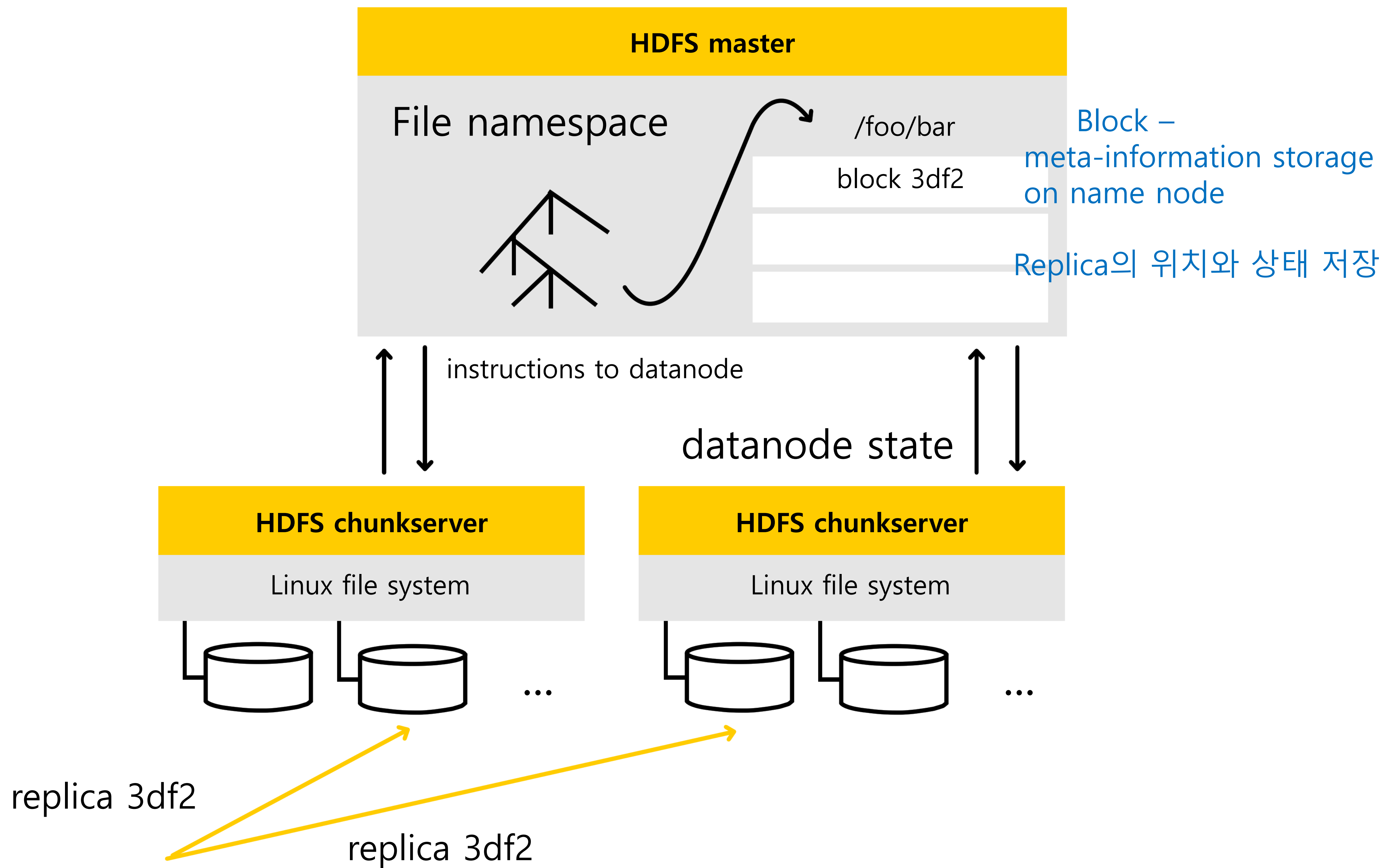


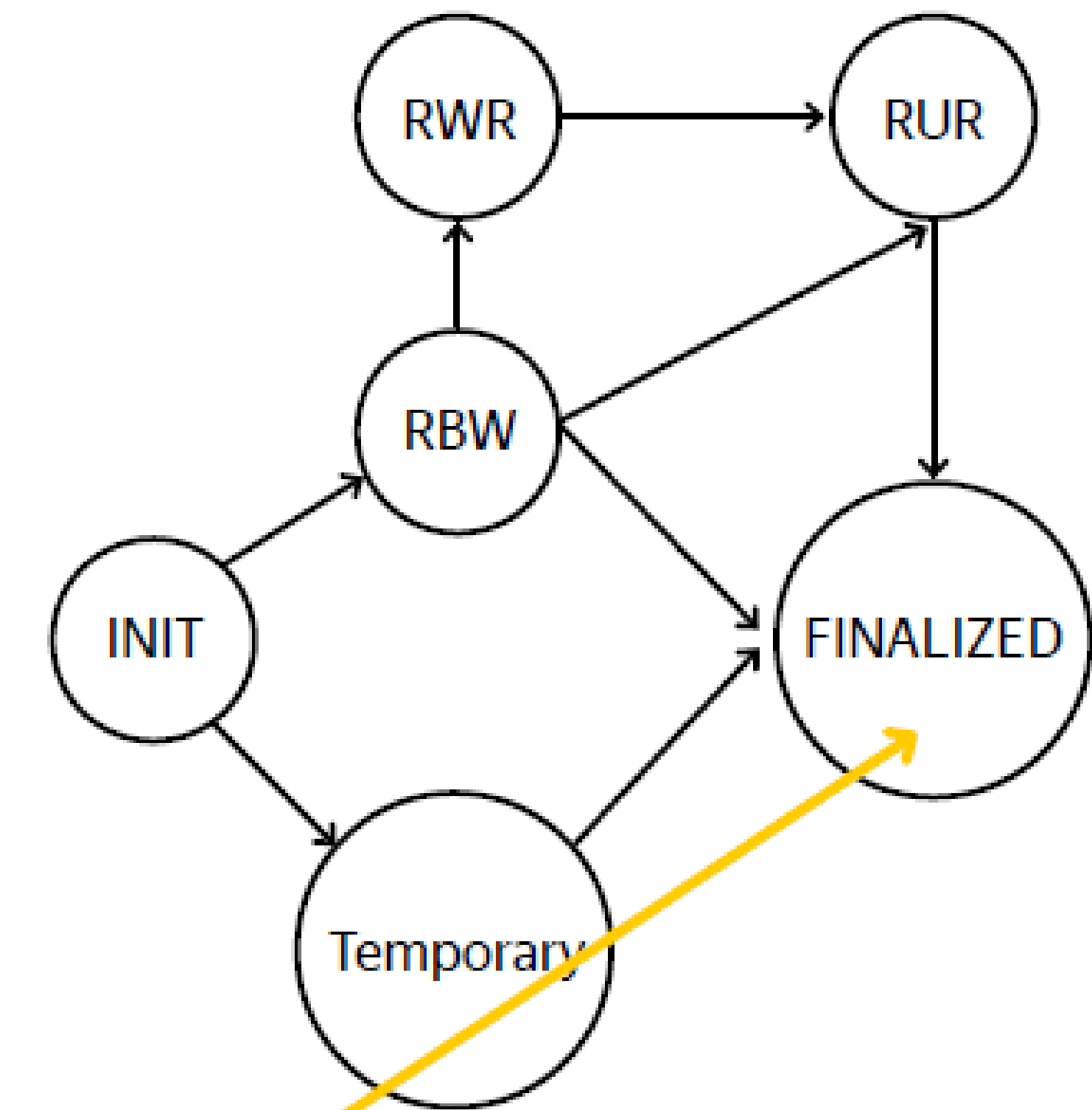
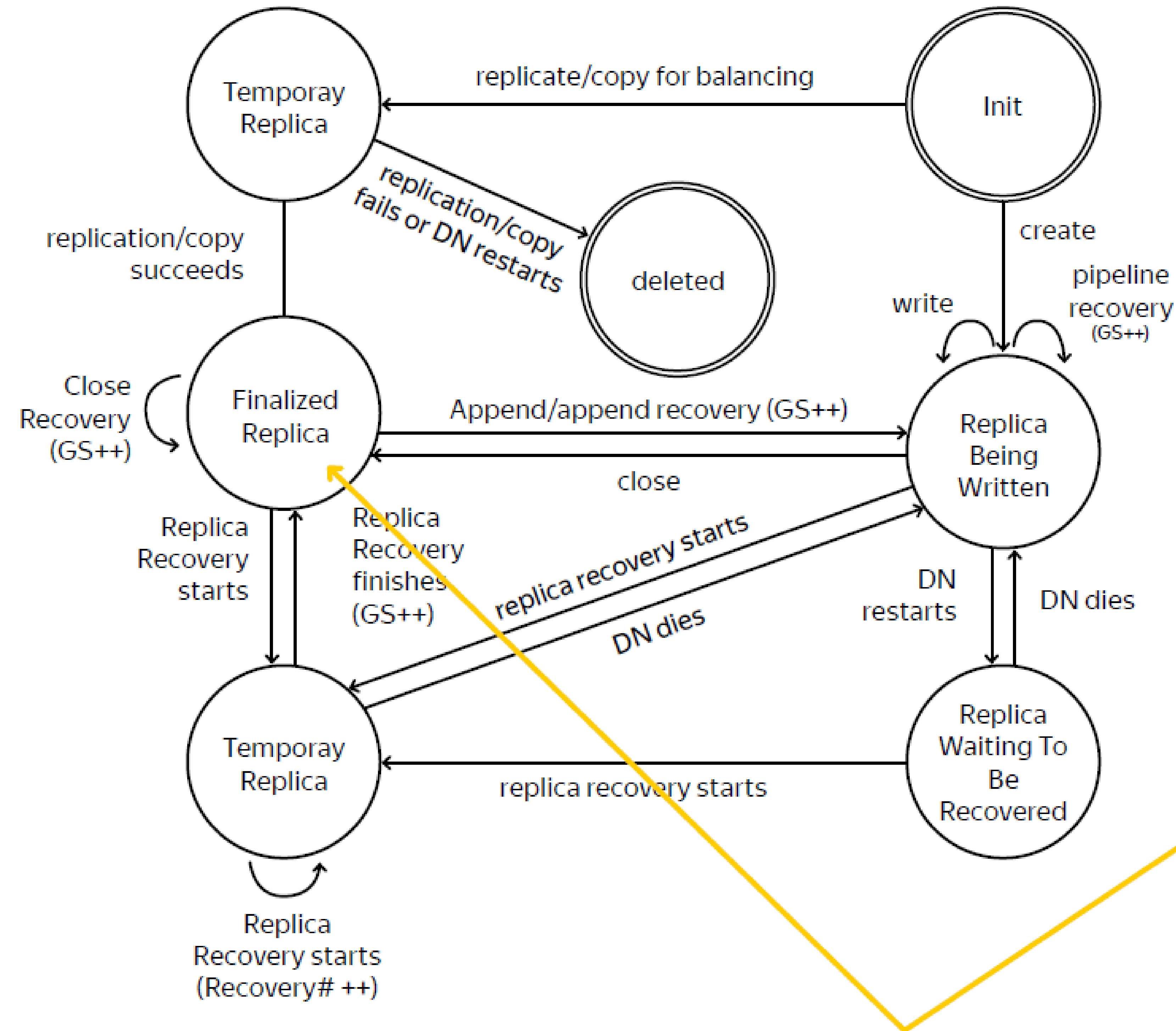
Yandex

HDFS

Block and Replica States, Recovery Process



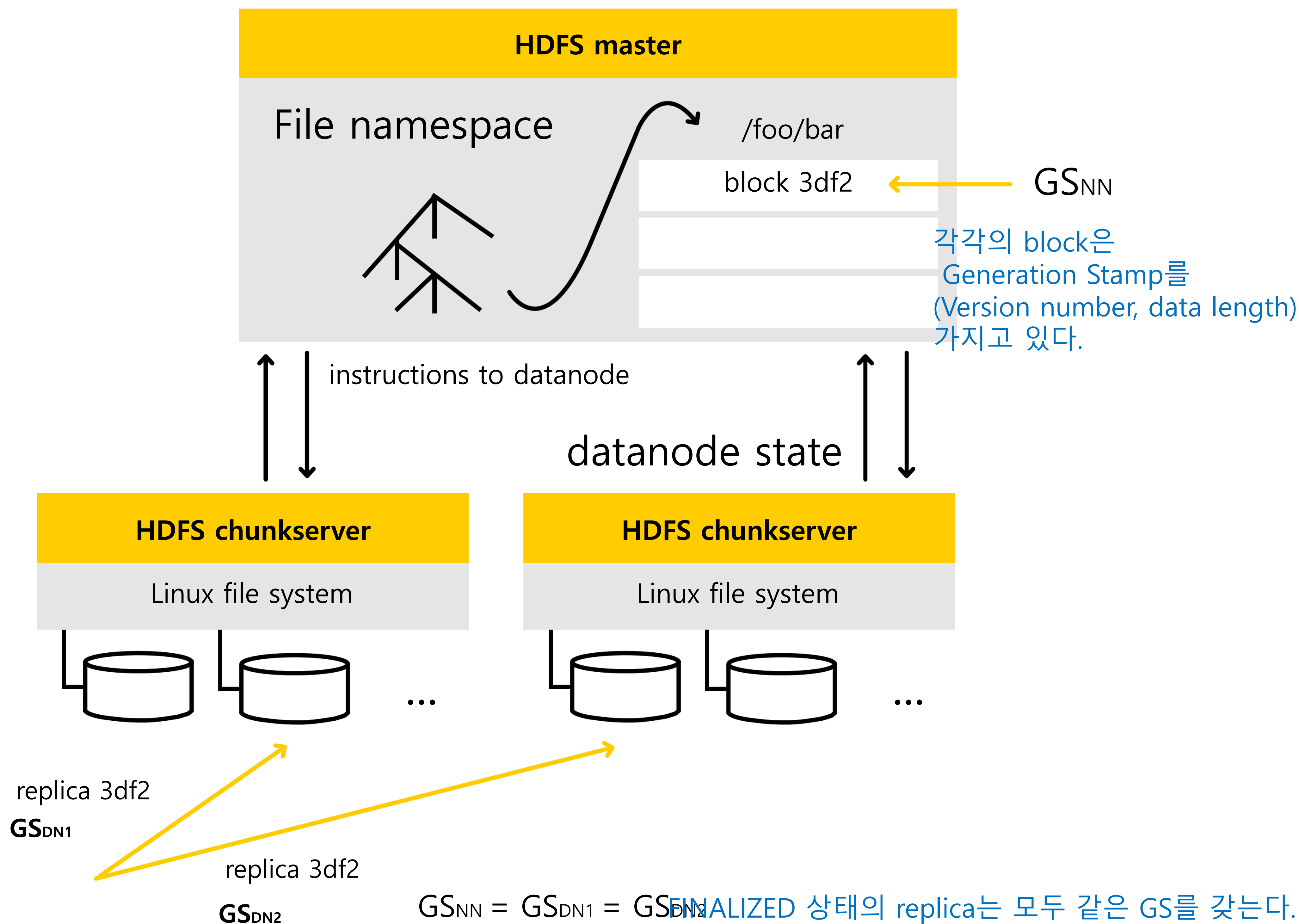
Replica state

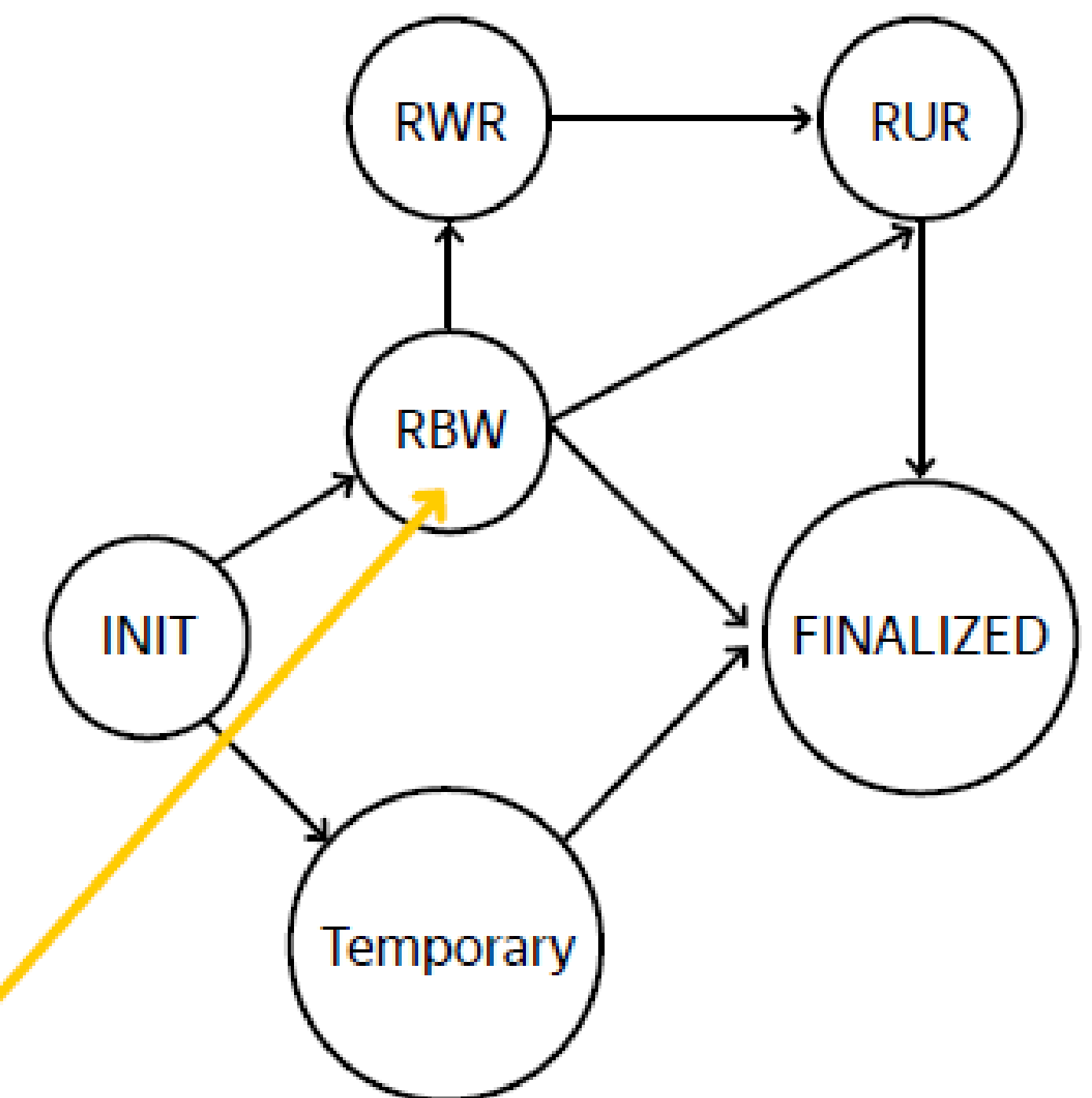
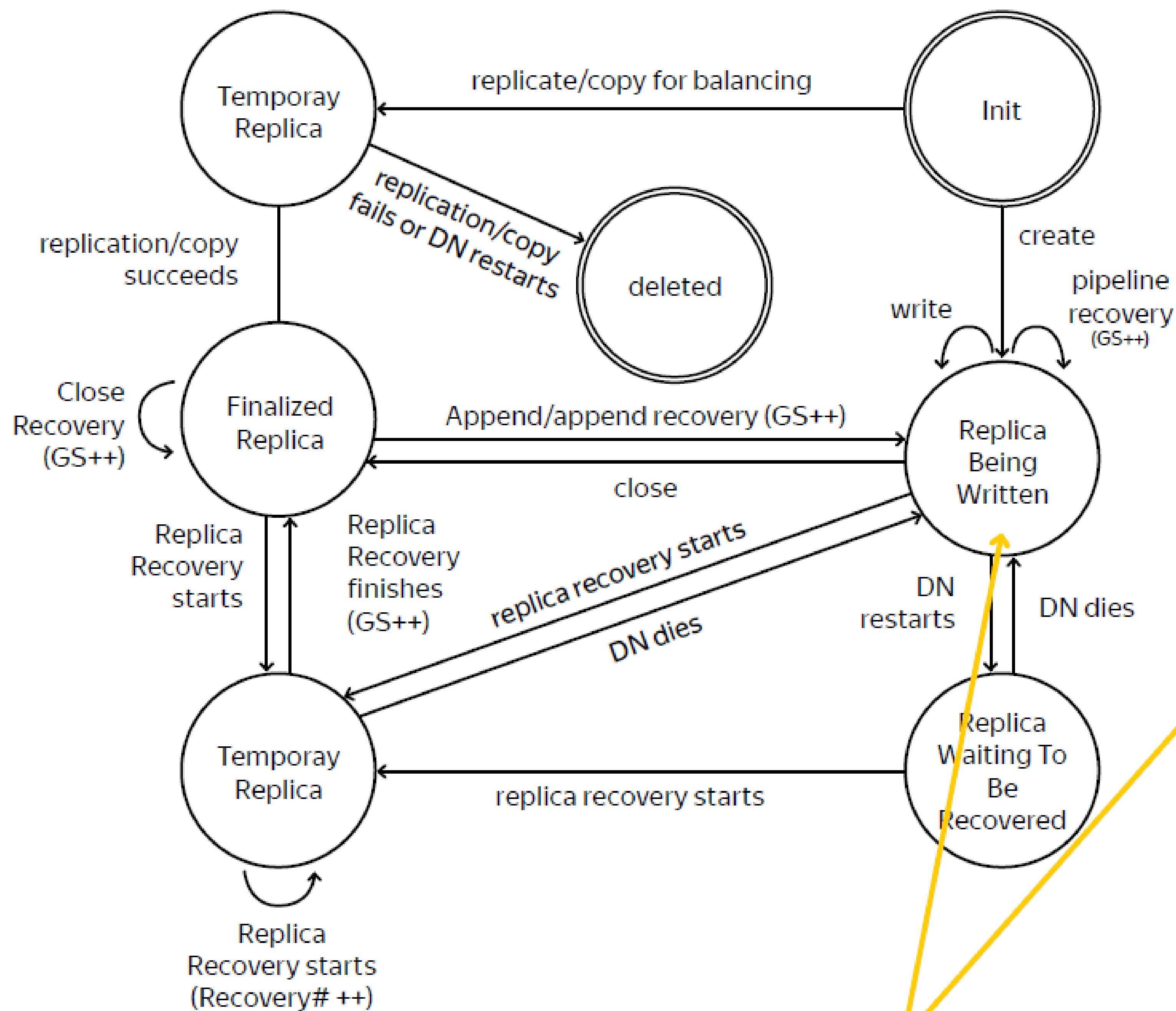


