B-TREEフル連想型大容量キャッシュ技術

- 先進のキャッシュメカニズム -
- 専用装置(Super CACHE/Super SSD)の紹介 -

“ストレージ・ソリューションのリーディング・プロバイダ”　コアマイクロシステムズ株式会社

Rev-2010-04-15B

アジェンダ

[1] 製品開発の動機
[2] 技術コンセプトとアイデア
[3] 従来キャッシュ方式の課題
[4] B+TREEキャッシュ検索による解決
[5] 技術特徴/技術仕様の紹介
[6] 性能特性
[7] 効果
[8] まとめ（***Super CACHE/Super SSD***概略仕様）

製品開発の動機

○ 背景

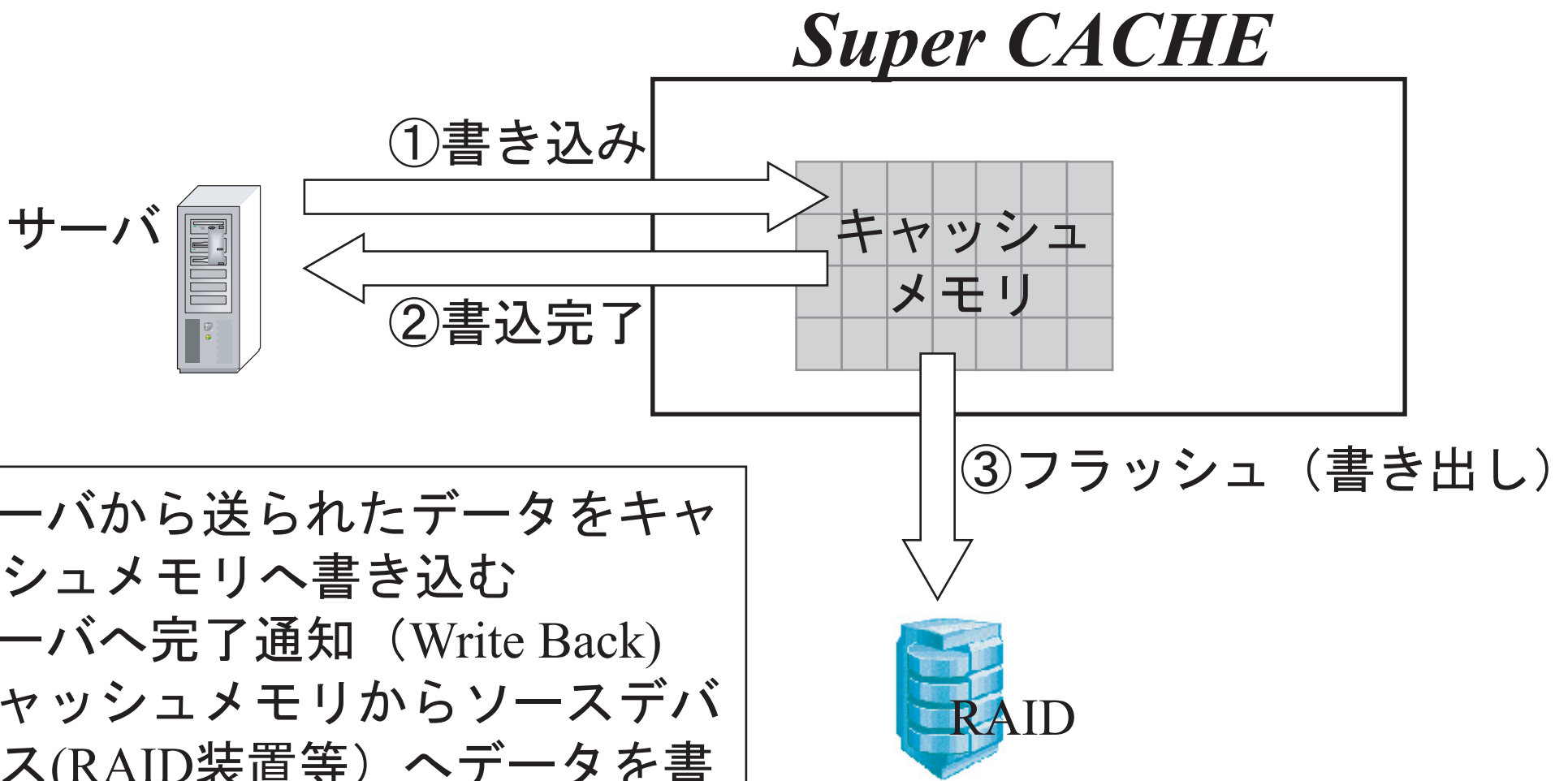
・DRAMベース半導体ディスクを開発。
・「***Solid STOR***」として2004年より販売開始。
・データベースアプリケーション、エンジニアリング用途へ導入。

