神经网络加速赋能端侧智能

17th June 2019

Imagination Technologies

10 billion units shipped

1.2 billion in past year

- **Developing innovative IP**
  - Industry leading graphics, vision & AI and communications
- **Delivering exceptional service**
  - Enabling fast time to market to customers around the world
- **Driving major new markets**
  - Helping our partners create successful solutions across many segments

The Imagination IP family

Imagination

The best solution for embedded graphics, vision, AI and communications

PowerVR GPU

Leading graphics IP cores for embedded devices

- XEP/XMP GPU
  - Focused Features
  - Fillrate & perf/mm2 focus
- XTP GPU
  - Feature rich
  - Performance/mW

PowerVR Vision & AI

Dedicated AI and Computer Vision IP Products

- PowerVR 3NX
  - Neural Network Accelerator
  - Performance/mm2
  - Performance/mW
- Video codec
  - ISP

Ensigma

Connectivity and broadcast communications

High performance, low power

- RF
  - Wi-Fi, Bluetooth
- Connectivity
  - Wi-Fi, Bluetooth
  - IEEE 802.15.4
  - GNSS
- Broadcast
  - TV, Digital Radio

Enabling the Edge for AIoT Enables New Worlds

Our Vision

Embedded intelligence at low power enables the EDGE

IoT

Retail PoS automation
Smart Cities
Asset tracking
Security
Smart Buildings
Foot-fall analytics

Mobile

Facial identification
Application monitoring
Advance camera filters
Picture annotations
Speech recognition
Smart assistance

Consumer / Media / AR VR

Image enhancement
Viewer analytics
Automatic subtitling
Environment awareness
Gesture control
Voice interface
Content filtering
User interaction

Industrial & Robotics

Defect detection
Object detection
Context awareness

Automotive

Auto emergency breaking
Lane departure warning
Collision warning
Driver monitoring
Navigation and safety
Traffic-sign recognition

Smart Surveillance

People tracking
Queue monitoring
Engagement recognition
Abnormal behaviour alerts
Event detection
Vehicle detection

1. Smart Automotive

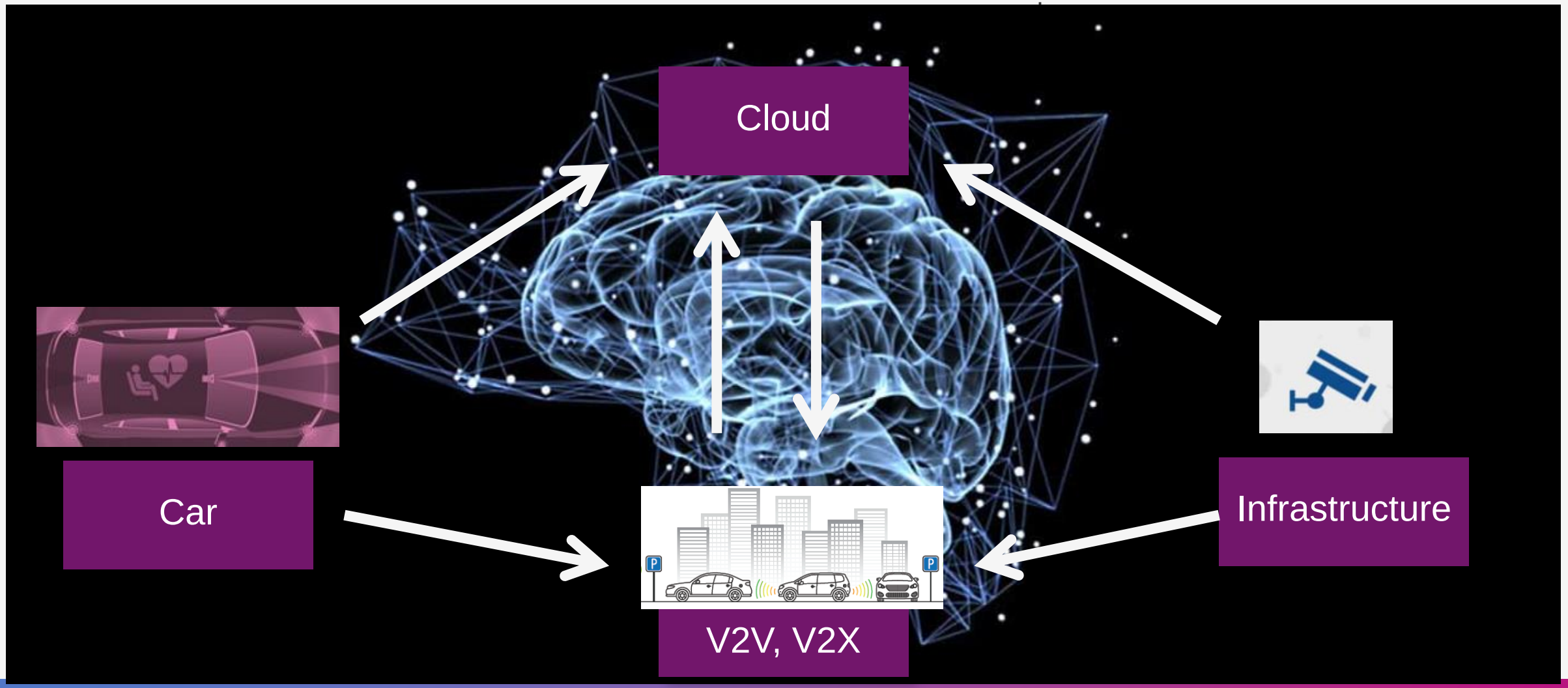
Enabling high performance Neural Networks in Automotive

