Faults

CGS - Hugo Miranda

2021

Ciências ULisboa

Faculdade de Ciências da Universidade de Lisboa

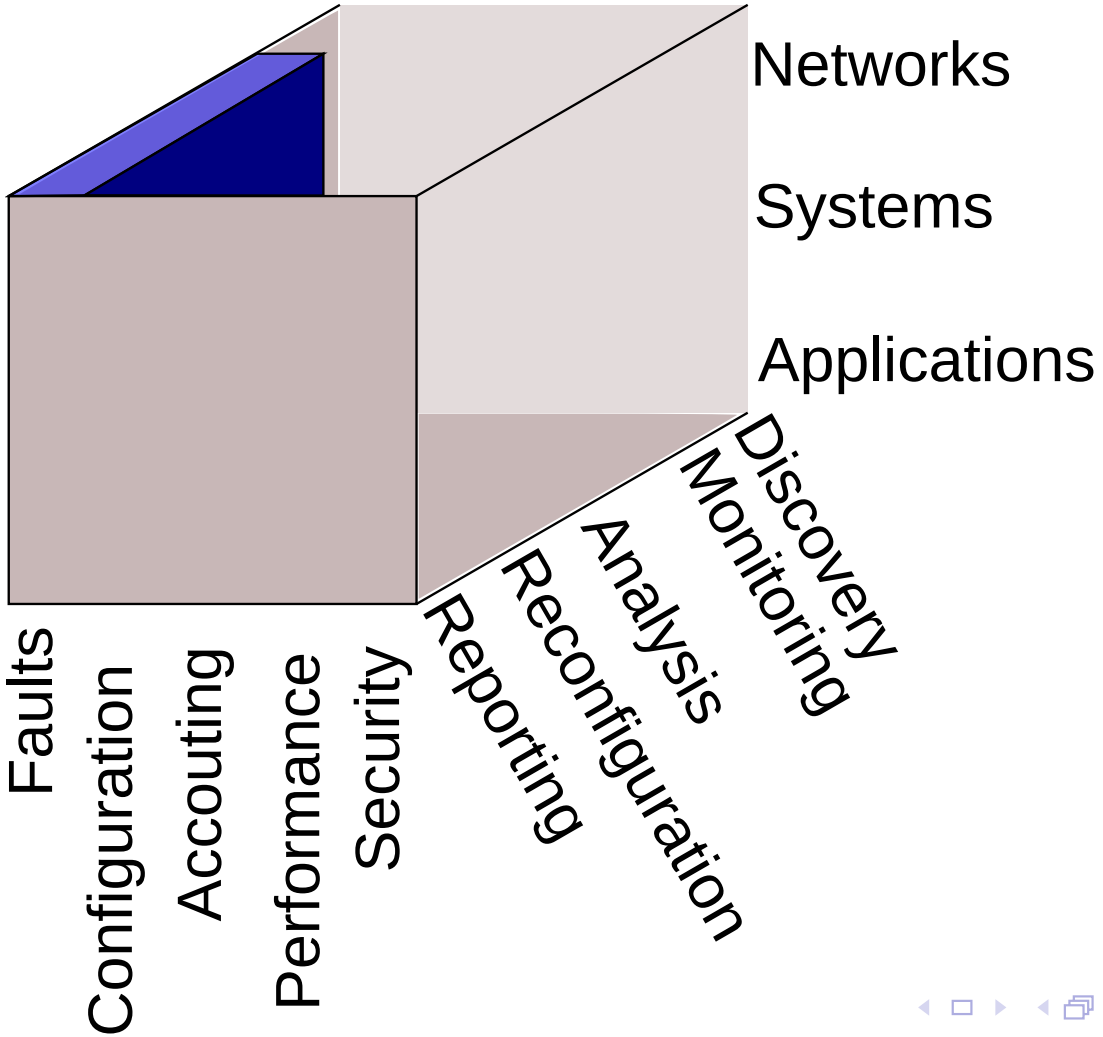
Faults

A distributed system is one that prevents you from working because of the failure of a machine that you had never heard of.
Leslie Lamport

The impact of faults

- How users "judge" how good a team is
  - There's always a problem
  - Users don't care about the *complexity/time to fix* ratio

Faults of What?



