Computer scientists develop novel software to smartly balance data processing load in supercomputers

April 30 2019, by Amy Loeffler

From left to right: Arnab K. Paul, second author and Ph.D. candidate in the Department of Computer Science; Ali Butt, professor of computer science; and first author Bharti Wadhwa, Ph.D. candidate in the Department of Computer Science. Credit: Virginia Tech

The modern-age adage "work smarter, not harder" stresses the

importance of not only working to produce, but also making efficient use of resources.

And it's not something that supercomputers currently do well all of the time, especially when it comes to managing huge amounts of data.

But a team of researchers in the Department of Computer Science in Virginia Tech's College of Engineering is helping supercomputers to work more efficiently in a novel way, using machine learning to properly distribute, or load balance, data processing tasks across the thousands of servers that comprise a supercomputer.

By incorporating machine learning to predict not only tasks but types of tasks, researchers found that load on various servers can be kept balanced throughout the entire system.The team will present its research in Rio de Janeiro, Brazil, at the 33rd International Parallel and Distributed Processing Symposium on May 22, 2019.

Current data management systems in supercomputing rely on approaches that assign tasks in a round-robin manner to servers without regard to the kind of task or amount of data it will burden the server with. When load on servers is not balanced, systems get bogged down by stragglers, and performance is severely degraded.

"Supercomputing systems are harbingers of American competitiveness in high-performance computing," said Ali R. Butt, professor of computer science. "They are crucial to not only achieving scientific breakthroughs but maintaining the efficacy of systems that allow us to conduct the business of our everyday lives, from using streaming services to watch movies to processing online financial transactions to forecasting weather systems using weather modeling."

In order to implement a system to use machine learning, the team built a

novel end-to-end control plane that combined the application-centric strengths of client-side approaches with the system-centric strengths of server-side approaches.

"This study was a giant leap in managing supercomputing systems. What we've done has given supercomputing a performance boost and proven these systems can be managed smartly in a cost-effective way through machine learning," said Bharti Wadhwa, first author on the paper and a Ph.D. candidate in the Department of Computer Science. "We have given users the capability of designing systems without incurring a lot of cost."

The novel technique gave the team the ability to have "eyes" to monitor the system and allowed the data storage system to learn and predict when larger loads might be coming down the pike or when the load became too great for one server. The system also provided real-time information in an application-agnostic way, creating a global view of what was happening in the system. Previously servers couldn't learn and software applications weren't nimble enough to be customized without major redesign.

"The algorithm predicted the future requests of applications via a time-series model," said Arnab K. Paul, second author and Ph.D. candidate also in the Department of Computer Science. "This ability to learn from data gave us a unique opportunity to see how we could place future requests in a load balanced manner."

The end-to-end system also allowed an unprecedented ability for users to benefit from the load balanced setup without changing the source code. In current traditional supercomputer systems this is a costly procedure as it requires the foundation of the application code to be altered

"It was a privilege to contribute to the field of supercomputing with this

team," said Sarah Neuwirth, a postdoctoral researcher from the University of Heidelberg's Institute of Computer Engineering. "For supercomputing to evolve and meet the challenges of a 21st-century society, we will need to lead international efforts such as this. My own work with commonly used supercomputing systems benefited greatly from this project."

The end-to-end control plane consisted of storage servers posting their usage information to the metadata server. An autoregressive integrated moving average time series model was used to predict future requests with approximately 99 percent accuracy and were sent to the metadata server in order to map to storage servers using minimum-cost maximum-flow graph algorithm.

Provided by Virginia Tech

Citation: Computer scientists develop novel software to smartly balance data processing load in supercomputers (2019, April 30) retrieved 25 April 2024 from https://techxplore.com/news/2019-04-scientists-software-smartly-supercomputers.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.