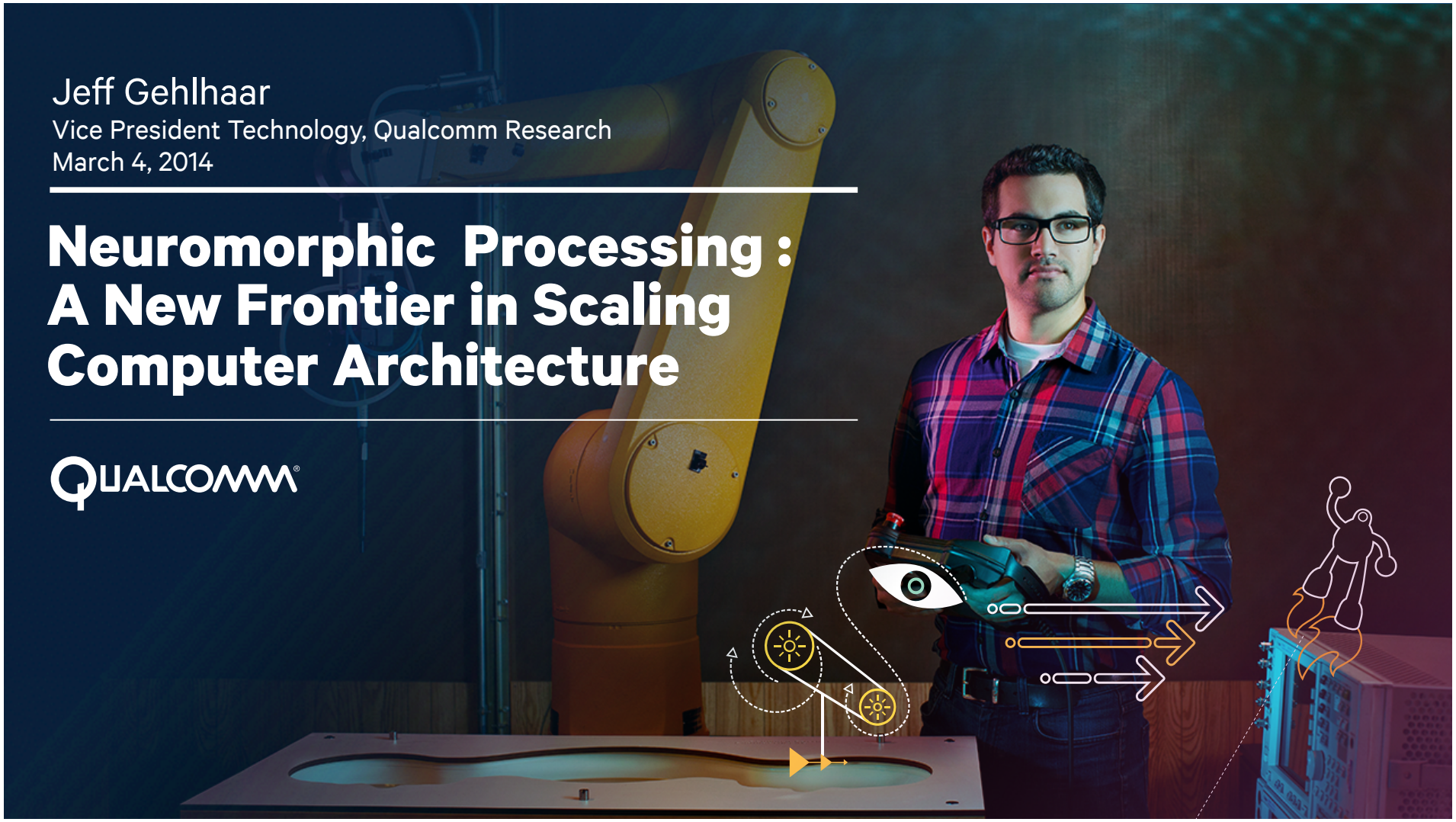
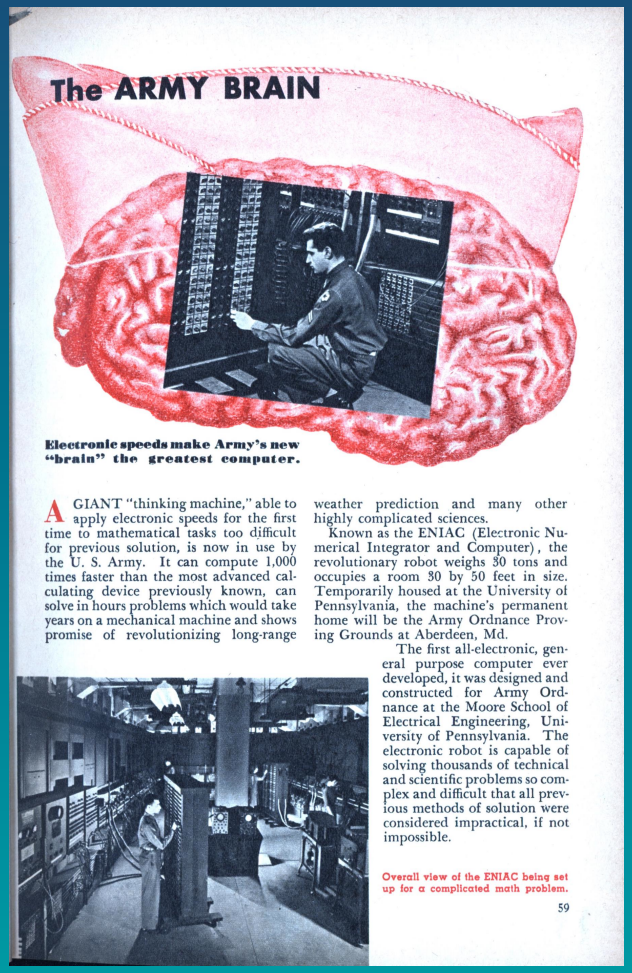
Jeff Gehlhaar
Vice President Technology, Qualcomm Research
March 4, 2014

Neuromorphic Processing : A New Frontier in Scaling Computer Architecture

QUALCOMM®



The desire to build brain-like computers is as old as the computer itself



The ARMY BRAIN

Electronic speeds make Army's new "brain" the greatest computer.

