ECE408 / CS483/CSE408 Fall 2022

Applied Parallel Programming

Lecture 26: Course Retrospective

Course Reminders

- Labs
  - Full set of grades will be available on Friday
- Project
  - PM3 is due this Friday Dec 2nd
  - Competition has been extended to Sun Dec 11th
- Midterm Exam 2
  - Tuesday 7-8:20pm December 6th,
    - **See wiki for details**
  - In Canvas, similar to MT1

Some notes

- Please fill out ICES forms for ECE 408
- Note on CampusWire about Physics Grad Student looking for CUDA help

ECE 408 Retrospective: What did we do this semester?

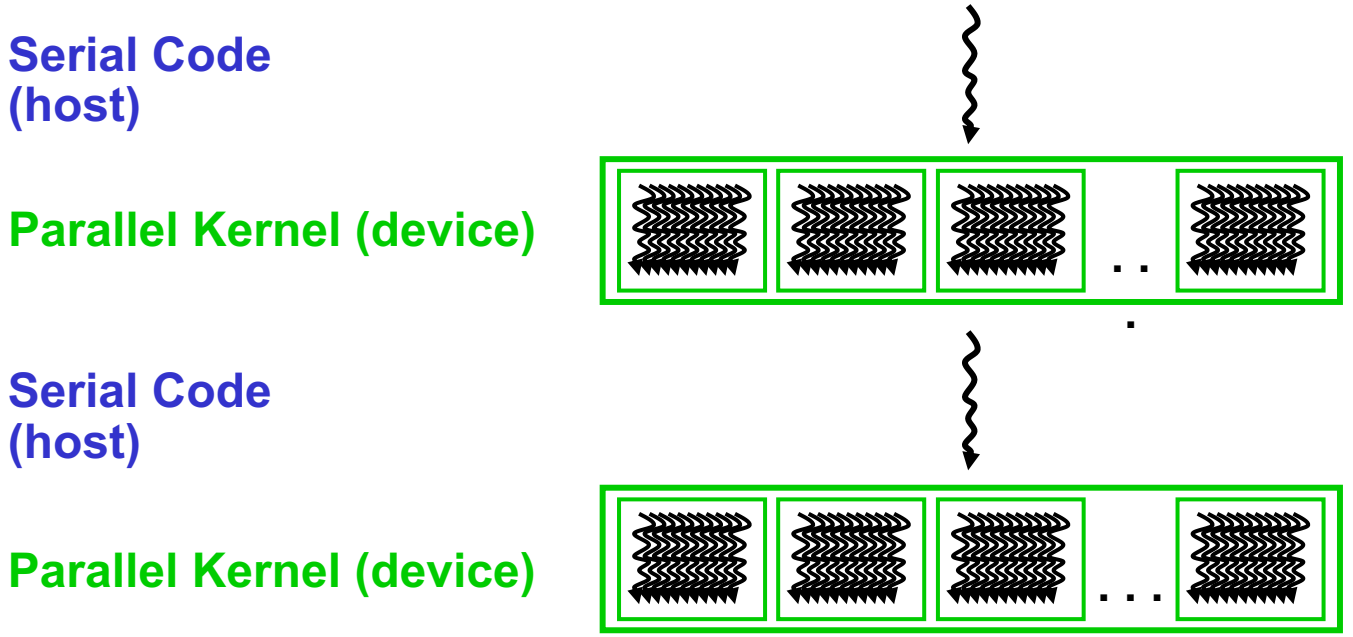
- Elementary Computational Patterns
  - Matrix Multiply, Convolution, Reduction, Scan, Histogram, Sparse Representations
- Parallel Optimization
  - Threading, Memory Management, Coalescing, Thread Divergence, Task Management, Profiling
- Programming Systems
  - CUDA, OpenACC, (OpenCL, DPC++, Hip, OpenMP,…)

Why did GPU Compute become a thing?