○ DRAMベース半導体ディスクの課題

・大容量アプリケーションへの適応時にコスト高。
・容量が限定、ホットデータ切り分け作業が必要。
・容量が制限されない半導体ディスクがほしい。

➡ 大容量キャッシュ装置の開発

キャッシュと書き込みデータの流れ



①サーバから送られたデータをキャッシュメモリへ書き込む
②サーバへ完了通知（Write Back)
③キャッシュメモリからソースデバイス(RAID装置等）へデータを書き込む

技術コンセプトとアイデア

○ キャッシュの大容量化

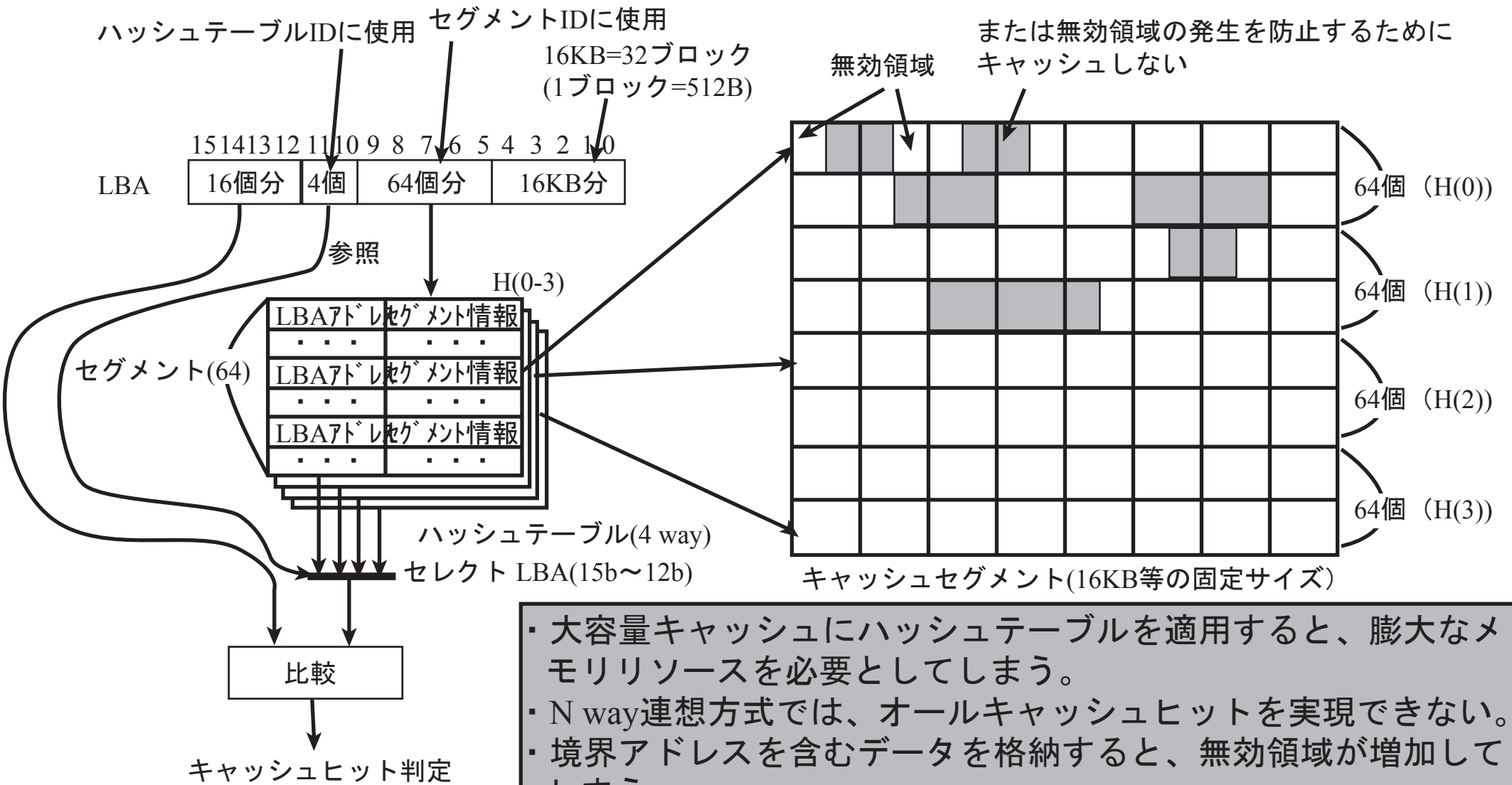
アプリケーションは、書き込んだデータまたは更新したデータを参照するはずである。また、ストレージサイズが大容量(Ex. 数10TB)でも、頻繁にアクセスされるホットデータは限定(Ex. 数100GB)されるはずである。よって、アプリケーションが書き込むまたは更新参照するホットデータサイズを越えるキャッシュ容量を実現すれば…

