HDFS 2015: Past, Present, and Future

9/30/2015

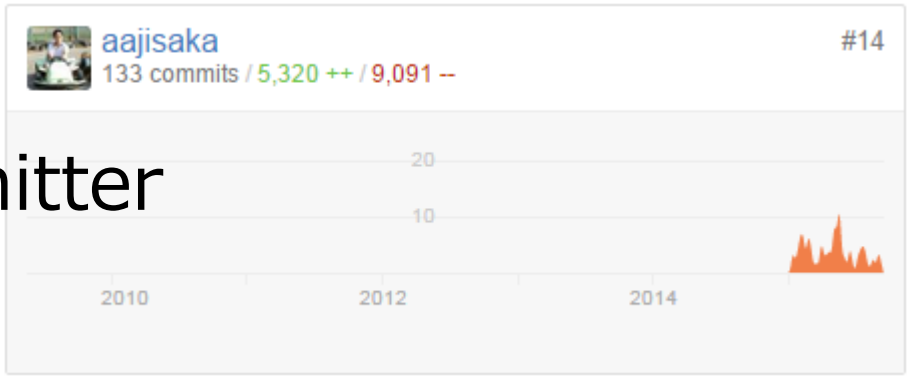
NTT DATA Corporation
Akira Ajisaka

NTT DATA

Self introduction

NTT DATA

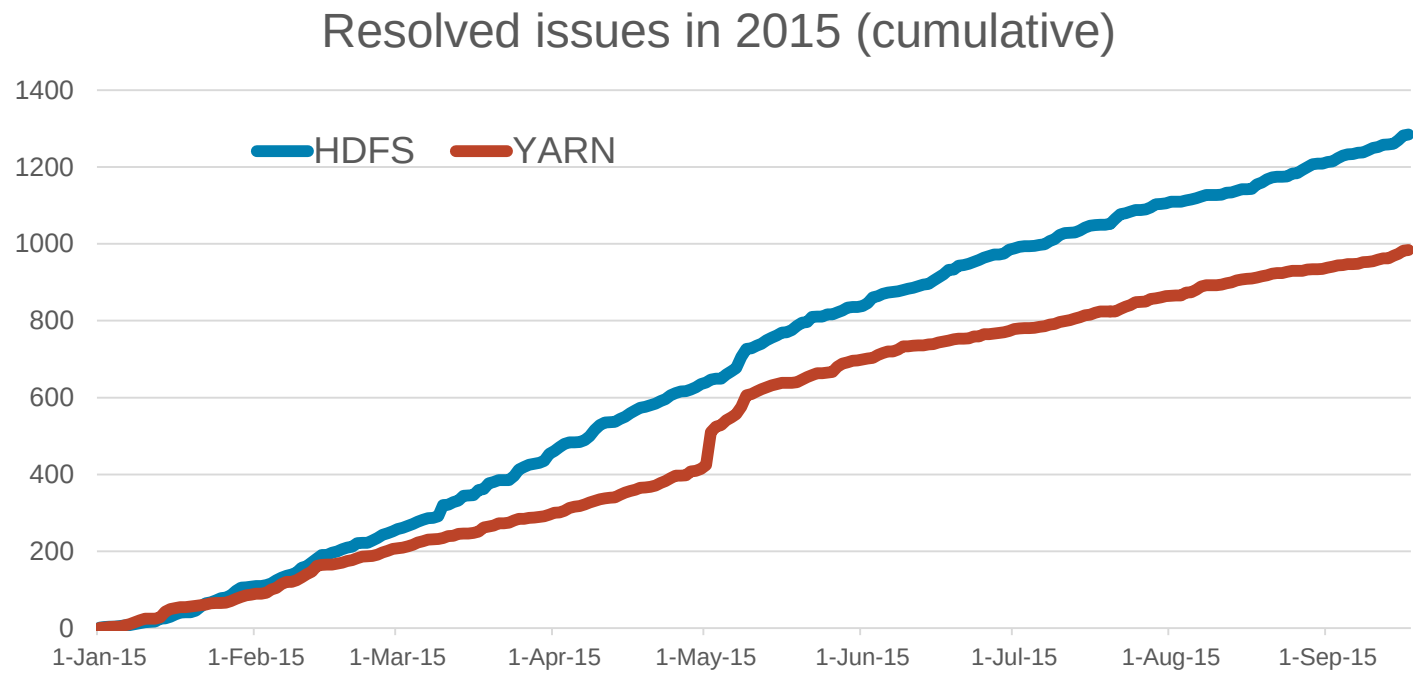
- Akira Ajisaka (NTT DATA)
  - Apache Hadoop Committer
    - 130+ commits in 2015
    - Working on usability
    - 80+ documentation patches
  - "Open-Source Professional Services" team
    - Has deployed and supported 10k+ nodes of Hadoop clusters overall for 7 years
    - Contributing to Apache Hadoop 6th in the world with NTT [1]



[1] The Activities of Apache Hadoop Community 2014
http://ajisakaa.blogspot.com/2015/02/the-activities-of-apache-hadoop.html

About

- Similar to "YARN 2015" presentation by @tshooter
- HDFS is developed faster than YARN



