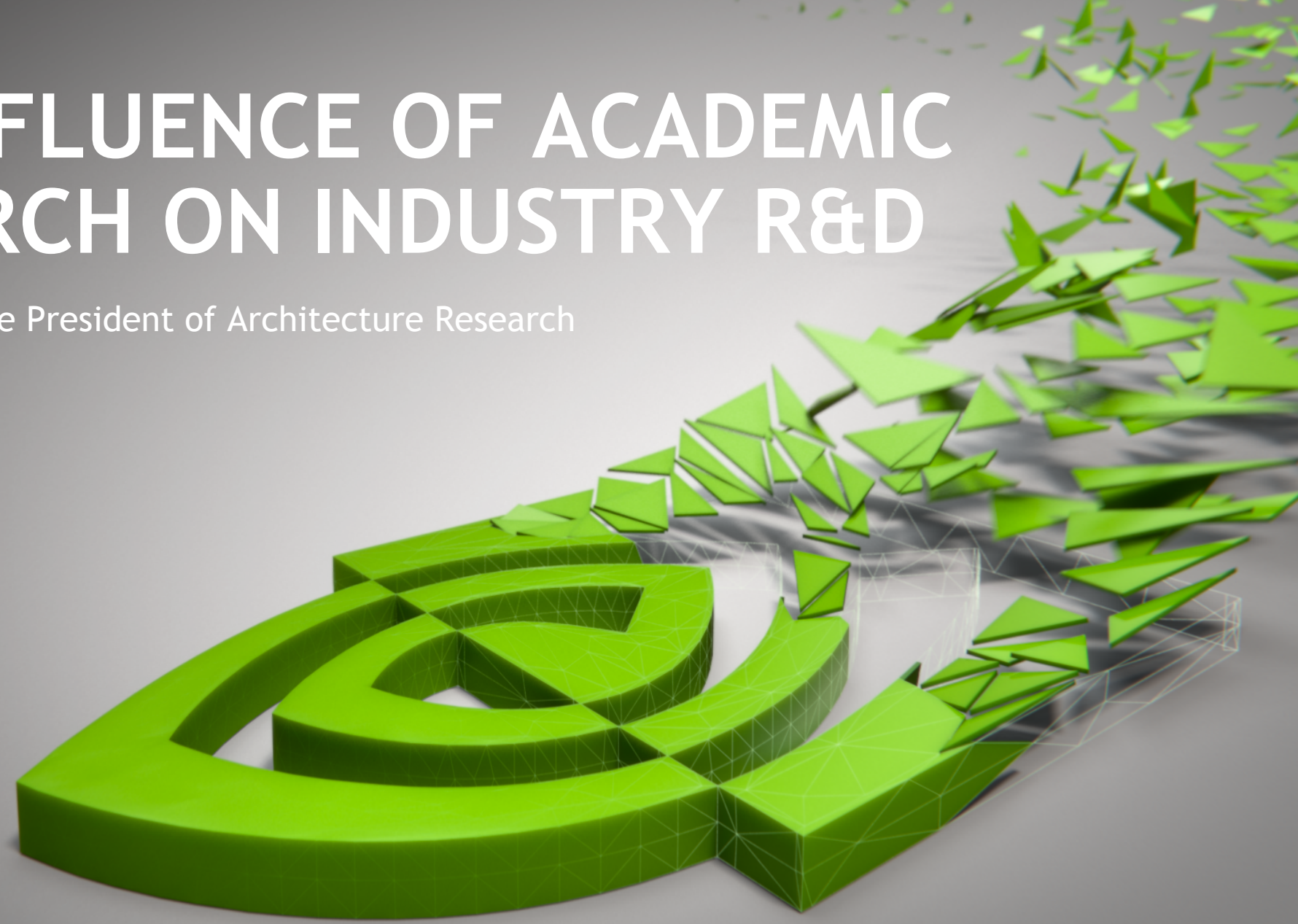
THE INFLUENCE OF ACADEMIC RESEARCH ON INDUSTRY R&D

Steve Keckler, Vice President of Architecture Research

June 19, 2016

NVIDIA



AGENDA

Academic/Industry Partnership

Architecture 2030

My Background/Experience

- 14 years as tenure-track professor at UT-Austin

- 6 years leading architecture research at NVIDIA
  - Drive architecture research beyond product time horizon (5-10 years)
  - Pay attention to trends and academic research
    - Not just architecture but applications, technology, programming systems, etc.
  - Collaborate with university researchers (and other companies)
  - Invest heavily in technology transfer with product teams

Disclaimer

These are my opinions.

