Efficiency Evaluation of Cray XT Parallel IO Stack

Weikuan Yu, Sarp Oral, Jeffrey Vetter, Richard Barrett

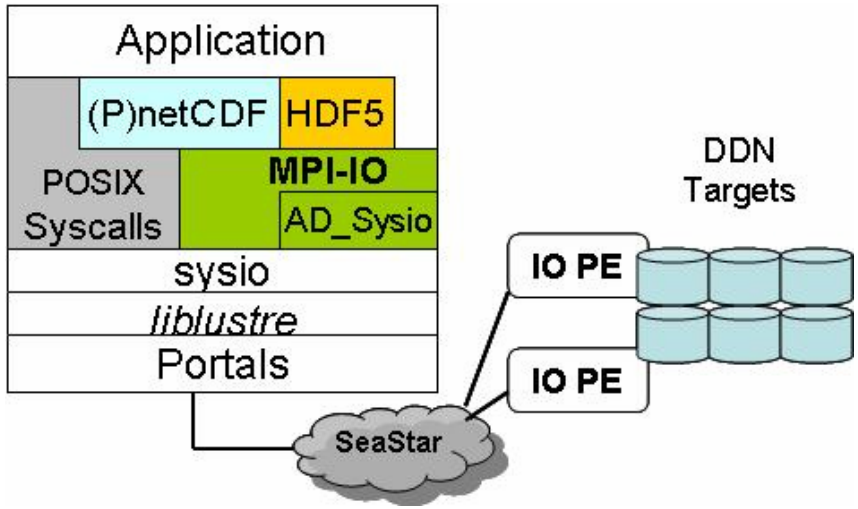
Future Technologies & Technology Integration
Oak Ridge National Laboratory
{wyu,oralhs,vetter,rbarrett}@ornl.gov

Presentation Outline

- Overview of Cray XT Parallel IO
- Posix Read/Write
- MPI-IO Independent IO
- Collective IO
- Sample Performance Tuning
  - BT-IO
  - Flash IO (HDF5)

Cray XT Parallel IO

- MPI-IO
  - IO interface for the message passing standard
  - Primary foundation for the parallel IO software stack
  - Over Cray XT platforms
    - vendor-supplied MPI-IO implementation
- Portable programming interface for higher layer IO models
  - HDF5
  - PnetCDF

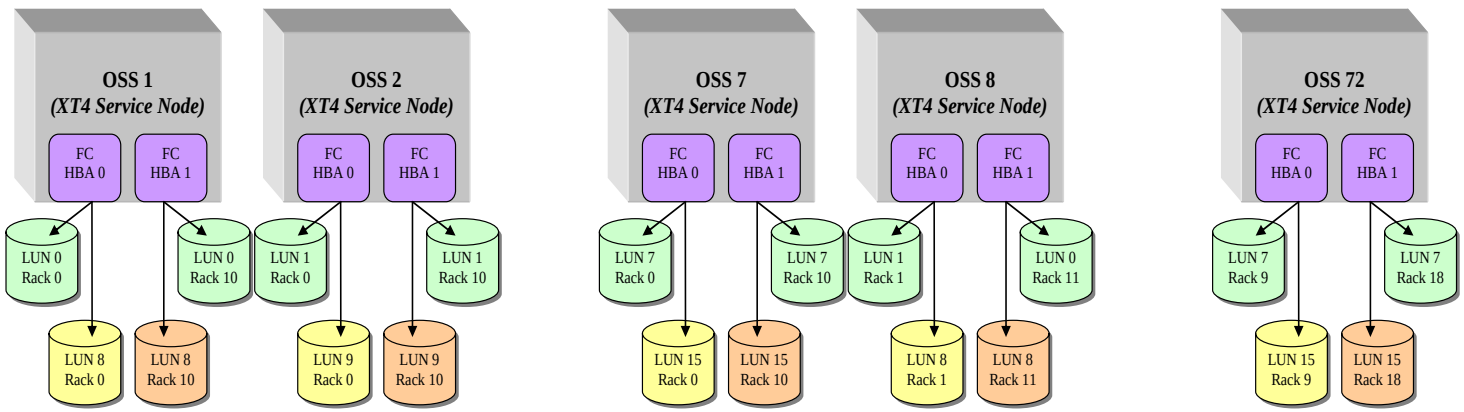


MPI-IO over Cray XT

- Generic techniques
  - Structured IO
  - Data sieving
  - Extended two-phase collective IO
  - Disk-directed aggregation
  - MPI-IO Prefecthing and Caching
  - Asynchronous, threaded IO
- Lustre Specific
  - Striping (manually with lfs {get,set}stripe)
  - Data Shipping (GPFS), List IO (PVFS)

ORNL Cray XT (Jaguar) Parallel IO Configuration

- Storage (DDN 9550)
  - 18 racks, each of 36 TB.
  - Connected direct Fibre Channel (FC) links.
  - Each LUN (2TB) spans two tiers of disks,
  - The write-back cache is set to 1MB on each controller.
- Three Lustre file systems
  - Each with its own MDS
  - Two with 72 OSTs, each of 150TB
  - One with 144 OSTs, 300 TB,
    - capable of 45 GB/s block IO bandwidth
  - 72 service nodes for OSSs, each supporting 4 OSTs

