2021년 IT21
Global Conference
Human in SW, SW in Human
Session 4-4

Building the next-generation high performance AI inference chip for cloud and data center

백준호 대표 (퓨리오사AI)

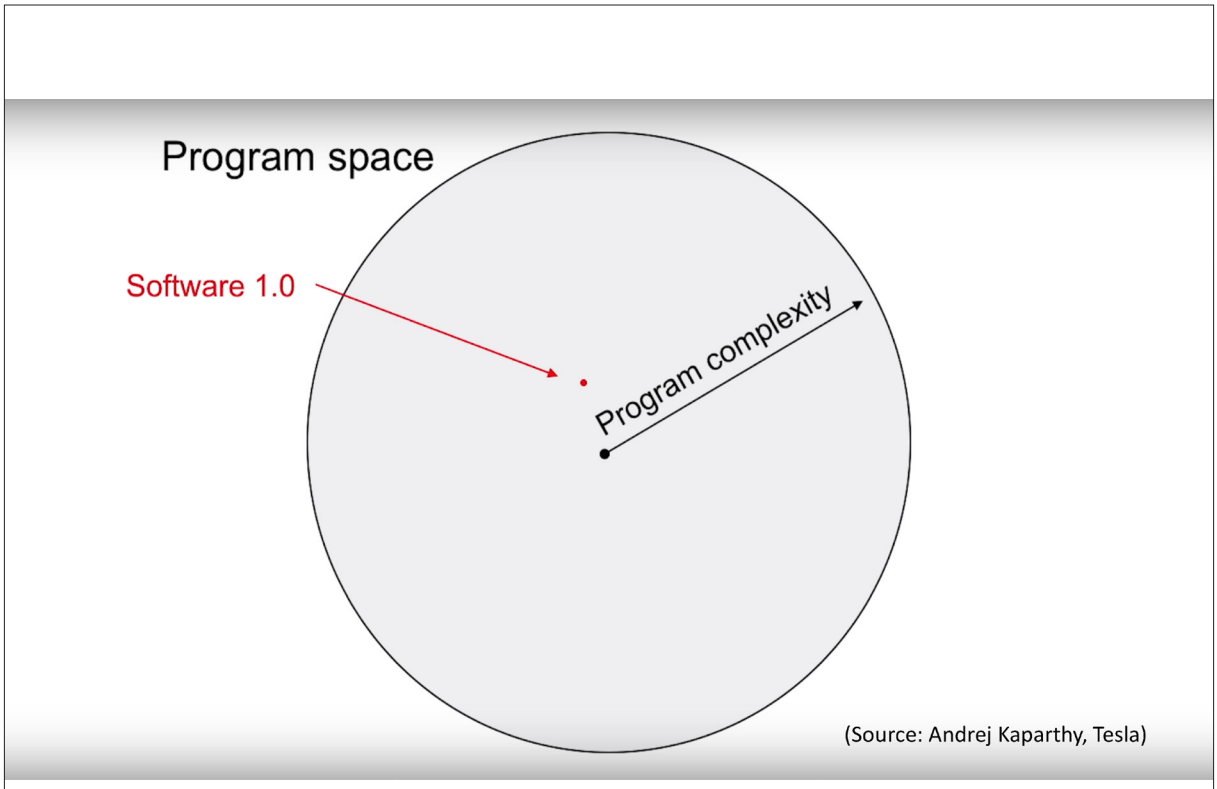
FuriosaAI builds high-performance AI inference coprocessors that can seamlessly integrate with the emerging data center computing infrastructure to provide scalable AI-driven services. The goal of FuriosaAI Renegade architecture is to natively run the emerging deep neural networks with the highest performance in real-time while maintaining the best energy efficiency possible. To achieve that goal, we first analyze the characteristics of underlying tensor operations, examine various shapes of tensor data and memory footprints of our main target models, and then optimize our architecture based on these analyses. To prevent our architecture from overfitting to a particular model, we generalize our architecture to be the superset of target models. To provide both flexibility and efficiency, Renegade provides ISA and programming models and many parts of data-paths in Renegade are configurable, allowing the software to control details of hardware operations during runtime. This talk will also introduce the approach, design and engineering challenges of our team.

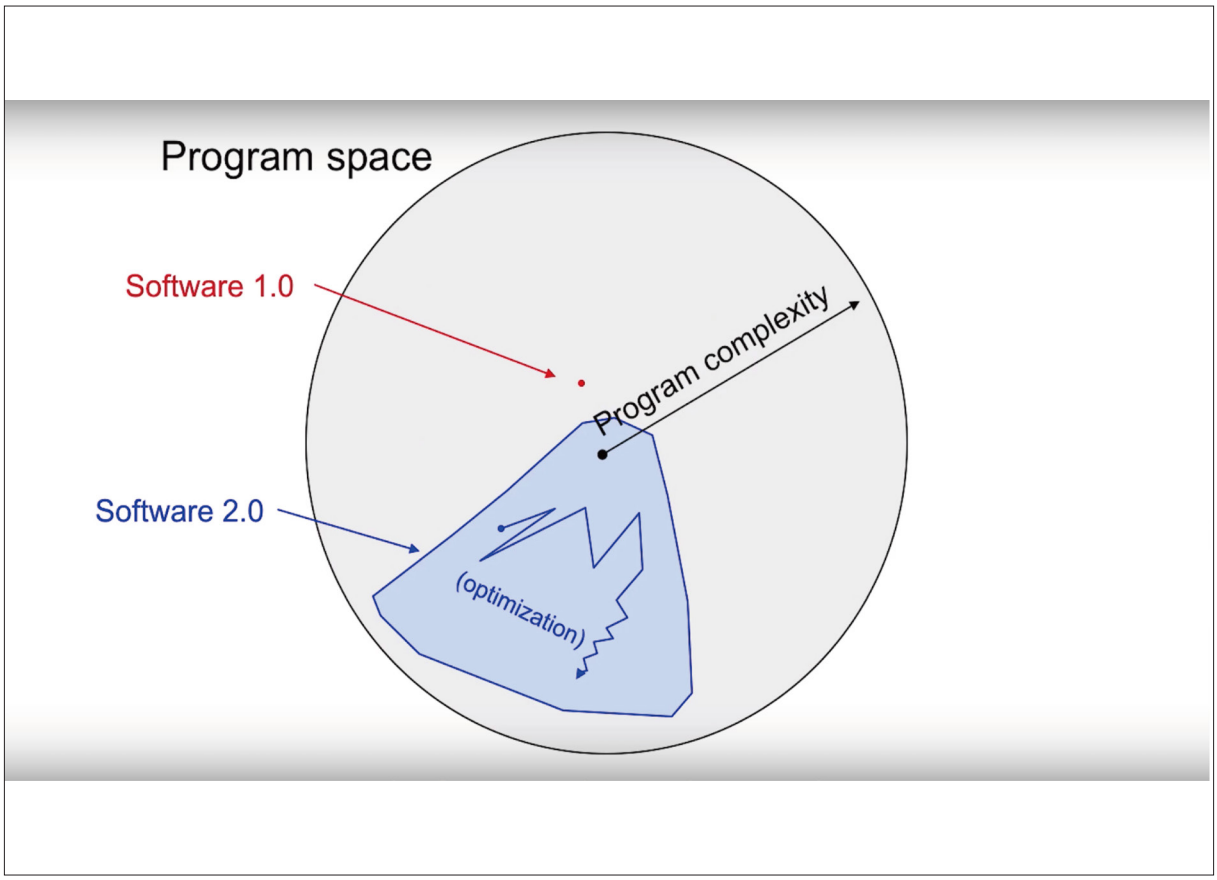
약 력

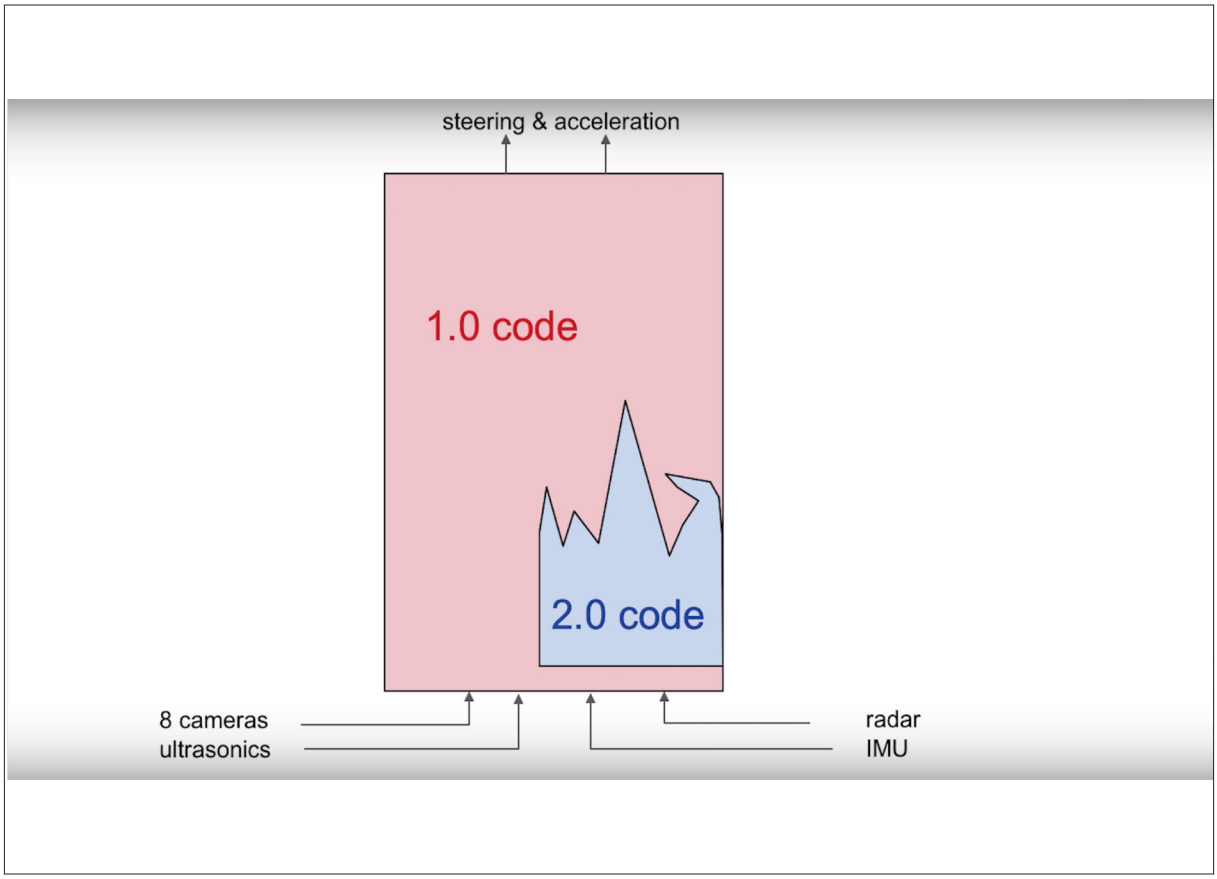
Joonho Baek is the founder and CEO of FuriosaAI Inc. He is leading Furiosa team to build the industry-defining AI chip products deployed in mass volume for emerging AI computing infrastructure. Previously he worked for Samsung and AMD, building world-class semiconductor products.

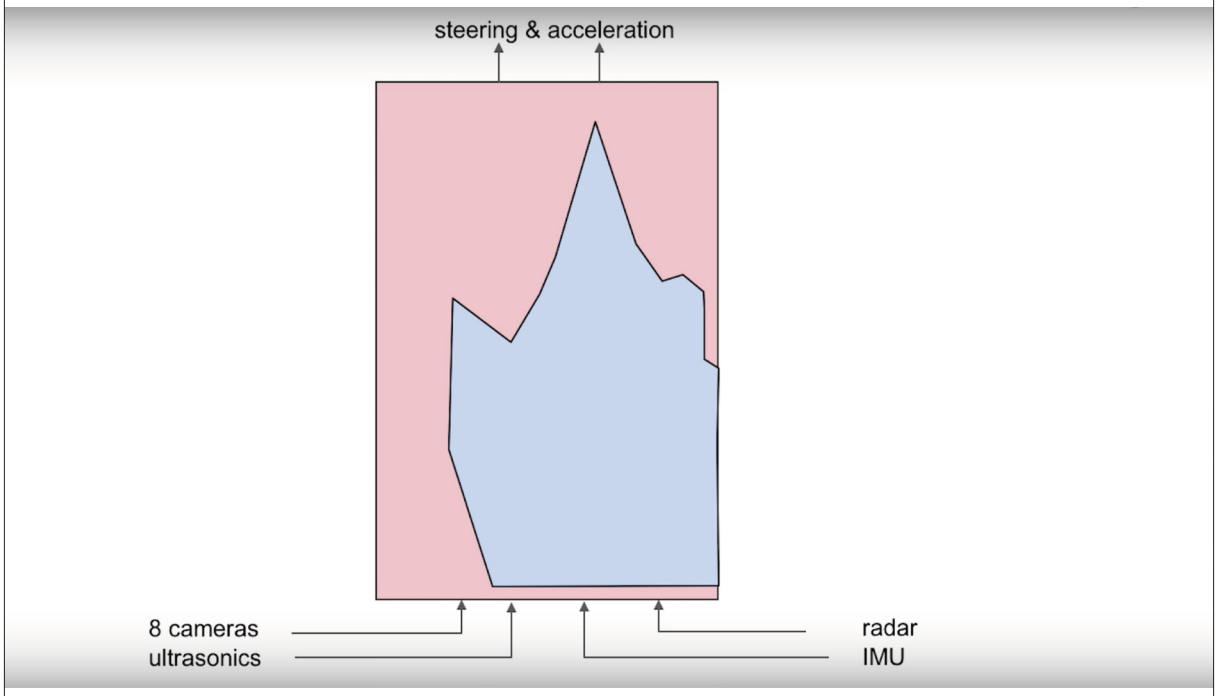
Challenges of Building AI chips driving future applications

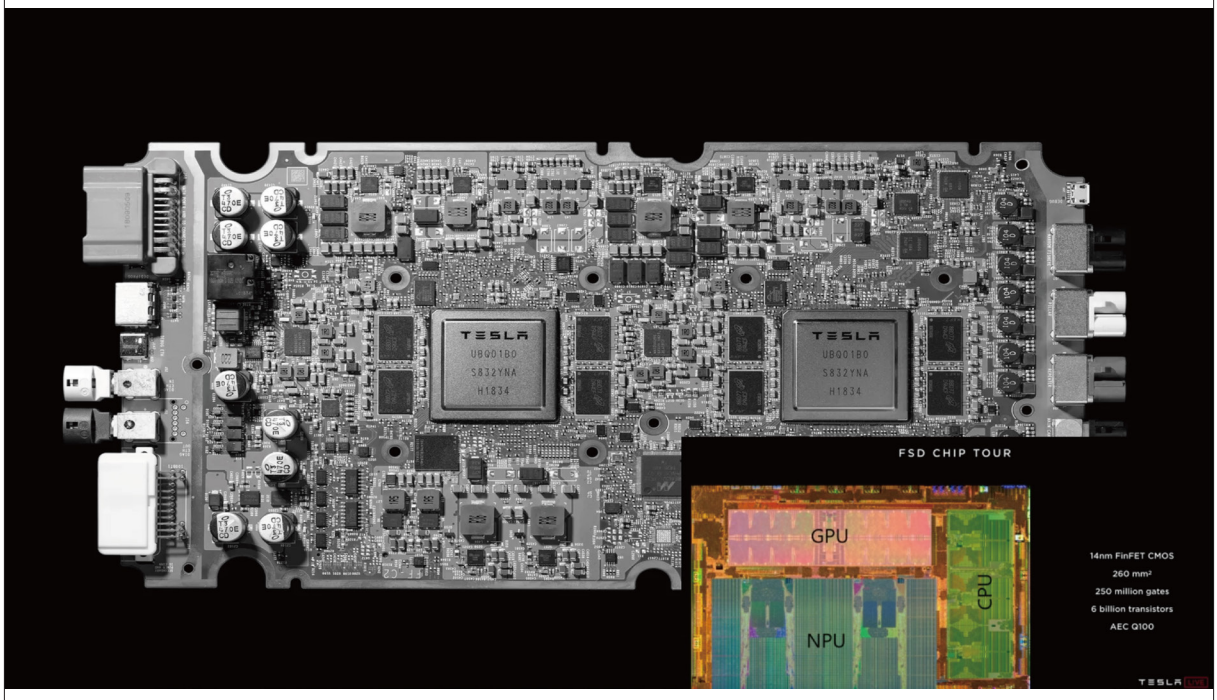
FuriosaAI 백준호 대표

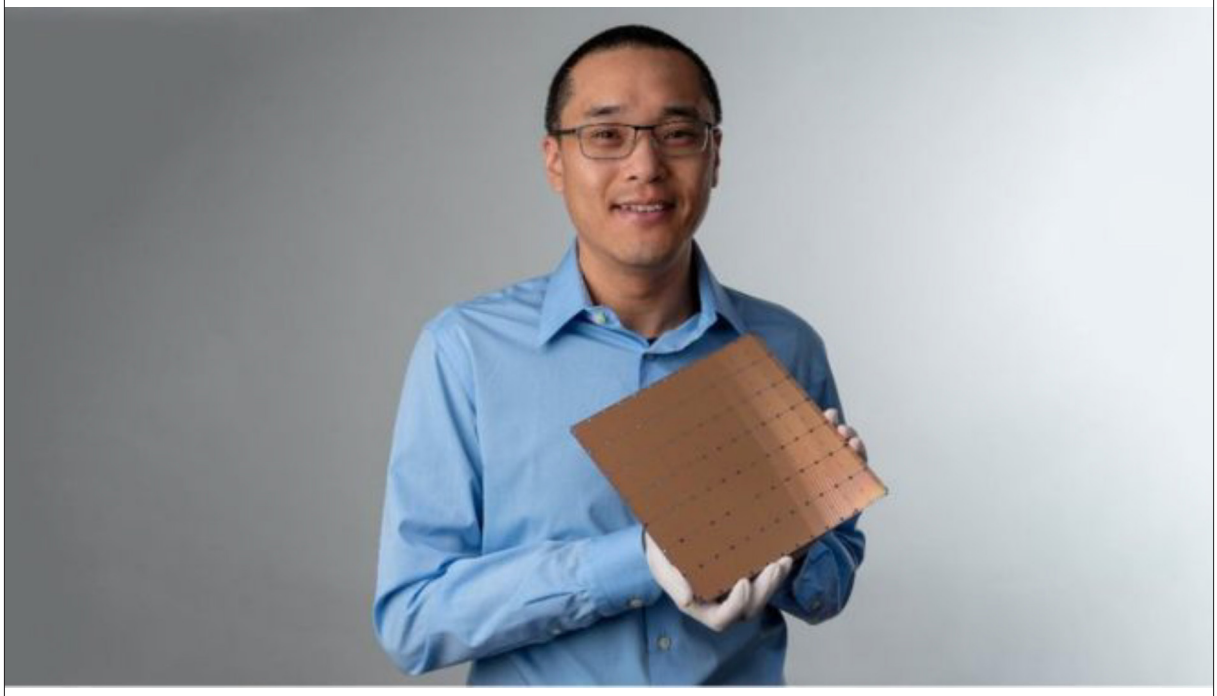


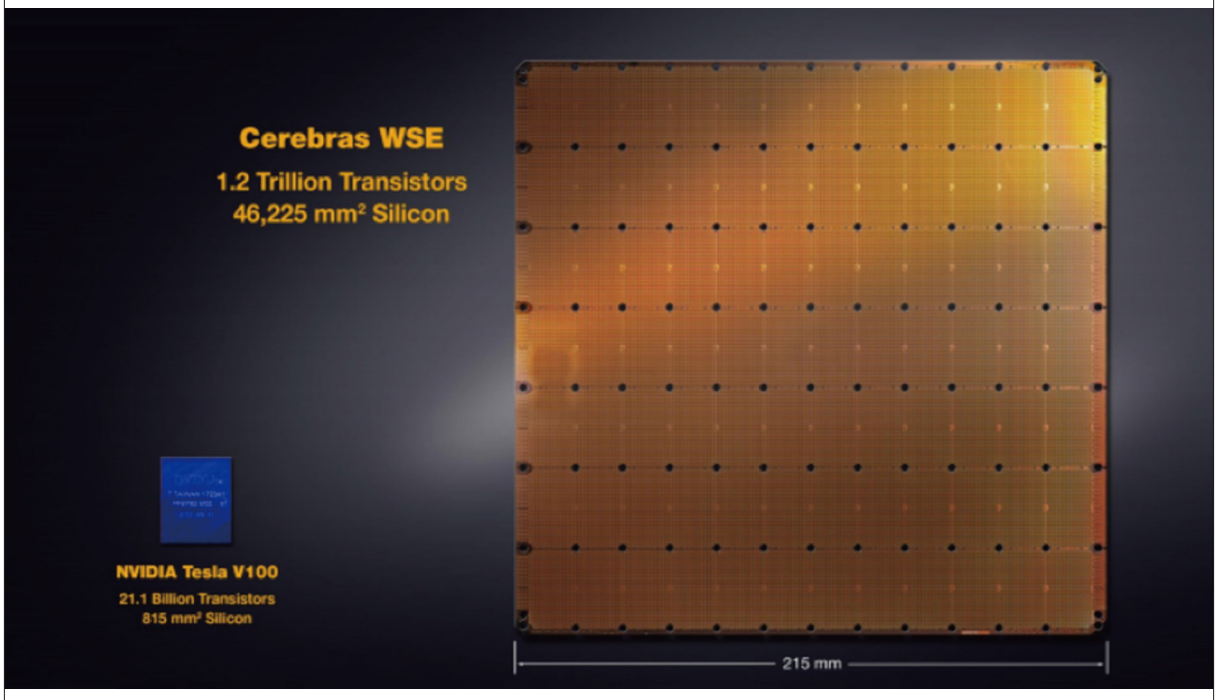

Program space
Software 1.0
Program complexity
Software 2.0
(optimization)


steering & acceleration
1.0 code
2.0 code
8 cameras
ultrasonics
radar
IMU

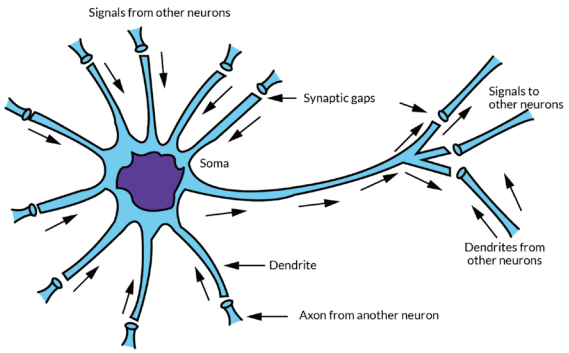

steering & acceleration
8 cameras
ultrasonics
radar
IMU

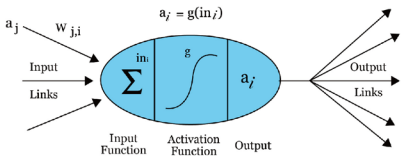

TESLA
UBQ01B0
S832YNA
H1834
TESLA
UBQ01B0
S832YNA
H1834
FSD CHIP TOUR
GPU
CPU
NPU
14nm FinFET CMOS
260 mm²
250 million gates
6 billion transistors
AEC Q100



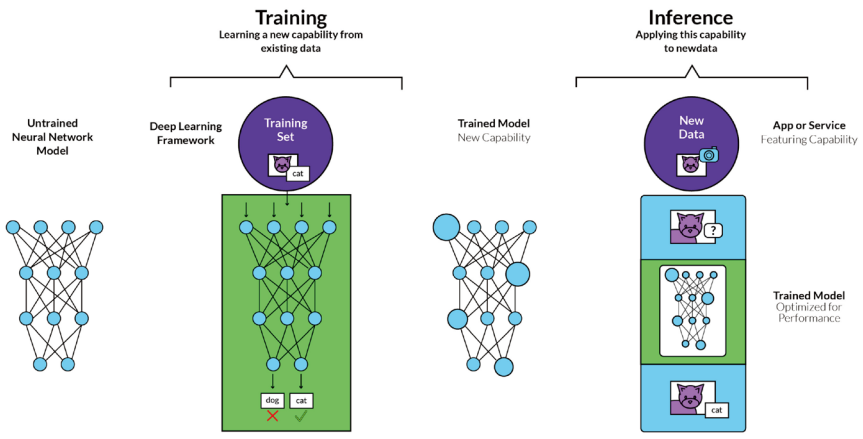

Cerebras WSE
1.2 Trillion Transistors
46,225 mm² Silicon
NVIDIA Tesla V100
21.1 Billion Transistors
815 mm² Silicon
215 mm

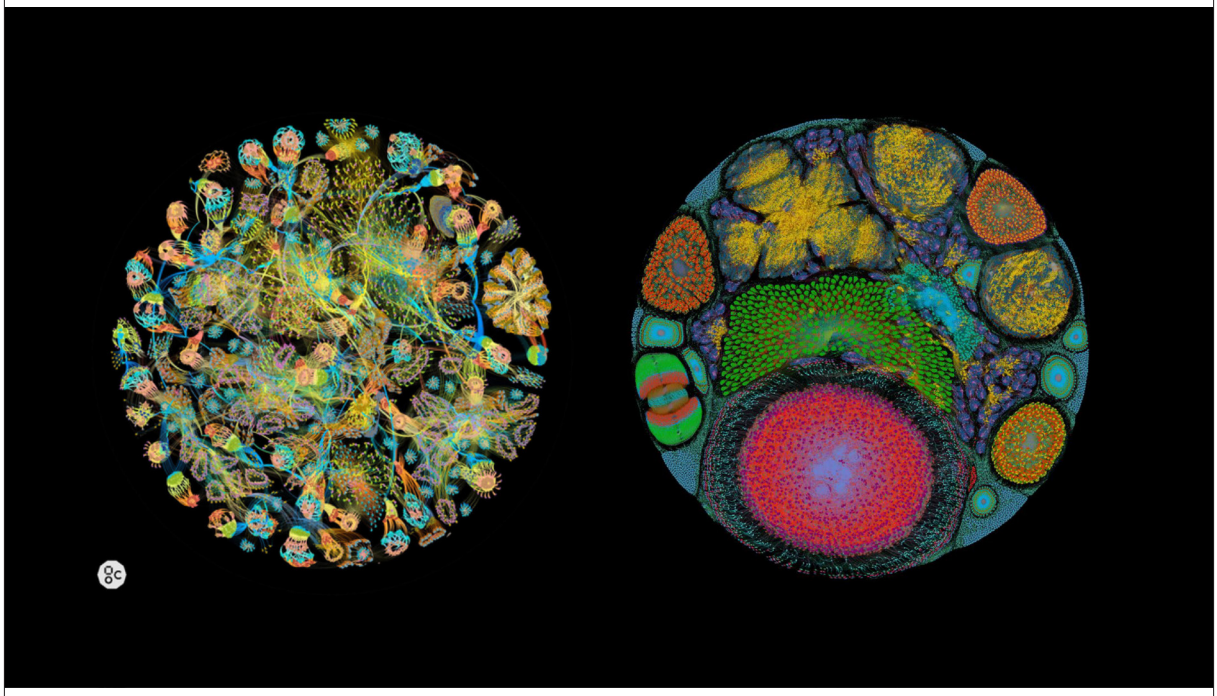
Neural Network In A Minute

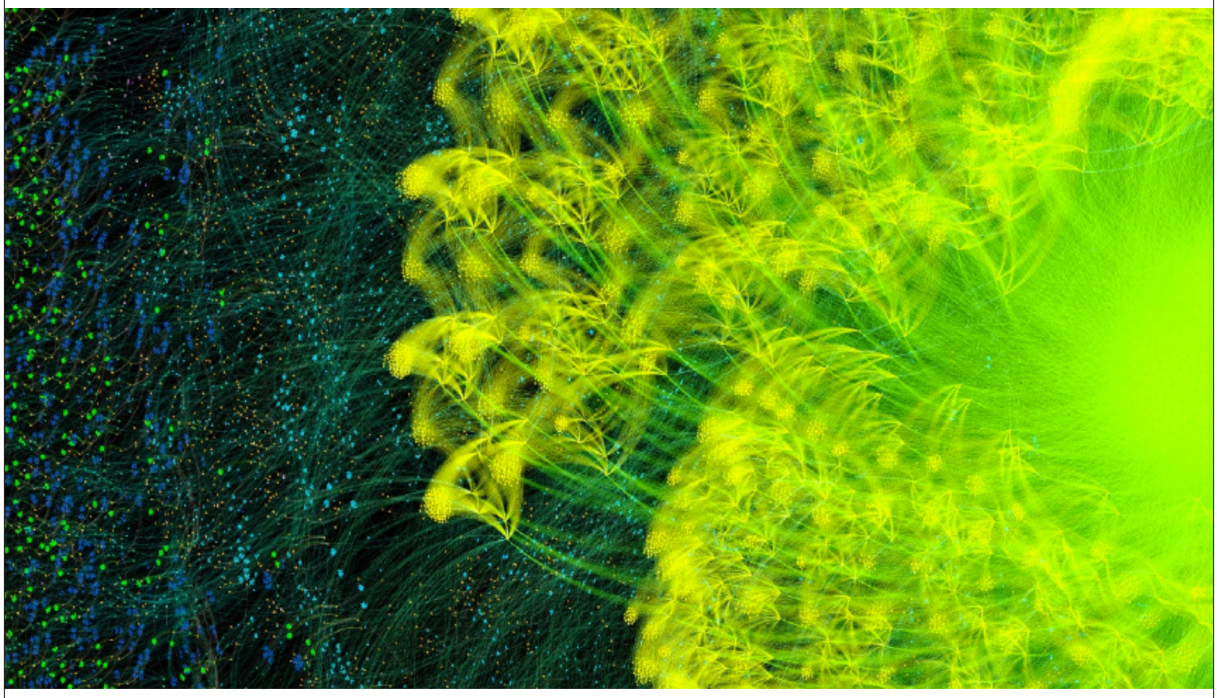


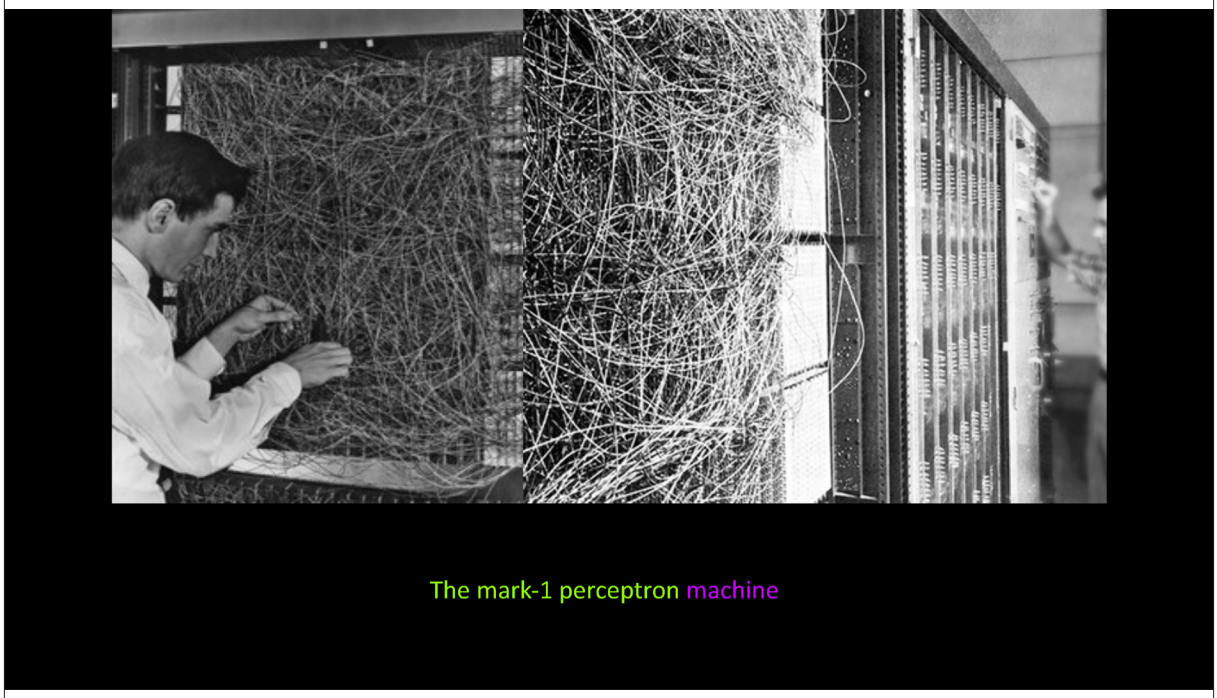


Training & Inference

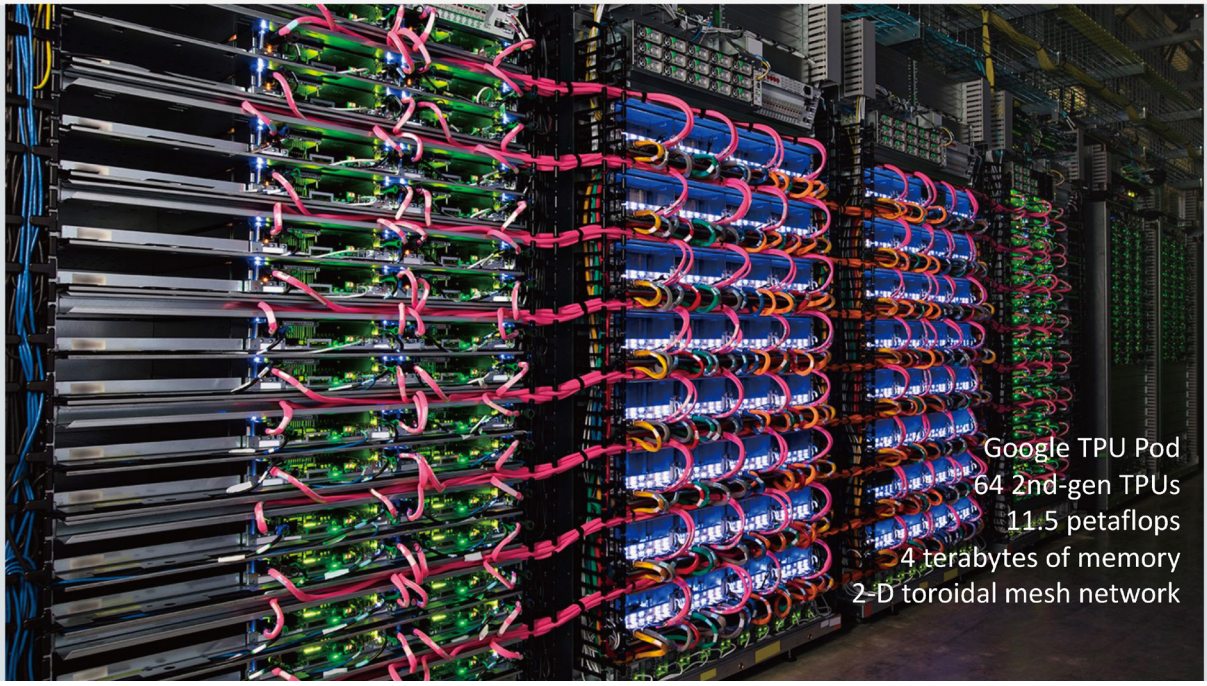




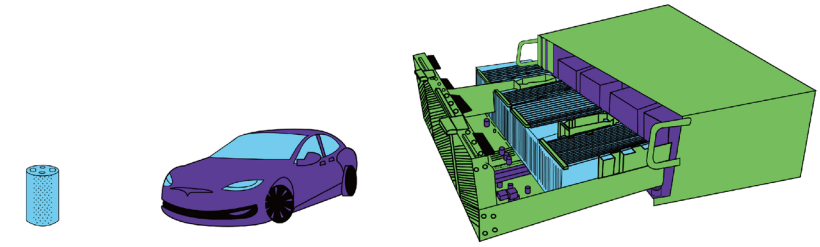




The mark-1 perceptron machine



AI Chip Scale of **Computation**



> 1 Tops > 10 Tops > 100 Tops

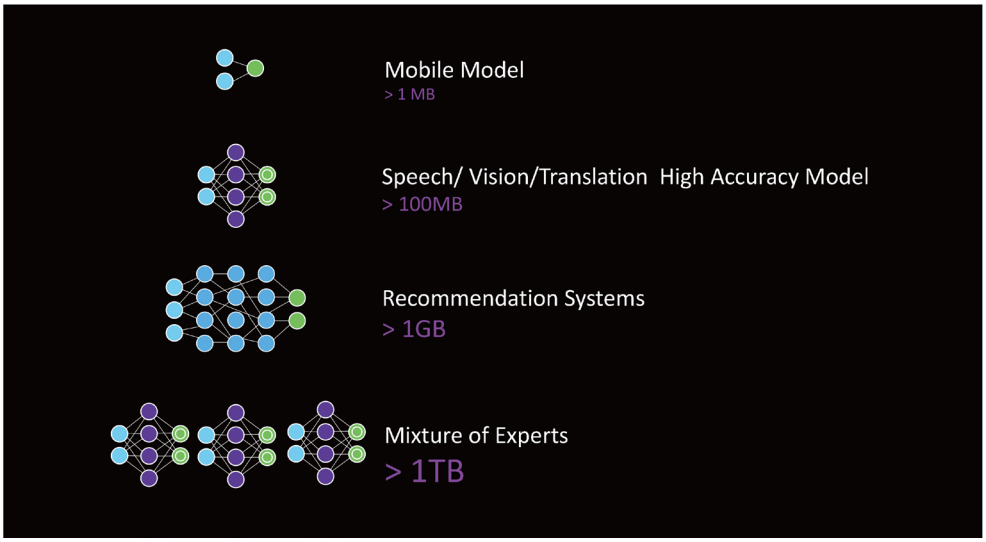
1 Tops = 1,000, 000, 000, 000 OP per Second

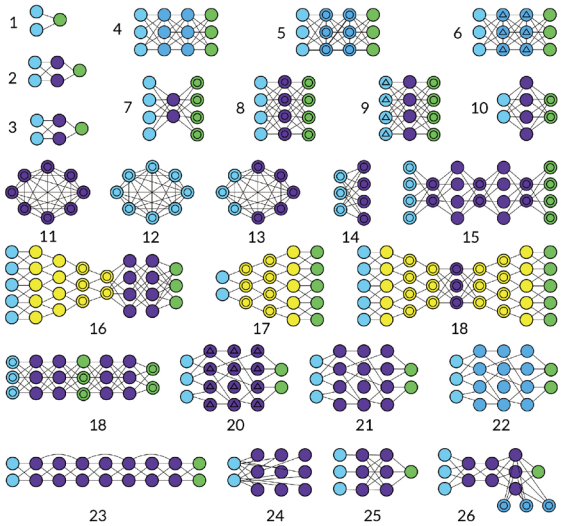
Scale of Storage : **Bandwidth**

R = 3, W = 112, N=64, p = 0.2	Fully Connected	3 x 3 Conv	Depthwise Seperarable Conv	Batch Norm Layer Norm
Compute	$W^4\ n^2$	$W^2\ r^2\ n^2$	$W^2\ r^2\ n + W^2\ n^2$	$5W^2\ n$
Data Access	$W^2\ n$	$W^2\ n$	$W^2\ n$	$W^2\ n$
Compute / Access	$W^2\ n$	$R^2\ n$	$R^2 + n$	5
BW per 125 TFLOP	**3 Gb/s**	**3 Tb/s**	**30 Tb/s**	**800 Tb/s**

(Source : Cerebras)

Scale of Storage: **Size**

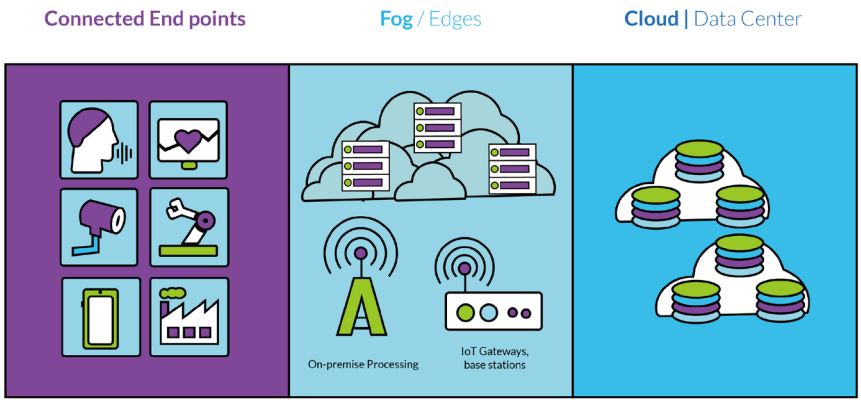




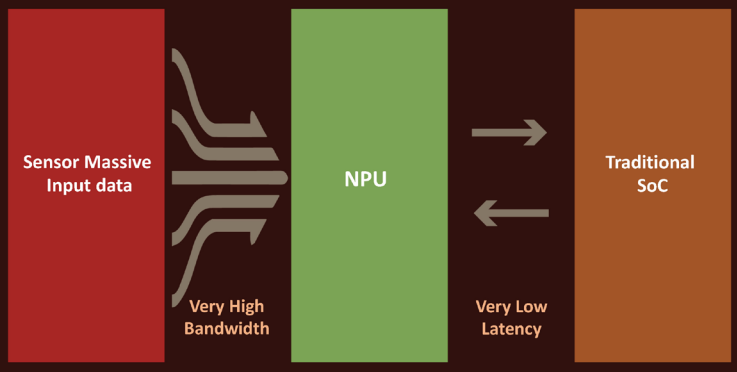
Scale of **Model Diversities**

1. Perceptron (P)
2. Feed forward (FF)
3. Radial Basis Network (RBF)
4. Recurrent Neural Network (RNN)
5. Long/ Short Term Memory (LSTM)
6. Gated Recurrent Unit (GRU)
7. Auto Encoder (AE)
8. Variational AE (VAE)
9. Denoising AE (DAE)
10. Sparse AE (SAE)
11. Markov Chain (MC)
12. Hopfield Network (HN)
13. Boltzmann Machine (BM)
14. Restricted BM (RBM)
15. Deep Brief Network (DBN)
16. Deep Convolutional Network (DCN)
17. Deconvolutional Network (DN)
18. Deep Convolutional Inverse Graphics Network (DCIGN)
19. Generative Adversarial Network (GAN)
20. Liquid State Machine (LSM)
21. Extreme Learning Machine (ELM)
22. Echo state Network (ESN)
23. Deep Residual Network (DRN)
24. Kohonen Network (KN)
25. Support Vector Machine (SVM)
26. Neural Turing Machine (NTM)

Scale of **Services**



Furiosa is building a NPU coprocessor for **AI computation 2.0**



AI computation 2.0

AI computation 2.0:
massive sensor-data driven real-time computing

AI computation 2.0, massive sensor-data driven real-time computing, brings the paradigm-shift in applications and computing. The revolutionary deep neural networks will process massive sensor data and extract key information. These inference results will be passed onto the traditional SoC to handle high-level policy logics. As more and more business logics move toward the rapidly-advancing deep neural networks with incredible capabilities, NPU will become universal, playing a central role in computing.

Furiosa's mission is to bring the most efficient and powerful NPU coprocessors into the world that can seamlessly integrate into existing infrastructures to make our life better and safe.

We target vision sensor-centric real-time future applications with mass markets

- **Autonomous Car**
- **Vision Server/8k TV**
- **Manufacturing Robot**
- **Intelligent Camera**
- **Medical Imaging Equipment**

The grand challenges of real-time inference computing:
Current solutions do not meet all these challenges

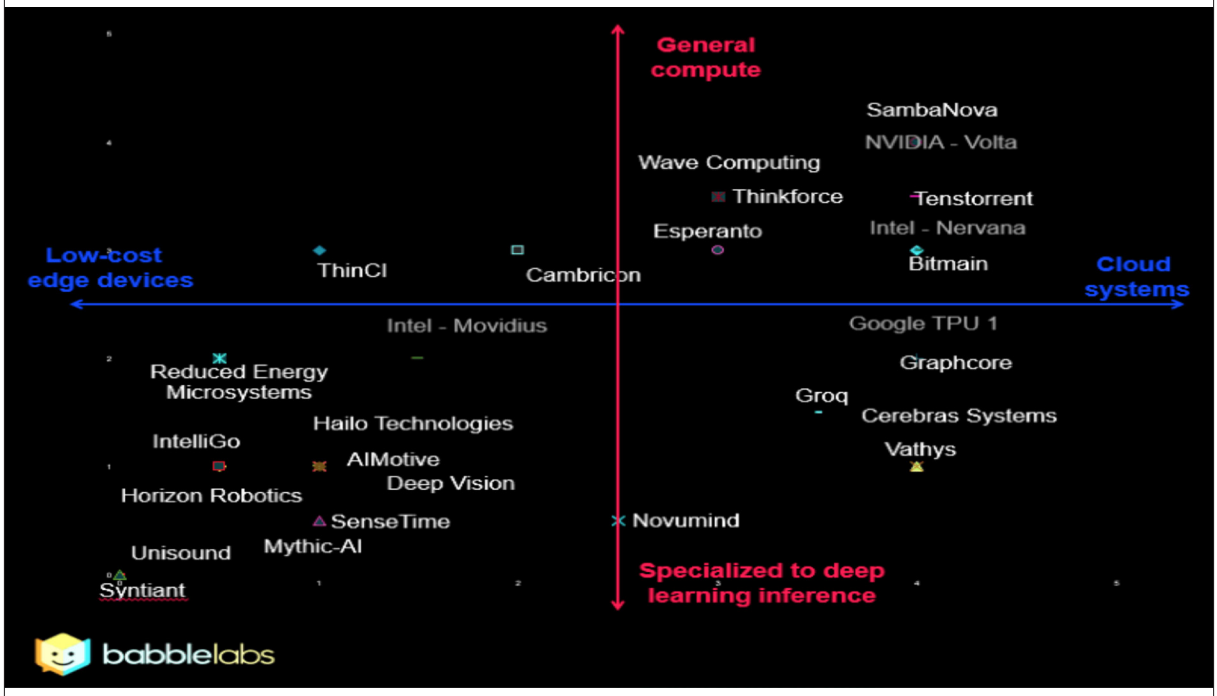
Challenges

1) Massive performance scale-up and extreme power efficiency

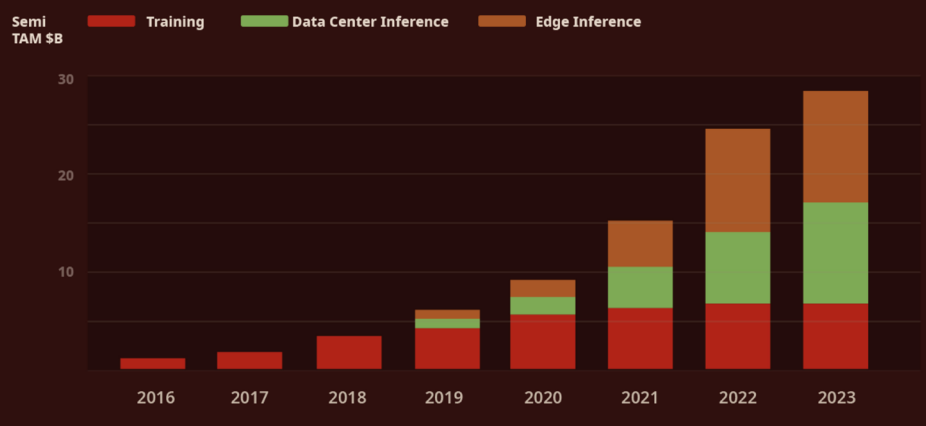
2) Rapid evolution of AI algorithms

3) Rapid evolution of AI software frameworks

Competition is heating up.
GPU is dominating the current coprocessor markets. But, large vendors and many startups are challenging the status quo.



AI inference chip market will explode as AI becomes more pervasive



Source: Barclays Research, Company Reports May 2018

How can we win in this space?

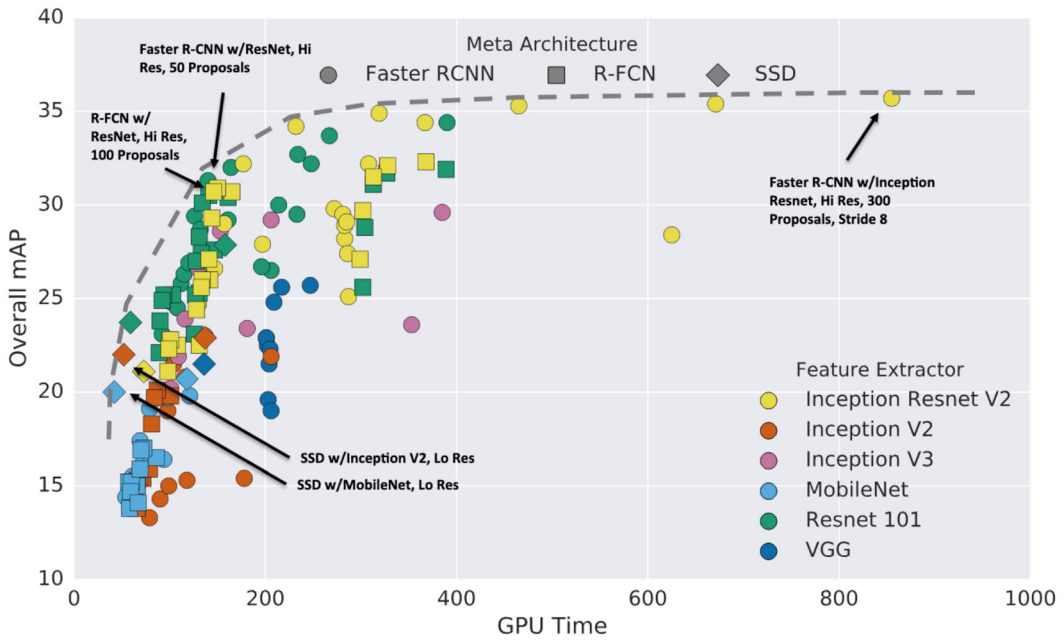
1) Existing accelerators have no competitive edge over our new architecture.

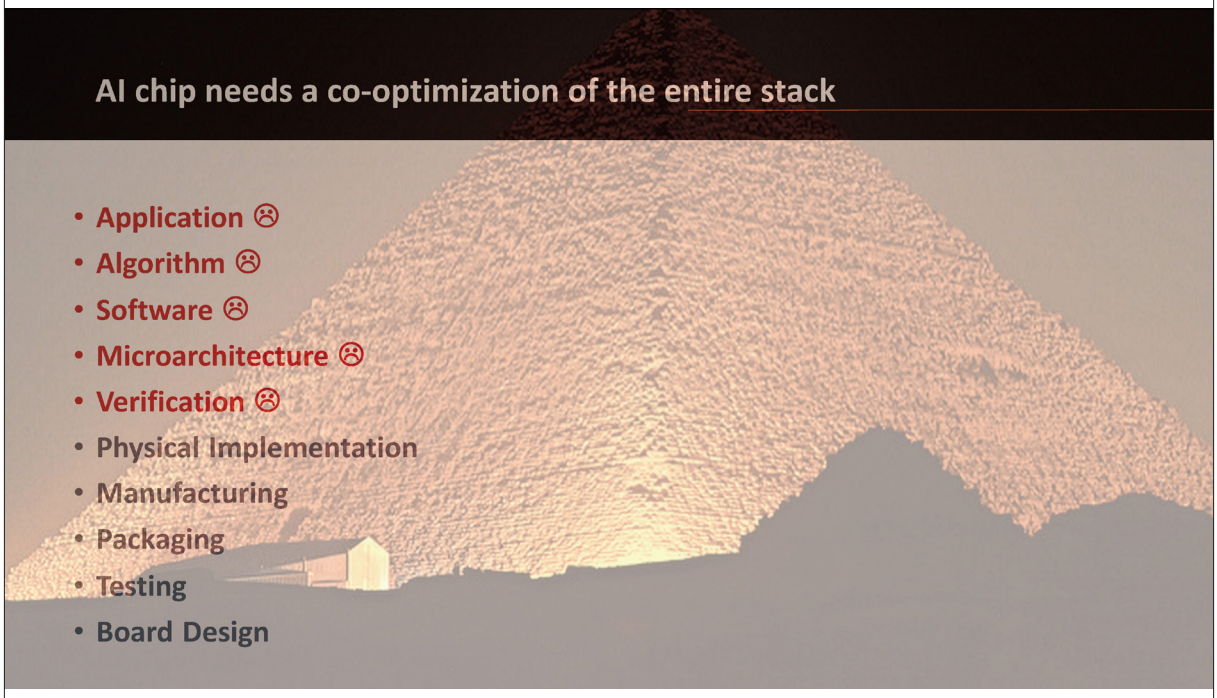
2) Customize for new emerging applications

3) Build our own ecosystem with strong execution and strong team.

It must start with algorithms.

	MobileNet V1	Inception V3
Computation	500 M	5000 M (10x)
Storage	5 MB	25 MB (5x)


Meta Architecture
Faster RCNN
R-FCN
SSD
Faster R-CNN w/ResNet, Hi Res, 50 Proposals
R-FCN w/ ResNet, Hi Res, 100 Proposals
Faster R-CNN w/Inception Resnet, Hi Res, 300 Proposals, Stride 8
SSD w/Inception V2, Lo Res
SSD w/MobileNet, Lo Res
Feature Extractor
Inception Resnet V2
Inception V2
Inception V3
MobileNet
Resnet 101
VGG
Overall mAP
GPU Time
40
35
30
25
20
15
10
0
200
400
600
800
1000


AI chip needs a co-optimization of the entire stack
• Application ☹
• Algorithm ☹
• Software ☹
• Microarchitecture ☹
• Verification ☹
• Physical Implementation
• Manufacturing
• Packaging
• Testing
• Board Design

Great AI chip comes with great architects.

필드의 근본 개념에 근거해서 설계해야 한다.

- Instruction Set Architecture
- VLIW, SIMD, VECTOR, Systolic Array
- SuperScale, Multithreading, DataFlow
- Pipelining
- Virtualization
- Prefetching, Caching
- IO, Memory subsystem
- Finite State Machine

We build a strong simulation and modeling infrastructure to develop the scalable AI chip.

- Everything should be modeled and simulated.
- Builds a scalable computation infrastructure and software that can push the limit
  - Cloud infrastructure
- Use the best tool available or build your own tools
  - Chisel
  - Rust
  - ...

New organization essential

Any orgnization that designs a system… will inevitably produce a design **whose structure is a copy of the organization's communication structure. – Conway**

Team = Algorithm + Hardware + Software

June Paik : CEO
(M.S, Georgia Tech, AMD/Samsung, GPU)

Han Kim : CTO
(Ph.D, KAIST, Samsung, Computer Architecture)

Rui Tam : COO
(Ph.D, NorthEastern, Sun/Apple, Server\Application Processor)

Michael Xu : GM of China, Chief SoC
(Ph.D, CUHK, Qualcomm/IBM, SoC)

Bon Gu : Chief Software Architect
(Ph.D, SNU, Samsung, NVMe SSD Runtime)

Jee H Kang : Chief Scientist
(Ph.D, SNU, Verified Compiler)

Hyung I Koo : Chief Algorithm
(Ph.D, SNU, Qualcomm, Faculty at Ajou University)

John Kim : Technical Advisor
(Ph.D, Stanford, Faculty at KAIST, Computer Architecture)

Chung Kil Hur : Technical Advisor
(Ph.D, Cambridge, Faculty at SNU, Programming language)

World class AP, GPU, SSD, DRAM product development experiences at Intel, Apple, Samsung, Qualcomm, etc

Multi-region R&D and sales with global talents

Strong collaboration with top R&D universities

글로벌 AI 칩 성능 수준 측정 MLPerf 벤치마크

MLPerf

2019년 7월
전 세계 26개 업체 참가 신청

Qualcomm Google Tencent 腾讯 NVIDIA Alibaba.com intel

2019년 10월
9개 업체 결과 제출

Startup H 기업 chip 335억 투자

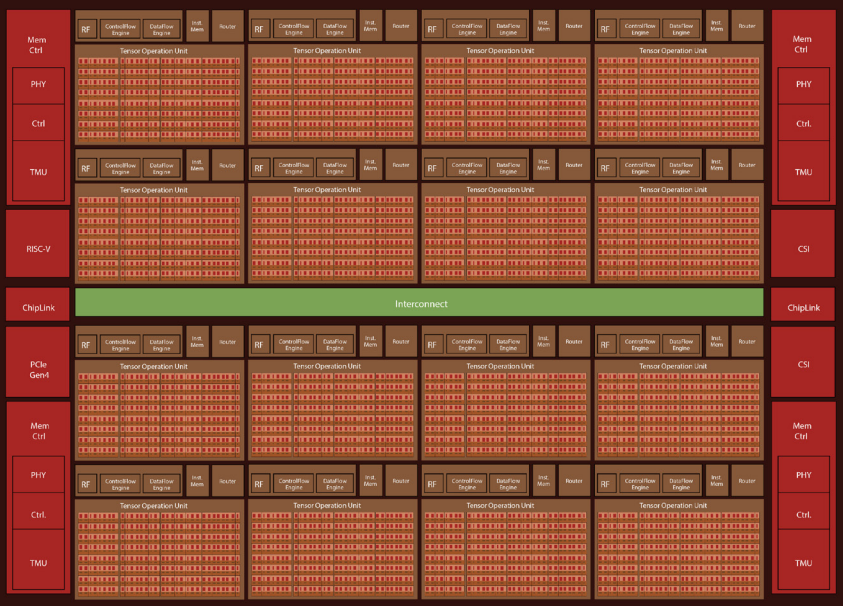
FURIOSA FPGA 13억 투자

H 기업 대비 2배 이상 성능 달성

Target applications

- 4K multi-camera 60 fps high-resolution object detection
- Facial tracking and 3D pose estimation model
- Sensor-fusing model
- Video prediction model
- Bayesian model

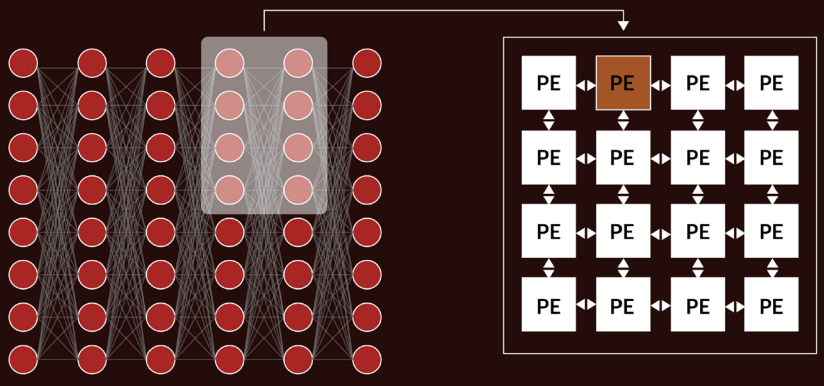
Autonomous cars will probably need all of these models



Furiosa Madrun Coprocessor

Technical Overview

We build our coprocessor on the firm foundations of distributed parallel computing. It is also domain-specific design to achieve the maximal performance-power efficiency and real-time latency for deep learning workloads. While maintaining the maximum efficiency, the software-driven flexibility of our architecture address challenges of fast-pacing changes of deep neural network models and deep learning software frameworks.



Deep Neural Network

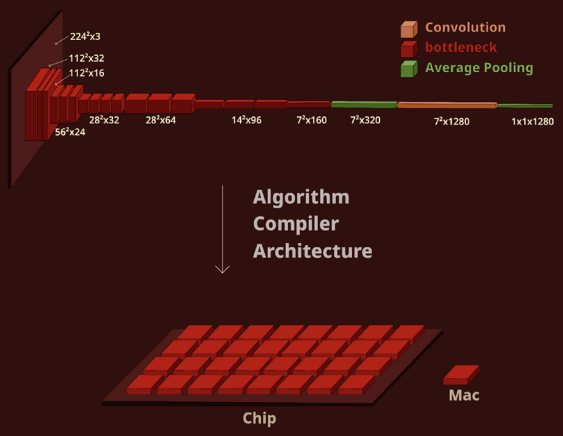
Distributed Parallel Computing

Performance Of Deep Learning Accelerators

Perf. = # MAC x Utilization

Utilization – mapping algorithms onto logics
- algorithms
- compiler + runtime software
- architecture

 # MAC – mapping logics onto physical substrate
- chip size
- transistor density
- metal layers (routing resources)
- power envelope

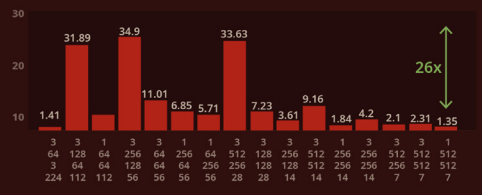


Utilization is what Matters

Algorithm (ex: MobileNet v2)

DNN algorithms have various types of operations and tensor shapes

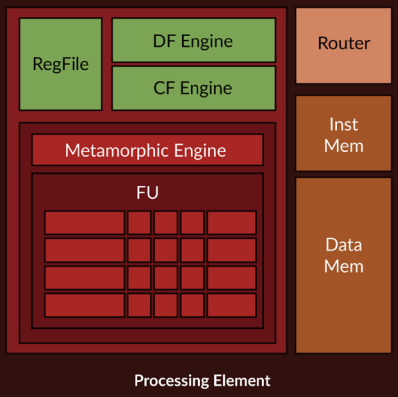
GPU Performance * (Titan X) - convolution



Filter width Out Channel In Channel Width, Height

Peak performance does not translate to real performance

Dataflow / Von Neumann
Hybrid Architecture for 2020s intelligence computing



Macro-level Data Flow

- programming model suitable for model-parallelism.
- Efficient data shuffling and task launching

Von Neumann Micro-execution

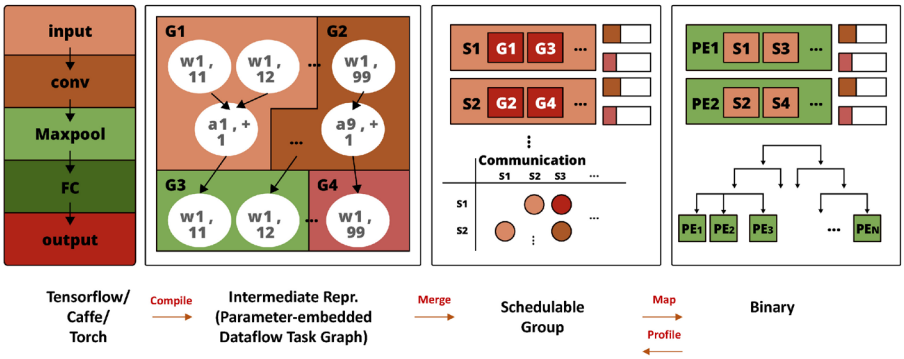
- MegaSIMD ISA specialized in massively parallel deep neural network computation: regular conv, depthwise-separable conv, matmul, nonlinear functions, et al.

Metamorphic Tensor Operation Unit

- Universal handling of tensor shapes and operations with strong compiler assistance
- Topologically reconfigurable and software-defined

SRAM-Centric Architecture

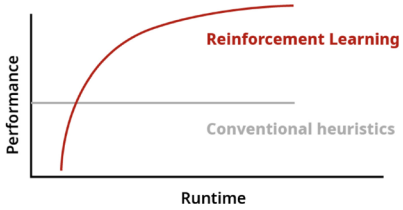
- Large distributed SRAMs balancing computation and memory
- Memory aware static scheduling: aggressive prefetching if model doesn't fit on SRAM
- 4/8/16 bit quantization support compatible with TensorFlow



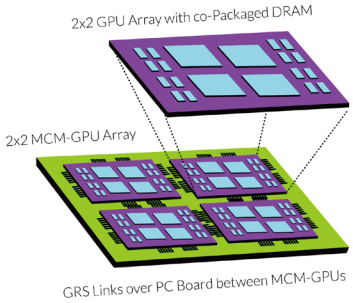
Software matters most

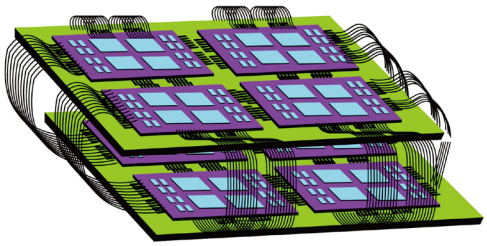
Furiosa software stack provides

- **Complete support for TensorFlow and PyTorch**
- **Multi-PE model parallel compiler optimization to best utilize hardware resources**
- **Robust and secure high performance runtime**

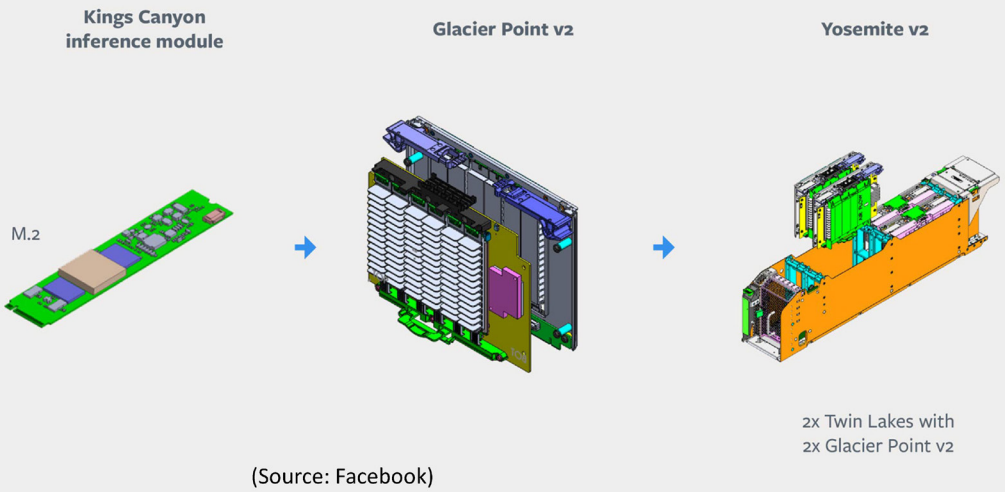


Co-Design of Channel and Signaling System





(Source: Nvidia)



(Source: Facebook)

Roadmap

2019 Q2

: 4K High Resolution object detection FPGA

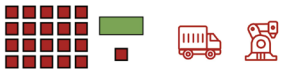
2020 Q1 (Series A Milestone)

: 14nm ASIC V0 MPW chip

PCIe Type

2021 Q3

: 14nm ASIC V1 coprocessor targeting cloud datacenter and edge server



Multi-chip systems

2022 Q4

: 5nm ASIC V2 coprocessor + system targeting mass market applications

Thank you!