BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

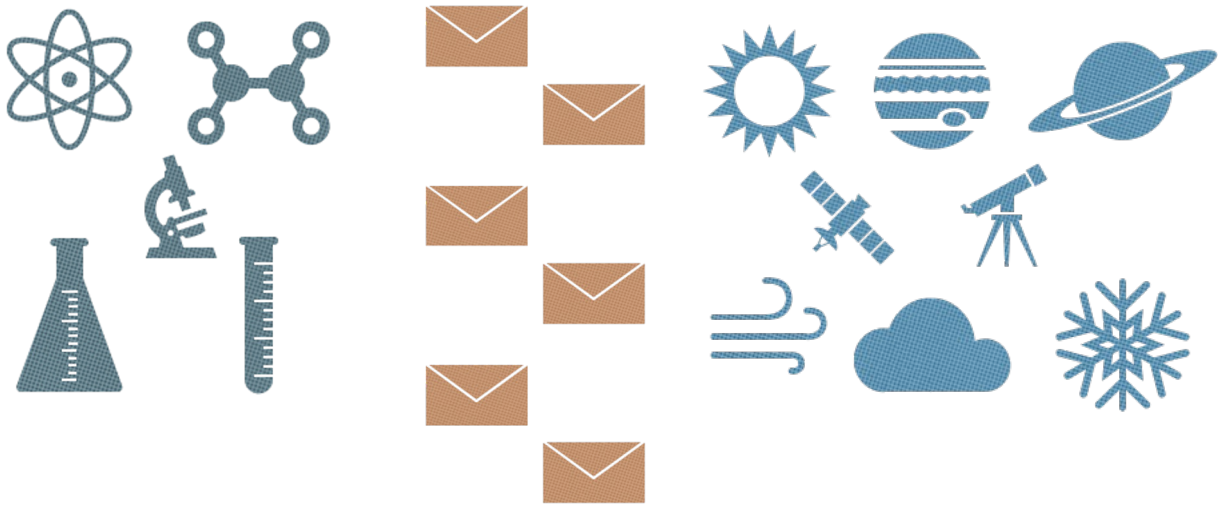
Towards Self-contained Metadata Search Capability for Self-describing File Formats

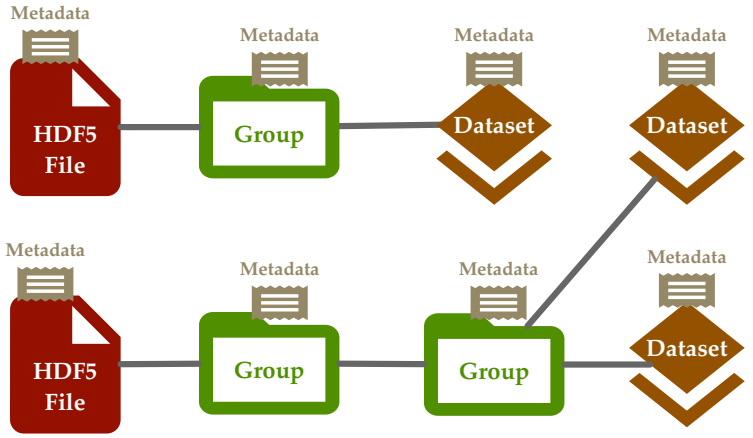
Wei Zhang – Lawrence Berkeley National Laboratory

Aug/17/2023

Data Management in Scientific Applications

- Scientific Applications
  - Experiments
  - Observations
- Ever-increasing Data
  - Size of the files
  - Number of the files
  - Variety of the files
- Self-describing Data Formats
  - HDF5, netCDF, ADIOS-BP, ASDF, etc.
  - Metadata is stored alongside the data objects

