Transformation of system failure life cycle

Montri Wiboonrat*

Graduate School of Information Technology, Assumption University, Bangkok, Thailand

(Received October 12 2008, Revised December 29 2008, Accepted January 17 2009)

Abstract. This research is an investigation of symptoms of Tier IV data center failures in cases of unplanned and planed downtimes. The paper examines the consequent impacts of system fault, error, and failure from data center operation. It is important to distinguish among fault, error, and failure of the systems. How does the syndrome of active failure and propagating failure associated with pervasive causes of disasters. Each type of system failure has its own characteristic warning signs. The system availability of the relevant indicators is discussed in some detail, and a comprehensive prevention strategy must take into account of each escalating failure.

Keywords: system failures, system reliability, mean time between failures (MTBF)

1 Introduction

In the 21st century, such as large-scale and high distributed networks of telecommunication systems, the Internet, Wi-Fi, WiMAX, voice over internet protocols (VoIP), data centers (DC), etc. improve the efficiency and effectiveness of business transactions. Critical infrastructures, including physical power distribution and communication systems are complex with numerous interconnected and interdependent components. A data center consists of a multitude of hardware parts that are bound to fail sooner or later. The DC Tier IV system is an automatic failover/ fault tolerant strategic concept that is designed to eliminate a single point of failure (SPOF). DC is a large and high complex integrated system. Proper planning maintenance is crucial to keep a system under normal operation for 8,760 hours per year. Unexpected downtimes from system failures are often caused by human omission and commission. Other causes may include natural disasters or component/ subsystem failures. Fault-tolerant system design is an established discipline for eliminating, or at least mitigating risks which cause by human errors, natural catastrophes, component failures, etc. that results in an SPOF in a system.

A step-by-step analysis of DC Tier IV, based on real cases, shed lights on the chronology of component faults and how these failures affect a fault tolerant system of the data center and the power distribution system. This research investigates the chronological impact of the utility outage and how it affects the data center’s maximum tolerable period of disruption (MTPD).

* Corresponding author. E-mail address: mwiboonrat@gmail.com.

Published by World Academic Press, World Academic Union
2 Background

2.1 Definitions

*Availability* is an ability of a component to be in a state to perform a required function at a given instant of time or at any instant of time within a given time interval, assuming that the external resources, if required, are provided.

*Dependability* is the ability to deliver service that can justifiably be trusted.

*Reliability* is an ability of a component to perform a required function under given conditions for a given time interval.

*Security* is the concurrent existence of the following attributes:

1. Availability for an authorized action only
2. Confidentiality
3. Integrity with improper meaning unauthorized

*Fault* is adjudged or hypothesized cause of a system malfunction; it can be internal or external, dormant or active.

*Error* is a deviation from the correct service state for a system or a subsystem.

*Failure* is the transition event that occurs when the delivered service deviates from the correct service state to an unwanted state.

2.2 Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Communication Channels</td>
</tr>
<tr>
<td>CIC</td>
<td>Component’s Inherent Characteristics</td>
</tr>
<tr>
<td>CRAC</td>
<td>Computer Room Air Conditioner</td>
</tr>
<tr>
<td>DC</td>
<td>Data Center</td>
</tr>
<tr>
<td>FPS</td>
<td>Fire Protection System</td>
</tr>
<tr>
<td>MDB</td>
<td>Main Distribution Boards</td>
</tr>
<tr>
<td>MTPD</td>
<td>Maximum Tolerable Period of Disruption</td>
</tr>
<tr>
<td>MTBF</td>
<td>Mean Time between Failures</td>
</tr>
<tr>
<td>MTTF</td>
<td>Mean Time to Failure</td>
</tr>
<tr>
<td>MTTR</td>
<td>Mean Time to Repair</td>
</tr>
<tr>
<td>PDS</td>
<td>Power Distribution System</td>
</tr>
<tr>
<td>SCT</td>
<td>System Connectivity Topology</td>
</tr>
<tr>
<td>SPOF</td>
<td>Single Point of Failures</td>
</tr>
<tr>
<td>SS</td>
<td>Security Systems</td>
</tr>
<tr>
<td>SSN</td>
<td>Servers, Storages, and Networking</td>
</tr>
<tr>
<td>STS</td>
<td>Static Transfer Switch</td>
</tr>
</tbody>
</table>

