An energy-aware job scheduling method supporting on-demand job execution

Daiki Nakai[1], Keichi Takahashi[1,2], Yoichi Shimomura[2], and Hiroyuki Takizawa[1,2]
1 Graduate School of Information Sciences, Tohoku University
2 Cyberscience Center, Tohoku University
daiki.nakai.q4@dc.tohoku.ac.jp

Introduction

Background

- High performance computing (HPC) systems are not only used for **traditional batch jobs but also for on-demand jobs with deadlines.**
- An HPC system usually **has a power management mechanism**, and thus a node could be **powered off or turned into a low-power mode** based on the operation policy of the system.

Assumptions

- There are two potential reasons that a node is unavailable for on-demand job execution upon the job request.
  - The node is **running another job.**
  - The node is **in a low-power mode or powered off**, referred to as sleeping.

Suspend overhead

Model of the suspend overhead

- Suspending a job is accomplished by **writing out the memory content to persistent storage** and resuming the job is by **restoring the memory content from storage.**
- The overhead required to suspend and resume a job are hence **proportional to the memory usage of the job**.
- **If multiple jobs need to be suspended**, their memory contents are sequentially **written to persistent storage one by one**.

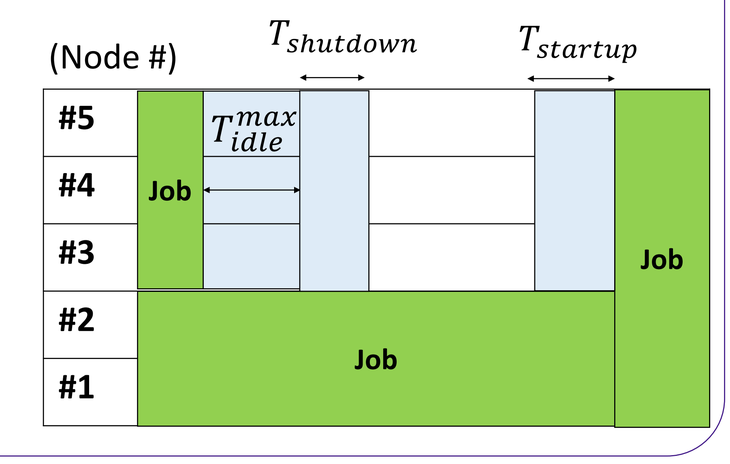
$$O = \underbrace{\frac{1}{\beta_w}\sum_{i=1}^{n} M_i}_{\text{suspend}} + \underbrace{\frac{1}{\beta_r}\sum_{i=1}^{n} M_i}_{\text{restore}}$$

O : suspend overhead
M_i : Memory usage of job i
n : Number of jobs
β_w : Write bandwidth
β_r : Read bandwidth

Shared computing resources with power-saving

Each node is in one of the following three modes.

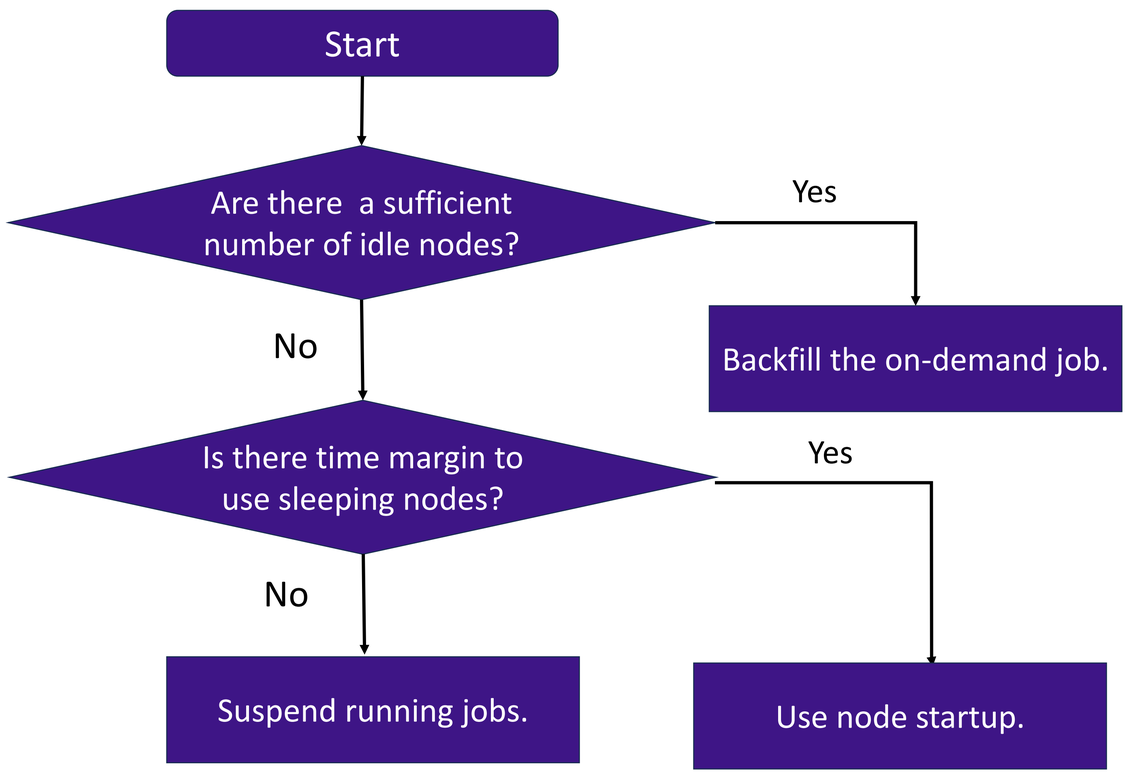
- Running
  - A job is running on the node.
- Idle
  - Any job is assigned to the node.
- Sleep
  - The node is **powered off** since it has not been unused for **a certain period**.



