XDF XILINX DEVELOPER FORUM

LSTM Deep Dive

Presented By

Jingxiu Liu, Director of Product Marketing, AI and Edge Computing
2018/10/1

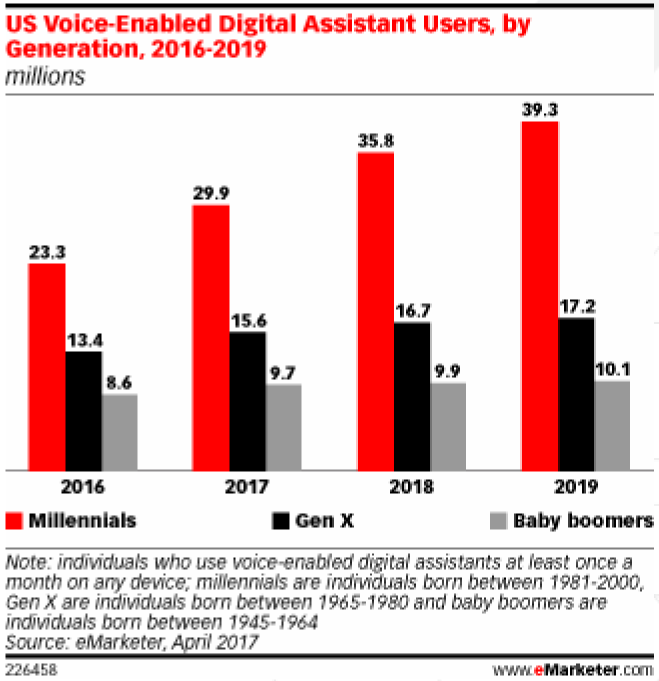
XILINX

© Copyright 2018 Xilinx

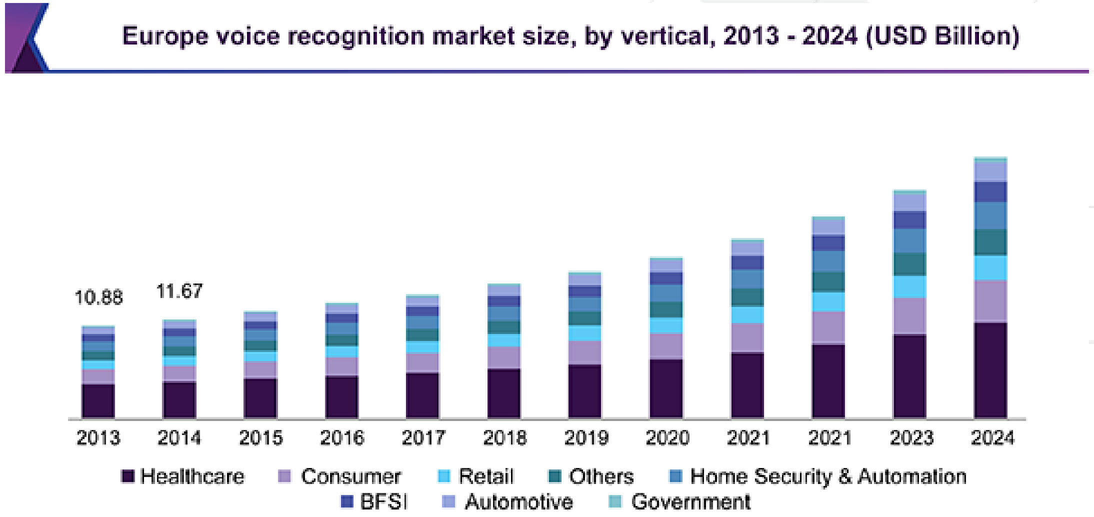
Speech to Text Application

How it drive business values?

- Usability : the accurate is reported on par with humans
  - the word error rate for Microsoft's 5.1%, while Google 4.9%.
- virtual assistants with speech recognition capabilities keeps increasing
- Change the way we interact with and build electronics
  - Voicebox : worked on voice recognition for partners including Samsung, AT&T, and Toyota.



The global voice recognition market size was valued at USD 55.17 billion in 2016 and is expected to grow at a CAGR of 11.0% during the forecast period of 2016 to 2024.



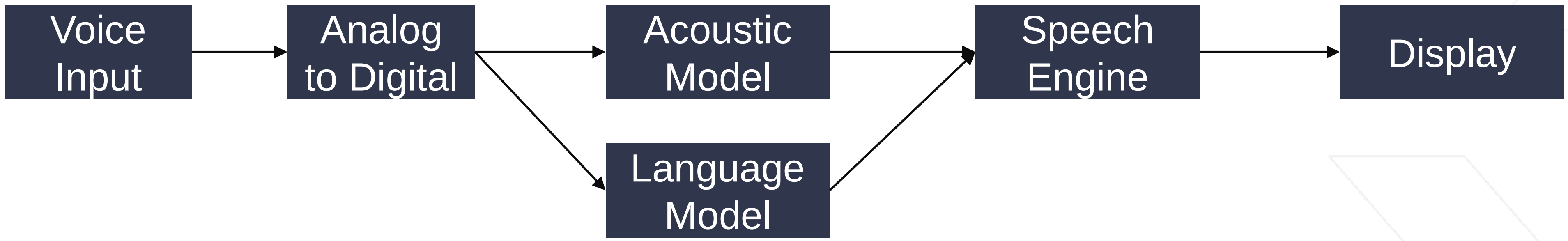
https://www.techemergence.com/ai-for-speech-recognition/

© Copyright 2018 Xilinx

Speech to Text Application

What is Speech to Text (Speech Recognition) ?

- understand voice by the computer and performing any required task.



What can it be used ?

- Dictation
- System control/navigation
- Voice dialing
- Commercial/Industrial applications

Facebook, Amazon, Microsoft, Google and Apple — are already offering this feature on various devices through services like Google Home, Amazon Echo and Siri.

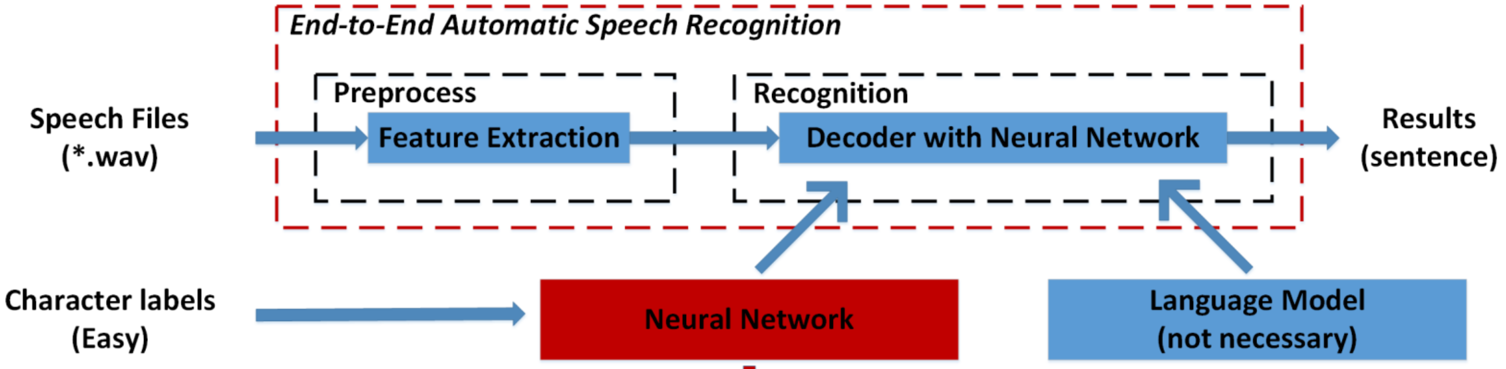
The try to make speech recognition a standard for most product.

