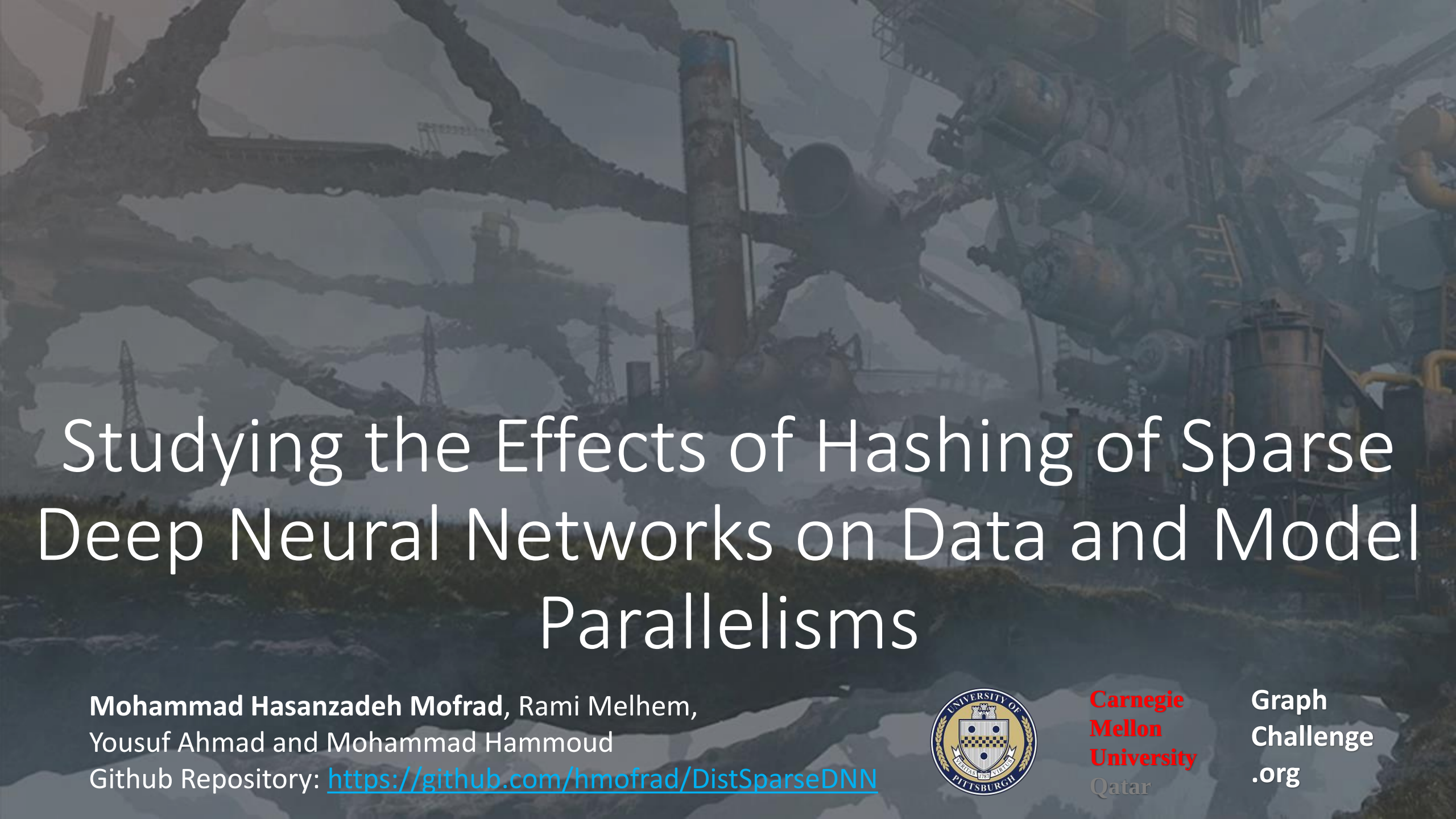


Studying the Effects of Hashing of Sparse Deep Neural Networks on Data and Model Parallelisms

Mohammad Hasanzadeh Mofrad, Rami Melhem,
Yousuf Ahmad and Mohammad Hammoud
Github Repository: https://github.com/hmofrad/DistSparseDNN

