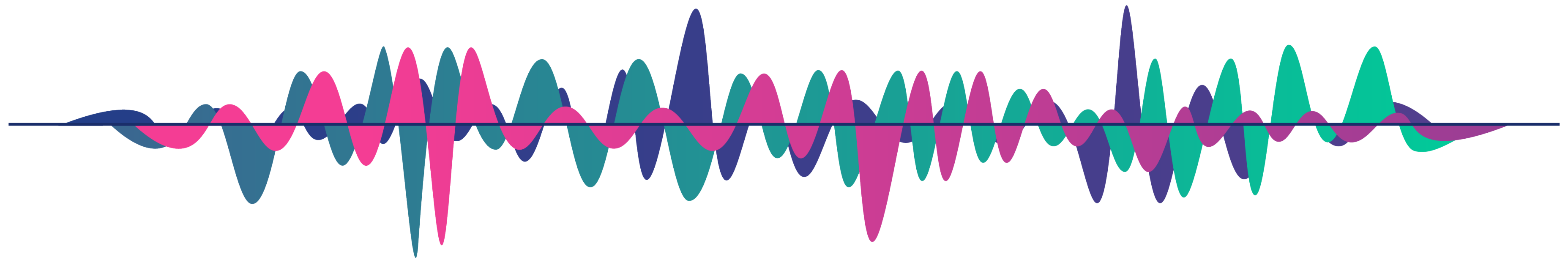
CSB-RNN: A Faster-than-Realtime RNN Acceleration Framework with **Compressed Structured Blocks**

Runbin Shi[1,3], Peiyan Dong[1], Tong Geng[2], Martin Herbordt[2], Hayden So[3], Yanzhi Wang[1]
[1]Northeastern University, [2]Boston University, [3]The University of Hong Kong

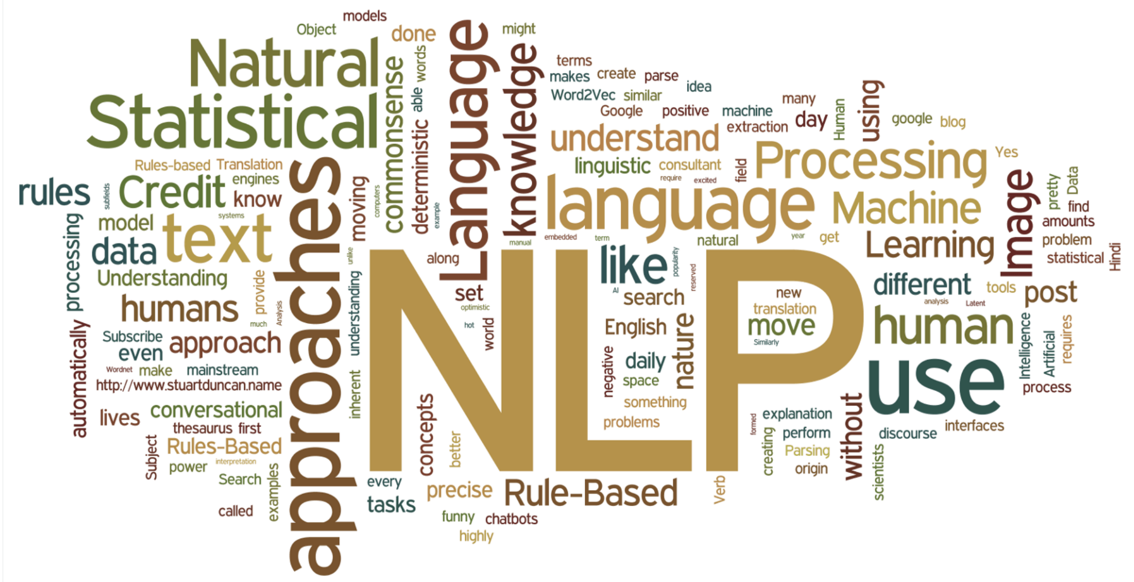
BARC'20, Jan. 31, 2020, Boston

Recurrent Neural Networks for Sequence Processing



✓ Speech Recognition

✓ Stock Price Prediction



✓ Natural Language Processing
(e.g., translation, sentiment classification)

