I/O Performance with Compression Enabled

Haiying Xu, John Dennis
NCAR/ASAP/IOWA

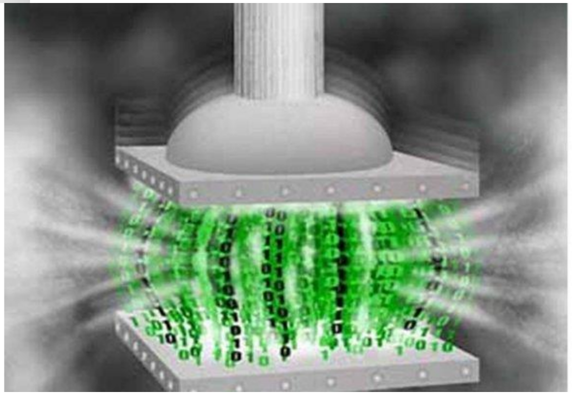
June 19, 2019

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

CESM Data Volume Increasing

CMIP5 175TB

CMIP6 230TB (compressed)



Introduction

- CESM need compression with good IO performance
  - Enabled parallel compression
  - Enabled asynchronous read/write in PIO
  - Evaluate cloud-friendly data format
  - Compare IO alternatives with IOR benchmark results
  - Conclusion

Impact of Compression on IO Performance

- Compression performance analysis
  - Configuration:
    - Pecount = 288, ompthreads = 3 => 8 io processors
  - IO format:
    - Pnetcdf: parallel NetCDF in 64bit. ← default in CESM
    - NetCDF4c: NetCDF4 with serial compression
    - NetCDF4p: NetCDF4 with parallel
    - NetCDF4pc: NetCDF4 with parallel compression
    - Async NetCDF4pc: NetCDF4 with parallel compression in asynchronous mode

Enabled Parallel Compression in CESM/PIO

Grid	Pnetcdf	Netcdf4c	Netcdf4p	Netcdf4pc	Netcdf4pc	Async
f09_f09 (total/history io)(sec)	560/2.78	577/10.36	600/12.68	609/23.98	590.7/13.09	584.7/14.78
f09_f09 (io percent)	0.5%	1.8%	2.1%	3.9%	2.2%	2.5%
Total increased		3.0%	7.1%	8.8%	5.4%	4.3%

Enabled Async Mode in CESM/PIO

Grid	Pnetcdf	Netcdf4c	Netcdf4p	Netcdf4pc	Netcdf4pc	Async
f09_f09 (total/history io)(sec)	560/2.78	577/10.36	600/12.68	609/23.98	590.7/13.09	584.7/14.78
f09_f09 (io percent)	0.5%	1.8%	2.1%	3.9%	2.2%	2.5%
Total increased		3.0%	7.1%	8.8%	5.4%	4.3%