- オールキャッシュヒット
- 大容量の場合に非常に高価なDRAM SSDと等価な半導体ストレージを低価格で提供可能

○ 課題

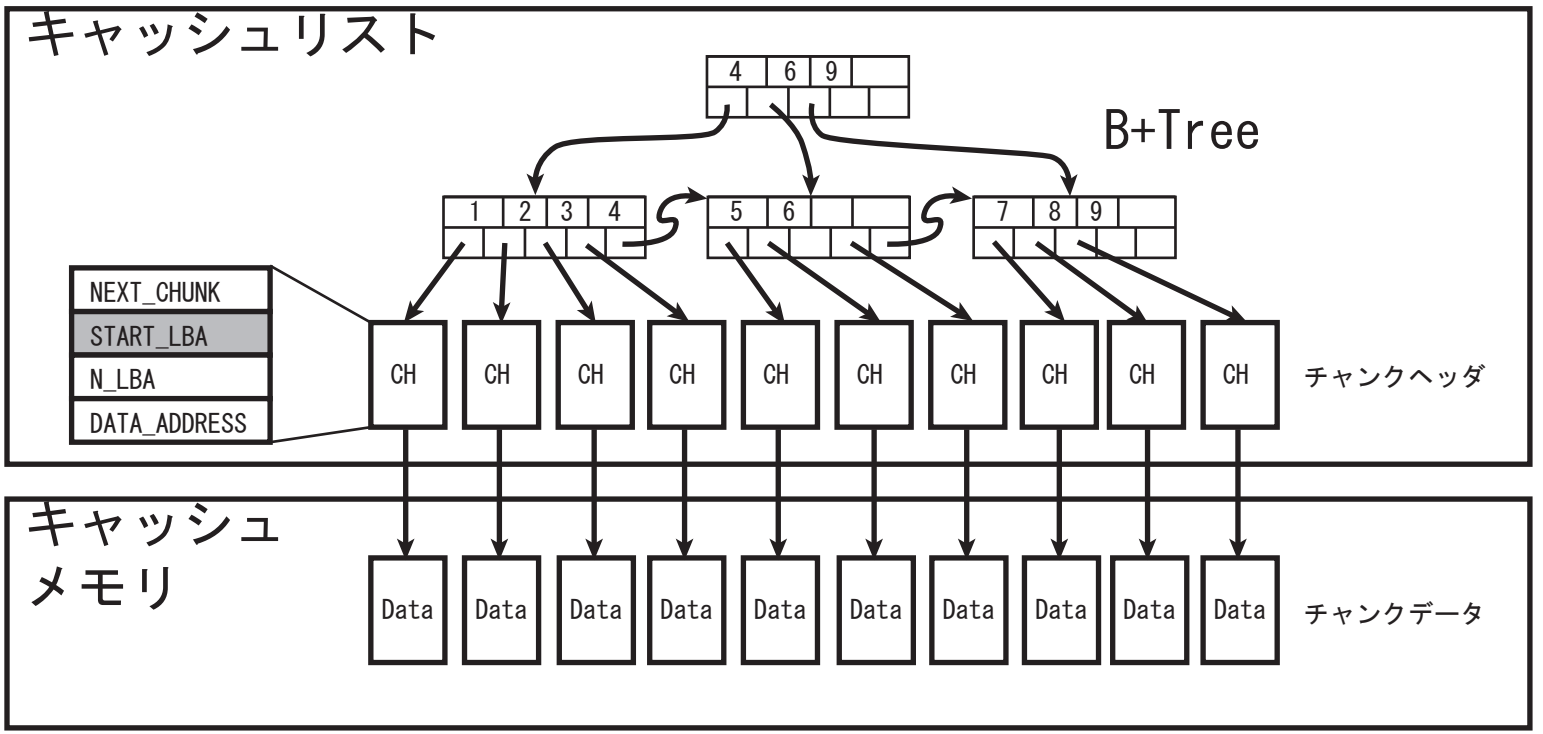
- – オールヒットを実現するキャッシュメカニズム

従来キャッシュ方式の問題点



- 大容量キャッシュにハッシュテーブルを適用すると、膨大なメモリリソースを必要としてしまう。
- N way連想方式では、オールキャッシュヒットを実現できない。
- 境界アドレスを含むデータを格納すると、無効領域が増加してしまう。

B+Treeによるキャッシュエントリ検索



・フル連想を実現。-> オールキャッシュヒット化が可能。
・データ管理上境界アドレスを持たない。
・　-> オールキャッシュヒット及びキャッシュメモリの有効利用が可能。

Super CACHEの特徴(1)

項目	特徴
大容量キャッシュ	・B+TREE検索によりフル連想大容量(数100GB)キャッシュを実現。 ・高速キャッシュエントリ検索(1μs以下のオーバヘッド)。 ・フル連想方式により、キャッシュ容量の有効利用性を向上。（対nWay連想方式） ・キャッシュデータのインデキシング方式によりキャッシュメモリのアドレス境界制限をなくし、キャッシュ容量の効率を向上。
オブジェクト志向なキャッシュ管理	・個々のソースデバイス（ストレージ装置/LUN）に対して、個々のキャッシュの作成が可能。 ・各キャッシュごとに異なる構成（容量、パラメータ）が可能。
キャッシュデータの常駐	・キャッシュフルになるまでキャッシュデータをパージしないことで、容量有効性を向上。これによりキャッシュヒット率を向上。
Dirtyデータの高速フラッシュ	・フロントIO優先のバックグランド先行フラッシュ機能。 ・連続アドレスデータをまとめ書きする高効率フラッシュ方式。 ・ランダムライトが発生するアプリケーション用には、ダーティデータをLBA(*1)順にフラッシュするLBAソートフラッシュ機能。

*1) LBA : Logical Block Address

Super CACHEの特徴(2)

項目	特徴
変更可能なキャッシュチャンクサイズ	・アプリケーションのアクセスパターンに従って、キャッシュデータ管理サイズを4KB、8KB、16KB、32KB、64KBへ変更可能。 ・チャンクサイズより小さいライトデータに対しても、レスキュー機能を提供（チャンク閾値：パラメータ）。
リードアヘッド	・Enable/Disableの変更が可能（アプリケーションによっては、リードアヘッドが逆効果の場合がある）。
リードムーブインディスエーブル	・リードアクセスデータのキャッシュへのムーブインを不可にする機能（フラッシュSSD時に利用）。
スタティスティックス	・チャンクカウンター：使用、未使用、Dirty、Clean数を表示。 ・ヒットカウンター：キャッシュヒット率、ミス率、ギブアップムーブイン数、パーシャルデータチャンク数を表示。 ・IOPS/Throughput：ホスト側、ソースデバイス側の性能を表示。 ・IOアクセスサイズ分散：ホストからIO及びソースデバイスへのIOサイズの分散を表示。 ・IOアクセスLBA分散：IOアクセスのアドレス分散を表示。

Super CACHEの特徴(3)

項目	特徴
キャッシュア リプレースメント ルゴリズム	・LRU(Least Recently Used)。 ・LRUアルゴリズムは、キャッシュサイズを超えるレンジで、毎回一通りのデータを巡回するようなIOアクセスに対して、オールキャッシュミスを引き起こしてしまう弱点を持っています。 ・LIRS(Low InterReference Recency Set)、LRUのその弱点を補うアルゴリズムです。LIRSは、（キャッシュサイズ）／（アクセスレンジ(アクセスデータ量））のキャッシュヒット率を期待できます。 ・キャッシュごとに選択できます。
フラッシュ アルゴリズム(2)	・Dirty&Cleanデータマージドフラッシュ。DirtyデータとCleanデータが連続アドレス上に混在している場合は、Cleanデータも一緒にマージしてフラッシュすることでフラッシュ効率を向上。 ・また、どの程度(割合)のCleanデータもマージしてフラッシュするかパラメータにより調整可能。
デバイス・ホスト間 プロビジョニング	・ソースデバイス（ストレージデバイス）とホスト間の接続可否(No Access/Read only/Read&Write)の設定が可能。

＊キャッシュ制御方法に関して、特許出願済み。

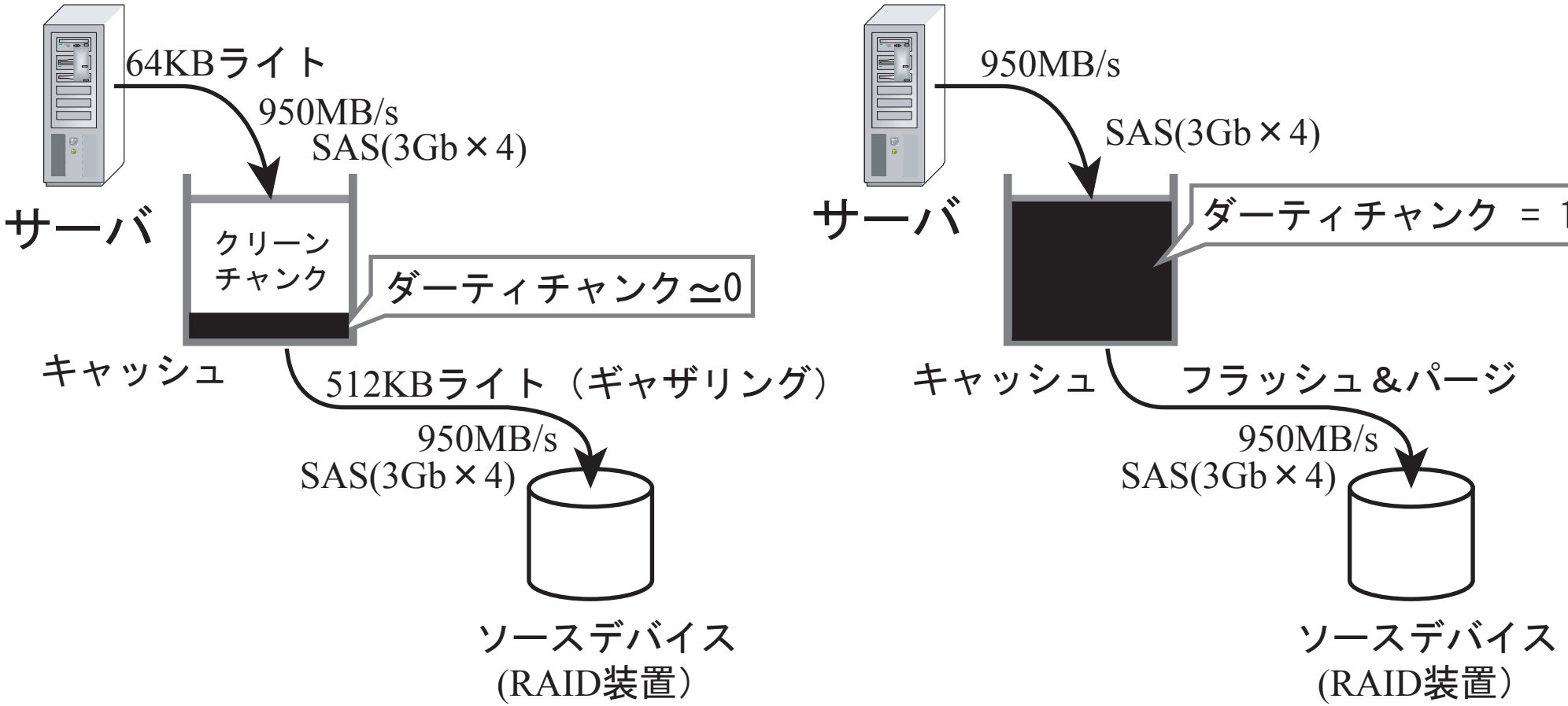
キャッシュヒット時のランダムアクセス性能

Chunkサイズ=8KB

アクセス方法	アクセスサイズ	キャッシュヒットライト*1		キャッシュヒットリード	
マルチ I/F アクセス FC(4Gb) x 8	512B	221,670 IOPS	108 MB/s	301,500 IOPS	147 MB/s
	4KB	211,700 IOPS	827 MB/s	279,800 IOPS	1,093 MB/s
	8KB	198,900 IOPS	1,554MB/s	260,890 IOPS	2,038 MB/s
	64KB	43,680 IOPS	2,730 MB/s	47,360 IOPS	2,960 MB/s
	512KB	5,700 IOPS	2,850 MB/s	6,200 IOPS	3,100 MB/s
	512KB	4,085MB/s (リード&ライト)			
シングル I/F アクセス FC(4Gb) x 1	512B	51,500 IOPS	25 MB/s	73,500 IOPS	36 MB/s
	4KB	36,600 IOPS	143 MB/s	50,880 IOPS	198 MB/s
	8KB	27,400 IOPS	214 MB/s	32,620 IOPS	254 MB/s
シングル I/F アクセス Wide SAS(3Gb x 4) x 1	64KB	15,250 IOPS	953 MB/s	16,850 IOPS	1,055 MB/s
	512KB	2,060 IOPS	1,030 MB/s	2,150 IOPS	1,075 MB/s
	512KB	1,155MB/s (リード&ライト)			

