Dynamic Sparse Training of Diagonally Sparse Networks

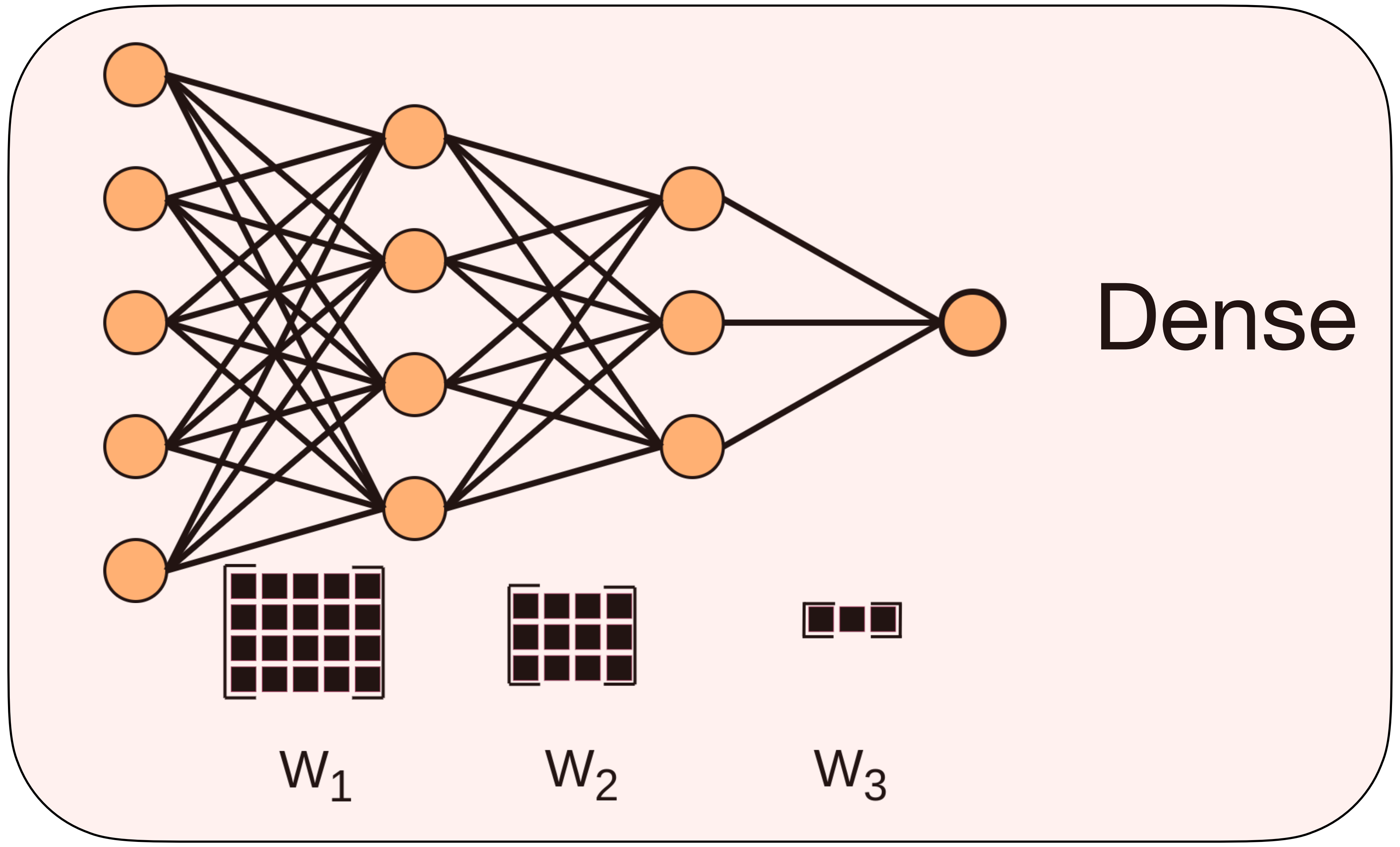
Abhishek Tyagi*, Arjun Iyer+, Christopher Kanan*, William H. Renninger+, and Yuhao Zhu*

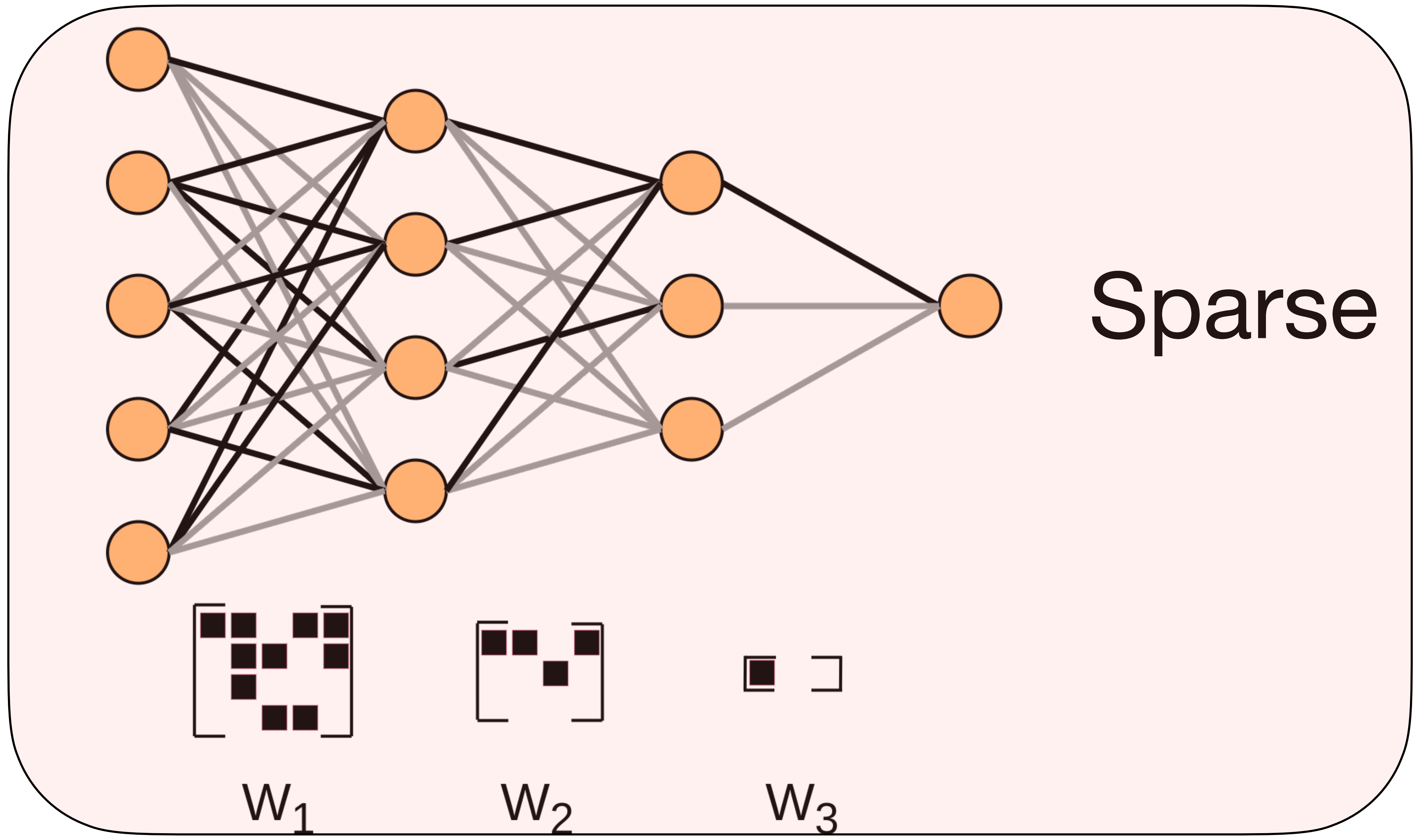
*Department of Computer Science

+The Institute of Optics

UNIVERSITY of ROCHESTER

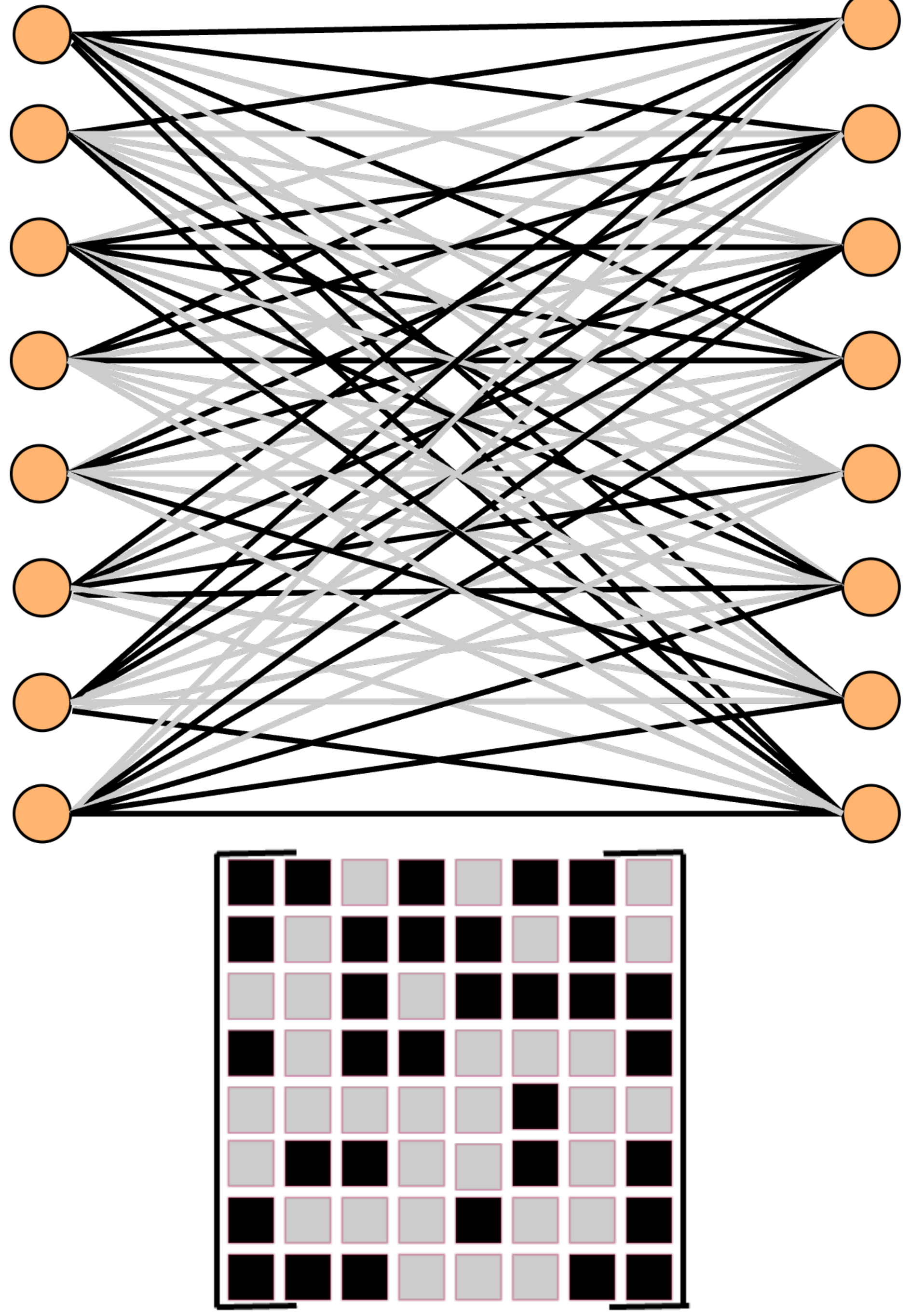
Motivation

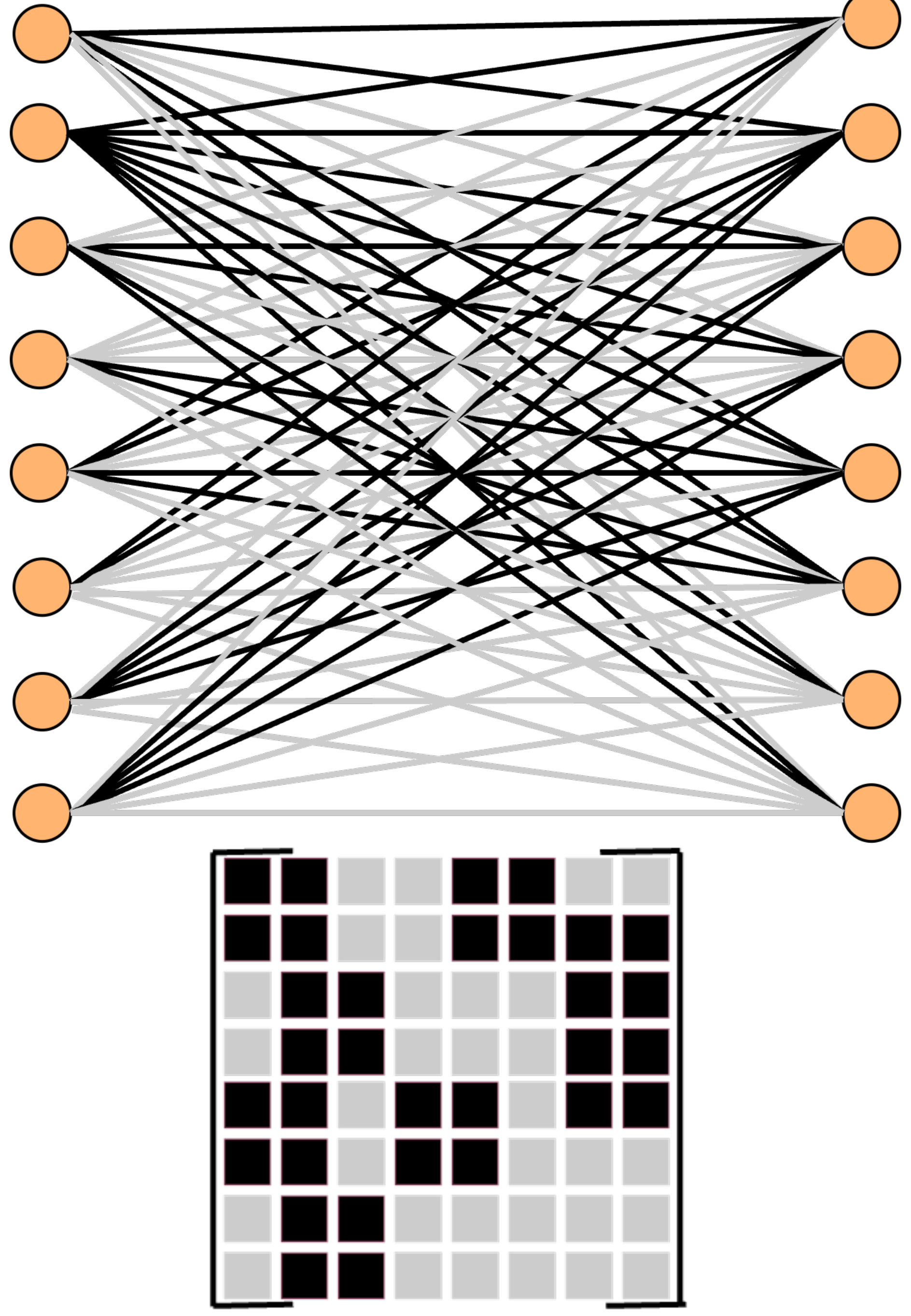


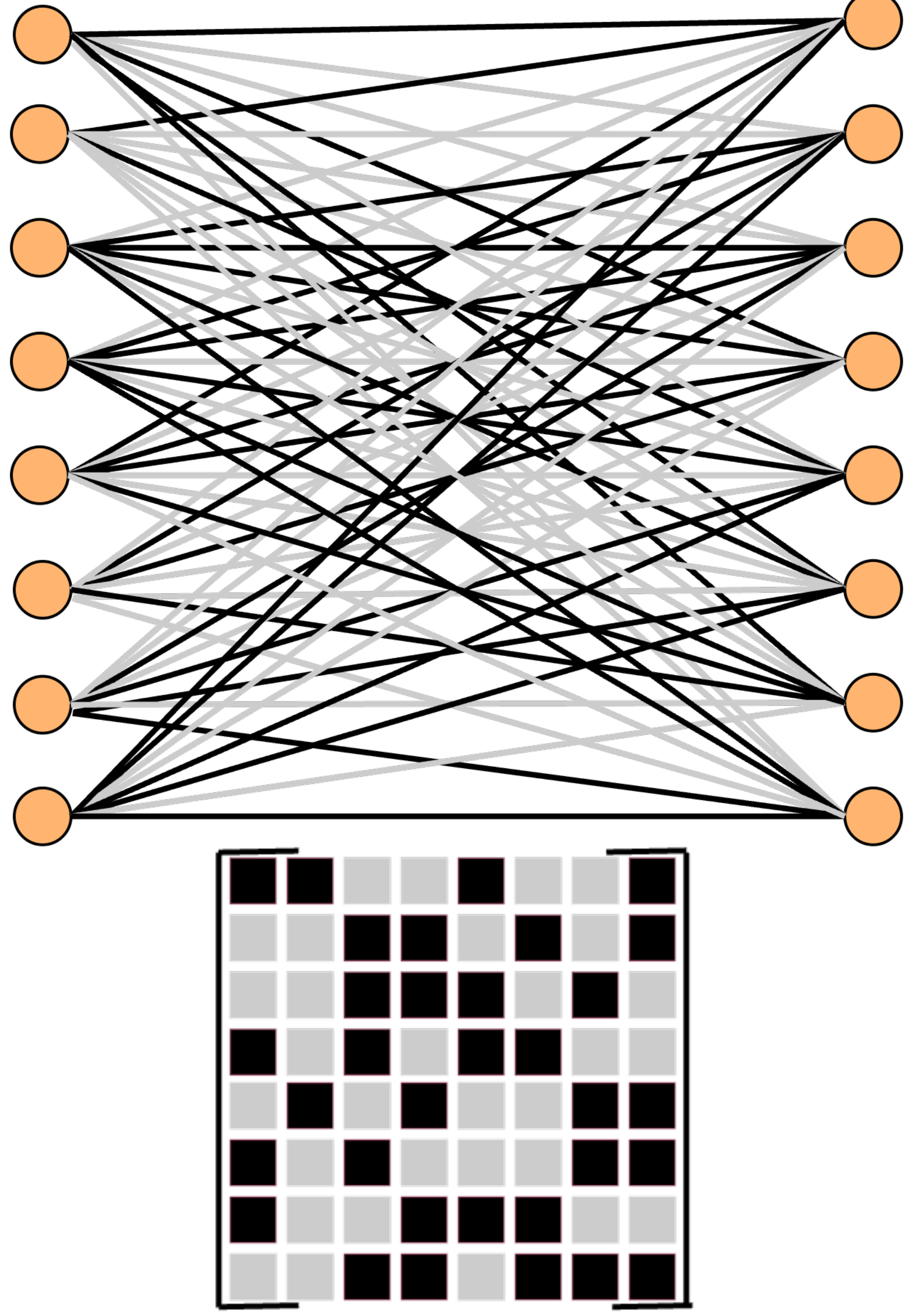


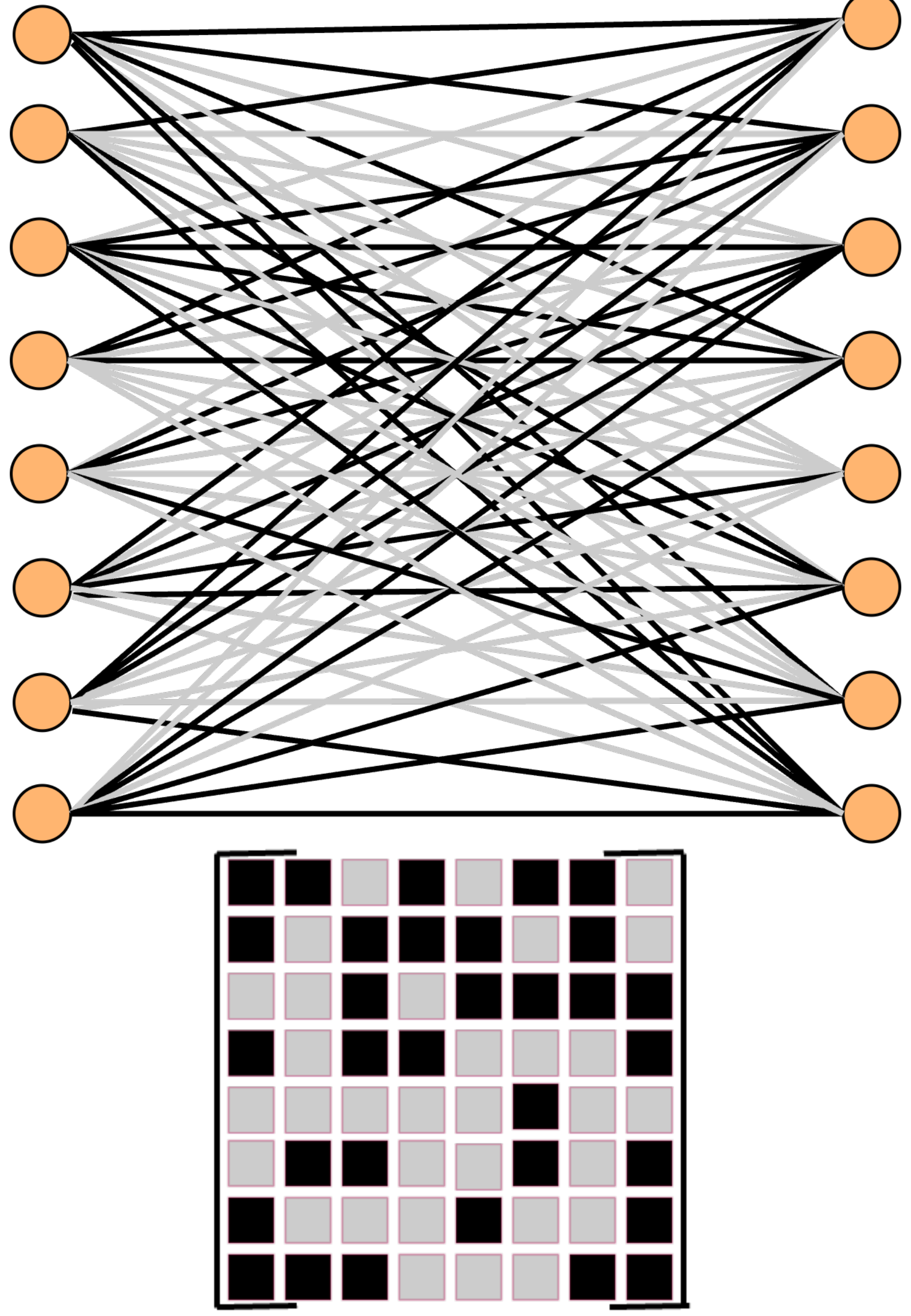
- **Training and Inference Acceleration**
  - Reduces the number of FLOPs without degrading accuracy
- **Reduces Network Size**
  - Leads to smaller memory footprint
- **Interestingly**
  - Better Generalization and Interpretability

Challenge: Speed vs Accuracy



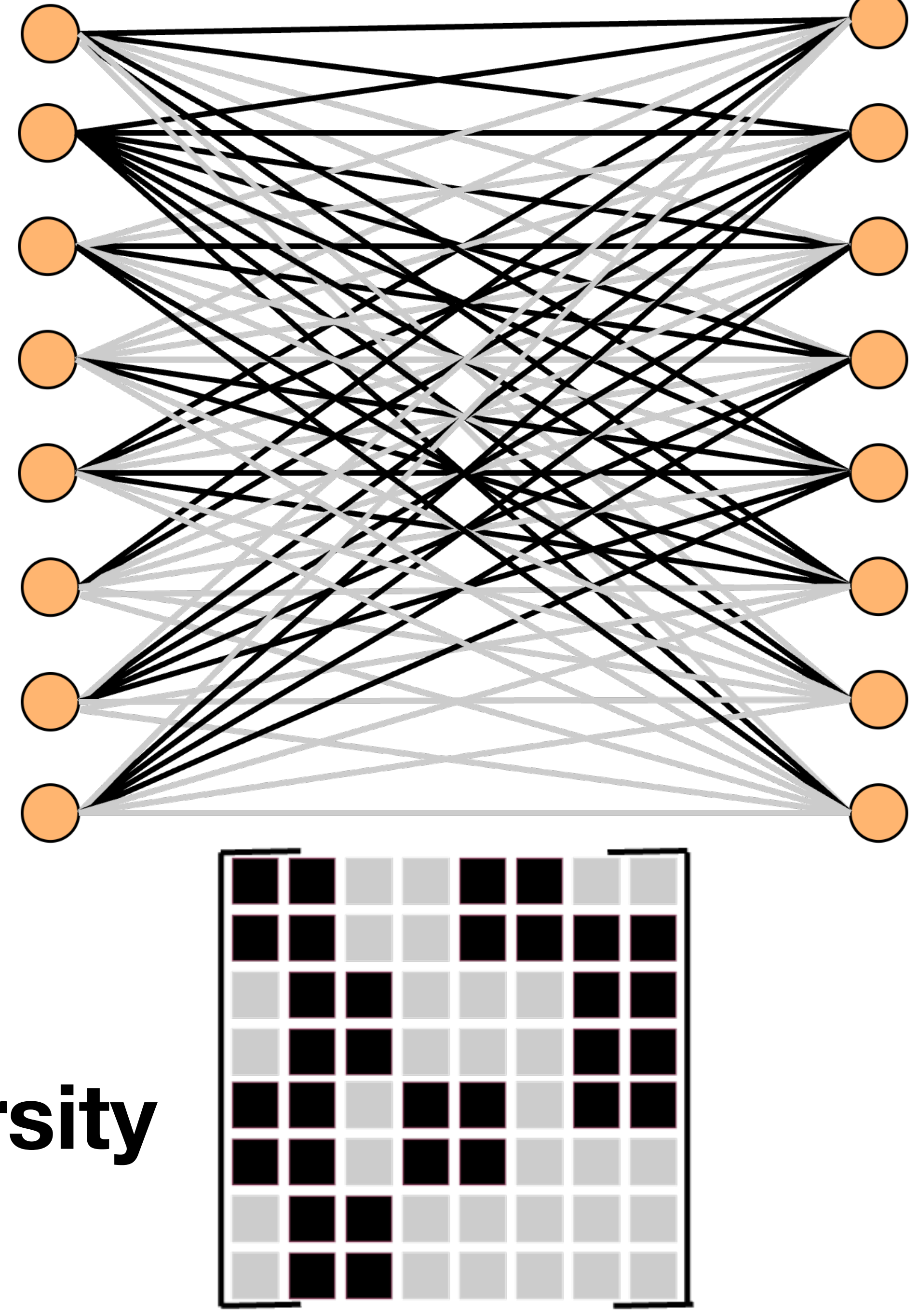


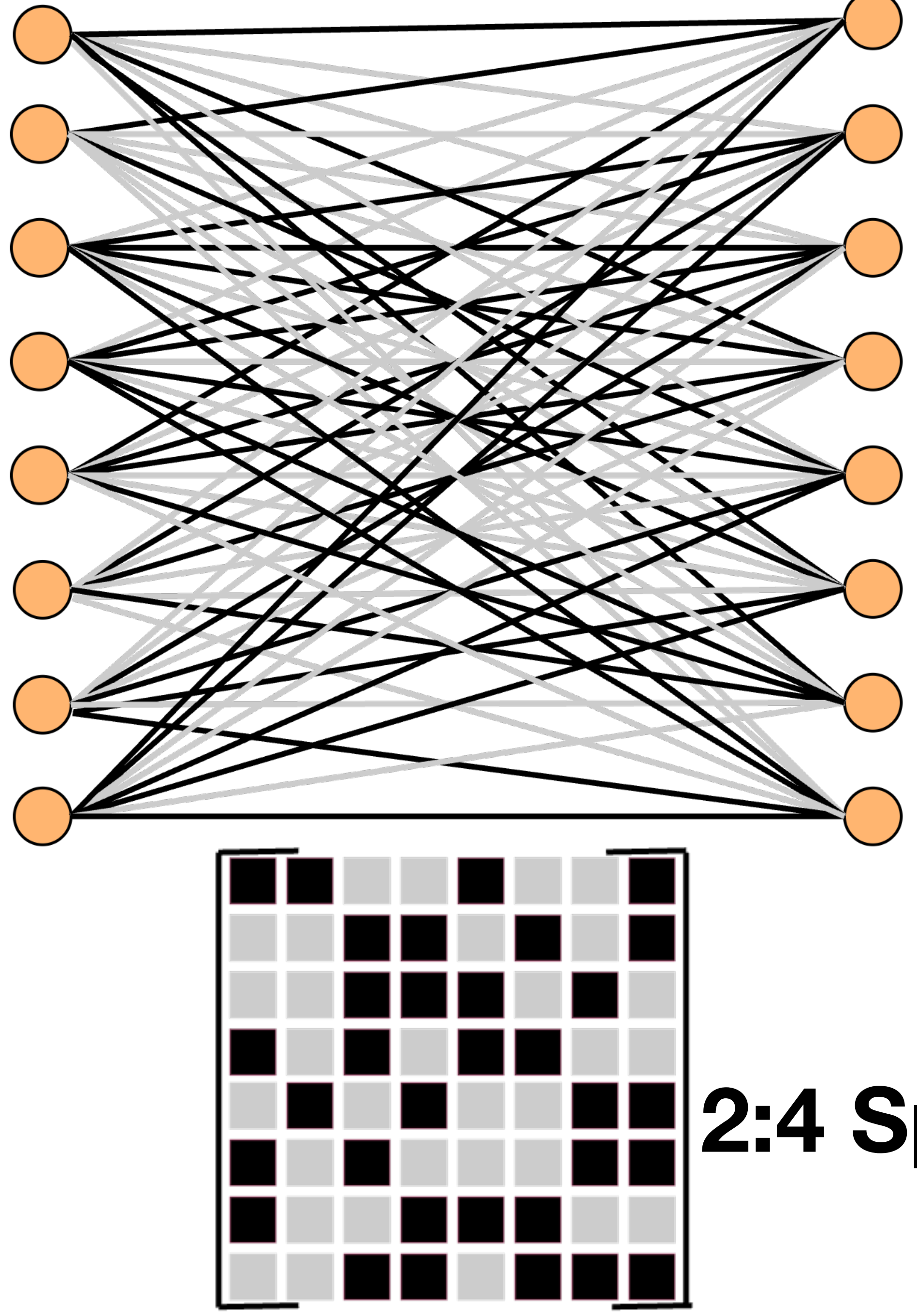




Unstructured

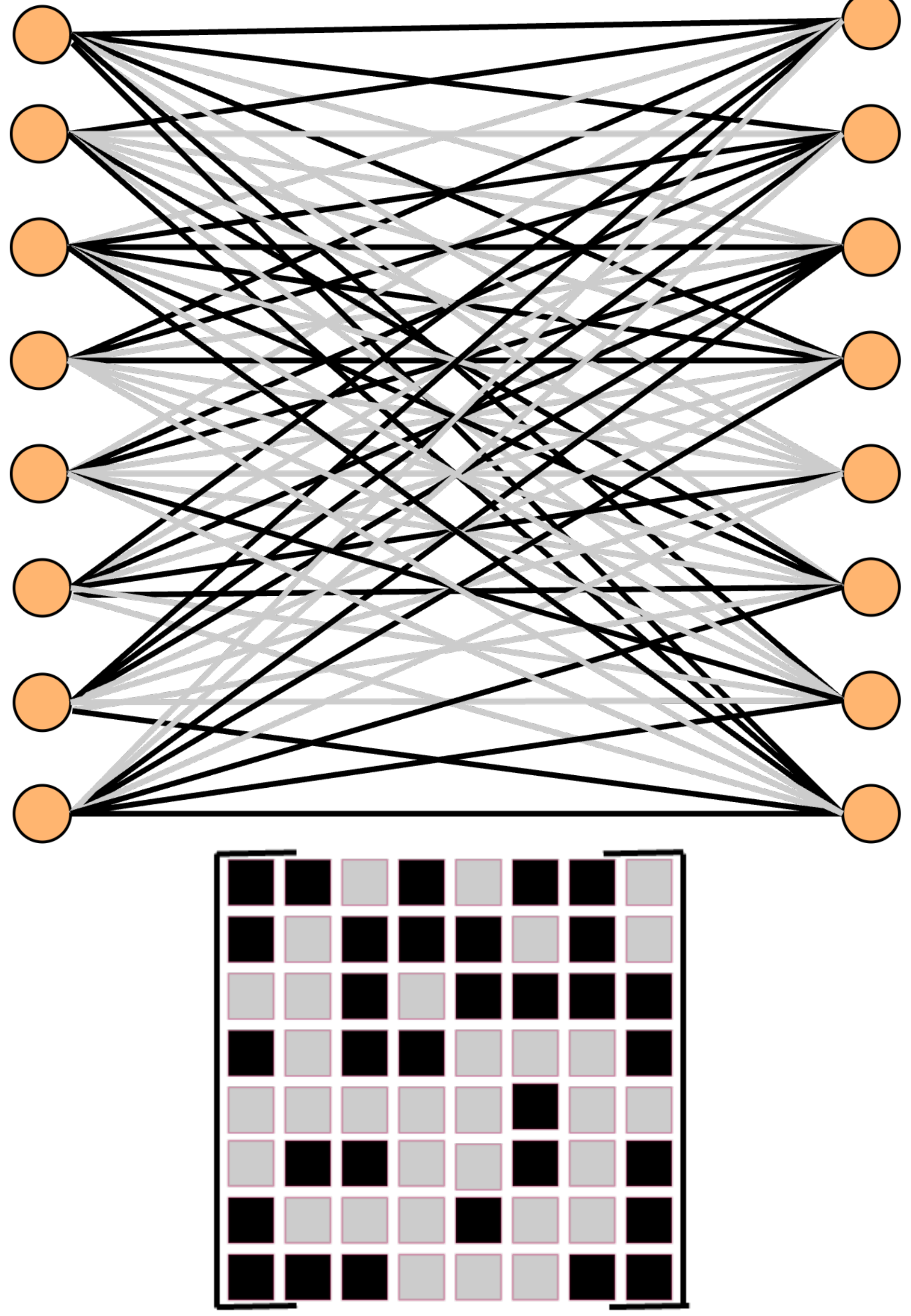
Block Sparsity





2:4 Sparsity

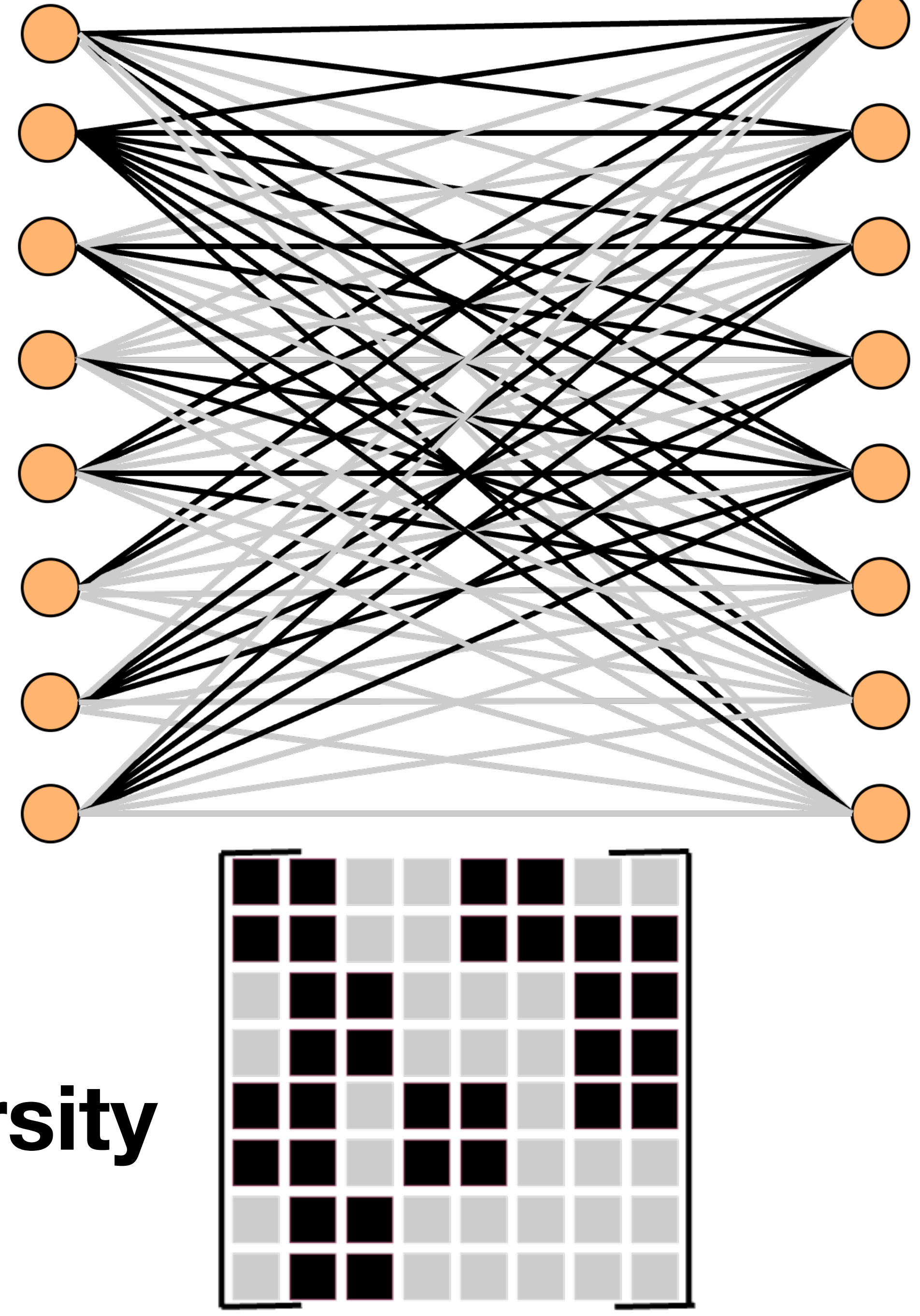
Structured

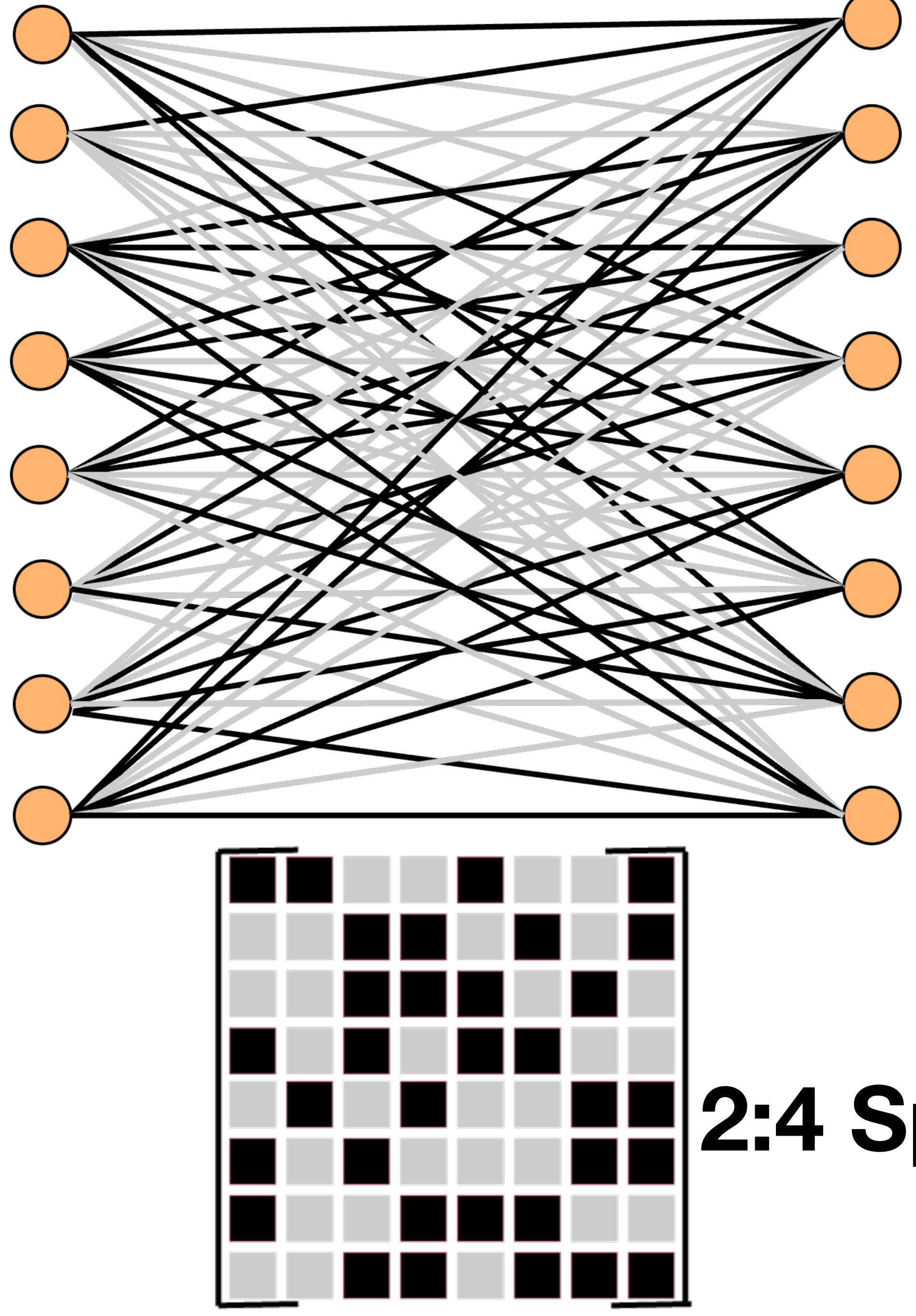


Unstructured

- Pros
  - $\text{Accuracy}_{\text{Unstruc}} \sim \text{Accuracy}_{\text{Dense}}$
  - Achieve high sparsity
- Cons
  - Hardware unfriendly

