# ECE 574 – Cluster Computing
# Lecture 24

Vince Weaver

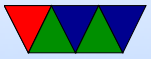
http://www.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

3 December 2015

# Announcements

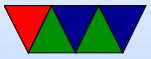
- Project presentations next week. If you haven't already, let me know if you are willing to present on Tues vs Thurs. Looking for one more group so it will be 4+4.
- One more (really brief) homework. Maybe.



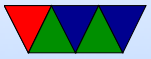
# Final

- Thursday the 17th
- Like the midterm. Cumulative, but focusing on more recent material.
- One page of notes again if it helps
- Topics

1. Performance (strong/weak scaling, parallel efficiency)
2. Pthread/OpenMP/MPI tradeoffs

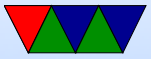


3. CUDA
4. Big Data
5. Fault Tolerance
6. Power/Energy



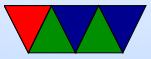
# New 100 Petaflop Supercomputers

- "Providing a Robust Tools Landscape for CORAL Machines" by Michael J. Brim, Dong H. Ahn, Scott Parker, and Gregory Watson
  Extreme Scale Programming Tools workshop at SC15



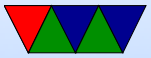
# CORAL

- CORAL = Collaboration of ORNL, ANL, LLNL
- "Leadership" computers to be at Argonne, Oak Ridge, and LLNL in 2017. Current leaders are Titan (ORNL), Sequoia (LLNL) and Mira (ANL)
- Request was greater than 100PFLOPS, 2GB/core, local NVRAM, performance 4-8x better than existing
- Three computers ordered, but at least one must be different than others for risk reduction



# Summit (ORNL) / Sierra(LLNL)

- "Data-centric" buzzword – put computers where the data collection is? expensive to move data around (power and otherwise)
- IBM, can they deliver (had to pull out of BlueWaters project)
- greater than 100PFLOPS peak
- Power 10MW

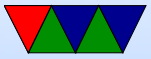


- 3400 nodes (40TFLOPS/node) [Titan had 18k nodes]
  Why is fewer nodes better? Shared memory?

- Multiple IBM Power9 Processors

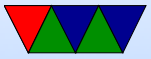
- Multiple NVIDIA Volta GPUs, NVLINK
  NVLINK is "high bandwidth energy-efficient CPU/GPU interconnect" 5-10x faster than PCIe
  3d-stacked memory?

- Memory: greater than 800GB (HBM+DDR4) per node, also NVRAM (each node 512GB+ coherent memory,



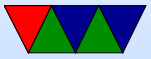
800GB NV) HBM = High Bandwidth Memory= 3d stacked

- Dual-rail Mellanox EDR-IB interconnect 23GB/s
- GPFS? 120Petabytes Spectrum scale?



# Software Stack

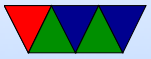
- Linux
- System Management: PCM (platform cluster management), xCAT (extreme cloud administration toolkit), GPFS
- Resource management: IBM Platform LSF Scheduling software
- Cuda, MPI, OpenMP, OpenACC, GA/ARMCI (global array, aggregate remote memory copy interface) (sort of



like MPI, cross-platform library for communication), CCI (common communication interface) (library for network transparent programming of tasks that outlast jobs, such as distributed FS, debugging, perf monitoring, etc), GASNet (anote high-speed networking library), OpenSHMEM (PGAS? partitioned global address space)

- Compilers LLVM, IBM XL, PGI

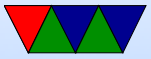
- Communication Libraries: MRnet – multicast reduction network, (efficient broadcast and reduction)



- Math libraries: ESSL (IBM Engineering and Scientific Subroutine Library), FFTW (Fast Fourier Transfer Library), ScaLAPACK (Scalable Linear Algebra Package), PETSc (Portable, Extensible Toolkit for Scientific Computation), Trilinos (C++ algorithms)

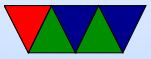
- I/O libraries: netCDF (network Common Data Form, machine independent file format), HDF5 (Hierarchical Data Format ) file format for large datasets

- MPI Library: OpenMPI



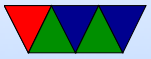
# Codebases

- ASC applications can take 10 years to complete
  Need to be maintainable and portable
- Moving to heterogeneous and complex memory will be difficult
- RAJA on top of OpenMP/MPI? RAJA is a C++ abstraction layer
- OpenMP 4.0 “target” accelerator offloading



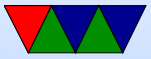
# Non-IBM system, Argonne

- Theta:
  8.5 Petaflops
  Knights Landing (Xeon Phi, Silvermont, 60+proc, AVX512)
  2500 nodes
  CrayCX
  Cray Aries Interconnect
  Lustre/10 Petabytes
  1.7MW power



2016

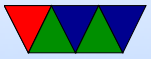
- Aurora:
  $>$ 180 Petaflops
  Knights Hill (Xeon Phi Next)
  50,000 nodes
  CrayShasta
  Intel Omni-Path (Si photonics) Interconnect
  Lustre/150 Petabytes
  13MW power (13GFLOPS/W)
  2018



# Other 100PFLOP machines

- Tianhe-2 (currently 38PFLOP) plan to have a 100PFLOP machine in 2016 (originally hoped for 2016 but blocked from more Xeon-Phi by US export controls). Looking to use Haswell+custom DSP chips instead? Tianhe-2a, 100 PFLOP, 18MW power, 3PB RAM, 6TFLOP/node, Custom accelerator, 18k nodes TH-Express 2+ interconnect

- Possibly another 100PFLOP machine using custom Chinese RISC processor, ShenWei SW1600



- Cori at NERSC (National Energy Research Scientific Computing Center) (in Oakland), Knights Landing, 30Petaflop

- Others? Army ARL?