CS/EE 217 – GPU Architecture and Parallel Programming

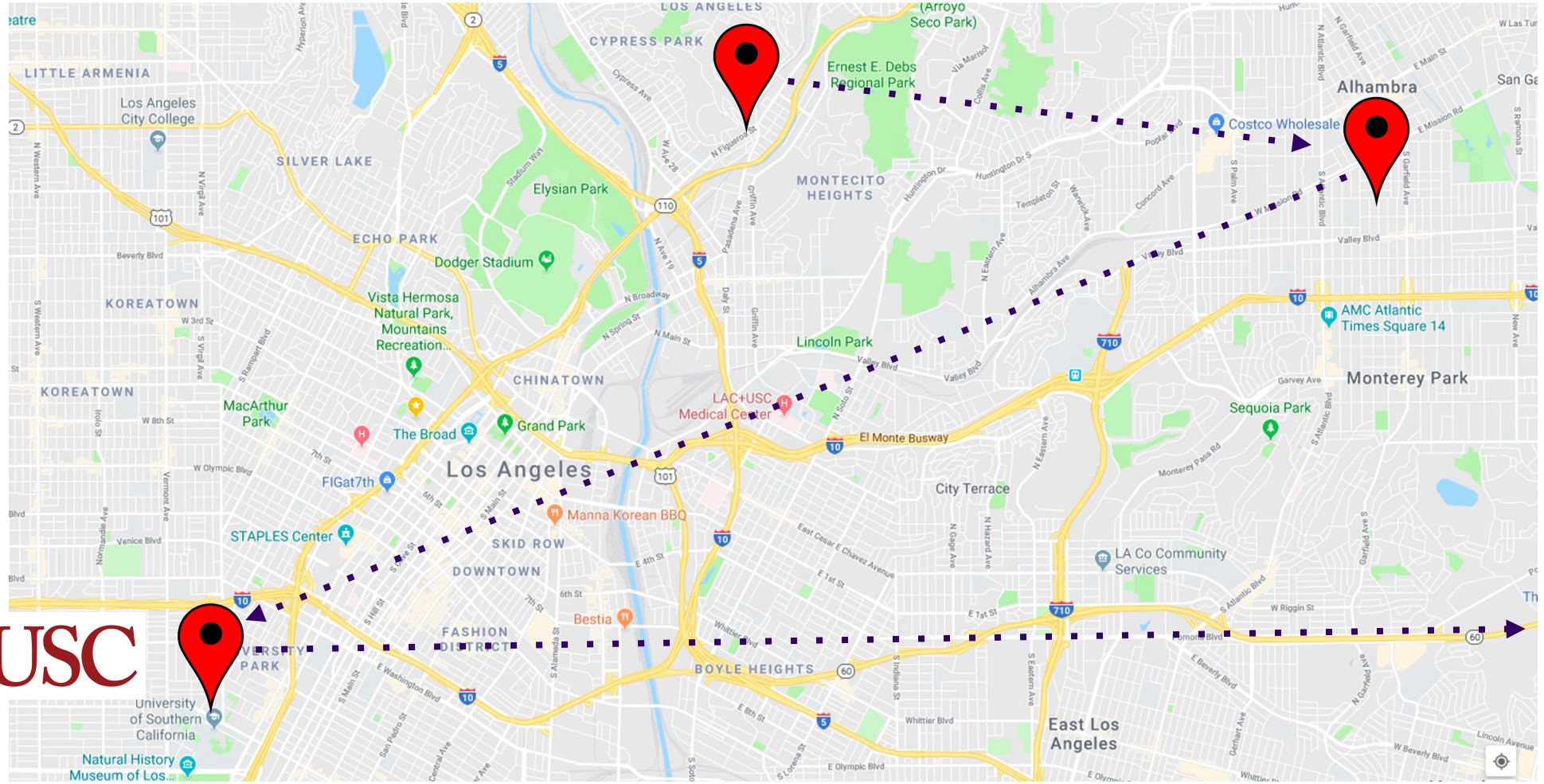
Introduction

Daniel Wong, Associate Professor
Department of Electrical and Computer Engineering
Cooperating Faculty,
Department of Computer Science and Engineering
danwong@ucr.edu

UNIVERSITY OF CALIFORNIA, RIVERSIDE

Welcome to CS/EE 217!

