PRACTICAL PERFORMANCE OF CKKS AND ENCRYPTED TRAINING AND INFERENCE FOR CLASSIFICATION

Damien Stehlé & Junbum Shin

{DAMIEN.STEHLE,JUNBUM.SHIN}@CRYPTOLAB.CO.KR

NIST WPEC24 - Online - September 25, 2024

HEAAN
CRYPTO LAB

REMINDERS

CKKS is a fully homomorphic encryption scheme:

$$\forall f, m_1, \ldots, m_k : \quad \mathrm{Dec}\left(\mathrm{Eval}\left(f;\ \mathrm{Enc}(m_1), \ldots, \mathrm{Enc}(m_k)\right)\right) \approx f(m_1, \ldots, m_k)$$

Plaintext space: vectors of $\mathbb{C}^{N/2}$ (up to some precision)

- add in //
- multiply in //
- conjugate in //
- rotate the coordinates

CKKS is **level-based**

• mult	consumes 1 level
• add, conj & rot	consume 0 level
• bootstrapping (BTS)	regains level

IND-CPA security from

RingLWE & a circular security assumption

(closely related to NIST choices for pq-crypto standardization)

BOOTSTRAPPING IS FAST

$N = 2^{16}$ Precision ≈ 22 bits Remaining levels: 10	**CPU** Single-thread, AVX512 Intel Xeon Gold 6342 @2.8GHz
Real-BTS ($N/2$ real numbers)	**5.3 s**
Complex-BTS ($N/2$ complex numbers or N real numbers)	**6.9 s**

HEaaN library, binaries available at heaan.it

[JKACL21] W. Jung, S. Kim, J. H. Ahn, J. H. Cheon, and Y. Lee. Over 100x faster bootstrapping in fully homomorphic encryption through memory-centric optimization with GPUs. TCHES, 2021.

BOOTSTRAPPING IS FAST

$N = 2^{16}$ Precision ≈ 22 bits Remaining levels: 10	**CPU** Single-thread, AVX512 Intel Xeon Gold 6342 @2.8GHz	**GPU** NVIDIA GeForce RTX 4090 (based on [JKACL21])
Real-BTS ($N/2$ real numbers)	**5.3 s**	**49 ms**
Complex-BTS ($N/2$ complex numbers or N real numbers)	**6.9 s**	**61 ms**

HEaaN library, binaries available at heaan.it

AMORTIZED MULTIPLICATION COST

Cost of a ct-ct multiplication, amortized over:

- slots
- a full BTS loop iteration
- several ciphertexts

$$73.6\ ns$$

(GPU, NVIDIA GeForce RTX 4090)

[DMPS24] N. Drucker, G. Moshkowich, T. Pelleg, and H. Shaul. BLEACH: Cleaning errors in discrete computations over CKKS. J. Cryptol., 2024
[BCKS24] Y. Bae, J. H. Cheon, J. Kim, and D. Stehlé. Bootstrapping bits with CKKS. Eurocrypt 2024.
[BKSS24] Y. Bae, J. Kim, D. Stehlé, and E. Suvanto. Bootstrapping integers with CKKS. Asiacrypt 2024.

BINARY CIRCUITS

CKKS is usually thought of as designed for **real/complex numbers**
But it can be used for **binary** computations! [DMPS24]

Bootstrapping can be optimized for such plaintext formats [BCKS24,BKSS24]

	CGGI	[DMPS24] (naive, our code)	[DMPS24] (optimized, our code)	[BCKS24]	[BKSS24]
Throughput (amortized time / binary gate) **single-thread CPU**	~10ms	$92.6\mu s$	$27.7\mu s$	$17.6\mu s$	$7.39\mu s$

LLAMA2-7B... HOMOMORPHICALLY!

One of Meta’s transformer-based LLMs (with 2^7 tokens)

LLAMA2-7B… HOMOMORPHICALLY!

