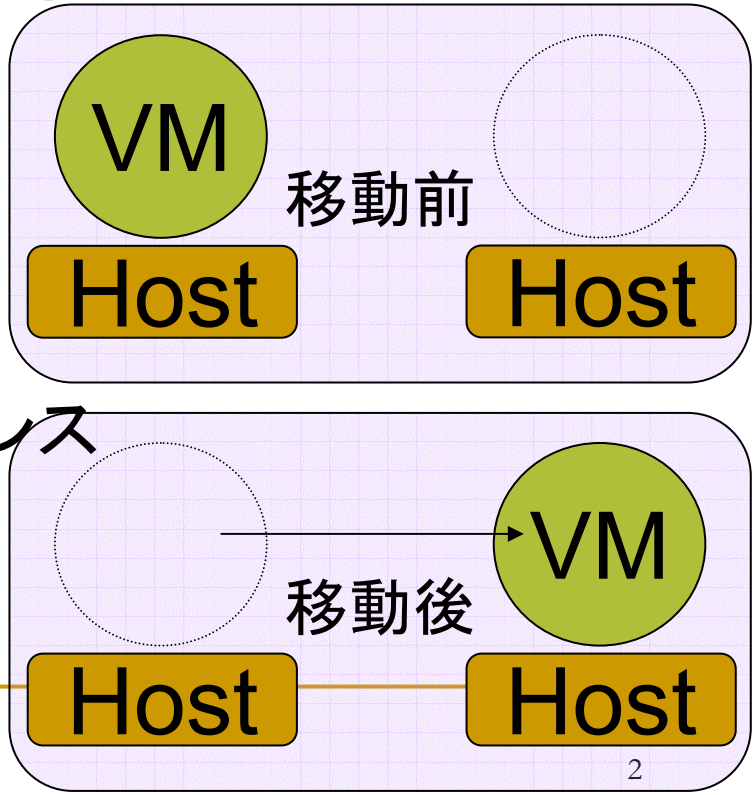
仮想クラスタ遠隔ライブマイグレーションにおけるストレージアクセス最適化機構

産業技術総合研究所 情報技術研究部門
広渕崇宏　小川宏高　中田秀基
伊藤智　関口智嗣

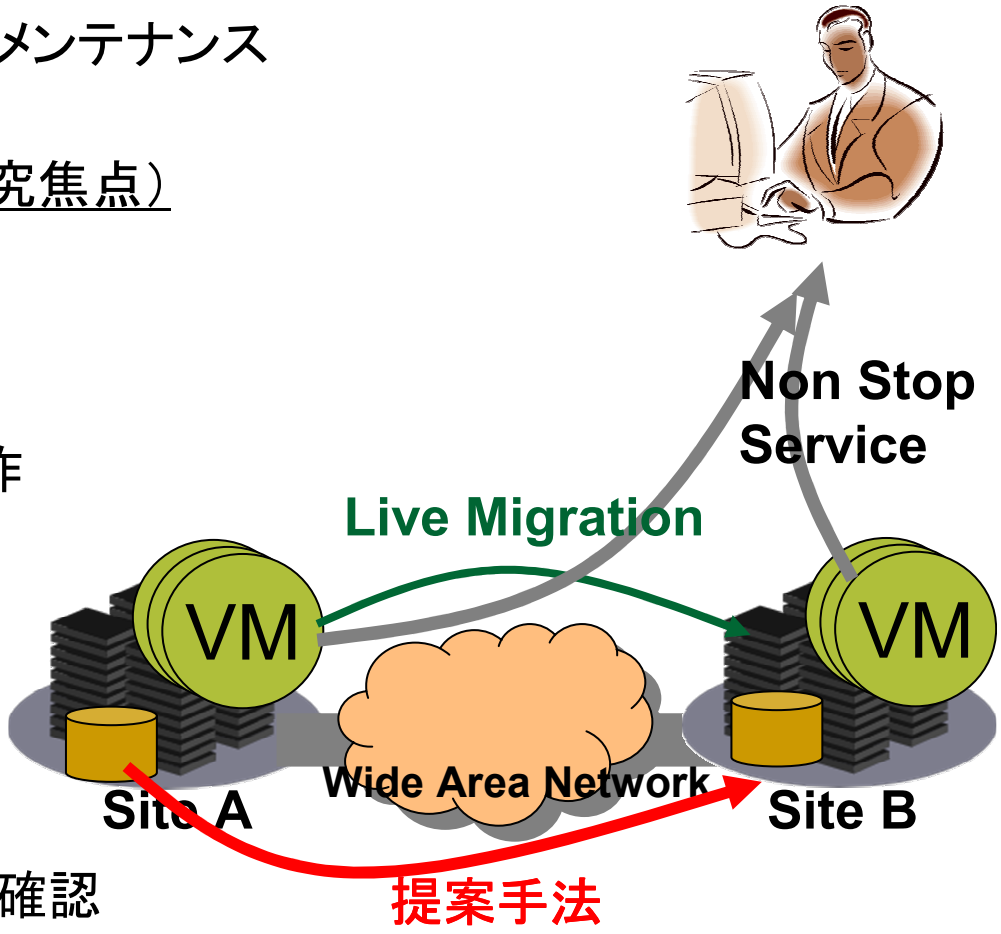
背景

- 仮想計算機技術
  - 計算資源の論理的分割、共有
  - 資源利用効率や運用柔軟性の向上
  - 仮想化データセンタ
- ライブマイグレーション
  - 仮想マシンの動的再配置
    - OSを起動したまま移動
  - 負荷分散、省電力化、メンテナンス
  - 単一拠点内での実用化のみ
    - 運用柔軟性に限界



研究目的と成果

- 遠隔ライブマイグレーション
  - 仮想計算機を遠隔拠点に対して再配置
  - 施設全体にわたる省電力化、メンテナンス
  - 拠点横断的な負荷バランス
  - ストレージアクセスに問題（研究焦点）
- 提案手法
  - VMストレージの再配置手法
  - 透過的振る舞い
    - 仮想マシンは継続的に動作
  - 完全な再配置
    - I/O性能の維持
    - 移動元の停止が可能
- 成果
  - 再配置手法の提示
  - プロトタイプ実装を通した動作確認

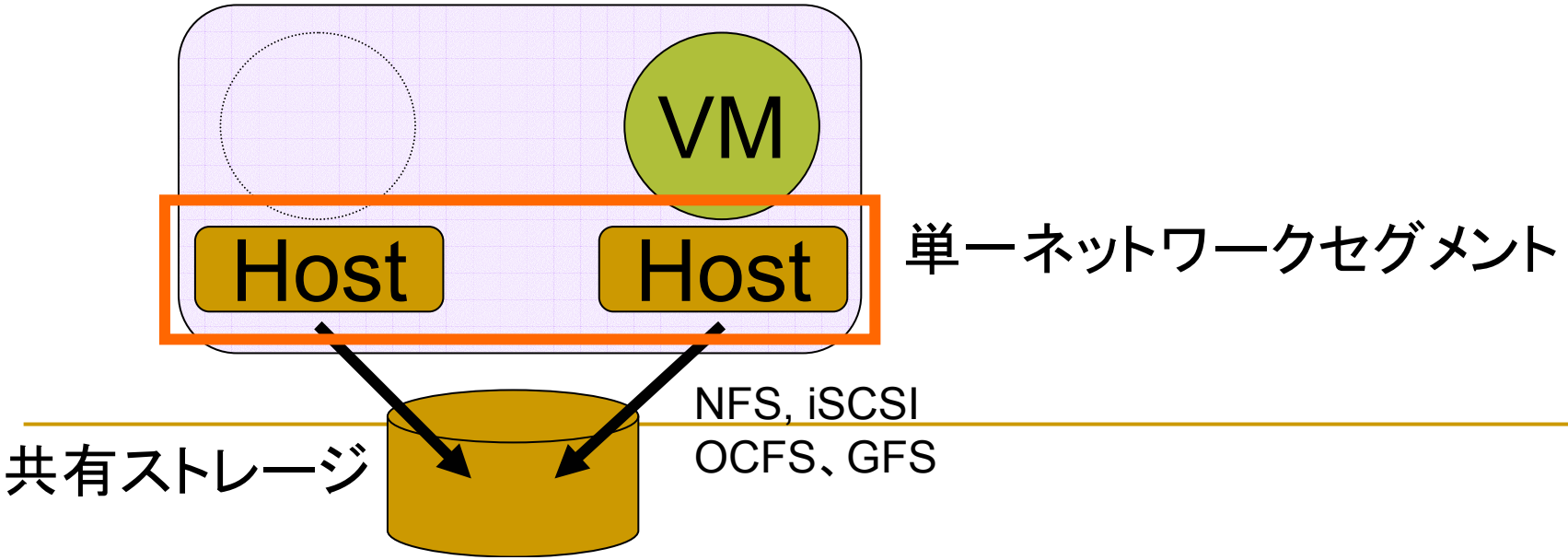


発表構成

- 動的再配置と広域環境
  - 仮想計算機ストレージに焦点を当てる理由
- 問題点
- 既存研究・技術
  - 遠隔マイグレーションとストレージ
- 要求事項
- 提案手法
- プロトタイプ実装
- 今後の課題
- 結論

動的再配置の構成要素

- 仮想マシンモニタのサポート
  - メモリイメージの動的再配置
- 単一ネットワークセグメント
  - MACアドレスやIPアドレスが不変
- 共有ストレージ
  - 同一ストレージへアクセスを継続



