ECE 574 – Cluster Computing
Lecture 2

Vince Weaver

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

vincent.weaver@maine.edu

11am, Barrows 133

18 January 2024

Announcements

- HW#1 was due!
- A break on homeworks until next week.
- Missing from last lecture:
  - Office Hours, Barrows 203, 11am-noon Mon/Wed

Other stuff

- 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

Historical Note

- 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

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	Frontier	US/ORNL	AMD EPYC	8,699,904	1.2k/1.7k	AMD Instinct	23MW
2	Aurora	US/Argonne	Intel SPR	4,742,808	585/1.0k	Intel MAX	25MW
3	Eagle	US/Microsoft	Intel SPR	1,123,200	561/846	NVD H100	?MW
4	SC Fugaku	Japan/Riken	ARM64	7,630,848	442/537	N/A	30MW
5	LUMI	Finland	AMD EPYC	2,752,704	380/531	AMD Instinct	7MW
6	Leonardo	Italy	Intel ICL	1,824,768	239/304	NVD A100	7MW
7	Summit (IBM)	US/ORNL	IBM Power9	2,414,592	148/200	NVD Volta	10MW
8	MareNostrum5	Spain/BSC	Intel SPR	680,960	138/266	NVD H100	2.6MW
9	EOS SuperPodDGX	US/NVIDIA	Intel SPR	485,888	121/189	NVD H100	?MW
10	Sierra (IBM)	US/LLNL	IBM Power9	1,572,480	94/125	NVD Volta	7MW
11	Sunway TaihuLight	China	Sunway	10,649,600	93/125	?	15.3MW
12	Perlmutter	US/LBNL	AMD EPYC	888,832	79/113	NVD A100	2.9MW
13	Selene	US/NVIDIA	AMD EPYC	555,520	63/79	NVD A100	2.6MW
14	Tianhe-2A	China	Intel IVB	4,981,760	61/101	MatrixDSP	18.5MW
15	Explorer-WUS3	US/Microsoft	AMD EPYC	445,440	54/87	AMD Instinct	?MW
16	ISEG	Netherlands	Intel SPR	218,880	47/87	NVD H100	1.3MW
17	Adastra	France	AMD EPYC	319,072	46/61	AMD Instinct	0.9MW
18	JUWLES Booster	Germany	AMD EPYC	449,280	44/71	NVD A100	1.7MW
19	MareNostrum	Spain/BSC	Intel	725,760	40/46	?	5.7MW
20	Shaheen III	Saudi Arabia	AMD EPYC	877,824	36/40	?	5.3MW
21	HPC5	Italy	Intel CSL	669,760	35/51	NVD Tesla	2.25MW
22	Sejong	South Korea	AMD EPYC	277,760	33/41	NVD A100	?MW
23	Voyager-EUS2	US/Microsoft	AMD EPYC	253,440	30/40	A100	?MW
24	Crossroads	US/LANL/SNL	Intel	660,800	30/40	?	6.2MW
25	Setonix GPU	Australia	AMD EPYC	182,248	27/35	AMD Instinct	0.5MW

What goes into a top supercomputer?

- Commodity or custom
- Architecture: x86? SPARC? Power? ARM
  embedded vs high-speed?
- Memory
- 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?

- 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
- Power / Cooling
  Big river nearby?
- Accelerator cards / Heterogeneous Systems
- Network
  How fast? Latency? Interconnect? (torus, cube,

hypercube, etc)
Ethernet? Infiniband? Custom?

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

problems

- Left off my summary: RAM? Interconnect?
- Operating system? Mostly Linux these days
- Power: does this include cooling or not?
  Cost of power over lifetime of use is often higher than the cost to build it.
- Power comparison: small town? 1MW around 1000 homes? (this varies)
- How long does it take to run LINPACK? How much money does it cost to run LINPACK?
- Cost to run computer more than cost to build it?

Notes on the Top500 BoF Video – 2023

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

Top500 BoF Video – Aurora

- 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

- 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 Church, 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

Introduction to Performance Analysis

What is Performance?

- Getting results as quickly as possible?
- Getting *correct* results as quickly as possible?
- What about Budget?
- What about Development Time?
- What about Hardware Usage?
- What about Power Consumption?

