

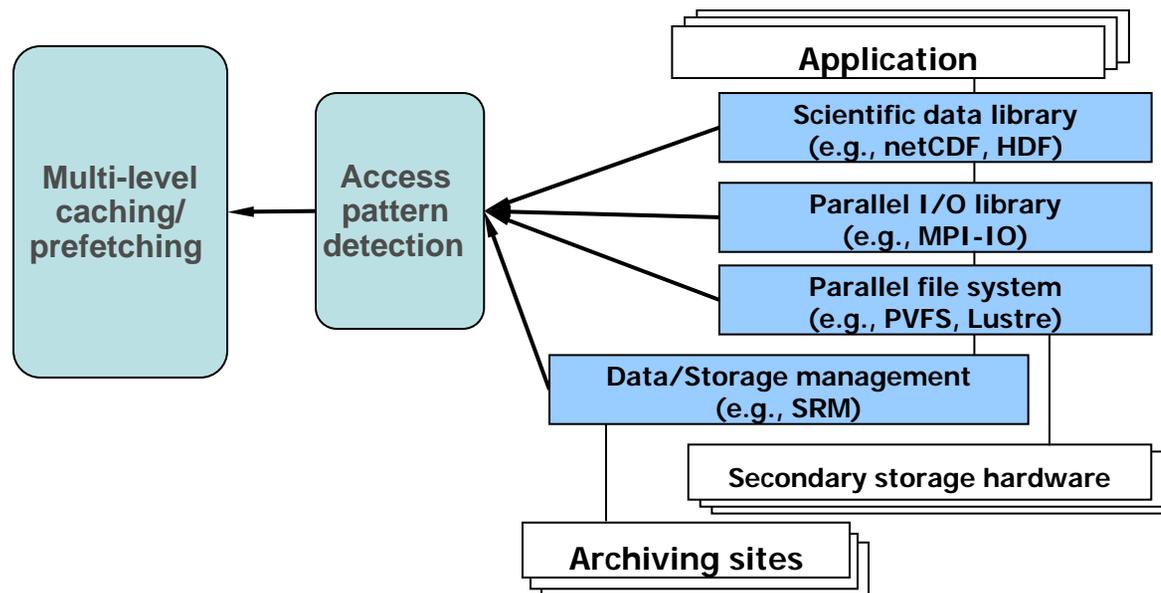
Collaborative Research:
***Improving Coordination in
HEC I/O Systems***

Xiaosong Ma*, Vincent Freeh, John Blondin (NC State U.)
Yuanyuan Zhou (U. of Illinois)
Anand Sivasubramaniam (Penn State U.)
Sudharshan Vazhkudai (Oak Ridge National Lab)

(* Joint faculty with Oak Ridge National Lab)

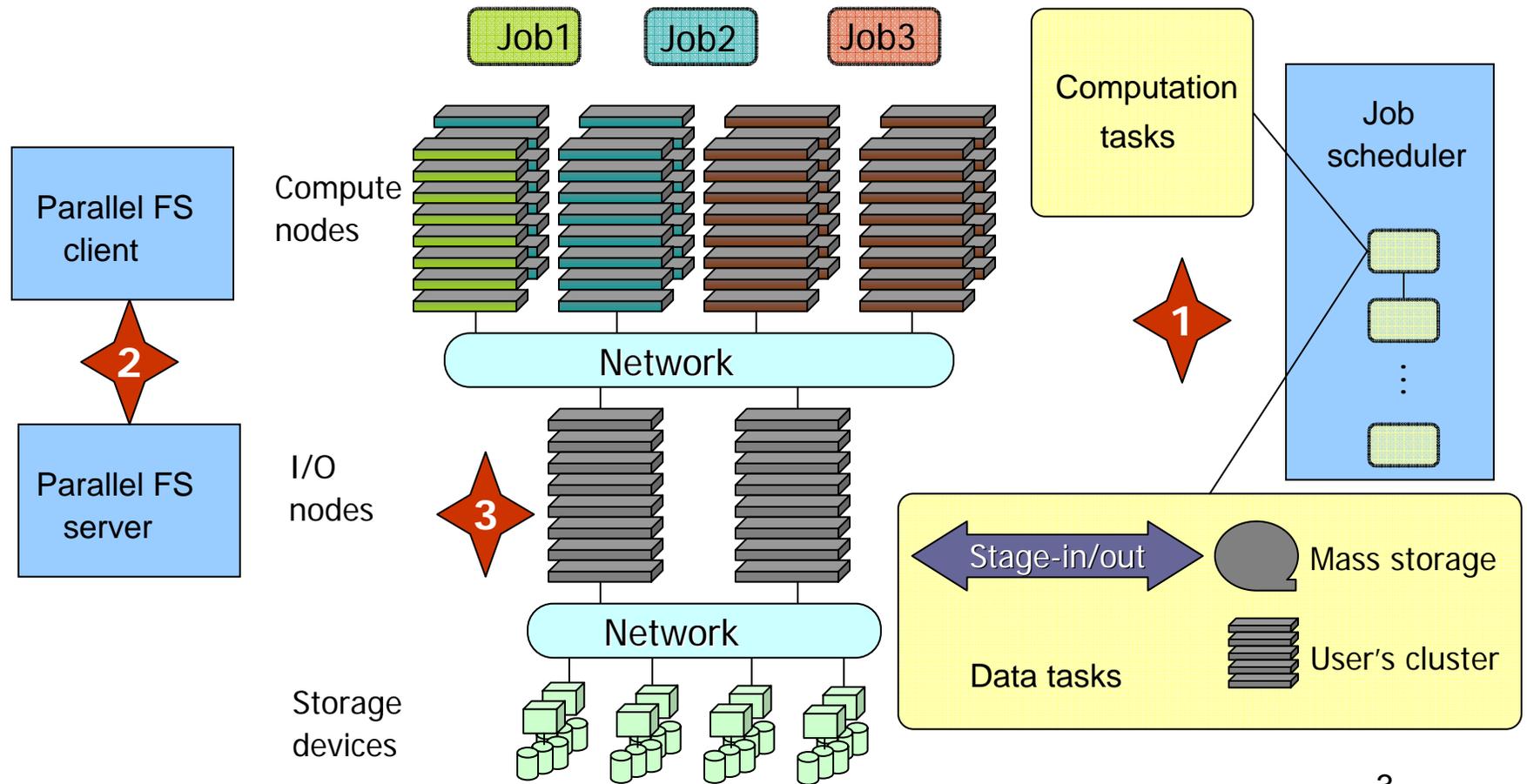
Project Overview

- PATIO (Parallel AdapTive I/O) Framework
- Focus
 - Automatic access pattern recognition
 - Multi-layer caching/prefetching

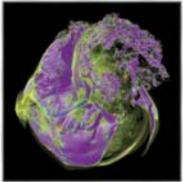


Theme: Enable coordination between different layers and components of HEC I/O system

- 1 *Coordinating storage w. job scheduling*
- 2 *Multi-layer, collaborative caching and prefetching*
- 3 *Coordinating service for multiple jobs at parallel FS servers*



Progress Overview

- Preliminary work on coordinating scheduling and I/O
 - Coordinating job scheduling with input data staging/reconstruction
- Application studies
 - Multiple types of applications
 - Simulations (FLASH) 
 - Visualization codes (Supernova) 
 - Parallel bio-sequence database search (BLAST)
- Setting stage for multi-layer caching/prefetching
 - Trace collection
 - Existing solutions in existing parallel file systems
 - Parallel file system simulator