© Copyright 2018 Xilinx

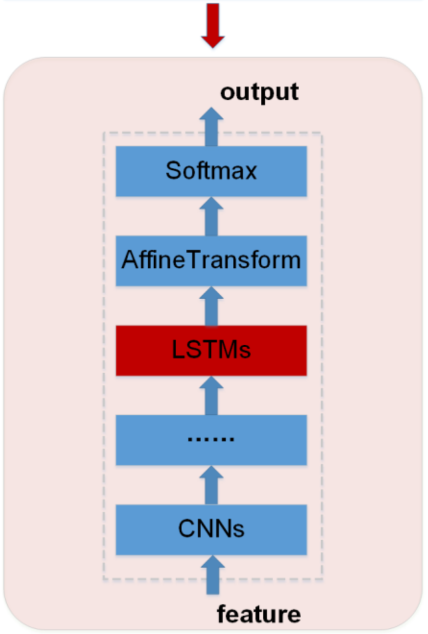
XILINX

Speech to Text Application

- **AI-based Solutions (LSTM)**



$$i_t = g(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$
$$f_t = g(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$
$$o_t = g(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$
$$c\_in_t = \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c)$$
$$c_t = f_t \bullet c_{t-1} + i_t \bullet c\_in_t$$
$$h_t = o_t \bullet \tanh(c_t)$$



Abilites:
(1) We have a demo base on Librispeech 1000h dataset under Deepspeech2 framework, which could show the entire flow of our algorithm, software and hardware co-design (containing pruning, quantization, compilation and FPGA inference);

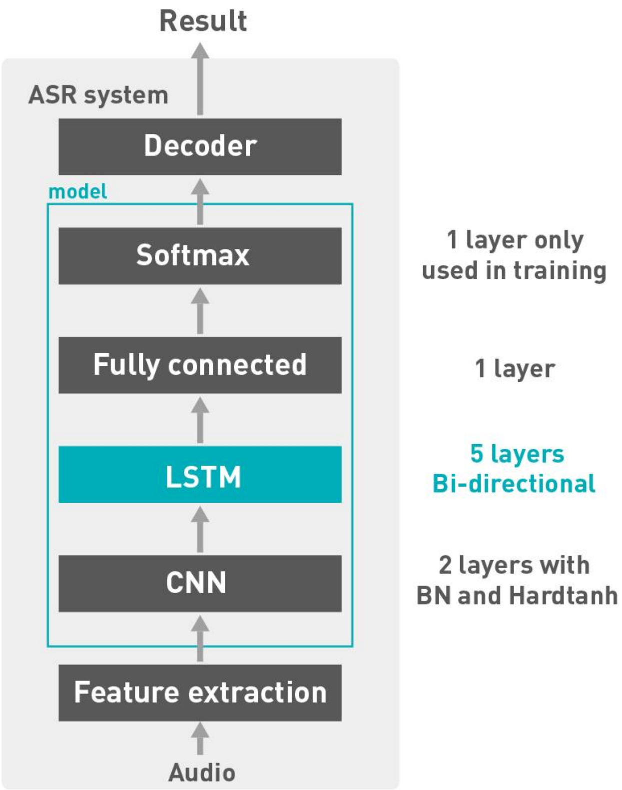
Notes:
(1) Compared to the traditional ASR system, a neural network with the same high accuracy is much more difficult to obtain. It means we haven't got a baseline with high accuracy.

Applications in end-to-end ASR

© Copyright 2018 Xilinx

Speech to Text Application

- **AI-based Solutions (LSTM)**



<table>
<tr><th>Time (ms)</th><th>Only CPU on AWS</th><th>CPU + FPGA on AWS</th><th>CPU + GPU P4 locally (cudnn)</th></tr>
<tr><td>Decoding</td><td>0.09</td><td>0.08</td><td>0.40</td></tr>
<tr><td>Softmax</td><td>0.07</td><td>0.08</td><td>0.08</td></tr>
<tr><td>Fully connected</td><td>0.11</td><td>0.73</td><td>0.12</td></tr>
<tr><td>LSTM</td><td>118.31</td><td rowspan="2">16.97</td><td>38.58</td></tr>
<tr><td>CNN</td><td>14.16</td><td>1.21</td></tr>
<tr><td>Feature extraction</td><td>2.78</td><td>2.73</td><td>1.95</td></tr>
<tr><td>E2E time</td><td>135.61</td><td>20.59 ★</td><td>42.35</td></tr>
</table>

2.06X speedup!

solutions	Devices and versions
CPU + FPGA on AWS	CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz (8 processors) FPGA: VU9P
Only CPU on AWS	CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz (8 processors)
CPU + GPU(P4) locally	CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz (56 processors) cuda: 8.0.44 cudnn: 6020

LSTM Solution Overview

XILINX

© Copyright 2018 Xilinx

Overview of LSTM solution

Algorithm

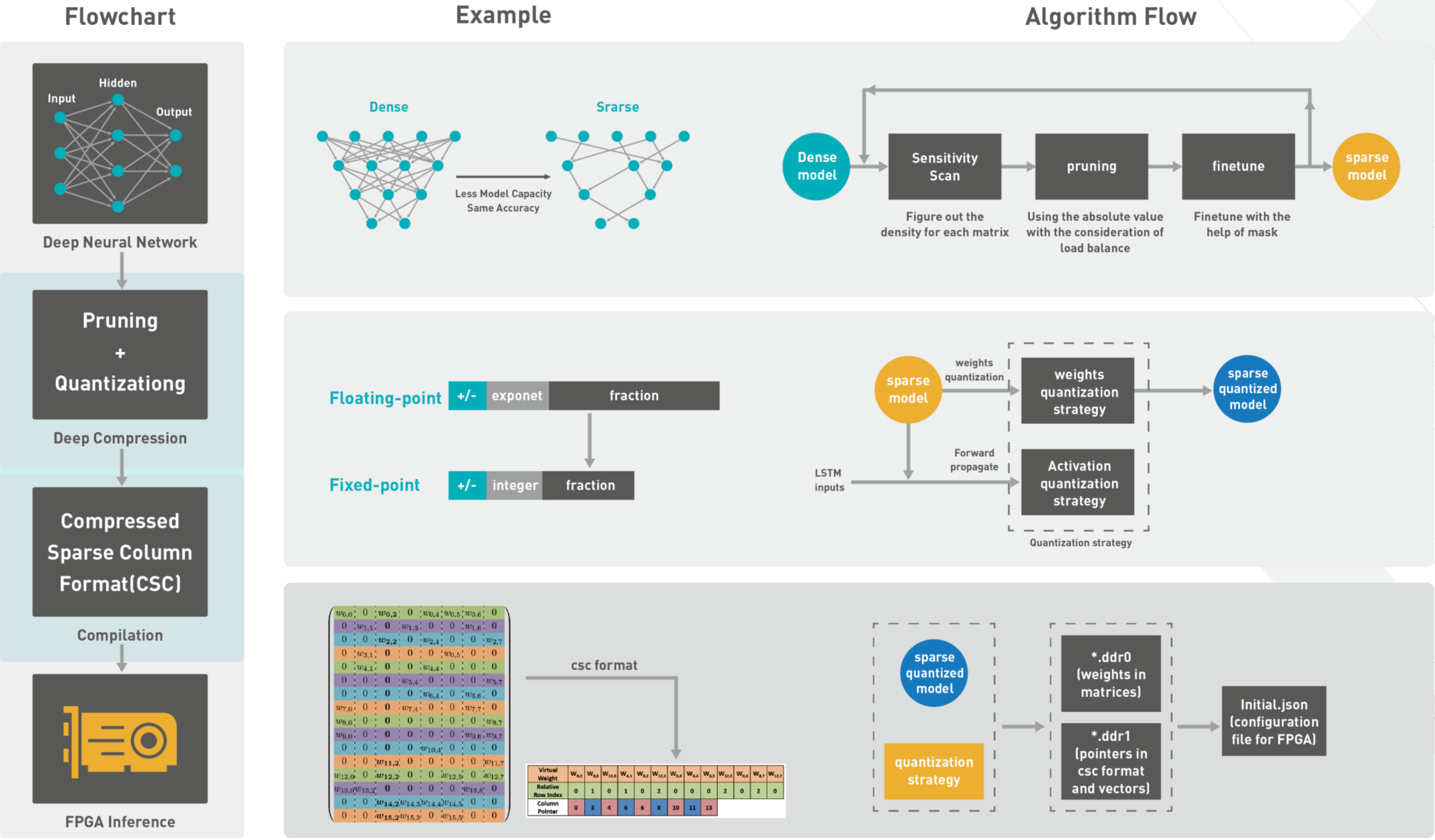
Deep Compression (Best paper of ICLR2016)

