jh200025-ISH

広域分散プラットフォーム Distcloud を用いたレジリエンスの定量的評価

柏崎 礼生 (国立情報学研究所)

概要

日本をはじめ、環太平洋地域の島嶼国においては、自然災害による情報インフラストラクチャの破壊が他地域と比較して高い頻度で発生している。センサー端末やモバイル端末から収集された時系列データを用いた防災・減災のための取り組みがこれらの地域では重要視されているが、収集する基盤が遠方にあるクラウドコンピューティング環境上にある場合、その途中にあるインターネット回線は大規模自然災害による影響を受けやすい。

そこで本研究ではまず、本研究提案者らが十年来運用している広域分散プラットフォーム「Distcloud」を拡大し、より多くのユーザが低遅延で Distcloud のサービスに接続できるようにした。次いで、この Distcloud を応用し、ユーザや情報ソースとの遅延が小さいことに優位性のあるアプリケーションに対して SRv6 を用いた経路最適化の寄与を評価した。また、Distcloud 上で展開されるアプリケーションに対し、意図的な障害を発生させることにより耐障害性の検証を行う SDN-FIT サービスを提供し、このサービスの有効性を検証した。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

東京工業大学　京都大学　大阪大学

1.2 共同研究分野

■超大規模情報システム関連研究分野

1.3 参加研究者の役割分担

- 柏崎礼生 (NII) SDN-FIT 評価
- 小谷大祐 (京大) スケーラブル SDN 技術
- 市川昊平 (NAIST) 仮想ネットワーク技術
- 大平健司 (阪大) IPv6 技術
- 北口善明 (東工大) ネットワーク評価技術
- 近堂徹 (広大) インタークラウドマイグレーション

2 研究の目的と意義

本研究の目的は以下の三点からなる。

1. 広域に分散した研究組織が計算機資源を提供し合うことにより構築される広域分散プラットフォーム「Distcloud」を拡大し、より多くのユーザが低遅延で Distcloud のサービスに接続できるようにする。
2. Distcloud を利用した様々な応用、特にユーザや情報ソースとの遅延が小さいことに優位性のあるアプリケーションに対して SRv6 を用いた経路最適化の寄与を評価する。
3. Distcloud 上で展開されるアプリケーションに対し、多様な障害シナリオに基づく意

図的な障害を発生させることにより耐障害性の検証を行う SDN-FIT サービスを提供し、このサービスの有効性を検証する。

日本をはじめ、環太平洋地域の島嶼国においては特に、自然災害による情報インフラストラクチャの破壊が他地域と比較して高い頻度で発生している。センサー端末やモバイル端末から収集された時系列データを用いた防災・減災のための取り組みがこれらの地域では重要視されているが、収集する基盤が遠方にあるクラウドコンピューティング環境上にある場合、その途中にあるインターネット回線は大規模自然災害による影響を受けやすい。

利用者や情報源により近い場所に計算機・ネットワーク資源を配置するエッジ・コンピューティングの応用に期待が集まっており、一方、広域に分散したエッジ・コンピューティング同士が情報共有を行うことで、様々な応用が期待される。このような需要を受けて、広域分散システムは単に構築し正常系における実証実験と評価実験を行うだけでなく、その耐障害性を多様な評価方法で検証し、把握しておくことが求められる。

3　当拠点公募型研究として実施した意義

学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) が提供する北海道大学のインタークラウドシステムは、インターネットやキャンパスネットワークに接続したユーザやセンサーデバイスが提供するデータを蓄積・解析するインターフェイスサービスを提供するのに適している。本研究提案では JHPCN の計算機を、当方が既に構築している広域分散プラットフォーム Distcloud と連携させることにより、より地理的に分散したエッジ・コンピューティング基盤を構築する。この基盤上でエッジ・サービスを提供し、広域分散エッジ・コンピューティング基盤の性能を評価する。