One of Meta’s transformer-based LLMs (with 2^7 tokens)

RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	3 in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
Softmax	dim = 2^7	2^{12} in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	once
RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{13.4} \times 2^{12} \times 2^7$	2 in //
SILU		$2^{20.4}$ in //
pt-ct matrix mult	$2^{12} \times 2^{13.4} \times 2^7$	once

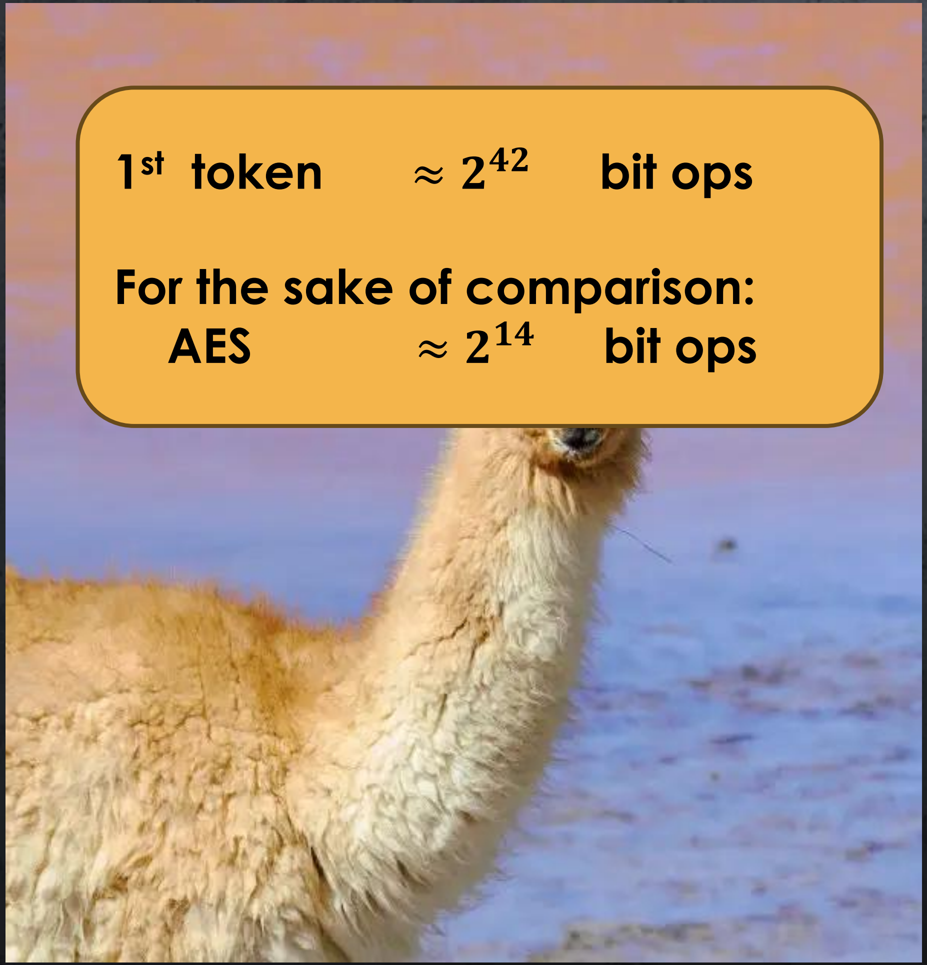
LLAMA2-7B… HOMOMORPHICALLY!

One of Meta’s transformer-based LLMs (with 2^7 tokens)

RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	3 in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
Softmax	dim = 2^7	2^{12} in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	once
RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{13.4} \times 2^{12} \times 2^7$	2 in //
SILU		$2^{20.4}$ in //
pt-ct matrix mult	$2^{12} \times 2^{13.4} \times 2^7$	once

Repeat 32 times (!#?!!)
This gives the first output token

LLAMA2-7B… HOMOMORPHICALLY!



