QCon 全球软件开发大会
INTERNATIONAL SOFTWARE
DEVELOPMENT CONFERENCE

BEIJING 2017

Cloud for Cognitive Computing
(AI, Deep Learning …)

林咏华 (IBM研究院认知系统技术总监)

主办方 Geekbang> 极客邦科技 InfoQ

Who am I : Pioneer of Innovation

LIN Yonghua (林咏华)

linyh@cn.ibm.com

- 15 years in IBM Research
- Leader of System and Cloud Research direction in IBM Research China
- Global Leader of Cognitive System in IBM Research
- Founder of IBM Supervessel Innovation Cloud (超能云)
- Led the build, deployment and operation of Cognitive Services on IBM Bluemix in China
- ~ 50 Technical patents, ~ 10 papers
- Chair of IEEE Women in Engineering Beijing

IBM Cloud – Message from CEO in InterConnect 2017

- IBM Cloud is **Enterprise** Strong
- IBM Cloud is **Data** Frist
- IBM Cloud is **Cognitive** to the core

*"**1.4Trillion** dollars for **IT**, but **2 Trillion** dollars for business to **make better decision**."*

*"100M cusumers being touched by Waston by end of 2016, and **1B people** being touched by Waston by end of 2017"*

What is the Major Difference for Cognitive Computing on Cloud

- The System for Cognitive Computation – New type of hardware will be required in data center and cloud

Image Classification

Object Detection

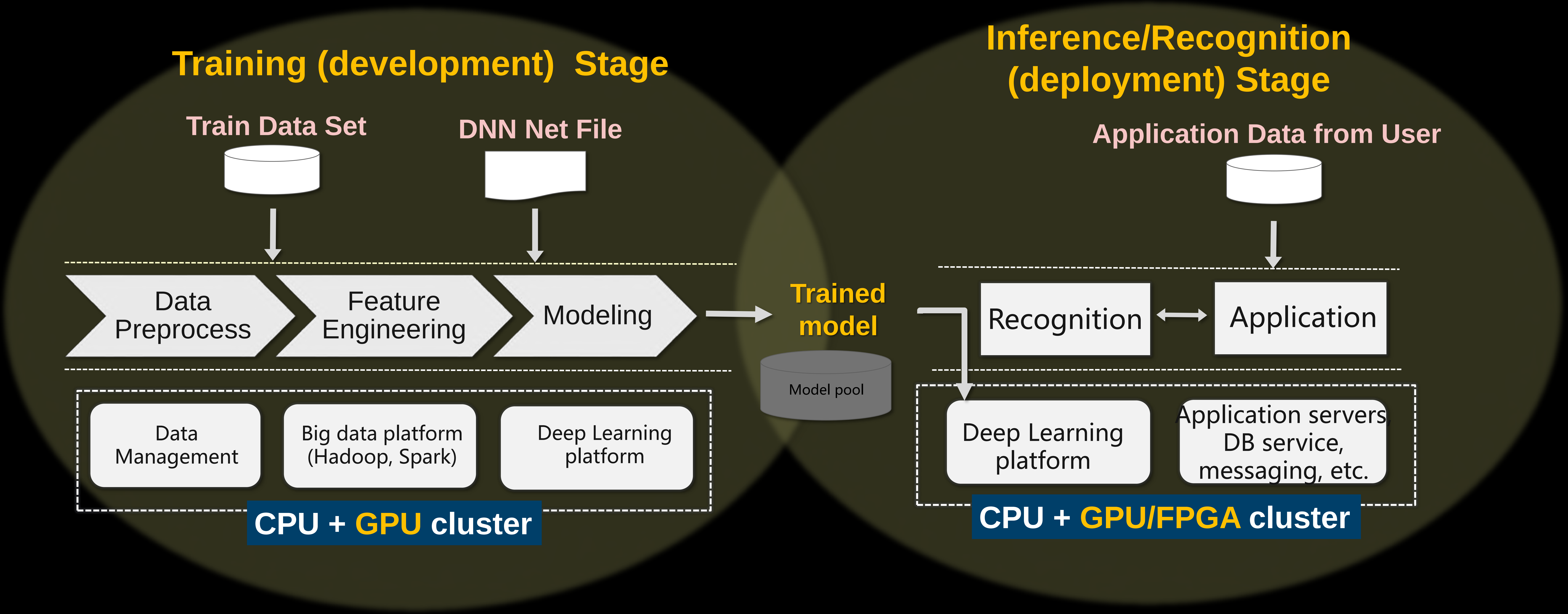
For image classification :

- CPU + FPGA vs. CPU : cost efficiency **2.5x ~ 8x**
- Included all the processing and the whole system cost)

For object detection :

- With VGG16: Processing latency on CPU 41.950s *VS.* latency 0.24s on GPU = **175times**
- With ZF: Processing latency on CPU 9.516s *VS.* latency 0.076s on GPU = **125times**

2 Stages Deep Learning for Cognitive Solution Build

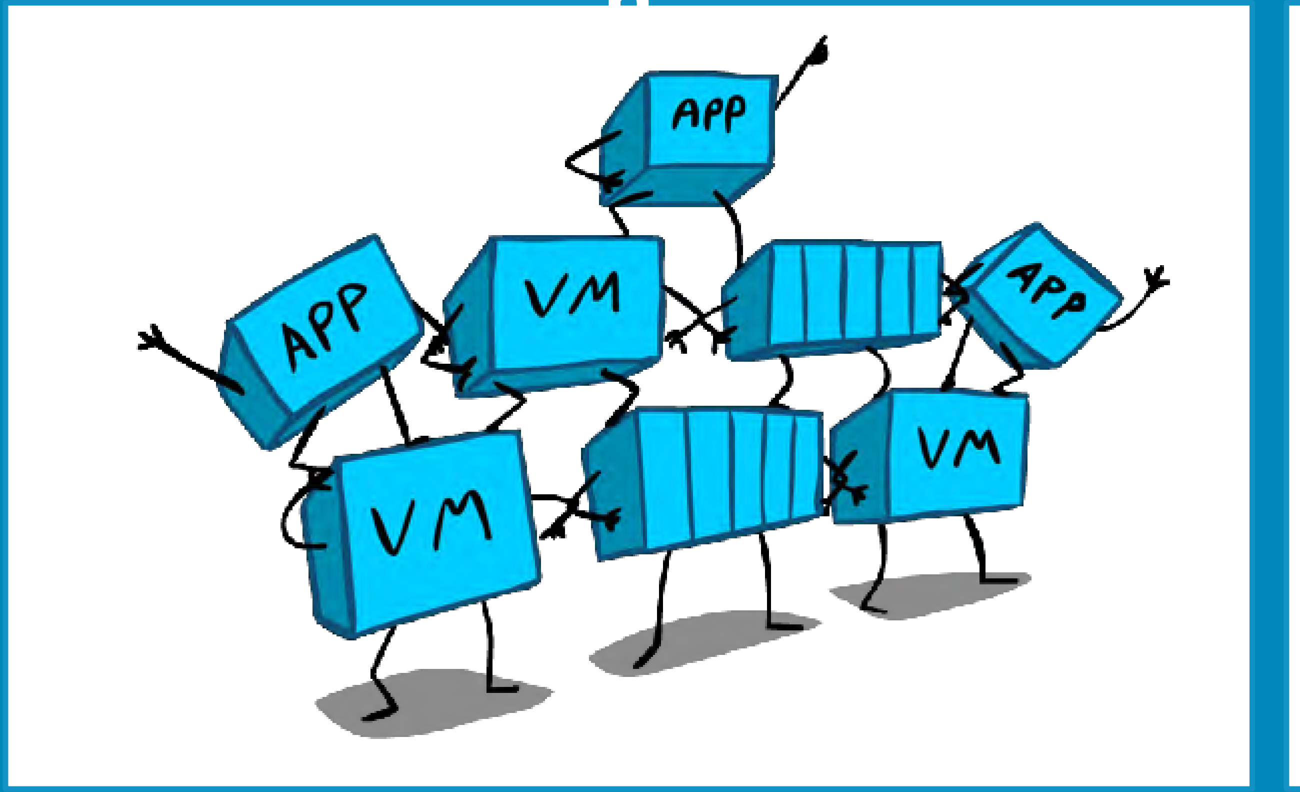


Time-to Insight

Time-to Market

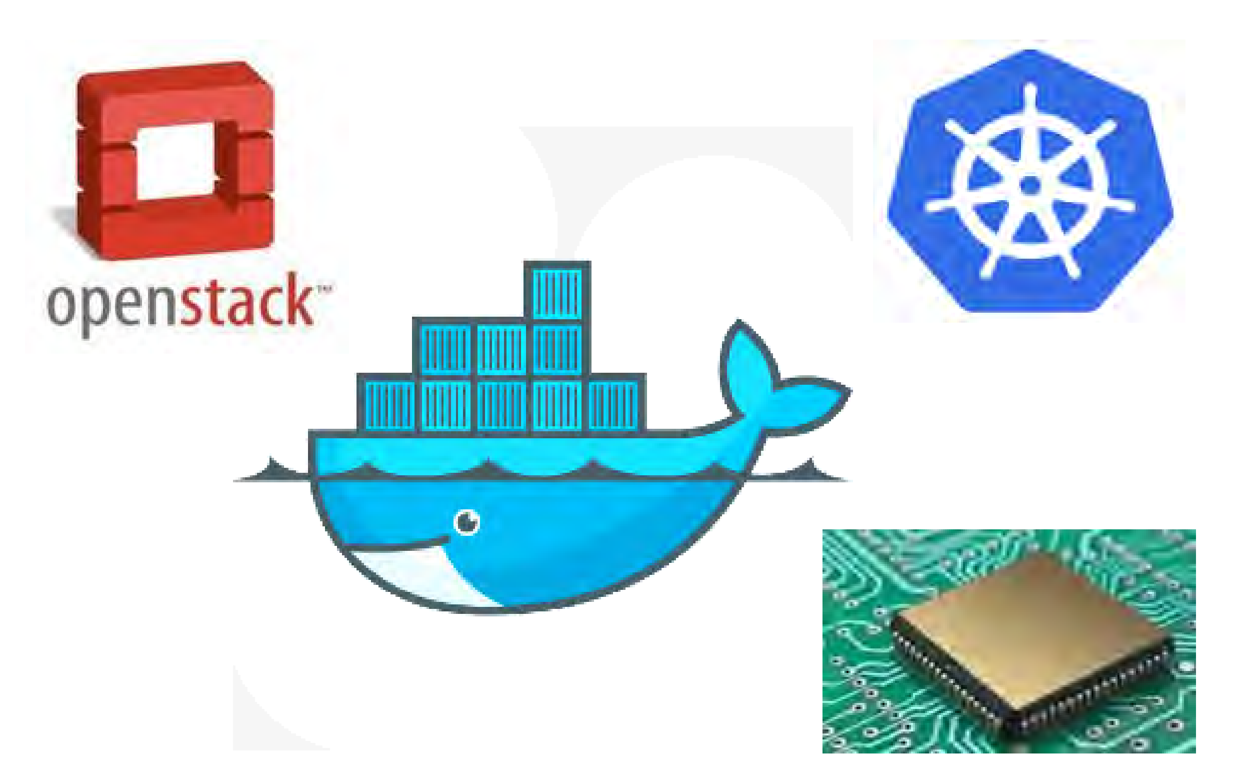
New Design Requirement with Accelerators in Cloud

Accelerator Sharing/Virtualizatio n



Cost Efficiency

Cloud Stack for Accelerator



Manageability

DevOps for Accelerator Project

Productivity

Deep Learning Platform : Coding for Accelerator

Caffe
TensorFlow
torch
theano
Spark
MLlib

Usability

Accelerators in Cloud (GPU, FPGA, etc.)

Why to Share Accelerator Resource?

For Image Object Detection using GPU to accelerate (15000 pics/hour ~ 50000 pics/hour) :

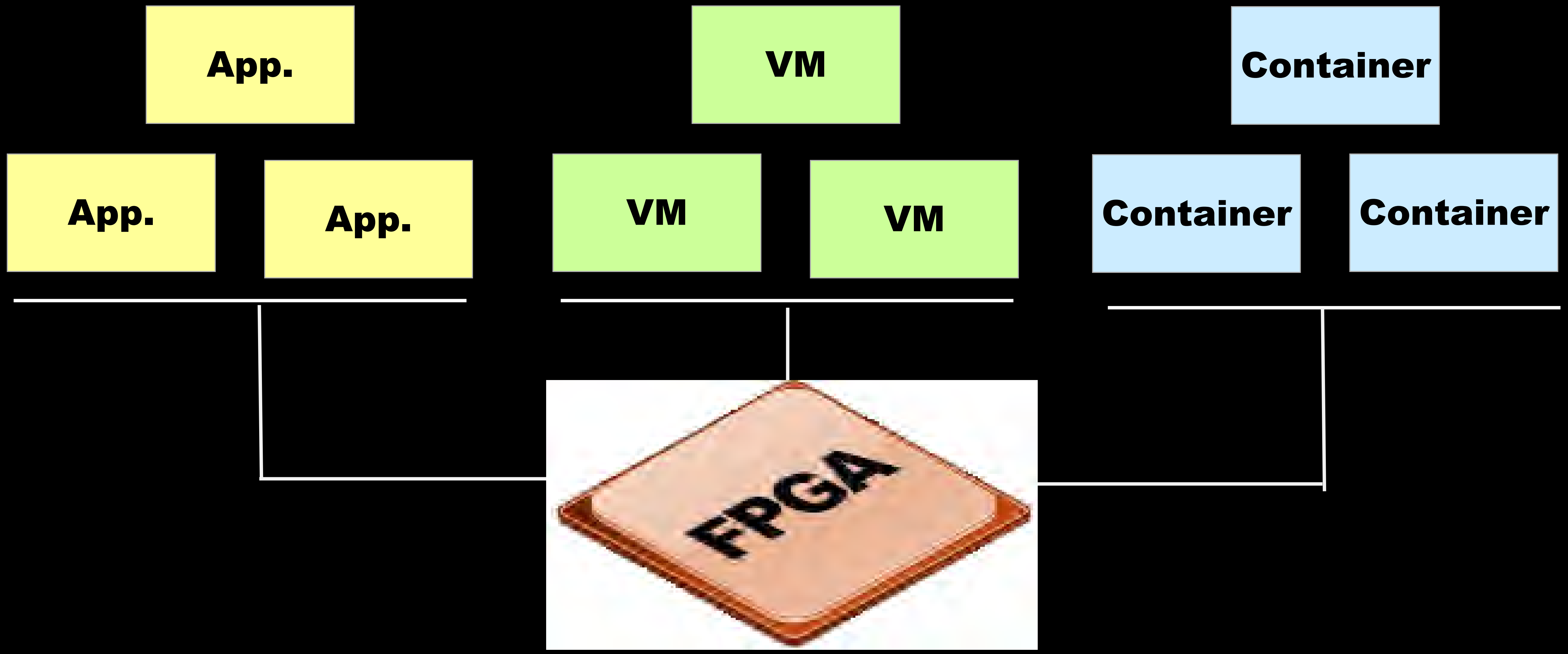
	GPU memory used	% of K80
Object Detection with VGG16	~ 1.7 GB	**~7%**
Object Detection with ZF	~ 1 GB	**~4%**

For Image Classification using FPGA to accelerate (300,000 pics/hour) :

	FPGA resource used	% of Xilinx KU115
Classification with AlexNet	DSP: 434, BRAM: 192	**8%** of DSP, **9%** of BRAM

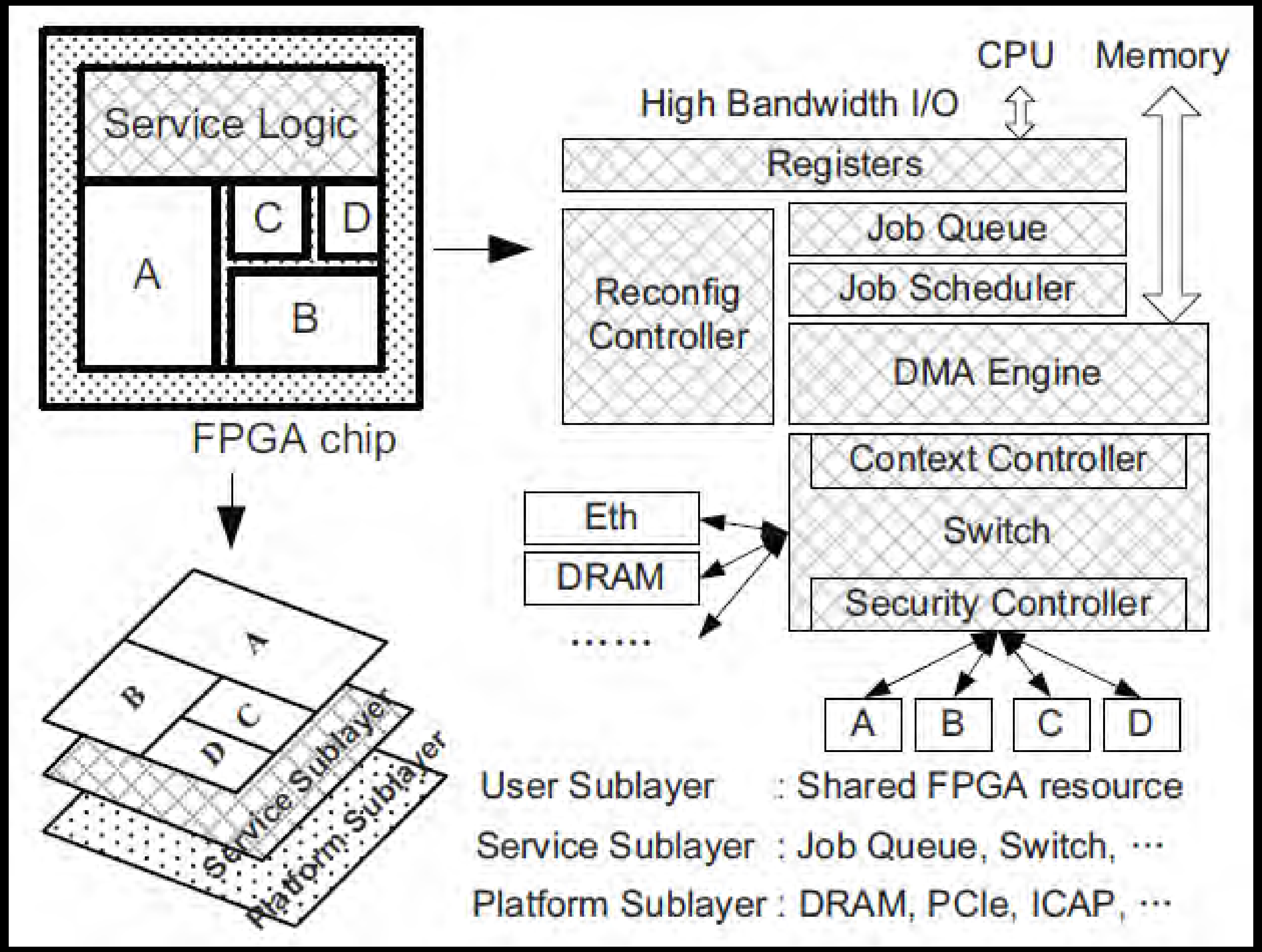
FPGA Virtualization for Multi-Tasks in Cloud

- FPGA resource could be shared by multiple applications, VM or container instances.



