ECE 574 – Cluster Computing
Lecture 2

Vince Weaver

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

vincent.weaver@maine.edu

23 January 2025

Announcements

- HW#1 was due!
- A break on homeworks until next week.

Some Supercomputing (SC24) notes

- November 2024, in Atlanta Georgia (show pictures)
- Large academic / industry conference, 17,500 people
- Competitive conference, only 21% paper acceptance rate
- Even workshops and poster sessions reject submissions
  - Attended Security workshop. Only recently have people started about caring about security in computer systems.
  - ProTools workshop, Willow and I had a paper there on Hybrid CPU performance measurement with PAPI

- Large vendor showfloor with lots of companies, hardware and software vendors, universities and research centers
  - SCInet infrastructure, 8.14TB/s
  - Student cluster competition
  - Lots of CLX RAM, Quantum, cooling/rack solutions
- Keynote. NASA talk.
- Attended various talks
  - Quantum Computing BoF

    There is hardware? Not sure how to use it? Different hardware (lots of N2 tanks?) TUM (Reinaldo). Schulzs Quantum Valley. Good at brute force problems, traveling salesman, prime factoring. How to integrate with SC setup. Do you need Quantum Physicists as Sysadmins?

- Talk on detecting inappropriate use of HPC resources How do you stop people running bitcoin miners on your hardware. Talk to trying to watch which programs are run and stop if signature doesn't match typical HPC programs. Many issues with this including eventually people can modify code to look more similar
- IPv6 in HPC. Use in clusters, things like Lustre, RDMA
- RISC-V accelerators. Coming in a few years?

Dave Ditzel (Transmeta) was there with fancy RISC-V accelerator plans but not shipping until 2027? People concerned about FORTRAN support

- ARM Firmware... everything chiplets these days
- Green500, didn't realize it was so hard to get measurements, most still aren't full system, many just 20% and extrapolating

  Level1/Level2/Level3 reporting rules. Most of top computers in Europe. 72 GFLOP/W to hit top (note: m1 laptop 6 GFLOP/w, Pi5 3.6 GFLOP/W, haswell-ep 2.1 GFLOP/W)

Top500 List Notes

- Rmax vs Rpeak – Rmax is max measured, Rpeak is theoretical best
- HPL Linpack
  - Embarrassingly parallel linear algebra
  - Solves a (random) dense linear system in double precision (64 bits) arithmetic
- HP Conjugate gradient benchmark
  - More realistic? Does more memory access, more I/O bound.

  - #1 on HPCG usually not same as #1 HPL
- Green 500

Example Top500 Listing

- From the November 2002 list, entry #332
- Location: Orono, ME
- Proc Arch: x86
- Proc Type: Pentium III, 1GHz
- Total cores: 416
- RMax/RPeak: 225/416 GFLOPS
- Power: ???
- Accelerators: None

UMaine Supercomputer Details

- Located at Target Tech Center (Orono Business Park)
- 208 desktop PIIIs, 100Mb eth admin, 1G Myrinet
- Originally single socket. With that they got #501 on list (briefly on before getting kicked off)
- Populated rest of the sockets, made the list

My Lab's top Computer List

- https://web.eece.maine.edu/~vweaver/group/machines.html
- Haswell-EP (we'll use for homeworks)
  436 GFLOPs, 2.3GFLOP/W, would have been #1 in 1996
- Raspberry Pi 4 – 13 GFLOPs, 2.0 GFLOP/W, #10 in 1993

Top500 List – November 2023

#	Name	Country	Arch /Proc	Cores	PFLOPS Max/P	Accel	Power
1	El Capitan	US/LLNL	AMD EPYC4	11,039,616	1.7k/2.7k	AMD Instinct	30MW
2	Frontier	US/ORNL	AMD EPYC	9,066,176	1.4k/2.1k	AMD Instinct	25MW
3	Aurora	US/Argonne	Intel SPR	9,264,128	1.0/2.0k	Intel MAX	39MW
4	Eagle	US/Microsoft	Intel SPR	2,073,600	561/846	NVD H100	?MW
5	HPC6	Italy	AMD EPYC3	3,143,520	478/607	AMD Instinct	8MW
6	SC Fugaku	Japan/Riken	ARM64	7,630,848	442/537	N/A	30MW
7	Alps	Switzerland	NVD Grade	2,121,600	435/575	NV GH200	7MW
8	LUMI	Finland	AMD EPYC3	2,752,704	380/531	AMD Instinct	7MW
9	Leonardo	Italy	Intel ICL	1,824,768	239/304	NVD A100	7MW
10	Tuolumne	US/LLNL	AMD EPYC4	1,161,216	208/289	AMD Instinct	3MW
11	MareNostrum5	Spain/BSC	Intel SPR	663,040	175/249	NVD H100	4.2MW
12	EOS SuperPodDGX	US/NVIDIA	Intel SPR	485,888	121/189	NVD H100	?MW
13	Venado	US/LANL	NVD Grace	481,440	99/130	NVD GH200	1.7MW
14	Sierra (IBM)	US/LLNL	IBM Power9	1,572,480	94/125	NVD Volta	7MW
15	Sunway TaihuLight	China	Sunway	10,649,600	93/125	?	15.3MW
16	CHIE-3	Japan	Intel SPR	297,840	92/138	NVD H100	?
17	CHIE-2	Japan	Intel SPR	297,840	90/138	NVD H100	?
18	JETI	Germany	NVD Grace	391,680	83/94	NVD GH200	1.3MW
19	Perlmutter	US/LBNL	AMD EPYC	888,832	79/113	NVD A100	2.9MW
20	El Dorado	US/Sandia	AMD EPYC4	383,040	68/95	AMD Instinct	1.1MW
21	Gefion	Denmark	Intel SPR	223,088	67/101	NVD H100	?
22	CEA-EA	Franch	NVD Grace	389,232	64/103	NVD GH200	1.2MW
23	Selene	US/NVIDIA	AMD EPYC	555,520	63/79	NVD A100	2.6MW
24	Tianhe-2A	China	Intel IVB	4,981,760	61/101	MatrixDSP	18.5MW

