ECE 574 – Cluster Computing
Lecture 19

Vince Weaver

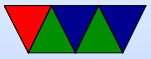
https://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

30 March 2023

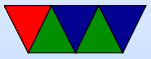
Announcements

- HW#8 was posted



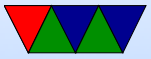
Project Topic Notes

- I responded to everyone's e-mail. If your group didn't get one let me know
- Have a wide variety of machines to run on.
- If interested in power measurement let me know.



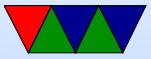
Cluster Computing Power

Can spend a whole class (i.e. ECE571) discussing where power goes in a modern computing system.



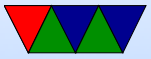
Computing Power Motivation

Why is low-power super-computing important?



Green500

- Green 500 list (Kirk Cameron?)
- Push for more accurate power reporting in the Top500 list



Green500 Highlights, Nov 2022

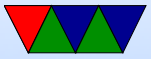
1. Henri (#293 top 500), Intel Xeon, NVIDIA H100
   2.8PFLOPs, 44kW, **65 GFLOPs/W**

2. Frontier TDS (Test and Devel) (#44), AMD EPYC/Instinct
   19 PFLOPs, 309kW, **62 GFLOPs/W**

3. Adastra (#17) **58 GFLOPs/W**, AMD

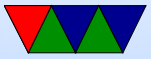
. . .

8. Frontier (#1) **52 GFLOPs/W**



My Personal Green 500 List

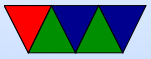
- https://web.eece.maine.edu/~vweaver/group/green_machines.html
- #1 is Apple M1 laptop, 6 GFLOPs/W
- haswell-ep is 2 GFLOPs/W (CPU only)
- pi-4 is also roughly 2GFLOPs/W



Pi-cluster Power

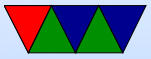
If we had more time I would have had you read *A Raspberry Pi Cluster Instrumented for Fine-Grained Power Measurement* by Cloutier, Paradis, and Weaver.

- ARM supercomputer has been #1. SVE-2 becoming mainstream.
- Low power, but floating-point so-so. Even worse is I/O (networking)
- Finally getting close



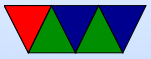
Pi-cluster Power

Machine	N	GFLOPS	Idle	Active	GFLOPS/W	MFLOPS/$
Pi 2B	10,000	1.47	1.8	3.4	0.432	42.0
Pi 3B	10,000	3.7/6.4	1.8	4.4	0.844	106
Pi 4B	20,000	13.5	2.5	7.3	1.85	385
Jetson TX-1	20,000	16.0	2.1	13.4	1.20	26.7
16x Haswell-EP	80,000	428	58.7	201	2.13	107
5xpi4-cluster	40,000	50.7	33.6	64.8	0.863	??
24xpi2-cluster	48,000	15.5	71.3	93.1	0.166	7.75



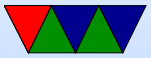
Pi-cluster Notes

- Per-node power measurement
- Network I/O big problem
- Why not OrangePi? (UTK)
- What if we could use the Pi GPU? No OpenCL, but people have reverse engineered part, also QPU?



SuperComputer Power

- Cooling
- DVFS
- Power-capping
- Up to 12% spent by the interconnect
  Pi2 cluster, 90W, 20W or so is the ethernet switch
- Accelerators like GPUs can save a lot of power

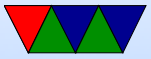


Fujitsu K Computer, 2012

- https://www.extremetech.com/extreme/120071-how-the-worlds-fastest-supercomputer-fujitsus-k-saves-on-power-and-money

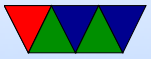
- Fine tune voltage for each CPU (variation in production). Save 7W/CPU (one Megawatt total)