A GIANT "thinking machine," able to apply electronic speeds for the first time to mathematical tasks too difficult for previous solution, is now in use by the U. S. Army. It can compute 1,000 times faster than the most advanced calculating device previously known, can solve in hours problems which would take years on a mechanical machine and shows promise of revolutionizing long-range weather prediction and many other highly complicated sciences.

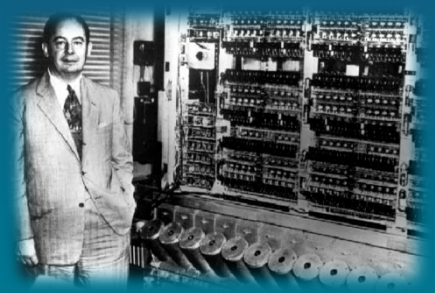
Known as the ENIAC (Electronic Numerical Integrator and Computer), the revolutionary robot weighs 30 tons and occupies a room 30 by 50 feet in size. Temporarily housed at the University of Pennsylvania, the machine's permanent home will be the Army Ordnance Proving Grounds at Aberdeen, Md.

The first all-electronic, general purpose computer ever developed, it was designed and constructed for Army Ordnance at the Moore School of Electrical Engineering, University of Pennsylvania. The electronic robot is capable of solving thousands of technical and scientific problems so complex and difficult that all previous methods of solution were considered impractical, if not impossible.

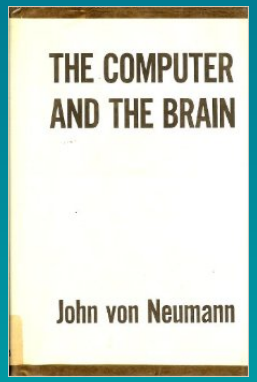
Overall view of the ENIAC being set up for a complicated math problem.

59

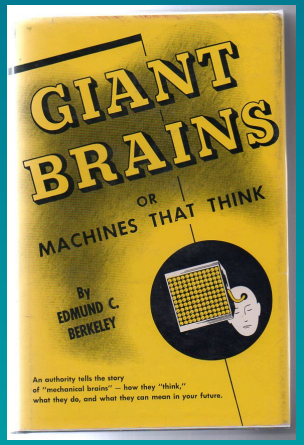
ENIAC – 1946



IAS - 1945

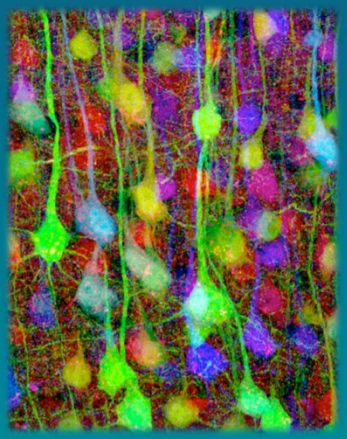


1958



Brain-inspired computing: why?

What makes the brain special?



Fault tolerance

Inputs may be noisy
Neurons may die

Power efficiency

20 W power consumption
Always On

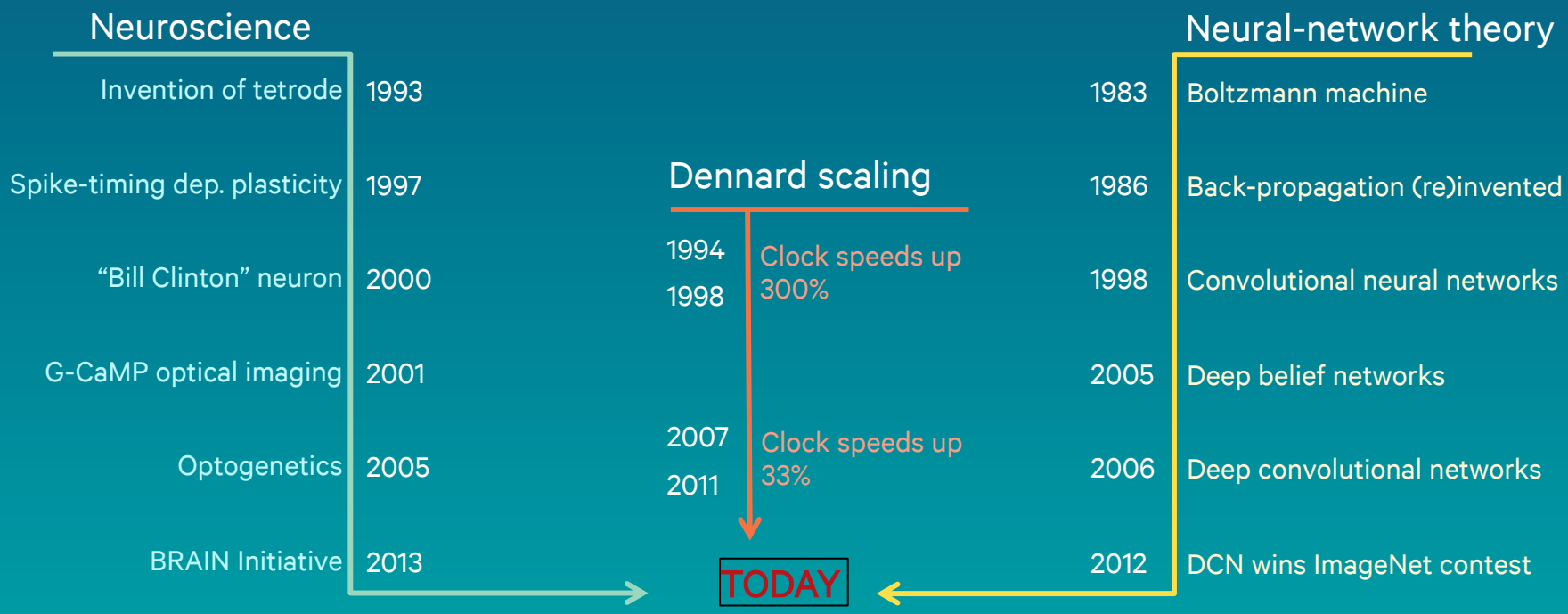
Learning ability

Supervised learning
Unsupervised learning
Reinforcement learning

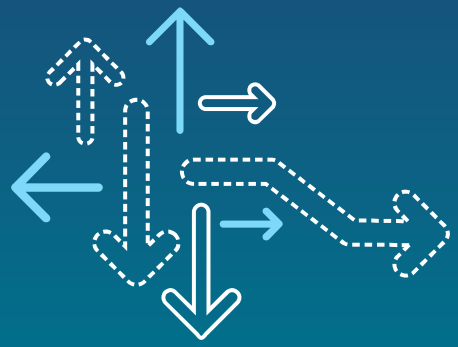
Object identification
Language acquisition
Motor learning

Brain-inspired Computing: Why Now?

A merger of three trends



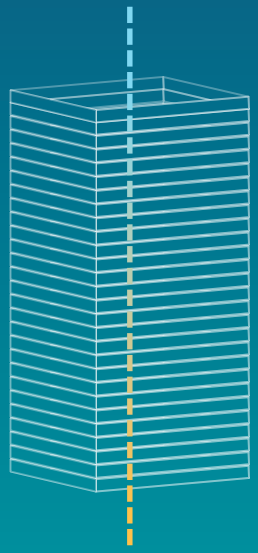
So is now finally the time for artificial cognition?



The brain is a massively parallel machine

$>10^6$ processing steps

$<10^1$ parallelism



Modern computer
Dense, real-valued data

$<10^1$ processing steps

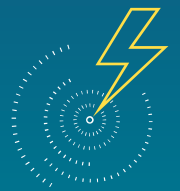
$>10^6$ parallelism

Human brain
Sparse ‘Events” or “Spikes”

Mobile is the most challenging design environment

Power efficiency

Always on



Performance

Small form factor

Qualcomm Zeroth™

Enabling perception and cognition for smart devices

Power efficiency

Performance

Natural, contextual sensor processing

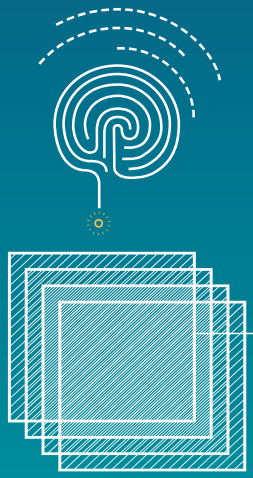
Cognitive algorithms

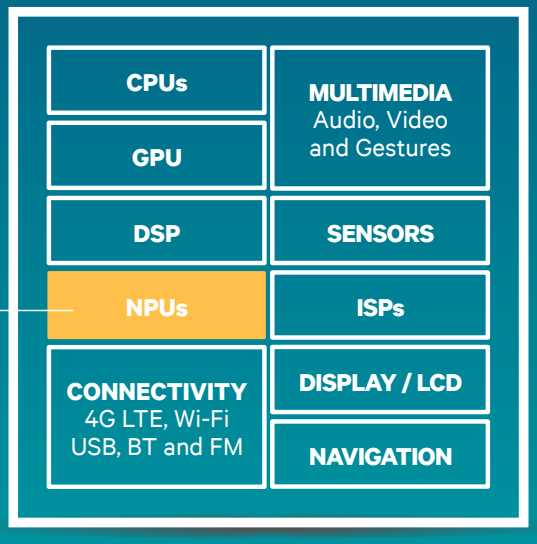
Continuous machine learning

Qualcomm Zeroth is a product of Qualcomm Technologies, Inc.

Neural Processing Units (NPUs)

A new class of processors mimicking human perception and cognition



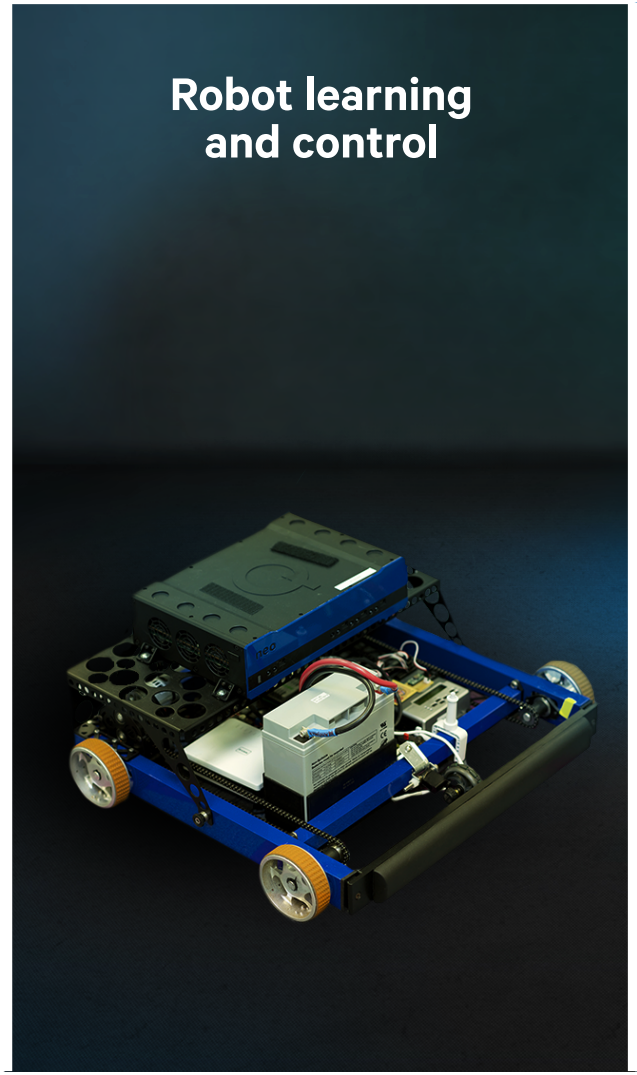


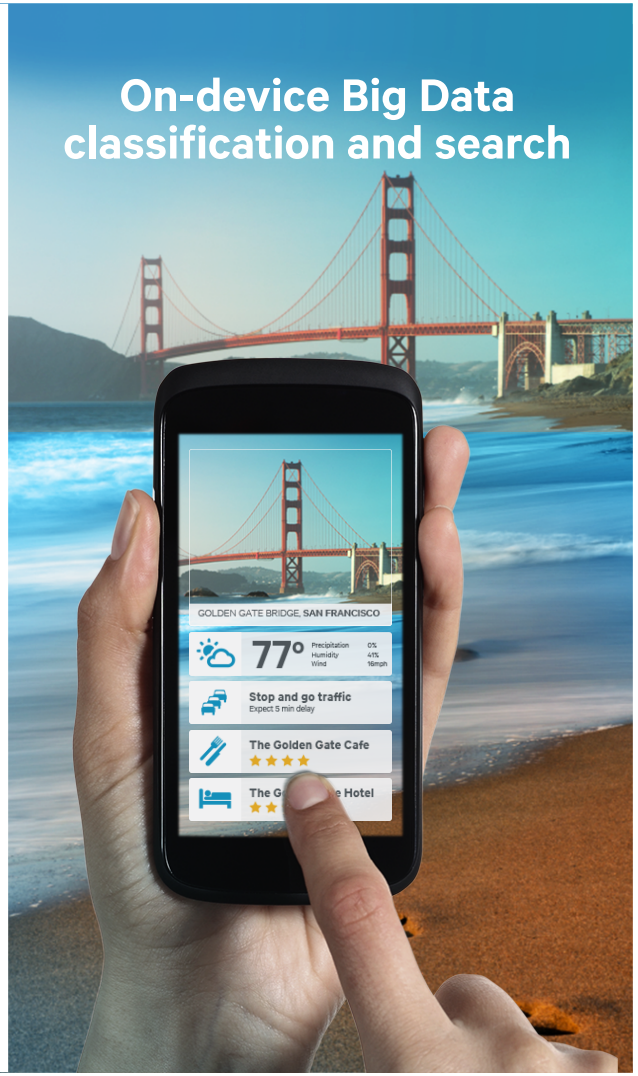
Massively parallel, reprogrammable

Comprehensive tools

Human-like functions


Natural interfaces and context awareness
00:01
Talk time


Robot learning and control


On-device Big Data classification and search
GOLDEN GATE BRIDGE, SAN FRANCISCO
77°
Stop and go traffic
Expect 5 min delay
The Golden Gate Cafe

Key Challenges – Building A Neural Processor

- Networks and Algorithms
- Hardware Representation
- Software Tools and Languages