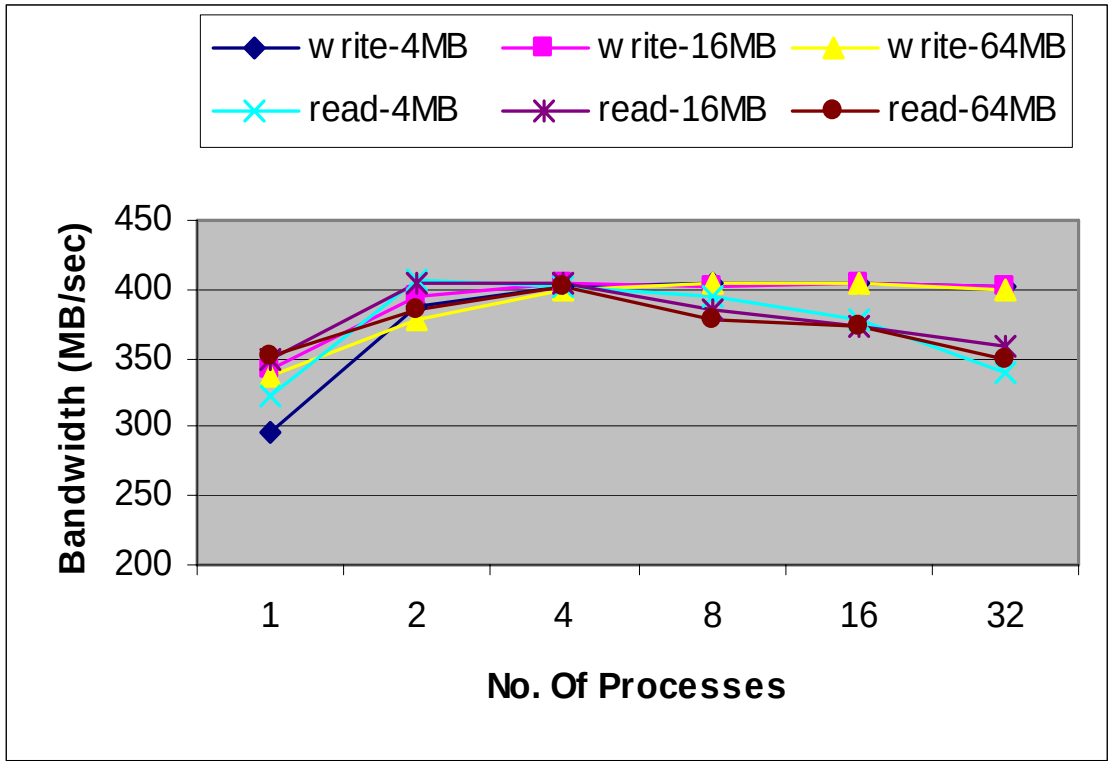


Presentation Outline

- Overview of Cray XT Parallel IO
- POSIX Read/Write
- MPI-IO Independent IO
- Collective IO
- Sample Performance Tuning
  - BT-IO
  - Flash IO (HDF5)

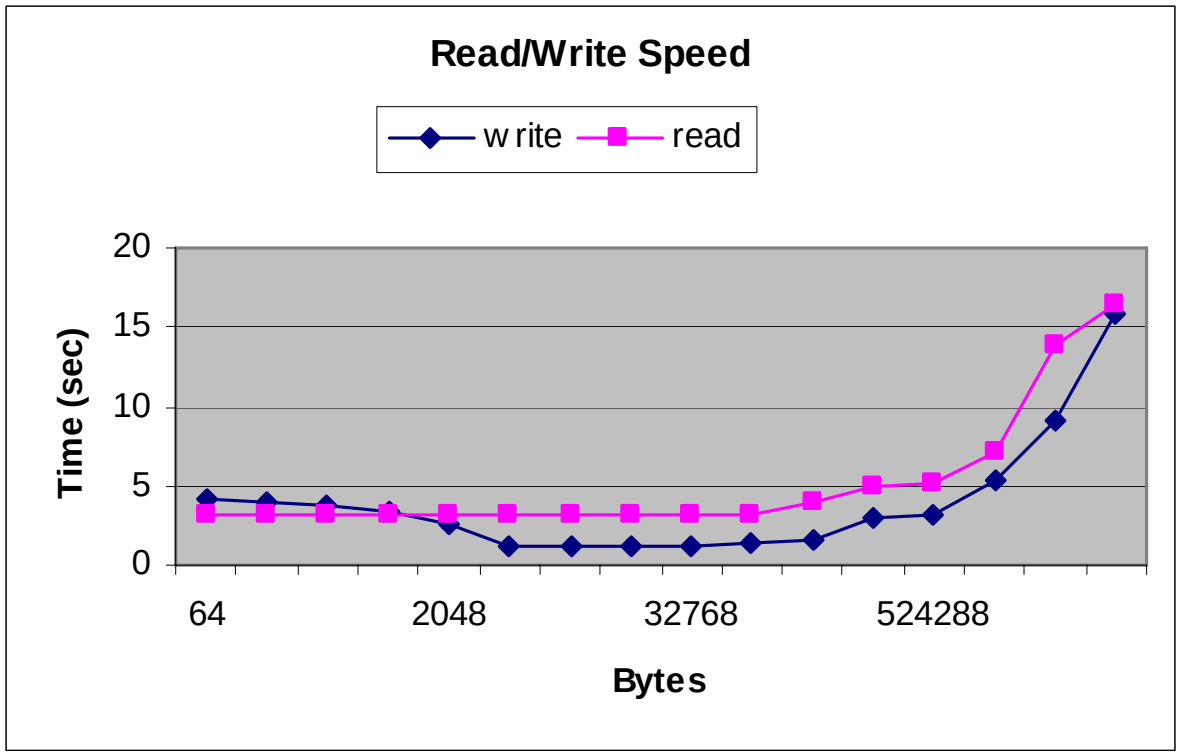
POSIX Read/Write – HotSpot

- Hotspot tests
  - Multiple processes read/write to a file over one OST
  - Test program: IOR benchmark from LLNL
  - transfer size, 4MB – 64MB
  - Max Read/Write per OST: 400MB/sec
  - Graceful write; more vulnerable to hotspot read pressure



