

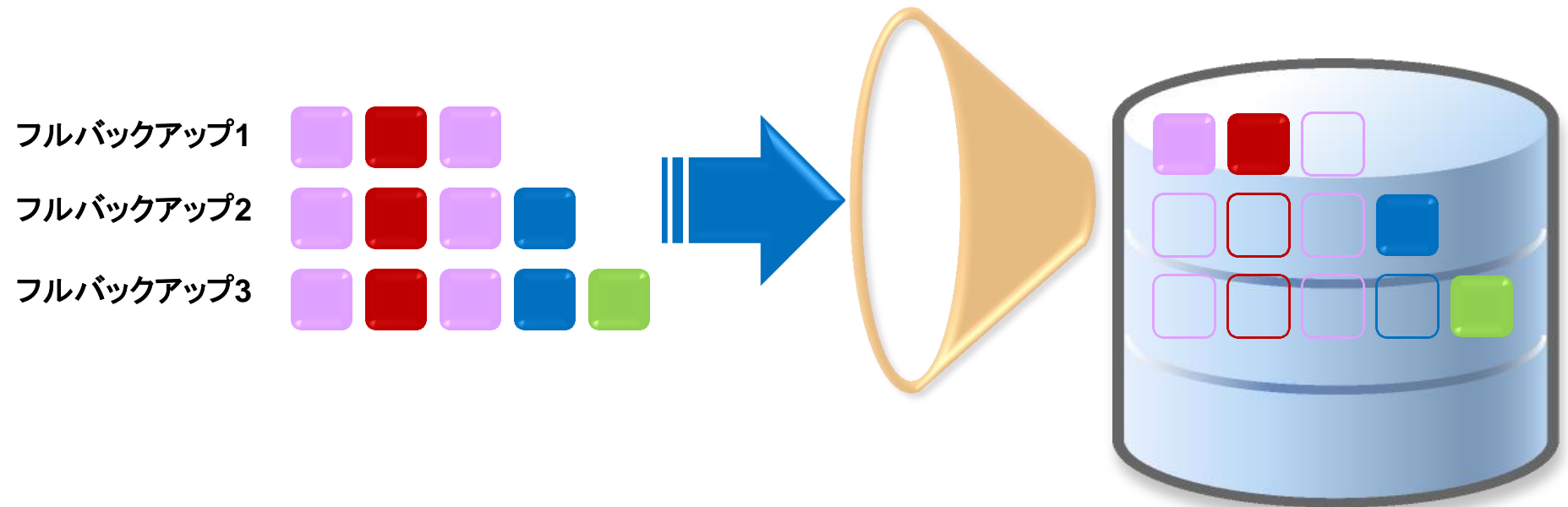
Arkeia Network Backup v.9

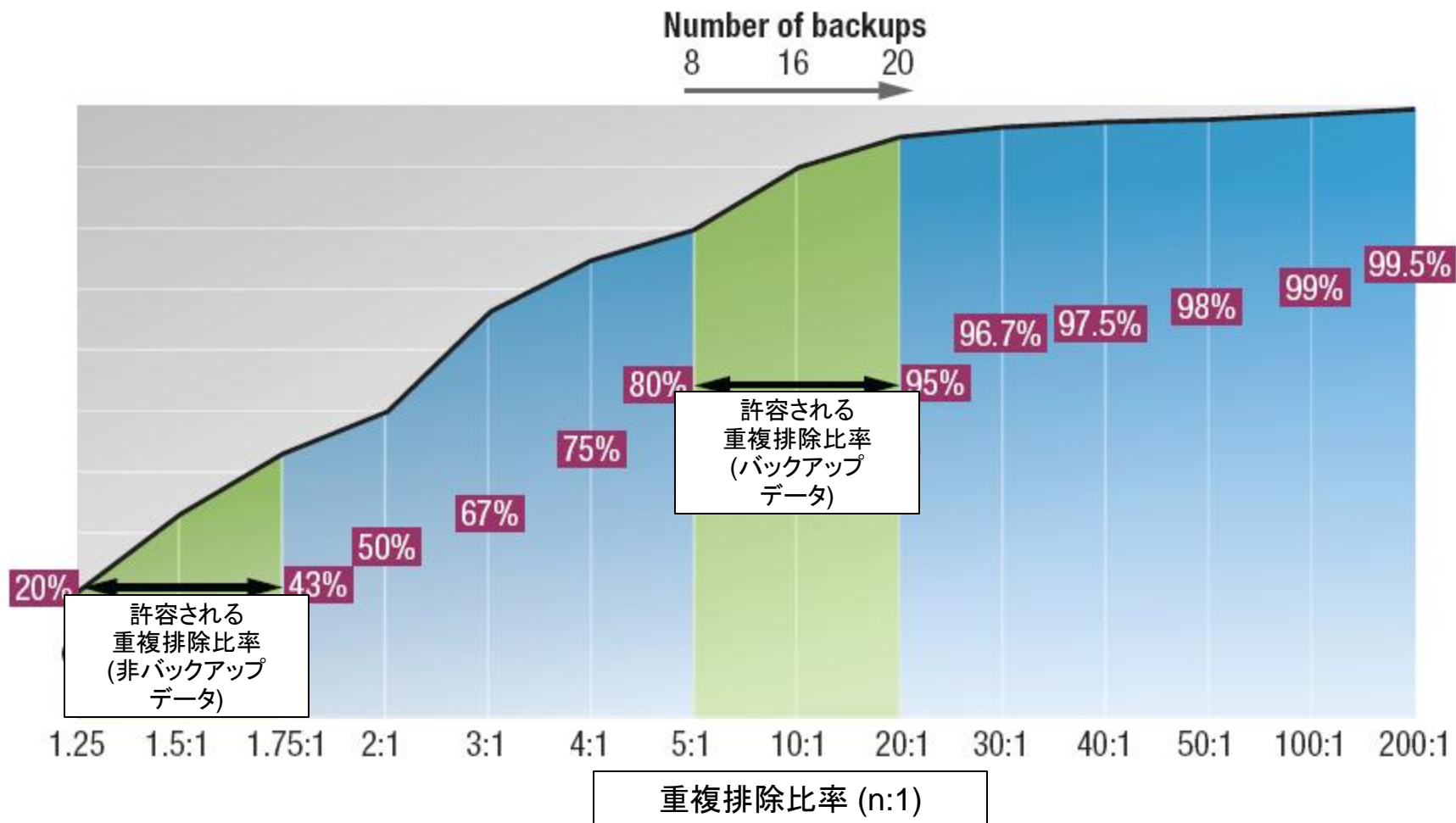
プログレッシブ重複排除技術 のご紹介

コンピュータダイナミックス株式会社

