
# Microarchitecture, Computing Systems, High Performance Computing and End-to-End Deep Neural Networks

**Mauricio Breternitz**
**Instituto Superior Tecnico & INESC-ID**
**Lisbon University**

WSCAD 2018

**0ct.02.2018**

INSTITUTO SUPERIOR TÉCNICO

# Outline

- Introduction
  - Microarchitecture Trends
  - HPC systems to Exascale
  - Research Topics
    - Novel applications of end-to-end DNN
    - System Organization
    - Research Opportunities

Research Support by Fundação para a Ciência e a Tecnologia (FCT) under project UID/CEC/50021/2013

# Brief BIO, Publications, Patents

PhD – Carnegie-Mellon, ECE
MSc – UNICAMP/Brazil
BSc – ITA-Brazil

Work: IBM Research, Motorola, Times N, Intel Labs, AMD Research

50 U.S. Patents Issued, 54 U.S. Patents Pending

Publications

Citations 1393 H-index 22, i10-index 39
Computer Architecture, Computer Systems, Performance Tuning
Big Data, Machine Learning

Creator /General Chair : AMAS-BT International Workshop on Architectural/Microarchitectural Support for Binary Translation, joint with ISCA and CGO.

INSTITUTO SUPERIOR TÉCNICO
Universidade de Lisboa

# Programas de Doutoramento
## (3º Ciclo)


- Bioengenharia ✹*
- Engª Biomédica
- Biotecnologia
- Engª Química
- Química
- Engª Electrotécnica e Computadores ✹✹*
- Engª Informática e Computadores ✹✹*
- Engª Mecânica ✹
- Engª dos Materiais
- Engª Aeroespacial
- Engª Naval

- Matemática ✹✹☼*
- Estatística e processos estocásticos
- Segurança de informação
- Arquitectura *
- Eng. Civil *
- Engª do Território
- Transportes ✹
- Engª do Ambiente
- Minas e Georecursos
- Física *
- Engª Física Tecnológica
- Engª da Computação
- Engª e Gestão
- Sistemas Energia Sustentáveis ✹
- Engª e Políticas Públicas ✹✹
- Mudança Tecnológica e Empreendedorismo ✹✹
- Leaders para as indústrias de tecnologia

✹ *Com Massachusetts Institute of Technology - MIT*
✹✹ *Com Carnegie Mellon University - CMU*
☼ *Com University Texas – Austin - UTA*

* *Com Ecole Polytech. Fédérale de Lausanne - EPFL*

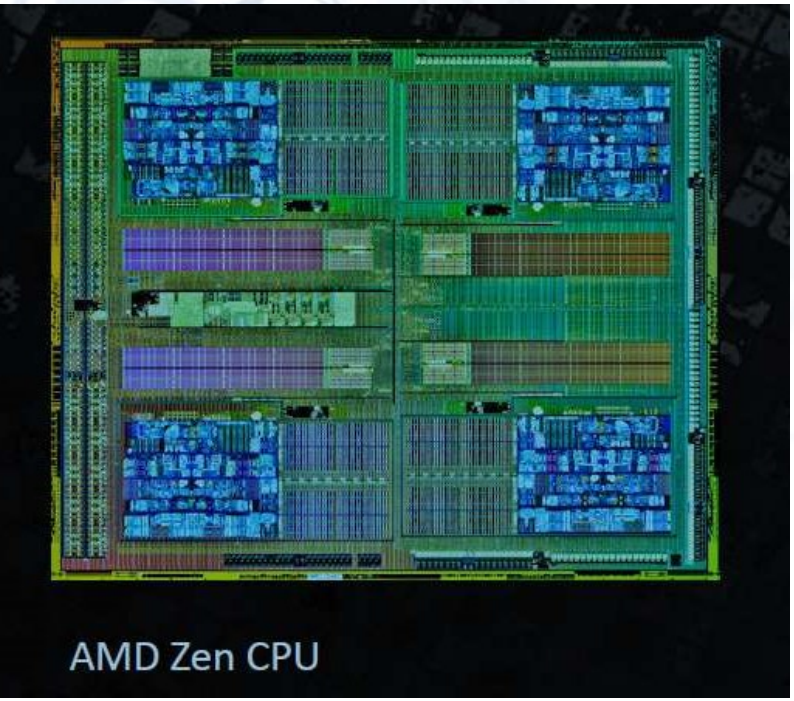
# Departamento de Engenharia Informática
# Docentes (faculty)

- Cerca de 80 professores
- PhDs das melhores escolas mundiais:
  - MIT (3 professores)
  - UC Berkeley (2 professores)
  - CMU (1 professor)
  - 1/3 dos professores com PhDs internacionais
- Reconhecimento internacional do seu trabalho:
  - ACM Distinguished Member
  - ERC Grant
  - 2016 ACM Computing Review "21st annual list of notable items published in computing”
  - Program chairs de conferências de topo (EuroSys, Eurographics, etc.)
  - Organização da Lisbon Machine Learning School desde 2011

# MICROARCHITECTURE and EXASCALE

- CPU trends
- Modern Microarchitecture
- Accelerators - GPU

# CPU Trends



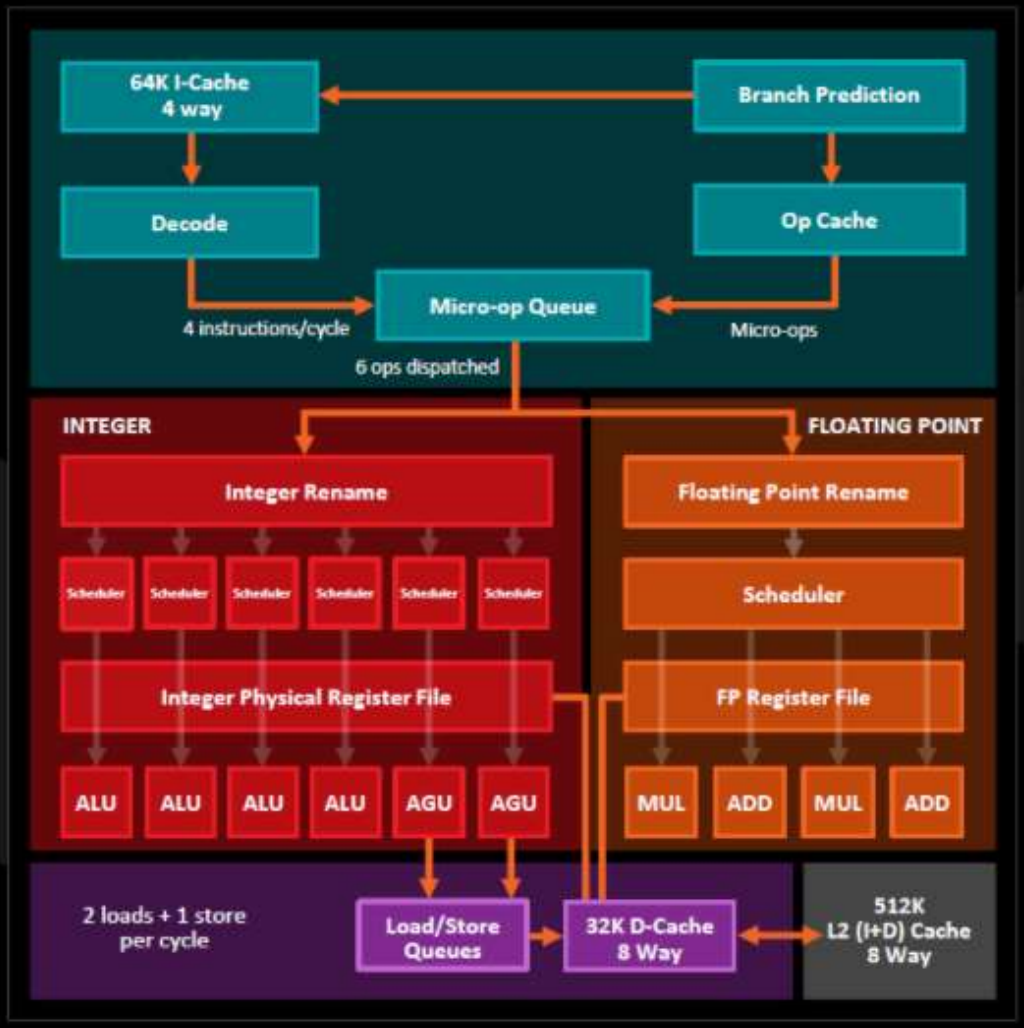

CPU architecture trends:

- Bigger pipelines
- Increased out of order execution
- Improved speculative execution
- Wider vector operations
- Memory scatter/gather instructions (vectored I/O)

These architectural features improve performance at the cost of die space and reduced energy efficiency

Increasing core counts enables parallel thread execution

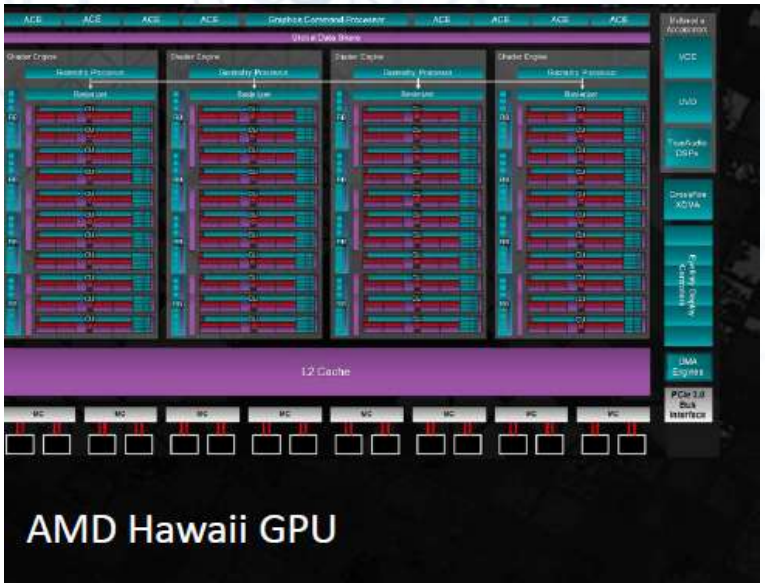
Credit: Alan Lee, AMD Research CVP, 2017




# ZEN MICROARCHITECTURE

- Fetch Four x86 instructions
- Op Cache instructions
- 4 Integer units
  - Large rename space – 168 Registers
  - 192 instructions in flight/8 wide retire
- 2 Load/Store units
  - 72 Out-of-Order Loads supported
- 2 Floating Point units x 128 FMACs
  - built as 4 pipes, 2 Fadd, 2 Fmul
- I-Cache 64K, 4-way
- D-Cache 32K, 8-way
- L2 Cache 512K, 8-way
- Large shared L3 cache
- 2 threads per core

# GPU Overview




Very high core count with highly parallel architectures

- Simplified Core architecture to reduce die space and improve energy efficiency
- Sequential code runs poorly on the GPU, although current GPUs have better support for general purpose compute
- Excellent floating point capability
- High throughput memory architecture

Programmable using OpenCL, C++ and other high level languages via OpenMP and OpenACC.

GPUs are good choices for highly parallel data processing such as signal and image processing.

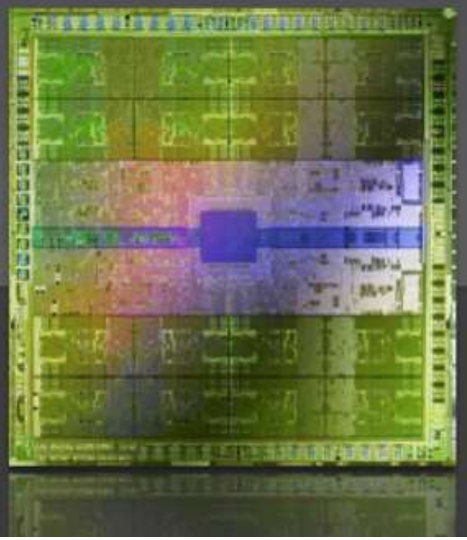
# CPU, GPU Comparison

Differences in GPU/CPU for Power Consumption
from "GPU Computing To Exascale and Beyond", Bill Dally, SC10

## GPU
200pJ/Instruction

Optimized for Throughput

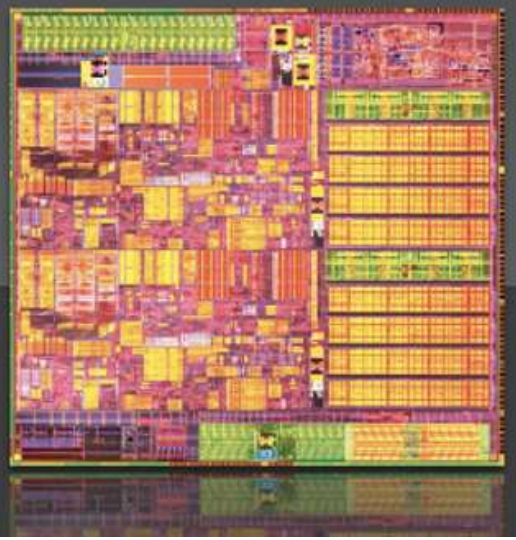
Explicit Management of On-chip Memory



## CPU
2nJ/Instruction

Optimized for Latency

Caches

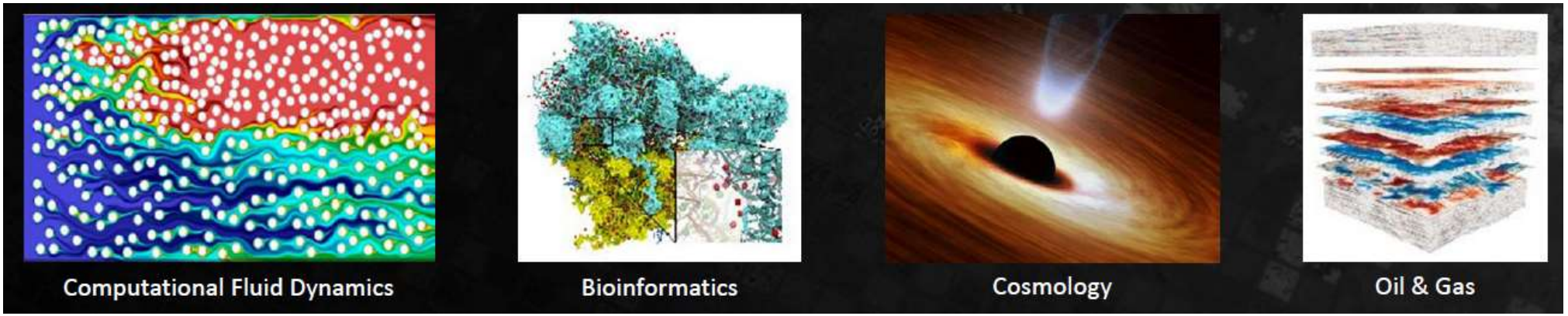


# Handling Large Data Sets at High Speed

- A conventional CPU executes *one* thread at a time
  A multi-core CPU might execute *tens* of threads at a time
  A GPU can process *thousands* of threads concurrently
  (Repurpose pixel processing for general purpose processing)

Result: Huge increase in power-performance efficiency
Highly parallel algorithms (e.g., X-correlation) experience massive acceleration

**Trend**: accelerators are increasingly deployed to attack more algorithms and problems:



# AMD High Performance Computing

- Research funded by U.S. Department of Energy
- Technology towards Exascale
- Several Programs: FastForward, FastForward2, DesignForward, PathForward
- Total AMD Awards exceed US$40M

# AAR Fast Forward 2 Node Architectures

## Investigators

- Michael Schulte (PI - Technical lead)
- John Keaty (Director)

## DOE representatives

- Scott Atchley, ORNL
- John May, LLNL

## Total Funding

- Approximately $20M

## Contract Period of Performance

- Sept 2014-Dec 2016

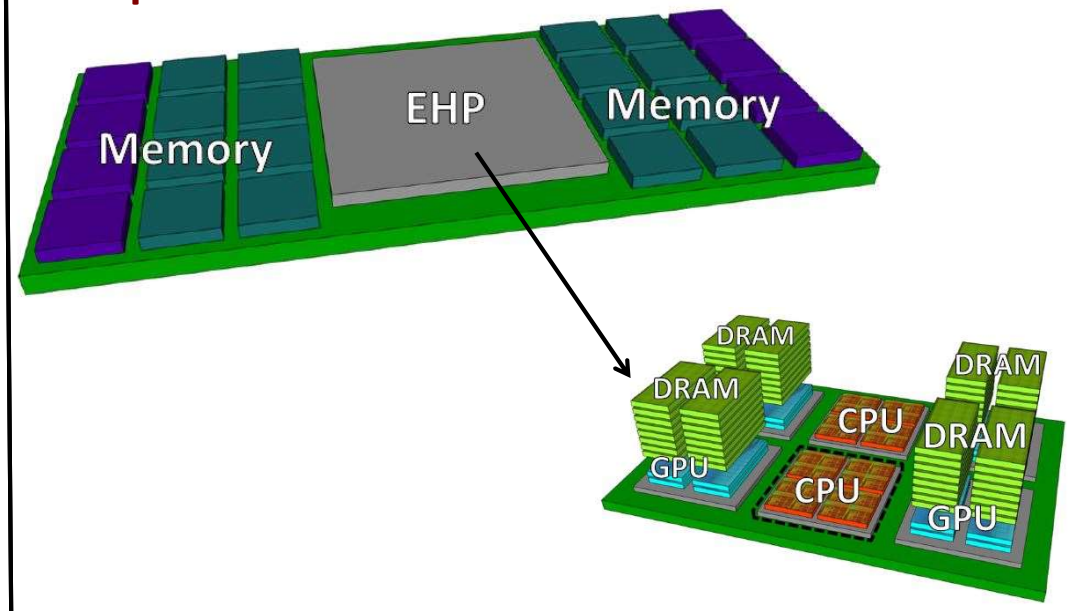
**Research and Development in:**

- Node Architecture, Integration and Evaluation
- Processor and SoC Design Enhancements
- Energy Utilization Techniques
- Resilience and Reliability Enhancements
- Data Movement and Processing-in-Memory
- Enhanced Programmability and Applications
- Simulation and Modeling Infrastructure

**Co-Design, Technology Transfers, and Interactions with System Integrators**

- Driving Research Results into Future Products

## Proposed Exascale Node Architectur

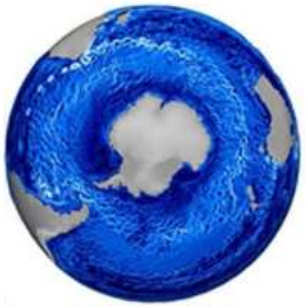


# EXASCALE SUPERCOMPUTING

# Exascale Research Areas

NUCLEAR ENERGY

Accelerate design and commercialization of next-generation small modular reactors.

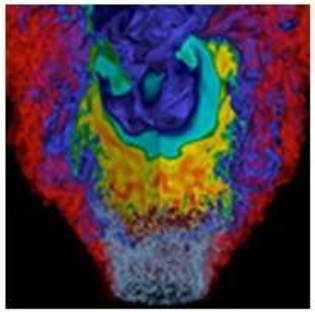


CLIMATE

Accurate regional impact assessment of climate change.

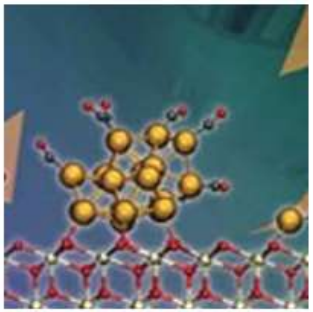
WIND ENERGY

Increase efficiency and reduce cost of turbine wind plants sited in complex terrains.



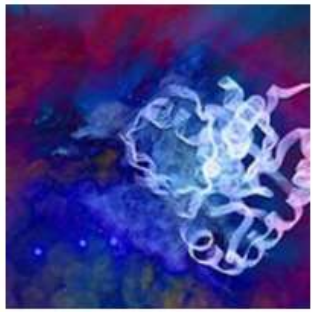
COMBUSTION

Design high-efficiency, low-emission combustion engines and gas turbines.



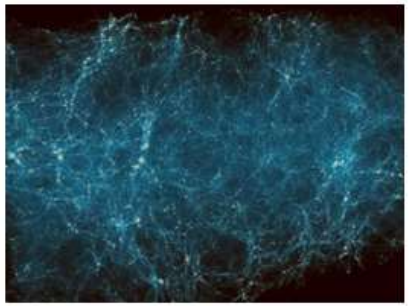
CHEMICAL SCIENCE

Biofuel catalysts design; stress-resistant crops.



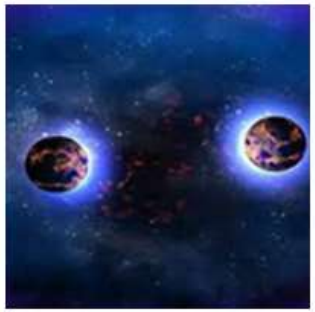
PRECISION MEDICINE FOR CANCER

Accelerate and translate cancer research in RAS pathways, drug responses, and treatment strategies.



COSMOLOGY

Cosmological probe of standard model (SM) of particle physics: inflation, dark matter, and dark energy.



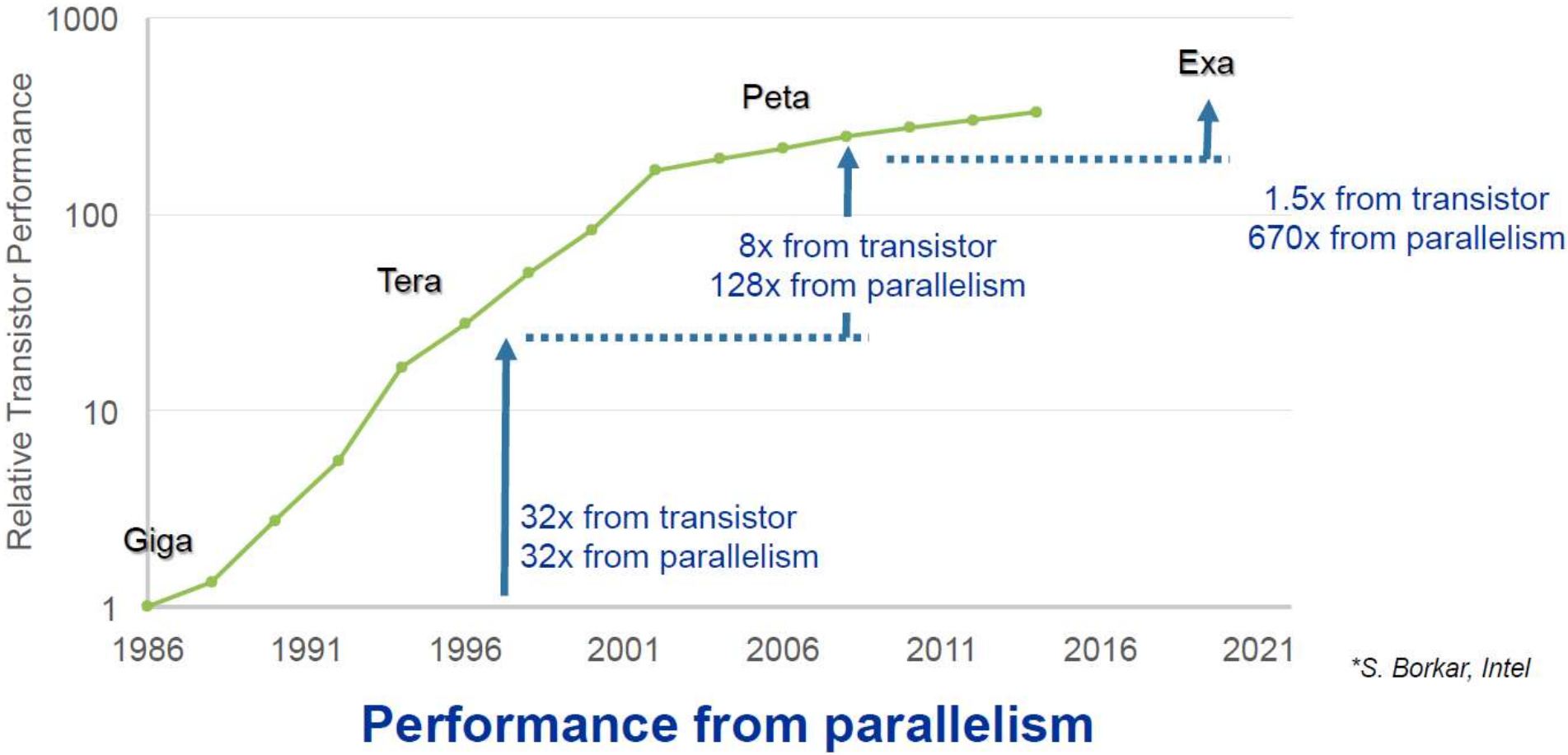
ASTROPHYSICS

Demystify origin of chemical elements (> Fe); confirm LIGO gravitational wave and DUNE neutrino signatures.

credit: Exascale Computing Project, 2017

# Achieving Exascale

## From Giga to Exa, via Tera & Peta*




Credit: Paul Messina, Exascale Computing Project, 2017

# Exascale System Specification

| Exascale System | Goal |
|---|---|
| Delivery Date | 2019-2020 |
| Performance | 1000 PF LINPACK and 300 PF on to-be-specified applications |
| Power Consumption* | 20 MW |
| MTBAI** | 6 days |
| Memory including NVRAM | 128 PB |
| Node Memory Bandwidth | 4 TB/s |
| Node Interconnect Bandwidth | 400 GB/s |

*Power consumption includes only power to the compute system, not associated storage or cooling systems.

**The mean time to application failure requiring any user or administrator action must be greater than 24 hours, and the asymptotic target is improvement to 6 days over time. The system overhead to handle automatic fault recovery must not reduce application efficiency by more than half.

PF = petaflop/s, MW = megawatts, PB = petabytes, TB/s = terabytes per second, GB/s = gigabytes per second, NVRAM = non-volatile memory.

# Computing Devices

| Device | Ease of Programmability | Application Flexibility | Floating Point Capability | Energy Efficiency |
|---|---|---|---|---|
| CPU | Easy | High | 100's GFLOPS range | Low |
| GPU | Moderate | High | 10's TFLOPS range | Low |
| DSP | Moderate | High | 10's GFLOPS | Moderate to High |
| FPGA | Difficult | Moderate | Algorithm Specific | Moderate |
| ASIC | Very difficult | Low | Algorithm Specific | High |

- CPU, GPU, and DSP architectures are closest, differing on parallelism and control. What differentiates them is how the microarchitectures are combined with each other (parallelism and control) and with memory and I/O.

- FPGAs provide a semi-flexible solution where digital logic design is used to implement algorithms and I/O for a specific task. Modern FPGAs include a number of hardware multiply units that make them suitable for algorithms such as the FFT.

- ASICs are custom chips that can achieve better performance than FPGAs. They are suitable for well defined algorithms.

Credit: Alan Lee, AMD Research SVP, 2017

# Technology Trends

- Heterogeneous computing and accelerators
- Increased on-chip integration (e.g. CPU + GPU on the same die)
- Increased on-package integration using multiple chips on an interposer. E.g: CPU + NIC + memory
- 3D or Die-stacking: Stacked memory chips and logic chips (as seen in recent GPU products such as AMD's Fiji)
- Higher core counts
- New memory technologies (e.g. NVRAM, stacked memory)
- Faster interconnects
- New programming paradigms: C++ AMP, OpenMP and OpenACC standards

Credit: AMD Research, 2017

# U.S.Dept of Energy Exascale Project

- 10 year project – two Exascale Computers by 2023
- National Strategic Computing Initiative (NSCI) launched by the Obama Administration in July 2015
  - FastForward
  - FastForward2
  - DesignForward
  - PathForward

# U.S.Dept of Energy Exascale Project

- FastForward
  - initiate partnerships with multiple companies to accelerate the R&D of critical component technologies needed for extreme-scale computing
  - **NVIDIA, IBM**, **Intel**, **AMD**,**WhamCloud**
- FastForward2
  - focuses on two areas: Node Architecture and Memory Technology
- DesignForward
- PathForward

# U.S. Department of Energy Exascale Program

| | RFP | Awardees | Total (US$ million) |
|---|---|---|---|
| • FastForward | 2011 | NVIDIA AMD Intel IBM WhamCloud | 62.0 |
| • DesignForward | 2013 | AMD Cray IBM Intel NVIDIA | 25.4 |
| • FastForward2 | | NVIDIA AMD Intel IBM | 100.0 |
| • DesignForward2 | 2014 | AMD, Cray, IBM | 20.0 |
| • PathForward | 2016 | AMD, Cray, IBM, HP, Intel, Nvidia | 258.0 |

# ON THE PATH TO THE NATION'S FIRST EXASCALE SUPERCOMPUTERS: PATHFORWARD

**06/15/17**

## Department of Energy Awards Six Research Contracts Totaling $258 Million to Accelerate U.S. Supercomputing Technology

**June 15, 2016**

**WASHINGTON, D.C.** – Today **U.S. Secretary of Energy Rick Perry** announced that six leading U.S. technology companies will receive funding from the Department of Energy's Exascale Computing Project (ECP) as part of its new PathForward program, accelerating the research necessary to deploy the nation's first exascale supercomputers.

The awardees will receive funding for research and development to maximize the energy efficiency and overall performance of future large-scale supercomputers, which are critical for U.S. leadership in areas such as national security, manufacturing, industrial competitiveness, and energy and earth sciences. The $258 million in funding will be allocated over a three-year contract period, with companies providing additional funding amounting to at least 40 percent of their total project cost, bringing the total investment to at least $430 million.

"Continued U.S. leadership in high performance computing is essential to our security, prosperity, and economic competitiveness as a nation," said Secretary Perry.

"These awards will enable leading U.S. technology firms to marshal their formidable skills, expertise, and resources in the global race for the next stage in supercomputing—exascale-capable systems."

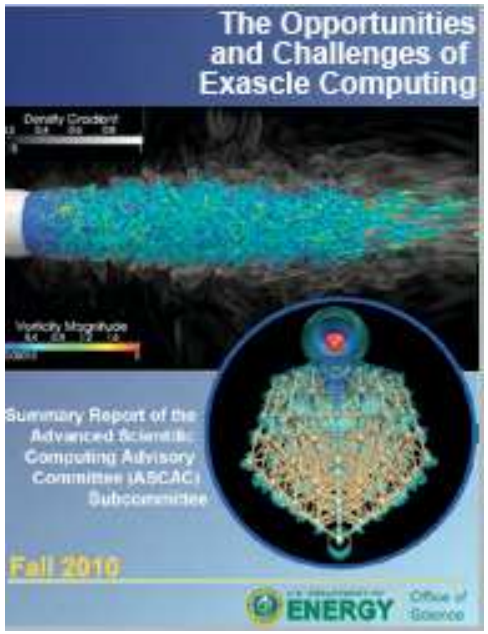
"The PathForward program is critical to the ECP's co-design process, which brings together expertise from diverse sources to address the four key challenges: parallelism, memory and storage, reliability and energy consumption," ECP Director Paul Messina said. "The work funded by PathForward will include development of innovative memory architectures, higher-speed interconnects, improved reliability systems, and approaches for increasing computing power without prohibitive increases in energy demand. It is essential that private industry play a role in this work going forward: advances in computer hardware and architecture will contribute to meeting all four challenges."

The following U.S. technology companies are the award recipients:

- Advanced Micro Devices (AMD)
- Cray Inc. (CRAY)
- Hewlett Packard Enterprise (HPE)
- International Business Machines (IBM)
- Intel Corp. (Intel)
- NVIDIA Corp. (NVIDIA)

# Exascale initiatives are advancing the computational power of supercomputers

- NSCI (National Strategic Computing Initiative) announced by President Obama in June 2015
- Exascale Computing Project ECP started by DOE in the US
- Similar initiatives in Europe, Japan, and China






English.news.cn
新华网 NEWS
Search
Advanced Sea
Services | Database | Markets | Weather | Site Index | RSS | Feedback
Global Edition | China | World | Business | Culture & Edu | Sports | Entertainment | Science & Tech
Most S

**China planning new supercomputer**

Source: Xinhua 2016-01-22 19:26:01

TIANJIN, Jan. 22 (Xinhua) -- China is planning a supercomputer 1,000 times more powerful than its groundbreaking Tianhe-1A as it faces rising demand for next-generation computing.

Meng Xiangfei, head of the applications department of the National Supercomputer Center, said on Friday that the center will release a prototype in 2017 or 2018 of an "exascale" computer -- one capable of at least a billion billion calculations per second

Exascale computing is considered the next frontier in the development of supercomputers.

Thanks: Horst Simon, Berkeley Lab

# MACHINE LEARNING

- “Program From Data”
- Software 2.0
- Key Algorithms – Deep Neural Networks

# DEFINITION OF MACHINE LEARNING

- Simple Definition: "Algorithms that Learn

Traditional Programming


In Traditional Programming, a Human expert encodes his knowledge of the relationship of data and desired output as a program to process input data to generate the desired output

Machine Learning


In Machine Learning, the system autonomously learns the relationship of data and the desired output, creating classification rules (inference) to provide the desired output from similar input

◢ Machine Learning: A system capable of the autonomous acquisition and integration of knowledge

**Challenge: Black Box -> Hardware**

# Software 1.0 vs Software 2.0


- Written in code (C++, …)
- Requires domain expertise
  1. Decompose the problem
  2. Design algorithms
  3. Compose into a system

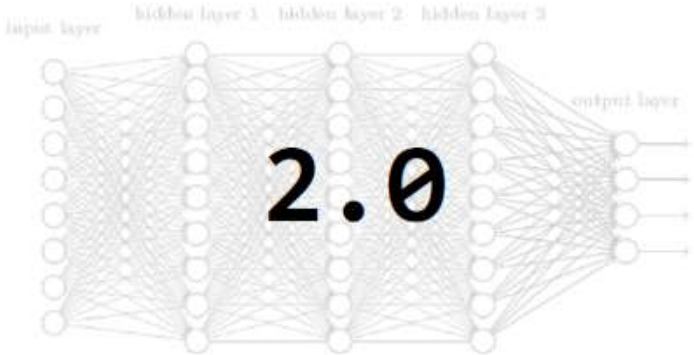
- Written in the weights of a neural network model by optimization

Andrej Karpathy
Scaled ML 2018 talk

ISCA 2017 Keynote – Kunle Olukotun

# Training Data: The New Input to Software 2.0


- Input: Algorithms in code
- Compiled to: Machine instructions




- Input: Training data
- Compiled to: Learned parameters

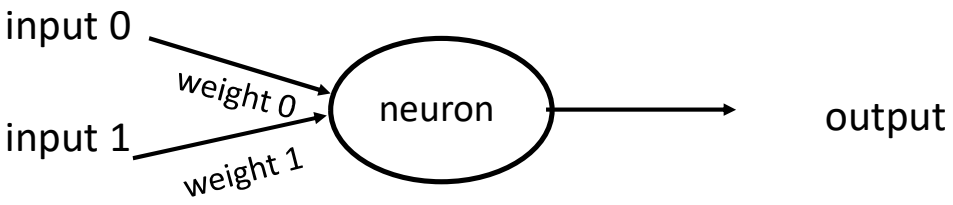
https://medium.com/@karpathy/

# KEY MACHINE LEARNING ALGORITHMS

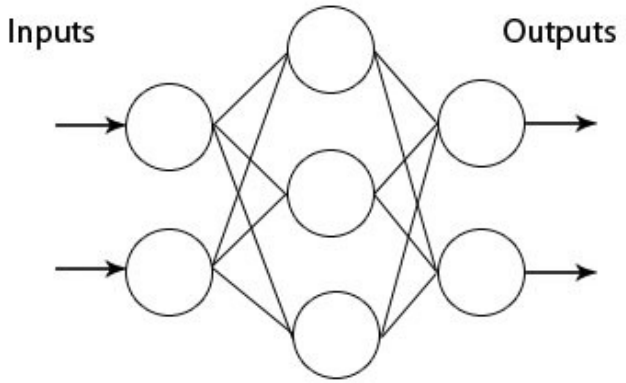
- *Classes of Machine Learning Algorithms:*
  - *Statistically-inspired algorithms: Bayesian Networks, Logistic Regression, Decision Trees, etc.*
  - *Deep Neural Networks(DNN)*
    - *rapidly becoming the preferred algorithm, currently provide the best solutions for image/speech/natural language processing*
    - *Biologically-inspired: simulated neurons*
    - *DNNs are a good match for heterogeneous (GPU, FPGA) acceleration because the mathematical operation to compute the effects of weighted inputs for multiple neurons is a matrix-vector multiplication.*

# DEEP NEURAL NETWORKS

- *rapidly becoming the preferred algorithm, currently the best solutions for image/speech/natural language processing*
- *Biologically-inspired: simulated neurons*
- *Good match for AMD GPU acceleration because the mathematical operation to compute the effects of weighted inputs for multiple neurons is a matrix-vector multiplication.*

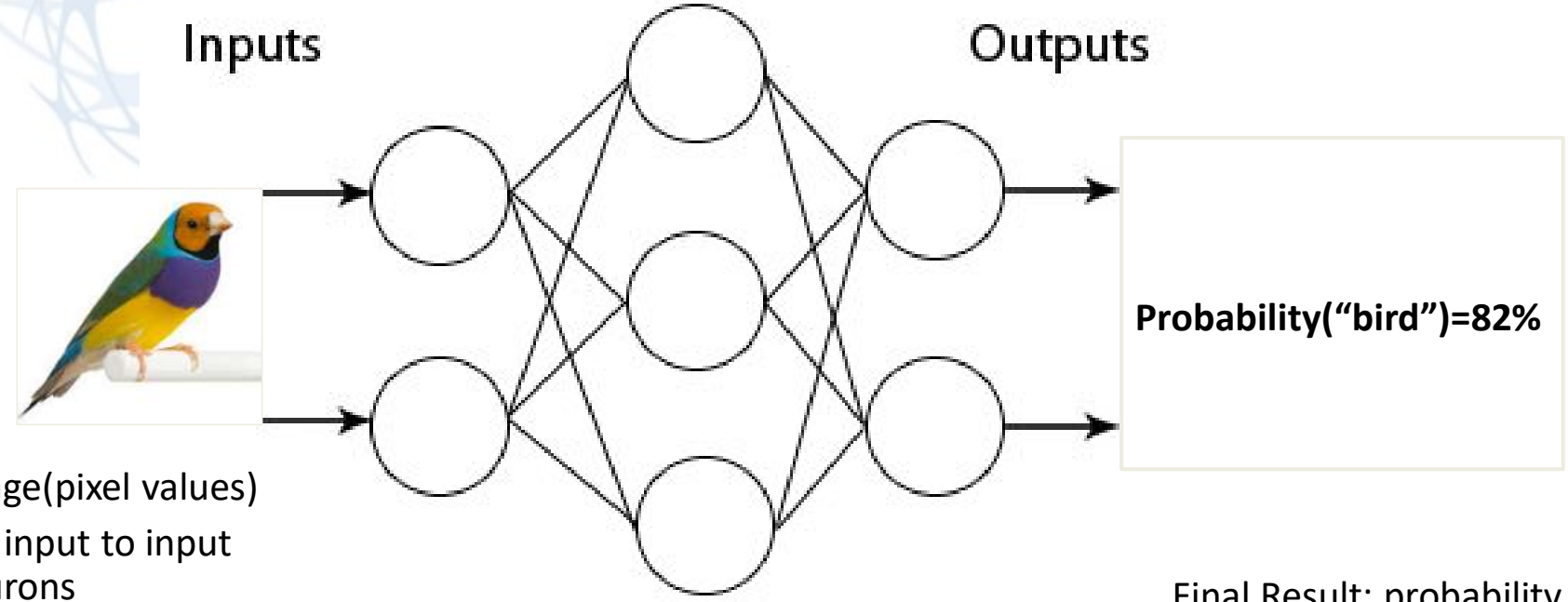



*A Simulated Neuron: A biologically inspired algorithm whereby a number of input values are provided to a simulated neuron, which computes an output based on a **weighted** combination of the input values*




*An Example Deep Neural Network(DNN): A multi-layered sequence of simulated neurons*

# EXAMPLE: DEEP NEURAL NETWORK CLASSIFYING AN IMAGE




Image(pixel values) are input to input neurons

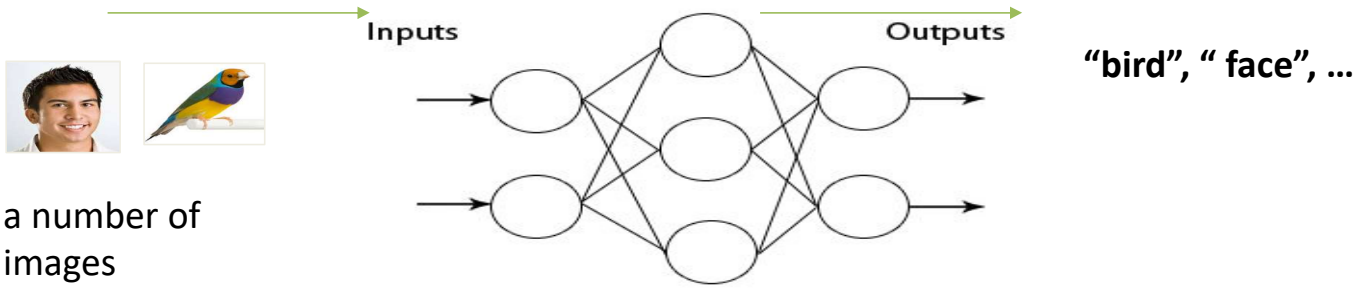
Groups of neurons are trained to detect specific features in image regions.

Subsequent layers reacts to combinations of features that compose the image.

Final Result: probability of input image being a given entity

# INFERENCE

- **Is** the problem of identifying to which categories a new observation belongs
  - Examples - Sort and Classify input into discrete categories
    - Create photo categories from input set (exemplified on previous slide)
    - Email: {message} classified as one of { spam, NOT-spam}
    - Diagnosis: { gender, symptoms} used to determine disease
  - Uses *Trained* deep networks for RECOGNITION tasks
  - Focus on efficient Forward computation
  - Focus on Latency: Minimize end-do-end response time: smaller mini-batches

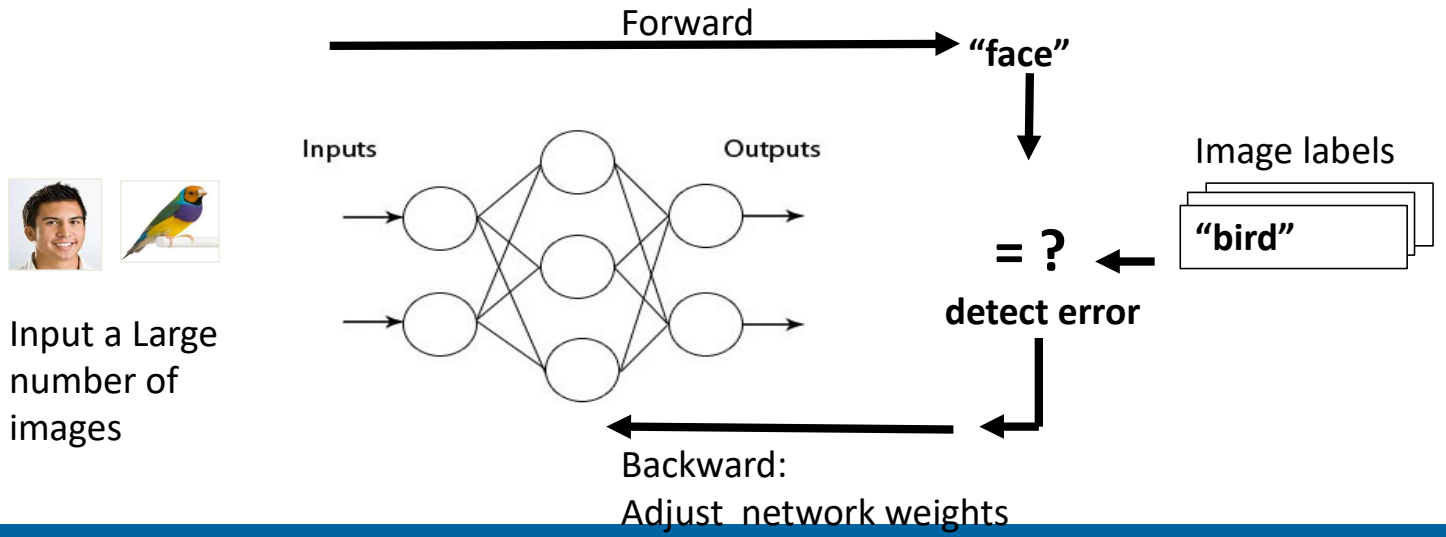



*Inference: images are presented to the network to determine what class of image it is*

# TRAINING

- or LEARNING is the computationally-demanding task of determining the parameters of a neural network
  - Has both Forward, Backward propagation phases
  - State of the Art is GPU Acceleration
  - Focus on High Throughput

*Images are presented to the network to determine class; erroneous outputs are propagated backwards correcting network parameters to achieve high accuracy:*



# MACHINE LEARNING:

## APPLICATION FRAMEWORK across MULTIPLE REGIONAL SCOPES

data travel distance

**Point of Data Capture**

Compute reaction nearby the place data is captured, using locally stored knowledge

Examples - Car warning system using front camera image; Factory Sensor Input

**Local/Regional Data Center**

Compute reaction using aggregate data from a few, logically neighboring compute sites

Example: Traffic management and scheduling of a city-wide fleet

**Scale-Out Data Center**

Compute reaction using knowledge from a worldwide and long term memory

Example: Netflix movie recommendation engine

# Accelerators for ML



**CPU**
- Threads
- SIMD

**GPU**
- Massive threads
- SIMD
- HBM

**FPGA**
- LUTs
- DSP
- BRAM

**TPU**
- MM unit
- BRAM

**What next?**

# End To End Machine Learning

Self Driving Car Example:

- Map camera pixels to steering command
- System learns internal representation
- Avoids explicit system decomposition
  - Lane marking, path planning, control

Efficiency:

- Optimized for maximal overall performance
- Enable smaller networks

# Training and Inference

## Data Collection

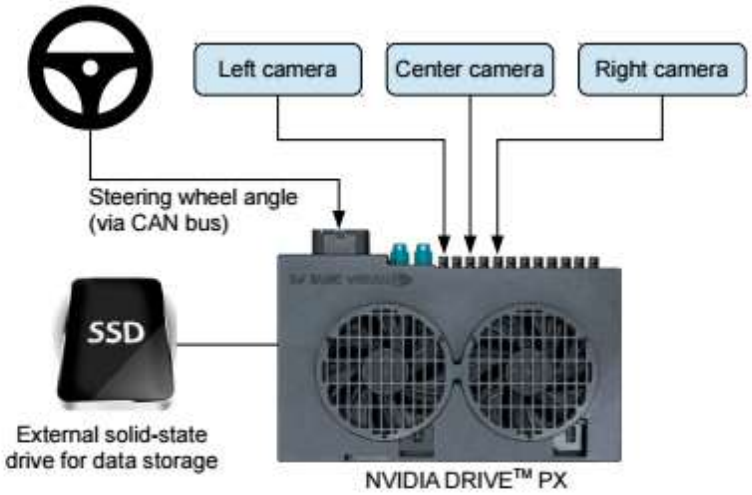
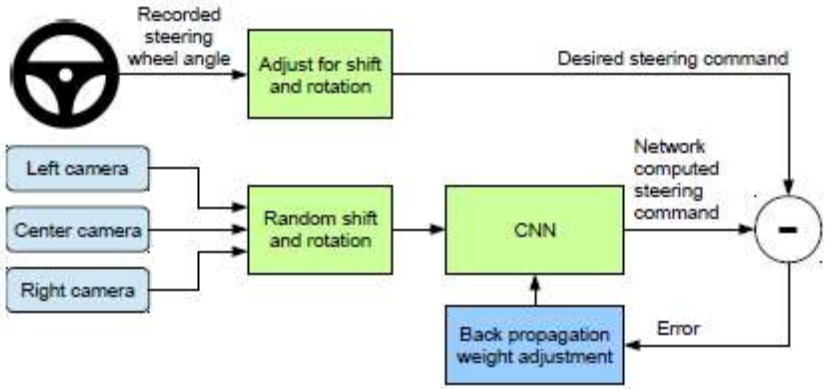



Figure 1: High-level view of the data collection system.

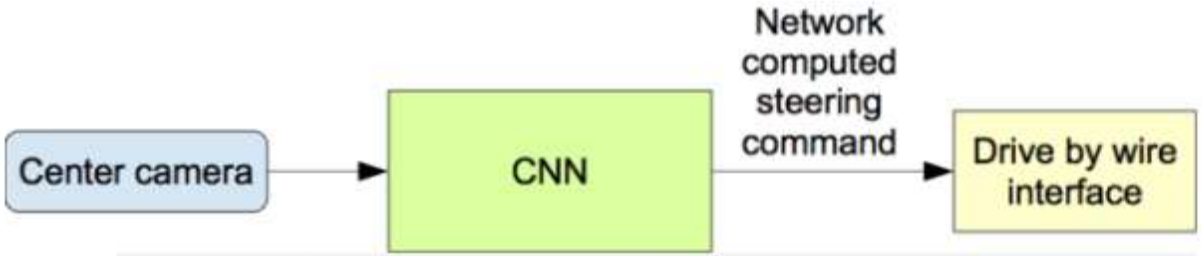
## Training




Training the neural network.

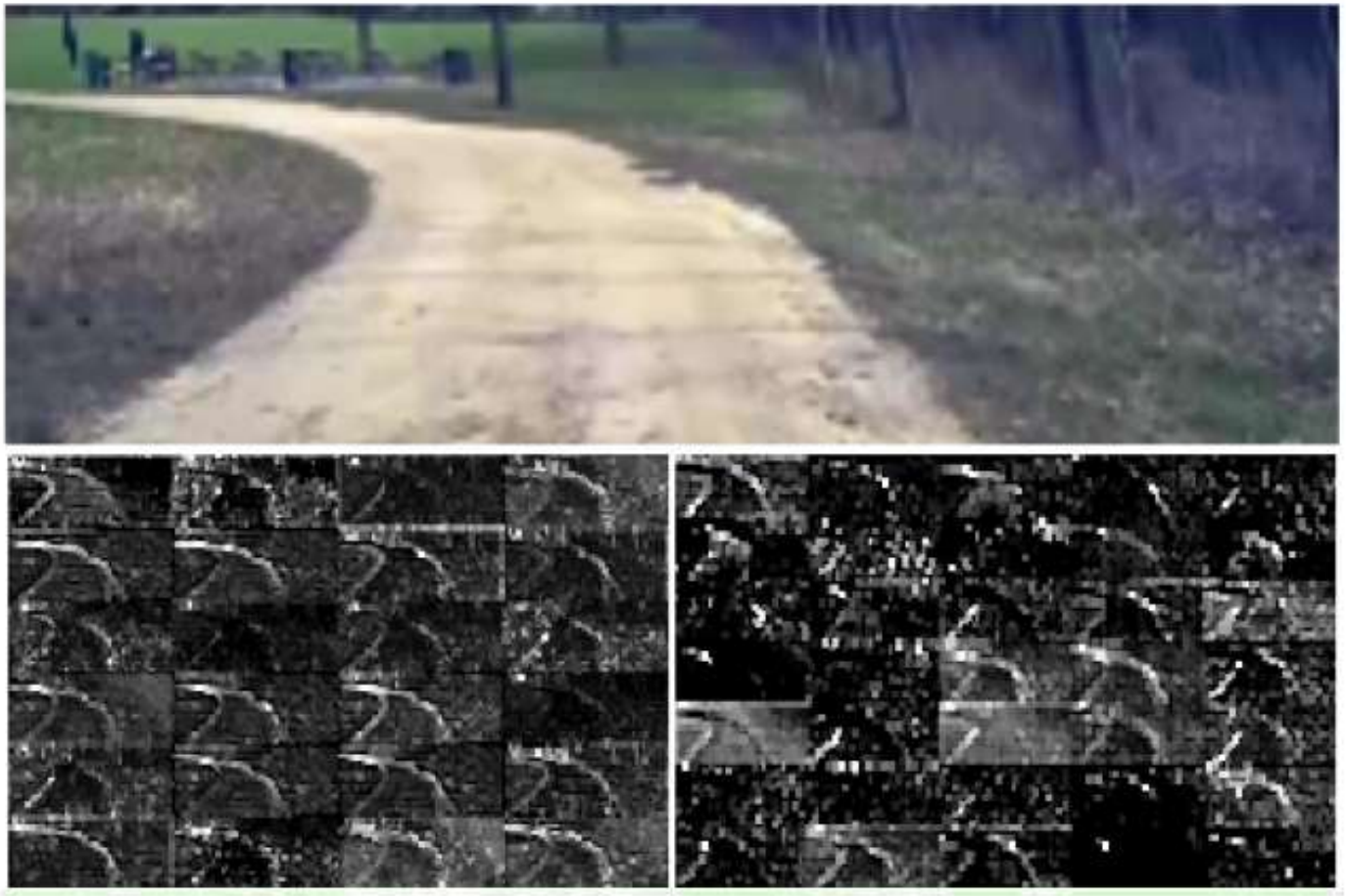
# Training and Inference

## Inference



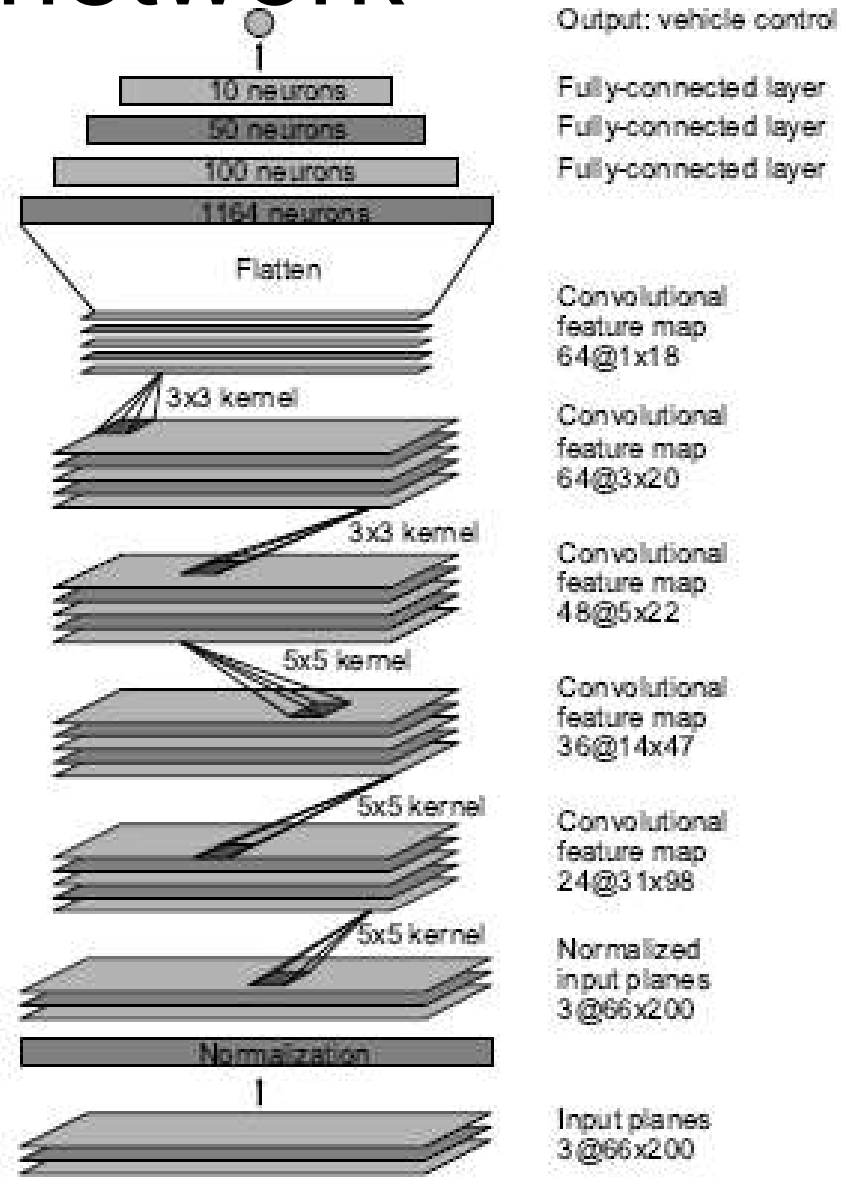

The trained network is used to generate steering commands from a single front-facing center camera.

# Road Image Example



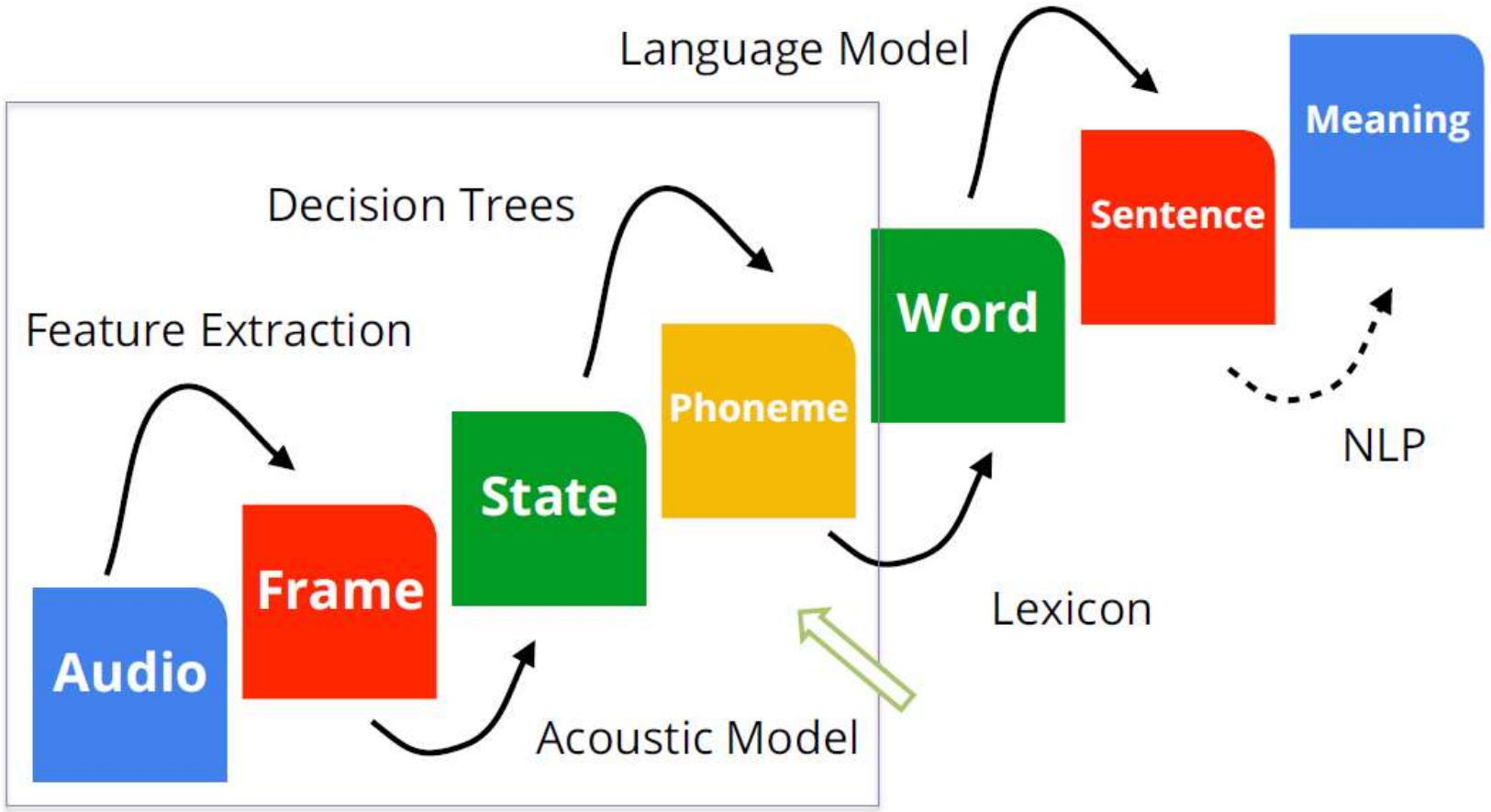
How the CNN "sees" an unpaved road. Top: subset of the camera image sent to the CNN. Bottom left: Activation of the first layer feature maps. Bottom right: Activation of the second layer feature maps. This demonstrates that the CNN learned to detect useful road features on its own, i. e., with only the human steering angle as training signal. We never explicitly trained it to detect the outlines of roads.

# CNN network



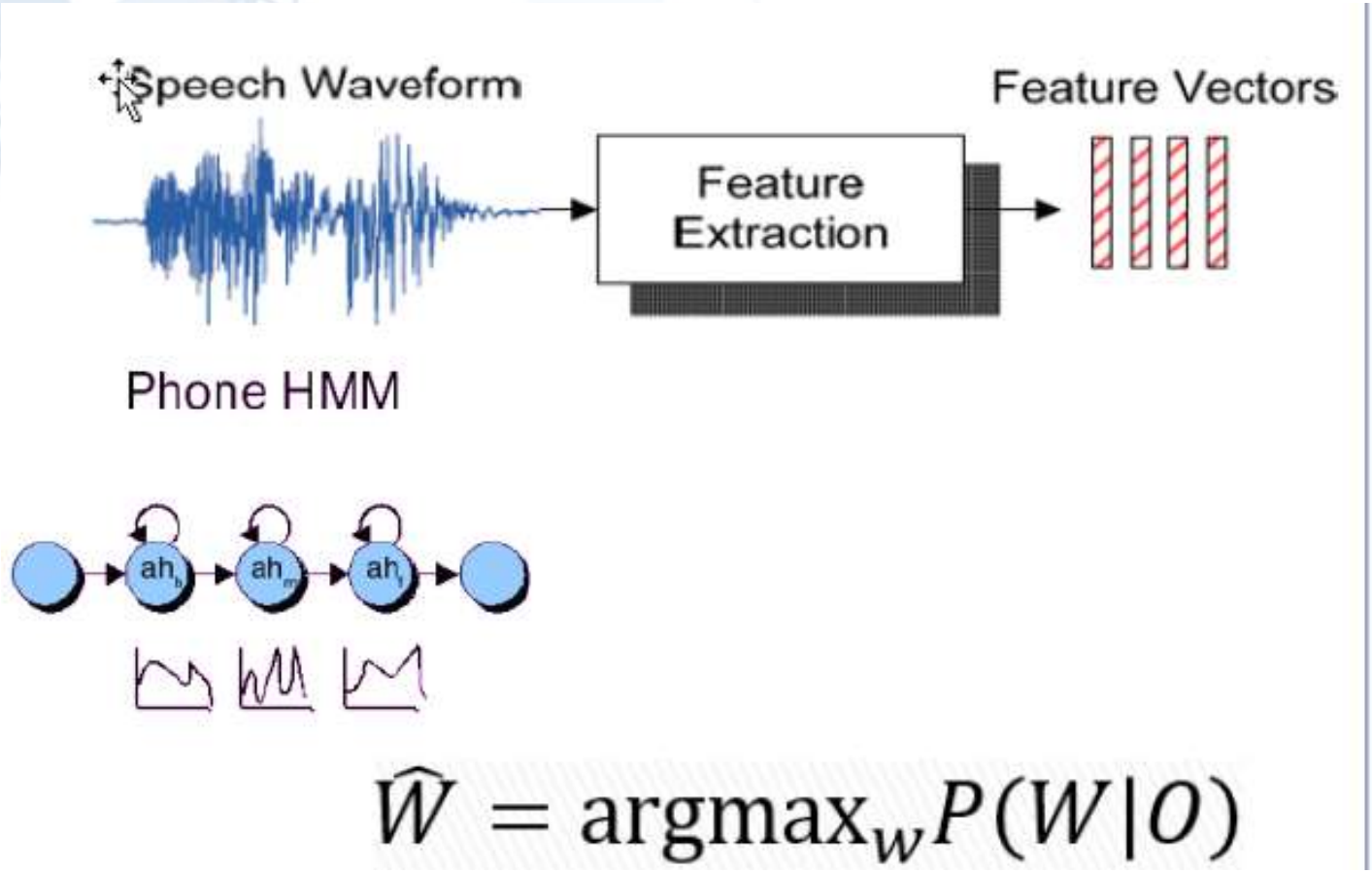

CNN architecture. The network has about 27 million connections and 250 thousand parameters.

