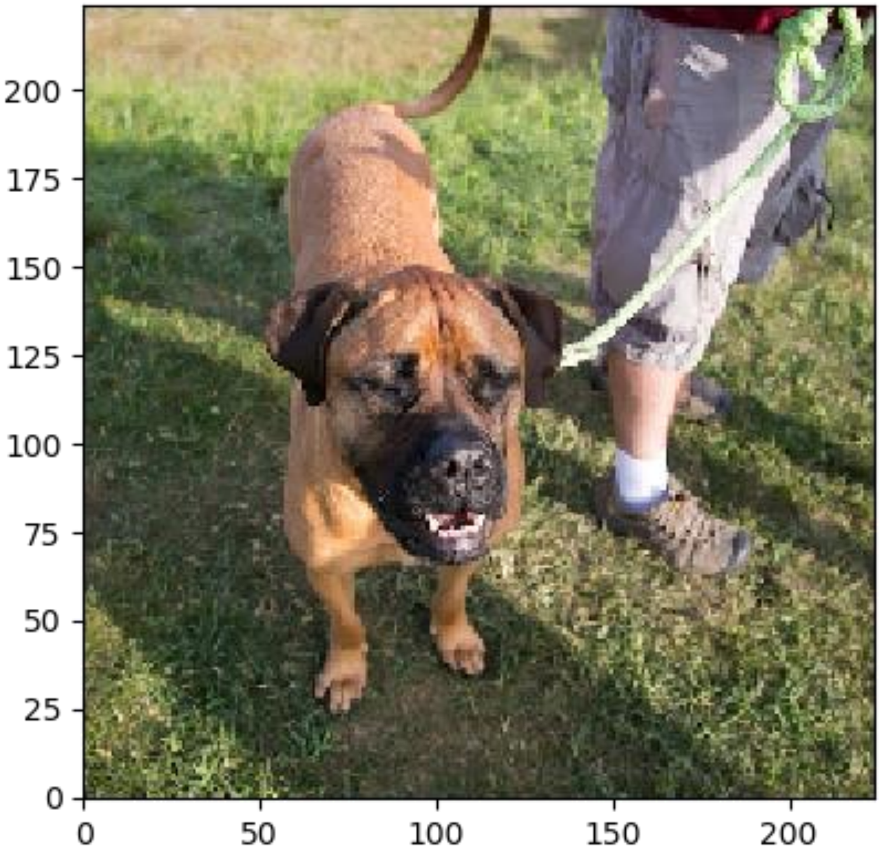
Band-limited Training and Inference for Convolutional Neural Networks

Adam Dziedzic*, John Paparrizos*, Sanjay Krishnan, Aaron Elmore, Michael Franklin

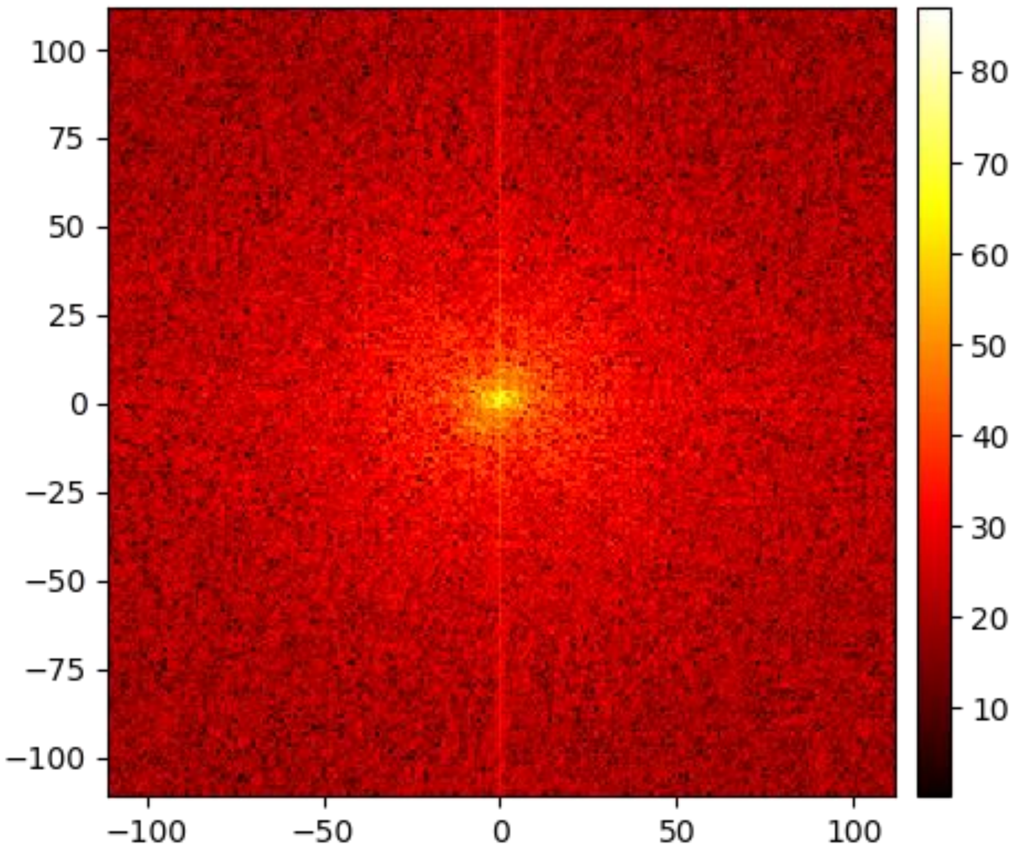
THE UNIVERSITY OF CHICAGO

Natural images
More information put in lower frequencies

Original image



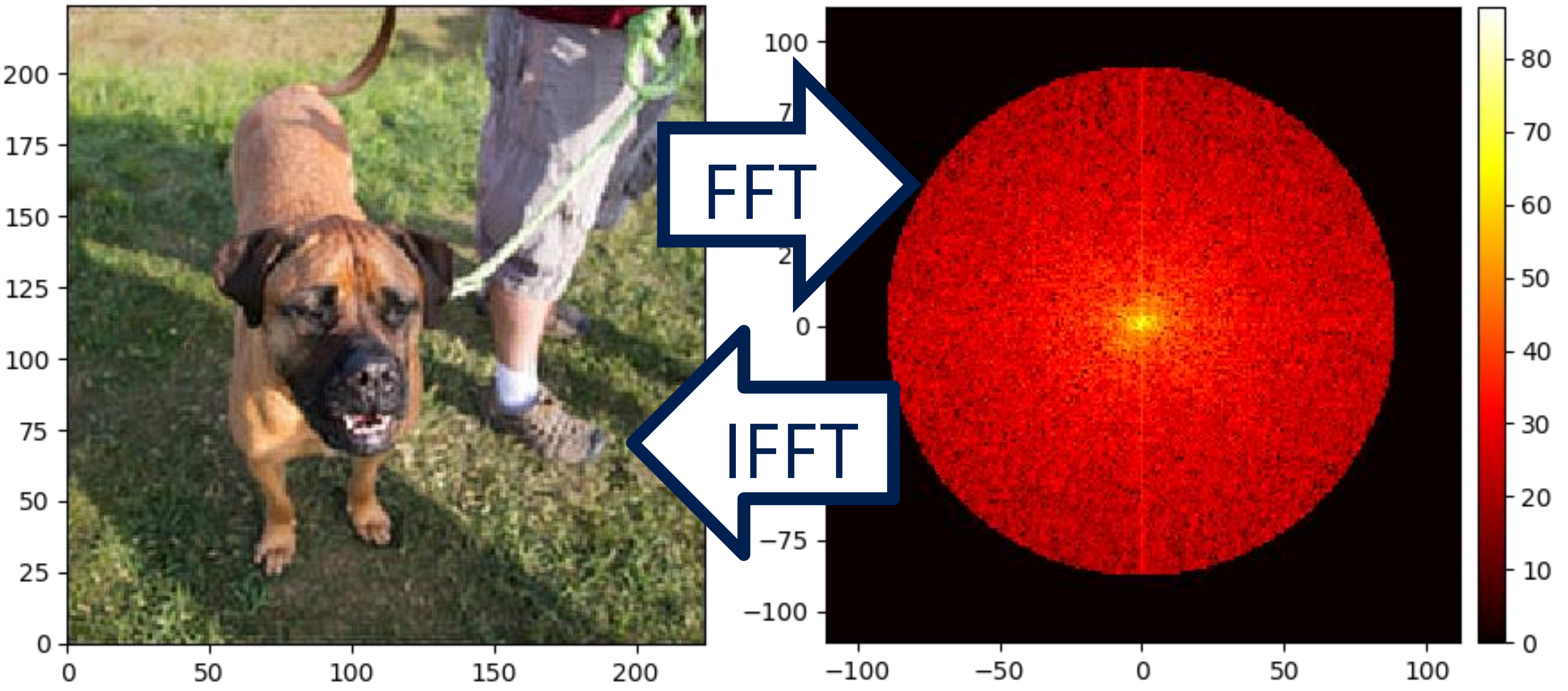
Spatial domain



Frequency domain

Natural images
Transformations between the domains

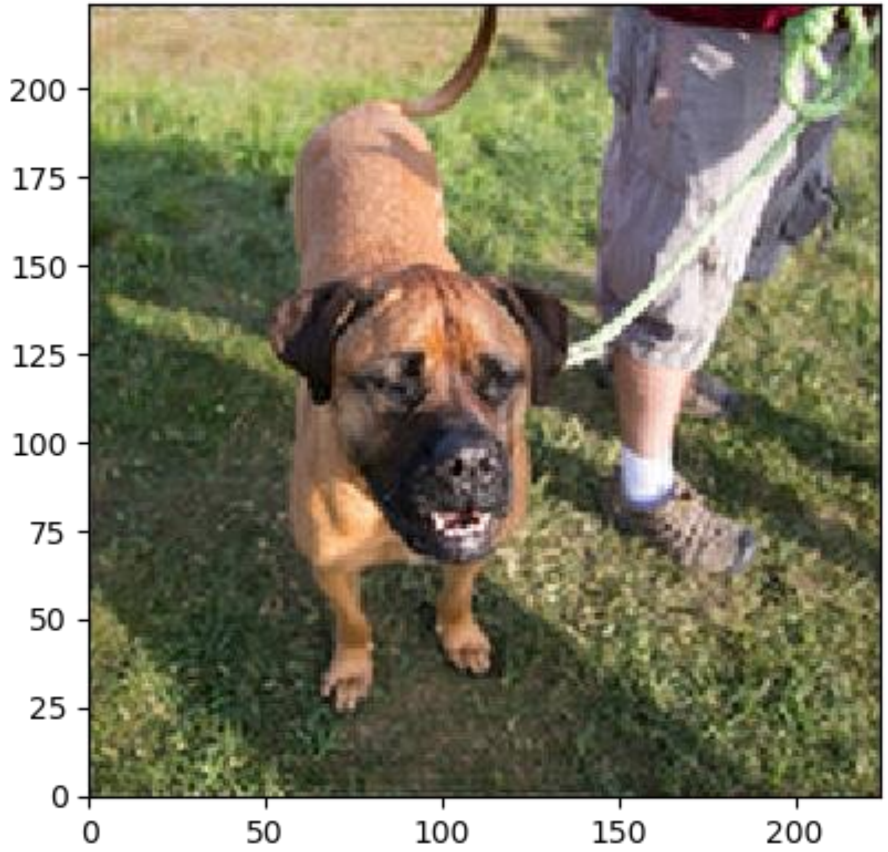
Compression 50%



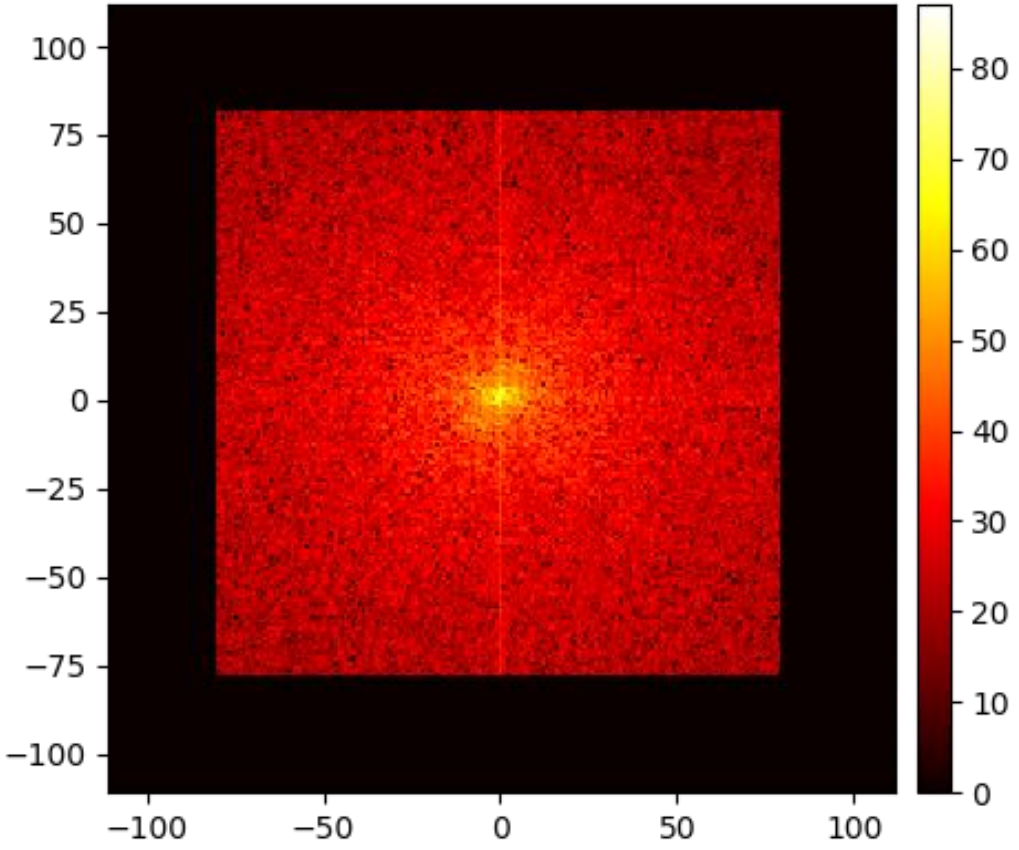
Spatial domain **Frequency domain**

Method for ConvNets to constrain the frequency band in convolution operation for efficiency

Compression 50% in practice



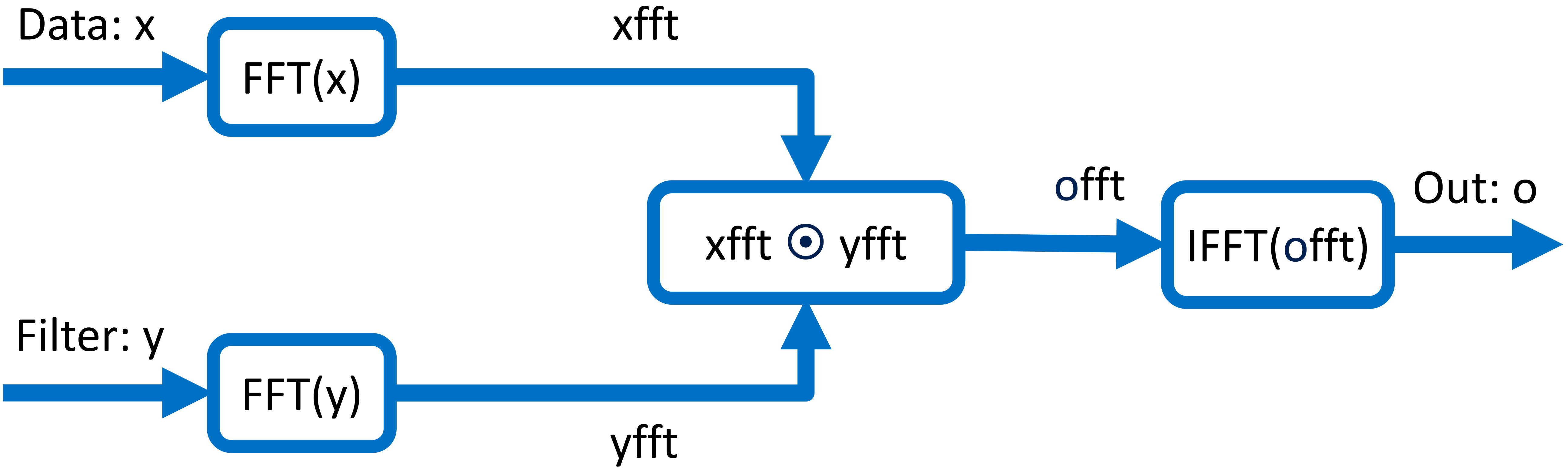
Spatial domain



Frequency domain

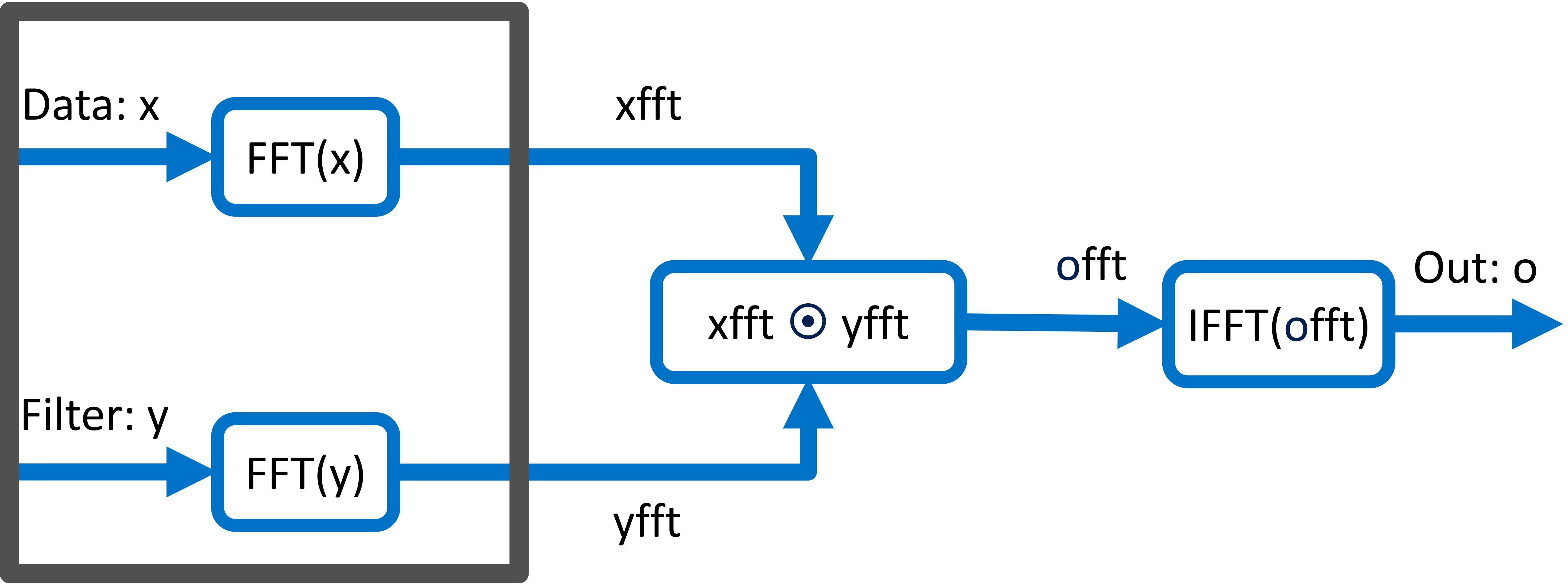
FFT based convolution

Mathieu et al.: "Fast Training of Convolutional Networks through FFTs"
Vasilache et al.: "Fast Convolutional Nets With fbfft: A GPU Performance Evaluation"