Ice Breaker – About me



Education

University of Southern California, Los Angeles, CA

BS Computer Engineering/Computer Science ’09

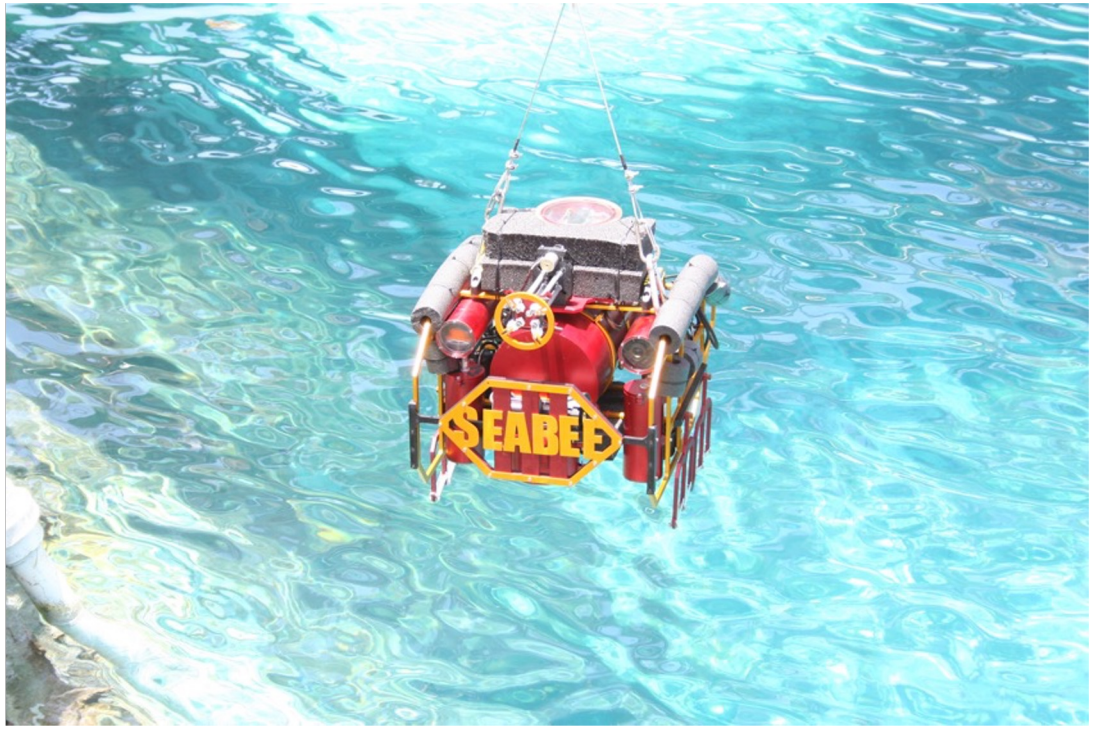
MS Electrical Engineering ’11

PhD Electrical Engineering ’15

USC

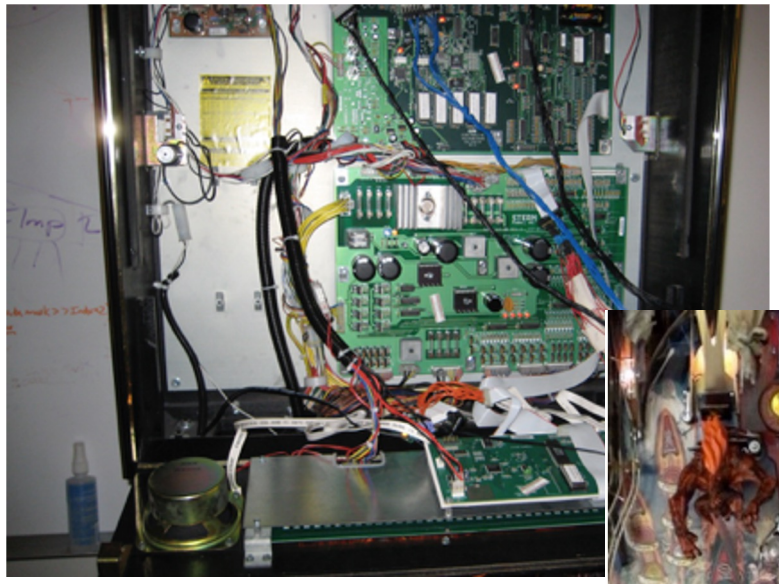
Why Computer Architecture?