1st token $\approx 2^{42}$ bit ops

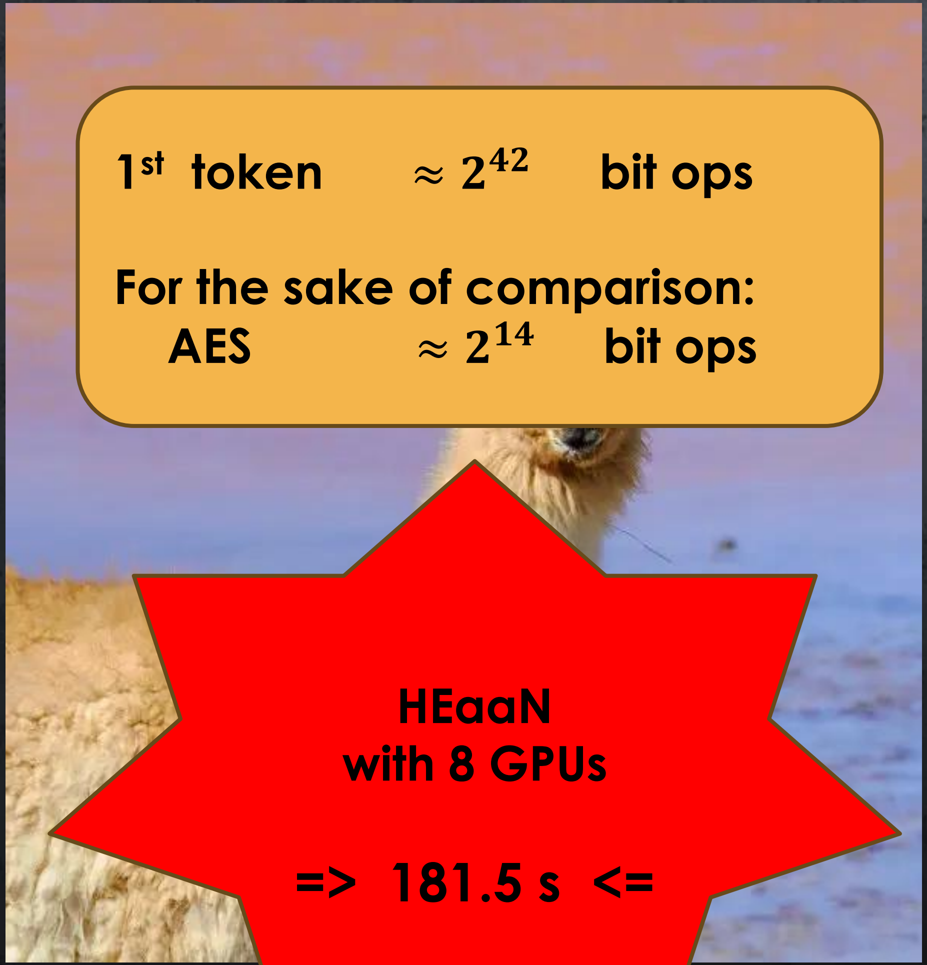
For the sake of comparison:
AES $\approx 2^{14}$ bit ops

One of Meta’s transformer-based LLMs (with 2^7 tokens)

RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	3 in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
Softmax	dim = 2^7	2^{12} in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	once
RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{13.4} \times 2^{12} \times 2^7$	2 in //
SILU		$2^{20.4}$ in //
pt-ct matrix mult	$2^{12} \times 2^{13.4} \times 2^7$	once

Repeat 32 times (!#?!!)
This gives the first output token

LLAMA2-7B… HOMOMORPHICALLY!



1st token $\approx 2^{42}$ bit ops

For the sake of comparison:
AES $\approx 2^{14}$ bit ops

HEaaN
with 8 GPUs

=> 181.5 s <=

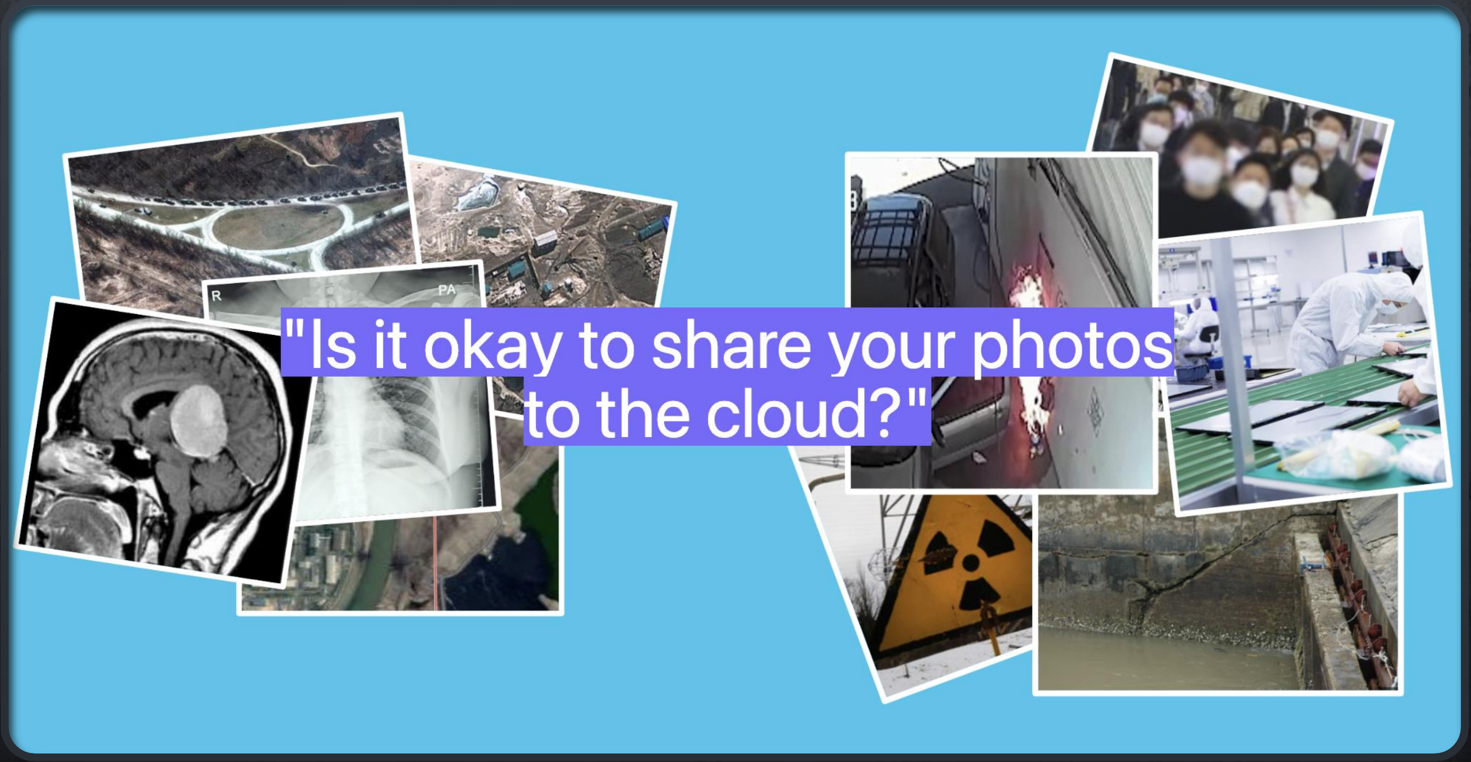
One of Meta's transformer-based LLMs (with 2^7 tokens)

RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	3 in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
Softmax	dim = 2^7	2^{12} in //
ct-ct matrix mult	$2^7 \times 2^7 \times 2^7$	2^5 in //
pt-ct matrix mult	$2^{12} \times 2^{12} \times 2^7$	once
RMSNorm	dim = 2^7	2^{12} in //
pt-ct matrix mult	$2^{13.4} \times 2^{12} \times 2^7$	2 in //
SILU		$2^{20.4}$ in //
pt-ct matrix mult	$2^{12} \times 2^{13.4} \times 2^7$	once

Repeat 32 times (!#?!!)
This gives the first output token

APPLICATION - PRIVACY PRESERVING CLASSIFICATION

Cloud service providers (e.g. AWS, Azure, GCP) provide automated machine learning for Classification



