ECE 574 – Cluster Computing
Lecture 23

Vince Weaver

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

vincent.weaver@maine.edu

1 December 2015

Announcements

- Project presentations next week
- There is a final. Maybe should change that for next time.
- One more (really brief) homework?

Exascale

- Exsacale computing – Exaflop $=$ 1000 Petaflops / 1M Terraflops
- Petascale in 2008, estimated Exascale in 2018 around that time? Others think 2020?
- Current fastest roughly 40 Petaflops
- Envision as having 100k nodes, each with 10Tflop; modern high-end GPUs only about 3Tflop (double percision)

- Many challenges

DOE

- US Department of Energys objectives and requirements for exascale systems
- Had series of workshops 2008 - 2012 to discuss what is needed

Power

- Biggest challenge going forward
- Power costs of largest system $5-10million
- Exascale with current tech would need 350 megawatts ($250 million/year)
- To be feasible really need to cap at 20 megawatts
- Data movement – historically 1byte/flop considered reasonable

But for current 2petaflops system that would take 1.25MW

Even if reduce to bare minimum (0.2byte/flop) would be 50MW for exascale.

Proposed: more energy-efficient hardware, Si-photonic communication, power-aware algorithms

Concurrency

- Already can't keep cores busy to mask long-latency (usually memory) events

- Flattening of CPU clock frequency is keeping things from getting worse, but having more cores making requests is not helping

- With exascale, costs more energy to transport data than to compute it.

Fault Tolerance

- Mean Time To Interrupt (MTTI)
- Improve MTTI so applications can run for hours without faults

RAM

- Current power levels unsustainable
- Slowing technology growth, from 4-times per 3 years to 2-times per three years
- Limiting factor in most applications
- Need 4TBpbs bandwidth and 1TB per node
  Current DIMMs have single-digit number of channels with 10s of GB/s

DRAM Performance Metrics

- Energy per bit
- Aggregate bandwidth per socket
- Memory capaity per socket
- FIT rate per node
- Error detection
- Processing in Memory

- Programmability

Programmability

- Three stages: algorithm capture, correctness debugging, performance optimization

- Parallelism – anticipated that 10-billion-way concurrency needed

- Distributed Resource Allocation – need to spread out to parallel, but also need to keep close for low-latency

- Latency Hiding – overlap communication with computation

- Hardware idiosyncrasies – allow using fast novel hardware without burdening programmer too much with the details
- Portability – use software across machine types
- Synchronization – barries and expesive operations replaed by lightweight (transactional memory?)

CPU/Network

Not really worried about CPU or Network?

AMD: Achieving Exascale Capabilities Through Heterogeneous Computing

- APU (CPU combined with GPU), 3D-RAM, connected to off-core NVRAM
- CPU handles serial sections, GPUs parallel sections
- APU – exascale heterogeneous processor (EHP) Supports HSA (Heterogeneous System Architecture) – CPU and GPU have same shared memory space, CPU and GPU can trade pointers w/o going over PCIe bus

CPU – 32 cores (ARM or x86)
CPU only possible, but probably not have high enough perf/W why integrated? lower overhead. Also higher FLOPS/volume (meters cubed)

- 2.5D interposer-based stacking vs 3d? (3d has through vias CPU to DRAM, 2.5d the dram stack is next to CPU with interposer board to connect)

- QuickRelease and HRF (heterogeneous-race-free) – need complex setup to get cache coherency between GPU and

CPU

- JEDEC high-bandwidth memory (HBM) standard 128 GBps per DRAM stack. With eight stacks, TBps with current tech.
- Three levels of memory (fast, NVRAM, flash?)
- How to use memory? Transparent like current, or expose to user?
- DRAM power – even if reduce from current 60pJ/bit of DDR3 to 2pJ/bit, 4TBps could consume half of the

power of entire cluster

- Processor in memory (PIM) can maybe prodide better energy efficiency
- How to program?
- Reliability? GPUs not typically as reliable as CPUs. Corruption in GPU output not considered as critical as in CPU

Intel's Exascale Plan

- Can't find a nice article like for AMD.
- Knight's Landing?
- 14nm successor to Knights Corner
- AVX-512, Multi-Channel DRAM (MCDRAM), Silvermont based CPU core
- 76 cores (72 with 4 spares)

- Omni-path interconnect, says it is more power effecient than infiniband

New IBM Supercomputers