

東北大学工学部 卒業論文

ウェブインタフェースを介した
スーパーコンピュータ利用環境に関する研究

機械知能・航空工学科 滝沢研究室

谷澤悠太

(令和6年3月)

目次

第 1 章	緒論	1
1.1	背景	1
1.2	目的	1
1.3	本論文の構成	2
第 2 章	関連研究	3
2.1	緒言	3
2.2	HPC システム利用方法	3
2.3	Open Ondemand	4
2.4	現行のウェブインタフェースにおける課題	5
2.5	結言	6
第 3 章	ウェブインタフェースを介した HPC システム利用環境	7
3.1	緒言	7
3.2	従来手法	7
3.3	提案手法	7
3.4	実装	8
3.4.1	実装の概要	8
3.4.2	スケジューラ抽象化機能と NQSV の連携	9
3.4.3	ウェブ機能とスケジューラ抽象化機能との連携	11
3.5	結言	11
第 4 章	実装評価	12
4.1	緒言	12
4.2	評価環境	12
4.3	評価条件	12
4.4	実行時オーバヘッドの評価	12
4.5	結言	12
第 5 章	結論	13
	参考文献	14

図目次

1	一般的な HPC システムの模式図	4
2	ダッシュボード画面	5
3	ホームディレクトリ画面	5
4	ジョブ管理画面	6
5	従来手法の模式図	7
6	提案手法の模式図	8
7	実装環境	9

表目次

コード目次

1	ジョブの状態取得メソッド	10
---	------------------------	----

第 1 章 緒論

1.1 背景

近年，高性能計算 (High Performance Computing, HPC) システムの用途は多様化し，専門知識を持たない利用者が容易に HPC システムを利用する需要が高まっている．一般的に，コマンド操作に基づいて HPC システムを操作する利用環境や利用する HPC システムごとに異なる操作方法により，HPC を専門としない研究者は HPC システムを使いこなすために多くの学習時間を費やす必要がある．実際に Ping らによると，学問のために HPC システムを初めて利用する学生などは HPC システムのための利用環境の構築に多くの時間を費やしてしまい，本来の目的である学問のための HPC システムの利用までの多大な時間を費やしてしまうという問題が挙げられている．[1] そこで，従来のコマンド操作に基づく利用環境や，システムごとに異なる利用方法を利用者から隠蔽し，ウェブブラウザを用いて容易かつ統一的に HPC システムを利用することが可能なウェブインタフェースの研究開発が行われている．

しかし，現行のウェブインタフェースはユーザへの簡易かつ統一的な HPC 利用環境とを提供するため，ウェブインタフェースの機能の改修を行う度にウェブインタフェース本体を改修する必要がある．そのためシステムの保守性に問題があるといえる．また，HPC システムの利用者にとって，様々なジョブスケジューラを統一的に扱うということは〇〇や〇〇など，様々な分野において有用な研究であり，関心が高い分野である．

そのため，ウェブインタフェースの機能をユーザがウェブブラウザ上で HPC 利用を可能とする機能と多様なジョブスケジューラを統一的に取り扱う機能の二つの機能に分離して実装することで前述した課題点を解決し，ジョブスケジューラ統一化の要望にも応用できる機能の分離利用が可能なウェブインタフェースを考えることができる．

1.2 目的

本研究では，HPC システムの利用難度の高さや HPC システムにおけるジョブスケジューラの多様化に伴い発生し得る HPC システム利用環境に関する課題点に着目する．インタフェースをウェブ機能とスケジューラ抽象化機能に分離することで操作の簡易化や保守性などの問題を解決することを目的とする．具体的には現行のインタフェースの機能を分離し，実装を行う．実装における動作の確認を行い，提案したインタフェースを定量的に評価することで提案手法の有用性を示す．

1.3 本論文の構成

本論文は全 5 章から構成される。第 1 章では，本研究の背景と目的について述べた。第 2 章では，関連研究について説明する。第 3 章では，ウェブインタフェースを介した HPC システム利用環境について説明し，提案手法の実装を行う。第 4 章では，実装の評価結果を示し，その考察を行う。第 5 章では，本研究の結論と今後の課題を述べる。

第 2 章 関連研究

2.1 緒言

本章では関連研究について述べる．はじめに一般的な HPC システムの利用方法，続いて OpenOnDemand と呼ばれるインタフェース，最後に現行のインタフェースにおける課題を述べる．