Carnegie Mellon University Qatar

Graph Challenge .org

Sparse Matrix-Matrix Multiplication (SpMM) C = A x B

- SpMM is the kernel behind training/inference of sparse DNNs
- **Sparse data structure** (1st part of our study)**:**
  - Row-major access: Compressed Sparse Row (CSR)
  - Column-major access: Compressed Sparse Column (CSC)
- **SpMM algorithm** (2nd part of our study)**:**
  - Multiplication type: Left multiplication, right multiplication
  - Accumulation: Sparse accumulator (SPA)
  - Memory allocation: Symbolic + real SpMM for memory (re)allocation
- **Scale up** using multithreading (3rd part of our study)
  - Data parallelism (parallelizes over the input)
  - Model parallelism (parallelizes over the DNN)
- **Scale out** using MPI (4th part of our study)
  - Data*data parallelism
  - Data*model parallelism

Neural Network Hashing (RadixNet)

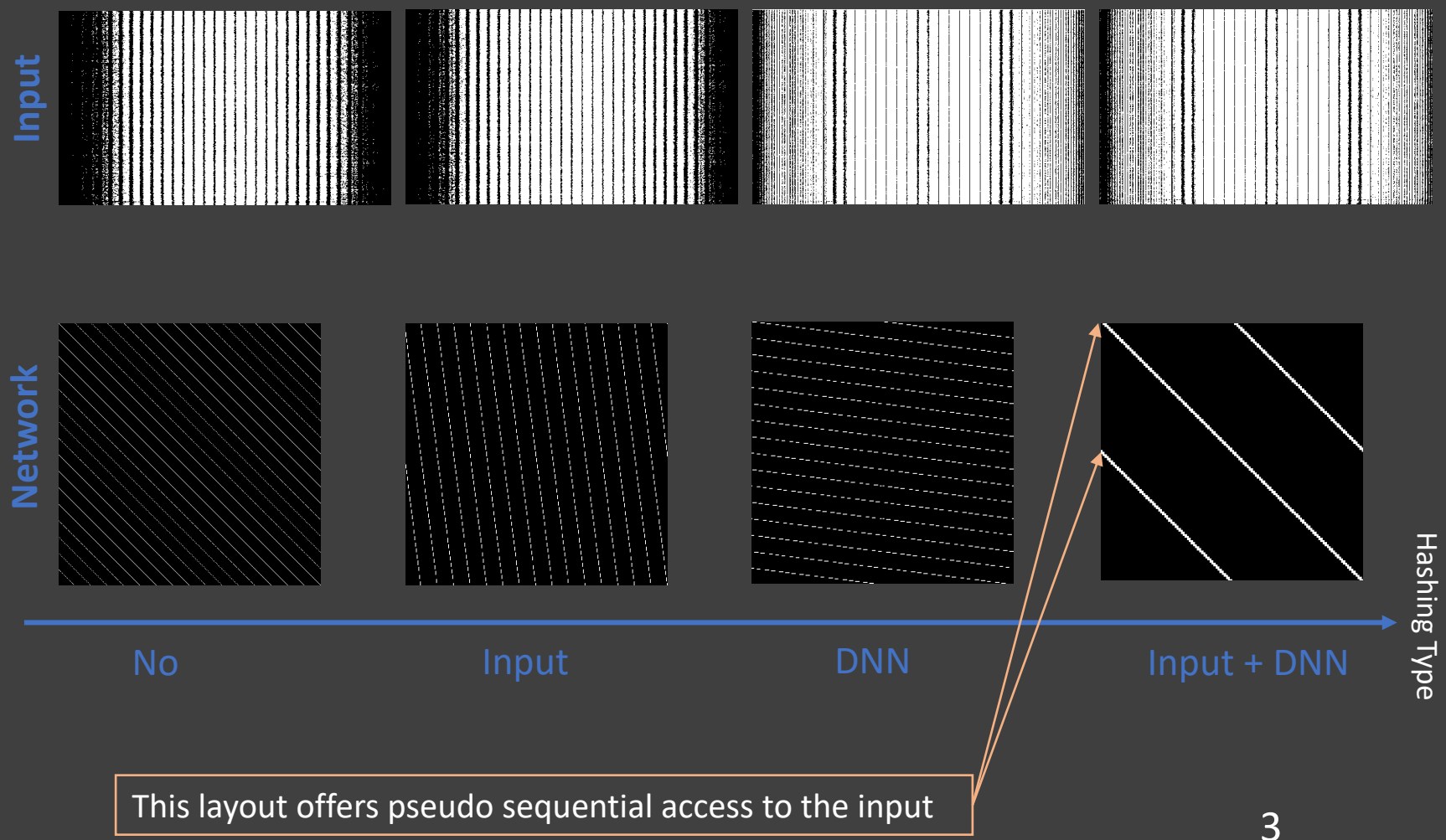
- Hashing mitigates data parallelism stragglers
- **Input hashing**:
  - Hashes rows of A