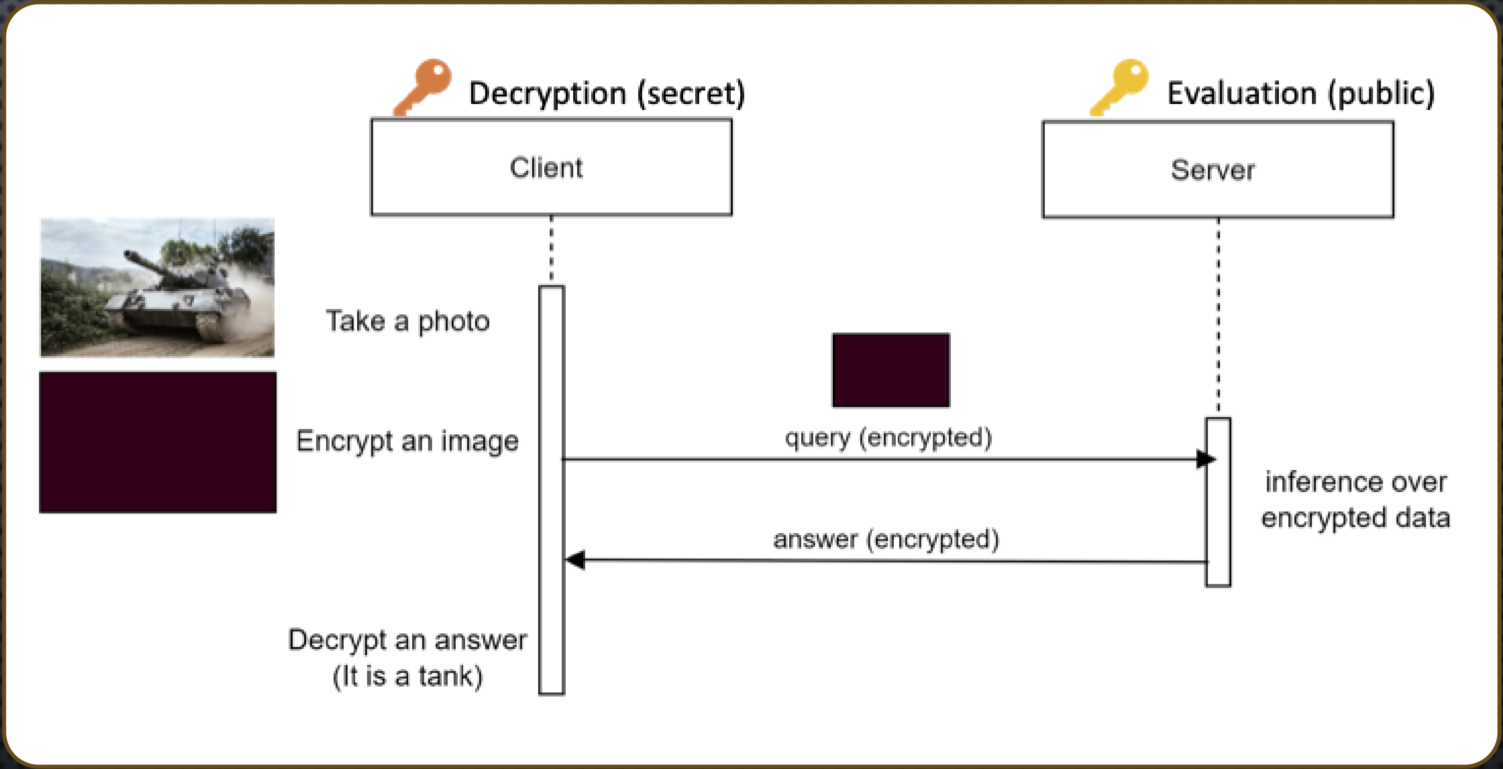
"Is it okay to share your photos to the cloud?"

HOW TO SHARE DATA SECURELY?

Encrypted Classification

- To send data after encrypting it using homomorphic encryption
  - Typical use case of homomorphic encryption.
- Too slow?

  0.2 sec. / Encrypted inference



* *GCP G2-STANDART-16 - INTEL® XEON® PLATINUM 8273CL*

AUTOFHE – PRACTICAL ENCRYPTED CLASSIFICATION

AutoFHE supports Image/Text Classification (https://www.autofhe.com/)

- Image classification: classifying military/medical photos
- Text CLASSIFICATION: sentiment/intent analysis