2.2 HPC システム利用方法

HPC システムとは，スーパーコンピュータやコンピュータクラスタの能力を利用して，ほかのコンピュータを遥かに凌ぐ速度で計算課題 (ジョブ) を処理し，実行するシステムを指す．このようなコンピューティング能力の集約によって，さまざまな科学分野において他の方法では対処できない大きな課題を解決できる．実際に，平均的なデスクトップコンピュータは毎秒数十億の計算を実行できる．これは，人間が複雑な計算を行うことができるスピードに比べれば，素晴らしい数字である．しかし，HPC システムは，1 秒に数千兆の計算を実行することができるため，大規模な課題に対してはより適しているといえる．

一般的に HPC システムの利用の流れを図 1 に示す．HPC システムは数種類のサーバやデータベースから構成される．ジョブを実行するために計算を行う計算サーバ (ワーカーノード)，ワーカーノードを管理するためのジョブスケジューラが搭載されたジョブ管理サーバ (マスターノード)，ユーザ情報が保存されたデータベースと連携してログイン情報の管理やユーザリクエストの受け渡しを行うログイン用サーバ，実行するジョブのファイルなどが保存されたファイルサーバなどがシステムの構成例である．ユーザはログイン用サーバが取り扱うユーザ情報を用いて HPC システムにログインする．ユーザはログインサーバを通じてマスターノードにジョブの実行を依頼する．マスターノードは依頼されたジョブのファイル情報などをファイルサーバと連携して受け渡しを行い，ワーカーノードにジョブの実行を依頼する

ユーザは利用したいクラスタを遠隔で操作するために自身のコンピュータからユーザ情報を用いて鍵の登録を行った後，ログイン用サーバに ssh 接続する．そして，ユーザは利用するジョブスケジューラの種類に応じた形式でジョブスクリプトを作成する．その後，与えられたコマンドを用いてジョブを投入を行う．マスターノードのジョブスケジューラが実行するジョブの管理を行い，ワーカーノードはマスターノードの命令に従ってジョブの実行を行う．実行が終了したジョブは標準出力とエラーファイルが出力され，ジョブの実行結果を確認することができる．

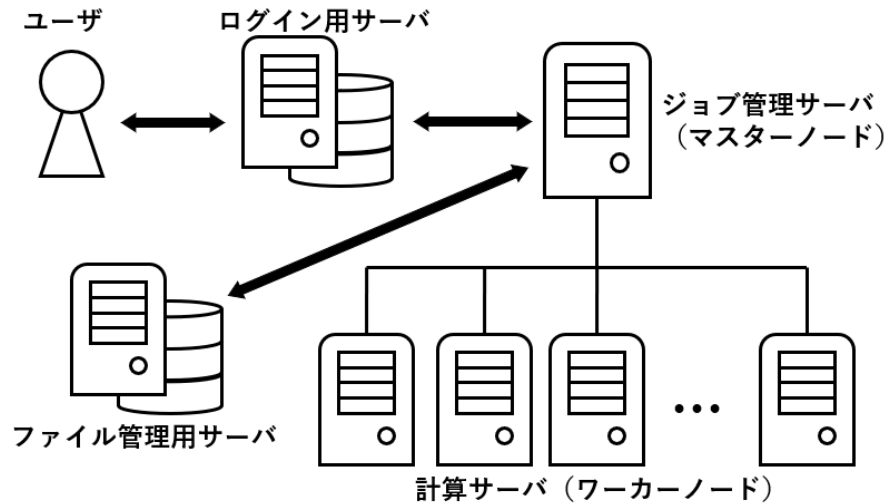


図 1: 一般的な HPC システムの模式図

2.3 Open Ondemand

代表的な関連研究として、Open OnDemand(OOD) とその機能や設計構成を紹介する [2][3]. OOD は米国オハイオ・スーパーコンピューティングセンターが開発したオープンソースソフトウェアであり、ウェブインタフェースを介して HPC システムを利用できる環境を提供する。ユーザはウェブブラウザ上で HPC クラスタを簡単に操作することができ、プラグインやほかのソフトウェアのインストールや設定は不要である。また、リモートデスクトップや jupyternotebook, VSCode などのグラフィカルな対話的操作もウェブブラウザ上から利用することができる。OOD は世界的に使われている様々なジョブスケジューラ (PBS Pro, Slurm, Grid Engine, Torque, LSF など) に対応しているためシステム間の利用方法の差異を隠蔽している。

