Ouroboros: An Inference Engine for Deep Learning Based TTS on Embedded Devices

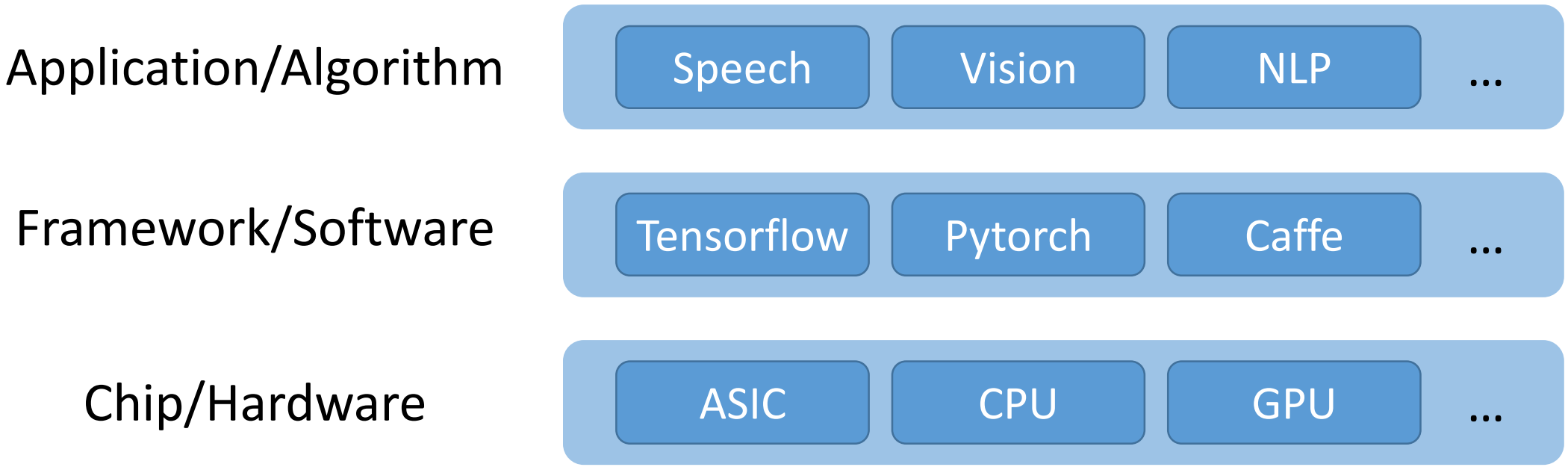
Jiansong Zhang, Lixue Xia, Zhao Jiang, Hao Liang, Jiaoyan Chen, Shouda Liu, Wei Lin, Yuan Xie

Alibaba Group

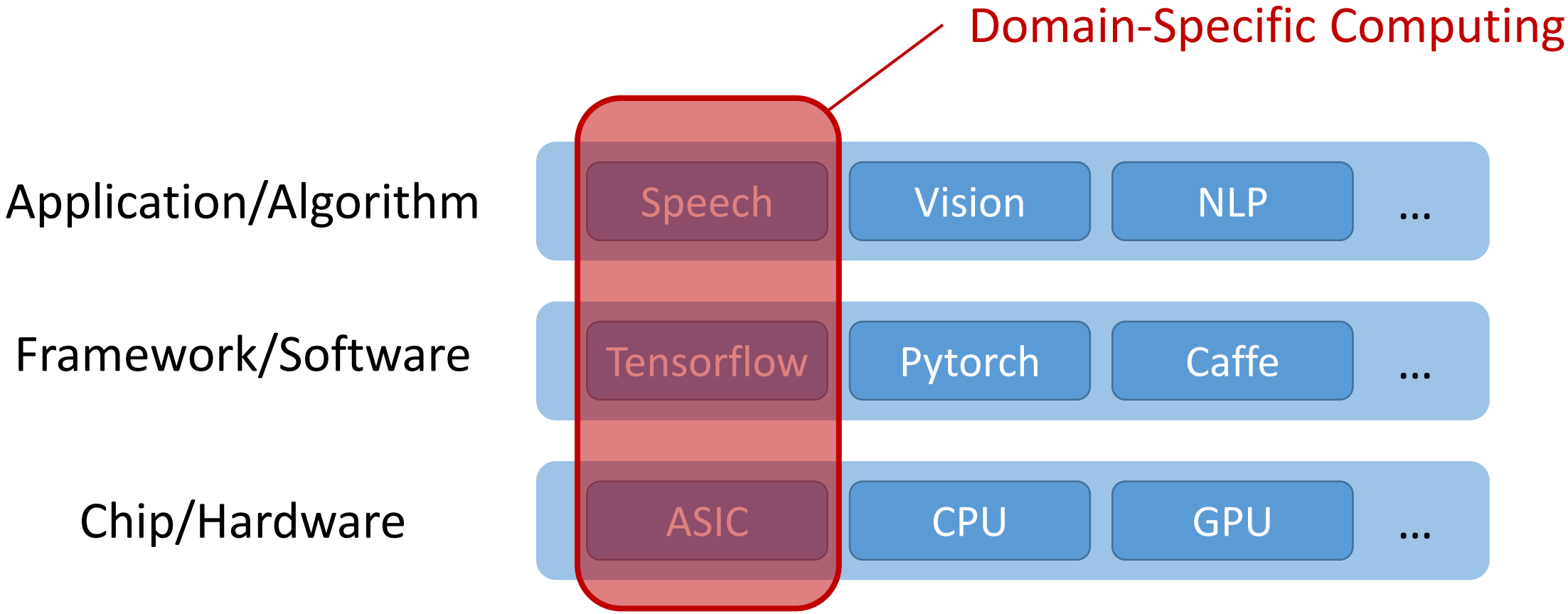
Algorithm
Big Data
Computing
Domain-Specific
Ecosystem

