リソース分離型 マイクロデータセンターの ネットワーク構成評価のための 資源間通信シミュレータの設計

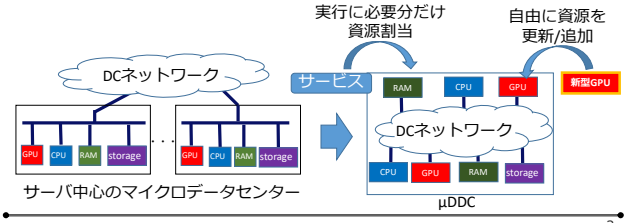
大阪大学 大学院情報科学研究科 村田研究室
博士後期課程 2 年
生駒 昭繁

2024/3/1 1

1

研究背景

- マイクロデータセンターを利用したエッジサービスの提供
  - 通信遅延の削減やトラヒックの削減を実現
  - 大規模データセンターと比べ保有資源に限りがある
- リソース分離型マイクロデータセンター (μDDC) が提案
  - 資源単位で構成されたマイクロデータセンター
  - 柔軟なスケーリングや資源利用率の向上の実現

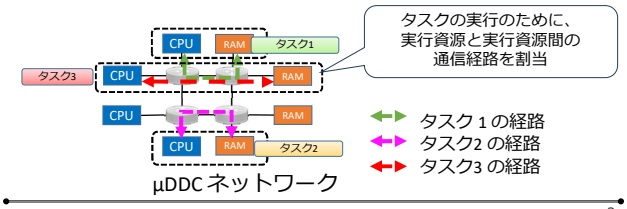


2

2

μDDC の課題

- 実行する資源間で通信をしながらタスクを実行
- 従来のアーキテクチャでは発生しない、資源間の通信遅延が性能を低下させる
  - 資源の性能が通信遅延に依存
- 資源間が、タスクの性能要件を満たすために十分低遅延で通信可能であることが重要

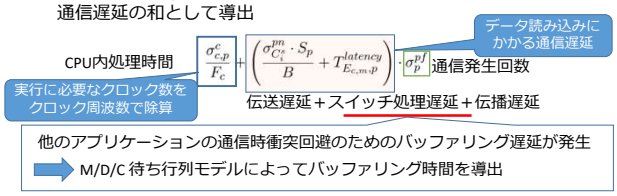


3

3

資源間の通信遅延

- 実行資源の通信頻度、経路設定やネットワーク環境に依存
  - 経路設定：経路のホップ数やトラヒック量
  - ネットワーク環境：伝搬遅延や伝送方式、スイッチの性能
- 複数資源間の通信が相互に影響を及ぼす
  - 通信量の増大による遅延の増加
- ネットワークがタスクに及ぼす影響をモデル化[1]
  - CPU 内処理時間とメモリからのデータ読み込みにかかる通信遅延の和として導出



他のアプリケーションの通信時衝突回避のためのバッファリング遅延が発生
M/D/C 待ち行列モデルによってバッファリング時間を導出

[1] Disaggregated Micro Data Center: Resource Allocation Considering Impact of Network on Performance

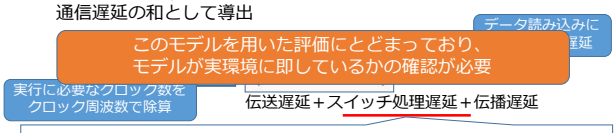
4

4

資源間の通信遅延

- 実行資源の通信頻度、経路設定やネットワーク環境に依存
  - 経路設定：経路のホップ数やトラヒック量
  - ネットワーク環境：伝搬遅延や伝送方式、スイッチの性能
- 複数資源間の通信が相互に影響を及ぼす
  - 通信量の増大による遅延の増加
- ネットワークがタスクに及ぼす影響をモデル化[1]
  - CPU 内処理時間とメモリからのデータ読み込みにかかる通信遅延の和として導出

このモデルを用いた評価にとどまっており、
モデルが実環境に即しているかの確認が必要



他のアプリケーションの通信時衝突回避のためのバッファリング遅延が発生
M/D/C 待ち行列モデルによってバッファリング時間を導出

[1] Disaggregated Micro Data Center: Resource Allocation Considering Impact of Network on Performance

5

5

μDDC に求められるネットワーク

- 資源間の通信遅延がサービスの性能に影響
  - 性能要件を満たすことができない可能性がある
- ある資源間の通信が他資源間の通信や資源割り当てを制限
- ネットワークが性能に及ぼす影響を考慮した μDDCに向けたネットワークの構成が必要
  - 資源間の通信が他の資源間の通信遅延を大きく増大させない
  - 各資源間は性能要件を満たすのに十分なホップ数で通信可能

