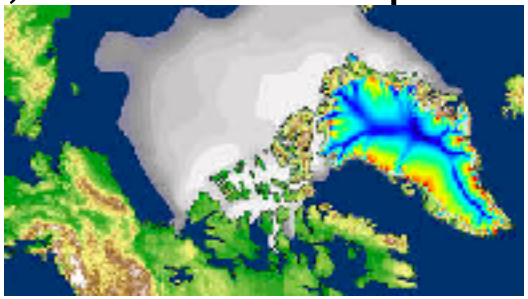
Error-Controlled, Progressive, and Adaptable Retrieval of Scientific Data with Multilevel Decomposition

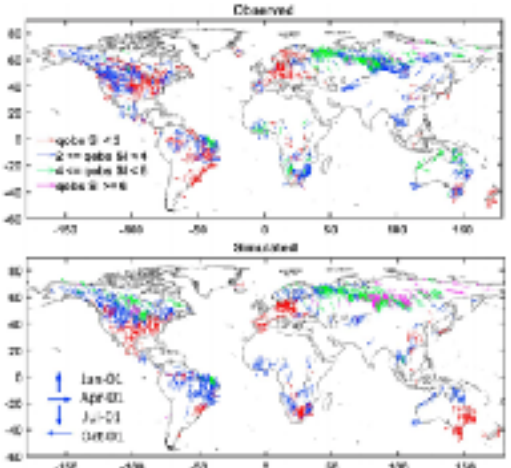
Xin Liang*, Qian Gong+, Jieyang Chen+, Ben Whitney+, Lipeng Wan+, Qing Liu^,
David Pugmire+, Rick Archibald+, Norbert Podhorszki+, and Scott Klasy+
*Missouri University of Science & Technology
+Oak Ridge National Laboratory
^New Jersey Institute of Technology

Data Challenges – Simulations

Climate simulation

- Climate and Earth System Model: a fully coupled numerical simulation of the Earth system consisting of atmospheric, ocean, ice, land surface, carbon cycle, and other components





Seasonality of observed (upper) and simulated (lower) streamflow at stream gauge stations in major river basins around the world in E3SM [1].

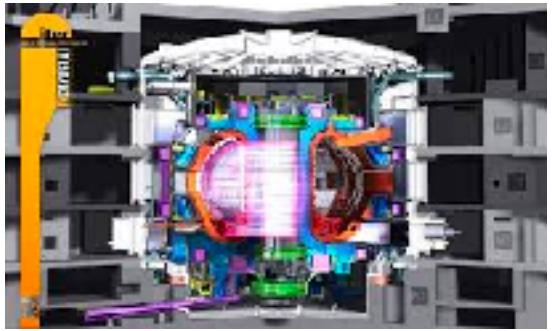
- Acquire thousands of high-resolution spectra every day and can cost a long time to transmit/retrieve for observation data
- Requirements to follow FAIR data principles will help scientists re-use and analysis results of data and help with reproducibility.

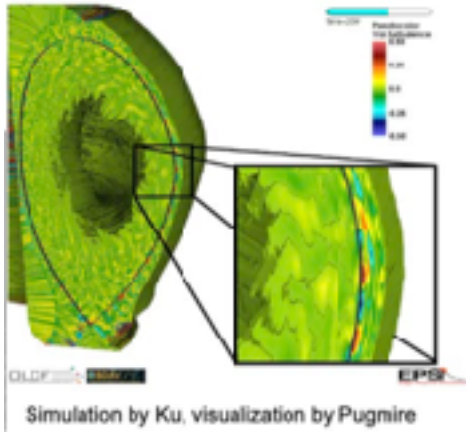
[1] Golaz, Jean-Christophe, et al. "The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution." Journal of Advances in Modeling Earth Systems 11.7 (2019): 2089-2129.

Data Challenges – Simulations

Fusion simulation

- Enables fusion reactor operation using magnetically confined plasma
- Long enough confinement of D-T plasma at ~15keV is key





"Streamer" and "blob" edge turbulence eddy patters from XGC kinetic particle code

- Simulation data will be Exa-Bytes per day
- Various synthetic diagnostics data from simulation and thousands of sensors from experiment need to get married with dynamic multiscale interaction in mind

Computing vs Storage and I/O

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Theta 2016	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	> 8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	17	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Volta GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s Lustre®	32 PB 1 TB/s Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

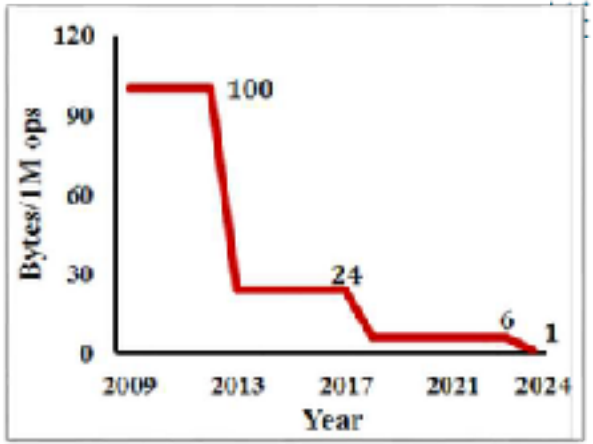


Figure and table from S. Klasky (ORNL)

- Limited storage capacity upgrade
  - ➢ Titan -> Summit: 4x in storage vs 6x in computing
- High computing-to-I/O ratio
  - ➢ Summit: 150 PF/s, 1 TB/s I/O → 1M FLOPS per double
  - ➢ Much more severe when data reside in secondary storage

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Lossless Compression: Challenges

- Cannot do much for scientific data
  - Mantissas are like noise…
- Compression ratio is usually less than 2
  - Far from what is desired!

Scheme	Transformation Applied	Algorithm	Compression Ratio
FPC [8]	not used	it first predicts values sequentially using two predictors (FCM and DFCM), and subsequently selects the closer predicted value to the actual. Lastly, it XORs the selected predicted value with the actual value, and leading-zero compresses the result.	1.02x~1.96x
ISOBAR [30]	divide byte-columns into compressible and incompressibles	apply zlib, bzlib2, (fpzip, FPC) on all compressible (after discarding noisy byte-columns). zlib is the main compression algorithm; others are for comparison purposes	1.12x~1.48x
PRIMACY [31]	frequency based permutation of ID values	apply zlib on transformed data	1.13x~2.16x
ALACRITY [19]	split floating-point values into sign, exponent, and significands	unique-value encoding of the most significant bytes (assuming high-order bytes (sign and exponents) are easy to compress); low-order bytes are compressed using ISOBAR	1.19x~1.58x
CC [6]	XOR on Δ of neighboring data point in the same iteration	apply zero-filled run length encoding	up to 2.13x
IOFSL [36]	not used	integration of LZO, bzip2, zlib within the I/O forwarding layer	~1.9x
Binary Masking [5]	bit masking (XOR)	apply zlib on bit masked data in order to partially decreases the entropy level	1.11x~1.33x
MCRENGINE [18]	variable merging in the same group	apply parallel gzip on the merged variables across processes	up to 1.18x

Deflate algorithm

Data compression for the Exascale Computing Era Survey.
S. Son et al. (Supercomputing frontiers and innovations 2014)

Floating point data set (numerical simulation of the brain):

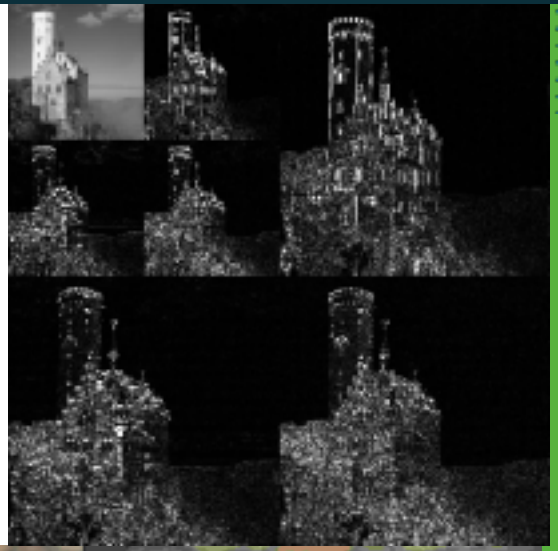
Random (noise)

Source: Leonardo Bautista Gomez (BSC)

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Lossy Compression: Advantages

- Tradeoff accuracy for compression ratio
  - The data is altered during compression: some piece of information is removed, the original data cannot be retrieved
- Tunable ratios
  - High compression ratio is achievable
- Multiscale representation



OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Lossy Compression: Requirements

- Error control
  - The difference between original data and decompressed data is bounded by user-specified tolerance
- Progressive
  - Ability of providing up to near lossless data with incremental update to allow for reproducibility with high efficiency
- Adaptable
  - To resolution: Ability of adjusting resolution for fast computation
  - To analysis: Ability of prioritizing data based on target analysis

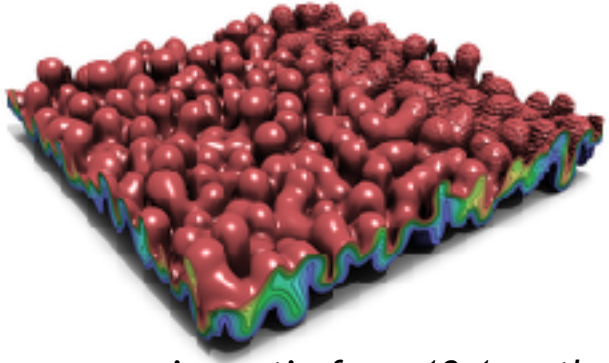
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Error-Controlled Lossy Compressors

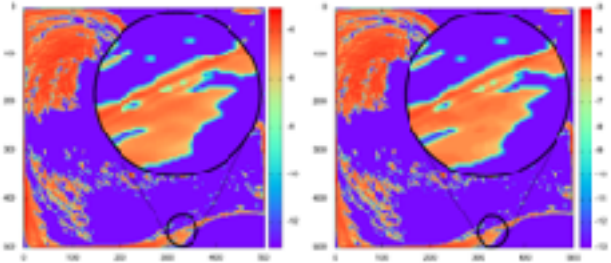
- Add error control to general lossy compression techniques to ensure "trust"
  - The distortion of decompressed data and original data is bounded by some metric

No/weak progressiveness

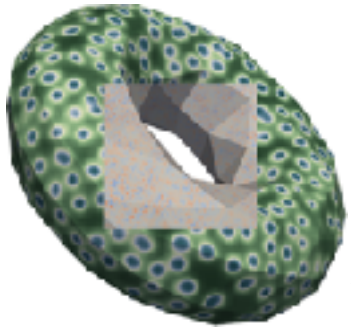
Not adaptable to analysis/resolution



Varying compression ratio from 10:1 on the left to 250:1 on the right for scientific data with error-controlled lossy compressor ZFP. Figure from P. Lindstrom (LLNL)



Visualization of SZ decompressed data when compression ratio goes up to 64 Figure from S. Di (ANL)



MGARD error-controlled lossy compression for unstructured data

Figure from B. Whitney (ORNL)

OAK RIDGE National Laboratory

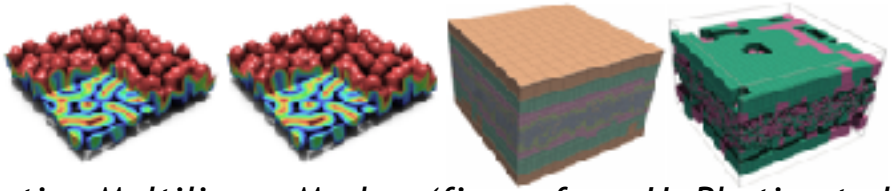
MISSOURI S&T

NJIT New Jersey Institute of Technology

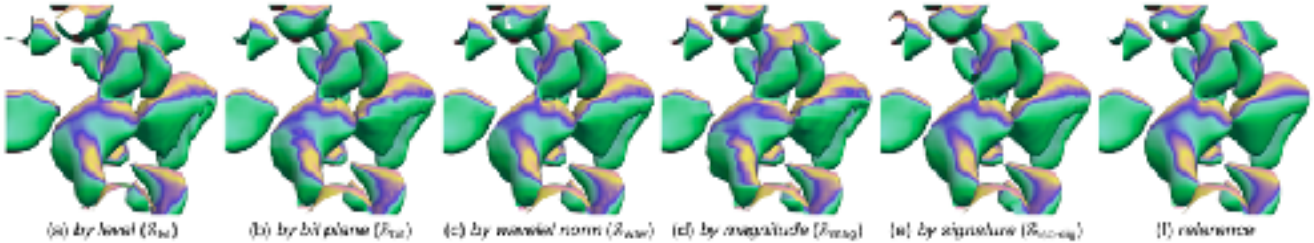
Progressive Compressor

- JPEG/JPEG2000
  - Wavelet methods with mixed precision and resolution
- Task-optimized streaming
  - Precision-resolution trade-off with varying data streams
- Adaptive Multilinear Meshes
  - Adaptive representations using multilinear wavelets

No/weak error control
Not adaptable to analysis



Adaptive Multilinear Meshes (figure from H. Bhatia et al)

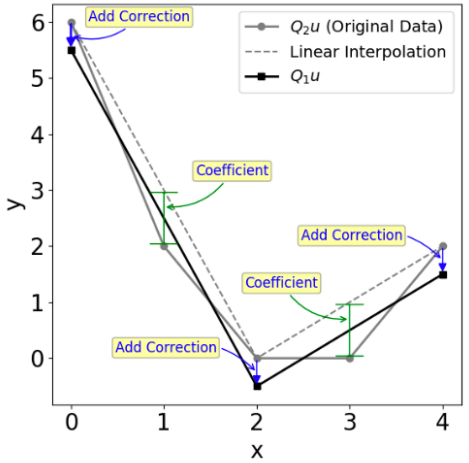


Task-optimized streaming (figure from D. Huang et al)

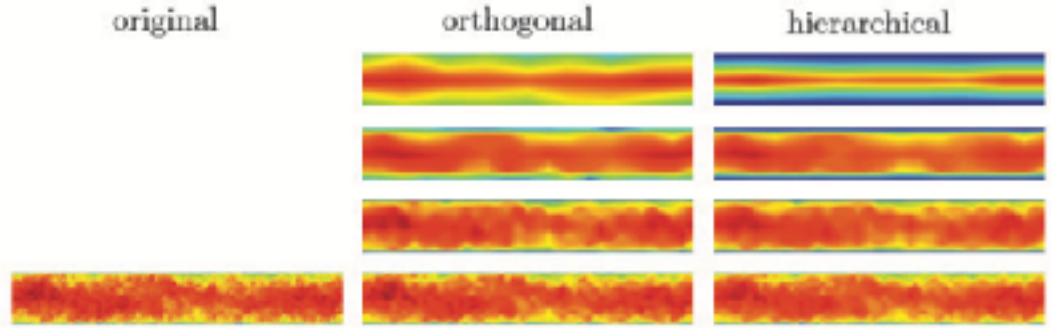
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

MGARD: Multilevel Data Reduction

- MGARD provides multilevel representations with error control
  - Works for non-uniform data grid
  - Offers multi-resolution representation
  - Supports most error-control metrics
    - L-infinity: $|u - \tilde{u}|_{L^\infty} \leq \tau$
    - L²: $|u - \tilde{u}|_{L^2} \leq \tau$
    - QoI: $|\mathcal{Q}(u) - \mathcal{Q}(\tilde{u})| \leq \tau$



1D Decomposition in MGARD

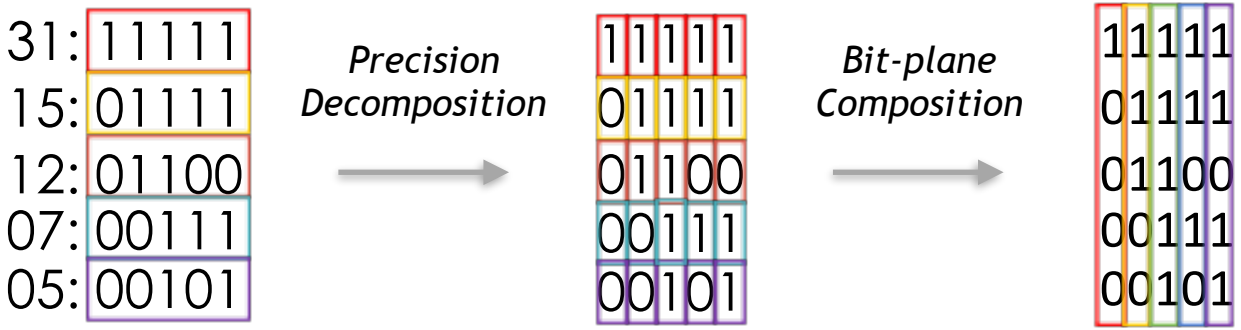


MGARD Decomposition (orthogonal) versus Multilinear Decomposition (hierarchical)

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Bit-Plane Encoding

- Represent data in sign-magnitude format
  - Progressive in precision/accuracy
  - Prioritized in information
  - Automatic error control
  - Works with floating-point data with exponent align



Bitplane representation

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Method

- Coupling MGARD decomposition with bit-plane (BP) encoding
  - Derivable error control
    - Combine error controls in MGARD and BP encoding
  - Progressive
    - In accuracy thanks to MGARD error control and BP encoding
    - In computation thanks to MGARD decomposition and BP formats
  - Adaptable
    - To resolution thanks to MGARD decomposition
    - To analysis by bit-plane segmentation with on-demand retrieval

Combine MGARD with bit-plane encoding

Derivable Error Control

- Collect error information using auxiliary arrays

$$A_{L^\infty}[l][b_l] = \|\Delta_l u - \Delta_l \tilde{u}^{b_l}\|_{L^\infty} \qquad A_{L^2}[l][b_l] = \|\Delta_l u - \Delta_l \tilde{u}^{b_l}\|_{L^2}^2$$

- Translate MGARD error control to number of required BPs

MGARD: $\|u - \tilde{u}\|_{L^\infty} \leq C_{L^\infty} \sum_{l=0}^{L} \|\Delta_l u - \Delta_l \tilde{u}\|_{L^\infty}$ *MGARD:* $\|u - \tilde{u}\|_{L^2}^2 \leq \sum_{l=0}^{L} vol(P_l) \|\Delta_l u - \Delta_l \tilde{u}\|_{L^2}$

MGARD-BP: $\|u - \tilde{u}\|_{L^\infty} \leq C_{L^\infty} \sum_{l=0}^{L} A_{L^\infty}[l][b_l]$ *MGARD-BP:* $\|u - \tilde{u}\|_{L^2}^2 \leq \sum_{l=0}^{L} vol(P_l) A_{L^2}[l][b_l]$

MGARD: $\|Q(u) - Q(\tilde{u})\| \leq \Upsilon_s(Q)^2 (\sum_{l=0}^{L} 2^{2sl} vol(P_l) \|\Delta_l u - \Delta_l \tilde{u}\|_{L^2}^2)$

MGARD-BP: $\|Q(u) - Q(\tilde{u})\|^2 \leq \Upsilon_s(Q)^2 (\sum_{l=0}^{L} 2^{2sl} vol(P_l) A_{L^2}[l][b_l])$

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Progressive Reconstruction

- Incremental update to save reconstruction cost
  - Current representation (with b_l BPs from level l):

$$u_1 = \sum_{l=0}^{L} (I - Q_{l-1}) \Delta_l u^{b_l}$$

  - Next representation (with b_l' BPs from level l):

$$u_2 = \sum_{l=0}^{L} (I - Q_{l-1}) \Delta_l u^{b_l'} = \sum_{l=0}^{L} (I - Q_{l-1}) (\Delta_l u^{b_l} + \Delta_l u^{b_l' - b_l})$$

$$= \sum_{l=0}^{L} (I - Q_{l-1}) \Delta_l u^{b_l} + \sum_{l=0}^{L} (I - Q_{l-1}) \Delta_l u^{b_l' - b_l}$$

$$= u_1 + \sum_{l=0}^{L} (I - Q_{l-1}) \Delta_l u^{b_l' - b_l}$$

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Refactoring Phase

Table 2: Metadata recorded during refactoring/writing

Name	Type	Size	Description
$\|\|\delta_l\|\|_{L^\infty}$	double	L	Maximum level-wise coefficient value.
A_{L^2}	double	$L \times (B+1)$	Squared L^2 error matrix.
S	integer	$L \times B$	Size of encoded bit-planes in each level.

Algorithm 2 Multilevel data refactoring with bit-plane encoding.

Input: original data u; maximum level L; number of encoding bit-planes B.

```
Q_L ← I                                          ▷ initialization
for l = L → 0 do
    Q_{l-1}u, u_mc[N_l^*] = decompose(Q_l u)     ▷ decompose the current data
    buffer ← interleave(u_mc[N_l^*])             ▷ collect multilevel components
    ||δ_l u||_{L^∞} ← max(buffer)                ▷ compute ||δ_l||_{L^∞}
    A_{L^2}[l] ← compute_A_{L^2}(buffer)         ▷ compute A_{L^2}[l]
    streams[l], S[l] ← encoding(buffer, ||δ_l u||_{L^∞})  ▷ bit-plane encoding and lossless compression
end for
write_to_PFS(metadata)                           ▷ write metadata to PFS
write_to_PFS(streams)                            ▷ write data to PFS
select_and_write_to_tape(streams)                ▷ move low-precision portions to tapes
```

Key difference with existing approaches

- Metadata collection
  - For error-controlled retrieval
- Level-wise segmentation
  - For adaptable retrieval to given data analytics

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Retrieval Phase: Size Interpretation

For $M \in \{L^\infty, L^2\} \bigcup \{s\}$, *define:*

$$\epsilon_{L^\infty}(l,k) = C_{L^\infty} A_{L^\infty}[l][k]$$

$$\epsilon_{L^2}(l,k) = vol(P_l) A_{L^2}[l][k]$$

$$\epsilon_s(l,k) = \Upsilon_s(Q)^2 2^{2sl} vol(P_l) A_{L^2}[l][k]$$

Make greedy decisions on demand based on:

$$\eta_M(l,k) = \frac{\epsilon_M[l][k] - \epsilon_M[l][k+1]}{S[l][k]}$$

Key difference with existing approaches

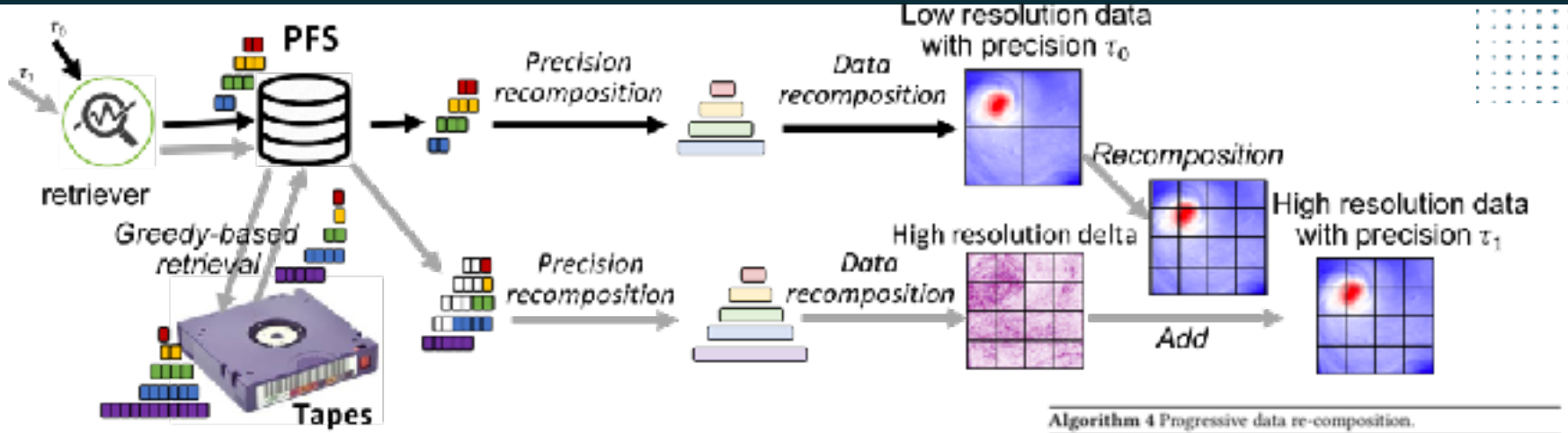
- Enables error control
- Accuracy checking
  - For resolution adaptability
- On-demand retrieval based on s
  - For analysis adaptability

Algorithm 3 Interpretation of bit-plane retrieval for required error tolerance.

Input: required error tolerance τ; current recomposed level L'; total number of decomposition levels L; number of bit planes fetched by now $index[0:L]$; number of recorded bit-planes B; error metric $M \in \{L^\infty, L^2\} \bigcup \{s\}$; level sizes S; error matrix A_M ($A_s = A_{L^2}$). (the last fives are read from metadata)

```
1  read(metadata)                                   ▷ read metadata
2  τ_min ← 0, τ_0 ← 0, max_heap ← ∅
3  for l = 0 → L do                                 ▷ initialize index and heap
4      τ_min ← τ_min + ε_M(l, B)                    ▷ compute the achievable minimal error
5      τ_0 ← τ_0 + ε_M(l, index[l])                 ▷ accumulate current errors in each level
6      max_heap.push({η_M(l, index[l]), l})         ▷ push bit-plane to heap
7      if resolution_first and l ≥ L' and τ_min ≤ τ then
8          L ← l                                    ▷ limit the number of levels to l if tolerance can be met
9          break
10     end if
11 end for
12 while max_heap ≠ ∅ and τ_0 ≤ τ do
13     l ← max_heap.pop()                           ▷ get index of level with the largest efficiency
14     τ_0 ← τ_0 − ε_M(l, index[l]) + ε_M(l, index[l] + 1)
15     index[l] ← index[l] + 1                      ▷ increment bit-plane index for the selected level
16     if index[l] ≠ B − 1 then                     ▷ push next bit-plane to heap if exists
17         max_heap.push({η_M(l, index[l]), l})
18     end if
19 end while
20 return index, L
```

Retrieval Phase



Features

- Error control for all supported metrics in MGARD
- Progressive in both precision and recomposition
- Adaptable to both resolution and analysis

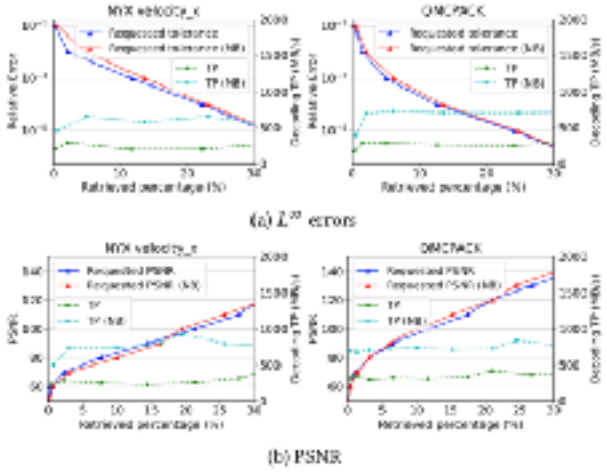
Algorithm 4 Progressive data re-composition.

Input: required error tolerance τ; current data representation u'; current recomposed level L'; total number of decomposition levels L; number of bit-planes fetched for current representation $index[0:L]$.

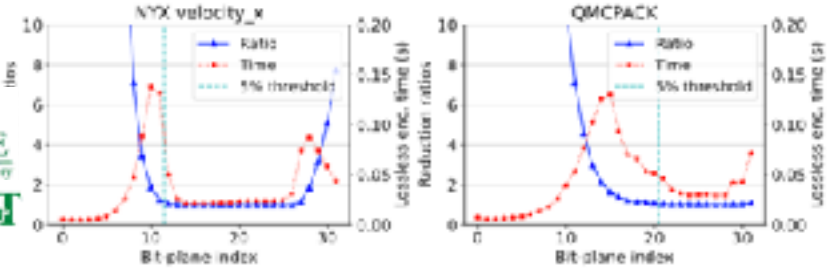
```
next_index, L' ← size_interpretation(τ, L', index)        ▸ Interpret size
Q_{-1} ← NULL
for l = 0 → L do
    streams[l] ← read(index[l], next_index[l])            ▸ read refactored data (may from PFS directly or from tapes)
    buffer ← decoding(streams[l], pre_index[l], index[l])  ▸ lossless decompression and bit-plane decoding
    Q_mc[N_l'] ← deinterleave(buffer)                     ▸ put multilevel components in place
    Q_l û = recompose(Q_{l-1} û, Q_mc[N_l'])              ▸ recompose the delta
end for
for l = L' → L'' do                                       ▸ recompose current data to the same resolution if need
    u' ← recompose(u', {l})
end for
L' ← L'', index ← next_index                              ▸ maintain necessary variables
return u' + Q_l û                                         ▸ add delta to current data
```

Optimizations

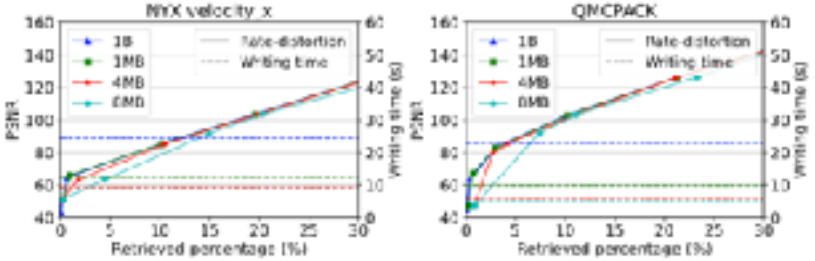
- Adaptive encoding algorithms
  - General BP encoding
  - Negabinary encoding
- Adaptive lossless compression
  - Skip hard-to-compress BPs
- Adaptive BP aggregation
  - Merge adjacent BPs to one file



Adaptive encoding



Adaptive lossless compression



Adaptive aggregation

Evaluation: Setup

- Platform: Andes @ OLCF
  - 704 nodes with AMD EPYC 7302
  - 32 cores/node with 256 GB memory
  - Alpine IBM Spectrum Scale Filesystem
  - HPSS Data Archival System
- Datasets
  - Climate, cosmology, weather, quantum MC

Andes cluster

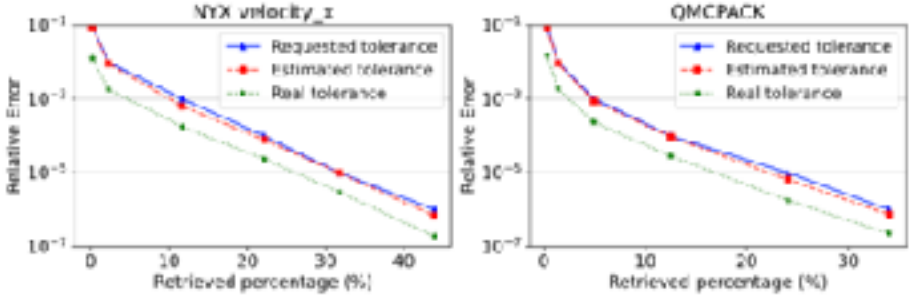
Table 3: Datasets for evaluation

Dataset	#Fields	Dimension/core	Size/core
Hurricane Isabel	13	100 × 500 × 500	1.21 GB
NYX	6	512 × 512 × 512	3 GB
SCALE-LETKF	12	98 × 1200 × 1200	6.31 GB
QMCPACK	1	288 × 115 × 69 × 69	0.59 GB

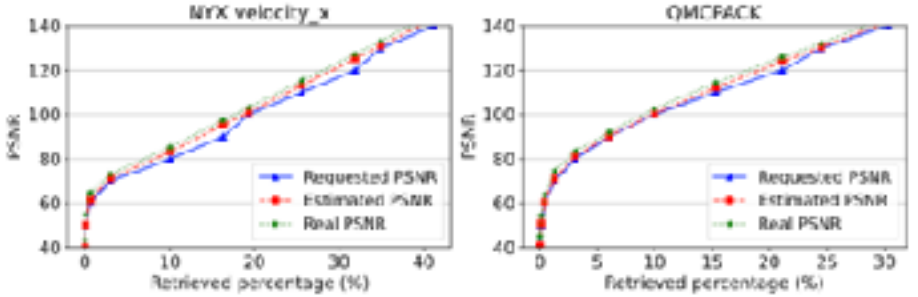
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Evaluation: Error Control

- Error control
  - L-infinity norm
  - L2 norm



(a) L^∞ errors (using general bit-plane encoding)



(b) PSNR (using negabinary encoding)

Guaranteed error control

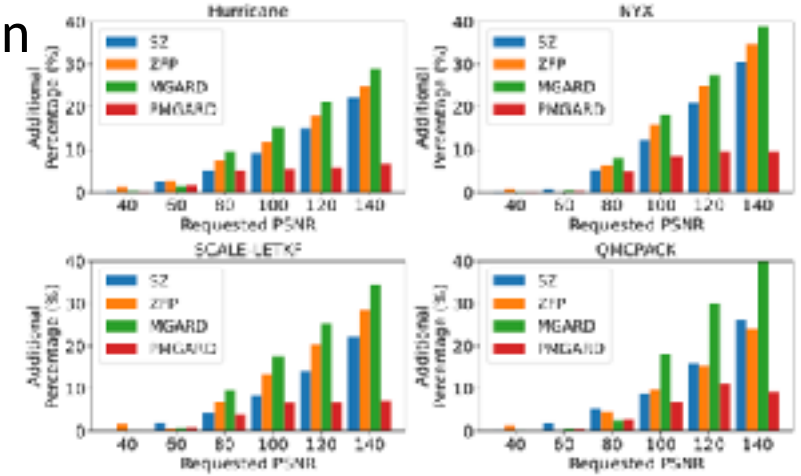
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Evaluation: Progressiveness

- Error control
  - L-infinity norm
  - L2 norm
- Retrieval percentage
  - Single precision
  - Progressive precision
  - Parallel evaluation

Table 4: Retrieval percentages for one precision

Dataset	Method	PSNR					
		40	60	80	100	120	140
Hurricane	MGARD	0.26%	1.48%	5.58%	15.19%	21.26%	28.83%
	PMGARD	0.08%	1.75%	6.92%	12.41%	18.15%	24.94%
NYX	MGARD	0.05%	0.41%	8.14%	18.39%	27.40%	39.04%
	PMGARD	<0.01%	0.28%	5.05%	13.46%	23.06%	32.63%
SCALE-LETKF	MGARD	0.12%	0.83%	9.52%	17.50%	25.34%	34.45%
	PMGARD	0.03%	1.00%	4.90%	11.66%	18.46%	25.83%
QMCPACK	MGARD	0.13%	0.64%	2.53%	17.99%	30.15%	41.85%
	PMGARD	0.05%	0.43%	3.13%	10.06%	21.10%	30.96%

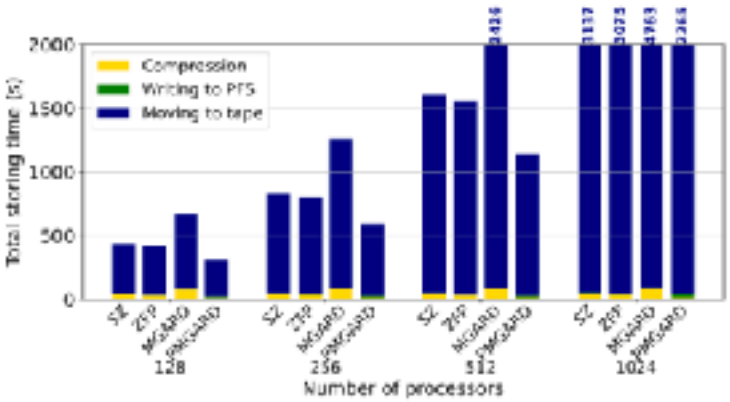


Retrieval percentages under progressive requirements

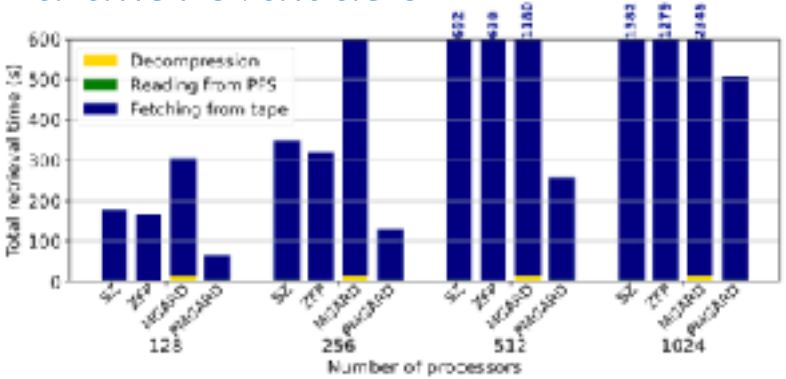
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Evaluation: Progressiveness

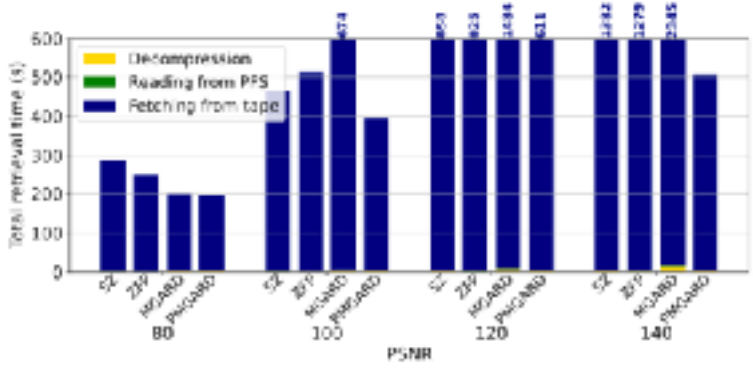
- Error control
  - L-infinity norm
  - L2 norm
- Retrieval percentage
  - Single precision
  - Progressive precision
  - Parallel evaluation



Refactoring and writing



Progressive retrieval: PSNR 120—>140



Progressive retrieval: 1,024 cores

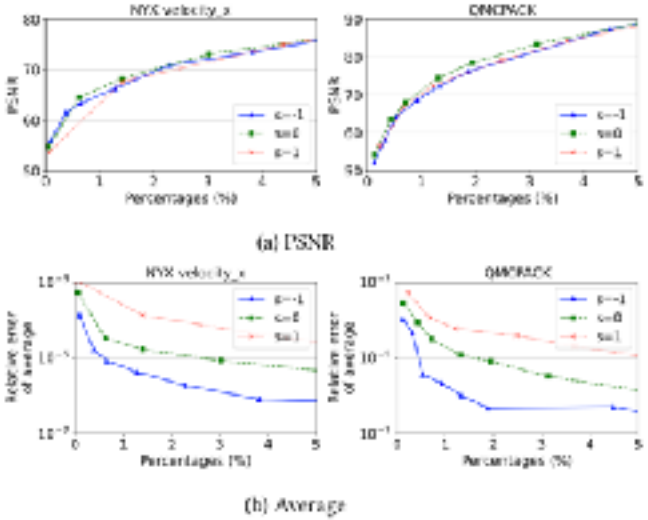
OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Evaluation: Adaptability

- Error control
  - L-infinity norm
  - L2 norm
- Retrieval percentage
  - Single precision
  - Progressive precision
  - Parallel evaluation
- Adaptability
  - Resolution
  - Analysis

Table 5: Performance of iso-surface analysis when target PSNR is 60 (NYX velocity_x)

Method	Percentage	Resolution	Decompression time (s)	Analysis time (s)	Analysis error
SZ	2.19%	512 × 512 × 512	0.96	60.83	2.25%
ZFP	1.78%	512 × 512 × 512	0.51	59.99	5.89%
MGARD[1]	10.62%	257 × 257 × 257	0.79	10.99	5.08%
PMGARD[2]	0.81%	257 × 257 × 257	0.46	11.16	5.67%



Impact of smooth parameter s on different analysis

OAK RIDGE National Laboratory
MISSOURI S&T
NJIT New Jersey Institute of Technology

Future Work

- Extension to other decomposition algorithms
- Exploration of other progressive formats
- Extension to more analytics
- Extension to error-controlled AMR representation

- Acknowledgment


U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

SIRIUS

RAPIDS

OAK RIDGE National Laboratory

MISSOURI S&T

NJIT New Jersey Institute of Technology

Q&A

Thank you for listening!

ANY QUESTIONS?

OAK RIDGE National Laboratory

MISSOURI S&T

NJIT New Jersey Institute of Technology