

On the Duality of Data-intensive File System Design: Reconciling HDFS and PVFS

Wittawat Tantisiriroj, Swapnil Patil, Garth Gibson (CMU) - Seung Woo Son, Samuel J. Lang, Robert B. Ross (ANL)

Overview

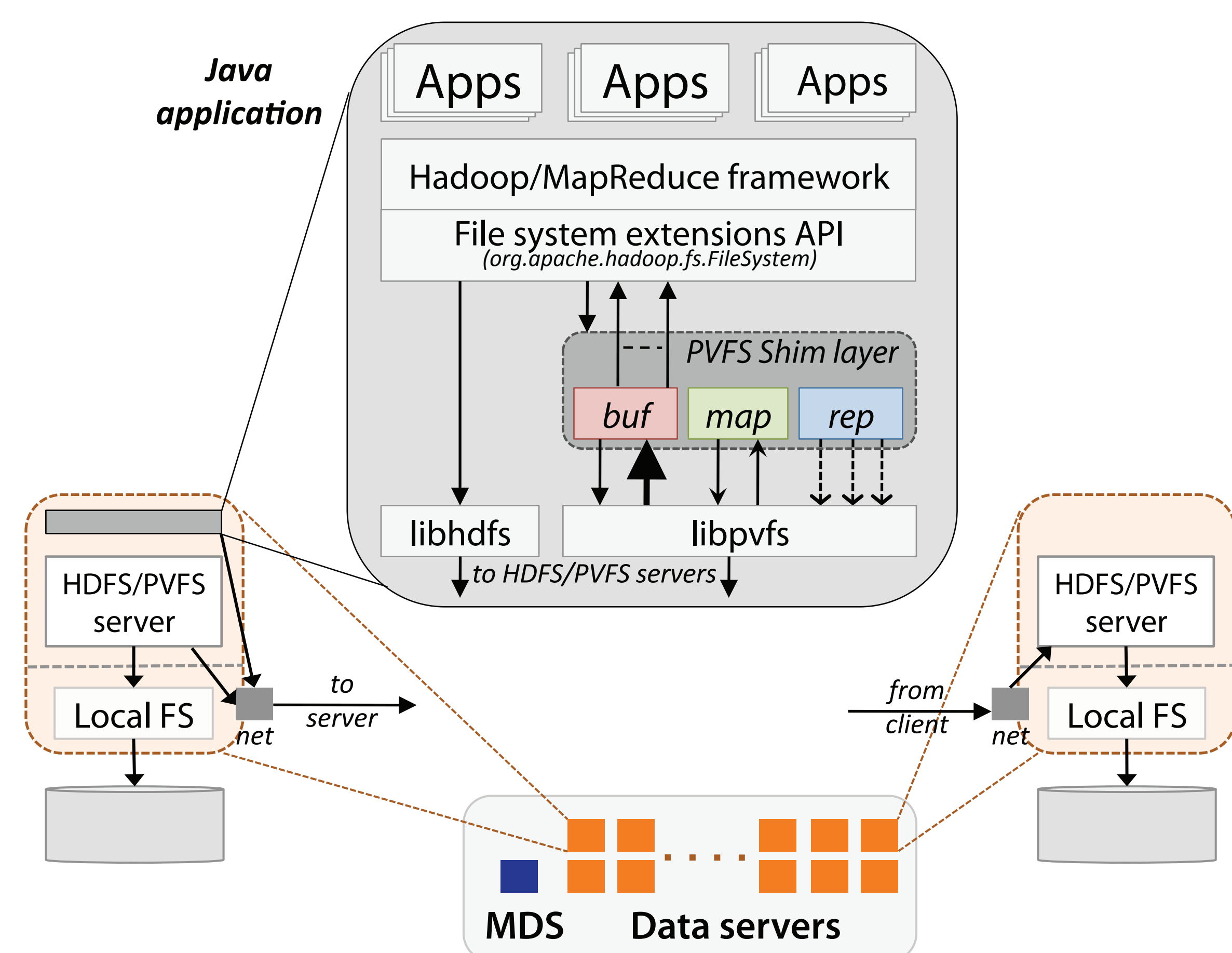
Internet Services

- Distributed file system
 - Purpose-built for anticipated workloads
- Hadoop & Hadoop distributed file system (HDFS)
 - Use triplication for reliability
 - Use file layout to collocate computation and data

High performance computing (HPC) [e.g. PVFS]

