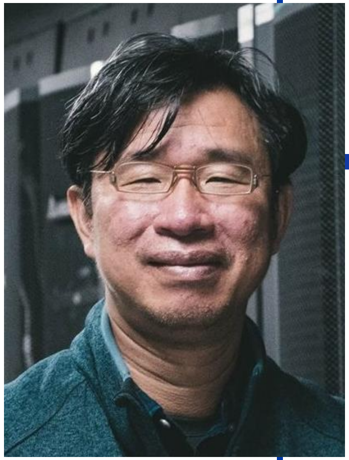


Introduction (2020)

Shih-Hao Hung, Ph.D.

Dept. of Computer Science and Information Engineering

National Taiwan University

Computer Organization

- Q: Why do you want to learn about computer organization?

- A1: To know how computers work
- A2: To learn the concept of engineering
- A3: To be able to design computer systems
- A4: To be able to optimize computer software
- A5: To become a good computer architect

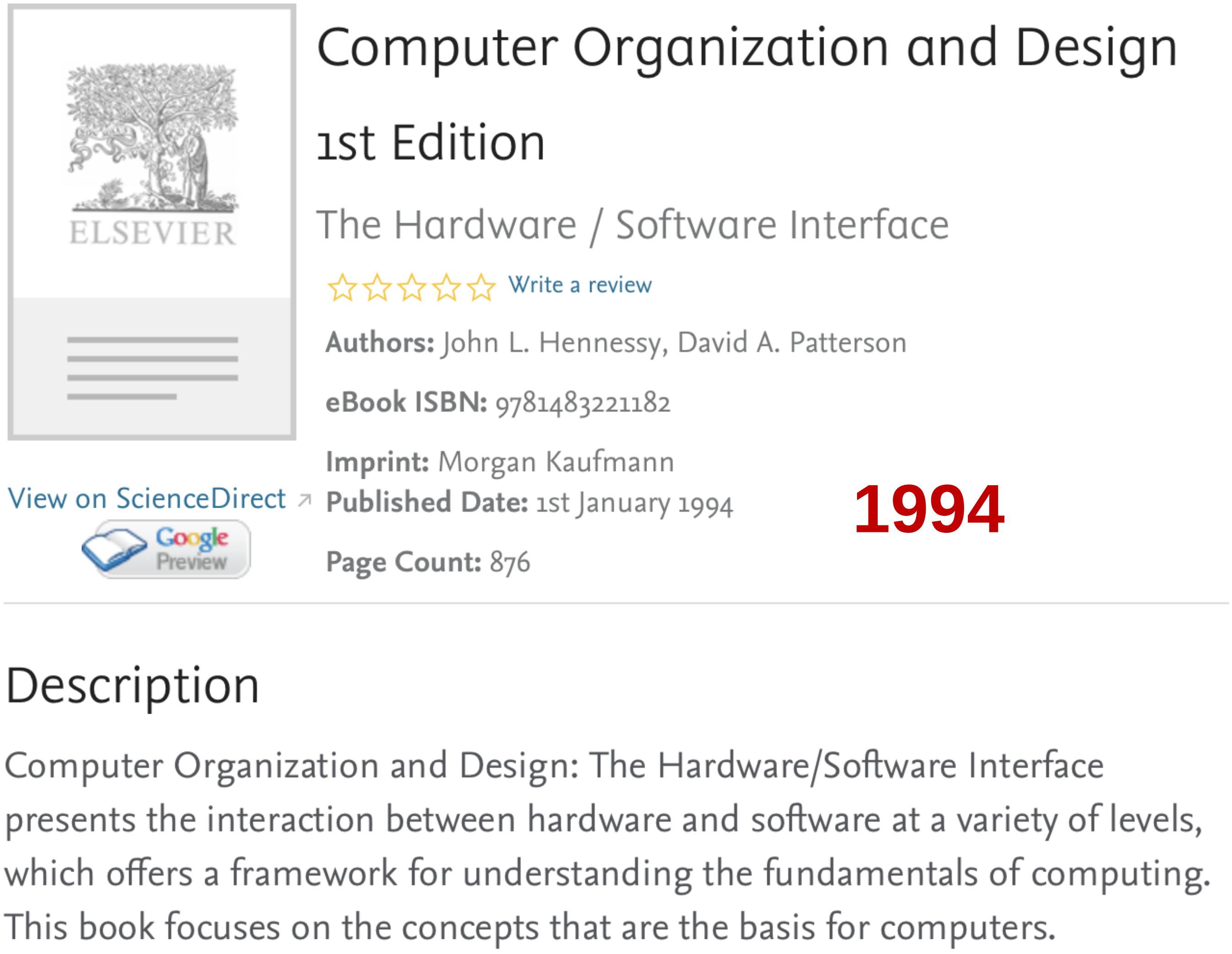
The Hardware/Software Interface

- Understanding the computer organization means that you are familiar with
  - The interface between hardware and software
  - The HW/SW components near the interface
  - The costs of the HW components
  - The interactions between these HW/SW components
  - The events occur in these HW/SW components
  - How these events affect the performance
  - How to improve the performance

Approach

- Modern computer organization is very complex
  - Design space is huge
  - Many important design principles
  - Real implementations to illustrate design principles
  - Apply the principles in design projects
- Computer performance is very complex, too
  - Can be affected by many factors
  - Few can be calculated with equations in real world
  - Tools and guidelines to improve performance with HW/SW techniques
- Lots of things to learn, but the more you learn, the faster you learn. Experience matters.