2.3 System determination and failure models

\[ R(t) : \] The reliability of the system
\[ \lambda : \] Constant failure rate of component

\[ MTBF = \int_0^\infty \exp(\lambda t)dt = \frac{1}{\lambda} \]

\[ R(t) = \exp(-\lambda t) = \exp\left(\frac{-t}{MTBF}\right) \]

\[ R(t) = 1 - \lambda t = 1 - \left(\frac{t}{MTBF}\right) \]

\[ MTBF = \frac{t}{1 - R(t)} \]

MSEM email for contribution: submit@msem.org.uk
A system life cycle of uptime and downtime is stated in an equation form as $MTBF = MTTF + MTTR$. The system availability can be expressed in Eq. (1).

$$\text{System Availability} = \frac{MTTF}{MTTF + MTTR}$$ (1)

Fig. 1 illustrates the life cycle of system operation in terms of state modes: system operation and system breakdown. The $X_i$ is an average time of system failure and $R_i$ is an average time which system repair is complete. $MTBF$ is an average of $X_i - X(i-1)$; whereby $i > 1$. $MTTR$ is a difference between $R_i$ and $X_i$, i.e. $R_i - X_i$.

An increase in $MTTF$ or a decrease in $MTTR$ improves the system availability (SA). $\triangle MTTF$ can be extended dramatically by adopting appropriate technologies, e.g. CIC and SCT\cite{17, 18}. However, $\triangle MTTR$ needs more thorough understanding of human activities and proper technological deployment strategies to handle recovery systems.

Reliability distributions with increasing, decreasing, and constant failure rates are depicted in Fig. 2. The composite curve is bathtub-shaped. The equipment failure rate varies upon respecting of equipment lifetime. A failure rate may serve as an indicator of risk proneness\cite{7}. When the failure rate increases beyond the constant failure rate ($\lambda$), the chance of equipment failure becomes higher. Part change (P.C) is thus recommended to prevent possible future system failures.

An $MTBF$ cycle is not always constant, i.e. $MTBF(i) \geq MTBF(i + 1)$ whereby $i > 1$. This results from change in the failure rate over time, as depicted in Fig. 2. The relationship between system’s normal operating time and successive $MTBF$ is illustrated in Fig. 3.

A system is defined as dependable, when the backup component continues its normal operation immediately following the active component’s failure. Typical faults occur due to parts failures or adverse operating environment. Fig. 4. lists a tree diagram of system dependability and security in terms of three fundamentally related factors, i.e. attributes, threats and means. Understanding of system faults and causes of system faults helps devise solutions to reduce system downtime, because a system has faults latency or tolerance. This faults tolerance allows time for service personnel to make an in-time corrective action before system breakdown. MTPD or ($MTTR + \text{latency}$) is the time allowed for corrective actions to recover or replace components/services and subsystems/modules, as depicted in Fig. 5.

### 3 Data center model assumption

#### 3.1 Conditional validation

Controllable condition of data center operation is classified into 4 operating processes: normal routine human operation activities, temperature condition around $20 - 25^\circ C$, maintenance agreement from components and systems warrantee procedures, and physical regular audits.

Predicable condition of data center operation is classified into 3 conditional statistics: $MTBF$ on this condition part changing needed to take action, on site part readiness to reduce $MTTR$, and corrective action plans to handle foreseen incidents.
Uncontrollable symptoms of components and systems are concern with active or passive failures. The failure of active and passive CICs are different in the process level, as shown in Fig. 6.

Humans are more sensitive to motion and feeling to such a degree that they are less predictable and controllable in precise patterns. As well as, natural disasters are uncontrolled and cannot be predicted exactly in level of severity before incidents occur. They can perceive the results only after the occurrences. The patterns which are controllable, uncontrollable, predictable, and unpredictable events are illustrated in Fig. 6.
3.2 Model validation

The TIA 942 Tier IV Standard of data center connectivity diagram is depicted in Fig. 7. Fig. 7 shows how a failure rate may be reduced by including a redundant system. This Tier IV is called fully redundant component or fault tolerance system that is the highest MTPD in DC system availability design. It is derived from high availability/ reliability data centers at level 99.995% uptimes or only 48 minutes downtime during 5 years\textsuperscript{[15]}

Investigation of data center system failure is conducted with 3 real case studies: planned downtime with regular maintenance or upgraded system, unplanned downtime with condition of human made, and unplanned downtime with system failures.

Planned downtime is under controllable and predictable conditions. It therefore covers standard processes under experts of each field during operation. Three experimental cases; upgraded networking switch, changed uninterruptible power supply (UPS) batteries, and maintenance MDB, are followed the plans of system downtime under MTTR conditions. It does not, however, prove the system has more system reliability or fault tolerance 100% yet.

