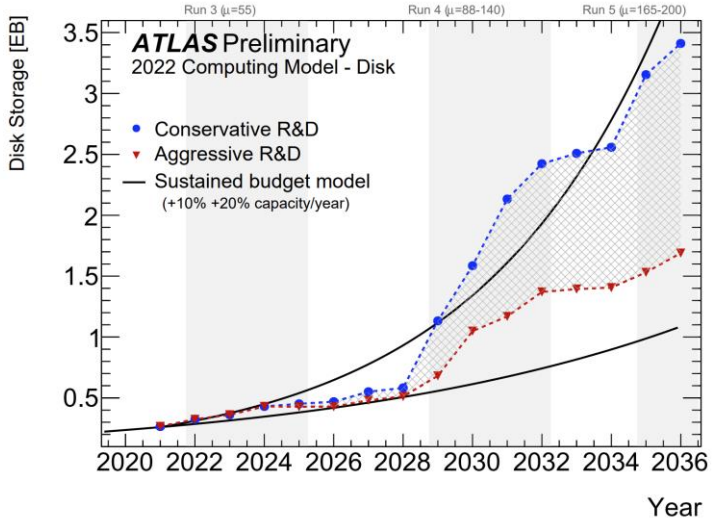
WE START WITH ~~YES,~~ DATA

HEP-CCE I/O STRATEGIES STORAGE & COMPRESSION OUTLOOK



PETER VAN GEMMEREN

ROB ROSS

HEP-CCE/IOS

April 11th, 2023

HEP-CCE/IOS: STORAGE AND COMPRESSION

Future Priorities

- The current cycle of HEP-CCE has been mainly focused on making HEP applications make [efficient] use of High Performance Computer
- This addresses the crucial need for **CPU cycles** expected for HEP experiment at the HL-LHC, DUNE and beyond.
- HEP, however, faces similar challenges for **disk and tape storage**, which also need to be addressed
  - Additional compute cycles may help, but won't solve this issue
- Most experiment HEP data is stored in a compressed format using **standard loss-less compression** algorithms
- More advanced/intelligent, but often lossy compression algorithms are less common
  - Exception: CMS Nano-AOD, soon? ATLAS PHYSLITE

INTELLIGENT LOSSY COMPRESSION

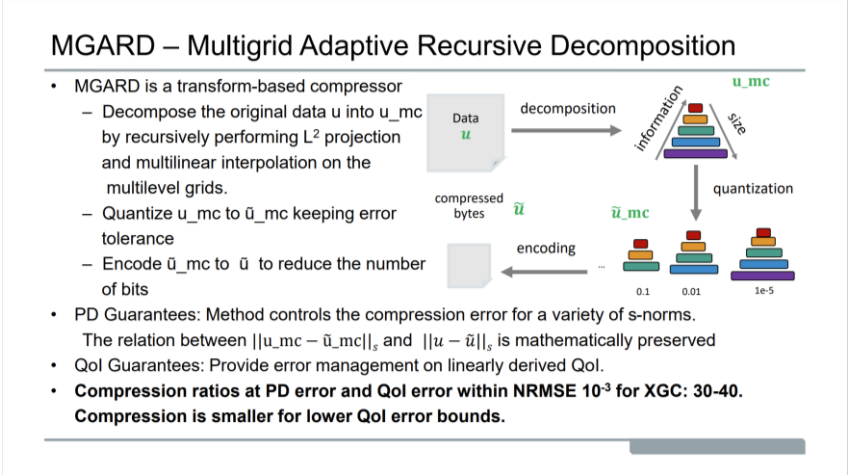
Computing Science

- Overview of "intelligent" data compression:
- Oct 18, 2022: **Speakers**: Prof. Anand Rangarajan , Prof. Sanjay Ranka:
  Hybrid Learning Techniques for Scientific Data Reduction with Performance Guarantees
- Nov 29, 2022: **Speakers**: Dr Franck Cappello (ANL), Dr Sheng Di (ANL):
  Compression of Scientific Data with SZ
- Mar 21, 2023: **Speaker**: John Wu (LBNL):
  Statistical Similarity for Data Compression

HYBRID LEARNING TECHNIQUES FOR SCIENTIFIC DATA REDUCTION

Prof. Anand Rangarajan , Prof. Sanjay Ranka: **GitHub - CODARcode/MGARD: MGARD: MultiGrid Adaptive Reduction of Data**