The Original Textbook



Computer Organization and Design

1st Edition

The Hardware / Software Interface

☆☆☆☆☆ Write a review

Authors: John L. Hennessy, David A. Patterson

eBook ISBN: 9781483221182

Imprint: Morgan Kaufmann

Published Date: 1st January 1994

Page Count: 876

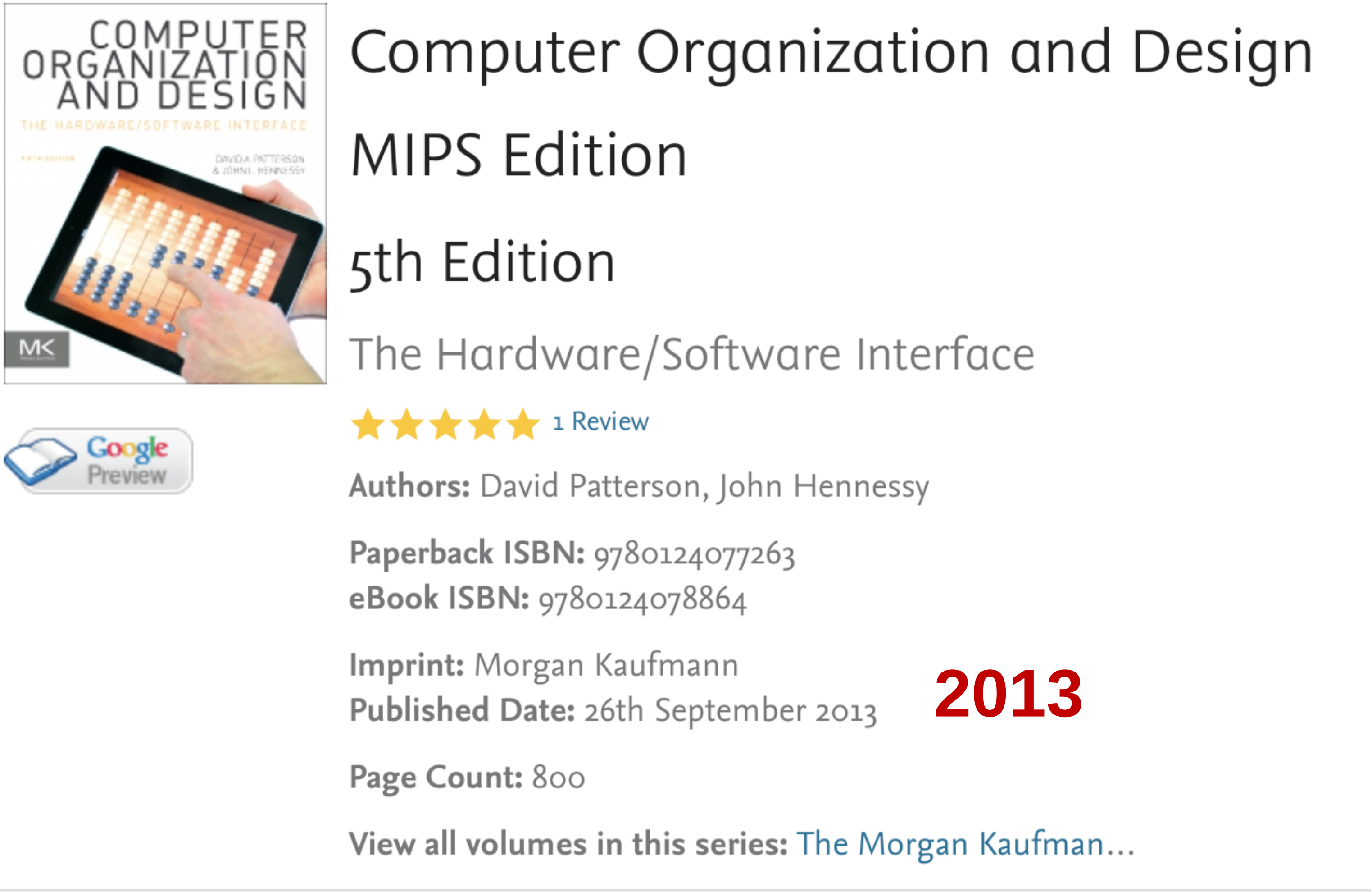
View on ScienceDirect ↗

Description

Computer Organization and Design: The Hardware/Software Interface presents the interaction between hardware and software at a variety of levels, which offers a framework for understanding the fundamentals of computing. This book focuses on the concepts that are the basis for computers.

1994

Fifth Edition in 19 Years!


COMPUTER ORGANIZATION AND DESIGN
THE HARDWARE/SOFTWARE INTERFACE
FIFTH EDITION
DAVID A. PATTERSON & JOHN L. HENNESSY
MK
Google Preview

Computer Organization and Design

MIPS Edition

5th Edition

The Hardware/Software Interface

★★★★★ 1 Review

Authors: David Patterson, John Hennessy

Paperback ISBN: 9780124077263
eBook ISBN: 9780124078864

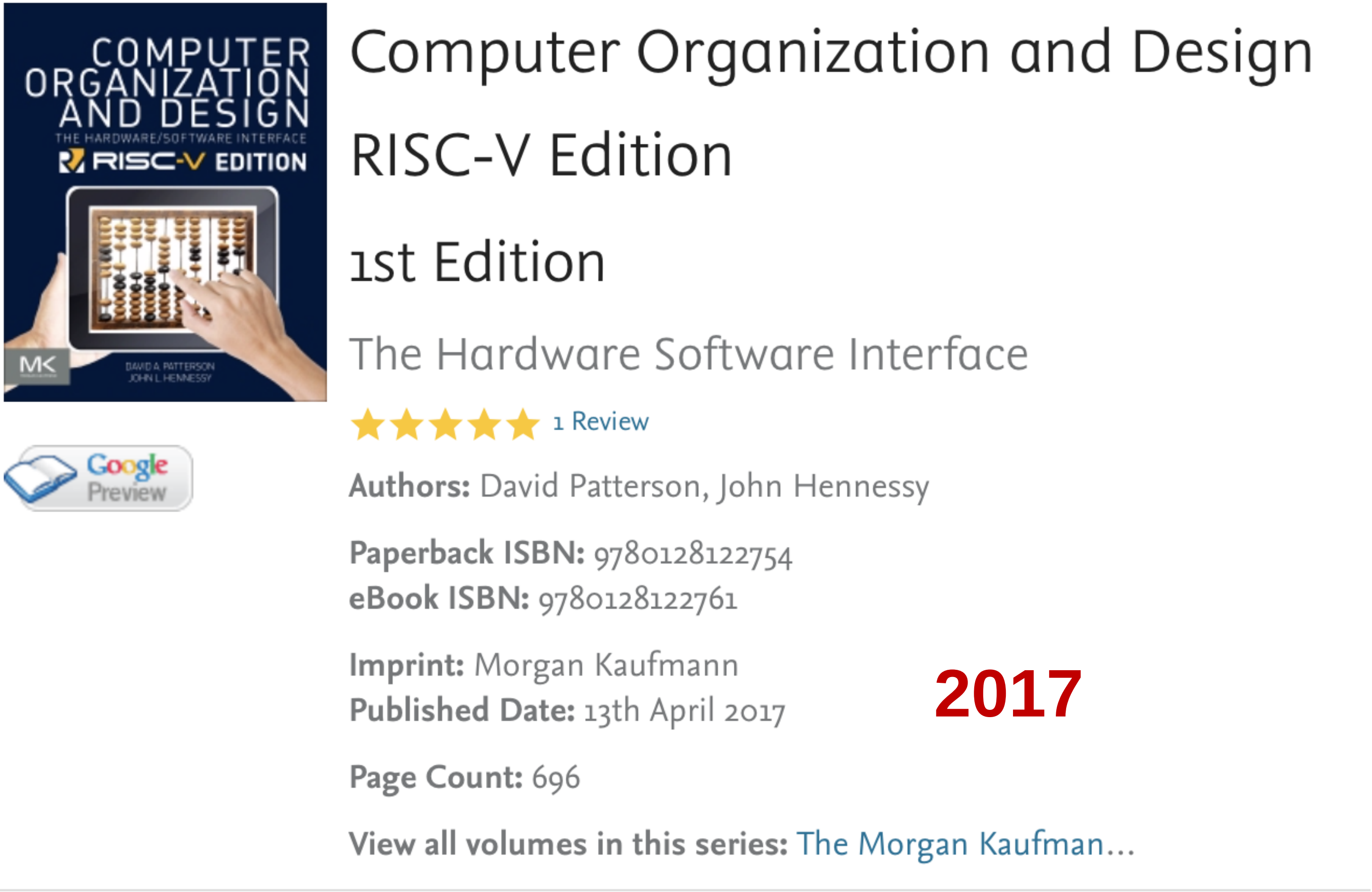
Imprint: Morgan Kaufmann
Published Date: 26th September 2013

Page Count: 800

View all volumes in this series: The Morgan Kaufman…

2013

RISC-V Edition (~5th Ed.)