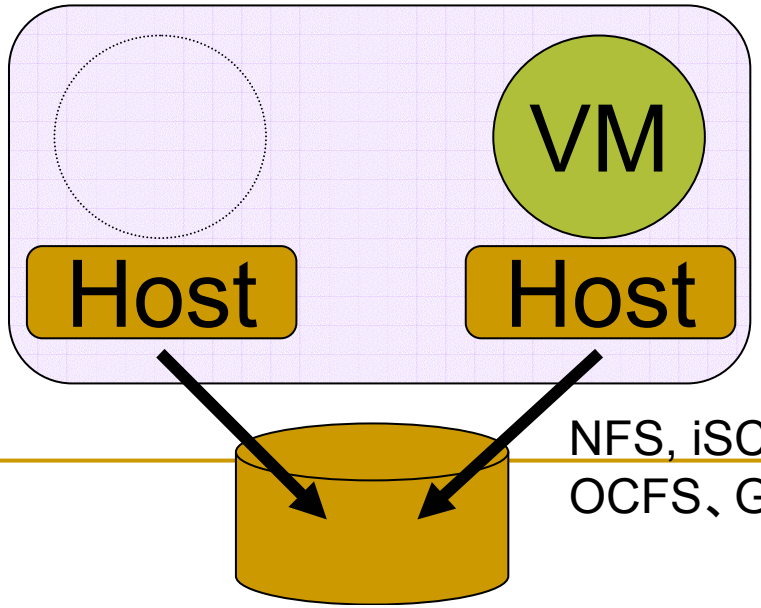
動的再配置の構成要素と広域環境

- 仮想マシンモニタのサポート
  - メモリイメージの動的再配置
- 単一ネットワークセグメント
  - MACアドレスやIPアドレスが不変
- 共有ストレージ
  - 同一ストレージへアクセスを継続

広帯域WANでは迅速に
使用中ページのコピー可能

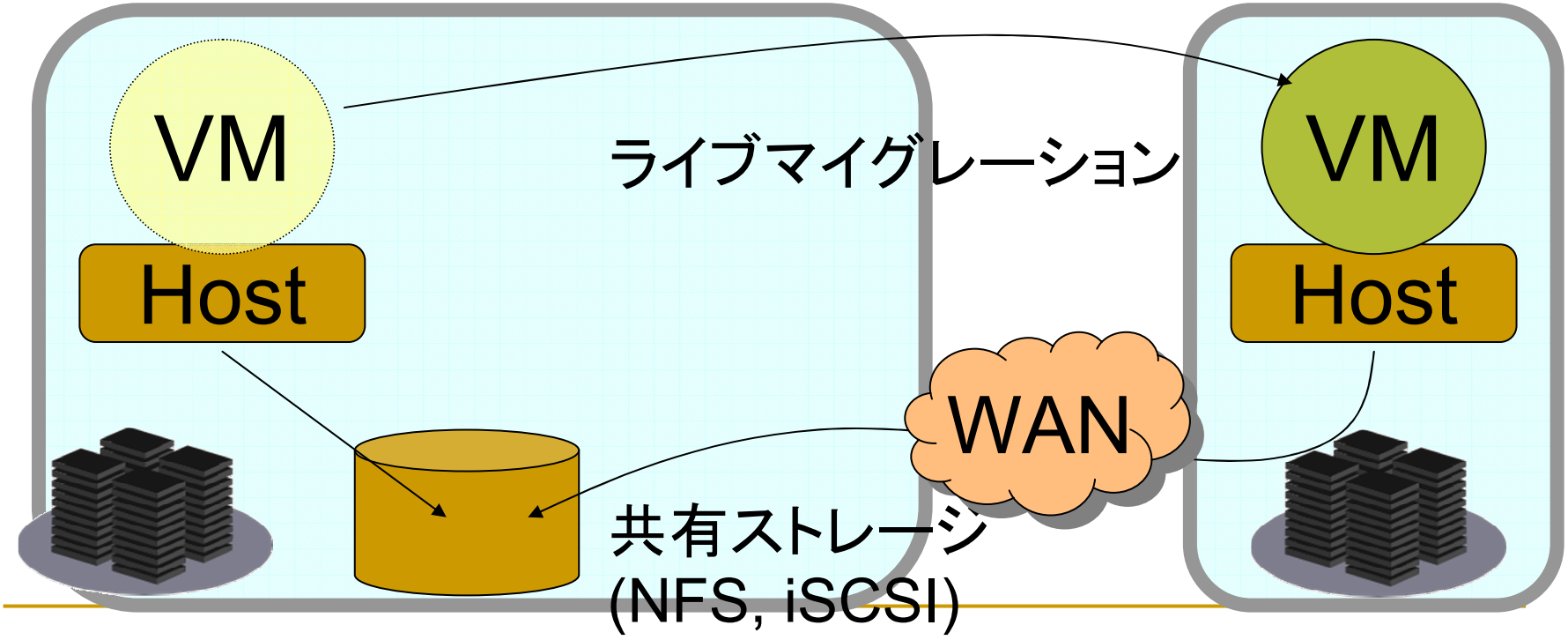
イーサネットVPN

WAN環境に不向き



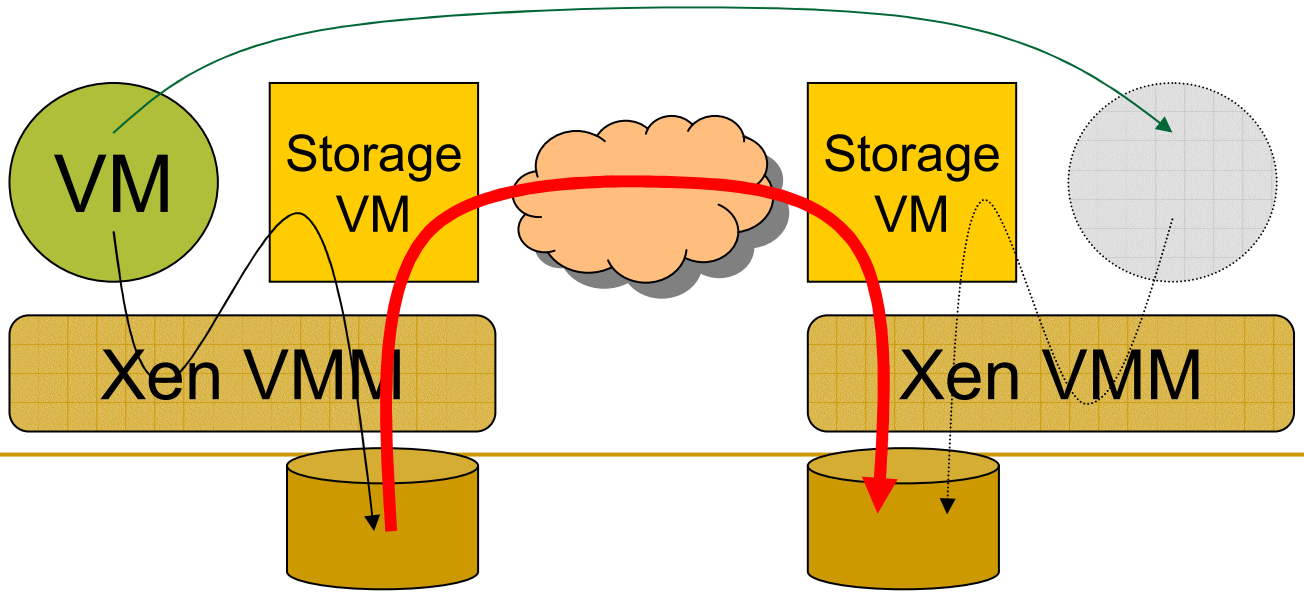
遠隔ライブマイグレーションにおける共有ストレージの問題点

- ネットワーク遅延による性能低下
- 移動後も移動元拠点に依存



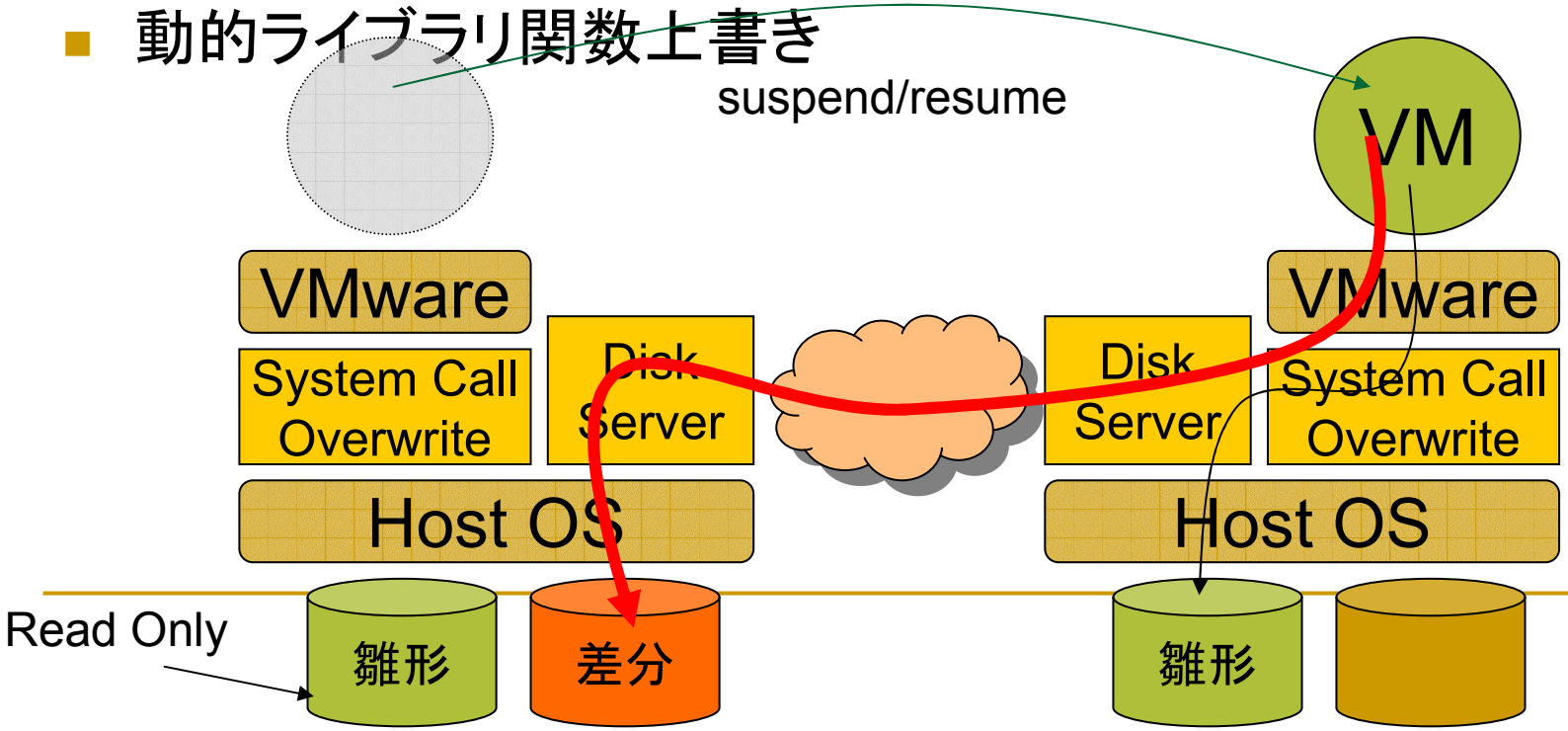
既存研究・技術（1）

- **“Live Wide-Area Migration of Virtual Machines Including Local Persistent State”** by Bradford, 2007
- 仮想ディスクイメージを先行コピー
  - 移動先にディスクイメージが再現できたら移動先でVM起動
  - コピー中はVMのI/O速度を意図的に低下
- 特殊なバックエンドストレージドライバ
- VM実行ホストの迅速な変更が不可能



既存研究・技術（2）

- **“Optimizing the Migration of Virtual Computers”** by Sapuntzakis, 2002
- あらかじめ雛形ディスクイメージを準備
- 移動後に差分のみをオンデマンドに取得
- 移動先ホストが固定的、共通イメージの存在を前提
  - ソフトウェアライセンス問題
- 動的ライブラリ関数上書き



既存研究・技術（3）

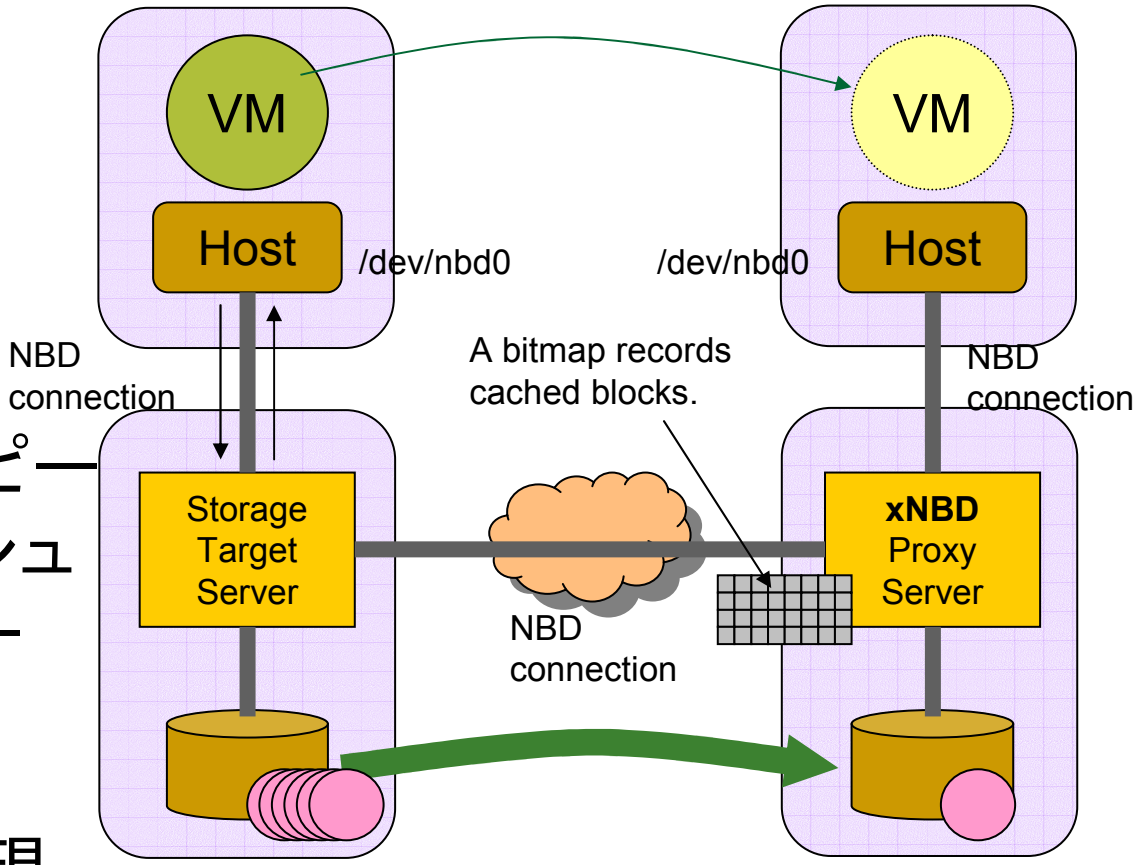
- **“Efficient State Transfer for Internet suspend/resume”** by Kozuch, 2002
  - VMディスクイメージ on Coda FS
  - 拠点外Codaサーバに常に依存
  - オーバヘッド
- **“Storage VMotion”** by VMware
  - VMディスクイメージを別のストレージに移動
  - 単一拠点内でのSANを想定
- ストレージの遠隔ミラーリング
  - 高コスト
    - ディスク、ネットワーク

遠隔ライブマイグレーションに対応するストレージアクセス機構への要求事項

- 仮想ディスクイメージの完全な再配置
  - 信頼性の向上
  - I/O性能の維持
- 仮想マシンやVMMにとって透過的な仕組み
  - OSが継続的に動作
  - オーバヘッドの最小化
  - 実装の非依存性

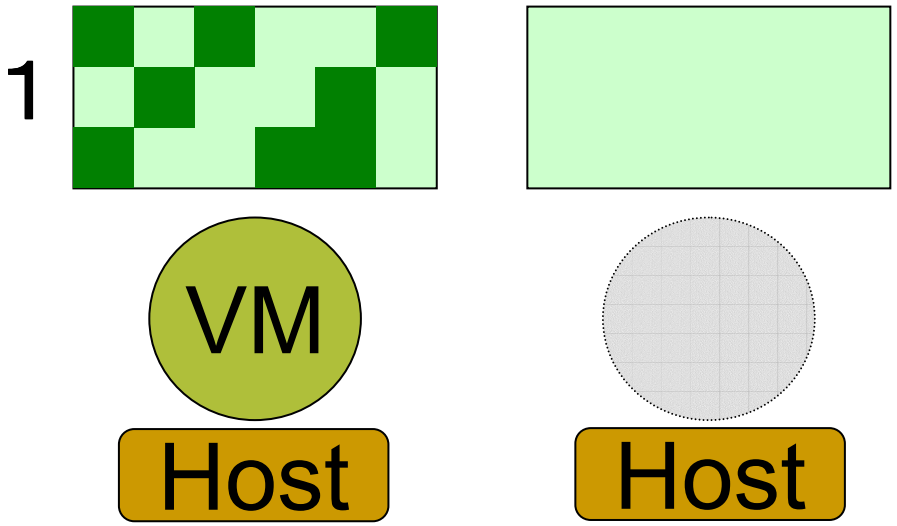
提案手法

- ストレージターゲット
  サーバ型
  - iSCSI, NBD
- メモリ再配置と連動
- オンデマンドデータコピー
- ブロックデータキャッシュ
- バックグラウンドコピー
- ストレージの完全な
  再配置を透過的に実現



メモリイメージの再配置とI/O（1）

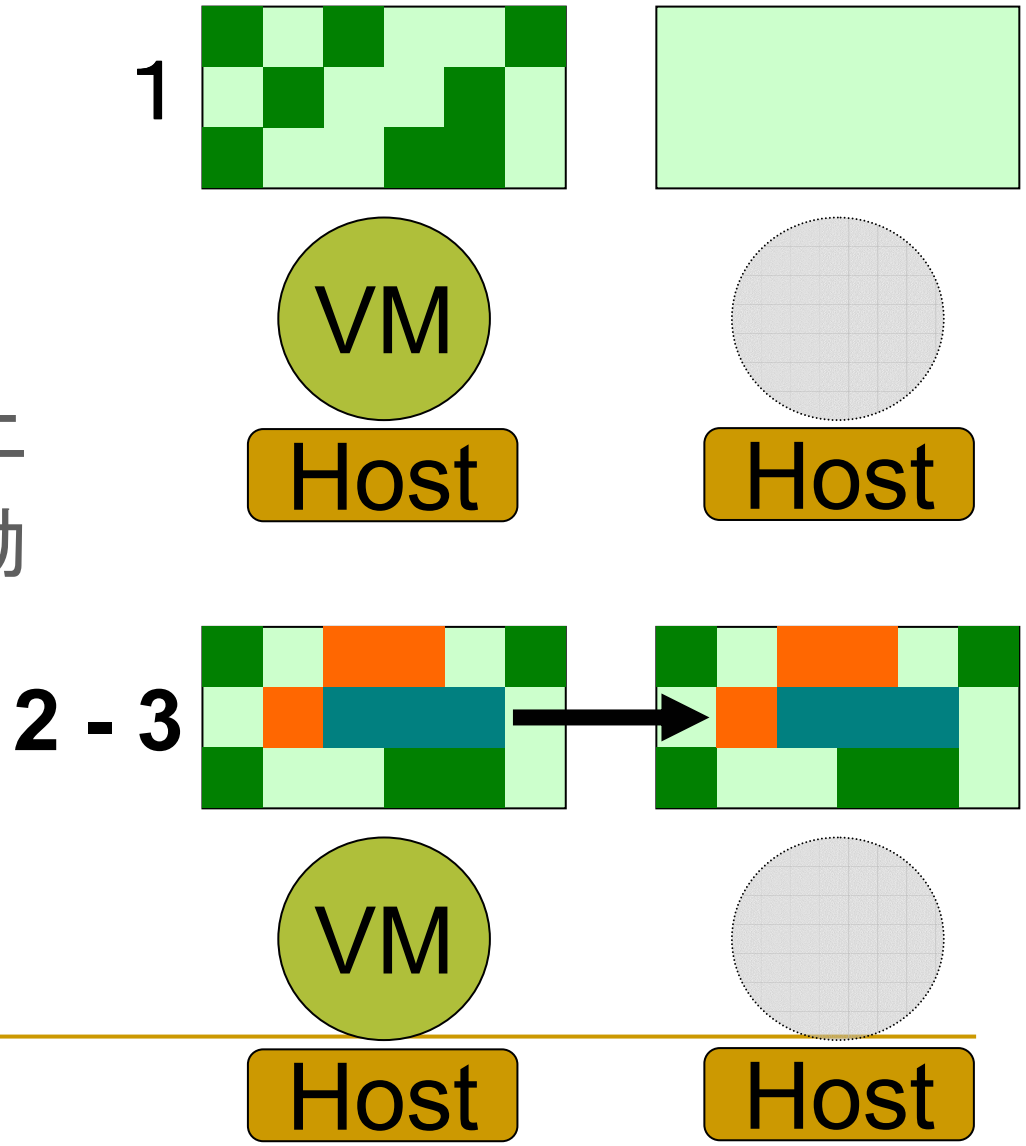
1. 資源予約
2. メモリコピーの開始
3. 更新差分のコピー
4. 移動元でのVM停止
5. 移動先でのVM起動



＊Xen “Live migration of virtual machines” by Clark, 2005

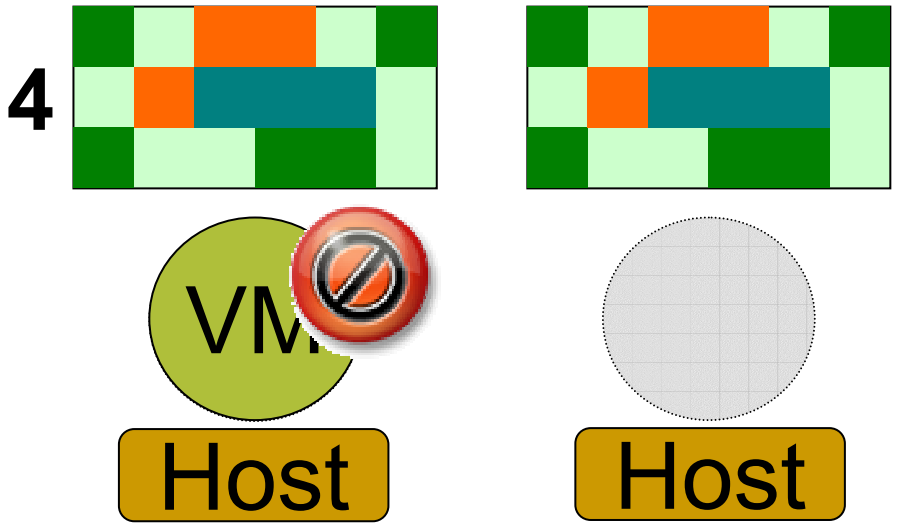
メモリイメージの再配置とI/O（2）

1. 資源予約
2. メモリコピーの開始
3. 更新差分のコピー
4. 移動元でのVM停止
5. 移動先でのVM起動



メモリイメージの再配置とI/O（3）

1. 資源予約
2. メモリコピーの開始
3. 更新差分のコピー
4. 移動元でのVM停止
5. 移動先でのVM起動

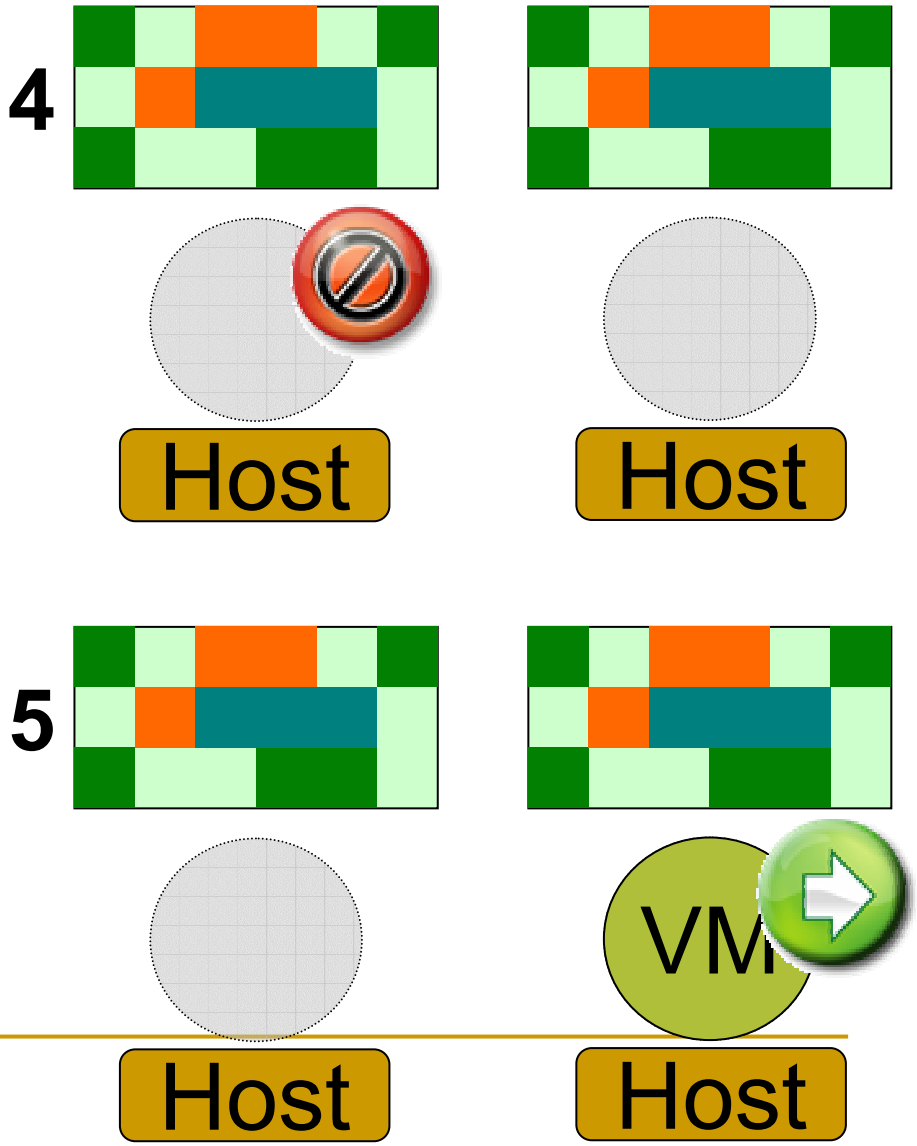


メモリイメージの再配置とI/O（4）

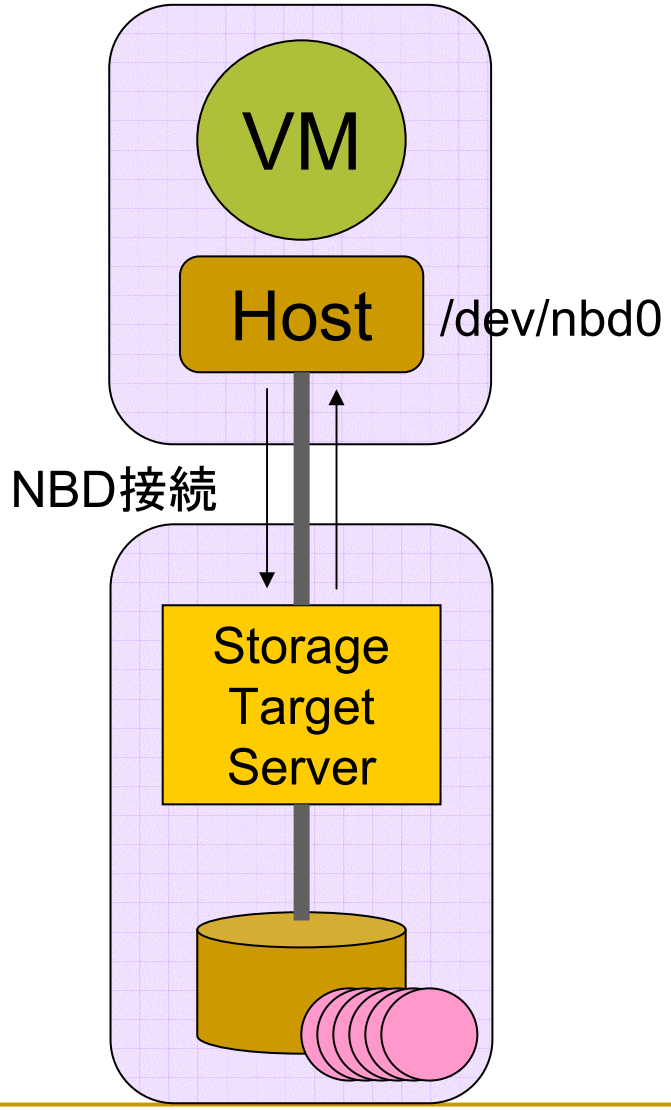
1. 資源予約
2. メモリコピーの開始
3. 更新差分のコピー
4. 移動元でのVM停止
5. 移動先でのVM起動

１－４までは移動元で
I/Oが行われる

5以降は移動先で
I/Oが行われる

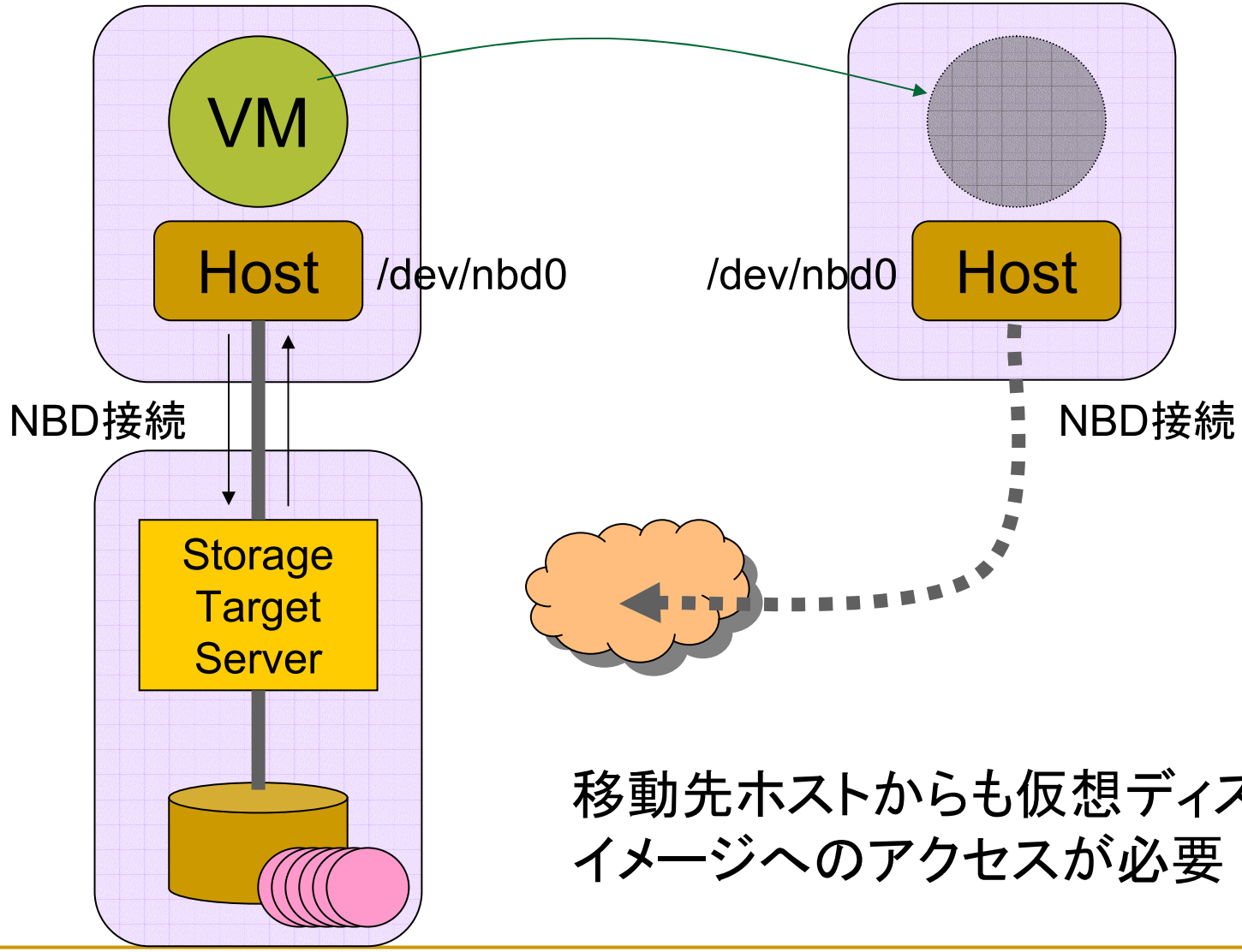


提案機構(1)



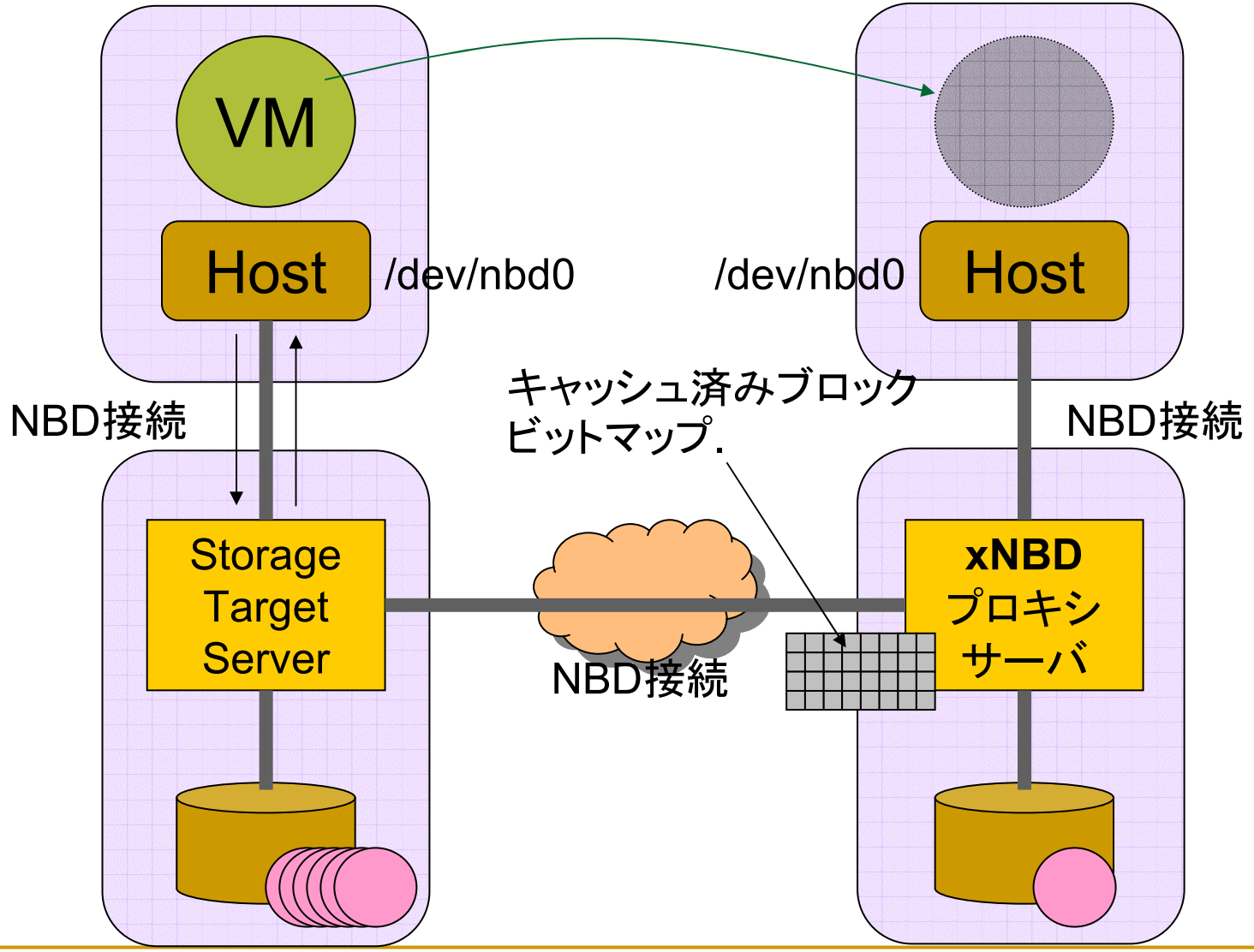
移動前は通常のNBD利用形態

提案機構(2)



移動先ホストからも仮想ディスクイメージへのアクセスが必要

提案機構(3)

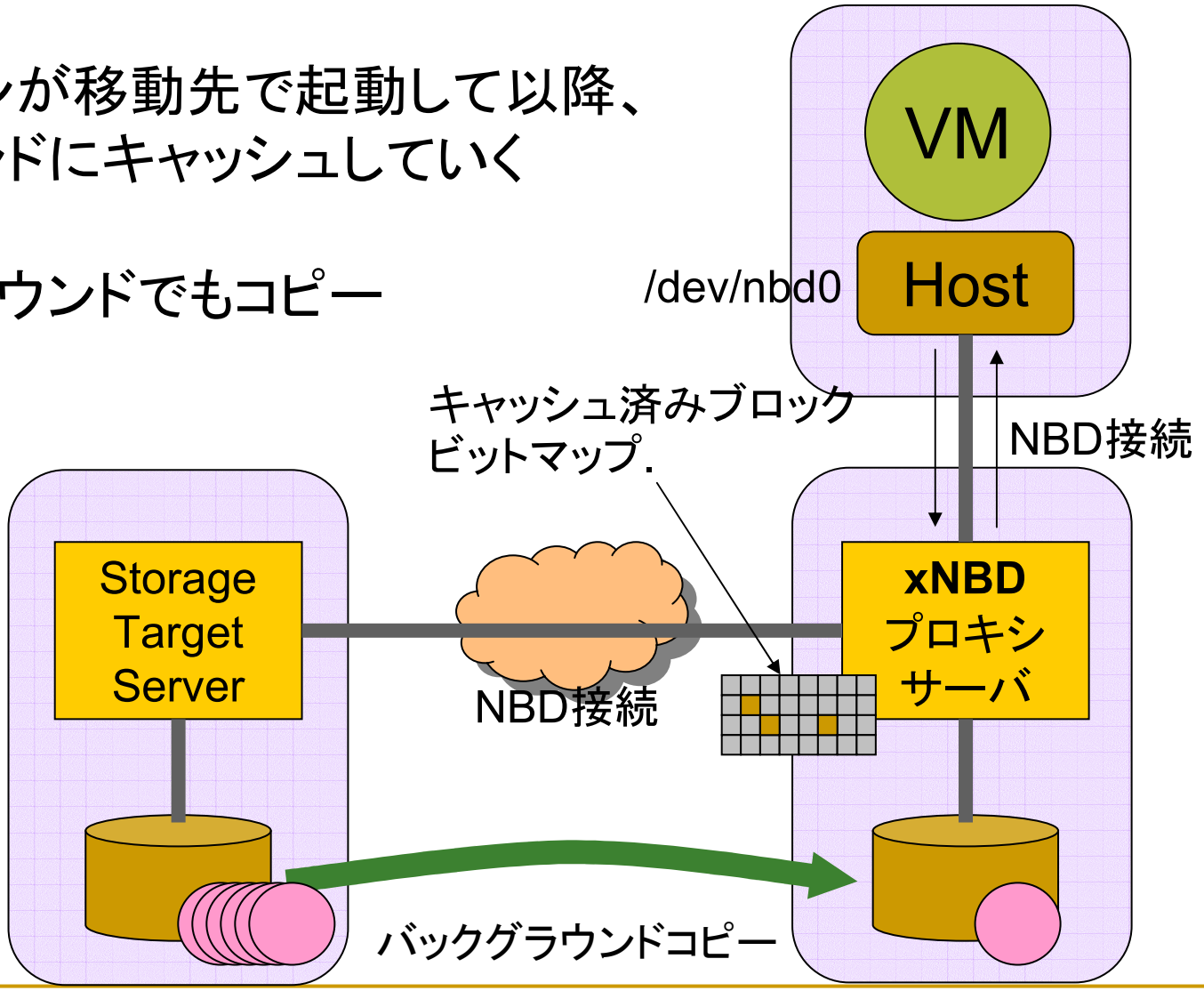


I/Oリクエストをリダイレクト

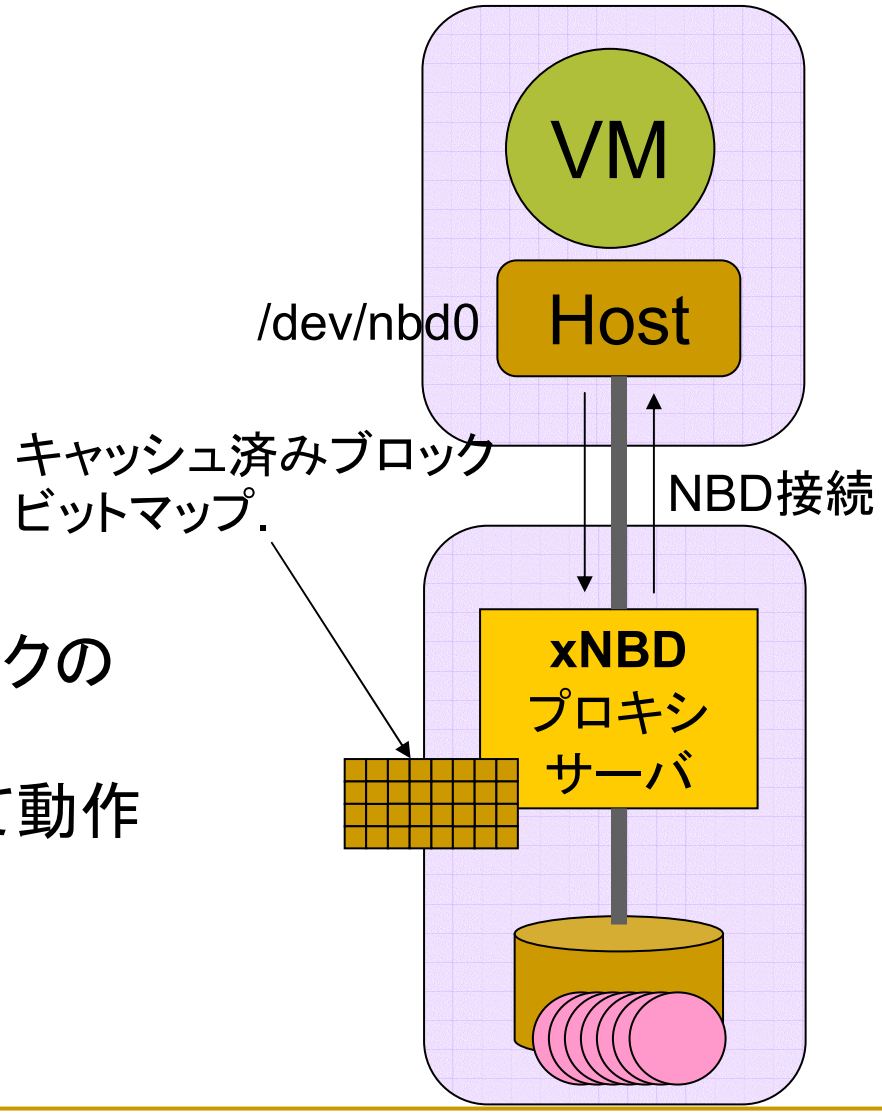
提案機構(4)

仮想マシンが移動先で起動して以降、
オンデマンドにキャッシュしていく

バックグラウンドでもコピー

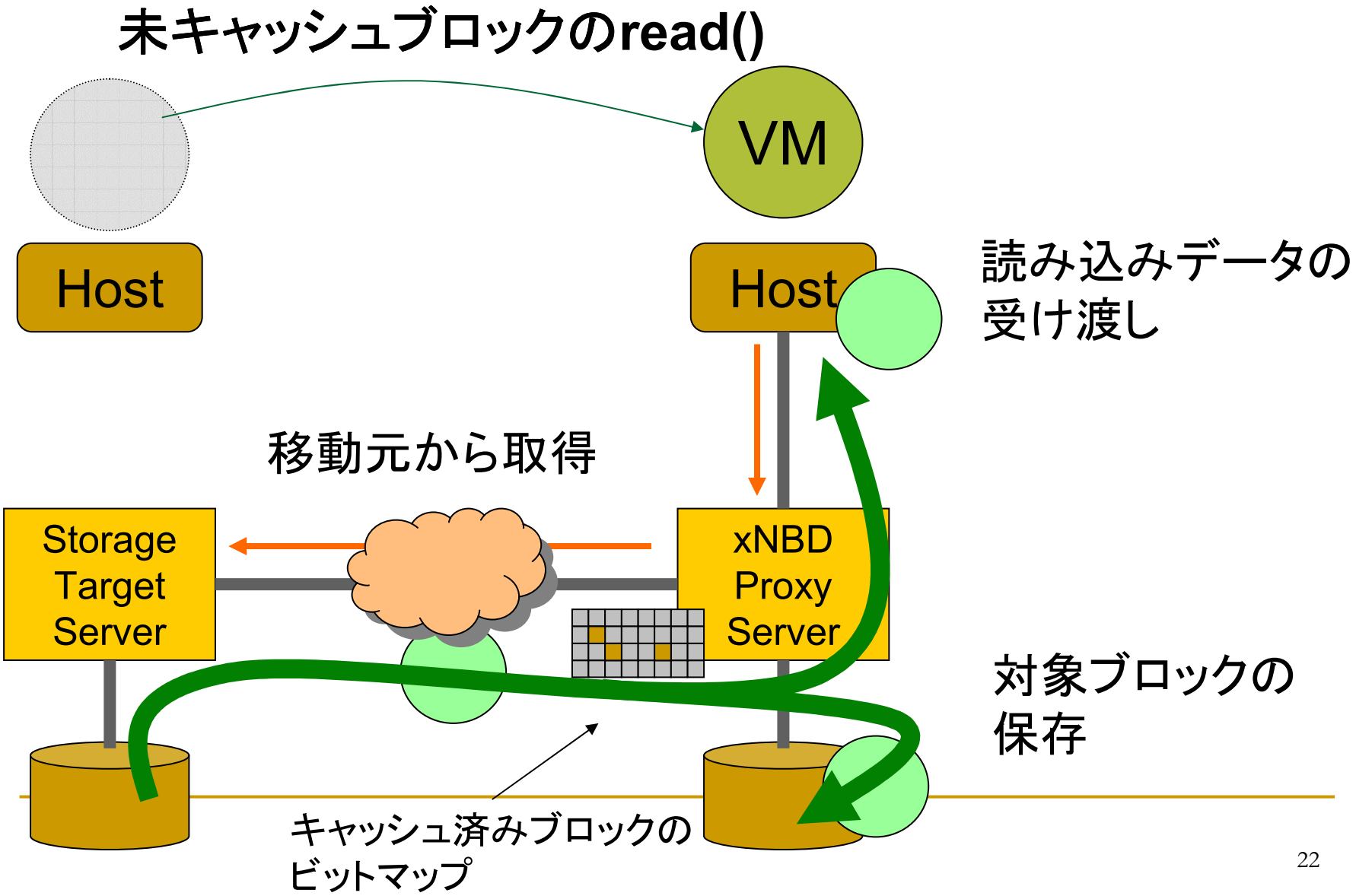


提案機構(5)

