Evaluation and Analysis of GreenHDFS: A Self-Adaptive, Energy-Conserving Variant of the Hadoop Distributed File System

Rini T Kaushik, Milind Bhandarkar*, Klara Nahrstedt
University of Illinois, Urbana-Champaign, *Yahoo Inc.
Motivation

Data-intensive Computing Rapidly Popular
Advertising optimizations, Mail anti-spam, Data Analytics

Growing Hadoop Deployment
Open-source Hadoop platform of choice, Yahoo! 38000 servers, 170 PB

Escalating Energy Costs
Operating energy costs >= acquisition costs, Environmentally (Un)-friendly

Energy-Conservation in Hadoop Clusters Necessary
Energy-Conservation Techniques

Server & Cooling

• Scale-down
• CPU (DVFS, DFS, DVS)
• Disks
• Smart cooling

Possible power states: Active, Idle, Inactive (Sleep)
- Idle Power = 30-40% Active Power
- Sleep Power = 3-10% Active Power

Scale-down transitions servers from active to inactive (Sleep) power state → most energy-proportional

Scale-Down Very Attractive
Scale-down Mandates

Sufficient idleness
- To mitigate power state transition time, energy expended

No performance degradation

Few power state transitions
- To not reduce lifetime of disks
Existing Scale-Down Techniques

Workload Migration
- Chase et. al., SOSP’01
- G. Chen et. al., NSDI’08, …
- Con - Works if servers stateless

Always-ON Covering Primary Replica Set
- Leverich et. al., HotPower’09
- Amur et. Al., SOCC’10
- Con - Write performance impact
Unique Hadoop Scale-Down Challenges

- Hard to generate significant idleness
  Replicas and chunks distributed across cluster
- Workload migration not an option
  Servers NOT state-less
- Data-locality: Computations reside with data
Unique Hadoop Scale-Down Challenges

- Write Performance Important
  - Reduce phase of Map-reduce task
  - Production workloads such as click-stream processing operate on newly written data

Need More Scale-Down Approaches in a Hadoop Cluster
Our Approach - GreenHDFS

- Focus on energy-aware **data** placement instead of **workload** placement

- Exploit **heterogeneity** in data access patterns towards data-differentiated data placement

Meets all scale-down mandates and works for Hadoop
GreenHDFS

Opportunities for consolidation:
10-50% CPU Utilization *

In peak loads:
Compute capacity of Cold zone servers can be used

*Barasso et. al.
GreenHDFS Logical Zones

Hot Zone
- Performance-Driven Policies

Cold Zone
- Aggressive Energy-Driven Policies
- Minimize server wakeups
- No data chunking
- In-order file placement
- On-demand power-on
- Storage-heavy servers
- Reduces cold zone’s footprint

Zones Trade-off Energy and Performance
GreenHDFS Hot Zone Policy

- File Migration Policy
  - Dormant, low temperature data moved to Cold zone
  - Run during low periods of load

$\text{Coldness} > \text{Threshold}_{FMP}$
GreenHDFS Cold Zone Policies

- Server Power Conservation Policy
  - Server (CPU, DRAM & Memory) level

Dormant > Threshold_{SCP}

Active

Sleep

Wake-on-LAN
- File Access
- Data Placement
- Bit-Rot Scanning
- File Deletion
File Reversal Policy
- Ensures QoS of data that becomes hot after period of dormancy

$\text{Hotness > Threshold}_{FRP}$
GreenHDFS Goals

- Maximize energy savings
- Minimize data oscillations
- Minimize performance degradation

Can be achieved if none or few accesses to the Cold Zone
Policy Threshold Selection

- **Hot Space**
  - Low: File Migration Policy
  - High: Data Oscillations, Performance, Energy Savings

- **Energy Savings**
  - Low: Server Power Policy
  - High: Performance, State Changes

- **Performance**
  - Low: File Reversal Policy
  - High: Data Oscillations
Yahoo! Hadoop Cluster Analysis

- 2600 servers, 5Petabytes, 34 millions files
- 1-month of HDFS traces and metadata snapshots
- Multi-tenant production cluster
  - Analyzed 6 top-level directories → each signifies a tenant
    - Directories d, p, u, m
63.16% of total file count and 56.23% of total used capacity is cold (not accessed in 1-month)
Evolution of File

1. Create
2. First Read
3. Last Read
4. Delete

- Create
- Hot Lifespan\textsubscript{CLR}

- Dormant Lifespan\textsubscript{LRD}
  - Last Read
  - Delete

19
Lifespan$_{CFR}$ (Create -> First Read)

90% of data’s first read happens within 2 days of creation
89% of data is accessed for less than 10 days after creation.

Threshold_{FMP} should be > Lifespan_{CLR}
Lifespan_{LRD} (Last Read -> Delete)

80% of data in dir $d$ dormant for $> 20$ days
20% of data in dir $p$ dormant for $> 10$ days
0.02% of data in dir $m$ dormant beyond 1 day
Key Observations

- 89% of data in Yahoo! Hadoop compute cluster has a **news-server-like** access pattern
- Once data is deemed cold, low probability of it getting accessed again
  - Significant idleness in Cold Zone $\rightarrow$ high energy savings
  - Few accesses to Cold Zone $\rightarrow$ less performance degradation
  - System stable – less data oscillations

Great for GreenHDFS Goals
Evaluation

- Trace-driven simulation driven by 1-month long hdfs traces from a 2600 server/ 5pb cluster for main directory dir \( d \)
- Hot zone \( \rightarrow 1170 \)
- Cold zone \( \rightarrow 390 \)
- Assumed 3-way replication in both zones
- Used power and transition penalties from datasheets of Quad Core Intel Xeon, Seagate Barracuda SATA disk

*not representative of Yahoo H/W Configuration*
24% cost savings, $2.1 Million $\rightarrow$ 38000 servers, in reality more savings (cooling, idle power in Hot zone)

Minimally sensitive
Only 6.38TB worth of data migrated daily
Migrations/Reversals

Insignificant file reversals
Data oscillations & energy savings insensitive to the File Migration Policy threshold.
Improved Free Space in Hot Zone

More free space in Hot zone → more hot data
Max power state transitions observed = 11, no risk to disk longevity
Results in significant energy cost reduction as shown with real-world large-scale traces from Yahoo! Hadoop Cluster

- Insensitive to thresholds

Allows effective server-level scale-down in Hadoop Cluster

- Generates significant idleness in Cold Zone
- Few power state transitions
- No write performance impact
Thank You