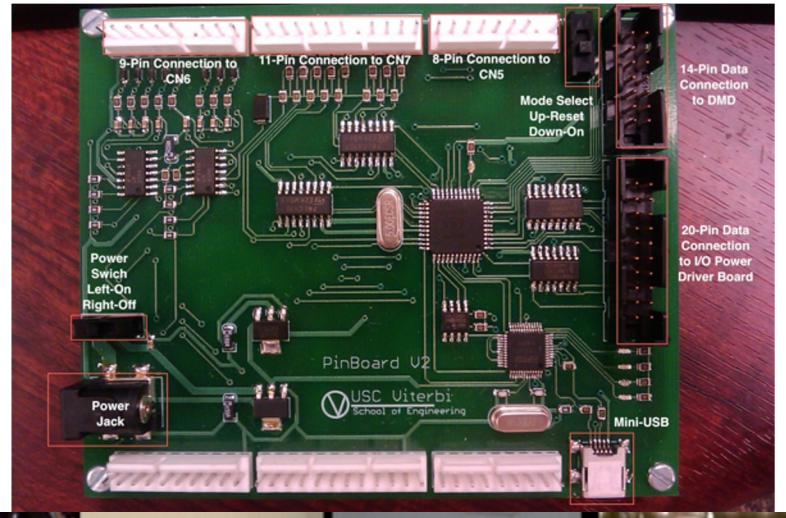
AUVSI 2009 (RoboSub)

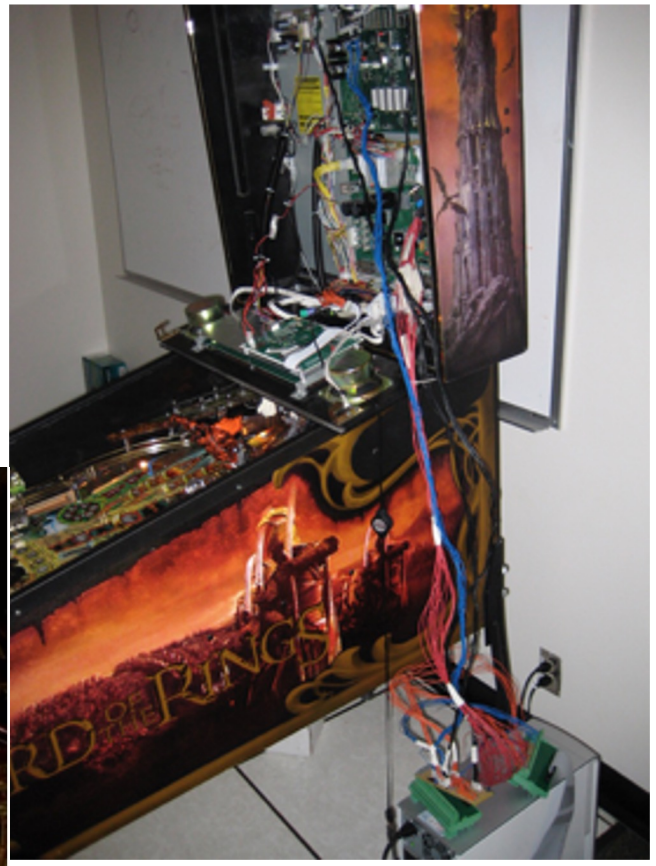


Why Computer Architecture?

Pinball Machines!







https://www.pinballnews.com/learn/pinhorse/

Why Computer Architecture?

Lawrence Livermore National Laboratory

SAMSUNG
SEMICONDUCTORS

Why Computer Architecture?

CSRankings: Computer Science Rankings

CSRankings is a metrics-based ranking of top computer science institutions around the world. **Click on a triangle** (▶) to expand areas or institutions. **Click on a name** to go to a faculty member's home page. **Click on a chart icon** (the after a name or institution) to see the distribution of their publication areas as a bar chart. **Click on a Google Scholar icon** () to see publications, and **click on the DBLP logo** () to go to a DBLP entry. *Applying to grad school? Read this first.* **Do you find CSrankings useful? Sponsor CSrankings on GitHub.**

Rank institutions in USA by publications from 2013 to 2023

All Areas [off | on]

AI [off | on]

- ▶ Artificial intelligence
- ▶ Computer vision
- ▶ Machine learning
- ▶ Natural language processing
- ▶ The Web & information retrieval

Systems [off | on]

- ▼ Computer architecture

ACM SIGARCH

- ASPLOS
- ISCA
- MICRO

- HPCA

#	Institution	Count	Faculty
1	▶ Univ. of Illinois at Urbana-Champaign	56.2	24
2	▶ University of Michigan	48.5	24
3	▶ Georgia Institute of Technology	34.8	17
4	▶ Massachusetts Institute of Technology	30.8	14
5	▶ University of Wisconsin - Madison	27.2	14
6	▶ Pennsylvania State University	25.7	15
7	▶ North Carolina State University	22.4	11
8	▶ Purdue University	21.6	12
9	▶ Princeton University	21.2	9
10	▶ Carnegie Mellon University	20.5	15
11	▶ Univ. of California - Riverside	20.4	11
12	▶ Cornell University	18.6	11