- Need a summary of HDFS new features

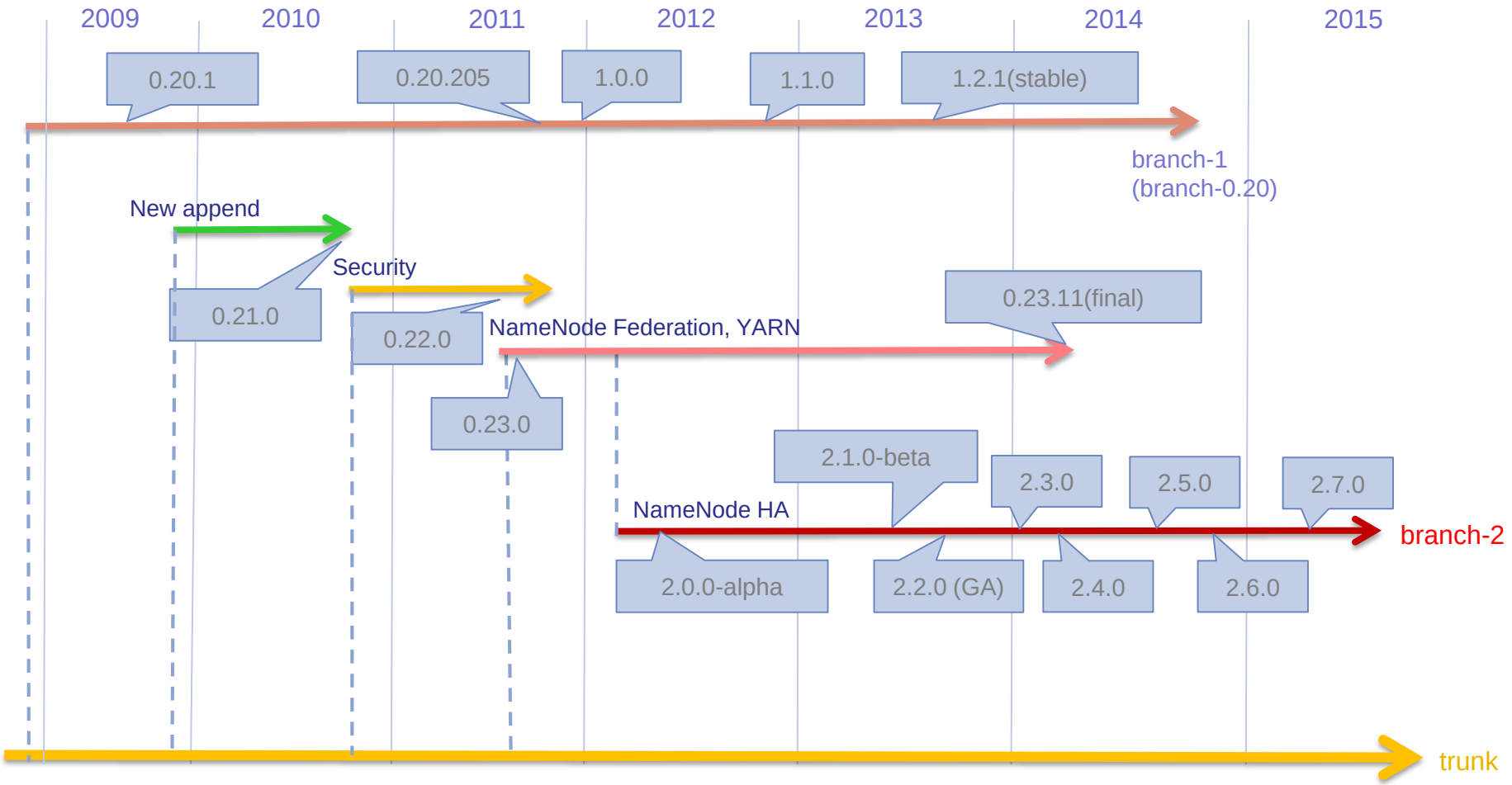
Agenda

- Past
- Present
- Future

Past

Past releases

- 2.X is the release branch
- 1.X and 0.23.X are no longer maintained

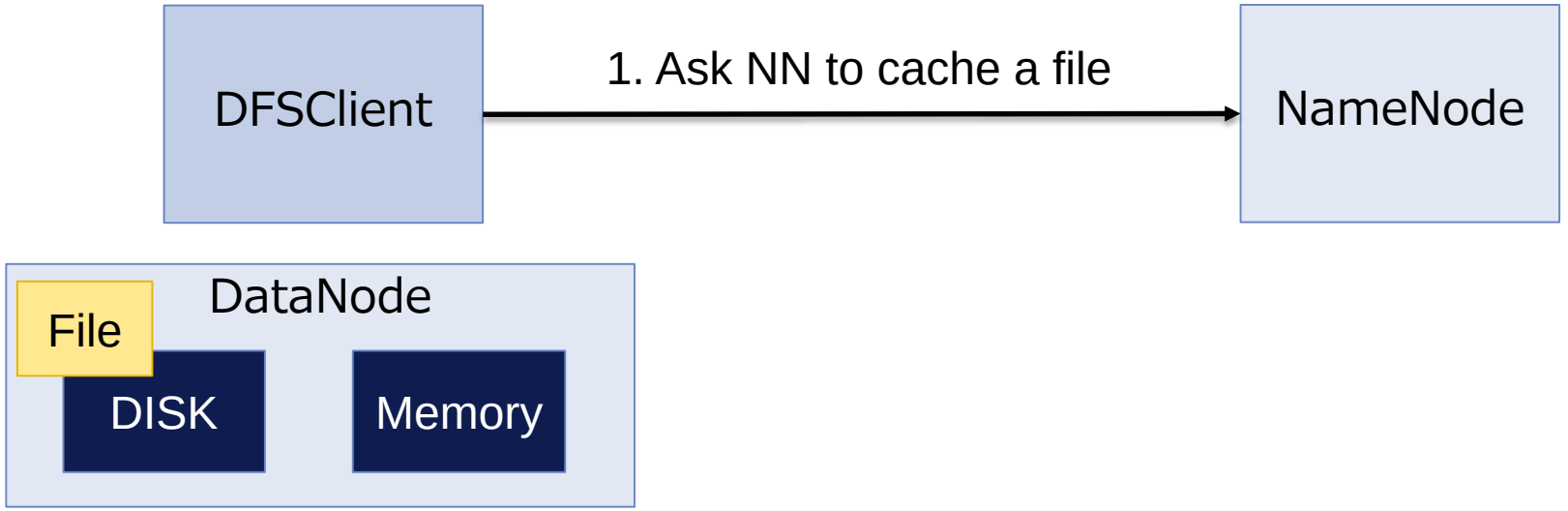


Hadoop 2.2 (2013-10-13)

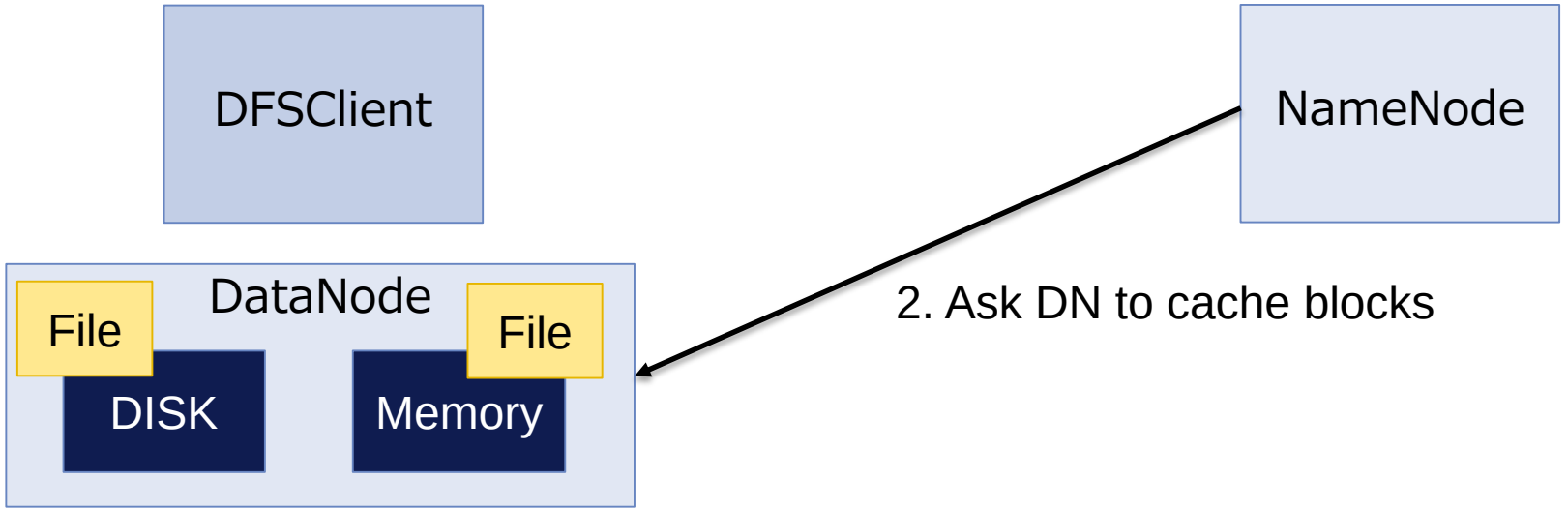
NTT DATA

- NameNode High-Availability
  - No Single Point of Failure
- Federation
  - Multiple NameNodes, multiple namespaces
  - Improve scalability
- Snapshots
  - Read only point-in-time copy (Copy on Write)
- NFSv3 mount