Software

+

ESE (Best paper of FPGA2017)

Hardware



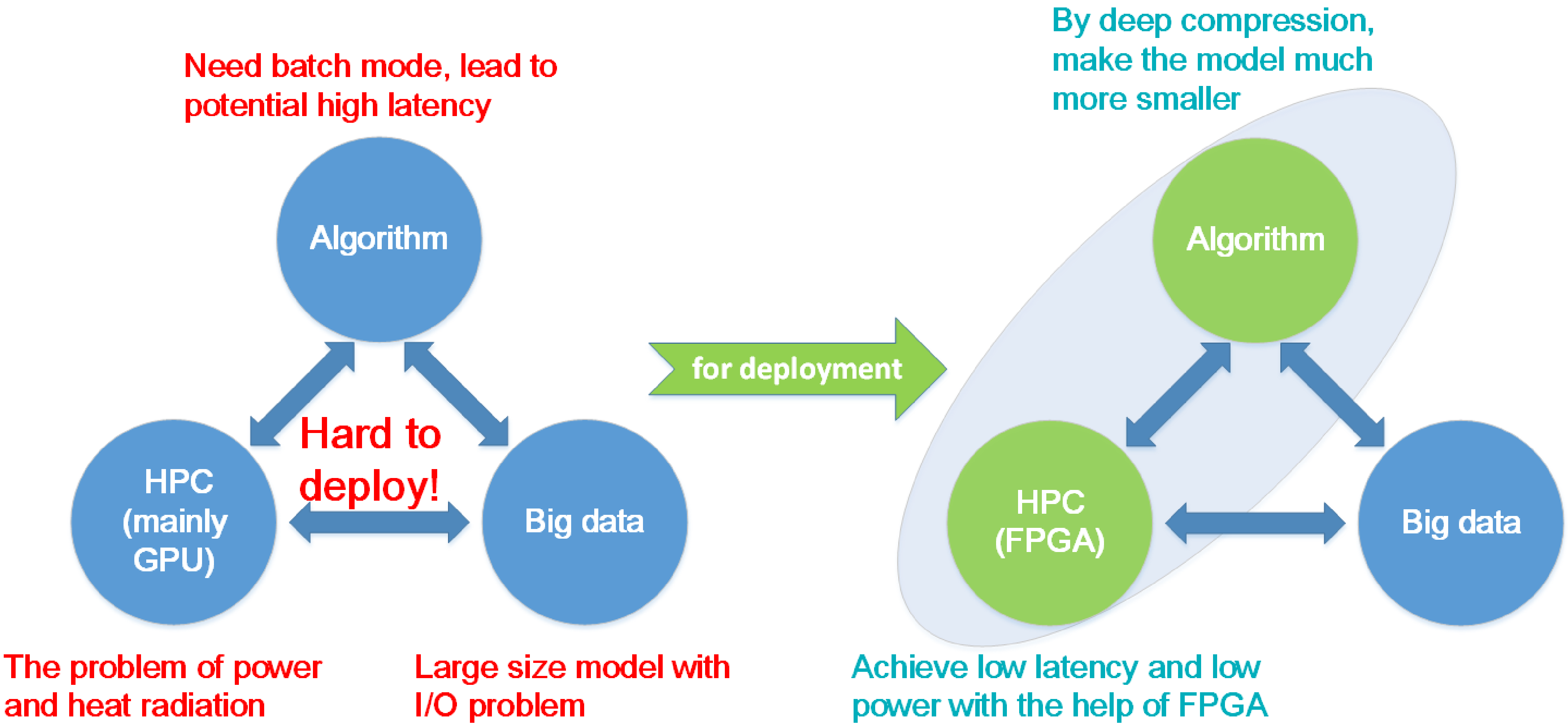
Our algorithm, software and hardware co-design flow for RNN(LSTM) acceleration

© Copyright 2018 Xilinx

XILINX

Overview of LSTM solution

What is the role of DeePhi in this game for LSTMs?



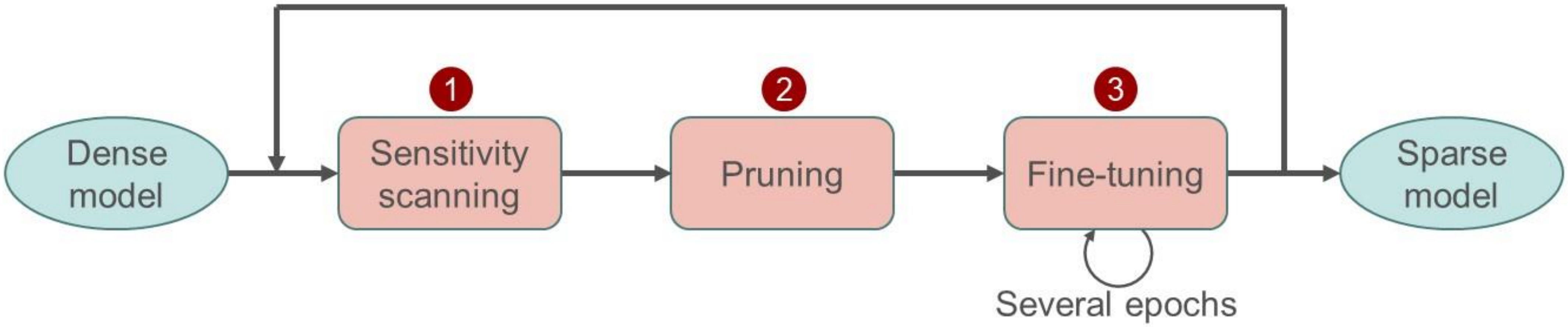
© Copyright 2018 Xilinx

Deep Compression for LSTM

© Copyright 2018 Xilinx

Deep Compression for LSTM

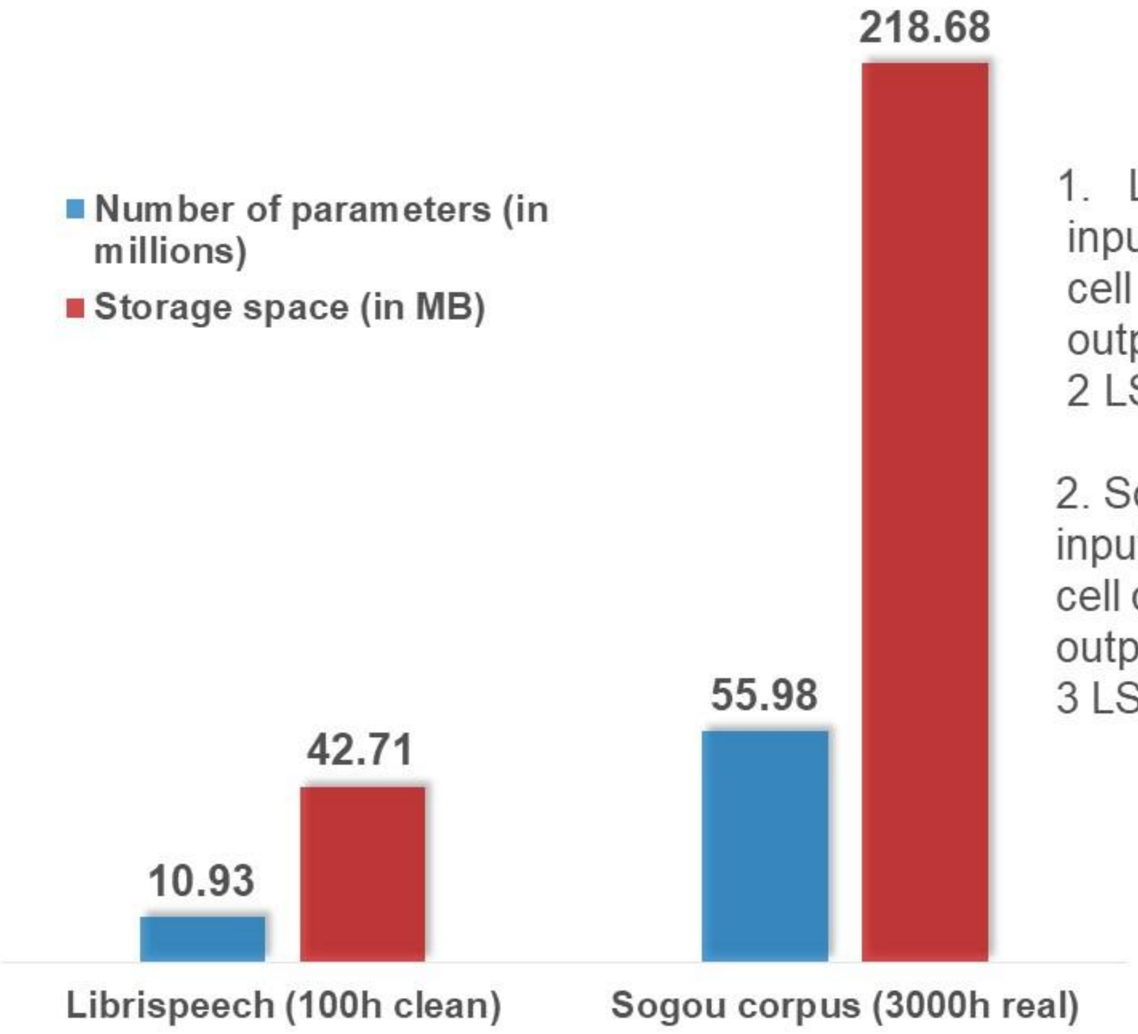
- Step 1: Perform sensitivity scanning to determine how much to be pruned
- Step 2: Prune parameters based on absolute values
- Step 3: Fine-tune to enhance rest parameters
- Repeat step 1-3 to achieve higher sparsity