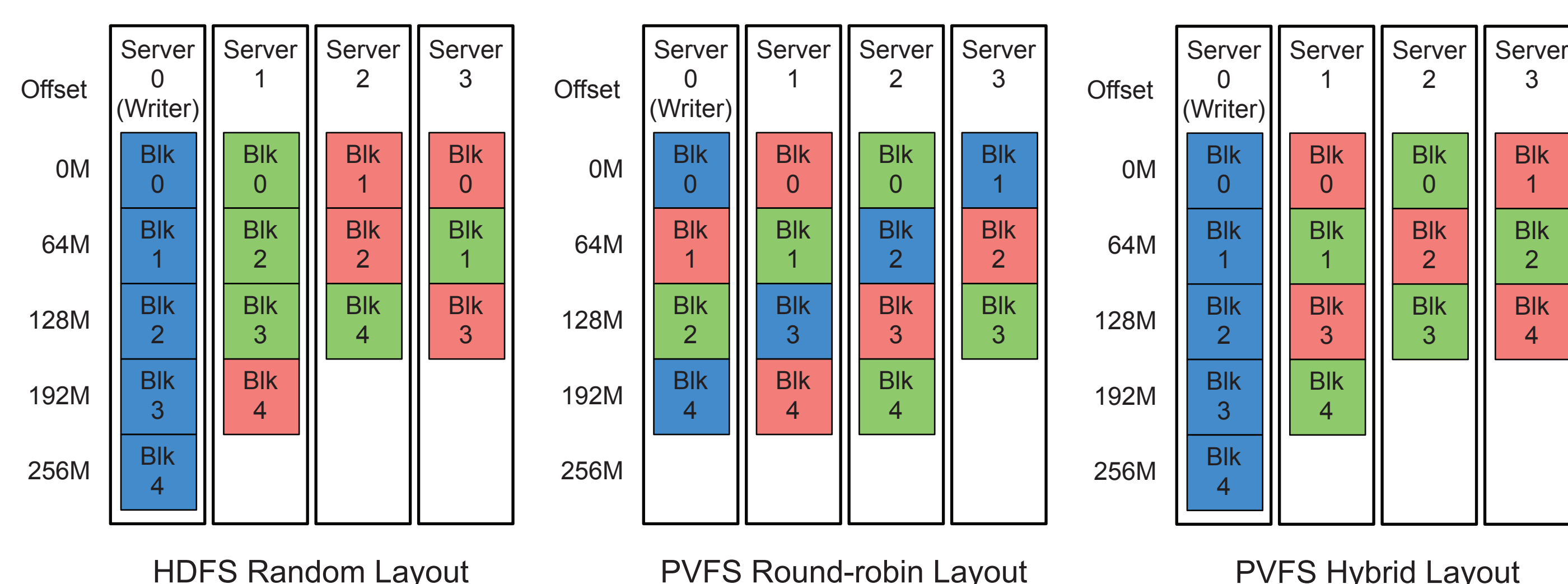
- Equally large scale applications
- Parallel file system
 - Concurrent reads and writes
 - Typically support POSIX and VFS interface

PVFS Plug-in under Hadoop Stack



PVFS Shim responsibilities:

- Readahead buffer: reads from PVFS in 4MB requests
- File layout: file layout exposed as extended attributes
- Replication: triplicates data in one PVFS file



HDFS/PVFS data layout schemes:

- HDFS Random: 1 copy on writer's disks, 2 copies random
- PVFS Round-robin: 3 copies striped in file
- PVFS Hybrid: 1 copy on writer's disks, 2 striped

