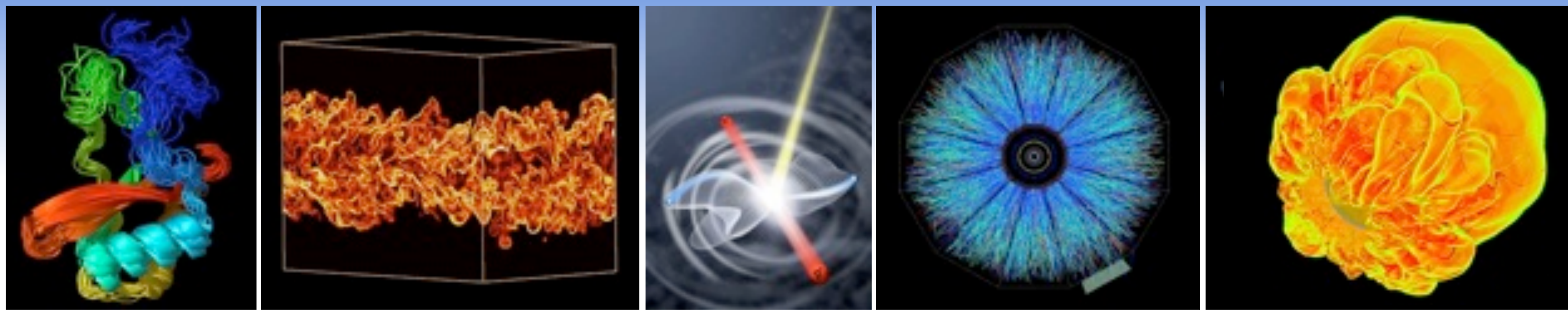


Tuning HDF5 for Lustre

John Shalf, Mark Howison
LBNL/NERSC
jshalf@lbl.gov , mhowison@lbl.gov

Quincey Koziol
HDF Group
koziol@hdfgroup.org

SuperComputing 2009
HDF5 BoF

Summary of HDF5 Extreme Scale I/O Effort

- **HDF5 is the most commonly used parallel I/O library in both DOE SC and DOE SciDAC applications**
  - 3rd most popular library according to NERSC ERCAP (MPI and ScaLAPACK are #1 and #2)
  - Consistently most popular in SciDAC survey
- **HDF5 performance has been declining on recent systems**
  - Corresponds to decline of investment in HDF Group for performance tuning
  - Formerly central to DOE ASCI program
- **NERSC workshop in June 2009 to assess HDF5 performance issues**
  - Meeting brought together DOE SC applications scientists, Cray Developers, MPI-IO developers
  - Developed strategy for Performance tuning HDF5
- **NERSC funded pilot effort on HDF5 performance tuning**
  - 50% FTE at HDF Group and 50% at NERSC
  - Demonstrated 8x-10x improvement and scaling to 32,000 processors

