Proposed Research Plan to Tackle Memory and Communication Walls as a Margaret Butler Fellow

As we are going into the exascale era of computing, large-scale supercomputing is solving grand problems that once were thought unsolvable. The race to exascale can be seen as the space race in the 60s, but in another area. Now we are racing to demonstrate technological capacity to solve critical problems today, such as understanding climate change or simulating COVID-19 mechanisms at molecular level. In fact, exascale computing may be considered as a psychological milestone because the LINPACK benchmark is based on regular and dense computations. However, only a handful of applications will enjoy exascale deployment on early systems. Because, a plethora of today's relevant scientific, AI, and graph-analytics applications involve irregular and sparse computations, and hence they suffer from memory-wall and communication-wall bottlenecks. As a result, they utilize only a tiny portion of the peak performance in practice.

To obtain higher utilization on exascale computer, my research plan tackles inefficiencies due to irregular and sparse memory accesses and communications. Moreover, the next generation of exascale systems-Frontier, Aurora, and El Capitan-will involve multi-GPU nodes connected with a hierarchical communication network with no exception. Therefore, the proposed algorithms target multi-GPU node architecture and the accompanying interconnect topology. Specifically, I will seek applications for dissemination and generalization of the Tiled SpMM and hierarchical communication techniques that I propose in my doctoral dissertation. Both techniques embrace an inspection/execution model that preprocess the memory access and communication patterns and optimize them to perform distributed matrix multiplication with high performance.

The proposed techniques have already been applied to several award-winning applications to solve unprecedently-large problems. A good example is the X-ray imaging problem to reconstruct 3D images with sub-micron resolution from TB-scale scan data collected at the Advanced Photon Source of Argonne, where I worked with an interdisciplinary team from XSD and MCS divisions. Our SC20 paper reconstructs a 3D mouse brain on 24,576 GPUs on OLCF Summit. The Tiled SpMM throughput reaches 65 mixed-precision PFLOPS and the hierarchical communications reduces the dominating communication time by 60%. To enhance the performance, we reorder rows and columns of the sparse matrix for modifying irregular data-access patterns with space-filling algorithms. This work won the best paper award at SC20. I have also applied the proposed techniques at the IBM-Illinois center to accelerate HPCG benchmark on POWER9 processors by 17%, and also to accelerate sparse deep neural network inference up to 180 TeraEdges/Second on Summit. Our HPEC20 paper obtained the championship title at MIT/Amazon/IEEE Graph Challenge. I am currently collaborating with NVIDIA to contribute to the cuSPARSE library with the SpMM tiling techniques.

My research plan involves collaboration with domain scientists not only at Argonne, but also other national laboratories, industry partners, and academia to apply the proposed techniques and/or developing novel algorithms to solve grand problems on exascale systems. These techniques will be generalized by handling corner cases and will be released as a performant library for the benefit of other exascale application developers. Throughout my graduate programs, I have applied similar techniques on various application domains involving nonlinear inverse problems, fast algorithms, n-body problems, multigrid method, computational imaging, sparse deep neural networks, stencil computations, and graph neural networks, as well as optimized applications such as HPCG, SETSM, and ChaNGa on petascale systems. I believe my previous experience will help me to communicate with domain scientists in the language of applied mathematics and computational science.

In summary, as a Margaret Butler Fellow at Argonne, my research plan is developing novel algorithms for critical applications at scale, especially for those involving irregular and sparse computations and communications. I will optimize these algorithms on multi-GPU node architecture and heterogeneous communication topologies that the exascale computers will embrace. My research will alleviate the memory-wall bottleneck on single GPU and communication-wall bottleneck on thousands of GPUs by exploiting the underlying exascale supercomputer architecture. This will accelerate a plethora of large-scale sparse applications. As a result, my research will yield high-throughput science production at exascale systems and contribute solving grand problems of the next decade.