PLFS: A Checkpoint Filesystem for Parallel Applications

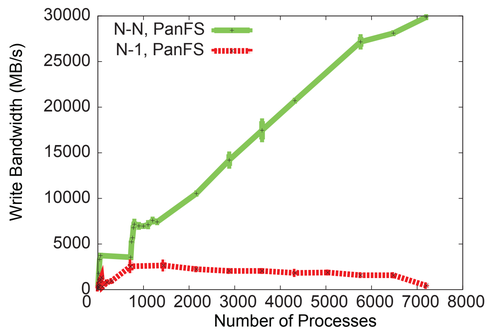
John Bent*, Garth Gibson†, Gary Grider*, Ben McClelland*, Paul Nowoczynski‡,
James Nunez*, Milo Polte†, Meghan Wingate*

**Los Alamos National Laboratory †Carnegie Mellon University ‡Pittsburgh Supercomputing*

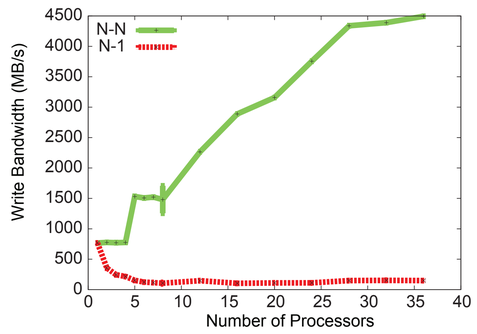
Problem

- **Many important scientific applications create checkpoints using small, strided, concurrent writes to a shared file (N-1 checkpointing)**
- **Filesystems perform best on non-concurrent sequential workloads, such as N-N checkpointing**
- **Small-strided writes to a shared file often suffer from seeks and false sharing**
- **Unfortunately, we can't change the applications, but we can modify our filesystems**

Checkpoint Bandwidth



PanFS



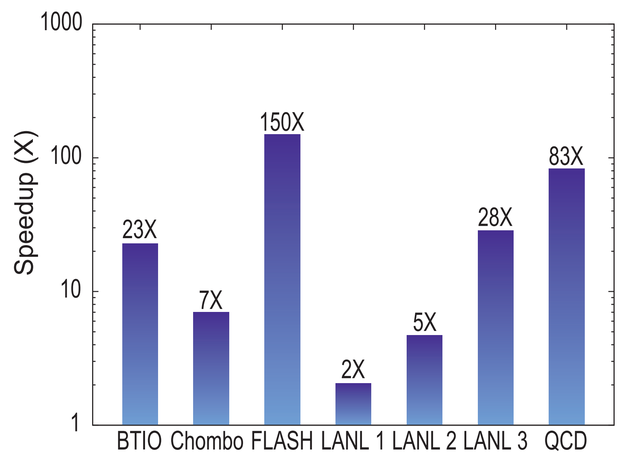
GPFS

PLFS – Parallel Log-Structured File System

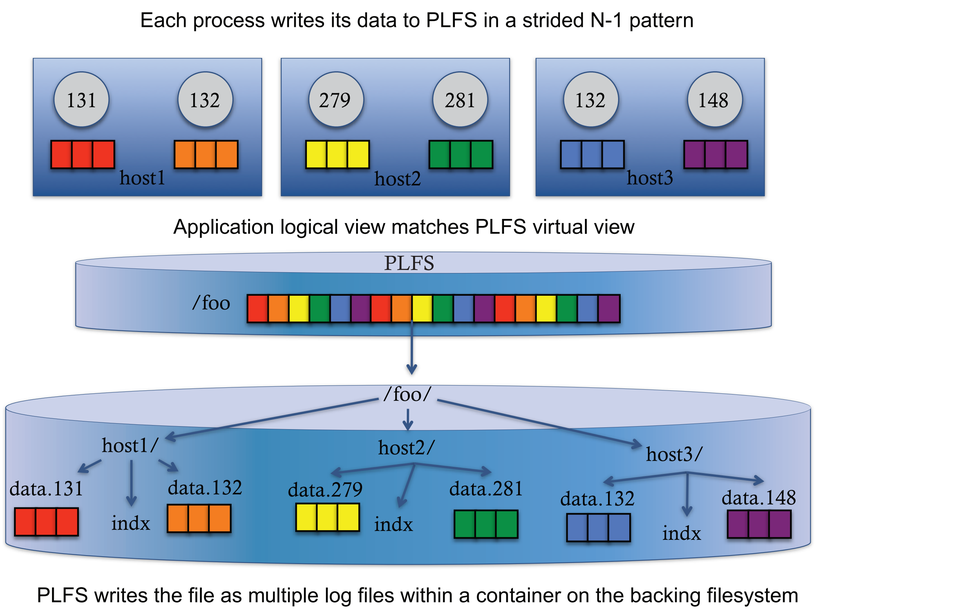
- **A project developing a filesystem level improvement to N-1 checkpointing, led by John Bent (LANL)**
- **FUSE based filesystem mounted on top of any existing parallel filesystem on clients**
- **Decouples a concurrent N-1 checkpoint into a non-concurrent N-N checkpoint**
- **Redirects strided writes from multiple processes accessing a single file to sequential writes to data logs and index files**

PLFS Speedup

- **2x – 150x speedups for important HPC applications at LANL scale!**

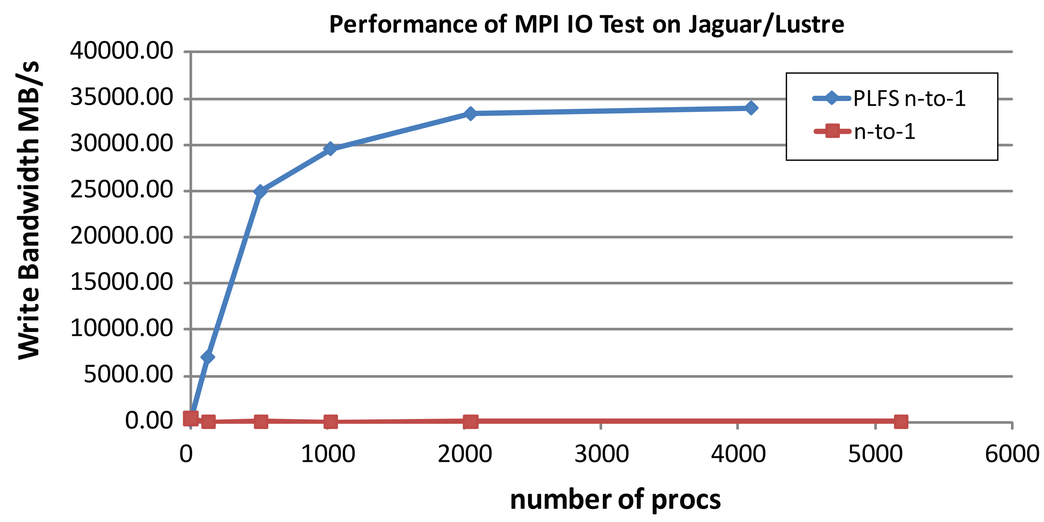


Layout of a PLFS Container



Elaborations and Future Work

- **PLFS can generate light-weight write map traces**
  - **http://institute.lanl.gov/plfs/maps/**
- **MPI-IO and Library Interfaces**
  - **Eliminate FUSE overheads**
  - **Enabling PLFS at sites without FUSE**
  - **Currently working on deployment on Jaguar XT at ORNL**
- **Analyzing read performance on non-checkpoint workloads**
- **In-memory index servers for Read/Write mode**
- **Project homepage: http://sourceforge.net/projects/plfs/**



N-to-1 checkpoint MPI-IO benchmark
running on Jaguar (Cray XT, Lustre filesystem)
with and without PLFS

Carnegie Mellon
with information from LANL Technical Release LA-UR 09-02117

pdsi

IRHPIT INSTITUTE FOR RELIABLE HIGH PERFORMANCE INFORMATION TECHNOLOGY

Parallel Data Laboratory
PDLR10