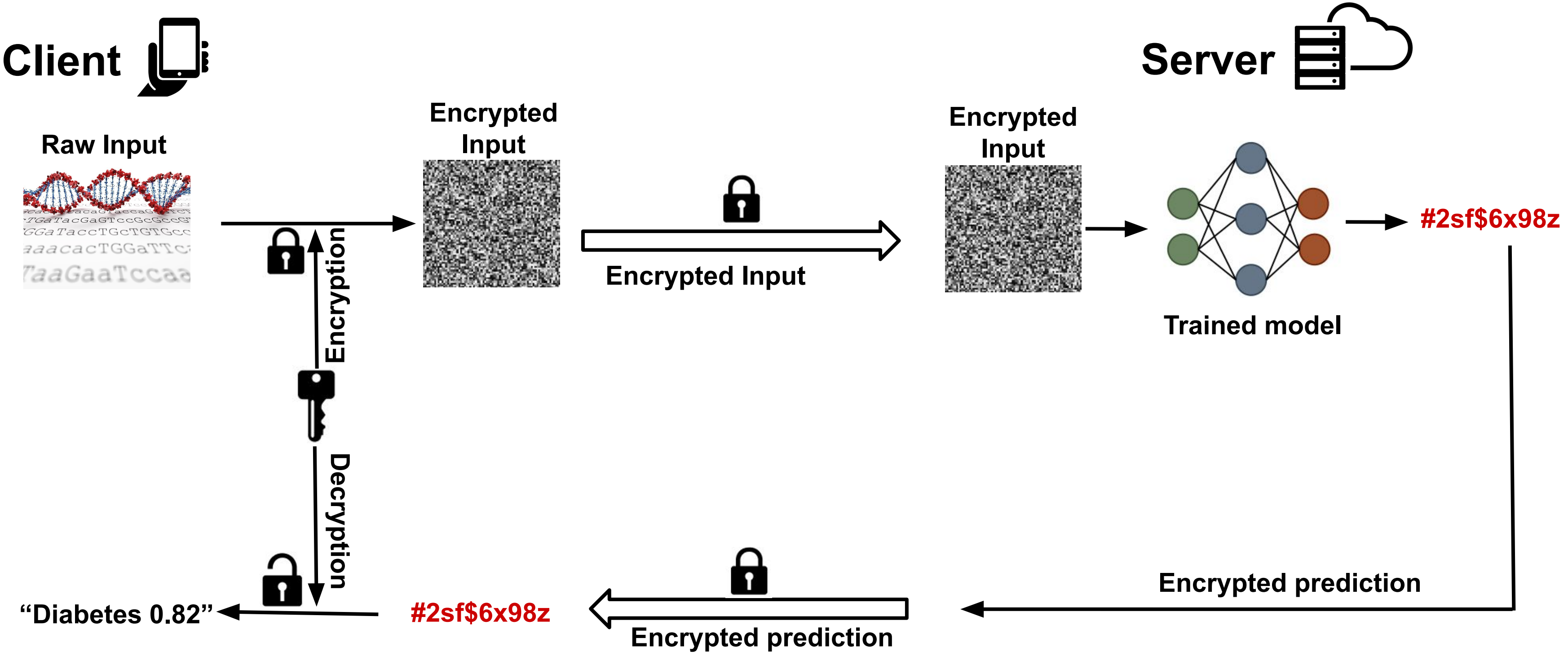
DeepReShape: Redesigning Neural Networks for Efficient Private Inference

