Research Statement

Li Chen

Thanks to the exponential growth of data that needs to be processed in cloud datacenters, data parallel frameworks have emerged as foundations of cloud computing. It becomes increasingly significant to improve the performance of big data jobs running in a shared cluster. My research interests lie in this broad area of big data and cloud computing, which involve the optimization of machine learning frameworks, datacenter networking, network optimization and resource allocation. My thesis research work has focused on improving the performance of sharing jobs with fairness taken into consideration, through the optimal design of resource allocation and network optimization. In particular, the primary themes of my work are threefold: first, understanding the performance requirements, fairness constraints and their tradeoffs for the sharing big data jobs; second, designing optimal and practical solutions by leveraging these insights; third, implementing the solutions for real-world deployment and performance evaluation. Valuable research problems centering on these themes have been investigated in the contexts of both the intra-datacenter network and the inter-datacenter network. These studies have resulted in a disciplined performance-centric fair allocation of bandwidth, scheduling of tasks as well as optimal application-layer multipath routing.

I Current Research

Performance-Centric Resource Allocation in Datacenter Networks. In data parallel frameworks, a job typically proceeds in multiple computation stages, each consisting of a number of parallel tasks. Between consecutive stages, multiple network flows need to be initiated in parallel to transfer a large volume of intermediate data. In datacenter network, such network transfers usually have a significant impact on the performance of typical jobs in big data processing.

As the datacenter network is shared by multiple concurrent jobs, how should we allocate link bandwidth among their flows? It is commonly accepted that bandwidth should be shared in a fair manner, yet there has been no consensus on how the notion of fairness should be defined. The traditional wisdom on fair bandwidth sharing has largely focused on datacenters in a public cloud, with an emphasis on payment proportionality, which is not applicable to the context of a private datacenter. To fill this gap, we have proposed the principle of weighted performance-centric fairness, regulating that fairness should be maintained with respect to the performance across multiple jobs. In particular, this notion of fairness indicates that the reciprocal of the transfer times should be proportional to the weights across sharing jobs.

Guided by this principle, we have designed an optimal bandwidth allocation strategy that maximizes the social welfare with a tunable degree of relaxation on fairness, which can be practically implemented in a distributed fashion. This work has been published on IEEE

INFOCOM 2014. Complementary to the bandwidth allocation, we have extended the application of our fairness in both the task placement and the multipath routing in datacenter networks, resulting in two papers published on IEEE IC2E 2016. We have also investigated the tradeoff between our proposed fairness and efficiency, and proposed algorithms to arbitrate the tradeoff. This work will appear in IEEE Transactions on Parallel and Distributed Systems, 2018.

Utility-Optimal Coflow Scheduling with Max-Min Fairness in Datacenter Networks. For a typical data parallel job, the network transfer is not considered complete until all of its constituent flows have finished. It is the collective behavior of all of these flows that matters, rather the individual behavior of each flow. These flows are hence referred to as a coflow.

As the datacenter network is shared by active coflows from multiple competing jobs, it is critical to schedule these coflows efficiently and fairly. Existing research efforts on coflow scheduling focused on minimizing coflow completion times and meeting coflow deadlines, which treat coflows identically as equal citizens. However, due to their inherent nature, different jobs have widely diverging requirements with respect to their completion times: an interactive query in a web application should not be similarly treated as a background job for data analytics. Intuitively, different jobs have different sensitivity to their completion times, which can be characterized by different utility functions. With this observation, we argue that more time-sensitive coflows should be allocated more network resources, allowing them to complete earlier, achieving a higher utility based on their utility functions.

Therefore, we have designed and implemented a new utility optimal scheduler to provide differential treatment to coflows with different degrees of sensitivity, yet still satisfying max-min fairness across these coflows. The scheduling algorithm is designed based on the theoretical study of a lexicographical optimization problem. We have decoupled the problem into subproblems with integer variables, which is rigorously proved to be equivalent to linear programming problems that can be efficiently solved. This work has been published on IEEE INFOCOM 2016.

Optimal Job Scheduling with Max-Min Fairness in Inter-Datacenter Networks. Large volumes of data are generated across multiple datacenters around the world, which is increasingly typical for global services deployed by Microsoft and Google generating user activity and system monitoring logs. It becomes a challenging issue to optimize data analytic jobs deployed in such an inter-datacenter network. To process geo-distributed data, a naive approach is to gather all the data to be processed locally within a single datacenter. Naturally, transferring huge amounts of data across datacenters may be slow and inefficient, since bandwidth on inter-datacenter network links is limited. Existing research has shown that better performance can be achieved if tasks in an analytic job can be distributed across datacenters, and located closer to the data to be processed. In this case, designing the best possible task assignment strategy to assign tasks to datacenters is critical, since different strategies lead to different flow patterns across datacenters, and ultimately, different job completion times.

