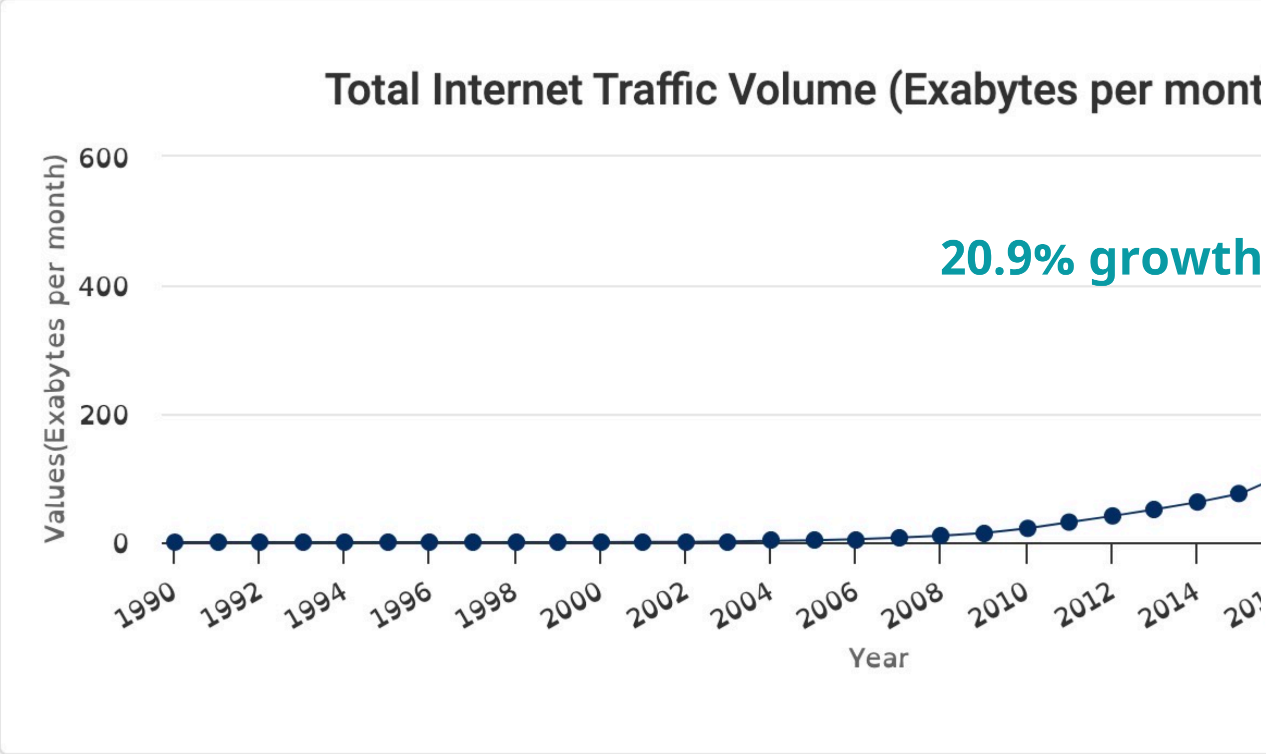
Faster and Stronger Lossless Compression with Optimized Autoregressive Framework