Nandan Kumar Jha, Brandon Reagen

New York University

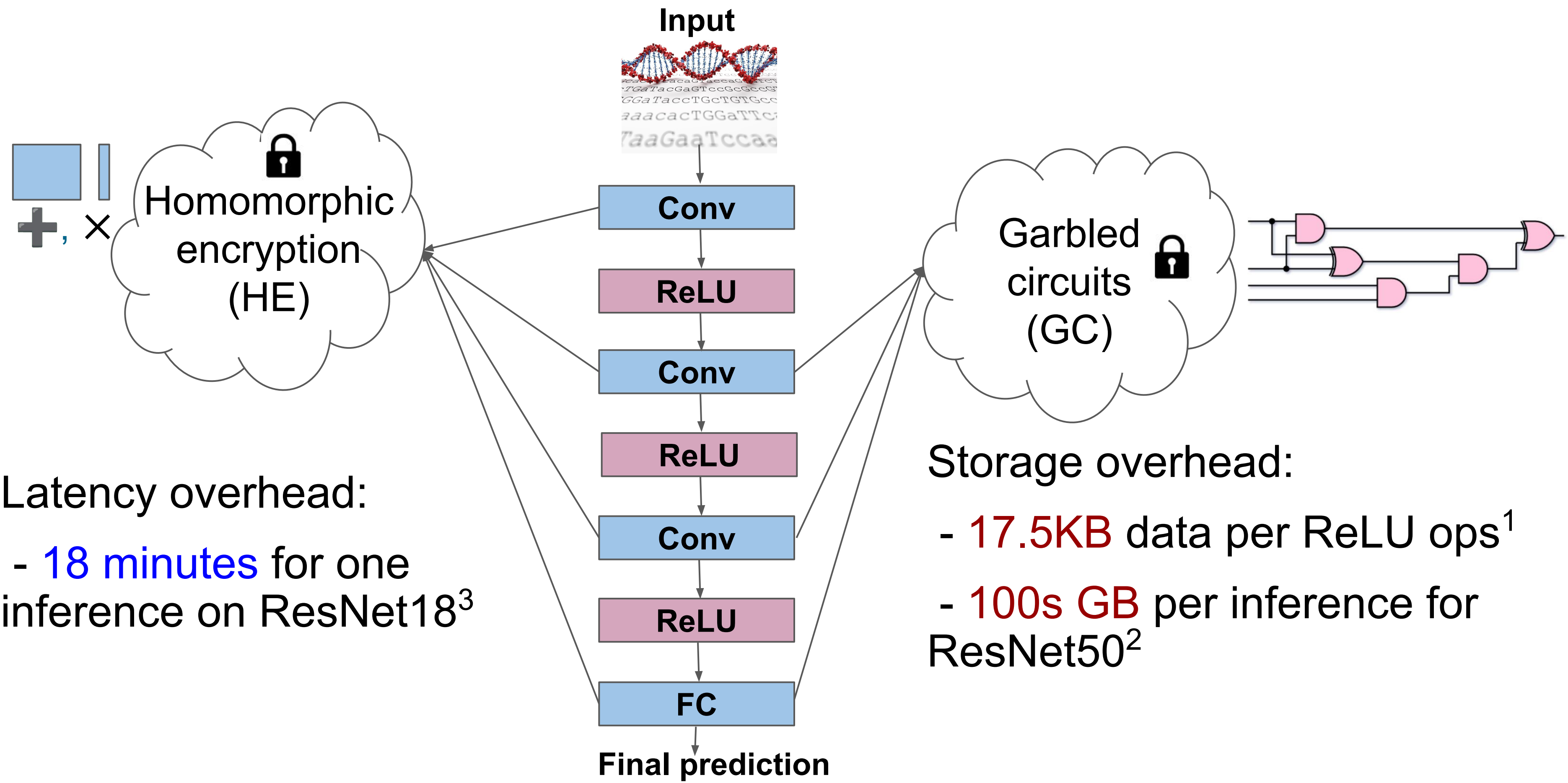
TMLR'24

Private Inference (PI)



Client's input privacy is preserved, and the server's model is protected

Overheads of Private Inference



Latency overhead:

- 18 minutes for one inference on ResNet18[3]

Storage overhead:

- 17.5KB data per ReLU ops[1]
- 100s GB per inference for ResNet50[2]

