Exploration of Computing Technologies

Relevant for FCPPL

Fabio Hernandez

fabio@in2p3.fr

4th FCPPL Workshop
Jinan (China), April 2011

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DES AFFAIRES ÉTRANGÈRES ET EUROPÉENNES

中國科學院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

CCIN2P3

Introduction

- This presentation is a short overview of a new project in the *Related Technologies* chapter submitted for funding to FCPPL in 2011

- Topics
  1. *France-Asia computing platform*
  2. *Usage analysis of the international network links*
  3. *Cloud-based file storage*

- Partners

  IHEP computing centre

  CC-IN2P3

France-Asia Grid Platform

- To join the effort initiated by KISTI (Korea), CCIN2P3 (France), recently joined by KEK (Japan) for deploying a grid-based computing platform

  *Aiming to support the joint research activities of the involved countries, in the framework of the F*PPLs*

- Concrete goals

  *to help interested Chinese **organizations** join this platform by contributing computing and storage resources*

  *to help **researchers** effectively use this facility by providing the necessary support*

France-Asia Grid Platform (cont.)

- Available services

  user interface, virtual organization management, job submission, storage elements, computing elements, file and replica catalogues

- Middleware: gLite

- Current usage: in-silico docking, QCD simulations, Geant4/GATE applications

Source: S.Hwang, FKPPL Workshop, March 2011 http://indico.in2p3.fr/conferenceTimeTable.py?confId=4516

France-Asia Grid Platform (cont.)

- TREND* interested in using this platform for its data processing needs
- We intend to explore DIRAC for centrally managing the workload of TREND on the France-Asia grid platform

Will benefit from the support of the DIRAC experts at IN2P3/ CPPM

http://dirac.in2p3.fr/DIRAC/

*Tianshan Radio Experiment for Neutrino Detection

Network usage analysis

- Goal: build a platform for collecting and analyzing data related to the usage of the international links of IHEP computing centre

- Two main expected benefits: collect information for **capacity planning** purposes and for day-to-day **network operations**

To be able to answer questions such as:

1. *what are the top 10 sites (in terms of volume) we exchange data with?*
2. *what is the network usage associated to each major experiment supported by IHEP CC?*
3. *what network protocols are actually using the links?*

In addition, provide means for detecting changes in the usage patterns, for instance for helping detecting intrusions, abnormal usage or abuse

Network usage analysis (cont.)

- Current status

  Collecting data since late 2010 and storing them in a relational database

  The storage component needs to be revisited to better suit the type of queries and to scale up. Will explore distributed hash tables for this (more on this later).