Domain-Specific Computing

Alibaba Group



Domain-Specific Computing



A Case of Domain-Specific Computing in Alibaba

Alibaba Group

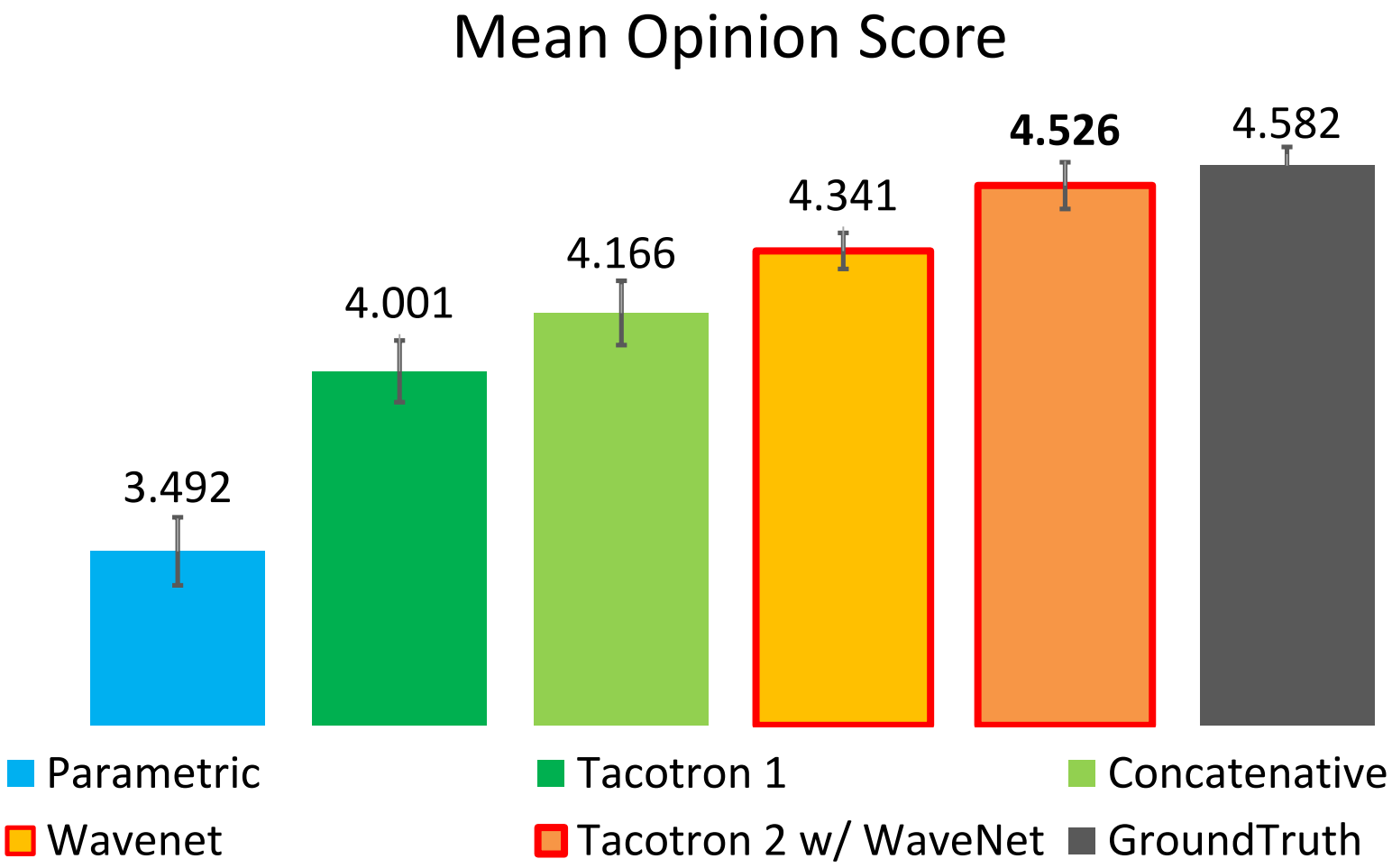
DAMO

ALIBABA DAMO ACADEMY

Founded in Oct. 2017
Computing Technology Lab

Smart Speakers, 10M+ units sold
A lot of other speech scenarios

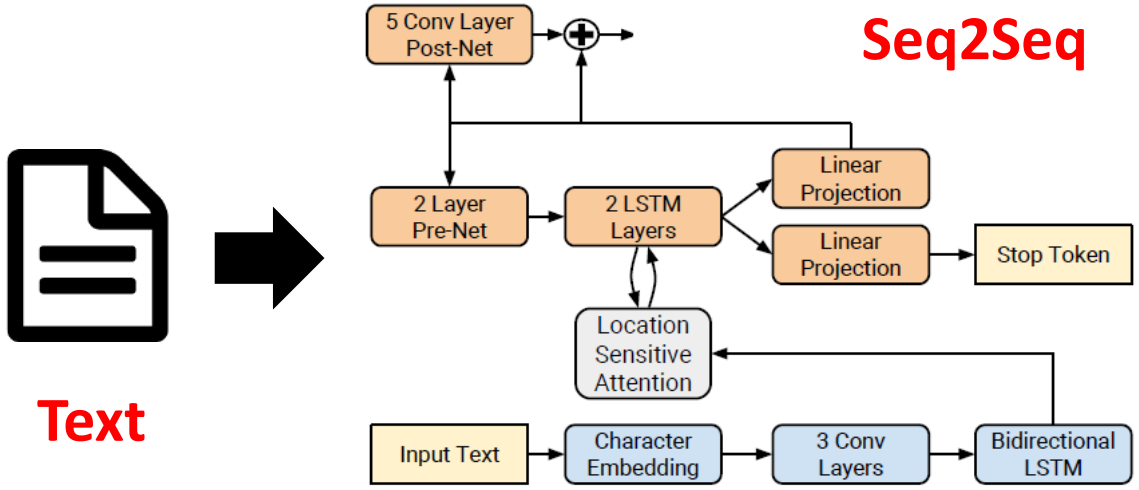
Background – TTS Quality with Deep Learning



