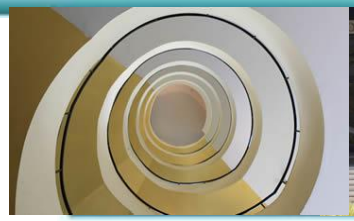
PES UNIVERSITY






Mini Project Progress Review #2

Project Title: Energy efficiency for HDFS

Project ID: MPW20HLP03

Project Guide: Prof. H L Phalachandra

Project Team: Abhishek Das (PES1201800177)
N Sanketh Reddy (PES1201800389)
Bhargav SNV (PES1201800308)

Project Abstract and Scope

Abstract:

- With growth of big data, the need for distributed computing and clusters is higher than ever.
- By using distributed computing, tasks can be split and processed in parallel. But the energy consumed by clusters to compute is high.
- Almost 55% of resources aren't used in clusters. They remain in their idle state, wasting a lot of energy.
- Our project aims to conserve energy in these clusters.

Project Abstract and Scope

Scope:

- The scope of this project is to reduce energy consumption in HDFS. We will simulate various conditions to devise effective means of energy conservation.
- This project will not involve working on the Hadoop codebase.
- Project will use simulation software to test HDFS power consumptions and further analysis will be done on the same.

Further Literature Survey

Relevant Research Papers:

- **Dynamic Energy Efficient Data Placement and Cluster Reconfiguration Algorithm for MapReduce Framework. N. Maheshwari, R. Nanduri, V. Varma**

  This paper discusses energy ***efficient data placement*** and ***cluster reconfiguration(balancing)*** which dynamically scales clusters in accordance to the workload imposed on it.

Shortcomings: This methodology requires frequent dynamic reconfiguration of the cluster. This is a resource heavy operation.

- **GreenHDFS: Towards An Energy-Conserving, Storage-Efficient, Hybrid Hadoop Compute Cluster. R. T. Kaushik, M. Bhandarkar**

  This paper describes the concept of using ***"Hot and Cold zones"***, where Hot Zones have high computational hardware and contain data that is frequently used. Cold Zones have larger disks and energy conserving hardware which store large amounts of data that are not frequently used.

Shortcomings: There is no dynamic policy to shift data among Zones. Scaling zones is also an issue.

Further Literature Survey

Relevant Research Papers:

- **Hybrid HDFS: Decreasing Energy Consumption and Speeding up Hadoop using SSDs. I. Polato, F. Kon, D. Barbosa and A. Hindle**

  This paper describes the use of ***different types of disks*** to improve energy and compute efficiency. The file system is split into two storage zones, one ***using SSD and the other with HD.*** A predefined policy determines which zone a block is stored in.

Shortcomings: This requires a lot of manual configuration and policies for storage must be updated frequently. These processes are not dynamic.

- **Scheduling and Energy Efficiency Improvement Techniques for Hadoop Map-reduce. N. Tiwari**

  This paper discusses various ***scheduling algorithms*** to save energy for MapReduce tasks. A few algorithms discussed are Delay algorithm, Constraint scheduling algorithm, schedulability test algorithm.

Shortcomings: All algorithms discussed work on Map Reduce layer (Application) of Hadoop, it does not help with energy conservation at the HDFS (Storage) layer.

User Characteristics

- Companies running large scale Hadoop clusters would find it beneficial to adapt our model.
- Several types of data streaming into a large amount of servers.
- Server running/maintenance power costs substantial
- Big Data servers have constant middle to high resource utilization

Dependencies / Assumptions / Risks

Legal Implications:

All software, tools and frameworks used are open source and have no legal implications

Software Dependencies:

- Java environment with IDE.
- HDFS-Sim setup.

Hardware Dependencies:

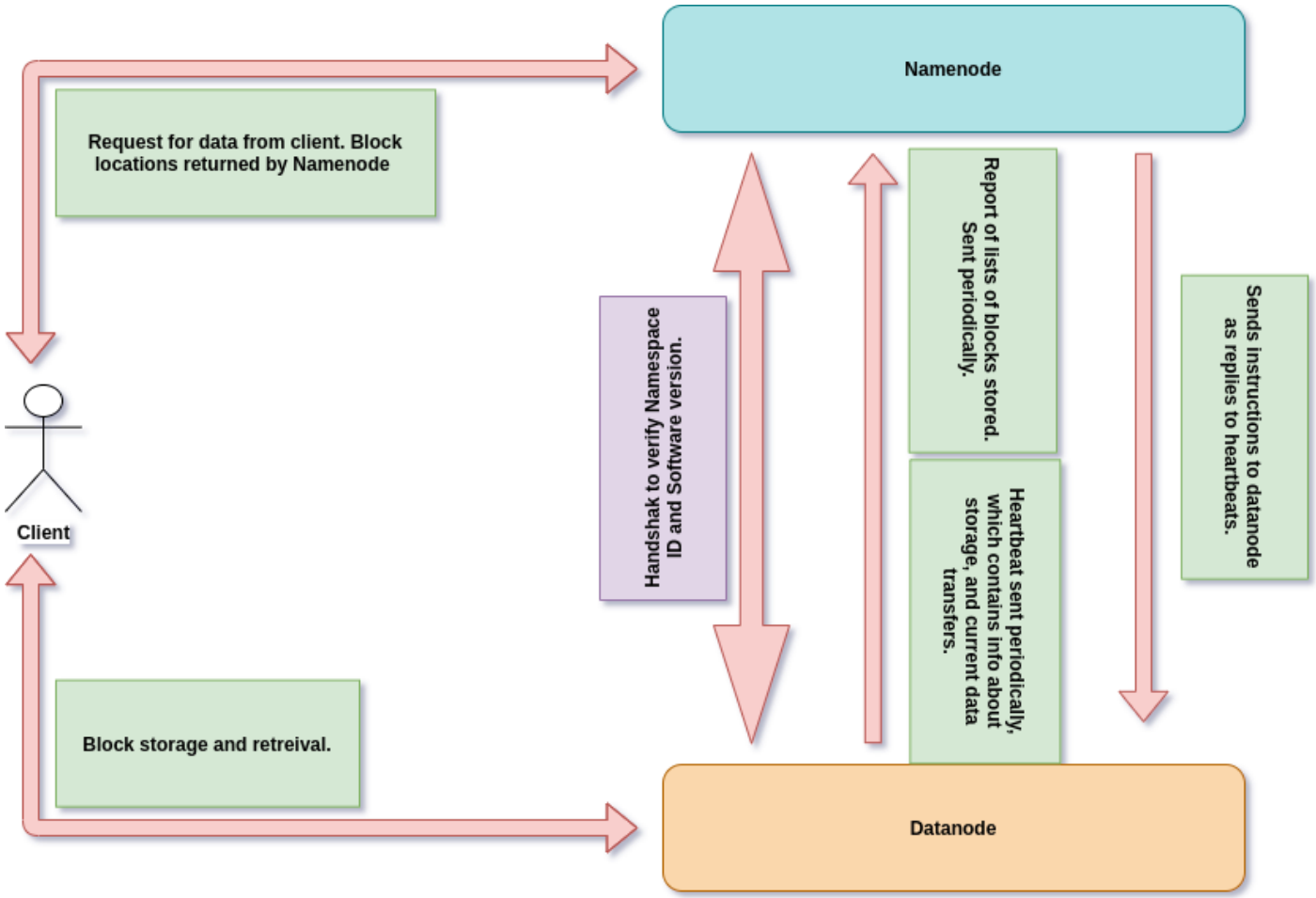
- Systems used have a minimum of Intel i5 cores with 8 GB ram and sufficient storage space.

