高可用化ミドルソフト HA Booster の開発

市川　正也†，春名　高明†，二瀬　健太†，真矢　譲†，三瓶　英智‡
(株)日立製作所システム開発研究所†　(株)日立製作所ソフトウェア事業部‡

1．はじめに

近年，ネットワークの普及に伴い，計算機システムを用いて提供されるサービスが多様化しており，オンラインサービスも 24 時間連続運転が求められている。このようなシステムでは，システムダウンを防止し，障害が発生しても直ちに処理を継続できる可用性が要求される[1]。

本研究では，ホットスタンバイ方式において，系切り替え時間を大幅に短縮する共有ディスク切り替え方式を開発した。

2．ホットスタンバイ方式

2.1 システム構成

対象とするシステム構成を**図 1** に示す。これは，現用系とこれをバックアップする待機系，及び共有ディスクから構成される。現用系と待機系はそれぞれ，高可用化ミドルソフト(HA Booster)を搭載する。HA Booster は，系切り替え処理時に共有ディスクへのアクセスを制御する。

2.2 従来方式

従来方式は，デバイスドライバが有するディスクの活性化，及び非活性化の機能を用いて共有ディスクを切り替えていた。これは，現用系のみデバイスドライバがディスクへの読み書き可能な状態(活性状態)に設定し，待機系はデバイスドライバがディスクへの読み書き不可能な状態(非活性状態)に設定しておき，系切り替えの際に VG の状態を再設定する。非活性状態から活性状態に再設定する際に，ディスク構成情報を取得する IO が発生し，ディスク数に比例して切り替え時間が増加する。

2.3 提案方式の概要

ここでは，上記の問題点を解決するために，共有ディスクの高速切り替え技術として，制御グループ方式と擬似オフライン方式を提案する。

2.3.1 制御グループ方式

共有ディスクへのアクセス制御を行う単位として，「制御グループ」という概念を導入する。この制御グループは，1 つまたは複数のボリュームグループ(VG)から構成される集合である(**図 1** 右側)。各制御グループは，共有ディスクへのアクセスの可否を設定できる。

2.3.2 擬似オフライン方式

システム起動時に現用系と待機系は共に共有ディスクを活性状態に設定し，以下の処理によりデバイスドライバへのアクセスを制御する。

本方式の概要を**図 2** に示す。HA Booster は，デバイスドライバの処理ルーチンと同一のインタフェースを持つアクセス制御判定ルーチン(以降，

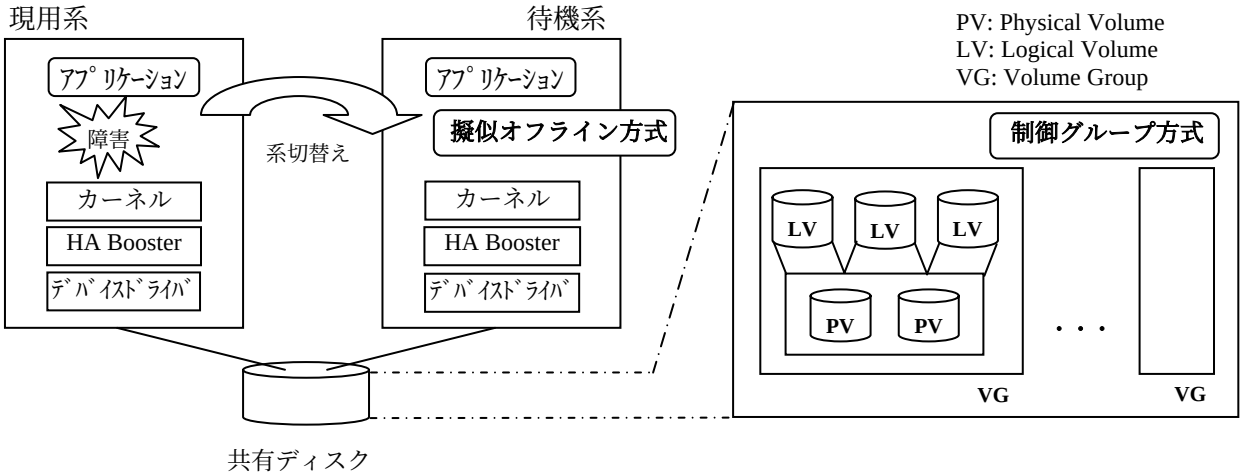

図 1 システム構成図

A Development of High-Availability Middle Software HA Booster
Masaya ICHIKAWA†, Takaaki HARUNA†, Kenta NINOSE†, Yuzuru MAYA† and Hideaki SANPEI‡.
† Hitachi, Ltd., Systems Development Laboratory.
‡ Hitachi, Ltd., Software Division.

