Parallel IO Library Benchmarking on GPFS

NERSC

Lawrence Berkeley National Lab

Katie Antypas

July 18, 2007

NERSC is supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract number DE-AC02-05CH11231.

Acknowledgements

Many thanks to Hongzhang Shan, John Shalf and Jay Srinivasan for input and feedback

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Motivation

- With simulations creating larger and larger data files we are approaching the point where a naive one file per processor IO approach is no longer feasible
  - Difficult for post processing
  - Not portable
  - Many small files, bad for storage systems
- Parallel IO approaches to a single file offer an alternative, but do have an overhead

Objective

- Explore overhead from Parallel IO libraries HDF5 and Parallel NetCDF compared to one file per processor IO
- Examine effects of GPFS file hints on application performance.

Outline

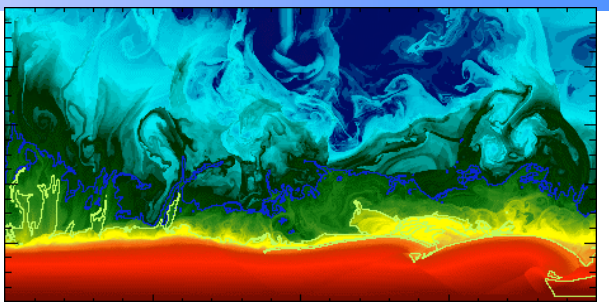
- Benchmark Application Methodology
- Bassi Benchmarking
  - IO File System Hints
  - IO Library Comparison
    - HDF5
    - Parallel NetCDF
    - Fortran one-file-per-processor
  - Overhead Compared to direct IO
- Jacquard Benchmarking
  - IO Library Comparison
  - Overhead compared to direct IO

Bassi and Jacquard Details

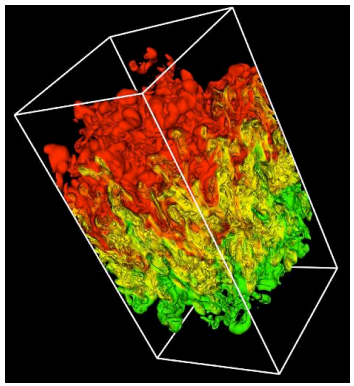
Machine	Vendor	Proc Arch	Total Procs	File system	Interconnect	Peak IO Bandwidth
Bassi	IBM	Power 5	888 (111 8 proc nodes)	GPFS	Federation	~ 8GB/sec 6 VSD * ~1-2GB/sec
Jacquard	Linux Networks	Opteron	712 (356 2 proc nodes)	GPFS	Infiniband	~3.1GB/sec 5 DDN couplets * 620 MB/sec

FLASH3 IO Benchmark

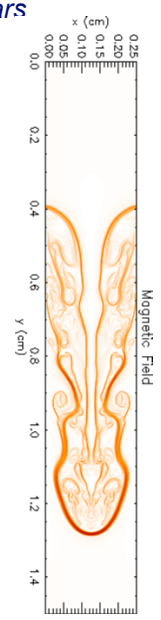
- IO from FLASH3 code
  - Astrophysics code designed primarily for compressible reactive flow
  - Parallel, scales well to thousands of processors
  - Typical IO pattern of many large physics codes
  - Writes large contiguous chunks of grid data
- Multiple IO output formats
  - Parallel IO libraries built on top of MPI-IO
    - HDF5 to a single file
    - Parallel-NetCDF to a single file
  - One file per processor Fortran unformatted write
- Similar data format to IOR benchmark



