embedded VISION SUMMIT 2018

NovuTensor: Hardware Acceleration of Deep Convolutional Neural Networks for AI

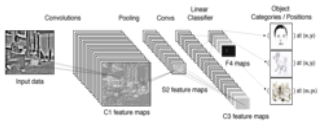
Miao (Mike) Li

May 23, 2018

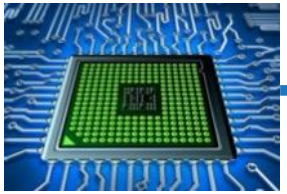
NovuMind --- Super Computing + Algorithm + IC

NovuStar
Supercomputer

Training



Neural
Network
Models



NovuTensor
AI Chip

NovuBrain
AI Module

Full-Stack AI
Solutions
NovuMind

Training ability

Model accuracy

Get Data

Industry

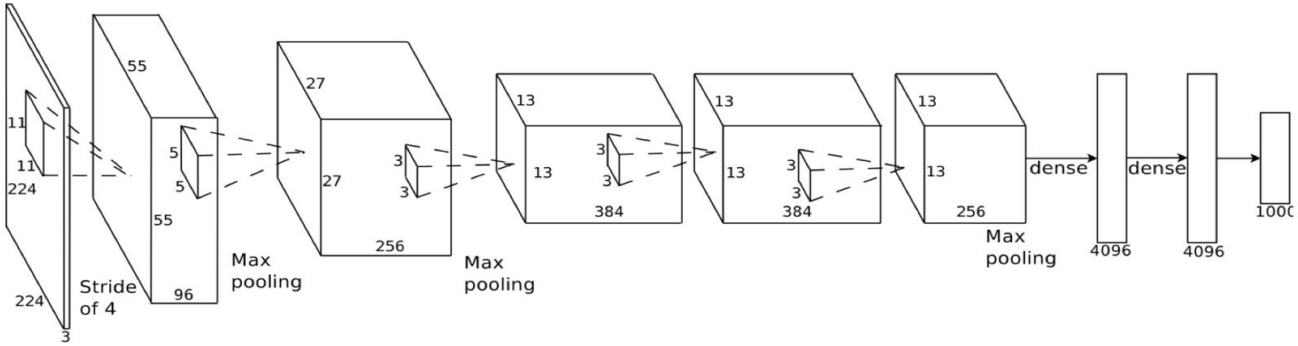
User Friendly

Corporation

Customer
& Alliance

Copyright © 2018 NovuMind Inc.

Deep Learning CNN Is Based on 3D Convolution

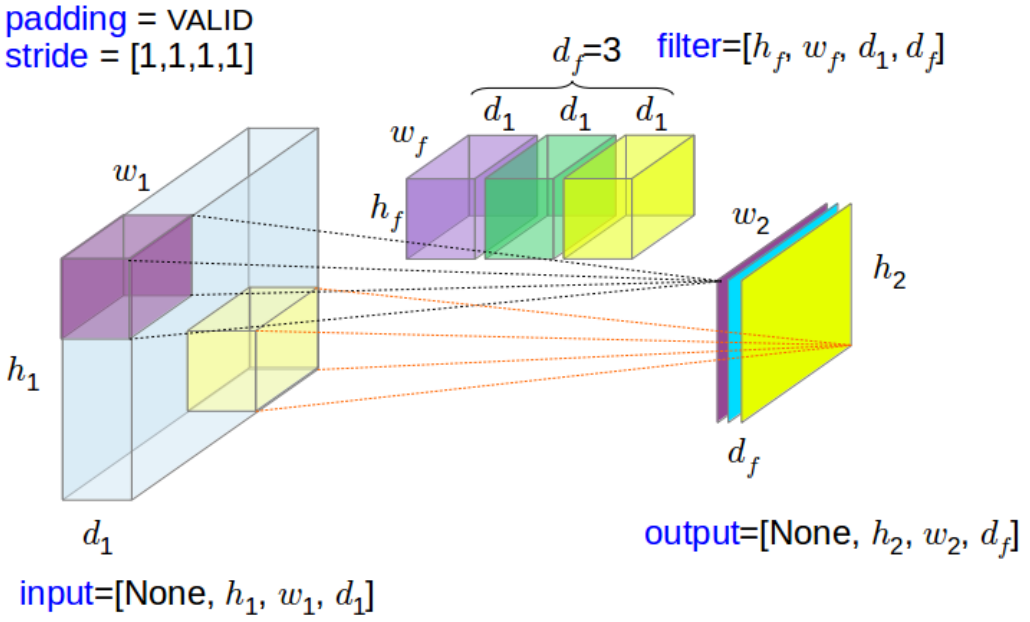


- The core of CNN is 3-D Convolution Computation
- Deeper CNN needs more 3-D Convolution Computation
  - AlexNet -> VGG -> ResNet -> …
- DCNN is a network topology composed of basic computation elements (e.g., 3x3 3-D Convolutions)

Copyright © 2018 NovuMind Inc.

