Chris Fry

Google Cloud

Forecasting Advances and Lessons Learned in Google Infrastructure Planning

Over the past decade, the scale and complexity of Google's technical infrastructure has increased dramatically. Our Cloud business now represents a double-digit fraction of the company's revenue, bringing with it the need to build, plan, and run reliable services for hundreds of thousands of customers. We also support more first-party and third-party products and consumption models than ever before. And in recent years, competition to lead in the AI space has reached unprecedented levels, with capacity and supply constraints making efficiency, reliability, and predictability even more critical. Staying ahead of these dynamics requires continuous innovation in forecasting and planning. We are developing new capabilities to support cross-resource planning in a holistic way, enabling us to plan workloads consistently across accelerators, compute, storage, network, space, power, and other resources. We are investing in forecasting and capacity planning models that exploit demand and supply fungibility across products and platforms. And we continue to invest in research to develop and refine our models to improve accuracy, efficiency, and resource obtainability.