3. Garimella et al., Characterizing and optimizing end-to-end systems for private inference, ASPLOS'23

1. Mishra et al., Delphi: A cryptographic inference service for neural networks, USENIX Security'20
2. Rathee et al., CrypTFlow2: Practical 2-party secure inference, ACM CCS'20

The Era of Offline-online Phases

Prior cryptographic frameworks for PI used hybrid protocols, splitting evaluation into offline and online phases[1]

Offline phase:

- Input-independent tasks
- Compute-heavy HE tasks

Online phase:

- Input-dependent tasks
- Linear layer evaluation: Near-plaintext latency using additive secret sharing
- **99%** of the online cost stems from ReLUs[2]

1. Mishra et al., Delphi: A cryptographic inference service for neural networks, USENIX Security'20

2. Lou et al., SAFENet: A Secure, Accurate and Fast Neural Network Inference, ICLR'21

Fallacies of Offline-online Phases

Single PI Isolation:

- Assumed FLOPs are free
- Primarily optimized for ReLU efficiency

Multiple Requests Impact:

- Time gap between consecutive client requests matters
- Network complexity further worsen this impact

Implications:

- FLOPs do carry significant penalties for e2e performance
- Offline cost starts affecting the real-time performance

Challenges in Simultaneous Optimization of ReLU and FLOPs

Layer-Specific Distribution:

- ReLUs are concentrated in early layers
- Critical ReLUs for network accuracy are in deeper layers

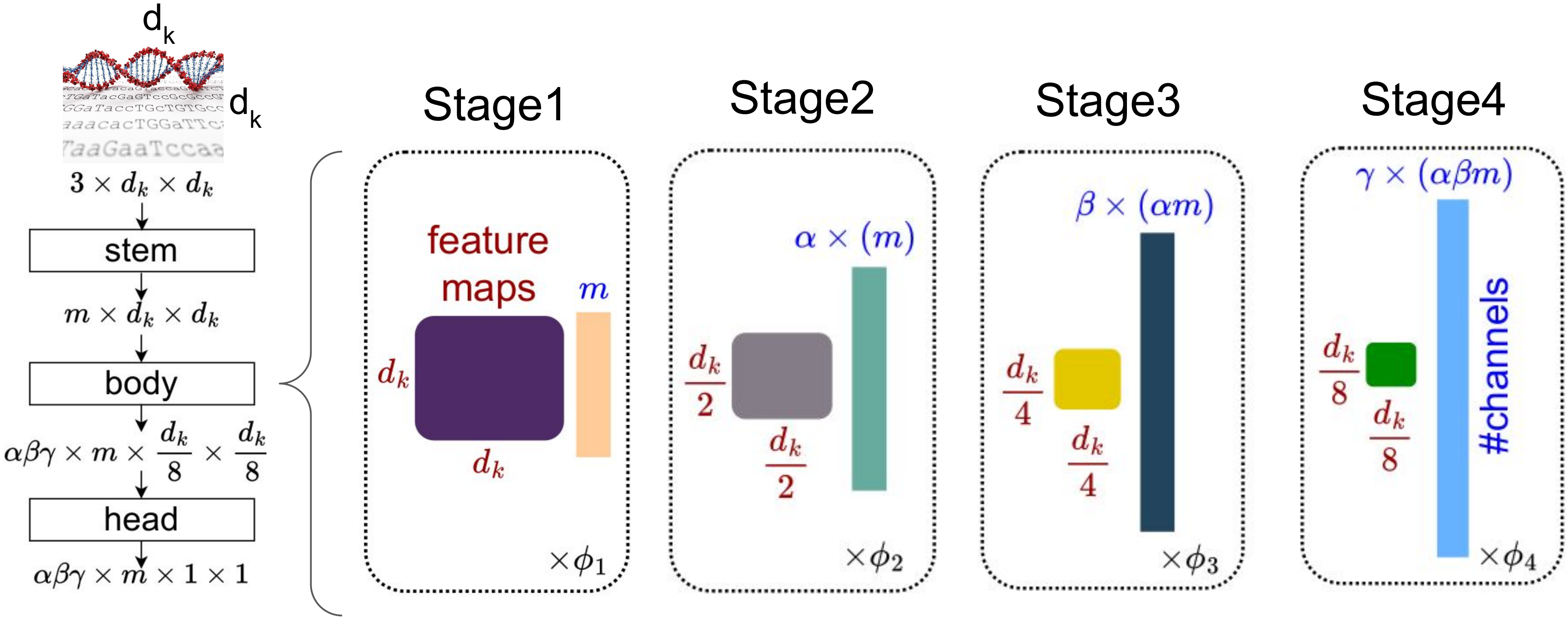
Pruning Conflicts:

- ReLU pruning often removes ReLUs from early layers
- FLOPs pruning targets deeper layers due to higher channel counts

Design Conflicts:

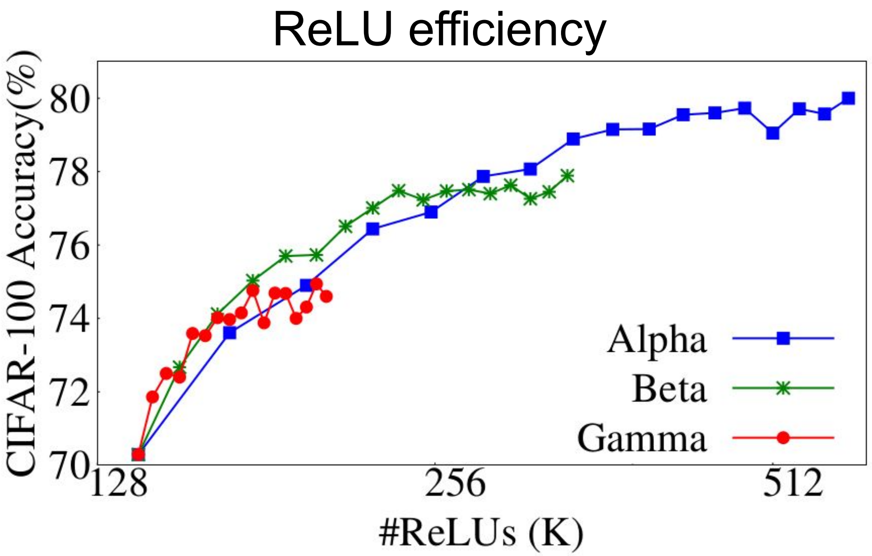
- ReLU-efficient networks require different hyper-parameters than FLOPs-efficient networks

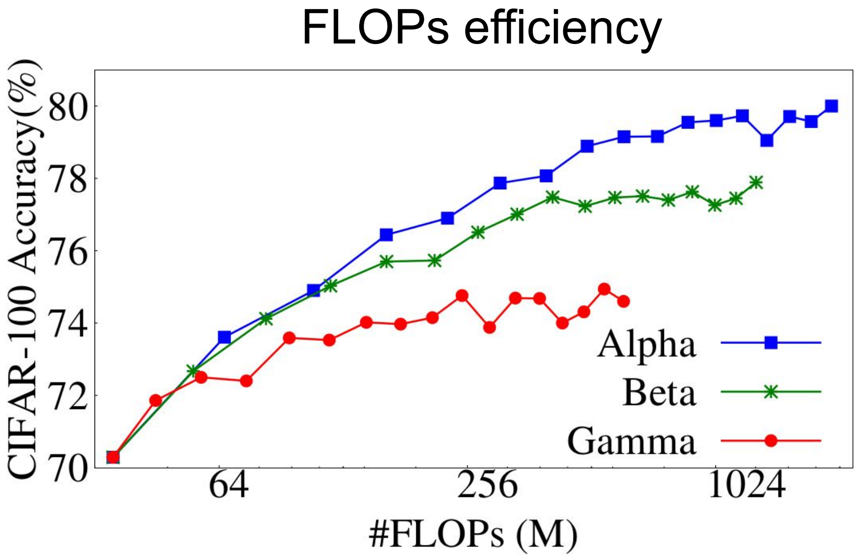
Network Design Hyper-parameters



For ResNet18, $\alpha = \beta = \gamma = 2$ and $\phi_1 = \phi_2 = \phi_3 = \phi_4 = 2$