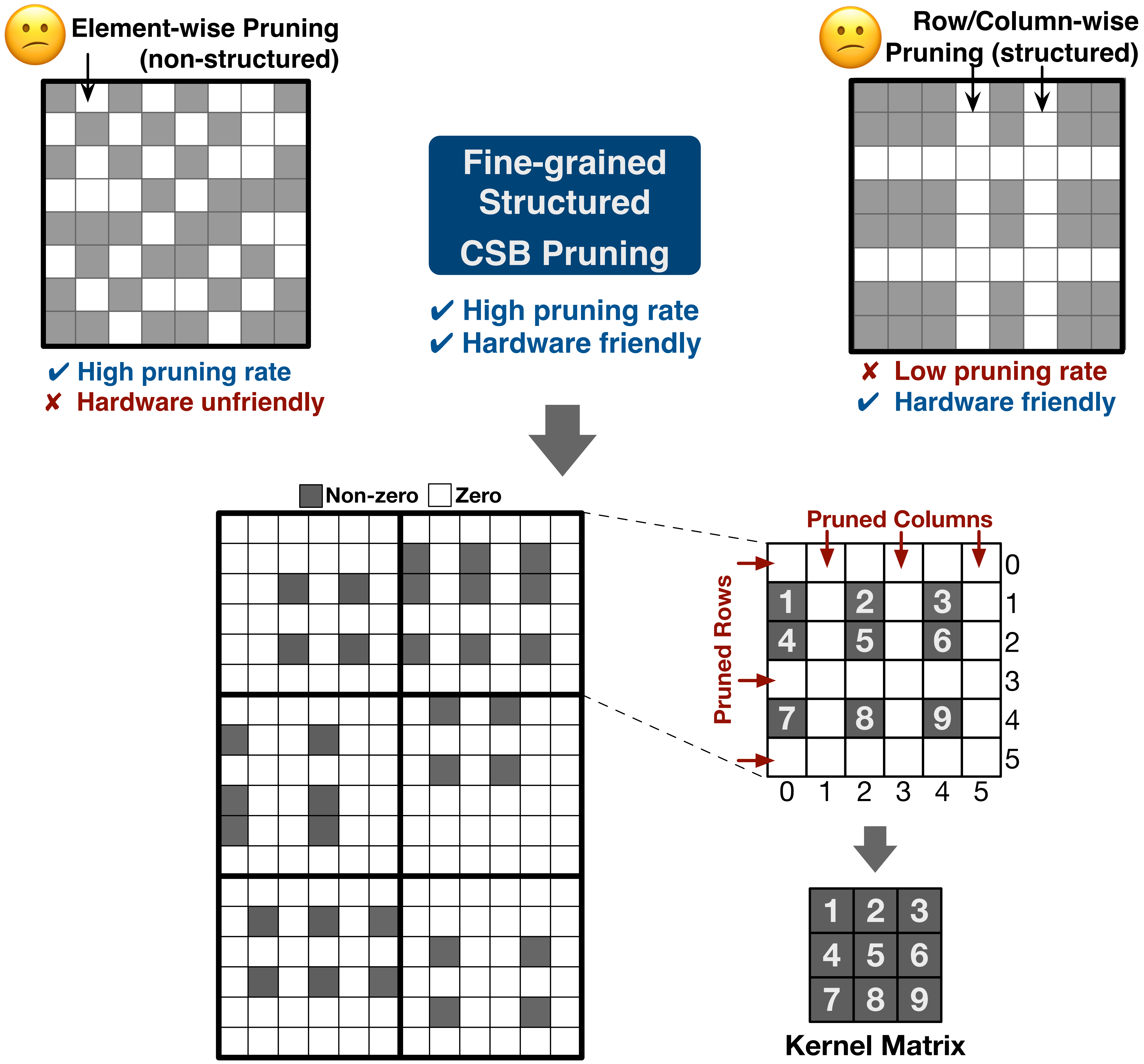
Real World Issues 🤔

- **Tremendous MVM Computation** (10**ms** latency on CPU *v.s.* 500**μs** realtime requirement)
- **Sequential Processing** (low concurrency)
- **Large Model Size** (30M+ Weights)

Techniques in CSB-RNN

- **Roadmap: Algorithm** + **Acceleration** Co-optimization
- **Algorithm side: Pruning** with a novel **structured sparse pattern, CSB**
- **Acceleration side:** ISA **Architecture** with **Compilation**

