CATEGORY: SUPERCOMPUTING & HPC - SC02
POSTER P6207 | CONTACT NAME Gavin Baker: gabaker@calpoly.edu
GPU TECHNOLOGY CONFERENCE

Astro: A Low Cost Low Power Computing Cluster

CAL POLY SAN LUIS OBISPO

Gavin Baker, Sean Sheen, Dr. John Oliver, Dr. Chris Lupo.
California Polytechnic State University, San Luis Obispo, CA
{gabaker, ssheen, jyoliver, clupo}@calpoly.edu

Sandia National Laboratories

Abstract

On the path to exascale computing, the trend toward more expensive, power hungry High Performance Computing (HPC) clusters presents a number of challenges to academic institutions, with limited funding, who wish to contribute to the study of scalable and resilient HPC architectures and software runtime systems. These future supercomputer architectures will likely contain heavy-duty hardware accelerators that are used to maximize floating-point performance (FLOPS) while minimizing power consumption. The cost of procurement for sample systems (with few nodes) can easily exceed the budget of most academic institutions. Additionally, communication bottlenecks introduced by PCIe can limit the efficiency of accelerator cards due to memory transfer costs. This poster will demonstrate the efficacy of hybrid CPU-GPU systems within commodity computing clusters through our design, implementation, and benchmarking of a hybrid CPU-GPU cluster based on the NVIDIA Jetson TK1 embedded system. The Jetson TK1 development board consists of a hybrid CPU-GPU, system on a chip (SOC), that runs on minimal power and provides the potential for massively parallel acceleration from 192 CUDA cores. Our new computing system, which we will refer to as "Astro", has a theoretical peak floating-point performance of ~24 TFLOPS and operates at less than 1KW power consumption while under full load. For practical application performance, we are able to see a parallel efficiency for the LAMMPS molecular dynamics simulation of above 60%. The cost of the entire Astro cluster was approximately $13,000, less than two high end workstation computers with Kepler GPUs.

Requirements

- Maximize system-performance and node-count while maintaining a total budget of $15,000 including cabling, rack hardware, network equipment and Jetson boards.
- Minimize power consumption of cluster equipment to allow simple power measurement and allow for entire system to use **one** standard 110 Volt wall socket as a power source.
- Utilize commodity CPUs + GPU accelerators (Jetson TK1 Tegra Development Board) and networking equipment (ethernet switches).
- Design a system with a reasonable node count to simulate modern HPC systems and allow for the study of the efficacy of hybrid architectures in full scale HPC systems as well as resilience in HPC runtime systems (MPI, Charm++, etc...).
- Benchmark system performance in the following categories: floating point computation performance, network bandwidth, memory access rates and overall practical performance availability of the cluster on real HPC applications.
- Provide an easy to use platform for deployment of HPC applications and the study of network topology on application performance.

Benchmarks

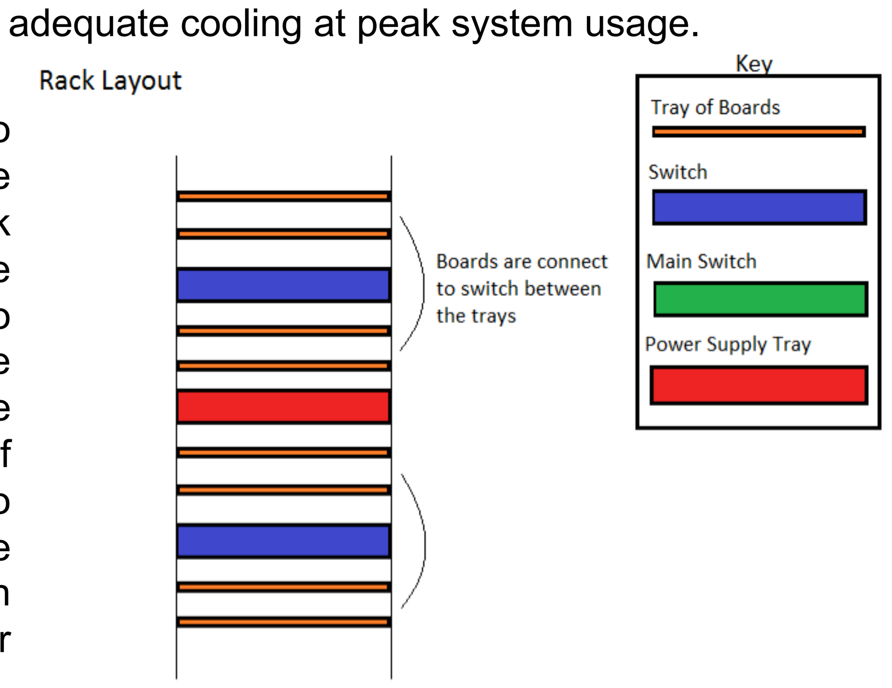
This section provides a brief overview of the system benchmarks used as well as their purported profiling purpose.