Complete Solution Requires Consideration of the Entire Problem

Zeroth™ approach: start with Spiking Neurons

Developing complex neuron models that can be implemented in hardware

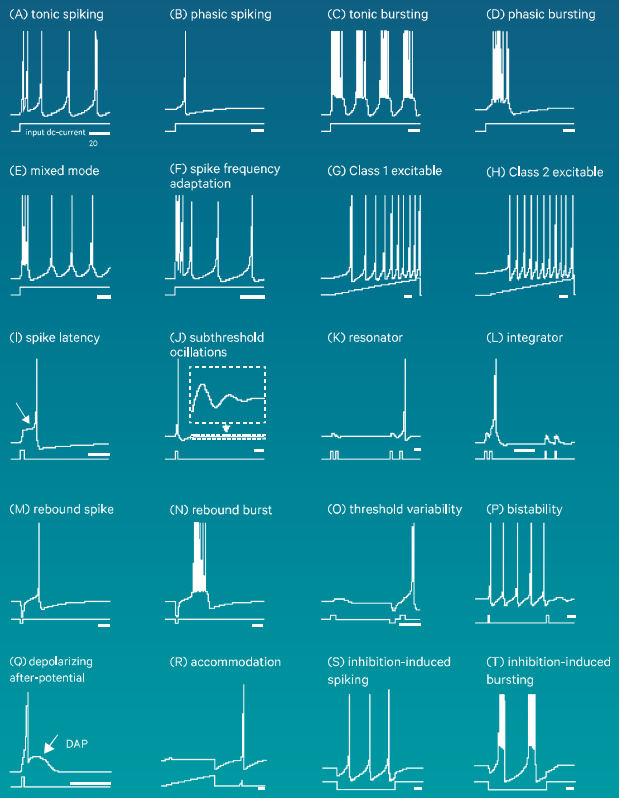
Real neuron

Electrical measurement

$$C\frac{dv}{dt} = F(v) - u + I$$

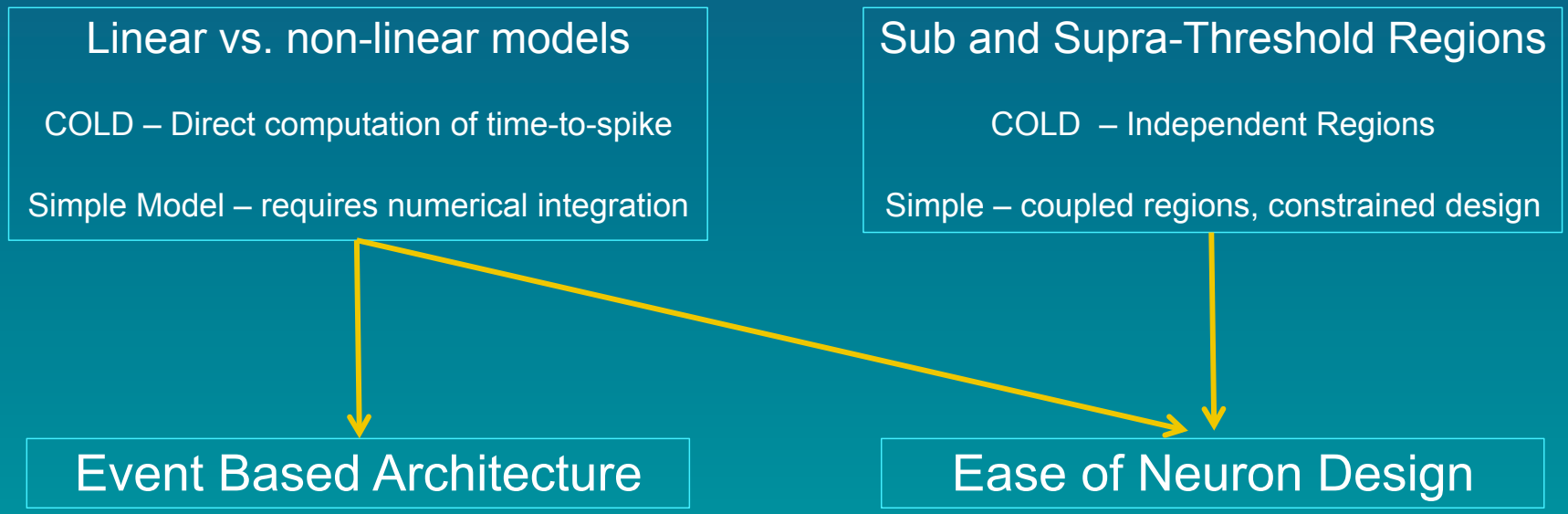
$$\frac{du}{dt} = a(b(v - v_r) - u)$$

Individual neuron modeling



Family of models

Designing a Better Neuron Model – COLD



[HunzingerSFN13] J. F. Hunzinger, Y. C. Yoon, T. Zheng, "A Spiking Neuron Model based on Two-dimensional, Two-piece, Linear Dynamical Systems", SfN 2013 Nanosymposium, Nov 11, 13.

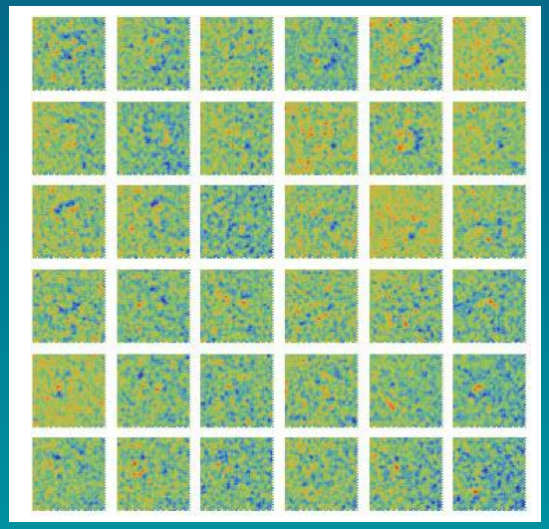
Color and Orientation Pop-out Networks

Results with COLD and Event Based Approaches

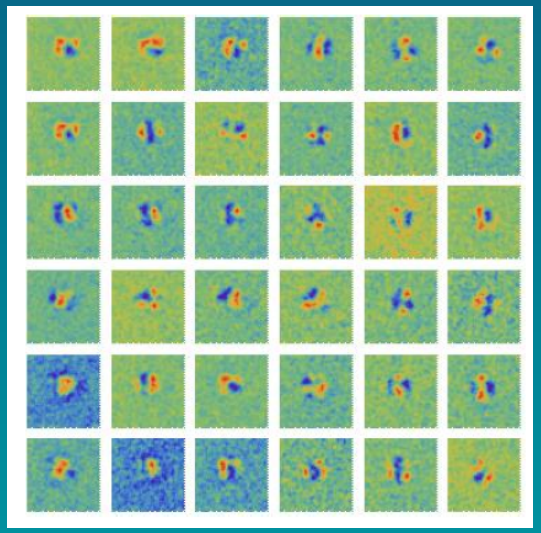
- Emergent networks for color pop-out (parvo) and orientation (magno) were constructed
- Using Qualcomm's COLD neuron formulation and an early event-based simulation approach
- Receptive field emergence for orientation was achieved in much shorter simulation times than other approaches

