XILINX
ALL PROGRAMMABLE™

FPGA Acceleration:

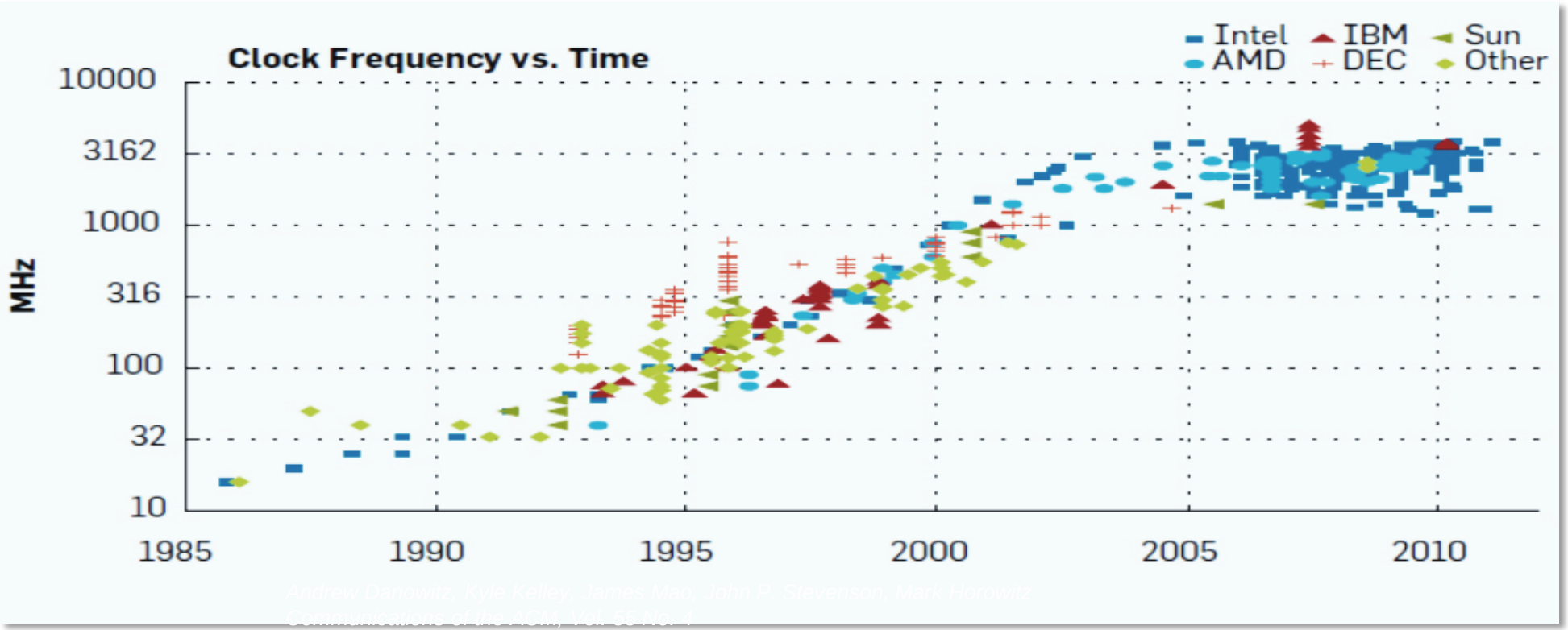
The Next Wave of Cloud and Edge Computing