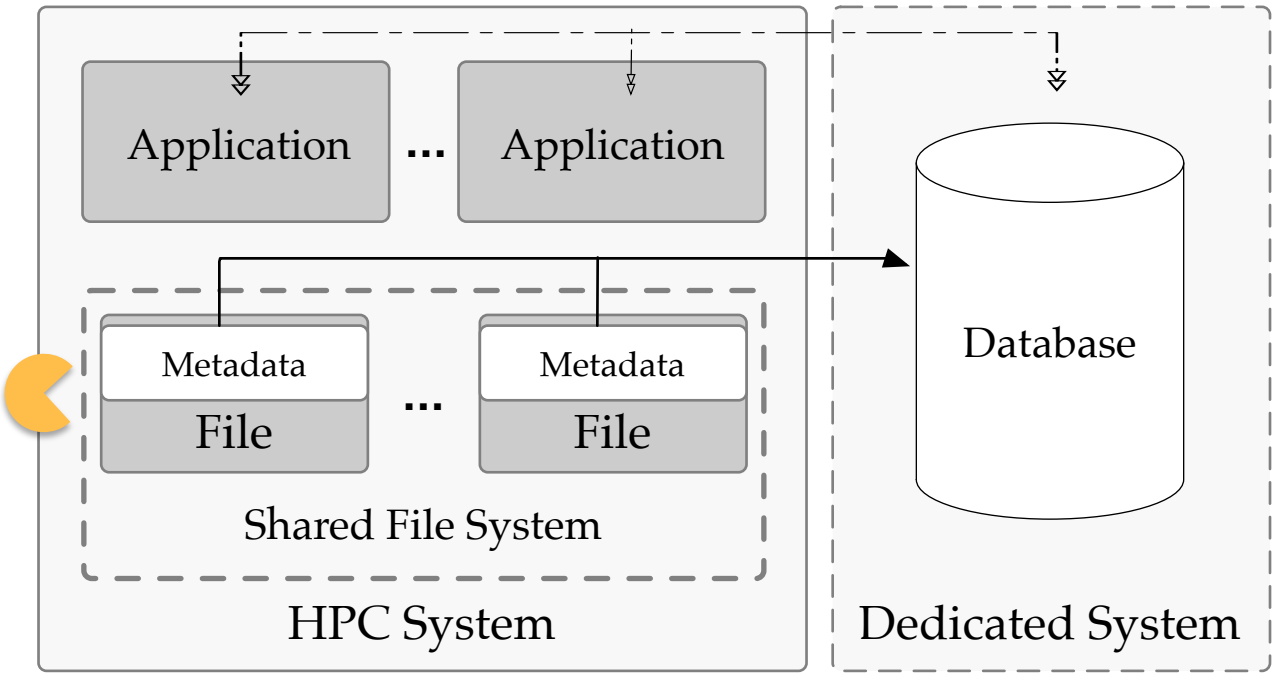




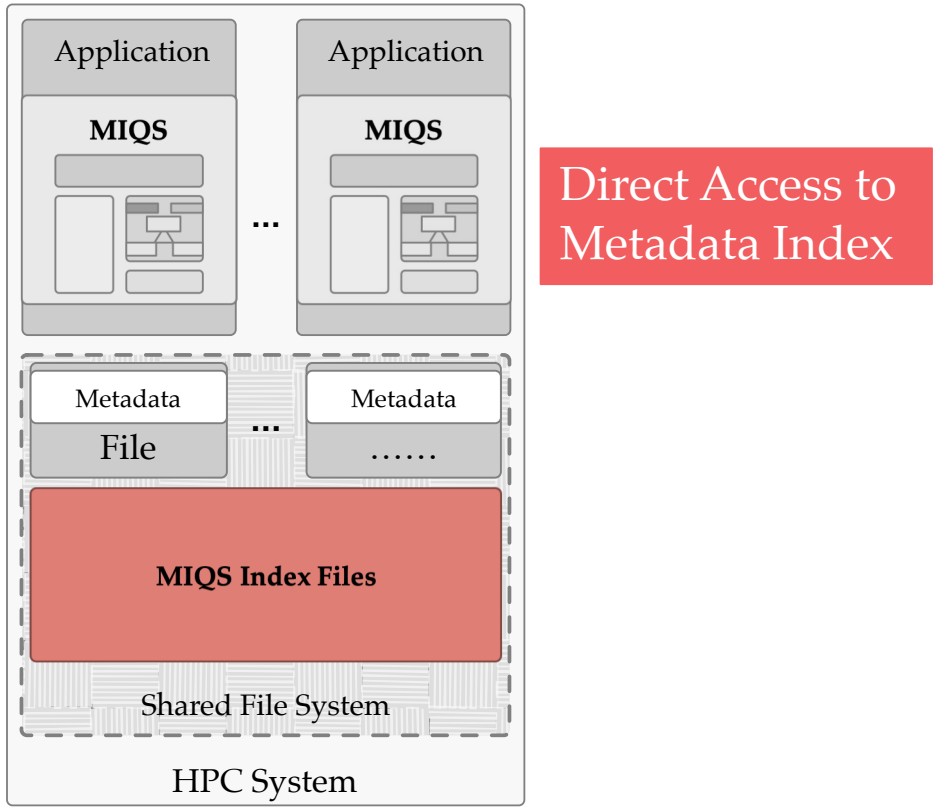
Metadata Search

- **Scanning Files?**
  - Time-consuming file scanning process
    - Size of the data objects
    - Number of the data files
- **Using Databases?**
  - Deployment Effort
  - Maintenance Demand
  - Data Model Adaption
  - Storage Redundancy
  - Performance Issues
  - Poor Portability & Mobility

Name	Type	Data Source	DB System	DB Type
SPOT Suite	Tomography	HDF5	MongoDB	NoSQL
JAMO	Genomics	HDF5	MongoDB	
BIMM	Biomedical	Biomedical Image	MySQL	RDBMS
EMPRESS 2.0	General	HDF5, netCDF	SQLite	



A New Era of Metadata Search

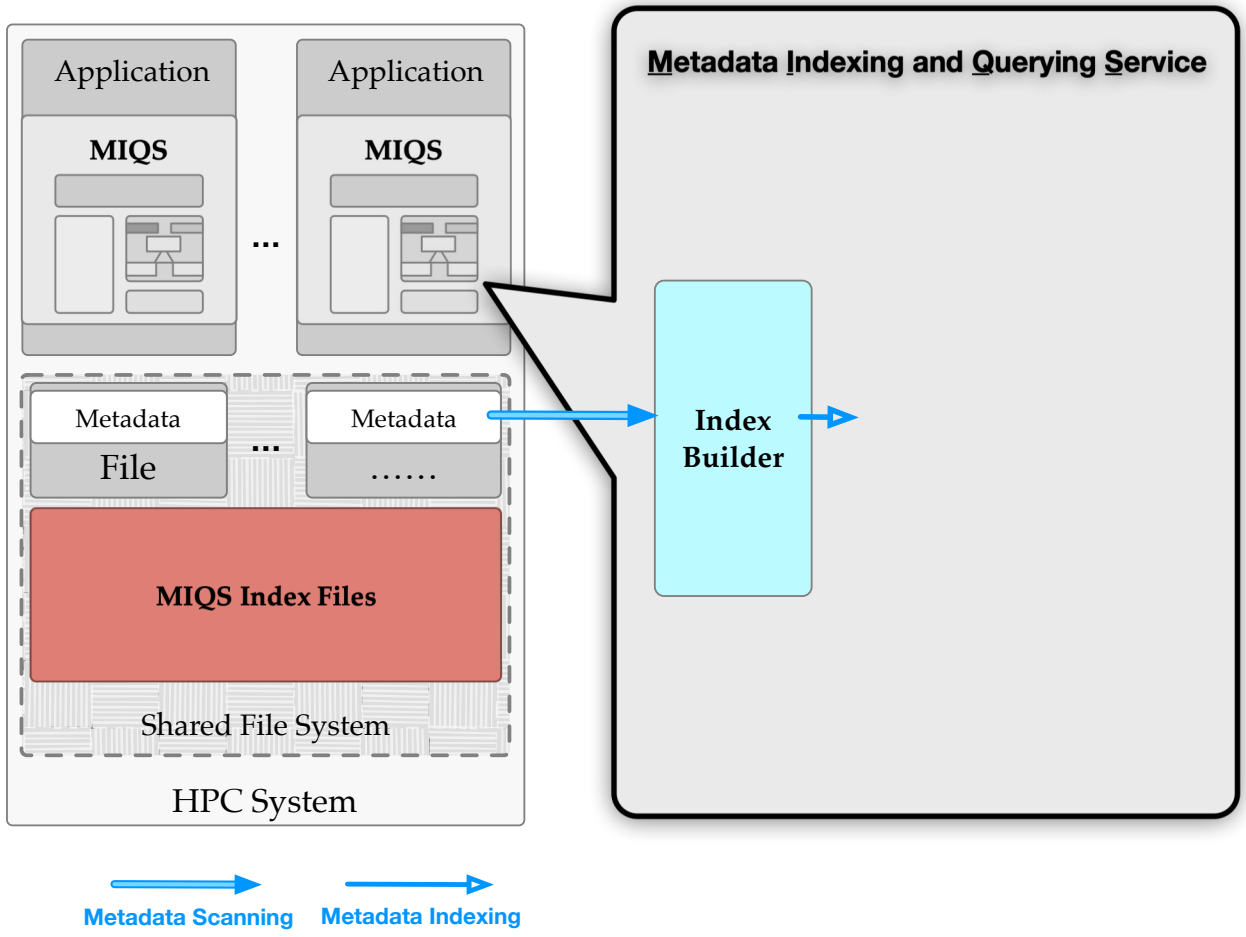


Direct Access to Metadata Index

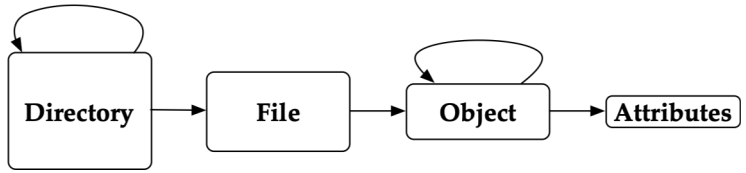
– Self-contained Metadata Index

- Minimal Complexity:
  - Metadata schema
  - Metadata search process
- Portability & Mobility
- Minimal Storage Requirement
- Excellent Performance