- Watercooled



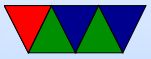
Titan Supercomputer, 2012

- http://www.anandtech.com/show/6421/inside-the-titan-supercomputer-299k-amd-x86-cores-and-186k-nvidia-gpu-cores/2
- Was upgrade: install 18,688 CPUs/GPUs manually
- 480V input to cabinets (rather than 208V) to reduce cable thickness
- 9MW, building gets 25MW
- Not big enough UPS for whole machine, flywheel UPS to keep I/O nodes up until diesel kicks in
- Cabinets are air cooled, but air is water-cooled first



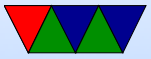
Frontier Supercomputer Paper, 2022

- https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202203/ASCAC_202203-Geist.pdf
- "Since 2009 biggest concern with reaching exascale has been energy consumption"
- 200x reduction in energy from Jaguar to Frontier
- Warm water cooling 32C (85F)



Frontier Supercomputer Paper – Metrics

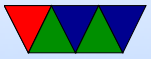
- Power Usage Effectiveness
- Efficiency of data center, total power coming in divided by power used by computing
- Higher is worse, close to 1 good
- Frontier (?) Data Center PUE = 1.03



Frontier Supercomputer Paper – The Rise of GPUs

Note: GF/W is 1/(MW/EF)*1000

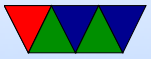
Computer	Megawatts/ExaFLOP	GFLOPs/W	GPU/CPU ratio
Jaguar	3043	0.3	0
Titan	330	3	1
Summit	65	15	3
Frontier	15	60	4



Power-Capping

Power Capping: a Prelude to Power Shifting by Lefurgy Wang, and Ware

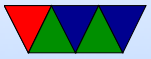
- Traditionally you have to design for the "worst-case" thermal and power behavior
- Often this will leave some resources underutilized "over-provisioned"
- Power-capping – let you design cheaper power/thermal setup, and if the CPU detects it is getting too hot/too much power automatically slows things down



Definitions

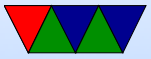
People often say Power when they mean Energy

- Dynamic Power – only consumed while computing
- Static Power – consumed all the time.
  Sets the lower limit of optimization

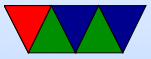


Units

- Energy – Joules, kWH (3.6MJ), Therm (105.5MJ), 1 Ton TNT (4.2GJ), eV (1.6×10^{-19} J), BTU (1055 J), horsepower-hour (2.68 MJ), calorie (4.184 J)
- Power – Energy/Time – Watts (1 J/s), Horsepower (746W), Ton of Refrigeration (12,000 Btu/h)
- Volt-Amps (for A/C) – same units as Watts, but not same thing
- Charge – mAh (batteries) – need voltage to convert to Energy



CPU Power and Energy

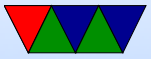


CMOS Dynamic Power

- $P = C\Delta V V_{dd} \alpha f$
  Charging and discharging capacitors big factor ($C\Delta V V_{dd}$) from V_{dd} to ground
  α is activity factor, transitions per clock cycle
  f is frequency

- α often approximated as $\frac{1}{2}$, $\Delta V V_{dd}$ as V_{dd}^2 leading to $P \approx \frac{1}{2} C V_{dd}^2 f$

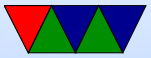
- Some pass-through loss (V momentarily shorted)



CMOS Dynamic Power Reduction

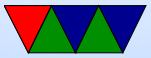
How can you reduce Dynamic Power?

- Reduce C – scaling
- Reduce V_{dd} – eventually hit transistor limit
- Reduce α (design level)
- Reduce f – makes processor slower



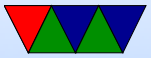
CMOS Static Power

- Leakage Current – bigger issue as scaling smaller. Forecast at one point to be 20-50% of all chip power before mitigations were taken.
- Various kinds of leakage (Substrate, Gate, etc)
- Linear with Voltage: $P_{static} = I_{leakage} V_{dd}$



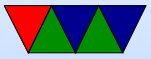
Leakage Mitigation

- SOI – Silicon on Insulator (AMD, IBM but not Intel)
- High-k dielectric – instead of SO2 use some other material for gate oxide (Hafnium)
- Transistor sizing – make only critical transistors fast; non-critical can be made slower and less leakage prone
- Body-biasing
- Sleep transistors