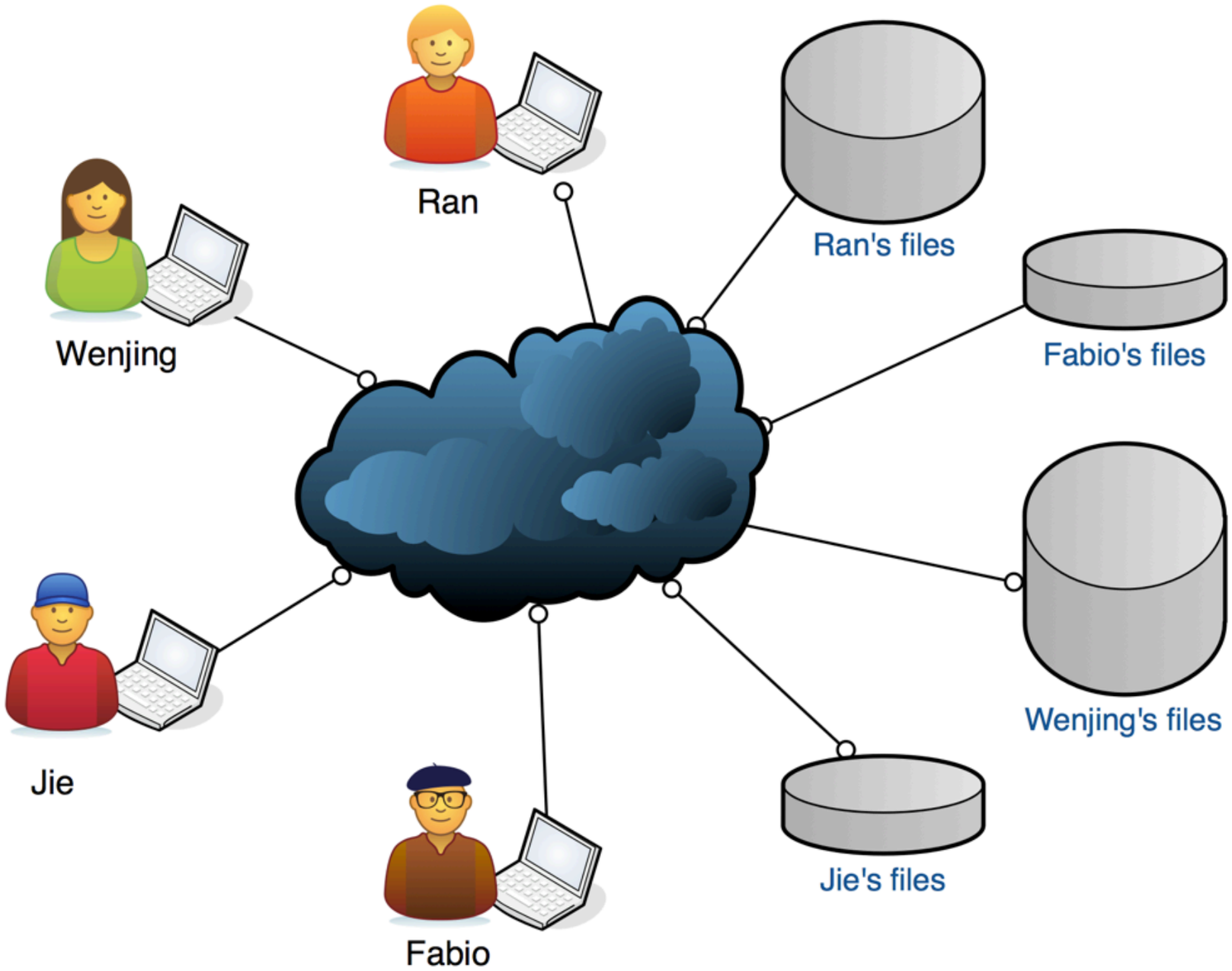
- CC-IN2P3 expressed interest in the results of this activity

FileStore: cloud-based file storage

- Goal: to provide individual researchers a **convenient** service for storing their files

*Think of it as a **personal storage element**, accessible both from the user's desktop and from the grid jobs*

FileStore (cont.)



Your files are physically stored on remote servers accessible seamless through the network.

You interact with your remote files as you usually do with your local files, using your desktop's metaphors (file explorer, drag & drop, etc.).

You are free to organize your own storage space.

The system provides you significantly more storage capacity than is locally available in your personal computer.

You can share your files with selected users of the system.

FileStore (cont.)

- Who this service is intended for?

  individual *scientists, initially those of the HEP community*

- Why would you use this service?

  because it provides a ***convenient*** *way for storing your individual files, for instance, the logs of your (grid) jobs, datasets you are analyzing, etc.*

  because you may want to ***share*** *some files with some colleagues sitting at the next door or across the world*

  because you would have total ***flexibility*** *to manage your own space according to your needs and working methods*

  because it provides you more storage ***capacity*** *than you have available on your personal computer*

FileStore (cont.)

- Who would operate the service?

  computing centres of the HEP community: good national and international connectivity, 24x7 service, expertise in running IT services, trustable, reliable, ...

  files would be physically located in disk servers managed and operated by one or more of those centres

- Reduced set of basic operations
  - ***list, store, retrieve, delete files***
  - ***create, delete directories***

  although the system does not expose complete POSIX semantics, emulation will be possible for compatibility

FileStore (cont.)