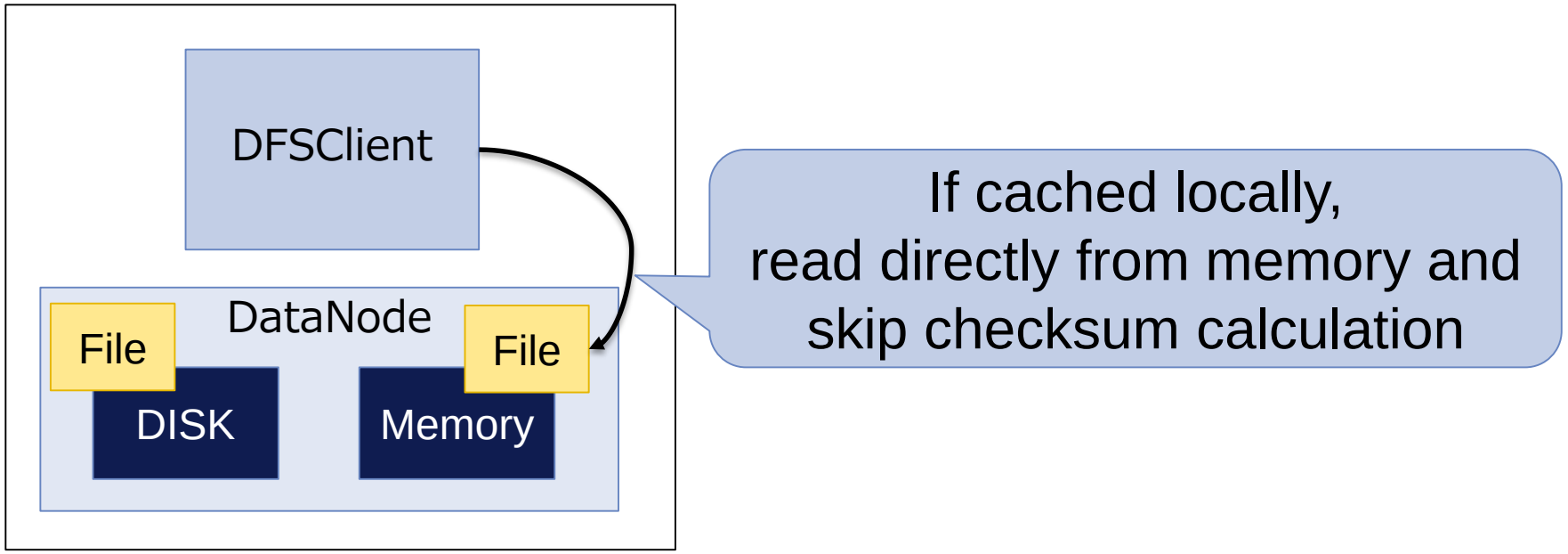
- Heterogeneous Storages (Phase 1)
- In-memory caching
  - Introduce memory-locality
  - Make efficient use of memory in DNs



- Heterogeneous Storages (Phase 1)
- In-memory caching
  - Introduce memory-locality
  - Make efficient use of memory in DNs



- Heterogeneous Storages (Phase 1)
- In-memory caching
  - Introduce memory-locality
  - Make efficient use of memory in DNs



Hadoop 2.4 (2014-04-07)

NTT DATA

- Rolling Upgrades
  - No need to wait for hours
- ACLs
  - More fine-grained permissions
  - Similar to POSIX ACL

```
-rw-rw-r--   3 tester hadoop   129 2015-09-15 12:00 /user/tester/test.txt
```

$ hdfs dfs -setfacl -m group:hive:rw- /user/tester/test.txt
gives write permission to hive group

Hadoop 2.5 (2014-08-11)

NTT DATA

- Extended Attributes (XAttrs)
  - Similar to extended attributes in Linux

```
-rw-r--r--   3 tester hadoop   129 2015-09-15 12:00 /user/tester/test.txt
```

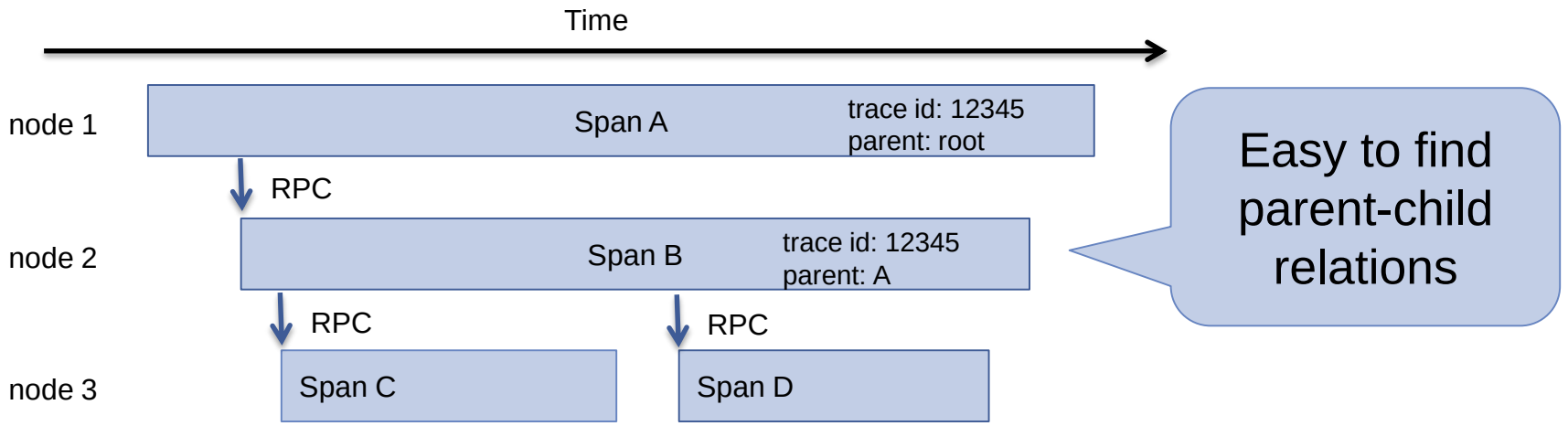
Set XAttrs
$ hdfs dfs -setfattr -n user.locale -v jp /user/tester/test.txt
$ hdfs dfs -setfattr -n user.city -v tokyo /user/tester/test.txt

Get XAttrs
$ hdfs dfs -getfattr -d /user/tester/test.txt
# file: /user/tester/test.txt
user.locale="jp"
user.city="tokyo"

  - Currently used by transparent encryption

Hadoop 2.6 (2014-11-18)

- Hot swap volumes
  - Recover from disk failures w/o stopping DNs
- Integrate Apache HTrace (incubating)
  - Trace RPCs inside HDFS



  - Finding bottlenecks becomes easier

Hadoop 2.6 (2014-11-18) (Cont.d)

