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

令和　２ 年　 ６ 月　 ８ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００１０１

研究課題名（和文）省電力運用に向けたSDN相互結合網でのクラスタ資源管理技法に関する研究

研究課題名（英文）Research for Resource Management Techniques Toward Power-Saving Operation on Cluster System with SDN Available Interconnect

研究代表者
渡場　康弘（Watashiba, Yasuhiro）

大阪大学・サイバーメディアセンター・特任講師（常勤）

研究者番号：６０７５８２７５

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、大規模クラスタシステムにおいて、個々の計算機だけでなくそれらの相互結合網も考慮した省電力運用のための資源管理技法の実現を目指す。相互結合網は、SDN (Software Defined Networking) を利用することでネットワーク資源として管理可能である。本課題では、計算資源およびネットワーク資源を考慮した省電力運用のための資源管理技法の開発･検証を可能とするため、ローカル環境および広域環境における資源管理システムおよびそのシミュレータの研究開発を行った。本成果により、省電力運用下における新たな資源管理技法の研究開発を促進することが可能となる。

研究成果の学術的意義や社会的意義
本課題で研究開発を行った資源管理システムでは、両資源を扱った新たな資源管理技法を配備することができるだけでなく、ジョブキューの構成や扱うアプリケーションといった実際の運用に即した機能についても着目して研究開発を行った。これにより、本システムを用いた省電力化技法の研究開発は、実践的プロトタイプとして実装することで実際のクラスタシステムにおける実用性の評価までを可能とする。このことは、実際のクラスタシステムの運用における本研究成果の利用を促進し、そこからのフィードバックを新たな手法の研究開発に取り入れることが可能となる。

研究成果の概要（英文）：This research aims to realize new resource management techniques in large-scale cluster system under power-saving operation by taking not only computing nodes but also interconnect into account. SDN (Software Defined Networking) enables to manage and control interconnect as network resources. In order to support the development and verification of efficient resource management techniques under power-saving operation by handling both computing resources and network resources, this research has been studied and developed new resource management system and its simulator in local-area and wide-area environment. The outcomes of this research enable to facilitate the research and development of new resource management techniques for large-scale cluster system under power-saving operation.

研究分野：情報学

キーワード：資源管理システム　大規模クラスタシステム　SDN　省電力運用

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

今日の多くの高性能計算環境は、多数の計算ノードを、冗長経路を有したトポロジで構成された相互結合網（インターコネクト）で接続したクラスタシステムとして構成されており、システムの性能向上のために計算ノードの数は増加して大規模化していく傾向にある。それゆえ、このような大規模クラスタシステムでは消費電力が増加していくため、計算資源提供サービスの品質を維持しつつ消費電力を削減することは運用コストの観点で非常に重要な問題となっている。また、クラスタシステムの大規模化に伴い、相互結合網も同様に大規模化・複雑化しており、相互結合網がクラスタシステムの消費電力に占める割合は増加の傾向にある。一方、このような大規模クラスタシステムにおけるジョブの多くは、より高い実効性能を得るために分散並列計算として実行される。分散並列計算ではプロセス間の通信性能が実行性能に影響を与えるため、ジョブの通信特性にあわせて適切な資源を割り当てることが資源提供サービスとして重要である。

上述の状況から高性能計算環境における消費電力の増加は重要な問題として着目されており、低消費電力なハードウェアや電力管理を考慮したソフトウェアなどさまざまな研究が進められている。また、高性能計算環境でジョブへの資源割当を制御するジョブ管理システム（Job Management System, JMS）を利用した研究も行われており、その一つとしてジョブへの計算ノードと電力の割当を最適化する研究[1]があげられる。このような資源割当に着目した省電力化手法はさまざまな研究が行われているが、現在利用されているジョブ管理システムの多くは計算資源のみを対象として管理・制御を行うため、これまでの資源割当の観点で行われた研究は主に CPU を対象とした電力管理が対象となる。しかし、クラスタシステムにおける相互結合網の消費電力も増加していることから、計算資源だけでなく相互結合網をネットワーク資源として考慮した省電力化技法が必要である。

