SEMICONDUCTOR PROCESSING CHALLENGES POSED BY BURGEONING AI CHIPSET MARKETS

November 2019

ANAND JOSHI

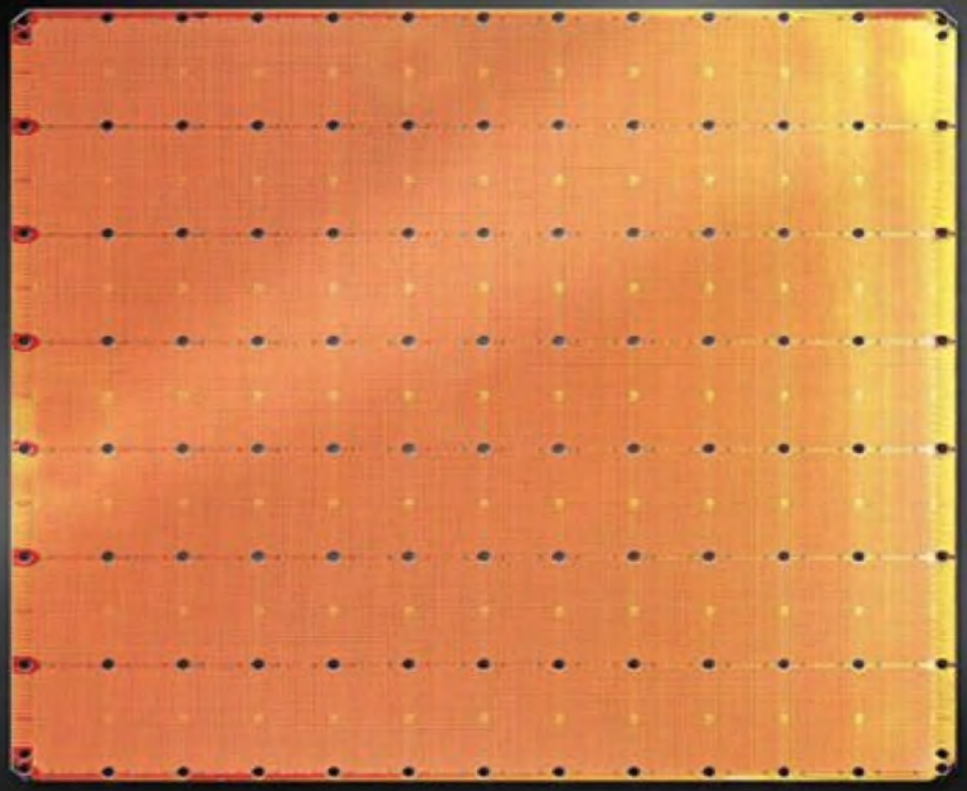
anand@joshipartners.com

About me

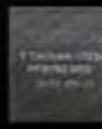
- Independent Consultant, Industry Executive and Analyst, AI Computer Vision
    - Tractica, Wave Computing, Thinci, Alten Calsoft, Synopsys, LSI Logic Adobe Flash BU
- Involved in many interesting projects
  - Reports on AI chipsets, systems, computer vision, visual analytics
  - Product management, marketing and engineering executive
  - Investors and start-ups – business plans, due diligence, market strategy

