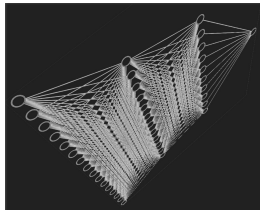
RUNNING 3D FINITE DIFFERENCE SEISMIC IMAGING ON THE GROQ AI INFERENCE ACCELERATOR

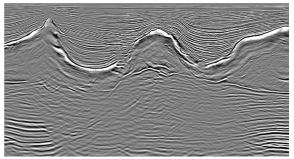
From Stencils To Tensors

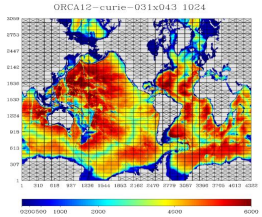
Tobias Becker
November 2023

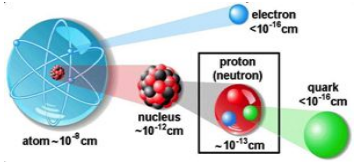
Idea: Use AI Accelerators for HPC Problems

Emergence of new AI chips sparks interest in wider use cases



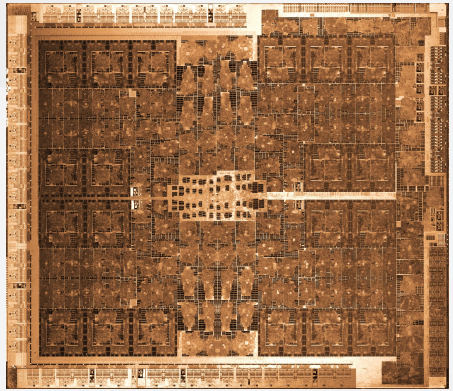


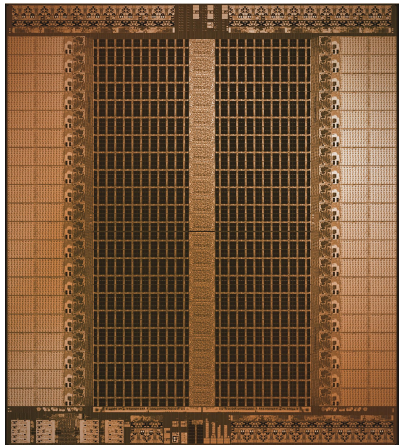




$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Groq Simplifies Compute





Typical GPU
Graphic Processor

COMPLEX
Difficult programming
Less responsiveness
Non-deterministic execution
Higher costs

GroqChip™ 1
First LPU™ Accelerator

SIMPLIFIED

GroqChip™ 1 Overview

Scalable compute architecture

SRAM Memory

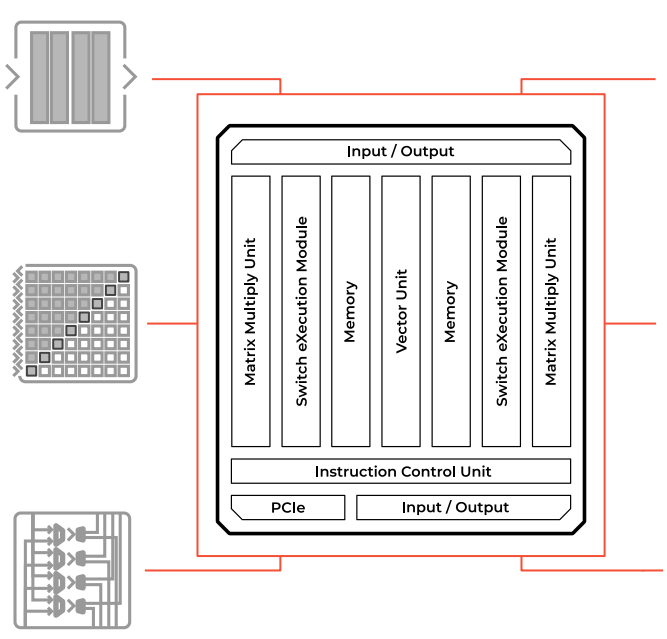
Massive concurrency
80 TB/s of BW
230MB capacity
Stride insensitive

Groq TruePoint™ Matrix

4x Engines
750 TOP/s int8
188 TFLOP/s fp16
320x320 fused dot product

Programmable Vector Units

5,120 Vector ALUs for high performance



Networking

480 GB/s bandwidth
Extensible network scalability
Multiple topologies

Data Switch

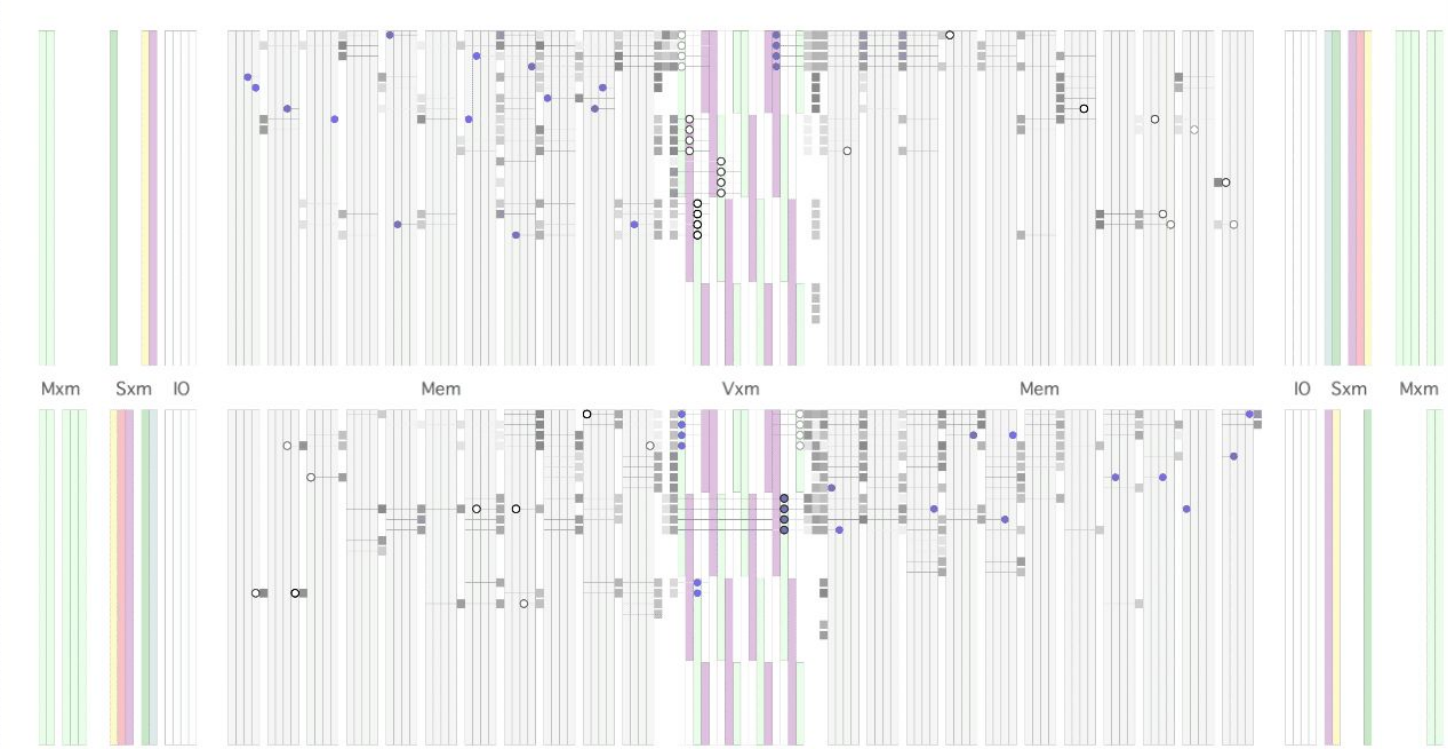
Shift, Transpose, Permuter for improved data movement and data reshapes

Instruction Control

Multiple instruction queues for instruction parallelism

Visualizing Data Orchestration

Given to Groq Compiler



Groq Workloads at Scale

TSP Architecture
Provides Near-linear Scaling Performance

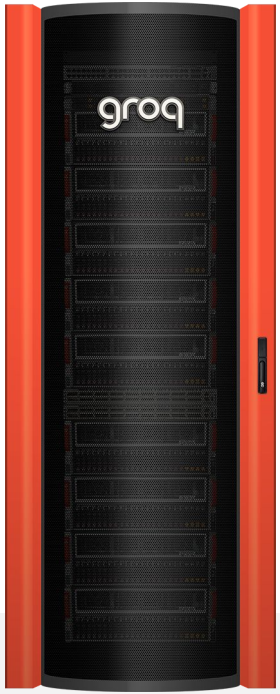
Cards Scale to Nodes
GroqNode contains 8 cards