Benchmarking I/O kernels

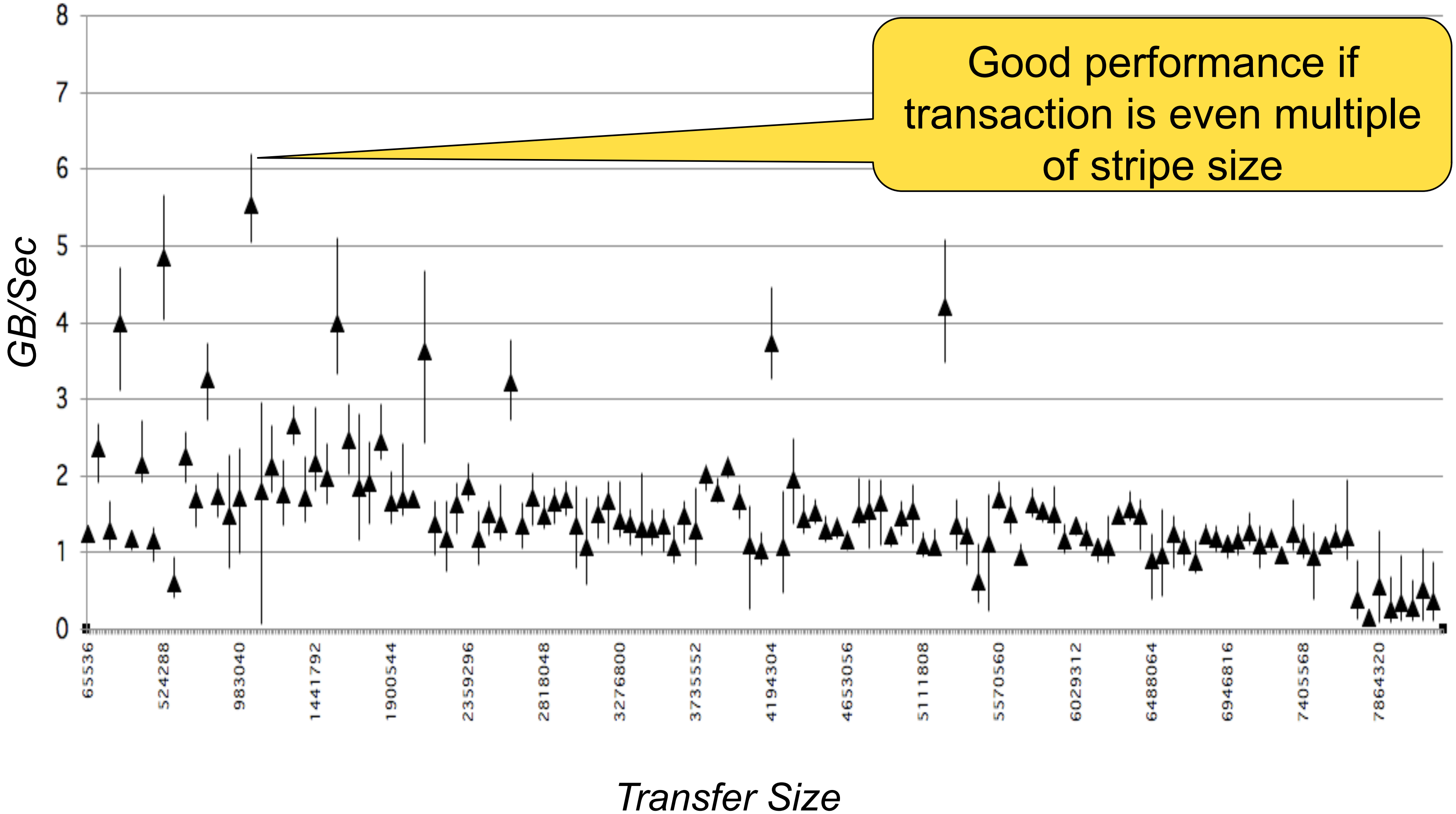
- **GCRM** (regular 1D/2D/3D)
  - Global Cloud Resolving Model
  - David Randall Group Icosahedral model from Colorado State University
- **Chombo** (irregular 1D)
  - AMR framework and SciDAC application
  - Phil Collela's APDEC group at LBNL
- **VORPAL** (irregular 3D + irregular 1D)
  - Particle-in-Cell: Fusion and Accelerator Modeling
  - Particles OK, but 1D
  - Tech X Corporation and SciDAC COMPASS

Optimizations

- **Lustre**
  - select correct stripe count
  - align I/O operations to stripe boundaries
- **MPI-IO**
  - improve collective buffering (2-phase) performance
- **HDF5**
  - remove serialization points (e.g. ftruncate)
  - aggregate small operations (e.g. metadata)
  - linearize data with chunking