They represent my experiences.

My observations may not necessarily be consistent with experiences at other organizations.

Current NVIDIA/Academic Collaborations

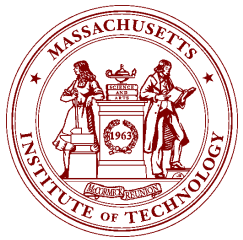
That I know of

THE UNIVERSITY OF UTAH

NC STATE UNIVERSITY

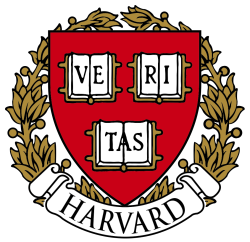
Aalto University
School of Science

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386



Carnegie Mellon University

UF UNIVERSITY of FLORIDA



THE UNIVERSITY OF TEXAS AT AUSTIN

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

UNIVERSITY OF CAMBRIDGE

PRINCETON UNIVERSITY

Examples of Technology Transfer

At NVIDIA

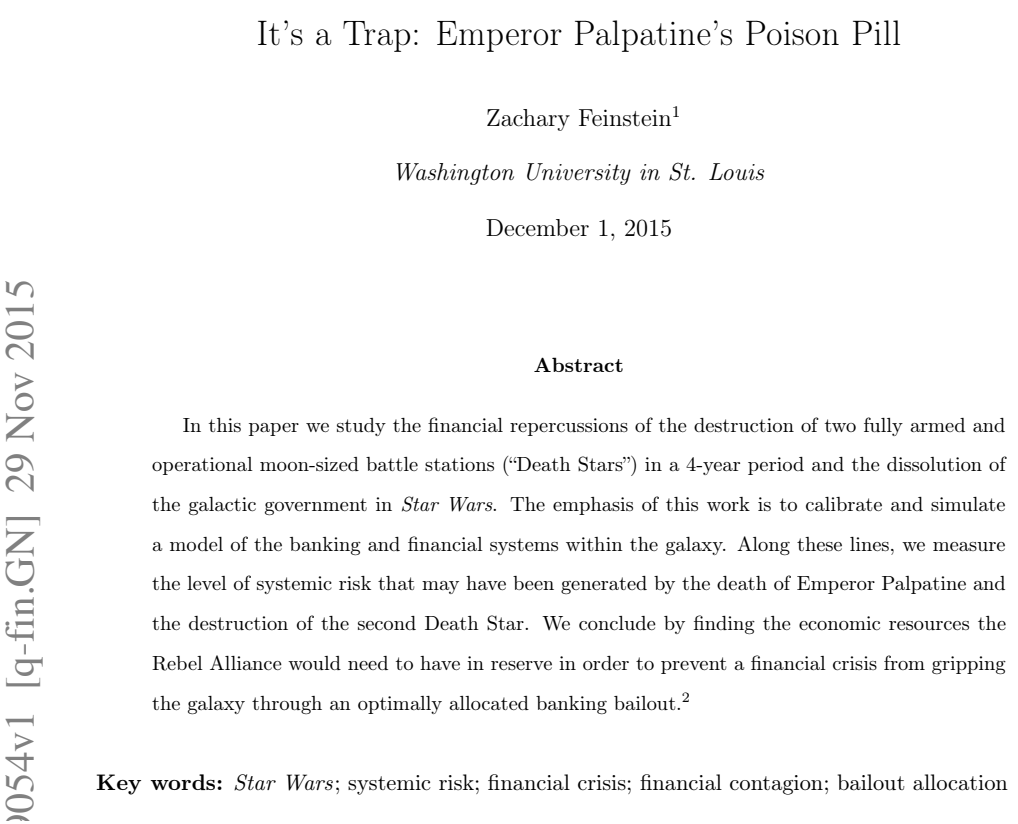
- Streaming architectures (Merrimac, etc.)
- Brook/Cuda
- CuBLAS
- CuDNN
- Machine learning frameworks (Caffe, Torch, Theano, etc.)
- A lot more in the pipe...