Cluster Ready
8 x RealScale external ports

2U Server with 4x GroqCard

Nodes Scale to Racks
GroqRack: 8 compute nodes

+ 1 redundant node



GroqChip™ 1

Synchronous Tensor Streaming Processor architecture

RealScale™ enabled

GroqCard™

Up to 750 TOPs, 188 TFLOPs (INT8, FP16 @ 900MHz), 240W*

Dell R750XA

Up to 3 POPs, 752 TFLOPs

(INT8, FP16)

GroqNode™

Up to 6 POPs, 1.5 PFLOPs (INT8, FP16)

GroqRack™

Up to 48 POPs, 12 PFLOPs (INT8, FP16)

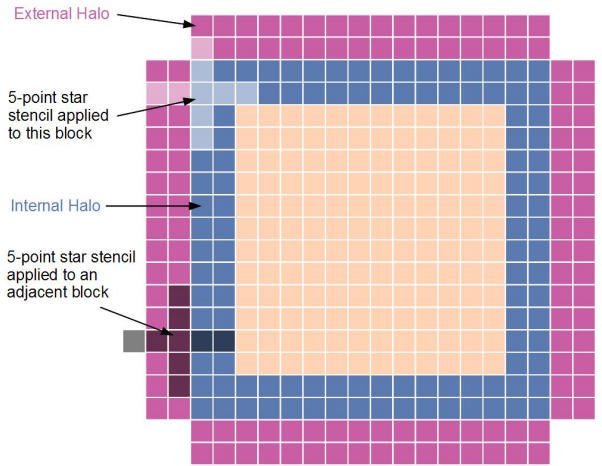
Seismic Modelling

- Simulate the propagation of an acoustic wave through earth / water by solving the acoustic wave equation:

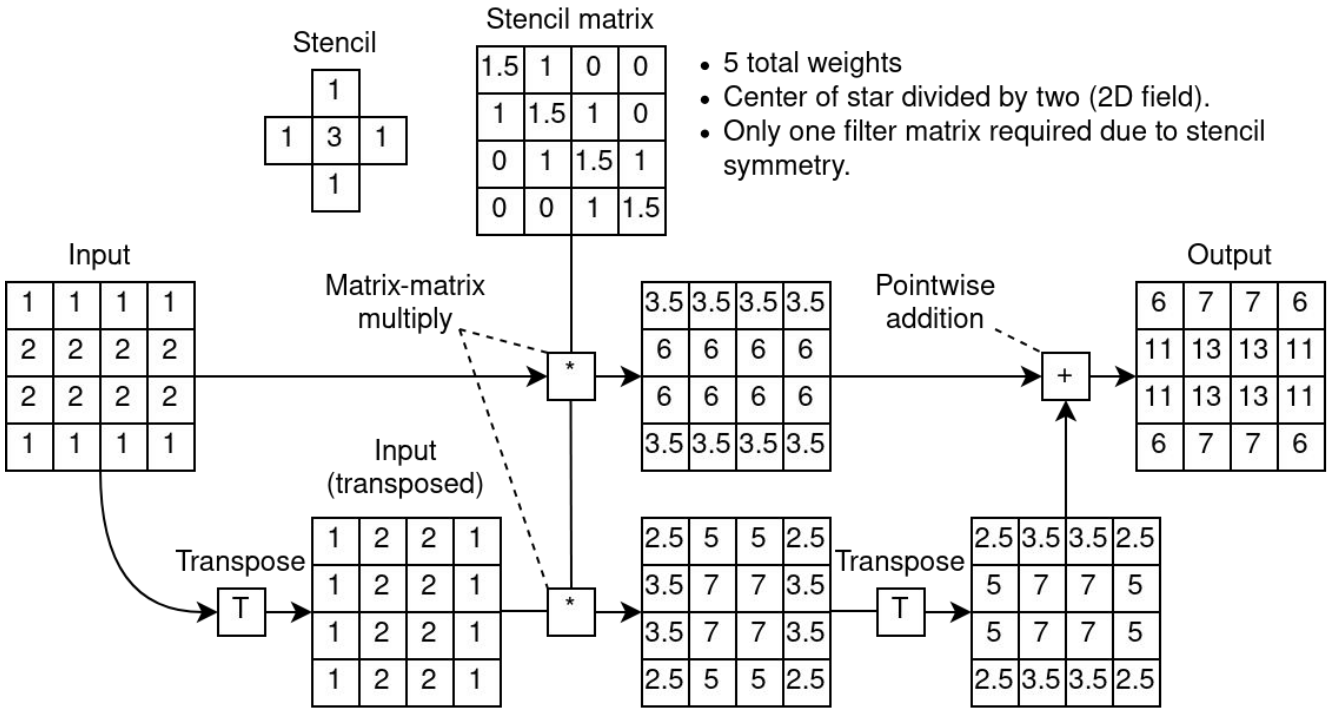
$$\frac{\partial^2 p}{\partial t^2} = v^2 \nabla^2 p + s(t)$$

- Used in Reverse Time Migration (RTM) and Full Waveform Inversion

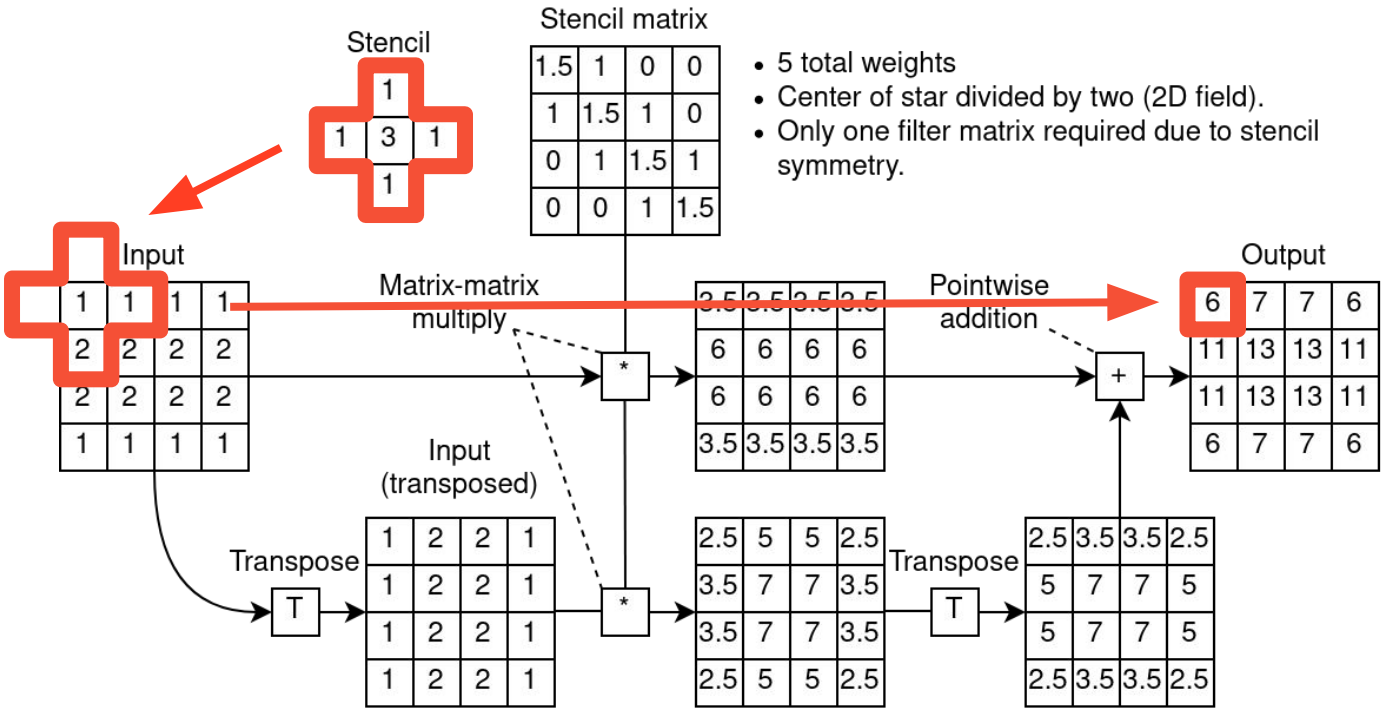
- Finite difference solver with 3D stencil



