What a great time to be a student in computing!

Wen-mei Hwu

Professor and Sanders-AMD Chair, ECE, NCSA

University of Illinois at Urbana-Champaign

ECE ILLINOIS

The IMPACT Research Group

NCSA

ILLINOIS

Agenda

- Revolutionary paradigm shift in applications
- An commercial example of positive application-technology spiral
- Post-Dennard technology pivot - heterogeneity
- Lessons learned and outlook

A major paradigm shift

- **In the 20th Century, we were able to understand, design, and manufacture what we can measure**
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, understand natural processes, control artificial processes…

A major paradigm shift

- In the 20th Century, we were able to understand, design, and manufacture what we can measure
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, understand natural processes, control artificial processes…
- **In the 21st Century, we are able to understand, design, and create what we can compute**
  - Computational models are allowing us to see even farther, going back and forth in time, learn better, test hypothesis that cannot be verified any other way, create safe artificial processes…

Examples of Paradigm Shift

20th Century

- Small mask patterns
- Electronic microscope and Crystallography with computational image processing
- Anatomic imaging with computational image processing
- Teleconference
- GPS

21st Century

- Optical proximity correction
- Computational microscope with initial conditions from Crystallography
- Metabolic imaging sees disease before visible anatomic change
- Tele-emersion
- Self-driving cars

What is causing the paradigm shift?

- Most of the methods for the new paradigm have existed for decades
  - But they are considered to be too expensive and thus impractical
- The memory capacity and computing throughput improved by more than a million fold in the past three decades
  - Many of these methods have just become practical

- That is, the field of computing has finally grown up!

Diving deeper into computational microscope

- Large clusters (scale out) allow simulation of biological systems of realistic space dimensions
  - 0.5Å (0.05 nm) lattice spacing needed for accuracy
  - Interesting biological systems have dimensions of mm or larger
  - Thousands of nodes are required to hold and update all the grid points.

- Fast nodes (scale up) allow simulation at realistic time scales
  - Simulation time steps at femtosecond (10^{-15} second) level needed for accuracy
  - Biological processes take milliseconds or longer
  - Current molecular dynamics simulations progress at about one day for each 10-100 microseconds of the simulated process.

Blue Waters Science Breakthrough Example

- Determination of the structure of the HIV capsid at atomic-level
- Collaborative effort of experimental groups at the U. of Pittsburgh and Vanderbilt U., and the Schulten's computational team at the U. of Illinois.
- 64-million-atom HIV capsid simulation of the process through which the capsid disassembles, releasing its genetic material
- a critical step in understanding HIV infection and finding a target for antiviral drugs.

A commercial example of positive application-technology spiral

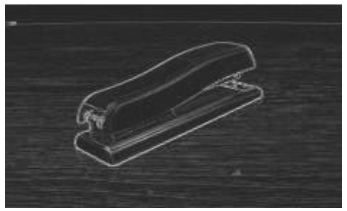
Machine Learning

- An important way of building applications whose logic is not fully understood.
  - Use labeled data – data that come with the input values and their desired output values – to learn what the logic should be produce
  - Capture each labeled data item by adjusting the program logic
  - Learn by example!
- Training Phase
  - The system learns the logic for the application from labeled data.
- Deployment (inference) Phase
  - The system applies the learned program logic in processing data