Evaluate Cloud-Friendly Data Format

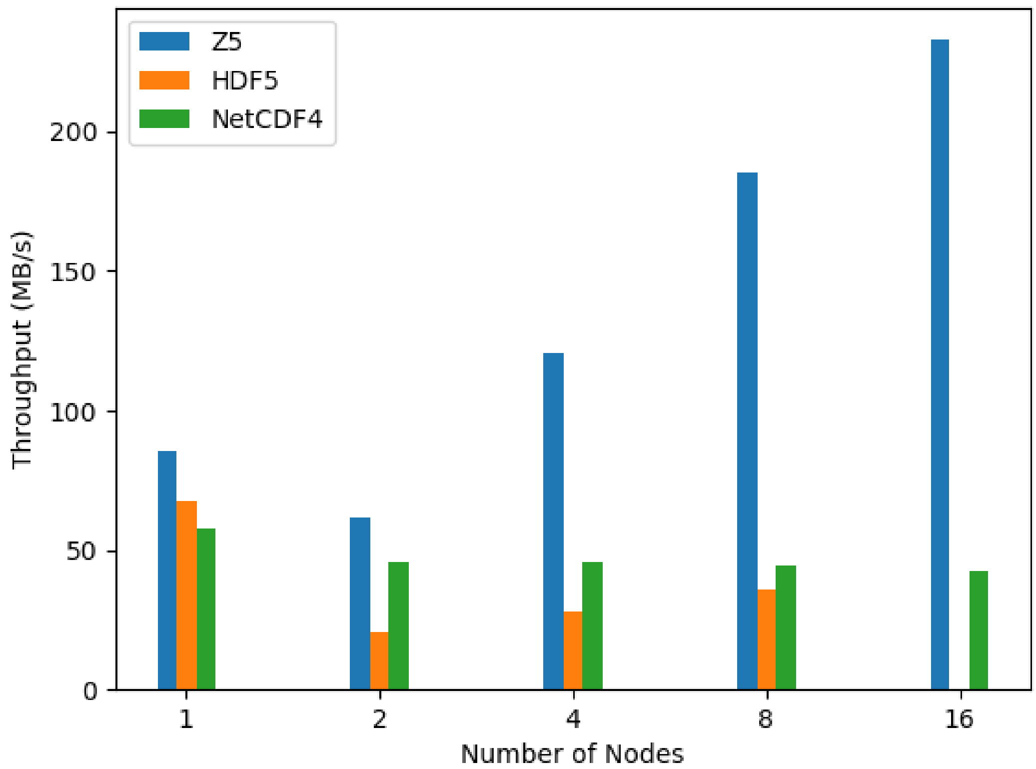
- Zarr/Z5
  - Provides Python/C++ classes and functions for N-dimensional array
  - Writes out array data in chunks and each chunk is compressed
  - Convert to/from NetCDF files easily
  - Uses distributed storage systems: S3Map, HDFSMap, GCSMap
- Working on a SIParCS project with Weile to enable Z5 backend in PIO

IOR Benchmark

- IOR: Interleaved or Random
  - File system benchmarking tool
  - Well-suited for evaluating the performance of parallel system
  - Have several IO backend: POSIX, MPIIO, NCMPI, HDF5, S3
- Purpose
  - IOR has so many IO backend
  - Can compare performance among different IO backends
- Prepare for IOR benchmarking
  - Added c wrapper of Z5
  - Integrated in IOR benchmark tool

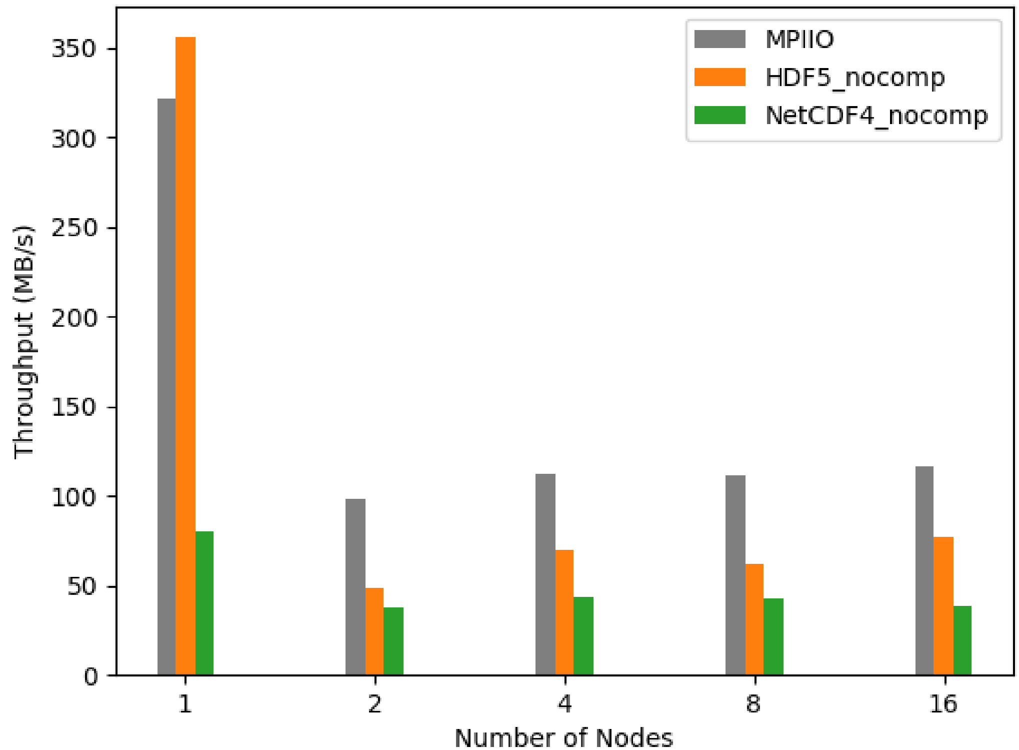
IOR Benchmark
Write 16 Segments of 256KB Blocks