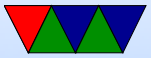
Total Energy

- $E_{tot} = [P_{dyanmic} + P_{static}]t$
- $E_{tot} = [(C_{tot}V_{dd}^2\alpha f) + (N_{tot}I_{leakage}V_{dd})]t$



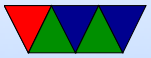
Delay

- $T_d = \frac{C_L V_{dd}}{\mu C_{ox}(\frac{W}{L})(V_{dd}-V_t)}$
- Simplifies to $f_{MAX} \sim \frac{(V_{dd}-V_t)^2}{V_{dd}}$
- If you lower f, you can lower V_{dd}



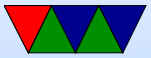
Thermal Issues

- Temperature and Heat Dissipation are closely related to Power
- If thermal issues, need heatsinks, fans, cooling



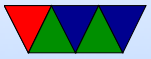
Metrics to Optimize

- Power
- Energy
- MIPS/W, FLOPS/W (don't handle quadratic V well)
- $Energy * Delay$
- $Energy * Delay^2$



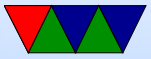
Power Optimization

- Does not take into account time. Lowering power does no good if it increases runtime.



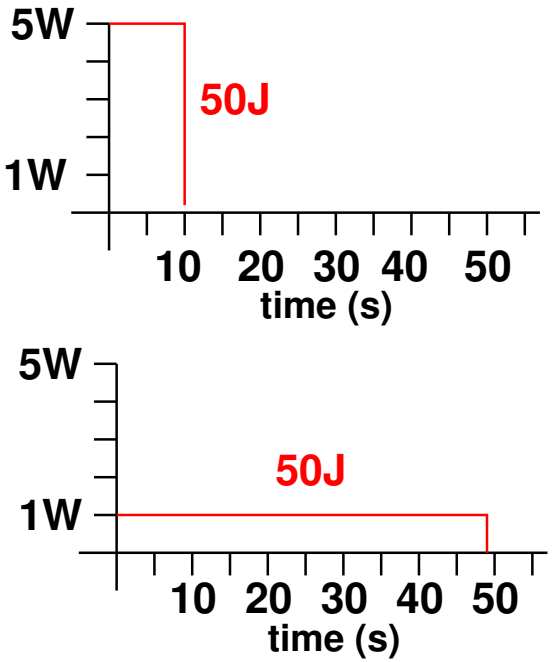
Energy Optimization

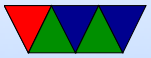
- Lowering energy can affect time too, as parts can run slower at lower voltages



Energy Optimization

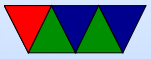
Which is better?



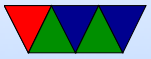


Energy Delay – Watt/t*t

- Horowitz, Indermaur, Gonzalez (Low Power Electronics, 1994)
- Need to account for delay, so that lowering Energy does not made delay (time) worse
- Voltage Scaling – in general scaling low makes transistors slower
- Transistor Sizing – reduces Capacitance, also makes transistors slower

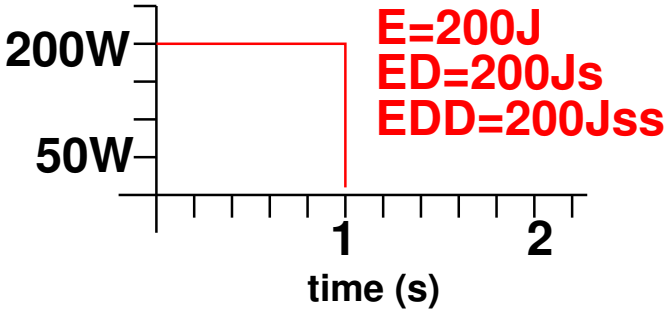


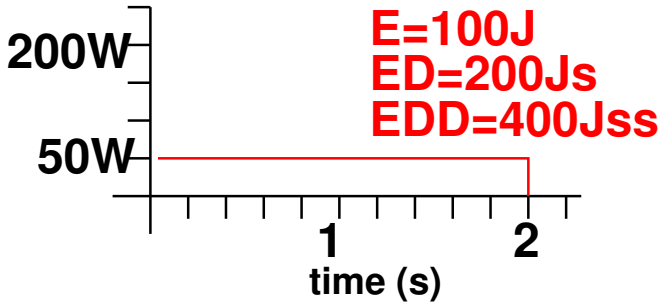
- Technology Scaling – reduces V and power.
- Transition Reduction – better logic design, have fewer transitions
  Get rid of clocks? Asynchronous? Clock-gating?