Learn → Identify → Resolve

- Monitoring infrastructure reports Events
- One or more events are translated into Symptoms
- Symptoms point to Causes
- Causes are fixed

Events

- The information arriving to the monitoring infrastructure
- Delivered by the system
  - Users $\subset$ System
- "No News" $\not\Rightarrow$ "Good News"
  - "No News" must be converted in "Bad News"
  - E.g. ping

Symptoms

- The expression of a problem
  - Typically expressed by events
  - Not necessarily 1 $\Longleftrightarrow$ 1
- Example:

  Symptom Server down
  Event Ping failed
  Event Webserver cannot reach the database
  Event Phone call from user
  - complaining to be unable to read client record

Causes

- The reason why symptom(s) exist
  - The problem to be addressed
  - Diagnostic: identifying the cause from symptoms

Classes of Causes

Permanent

- Infrastructure failure
  - Power outage
  - Overheating
- Component failure
  - Hardware, software, user
- Configuration problem
  - Wrong configuration
  - Software bugs
  - Wrong requests

Classes of Causes

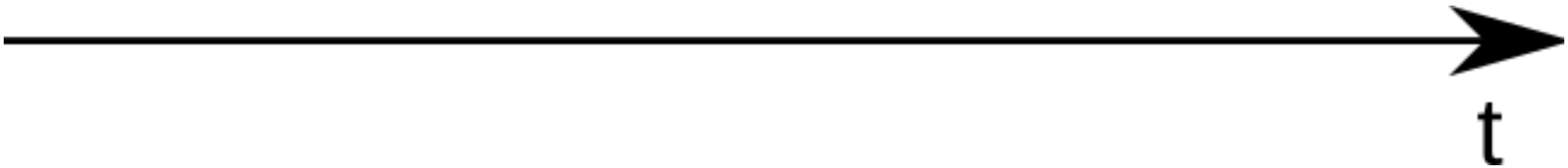
Transient

- Component restarting
- Component or system reconfiguration
  - Route changes
  - Request redirection

Making a Diagnostic

- Given a set of symptoms $s_0,\ldots,s_n$
  - observed on moments $ts_0,\ldots,ts_n$
- Identify a set of root causes $c_0,\ldots,c_m$
  - that occurred on moments $tc_0,\ldots tc_m$
- Such that tc_i happens before ts_j, $\forall$ s_j associated to c_i

