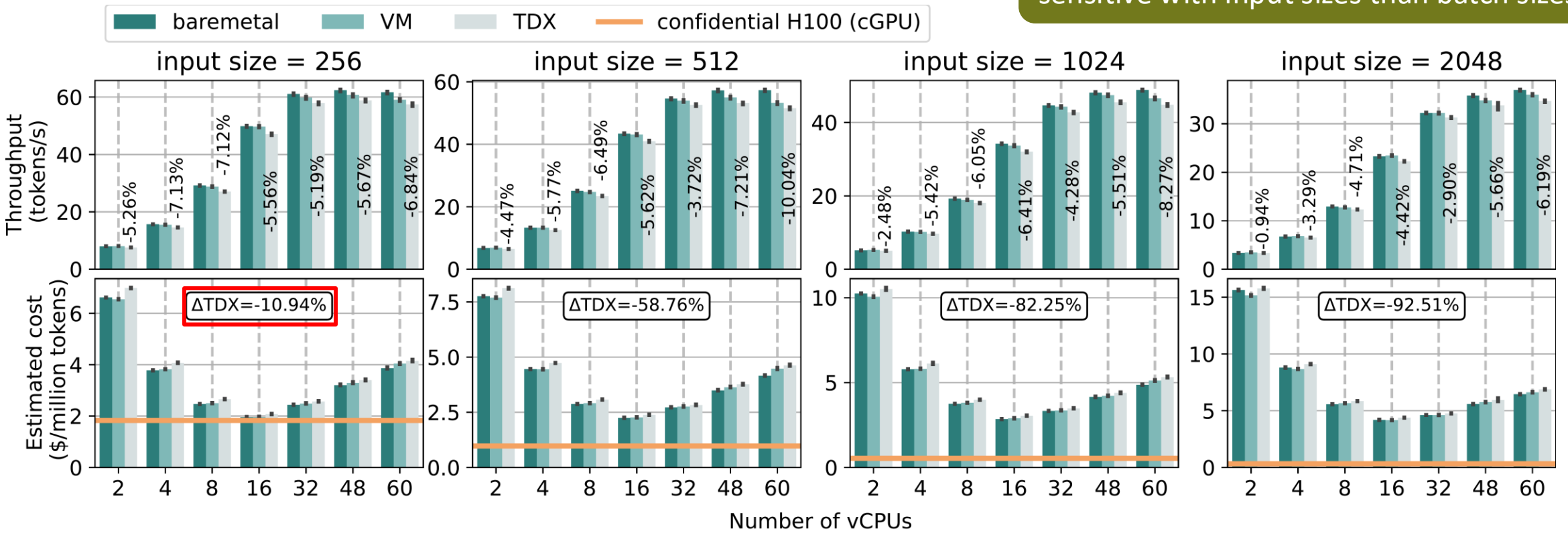
TDX saturating at 32 cores (workload turns into a memory bound problem),
16 core most efficient cost wise
No need for a whole large CPU to achieve top performance!



TDX beats cGPUs in cost competitiveness until bs=128

Cost efficiency – when CPUs start losing out

Performance of TDX considerably more sensitive with input sizes than batch sizes.



CPUs win when the GPU is unsaturated (small models/batch sizes/input sizes). No support for weaker, more cost-efficient GPUs such as A100.

Another time when CPU TEEs win is in hybrid setups where model is too large for the cGPU and partially resides on the host memory.

Conclusions

LLMs make secure and private computations critical

Deploying practical confidential LLMs currently can be done only using TEEs and can cost <10%

CPU TEEs can be even 100% cheaper than GPU TEEs in the case of confidential LLMs

Maximize algorithmic intensity for lower overhead (e.g., minimize data movement, use AMX, large input/batch sizes)

Scaling is challenging:
- CPU TEEs not supporting NUMA/hugepages
- GPU TEEs not protecting communication

RAG can achieve a similar overhead to inference

More of SPCL's research:

youtube.com/@spcl — 210+ Talks

twitter.com/spcl_eth — 1.6K+ Followers

github.com/spcl — 5.6K+ Stars

... or spcl.ethz.ch

arxiv.org/abs/2509.18886

github.com/spcl/confidential-llms-in-tees