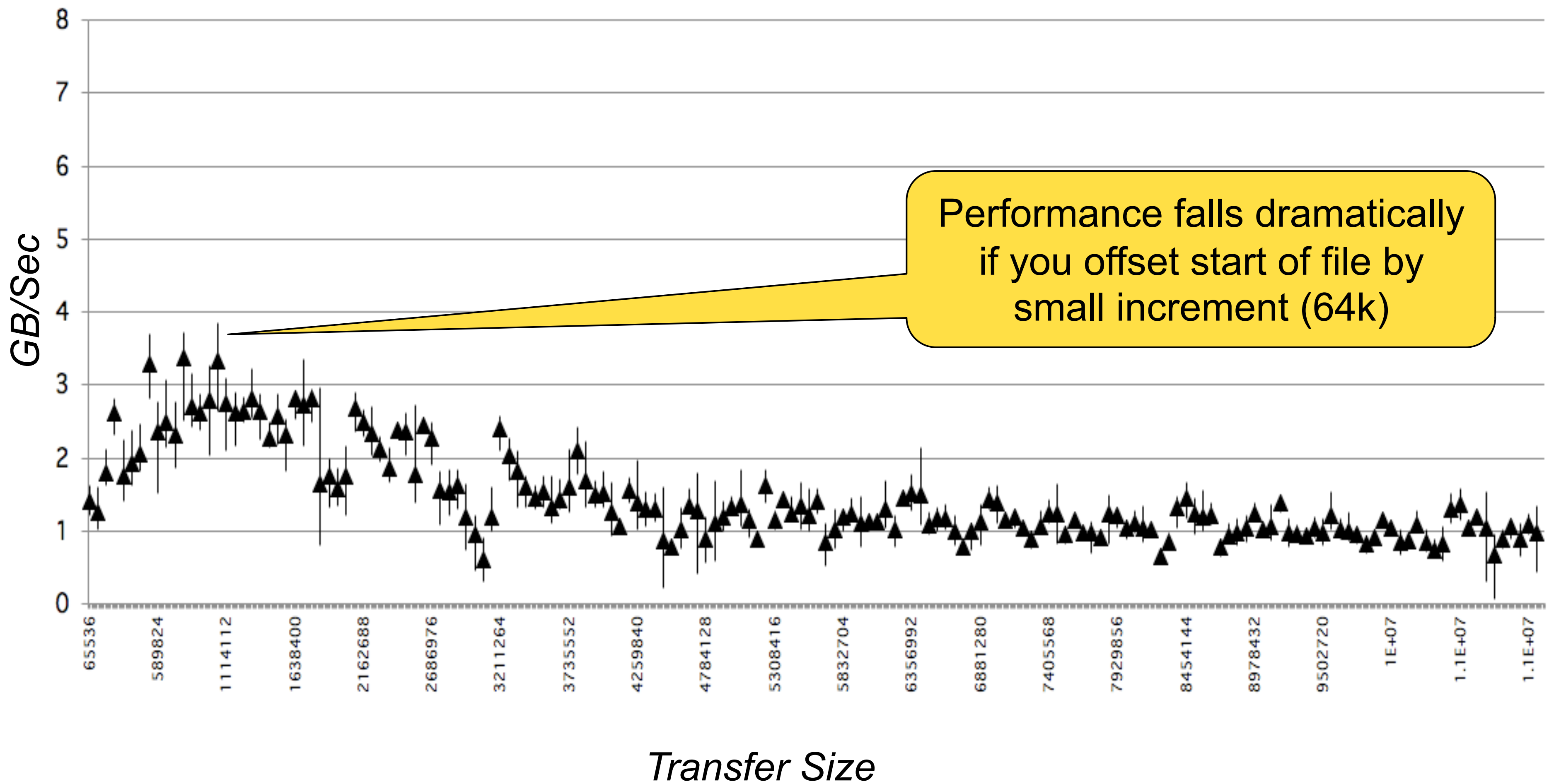
I/O Performance Sensitivity to Transfer Size

