3.23 障害監視機能のあり方に関する教訓（T23）

教訓 T23

障害監視は、複数の観点から実装し、障害の見逃しを防げ！

問題

A 社の基幹システムは 24 時間 365 日稼働のオンラインシステムであり、業務の特性から瞬時の停止も許されない。DB サーバは 4 重化されており、それぞれ障害監視機能を持っている。

ある日、DB サーバ 4 台すべてが順番に停止した。

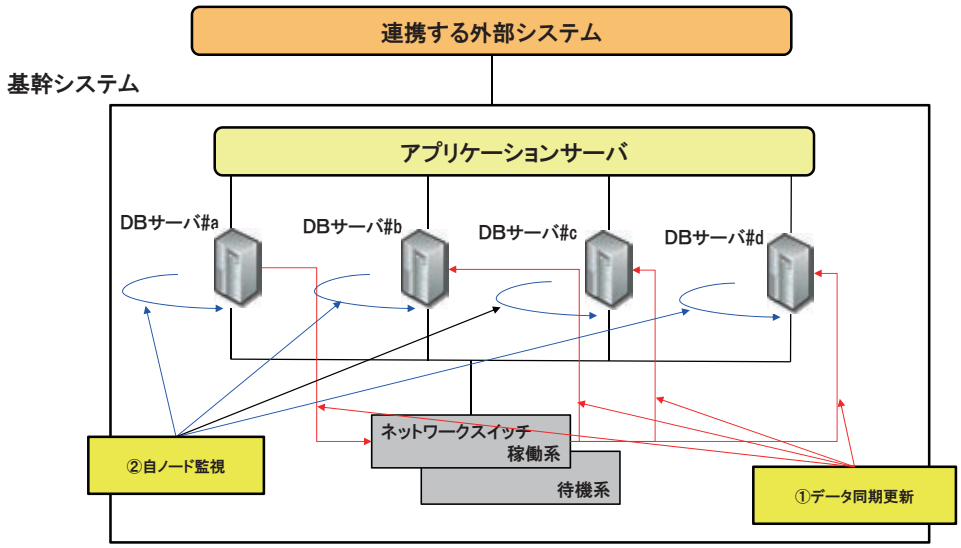

図 3.23－1　DB サーバの障害監視

DB サーバの障害監視には 4 つの機能が用意されていたが、今回の障害に関係した機能は以下の 2 つであった（図 3.23－1）。

（1）データ同期更新

DB サーバは 4 台の物理サーバでデータ同期機能を構成しており、1 台の DB サーバでデータ更新があれば、DB のバッファ・キャッシュと呼ばれるメモリ領域にあるデータをコピーし、ネットワークスイッチを介在して、他の DB サーバにデータ同期更新を実行する。このデータ同期更新処理がタイムアウトした場合、DB ミドルウェアに付随する監視プロセスから監視サーバに通知を行う。監視サーバはこれを受けて、データ同期更新の発信側の DB サーバを停止させる（図 3.23－1 ①）。

(2) 自ノード監視

これは、各 DB サーバが自ら稼働しているかどうかを監視する機能である。各 DB サーバが、サーバ OS の障害保護機能を使って自分自身に SQL を投入し、返信の有り無しで死活を調べる監視を 45 秒ごとに行っている（図 3.23－1 ②）。

原因

サーバ停止の直接の原因は、ネットワークスイッチのキャッシュメモリ故障であった。このスイッチは、本件に関してエラーメッセージを出力しなかったので障害検知ができなかった。以下、全 DB サーバが停止した経緯を示す（図 3.23－2）。

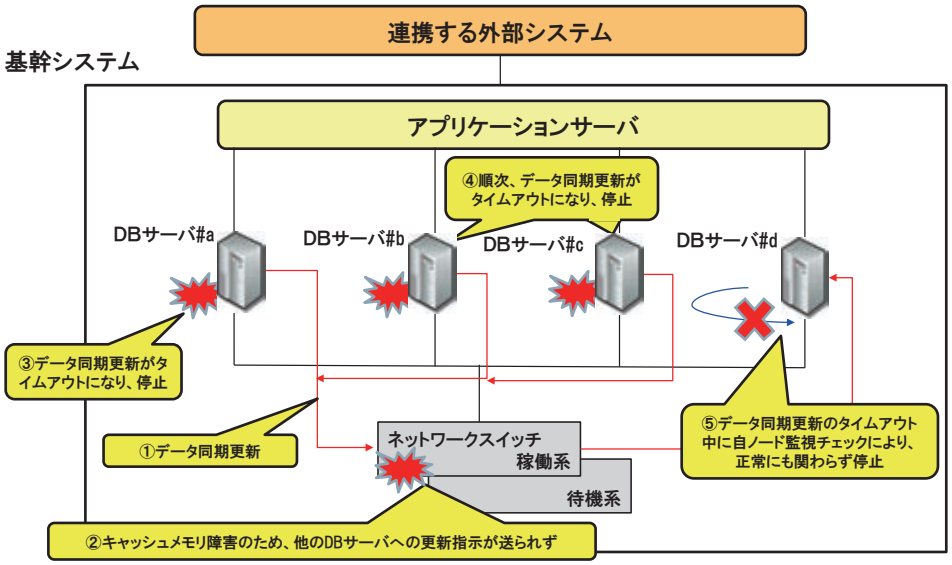

図 3.23－2　障害経緯

DB サーバ #a にデータ更新が行われた際、DB サーバ #a はデータ同期更新機能によって他の DB サーバ 3 台との間でセッションを張り、DB サーバ #a から他の DB サーバ 3 台へデータ更新の送信を行った（図 3.23－2 ①）。この時点でネットワークスイッチが正常動作しておらず他の DB への同期更新が正常に行われなかったため（図 3.23－2 ②）、DB サーバ #a と受信側の DB サーバがタイムアウトになった。監視機能は DB サーバ #a が DB 更新エラーになった通知を受け取り、DB サーバ #a を止めた（図 3.23－2 ③）。DB サーバ #b 、#c でも DB サーバ #a と同様の DB 更新エラーが発生し、順次停止していった（図 3.23－2 ④）。