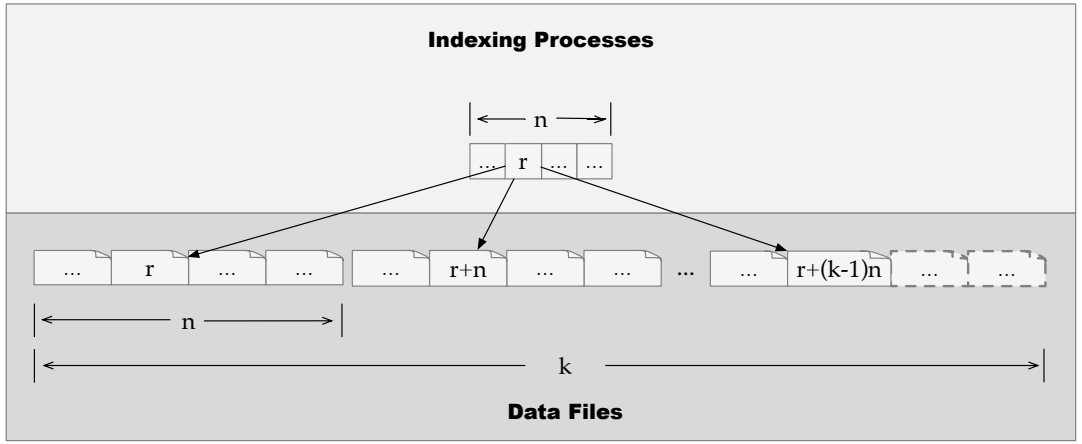
MIQS – Metadata Indexing and Querying Service



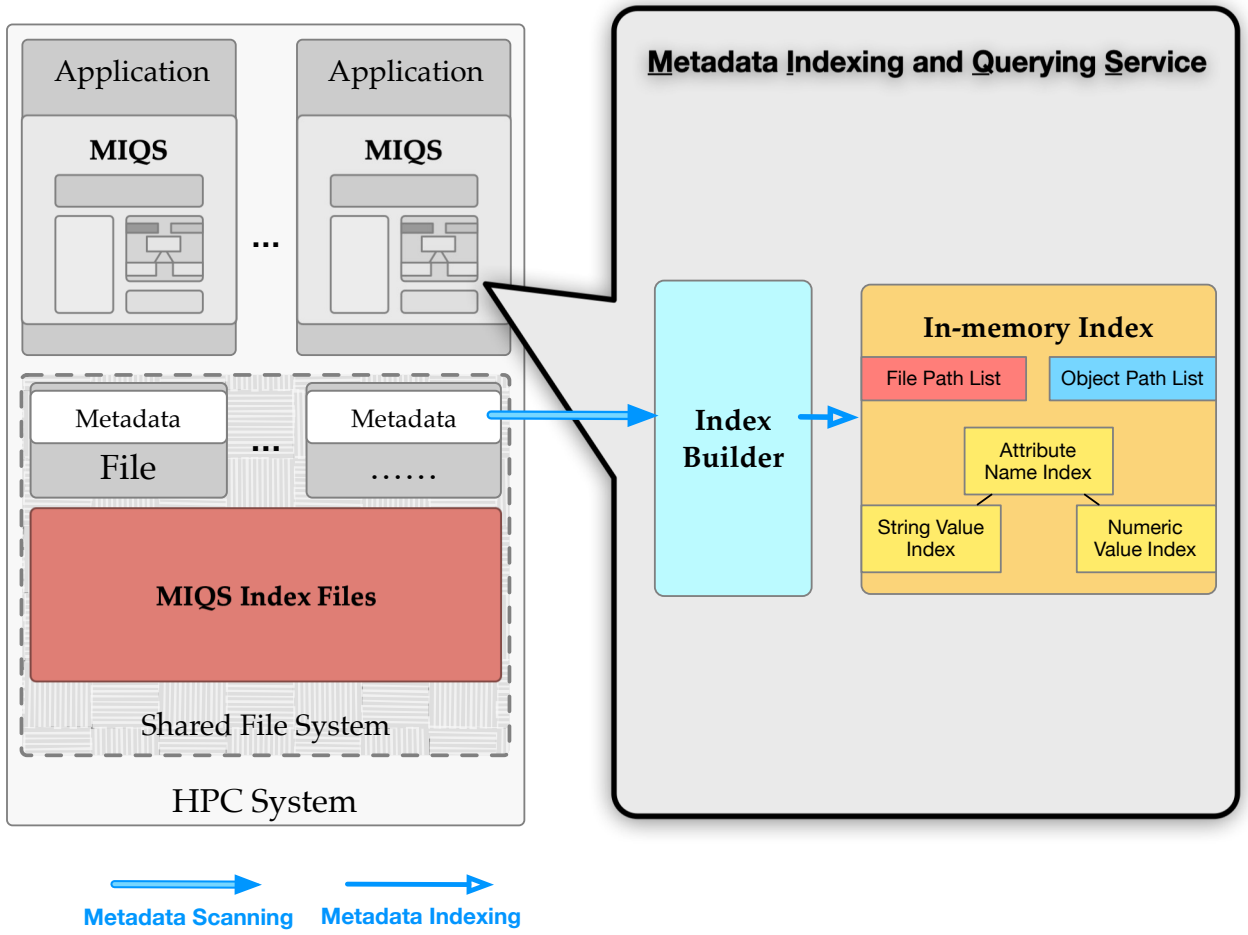
- Index Builder
  - Initial Index Construction



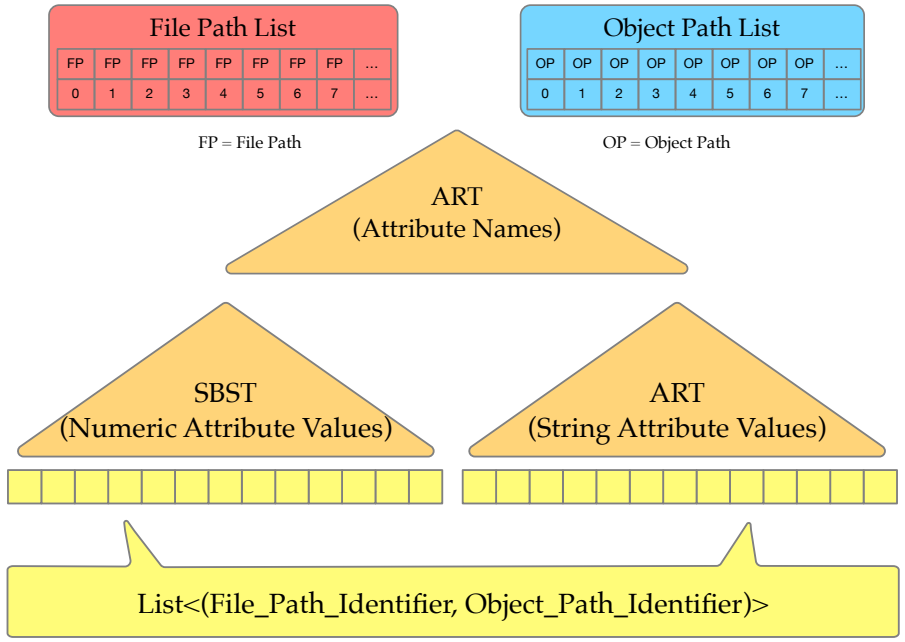
Process r only index when $file\_counter_r \% n == r$



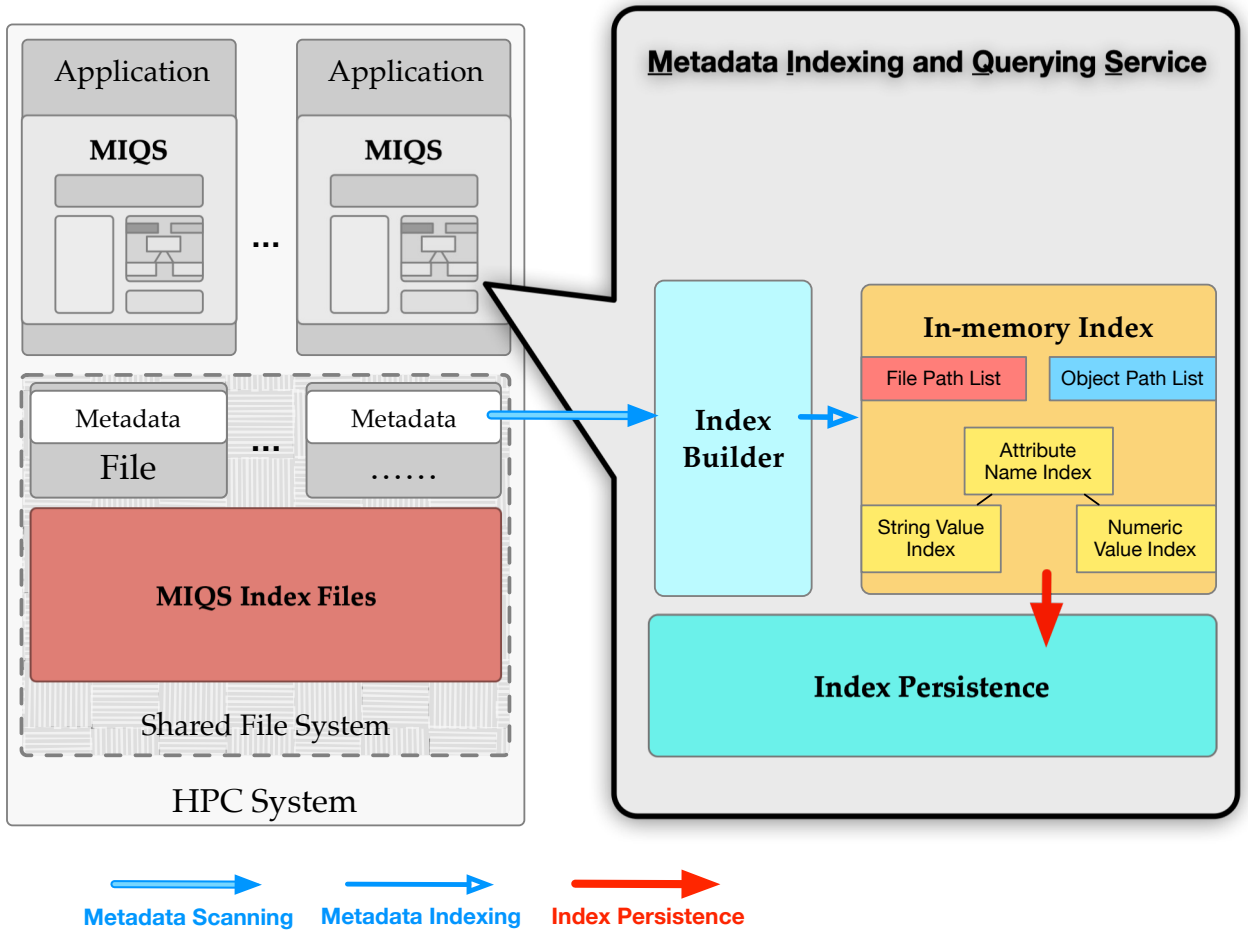
MIQS – Metadata Indexing and Querying Service



