Cray IO COE

Performance of MPIIO on DVS+GPFS

Yushu Yao

Collaboration with:

Mike Aamodt, Katie Antypas, Tina Butler, Mark Cruciani, Jason Hick, David Knaak, Rei Lee, Rose Olson, Mike Welcome

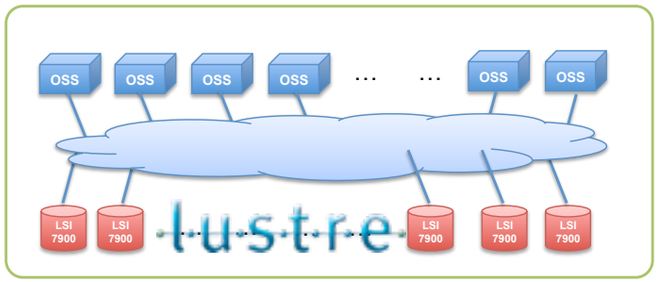
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

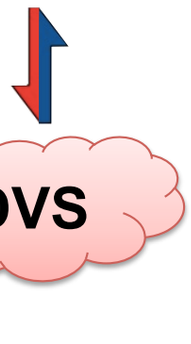
WHY

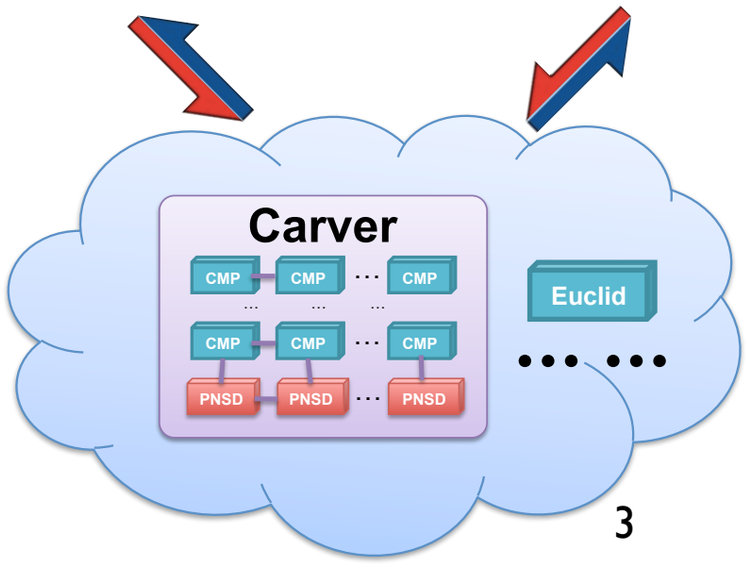
Reason 1. Users Love Global File Systems

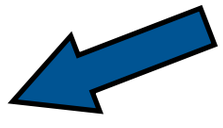


Scratch 35GB/S

FAST MPIIO!!!