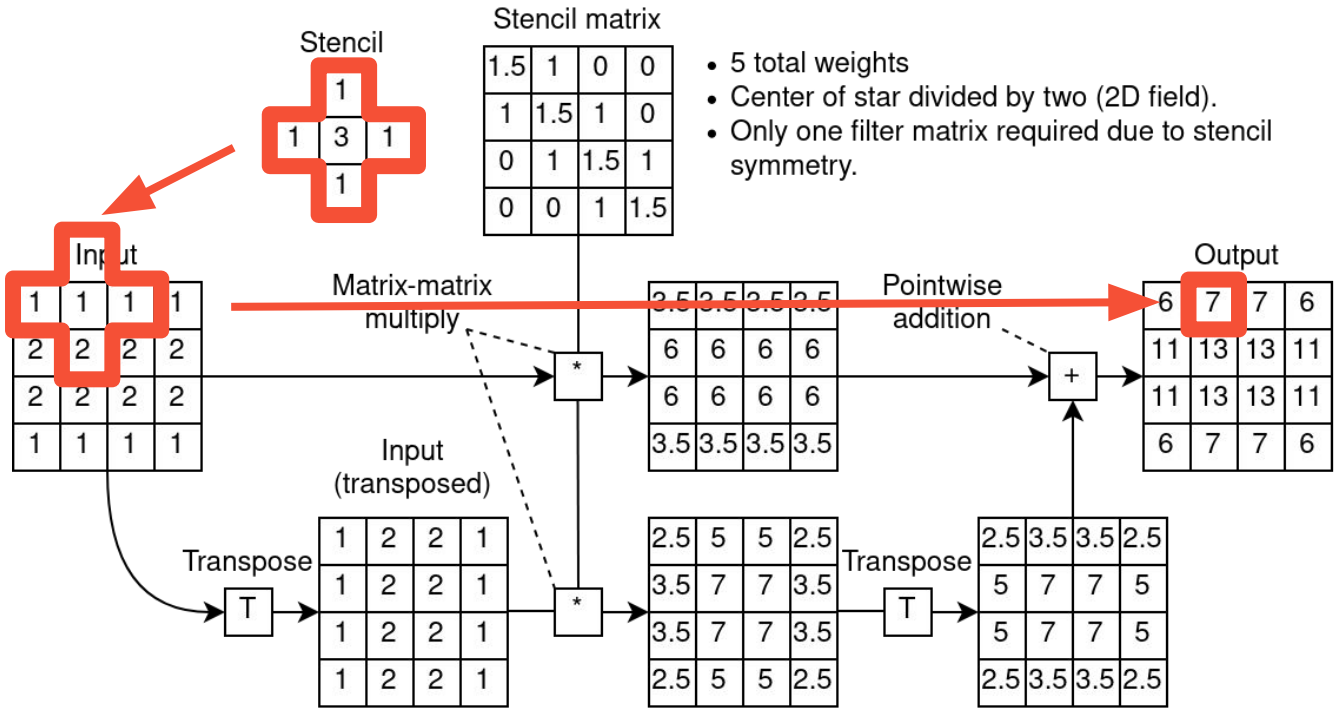
From Stencils to Tensors



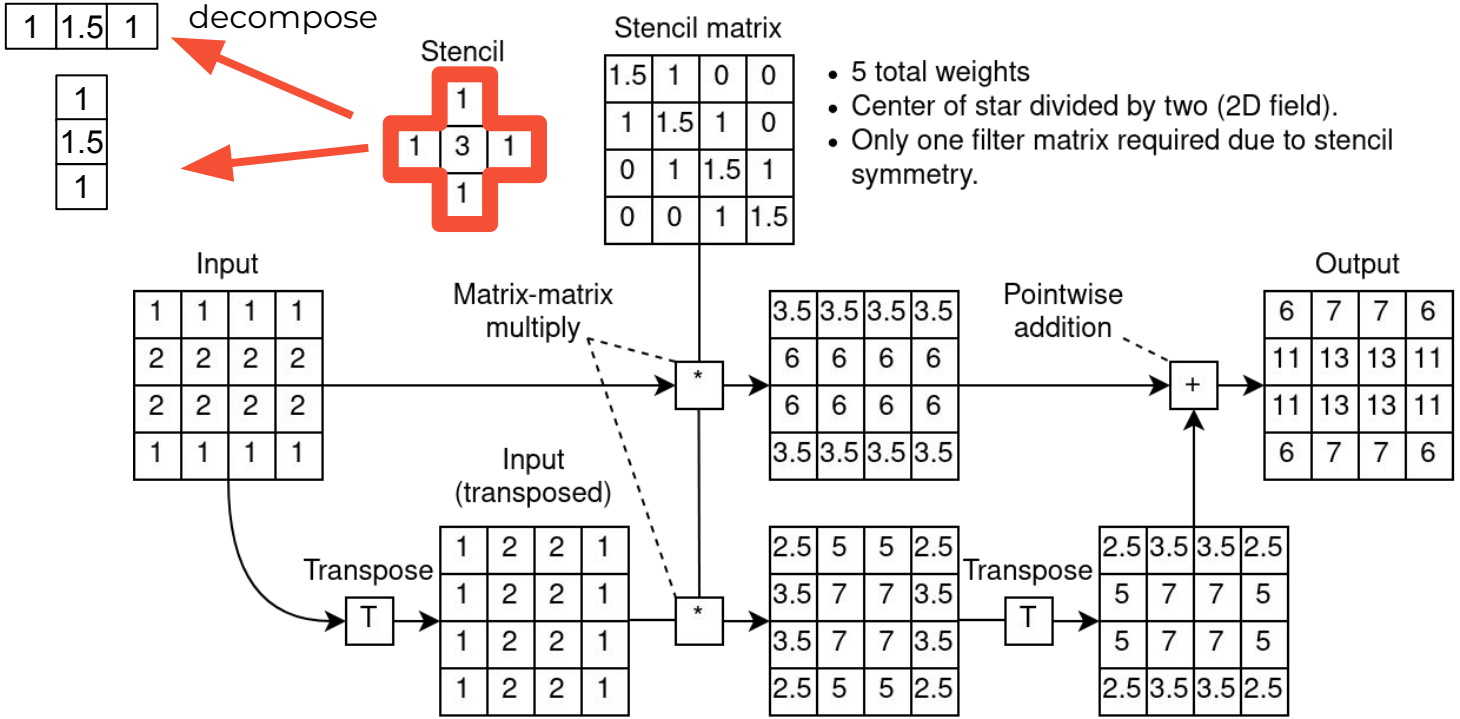
From Stencils to Tensors



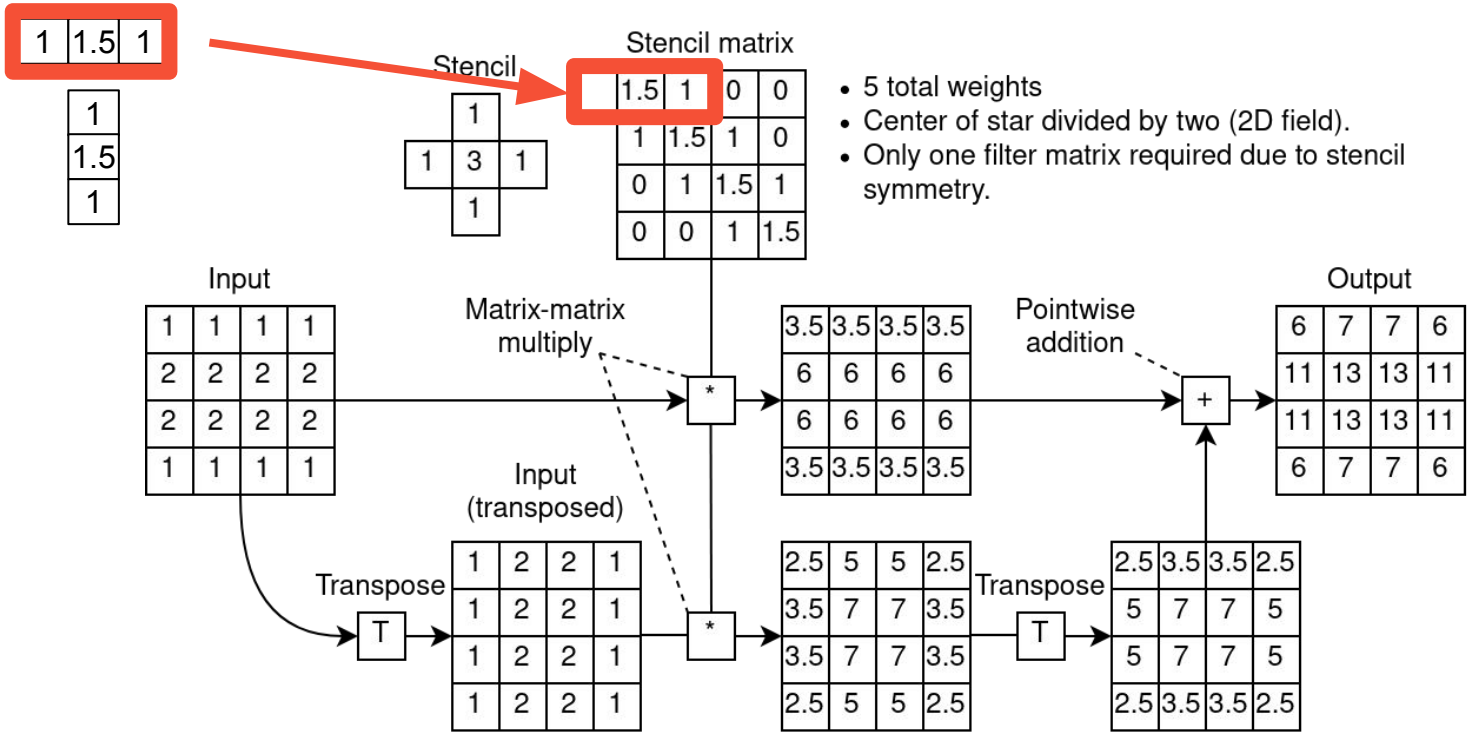
From Stencils to Tensors



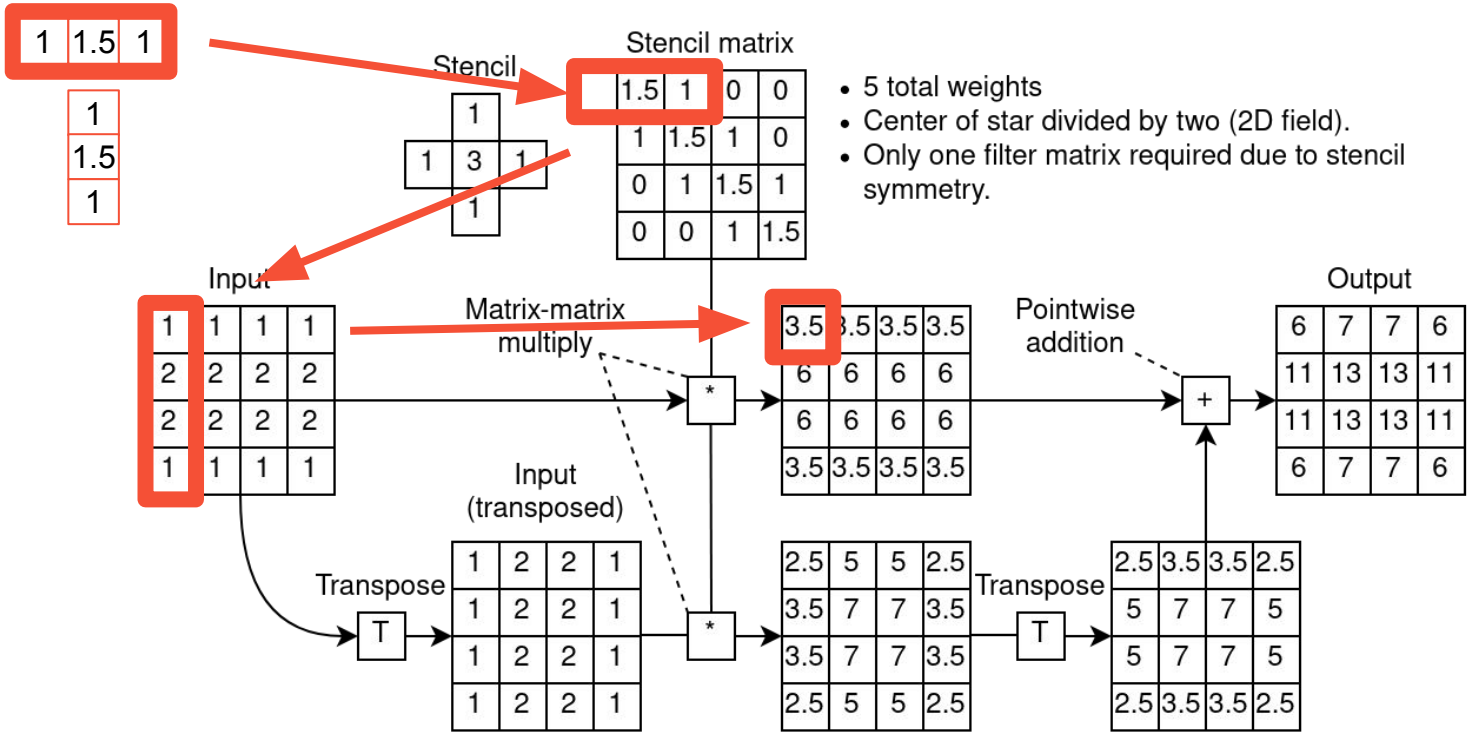
From Stencils to Tensors



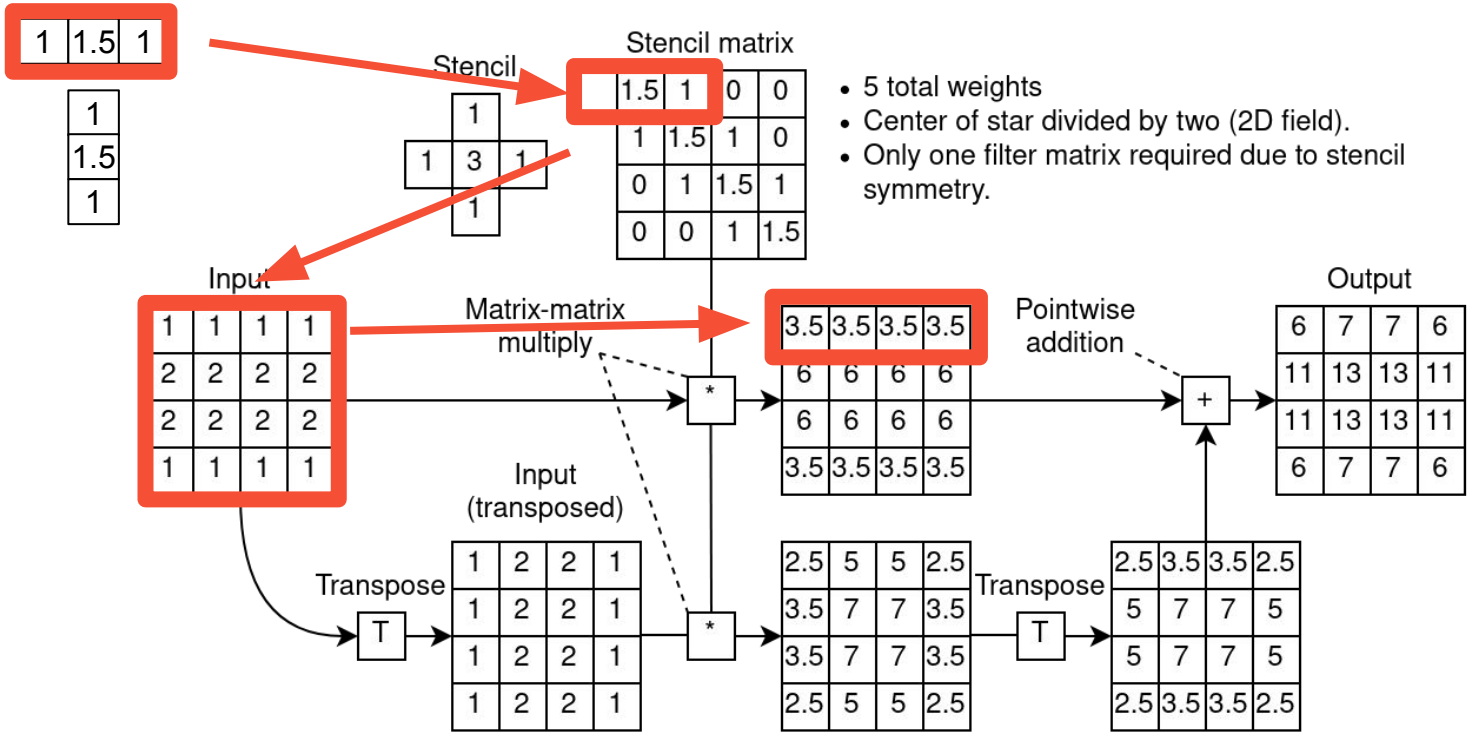
From Stencils to Tensors



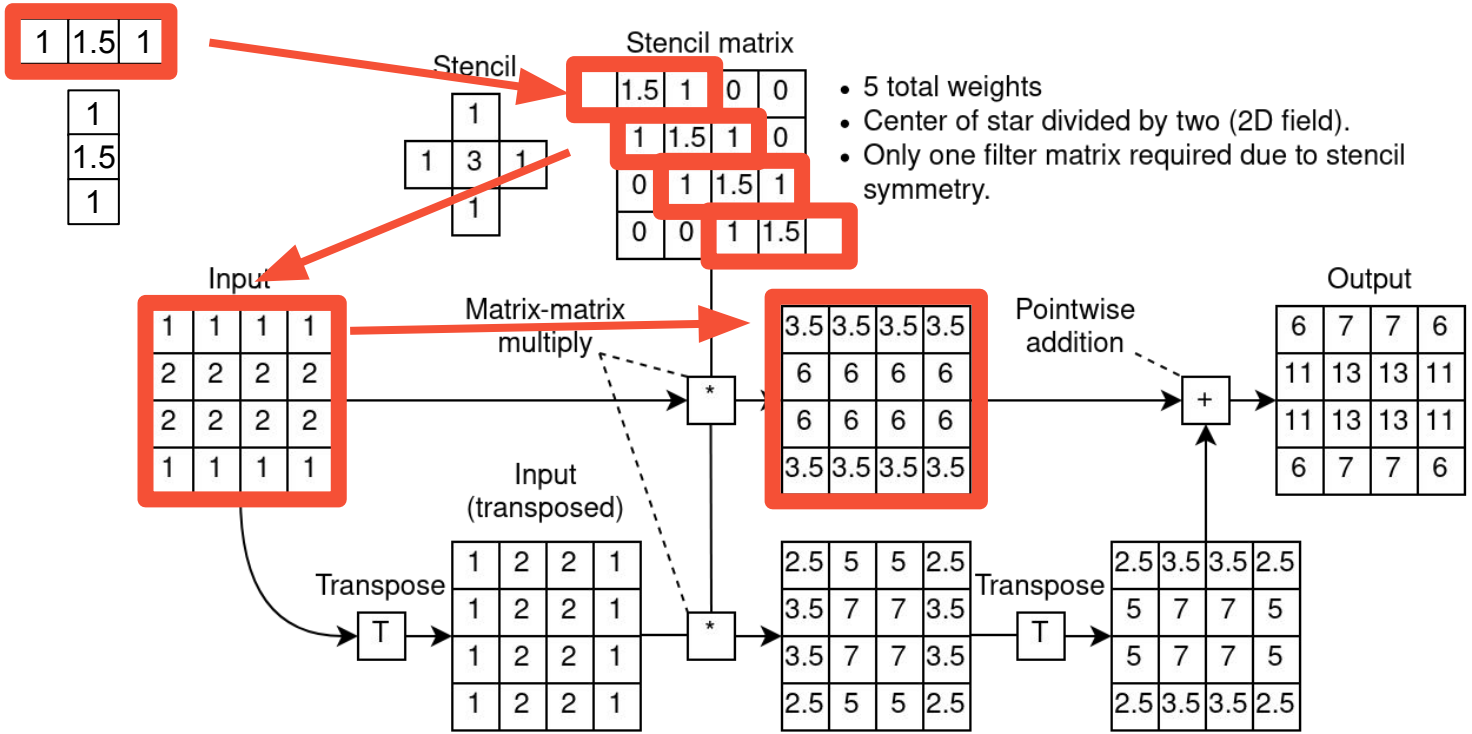
From Stencils to Tensors



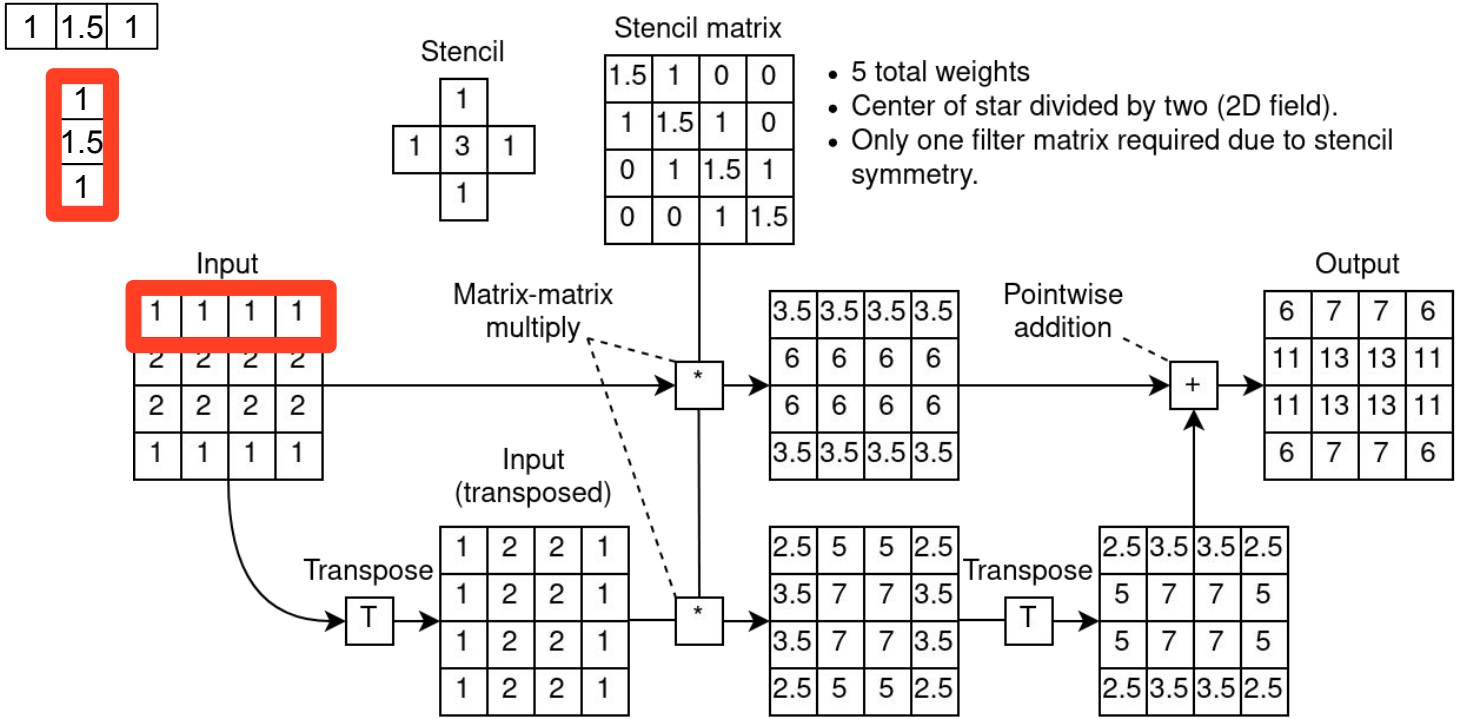