DIFFERENT MODALITIES OF REAL-WORLD DATA

Images/video

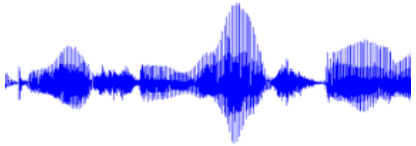
Image



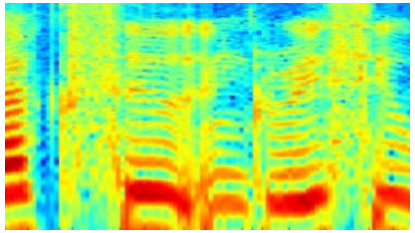
Vision features

Detection

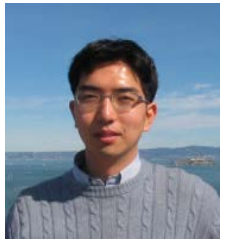
Audio



Audio



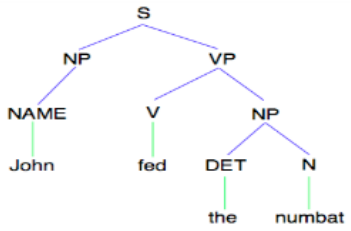
Audio features



Speaker ID

Text

Text



Text features

Text classification, machine translation, information retrieval, ...

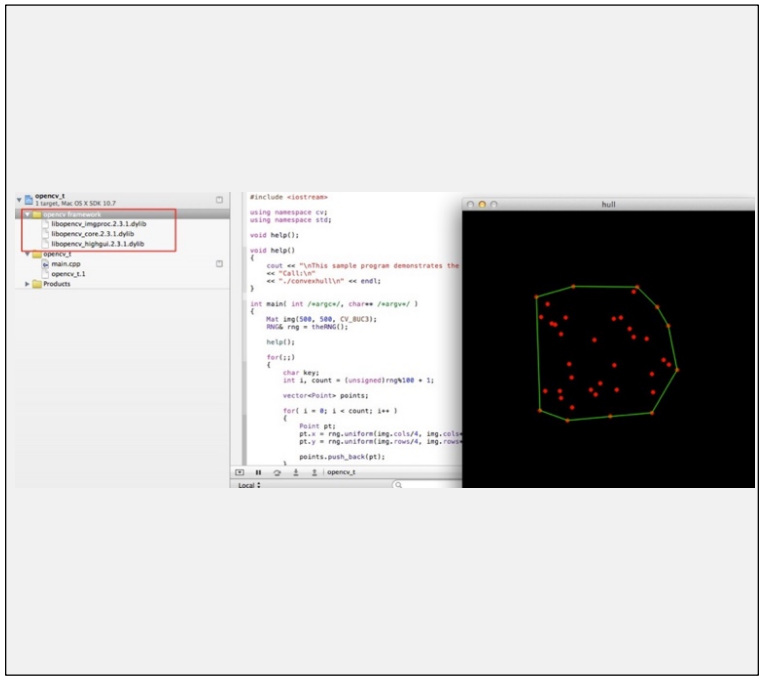
Slide courtesy of Andrew Ng, Stanford University

Recent Explosion of Deep Learning Applications

- GPU computing hardware and programming interfaces such as CUDA has enabled effective use of massive data sets in very fast research cycle of deep neural net training

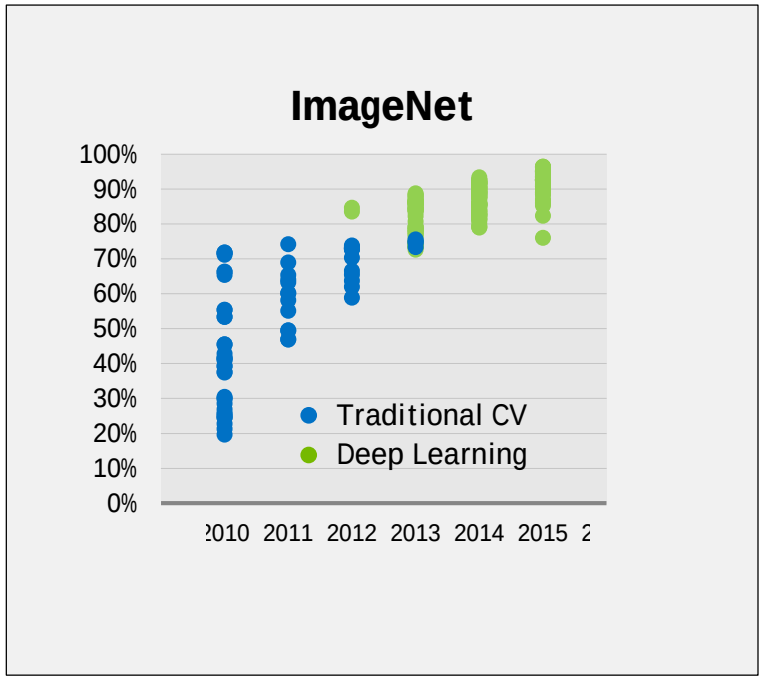
- Using big labeled data to train and specialize DNN based classifiers
  - Deriving a large quantity of quality labeled data is a challenge