Demo: Training (Image)
Model training to detect military vehicle

* Vision transformer / Custom dataset

Demo: Inference (text)
Model Inference for sentiment analysis

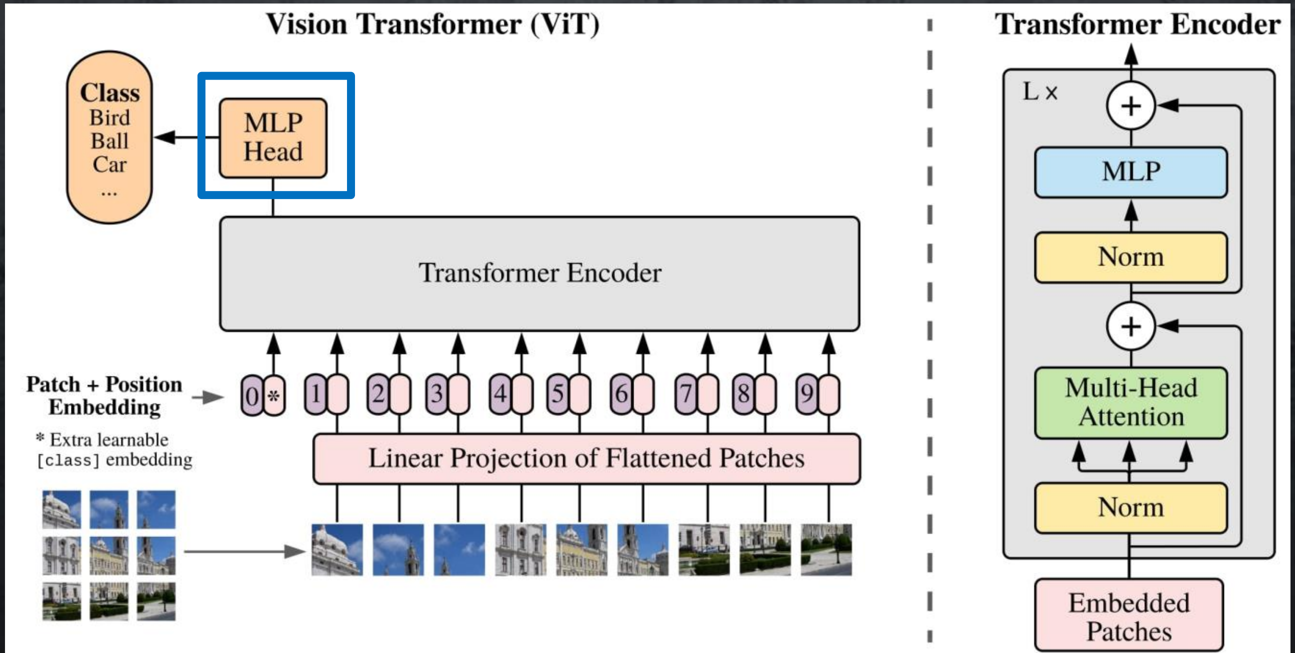
* E5 / Amazon PoPolarity dataset (5000)

[ICML23] S. Lee, G. Lee, J. Kim, J. Shin, M. Lee, HETAL: Efficient Privacy-preserving Transfer Learning with Homomorphic Encryption, ICML 2023

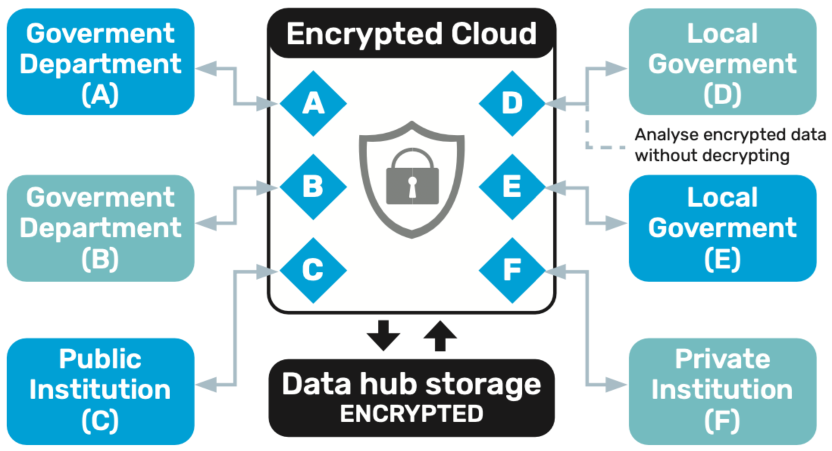
AUTOFHE: FHE + TRANSFORMER ENCODER

- Image: Vision Transformer
- Text: BERT, MPNET, E5

Public transformer encoder runs on client
→ Encrypt the encoder output @ client
→ MLP runs on the server / without decryption
(No attack surface on server)



Hub&Spoke Model from Statistics Korea



Goverment Department (A)

Goverment Department (B)

Public Institution (C)

Encrypted Cloud

Data hub storage
ENCRYPTED

Local Goverment (D)

