Architecting for Reliability – Detection and Recovery Mechanisms

Robert S. Hanmer
LCP Architecture
Alcatel-Lucent
hanmer@alcatel-lucent.com

Veena B. Mendiratta
Network Performance & Reliability
Bell Labs, Alcatel-Lucent
veena@alcatel-lucent.com

International Symposium on Software Reliability Engineering (ISSRE) **November 2009**

Abstract

High availability, in the form of continuous service availability, is achieved in telecommunications systems by implementing extensive and effective error detection and recovery mechanisms with high coverage. Escalating detection and recovery mechanisms start with those that can deal with targeted errors with very low latency and impact can escalate to actions with longer recovery times and broader system impact. In this work we extend previous studies, combining Markov models for escalating detection and recovery into a unified model. The results of this model show that a unified view, such as this, produces results that more closely align with our experience. It also non-intuitively shows that detection and recovery coverage should be balanced. Designers can use these models to evaluate alternative schemes for error detection and recovery to achieve a given system/service availability target.

Escalation is a technique where the system attempts a local error recovery action, followed by more severe and wide-ranging actions if the local actions do not succeed. When the fault and error are not covered the system might enter a failed state that requires human intervention to recover. Whenever the system requires human intervention the period of unavailability is increased and is generally a long outage.

We combined the escalating detection and recovery Markov models from the previous papers into a comprehensive Markov model to provide insight into how escalating fault recovery mechanisms can be used optimally to achieve high system availability. The recovery model begins with three levels of detection followed by three levels of recovery. If the earlier levels of detection and recovery detect and recover from the error the system returns to a working state more quickly and the later levels of detection and recovery escalation are avoided thus resulting in higher system availability.

The results of the combined detection and recovery escalation model exhibits more of the expected behaviour in terms of system availability and recovery than the individual models showed. For example, the results of the combined model show the range of availability that we expect from varying coverage factors. The model shows that both detection and recovery coverage factors must be high to achieve high availability. Building a system with a low coverage factor for either detection or recovery will not result in a system with high levels of availability, which is counter to conventional wisdom. Future work includes analyzing additional scenarios, as well as considering the cases when the detection and recovery rates and coverage factors vary independently based upon the type of fault and error that is present.

The Problem

Telecom systems in the past were custom built and designed for high availability.

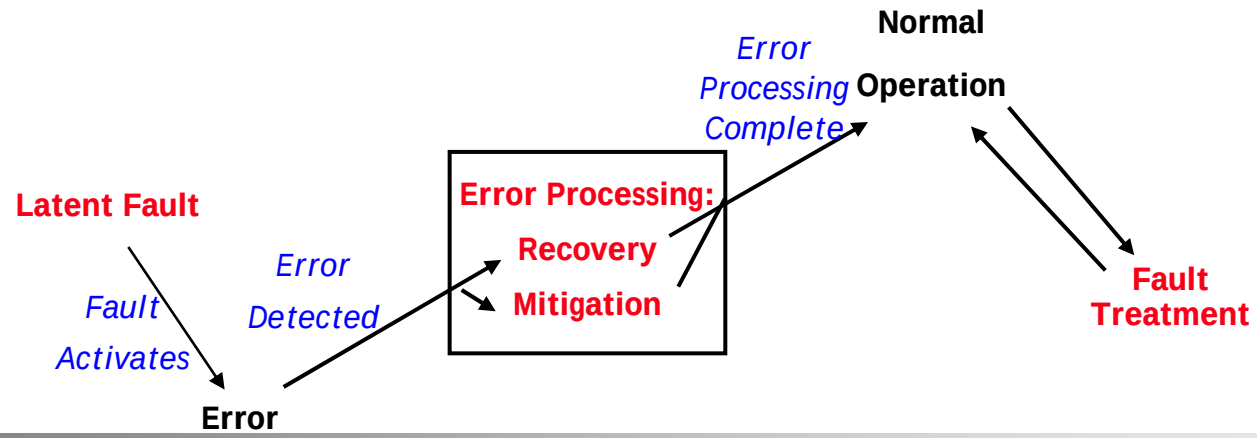
Modern networks utilize COTS components that may not be designed for high availability.