DEEP LEARNING IN COMPUTER VISION



Traditional Computer Vision
Experts + Time

Deep Learning Object Detection
DNN + Data + HPC



Deep Learning Achieves
"Superhuman" Results

Slide courtesy of Steve Oberlin, NVIDIA

Behind the Scenes

- In 2010 Prof. Andreas Moshovos at University of Toronto adopted the ECE498AL Programming Massively Parallel Programming Class
- Several of Prof. Geoffrey Hinton's graduate students took the course
- These students developed the GPU implementation of the DNN that was trained with 1.2M images to win the ImageNet competition in 2012

A long way to go towards cognitive computing

▼ **Social Sciences**

Use the cartoon to answer the next TWO questions.

24

What economic condition motivated Phil to request a raise?

A. Inflation
B. Specialization
C. Unemployment
D. Embargo

25

Without his raise, which would typify Phil's behavior in the marketplace?

A. He will increase his interest for higher priced items.
B. He will increase his demand for higher priced items.
C. He will decrease his demand for lower priced substitutes.
D. He will increase his demand for lower priced substitutes.

Taken from http://www.ode.state.
chlearn/testing/samp
ocsci_sampletest_en

Human Instructions

Image Recognition

Text Extraction

Speech Recognition

Diagram Understanding

Natural Language Processing

Knowledge Indexing

IR

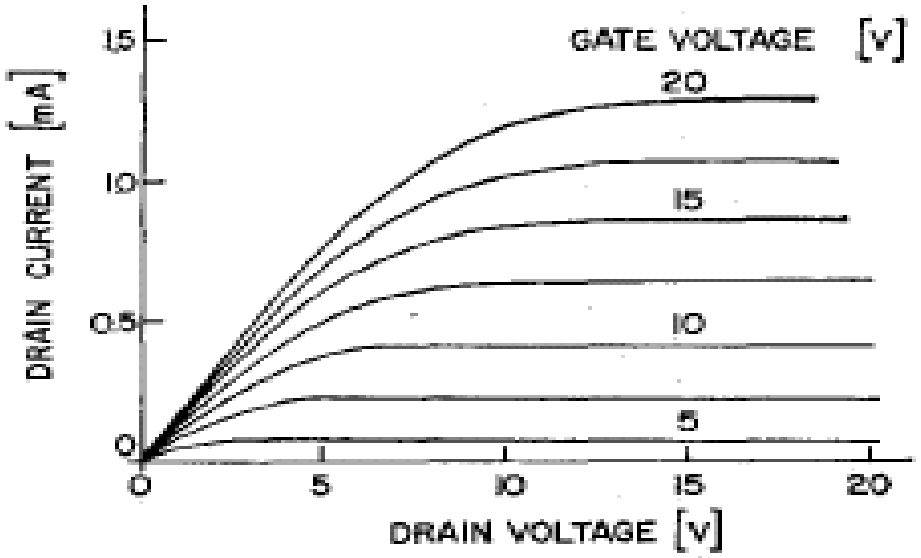
Knowledge Inferencing

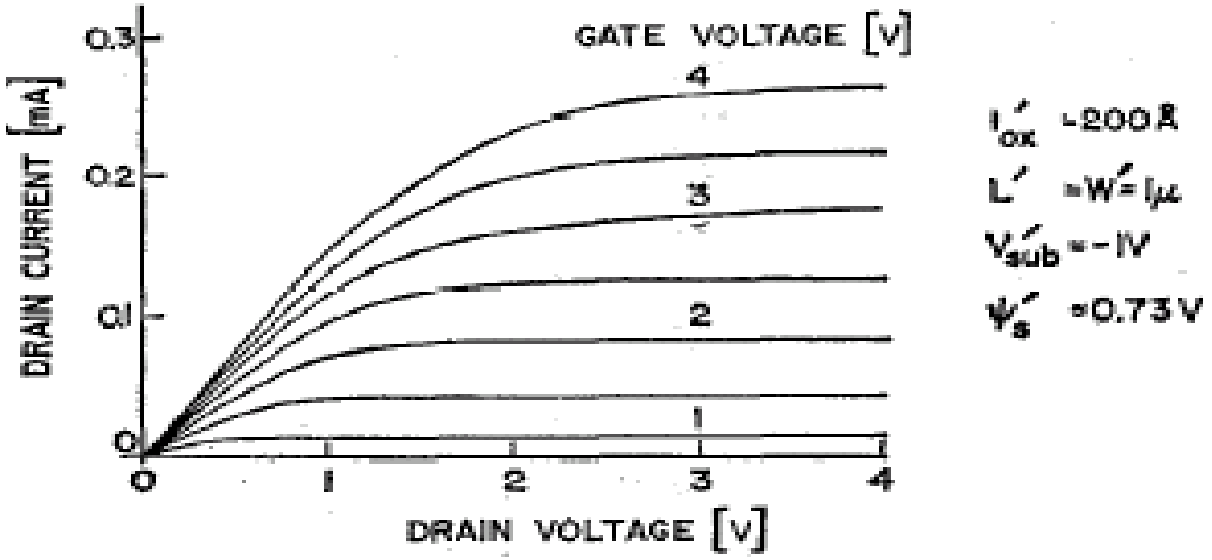
Programming Framework

Hardware Platform

Post-Dennard technology pivot - heterogeneity

Dennard Scaling of MOS Devices

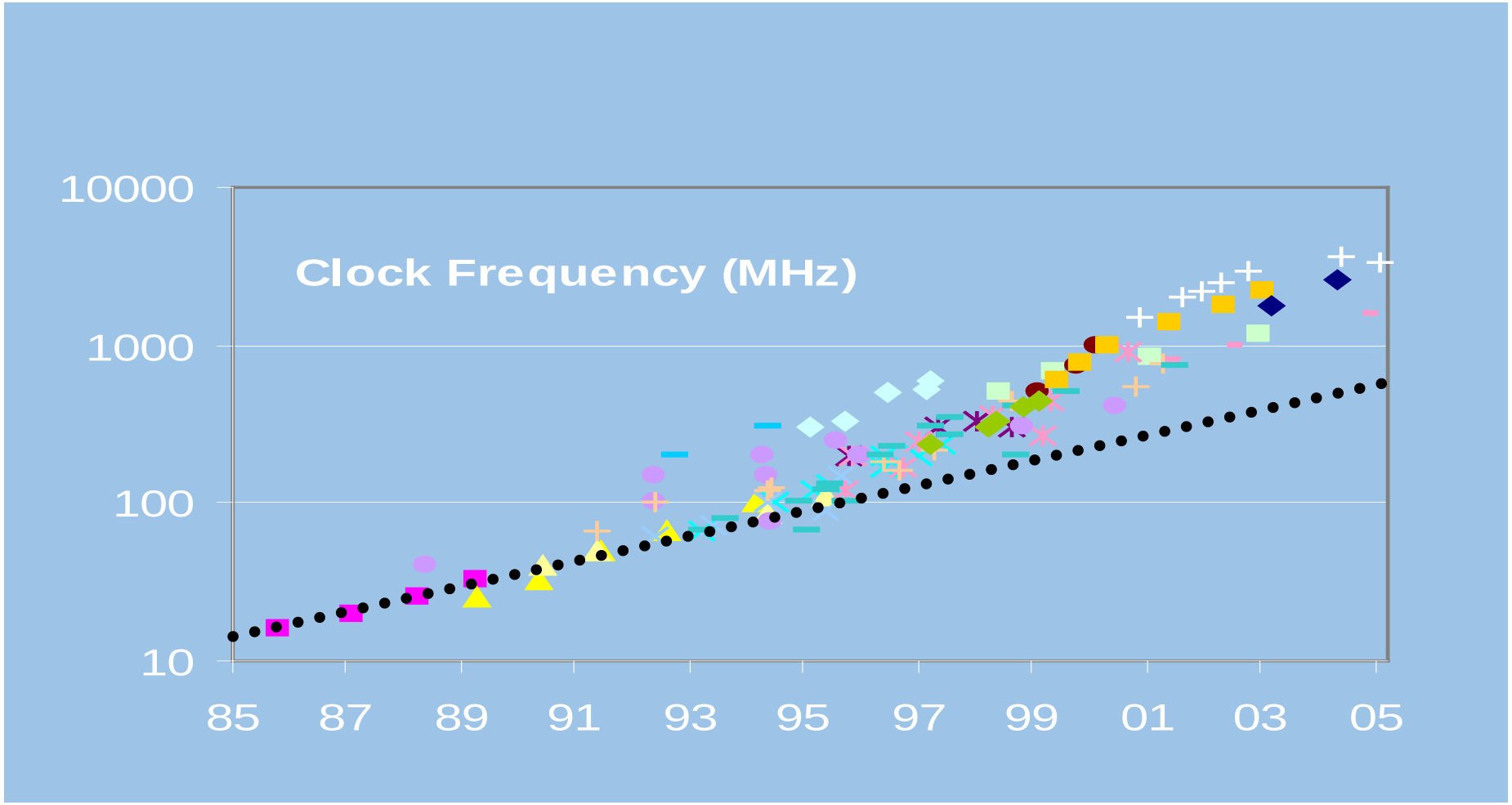




JSSC Oct **1974**, page 256

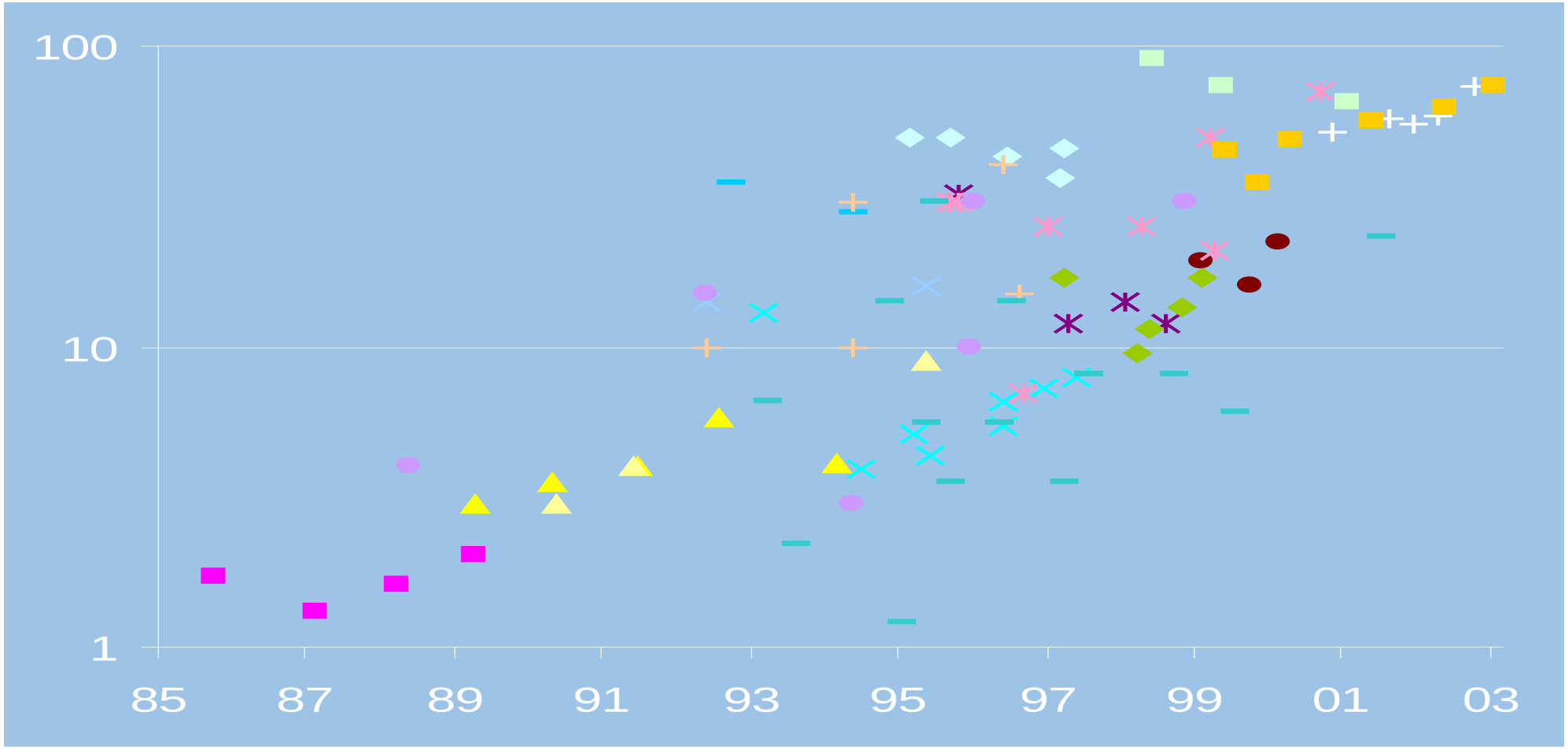
- In this ideal scaling, as L ⟶ α*L