- One processor per node



MPIIO Collective Write On Multiple Nodes

- Write throughput decreased to one tenth
- One processor per node



Conclusion on IO with Compression

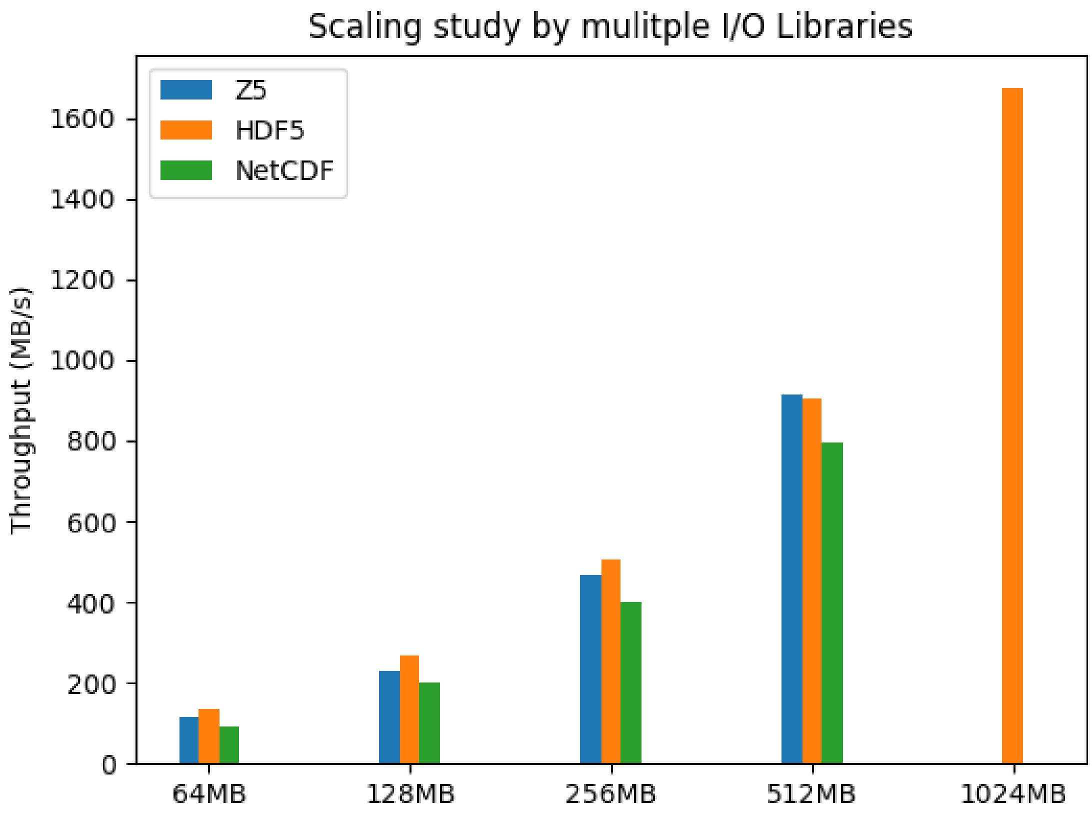
- Write out from one node instead of multiple nodes
- Evaluate Z5 backend
  - Flexible chunking, compression and metadata management
  - Both pthread and MPI
  - Better with cloud storage
- Enable async mode with large buffers