Dependencies / Assumptions / Risks

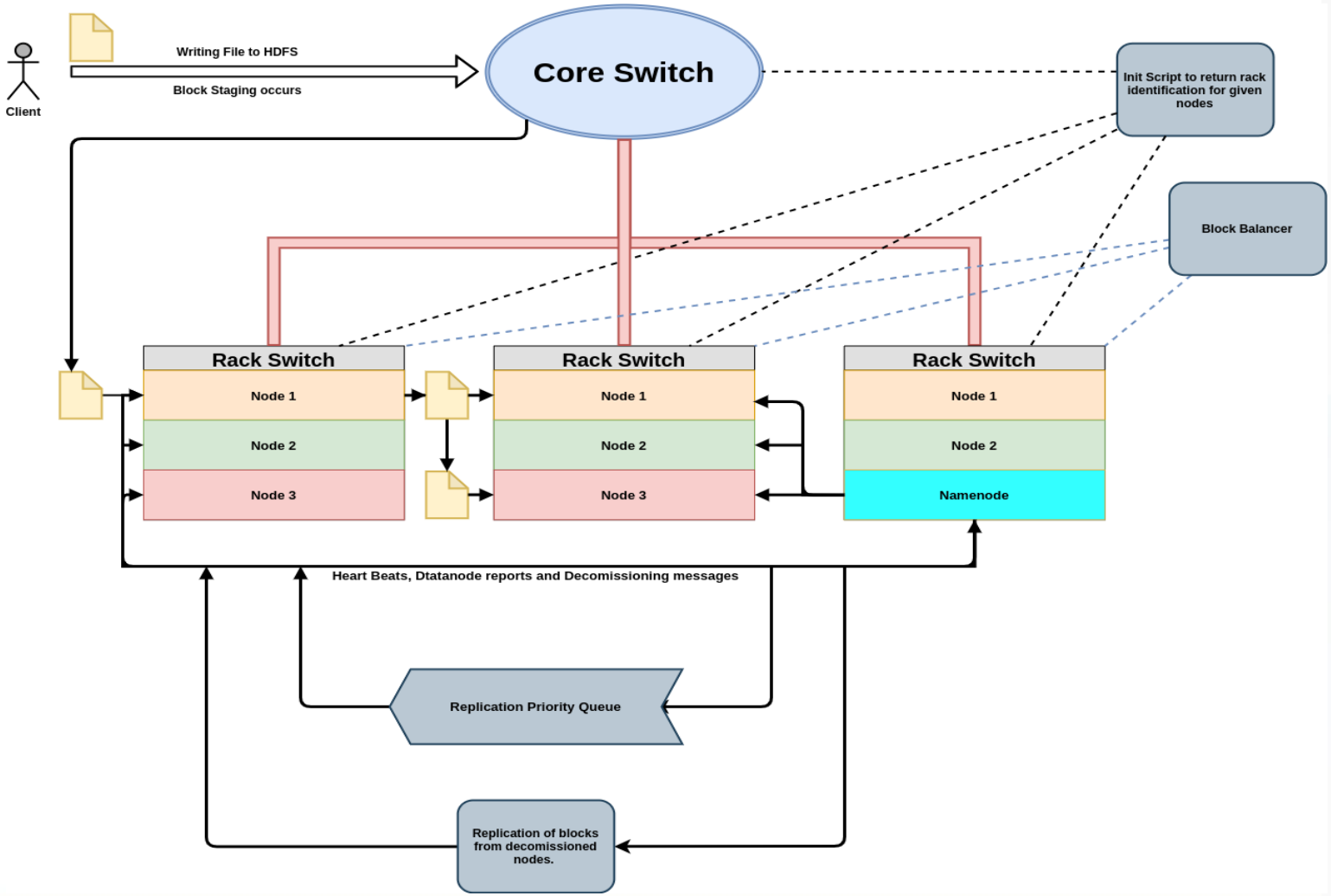
Risks

- Algorithms works only in theory but not in real-world scenarios.
- Failure of integration of algorithm with Hadoop code base.
- Simulator cannot measure energy consumption for our setup.

