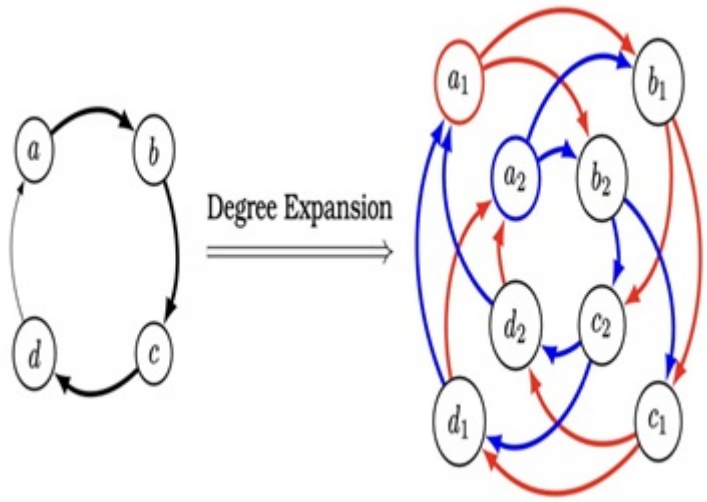
Optimizing Network Topology for Collective Communications

A general, highly scalable algorithmic framework is proposed to synthesize a network topology and schedule optimized for a given collective operation run on a network. The approach tailors efficient designs for distributed systems, demonstrating significant performance benefits even for large-scale operations.



Technology ID

BDP 8857

Category

Software/Network Infrastructure

Selection of Available Technologies

Authors

Arvind Krishnamurthy

Learn more

What is the Problem?

In the dynamic landscape of modern distributed computing and high performance computing (HPC), efficient communication among nodes within a cluster is critical. As advancing hardware has opened up network topology as a new way to optimize collective operations, this challenge becomes increasingly important when considering applications such as the training of machine learning (ML) models.

However, existing network topologies do not use this variability and underutilize regular patterns of data flow during ML training. This results in falling short of achieving the optimal balance between latency and bandwidth. Moreover, the quest for optimal network topologies and corresponding communication schedules is highly complex; one that involves navigating a vast parameter space while considering the various tradeoffs. Bridging this gap requires innovative approaches that tailor network topologies specifically for a given network and collective communications operation, ensuring seamless data exchange while minimizing bottlenecks. The challenge lies not only in identifying efficient topologies but also in doing so in a reasonable time.

What is the Solution?

The software introduces a sophisticated and adaptable algorithmic framework designed to enhance collective communications within distributed networks. At its core, the solution aims to optimize both the network topology and communication schedule, leveraging scalable graph-theoretic approaches.

To address larger-scale scenarios, the software employs two essential tools: expansion techniques, and breadth-first broadcast (BFB) schedule generation. The first makes use of small-scale known optimal combinations of topologies and schedules and expands them in size and complexity to fit the requirement. The second makes use of known large-scale topologies from graph theory and generates a schedule for them in polynomial time.

After using these two approaches to generate a list of viable options, the system explores the range of topology-schedule pairs using a searching algorithm. It identifies the most efficient combination tailored to the workload and compiles it down to the hardware level.

What is the Competitive Advantage?

The technology offers several advantages over related work in the field of collective communications in direct-connect networks. While some methods rely on pre-defined network topologies, this algorithm synthesizes optimized topologies and schedules, addressing the limitations of traditional topologies. It optimizes both the topology and the associated schedule, ensuring seamless data exchange while minimizing latency. This holistic approach directly translates to improved system performance. Evaluations across various clusters and simulations demonstrate these tangible improvements, leading to a reduction in cost for running collective operations.

References

1. Liangyu Zhao, Siddharth Pal, Tapan Chugh, Weiyang Wang, Jason Fantl, Prithwish Basu, Joud Khoury, Arvind Krishnamurthy(45424) , http://arxiv.org/abs/2202.03356, http://arxiv.org