OpenTagを用いた網内データ複製による クラウドアクセス堅牢化

東京大学
古橋 亮慈, 中尾 彰宏

This work is supported by Ministry of Internal Affairs and Communications of the Japanese Government.

Agenda

1. Introduction
2. Design
3. Implementation
4. Evaluation
5. Conclusion

1. Introduction

研究の背景

- クラウド・プラットフォームへのアクセスにおける課題
  - アクセスの堅牢性向上
    e.g.) 災害時のデータ保護など
  - アクセス効率の向上
    e.g.) ピークタイムの集中など

→ ルータ上でのIn-Network Processing(ネットワーク内部パケット処理)が必要
（データ複製／ピークタイムのオフロードのためのデータバッファリングなど）

- ネットワーク内資源のスライス化
  - 任意の処理をスライスとして挿入可能
  - ネットワーク内でスライス毎に様々なパケット処理を実現

関連研究

- Active Networking
  - パケットへの実行コード埋め込みによるネットワーク内処理
  - Active Router上でパケット内実行コードを処理
  - 広帯域の要求に応え難い

  → *OpenTag[†] ではシンプルなリダイレクションを基本概念として採用*
- OpenFlow
  - フローに基づくスイッチング／リダイレクション
  - オペレータ主導のフロー認識によるIn-Network Processingに応用可能
  - P2Pなど非クライアント・サーバ型通信ではフロー認識が難しい
  - セキュリティ確保を伴う選択的リダイレクションの仕組みは特に無し

  → *OpenTag[†] では*
  - *ユーザー主導のタグ挿入によるIn-Network Processing*

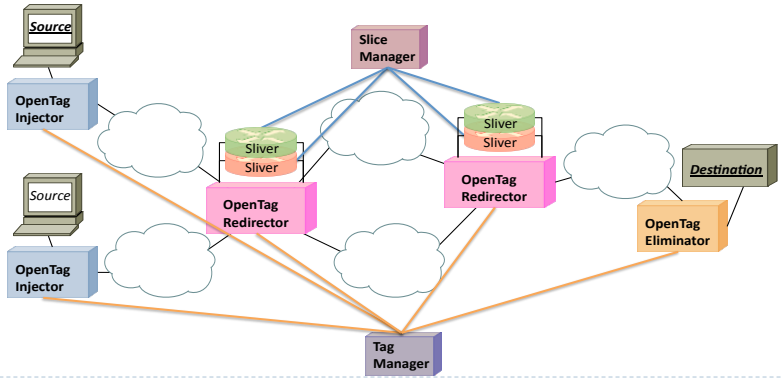
  （⇔オペレータ主導のフロー認識）
  - *パケット毎のスライス認証メカニズム*
  - *タグに基づくリダイレクションを採用*

[†]R. Furuhashi and A. Nakao. Opentag: Tag-based network slicing for wide-area coordinated in-network packet processing. In Communications Workshops (ICC), 2011 IEEE International Conference on, pages 1–5, June 2011.

2. Design

OpenTagのシステム構成

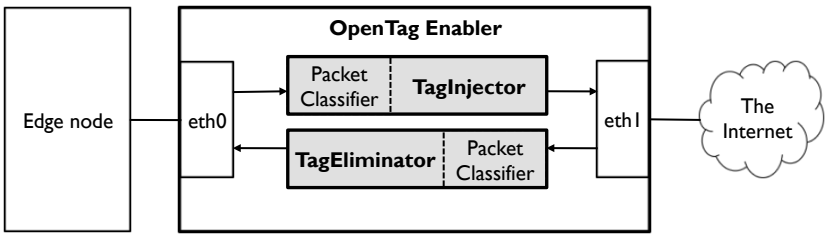
- Tag Manager：タグの登録受付・配布、暗号化キーの登録受付・配布
- Tag Injector：ユーザー主導でパケットにタグを挿入
- Tag Redirector：タグに基づいてパケットを振り分け
- Tag Eliminator：パケットからタグを除去
- Slice Manager：スライスのリソース管理（PlanetLab, CoreLab等*を活用）



*コンピュータ資源を広域で予約する仕組み

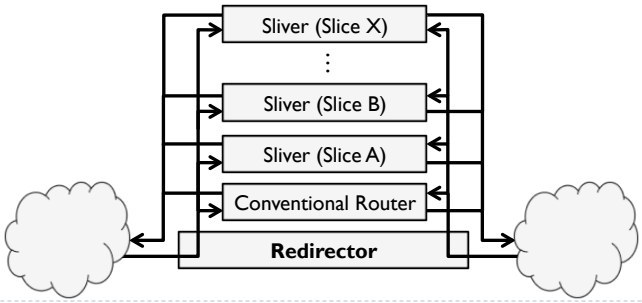
OpenTagにおけるEdgeの構造

- TagInjectorとTagEliminator
  - Edgeノードと外部ネットワークの間に配置
  - Packet classifierで対象パケットを抽出しタグを挿入／除去
  - 対象パケット以外にはBridgeとして振る舞う



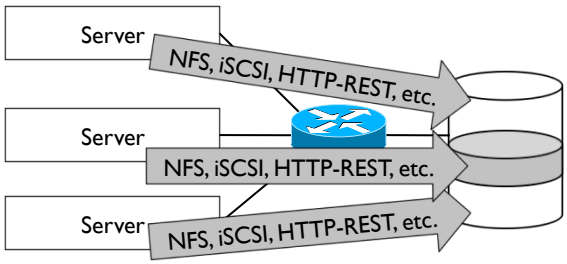
OpenTagにおける中継点の構造

- パケットのリダイレクション
  - リダイレクタがタグに基づいて各スライスのsliverに振り分け
  - タグが付いていない場合はconventional routerへ
- リダイレクション性能の考慮
  - パケットは転送後リダイレクタを通らず直接送出



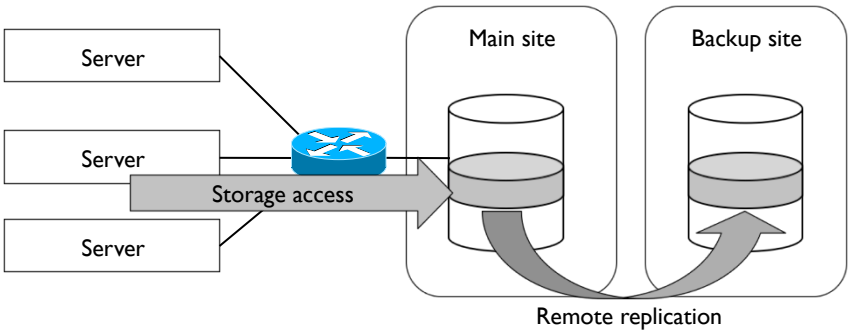
アプリケーションの検討

- ネットワークを介したストレージアクセス
  - NFS, iSCSI, HTTP-RESTを用いたクラウドストレージサービス (Amazon S3など)
- ネットワーク上のストレージを活用するアプリケーション
  - 災害時のデータ保護
  - ピーク時帯域使用のオフロード



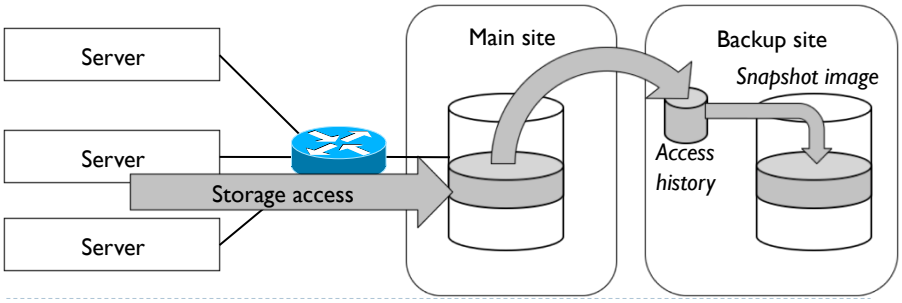
従来のリモートレプリケーション

- 災害復旧のためのデータ保護
  - メインサイトの遠隔地にバックアップサイトを設置
  - メインサイトにおけるストレージのデータをバックアップサイトのストレージにコピー