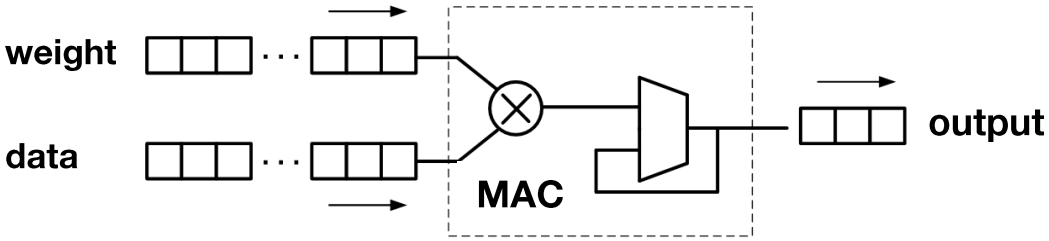
Basic Computation Element: 3D Convolution

- Input Feature Map (IFM), Filter and Output Feature Map (OFM) are all 3D data
- 3D Convolution:
  - Inner product, Summation for each OFM pixel
  - Stride filter over IFM to form other OFM pixels
- Repeat with another filter on the same IFM

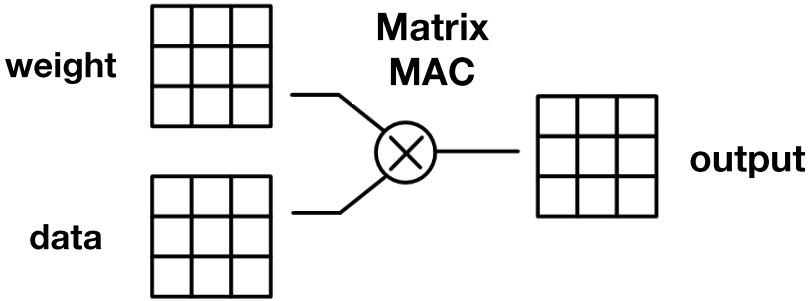


Copyright © 2018 NovuMind Inc.

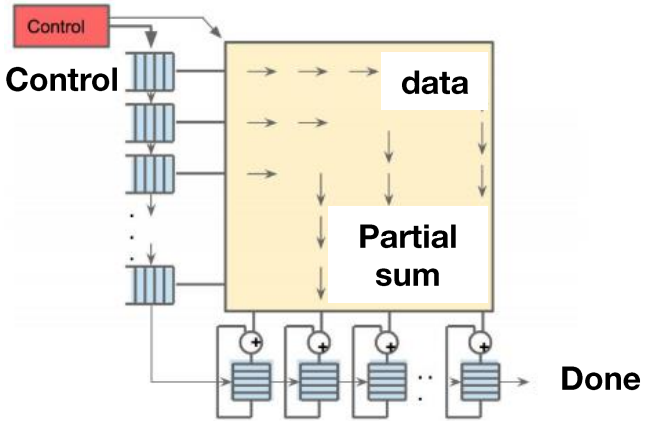
Conventional 1D, 2D Implementation of 3D Convolution



1D Operation (DSP)



2D Operation (GPU, TPU)

