McCormick
Northwestern Engineering
Electrical Engineering & Computer Science

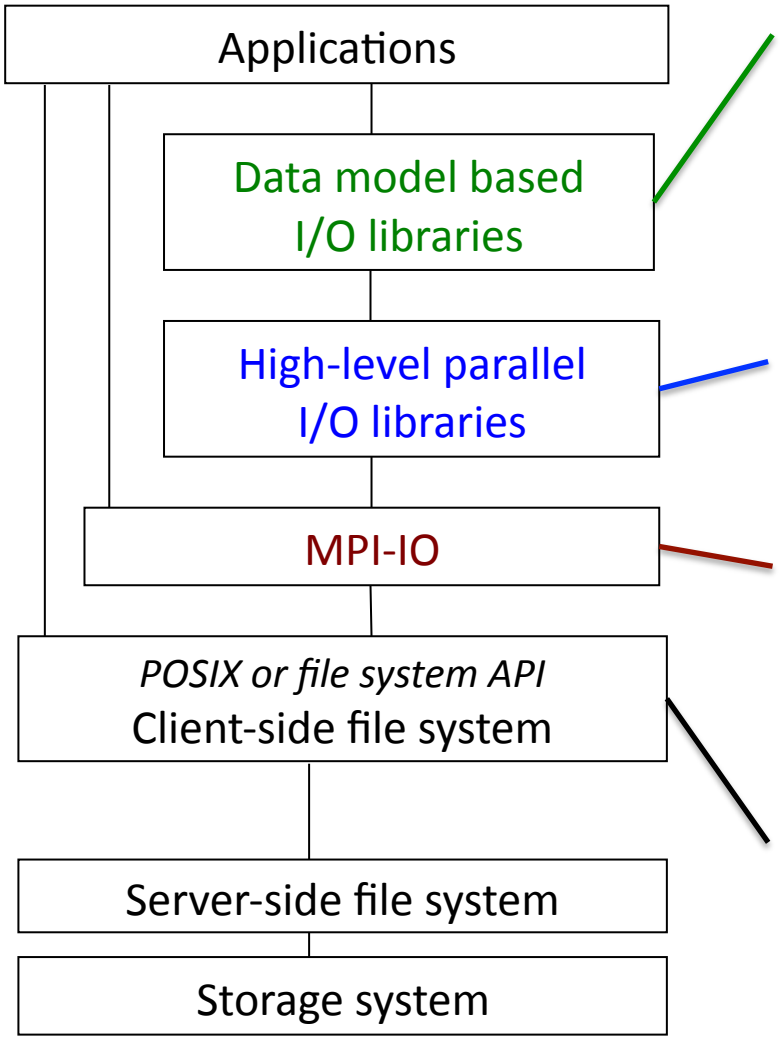
I/O Software and Data

Alok Choudhary

Northwestern University

Large Scale Computing and Storage Requirements for Advanced Scientific Computing Research, ASCR / NERSC Workshop, January 5-6, 2011

I/O software stack in HPC



- Domain-specific
- Encapsulate multiple lower-level I/O interfaces (e.g. HDF5, netCDF)
- Examples:
  - PIO: Community Climate System Model (CCSM)
  - GIO: Global Cloud Resolve Model (GCRM)
- Portable file format, self-describing, metadata-rich
- Examples:
  - PnetCDF, netCDF4, HDF5
- MPI standard, uniform API to all file systems
- Lower-level parallel I/O optimizations
  - Coordinate processes to rearrange I/O requests
  - Collective I/O, data sieving, caching, lock alignment
- Parallel file systems
  - Lustre, PVFS2, GPFS, PanFS
  - File striping, caching, consistency controls, scalable metadata operations, data reliability, fault tolerance

I/O resources in NERSC

- NERSC Global Filesystem (NGF), an IBM GPFS
  - HPC
    - Franklin (Cray XT4), Hopper (Cray XT5), Hopper II (Cray XE6), Carver (IBM)
  - Analytics clusters
    - PDSF (Linux), Euclid (Sun)
  - Others
    - Tesla/Tuning, Dirac (GPU cluster), Magellan (Cloud)
- Parallel file systems
  - Lustre, GPFS
- Archival Storage (HPSS)
- Consulting team
  - Very helpful for answering I/O related questions, including hardware configuration, software availability, run-time environment setup, system performance numbers, communication with Cray, etc.

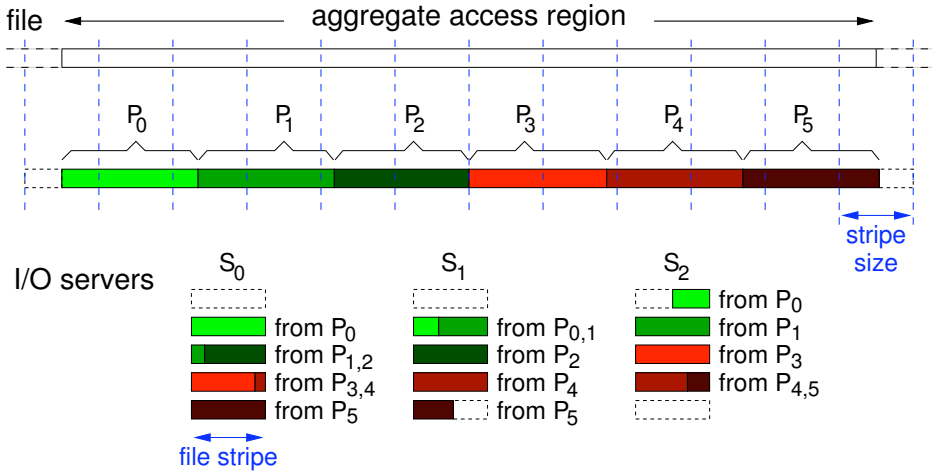
I/O software at NERSC

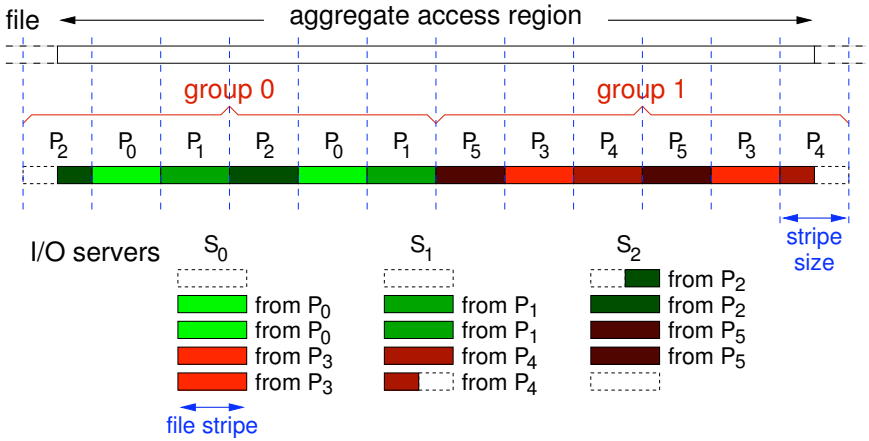
- High-level I/O libraries
  - HDF5, netCDF4, Parallel netCDF
- I/O tracing libraries
  - IPMIO (profiling POSIX calls)
- Parallel I/O middleware
  - MPI-IO (Cray's implementation)
  - ROMIO with NWU's optimization (file domain alignment)
- Lustre parallel file systems
  - User customizable striping configuration, an important feature for I/O developers

Case studies

- MPI-IO file domain alignments
  - Reorganize I/O requests to match the Lustre locking protocol
  - Significantly improve performance on Franklin
- I/O delegation
  - Additional set of compute nodes to enable caching, prefetching, aggregation
  - I/O requests are forwarded from application processes to delegates
  - Boost independent I/O competitively to collective I/O
- Parallel netCDF non-blocking I/O
  - Aggregates many small requests for better bandwidths
- Data-model based I/O library
  - Next generation high-level I/O library
  - Supports data models (seven dwarfs) with new file formats and data layouts