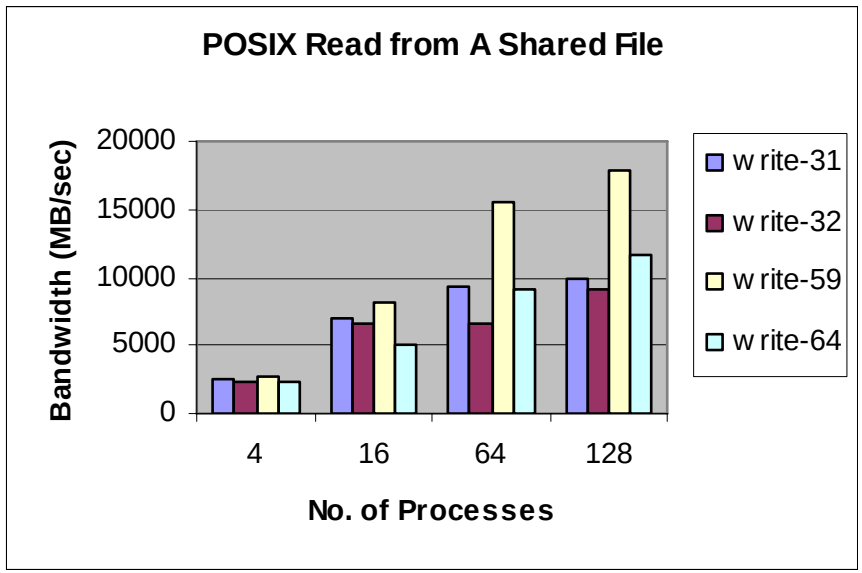
POSIX Read/Write – Operation Latency Test

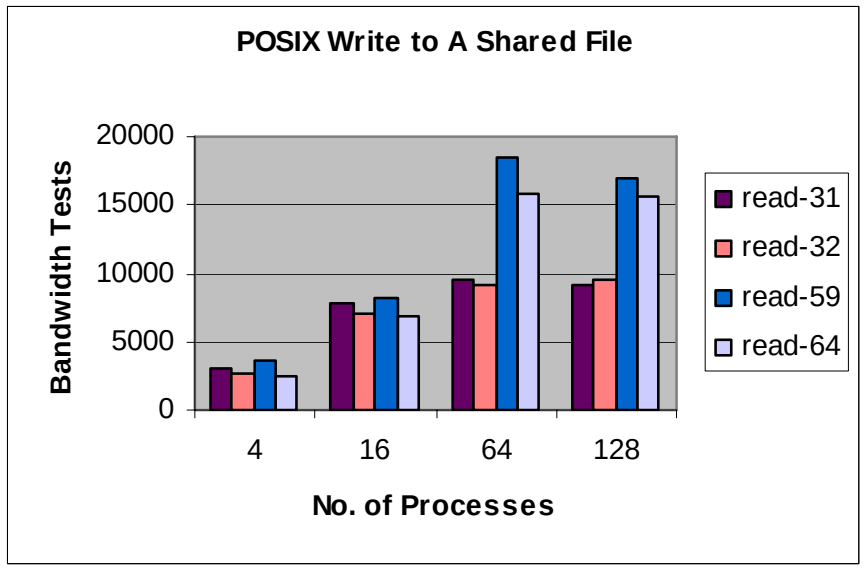
- Operation Latency Tests
  - Test program: IOR, transfer size, 64B – 2GB
  - 1000 Iterations of read/write. One process to one OST
  - Slow small write for very small messages
  - Faster Write for 2KB++



POSIX Read/Write – Bandwidth with Shared File

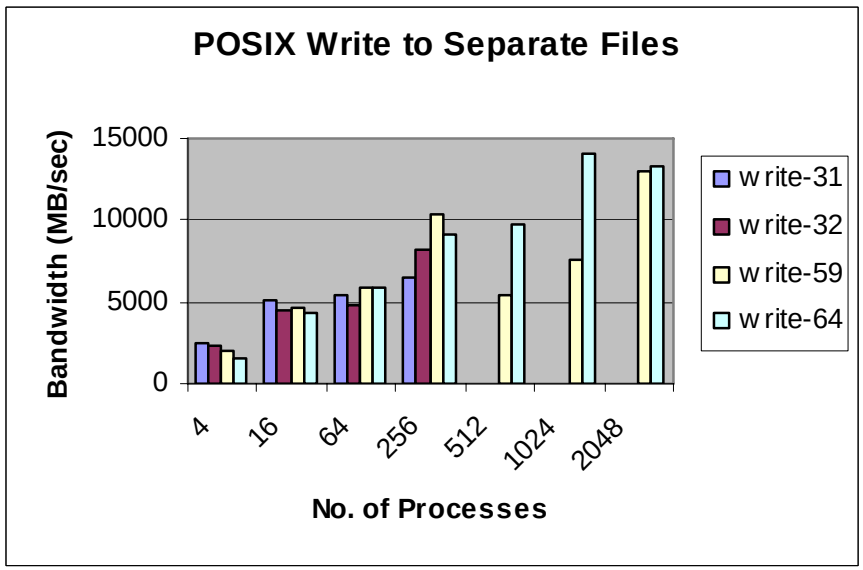
- Bandwidth tests
  - Test program: IOR
    - file size (512MB), transfer unit: 64MB
  - 31/32/59/64 stripes
  - Prime stripe counts seem to help read with small number of processes

