ABSTRACT: The devices and software systems we interact with on a daily basis are more intelligent than ever. The computing required to deliver these experiences for end-users is hosted in Warehouse Scale Computers (WSC) where intelligent web services are employed to process user images, speech, and text. These intelligent web services are emerging as one of the fastest growing class of web services. Given the expectation of users moving forward is an experience that uses intelligent web services, the demand for this type of processing is going to drastically increase. Today’s cloud infrastructures, tuned for traditional workloads such as Web Search and social networks, are not adequately equipped to sustain this demand.

This dissertation shows that applications that use intelligent web service processing on the path of a single query require orders of magnitude more computational resources than traditional Web Search. Intelligent web services use large pretrained machine learning models to process image, speech, and text based inputs and generate a prediction. As this dissertation investigates, we find that hosting intelligent web services in today’s infrastructure exposes three critical problems: 1) current infrastructures are computationally inadequate to host this new class of services, 2) system designers are unaware of the bottlenecks exposed by these services and the implications on future designs, 3) the rapid algorithmic churn of these intelligent services deprecates current designs at an even faster rate.

This dissertation investigates and addresses each of these problems. After building a representative workload to show the computational resources that would be required by an application composed of three intelligent web services, this dissertation first argues that hardware acceleration is required on the path of a query to sustain demand moving forward. Second, we focus on Deep Neural Networks (DNN), a state-of-the-art algorithm for intelligent web services, identifying critical compute bottlenecks that inform the design of a system based on Graphics Processing Units (GPUs). Finally, we design a runtime system based on a GPU equipped server that improves upon hitherto designed systems accounting for recent advances in intelligent service algorithms.

By thoroughly addressing these problems, we produce designs for WSC that are equipped to handle the future demand for intelligent web services. The investigations in this thesis address significant computational bottlenecks and lead to system designs that are more efficient and cost-effective for this new class of web services.

Advisors: Prof. Jason Mars and Lingjia Tang