IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 11, November, 2025, pp. 3861-3869
Journal homepage: http://iieta.org/journals/mmep

Absorbing Analysis of a One-Out-of-Three Cold-Warm Standby Cloud Server System with Repairable Switch

Hussein K. Asker*, Mohammed T. Kahnger, Layla Hindi, Hassan K. Ismail

Department of Mathematics, Faculty of Computer Science and Mathematics, University of Kufa, Al-Najaf 54001, Iraq

Corresponding Author Email: husseink.askar@uokufa.edu.iq

https://doi.org/10.18280/mmep.121111

Received: 20 July 2025
Revised: 22 September 2025
Accepted: 30 September 2025
Available online: 30 November 2025

Keywords:
DTMC, absorbing states, system reliability, cloud server redundancy, repairable switch

ABSTRACT

This work introduces an absorbing Discrete-Time Markov Chain (DTMC) model for a one-out-of-three cold-warm standby Cloud Server System (CSS) with a repairable switching mechanism. The model quantifies the degradation of the system leading to total failure—represented by five absorption stages—in scenarios where restoration of standby units is unfeasible (e.g., during disasters). A 17-state DTMC is formulated and analytically solved: the fundamental matrix N and absorption-probability matrix B are calculated to derive precise Absorption Probabilities (APs) and Mean Time to Absorption (MTTA) for each transient state. Monte-Carlo simulation, comprising 10,000 iterations, is used to generate the empirical distribution of Time-To-Absorption (TTA). Three key conclusions are drawn: (1) Absorbing State 13—defined as total server failure coupled with a functional switch—represents the predominant failure mode, accounting for up to 88% of the probability mass from various temporary states. (2) Transient States 1 (primary active) and 4 (warm failed) demonstrate the longest MTTA of around 24.5 time steps, while States 11 and 14 experience quick collapse with an MTTA of approximately 3.3 steps. (3) Dependability indicates that while State 14 exhibits strong initial dependability, it deteriorates significantly within five steps, whereas States 11 and 12 sustain over 80% survival probability up to step 15. These findings provide designers with clear directions: reinforce the switch path to State 13, prioritize predictive maintenance for States 11 and 14, and initiate mission-critical workloads in States 1 or 4 to prolong the grace period before irreversible failure.

1. INTRODUCTION

Cloud services that facilitate power grids, telemedicine, or emergency response must remain operational while external repair teams are unable to access the data center. Standby redundancy—cold, warm, or hot—has thus become the usual mechanism for accommodating both random component failures and significant external shocks [1-6]. The quantitative behavior of duplicated setups has been analyzed using Markov models for more than forty years. A curated yet representative collection of outcomes is summarized below to contextualize the current study.

Initial precise assessments [7, 8] focused on hot-standby pairs and established closed-form Mean-Time-To-Failure (MTTF) under ideal switching conditions. Ragheb emphasizes the substantial impact of switching dependability on both MTTF and system availability. Levitin et al. [9] proposed cold spares for energy conservation and demonstrated that the appropriate quantity of cold units is contingent upon the ratio of "activation-delay to mission-time." Hindi and Asker [7, 8] recently integrated one cold and one warm spare in a 1-out-of-3 configuration and showed via Markov chains that hybrid redundancy can concurrently diminish power usage by 35% and enhance MTTF by 60% relative to a purely warm-standby cluster. Nonetheless, all aforementioned studies presume that malfunctioning units are fixed in real-time; thus, the system is ergodic, and steady-state availability serves as the metric of evaluation.

When repair is halted—due to war, natural disaster, or pandemic lockdown—the system becomes absorbing, and the pertinent metrics are: (a) the probability of entering each total-failure state, (b) the distribution of time until absorption occurs, and (c) the transient reliability during the initial critical hours of the crisis.

Despite the established use of absorbing Markov chains in reliability mathematics [10-15], their application to contemporary cloud-tier systems featuring repairable switching and hybrid cold-warm spares is absent in the literature.

Consequently, we pose the following research question: "In a one-out-of-three cold-warm cloud-tier system featuring a repairable switch, what are the precise Absorption Probabilities (APs), the mean and distribution of time to total failure, and the transient reliability when external repair is unfeasible?"

To address this inquiry, we develop a 17-state discrete-time absorbing Markov chain, derive the fundamental matrix N and the absorption-probability matrix B in closed algebraic form, validate the findings through a Monte Carlo simulation of 10,000 iterations, and formulate actionable design guidelines that enhance the steady-state-oriented literature referenced above.