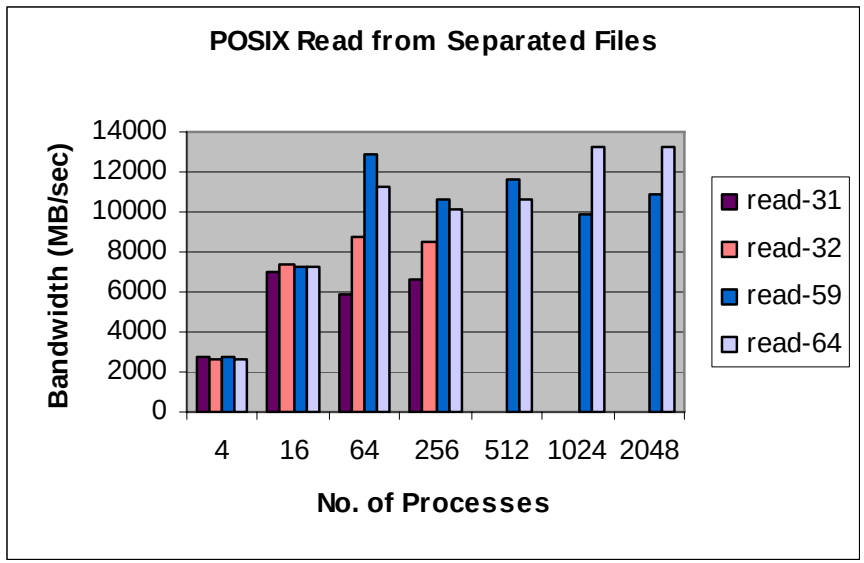




POSIX Read/Write – Bandwidth with Separated File

- Bandwidth tests
  - Test program: IOR
    - file size (512MB), transfer unit: 64MB
  - Read/Write with separated files performs worse than that with a shared file
  - 31/32/59/64 stripes; Prime stripe counts seem to help read
  - Too many processes may hurt aggregated bandwidth



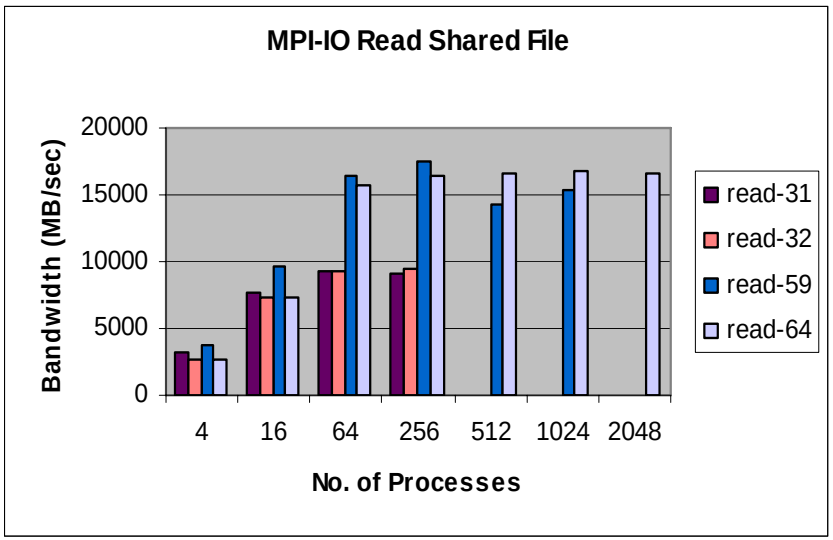


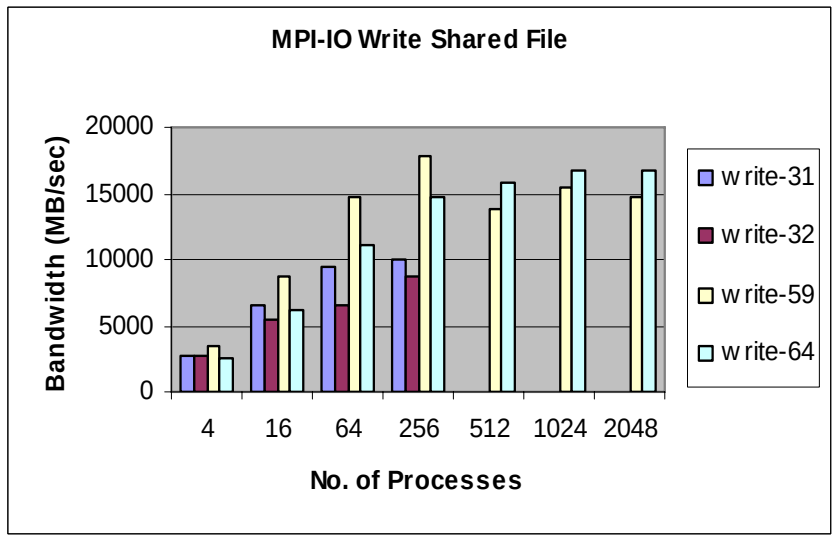
Presentation Outline

- Overview of Cray XT Parallel IO
- POSIX Read/Write
- MPI-IO Independent IO
- Collective IO
- Sample Performance Tuning
  - BT-IO
  - Flash IO (HDF5)

MPI-IO Read/Write – Shared file

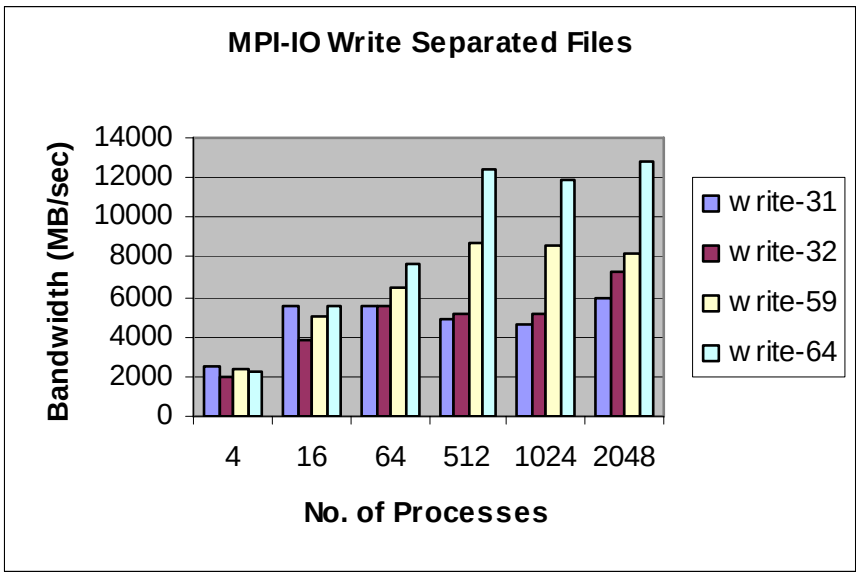
- Bandwidth tests
  - Test program: IOR
    - file size (512MB), transfer units 64MB
  - 31/32/59/64 stripes. Prime stripes help? Maybe for read
  - Too many processes can degrade aggregated bandwidth

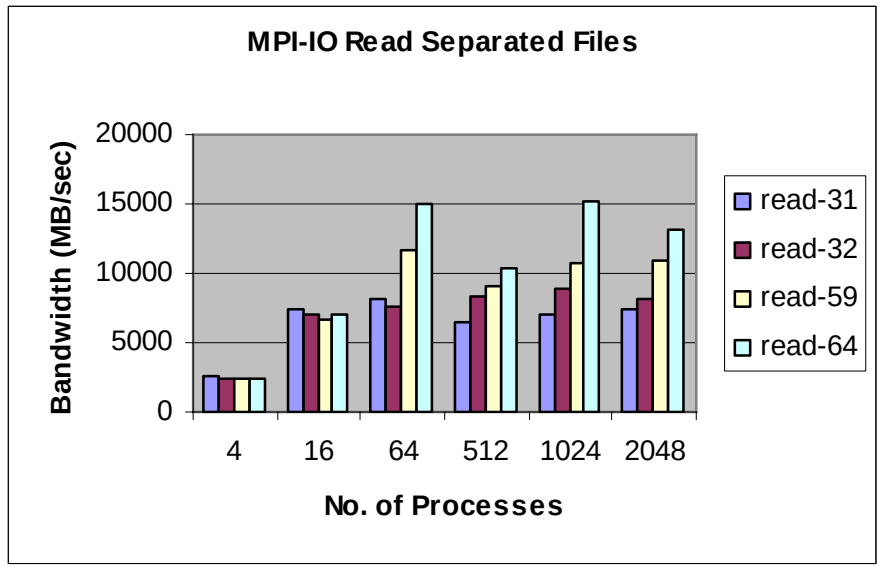




MPI-IO Read/Write – Separated file

- Bandwidth tests
  - Test program: IOR
    - file size (512MB), transfer units 64MB; 31/32/59/64 stripes
  - Performance with separated files is worse than a share file
  - Prime stripes do not help for writes



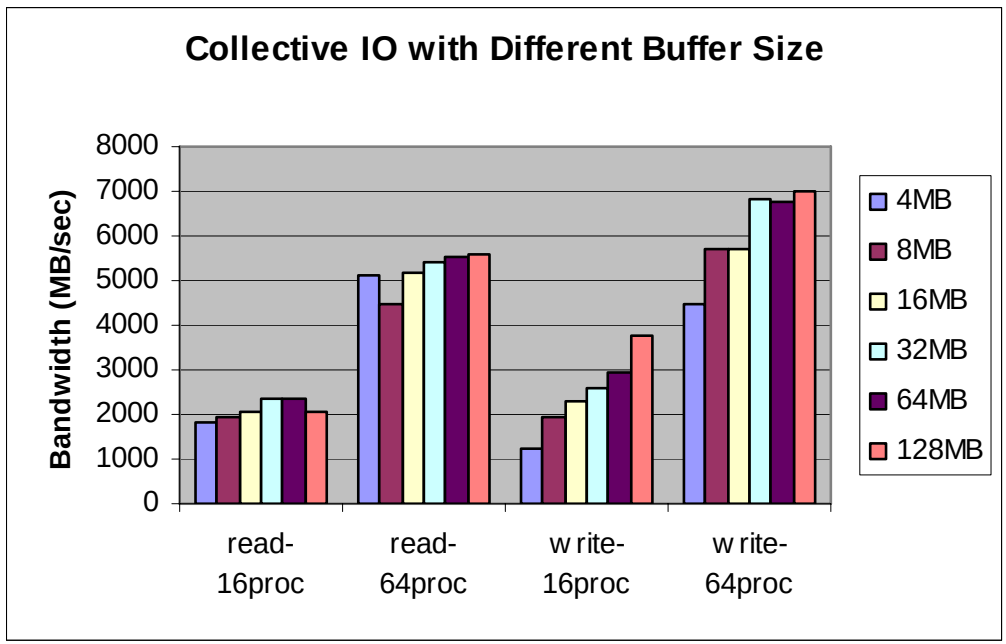


Presentation Outline

- Overview of Cray XT Parallel IO
- POSIX Read/Write
- MPI-IO Independent IO
- Collective IO
- Sample Performance Tuning
  - BT-IO
  - Flash IO (HDF5)