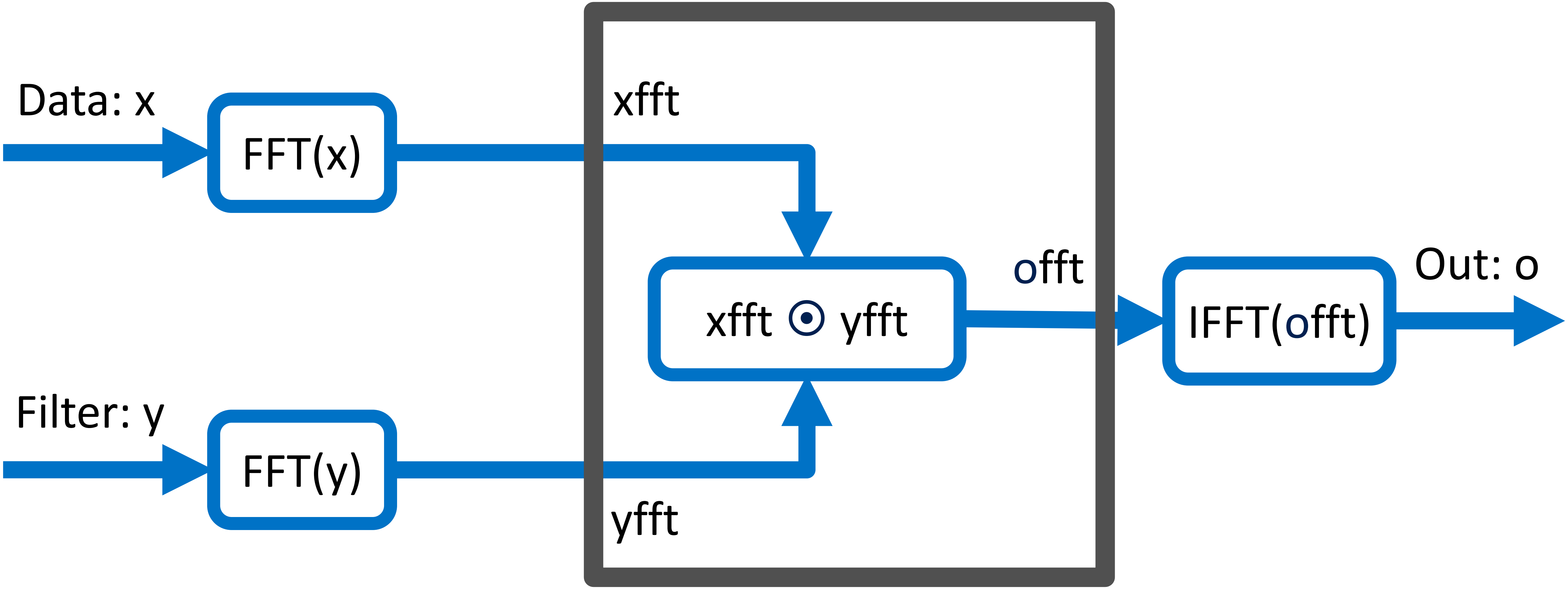
FFT based convolution

Mathieu et al.: "Fast Training of Convolutional Networks through FFTs"
Vasilache et al.: "Fast Convolutional Nets With fbfft: A GPU Performance Evaluation"



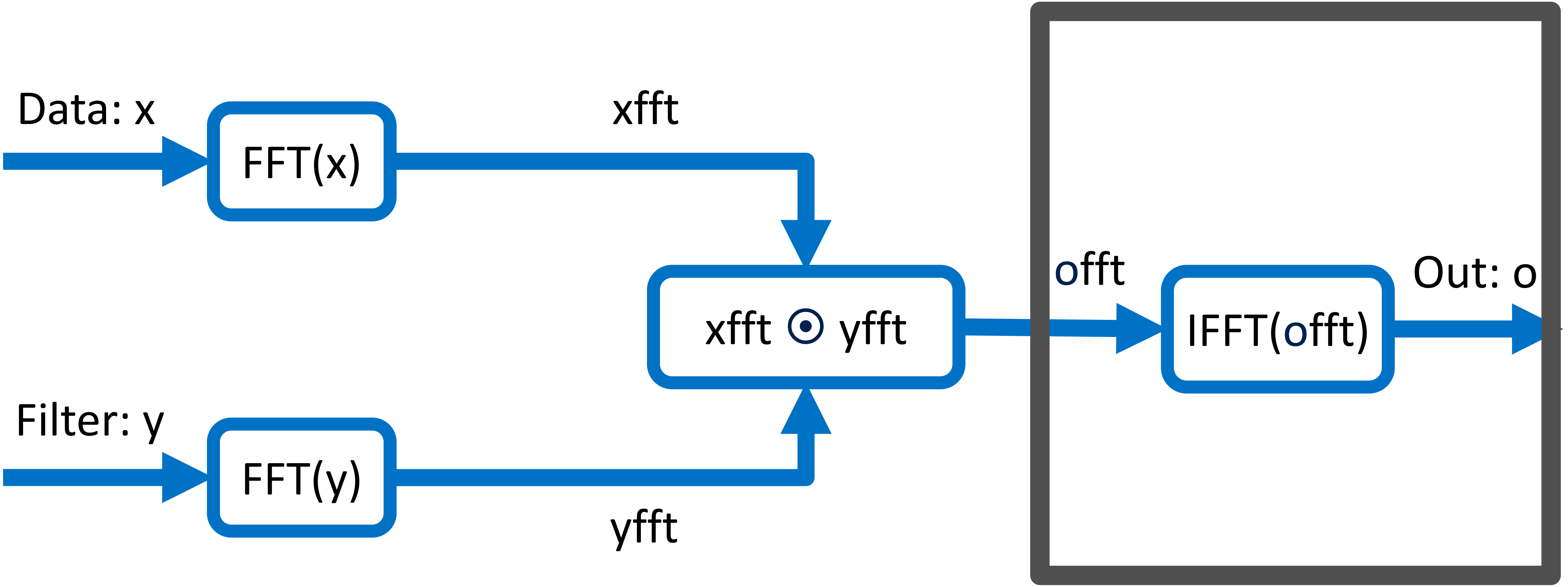
FFT based convolution

Mathieu et al.: "Fast Training of Convolutional Networks through FFTs"
Vasilache et al.: "Fast Convolutional Nets With fbfft: A GPU Performance Evaluation"



FFT based convolution

Mathieu et al.: "Fast Training of Convolutional Networks through FFTs"
Vasilache et al.: "Fast Convolutional Nets With fbfft: A GPU Performance Evaluation"

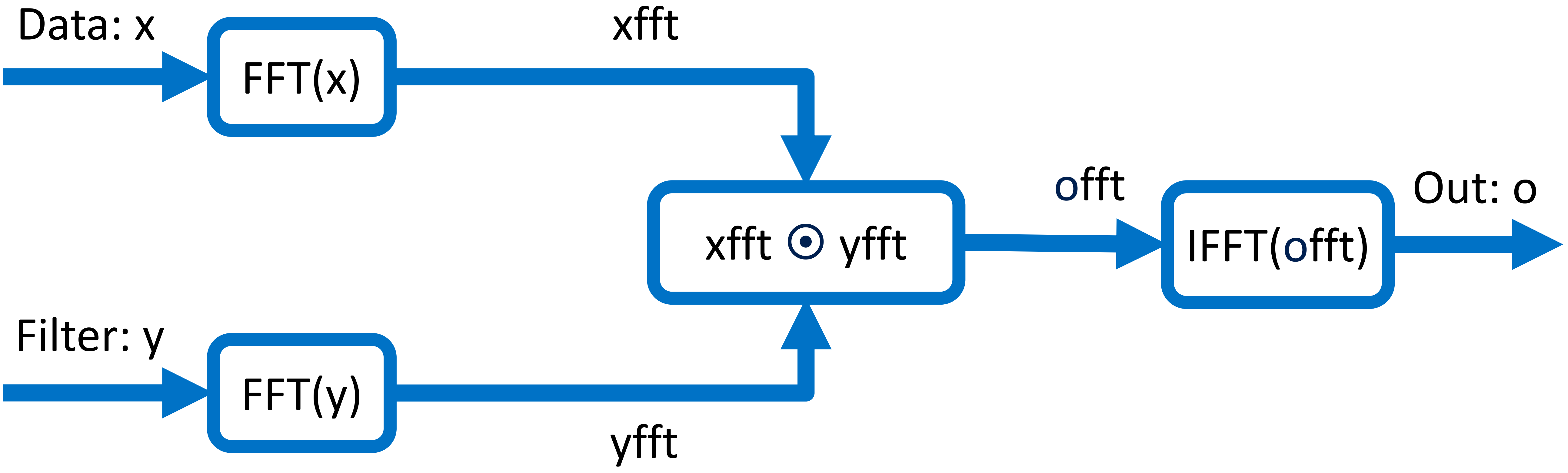


FFT based convolution

Mathieu et al.: "Fast Training of Convolutional Networks through FFTs"

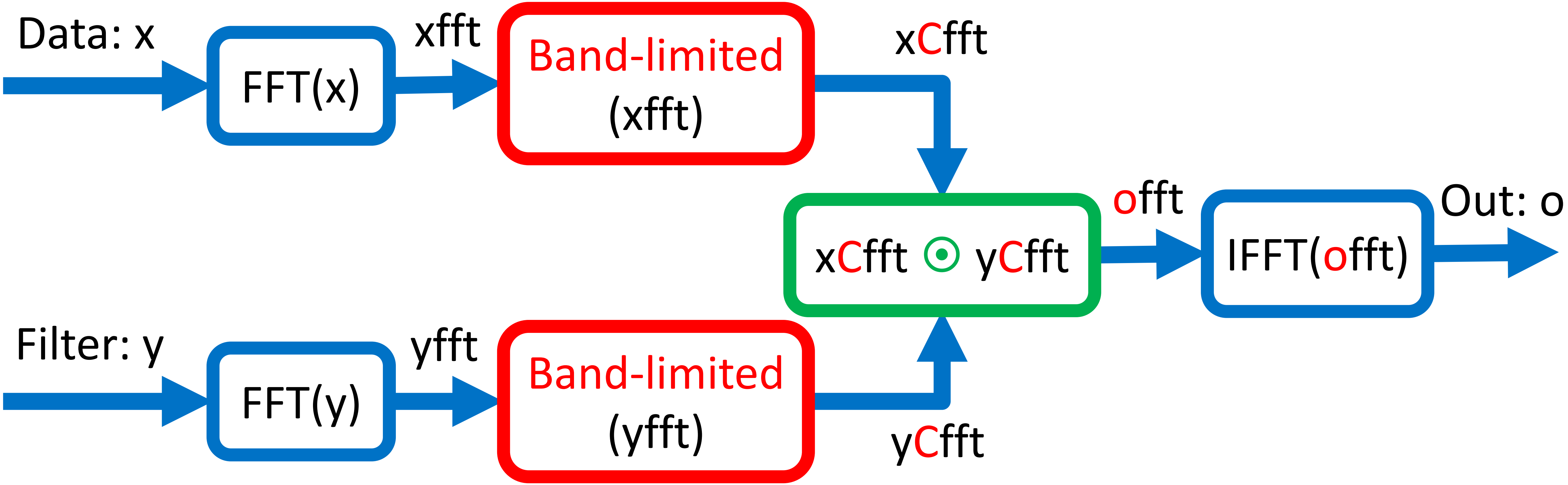
Vasilache et al.: "Fast Convolutional Nets With fbfft: A GPU Performance Evaluation"

cuDNN: Substantial memory workspace needed for intermediate results.

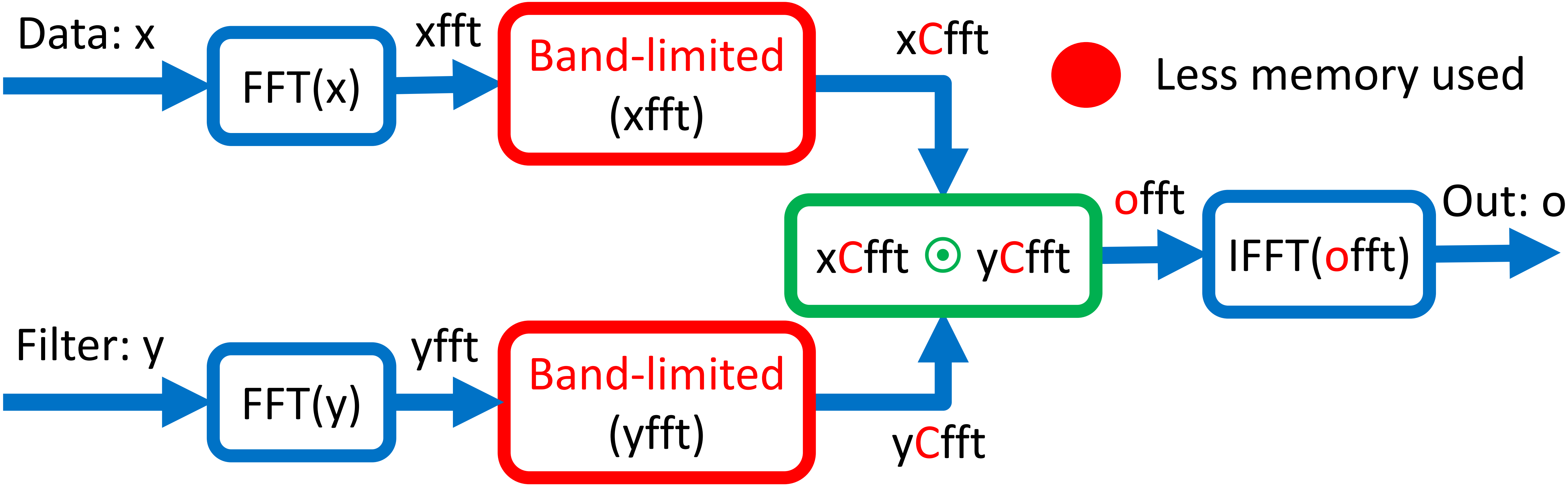


Band-limited FFT based convolution

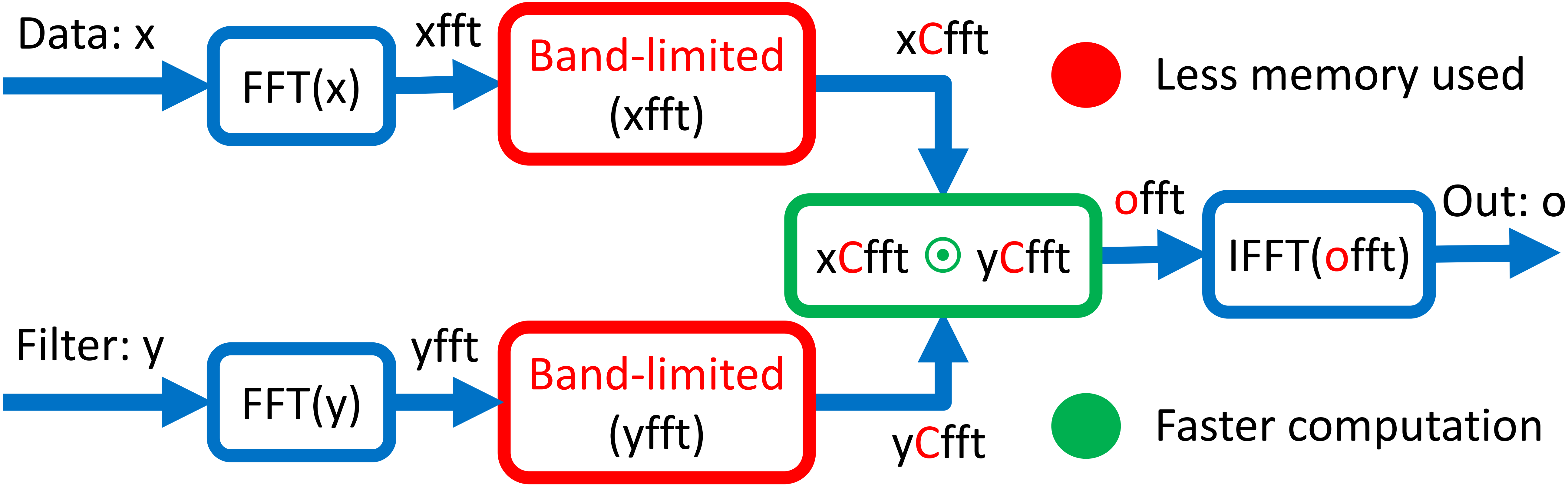
Band-limiting = masking out high frequencies