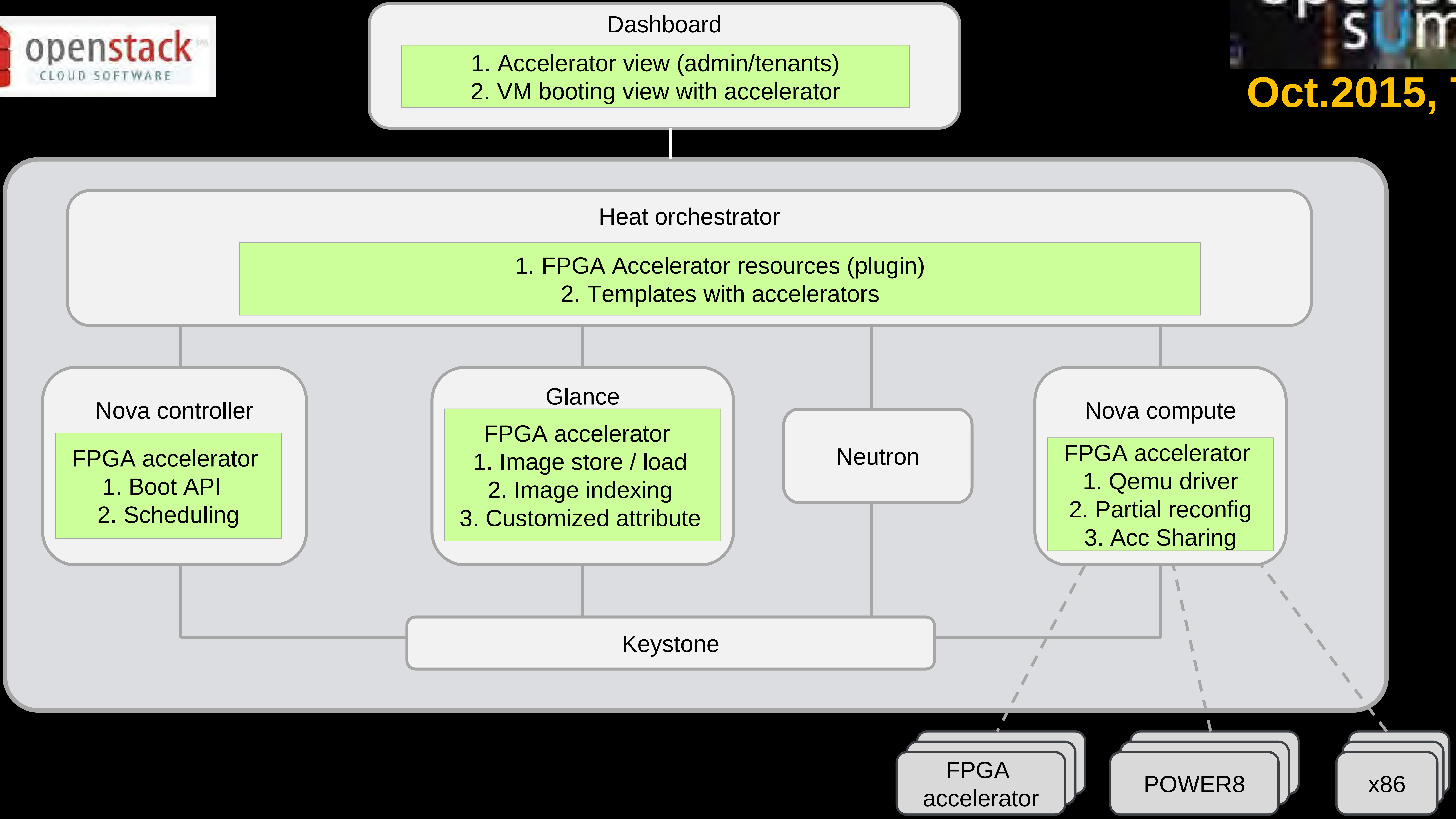
“Enabling FPGAs in Cloud”

ACM Computing Frontiers 2014, Italy



Extended OpenStack to support accelerator/FPGA as service

Oct.2015, TOKYO



Dashboard
1. Accelerator view (admin/tenants)
2. VM booting view with accelerator

Heat orchestrator
1. FPGA Accelerator resources (plugin)
2. Templates with accelerators

Nova controller
FPGA accelerator
1. Boot API
2. Scheduling

Glance
FPGA accelerator
1. Image store / load
2. Image indexing
3. Customized attribute

Neutron

Nova compute
FPGA accelerator
1. Qemu driver
2. Partial reconfig
3. Acc Sharing

Keystone

IBM launched the first FPGA service on cloud (in Apr.2015)