Sensitivity-based pruning flow for LSTM

© Copyright 2018 Xilinx

Deep Compression for LSTM



1. Librispeech: input dim=440, cell dim=1024, output dim=512, 2 LSTM layers

2. Sogou: input dim=153, cell dim=2048, output dim=1024, 3 LSTM layers

Examples

© Copyright 2018 Xilinx

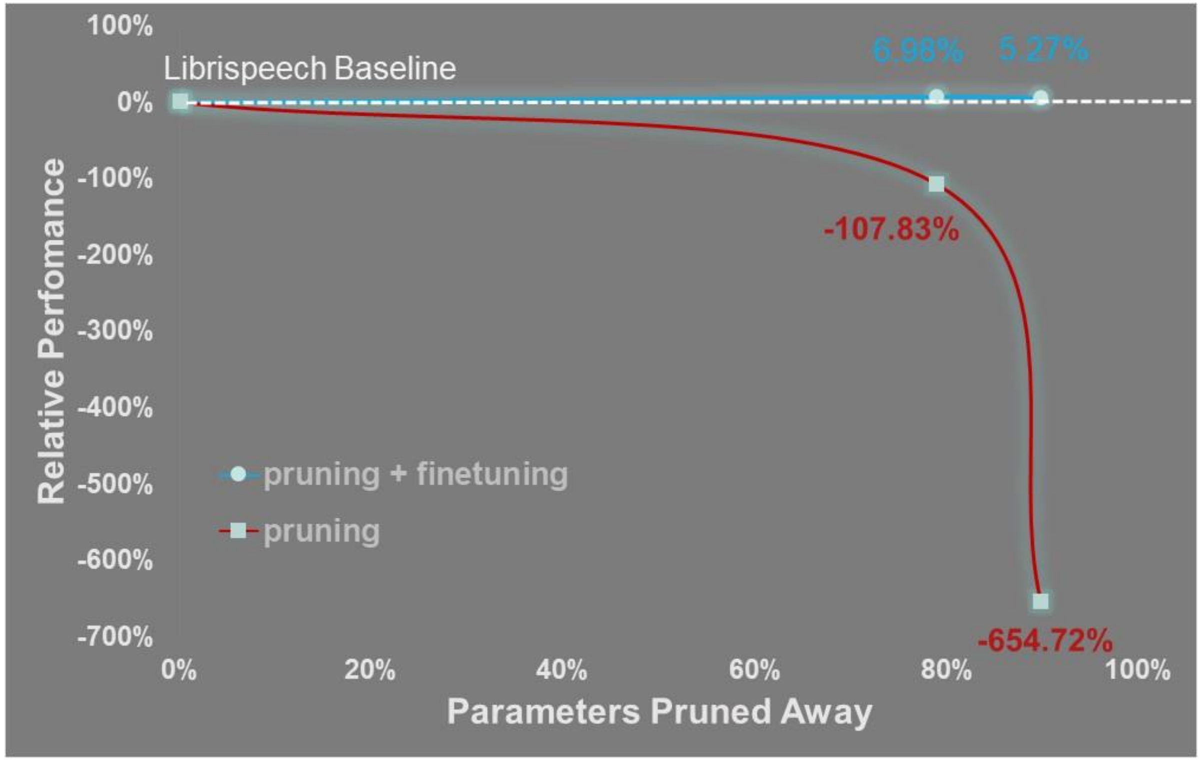
Deep Compression for LSTM

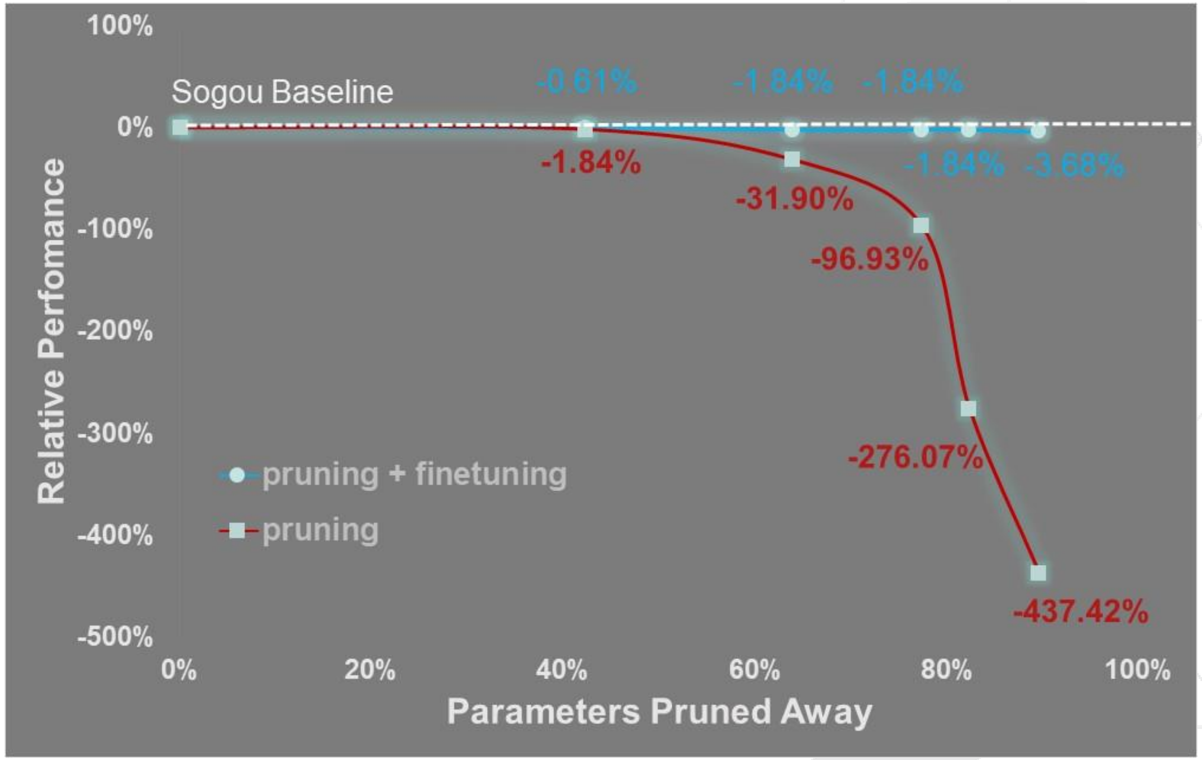
Dataset	Model	density	WER	Relative Perfomance
Small (100h)	Librispeech dense	100%	12.90	0.0%
	Librispeech sparse	21.1%	12.00	+6.98%
	Librispeech sparse	10.30%	12.22	+5.27%
Big (3000h)	Sogou dense	100%	16.3	0.0%
	Sogou sparse	17.76%	16.6	-1.84%
	Sogou sparse	10.47%	16.9	-3.68%