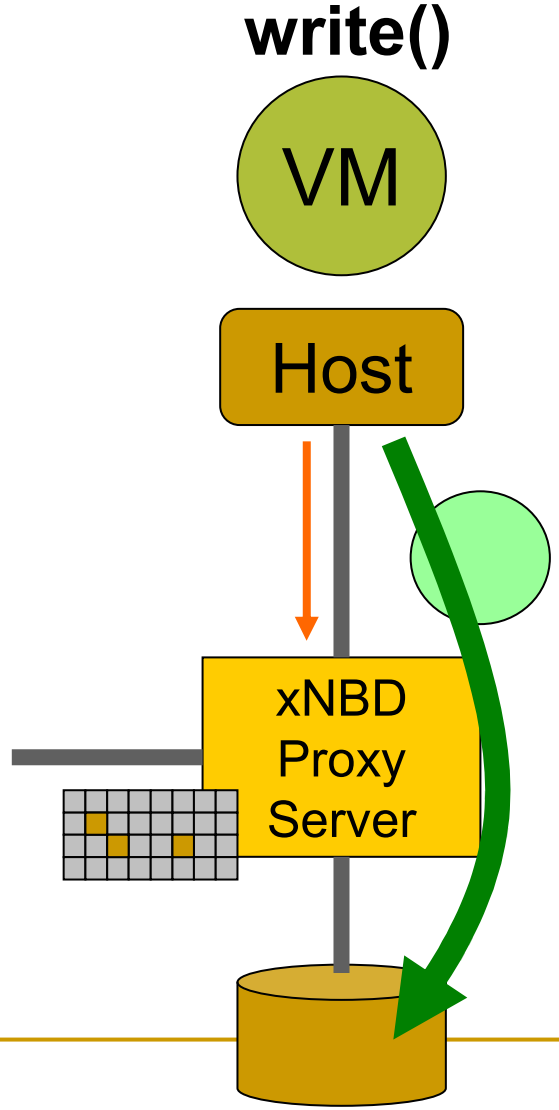


最終的にすべてのブロックの
キャッシュが完了すると、
通常のNBDサーバとして動作

基本動作(1)


未キャッシュブロックのread()
VM
Host
Host
読み込みデータの
受け渡し
移動元から取得
Storage
Target
Server
xNBD
Proxy
Server
対象ブロックの
保存
キャッシュ済みブロックの
ビットマップ

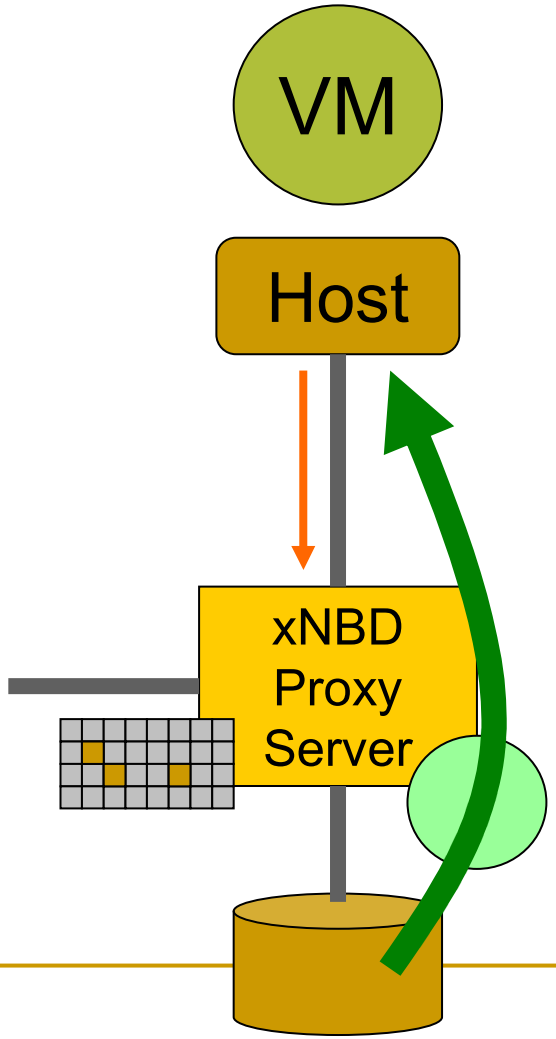
基本動作(2)



書き込みデータを保存し、
未キャッシュブロックだったならば
ビットマップを更新

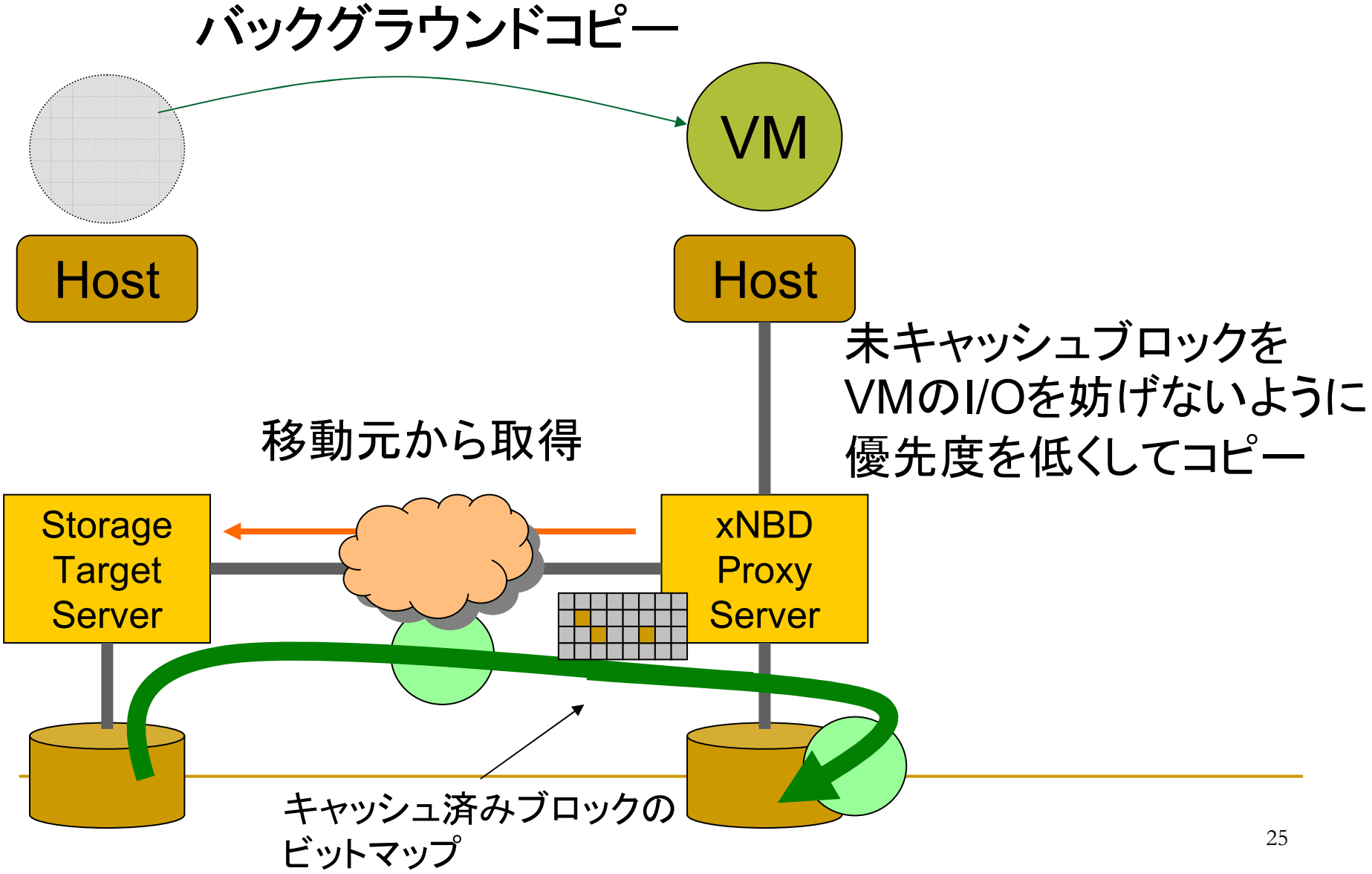
基本動作(3)

キャッシュ済みブロックのread()