Challenge – determining the detection and recovery techniques to implement, the sequence of implementation, and the coverage factor.

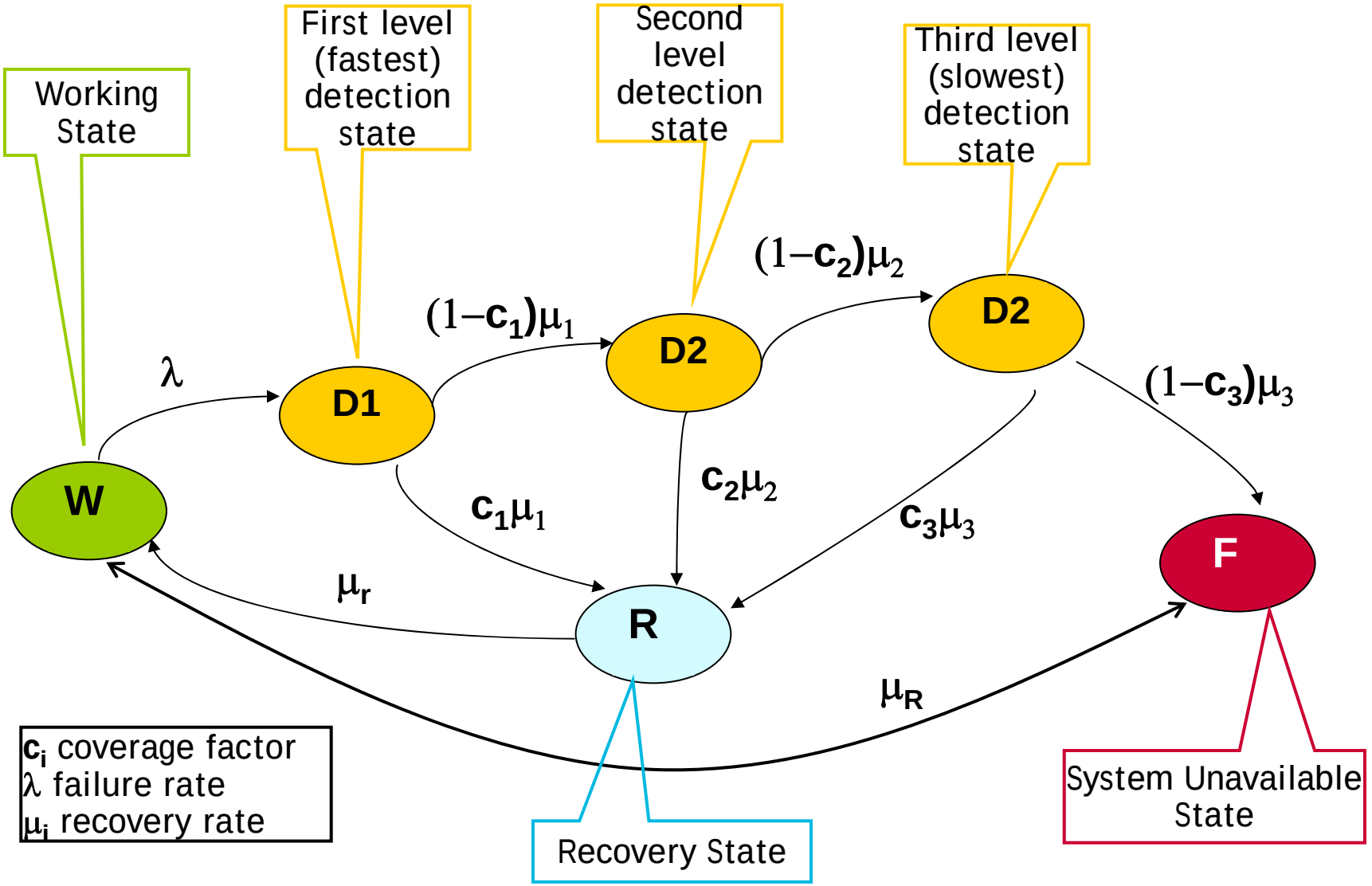
Detection: different errors require different techniques for fast detection; techniques can be nested in scope and chronology (DSN WADS 2008).