  - V_{DD} ⟶ α*V_{DD}, C ⟶ α*C, i ⟶ α*i
  - Delay = CV_{DD}/I scales by α, so f ⟶ 1/α
  - Power for each transistor is CV^2*f and scales by α^2
    - keeping total power constant for same chip area

Frequency Scaled Too Fast 1993-2003



Total Processor Power Increased

(super-scaling of frequency and chip size)



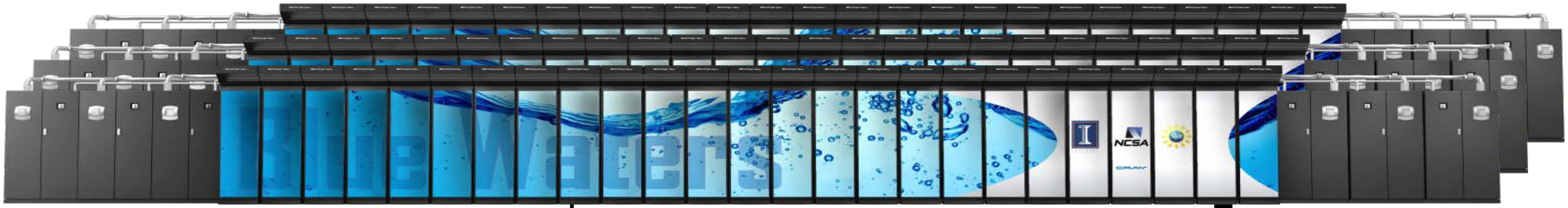
Post-Dennard Pivoting

- Multiple cores with more moderate clock frequencies
- Heavy use of vector execution
- Employ both latency-oriented and throughput-oriented cores
- 3D packaging for more memory bandwidth and increased system capacity

Blue Waters Computing System

Operational at Illinois since 3/2013

49,504 CPUs -- 4,224 GPUs



12.5 PF
1.6 PB DRAM
$250M

10/40/100 Gb Ethernet Switch

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

WAN

Spectra Logic: 300 PBs

Sonexion: 26 PBs

Initial Production Use Results

	Application Description	Application Speedup
NAMD	100 million atom benchmark with Langevin dynamics and PME once every 4 steps, from launch to finish, all I/O included	1.8
Chroma	Lattice QCD parameters: grid size of 483 x 512 running at the physical values of the quark masses	2.4
QMCPACK	Full run Graphite 4x4x1 (256 electrons), QMC followed by VMC	2.7
ChaNGa	Collisionless N-body stellar dynamics with multipole expansion and hydrodynamics	2.1
AWP	Anelastic wave propagation with staggered-grid finite-difference and realistic plastic yielding	3.7-5.0

More Heterogeneity Is Coming

- Beyond traditional CPUs and GPUs
  - FPGAs (e.g., Microsoft FPGA cloud)
  - ASICs (e.g., Google's TPU)

- Beyond traditional DRAM
  - Stacked DRAM for more memory bandwidth
  - Non-volatile RAM for memory capacity
  - Near/in memory computing for increased throughput and reduced power used in data movement

Summary and Outlook

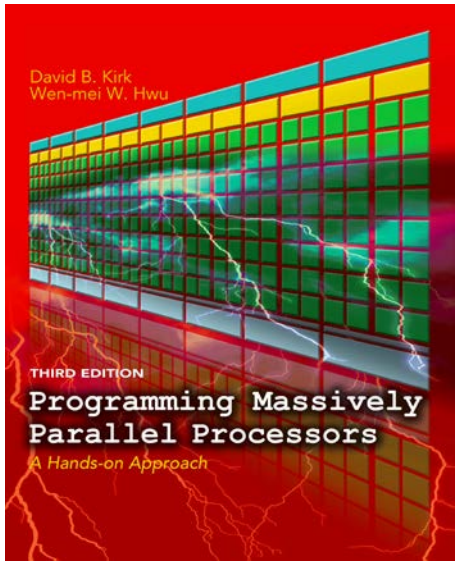
- Throughput computing using GPUs can result in 2-3X end-to-end application-level performance improvement
  - It is NOT too expensive!

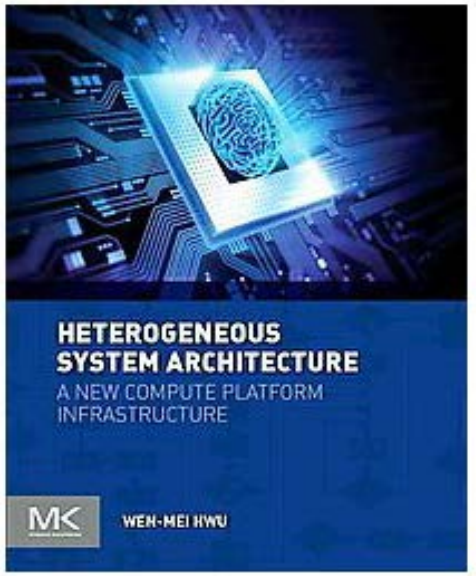
- GPUs, big data and deep learning have formed a positive spiral for the industry

- This is an exceptional time to be a graduate student
  - Paradigm shift, partly thanks to the generations of super-Dennard scaling
  - **But you have to work much harder, also thanks to the generations of super-Dennard scaling**

Thank you!

