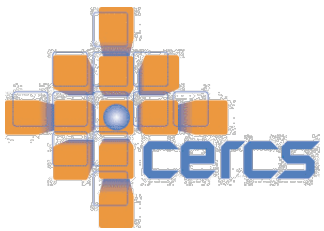
# High Performance I/O

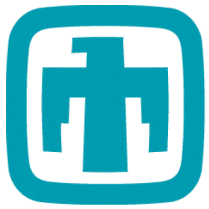
## Addressing Petascale Scientific Needs


Hasan Abbasi
Matthew Wolf
Jay Lofstead
Fang Zheng
Greg Eisenhauer
Karsten Schwan

Scott Klasky
Norbert Podhorszki
Qing Liu

Ron Oldfield

Sandia National Laboratories



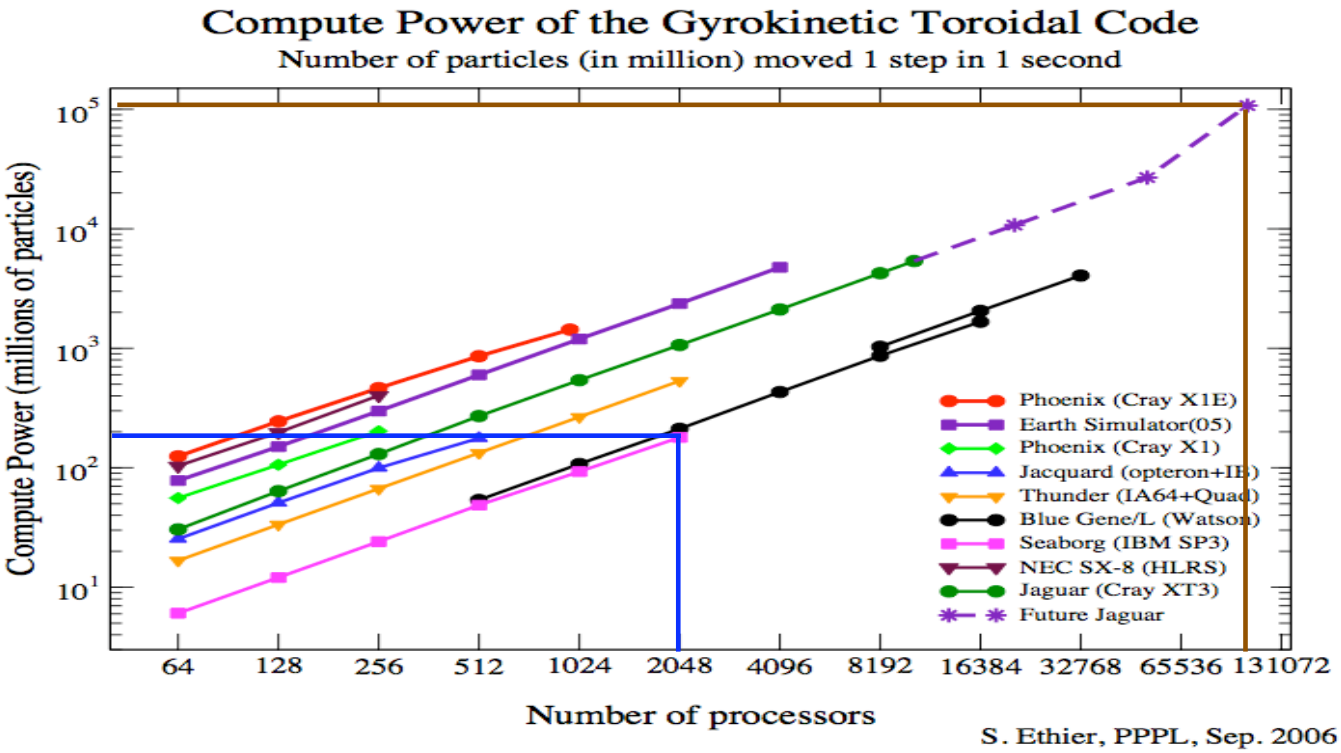
# I/O as a HPC bottleneck

- **Larger Data**

  Requires faster I/O to reduce data output overhead

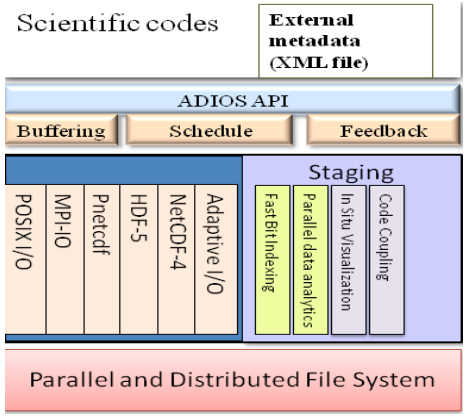
- **Big Data**

  Finding "information" is like hunting for a needle in the hay stack

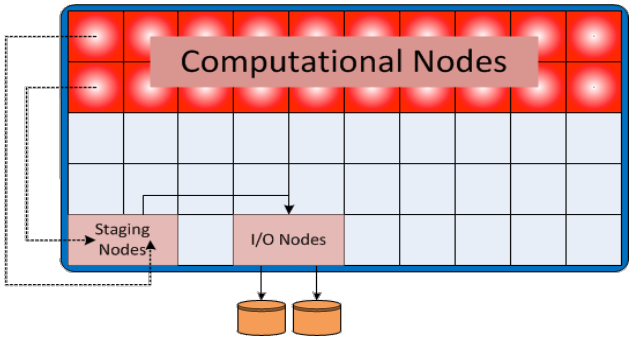


# Outline view

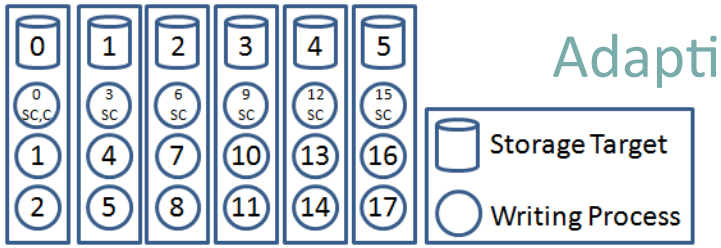
ADIOS



