Accelerators

What is an accelerator?

- **Additional hardware to "offload" common software** operations
  - Recall a Turing machine can do anything
    - Explain Turing machines
  - Past accelerators:
    - x87 series – floating point and math operations
      - Otherwise emulate in software (c.f., datalab)
    - GPUs
      - Always present, role / capabilities changed

Rewrite Amdahl's law in terms of resource limits

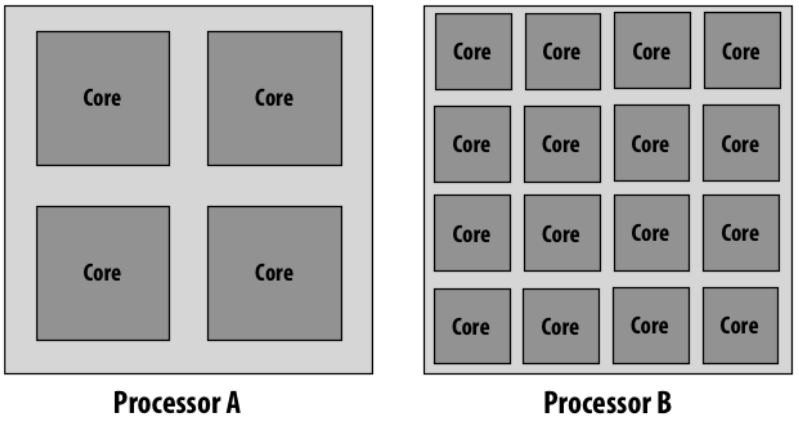
$$\text{speedup}(f, n, r) = \frac{1}{\frac{1-f}{\text{perf}(r)} + \frac{f}{\text{perf}(r) \cdot \frac{n}{r}}}$$

Relative to processor with 1 unit of resources, n=1.
Assume perf(1) = 1

f = *fraction of program that is parallelizable*
n = *total processing resources (e.g., transistors on a chip)*
r = *resources dedicated to each processing core,*
(each of the n/r cores has sequential performance $perf(r)$

More general form of **Amdahl's Law in terms of f, n, r**

Two examples where n=16
r_A = ***4***
r_B = ***1***



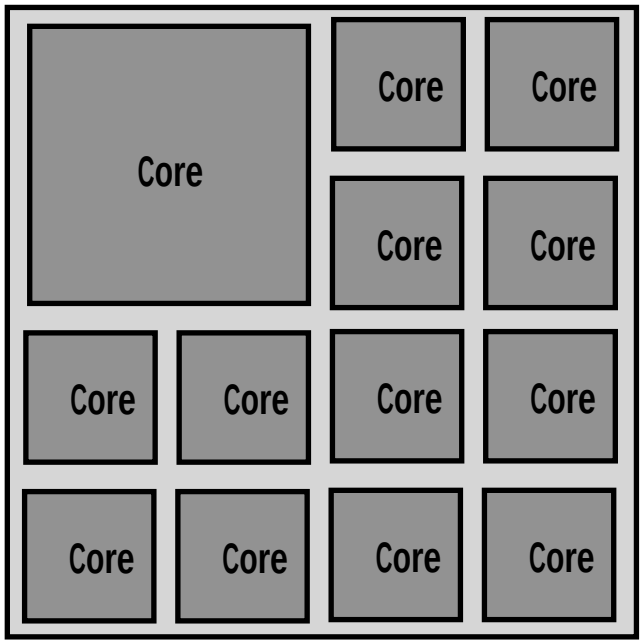
[Hill and Marty 08]

Asymmetric set of processing cores

Example: n=16

One core: r = 4

Other 12 cores: r = 1



$$\text{speedup}(f, n, r) = \frac{1}{\frac{1-f}{\text{perf}(r)} + \frac{f}{\text{perf}(r)+(n-r)}}$$

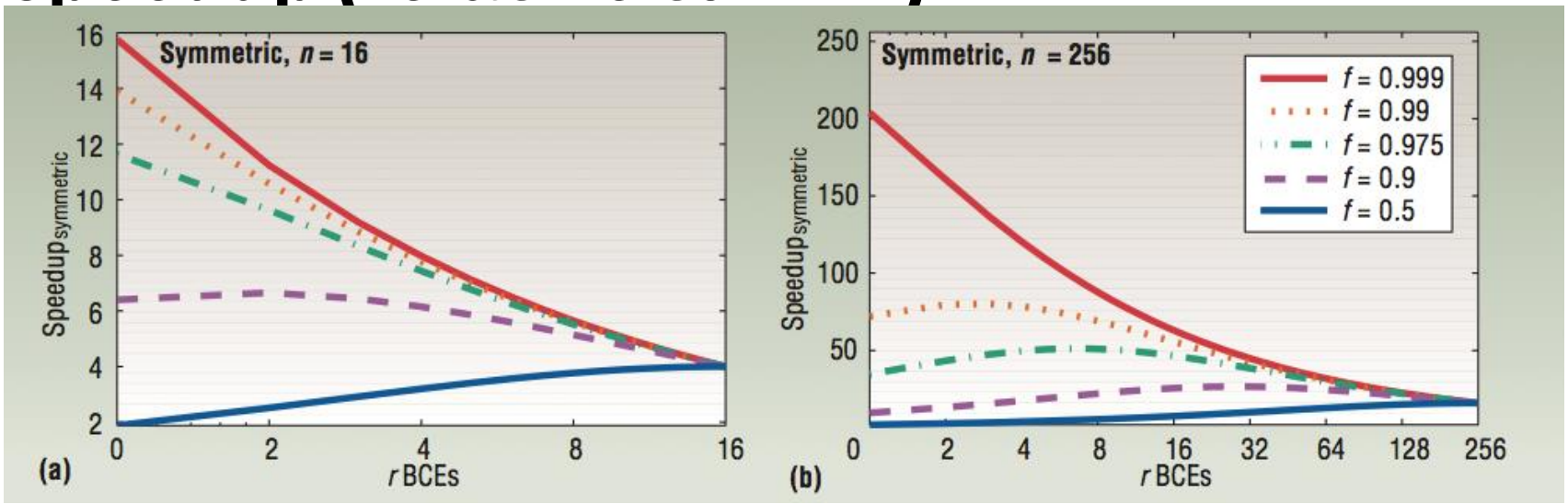
(of heterogeneous processor with *n* recourses, relative to uniprocessor with one unit worth of resources, n=1)

one perf(r) processor + (n-r) perf(1)=1 processors

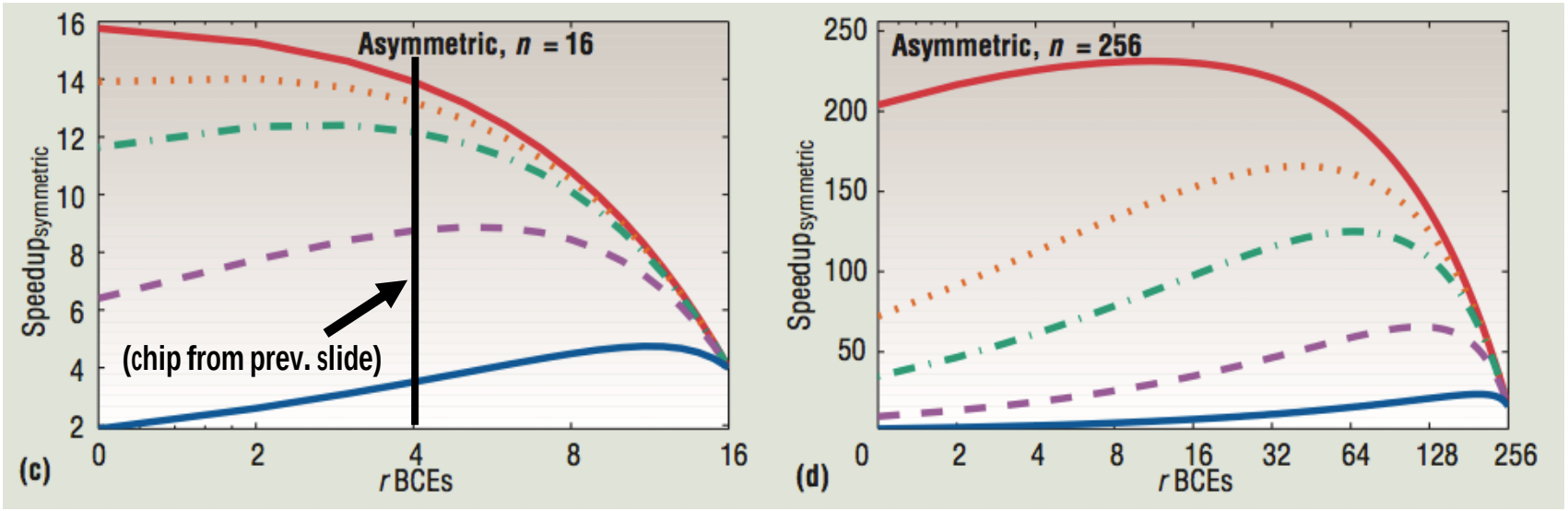
[Hill and Marty 08]

Speedup (relative to n=1)

[Source: Hill and Marty 08]



X-axis for symmetric architectures gives r for all cores (many small cores to left, few "fat" cores to right)



X-axis for asymmetric architectures gives r for the single "fat" core (assume rest of cores are $r = 1$)

Multiamdahl Takeaway

- Heterogeneous CPUs

Saving Energy, Winning Performance

- 16b int add ~ 32 fJ
  - ARM add ~ 250 pJ
- Sending a word from local cache ~ 5pJ
  - Sending a word from DRM ~ 640 pJ

https://www.microarch.org/micro52/media/dally_keynote.pdf

Where is energy being spent?

- Bitcoin
- DNNs
- Search

How FPGAs work (briefly)

- Key component is LUTs
  - Each is n-bit input and output (often 5)
  - Can compute any arbitrary n-bit Boolean function

[Putnam et al. ISCA 2014]

Project Catapult

- Microsoft Research investigation of use of FPGAs to accelerate datacenter workloads
- **Demonstrated offload of part of Bing Search's document ranking logic**
- Now widely used to accelerate DNNs across Microsoft services

FPGA board

1U server (Dual socket CPU + FPGA connected via PCIe bus)

- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow

200 LFM

68 °C Inlet

DNNs

- Can approximate other functions

https://www.cs.utexas.edu/~cart/publications/dissertations/hadi.pdf