From Stencils to Tensors



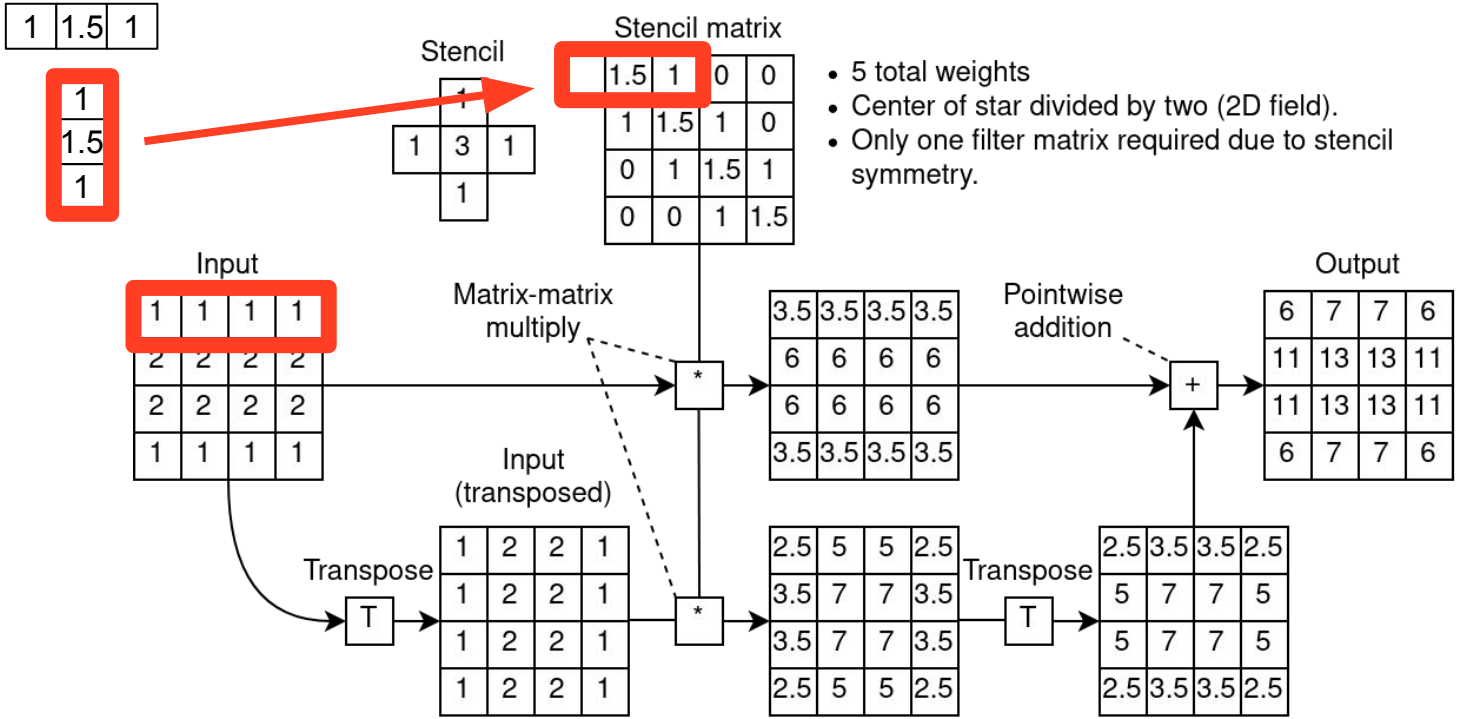
From Stencils to Tensors



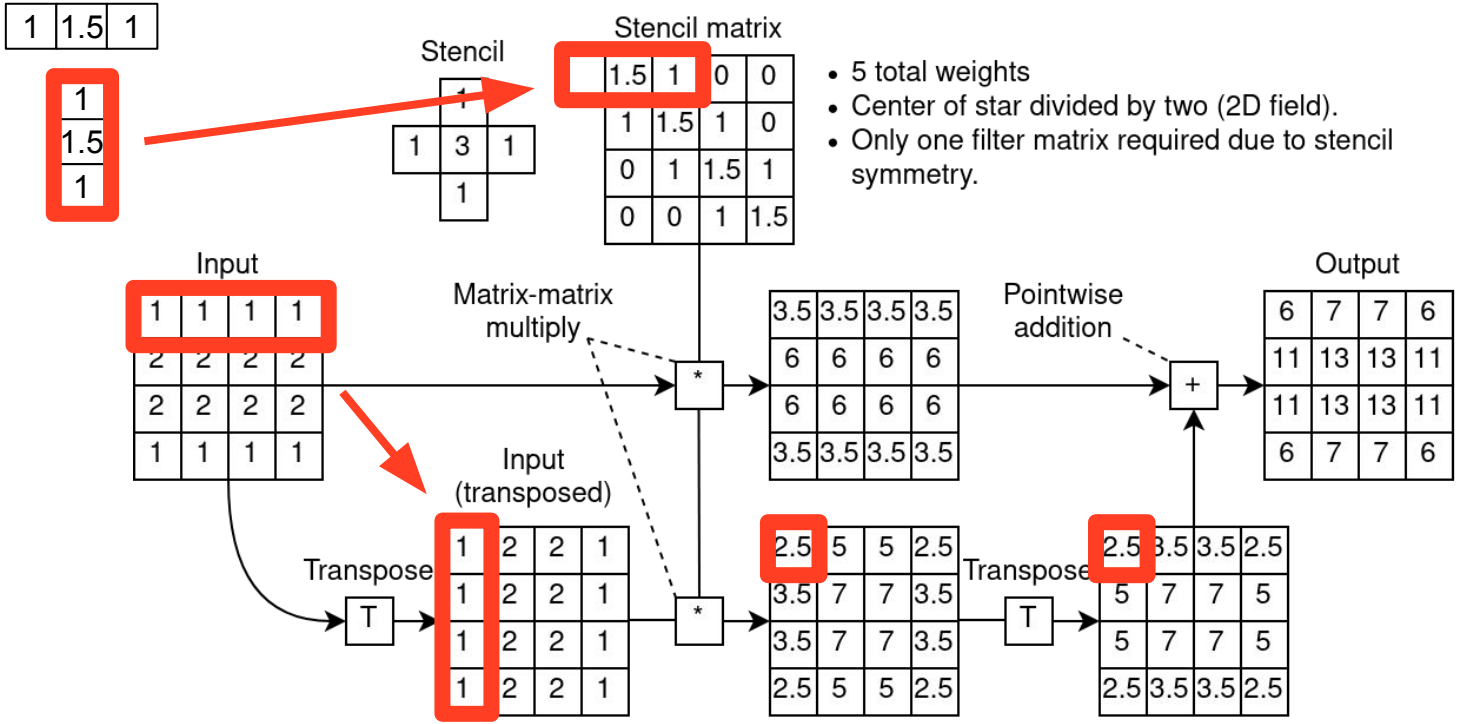
From Stencils to Tensors



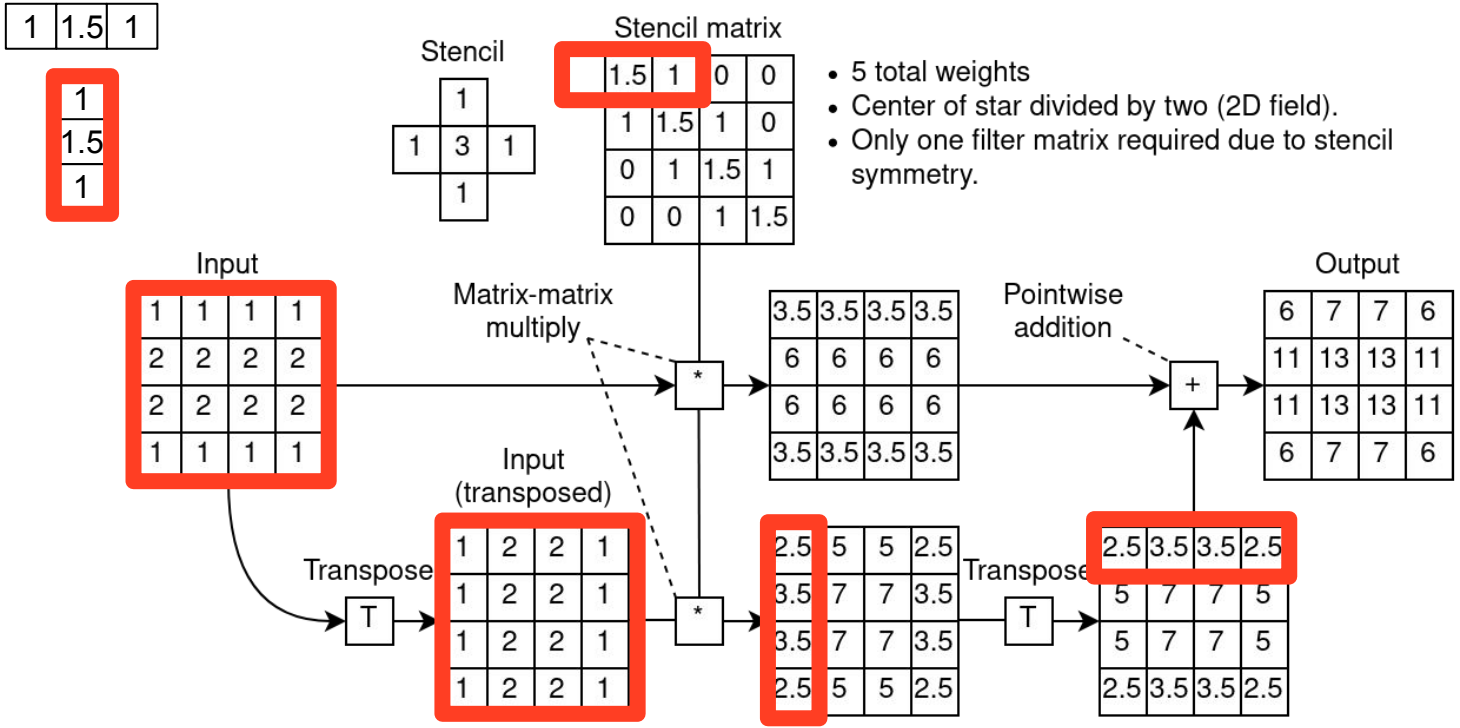
From Stencils to Tensors



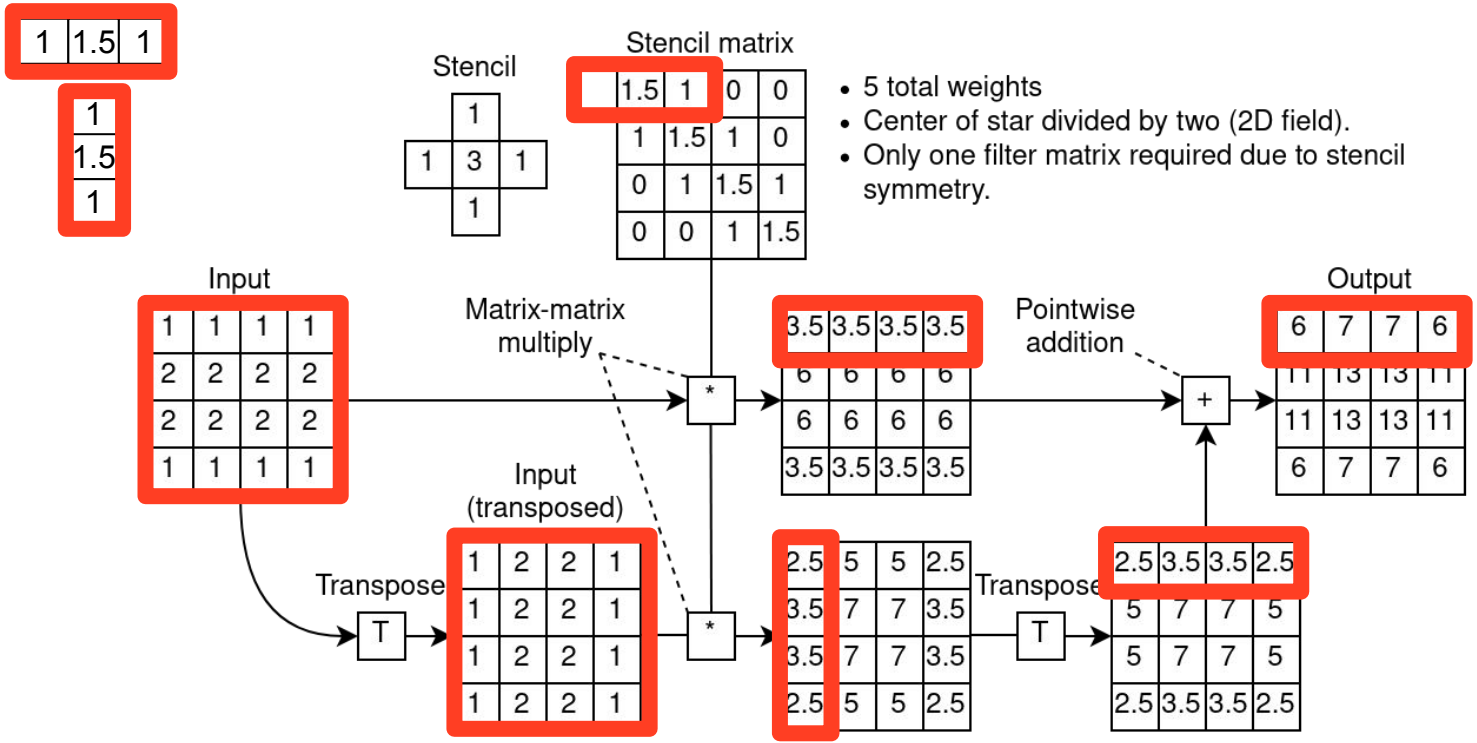
From Stencils to Tensors



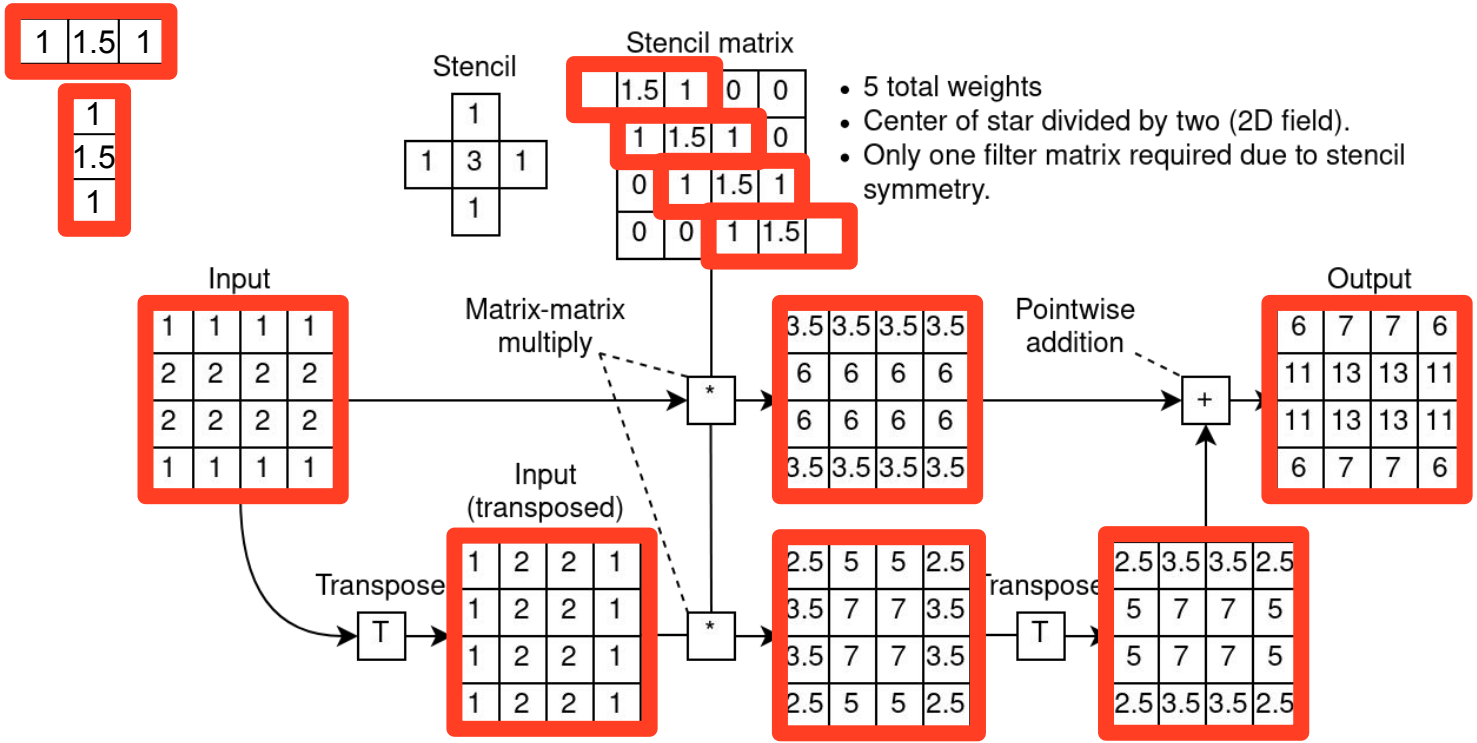
From Stencils to Tensors



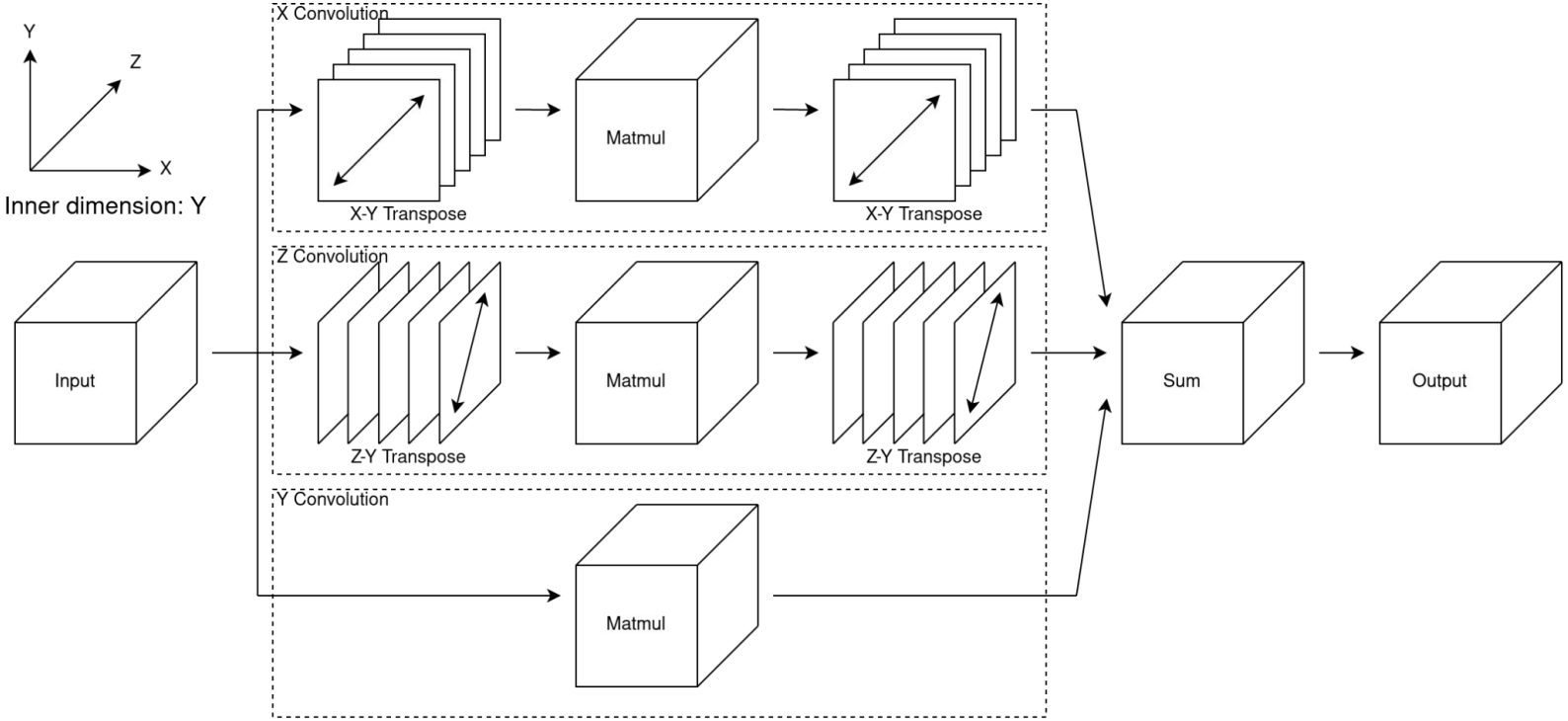
From Stencils to Tensors



