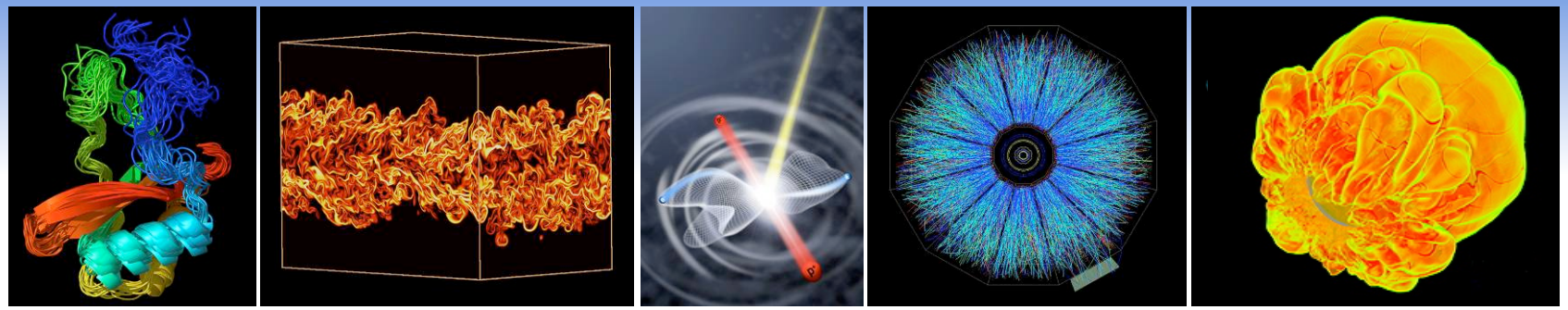


I/O Using H5Part

Mark Howison
NERSC Analytics / SDSA / Visualizaton Group
mhowison@lbl.gov

NERSC User Group Meeting
October 8, 2009

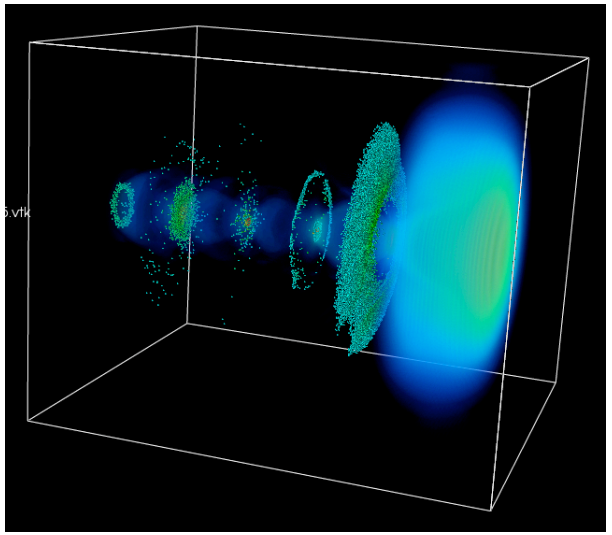
U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

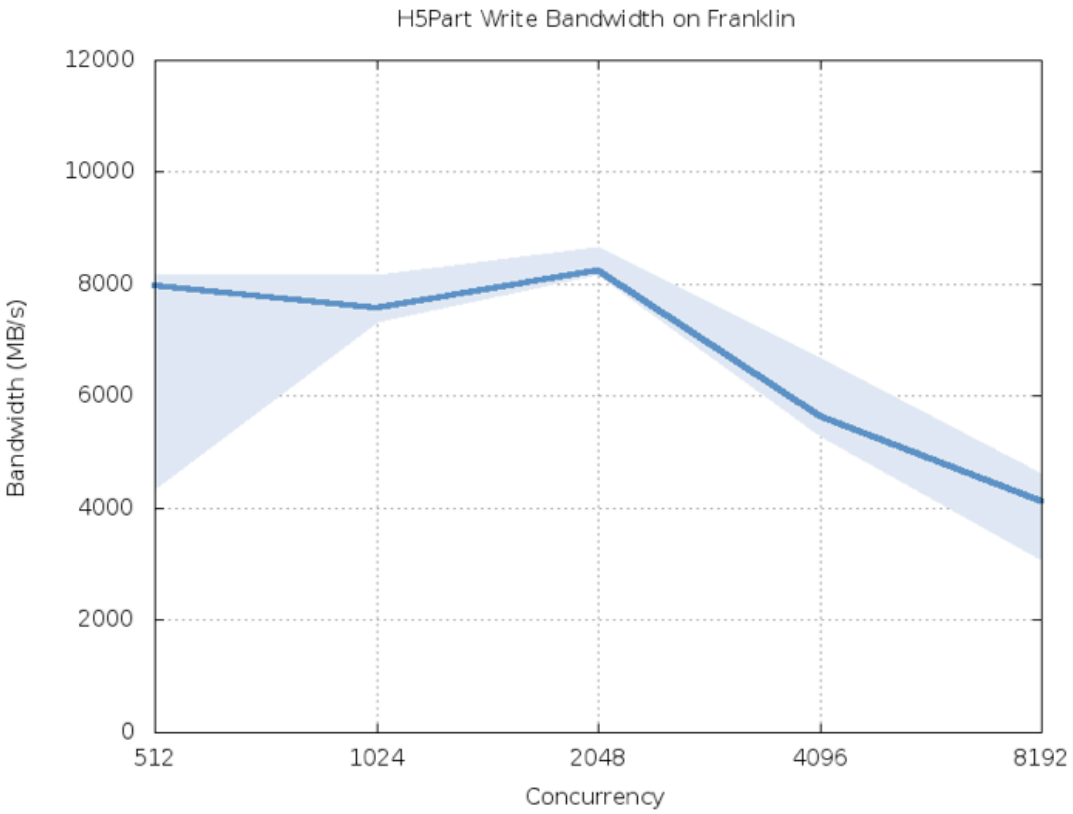
Overview of H5Part

- A simplified 'veneer' API for HDF5
  - C/C++ and Fortran bindings
  - Datasets are stored across 'timesteps'
  - Scalable parallel I/O to single shared files



- Supports 3 basic data models:
  - H5Part for regular 1D arrays, e.g. particles
  - H5Block for irregular 3D grids, e.g. fields
  - H5MultiBlock for regular 3D grids partitioned for distributed memory with halo regions
  - (A fourth component for irregular meshes is under development at PSI)

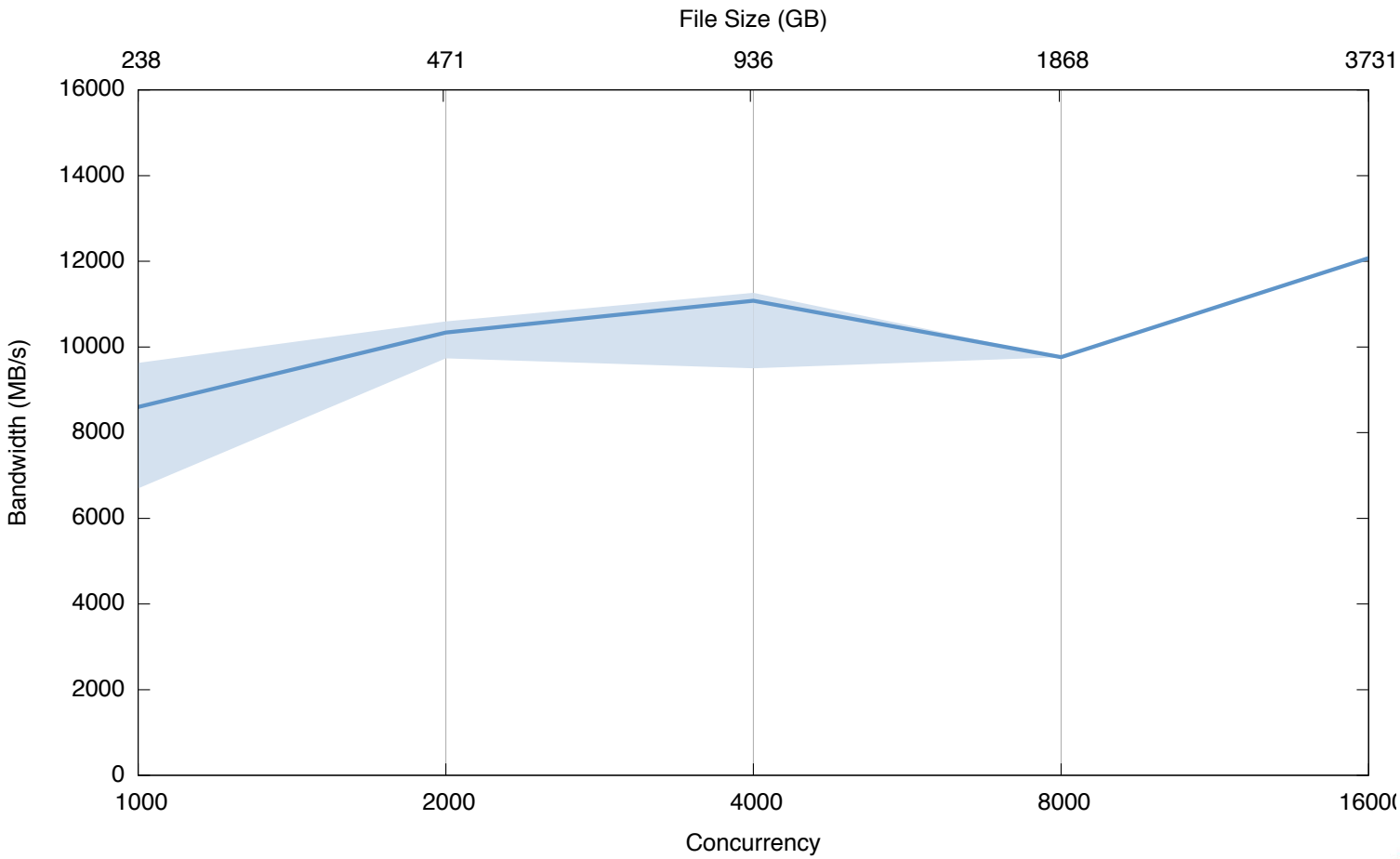
Strong Scaling of H5Part



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

