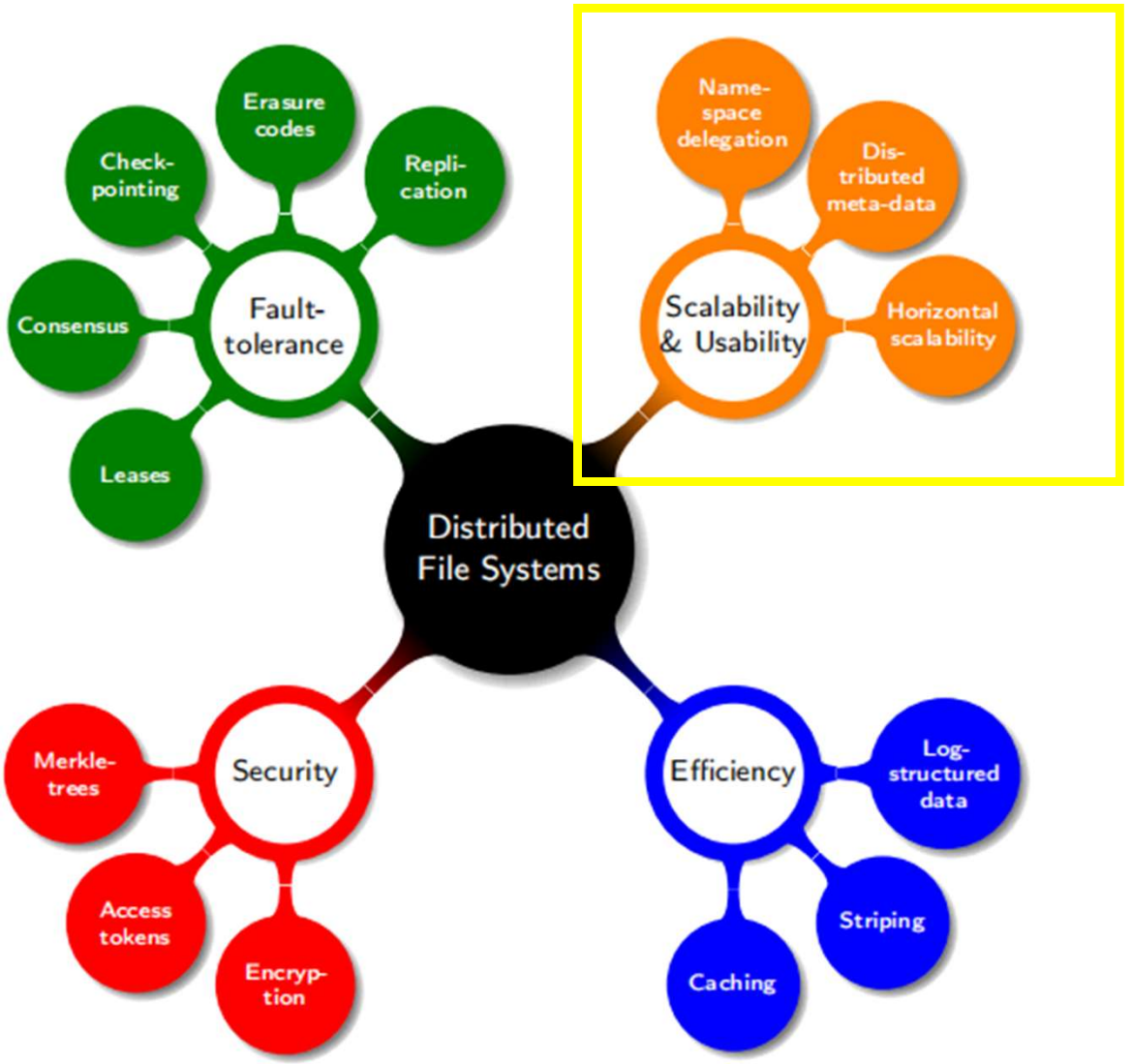
相关技术


Distributed File Systems
Fault-tolerance
Erasure codes
Check-pointing
Repli-cation
Consensus
Leases
Scalability & Usability
Name-space delegation
Dis-tributed meta-data
Horizontal scalability
Security
Merkle-trees
Access tokens
Encryp-tion
Efficiency
Log-structured data
Striping
Caching

相关技术
Scalability&Usability : Distributed metadata

- Metadata includes:
  - File namespace
  - Locations of each file's replicas
  - Versions of replicas
  - ……
- A central metadata server (common approach)
  - Decoupling metadata and data
  - High throughput
- Metadata distributed in all nodes
  - Hard to manage

相关技术
Scalability&Usability : Distributed metadata

- In GFS[1], master stores three major types of metadata:
  - File and chunk namespaces
  - Mapping from files to chunks
  - Locations of each chunk's replicas
- In-memory metadata
- How to read
  - Control messages
  - Data messages

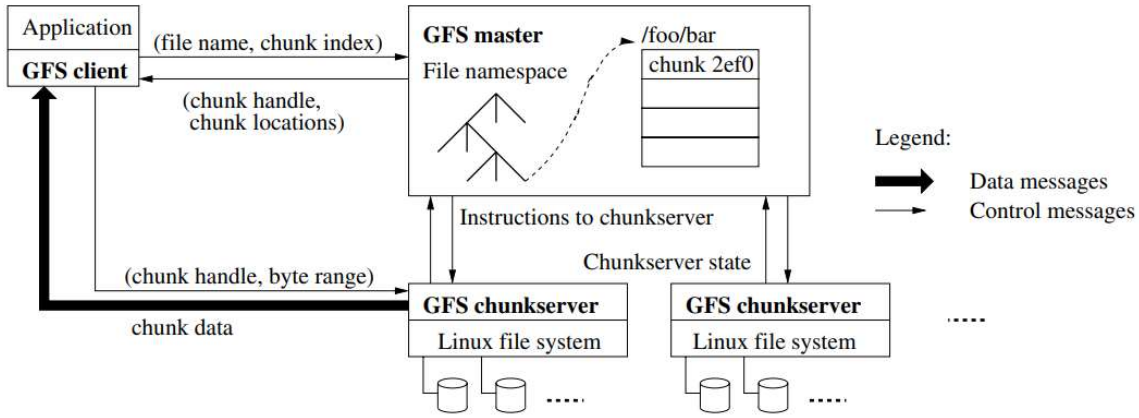


Figure 1: GFS Architecture

[1] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03). Association for Computing Machinery, New York, NY, USA, 29–43.

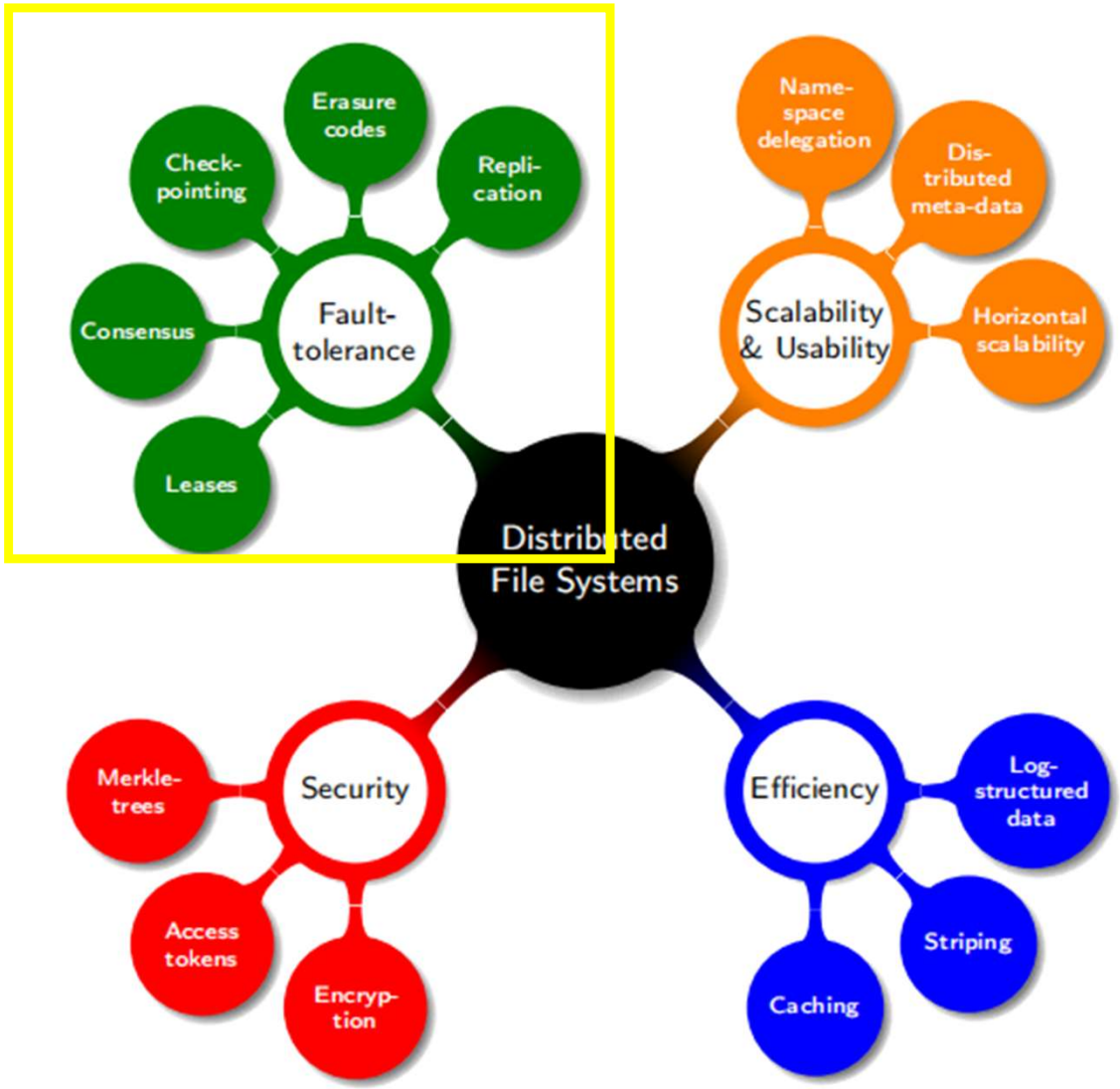
相关技术
Scalability&Usability : Namespace delegation