Density vs accuracy

© Copyright 2018 Xilinx

Deep Compression for LSTM



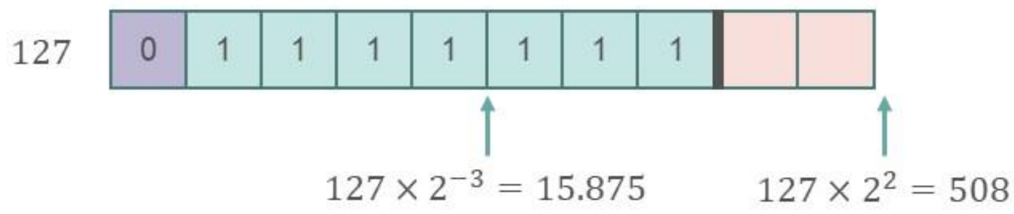


Sparsity vs accuracy

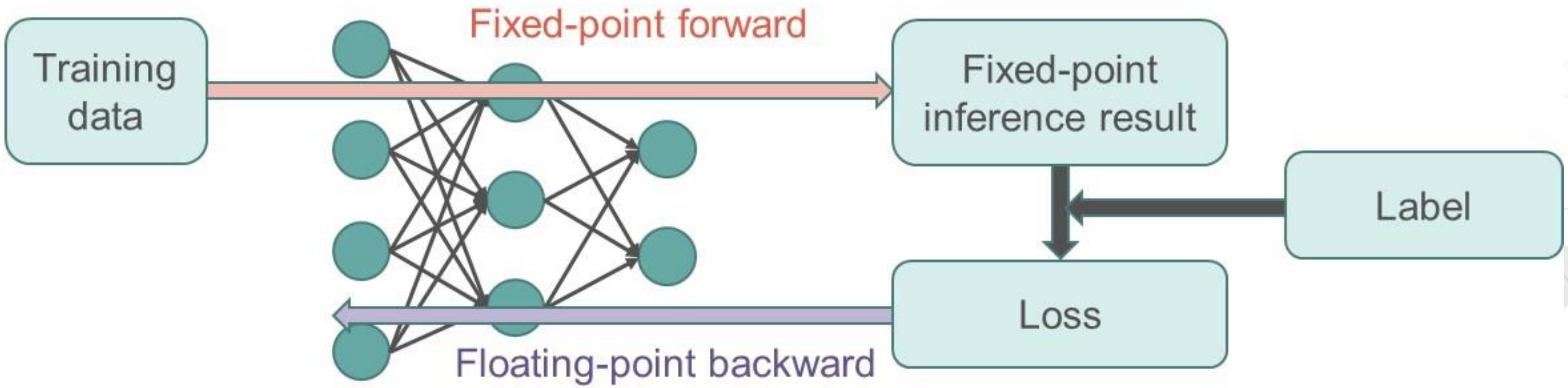
© Copyright 2018 Xilinx

Deep Compression for LSTM

Quantization strategy (e.g. 8bit)



Fixed-point re-training



Quantization and fixed-training for LSTM

© Copyright 2018 Xilinx

Deep Compression for LSTM

Dataset	Model	density	WER	Relative Perfomance
Small (100h) English	Librispeech dense	100%	12.90	0.0%
	Librispeech sparse	10.30%	12.22	+6.98%
	Librispeech sparse and quantized	10.30%	11.96	+7.29%
Big (3000h) Chinese	Sogou dense	100%	16.3	0.0%
	Sogou sparse	17.76%	16.6	-1.84%
	Sogou sparse and quantized	17.76%	16.6	-1.84%

1. Quantization strategy:
 for Librispeech, weights and x/y/m vectors are both directly quantized to 8bits.
 for Sogou network, weights are quantized to 12bits and all vectors are quantized to 16bits;
2. WER: word error rate