From Stencils to Tensors

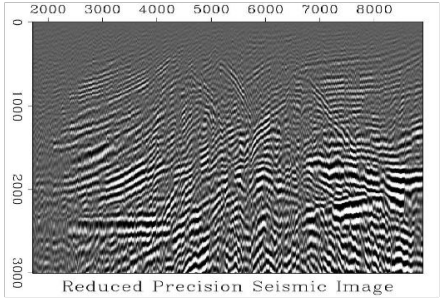


Tensor-based Solution in 3D

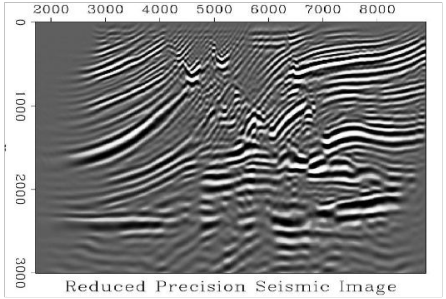


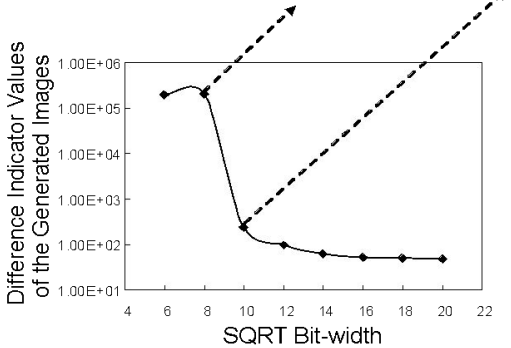
Numerical Considerations

8-bit fixed-point

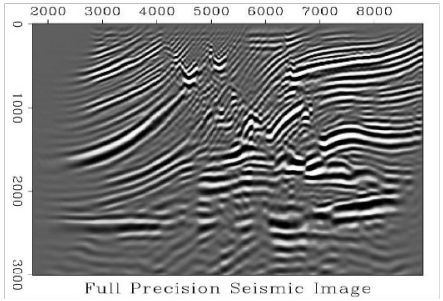


10-bit fixed-point





'true' image: single-precision floating-point



AI accelerators leverage low-precision modes

Previous FPGA work demonstrated feasibility of low bit fixed point computations