Recovery: escalating recovery techniques with varying duration and coverage.

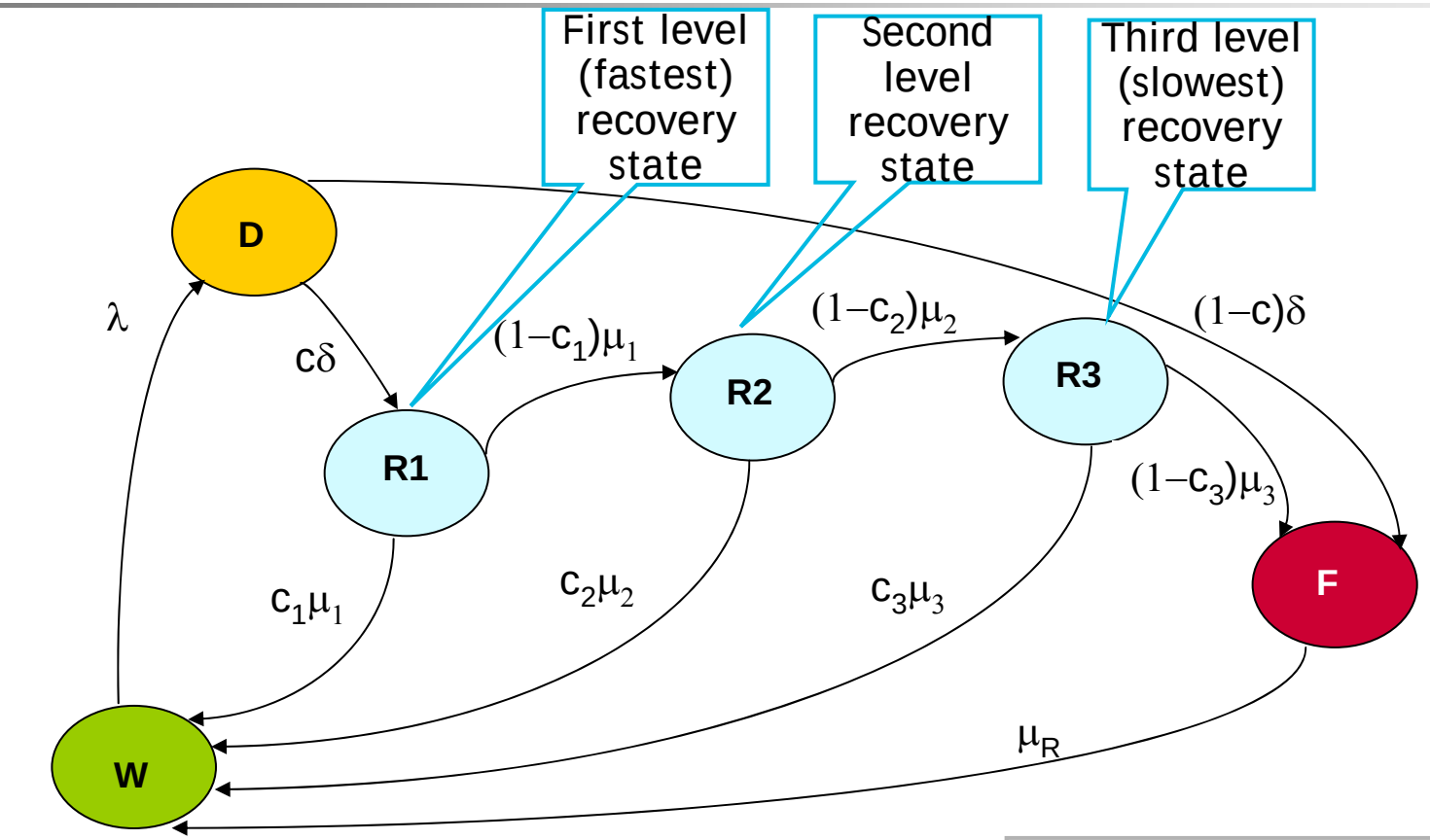
Solution – developed simple Markov model to compare alternatives.