Job-data Co-scheduling

- Data staging performed manually or in job script

Wish list for traces

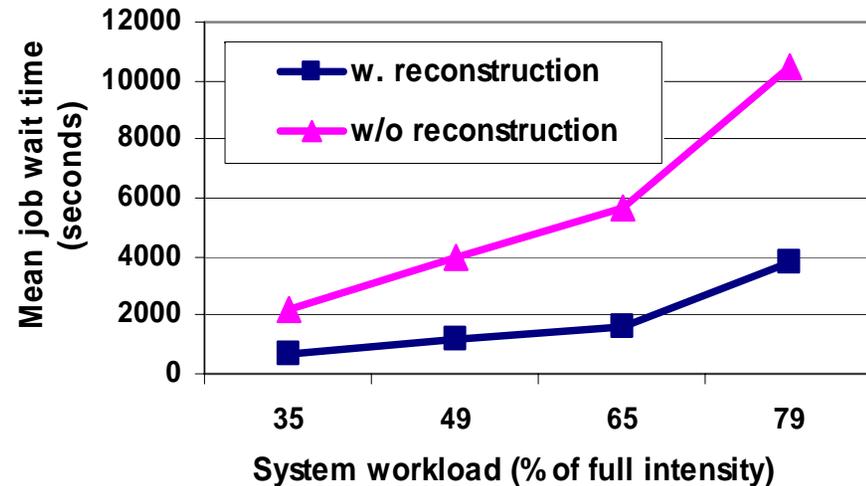
- - *Input/output file attributes for jobs*
 - *File size*
 - *Important timestamps (creation, deletion, first and last accesses)*
 - *Where from, where to*

ing alongside

- Preliminary results (SC|07)

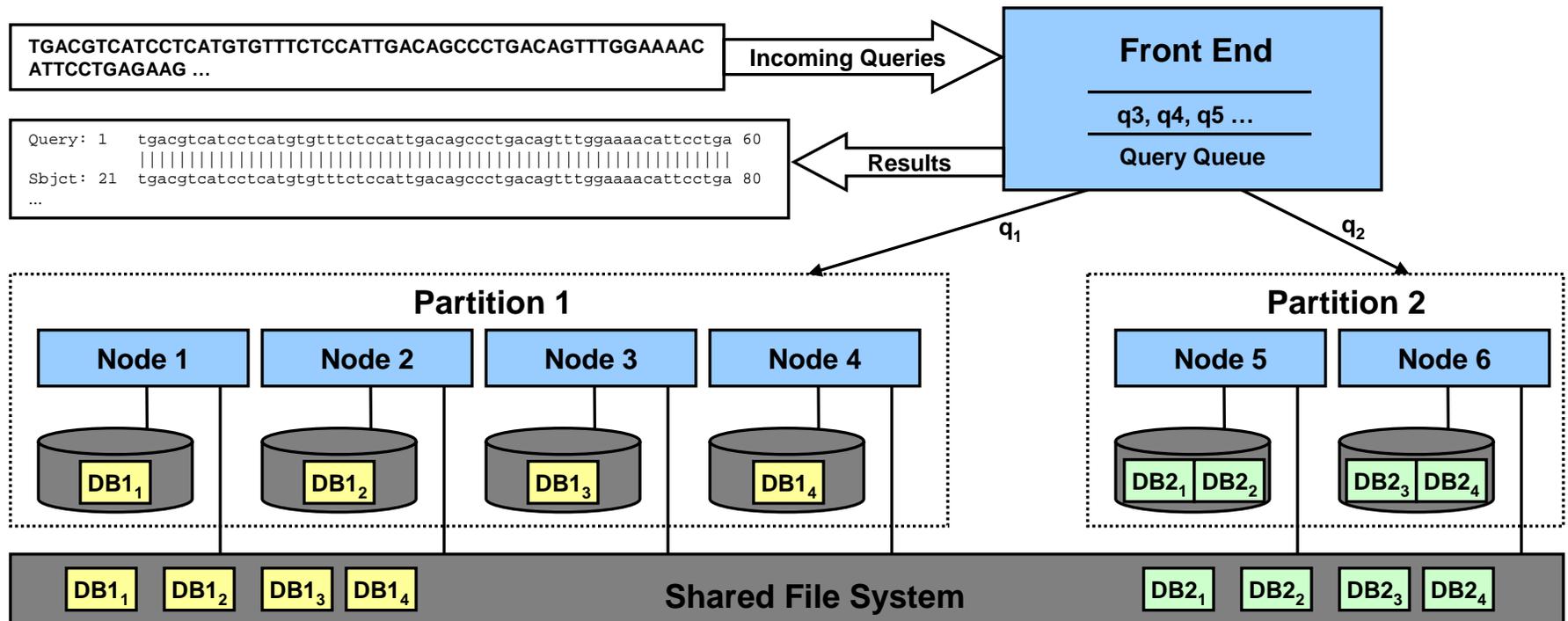
- Job and failure data from LANL
- Data staging records created based on ORNL scratch system

