THE UNIVERSITY OF MELBOURNE

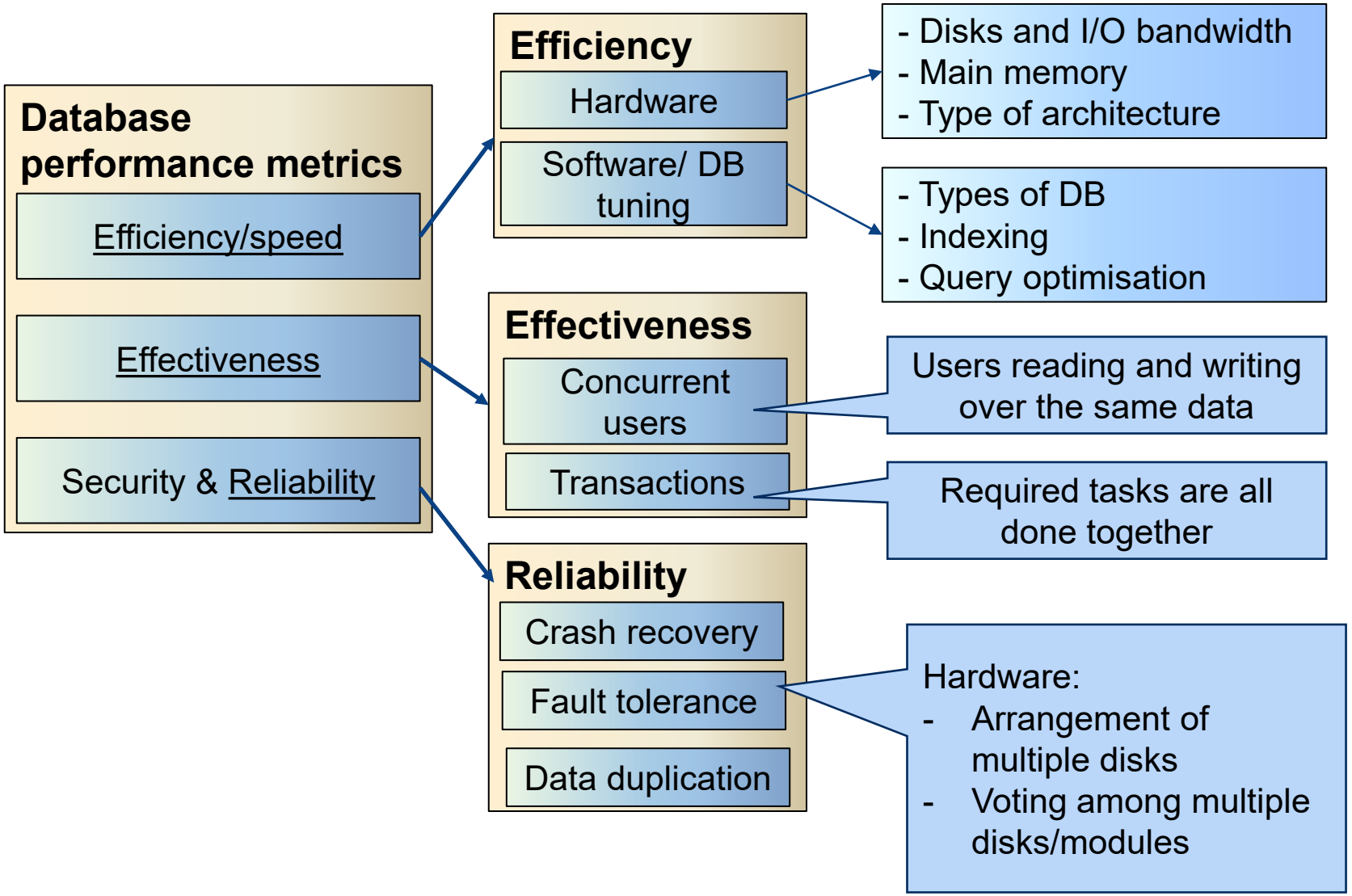
COMP90050 Advanced Database Systems

Winter Semester, 2023

Lecturer: Farhana Choudhury (PhD)

Week 1 part 3

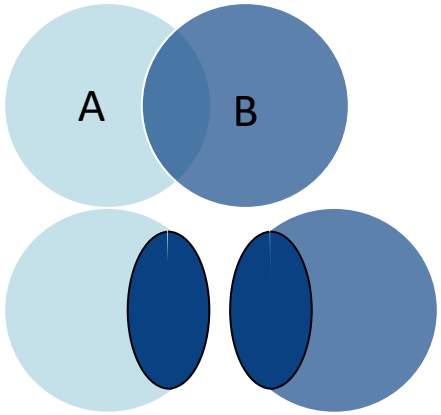
Core Concepts of Database management system



Fault Tolerance

The property that enables a system to continue operating properly in the event of the failure of some of its components.

Some statistics first!



P(A) = probability of an event A is happening in a **certain period.**

P(A and B) = probability both A and B happening in that period

= P(A)*P(B) assuming A and B are independent events.

P(A or B) = P(A) + P(B) - P(A and B)

= P(A) + P(B) - P(A)*P(B) [Assuming A and B are independent]

≈ P(A) + P(B) [if P(A) and P(B) are very small]

Mean time to event A is, MT(A) = 1/P(A)

If events A, B, have mean time MT(A), MT(B), then the mean time to the first event, MT(A or B) = 1/P(A or B)

If there are n events, each with the same probability p, then

1. Probability that one of the events occur = p + p + …. (n times)

   = n*p [assuming p is small]