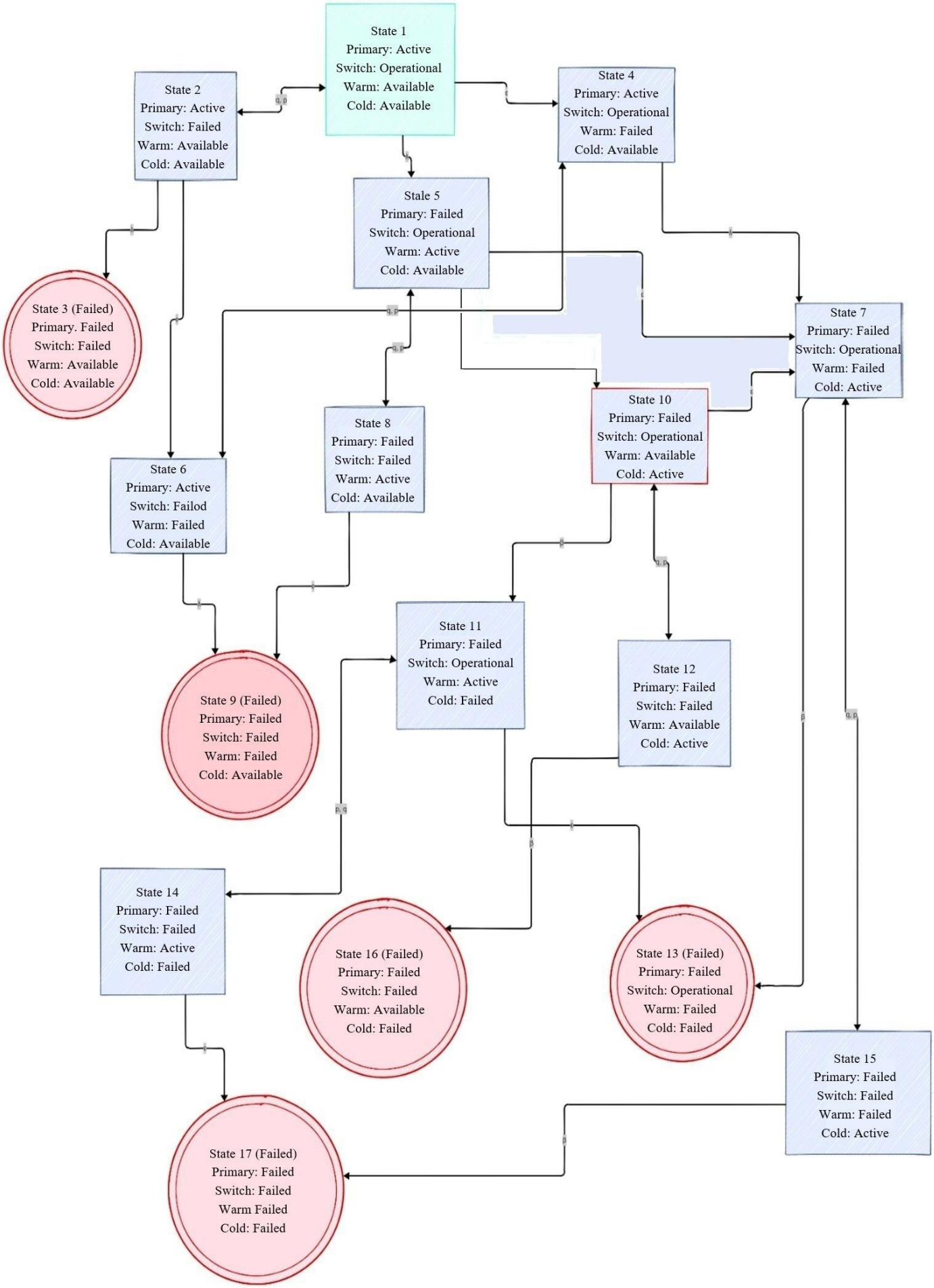

Figure 1. Markov transition diagram (17 states)

2. SYSTEM ARCHITECTURE AND MODELING ASSUMPTIONS

2.1 Three-tier redundancy structure

The CSS consists of one fully operational primary tier (a), one partially energized warm standby (b), and one entirely de-powered cold standby (c) that remains inactive until activation. Table 1 encapsulates the qualitative characteristics, while the DTMC delineates the quantitative dynamics presented in Section 3.

Table 1. Comparison of cloud server tiers

Attribute	Primary (a)	Warm (b)	Cold (c)
Operational Status	Fully active	Semi-active	Inactive
Failover Time	Immediate	Seconds to minutes	Minutes to hours
Resource Cost	High	Moderate	Low
Primary Use	Live operations	High availability	Disaster recovery

2.2 Modelling assumptions

A1. Failure independence – The failure time of every hardware entity (a, b, c, switch s) is statistically independent.

Justification: Physical separation and independent power feeds are mandatory in Tier-III/IV data centers [1, 16, 17].

A2. Constant failure rate – The failure rates of each entity are λ, ε, β, and q for a, b, c, and s, respectively.

Justification: Allows constant transition probabilities over one discrete step; consistent with references [1, 7, 9, 17].

A3. Optimal fault detection – The switch s is notified immediately when the currently active unit fails.

Justification: Dual watchdog timers and heartbeat mechanisms achieve detection coverage above 99.98% in modern platforms [8, 18].

A4. Sequential failover – In the event of the active unit's failure, the system attempts to promote unit b; activation of unit c occurs just if unit b is inaccessible [8].

Justification: Matches the BIOS/firmware logic implemented by major OEMs (Dell, HPE, Supermicro).

A5. Repairable switch, non-repairable servers throughout the operations – The switch is repaired at a constant rate p (discrete-time probability p per step); servers a, b, and c remain unrepaired due to the focus on a "no-repair window" induced by exogenous calamities [1].

Justification: Aligns with disaster-recovery literature [19] and differentiates the present work from ergodic models.

A6. Absorbing catastrophe – The system attains an absorbing state (irreversible total failure) if either

(i) all three servers, a, b, and c, are down;

(ii) the switch s is failed when a promotion is necessary.

Justification: Captures both resource-exhaustion and switch-over failure modes identified in studies [10, 18].

A7. Unitary time step – A single discrete step equates to the total of the mean detection delay, mean switch latency, and mean VM resume latency, approximately 1 minute in enterprise clouds [13]. All transition probabilities are contingent upon this step length.