Yu Mao∗, Jingzong Li∗, Yufei Cui†, and Jason Chun Xue∗

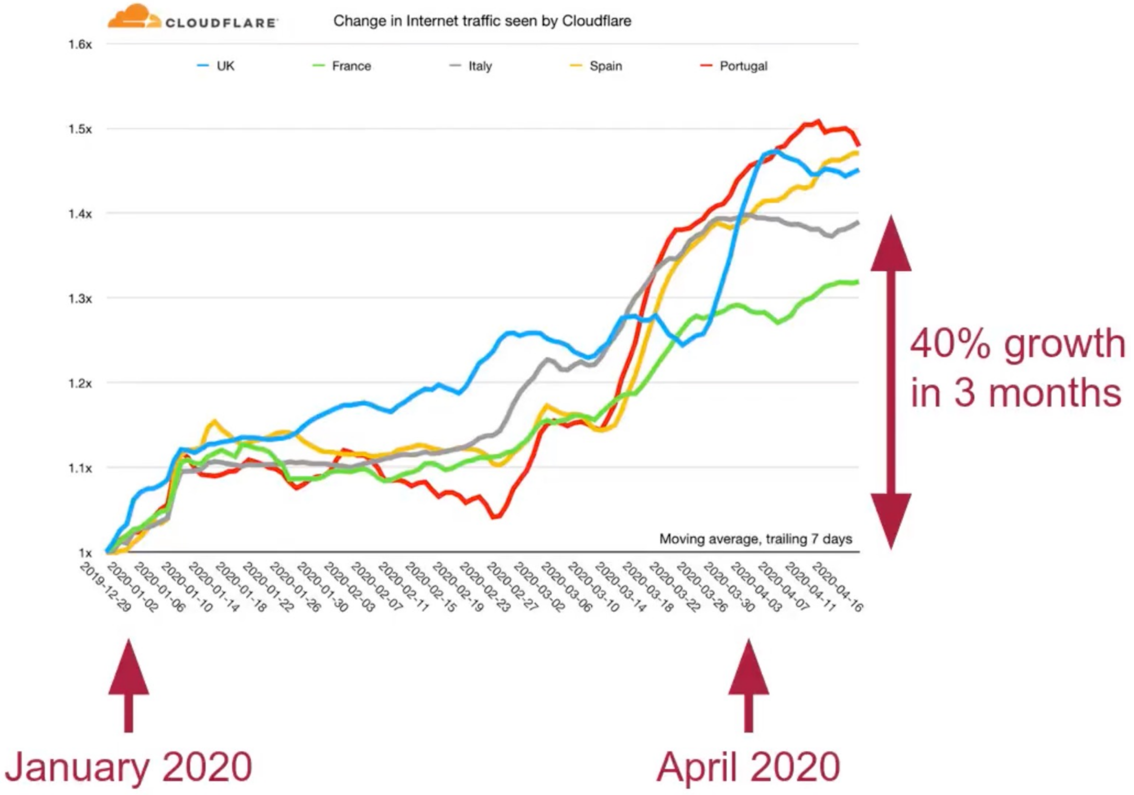
∗Department of Computer Science, City University of Hong Kong
†School of Computer Science, McGill University

Why Data Compression?

Projection by IBISWorld by 1990



Observation by Cloudflare in 2020



[https://blog.cloudflare.com/recent-trends-in-internet-traffic/]

We need a stronger compression algorithm to deal with the rapid growing trend.

Lossless compressors

– General-Purpose Lossless Compressors

Traditional	Deep-learning based
• Gzip, 7z, Zstandard • ...	• Cmix, NNCP, Dzip, TRACE • ...

Compression Ratio comparisons between traditional methods and deep-learning methods

Methods	Homogeneous Data					Heterogeneous Data	
	Enwik9	*Book*	*Sound*	*Image*	*Float*	*Silesia*	*Backup*
Gzip	3.09	2.77	1.37	1.14	1.06	3.10	1.28
7z	4.35	3.80	1.59	1.38	1.14	4.25	1.56
Zstd-19	4.24	3.73	1.40	1.16	1.10	3.97	1.36
Dzip	4.47	3.95	2.04	1.72	1.26	4.78	1.78
TRACE	5.29	4.58	2.16	1.81	1.28	4.63	1.78
OREO	5.68	4.94	2.25	1.86	1.28	4.86	1.87
PAC	5.97	5.05	2.25	1.96	1.29	4.99	1.92

Deep-learning based compressor can obtain **much higher performance** than traditional methods, but with **significantly slow compression speeds.**

To Compress 1GB Data:

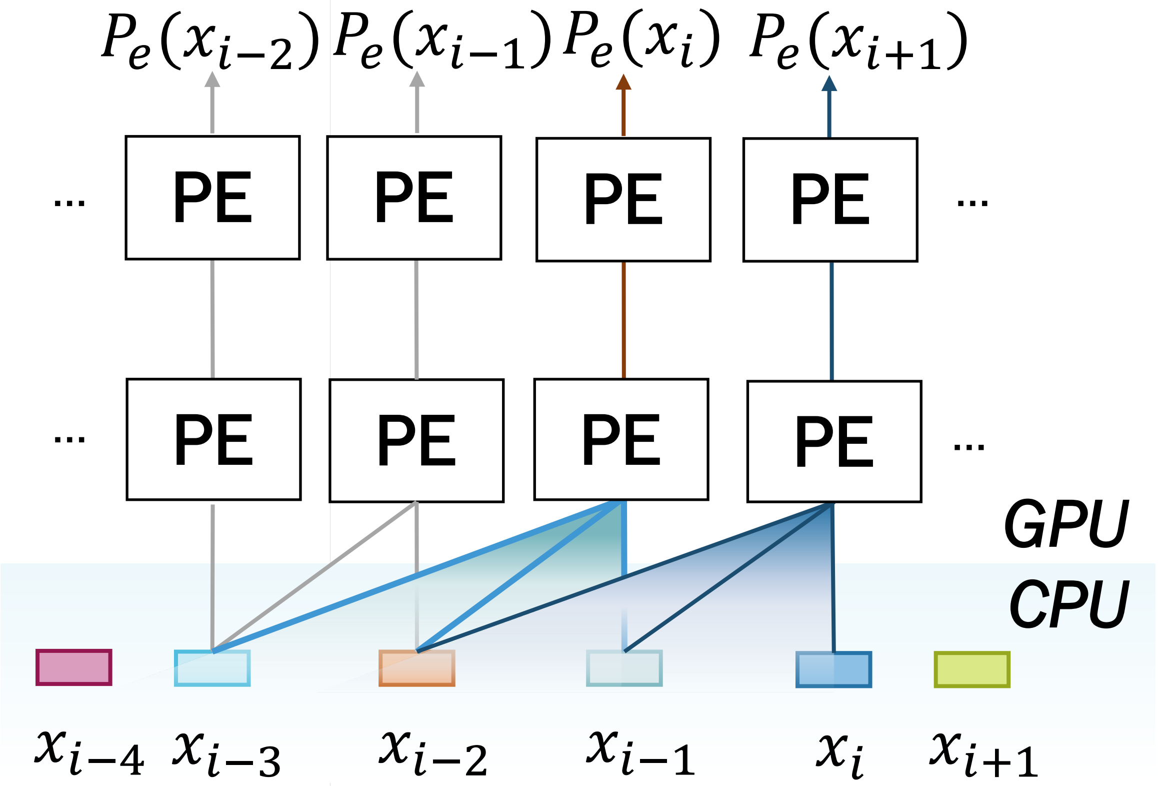
None Deep-learning compressor needs
Several tens of seconds

- Gzip→33s
- 7z→60s
- Zstd-19→321s
- Zstd-FPGA→1.28s
- LPAQ→85s

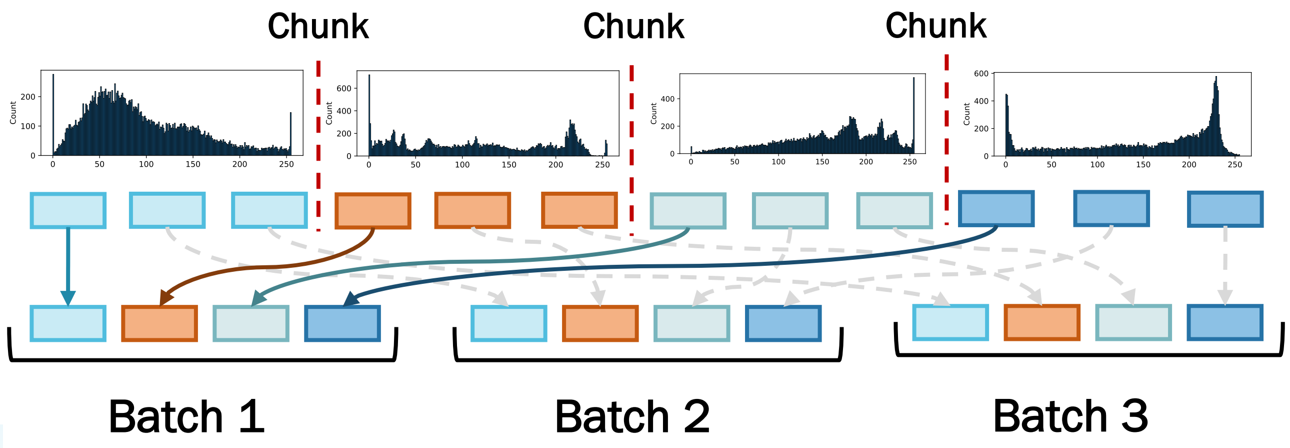
Deep-learning compressor needs
several days

- Cmix →25days
- Tensorflow-compress→8days
- NNCP→8days
- Dzip→1.7days
- TRACE→14h
- OREO→10h
- *PAC→5h (2080Ti)*