NTT DATA

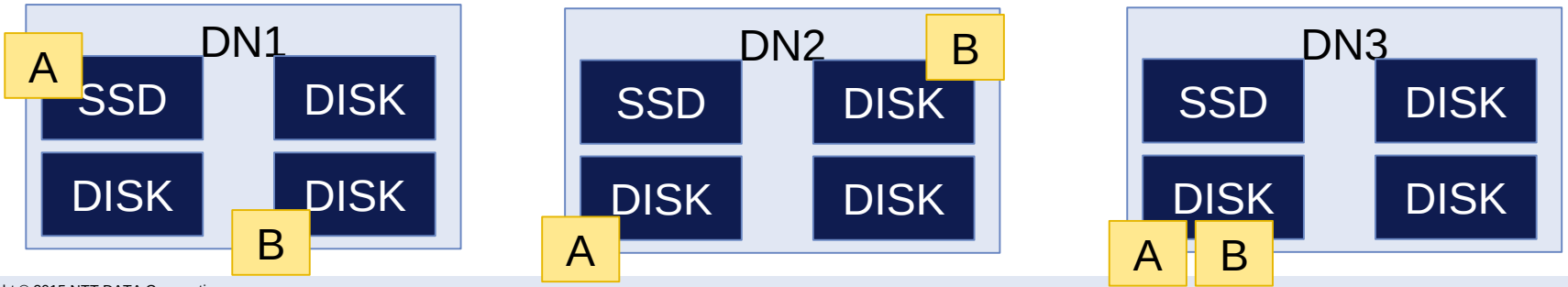
- Heterogeneous Storages (Phase 2)
  - Archival Storage
  - Memory as storage tier
- Transparent Encryption

Heterogeneous Storages

NTT DATA

Hadoop 2.6

- Problem
  - SSD is getting cheaper
  - Want to store hot data in SSD to achieve higher throughput
- Solution: Introduce storage type and block placement policy
  - Storage: HDD, SSD, ARCHIVE, ...
  - Policy: One_SSD, HOT, WARM, COLD, ...
  - Example: A -> One_SSD, B -> HOT



Heterogeneous Storages

NTT DATA

Hadoop 2.6

- How to use
  - Configure HDFS to recognize storage type for each disk

```
<parameter>
  <name>dfs.datanode.data.dir</name>
  <value>[SSD]file:///data/ssd,[HDD]file:///data/hdd</value>
</parameter>
```

  - Set block placement policy to HDFS path

```
$ hdfs setstoragepolicies -setStoragePolicy -path <path> -policy <policy>
```

    - Reset policy after putting data is possible
    - Mover will move blocks to satisfy the policy considering rack awareness

Archival Storage

NTT DATA

Hadoop 2.6

- DISK or ARCHIVE?
  - ARCHIVE is for cold data
- eBay reduces cost/GB by 5x [1]
  - Use low-spec DNs for ARCHIVE

	Regular Node	Archival Node
Drives	12 HDDs	60 HDDs
CPU	32 Cores	4 Cores
Memory	128GB	64GB
Run NodeManager	Yes	No

- No need to split cluster!

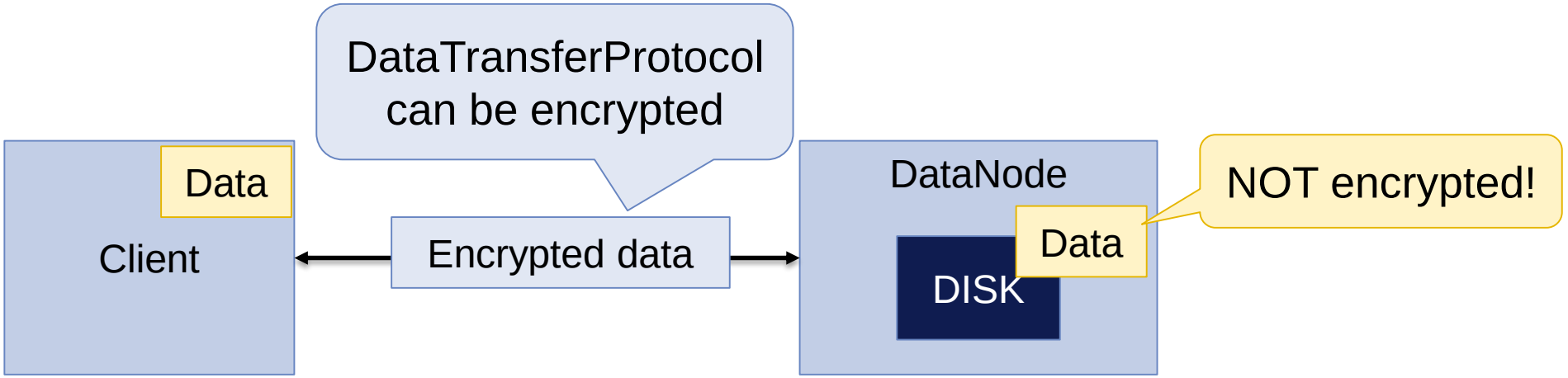
[1] Reduce Storage Costs by 5x Using The New HDFS Tierd Storage Feature
http://www.slideshare.net/Hadoop_Summit/reduce-storage-costs-by-5x-using-the-new-hdfs-tiered-storage-feature

Transparent Encryption

NTT DATA

Hadoop 2.6

- Problem
  - Cannot guard data from OS-level attacks



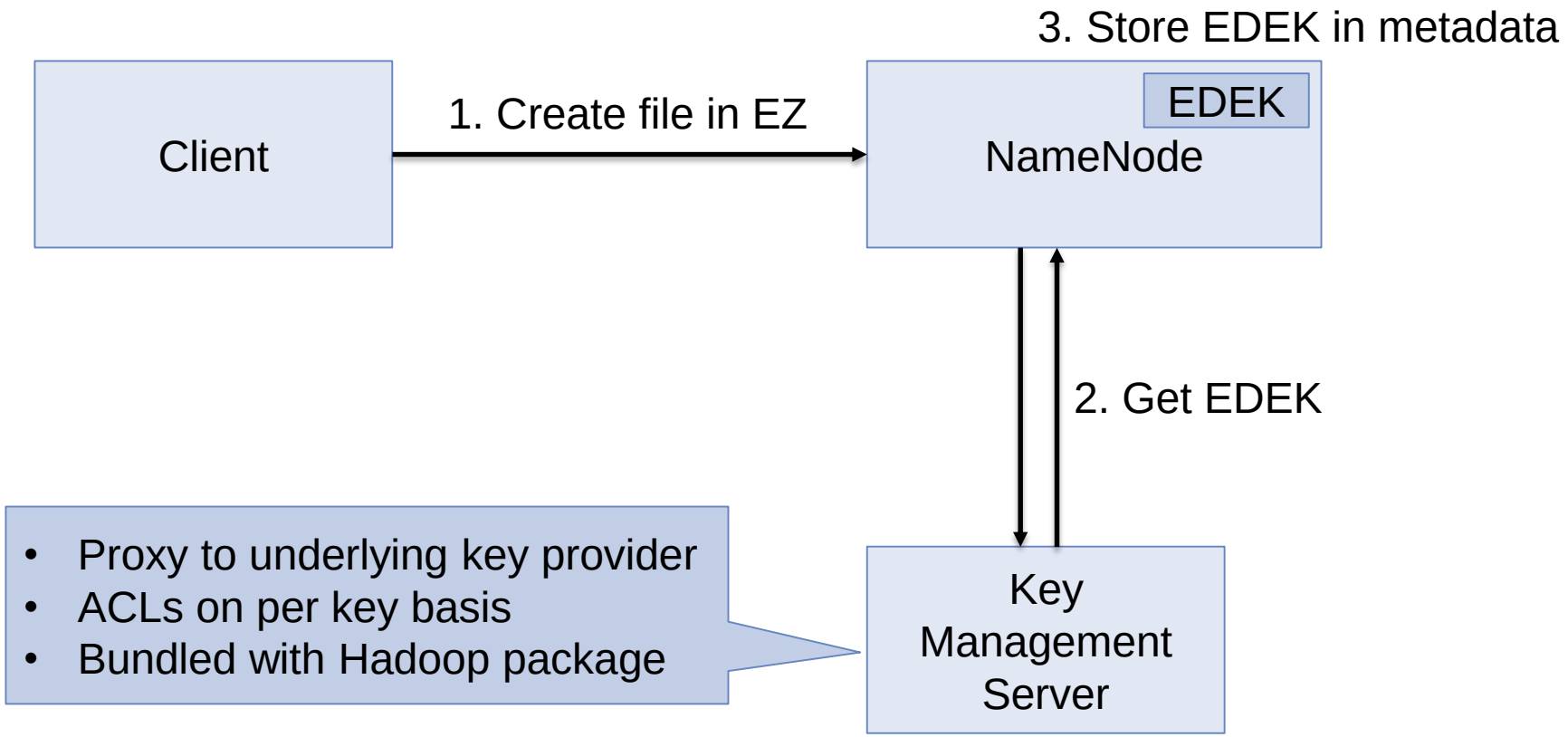
- Solution
  - Provide end-to-end encryption
  - Encrypt/decrypt data transparently
    - No need to rewrite user application