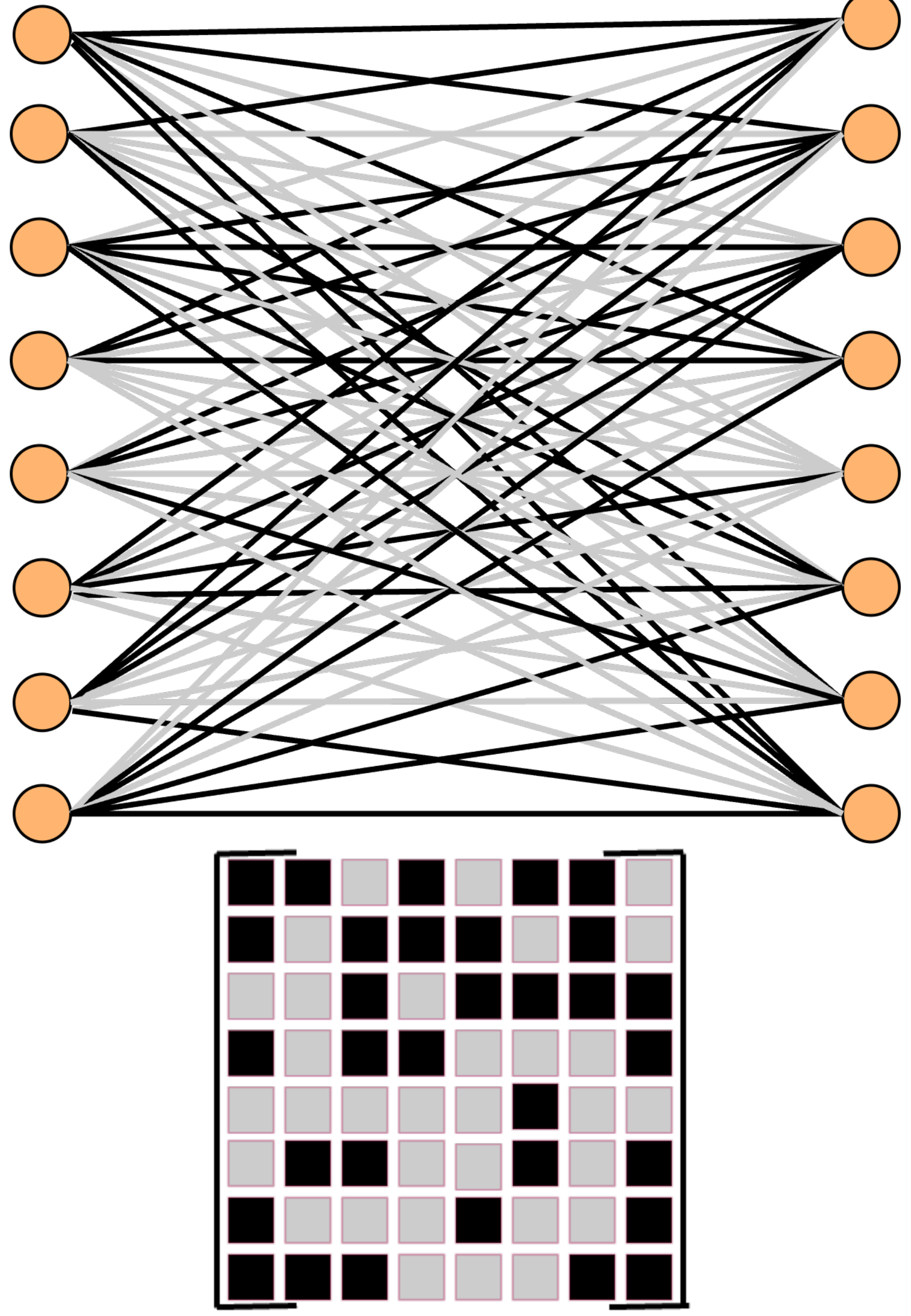
Block Sparsity





2:4 Sparsity

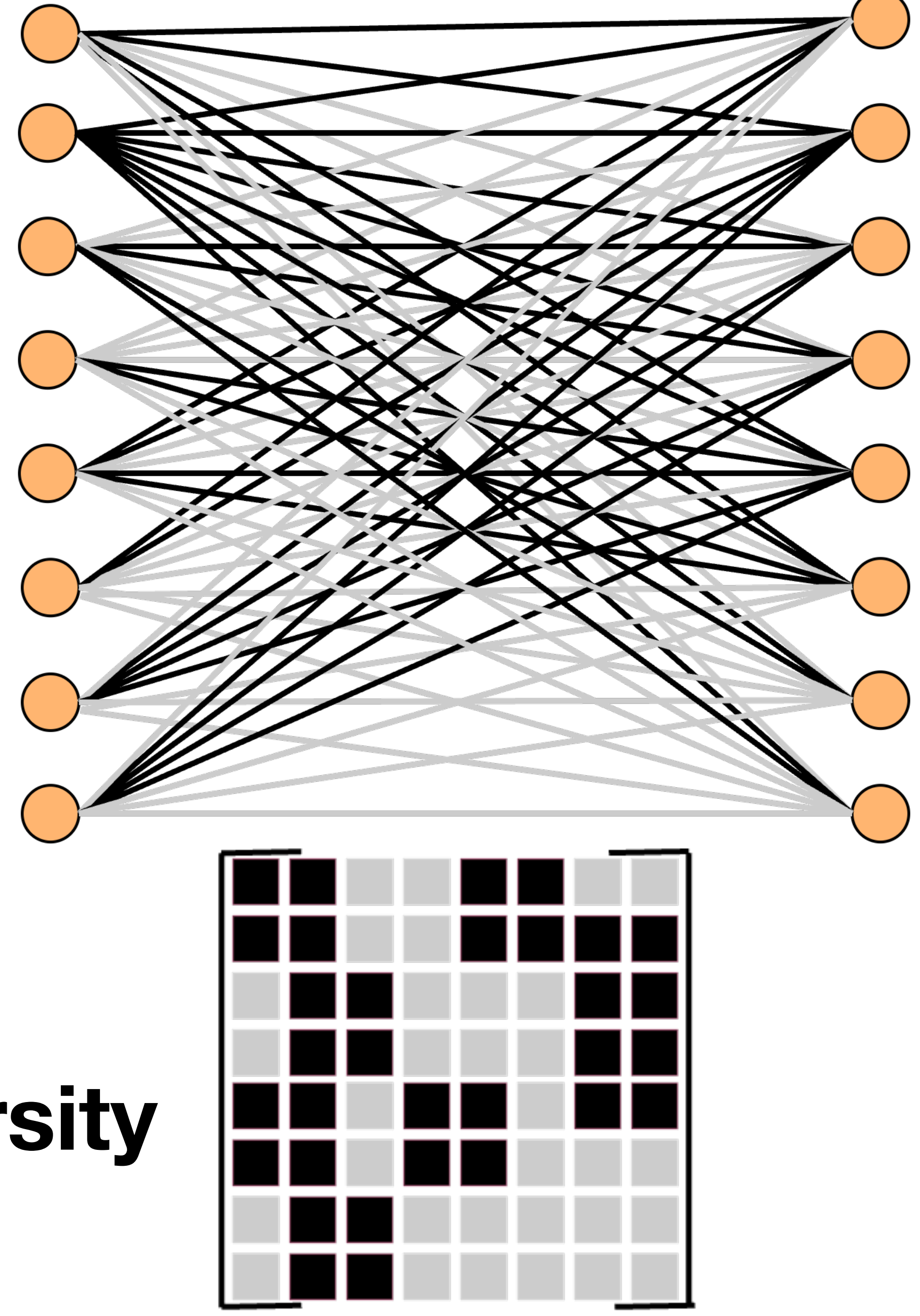
Structured

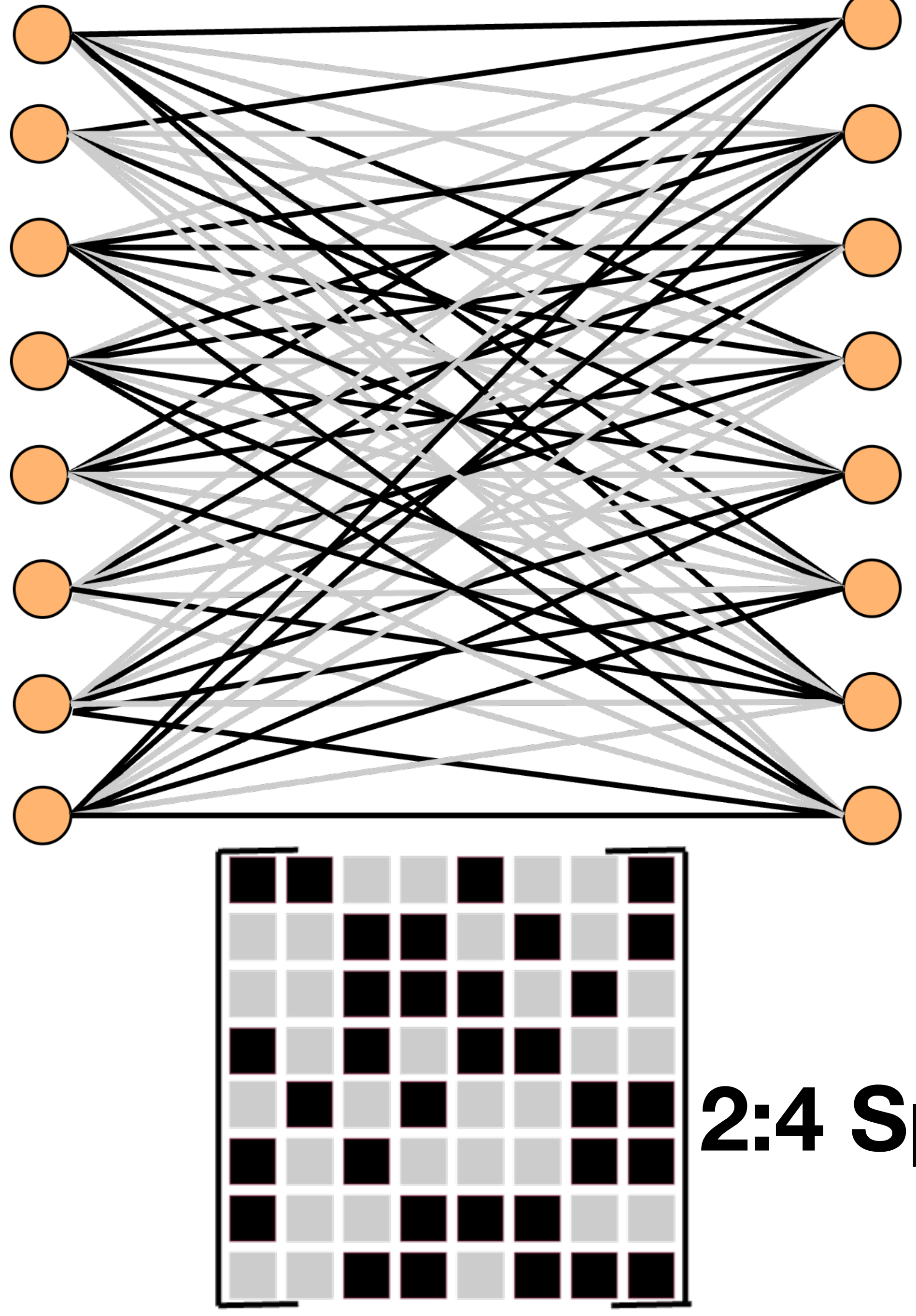


Unstructured

- Pros
  - $\text{Accuracy}_{\text{Unstruc}} \sim \text{Accuracy}_{\text{Dense}}$
  - Achieve high sparsity
- Cons
  - Hardware unfriendly

Block Sparsity

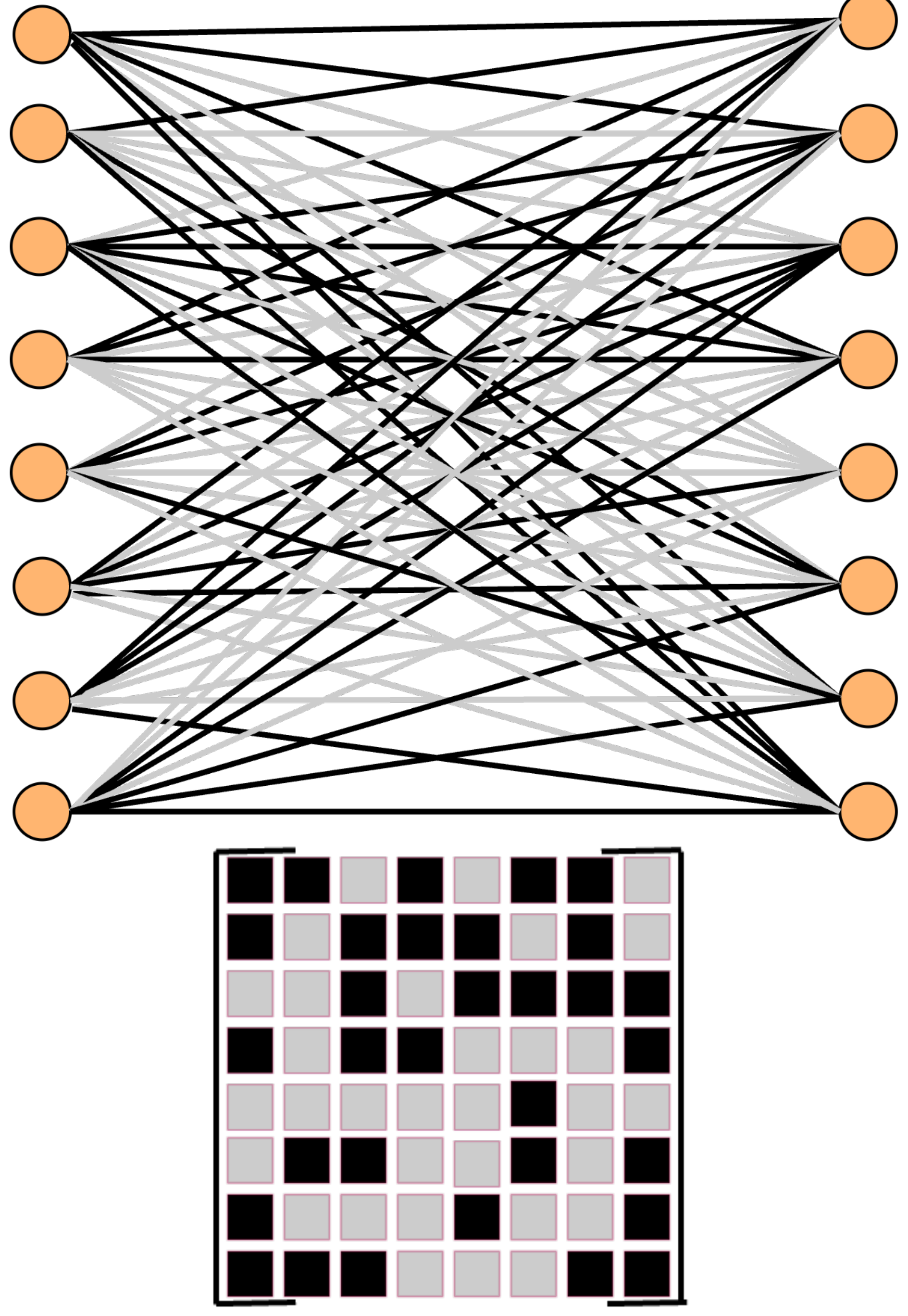




2:4 Sparsity

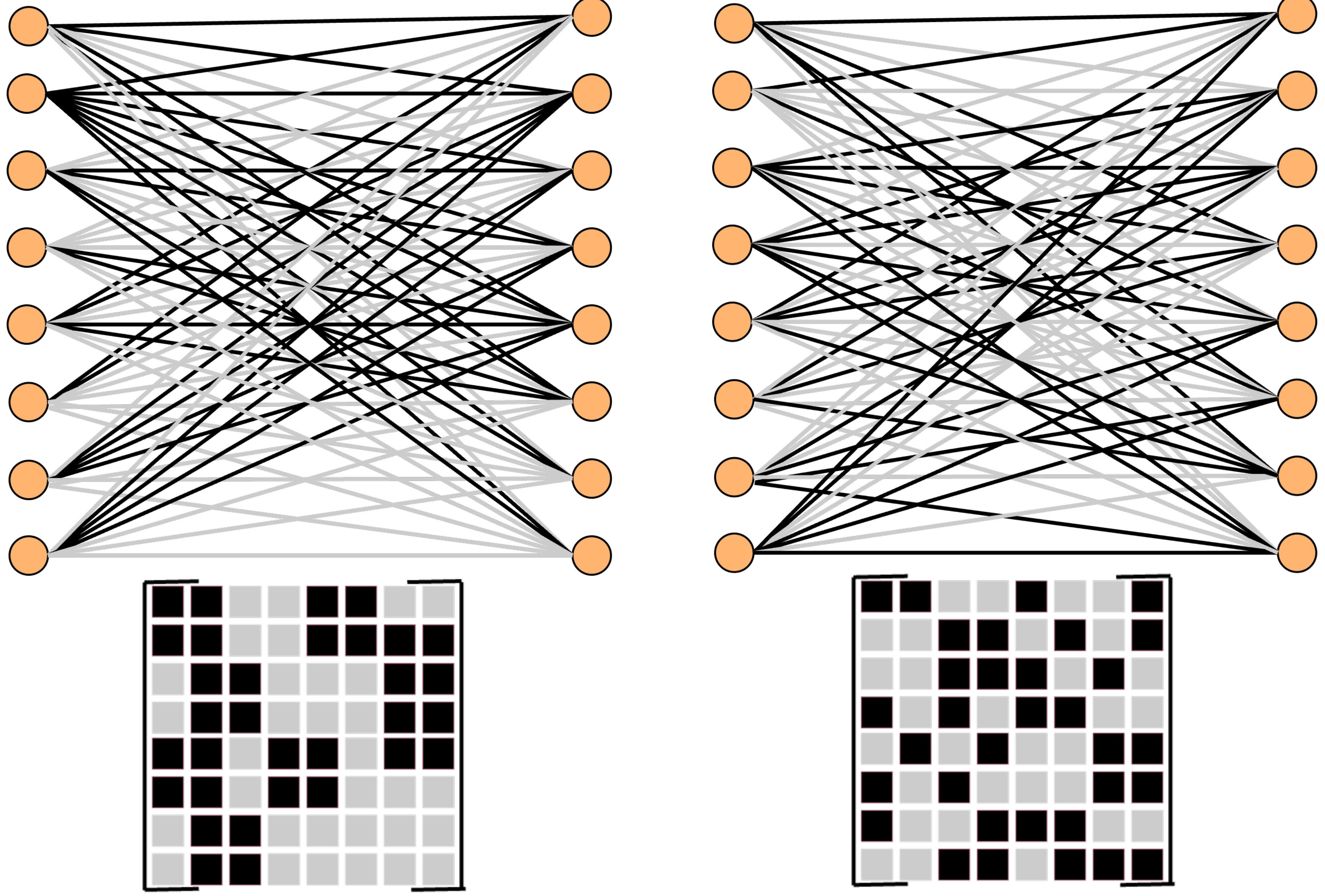
Structured

- Pros
  - Hardware friendly
- Cons
  - $\text{Accuracy}_{\text{Struc}} < \text{Accuracy}_{\text{Dense}}$
  - Degradation at high sparsity levels



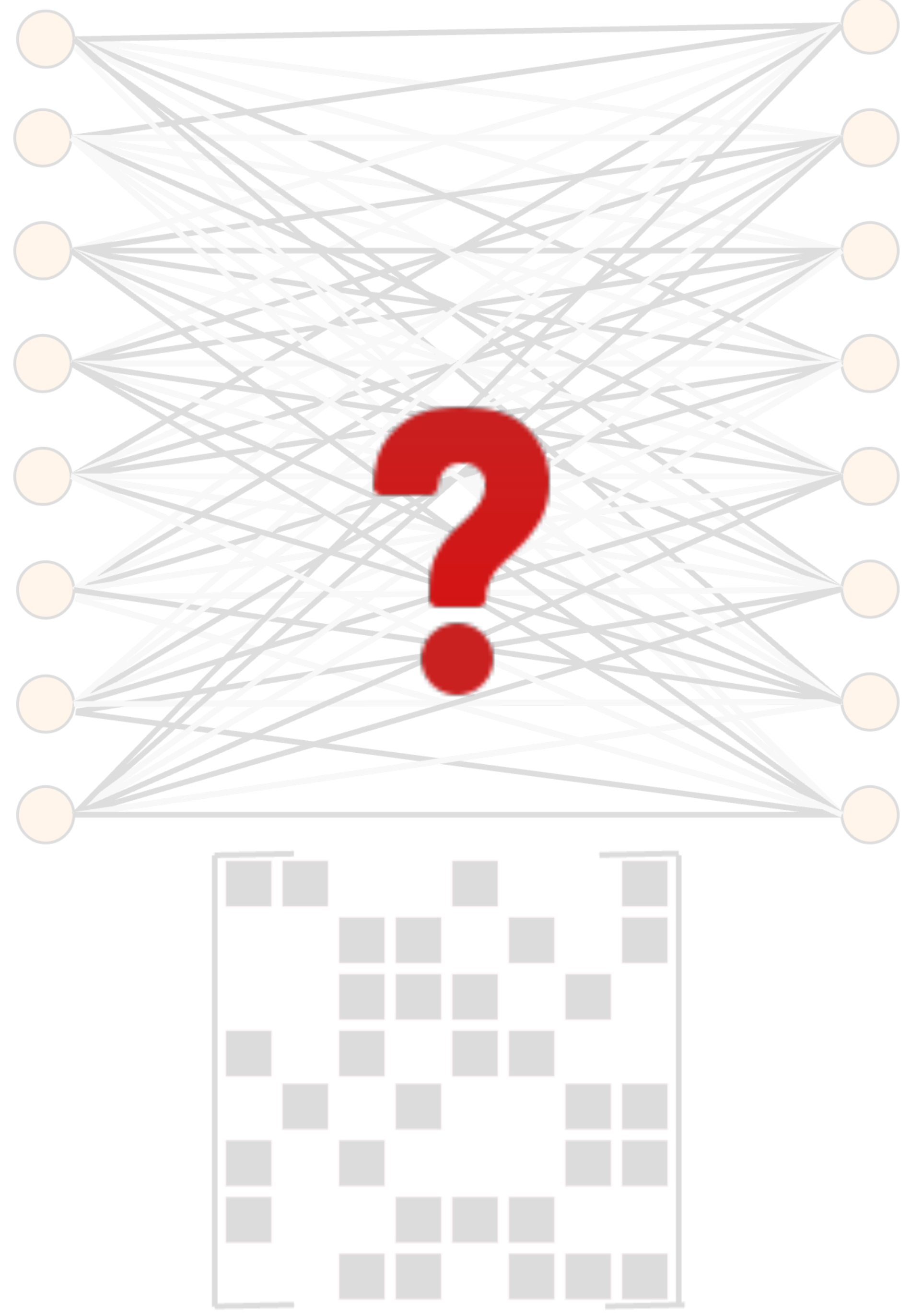
Unstructured

- Pros ✓
  - $\text{Accuracy}_{\text{Unstruc}} \sim \text{Accuracy}_{\text{Dense}}$
  - Achieve high sparsity
- Cons ✗
  - Hardware unfriendly

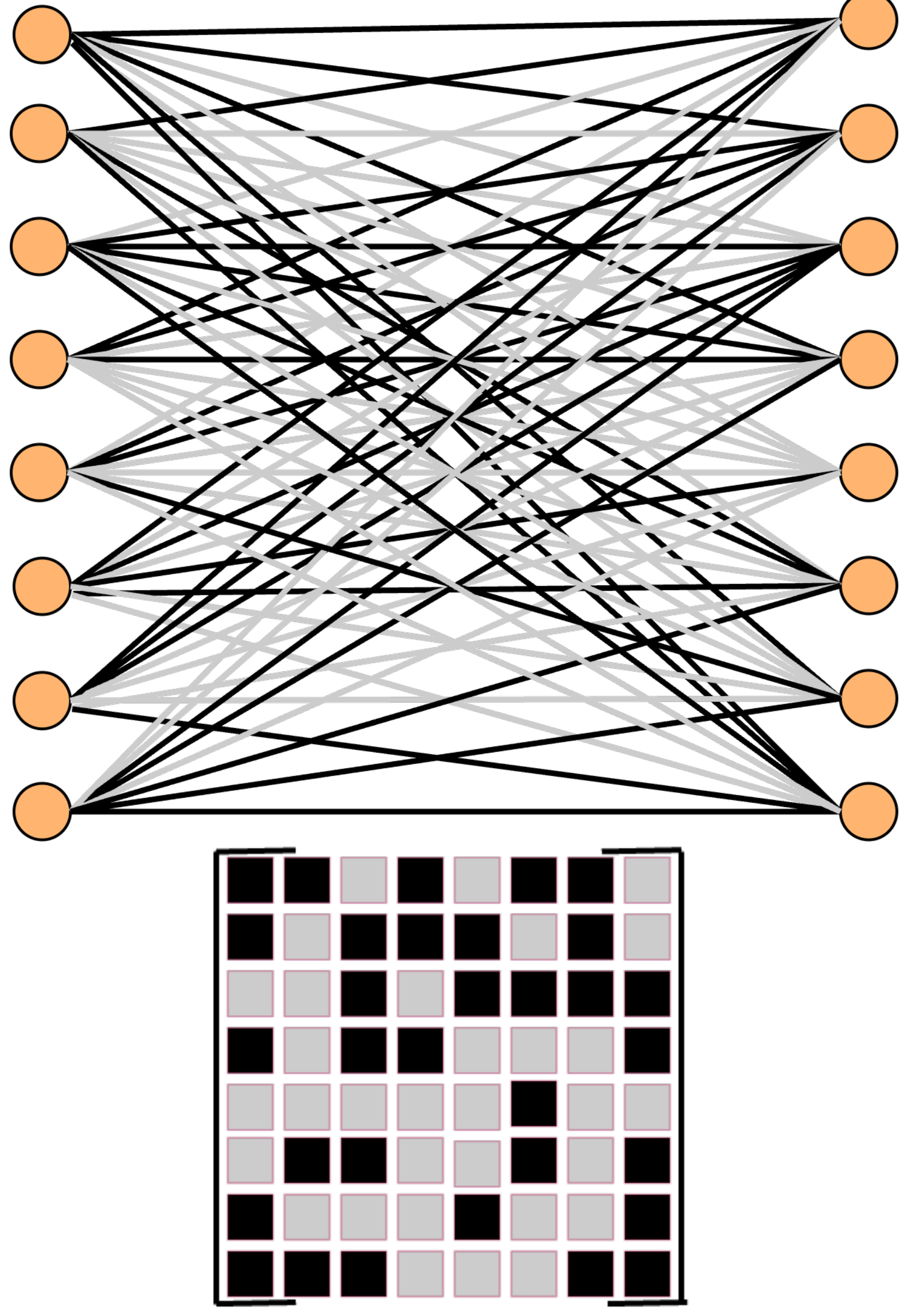


Structured

- Pros ✓
  - Hardware friendly
- Cons ✗
  - $\text{Accuracy}_{\text{Struc}} < \text{Accuracy}_{\text{Dense}}$
  - Degradation at high sparsity levels

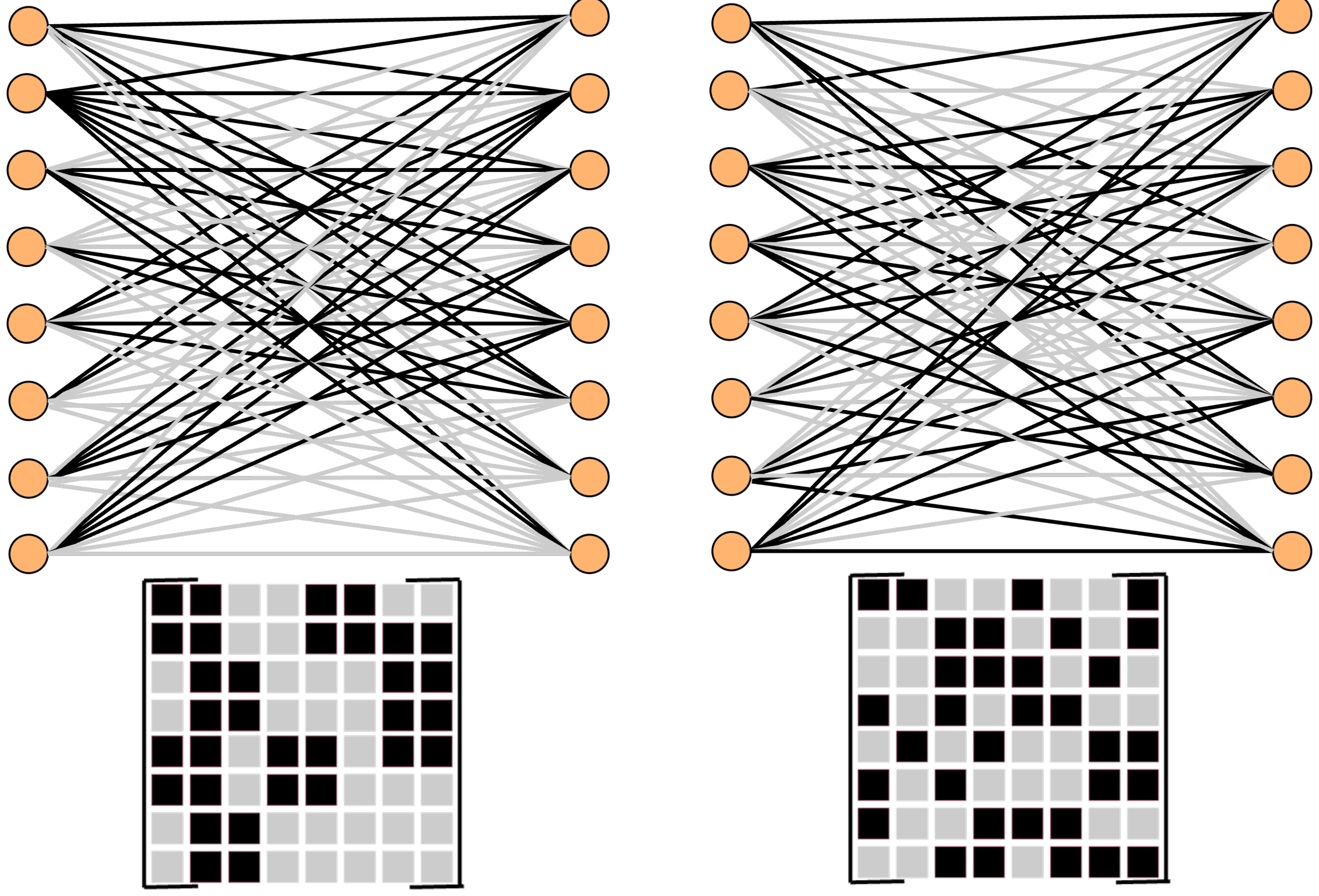


Structured
(Best of Both Worlds)



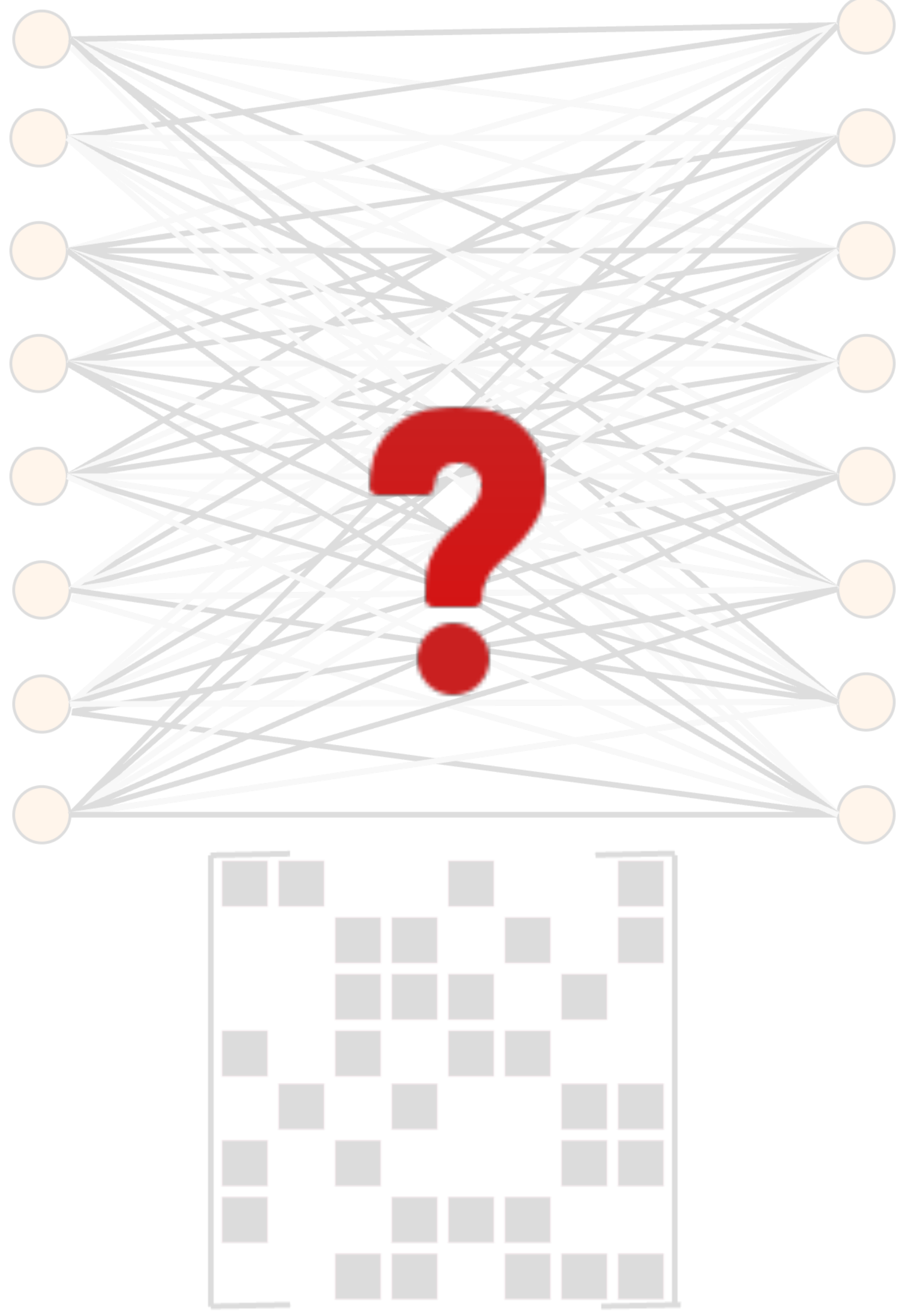
Unstructured

- Pros
  - $\text{Accuracy}_{\text{Unstruc}} \sim \text{Accuracy}_{\text{Dense}}$
  - Achieve high sparsity
- Cons
  - Hardware unfriendly



Structured

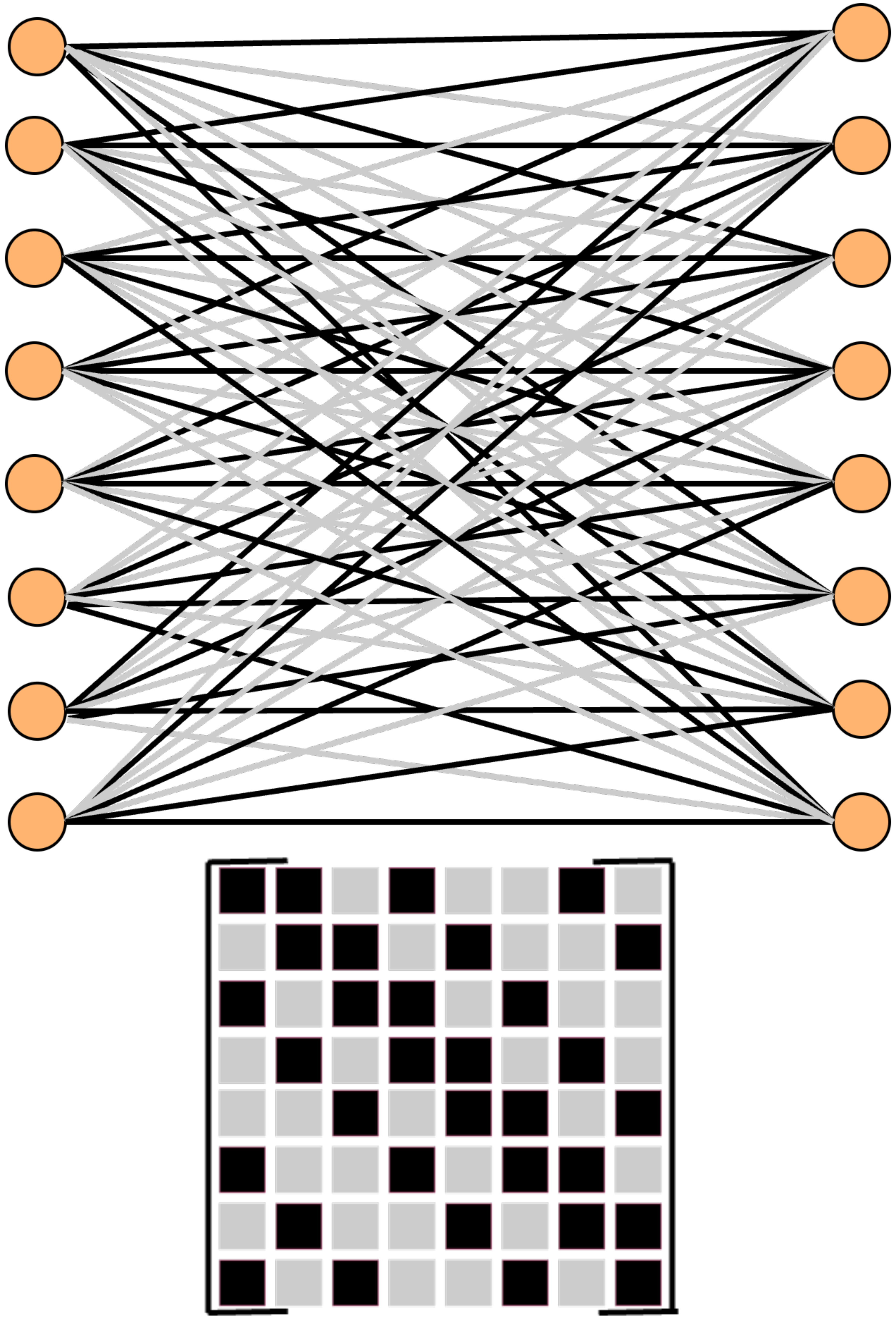
- Pros
  - Hardware friendly
- Cons
  - $\text{Accuracy}_{\text{Struc}} < \text{Accuracy}_{\text{Dense}}$
  - Degradation at high sparsity levels



Structured
(Best of Both Worlds)

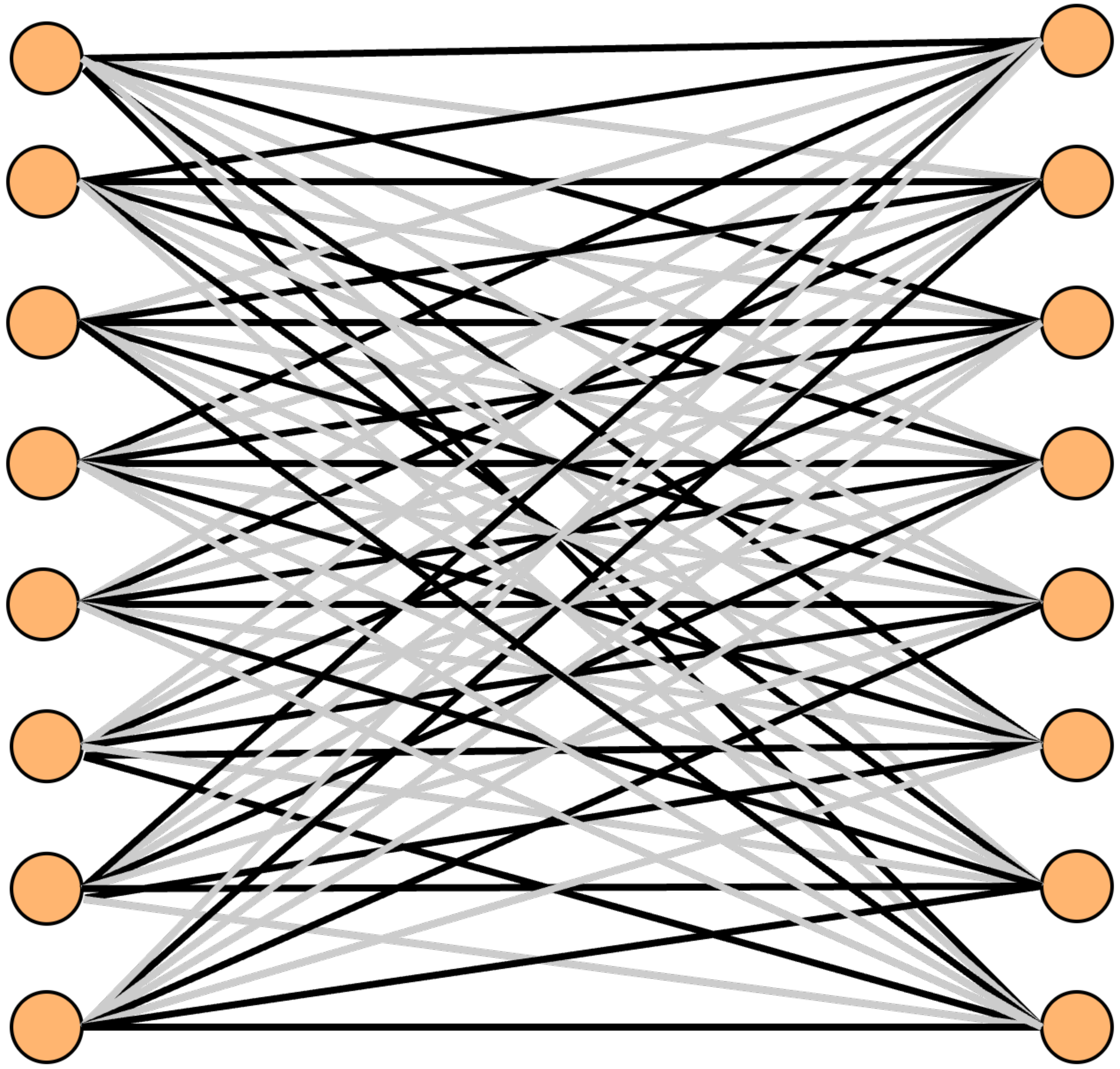
- Pros
  - $\text{Accuracy}_{\text{Struc}} \sim \text{Accuracy}_{\text{Dense}}$
  - Hardware friendly
- Cons
  - Ideally, none!

Our Proposal

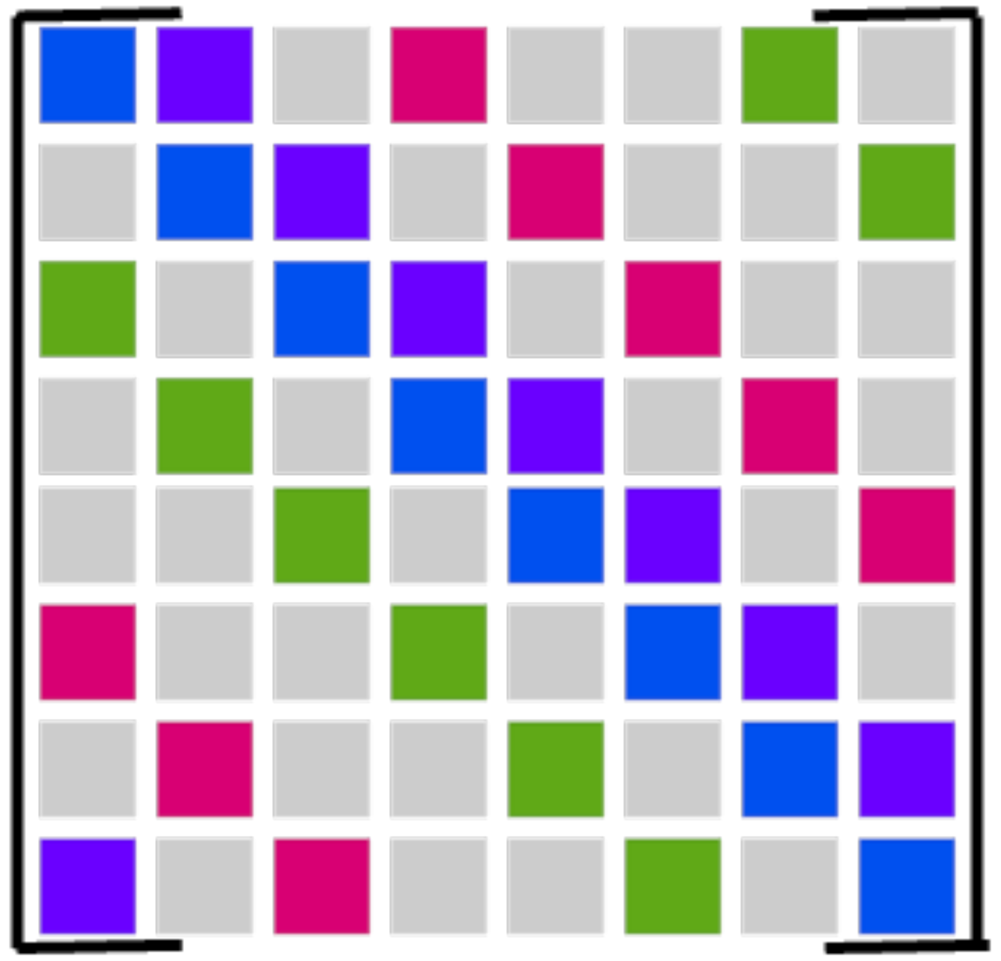


Diagonal Sparsity

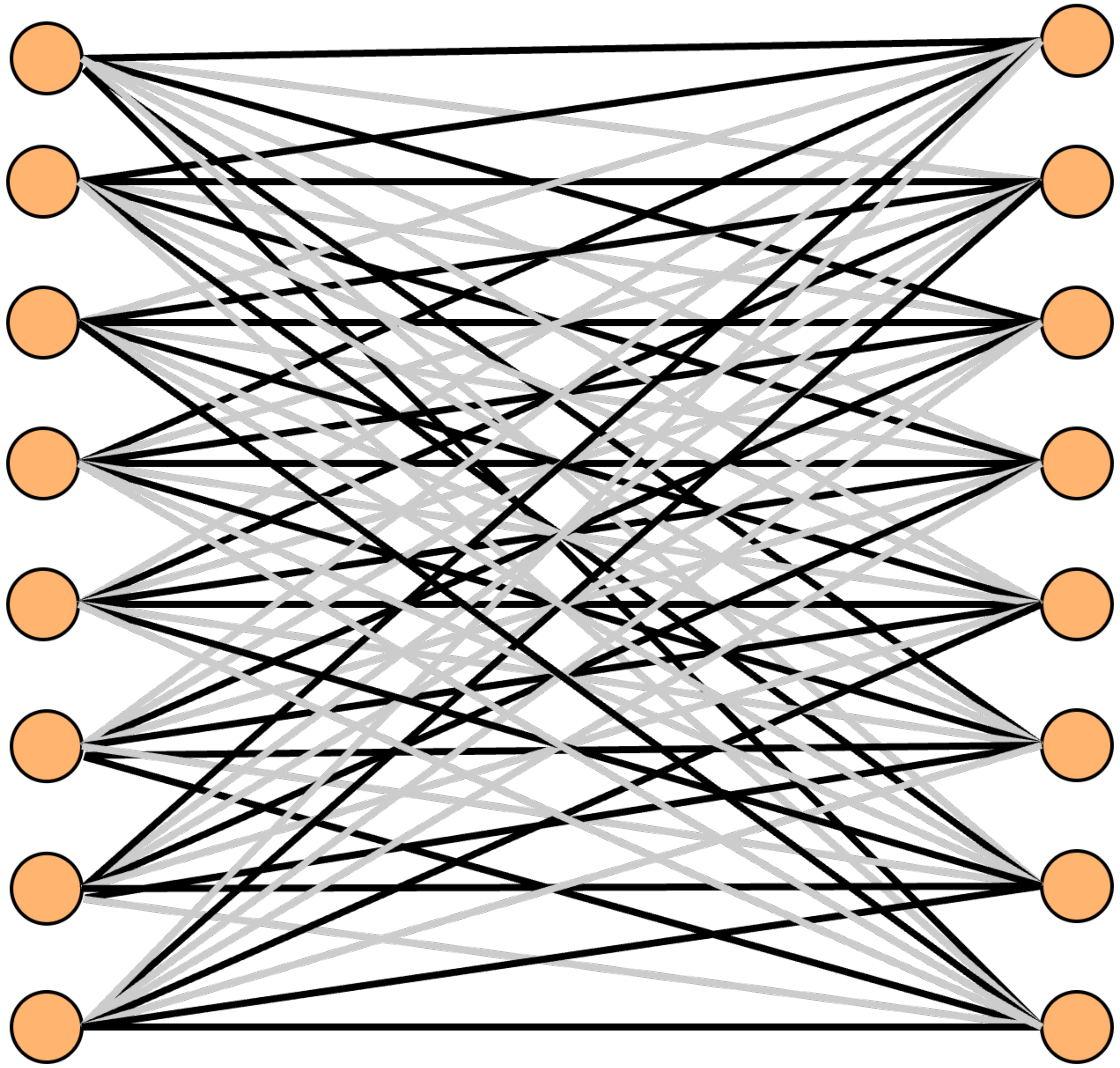
Our Proposal



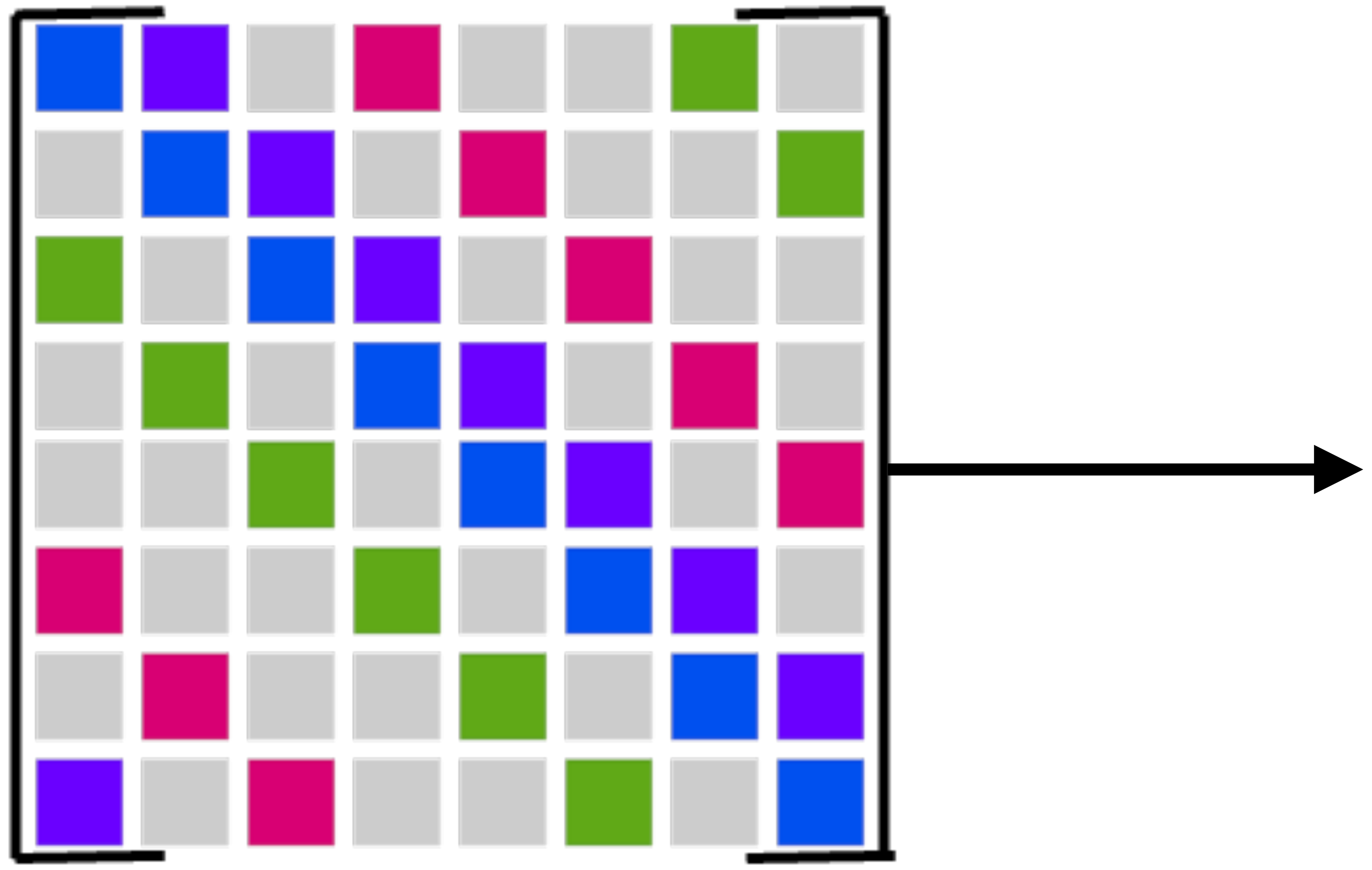
Diagonal Sparsity



Our Proposal

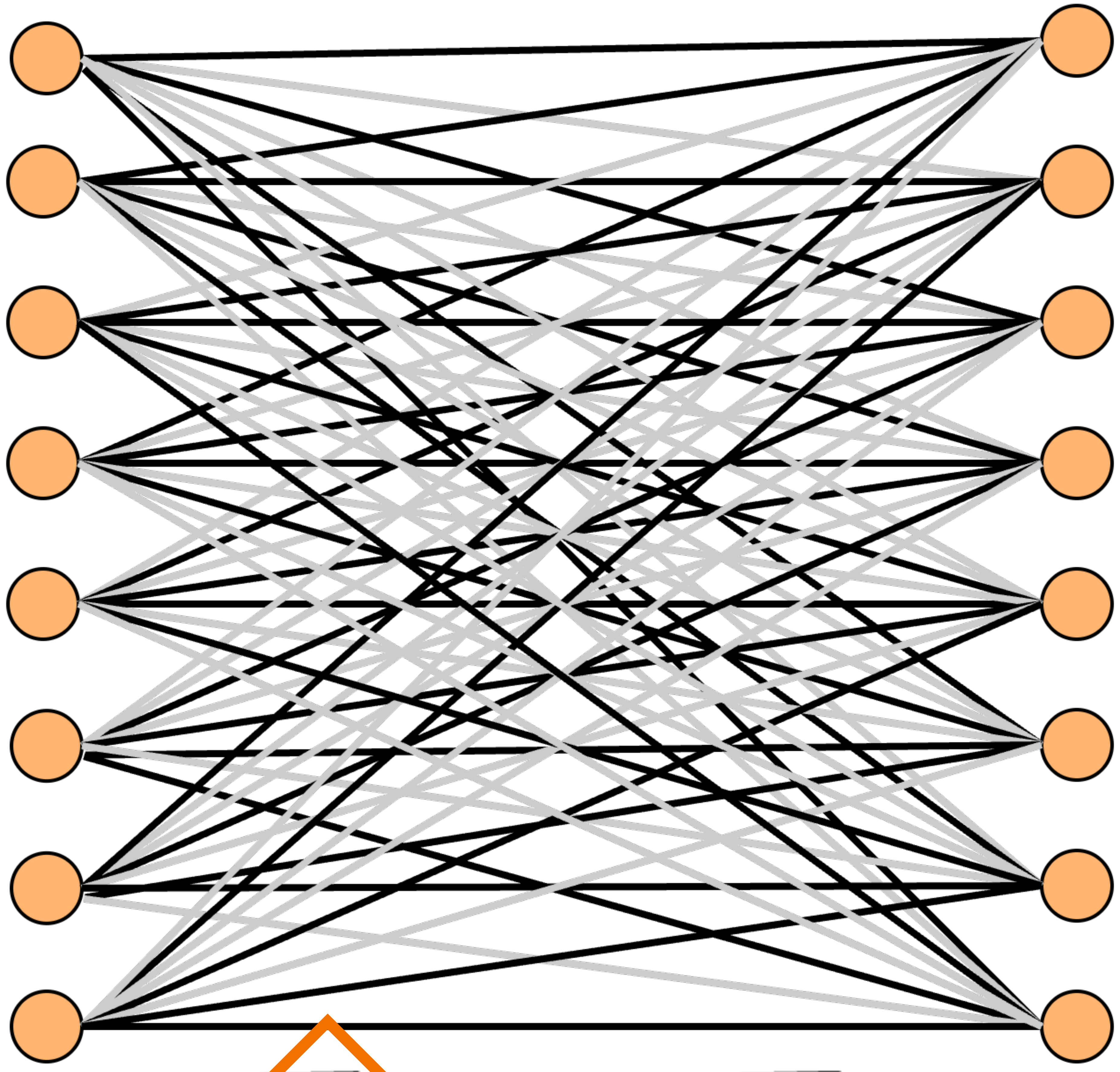


Diagonal Sparsity

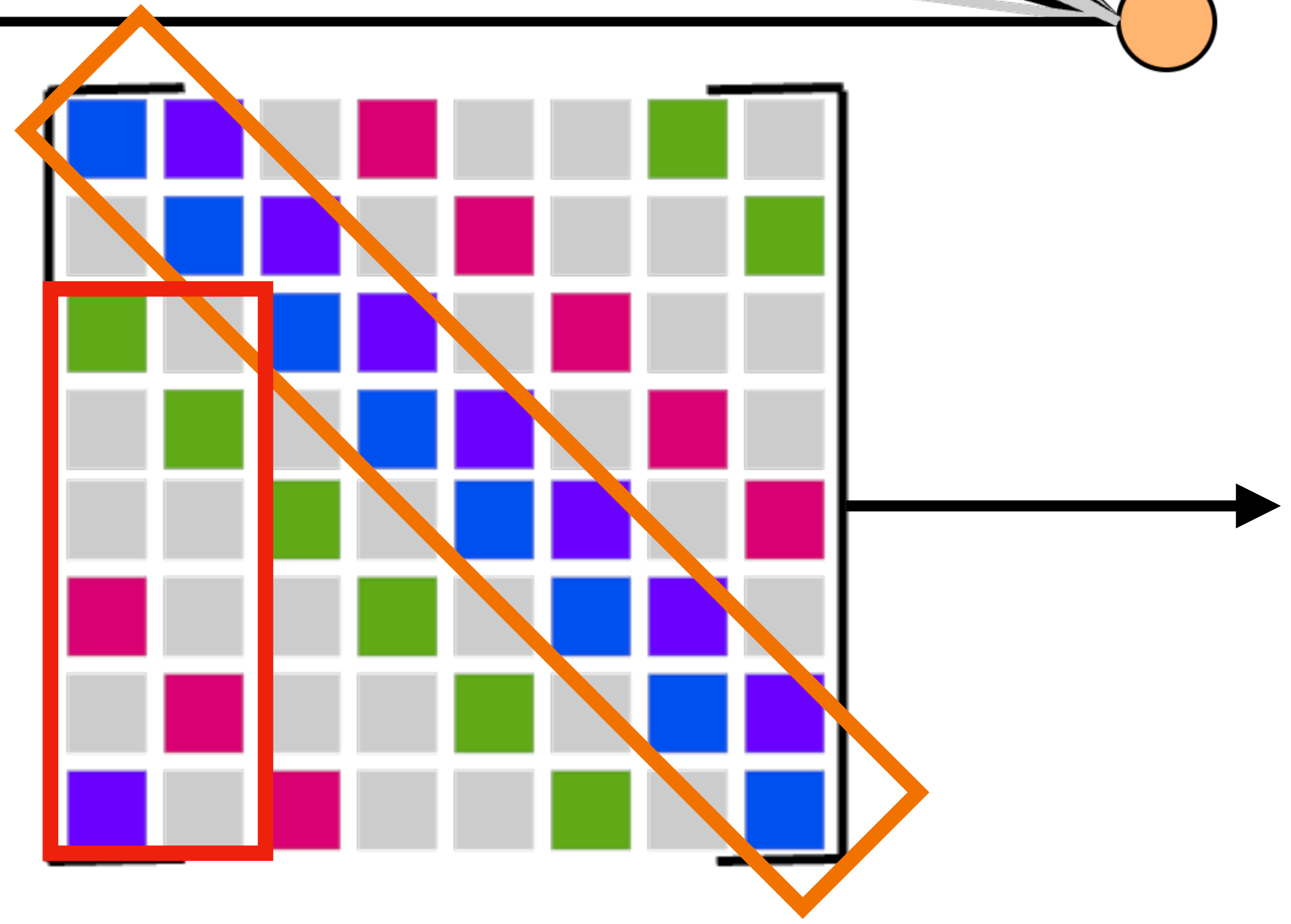


Small World Connectivity

Our Proposal



Diagonal Sparsity



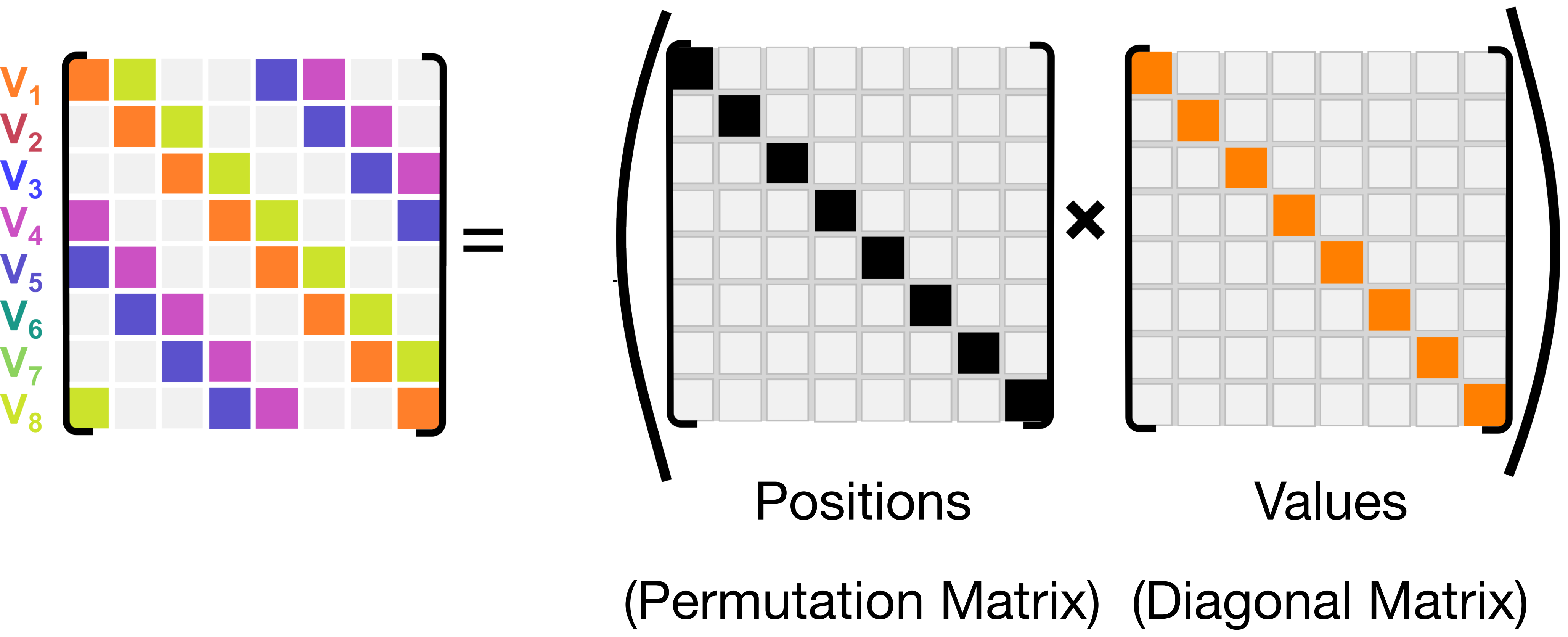
Small World Connectivity

=

Local Clustering + Global Connectivity

Diagonal Sparse Matrix Formulation

Diagonal Sparse Matrix Formulation



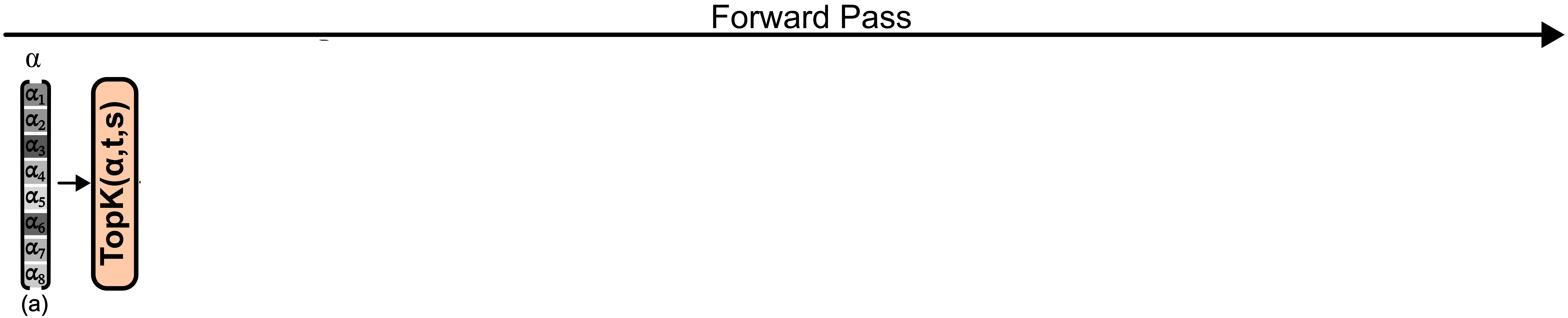
Diagonal Sparse Matrix Formulation

Diagonal Sparse Matrix Formulation

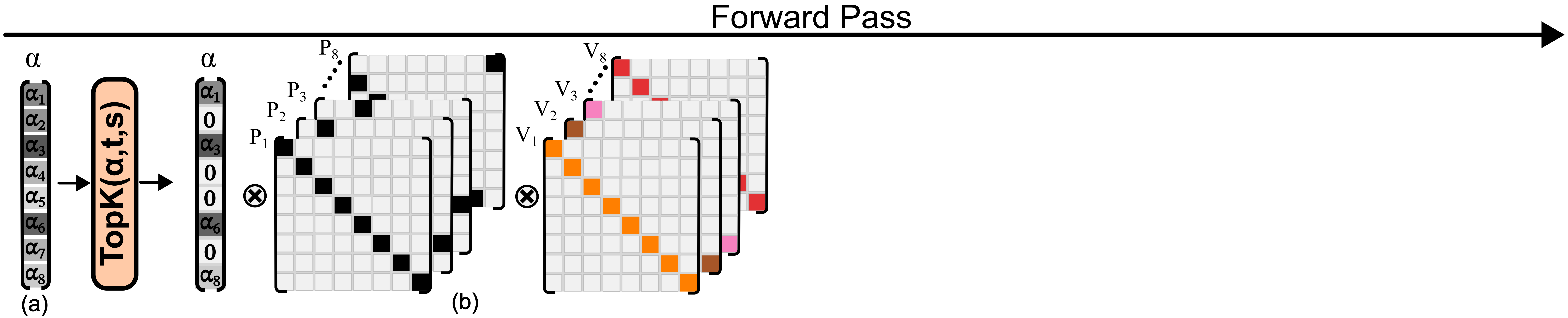
Diagonal Sparse Matrix Formulation

$$W_K = \sum_{j=1}^{K} \tilde{\alpha}_j \, P_j \, \mathrm{diag}(V_j)$$

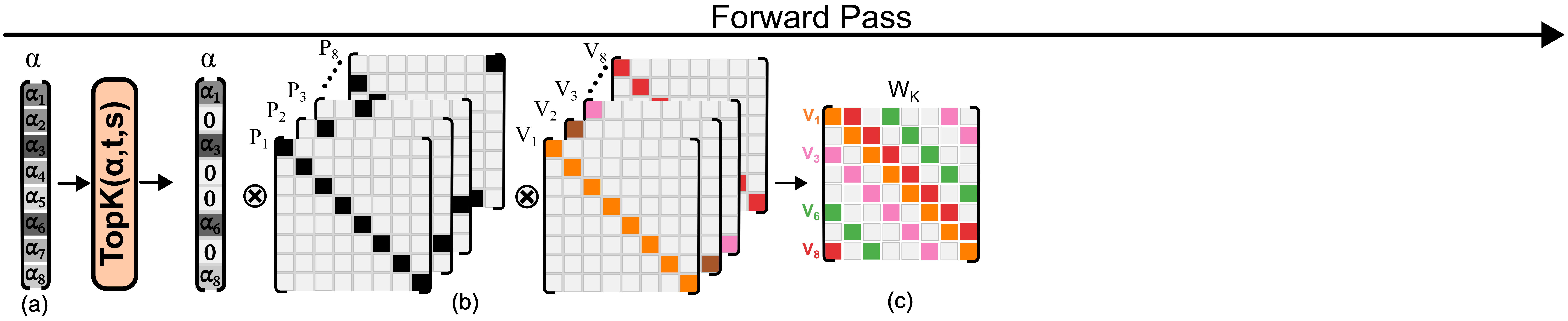
End-To-End Training



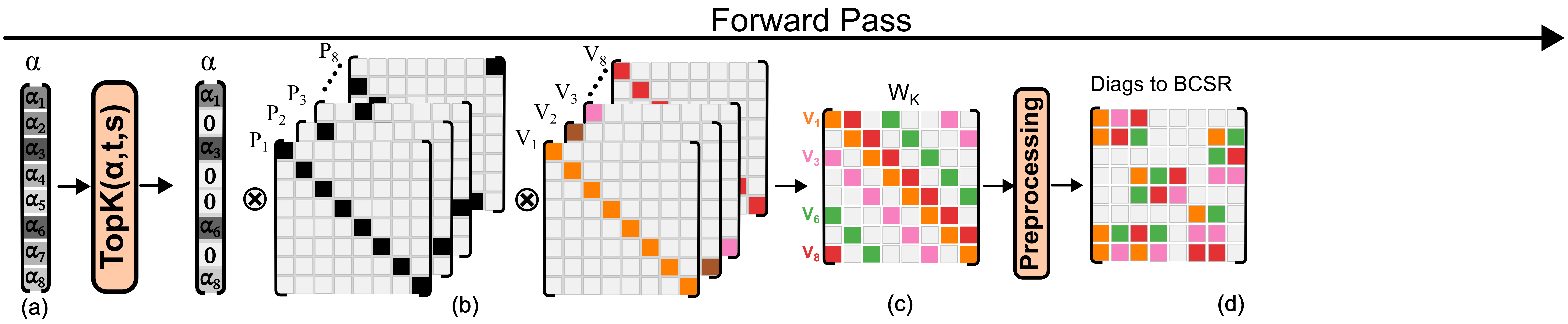
End-To-End Training



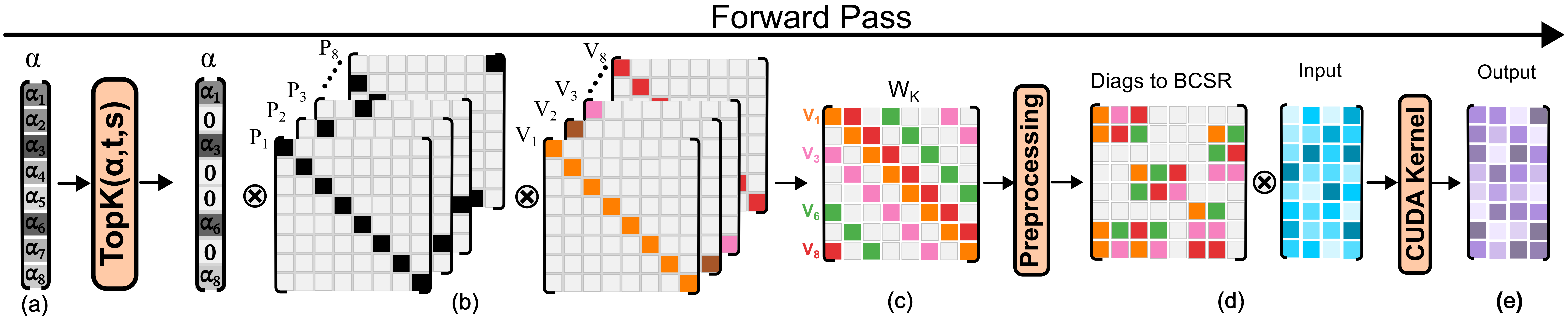
End-To-End Training



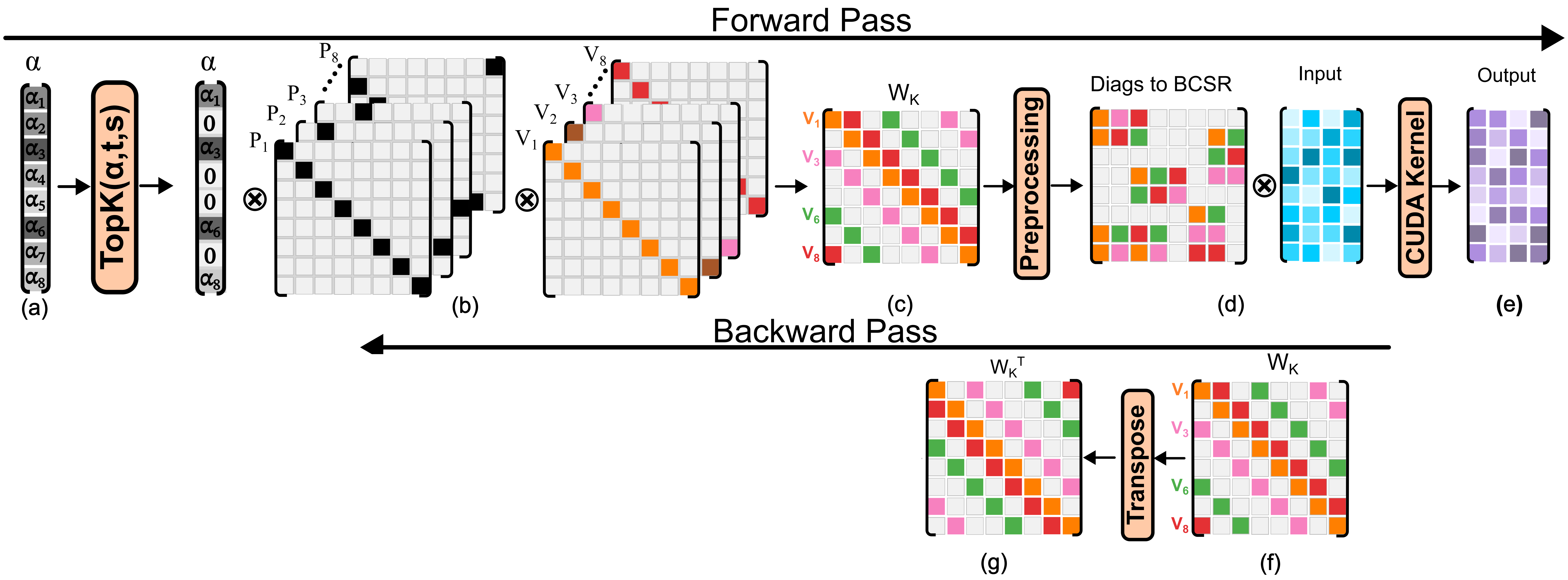
End-To-End Training

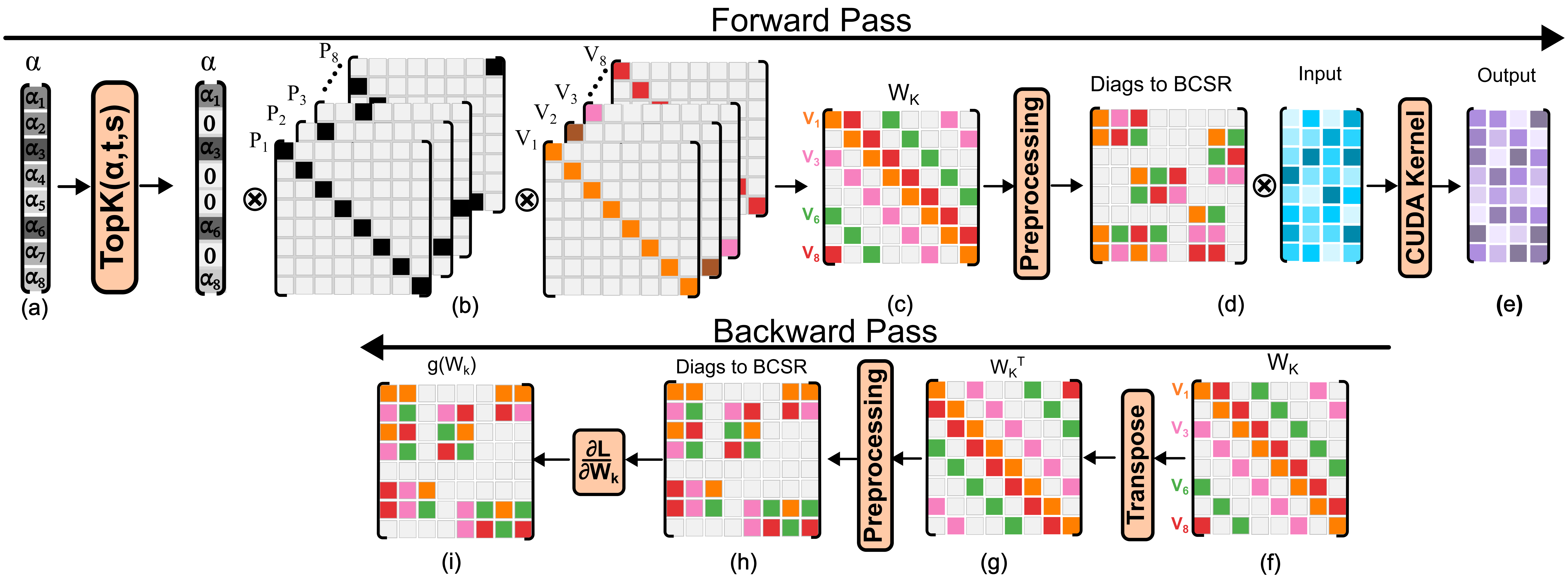


End-To-End Training



End-To-End Training

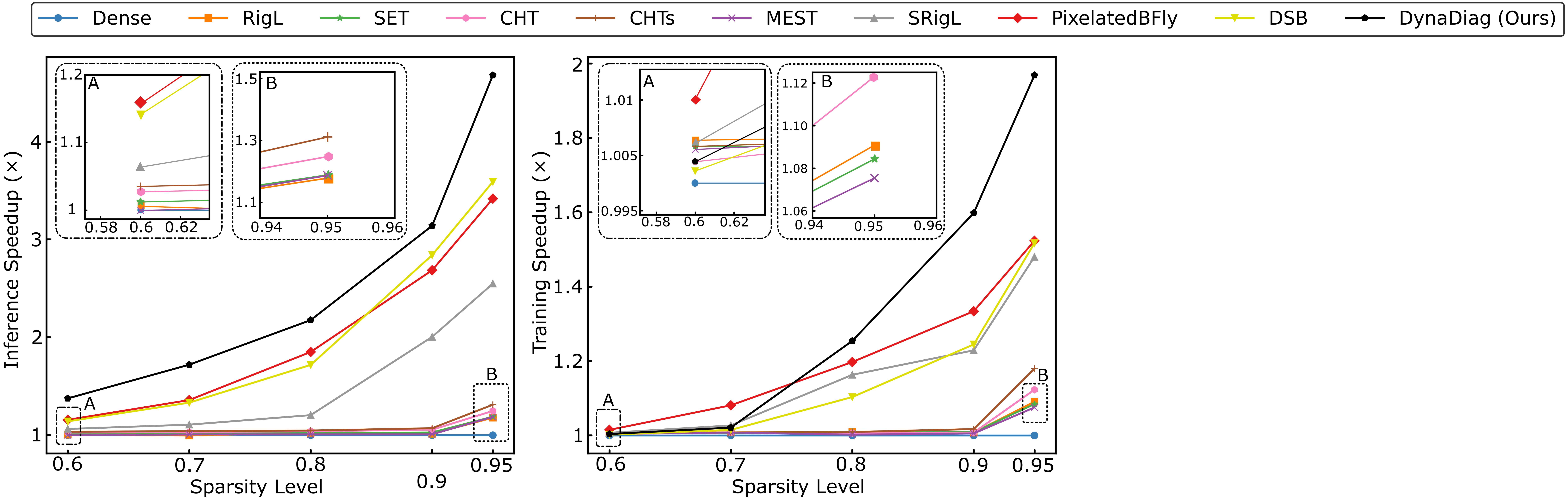




Forward Pass
α
TopK(α,t,s)
α_1 α_2 α_3 α_4 α_5 α_6 α_7 α_8
α_1 0 α_3 0 0 α_6 0 α_8
P_1 P_2 P_3 P_8
V_1 V_2 V_3 V_8
W_K
Preprocessing
Diags to BCSR
Input
CUDA Kernel
Output
(a)
(b)
(c)
(d)
(e)
Backward Pass
$g(W_k)$
Diags to BCSR
Preprocessing
W_K^T
Transpose
W_K
$\frac{\partial L}{\partial W_k}$
V_1 V_3 V_6 V_8
(i)
(h)
(g)
(f)

Results: ViT-B/16 on ImageNet-1K

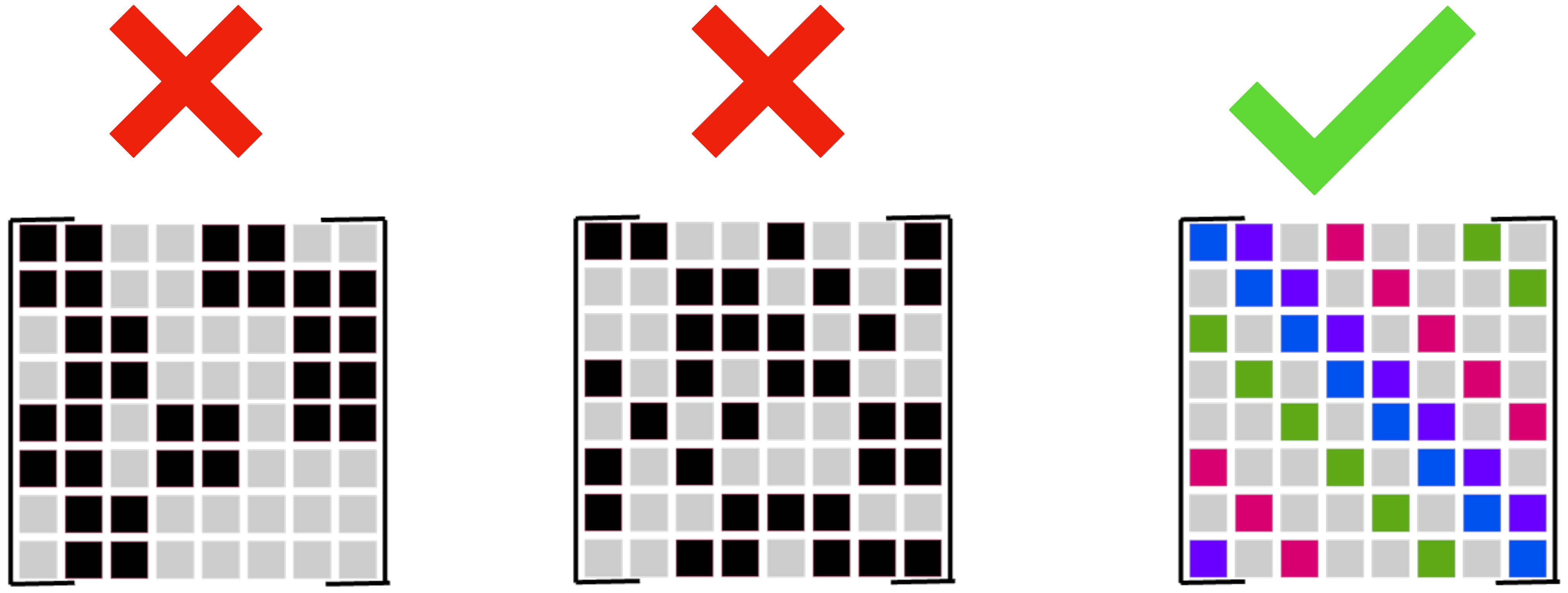
Results: ViT-B/16 on ImageNet-1K



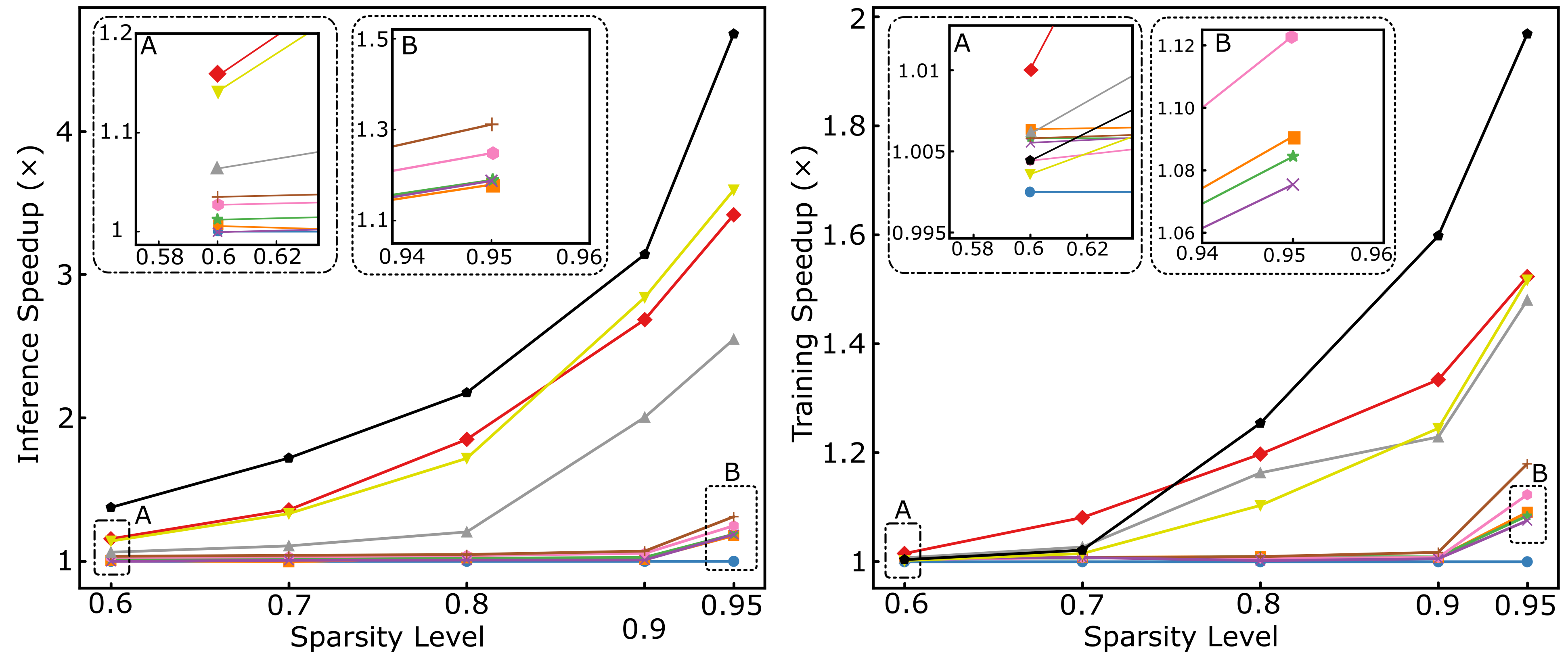
Results: ViT-B/16 on ImageNet-1K

More Results In The Paper!

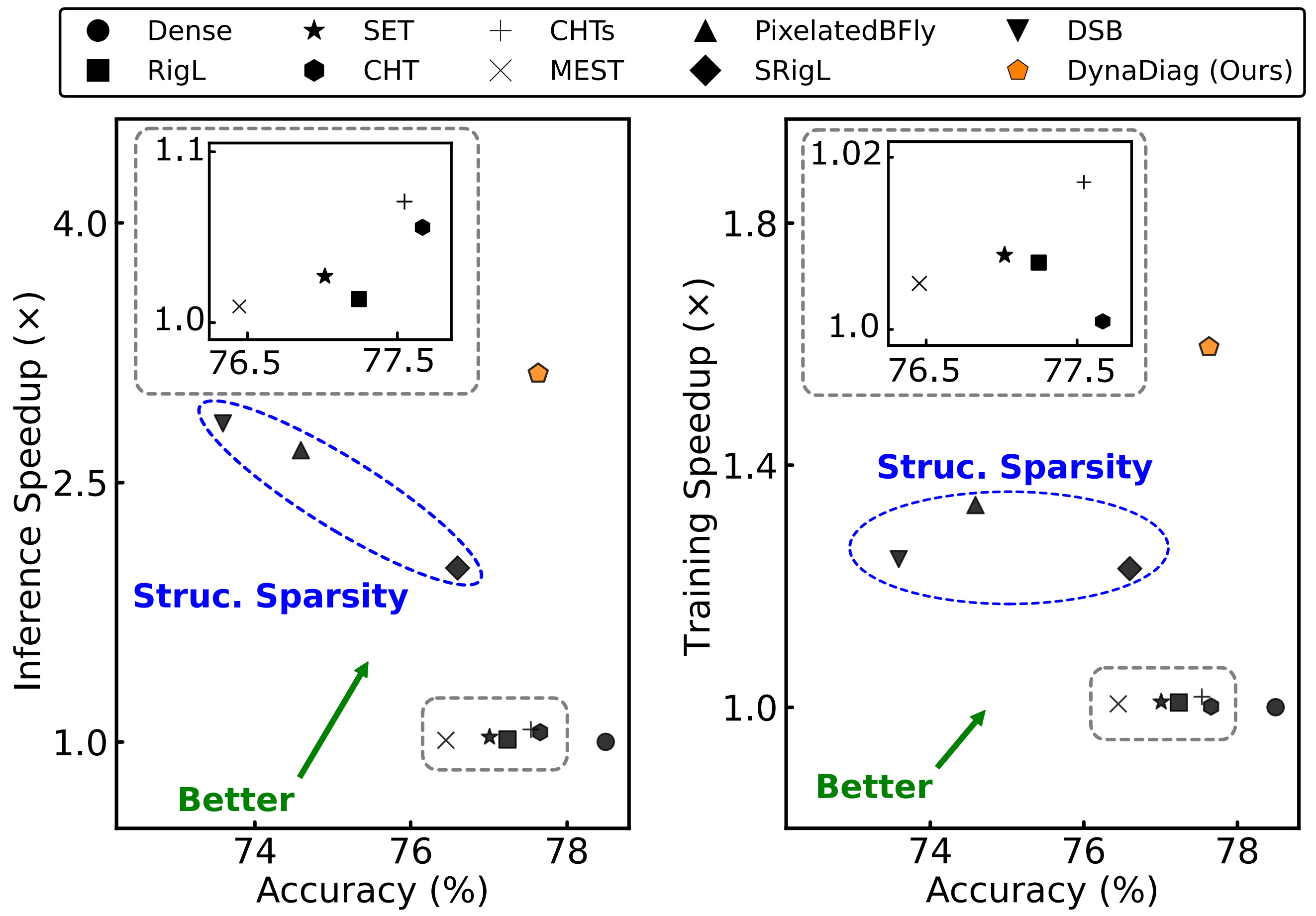
Takeaways



DynaDiag Delievers Unstructured Sparsity Accuracies



Diagonal Sparsity + Custom Kernel Accelerates Training and Inference



DynaDiag Closest To The Ideal Combination of Speed and Accuracy.

Article

Code

Thank you!

abhishektyagi@rochester.edu