Deep compression vs accuracy

© Copyright 2018 Xilinx

XILINX

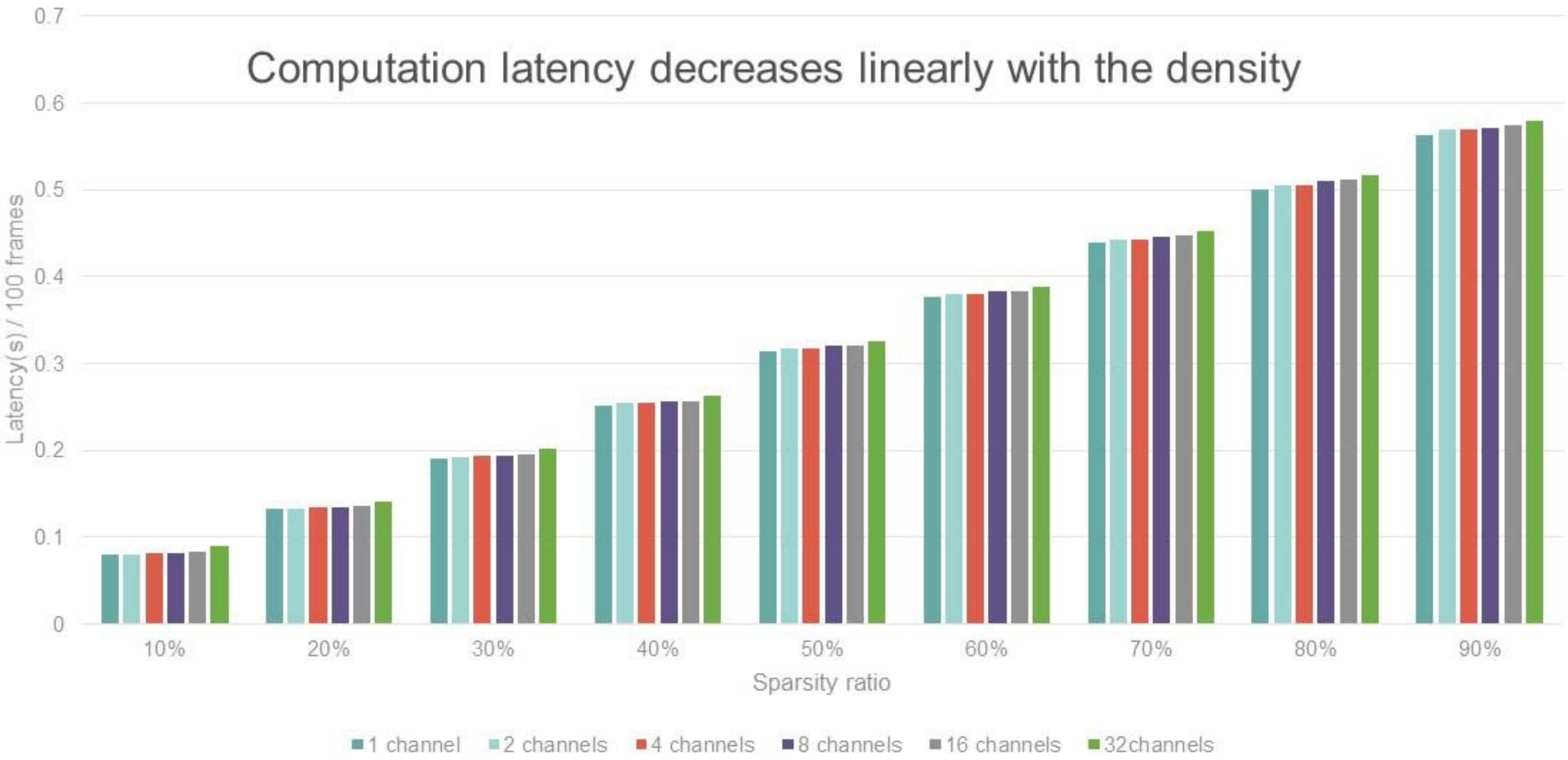
Compressed LSTM on ESE

XILINX

© Copyright 2018 Xilinx

Compressed LSTM on ESE

Three-layers LSTM



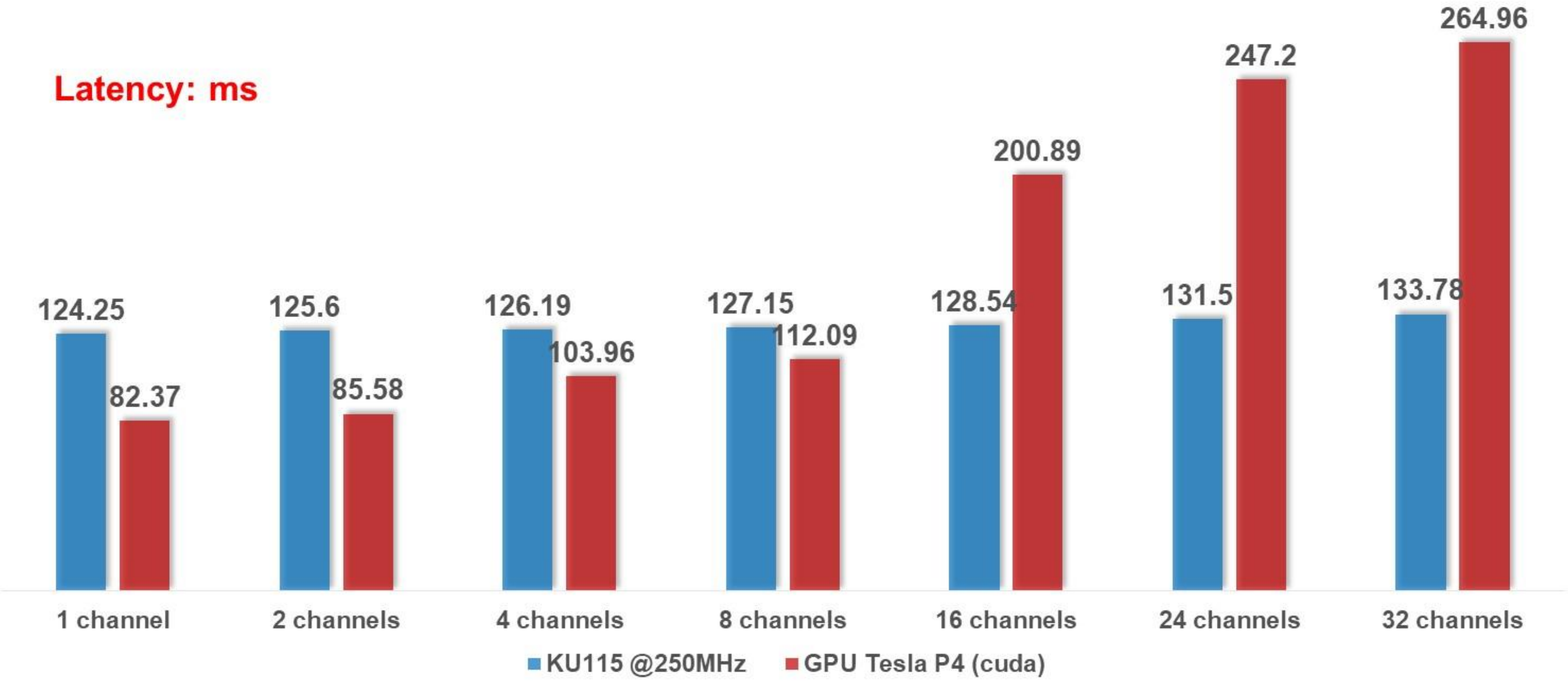
1. Net parameters: input dim=153, cell dim=2048, output dim=1024, 3 LSTM layers;
2. KU115 is clocked at 250MHz;
3. CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @2.60GHz 14 Cores 56 Threads

Latency under different density

© Copyright 2018 Xilinx

XILINX

Compressed LSTM on ESE

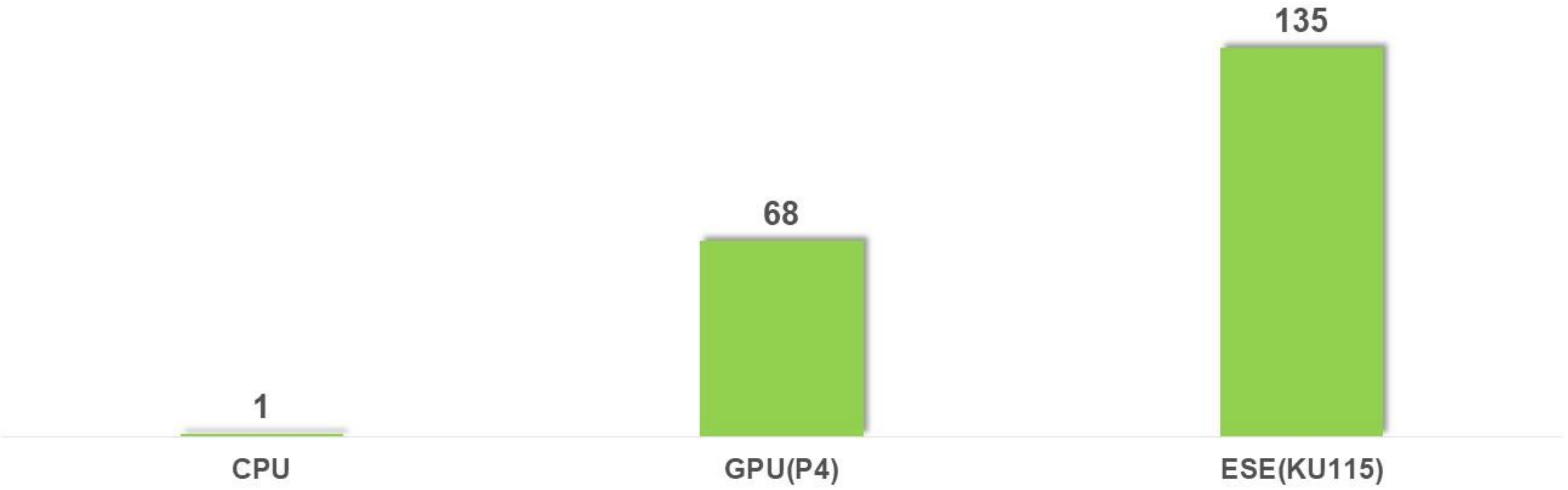


Parameters: input dim=153, cell dim=2048, output dim=1024, 3 LSTM layers , density=17.76%;
Local CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz (56 processors)
Non-merged version, 100 frames of input

Performance of LSTMP model for different channels based on Sogou 3000h dataset

© Copyright 2018 Xilinx

Compressed LSTM on ESE



Platforms	CPU(E5-2690 v4)	GPU(P4)	ESE(KU115)
Latency	18155.22ms	264.9576ms	133.778ms
Power	--	62 W	53 W
Speedup	1x	68x	135x

1. Nnet parameters: input dim=153, cell dim=2048, output dim=1024, 3 LSTM layers, density=17.76%;
2. ESE-KU115 is clocked at 250MHz;
3. 32 channel speech inputs in parallel.