AI is leading chipsets to scale never seen before

Cerebras Wafer Scale Engine

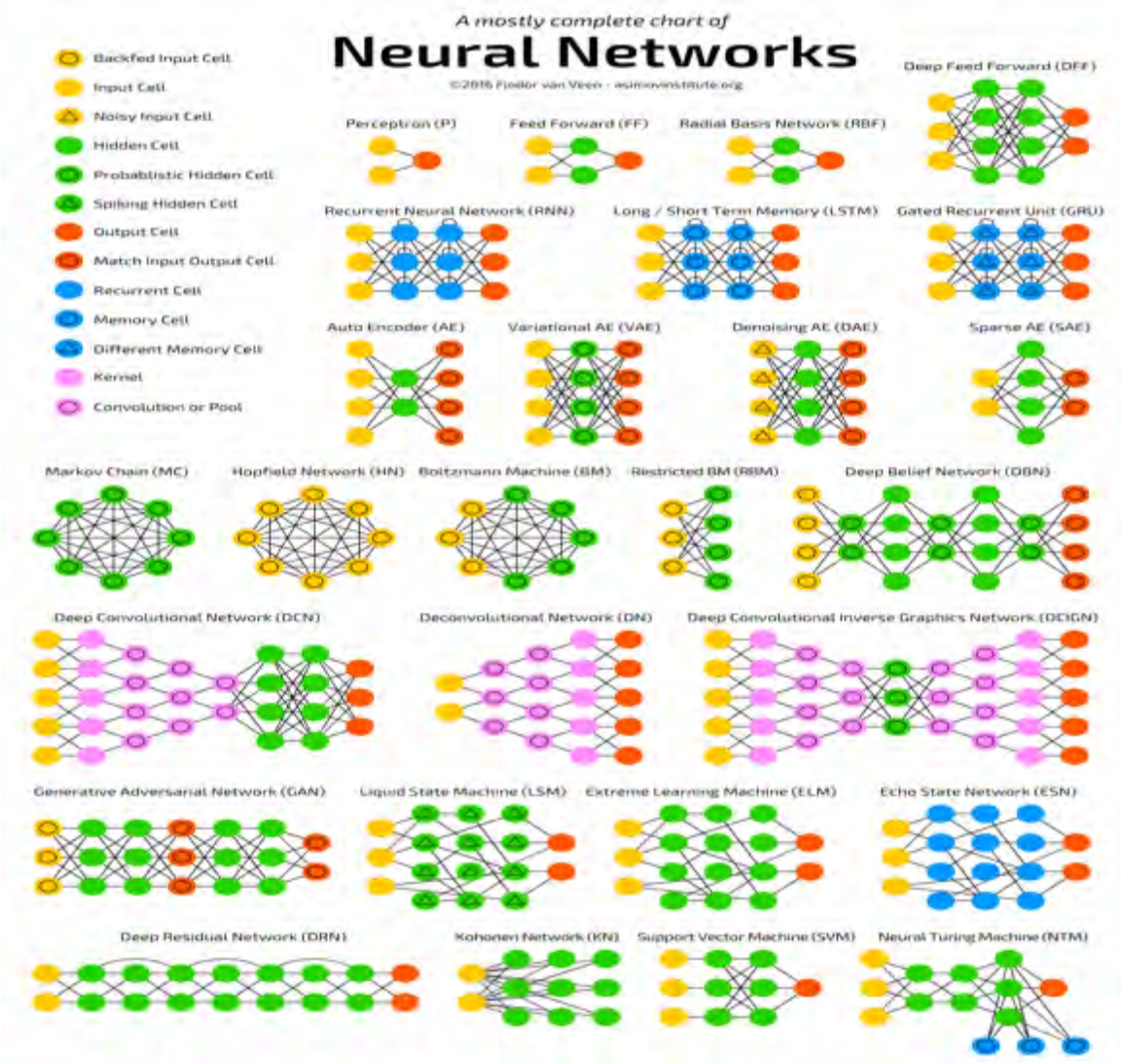


Cerebras WSE
1.2 Trillion Transistors
46,225 mm² Silicon

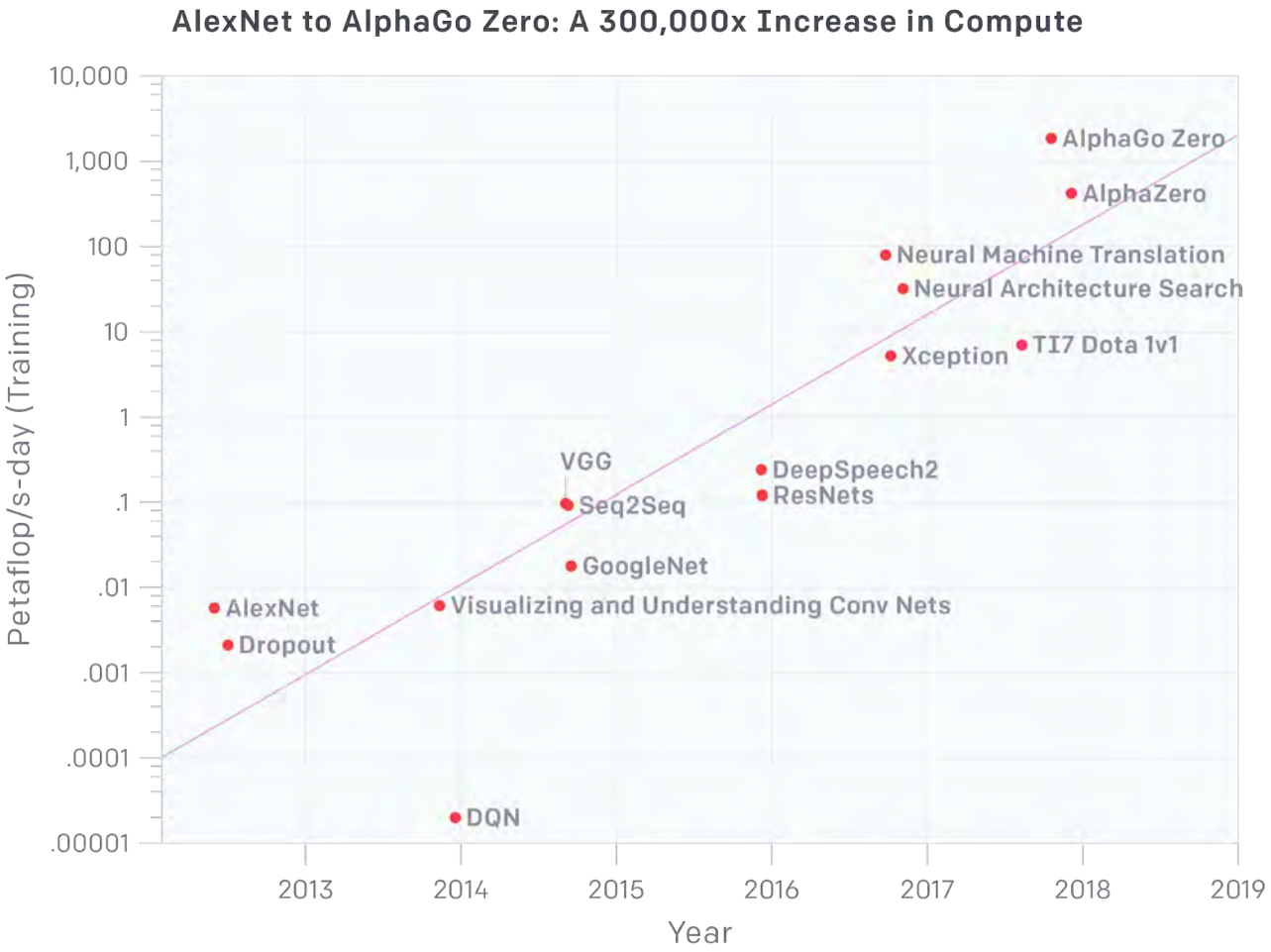