※ Numbers from the Tacotron 2 paper: https://arxiv.org/abs/1712.05884

Background – End-to-End Deep Learning Model

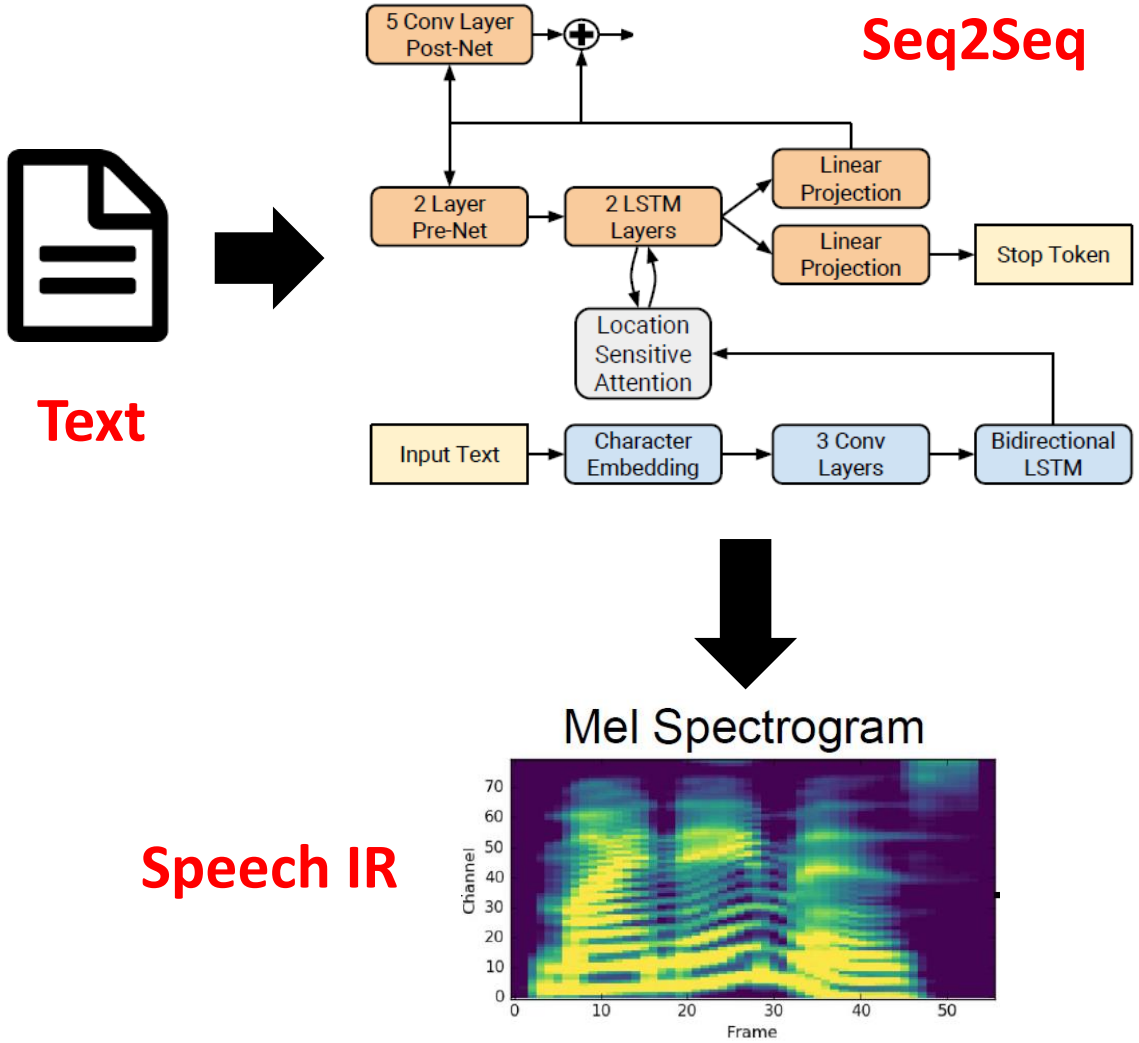
Alibaba Group



※ Tacotron 2 as in this example

Background – End-to-End Deep Learning Model

Alibaba Group



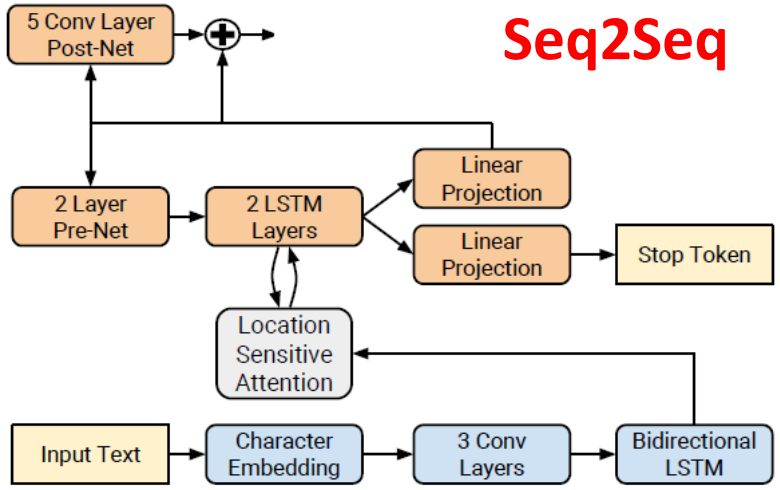
※ Tacotron 2 as in this example

Background – End-to-End Deep Learning Model

Alibaba Group

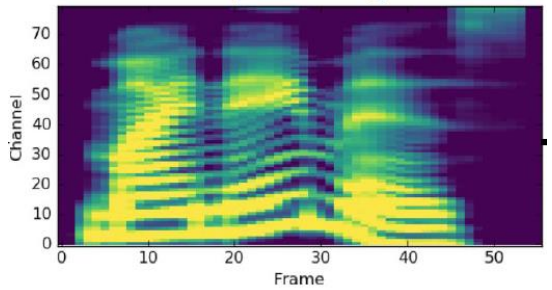
Text

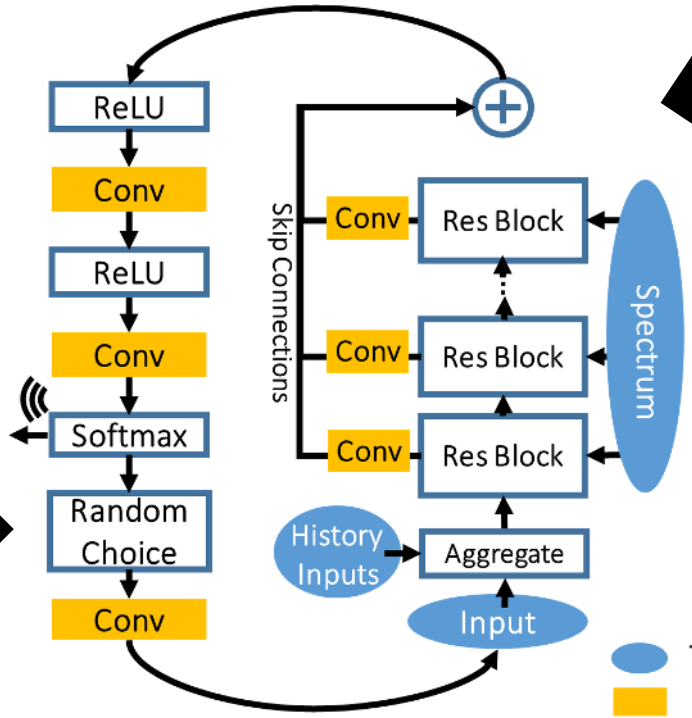
Seq2Seq



Speech IR

Mel Spectrogram



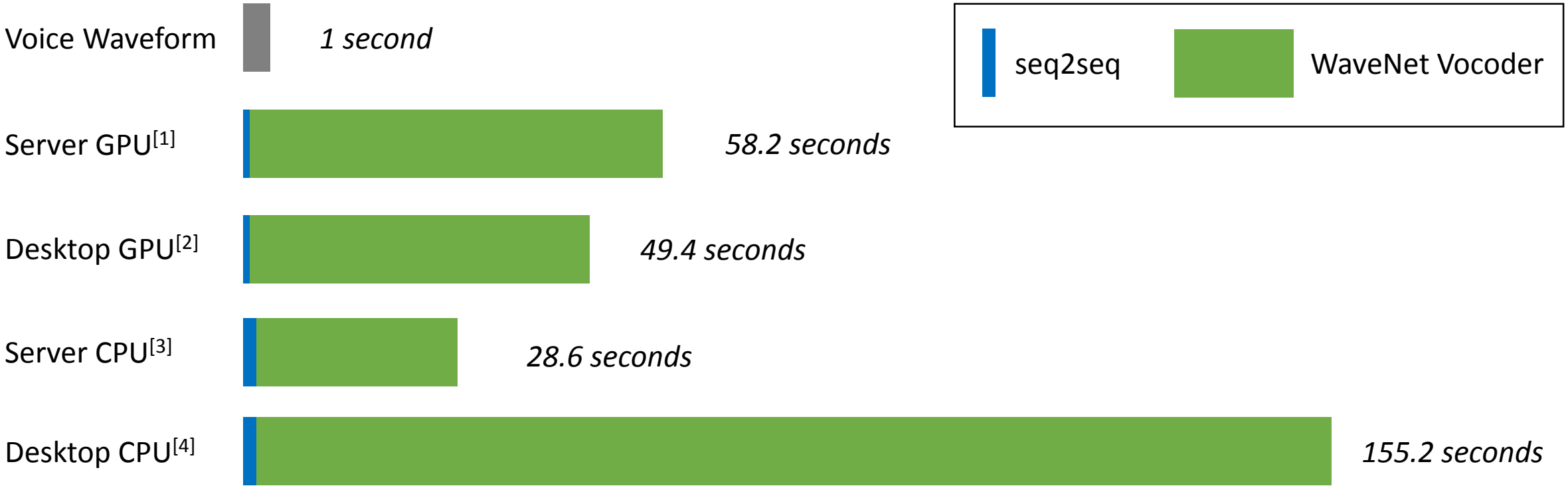


Vocoder (WaveNet)

※ Tacotron 2 as in this example

Deep Learning Based TTS is Costly to Compute

Alibaba Group

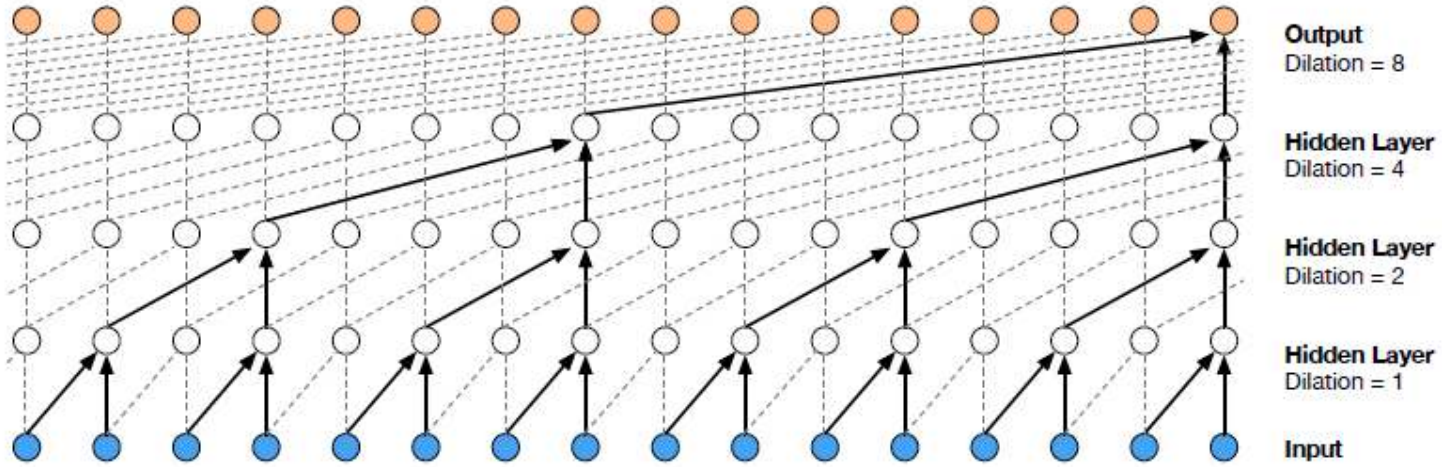


[1] Server GPU: Nvidia T4; [2] Desktop GPU: Nvidia 1070
[3] Server CPU: Intel Xeon 8163 x2; [4] Desktop CPU: Intel Core i7-5930K;
※ Measured using Tacotron-2's Tensorflow implementation (TensorRT makes little difference)

WaveNet Vocoder In-depth

Alibaba Group

- Autoregressive, Generative Model for Raw Audio
  - 16k samples/second → (**heavy**) dense sampling, NN inference for every sample
  - Casual convolutions → (**iterative**) dependency on previous samples

