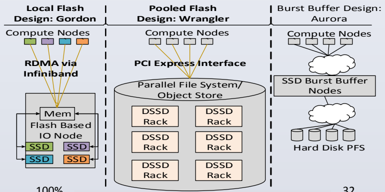
Lessons from Post-processing Climate Data on Modern Flash-based HPC Systems

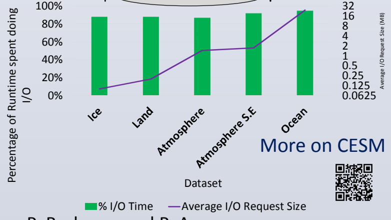
Adnan Haider[1], Sheri Mickelson(Advisor)[2], John Dennis(Advisor) [2], Xian-He Sun (Advisor) [1]

[1]Illinois Institute of Technology, USA; [2]National Center of Atmospheric Research, USA

Flash-based Systems and Post-processing Software

Flash devices are a plausible solution to accelerate I/O bound applications. However, the tradeoffs associated with different flash architectures is unclear. We quantitatively assess two modern flash architectures using post-processing climate data applications to **facilitate correct matching between I/O workloads and flash storage architectures.**





More on CESM

- PyReshaper and PyAverager
- 90% of execution time is spent waiting for I/O to complete.
- Different datasets have vastly different I/O workloads (i.e. request size).
- IOR used for comparison with other workloads

Gordon System Results: Local Flash Architecture

- Each compute node has access to a single solid state drive (SSD)
- Remote direct memory access via Infiniband.
- Can cause accesses to become queued

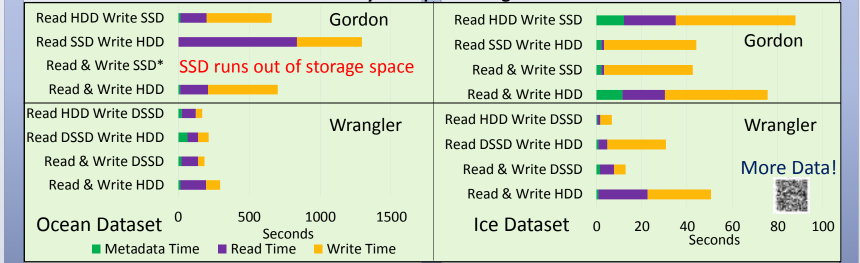
1) Single SSD cannot handle rate of parallel accesses and interconnect causes latency.

Wrangler System Results: Pooled Flash Architecture

- Uses DSSD devices which are faster than SSD.
- Each compute node has access to all DSSD devices (Pooled) via PCI Express
- Deploys parallel file system

1) Multiple DSSD and high throughput interconnect provide 2x to 6x improvements.

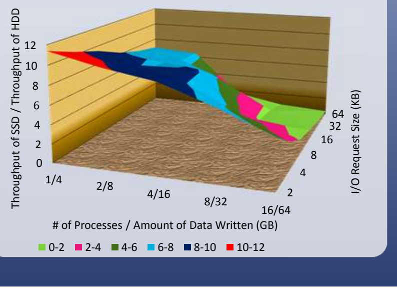
PyReshaper Timings



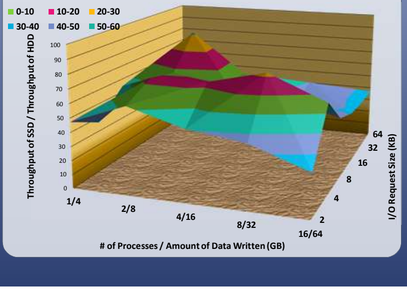
More Data!

IOR Benchmark

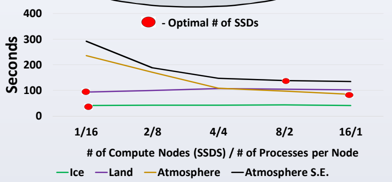
2)Benefits of flash **decreases** at moderate scale and relatively small request sizes.



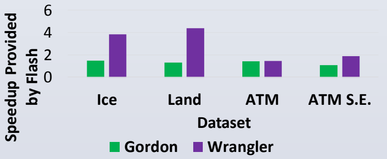
2) **Consistent** benefits for all configurations when using flash.



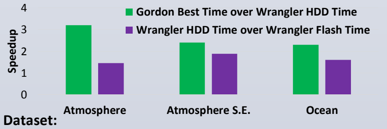
Comparison of I/O Architectures



Multiple flash devices per compute node are needed to accommodate rate of parallel accesses issued by post-processing applications.



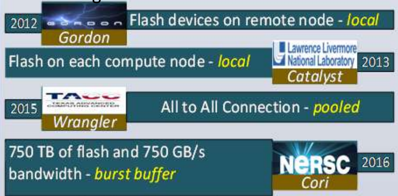
A local architecture provides similar speedups as a pooled architecture if using multiple flash devices per compute node.



Using a three-year newer system while not using flash (green bar) provides **more** speedup than using flash while keeping other hardware constant (purple bar)

Lessons Learned

- An incorrect matching between storage architecture and I/O workload can hide the benefits of flash devices by increasing runtime by 2x.
- Hybrid I/O decreases flash storage consumption by half while decreasing runtime by 6x. Video Presentation→
- Local flash could be a cheaper alternative to a pooled architecture if scalability and interconnect bottlenecks are alleviated.
- Three main criteria which determine performance on flash systems. 1) Number of flash devices in job. 2) Interconnect 3) Data availability of data stored on flash.
- Three years of more advanced hardware without flash devices provides more speedup than flash devices for some datasets, lessening the need for flash.



Flash devices on remote node - *local*

Flash on each compute node - *local*

All to All Connection - *pooled*

750 TB of flash and 750 GB/s bandwidth - *burst buffer*

Acknowledgements

We would like to thank XSEDE, TACC, and SDSC for the use of their resources and support. XSEDE allocation # TG-ASC150025