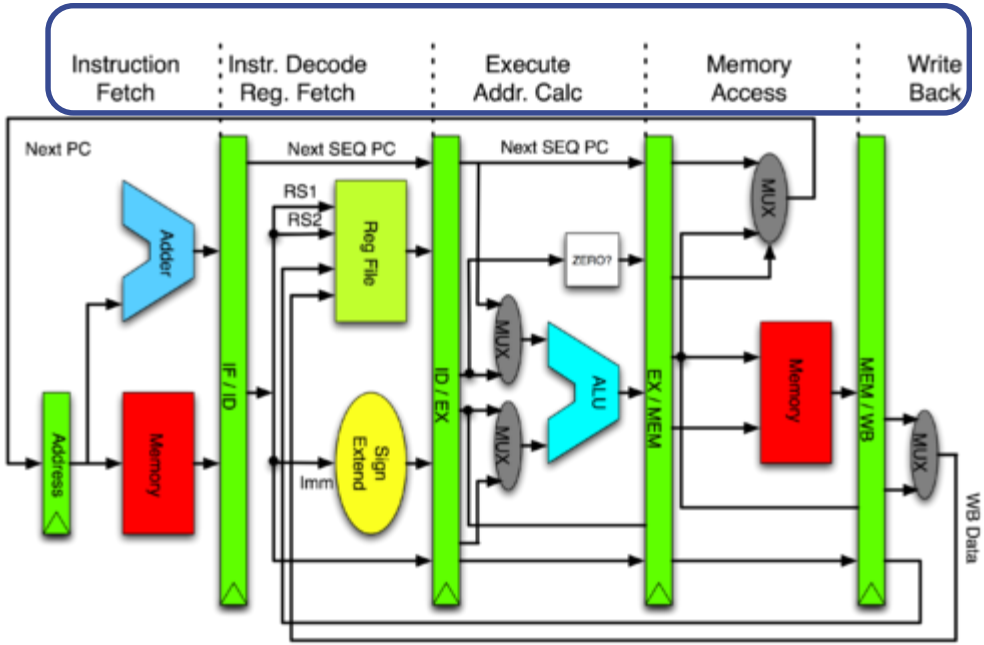


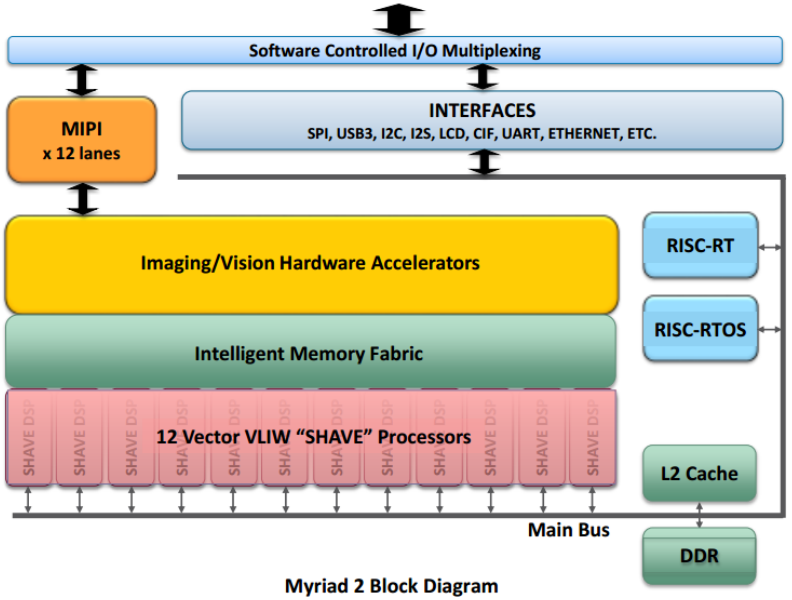
TPU with 2D Systolic Array

Copyright © 2018 NovuMind Inc.

1D Pipelining, Multi-Core

Processor Pipelining Speed Up Opportunity



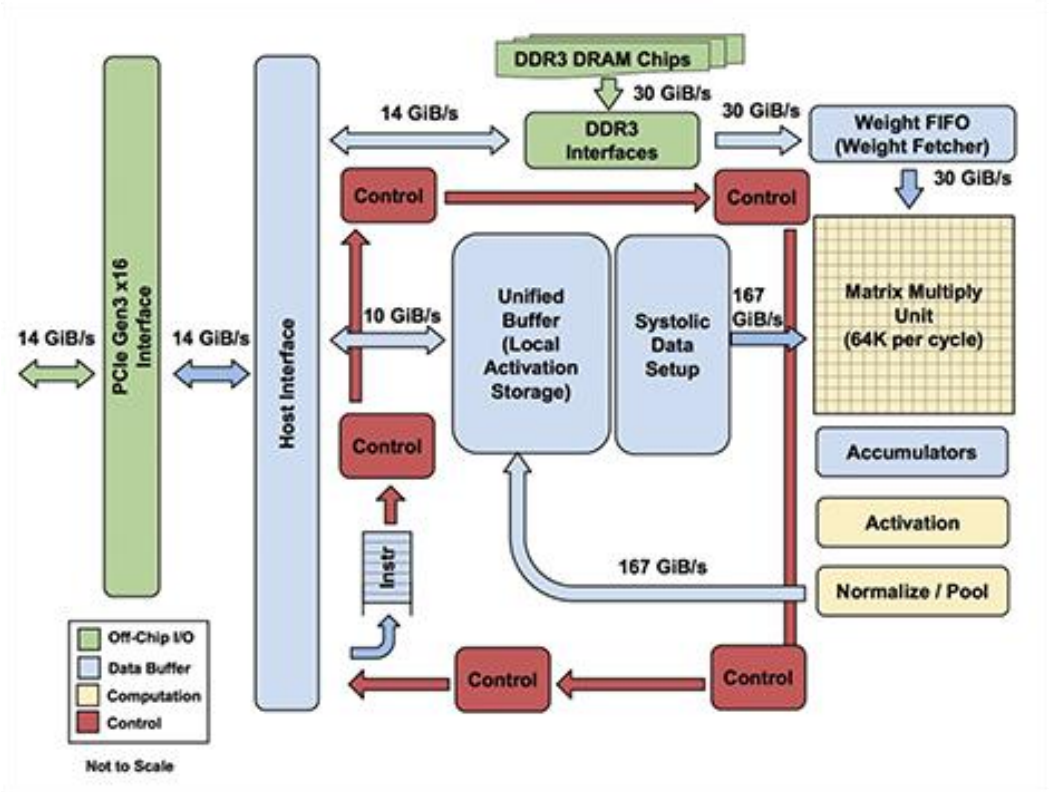


Myriad 2 Block Diagram

AI DSP Example Block Diagram
Parallel Cores for Acceleration
Peak 1T/s, Average 0.2T/s
Efficiency: 20%

Copyright © 2018 NovuMind Inc.

2D Overhead in Computing 3D Data (TPU)

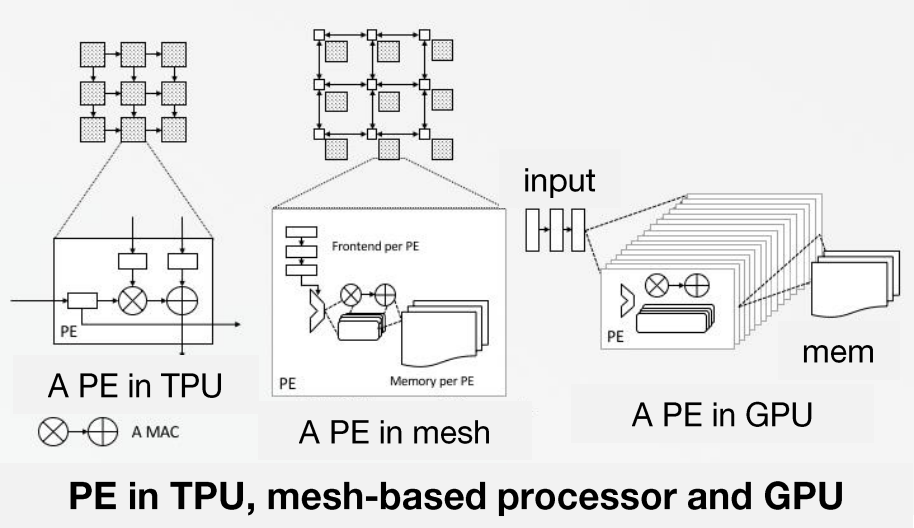


1080Ti Throughput:
Peak: 11TOPS/s
Effective: 2.4TOPS/s
(ResNet18)
Efficiency: 22%