MPI-IO file domain alignments

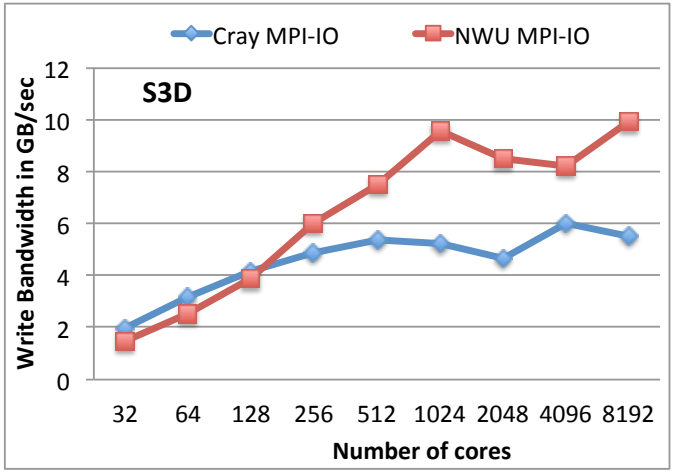


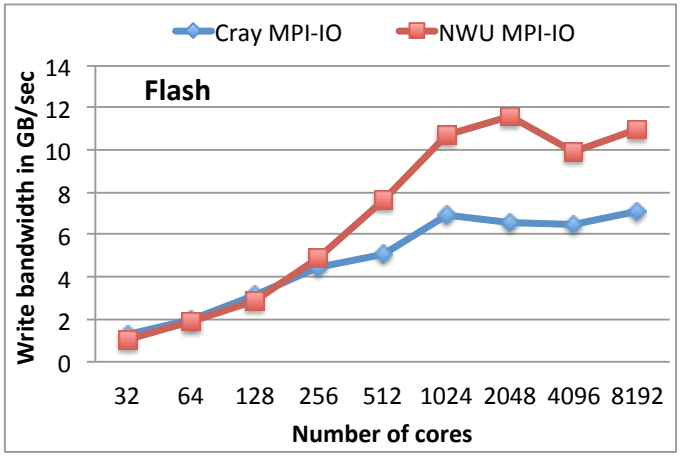


- Lustre
  - File system uses locks to keep data consistent
- File domain alignment in collective I/O
  - Minimize the number of clients accessing each file server
- Two implementations
  - NWU (single-stage) published in SC08*
  - Cray MPI-IO (multi-stage), available in June 2009

*W. Liao, and A. Choudhary. Dynamically Adapting File Domain Partitioning Methods for Collective I/O Based on Underlying Parallel File System Locking Protocols, SC 2008

Improvement from file domain alignment

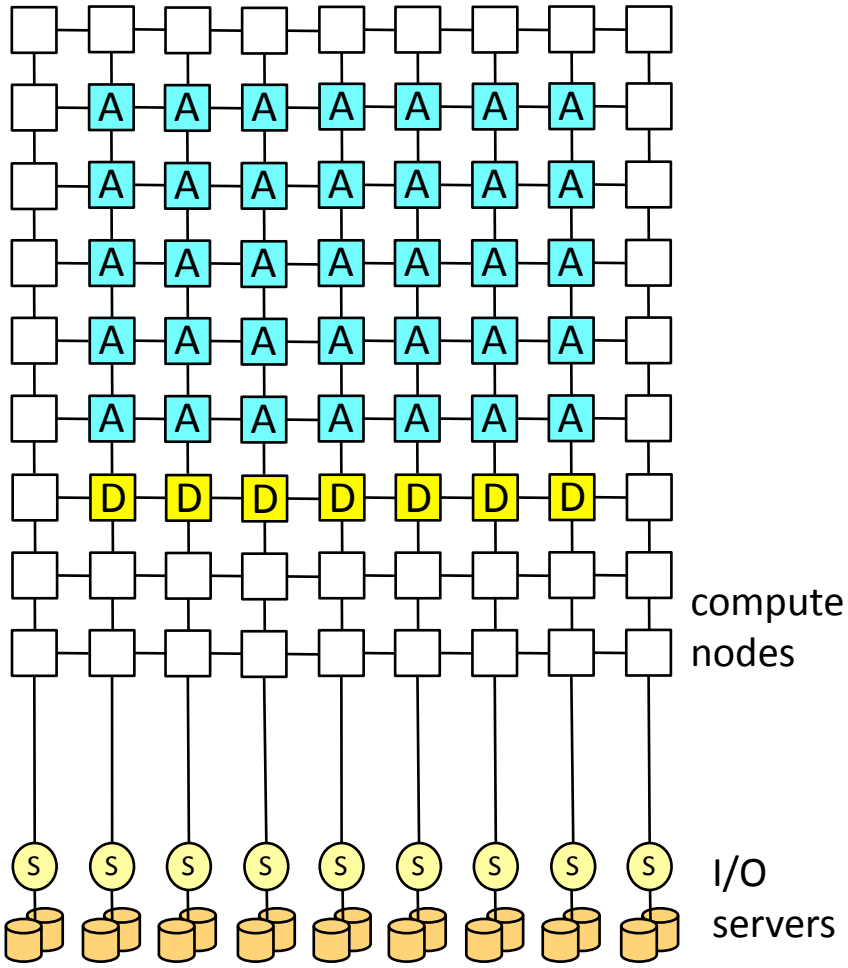




- Franklin @ NERSC
  - Compare Cray's and NWU's MPI-IO implementations
  - Measured peak for write is 16GB/sec
- S3D
  - Combustion application from Sandia Lab.
  - Four global arrays: two 3D and two 4D
  - Each process subarray size 50x50x50
- Flash
  - Astrophysics application from U. of Chicago
  - I/O method: HDF5
  - Each process writes 80~82 32x32x32 arrays