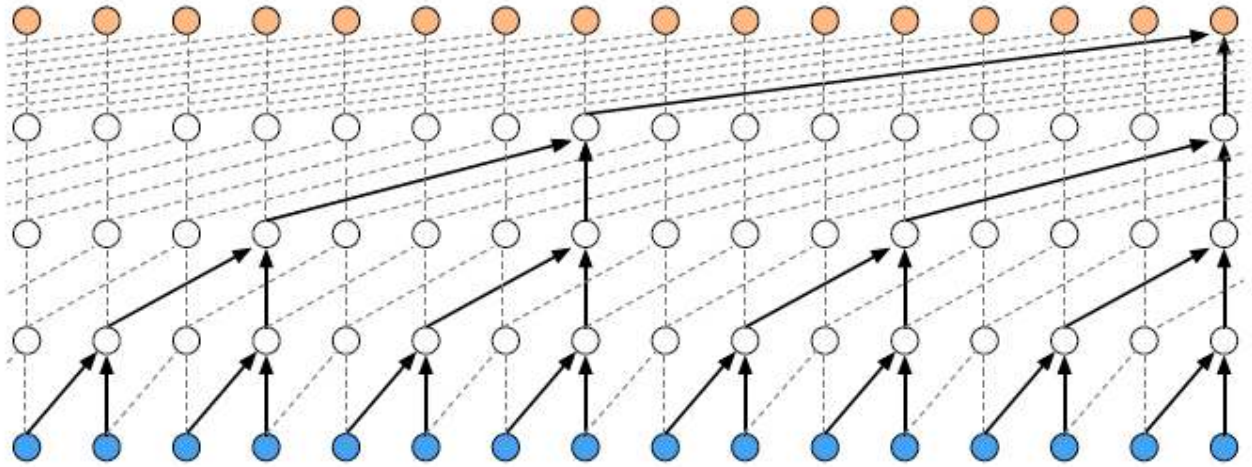


※ Figure from the WaveNet paper: https://arxiv.org/abs/1609.03499

WaveNet Vocoder In-depth

Alibaba Group

- Autoregressive, Generative Model for Raw Audio
  - 16k samples/second → (**heavy**) dense sampling, NN inference for every sample
  - Casual convolutions → (**iterative**) dependency on previous samples



Million-level sequential operators per second, huge kernel-launch overhead!

※ Figure from the WaveNet paper: https://arxiv.org/abs/1609.03499

Related Work

Alibaba Group

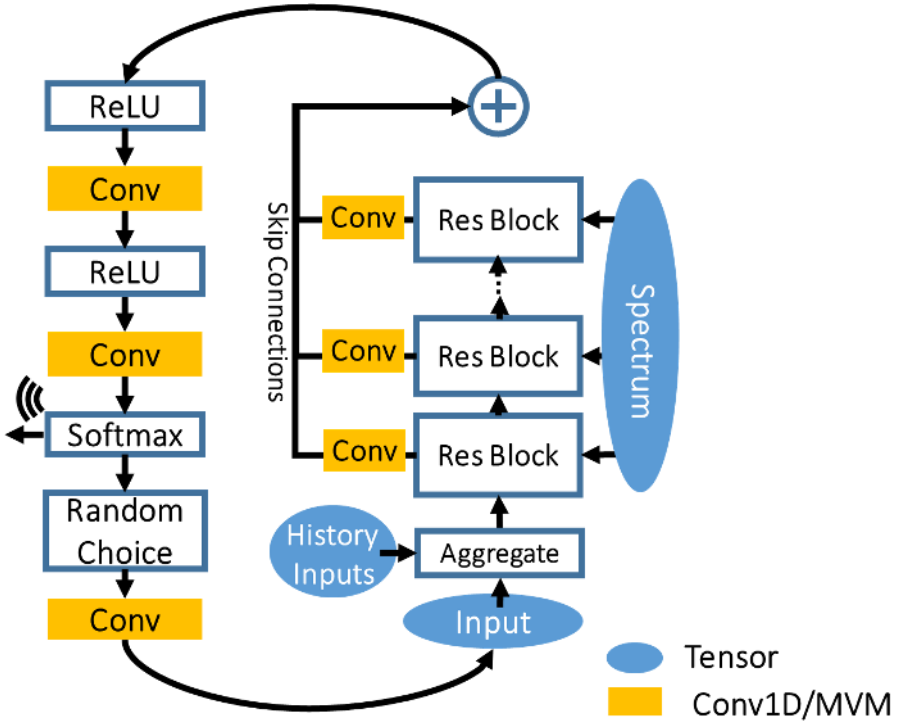
- Algorithm-layer Solution
  - Parallel Wavenet[1] – adding a transfer learning step from trained WaveNet model
  - WaveRNN[2], LPCNet[3] – adding hidden states, only relying on the last sample

- Software-layer Solution
  - Nv-wavenet[4] – deeply customized cuda kernel
  - Other CPU/GPU implementations[5]

[1] Parallel Wavenet: https://arxiv.org/abs/1711.10433;
[2] WaveRNN: https://arxiv.org/abs/1802.08435;
[3] LPCNet: https://arxiv.org/abs/1810.11846;
[4] Nv-wavenet: https://github.com/NVIDIA/nv-wavenet;
[5] http://on-demand.gputechconf.com/gtc/2017/presentation/s7544-andrew-gibiansky-efficient-inference-for-wavenet.pdf

Hardware-Layer Solution – Ouroboros

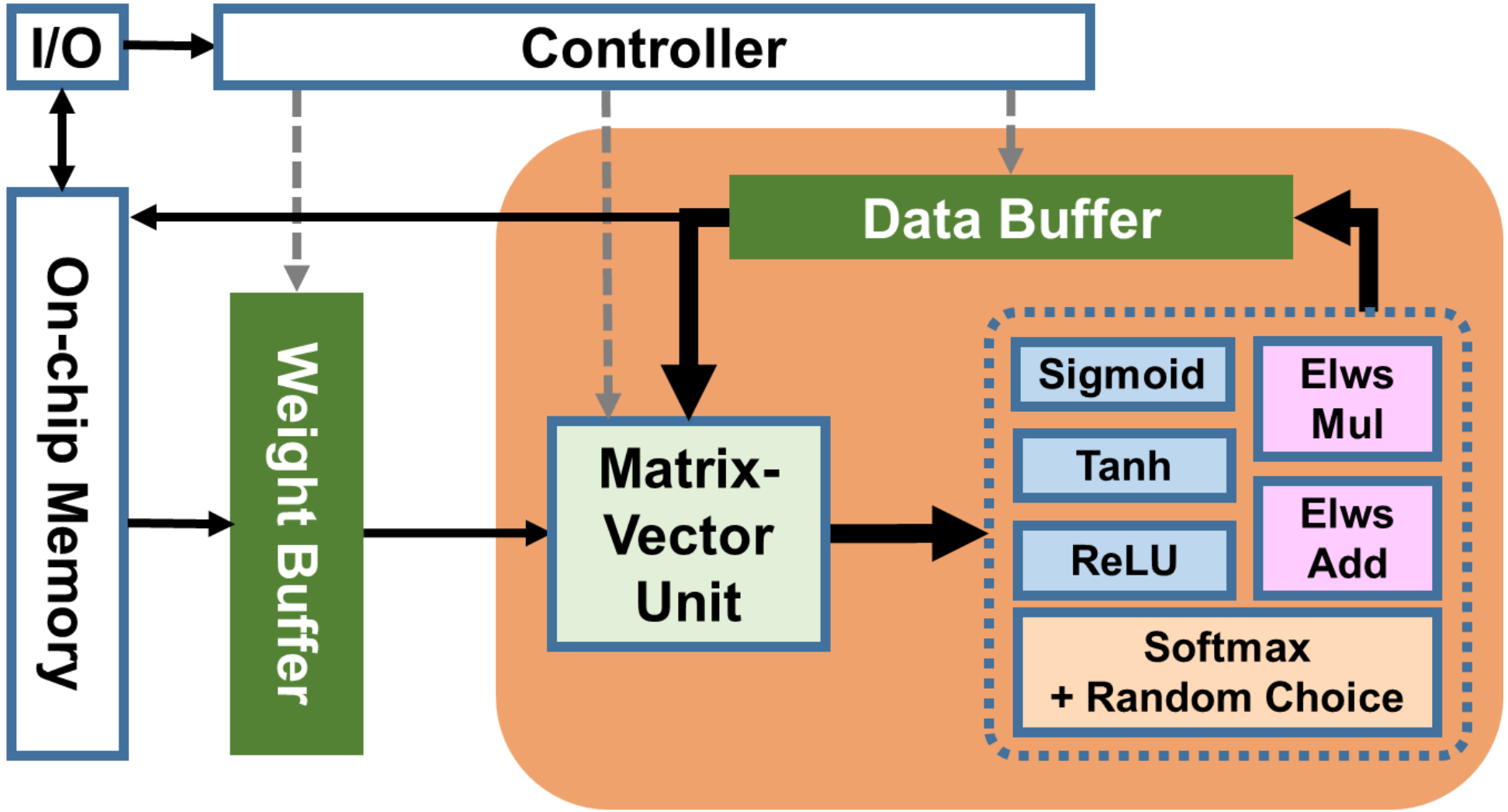
- Design Idea – supporting iterative algorithm using on-chip loop structure
- Benefits – no additional model training overhead; versatility