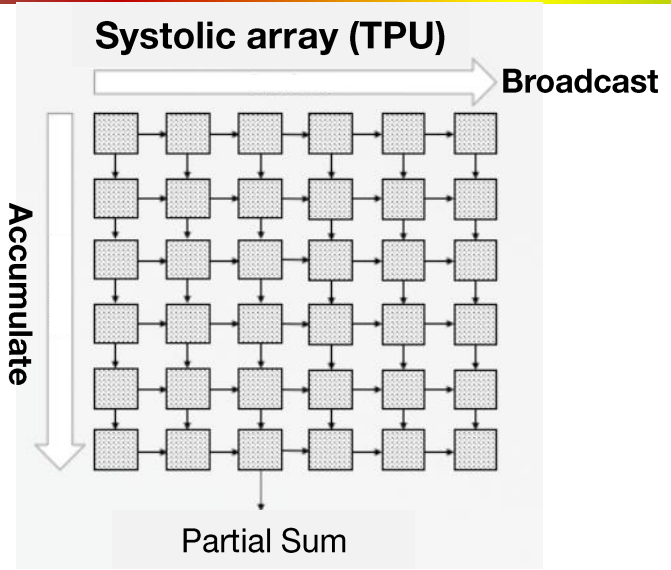
TPU Throughput:
Peak: 96TOPS/s
Average: 23TOPS/s
Efficiency: 24%

Copyright © 2018 NovuMind Inc.

2D Accelerators for Deep Learning



PE in TPU, mesh-based processor and GPU



Mesh Structure and Systolic Array

- Overhead to prepare the 3D data into 2D data
- Overhead to feed data
  - TPU uses 256x256 systolic array->needs 65536 data to fill
- Systolic Array got 2D parallel acceleration – only when data is loaded

Copyright © 2018 NovuMind Inc.

1D, 2D, 3D – An Ice Cream Example

1-D [Point/Line-Vector]

CPU / DSP

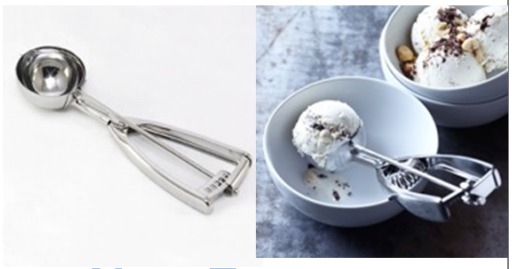
Low utilization (20-30%)

2-D[surface-Matrix]

GPU, TPU

Low utilization (20-30%)
High Latency (50-300ms)

3D[Tensor]



NovuTensor

High utilization (70-80%)
Low latency (<3ms)
Low power

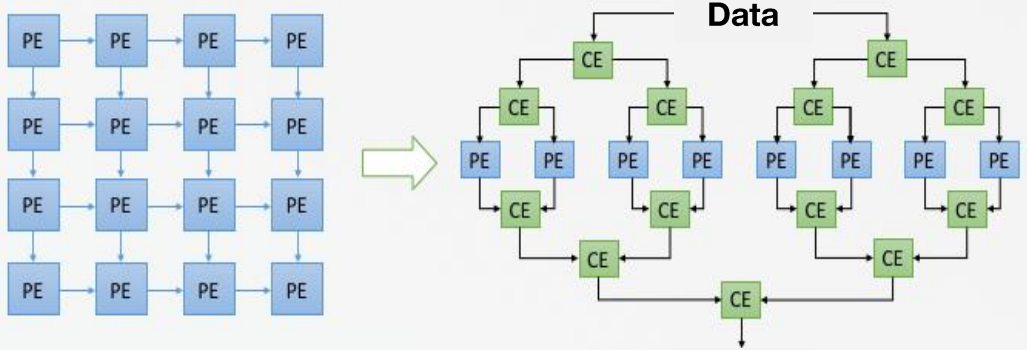
Copyright © 2018 NovuMind Inc.

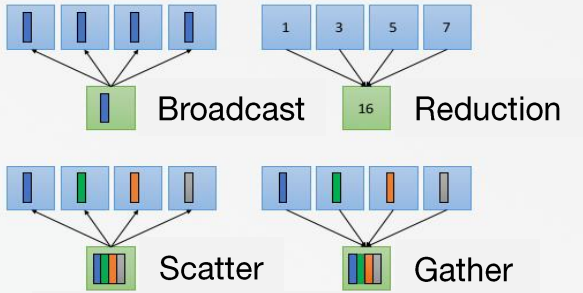
How Do You Eat Ice Cream?

- Data Movement:
  - Need a scoop,then the art of scooping
  - Need a spoon or cone that is mouth size
- Arithmetic:
  - Need a mouth to consume the ice cream

Copyright © 2018 NovuMind Inc.

From Mesh to Collective Streaming





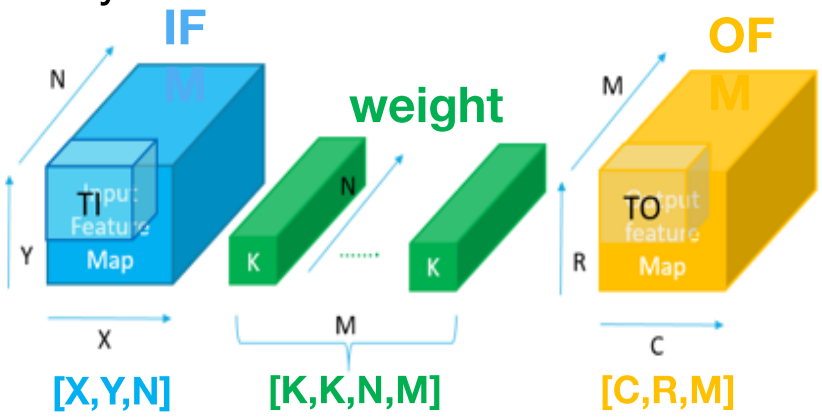
Four types of operations

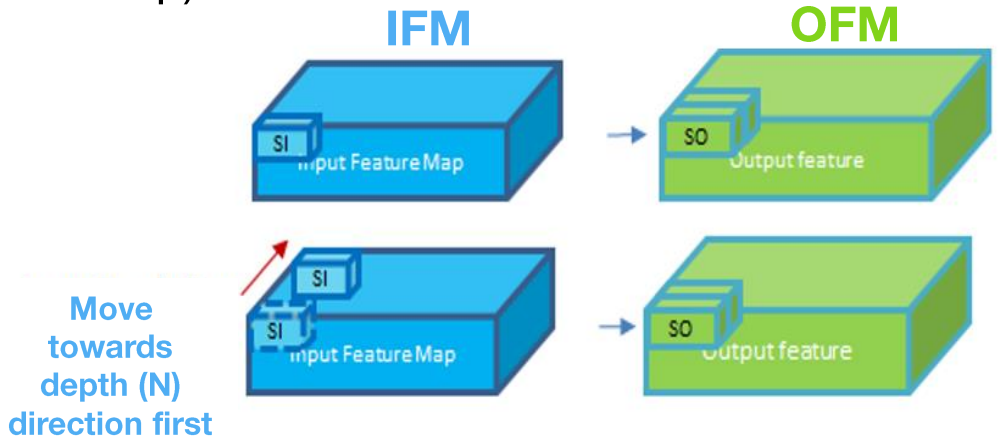
- Unnecessary Assumptions with Mesh:
  - dqual distance; deighboring data dependency
- Collective Streaming, and CNN 3D Data Processing: Data independency b/w PEs:
  - no data flow
  - no need to be rectangular, PEs can be flexible to have dozens to thousands of MACs

Copyright © 2018 NovuMind Inc.

The Scooping: Tile Based Operation (patented)

- NovuTensor does not need all tensor data at once
- Instead, it divides input/output of one DCNN layer into may tiles, and each time works on one tile
- When all tiles of one layer finishes, (the scoop) moves on to next layer





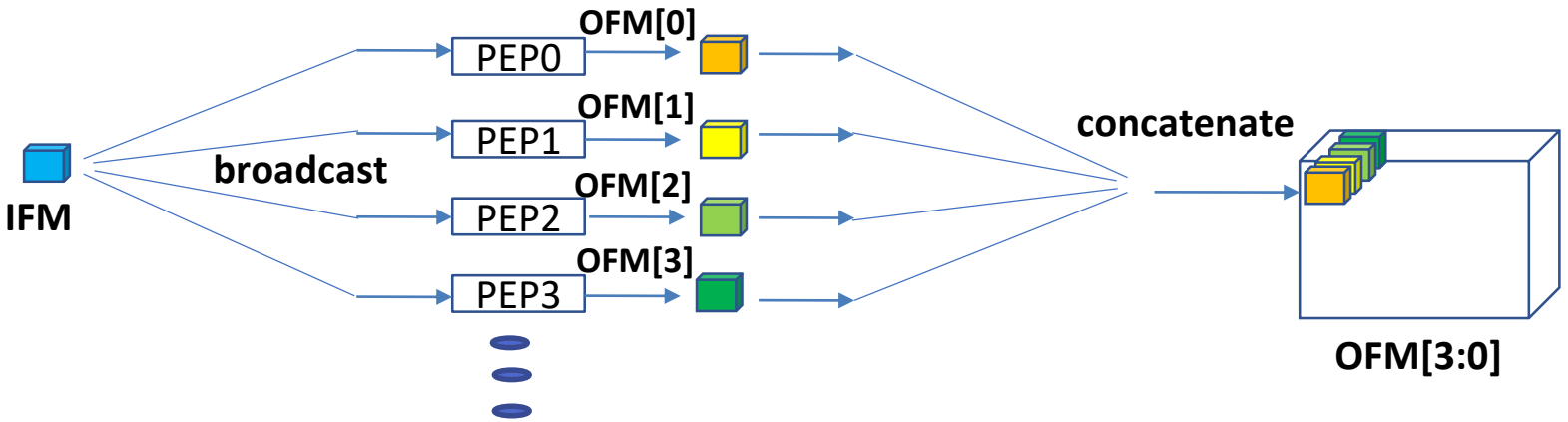
Copyright © 2018 NovuMind Inc.

