Embedding Programmable DNNs in Low Power SOCs

XPERI

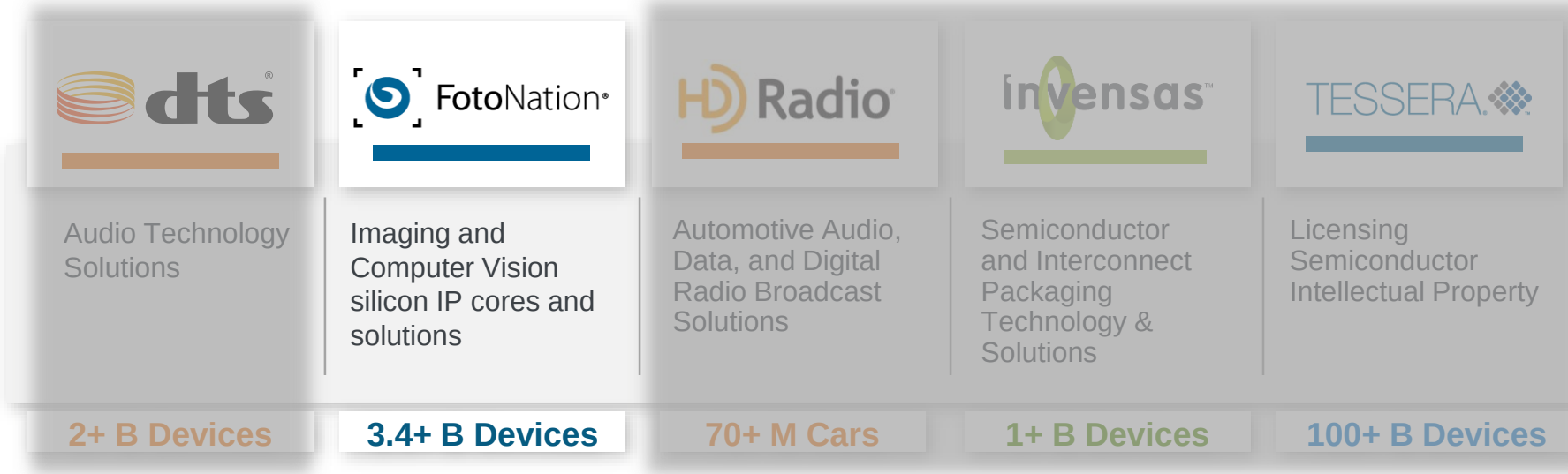
Steve Teig

22nd May 2018

Company Overview

FotoNation – trusted brand of XPERI

Portfolio of Trusted Brands



What, How and Why

What – AI Enabled Imaging at the Edge

Always-on inference: operates even while the device is “off”

(e.g., ultra low power FD/FR as an enabler)

Driver State Sensing for autonomous driving

(e.g., always ON driver assistant)

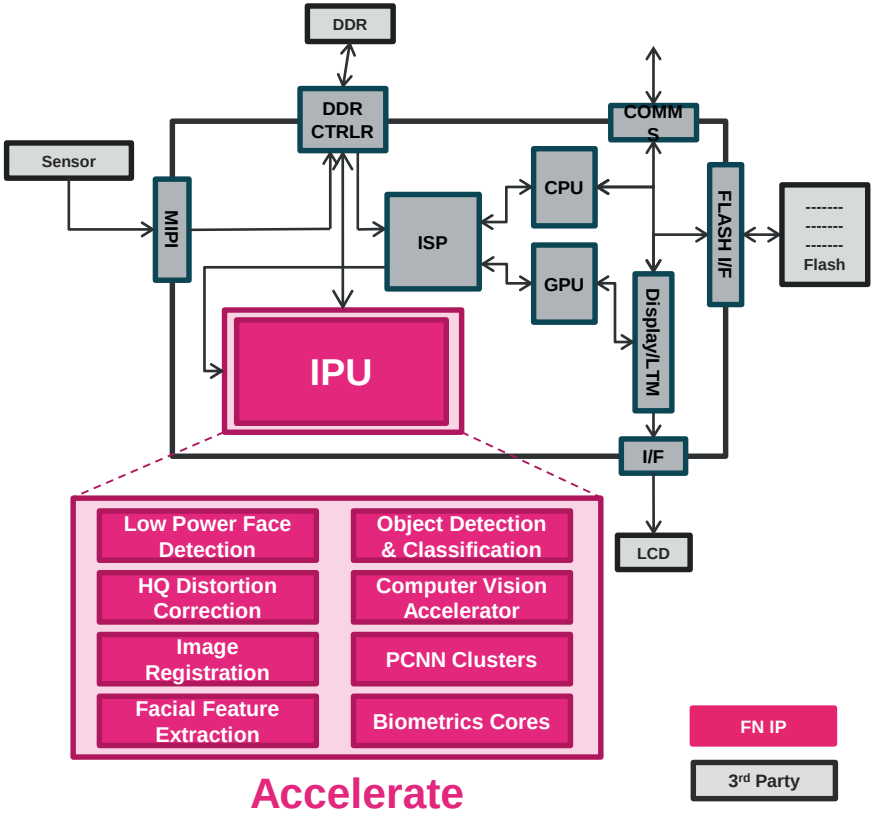
Smart IoT: TVs to drones to microwave ovens to …

(e.g., ultra low power people detection)

Head-mounted displays for AR or MR

(e.g., ultra low power IRIS, eye gaze, scene understanding,, etc..)

How – Image Processing Unit hybrid deployment



Understand

- Face, People, Object Detection, Segmentation & Tracking
- Scene Classification and recognition

Enhance

- De-warping & Stitching
- Stabilization with Rolling Shutter Correction
- HDR & LTM

Personalize

- Visible and NIR 3D FR
- IRIS recognition
- Liveliness Detection & Continuous recognition (hand jitter, facial, etc.)

Why - Advantages

- Ultra high performance for always ON features via dedicated accelerators (e.g., facial detection and recognition, resampling, LTM, etc.)
- High flexibility and configurability after silicon via programmable inference engines
- Flexibility and configurability for various markets via the concept of clusters
- Unmatched performance and power consumption for concurrent understanding, enhancing & personalization imaging features

Examples

UNDERSTAND

ENHANCE

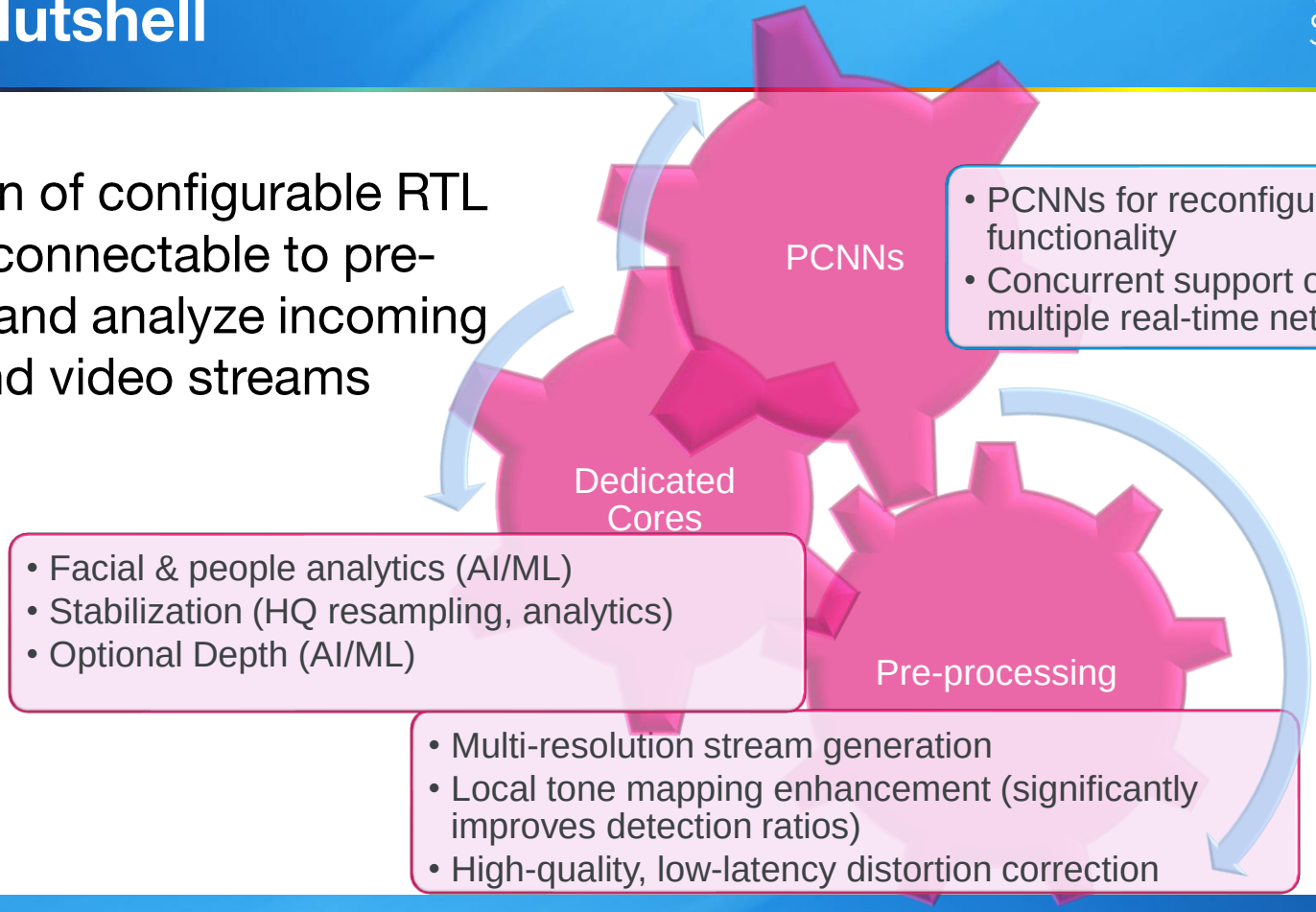
PERSONALIZE

IMAGE PROCESSING UNIT (IPU)

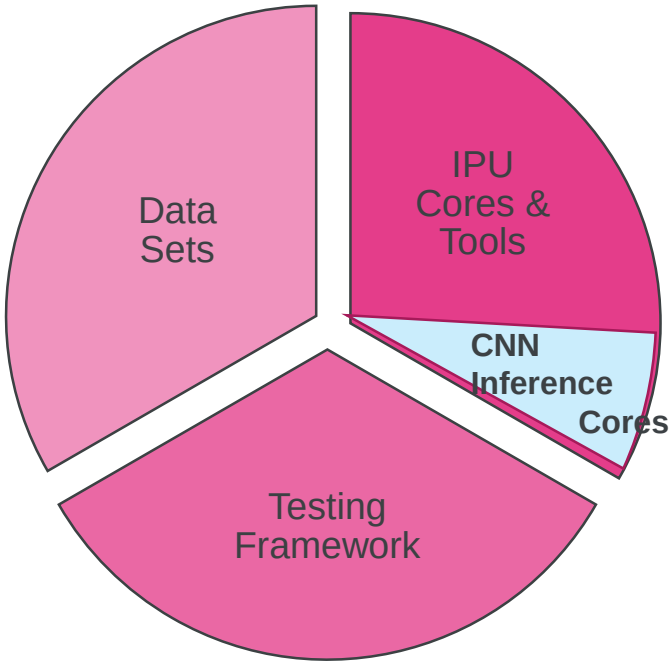
What is the IPU?

IPU in a Nutshell

Collection of configurable RTL engines connectable to pre-process and analyze incoming image and video streams



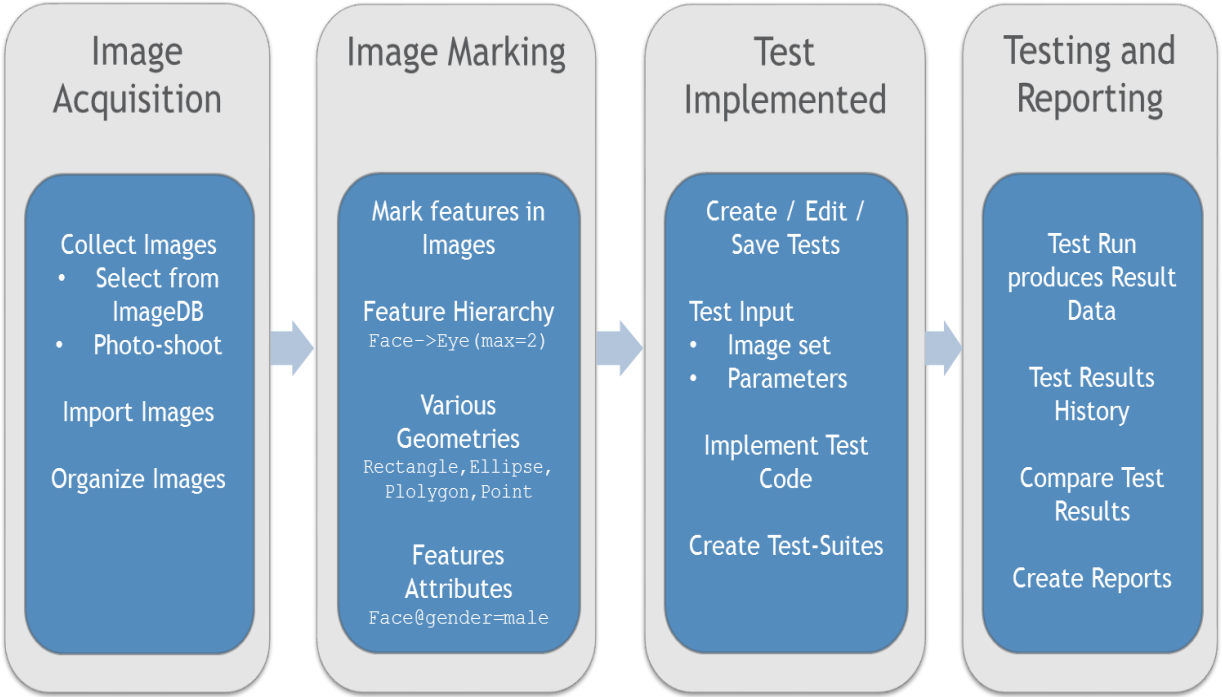
IPU – the supporting ecosystem

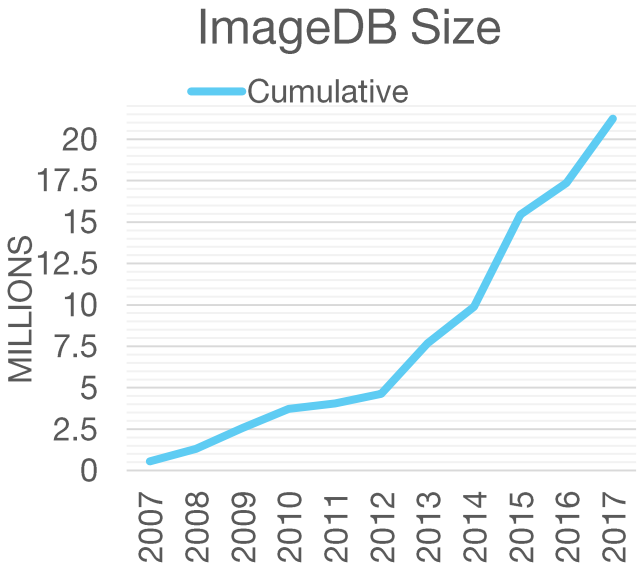


… inference cores are a small part of the total

- IPU
  - RTL (image pre-processing, dedicated analytics and inference cores)
  - Tools for programming, training and debugging
- Testing Framework (a.k.a. ImageDB) – test framework supporting acquisition, marking, testing and reporting
- Data Sets (a.k.a. CV Infra) – computer vision infrastructure supporting 2D and 3D image sets acquisition, annotation, marking and training sets generation with ground truth

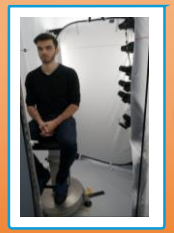
Testing Framework (a.k.a. Image DB)

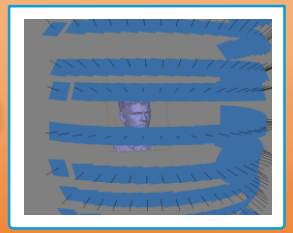




Computer Vision Infrastructure

Acquisition – 3D Images





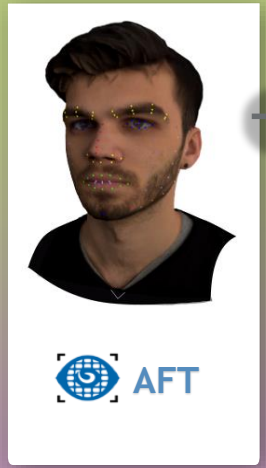
Mesh

Texture *(vis)*

Texture *(nfr)*

Setup (visible «& NIR») | **Photogrammetry** | **3D Model**

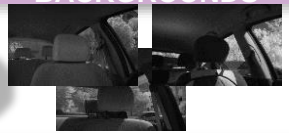
Auto-marking 3D Model



AFT

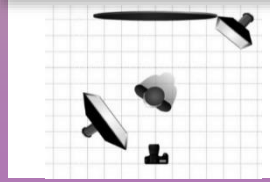
3D Model

CUSTOM BACKGROUNDS



3D ACCESSORIES

CUSTOM LIGHTING



Rendering (lens/sensor/ISP model)

Generating Marked 2D Renders

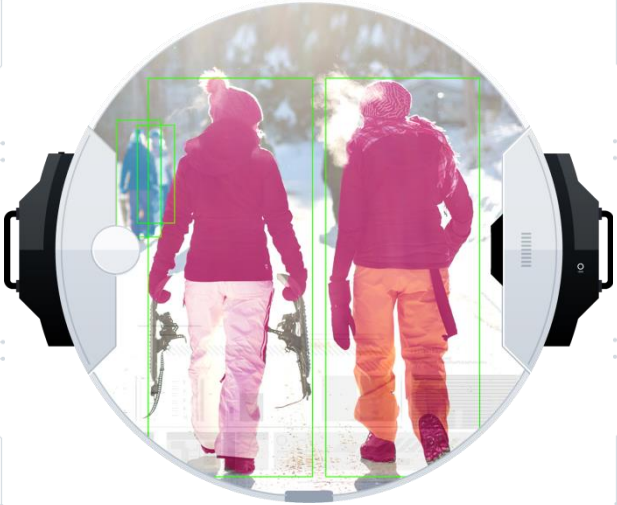
Marked 2D Samples

PCNNs and PCNN-Clusters

FotoNation's Programmable CNN Engine (PCNN)

Built in pre-processing on the fly imaging engine for (e.g., layer 0)

Support for compression, quantization, decryption



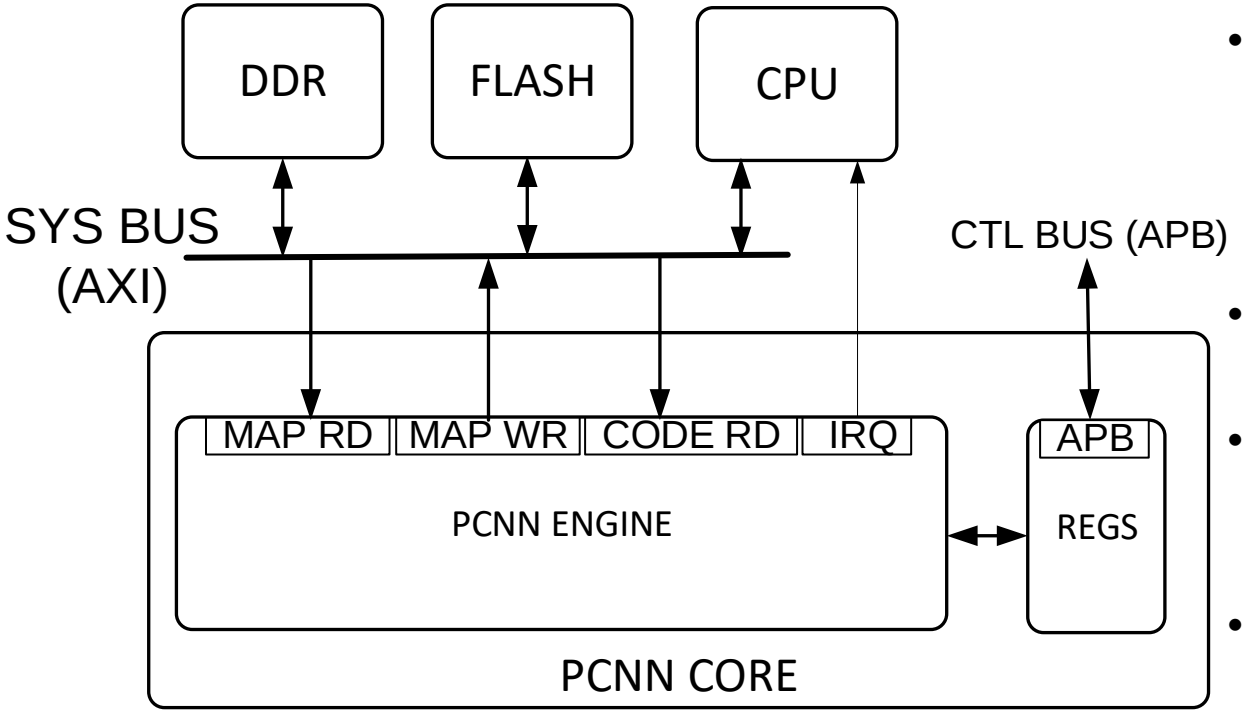
Designed for very low latency and real-time network inference

Local memory for fast data access and reduced memory BW

PCNN 1.2 - 36 MAC/cycle
or
PCNN 2.0 - 512 MAC/cycle

Supporting toolchain to implement customer defined network architectures

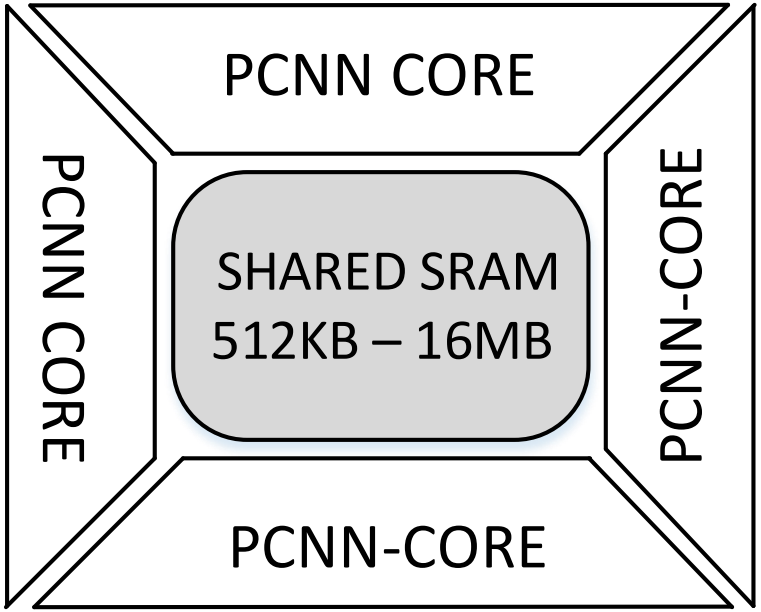
FotoNation's PCNN Highlights



- Low power consumption (22n FDSOI tech)
  - 18 mW for PCNN 1.2
  - 120 mW for PCNN 2.0
- Built in real-time, on-the-fly NN decryption engine
- Separate memory channels for network fetch and intermediate layers/input
- Separate cache for code/net & intermediate layers

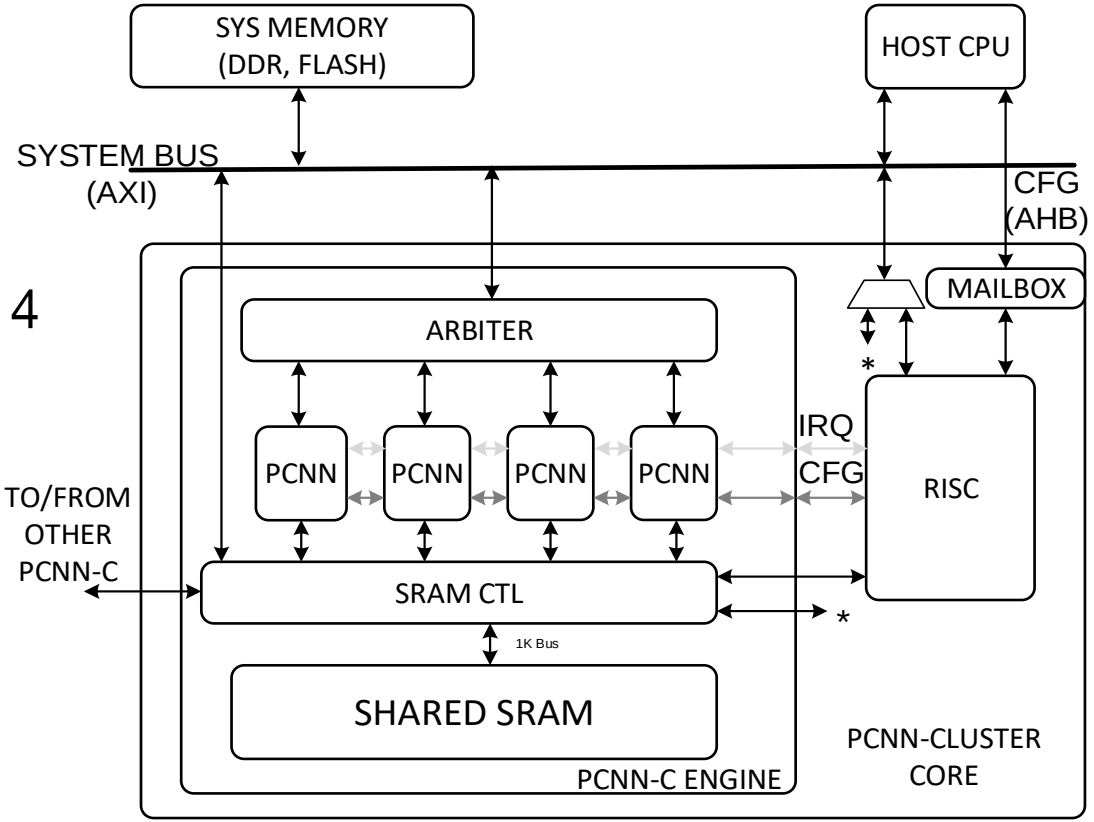
PCNN-Cluster Concept

- PCNN–C accelerates any type of CNN
- Highly configurable and scalable architecture
- Multiple PCNN IP cores connected to a shared common memory (e.g., SRAM)
- Configurable with up to 4 PCNN cores

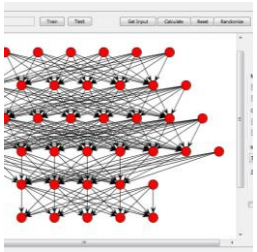


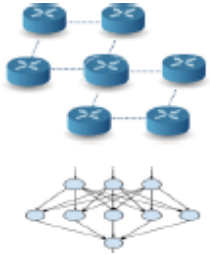
PCNN-Cluster Highlights

- Clusters can have a mix of PCNN 2.0 or 1.2 cores
- Can execute single or up to 4 individual networks at the same time
- Several clusters can be chained for increased computing power and capabilities



Security and Protection – the Danger





Networks sit in the device's permanent storage

Device can be accessed and storage contents read

Neural processor makers offer network transfer tools

Networks representation patterns can be identified and localized in the storage contents

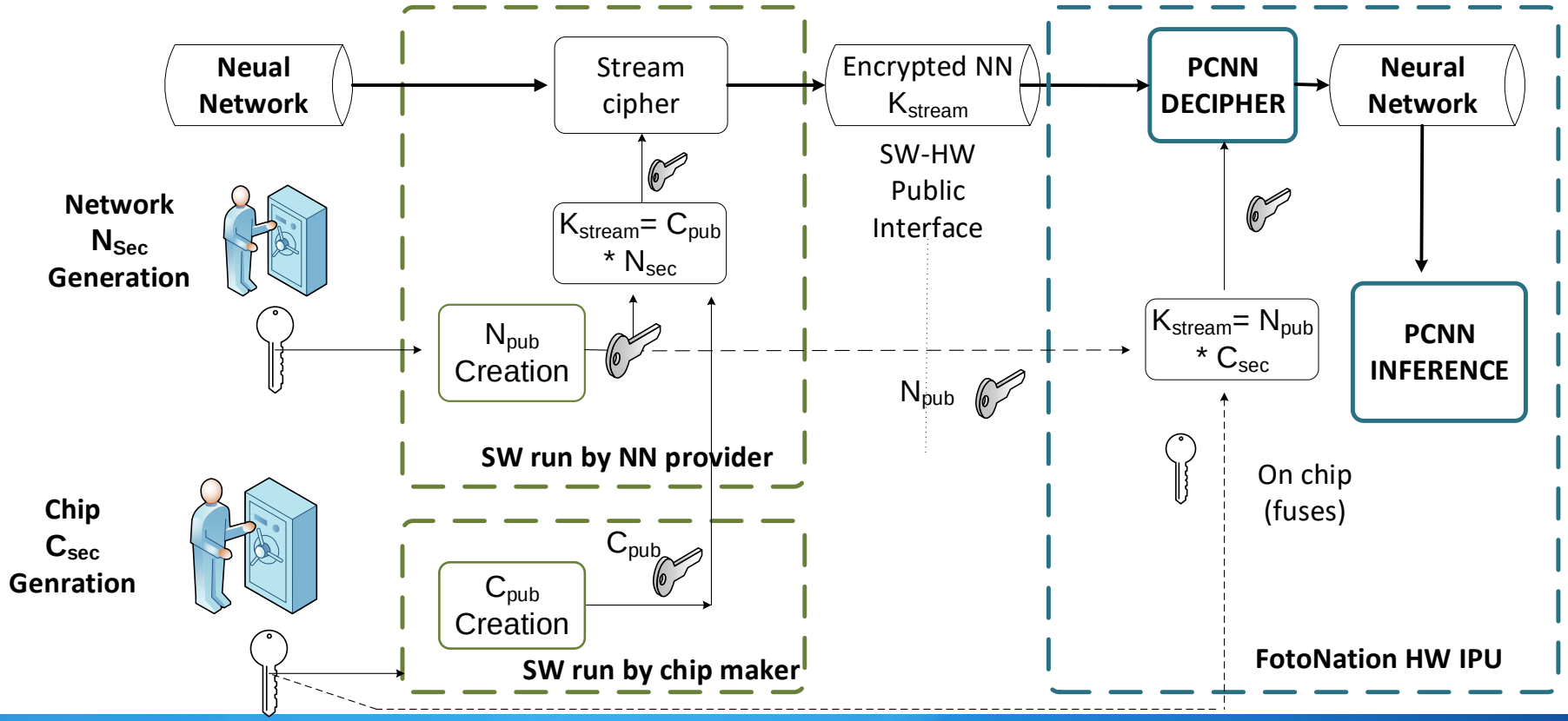
Network architecture & weights extraction

Once the network representation is known, architecture and weights values can be obtained

Network re-map and inference on alternative architectures

Networks can be remapped and inferred on alternative architectures as own

Solution

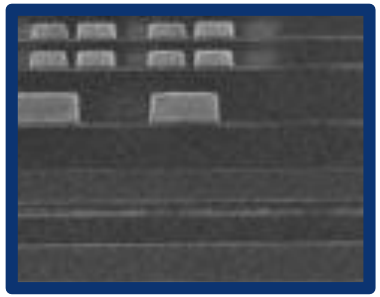


DBI® 3D AI Enabled Chip Architectures using IPU

Digital Bond Interconnect platforms for 3D semiconductors – DBI®

ZiBond®

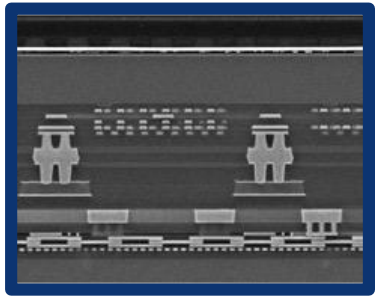
Homogeneous Bonding



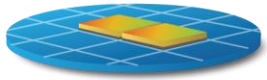
Scalable to 1,000,000 interconnects per mm^2

DBI®

Hybrid Bonding



Wafer to Wafer (W2W)

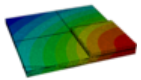


Die to Wafer (D2W)

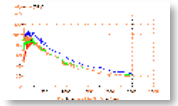
Die to Die (D2D)

Technology Development & Optimization

3D Design & Architecture

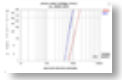


Simulation



Materials Characterization

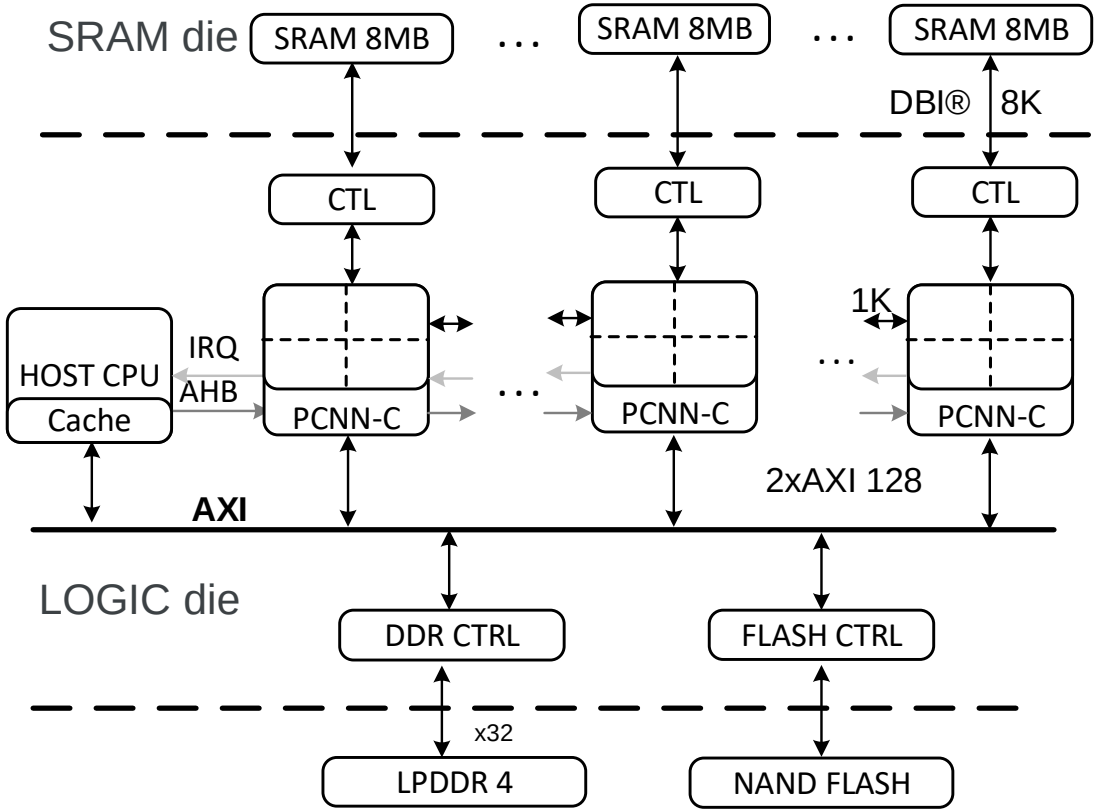
Wafer/Die Bonding & Processing



Reliability

Failure Analysis

AI Enabled Chip Architecture with IPU & DBI®



- Array of 2µm pitch DBI® interconnects (between silicon dies) - up to 250,000 vertical interconnects per mm^2 enables groundbreaking computing architecture
- Very short vertical interconnects offer ultra high performance at very low power

Conclusions

Conclusions

- Edge computing requires low power and real-time response
- Huge investment beyond inference cores in development, data preparation, testing and validation of the algorithms
- Edge computing brings IP security complications with regard to reverse engineering
- Edge computing is recommended to have real time, on-the fly network protection with no meaningful compromise to latency, area, and power consumption; FotoNation Cores do
- Edge intelligence with IPU is the way forward: for power efficiency, latency and privacy/security reasons

XPERI enables ultimate user experience making possible
AI enabled imaging in the device – a.k.a. the Edge …

Resources

- DBI®: https://www.invensas.com/solutions/waferdie-bonding/
- IPU: https://www.fotonation.com/products/image-processing-unit/
- Deep Learning course: https://www.coursera.org/specializations/deep-learning#courses
- Deep Learning book: http://www.deeplearningbook.org/

Thank you!

Additional Slides

PCNN-C performance

PCNN-C Configuration scenarios

	PCNN-C Lite	PCNN-C Base	PCNN-C Performance	PCNN-C Future
Number of PCNN 2.0 Cores	0	1	2	4
Number of PCNN 1.2 Cores	2	3	2	0

PCNN-C Performance examples on various networks

	PCNN-C Lite	PCNN-C Base
Yolov2 [FPS]	30	250
Body Pose [FPS]	-	15
CCOT [FPS]	450	4700
SEGNET [FPS]	10	80
AlexNet [FPS]	60	400
LeNet [FPS]	20000	100000
GoogleNetVx [FPS]	30	270
VGG16 [FPS]	3	12
FotoNation BD [FPS]	44	300
FotoNation FR (largest network on visible) [FPS]	44	300

DBI® explained

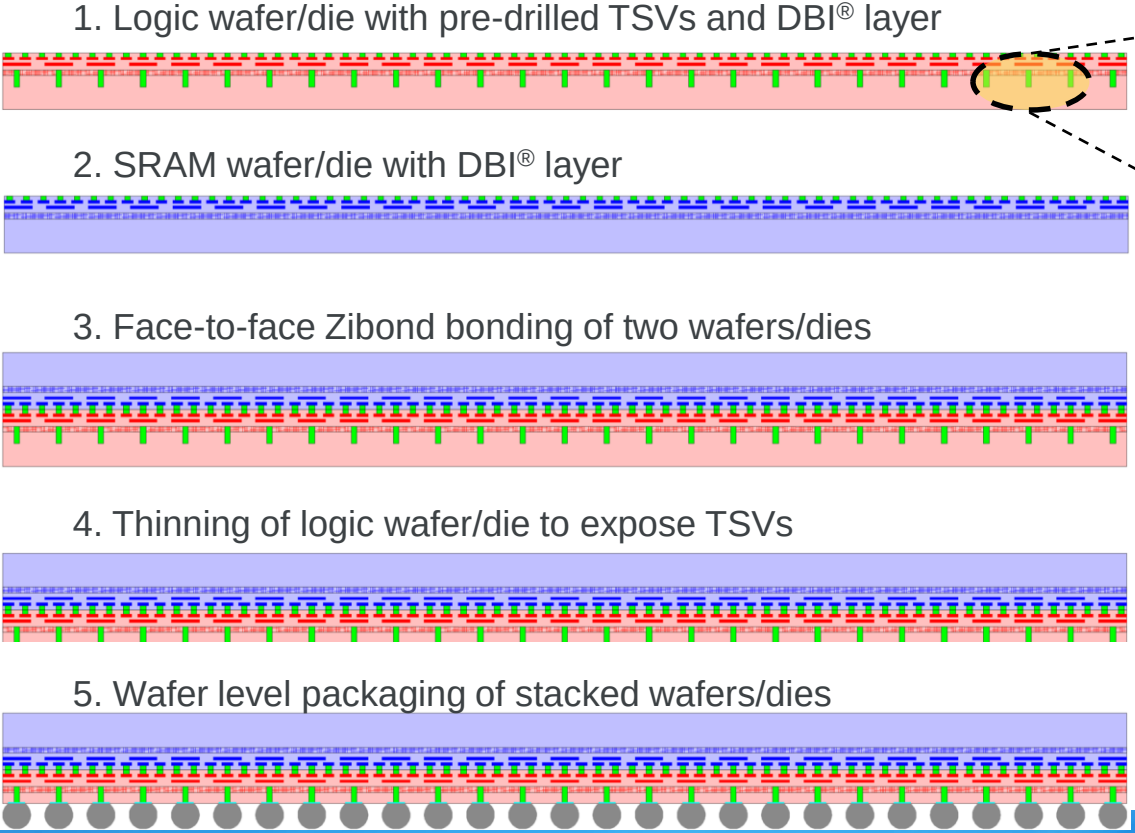
1. Logic wafer/die with pre-drilled TSVs and DBI® layer

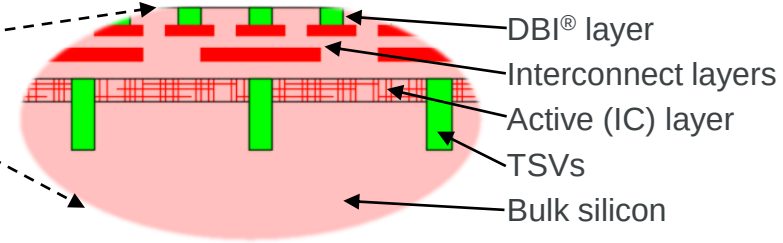
2. SRAM wafer/die with DBI® layer

3. Face-to-face Zibond bonding of two wafers/dies

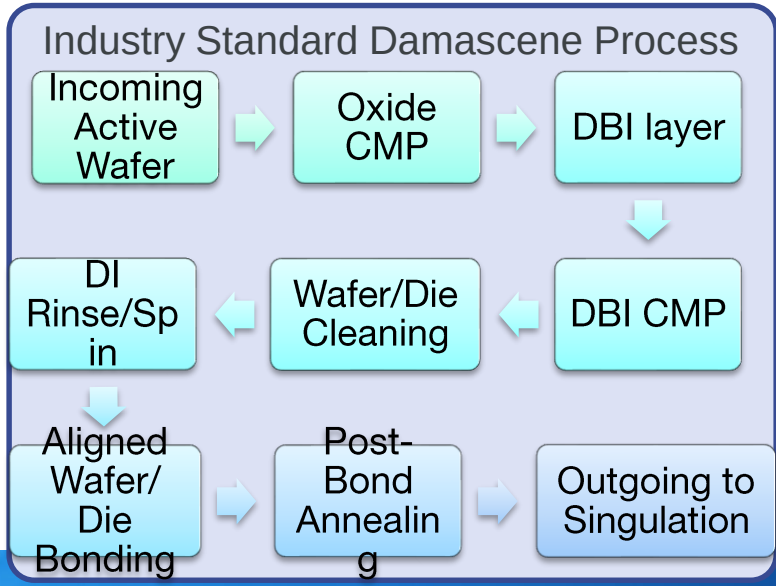
4. Thinning of logic wafer/die to expose TSVs

5. Wafer level packaging of stacked wafers/dies



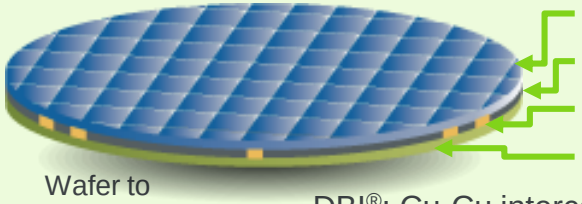


Industry Standard Damascene Process



DBI® with Zibond: the most advanced 3D interconnects

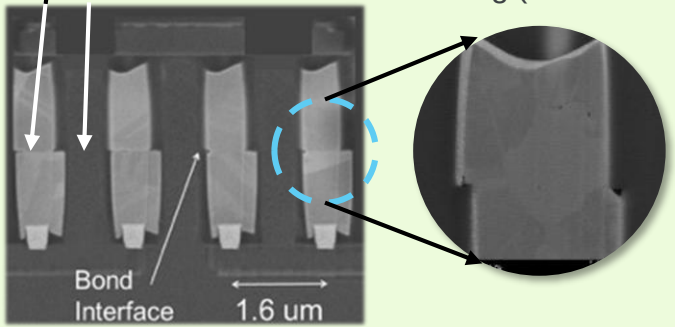
3D Wafer Bonding with Copper Interconnects



Wafer 1
Full surface oxide bond
µm scale Cu interconnects
Wafer 2

Wafer to wafer bonding

DBI®: Cu-Cu interconnect joining (no solder)
ZiBond: Full surface bonding (no underfill)



Chip to wafer bonding

Chip to chip bonding

Low Temperature | Fine Pitch | Thin Profile | Fast Throughput | Low Cost

Sony's latest image sensor uses DBI®