一般的な JMS の多くは、計算ノードにおける CPU やメモリなどの計算資源だけを扱い、他の資源は考慮しない。特に、分散並列計算の実行性能に影響を与える相互結合網をネットワーク資源として管理・制御するための機能を備えていない。そこで、申請者はこれまでの研究で、計算資源と同様にインターコネクトをネットワーク資源として管理・割当が可能な JMS を実現するため、SDN（Software Defined Networking）のネットワークプログラミング性を利用したネットワーク資源制御機構（Network Management Module）を実現し、従来の JMS と連携させた SDN-enhanced JMS Framework を研究開発してきた[2]。本システムにより、計算資源と同様にネットワーク資源も明示的にジョブに割り当てることが可能となった。しかし、従来の JMS に備えられた割当資源決定アルゴリズムではネットワーク資源は扱えないため、新たなクラスタ資源管理技法の整備が両資源を考慮した効率的なクラスタ資源の運用のために必要不可欠であると考える。そこで、SDN-enhanced JMS Framework のネットワーク資源管理機能を利用することで、両資源を考慮した効率的な省電力化技法の研究開発基盤が構築できると考え本研究の着想に至った。

２．研究の目的

大規模クラスタシステムにおける省電力運用方法の 1 つとして、管理者によるクラスタ資源の利用状況に応じて資源の一部を停止した環境で資源提供サービスを行う縮退運転があげられる。今日の大規模クラスタシステムでは、計算ノード間の相互結合網における消費電力が全体に占める割合は増加しているため、相互結合網についても考慮する必要がある。しかし、一般的な資源管理システムでは相互結合網をネットワーク資源として管理していないため、クラスタシステム全体での縮小運用およびその際のジョブの要求資源量の保証を行うための仕組みは十分に整備されていない。本研究では、SDN (Software Defined Networking) を利用した相互結合網におけるネットワーク資源管理機能を活用して、相互結合網も考慮した省電力運用に向けたクラスタ資源管理技法の研究開発基盤の実現を目指す。

３．研究の方法

研究開発を行っている SDN-enhanced JMS Framework の特徴の一つとして、任意の資源割当アルゴリズムを資源割当ポリシとして配備可能な機能があげられる。この仕組みを活用し、縮退運転による資源量の減少を考慮した実践的な省電力下における資源割当技法の開発・検証を可能とする基盤を構築する。また、さまざまなシステム構成での評価の観点からジョブスケジューリングシミュレータ、およびネットワーク資源管理の効果が大きく出るとの観点から広域分散型環境に対応可能とするための SDN-enhanced JMS Framework の拡張も行う。

４．研究成果

(1)

クラスタシステムの実運用に適用されているジョブ管理システムでは、一般的にユーザから投入されたジョブを格納するためのジョブキューが複数設定される。各ジョブキューには資源量等の制約に基づき分類されており、ジョブはユーザの要求に応じて適切なジョブキューに振り分けられる。また、ジョブキューはスループット向上や負荷分散等の観点から利用可能な計算ノードへの割当が決められており、この設定をマッピングと呼ぶ。ジョブキューの構成および各ジョブキューのマッピング設定は、どの計算資源およびネットワーク資源をジョブに割り当て

るのかを決定する一つの要因である。

一方、ジョブ管理システムの振る舞いを検証するためのジョブスケジューリングシミュレータの多くでは、マッピング設定を行うための機能を有していない。その理由としては、ジョブスケジューリングシミュレータの目的が資源割当アルゴリズムの検証であるため、ジョブキューによるジョブの分類は不要である点が考えられる。しかし、実運用における縮退運転を考慮した資源管理技法の設計には、資源割当手法だけでなくジョブキューによるジョブの分類および利用可能な計算資源の制御も必要な機能である。そこで、本目的に対応可能なジョブスケジューリングシミュレータを構築するため、既存のジョブスケジューリングシミュレータの 1 つである ALEA[3]を基盤技術とし、マッピング機能を拡張したジョブスケジューリングシミュレータの構築を行った。その構成を図 1 に示す。