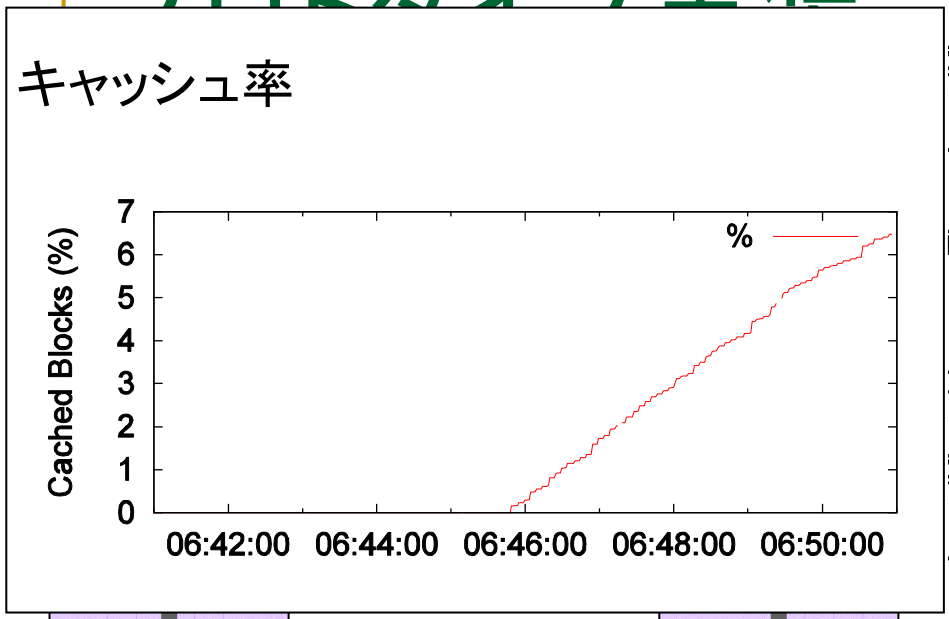


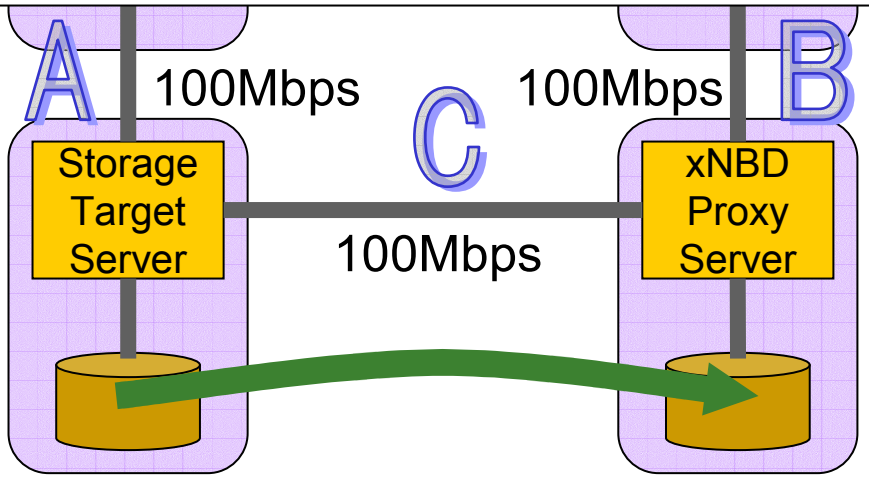
データの読み込み

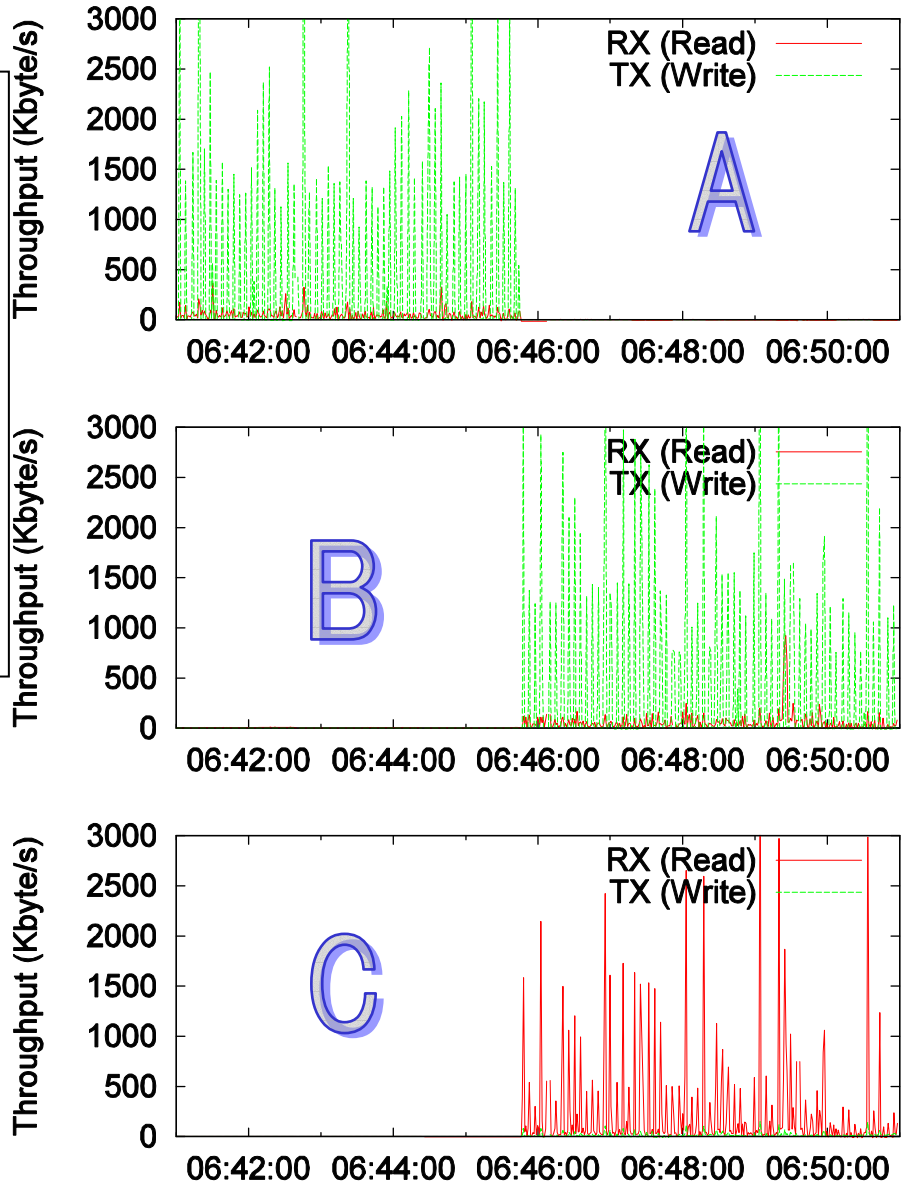
基本動作(4)


バックグラウンドコピー
VM
Host
Host
未キャッシュブロックを
VMのI/Oを妨げないように
優先度を低くしてコピー
移動元から取得
Storage
Target
Server
xNBD
Proxy
Server
キャッシュ済みブロックの
ビットマップ

プロトタイプ実装



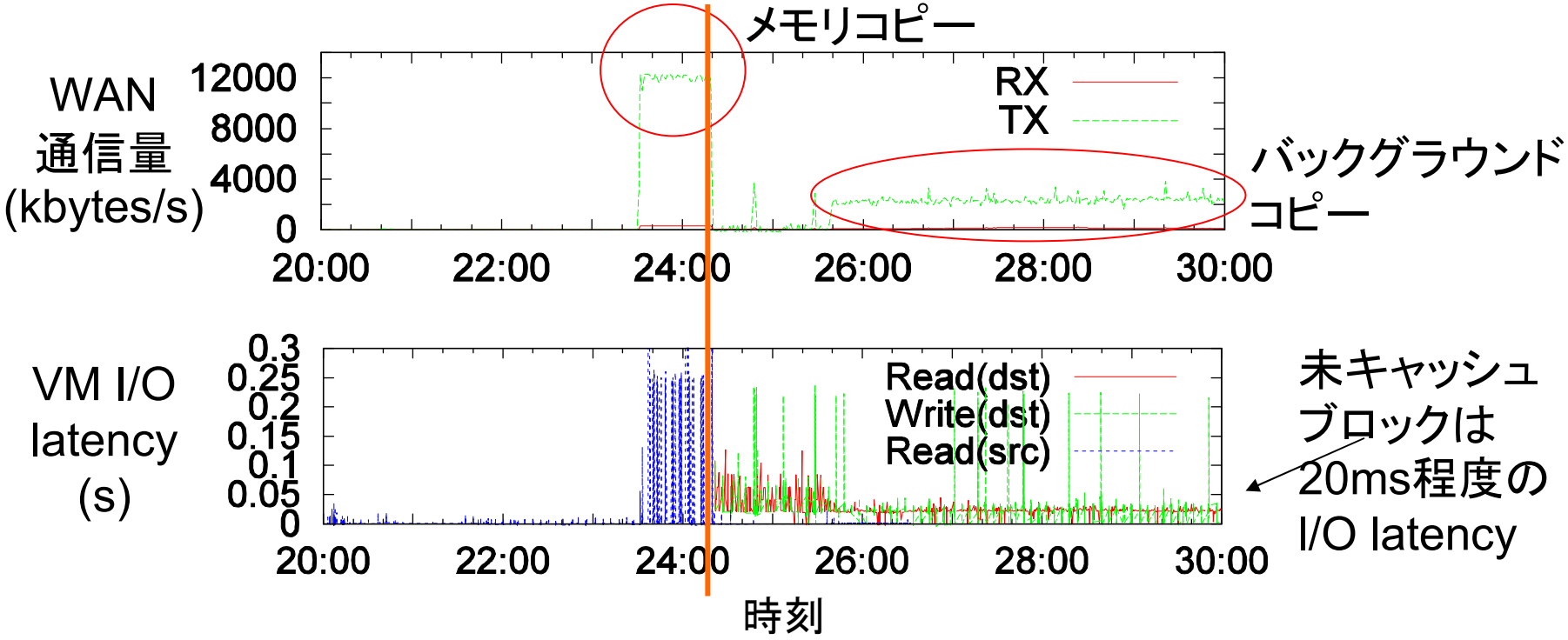




＊メモリコピートラフィックは別リンクを経由
＊バックグラウンドコピーは無効

今後の課題

- WAN環境における評価
- バックグラウンドコピー戦略の検討
  - VM I/Oへの影響を抑える



カーネルコンパイル中の遠隔マイグレーション
RTT = 20ms

結論

- 遠隔拠点間VMストレージ再配置機構の提案
  - ストレージデータの完全な再配置
  - SANへの統合
- ブロックレベルのストレージサーバ
  - オンデマンドフェッチ
  - ディスクブロックキャッシュ
  - バックグラウンドコピー
- プロトタイプ実装
  - 基本動作の確認