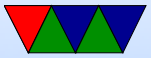


ED Optimization

Which is better?

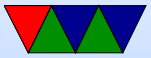




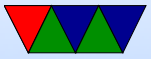


Energy Delay Squared– E*t*t

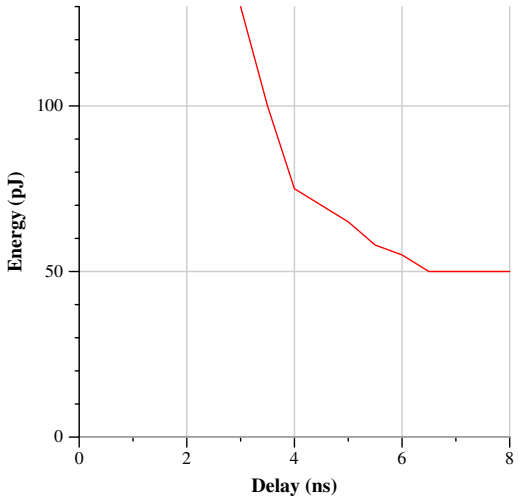
- Martin, Nyström, Pénzes – Power Aware Computing, 2002
- Independent of Voltage in CMOS
- ED can be misleading
  $E_a = 2E_b$, $t_a = \frac{t_b}{2}$
  Reduce voltage by half, $E_a = \frac{E_a}{4}$, $t_a = 2t_a$, $E_a = \frac{E_b}{2}$, $t_a = t_b$

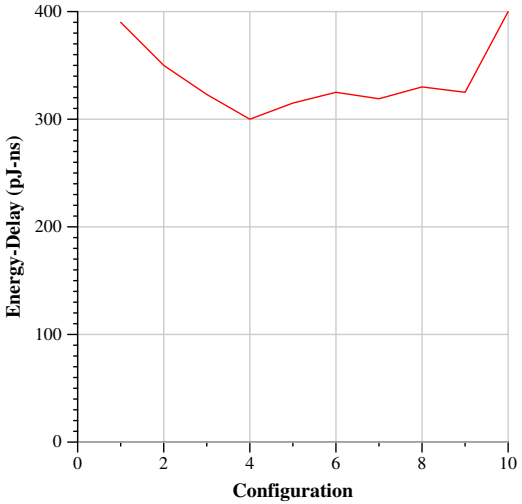


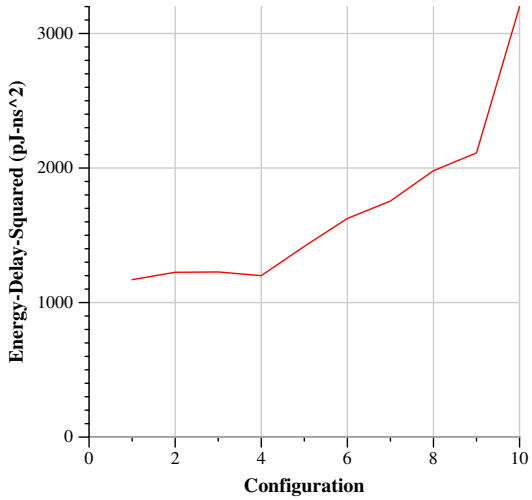
- Can have arbitrary large number of delay terms in Energy product, squared seems to be good enough



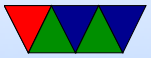
Energy-Delay Product Redux





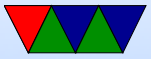


Roughly based on data from "Energy-Delay Tradeoffs in CMOS Multipliers" by Brown et al.