Largest GPU
21.1 Billion Transistors
815 mm² Silicon

Neural Networks come in all shapes and sizes



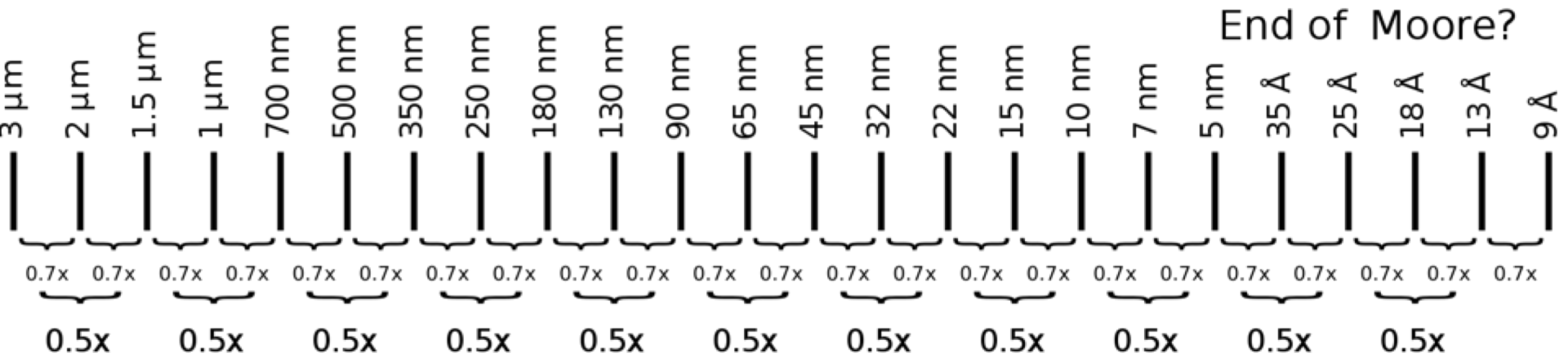
And they are getting complex every day



