様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　４　日現在

機関番号：　１２６０８
研究種目：　若手研究(A)
研究期間：　2013～2017
課題番号：　２５７００００８

研究課題名（和文）100億ノードからなる自律分散システムのシミュレーション手法

研究課題名（英文）Techniques for simulating an autonomous system with over 10 billion nodes

研究代表者
首藤　一幸（SHUDO, KAZUYUKI）

東京工業大学・情報理工学院・准教授

研究者番号：９０３０８２７１

交付決定額（研究期間全体）：（直接経費）　　16,300,000円

研究成果の概要（和文）：インターネット上では1,000万台から成る分散システムが稼働している。インターネット接続機器（IoT）の数は数百億に達しようとしている。にも関わらず、我々が持つ手段で実験可能な規模は数百万ノードにとどまる。そこで我々は、100億ノード規模を扱うことのできるシミュレーション手法を研究した。
分散データ処理システム、例えばApache HadoopやSparkを用いて、イベント駆動で時刻を正確に取り扱いつつ大規模なシミュレーションを行う手法を考案した。その手法を実装したソフトウェアを用いて、一般的なPC 10台で1億ノードという規模のシミュレーションを達成した。

研究成果の概要（英文）：There are large-scale distributed systems with 10 million computers working on Internet. And, the number of computing devices connected to the Internet is going to reach 10 billions. Nevertheless, known techniques can simulate only millions of nodes. Our goal was inventing techniques to simulate 10 billion nodes.
The results include an event-driven simulation technique running on a distributed data processing system such as Apache Hadoop and Spark. The technique handles timing precisely thanks to its event-driven nature. A software implementing the technique could simulate 100 millions of nodes with a commodity PC.

研究分野：　分散システム

キーワード：　シミュレーション　大規模分散システム

１．研究開始当初の背景

インターネット上には 100 万台以上の PC からなる自律分散システムが少なくとも 3 つ稼働しており、最大のものは 1,000 万台に達している。インターネットに接続される機器は多様化しつつ増大を続けており、多くの予想が、その、Internet of Things（IoT）の規模が 2020 年には数百億に達すると見ている。これを想定すると、我々は今後、少なくとも数百億という規模を研究対象としなければならない。しかし、現在手に入る実験方法・手段では、10 万～100 万ノードという規模が限界である（表１）。

実験プラットフォーム	ノード数	
本研究のターゲット	100 億	
Overlay Weaver	約 100 万	1 台上で通信遅延とパケットロスを模擬
PerfectSim.KOM	10 万以上	推奨は 1 万程度
OverSim	10 万	1 台上で通信遅延を模擬
p2psim	1 万	同上
Peeremu	1120	PC 14 台で遅延とパケットロスを模擬

表１　各実験プラットフォームで可能な実験の規模

つまり、我々は今日、現れつつある規模（億～）はおろか、すでに稼働している規模（～1,000 万）すら実験できないという状況を迎えている。例えば、以下を調べる手段を持たないということである。

- 負荷：インターネット（ルータ、回線、組織間接続、…）にどの程度の負荷をかけるのか？
- 安定性：何かを契機に通信量が爆発的に増えてインターネット自体を阻害することはないのか？

ネットワークや分散システムは、モデルの上ではスケールフリーであることも多いが、実地はそうではない。1 万ノードのシミュレーションで健全に動作したからといって、インターネット上の 1,000 万台で動作するかはわからない。

インターネットは、産業、商取引、教育など社会のあらゆる面を支える基盤となった。にもかかわらず、これは危険な状況である。

２．研究の目的

実地で稼働している規模（1,000 万 = 10^7）の実験を可能とするシミュレーション手法を確立し、ソフトウェアを用意することが火急の課題である。続いて、現れつつある規模（100 億 = 10^10）を目指すべきであることは言うまでもない。

３．研究の方法

汎用の分散データ処理システムを基盤として用いて、従来より数桁大きい規模のシミュレーションを可能にする。

分散データ処理システム、具体的には Apache Hadoop や Spark などを用いることで、これまでのアプローチでは得難かった「スケーラビリティ」「耐故障性」「ソフトウェアとしての成熟」を得られる。その一方、「ノードおよびノード間通信の表現」や「シミュレーション時間の取り扱い」といった別の困難がある。

