intel.

TECH tour.TW

Lunar Lake AI Hardware Accelerators

TAP
Intel Fellow

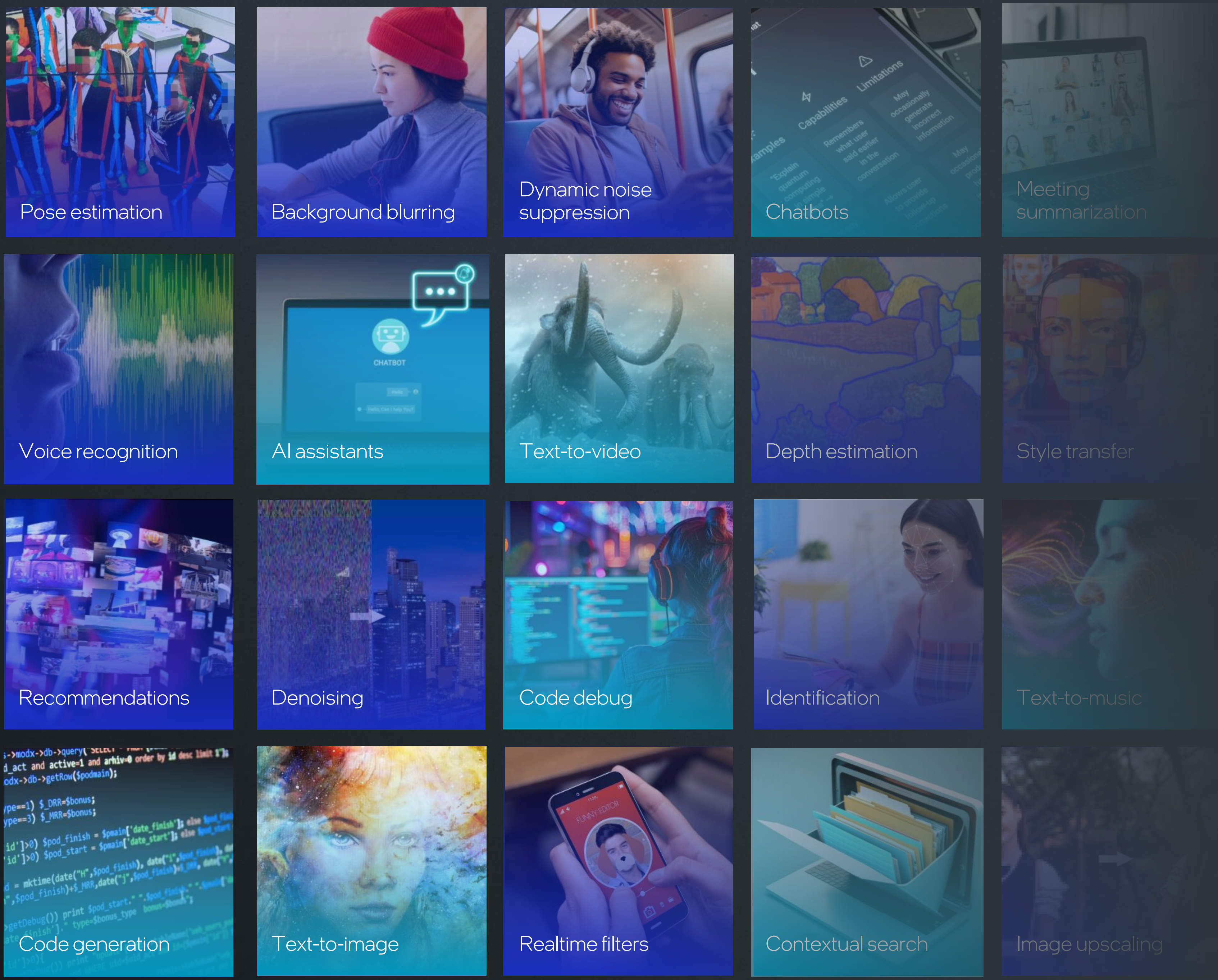
AI PC Landscape

AI Client Workload Trends

Growing in diversity
From background blurring to gen AI

App & OS integration
from features in apps to OS co-pilots

Multi Modal
Transformers and diffusion



TECH tour.TW

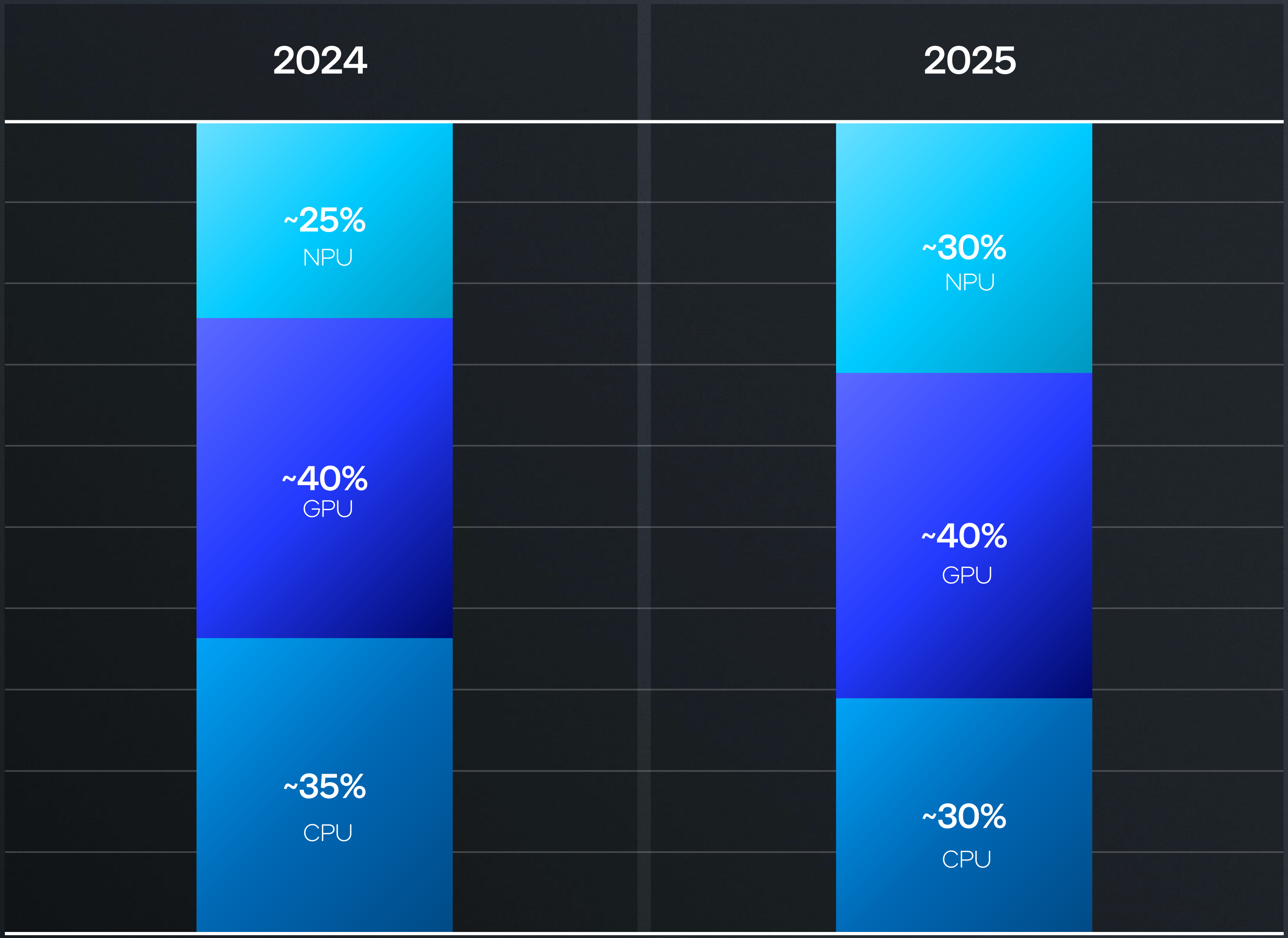
AI PC Trends

AI Engine Adoption

Multi-engine adoption
Macro ISV trend

GPU role significant
In ISV feature plans through ‘25

Multiple performant engines
Are best fit for enabling ISV efforts



Based on internal Intel research as of May 2024.