Computer Organization and Design

RISC-V Edition

1st Edition

The Hardware Software Interface

★★★★★ 1 Review

Authors: David Patterson, John Hennessy

Paperback ISBN: 9780128122754
eBook ISBN: 9780128122761

Imprint: Morgan Kaufmann
Published Date: 13th April 2017

Page Count: 696

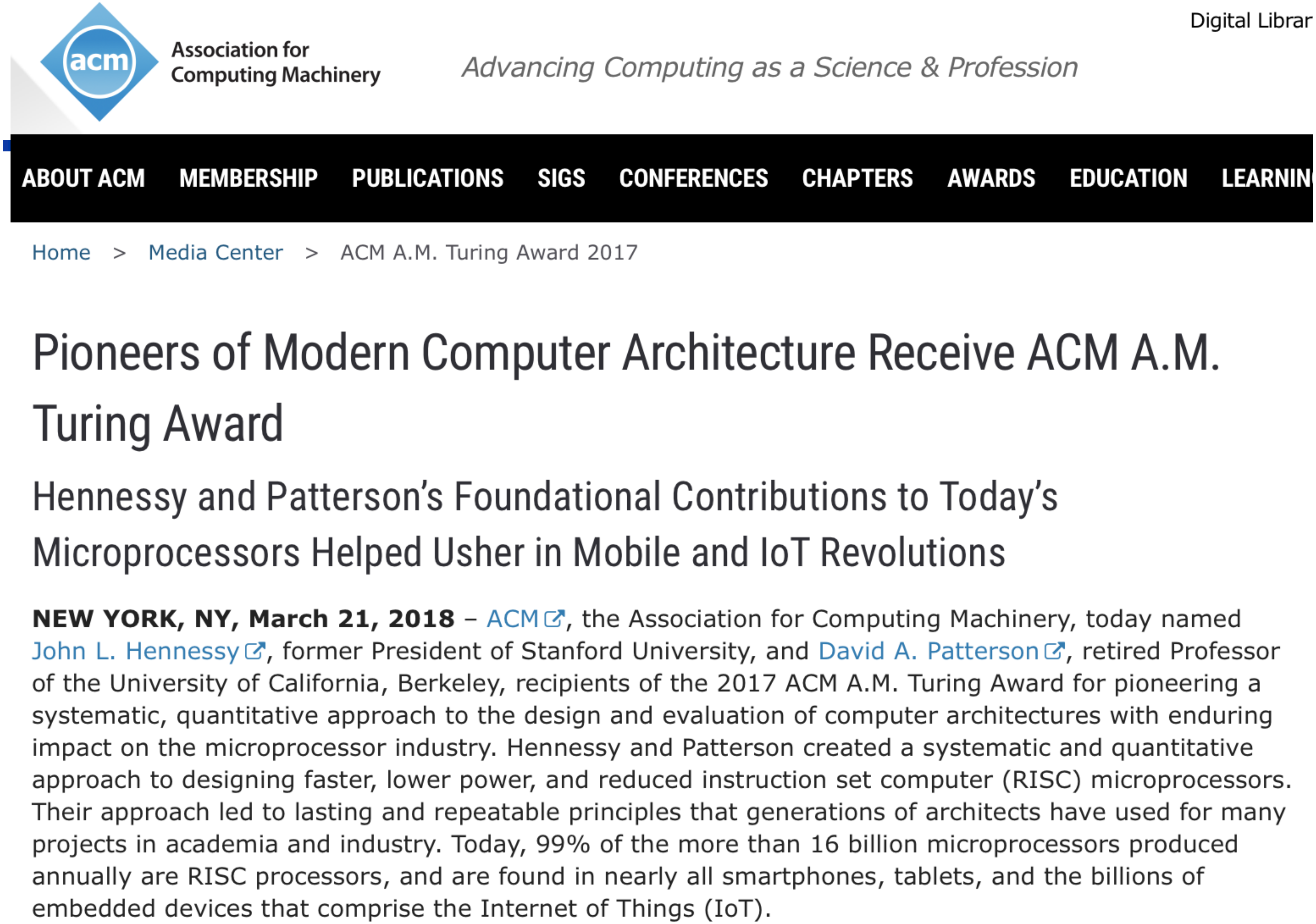
View all volumes in this series: The Morgan Kaufman…

2017

About the Authors

- David Patterson
  - https://en.wikipedia.org/wiki/David_Patterson_(computer_scientist)
  - https://www.eecs.berkeley.edu/Faculty/Homepages/patterson.html

- John Hennessy
  - https://en.wikipedia.org/wiki/John_L._Hennessy
  - http://web.stanford.edu/~hennessy/

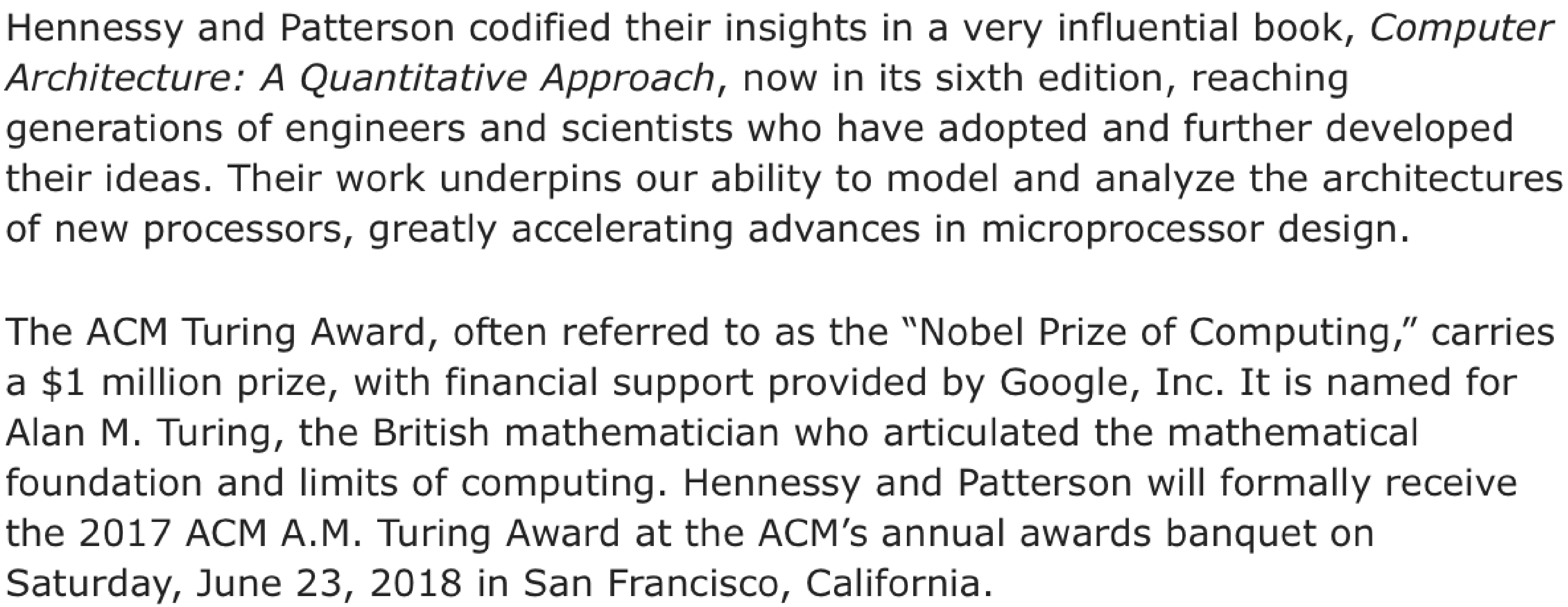

Association for Computing Machinery
Advancing Computing as a Science & Profession
Digital Librar
ABOUT ACM
MEMBERSHIP
PUBLICATIONS
SIGS
CONFERENCES
CHAPTERS
AWARDS
EDUCATION
LEARNIN
Home > Media Center > ACM A.M. Turing Award 2017

Pioneers of Modern Computer Architecture Receive ACM A.M. Turing Award

Hennessy and Patterson's Foundational Contributions to Today's Microprocessors Helped Usher in Mobile and IoT Revolutions

NEW YORK, NY, March 21, 2018 – ACM, the Association for Computing Machinery, today named John L. Hennessy, former President of Stanford University, and David A. Patterson, retired Professor of the University of California, Berkeley, recipients of the 2017 ACM A.M. Turing Award for pioneering a systematic, quantitative approach to the design and evaluation of computer architectures with enduring impact on the microprocessor industry. Hennessy and Patterson created a systematic and quantitative approach to designing faster, lower power, and reduced instruction set computer (RISC) microprocessors. Their approach led to lasting and repeatable principles that generations of architects have used for many projects in academia and industry. Today, 99% of the more than 16 billion microprocessors produced annually are RISC processors, and are found in nearly all smartphones, tablets, and the billions of embedded devices that comprise the Internet of Things (IoT).