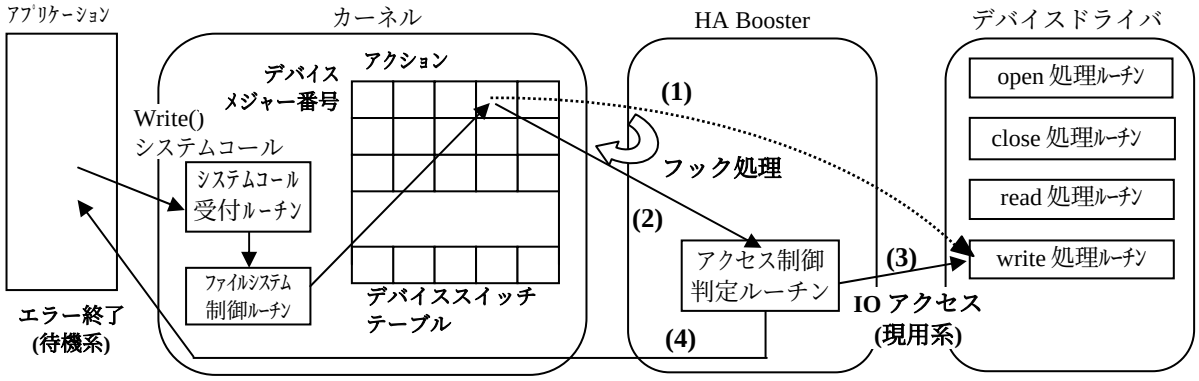

図 2 擬似オフライン方式

判定ルーチン)と，各制御グループに属するディスクへのアクセスの可否を表すアクセス制御状態フラグ(以降，状態フラグ)を有する。

HA Booster は，システム起動時に，OS が持つデバイススイッチテーブルを書き換え，判定ルーチンを登録する(**図 2** の(1))。デバイススイッチテーブルとは，各デバイスに対し OS が割り当てたデバイスメジャー番号と，当該デバイスに対するアクション(open, close, read, write 等の IO の種類)を実施する処理ルーチンのアドレスを対応付けたテーブルである。この書き換え処理により，OS やアプリケーションから VG へ IO 要求が発生した時に，カーネルは判定ルーチンを呼び出す(同(2))。

判定ルーチンは，VG が所属する制御グループの状態フラグを参照し，これがアクセス許可を示す場合(現用系の場合)，デバイスドライバの処理ルーチンを呼び出す(同(3))。一方，このフラグがアクセス禁止を示している場合(待機系の場合)，デバイスドライバを呼び出さず，アプリケーションにエラー終了を通知する(同(4))。

提案方式と従来方式の特徴を**表 1** に示す。本方式により，VG を活性状態に保ったままアクセス制御が可能となるため，VG の活性化処理で発生していた IO 処理が不要となる。また，制御グループの状態フラグの書き換える処理のみで，制御グループに属する VG のアクセス制御を一括して実施できるため，従来方式に比べ系切り替え時間を大幅に短縮できる。

表1　方式比較

方式名	従来方式	提案方式
現用(障害)系の VG 非活性化処理	要	不要
待機(新現用)系の VG 活性化処理	要	不要
制御グループ単位の一括制御	無	有

3．評価と考察

提案方式と従来方式の共有ディスク切り替え時間の実測結果を**図 3** に示す。

(1) 提案方式

この切り替え時間は VG 数が増加しても、数十 m 秒程度と一定である。これは、擬似オフライン方式のシステム立ち上げ時の VG 活性化処理によりディスク切り替え時の VG 活性化処理が不要となり，切り替え時間が短縮できたこと，およびグループ制御方式の VG 一括切り替えにより切り替え時間が VG 数に依存しないためである。

(2) 従来方式

VG 毎に，現用系の VG 非活性処理と待機系の VG 活性化処理を行うため，切り替え時間は VG 当り約 1.5 秒の割合で増加する。

4．おわりに

本論文では，高速ディスク切り替え方式として，擬似オフライン方式とグループ制御方式を提案した。この結果，提案方式は共有ディスクの切り替え時間を VG 数に比例することなく，数十 m 秒に抑えられる。

[参考文献]

[1] 内藤祥雄,他:高信頼UNIXｼｽﾃﾑ;ﾏｸﾞﾛｳﾋﾙ,1994

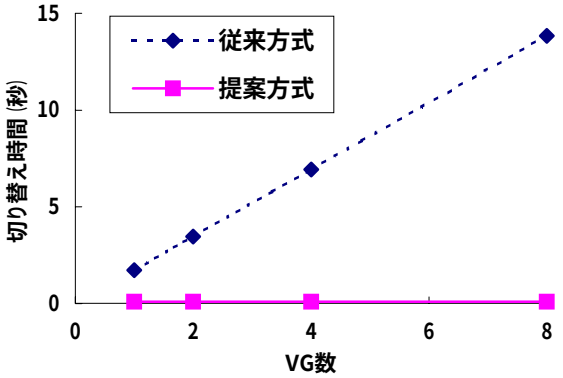

図 3 共有ディスクの切り替え時間