Observation 1: Technology Transfer Gap

Observation 1: Technology Transfer Gap

Observation #2: Academic Papers

Yes – we do read LOTS of them



It's a Trap: Emperor Palpatine's Poison Pill

Zachary Feinstein[1]

Washington University in St. Louis

December 1, 2015

Abstract

In this paper we study the financial repercussions of the destruction of two fully armed and operational moon-sized battle stations ("Death Stars") in a 4-year period and the dissolution of the galactic government in *Star Wars*. The emphasis of this work is to calibrate and simulate a model of the banking and financial systems within the galaxy. Along these lines, we measure the level of systemic risk that may have been generated by the death of Emperor Palpatine and the destruction of the second Death Star. We conclude by finding the economic resources the Rebel Alliance would need to have in reserve in order to prevent a financial crisis from gripping the galaxy through an optimally allocated banking bailout.[2]

Key words: *Star Wars*; systemic risk; financial crisis; financial contagion; bailout allocation

- Plus guest lectures, intern talks, etc.
- Wide audience: research + product teams
- No institutional ignorance of good research

Observation #3: Value of an Individual Idea

A Product Consists of Thousands of Ideas/Inventions

- Reasonable to explore an idea in “isolation”
- But acceptance depends on interaction between the idea and all of the other factors in the design

Observation #4: Experimental Results

We don’t believe your simulator

- But don’t despair – the product teams don’t believe ours either
- More important than precise results include
  - Quality of the idea
  - Characterization of opportunity
  - Insight into range of solutions
- Good ideas will get re-examined in context of product roadmap

Observation #5: Field Can Advance Quickly

Product Can Be Ahead of Academic Research

- Can point to papers that have been superseded by product features at time of publication
- Incremental research in well-trodden area is not usually relevant

How to Minimize the Impact of Your Research

- Work on well-trodden and near-term areas
  - More warp scheduling papers please
- Optimize research for maximizing paper count
- Don’t develop direct relationships with industry research and product teams
- Don’t visit or take sabbatical time in industry
- Focus your papers/presentations on the results at the expense of ideas and characterizations
- Expect that your ideas are so good that they will be adopted all by themselves

Architecture 2030

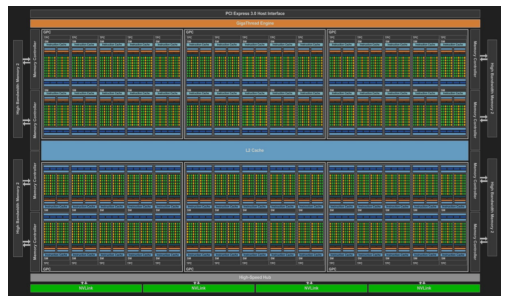
2002

Pentium4

130nm
5.5M xtors
1 core
4-way HT
~6 GF

2016

GP100 GPU



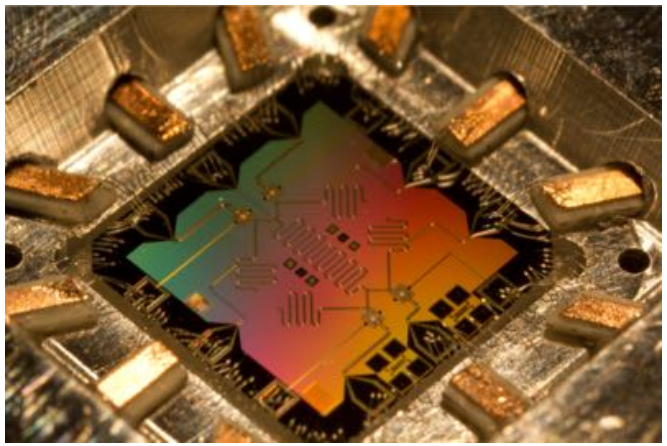
16nm
15.3B xtors (300x)
~2K math units (1000x)
~5.3TF (~1000x)