Mean wait time for all jobs w. storage node failures



Extending Coordination to Within Application

- New category of data-intensive application
 - Parallel bio-sequence database search



- Unlike traditional HPC applications
 - Repeated, sequential reads
- Unlike traditional cluster web servers
 - Computation-intensive, large access granularity
 - One server node working on one query at a time
- Node-attached local storage as cache

Extending Coordination to Within Application (cont'd)

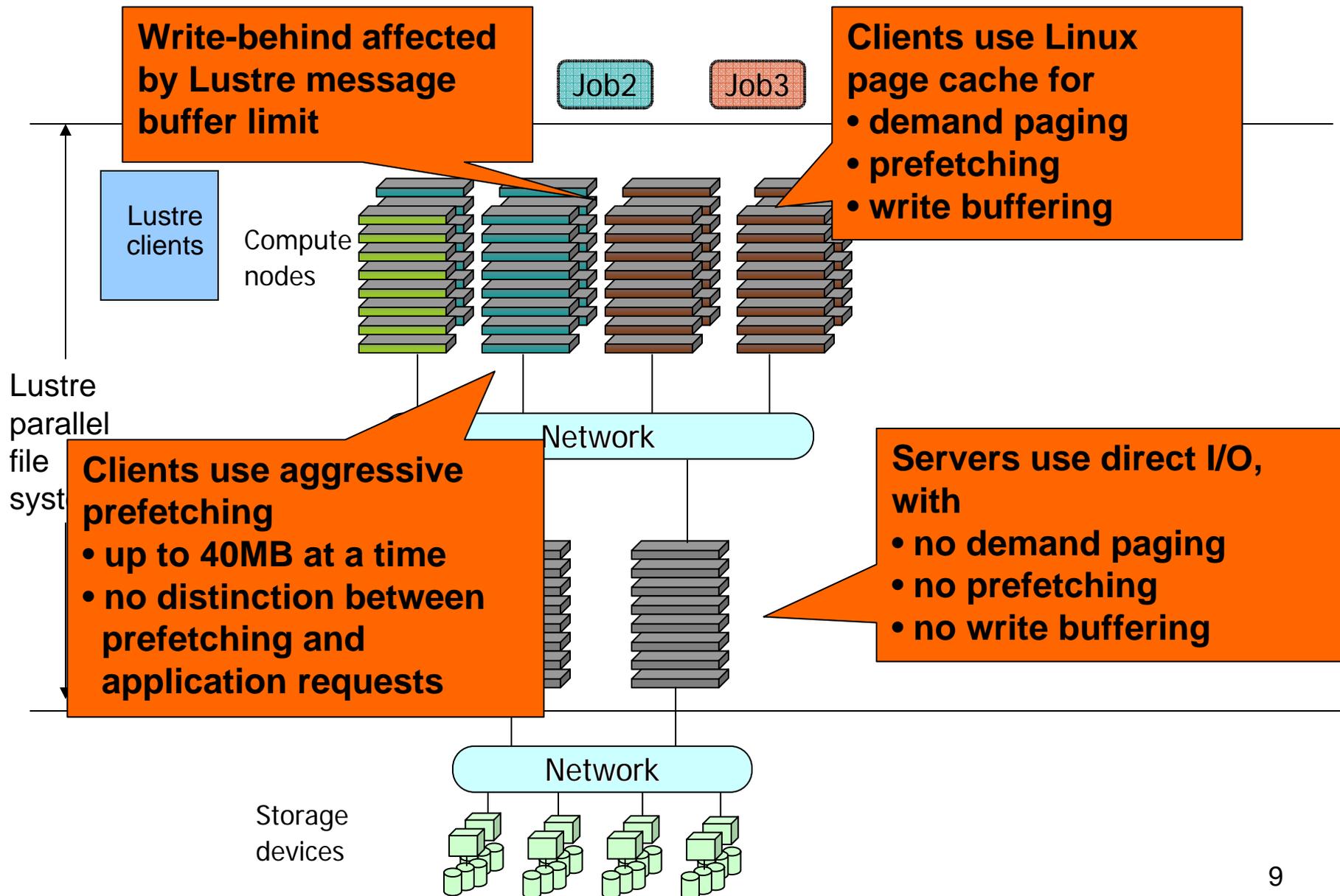
- Solution for traditional parallel web servers
 - Locality-aware request scheduling (e.g., LARD)
 - Still, I/O cost remains significant
- Proposed approach
 - Scheduling-aware I/O
 - Multi-layer prefetching based on expected scheduling decisions