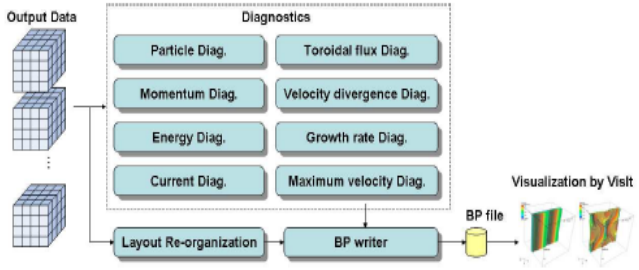

Staging

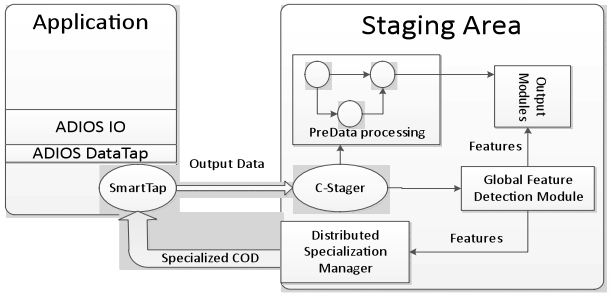



Adaptive I/O



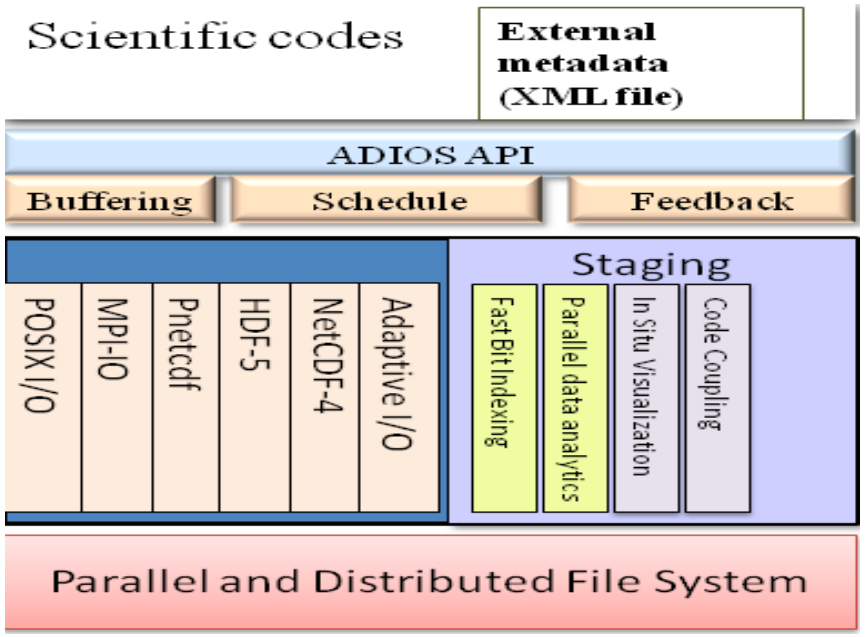

PreDatA




EnStage

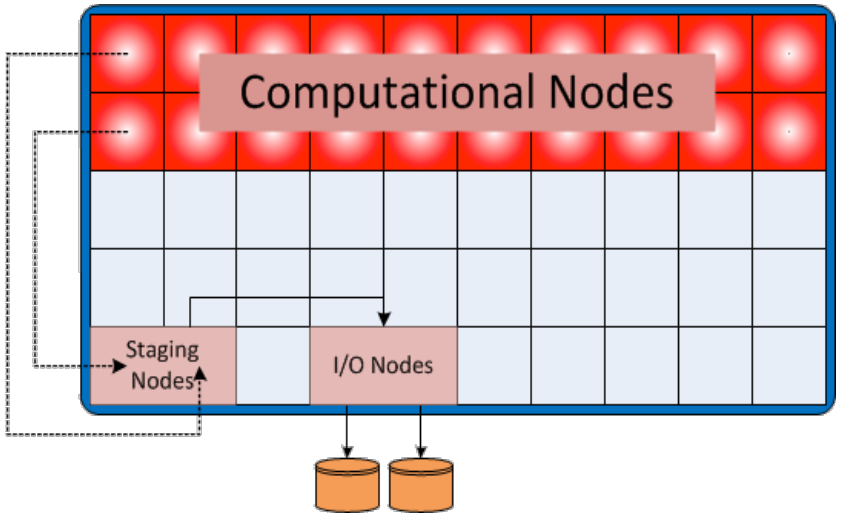
# ADIOS




- Componentized I/O layer
- Multiple transport methods
- Optimize for platforms and application scenarios
- Runtime selection of I/O methods
- New research methods to explore the cutting edge
- Version 1.2 release this summer.


1. *Lofstead, J and Zheng, F and Klasky, S and Schwan, K. "Adaptable, Metadata Rich IO Methods for Portable High Performance IO.", Rome, Italy, May, 2009.*
2. *Lofstead, J and Klasky, S and Schwan K and Podhorszki, N and Jin, C. "Flexible IO and Integration for Scientific Codes Through The Adaptable IO System (ADIOS).", CLADE 2008 at HPDC, Boston, Massachusetts, June, 2008.*
3. *Lofstead, J and Zheng, F and Klasky, S and Schwan, K. Input/Output APIs and Data Organization for High Performance Scientific Computing. , Austin, Texas, November, 2008.*
4. *Milo Polte and Jay Lofstead and John Bent and Garth Gibson and Scott Klasky and Qing Liu and Manish Parashar and Norbert Podhorszki and Karsten Schwan and Meghan Wingate and Matthew Wolf. "...And Eat it Too: High Read Performance in Write-Optimized HPC I/O Middleware File Formats.", In Proceedings of Petascale Data Storage Workshop 2009 at Supercomputing 2009, Portland, Oregon.*