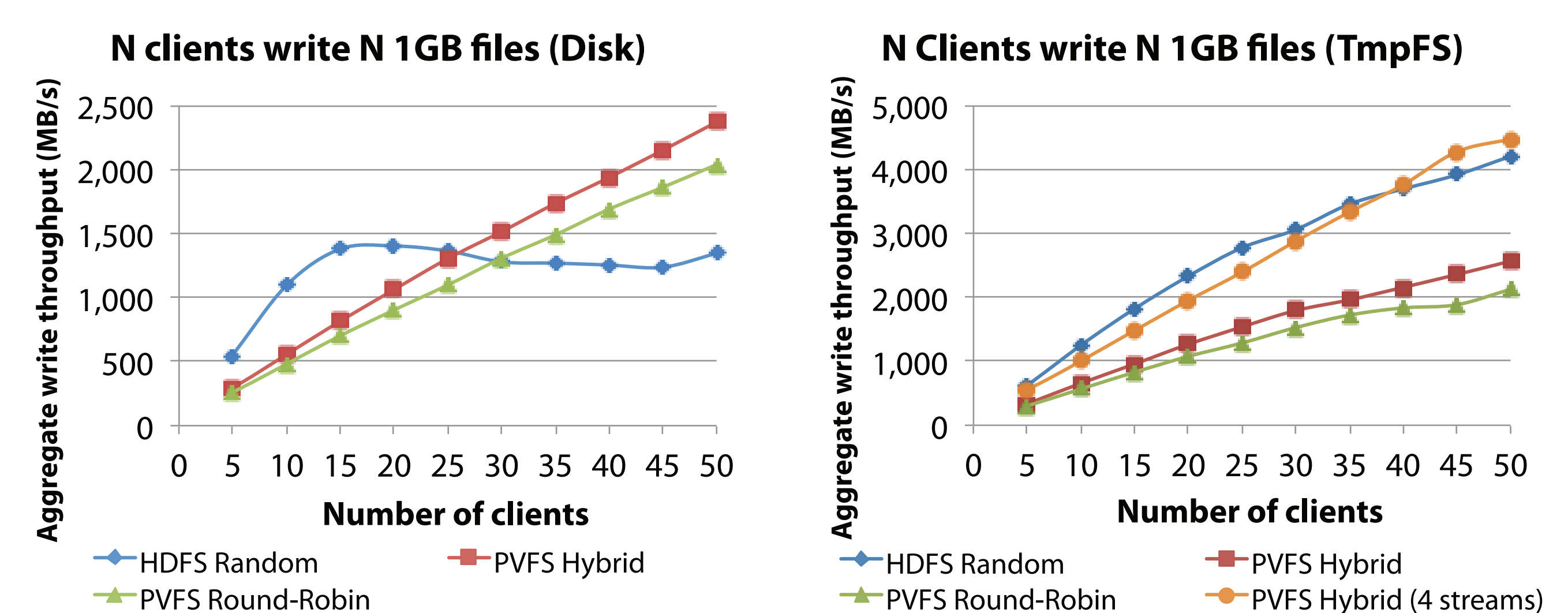
Acknowledgements: Robert Chansler, Tsz Wo Sze, Nathan Roberts, Bin Fu, and Brendan Meeder