2030

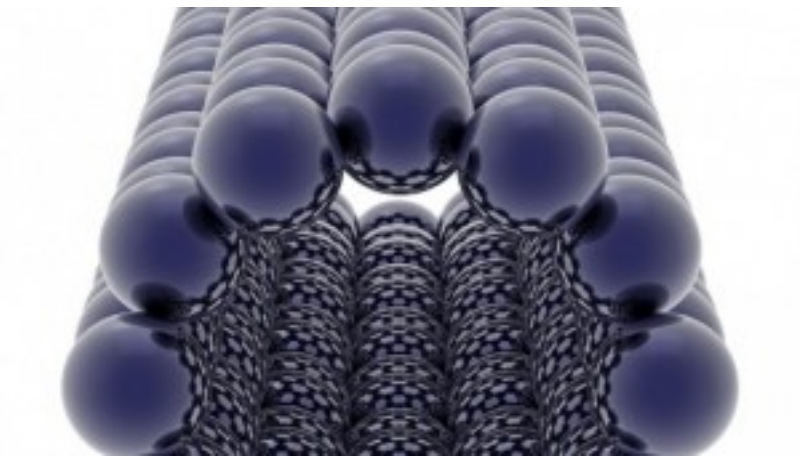
?

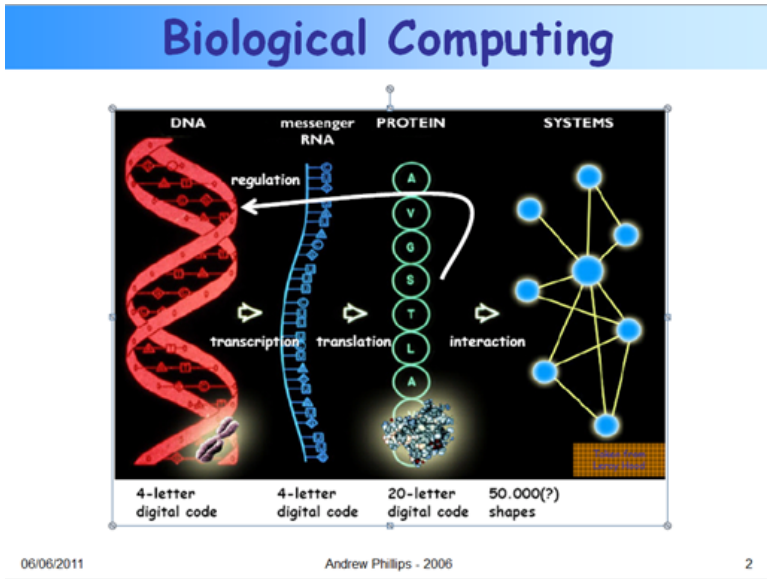
Emerging (Esoteric) Technologies

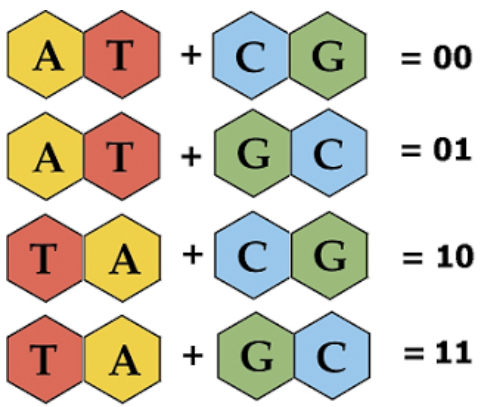
Quantum



Molecular







DNA

Cost

Manufacturability

Programmability

Reliability

Predictability

Innovator's Dilemma

Christensen, 1997

- Inferior disruptive technology can eventually displace incumbent if it can leverage high growth sector of market.

- So-called esoteric technologies need high-volume "killer app".

“I would predict that in 10 years there’s nothing but quantum machine learning—you don’t do the conventional way anymore.”

- Hartmut Neven (Google); MIT Technology Review 6/9/16

Is this plausible?

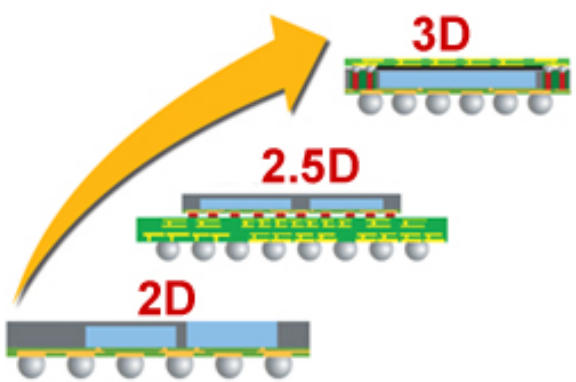
The Death of “New” Chips?