CS/EE 217 GPU ARCHITECTURE AND PARALLEL PROGRAMMING

Contact

- Instructor: Daniel Wong
  - Email: danwong@ucr.edu
  - Homepage: http://www.danielwong.org
  - Office Hours: TBD
- TA: Ali Jahanshahi, Mohammadreza Rezvani
  - Email: ajaha004@ucr.edu@ucr.edu, mrezv002@ucr.edu
  - Office Hours: TBD
- Note: Monday discussion section will not be used

Course Goals

- Learn how to program GPGPU processors and achieve
  - High performance
  - Functionality and maintainability
- Learn how GPGPUs work
  - Microarchitecture
  - Design challenges
- Technical subjects
  - principles and patterns of parallel algorithms
  - processor architecture features and constraints
  - programming API, tools and techniques

Logistics

- Course Website
  - http://teaching.danielwong.org/csee217/fall23
  - Check often for announcements
- eLearn (eLearn.ucr.edu)
  - Grades / Exams
  - Yuja for hosting lecture videos
- Piazza for discussion board
  - Please use Piazza for all forms of communication!
- ENGR Account Setup
  https://www.engr.ucr.edu/secured/systems/login.php

Textbook

1. D. Kirk and W. Hwu, “Programming Massively Parallel Processors – A Hands-on Approach, Second Edition”

2. CUDA by example, *Sanders and Kandrot*

3. *Nvidia CUDA C Programming Guide*
   - *https://docs.nvidia.com/cuda/cuda-c-programming-guide/*

Attendance/Recordings

- Attendance
  - You are expected to attend all lectures.
  - However, stay home if you're sick or feeling unwell. ☺
- Lectures will be recorded and made available after lecture.
- You are responsible for keeping up with course lectures and assignments!

Grading

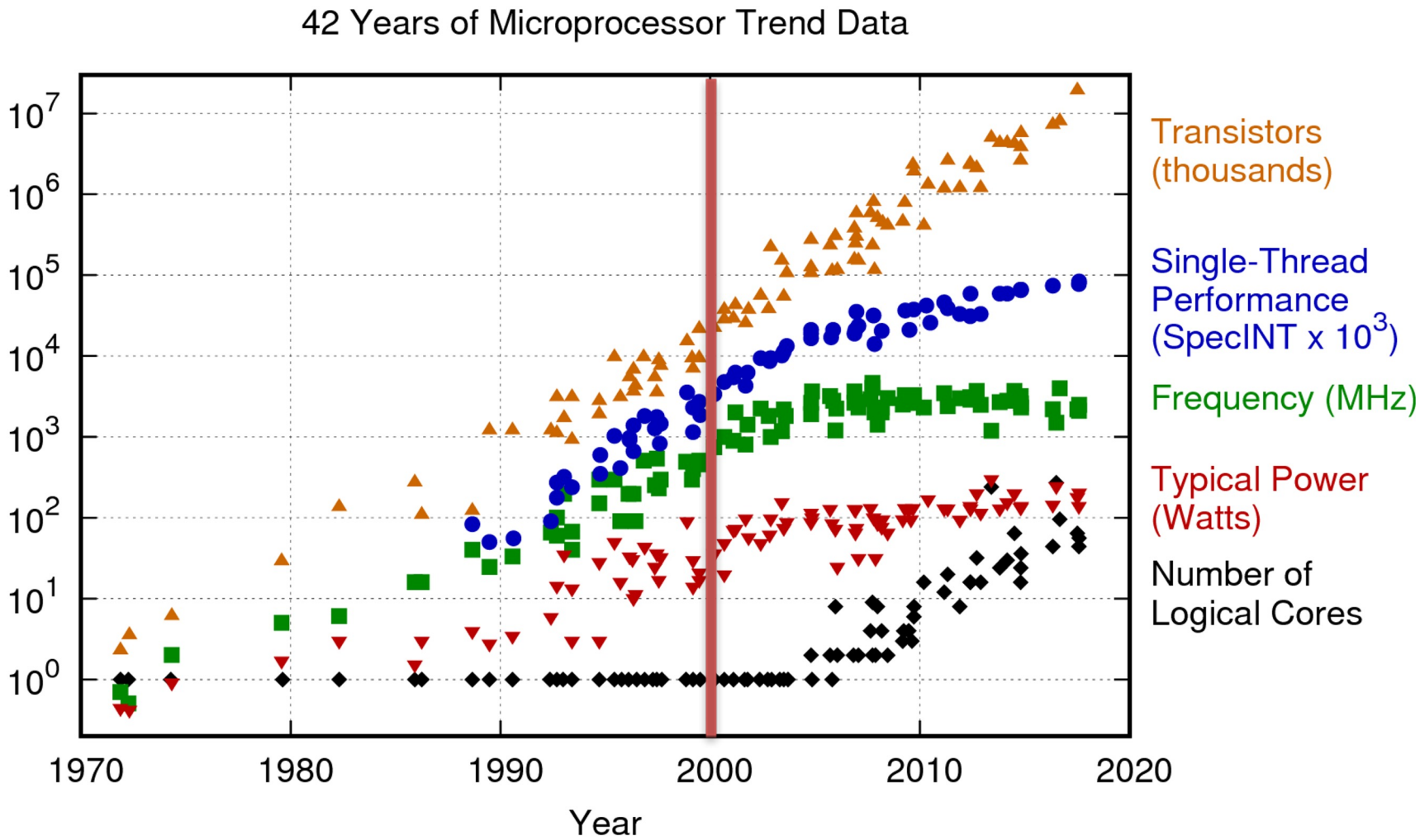
- Grade Breakdown
  - Assignments: 32% (8% each assignment)
  - Final Project: 20% (Implementation: 10%, Final report: 10%)
  - Midterm Exam: 25%
  - Final Exam: 20%
  - Participation/Discussion: 3%
  - Extra credit: 2%