H.Fu, W. Osborne, R. G. Clapp, O. Pell, Accelerating Seismic Computations, 70th EAGE Conference, Italy, 2008

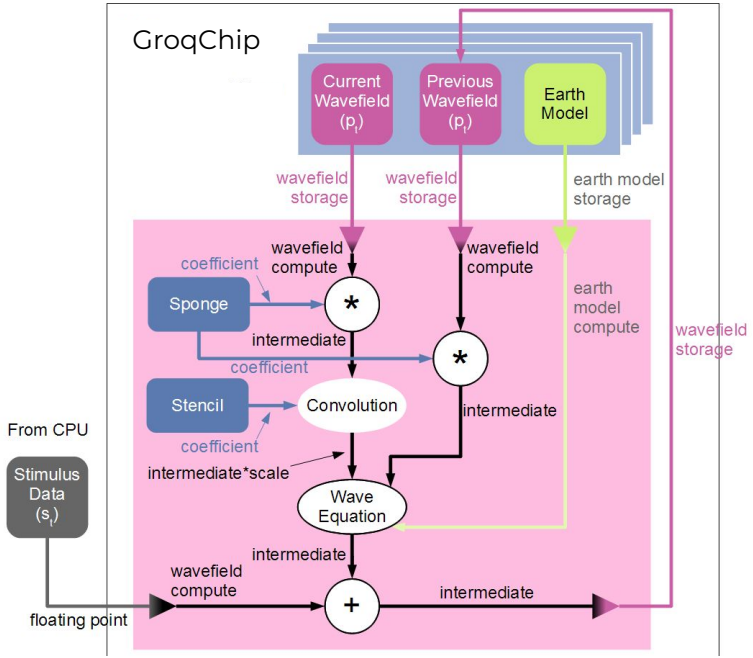
Scalability: Single Chip

Core Compute:

- Compute the wave's pressure field at the next time step based on the previous and current time steps
- Use explicit finite difference to approximate derivatives
- Earth model gives the velocity v at each point
- Apply sponging at the boundaries to prevent wave reflections