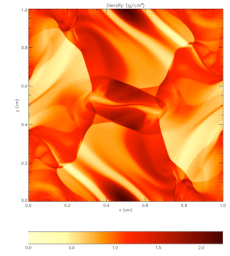
Helium burning on neutron stars



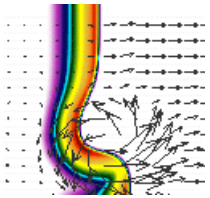
Rayleigh-Taylor instability



Magnetic Rayleigh-Taylor



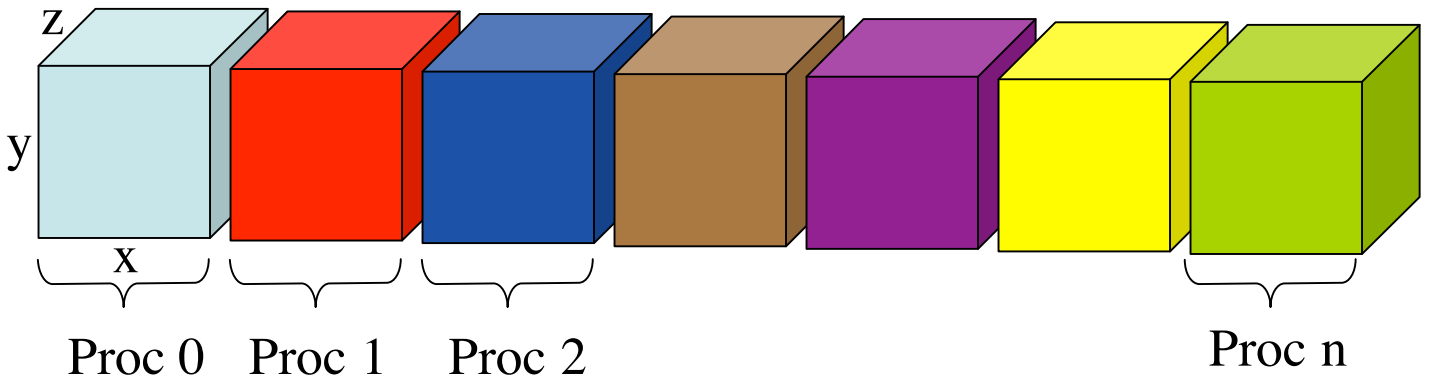
Orzag/Tang MHD vortex



Flame-vortex interactions

Modified FLASH3 IO Benchmark

- Only writing out 4 dimensional datasets x,y,z,proc
- Layout of data is contiguous
- 2 experiments weak scaling IO
  - Each processors writes 64 MB of data
    - Fits in block buffer cache
    - 96x * 96y * 96z * 8bytes * 9vars = 64MB
  - Each processor writes 576 MB of data
    - Above 256 MB/proc no longer see caching effect
    - 200x * 200y * 200z * 8 bytes * 9 vars = 576 MB

