ECE 574 – Cluster Computing
Lecture 16

Vince Weaver

`https://web.eece.maine.edu/~vweaver`

`vincent.weaver@maine.edu`

11am, Barrows 133

19 March 2024

Announcements

- HW#7 was posted
- Don't forget project topics due Thursday (21st)!
- Midterms handed back, average was 92%
- Office hours cancelled this Monday and next due to Faculty Interviews
- Talk by Joseph Olivas from Intel on Friday (March 22nd) at 1pm in Hill Auditorium

Old Pi2 cluster

- 1 head node (16GB SD card), 24 sub-nodes. One currently seems to be down (reliability!)
- Read up on the cluster here:
  `https://www.mdpi.com/2079-9292/5/4/61/htm`

Old Pi2 Cluster Power Usage

- It's not quite a commodity cluster as it has a fairly complicated power distribution system (ATX power supply to power boards to provide measured 5V to the USB power sockets)

  A bit time consuming to wire up all the cables.
- Power distribution issues

  An ATX power supply runs best when it has a PC-like power draw

  Drawing too much 5V without a 12V low and the 5V line droops low enough that the Pis won't boot.

- Draws 90W at idle, which is 20W for ethernet switch, a few watts for fan/lights, and rest for the boards

New Pi4 Cluster

- Five Pi4 nodes (one is 8GB, others 4GB)
- Gigabit ethernet (Pi4s have PCIe and can handle)
- Only 5-nodes so manually updating IP addresses and password files
- Also haven't set up ssh-agent yet
- Set up slurm which can be a pain, especially getting a workable configuration file

New Pi4 Cluster Performance / Power Usage

- 50 GFLOPS
- Single node is 13 GFLOPS or so, so scaling reasonably
- Same performance as 10 year old macbook air
- Power over Ethernet
- 33W idle, 64W linpack, 0.863 GFLOPS/W
- Three times as fast as Pi2 cluster while using less power

MPI and slurm

- A bit hard to get working, provided script for sbatch should work
- Use “-n” to specify number of cores
- Alternate use “-N” to specify number of nodes maybe in conjunction with - -tasks-per-node
- Not sure how OpenMP and MPI interact here

MPI and Linpack

- Running MPI on your own cluster can be a pain, especially making sure it is properly running on all nodes
- Essentially it uses ssh or similar to log into each node and launch your executable. Need to have a copy in the same place on each node (using NFS or similar helps)
- At MPI init it will set things up and communicate between them using sockets. (Details???0
- If running Linpack, unlike previous times you will need to mess with P and Q settings in HPL.dat to get it to

run on more than one node.

Why use slurm?

- Can set account to charge
- Can handle checkpointing
- Can set constraints (run on machine with gpu, certain proc type)
- Contiguous allocations
- CPU freq, power capping
- Licenses avail (things like Matlab etc)
- Memory avail

Homework #7 Note

- Didn't really give advice on Combine
- You need to make sure the ranks have all the data they need for combine. You might have to re-broadcast it back, or do a GatherAll that does that in one step.
- Alternately if you keep the same split up of work you can do sobelx and sobely and combine all on the ranks and only gather once at end

Accelerators

What if CPU power isn't enough?

- We've been mostly looking at ways to get the most performance out of CPUs
- What else is there?

Accelerator Options – ASIC

- hard-coded custom hardware for acceleration
- quite possibly the fastest, as custom made for your workload
- expensive to make, as one-off
- need to hire ASIC designers and get things fabbed
- found in BitCoin mining?

Accelerator Options – FPGA

- Reprogramable logic
- can have fast in-hardware designs but can re-program when workload changes
- Need to have someone who can write FPGA code
- There has been work for having OpenMP and such be able to handle FPGAs

Accelerator Options – DSP

- Digital Signal Processors
- Can be good at certain workloads
- Some supercomputers have had them

Google Tensor Processing Unit (TPU)

- For accelerating machine learning tasks
  - TPUs good at CNN (convolutional neural networks)
  - GPUs good at fully connected
  - CPUs good at RNN (recurrent)
- ISCA paper – In Datacenter Performance Analysis of a Tensor Processing Unit
- For high-volume low-prevision FP calculations (8 and 16-bit)
- Unlike GPU has no rasterizer or texture processor

- Some recent NVIDIA GPUs have tensor units

Accelerator Options – Cell Processor (Obsolete)