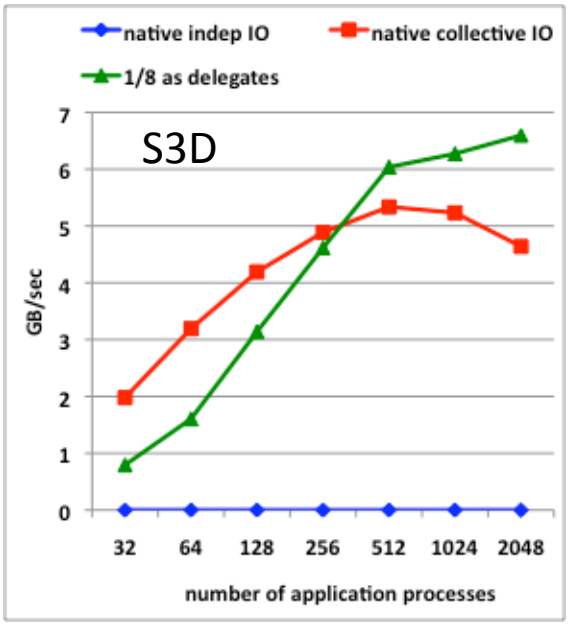
I/O delegation

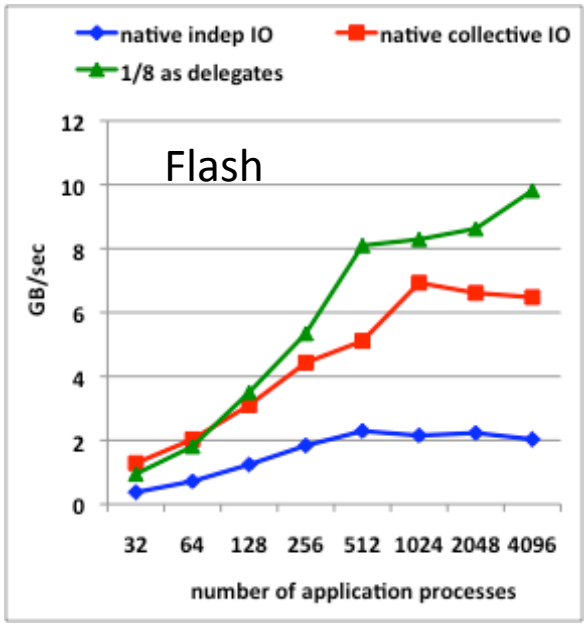
A MPI application nodes D I/O delegate nodes



- Run inside of MPI-IO
- Run on a small set of additional MPI processes
- All I/O delegates collaborate for better performance
- Related work
  - I/O forwarding developed by Rob Ross's team at ANL

Improvement from I/O delegation





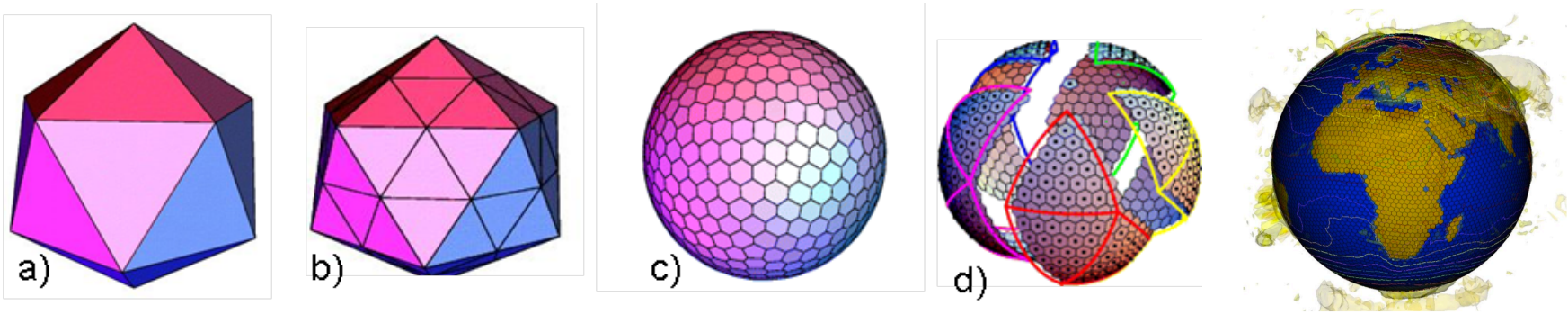
- Franklin
  - Compared with Cray's collective I/O
- I/O delegation
  - 1/8 of application processes as I/O delegates
  - Enhances independent I/O performance to be similar to the collective
  - Using independent I/O eases I/O programming
  - The above results presented in a paper accepted by IEEE TPDS*