Workload		0.1	0.5	1.0
Average Response Time (secs)	No Locality-aware optimization	8869	37795	39195
	LARD	7	105	769
	LARD + LC	5	59	505
Cache Misses (#misses/query/pro c)	No Locality-aware optimization	5.032	5.008	5.008
	LARD	0.128	0.787	0.056
	LARD + LC	0.043	1.435	0.048
Copy Volume (GB)	No Locality-aware optimization	3317	3477	3477
	LARD	124	302	31
	LARD + LC	4	242	23

Extending Coordination to Multiple Workloads

- Optimize center-wide high-performance storage space utilization
 - Manage scratch space as cache
 - Data not equally important
 - Associate files w. compute job and data staging status
 - Differentiate QoS for data associated w. different jobs
 - Expected benefits
 - Overall better storage resource utilization
 - Improved data reliability with small resource costs
 - Reduced usage charges if storage space gets billed

Multi-layer Prefetching/caching: Lustre Status



Enhancing Prefetching/caching for PFSs

- Why at parallel file system layer?
 - Common platform for diverse applications
 - No need to worry about high-level access consistency
- Adapting client-side strategies to HEC applications
 - More aggressive write-behind, less aggressive prefetching
- Adding server-side caching/prefetching
 - Client-side cache space availability: majority of HEC applications memory-intensive
 - Many small- or medium-size machines run single workload
 - For simulations, data need to be cached at server-side may be of small sizes
 - With smart prefetching strategies, cache space consumption may be reduced
- Coordinating client and server layers
 - Multi-layer prefetching/caching

Supported Personnel

- NCSU
 - Faculty: Xiaosong Ma, John Blondin, Vincent Freeh
 - Graduate student: Zhe Zhang
 - Undergraduate: Andrew Brown
- UIUC
 - Faculty: Yuanyuan Zhou
 - Graduate student: Kyu Hyung Lee
- ORNL
 - Sudharshan Vazhkudai, Greg Pike, John Cobb (supported by ORNL LDRD program)
- PSU
 - Anand Sivasubramaniam

Publications

- [SC07] Z. Zhang, C. Wang, S. S. Vazhkudai, X. Ma, G. Pike, J.W. Cobb, F. Mueller, "**Optimizing Center Performance through Coordinated Data Staging, Scheduling and Recovery**", to appear, *Proceedings of Supercomputing 2007*, Reno, Nevada, November 2007
- [OSR07] S. Vazhkudai, X. Ma, "**Recovering Transient Data: Automated On-demand Data Reconstruction and Offloading on Supercomputers**", *ACM SIGOPS Operating Systems Review: Special Issue on File and Storage Systems*, 41(1), January 2007
- [Frontiers07] O. Thorsen, K. Jian, A. Peters, B. Smith, H. Lin, W. Feng, Carlos P. Sosa, "**Parallel Genomic Sequence-Search on a Massively Parallel System**", ACM International Conference on Computing Frontiers, May 2007.