Accelerator developers :
Easily develop and deploy accelerator on cloud

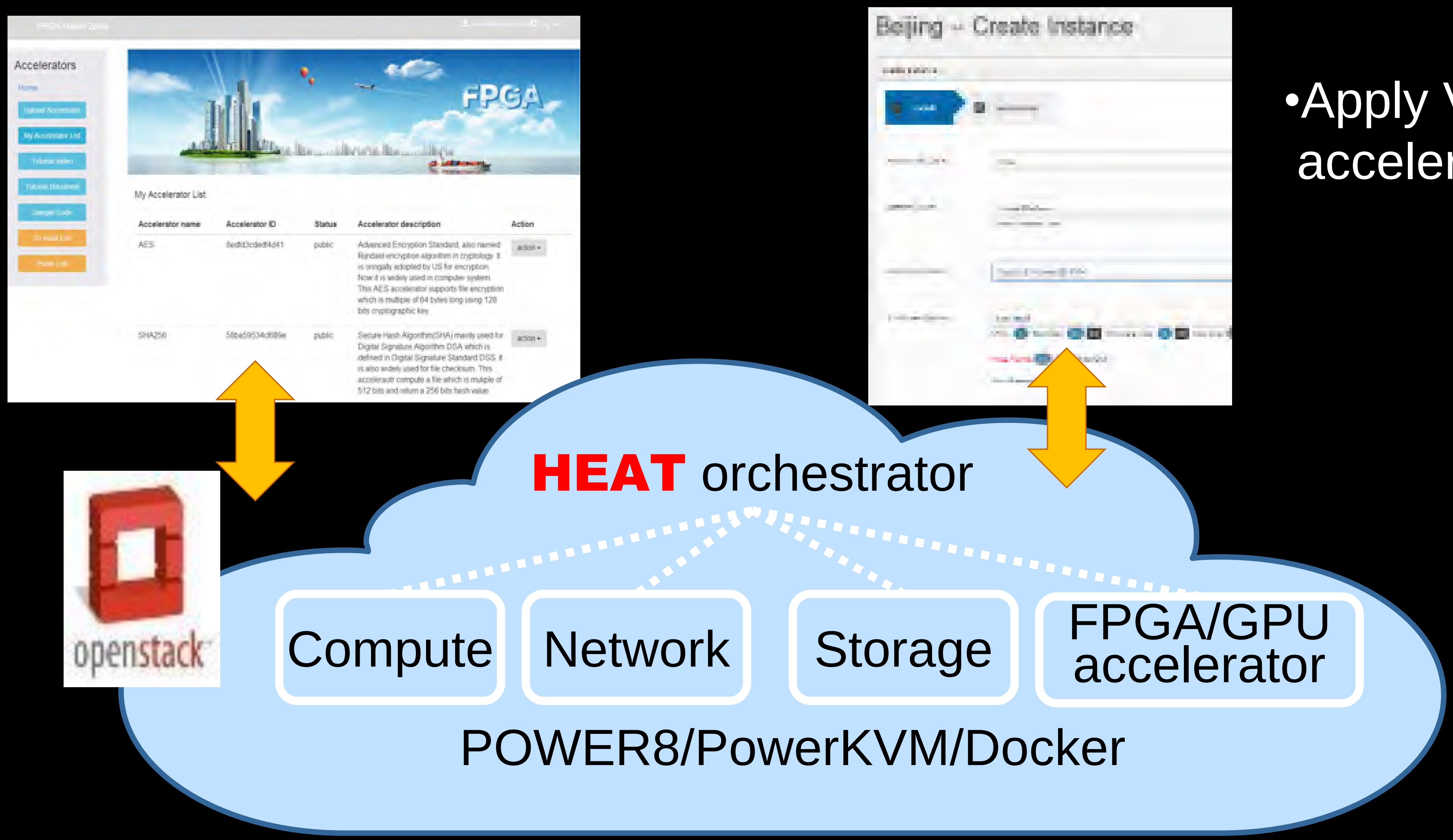
Application developers :
Easily use accelerator for application

Accelerator Maker Zone

Accelerator Service in Cloud

•Upload accelerator
•Cloudify

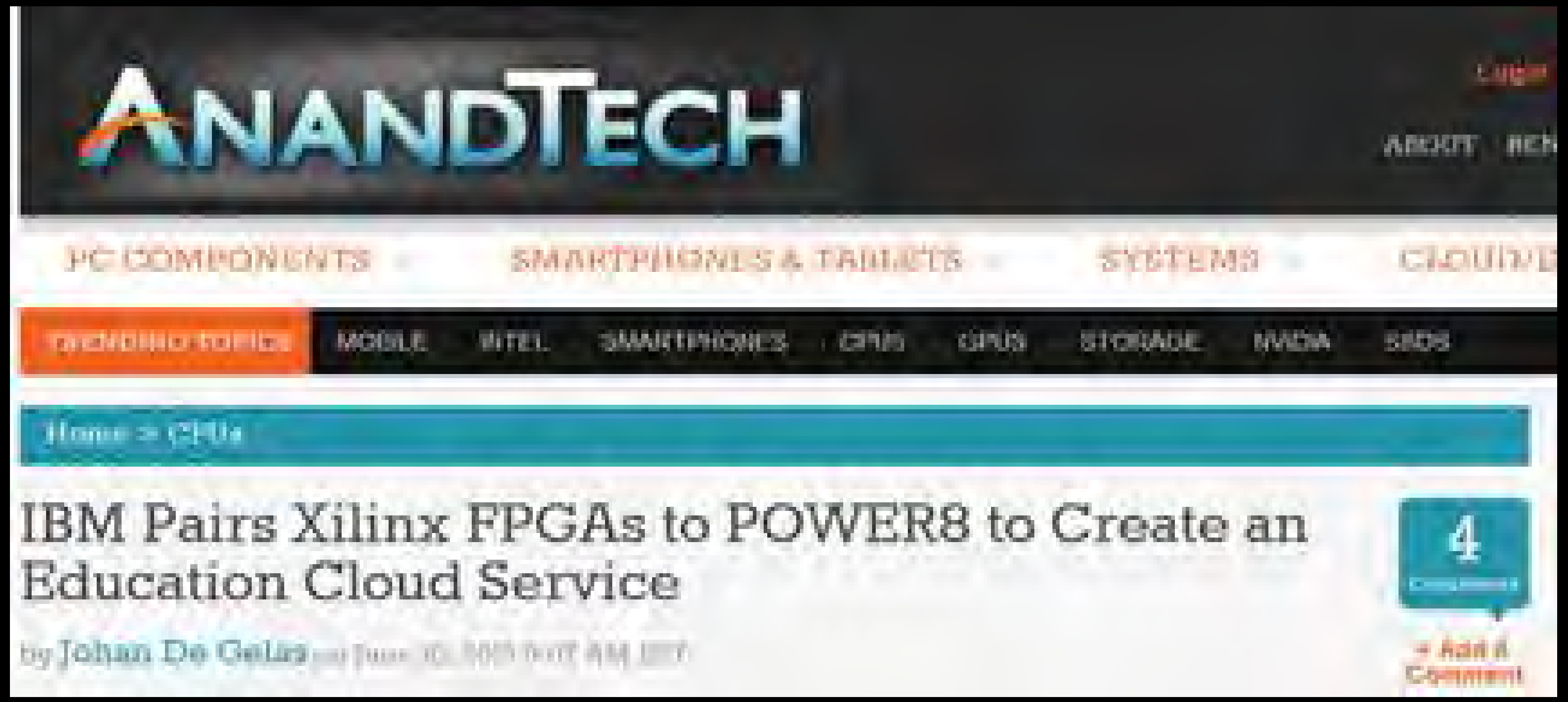
•Apply VM with accelerator



Acceleration Hardware

FPGA

- Supported FPGA developers: >200
- Supported accelerated application users : > 10000 (DB acceleration)
- Accelerated workloads
  - Deep learning
  - Genomics
  - Database
  - Data processing: compression, KVS, FFT
  - …