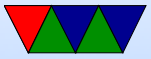
Raw Data

Delay	Energy	ED	ED^2
3	130	390	**1170**
3.5	100	350	1225
3.8	85	323	1227
4	75	**300**	1200
4.5	70	315	1418
5	65	325	1625
5.5	58	319	1755
6	55	330	1980
6.5	**50**	390	2535
8	50	400	3200

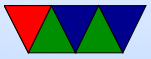


Other Metrics

- $Energy - Delay^n$ – choose appropriate factor
- $Energy - Delay - Area^2$ – takes into account cost (die area) [McPAT]
- Power-Delay – units of Energy – used to measure switching
- Energy Delay Diagram – [SWEEP]

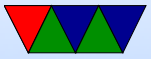


Measuring Power and Energy



Why?

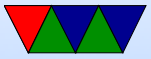
- New, massive, HPC machines use impressive amounts of power
- When you have 100k+ cores, saving a few Joules per core quickly adds up
- To improve power/energy draw, you need some way of measuring it



Energy/Power Measurement is Already Possible

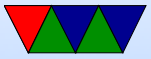
Three common ways of doing this:

- Hand-instrumenting a system by tapping all power inputs to CPU, memory, disk, etc., and using a data logger
- Using a pass-through power meter that you plug your server into. Often these will log over USB
- Estimating power/energy with a software model based on system behavior



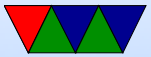
Measuring Power – Sense Resistor

- Sense resistor or Hall Effect sensor gives you the current
- Sense resistor is small resistor. Measure voltage drop. Current V=IR Ohm's Law, so V/R=I
- Voltage drops are often small (why?) so you made need to amplify with instrumentation amplifier
- Then you need to measure with A/D converter
- $P = IV$ and you know the voltage
- How to get Energy from Power?



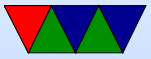
Where does the power go in a system?

- CPU
- DRAM
- GPU
- Disk
- Network
- Cooling/Fan
- Power Supply



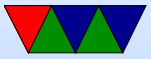
Measuring system wide

- Can use kill-o-watt, WattsUpPro, or similar



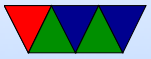
Measuring hardware

- Really hard
- DRAM, PCI, USB, fans: can put sense resistor in line with power supply, measure current and voltage to calculate power
- CPU harder, how do you intercept? There is the P4 line (a 12V special power cable from PSU on recent systems) but it might also power other parts of the system



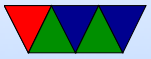
Estimating Power

- Can construct a model to estimate power
- Inputs like performance counters
  - CPU might be related to instructions/cycle, pipeline stalls, FPU instructions retired, etc
  - DRAM might be related to cache misses, bytes/second
- Hopefully you validate the model



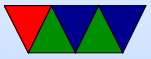
RAPL Power Estimates

- Running-Average Power Limit
- Recent Intel CPUs
- Need to estimate power usage for power-capping and turbo-boost
- Nicely provide the values to userspace
- On most systems (excluding some Haswell models) is an estimate from an on-chip power model based on various inputs (temperature, perf counters, etc)
- Very easy to read, using the perf tool

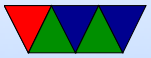


Energy Aware Algorithms

- It's easy to optimize for speed
- Can you optimize an algorithm for power/energy? Does anyone do this? HPC? Cellphone apps?
- With HPC can be tricky
  - FPU calculation so many flops/J
  - Copying from memory so many bits/J
  - Copying across network many more bits/J
  - Does it make more sense to duplicate calculations if it saves a trip across the network?



- Another problem, fastest doesn't necessarily mean least energy
- Conversely, easy to play games where you use less energy but takes so much extra time it's not worth it (see energy delay)



Other issues

- Cost of electricity. Much higher in other countries. Even varies state to state in US. Maybe good reason for power-capping
- Many of these issues also apply to large datacenters, like Google, or cloud-computing like Amazon/Microsoft
- It's easier to push problems off into the cloud