2GB File Size, 80 Processors, 40 OSTs

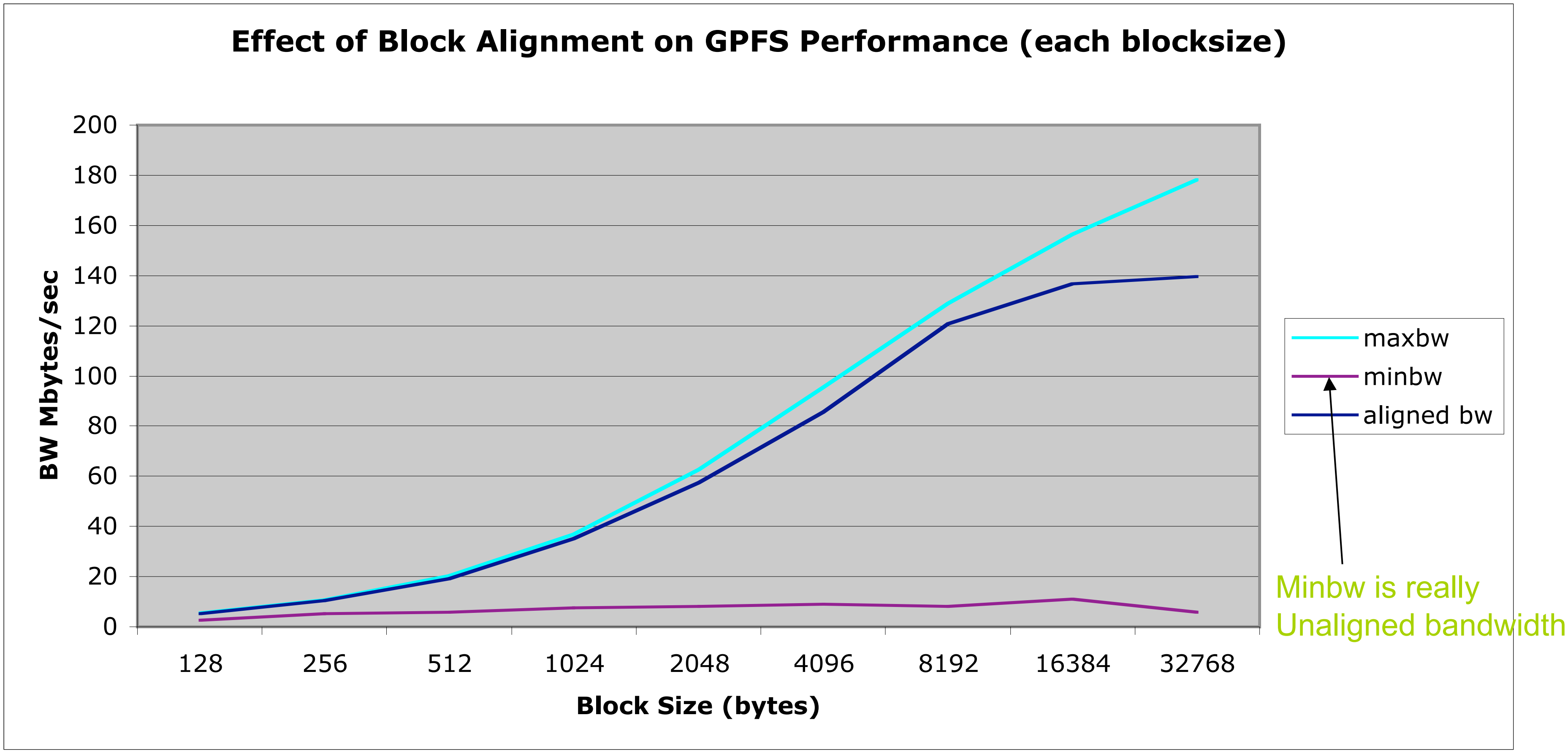


I/O Performance Sensitivity to Transfer Size

2GB File Size, 80 Processors 40 OSTs: Offset file start by 64k



Streaming Unaligned Accesses
(not to pick on Lustre… GPFS suffers too)



IPM I/O Profile of GCRM

Baseline

Reduce writers
(2-phase I/O)