IBM Pairs Xilinx FPGAs to POWER8 to Create an Education Cloud Service

by Johan De Gelas

Enabled GPU in both OpenStack and Container Cloud

Support and Contribution for major open source cloud stack

- Enabled GPU sharing service on OpenStack Cloud
- Enabled GPU on Mesos, Marathon and Kuberntes – Contributing to communities
- Supported latest Kubernetest release (v1.6.1 now) with 3000+ lines of code extension

Enabled easy management for cloud:

- GPU resource discovery and management
- GPU topology aware scheduling
- GPU resource sharing

dockercon16
June 19-21, 2016 ▪ Seattle

kubernetes

Apache MESOS

IBM internal users

- Deep Learning as a Service (Deep Learning platform for Waston)
- VisionBrain (Deep Learning platform for Computer Vision)
- IBM Container Cloud for Bluemix
- Spectrum Conductor for Containers (IBM Container platform product)

GPU Technical Conference 2017 (May.8~11 @ Silicon Valley)

- 50 min. Talk: Speed Up Deep Learning Service -- When GPU meets Container Cloud

DevOps Service for FPGA Accelerator

- IBM and Xilinx launched the first Accelerator DevOps Service on cloud for FPGA developers in Apr.2016.

Online Accelerator project management

ibm-capi / pslse

Watch 18 · Star 11 · Fork 18

Code · Issues 0 · Pull requests 0 · Wiki · Pulse · Graphs

Power Service Layer Simulation Engine

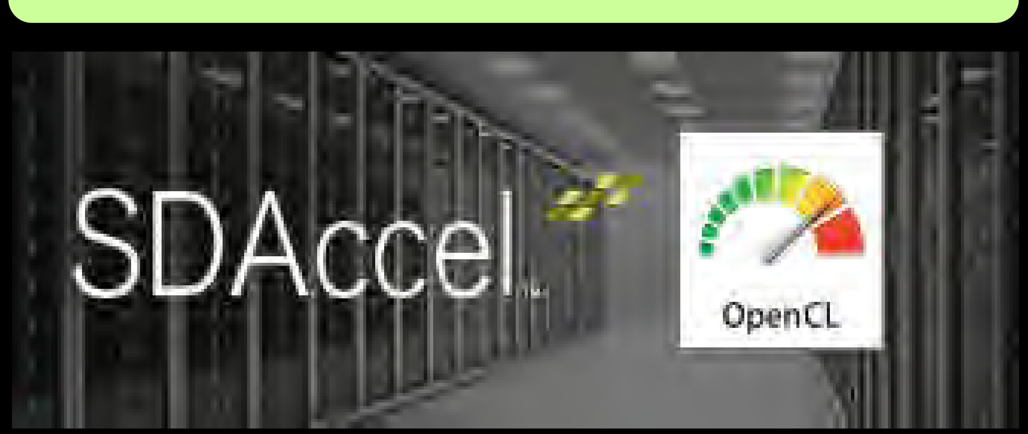
128 commits · 3 branches · 2 releases · 6 contributors

Branch: master · New pull request · New file · Find file · HTTPS · https://github.com/ibm-capi · Download ZIP

kirkmorrow Merge branch 'develop' · Latest commit on 30 Sep 2015

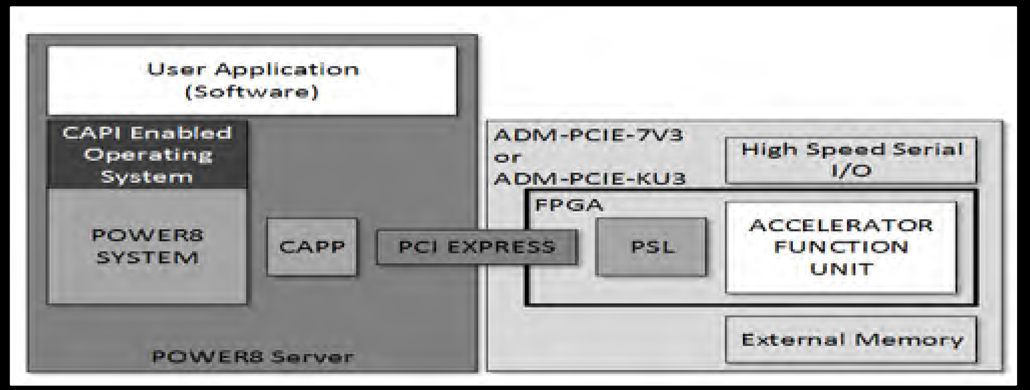
afu_driver	Cleaned up Makefile vars to only include debug and performance info f…	4 months ago
common	Added get_response() and set_machine_config_disable() functions	3 months ago
debug	Cleaned up Makefile vars to only include debug and performance info f…	4 months ago
libcxl	Fixed oversight in last commit	4 months ago
pslse	Fixed command debug message	3 months ago
sample_app	Whitespace cleanup	6 months ago
test	Added timeout support and handler for ctrl+C	3 months ago
.gitignore	Added Test API executable to ignore list	4 months ago
NOTICES	Update NOTICES	a year ago
QUICK_START	Documentation improvements	6 months ago
README	Documentation improvements	6 months ago
SIMULATOR_SETUP	Slightly fixed wording	5 months ago
vcs_include	VCS Support	5 months ago