- Use-case profile

  retrieval of files more frequent than storage

  repository used as a high-capacity highly-available archive system: ***not intended to directly serve I/O-intensive applications***

  in addition to access files from the desktop, grid jobs and jobs submitted to a particular site can also interact with the repository for storing/retrieving files

- Scale

  initially to provide each user an individual storage capacity of 1TB to 2TB

  thousands of directories, tens of thousands of files, per user

  file size in the region of 1B to 5GB, but most of the files expected in the area of a few hundreds MB

FileStore (cont.)

- Current status

  Evaluating candidate open-source implementations of Key-Value stores as the back-end for storing the files and the associated meta-data

  Candidates: Cassandra, OpenStack's Swift, Project Voldemort, and possibly Riak

  Building a demonstrator of the concept by using in-memory key-value stores

- What is next

  Select the back-end store and perform scalability tests

Questions & Comments

fabio hernandez
fabio@in2p3.fr
CAS/IHEP Computing Centre
CNRS/IN2P3 Computing Centre

Backup Slides

fabio hernandez
fabio@in2p3.fr
CAS/IHEP Computing Centre
CNRS/IN2P3 Computing Centre

Key-Value Store

- Data store exposing a simplified interface, basically composed of 3 operations

```
bool put(string key, byte[] value);

byte[] get(string key);

bool delete(string key);
```

- Data structure supported by several programming languages, known also as associative arrays

- The semantics of *value* are client-defined: no predefined schema

Key-Value Store (cont.)

- Heavily used as a storage back-end for (large scale) web applications

  user profiles, user sessions, shopping carts, blog entries and their associated threaded discussions, streams of data, …

  Amazon, Google, Twitter, Baidu , Facebook, LinkedIn, Yahoo, ...

- Benefits

  very fast, scalable, flexible schema, flexible datatypes, programmer-friendliness, …

- Constraints

  modeling the problem domain in terms of key-value mappings

  no transactions, no joins, ...

  logic for resolving potential inconsistencies may need to be added to the application

Key-Value Store (cont.)

Cassandra

riak

Project Voldemort
A distributed database.

Google labs

amazon web services™

redis

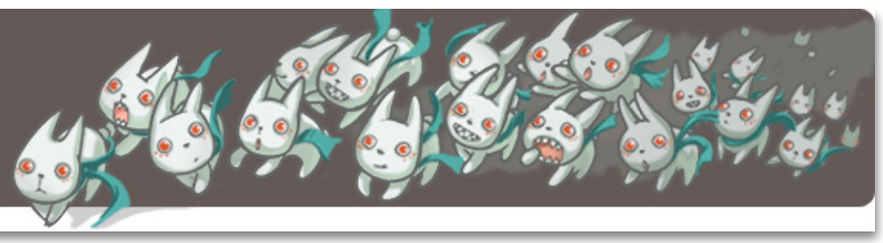
mongoDB
{name: "mongo", type:"DB"}

Tokyo Cabinet 8192PiB

openstack

apache CouchDB relax

MEMCACHED



Dynamo Characteristics

- Availability

  data is automatically (and asynchronously) replicated to multiple file servers within the same data centre or across data centres

- Symmetry & decentralization

  every file server plays an identical role: no single point of failure, no network bottleneck

- Manageability

  failed file servers can be replaced with no downtime

  additional servers can be added without service interruption

- Elasticity

  read and write throughput both increase linearly as more file servers are added to the system

- Heterogeneity

  system exploits the heterogeneity of the infrastructure (network throughput and latencies, file server capacities, etc.). This allows for adding new more powerful file servers without having to upgrade them at once

- Eventual consistency

  replicated copies of the same data may not all be consistent at any given moment in time, but the system is designed for them to be eventually consistent