*1) キャッシュミス時の空きチャンクまたはクリーンチャンクありケースも同性能。

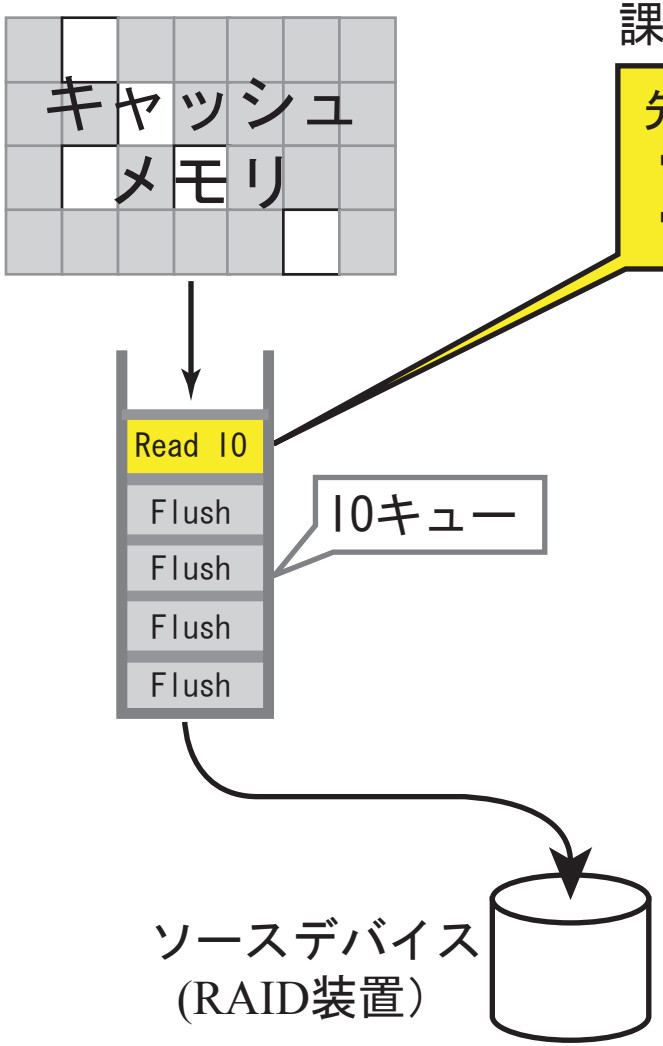
データフラッシュの振る舞い



データギャザリングによるまとめ書きにより高効率化

キャッシュフルでもソースデバイスから性能劣化無し

遅延フラッシュ機能（フロントIO優先）



課題

先行フラッシュ実行中にキャッシュミスリードが発生
- ->フラッシュIOがフロントIOを妨害
- ->リードIOレスポンスが大

対策①：遅延フラッシュ

ソースデバイスへのIO発生時は、先行フラッシュ停止

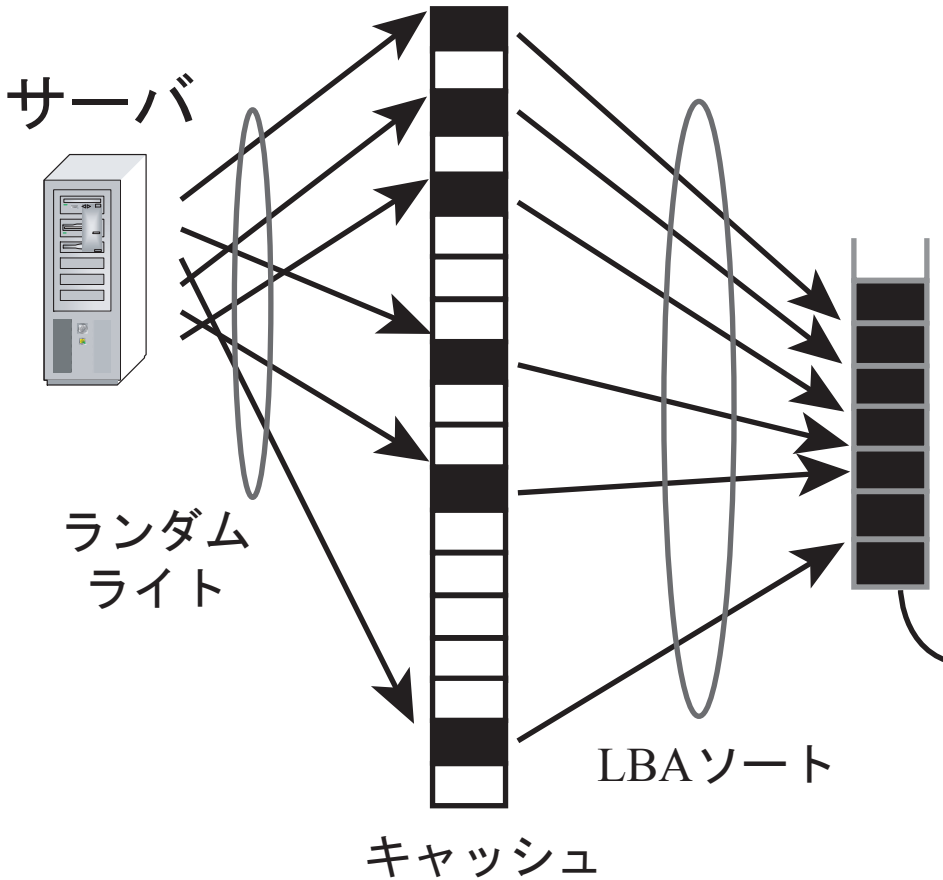
Background Flush : DELAYED IO
Flush Delay Timer : 5000 mili seconds
Delay timer reset threshold : 5 IOs