Simplified
Replica State
Transition

Finalized 완전한 상태 (read consistency를 유지한다.)
= frozen

Finalized



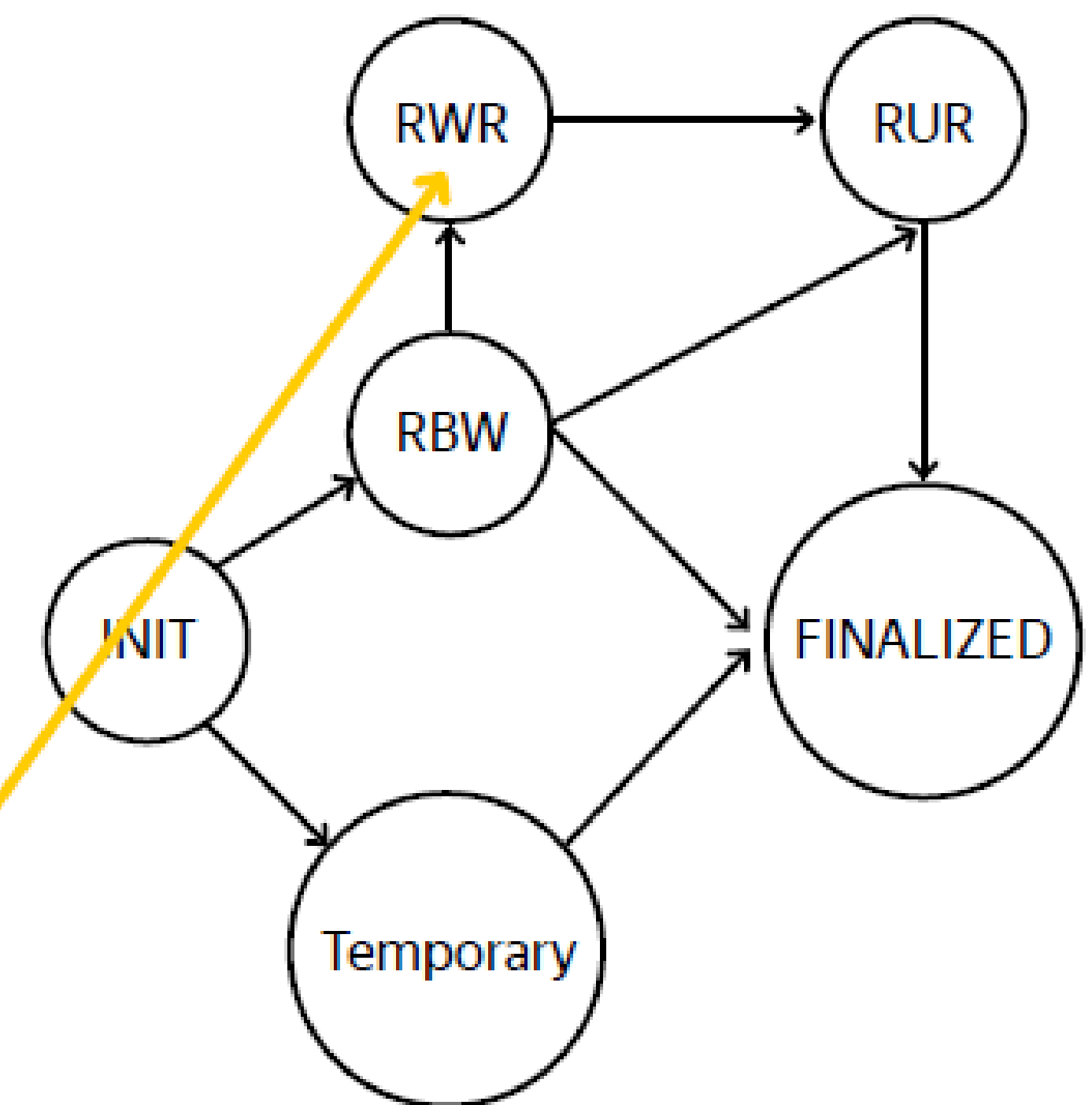
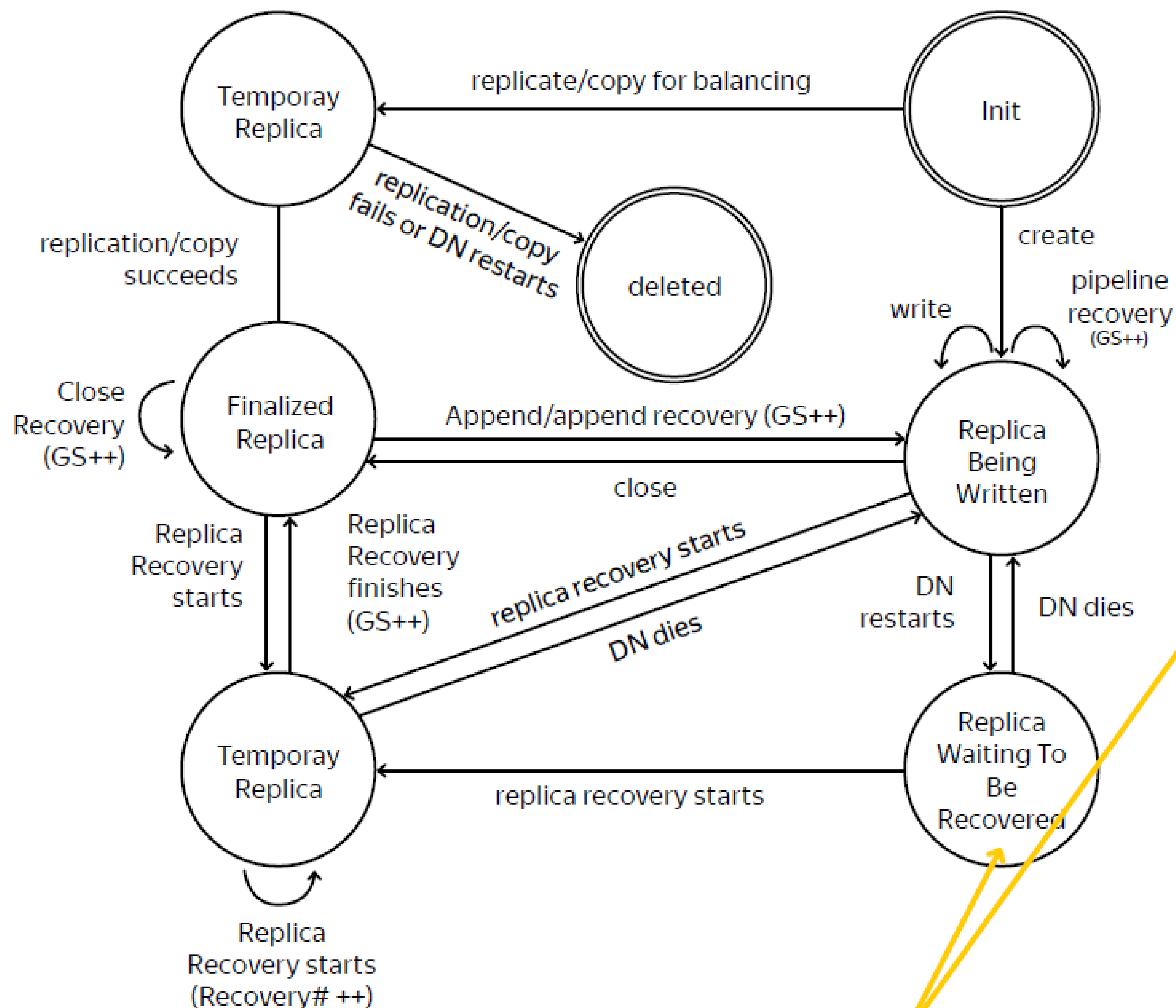


Simplified
Replica State
Transition

DATA를 append 할 때의 상태

Replica Being Written to

RBW 상태에서는 namenode의 meta 정보와 다를 수 있음

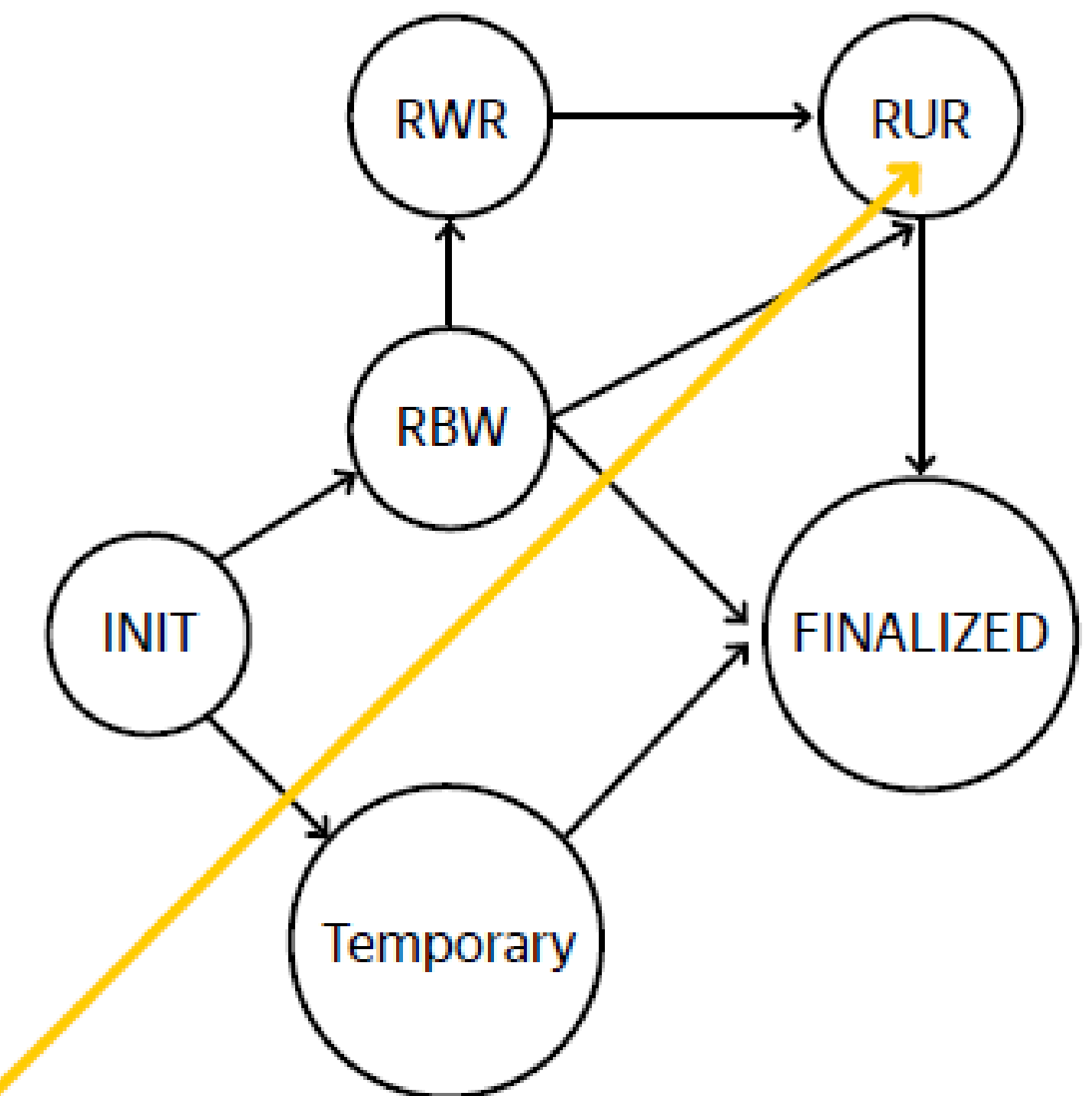
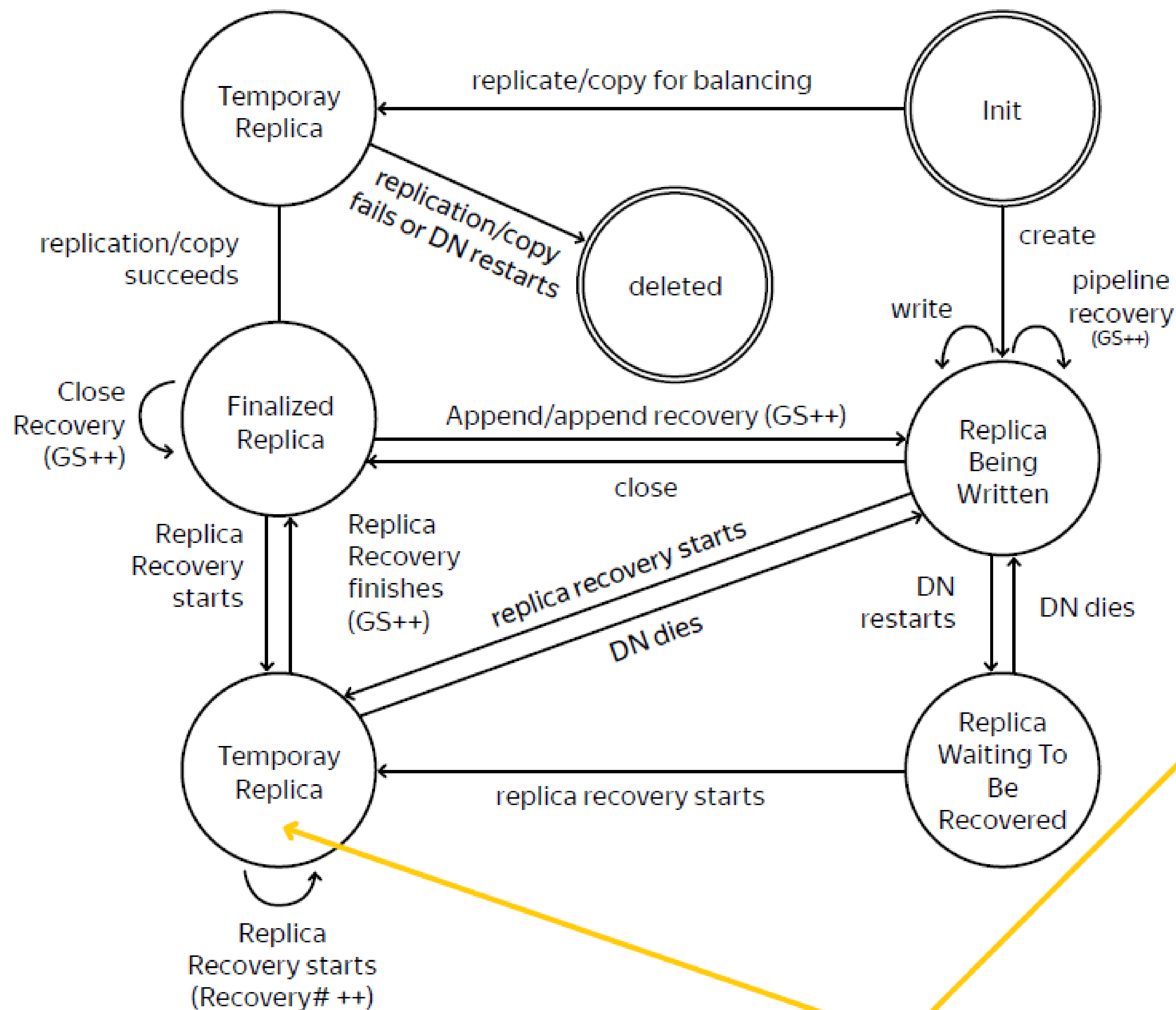


Simplified
Replica State
Transition

Datanode의 Failure 상태

Replica Waiting to be Recovered

RWR 상태는 datanode 파이프라인에 연결되지 않는다.
이 replica는 버려지거나 특별한 recovery를 실행할 수 있다.

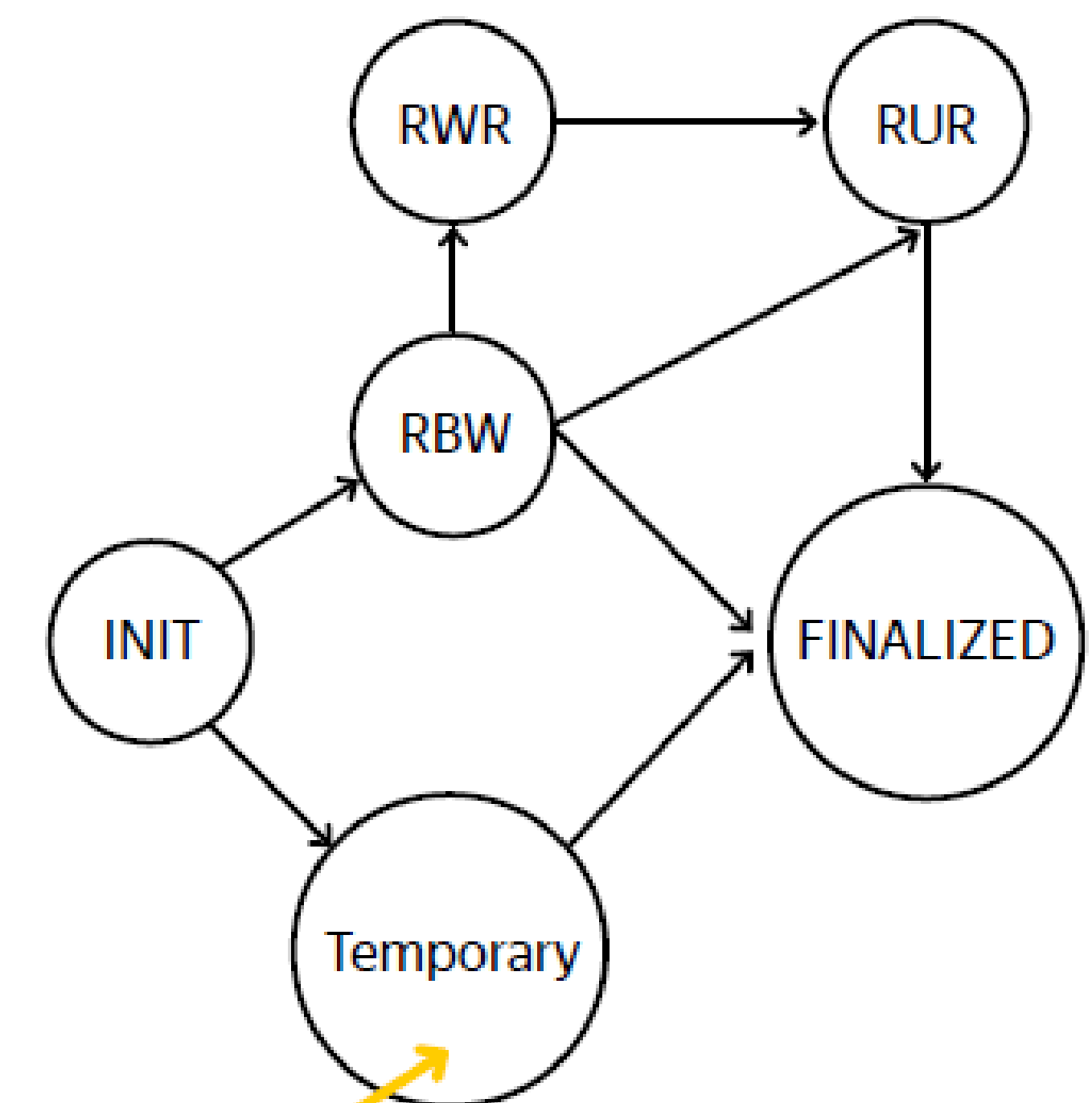
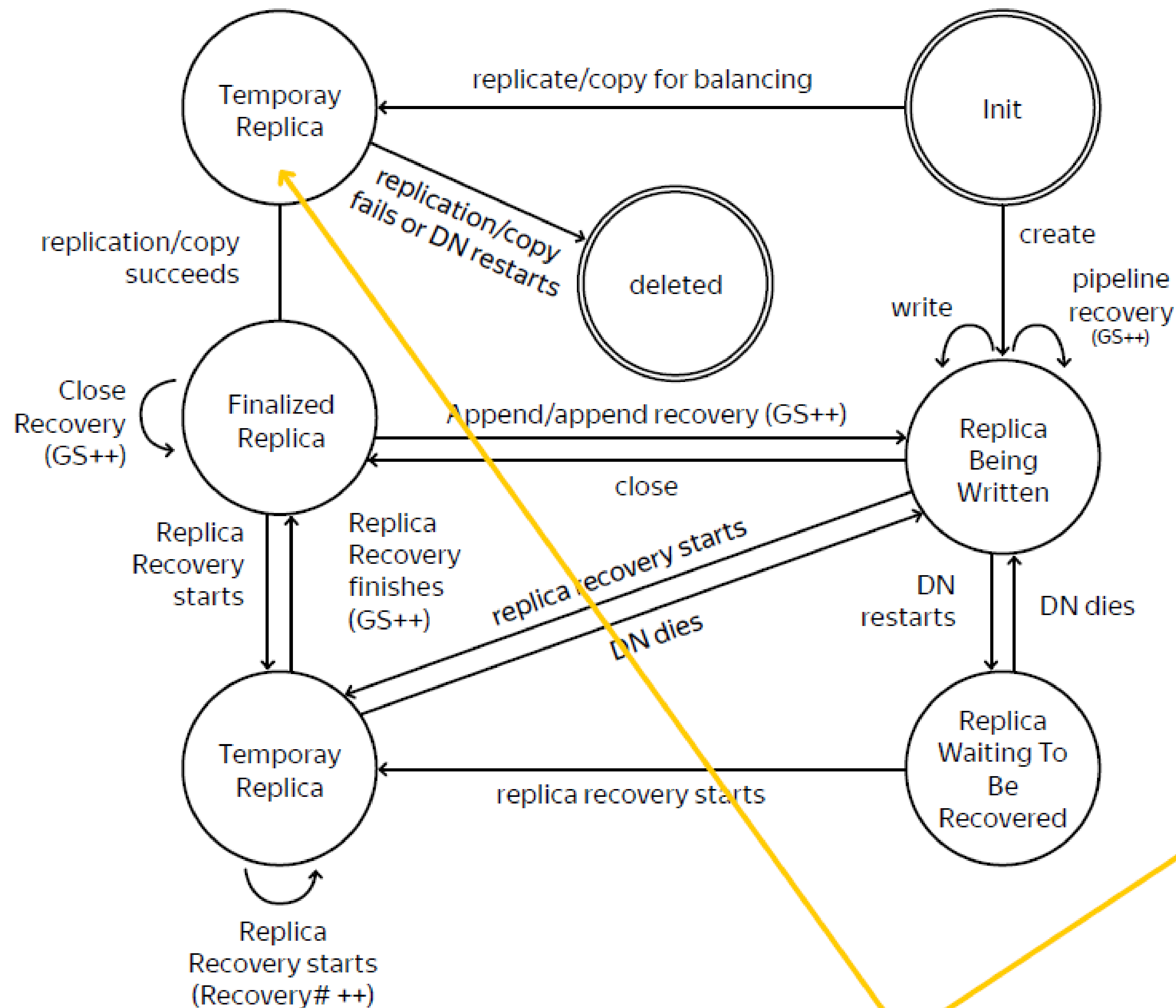


Simplified
Replica State
Transition

RBW -> RUR 로 가는 경우는 lease expiration을 의미한다.

Replica Under Recovery

Lease : Client는 파일을 write하는 독점 권한 (exclusive access) 을 요청한다. 단, read에는 독점 권한이 없음
 Lease expiration : write하기 위한 접근 중 에러 발생(주로 site failure로 인한 문제)



Simplified Replica State Transition



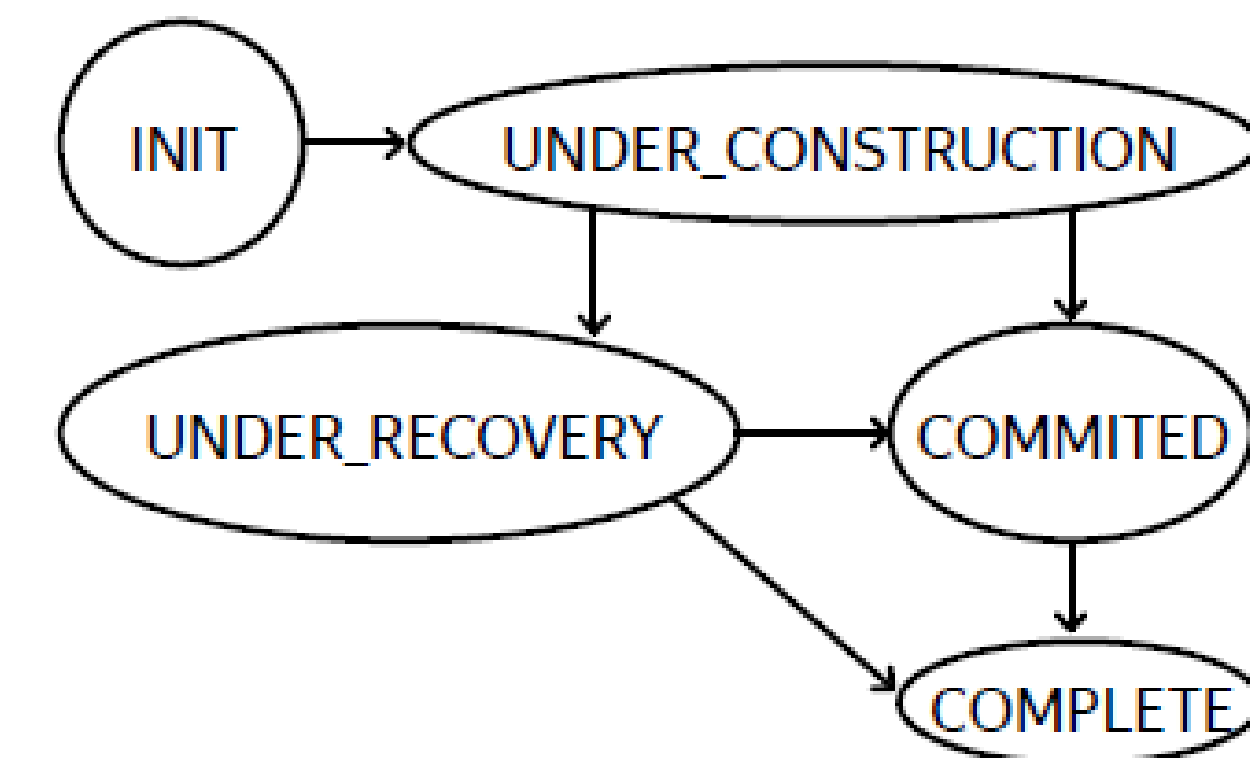
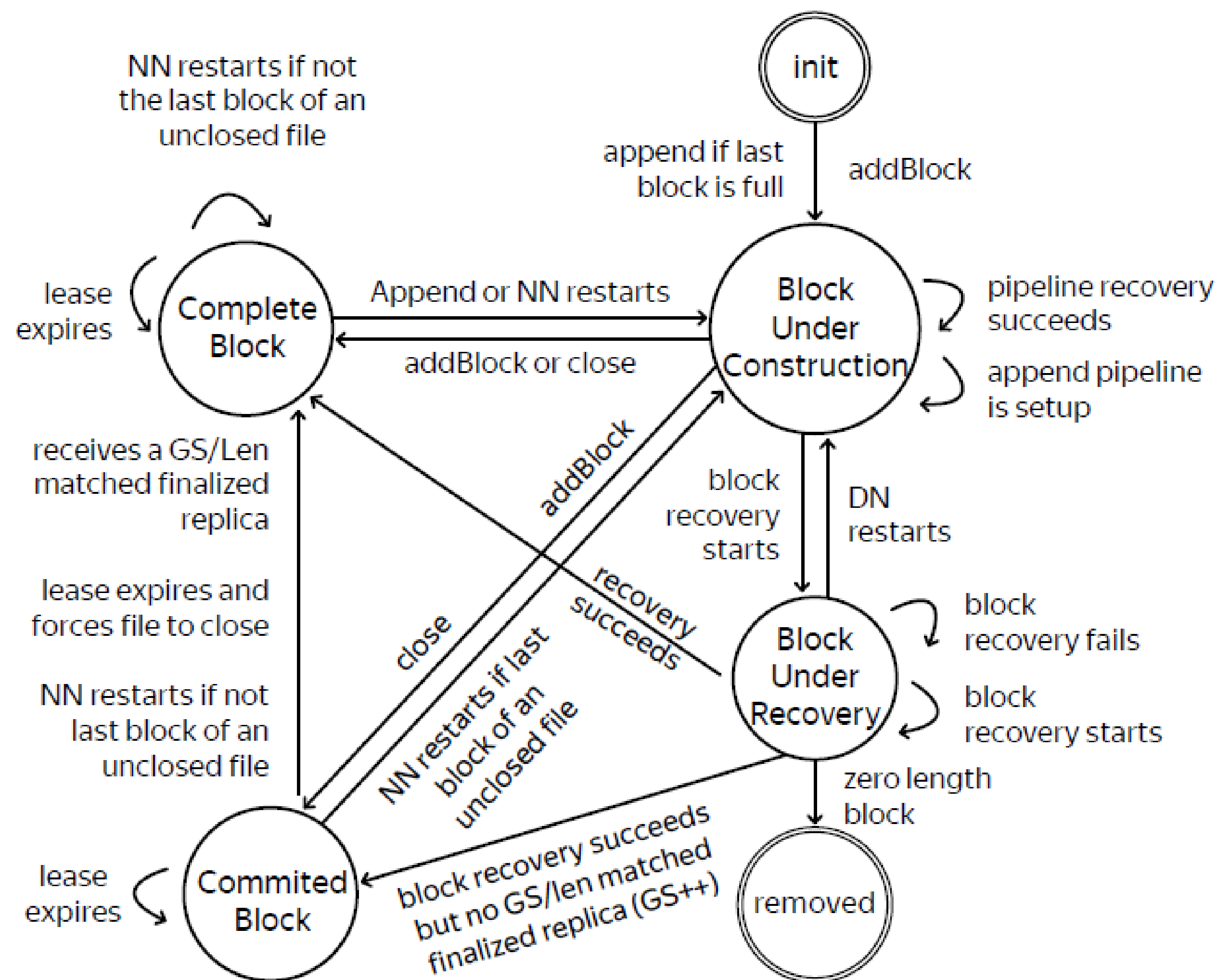
Temporary

Rebalancing 또는 replica 확장을 위해 새로 생긴 replica 상태

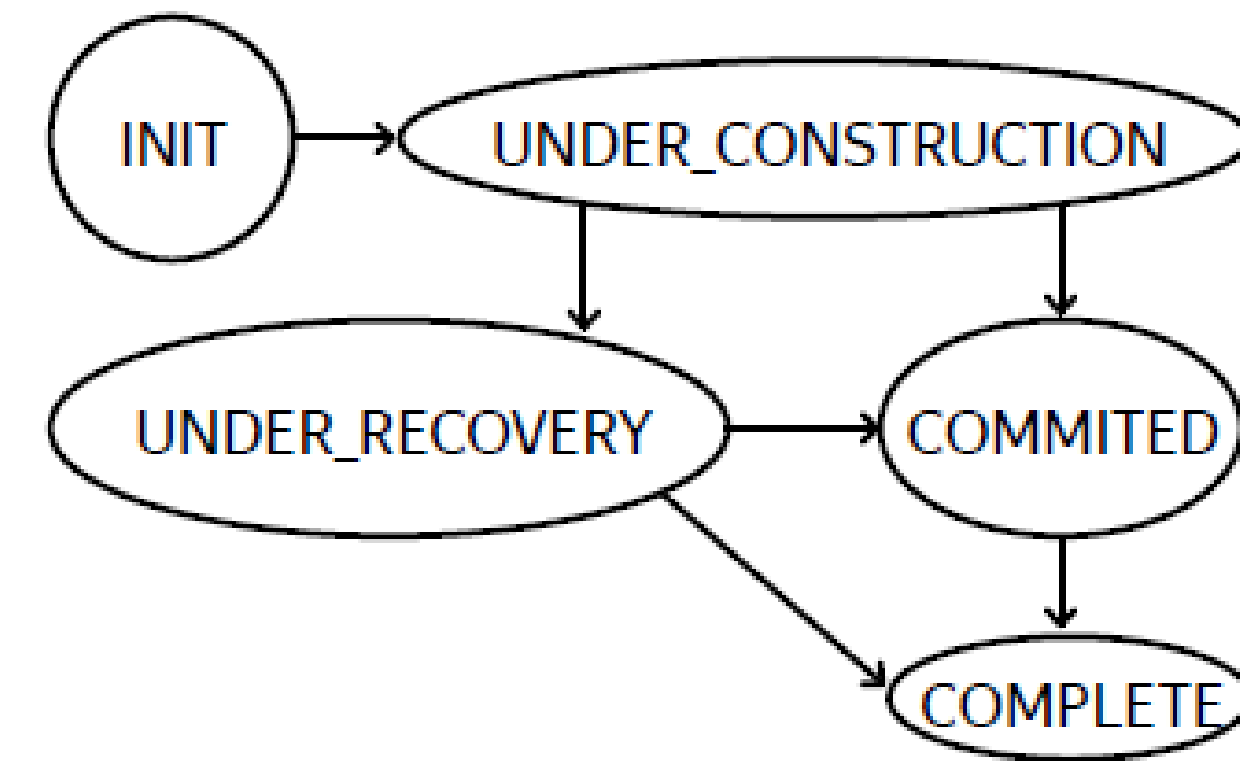
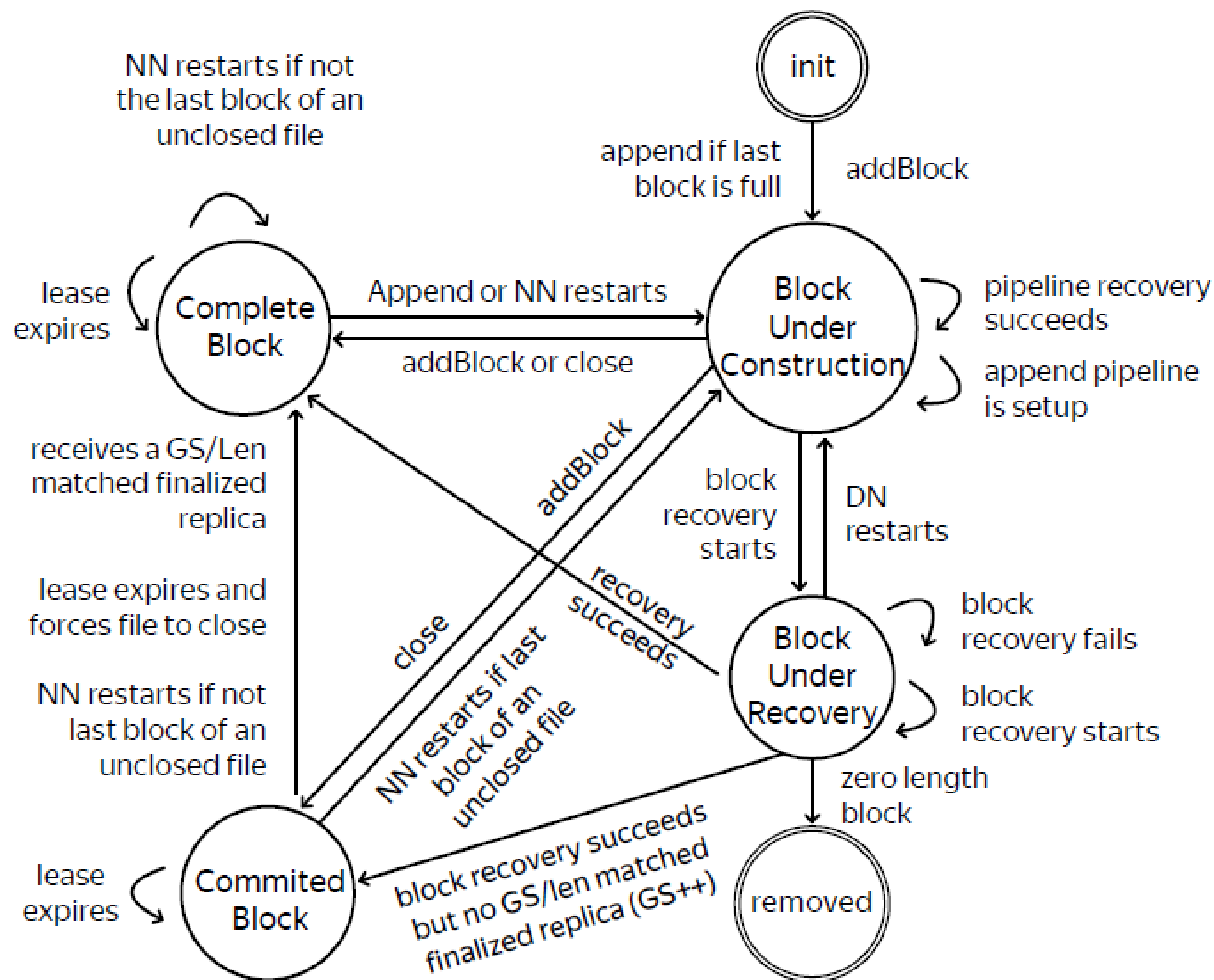
데이터가 user에게 보이지 않는 점만 빼고는 RBW 상태와 같다. (finalized가 되어야 보인다.)

Rebalancing : 데이터가 cluster에 공평하게 분산된다는 보장을 하지 않기 때문에 하는 작업

Namenode Block State



Simplified
Block State
Transition

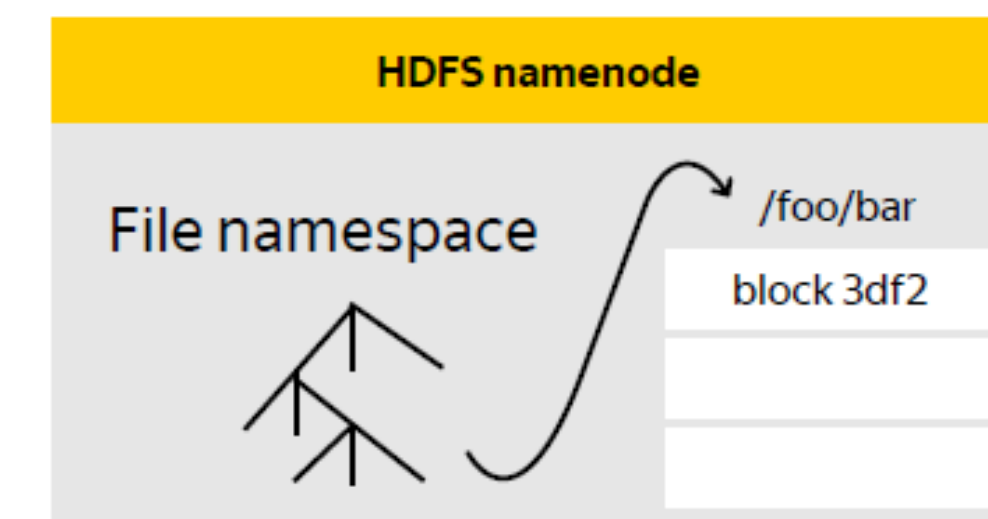


Simplified Block State Transition

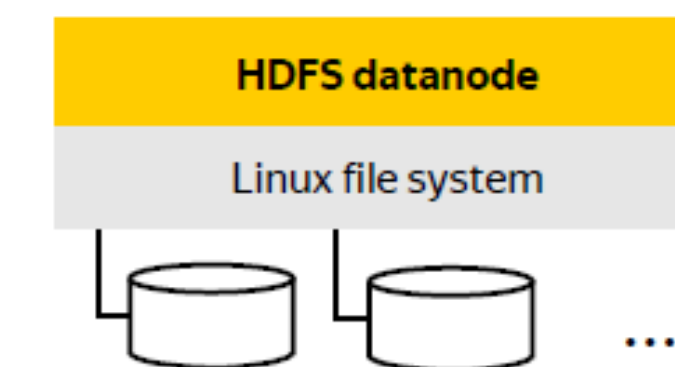
Block state는 메모리에,
Replica state는 disk에 저장된다.