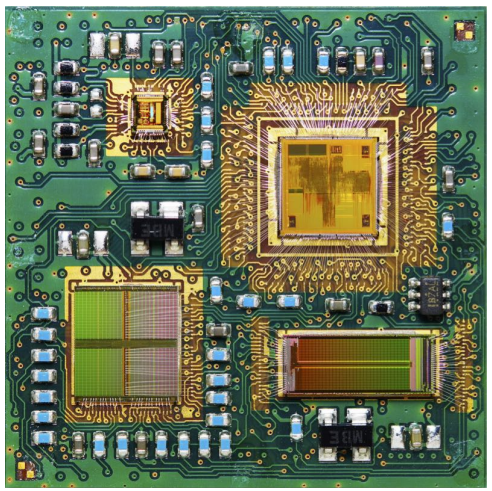
Slowing of Technology Opens up Architecture Competitive Landscape

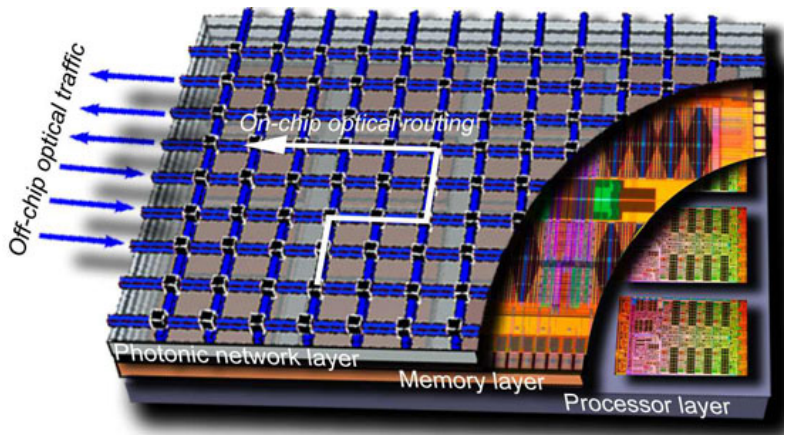
2030 Predictions

No wholesale replacement of CMOS
(and its direct derivatives)

Ample room for innovation in packaging, circuits,
heterogeneous systems (electrical, optical)...and Software







2030 Predictions

No wholesale replacement of CMOS

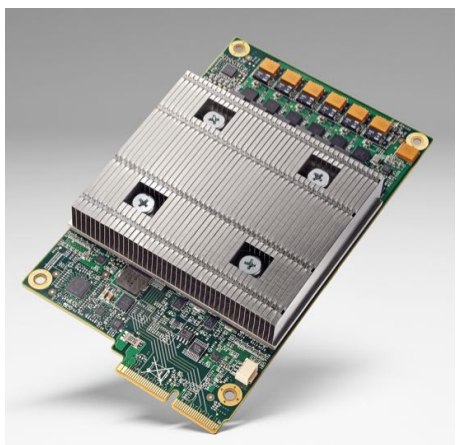
The system will be more important than the chip

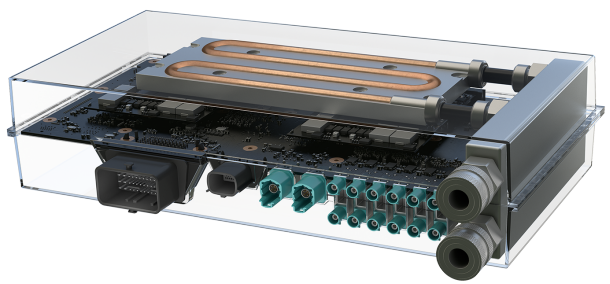
2030 Predictions

No wholesale replacement of CMOS

The system will be more important than the chip

Ample room for domain-specific acceleration





2030 Predictions

No wholesale replacement of CMOS

The system will be more important than the chip

Ample room for domain-specific acceleration

We will still be struggling with programmability

Parallel and Heterogeneous Systems

Programmability vs. Fixed-Function

2030 Predictions

No wholesale replacement of CMOS

The system will be more important than the chip

Ample room for domain-specific acceleration

We will still be struggling with programmability

Chip design will be even more like SW design

PyMTL

Catapult HLS

Stratus HLS

Summary

- “Rennaissance” for architecture research
  - Architecture will continue to increase in importance
  - But needs to span stack (circuits to applications)
- Stay the course on architecture principles
  - Data transformation, date movement, data storage
  - Parallelism, locality, etc.
- Key opportunities
  - Scalability – at multiple levels
  - Domain-specific acceleration