- Special IBM Power core that had many smaller helper cores
- Could be really fast if programmed well, hard to program
- In end people found it not worth the extra effort
- Was also in PlayStation 3
- Some groups would buy them up and make fast clusters with them. This annoyed Sony who eventually dropped Linux support

Accelerator Options – Xeon Phi (Obsolete)

- Intel, came out of the larabee design (effort to do a GPU powered by x86 chips)
- Large array of x86 chips(p5 class on older models, atom on newer) on PCIe card.
- Sort of like an internal mini cluster
- Runs Linux, can ssh into the boards over PCIe.
- Benefit: can use existing x86 programming tools and knowledge.
- Intel cancelled this

Graphics and Video Cards / History

Old CRT Days

- Electron gun
- Horizontal Blank, Vertical Blank

LCD Displays (sic)

- Crystals twist in presence of electric field
- Asymmetric on/off times
- Passive (crossing wires) vs Active (Transistor at each pixel)
- Passive have to be refreshed constantly
- Use only 10% of power of equivalent CRT
- Circuitry inside to scale image and other post-processing
- Need to be refreshed periodically to keep their image
- New "bistable" display under development, requires no

power to hold state

Coding for CRTs

- Atari 2600 – only enough RAM to do one scanline at a time
- Apple II – video on alternate cycles, refresh RAM for free
- Bandwidth key issue. SNES / NES, tiles. Double buffering vs only updating during refresh
- Multibanks of graphics (VGA and older) another way to deal with lack of bandwidth

Old 2D Video Cards

- Framebuffer (possibly multi-plane), Palette
- Dual-ported RAM, RAMDAC (Digital-Analog Converter)
- Interface (on PC) various io ports and a 64kB RAM window
- Mode 13h
- Acceleration – often commands for drawing lines, rectangles, blitting sprites, mouse cursors, video overlay

Old 3D Video Cards

- At first only in high-end workstations (like SGI)
- 3dfx cards, with passthrough cable
- Became more mainstream

Modern Graphics Cards

- Essentially high-end linear algebra / 3D rendering supercomputers
- Can draw a lot of power
- 2D (optional afterthought these days)
- Can contain other hardware accelerators (such as Video decoders)

Interface – Integrated vs Standalone

- Integrated
  - Built into motherboard/chipset/processor
  - Can share memory (and bandwidth) with CPU
  - Traditionally less capable, but that is changing
- Standalone
  - Usually in PCIe slot, bandwidth constrained
  - Can draw lots of power
  - Can have multiple

Video RAM

- VRAM (old) – dual ported. Could read out full 1024Bit line and latch for drawing, previously most would be discarded (cache line read)
- GDDR3/4/5 – traditional one-port RAM. More overhead, but things are fast enough these days it is worth it.
- Confusing naming, GDDR3 is equivalent of DDR2 but with some speed optimization and lower voltage (so higher frequency)

Busses

- DDC – i2c bus connection to monitor, giving screen size, timing info, etc.
- PCIe (PCI-Express) – most common bus in x86 systems
  Original PCI and PCI-X was 32/64-bit parallel bus
  PCIe is a serial bus, sends packets
  Can power 25W, additional power connectors to supply can have 75W, 150W and more
  Can transfer 8GT/s (giga-transfers) a second
  In general PCIe is limiting factor to getting data to GPU.

Connectors

CRTC (CRT Controller) Can point to same part of memory (mirror) or different.

- RCA – composite/analog TV
- VGA – 15 pin, analog
- DVI – digital and/or analog. DVI-D, DVD-I, DVD-A
- HDMI – compatible with DVI (though content restrictions). Also audio. HDMI 1.0 – 165MHz, 1080p

or 1920x1200 at 60Hz. TMDS differential signaling. Packets. Audio sent during blanking.

- Display Port – similar but not the same as HDMI
- Thunderbolt – combines PCIe and DisplayPort. Intel/Apple. Originally optical, but also Copper. Can send 10W of power.
- LVDS – Low Voltage Differential Signaling – used to connect laptop LCD

Interfaces for 3D Graphics

- OpenGL – SGI (Khronos)
- DirectX – Microsoft (Direct3d)
- Vulkan (sort of next gen OpenGL. Lower level, closer to hardware)
- Metal – from Apple
- WebGL – javascript/web
- OpenGL ES – embedded subset