Transparent Encryption: How to encrypt data

NTT DATA

Hadoop 2.6

- DEK (Data Encryption Key)
  - A unique key for each file in EZ (Encryption Zone)
  - Stored in an Xattr of the file, encrypted (EDEK)

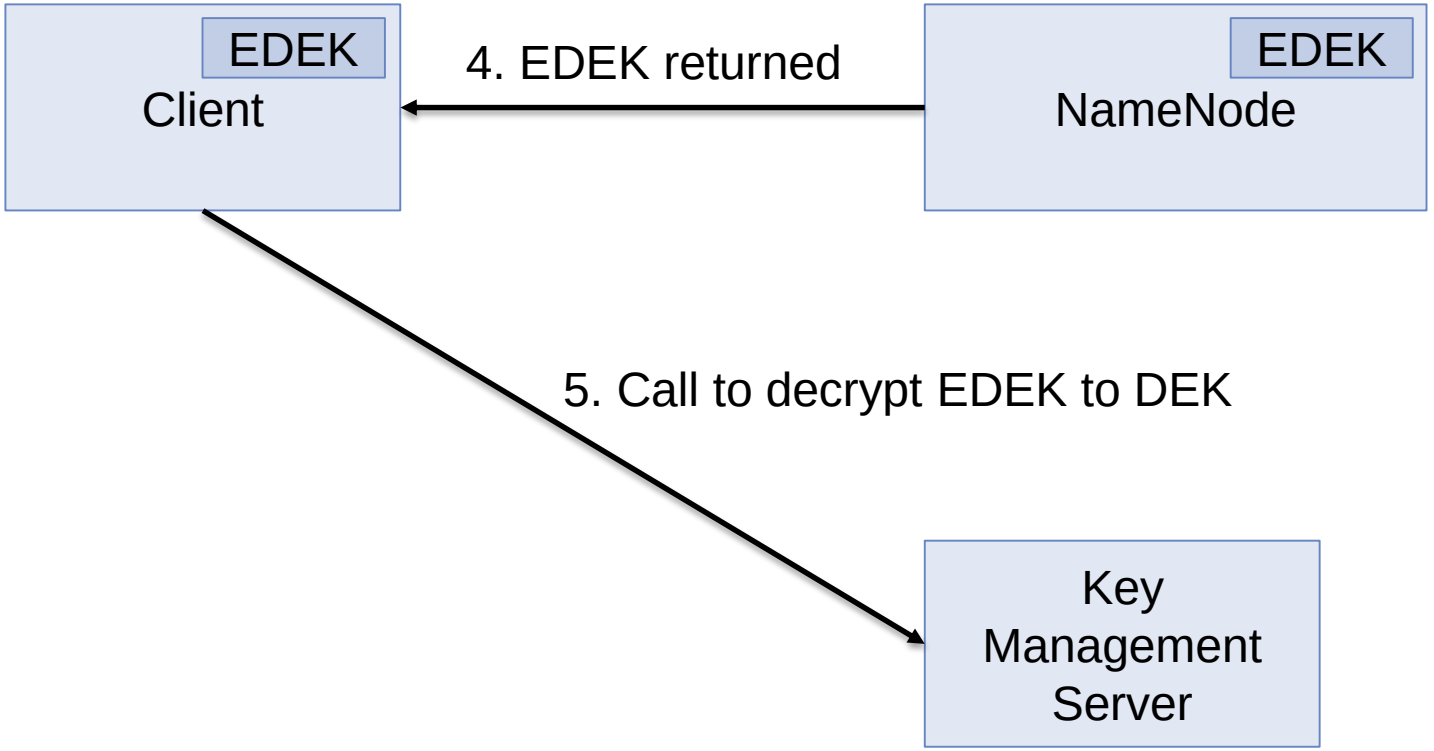


Transparent Encryption: How to encrypt data

NTT DATA

Hadoop 2.6

- DEK (Data Encryption Key)
  - A unique key for each file in EZ (Encryption Zone)
  - Stored in an Xattr of the file, encrypted (EDEK)