Two blind spot of current NN-based compressors



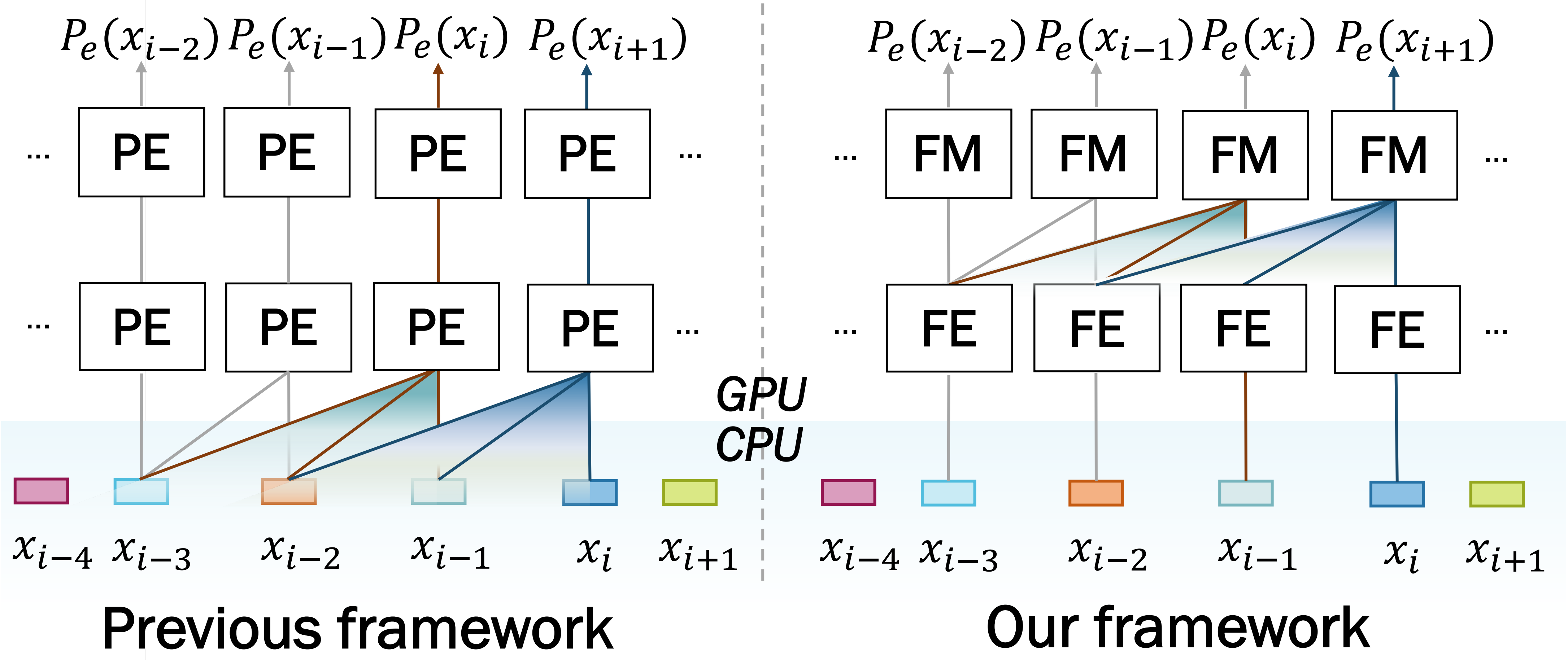
Duplicated processing problem



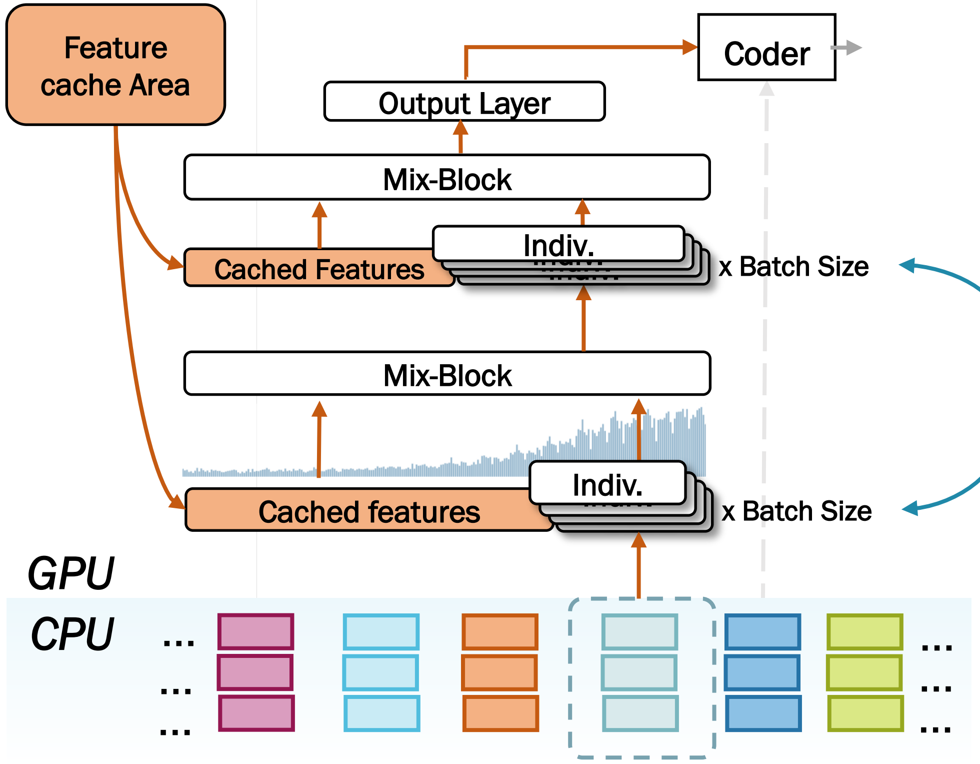
In-batch distribution variation problem

Duplicated processing problem

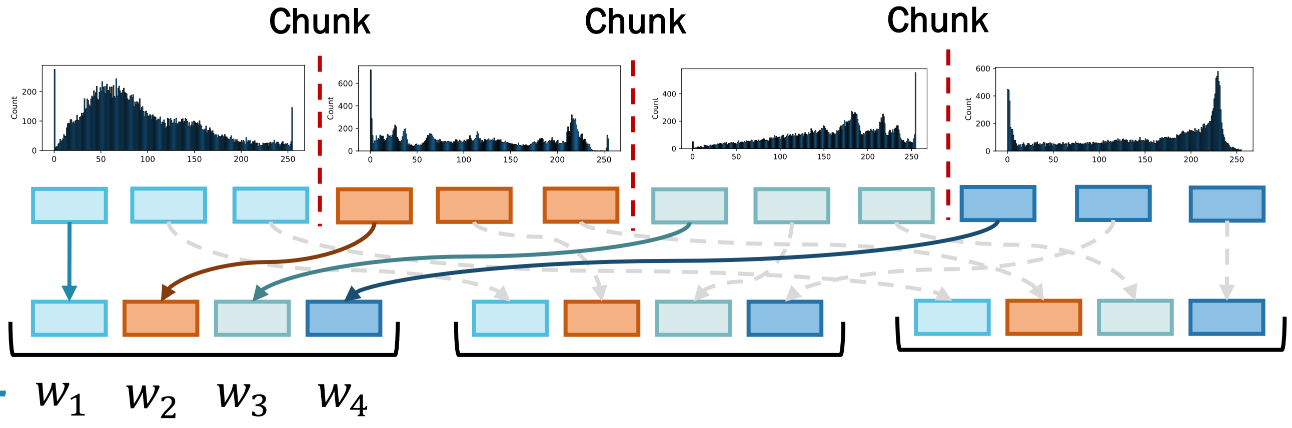
PE: Probability Estimation, **FE:** Feature Extraction, **FM:** Feature Mixture



In-batch distribution variation problem



Whole compression framework



Batch-aware model design: N set of parameter in individual layer corresponds to four item in a batch.

Learned Ordered Mask

A trainable 1D vector is introduced to dynamically learn the order information. The ordered importance is modeled as:

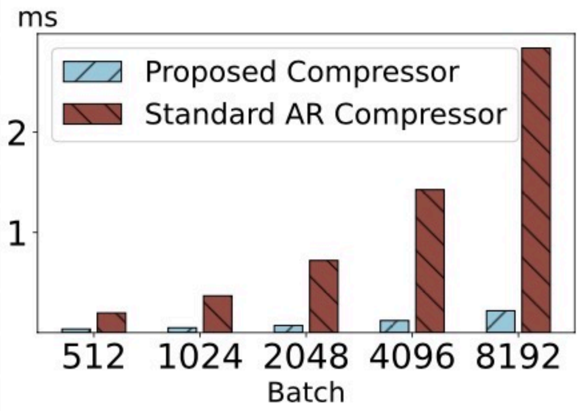
$$F(x_{i-1}),\ldots,F(x_{i-k}) = W * \{F(x_{i-1}),\ldots,F(x_{i-k})\}$$