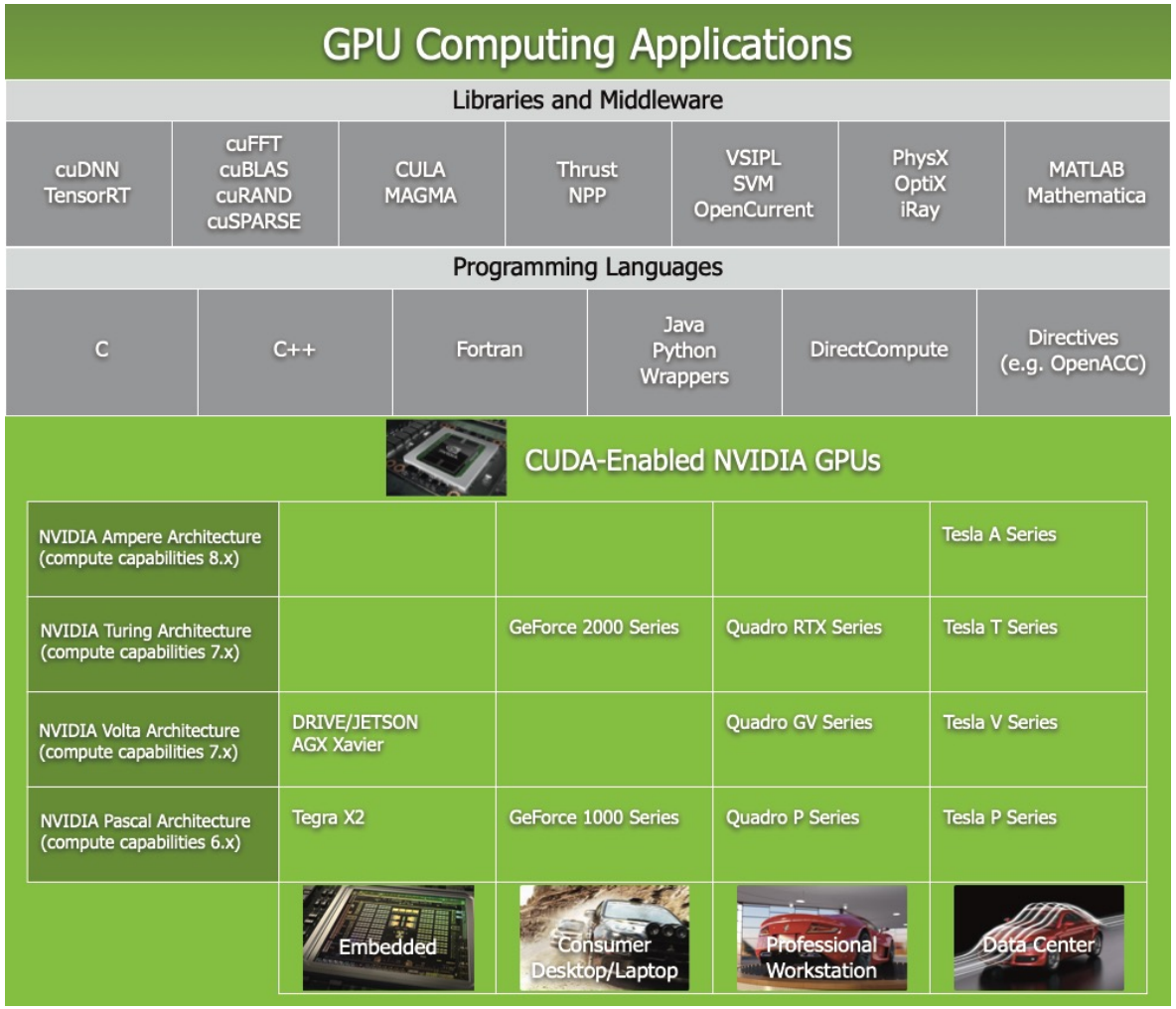
- And if its such an important concept, why did it take 15 years?