Analyse encrypted data without decrypting

Local Goverment (E)

Private Institution (F)

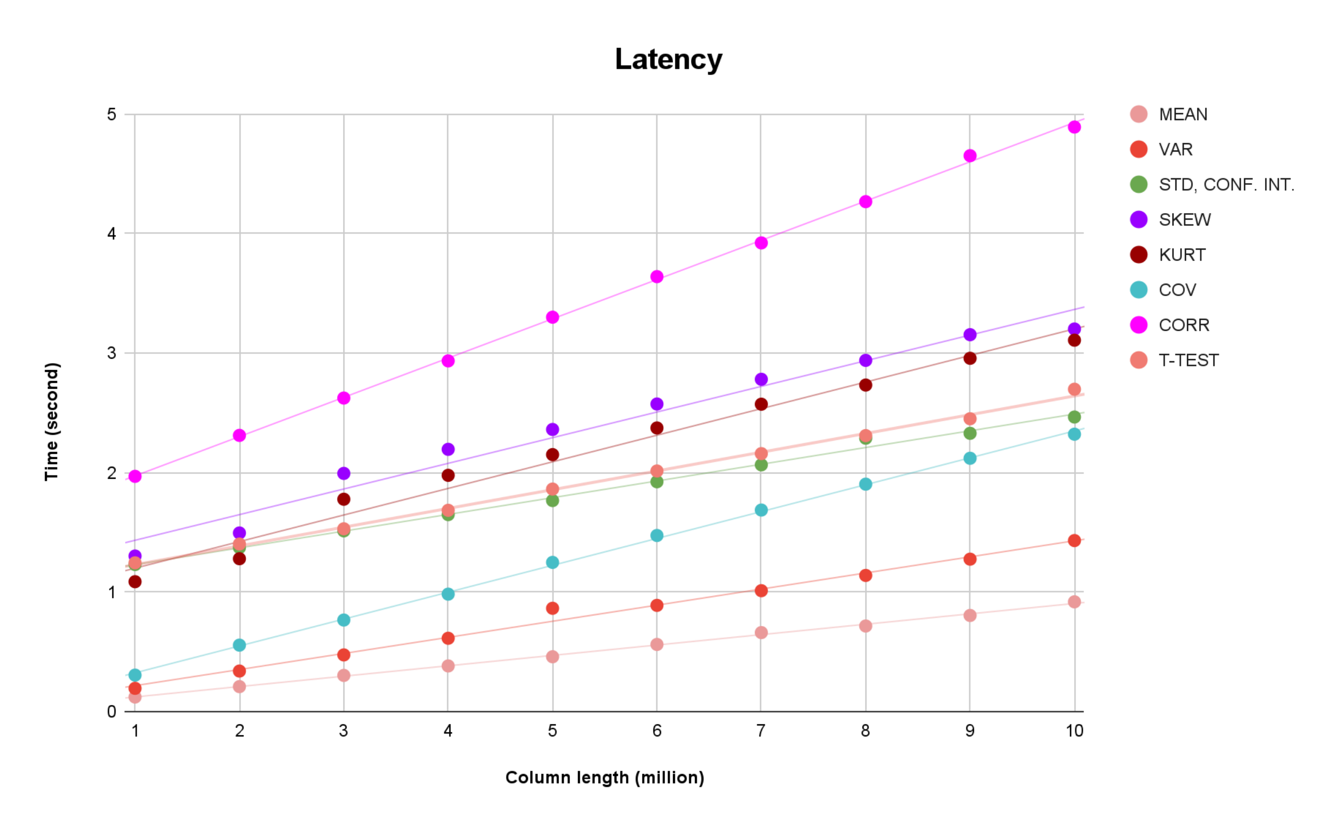
OTHER APPLICATION – ENCRYPTED STATISTICS

- Performance
  - 1M data < 2 seconds
  - 10M data < 5 seconds

CPU: INTEL® XEON® GOLD 6248 CPU @ 2.50GHZ,

• CPU: Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz,

GPU: NVIDIA A40



CONCLUSION

CKKS will be **even faster** soon!

- Algorithmic improvements (arithmetic, bootstrapping, HE-softmax, hom. linear algebra)
- Exploit parallelism more
- Towards dedicated chips

=> Even more privacy-preserving applications

QUESTIONS?