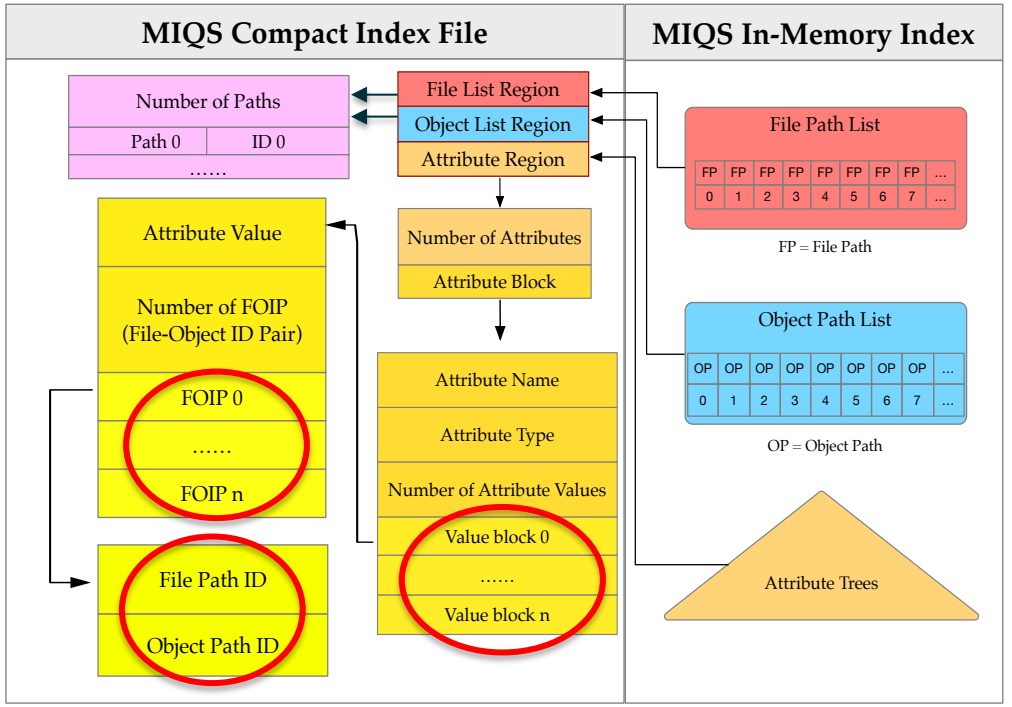
- In-memory Index



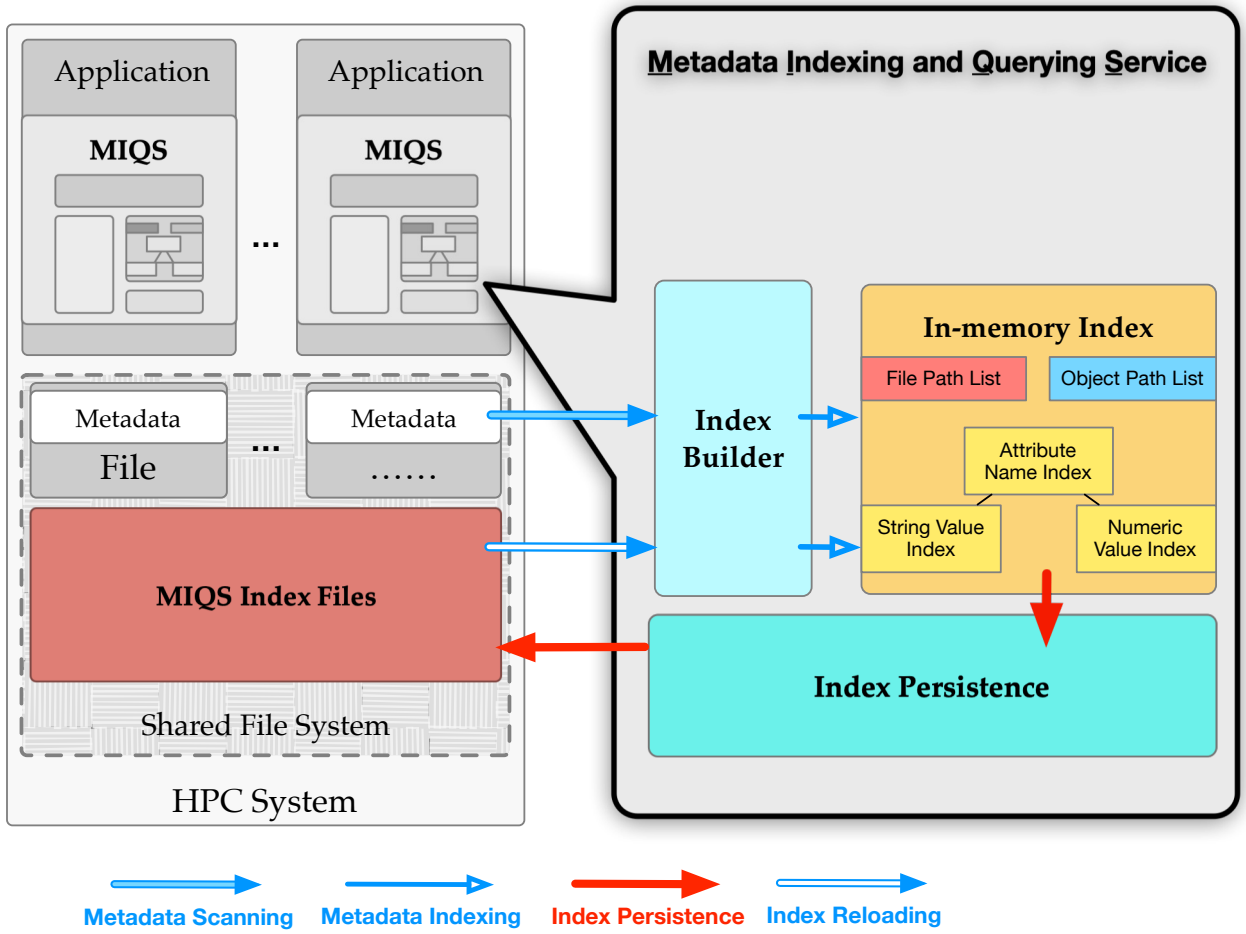
MIQS – Metadata Indexing and Querying Service



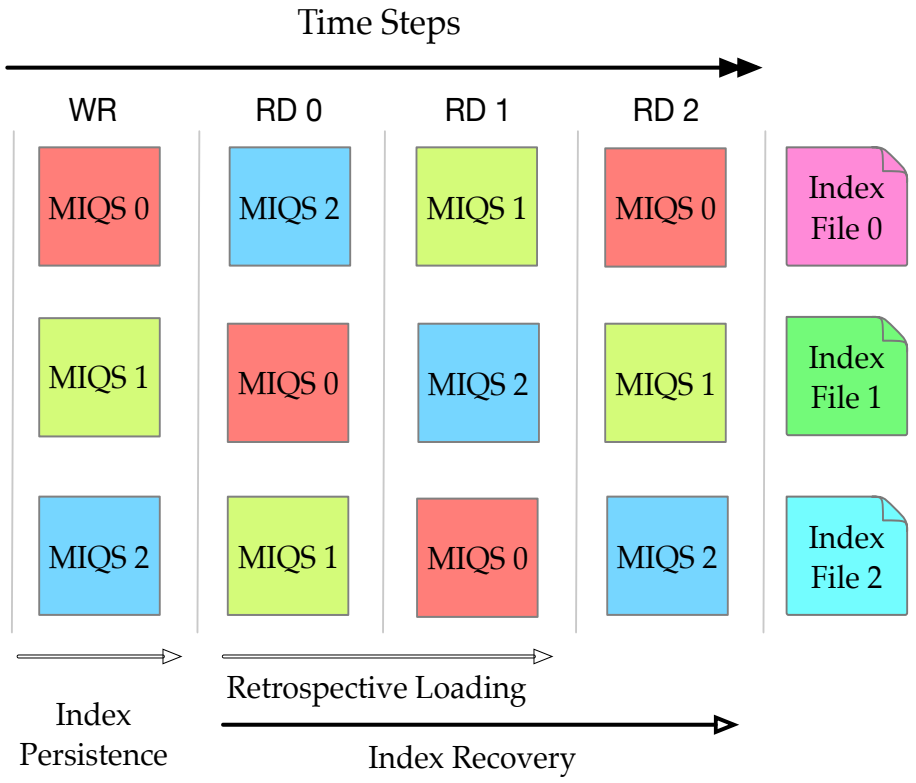
- Index Persistence
  - Compact Index File



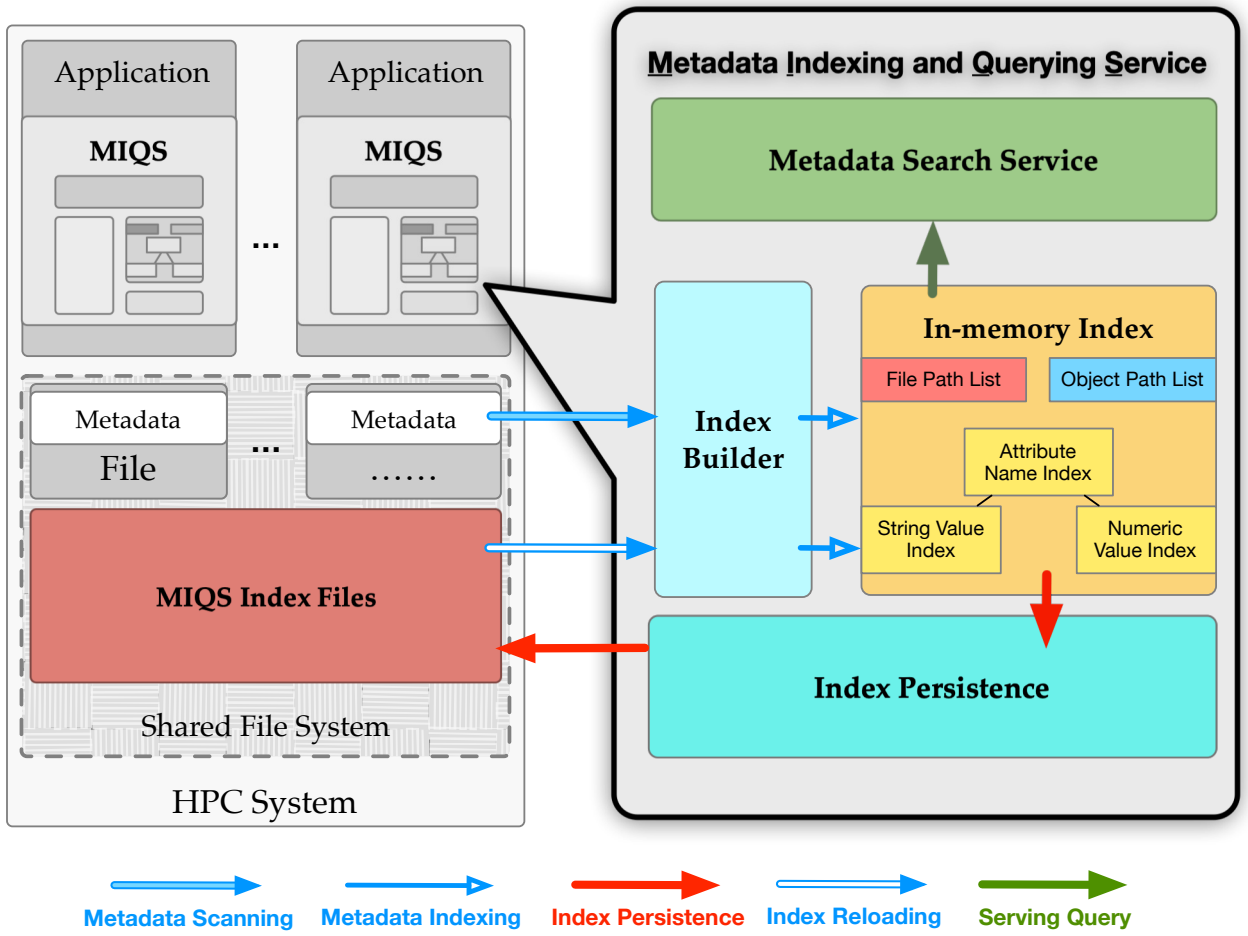
MIQS – Metadata Indexing and Querying Service