Bulk Synchronous Model Contributed to the Success

- In bulk synchronous, barriers separate temporal regions of code. interleaving / data sharing occurs only within regions (called phases).



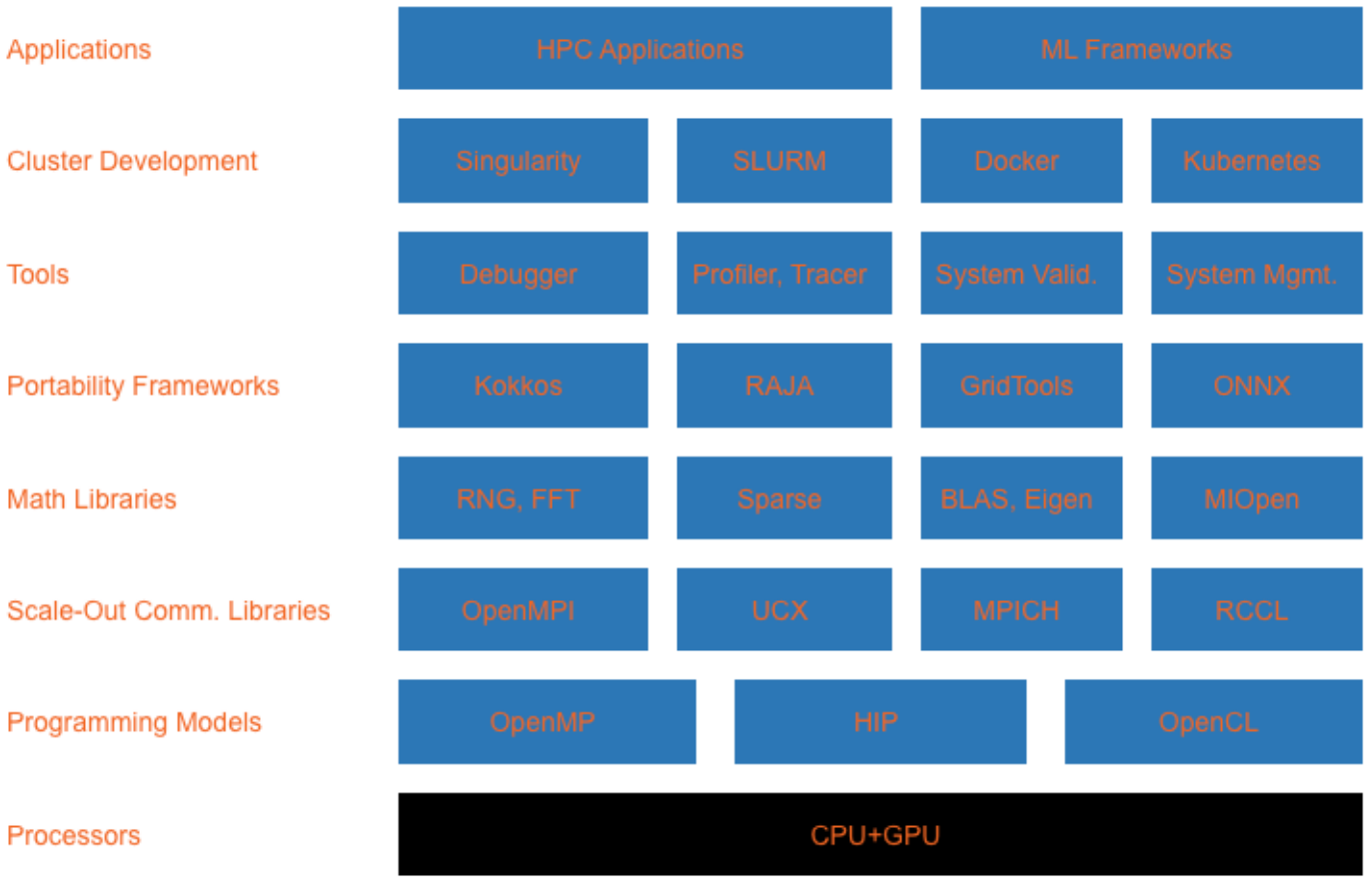
The GPU Computing EcoSystem 2021 (Nvidia Perspective)