Smart Automotive – ADAS and Autonomous

ADAS – Advanced Driver-Assistance Systems

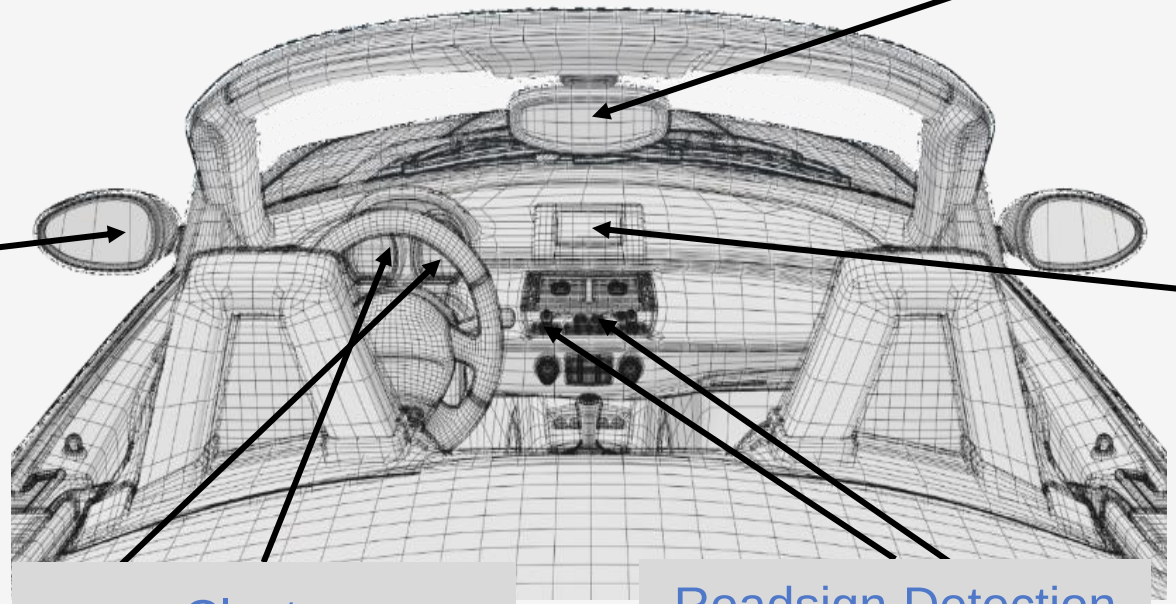
Device + Cloud + Urban infrastructure

Hybrid model



PowerVR applications in the e-cockpit

Combining graphics, compute and neural networks



Distracted Driver

NNA: Detection

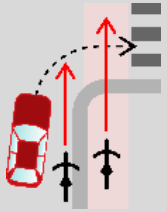
Mirror Replacement

GPU: Render
NNA: Object detect

Surround View

GPU: De-warp, Stitch
NNA: Object detection