初めに、OOD の機能について説明する。OOD は、図 2 に示すダッシュボードとユーザのディレクトリを管理する Files アプリ、ジョブの管理を行う jobs アプリ、シェルの操作を行う Clusters アプリ、開発者が任意の追加アプリを導入することができる interactiveapps から構成される。図 3 に示す files からはユーザのディレクトリをグラフィカルに操作ことができ、ファイルやディレクトリの削除追加編集なども容易に行うことができる。図 4 には Jobs のジョブ管理を行う JobComposer の画面を示す。ユーザはジョブの作成と投入削除などをすべてこの画面から行うことができる。また、Jobs の ActiveJobs 画面からは投入したジョブの状態を確認することができ、HPC クラスタの様子を OOD を介して把握することができる。また、Shell アプリからは連携した HPC クラスタのシェルをウェブブラウザ上から操作することができる。このように、OOD は様々なアプリケーションと連携し

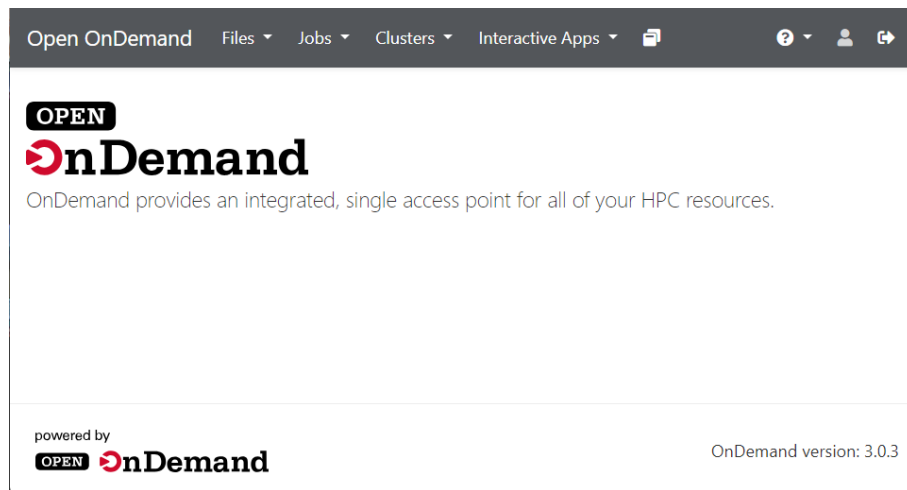


図 2: ダッシュボード画面

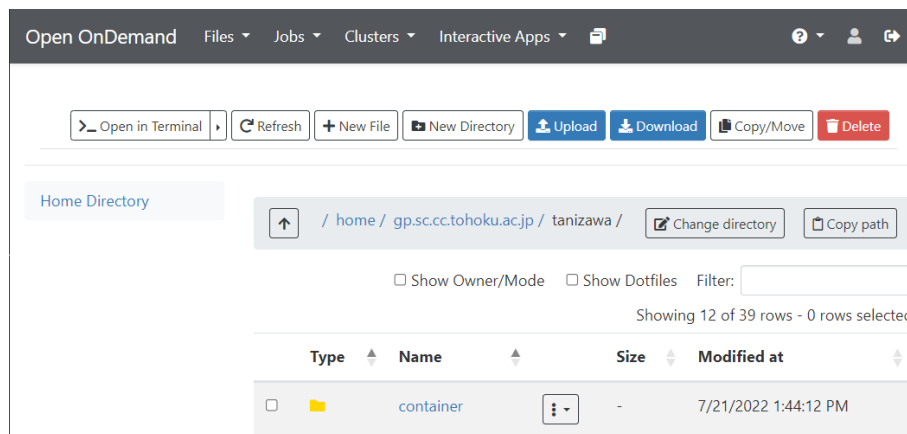


図 3: ホームディレクトリ画面

て HPC ユーザの利用支援を行っている。

続いて、OOD の設計について説明する。OOD は現在多様なジョブスケジューラに対応しているが、各ジョブスケジューラへの対応は adapters ディレクトリ下に配置される。指定したジョブスケジューラに対応するために、各々で Adapter クラスのサブクラスが宣言され、内部ではジョブスケジューラに合わせてジョブの投入を行うメソッドやジョブの削除を行うメソッド、ジョブの情報を取得するメソッドが再定義されている。対応する Adapter ファイルを呼び出して参照することで OOD は多様なジョブスケジューラに対応することができる。