Compute needs for AI have doubled every 3.5 months

Source: OpenAI Blog

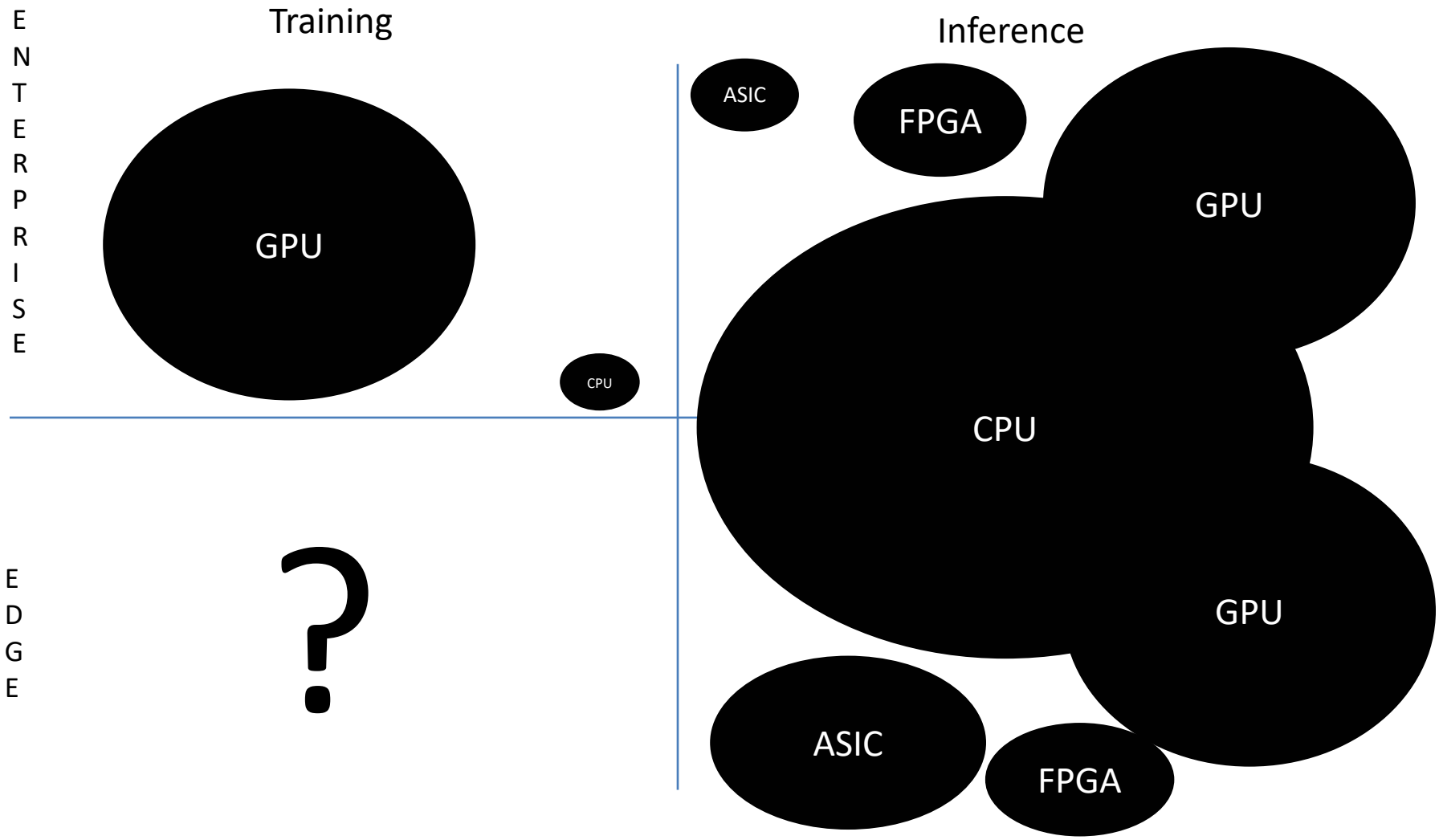
Moore’s law is ending



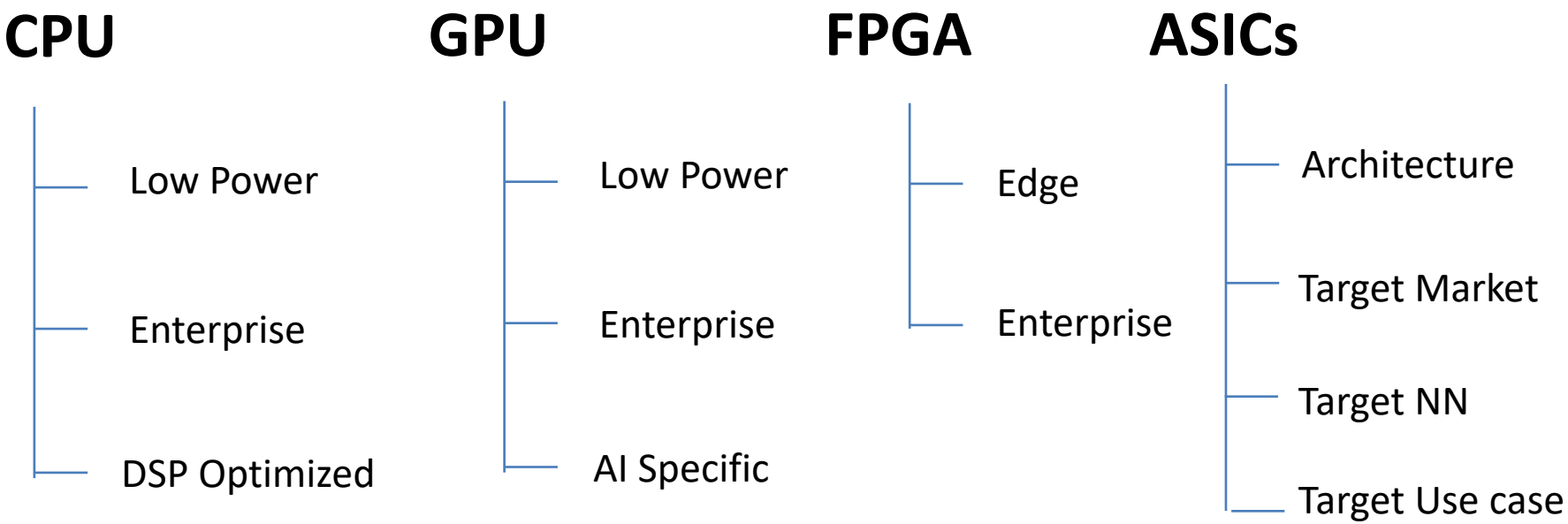
Chipset performance doubles every ~18 months