Assignment Policies

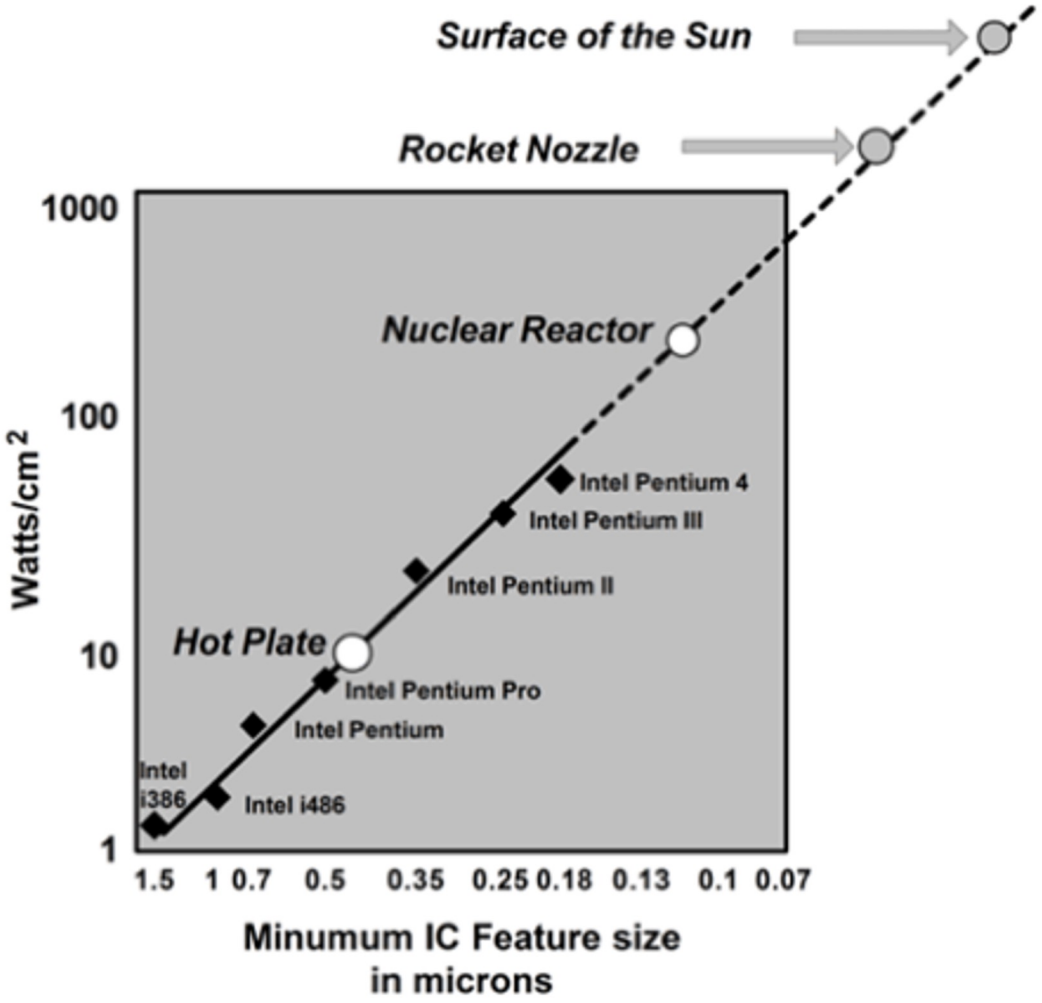
- 3 slip days
- 15% penalty per late day
- If it's one minute late, it's still late
- No extensions will be given
- All assignments/projects are due at the end of the due date (midnight Pacific time)

QUESTIONS?