Band-limited FFT based convolution

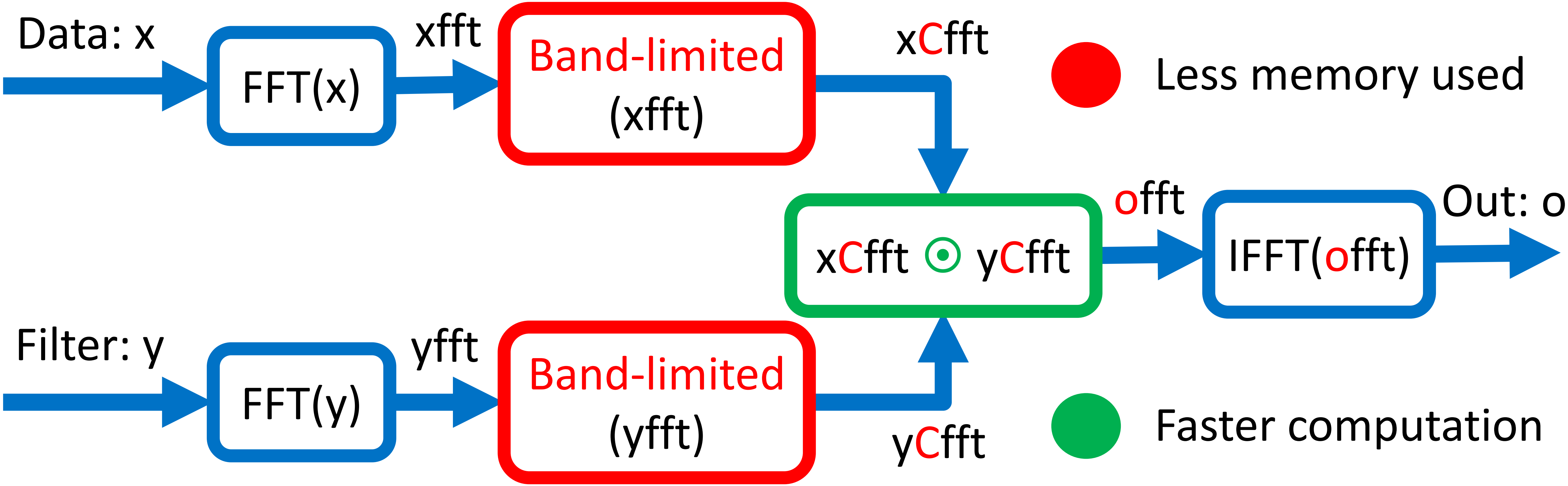


Band-limited FFT based convolution



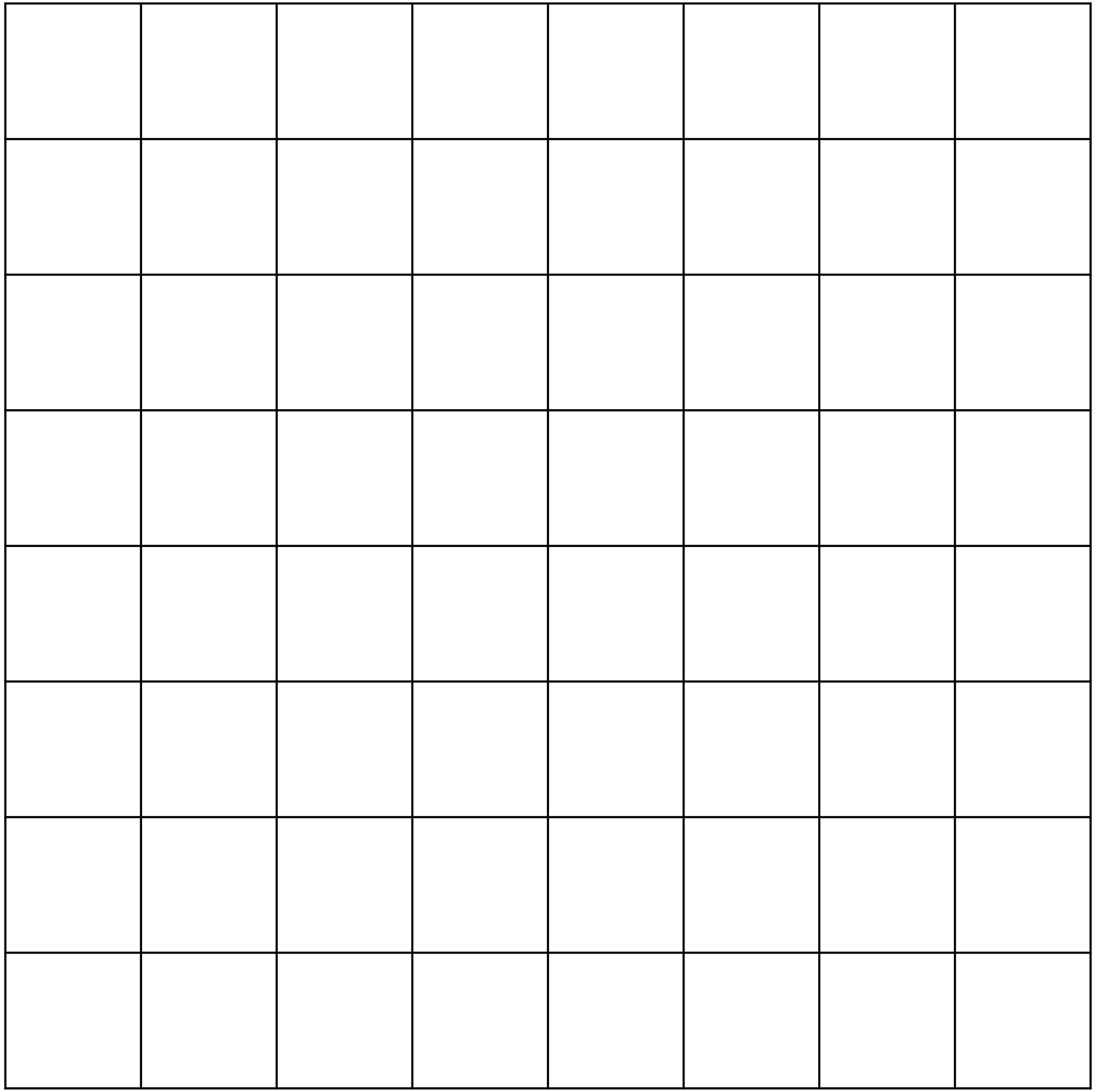
Band-limited FFT based convolution

Preserve enough of the spectrum to retain high accuracy of models.



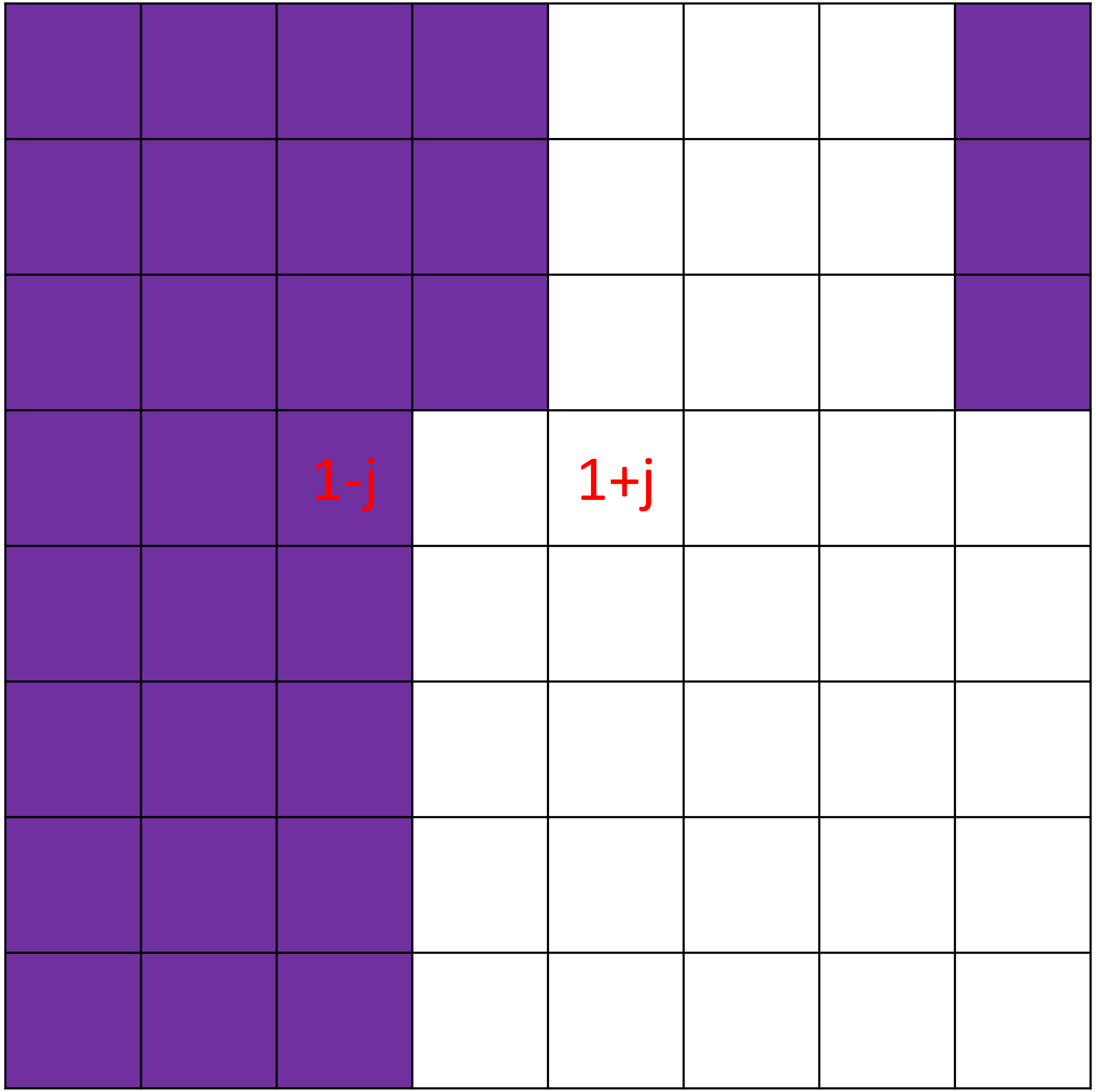
Band-limiting Technique

1. FFT of an input data



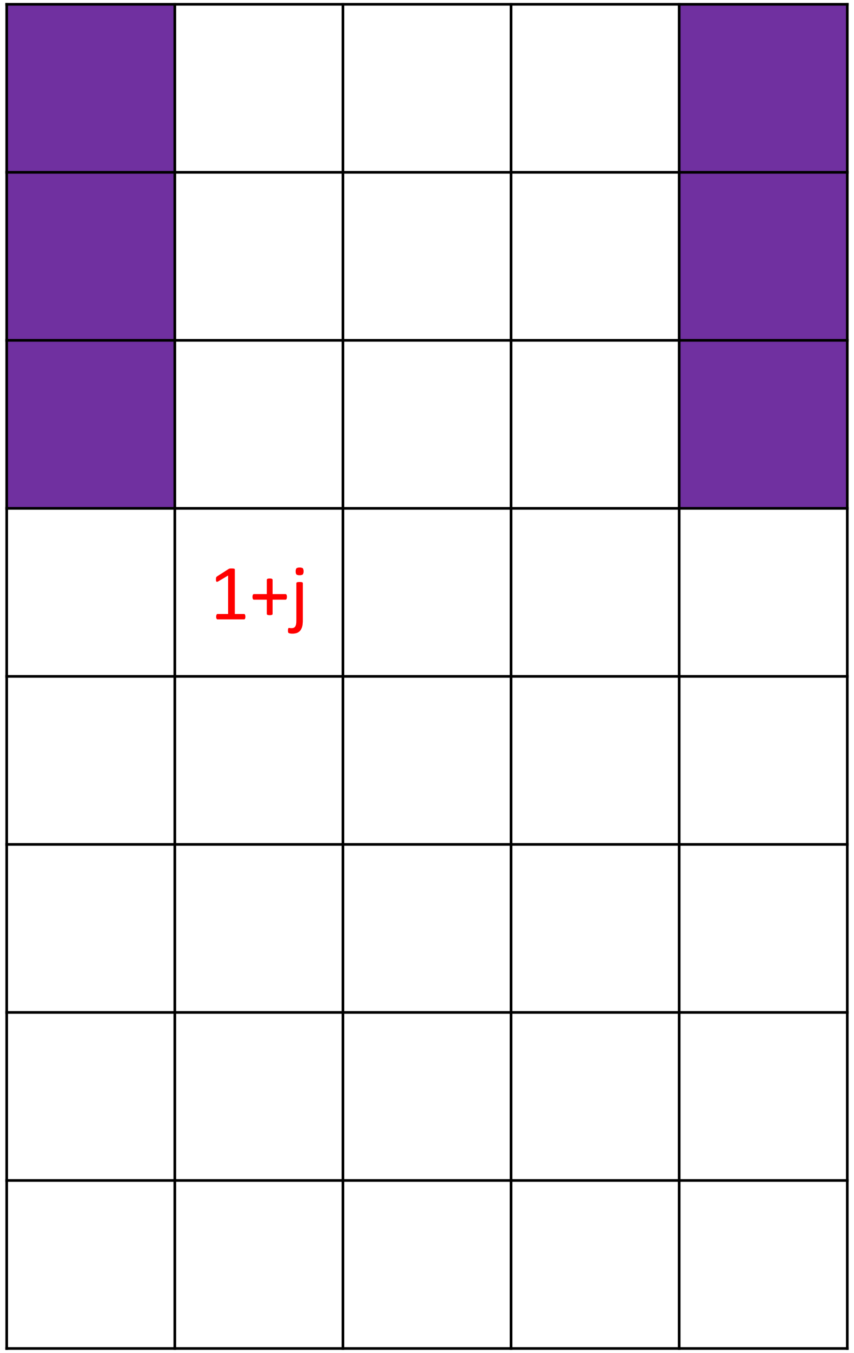
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry



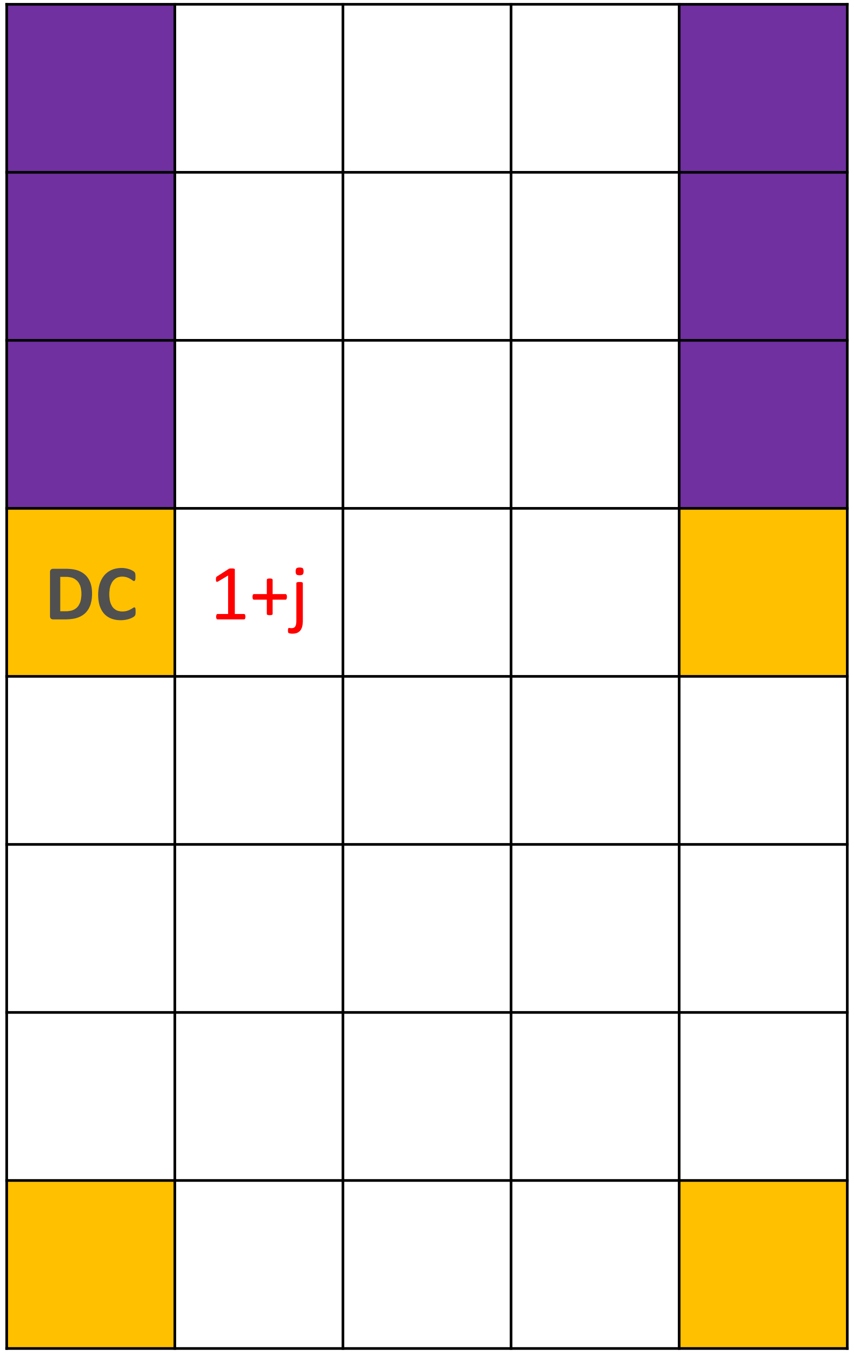
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry



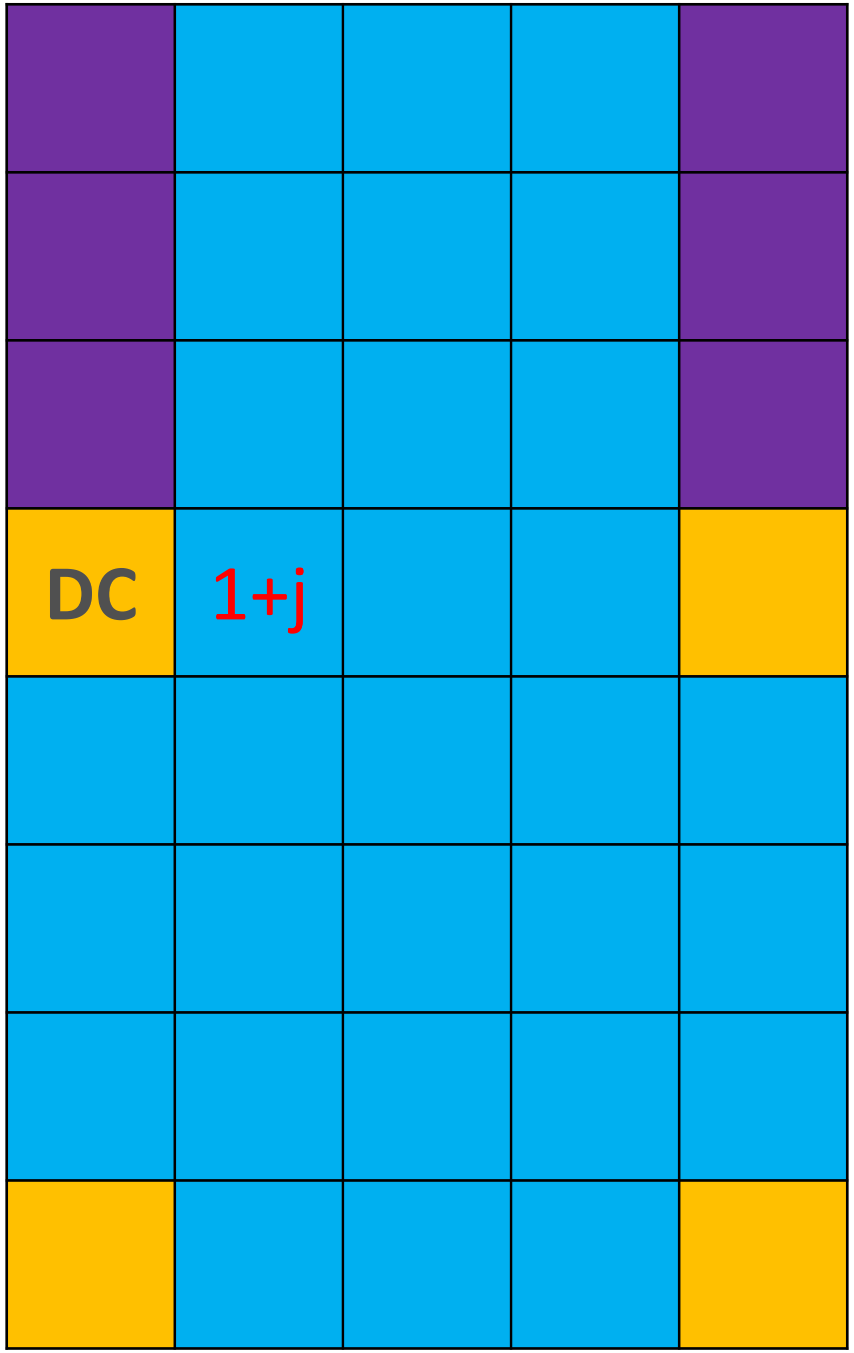
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry
3. Real values



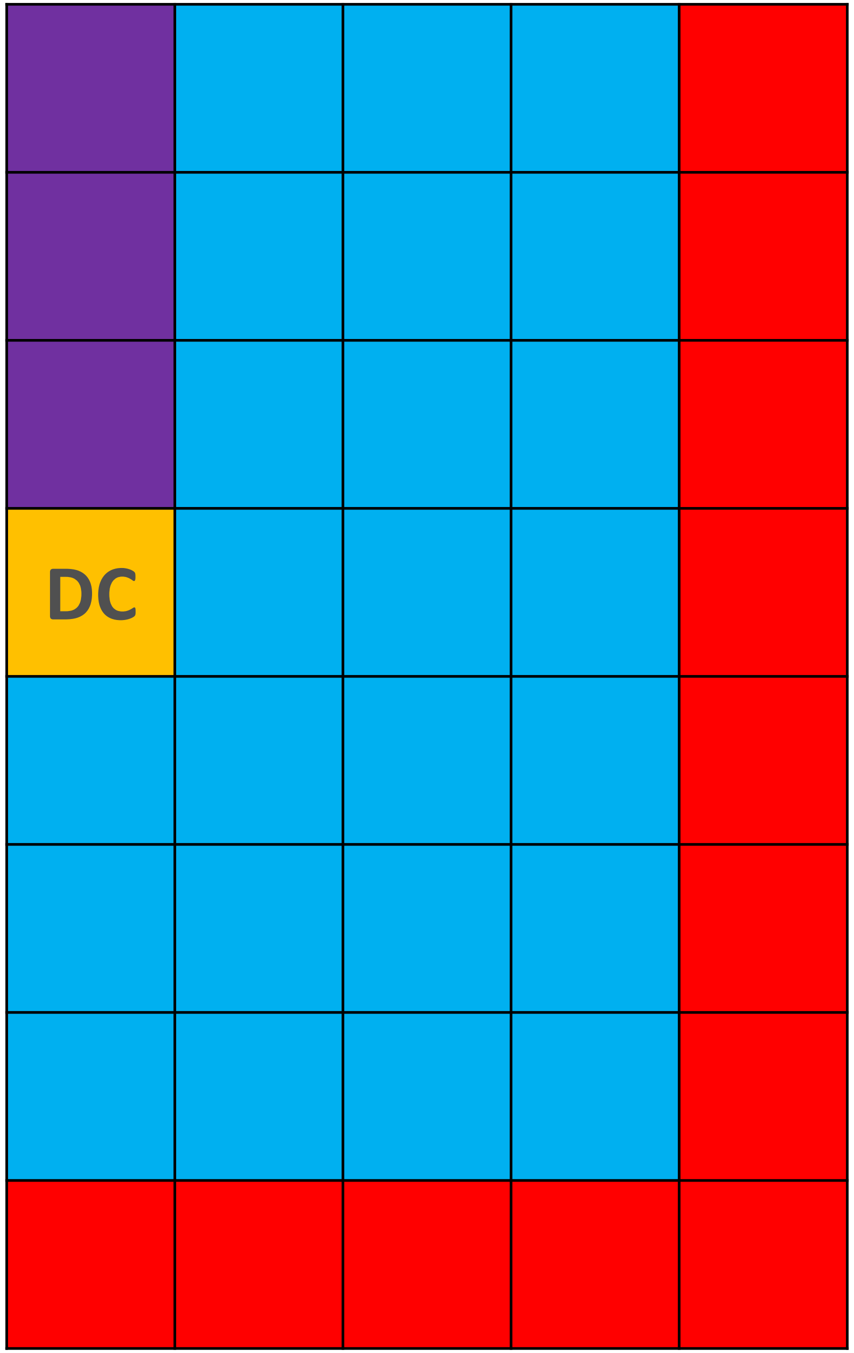
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry
3. Real values
4. No constraints



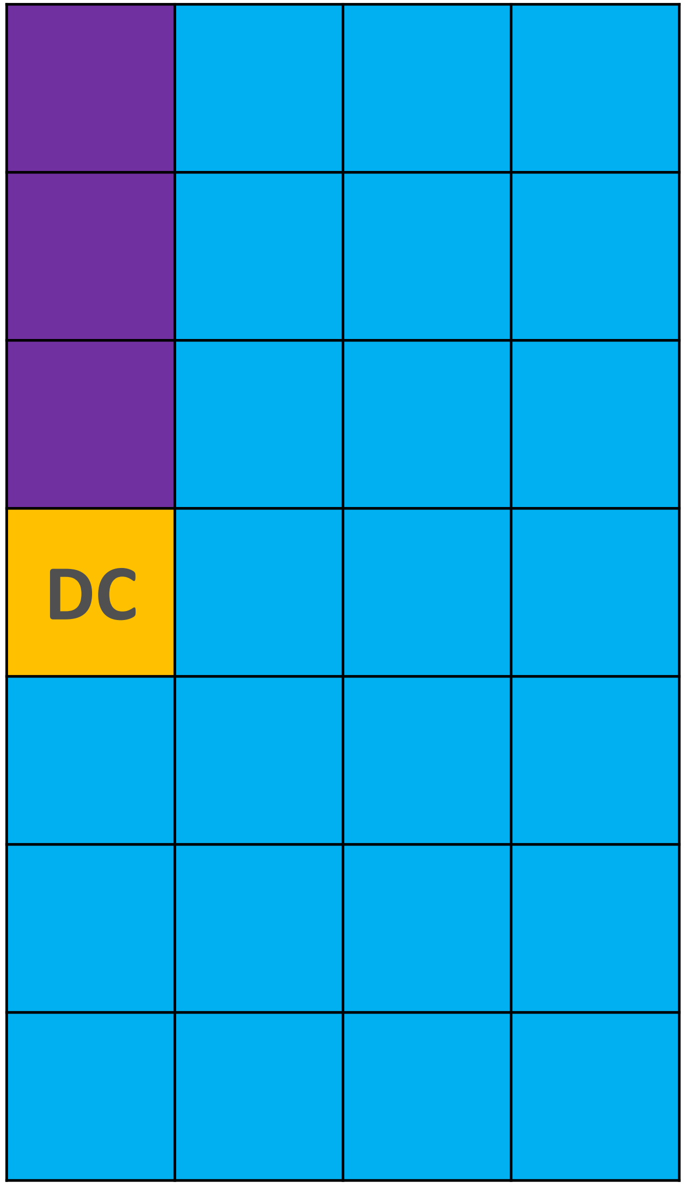
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry
3. Real values
4. No constraints
5. 1st compression



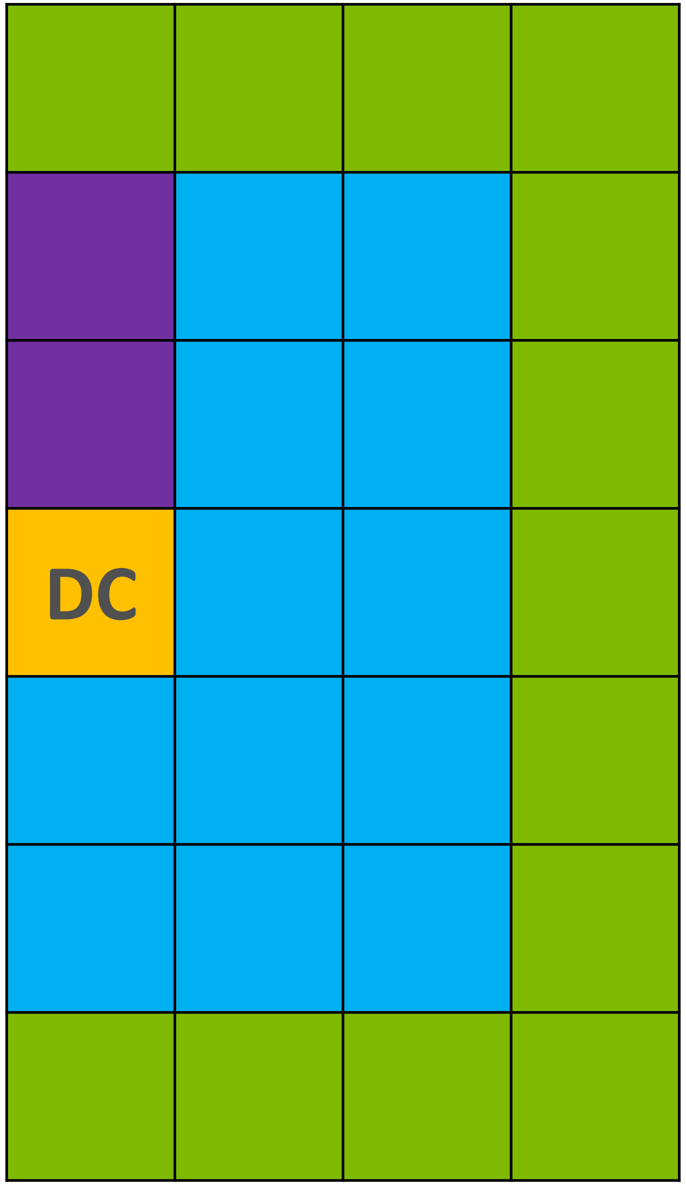
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry
3. Real values
4. No constraints
5. 1st compression



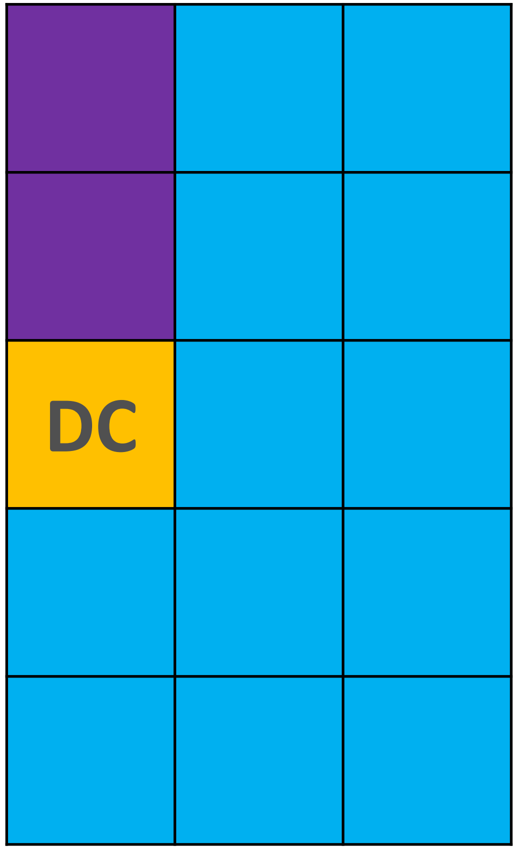
Band-limiting Technique

1. FFT of an input data
2. Conjugate symmetry
3. Real values
4. No constraints
5. 1st compression
6. 2nd compression

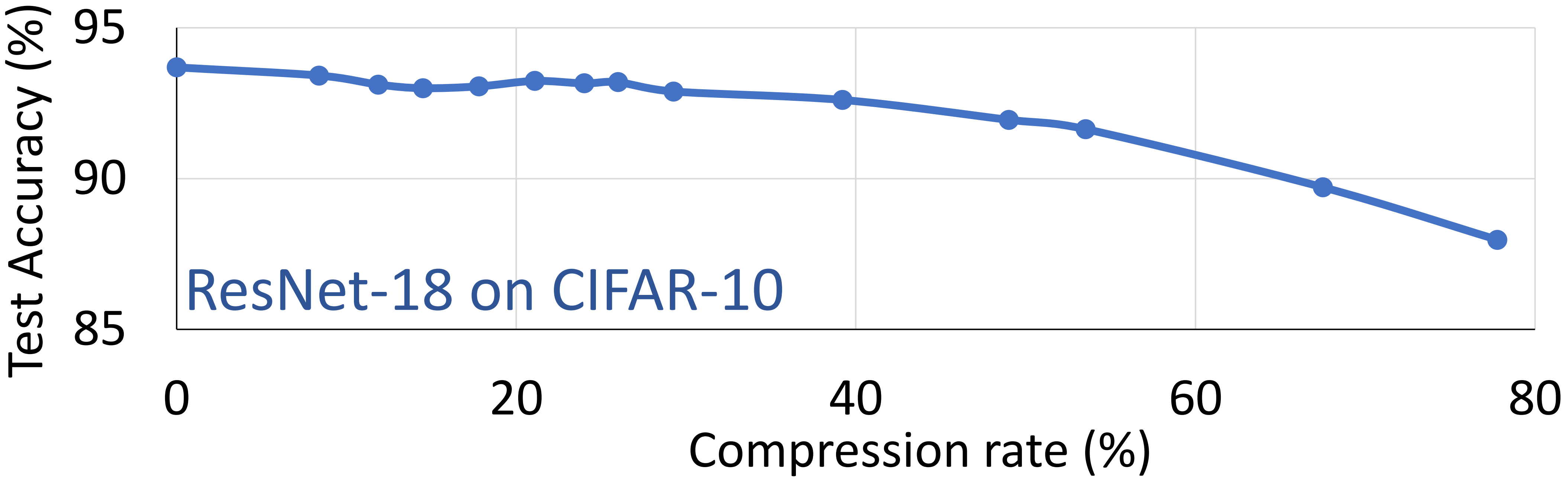


Band-limiting Technique

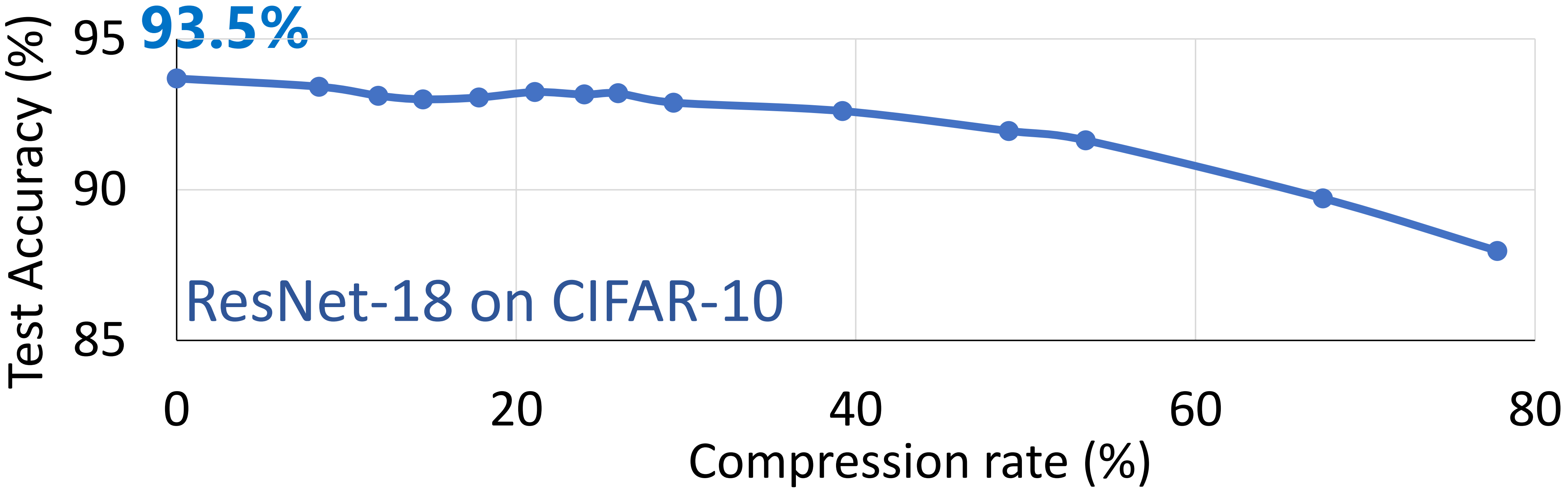
1. FFT of an input data
2. Conjugate symmetry
3. Real values
4. No constraints
5. 1st compression
6. 2nd compression