# Automated Speech Recognition




* Slide from V. Vanhoucke, ICML 2013 Keynote

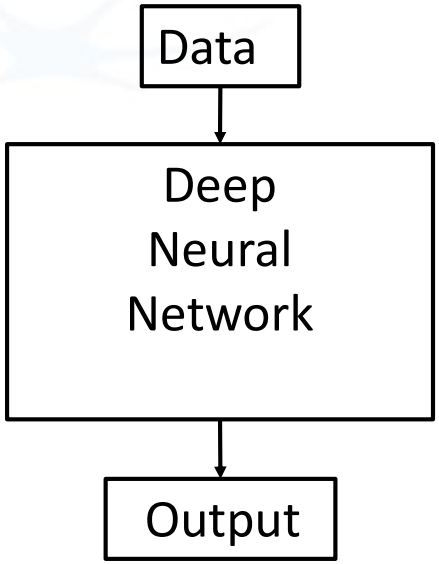
# Automated Speech Recognition



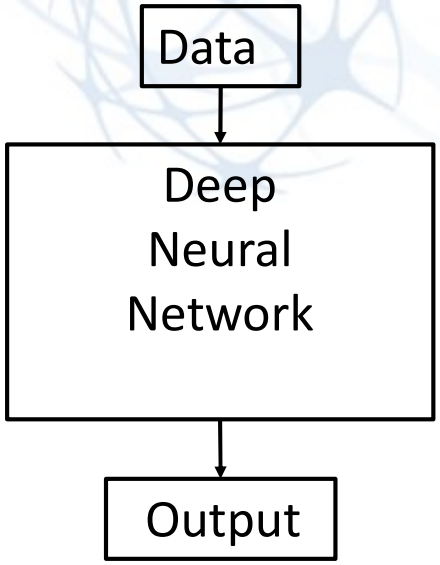

$$\hat{W} = \text{argmax}_w P(W|O)$$

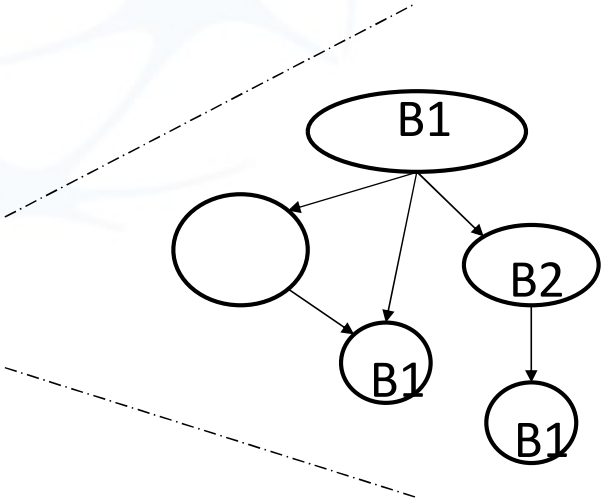
# End-To-End DNN

- Idealized Framework



# DNN Inspection & Introspection




B1 -parallel

B2 –latency sensitive

B3 – allows low precision

# End-to-End ML Challenges

-Network Inspection

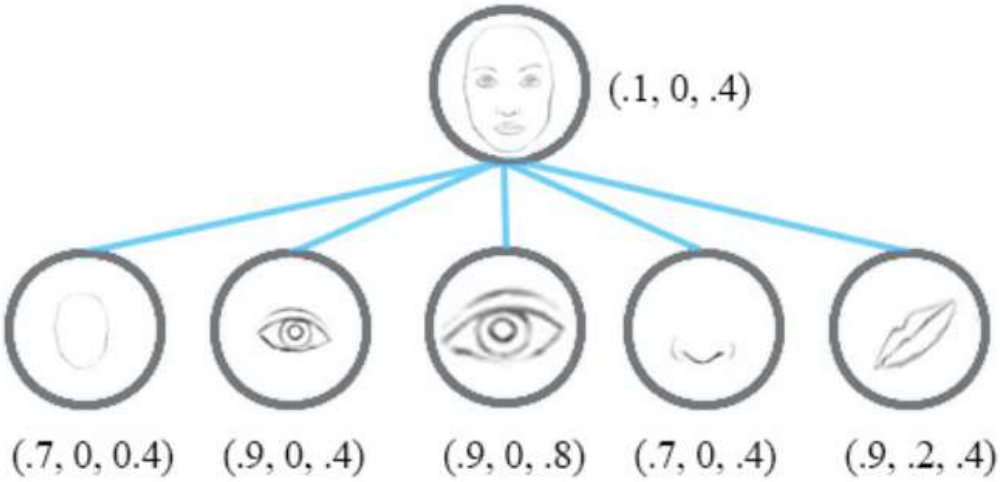
-Who does what?

-Adversarial Input

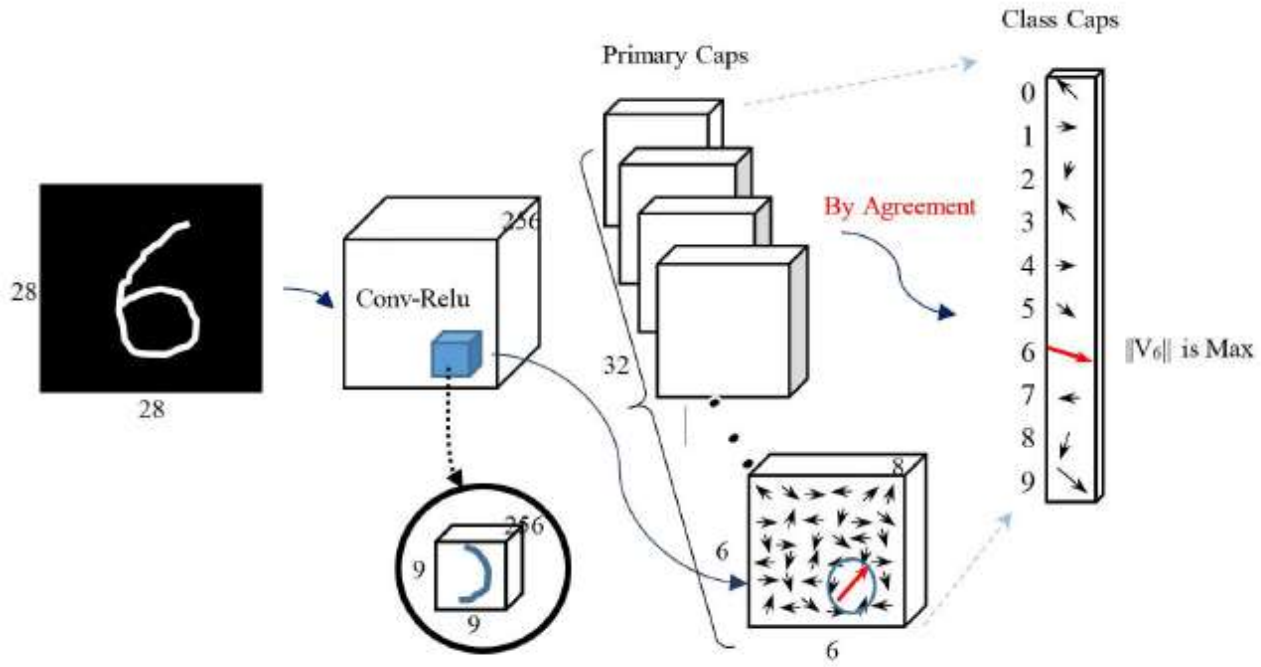
# Capsule Networks

Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." *Advances in Neural Information Processing Systems*. 2017.

*A capsule is a group of neurons that not only capture the likelihood but also the parameters of the specific feature.*




# Capsule Networks connectivity



# Conjecture

- End-To-End DNN may lead to globally optimized results
  - Tailored code generator optimizing across layers may be better than hand-tuned generic libraries
  - Trained Network: tailored interconnect

# Conjecture2

- Capsule Networks enable improved introspection
  - Tailored HW and SW stack
    - Precision –
      - how many precision bits? Inter-phase data communication
    - Interconnection – data flow
    - Code Generation
    - Reliability – tolerance to HW failure
    - Resilience – resist adversarial data

# ML Application Components

- Data Preparation

  *acquiring, producing, cleaning*

  *ENOUGH data to feed ML algorithm*

- Feature Selection and Extraction

  *identify data characteristics and behaviors of interest: what key data aspects*

- Productization/Deployment

  *deploy a stable system at SCALE;*

  *deal with data variations over time*

*(model drift: phenomena evolve & models dont)*

# End-to-End ML Components

-The **data pipeline**

clean, perhaps labeled, accessible dataset;
message queue,
storage,
preprocessing (such as normalization and vectorization)

-The choice of **algorithms and their tuning**

choice of deep network topology
hyper-parameter optimization

-The **hardware** associated with the training of algorithms;

- Visualization / actuation/ communication **of results**

# Data Preparation Challenges

- End-to-End DNN -> handling multi-modal data types
  - Video, Audio, Streaming, IoT, etc…
- Handling Data Artifacts
  - Data Normalization

# Protocol Buffers as Canonical Data Representation

- Protocol Buffers
  - Standardized data transfer format for data centers

- DER (Distinguished Encoding Rules)

# Research Approach

- Canonical Data Representation
- Inspect/Introspection of Trained Network
- Global Code Generation Approach

- Related Work: TVM, Dawn

# Software Stack

Data Ingestion & Canonical Representation

Graph Optimizer – connectivity, operator merging

Tensor-Level Optimizer – memory, precision, schedule

JIT Runtime

ISA

- Related: TVM, DAWN, DLVM

# Graph Optimizer

- Computation
  - Operator Merging
  - Precision
    - Bit-width determination
    - Data Transducer
- Memory
  - Data Access
  - Data Layout
  - Reuse
- Partitioning & Scheduling

# IR

- Potential Reuse

- Halide, TVM, Weld, Tensorflow XLA, Intel GraphN, DLVM, Glow, DLVM

# Conclusion

- Research topics:
  - Full-stack:  HW, SW, Application, system
  - Cross-layer optimization
  - Societal Applications