ACM Turing Award



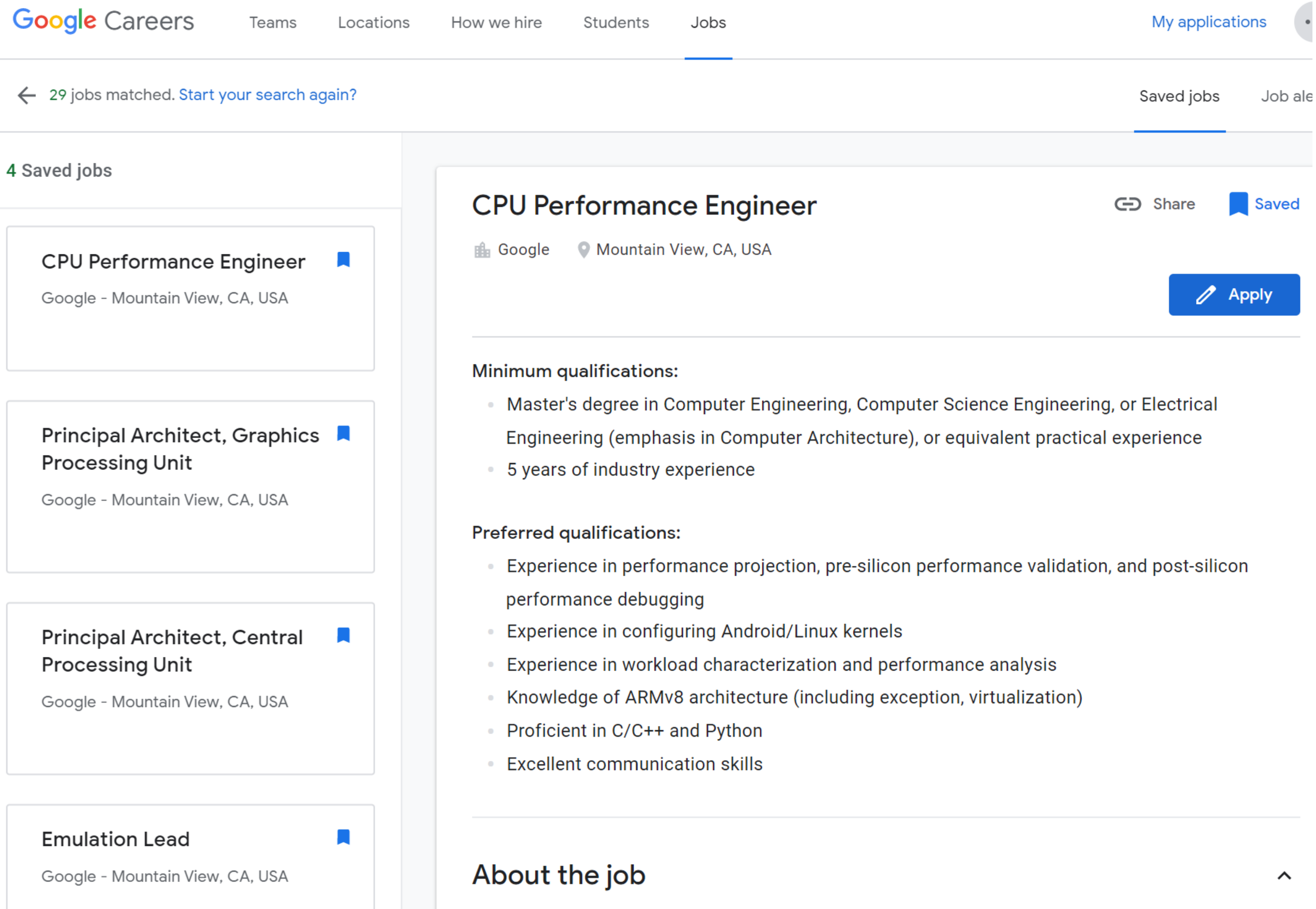
Hennessy and Patterson codified their insights in a very influential book, *Computer Architecture: A Quantitative Approach*, now in its sixth edition, reaching generations of engineers and scientists who have adopted and further developed their ideas. Their work underpins our ability to model and analyze the architectures of new processors, greatly accelerating advances in microprocessor design.

The ACM Turing Award, often referred to as the "Nobel Prize of Computing," carries a $1 million prize, with financial support provided by Google, Inc. It is named for Alan M. Turing, the British mathematician who articulated the mathematical foundation and limits of computing. Hennessy and Patterson will formally receive the 2017 ACM A.M. Turing Award at the ACM's annual awards banquet on Saturday, June 23, 2018 in San Francisco, California.

Job Opportunities

- If you are really good, jobs are abundant
- Well-trained computer architects are wanted in the industry
- Not just the knowledge, but also
  - Practical skills
  - Creativity
  - Methodology
  - Communication skills
  - …

https://careers.google.com/jobs/



29 jobs matched. Start your search again?

4 Saved jobs

CPU Performance Engineer
Google - Mountain View, CA, USA

Principal Architect, Graphics Processing Unit
Google - Mountain View, CA, USA

Principal Architect, Central Processing Unit
Google - Mountain View, CA, USA

Emulation Lead
Google - Mountain View, CA, USA

CPU Performance Engineer

Google Mountain View, CA, USA

Minimum qualifications:

- Master's degree in Computer Engineering, Computer Science Engineering, or Electrical Engineering (emphasis in Computer Architecture), or equivalent practical experience
- 5 years of industry experience

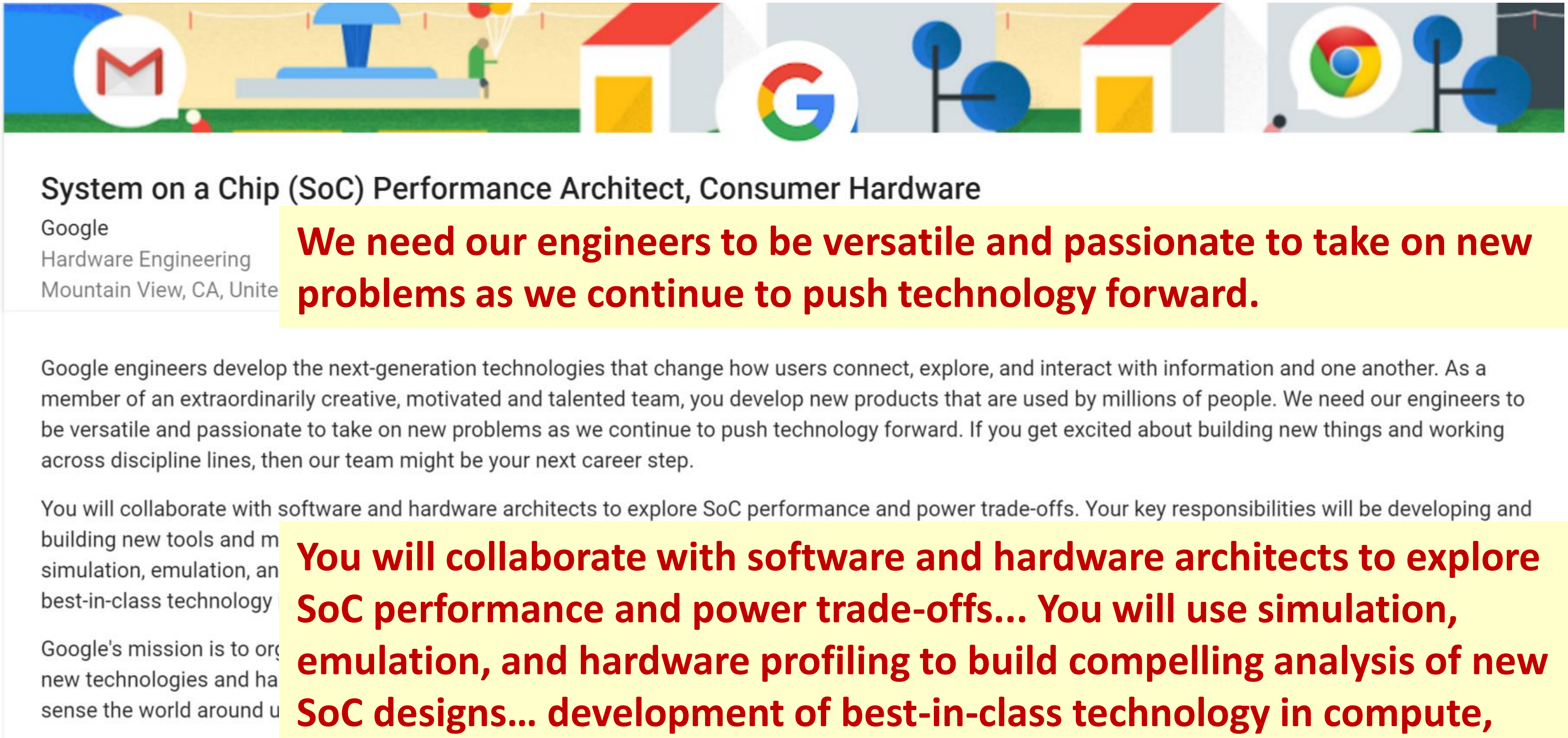
Preferred qualifications:

- Experience in performance projection, pre-silicon performance validation, and post-silicon performance debugging
- Experience in configuring Android/Linux kernels
- Experience in workload characterization and performance analysis
- Knowledge of ARMv8 architecture (including exception, virtualization)
- Proficient in C/C++ and Python
- Excellent communication skills

About the job

Performance Architect

https://careers.google.com/jobs#!t=jo&jid=/google/system-on-a-chip-soc-performance-1600-amphitheatre-pkwy-mountain-view-ca-3407250819&f=true&



System on a Chip (SoC) Performance Architect, Consumer Hardware

Google
Hardware Engineering
Mountain View, CA, Unite

Google engineers develop the next-generation technologies that change how users connect, explore, and interact with information and one another. As a member of an extraordinarily creative, motivated and talented team, you develop new products that are used by millions of people. We need our engineers to be versatile and passionate to take on new problems as we continue to push technology forward. If you get excited about building new things and working across discipline lines, then our team might be your next career step.

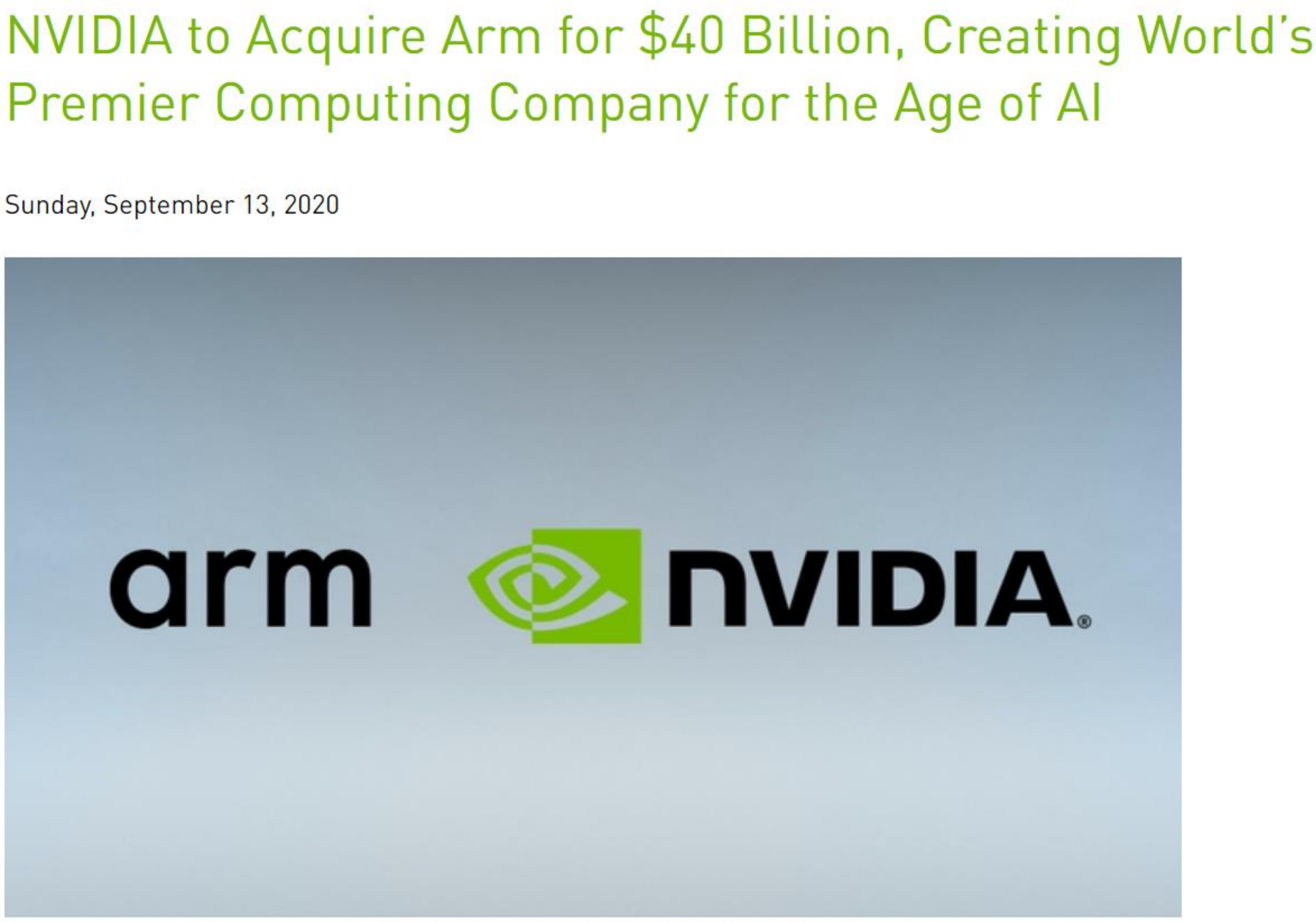
You will collaborate with software and hardware architects to explore SoC performance and power trade-offs. Your key responsibilities will be developing and
building new tools and m
simulation, emulation, an
best-in-class technology

Google's mission is to org
new technologies and ha
sense the world around u

We need our engineers to be versatile and passionate to take on new problems as we continue to push technology forward.