A Timeline



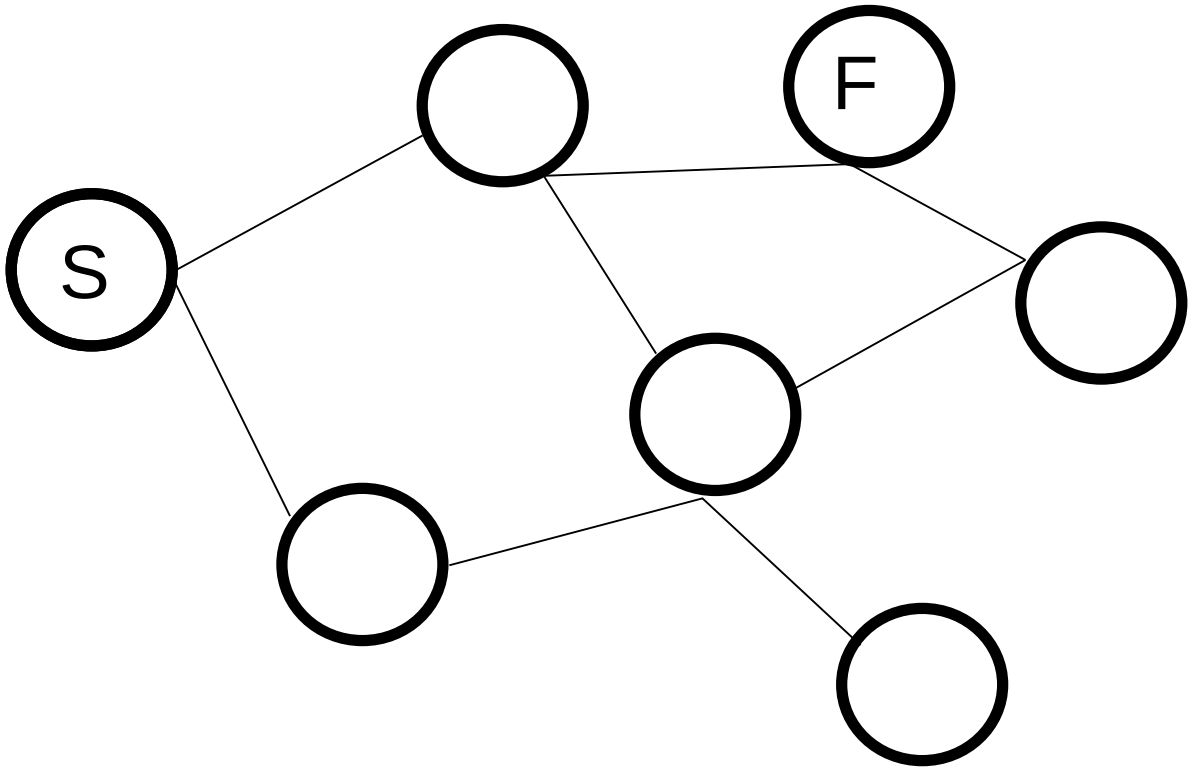
Diagnostic Approaches Library

- No single approach fits all
- Sometimes cause is obvious
- Frameworks can/should help

Topology Analysis

- Graph shows dependencies/connectivity status
- Narrow down possibilities until causes can be found

Topology Analysis



Library of Rules

- Set of if/then clauses
- Then returns either root causes or other symptoms

Example

- If $s_0 \wedge s_1 \rightarrow s_{10}$
- If $(s_2 \wedge s_{10}) \vee (s_9 \wedge s_5) \rightarrow c_5$

Decision Tree

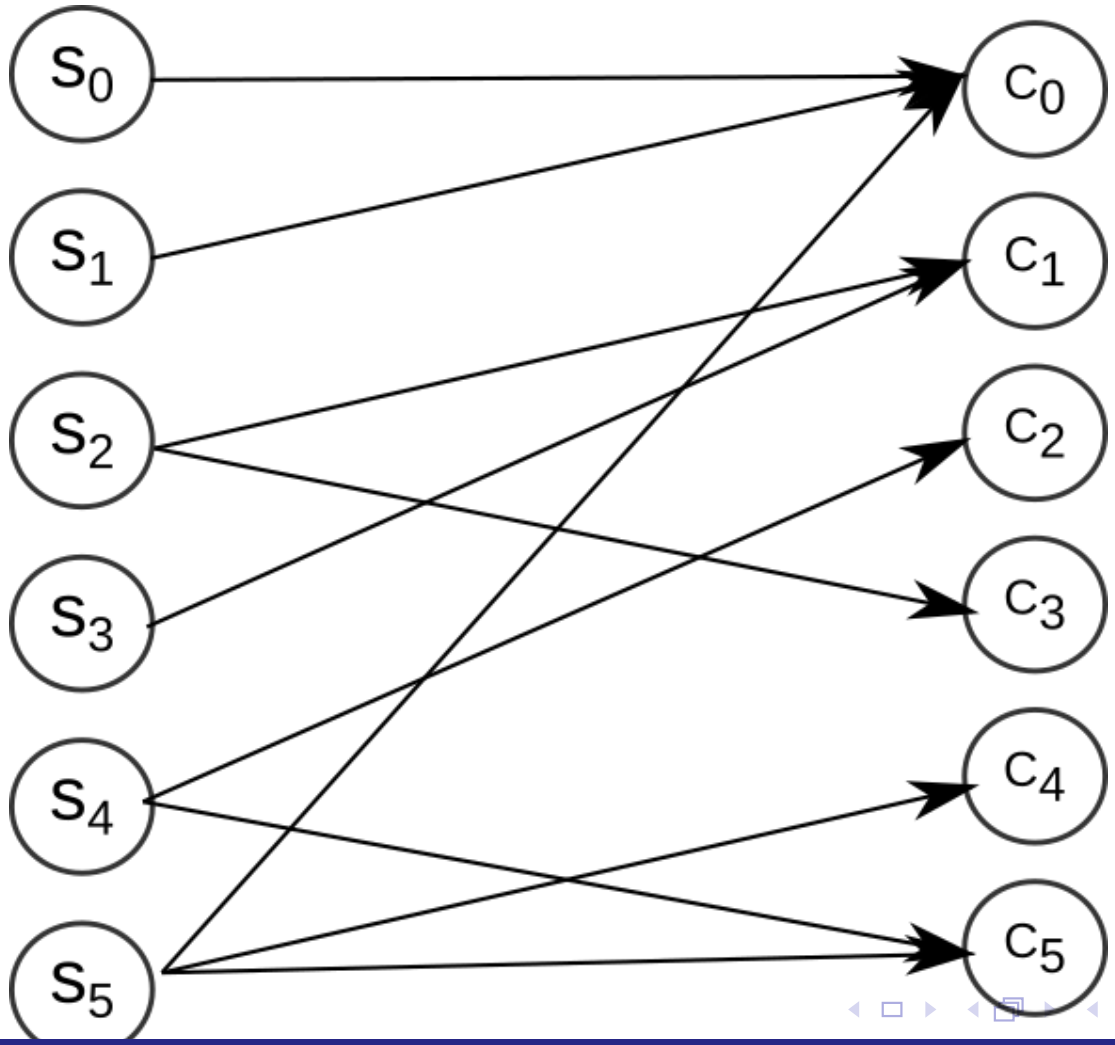
- Workflow tree
- Decisions on symptoms
- Root causes are leafs