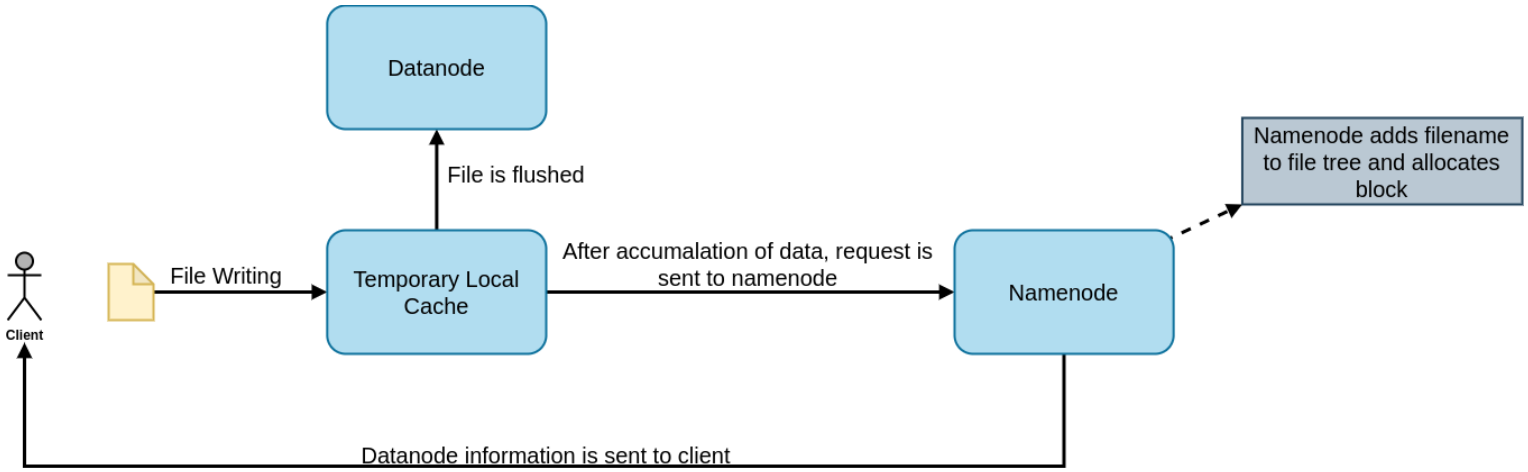
Hadoop Architecture



Hadoop Architecture



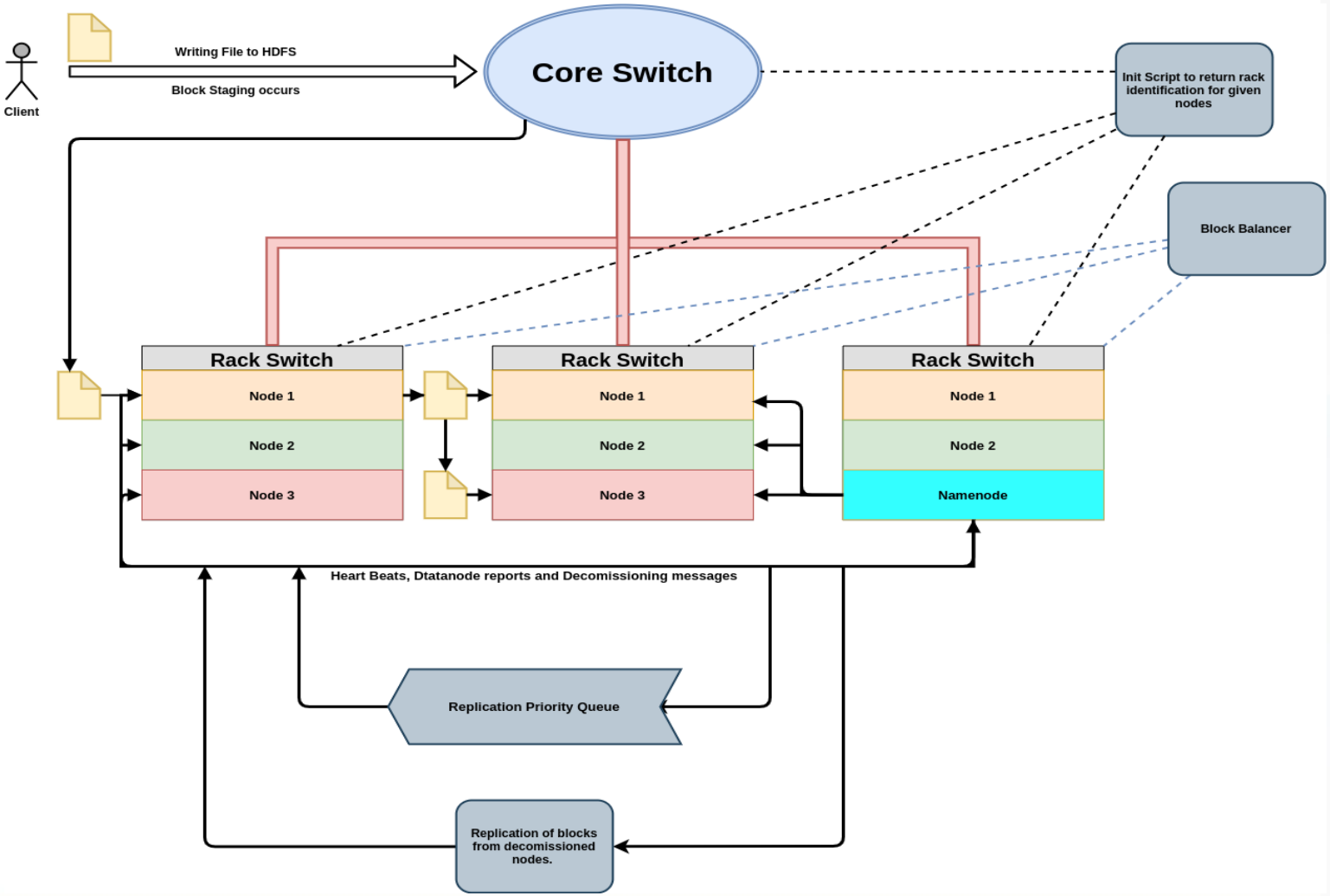
Hadoop Architecture (Staging)