ネットワークスイッチの障害だけであれば、DB サーバ #a、#b 、#c が停止した時点でデータ同期更新機能は稼働しないため、最後の 1 台の DB サーバ #d で運用が継続できるはずであった。しかし、タイミング悪く DB サーバ #d は、停止する直前の DB サーバ #c のデータ同期更新中に自ノード

監視機能の SQL を投入したため、この SQL がデータ同期更新のタイムアウトに巻き込まれてタイムアウトになってしまい、DB サーバ #d も停止してしまった（図 3.23－2 ⑤）。

このように、2 つの監視機能（データ同期更新、自ノード監視）の実行タイミングの重複によって全 DB サーバが停止してしまった。

このような障害が発生した根本原因は、ネットワークスイッチを含んだ DB サーバの障害監視が適切でなかったことである。以下にその原因を示す。

【原因 1】 ネットワークスイッチの死活監視でつかめないエラーの発生

本来であれば、稼働系スイッチが障害を起こせば、待機系スイッチに切り替えるのだが、今回は、稼働中のネットワークスイッチが完全に動作を停止しておらず、障害メッセージを出力しなかったため、監視機能でエラーを摘出できなかった。

ネットワークスイッチは、「動くのが当然」と思っていたため、設計段階でスイッチの観点でのリスク抽出が十分網羅されておらず、十分な対応策が練られていなかった。そのため、せっかく冗長構成における予備系への切替え機能を用意しながら、実行することができなかった。

【原因 2】 DB サーバの監視機能の不整合

データ同期更新と自ノード監視が同時に起動したときのエラー判断に考慮漏れがあった。このパターンに対応できていれば、ネットワークスイッチの障害（データ同期更新ができない）によって 4 台の DB サーバの内 3 台が停止したとしても、1 台で運用を継続できた。

さらに障害原因を調査すると、A 社のシステム部門では、業務部門から「瞬時の業務停止も許されない」との過酷な要件を課せられていた。点検内容については製造ベンダも含めて設計を行っており、システム部門が必要と判断した保守作業は業務要望よりプライオリティを上げて対応を行っていた。しかし、今回の障害を踏まえ、敢えて以下の点を指摘したい。

【原因 3】 保守（点検作業）不足

ネットワーク機器の点検の実施範囲が最小限に抑え込まれており、十分ではなかった。障害を起こしたネットワークスイッチは、メーカ推奨に基づき、保守点検を行っていたが、電源 off/on、稼働系 / 待機系の入替えなどは実施していなかった。

対策

直接の原因であるネットワークスイッチは、交換することにより対応した。

根本原因としてあげた「ネットワークスイッチを含んだ DB サーバの障害監視が適切でなかった」ことに対する対策をまとめると以下のようになる。

【対策 1】個々の機器の監視を複数のツールで行う

今回のように監視メッセージが出力されないため障害を見逃すことがあることを想定し、複数の観点から監視機能を検討し、実装する。予備機への切替えは、いずれの監視機能でも検知し実行できるように設計する。

【対策 2】複数の監視機能の組み合わせで動作に問題がないか確認する

この障害の対策としては、データ同期監視が完了してから自ノード監視の判定を行うように、自ノード監視機能に、「SQL 投入からの死活監視のデータのタイムアウト待ち時間を延ばす」、「リトライを数回行う」などの対策を行う。これにより、データ同期監視機能と自ノード監視機能の不整合を解消する。さらに、他に同様な問題がないか監視機能の組み合わせ確認を行い、テストを実施する。

【対策 3】十分な保守時間の確保

システム部門は、業務部門との調整を行った上で、システムの安定稼働のための保守時間を確保する。その保守時間を使って、ネットワークスイッチの電源 off/on、パッチの適用などの定期保守点検を十分に行う。さらに、バックアップ切替え確認、正副機器の切替えによる入替えなど、日ごろ待機系になっている機器を動作確認する、あるいは稼働系として使うなど行う。

効果

ネットワークスイッチなどの機器の監視は見逃される傾向があるため、そのスイッチが障害になると重大な影響を及ぼす。システム部門は、ネットワークスイッチを含めた複数の監視方法を実装し、監視漏れがないように、またそれぞれの監視機能に矛盾が生じないようにすることで、障害を減らすことができる。

また、業務部門は「瞬時の業務停止も許さない」運用を行うことを求める傾向があるが、今回のような事例を考えれば、システム部門は、十分な保守時間の確保を業務部門に提案することができる。これにより、システム障害が減り、トータルのサービス提供時間が増えることになる。

教訓

ネットワーク機器が高度になり、接続される機器も増えている中、システム障害対策も複雑になっている。そのため、監視技術が重要度を増している一方で、監視の不具合による障害も増えている。

この事例は、監視ミス（自ノード監視がエラーでないものをエラーとした）と監視漏れ（ネットワークスイッチのエラーを見過ごした）が同時発生した事例である。

教訓は、「障害監視は、複数の観点から実装し、障害の見逃しを防げ！」とした。