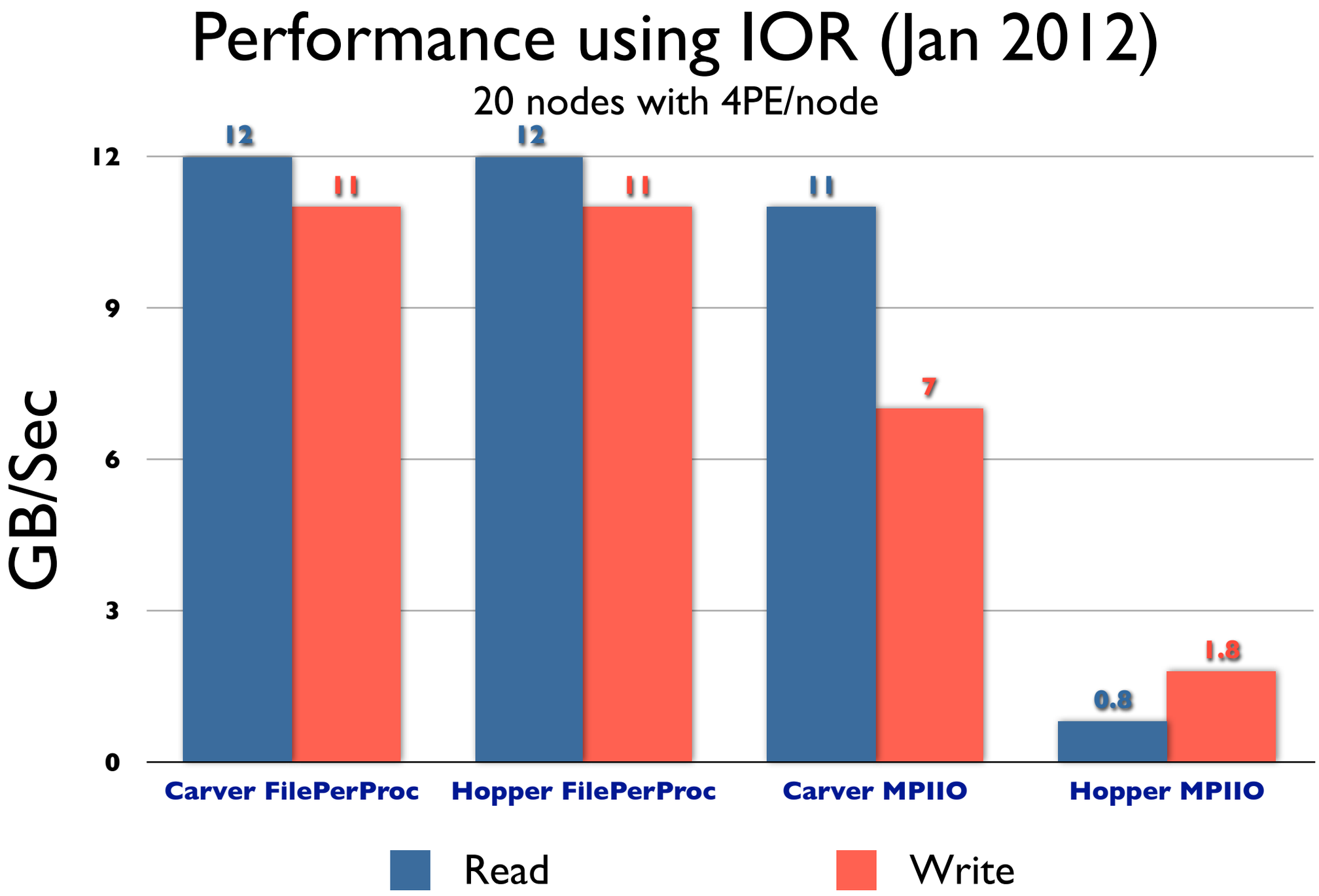




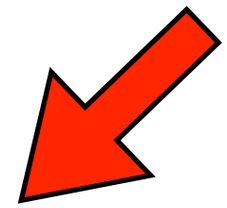
Data Analysis
Visualization

… …

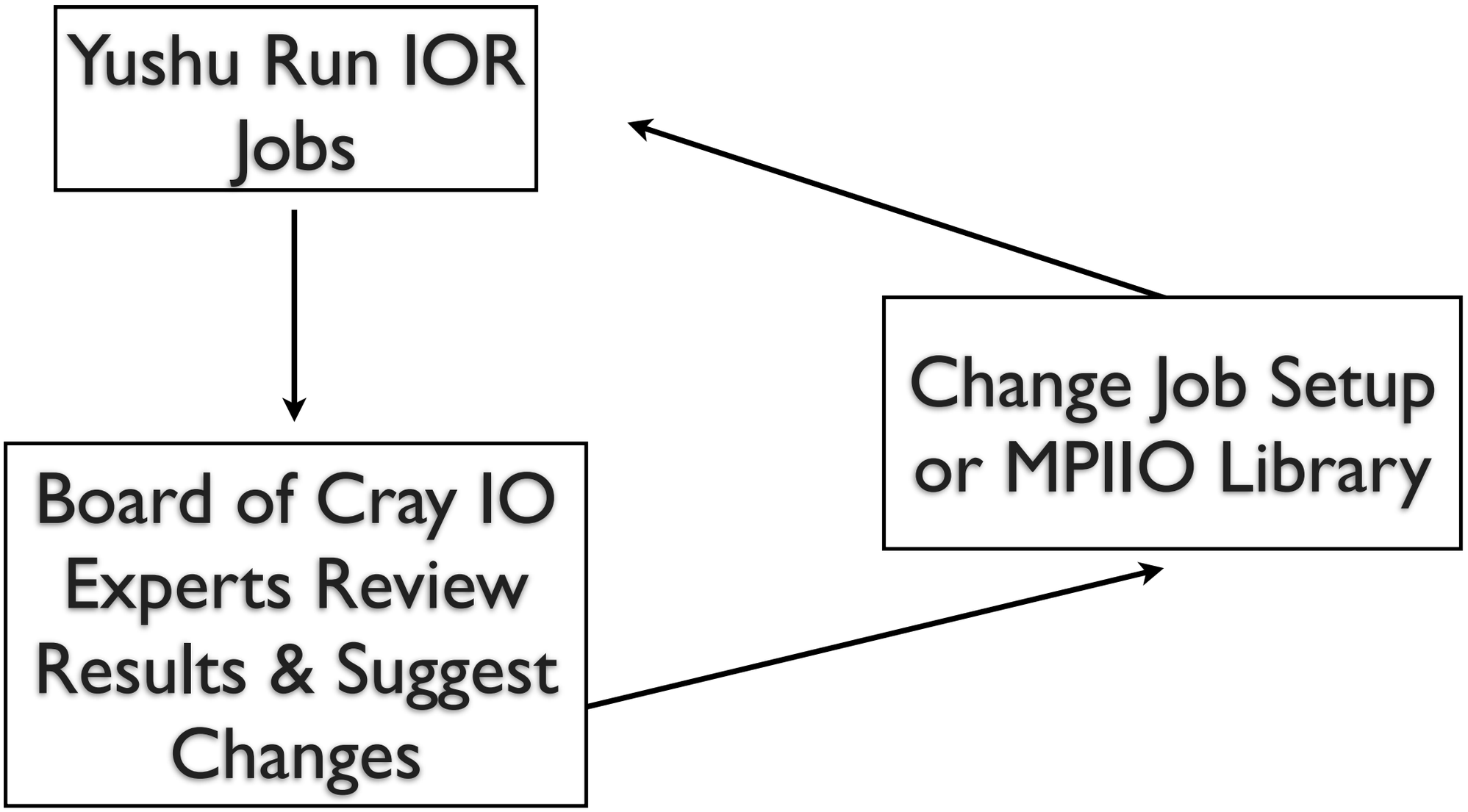
Reason 2. Well, DVS+MPIIO Was Super SLOW