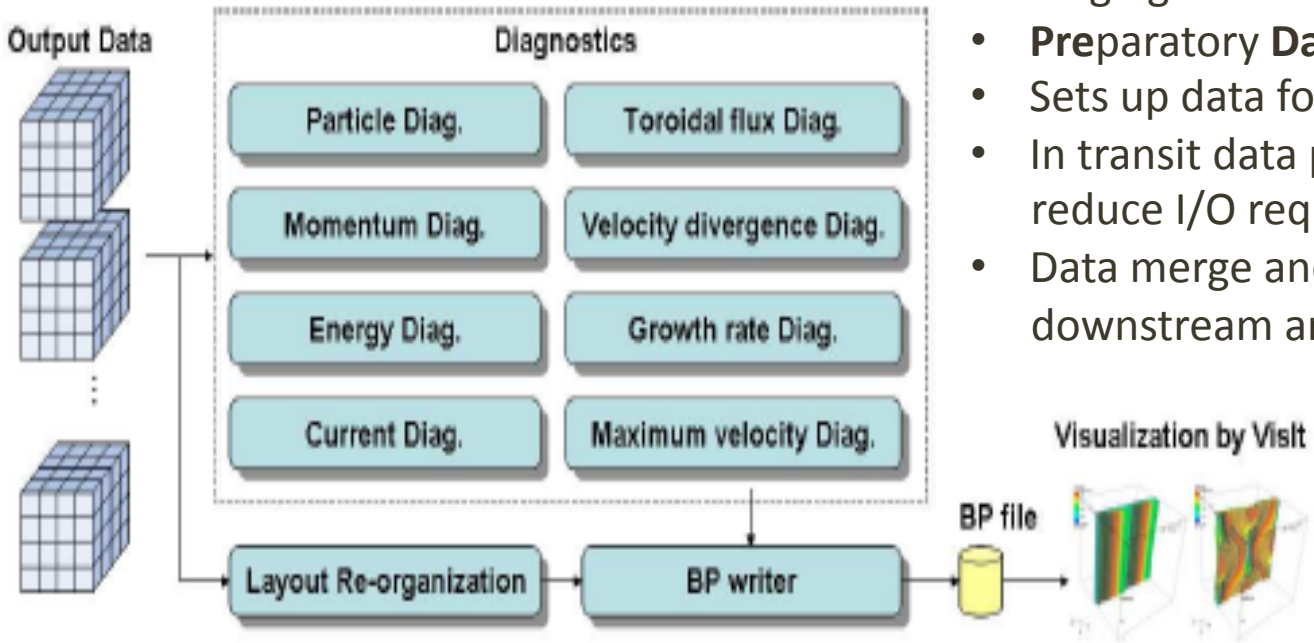
# Staging




- Use additional resources in the compute node
- ADIOS staging method – included in version 1.2 release
- High performance asynchronous output
- State aware schedulers for limiting interference


1. *Abbasi, H., Wolf, M., Eisenhauer, G., Klasky, S., Schwan, K., and Zheng, F. 2009. DataStager: scalable data staging services for petascale applications. In Proceedings of the 18th ACM international Symposium on High Performance Distributed Computing (Garching, Germany, June 11 - 13, 2009). HPDC '09.*
2. *Hasan Abbasi, Jay Lofstead, Fang Zheng, Scott Klasky, Karsten Schwan, Matthew Wolf. "Extending I/O through High Performance Data Services." Cluster Computing 2009, New Orleans, LA. August 2009.*
3. *Julian Cummings, Alexander Sim, Arie Shoshani, Jay Lofstead, Karsten Schwan, Ciprian Docan, Manish Parashar, Scott Klasky, Norbert Podhorszki and Roselyne Barreto. "EFFIS: an End-to-end Framework for Fusion Integrated Simulation". PDP 2010 - Th 18th Euromicro International Conference on Parallel, Distributed and Network-Based Computing, February 2010, Pisa, Italy*