Performance of LSTMP model for 32 channels compared to CPU and GPU

© Copyright 2018 Xilinx

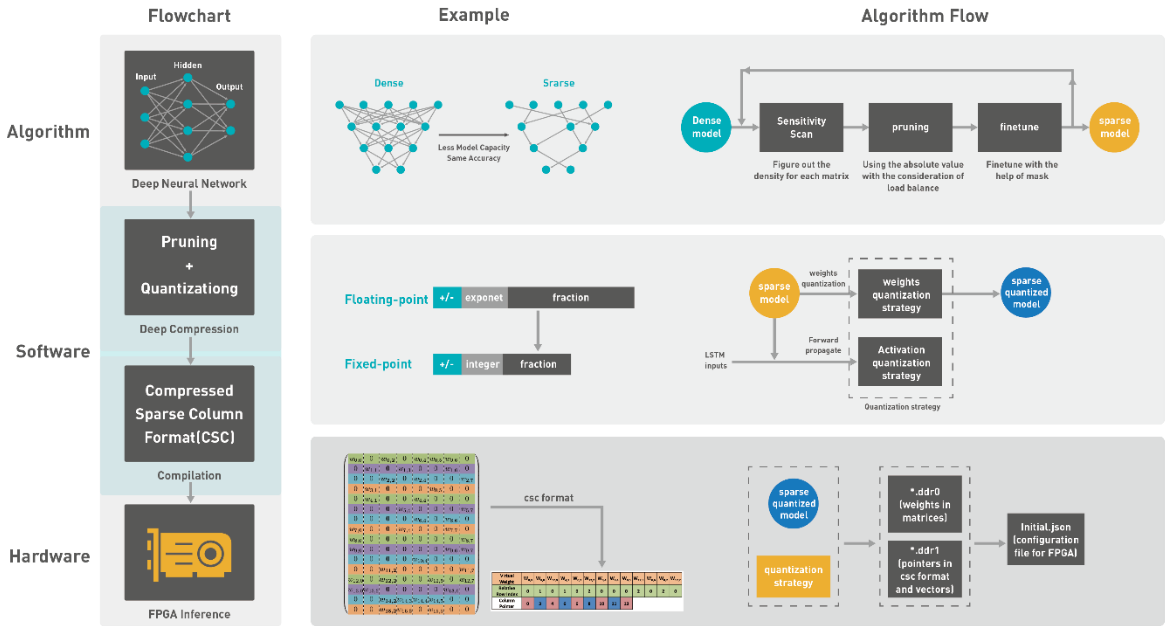
DDESE on Cloud

XILINX

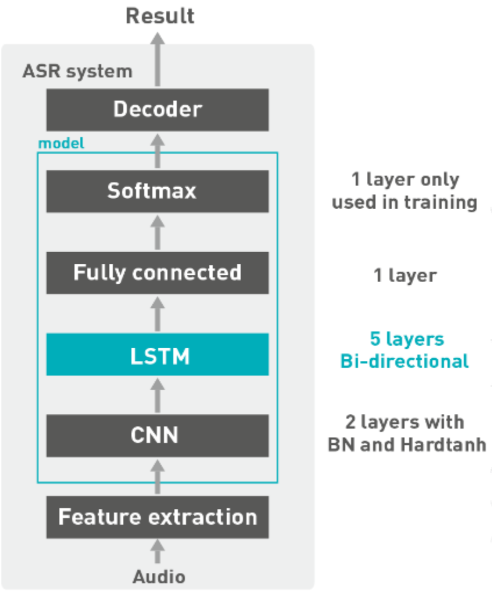
© Copyright 2018 Xilinx

DDESE on Clouds

Solution



for end-to-end speech recognition



Partners

- ✓ On clouds, aiming at customers all over the world
- ✓ Already officially launched in AWS Marketplace and HUAWEI cloud (http://www.deephi.com/ddese.html)
- ✓ Now transplanting to Alibaba cloud

Features

Low storage	Model compressed more than 10X with negligible loss of accuracy
Low latency	More than 2X speedup compared to GPU (P4)
Programmable	Reconfigurable for different requirements

© Copyright 2018 Xilinx

XILINX

DDESE on Clouds

The same model

	amazon web services EC2	HUAWEI	Alibaba Cloud aliyun.com
Algorithm	Accelerating CNN/BLSTM Pruning BLSTM to 15% 16bit weights/activations	Accelerating CNN/BLSTM Pruning BLSTM to 15% 16bit weights/activations	Accelerating LSTMP Pruning LSTMP to 26.9% 16bit weights/activations
Software	DeepSpeech2 + PyTorch Tools for compression and compilation	DeepSpeech2 + PyTorch Tools for compression and compilation	DeepSpeech2 + PyTorch Tools for compression and compilation
Hardware	Based on VU9P 220MHz 1 channel	Based on VU9P 200MHz 1 channel	Based on KU115 300MHz 1 channel

© Copyright 2018 Xilinx

DDESE on Clouds

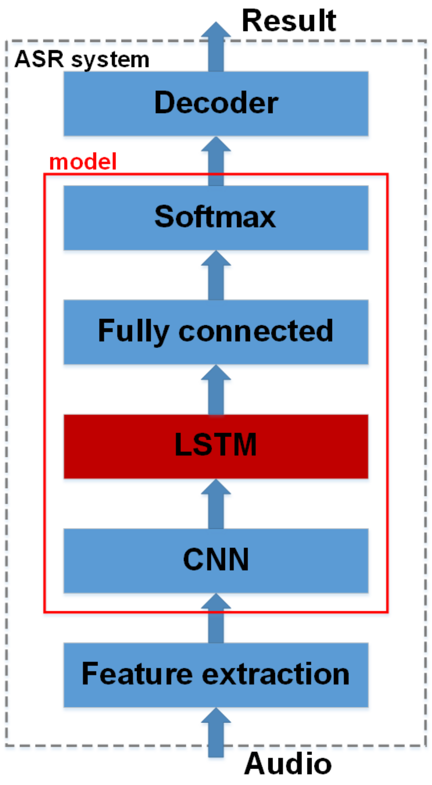
Features

Innovative full-stack acceleration solution for deep learning in acoustic speech recognition (ESE: best paper of FPGA2017)

- Support both unidirectional and bi-directional LSTM acceleration on FPGA for model inference
- Support CNN layers, Fully-Connected (FC) layers, Batch Normalization layers and varieties of activation functions such as Sigmoid, Tanh and HardTanh
- Support testing for both performance comparison of CPU/FPGA and single sentence recognition
- Supporting user's own test audio recognition (English, 16kHz sample rate, no longer than 3 seconds)

Usage	Hardware PCIE interface, software API
Supported layer	CNN、uni/bi-directional LSTM(P)、FC、BN
LSTM layer number	According to the requirements and source
Channel number	According to the requirements and source
Quantization	16bit
Maximum input of LSTM	1024
Maximum size of LSTM	2048
Density of LSTM	Any, typically 10%~20%
Peephole in LSTM	Selectable
Projection in LSTM	Selectable
Activation function	Sigmoid、Tanh、HardTanh

Note: It is available to change the hardware configuration for different requirements such as layer number and model size



1 layer
only used in training

1 layer

5 layers
uni- or bi- directional
input = 672, cell = 800

2 layers
with BN and Hardtanh

DDESE on AWS

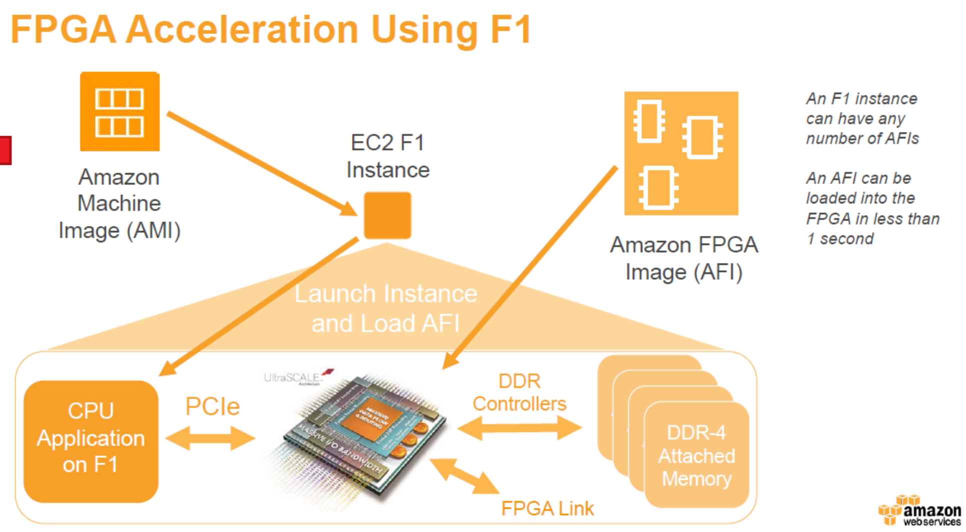
✓ **Already officially launched in AWS Marketplace**

https://aws.amazon.com/marketplace/pp/B079N2J42R?qid=1523443241195&sr=0-1&ref_=srh_res_product_title

DDESE

Version	Launch	Description
V1.0	2017.12	acceleration for unidirectional and bi-directional LSTM model
V2.0	2018.02	acceleration for CNN + bi-directional LSTM model