Orientation Receptive Field Formation

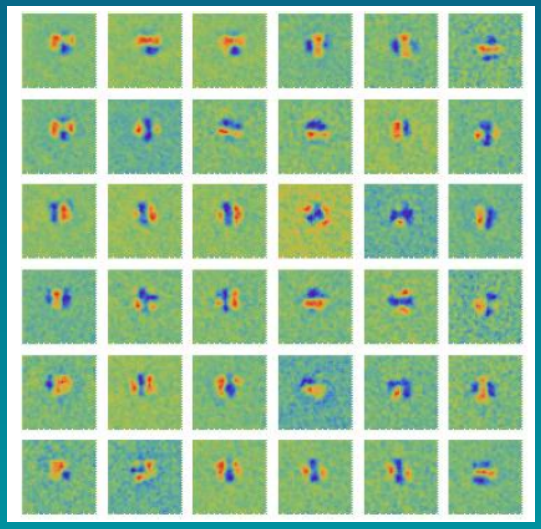
Rapid Specialization



Undifferentiated Fields



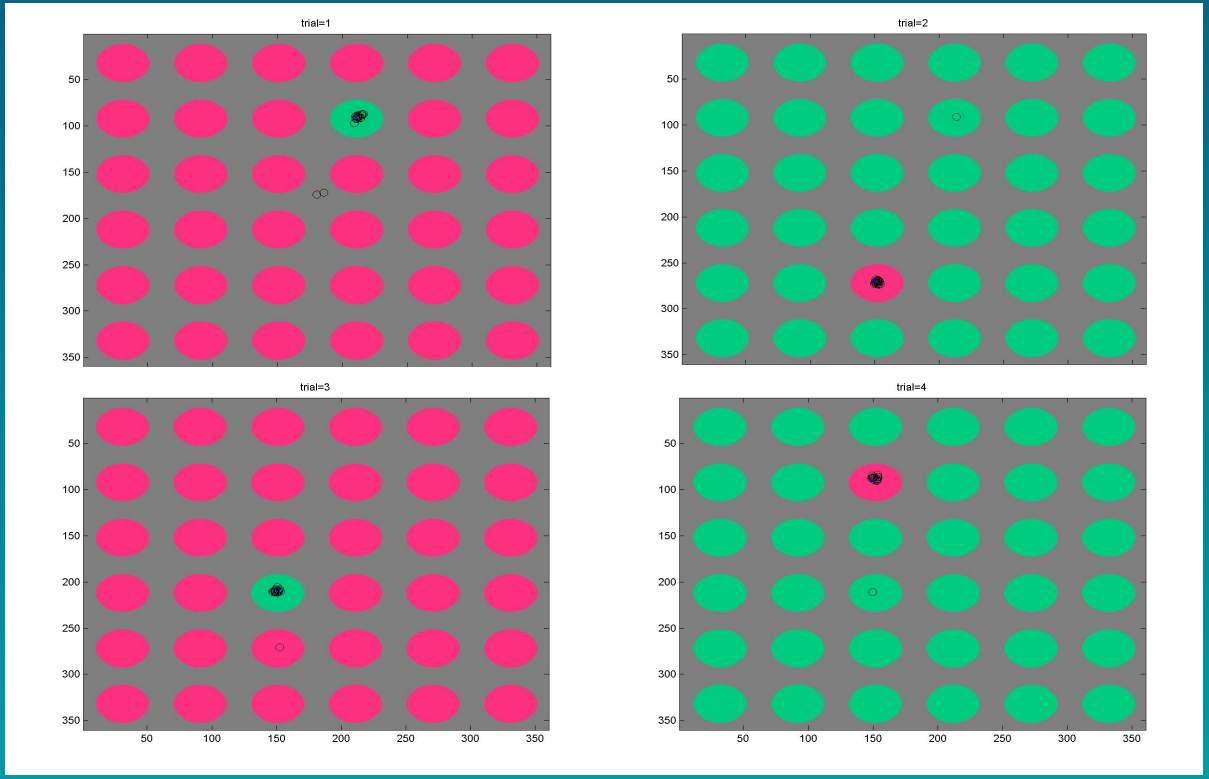
After 30 Seconds



After 100 Seconds

Color Pop-Out with Saccade Locations

The Red – Green Test

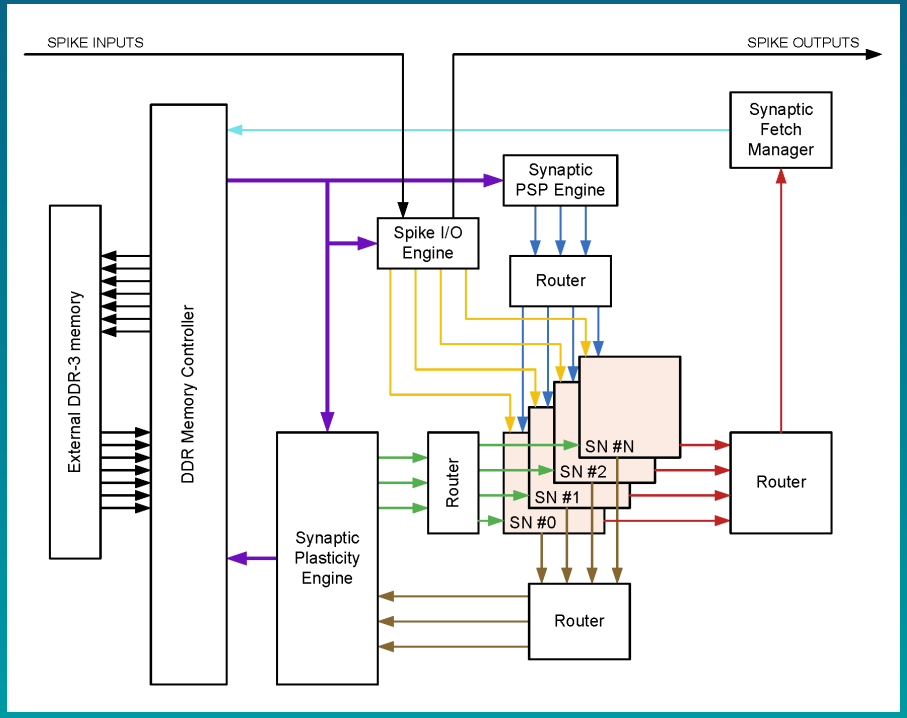


High Level Hardware Considerations

- Analog or Digital Architecture
- Fixed Versus Programmable Neuron Behavior
- High Density Versus Power Considerations

Hardware Overview

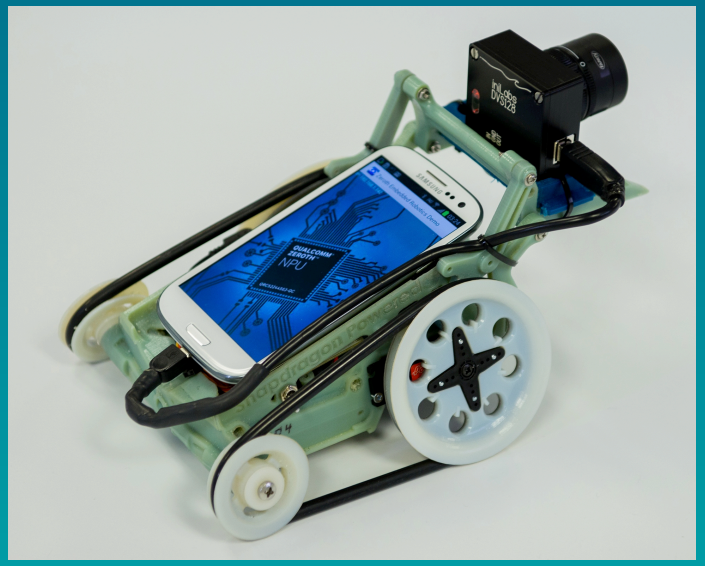
- **Neural Accelerator:** Custom tuned for efficient execution of spiking neural networks
- Programmable neuron and synapse dynamics
- Time multiplexed implementation across a scalable number of “Super Neurons”
- On board neuronal state, off-chip DRAM used for high density storage of synapatic connections



Key Future Hardware Challenges

Further Avenues of Investigation

- Future Memories
- Asynchronous logic and Event Based Processing
- The right level of programmability
- New Classes of Sensors



Software and Programming Considerations

The Brain is Parallel, Not Linear!

