3.25 原因不明障害への対応に関する教訓（T25）

教訓 T25

障害原因が不明でも再発予防と発生時対策はできる

問題

A 社の社内共通基盤システムは Web での社内向け／顧客向けの様々な情報公開や各従業員の仮想化 PC のドメイン管理などの機能を保有している。システムの各サーバは仮想化されており、外部連携用の基幹スイッチ、自社内連携用の集約スイッチ、仮想ネットワークスイッチにより機器間の通信を行っている。また、業務の特性上、システム障害発生時に銀行オンラインや航空券予約システムのような高レベルの即時回復は要求されない。

ある朝、業務開始後に物理 PC や社内共通基盤上に構築された仮想化 PC から使用していた業務画面からのレスポンスがなくなる現象が発生し始めた。Web で提供される各種サービスは動作せず、仮想化 PC はすべて動作しなくなった。物理 PC は単独では動作したが、ネットワークに接続する業務は動作せず、外部とのメールも送受信できない状態になった。

社内システム管理者は個別の業務ではなく、ネットワークを経由するシステム全体が動作しない状況からスイッチやストレージなどの基盤系処理装置の障害を疑い、各層のスイッチの調査およびコマンドによる再起動を実施したが業務時間終了までの回復が期待できず、17 時になって当日の業務再開を断念し全社に通知した上で、ハード／ソフト両面からの再調査に着手した。

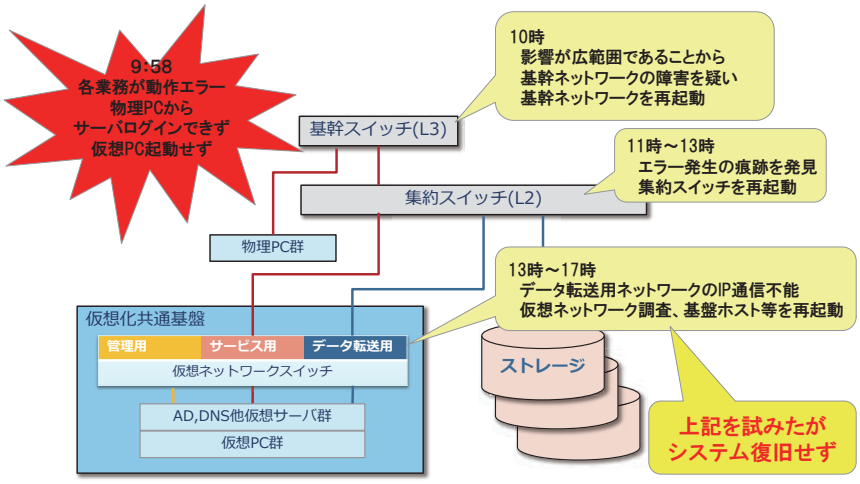

図 3.25－1　障害状況

ハード／ソフト両面からの全面再調査の過程で、この会社に設置していたサーバやスイッチ機器を物理層からすべて再点検した結果、午前中に再起動したはずの集約スイッチの状態表示 LED の点灯状況が通常と異なることが判明した。コマンドによる再起動では状態が改善していないことから、電源コードオフによる物理的な再起動を試みたところ、今回は通信が回復し、当日の 22 時頃にシステムの動作は正常に戻った。

この結果、いずれかのスイッチ機器の動作異常がトラブルの原因であるとの、トラブル発生後の初期調査時の判断が正しかったことが証明されたが、一方でこの障害はコマンドによる再起動では解消しないという別の問題があることも判明した。