Stripe Alignment
& Chunking

Aggregate
Metadata

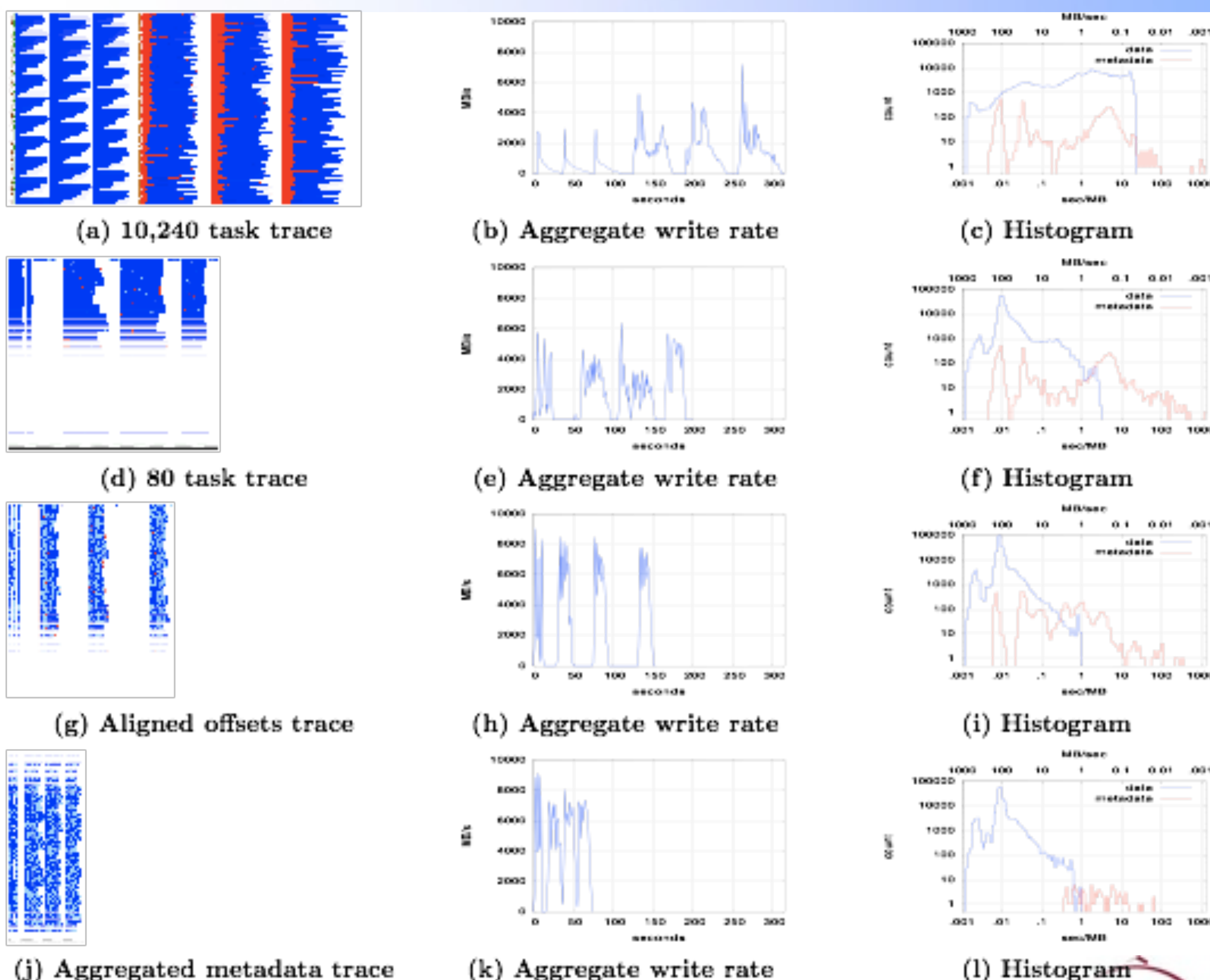
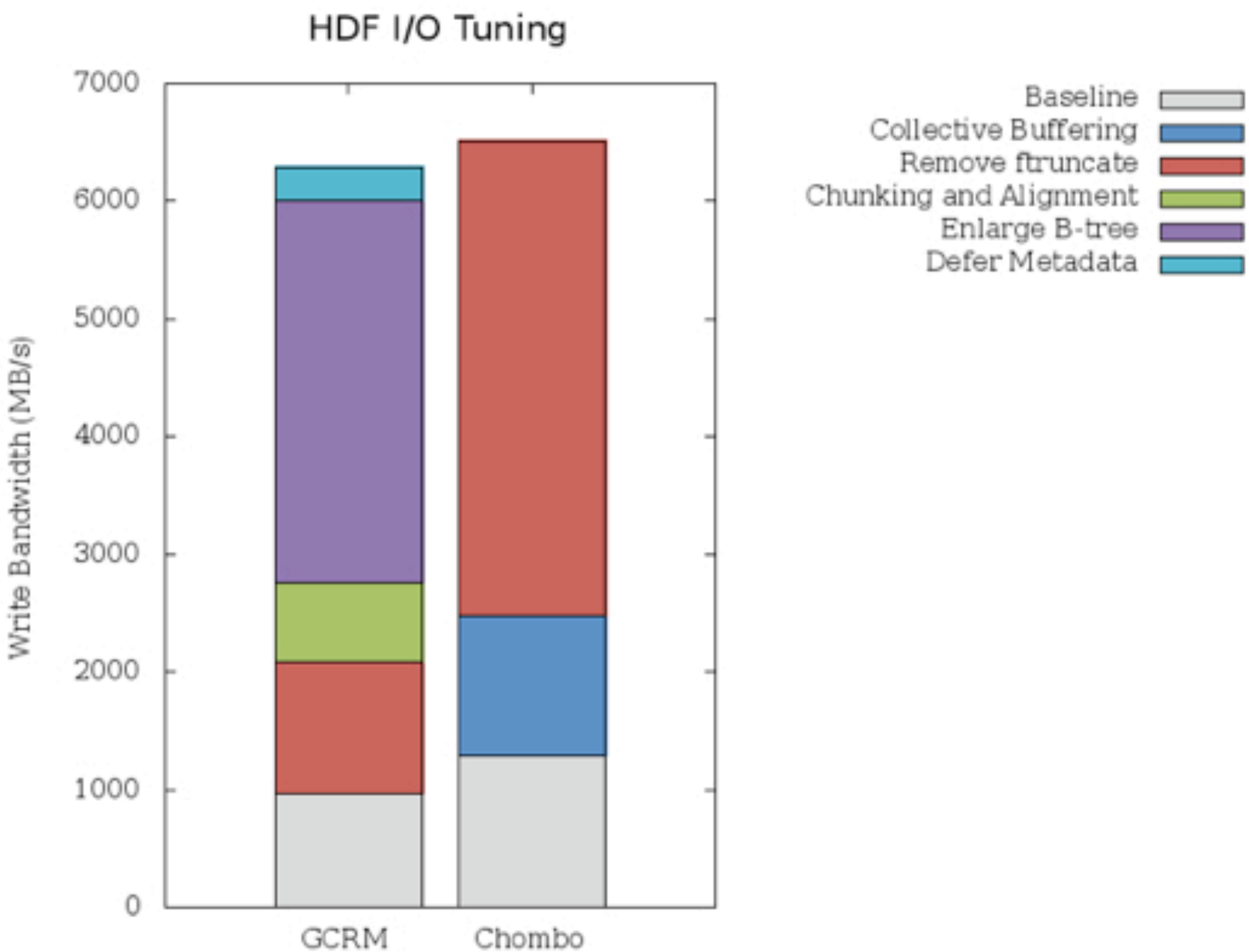
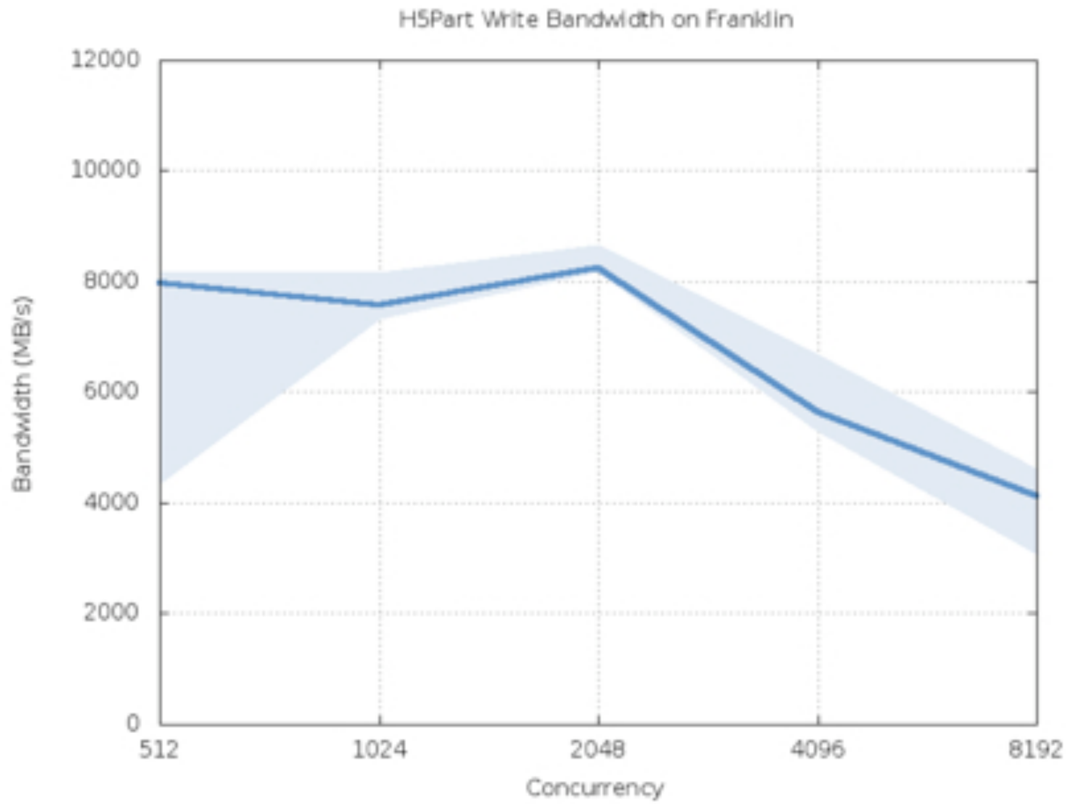