- Programs no longer linear
- Program Logic Encoded into the Network Itself
- Common Design Flow Across Output Targets
- Leveraged LLVM

Nature of the Programming Language

Basic Building Blocks

```
define unit exc_neuron from cold {
   init {
      a = 0.02;
      b = 0.1;
      d = 10;
   }
}

define unit inh_neuron from cold {
   init {
      a = 0.1;
      b = 0.4;
      d = 8;
   }
}
```

Define Units

Create Populations

```
// create excitatory (exc) population
create exc_neuron on Grid( countX=Nex, countY=Ney) as exc, all;
create random_input set prob=1.0/N on Grid( countX=Nex, countY=Ney ) as excInput;


// create inhibitory (inh) population
create inh_neuron on Grid( countX=Nix, countY=Niy ) as inh, all;
create random_input set prob=1.0/N on Grid( countX=Nix, countY=Niy ) as inhInput;
```

Connect

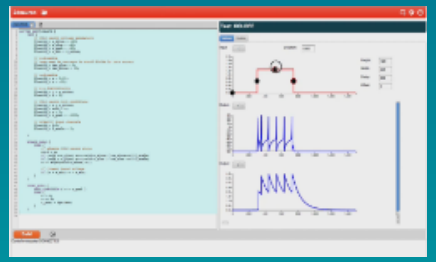
```
connect from excInput to 1 nearest exc with fixed_syn set weight=20, delay=1;

connect from inhInput to 1 nearest inh with fixed_syn set weight=20, delay=1;
```

Zeroth Neural Network Tools Platform

Zeroth Development Studio (ZDS): end to end package

- Programming Language: High Level Network Description language **(HLND)**
- Development Environment: Integrated Development Environment **(IDE)**
- Libraries- Neuromorphic Development Kit **(NDK)**
- Robotics Simulator- Virtualization World **(VW)**



HLND example IDE/NDK

Supports multiple neuron types

NST Compiler

automated engine specific format support

Support multiple execution targets:

- Software simulation: Linux workstation or equivalent *(today)*
- Hardware emulation: Custom Xilinx-K7 based FPGA board *(today)*
- ASIC Chip: Zeroth Neural Processing Unit (NPU) *(future)*

CPU (today)
GPU (future)

1-N Compute Cluster

Automatic distribution onto multiple cores → *scalable*

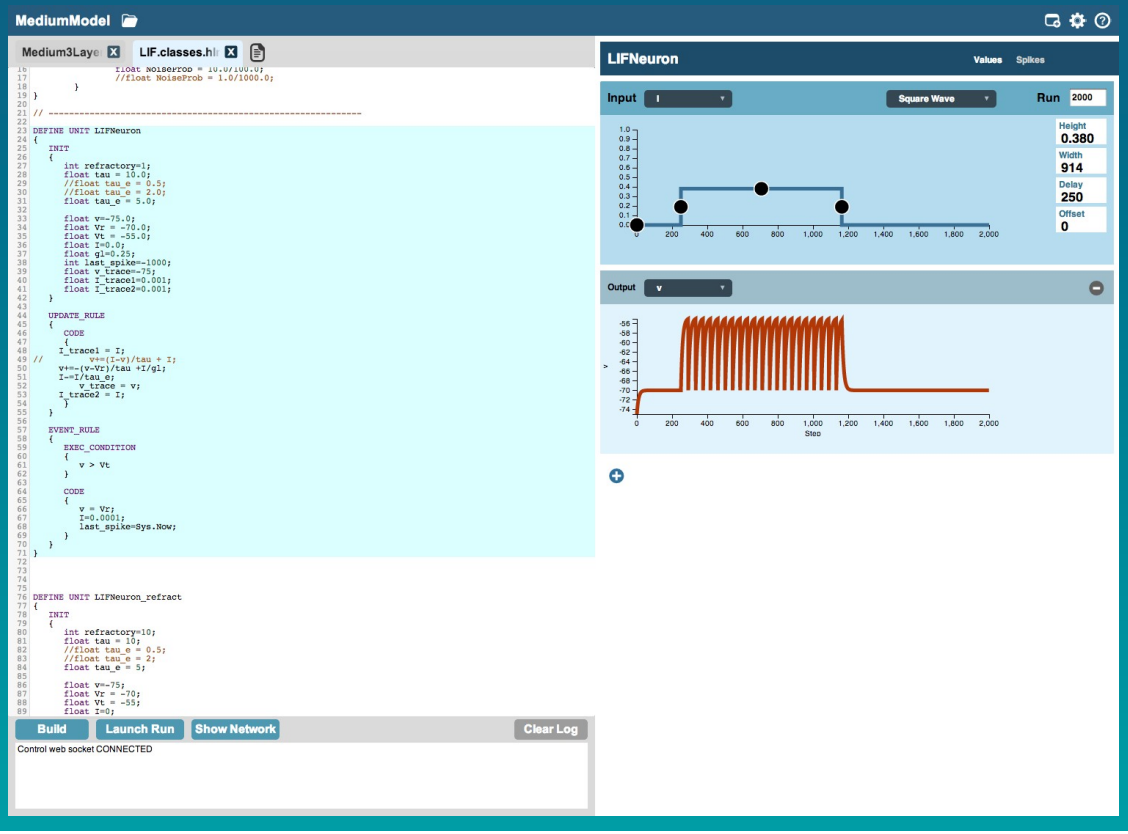
FPGA

NPU ASIC (Future)