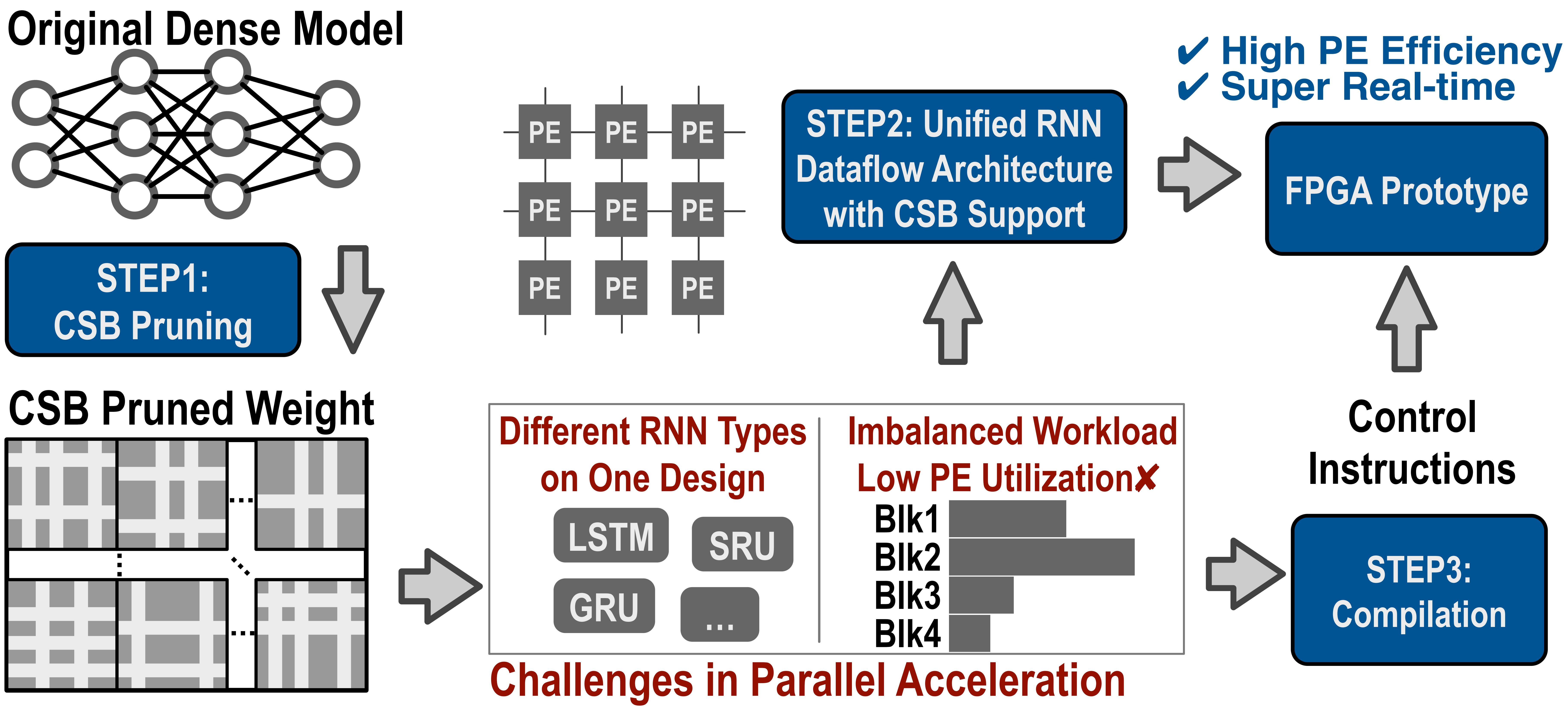
Fine-grained Structured Pruning



Advantages of CSB 🤩

- **Structure:** friendly to parallel hardware
- **Fine-grained pruning block:** (high pruning rate)

Acceleration Framework



Performance

- **Pruning rate: 3.5×-25×** Lossless pruning, **1.3×-2.8×** to existing works.
- **Acceleration:** 0.12μs - 4.79μs inference latency, **2.6×-38.66×** more power efficient.

Table 1: Comparison of Pruning Rate

Abbr.	Compression Technique	Prune Rate	Weight Width	Metric	Result	Impro-vement
MT1	column pruning [25]	8×	16-bit	PPL	112.73	1×
	CSB pruning	**12.5×**	**16-bit**		**112.02**	**1.6×**
MT2	row-column [26]	4.35×	floating	PPL	82.59	1×
	bank-balanced [3]	5×	16-bit		82.59	1.1×
	CSB pruning	**12×**	**16-bit**		**82.33**	**2.8×**
SR1	block-circulant [24]	8×	16-bit	PER	24.57%	1×
	row-balanced [10]	8.9×	16-bit		20.70%	1.1×
	bank-balanced [3]	10×	16-bit		23.50%	1.3×
	CSB pruning	**13×**	**16-bit**		**20.10%**	**1.6×**
SR2	block-circulant [14]	8×	16-bit	PER	20.02%	1×
	CSB pruning	**20×**	**16-bit**		**20.01%**	**2.5×**
SR4	column pruning [7]	14.28×	16-bit	Accu	98.43%	1×
	CSB pruning	**23×**	**16-bit**		**99.01%**	**1.61×**

Table 2: Comparison of Inference Latency and Power Efficiency

Abbr.	Work	#PE	Freq. (MHz)	Latency (μs)	Power (Watt)	Power Eff. (k-frames/W)	Power Eff. Improv.
MT1	BBS [3]	1518	200	1.30	19	40.49	1×
	CSB-RNN	**512**	**200**	**0.79**	**8.9**	**142.72**	**3.53×**
SR1	C-LSTM [24]	2680	200	8.10	23	5.37	18.20×
	E-RNN [14]	2660	200	7.40	29	4.66	15.80×
	ESE [10]	1504	200	82.70	41	0.29	1×
	CSB-RNN	**1024**	**200**	**4.79**	**18.3**	**11.40**	**38.66×**
SR2	E-RNN [14]	2280	200	6.70	25	5.97	1×
	CSB-RNN	**1024**	**200**	**2.60**	**18.3**	**20.98**	**3.51×**
SR4	DeltaRNN [7]	768	125	0.38	7.3	360.49	1×
	CSB-RNN	**512**	**200**	**0.12**	**8.9**	**936.74**	**2.60×**

Thank you.