Figure 1: Trace graphs, aggregate write rates, and histograms for the GCRM I/O kernel with a baseline configuration and three progressive optimizations.

GCRM and Chombo Benchmarks



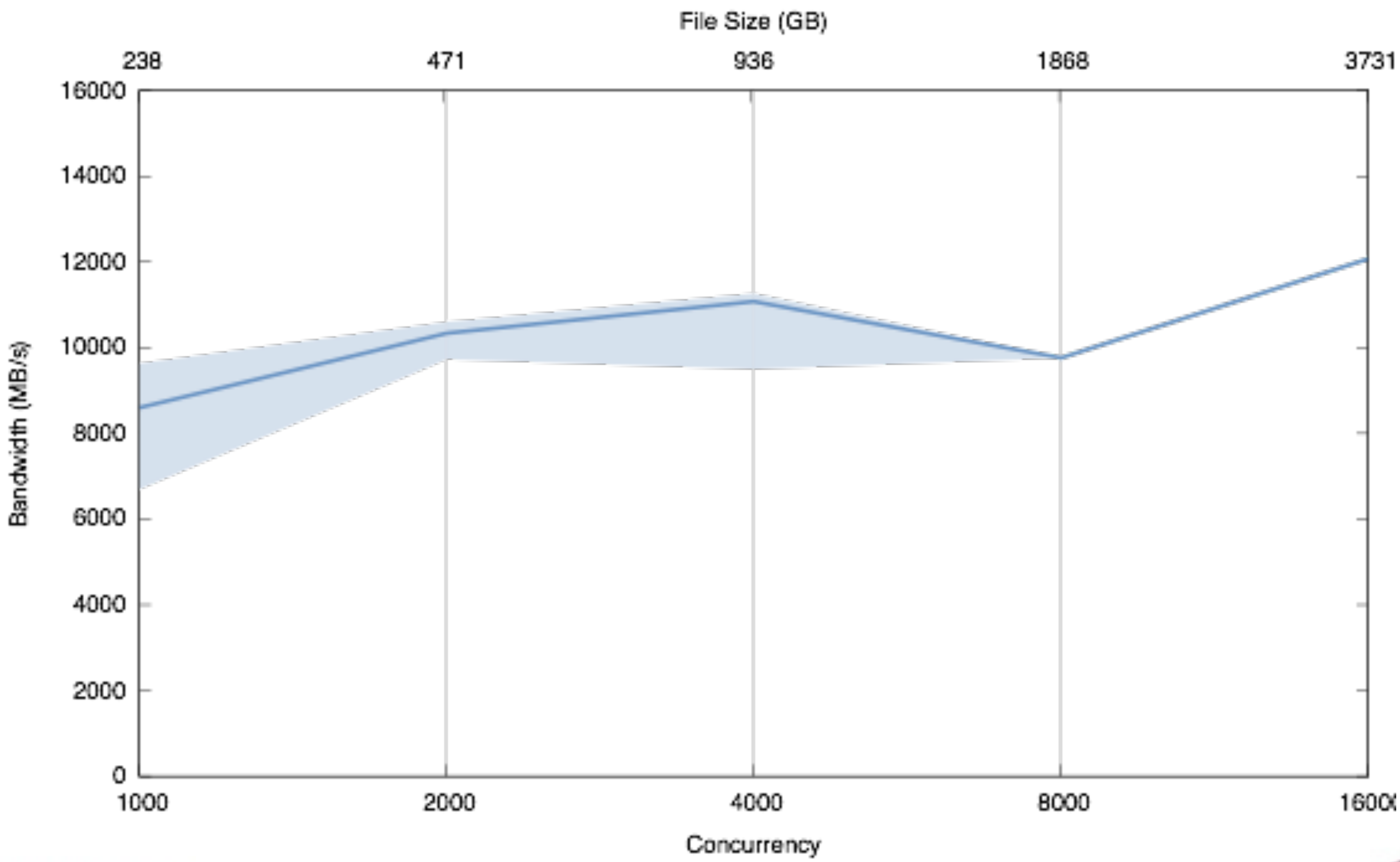
Strong Scaling



Particles (float) per core	MB per core	Timesteps	Cores	GB
20,000,000	76.29	16	512	610.35
10,000,000	38.15	16	1024	610.35
5,000,000	19.07	16	2048	610.35
2,500,000	9.54	16	4096	610.35
1,250,000	4.77	16	8192	610.35

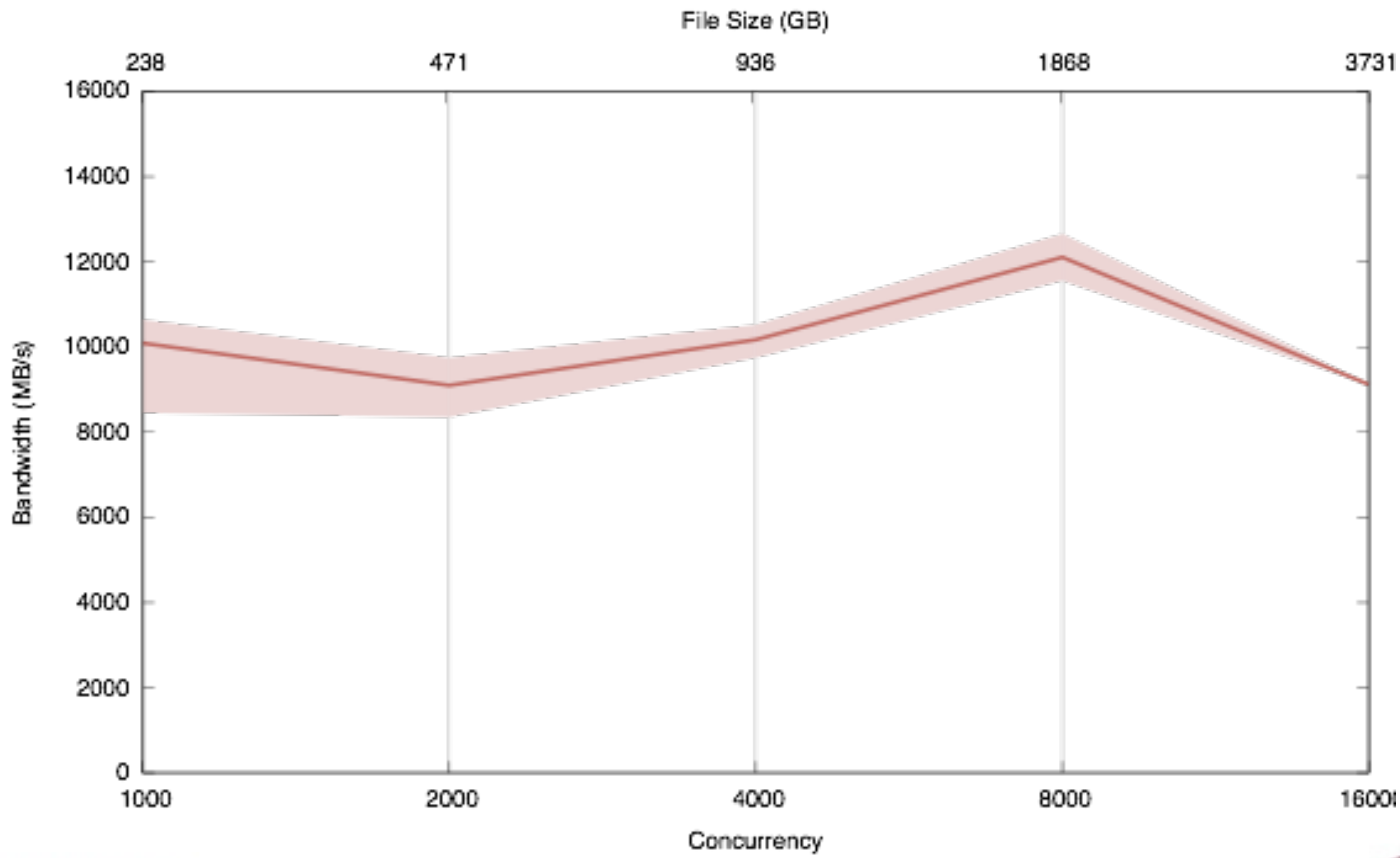
Weak Scaling

Write with MPI-POSIX on Franklin (scratch2)



Weak Scaling

Read with MPI-POSIX (plus halo exchange via MPI) on Franklin (scratch2)



Whats Next?

- Automatic Tuning for Lustre
  - First expose tunable parameters to expert users
  - Then use tunable parameter interfaces to introspect filesystem configuration to tune automatically
- Working on multi-lab whitepaper to sustain support for HPC-class HDF5
  - LLNL, LBNL, HDF-Group, ….

Acknowlegements

- LBL/NERSC
  - Prabhat
  - Wes Bethel
  - Andrew Uselton
  - Noel Keen
  - Hongzhang Shan
  - Katie Antypas
  - Shane Canon
  - David Skinner
  - Nick Wright
- HDF Group
  - John Mainzer
- Cray
  - David Knaak
- ANL
  - Rob Latham
  - Rob Ross
- Lustre Center For Excellence at ORNL