CMSC 714
Lecture 12
Cray XT4 and IBM Blue Gene/L

Alan Sussman

Notes

- MPI project grading in progress
  - Grades should be posted by end of spring break
- CUDA project due Monday after spring break
  - Questions?
- Timeline for group research project will be posted soon, along with sample topics from previous years
  - Should start thinking about forming groups, potential topics
  - Proposal deadline will be a week after CUDA project is due
  - Questions?
- More readings posted, for week after spring break

Cray XT4 vs. XT3

- Dual core AMD Opteron nodes (later upgraded to quad core) with integrated memory controller for DDR2 RAM (up to 12.8 GB/sec)
  - VN mode splits memory between cores, but only one core handles OS functions and network communication (MPI)
  - SN mode uses only 1 core that accesses all memory, and can do communication
- Nodes connected via SeaStar2 3D toroidal network
  - sustained bi-directional bandwidth of 6GB/sec over 6 links
  - Opteron memory bus directly connected to SeaStar chip (like Intel QPI), not to I/O bus
- Compute nodes run a stripped down version of Linux from Sandia, for scaling
  - to limit effects of OS activities on computations
  - and limit effects of contention between multiple processes/threads running on multiple cores within a node
  - two modes – virtual node (VN) and single/serial node (SN)
- Lustre parallel filesystem
  - multiple I/O nodes on SeaStar network, running full Linux, called object store servers (OSSs), which connect to object store targets (OSTs) that directly talk to I/O devices (disks, RAIDs, etc.)
  - files can be striped across multiple OSTs
  - Paper gets OSS and OST backwards in Figure 1

Cray XT4 evaluation

- Micro-benchmarks
  - HPCC benchmarks to measure network, node-local and global performance
  - network – latency and bandwidth, with different patterns
  - node local – different measures for combinations of spatial and temporal locality
  - global – SP mode uses 1 processor per node (other does communication and OS functions) , EP mode runs same computation on all processors on node (w/o communication between them)
- Application benchmarks
  - Community Atmospheric Model – climate/weather modeling
  - Parallel Ocean Program – ocean circulation model
  - Nanoscale Molecular Dynamics – biomolecule simulations
  - Turbulent combustion – coupled fluid dynamics, chemistry, molecular transport (direct numerical simulation)
  - Fusion plasma heating – all orders spectral algorithms

Blue Gene/L

- Scalable high performance distributed memory machine, with interesting design decisions
  - main goal is high performance with low power consumption, and high reliability
  - idea is to scale to large configurations of low power, less powerful individual components
  - Distributed memory system, with up to 64K nodes
- Each node is a dual processor chip, with integrated memory controller, network interface, caches, etc.
  - System on chip (SoC) design
  - 2 PowerPC cores, split L1 cache per core, L2 cache/prefetch buffer per core, 1 shared L3 cache, 512MB shared memory
    - L1 caches not hardware coherent, so need software help, other cache levels are coherent
  - Floating point multiply-add instructions for improved power/performance (when used)
  - Plus link chip for network component, for all routing and other network ops between nodes

Blue Gene/L

- 5 distinct networks, all interfaced to node through link chip
  - 3D torus is main message passing network – each node has 6 bi-directional nearest-neighbor links, with cut-through routing
  - Collective network for broadcasts, reductions, etc. – each node has 3 links (parent, 2 children, in a binary tree)
    - also forwards I/O requests to I/O nodes
  - Barrier network – for global sync operations (or other OR or AND operations from a set of nodes)
  - Gigabit Ethernet for file system access (I/O nodes only)
  - Fast Ethernet (100Mb) for initialization, diagnostics, debugging
- Programming model allows for either using each core for a separate process (and split the physical memory, but allow sharing of read-only data), or use 1 core as a communication co-processor