Performance Prediction of Large Data Transfers

Jason Cheung[1], Alex Sim (advisor) [2], Jinoh Kim (advisor) [2], John Wu (advisor) [2]
[1]Stony Brook University, [2]Lawrence Berkeley National Laboratory

ABSTRACT

Scientific facilities around the world transfer terabytes of data to Berkeley Lab's National Energy Research Scientific Computing Center (NERSC) for processing. These large data transfers can cause congestion on the computer network. To better manage these large transfers, we plan to predict their expected transfer time using machine learning techniques. Through a careful study of traffic logs (Tstat), we find an effective way of utilizing information from recently completed transfers to improve the prediction accuracy by up to 30%.

INTRODUCTION

- NERSC uses specialized Data Transfer Nodes (DTNs) for high-speed data transfers
- Flows traveling to and from the DTNs are measured using Transfer Control Protocol Statistics (Tstat)
- Our Tstat logs contain over 300 million flows in 2019 and 2020
- Large data transfers are defined as the largest 5% of flows (larger than 5 GB)

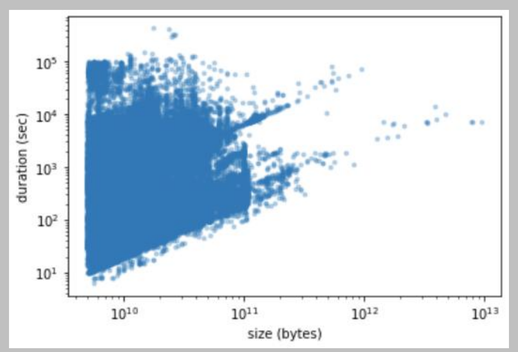


Figure 1: Duration vs size of large transfers

RESEARCH QUESTION

Can historical information improve the predicted duration of large data transfers?

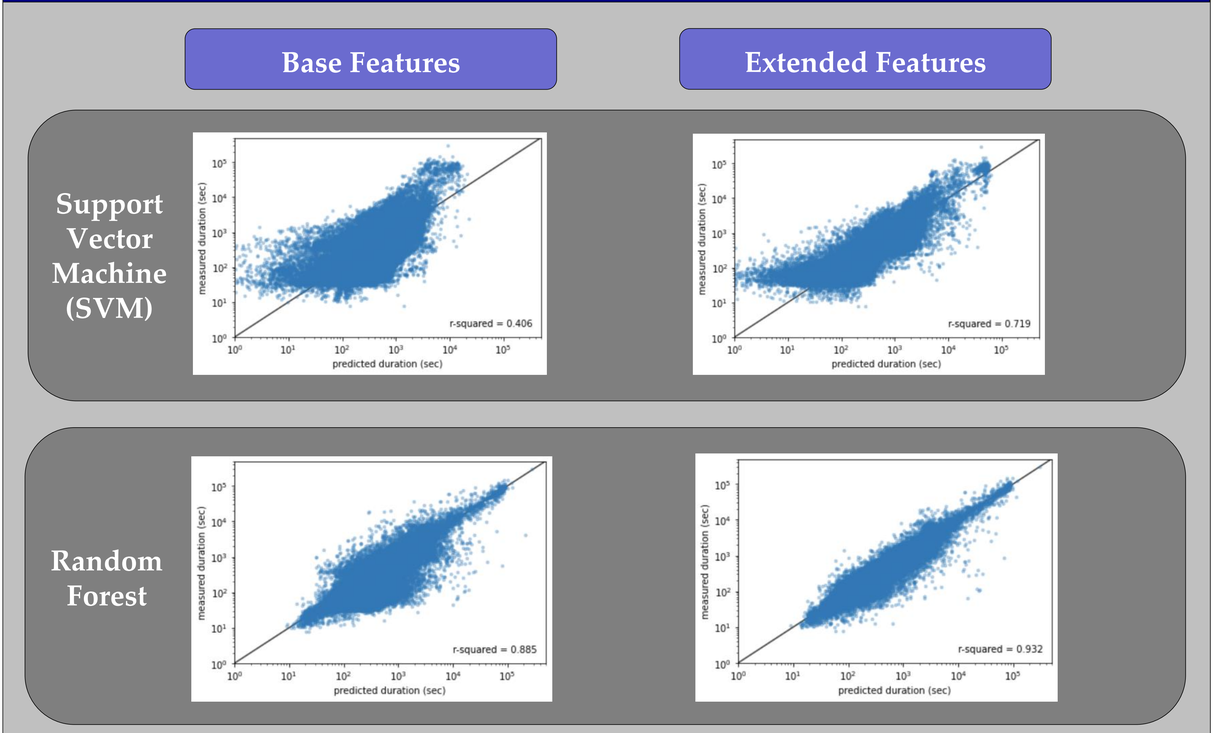
METHODS

- Many features are only known after a transfer is complete and cannot be used
- Predictions can reference recently completed transfers for unknown features

Sample Base Features:
- Source / Destination IP
- Size
- Pinged round trip time (RTT)

Sample Extended Features:
- RTT standard deviation
- Throughput
- Congestion window size

RESULTS



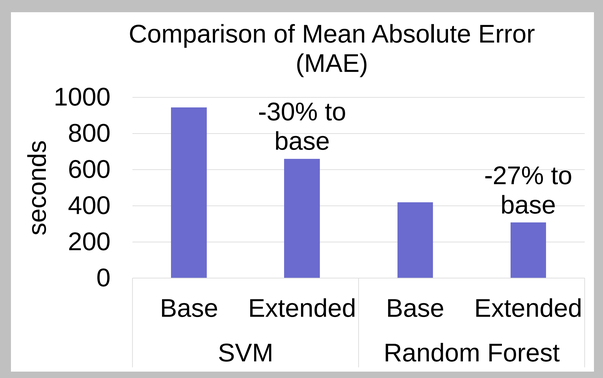


Figure 2:
(Top) Measured vs predicted duration for support vector machine and random forest trained on both feature sets

(Left) The extended features improve model performance by up to 30%, and support vector machine is outperformed by random forest. An MAE of 300 seconds is approximately 3% of the longest transfer times.

FURTHER READING

CONCLUSION

- Random forest is the best performing model on Tstat data
- Referencing recent transfers improves model performance by up to 30%
- The extended features add awareness to network conditions and anomalies
- In the future, we plan to train deep learning models using the two feature sets and classify low performance flows

ACKNOWLEDGMENTS

This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI). This work was supported in-part by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC).