Stanford | ENGINEERING
Electrical Engineering
Computer Science

PERVASIVE PARALLELISM LABORATORY PPL

Scaling Machine Learning with Moore's Law

Kunle Olukotun
Stanford University

EE and CS

Machine Learning Becoming Dominant

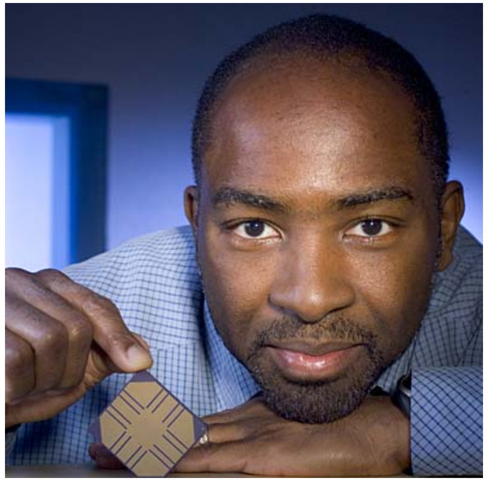
- Recent advances in image recognition, natural language processing, planning, knowledge base construction are driven by machine learning
- Society-scale impact: autonomous vehicles, personalized medicine, personalized recommendations
- Developing high-quality ML applications is challenging
  - Requires deep ML knowledge, custom tools and high-performance computing

The DAWN (Data Analytics for What's Next) Project

Peter Bailis
Streaming &
Databases

Chris Ré
MacArthur Genius
Databases + ML

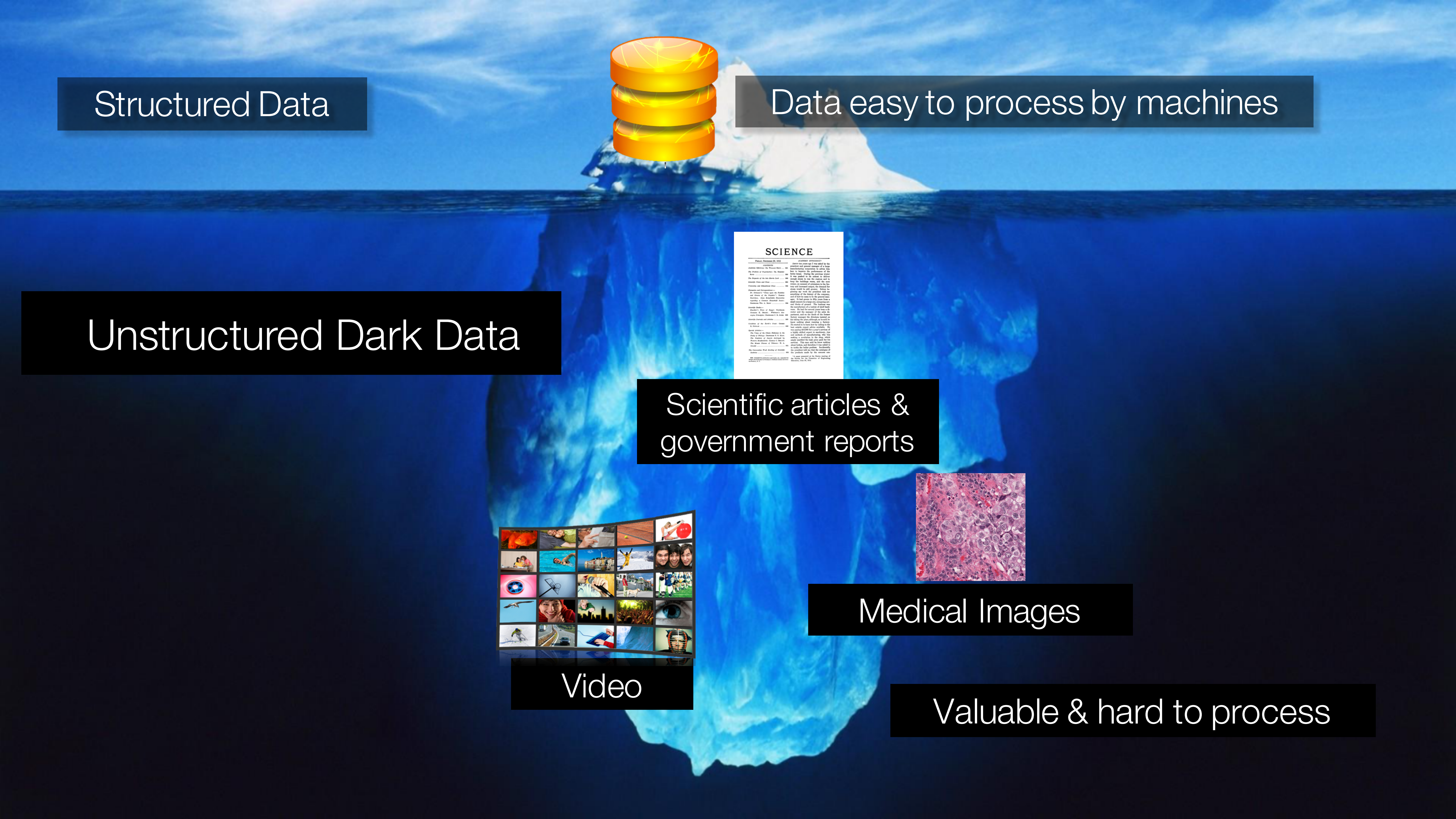
Kunle
Olukotun
Father of Multicore
Domain Specific
Languages



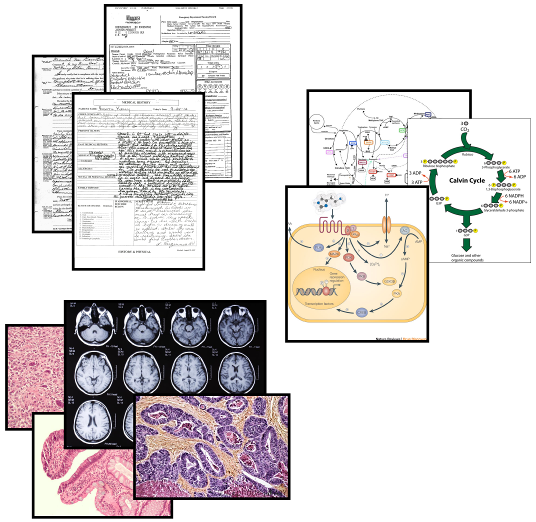
Matei
Zaharia
Co-Creator of
Spark and Mesos

The DAWN Proposal

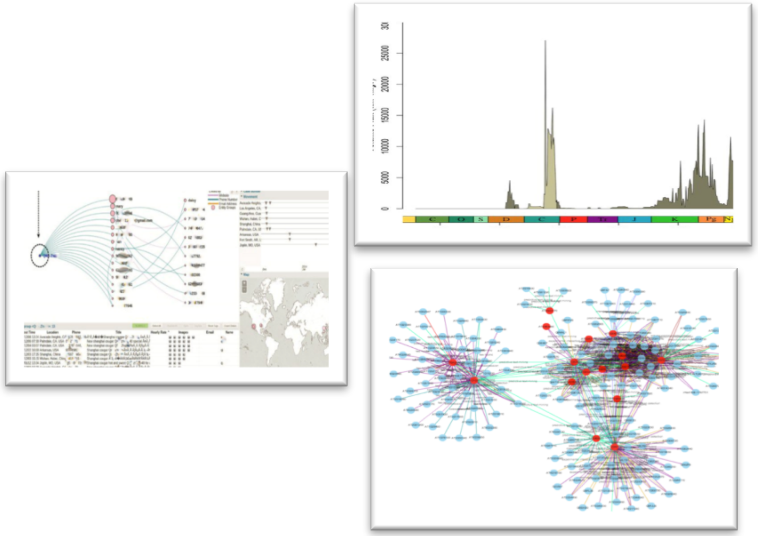
- What if *anyone* with domain expertise could build their own production-quality ML products?
  - Without a PhD in machine learning
  - Without being an expert in DB + systems
  - Without understanding the latest hardware


Structured Data
Data easy to process by machines
Unstructured Dark Data
SCIENCE
Scientific articles & government reports
Medical Images
Video
Valuable & hard to process

Dark Data Extraction (DDE)



Dark Data: Text, Tables, Images, Diagrams, etc.



Structured Data: Enables analyses, interfaces, etc.

A critical and difficult step in many data analysis pipelines

DeepDive

Dark Data System

Human-caliber **quality**
with machine-caliber **scale**

Extraction from the Scientific Literature

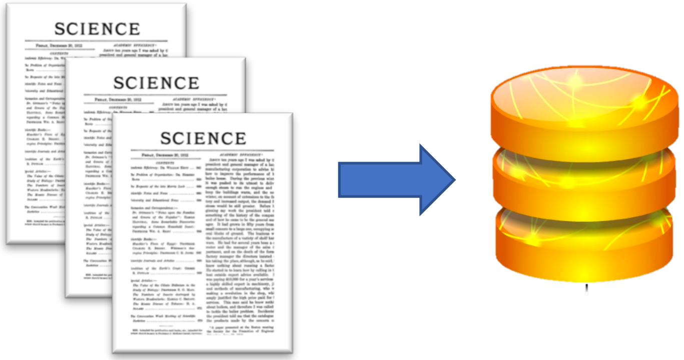
Scientific data accessible, but not *readable*

- What is the impact of human genetic variation on drug responses?

Helix Group

PharmGKB
The Pharmacogenomics Knowledgebase

- What drugs may have unsafe reaction with which gut bacteria?



Dark Data Helps with Societal Problems

Anti-human Trafficking

100M sex ads read with human-caliber quality

- Child predators & human traffickers arrested in multiple jurisdictions across the US

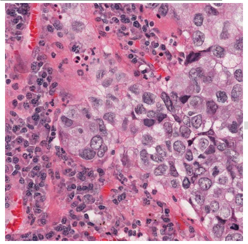
2016 Presidential Award for Extraordinary Efforts to Combat Trafficking in Persons

the WHITE HOUSE *PRESIDENT BARACK OBAMA*

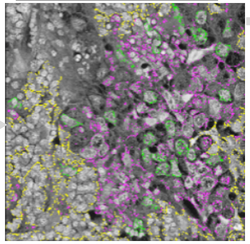
Dark Data Extraction: Beyond Text!

- Example: Tumor grade & stage classification from histopathology slides (Nature Comm., Hsing-Yu et. al.)

Pathlogy Images



Visual Features



Probability of **Survival**

Pathology Reports

A.
Histologic
type:

Adenocarcinoma
Histologic
grade:
Moderately

differentiated
... ...

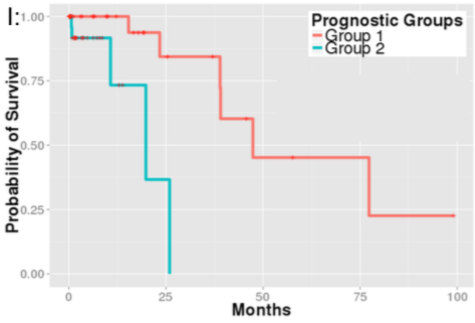
Text Extractions

A.
Histologic
type:

Adenocarcinoma
Histologic
grade:
Moderately

differentiated
... ...

Cleanly separated

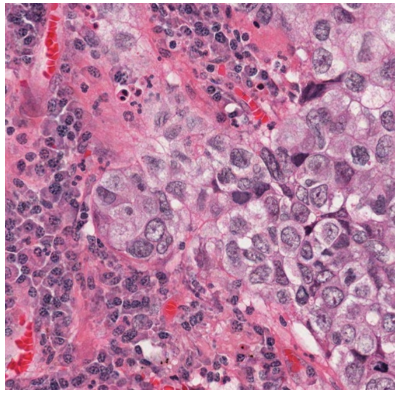


Images + patient data **outperform** expert pathologists at prognosis

Dark data can help improve science, business, and society

Biodiversity

Drug Response



Lung Cancer Prognosis

Fight against human trafficking

http://deepdive.stanford.edu/
http://lattice.io/

Data Programming Pipeline in Snorkel

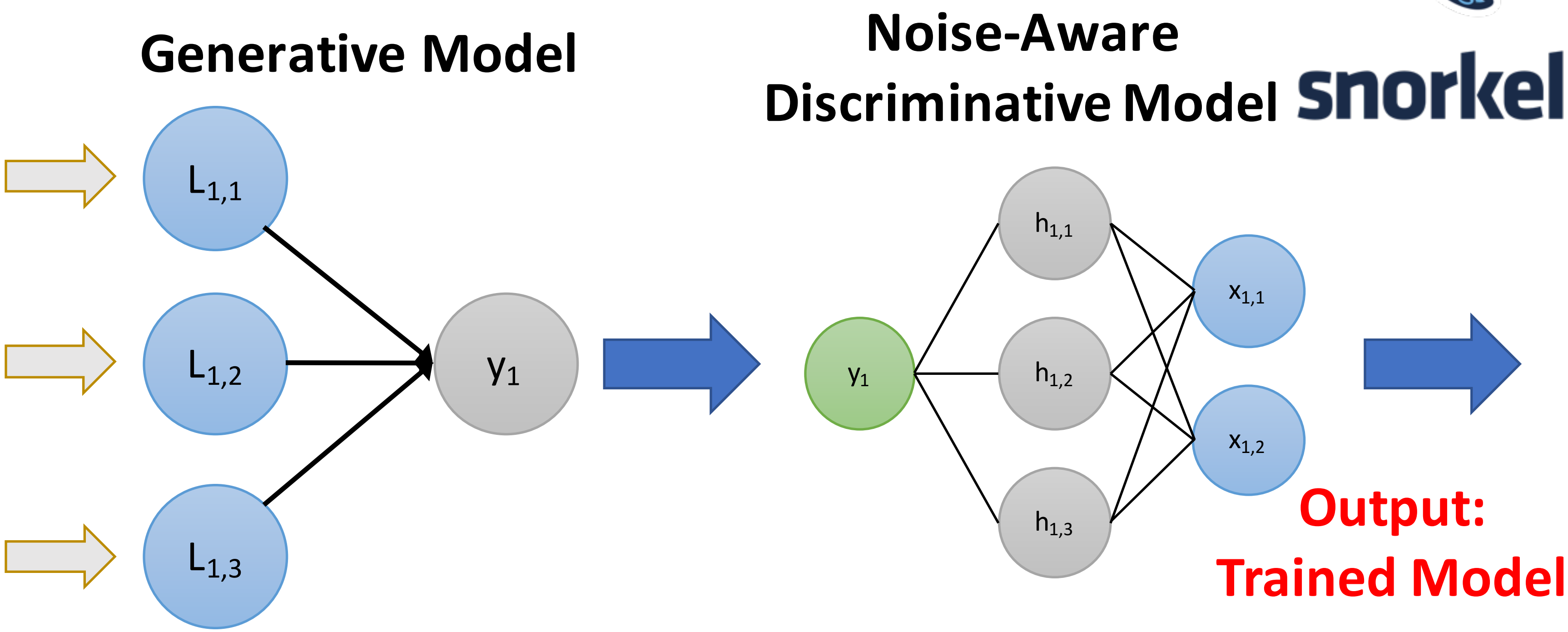
Input: Labeling Functions

DOMAIN EXPERT

```
def lf1(x):
  cid = (x.chemical_id,
x.disease_id)
  return 1 if cid in KB else 0
```

```
def lf2(x):
  m = re.search(r'.*cause.*',
x.between)
  return 1 if m else 0
```

```
def lf3(x):
  m = re.search(r'.*not
cause.*', x.between)
  return 1 if m else 0
```



Users write scripts to label training data

We model this process to denoise it

We use this to train e.g. a deep learning model!

From EE Times – September 27, 2016

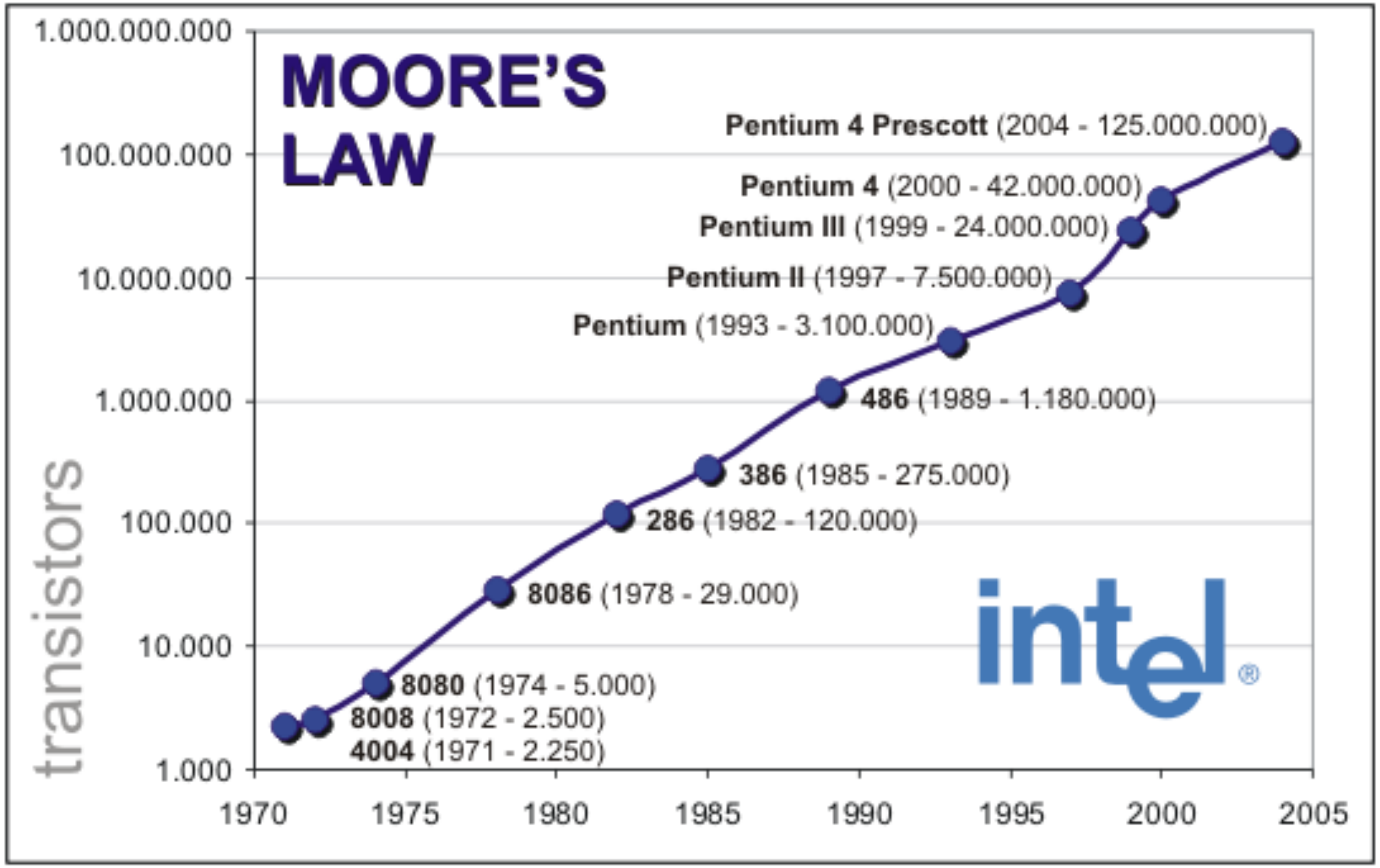
”Today the job of training machine learning models is limited by compute, if we had faster processors we’d run bigger models...in practice we train on a reasonable subset of data that can finish in a matter of months. We could use improvements of several orders of magnitude – 100x or greater.”

Greg Diamos, Senior Researcher, SVAIL, Baidu

DAWN Goals

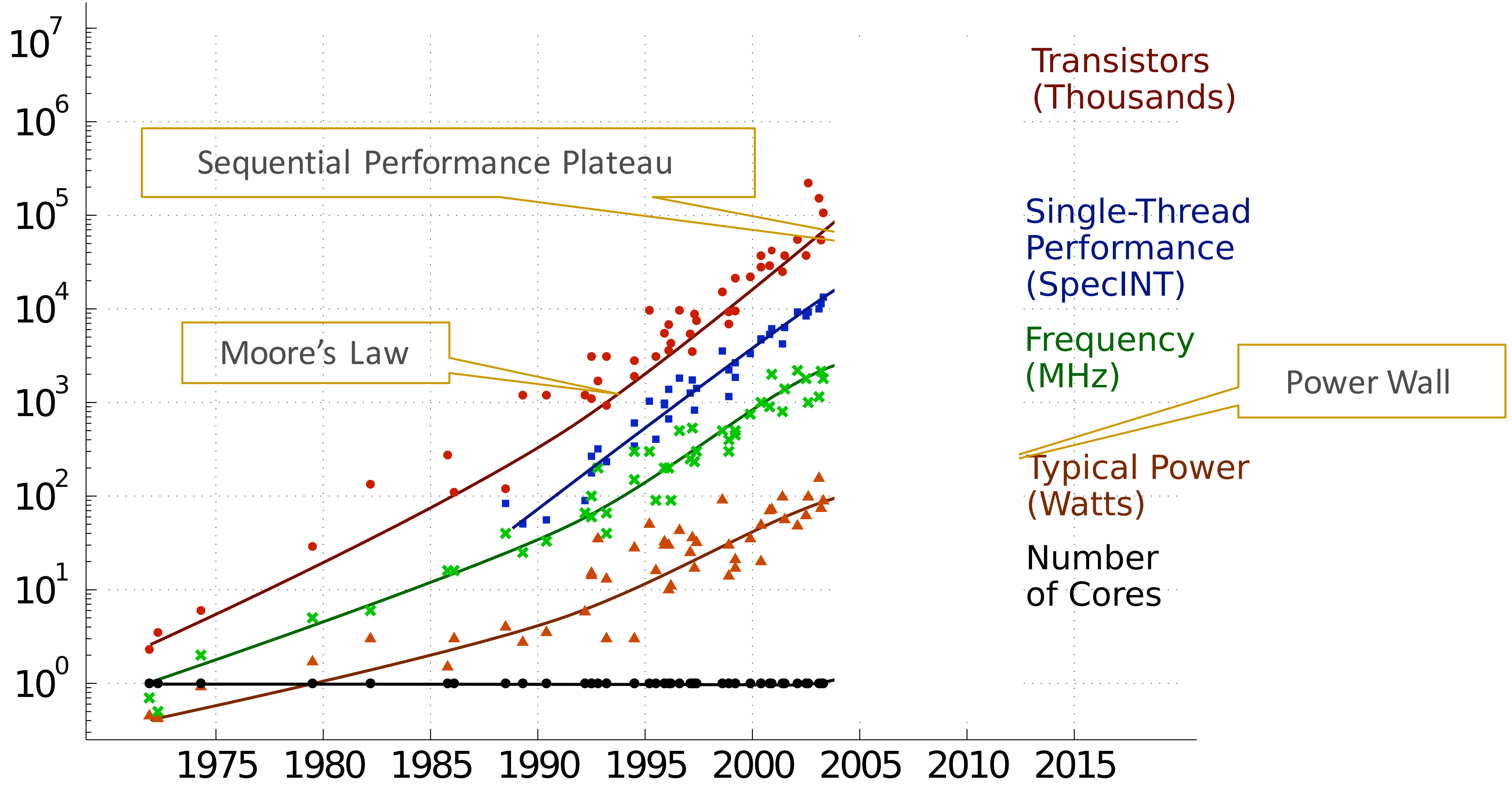
- Speed up machine learning by 100x
  - 1000x improvement in performance/watt
- Enable real-time and interactive ML on big data
  - Data center
  - Mobile
- Full stack approach:
  1. Algorithms
  2. Programming Languages and Compilers
  3. Hardware

Moore's law: The Good Old Days



More transistors… used to mean faster!

Moore's Law Today ⇒ More Cores



Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

Machine Learning Computational Model

- Underlying model for modern applications

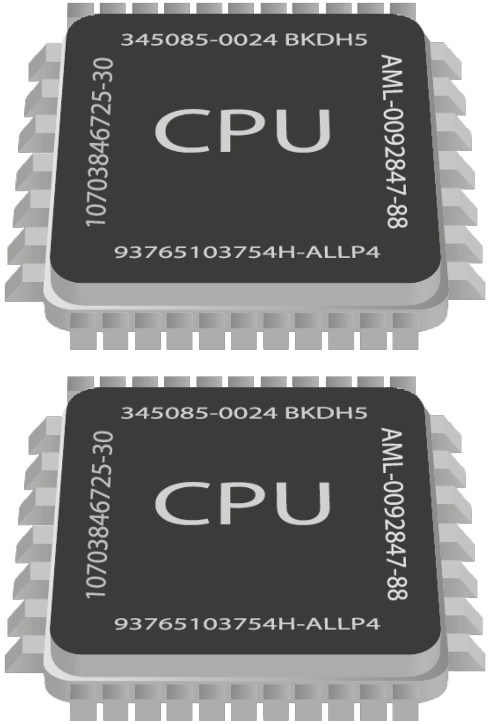
- New computational model
  - Old: Classical deterministic computations with algorithms
  - New: Probabilistic machine-learned models from data

- Statistical correctness creates many opportunities for improved parallel performance

Everything You Learned About Parallel Computing is Wrong for Machine Learning!

A Crash Course in Parallel Computing

Multicore: No Data Sharing Case



Jobs with little data sharing, 2 cores execute twice as fast!

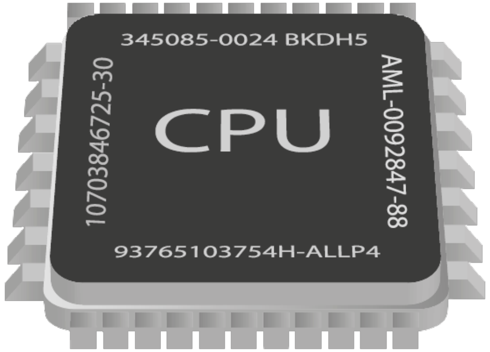
Multicore: Shared Data Case

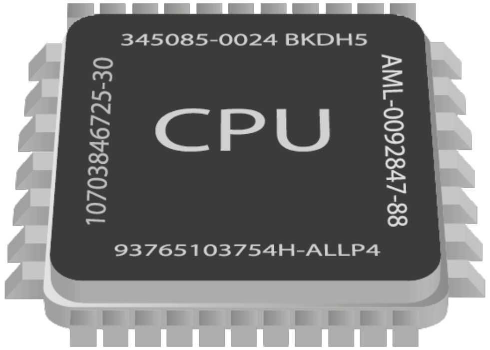
Job 1

Job 2

Job 3

Job 4





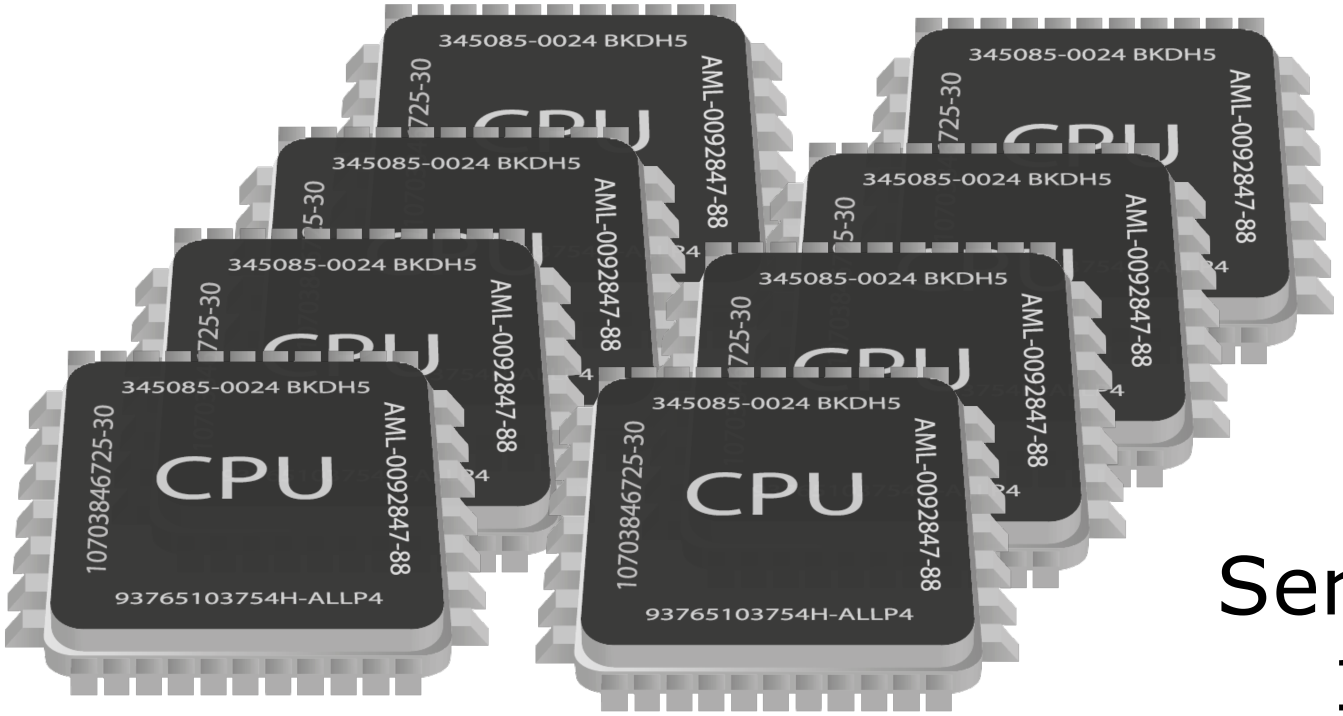
Is it my turn?

Protocol for "whose turn," called **locking,** takes 100s of CPU clock cycles

Locking Overhead Scales Quadratically

Suppose it takes 1 second to synchronize with 2 cores

4 cores takes 4 seconds



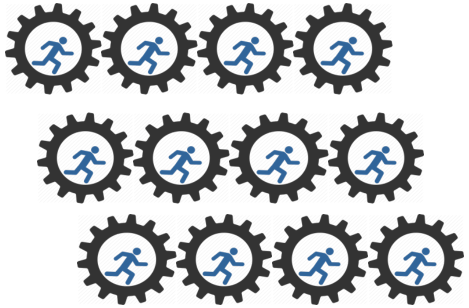
8 cores takes 16 seconds

k cores takes $(k/2)^2$ seconds

Server may have 100+ cores

SGD: The Key Algorithm in Machine Learning

The core algorithm of modern learning is called **Stochastic Gradient Descent (SGD)**



SGD consists of **BILLIONS** of tiny jobs that share a single data stucture!

Implemented in a classical way (locking) SGD actually gets *slower* with more cores

So what can we do?

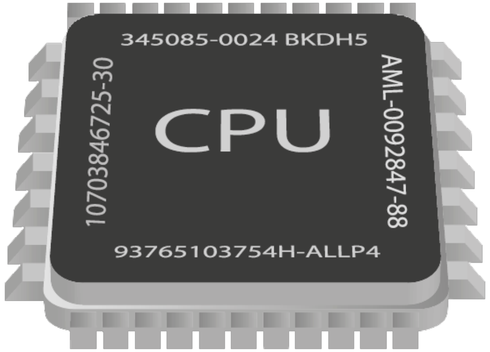
Multicore: Hogwild! Case

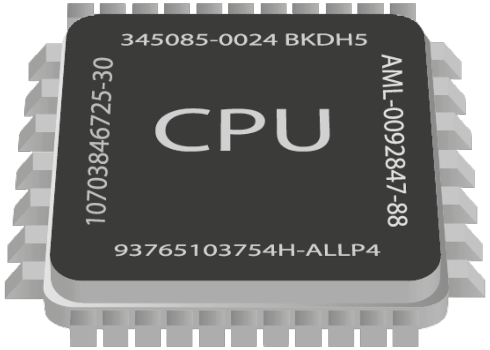
Job 1

Job 2

Job 3

Job 4





Is it my turn? Yes!

Ignore the locks!

How do we run SGD in Parallel?

Just ignore the locking protocol…
As we say, go **Hogwild**!

This is computer science
heresy!

Theorem (roughly, NIPS11): If we do ***no locking***, SGD converges to *correct* answer—*at essentially the same rate*!

Cortana: Microsoft's Digital Assistant

WIRED

AI breakthrough: Microsoft's 'Project Adam' identifies dog breeds, points to future of machine learning

All web companies have similar: image rec, voice, mobile, search, etc.

"…using a technology called, of all things, ***Hogwild!"***

http://www.wired.com/2014/07/microsoft-adam/

http://www.geekwire.com/2014/artificial-intelligence-breakthrough-microsofts-project-adam-identifies-dog-breeds/

Single Instruction Multiple Data (SIMD)

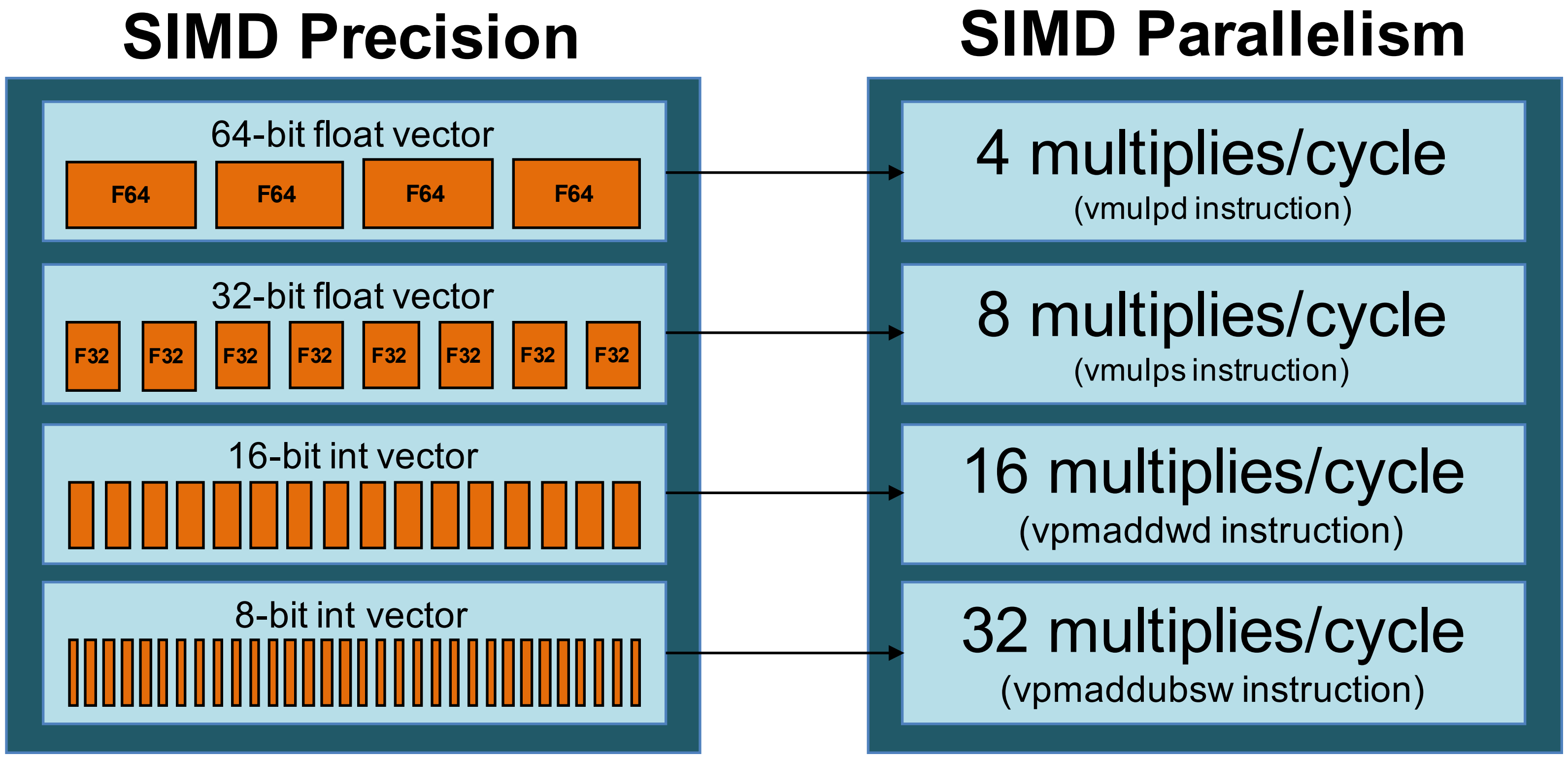
- Like vector processing
- Single instruction can process multiple values at once

$$\underbrace{\begin{array}{ccccc} \boxed{x_1} & \times & \boxed{y_1} & = & \boxed{x_1 y_1} \\ \boxed{x_2} & \times & \boxed{y_2} & = & \boxed{x_2 y_2} \\ \boxed{x_3} & \times & \boxed{y_3} & = & \boxed{x_3 y_3} \\ \boxed{x_4} & \times & \boxed{y_4} & = & \boxed{x_4 y_4} \end{array}}_{\text{one } \textbf{\textit{vmulpd}} \text{ instruction}}$$

- Source of parallelism independent of multicore

Low Precision and SIMD Parallelism

- Major benefit of low-precision: use SIMD instructions to get more parallelism on CPU

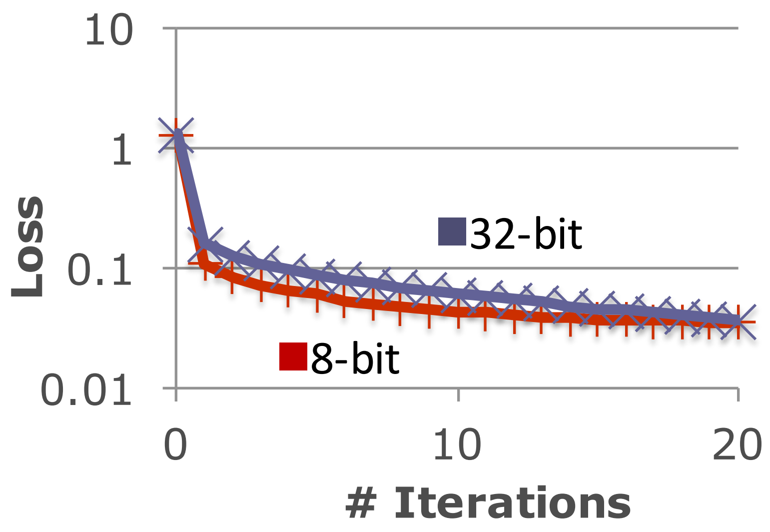


The Buckwild! Strategy

- Use 8- or 16-bit fixed point numbers for computing SGD rather than 32-bit floating point
  - Fewer bits of data → better use of SIMD → higher performance
  - Fewer bits of data → same convergence behavior with SGD
    - Theory: [De Sa, Zhang, Olukotun, Ré: NIPS 2015]

Buckwild!
Statistical vs. Hardware Efficiency

Same statistical efficiency



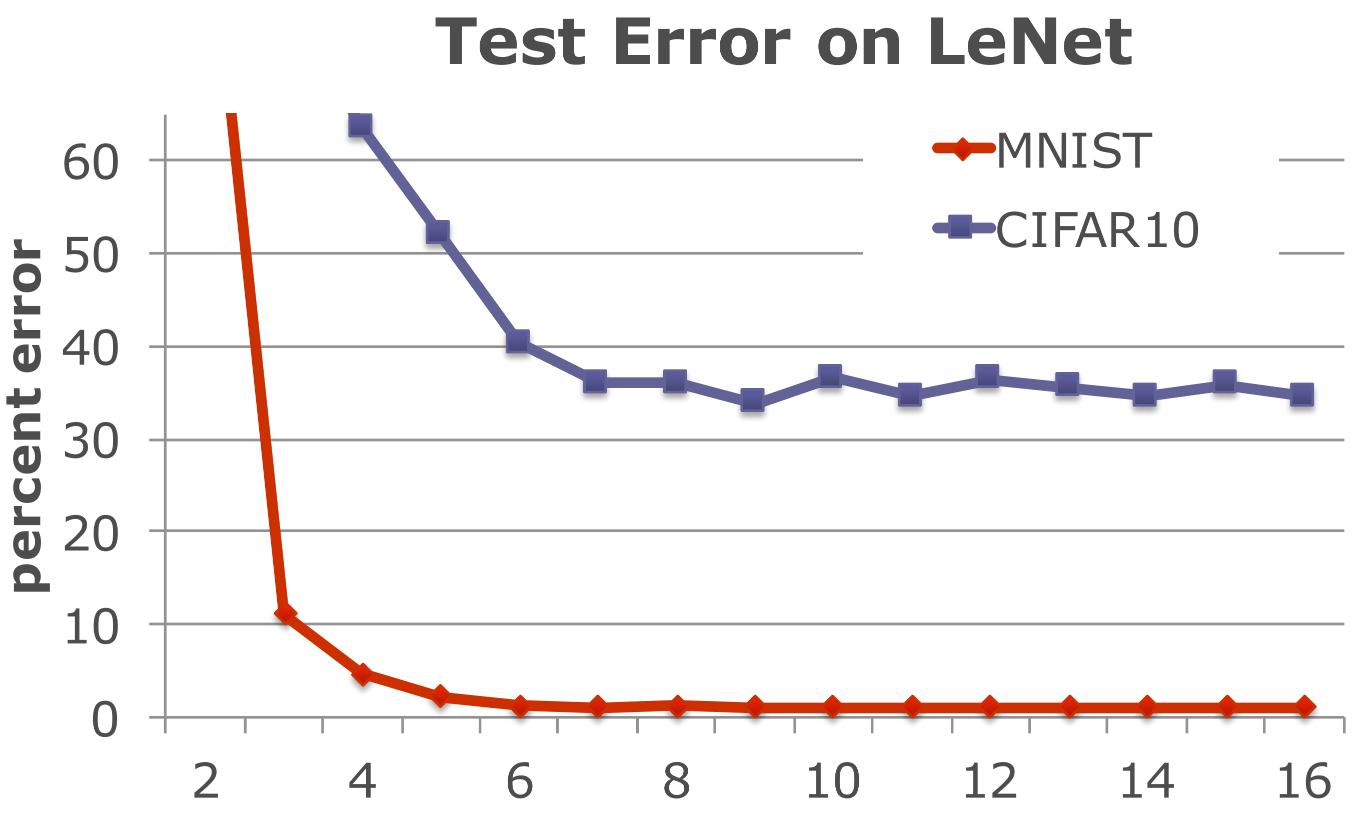
Logistic Regression using SGD

Improved hardware efficiency

- 8-bit gives about **3x** speed up!
- Lower precision is possible
- Good match to specialized/reconfigurable HW?

BUCKWILD! has same **statistical efficiency** with greater **hardware efficiency**

Low Precision for Convolutional Neural Network



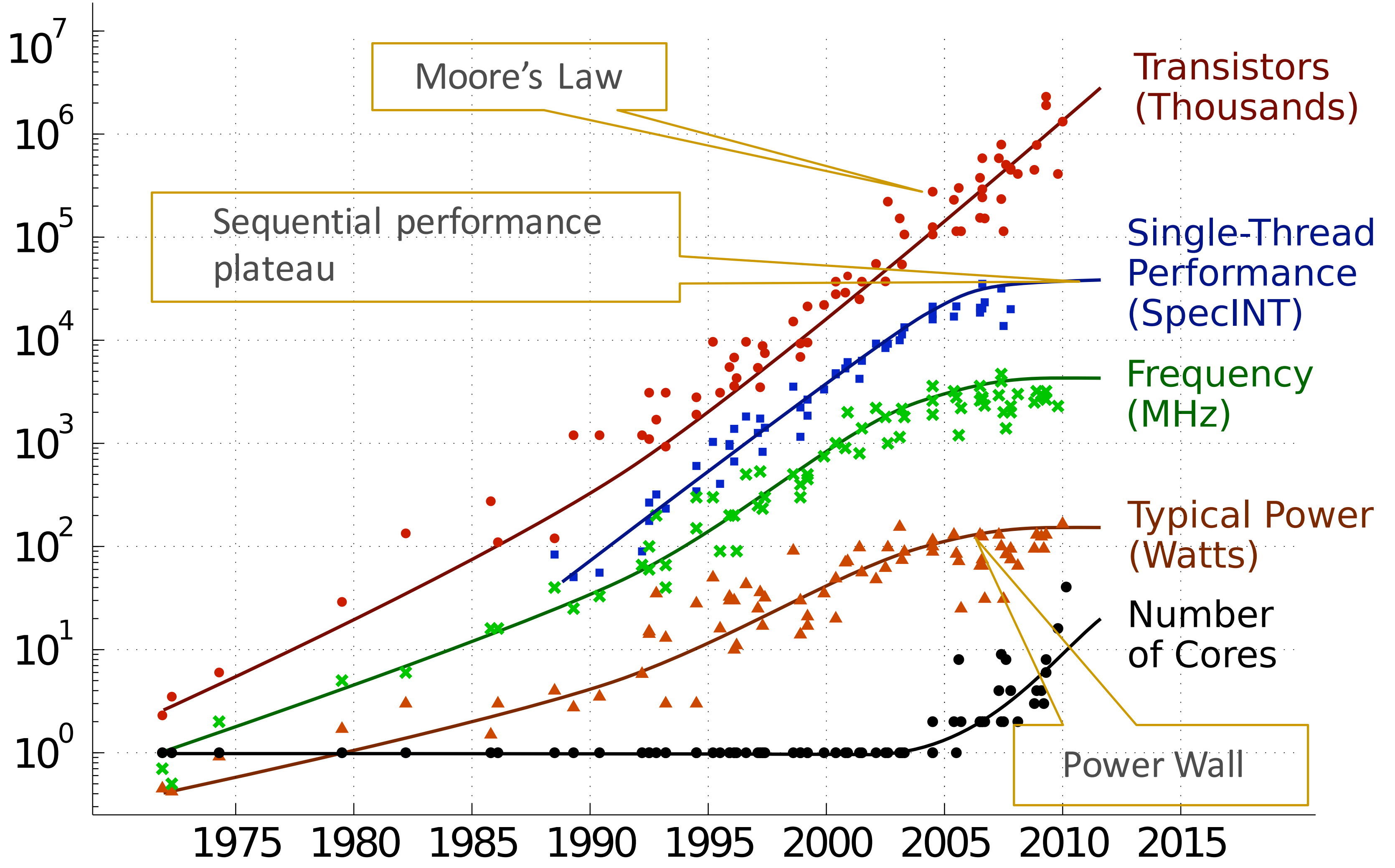
We can go down below 8-bits of precision without sacrificing accuracy

Relax, It's Only Machine Learning

- Relax locking: data races are better
  - HogWild! [De Sa, Olukotun, Ré: *ICML 2016*, ICML Best Paper]
- Relax precision: small integers are better
  - BuckWild! [De Sa, Zhang, Olukotun, Ré: *NIPS 2015*]
- Relax cache coherence: incoherence is better
  - [De Sa, Feldman, Ré, Olukotun: *ISCA 2017*]

Better hardware efficiency
with negligible impact on statistical efficiency

End of Dennard Scaling ⇒ End of Multicore



Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

11 nm process in 2022
96 cores @ 4.9 GHz ⇒ 300W

Power Limit	Active Cores
165 W	53/96
180 W	58/96
200 W	64/96

Dark Silicon

Power and Performance

Energy efficiency **Performance**

$$Power = \frac{Joules}{Op} \times \frac{Ops}{\text{second}}$$

FIXED

Specialized accelerators improve energy efficiency

FPGA Based Accelerators

- Increasing interest in use of FPGAs as application accelerators in data centers

Key advantage: Performance/Watt

Microsoft®

Bing

Baidu 百度

amazon webservices

intel®

ALTERA®
now part of Intel

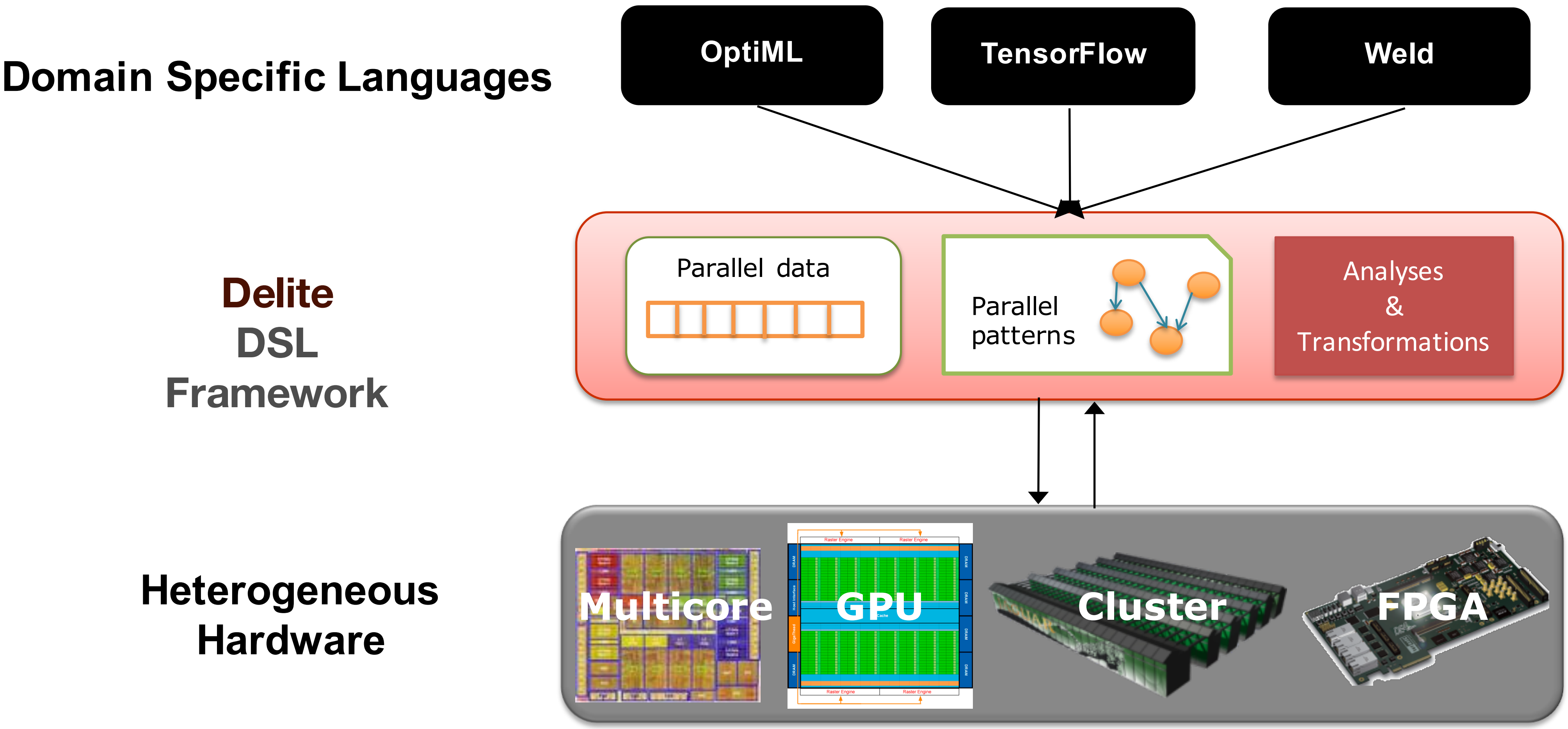
FPGA Problems: Programmability and Design

- Verilog and VHDL poor match for software developers
  - High quality designs, but low productivity

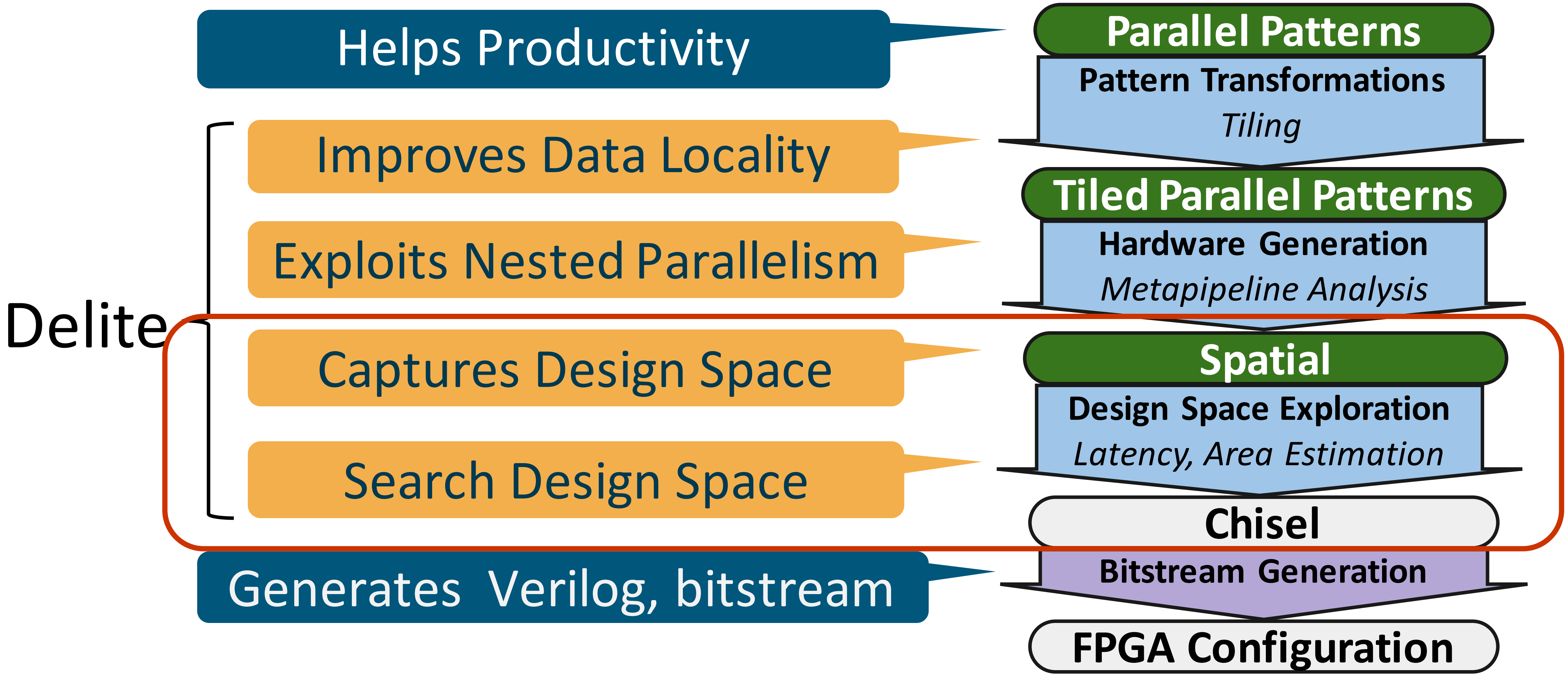
- High level synthesis (HLS) tools with C interface
  - Medium/low quality designs
  - Need hardware knowledge to build good accelerators

- FPGA design space grows exponentially with the number of parameters
  - Even relatively small designs can have very large spaces
  - Manual exploration is tedious, usually results in suboptimal designs

DSLs, Parallel Patterns and Delite



Parallel Patterns to Hardware



Spatial Performance vs. HLS

Spatial:

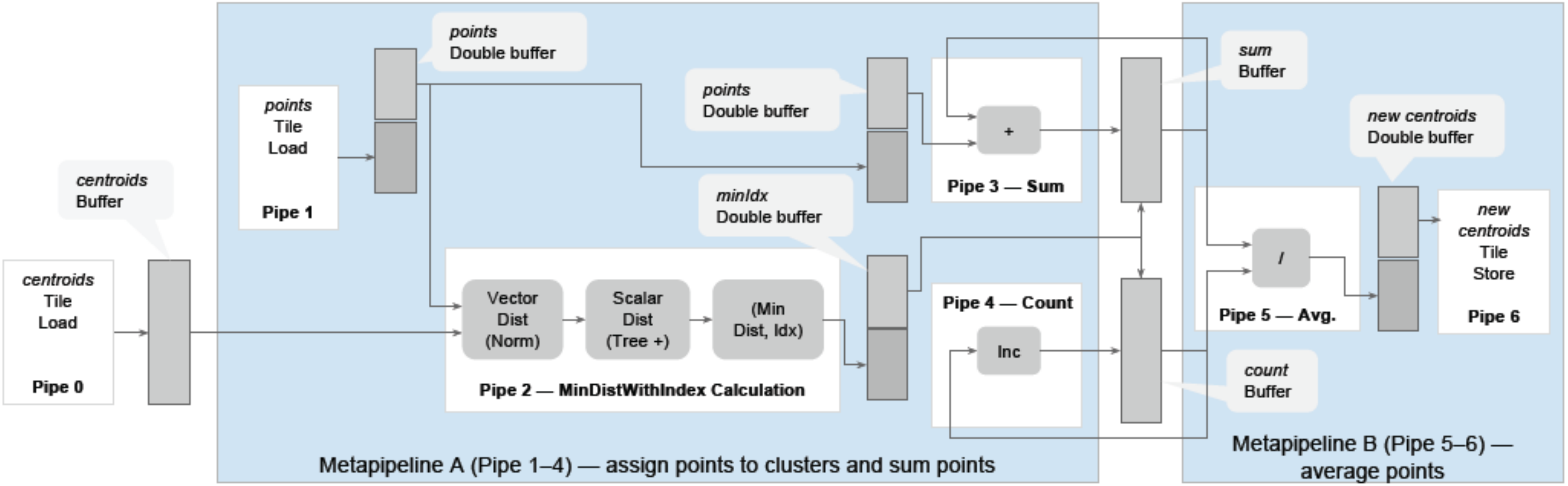
Benchmark	Designs	Search Time
Dot Product	5,426	**5.3 ms / design**
Outer Product	1,702	**30 ms / design**
TPCHQ6	[illegible]	**[illegible].2 ms / design**
Blackscholes	[illegible]	**[illegible]7 ms / design**
Matrix Multiply	[illegible]	**[illegible]1 ms / design**
K-Means	75,200	**20 ms / design**
GDA	42,800	**17 ms / design**

6500x Speedup Over HLS!

Vivado HLS:

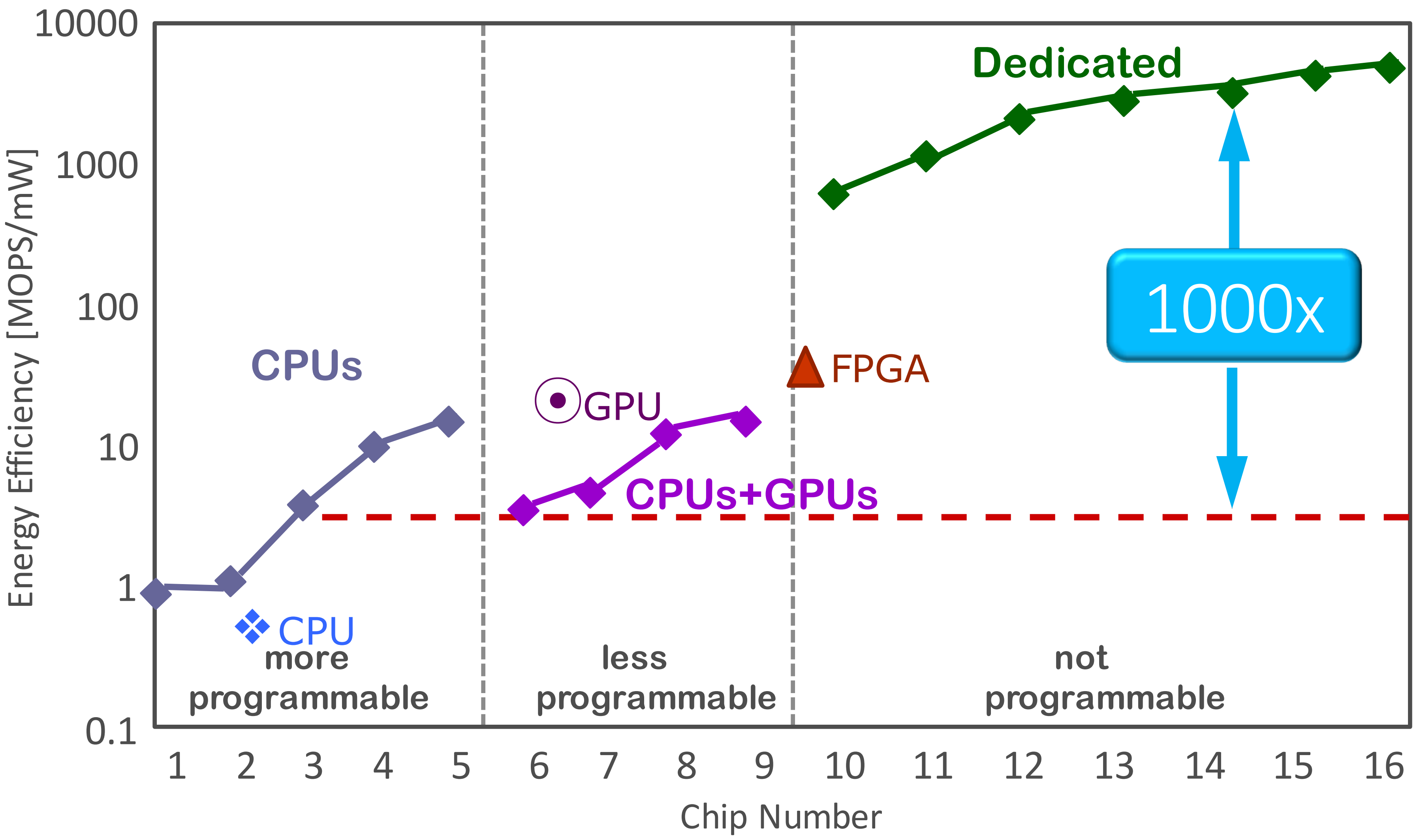
	Designs	Search Time
GDA	250	**1.85 min / design**

