FOTOS –
Fail Over and Take Over System

Lennart Öhman

Sjöland & Thyselius

lennart.ohman@st.se

Sjöland&Thyselius

Experts in Advanced Technology

Erlang/OTP selling point

- **It is very easy to create multi node systems with load balancing and failover in Erlang/OTP!**

  or…

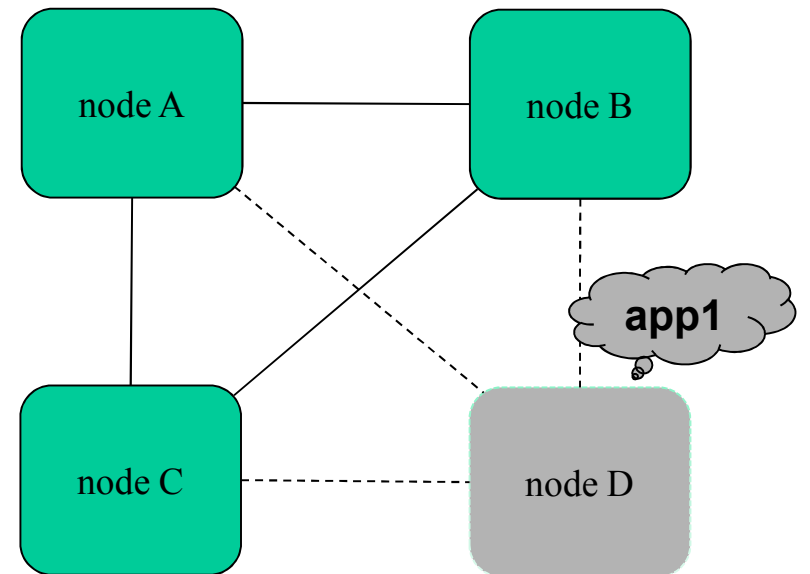**Sjöland&Thyselius**
Experts in Advanced Technology

It is very easy to create…

```
[{kernel,
  [{distributed,
    [{app1,
      [nodeD,
       nodeC,
       nodeB,
       nodeA]}]
  }]}].
```


Sjöland&Thyselius
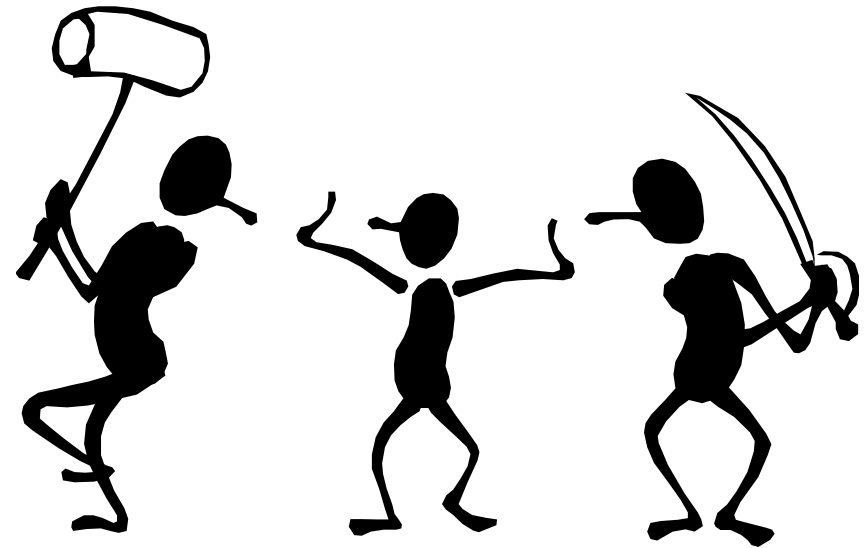
Experts in Advanced Technology
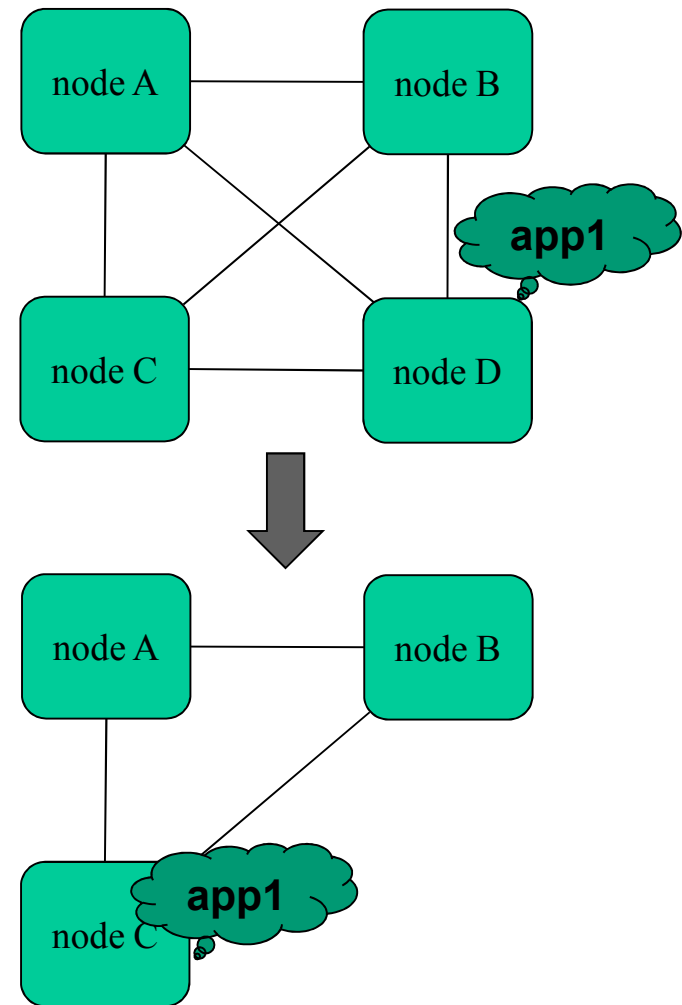
What is failover?