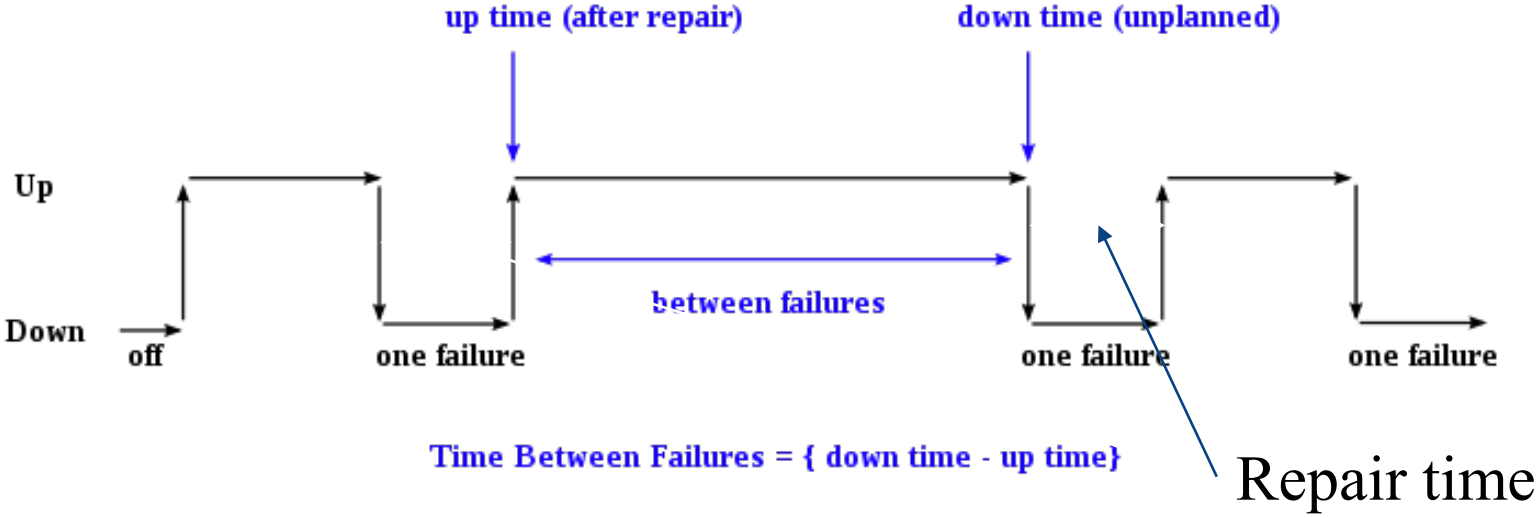
2. Mean time to one of the events (i.e., mean time to the first event)

   = 1/(n*p)

   = (1/p) * (1/n)

   = m * (1/n) = m/n

A system's lifecycle



Module availability : measures the ratio of service accomplishment to elapsed time

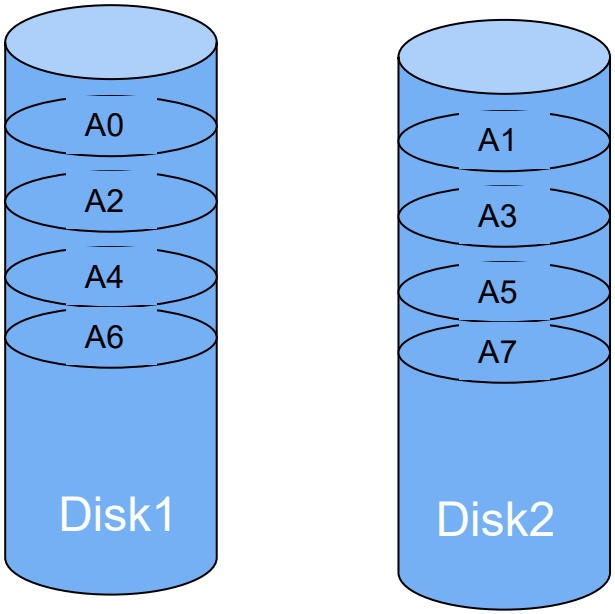
the **time** elapsing before a failure is experienced

$$= \frac{\text{Mean time to failure}}{\text{Mean time to failure} + \text{mean time to repair}}$$

Fault tolerance by RAID

Redundant Array of Independent Disks – different ways to combine multiple disks as a unit for fault tolerance or performance improvement, or both of a database system

RAID 0 (Block level Striping)



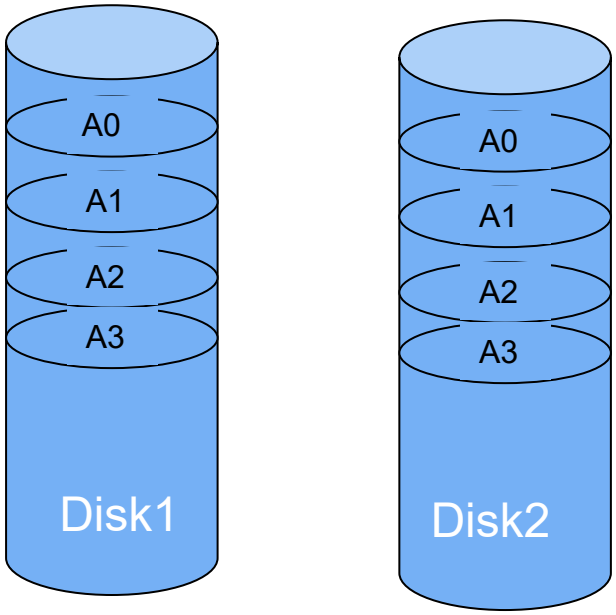
A0, A1, A2, ... are contiguous blocks of data of a file

- Provides balanced I/O of disk drives –throughput ~doubles
- Any disk failure will be catastrophic and MTTF reduces by a factor of 2

Higher throughput at the cost of increased vulnerability to failures

- A means Block (4K or 8K bytes of storage)
- MTTF = Mean Time To Failure

RAID 1 (mirroring)



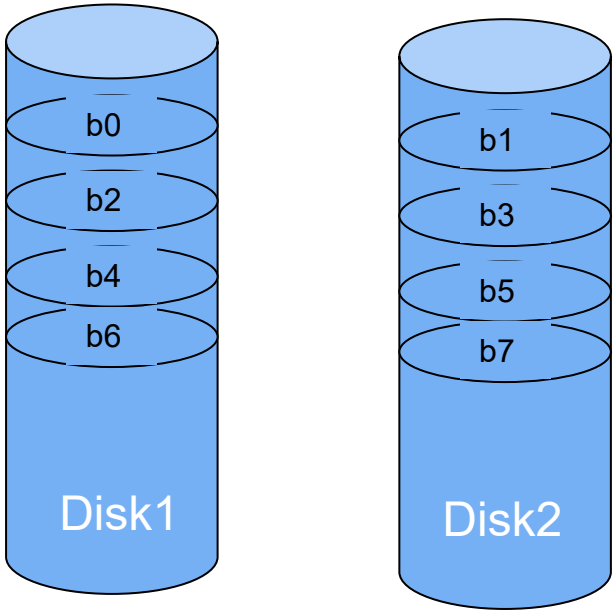
- Provides higher read throughput but lower write throughput (half of the total speed –i.e. single disk speed)
- Half storage utilization.
- MTTF increases substantially (quadratic improvement –i.e. $MTTF^2$!)