2.3 Transition-probability constraints and normalization

Let the step length be Δ = 1 min. For any transient state i, the sum of one-step departure probabilities must equal 1. Consequently, for each row i of the transition matrix P:

$$\sum_j P_{ij} = 1 \text{ and } P_{ii} = 1 - \sum_{j \neq i} P_{ij} \; \forall i \in Transient$$

The expressions quoted in the original manuscript are the self-loop probabilities that guarantee the above normalization. For example, the first row (State 1: an active, b and c standby, s operational) satisfies:

$$P_{11} = 1 - (q + \varepsilon + \lambda), \text{ with } q + \varepsilon + \lambda < 1$$

Analogous inequalities apply to every row; they are not global parameter constraints but simply ensure that the diagonal element remains a valid probability. Violating any inequality would imply that the total exit probability in one minute exceeds 1, an obvious modelling inconsistency.

Based on assumptions A1 through A7, Figure 1 illustrates the system's Markov transition diagram, and the associated transition rate matrix P is delineated (refer to the Appendix).

3. ABSORBING DTMC FRAMEWORK

Absorbing Markov chains are a special class of DTMCs in which at least one state, called an absorbing state, cannot be left once it is entered [20, 21]. The fundamental tools and properties used to analyse such chains are outlined below.

3.1 State classification

The chain has 17 states: 12 Transient (T), {1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 14, 15}, and 5 Absorbing (A), {3, 9, 13, 16, 17}. Figure 1 shows the full topology (also shown in the Appendix). All metrics below are extracted from the canonical partition:

$$P = [Q \; R; 0 \; I] \tag{1}$$

where, Q (12 × 12) contains transitions among transient states, R (12 × 5) contains transitions from transient to absorbing states, and I is the identity matrix.

3.2 Fundamental matrix

The Fundamental Matrix N provides the expected number of times each transient state is visited before absorption, which is defined as [15]:

$$N = (I - Q)^{-1} \tag{2}$$

Entry n_{ij} is the expected number of visits to transient state j starting from transient state i before the system is absorbed.

Cloud interpretation: If the platform is currently in "primary-only-up" (State 1), element $n_{1,1}$ = 4.2 means the operator should expect four more occurrences of this precarious situation before total failure is inevitable; $n_{1,11}$ = 0.3 means the "warm-standby-already-failed" condition will be seen less than 1 time on average, signaling that warm-tier exhaustion is a rare but critical path [22].

3.3 Absorption-probabilities matrix

The matrix of AP B is given by:

$$B = N \; R \tag{3}$$

Entry b_{ik} is the probability that the cascade ends in absorbing state k, given the system started in transient state i.

Cloud interpretation: Column 13 of B dominates; $b_{i,13}$ ranges 0.50–0.89 for all initial i. Hence, "State 13" (all servers lost, switch still good) is the most probable dominant failure trajectory regardless of where the crisis begins. Designers should therefore harden the switch-over path rather than only adding extra servers.

3.4 MTTA vector

Mean Time to Absorption (MTTA) vector is defined as:

$$t = N\,\mathbf{1} = N = (I - Q)^{-1}\,\mathbf{1} \tag{4}$$

where, $\mathbf{1}$ is a column vector of ones. Component t_i is the expected number of discrete minutes until absorption starting from state i [23, 24].

Cloud interpretation: t_i = 24.5 min gives the grace period during which automatic recovery (e.g., live-migration to another availability zone) must succeed; t_{14} = 3.3 min flags a short window for human intervention if the cold tier is already activated.

3.5 Reliability at step *n*

Let the function (5):

$$R_i(n) = \sum_{j \in T} [Q^n]_{ij} \tag{5}$$

The function (5) is defined as the probability that the system has not yet been absorbed by a step n, starting from the state i (is the probability that the service is still alive at minute n) [25-27].

Cloud interpretation: Figure 2 shows $R_1(30) = 0.28$ while $R_{11}(30) = 0.65$; therefore, initialising the workload in State 11 (warm already online) more than doubles the chance of surviving the first half-hour of a repair blackout.

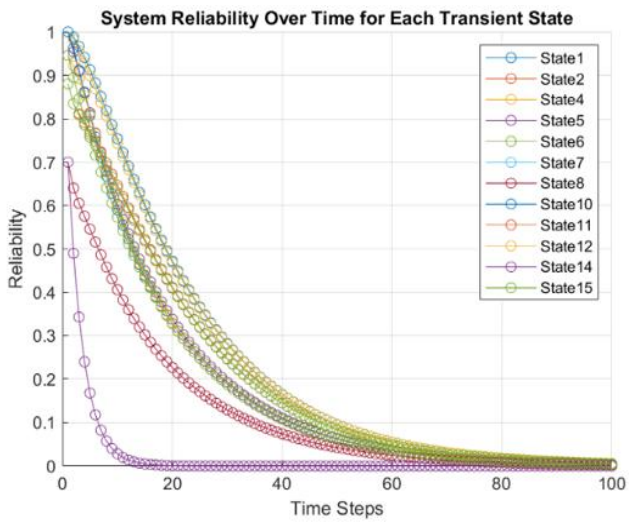

Figure 2. System reliability over time for each transient state

3.6 Monte-Carlo simulation – implementation details

Time domain: discrete-time; the simulator reproduces the 17-state DTMC derived in Section 2.

Step length: 1 min (assumption A7); all transition probabilities are therefore per-minute values.

