

# Collaborative Research: SAM<sup>2</sup> Toolkit: Scalable and Adaptive Metadata Management for High-end Computing

Hong Jiang<sup>1</sup> Yifeng Zhu<sup>2</sup> Jun Wang<sup>3</sup>  
David Swanson<sup>1</sup>

<sup>1</sup>Computer Science and Engineering  
University of Nebraska - Lincoln

<sup>2</sup>Electrical and Computer Engineering  
University of Maine

<sup>3</sup>Electrical Engineering and Computer Science  
University of Central Florida



# Outline

- 1 Metadata Management
  - Information avalanche
  - New challenges
- 2 Our Research Plan
  - Develop multi-variable forecasting models
  - Scalable file mapping schemes
  - Locality-aware metadata grouping
  - Cache coherence protocol
  - Prototype SAM<sup>2</sup> into PVFS<sup>2</sup> and dCache

# Kilo, Mega, Giga, Tera, Peta, Exa, ...

In part due to

- More powerful and higher precision observational **instruments**
- Larger-scale **simulations**
- Rapid advances in **networking**



## Suggested Topics

“It is vital that continued R&D investments be made in the scaling of metadata performance.”

- *Suggested R&D Topics for 2005-2009 (DOE, DOD)*

# New Challenges in Metadata Management

- Lack of understanding to metadata access **patterns** in shared file systems
- Trillions of files in **Exa-byte** ( $10^{18}$  bytes) storage
- **Extremely** high volumes of metadata activities
- Limitations of traditional caching, prefetching and coherence control schemes in **scalability**
- Lack of solid **benchmark** and **validation** methodology

# Metadata Patterns and Forecasting Models

- File access and metadata access have **different** characteristics.
- File access has been studied for two decades.
- But **little** study on metadata usage patterns.
- We will investigate metadata patterns and develop stochastic models to forecast metadata traffic.

# Scalable File Namespace Management

- Heavy traffic in a system with **trillions** of files:  $> 10,000$  ops/second
- Design **objectives**:
  - Single shared namespace
  - Scalable services
  - Balancing the load of metadata accesses
  - Flexibility of storing file metadata on any server
- Key techniques
  - Distributed metadata queries
  - Migrated or replicated to avoid hot-spots
  - Exploit localities to perform fast lookup

# Locality-aware Metadata Grouping

- Small I/Os are detrimental.
- Identify metadata access localities and form **groups**
  - Improve metadata update efficiency
  - Facilitate aggressive metadata prefetching
- Exploit temporal relationship between groups

# Design Scalable Coherence Protocols

- Large parallelism leads to extensive negotiations for metadata.
- **Decentralized** coherence protocols for metadata caching for extreme-scale storage systems.
- Different portions of metadata have different updating **frequencies** and different levels of **detriment**.
- Adaptive protocols to optimize the tradeoff between **false-sharing** and management **overheads** by dynamically changing sharing granularity, coherence levels, etc.

# Prototype Design

- UNL hosts a DOE **CMS** (high-energy physics) Tier-2 site.
- CMS deploys **dCache** and **PVFS<sup>2</sup>**.
- Integrating **SAM<sup>2</sup>** into CMS experiments

# Summary

- Develop multi-variable **forecasting models** to analyze and predict file metadata access patterns
- Develop scalable and adaptive file **name mapping** schemes to enforce load balance and increase scalability
- Develop decentralized, locality-aware **metadata grouping** schemes to facilitate the bulk metadata operations such as prefetching
- Develop an adaptive **cache coherence** protocol using a distributed shared object model
- **Prototype** the SAM<sup>2</sup> toolkit into PVFS<sup>2</sup> and dCache