Main Difficulty

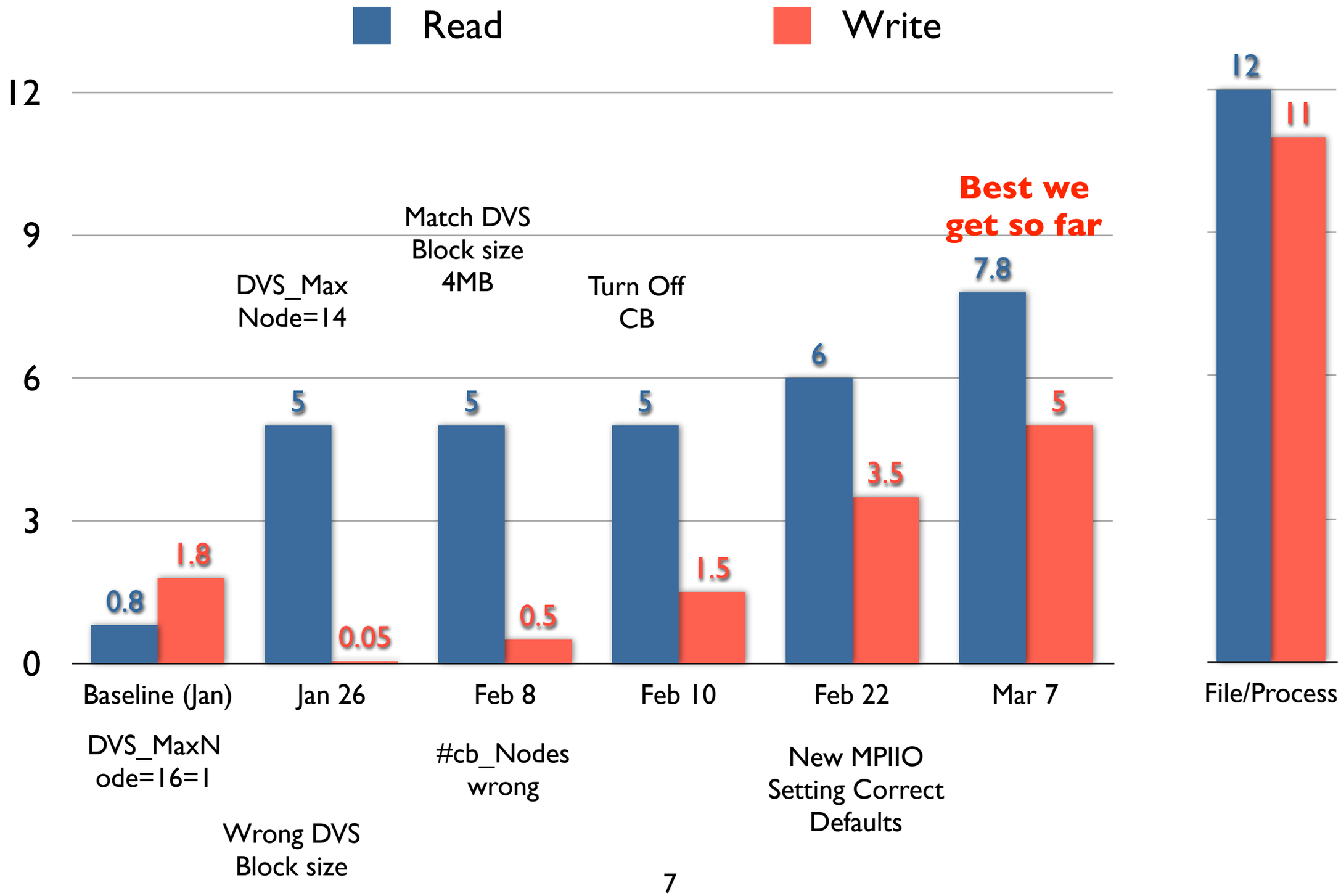
- **For Users:** Setting the right Parameters 
- **For DVS/MPIIO developer:** Not sure what GPFS is doing