Motivation for HPC Optimization

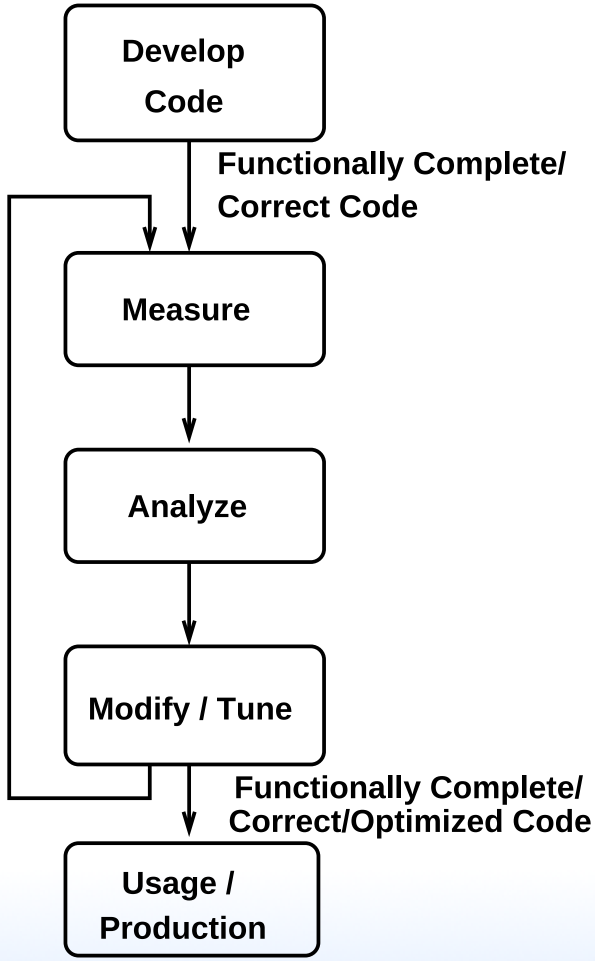
HPC environments are expensive:

- Procurement costs: $\sim\$40$ million
- Operational costs: $\sim\$5$ million/year
- Electricity costs: 1 MW / year $\sim\$1$ million
- Air Conditioning costs: ??

Know Your Limitation

- CPU Constrained
- Memory Constrained (Memory Wall)
- I/O Constrained
- Thermal Constrained
- Energy Constrained

Performance Optimization Cycle



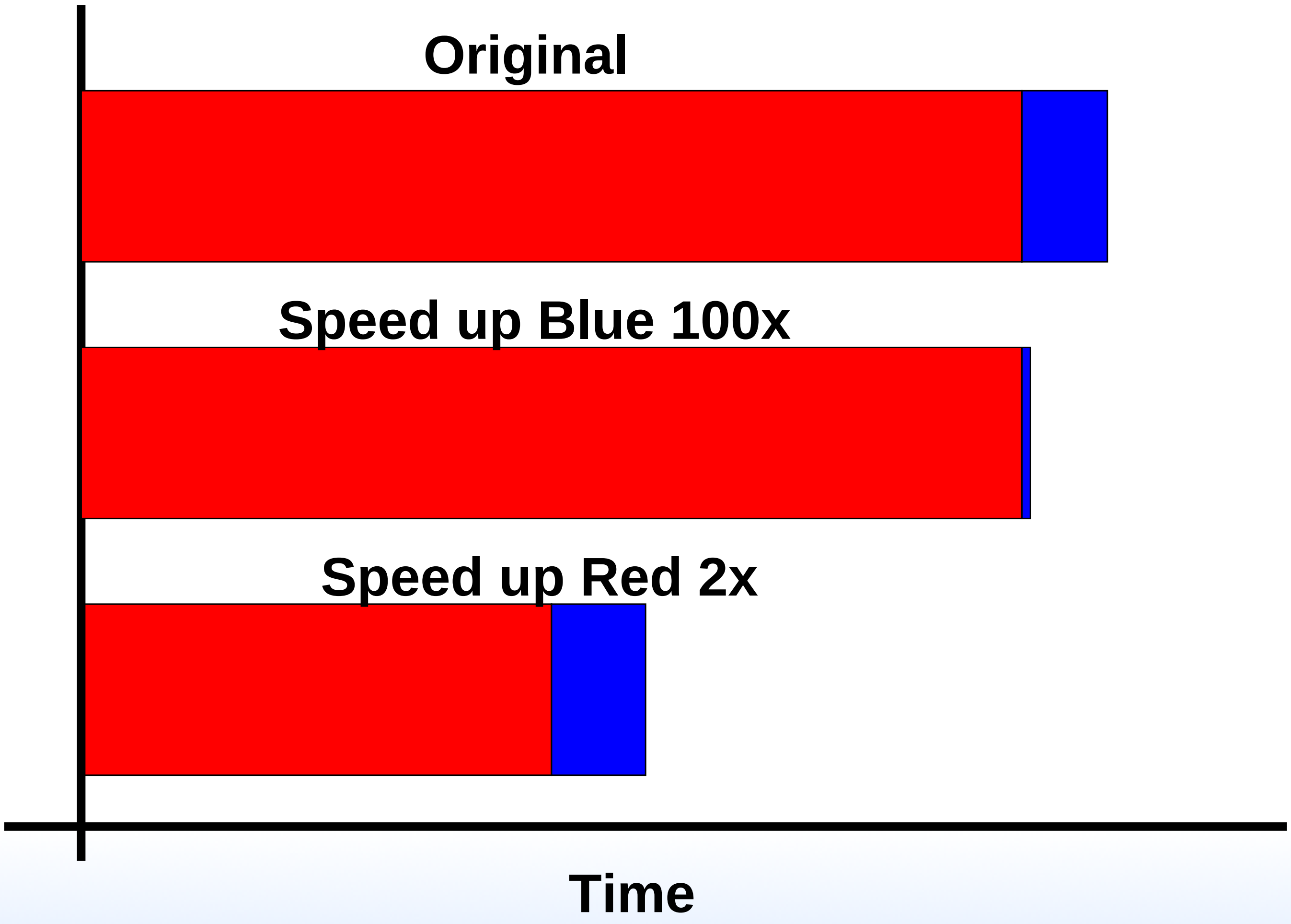
Wisdom from Knuth

"We should forget about small efficiencies, say about 97% of the time:

premature optimization is the root of all evil.

Yet we should not pass up our opportunities in that critical 3%. A good programmer will not be lulled into complacency by such reasoning, he will be wise to look carefully at the critical code; but only after that code has been identified" — Donald Knuth

Amdahl's Law



Speedup

- Speedup is the improvement in latency (time to run)

$$S = \frac{t_{old}}{t_{new}}$$

So if originally took 10s, new took 5s, then speedup=2.

Scalability

- How a workload behaves as more processors are added

- Parallel efficiency: $E_p = \frac{S_p}{p} = \frac{T_s}{pT_p}$
  p=number of processes (threads)
  T_s is execution time of serial code
  T_p is execution time with p processes

- Linear scaling, ideal: $S_p = p$

- Super-linear scaling – possible but unusual

Strong vs Weak Scaling

- Strong Scaling –for fixed program size, how does adding more processors help
- Weak Scaling – how does adding processors help with the same per-processor workload

Strong Scaling

- Have a problem of a certain size, want it to get done faster.
- Ideally with problem size N, with 2 cores it runs twice as fast as with 1 core (linear speedup)
- Even if not ideal linear scaling, if there's any speedup then some strong scaling is happening
- Often processor bound; adding more processing helps, as communication doesn't dominate

- Hard to achieve for large number of nodes, as many algorithms communication costs get larger the more nodes involved

- Amdahl's Law limits things, as more cores don't help serial code

- Strong scaling efficiency: t1 / (N * tN) * 100%

- Improve by throwing CPUs at the problem.

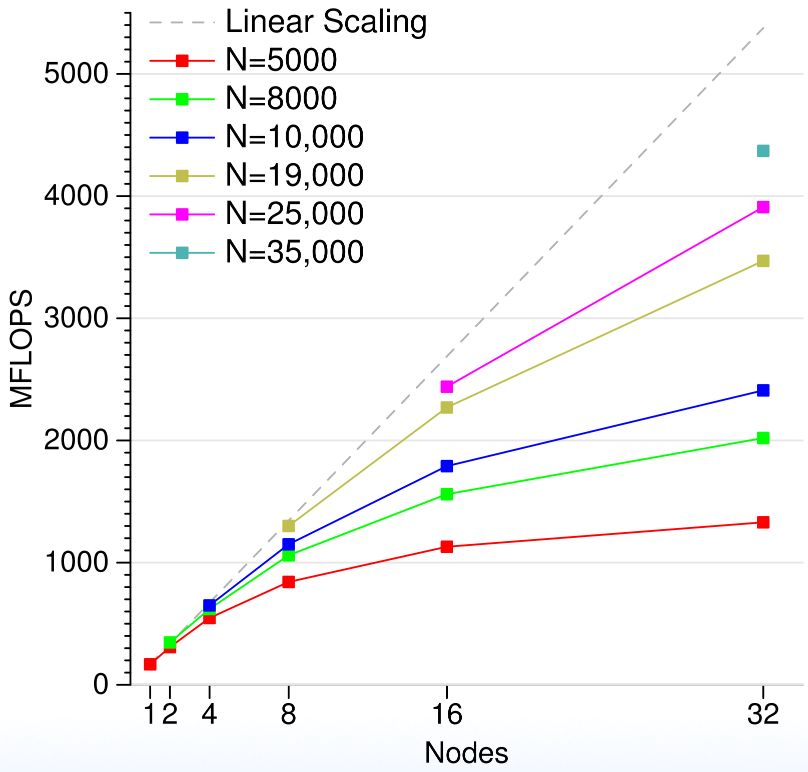
Weak Scaling

- Have a problem, want to increase problem size without slowing down.
- Ideally with problem size N with 1 core, a problem of size 2*n just as fast with 2 cores.
- Often memory or communication bound.
- Gustafson's Law (rough paraphrase)
  No matter how much you parallelize your code, there will be serial sections that just can't be made parallel

- Weak scaling efficiency: (t1 / tN) * 100%
- Improve by adding memory, or improving communication?

Scaling Example

LINPACK on Rasp-pi cluster. What kind of scaling is here?



Weak scaling. To get linear speedup need to increase problem size.

If it were strong scaling, the individual colored lines would increase rather than dropping off.