Desirable Network Attributes for ReLU and FLOPs Efficiency





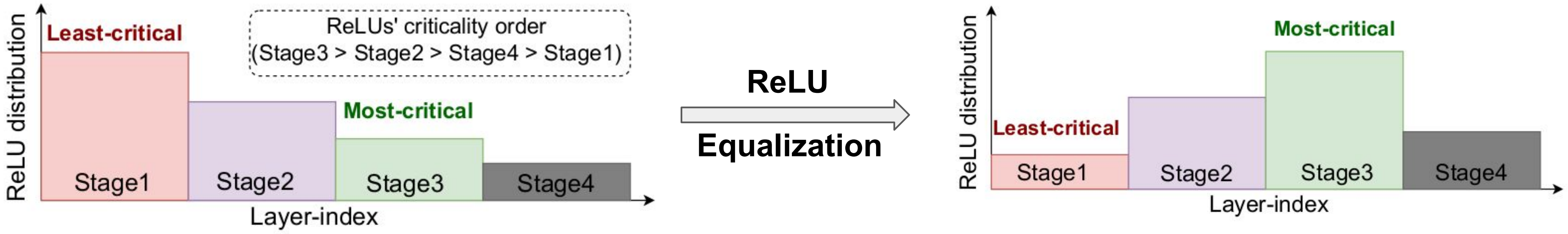
Not all stages *equally* affect ReLU and FLOPs efficiency of the network!

Achieving the right balance requires higher alpha and beta values, and a lower gamma value.

How can we design a network that balances ReLU and FLOPs efficiency under PI constraints?

Our Solution: ReLU Equalization

ReLU Equalization



Baseline Network

ReLU-equalized Network

ReLUs are redistributed based on their criticality
by adjusting network design parameters.

Design of PI-Efficient HybReNets Networks

$$\#ReLUs(S_3) > \#ReLUs(S_2) > \#ReLUs(S_4) > \#ReLUs(S_1)$$

$$\phi_3\left(\frac{\alpha\beta}{16}\right) > \phi_2\left(\frac{\alpha}{4}\right) > \phi_4\left(\frac{\alpha\beta\gamma}{64}\right) > \phi_1$$

$$\alpha\beta > 16,\ \alpha > 4,\ \alpha\beta\gamma > 64,\ \beta > 4,\ \beta\gamma < 16,\ \text{and}\ \gamma < 4$$

$$(5,2)\ \&\ \alpha \geq 7;\ (5,3)\ \&\ \alpha \geq 5;\ (6,2)\ \&\ \alpha \geq 6;\ (7,2)\ \&\ \alpha \geq 5$$

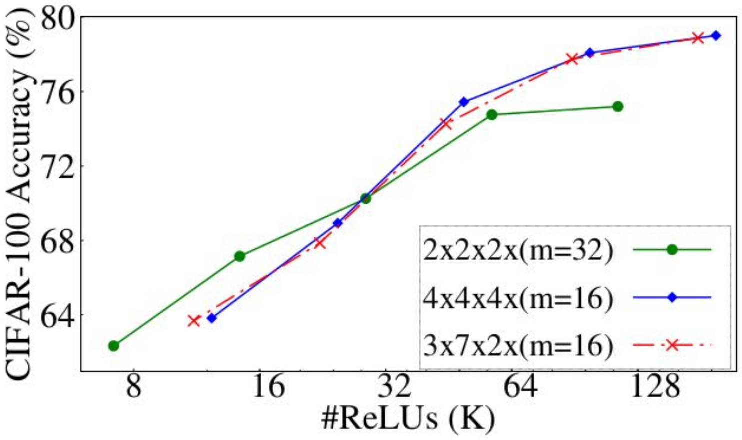
Bound on γ prevents excessive FLOPs in deeper layers while *maintaining* ReLU efficiency

HRN-7x5x2x
HRN-5x5x3x
HRN-6x6x2x
HRN-5x7x2x

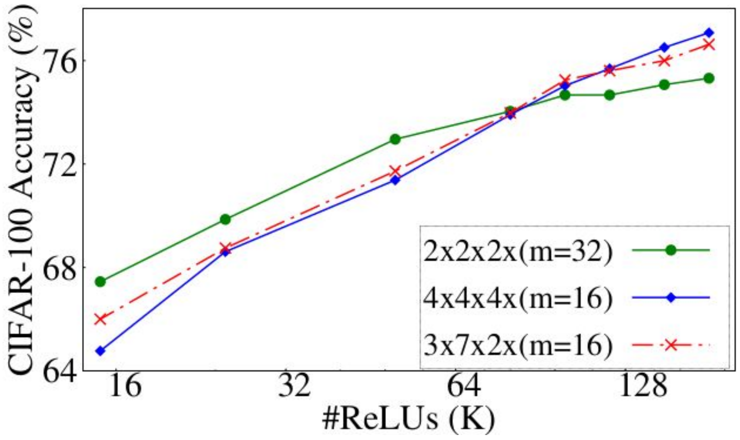
Can one baseline network excel across all ReLU counts, when using ReLU optimization techniques?

Impact of Baseline Network on ReLU Optimization

Coarse-grained optimization



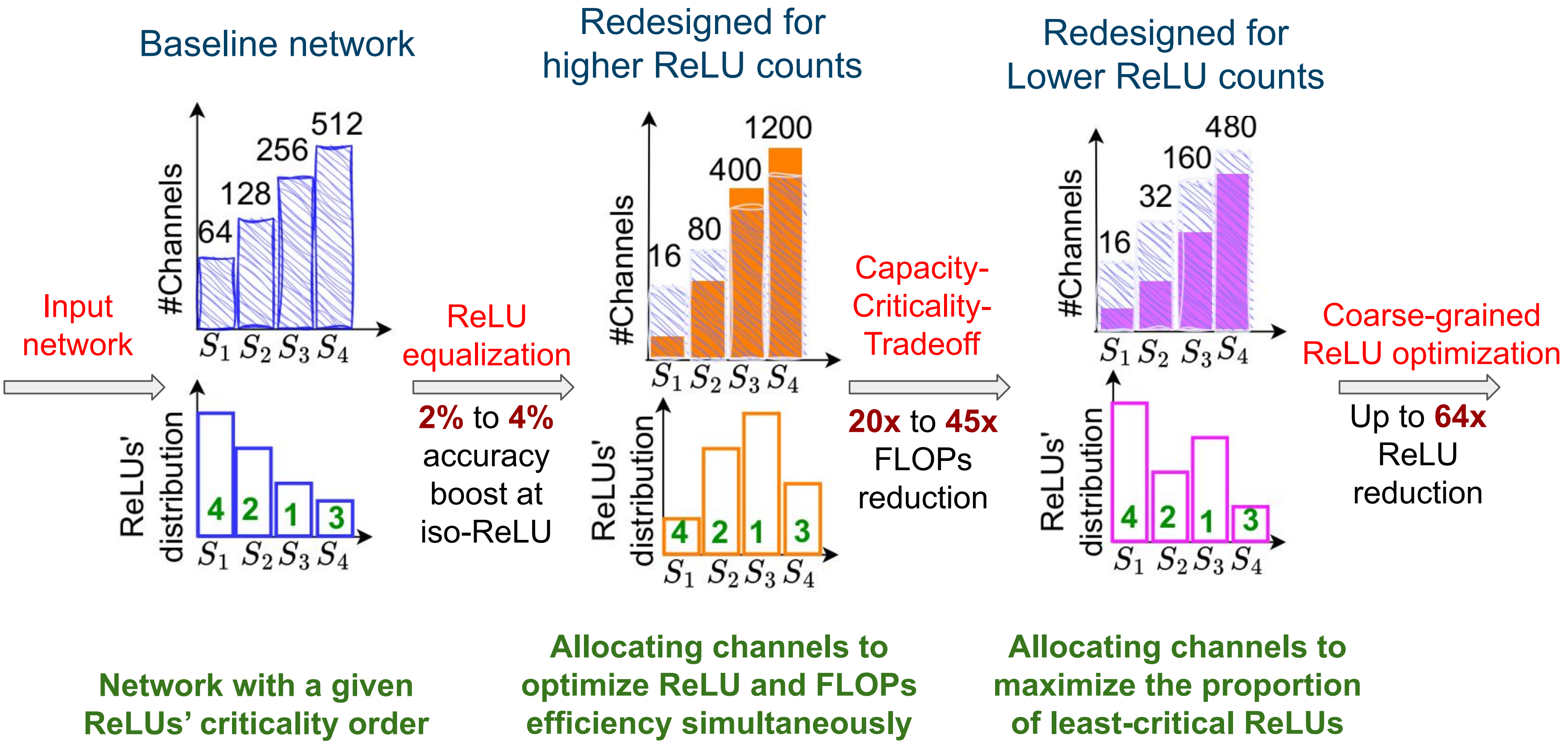