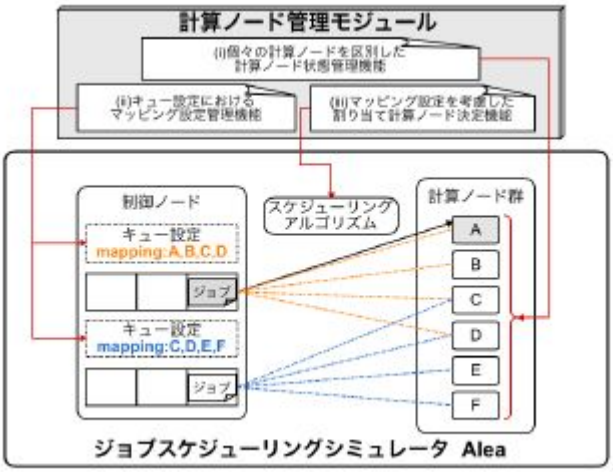
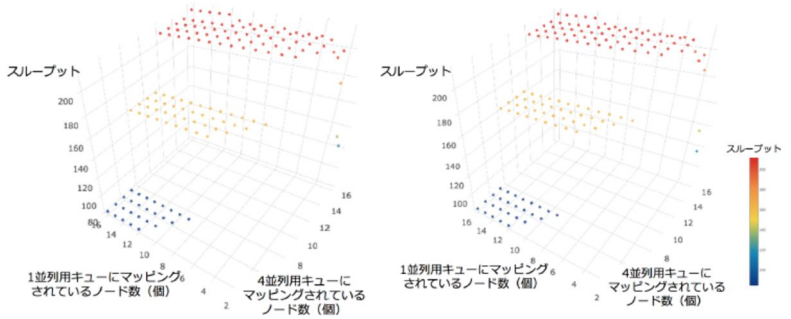


図 1：複数のジョブキューに対応したジョブスケジューリングシミュレータの構成

本ジョブスケジューリングシミュレータでは、計算ノード管理モジュールを設計・開発し、ALEA と連携させることでジョブスケジューリングシミュレータにマッピング設定機能を実現した。計算ノード管理モジュールは 3 つの機能で構成される。計算ノード状態管理機能(i) は、計算ノードの使用状況を個々に管理し、計算機クラスタ管理を行う。マッピング設定管理機能(ii)は、計算ノード状態管理機能が管理している計算ノード単位で設定管理を行う。割り当て計算ノード決定機能(iii) は、計算ノード状態管理機能が管理している各計算ノードの状態とマッピング設定管理機能の情報をもとに、ジョブへの割り当て計算ノードの決定を行う。

キューから取り出されたジョブは、スケジューリングアルゴリズムによって割当計算ノードが決定される。この時、割当計算ノード決定機能が働き、マッピング設定管理機能によりキューに設定されていたマッピング情報を考慮した計算ノード決定が行われる。計算ノード状態管理機能は、割当計算ノード決定機能からジョブの情報を受け取り、ジョブの実行開始・終了に合わせて、対応する計算ノードの使用状況を変更する。

本ジョブスケジューリングシミュレータの挙動の正確性を検証するため、実際のクラスタシステムと同じ構成を本ジョブスケジューリングシミュレータに設定し、同じ構成のジョブセットを投入した際の各マッピング設定におけるジョブスループットを比較した。その結果を図 2 に



(a)シミュレータ　　(b)実クラスタシステム

図 2：シミュレータと実クラスタシステムのジョブスループット比較

示す。図 2 よりどちらの環境においても同様の傾向が見て取れることから、本ジョブスケジューリングシミュレータによるマッピング設定の検証は、実環境での挙動を再現できていることが確認できた。また、計算ノード管理モジュールを追加したことによるシミュレータの実行速度についても、実用上問題ない程度であることを確認した。