※ The right figure: A copy of a 1478 drawing by Theodoros Pelecanos of an alchemical tract attributed to Synesius – From Wikipedia

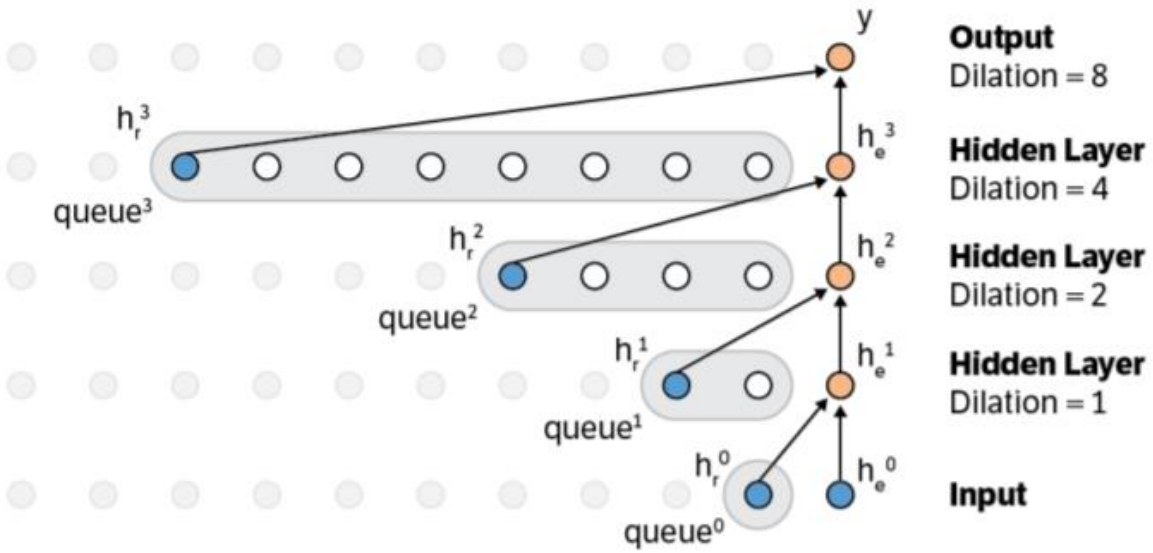
Ouroboros – On-Chip Loop Structure

Alibaba Group



Data Buffer – Caching Intermediate Results

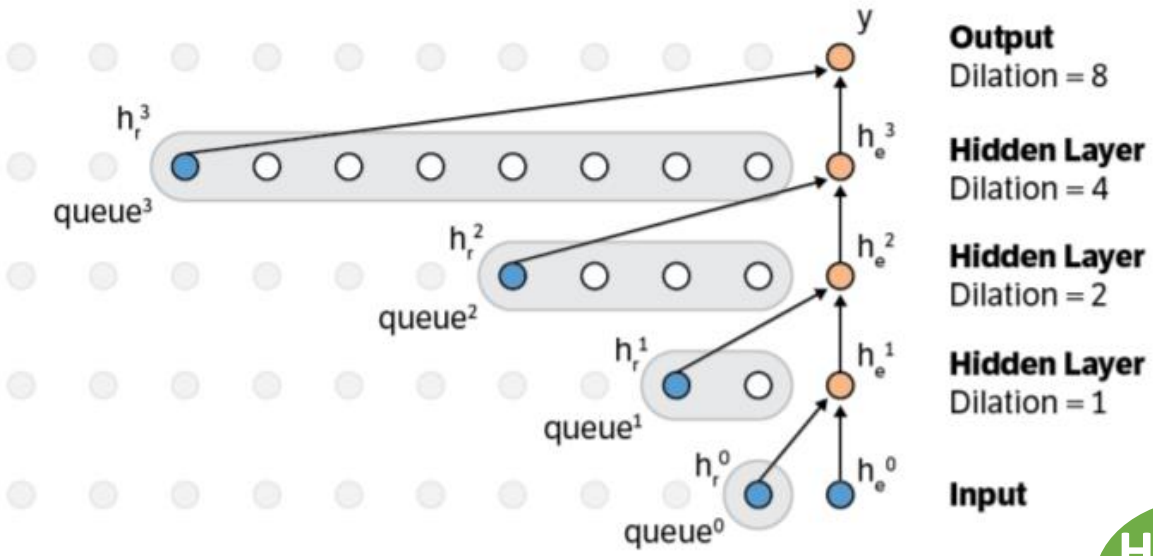
- Save Computation without Adding Buffer
  - In this example, the total queue size is the same as storing raw inputs
    - The size of the queue at the l^{th} hidden layer is 2^l

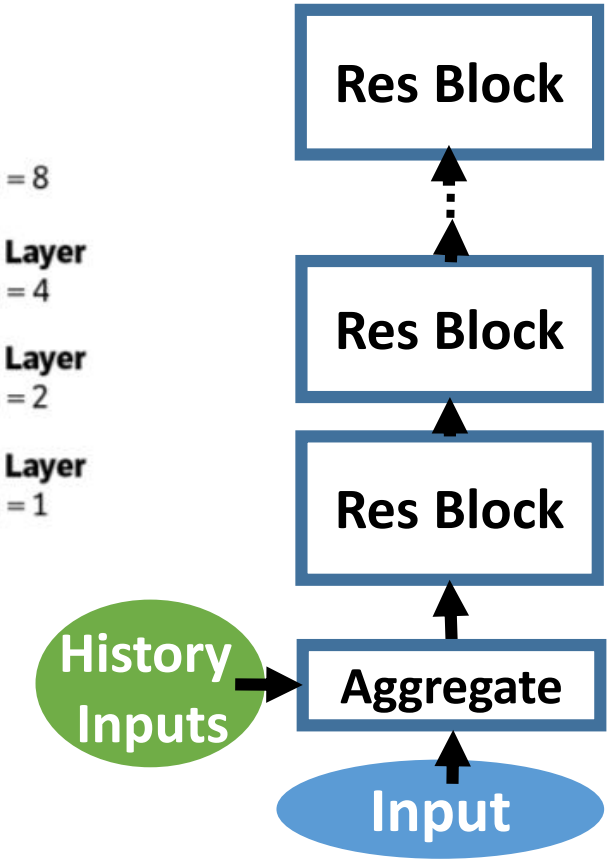


※ Inspired by Fast-WaveNet (the left figure): https://arxiv.org/abs/1611.09482

Data Buffer – Caching Intermediate Results

- Save Computation without Adding Buffer
  - In this example, the total queue size is the same as storing raw inputs
    - The size of the queue at the l^{th} hidden layer is 2^l