*A. Nisar, W. Liao, and A. Choudhary. Scaling Parallel I/O Performance through I/O Delegate and Caching System, SC 2008

PnetCDF

- A parallel I/O library for accessing files in CDF format
  - Version 1.0 released in 2005, now in version 1.2
- Optimizations
  - Built on top of MPI-IO
  - File header and dataset alignment
  - Non-blocking I/O enables aggregation of multiple requests
- Developers
  - NU: Jianwei Li (graduated in 2006), Kui Gao (postdoc), Wei-keng Liao, Alok Choudhary
  - ANL: Rob Ross, Rob Latham, Rajeev Thakur, Bill Gropp
- Recent application collaborators
  - GIO (PNNL), FLASH (U. Chicago), CCSM (NCAR)

GCRM I/O

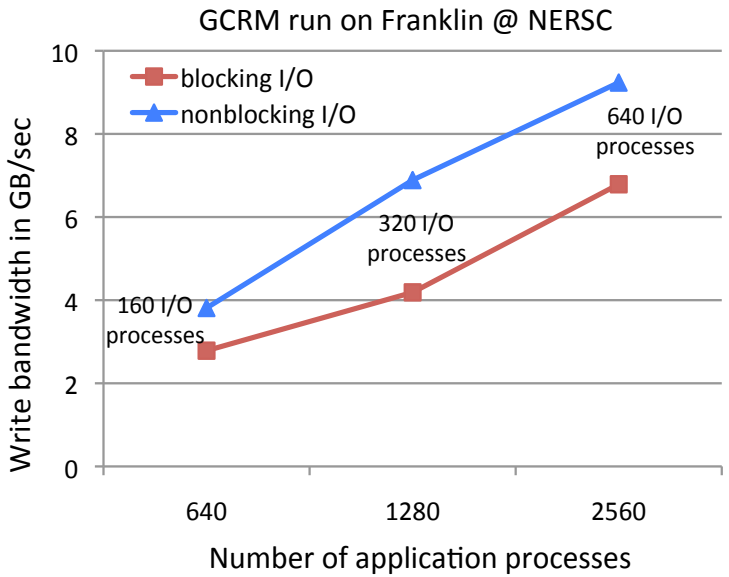


Pictures by courtesy of Karen Schuchardt

- GIO (Geodesic Parallel IO API) developed at PNNL
  - Is an I/O library developed by Karen Schuchardt from PNNL
  - Used by Global Cloud Resolve Model (GCRM) developed at Colorado State University that simulates the global climate
  - Provides user configurable I/O method
    - PnetCDF
    - HDF5
    - NetCDF4

GCRM I/O performance

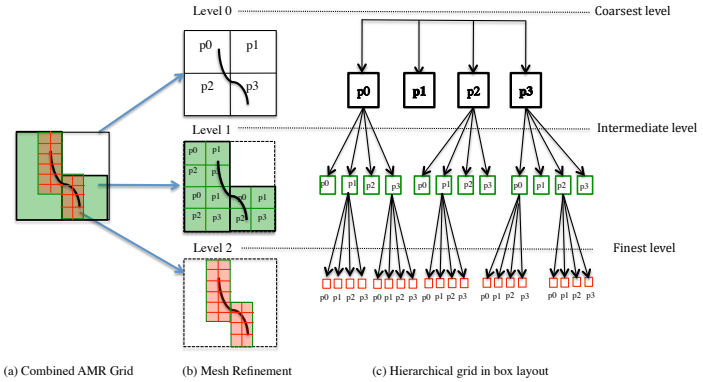
Resolution level	Number of cells	Grid-point spacing
9	2.6 Million	15.6 km
10	10.5 Million	7.8 km
11	41.9 Million	3.9 km



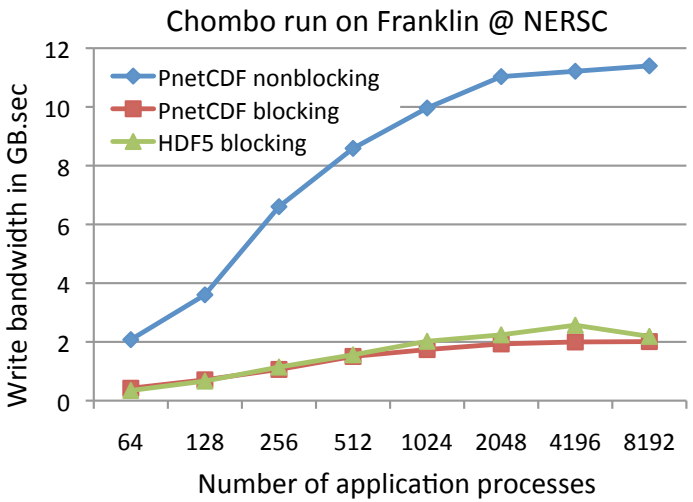
- I/O pattern
  - Each process writes many, noncontiguous data blocks for each variable
- GIO strategies
  - Direct and interleaved messaging methods to exchange I/O requests in order to get better performance
- PnetCDF non-blocking I/O
  - Delay request so small-sized requests can be aggregated into large ones
  - Simplifies the programming task
  - Results were presented in the Workshop on High-Resolution Climate Modeling 2010*

*B. Palmer, K. Schuchardt, A. Koontz, R. Jacob, R. Latham, and W. Liao. IO for High Resolution Climate Models. Workshop on High-Resolution Climate Modeling, 2010

Chombo I/O



A 2-D patch-based hierarchical grid with three levels of refinement on 4 processes

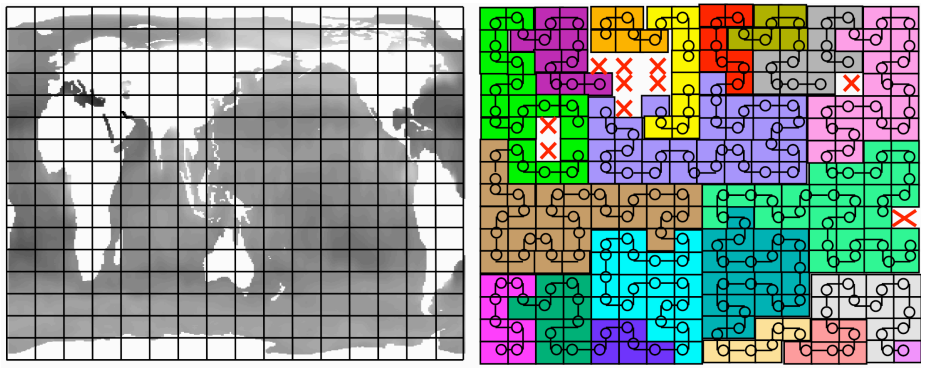


- PDE tool from LBNL
  - Supports block-structured AMR grids
- I/O pattern
  - Array variables are partitioned among a subset of processes
  - Calls MPI independent API as the collective is not feasible
- PnetCDF non-blocking I/O
  - Aggregate requests to multiple variables
  - One collective I/O carries out the aggregated request

Data-model based I/O library (X-stack)

(Pis Alok Choudhary, Wei-Keng Liao, Northwestern; Rob Ross, Tim Tautges, ANL; Nagiza Samatova, NCSU; Quincy Koziol, HDF Group)

- DAMSEL: next generation of high-level I/O library
- Supports various data models
  - Describe unstructured data relationships: trees, graph-based, space-filling curve, etc.
  - Scalable I/O for irregular distributed data objects
  - More sophisticate data query API
- Virtual filing
  - A file container of multiple files appears as a single file
  - Balance concurrent access (to reduce contention) and the number of files created (to ease file manageability)



Space filling curves are used in climate codes when partitioning the grid to improve scalability. Image courtesy John Dennis (NCAR).

Computational motifs

Table 1: The expanded list of Computational Motifs (Dwarfs). Here, we have identified data models used in the motifs and provided illustrative examples. Some codes employ more than one motif. This project focuses on the top six (blue).