Why GPUs?



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

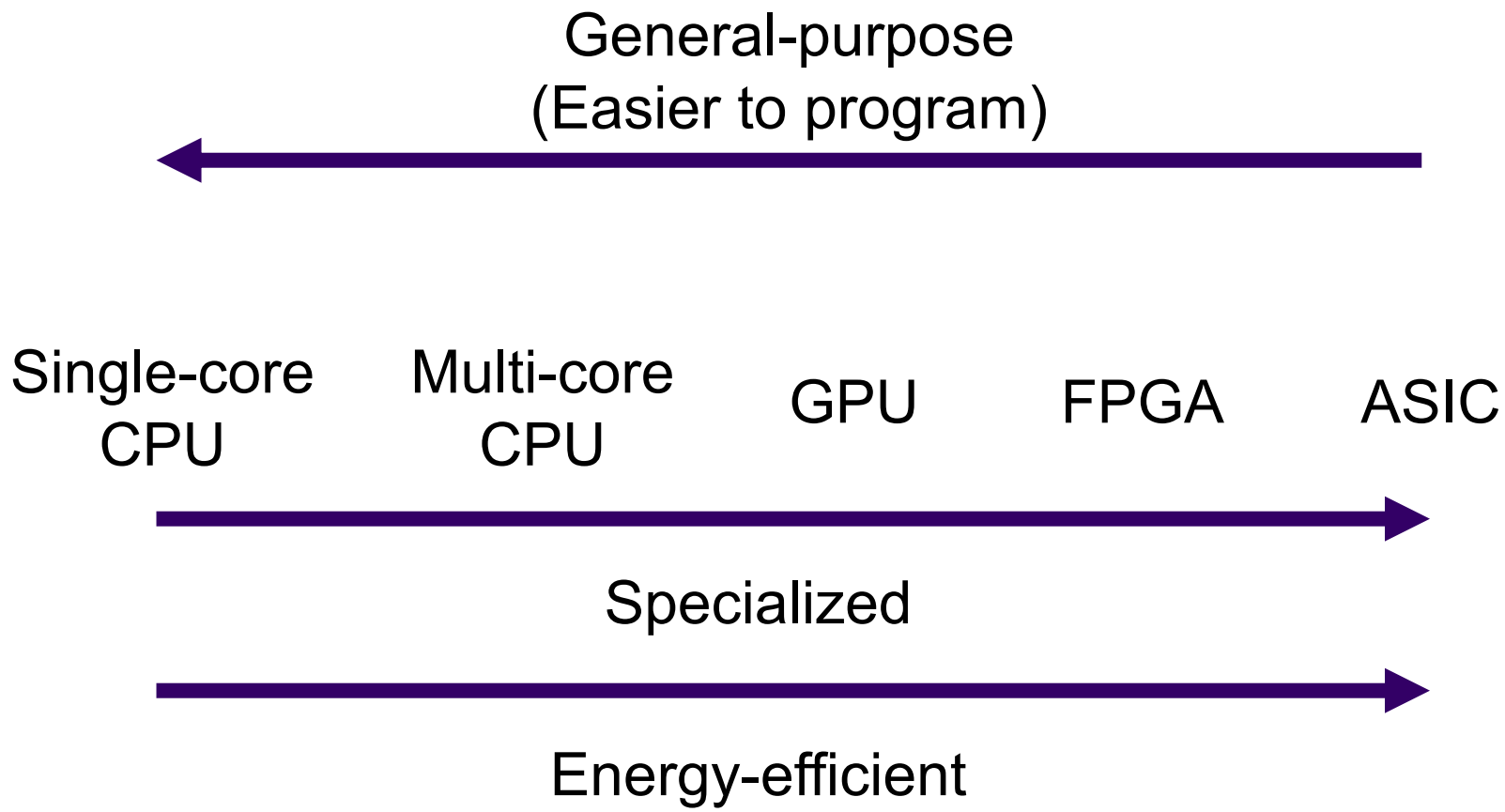

Surface of the Sun
Rocket Nozzle
Nuclear Reactor
Hot Plate
Intel Pentium 4
Intel Pentium III
Intel Pentium II
Intel Pentium Pro
Intel Pentium
Intel i486
Intel i386
1000
100
10
1
Watts/cm 2
1.5 1 0.7 0.5 0.35 0.25 0.18 0.13 0.1 0.07
Minumum IC Feature size in microns

Modern computer architecture is limited by:

a. Process Technology / Transistor density
   (End of Moore's Law)

