CSCI 6360: Parallel Computing Lecture Summary - 5

Anirban Das (dasa2@rpi.edu)

January 30, 2018

Summary of Paper: ”Overview of the Blue Gene/L System Architecture” by Gara et al. IBM J. Research and Development

The paper from IBM gives an insight into the goals and core design decisions behind developing the Blue Gene series of supercomputers and the architectural implementations achieving them. The authors elaborated on the challenges they faced while building the system, such as achieving low latency, low cost and high power/performance ratio at the same time, and also achieving extreme scalability, reliability, monitoring facilities etc.

They used specific high bandwidth low latency network architectures inside BG/L to suit specific application use cases, such as using torus network to assign task to nodes in a geometry mirroring the geometry of the actual physics problem, global tree network to handle one to all broadcast etc. Essentially , the goal is to maintain task and data locality, such that an executing process owns everything including nodes to communication channels in the network partition it is allocated. This dramatically improves efficiency and throughput.

Each node among the mammoth 65,536 BG/L nodes, is a single ASIC containing 2 700Mhz processors (PPC400) and 512MB DDR SDRAM, 4 nodes make a compute card, 32 compute cards make a node card, and 32 node cards make up the rack. Several racks make up the supercomputer. The nodes are low powered, physically small and densely packed to give high cost/performance. Apart from compute nodes, there are I/O nodes, service nodes, and front-end nodes to provide various io/ file handling, scheduling, load balancing jobs, providing the UI, etc.

BG/L is designed to use a distributed memory message passing programming model, mainly MPI interface, and supports some standard programming languages like C, C++ and Fortran. Moreover, the Power PC 440 core harware micro-architecture changes improves floating point calculation throughput. Also, large effort was given to design the memory systems and the software assisted memory coherence of the BG/L nodes.

Finally the authors elaborates about the fault tolerance and availability of the system, achieved using, checkpointing and redundancy at the system level, to low hardware level ECC memory protection protocols, network error detection, parity checking on buses etc. at the node level. These again allows excellent fault tolerance and isolation capabilities.

The whole driving force for such architecture is the idea that the HPC usecases targeted by BG/L, are highly parallelizable. Hence, they will scale up to hundreds of nodes. The performance gain thus obtained by dividing the job into hundreds of low powered low frequency nodes, will eclipse the performance obtained from running it on few high frequency , high power processors. Also, cheaper low frequency nodes can be easily added and with efficient and high speed network backhaul, communication jitter can also be highly improved. The performance/ Watt of BG/L was so dramatically superior to the other supercomputers at that point, that their efficiency oriented system architecture became the standard blue print for the future supercomputers in the years to come.