Cybermedia Center

Osaka University, Japan

SC22 BOOTH 1613

AI assisted job scheduler / Profile guided vector optimization

AI assisted job scheduler: Cloud Burst Optimization with Deep Q Network

Background

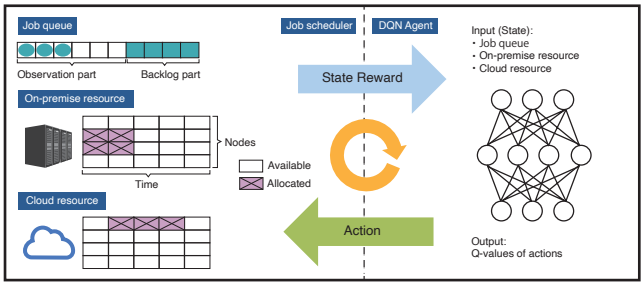
- Cloud bursting becomes attractive for HPC systems to prevent an increase of job waiting time under high load.
- However, it is still difficult to control the tradeoff between job waiting time and cloud cost.

Proposal

- Job scheduler with DQN (Deep Q Network) that can optimize the tradeoff of cloud bursting

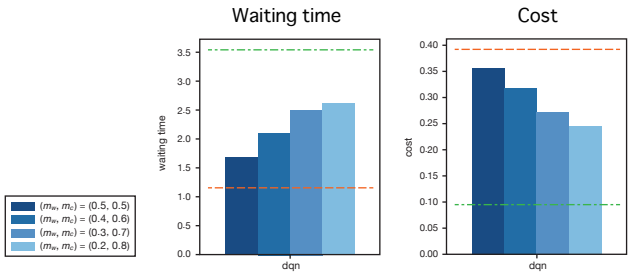
Architecture

- Job scheduler provides state of job queue and on-premise and cloud resources to DQN when scheduling a job
- DQN returns action that shows the scheduler should assign on-premise or cloud resources to the job or skip scheduling
- Job scheduler schedules the job based on the action
- Job scheduler provides reward to DQN for evaluating the action based on a waiting time and a cloud cost



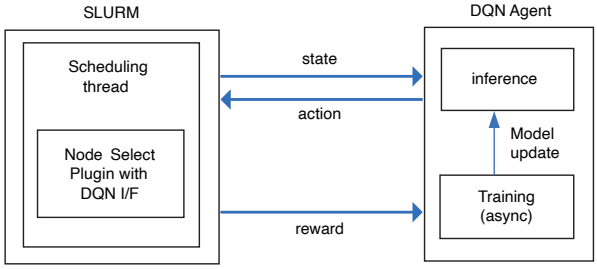
Results

- Proposed architecture can control tradeoff between waiting time and cloud costs



Future Work

- Evaluation and implementation of the proposed method into the SLURM scheduler



Profile guided vector optimization for SX-Aurora TSUBASA

Background

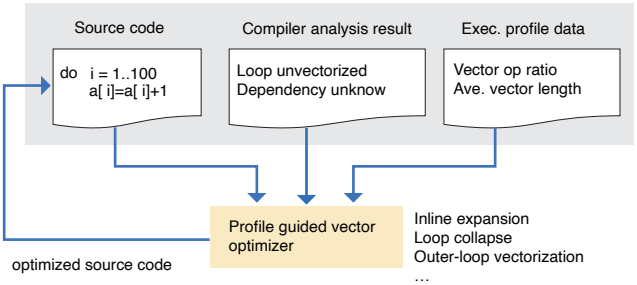
- Realization of vector optimization by users without HW knowledge

Proposal

- Automatic source-to-source translation tool by Profile Guided Vector Optimization (PGVO) for SX-Aurora TSUBASA

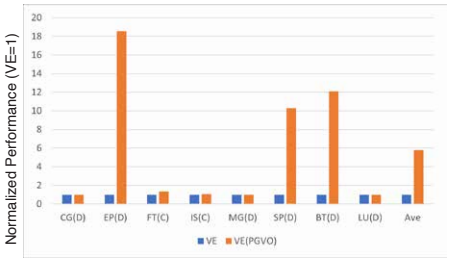
Architecture

- PGVO uses source codes, compiler analysis results and execution profile data as inputs
- PGVO outputs translated source codes



Results

- Significant performance improvement by PGVO compared to automatic vectorization compiler with 3/8 workloads* of NPB
  (* Human-optimized codes that assumes tool behavior are evaluated)



- NPB 3.4.2 (omp version)
- VE: SX-Aurora 20B
  1VE(8core) + ncc/nfort

- ☐ EP/SP/BT: PGO achieves great improvement
- ☐ CG/MG: Compiler already achieves good performance. No room for PGVO.
- ☐ FT/IS/LU: Some room for optimization, but current PGVO achieves little or no improvement.

Future Work

- Implement as a tool and confirm the feasibility
- Evaluate with more workloads
- Support more vectorization technologies

These works were carried out in Joint Research Laboratory for Integrated Infrastructure of High Performance Computing and Data Analysis
https://www.nri.cmc.osaka-u.ac.jp/

Contact : sc22@ais.cmc.osaka-u.ac.jp https://www.cmc.osaka-u.ac.jp/