Autonomous Drive



GPU: Pre-processing
NNA: Path planning

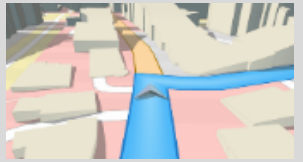
Cluster

GPU: Render

Roadsign Detection

GPU: Render
NNA: Identification

IVI/Navigation

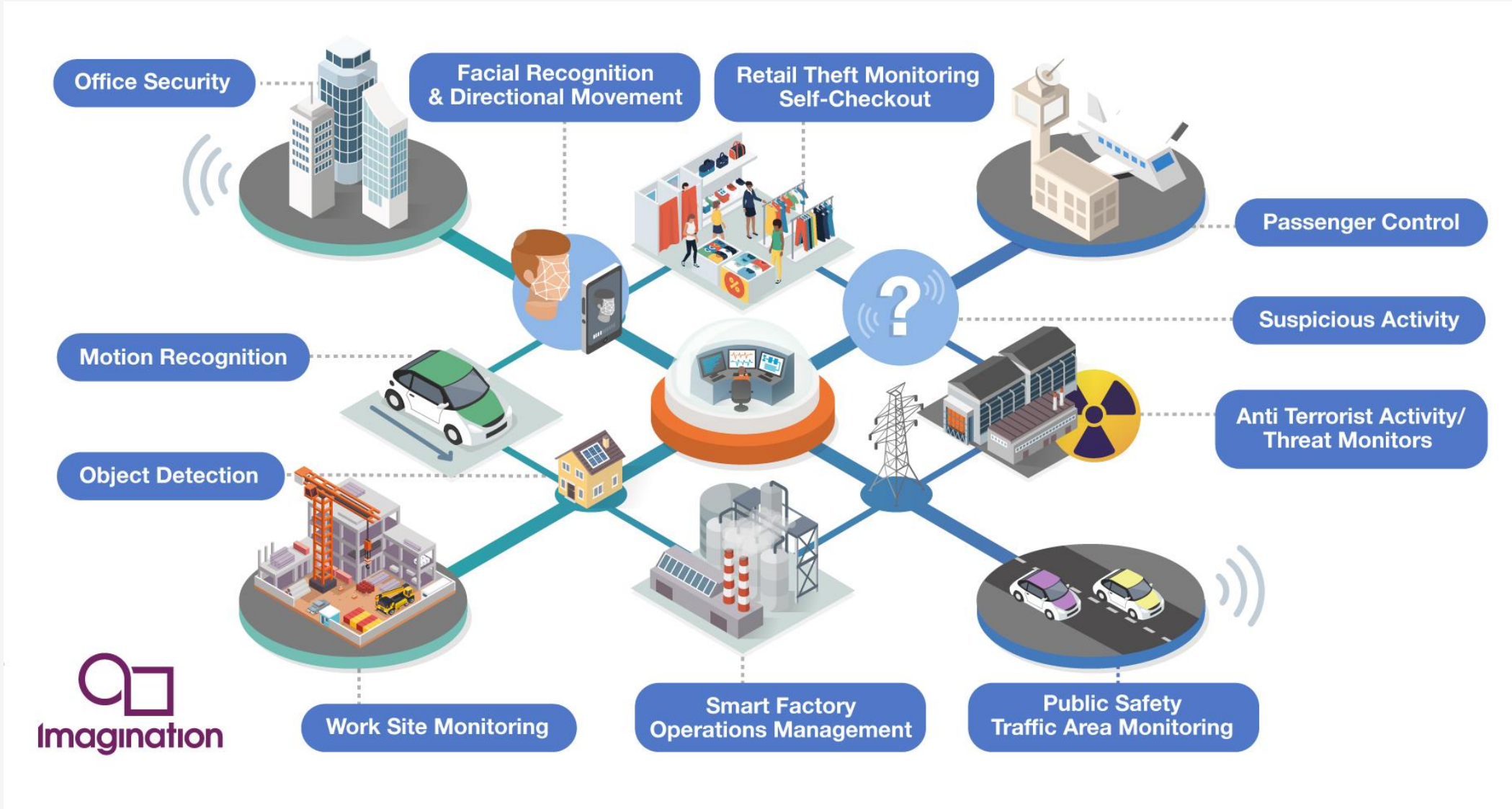


GPU: Render
NNA: Route planning,
Voice understanding

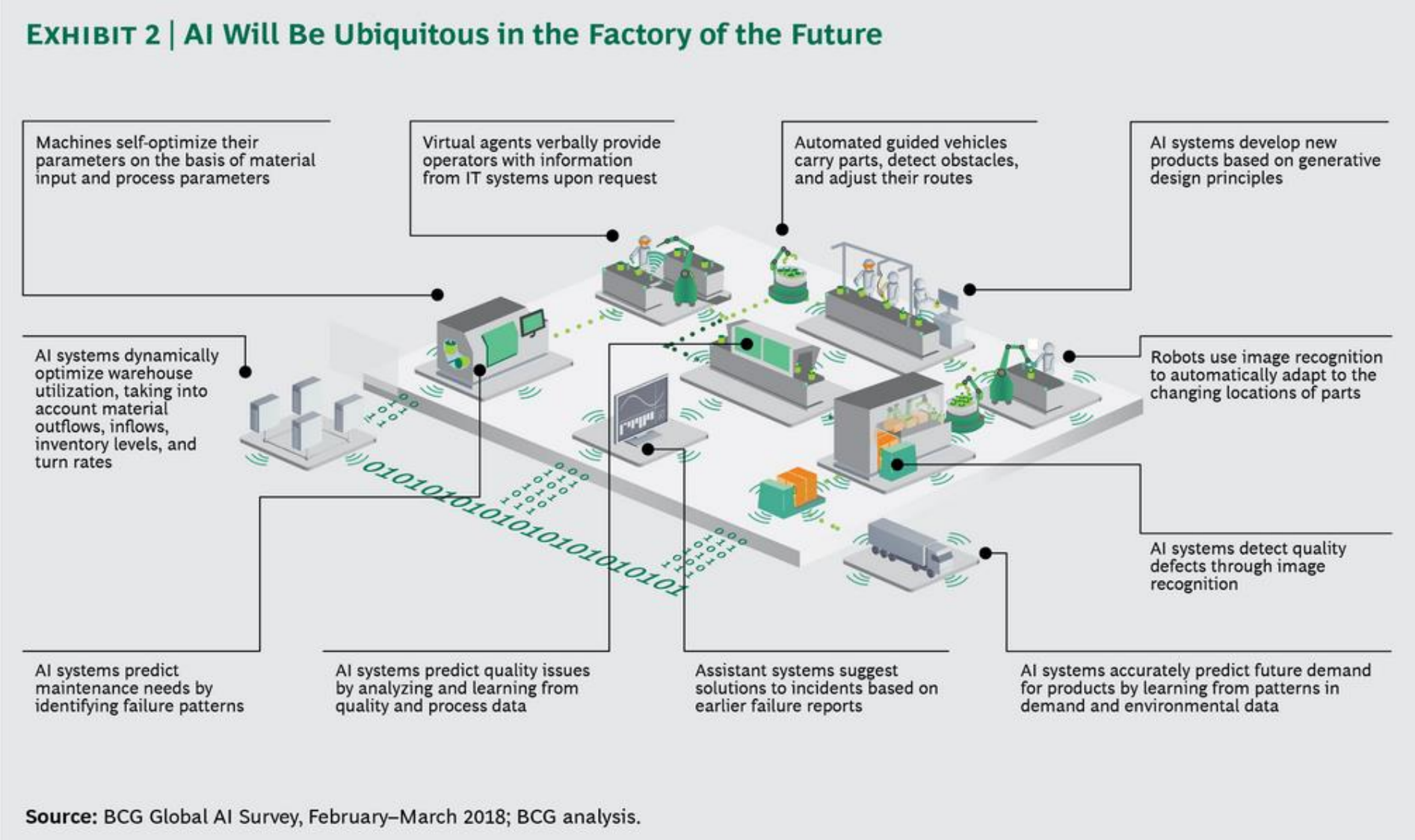
Imagination

2. Smart Cities

AIoT enables the Smart City



Smart Workplace + Robotics



Smart Camera Urban Infrastructure

