APEX Lab @ Duke

Duke Systems Group

Infini-AI Lab @ CMU

Duke DEPARTMENT of COMPUTER SCIENCE

Carnegie Mellon University
School of Computer Science

LLM.265: Video Codecs Are Secretly Tensor Codecs

Ceyu Xu*
eeentropy@ust.hk

Yongji Wu*
yongji.wu769@duke.edu

Xinyu Yang*
xinyuya2@andrew.cmu.edu

Beidi Chen
beidic@andrew.cmu.edu

Matthew Lentz
mlentz@cs.duke.edu

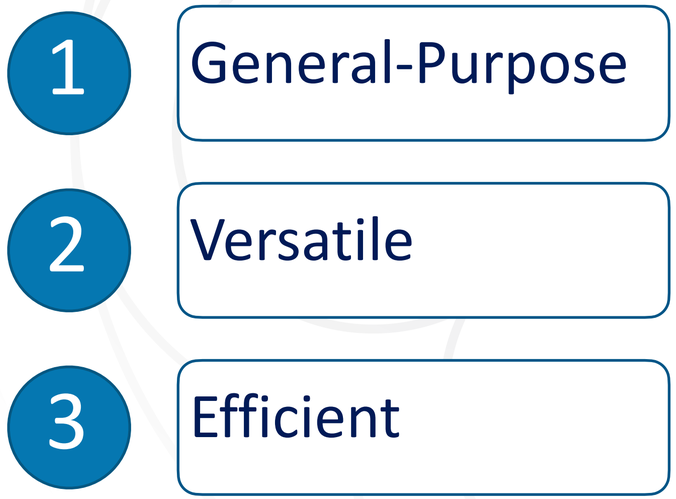
Danyang Zhuo
danyang@cs.duke.edu

Lisa Wu Wills
lisa@cs.duke.edu

Duke

Outline

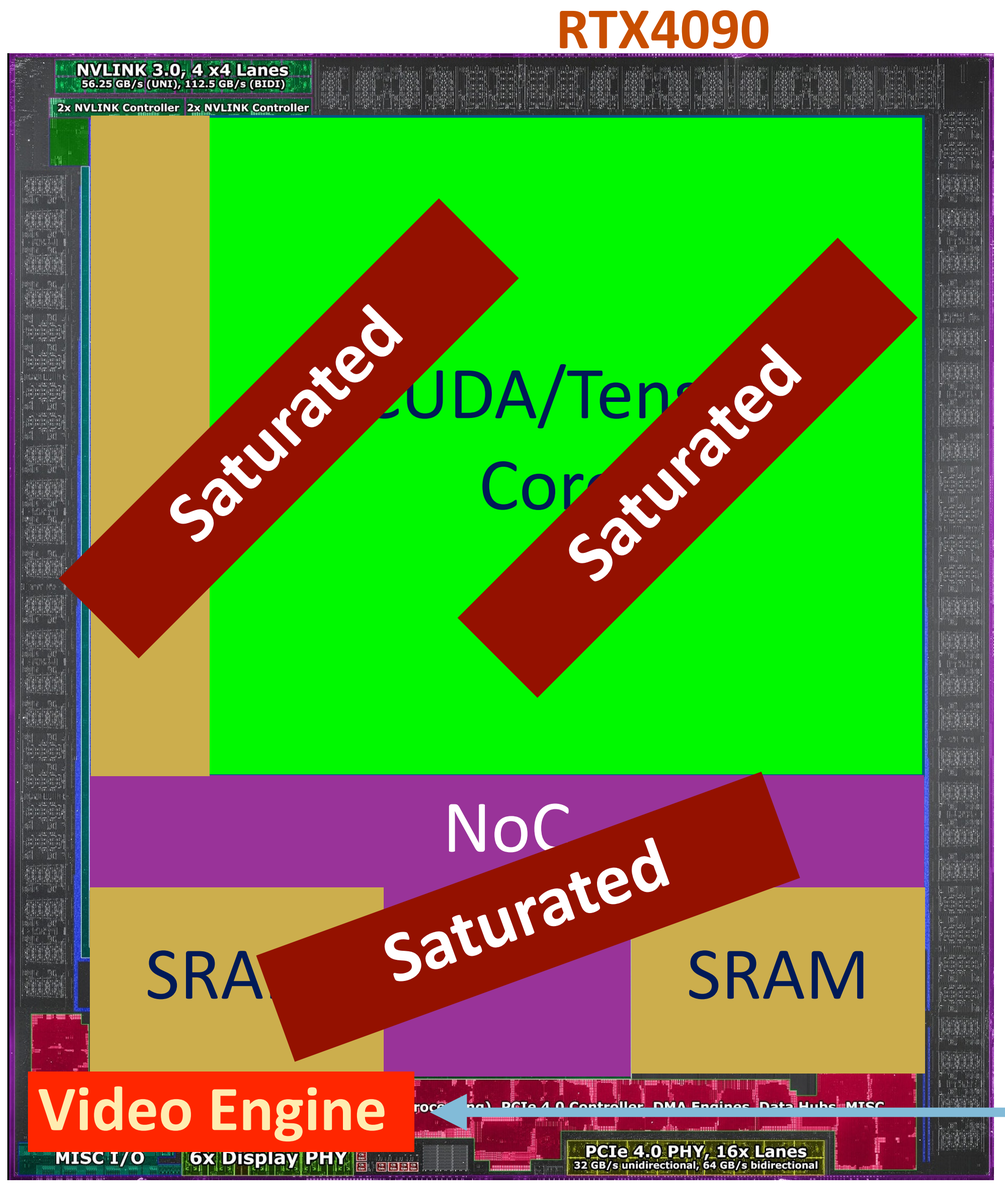
1. Motivation: Need for compression and the unused video engines on GPUs.
2. Why Video Codec Works for Tensors?
   1. Video Codec Basics
   2. Evidence why it works for tensors
3. How does it Compare with Other LLM Compression Algorithms?
   1. Weight, KV-Cache
   2. Training Gradient
   3. Non-LLM Models
4. Insights for Future Architecture Design.