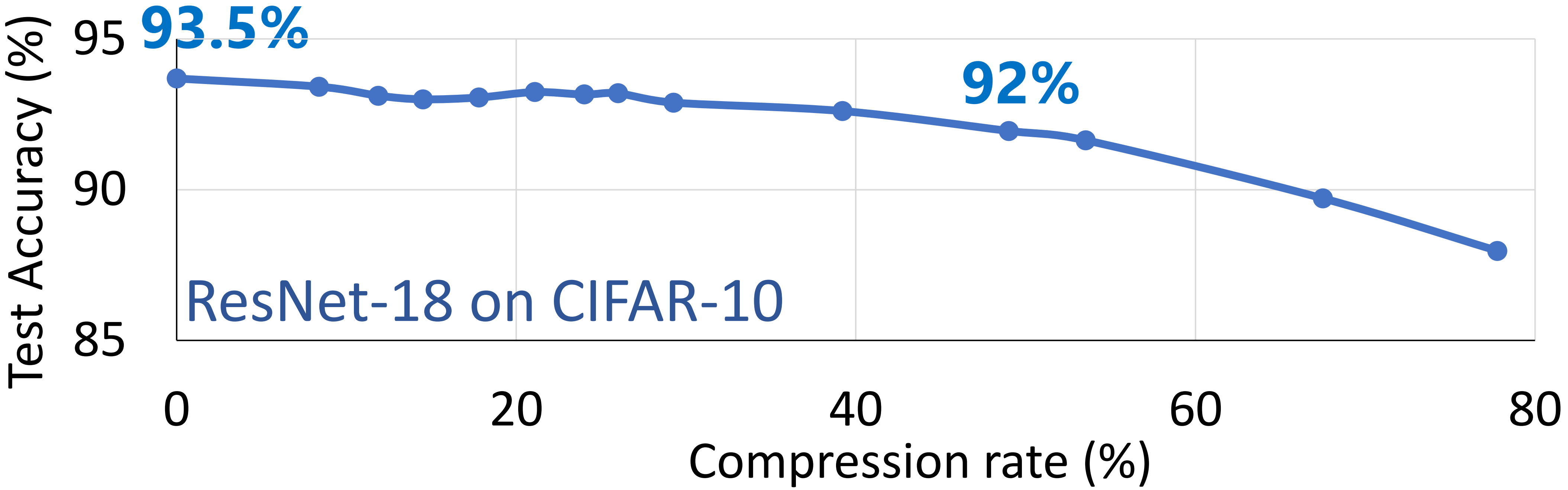
Effects of band-limiting on accuracy



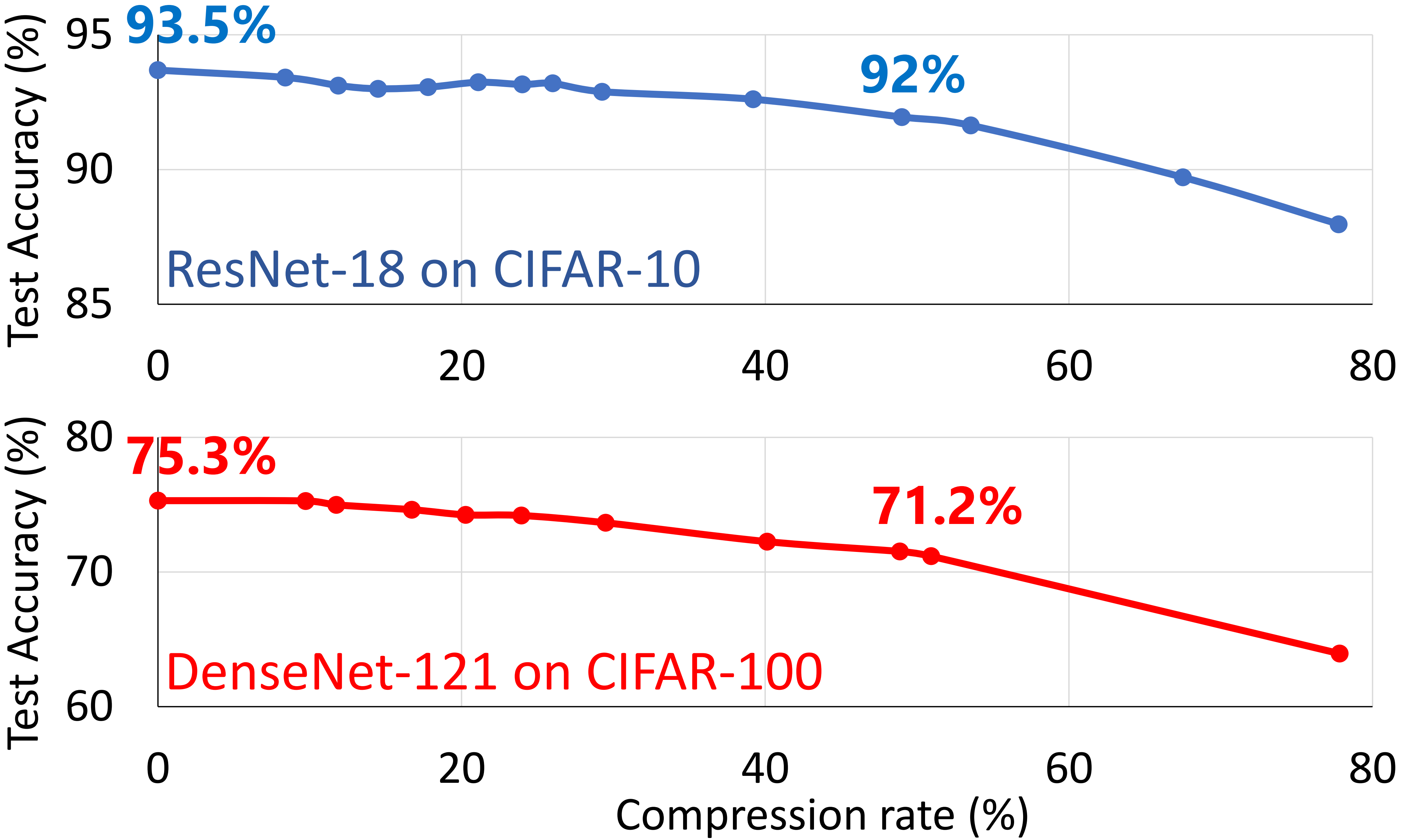
Effects of band-limiting on accuracy



Effects of band-limiting on accuracy



Effects of band-limiting on accuracy



Main take-aways from Band-limited CNNs

- Method to constrain the frequency band in convolution.
- Models trained with band-limiting **gracefully degrade** the accuracy as the function of the compression rate.
- Effectively **control resource usage** (GPU/CPU and memory).
- The **low frequency** coefficients learned first during training.
- The **same compression rate** applied to training and inference.
- The more band-limited model, the more **robust to attacks**.
- Applicable to **other domains:** time-series & speech data.

Thank you

Poster: 6:30-9:00 PM @ Pacific Ballroom #132

github.com/adam-dziedzic/bandlimited-cnns

ady@uchicago.edu

Backup

THE UNIVERSITY OF CHICAGO

Why is FFT based convolution important?

- The theoretical properties of the Fourier domain are well understood. No such properties in other domains (Winograd).
- ResNet and DenseNet architectures use 7x7 filters in first layers.
- FFT based convolution can be combined with spectral pooling.
- Band-limiting minimize aliasing & serves as a simple defense.
- A standard algorithm included in popular frameworks (cuDNN).
- Gradient acts as a large filter in the backward pass.
- Zlateski et al. suggest using FFT based convolution on CPUs.
- The 1D FFT convolution for DSP where large filters are used.

Band-limited FFT based convolution *formally*

Cross-correlate input data and filter: $x *_c y$

$$F_x[\omega] = F(x[n]) \qquad F_y[\omega] = F(y[n])$$

$$x *_c y = F^{-1}(F_x[\omega] \odot F_y[\omega])$$

Spectrum of convolution: $S[\omega] = F_x[\omega] \odot F_y[\omega]$

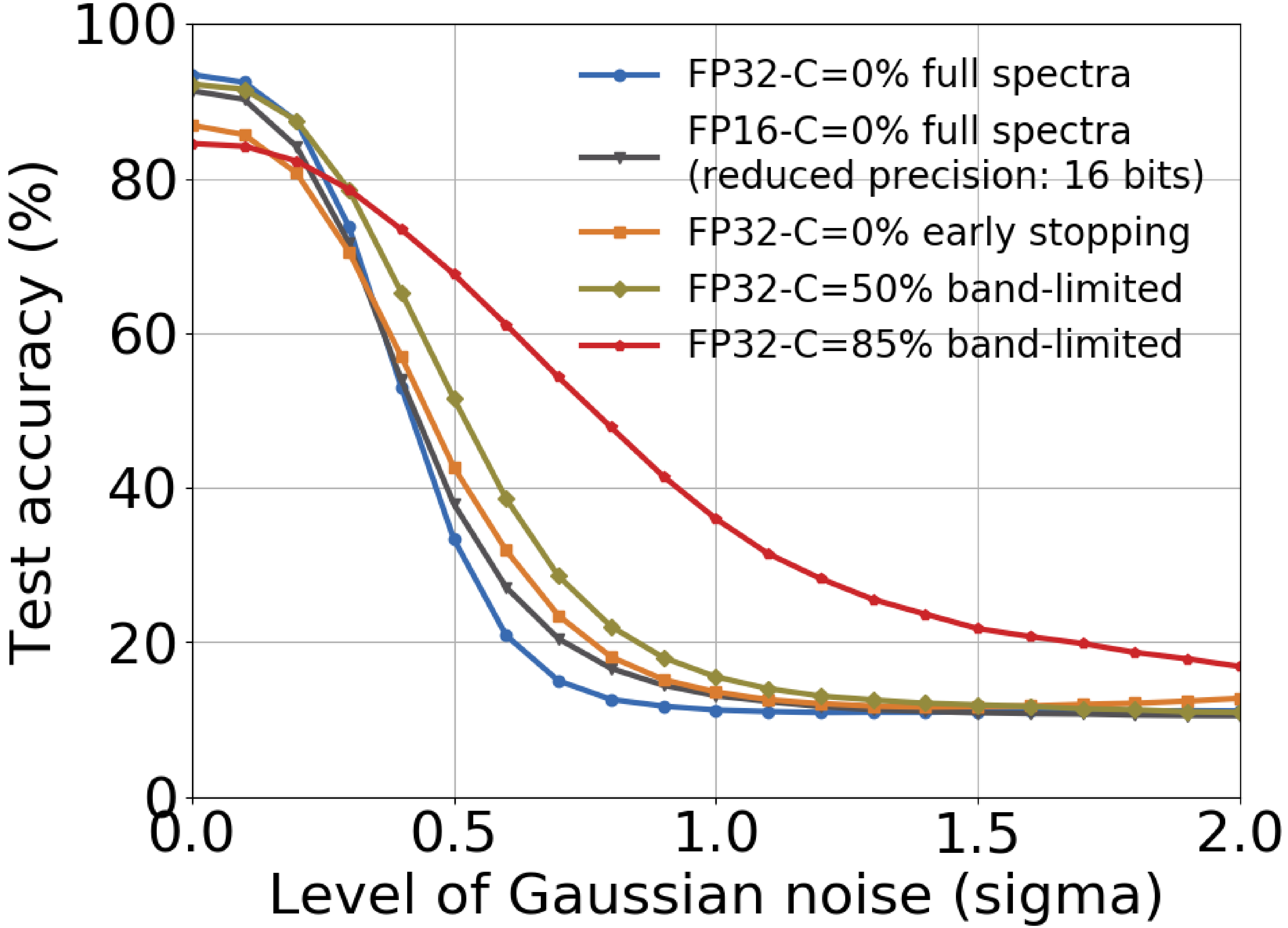
$$\mathbf{M_c}[\boldsymbol{\omega}] = \begin{cases} \mathbf{1}, \boldsymbol{\omega} \le \mathbf{c} \\ \mathbf{0}, \boldsymbol{\omega} > \mathbf{c} \end{cases}$$

$$x *_c y = F^{-1}\left[\left(F_x[\omega] \odot M_c[\omega]\right) \odot \left(F_y[\omega] \odot M_c[\omega]\right)\right]$$

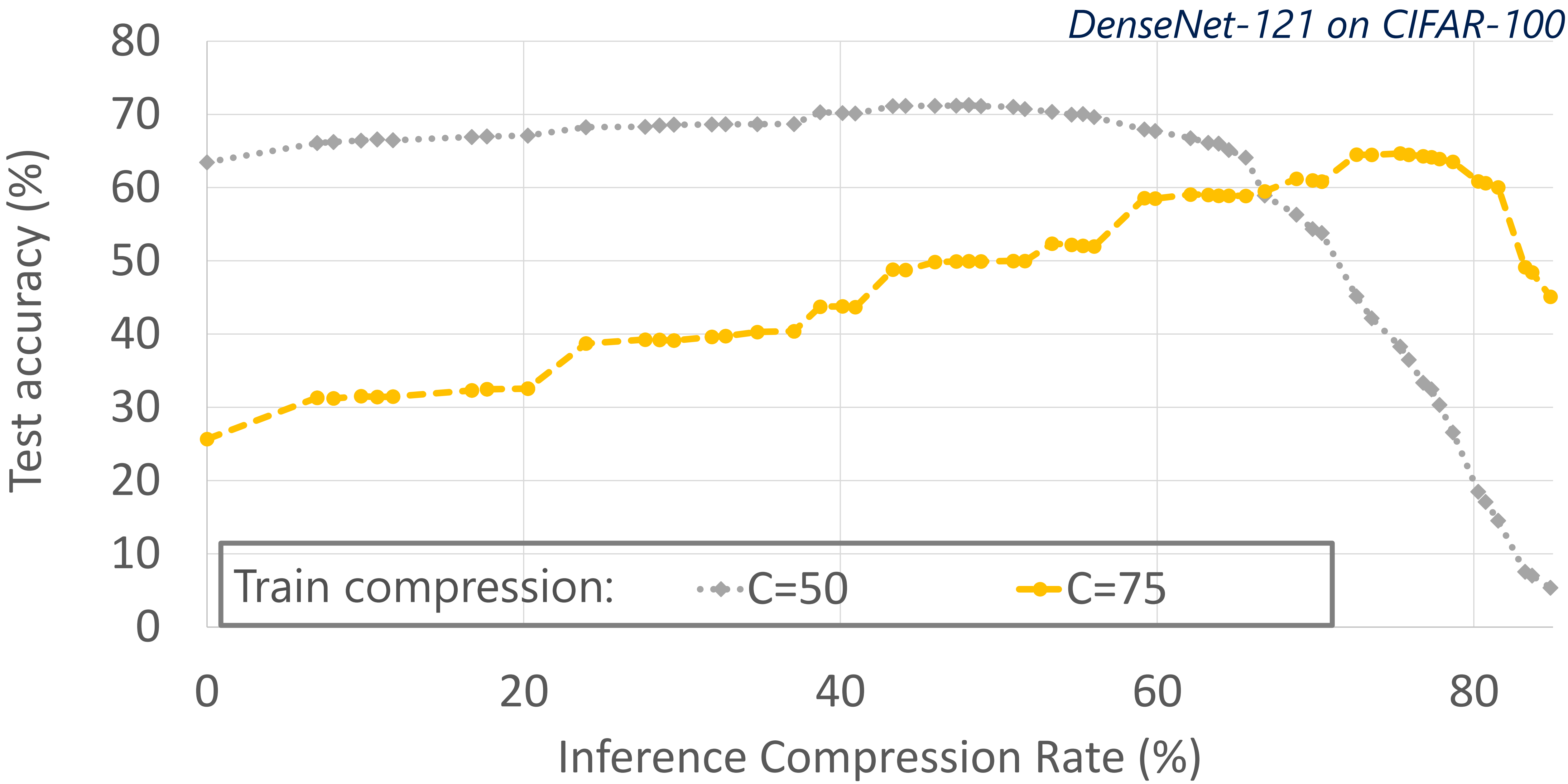
$$x *_c y = F^{-1}(S[\omega] \odot M_c[\omega])$$

Energy (Parseval's theorem): $\sum_{n=0}^{N-1} |x[n]|^2 = \sum_{\omega=0}^{2\pi} |F_x(\omega)|^2$

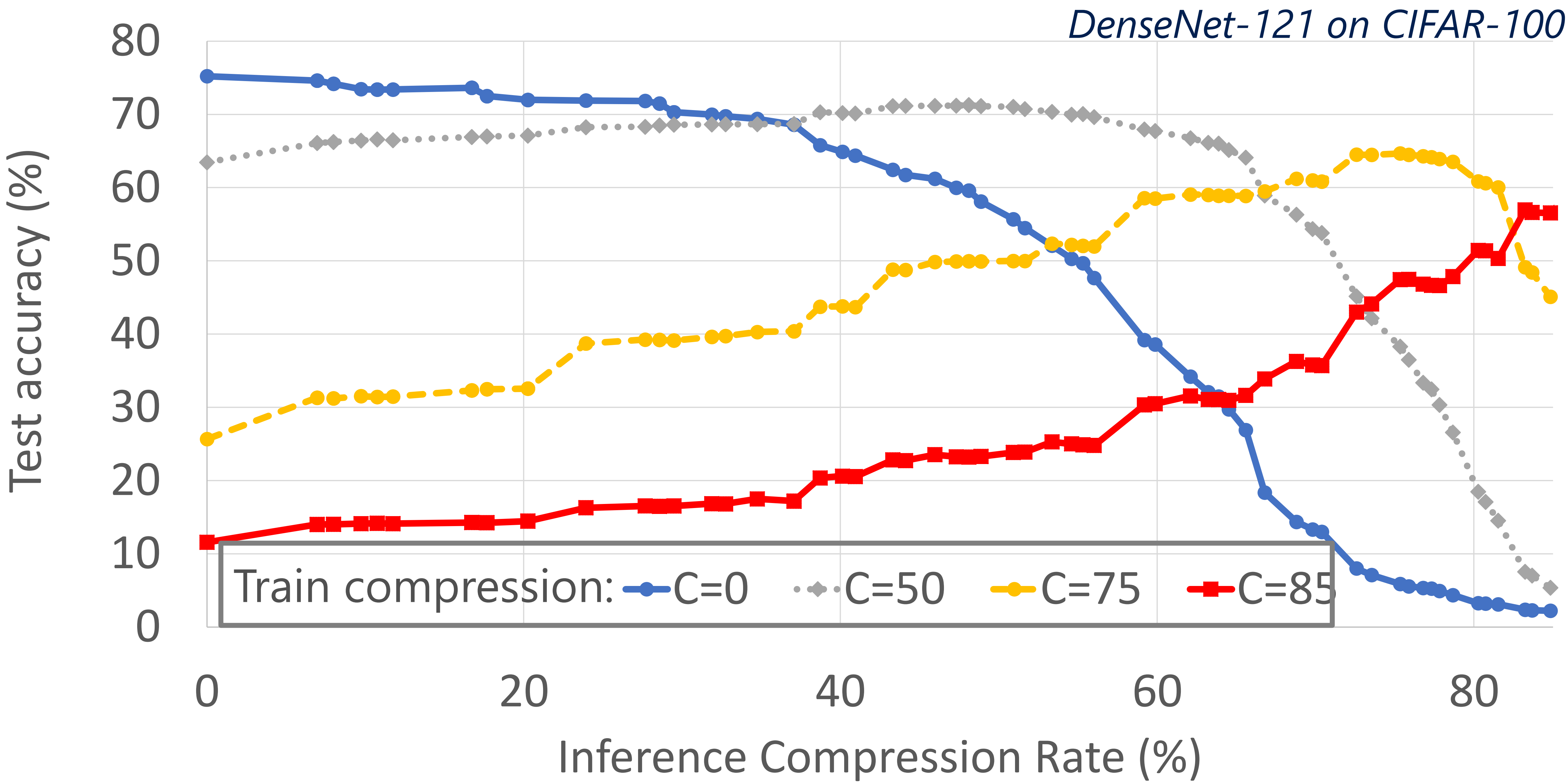
Robustness to noise



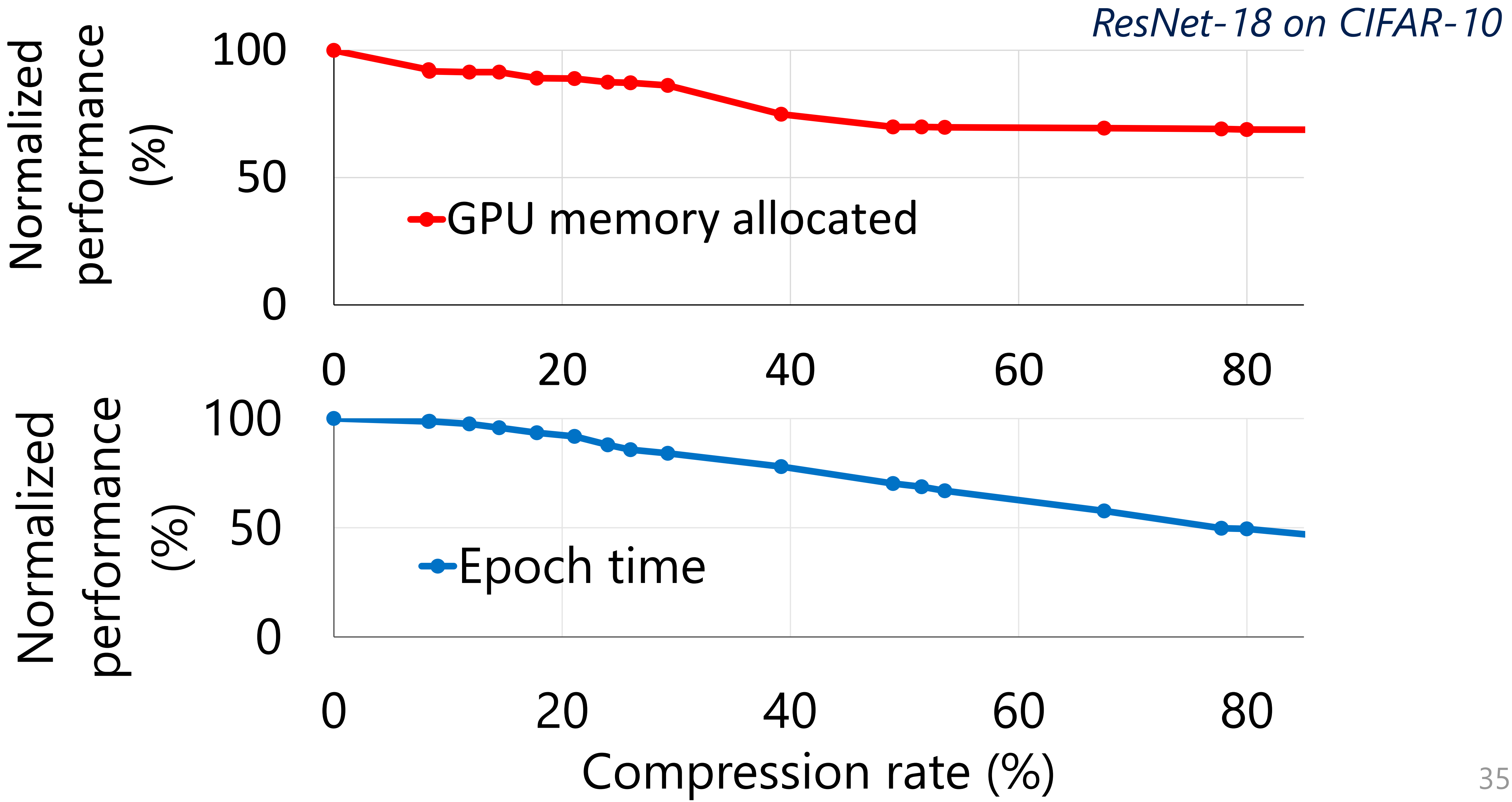
Compression Rate for Training vs Inference



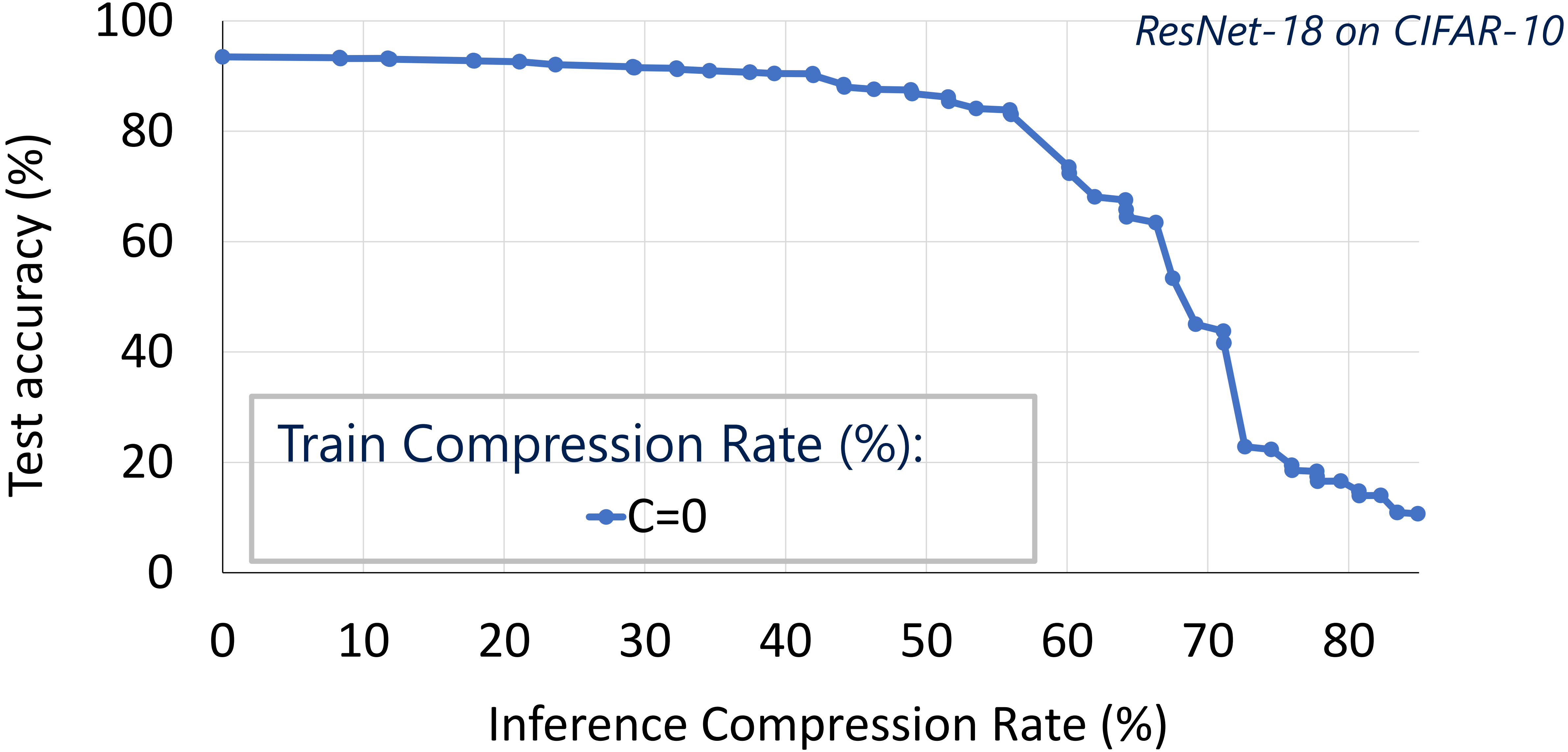
Compression Rate for Training vs Inference



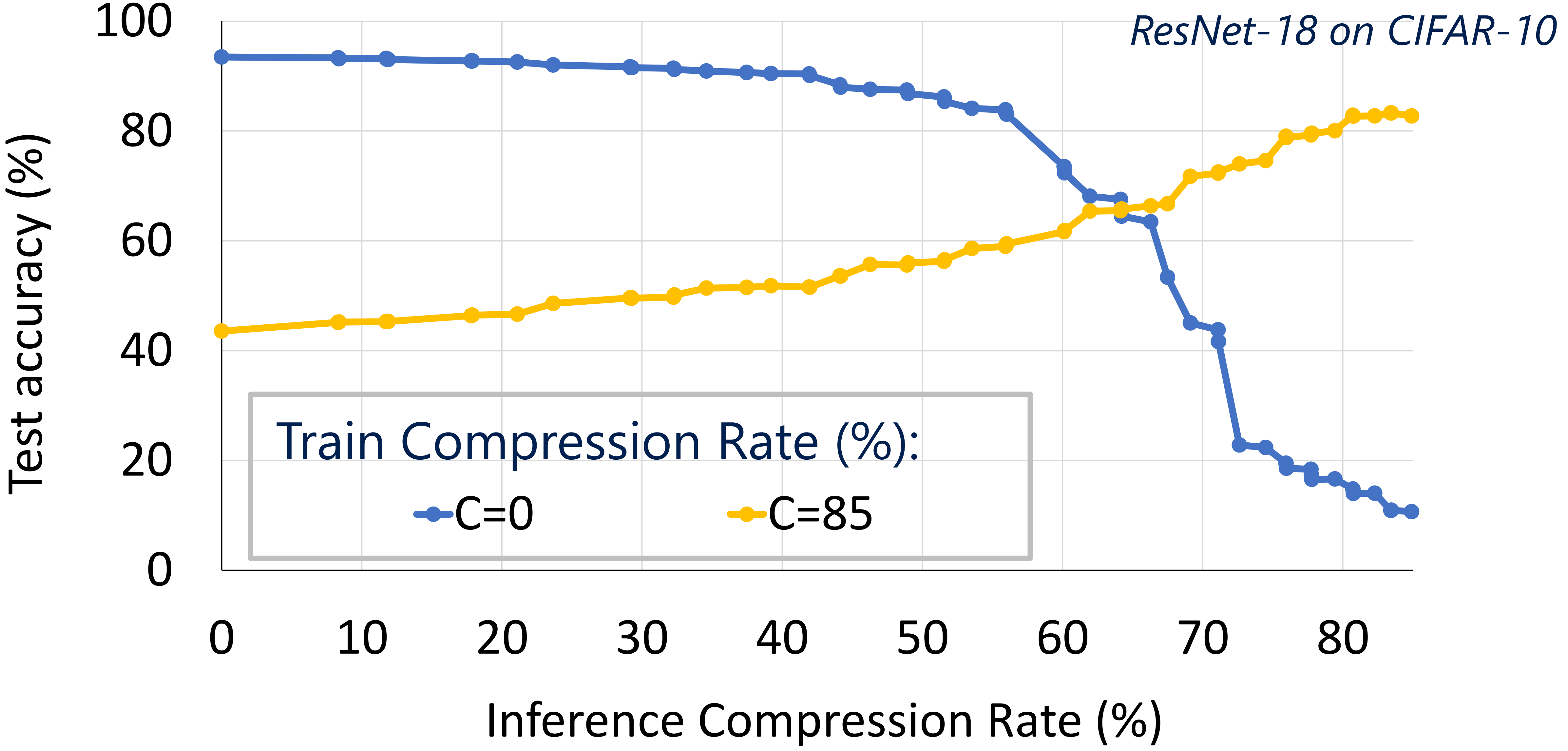
Effectively control resource usage



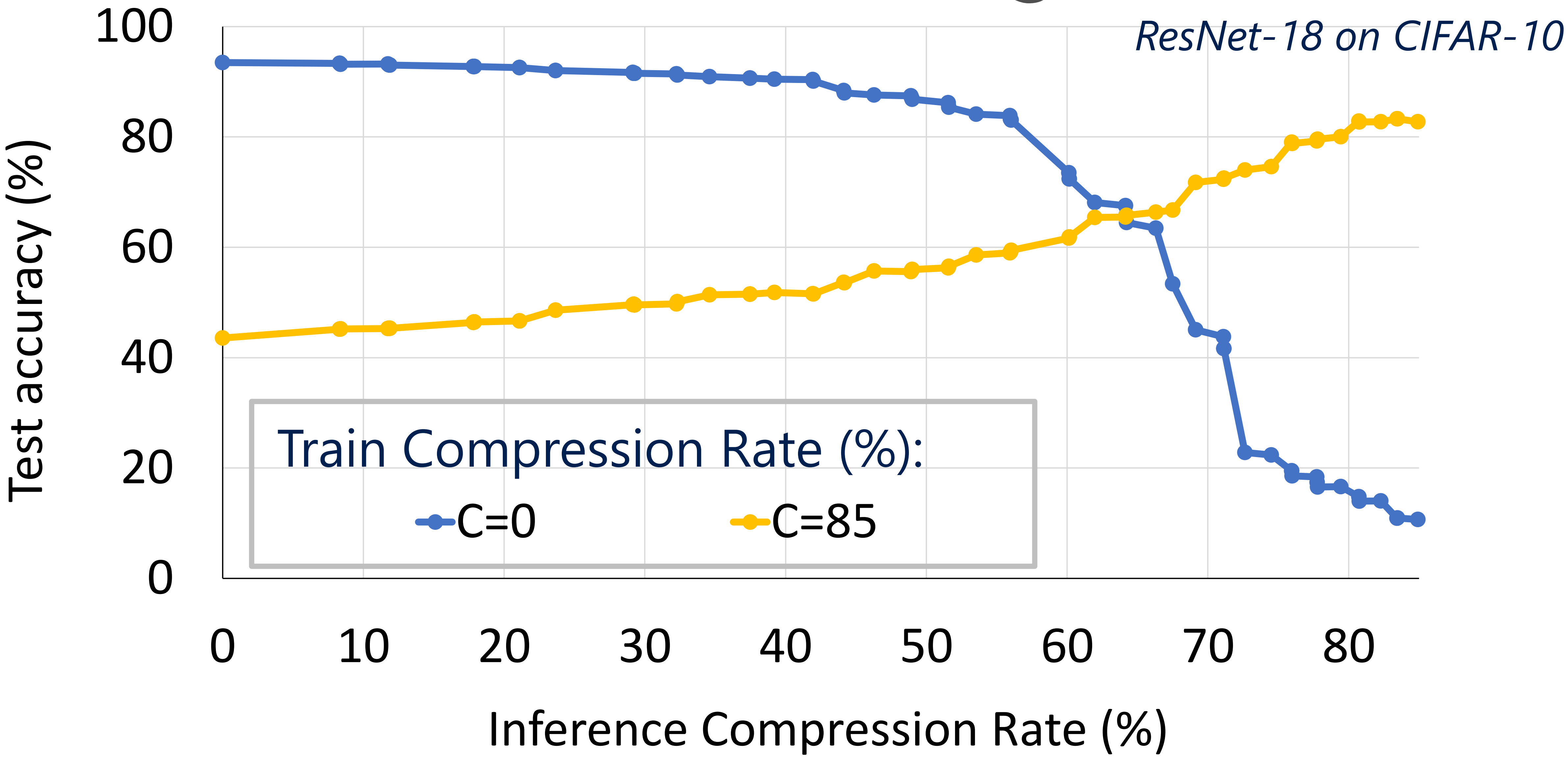
Compression Rate for **Training vs Inference**



