Towards Tackling Sparse Communication Wall on Exascale Systems

Motivation of the Proposed Research and its Importance

As we are going into the exascale era of computing, large-scale massively-parallel systems are addressing challenging problems that once were thought unsolvable. Now we are racing to demonstrate technological capabilities for solving critical problems for the well-being of citizens of all countries today, such as understanding climate change in global level 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 workloads, and they therefore suffer from memory-wall and communication-wall bottlenecks. As a result, these applications utilize only a tiny portion of the theoretical performance of large-scale computing systems in practice.

To obtain higher utilization on exascale computer, my research plan is to address 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 by a hierarchical communication network with no exception. Therefore, the proposed algorithms target multi-GPU node architecture and the accompanying heterogeneous interconnect topology. More specifically, I will seek applications that can benefit from the Tiled SpMM and hierarchical communication techniques that I have developed in my doctoral dissertation research. Both techniques embrace an inspection/execution model that preprocess the memory access and communication patterns and optimize them for the underlying architecture to perform distributed matrix multiplication with high performance.

Progress to Date: Applications

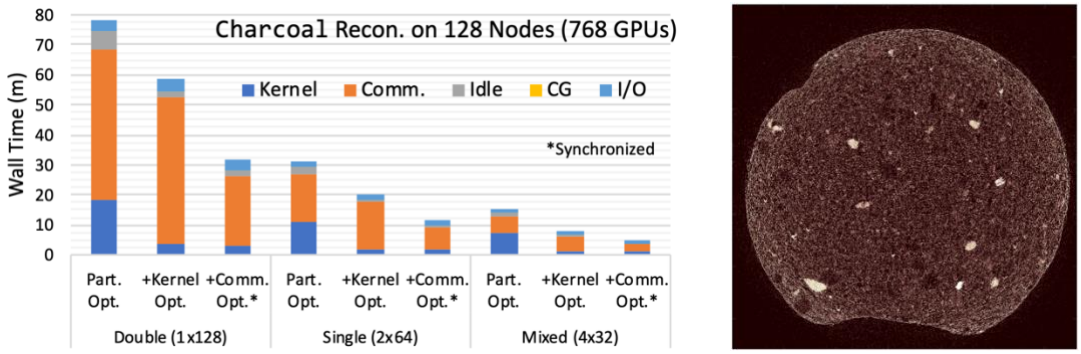
The proposed techniques have already been applied to several award-winning applications to solve problems at unprecedented scales. A good example is the X-ray imaging problem to reconstruct 3D images with sub-micron resolution from TB-scale scan data collected at the Advanced Photon Source of Argonne, where I worked with an interdisciplinary team from MCS/DSL and XSD divisions. I was leading SC19 and SC20 papers demonstrated the reconstruction of a 3D mouse brain on 4,096 KNLs (256k cores) of ALCF Theta and on 24,576 GPUs of 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-curve-based data layout algorithms. My SC19 work has been chosen as the reproducibility benchmark for the student cluster competition thanks to the extensive evaluation and benchmarking of the Tiled SpMM technique across systems, and my SC20 work won the best paper award for the proposed hierarchical communications for sparse data reduction.

I have also applied the proposed Tiled SpMM technique at the IBM-Illinois center 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 Sparse Challenge. I am currently collaborating with NVIDIA to contribute to the cuSPARSE library with the SpMM tiling techniques. For dissemination of my work with my collaborators, we are preparing a submission for a hands-on tutorial on sparse neural network inference at MLSys in 2022.