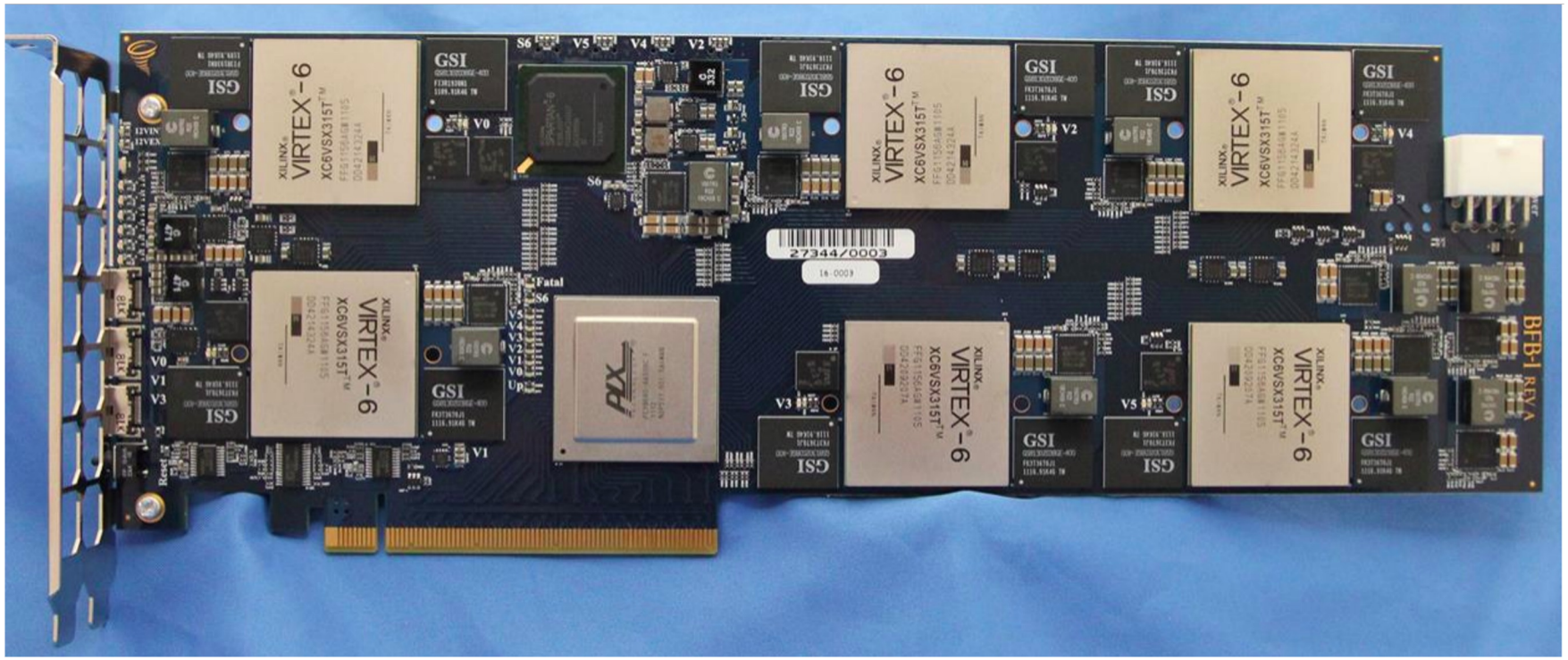
b. Power
   (End of Dennard Scaling)

c. Temperature


General-purpose
(Easier to program)
Single-core CPU
Multi-core CPU
GPU
FPGA
ASIC
Specialized
Energy-efficient

