止まらないシステムにしたいなら、止まらないストレージを選べ

ビジネスのデジタル化が進むにつれ、サービス停止は企業にとって致命傷になりかねない。システム障害はなんとしてでも防ぎたいところだ。どの現場でも、あらゆる要素を冗長化して障害が発生しても乗り切れるように可用性を高めている。

それでも何度かシステム障害のニュースを目にする。近年金融インフラ周辺で起きたシステム障害に目を向けると、意外なところで共通点が見いだせる。盲点と言えるかもしれない。それがストレージだ。より正確に言うと、冗長構成を取り可用性が高いはずのストレージが障害時のフェイルオーバーに失敗し、サービス停止を引き起こしているのだ。止まらないストレージを選ぶには？

2021年3月12日、ある大手銀行にて外貨建て送金約300件で遅延が生じた。3月12日付送金とするには当日午後1時までに送金処理を終えなくてはならないところ、最終的に処理が終了したのは当日午後7時過ぎ。こちらは企業間取引が中心となるため、銀行の信用問題に影響を及ぼす障害へと発展してしまった。システム障害の起因となったのは、統合ファイル授受基盤にあるストレージのハードウェア故障だ。

もちろん障害を想定した準備はしていた。ディスク装置内で故障が発生すると、バックアップ機器(スタンバイコントローラ)へ自動切り替えされるはずだった。しかし当日はフェイルオーバーがうまく機能せず、手動での強制切り替えも機能せず、装置交換やシステム再起動で時間がかかり、結果的に送金業務に遅延が生じてしまった。

2020年10月1日には、東京証券取引所にて株式売買システム「arrowhead」で障害が発生した。メモリ障害を発端として、NASの制御機構がダウン。予備機に自動切り替えが行われず、NAS全体が使用不能の状態となった。

こちらもストレージ(NAS)の切り替え失敗が大きく影響している。本来は故障時でも30秒以内に切り替え、業務継続できることを要件に定めていた。しかし実際はメモリ障害時に自動で切り替わらない仕様となっていた。現場では自動切り替えが進むことを前提として障害対応手順を整備していたため、復旧に手間取り、結果的に終日取引停止という事態に陥ってしまった。

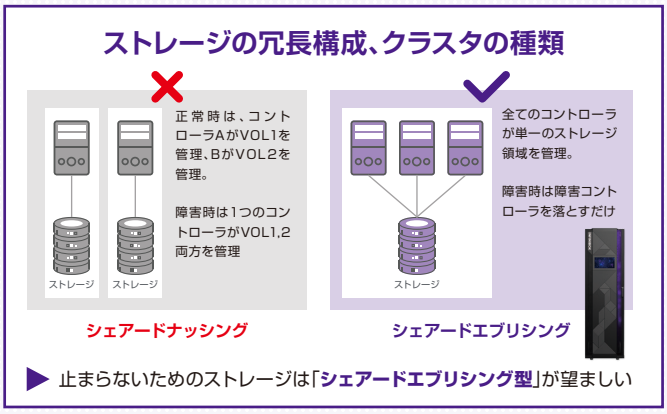
もう1つ、ストレージが関連する深刻な障害を挙げておこう。2019年12月4月、自治体クラウドにて自治体クラウド基盤「Jip-Base」で障害が発生した。サービス停止で47自治体のシステムがダウンし、自治体業務や住民サービスに影響を及ぼした。こちらはストレージ機器のファームウェア不具合が発端だったものの、バックアップにも問題がありデータのリストアに失敗したためデータを一部消失する事態となった。

「これらは旧式のストレージ構造に起因する問題です」とInfinidat Japan東一欣氏は指摘する。

■ シェアードナッシング型ストレージは何が問題か？

旧式とはシェアードナッシング型を指す。ストレージシステムは制御を行うコントローラーと記憶装置のストレージからなり、コントローラーがストレージを共有しないのがシェアードナッシング型となる。そのためコントローラーごとにストレージがぶら下がるような形になる。一方、近年ではシェアードエブリシング型が増えてきている。こちらは複数のコントローラーがストレージを共有する形となる。

コントローラーごとにストレージを分けるべきか、コントローラー全体でストレージを共有すべきか。どちらがいいかと考えるよりは、現時点でどちらのタイプの製品を選んだほうがより安全か、より障害に耐えうるかと考えるといいだろう。



なぜ多くのベンダー製品がシェアードナッシング型を選んできたのだろうか。東氏は排他制御を挙げる。シェアードエブリシング型のようにコントローラーがストレージを共有するとなると、大きなストレージの塊に複数のコントローラーからの手が伸びてきて一斉に書き込むようなイメージとになる。そのため排他制御が難しくなり、性能を高める阻害要因にもなりうる。それならコントローラーごとにストレージを分けてしまえばいいという発想になる。排他制御の観点からするとシンプルになり安易に実装できるため、旧式のストレージや200x年代のストレージのベンチャーの多くはシェアードナッシング型を採用してきた。

ただしCPU、メモリ、I/Oインターフェース性能が高まり排他制御で生じるデメリットはそう大きくなくなってきた。それよりもシェアードナッシング型にしていると、コントローラーの一部に障害が発生すると、残りのコントローラーの負荷が高くなる。性能で縮退するリスクになる。データ量が増え、ストレージ全体のサイズが巨大化している近年では何よりもコントローラーを切り替える処理が複雑で長く難しくなる。下手すれば切り替え失敗という最悪の事態も起こりうる。