Jie Tang System Architect

- Why FPGA matter
- New Workload
- Software & Hardware
- FPGA as a Service

After Moore

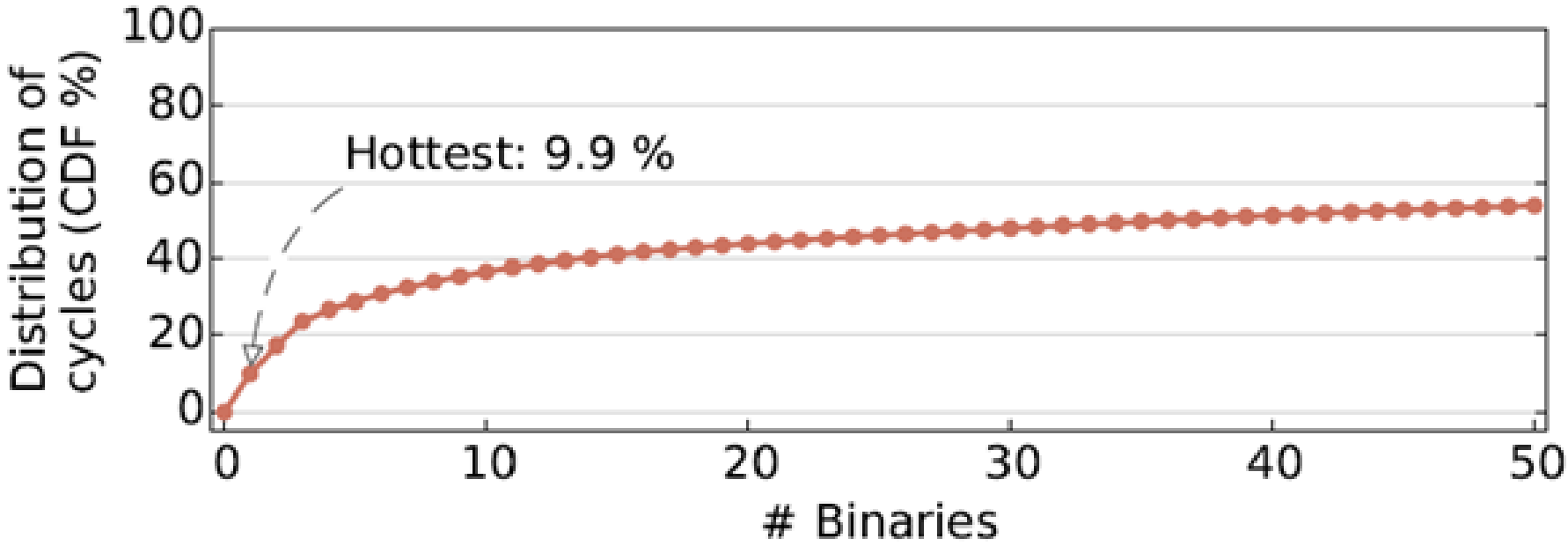
- CPU Architectures Not Scaling with Workloads



- Processor frequency scaling ended in 2007
- Multicore architecture scaling has flattened

Workload Diversify

- Datacenter Workloads are Increasingly Diverse



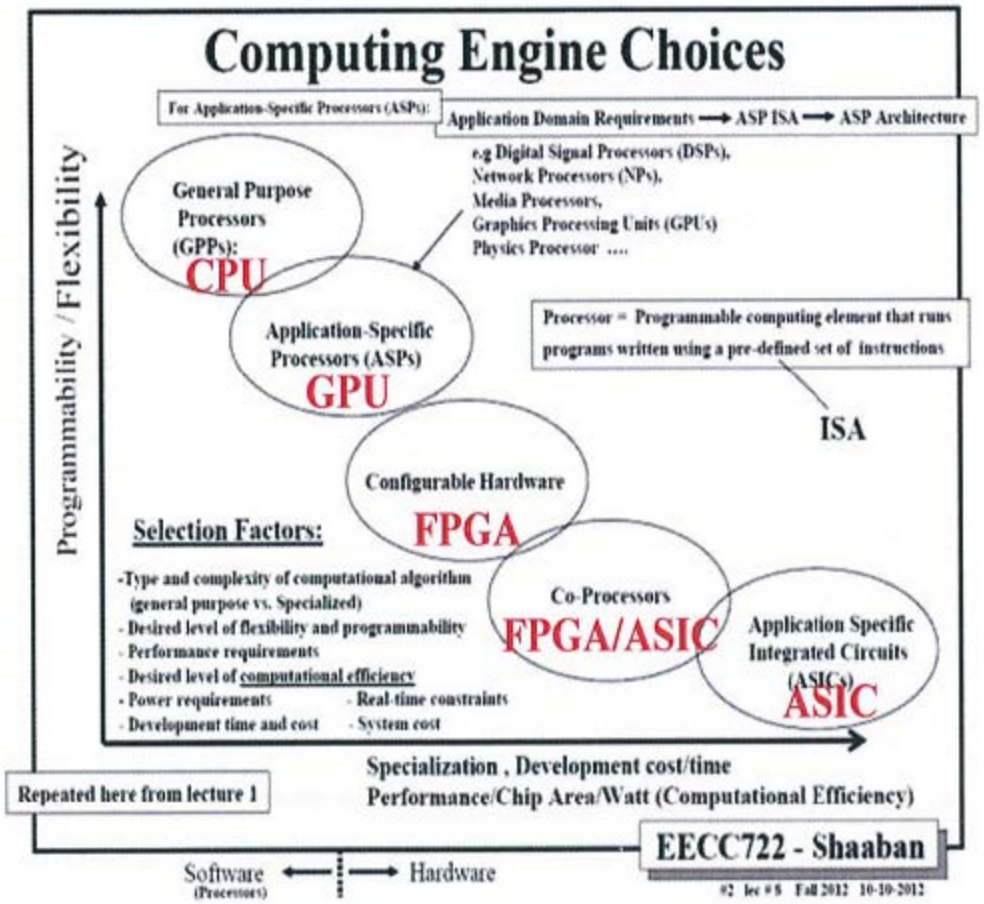
No “killer application” to optimize for.