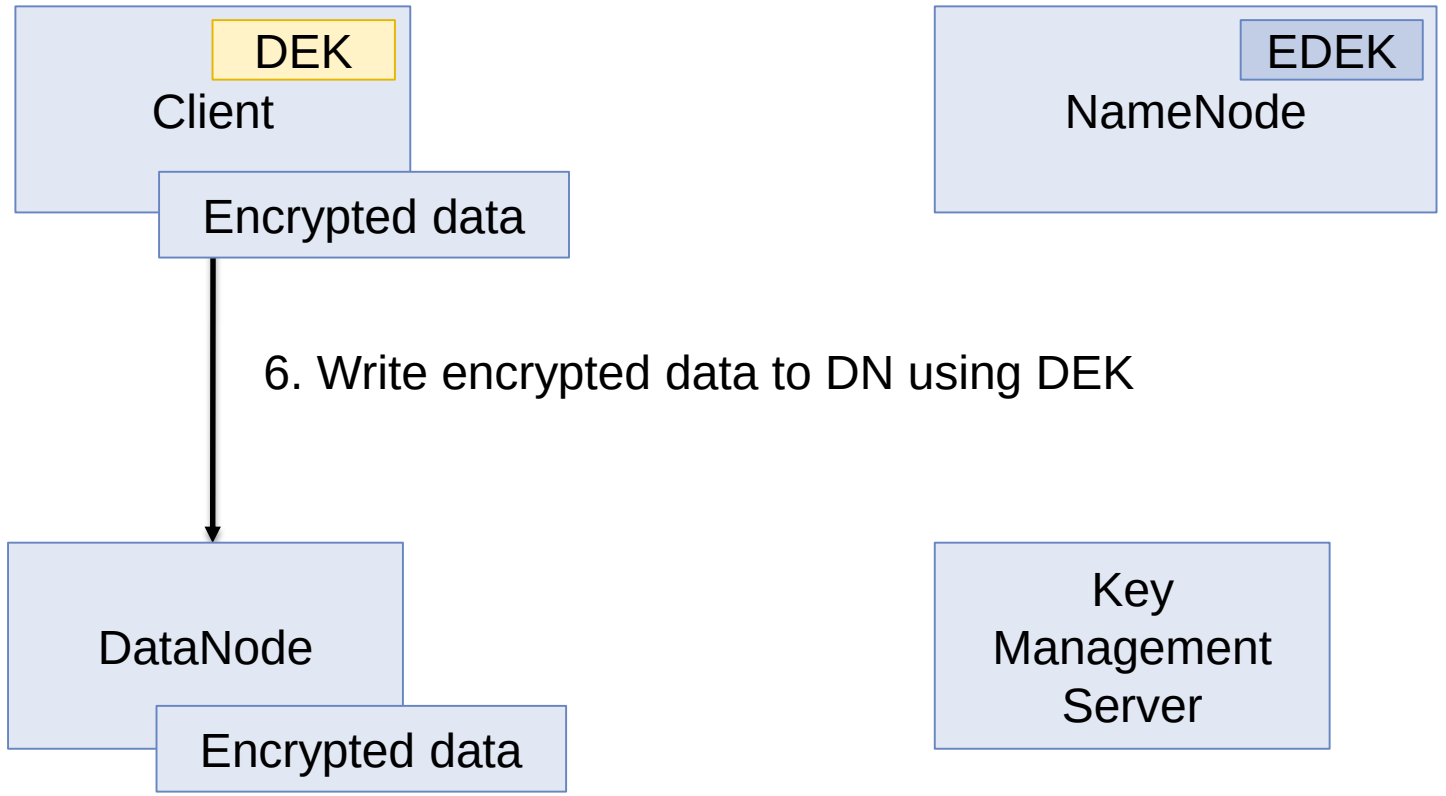


Transparent Encryption: How to encrypt data

NTT DATA

Hadoop 2.6

- DEK (Data Encryption Key)
  - A unique key for each file in EZ (Encryption Zone)
  - Stored in an Xattr of the file, encrypted (EDEK)



Transparent Encryption: Very low overhead

NTT DATA

Hadoop 2.6

- Very low overhead
  - Simple benchmark with 3 slaves (m3.xlarge, 4 core Xeon E5-2670 v2)
  - Use AES-NI

	Encryption Off	Encryption On
1GB Teragen	17 sec	18 sec
1GB Terasort	47 sec	49 sec

- Known issue
  - Encryption is sometimes done incorrectly (HADOOP-11343)
  - Recommend 2.7.1 or 2.6.1

Present

Hadoop 2.7 (2014-11-18)

NTT DATA

- Quota per storage type
- Truncate API
- Files with variable-length blocks
- Web UI for NFS gateway
- NNTop: top-like tool for NameNode
  - List top users for each operation
  - Exposed via metric
- fsck -blockId option
  - Print the file which the blockId belongs to
- Inotify

INotify for HDFS

NTT DATA

Hadoop 2.7

- Problem
  - Some components do caching
    - Hive caches path names
    - Impala caches block locations
  - When to invalidate cache?

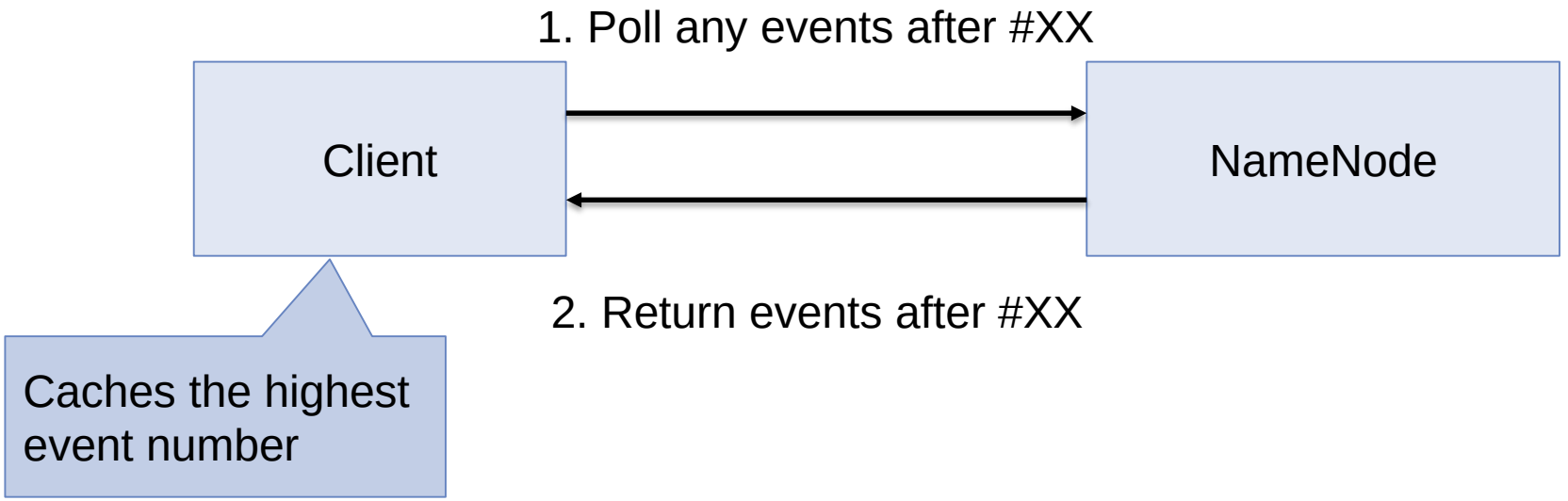
- Solution
  - Introduce a tool similar to Linux inotify
  - Client can monitor the events without parsing NN log or edits

INotify for HDFS: Technical Approach

NTT DATA

Hadoop 2.7

- Client polls NameNode periodically
  - Not push model



- Known issue
  - Truncate is not notified (HDFS-8742)
  - Fixed in 2.8.0

Future

Many features are being developed

NTT DATA

- 2.8 (not released)
  - Support OAuth2 in WebHDFS
  - RPC Congestion control
- Feature branches
  - Erasure Coding (HDFS-7285)
  - Ozone: Object store (HDFS-7240)
  - BlockManager Scalability Improvements (HDFS-7836)
  - HTTP/2 support for DataTransferProtocol (HDFS-7966)
  - Implement an async pure c++ HDFS client (HDFS-8707)

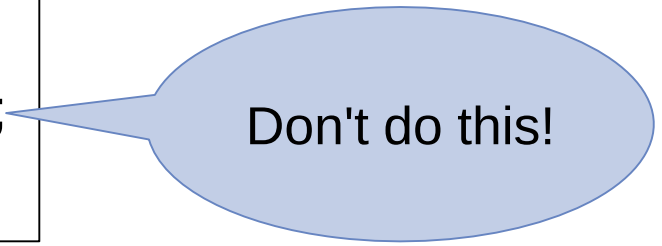
RPC Congestion Control

NTT DATA

Hadoop 2.8

- Problem
  - NameNode RPC queue is FIFO
  - DDoS can kill entire cluster

```
while (true) {
  dfs.exists("/data");
}
```