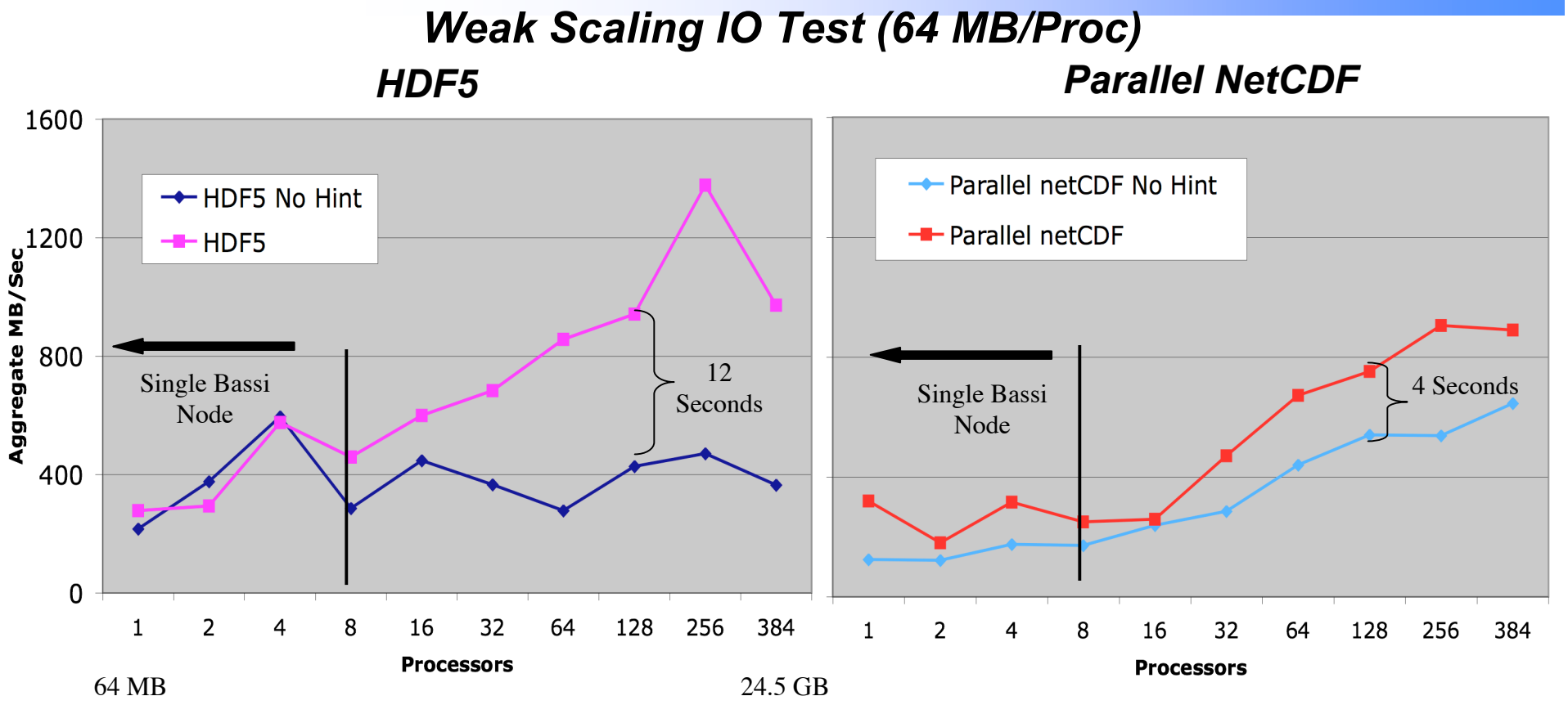


MPI-IO/GPFS File Hints

- MPI-IO allows the user to pass file hints to optimize performance
- IBM has taken this a step farther by implementing additional file hints in their implementation of MPI-IO to take advantage of GPFS features
- File hint IBM_largeblock_io = true
  - Disables data shipping
  - Data shipping used to prevent multiple MPI tasks from accessing conflicting GPFS file blocks
  - Each GPFS file block bound to single task
  - Aggregates small write calls
  - Data shipping disabled saves overhead on MPI-IO data shipping but only recommended when writing/reading large contiguous IO chunks.

```
MPI_Info_set(FILE_INFO_TEMPLATE, "IBM_largeblock_io", "true");
```

Effects of IBM_largeblock_io = true on Bassi



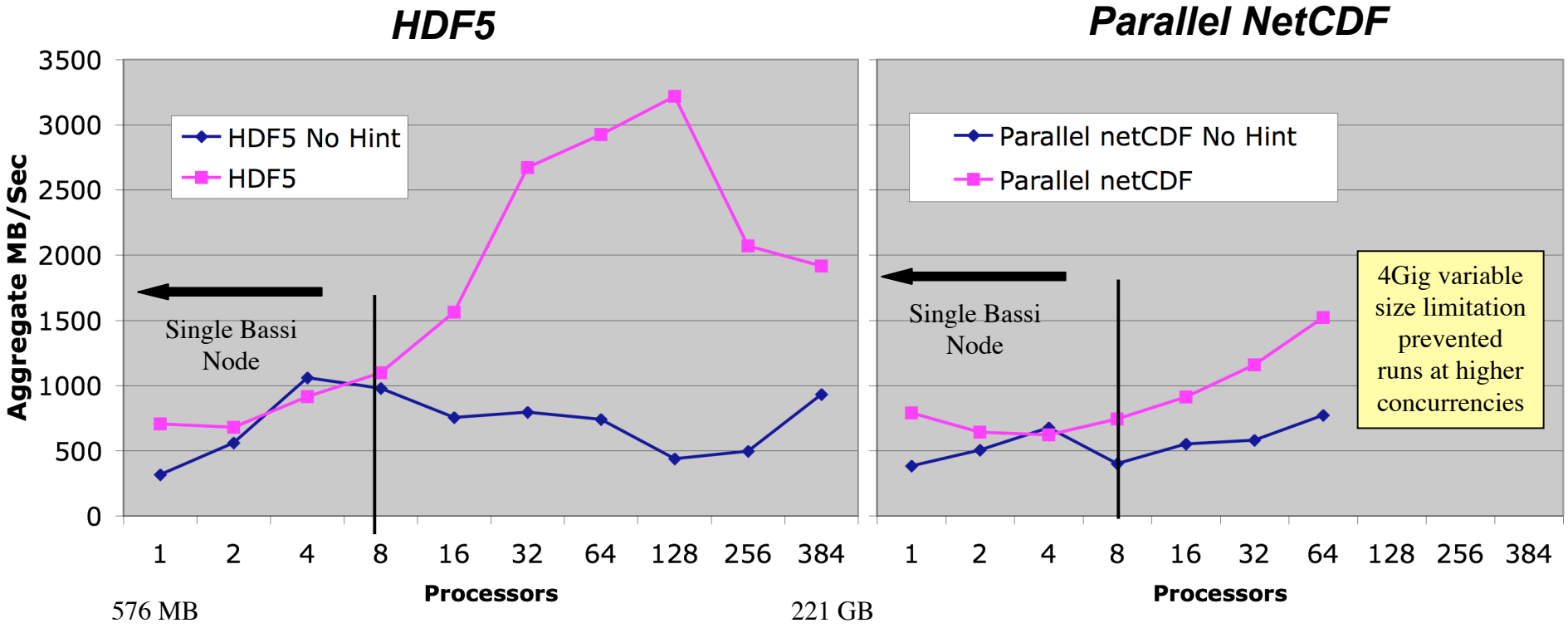
On average, for runs on more than 8 processors HDF5 received a 135% performance increase compared with a 45% improvement for Parallel NetCDF