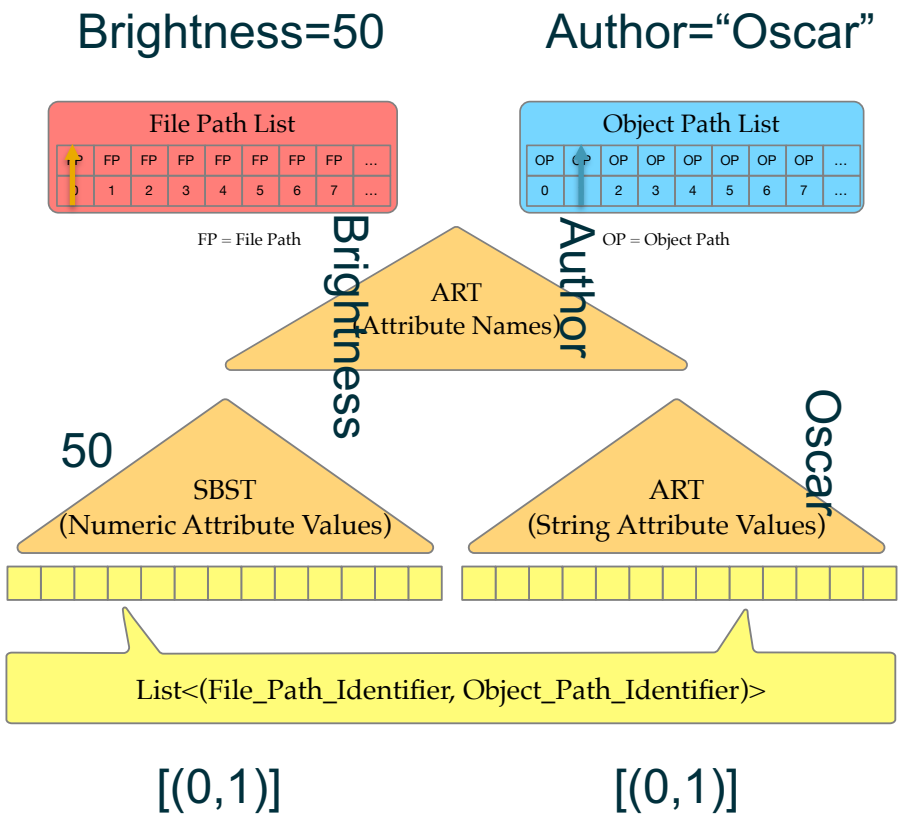
- Index File Read/Write



MIQS – Metadata Indexing and Querying Service



- Serving Metadata Queries



"/home/Oscar/data/test.hdf5","/2019/05/2/B/pixel.fit"

Evaluation – Platform & Control Experiment

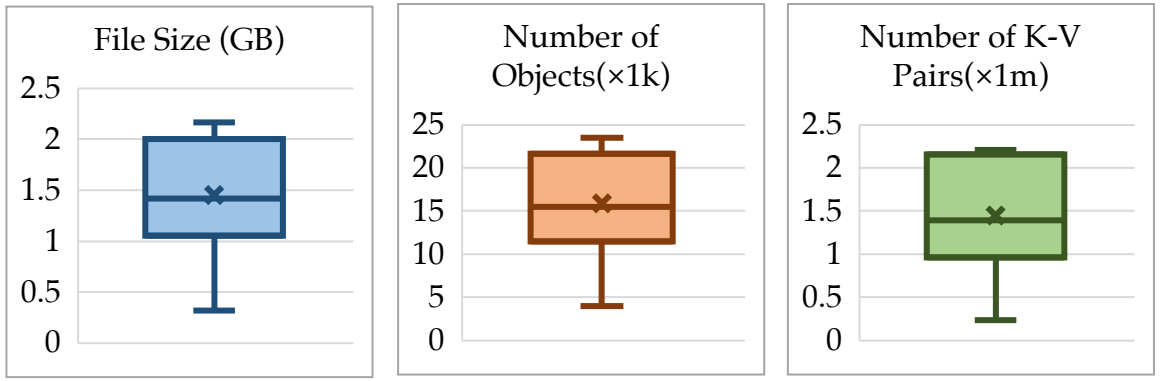
- MIQS v.s. MongoDB
  - NoSQL
  - Flexible Data Schema
  - State-of-the-art

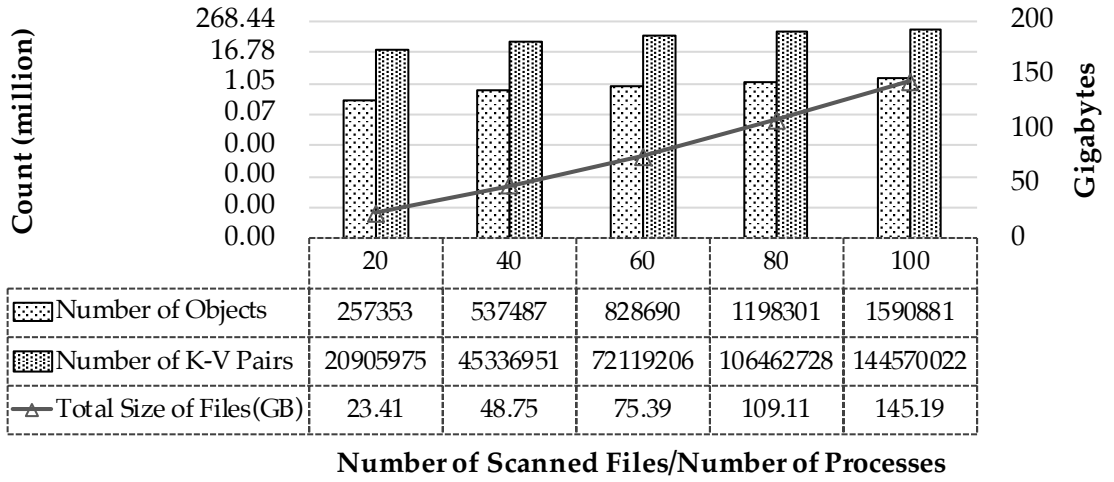
MIQS Evaluation Platform	
Supercomputer	Edison
CPU	12 * Ivy Bridge @2.4GHz
Memory	64GB DDR3 1866
Network	23.7TB/s global bandwidth
Lustre	30PB @ 700GB peak I/O

MongoDB Evaluation Platform	
Host machine	Different from Edison
CPU	16 * Haswell @2.3GHz
Memory	128GB DDR4 2133
Network	56Gb/s bandwidth
HDD	6TB 7200rpm 6Gb/s SAS
MongoDB Storage Engine	WiredTiger with data compression

Evaluation – Dataset

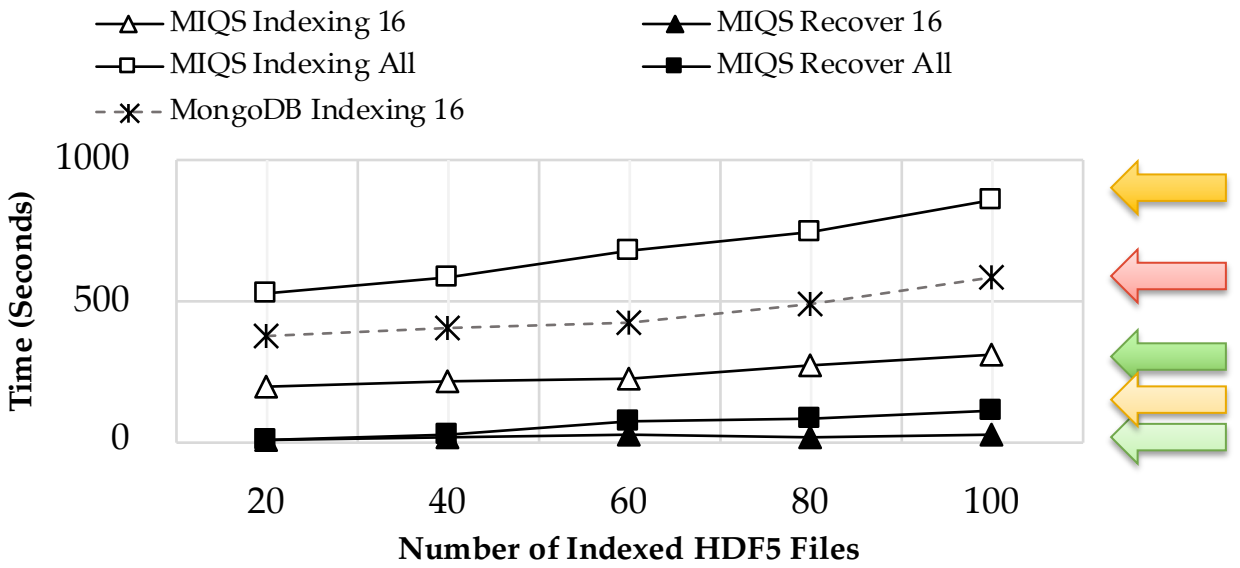
- 100 HDF5 files
- Baryon Oscillation Spectroscopic Survey(BOSS)
- 145 GB
- 250 attributes
- 144 million key-value pairs
- 1.5 million data objects.





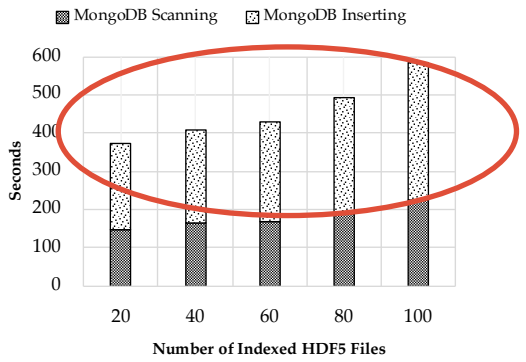
Evaluation – Indexing Time

- 16 attributes in MongoDB
  - 5-9 min
- 16 attributes in MIQS
  - 50% Indexing Time Reduction (initial indexing)
  - 99% Indexing Time Reduction (index recovering)
- You can also:
  - Index all 250 attributes in MIQS in 8-14min.
  - Recover index of all 250 attributes within 2min.

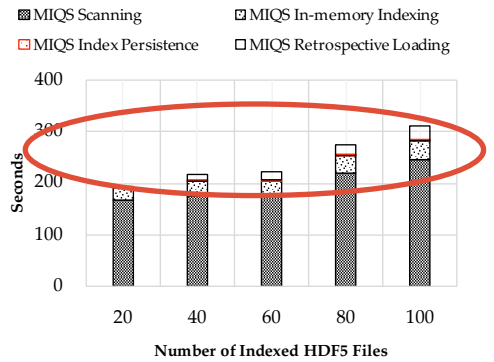


Evaluation – Indexing Time (Break-down)

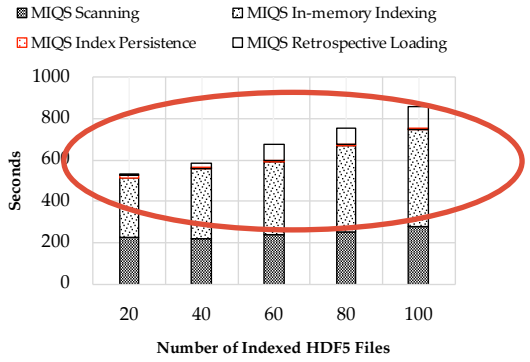
- Scanning Time : roughly equal (MIQS v.s. MongoDB)
- MongoDB 16 attributes
  - Inserting BSON (3 - 6min)
- MIQS 16 attributes
  - In-memory index: 0.5 – 1min
  - Persistent index : ignorable
- MIQS all 250 attributes:
  - In-memory index: 5 – 8min
  - Persistent index : ignorable
- MIQS index recovery
  - 16 attributes: < 40s
  - All 250 attributes: < 2min.



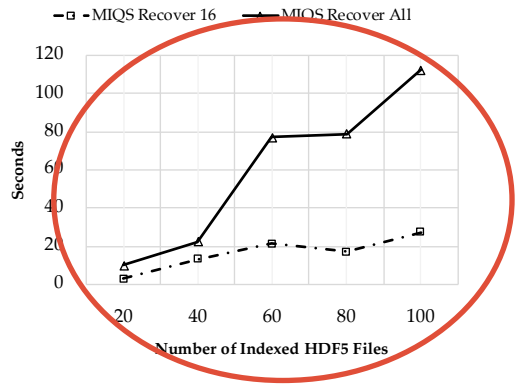
MongoDB Indexing Time (16 attributes)



MIQS Indexing Time (16 attributes)



MIQS Indexing Time (all attributes)



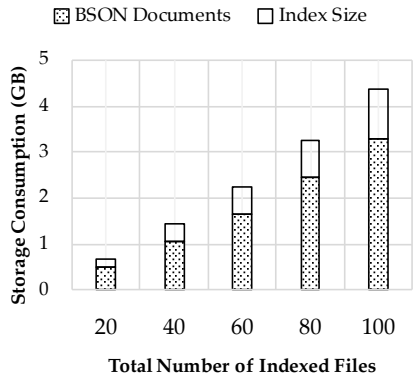
MIQS Index Recovery Time

Evaluation – Storage Consumption

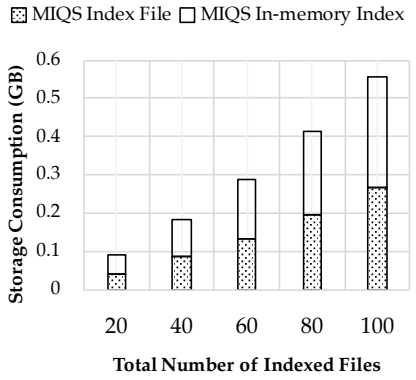
- MongoDB:
  - 16 attributes: Up to 4.2GB
- MIQS:
  - 16 attributes: up to 600 MB
  - All 250 attributes: up to 7.8GB

- You can:
  - Save space
  - Index more attributes

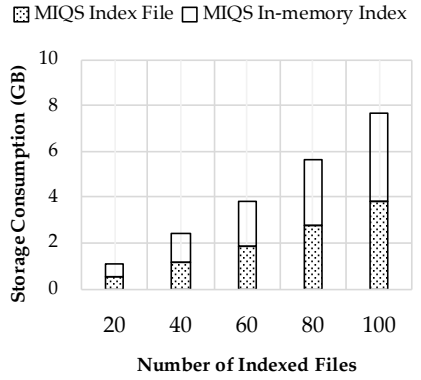
75% Storage Reduction



MongoDB Storage Consumption (16 attributes)



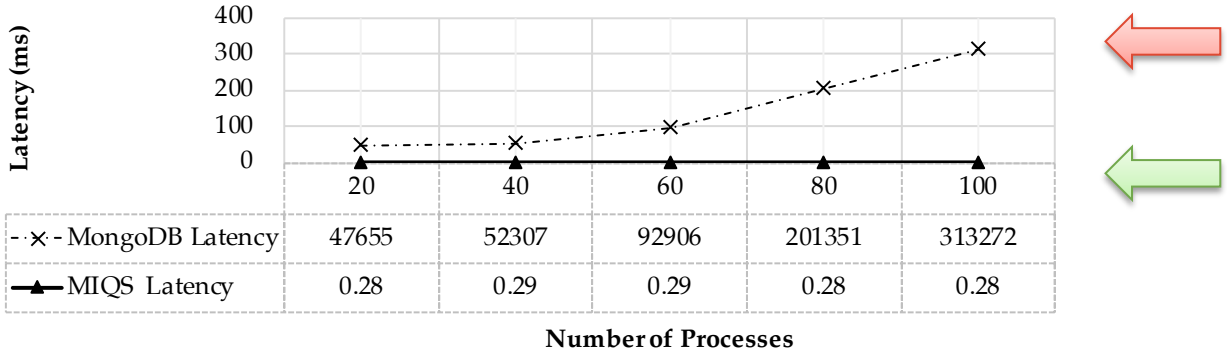
MIQS Storage Consumption (16 attributes)



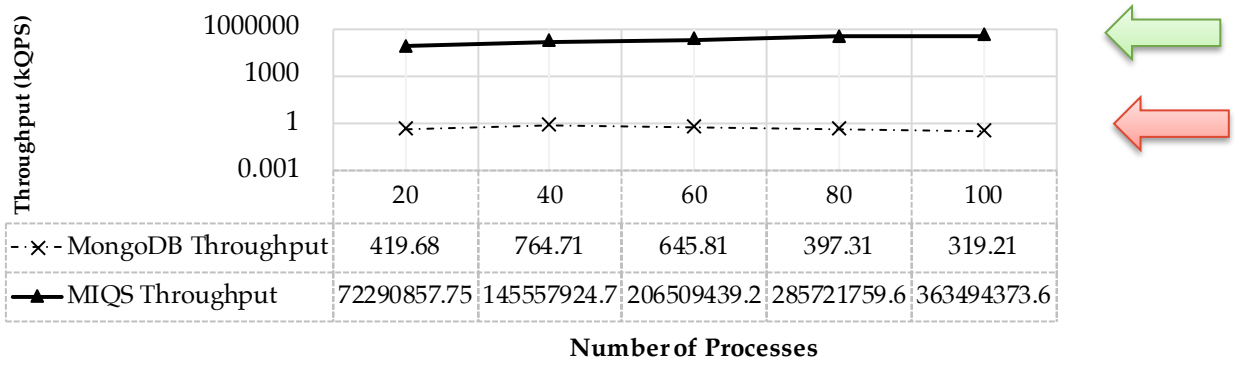
MIQS Storage Consumption (all attributes)

Evaluation – Query Performance

- Latency
  - MongoDB: 5 min at maximum scale
  - MIQS: 0.29 ms at maximum scale
- Throughput:
  - MongoDB: 319 kQPS at maximum scale
  - MIQS: 363 billion QPS at maximum scale



Query Latency Comparison (16 attributes)



Query Throughput Comparison (16 attributes)

- Discussion
  - Full Functionality of DBMS
    - Do we need it in most of the cases?
    - What does it take to deploy/maintain/learn about DBMS
    - Compound queries
      - Creatively build composite index
      - Query execution plan
  - What about embedded DBMS
    - SQLite?
    - Complexity
      - Data model adaption
      - Learning curve
    - Storage overhead
    - Performance
- Conclusion
  - Problems: No metadata indexing or Not self-contained.
  - MIQS – Self-contained Metadata Indexing and Querying Service
  - Benefits:
    - Minimal Complexity
    - Minimal Storage Requirements
    - Portability & Mobility
    - Impressive Performance
- Future Work
  - More types of queries
  - Performance improvement

Follow Up

Paper:

ACM Digital Library:

http://bit.ly/SC19-MIQS

Contact Us:

SDM Group @ LBNL

https://sdm.lbl.gov

zhangwei217245@lbl.gov
htang4@lbl.gov
sbyna@lbl.gov

DISCL @ TTU:

https://discl.cs.ttu.edu

Brody.Williams@ttu.edu
Yong.Chen@ttu.edu

Acknowledgement:

Many thanks to the audience and also those paper reviewers who provided valuable comments.

This research is supported in part by the National Science Foundation under grant CNS-1338078, CNS-1362134, CCF-1409946, CCF-1718336, OAC-1835892, and CNS-1817094. This work is sup- ported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. (Project: EOD-HDF5: Experimental and Observational Data enhancements to HDF5, Program managers: Dr. Laura Biven and Dr. Lucy Nowell). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility.

Q & A

Thank You