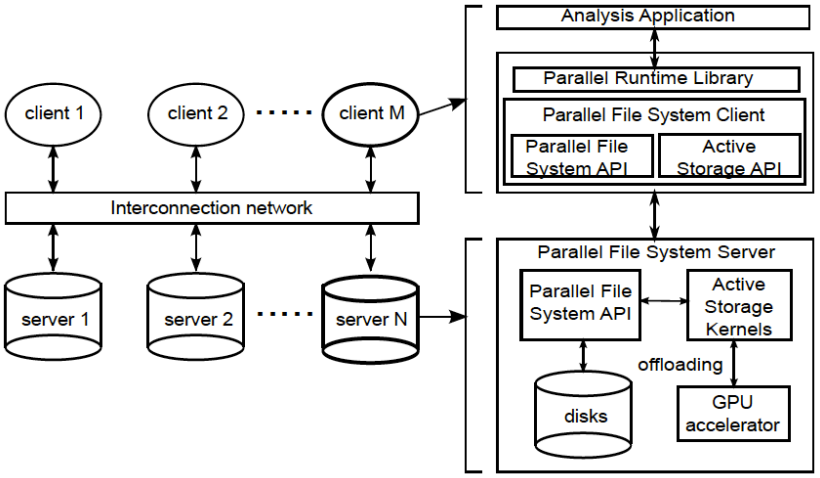
Motif	Data Model/ Data Structure	Examples
Dense Linear Algebra	**a**	**BLAS, LAPACK, ScaLAPACK, Matlab, S3D**
Sparse Linear Algebra	**f**	**OSKI, SuperLU, SpMV**
Spectral Methods	**a**	**FFT, Nek5000 (Nuclear Energy)**
N-Body Methods	**b, e, j**	**Molecular Dynamics, NN-Search**
Structured Grids (+ AMR)	**a, b, c**	**FLASH (Astrophysics), Chombo-based codes**
Unstructured Grids (+ AMR)	**c**	**UNIC, Phasta, SELFE numerical tsunami models**
Monte Carlo, MapReduce	**a-l**	GFMC, EM, POV-Ray
Combinational Logic	**g, i**	RSA encryption, FastBit
Graph Traversal	**f, h**	S3D, Boost Graph Library (BGL), C4.5
Dynamic Programming	**a**	Smith-Waterman
String Searches	**d, e**	BLAST, HMMER
Backtrack and Branch-and-Bound	**f, i, g**	Clique, Kernel regression
Probabilistic Graphical Models	**h, k**	BBN, HMM, CRF
Finite State Machines	**l**	Collision detection

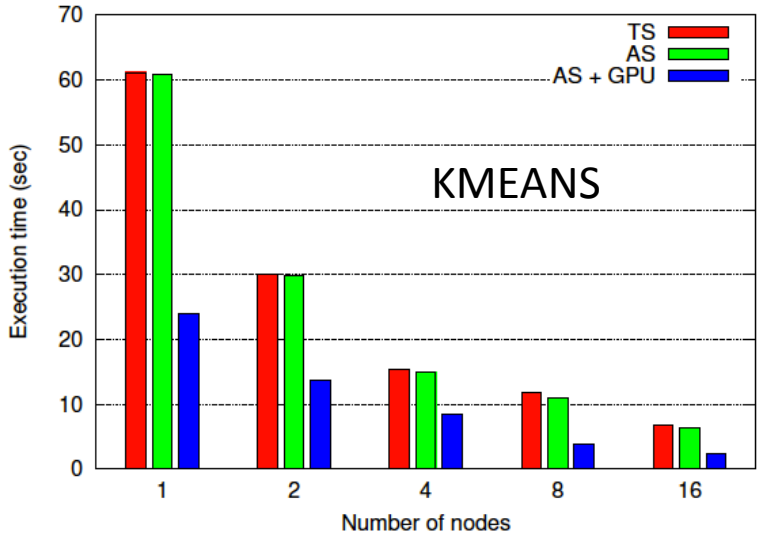
a–Multidimensional array, e.g., dense matrix in 2D; **b**–Point- or region-based quadtree, octree, compressed octree, or hyperoctree; **c**–Lattice model; **d**–Suffix tree, suffix array; **e**–R-tree, B-tree, X-tree, and their variants; **f**–Sparse matrix, e.g., block compressed sparse row (BCSR); **g**–Bitmap index, bitvector; **h**–Direct Acyclic Graph (DAG); **i**–Hash table, grid file; **j**–K-d tree; **k**–Junction tree; **l**–Transition table, Petri net.

Future

- Challenges:
  - Reads in data analysis, as most of the case only subsets of data are read
- Hardware accelerators
  - GPU for on-line data compression, analysis
- Faster storage device
  - SSD as a read cache at compute nodes or I/O servers