There’s a large gap in NN performance needs and what chipsets can offer

Many AI chipsets are needed to address market needs



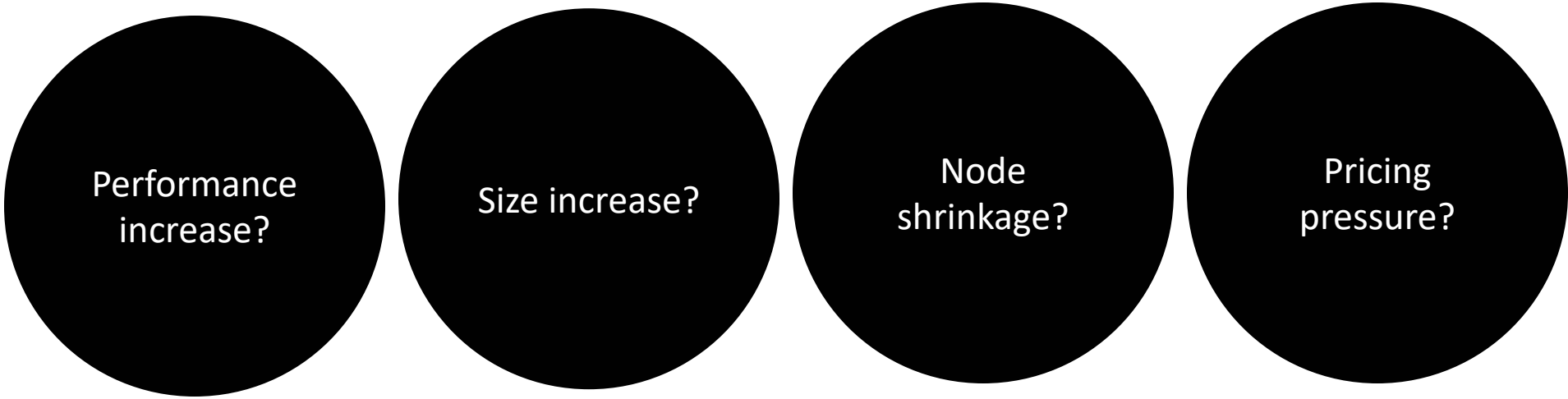
AI Chipsets come in all shapes and forms



