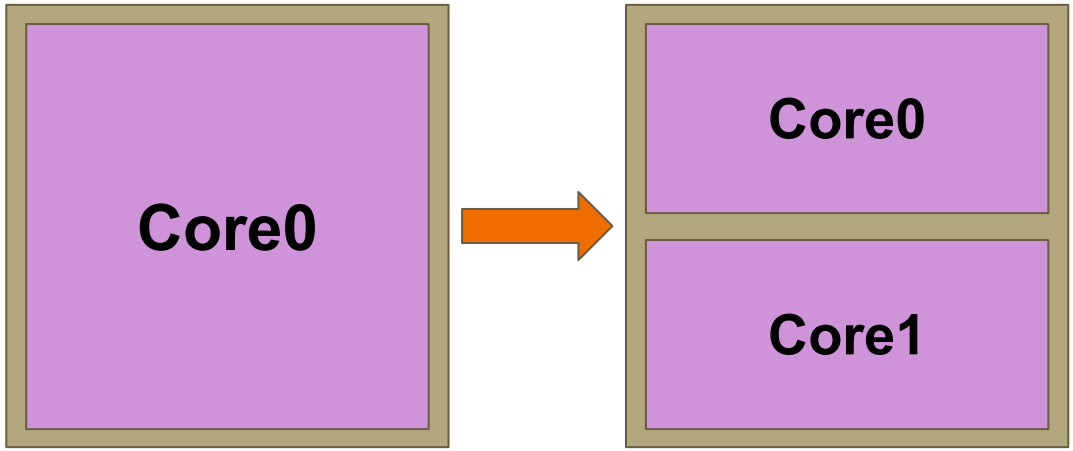
Bits of Architecture

The Multi-Core Era

We Can’t Scale Frequency - Now What?

No More Performance?

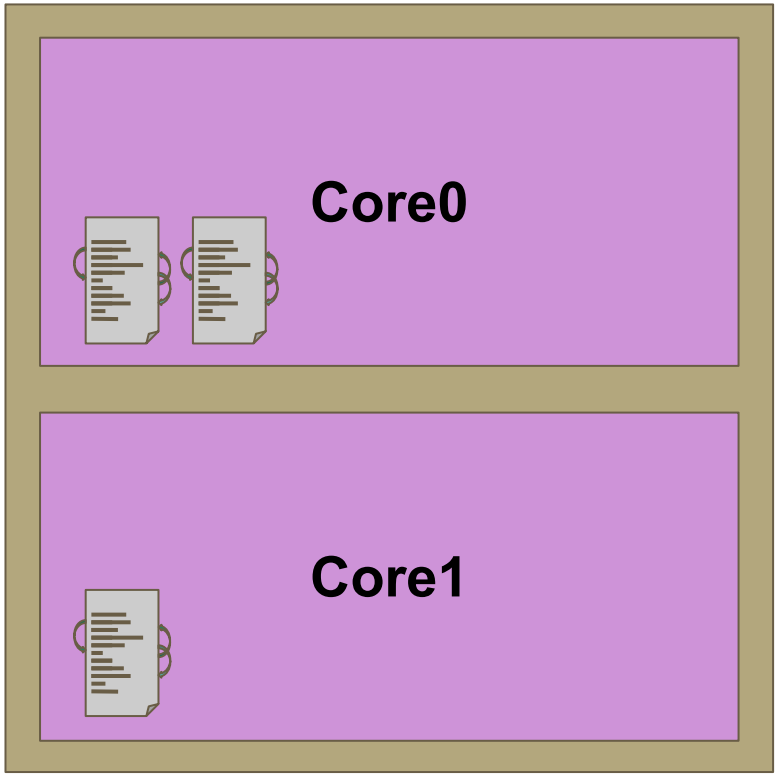
- Frequency drove a lot of single-threaded performance
- Transistors are still getting smaller (slowly)
- How do we allocate our silicon?
  - Branch predictors?
  - Reorder buffers?
  - Vector Hardware?
- >1 Core?



Why Multi-Core Works

Exploiting Parallelism

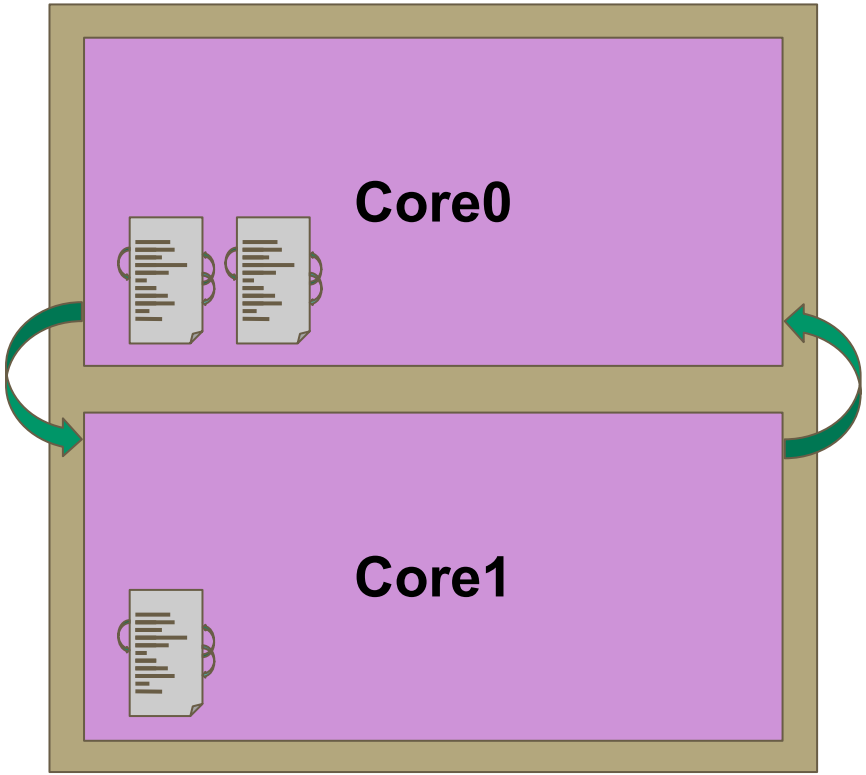
- We often want to run many programs at once
- Our programs often have independent sections
- Multiple cores allow us to exploit these things



Why Multi-Core Is Difficult

Parallel Programming is Performance Programming

- Parallel programming breaks our layers of abstraction
  - Programs need to to know about hardware
- Tricky to get right
  - Functionality
  - Performance
- New problems
  - Coherence



Multi-Core Forever?

Not So Fast…

- Transistors keep getting smaller (for now)
- More transistors = More Cores
  - But wait! What about the end of **Dennard Scaling**?
- More transistors, but we can't turn them all on at once (**Dark Silicon**)
  - Thermal Design Power constraint

Dark Silicon and the End of Multicore Scaling

Hadi Esmaeilzadeh[†] Emily Blem[‡] Renée St. Amant[§] Karthikeyan Sankaralingam[‡] Doug Burger[◇]
[†]University of Washington [‡]University of Wisconsin-Madison
[§]The University of Texas at Austin [◇]Microsoft Research
hadianeh@cs.washington.edu blem@cs.wisc.edu stamant@cs.utexas.edu karu@cs.wisc.edu dburger@microsoft.com