Unmatched AI Compute

With our Multi-Engine approach

Up to
120
platform TOPS

GPU
Creator & gamer AI

NPU
AI assistants & gen AI

CPU
Light "embedded" AI

Lunar Lake

GPU

AI Engine

Xe2

GPU architecture

XMX

Xe Matrix Extensions

67

peak TOPs

All tops are Int8 on high end SKU, will vary based on SKU

Lunar Lake

CPU

AI Engine

P-core & E-core
CPU architecture

VNNI & AVX
AI instructions

5
peak TOPs

All tops are Int8 on high end SKU, will vary based on SKU

TECH tour.TW

Lunar Lake NPU Deepdive

Darren Crews
Sr Principal Engineer, NPU Lead Architect

Lunar Lake

NPU

AI Engine

NPU 4
Architecture

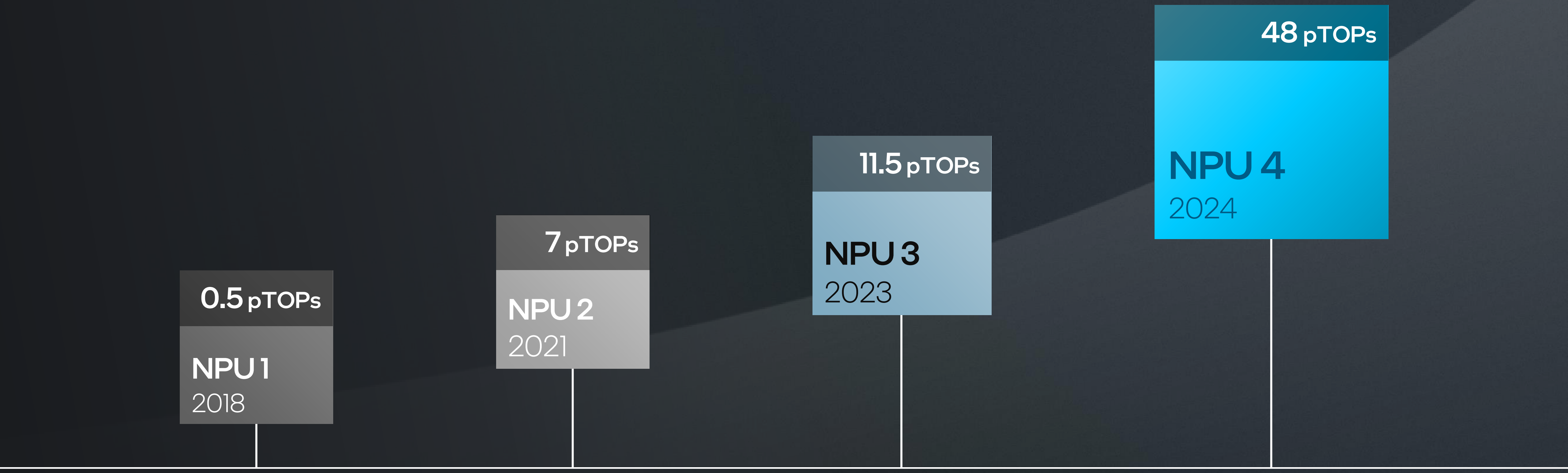
2x
Power
efficiency