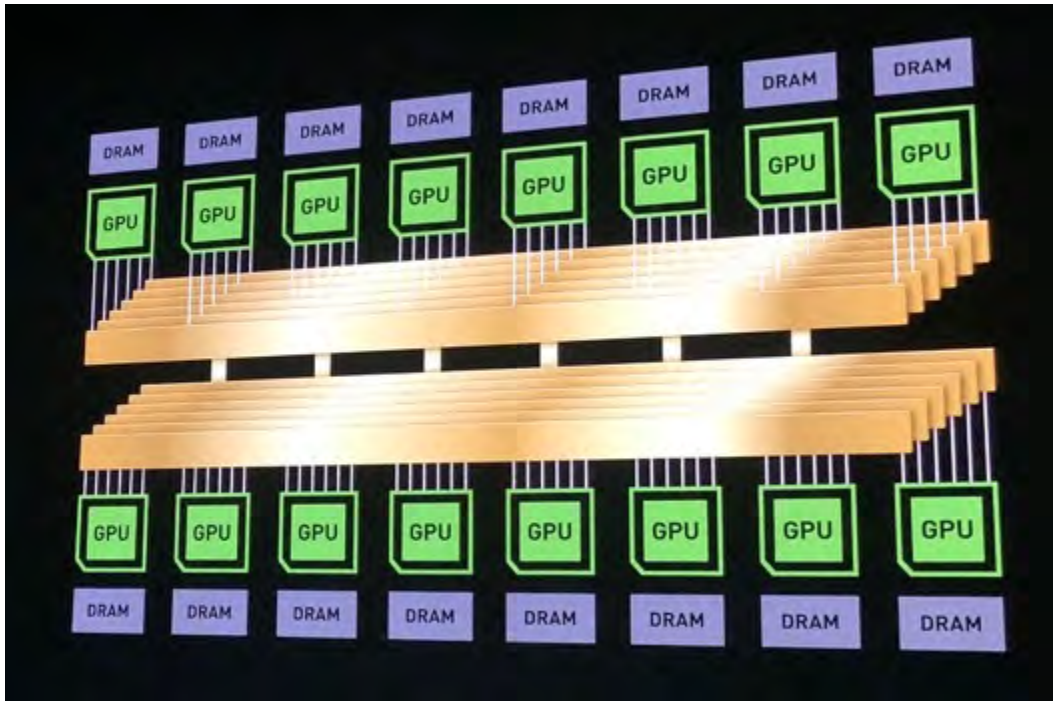
= Many new challenges (and opportunities) in semiconductor processing

Challenges for semiconductor processing – performance and size

Feature	Pascal GPU (2016)	Volta GPU (2017)
Technology	16/14nm	12nm
Size	471mm2	815mm2
Power	250W	250W
Performance	10 TFlops	120 TensorFlops
Price	~$10K	~$10K



Challenges for semiconductor processing – power and heat



DGX2

GPU power = 250W
System power = 10 kW
Datacenter Power = 10MW

Chip cooling? System Cooling?

Air vs Liquid?

Noise?

Challenges for semiconductor processing – manufacturing technology

Analog

Optical

Memristor

Alternate technologies?

Yield?

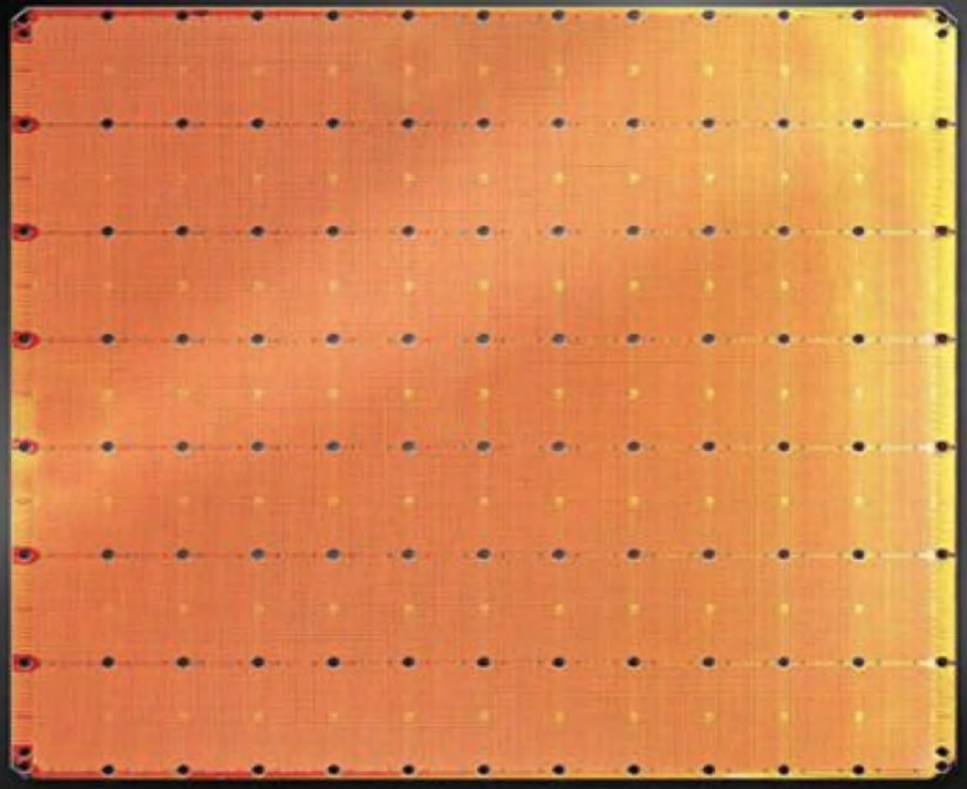
Price?

Power?

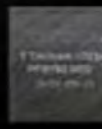
Edge vs Enterprise?

Finally – how to make this work?

Cerebras Wafer Scale Engine



- Yield?
- Die-to-die connectivity ?
- Package assembly?
- Power and cooling?