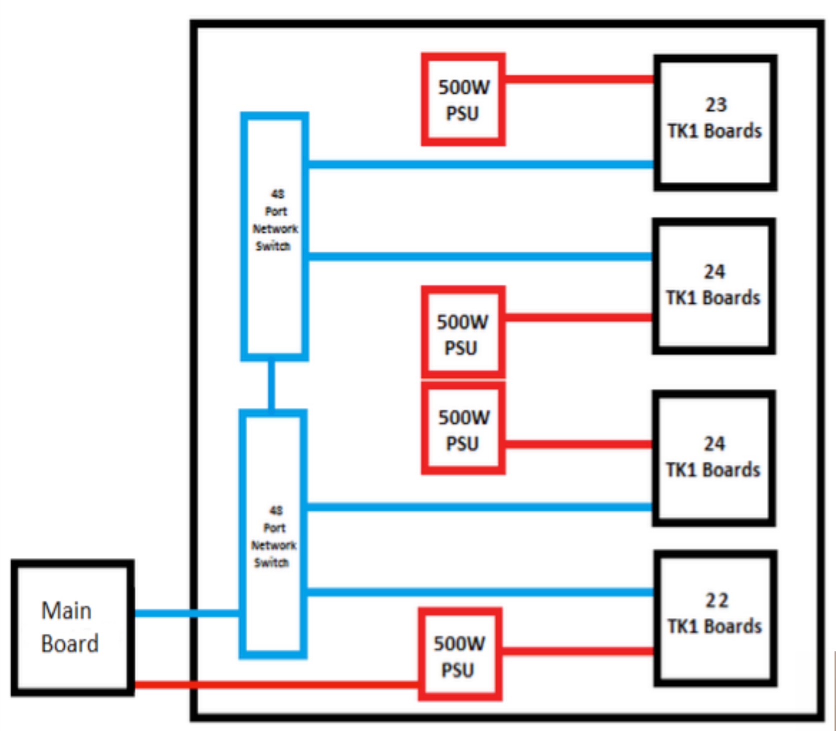
- LAMMPS (Lennard-Jones Liquid Benchmark) - The LAMMPS benchmark is a molecular dynamics simulator that reflects other classic HPC application codes. In our runs, the system is benchmarked with 256k atoms with between 1 and 72 nodes (strong scaling). Each process uses a single GPU (with 192 CUDA cores) and corresponds to a single node. A CPU only run (with multiple processes per device) was also run but was significantly slower than the GPU.
- STREAM (Sustainable Memory Bandwidth Benchmark) - This benchmark was run with OpenMP enabled and 2 million elements. For a single node run using all 4 processors we are able to get around 3500 MB/s for all operations. An OpenMPI version was also tested but resulted in a linear increase in bandwidth as expected since (no network I/O).
- LINPACK - Currently Linpack is not supported on the initial release of Jetson-Tk1 boards because of it's required 64-bit support.

Design

The design of the Astro cluster was based upon the idea of minimizing costs and power, maximizing computational performance and providing physical portability. The entire system is self contained on a single open-frame server rack that can be rolled between locations. Each custom-made metal shelf can contain twelve Jetson boards each of which is secured via brass computer spacers. The Jetson nodes along with the head node are each connected via Gigabit ethernet to one of two unmanaged switches. The head node consists of a commodity Ubuntu desktop PC containing a dual-core CPU and 4GB of RAM. The Jetson's self contained fans along with shelf spacing allow for adequate cooling at peak system usage.

Network Topology:
The network backbone of the Astro cluster is a simple bus architecture consisting of two 48 port TP-Link unmanaged ethernet switches. The unmanaged switches used to connect each of the nodes limit the possible network topologies (ie ring). Further study of the effects of network topology on neighbor to neighbor communication will be carried out after our current switch hardware is upgraded to their managed equivalents.





Power Topology
Each shelf of the Astro cluster (12 boards) is connected to one of three 500W power supplies. Keeping the total PSU power rating near the halfway point to the average power consumption allows for improved power efficiency. The maximum power consumption under load of our 72 node cluster is just under 1KW while the idle power consumption is a steady 320W. Our cluster design supports up to 96 total nodes with the addition of one more 500W PSU.

Astro Specifications:

- 72 x Jetson TK1 Development Boards (94 Supported)
- 2 x 48 Port TP-Link Ethernet Switches
- 3 x 500W Computer Power Supplies (4 supported)
- 1 Commodity PC (Dual core + 4Gig RAM) head node/ NFS server

Jetson (Node) Specifications:

- Tegra K1 SOC: Arm A15 Quad core CPU + Kepler GPU with 192 CUDA cores
- 2 GB System RAM
- 16 GB eMMC Persistent memory.
- 1 x GigE LAN Port
- Various I/O ports (HDMI, USB, eSATA)

Results

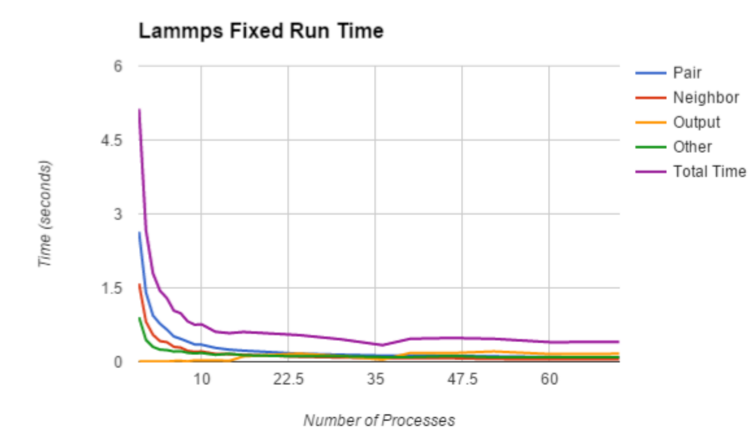


Figure One: Strong scaling Lammps run times with 256K atoms. As we increase the number of processes we see a corresponding decrease in compute time until we reach the application's Amdahl scaling limit.

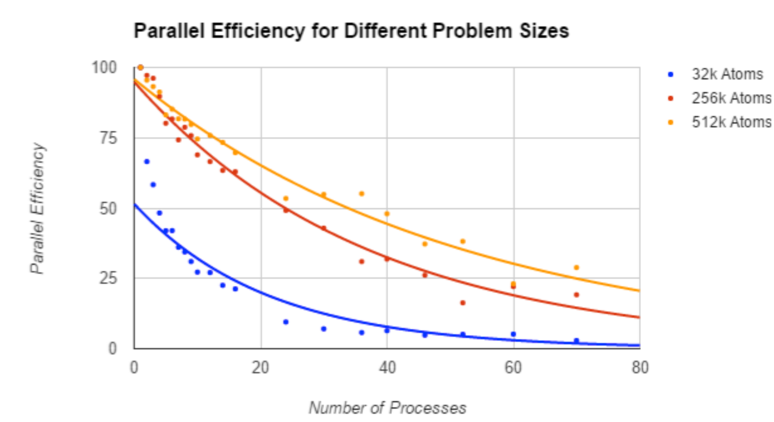


Figure Two: The Astro cluster demonstrates high parallel efficiency matching comparable systems. As we increase the problem size parallel efficiency also increases reflecting the applications scaling profile.

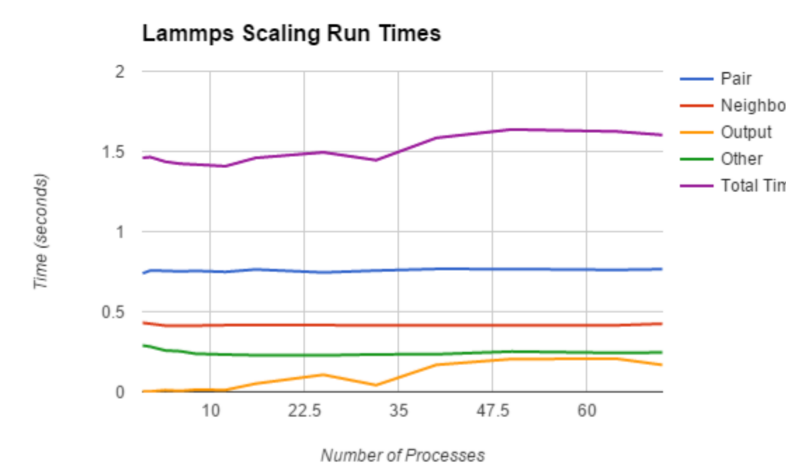


Figure Three: Lammps weak scaling, where problem size increases proportionately with the number of nodes. The system scales to the max node count of 72 devices (85% parallel efficiency). During the scaling run, 32K atoms were added for each additional node up until the maximum of 72 processes total (1 proc/node) nodes was reached.

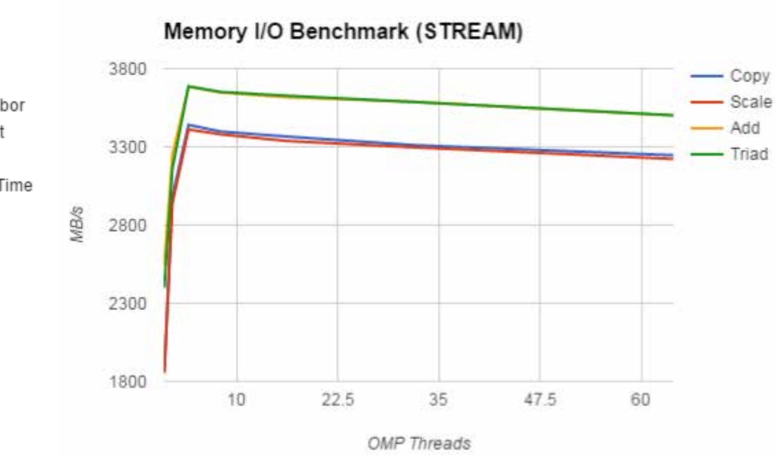


Figure Four: With 4 OMP threads the Jetson-Tk1 is able to reach around 3500 MB/s on all operations. We see the same performance when running this benchmark using the distributed version due to the lack of use of a distributed file system (DFS). NFS is used mainly for system/library updates and distribution.

Conclusions

The Tegra chipset shows great promise for use in a variety of compute intensive applications (computer vision, HPC, etc...) due to its relative low cost, lower power consumption and hybrid CPU/GPU architecture. Integration of the CPU and GPU onto a single chip eliminates the PCI Express bus bandwidth bottleneck, allowing GPU acceleration of kernel codes with limited overhead above conventional CPU multithreading methods. The Astro cluster has demonstrated, in an educational setting, the viability of both the Jetson TK1 board as a hardware accelerator and the use of commodity hybrid CPU-GPU systems for use with HPC application workloads. As we continue our performance analysis efforts we will continue to study memory access patterns of hybrid architectures as well as how network topology affects parallel efficiency in typical HPC workloads on commodity clusters.

Future Work

- **Floating Point Performance Characterization:** Our experience with the GPU enabled version of the LINPACK benchmark indicates that only 64-bit systems are supported. We are working on a conversion of NVIDIA's CUDA enabled version of LINPACK to 32-bit with zero-copy memory transfers for improved performance on hybrid CPU/GPU.
- **Network Topology Study** and replacement of unmanaged ethernet switches with managed equivalent versions to optimize for communication burdened parallel efficiency.