2.4 現行のウェブインタフェースにおける課題

現行のウェブインタフェースにおける課題について説明する。国内でのウェブインタフェースの実装事例として、スーパーコンピュータ富岳での OOD の実装が挙げられる。

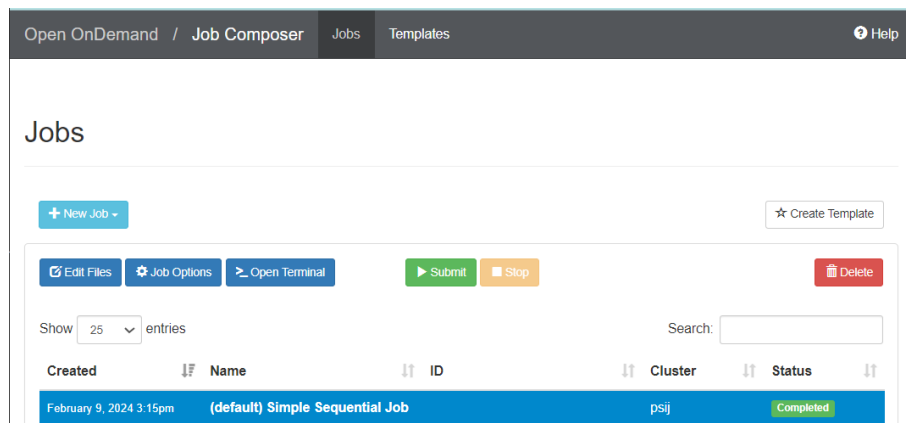


図 4: ジョブ管理画面

OOD は汎用的なツールであるが、富岳で用いられているジョブスケジューラ (Fujitsu Technical Computing Suite, Fujitsu-TCS) に対応していなかったことから、中尾らは OOD を Fujitsu-TCS 向けに改修した事例を報告している [4]. Fujitsu-TCS に対応するために、新たな Adapter ファイルを作成し、Fujitsu-TCS 用のメソッドを再定義することにより OOD は Fujitsu-TCS への対応を行うが、このような改修方法はシステムの基幹部分を直接改修しなければいけないため、慎重に作業を行う必要がある。このように、ほかにも様々なジョブスケジューラが存在し、今後も登場することを考えると、ジョブスケジューラの種類が増えるごとに OOD 本体を直接改修する方法では問題があるといえる。

2.5 結言

本章では、関連研究について述べた。はじめに一般的な HPC システムの利用方法について説明した。その後、Open OnDemand と呼ばれるインタフェースについてその機能と設計について説明した。最後に、現行のインタフェースについて、国内での OOD の実装例を参考にして課題点を述べた。次章では、本章で述べた現行のウェブインタフェースにおける課題である保守性を考慮した手法を提案し、その実装を行う。

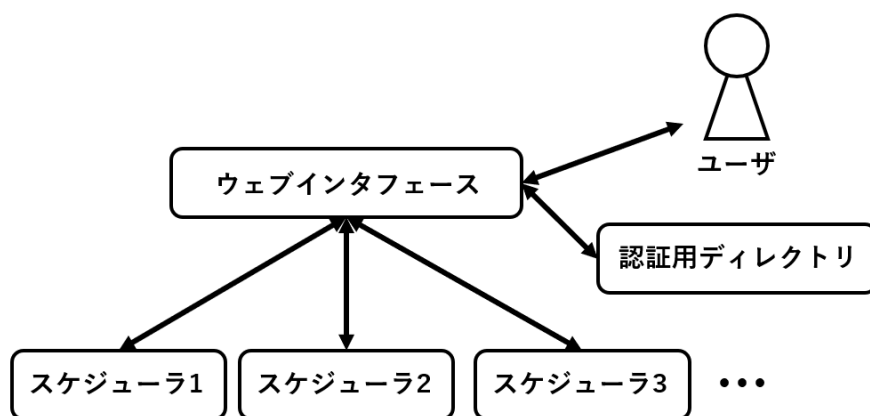


図 5: 従来手法の模式図

第 3 章 ウェブインタフェースを介した HPC システム利用環境

3.1 緒言

前章では，本研究の関連研究について述べた．本章では，ウェブインタフェースを介した HPC システム利用環境の提案手法について説明し，その実装を行う．はじめに，提案手法の概要を説明する．その後，実装の概要，具体的な実装の手順について説明する．

3.2 従来手法

従来のウェブインタフェースを介した HPC システム利用環境について説明する．従来手法の模式図を図 5 に示す．ウェブインタフェースはユーザとスケジューラのを直接結ぶ構成になっている．外部の認証用ディレクトリと連携してユーザ情報を管理することで，ウェブブラウザ上での利用空間の提供を行う．また，主要なスケジューラの抽象化を行うことで，統一的な操作環境を提供する．前章で述べたように，このシステム構成はシステムを改修する際にウェブインタフェース本体を改修する必要があるため，改修中に技術者の予期しない影響を与えてしまう可能性があり，保守性に問題がある．

3.3 提案手法

本研究の目的は，HPC 利用環境をウェブインタフェースに提供する機能 (ウェブ機能) と，ジョブスケジューラ間の差異を抽象化する機能 (スケジューラ抽象化機能) に切り分け，それぞれ独立に保守できる構成を実現することである．このために，本研究ではウェブ機能

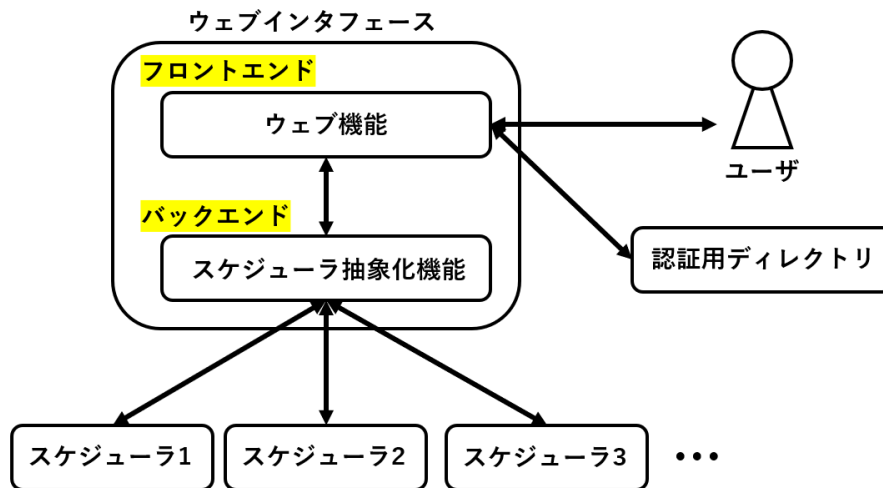


図 6: 提案手法の模式図

からは統一的にシステムを利用し，スケジューラ抽象化機能でシステム間の差異を埋める構成の利用環境を提案する．

この提案手法の模式図を図 6 に示す．フロントエンドでは，ユーザはウェブ機能のみとやり取りを行い，ユーザ情報を管理する外部の認証用ディレクトリを用いて安全に HPC システムを利用することができる．バックエンドでは，ウェブ機能から得られた様々なスケジューラに対するリクエストをスケジューラ抽象化機能が受け取り，処理を行う．この実現のためには，ウェブ機能とスケジューラ抽象化機能を連携させる必要があることから，両者間に求められる情報のやり取りを整理し，適切な連携方法を検討する．

3.4 実装

3.4.1 実装の概要

ウェブ機能とスケジューラ抽象化機能をそれぞれ独立に実装し，連携させることでウェブインタフェースを介して様々なシステムを統一的に利用できる環境を実現する．そのために，ウェブ機能の基盤として OOD，スケジューラ抽象化機能の基盤として PSI/J[5] と呼ばれる Python ライブラリを利用し，両者を組み合わせることで提案手法を実装する．

本研究では，東北大学のスーパーコンピュータ「AOBA」で運用されているジョブスケジューラ (NEC Network Queuing System V, NQSV) が OOD に対応していないという事実に着目して，NQSV をスケジューラ抽象化機能側に実装することと，それをウェブ機能側から利用できることを検証する．実装環境として，OOD 用のホストサーバとスーパーコンピュータ AOBA を模した HPC クラスタ (疑似 AOBA クラスタ) を考える．模式図を図 7 に示す．OOD 用のホストサーバでは，OOD の動作が保証されている Ubuntu20.04 LST を OS として用いる．疑似 AOBA クラスタは，マスターノードと二つのワーカーノードか

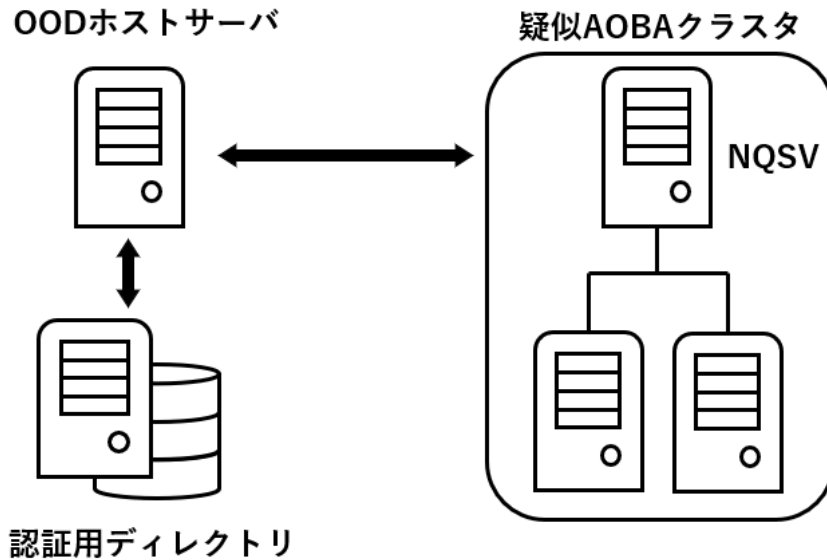


図 7: 実装環境

ら構成される小規模なクラスタであり，AOBA と同様に NQSV がジョブスケジューラとして利用されている．疑似 SOBS クラスタでは QSV の動作確認が行われている CentOS7 を用いる．OOD はログイン時に認証機構を必要としており，dex との OpeID コネクトや Shibboleth，CAS などの認証方法がある．本研究の実装では，公式が推奨している Dex との OpenID コネクトを用いて LDAP 認証を行う．認証用ディレクトリである LDAP サーバと連携して認証機構を設計する．

3.4.2 スケジューラ抽象化機能と NQSV の連携

はじめに，スケジューラ抽象化機能と NQSV の連携を考える．スケジューラ抽象化機能の基板である PSI/J は，ジョブの情報を格納する Job クラスとジョブの投入や削除などのメソッドをスケジューラごとに再定義している JobExecutor クラスにより構成されている．本研究では新たに NQSV 用の JobExecutor クラスを作成し，ジョブの投入，削除，ジョブの状態確認を行うための三つのメソッドを実装する．PSI/J が対応している他のスケジューラ (Slurm, PBS Pro, LSF, Flux, Cobalt) はジョブの終了後にジョブの状態 (COMPLETED, CANCELED, FAILED) をコマンドの出力結果から確認できる．しかし，NQSV ではジョブの終了後に COMPLETED, CANCELED, FAILED の状態を確認できない．そのため，NQSV に対応するためには，ジョブの投入，ジョブの削除，および待機中のジョブの存在確認に基づいてジョブの状態を PSI/J 側で把握する必要がある．この機能を実現するため，本研究の実装ではジョブが投入された後にジョブキューからジョブが無くなった際に，そのジョブの状態を COMPLETED に変更する．また，ジョブが削除

された際には、そのジョブの状態を CANCELED に変更する。それ以外にジョブキューの状態をコマンドを用いて定期的に確認し、出力結果に応じて QUEUED あるいは ACTIVE という状態にする。

具体的な実装をコード 3.4.2 に示す。必要なライブラリのインポートや関数外の変数の宣言などは省略する。JobExecutor クラス内で定義されている `get_status_now` メソッドは Job クラスのインスタンスを引数にとり、ジョブの状態を保持する JobStatus クラスを返り値に持つ。引数から受け取った Job クラスのインスタンスからジョブの ID を抽出し、書くスケジューラごとに指定されているジョブの状態確認コマンドをジョブ ID を指定して投げる。出力は行ごとに `lines` に代入され、中身を for ループでまわす。指定した ID のジョブがキューにない場合は「Batch Request: ジョブ ID does not exist on クラスタのホスト名.」という出力が帰ってくるため、この出力が取得され、cancel 時に立つ `cancel_frag` が立っていれば CANCELED 状態にする。一方、この出力が取得され、`cancel_frag` が立っていない場合は、キュー内部にジョブがないが、ジョブの削除が行われていないので、ジョブの状態は COMPLETED にする。また、ジョブの投入時に立つ `submit_frag` が立っていない場合はそもそもジョブの投入が成功していないので状態を FAILED にする。なお、他のスケジューラではジョブ状態の詳細メッセージ出力される場合があるが、NQSV にはメッセージ出力機能がないため `message` 変数は None としている。