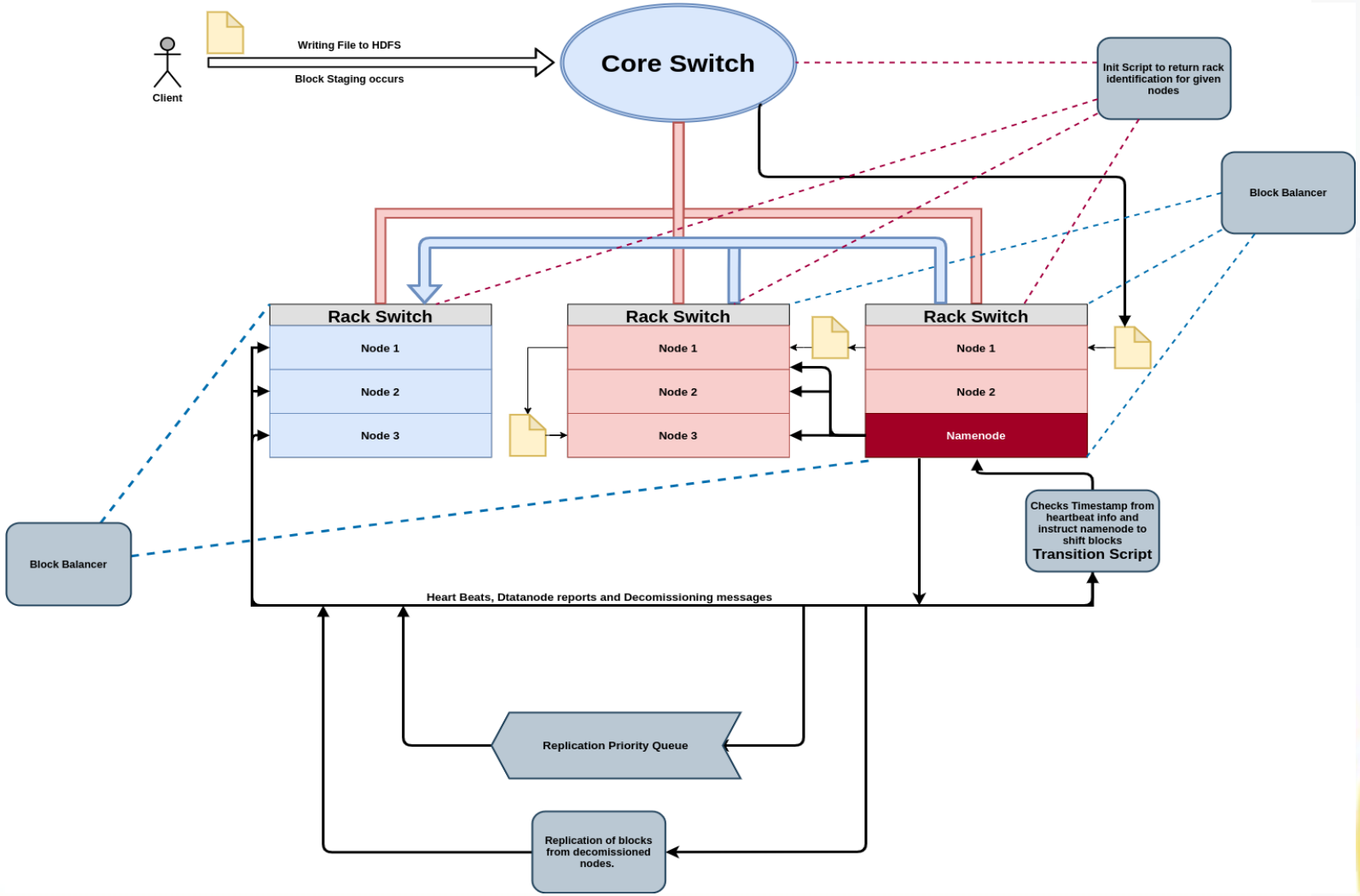
Modules

- HDFS replication simulator: To evaluate effect of deployment platform and workload characteristics on data availability.
- A new block placement policy and heartbeat mechanism.
- Dynamic data storage/transfer mechanism between hot and cold zones.