Decision Tree

Dependency Graph

- Symptoms on the left
- Root causes on the right
- 1 symptom $\rightarrow$ N causes
- 1 root cause $\rightarrow$ N symptoms
- Typically represented as a bipartite graph

Bipartite graph



Codebook

- Projection of bipartite graph on a table

Rows symptoms

Columns Root causes

Cells 1 if the link exists/0 otherwise

- Hamming Distance helps to automate diagnostic

Codebook

Example

	c_0	c_1	c_2	c_3	c_4	c_5
s_0	1	0	0	0	0	0
s_1	1	0	0	0	0	0
s_2	0	1	0	1	0	0
s_3	0	1	0	0	0	0
s_4	0	0	1	0	0	1
s_5	1	0	0	0	1	1

Knowledge Base

- Local
  - Ticket Troubling Service (TTS)
  - Incident reports
  - Documentation produced by the team
    - If it exists
- Vendors
  - Web service mapping error codes on problems and solutions
- Web search engines
  - Google the error code or symptoms
  - Developers forum
  - ...

Case Based Reasoning

- If the symptoms were observed in the past
  - the cause is likely the same
- TTS/Incident reports
  - Knowledge Base

Redundancy

cold standby boot when necessary

warm standby periodic updates

- E.g. DNS servers, databases

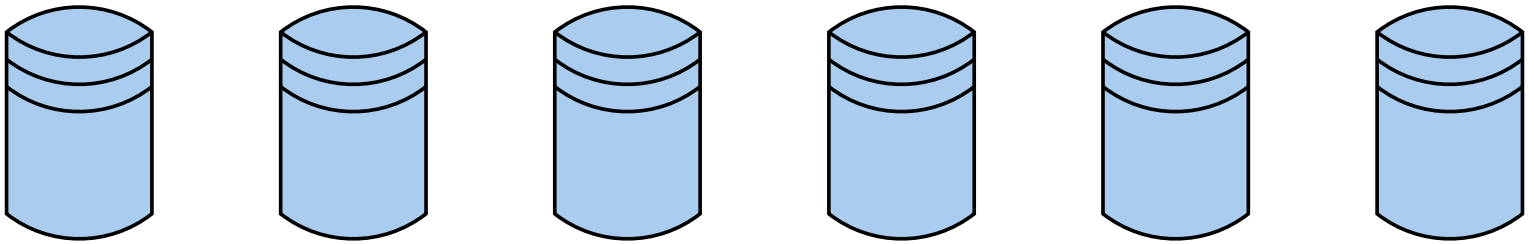
hot standby permanently updated

- E.g. RAID

RAID Levels

- 0 Merge hard drives on a single virtual one. No redundancy.
- 1 Mirror pairs of hard drives. 50% space for redundancy.
- 5 Recovers from the failure of one drive from the group. 1/N for redundancy
- 6 Recovers from the failure of two drives of the group. 2/N for redundancy
- 10 1+0

RAID Levels



Prevention

- Automatic reboots
- Redundancy management
  - centralized on a single master node
  - distributed

Conclusions

- Faults are the most visible activity of IT administration
  - Most users believe it is (almost) the only one
- Infrastructure doesn't choose the adequate moment for failing
  - That's why "Prevention status" exist
- Redundancy is key to mitigate problems
- Many things can happen and symptoms can be confusing