You will collaborate with software and hardware architects to explore SoC performance and power trade-offs... You will use simulation, emulation, and hardware profiling to build compelling analysis of new SoC designs… development of best-in-class technology in compute, media, fabric, memory, etc., and filing associated patents.

Recent Industrial Trends



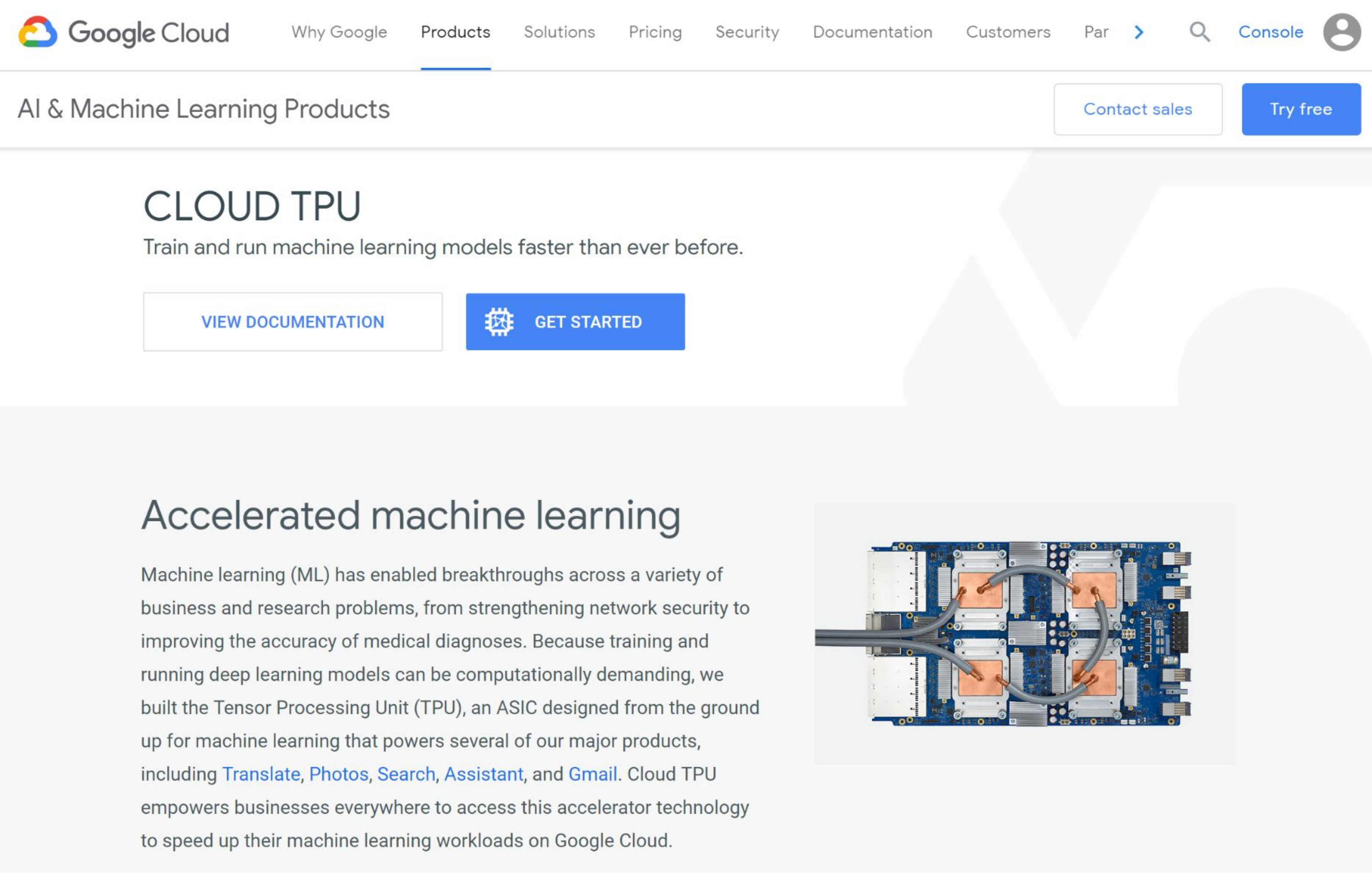
NVIDIA to Acquire Arm for $40 Billion, Creating World's Premier Computing Company for the Age of AI

Sunday, September 13, 2020

Deep Learning x Computer Architecture

- ***Deep Learning*** and ***Deep Neural Network*** were nearly impossible in ‘90s, due to **intensive computational requirements**
- Even fastest multicore CPUs in early ‘00s could not satisfy the requirements
- **GPUs** came to rescue in late ‘00s, reducing training time from weeks to days
- Google designed a special-purpose processor, called **TPU**, to make deep learning far more **affordable**
- Apple, Qualcomm, Huawei, etc. also designed special-purpose processors, a.k.a. **neural engines**, for mobile phones