Efficiency Potentials

- Efficient Data Movement
  - Reduce unnecessary data movement off chip and on chip
  - Fully exploit data sharing, re-using:
    - IFM sharing between OFMs
    - Weight sharing within IFM (X-Y)
    - Weight sharing within batch
- Efficient Arithmetic Logic
  - Reduce unnecessary computation

Copyright © 2018 NovuMind Inc.

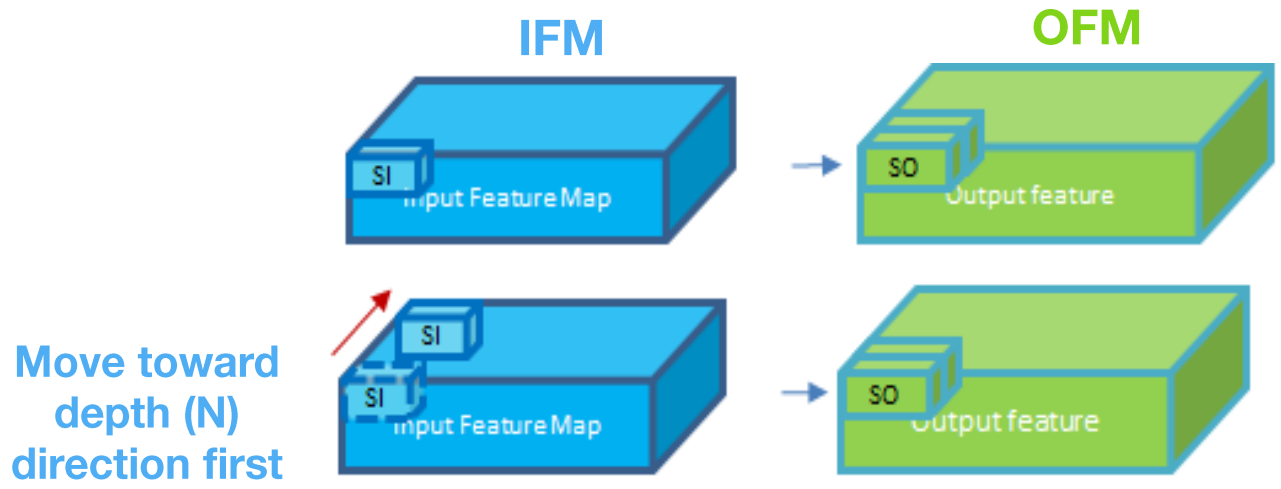
IFM Sharing Among Multiple PEPs



For multiple PEPs, each PEP is sharing the same IFM while generating different OFMs (IFM sharing)

Copyright © 2018 NovuMind Inc.

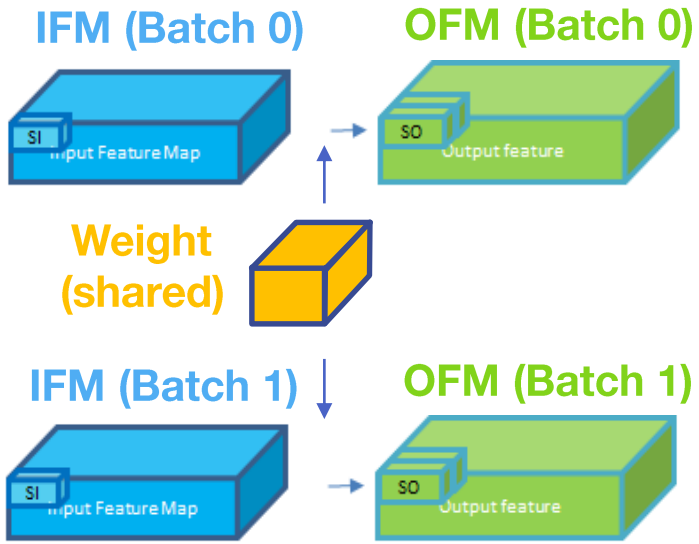
Weight Sharing Inside Tile



- Inside each tile, weight is shared in X-Y direction
  - Taking advantage of mathematic characteristics of convolution

Copyright © 2018 NovuMind Inc.

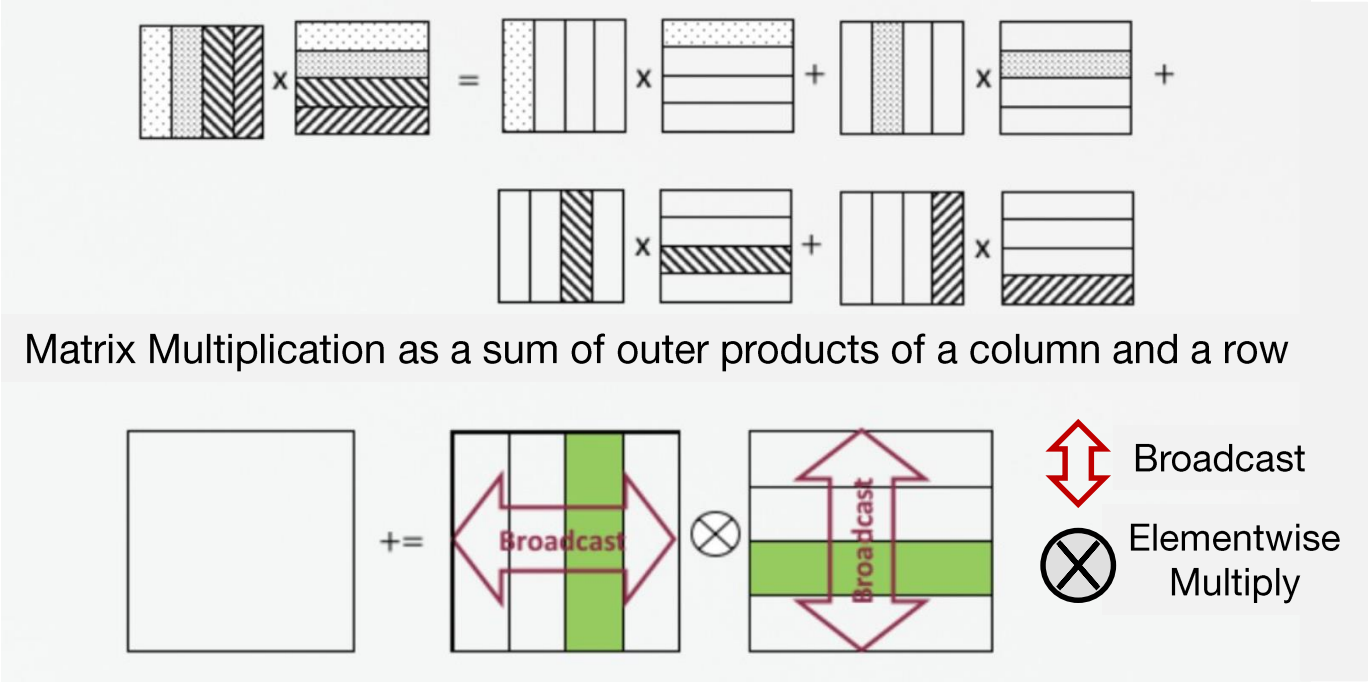
Weight Sharing with Batches



- For multi-batch input, different batches share same weight

Copyright © 2018 NovuMind Inc.

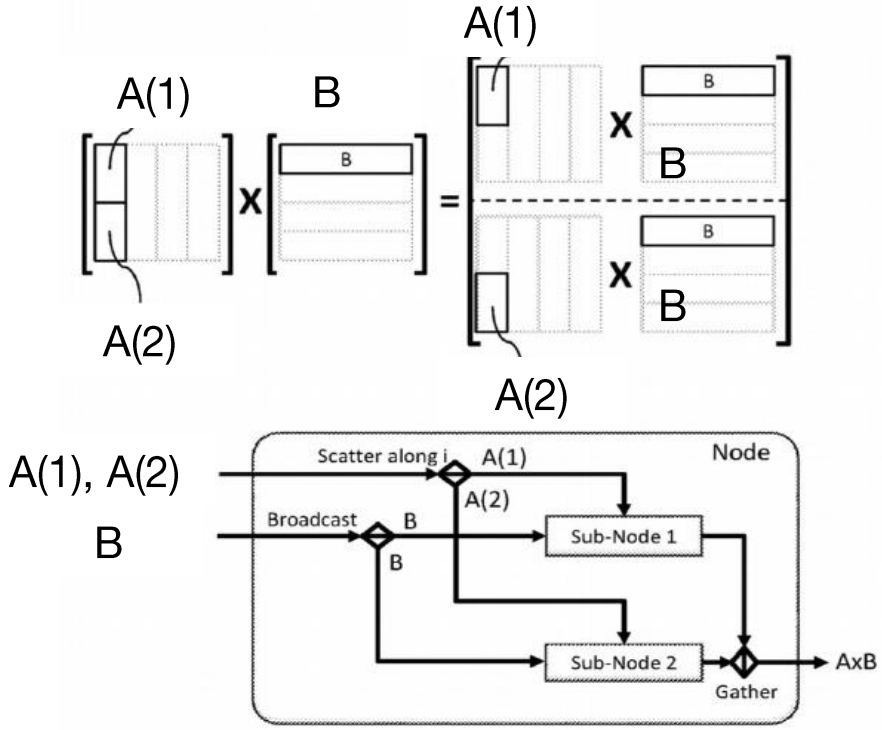
Math Behind Tiling: Outer Product (Patented)



Each outer product consists of 1-D data, 1-D broadcasts and 2-D computations

Copyright © 2018 NovuMind Inc.

Hardware Architecture for Tiling



Each A and B is broadcast to multiple processing engines (sub-nodes)

Copyright © 2018 NovuMind Inc.

NovuTensor at CES2018

	NT-BB FPGA	Server GPU
Performance (FPS)	301	666
Power (W)	12	250

1/20 Power Consumption

1/2 Performance

ResNet18

ImageNet Classification

Copyright © 2018 NovuMind Inc.

Major AI Chips in Market

Name	Power Consumption (w)	Performance (Effective TOPS, VGG16)	Performance Power Ratio (TOPS/W)
#1 Embedded	1.5 (500MHz)	0.2	0.13
#2 Smartphone	NA (use as IP)	0.19	NA
#3 Smartphone	NA (use as IP)	0.3	NA
NovuMind NT-BB ("BlackBear")	**12~18**	**2.5**	**0.21 (FPGA)**
NovuMind NT-GZ ("Grizzly")	**5~10**	**11 (peak 14)**	**2.25 (ASIC)**
Embedded GPU	10	0.4	0.04
Data Center AI ASIC	40	23	0.58
Server GPU	250	6.6	0.026

Low Power
Low Performance
Cannot meet the requirements of most AI computing needs.