Method

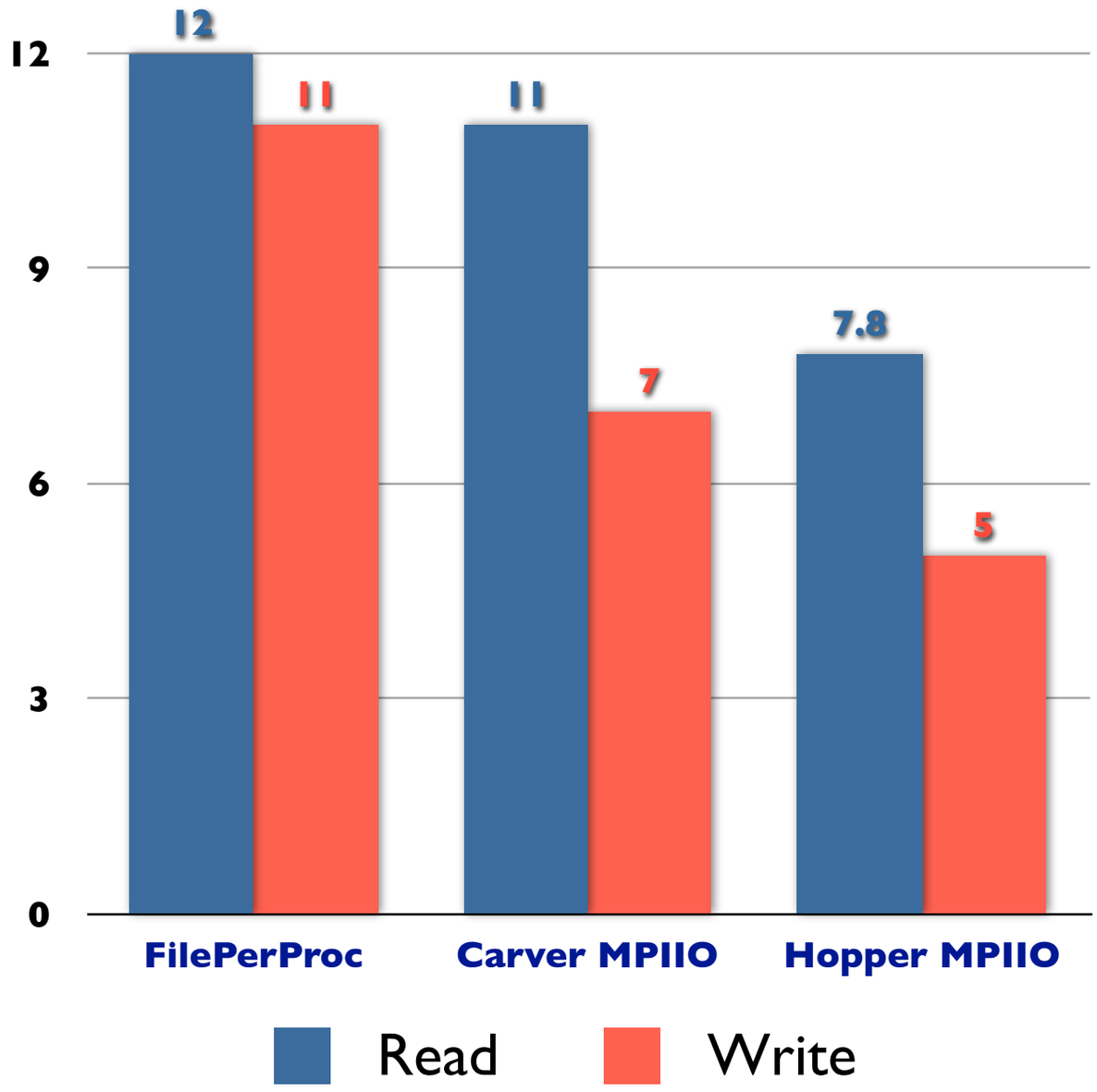


- Experts can quickly point out setup problems
- Give feedback to developers to quickly implement library changes

Progress over time ...