本ジョブスケジューリングシミュレータの構築により、省電力のための縮退運転時における新たな資源管理技法の設計において、ジョブキューの構成とそれらのマッピング設定を設計要素として加えることを可能とできた。

(2)

近年、クラウドに代表されるような複数の拠点に分散した高性能計算機環境を 1 つの計算機環境と見做し、その資源からユーザに資源提供を行うサービス形態がある。このような広域分散環境では、ネットワーク資源は各拠点におけるクラスタシステム内の相互結合網だけでなく、拠点間を接続するネットワークも含まれる。加えて、どの拠点の計算資源を割り当てるのか、分散ストレージ上におけるデータの配置なども考慮する必要があり資源管理はより複雑になる。また、省電力運用の観点においても、拠点ごとの電力バジェットの違い等も考慮する必要があり、さまざまなモデルに対する省電力運用における資源管理技法の研究開発が可能となる。

一方、これまで研究開発してきた SDN-enhanced JMS Framework は単一のクラスタシステムにおける従来の HPC 資源提供サービスを想定した設計・実装となっているため、拠点間ネットワークの資源管理やストレージの位置を考慮する仕組みを備えていない。また、仮想マシンやコンテナ技術に代表される計算資源の仮想化技術にも対応しておらず、今後の実運用された高性能計算環境を想定した省電力運用における資源管理技法の開発・検証基盤としては不十分であると考えた。そこで、SDN-enhanced JMS Framework の設計コンセプトに則り、広域分散環境における資源管理技法の開発・検証を可能とする新たな資源管理システムの構築を行った。その資源管理システムの構成を図 3 に示す。なお、本資源管理システムは広域分散環境における耐障害性も考慮した機能についても含まれている。