2011年12月

重複排除エンジン



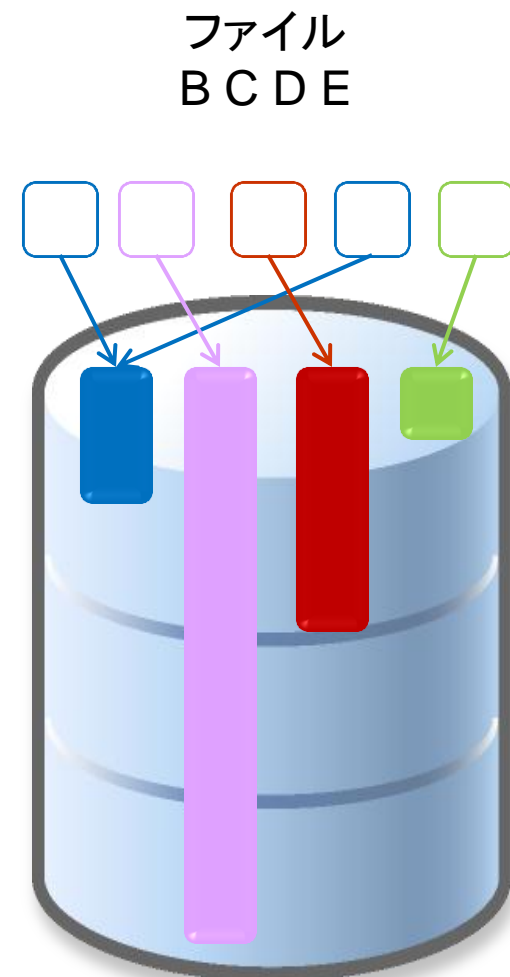
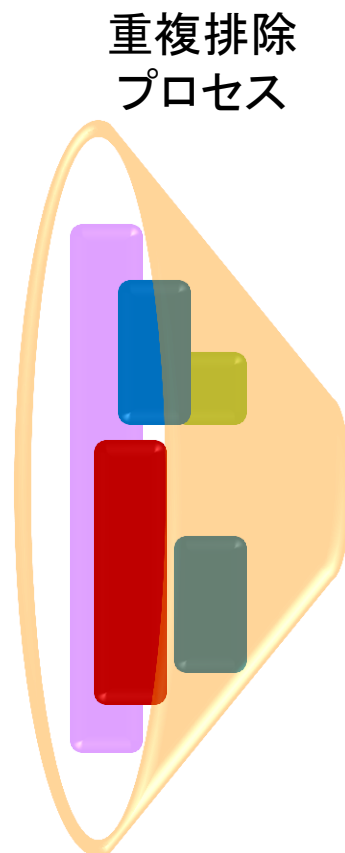
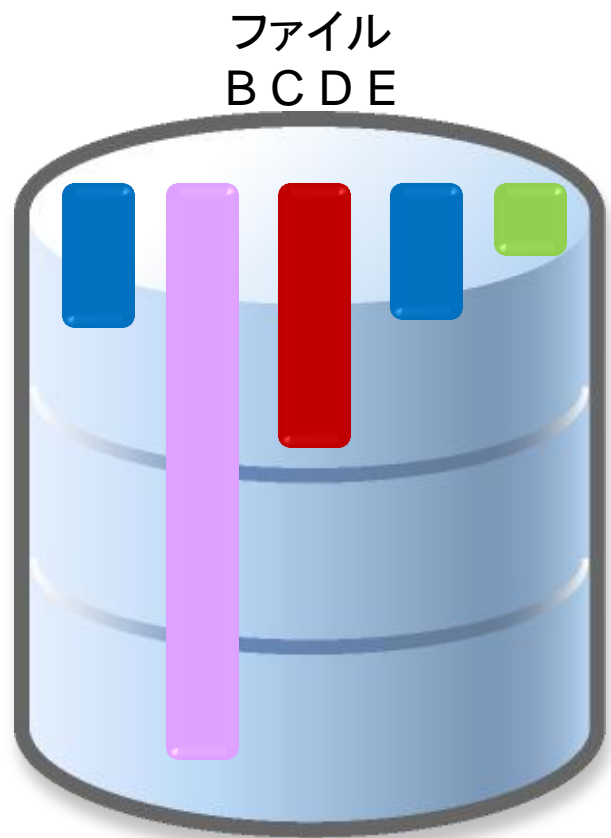


ソース: SNIA及びINFOSTOR、2007年12月

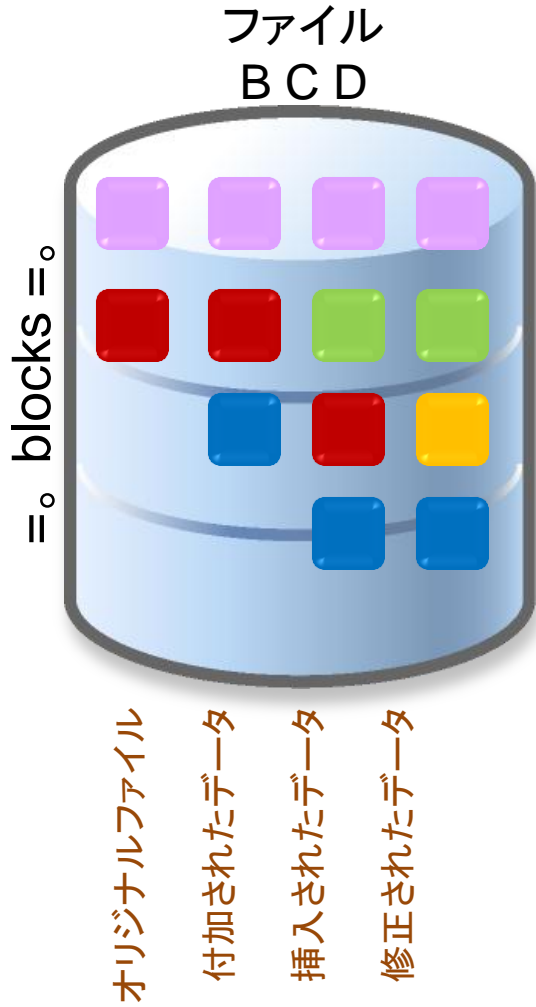
- ファイル単位("シングル・インスタンス"ストレージ)
 - 長所: 速い
 - 短所: ファイルの増分修正ができない。一般に1台のマシンに制限される。
- ブロック単位
 - 長所: ファイルの増分修正を管理できる。
 - 短所: **CPU**負荷が高い

ファイルの増分修正:

- ファイル中の何バイトかを修正。(すべてのアルゴリズム)
- ファイル中に何バイトかを付加。(すべてのアルゴリズム)
- ファイル中に何バイトかを挿入 (非固定ブロックアルゴリズム)



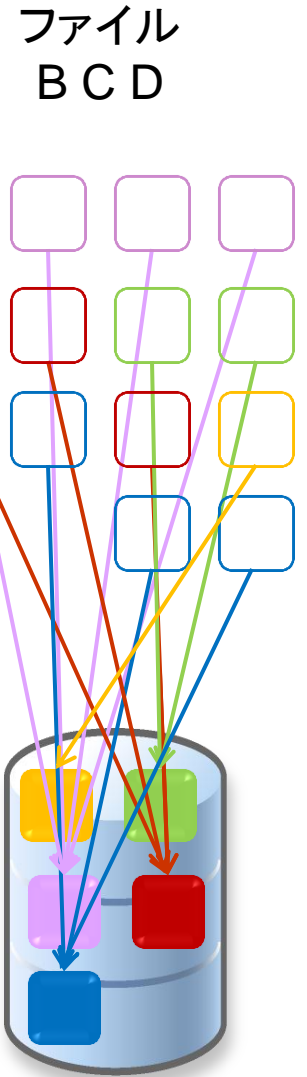
他に無いファイルだけがディスクに保存される。



重複排除プロセス



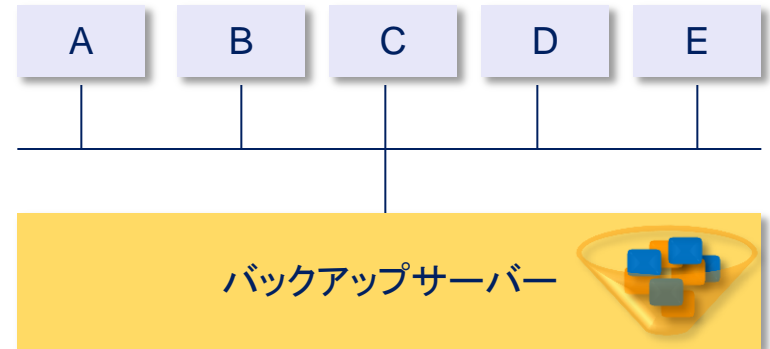
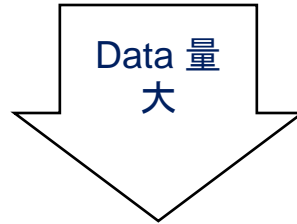
他に無いブロックだけが
ディスクに保存される。



重複排除処理を行う場所

- ターゲット重複排除

- サーバー上での処理
- ネットワークトラフィックをフルに使用

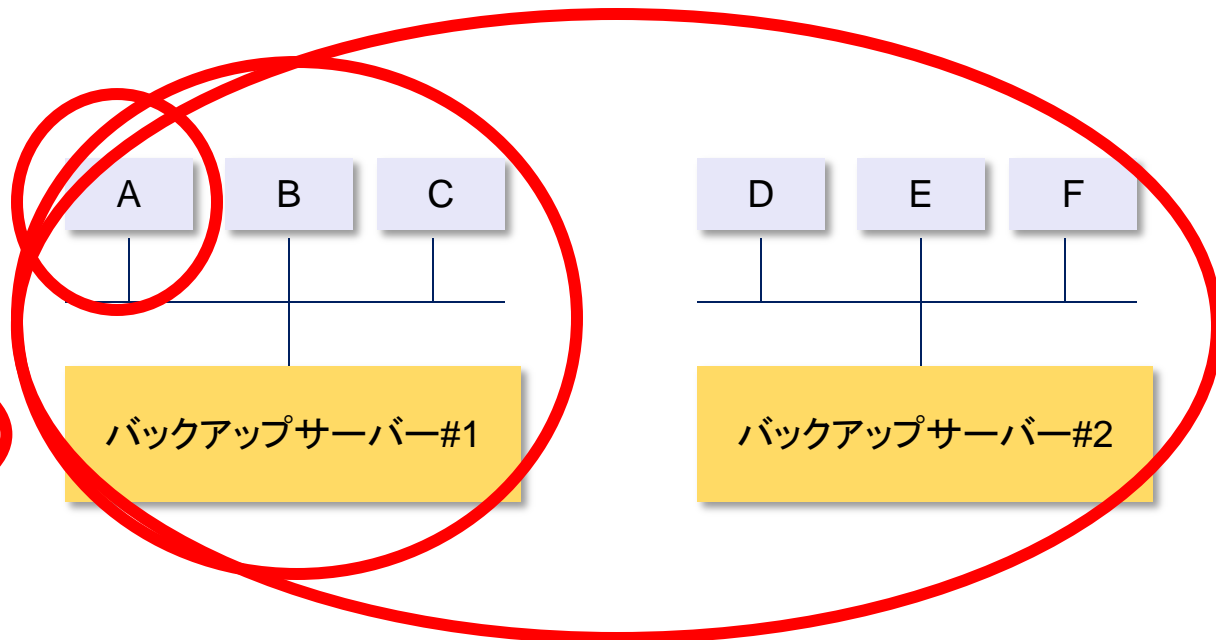


- ソース重複排除

- クライアント上での処理
- ネットワークトラフィックは削減

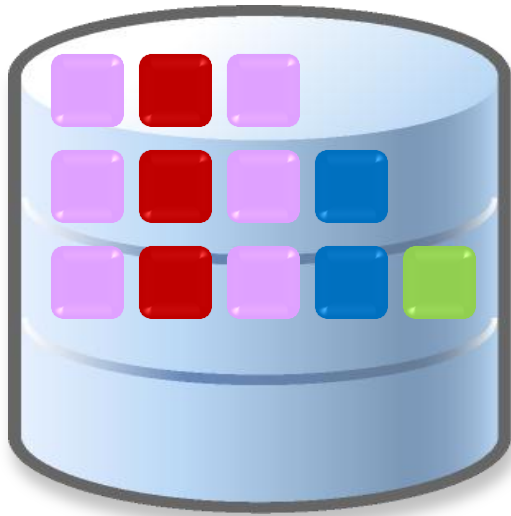


- ローカル
 - 単一のマシン
- グローバル
 - 複数クライアント
 - 1台のサーバー
- ユニバーサル
 - 複数クライアント
 - 複数台のサーバー



- インライン(別名「インバンド」)
 - データがディスクに格納される前に重複排除が行われる

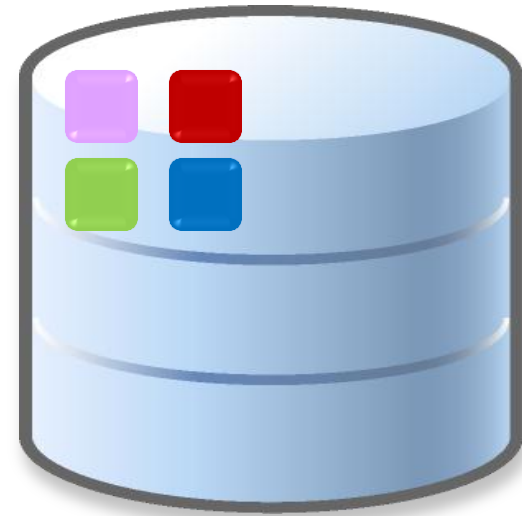
ソースデータ



重複排除プロセス

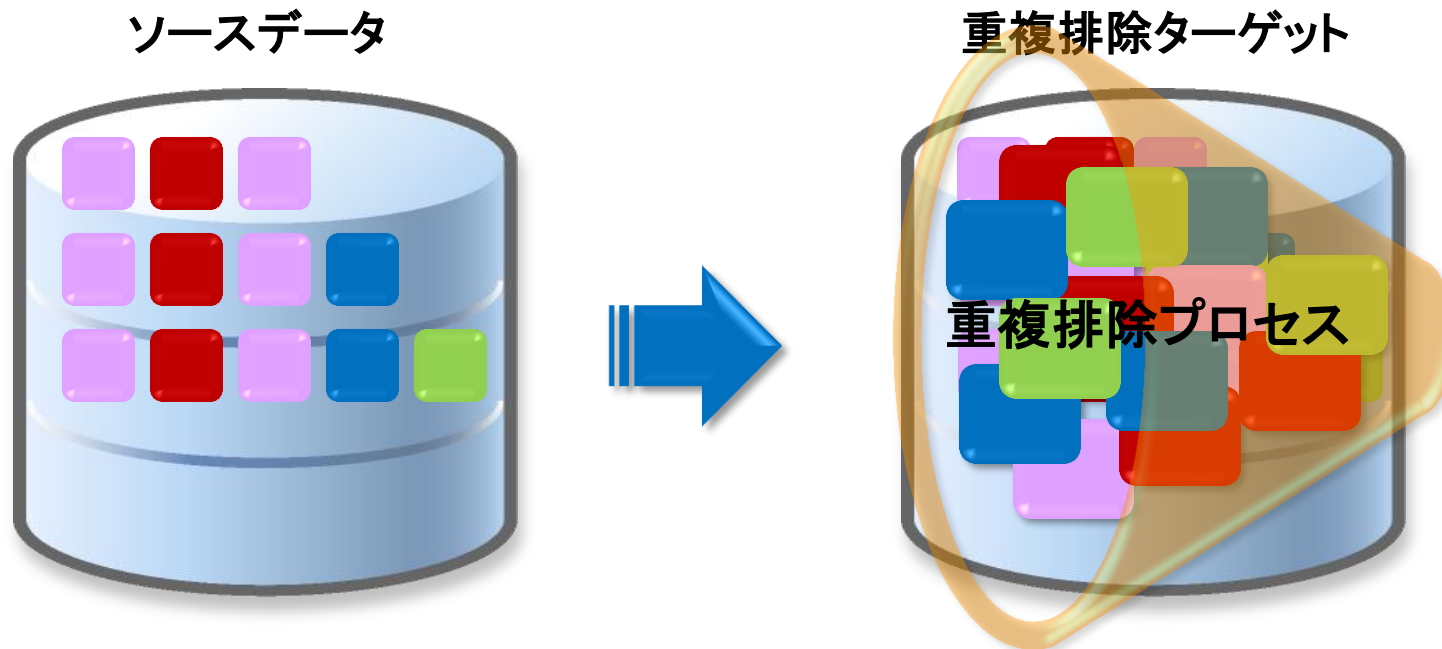


重複排除ターゲット



- 後処理(別名「アウトバンド」)

- データがディスクに格納された後に重複排除が行われる。



インラインvs.後処理

● インライン重複排除

－ 長所

- より少ない必要ストレージ
- より管理しやすい

－ 短所

- より処理時間が必要

● 後処理重複排除

－ 長所

- より短いバックアップウィンドウ

－ 短所

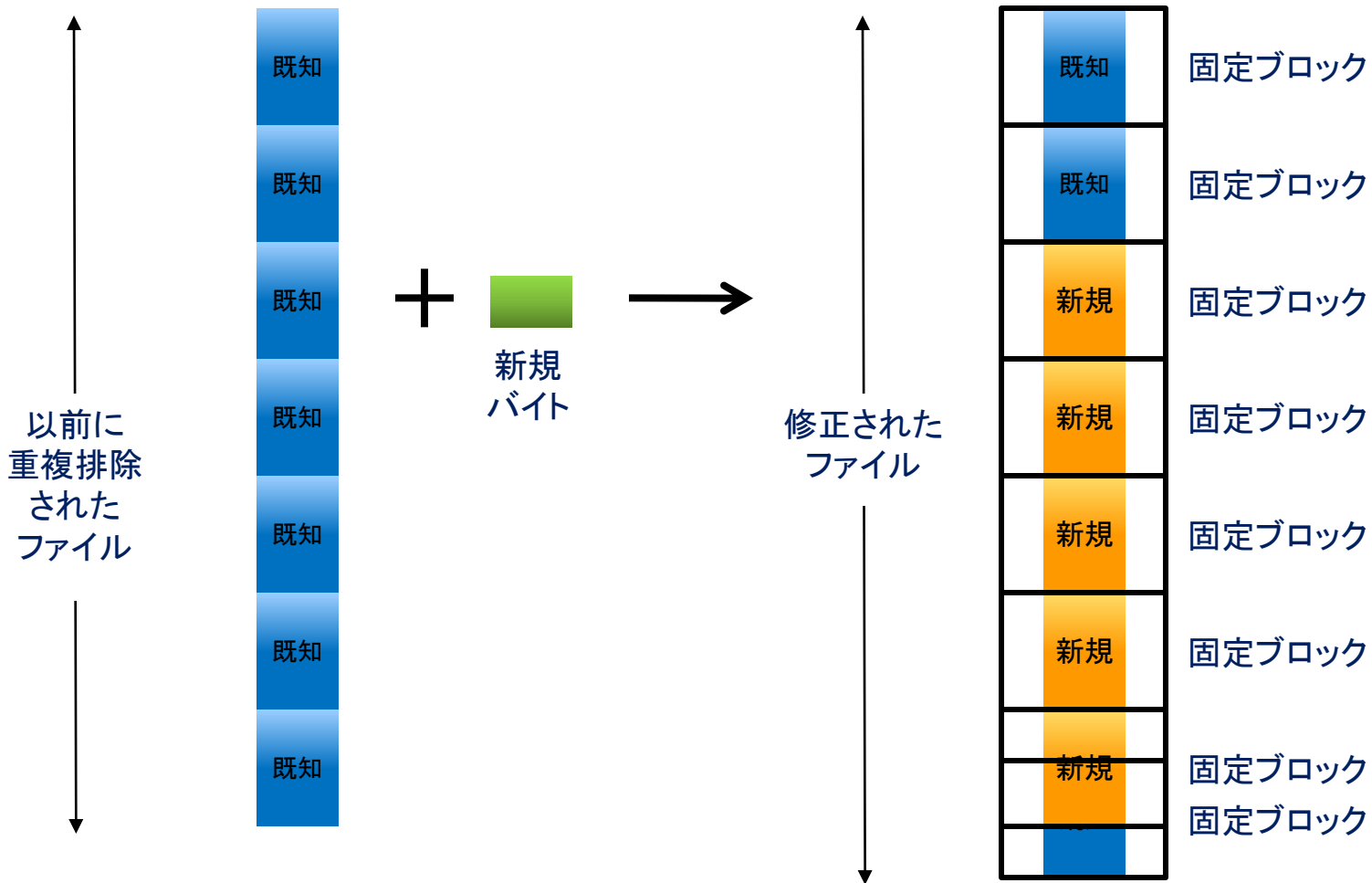
- より大きな必要ストレージ

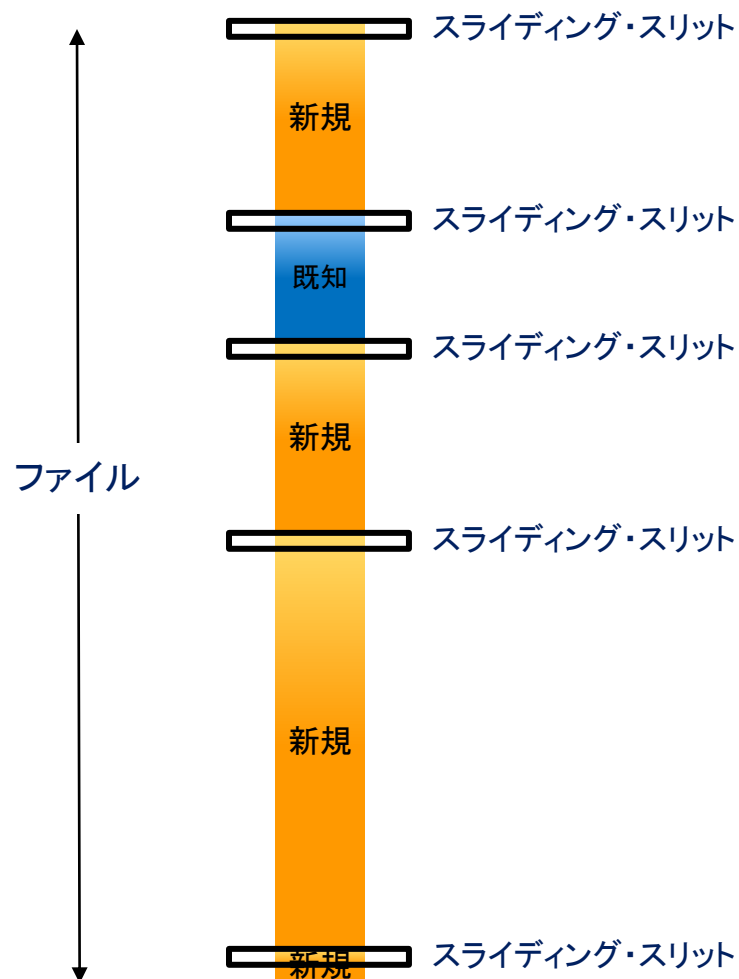
インラインと後処理重複排除の間の処理時間のギャップは、急速に縮小中！

- 固定ブロック
 - 長所: 速い
 - 短所: データ挿入に対応できない
 - 例: Commvault、PureDisk
- 可変長ブロック
 - 長所: 圧縮率高い
 - 短所: CPU負荷大、内容に応じて処理できない、特殊なケースが頻出。
 - 例: Data Domain、Quantam
- スライディング・ウィンドウ
 - 長所: 圧縮率高い、内容に応じて処理、ほとんどない特殊なケース
 - 短所: 非常にCPU負荷大 (プログレッシブ・マッチング技術無しの場合)
 - 例: Kadena System



- 重複排除プロセス(簡易的な説明)
 - 内容に基づいてブロックサイズを設定
 - ブロックに対する指紋を計算
 - 既知のブロックの指紋とそのブロックの指紋を照合することで重複を検出
- ソース側バックアッププロセス
 - 新規ブロックをzipしバックアップサーバに送付
 - 既知のブロックに対するリファレンスをバックアップサーバに送付