４．研究成果

研究成果は、次の 3 項目に渡る：

① 分散システムのモデル化
② イベント駆動シミュレーションの手法
③ シミュレータ

以下、順に述べる。

① 分散システムのモデル化

シミュレーションのためには、分散データ処理システムで取り扱える形で分散システムを表現する必要がある。分散データ処理システムがサポートする代表的な処理モデル MapReduce で扱うことのできる形で、分散システムをモデル化した（図１）。

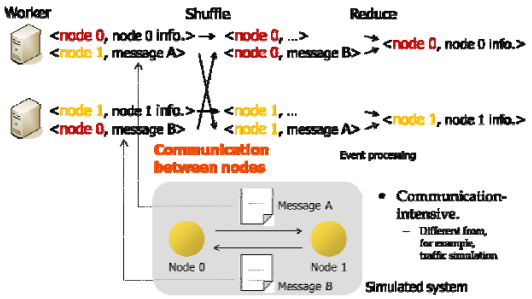

図１　MapReduce 上での分散システムのモデル化

peer-to-peer 分散システムをモデル化・実装・実験したことに加え、ノードの位置も重要となる無線アドホックネットワークをモデル化した（図２）。

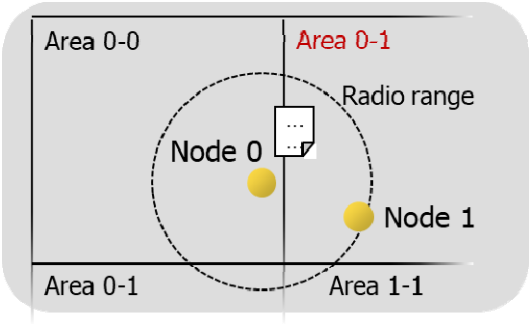

図２　MapReduce 上での無線アドホックネットワークのモデル化

② イベント駆動シミュレーションの手法

モデルの上では、分散システムは、MapReduce の shuffle フェーズで通信を行う。しかし shuffle フェーズは全計算機が同期して通信を行うため、そのままでは任意タイミングでの通信を表現できない（図３）。

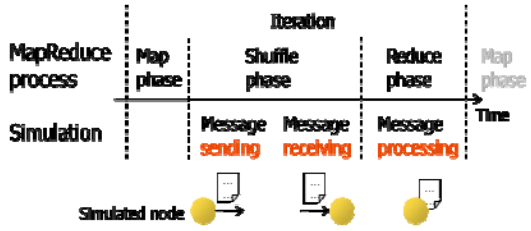

図３　MapReduce shuffle フェーズでの通信を表現

そこで、通常のシミュレータが採用する悲観的な並列シミュレーション手法ではなく、楽観的な並列シミュレーション手法を採用した。これは、Time Warp [Jefferson 1985] に基づく。これにより、任意タイミングでの通信を表現できるようになった（図４）。

楽観的な並列シミュレーションでは、巻き戻しのためにイベントログを保存する必要があり、それがメモリや二次記憶装置を圧迫するというという課題がある。そこで、ログの量を抑制する手法 Moving Time Window (MTW) と Adaptive Time Warp (ATW) を適用した。狙い通りに抑制できた。

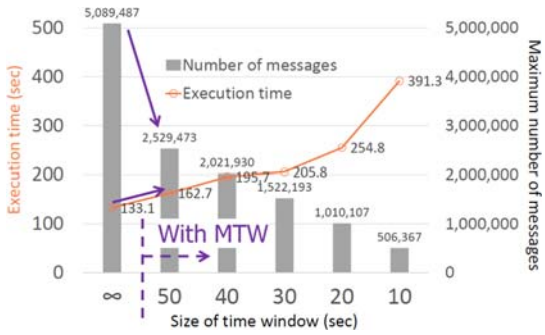

図４　楽観的シミュレーションの所要時間

③ シミュレータ

上記の提案手法を、実際に Apache Hadoop と Apache Spark 上に実装し、評価した。

スケーラビリティという点では、一般的な PC 10 台で 1 億ノードのシミュレートに成功した（図５）。スケーラビリティを制約する要因は見つかっておらず、1,000 台の PC を用いることで 100 億ノードをシミュレートできる見込みである。この際にシミュレートした分散システムは flooding を行う Gnutella であり、分散システムの中では通信が非常に多く発生してシミュレータに多大な負荷をかけるものであった。

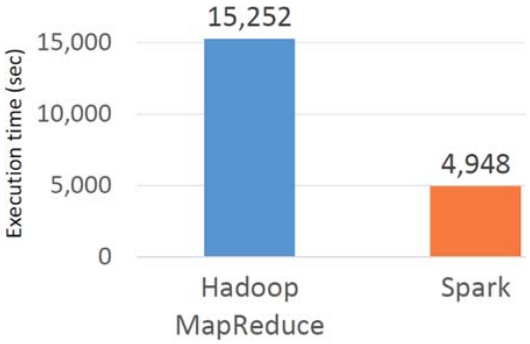

図５　10 億ノードから成る分散ハッシュテーブル（Chord）のシミュレーション所要時間

性能という点では、分散シミュレータは 1 台上のシミュレータに比べ、大幅に低下することが通常である。計算機間での通信がその原因である。例えば、PeerSim と比較して、その分散版である dPeerSim は 100 倍に遅くなっていた。それに対し、我々のシミュレータは 4 倍にしか遅くならなかった。つまり、dPeerSim の約 20 倍の性能を達成した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１１件）

① 首藤一幸、加藤裕也、杉野好宏、華井雅俊、データ処理エンジン上での分散システムシミュレーション、第 9 回 広域センサネットワークとオーバレイネットワークに関するワークショップ、2017 年 6 月 10 日～11 日

② 首藤一幸、データ処理エンジン上での分散システムシミュレーション、ビッグデータ基盤研究会、2016 年 12 月 22 日

③ Kazuyuki Shudo、Yuya Kato、Takahiro Sugino、Masatoshi Hanai、Parallel Discrete-Event Simulation on Data Processing Engines、Proc. 20th IEEE/ACM Int'l Symposium on Distributed Simulation and Real Time Applications (IEEE/ACM DS-RT 2016)、pp. 69-76、2016 年 9 月 21～23 日

④ 加藤裕也、杉野好宏、華井雅俊、首藤一幸、汎用分散処理システムでの大規模分散システムシミュレーション、第8回 広域センサネットワークとオーバレイネットワークに関するワークショップ、2016年 3月 22日～23日

⑤ 加藤裕也、杉野好宏、華井雅俊、首藤一幸、汎用分散処理システムでの大規模分散システムシミュレーション、第8回データ工学と情報マネジメントに関するフォーラム (DEIM2016)、2016 年 2 月 29日～3月 2日

⑥ 加藤裕也、杉野好宏、華井雅俊、首藤一幸、汎用分散処理システムでの大規模分散システムシミュレーションの予備評価、電子情報通信学会 技術研究報告、Vol.115、No.404、NS2015-162、pp.93-96、2016年 1月 21日～22日

⑦ Masatoshi Hanai、Toyotaro Suzumura、Anthony Ventresque、Kazuyuki Shudo、An Adaptive VM Provisioning Method for Large-Scale Agent-based Traffic Simulations on the Cloud、Proc. 6th IEEE Int'l Conf. on Cloud Computing Technology and Science (IEEE CloudCom 2014)、pp.130-137、2014年 12月 15～18日

⑧ 加藤裕也、首藤一幸、Apache Giraph におけるソーシャルネットワーク上の情報伝達モデルの実装、第7回 広域センサネットワークとオーバレイネットワークに関するワークショップ、2014年 11月 29日～30日

⑨ 三津山修平、華井雅俊、首藤一幸、噂の伝播モデルを用いたツイート拡散のシミュレーション、第7回 広域センサネットワークとオーバレイネットワークに関するワークショップ、2014 年 11 月 29 日～30日

⑩ Masatoshi Hanai、Kazuyuki Shudo、Optimistic Parallel Simulation of Very Large-Scale Peer-to-Peer Systems、Proc. 18th IEEE/ACM Int'l Symposium on Distributed Simulation and Real Time Applications (IEEE/ACM DS-RT 2014)、pp.35-42、2014年 10月 1～3日

⑪ Masatoshi Hanai、Anthony Ventresque、Kazuyuki Shudo、Toyotaro Suzumura、Towards a Framework for Adaptive Resource Provisioning in Large-Scale Distributed Agent-based Simulation、Proc. 2nd Workshop on Parallel and Distributed Agent-Based Simulations (PADABS 2014) (in conj. with Euro-Par 2014)、2014年 8月 25日

６．研究組織

(1)研究代表者

首藤　一幸 (SHUDO KAZUYUKI)

東京工業大学・情報理工学院

准教授

研究者番号：90308271

(2)研究分担者

なし

(3)連携研究者

なし