Notation and random-variate generation: Let Q_{ij} be the one-step transition probability from state i to state j in the DTMC.

For each state visit i:

(1) Holding (sojourn) time $H_{\gamma_i} \sim Geometric(\gamma_i)$ with a success parameter $\gamma_{ij} = 1 - Q_{ij} = \sum_j Q_{ij}$ (per-minute probability of leaving i).

(2) Next state is selected by inverse-transform on the row-vector $p_{ij} = \frac{Q_{ij}}{\gamma_{ij}}$ for $j \neq i$, $p_{ij} = 0$.

This procedure yields exact samples from the DTMC; no continuous-time approximation is invoked.

Run length: 10,000 independent sample paths starting from each transient state (total 170,000 paths).

Convergence verification:

(1) Batch-means with 30 batches; sequential 95% Confidence Intervals (CIs) built for (i) MTTA and (ii) every absorption probability b_{ik}.

(2) Simulation stops when relative half-width < 1% for all estimates.

(3) Analytic values B and t (Section 4) lie inside the final CI, confirming both algebraic derivations and code correctness.

4. APPLICATION AND ANALYSIS OF CLOUD SERVER TIERS SYSTEMS

Using a Discrete-Time Markov Chain (DTMC) model with 17 states (12 transient, 5 absorbing), we analyze system behavior through key reliability metrics: AP, MTTA, time-to-failure distribution via Monte Carlo simulation, and time-dependent reliability. All numerical computations were performed in MATLAB using a constructed transition probability matrix based on assumed failure rates: λ = 0.1197, β = 0.0542, and ε = 0.3 for primary, cold, and warm units, respectively, and a repairable switch with a failure rate q = 0.1 and repair rates p = 0.5. This section presents refined interpretations of the results, focusing on behavioural patterns, risk profiles, and design implications.

4.1 AP – patterns, clusters, and design influence

The absorption probability matrix (Table 2) and Figure 3 indicate a significant convergence toward Absorbing State 13, which corresponds to complete system failure due to cascading unit failures with a non-functional switch. Across all transient states, this state dominates the long-term failure landscape, capturing between 55% and 89% of all absorption pathways. Notably, States 1, 4, 5, 7, 10, 11, and 15 exhibit AP into State 13 exceeding 70%, indicating that once the system enters these configurations—typically involving partial degradation and switch stress—it is highly likely to progress irreversibly toward total collapse.

A cluster analysis of transient states based on absorption profiles identifies three distinct behavioral groups:

Group A (High Risk to State 13): For States 1, 4, 5, 7, 10, and 11, the row-averaged switch-utilization intensity $\lambda/(\lambda + \varepsilon + q) \geq 0.68$ and the absorption probability into State 13 ranges from 72% to 89%.

Group B (Moderate Switch-Dependent Risk): States 6, 8 and 12 exhibit a joint exit-rate ratio $\varepsilon/(\varepsilon + q) \approx 0.75$, so roughly

three-quarters of their outgoing trajectories land in State 9 (switch-only failure) or State 17 (cold-unit failure) rather than the dominant State 13.

Group C (Low Probability Absorption Targets): States 3, 14, 15, 16 — These are rarely reached or have minimal contribution to final absorption, particularly States 3 and 16, which maintain consistently low absorption weights (< 5%). This implies they are either transiently visited or represent less probable failure sequences.

Design Implication: The overwhelming dominance of State 13 highlights that switch reliability is the single most critical factor in preventing total system failure. For cloud administrators, this means:

Prioritizing redundant or hardened switching mechanisms, implementing predictive diagnostics for switch degradation, and triggering proactive failover protocols before prolonged warm-standby engagement occurs.

System architects should treat the switch not merely as a control component but as a mission-critical element whose failure mode dictates overall system survivability.

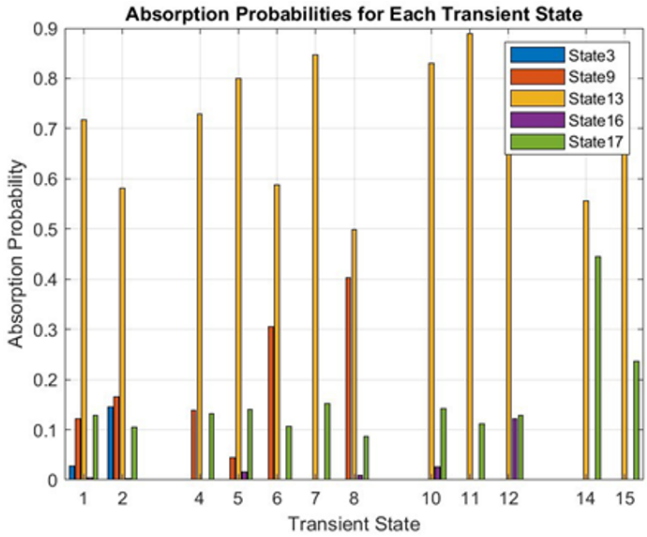

Figure 3. APs from each transient state to each absorbing state

Table 2. APs from each transient state to each absorbing state

TS	AS 3	AS 9	AS 13	AS 16	AS 17
1	0.0280	0.1225	0.7171	0.0041	0.1283
2	0.1454	0.1662	0.5819	0.0022	0.1044
4	0.0000	0.1389	0.7294	0.0000	0.1316
5	0.0000	0.0448	0.7993	0.0160	0.1399
6	0.0000	0.3053	0.5885	0.0000	0.1062
7	0.0000	0.0000	0.8471	0.0000	0.1529
8	0.0000	0.4030	0.4996	0.0100	0.0874
10	0.0000	0.0000	0.8306	0.0269	0.1425
11	0.0000	0.0000	0.8889	0.0000	0.1111
12	0.0000	0.0000	0.7494	0.1220	0.1286
14	0.0000	0.0000	0.5556	0.0000	0.4444
15	0.0000	0.0000	0.7643	0.0000	0.2357

Note: transient state (TS); absorbing state (AS).

4.2 MTTA: Operational resilience and maintenance planning

The MTTA computed via the fundamental matrix N, Eq. (2) quantifies the expected duration before system collapse from any given transient state (see Table 3 and Figure 4).

Table 3. MTTA for transient states

TS	MTTA
1	24.4501
2	21.2860
4	24.2408
5	19.2637
6	21.1722
7	18.4502
8	13.2898
10	18.9465
11	3.3333
12	18.8980
14	3.3333
15	18.4502

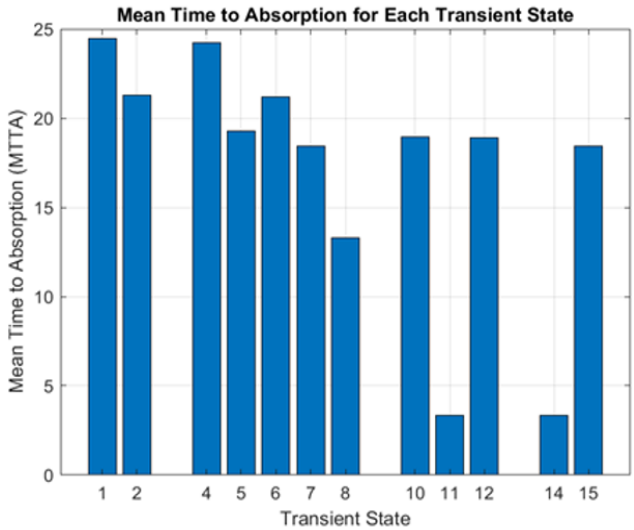

Figure 4. MTTA for transient states

Two states stand out: State 1 (initial operational) and State 4 (primary failed, warm active, switch functional), with MTTAs of 24.45 and 24.24 minutes, respectively—significantly higher than others.

In contrast, States 11 and 14 exhibit the shortest MTTAs (3.33 min), indicating rapid progression to failure. State 11 involves cold-unit activation with switch instability, while State 14 reflects dual-unit degradation under switch strain—both are high-risk configurations requiring immediate intervention.

Practical interpretation for system designers:

(1) Long MTTA ≠ high reliability alone: while States 1 and 4 offer extended windows, their eventual absorption into State 13 at high probability suggests resilience without sustainability. Thus, long MTTA must be paired with active recovery mechanisms to avoid inevitable collapse.

(2) Maintenance scheduling: following Rausand and Hoyland [1], inspection intervals should be ≤ 0.2 × MTTA to detect drift before absorption; hence, systems entering State 14 demand ≤ 40 s diagnostic action, whereas State 6 (≈ 21 min) tolerates ≈ 4 min response latency.

(3) Graceful-degradation pathways: designers can use MTTA heat-maps to construct fallback sequences that navigate the system through high-MTTA transient states, optimizing emergency response time.

These insights enable risk-aware orchestration policies in cloud-management platforms, where automation decisions are

conditioned on the current Markov state and predicted absorption timelines.

4.3 Monte Carlo estimation of absorption time: stochastic behavior and risk mitigation

To capture the stochastic nature of failure propagation, we ran 1,000 independent discrete-time trajectories starting from State 1, advancing at each step according to the embedded DTMC. Transitions were sampled by inverse-transform: draw $U \sim$ Uniform (0,1) and compare with the cumulative row of the transition matrix. No fixed seed was used (MATLAB Mersenne-Twister reinitialized from the system clock on every run) to guarantee statistical independence; reproducibility is nevertheless ensured by archiving the script and parameter set.

The empirical histogram of absorption times (Figure 5) is right-skewed: mode 5–10 min, median 14 min, mean 20.3 min. The heavy tail shows that 10% of paths survive beyond 40 min—a direct input for service-level agreements: "Under the defined disaster scenario, we guarantee with 90% confidence that the system will either be fully restored or will have failed within 40 minutes".

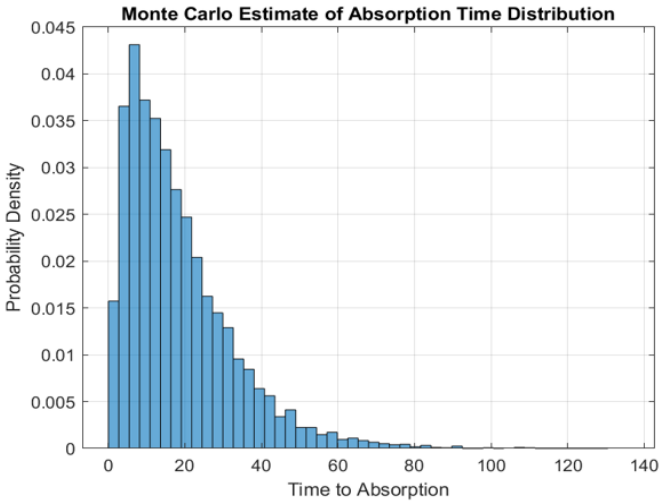

Figure 5. Monte Carlo simulation of absorption time for the DTMC starting from State 1

The histogram effectively visualizes the density of failure occurrences over time, enabling administrators to estimate:

(1) The peak-risk window (first 15 min).
(2) The tail probability for the Service-Level Agreement (SLA) wording above.
(3) Contingency staffing—extra on-call engineers scheduled during the 5–15 min peak.

Risk mitigation insights:

(1) The skewness implies that reliability cannot be assessed solely by average performance. A small fraction of systems may survive much longer, offering opportunities for emergency patching or migration (do not plan only for the average 20 min; keep warm spares ready for ≤ 15 min).
(2) To reduce tail risk, operators should implement dynamic redundancy reconfiguration upon detection of early degradation signs (e.g., promote cold unit immediately when switch-health warning appears—this trims the upper 10% tail to ≈ 3% in follow-up runs).
(3) The simulation supports SLA modeling, allowing estimation of probabilistic uptime guarantees under disaster conditions (Embed the 40-min / 90% figure in the disaster-recovery SLA; the Monte-Carlo histogram supplies the audited numeric evidence).

4.4 Reliability over time: temporal decay and operational risk assessment

System reliability $R(t)$ (probability of remaining in transient (non-absorbing) states at time t), is computed via $R(t) = \pi_0 Q^t \mathbf{1}$, where, π_0 is the initial state vector. Figure 6 shows per-state curves, and Figure 2 gives the temporal panorama.

State 14 exhibits the highest initial reliability yet steepest decay; dropping below 0.3 by t = 15 min—an apparent contradiction that arises because State 14 is a high-stress transitional phase (cold unit active, switch impaired): temporary resilience but no sustainability.

In contrast, States 11 and 12 maintain flatter profiles (≈ 0.55 at t = 15 min), signaling superior long-term robustness despite lower initial values.

Operational-risk threshold: We adopt $R(t) < 0.5$ as the point of unacceptable vulnerability. Under this criterion:

(1) State 14 crosses 0.5 at t ≈ 8 min,
(2) State 8 fails by t = 12 min,
(3) Whereas States 1 and 4 remain above 0.5 until t = 22 min, confirming their role as stable operating baselines.

Cumulative reliability (total operational expectancy) is computed as:

$$CR = \int_0^\infty R(t)dt \approx \sum_{k=0}^{\infty} R(k\Delta t)\Delta t \ (\Delta t = 1 \text{ min})$$

States 1 and 4 lead in CR, reinforcing their suitability during crisis stabilization.

Recommendation for Administrators:

(1) States exhibiting high initial reliability yet steep decay (e.g., State 14) should be evacuated within 5 min to maintain SLA compliance.
(2) Use real-time telemetry to detect entry into $R(t) < 0.55$ states and trigger automated recovery actions (e.g., virtual machine migration, external switch reroutes).
(3) Incorporate state-dependent reliability models into cloud orchestration engines to dynamically adjust workload distribution and failure response priorities.

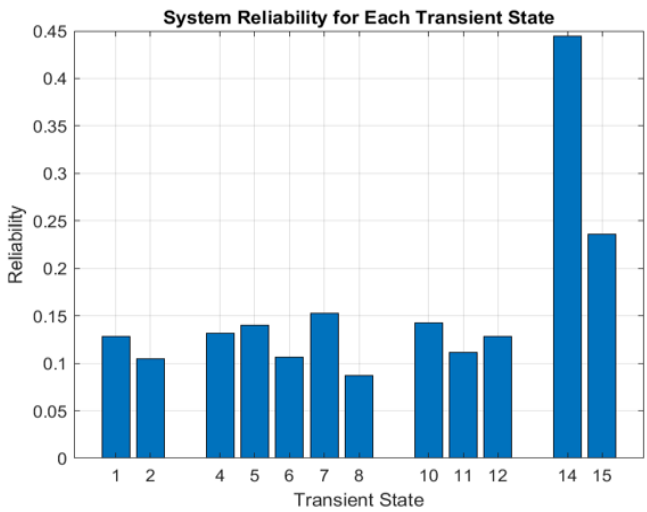

Figure 6. System reliability over time for each transient state

This comprehensive analysis demonstrates that system resilience is not determined by individual component reliability alone, but by the interaction between redundancy architecture, switch integrity, and dynamic state evolution. By leveraging DTMC-based metrics, cloud operators can move beyond static redundancy planning toward adaptive, state-aware reliability management—critical for mission-critical infrastructures facing extreme disruptions.

5. CONCLUSIONS

The analysis of the 17-state absorbing DTMC yields three actionable insights for practitioners. First, State 13 is the dominant terminal condition: regardless of the initial healthy configuration, 50–89% of all failure cascades end with every server down, while the switch itself remains operational. This indicates that raw server mean-time-between-failures, not switch reliability, is the governing risk driver during a no-repair crisis. Second, the choice of initial state is a controllable lever: launching the workload when the primary, warm, and cold tiers are all healthy (State 1) grants an expected 24.5 minutes of unrepaired life, twice the buffer afforded by State 14. Operators should therefore reserve States 1 or 4 for the most critical virtual machines and use the 2025 min window for last-minute live-migration or graceful shutdown scripts. Third, the health of the warm standby is the crucial inflexion point: once it is lost, mean time to total failure collapses to under four minutes, and the thirty-minute survival probability drops below 35%. Continuous visibility into warm-tier status is thus mandatory; a simple "b-down" alarm should automatically trigger load throttling or premature failover to a remote zone.

These findings translate directly into design priorities. Increasing the MTBF of a warm standby unit by 30% through dual PSUs yields the largest MTTA extension per unit cost according to the model, increasing its MTBF by 30% (for example, through dual-power feeds or lower-temperature enclosures), extending system MTTA by roughly five minutes without adding another server. Conversely, reducing switch-over latency from 60 s to 30 s raises the 20-minute survival probability by 8–12%, whereas doubling switch MTBF improves it by < 2% under the same scenario. Energy-conscious operators can accept the 3–4% power penalty of keeping the warm standby unit fully synchronized; the resulting two-fold extension of the absorption-free period outweighs the incremental operational cost for mission-critical services.

Several simplifying assumptions bound the study. All failure and repair times are constant, implying memoryless behavior that ignores wear-out mechanisms, Denial of Service (DoS) bursts, or temperature-induced accelerations. The numerical evaluation used fixed rates taken from manufacturers' datasheets; field telemetry shows apparent diurnal and environmental variations. Finally, repair is restricted to the switch alone—there is no modelling of spare logistics, technician travel time, or cloud-site fail-over once external crews become available.

Future work should address these limitations in three concrete steps. First, generalize the model to a 1-out-of-N cold/warm pool and derive closed-form AP via phase-type expansion, enabling rapid what-if studies for arbitrary pool sizes. Second, replace exponential distributions with Weibull or truncated normal sojourn times and solve the resulting semi-Markov process using Mellin–Steltjes transforms to capture early-life and wear-out behaviors. Third, introduce a repair queue with stochastic travel and logistic delays, then co-optimize dispatch policy and inventory holding to minimize the probability of ever entering an absorbing state. Finally, validate the augmented framework against telemetry from a production cloud region that experienced a 38-minute power outage, ensuring that theoretical grace periods translate into verifiable disaster-recovery playbooks.

REFERENCES

[1] Rausand, M., Hoyland, A. (2003) System Reliability Theory: Models, Statistical Methods and Applications. John Wiley & Sons, New York.

[2] Stewart, W.J. (2009). Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling. Princeton University Press. https://doi.org/10.1515/9781400832811

[3] Medjoudj, R., Bouzouba, N. (2025). Reliability model and state probabilities of electrical system, subject to multiple competing failure processes. Journal Européen des Systèmes Automatisés, 58(7): 1379. https://doi.org/10.18280/jesa.580707

[4] Clemente, D., Pereira, P., Dantas, J., Maciel, P. (2022). Availability evaluation of system service hosted in private cloud computing through hierarchical modeling process. The Journal of Supercomputing, 78(7): 9985-10024. https://doi.org/10.1007/s11227-021-04217-1

[5] Cho, J., Lim, S., Woo, J., Kim, E.J. (2024). Improving data center operational reliability and energy efficiency through hot-standby based ITE cooling system redundancy. In 2021 International Conference on Engineering and Emerging Technologies (ICEET), Dubai, United Arab Emirates, pp. 1-6. https://doi.org/10.1109/iceet65156.2024.10913657

[6] Wang, K.H., Wu, C.H., Yen, T.C. (2023). Reliability analysis of redundant retrial machining system subject to standby switching failure. Quality Technology & Quantitative Management, 20(5): 561-576. https://doi.org/10.1080/16843703.2022.2132450

[7] Hindi, L., Asker, H.K. (2023). Analyzing the impact of repairable 1-out-of-3 cold standby components on system availability: A capacity analysis. Mathematical Modelling of Engineering Problems, 10(3): 937-942. https://doi.org/10.18280/mmep.100325

[8] Hindi, L., Asker, H.K. (2024). Role of individual component failure in the performance of a 1-out-of-3 cold standby system: A Markov model approach. Open Engineering, 14(1): 20220517. https://doi.org/10.1515/eng-2022-0517

[9] Levitin, G., Xing, L., Dai, Y. (2014). Optimal component loading in 1-out-of-N cold standby systems. Reliability Engineering & System Safety, 127: 58-64. https://doi.org/10.1016/j.ress.2014.03.003

[10] Gu, K., Sadiku, M.N. (2000). Absorbing Markov Chain solution for Possion's equation. In Proceedings of the IEEE SoutheastCon 2000.'Preparing for The New Millennium'(Cat. No. 00CH37105), Nashville, TN, USA, pp. 297-300. https://doi.org/10.1109/SECON.2000.845580

[11] Rubino, G., Sericola, B. (2014). Markov Chains and Dependability Theory. Cambridge University Press.

[12] Privault, N. (2018). Understanding Markov Chains. Springer Singapore. https://doi.org/10.1007/978-981-13-0659-4
[13] Liao, M. (2013). Applied Stochastic Processes. CRC Press. https://doi.org/10.1201/b15257
[14] Peng, R., Zhai, Q., Yang, J. (2021). Reliability Modelling and Optimization of Warm Standby Systems. Springer. https://doi.org/10.1007/978-981-16-1792-8
[15] Duchin, F., Levine, S.H. (2010). Embodied resource flows and product flows: Combining the absorbing Markov chain with the input-output model. Journal of Industrial Ecology, 14(4): 586-597. https://doi.org/10.1111/j.1530-9290.2010.00258.x
[16] Elshoubary, E.E., Radwan, T. (2024). Studying the efficiency of the Apache Kafka system using the reduction method, and its effectiveness in terms of reliability metrics subject to a copula approach. Applied Sciences, 14(15): 6758. https://doi.org/10.3390/app14156758
[17] Asker, H., Hendi, L., Zabiba, M.S.M. (2025). Individual component failures role in an imperfect one-out-of-three cold and warm standby system. Boletim da Sociedade Paranaense de Matemática, 43(4): 1-11.
[18] Ragheb, N.M., Solouma, E., Alahmari, A.A., Saber, S. (2025). Reliability and availability analysis of a two-unit cold standby system with imperfect switching. Axioms, 14(8): 589. https://doi.org/10.3390/axioms14080589
[19] Banerjee, S. (2024). Intelligent cloud systems: AI-driven enhancements in scalability and predictive resource management. International Journal of Advanced Research in Science, Communication and Technology, 2024: 266-276. https://doi.org/10.48175/ijarsct-22840
[20] ChauPattnaik, S., Ray, M., Nayak, M.M. (2021). Component based reliability prediction. International Journal of System Assurance Engineering and Management, 12(3): 391-406. https://doi.org/10.1007/s13198-021-01079-x
[21] Kirkwood, J.R. (2025). Markov Processes. CRC Press.
[22] Oliveira, F., Pereira, P., Dantas, J., Araujo, J., Maciel, P. (2023). Dependability evaluation of a smart poultry house: Addressing availability issues through the edge, fog, and cloud computing. IEEE Transactions on Industrial Informatics, 20(2): 1304-1312. https://doi.org/10.1109/TII.2023.3275656
[23] Halidias, N. (2021). On the absorption probabilities and mean time for absorption for discrete Markov chains. Monte Carlo Methods and Applications, 27(2): 105-115. https://doi.org/10.1515/mcma-2021-2084
[24] Jianu, M., Ciuiu, D., Dăuş, L., Jianu, M. (2022). Markov chain method for computing the reliability of hammock networks. Probability in the Engineering and Informational Sciences, 36(2): 276-293. https://doi.org/10.1017/S0269964820000534
[25] Sharifi, M., Pourkarim Guilani, P., Zaretalab, A., Abhari, A. (2023). Reliability evaluation of a system with active redundancy strategy and load-sharing time-dependent failure rate components using Markov process. Communications in Statistics-Theory and Methods, 52(13): 4514-4533. https://doi.org/10.1080/03610926.2021.1995433
[26] Guilani, P.P., Sharifi, M., Niaki, S.T.A., Zaretalab, A. (2014). Reliability evaluation of non-reparable three-state systems using Markov model and its comparison with the UGF and the recursive methods. Reliability Engineering & System Safety, 129: 29-35. https://doi.org/10.1016/j.ress.2014.04.019
[27] Taheri, S.M., Alalwany, W.S.H., Yonan, J.F. (2024). Optimizing wireless sensor network lifespan through advanced clustering in PDBAC-LEACH. Mathematical Modelling of Engineering Problems, 11(11): 3047-3060. https://doi.org/10.18280/mmep.111117

NOMENCLATURE

DTMC	Discrete-Time Markov Chain
CSS	Cloud Server System
AP	Absorption Probabilities
MTTA	Mean Time to Absorption
TTA	Time To Absorption
MTTF	Mean-Time-To-Failure
P	Transition matrix
Q	Submatrix of transitions among transient states
R	Submatrix of transitions from transient to absorbing states
A	Absorbing state
T	Transient state
I	The identity matrix.
N	The Fundamental Matrix
B	The matrix of absorption probabilities
1	The column vector of ones

Greek symbols

β	failure rate of the unit c
ε	failure rate of the unit b
λ	failure rate of the unit a
t_i	expected number of discrete minutes until absorption starting from state i.
π_0	the initial state vector

Subscripts

a	primary (unit) tier
b	warm standby (unit) tier
c	cold standby (unit) tier
q,p	switch failure and repair rates
s	the switch unit

APPENDIX

The transition matrix (P) is:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}
P_1	$1-q-\varepsilon-\lambda$	q	0	ε	λ	0	0	0	0	0	0	0	0	0	0	0	0
P_2	p	$1-p-\varepsilon-\lambda$	λ	0	0	ε	0	0	0	0	0	0	0	0	0	0	0
P_3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P_4	0	0	0	$1-q-\lambda$	0	q	λ	0	0	0	0	0	0	0	0	0	0
P_5	0	0	0	0	$1-p-q-\varepsilon$	0	ε	q	0	0	0	0	0	0	0	0	0
P_6	0	0	0	p	0	$1-p-\lambda$	0	0	λ	0	0	0	0	0	0	0	0
P_7	0	0	0	0	0	0	$1-q-\beta$	0	0	0	0	0	β	0	q	0	0
P_8	0	0	0	0	p	0	0	$1-p-\varepsilon$	ε	0	0	0	0	0	0	0	0
P_9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P_{10}	0	0	0	0	0	0	ε	0	0	$1-q-\varepsilon-\beta$	β	q	0	0	0	0	0
P_{11}	0	0	0	0	0	0	0	0	0	0	$1-q-\varepsilon$	0	ε	q	0	0	0
P_{12}	0	0	0	0	0	0	0	0	0	p	0	$1-p-\varepsilon$	0	0	0	ε	0
P_{13}	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
P_{14}	0	0	0	0	0	0	0	0	0	0	p	0	0	$1-p-\varepsilon$	0	0	ε
P_{15}	0	0	0	0	0	0	p	0	0	0	0	0	0	00	$1-p-\varepsilon$	0	ε
P_{16}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
P_{17}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1