- **The ability of peer nodes to detect the failure of a node and resume its responsibilities.**
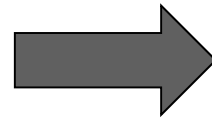
Failover…

- **Detect failing node, and**

- **Decide who resumes which responsibilities.**

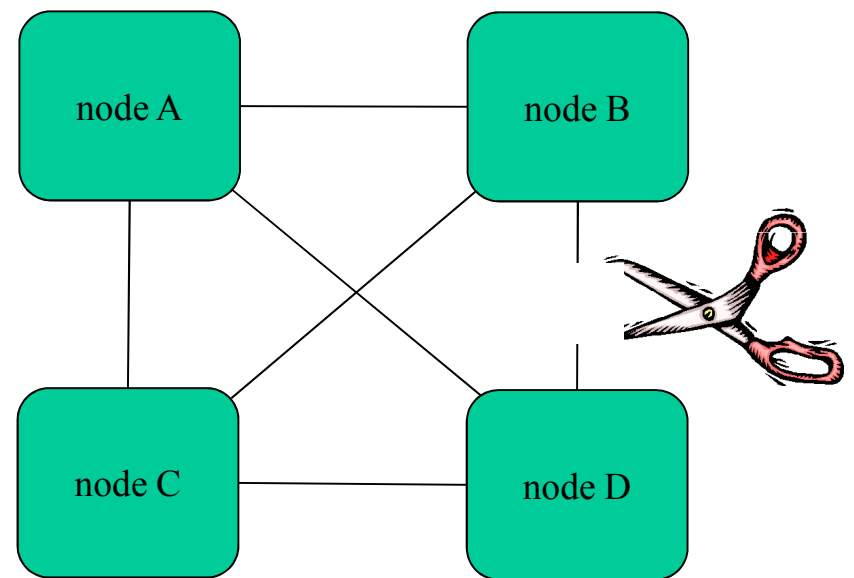# Failover…

- **Decide who resumes which responsibilities.**

→

All nodes must agree on which nodes are in fact part of the network!