Single chip case: keep domain data in on-chip memory

Support up to 128^3 domain size



Scalability: Multi Chip

Larger Domains:

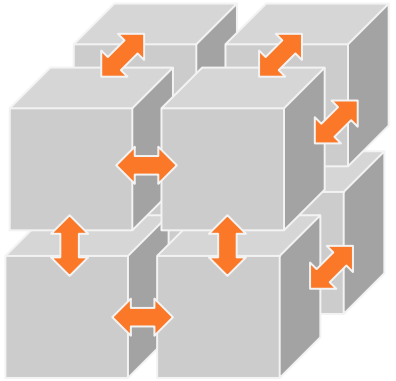
Split into subcubes

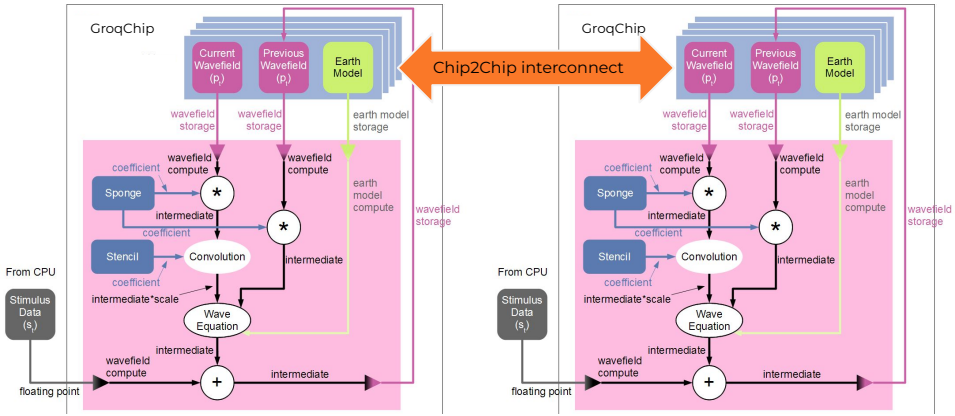
Requires halo data exchange at the edge

Use Groq RealScale Chip2Chip interconnect to avoid PCIe bottlenecks

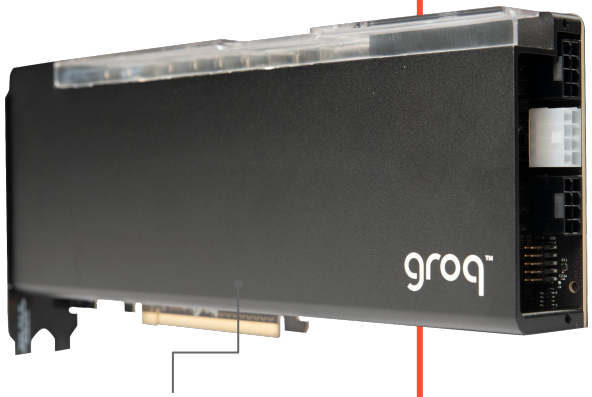
Single-chip (1) performance for 128^3: 10 Gpt/s

Multi-chip (8) performance for 512^3: 400 Gpt/s





Expansion with FPGAs



Overlapped I/O (Read & Write)
Deterministically Synchronized

Full duplex
x4 each up to 100Gbps

- Real-time inference
- Compute intensive offload

We have developed a high speed interface to an FPGA via direct chip-to-chip interconnect

- Low-latency IO via Ethernet
- Memory expansion
- Data preprocessing

groq™

Tobias Becker
tbecker@groq.com

LEARN MORE AT GROQ.COM





groq™