- Potentially get a compression IO with good performance

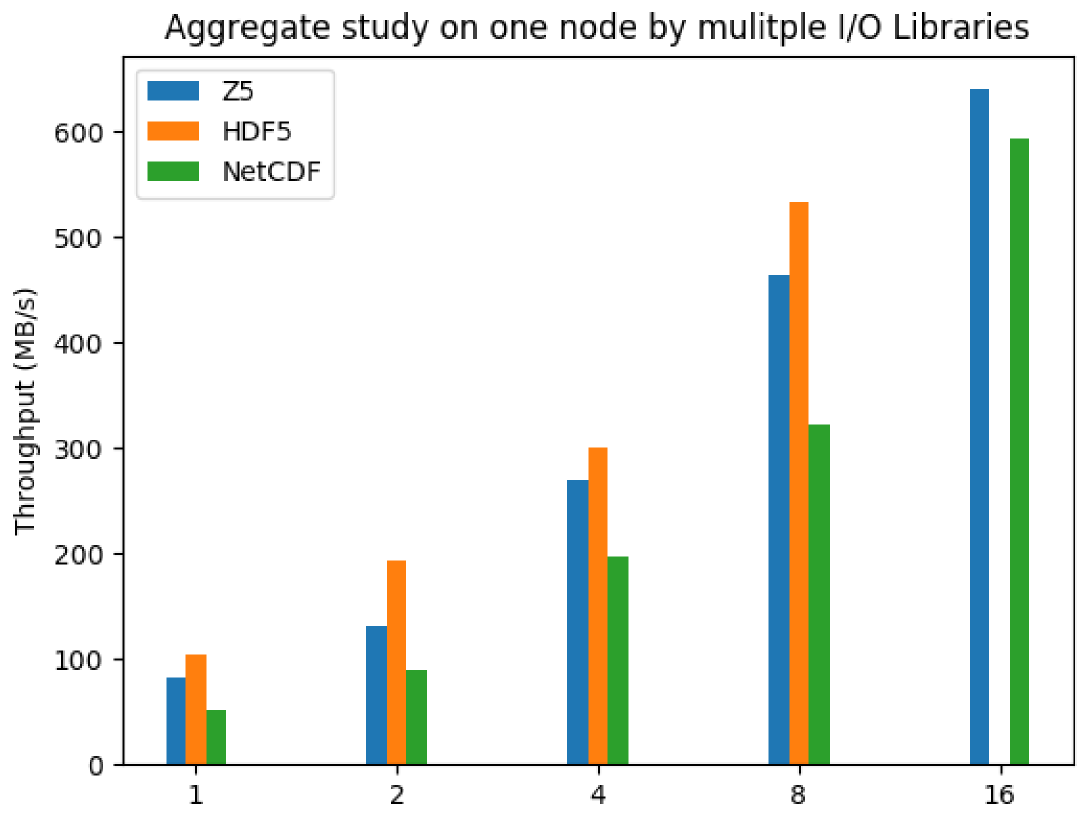
Question?

- Contact me at haiyingx@ucar.edu

IOR Benchmark Scaling Study



IOR Benchmark Aggregate Study on 1 Node



NCAR
UCAR

XIOS Server

- XML-IO-SERVER
  - Two ways: attached mode, and server mode
  - Accept one-to-one file mode and multiple-to-one file mode
  - Use double buffers on client side, and circular buffer on server side
  - Very good performance
    - 1.5% IO for daily mean output (4322x2882x31, 8160 cores, 32 XIOS)
    - 5% IO for 6 hours mean output
    - 15%-20% for hourly mean output (128 XIOS)

  - No API for inquiring dimension, variable id, etc.
  - Need to make sure all communications overlapped by computation
  - All writing overlapped by computation
  - Otherwise, blocking time will take longer

IOR Benchmark Conclusion

- Z5, NetCDF, HDF5 with compression
  - same performance when write once from multiple processors
  - Z5 and NetCDF need more memories
  - Z5 and HDF5 are better when write aggregately in one node
  - Z5 is five times better when write aggregately from multiple nodes
- IO processors of CESM/PIO now is distributed on processors on multiple nodes
- Z5 has the advantages on chunking, compression and metadata