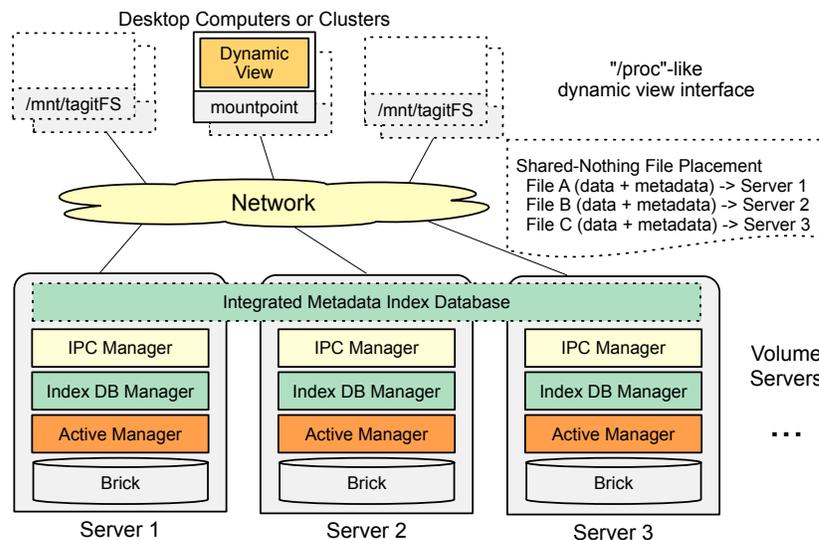


# TagIt: An Integrated Indexing and Search Service for File Systems

**Achievement:** Developed a scalable data management service framework aimed at scientific datasets, which is tightly integrated into a shared-nothing distributed file system. A key feature of TagIt is a scalable, distributed metadata indexing framework, using which we implement a flexible tagging capability to support data discovery. The tags can also be associated with an active operator, for pre-processing, filtering, or automatic metadata extraction, which we seamlessly offload to the servers in a load-aware fashion.

**Significance and Impact:** This work demonstrates a case for tightly integrating data management services within file systems in an attempt to enable rich search semantics therein. Traditionally, such services are provided via database catalogs external to the file system. However, emerging trends in data production enable us to revisit this design philosophy. We have shown how data management capabilities can be tightly integrated into the GlusterFS shared-nothing distributed file system. Our evaluation shows that TagIt can expedite data search by up to 10× over the extant decoupled approach.



## Research Details:

- **Tagging** TagIt extrapolates indexing and taxing capabilities to petabyte-scale file systems, wherein users can associate a richer context to collections of files by adding their own tags in order to quickly discover them.
- **Distributed Metadata Indexing** We have designed a consistent and scalable metadata indexing service that indexes user-defined extended attributes, and is tightly integrated into a shared-nothing distributed file system.
- **Active Operators** We have developed the ability to apply an operation or a filter on the file collections or specific portions of a file such as a stored variable, that is executed on the file system servers.
- **Automatically Extracting Metadata and Indexing** TagIt automatically extracts metadata from files.

**Sponsor/Facility:** This work was a collaboration between ORNL, Sogang University and Virginia Tech. It was sponsored by the US Department of Energy (DOE) Office of Advanced Scientific Computing Research (ASCR).

**PI and affiliation:** Sudharshan S. Vazhkudai – Center for Computational Sciences (CCS), Oak Ridge National Laboratory (ORNL)

**Publication:** Hyogi Sim, Youngjae Kim, Sudharshan S. Vazhkudai, Geoffroy R. Vallee, Seung-Hwan Lim, and Ali R. Butt. "TagIt: An Integrated Indexing and Search Service for File Systems", **The International Conference**

## **Overview:**

Data services such as search, discovery, and management in scalable distributed environments have traditionally been decoupled from the underlying file systems, and are often deployed using external databases and indexing services. However, modern data production rates, looming data movement costs, and the lack of metadata, entail revisiting the decoupled file system-data services design philosophy.

TagIt, a scalable data management service framework aimed at scientific datasets, is tightly integrated into a shared-nothing distributed file system. A key feature of TagIt is a scalable, distributed metadata indexing framework, using which we implement a flexible tagging capability to support data discovery. The tags can also be associated with an active operator, for pre-processing, filtering, or automatic metadata extraction, which we seamlessly offload to file servers in a load-aware fashion. Our evaluation shows that TagIt can expedite data search by up to 10× over the extant decoupled approach.