48
Peak
TOPs

All tops are Int8 on high end SKU, will vary based on SKU

Continuous NPU Improvements

Across 4 generations of IP



All tops are Int8 on high end SKU, will vary based on SKU

NPU 4

2024

Proven foundations

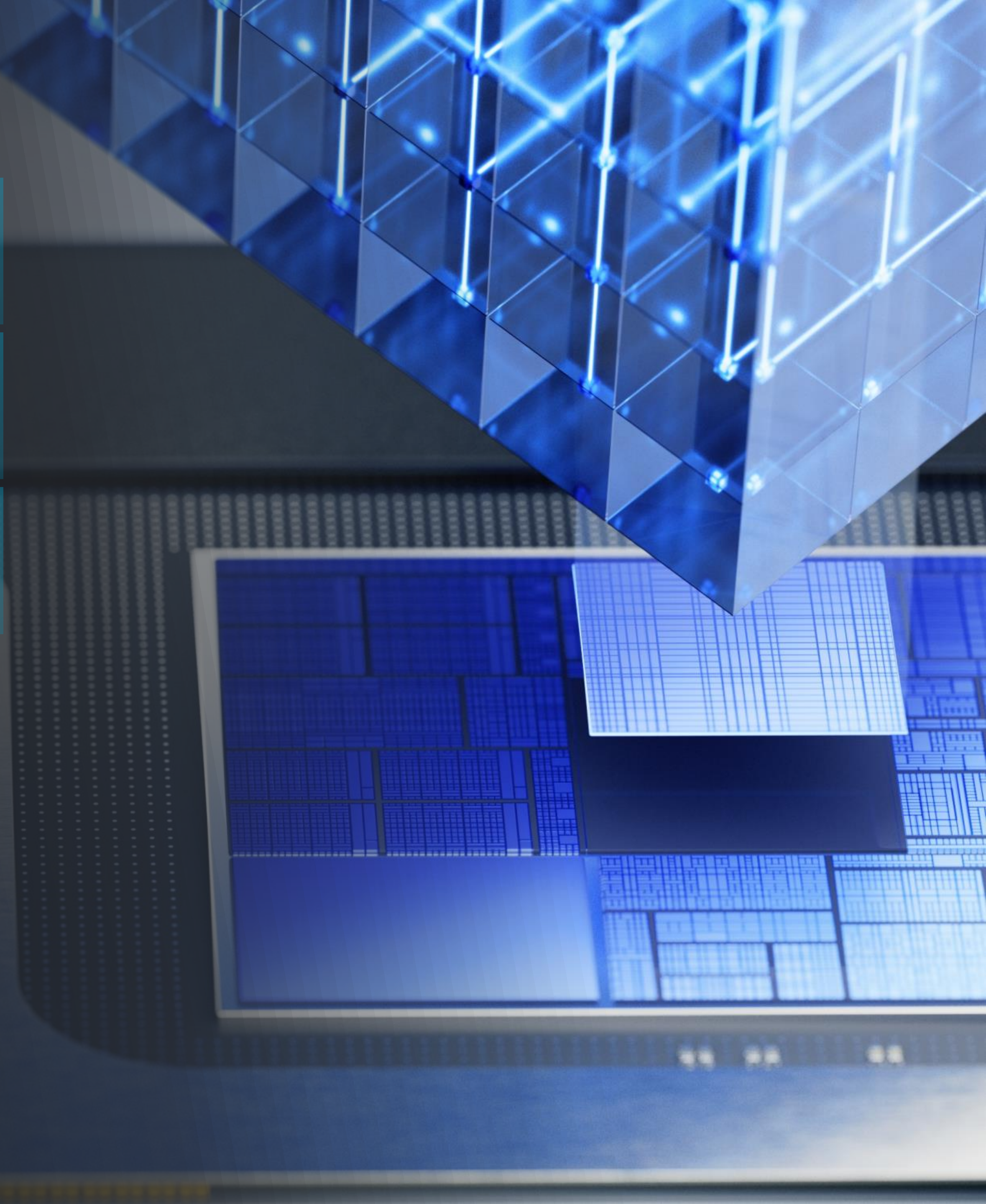
based on three prior generations

Higher compute capacity

to support growing number of use cases

Increased efficiency

to support longer battery life



Scaling the NPU

NPU 3

Increase number of engines

Increase frequency

Improve architecture

NPU 4

What is a TOP?

Trillions of Operations per Second

Operations: 1 Multiply & 1 Accumulate (MAC)

Second: Clock Frequency (Hz)

How Many AI TOPS?