Sjöland&Thyselius
Experts in Advanced Technology

It is very easy, or…

- **The distributed application controller does not handle network failures.**

FOTOS

- **Monitoring of the network and maintain a network of operational nodes.**

    - **Without having to elect a leader.**

- **A simple database with vector clock controlled updates.**

Sjöland&Thyselius
Experts in Advanced Technology

Assumptions

- **All nodes are loyal and follows "the rules".**

- **Messages between two nodes are delivered in the same order as sent.**

- **External means of detecting a partitioned network.**

Sjöland&Thyselius
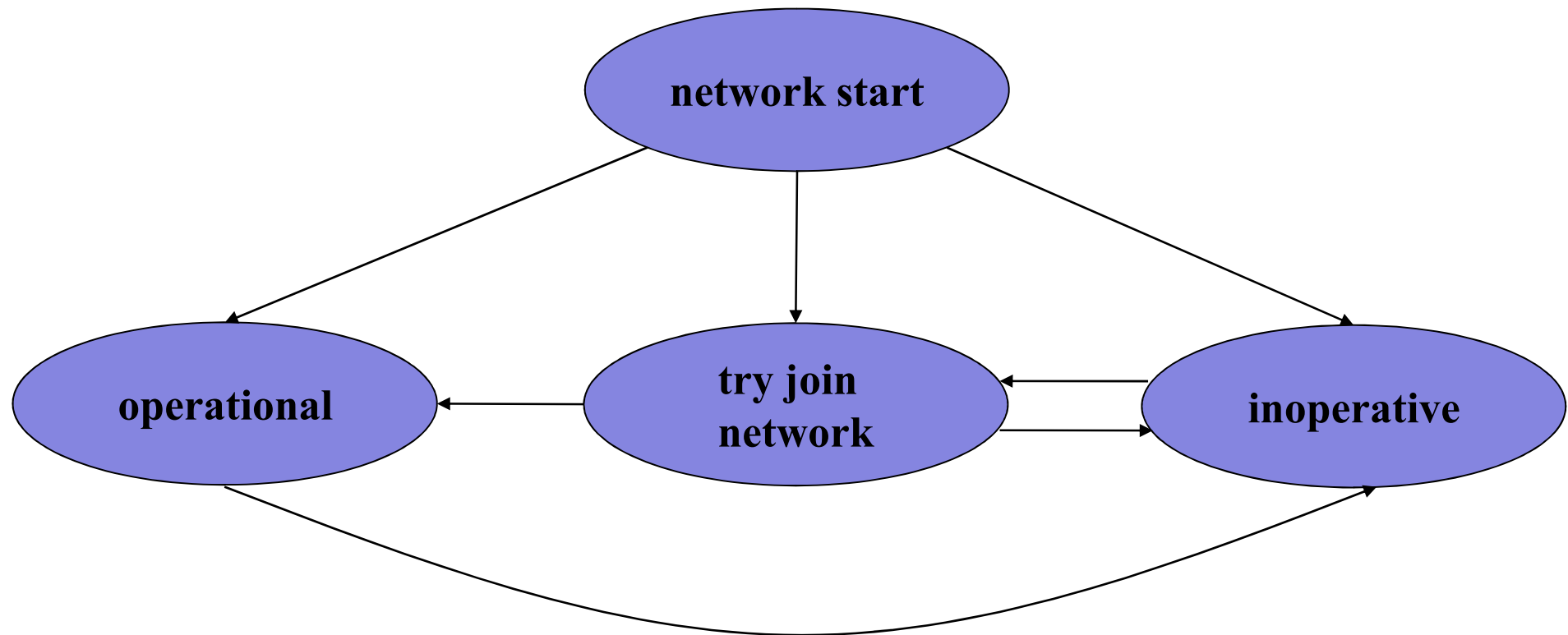
Experts in Advanced Technology