Best Performance



- 24PE/Node Each node reads/ writes 24GB DVSMaxNode=14 Custom MPIIOHints

- DVS_MAXNODES=14
  DVS_BLOCKSIZE=4194304
  IOR_HINT__MPI__romio_cb_read=disable
  IOR_HINT__MPI__romio_cb_write=enable
  IOR_HINT__MPI__romio_ds_read=disable
  IOR_HINT__MPI__romio_ds_write=disable
  IOR_HINT__MPI__striping_unit=4194304
  IOR_HINT__MPI__cb_nodes=14

Still, too complicated for a user to set, a naive user will 100% guess them wrongly

Best Solution:

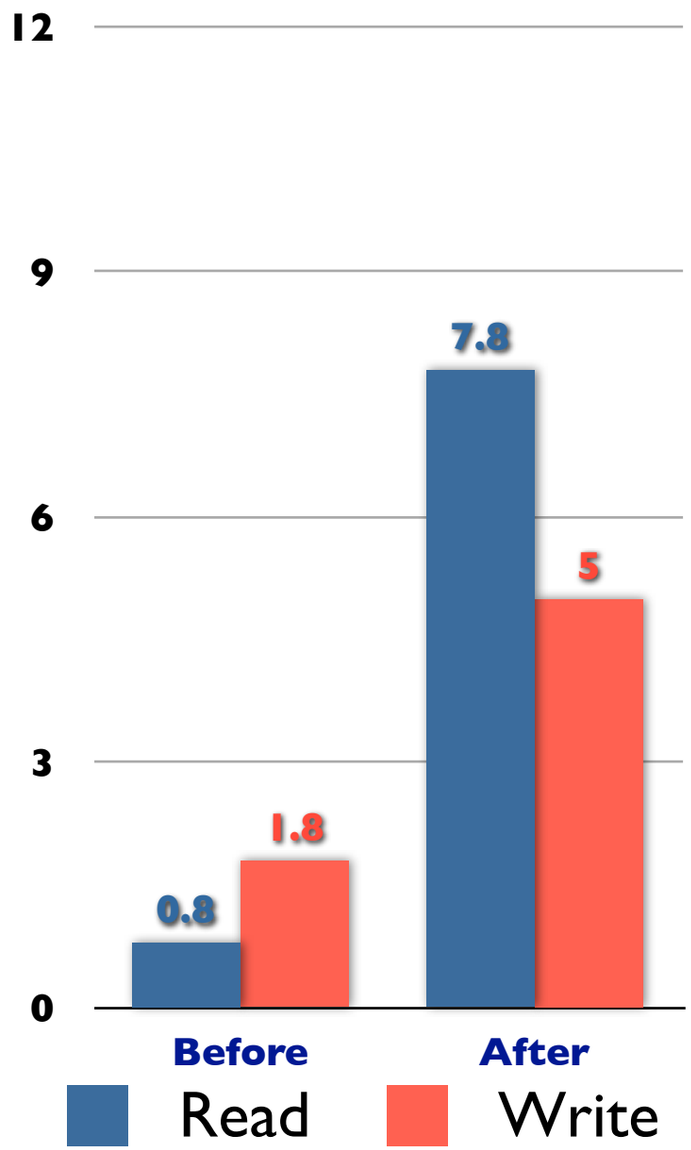
Setting Defaults for All Users

- For all users we set default environment variable:
  MPICH_MPIIO_DVS_MAXNODES=14
  DVS_BLOCKSIZE 4194304

- Non-intrusive.: This will not affect anything else
- Work-less: A user don't need to set any MPIIO hints to get (relatively) good performance

Will be on Hopper from MPT/5.5.0

Conclusion



- 10 X performance improvement on read, 3X write, after changing both run setup and MPIIO library
- Setting DEFAULT values for users so that they can get best performance (in most cases) automatically

Next Step

- **For DVS/MPIIO developer:** Not sure what GPFS is doing
- IO benchmarking on DVS nodes to figure out where the bottlenecks are
- Maybe carried out in a less formal way?