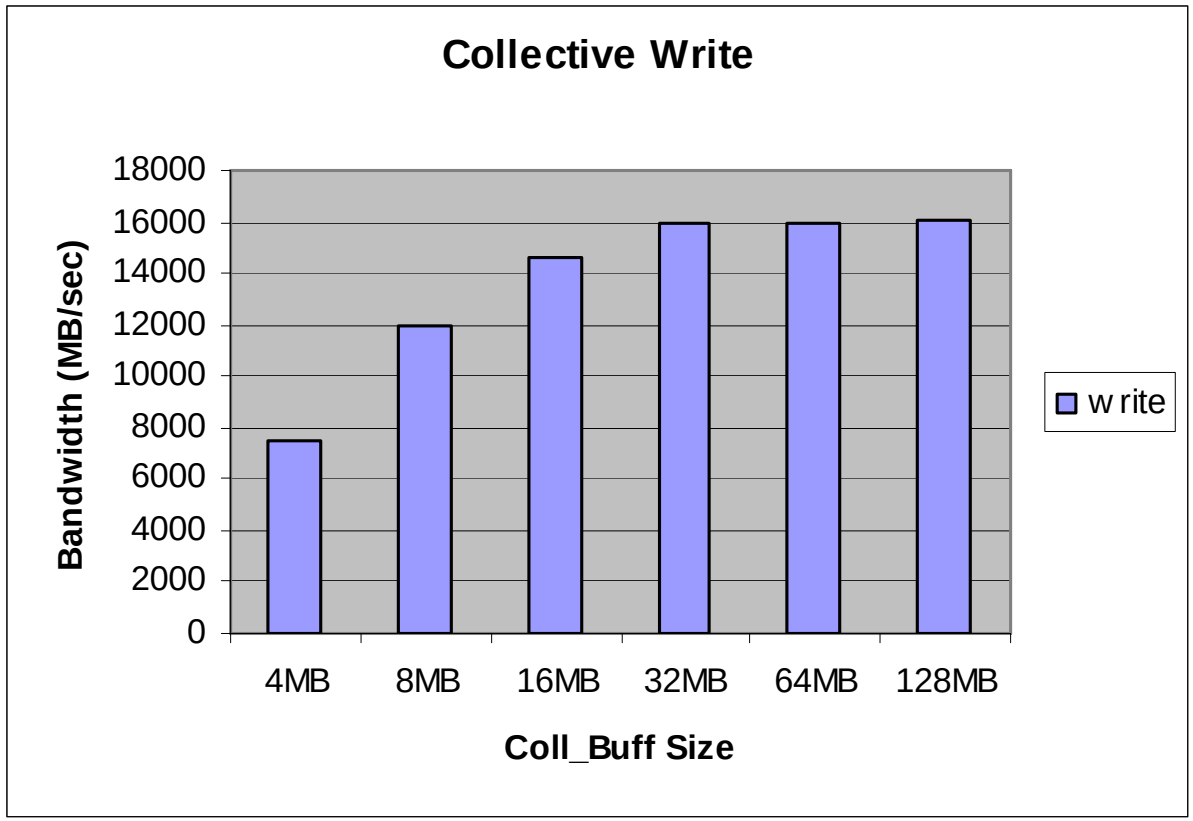
Collective IO

- Collective IO tests
  - Test program: MPI-Tile-IO
    - Interleaved Tile Read/Write to a file striped to 64 OSTs
  - Collective write performs better than collective read



Collective IO -- continued

- Collective write needs large collective buffer
  - MPI-Tile-IO, 64 stripes
  - 32MB is sufficient