Monitoring of the network

- ## A network of operational nodes

  - **A fully connected mesh of Erlang nodes.**

  - **Only nodes that can "see" all other participating nodes may participate.**

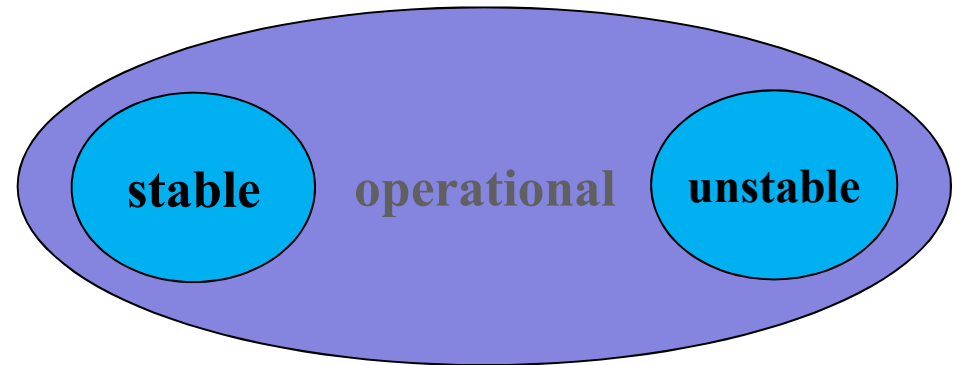**Sjöland&Thyselius**

Experts in Advanced Technology
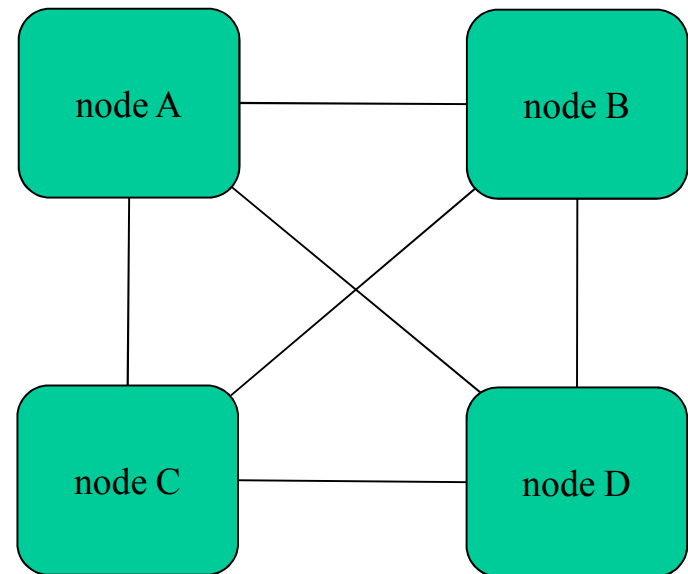
The node state machine

The node state machine

- **An operational node may be either stable or unstable.**

Network start

- **The first thing a newly started node does is to try to start a network of operational nodes.**

  - It contacts all known nodes to see if they want to help create a network of operational nodes.
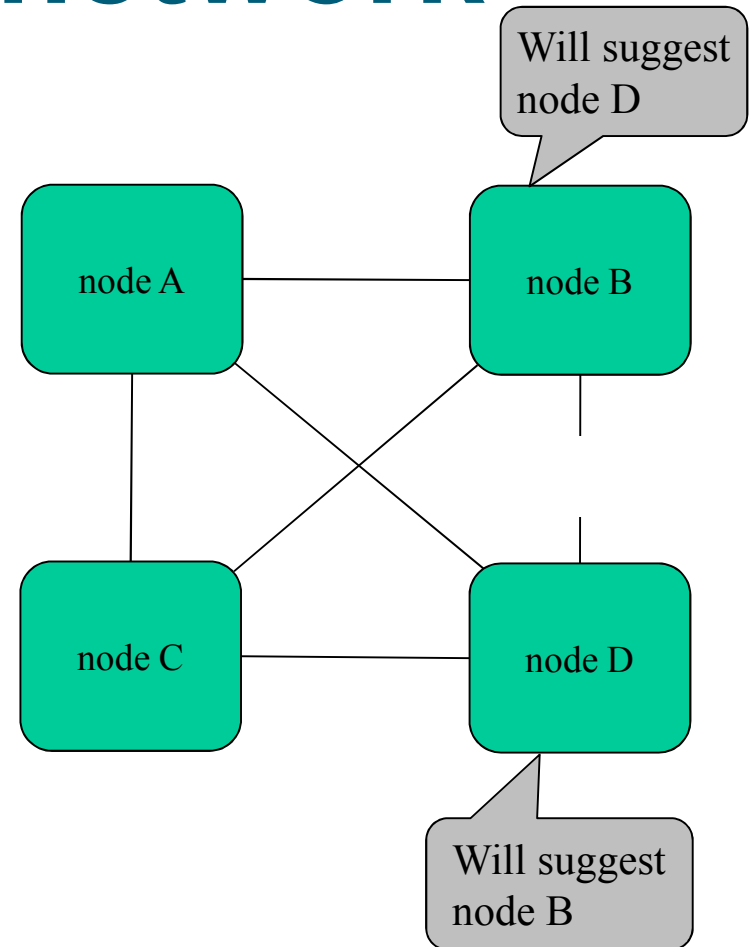
Joining a network of operational nodes