Top 50 binaries only cover ~60% of data center workloads.

Reconfigurable accelerators are key to scalable compute architectures

Attribution:
Profiling a Warehouse-scale Computer
Svilen Kanev Harvard University et. al, ISCA, June 2015.

Silicon in Data Center



- **Make FPGA more easy**
  - All programmable for Software and Hardware from Silicon
  - High efficiency Interface for co-processor
  - Program framework for workload
  - FPGA as a service

Deep Learning

Algorithms & Models

CNN, R-CNN, RNN, LSTM, Feed Forward, BP, SGD, Deep Reinforcement learning, etc

AlexNet, VGG, GoogLeNet, ResNet, SqueezeNet, SyntaxNet, MobileNet, SSD, YOLO, etc

+

Computing Platforms

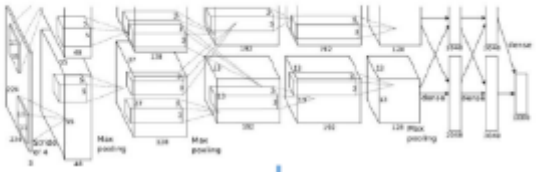
CPU, GPU, FPGA, TPU, etc

Deep learning software stack

+

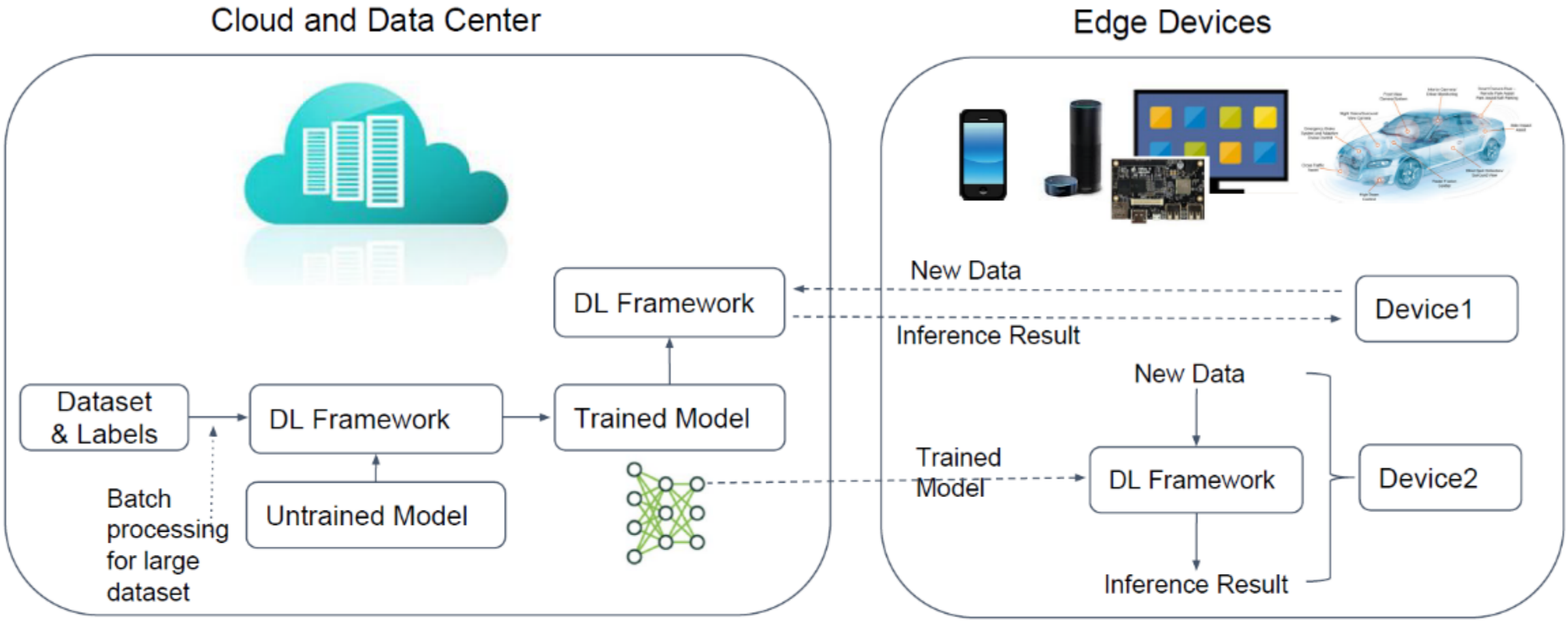
Datasets

MNIST, ImageNet, CIFAR10, CIFAR100, AudioSet, DET, CLS-LOC, ActivityNet, etc



Deep Learning

Impact new Architect



- **Training** on cloud & data center (desktop workstation can be used as well for development)
- **Inference** on cloud & data center or edge device

New Movement

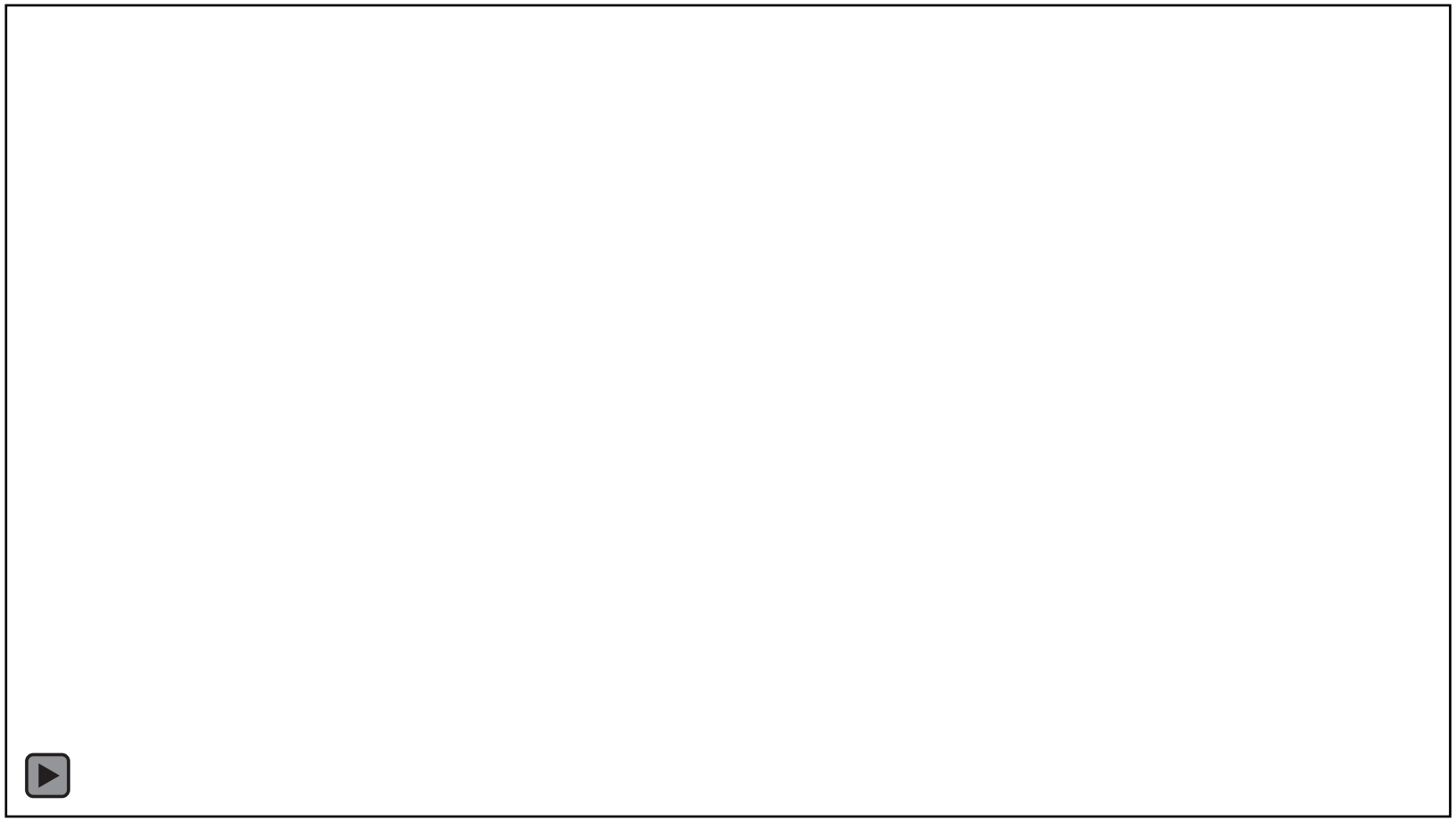
Cloud and Data Center

- Provide the AI service with Cloud aware
- Build on-premises Data Center for AI
- CPU+FPGA co-processor
- ASIC in Data Center