Fine-grained optimization



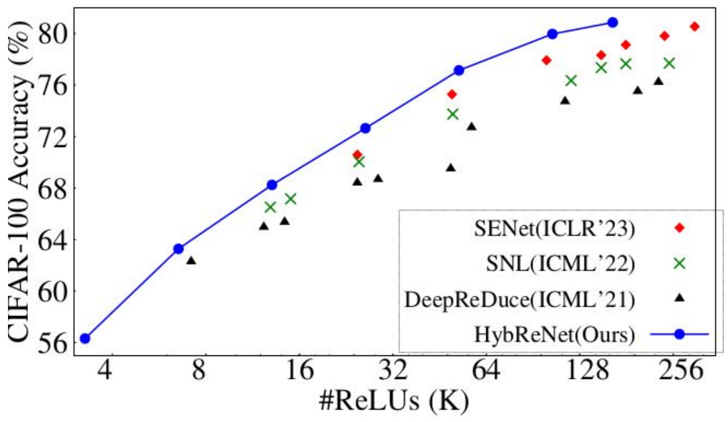
Model	Acc(%)	FLOPs	ReLUs	Stagewise ReLUs' distribution			
				Stage1	Stage2	Stage3	Stage4
2x2x2x(m=32)	75.60	141M	279K	58.82%	23.53%	11.76%	5.88%
4x4x4x(m=16)	78.16	661M	279K	29.41%	23.53%	23.53%	23.53%
3x7x2x(m=16)	78.02	466M	260K	31.50%	18.90%	33.07%	16.54%

