

ECE 574 – Cluster Computing

Lecture 17

Vince Weaver

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

vincent.weaver@maine.edu

14 October 2026

Announcements

- Midterm Monday, October 19th, during class
- If you need any special accommodations let me (or the accessibility office) know sooner rather than later
- Don't forget HW#6
- Reminder on Academic Honesty
Remember: using AI for assignments not allowed

Midterm on 19 October 2026

- Can bring one page (8.5" by 11" one sided) of notes. Otherwise closed notes. No computers, cell-phones, Beowulf cluster, etc.
- There is a speedup question which requires some math. Should be able to do it w/o a calculator, but you can have one if you must.

Midterm Topics

- Performance / Definitions
 - Speedup, Parallel efficiency
 - Strong and Weak scaling
 - Distributed vs Shared Memory Systems
- Computer Architecture Review
 - Know that memory bandwidth (and caches) can affect performance
For example, changing loop order
 - Know hardware limitations like number of cores and hardware (hyper) threads can limit scaling
 - Not going to ask low-level architecture questions

More Midterm Topics

- Pthread Programming
 - Know about race condition, deadlock
 - Know roughly the layout of a pthreads program.
(define pthread_t thread structures,
pthread_create, pthread_join)
 - Know why you'd use a mutex.
- OpenMP Programming
 - parallel directive
 - scope (private vs shared)
 - section
 - for directive

More Midterm Topics

- Limited to no SIMD questions, but be aware of the concept
- No MPI material

Shared Memory vs Distributed Systems

- So far we've mostly talked about shared memory systems

Shared Memory Systems

- Can have many cores, but has only one copy of Operating System running
- All programs see one unified memory space
- Parallel code uses threads (pthreads, OpenMP)
- Threads can communicate simply by writing to memory

Shared Memory Limitations

- OpenMP is nice to use. But what if your problem won't fit on a single machine?
- How big can a shared-memory machine be?

Niche Large Shared Memory Systems

- SGI Altix/UV systems at least 4096 cores and 16TB running one Linux image

<http://www.techeye.net/hardware-2/>

`sgi-builds-pittsburgh-4096-processor-core-16tb-shared-memory-supercomputer`

- Digression about SGI (this was 2006-2012 or so)
- Use special NUMA-Linux architecture to spread cache coherence across multiple machines.
- Origin TM and Onyx2 TM Theory of Operations Manual
- I have part of an Altix system that some students rescued

More Recent Large Shared-Memory Systems

- Intel Granite Rapids servers can have 144 cores (288 threads) per socket, with two sockets, so 576 threads
- AMD Zen5c can have 192 cores (384 threads) by two sockets so 768 threads
- AmpereOne Aurora: 512 arm64 cores

Linux Scaling Limitations

- Linux currently(?) maxes out to 4096 cores or so
- Somewhat dated “Scaling Linux to the Extreme” paper problems: cache contention could bring machine to halt (if a global idle counter, each thread trying to increment once a second)
lock contention, page cache
- What are the challenges? Locking contention?

Eventually you hit the limit

What's the alternative?

Moving to a distributed system

Distributed System

- Many systems each with own memory, communicate via Message passing
- Communicate over a network
- Each node has own copy of Operating System
- Need to explicitly manage moving data between nodes (MPI)

Could you Open Code by Hand?

- Sort of how you can use pthread directly?
 - Need to have copy of executable on each node (network filesystem helps)
 - Need to launch executable on remote system (ssh can do this) authentication is a challenge
 - Then write custom network code to open sockets and communicate among them all
- Network code is a pain
- Just crying out for abstraction

Networks – Physical Layer

- Copper Wire vs Fiber Optics
- Which is faster? Which is lower latency? Which is cheaper?
- Which can go further distance? Which can bend around corners?
- Which is lower power?
- Latency vs Bandwidth
- Speed of Light in each? (0.6 - 0.85c for copper at RF frequencies, 0.7c for fiber)

Networks – Copper

- Usually Twisted Pair (avoids noise)
- The faster you go you might need fancy connectors
- Trouble with high speeds, transmission line effects, eventually hit microwave frequencies

Networks – Fiber

- Single-mode
 - More expensive?
 - Only one frequency?
- Multi-mode
 - Shorter distance

Networks – Other

- Can you just shoot lasers around the room?