- Considering that we create multiple files in the same directory concurrently
- Namespace Lock
  - Read lock
  - Write lock
- Operations in the directory " /d1/d2/.../dn/leaf "
  - Requiring read locks of "/d1", "/d1/d2", "/d1/d2/…/dn"
  - Requiring read/write lock of "/d1/d2/…/dn/leaf"

相关技术
Scalability&Usability : Horizontal scalability

- How to rebalance replicas periodically?
  - Identified by **disk utilization**, CPU utilization, network, etc.
  - Moving replicas for better diskspace and load balancing
- How to balance workload if we add a new server?
  - Limit the number of “recent” creations on each server
  - Avoid heavy traffic in the beginning
  - Remain a warm-up time
- How to update metadata?
  - For central metadata server, update related metadata directly.
  - For distributed metadata, logical location and physical location are separated.

相关技术


Distributed File Systems
Fault-tolerance
Erasure codes
Replication
Check-pointing
Consensus
Leases
Scalability & Usability
Name-space delegation
Distributed meta-data
Horizontal scalability
Security
Merkle-trees
Access tokens
Encryption
Efficiency
Log-structured data
Striping
Caching

相关技术
Fault-tolerance: Checkpointing

- Operation log: persistent record of metadata
  - File and chunk namespaces
  - Mapping from files to chunks
  - Both locally and remotely
- Checkpoints
  - To minimize startup time consumed by the master recovering
  - Build a checkpoint when the log grows beyond a certain size
  - In a separate thread

相关技术
Fault-tolerance: Leases

- Leases: maintain a consistent mutation order
  - Primary replica and secondary replicas
  - Picked by master
  - 60s timeout
- How to write
  - 1. Client requests for chunk information
  - 2. Master replies to Client (cached locally)
  - 3. Client pushes data to all the replicas
  - 4. Client sends a write request to Primary
  - 5. Primary forwards the write request to all secondary replicas
  - 6. Secondaries reply to Primary
  - 7. Primary replies to Client

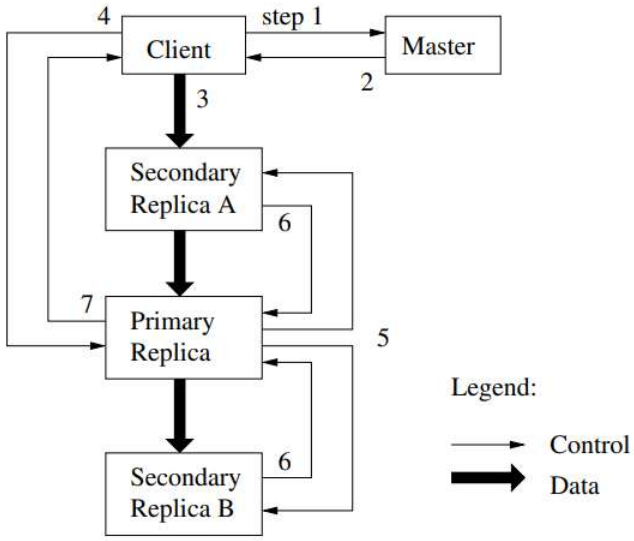


Figure 2: Write Control and Data Flow

[1] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03). Association for Computing Machinery, New York, NY, USA, 29–43.

相关技术
Fault-tolerance: Consensus&Replication

- Consistency model
  - Consistent & Defined
  - Ensure file region **defined** by performing same serial operations in all replicas
- Detect the state of servers
  - Handshake periodically
- Detect damaged or expired data
  - Checksum
  - Version control
- Replica choose
  - Nearest, fastest, lowest CPU utilization, Round-Robin, etc.