- 最適なネットワークの構築のためには資源間の通信遅延について正確な把握が必要

6

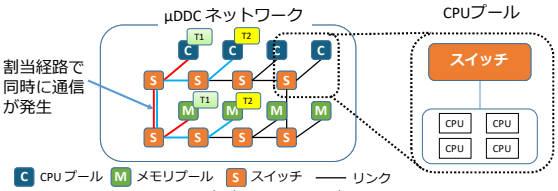
6

資源間通信シミュレータの検討

- 研究の目的
  - μDDC に向けたネットワーク構成の評価のための、資源間通信シミュレータの設計
- アプローチ
  - CPU・メモリ・スイッチの各動作をコンテナとして管理
    - 各機器の動作を模倣
  - 各機器にキューを配置し、待ち時間として資源間の通信遅延を模擬
  - 資源間の通信遅延と資源割当の影響について動作を確認

想定する μDDC

- CPU とメモリの分離が行われた μDDC を想定
  - CPU とメモリは性能低下へ大きく影響 [2]
- 同種の資源は資源プールとして集約
- 資源プールがスイッチ間ネットワークによって接続
- タスク実行時、CPUとメモリとその間の経路が割当
  - 複数のタスクが同時的に実行し、資源間で通信が発生



想定する μDDC の概要

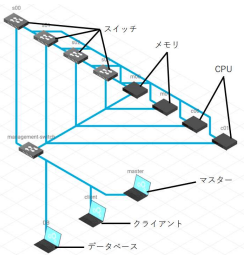
[2] P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker, "Network requirements for resource disaggregation," in *Proceedings of 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, (Savannah, GA), pp. 249–264, USENIX Association, Nov. 2016.

シミュレータの概要

- 各資源とマスター・クライアント・データベースで構成
  - それぞれがコンテナとして動作
  - CPU・メモリ・スイッチがネットワークのノードに相当
- マスターのクロックに合わせて、各構成要素は同期される
- 各資源の動作クロックを個別指定することで、様々な環境へ対応

構成要素	動作内容
CPU	一定量のデータをメモリから取得/書き込み
メモリ	対応する CPU からデータを、受け取り保存/読み出し送信
スイッチ	CPU-メモリ間の通信を保有するキューに基づき中継
マスター	シミュレータ全体のクロックを管理
クライアント	資源間の経路設定や処理内容を決定
データベース	ネットワーク内のノード情報を管理

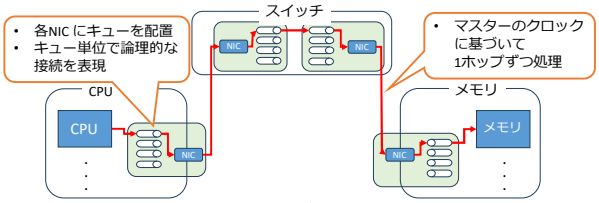
シミュレータの構成要素



シミュレータの操作画面

通信の模擬方法

- マスターのクロックに合わせて、1 ホップずつ処理
- 各ノードに複数のキューを配置し、キュー内の待ち時間をクロックとして表現
  - 複数のキューを用いることで複数資源間の論理的なつながりを再現
  - 待ち時間はマスターのクロックで管理
- CPU からメモリへの読み込み/書き込みにかかる時間を測定
  - 1 クロックあたりの時間に基づき測定



CPU からメモリへのデータ書き込み処理の概要

シミュレーションの操作方法

1. ノード間の接続情報に基づき物理トポロジを構築
2. マスターへ各ノードの種別、アドレス、利用ポートを送信
3. マスターからデータベースへノード情報を集約
4. クライアントから経路情報、模擬内容をマスターへ送信
   - やり取りするデータ量・メモリへの読み込み/書き込みの有無
   - 通信する資源ペアの指定
5. マスターから対応する各ノードへ処理内容と経由ノードを送信し、資源間の通信をシミュレーション

パラメータ設定

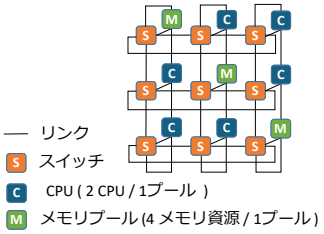
- 以下のパラメータを設定することで、様々なネットワーク環境に対応

パラメータ	内容
スイッチの処理遅延	通信時のパケット転送処理にかかる遅延
キューの数	スイッチが保有するキューの個数
FLOPS	各CPU の演算速度
クロック周波数	各CPU のクロック周波数
キューの数	CPU が保有するキューの個数
資源数	各 CPU プール が保有する CPU 数
キューの数	メモリが保有するキューの個数
メモリ資源数	各メモリプールが保有するメモリ資源数
帯域幅	ノード間を接続するリンクの帯域幅

シミュレーション時に設定できるパラメータ項目

計測環境

- 3 × 3 2Dトーラスネットワークにおいて、動作確認を実施
  - ネットワーク内の CPU とメモリ資源は同数
- 1 つのスイッチに 1 つの資源プールを接続
- 資源同士はスイッチを中継することで通信可能



2D トーラスネットワーク

パラメータ	内容
スイッチ処理遅延	3 µs
キューの数	5
FLOPS	76.8 GFLOPS
クロック周波数	2.9 GHz
キューの数	5
資源数	2
キューの数	5
メモリ資源数	4
帯域幅	10 Gbps

パラメータ設定

13

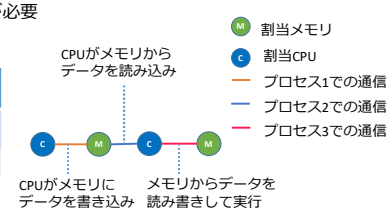
13

割当タスク

- 3 プロセスで実行される 3 種のタスクを想定
  - プロセス 1 : CPU がメモリへ処理命令を送信 (書き込み)
  - プロセス 2 : 実行 CPU がメモリから命令を読み込み (読み込み)
  - プロセス 3 : 命令実行用の CPU とメモリで処理を実行 (読み込み)
- タスクを µDDC の資源の限界まで割り当て
  - 6 タスクを割り当て
- タスクの性能要件として資源間のホップ数を設定
  - ホップ数以内の通信が必要

	制約 1	制約 2	制約 3
ホップ制約	6	4	3
要求数	2	2	2

各タスクのホップ数制約と生成数



割当タスクの概要

14

14

資源割当結果

- 資源割当手法 [1] (RA-CNP)と従来手法 (NP)で比較
  - RA-CNP: ネットワークが実行性能に与える影響を考慮した手法
  - NP: 資源間でできる限り最短経路を割り当てる手法
- 割り当て結果：
  - RA-CNP では、全てのタスクの制約を満たす割当に成功
  - NP において、タスク 6 で性能要件を満たす割当に失敗
    - 最短経路を優先した結果、最短経路で通信可能な経路が枯渇したため

生成タスク (ホップ数制約)	ホップ数 (プロセス1/2/3)	
	RA-CNP	NP
タスク 1 (6)	3/3/3	3/3/3
タスク 2 (6)	3/3/3	3/3/3
タスク 3 (4)	4/4/4	3/3/3
タスク 4 (4)	3/3/3	3/3/3
タスク 5 (3)	3/3/3	3/3/3
タスク 6 (3)	3/3/3	4/4/4

各手法におけるプロセスごとの資源間のホップ数

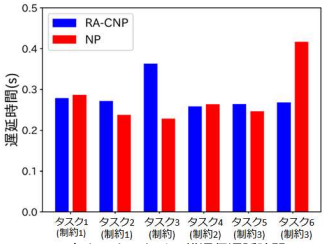
[1] Disaggregated Micro Data Center: Resource Allocation Considering Impact of Network on Performance

15

15

シミュレータによる遅延の計測

- 資源間の通信遅延時間の合計を計測
- ホップ数の増大とともに、遅延時間が増加することを確認
  - 1.5 倍程度増加しており、性能要件を考慮した経路設定が重要
- 全タスクにおいて、ミリ秒単位の通信遅延を計測
  - 通信遅延が大きく、低遅延なネットワークの構築が必要
- 同一ホップにおいて
  最大50ミリ秒の差が発生
  - ネットワーク状況に応じて
    通信遅延は大きく影響
  - 遅延の揺らぎを考慮した
    割当が必要



各タスクにおける総通信遅延時間

16

16

まとめと今後の課題

- まとめ
  - µDDC の資源間の通信遅延シミュレータを設計し、動作を確認した
  - 資源割当を行い、その時の資源間の通信遅延について計測
    - 資源間の通信遅延が大きく、低遅延通信の実現の必要性を確認
    - 遅延時間の揺らぎが大きく、それらを考慮した割当が必要
- 今後の課題
  - パラメータを変更し、様々なネットワーク環境での遅延時間を計測
  - シミュレーションをもとに、
    µDDC に求められるネットワーク要件を調査
  - 要件を満たすネットワーク構成について検討

17

17