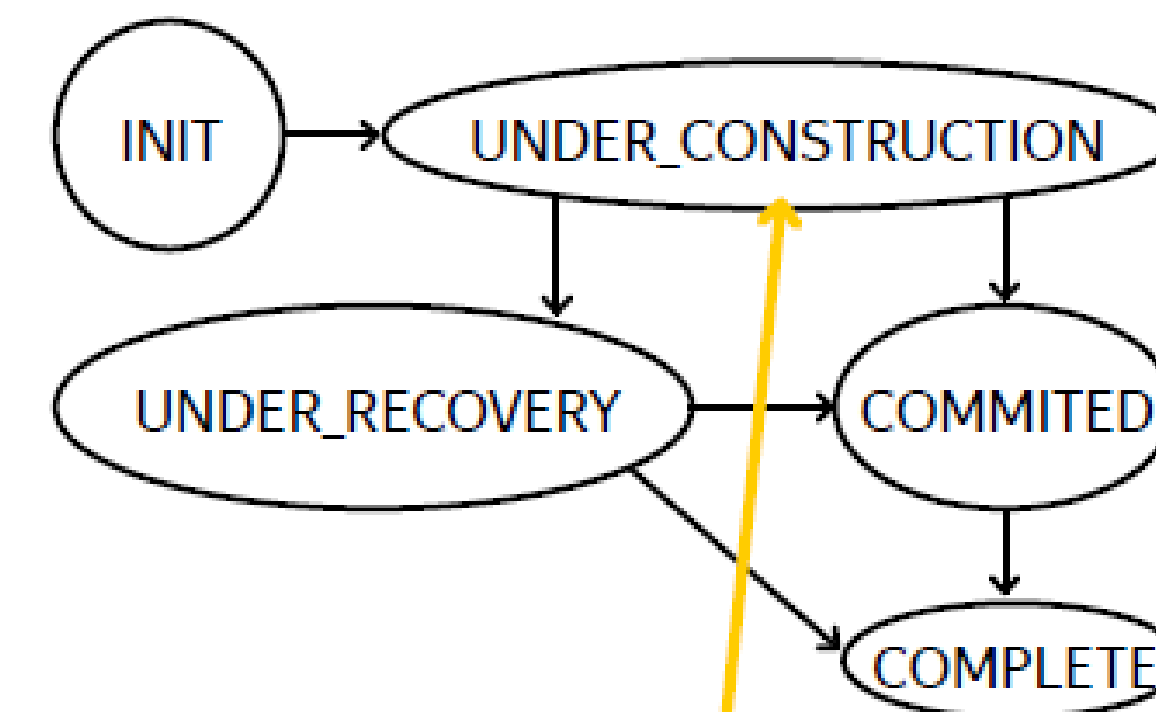
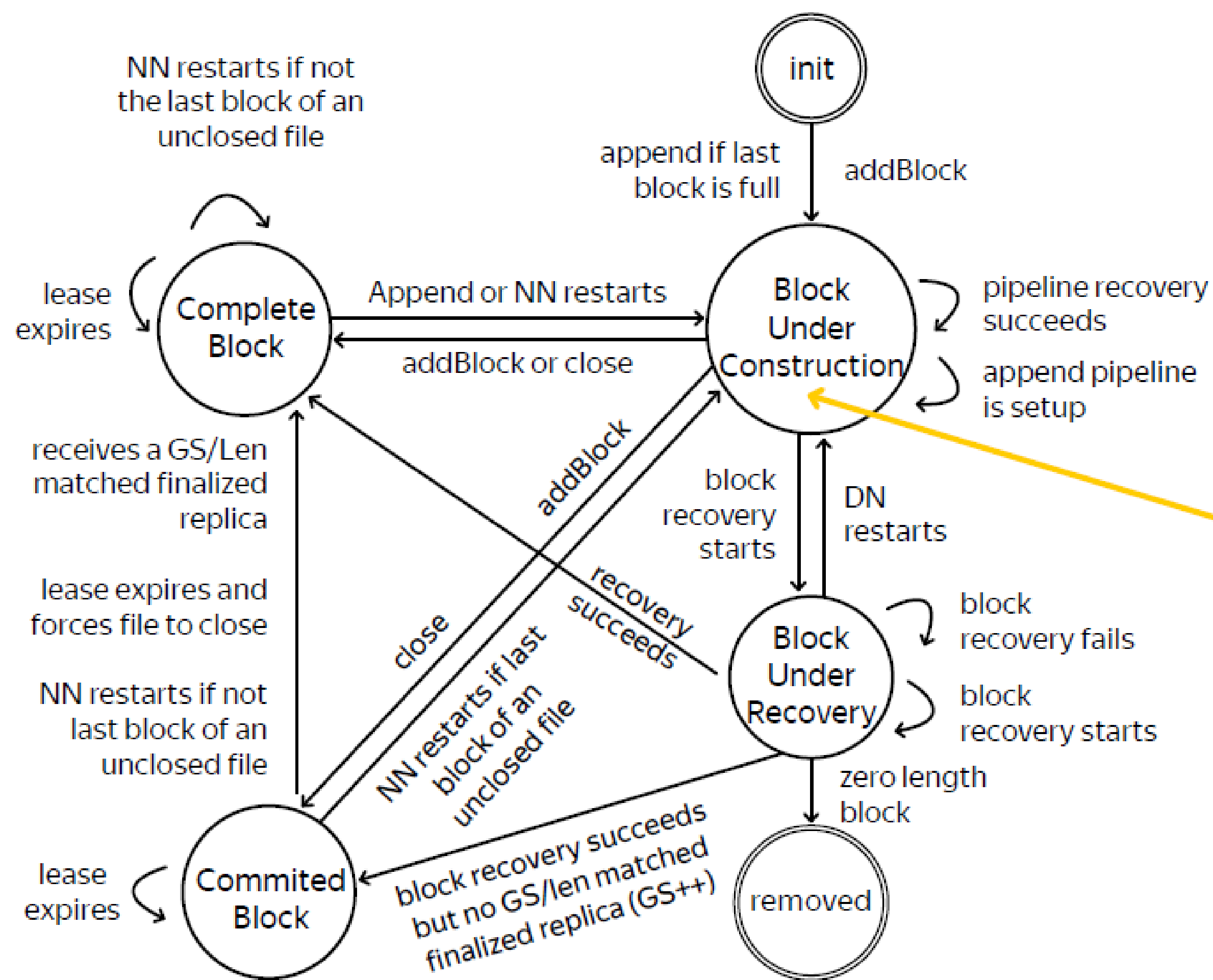


block
state



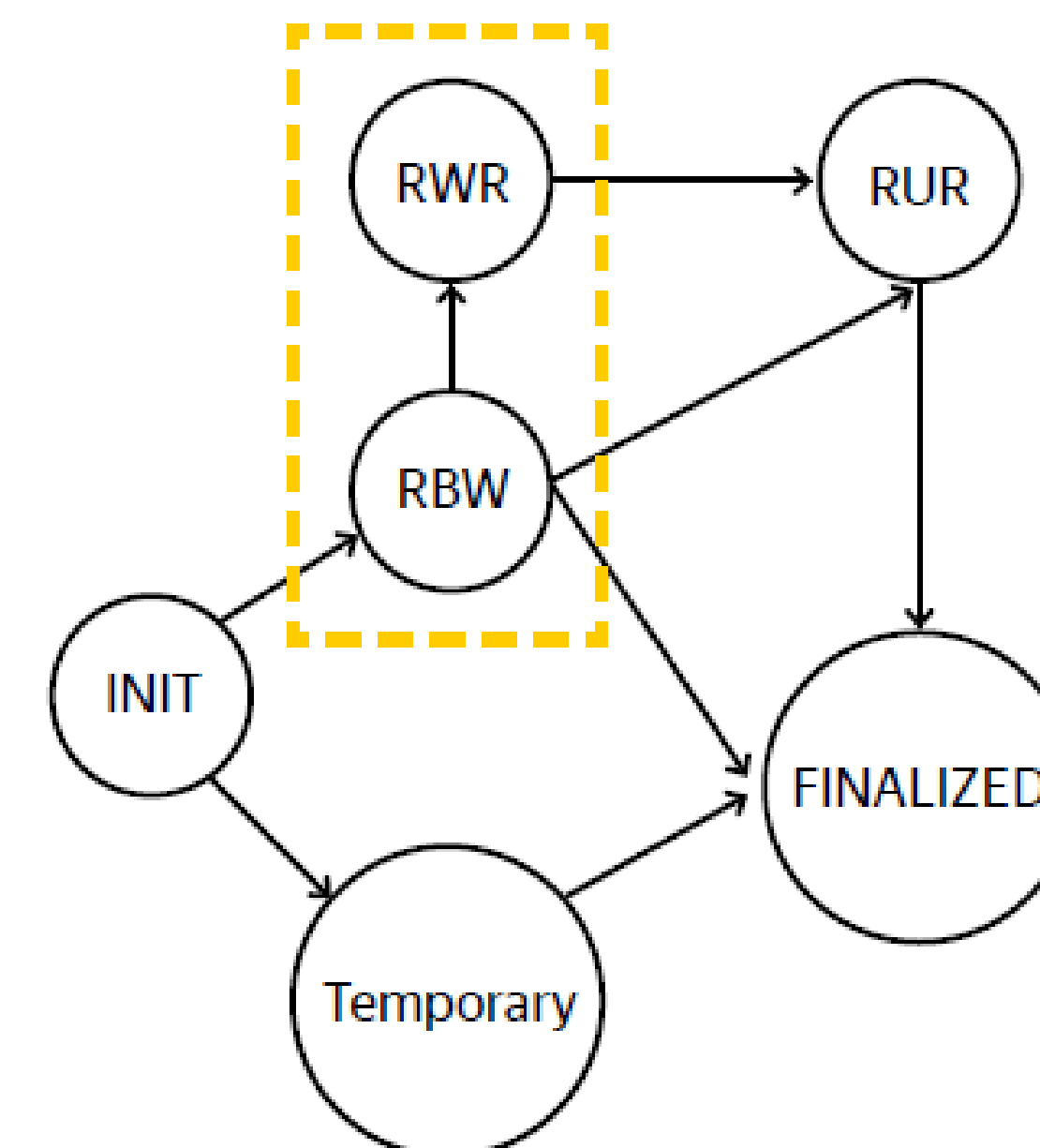
replica
state



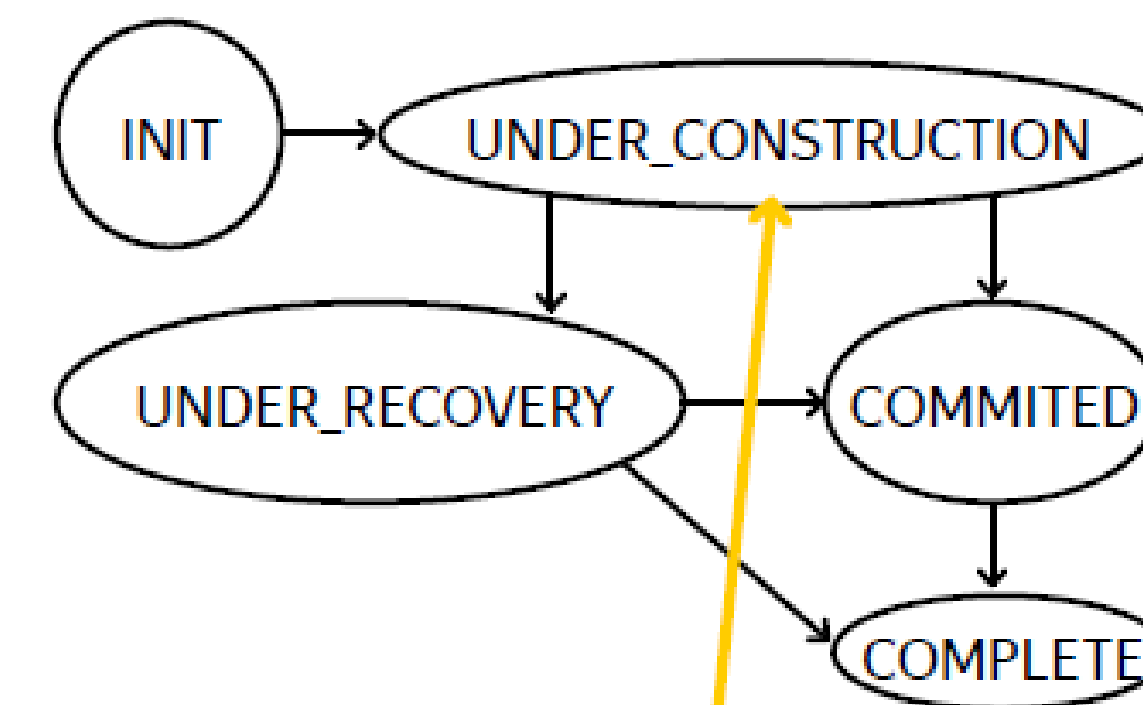
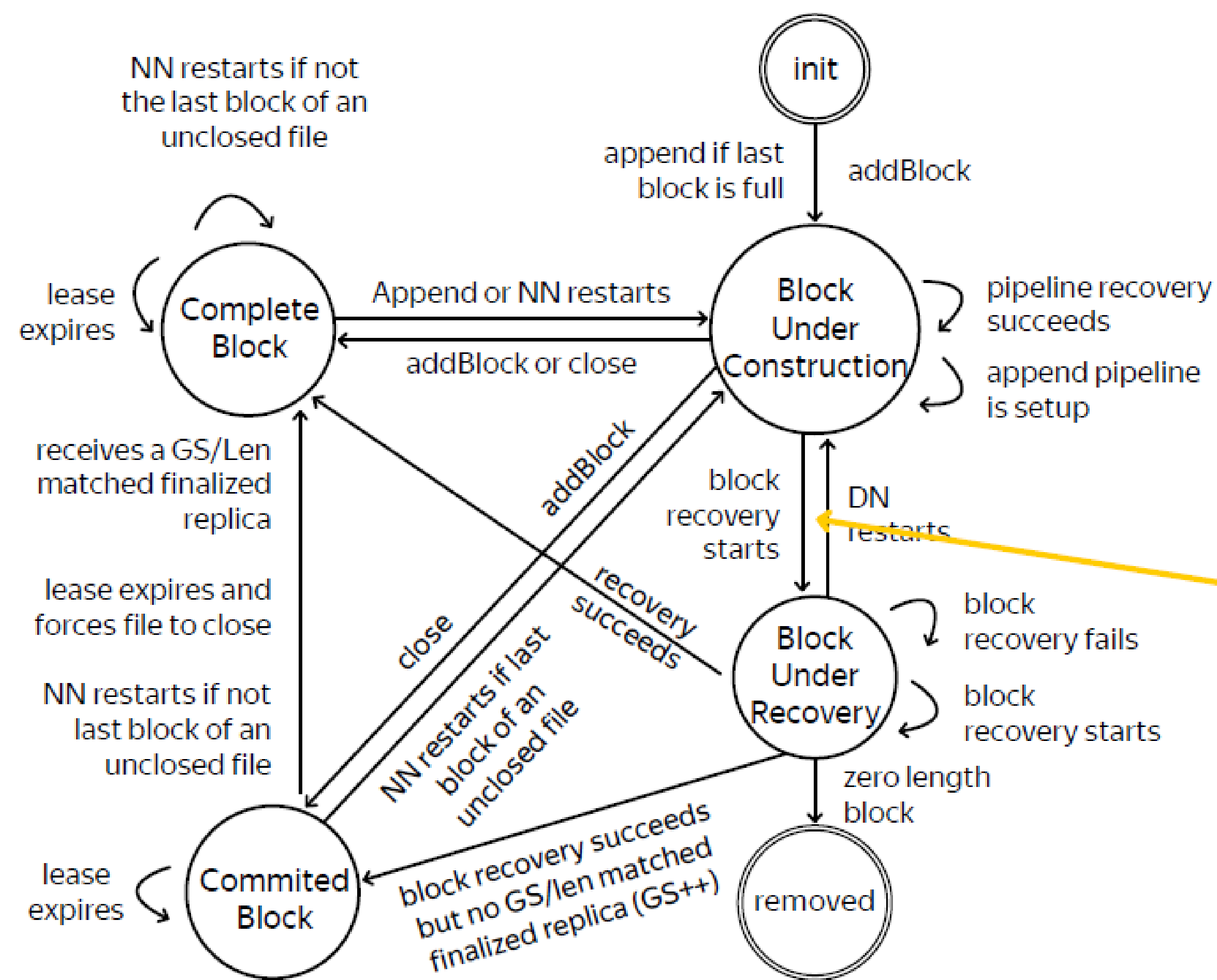


Simplified
Block State
Transition

under_construction

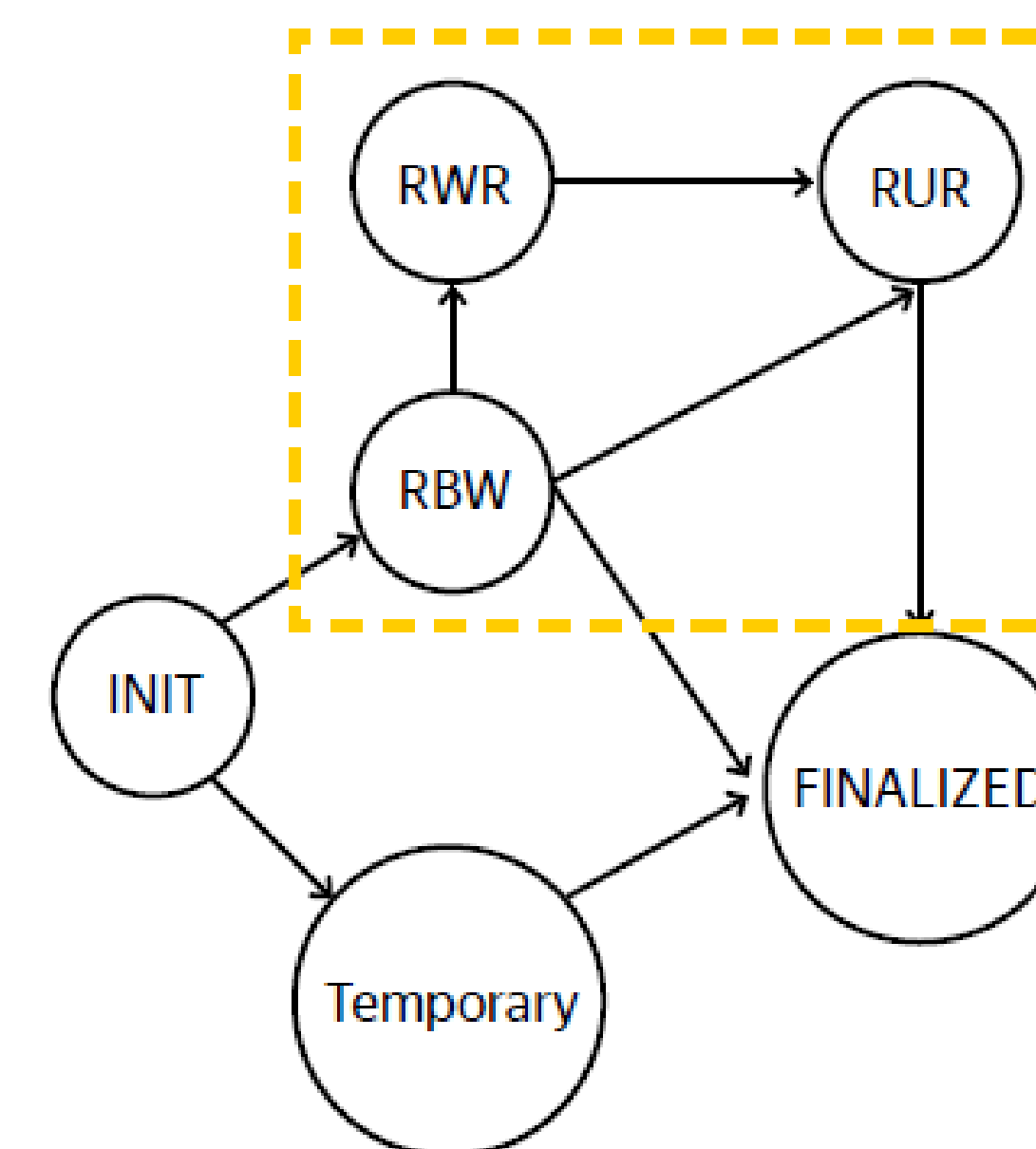


Under_construction 상태는 RBW와 RWR 상태를 포함한다.