Continues to operate as long as 1 disk is functional

Calculation of MTTF values – in tutorials

- A means Block (4K or 8K bytes of storage)
- MTTF = Mean Time To Failure

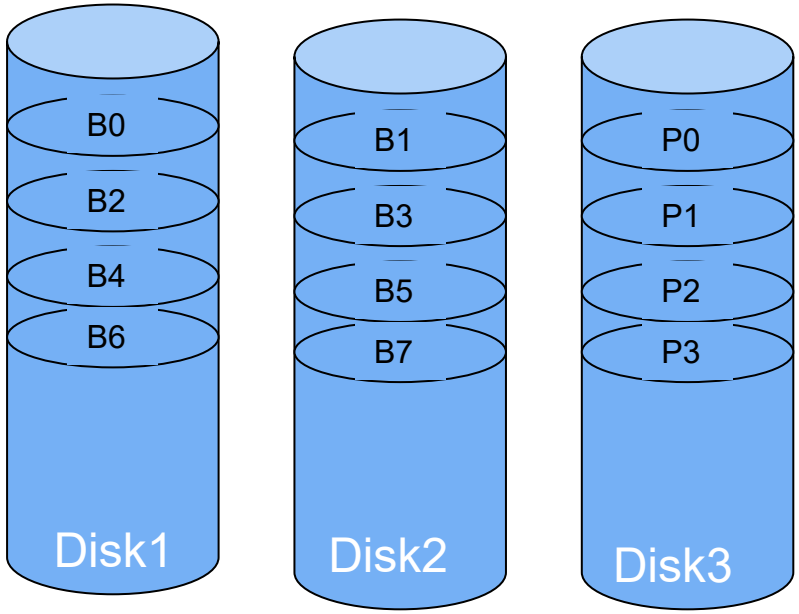
RAID 2 (bit level striping)



- Striping takes place at bit level
- Provides higher transfer rate (double the single disk)
- MTTF reduced by half as in RAID 0
- rarely used

- b means bit
- MTTF = Mean Time To Failure

RAID 3 (Byte level striping)



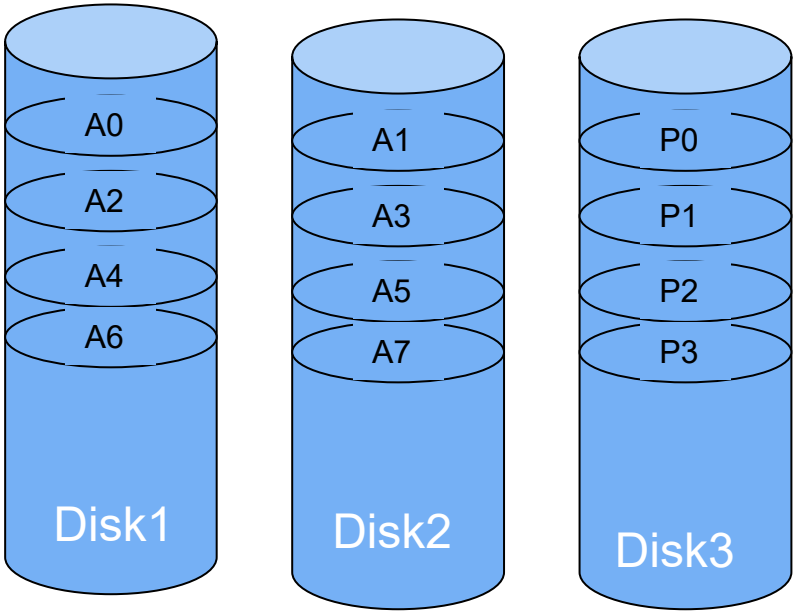
- B means Byte
- P is parity
- MTTF = Mean Time To Failure
- Parity (or check bits) are used for error detection

B0, B1, B2, B3, .. are bytes of data of file

- Striping takes place at byte level
- Rarely used
- Provides higher transfer rate as in RAID 0
- P0 is parity for bytes B0 and B1
- $P_i = B_{2i} \oplus B_{2i+1}$, here $\oplus$ is exclusive-or operator

MTTF increases substantially (1/3 of RAID1 = $MTTF^2/3$), as 1 disk failure can be recovered from the data of the other 2 disks

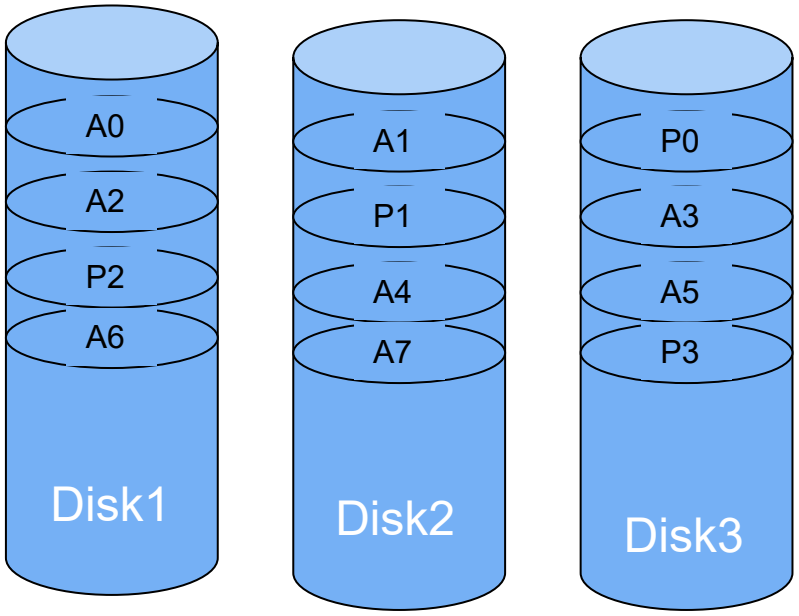
RAID 4 (Block level level striping)



- Striping takes place at block level
- Dedicated disk for parity blocks
- Provides higher throughput but very slow writes. Disk3 has more writes as Parity needs to be updated for every data write.
- MTTF increases substantially (same as RAID3)
- $Pi = A_{2i} \oplus A_{2i+1}$, here $\oplus$ is an exclusive-or operator

- A means Block (4K or 8K bytes of storage)
- P is parity
- MTTF = Mean Time To Failure

RAID 5 with 3 disks (Block level striping)

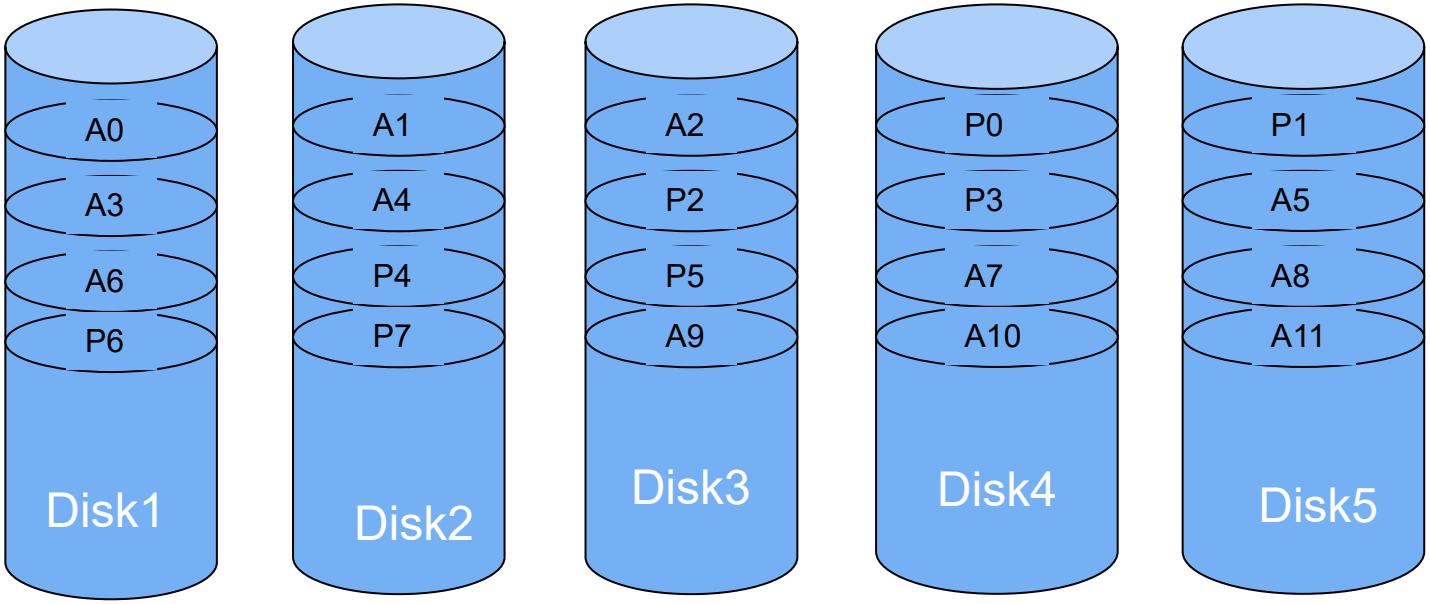


- A0,A1, A2, A3, .. are contiguous blocks of data of a file
- Striping takes place at block level
- Parity blocks are also striped
- Provides higher throughput but slower writes but better than RAID 4 as Parity bits are distributed among all disks and the number of write operations on average equal among all 3 disks.
- MTTF increases substantially (same as RAID3)
- $Pi = A_{2i} \oplus A_{2i+1}$, here $\oplus$ is an exclusive-or operator

•A means Block (4K or 8K bytes of storage)
•P is parity
• MTTF = Mean Time To Failure

RAID 6 (Block level level striping)

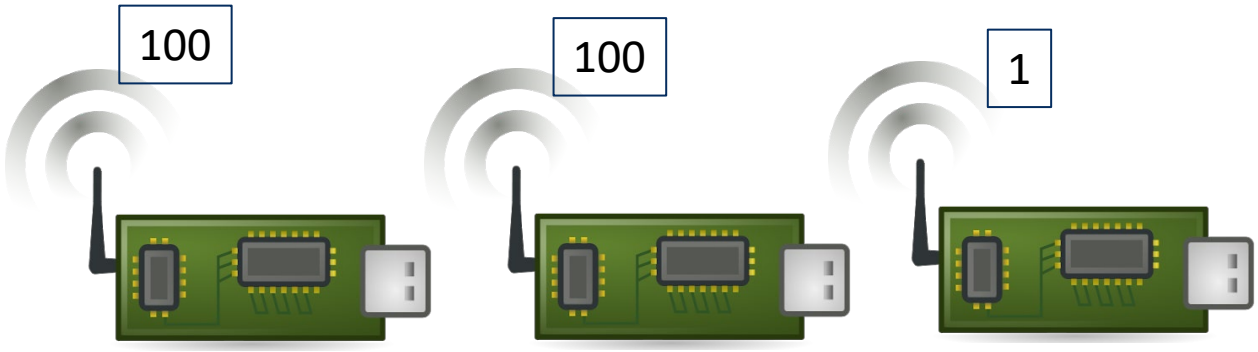
- A means Block (4K or 8K bytes of storage)
- P is parity



- Similar to RAID 5 except two parity bocks used.
- Reliability is of the order of $MTTF^3/10$
- P0 and P1 are parity blocks for blocks A0, A1 and A2. These are computed in such way that any two disk failures can be safe to recover the data.

Fault Tolerance by voting

Use more than one module, voting for higher reliability



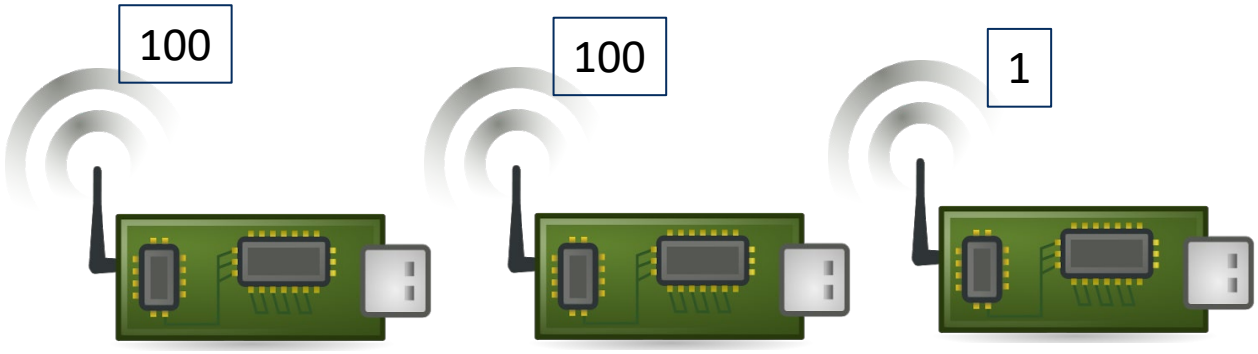
Failvote - Stops if there are no majority agreement

Failfast (voting)- Similar to failvote except the system senses which modules are available and uses the majority of the available modules.

- A 10 module Failfast system continues to operate until the failure of 9 modules where as Failvote stops when 5 modules fail.
- Failfast system has better availability than failvoting (since failvote stops when there is no majority agreement).