Application

People Tracking

AI Technology

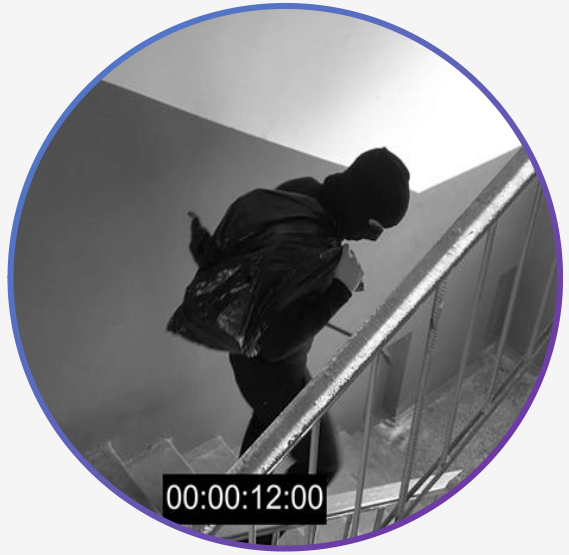
Face Identification

Application

Queue Monitoring

AI Technology

Scene Recognition



Application

Suspicious Behaviour

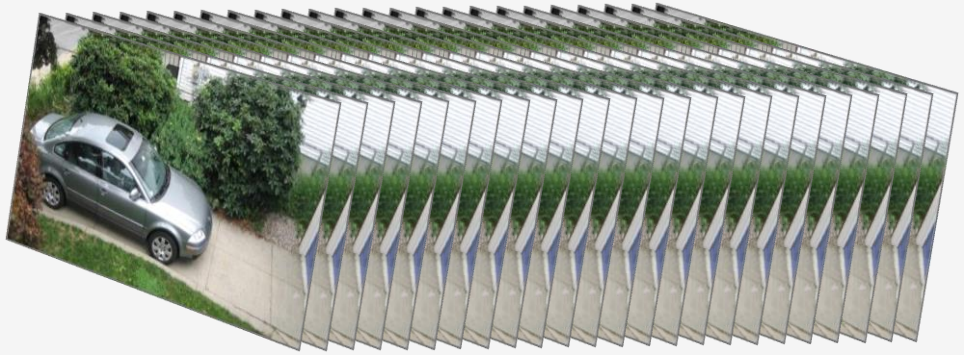
AI Technology

Object Detection

Edge Device + Cloud Brain = Efficiency – only ship vital data

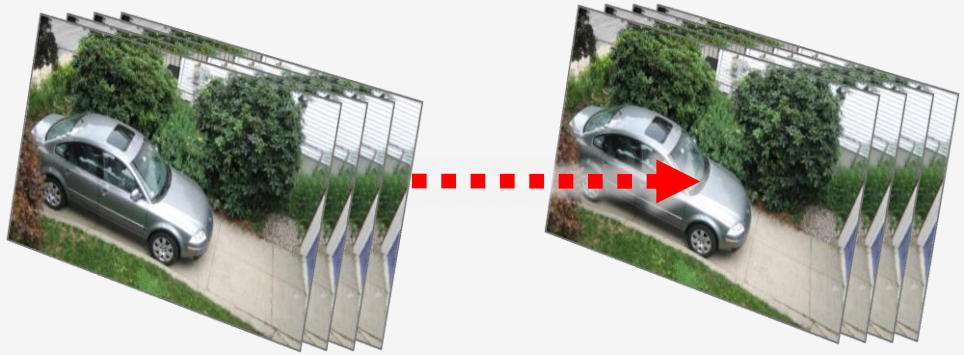
How local analytics can improve effectiveness

Standard camera



Send **all** video frames

Camera with motion sensor



Only **interesting** video frames

Smart camera with GPU + NNA

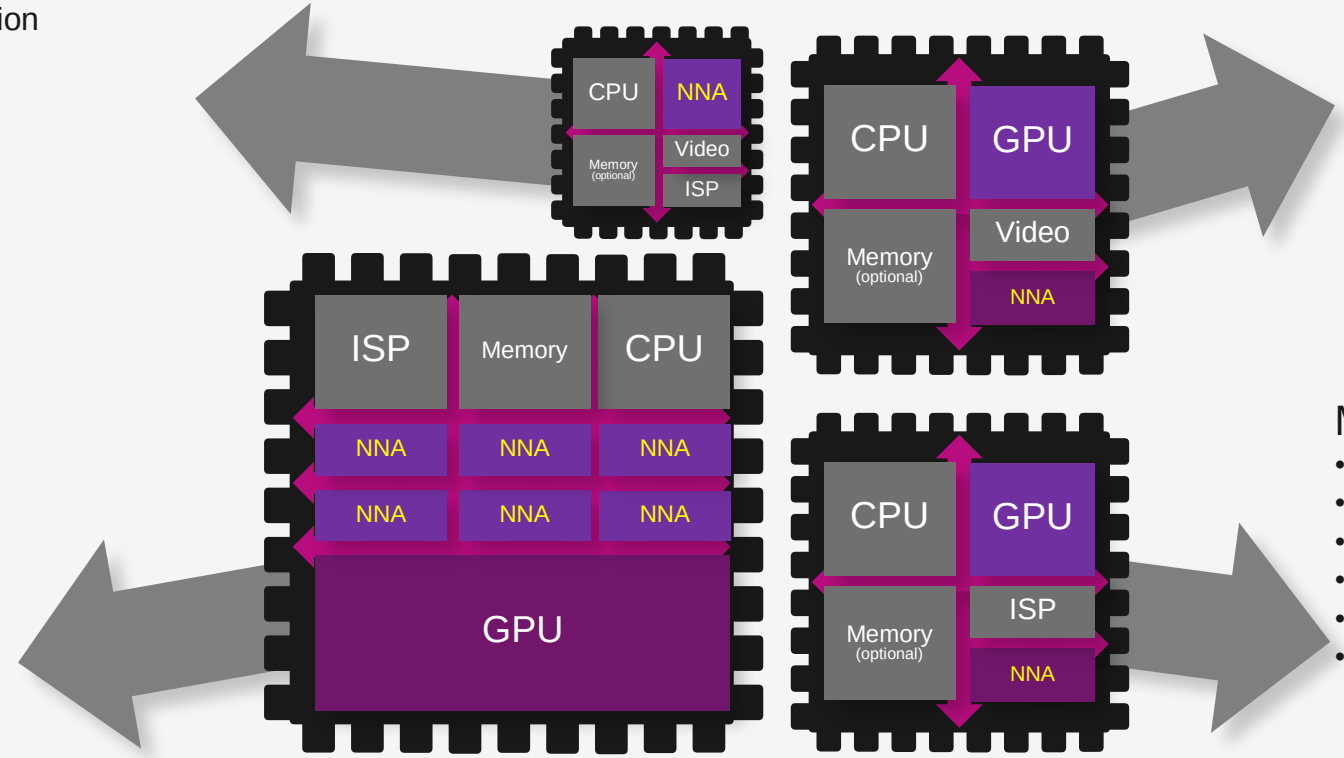
```
0  0  0  1111  0  0  0  0  111
```

Convert video to **metadata**

Choose the future

PowerVR flexibility: Addressing the markets

Security Camera
- People detection & recognition
- Emergency detection
- Emergency identification
  - Breaks
  - Fire/Smoke
  - Leak Identification



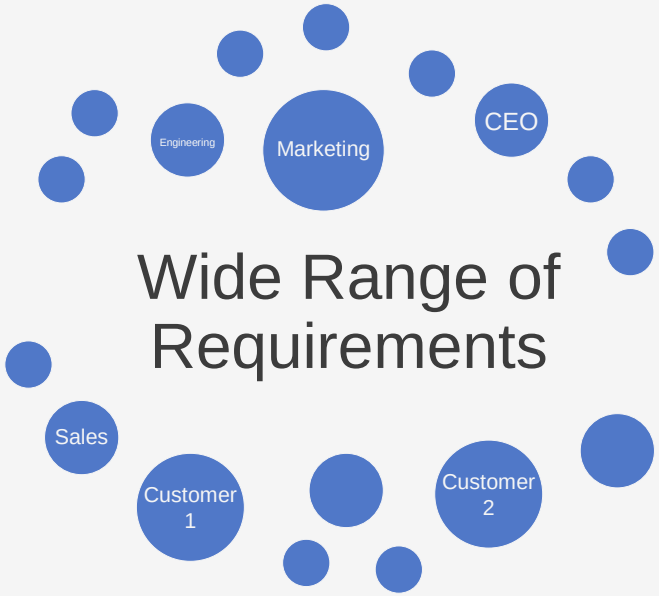
STB/DTV Platform
- User Recognition/Detection
- Video scene recognition
- Commercial detection

Automotive
<u>Non-ADAS</u>
- Lane detection warning
- Driver distraction warning
- Street Sign Detection
- Blind spot warning

<u>ADAS</u>
- Environment Recognition
- Obstacle detection
- Feed into ASIL Brain

Mobile
- Scene Detection/Recognition
- Image Depth Estimation
- Super Resolution
- Speech Recognition
- Noise Reduction
- GPU post processing
  - MSAA
  - Depth of field

Balancing Requirements with Reality: GPU

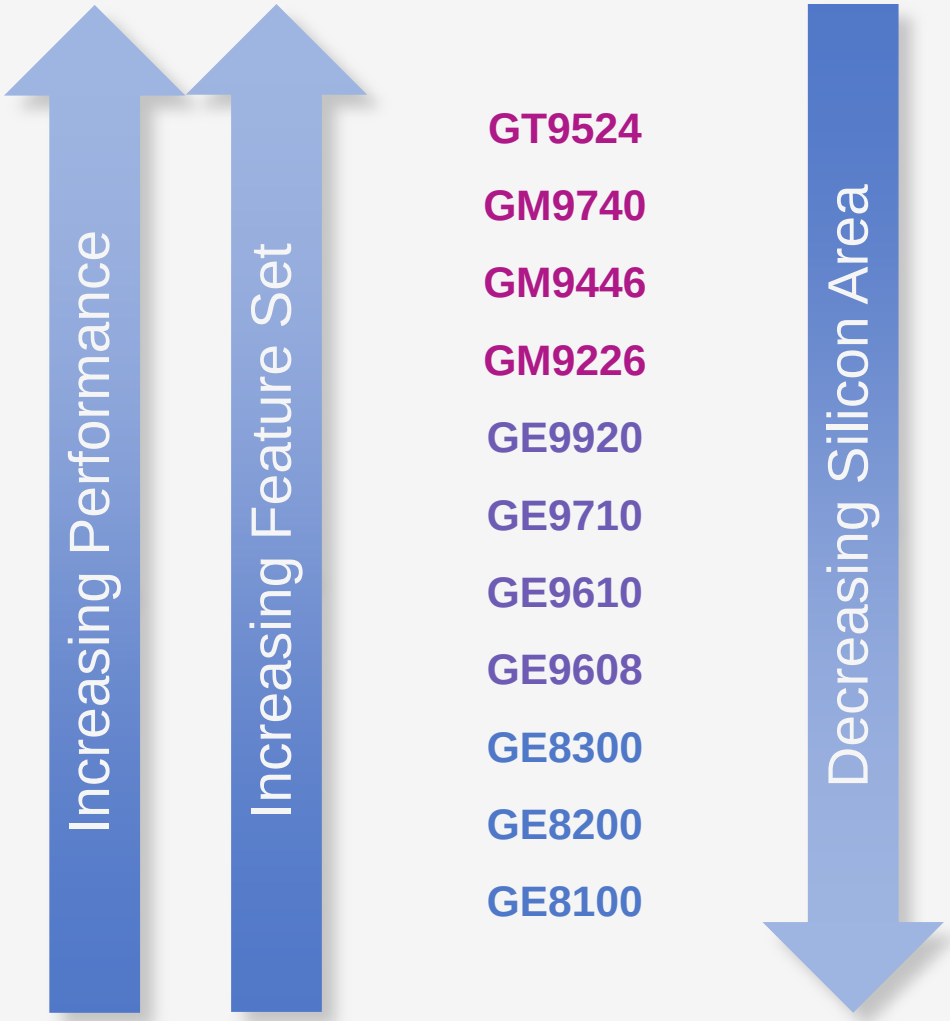


Real World Practical GPU

- Extreme Performance Expectations
- Tiny silicon Area
- All the features in the universe
- Virtually no bandwidth usage
- Photo-realistic
- Market-leading
- Perpetual Motion Machine
- …

- Performance matching true requirements
- Balanced silicon area usage
- Sensible feature set
- Minimal bandwidth usage
- Industry Standard Conformance
- Market-leading within segment/budget
- Power efficiency
- …

Balancing Requirements – Silicon Area



Balance requirements versus Area Costs

Target Chip Cost determines Silicon Budget

- Excessive GPU area prices SoC out of target segment

GPUs take in a growing % of SoC die

- Never-ending user expectations
- Continues to scale linear (unlike CPU)
- Largest component in most SoCs already

The right core for your requirements

PowerVR GPU Product line-up offers a wide range of performance/area/feature trade-offs

Case Study – UHDTV SoC – Low Cost

GPU Selection Process Example

Requirements → *Balancing* ← *Limitations*

4K Screens

Smooth 30 FPS

Performance

GUI
'Not' Games

YUV handling

OpenGL ES 3.x

Features

Android & Linux

Gestures Camera Vision OpenCL

PowerVR GE8300

Sensible Power

Low Bandwidth

Cost & Silicon Area

Cost dominates in the TV Market
GPU Size and Bandwidth driven by Cost
Conflict with very high fillrate requirement (4K)
OpenGL ES3.x Area Efficient GPU

PowerVR Graphics Strengths - Summary

Software & Ecosystem

Leading API Support

Free Tools & SDK

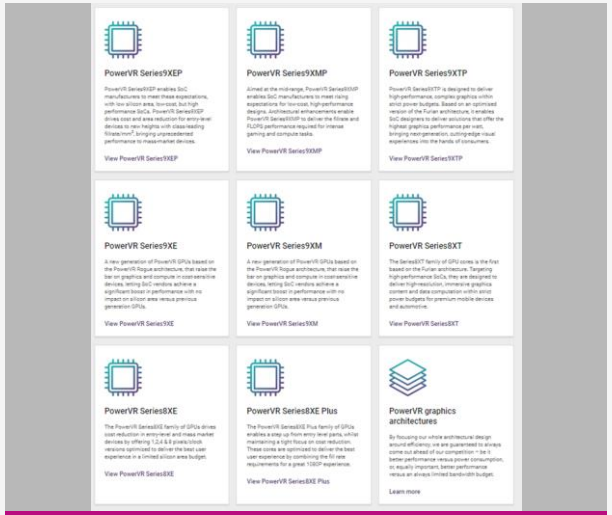
Ecosystem

Efficiency Focussed

Performance/mW

Performance/mm²

Bandwidth Reduction



Range of Solutions

Performance Scalability

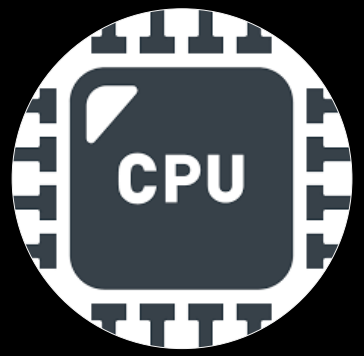
Feature Scalability

NNA Synergy

How we enable the Edge

Smart AI

Why GPU and Neural Network Accelerators provide the optimum solution

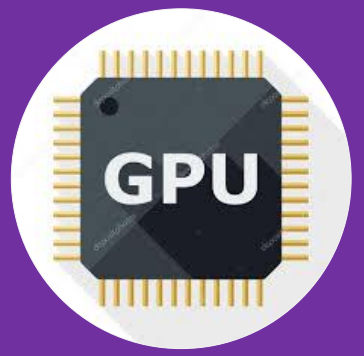


CPU

- Fully Flexible
- BUT inefficient and slow for high compute workloads

DSP

- Fully Flexible
- BUT hard to program – no standardisation, INT focussed

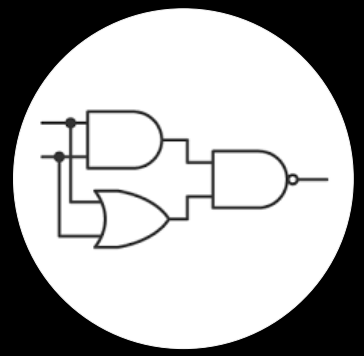


GPU

- Fully Flexible
- Standardised APIs for Compute, Float and INT support

Neural Network Accelerator

- Configurable
- Lowest power with domain specific flexibility



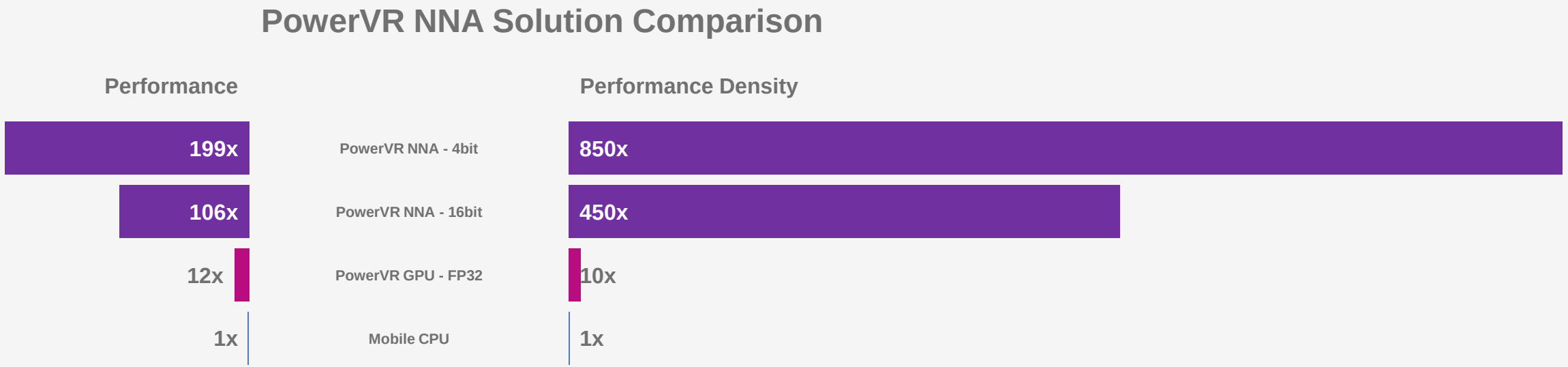
Fixed Function

- Single usage case
- Lowest power BUT zero flexibility

Best partners

Why the dedicated accelerator approach

Extracting the best performance for the area budget



Neural Networks have high bandwidth and computation requirements