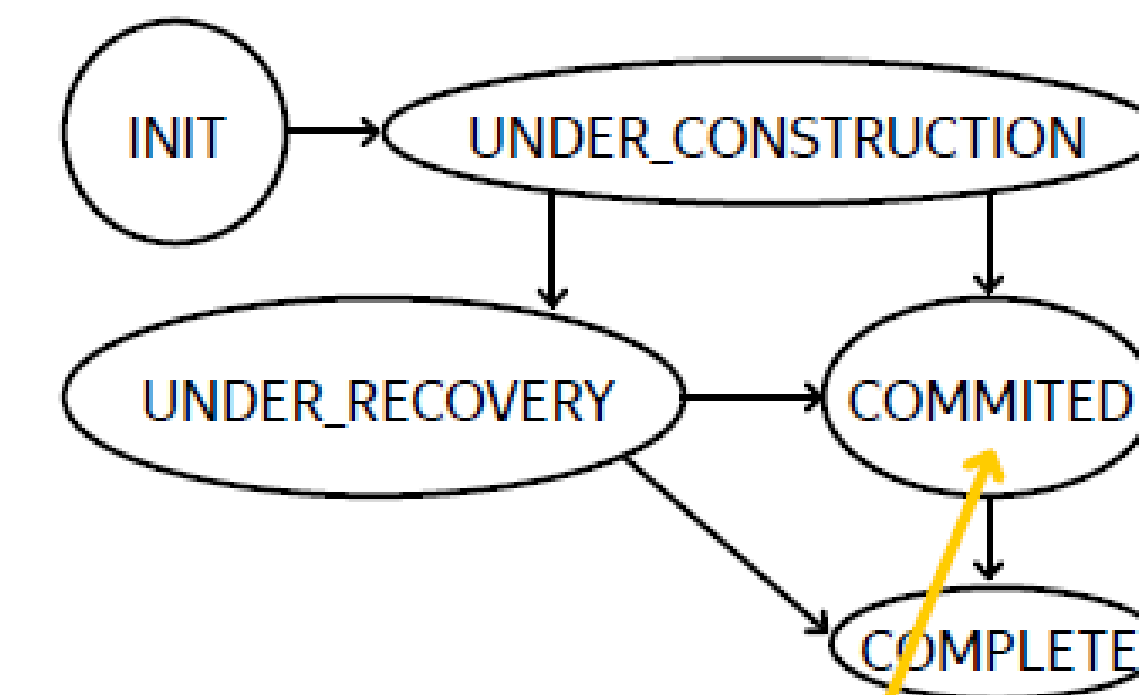
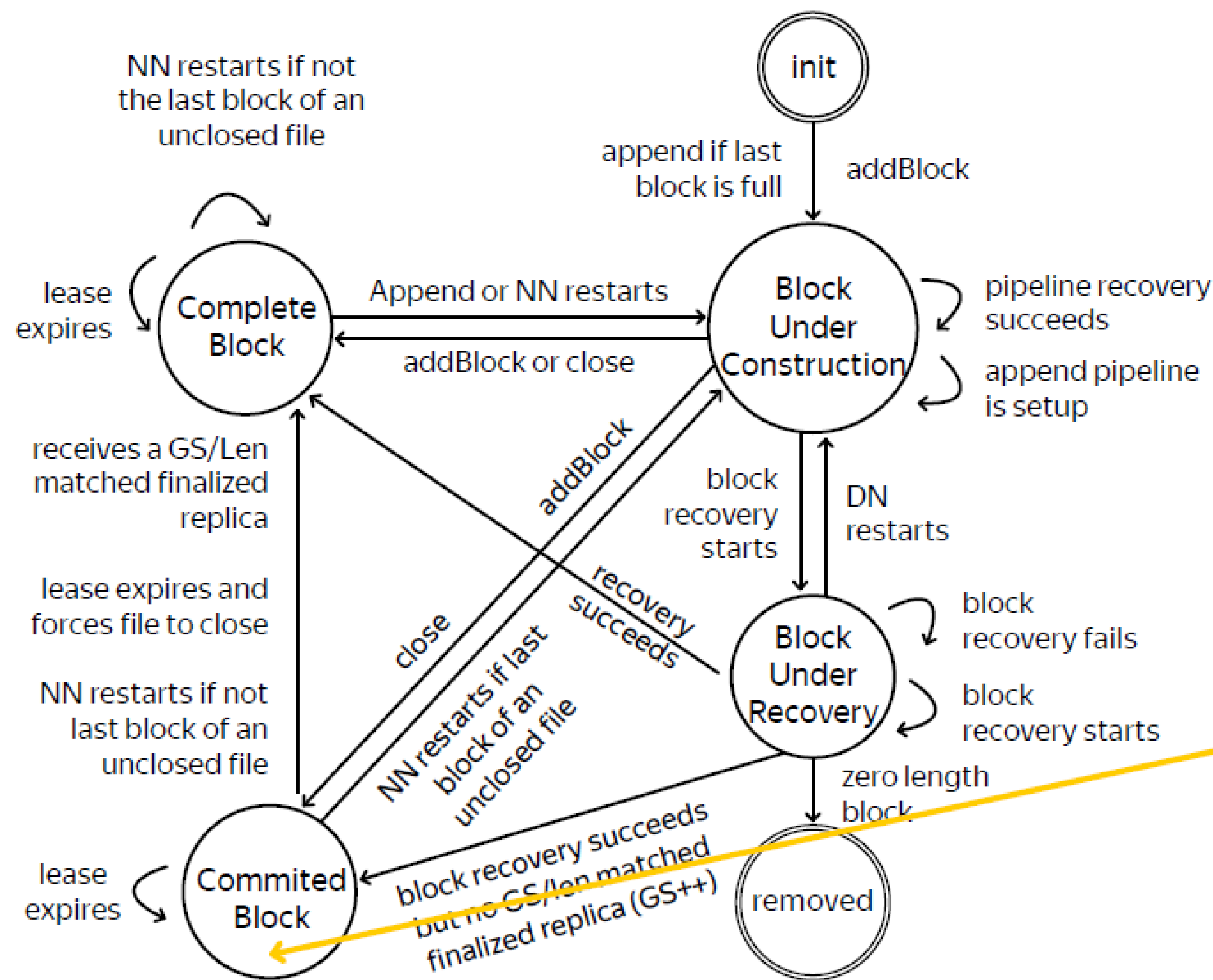


Simplified
Block State
Transition

under_recovery

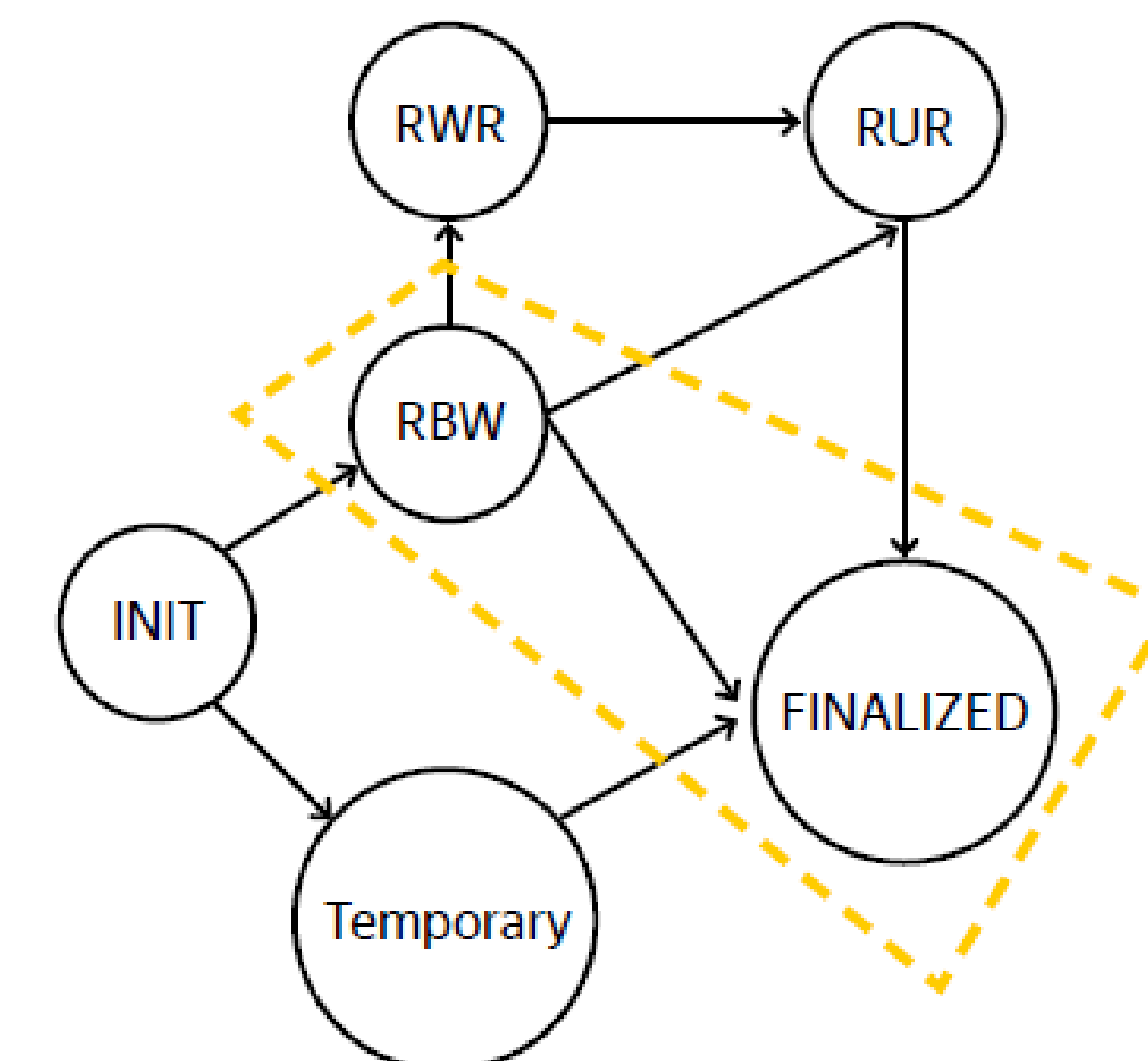


Replica가 RUR 상태가 될 때 under_recovery로 전환된다.

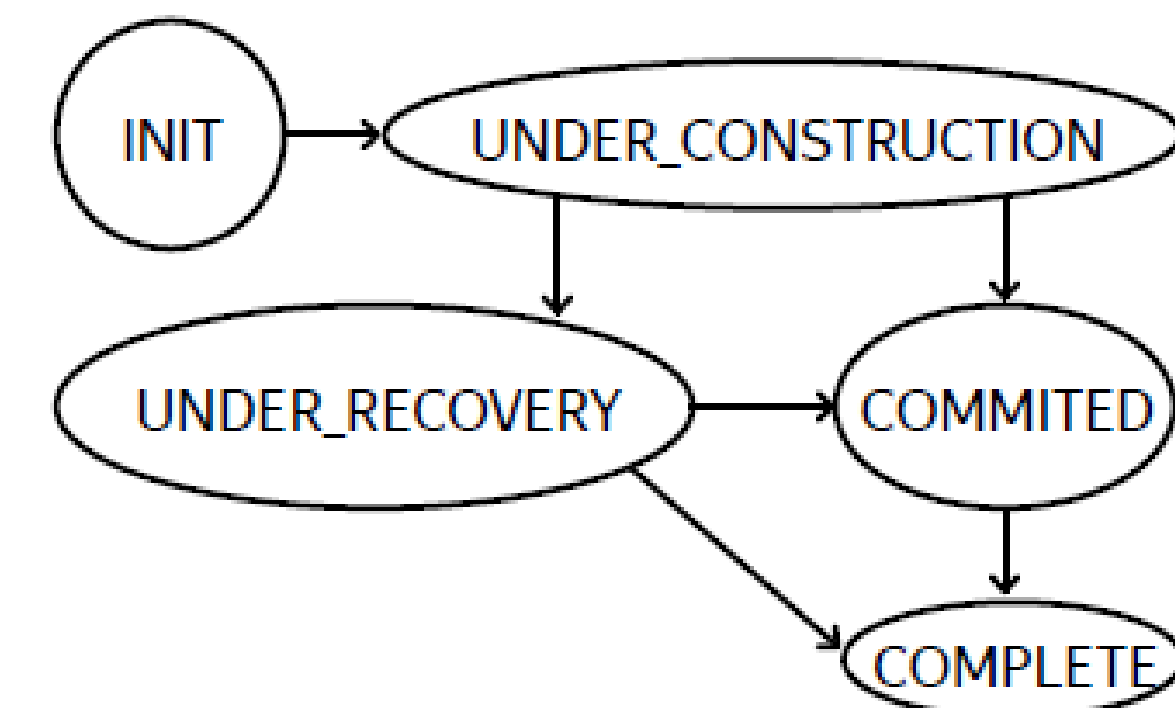
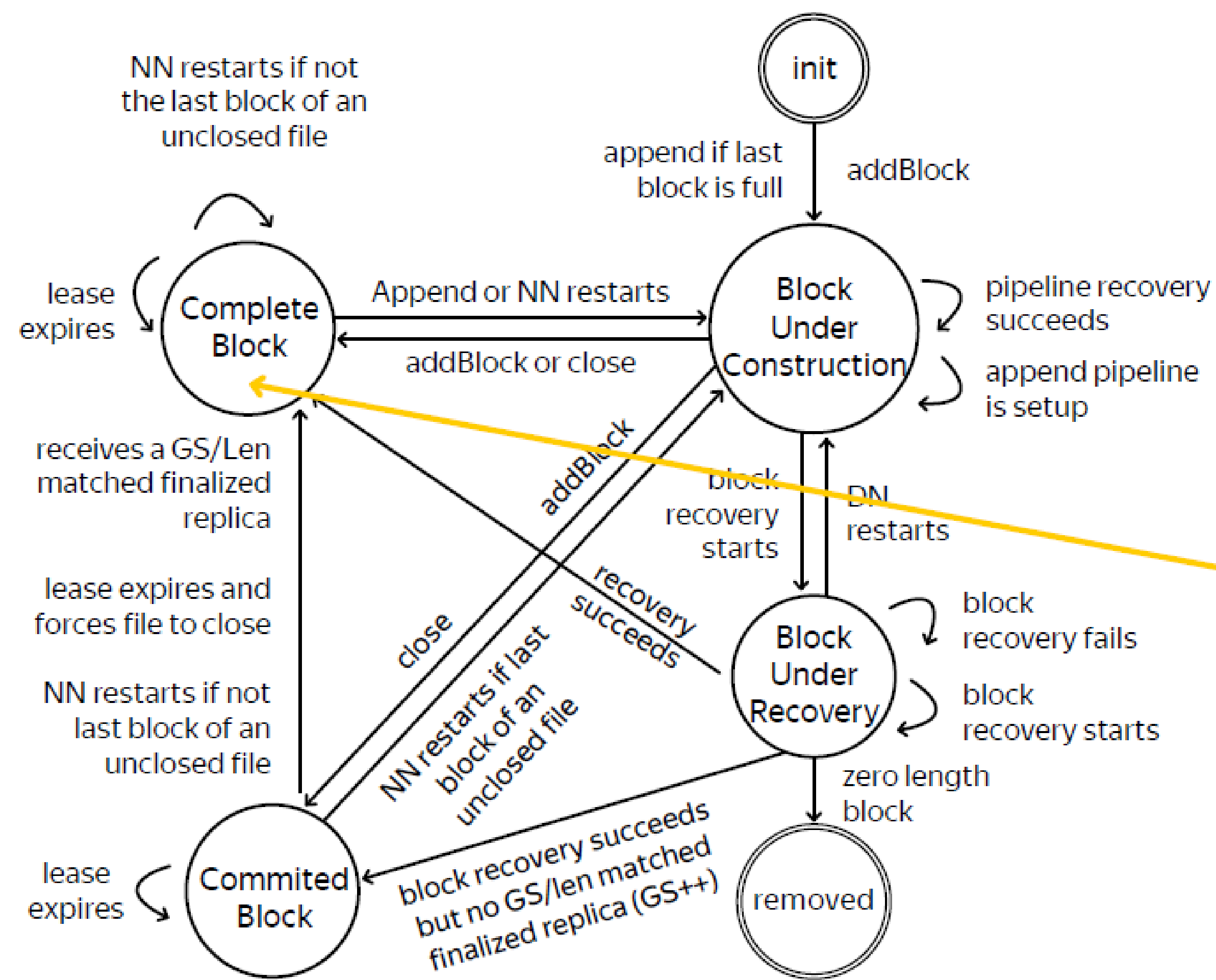


Simplified
Block State
Transition

committed

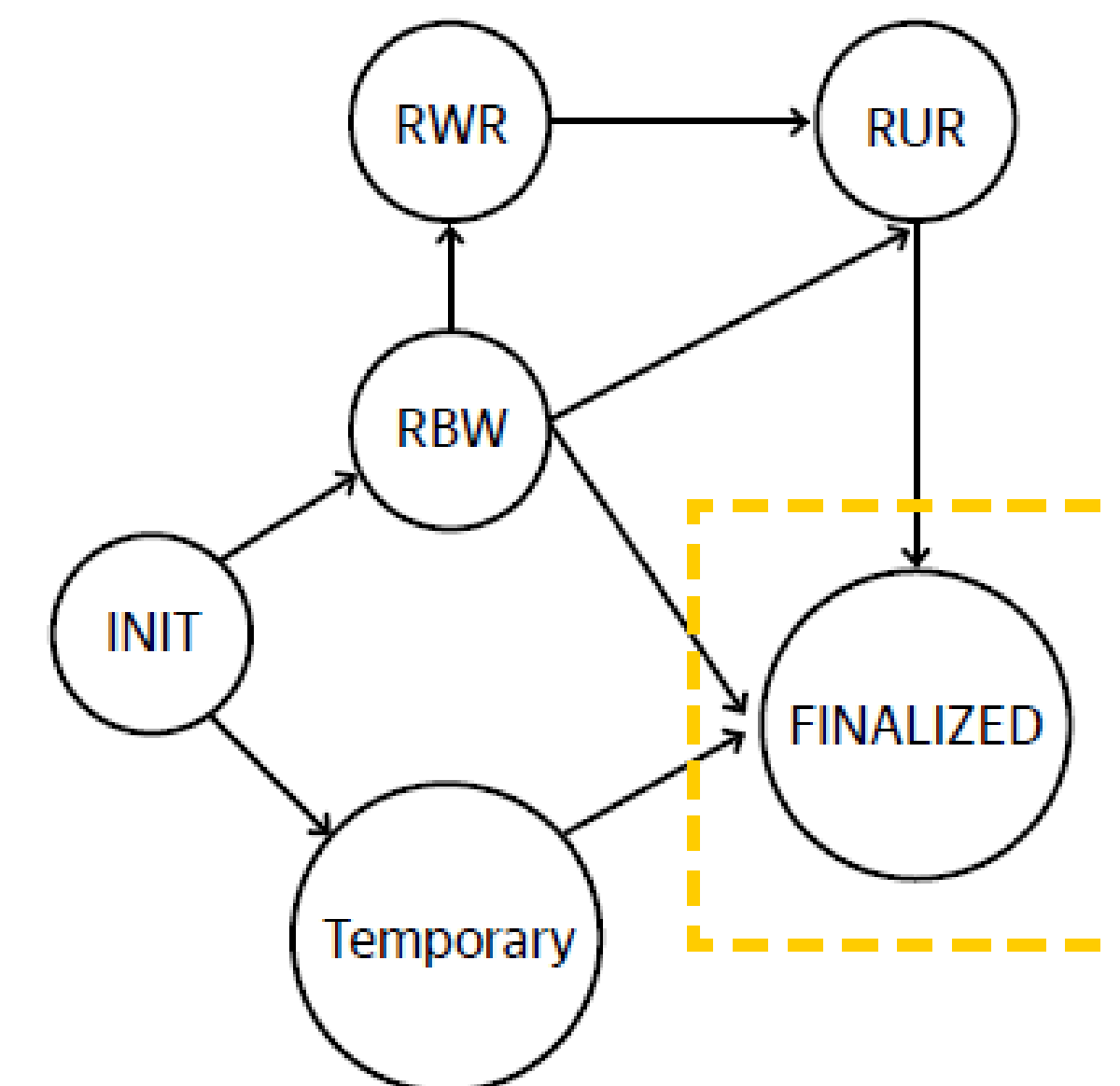


Committed 상태는 일부 replica가 finalized 상태가 되었다는 의미
남은 RBW들을 finalized 될 때까지 track한다.



Simplified
Block State
Transition

complete



모든 replica가 finalized 상태임을 의미한다.

Recovery

- Replica Recovery

- Block Recovery

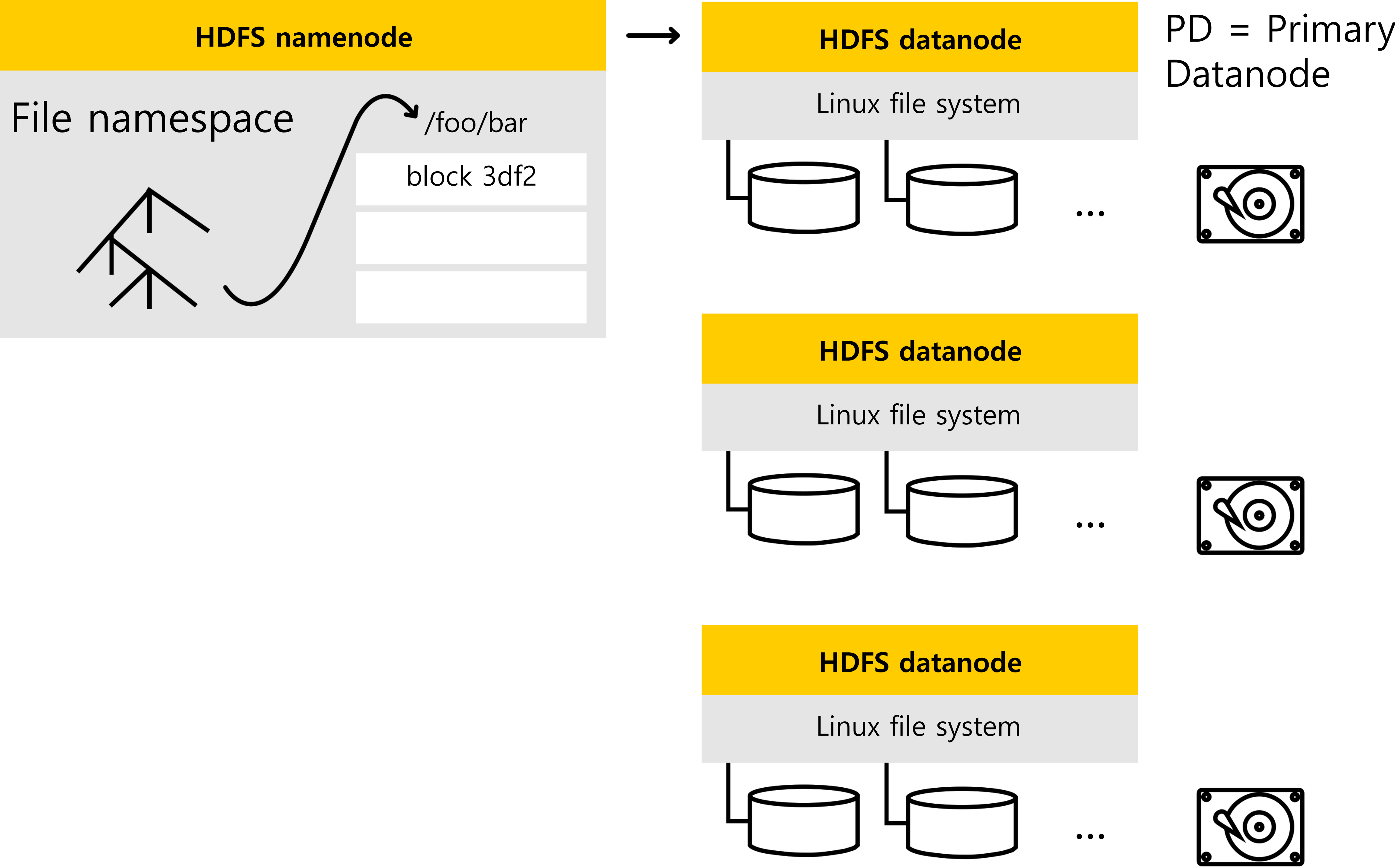
 - Block Recovery는 항상 Lease Recovery의 일부분이다.