PDS is the most sensitive system for DC downtime; thus, the research tested fault injection by the cutting of the main utility system as human error or unplanned downtime. The fault tolerant system reacted as shown in Fig. 8. UPSs will take action to recharge power back to system immediately as long as the battery can handle the loaded points. Gen-Set will activate within 15 seconds to change power back to the UPSs to re-supply on loaded points\textsuperscript{[11, 16]}

By TIA 942 Gen-Set availability is 84 hours for consecutive operation without interruptions\textsuperscript{[15]}. Natural faults of Data Center are categorized into four patterns as follows\textsuperscript{[6]}:
(1) **Hardware faults**: Failing devices (CIC) which failure rate is changed overtimes as shown in Fig. 3.

(2) **Design faults**: Faults in system topology (SCT) and CIC incompatibility selection.

(3) **Operations faults**: Mistakes made by personal omission and commission during operations and maintenances, i.e. skip-to-do list, or sabotage.

(4) **Environmental faults**: Temperature, fire, flood, earthquake, lightning, utility outage, tornado, hurricane, tsunami, etc.

### 3.3 Severity level model

To limit the scope of this investigation, the researcher focuses on Tier IV data center, and power distribution system (PDS) from incoming utility throughout loaded consumptions, i.e. server racks or networking equipment racks. This research assumes the DC natural faults in Fig. 2 to Fig. 4, namely hardware faults and design faults. The DC Tier IV model assumes power distribution system, e.g. power cables, circuit breakers, power distribution unit (PDU), and bus-bars are performed 100%.

Normal operation of DC-PDS is depicted in Fig. 7 and Fig. 8. Utility will supply power throughout the PDS: Transformer, automatic transfer switch (ATS), bus-bar, UPSs, bus-bar, and loaded points. After PDS failure, normal operation recovers when either of the two strategies is accomplished, as illustrated in Fig. 8 and Fig. 9:

1. **Automatic Failover**: (I). UPS and Gen-Set is going to handle the loaded points immediately after utility outage. (II). Secondary Site takes over the action as the primary/active when Gen-Set or UPSs on Primary Site failure occurs at the same time or over latency.

*MSEM email for contribution: submit@msem.org.uk*
Fig. 9. Transformation of system failure life cycle model

a. Stage I: when utility outage UPSs will activate automatic immediately to handle load points, as seen in Fig. 8 and latency time of each stage in Fig. 10. Within 15 seconds, Gen-Set will fully resupply power to the system, as illustrated in Fig. 10 and latency time in Fig. 8 and Fig. 9. The system failure is considered as Severity Level 1, as shown in Tab. 1.

b. Stage II: When the Primary Site is totally destroyed by any causes, e.g. natural disasters, sabotage, or Gen-Set or UPS failures, etc. The Primary Site will transfer operations to the Secondary Site within the specified response time ($\Delta rt$), as shown in Fig. 9. This system failover is classified as Severity Level 4.

(2) Reconstruction: The Primary Site is restored and usable again, under limitation of $\text{MTPD} \leq \text{MTTR}$, either at its original site or at another site recovery disaster of the scope of the failure.

a. Scenario I: when the utility outage occurs as assumption on Automatic Failover, the acceptability of MTPD for reconstruction must complied with TIA-942. MTPD of Tier IV is equal to 84 hours; thus the Gen-Set must be perform greater than 84 hours each time DC utility outage occurs. This system reconstruction judges as Severity Level 1-2, as demonstrated in Tab. 1.

b. Scenario II: UPSs will recharge power to the load point immediately after utility outage. A latency time within 2-3 seconds for Flywheel restart Gen-Set, taken around 15 seconds, to charge power back to Flywheel and UPSs + Batteries, as shown in Fig. 8. This system reconstruction is considered as Severity Level 0.

c. Scenario III: Transfer operation back to the Primary Site, when operation managements agree on transfer back for many reasons, e.g. operation costs, capacity, efficiency, regulation, and performance of systems and applications. This system reconstruction is regarded as Severity Level 2-3. It depends on resuming operational readiness on the Primary Site.