  + Load balance in data parallelism
- **Network hashing**:
  - Hashes columns of A & rows of B

  + Better access pattern in model parallelism
- **Input + network hashing**:
  - Hashes rows of A
  - Hashes columns of A & rows of B
  - Hashes columns of B

  + Both above advantages

  + Pseudo sequential access pattern



Experiments

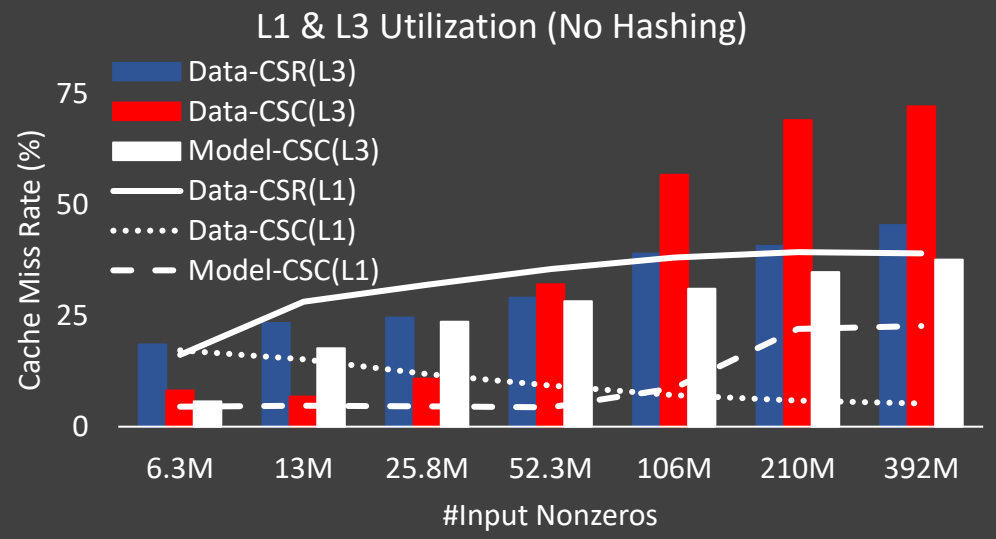
Dataset (RadixNet Sparse DNN, MNIST)					
Input		Network			
		Each Layer		All Layers	
Size $_{m \times n}$	NNZ	Size $_{n \times n}$	NNZ	L	NNZ
60K x 1K	6.3M	1K x 1K	32K	120	3.9M
				480	15.7M
				1920	62.9M
60K x 4K	25M	4K x 4K	131K	120	15.7M
				480	62.7M
				1920	251M
60K x 16K	99M	16K x 16K	524K	120	62.9M
				480	251M
				1920	1B
60K x 65K	392M	65K x 65K	21 M	120	251M
				480	1B
				1920	4B

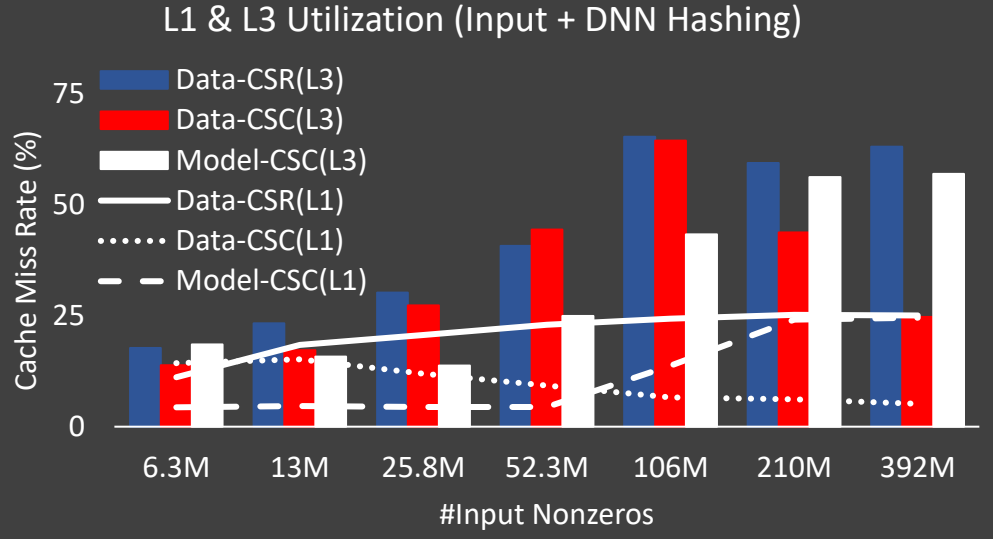
Node Specification (32)	
CPU	28-core @ 2.6 GHZ
Memory	192 GB
OS	Linux
MPI	Intel
Network	Intel Omni-path Fabric

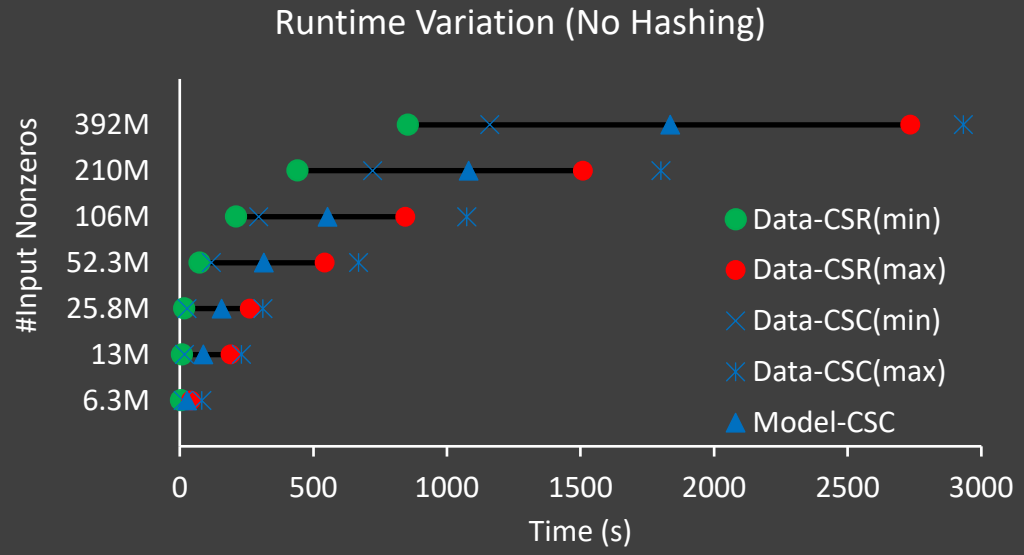
Parallelism/ Multiplication	Right Multiplication (CSC)	Left Multiplication (CSR)
Data Parallelism	Yes	Yes
Model Parallelism	Yes	N/A

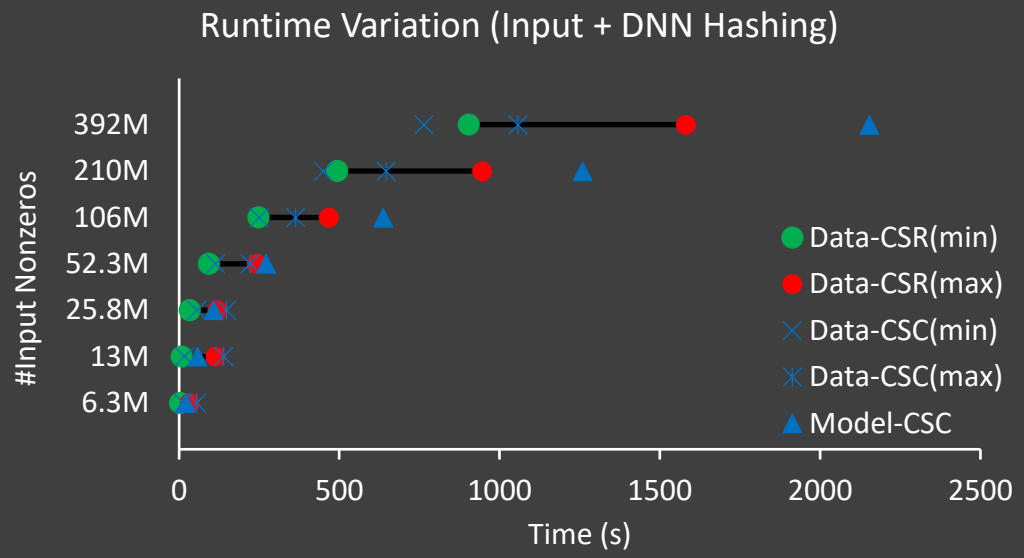
→ **51 Seconds on 32 machines (896 cores)**

