



Quadric

Quadric Processors for Edge AI

DAC 2023

Panel Session 102: AI on the Edge

Monday 10 July, 2023

About Quadric



Pure play Semiconductor IP Licensing

- Processor IP & Software Tools

Edge / device AI/ML Inference + DSP processing

Founded: 2017
HQ: Silicon Valley – Burlingame CA
Series B: \$31M in 2022

IP product announced Nov-2022
First IP deliveries Q2-2023
DevStudio Online May-2023

Patents: 10 Granted, +4 Filed

Silicon Proven Test Chip



Quadric Team

Startup & IP Licensing Veterans

Management Team

- Multiple startup exits
- Extensive experience in IP licensing, Machine Learning & Processor IP

Engineering Leadership

- Senior talent averaging 10+ years of experience at companies such as Nvidia, Intel, Nirvana, NXP

Board & Investors

- Early product to IPO
- Enterprise Software & Silicon Expertise



Co-Founders



Veerbhan Kheterpal, CEO
3 startups, 2 exits
15+ Patents
CMU Ph.D.



Nigel Drego, CTO
2 startups, 1 exit
15+ Patents
MIT Ph.D.



Daniel Firu, CPO
2 startups, 1 exit
10+ Patents
UF M.S

Semi IP Industry Veterans on Exec Staff



Tim Smith, VP Sales
Ex VP of Sales at
Rambus, Coware, Sonics,
Founder & MD at Memec



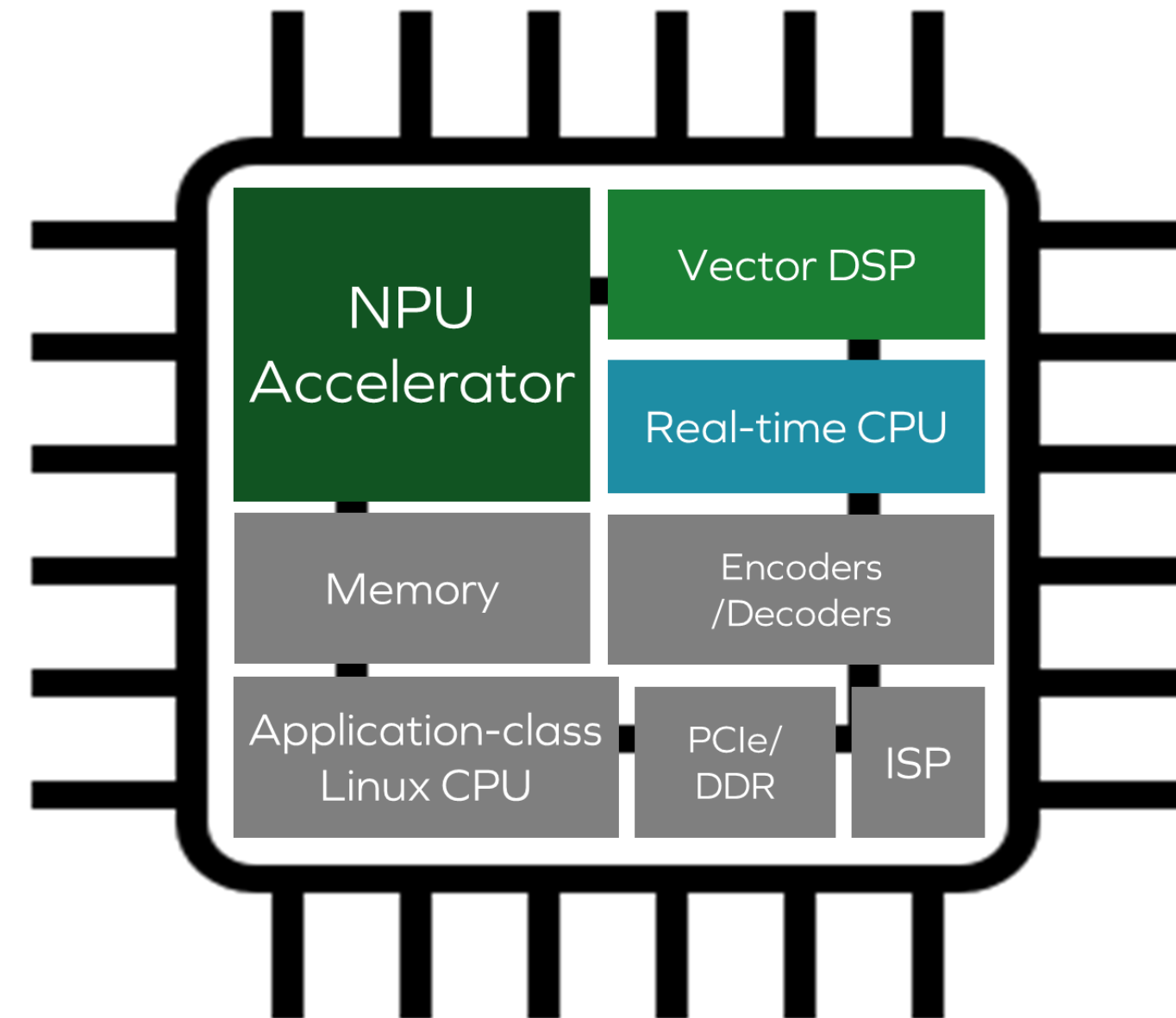
Steve Roddy, CMO
Ex Arm VP Machine
Learning; VP/GM Tensilica,
4 Startups