- Solution
  - Fair scheduling for RPC queue (2.6.0)
  - Retriable exception with exponential backoff (2.8.0)
- Enable by default in 2.8

Erasure Coding

NTT DATA

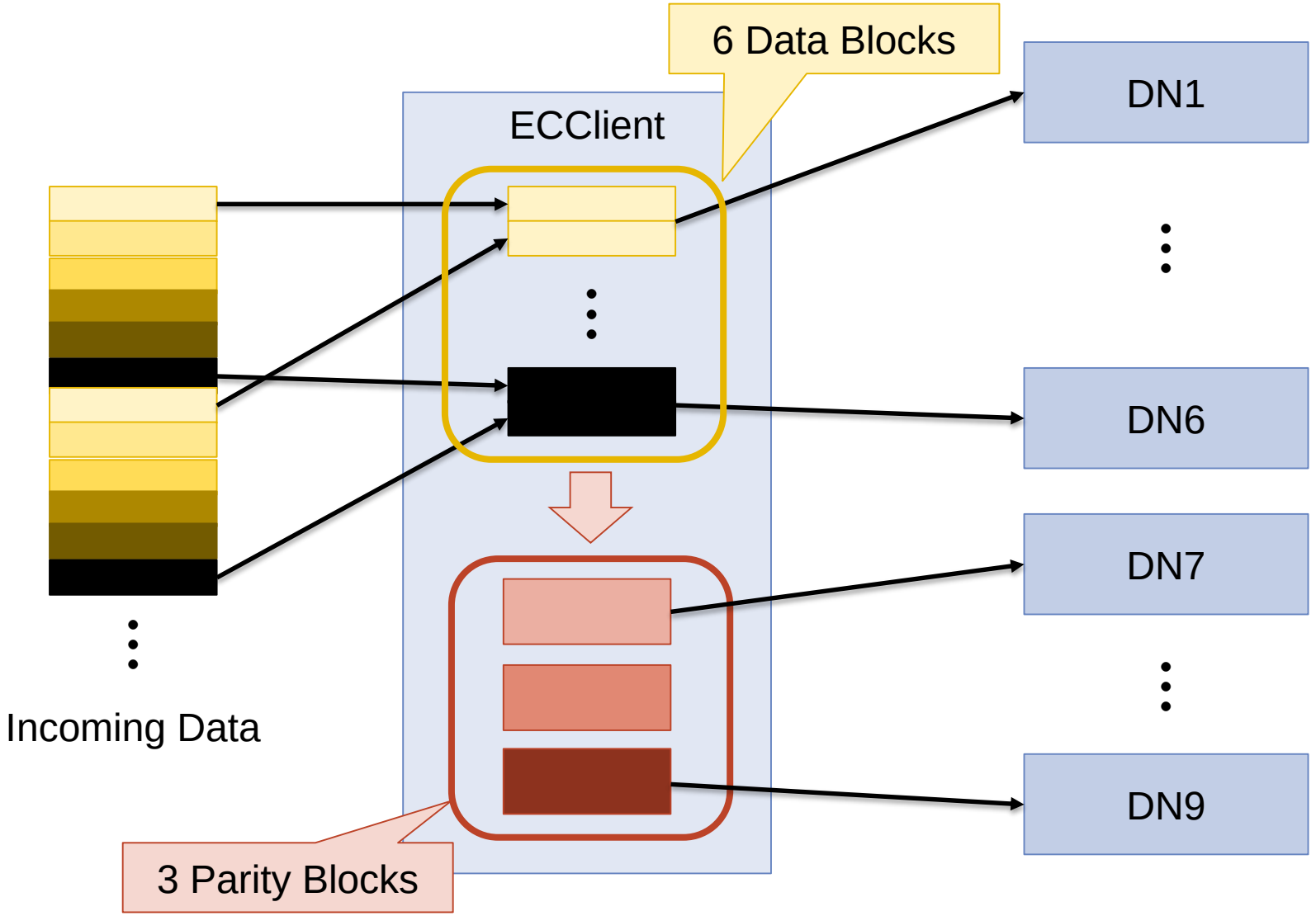
- Problem
  - Reduce costs of storage
  - Blocks are replicated to 3 DNs
    - 3x storage overhead is costly
- Solution
  - Use Erasure Code

	3-replication	(6,3)-Reed-Solomon
Tolerates	2 failures	3 failures
Disk Usage	3x	1.5x