High Power
High Performance
Cannot meet the requirements of most edge application scenarios.

NovuTensor: Server performance within embedded power

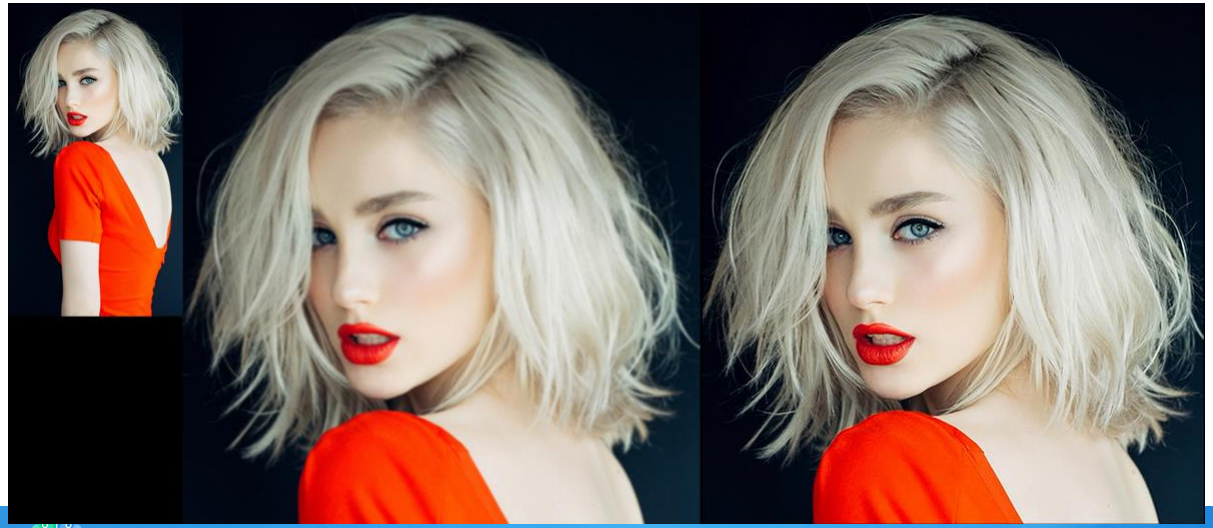
The most power-efficient chip on the market!

Copyright © 2018 NovuMind Inc.

Super Resolution at CES2018: Beauty Made by Muscle

The world's 1st in

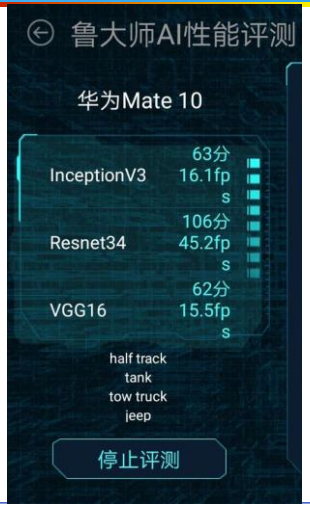
using AI super resolution technology to upscale low-resolution video into 4k or 8k UHD video in real time



- Intelligently fill the details to ensure ultra-high quality
- Solve the problem of media content lacking 4K / 8K resolutions for online video and air broadcasting

Copyright © 2018 NovuMind Inc.

NovuTensor Kestrel: When Grizzly Flies



Mobile AI IC Name	Power Consumption (W)	Performance (Effective TOPS, VGG16)
NovuTensor Kestrel	**1**	**5 (160 fps)**
iPhone X A11 Bionic	NA	0.33
HUAWEI HiSilicon Kirin970	NA	0.5
Samsung Galaxy S8 Snapdragon 835	NA	0.2

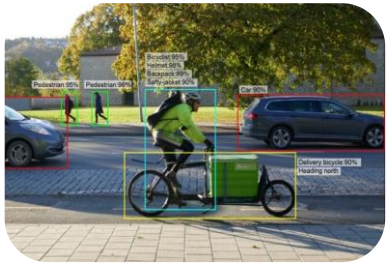
Copyright © 2018 NovuMind Inc.

Call to Action

Please visit Booth #704 and our website
http://www.novumind.com

Copyright © 2018 NovuMind Inc.

Resources



- **CES January 2018:**
  http://share.gmw.cn/tech/2018-01/11/content_27322512.htm?from=message&isappinstalled=0

- **HP Enterprise + NovuMind at GTC October 2017**
  "Improving Deep Learning Scalability on HPE Servers With NovuMind: GPU RDMA Made Easy"
  https://www.leiphone.com/news/201710/GG9umC93Gtav2Eac.html

- **Real Time Endoscopic Application at West China Hospital, July~August 2017**
  SCTV Video: http://www.iqiyi.com/v_19rrefrw1g.html
  People's Daily News: http://health.people.com.cn/n1/2017/0801/c14739-29440444.html
  CCTV: http://jiankang.cctv.com/2017/07/28/ARTIE26PkbwL9t34pFS8jbIW170728.shtml

- **AI Start-up NovuMind Helps European City Improve Their Green Life**
  https://smartcitiesworld.net/news/news/ai-equals-a-greener-life-for-trondheim-2407

Copyright © 2018 NovuMind Inc.