47.5mm × 47.5mm Package

2.5mm × 2.4mm Die

Chip-to-chip interconnect

Activation storage

Processing elements with distributed weights

On-chip weight storage (8-bit parameters)

Datacenter

Automotive

Mobile

18M

8M

2M

0.5M

1 Chip (1×1)

4 Chips (2×2)

16 Chips (4×4)

36 Chips (6×6)

Fig. 1: System overview of the proposed MCM-based DL accelerator.