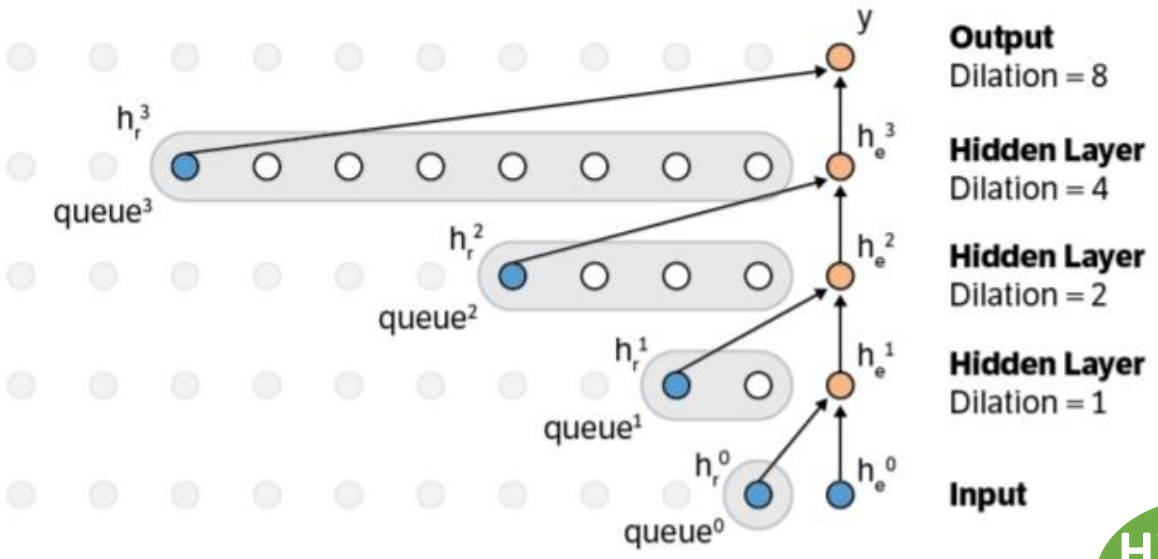


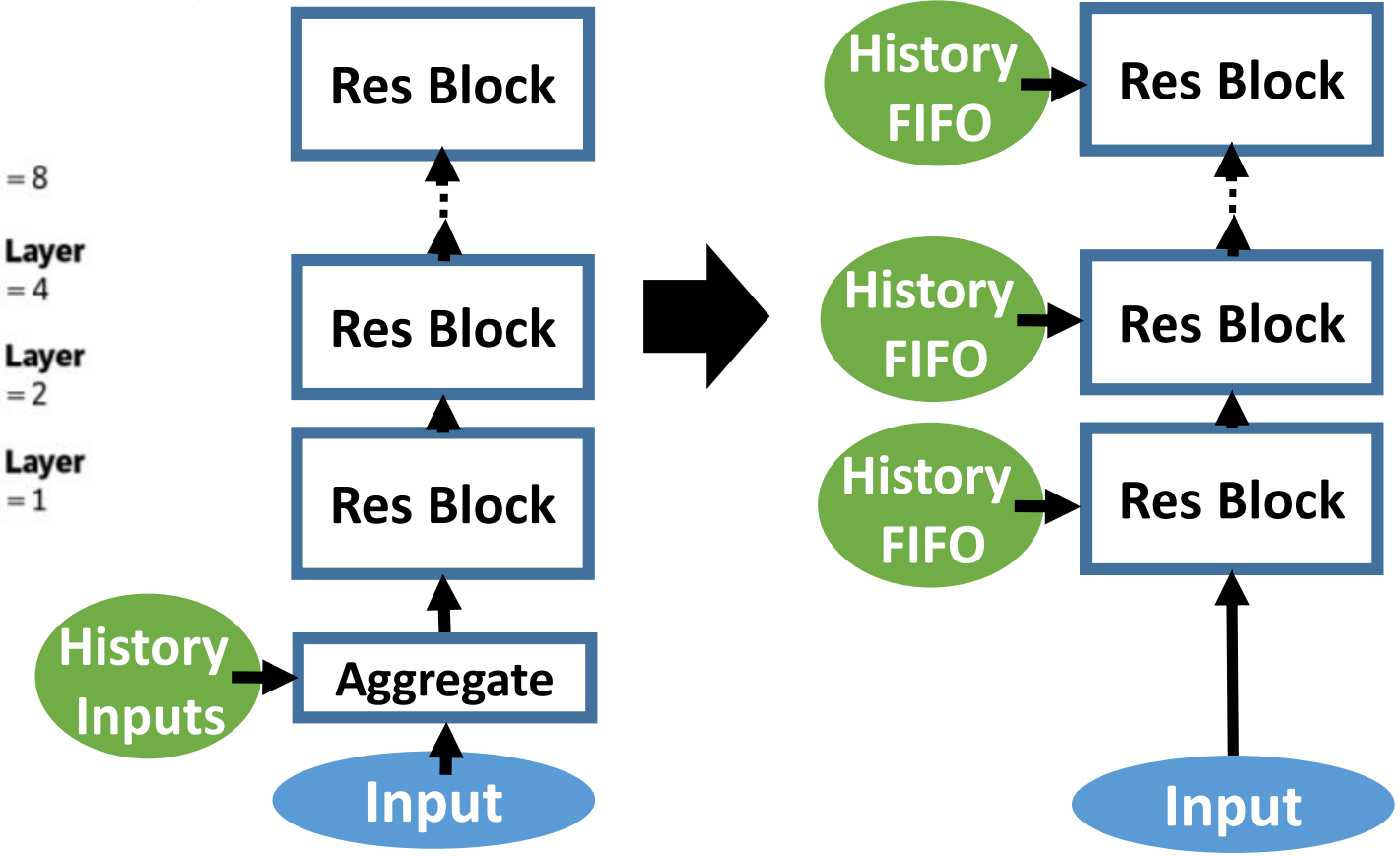


※ Inspired by Fast-WaveNet (the left figure): https://arxiv.org/abs/1611.09482

Data Buffer – Caching Intermediate Results

- Save Computation without Adding Buffer
  - In this example, the total queue size is the same as storing raw inputs
    - The size of the queue at the l^{th} hidden layer is 2^l

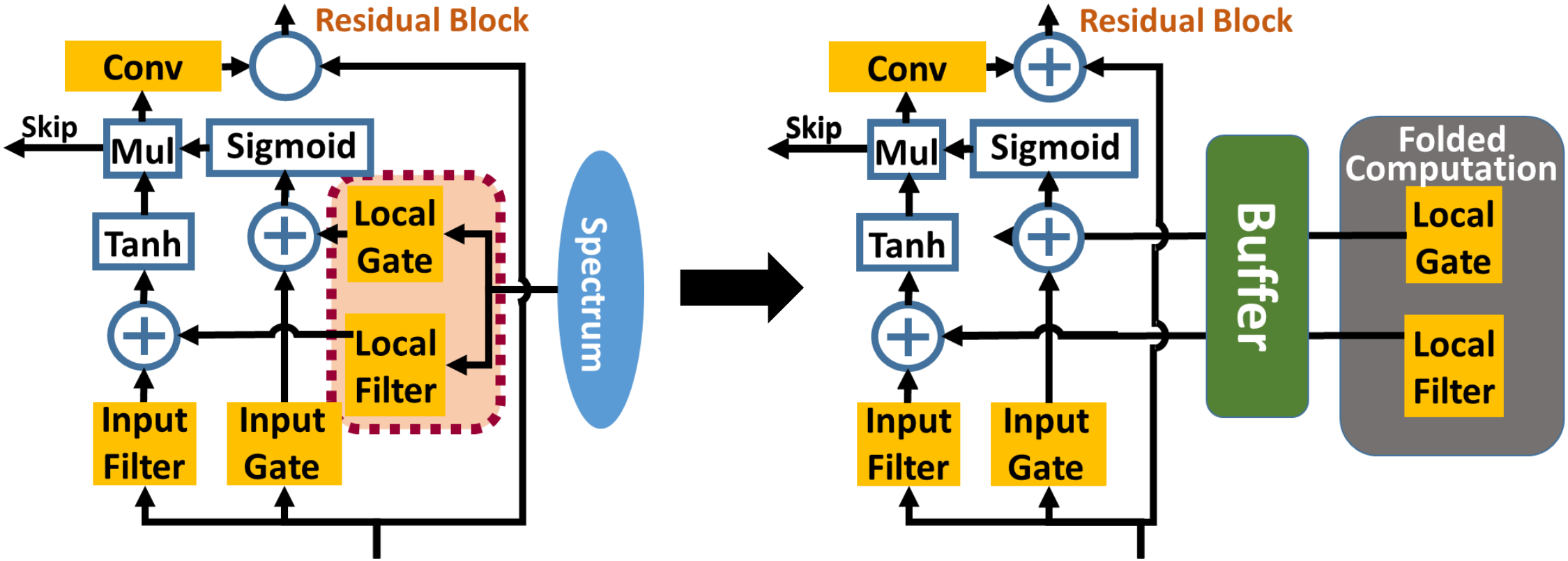




※ Inspired by Fast-WaveNet (the left figure): https://arxiv.org/abs/1611.09482

Data Buffer – Conditional Folding

- Mel Spectrogram usually updates every 10ms (160 samples)
- Add Small Buffer to Save Computation



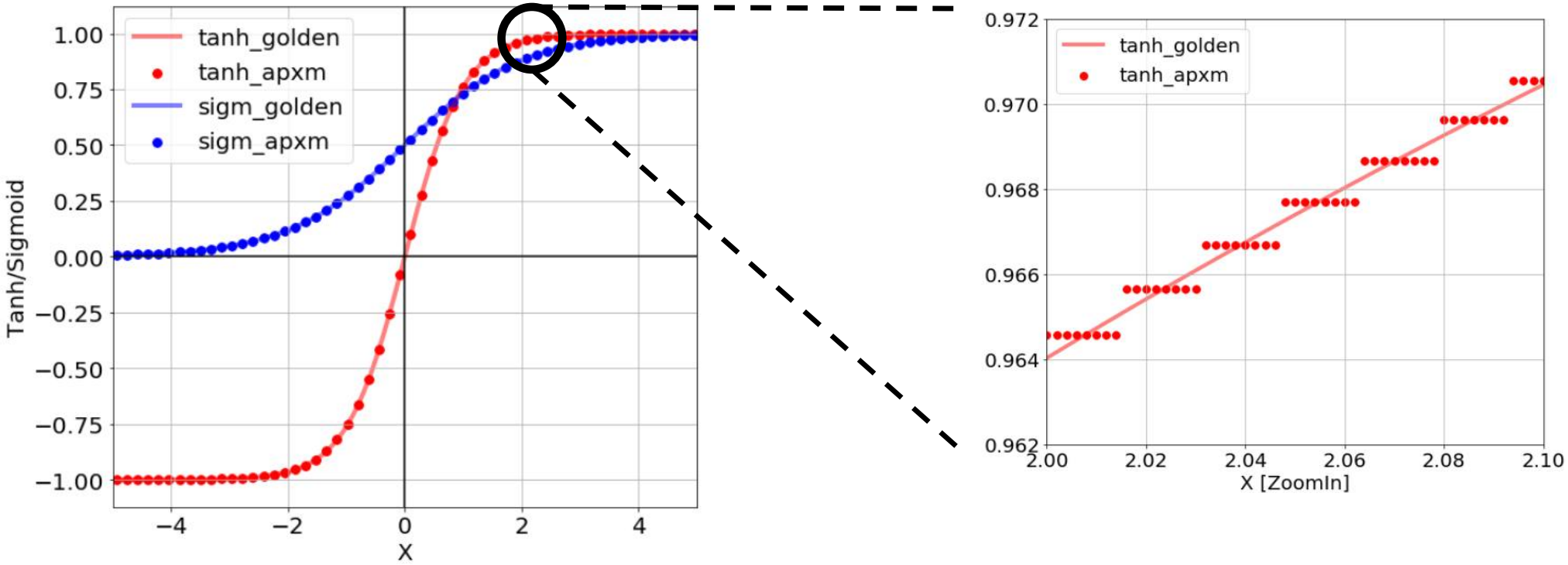
※ Illustrated using algorithm diagram

Operators – Activation Function Quantization

- Piecewise Look-Up Table + Piecewise Linear Function
  - < 1024 table items
  - < 0.001 quantization errors

Sigmoid: $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

Tanh: $\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}-1}{e^{2x}+1}$



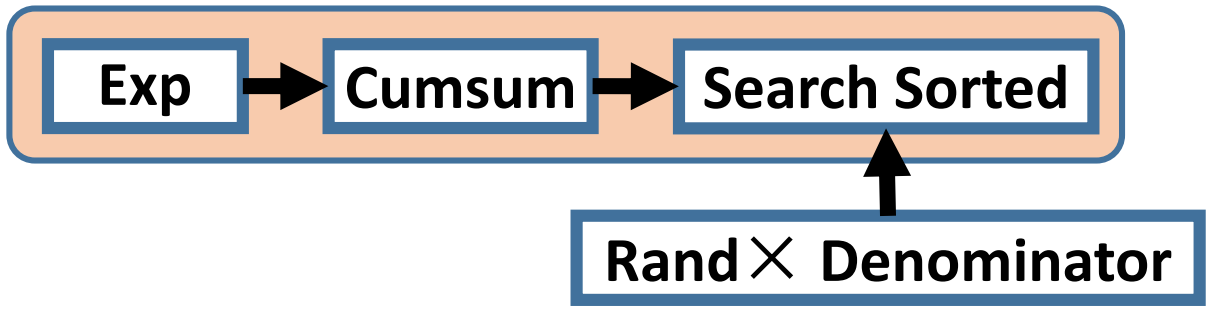
Operators – Softmax + Random Choice

- Operator Basics
  - Softmax – normalize the tensor into a probability distribution

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$$

  - Random Choice – randomly choose the output according to the probability distribution

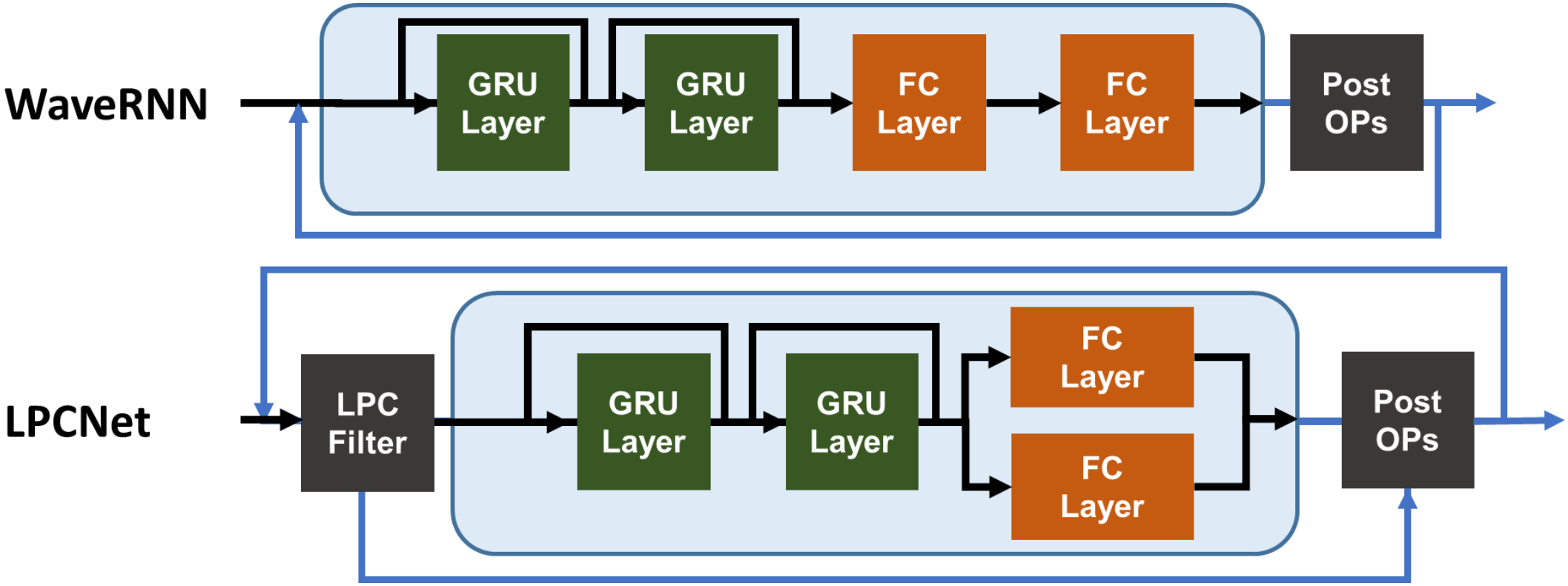
- Optimizations
  - Division to multiplication – Remove division of Softmax; multiply the denominator to Rand
  - Operator fusion – ***Exp + CumSum + Search Sorted*** with random value generator



Versatility Beyond WaveNet Configurations

Alibaba Group

- Other Deep Learning Based Vocoders
  - Using RNN to replace 1D-Conv, e.g., WaveRNN[1], LPCNet[2], etc.