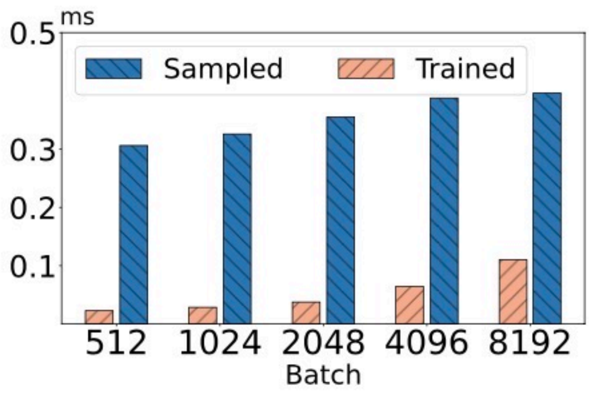
where $F(x_{i-1})$ is the extracted feature of x_{i-1} and W is learned ordered importance.

This makes PAC's probability estimator a pure MLP architecture, which gives possibility for current general-purpose compressor to implement on other hardware.

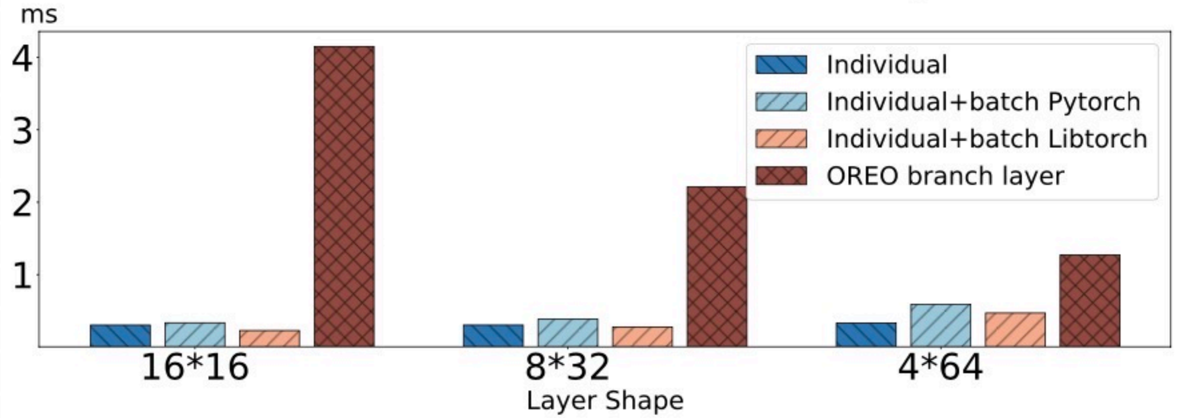
Performance Evaluation

Compressor	Peak GPU Memory Usage (GB)	Inference (ms)	FLOPs
NNCP	7.75	95.67	15.83×10^{10}
Dzip	6.39	5.82	7.48×10^{10}
TRACE	2.02	2.08	0.34×10^{10}
OREO	1.18	1.54	0.12×10^{10}
PAC	1.07	1.54	0.1×10^{10}





(a) Host-GPU data transmit time. (b) Ordered mask generation time.



(c) Individual layer inference time.

Future Direction

- Future directions in the field of compression could include:

- 1. Specialized Hardware Acceleration
- 2. Hybrid Compression Approaches
- 3. Fast Image Compression—We are working on it!

Our works

Mao Y, Li J, Cui Y, et al. Faster and Stronger Lossless Compression with Optimized Autoregressive Framework[C]//60th Design Automation Conference (DAC 2023): From Chips to Systems-Learn Today, Create Tomorrow. 2023.

Mao Y, Cui Y, Kuo T W, et al. TRACE: A Fast Transformer-based General-Purpose Lossless Compressor[C]//Proceedings of the ACM Web Conference 2022. 2022: 1829-1838.

Mao Y, Cui Y, Kuo T W, et al. Accelerating General-Purpose Lossless Compression via Simple and Scalable Parameterization[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 3205-3213.

Cui Y, Mao Y, Liu Z, et al. Variational Nested Dropout[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

WHERE **INNOVATION** BEGINS

DESIGN AUTOMATION CONFERENCE

Thank you!

JULY 9–13, 2023

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA