大阪大学
OSAKA UNIVERSITY

プログラマブルスイッチと計算機を組み合わせた ICN ルータにおけるキャッシュの高速化に関する一考察

2022/12/13

武政 淳二、小泉 佑揮、長谷川 亨

大阪大学

本研究は、科研費 20H04176 によるものである。

高速な ICN ルータ実装

[1]Junji Takemasa, et.al., “Vision: toward 10 Tbps NDN forwarding with billion prefixes by programmable switches,” ACM ICN 2021.

■ 目標

- テラバイトのキャッシュ容量とテラビット/秒のフォワーディング速度を両立する ICN ルータ

■ 有望なハードウェア基盤

- 計算機とプログラマブルスイッチ(以降スイッチ)の組み合わせ
  - 単一のハードウェアのみだと、速度と容量の両立が難しい[1]
    - 計算機: 100 Gbps (CPU)、1TBのメモリ容量 (DRAM)
    - スイッチ: 10 Tbps (ASIC)、100MBのメモリ容量 (SRAM)

■ 課題

- ボトルネックの明確化
  - 異なるハードウェアをまたがるパケット処理が律速箇所を複雑化
- キャッシュ機能の設計
  - 計算機のメモリとスイッチの速度を最大限活かす機能の分離・配置が必要

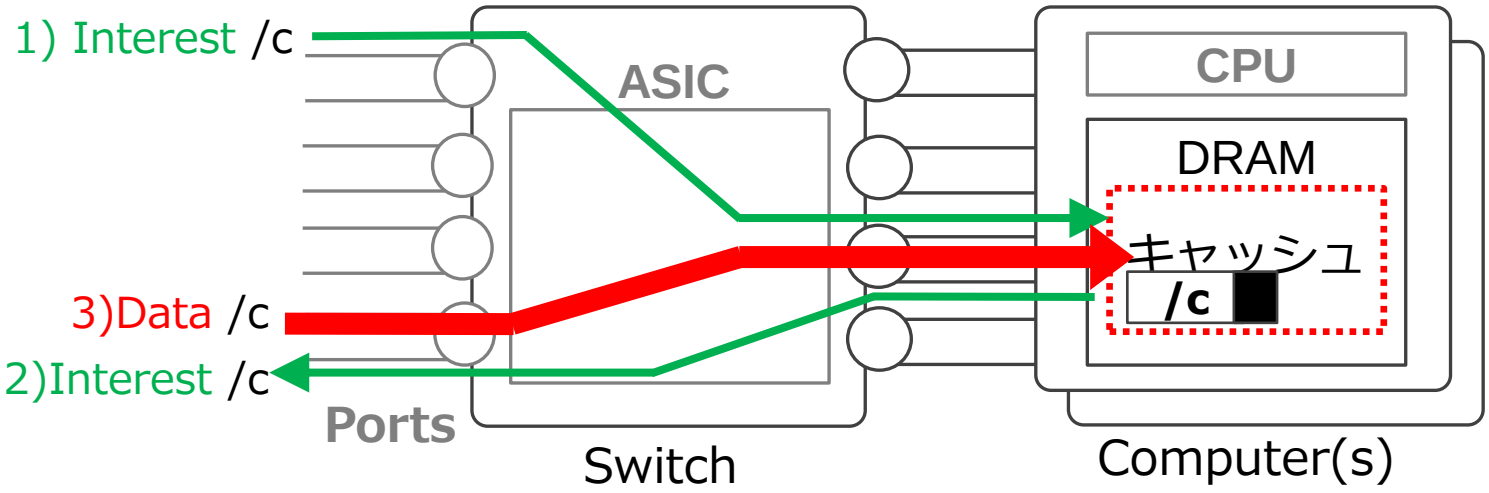
計算機とスイッチを組み合わせたルータ

- **システムアーキテクチャ**
  - スイッチ: 外部とのパケット転送
  - 計算機: DRAM上にData パケットをキャッシュ
- **処理フロー (キャッシュミスする場合)**