Carnegie Mellon

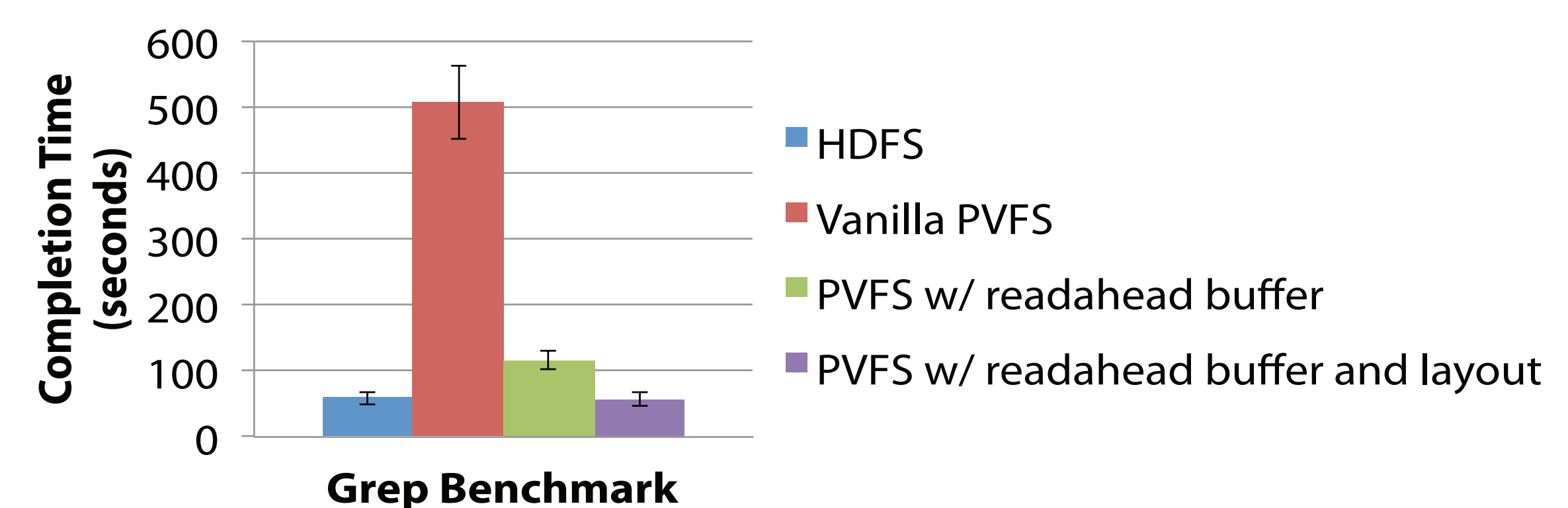
Experiment Setup

- OpenCloud cluster - 51 nodes (8-core 2.8 GHz, 16GB DRAM, 4 SATA disks, 1 used in experiments, 10 GE)
- Benchmarks
 - Data-set: 50 million 100-byte records (50GB)
 - Workload: write, read, grep (for a rare pattern), sort
- Applications
 - Sampling (B. FU): Read 71GB astronomy data-set
 - FoF (B. FU): Cluster & join astronomical objects
 - Twitter (B. Meeder): Reformat 24GB to be 56GB

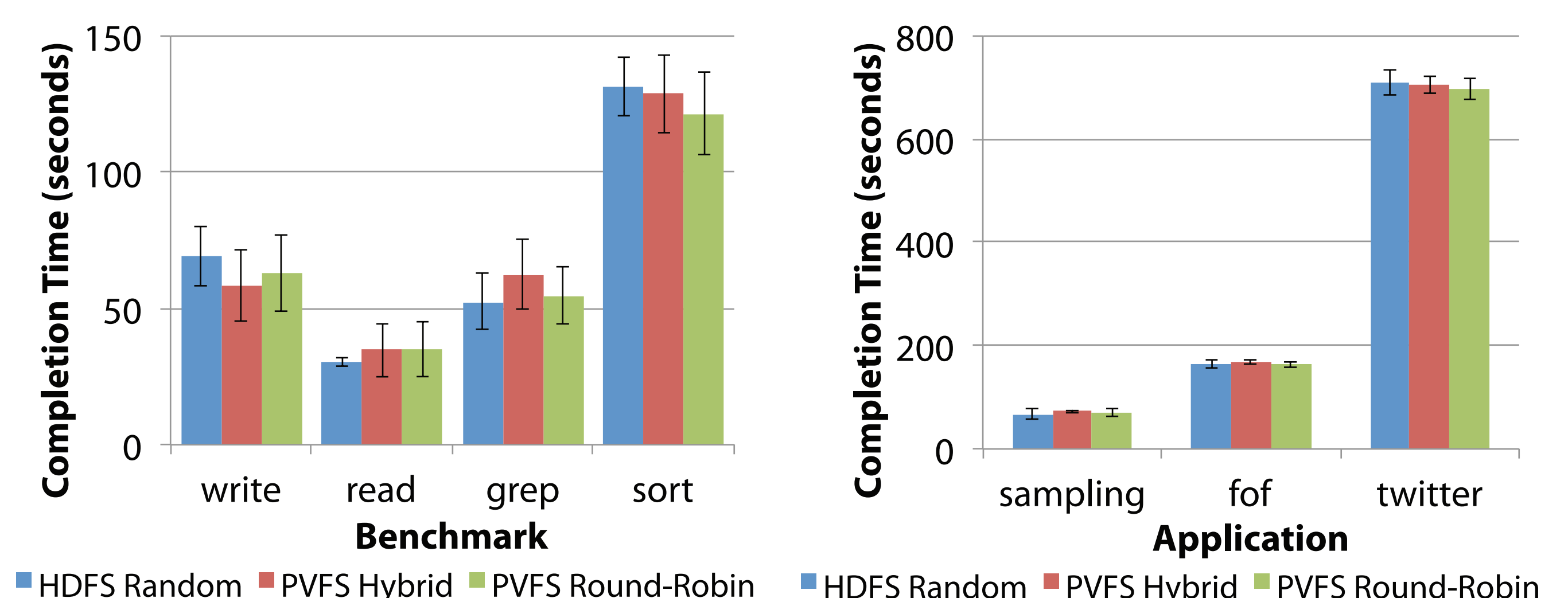
Experiment Results



- [Left] HDFS performance is limited by head-of-line blocking when creating file while the disk is busy flushing write-back buffer
- [Right] HDFS pipelined replication improves parallelism and resource utilization



- By using both readahead buffer and file layout information, PVFS performance is comparable to HDFS



- PVFS performance is comparable to HDFS for both Hadoop benchmarks and scientific applications

Conclusions

- With a few modifications in a non-intrusive shim layer, PVFS matches performance for Hadoop applications
- File layout information is essential for Hadoop to collocate computation and data