- Lease Recovery

- Pipeline Recovery

Block Recovery

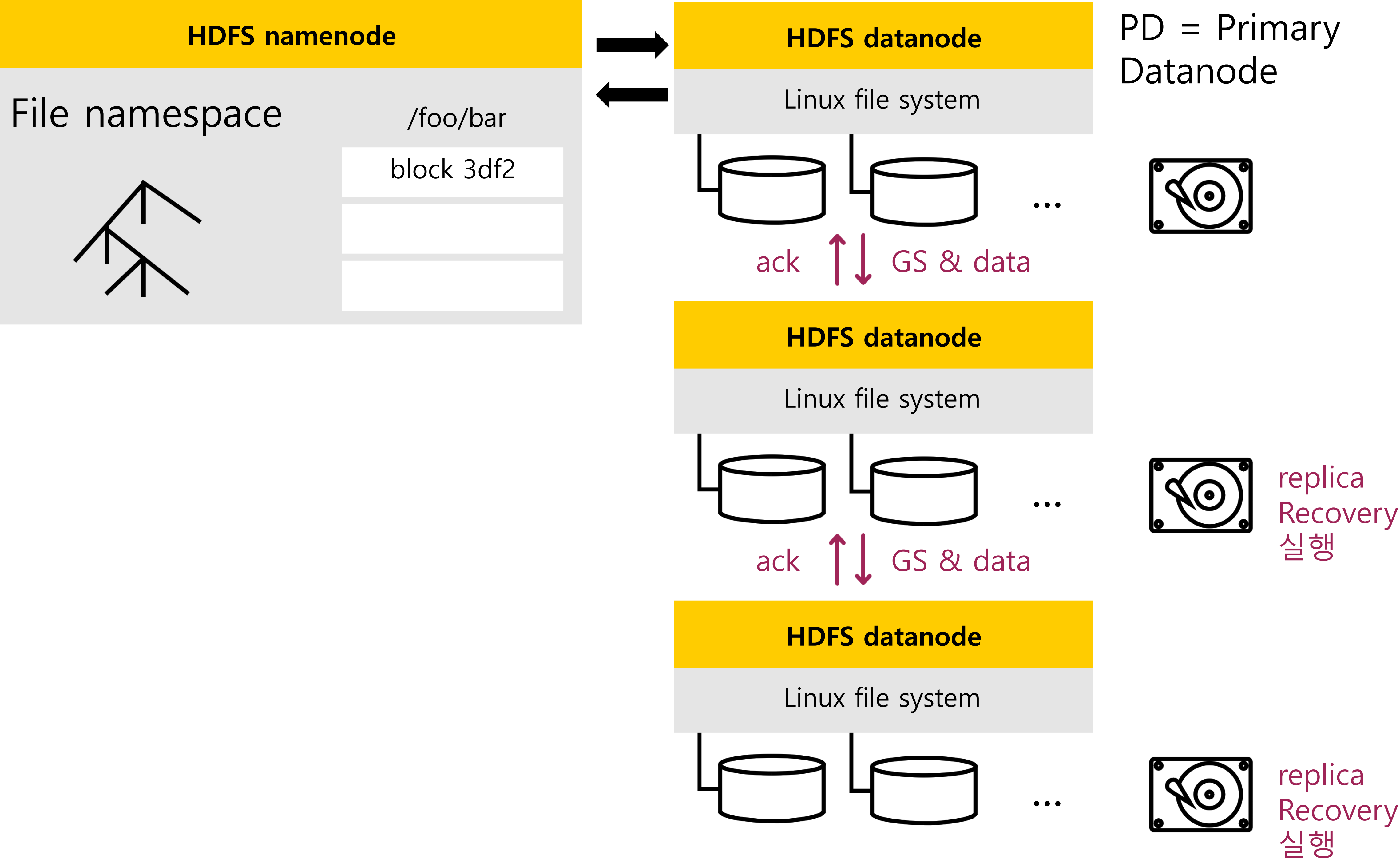
Recovery를 위해 해당 데이터를 가진 node 중 하나를 PD로 선택한다.



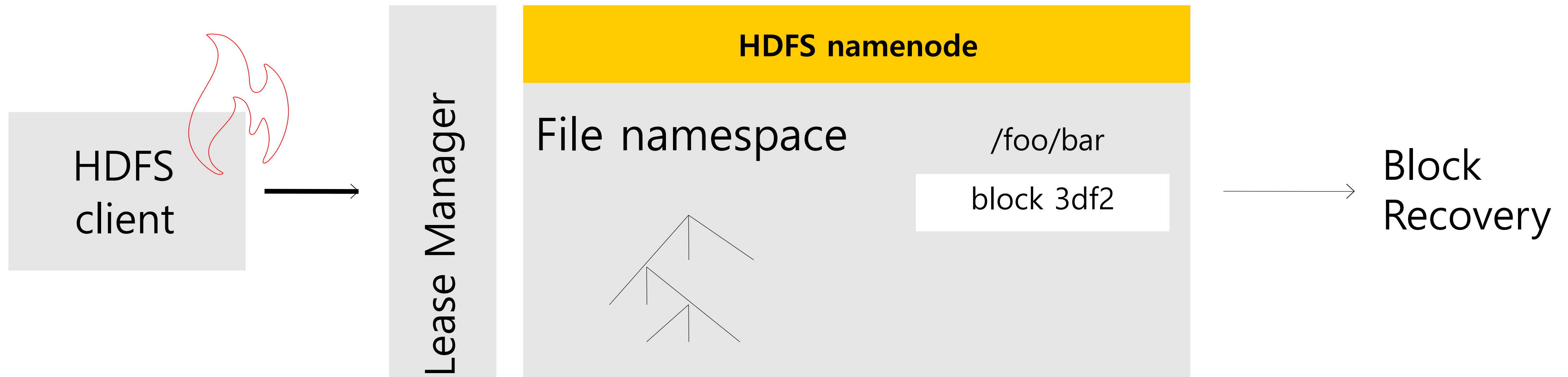
Block Recovery

PD가 GS와 복구할 datanode의 정보를 namenode에 요청한다.

Request & GET GS



Lease Recovery



user₁, /path/1, lease (soft + hard)
user₁, /path/2, lease (soft + hard)
user₂, /path/3, lease (soft + hard)
user₃, /path/4, lease (soft + hard)
...

Linux에서 시스템 리소스 제한 방식

Soft limit : user가 정한 최대 실행가능한 프로세스 수

Hard limit : 관리자(root)가 정한 최대 실행가능한 프로세스 수

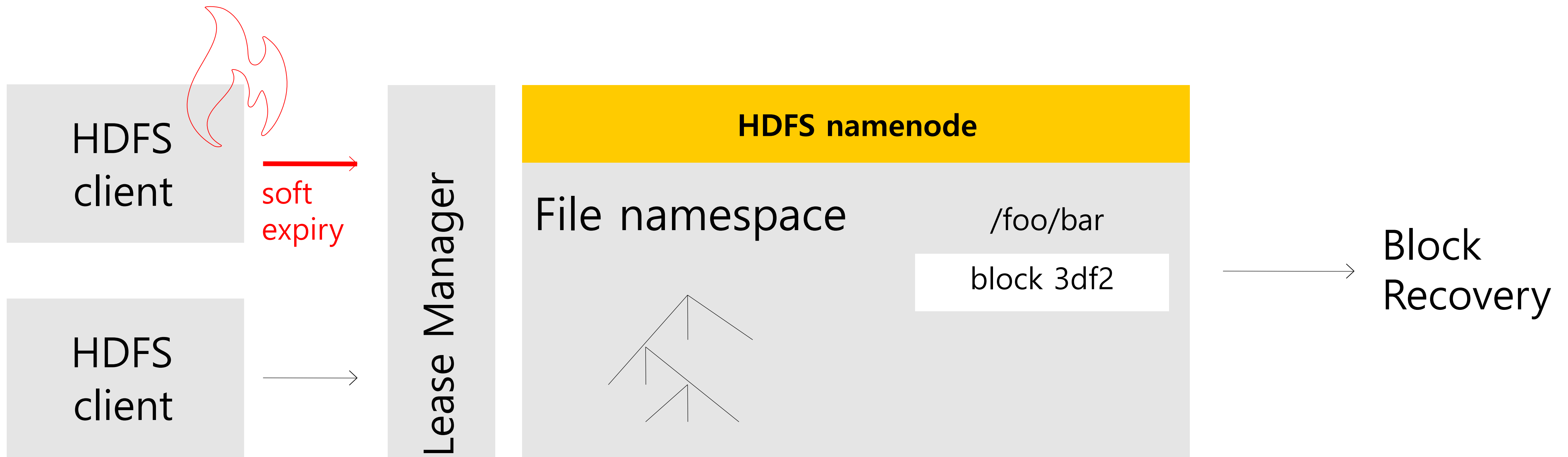
Soft limit ≤ Hard Limit

Lease Manager가 clients의 lease를 관리한다.

Lease Manager는 soft limit(1분) and hard limit(1시간) timeout을 유

User는 해당 시간 안에 lease를 renew하거나 파일을 닫아야 한다.

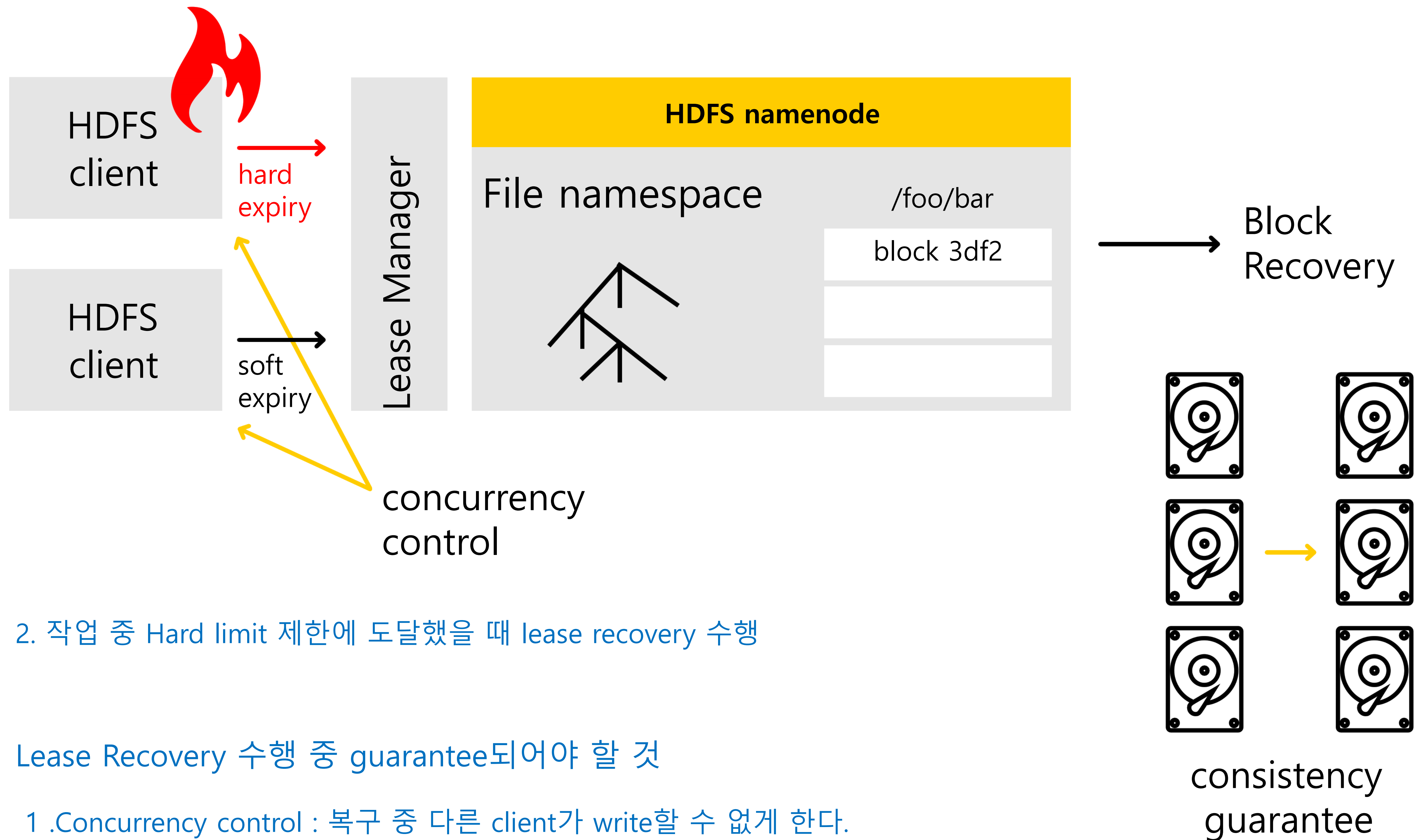
Lease Recovery



Soft limit expiry(timeout) 의 경우 다른 client가 lease를 가져갈 수 있다.