Experiments (Single Node): Hashing



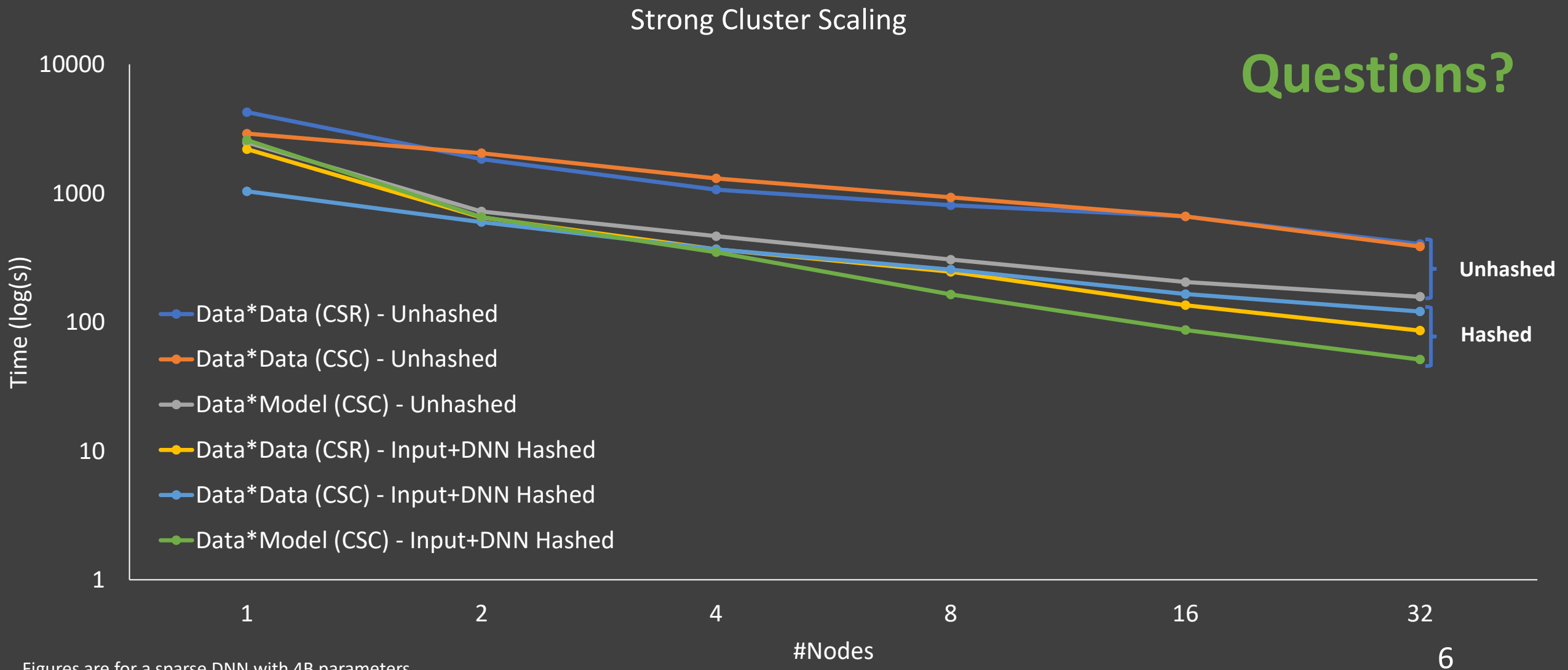






Figures are for a sparse DNN with 4B parameters

Experiments (Distributed, up to 32 Nodes): Scaling



Figures are for a sparse DNN with 4B parameters