Dhanendra Jani, VP Eng
Ex Cadence VP,
Tensilica, Intel
Virginia Tech M.S.

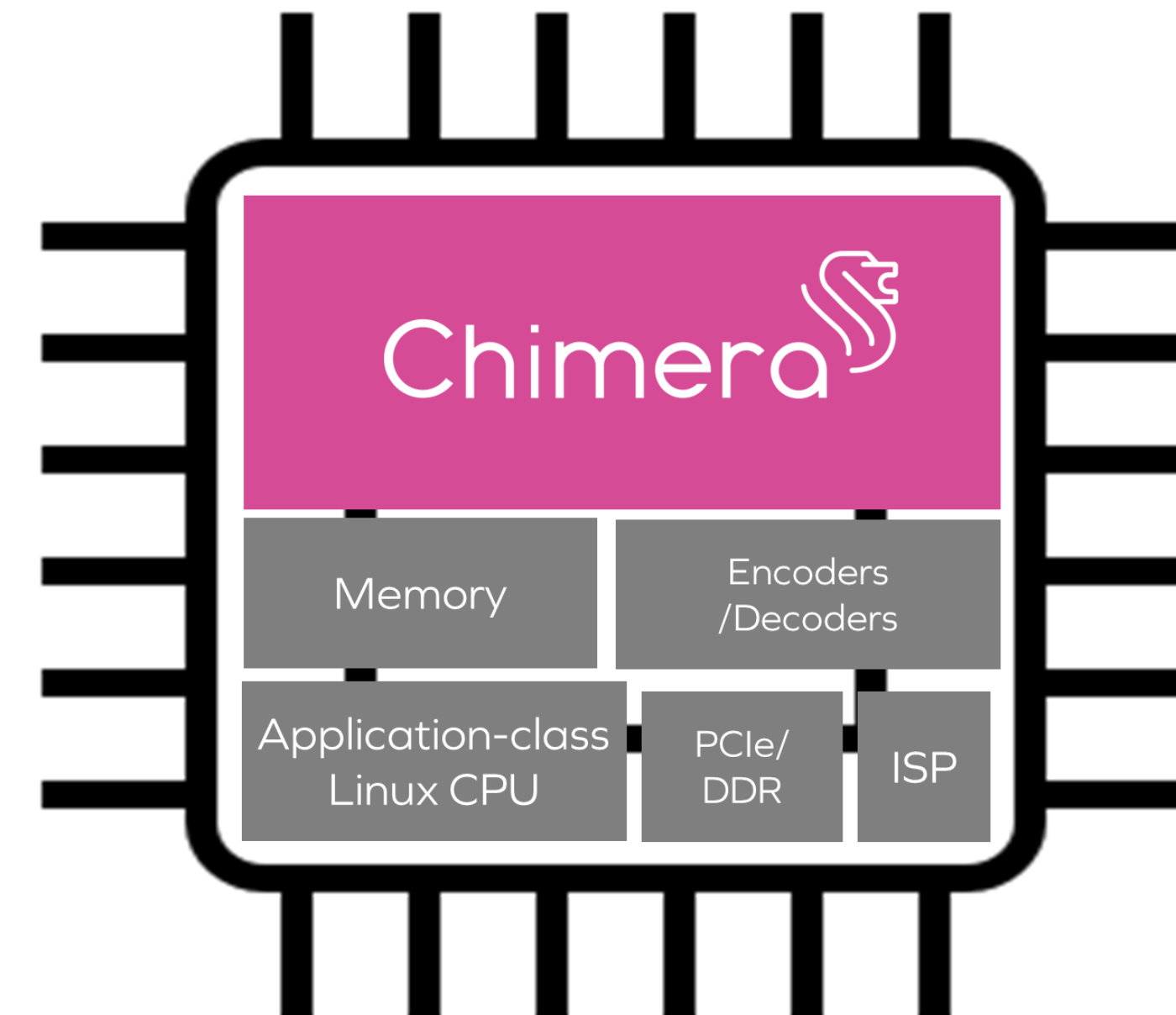
Chimera GPNPU – Speeding Edge AI Programming

Before Quadric Chimera GPNPU



Conceptual block diagram of a Smart Camera SoC

Chimera GPNPU from 1 TOP to 16 TOPs



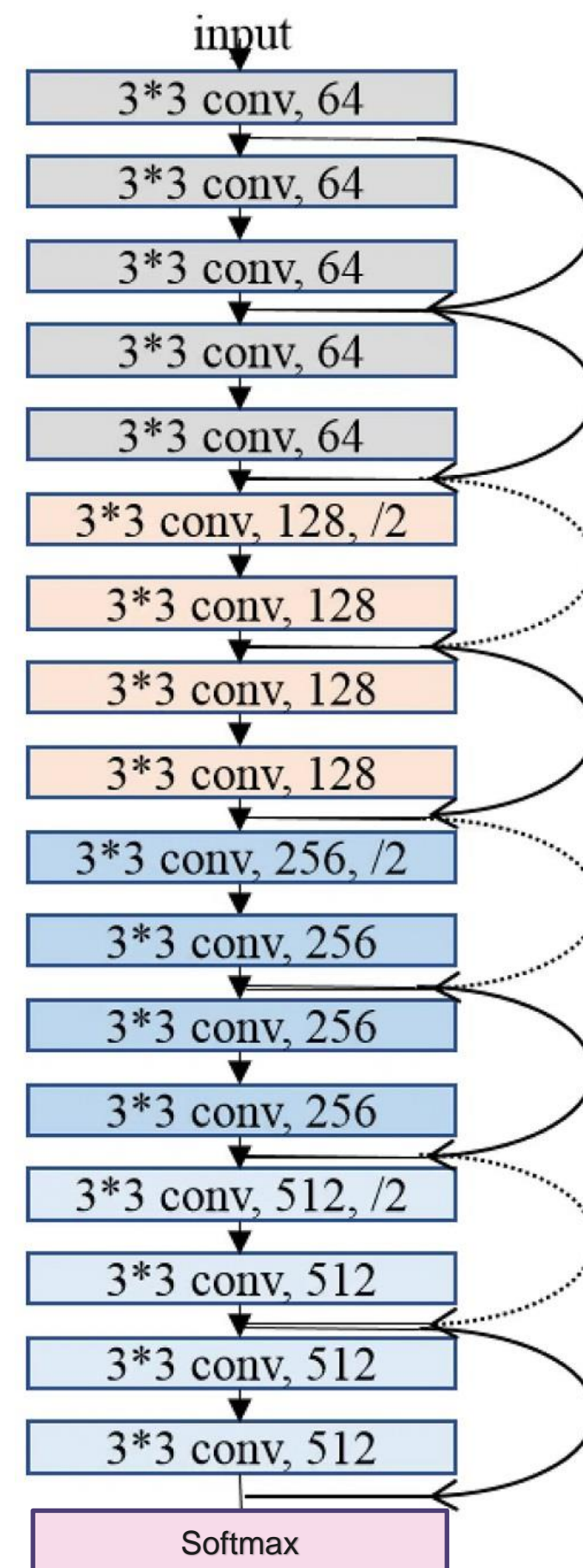
↓ System Complexity / Power
↓ Programming Complexity (not productive)
↓ Accelerator Brittleness – High risk of obsolescence

↑ Simpler SoC Architecture
↑ Dramatically Easier SW Programming
↑ Long SoC Lifespan – Easy ML Operator support

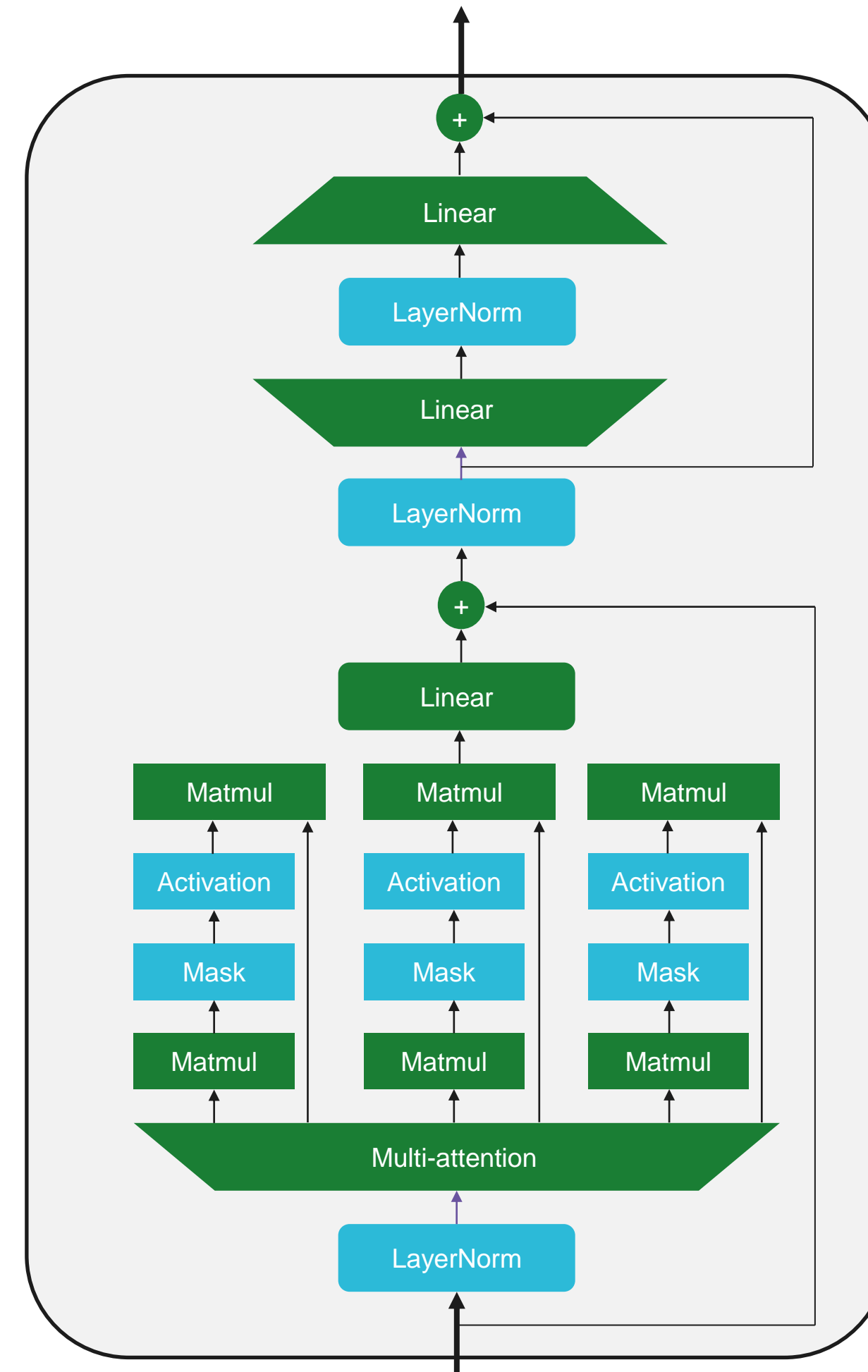
Generative AI/Transformers Broke Your NPU Accelerator