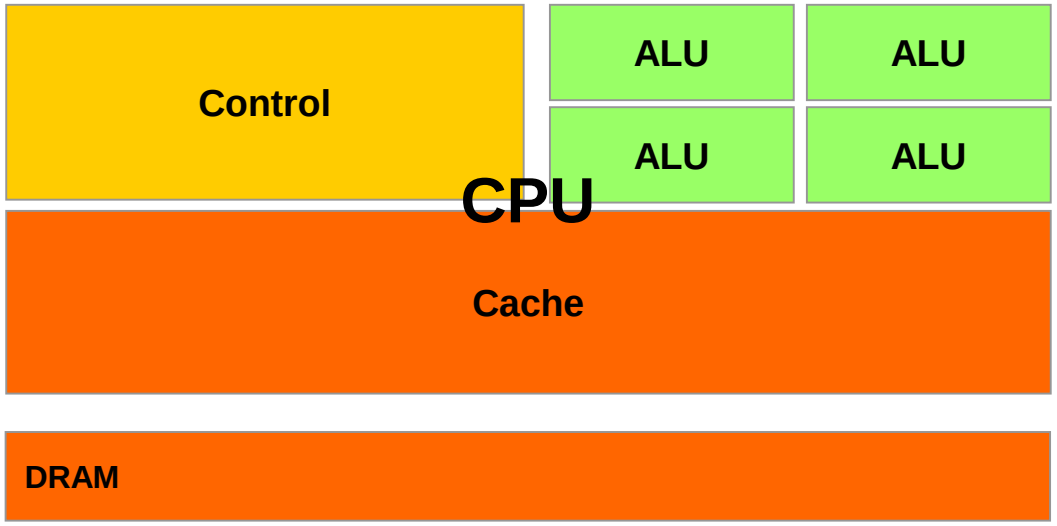
Any questions?





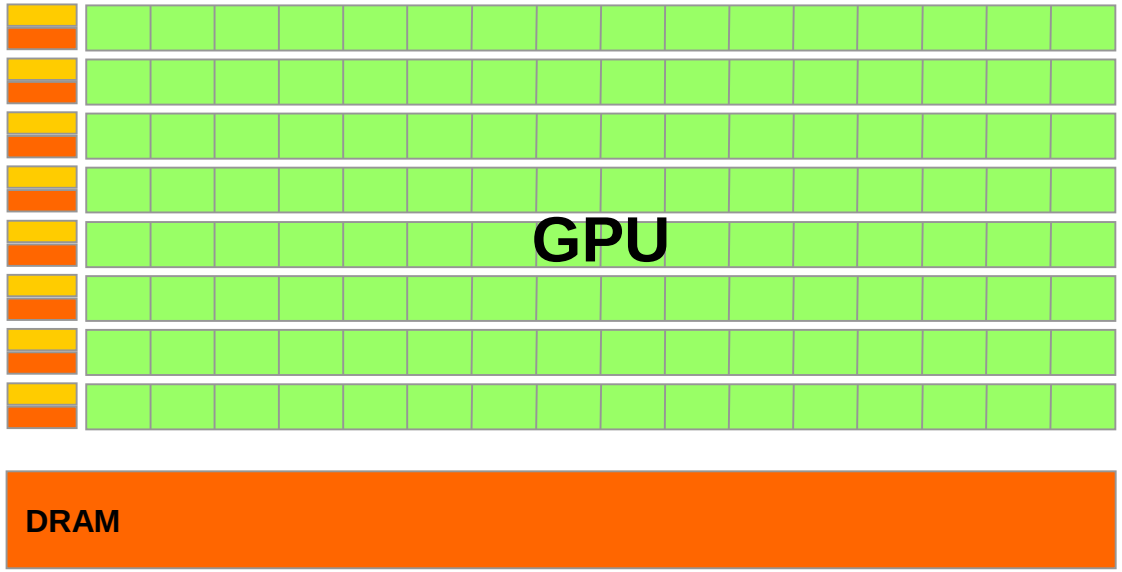
CPUs: Latency Oriented Design

- High clock frequency
- Large caches
  - Convert long latency memory accesses to short latency cache accesses
- Sophisticated control
  - Branch prediction for reduced branch latency
  - Data forwarding for reduced data latency
- Powerful ALU
  - Reduced operation latency



GPUs: Throughput Oriented Design

- Moderate clock frequency
- Small caches
  - To boost memory throughput
- Simple control
  - No branch prediction
  - No data forwarding
- Energy efficient ALUs
  - Many, long latency but heavily pipelined for high throughput
- Require massive number of threads to tolerate latencies



Applications Benefit from Both CPU and GPU

- CPUs for sequential parts where latency matters
  - CPUs can be 10+X faster than GPUs for sequential code
- GPUs for parallel parts where throughput wins
  - GPUs can be 10+X faster than CPUs for parallel code