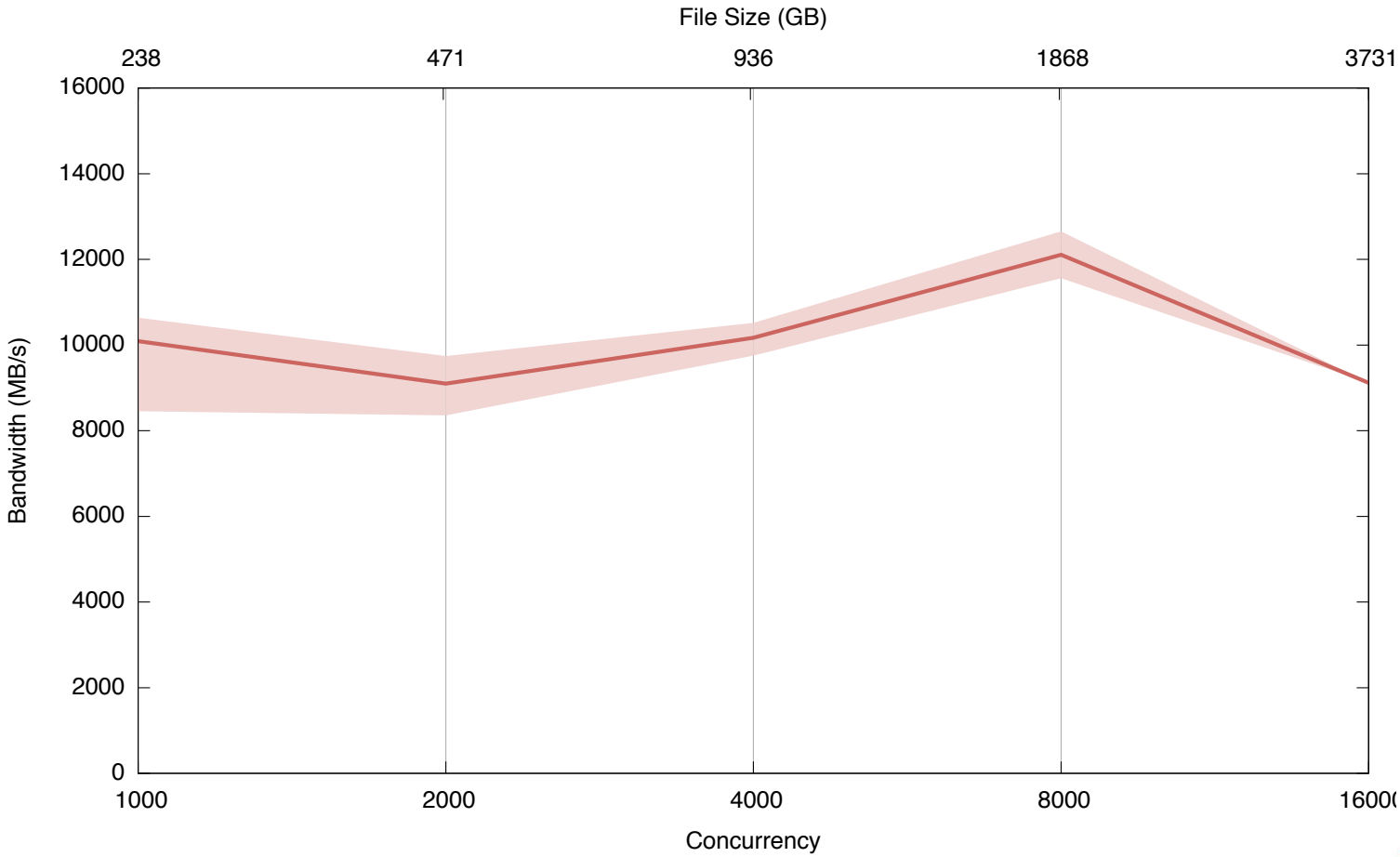
Weak Scaling of H5MultiBlock

Write with MPI-POSIX on Franklin (scratch2)