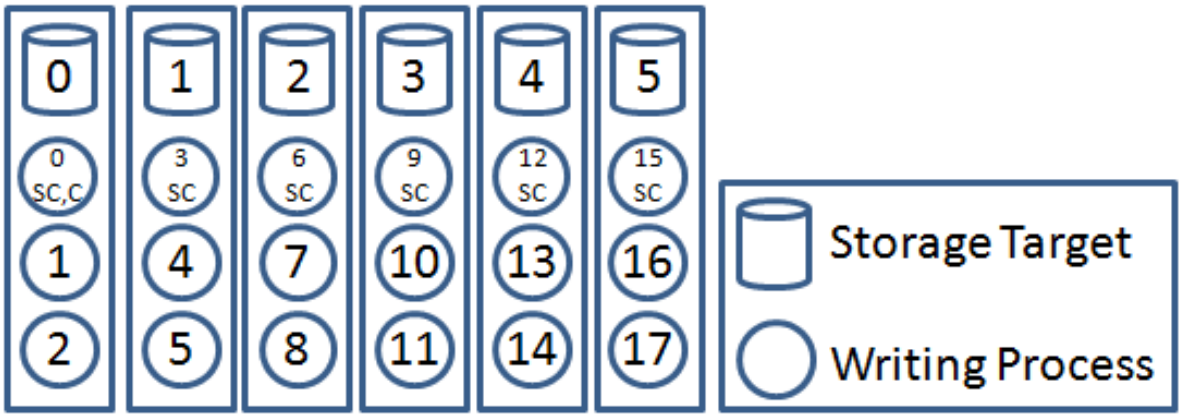
# PreDatA




- Staging Area computations
- **Pre**paratory **Dat**a **A**nalytics
- Sets up data for future analytics
- In transit data processing to reduce I/O requirements
- Data merge and sort aids downstream analysis

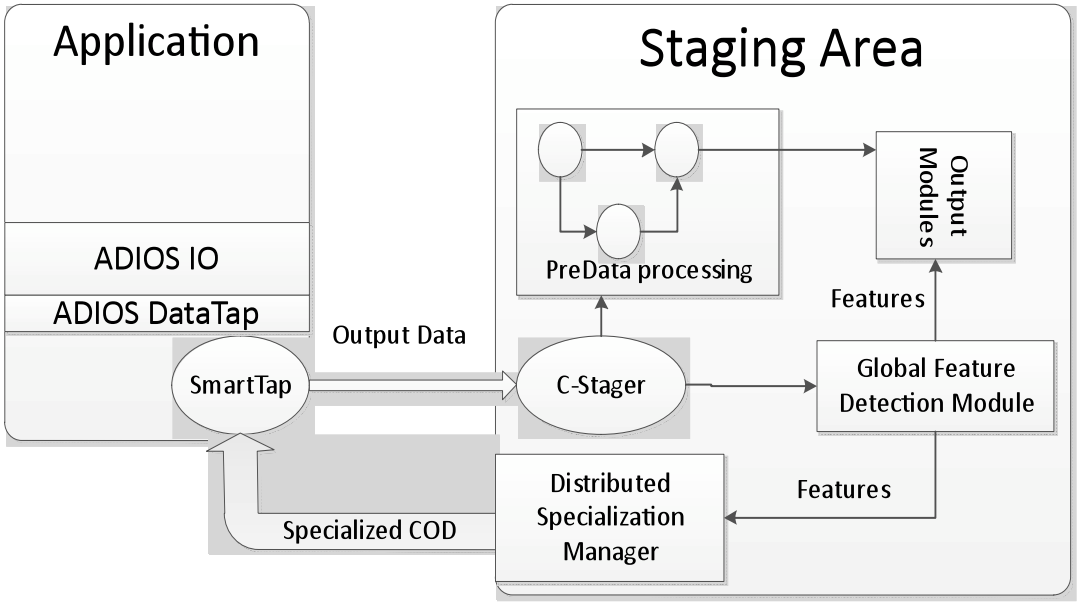

1. *Fang Zheng, Hasan Abbasi, Ciprian Docan, Jay Lofstead, Scott Klasky, Qing Liu, Manish Parashar, Norbert Podhorszki, Karsten Schwan, Matthew Wolf. "PreDatA - Preparatory Data Analytics on Peta-Scale Machines". In Proceedings of 24th IEEE International Parallel and Distributed Processing Symposium. Atlanta, GA. April 2010.*

# Adaptive I/O



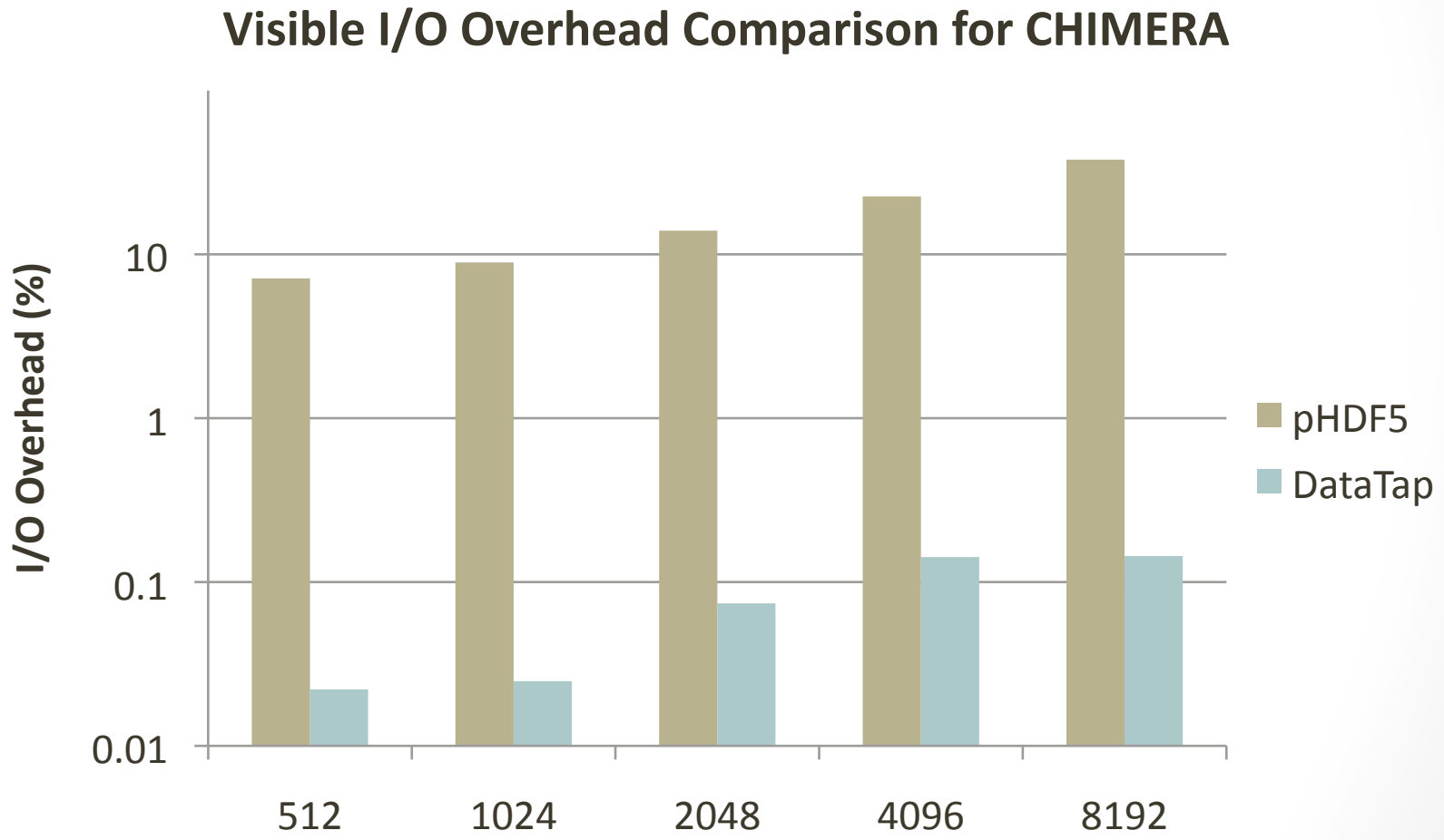

- Large shared storage is useful, but at a cost of unpredictability
  - Multiple users, not even on the same machine
  - Unpredictable hardware behavior
- Upgrade I/O methods so that they can monitor this behavior and adapt the striping of the writes.
  - Requires some additional changes to file formats for full benefit