対策②：スケジュールフラッシュ

低稼働時間帯に先行フラッシュを実行

Flush start time : 02:00
Flush stop time : 05:00

LBAオーダーによるフラッシュの効率化



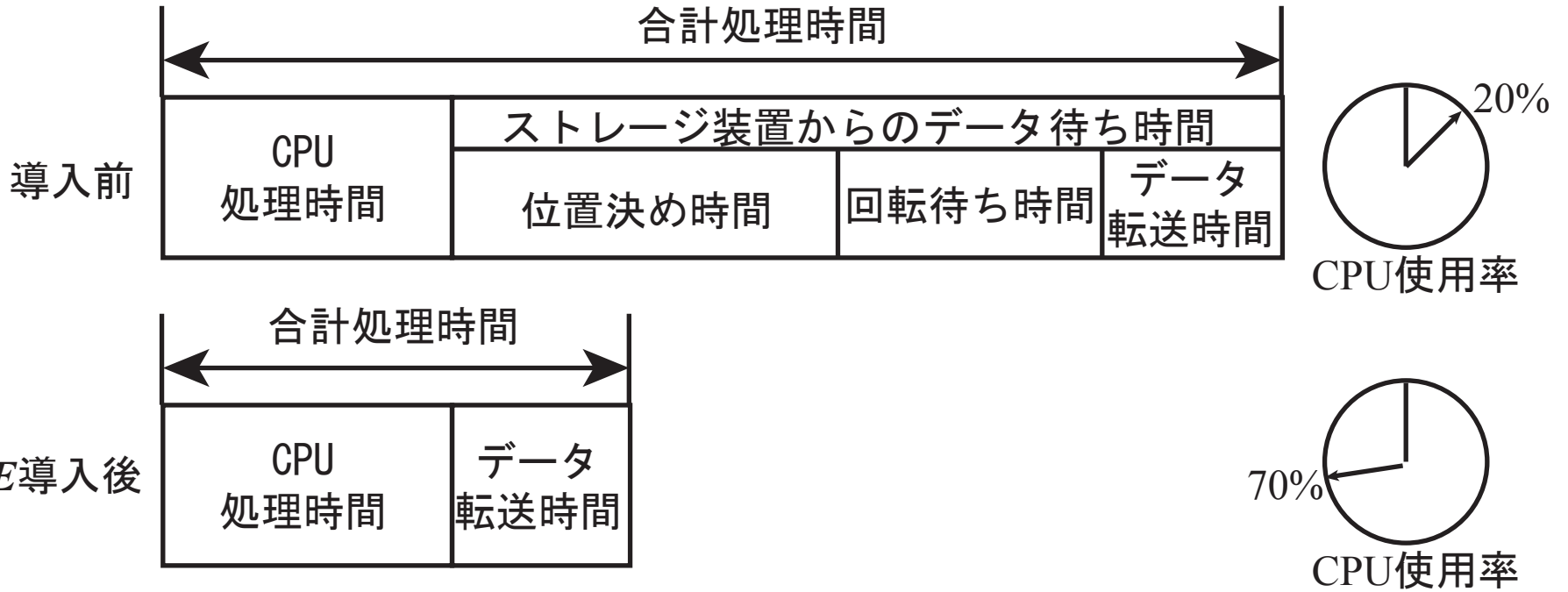
○ フラッシュ性能参考値*1(IOPS)

	LBAオーダー	LRUオーダー
512B	5,500～6,500	1,300～1,600
8KB	2,200～2,700	1,300～1,500
64KB	1,800～2,100	1,200～1,400

*1)ソースデバイスの性能に依存します。

1.5倍～4倍の効率化

効果例



・これまで10時間かかっていた解析処理が3時間に短縮できる。
・これまで、夜間バッチでしか行えなかった処理が昼間に行えるようになる。
・レスポンス時間が数十秒かかっていたオンライン処理が、数秒でレスポンスするようになる。
・その他、***Solid STOR***(DRAM型半導体ディスク）の事例や効果例をご参照ください

効果(2)

○ エコ対策

・これまで、性能処理を向上させるために、サーバ増設やストレージ装置の増設やHDDの増設を行ってきませんでしたか？
・***Super CACHE***を導入することにより、サーバ数、ディスク数を減らし、消費電力を低減することができます。

○ コスト低減

・データベースのチューニングやシステムのチューニングにSEコストを掛けすぎていませんか？
・***Super CACHE***の導入により、アプリケーションのチューニング費用を軽減することができます。

*Super CACHE*の概略仕様

項目	タイプ	仕様
インターフェース	ホストポート	4.0Gb FC x 2 (最大4)、SASオプション
	ストレージポート	4.0Gb FC x 2 (最大4)、SASオプション
キャッシュ容量	ー	20GB/45GB/100GB/(210GB*1)
IOPS性能	キャッシュヒット時最高性能	300K IOPS
スループット性能	キャッシュヒット時最高性能	4 GB/s
Cache／数	Cache オブジェクト	最大16個
Cacheアルゴリズム	ー	LRU/LIRS
Cacheフラッシュ	バックグラウンド	Dirtyチャンクマージ&フラッシュ
		Noソート/LRUソート/LBAソート
Volume共有機能	LUN MASK/Provisioning	WWNによる共有制御
設定ツール	シリアルコンソール/SSH	コマンドインターフェースユーティリティ
モニター機能	ー	LED、LCD、他ロギング機能
ケース	基本ユニット	19インチ EIA 2Uラックマウント
電源	ホットスワッパブルタイプ	二重化電源 (500W x 2)
管理	ー	SNMP及びSNMPトラップによる故障通知

＊ 1) 210GBは検証中。
＊ 2) 本仕様は変更することがあります。

Super SSDの概略

Super CACHE

＋

Super RAID/SSD

・FlashSSD搭載RAID装置

⇨

Super SSD

・ソースデバイスにFlashSSDを応用。
・キャッシュミス時のランダムリードを高速化。
・全文検索アプリ、フル検索参照アプリの処理時間短縮。

*Super SSD*の概略仕様

項目	タイプ	仕様
インターフェース	ホストポート	4.0Gb FC x 2 (最大4)、 3Gbx4/wide SAS×2 (最大4)
キャッシュ容量	—	5GB/14GB/32GB/50GB(110GB*1)
SSD容量	Flash/SATA/2.5インチ×n	500GB/1TB/2TB/4TB
IOPS性能	キャッシュヒット時最高性能	300K IOPS
スループット性能	キャッシュヒット時最高性能	4 GB/s
Cache／数	Cache オブジェクト	最大16個
Cacheアルゴリズム	-	LRU/LIRS
Cacheフラッシュ	バックグラウンド	Dirtyチャンクマージ&フラッシュ
		Noソート/LRUソート/LBAソート
Volume共有機能	LUN MASK/Provisioning	WWNによる共有制御
SSDプロテクション	RAID	RAID 0,1,5,6,10,50,60/Hot-spair
設定ツール	シリアルコンソール/SSH	コマンドインターフェースユーティリティ
モニター機能	—	LED、LCD、他ロギング機能
ケース	基本ユニット	19インチ EIA 2Uラックマウント
管理	—	SNMP及びSNMPトラップによる故障通知

*1) 110GBは検証中。
*2) 本仕様は暫定仕様です。変更することがあります。

コアマイクロシステムズ株式会社

Core Micro Systems, Inc.

URL：http://www.cmsinc.co.jp/　Mail：sales@cmsinc.co.jp

TEL：　03-5917-6451　IP Phone：050-5558-5410　　FAX 03-5917-6452

本社 〒173-0026 東京都板橋区中丸町11-2 ワコーレ要町ビル9F