Online development service with Cloud-based IDE

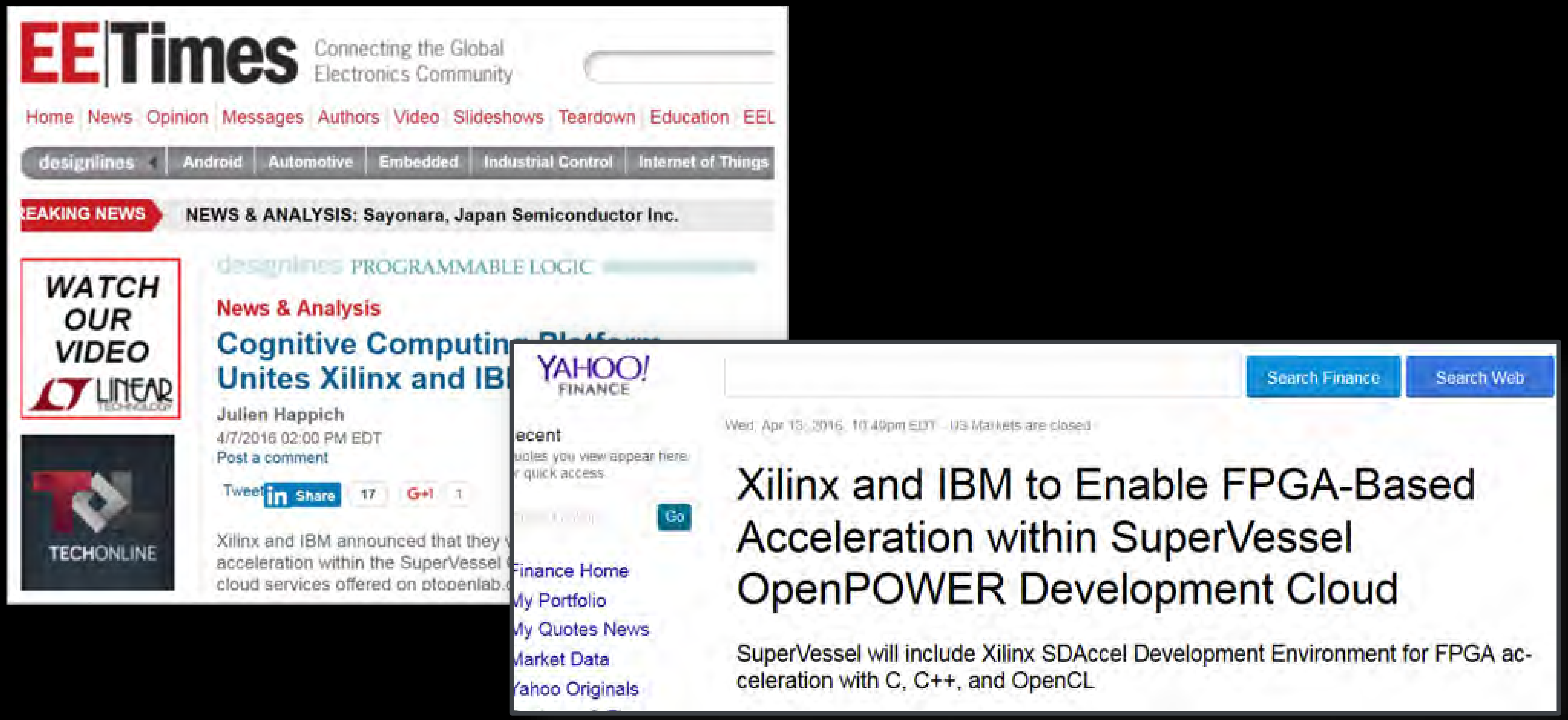


(Collaboration with Xilinx)

Test in VM/Docker equipped with FPGA (for POWER8 & CAPI)


User Application (Software)
CAPI Enabled Operating System
POWER8 SYSTEM
CAPP
PCI EXPRESS
POWER8 Server
ADM-PCIE-7V3 or ADM-PCIE-KU3
FPGA
PSL
High Speed Serial I/O
ACCELERATOR FUNCTION UNIT
External Memory

Publish to Accelerator App. Store and deployment for application on cloud



EE|Times Connecting the Global Electronics Community

Home | News | Opinion | Messages | Authors | Video | Slideshows | Teardown | Education | EEL

designlines | Android | Automotive | Embedded | Industrial Control | Internet of Things

BREAKING NEWS NEWS & ANALYSIS: Sayonara, Japan Semiconductor Inc.

WATCH OUR VIDEO LINEAR TECHNOLOGY

TECHONLINE

designlines PROGRAMMABLE LOGIC

News & Analysis

Cognitive Computing Platform Unites Xilinx and IBM

Julien Happich
4/7/2016 02:00 PM EDT
Post a comment

Tweet · in Share 17 · G+1 1

Xilinx and IBM announced that they … acceleration within the SuperVessel … cloud services offered on ptopenlab.…

YAHOO! FINANCE

Search Finance · Search Web

Recent

Quotes you view appear here for quick access.

Go

Finance Home
My Portfolio
My Quotes News
Market Data
Yahoo Originals

Wed, Apr 13, 2016, 10:49pm EDT - US Markets are closed

Xilinx and IBM to Enable FPGA-Based Acceleration within SuperVessel OpenPOWER Development Cloud

SuperVessel will include Xilinx SDAccel Development Environment for FPGA acceleration with C, C++, and OpenCL