Proposed Method

- This work proposes a job scheduling method **for selecting whether to suspend a running job or to start up a sleeping node.**
- The proposed method **reduces the total overhead by appropriately selecting** how to secure computing resources for on-demand jobs.

Algorithm



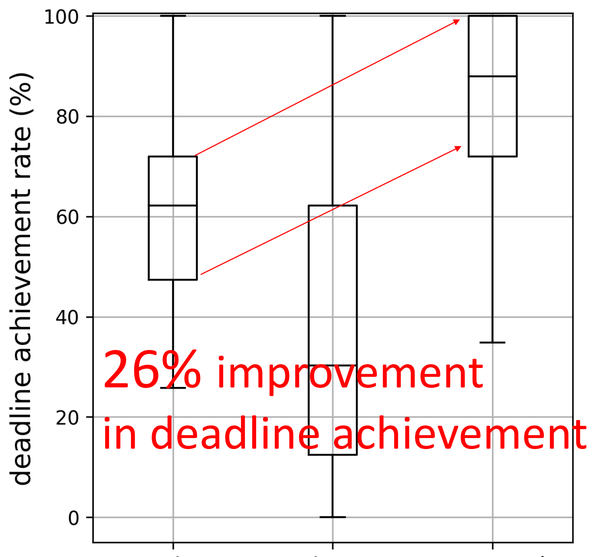
Evaluation

- The simulation in this work assumes the system configuration of **AOBA-A**[1] installed at the Cyberscience Center, Tohoku University.
- Two types of jobs used in the evaluation
  - **On-demand jobs with tight deadlines** (e.g., urgent jobs for disaster mitigation)
  - **On-demand jobs with loose deadlines** (e.g., interactive jobs for user commands)

Job parameter

	node	Excution time	Deadline
Urgent job	64	500 (s)	600 (s)
Interactive job	8～32	1,200 ～ 1,800 (s)	2,400 (s)

Urgent job Evaluation



Interactive job Evaluation

- **The execution efficiency of normal jobs** is improved by **8.2%.**
- **The power consumption** is improved by **0.56%.**
- **Deadline achievement rate** is maintained at **100%.**

[1] Hiroyuki Takizawa,2023. AOBA: The Most Powerful Vector Supercomputer in the World. In Sustained Simulation Performance 2022. Springer Nature

Conclusion

In this work, we proposed a method that uses **node start up for on-demand jobs with loose deadlines** and **suspension for jobs with tight deadlines.**
This study makes two contributions

- **The deadline achievement rate is improved** thanks to the proper selection of running jobs to be suspended.
- **Node start up improves the execution efficiency of normal jobs, and suppresses the increase in power consumption while maintaining a deadline achievement rate.**

HPC Asia 2024

This study was partially supported by JSPS KAKENHI Grant Numbers JP21H03449 and JP22K19764.

https://www.hpc.is.tohoku.ac.jp/home-en/
daiki.nakai.q4@dc.tohoku.ac.jp