このように PSI/J 自身がジョブの状態を管理することにより、さらに広い範囲のジョブスケジューラに対応することができることから、ジョブスケジューラ抽象化機能の汎用性を高めることができたといえる。

コード 1: ジョブの状態取得メソッド

```
1 class NQSVJobExecutor(BatchSchedulerExecutor):
2     def get_status_now(self, job: Job) -> Job.status:
3         native_ids = ''.join(str(job.native_id))
4         command = ['qstat', '-F', 'rid,slt', '-n', '-l', native_ids]
5         out = subprocess.run(command, capture_output=True, text=True).stdout
6         r = {}
7         lines = iter(out.split('\n'))
8
9         for line in lines:
10             if not line:
11                 continue
12             cols = line.split()
13
14             if(len(cols) == 8 and self.cancel_frag):
15                 s = cols[2]
16                 native_id = ""
17                 for char in s:
18                     if char.isdigit():
```

```

19         native_id += char
20         state = JobState.CANCELED
21         r[native_id] = JobStatus(state=state, message=None)
22         return r[native_id]
23
24     elif(len(cols) == 8 and not(self.cancel_frag)):
25         s = cols[2]
26         native_id = ""
27         for char in s:
28             if char.isdigit():
29                 native_id += char
30         state = JobState.COMPLETED
31         r[native_id] = JobStatus(state=state, message=None)
32         return r[native_id]
33
34     elif(not(self.submit_frag)):
35         s = cols[2]
36         native_id = ""
37         for char in s:
38             if char.isdigit():
39                 native_id += char
40         state = JobState.FAILED
41         r[native_id] = JobStatus(state=state, message=None)
42         return r[native_id]
43
44     else:
45         assert len(cols) == 2
46         match = re.search(r'\b(\d+)\b', cols[0])
47         native_id = match.group(1) if match else None
48         native_state = cols[1]
49         state = self._get_state(native_state)
50         msg = None
51         r[native_id] = JobStatus(state=state, message=msg)
52         return r[native_id]

```

3.4.3 ウェブ機能とスケジューラ抽象化機能との連携

3.5 結言

第 4 章 実装評価

4.1 緒言

4.2 評価環境

4.3 評価条件

4.4 実行時オーバヘッドの評価

4.5 結言

第 5 章 結論

参考文献

- [1] Ping Luo, Benjamin Evans, Tyler Trafford, Kaylea Nelson, Thomas J. Langford, Jay Kubeck, and Andrew Sherman. Using Single Sign-On Authentication with Multiple Open OnDemand Accounts: A Solution for HPC Hosted Courses. *IEICE TRANS. INF. SYST*, No. 9, pp. 2307–2314, 9 2018.
- [2] David E. Hudak, Thomas Bitterman, Patricia Carey, Douglas Johnson, Eric Franz, Shaun Brady, and Piyush Diwan. OSC OnDemand: A Web Platform Integrating Access to HPC Systems, Web and VNC Applications. *XSEDE '13*, No. 49, pp. 1–6, 7 2013.
- [3] Robert Settlage, Eric Franz, Doug Johnson, Steve Gallo, Edgar Moore, and David Hudak. Open OnDemand: HPC for Everyone. *ISC 2019 Workshops*, Vol. 11887, pp. 504–513, 12 2019.
- [4] Masahiro Nakao, Masaru Nagaku, Shinichi Miura, Hidetomo Kaneyama, Ikki Fujiwara, Keiji Yamamoto, and Atsuko Takefusa. Introducing Open OnDemand to Supercomputer Fugaku. *SC-W 2023*, pp. 720–727, 12 2023.
- [5] Mihael Hategan-Marandiuc, Andre Merzky, Nicholson Collier, Ketan Maheshwari, Jonathan Ozik, Matteo Turilli, Andreas Wilke, Justin M. Wozniak, Kyle Chard, Ian Foster, Rafael Ferreira da Silva, Shantenu Jha, and Daniel Laney. PSI/J: A Portable Interface for Submitting, Monitoring, and Managing Jobs. *IEEE 19th International Conference on e-Science*, 2023.

謝辭