<table>
<thead>
<tr>
<th>Severity Levels</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Primary site faults, error, failure on a level of fault tolerance can handle</td>
</tr>
<tr>
<td>1</td>
<td>Primary site failure but not activated secondary site wait for management decision</td>
</tr>
<tr>
<td>2</td>
<td>Primary site failure, Activated secondary site, recovery primary site may takes hours to days</td>
</tr>
<tr>
<td>3</td>
<td>Primary site failed, Activated secondary site, recovery primary site may takes weeks to months</td>
</tr>
<tr>
<td>4</td>
<td>Primary or Secondary site destroyed permanently</td>
</tr>
<tr>
<td>5</td>
<td>All site destroyed permanently</td>
</tr>
</tbody>
</table>

The original level of redundancy system is restored back to normal services on the Primary Site. MTTR is defined as a completed resumption within an acceptable time. The Secondary Site is independent tolerant of another failure but not $\geq$ MTPD, as seen in Fig. 9. For example, as in Fig. 10, when UPS A.1$^{[9]}$ fails during operation, UPS A.2 will take over immediately without any human intervention. Even though, UPS A.2 fails...
at the same time with UPS A.1, UPS B.1 will take over instantaneously, which is called automatic failover. For continuous operation of DC MTTR of UPS A.1 and UPS A.2, must be ≤ MTPD or capability to recharge to system load points at time \( t \) of UPS B.1, as shown in Fig. 10. DC Tire IV parallel connectivity topology is designed for resolving single point of failures (SPOF). N-tier PDS operation degraded architecture was verified and evaluated by chronological theory\(^{[19]}\) as illustrated in Fig. 10. In Fig. 10, static transfer switch (STS) proposes for concurrent maintenance on UPS and automatic failover when UPS is malfunction or failure. During UPS system faults or/ and have errors on battery systems, operation on automatic mode can immediately swap to by-pass mode through STS before UPS system failure.

![Fig. 10. Utility outage affected to DC-PDS](image)

In Fig. 10, \( \triangle G \) could be set up as delay times to comply with local utility reliability, but not over time \( t \) of \(((\text{UPS + Batteries recharged}) - 15\text{seconds})\), that depends on the maximum current load of DC systems.

The advantage of the system failure life cycle model is given the designer and consultant a chance to predict and foresee the results of DC-PDS chronological failure symptoms, which in turn depends on each CIC and SCT degraded stage.

4 Discussion

MTTF of equipment is depended on 4 conditions; first, selected on CIC and designed on SCT of a DC system, second, operation of data center site conditions, third, related risks on daily operation of human activities, and last, procedural maintenance\(^{[16]}\).

*MSEM email for contribution: submit@msem.org.uk*
During the design process, engineers need to understand throughout the metamorphosis of a system failure cycle. The research results are derived from the root cause analysis of each system failure cycle to prepare the preventive action. Severity Levels in Tab. 1 identifies to complex design of preventive action and level of investment to protect the downtime costs.

The research found that a rapid response to failure problems reduces the total cost of downtime losses. Furthermore, skills and knowledge of system operators help to reduce the recursive times of corrective actions. The facility’s monitoring and alarm system provides a real-time reporting (locally and/or remotely) of system failures, which will enable the IT team or operating engineer to troubleshoot the system within a timely manner. By subjecting the critical support systems to real world or “worst case” operating conditions, and exposing problems of failure could affect the reliability and efficiency of data center systems. Comprehensive understanding on the weakness of DC-PDS helps to increase MTTF and reduce MTTR of the system, such as the relationship of operation temperature and USP battery life time\[1, 2, 4\]. Prediction and simulation model assists DC operators to prevent system failures and plan for the right corrective actions\[10, 13, 14\]. Besides, cognitive design on power effectiveness contributes to operation cost and reduce heat emission\[12\].

The testing and commissioning system must be ensured that the critical support systems are operating at optimum levels. Previous researches are more concerned on design processes or best practice guidelines only. This research believes it is useful of system operators and consultants in understanding the system symptoms: fault, error, and failure of metamorphosis life cycle before applying this knowledge to improve their data center system reliability/ availability.

5 Conclusion

A dependable system needs to understand on the levels of components and systems: fault, error, and failure affected. A comprehensive understanding of natural causes of the data center power distribution system failure does improve the data center system availability and reduce the system risk by in-time corrective action. A redundancy system reduces the variance in failure rates. The finding of system failure life cycle and their possible causes allows the proper corrective action and the creation of knowledge as a base for improving system reliability. Creating the corrective action bible, e.g. collaboration with vendors, system integrators, and consultants on life cycle of system failure symptoms: faults, errors, and failures, improves the system reliability/ availability.

References