# EnStage



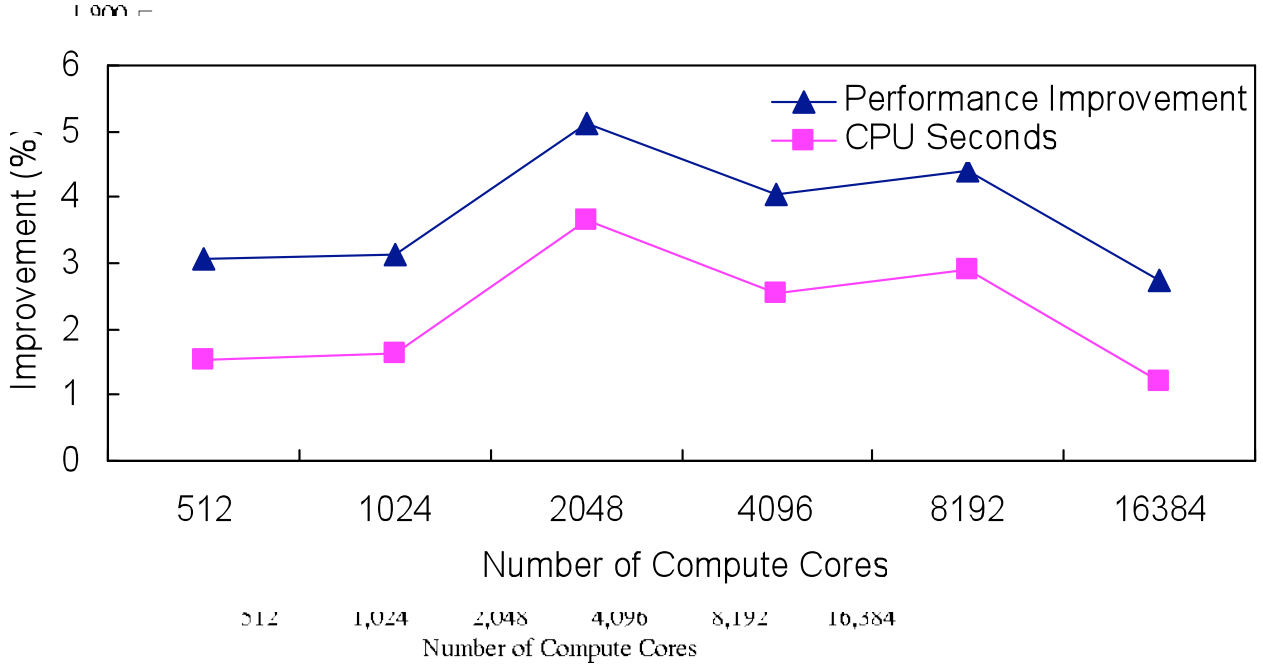

- Extends the staging concept to allow computation within the application
- Used C-o-D with binary code generation to flexibly move computation
- Global operations are performed in the staging area
- Feature extraction and function specialization enable pseudo-collective operations in independent SmartTaps

# Staging advantage



# PreDatA performance

- GTC (all three operations + simulation + BP writer)
  - Performance & Cost

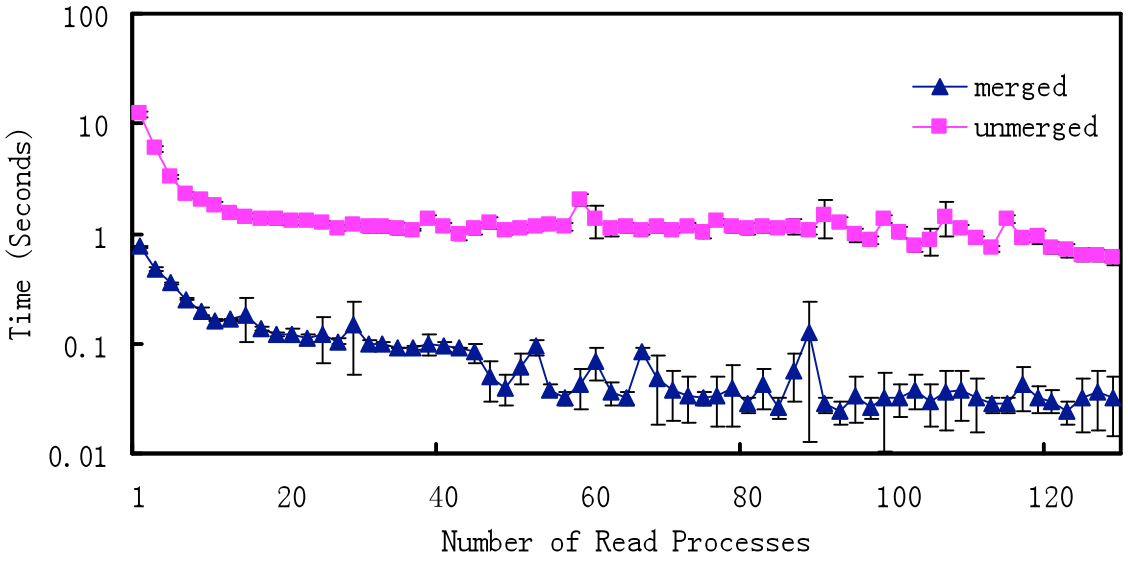



In-Staging approach improve total simulation time by 2.7~5% by hiding I/O and operation latency