- **A newly started node tries to join a network if it "sees" all nodes.**

  - An operational node may only object to a JOIN REQUEST if it does not include exactly all nodes that are operational, or

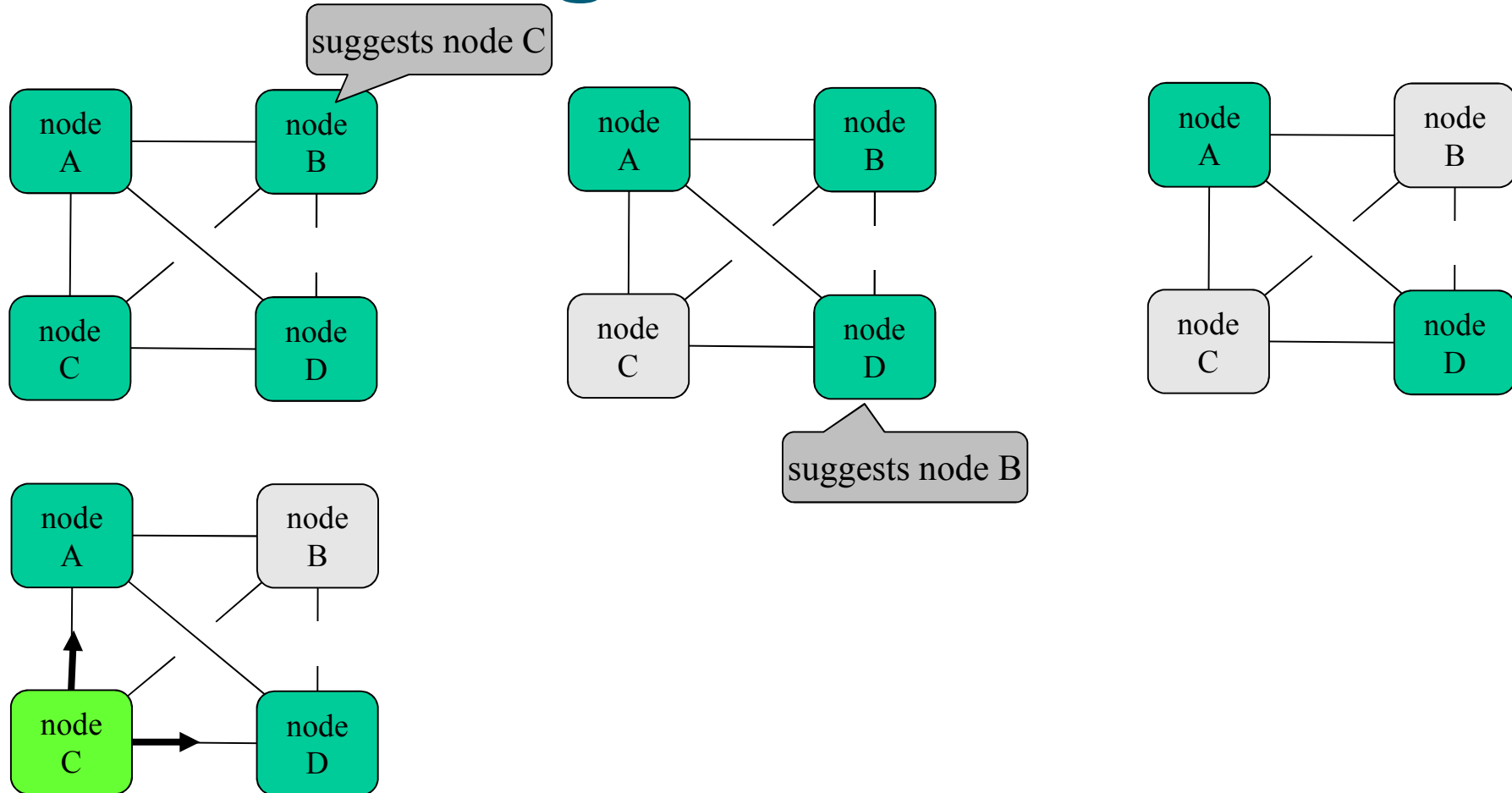 - If it is participating in another procedure currently.

