2.11 システムの運用・保守に関する教訓（G11）

教訓 G11 システムの重要度に応じて運用・保守の体制・作業に濃淡をつけるべし

問題

A 社の社内ワークフローシステムに障害が発生し、連携している顧客向けサービスが終日全面停止した。障害発生後、システム関係者が集まったものの、状況を把握できず、障害個所の特定に多くの時間が掛かった。途中、様々なリカバリ手順を実施したが、なかなか復旧せず、システム停止時間は長時間に及んだ。

本ワークフローシステムの構成図（概要）を以下の図 2.11－1 に示す。また、図中には障害発生時の状況を順番に番号で示している。

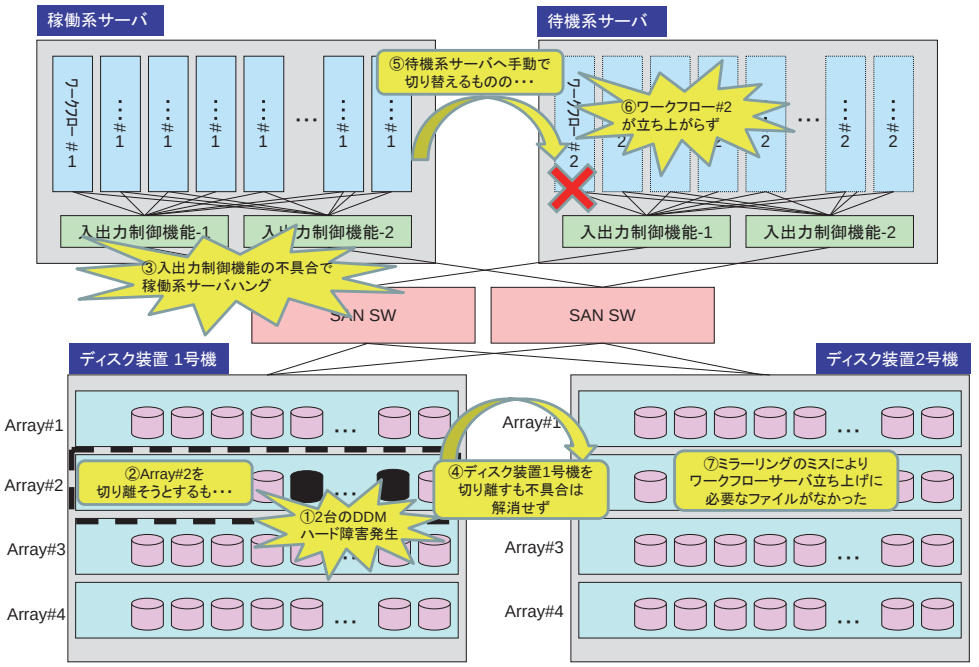

図 2.11－1　システム構成図（概要）

物理的なサーバ装置は、稼働系と待機系の 2 台にて構成されている。なお、この 2 台の物理的なサーバ装置内を入出力制御機能にて論理区画分割することで、複数の論理サーバとして機能させている。

また、ディスク装置は 2 台あり、こちらも二重化されている。各ディスク装置は 4 つのディスクグループ（以下、Array）によって構成されており、それぞれ RAID5[9] で構成されている。1 つの Array に障害が発生した場合には、当該 Array を切り離して稼働を継続する。また、一方のディスク装置全体に障害が発生した際には、当該ディスク装置を切り離して稼働を継続する構成としている。

原因

（直接原因）

本障害の直接の原因は、以下の 3 点であった。ここでは、システムが停止した原因と、復旧に時間が掛かった原因に分けて記載する。

○システムが停止した原因

- DDM（Disk Drive Module）のハード故障

  ディスク装置 1 号機の Array#2 において、ほぼ同時刻に 2 台の DDM が故障した（図 2.11－1①）。RAID5 の仕様により、Array#2 は稼働を停止した。

  なお、今回障害が発生した 2 つの DDM のうち 1 つは、DDM が自己診断機能で異常を検知し、物理的には機能を継続できる状態であるにも関わらず、製品仕様により自ら機能を停止していた。

- 入出力制御機能の不具合

  停止した Array#2 を切り離す過程で、入出力制御機能の製品不具合により、稼働系サーバがハングアップした（図 2.11－1②、③）。稼働系サーバがハングアップした場合、待機系サーバに自動的に切り替わる仕様となっているが、切り替わらなかった。

○復旧に時間が掛かった原因

- ミラーリング[10] の誤設定

  ディスク装置 1 号機をシステム全体から切り離すことで、稼働系サーバのハングアップを解消しようと試みたが、成功しなかった（ディスク装置の切離し自体は成功した）（図 2.11－1④）。次に、稼働系サーバの強制シャットダウンにより、待機系サーバへの切替えを実施したところ、他のサーバは立ち上がったものの、ワークフローサーバだけが立ち上がらなかった（図 2.11－1⑤、⑥）。

  これは、結果的にはディスク装置 1 号機と 2 号機間のミラーリングが誤って設定されていたことにより、ワークフローサーバの稼働に必要なファイルの一部が、ディスク装置 2 号機上に存在しなかったからである。しかし、この原因究明に辿り着くまでに多くの時間を要したため、復旧に時間が掛かってしまった。