Operations per **Second** ÷ **Trillions**

Operations	Second	Trillions
1 Multiply & 1 Accumulate (MAC)	Clock frequency (Hz)	10^{12}

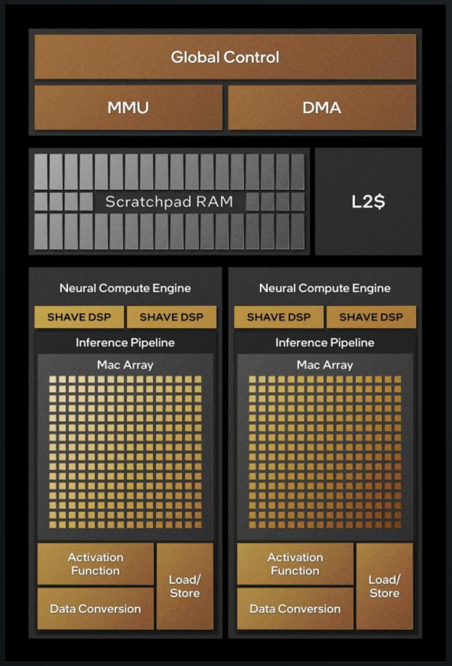
How Many AI TOPS?

Operations × **Frequency** ÷ **Trillions**

Operations: 1 Multiply & 1 Accumulate (MAC)

Trillions: 10^{12}

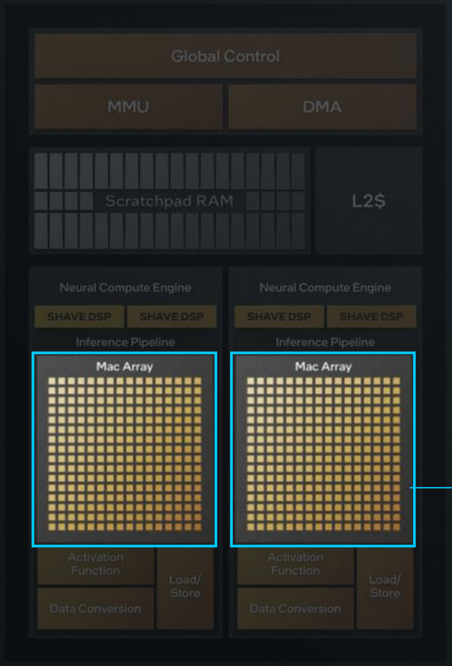
How Many AI TOPS in Meteor Lake's NPU?



Operations × **Frequency** ÷ **Trillions** = **TOPs**

- **Operations**: 1 Multiply & 1 Accumulate (MAC)
- **Frequency**: Clock Frequency (Hz)
- **Trillions**: 10^{12}

How Many AI TOPS in Meteor Lake's NPU?



Operations		Frequency		Trillions		
2 ops/MAC *4096 MAC/cycle	×	1.4 GHz	÷	10^{12}	=	**11.5 TOPs**

intel. TECH tour.TW

Operation Types
Overview

	Scalar	Vector	Matrix
Complexity	1	N	N^2
Example functions	Conditional Looping	SoftMax Activation functions	Convolution Matrix multiplication
Occurrence in AI	Low	Very high	Very high

Operation Types

Overview

	Scalar	Vector	Matrix
Complexity	1	N	N^2
Example functions	Conditional Looping	SoftMax Activation functions	Convolution Matrix multiplication
Occurrence in AI	Low	Very high	Very high