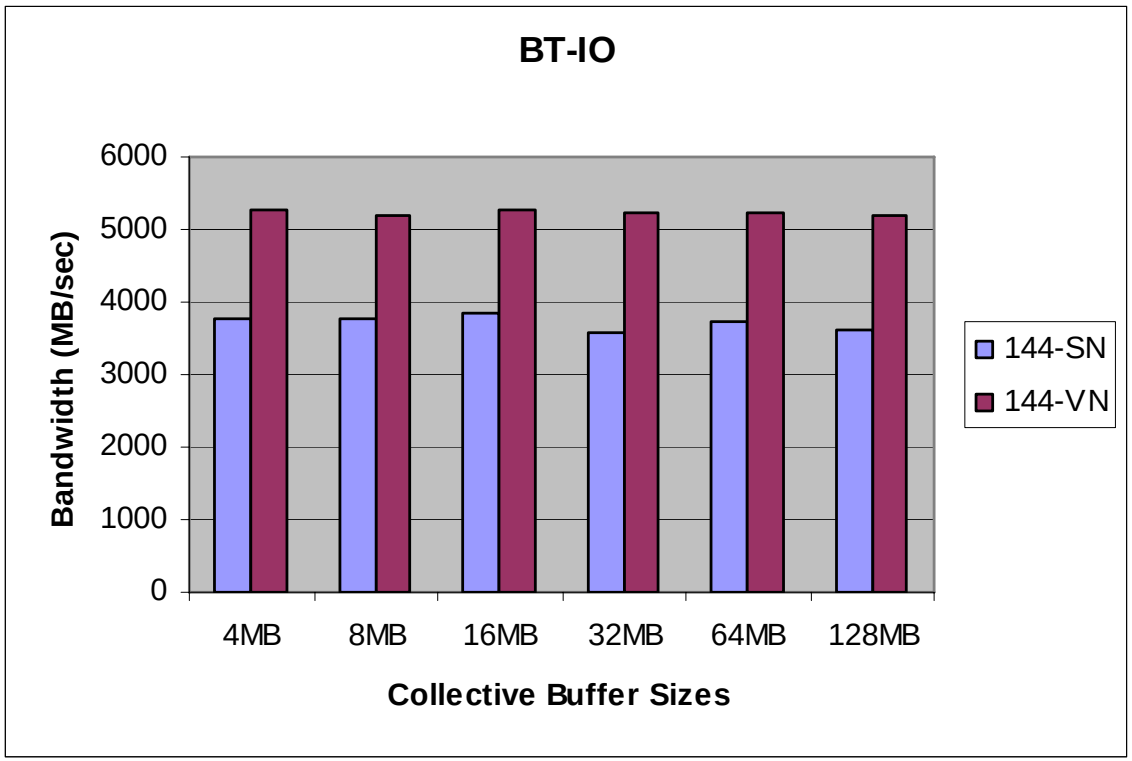


Presentation Outline

- Overview of Cray XT Parallel IO
- POSIX Read/Write
- MPI-IO Independent IO
- Collective IO
- Sample Performance Tuning
  - BT-IO
  - Flash IO (HDF5)

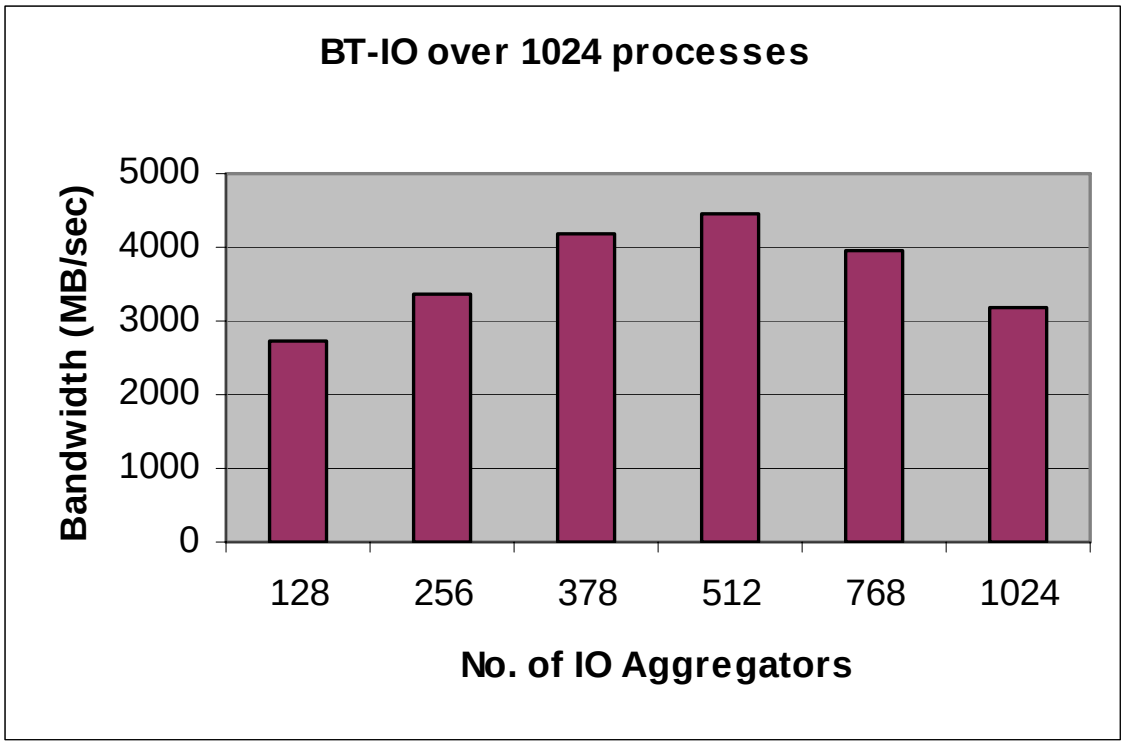
BT-IO – Different Buffer Size

- BT-IO does not require large collective buffer
  - BT-IO is the IO benchmark in the NPB suite. Many small read/writes for data partitioning
  - Used the full mode implementation
  - Collective buffer of 16MB is sufficient
  - Using dual-core (VN) mode is beneficial to collective IO