Edge Devices

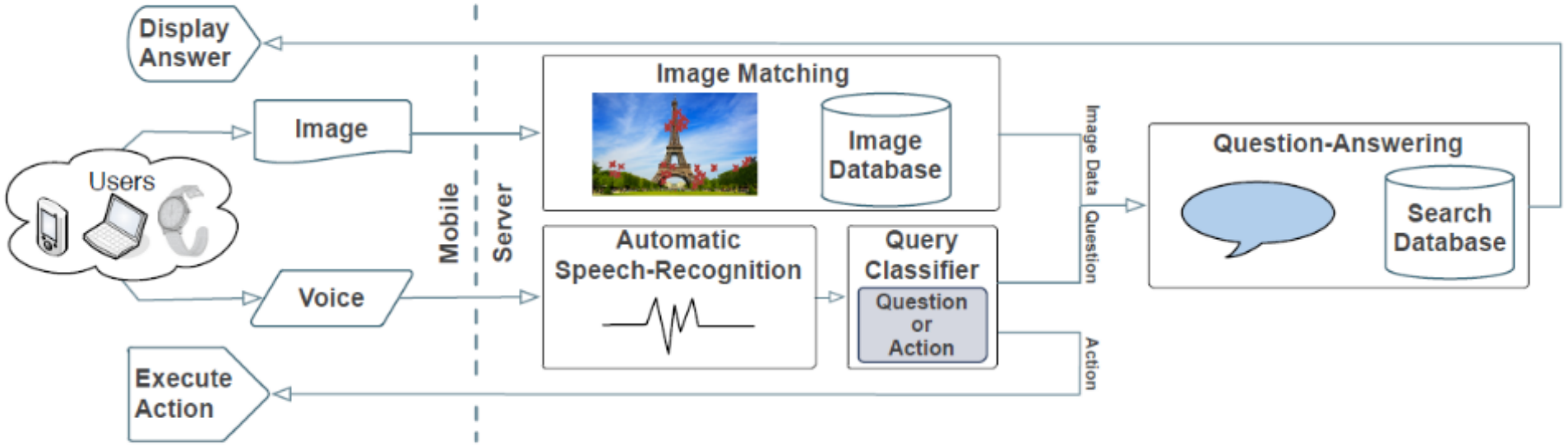
- ARM Computer Library
- Qualcomm Neural Processing Engine
- Apple AI SDK
- Huawei Meta 10 with Cambricon NPU
- Xilinx reVision