Compression Rate for **Training** vs **Inference**

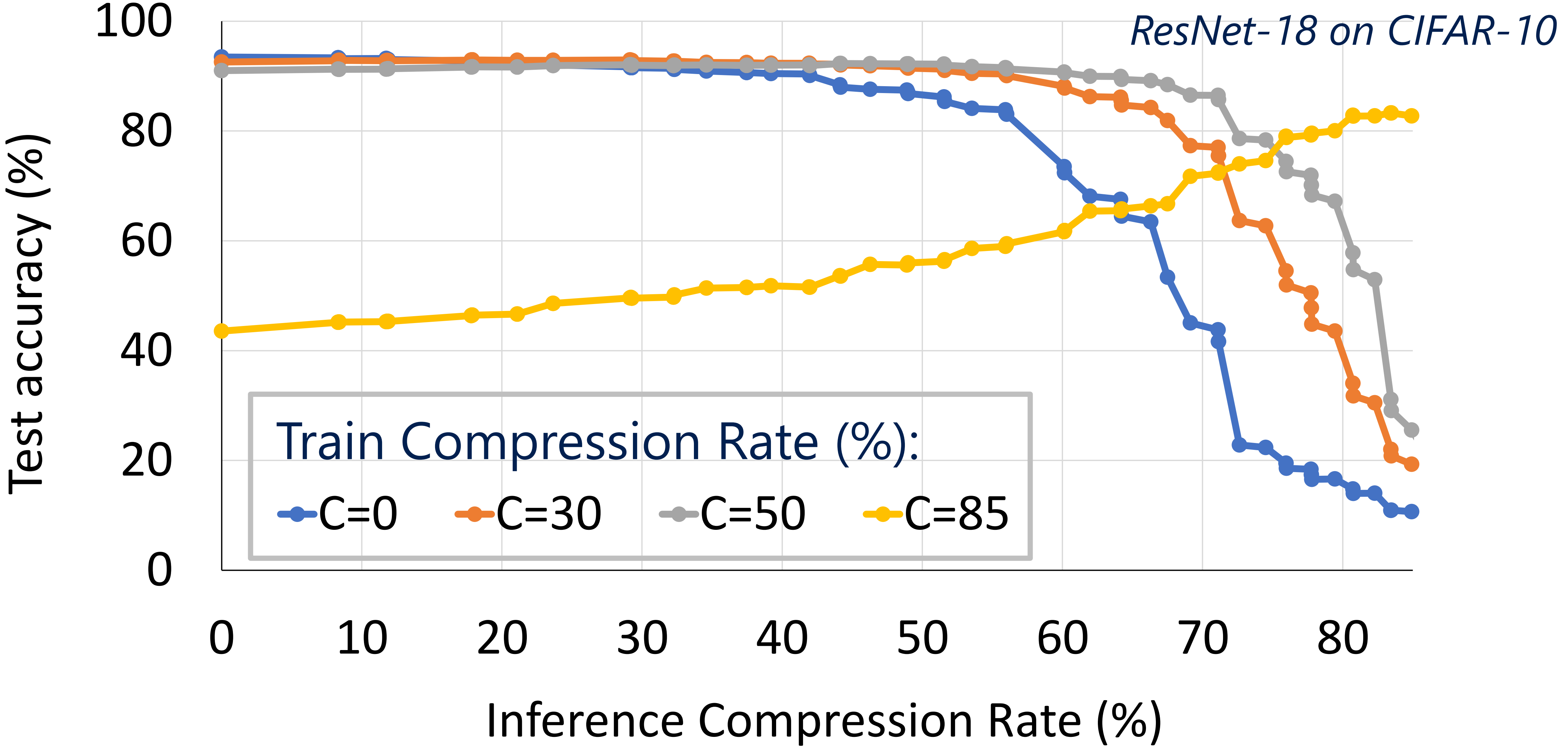


Compression Rate for **Training vs Inference**



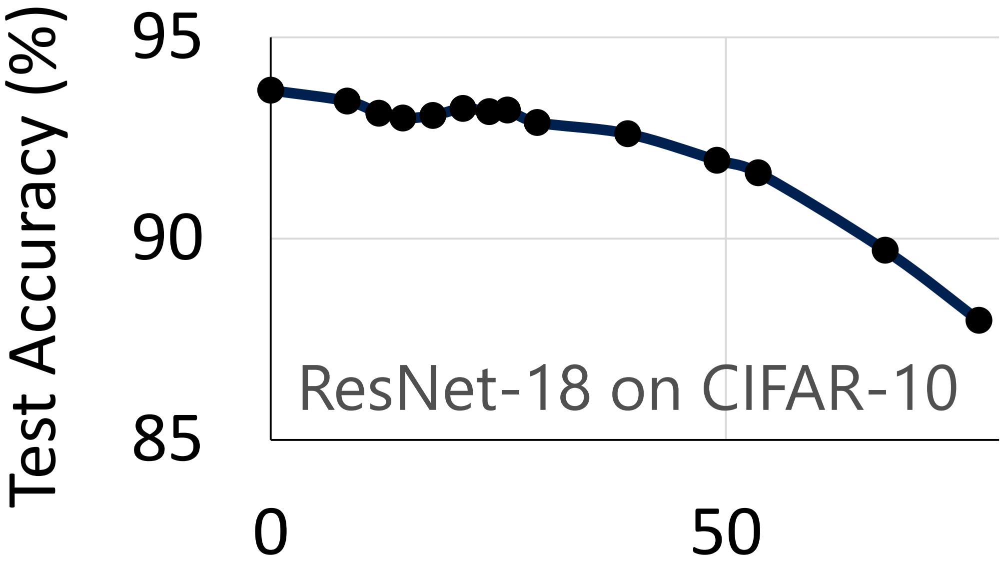
Smooth degradation of accuracy during inference

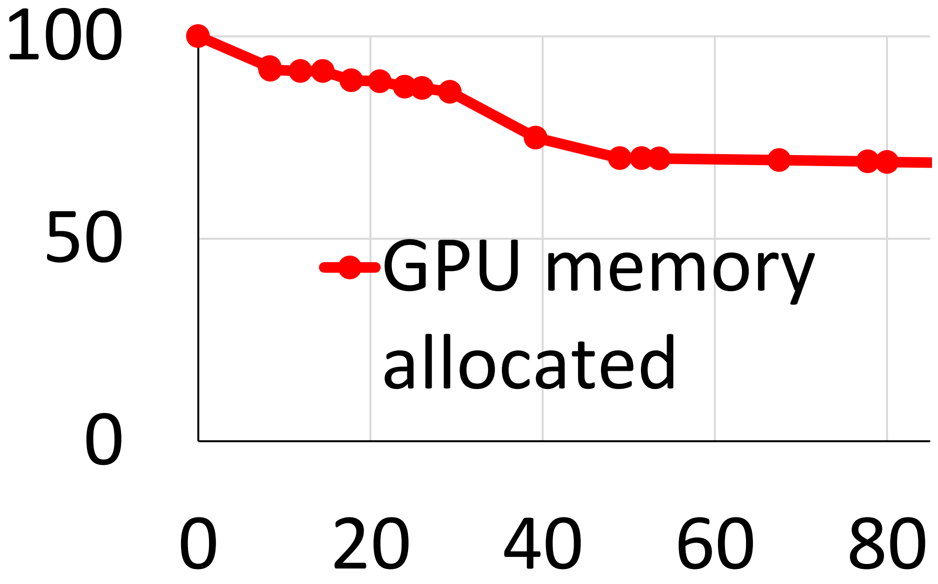
Compression Rate for **Training vs Inference**

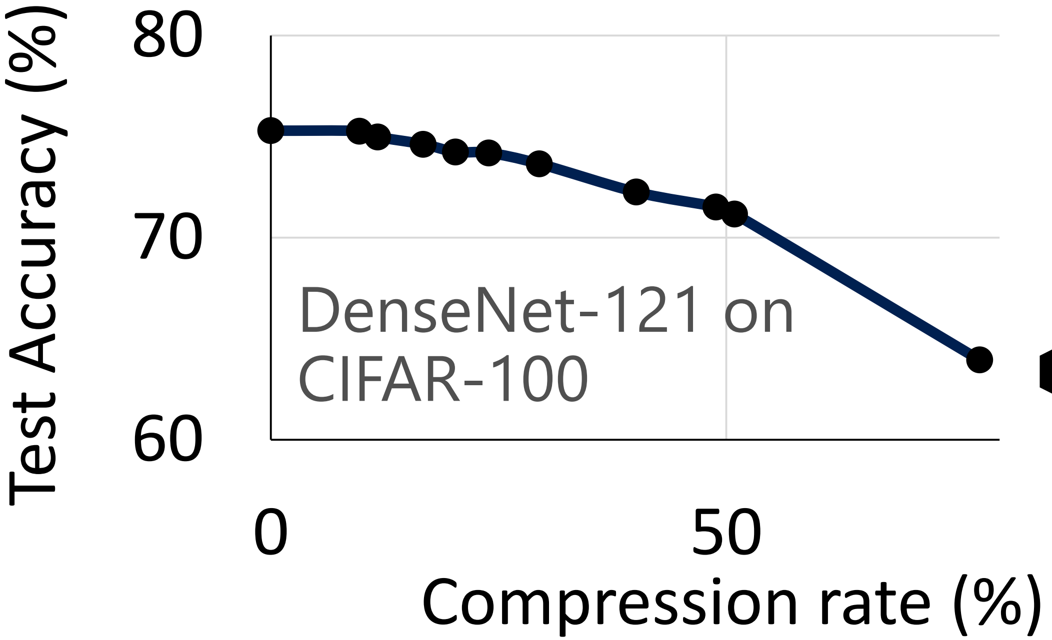


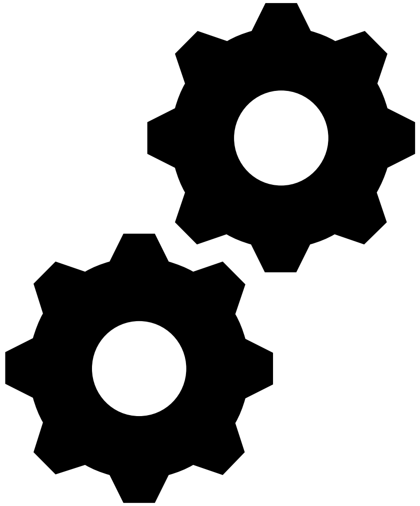
Apply the same compression rate to training and inference