Erasure Coding: Write files using (6,3)-Reed-Solomon

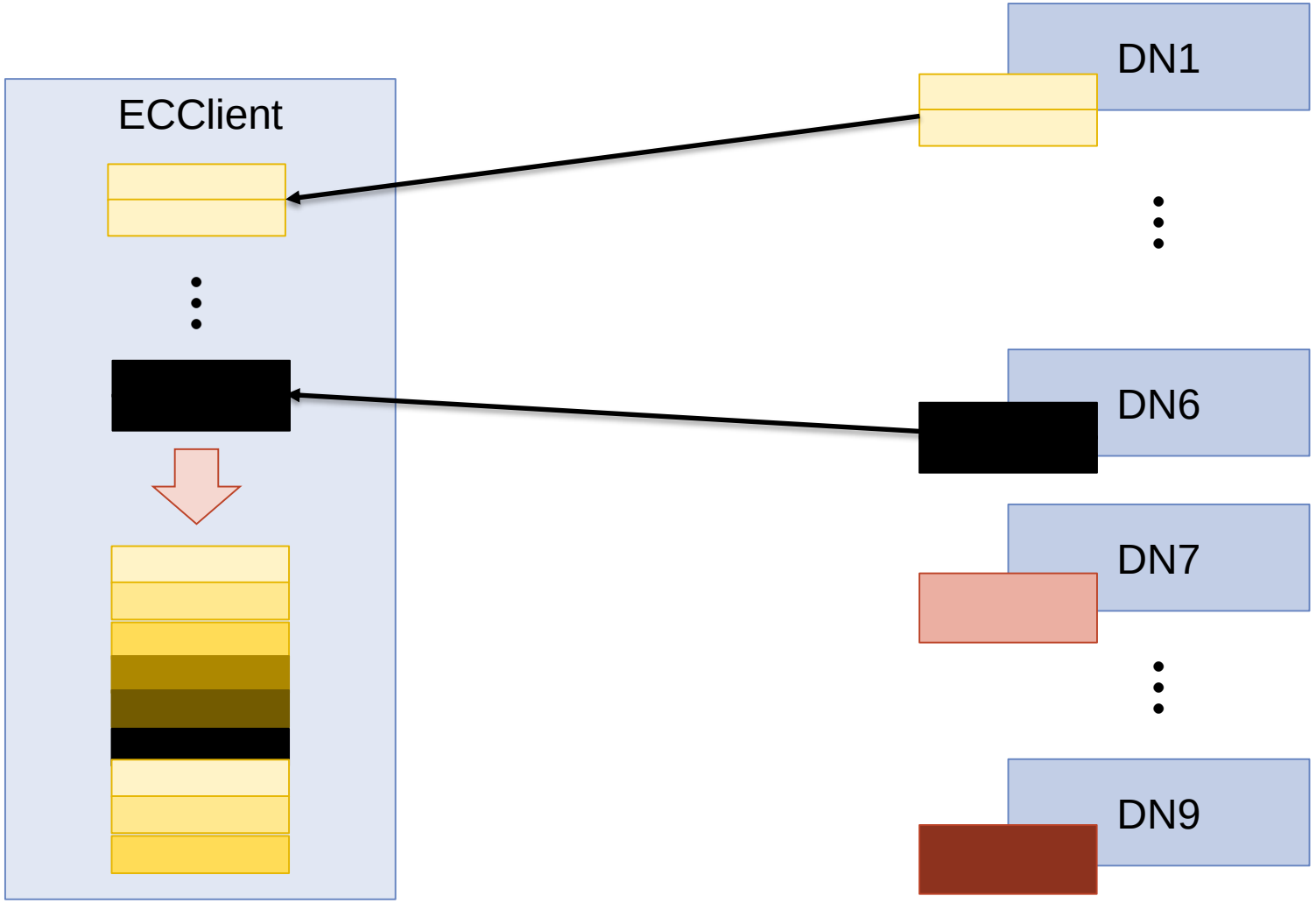
NTT DATA

- Write data to 9 DNs in parallel



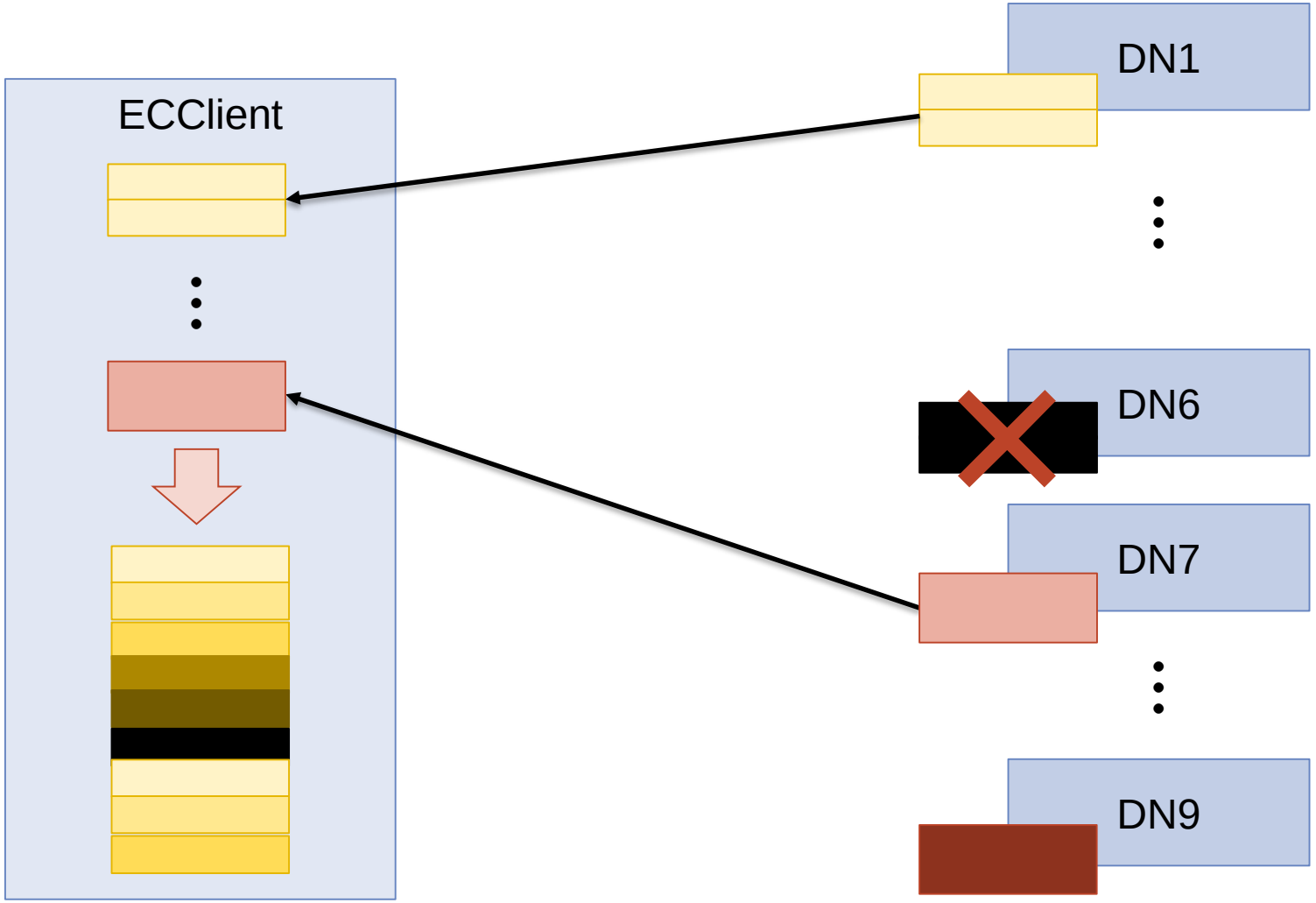
Erasure Coding: Read files

- Read data from 6 DNs in parallel



Erasure Coding: Read files when DN fails

- Read data from (arbitrary) 6 DNs in parallel



Erasure Coding: Current status

NTT DATA

- Suitable for cold data
  - No data locality
  - Very low cost/GB with archival storage
- Now preparing for merge
- Follow on work
  - Intel ISA-L support for faster encoding
  - Support append/truncate/hflush/hsync
  - More encoding schemas
  - Pipeline error handling
  - Support contiguous layout (HDFS EC Phase 2)

Summary

NTT DATA

- Many features are still in development
- I cannot predict when the feature will be available
- Recommend anyone who wants a feature to join contributing to it to make the development faster
- There are many ways to contribute
  - Creating/Testing/Reviewing patches
  - Reporting bugs
  - Writing documents
  - Discussing architecture design
  - https://wiki.apache.org/hadoop/HowToContribute

NTT DATA

Global IT Innovator

References

NTT DATA

- Apache Hadoop Docs: http://hadoop.apache.org/docs/current/
- In-memory caching (HDFS-4949)
  - In-memory Caching in HDFS: Lower Latency, Same Grate Taste: http://www.slideshare.net/Hadoop_Summit/inmemory-caching-in-hdfs-lower-latency-same-great-taste-33921794
- Heterogeneous Storages (HDFS-5682)
  - Reduce Storage Costs by 5x Using The New HDFS Tiered Storage Feature: http://www.slideshare.net/Hadoop_Summit/reduce-storage-costs-by-5x-using-the-new-hdfs-tiered-storage-feature
- Transparent Encryption (HDFS-6134)
  - Transparent Encryption in HDFS: http://www.slideshare.net/Hadoop_Summit/transparent-encryption-in-hdfs
- INotify (HDFS-6634)
  - Keep Me in the Loop: Introducing HDFS Inotify: http://www.slideshare.net/Hadoop_Summit/keep-me-in-the-loop-inotify-in-hdfs

References

- RPC congestion control (HADOOP-9640, HADOOP-10597, HDFS-8820)
  - Improving HDFS Availability with Hadoop RPC Quality of Service: http://www.slideshare.net/MingMa4/hadoop-rpcqoshadoopsummit2015
- Erasure Coding (HDFS-7285)
  - HDFS Erasure Code Storage - Same Reliability at Better Storage Efficiency: http://www.slideshare.net/Hadoop_Summit/hdfs-erasure-code-storage-same-reliability-at-better-storage-efficiency