Lucida



http:// lucida.ai

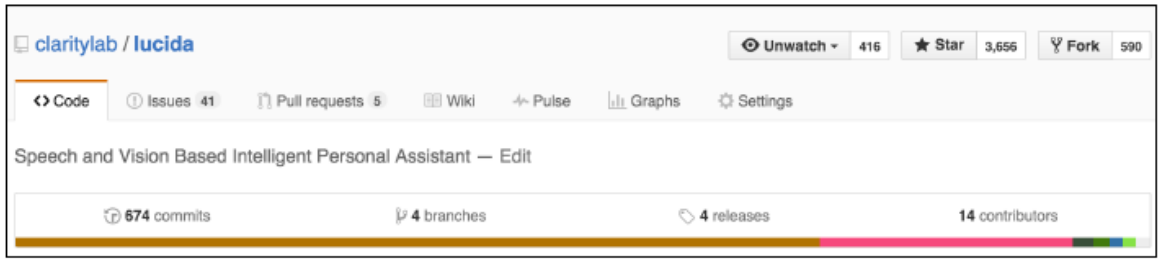
Work together

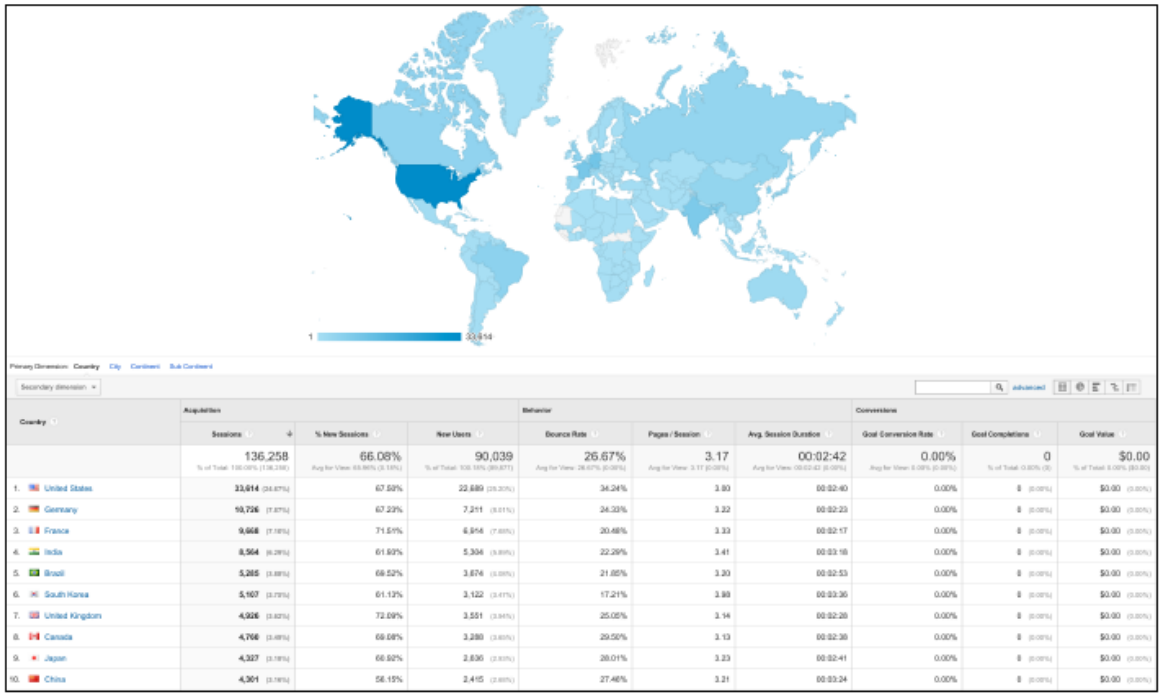


- AWS / Google /Microsoft start to deploy on **every** HOME.
- Mobile is new PC and voice is new keyboard
- New application is more easy to deploy and develop

Project status

- Recent Publication
  - BayMax [ASPLOS '16]
- Open source community
  - lucida-ai.slack.com
- Student projects
  - Independent studies
  - Summer interns