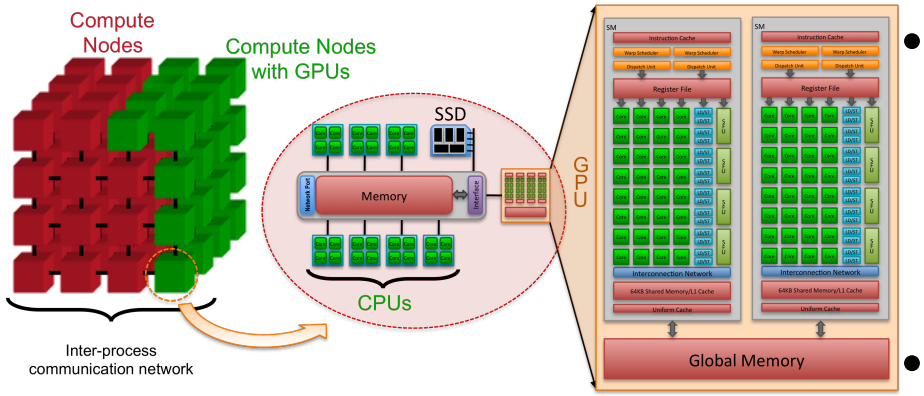
Active storage

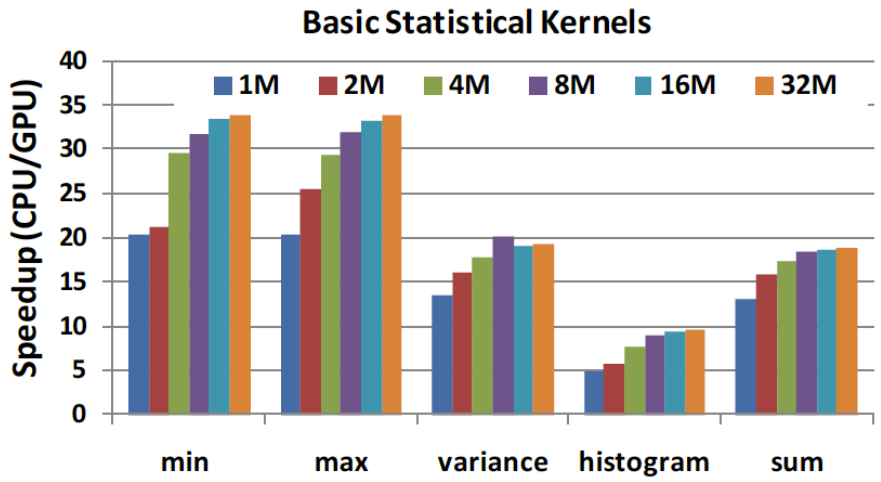




- Offloading I/O intensive computation to the file servers
- If servers were equipped GPUs, the operations can run faster
  - TS: traditional storage, task runs on clients
  - AS: active storage, task runs on server's CPU
  - AS + GPU: task runs on server's GPU

GPU to accelerate data analytics





- Our I/O delegation work demonstrates that set-aside processes improve I/O performance
- Delegate processes can also be used for off-loading data intensive computation
- HPC with a subset of compute nodes equipped with GPU and SSD provides a rich experimental platform

Publications

- Alok Choudhary, Wei-keng Liao, Kui Gao, Arifa Nisar, Robert Ross, Rajeev Thakur, and Robert Latham. **Scalable I/O and Analytics.** In the *Journal of Physics: Conference Series*, Volume 180, Number 012048 (10 pp), August 2009.
- Kui Gao, Wei-keng Liao, Arifa Nisar, Alok Choudhary, Robert Ross, and Robert Latham. **Using Subfiling to Improve Programming Flexibility and Performance of Parallel Shared-file I/O.** In the Proceedings of *the International Conference on Parallel Processing*, Vienna, Austria, September 2009.
- Kui Gao, Wei-keng Liao, Alok Choudhary, Robert Ross, and Robert Latham. **Combining I/O Operations for Multiple Array Variables in Parallel NetCDF.** In the Proceedings of *the Workshop on Interfaces and Archi- tectures for Scientific Data Storage*, held in conjunction with the the IEEE Cluster Conference, New Orleans, Louisiana, September 2009.
- B. Palmer, K. Schuchardt, A. Koontz, R. Jacob, R. Latham, and W. Liao. **IO for High Resolution Climate Models**. Workshop on High-Resolution Climate Modeling, 2010
- Arifa Nisar, Wei-keng Liao, and Alok Choudhary. **Delegation-based I/O Software Architecture for High Performance Computing Systems**. Accepted by IEEE TPDS, 2010.

I/O resources from NERSC

- NERSC Global Filesystem (NGF), an IBM GPFS
  - HPC
    - Franklin (Cray XT4), Hopper (Cray XT5), Hopper II (Cray XE6), Carver (IBM)
  - Analytics clusters
    - PDSF (Linux), Euclid (Sun)
  - Others
    - Tesla/Tuning, Dirac (GPU cluster), Magellan (Cloud)
- Parallel file systems
  - Lustre, GPFS
- Archival Storage (HPSS)
- Consulting team
  - Very helpful for answering I/O related questions, including hardware configuration, software availability, run-time environment setup, system performance numbers, communication with Cray, etc.