1. 작업중인 파일에 대한 Lease가 다른 client한테 넘어갔을 때 Lease Recovery 발생

Lease Recovery

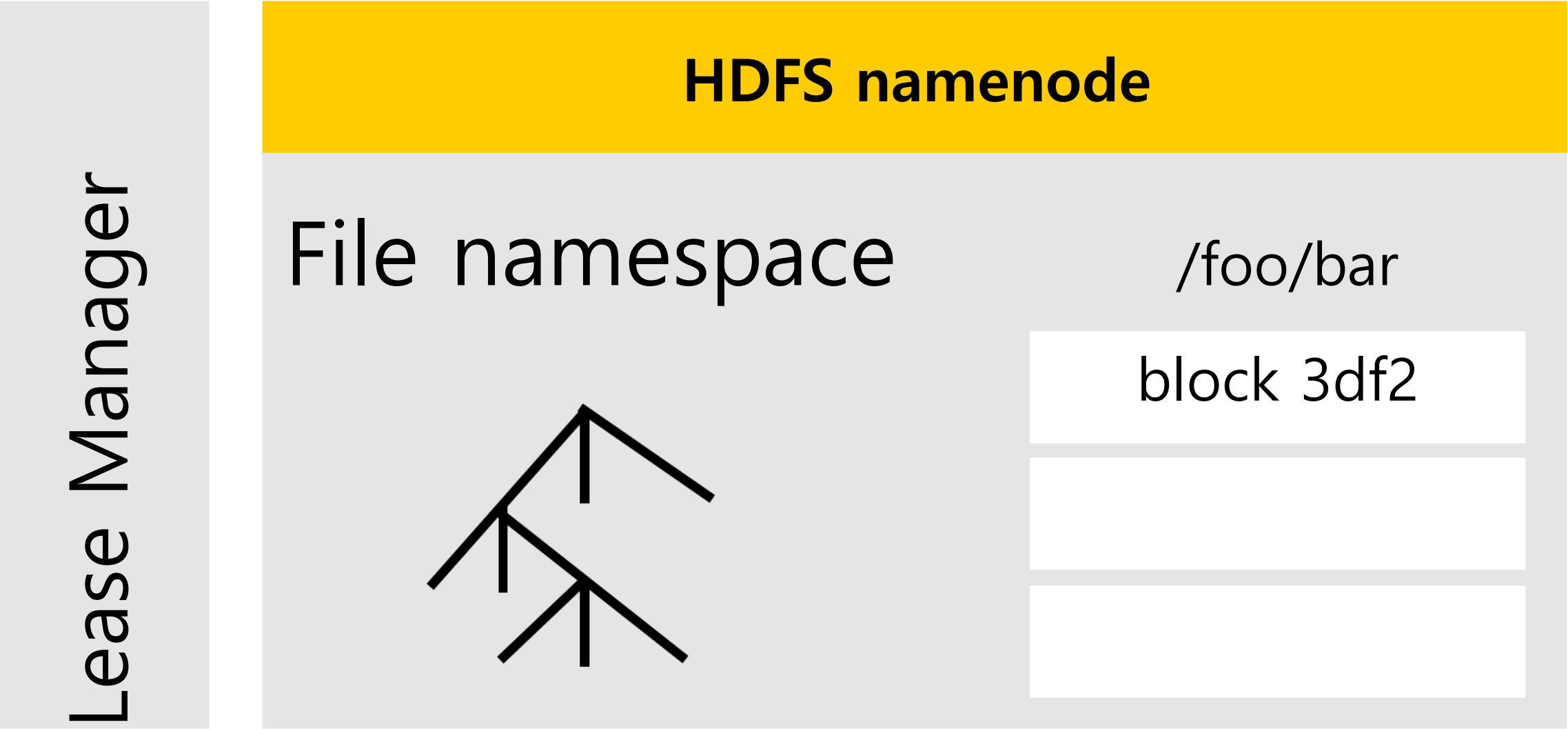


2. 작업 중 Hard limit 제한에 도달했을 때 lease recovery 수행

Lease Recovery 수행 중 guarantee되어야 할 것

1. Concurrency control : 복구 중 다른 client가 write할 수 없게 한다.
2. Consistency guarantee : 모든 replica가 동일하게 복구되어야 한다.

Lease Recovery



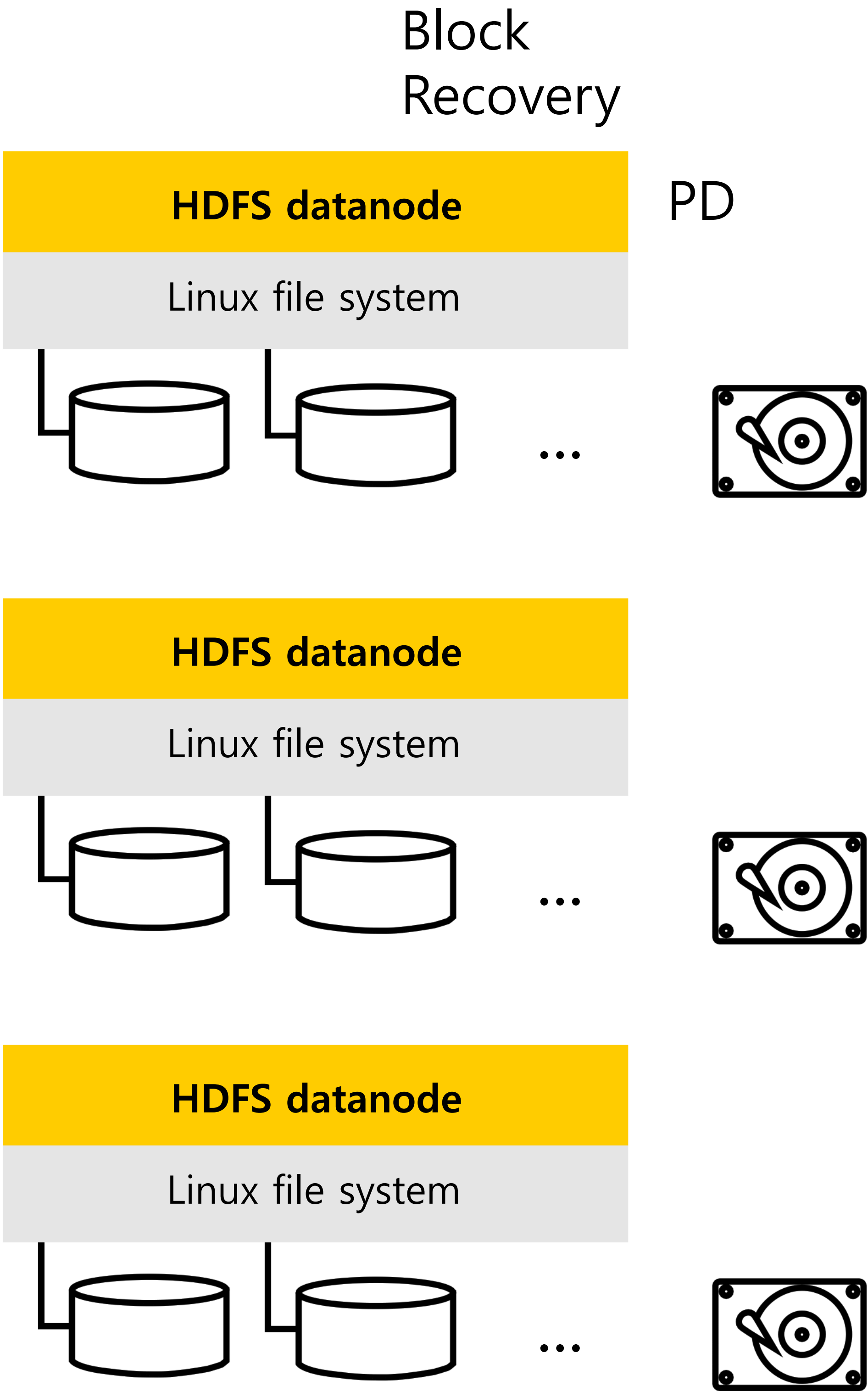
user₁, /path/1, expired lease



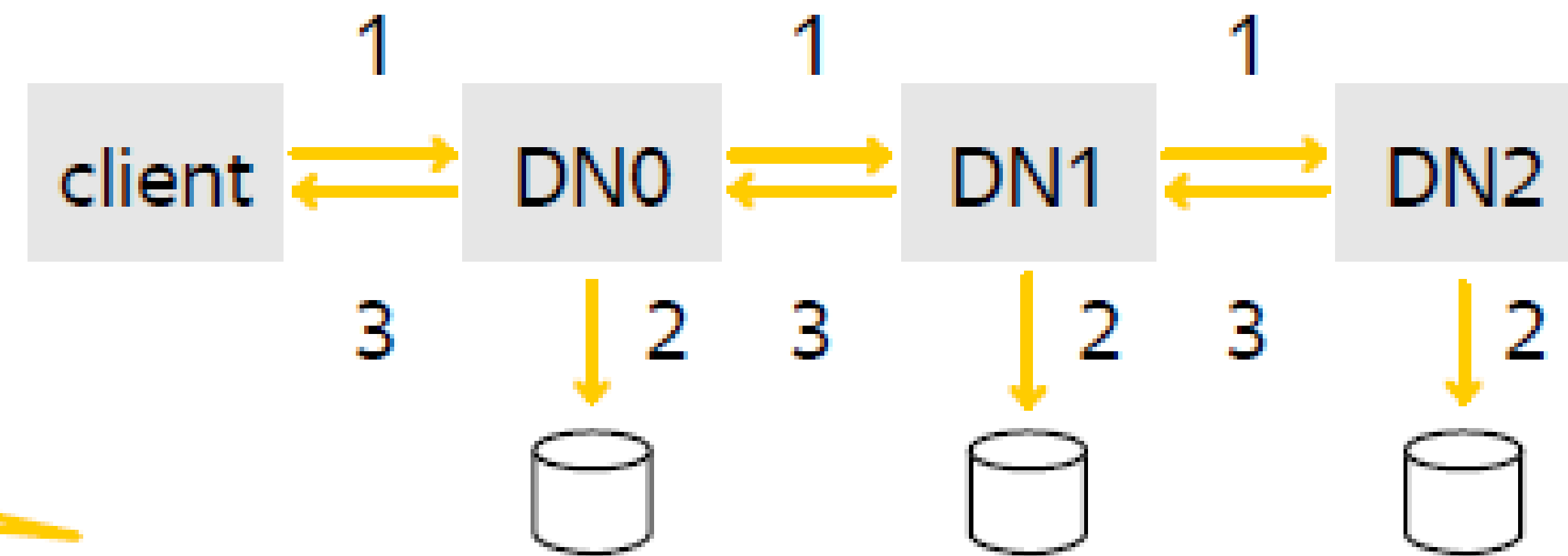
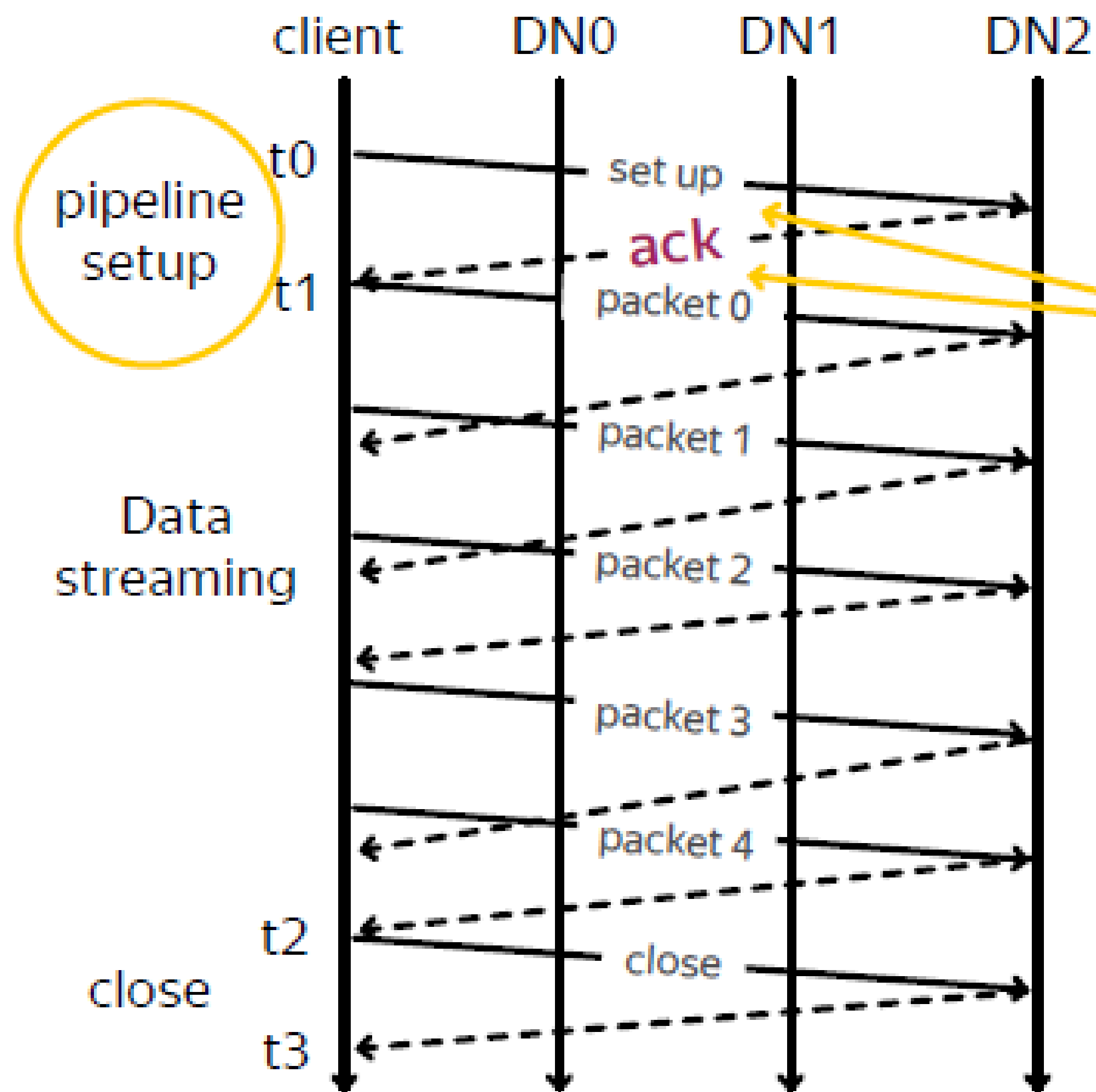
dfs, /path/1, new lease

Recovery가 실행되면,

dfs가 super user가 되어 다른 client의 모든 요청을 거부한다.



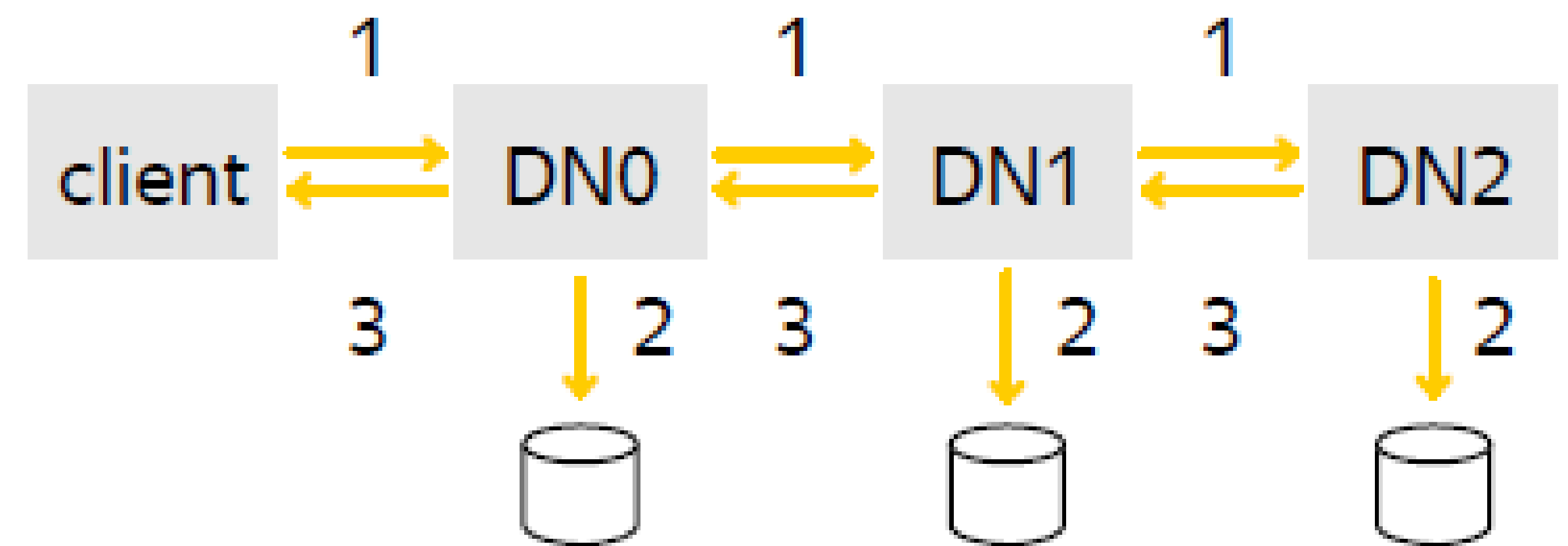
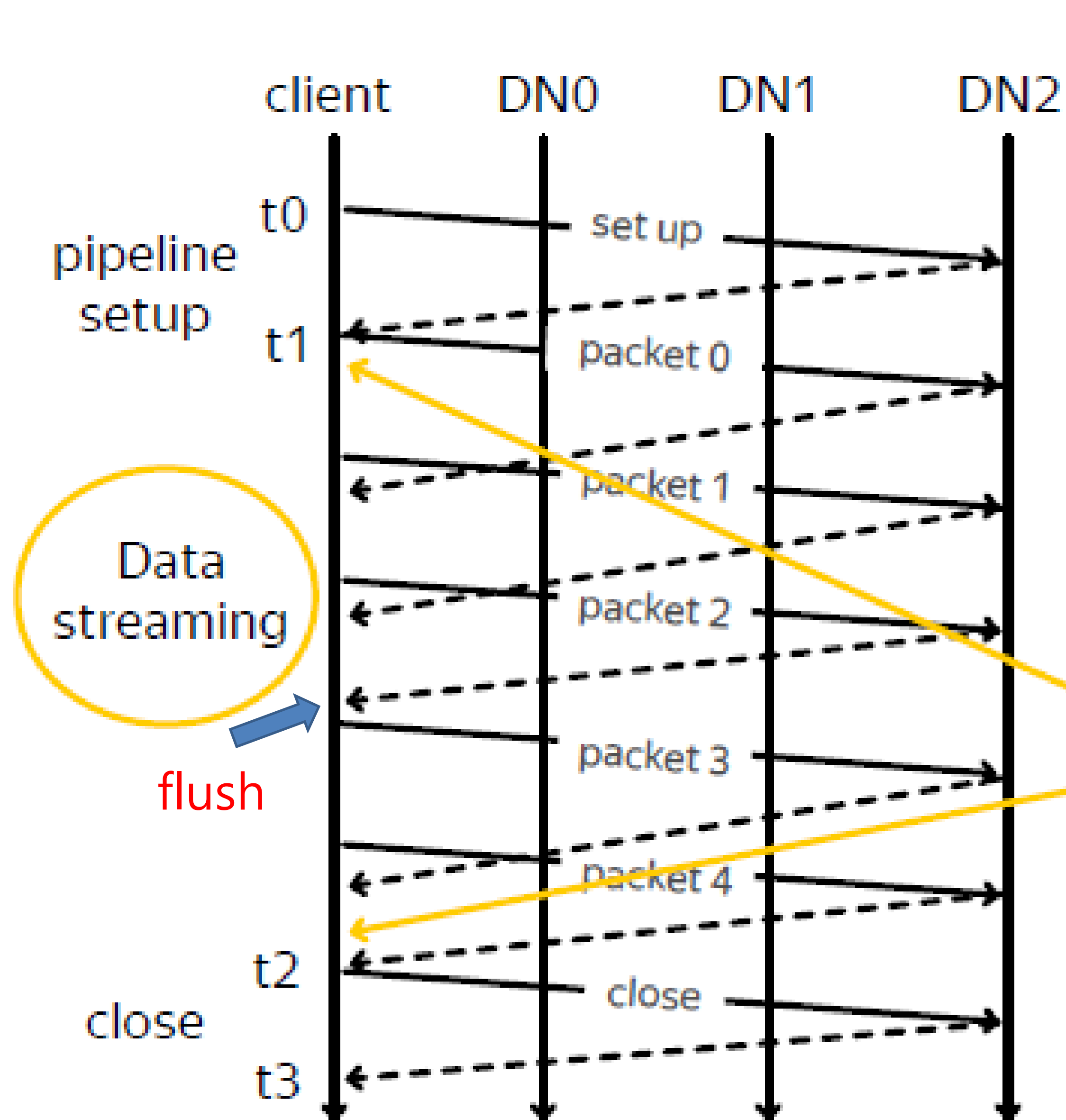
Pipeline



Data는 packet형태로 전송된다.

Ack packet으로 전송을 확인한다.

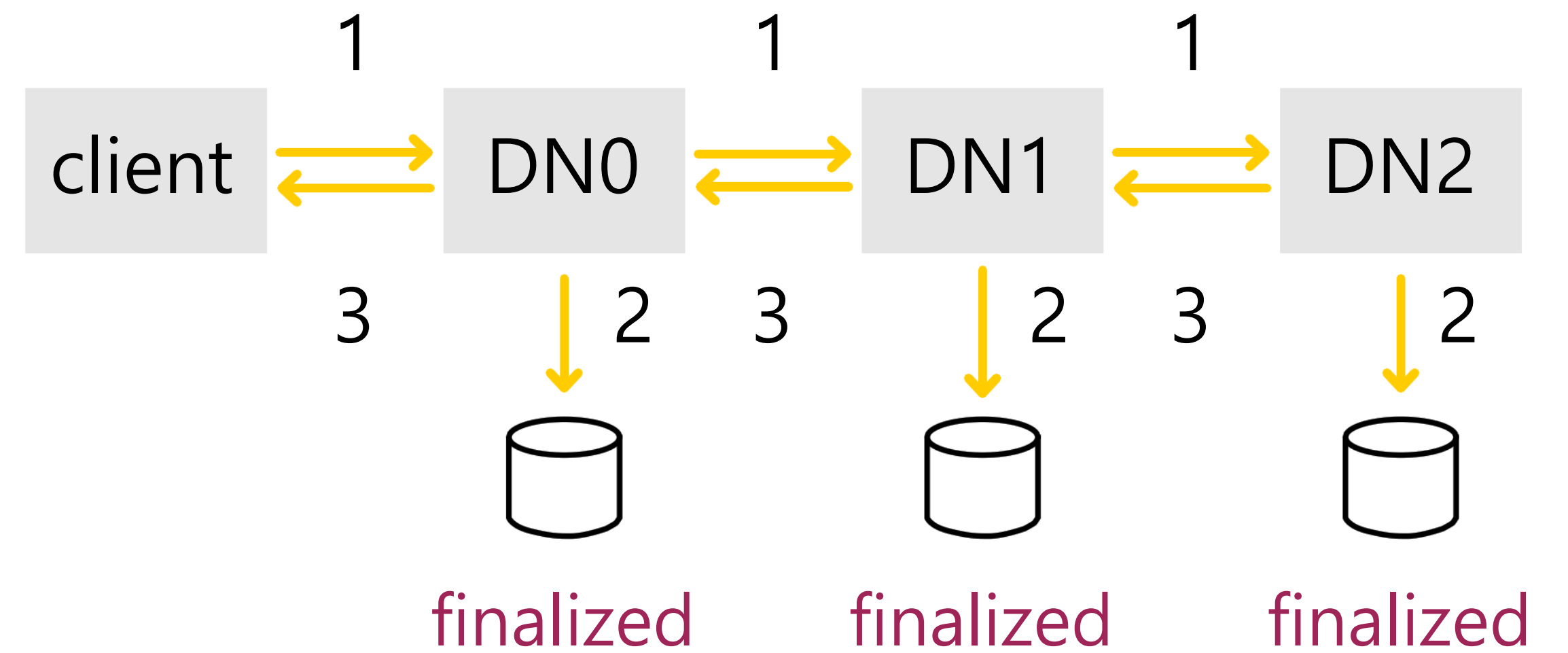
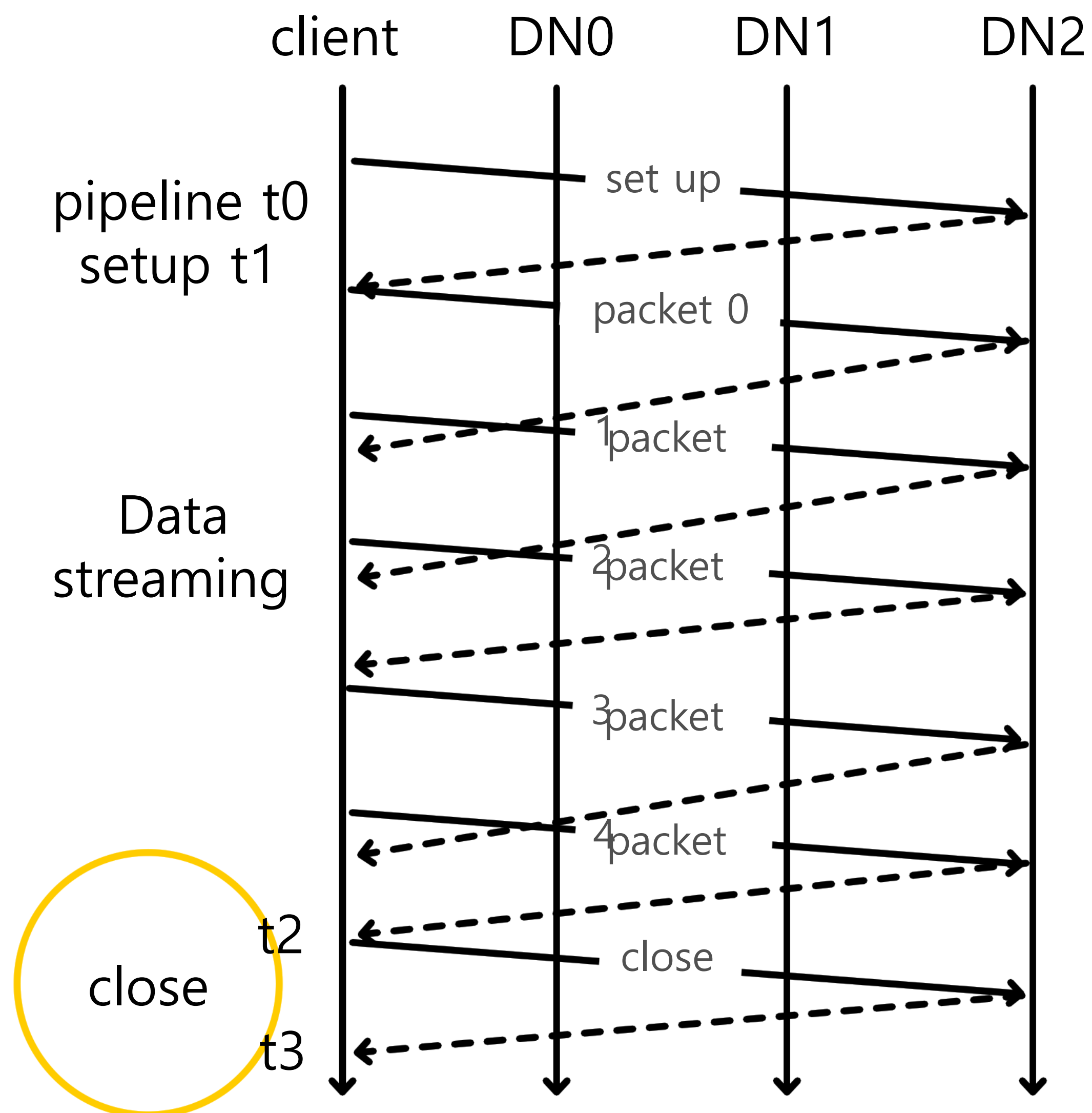
Pipeline



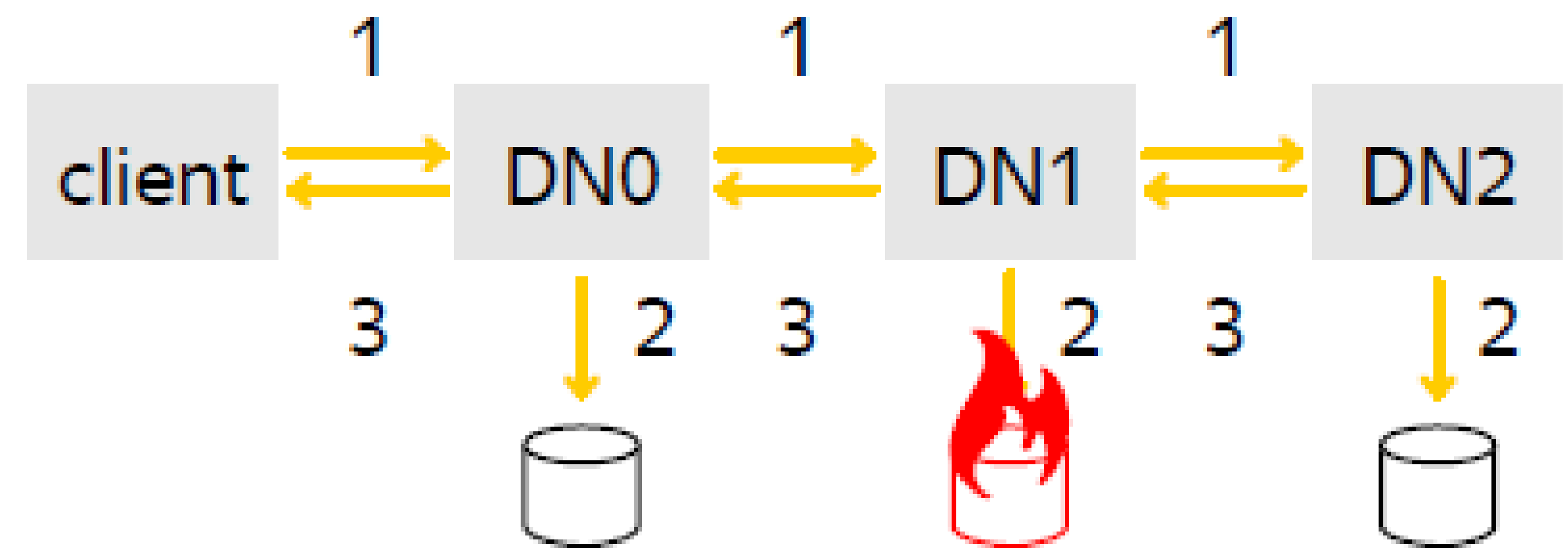
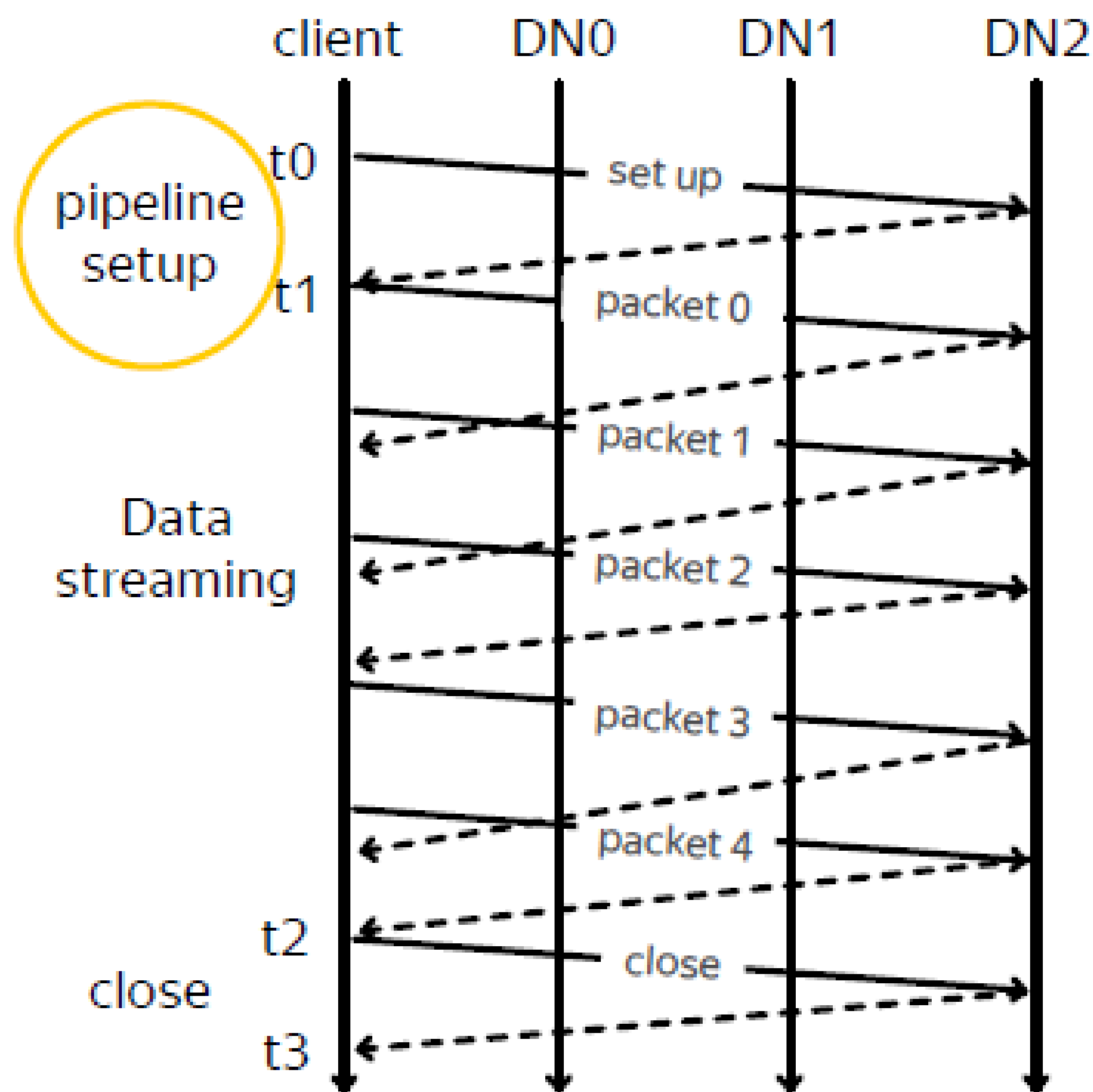
Ack 패킷이 도착하기 전에 data 패킷이 전송된다.

단, flush 패킷은 synchronous packet으로 데이터 동기화에 활용된다.

Pipeline



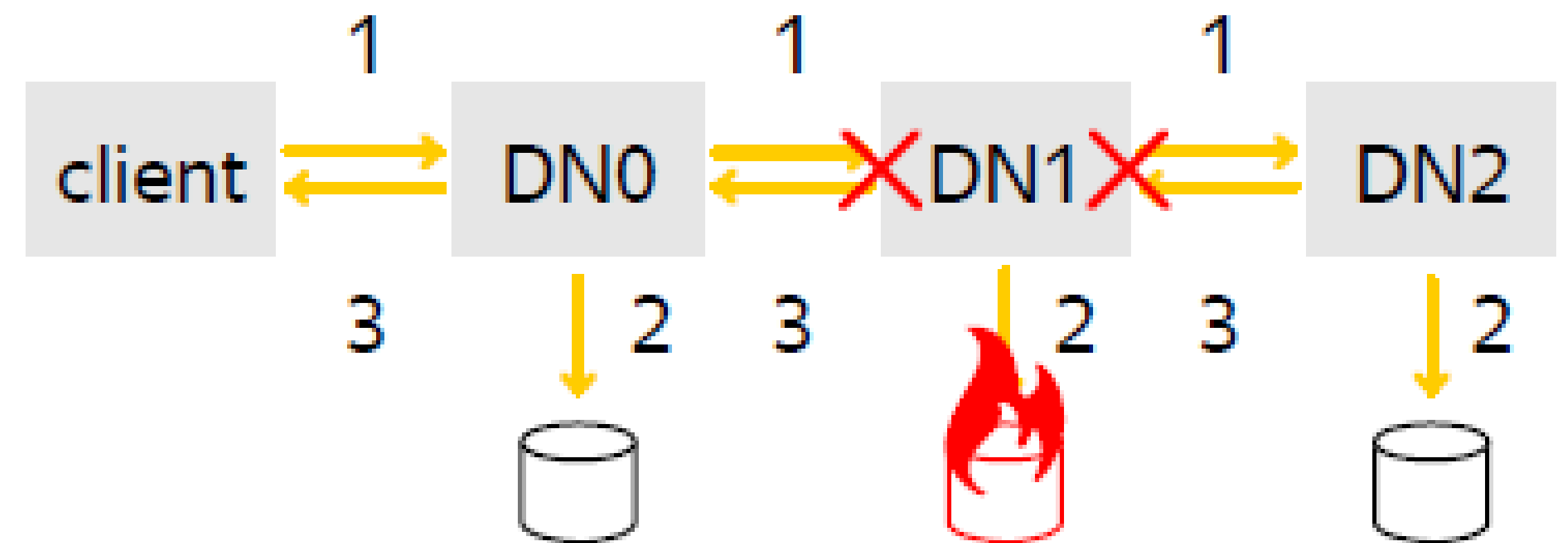
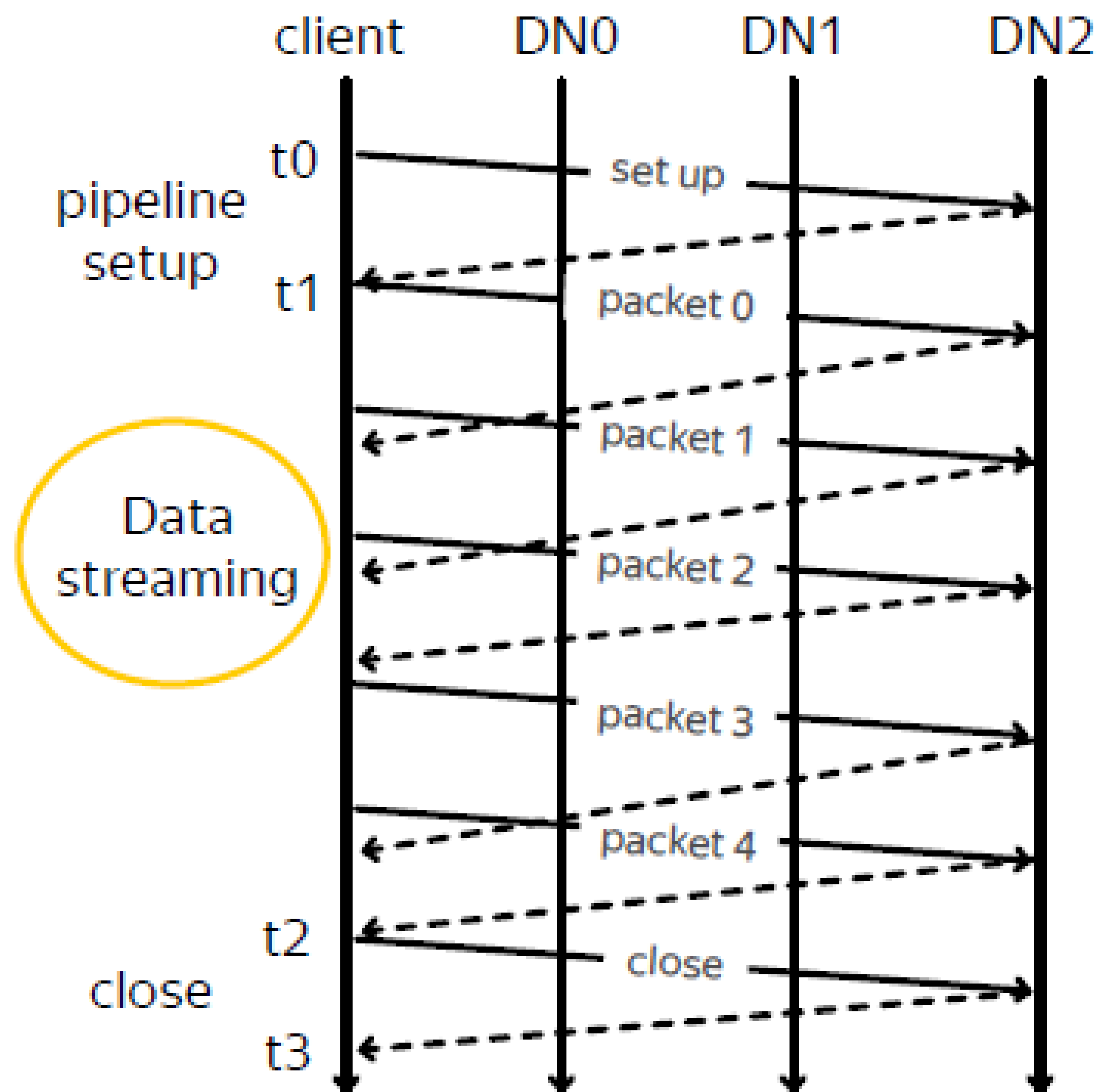
Pipeline Recovery



- cases:
1. new file
 2. append mode

Setup 중 failure 발생 시 새로 파이프라인을 만든다.

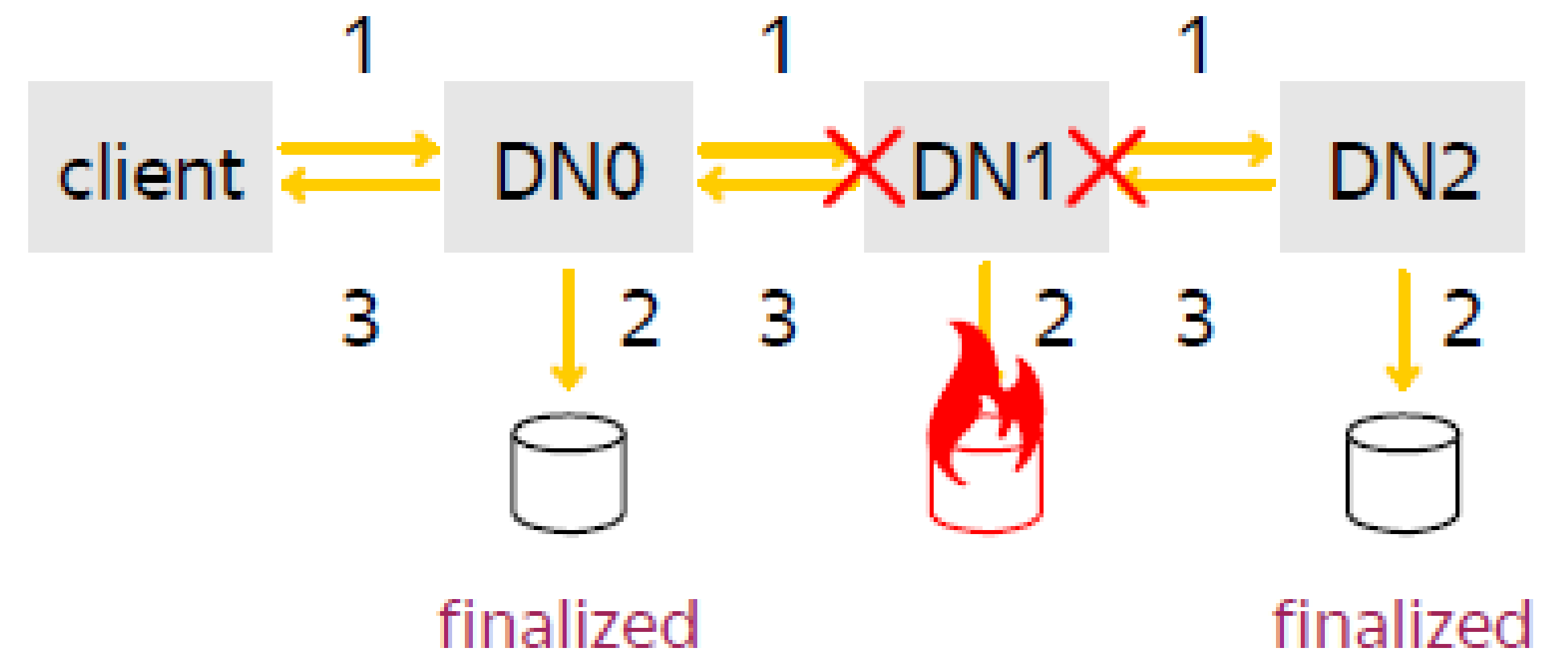
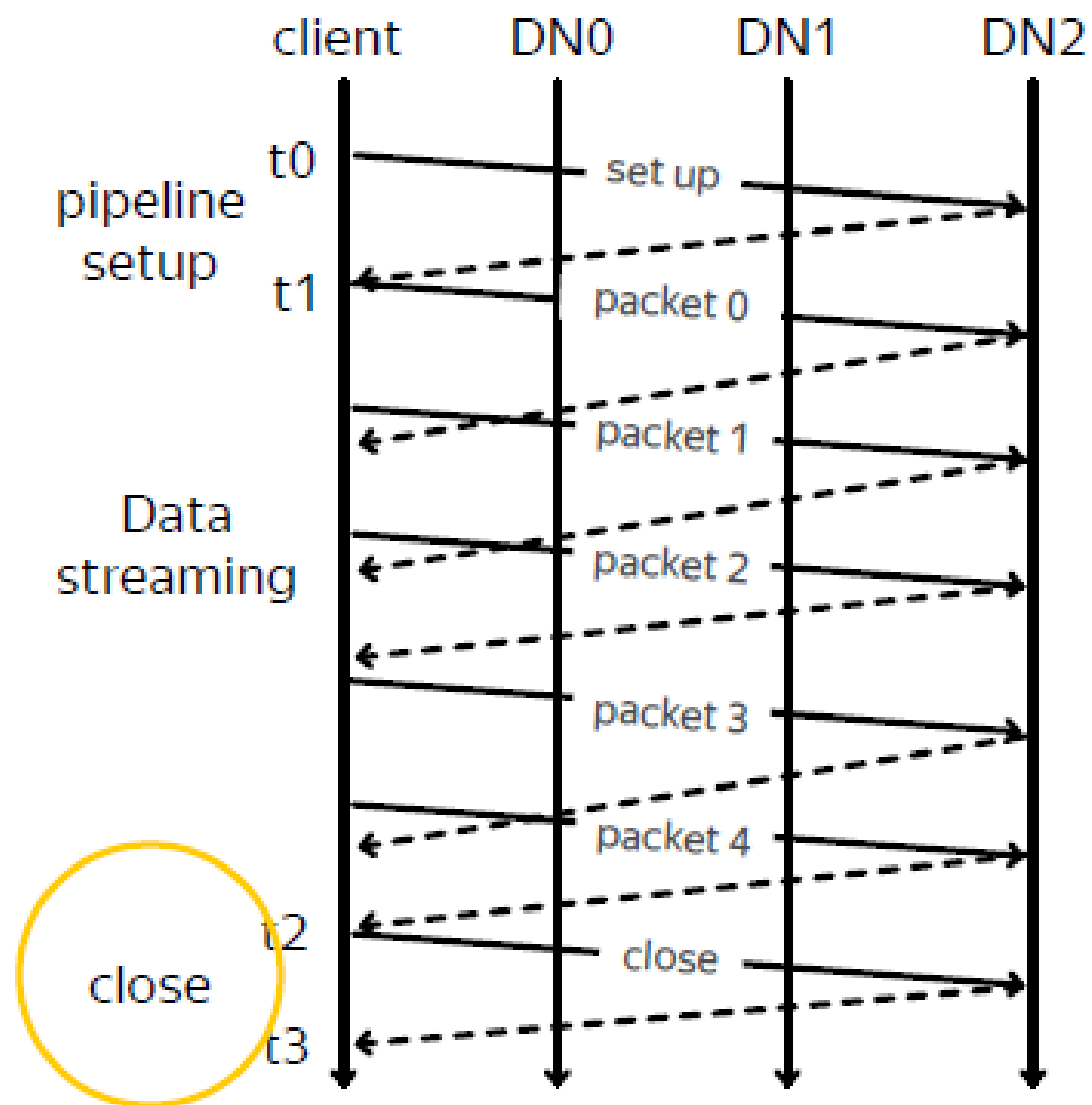
Pipeline Recovery



Data Streaming 중 failure가 발생하면, 패킷 전송을 중단하고 GS를 다시 받아 파이프라인을 구축한다.

이미 전송된 패킷은 디스크에 저장되기 때문에 재구축 시 IO overhead를 줄일 수 있다.


Pipeline Recovery



Close 시 failure가 발생하면 GS를 다시 발급받고 파이프라인을 다시 구축한다.

데이터는 이미 디스크에 저장되어 있기 때문에 추가적인 작업은 적다.

Pipeline Recovery



 Hadoop HDFS / HDFS-8344

NameNode doesn't recover lease for files with missing blocks





Agile Board

Export

Details

| | | | |
|--------------------|---------------------------------------------------------------------------------------------------------|----------------|------------------------|
| Type: |  Bug | Status: | PATCH AVAILABLE |
| Priority: |  Major | Resolution: | Unresolved |
| Affects Version/s: | 2.7.0 | Fix Version/s: | None |
| Component/s: | namenode | | |
| Labels: | None | | |
| Target Version/s: | 2.9.0 | | |
| Release Note: | Allow a configuration to specify the maximum number of recovery attempts for blocks under construction. | | |

People

| | |
|-----------|--------------------------------------------------------------------------------------------------------------------|
| Assignee: |  Ravi Prakash |
| Reporter: |  Ravi Prakash |
| Votes: |  0 Vote for this issue |
| Watchers: |  29 Start watching this issue |

Dates

| | |
|----------|-----------------|
| Created: | 07/May/15 17:17 |
| Updated: | 10/Mar/17 16:10 |

Summary

- > you can **draw** State block and replica transition tables
- > you can **identify** write pipeline stages and associated recovery process
- > you can **list** 4 recovery processes and **explain** their purpose (lease recovery → block recovery → replica recovery; pipeline recovery)

BigDATAteam