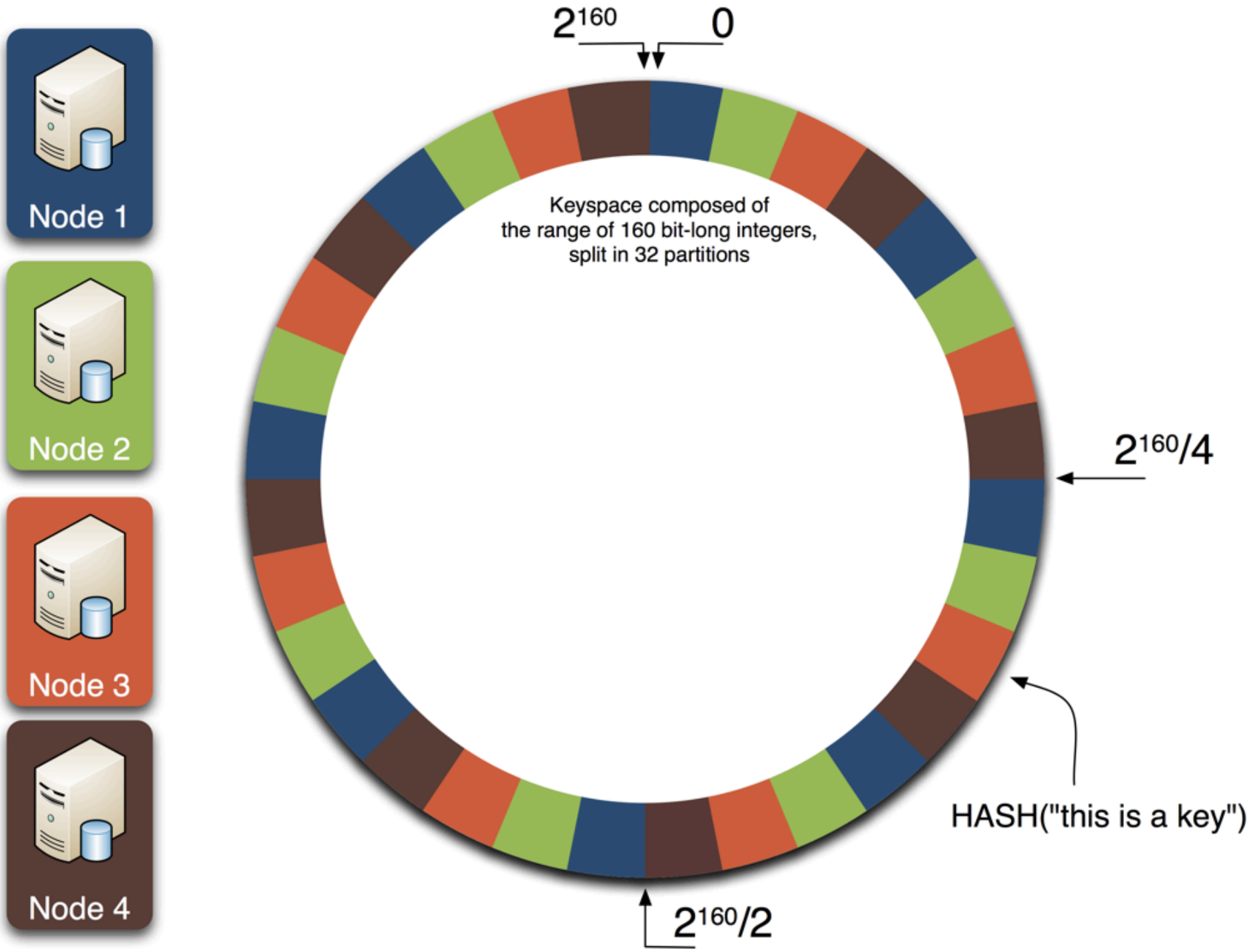
  conflict resolution may be implemented at the store level or at the application level, e.g. "last write wins"