JOIN REQUEST

Sjöland&Thyselius
Experts in Advanced Technology
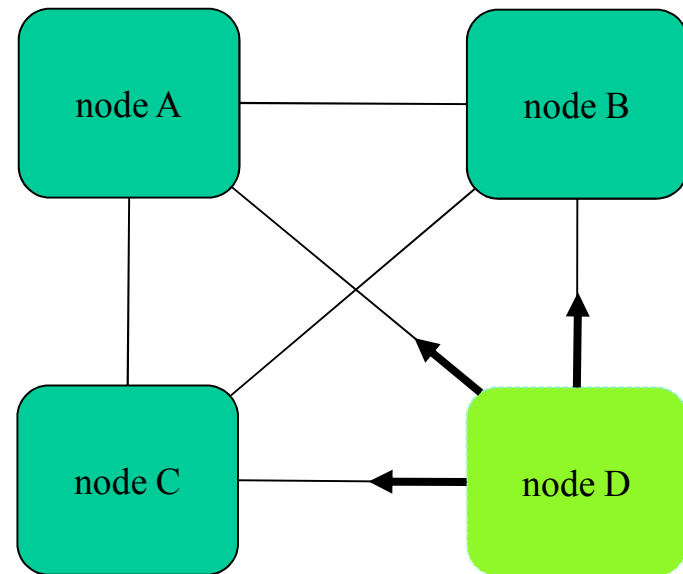
Monitoring of the network

- **A node that detects a peer node failure suggests its removal.**

    - **A node may not be removed unless all other nodes (except the failed node) agrees.**

Will suggest node D

Will suggest node B

node A

node B

node C

node D

Sjöland&Thyselius

Experts in Advanced Technology

Monitoring of the network

Single failure optimization

- **A node may only run one procedure at a time.**

    - **Removing or joining one node at a time.**

- **If multiple simultaneous failures, no guarantee the result is the most optimal possible network.**

Network unstable

- **As soon as a node learns of that a participating node is not visible from one or several nodes – the network is deemed unstable.**

- In first "person" ('DOWN' message).

- "Rumor" from another node that wishes to remove the now failed node from the network of operational nodes.

Sjöland&Thyselius
Experts in Advanced Technology

There is no magic!

■ **You must still make fail-over and load balancing decisions!**

    ■ **Based on stable/unstable.**

    ■ **Based on changes of the set of participating nodes.**

You must provide a "holy grail"

- **In case of a completely partitioned network – a "holy grail" is needed to decide which network is the "real network".**

Sjöland&Thyselius

Experts in Advanced Technology
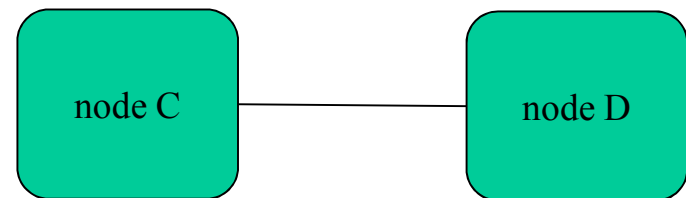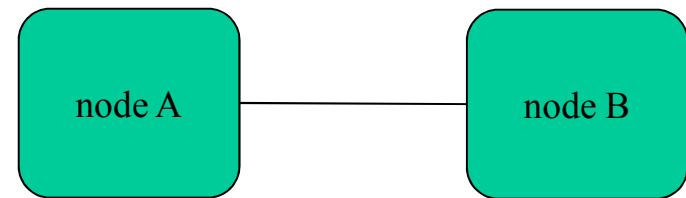
Some spin-offs

- **2 Phase Commit**

  - **2 phase commit with rendezvous**

- **A Vector Clock**

FOTOS summary

- **Subscribe to stability status.**

- **Subscribe to changes in the set of operational nodes.**

- **Optimized for single failure.**

- **A simple database with vector clock protected updates.**

Sjöland&Thyselius

Experts in Advanced Technology