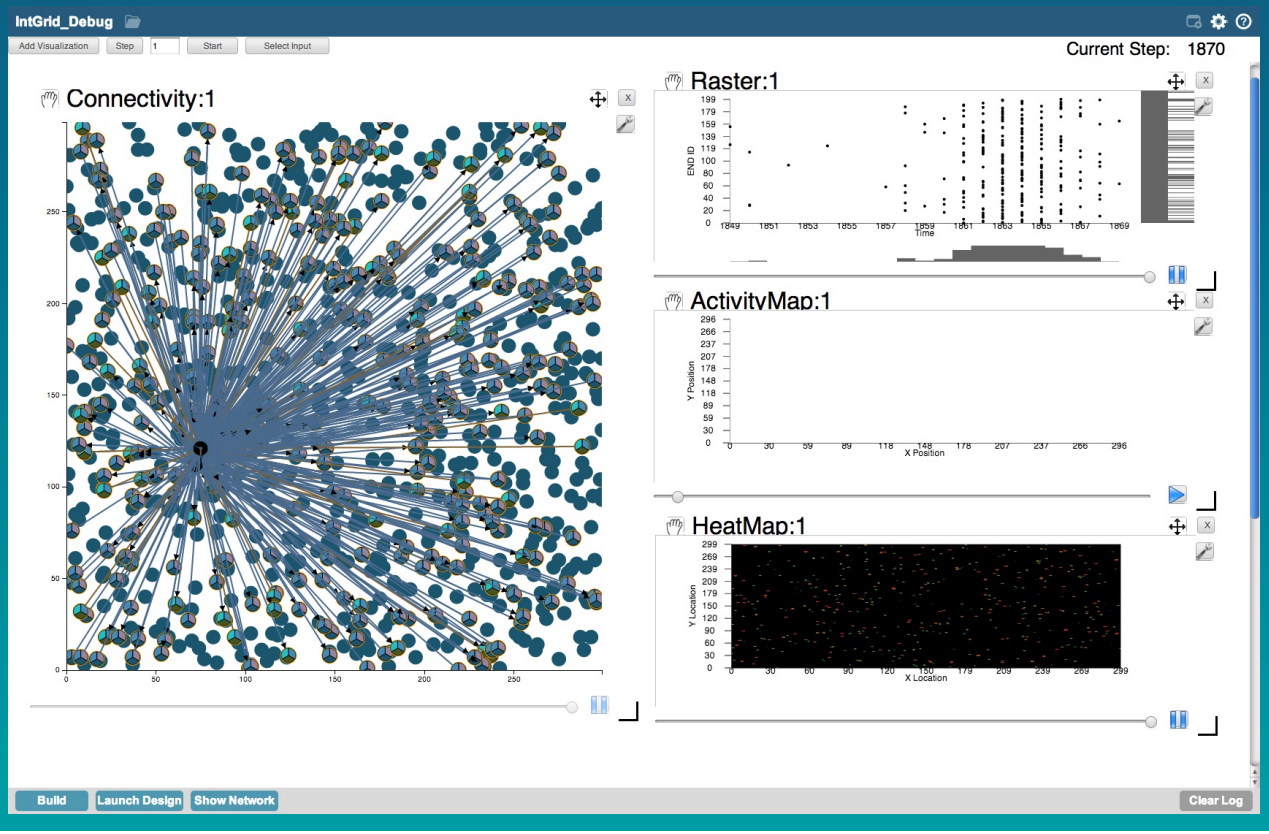
22

HTML 5 / Javascript IDE environment

Dynamic Code / Neuron parameter co-design

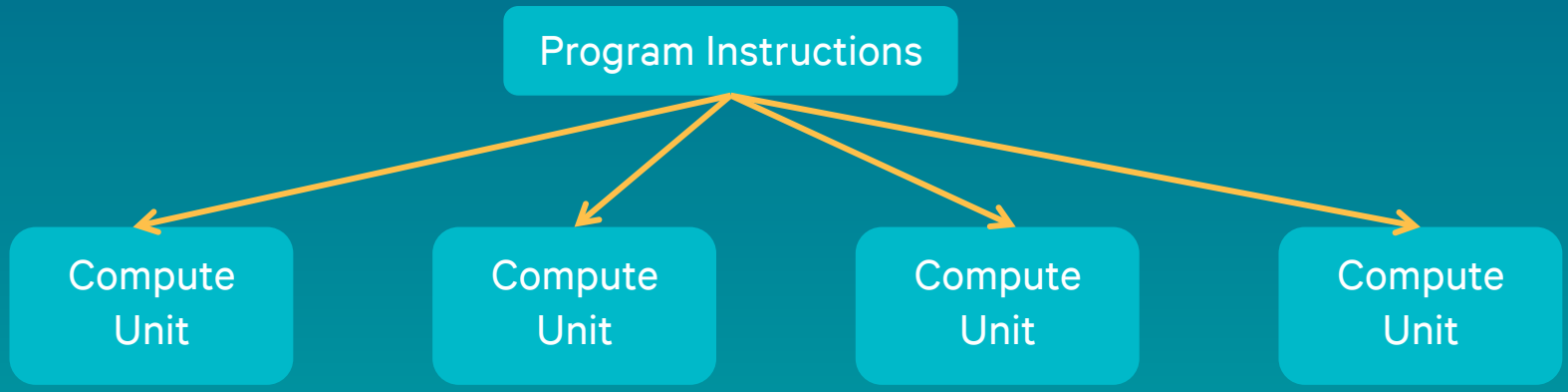


Real Time Data Visualizations



Hardware Design Creates Software Complexities

- Parallel and programmable nature of the hardware creates a complex “place and route” challenge



Neuron Types Synapse Types Communication Considerations

“Call to Action”

Research Opportunities Abound! – And We Can’t Do it Alone!

- Programming languages and abstractions to capture and represent these networks
- Better “routing and placement” algorithms for allocation to these massively parallel systems
- Breakthroughs in representation and training (e.g. DCNs take forever to train)
- HW architecture for asynch logic, continued advancements in memories
- Low power always-on sensors
- *Demonstrate “True Cognition”*

Thank you

Follow us on:

For more information on Qualcomm, visit us at:
www.qualcomm.com & www.qualcomm.com/blog