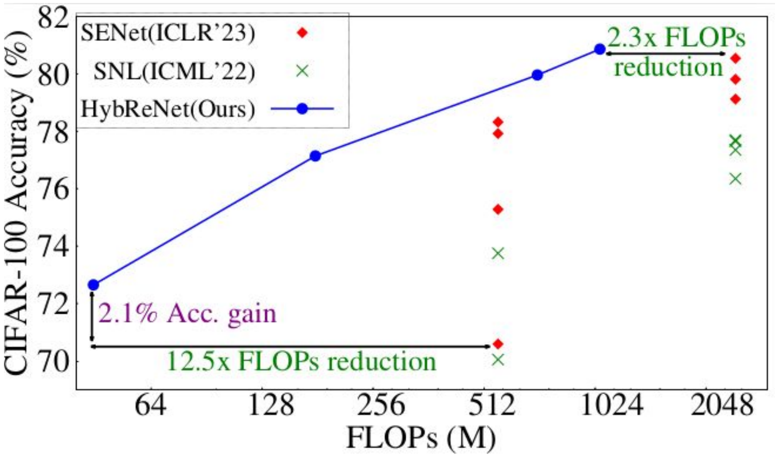
Capacity
Criticality
Tradeoff

DeepReShape



HybReNets Outperform SOTA in Private Inference



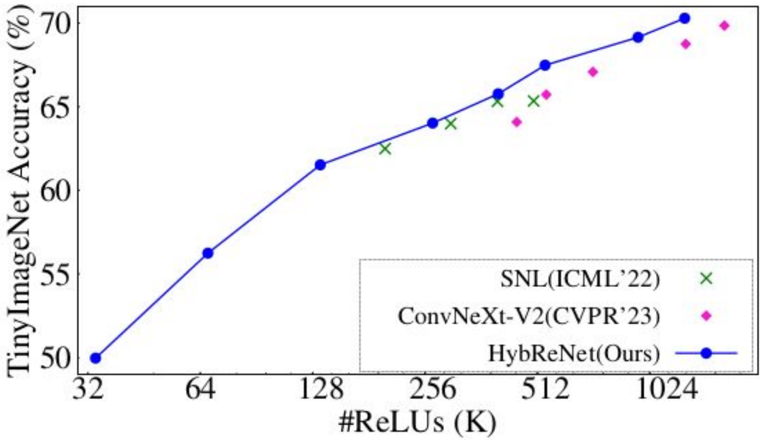


Iso-accuracy improvement:
2.3x ReLU savings and **3.4x** FLOPs reduction

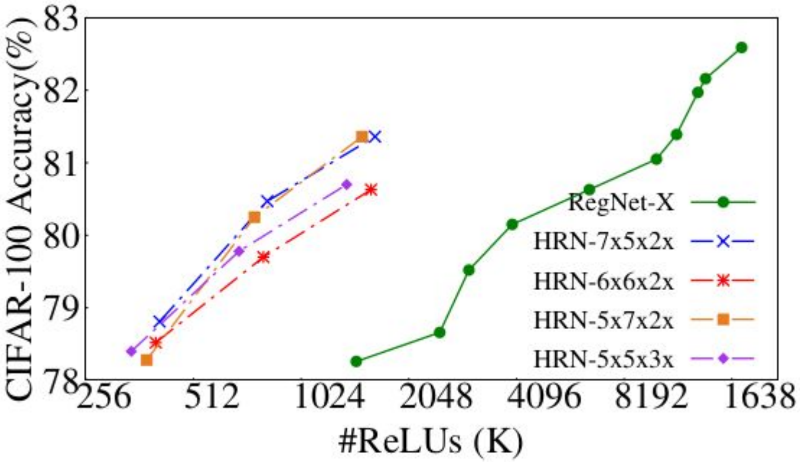
Iso-ReLU improvement:
2.1% accuracy gain and **12.5x** FLOPs reduction

HybReNets Outperform SOTA FLOPs-efficient Networks

ResNet34-based HRNs
vs ConvNeXt-V2



HRNs vs. RegNet-x



SOTA FLOPs-efficient networks exhibits inferior ReLU efficiency

DeepReShape shows generality beyond ResNet18

Key Takeaways from DeepReShape

1. Heterogeneous channel scaling is required to balance ReLU and FLOPs efficiency under PI constraints.

2. ReLU equalization positions ReLUs in their criticality order to prevent excessive FLOPs in deeper layers while maintaining ReLU efficiency.

3. Wider networks outperform at higher ReLU counts; least-critical ReLU proportion is crucial at lower counts.