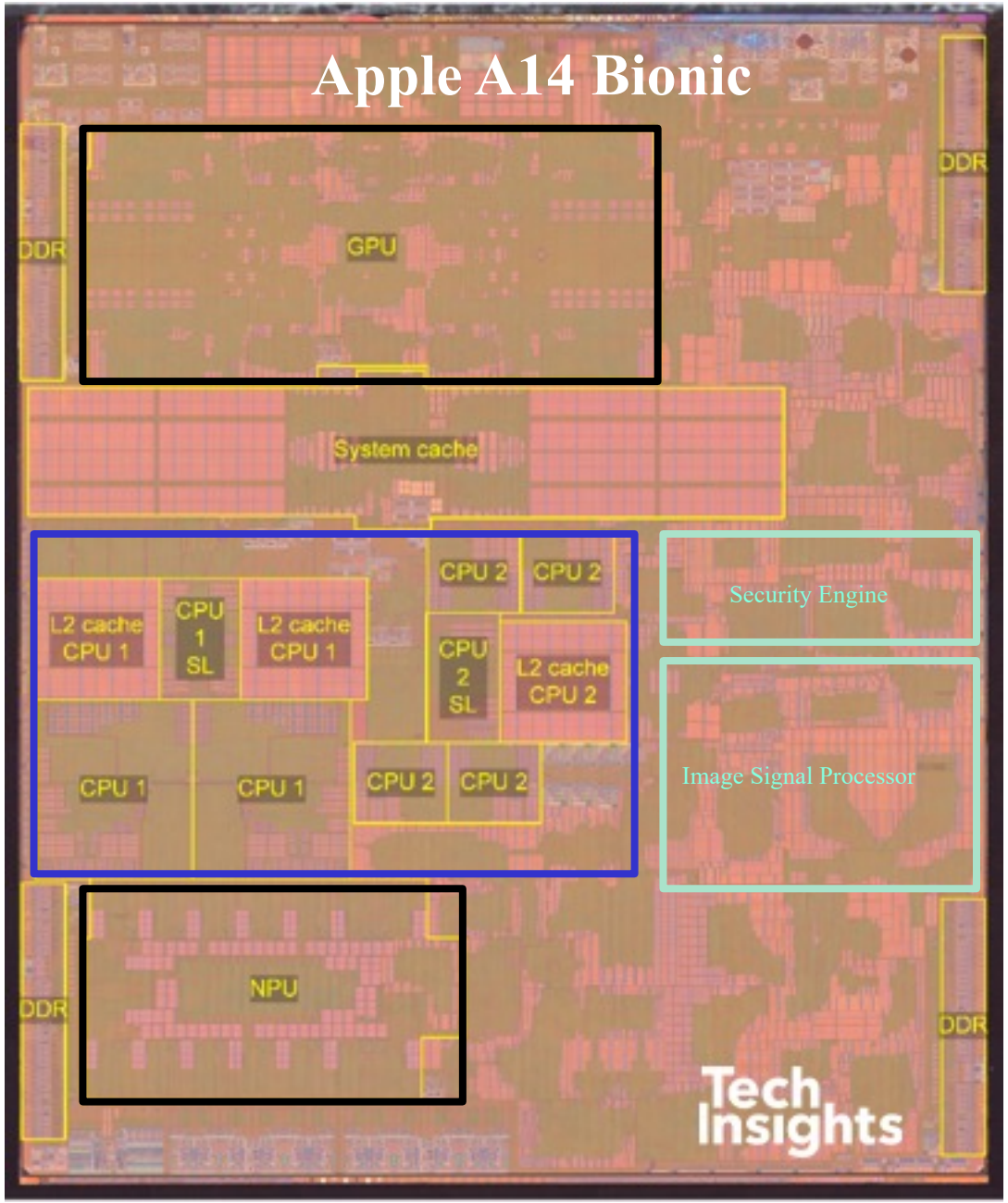
Since it's introduction in 2007, the level of developer abstraction has been steadily moving upward. More of the heavy GPU lifting is done by the companies developing the GPUs