Optimizing Sparse Communications for Exascale Systems

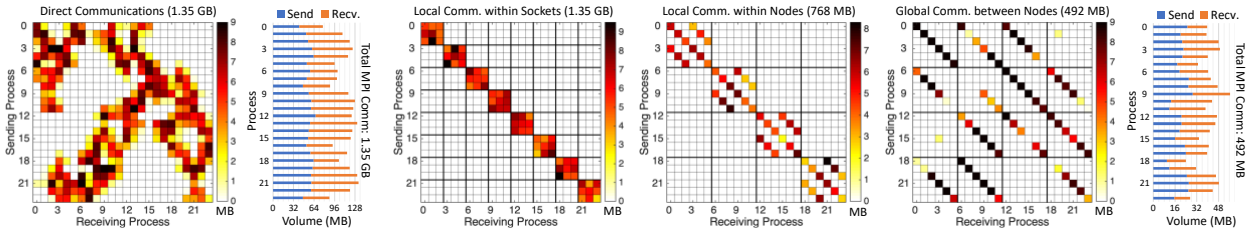
For my postdoctoral research, my plan is the generalization of sparse communication and data reduction for the hierarchical communication topology among GPUs. The hierarchical communications alleviates the communication bottleneck when solving problems with extremely-large sparse matrix A that fits on at least hundreds of multi-GPU nodes. As an example from previous work, Figure 1 shows the breakdown of end-to-end image reconstruction time for a large dataset on 128 nodes of the Summit supercomputer at Oak Ridge National Laboratory. The partitioned sparse matrix representing the forward model of the X-ray phenomena (based on ray tracing) has a memory footprint of 2.82 TB (with its transpose). In the iterative conjugate-gradient (CG) solution takes about 1.3 hours with double precision, and with direct communications (seen in Figure 2(a) for 24 GPUs), most of the time is spent for MPI communications between GPUs.

The hierarchical communications perform an extra two local data exchange and reduction within the three level hierarchy: the first level exchange and reduce data between every three GPUs which are connected through high-bandwidth Nvlinks, the second level similarly exchange and reduce data between every six GPUs within the same node. The local communication and reduction reduce the slow MPI communication between nodes by 60%. The details are provided in the SC20 paper.



(a) (b)

Figure 1. (a) Breakdown of end-to-end reconstruction time for reconstruction of the charcoal dataset on 128 nodes of Summit (on 768 V100 GPUs). (b) A single slice of 3D reconstruction of the charcoal dataset.



(a) (b)

Figure 2. Direct and hierarchical communication matrices and total MPI communications between nodes.

Previous work has been applied to Summit architecture with a fat-tree inter-node communication topology, where the cost of data movement between arbitrary pair of nodes is approximately the same. However, because of the high cost of the fat-tree interconnect, the upcoming exascale computers will involve increasingly heterogeneous topologies, such as the dragonfly in Frontier. My research plan is to extend the communication hierarchy up to the heterogeneous interconnects that takes advantage of the high-bandwidth communication capabilities in the hierarchy.

Reproducibility and Benchmarking

The proposed research will produce reproducible benchmarks for exascale systems. This will introduce an application-driven set of sparse communication benchmarks, e.g., for the top500 list, that currently lacks. The proposed sparse communication benchmark will complement the LINPACK and HPCG benchmarks to rank the communication performance of the exascale systems. It is more than relevant to today's applications involving distributed matrix multiplication and graph analytics with sparse data reduction which is bounded by the communication bottleneck.

Current communication libraries like MPI and NCCL do not support sparse communications and therefore a sparse extension for these libraries is desired. My research plan proposes introducing sparse communication routines optimized for multi-GPU heterogeneous systems and releasing them as a performant library or as an extension to the existing commonly used libraries. Mixed-precision communications with normalization and denormalization will be investigated to minimize the communication volume while avoiding overflows and maintaining the accuracy as much as possible. For sparse reduction routines distributed SpMM, pipelining communications for overlapping them with compute kernels will be applied to reduce end-to-end reduction time.

One reservation of the proposed inspection/execution model is its preprocessing overhead: Both Tiled SpMM and hierarchical communications require inspection of the sparse memory and communication patterns before the execution with optimally high performance. An integral part of the proposed research is quantifying the overhead and its amortization with the speedup in the iterative solution of inverse problems or sparse/graph neural network inference, where the tiling data structures and communication mappings are reused many times.

Collaboration for Science Production

My research plan involves collaborating with domain scientists not only at Berkeley Lab, but also at other national laboratories, industry partners, and academic institutions to apply the proposed techniques. I also plan to develop novel extensions/generalizations by handling application requirements 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 petascale applications such as HPCG, SETSM, and ChaNGa on leadership-class 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 postdoctoral fellow at Berkeley Lab as a Luis J. Alvarez or Admiral Grace M. Hopper Fellow, 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 communication topologies that the exascale systems will embrace. My research will alleviate the memory-wall bottleneck on individual GPUs, and communication-wall bottleneck on system level by exploiting the underlying exascale supercomputer architecture. This will accelerate and scale a plethora of sparse applications. As a result, my research will yield high-throughput science production at exascale systems and contribute solving challenging problems of the coming decade.