Top supercomputer Architecture

- Commodity or custom
- Architecture: x86? SPARC? Power? ARM embedded vs high-speed?
- Memory (how much? 1GB per core?)
- Accelerator cards / Heterogeneous Systems

Top supercomputer Storage

- How much?
- Large hadron collider one petabyte of data every day
- Shared? If each node wants same data, do you need to replicate it, have a network filesystem, copy it around with jobs, etc?
- Cluster filesystems?

Top supercomputer – Reliability

- Reliability. How long can it stay up without crashing?
  Can you checkpoint/restart jobs?
  Sequoia MTBF 1 day.
  Blue Waters 2 nodes failure per day.
  Titan MTBF less than 1 day

Top supercomputer – Power/Cooling

- Cost of Power over lifetime can be more than that of hardware cost
- Power comparison: small town? 1MW around 200 - 1000 homes? (this varies)
- Does Power include cooling
- Big river nearby?

Top supercomputer – Network

- How fast? Latency?
- Interconnect? (torus, cube, hypercube, etc)
- Ethernet? InfiniBand? Custom?

Top supercomputer – Software

- Operating System
  - Linux? Custom?
  - If just doing FP, do you need overhead of an OS?
- Job submission software, Authentication
- Tools – software that can help you find performance problems

Top supercomputer – Applications

- Software – how to program?
- Too hard to program can doom you.
- A lot of interest in the Cell processor. Great performance if programmed well, but hard to do.

Top supercomputer – Running Linpack

- How long does it take to run LINPACK?
- How much money does it cost to run LINPACK?
- Is it worth it?

Notes on the Top500 BoF Video

- What is a BOF anyway?
- Why is Exaflop a big deal?

Notes on the Top500 BoF Video 2024

- I was at SC, but missed the BoF as scheduled against poster session
- Jack Dongarra, Turing award winner, previous boss of mine
- Bronis who received El Capitan award also in theory former boss of mine
- I did attend the Green500 award ceremony

El Capitan (2024)

- Linpack under 30MW (20MW/Exaflop, which was goal)
- Slingshot interconnect
- Tri-lab OS (Redhat based)
- Near-node local storage “rabbits”
- Zen4 and AMD CNDA3, same package, share same RAM space
- Chiplets
- 256MB Last-level Cache
- Smaller system, still #10 on List, faster than previous

LLNL system

- Even smaller one at LANL #20, one rack system is #49

DGEMM with Tensor Cores (2024)

- Using AI hardware to do HPC (subsidized? like gaming was before?)
- Recovering 64bit precision from FP16
- If using FP32 no benefit going Hopper to Blackwell

Top500 Summary (2024)

- HPCG, no changes (summit discontinued)
- HPL-MxP
- Green500 – lots from Europe
- GFLOP/W growing exponentially
- Hyperscalers – gigawatt sized datacenter?
- Replacement rate low (46)
- China stopped giving numbers

Top500 BoF Video – Aurora (2023?)

- Aside, long history, supposed to be 200 PFLOPS in 2018 with lots of Xeon Phis
- Intel Sapphire Rapids, Intel having problems with chips
- Intel Xeon MAX (Ponte Vecchio) GPU, had to turn to TSMC to get parts made

  Roughly 2x A100 perf, not as fast as H100
- Only part of it running, goal is 2 Exaflops (mostly from GPUs)
- 160 Racks, 10,624 Nodes, 21,248 CPUs, 63,744 GPUs

- HPE Slingshot-11 interconnect (formerly Cray)

  `https://www.nextplatform.com/2022/01/31/crays-slingshot-interconnect-is-at-the-heart-of-hpes-hpc-`

  a lot of history on interconnects
- Dragonfly Topology?
- 10.9 PB of DDR5 RAM (512 GB/CPU?)
- 1.36 PB of CPU HBM
- 8.16 PB GPU HBM (100GB/GPU?)
- Storage 230 PB

Top500 BoF Video – MS Eagle (2023?)

- Azure HPC
- 14,400 H100 GPUs
- InfiniBand Quantum 7
- 1800 servers
- Any customer can use
- Ubuntu
- Generative AI

Top500 BoF Video – General

- HPCG results different as usual. Frontier only 16 PFLOPS
- HPL-MXP – Piotr working on. Mixed precision. Estimates can get 10x performance

  Complicated coding. Worth it? Old days would just wait, Moore's Law.

  Now maybe it is
- Green 500. Henri, 65 GFlops/W
- Less churn, computers on list for longer time and used

longer before getting rid of

- Unlikely to hit 10 Exaflops by end of decade

Notes on the Top500 BoF Video – 2022

- Frontier
  - 74 racks, 9.2PB RAM (half HB, half DDR4)
  - 90 miles of network cable
  - $600 million
  - Quiet, water cooled. Warm water (32C)
  - Trouble getting more than 600PFLOPs, turned out to be linear-time thing in Cray MPI library
  - 3 hours to run Linpack, nodes keep failing when try to do run

- Non-linpack results, HPGC Frontier #2 only 14 PFLOPs
- Green 500, GFLOPs/W. Frontier much lower. Top was machine with first NVIDIA H100 (Hopper)
- Systems appearing more slowly on list, aging more before dropping off