Weak Scaling of H5MultiBlock

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



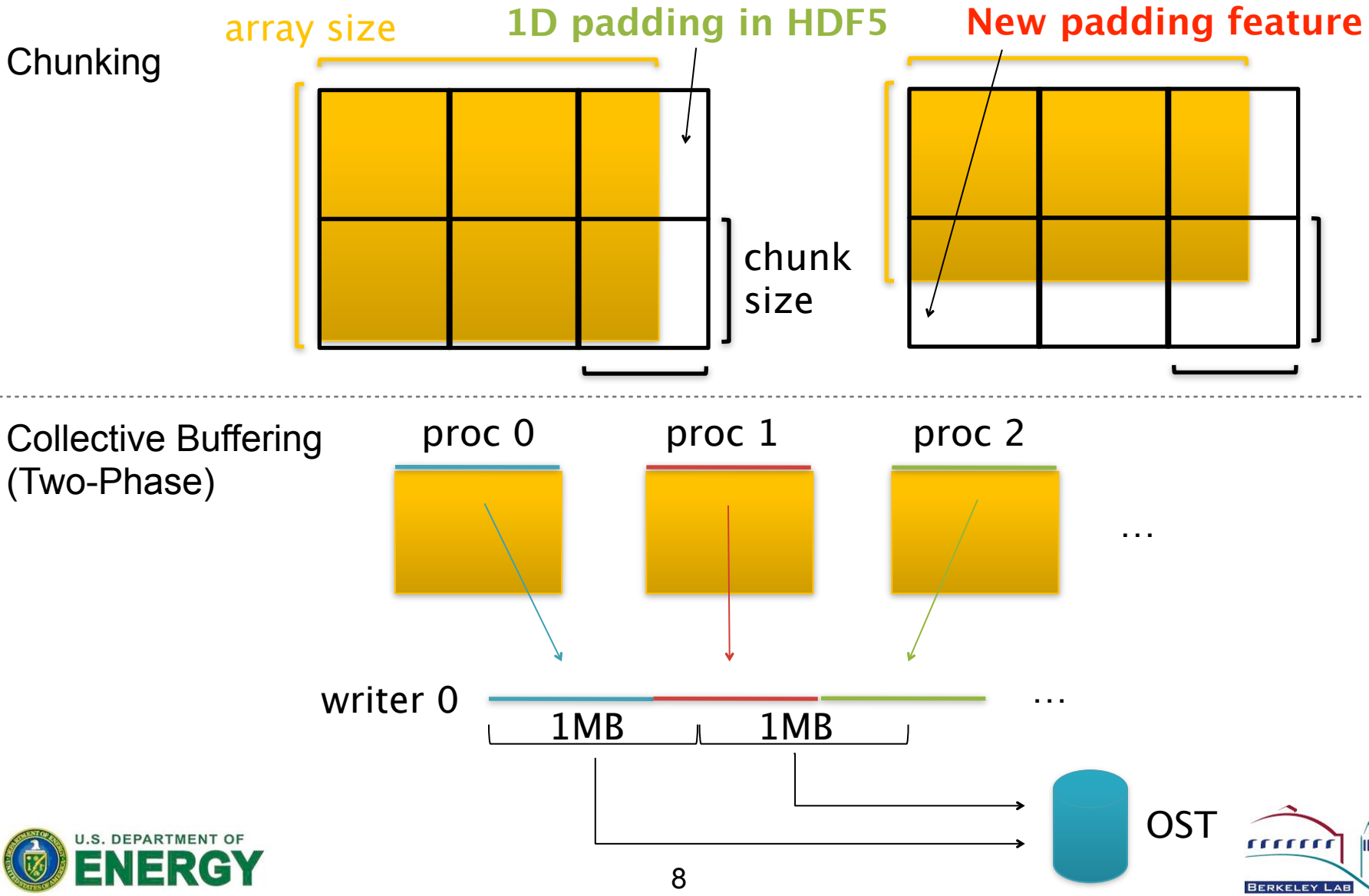
NERSC / HDF Group

- NERSC staff are collaborating with the HDF group to tune HDF5 for Franklin and the lustre filesystem (e.g. Hopper)
  - Optimizations will be merged into future public releases of HDF5
- Tuning I/O kernels for three apps:
  - GCRM (regular 1D/2D/3D)
  - Chombo (irregular 1D)
  - VORPAL (irregular 3D)

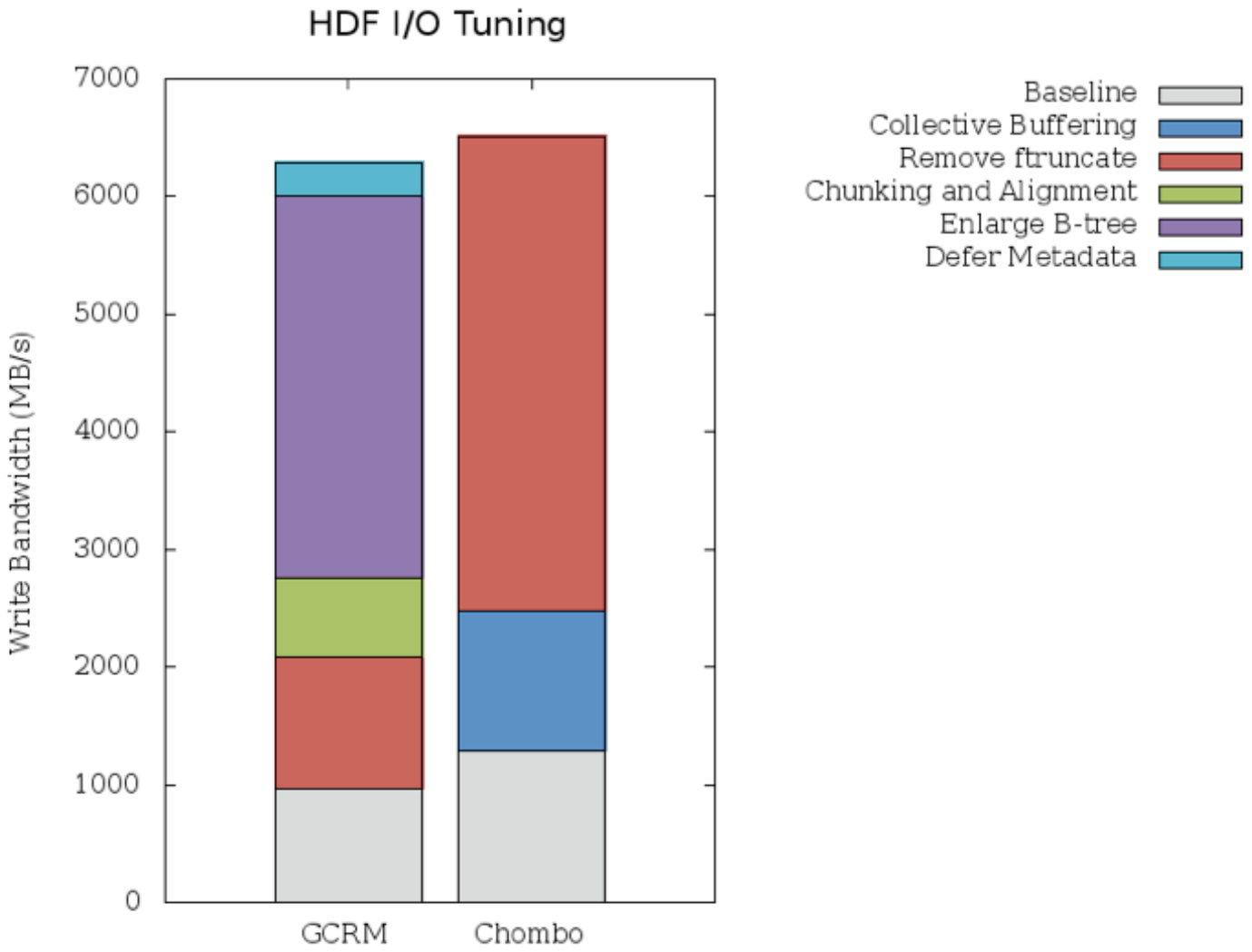
Optimizations

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

Chunking vs. Collective Buffering

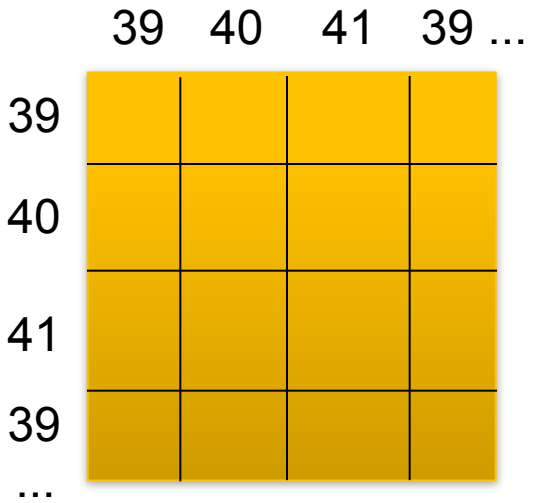


GCRM and Chombo Benchmarks



VORPAL Benchmark

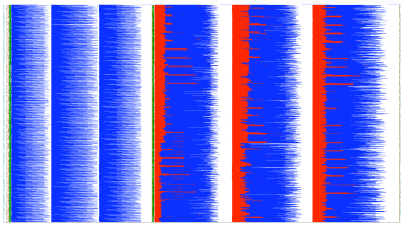
- **Irregular blocks:**
  - **dimensions differ slightly between adjacent grid cells**



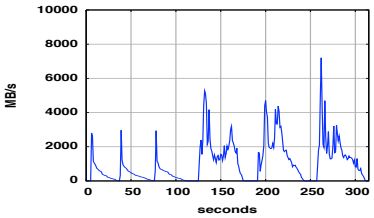
- **Initial results show problems with MPI communication phase**

Cores	Block	File Size	Collective	Min (MB/s)	Max (MB/s)
512	40x40x40x3	1.9GB	No		< 2
512	40x40x40x3	1.9GB	Yes	314.82	454.35
2048	40x40x40x3	30GB	Yes	486.97	1,816.57
2048	80x80x80x3	240GB	Yes	2,419.53	2,455.46

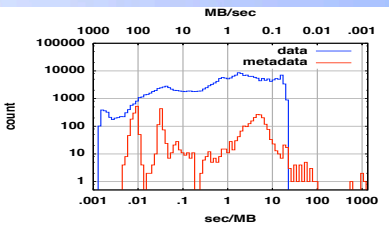
I/O Profiling with IPM



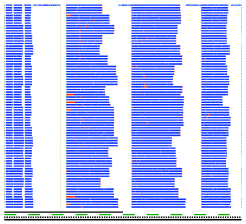
(a) 10,240 task trace



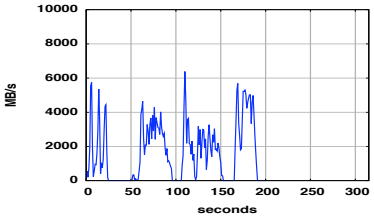
(b) Aggregate write rate



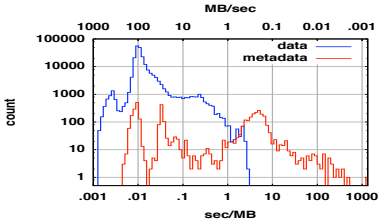
(c) Histogram



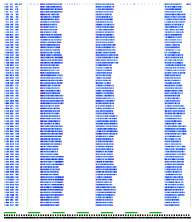
(d) 80 task trace



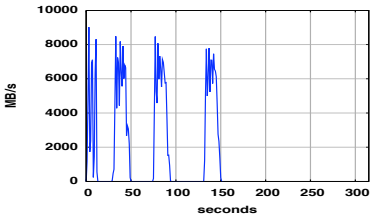
(e) Aggregate write rate



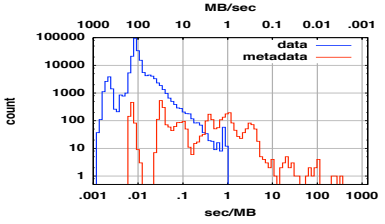
(f) Histogram



(g) Aligned offsets trace

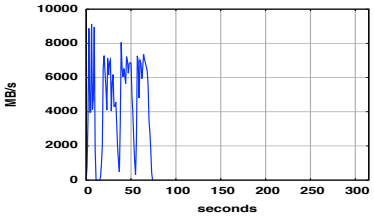


(h) Aggregate write rate

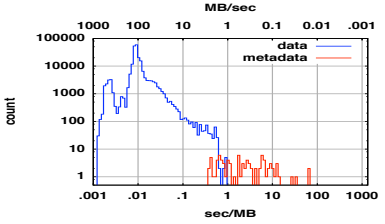


(i) Histogram

(j) Aggregated metadata trace



(k) Aggregate write rate



(l) Histogram

Acknowledgments

- LBL/NERSC
  - John Shalf
  - Prabhat
  - Wes Bethel
  - Andrew Uselton
  - Noel Keen
  - Hongzhang Shan
  - Katie Antypas
  - Shane Canon
  - David Skinner
  - Nick Wright
- HDF Group
  - Quincey Koziol
  - John Mainzer
- Cray
  - David Knaak
- PSI
  - Andreas Adelmann
  - Achim Gsell

Additional Resources

- On Franklin: `module help h5part`
- http://www.nersc.gov/nusers/resources/software/libs/io/h5part/
- http://vis.lbl.gov/Research/H5Part
- https://lists.web.psi.ch/mailman/listinfo/h5part

Demonstration on Franklin

- Writing a simple H5Part app...

- Loading H5Part data into VisIt...