2020 年度は Distcloud を構成する拠点間で SRv6 を用いた相互接続網を構築し、NTT コミュニケーションズとの共同研究開発の実施を予定している。この相互接続網は SINET5 の L2VPN(VPLS) を利用しており、今後の検証実験においても十分な帯域で拠点間が接続されていることが求められる。JHPCN は超広帯域ネットワークにより接続されているため、本研究開発と検証の環境には必須である。

4　今年度の研究成果の詳細

拠点間を接続する回線のサービスレベル同意 (SLA) を用いて複数の拠点からなる広域分散ネットワークのレジリエンスを確率的に定量化する提案を行った [文献 2, 3]。回線の増強により定量化されたレジリエンスがどの程度増大するかを導出することができる。これはすなわち JHPCN で提供される地理的に分散した計算機資源を SINET などの学術網を用いて接続し合い統一されたサービスとして提供する際に、拠点を追加する、回線を追加することによって耐障害性が単に「増す」と述べられるだけでなく「どの程度増すのか」を示すことができる。このことはとりもなおさず、拠点を追加する、回線を追加する費用の効果を定量的に示すことができる。

以下は本研究成果の詳細を解説した成果物の抜粋である。英文のままであることをお許し頂きたい。

The total number of failure patterns depends on the topology that constitutes the wide-area distributed system. It takes a lot of time to perform benchmarks on all the failure patterns and perform quantitative eval-

uations. Unless all failure patterns should be evaluated quantitatively, it is hard to obtain the result of a quantitative evaluation. Meanwhile, various designs are implemented for wide-area distributed services in order to improve fault tolerance, and these designs require some constraints for their proper operation. So we have proposed a pruning method to reduce the total number of failure scenarios.

For example, in a file system, there is a redundant design in which when a chunk is written to a node, a duplicate copy of this chunk is written to other n nodes to increase fault tolerance. In this design, there must be n or more other nodes connectable from a certain node. If the possibility of connection is lost due to the occurrence of failures and the number of other nodes that can be connected from a certain node falls below n, this writing process will fail. There are other measures against split-brain syndrome. In this case, when the total number of nodes is n, when write requests of chunks occur in a certain node, the requests will succeed only when the total number of nodes included in the cluster including the node is larger than $\frac{n}{2}$. Similarly, the requests will fail in a cluster where the total number of nodes is less than $\frac{n}{2}$.

There are no systems that can run under all the situation on the earth. The targeted system has its constraints for its expected environment. The behavior of the system under an arbitrary failure pattern can be classified into the following three by using the constraint.

1. Requests from all nodes are defined.
2. Requests from some (or all) nodes are not defined (therefore may return errors).

By matching the constraint conditions under which the wide-area distributed service operates and the given failure pattern, it is possible to know in advance which class the benchmark request belongs to before performing the benchmark. In the case of 1, the result obtained by the benchmark request may show a quantitative evaluation value of the wide-area distributed service in the failure pattern. In cases 2, the method of handling evaluation values for undefined results must be defined. That is, there can be a method of setting the evaluation value at the time of undefined operation to 0, or a method of excluding the evaluation value from the quantitative evaluation because the evaluation value is undefined because it is an undefined operation. By this exclusion, the time required for benchmarking for quantitative evaluation can be shortened.

Meanwhile, it is possible to quantitatively calculate the fault tolerance under the constraints of the topology and the design of the target system according to the number of failure patterns that can be expected as defined operations and not defined operations. When the identifier of each site is i, the nodes in the topology can be represented as n_i. N, the set of all nodes, can be expressed as follows.

$$N = \{n_1, n_2, ..., n_\nu\} \tag{1}$$

ν means a total number of nodes. In the

same way, the identifier of each interconnection is j (the number of i and j are not related), the edges in the topology can be represented as e_j. E, the set of all nodes, can be expressed as follows.

$$E = \{e_1, e_2, ..., e_\epsilon\} \tag{2}$$

ϵ means a total number of edges. A network failure f_k can be expressed as a subset of E (k is the identifier of each failure). The failures include simultaneous multiple failures. The set of all failures F can be expressed as the summation of single failures (expressed as a set F_1), double failures (F_2), and all ϵ-fold failures (F_ϵ). A number of F_n can be calculated as combinations of ϵ things, taken n at a time. Thus, $num(F)$ that is a total number of F can be expressed as follows
.

$$num(F) = \sum_{i=1}^{\epsilon} num(F_i) \tag{3}$$

$$= {}_\epsilon C_1 + {}_\epsilon C_2 +, ..., {}_\epsilon C_\epsilon \tag{4}$$

A set of all nodes N and a set of all edges E are defined. Then a probability of failure on the edge e_i is defined as p_{e_i}. A set of a probability on all edges P_E can be expressed as follows.

$$P_E = \{p_{e_1}, p_{e_2}, ..., p_{e_\epsilon}\} \tag{5}$$

f is also defined as a subset of E and it can express a failure pattern. F is a set of all failure patterns. An arbitrary f_k can be expressed as follows.

$$f_k = \{e_i, e_j, ..., e_\zeta\} \tag{6}$$

Then, $p(f_k)$, the probability of the failure pattern f_k can be calculated as a product of the probability of f_k multiplied by a product of the "non-failure" probability $(1 - p_{e_i})$ of remaining of f_k. So $p(f_k)$ can be expressed as follows.

$$\begin{aligned} p(f_k) &= \prod P_{f_k} \prod (1 - P_{E-f_k}) \\ &= \prod_{e \in f_k} p_e \prod_{e \in E-f_k} (1 - p_e) \end{aligned}$$

The set F can be separated to a set D that the system can run under a defined condition and U that the system can not run under the condition. P_D and P_U are defined as a summation of the probabilities of each failure pattern in D and U. P_D and P_U can be expressed as follows.

$$P_D = \sum_{f_k \in D} p(f_k) \tag{7}$$

$$P_U = \sum_{f_k \in U} p(f_k) \tag{8}$$

According these equations shown above, the value of resilience on the targeted system R can be expressed as follows.

$$R = log \frac{P_U}{P_D + P_U} \tag{9}$$

Table 1 shows a classification of the number of node groups that can reach each other by an arbitrary edge (cluster) and the number of multiple failure of failure patterns in the five node, full-mesh topology .

All failure patterns in the five-node, full-mesh topology are classified by the multiplicity of failures and the number of nodes (clusters) that can reach each other by any

edge. Is shown in Table 1. In this topology, the maximum multiplicity of failures is 10. All failure patterns with a multiplicity of failures of 3 or less have a cluster number of 1 and all nodes can reach each other. A failure pattern with a multiplicity of failures of 4 or more and a cluster count greater than 1 appears. When the multiplicity of failures is 7 or more, there is no failure pattern with 1 cluster.

In implementations of wide area distributed applications, such as Cloudian Hyperstore, , the result of an object creation request is undefined unless it is possible to connect to three or more locations. Therefore, a failure pattern with two or more clusters is undefined, and quantitative evaluation is performed with a failure pattern with one cluster. Since the total number of failure patterns with 1 cluster is 727 and the total number of failure patterns is 1023, 29% of benchmarks can be omitted compared to benchmarking all failure patterns.

Calculate the expected value of the predefined motion probability, weighted by the failure probability . Here, it is assumed that the probability that a failure occurs at any edge is uniformly p. At this time, the total W_d weighted by P_G for the number of failure patterns that can be expected to be defined is $W_d = 10p + 45p^2 + ..., +225p^5 + 125p^6$. On the other hand, the total W_u weighted by P_G for the number of failure patterns resulting in undefined behavior is $W_u = 5p^4 + 30p^5 + ..., +10p^9 + p^10$. The expected value R of the defined motion probability weighted by the failure probability is, for example, $R = 5.07 \times 10^{-7}\%$ when $p = 0.01$, $R = 5.07 \times 10^{-10}\%$ when $p = 0.001$ prospectively.

この評価実験は 10 ノード、20 エッジのトポロジで計算を行っているが、これを例えば SINET5 の現在のトポロジや SINET6 のトポロジに合わせると、計算量は等比級数的に爆発して現実的な時間では計算できない。この問題については数理的な知見を有する研究者と情報交換を行うことで解決していく予定である。

5 今年度の進捗状況と今後の展望

5.1 Distcloud の拡大

北海道大学のインタークラウドパッケージを利用させていただくことにより、北海道大学、東京大学、大阪大学、九州大学の 4 拠点が Distcloud に追加される形となった。しかしながら SINET L2VPN で接続する方法について事前に調整していた方式では困難であることが判明したため、まずは物理ノード 1 ノードが提供された 2020 年度の 1Q ではインタークラウドパッケージ環境と Distcloud とを、北大の環境に含まれる仮想ルータを境界として接続する方式で実現できる案に落ち着いた。しかしながら次に仮想ルータに接続される VLAN として既存の SINET5 L2VPN のプロジェクトを流用することはできないことが判明し[*1]、NII が提供する ABC クラウドをゲートウェイにすることで Distcloud と相互接続できることが論理的には明らかにできた。実環境として接続実験を行うまでに時間を要してしまい、かつ NII の ABC クラウドが設備移転のために年度内に

[*1] と記述していて、この点については再考の余地があることに気付いたので、本年度はその点を北大と協議して進めていくことを考えている (2021 年度 JHPCN の共同研究提案には漏れてしまったけれども (恨み節))。

multiplicity of failures	number of failure patterns	number of clusters				
		1	2	3	4	5
1	10	10	0	0	0	0
2	45	45	0	0	0	0
3	120	120	0	0	0	0
4	210	205	5	0	0	0
5	252	222	30	0	0	0
6	210	125	85	0	0	0
7	120	0	110	10	0	0
8	45	0	0	45	0	0
9	10	0	0	0	10	0
10	1	0	0	0	0	1

表 1　Classification of failure patterns with number of clusters

サービスを停止してしまったため、2020 年度期間内にこの接続実験を行うことが困難になってしまった。この点については減点せざるを得ない。しかし一方で東北大電気通信研究所で M1 プロセッサ搭載 Mac mini を 5 ノード追加するなど Distcloud の拡大は着実に進捗させることができたため、80% の達成率とする。

5.2　低遅延アプリケーションの評価

Distcloud で提供される計算機資源上で VM を動作させ、この VM を QinQ を用いて論理的に分離された SRv6 用の独自ネットワークで相互接続する環境を構築した。また、NTT コミュニケーションズが提供する NFV の実装である Kamuee を用いた SRv6 の検証実験に関する共同実験契約を東工大、NII、京都大、NAIST、大阪大、広島大、および NTT コミュニケーションズの間で結んだ[*2]。これにより Distcloud 内で利用可能なトラフィックエンジニアリング手法を、実装の改善も含めて評価することができる。NAIST ではこれを用いた SR ヘッダを用いた優先トラフィックの制御手法に関するデモンストレーションを行い、2020 年 3 月に開催された RICC-PIoT workshop で発表が行われた。また SINET SIM とも呼ばれる WADCI(広域データ収集基盤) を用いてシングルボードコンピュータやマイコンから直接 Distcloud に接続する検証実験とデモンストレーションも行った。これは非電化地域への IP ネットワーク延伸に関するフィジビリティスタディで利用され、電力が潤沢に提供されない非電化地域からのトラフィック要求量の大きな通信において優位性を示すことが期待できる。これらの成果は 2021 年度に公開される予定である。本件は 90% の達成率とする。

5.3　SDN-FIT サービスを用いた耐障害性の検証

Distcloud 上に配備された NFV 実装である VyOS を配備し、CLI-API を用いた制御により意図的な障害を発生する SDN-FIT サービスを提供している。また耐障害性の確率論的な

[*2] 契約手続きの締結は 2021 年度に入ってから。

評価指標を提案し、複数のトポロジにおける耐障害性や、回線を増やすことによる耐障害性の向上の定量的な評価について国際学会で 2 件発表している。当初は Cloudian 社が提供する HyperStore のような広域分散で配備可能なアプリケーションをインストールし、耐障害性実験を行うことを想定したが、この検証実験は 2021 年度に実施することを予定している。この点のみを減点とし、本件は 90% の達成率とする。

6 研究業績一覧（発表予定も含む）

学術論文 （査読あり）

なし。

国際会議プロシーディングス （査読あり）

- Hiroki Kashiwazaki, Takuro Ozaki, Hajime Shimada, Yusuke Komiya, Eisaku Sakane, Kazuhiro Mishima, Shiu Sakashita, Nariyoshi Yamai, Yoshiaki Kitaguchi, and Kensuke Miyashita: "Japanese Activities to bring online academic meetings against COVID-19: How We Learned to Stop Worrying and Love the Online Meetings," In ACM SIGUCCS Annual Conference (SIGUCCS '21). Association for Computing Machinery, New York, NY, USA, 2021, pp. 54–59. DOI:https://doi.org/10.1145/3419944.3441174
- H. Kashiwazaki, H. Takakura and S. Shimojo, "A Proposal of Stochastic Quantitative Resilience Index Based on SLAs for Communication Lines," 2021 International Conference on Information Networking (ICOIN), 2021, pp. 143-148, doi: 10.1109/ICOIN50884.2021.9333893.
- H. Kashiwazaki, H. Takakura and S. Shimojo, "A Quantitative Evaluation of a Wide-Area Distributed System with SDN-FIT," 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2020, pp. 607-612, doi: 10.1109/COMPSAC48688.2020.0-189.

国際会議発表 （査読なし）

国内会議発表 （査読なし）

- 杉浦智基, 高橋慧智, 市川昊平: SRv6 を用いたアプリケーションの特性を考慮した通信経路決定手法の提案, RICC-PIoT workshop 2021 (2021/3/5-6, 沖縄産業支援センター (沖縄県那覇市))
- 高名典雅, 柏崎礼生: 可搬性と機密性を両立した占有型情報環境の試作, RICC-PIoT workshop 2021 (2021/3/5-6, 沖縄産業支援センター (沖縄県那覇市))
- 北口善明: デュアルスタックスピードテストサイトを用いたインターネット環境評価, RICC-PIoT workshop 2021 (2021/3/5-6, 沖縄産業支援センター (沖縄県那覇市))
- 石原知洋, 北口善明, 阿部博: SINDAN システムを利用したコンテナプラットフォームにおけるネットワーク環境の計測, 情報処理学会研究報告, Vol.2021-IOT-52, No.33, pp.1-6, March 2021.
- 北口善明, IPv4/IPv6 同時計測可能なインターネットスピードテストサイトによるインターネット環境評価, 第 16 回地域間インタークラウドワークショップ (2020/9/11, 北海道大学情報基盤センター (北海道札幌市))
- 柏崎礼生: 持続可能な「スマート・ルー

ラル」を実現する自律無線メッシュネットワーク網の設計と実装, 第 16 回地域間インタークラウドワークショップ (2020/9/11, 北海道大学情報基盤センター (北海道札幌市))

- ハンズオンセミナー「書を捨てよ外に出よう (Raspberry Pi + SINET5 WADCI + Distcloud を持って)」: 第 16 回地域間インタークラウドワークショップ (2020/9/11, 北海道大学情報基盤センター (北海道札幌市))

公開したライブラリ等

- https://github.com/ITRC-RICC/

その他（特許，プレス発表，著書等）

なし。