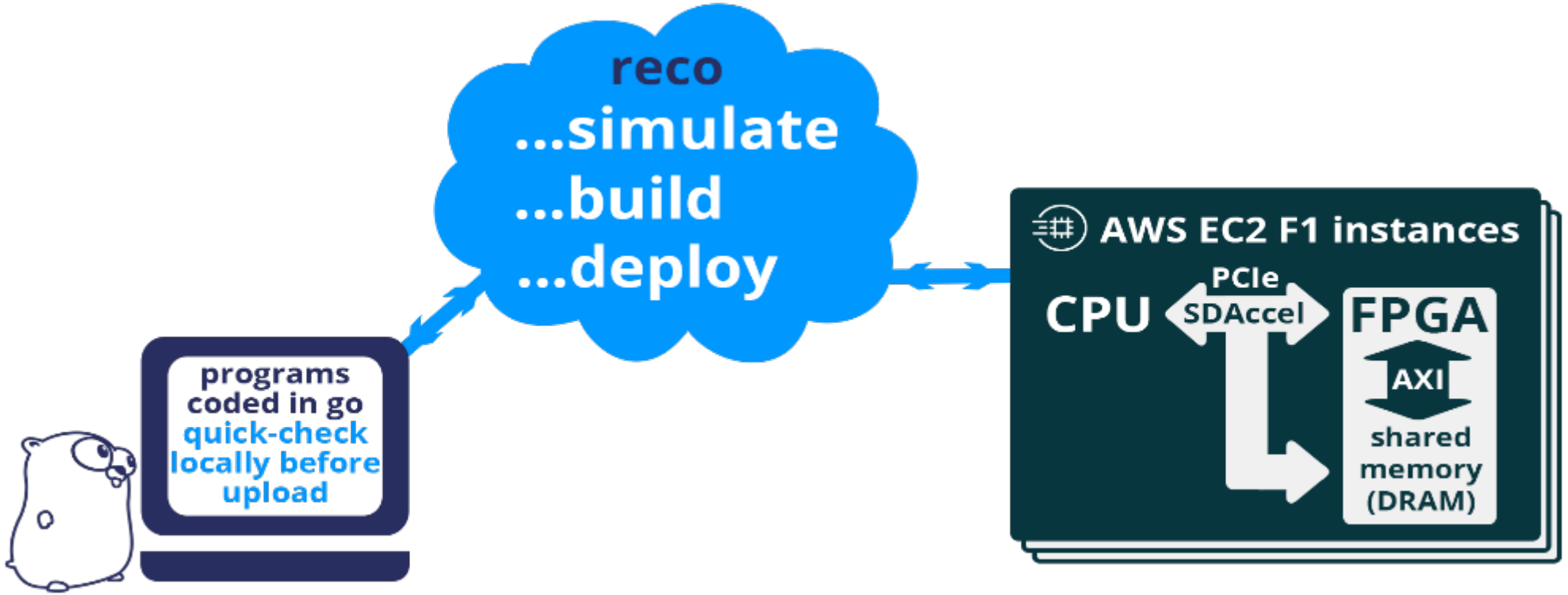
OpenCL enabled in FaaS

	Vendor	Languages	CUDA	OpenCL
Caffe	BVLC	C++, Python,Matlab	Yes	Yes
Caffe2	Facebook	C++, Pythin	Yes	No
TensorFlow	Google	Python,C++,Java,Go	Yes	Yes
MXNet	Apache	Python,R,C++,Perl	Yes	No
Torch	Community	Lua,Python, Matlab	Yes	Yes
Theano	Montreal	Python	Yes	Yes
Paddle	Baidu	Python Go	Yes	No
CNTK	Microsoft	Pythin,C++,C#, Net	Yes	No

Introduction to SDAccel Platforms

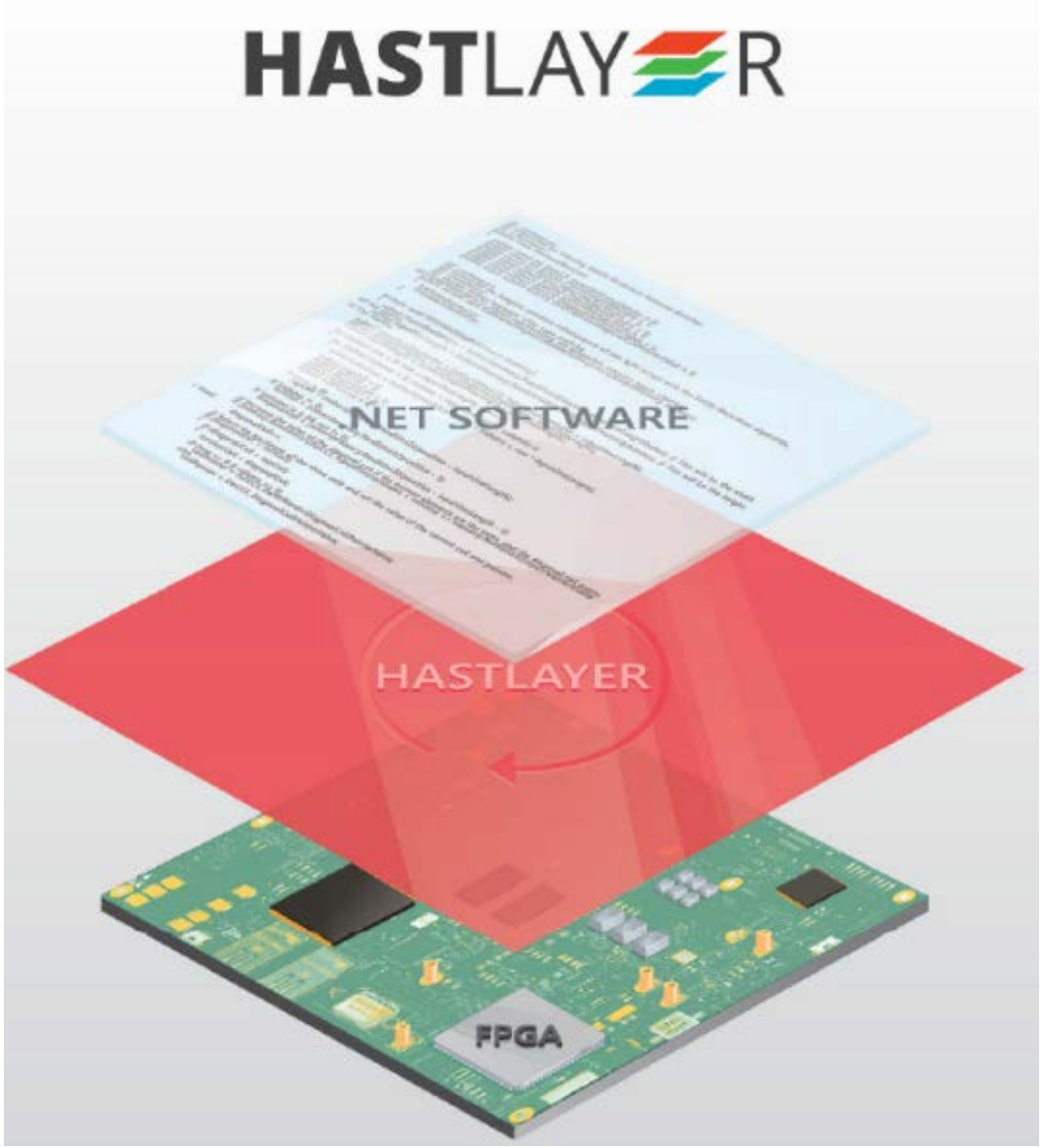
- Definition of Platform:
  - Infrastructure to enable communication to and from kernels to the outside world
  - IPs required:
    - PCIe
    - DMA
    - Memory Controllers
    - AXI Interconnects
    - OCL Region
    - AXI Performance Monitor
    - Flash programmer (optional)
    - Networking (10G/40G Ethernet) (optional)
  - Software required:
    - DMA driver
    - HAL API
    - SDAccel Runtime