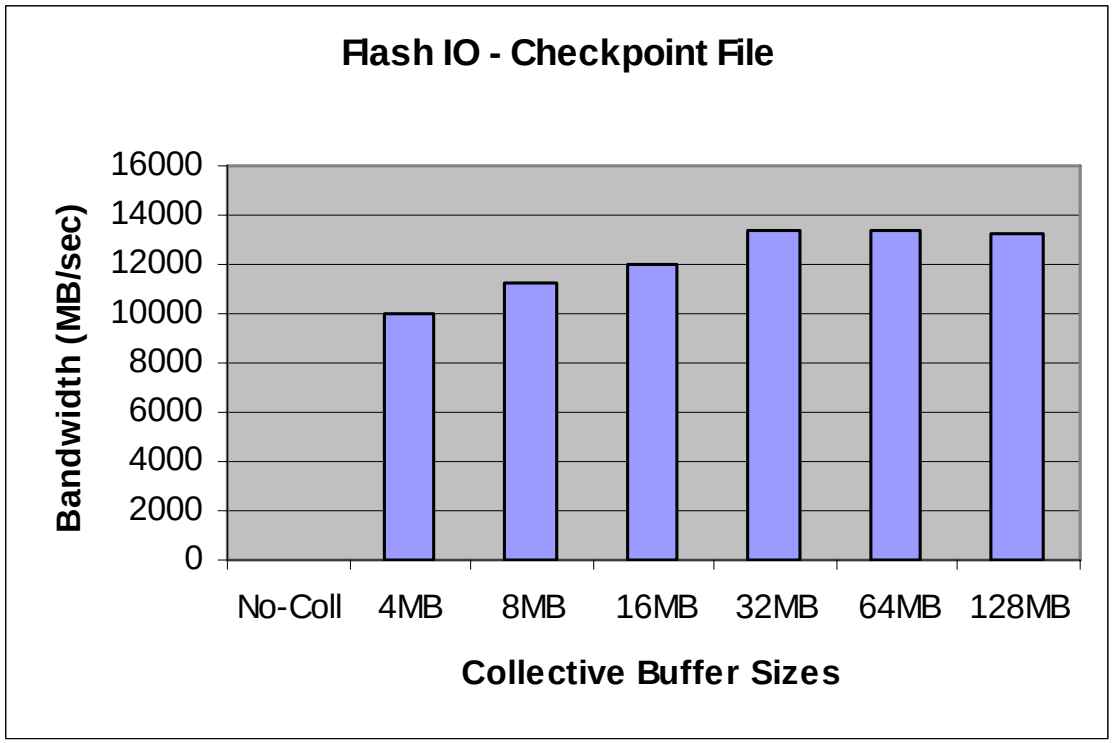
BT-IO – Number of IO aggregators

- The number of IO aggregators in BT-IO needs to be tuned
  - Writing to a file striped to 64 OSTs
  - 512 is optimal for a 1024-proc BT-IO program



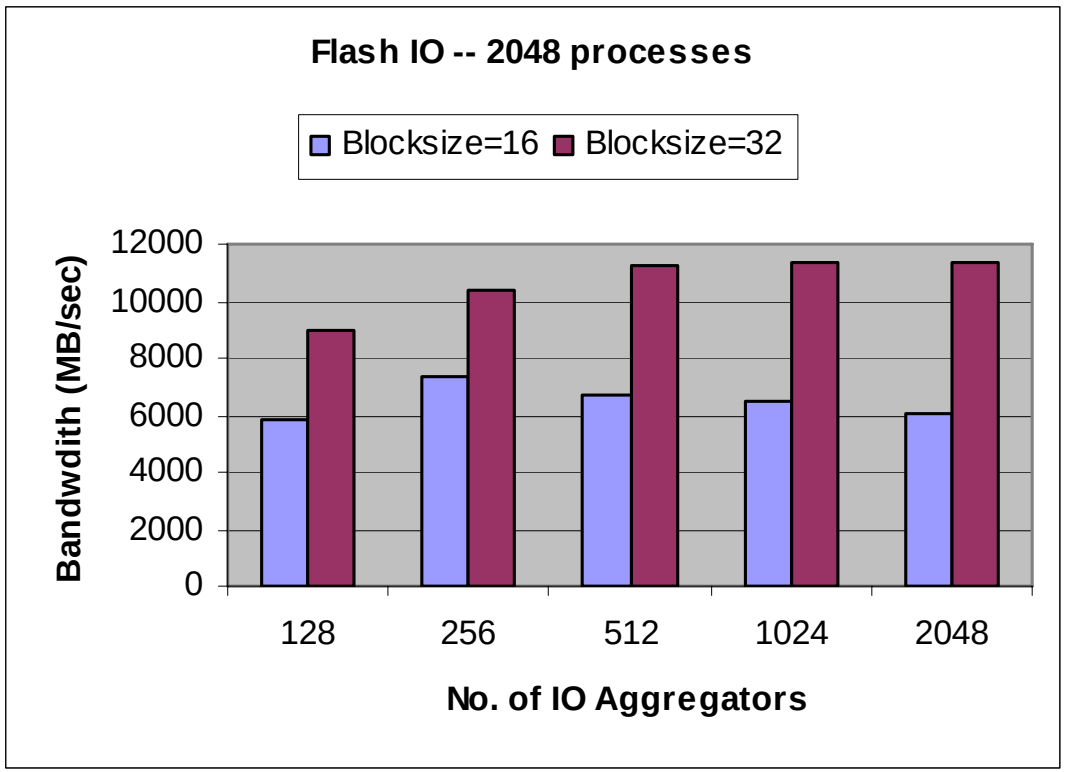
Flash IO – Collective IO and Buffer Size

- Flash IO is the IO kernel of the Flash application
- Focus on the performance of the biggest file: checkpoint file
- Important to enable collective IO for Flash IO
- Collective Buffer size: 32MB is sufficient



Flash IO – IO Aggregation

- Important to adjust the number of IO aggregators
- 256 or 512 IO aggregators would be sufficient, particularly for flash IO with smaller blocksize



Conclusions

- Evaluated the performance of Parallel IO stack
- Demonstrated the importance of tuning parallel IO
  - Collective IO
  - Collective Buffering
  - IO Aggregation
- Parallel IO over Cray XT
  - Use Shared file for Parallel IO
  - Enable collective IO for scientific benchmarks with interleaved IO pattern :
    - Examples: BT-IO and Flash -IO
    - Adjust your collective buffer size
      - 16-32MB recommended for 64OSTs
  - Delegate IO to smaller number of processes (256-512)

Acknowledgement