AMD Perspective

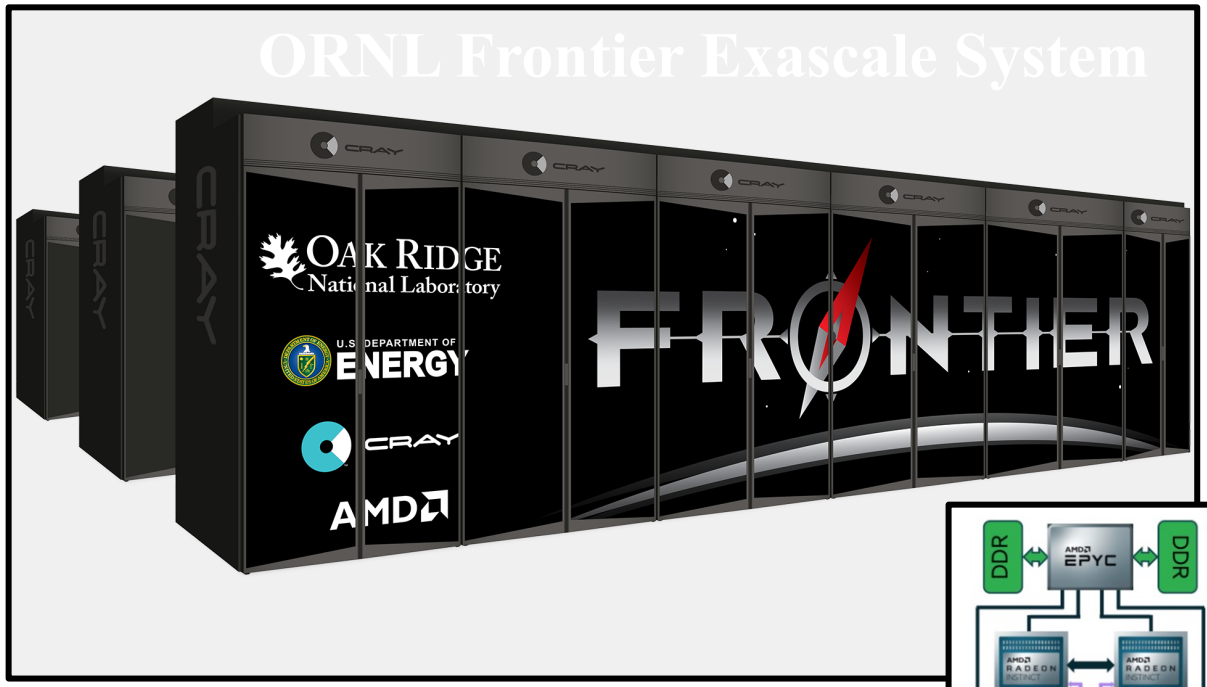
ROCm Components

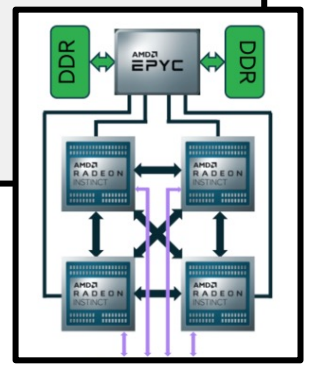


Apple A14 Bionic



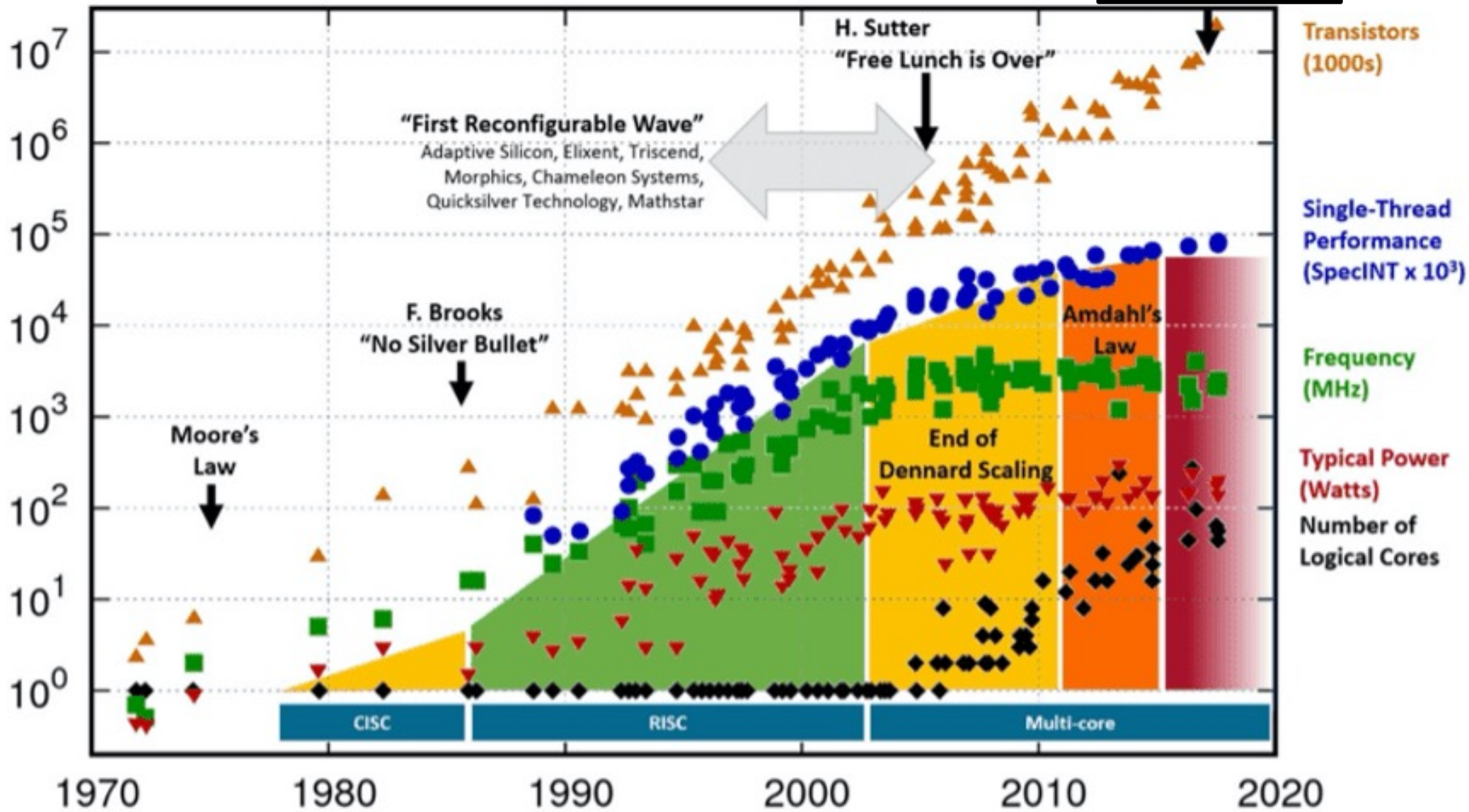
ORNL Frontier Exascale System





42 Years of Processor Data

The Accelerator Era



Hennessy and Patterson, Turing Lecture 2018, overlaid over "42 Years of Processors Data"
https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/;"First Wave" added by Les Wilson, Frank Schirrmeister
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp