From Error-Bounded to Fixed-Ratio: Efficient Sampling-Guided GPU Lossy Compression

Bohan Zhang[1] Yafan Huang[2] Sheng Di[3] Guanpeng Li[1]

[1] Department of Electrical and Computer Engineering, University of Florida

[2] Computer Science Department, University of Iowa

[3] Mathematics and Computer Science Division, Argonne National Laboratory

SC26 · November 15–20, 2026 · Chicago, Illinois

UF UNIVERSITY of FLORIDA IOWA Argonne NATIONAL LABORATORY

The big data issue



Frontier, Oak Ridge National Laboratory · ORNL, CC BY 2.0



Aurora, Argonne National Laboratory · Argonne NL, public domain

> 100 GB/s · ≈ 8.6 PB in a day

plasma physics diagnostics — must be **stored, transferred, and analyzed** under **limited storage capacity** and **restricted I/O bandwidth**

- **Lossy compression** has emerged as a key technique for data reduction
  - High compression ratios while preserving user-controllable data fidelity
- **GPU lossy compression** is a practical and increasingly necessary solution
  - Data is generated and processed directly on the device
  - Massive parallelism delivers high throughput

Why fixed-ratio compression

- An **error-bounded** mode cannot guarantee an output size
  - Most GPU lossy compressors are error-bounded: the compression ratio cannot be predetermined
  - Hitting a target ratio requires repeated **trial-and-error** re-compression, which significantly degrades performance
- In many practical scenarios, the binding constraint is not distortion but a **budget**
  - A fixed storage, communication, or memory budget
- **Fixed-ratio compression** enforces a user-specified budget
  - It provides a hard upper bound on output size
  - This enables buffer pre-allocation, stable I/O throughput, and predictable memory usage
  - In LLM inference, fixed-ratio KV cache compression directly controls GPU memory consumption

Existing solutions fall short

Error-bounded compressors

- **cuSZp, PFPL, MGARD-GPU**
  - Bound the maximum pointwise reconstruction error
  - Compressed size cannot be specified in advance

Fixed-ratio: cuZFP

- The **only** GPU compressor with native fixed-ratio control
  - Reduces fidelity, hurting post-hoc analysis
  - Struggles with irregular patterns in scientific data
  - Multi-pass transform: substantially lower throughput

CPU ratio-control frameworks

- **FRaZ, FXRZ, CAROL**
  - Iterative search or learned bound prediction
  - Host-driven control loops, incompatible with a GPU-resident in-situ pipeline

Challenges in GPU fixed-ratio compression

UF UNIVERSITY of FLORIDA

C-1 Variations in data pattern

Scientific datasets exhibit highly diverse structures — even within one field, different locations vary substantially.

C-2 High-speed requirements

Estimating the achievable ratio by repeated compression trials significantly degrades performance.

C-3 Strict parallel constraints

GPUs are used to accelerate throughput, yet any reliance on CPU-side operations incurs costly data movement.

MILIO: High-level Overview

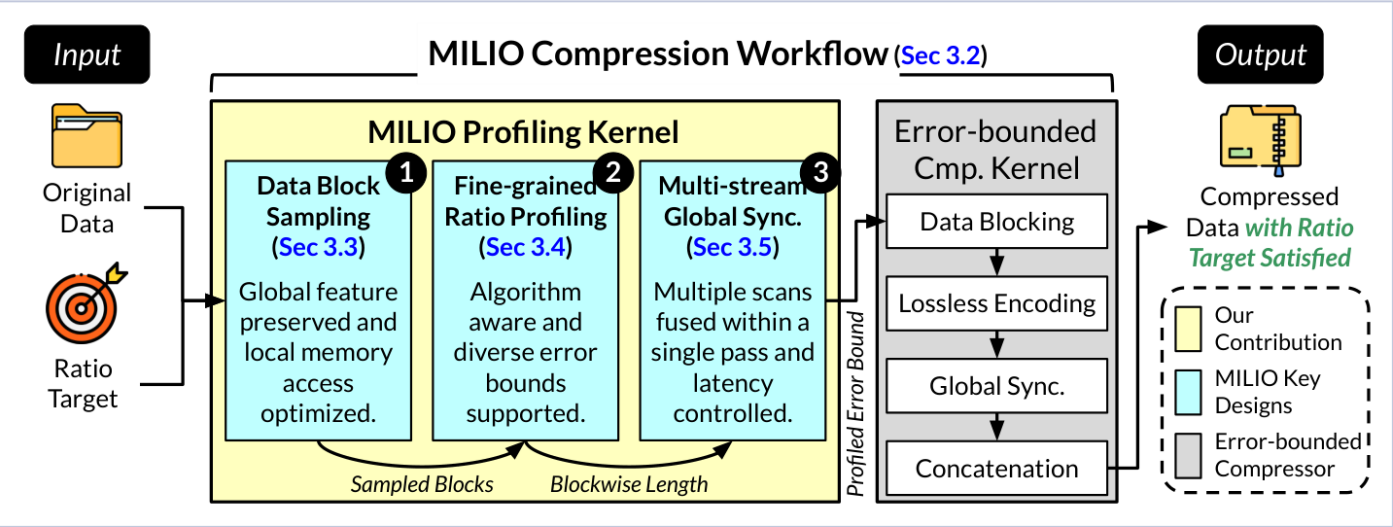


Fig. 1: Illustration of MILIO Compression Workflow

- Given a dataset and a target ratio, the **Profiling Kernel** generates an error bound; the **Compression Kernel** applies it.
- The compression step uses a publicly available single-kernel GPU compressor.

How MILIO responds

- **Data Block Sampling** — a small fraction of blocks is sampled evenly to capture global data patterns. (C-1)
- **Fine-grained Ratio Profiling** — each sampled block is profiled under candidate bounds, without full compression. (C-2)
- **Multi-stream Global Synchronization** — ratios across sampled blocks are fused in one pass. (C-3)

How to capture global data characteristics?

- HPC datasets, routinely GB-scale, exhibit substantial spatial heterogeneity, with data patterns varying significantly across different regions of a field.
- Regions (a) and (b) are both smooth, yet the same error bound can flush many values in (b) toward zero while yielding limited compressibility in (a).
- Region (c) contains sharp local variations and strong discontinuities.

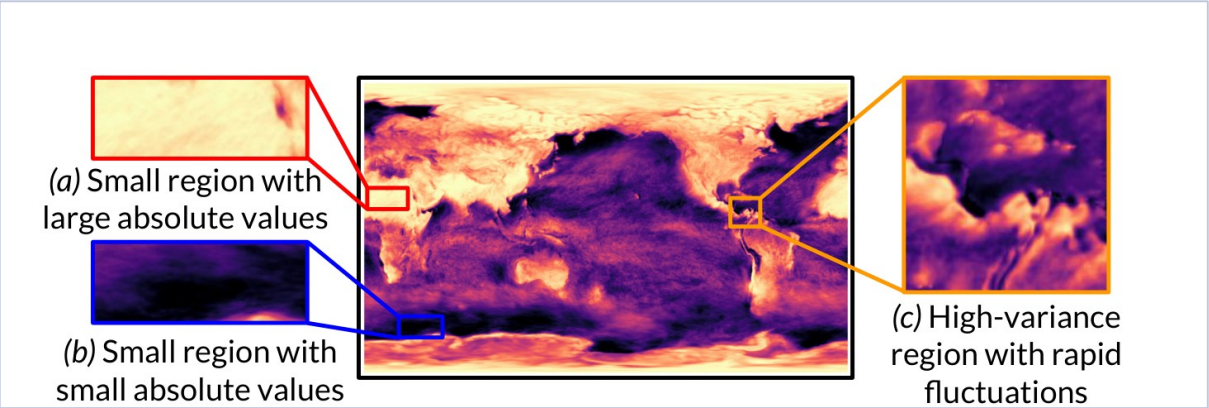


Fig. 2: Global data characteristics in a CESM climate field

Dimension-aware Uniform Sampling

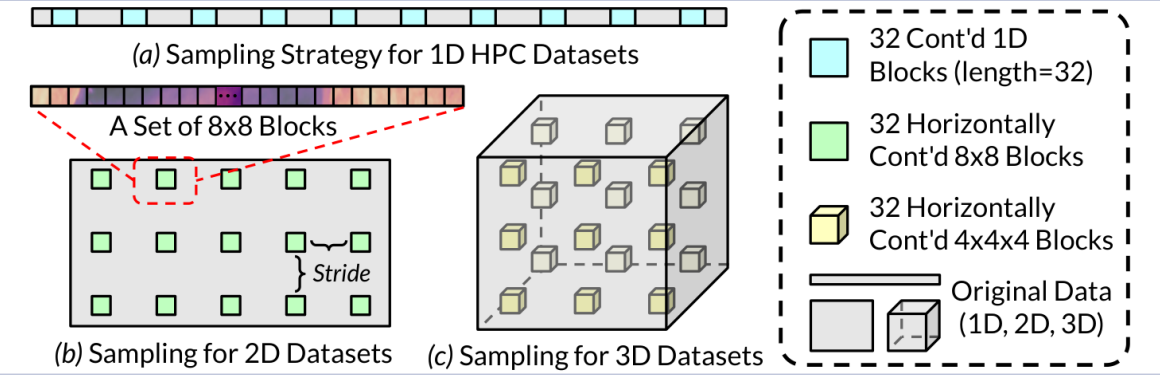


Fig. 3: MILIO sampling strategy for HPC datasets

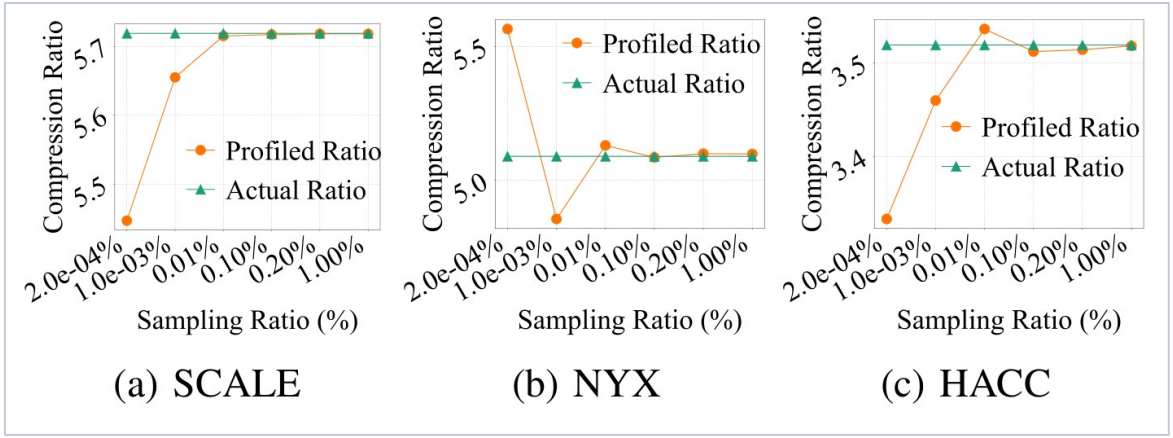


Fig. 4: Effect of the sampled fraction ρ on accuracy

- Each sampled portion consists of 32 horizontally contiguous data blocks — 32 for 1D, 8×8 for 2D, 4×4×4 for 3D — forming a chunk.
- Profiled ratios closely match actual ratios once ρ **≥ 0.1 %** — SCALE at 1E-2: 5.718 actual vs. 5.717 profiled.

Fine-grained Ratio Profiling

Algorithm 1: Fine-grained Ratio Profiling

Input: A CUDA warp with 32 threads $\{T_0, T_2, ..., T_{31}\}$, a sampled chunk that contains 32 (horizontally) continued data blocks $\{B_0, B_2, ..., B_{31}\}$, and 128 candidate error bounds $EB = \{eb_0, eb_2, ..., eb_{127}\}$.

Output: Compressed size for candidate error bounds $C = \{c_0, c_2, ..., c_{127}\}$.

1 Initialize all compressed sizes in C as 0;
2 **for each** T_i, *where* $i \in \{1, 2, ..., 32\}$ *// Executed in parallel.*
3 Read B_i from global memory; *// Shared by four candidates.*
4 Quantize B_i with eb_{4i} and delta encoding;
5 Identify the greatest quantized integer g_{4i} in B_i;
6 Calculate size c_{4i} based on g_{4i};
7 Repeat Line 4 to Line 6 for eb_{4i+1}, eb_{4i+2}, and eb_{4i+3}.

Dataset	CESM		HACC		EXAFEL		QMCPACK		RTM		SYNTH	
	min	max	min	max	min	max	min	max	min	max	min	max
MILIO-p	3.5	73.9	2.0	70.3	2.1	224.3	2.1	83.8	3.0	103.0	3.1	54.4
MILIO-o	3.6	75.9	2.1	70.6	2.1	224.3	2.1	83.8	3.1	103.0	3.2	58.5

TABLE I: Range of profiled ratios across 128 error bounds

- We sweep error bounds from 1E-1 down to 1E-6, yielding 128 candidate settings in total.
- On NVIDIA GPUs 32 consecutive threads form a warp; 128 candidates fully utilize four warps.
- Achievable ratios span from about **2×** to **224.3×** across datasets.

How to profile blockwise ratio?

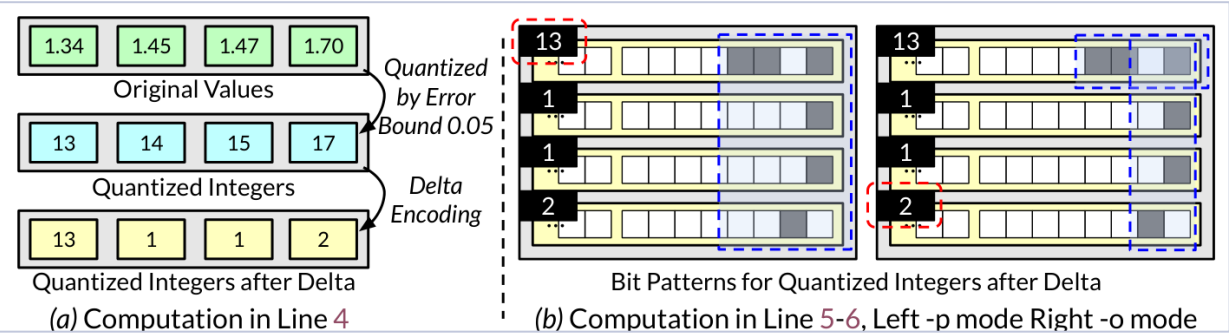


Fig. 5: Ratio profiling computations (Algorithm 1, lines 4–6)

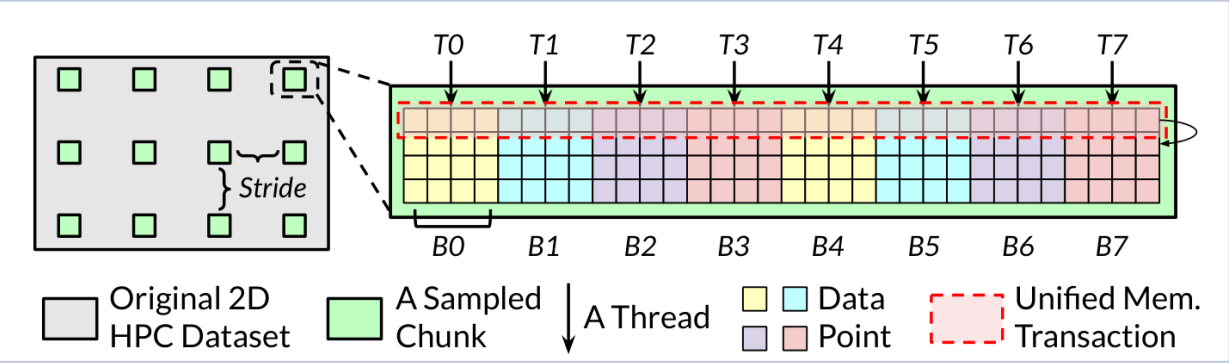


Fig. 6: Memory access optimization for 2D datasets

- Values are quantized by the error bound and delta encoded; one scan finds the greatest quantized integer.
- Its effective bitwidth, from the CUDA intrinsic _clz(), sets the block size: N × (L + 1) / 8 bytes.
- Consecutive threads in a warp access horizontally adjacent data within a row, enabling coalesced transactions.
- The plain and outlier fixed-length encodings are the “-p” and “-o” modes.

Multi-stream Global Synchronization

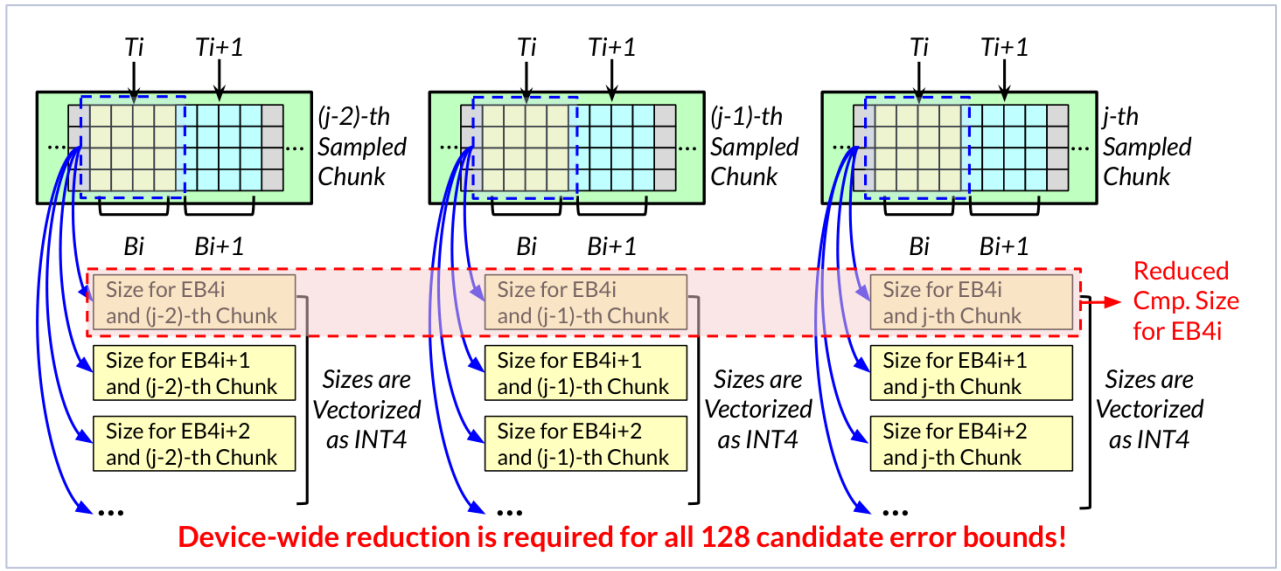


Fig. 7: Why multiple device-wide reductions are required

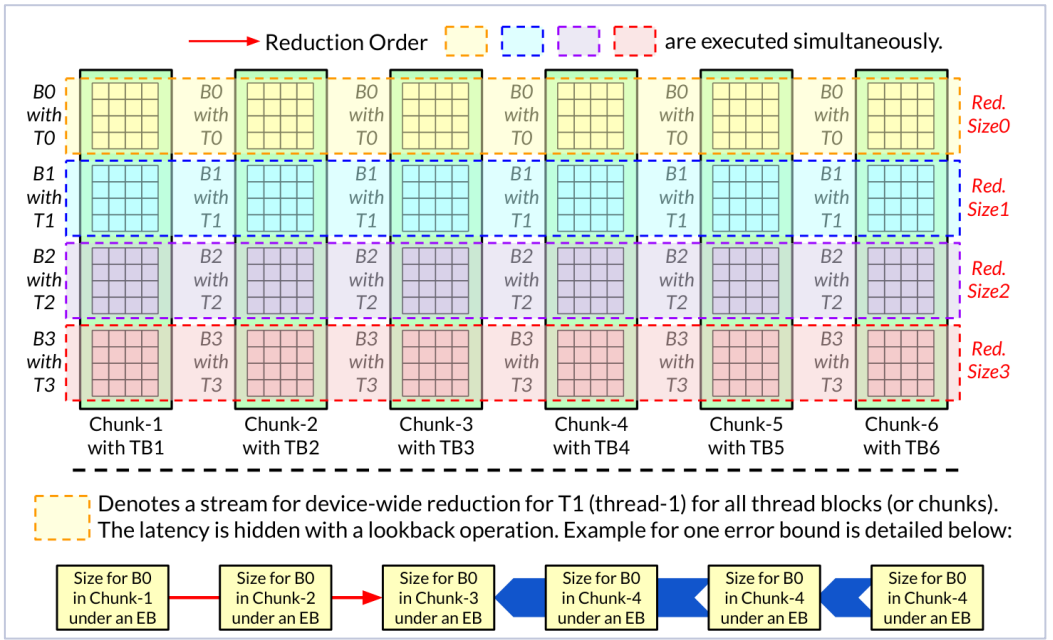


Fig. 8: Multi-stream global synchronization

- Every candidate bound needs a device-wide reduction; even with four bounds vectorized per block, 32 reductions remain.
- Threads with the same index across all thread blocks run independent serial reductions, overlapping the 32 latencies.

Implementation and Measurement

```
1  void Milio_comp(oridata, compdata, target_r){
2      // Initializing intermediate variables.
3      cudaMalloc(...);
4      cudaMemset(...);
5      // Step 1: Profiling Phase
6      Profiling_kernel<<<...>>>(...);
7      cudaMemcpy(...);
8      // Simple loop for error-bound search
9      float selected_eb = pick_best_eb(...);
10     // Step 2: Compression Phase
11     Compression_kernel<<<...>>>(...);
12     // Releasing intermediate variables.
13     cudaFree(...) ;
14 }
```

```
Milio_comp:     Compression API
oridata:        Original data on GPU
compdata:       Compressed data on GPU
target_r:       User-specified target ratio
selected_eb:    Profiled error bound
```

```
// Previous Science/AI workloads
...
timer.Start()
Milio_comp(oridata, compdata, target_r);
timer.End()
...
// Following Science/AI workloads
```

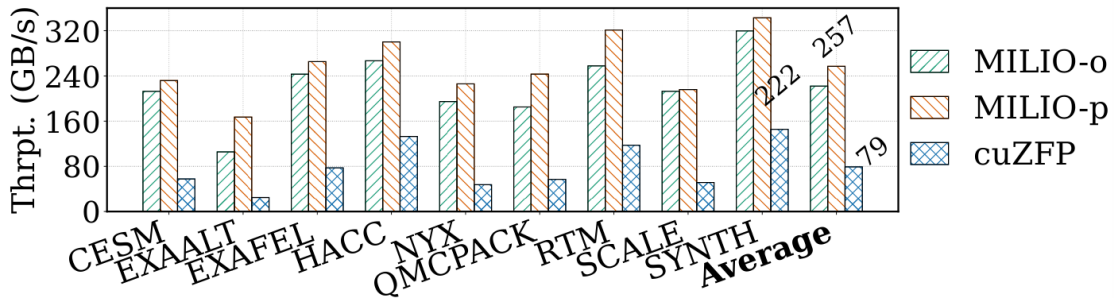
Fig. 9: Code snapshot of MILIO end-to-end measurement

Dataset	Domain	Fields	Dim.	Size
CESM [39]	Climate simulation	79	1,800×3,600	1.47 GB
HACC [37]	Particle simulation	6	10,7372,6487	25.8 GB
SCALE [40]	Climate simulation	12	98×1200×1200	6.8 GB
QMCPack [46]	Quantum Monte Carlo	2	288×115×69×69	1.3 GB
NYX [41]	Cosmology simulation	6	512×512×512	3.2 GB
EXAALT [47]	Molecular dynamics	6	83×107,7290	1.3 GB
EXAFEL [48]	LCLS instrument	1	130×1480×1552	1.2 GB
RTM [36]	Seismic imaging	3	1008×1008×352	4.2 GB
SYNTHESIS [38]	Structural mechanics	1	1200×1200×1200	6.9 GB

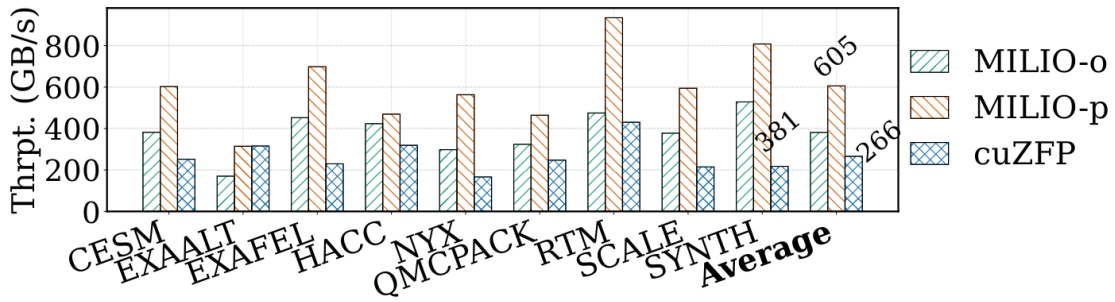
TABLE II: Scientific datasets used in evaluation

- Milio_comp(oridata, compdata, target_ratio) takes device-resident data and returns compressed output on the GPU.
- Dual Xeon E5-2695 v4 with an NVIDIA A100 (**40 GB**), CUDA 11.8 with -O3; nine datasets from SDRBench and Open-SciVis.

Compression and Decompression Throughput



(a) End-to-end compression throughput, averaged over $R \in \{4, 6, 8\}$ (fixed-rate 8, 5, 4 for cuZFP).

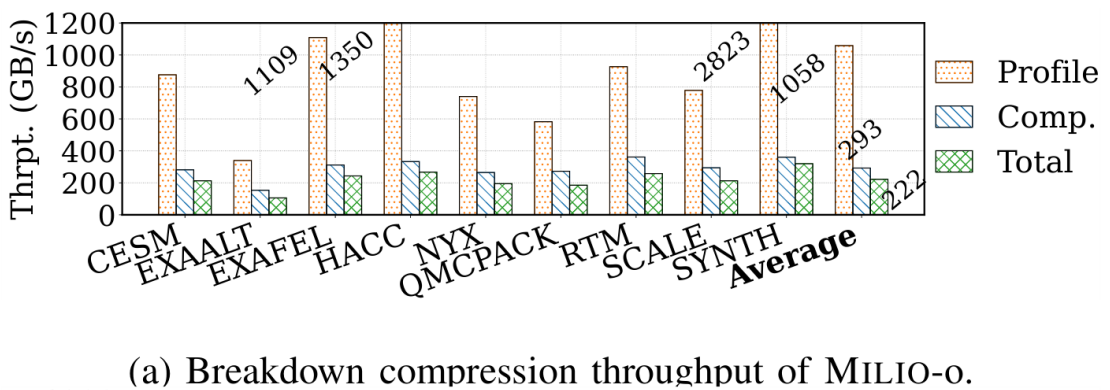


(b) End-to-end decompression throughput, averaged over $R \in \{4, 6, 8\}$ (fixed-rate 8, 5, 4 for cuZFP).

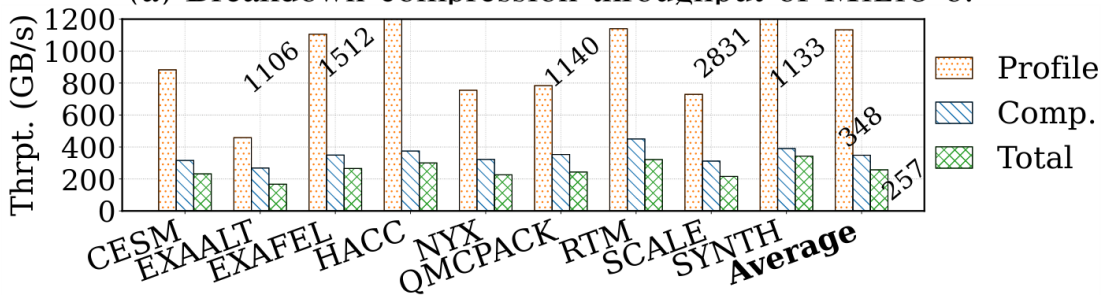
Fig. 10: Throughput of MILIO-p, MILIO-o and cuZFP

- On average, MILIO-p (plain mode) reaches **257.1 GB/s** and MILIO-o (outlier-aware mode) **222.0 GB/s**, compared to **79 GB/s** for cuZFP — **3.25×** and **2.81×** end-to-end speedups.
- On NYX, MILIO-p achieves **226 GB/s** versus only **47 GB/s** for cuZFP (**4.8×**).
- Decompression: MILIO-p averages **605.5 GB/s** and MILIO-o **381.2 GB/s**, versus **265.9 GB/s** for cuZFP.

Breakdown Throughput and Profiling Overhead

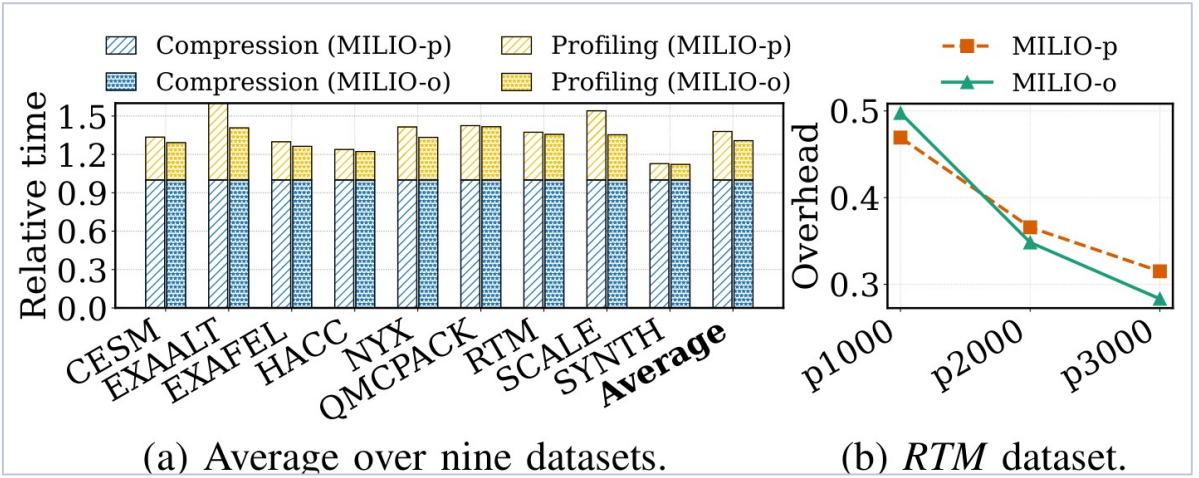


(a) Breakdown compression throughput of MILIO-o.



(b) Breakdown compression throughput of MILIO-p.

Fig. 11: Average breakdown compression throughput



(a) Average over nine datasets. (b) *RTM* dataset.

Fig. 12: Normalized profiling overhead of MILIO

- At R=8, profiling averages **1133.5 GB/s** for MILIO-p and **1058.2 GB/s** for MILIO-o; SYNTHESIS peaks near **2.83 TB/s**.
- Profiling overhead averages 0.395 of compression time for MILIO-p and 0.336 for MILIO-o.

Compression Ratio Accuracy

Dataset	Field	Target R=4				Target R=6				Target R=8			
		MILIO-p		MILIO-o		MILIO-p		MILIO-o		MILIO-p		MILIO-o	
		Ratio	Error-bound	Ratio	Error-bound	Ratio	Error-bound	Ratio	Error-bound	Ratio	Error-bound	Ratio	Error-bound
HACC	vy	**4.030**	2.50e-04	**4.130**	1.90e-04	**6.000**	1.60e-03	**6.010**	1.10e-03	**8.090**	4.60e-03	**8.150**	3.40e-03
CESM	GCLDLWP	**4.060**	1.50e-05	**4.170**	1.50e-05	**6.190**	1.90e-04	**6.230**	1.60e-04	**7.990**	6.40e-04	**8.040**	5.20e-04
EXAFEL	EXAFEL_data	**4.120**	1.60e-04	**4.140**	1.60e-04	**6.020**	8.80e-04	**6.000**	8.50e-04	**8.190**	2.50e-03	**8.270**	2.50e-03
NYX	velocity_x	**4.020**	2.80e-04	**4.210**	1.50e-05	**6.050**	1.90e-03	**6.010**	7.60e-05	**8.200**	5.50e-03	**8.210**	2.20e-04
QMCPack	einspline	**4.110**	1.60e-04	**4.110**	1.60e-04	**6.030**	9.70e-04	**6.030**	9.70e-04	**8.160**	2.80e-03	**8.160**	2.80e-03
SCALE	PRES	**4.060**	1.90e-03	**4.020**	6.40e-05	**6.120**	1.30e-02	**6.120**	4.60e-04	**8.350**	3.40e-02	**8.240**	1.30e-03
EXAALT	dataset1_x	**4.030**	3.10e-03	**4.300**	1.90e-03	**6.020**	1.90e-02	**6.000**	8.80e-03	**8.040**	5.50e-02	**8.690**	3.40e-02
RTM	pressure_3000	**4.010**	1.00e-05	**4.060**	8.00e-06	**6.030**	9.70e-05	**6.080**	7.30e-05	**7.940**	3.10e-04	**8.130**	2.50e-04
SYNTHESIS	synthetic_truss	**3.950**	6.40e-05	**3.983**	6.40e-05	**6.196**	8.80e-04	**6.241**	8.50e-04	**8.139**	2.80e-03	**8.052**	2.50e-03

TABLE III: Achieved ratios and error bounds under R = 4, 6, 8

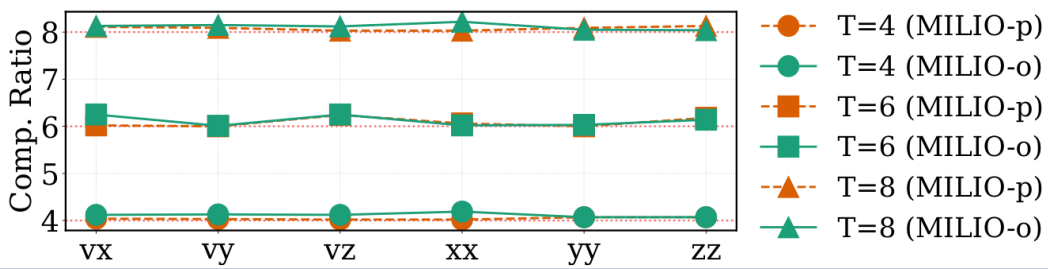


Fig. 13: Target vs. achieved ratios, six HACC fields

- 51 of the 54 configurations land within **5 %** of the target, including all 27 of MILIO-p.
- The three outliers all overshoot, producing smaller outputs than the nominal budget.

Rate–Distortion Analysis

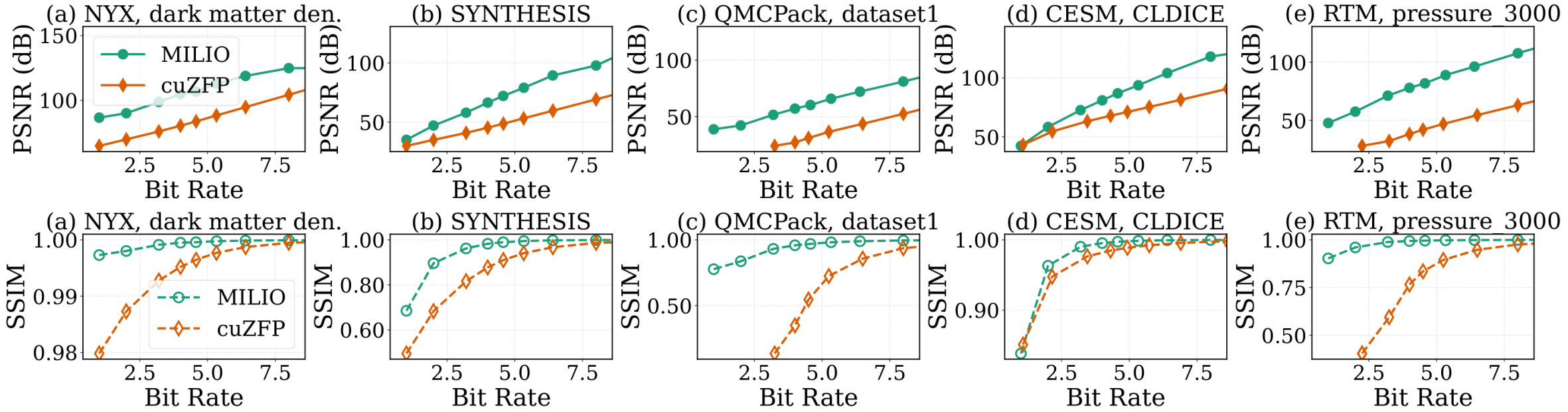
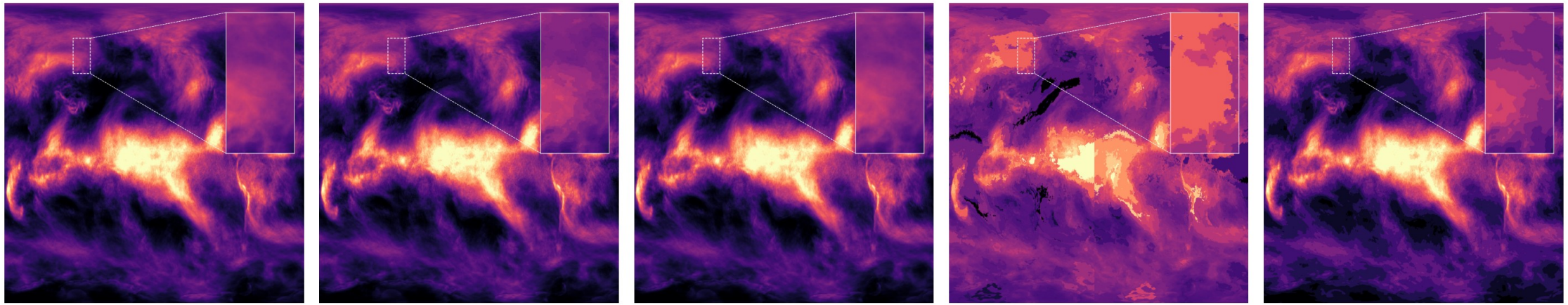


Fig. 14: PSNR (top) and SSIM (bottom) vs. bit rate — MILIO-o vs. cuZFP

- Across the five sparse fields, MILIO-o gives a higher PSNR than cuZFP at every bit rate, and its SSIM reaches near-saturation (≥ 0.99) much earlier.
- On NYX dark matter density MILIO-o is more than **12 dB** ahead; on CESM CLDICE it ends over **30 dB** ahead.

Visual comparison of reconstructed data



(a) Original (*CESM CLDHGH*, z=0 slice)

(b) cuZFP (CR=8.00, PSNR=41.41 dB, SSIM=0.960)

(c) MILIO-o (CR=8.03, PSNR=67.87 dB, SSIM=0.999)

(d) cuZFP (CR=16.0, PSNR=21.17 dB, SSIM=0.708)

(e) MILIO-o (CR=15.99, PSNR=35.82 dB, SSIM=0.874)

Fig. 15: CESM CLDHGH — original, cuZFP and MILIO-o at R≈8 and R≈16

- At R≈8, MILIO-o keeps fine-scale cloud structures (**67.87 dB**), while cuZFP shows blurring and block discontinuities (**41.41 dB**).
- At R≈16, cuZFP collapses under artifact amplification (**21.17 dB**); MILIO-o retains coherent patterns (**35.82 dB**).

Case Study: LLM Inference

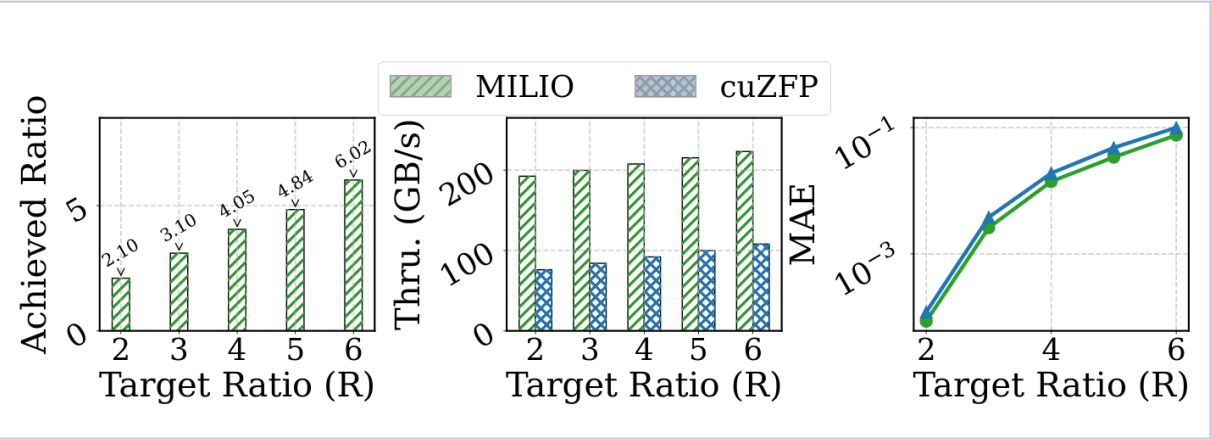


Fig. 16: Achieved ratio, throughput and MAE (LLaMA-2-7B)

Model	Metric	R=2	R=3	R=4	R=5	R=6
LLaMA-2-7B	SQAL-BLEU	100	82	53	26	25
	SQAL-ROUGE-L	100	75	66	66	65
	Qasper-BLEU	95	83	45	15	10
	Qasper-ROUGE-L	100	82	50	60	35
Qwen-2.5-7B	SQAL-BLEU	27	12	13	15	8
	SQAL-ROUGE-L	65	53	65	58	52
	Qasper-BLEU	96	78	25	8	7
	Qasper-ROUGE-L	90	80	90	60	30

TABLE IV: Generation quality under fixed-ratio KV cache

- MILIO runs inside the decoding loop of LLaMA-2-7B and Qwen-2.5-7B on SQuALITY and Qasper.
- Ratios stay within **1–5 %** of target, with **2–2.5×** cuZFP's throughput and lower MAE at every matched ratio.

Case Study: 3D Computational Fluid Dynamics

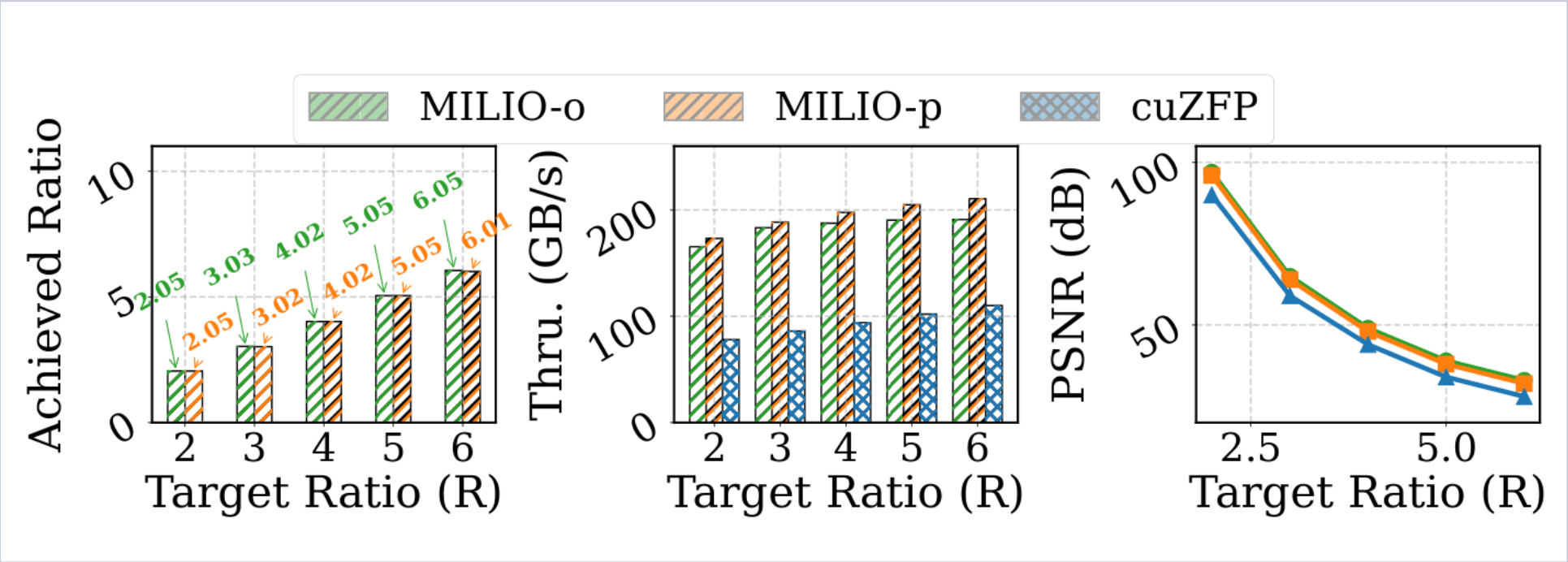


Fig. 17: Achieved ratio, throughput and PSNR (MILIO vs. cuZFP)

- A GPU leap-frog wave-propagation solver on a 1024^3 grid; achieved ratios land within **2.5 %** of every target.
- Both modes sustain **≥ 165 GB/s** while cuZFP stays around **78–110 GB/s**, with higher PSNR at every ratio.

Multi-GPU Scalability

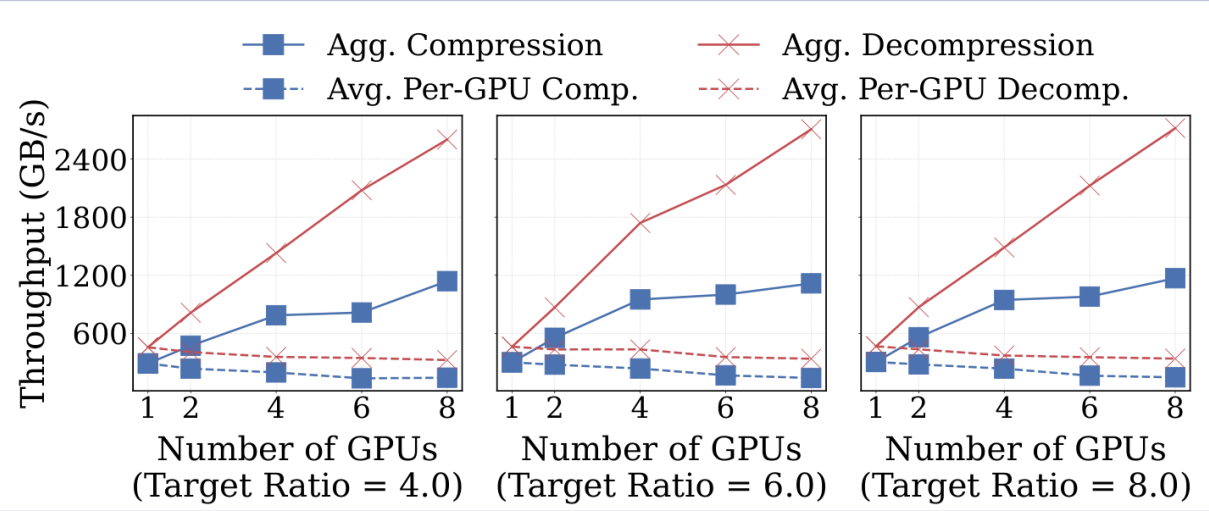


Fig. 18: Throughput on HACC (4.3 GB), 1–8 GPUs

#GPU	Target R=4		Target R=6		Target R=8	
	Avg.	Range	Avg.	Range	Avg.	Range
1	4.03	[4.03, 4.03]	6.00	[6.00, 6.00]	8.09	[8.09, 8.09]
2	4.03	[4.03, 4.03]	6.18	[6.09, 6.26]	8.10	[8.08, 8.12]
4	4.02	[4.00, 4.04]	6.13	[6.03, 6.22]	8.08	[8.01, 8.15]
6	4.05	[4.03, 4.07]	6.14	[6.09, 6.19]	8.07	[8.03, 8.12]
8	4.03	[4.01, 4.04]	6.12	[6.03, 6.18]	8.10	[8.05, 8.16]

TABLE V: Multi-GPU compression ratios on HACC

- Decompression scales from **~457–467 GB/s** on 1 GPU to **~2602–2719 GB/s** on 8 GPUs (**5.7–5.9×**).
- Compression reaches **~1116–1170 GB/s** on 8 GPUs; at R=4 and R=8 each worker stays within **~2 %** of target.

Limitation of MILIO

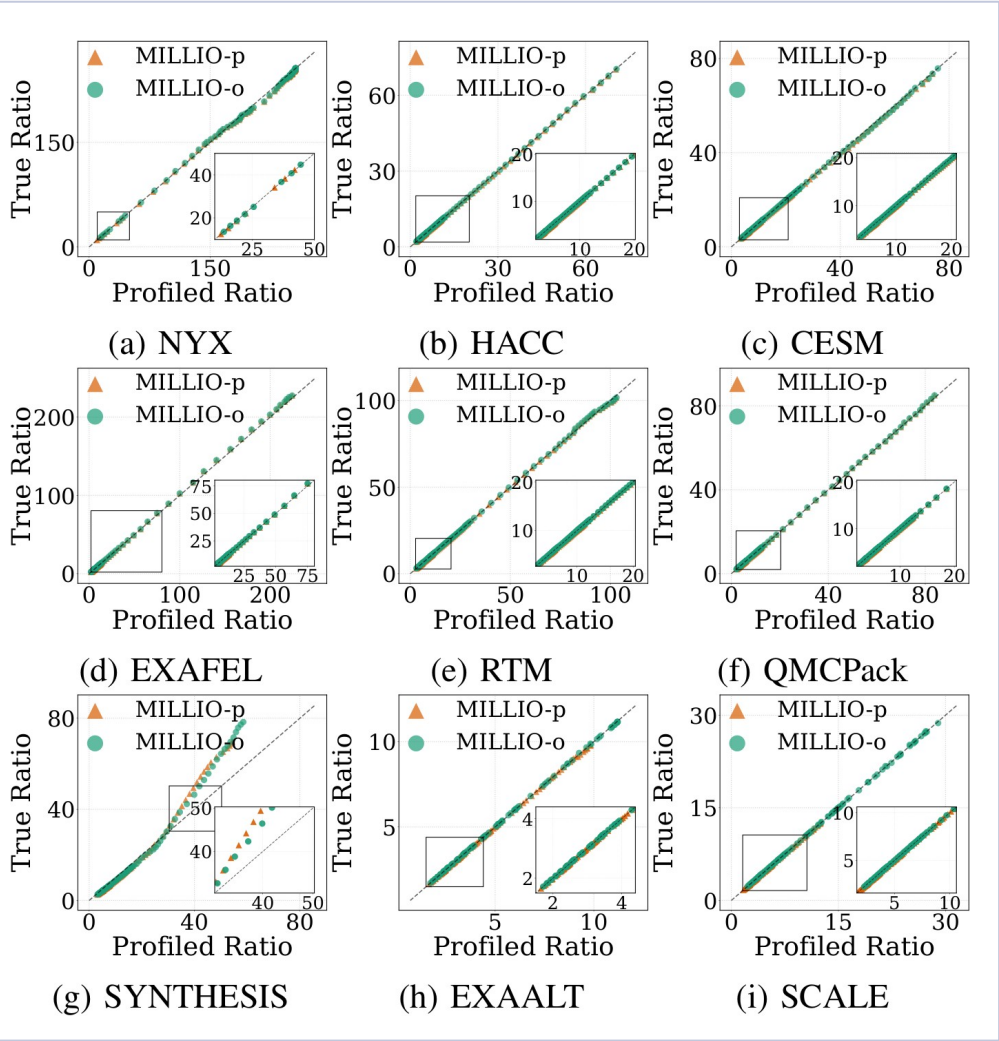


Fig. 19: Profiled vs. true ratios, all 128 bounds

- Profiled and true ratios hug the diagonal across all 128 candidate bounds on almost every field.
- On the eight non-SYNTHESIS datasets, mean absolute deviation stays under **~3 %**.
- SYNTHESIS at the highest ratios reaches up to **~20 %** under-prediction: its data is extremely irregular.

Summary

- MILIO is open source at **github.com/hyfshishen/MILIO**
- **257.1 GB/s** compression throughput on an NVIDIA A100 — **3.25×** cuZFP, measured end to end
- **51 of 54** configurations land within **5 %** of the target ratio
- Two CUDA kernels: a **TB/s-level** profiling kernel evaluates **128 candidate error bounds** in a single pass
- Error-bounded fidelity with fixed-ratio size predictability — **no trial-and-error re-compression**
- Validated inside LLM KV-cache, 3D CFD and multi-GPU workflows

Bohan Zhang
University of Florida
bohanzhang1@ufl.edu

UF Electrical & Computer Engineering UNIVERSITY of FLORIDA