1)スイッチで受信したInterest パケットを計算機へ伝送しCSを検索

2)CSでミスすると、Interestを外部へ転送

3)スイッチで受信した Data パケットを計算機へ伝送し CS へ挿入

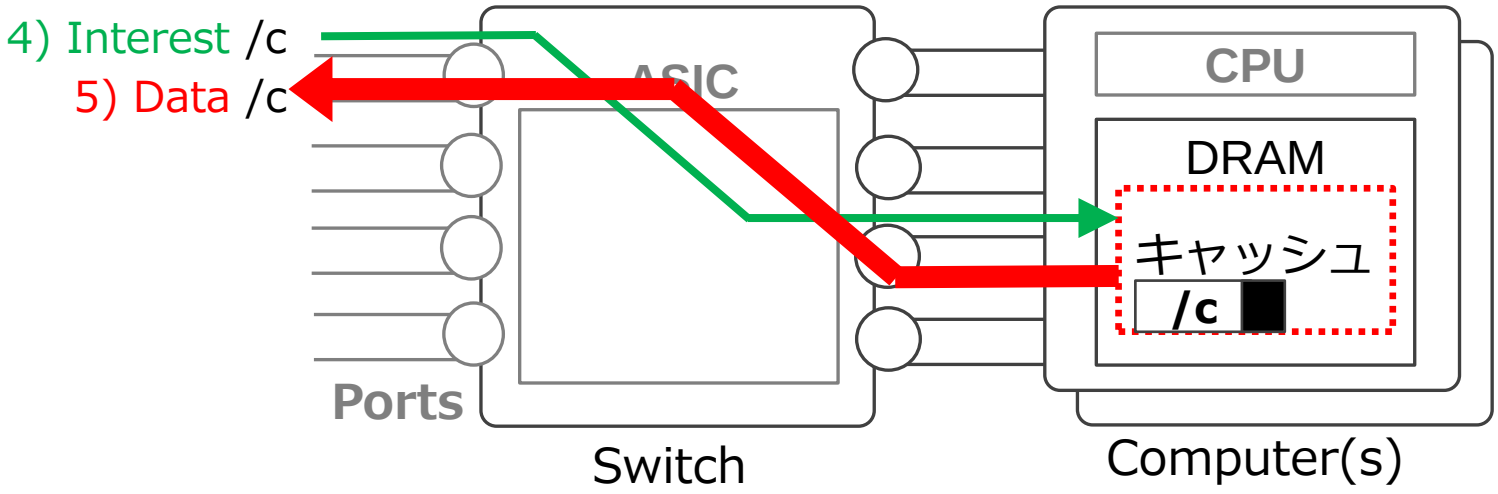


計算機とスイッチを組み合わせたルータ

- **システムアーキテクチャ**
  - スイッチ: 外部とのパケット転送
  - 計算機: DRAM上にData パケットをキャッシュ
- **処理フロー (キャッシュヒットする場合)**

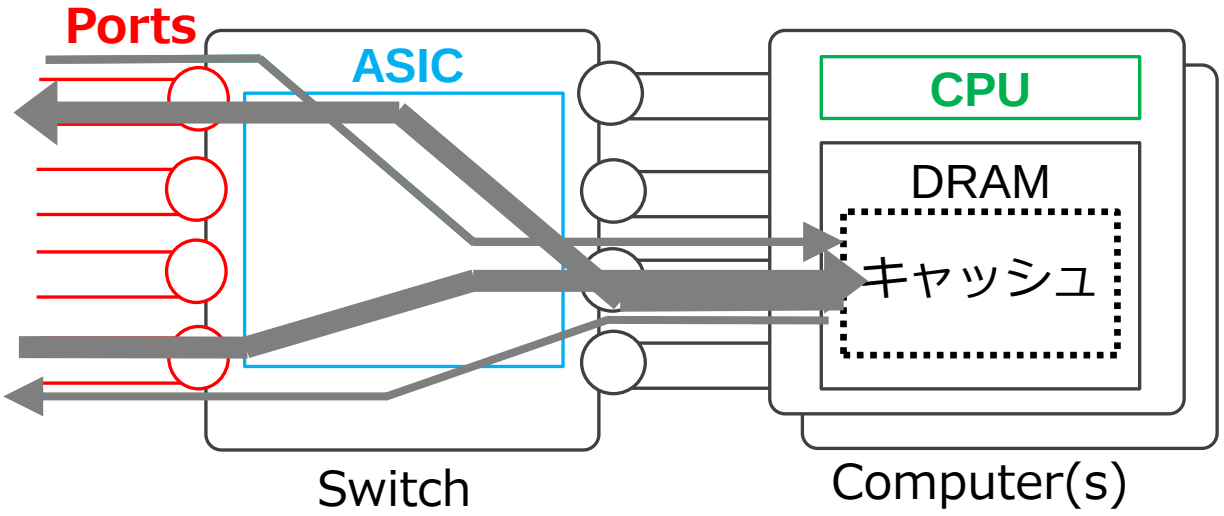
4)スイッチで受信したInterest パケットを計算機へ伝送しCSを検索

5)CSでヒットすると、計算機からスイッチへ Data パケットを読み出し外部へ転送



ボトルネックの解析方法

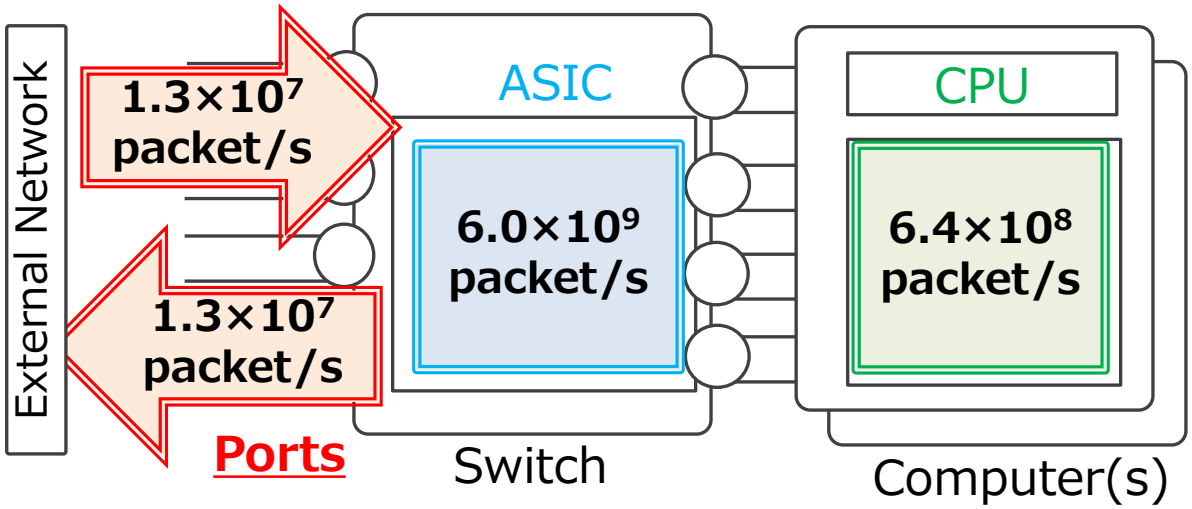
- **パケットが通過する各ハードウェアにおける、パケット処理レートの上限 (packet/s) をモデル化**
  - **ポート**: ポートの帯域 [bit/s] ÷ パケットサイズ [bit/packet]
  - **ASIC**: Intel Tofino ASICを想定し、カタログ値を利用[2]
  - **CPU**: 計算機にICNのフォワーディングとキャッシュを実装し測定[3]
- **律速箇所:パケット処理レートが最小のハードウェア**



[2]複数パイプラインに均一にパケットが分散され、各パイプラインが1クロック1パケットを処理する想定
[3]Intel DPDK でパケットI/O、ハッシュテーブルでフォワーディングテーブルとキャッシュのデータ構造を実装

解析結果

- **律速箇所: ポート**
  - フォワーディング速度はASICやCPUの速度でなく**ポートの帯域**により制限

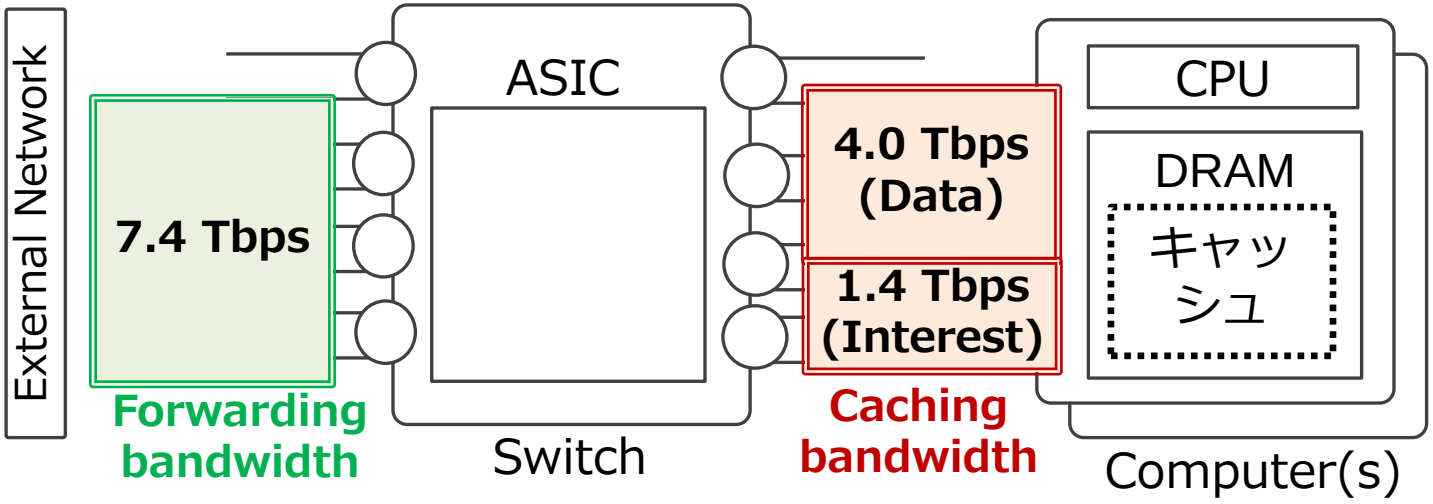


- **解析条件 (キャッシュヒット率30%を仮定)**

Switch: Intel Tofino 2	128 ports×100 Gbps 6×10^9 packet/s ASIC
CPU: AMD EPYC	128 cores
Measured processing rate per CPU core	5×10^6 packet/s
Packet size	128B Interest、1024B Data

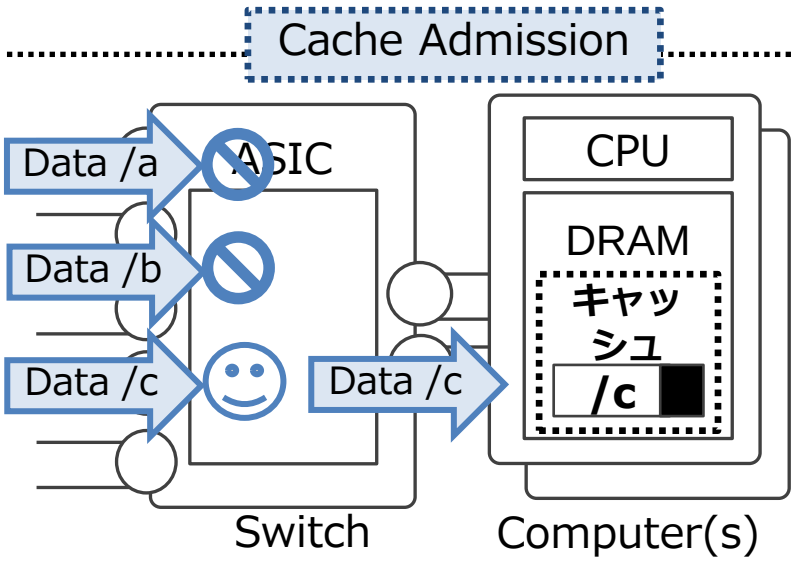
キャッシュへのアクセスによるフォワーディング速度の低下

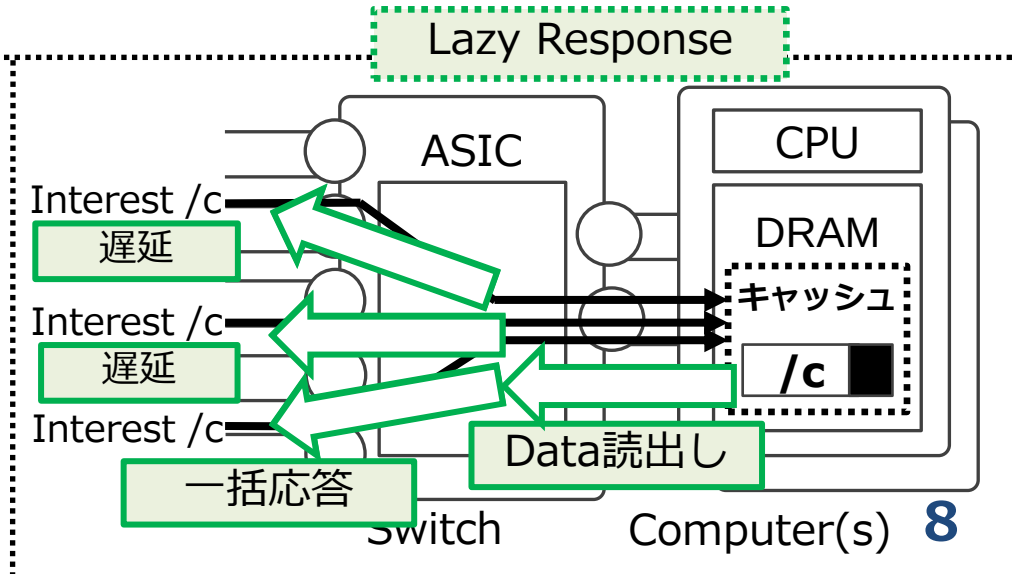
- **キャッシュへのアクセスにポートの帯域を割く必要**
  - 本来、外部とのパケット転送に割いていたポートの帯域を、計算機上のキャッシュへのアクセスに割く必要 (12.8 Tbpsのうち5.4 Tbps)
  - Interest よりもサイズの大きい Data パケットが帯域浪費の主要因
- **フォワーディング速度向上への要件**
  - **スイッチと計算機間の Data パケットの伝送量を削減する必要**
    - スイッチ→計算機: キャッシュミス時の Data パケットの挿入
    - スイッチ←計算機: キャッシュヒット時の Data パケットの読み出し



キャッシュ機能の設計方針

- **要件1: スイッチ→計算機のDataパケットの伝送量削減**
  - 方針: スイッチ上で不要な Data パケットの挿入をフィルタ (**Cache Admission**)
    - ミスする Data パケットのうち、90%程度はヒットせずに追い出される
    - 将来ヒットしない Data パケットをキャッシュに挿入するのが無駄
- **要件2: スイッチ←計算機の Data パケットの伝送量削減**
  - 方針:短時間内の同じ Data への要求を遅延させ一括で応答 (**Lazy Response**)
    - 人気な Data パケットはO(1)μs間隔で要求されるが、インターネットのRTTはO(10ms)
    - 応答を遅延させ複数の要求に対する応答を纏める余地がある





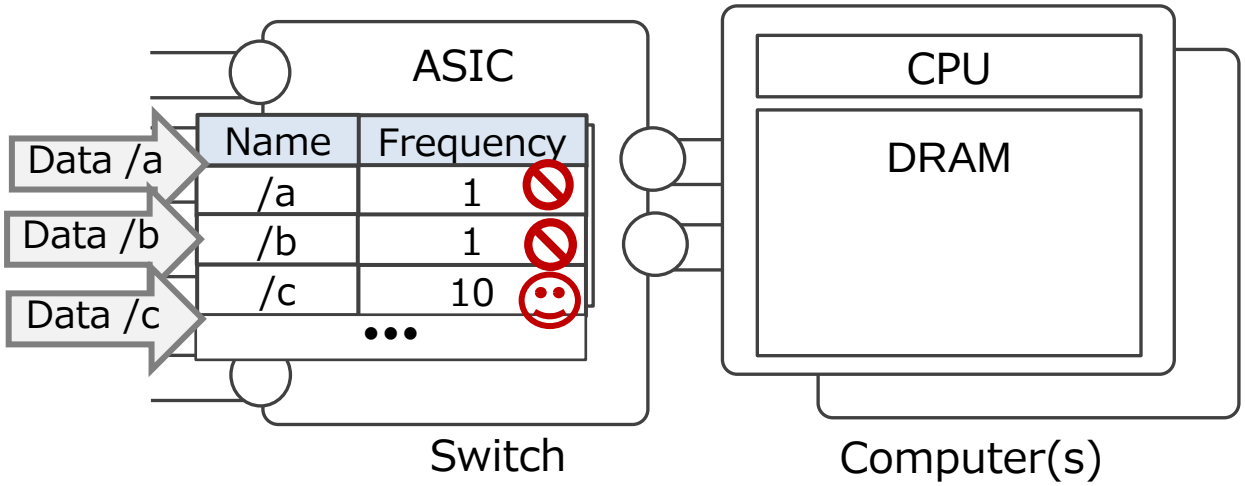
Cache Admission: 設計目標と課題

■ 設計目標

- 過去の要求頻度に基づき、挿入を判別する機能をスイッチで実行
  - 既存研究で高いヒット率と低い挿入率を検証済み (LFU、TinyLFU)

■ 設計上の課題

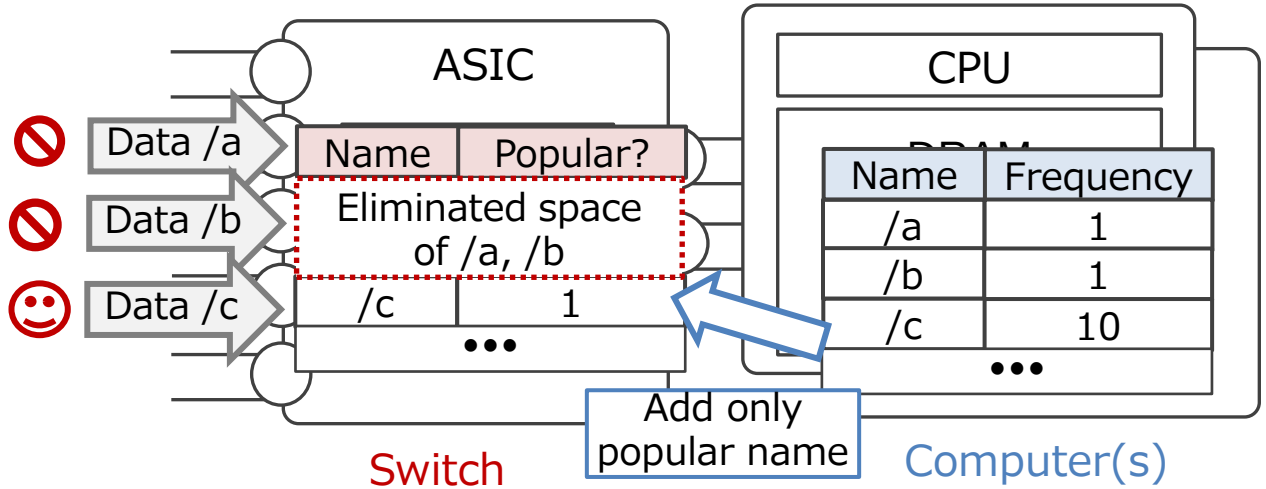
- メモリ容量が小さく、要求頻度を ASIC上に保持できない
  - ASICのメモリ容量: 100MB (SRAM)、要求頻度のデータ構造: 10 GB



Cache Admission: 設計

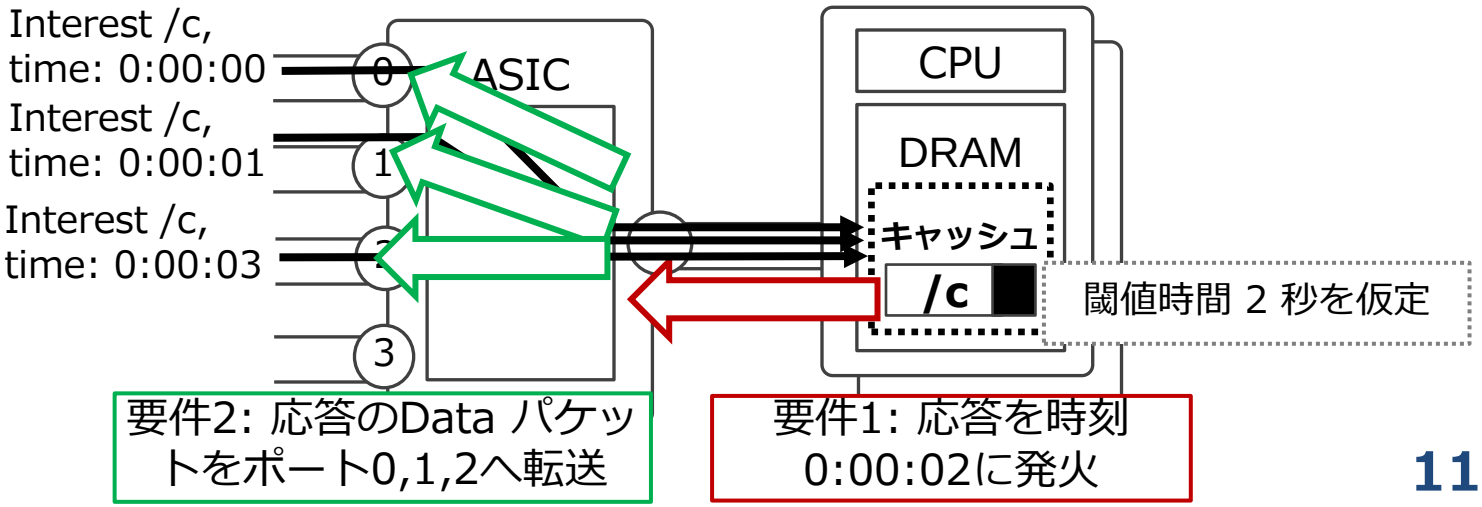
- **帯域の消費を抑えながら、メモリを消費する機能を計算機へオフロード**

1. 計算機: **要求頻度を記録**しスイッチへ人気な Data パケットを通知
   - Interest パケットのサイズが小さいため、スイッチと計算機間の帯域消費は問題ない
2. スイッチ: Data パケットが人気かを1bitのフラグで表現し**判別**
   - 不人気な Data パケットに対しメモリを消費する必要がない (94%)
   - 16-bit の要求頻度のカウンタを1-bit のフラグへ削減



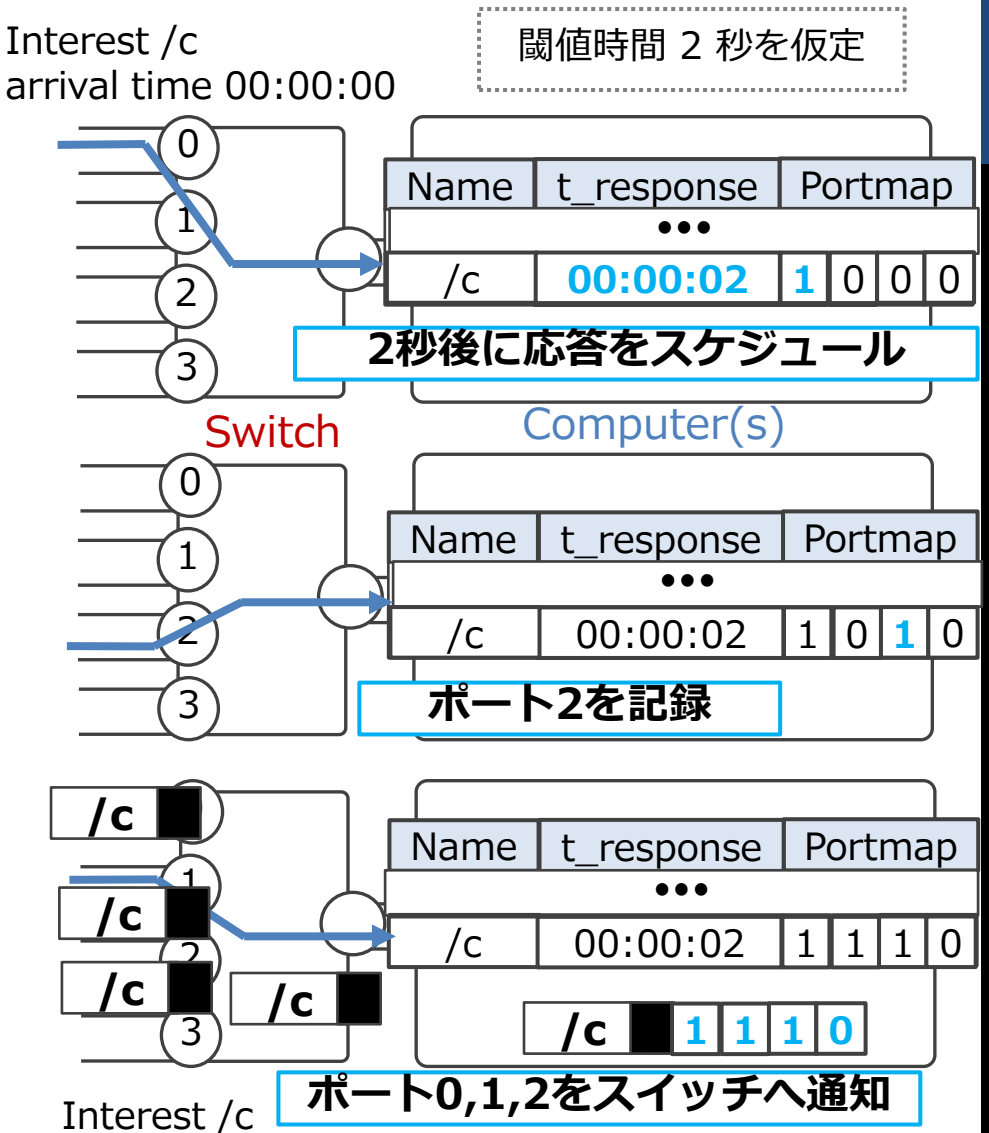
Lazy Response: 設計目標と要件

- **設計目標**
  - 一定時間内の同じ Data への要求を遅延させ、一括で応答
- **設計上の要件**
  1. 一定の時間(閾値時間と呼ぶ)が経過後に、応答を発火する必要
  2. Interest を受信した全てのポートに対し、応答を転送する必要



Lazy Response: 設計

1. **応答の発火**
   - 最初の Interest パケットの到着時、遅延時間後に応答をスケジュール
   - 後続の Interest パケット到着時に、応答時刻を超えると発火
2. **応答の転送**
   - Interest パケットを受信したポートを、応答の転送先としてビットマップへ記録
   - ビットマップに記録されたポートを全てスイッチへ通知
   - 通知されたポートへ Data パケットを複製し転送



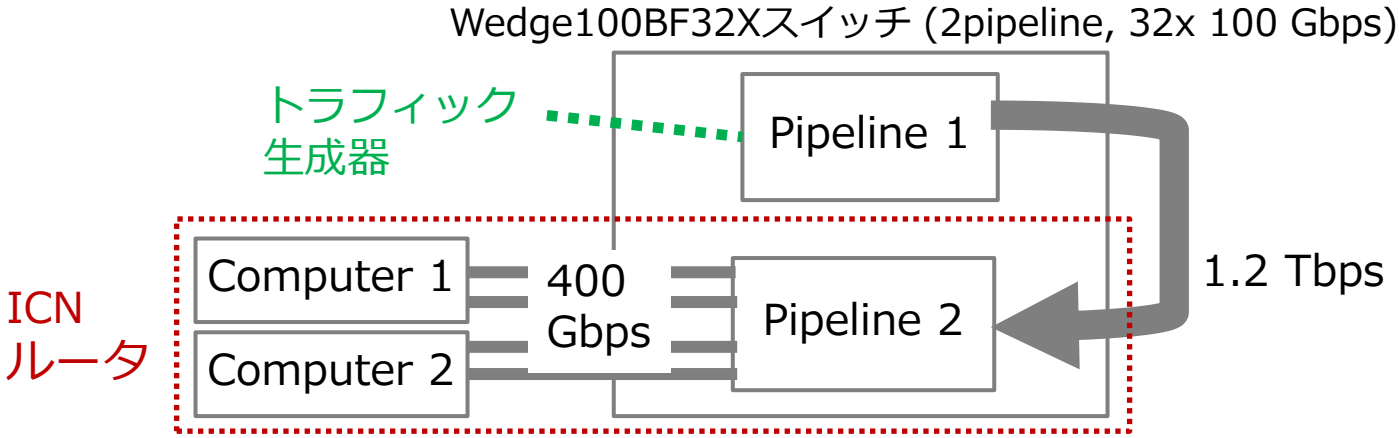
性能評価

評価環境

- Cache admission、Lazy Response を有するルータをスイッチと2台の計算機で実装
  - スイッチ (Wedge100BF32X): 100 Gbpsのポート 32 本、パケット処理パイプライン2基
  - 計算機: 2基の100 Gbps NIC (Intel E810)、2基の22コアCPUを備える
- 理想的なフォワーディング速度: 1.2 Tbps
  - 16ポートと1パイプラインで ICN ルータを実装 (残りはトラフィック生成器に利用)
  - 12ポートをフォワーディング用、 4ポートを計算機上のキャッシュとの接続に割り当て

評価シナリオ

- トラフィック生成器: 10^9 個のDataパケットをZipf分布(0.9)に従い要求するConsumer、Interestで要求された Dataを応答する Producer
- キャッシュ: 36%のキャッシュヒット率(128Bytesの容量)、Lazy Response で10msの応答遅延 (インターネットの平均RTTの1/6)



Naïve: キャッシュを実装した ICN ルータ
Proposal: Naïve + Cache admission + Lazy response

評価結果

■ フォワーディング速度

- Cache admission と Lazy response によりフォワーディング速度を向上し、1Tbpsに近い性能を達成
- プロトタイプは理想的なフォワーディング速度を達成できていない
  - 理想的: スイッチと計算機間の Data パケットの伝送量を0と仮定して予測したフォワーディング速度

	Prototype	Ideal speed (model)
Proposal	916 Gbps	1176 Gbps
Naïve	543 Gbps	900 Gbps

■ Dataパケットの伝送量

- スイッチ → 計算機: Cache admission により94%削減
- スイッチ ← 計算機: Lazy Response により30%削減
  - プロトタイプの速度が理想的な速度と乖離する要因

まとめ

- **スイッチと計算機を組み合わせ、大容量キャッシュと高速フォワーディングを両立する ICN ルータを構築**
  - ボトルネック解析
    - ASIC・CPU・ポートにより制限されるパケット処理レートの上限をモデル化し、ポートの帯域がパケット処理を律速することを明確化
    - スイッチと計算機間の Data パケットの伝送量削減を高速化の要件として抽出
  - Data パケットの伝送量を削減するキャッシュ機能の設計
    - Cache Admission: 不要な Data パケットの挿入をフィルタ
    - Lazy Response: 短時間内の同じ Data に対する要求を遅延させ、一括で応答
  - プロトタイプ実装
    - 1台のスイッチと2台の計算機で916 Gbps を実証