- Data model

  distributed key-value store

  partitioning of data among the file servers is done using consistent hashing

Dynamo: Ring



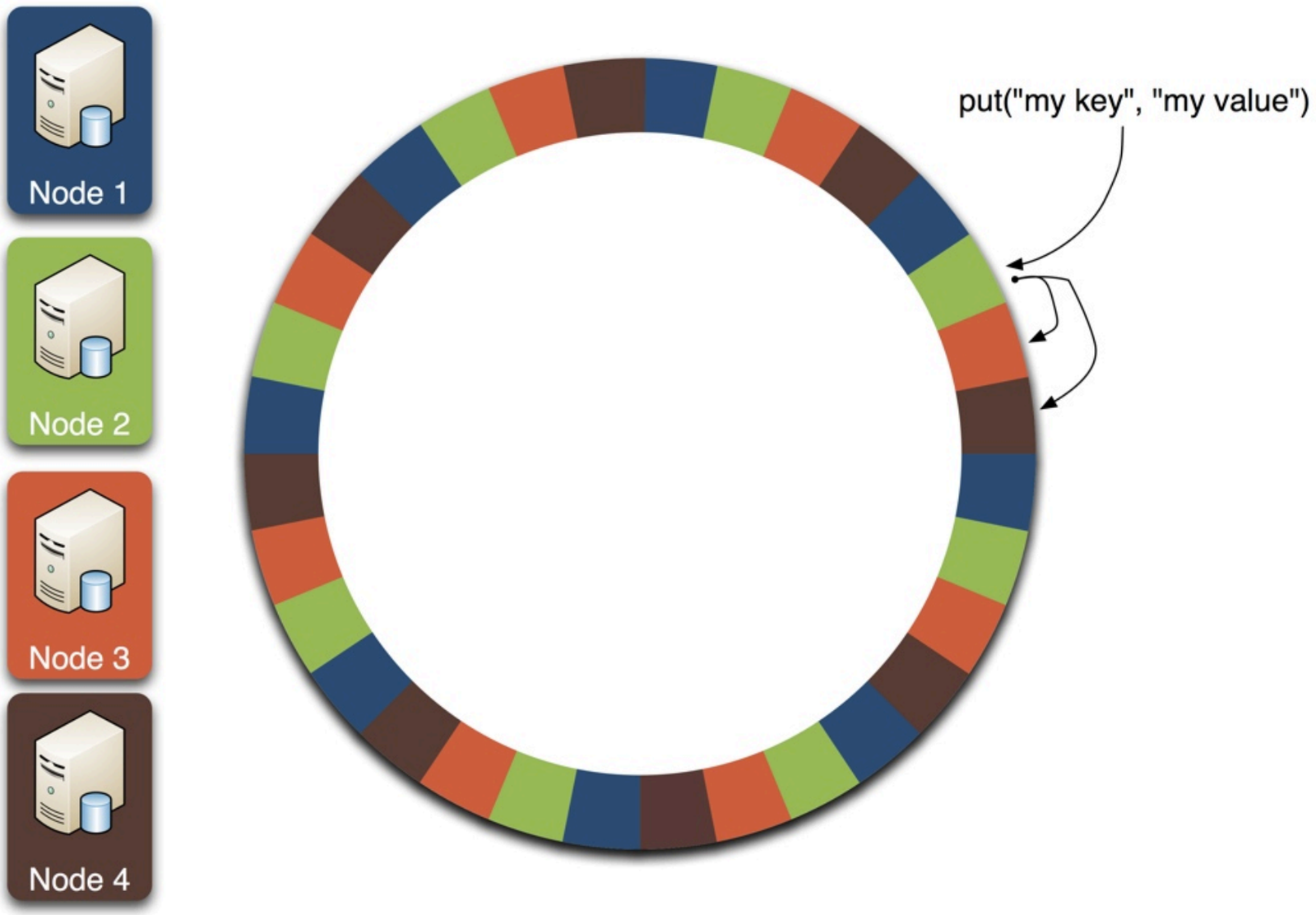
File servers are organized in a logical ring.

Each server is responsible for storing the values associated to one or more partitions in the key space.

Each node in the ring knows which peer node is responsible for storing & retrieving the value associated to a key.

Adapted from http://www.basho.com

Dynamo: Replication



The node responsible for handling operations associated to the key is selected as the coordinator of the request.

After storing the [key,value] pair locally, it replicates it to its available neighbors (N=3 in this case).

The number of replicas to keep (N) and the minimum number of replicas to read (R) or write (W) to consider a request successful is configurable.