However, existing works in the literature only considered a single data analytic job when designing optimal task assignment strategies. The problem of assigning tasks belonging to multiple jobs across datacenters remains open. Given the limited amount of resources in each datacenter, multiple jobs are inherently competing for resources with each other. It is, therefore, important to maintain fairness when allocating such a shared pool of resources,

which cannot be achieved if tasks from one job are assigned without considering the other jobs. To address this issue, we have designed a new task assignment strategy to achieve max-min fairness across multiple jobs with respect to their performance. We have rigorously formulated the scheduling problem, addressed the challenges brought by the integer variables and transformed the problem to a linear programming problem with proved equivalency. Our strategy has been implemented and evaluated with real-world experiments, demonstrated to be effective in minimizing the job completion times across all concurrent jobs while maintaining max-min fairness. This work has been published on INFOCOM 2017 and the extended version will appear in IEEE Transactions on Network Science and Engineering, 2018.

Optimal Multi-path Routing with Max-Min Fairness in Inter-Datacenter Networks. To optimize the performance of data analytic jobs deployed in the wide area, existing efforts either reduce the total volume of inter-datacenter traffic or design better task placement for load balancing, which all require modifying the generated traffic pattern across datacenters to alleviate the performance degradation of inter-datacenter transfers. Complementary and orthogonal to these work, we focus on directly optimizing the network transfers given certain traffic patterns. Particularly, we propose to optimize the inter-datacenter transfers by exploiting the path flexibility to better utilize the inter-datacenter link bandwidth and thus accelerate the geo-distributed jobs eventually.

Our strategy design is based on a rigorous formulation of bandwidth allocation and multipath routing problem among sharing flows across datacenters. With a theoretical study of the problem, we have designed an optimal strategy and implemented it in the central controller in the application-layer software-defined network. With extensive real-world experiments, our strategy has been demonstrated to achieve optimal job performance, with max-min fairness achieved among sharing jobs deployed in the wide area.

II Future Research

In the future, I look forward to exploring more research issues related to performance optimization for big data jobs in cloud computing. Beyond the topics of network optimization and resource allocation for data parallel jobs, I will expand my research areas into optimization for a broad range of distributed data analytic systems, from the perspectives of both framework optimization and transport acceleration, with more emphasis on system design and implementation. Moreover, due to the sharing nature of resources in cloud computing, a unified resource sharing system among a diverse range of machine learning applications is expected to be designed, to accommodate different performance requirements and fairness constraints across various applications. These research plans are elaborated as follows.

Distributed System for Machine Learning and Data Analytics in the Wide Area. In the era of big data where everything is personalized and connected, the amount of data to be processed is increasing faster than Moore's Law. With such volume and velocity, various machine learning systems at large scale are developed to handle the extensive amount of data at high speed. As it is inevitable that frequent exchange of intermediate data exists in these systems, network easily becomes the bottleneck, especially in the wide area deployment, which has become the new trend as many large companies have their data generated and stored across geographically distributed datacenters. It is intuitive that deploying the traditional data analytic systems in

the wide area leaves the application performance in the wild. Rather than rearranging the task collocation in the data parallelism system or the adaptive synchronization method in the model parallelism system, I believe we need to design a new system from scratch, with well-planed data ingestion, computing, storage, and synchronization, that is perfectly suitable for the inter-datacenter deployment.

Network Optimization for General Big Data Jobs. Distributed data analytics depends heavily on the reliability and the performance of the underlying networks, due mainly to its abundant demand for data exchange among tasks. However, the networks, being best-effort and shared among many, can hardly deliver consistent or predictable Quality of Service, especially in the wide area networks with low bandwidth and high latency. As such, complementary to the framework-level optimization aforementioned, the network-level optimization for transport acceleration also plays a significant role in improving the performance of big data jobs. Extending my past work which orchestrates inter-datacenter flows at the application-layer, I would like to further explore the rich space of performance improvement from the lower layer protocol design and rate allocation, given the availability of experimental testbed.

General Resource Sharing System for Diverse Applications. The cloud is a shared infrastructure which is expected to accommodate a broad range of applications with different characteristics. It is important to investigate the fundamental principles in sharing multi-dimensional resources across various types of competing applications. This is envisioned to be complementary to the existing schedulers which only focus on particular types of applications or resources. The challenges lie in the accurate modeling of application requirements. To efficiently utilize resources, we need to understand the relationship between performance gain and allocated resource, which becomes more complicated than the simplified linear relationship in previous efforts. Moreover, the principle of fairness also needs to be revisited when applications have different evaluations of the allocated resources. From the theoretical perspective, I would like to seek for an appropriate notion of fairness for this general context and explore the inherent tradeoff between fairness and efficiency. The ultimate goal is a practical resource sharing system which can be general and adaptive to a dynamic environment.

In summary, I plan to explore the theoretical and practical challenges in large-scale distributed systems for big data analytics. I am interested in the rigorous formulation and theoretical study of real-world problems, as well as practical system design and implementation, to have real impacts. To carry out the research in all these areas, I will seek for funding support from both academic agencies and industrial companies. My preliminary experience of proposal writing with my advisor and my rich experience of collaboration with Huawei for two years both help in my future funding application.