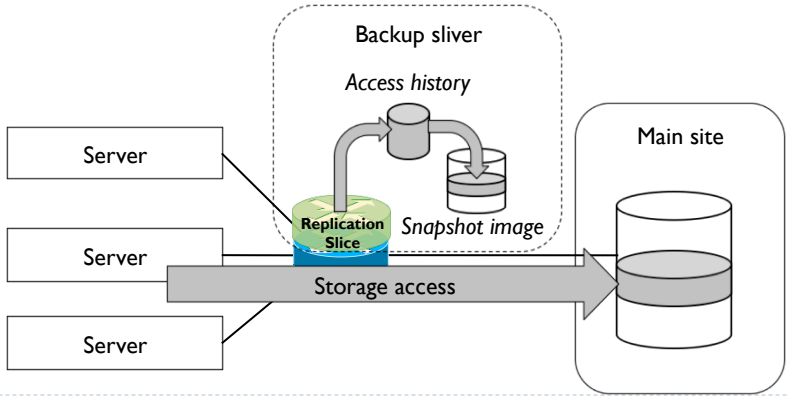
従来のリモートCDP (Continuous Data Protection)

- 継続的なデータ保護
  - アクセス履歴をバックアップサイトに送って蓄積
  - バックアップストレージのスナップショットイメージにアクセス履歴を適用することにより任意のリカバリポイントを使用可能
    (一般的なリモートレプリケーションでは一時点のリカバリポイントのみ)



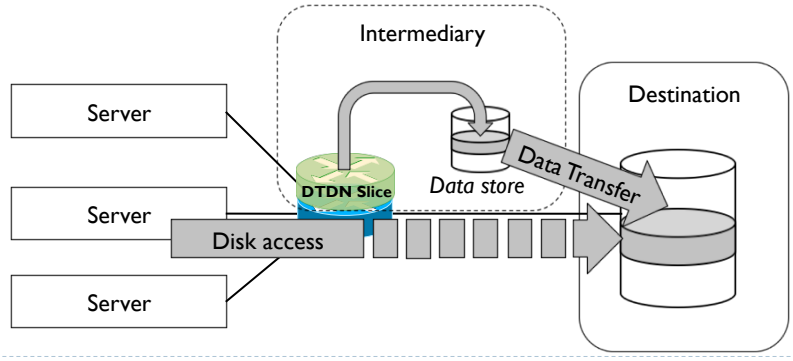
OpenTagを用いたリモートCDP

- OpenTagを用いてネットワーク内でアクセス履歴を蓄積
- ユーザー主導でリモートCDP機能を挿入
  - 既存ストレージサービスに変更無しで機能を追加可能



Delay Tolerant Data Networking (DTDN)

- “データ”に着目したDTN†アプリケーション
  - ピーク時帯域使用の”タイムシフト”を実施
  - OpenTagによるスライスにデータストアを配置
  - ストレージアクセスを一時的にデータストアで保持

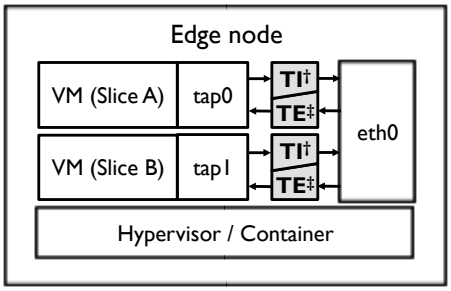


†DTN: Delay Tolerant Network

3. Implementation

エッジノードの実装

- 各スライスを仮想マシン(VM)として配置
- TagInjector/TagEliminatorをVM-NIC間のbridgeとして実装
- Click Modular Router[1]による実装

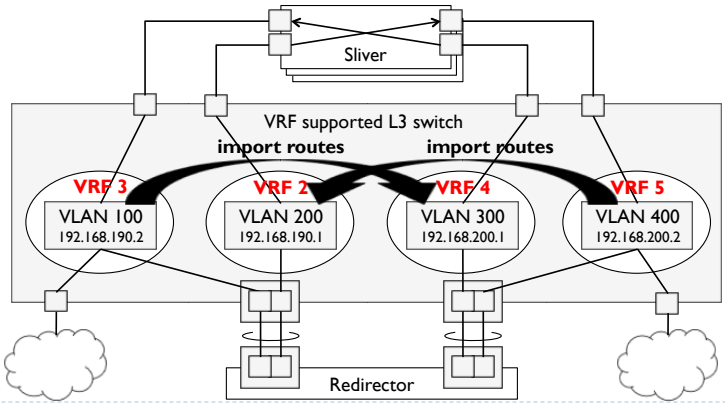


† Tag Injector
‡ Tag Eliminator

[1] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek. The click modular router. ACM Trans. Comput. Syst., 18(3):263–297, 2000.

中継点の実装

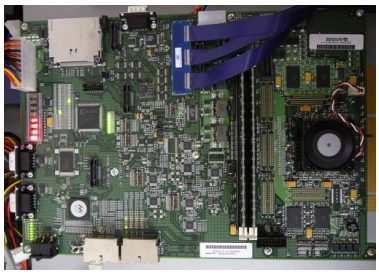
- スライスの機能はSliver上に実装
- 通常のパケットにはL3スイッチとして振る舞う
- VRF†機能を用いて構築
  - ルーティング情報のVRF間インポート機能を用いる

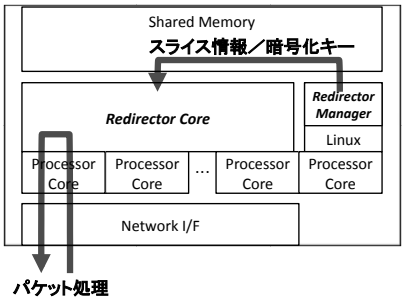


†Virtual routing and forwarding

リダイレクタの実装

- ネットワーク処理向けメニーコアSoC「OCTEON」使用
  - ハイエンドルータのアドオンカードに採用
  - 今回はOCTEON CN5860(800MHz x 16 cores)評価ボード上に実装
- OSレスのリダイレクタ・コアとLinux上管理プログラムを混載
  - 共有メモリを介してスライス情報／暗号化キーをやり取りする

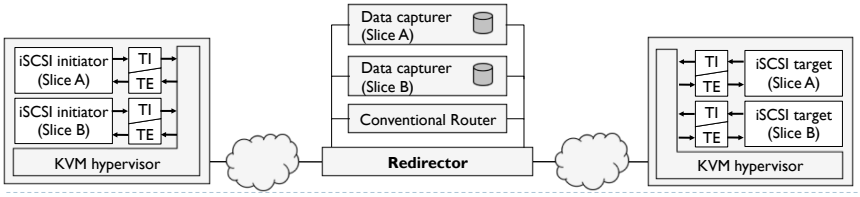




4. Evaluation

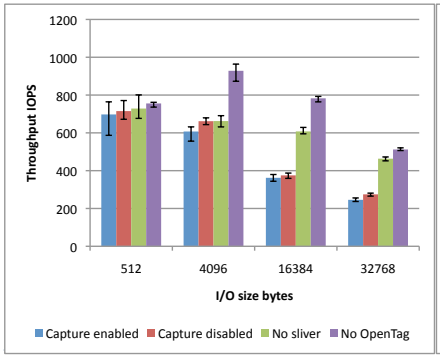
評価環境

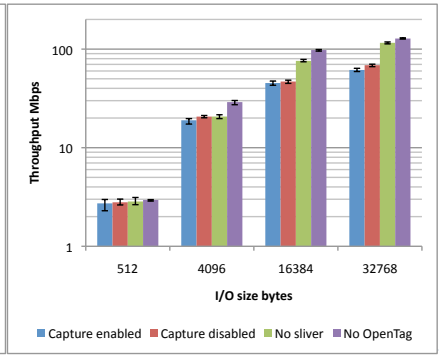
- Edgeノードとして**iSCSI**イニシエータ/ターゲットをKVM上に配置
  - Ubuntu LinuxのOpen-iSCSIによるiSCSIイニシエータ
  - OpenFilerによるiSCSIターゲット
- 中継点にiSCSIアクセス履歴保持スライスを実装
  - Ubuntu Linux上のClick Modular Routerによる実装
- IOMETERによりIOPS, Mbps値を測定
  - ブロックサイズ: 512, 4K, 16K, 32K bytes
  - シーケンシャル・アクセス、10秒間



1スライスでの評価

- **評価ケース**:
  キャプチャ有効／キャプチャ無効／Sliver無し／OpenTag無し
- **オーバーヘッドの測定 (ブロックサイズ=512B, 4KB, 16KB, 32KB)**
  キャプチャによるオーバーヘッド: 2.5%, 8.2%, 3.4%, 10.6%
  Sliverによるオーバーヘッド: 1.8%, 0.1%, 38.3%, 40.7%
  OpenTagによるオーバーヘッド: 3.6%, 28.6%, 22.4%, 9.5%

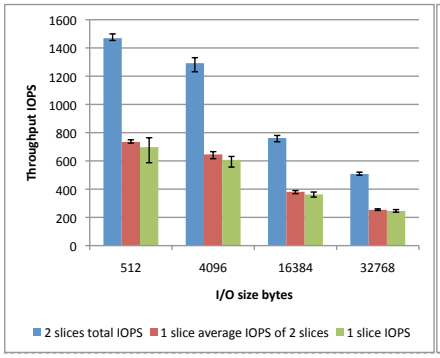


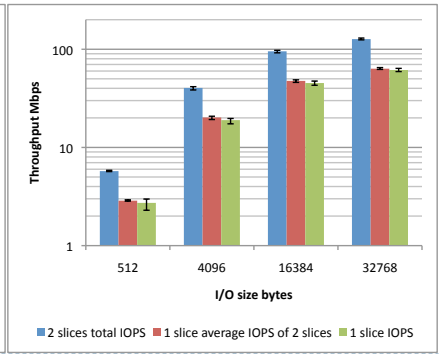


2スライスでの評価

- 2スライス／1スライスの比較(キャプチャ有効時)
  - 2スライス時の合計が1スライスの概ね2倍
- 2スライス／1スライス共にデータロス無し

→ スライス間の相互干渉無し

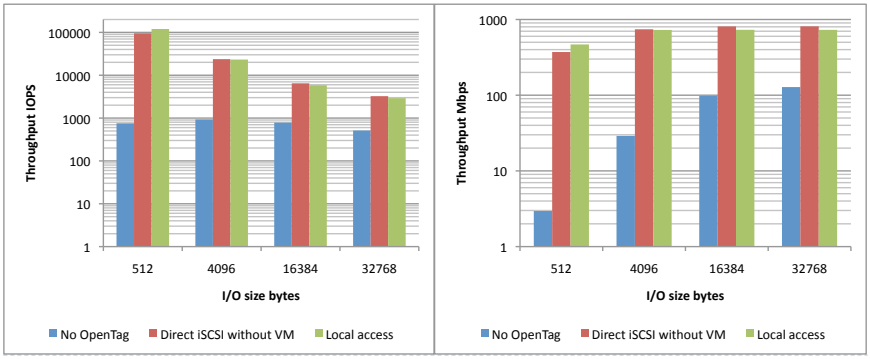




VM／iSCSIによるオーバーヘッド評価

- 評価ケース
  - OpenTag無し(但しVM有り)／iSCSI直接接続／ローカルアクセスの比較
- VMに関するオーバーヘッドにより5倍～150倍の差
- iSCSI直接接続とローカルアクセスには大きな差は無し

→VMに関するオーバーヘッド大



5. Conclusion

成果

- 実アプリケーションに向けたOpenTagアーキテクチャの設計
  - OpenTagアーキテクチャにおけるEnd-to-endデータパスの構成
  - リダイレクタ、Sliver、VRF機能付きルータからなる中継点の構成
  - EdgeノードにおけるTagInjector、TagEliminatorの配置
  - ネットワーク上のストレージを活用するアプリケーションの検討
- End-to-endデータパス上の構成要素のプロトタイプ実装
  - Click Modular Routerを用いたTagInjector/TagEliminatorの実装
  - 中継点におけるネットワーク構成とスライス機能の実装
- OpenTag上のクラウドアクセス堅牢化アプリケーションの評価
  - iSCSIを用いた実現可能性の確認
  - IOMETERによる性能評価とスライス間isolationの確認
  - データパスにおけるオーバーヘッドの明確化

今後の予定

- オーバーヘッドの除去と性能チューニング
- 商用ネットワーク機器への実装
- 広範なクラウドプラットフォームを用いた評価
- スライス上への多様なアプリケーションの実装

References

[1] D. Alexander, W. Arbaugh, M. Hicks, P. Kakkar, A. Keromytis, J. Moore, C. Gunter, S. Nettles, and J. Smith. The switchware active network architecture. Network, IEEE, 12(3):29 –36, may/jun 1998.

[2] S. Bhatia, M. Motiwala, W. Muhlbauer, Y. Mundada, V. Valancius, A. Bavier, N. Feamster, L. Peterson, and J. Rexford. Trellis: a platform for building flexible, fast virtual networks on commodity hardware. In CoNEXT '08: Proceedings of the 2008 ACM CoNEXT Conference, pages 1–6, New York, NY, USA, 2008. ACM.

[3] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman. PlanetLab: an overlay testbed for broad-coverage services. SIGCOMM Comput. Commun. Rev., 33(3):3–12, 2003.

[4] N. Egi, A. Greenhalgh, M. Handley, M. Hoerdt, F. Huici, and L. Mathy. Fairness issues in software virtual routers. In PRESTO '08: Proceedings of the ACM workshop on Programmable routers for extensible services of tomorrow, pages 33–38, New York, NY, USA, 2008. ACM.

[5] N. Egi, A. Greenhalgh, M. Handley, M. Hoerdt, F. Huici, and L. Mathy. Towards high performance virtual routers on commodity hardware. In CoNEXT '08: Proceedings of the 2008 ACM CoNEXT Conference, pages 1–12, New York, NY, USA, 2008. ACM.

[6] K. Fall. A delay-tolerant network architecture for challenged internets. In Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, SIGCOMM '03, pages 27–34, New York, NY, USA, 2003. ACM.

[7] R. Furuhashi and A. Nakao. Opentag: Tag-based network slicing for wide-area coordinated in-network packet processing. In Communications Workshops (ICC), 2011 IEEE International Conference on, pages 1–5, June 2011.

[8] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek. The click modular router. ACM Trans. Comput. Syst., 18(3):263–297, 2000.

[9] KVM: Kernel-based Virtual Machine. http://www.linux-kvm.org/.

[10] J. W. Lee, R. Francescangeli, J. Janak, S. Srinivasan, S. Baset, H. Schulzrinne, Z. Despotovic, and W. Kellerer. Netserv: Active networking 2.0. In Communications Workshops (ICC), 2011 IEEEInternational Conference on, pages 1 –6, june 2011.

[11] LXC: Linux Containers. http://lxc.sourceforge.net/.

[12] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. OpenFlow: enabling innovation in campus networks. SIGCOMM Comput. Commun. Rev., 38(2):69–74, 2008.

[13] A. Nakao, R. Ozaki, and Y. Nishida. CoreLab: an emerging network testbed employing hosted virtual machine monitor. In CoNEXT '08: Proceedings of the 2008 ACM CoNEXT Conference, pages 1–6, New York, NY, USA, 2008. ACM.

[14] E. Rosen, A. Viswanathan, and R. Callon. RFC3031: Multiprotocol Label Switching Architecture, 2001.

[15] B. Schwartz, A. Jackson, W. Strayer, W. Zhou, R. Rockwell, and C. Partridge. Smart packets for active networks. In Open Architectures and Network Programming Proceedings, 1999. OPENARCH '99. 1999 IEEE Second Conference on, pages 90 – 97, mar 1999.

[16] D. Wetherall, J. Guttag, and D. Tennenhouse. ANTS: a toolkit for building and dynamically deploying network protocols. pages 117 – 129, apr. 1998.