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Effects of IBM_largeblock_io = true

Weak Scaling IO Test (576 MB/Proc)



The effects of the file hint are more significant for both HDF5 and Parallel NetCDF at larger file sizes. And while the rates for HDF5 and Parallel NetCDF are similar without the file hint, the effects of data shipping turned off is again much larger with HDF5

Office of Science U.S. Department of Energy

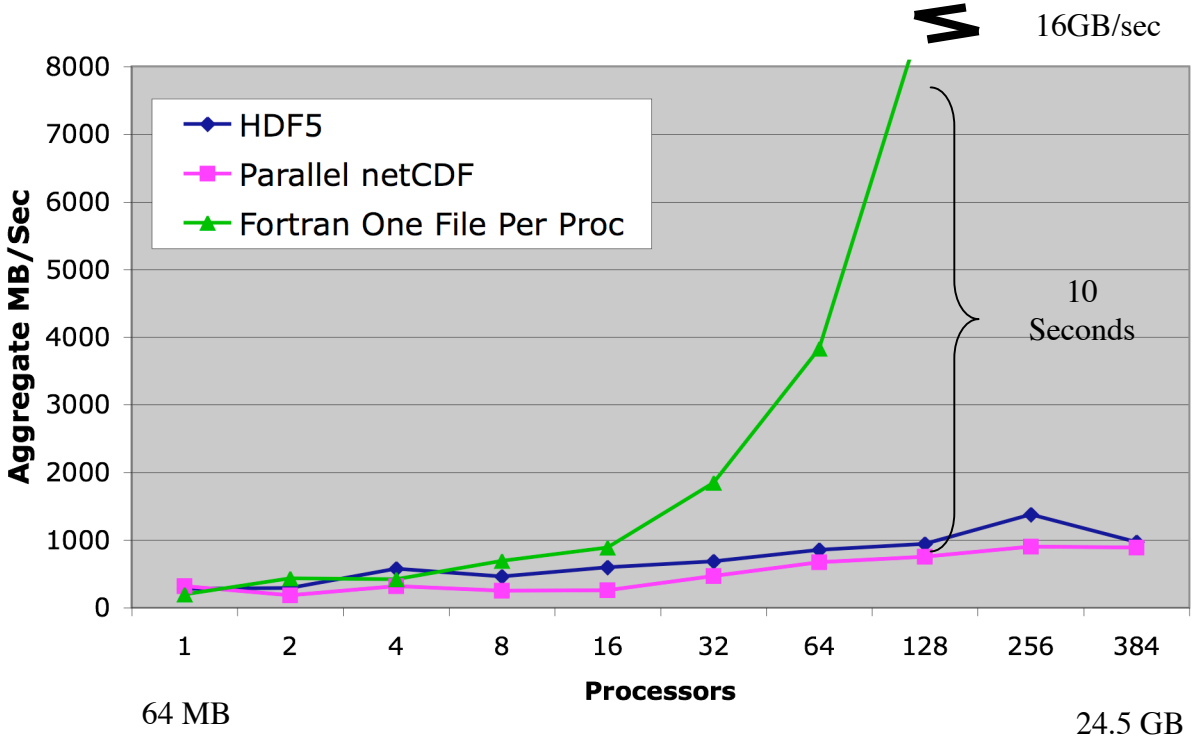
Berkeley Lab

Bassi

IO Library Comparison Weak Scaling IO (64MB/proc)

