Performance = Bandwidth divided by Latency

Yale N. Patt

The University of Texas at Austin

ICS 2016

June 1, 2016, Istanbul

Why this title (since it is not accurate!)

- ***The real equation is:***

  Performance = 1/(instructions)(CPI)(cycle time)

- ***Preoccupation with bandwidth.***

- ***What about latency? …and whose job is it anyway?***

Problem

Algorithm

Program

ISA (Instruction Set Arch)

Microarchitecture

Circuits

Electrons

At the layers (actually usually more than one):

- ***Algorithm layer***
  - ***Approximate computing***
  - ***Accelerators***
- ***Language layer***
  - ***Pragmas***
- ***Compiler layer***
  - ***Prefetch and Post-store of Cache data***
  - ***Dynamic recompilation***
  - ***Wish branches***
- ***Architecture layer***
  - ***Dense encoding of instructions***
  - ***EMT instruction***
  - ***Large Scratch Pad***

Microarchitecture layer

- ***Asynch for awhile, then synch***
- ***Big-little cores***
- ***Accelerators (ASICs and FPGAs)***
- ***Use of dark silicon***
- ***Allocation of Shared resources***
- ***Prefetching***
- ***Branch Prediction***
- ***Near-neighbor communication***
- ***Run-time system***

Thank you!