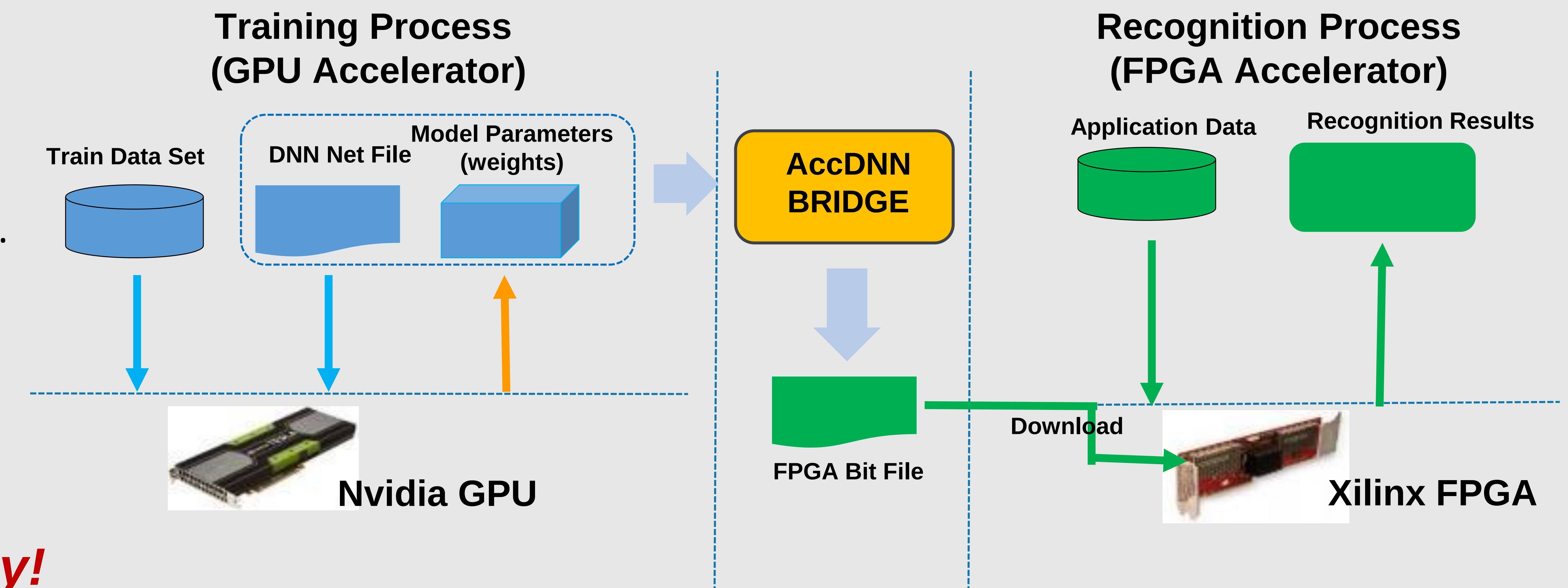
Allow FPGA Developers easily develop and build a new accelerator on Cloud

AccDNN : Tool to auto-generate accelerator for DNN

- **To solve programmability problem in deep learning**
  - A tool to generate DNN accelerator without FPGA programming and keep RTL level performance

DNN Application Design and Deployment Steps with AccDNN:

1. Design the specific deep neural network.
2. Training the network using GPU accelerator.
3. Use AccDNN to generate FPGA implementation
4. Deploy the recognition application using FPGA accelerator



Do it Automatically!

Illustration of training and recognition under *Caffe* framework

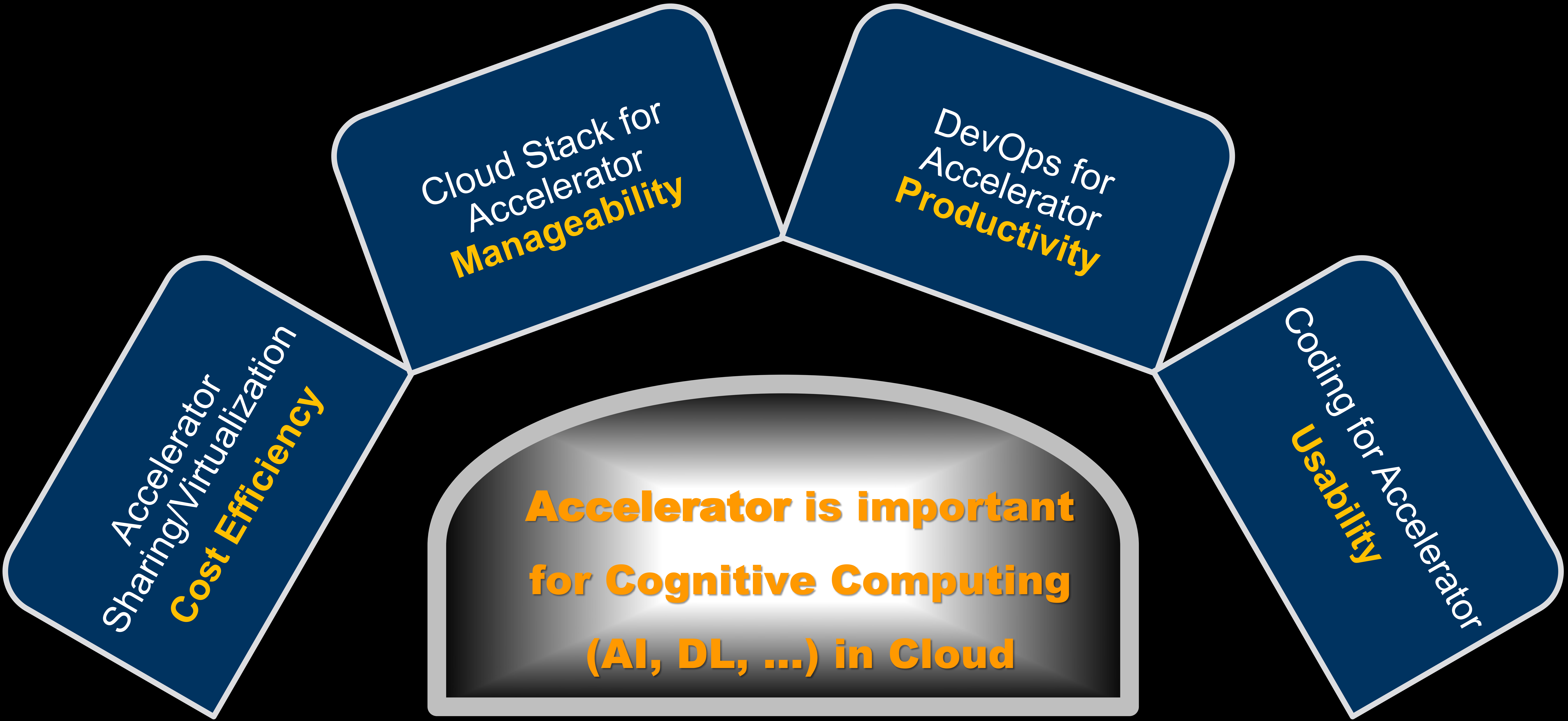
- AccDNN 0.1 was launched as cloud service to support OpenPOWER Global Challenge 2016

AccDNN: Accelerate Deep Neural Ne

in FPGA without Programming

Help you to convert your own trained DNN mode

FPGA RTL implementation automatically.



Accelerator Sharing/Virtualization
Cost Efficiency

Cloud Stack for Accelerator
Manageability

DevOps for Accelerator
Productivity

Coding for Accelerator
Usability

Accelerator is important for Cognitive Computing (AI, DL, ...) in Cloud

System and Cloud Innovation is driving a new wave for Cloud opportunities in AI era.