Network Topology

- Packet-switching vs bus
- Ring, mesh, star, line, tree, fat-tree fully connected
- Cube, hypercube
- Mesh networks and routing
- Routing. Fully connected? Crossbar?

For comparison – Top 500 Nov 2022

This is constantly changing, hard to keep up year to year

interconnect	#
Infiniband NDR	139
Infiniband HDR	112
100GB Ethernet	50
Infiniband EDR	27
25GB Ethernet	15
10GB Ethernet	11
400GB Ethernet	7
Infiniband Unknown	7
1GB Ethernet	3
NVD Spectrum-X	3
Infiniband XDR	1
LinQi	1
HPE Slingshot	?
...	-

For comparison – Top 500 Nov 2022

interconnect	#
100GB Ethernet	70
25GB Ethernet	70
10GB Ethernet	62
Infiniband EDR	35
Infiniband HDR	35
Intel Omnipath	34
Mellanox HDR Infiniband	26
Aries	25
Infiniband HDR 100	20
Slingshot-10	17
...	-

Network Types – Ethernet

- Low-end (10/100/1GB/2.5GB/10GB) commodity
- Regular computers support it
- Cat-5/Cat-6 cables relatively cheap?
- How many ports on a switch? What do you do when you need more?
- The faster schemes are complex, use a lot of power

Network Types – High-end Ethernet

- 25GB
 - Introduced as 10GB not fast enough, but 40GB expensive
 - SFP28 transceivers, both optical and copper
- 40GB
 - 802.3ba – 2010 – 1m backplane, 100m multi-mode fiber, 10km single-mode fiber
 - 802.3bg –
 - 802.3bq – 2016 – 4-pair balanced twisted pair 30m
 - QSFP+ transceivers, quad small form-factor pluggable, four 10GB lanes

Network Types – High-end Ethernet

- 100GB
 - 4 lanes of 25GB
- Anything above 100GB considered part of Terrabit (TbE)
- 400GB
- 800GB

Network Types – Infiniband

- low latency, most common in supercomputers
- Often a step ahead of ethernet (but more expensive)
- QSFP+ 3m 40Gb/s cable \$30?
- copper or fiber, GB/s (links aggregated, 4x common)

	SDR	DDR	QDR	FDR-10	FDR	EDR	HDR	NDR	XDR	GDR
4x	8	16	32	40	54	100	200	400	800	1600
12x	24	48	96	120	163	300	600	1200	2400	4800

- Note XDR/GDR just proposed (2023)
- RDMA, Ethernet over Infiniband

Network Types – LinQi

- In fastest Computer

Network Types – NVIDIA Spectrum-X

- Up to 800Gbps
- “AI Optimized” (???)

Network Types – HPE Slingshot

- HPE Cray Supercomputers
- Up to 200 Gb/s
- Handles lots of small packets
- 1.2 billion packets/second/port
- High Radix(?) 64-port 12.8TB/s switch
- Scale to 250k ports with max of 3 hops
- Price: “get quote” which means, a lot

Network Types – Fugaku Paper

- Hooking together 10k+ nodes?
- Mesh or 3d-torus?
- Blue-Genel L partition things so jobs run on subset partitions that are also 3d-torus. Issues if a node goes down
- Cray XT mesh and job can be scattered throughout, but an have higher latency
- Instead, 6-dimensional torus
- Tofu interconnect for K computer (torus fusion)

Network Types – Older/Other

- Cray Gemini – Mesh/torus – 64Gb/s
- Fibrechannel
- Older: custom, Myrinet

Remote DMA

- Zero copy of data from network card directly into memory of remote system (without CPU involvement)
- Can lead to really fast MPI transactions
- Avoids overhead of the OS TCP/IP stack
- Security?
- Infiniband can do it. Also RoCE (RDMA over Converged Ethernet) `https:`

`//www.fs.com/blog/rdma-over-converged-ethernet-guide-2208.html`