Add DeePhi solution



David Pellerin from AWS in HotChips2017

Model Description

	Released model from GitHub	Our model (low accuracy)	Our model (high accurcy)
RNN type	GRU	LSTM(without projection)	LSTM(without projection)
RNN layers	5	5	5
RNN input	672	672	672
RNN size	800	800	800
bi-directional	yes	no	yes
batch normalization	yes	no	yes
WER on libri-test-clean	11.20	15.608	10.764
Best model after compressing LSTM	---	20% density, LSTM 16bit WER=17.033	15% density, LSTM 16bit WER=11.51
Best model after further quantizing CNN	---	---	CNN 16bit weight/ACT WER=11.52

© Copyright 2018 Xilinx

XILINX

Performance of DDESE on AWS

➢ **For LSTM layers only:**

	Baseline (cudnn)	Compression	Hardware	Latency	Speedup
Unidirectional LSTM	WER = 15.608 E2E = 26.42ms **LSTM = 22.60ms**	WER = 17.033 Density = 20% 16bit weight/ACT	290MHz	E2E = 23.76ms **LSTM = 7.88ms**	1.11X **2.87X**
Bi-directional LSTM	WER = 10.764 E2E = 42.35ms **LSTM = 38.58ms**	WER = 11.51 Density = 15% 16bit weight/ACT	200MHz	E2E = 32.61ms **LSTM = 15.07ms**	1.30X **2.56X**

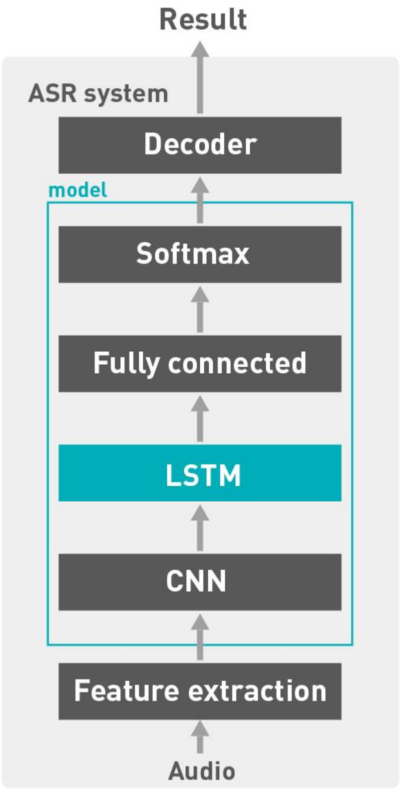
➢ **For CNN + bi-directional LSTM layers:**

	Baseline (cudnn)	Compression	Hardware	Latency	Speedup
CNN+ BLSTM	WER = 10.764 E2E = 42.35ms **CNN = 1.21ms** **LSTM = 38.58ms**	WER = 11.52 LSTM density = 15% All 16bit weight/ACT	**220MHz** （on going）	E2E 20.59ms **CNN+STM 16.97ms**	2.06X

Note: E2E is short for end-to-end, ACT is short for activation, WER is short for word error rate, input: 1 second.

© Copyright 2018 Xilinx

Details of CNN + bi-directional model on AWS



1 layer only used in training

1 layer

5 layers Bi-directional

2 layers with BN and Hardtanh

Time (ms)	Only CPU on AWS	CPU + FPGA on AWS	CPU + GPU P4 locally (cudnn)
Decoding	0.09	0.08	0.40
Softmax	0.07	0.08	0.08
Fully connected	0.11	0.73	0.12
LSTM	**118.31**	**16.97**	**38.58**
CNN	14.16		1.21
Feature extraction	2.78	2.73	1.95
E2E time	*135.61*	*20.59* ★	*42.35*

2.06X speedup!

solutions	Devices and versions
CPU + FPGA on AWS	CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz (8 processors) FPGA: VU9P
Only CPU on AWS	CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz (8 processors)
CPU + GPU(P4) locally	CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz (56 processors) cuda: 8.0.44 cudnn: 6020

© Copyright 2018 Xilinx

XILINX.

DDESE on HUAWEI Cloud

HUAWEI

- ✓ **Already officially launched on HUAWEI Cloud**
  - https://app.huaweicloud.com/product/00301-110982-0--0
- ✓ **Based on VU9P @200MHz, 1 channel**
- ✓ **Using CNN + bi-directional LSTM model (with high accuracy)**

Using this model

Model Description

	Released model from GitHub	Our model (low accuracy)	Our model (high accury)
RNN type	GRU	LSTM(without projection)	LSTM(without projection)
RNN layers	5	5	5
RNN input	672	672	672
RNN size	800	800	800
bi-directional	yes	no	yes
batch normalization	yes	no	yes
WER on libri-test-clean	11.20	15.608	10.764
Best model after compressing LSTM	---	20% density, LSTM 16bit WER=17.033	15% density, LSTM 16bit WER=11.51
Best model after further quantizing CNN	---	---	CNN 16bit weight/ACT WER=11.52

Performance of CNN+BLSTM

Solution	Latency (ms)	Speedup
GPU (P4)	39.79	1
AWS @220MHz	16.97	2.34
HUAWEI @200MHz	18.60	2.14

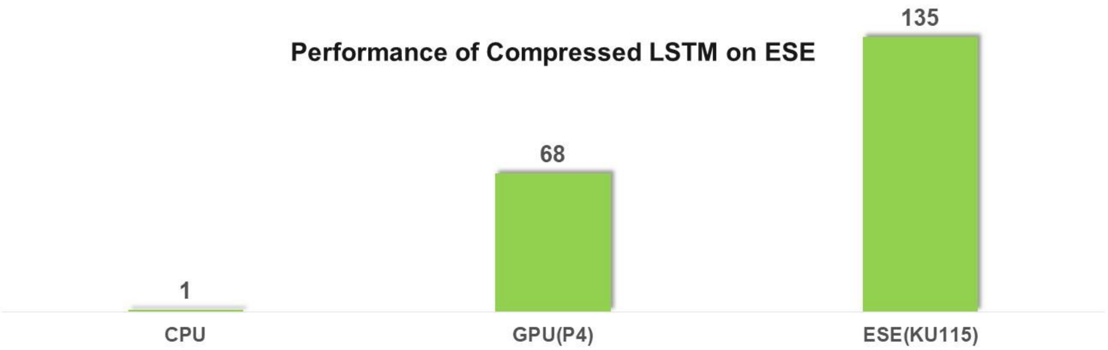
© Copyright 2018 Xilinx

XILINX

Review of Typical Products

Deep Compression vs Accuracy

Dataset	Model	density	WER	Relative Perfomance
Chinese (3000h real corpus)	Sogou dense	100%	16.3	0.0%
	Sogou sparse	17.76%	16.6	-1.84%
	Sogou sparse and quantized	17.76%	16.6	-1.84%



Platforms	CPU(E5-2690 v4)	GPU(P4)	ESE(KU115)
Latency	18155.22ms	264.9576ms	133.778ms
Power	--	62 W	53 W
Speedup	1x	68x	135x

1. Nnet parameters: input dim=153, cell dim=2048, output dim=1024, 3 LSTM layers, density=17.76%;
2. ESE-KU115 is clocked at 250MHz;
3. 32 channel speech inputs in parallel.

ESE for Sogou

Partners

amazon webservices EC2

Alibaba Cloud aliyun.com

HUAWEI

- ✓ On clouds, aiming at customers all over the world
- ✓ Already officially launched in AWS Marketplace and HUAWEI cloud (http://www.deephi.com/ddese.html)
- ✓ Soon available on Alibaba cloud

Features

Low storage	Model compressed more than 10X with negligible loss of accuracy
Low latency	More than 2X speedup compared to GPU (P4)
Programmable	Reconfigurable for different requirements

High performance	Corresponding to 2.52 TOPS for dense model

DDESE on Clouds

© Copyright 2018 Xilinx

XILINX

AI Boosting On-premise Solutions

DEEPHi

DEEPHi DNNDK

DEEPHi DPU

XILINX® ALVEO™

Adaptable.
Intelligent.

XILINX®

© Copyright 2018 Xilinx