A dedicated architecture addresses both of these issues

Cost efficient because of mobile experience

PowerVR NNA is designed to deliver the best performance per mm^2

NNA: Series3NX highlights + PowerVR8300 GPU

New core AX3595

- NNA new single core
- 4,096 MACs/clk
- Up to 10 TOPs

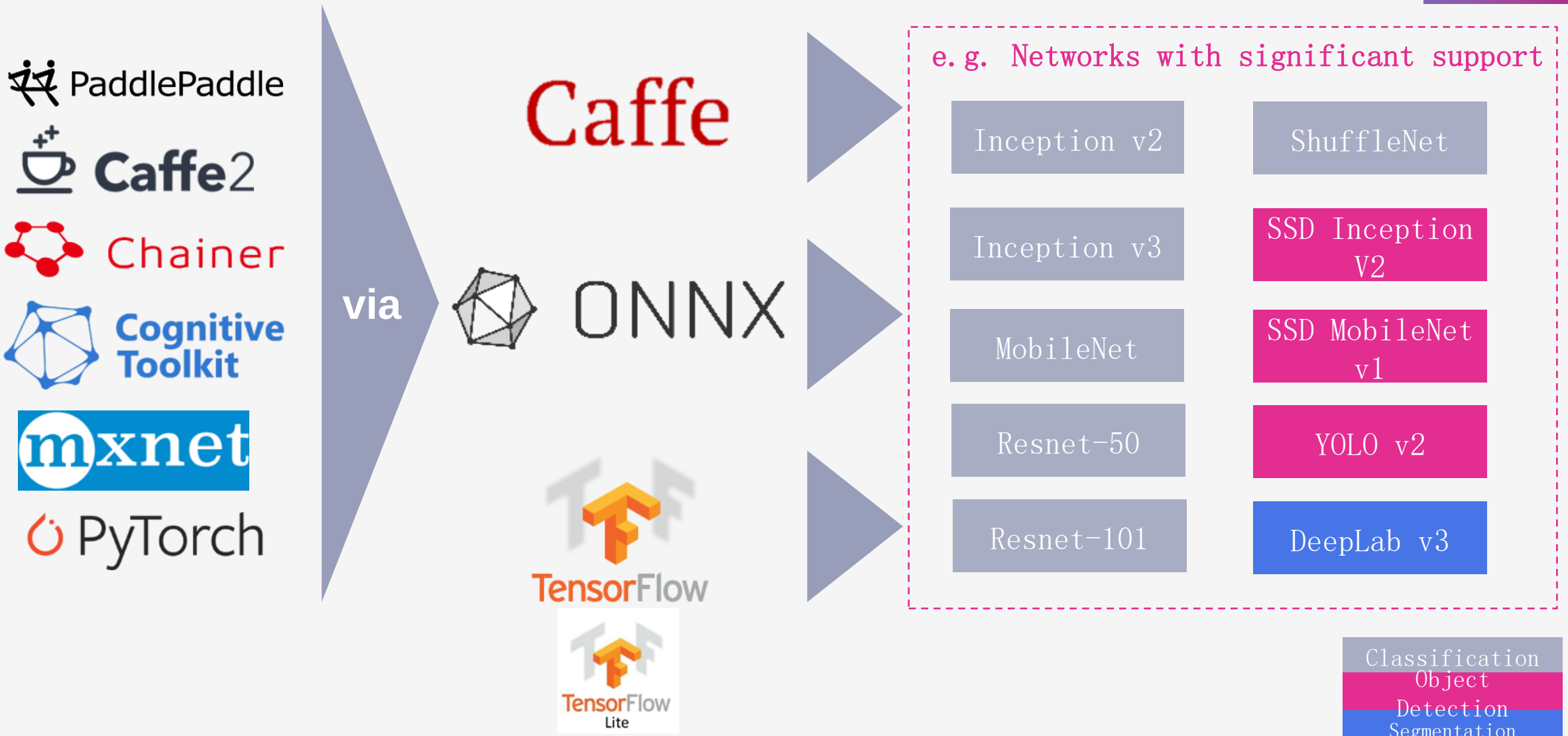
Improved PPA

- Up to 70% better inf/s
- Up to 40% better inf/s/mm^2
- Up to 35% lower bandwidth

New Features

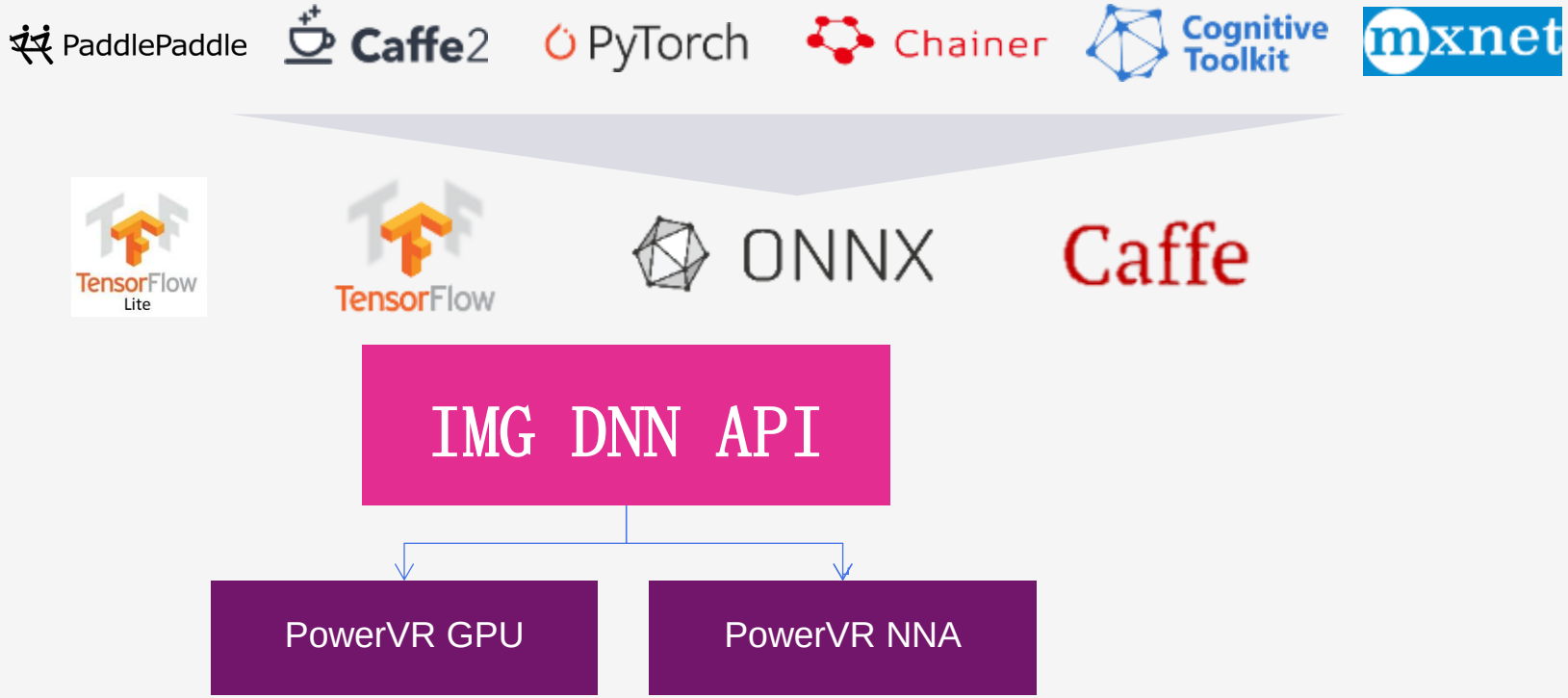
- Lossless weight compression
- Advanced security enablement

Acceleration for popular Neural Networks



PowerVR AI: One combined API

Your network, one API, optimised for GPU and NNA



Cost	Lowering effort to support a wide range of networks
Performance	Zero copy memory sharing between IP, driver-level synchronisation
Ease of use	Common tools across GPU and NNA

Award Winning

PowerVR NNA

- Team of the year 2018
- Design team of the year 2018
- Best IP product 2018

National Technology Awards

beea
british engineering excellence awards

Building the NNA ecosystem together

神经网络生态

Imagination

Thank you

www.imgtec.com