Motivation

Old wine in a new bottle. Or a new bottle for old wine?

A Fully Used GPU… Wait?



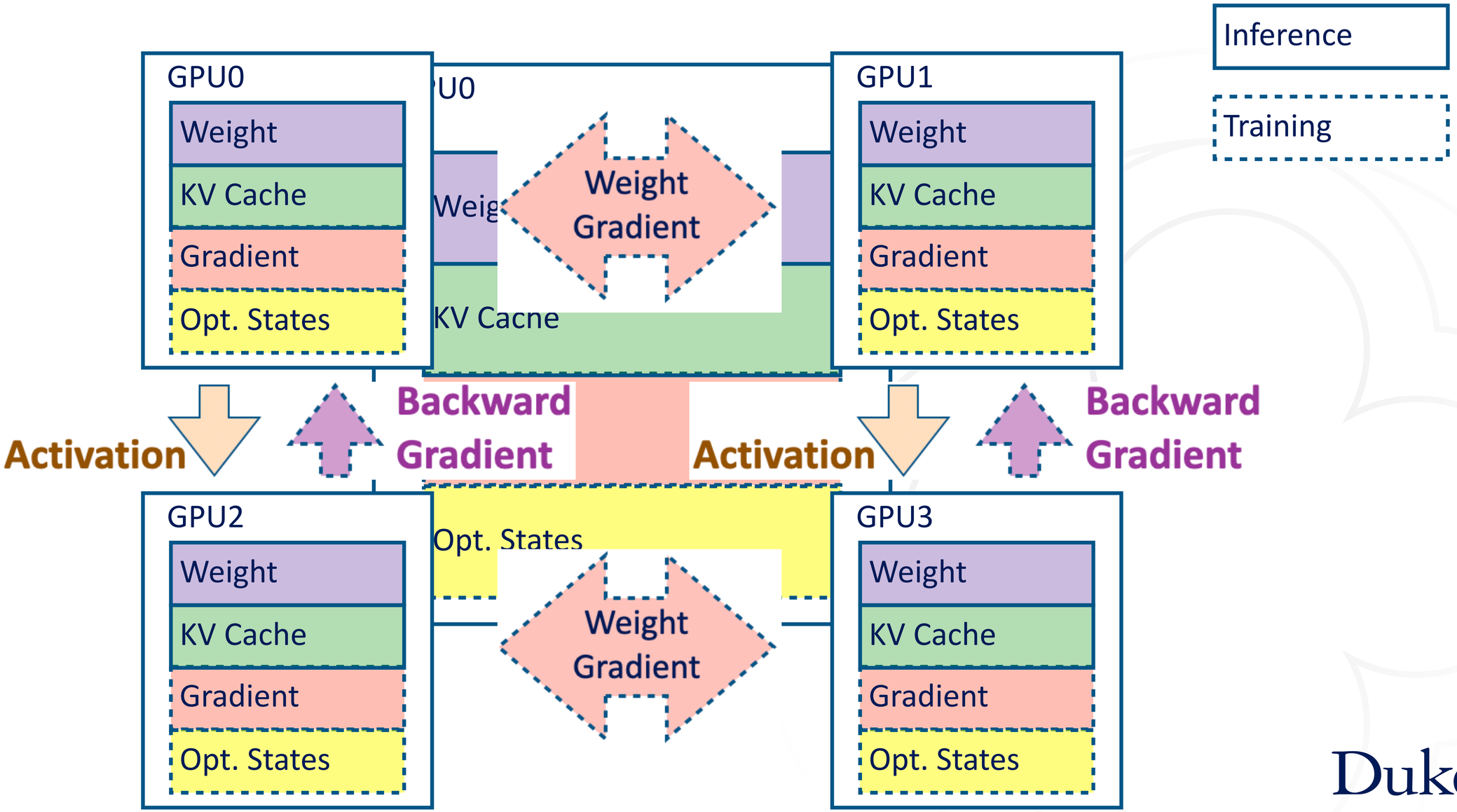
During LLM Inference/Training…

- Compute Cores Are been Saturated
- Memory and IO are All been saturated
- Are there anything on chip that is unused?

Can we use it for anything useful?

Duke

Various Types of Tensors in LLMs

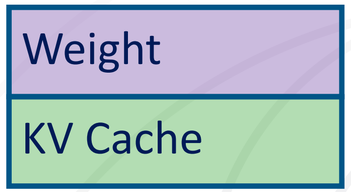


Compression Algorithms

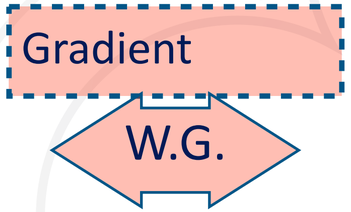
- Compression is essential for reducing the communication/storage overhead.
- Existing Compression/Quantization Algorithms have limited scope.
- **LLM.265** Provides one **Unified** Algorithm that works for all types of tensors.

GPT-Q [Frantar, ICLR2023],
AWQ [Lin, MLSys2024]:

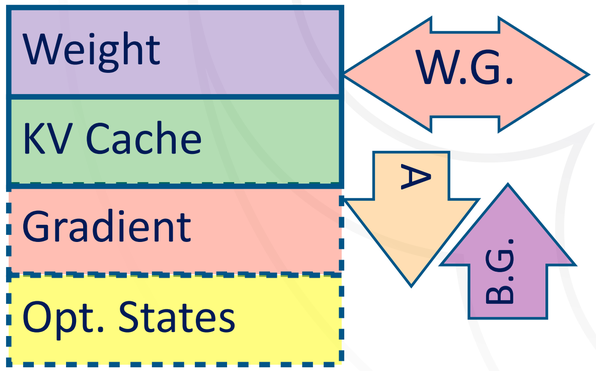
QuaRot [Ashkboos, NeurIPS 2024],
QServe [Lin, MLSys 2025]:



1b Adam[Tang, PMLR 2021],
1b Lamb[Li, HiPC 2021]:



LLM.265:

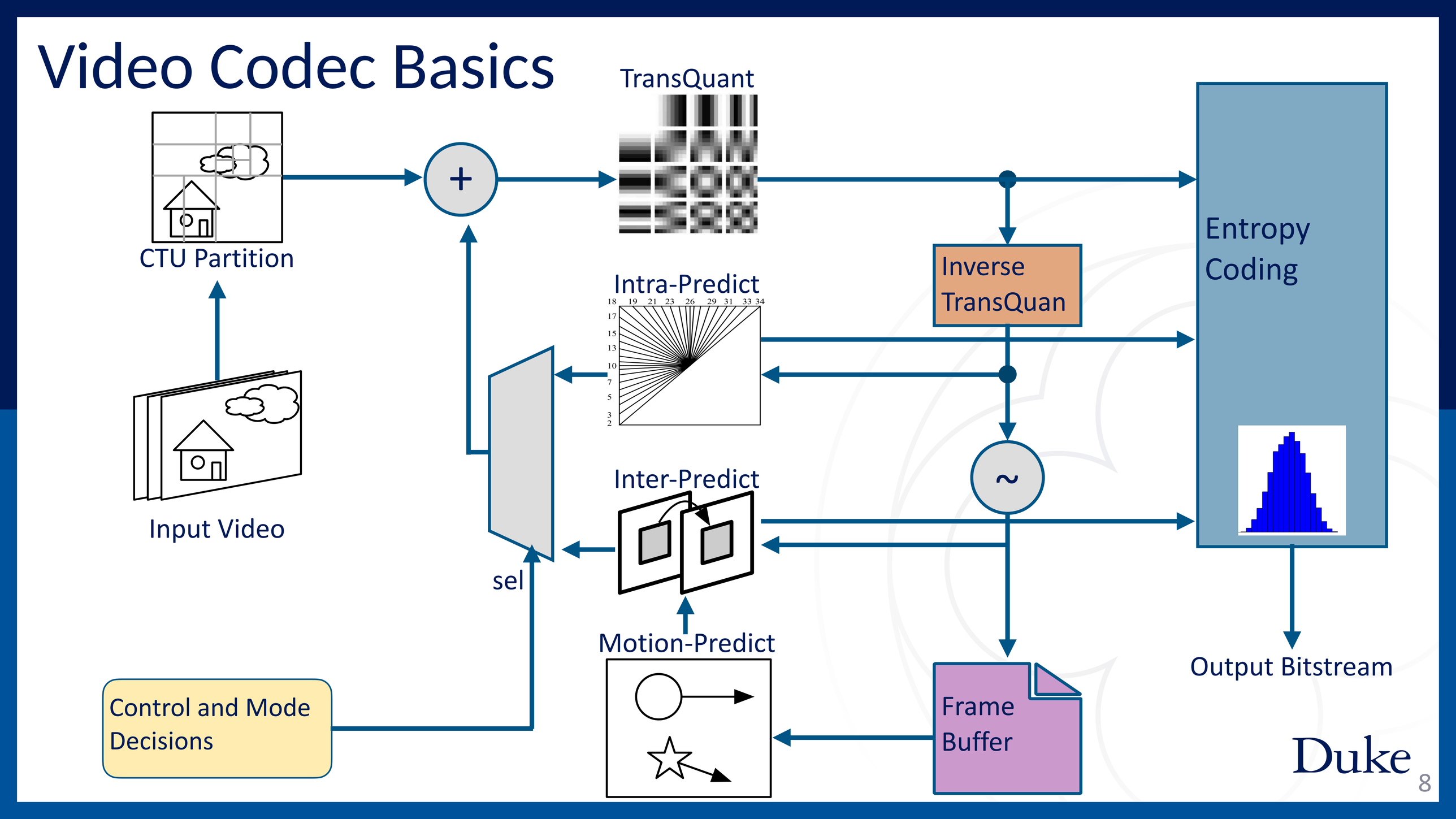


Duke

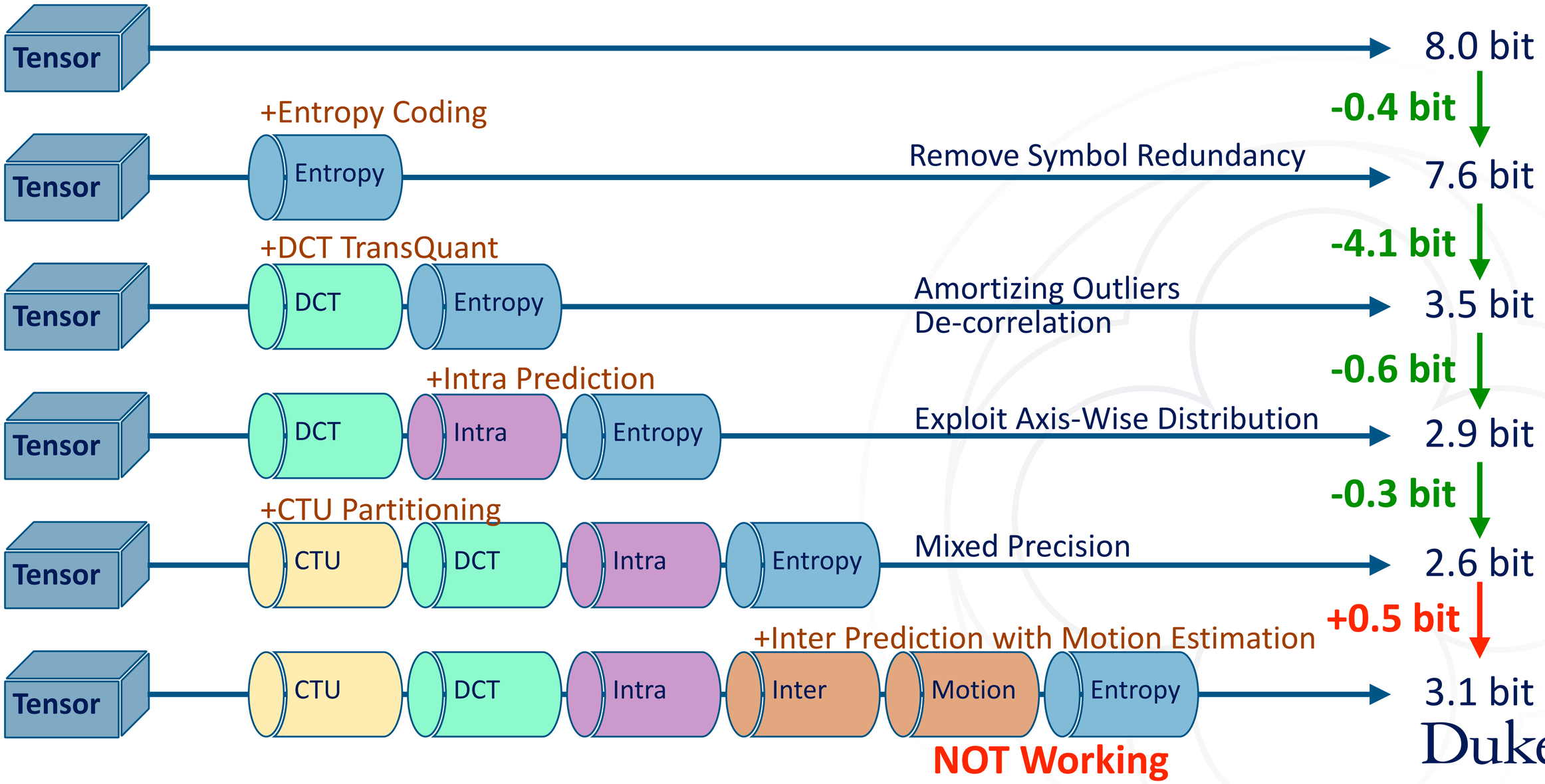
Why Video Codecs Work?

Don’t underestimate the old guard.

Video Codec Basics

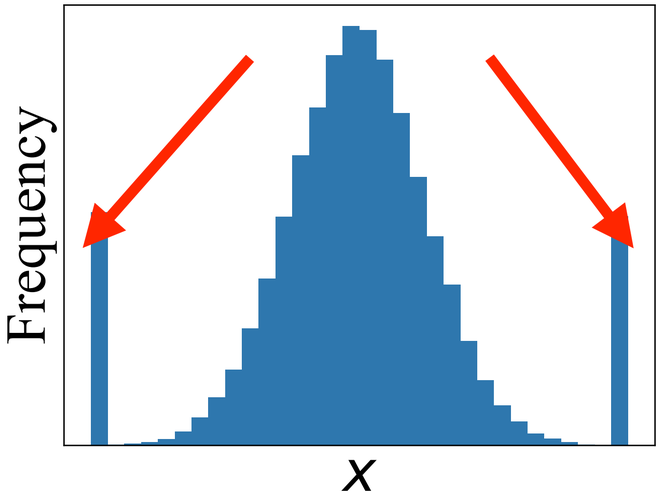


Does Video Codec Work?



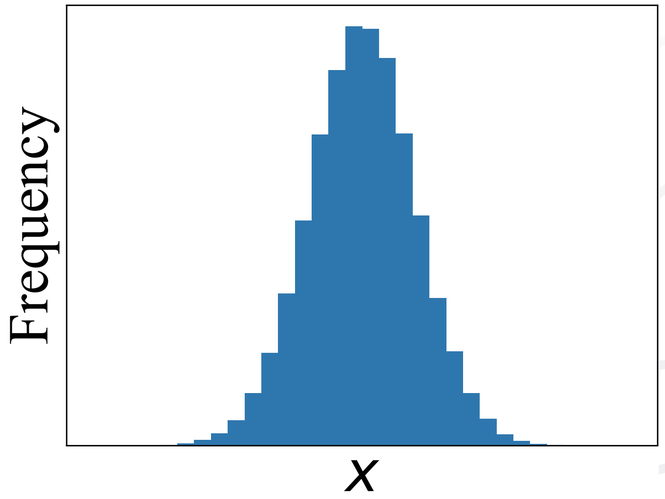
Example 1: Improve Outlier Coding Efficiency

Outliers



DCT Transform

Near Normal Distribution



-1	3	5	-8
2	0	4	1
3	-4	128	-3
2	1	5	3

Hard to Code

DCT Transform

12	-5	-13	17
-7	4	4	-8
-11	6	9	-15
13	-7	-15	20

Amortizing the encoding difficulties

Outliers → **Adjacent Elements**

Duke

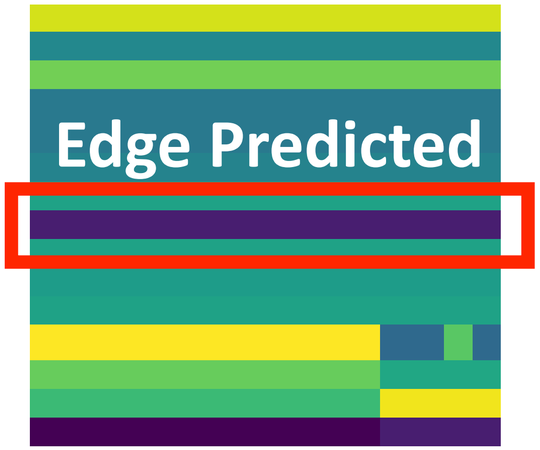
Example 2: Exploit Axis-Wise Pattern

Raw Weights

Channel-wise Distribution

-

Intra Prediction



Small prediction States to Replace dense values

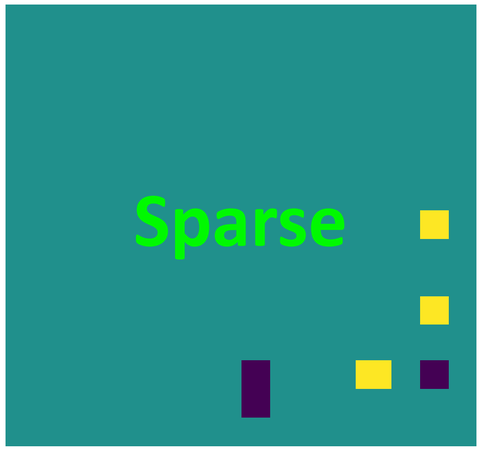
=

Residuals

Small Residuals

DCT

Coefficients



Sparse Coefficients

Duke

Compression for Inference

A Swiss Army Knife

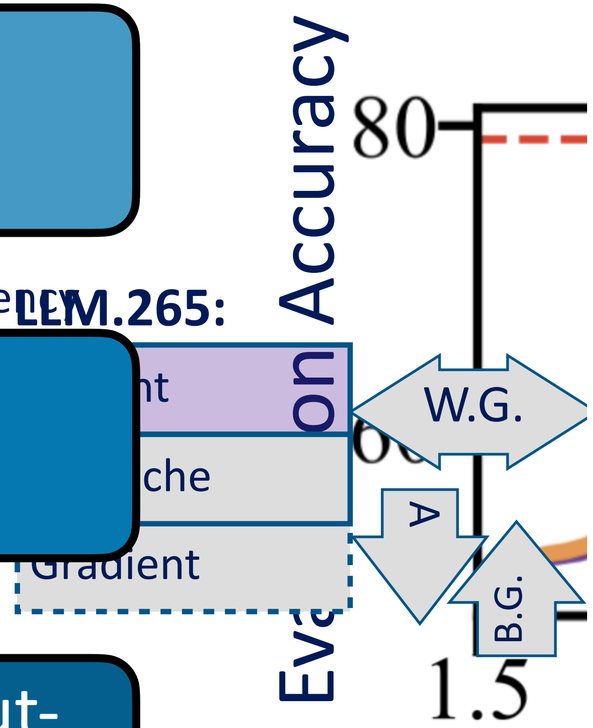
Weight-Only Compression

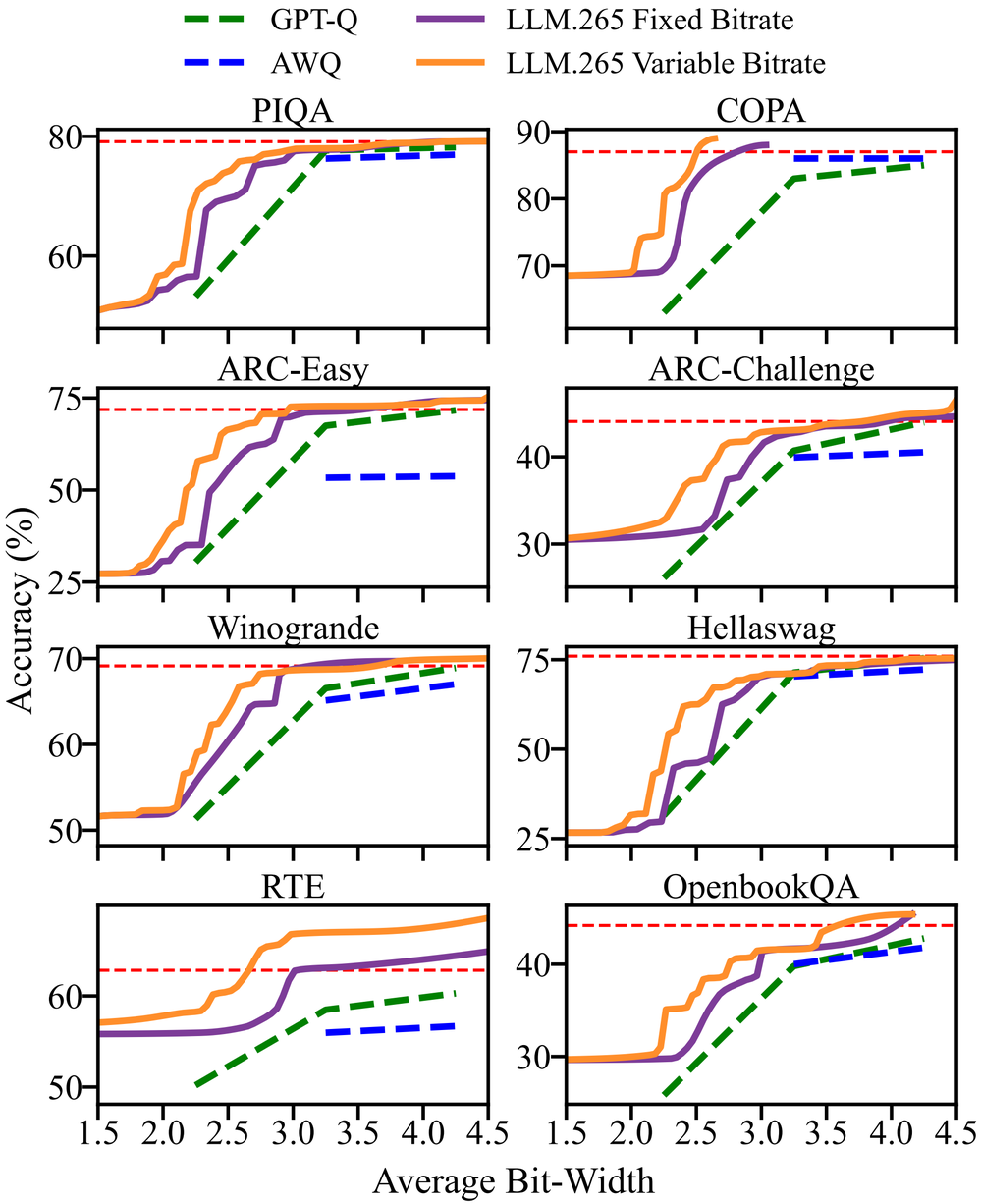
SOTA Information Efficiency

Integer → Fractional → Efficient

Versatile Bit-Rate

Calibration-Free, Input-Data-Agnostic



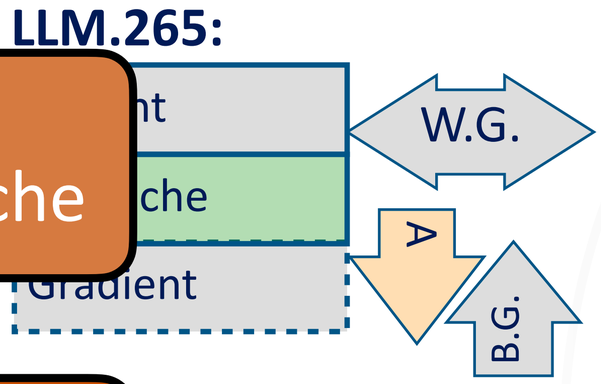


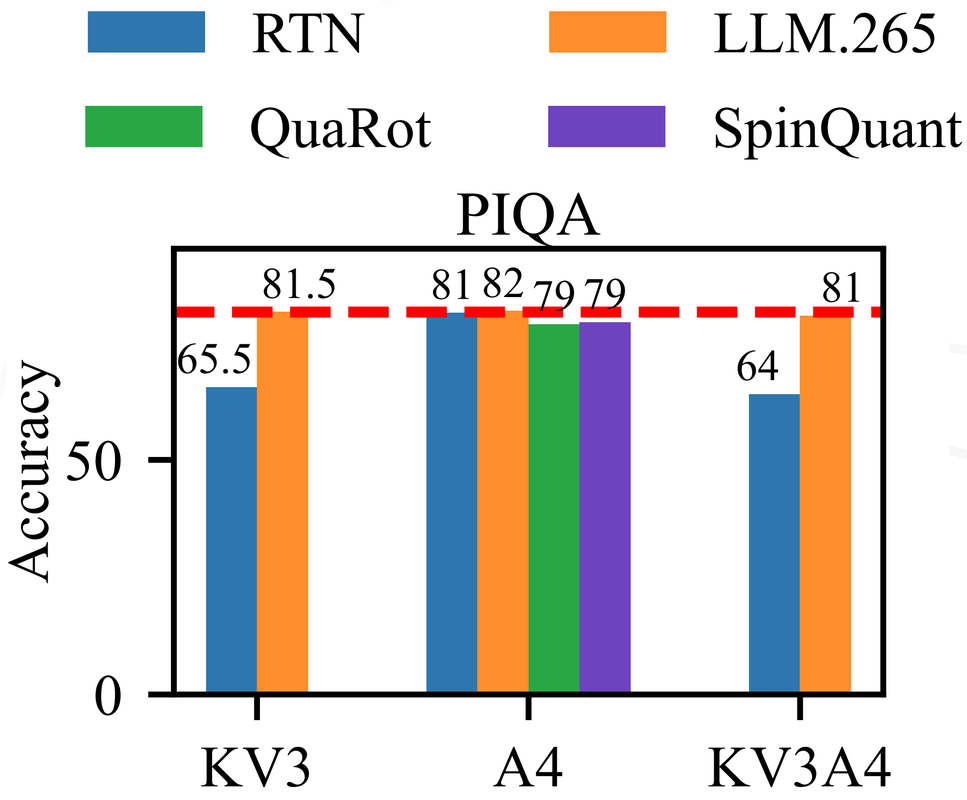
KV Cache and Activation Compression

SOTA Information Efficiency

Works for both Activation and KV-Cache

Hardware Accelerated Compression



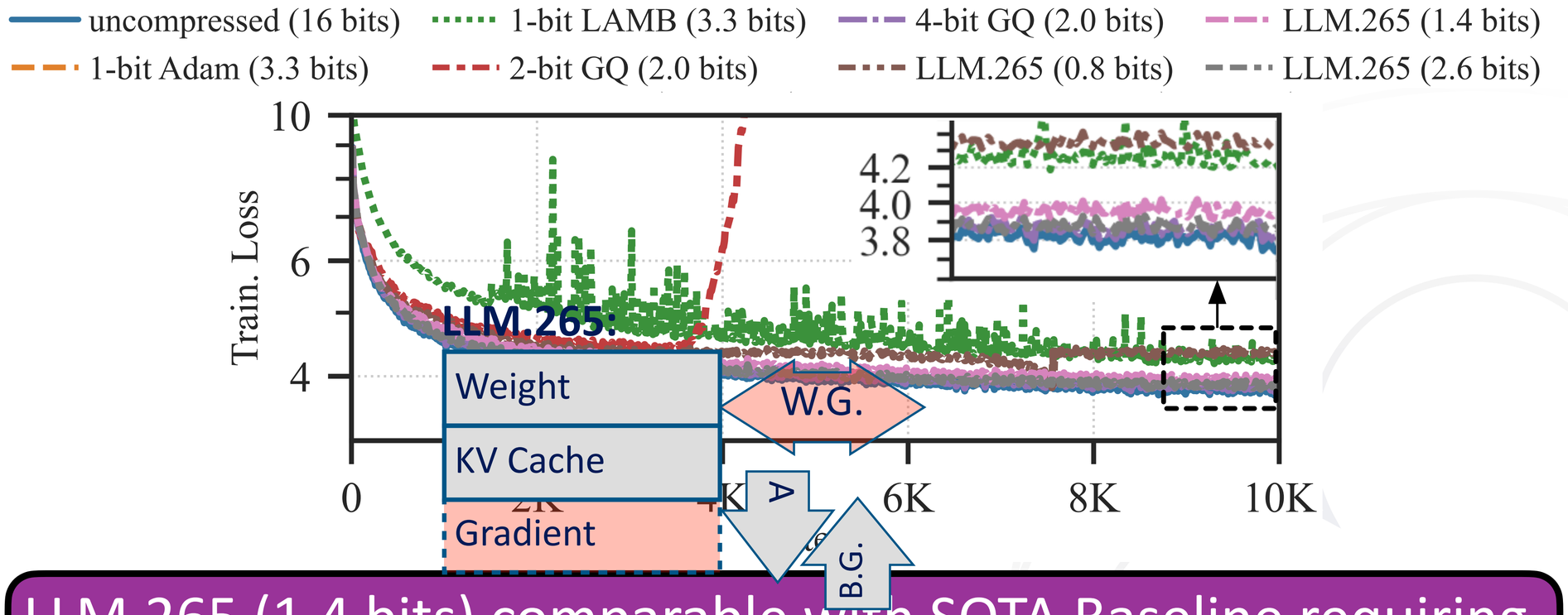


Duke

Compression for Training

To wear many hats

Data-Parallelism with Weight-Gradient Compression



LLM.265 (1.4 bits) comparable with SOTA Baseline requiring 3.3 bits. **(2.35x)**

LLM.265 (2.6 bits) comparable with Uncompressed Baseline requiring 16 bits. **(6.15x)**

Duke

Insights To Computer Architects

The more you compress the more you save.

Cost of Video Codecs

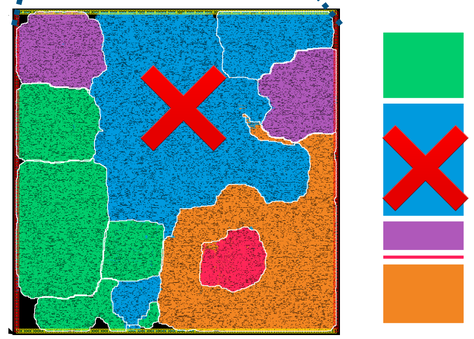
- Video Codecs are very small compared to other devices.
- Within Video Codecs. Inter-Frame Prediction consumes the majority of die area.
- Inter-Frame Prediction is useless for tensor. Could we do something better?

Nvidia GA-102
(RTX 3090)
628 mm^2

Mellanox CX5
100Gbps NIC
170 mm^2

H.264 Encoder (@100Gbps)
0.97mm^2

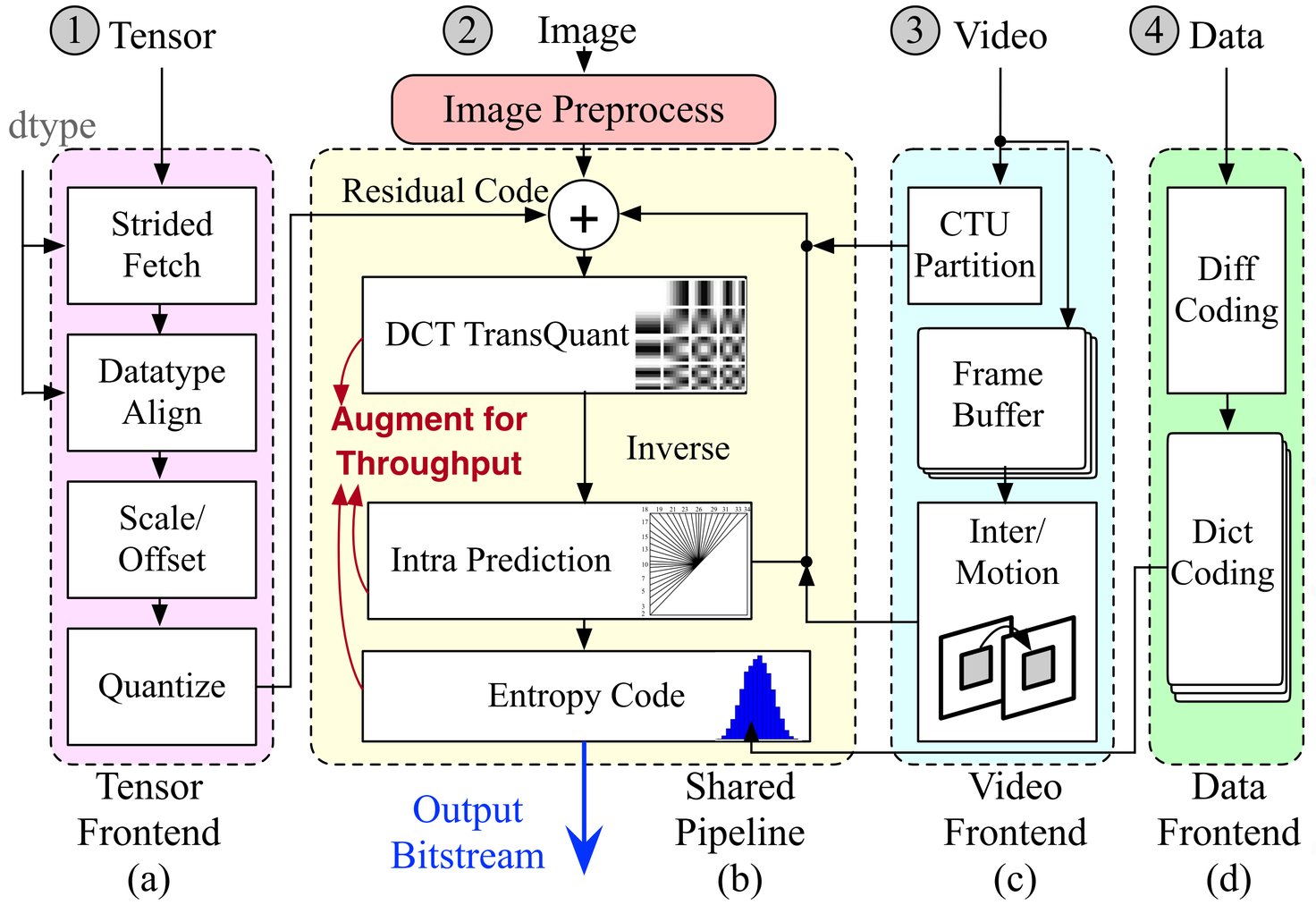
H.264 Encoder (@100Gbps)
0.96mm^2



Intra Prediction
Inter Prediction
MISC.
Entropy Coding
Buffer

Proposal: Trinity Codec

- Tensor, Image/Video, Data shares common compression pipelines.
- Video compression only needs 8K120fps at most.
- We can augment the tensor-required pipeline for better throughput.

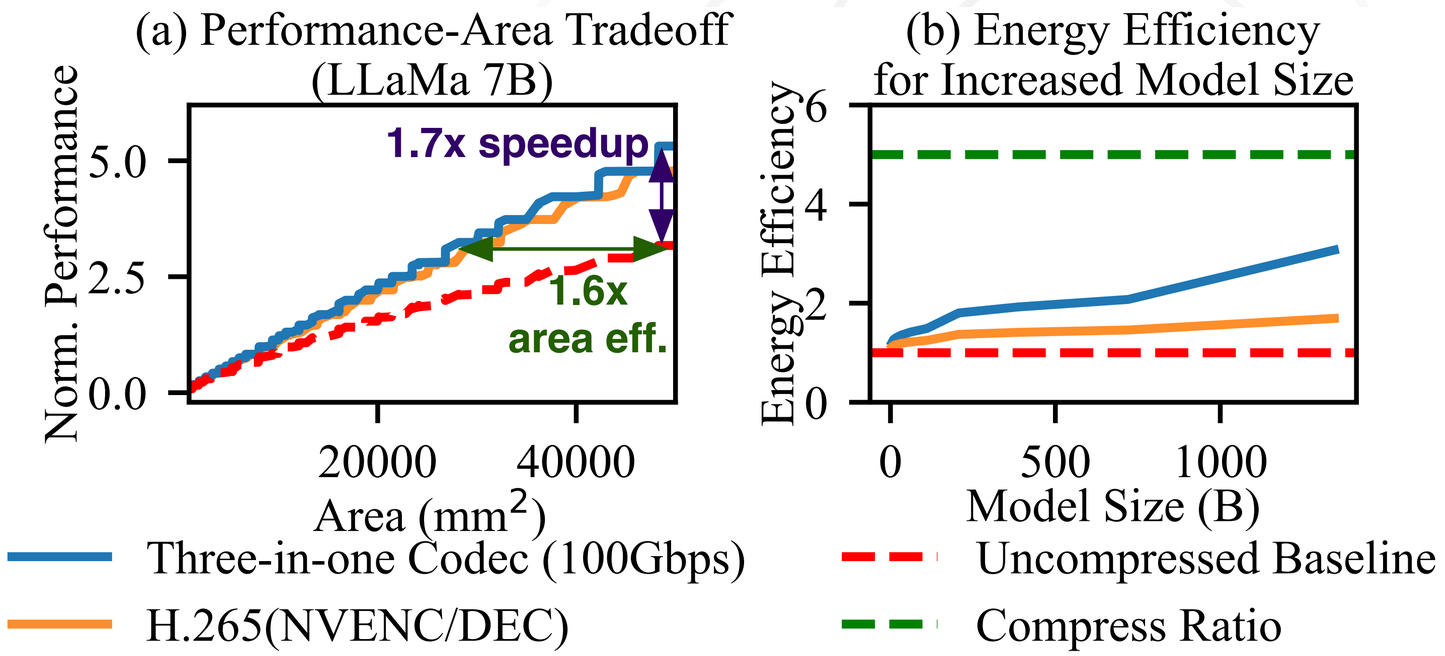


Benefits of Video Codecs

- When models getting larger and larger, compression is becoming more and more important.

- The codecs been built, the more die area and more energy you saved.

The larger the model size
The larger the datacenter scale



The larger the gain of compression.

Duke

Thank you!

References

- [1] Frantar, Elias, et al. "Gptq: Accurate post-training quantization for generative pre-trained transformers.”
- [2] Lin, Ji, et al. "Awq: Activation-aware weight quantization for on-device llm compression and acceleration." Proceedings of machine learning and systems
- [3] Ashkboos, Saleh, et al. "Quarot: Outlier-free 4-bit inference in rotated llms." Advances in Neural Information Processing Systems
- [4] Lin, Yujun, et al. "Qserve: W4a8kv4 quantization and system co-design for efficient llm serving.”
- [5] Tang, Hanlin, et al. "1-bit adam: Communication efficient large-scale training with adam’s convergence speed." International Conference on Machine Learning.
- [6] Li, Conglong, et al. "1-bit lamb: Communication efficient large-scale large-batch training with lamb’s convergence speed." 2022 IEEE 29th International Conference on High Performance Computing, Data, and Analytics (HiPC)
- [7] Liu, Zechun, et al. "Spinquant: Llm quantization with learned rotations." arXiv

Compatibility Matrix of Codecs vs. GPU Gen.

Encoding

	H.264	H.265	AV1	JPEG
Volta (RTX 20xx)	Yes	Limited	No	No
Ampere (RTX 30xx)	Yes	Yes	No	No
Ada (RTX 40xx)	Yes	Yes	Yes	No
V100	Yes	Yes	No	No
A100	No	No	No	No
H100	No	No	No	No

Decoding

	H.264	H.265	AV1	JPEG
Volta (RTX 20xx)	Yes	Yes	No	No
Ampere (RTX 30xx)	Yes	Yes	Yes	No
Ada (RTX 40xx)	Yes	Yes	Yes	No
V100	Yes	Yes	No	No
A100	Yes	Yes	No	Yes
H100	Yes	Yes	Yes	Yes

Duke

More Compression vs. More Compute

Turn compute into effective bandwidth

Energy efficient since it reduces the footprint directly

Better Scalability compared to Mem/IO/Compute

Hard to Saturate under bandwidth bottlenecks.

Energy Inefficient

Unable to scale due to communication overhead

Treating Memory/IO bottlenecks requires data-centric solutions.

Duke