- Compression of scientific applications differ from video and image compression
  - Guarantees on Quantities of Interest (QoI): Scientists are principally interested in QoI that are derived from raw data. The ability to quantify these with realistic bounds is essential.
- Compression Ratio of ~30-40 for fusion code data

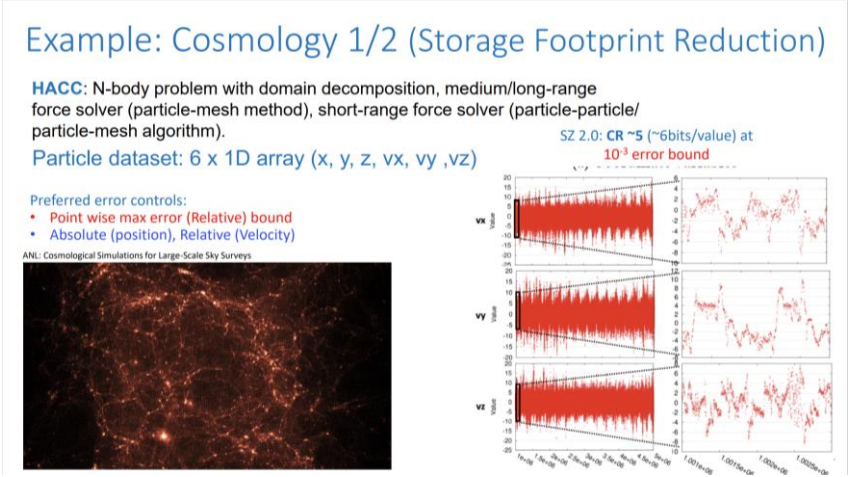


- Used for XGC Fusion Code that can produce 4.3 PB/day.

COMPRESSION OF SCIENTIFIC DATA WITH SZ

Dr Franck Cappello (ANL), Dr Sheng Di (ANL): **SZ Lossy Compression | SZ Lossy Compressor (szcompressor.org)**

- Consist in reducing scientific data volume by leveraging correlations and reducing precision
- Goal: keep the same science
  - Requires error bounds on observables
- Compression Ratio of ~5-100 for scientific data

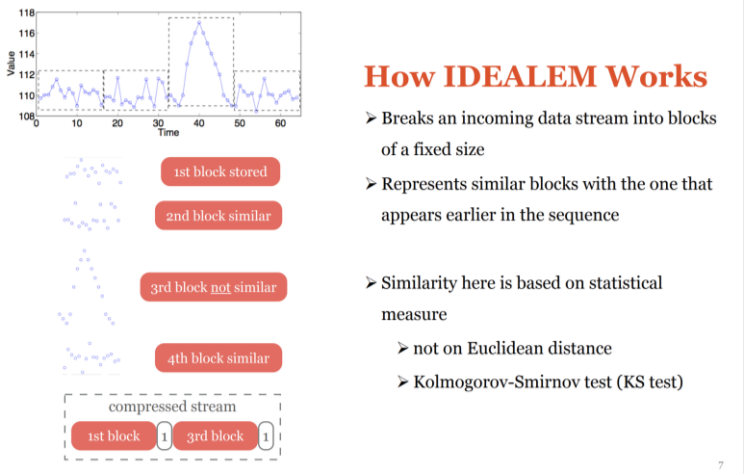


- An example application others understand better than me ;).

STATISTICAL SIMILARITY FOR DATA COMPRESSION

John Wu (LBNL): **IDEALEM at LBNL (lbl.gov)**

- Motivated by reading out many (1000s) of micro-Phaser Measurement Unit over time
  - Monitoring device is capable of sample dozens of measures many thousands of times a second
- That's for the power grid, don't ask me how, but does not sound so unsimilar to some detectors.
- Compression Ratio of ~100-200 in PMU example!



- One drawback/challenge: KS test is computational expensive.

OUTLOOK

- HEP-CCE/IOS has accomplished their main goals:
  - **Darshan for ROOT I/O** in HEP workflows on HPC
  - Investigate **HDF5 as intermediate event storage** for HPC processing using ROOT serialization
  - Start to Investigate **HPC/GPU friendly data model** (joined with PPS)
- Documentation (and CHEP publications) are being worked on and should be completed before project end.
- Foundation research for next HEP-CCE cycle work is well underway:
  - Investigate **HPC/GPU friendly data model** (joined with PPS)
  - Intelligent lossy **compression for storage** needs reduction
- Big Thanks to everyone involved, HEP-CCE and experiments!