Fault Tolerance ...

Failvote needs majority agreement to accept an action (eg, Read/write)

1. 

Accept

2. 100 100 ~~1~~ **Accept**

3. 100 ~~100~~ 1 **Do not accept**

If we start with 10 devices, the system works as long as 6 of them are working. Action is accepted when 6 or more agreeing on the decision.

The moment 5th one fails, system stops as there cannot be 6 devices agreeing

Fault Tolerance ...

In Failfast, we are only concerned of majority among the working ones. We are assuming that we can tell which ones are working. Hence we can continue to operate until 2 working ones and if both agree we can proceed with the action. But if they differ the system stops.

- 0 devices are faulty, we have 10 working and we need at least 6 to agree
- 1 device is faulty, we have 9 working and we need at least 5 to agree
- 2 devices are faulty, we have 8 working and we need at least 5 to agree
- 3 devices are faulty, we have 7 working and we need at least 4 to agree
- 4 devices are faulty, we have 6 working and we need at least 4 to agree
- 5 devices are faulty, we have 5 working and we need at least 3 to agree
- 6 devices are faulty, we have 4 working and we need at least 3 to agree
- 7 devices are faulty, we have 3 working and we need at least 2 to agree
- 8 devices are faulty, we have 2 working and we need both to agree
- 9 devices are faulty, we have 1working and we have to stop as nothing to compare!

Fault Tolerance for disks

Supermodule – Naturally, a system with multiple hard disk drives is expected to function with only one working disk (use voting when multiple disks are working/available, but still work even when only one is available)

Availability of failvote systems

If there are n events, mean time to the first event = m/n

Consider a system with modules each with MTTF of 10 years

Failvoting with 2 devices:

MTTF = 10/2 = 5 years (system fails with 1 device failure)

Failvoting with 3 devices:

MTTF = 10/3 for the first failure + 10/2 for 2nd failure = 8.3 years.

Lower availability for higher reliability (multiple modules agreeing on a value means that value is more likely to be accurate/reliable)

But, but, but…. cannot we have both??

Fault tolerance with repair

With repair of modules: the faulty equipment is repaired with an average time of MTTR (mean time to repair) as soon as a fault is detected (Sometimes MTTR is just time needed to replace)

Typical Values for recent disks:

MTTR = Few hours (assuming we stock spare disks) to 1 Day

MTTF = 750000 hours (~ 86 years) [hard fault]

Probability of a particular module is not available

= MTTR/(MTTF+MTTR)

≅ MTTR/MTTF if MTTF >> MTTR

Fault tolerance of a supermodule with repair

Probability that (n-1) modules are unavailable, $P_{n-1} = \left(\frac{MTTR}{MTTF}\right)^{n-1}$

Probability that a particular ith module fails, $P_f = \left(\frac{1}{MTTF}\right)$

Probability that the system fails with a particular ith module failing last =

$$P_f * P_{n-1} = \left(\frac{1}{MTTF}\right)\left(\frac{MTTR}{MTTF}\right)^{n-1}$$

Probability that a supermodule fails due to any one of the n modules failing last, when other (n-1) modules are unavailable $= \left(\frac{n}{MTTF}\right)\left(\frac{MTTR}{MTTF}\right)^{n-1}$

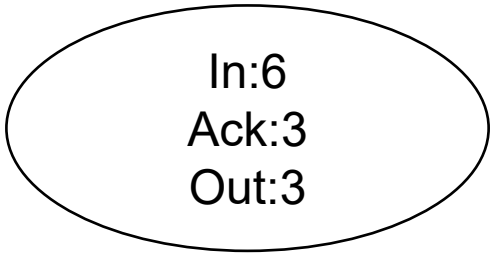
What will this value for failvote and for failfast?

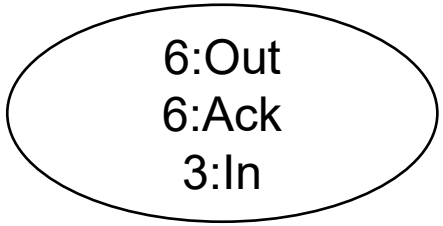
Communication reliability

Out = #messages sent

In= #messages received

Ack = #acknowledgements



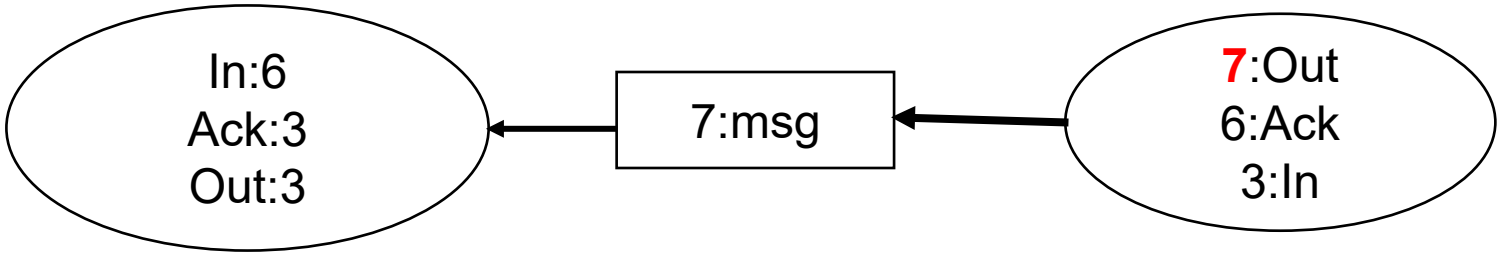


Communication reliability

Out = #messages sent

In= #messages received

Ack = #acknowledgements



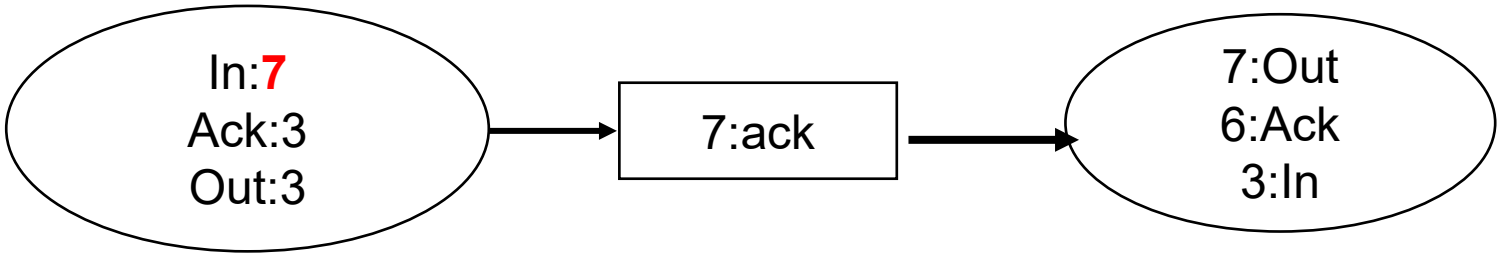
Communication reliability

Reliable message passing

Out = #messages sent