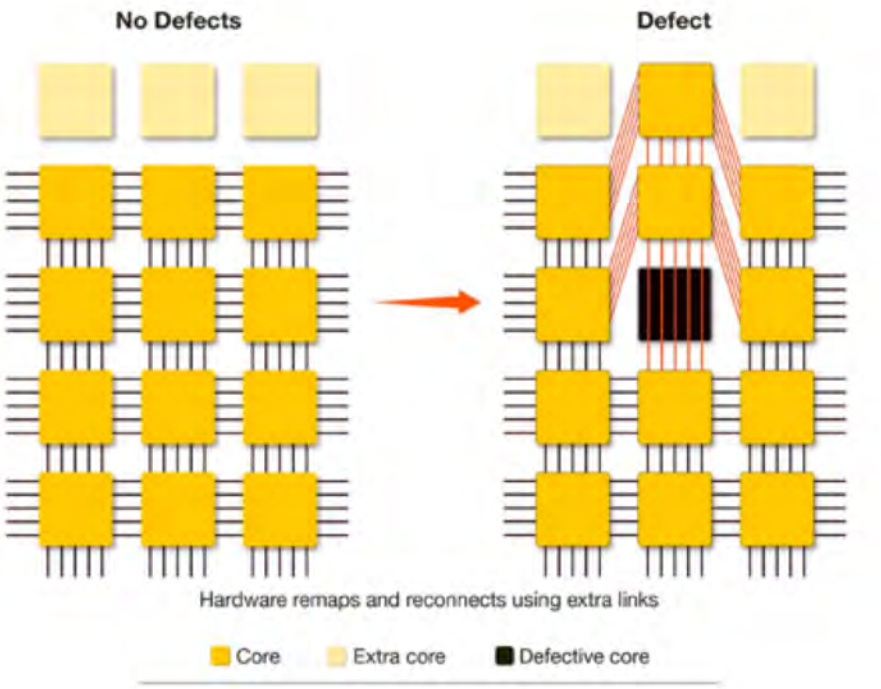


Cerebras WSE
1.2 Trillion Transistors
46,225 mm² Silicon

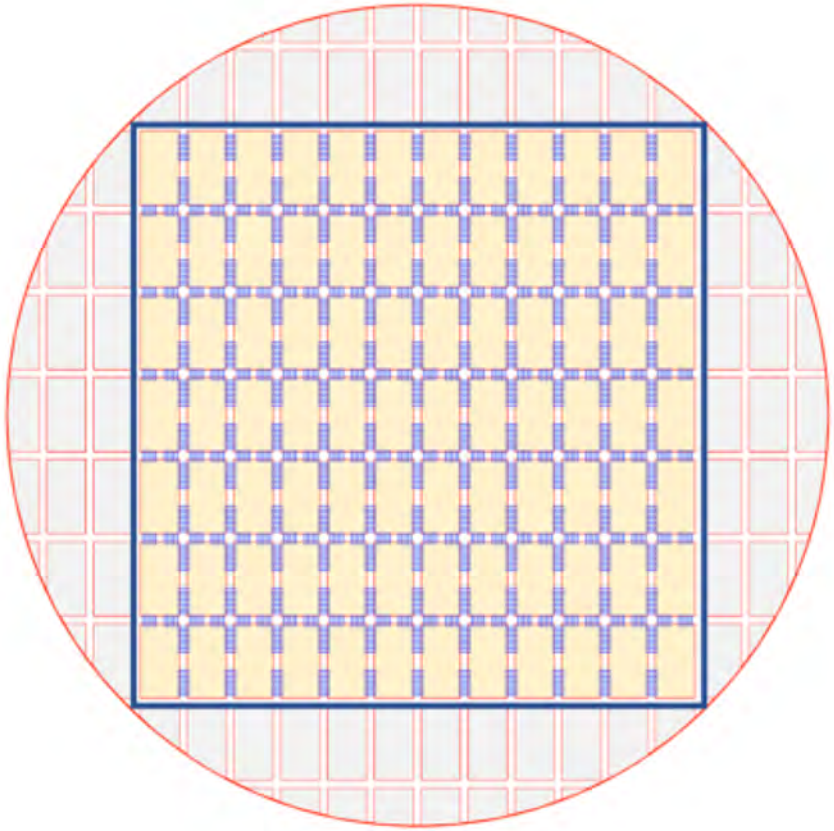
16 nm, 84 chips, 57X

Largest GPU
21.1 Billion Transistors
815 mm² Silicon

Challenges for Cerebras



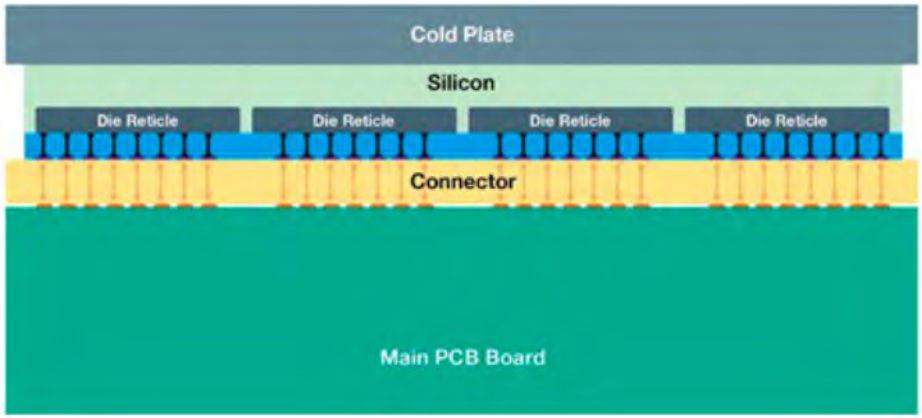
Rerouting for yield



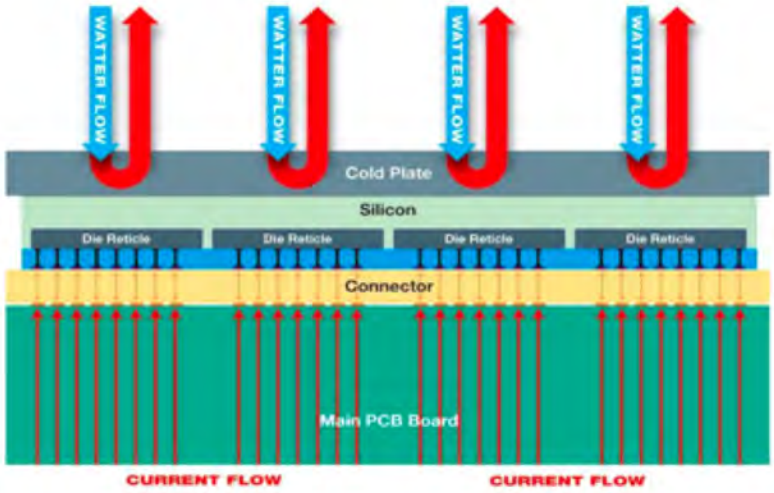
Die-to-die connectivity

Source: Cerebras

Challenges for Cerebras



Package Assembly



Cooling

Source: Cerebras

Questions and answers

anand@joshipartners.com