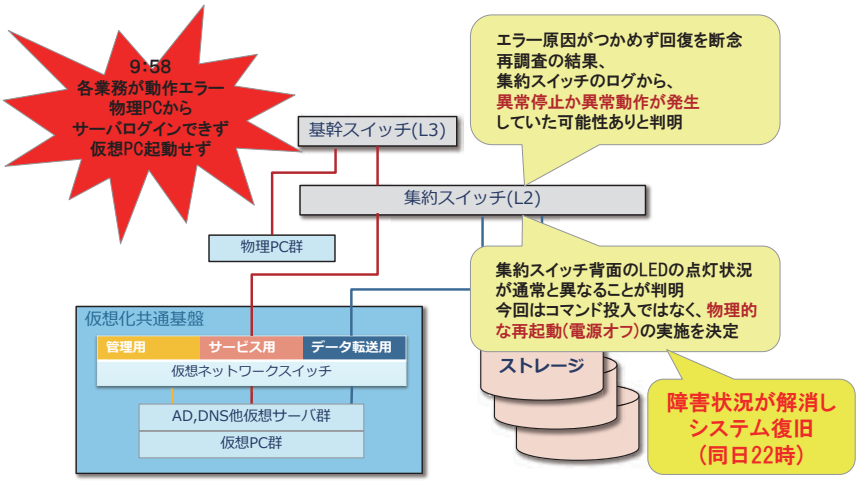

図 3.25－2　障害対応状況

原因

トラブル発生時にスイッチ機器から取得したログを機器の製造元に調査依頼した結果、ファームウェアの障害が原因である可能性が高いとの回答が寄せられたが、以下の理由から発生原因についての確証が得られず、本件の発生原因はこの教訓の作成時点で判明していない。

原因を特定できない理由

(1) 障害発生時のトラフィック（通信内容）の情報を取得できていない

(2) 障害発生の前後の時間帯に、異常が発生したスイッチに障害を示すログが出力されていない

(3) 調査用の情報が限られるため、スイッチ機器開発元でも自社に寄せられた過去の障害事例から原因を類推するまでが限度であり、その後、根本原因が判明したとの報告はない。

対策

上記の理由により発生原因を特定できず、この事例では再発防止に向けた恒久的対応を実施することができなかった。そのため、この会社では次善の策として障害再発リスク軽減のための予防保守【対策 1】と、それでも事象が再発したときの発生時対策の準備【対策 2】をそれぞれ実施した。

【対策 1】 障害再発リスクの軽減

この事例では、確定ではないながらも以下の障害原因が考えられたことから、以下の再発防止策を実施した。

(1) 集約スイッチ機器のハードウェア故障

未知の故障個所が内在している可能性もあるため、予防のために機器の交換を実施。ただし、機器を交換するためにはこのスイッチで通信を制御している社内共通基盤システムを一時的に停止する必要があったため、後日、休日にシステム停止日を設けて交換を実施した。

(2) 集約スイッチのファームウェア障害

当該のスイッチ機器は 1 回 / 年の法定点検日にその時点の最新のファームウェアを適用する運用にしており、障害発生時には最新のファームウェアを適用していなかった。ファームウェアを最新のものに更新していれば今回の障害の発生を予防できたかどうかは不明であるが、今回の障害の調査の過程で集約スイッチの開発元からはファームウェア障害が原因である可能性も報告されたことから、集約スイッチ機器の交換時にあわせてファームウェアを最新のものに更新した。

【対策 2】 障害再発時の対策の準備

このシステムは 365 日 24 時間の稼働を要求されるようなものではないが、トラブル発生時には早期に回復ができないと、回復までの間は社内の業務が停止する影響度がある。【対策 1】の実施により同一原因による障害の発生リスクを軽減できたと考えられるが、原因が判明していない以上、再発の危険性は解消されていないものとして、以下に示す対策を実施した。

(1) 障害原因切り分け基準の決定

今回と同様の事象が発生した際に、スイッチ装置の LED 点灯に異常が発生しているかどうかの判断が対応した要員によって異なることがないように、正常動作時の LED 点灯状況を文書化した。

(2) 障害対応手順の作成

スイッチ装置の動作異常が発生していると判断できた場合に、スイッチの物理的な電源切断から再起動、正常に再起動したかどうかの確認、正常動作しなかった場合の機器の切り離しまでの一連の動作を間違えずに実施できるよう、作業手順を文書化した。

(3) 根本原因調査のための施策の組込み

このシステムで提供する業務も障害発生時には回復最優先であるが、今後もこの障害が多発するリスクをなくすため、障害が今後再発したときに原因調査用に最低限取得するログを選択し、回復処置を実施する前に短時間で確実に取得できる手順を作成した。

(4) 障害対応マニュアルへの追加

上記で作成した (1) から (3) までの各文書を既存の障害対応マニュアルに追加し、印刷して設

備近くに常備することにより、障害発生時にすぐに参照できるようにした。その際には、今回の対策で追加したスイッチ点灯状況の切り分け基準に当てはまらなかった場合や、対応手順どおり実施してもスイッチが正常動作に復旧しなかった場合の対応、スイッチ機器製造元への障害発生の連絡などの対応ルールを追加した。

効果

上記の施策により、この事例では以下の効果を得た。

(1) 同一事象再発の抑止。

予防保守の効果があったかどうかは不明であるが、同一事象は再発していない。

(2) 障害発生時の早期回復

仮に同一の障害が発生した場合には、作成した手順に沿って調査から回復処置までを実施し、早期に業務を回復させられるプロセスを構築できた。

(3) 今後新たなパターンの障害があったときの基本動作の醸成

今後、今回と異なるパターンの未知の障害に直面した場合にも、今回と同様に予防保守の実施、発生時対策の構築により、サービス運用の改善を図る組織風土を醸成できた。

教訓

この事例からは、以下の事項について教訓として活用ができると考えられる。

(1) スイッチ機器がコマンドによる再起動で回復しないときは電源切断を試みる

コマンドによる再起動で正常動作に復帰しない場合には、機器の内部で論理エラーが発生していてコマンドを正常に処理できない状態になったと判断して、電源切断と再投入による機器の回復操作を行うことが必要になる。スイッチ機器のように電源投入用のハードウェアスイッチがない機器の場合には、電源プラグの抜線により上記を代替することが必要になる。スイッチ機器の物理的な電源断が可能な環境では、一見乱暴な手段に見えてもこれらの操作を回復手順に組み込む事が、障害からの早期回復に役立つ。ただし、判断基準と作業手順を文書で周知徹底し、作業を標準化することとペアで実施することが必要である。

(2) システムの運用要件に沿ったリスク軽減策と発生時対策を用意して障害に備える

発生原因不明の障害や原因解明済でも根本対策が実施できない障害の再発リスクがあるときには、とり得る対策を実施した上で、いざ発生したときにどう対応するかを決定して、再発に備える事が必要になる。稼働に対する要求はシステムごとに異なる。再発のリスクが残っても直ちに現場の業務を復旧させる必要があるシステムもあれば、原因調査を綿密に実施して障害を再発させないことが最優先のシステムもある。障害対策の立案においては、対象とするシステムの特性に合わせて実施することが必要になる。

(3) 根本解決のために必要最低限の調査資材を取得する

障害発生時における対応については、即時復旧と再発防止のバランスの考慮が必要である。即時復旧最優先のシステムであっても、原因究明用に取得が必要な必要最小限の調査資材を選択し、

それらを短時間で確実に取得する手段を用意しておくことが、障害の根本原因解明への糸口となり、最終的にシステム稼働品質の向上に結びつく。

この事例では、障害とその対応から得られた経験は相応の対価を払って得た貴重な財産であり、たとえ根本原因がわからず恒久的な解決策が実施できない状況にあったとしても、次回発生への備えとして必ず活用すべきものであることをまとめた。

昨今、システム構築においては自己の開発するアプリケーション以外に多種多様な機器を組み合わせることが多く、それらの機器には未摘出の障害が潜んでいるリスクも大きくなってきた。あらゆるリスクを事前に察知し、リスクの除去や軽減を行うことは事実上不可能である以上、システム運用においては障害再発の予防策と障害発生時の対応策を、そのシステムの要求レベルに応じて実施することが必要である。この教訓では、各システムにおいて上記を実施して障害に備えることの重要性を発信したい。

教訓タイトルは、「障害原因が不明でも再発予防と発生時対策はできる」とした。

なお、ネットワーク機器における未知の障害への予防策として、教訓 T17「長時間連続運転による不安定動作発生の回避には、定期的な再起動も有効！」もあわせて参照を願う。