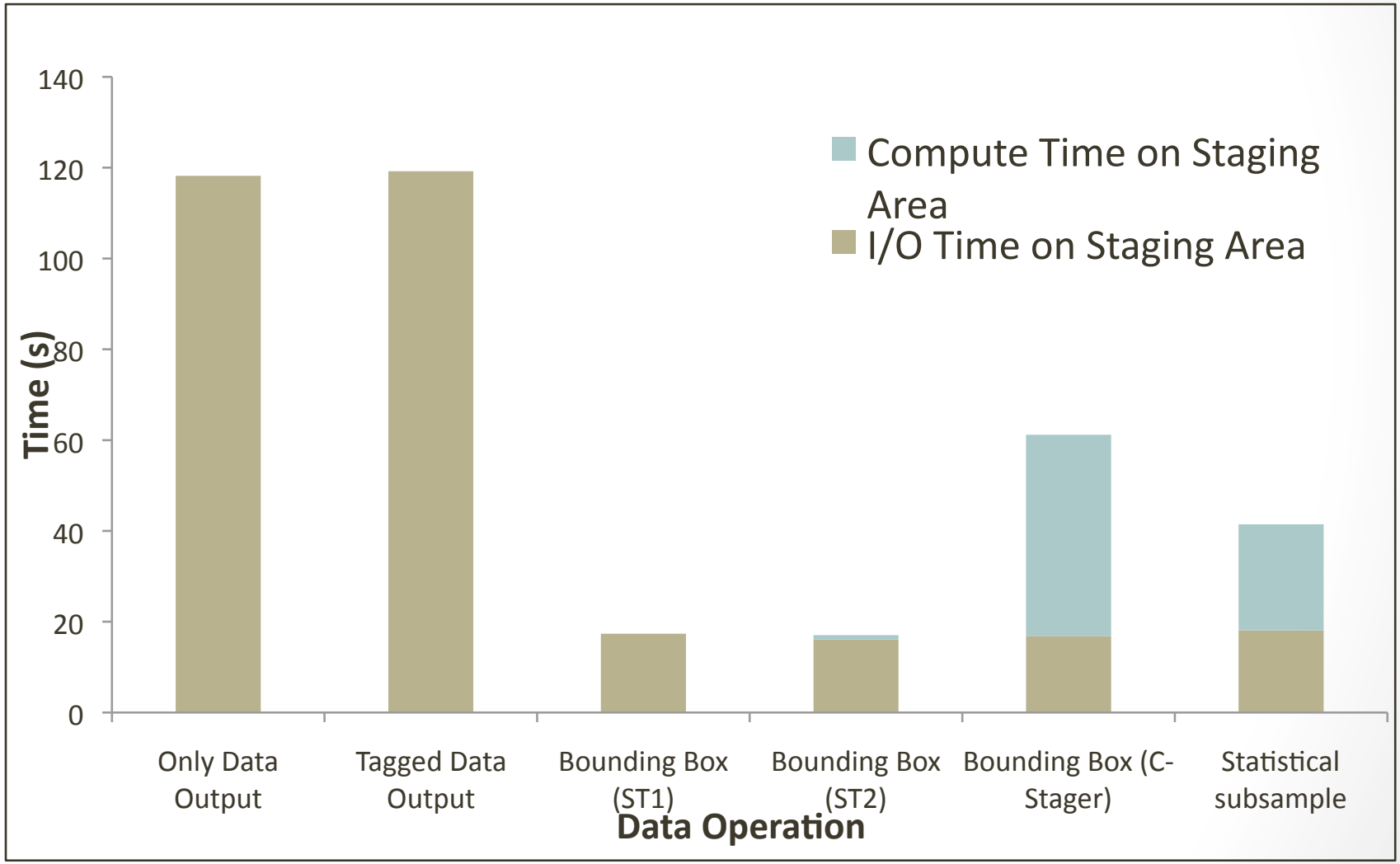
Also 0.7~3% improvement in resource usage (CPU seconds)

# Read Improvements

- Pixie3D:
  - Read performance:
    - Read one array out of BP file generated by 4096-core simulations (merged vs. unmerged)



# Staging Area utilization



# Adaptive Performance

- New adaptive method meant to handle the variability of the writes.