In= #messages received

Ack = #acknowledgements



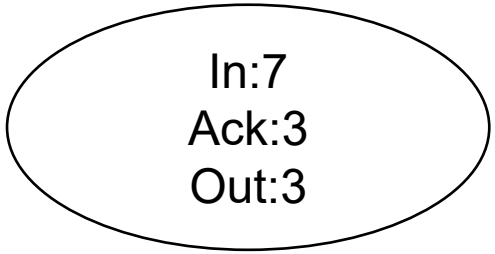
Communication reliability

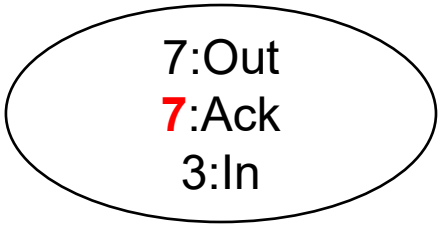
Reliable message passing

Out = #messages sent

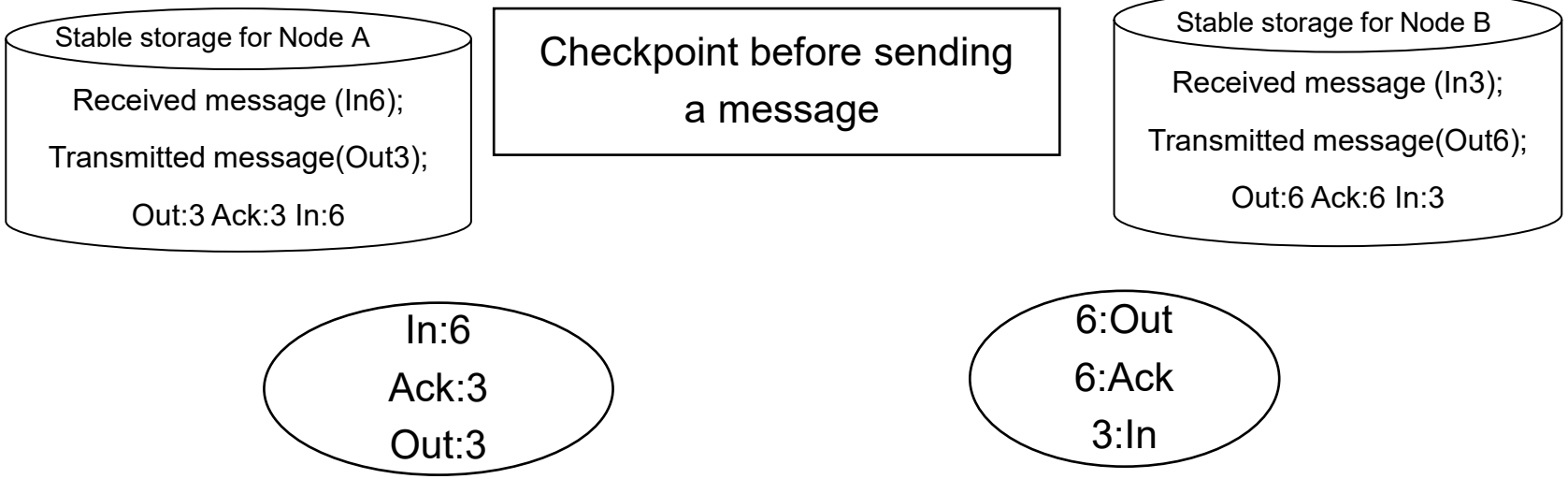
In= #messages received

Ack = #acknowledgements

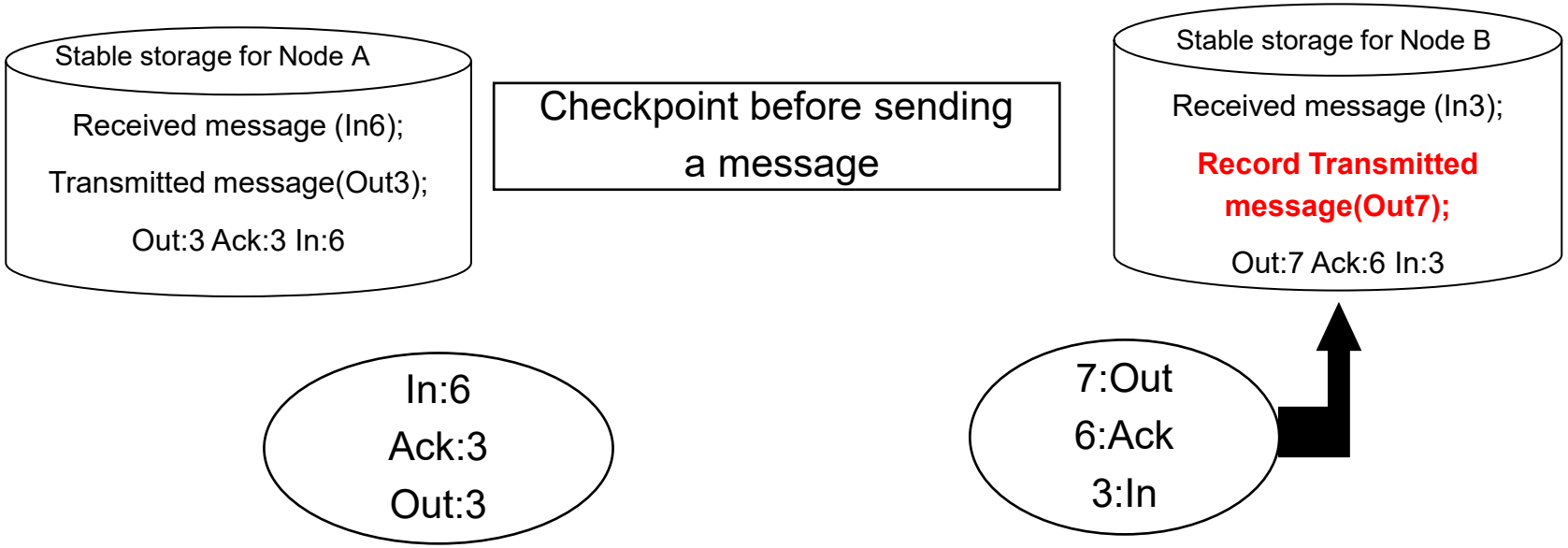




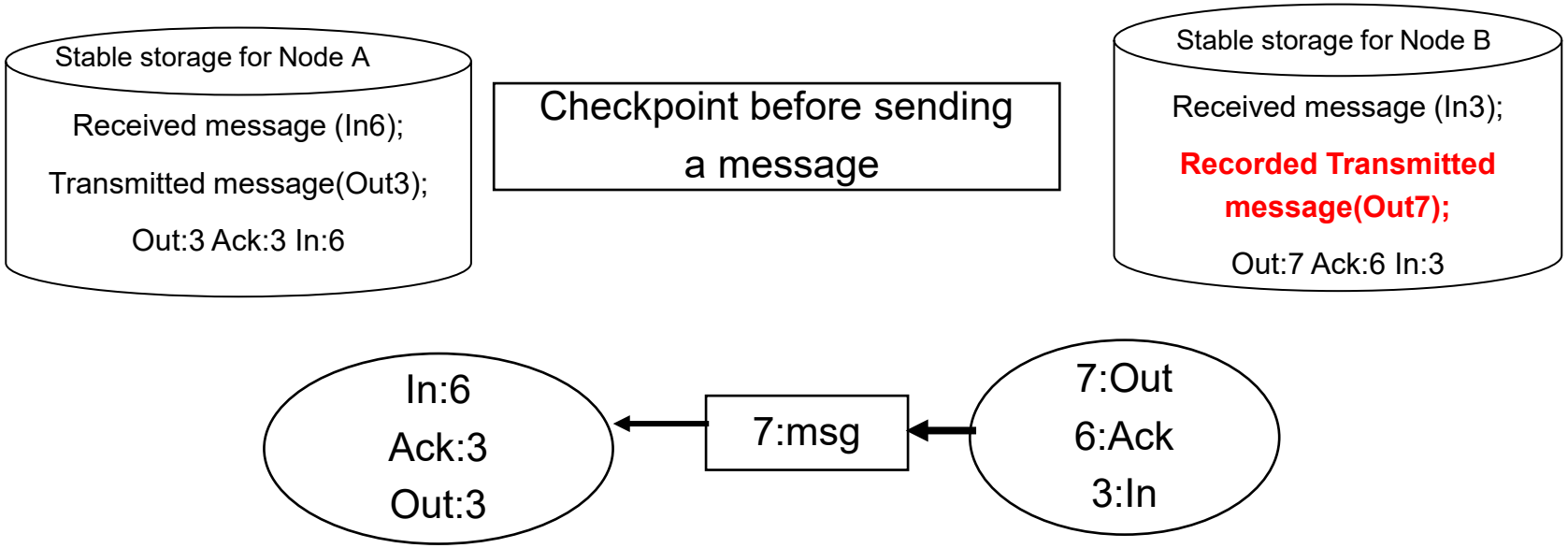