（根本原因）

○システムが停止した原因

[9] データ及びパリティを複数のハードディスクに分散することで、信頼性を向上させる技術。
[10] 複数台のハードディスクに、同時に同じ内容を書き込むこと。

- 入出力制御機能については、当該製品の不具合について販売元のベンダも認識しており、他社には修正プログラムが提供されていた（他社でも障害が発生していたため）。A 社は、重大な修正プログラムについては、ベンダから報告を受けることとしていたが、他社で発生していた障害は影響が大きいものではないと判断されたため、ベンダから A 社には修正プログラムの情報が伝わっていなかった。

○復旧に時間が掛かった原因

- ミラーリングの誤設定については、システムの構築時は正しくミラーリングがなされていたものの、その後の運用保守会社（以下、B 社とする）による保守作業において、誤ったミラーリングの設定がなされていた。当該保守作業では当日の作業指示書を作成していたが、記載に誤りがあった。作業指示書は、上位役職者（現場管理者レベル）のチェックを受けるルールとなっており、ルールどおりチェックを受けていたが、上位役職者も誤りを検知することはできなかった。
- 障害発生時に関係者が集まったとき、それぞれのチームが独自の資料を持ち寄ってきて、共通した資料（システム構成図など）を持ち合わせていなかった。それに加え、障害発生時の対応マニュアル等が十分整備されていなかったため、全員で意思疎通を図るのに多くの時間を要した。

対策

本障害に対する再発防止策として、以下の対策を実施した。

○ DDM 保守運用の改善

- DDM の停止を極力低減させる観点から、DDM が自己診断機能で異常を検知した後、自ら機能を停止する条件が見直された DDM の制御プログラムの適用を実施した。
- DDM のハード障害率そのものを低減させていく観点からも、DDM についてはベンダとともに品質評価を毎月実施し、相対的に障害率の高い DDM 製造ベンダや、製造ロットのものについては継続的に予防交換を行うこととした。

○システムの重要度に応じた修正プログラム適用ルールの見直し

- ベンダが A 社へ報告・連携する修正プログラムを選定する際に、これまでは他社での障害における影響の大きさが要否の重要な判断基準となっていた。今後はこうした基準に加え、システムの重要度に応じて、修正プログラムの適用範囲を拡大することとした。システムの冗長化に関する修正プログラムについては、より幅広く抽出するよう見直しを実施した。

○システムの重要度に応じた保守作業におけるチェック体制の見直し

- B 社における上位役職者（現場管理者レベル）によるチェックだけでは誤りを検知できなかったことを踏まえ、システムの重要度に応じて、（B 社の）上長レベル等による二重のチェックを行う体制に変更した。また、これをルールとして定め、作業マニュアルに明記した。

A 社では特に、復旧に多くの時間が掛かり、顧客への影響が大きくなってしまったことに注目し、以下の対策を実施した。

○システムの重要度に応じたシステム保守における運用面、体制面の見直し

- 障害対応メンバ間の意思疎通がしやすいよう、システム共通の資料を準備し、障害対応の迅速化を図ることとした。なお、この資料には、全面障害が発生した場合にまず確認すべきポイント等も記載されている。
- 本番システム稼働後、障害発生時に待機系システムに問題なく切り替わるかどうかを確認するための、実機を使ったシステム障害訓練・テストを行うこととした。今回のミラーリング設定の不備も、実際に実機でテストを行っていれば検知することができたと考えられる。
- 実機を使った訓練・テストに加え、システムのベンダと協業して作成した障害発生のシナリオに基づき、机上での障害対策演習を行うこととした。
- システム障害発生時に、組織内で適切に情報の共有を行うため、システム障害の対策本部を新規に立ち上げた。この組織には、システム企画部門、リスク管理部門、広報部門等の責任者がメンバとして参画している。

効果

上記対策を行ったことにより、当該ワークフローシステムを含め、A 社においてはその後、重大トラブルは発生していない。

教訓

ここでは、復旧に時間が掛かってしまったことを重大な問題と考え、その主たる原因について、改めて考える。

前述の根本原因では、保守作業の内容について、ルールどおりチェックを受けていたにもかかわらず、誤りを発見できなかったことをあげたが、A 社ではシステムによらず、一律にこのチェックのルールを適用していた。また、修正プログラムの適用範囲もシステムによらず一律であったことを考慮すると、一番のポイントは、「システムの重要度に優先順位を付けていなかった」ことにあると言える。

A 社ではその後、基幹システムを中心に、顧客への影響の有無、影響の範囲、推定される損害額など、システムの重要度に応じてランク付けを行い、そのランクに応じてシステム保守対応を行うことにした。このランク付けに基づき、重要なシステムに対して実機を使った障害訓練を定期的に実施したり、修正プログラムを優先して適用したりすることとした。

以上から、今回の障害に横串を通して得られる教訓は、「システムの重要度に応じて運用・保守の体制・作業に濃淡をつけるべし」となる。