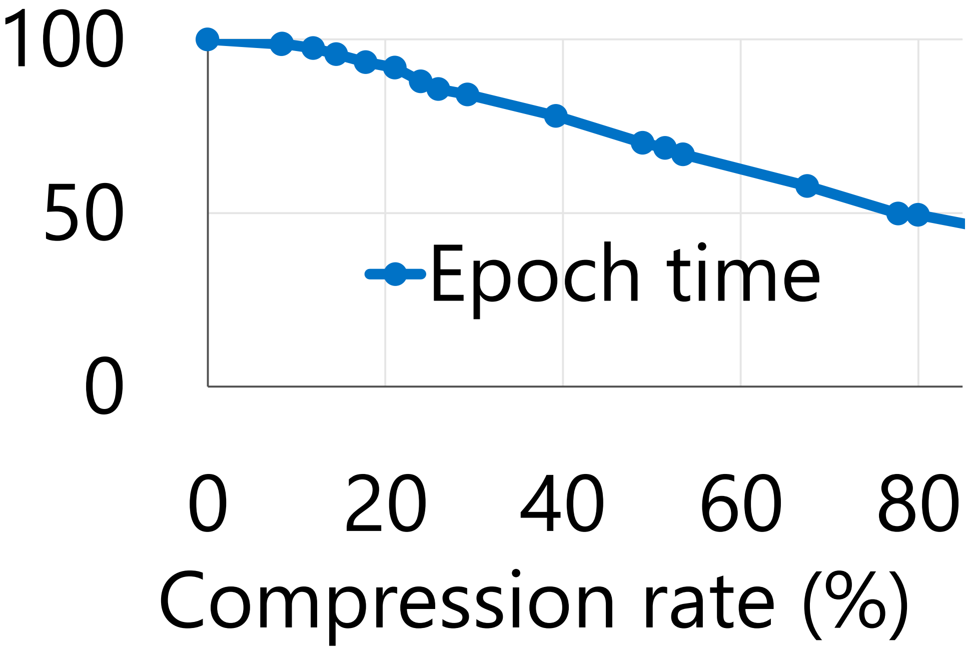
Tuning: Accuracy vs Higher Performance

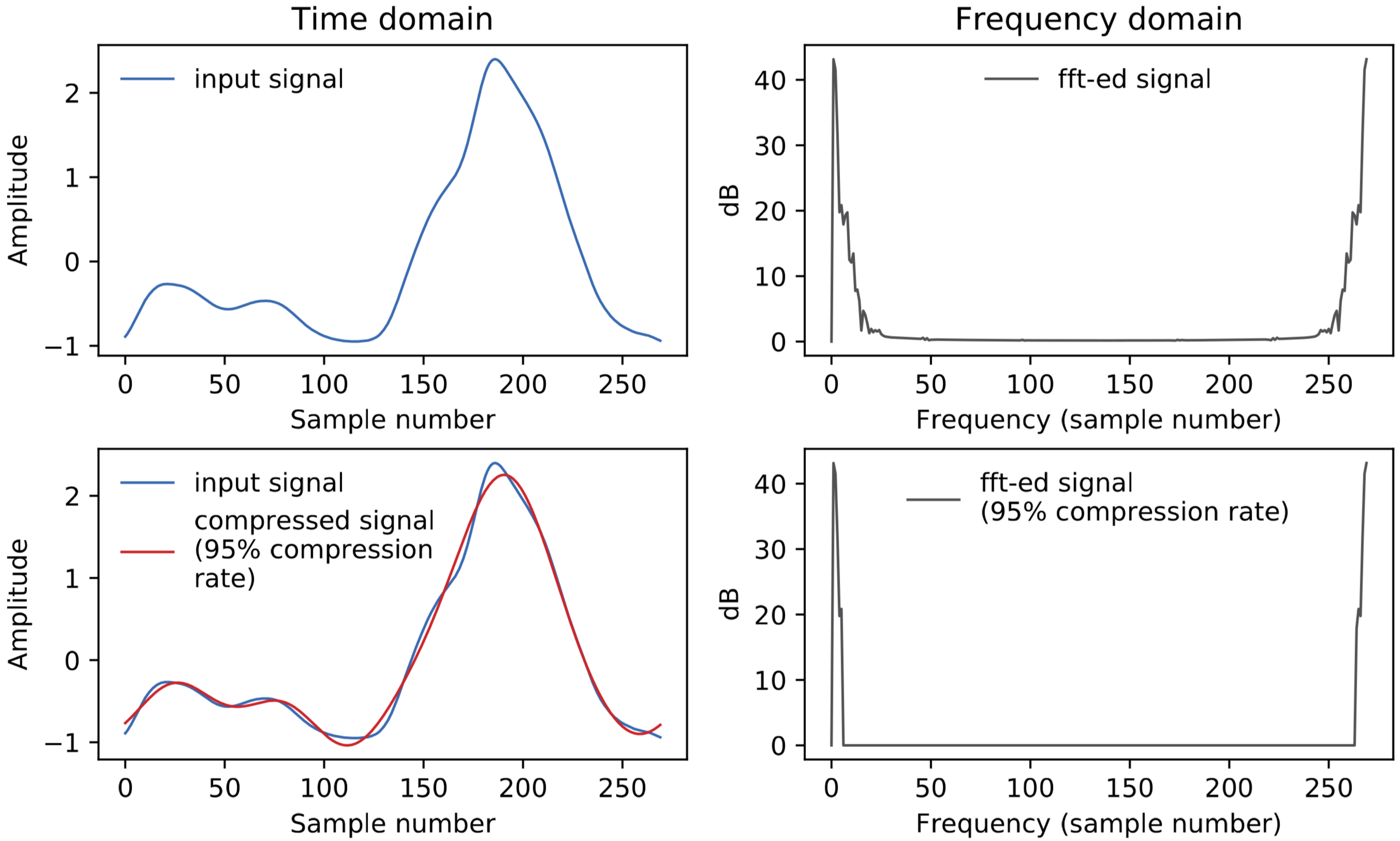




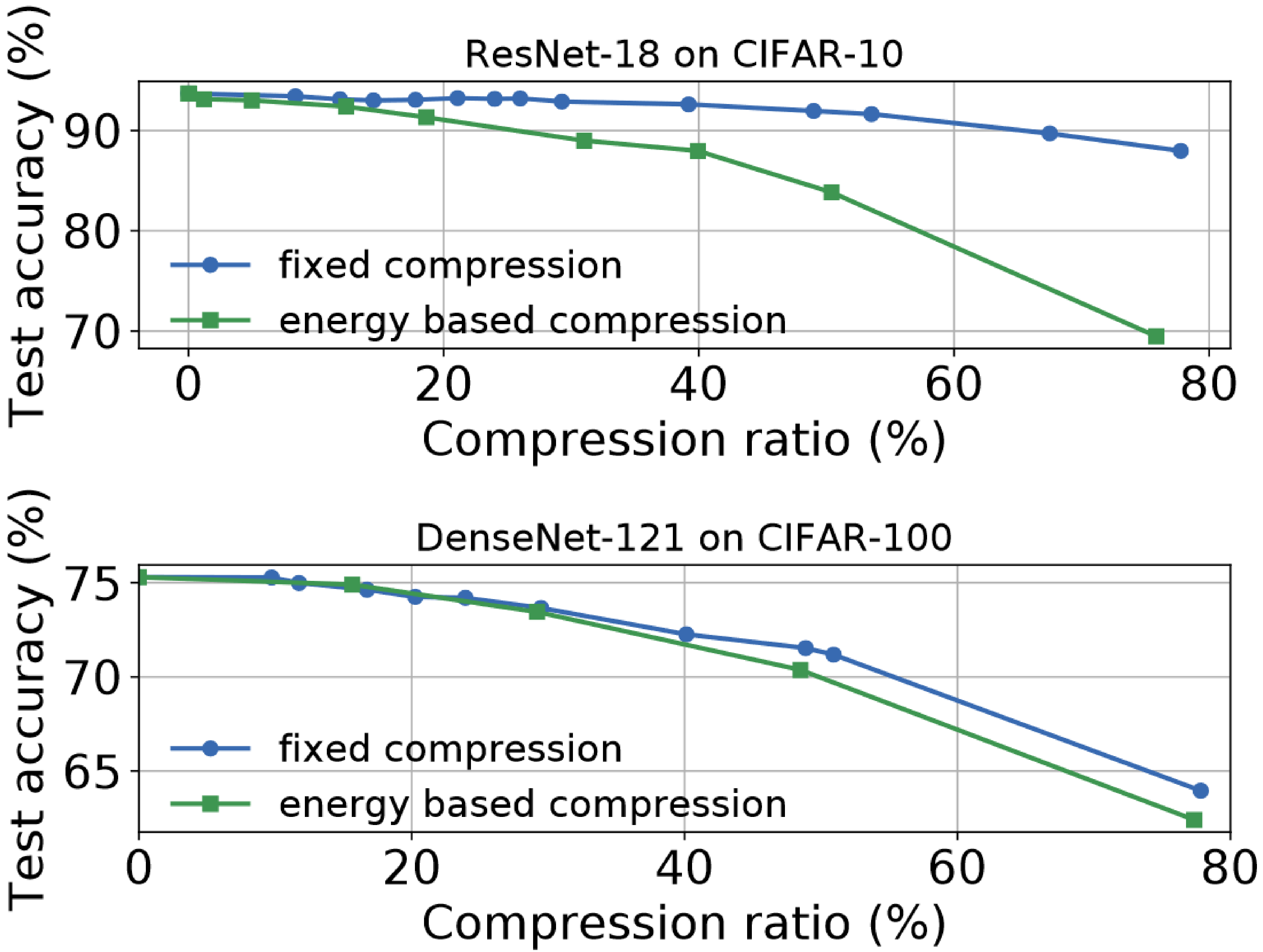


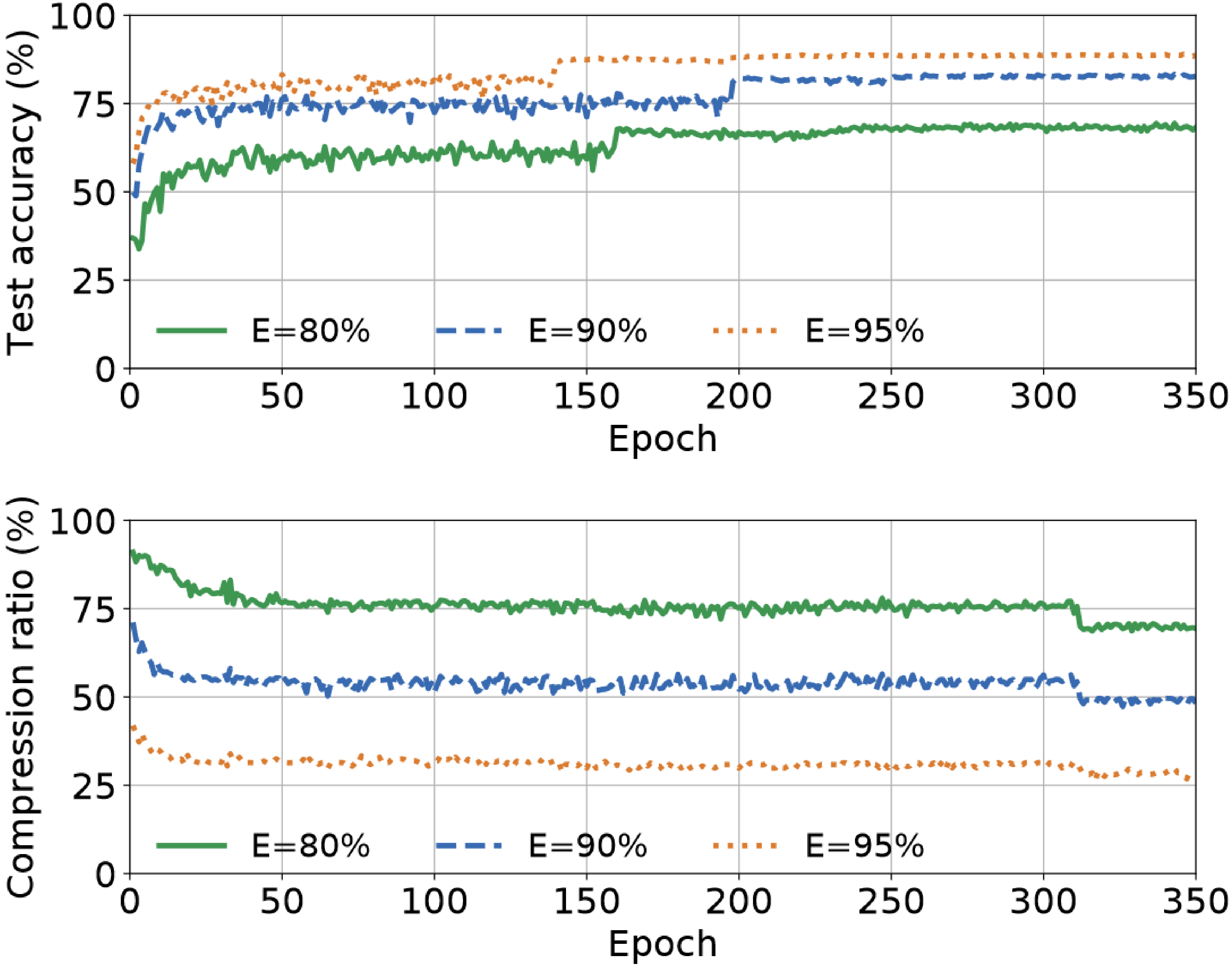


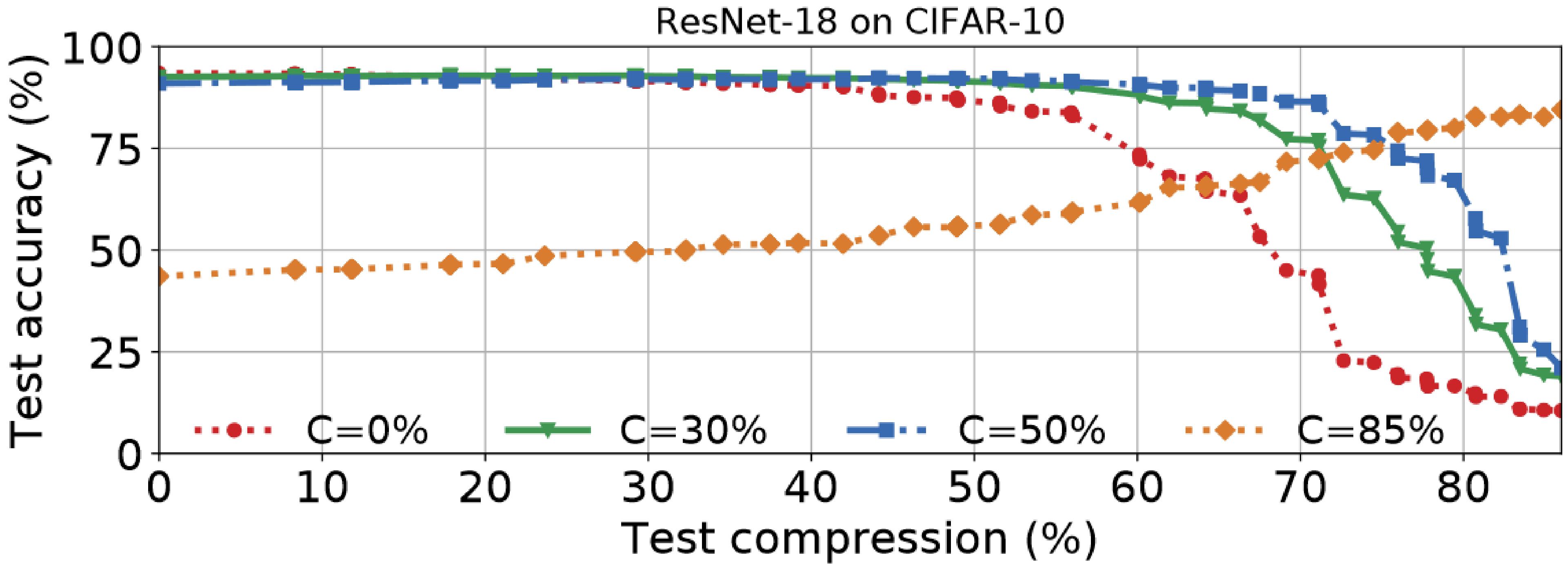


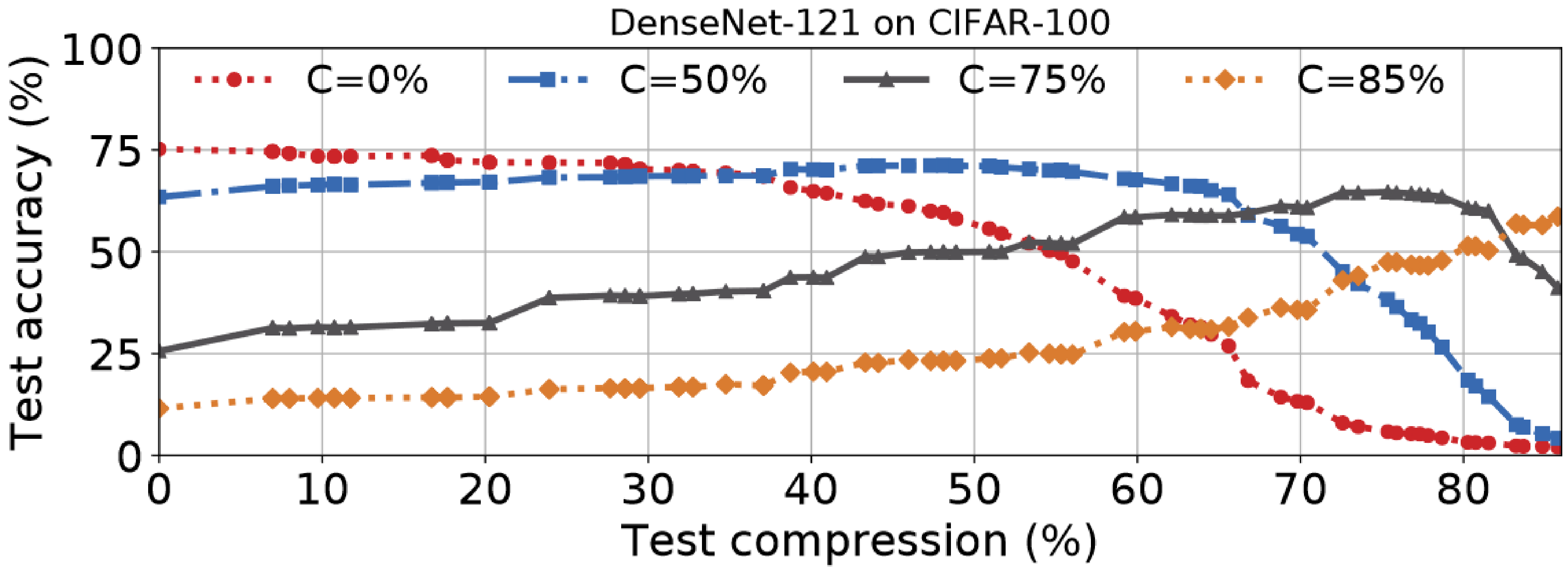


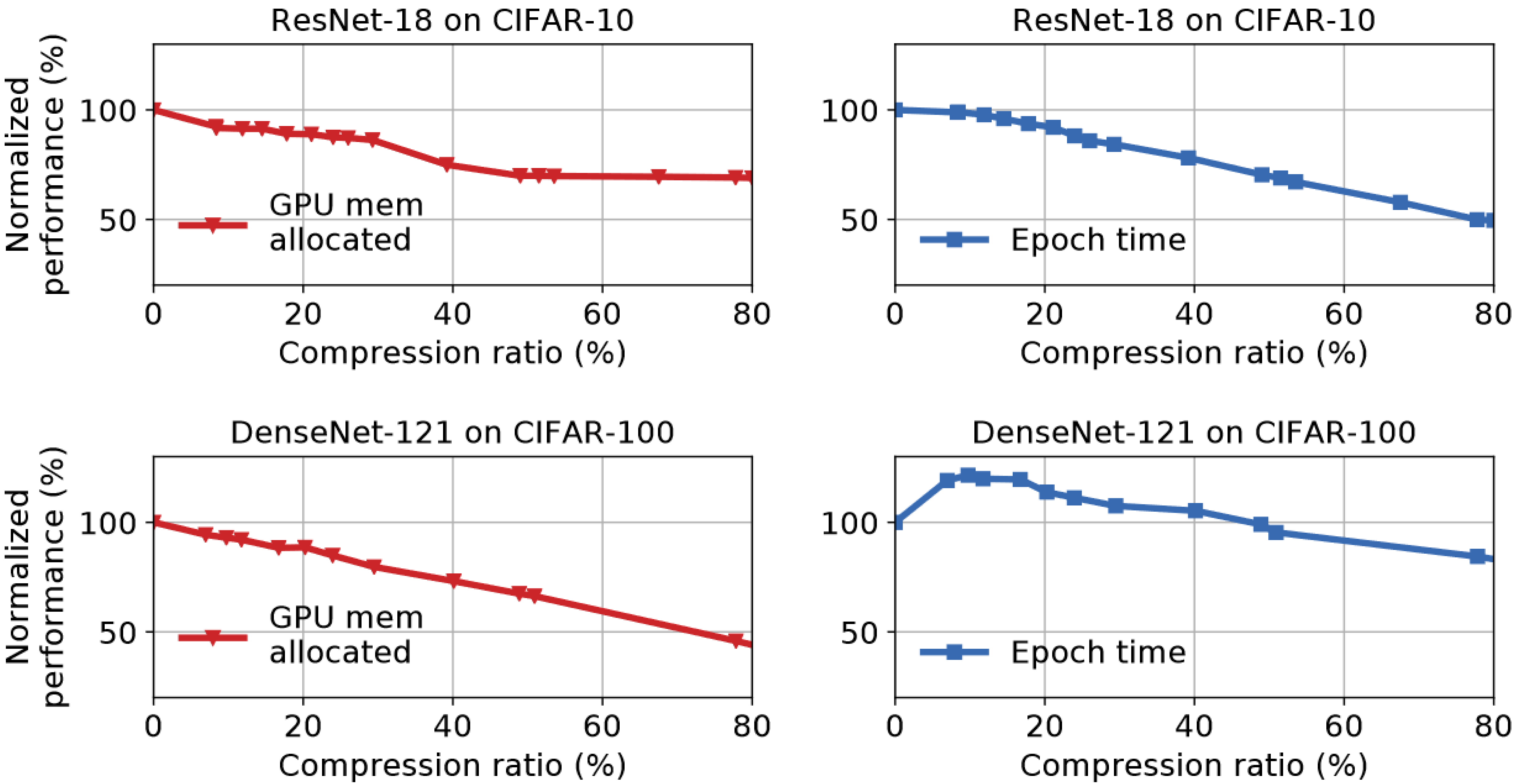
Time domain
Frequency domain
input signal
Amplitude
Sample number
fft-ed signal
dB
Frequency (sample number)
input signal
compressed signal (95% compression rate)
fft-ed signal (95% compression rate)


ResNet-18 on CIFAR-10
Test accuracy (%)
90
80
70
0
20
40
60
80
Compression ratio (%)
fixed compression
energy based compression
DenseNet-121 on CIFAR-100
Test accuracy (%)
75
70
65
0
20
40
60
80
Compression ratio (%)
fixed compression
energy based compression


Test accuracy (%)
100
75
50
25
0
0
50
100
150
200
250
300
350
Epoch
E=80%
E=90%
E=95%
Compression ratio (%)
100
75
50
25
0
0
50
100
150
200
250
300
350
Epoch
E=80%
E=90%
E=95%


ResNet-18 on CIFAR-10
Test accuracy (%)
Test compression (%)
C=0%
C=30%
C=50%
C=85%
0
25
50
75
100
0
10
20
30
40
50
60
70
80


DenseNet-121 on CIFAR-100
Test accuracy (%)
Test compression (%)
C=0%
C=50%
C=75%
C=85%
0
25
50
75
100
0
10
20
30
40
50
60
70
80


ResNet-18 on CIFAR-10
Normalized performance (%)
GPU mem allocated
Compression ratio (%)
ResNet-18 on CIFAR-10
Epoch time
Compression ratio (%)
DenseNet-121 on CIFAR-100
Normalized performance (%)
GPU mem allocated
Compression ratio (%)
DenseNet-121 on CIFAR-100
Epoch time
Compression ratio (%)

“Speaking of longer term, it would be nice if the community migrated to a fully open sourced implementation for all of this [convolution operations, etc.]. This stuff is just too important to the progress of the field for it to be **locked away in proprietary implementations**. The more people working together on this the better for everyone. There's plenty of room to compete on the hardware implementation side.”

Scott Gray

https://github.com/soumith/convnet-benchmarks/issues/93