Communication reliability



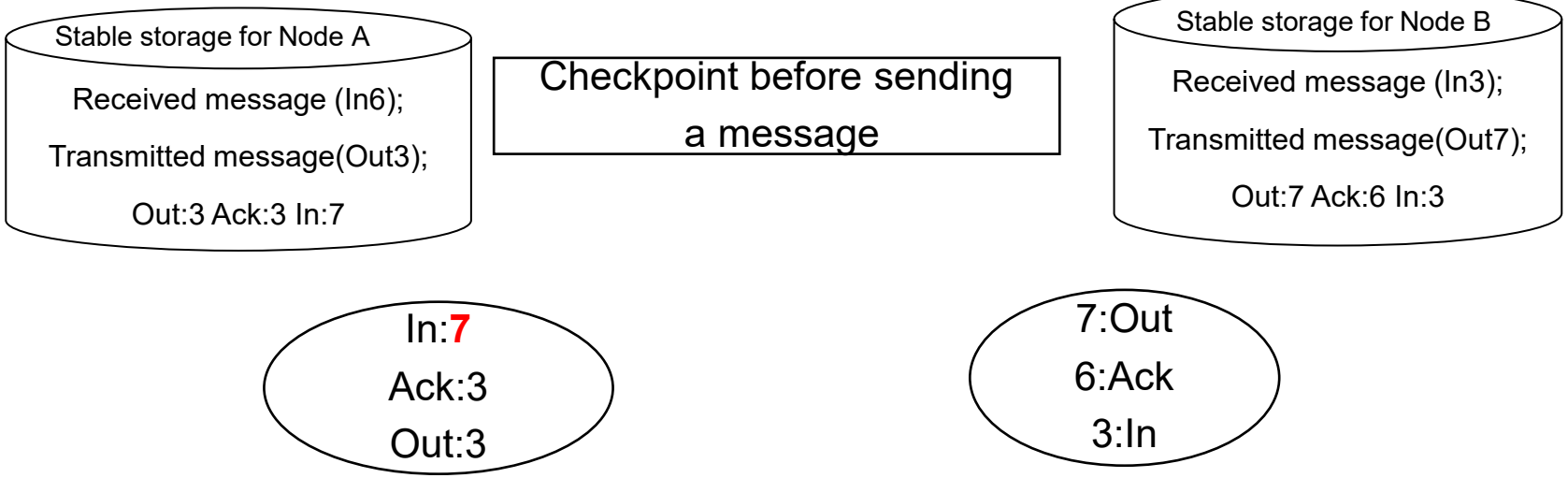
Communication reliability



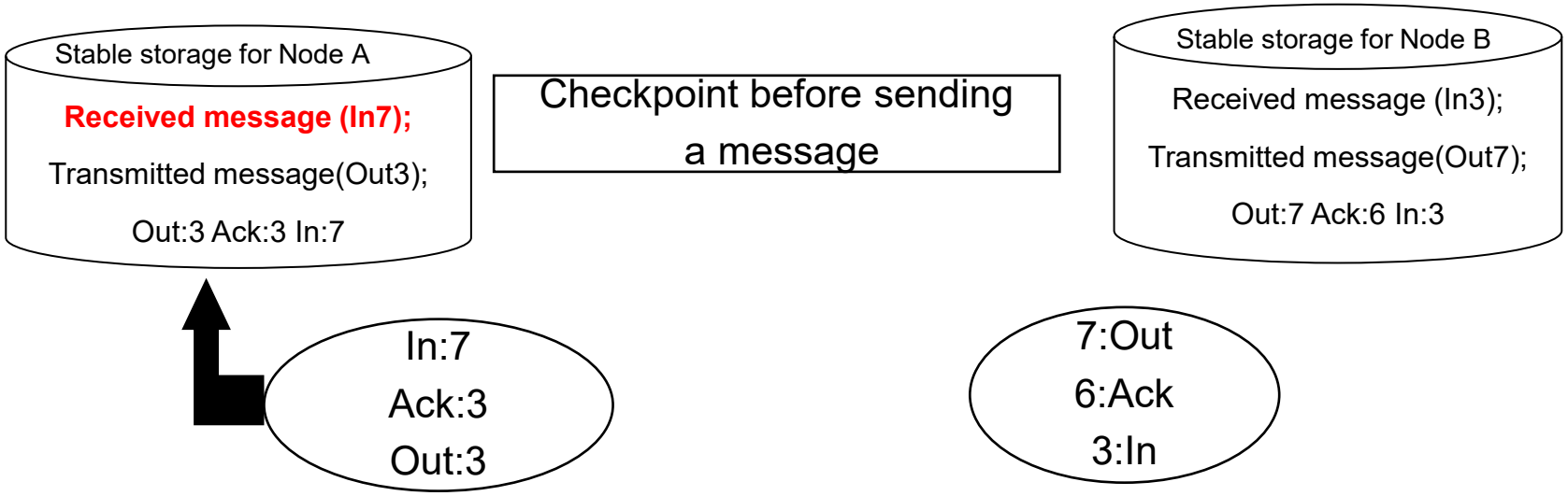
Communication reliability



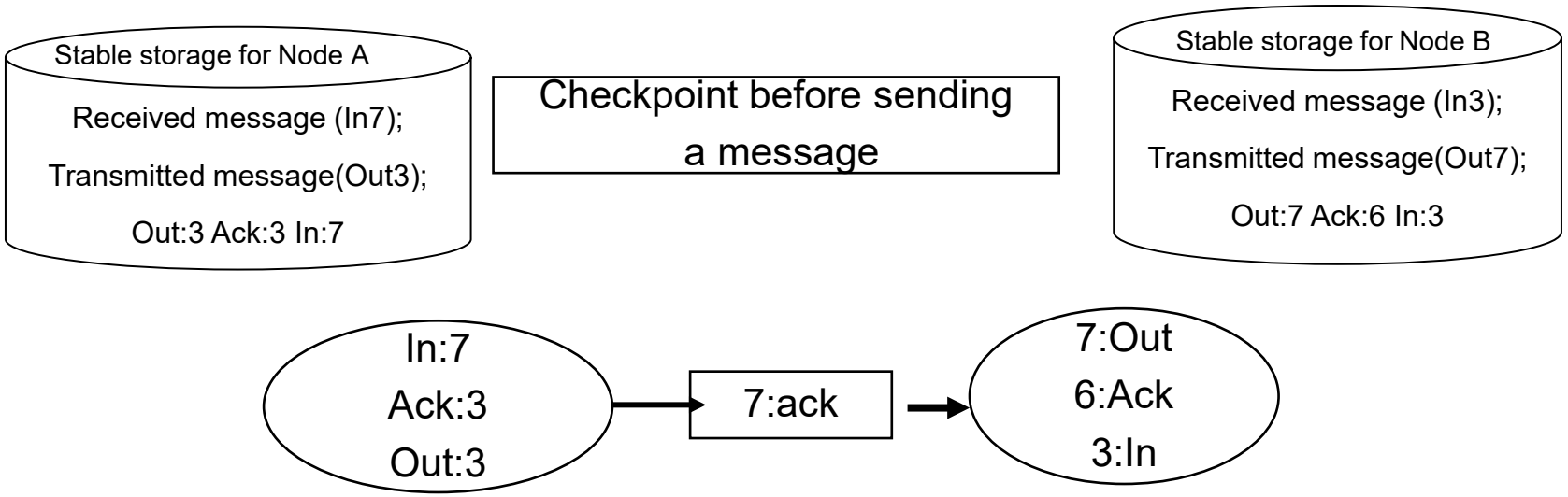
Communication reliability



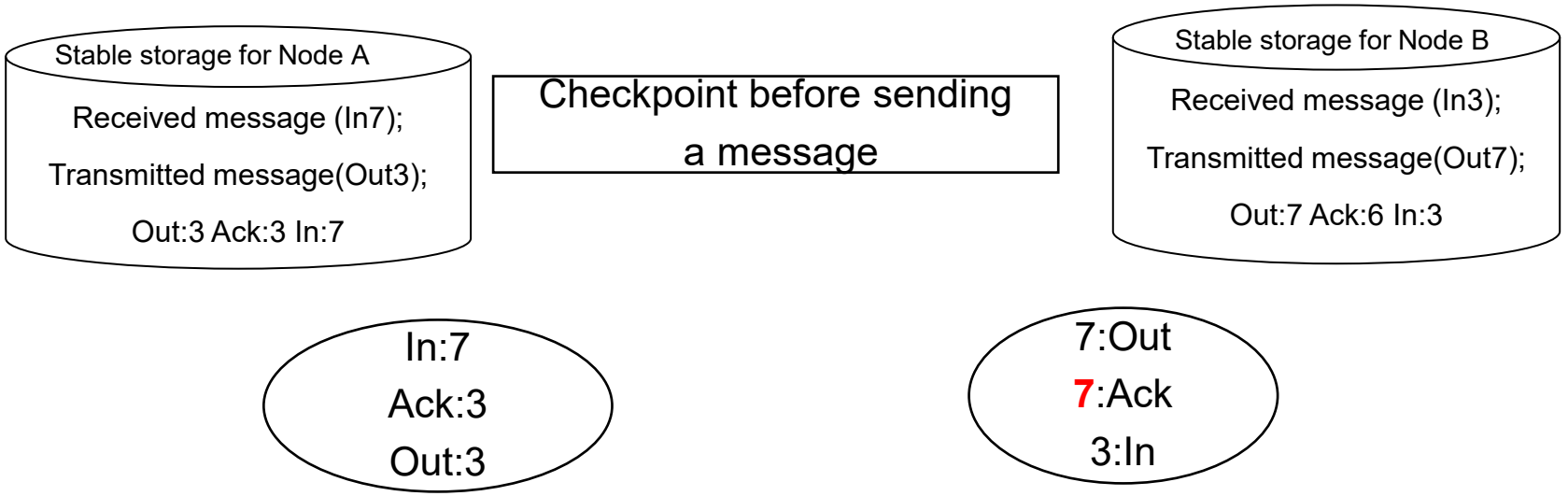
Communication reliability



Communication reliability



Communication reliability



Communication reliability