Core Technology#1: Processor Architecture



AI & Machine Learning Products

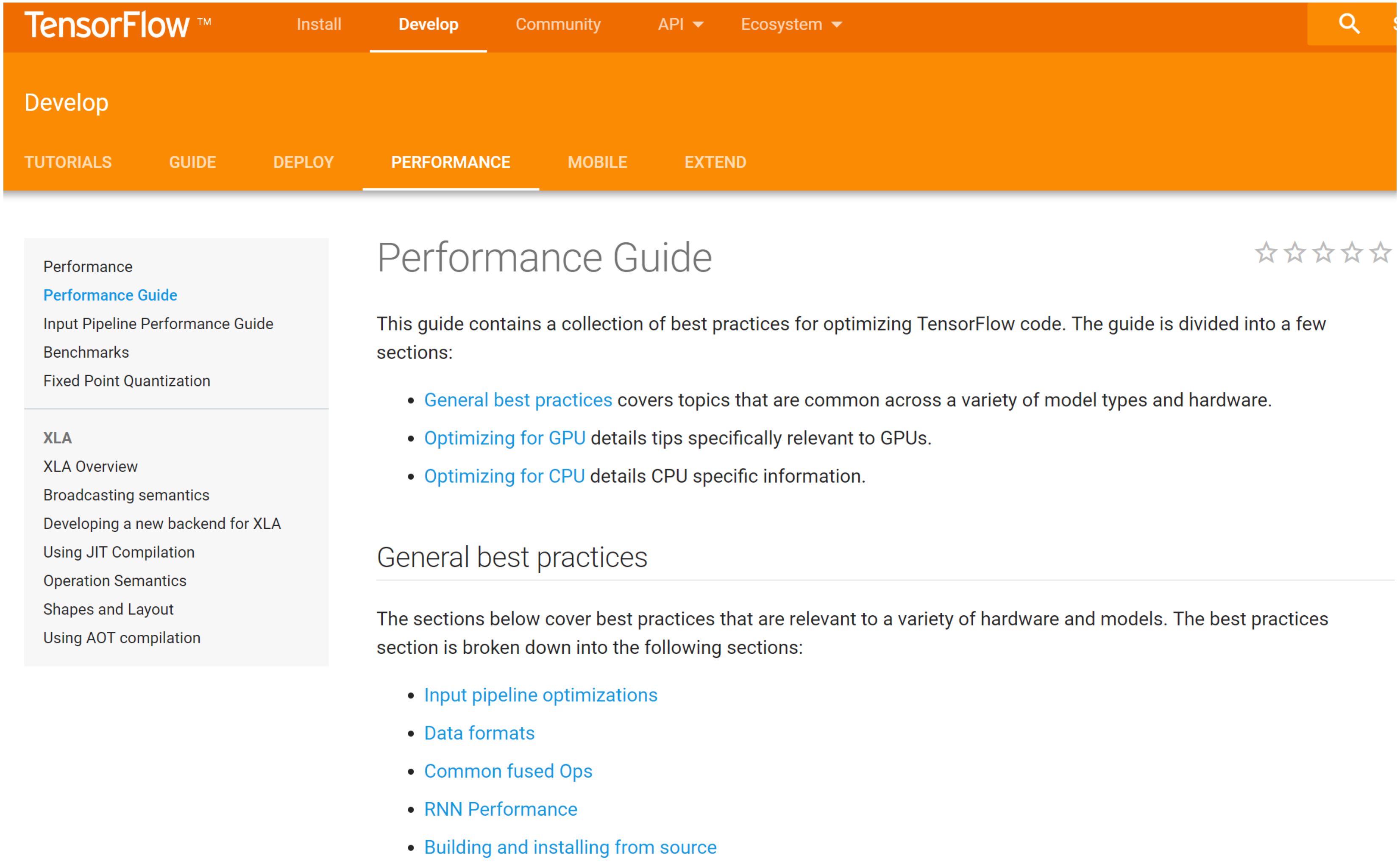
CLOUD TPU

Train and run machine learning models faster than ever before.

Accelerated machine learning

Machine learning (ML) has enabled breakthroughs across a variety of business and research problems, from strengthening network security to improving the accuracy of medical diagnoses. Because training and running deep learning models can be computationally demanding, we built the Tensor Processing Unit (TPU), an ASIC designed from the ground up for machine learning that powers several of our major products, including Translate, Photos, Search, Assistant, and Gmail. Cloud TPU empowers businesses everywhere to access this accelerator technology to speed up their machine learning workloads on Google Cloud.

Core Technology#2: Software Optimization



Performance
Performance Guide
Input Pipeline Performance Guide
Benchmarks
Fixed Point Quantization

XLA
XLA Overview
Broadcasting semantics
Developing a new backend for XLA
Using JIT Compilation
Operation Semantics
Shapes and Layout
Using AOT compilation

Performance Guide

☆☆☆☆☆

This guide contains a collection of best practices for optimizing TensorFlow code. The guide is divided into a few sections:

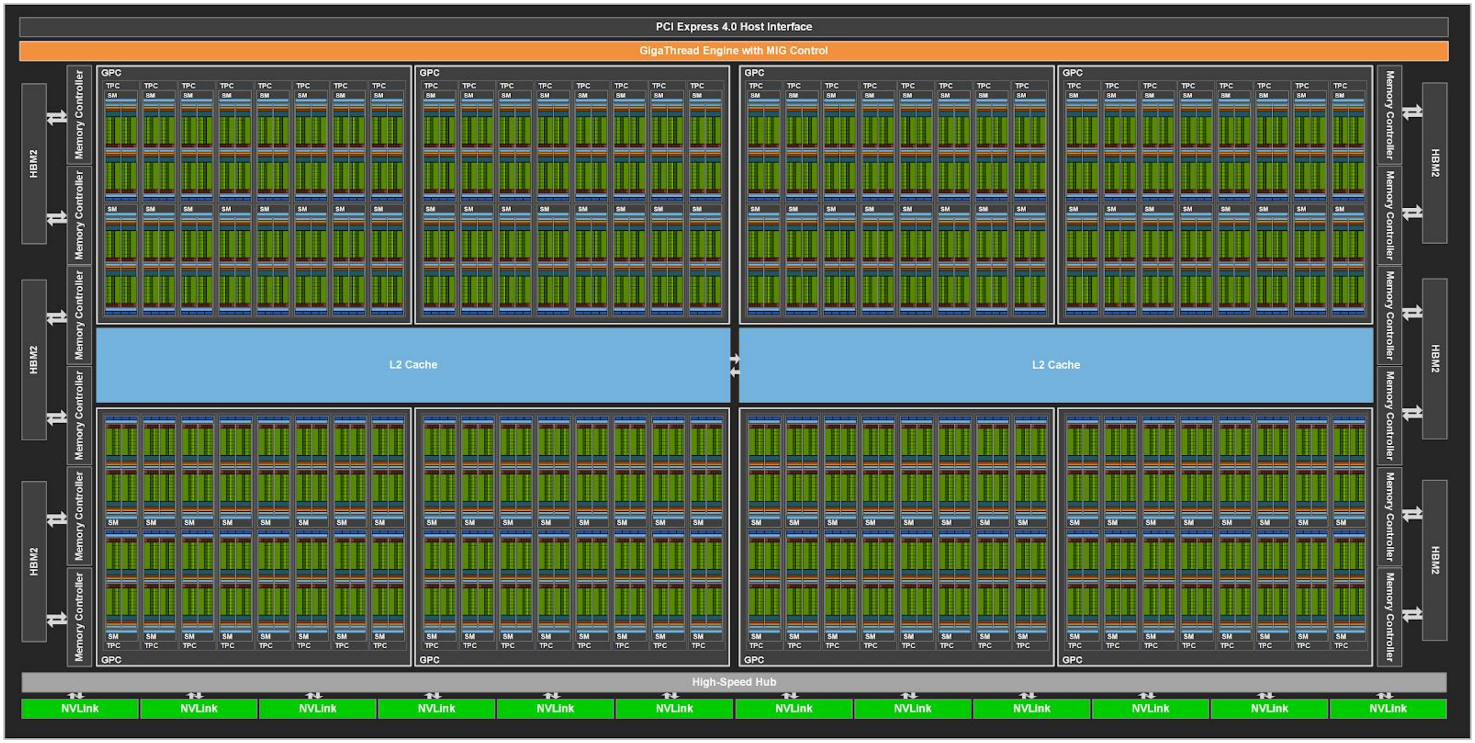
- General best practices covers topics that are common across a variety of model types and hardware.
- Optimizing for GPU details tips specifically relevant to GPUs.
- Optimizing for CPU details CPU specific information.

General best practices

The sections below cover best practices that are relevant to a variety of hardware and models. The best practices section is broken down into the following sections:

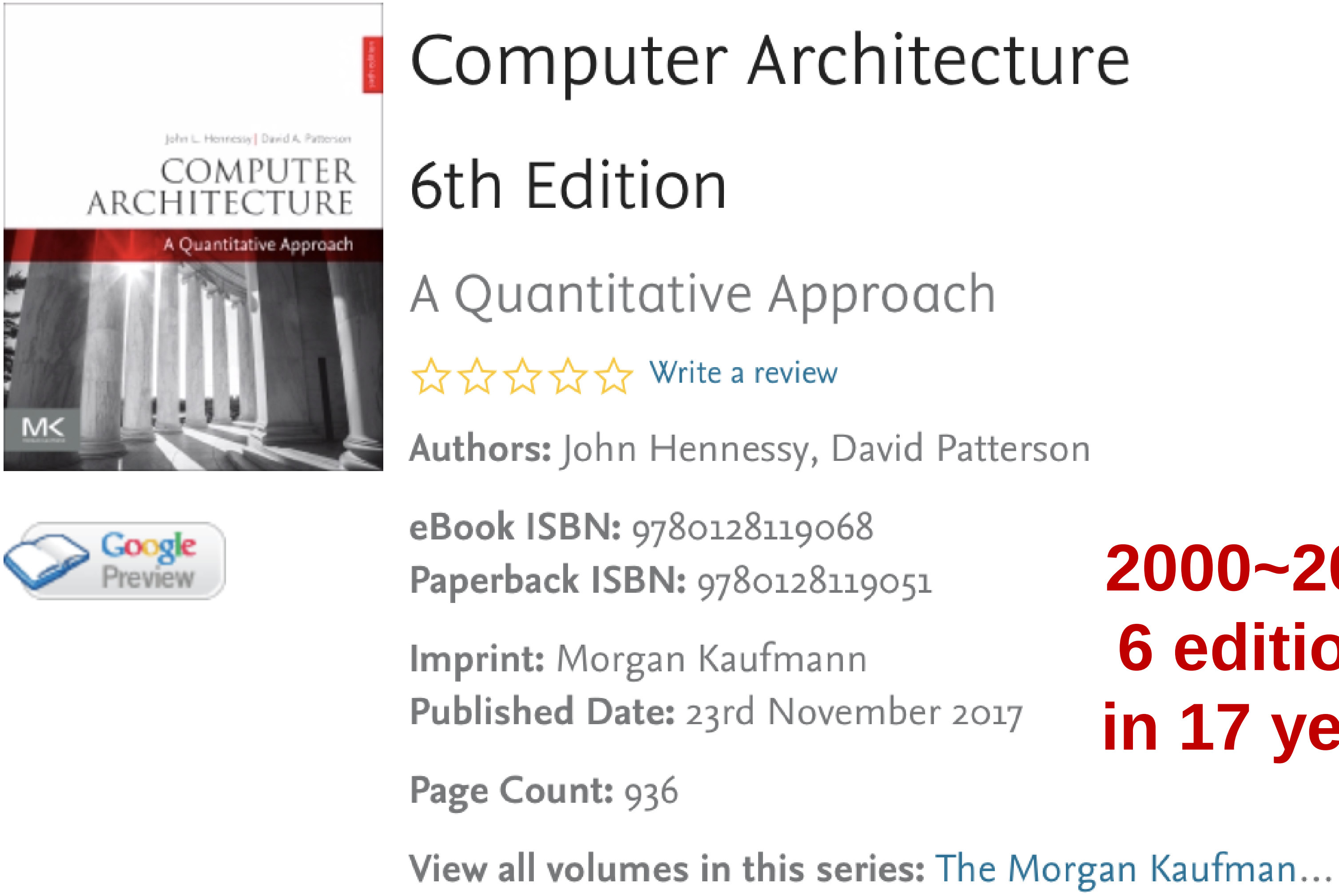
- Input pipeline optimizations
- Data formats
- Common fused Ops
- RNN Performance
- Building and installing from source

Core Technology#3: Parallel Computing


PCI Express 4.0 Host Interface
GigaThread Engine with MIG Control
GPC
TPC
SM
L2 Cache
Memory Controller
HBM2
High-Speed Hub
NVLink

https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

Advanced Contents



Computer Architecture

6th Edition

A Quantitative Approach

☆☆☆☆☆ Write a review

Authors: John Hennessy, David Patterson

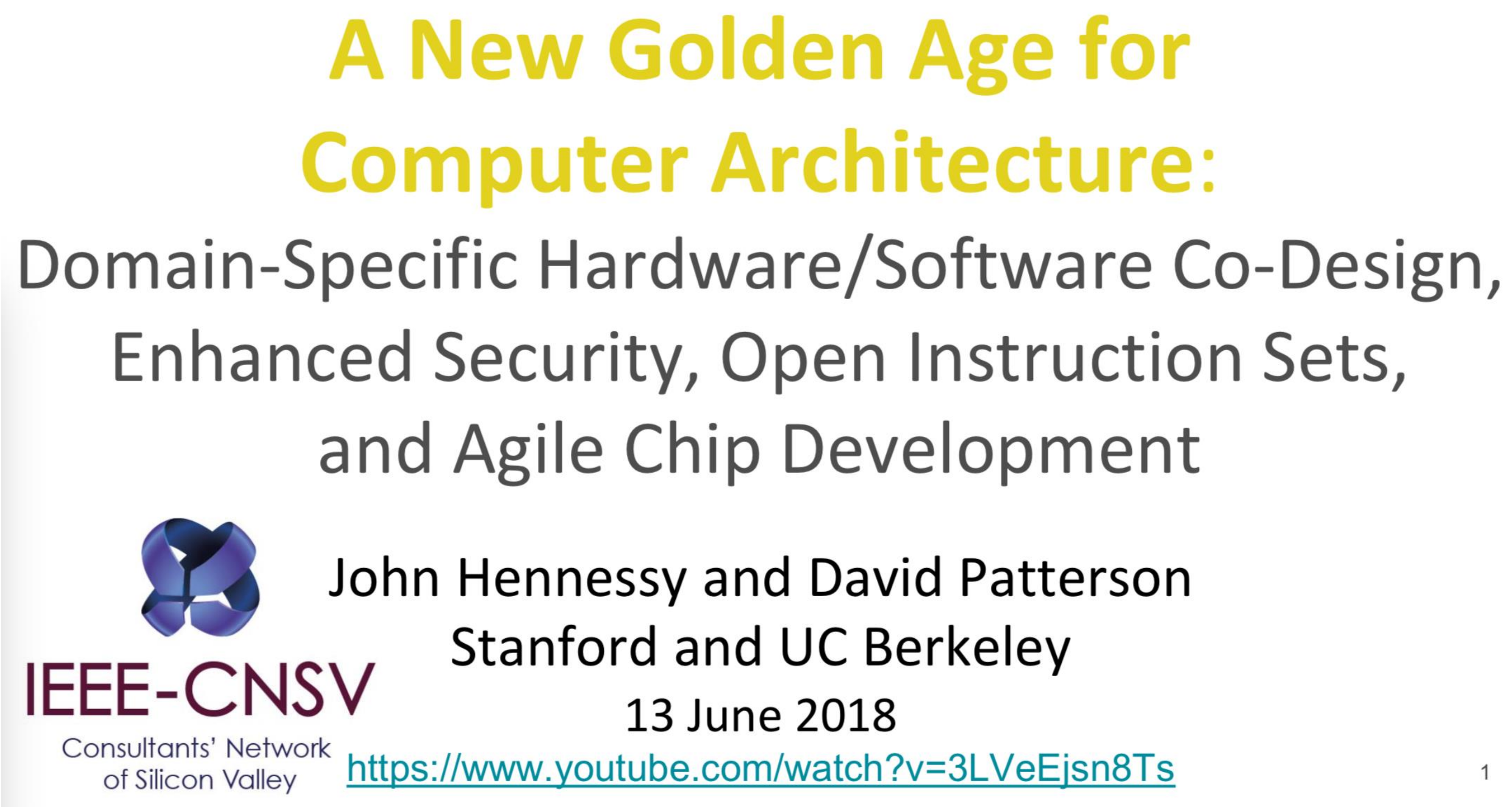
eBook ISBN: 9780128119068
Paperback ISBN: 9780128119051

Imprint: Morgan Kaufmann
Published Date: 23rd November 2017

Page Count: 936

View all volumes in this series: The Morgan Kaufman…

2000~2017
6 editions
in 17 years



A New Golden Age for Computer Architecture:

Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

IEEE-CNSV
Consultants' Network
of Silicon Valley

John Hennessy and David Patterson
Stanford and UC Berkeley

13 June 2018

https://www.youtube.com/watch?v=3LVeEjsn8Ts

1

https://amturing.acm.org/vp/patterson_2316693.cfm

DAVID PATTERSON

United States – 2017

Lecture Video

Course Outline

- CEIBA website: https://ceiba.ntu.edu.tw/1091CSIE3340_02
- Grading:
  - Assignments: 50%
  - Midterm Exam: 20%
  - Final Exam: 30%
- Office hour: Mondays, 15:00-17:00

QURESTIONS?