Assumptions/Setup

- Each processor writes 64MB data
- Contiguous chunks of data
- HDF5, Pnetcdf use IBM_largeblock_io = true to turn off data shipping



- *Buffering/cache effect in place*
- *Parallel IO libraries have overhead opening files, creating data objects and for synchronization*
- *One file per processor outperforms parallel IO libraries, but user must consider post processing and file management cost*

Office of Science U.S. DEPARTMENT OF ENERGY

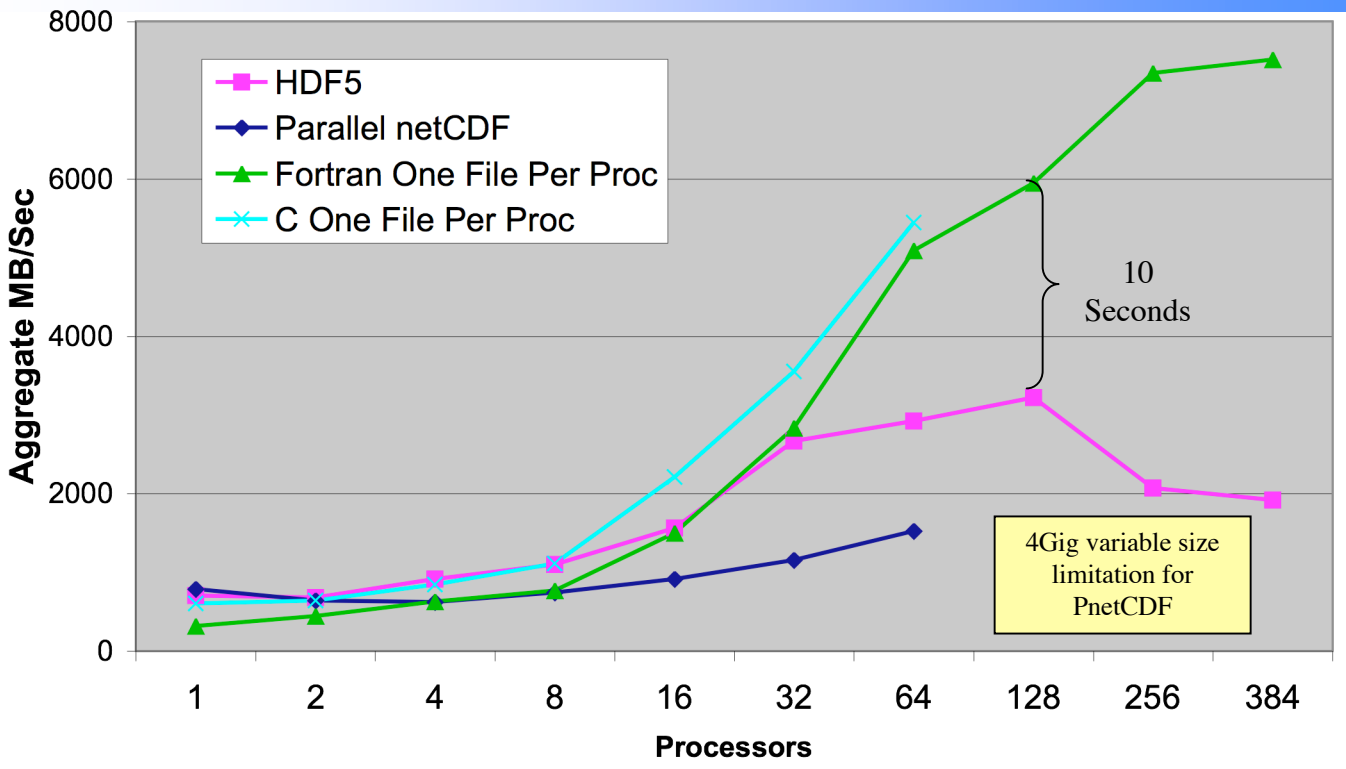
BERKELEY LAB

Bassi

IO Library Comparison Weak Scaling (576 MB/proc)

Assumptions/Setup

- Each processor writes 576MB data
- Contiguous chunks of data
- HDF5, Pnetcdf use IBM_largeblock_io = true to turn off data shipping



576 MB

221 GB

One file per processor IO begins to diverge from parallel IO strategies around 16 and 32 processors, however the absolute time difference between the two is still relatively low.

Office of Science U.S. DEPARTMENT OF ENERGY

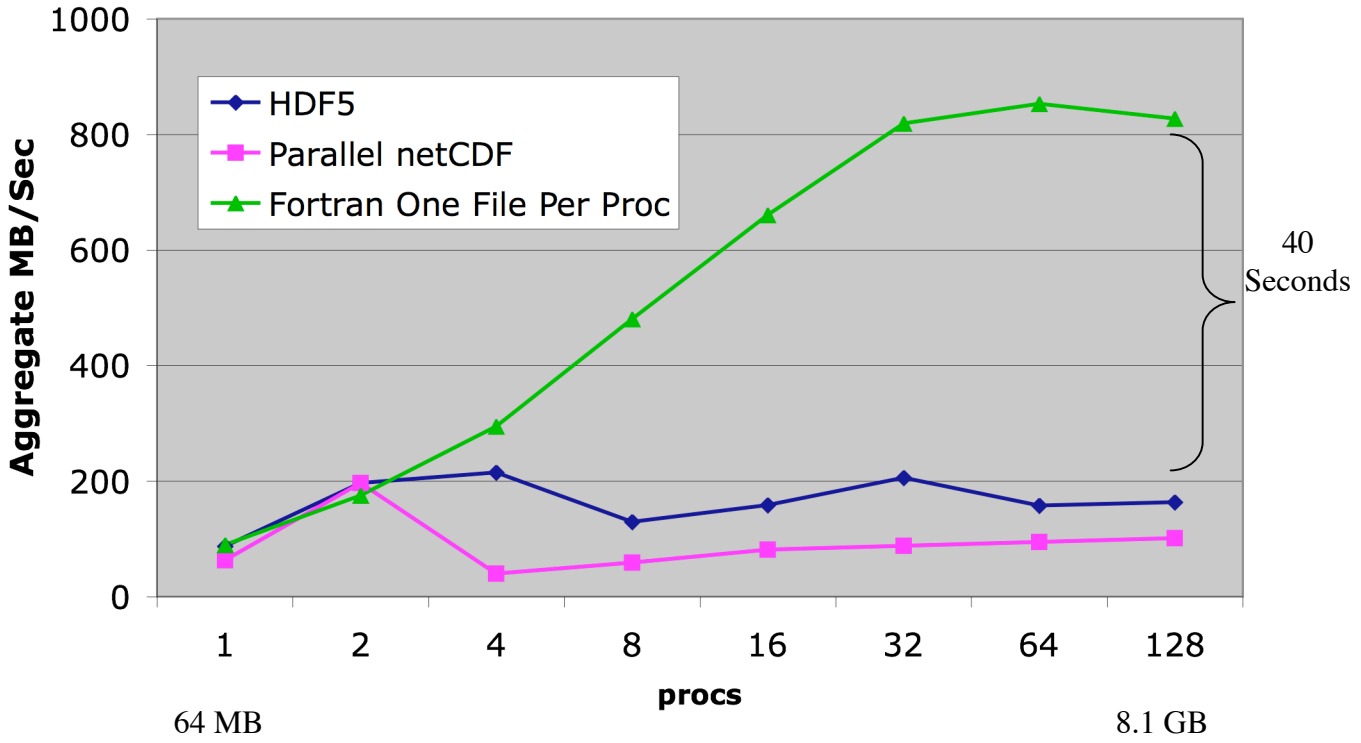
BERKELEY LAB

Jacquard

IO Library Comparison Weak Scaling IO (64 MB/proc)

Assumptions/Setup

- Each processor writes 64MB data
- Contiguous chunks of data
- No File Hints to mvapich implementation of MPI-IO



- *MPI-IO file hints for GPFS optimization are not implemented for mvapich MPI-IO*
- *Pure MPI-IO approach produces results similar to HDF5 and Parallel-NetCDF indicating performance bottleneck at MPI-IO implementation rather than higher level libraries*

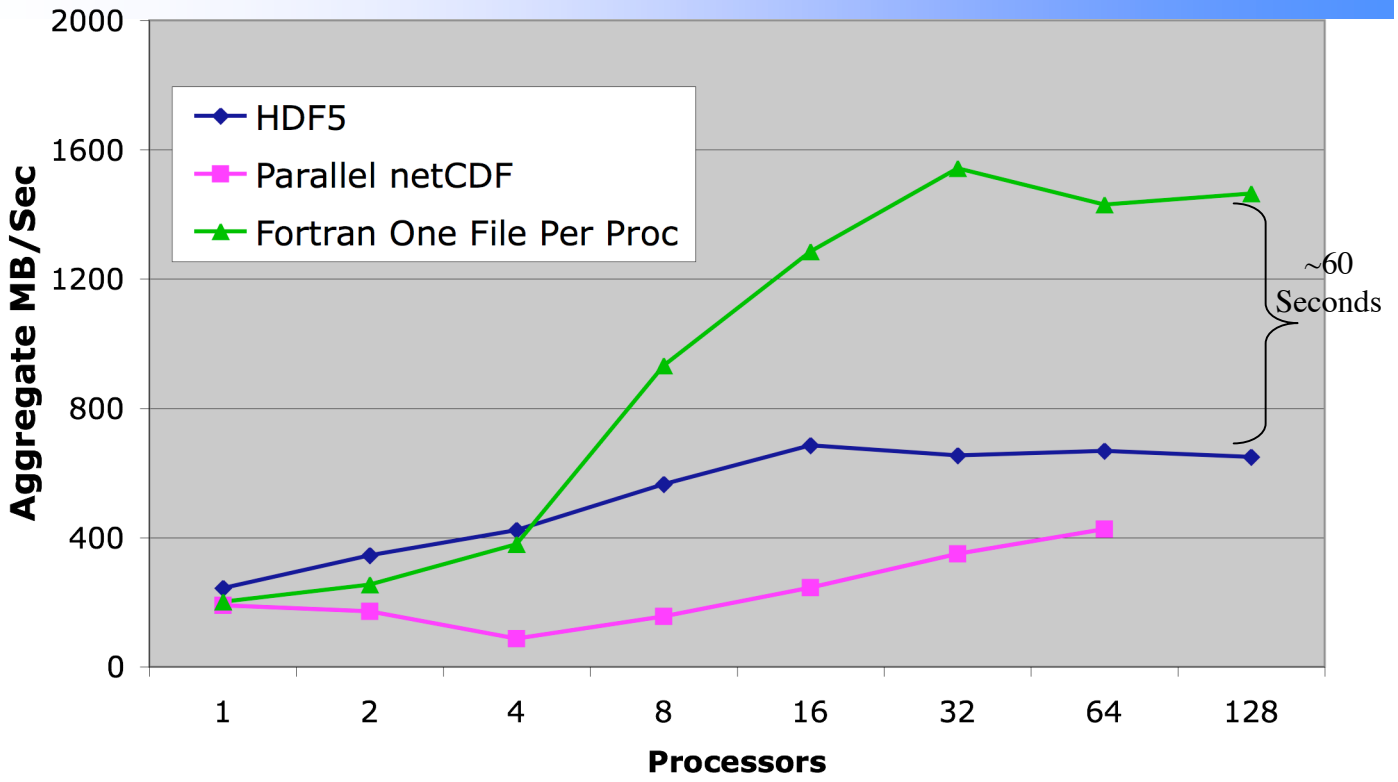
Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

IO Library Comparison Weak Scaling IO (576 MB/proc)

Assumptions/Setup

- Each processor writes 576MB data
- Contiguous chunks of data
- No File Hints to mvapich implementation of MPI-IO



IO Rates level off for HDF5 and Fortran one file per processor IO at 16 and 32 processors

Conclusions

- MPI-IO file hint IBM_largeblock_io gives significant performance boost for HDF5, less for Parallel NetCDF - exploring why.
- IBM_largeblock_io file hint must be used with IBM MPI-IO implementationa and GPFS (ie doesn't work for Jacquard)
- Yes, there is an overhead for using parallel IO libraries, but probably not as bad as users expect


NERSC


Office of
Science
U.S. DEPARTMENT OF ENERGY


BERKELEY LAB