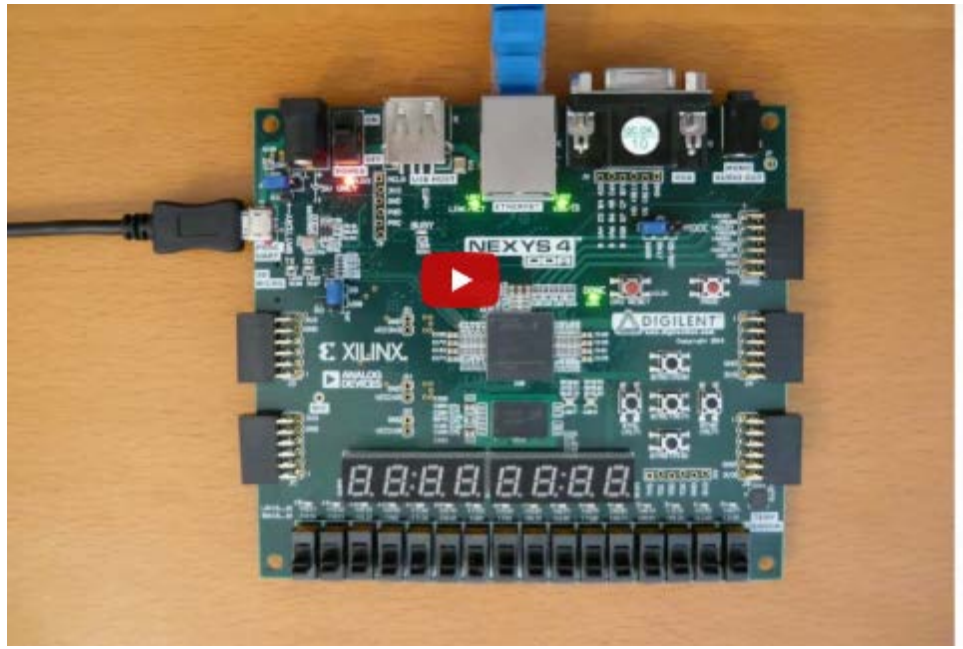
Project reconfigure.io



Reconfigure.io lets you program FPGAs with Go.

Project HastLayer





.Net -> VHDL -> FPGA Logic

More Application

F1 Use Cases and Partners

- Financial computing
- Genomics Sequencing
- Engineering simulations
- Image and video processing
- Big data and machine learning
- Security, Compression
- …and more

Your Application ?