冒頭に述べたような、ストレージのフェイルオーバーで失敗したケースはシェアードナッシング型のストレージだ。障害に備えて予備機を用意していたとしても、切り替えが正常にできるかどうかが難しく、リスクを抱えることになる。切り替えができたとしても、ストレージにあるデータを読み込みなおすところで時間がかかることも想定しなくてはならない。

■ Infinidatはシェアードエブリシング型で安定性と耐障害性を高める

こうしたシェアードナッシング型ストレージが抱える課題を解決するのがInfinidatのストレージだ。同社はシェアードナッシング型アーキテクチャの課題解決のために創業したと言っても過言ではない。長らくストレージのアーキテクチャデザインをしてきた創業者のモシェ・ヤナイ氏は、高速・大容量の時代に合わせシェアードエブリシング型に着目し、安定性と効率性を追求した設計を描いた。

繰り返しとなるが、Infinidatのストレージはシェアードエブリシング型だ。コントローラーは3台で固定。全部アクティブで機能する。同じシェアードエブリシング型のストレージ製品やHCI製品の中にはスケールアウトと称し拡張性を高めるためにコントローラー数を可変にしているものもある。しかし可変では安定性を阻害する要因にもなりかねない。そのためInfinidatではコントローラー数は固定とした。

シェアードエブリシング型なので、当然ながら全てのストレージのボリュームやファイルシステムに常時アクセスできる。コントローラー切り替えがスムーズにできるため、障害時の縮退度合いを最小限にできる。また障害時はノードを落としてマルチパスを減らしたりIPアドレスを移せばいいだけとなる。

階層化を極力排除しているところも特徴的だ。ストレージは記憶媒体により性能とコストが異なるため、複数の記憶媒体を混ぜることで最適化を図ることが可能になる。類似製品のなかにはデータの使用頻度に応じてSSDとHDDの間でデータを移動させるものもある。これだと複雑性を高めることになり、リスク要因になりかねない。InfinidatはDRAMとHDDをメインで使い、SSDは補完的なリードキャッシュとしてのみに用いている。これが階層化を最小限にして安定化につなげている。

細かいところだと、Infinidatでは「14Data+2Parity」を単位として、最大480本のディスクにデータを分散する。これは全てのディスクに散らして書き込むことで、障害に強く、リビルドも早くすむようになっている。

オールフラッシュストレージとの比較も補足しておこう。オールフラッシュはその名の通り、全てSSDなどのNANDフラッシュで構成したストレージだ。性能が魅力とされるが、SSDゆえの経年劣化や偏摩耗が起こりうる。アプリケーションのI/Oの単位と肥大化するNANDフラッシュのP/E(プログラム/イレース)の単位のミスマッチによりガベージコレクションが常態化し、性能劣化やプチフリーズが起きる製品もあるため、注意が必要だ。

これまで述べてきたように、Infinidatはシェアードナッシング型が抱える現状のリスクを排して、効率化と安定性を追求したストレージとなっている。コントローラー3台アクティブのシェアードエブリシング型、階層化は最小限にするなどして安定性を高めている。シェアードエブリシング型で重要になる排他制御に関しては独自の作り込みで性能と安定性を高めている。

これから企業を支えるシステムではサービスを止めないことがとても重要になる。冒頭に挙げたような社会インフラとなる重要なサービスになると、経営陣の引責辞任に発展する場合もある。システム障害で誰が責任をとるべきかは議論の余地があるものの、デジタル化が進んだ昨今ではサービス停止が起きると影響は計り知れない。

止まらないシステムを構築するにはストレージの安定性や耐障害性がとても重要になる。オンプレではなくクラウドを利用するとしてもストレージの重要性は変わらない。ストレージの安定性や耐障害性は注意深く確認しておこう。

また、ストレージ・コンピュート・ネットワークを含めた全体的な可用性の観点でクラウドに着目したい。外資メガクラウドなど多くのパブリッククラウドの目標稼働率は99.99%(年間に約1時間の停止)程度であるが、メガクラウド上でマルチAZ(アベイラビリティ ゾーン)の冗長構成を取っているはずのペイメントアプリやオンラインゲームなどで、より長時間のサービス停止が多発している。

対してオンプレ、特にストレージでは何十年も前から99.999%(ファイブナイン、年間5分の停止)、99.9999%(シックスナイン、年間30秒の停止)の目標稼働率が当たり前であり、さらにInfinidatはより厳しい99.99999%(セブンナイン、年間3秒の停止)の目標稼働率で設計されている。加えて2つの筐体を並べた構成により、100%のデータ可用性が保証されている。

クラウドのなかにはストレージにInfinidatのテクノロジーを採用しているパブリッククラウドもある。なかでもNeutrix Cloud JapanはInfinidatを採用しており、日本企業が運営している。データを安定した環境で、国内にとどめておけることを意味する(大手パブリッククラウドでは日本のリージョンを選んだとしてもメタデータは海外に流出している可能性がある)。

Infinidat 東氏は**「企業にとってデータは何よりも重要です。止まらないストレージ、止まらないクラウドを選びましょう」**と呼びかける。

INFINIDAT
The Standard in Enterprise Storage ｜ jp_sales@INFINIDAT.com ｜ 03-4243-6343
© INFINIDAT 20241201