Reference Error Detection Model



Reference Escalating Recovery Model

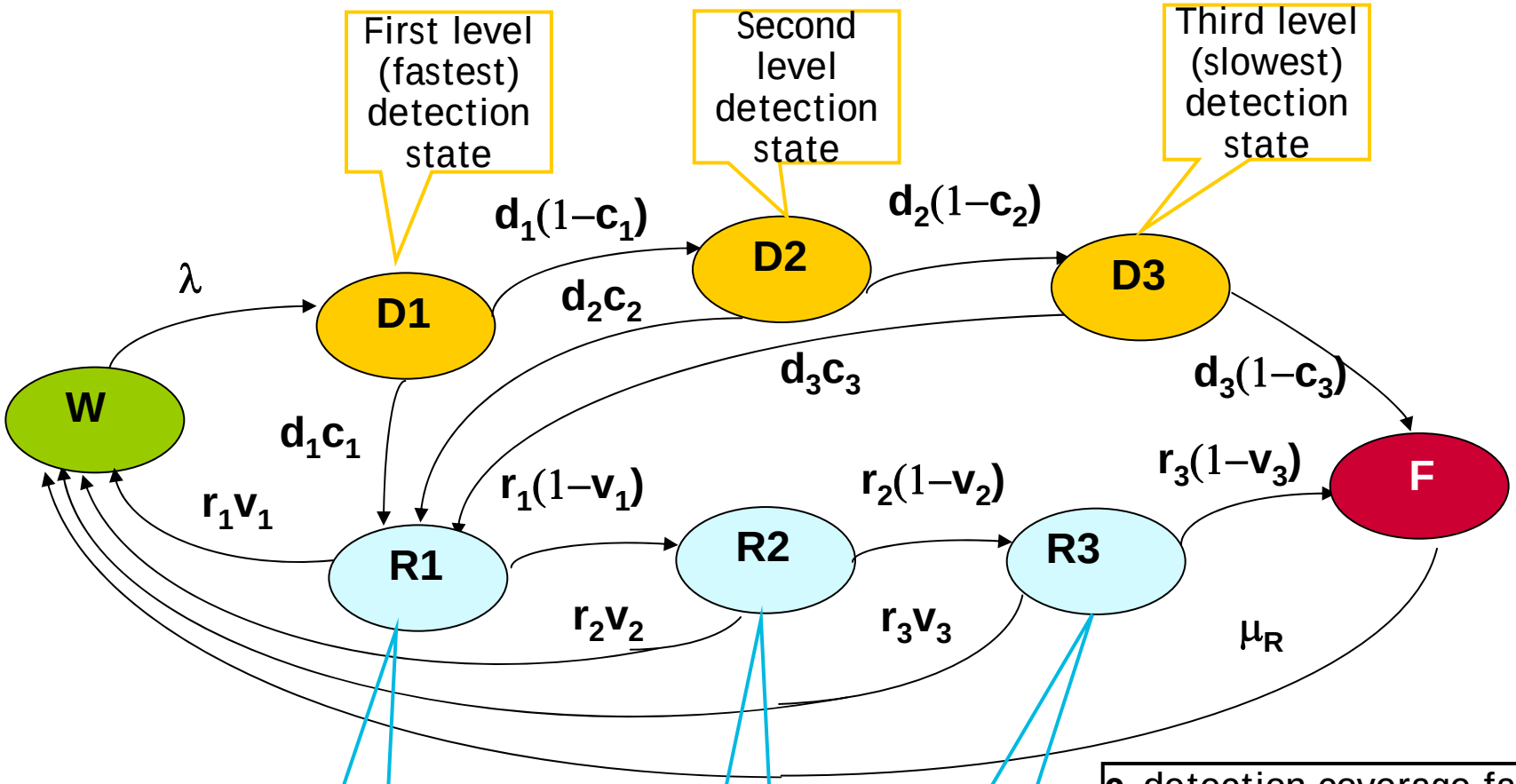


c_i	coverage factor
λ	failure rate
μ_i	recovery rate

W	working
D	detection
R_i	recovery
F	failed

c_i = 0.5 to 0.9, c = 0.9
λ = 100 error triggers/year
μ_1 = 1/10, μ_2 = 1/100,
μ_3 = 1/1000, δ =1/120 (/ sec)
μ_R = 1/3600 repairs/sec

Reference Error Detection and Recovery Model



Combined Detection & Recovery schemes ranked by steady-state probabilities

	W	F	D+R	F/(D+R)
Best ⟷ Worst	H H L	L H H	H H L	L H H
	H L H	H L H	H L L	H L H
	L H H	H H L	H L H	L L H
	H L L	L L H	L H L	H H L
	L H L	L H L	L H H	L H L
	L L H	H L L	L L H	H L L

- H indicates high coverage, modeled as 0.9
- L indicates low coverage, modeled as 0.5

- HHL indicates:
  - high coverage recovery method, escalating to another high coverage recovery method, that escalates to a low coverage recovery method

Detection Techniques and Costs

Many different techniques are well known in the industry.

They have varying:

- run-time cost (speed of recovery)
- development cost
- potential to provide high coverage solutions

Technique	Run-Time Cost	Development Cost	Coverage Potential	Detection State
Invalid Arithmetic detection	Medium	High	High	1
Software based Protocol checking	Medium	Medium	Medium	1
Complete Parameter Checking	High	Medium to High	Medium	2
Routine Hardware Exercises	High	Medium	High	3
Routine Correcting Audits	High	High	High	3

Recovery Techniques and Costs

Technique	Run-Time Cost	Development Cost	Coverage Potential	Recovery State
Data Reset	Low	Low	Low	1