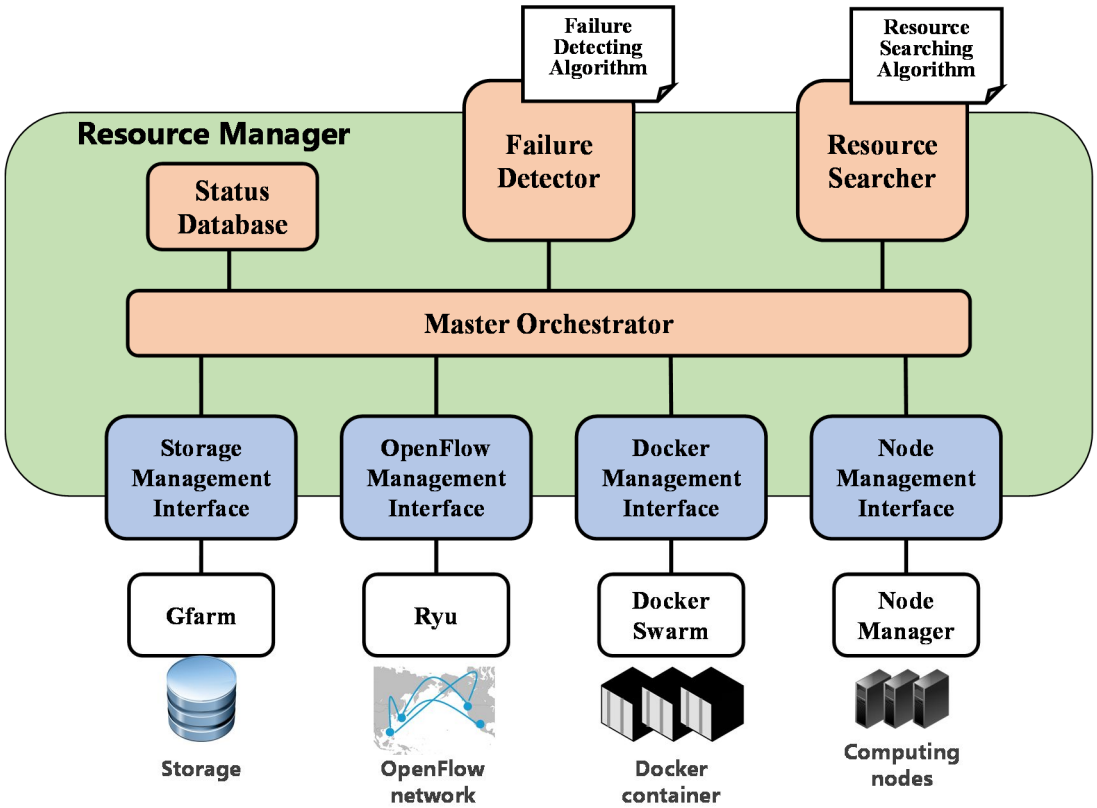


図 3：広域環境に対応した資源管理システムの構成

本資源管理システムでは、さまざまな資源を考慮した省電力運用のための資源管理技法を設計できるようにするため、個々の資源を管理しているマネージャを統括できる設計としている。そのプロトタイプとして、従来の SDN-enhanced JMS Framework と同様の機能である計算資源管理機能、および SDN によるネットワーク資源管理機能を備える。これらに加え、分散ストレージを管理する Gfarm を制御するための機能、および Docker コンテナを管理する Doker Swarm を制御する機能を実装した。これらは資源管理システム上のインターフェイスとして設計し、各資源

のマネージャと連携して情報収集および資源制御を行う。また、インターフェイスを置き換えることで多様な資源管理機構と接続可能となる。各資源のマネージャを制御するための機構として Master Orchestrator と呼ぶモジュールを実装した。Master Orchestrator モジュールは、各資源の情報を収集する機能、およびジョブへの資源の割り当てや切り替えを行うための機能を備える。Failure Detector モジュールおよび Resource Searcher モジュールは、SDN-enhanced JMS Framework における資源割当ポリシ機能に該当する機能であり、前者は電力事情やシステム障害といった割当資源の切り替えの必要性を判定するためのモジュールであり、後者はどのように資源を割り当てるかを決定するためのモジュールである。本資源管理システムの構成により、従来の静的な縮退運転だけでなく、動的な縮退運転による運用を想定した資源管理技法の設計・検証が可能となる。

本資源管理システムの動作検証として、図 4 に示す構成で構築したローカルテストベッドにおいて割当資源の一部で障害が発生した際の資源切り替えにおける各処理の所要時間を測定した。本実験において、Failure Detector モジュールによる検出は Docker Swarm のヘルスチェック機能の情報を利用した手法を、Resource Searcher モジュールにおける資源決定については、計算資源は任意の利用可能資源を選択、ネットワーク資源は最少ホップ数の経路の割当を行うアルゴリズムを適用した。その測定結果を図 4 に示す。本実験により、本資源管理システムは配備したアルゴリズムにより適切に動作することが確認できた。所要時間については各アルゴリズムに依存する部分が大半であり、資源管理システム自体の処理による時間は非常に小さかった。本実験の場合、所要時間の大半は Docker Swarm による Docker コンテナの制御であった。

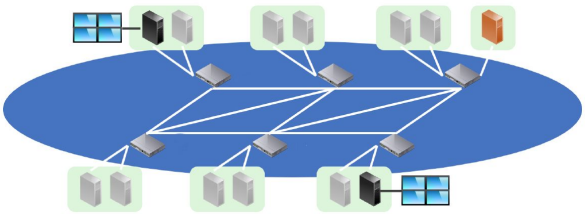


図 4：評価環境の構成

Action	Time (s)
Failure detection	12.20
Searching alternative resources	0.11
Reconfiguring network	0.31
Redeploying Docker container	16.20