Resnets – Circa 2018-2020 SOTA

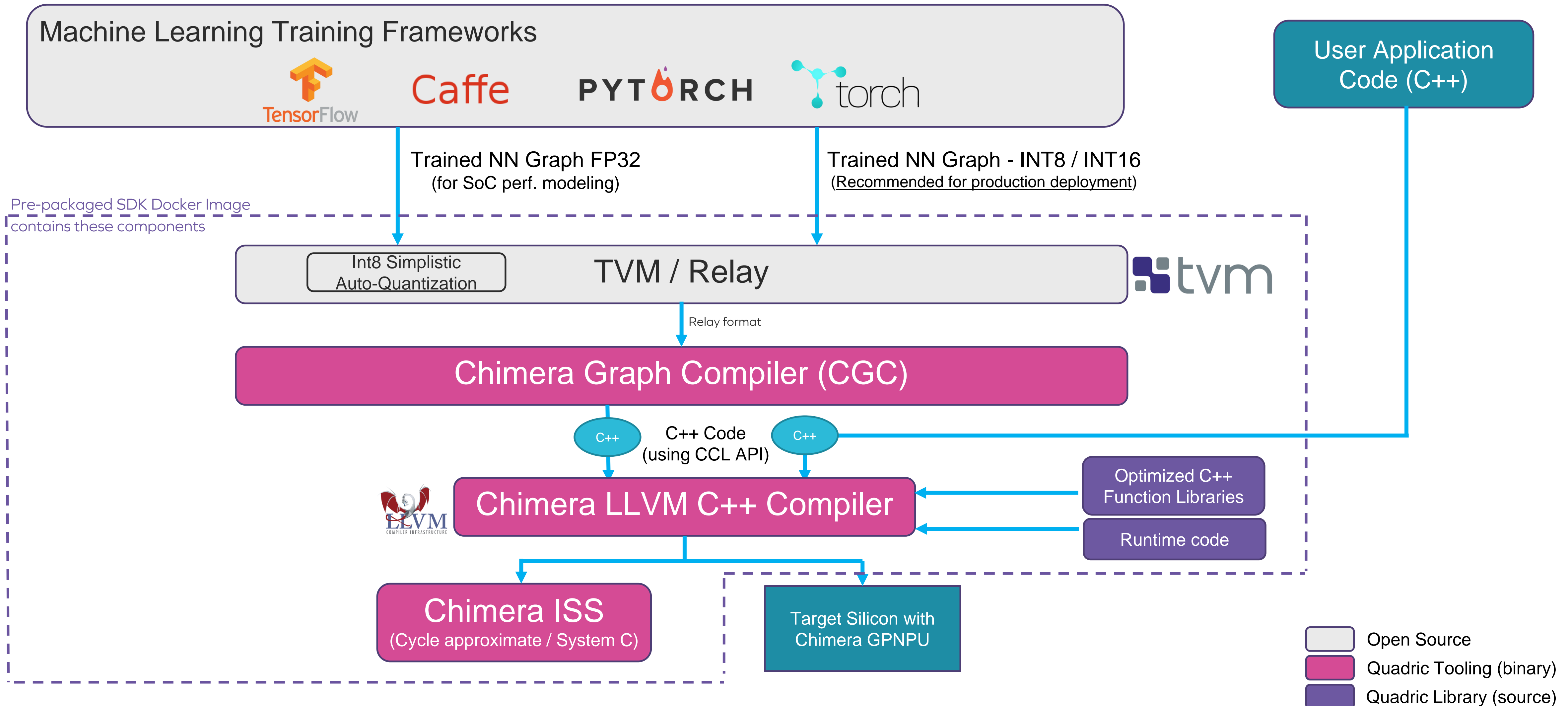
Resnet18 Architecture



Transformers – 2023 SOTA

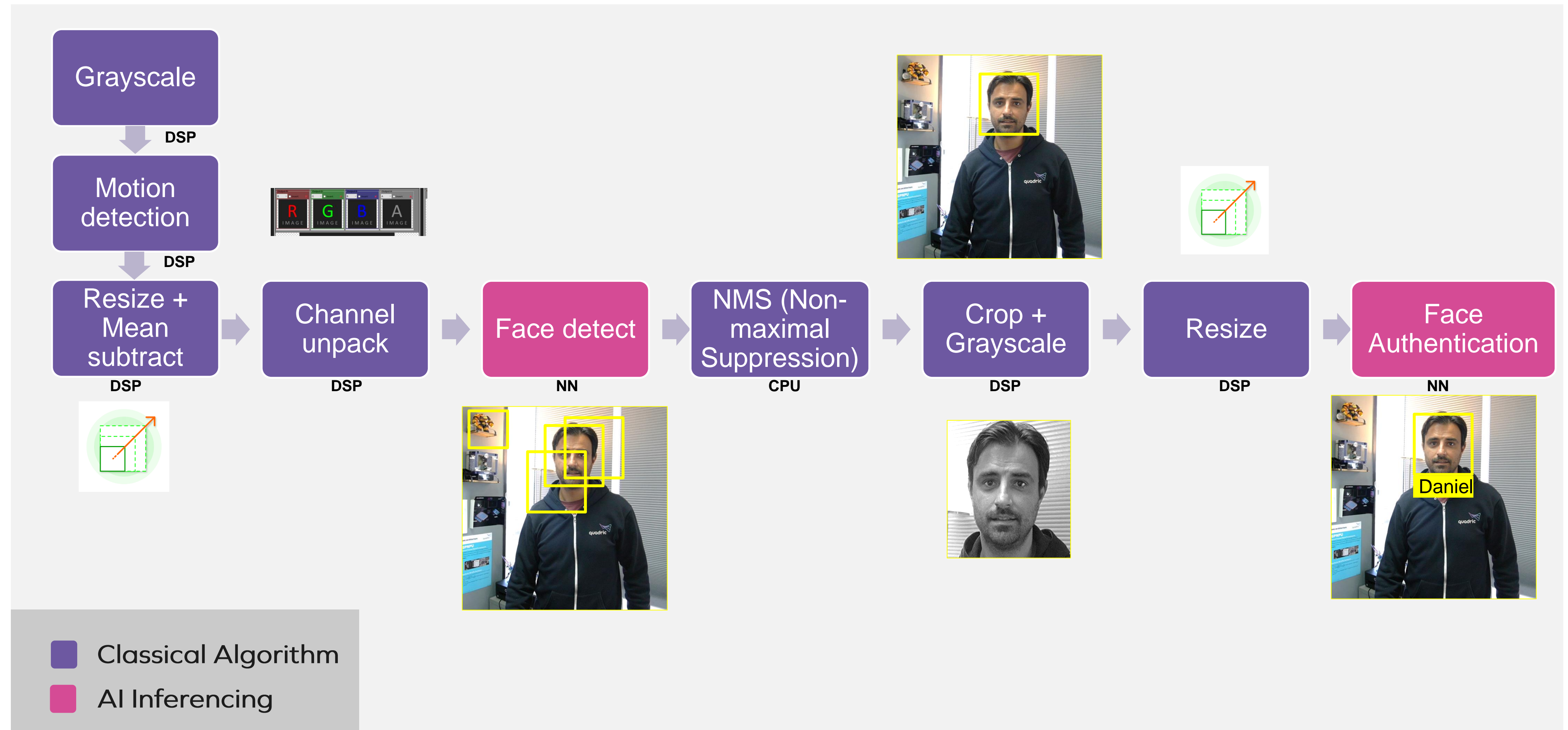


The Key Innovation: Two Code Styles -> One Target



Example – Full pipeline for person authentication

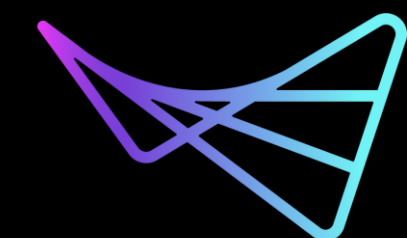
All 9 Kernels Run on the single Chimera GPNPU



Why Chimera^{\$} for Edge AI?



- High Inference Performance
 - . Architecture purpose-built optimized for convolution
- Programming simplicity
 - . Write and debug graph code & C++ code on one core, not three
 - . Edge AI → means lots of designers, coders, products. Ease of SW is key
- Future proof – without Fallback
 - . C++ programmability – at HIGH PERFORMANCE - for future NN operators



www.quadric.io

Thank you