Correcting Audits	Low	Low to High	Low to High	1
Checkpoint and Rollback	Medium	High	High	2
Failover	Medium	High	High	2
Complete Restart	High	Low to Medium	High	3

Combined Model Observations

The combined model exhibits more of the expected behaviour in terms of system availability and recovery than the individual models showed.

- The combined model show the range of availability that we expect from varying coverage factors.
- The model shows that both detection and recovery coverage factors must be high in order to achieve high availability.

Building a system with a low coverage factor for either detection or recovery will not result in a system with high levels of availability, which is counter to conventional wisdom.

Coverage Factors 3 D, 3R	Prob (Detection)	Prob (Recovery)	Prob (Detection or Recovery)	Prob (Failed)	Prob (Working)	Prob (Not Up)	Prob Failed / (Detection + Recovery)
.9 .9 .9 .5 .5 .5	0.00760	0.07847	0.08606	0.11483	0.79910	0.20090	1.33
.5 .5 .5 .9 .9 .9	0.07862	0.00666	0.08528	0.11493	0.79980	0.20020	1.35

Conclusion and Next Steps

- Extended earlier work to consider variable coverage factors in detection and recovery.
- Combined model produced results that were more in line with intuition than the individual models had previously provided.
- Detection and Recovery coverage must both be high to achieve high availability.
- Future work includes analyzing additional scenarios, as well as considering the cases when the detection and recovery rates and coverage factors vary independently based upon the type of fault and error that is present.