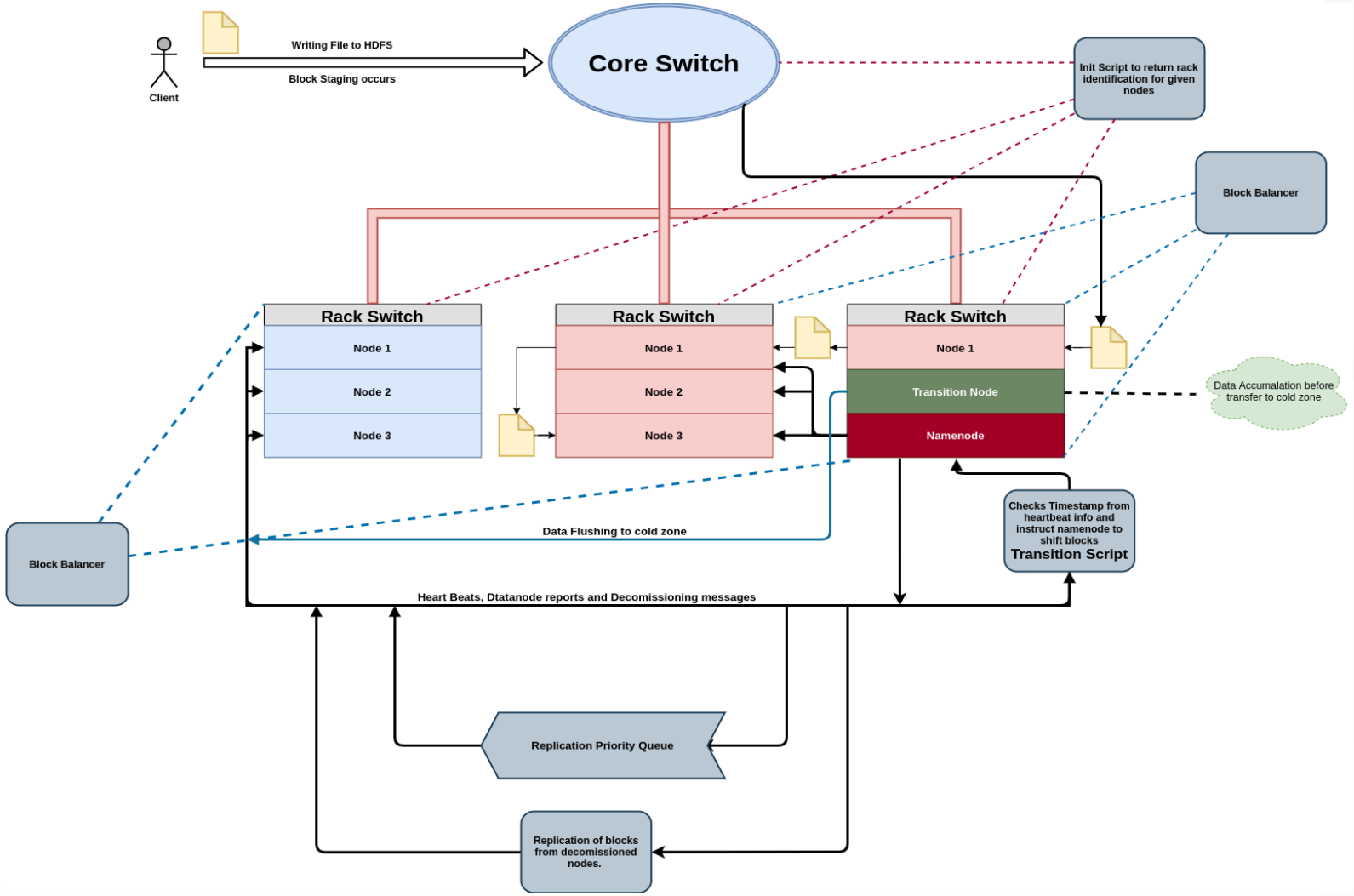
Actual Hadoop Architecture



Policy and Algorithm 1 of 2 (Infrequent Transfer)



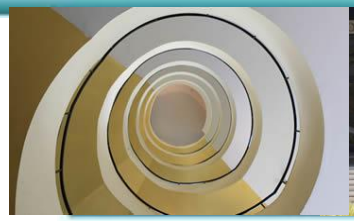
Policy and Algorithm 2 of 2 (Frequent Transfer)



Technologies Used

- HDFS replication simulator
- Java
- Draw.io

PES UNIVERSITY






Thank You