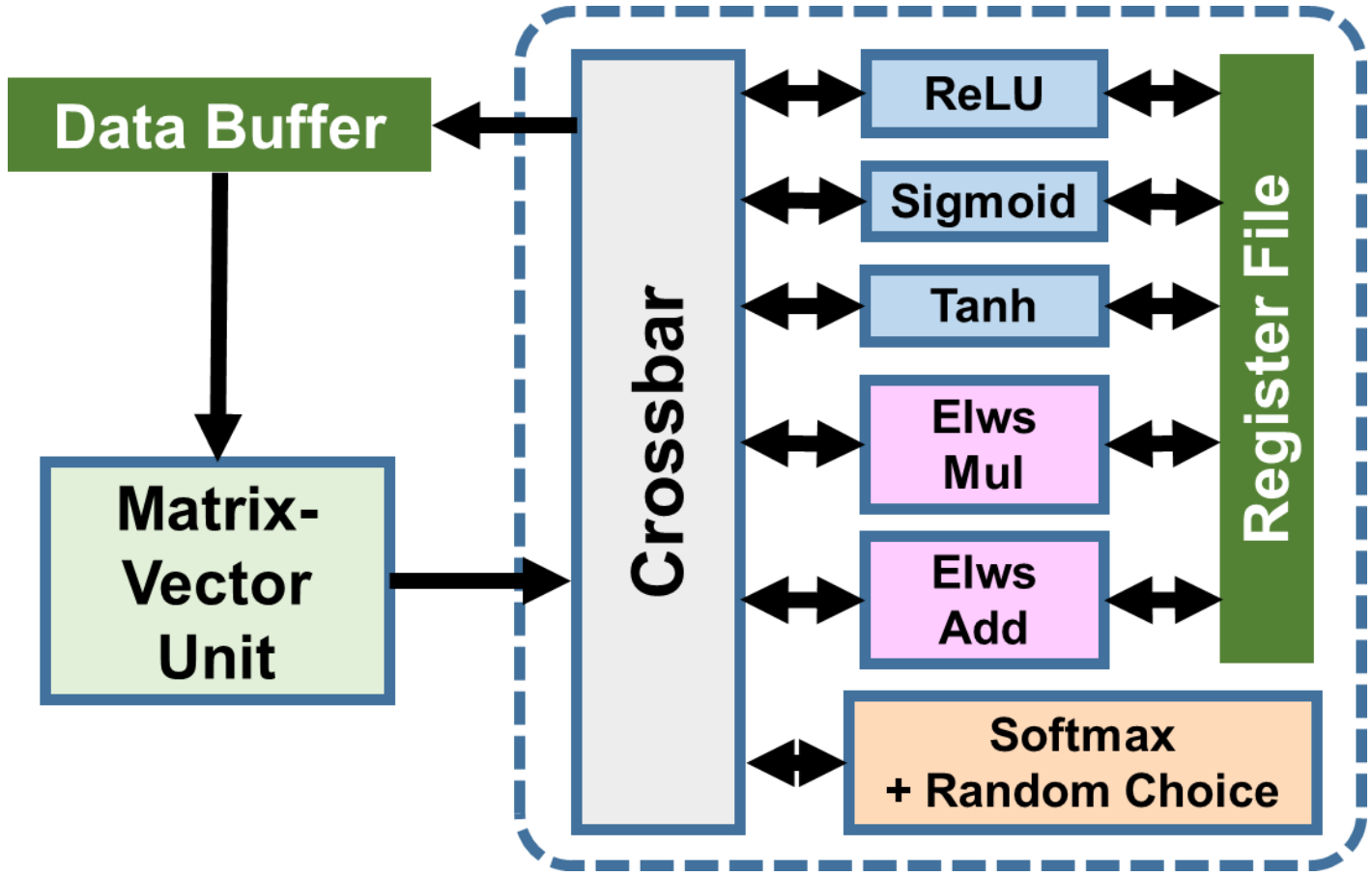
[1] WaveRNN: https://arxiv.org/abs/1802.08435; [2] LPCNet: https://arxiv.org/abs/1810.11846

Versatility Beyond WaveNet Configurations

Alibaba Group

- Unified Architecture
  - Matrix-Vector Unit
  - Flexible Post Module

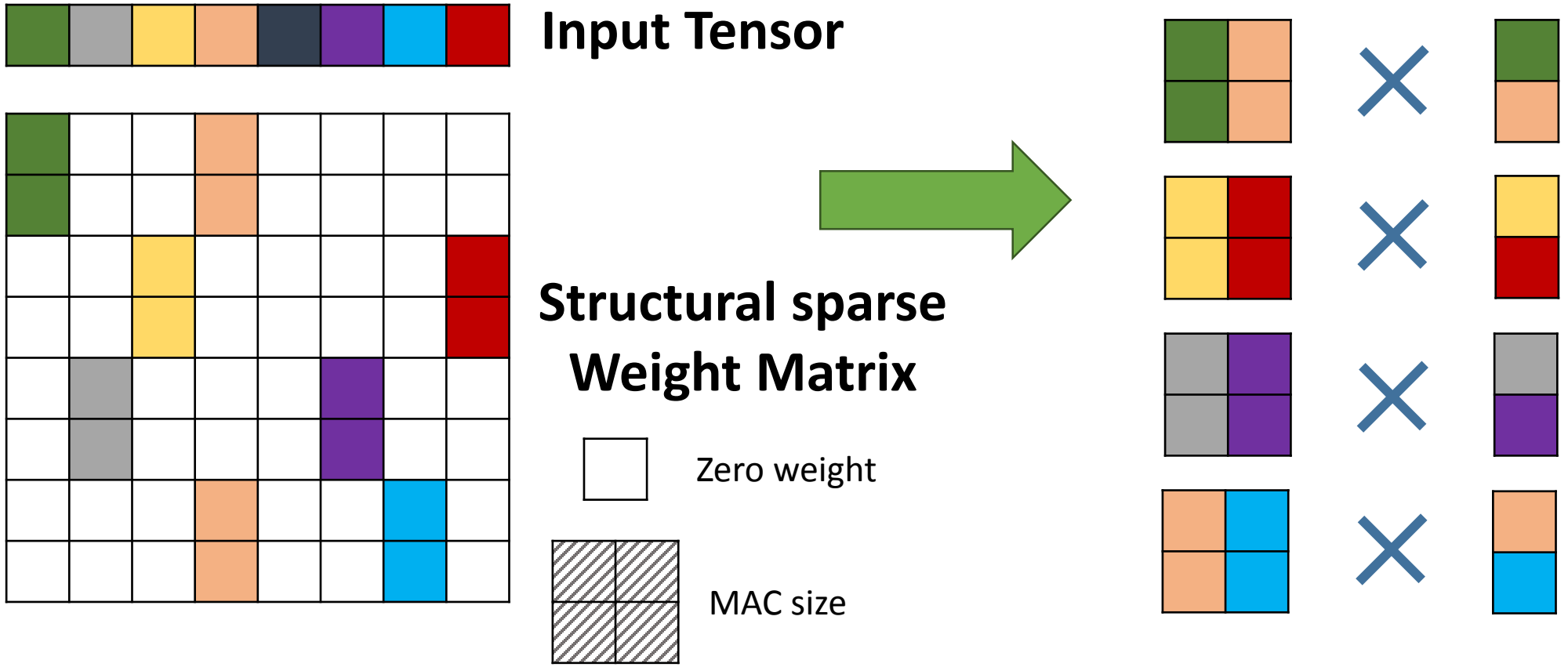
Fortunately, WaveRNN and LPCNet share the same basic operators with WaveNet



Versatility Beyond WaveNet Configurations

Alibaba Group

- Support Sparse Weight Matrix in WaveRNN and LPCNet
  - Divide and compress structural sparse weights to match the MAC array size



Performance – WaveNet (Tacotron 2)

Alibaba Group

Platform	CPU	GPU	FPGA		ASIC (16nm, simulation)		
Frequency (MHz)	3500	1506	300	300	800	800	300
MAC Precision	FP32	FP32	FP32	INT16	FP16	INT16	INT16
MAC Size	-	-	2048	2048	2048	2048	256
Real-Time Factor	155.2	49.4	0.97	0.33	0.05		0.47
GOP/s	0.5	1.6	81.3	242	1649		170
Power (W)	-	39	48	35	3.4	3.2	0.2
GOP/s/W	-	0.04	1.7	6.9	485	515	849

Performance – WaveNet Configurations

Alibaba Group

Model	WaveNet in Tacotron 2[1]				Basic WaveNet[2]			
Size of Residual Block	128 × 128 × 2				32 × 32 × 2			
GOPs (1s audio)	79.2				55.6			
Platform	GPU	FPGA	ASIC	ASIC-s[3]	GPU	FPGA	ASIC	ASIC-s
Real-time Factor	49.4	0.33	0.05	0.47	142	0.68	0.12	0.47
Speed-up over {GPU+Tensorflow}	-	151×	1000×	100×	-	210×	1200×	300×

[1] Tacotron 2: https://arxiv.org/abs/1712.05884;
[2] Basic WaveNet: https://arxiv.org/abs/1609.03499;
[3] ASIC-s means the smaller chip: *frequency=300MHz; MAC size=256*

Performance – LPCNet

Alibaba Group

Model	LPCNet Dense			LPCNet Sparse		
GOPs (1s- audio)	16.4			3.62		
Platform	CPU	FPGA	ASIC	CPU	FPGA	ASIC
Real-time Factor	5.86	0.15	0.107	1.30	0.08	0.028
Speed-up over Desktop CPU	-	38×	54×	-	16×	47×

Conclusion

- A Case of Domain-Specific Computing
  - Existing {framework + CPU/GPU} is inefficient to support deep learning based TTS
  - We propose a hardware accelerator to efficiently support deep learning based TTS
    - Design idea – on-chip loop structure for iterative algorithm
    - Benefits – no additional model training overhead; versatility