Generated *k*-means Hardware



- High quality hardware design
  - Hardware similar to Hussain et al. *Adapt. HW & Syst. 2011*
    - “FPGA implementation of k-means algorithm for bioinformatics application”
    - Implements a fixed number of clusters and a small input dataset
  - Tiling analysis automatically generates buffers and tile load units to handle arbitrarily sized data
  - Parallelizes across centroids and vectorizes the point distance calculations

Energy Efficiency vs. Programmability



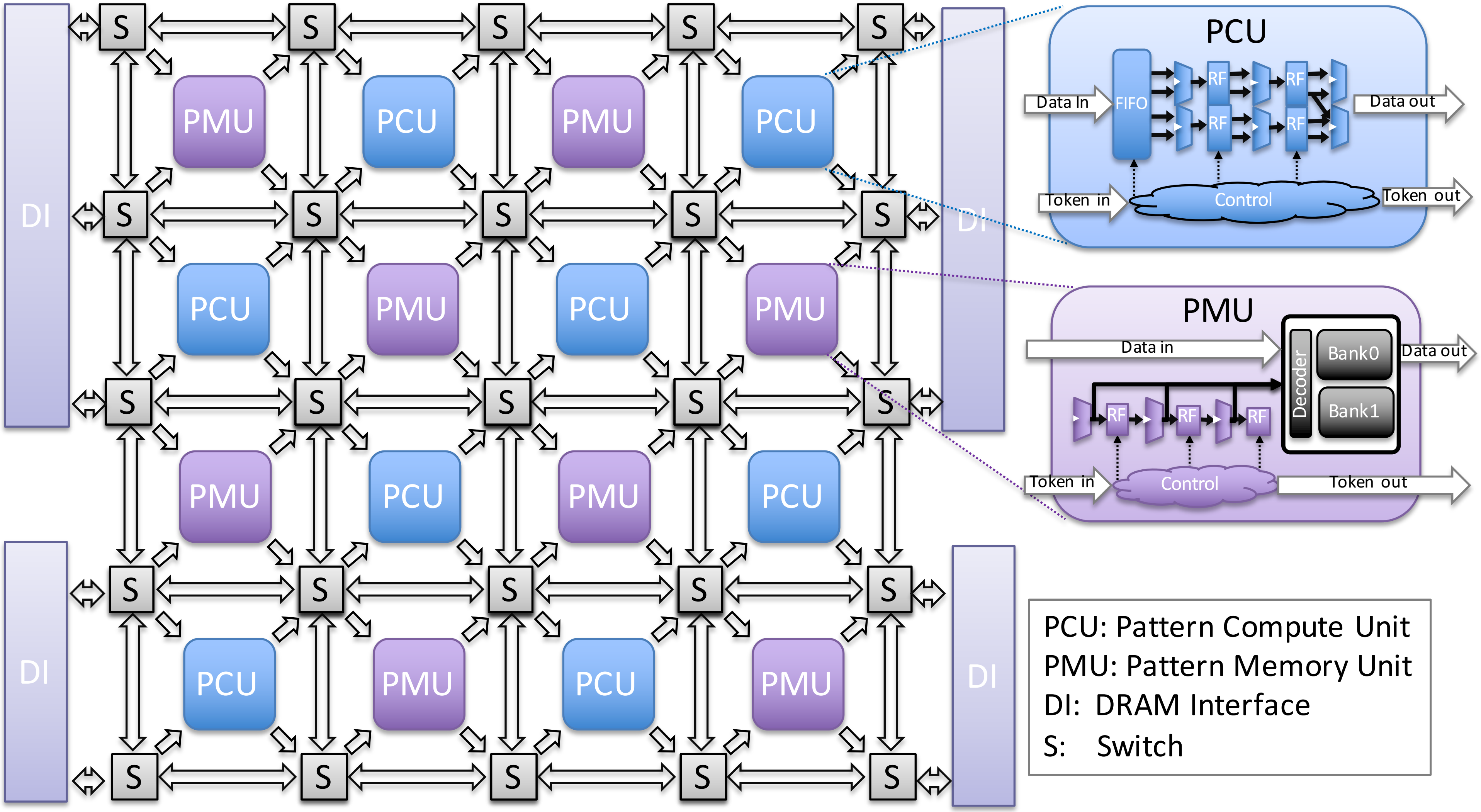
Specialized ML hardware that provides programmability of CPUs and energy efficiency of ASICs

Software Defined Hardware (SDH)

- All(Just) the advantages of conventional accelerators
  - Flexibility of FPGAs
  - Programmability of GPUs
  - Efficiency of ASICs
- SDH Goals
  - 100x performance/Watt vs. CPU
  - 10x performance/Watt vs. FPGAs/GPUs
  - 1000x programmability vs. FPGAs

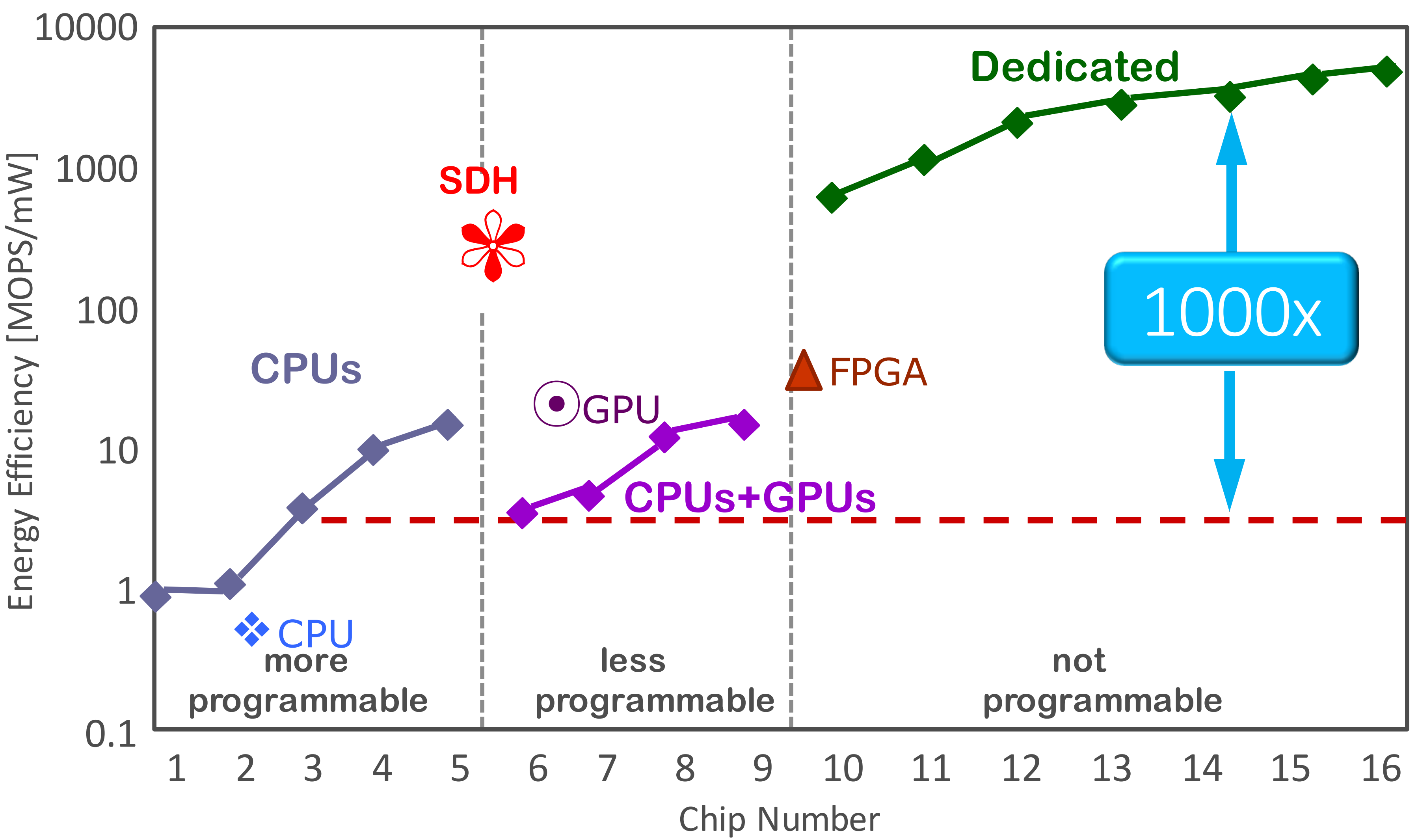
Plasticine: A SDH Architecture

Plasticine: A Reconfigurable Architecture for Parallel Patterns, ISCA 2017



PCU: Pattern Compute Unit
PMU: Pattern Memory Unit
DI: DRAM Interface
S: Switch

Software Defined Hardware



We Can Have It All!

- Performance
- Power
- Programmability
- Portability

ML Applications (DeepDive, Snorkel)
Algorithms (Hogwild!, Buckwild!)

App Developer

High Performance DSLs
(Tensorflow, OptiML …)

High Level Compiler (Delite)

Accelerators
(GPU, FPGA, SDH)