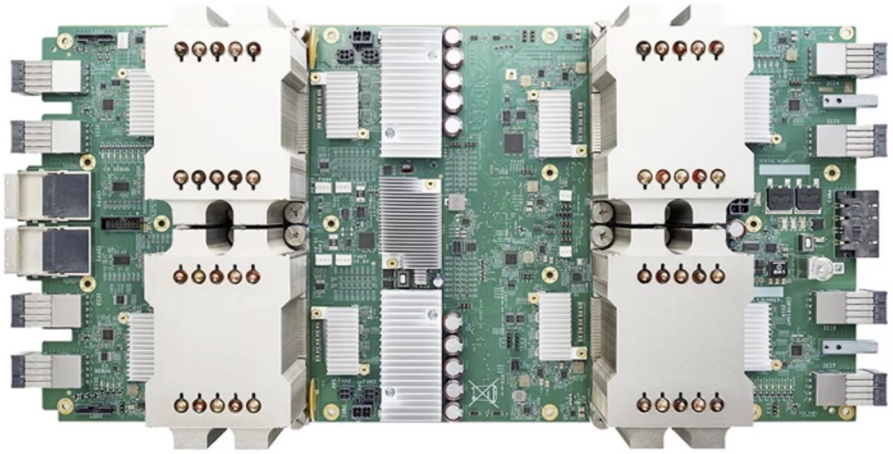
Examples of Specialization

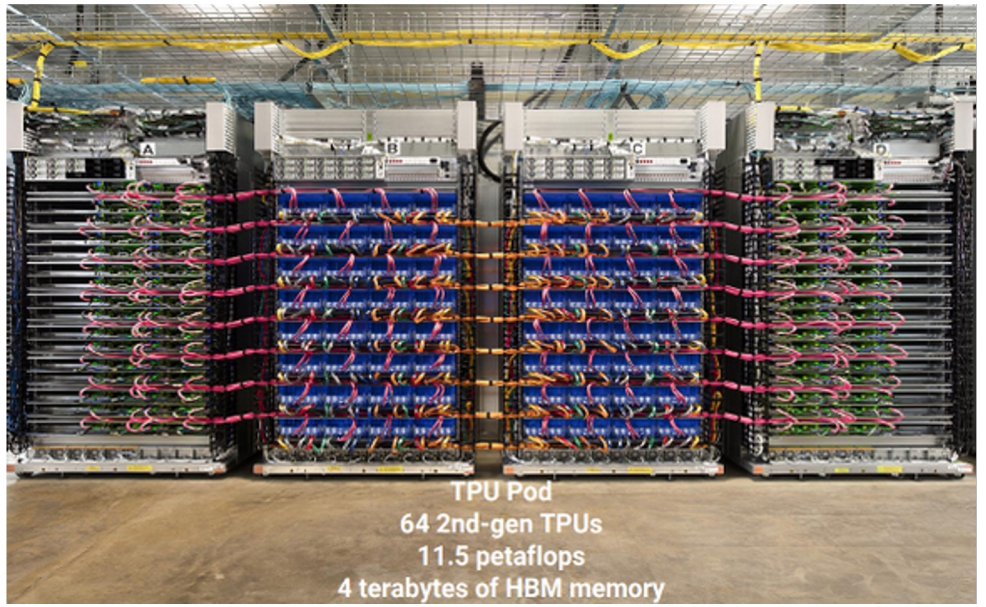
Microsoft Catapult



Examples of Specialization

Google TPU - Tensor Processing Unit



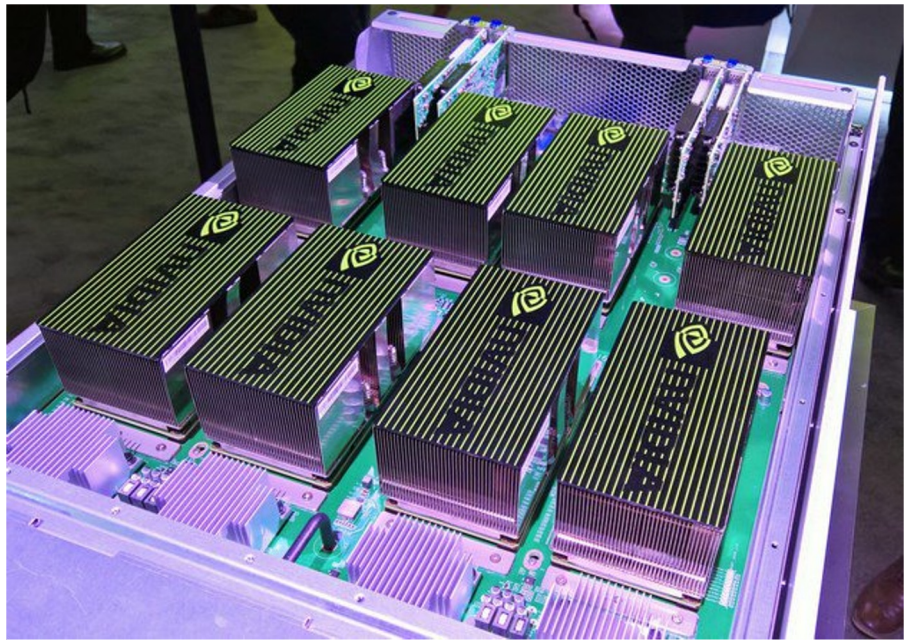


Examples of Specialization

Bitcoin Mining

Examples of Specialization

GPUs - Supercomputers



Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371

GPUs have wide applications in…

- Computer Science
- Robotics
- Computational Data Science
- Computer Engineering
- Electrical Engineering

QUESTIONS?