図 5：資源管理システムの性能評価

本資源管理システムの実現により、広域分散環境を対象とした多様な省電力モデルのための資源管理技法の開発・検証が可能となる。今後の課題として、各拠点の電力バジェット情報の収集や動的なパワーキャッピング制御のための機能の設計・実装があげられる。これにより、動的な省電力運用における新たな資源管理技法の開発が可能となると考える。

＜引用文献＞

[1] O. Sarood et al., “Maximizing Throughput of Over-provisioned HPC Data Centers under a Strict Power Budget”, SC14, pp.807–818, 2014.
[2] 渡場康弘 ほか, “計算資源とネットワーク資源を考慮した割当ポリシを配備可能とするジョブ管理フレームワーク”, 電子情報通信学会論文誌, vol.J97-D, No.6, pp.1082-1093, 2014.
[3] Luis P. Prieto, Mar´ıa Jes´us Rodr´ıguez-Triana, Marge Kusmin, and Mart Laanpere. Complex Job Scheduling Simulations with Alea 4. CEUR Workshop Proceedings, Vol. 1828, pp. 53–59, 2017.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計13件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 松井祐希，渡場康弘，伊達進，吉川隆士，下條真司
２．発表標題 細粒度マッピング設定に対応したジョブスケジューリングシミュレータの構築
３．学会等名 第143回 システムソフトウェアとオペレーティング・システム研究会
４．発表年 2018年

１．発表者名 Yuki Matsui, Yasuhiro Watashiba, Susumu Date, Takashi Yoshikawa and Shinji Shimojo
２．発表標題 Architecture of Job Scheduling Simulator for Evaluating Mapping Between Queue and Computing Node
３．学会等名 PRAGMA 34 workshop（国際学会）
４．発表年 2018年

１．発表者名 松井祐希,渡場康弘,伊達進,木戸善之,下條真司
２．発表標題 広域連携型災害管理アプリケーション基盤を提供する資源管理システムの検討
３．学会等名 日本ソフトウェア科学会 第16回ディペンダブルシステムワークショップ（DSW2018）
４．発表年 2018年

１．発表者名 Yuki Matsui, Yasuhiro Watashiba, Yoshiyuki Kido, Susumu Date and Shinji Shimojo
２．発表標題 Architecture of Resource Manager for Software-Defined IT Infrastructure
３．学会等名 International Symposium on Grids & Clouds 2019 (ISGC 2019)（国際学会）
４．発表年 2019年

１．発表者名
Yasuhiro Watashiba, Yuki Matsui, Yoshiyuki Kido, Susumu Date, and Shinji Shimojo

２．発表標題
Toward Orchestration on Software-Defined IT Infrastructure for Disaster Management Applications

３．学会等名
PRAGMA 36 workshop（国際学会）

４．発表年
2019年

１．発表者名
Yasuhiro Watashiba, Yoshiyuki Kido, Kazuya Ishida, Susumu Date, Kohei Ichikawa, Jason Haga, Hirotake Abe, Hiroaki Yamanaka, Ryousei Takano, Jason Leigh, Fang-Pang Lin, Jos? Fortes, and Shinji Shimojo

２．発表標題
Toward Resilient Software-Defined IT Infrastructure for Supporting Distributed Disaster Management Applications

３．学会等名
27th Workshop on Sustained Simulation Performance (WSSP27)（国際学会）

４．発表年
2018年

１．発表者名
Yuki Matsui, Yasuhiro Watashiba, Susumu Date, Takashi Yoshikawa and Shinji Shimojo

２．発表標題
Architecture of Job Scheduling Simulator for Evaluating Mapping Between Queue and Computing

３．学会等名
PRAGMA 34 Workshop（国際学会）

４．発表年
2018年

１．発表者名
松井祐希，渡場康弘，伊達進，吉川隆士，下條真司

２．発表標題
細粒度マッピング設定に対応したジョブスケジューリングシミュレータの構築

３．学会等名
第143回 システムソフトウェアとオペレーティング・システム研究会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考