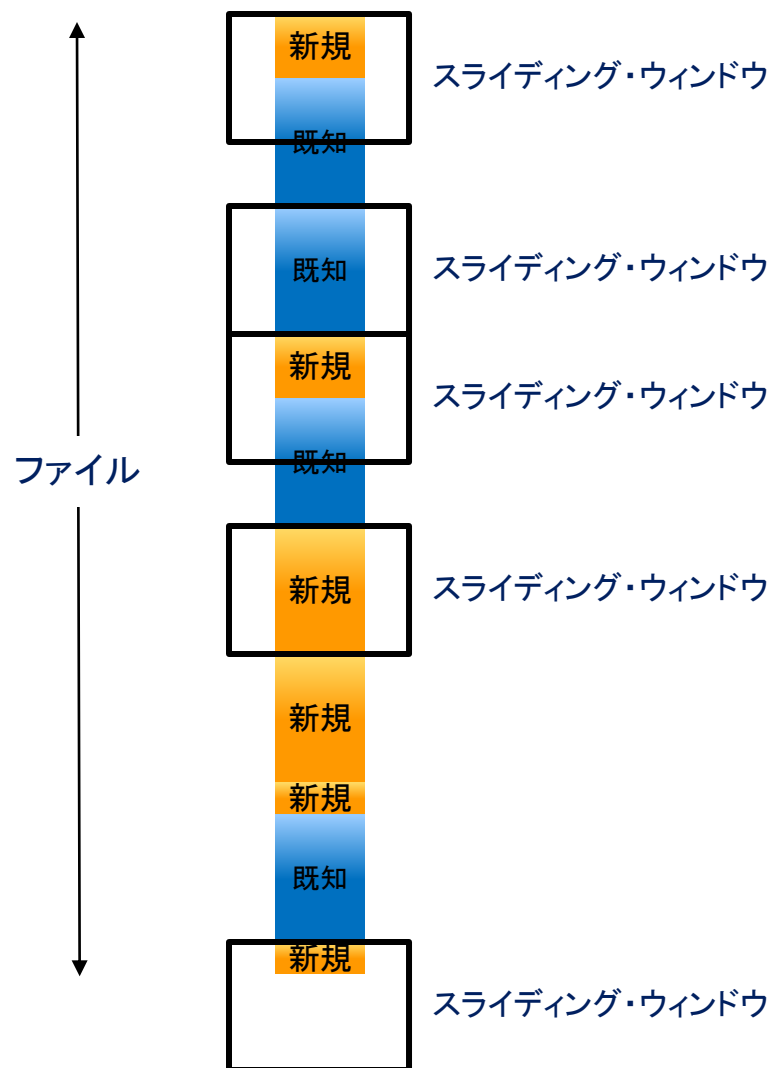


● 重複排除プロセス(簡易的説明)

- スライディング・スリットによってブロックの終わりを決める「魔法」のパターンを見つける
- ブロックに対する指紋を計算
- 既知のブロックの指紋とそのブロックの指紋を照合することで重複を検出

● ソース側バックアッププロセス

- 新規ブロックをzipしバックアップサーバに送付
- 既知のブロックに対するリファレンスをバックアップサーバに送付



- 重複排除プロセス(簡易的説明)
 - 内容に基づいてウィンドウサイズを設定
 - プログレッシブ・マッチング技術を用いて重複ブロックと成り得るものを先に検出
 - そのブロックの指紋を既知のブロックの指紋と比較することによって重複を確認
 - 比較する重複プールの範囲は"グローバル"か、"ユニバーサル"
- ソース側バックアッププロセス
 - 新規ブロックをzipしバックアップサーバに送付
 - 既知のブロックに対するリファレンスをバックアップサーバに送付

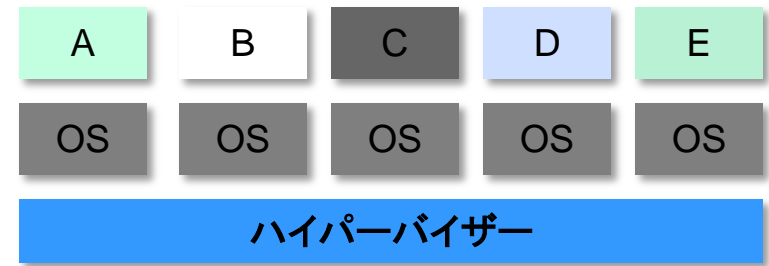
- なぜスライディング・ウィンドウは可変長ブロックより優れているか？
 - 内容に応じた処理(即ちウィンドウサイズを調整可能)のため、より高い圧縮比
 - ファイル中のすべてのブロックが同じサイズなのでより高速であり、データ管理もより簡単な(例えば中間処理が減少)
 - 特殊なケース(例えば、EOBも見つからない場合)が無いためより高速
- スライディング・ウィンドウを可変長ブロックに比べた場合どうか？
 - ハッシュ計算については同様なCPU負荷(Arkeiaではプログレッシブ・マッチングも併用)

可変長ブロックアルゴリズムにおける根本的な欠点

可変長ブロックアルゴリズムのブロックの終わりを検出するアルゴリズムは、データがランダムであると仮定しています。これはデータが重複排除できるという仮定と矛盾しています。結果として「特殊なケース」が多く生じる事となり、重複排除処理を遅くし、圧縮比を下げています。

	固定長ブロック 重複排除	可変長ブロック 重複排除	プログレッシブ 重複排除
付加(Append)や修正 (Modify)に対処	○	○	○
挿入(Insert)に対処	×	○	○
圧縮率	△	○	◎
処理速度	○	△	◎

- サーバ仮想化環境は大規模なデータ重複の問題を持っている
 - ホストOSあたり1セットのシステム・ファイルを保存
- インパクト
 - 遅いバックアップ
 - ネットワークとディスクトラフィック渋滞
- Arkeiaソリューション
 - ハイパーバイザーベースのソース側重複排除
- Arkeiaの利点
 - バックアップ時間を削減
 - ディスク要件を削減
 - 3つの導入モード



- 分散環境は大規模な帯域幅制限の問題を持っている。

- 大量データをWAN上に送ることは現実的でない

- インパクト

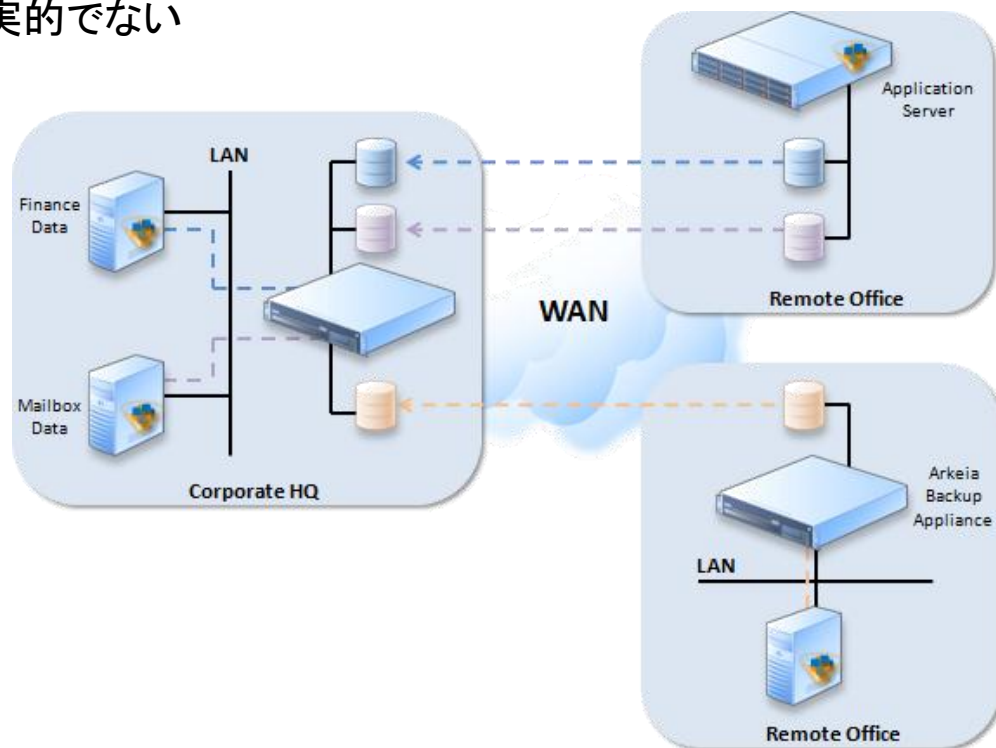
- 低速バックアップか、またはバックアップを行わない
- ネットワークの渋滞
- 管理の困難さ

- Arkeiaソリューション

- ソース側のグローバルな重複排除
- 重複排除の複製

- Arkeiaの利点

- 統合バックアップ
- バックアップ時間を削減
- データセンターにおけるソフトウェアまたはリモート・サイトにおけるアプライアンスとして簡単に導入可能



重複排除機能バックアップソフトウェア比較 **Computer Dynamics**

	Arkeia Network Backup v9	Symantec NetBackup 7.0 ¹	Symantec Backup Exec 2010 ¹	BakBone NetVault: Backup 8.2	CommVault Simpana 8	CA ArcServe 12.5	EMC Avamar 5.0	Atempo TN	Acronis B & R	Asigra Hybrid Cloud
重複排除粒度										
ファイルレベル粒度									○	
ブロックレベル粒度	○	○	○	○	○	○	○	○	○	○
タイミング										
ポストプロセッシング	○			○				○	?	
インライン	○	○	○		○	○	○	○	?	○
重複排除処理の場所										
ソース	○	○	○				○		○	○
ターゲット	○	○	○	○	○	○	○	○	○	
範囲										
ローカル	○	○	○	○	○	○	○	○	○	○
グローバル	○	○	○	○	○	○	○	○	○	○
ユニバーサル	○	○	○				○			?
アルゴリズム										
固定長ブロック		○	○		○			?	○	
可変長ブロック				○		○	○	?		○
スライディング・ウィンドウ	○									
その他オプション/機能										
重複排除レプリケーション	○	○	○		○		○	○		
テープへの重複排除	○				○		○			
コンテンツ認識	○			○	○		○			
プログレッシブ・マッチング	○									
物理アプライアンス	○						○			
仮想アプライアンス	○						○			