TOPs

Scaling the NPU

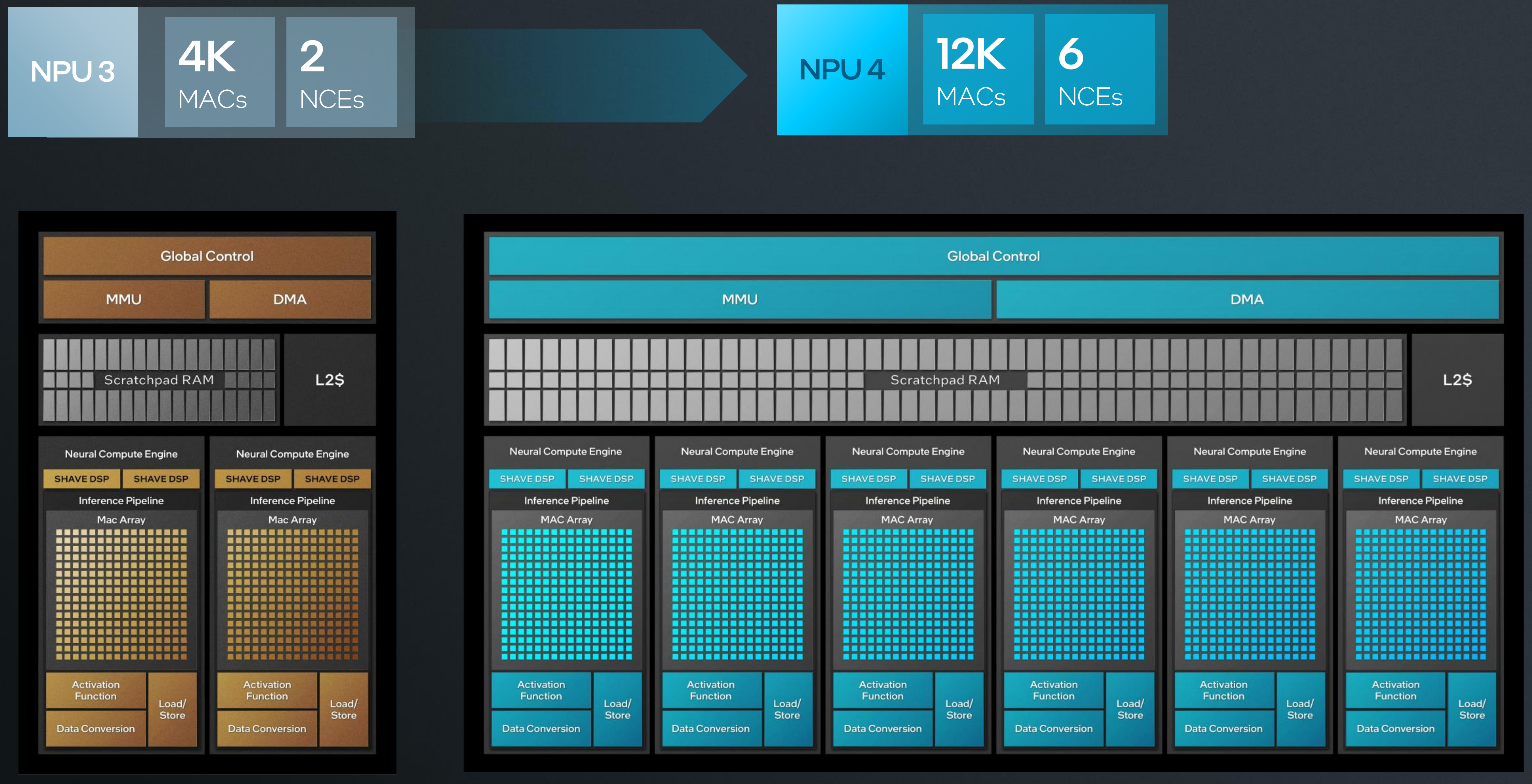
Increase number of engines

Increase frequency

Improve architecture

NPU 3

NPU 4



NPU 3
4K MACs
2 NCEs
NPU 4
12K MACs
6 NCEs
Global Control
MMU
DMA
Scratchpad RAM
L2$
Neural Compute Engine
SHAVE DSP
SHAVE DSP
Inference Pipeline
Mac Array
Activation Function
Load/ Store
Data Conversion
Neural Compute Engine
SHAVE DSP
SHAVE DSP
Inference Pipeline
Mac Array
Activation Function
Load/ Store
Data Conversion
Global Control
MMU
DMA
Scratchpad RAM
L2$
Neural Compute Engine
SHAVE DSP
SHAVE DSP
Inference Pipeline
MAC Array
Activation Function
Load/ Store
Data Conversion

Scaling the NPU

NPU 3

Increase number of engines

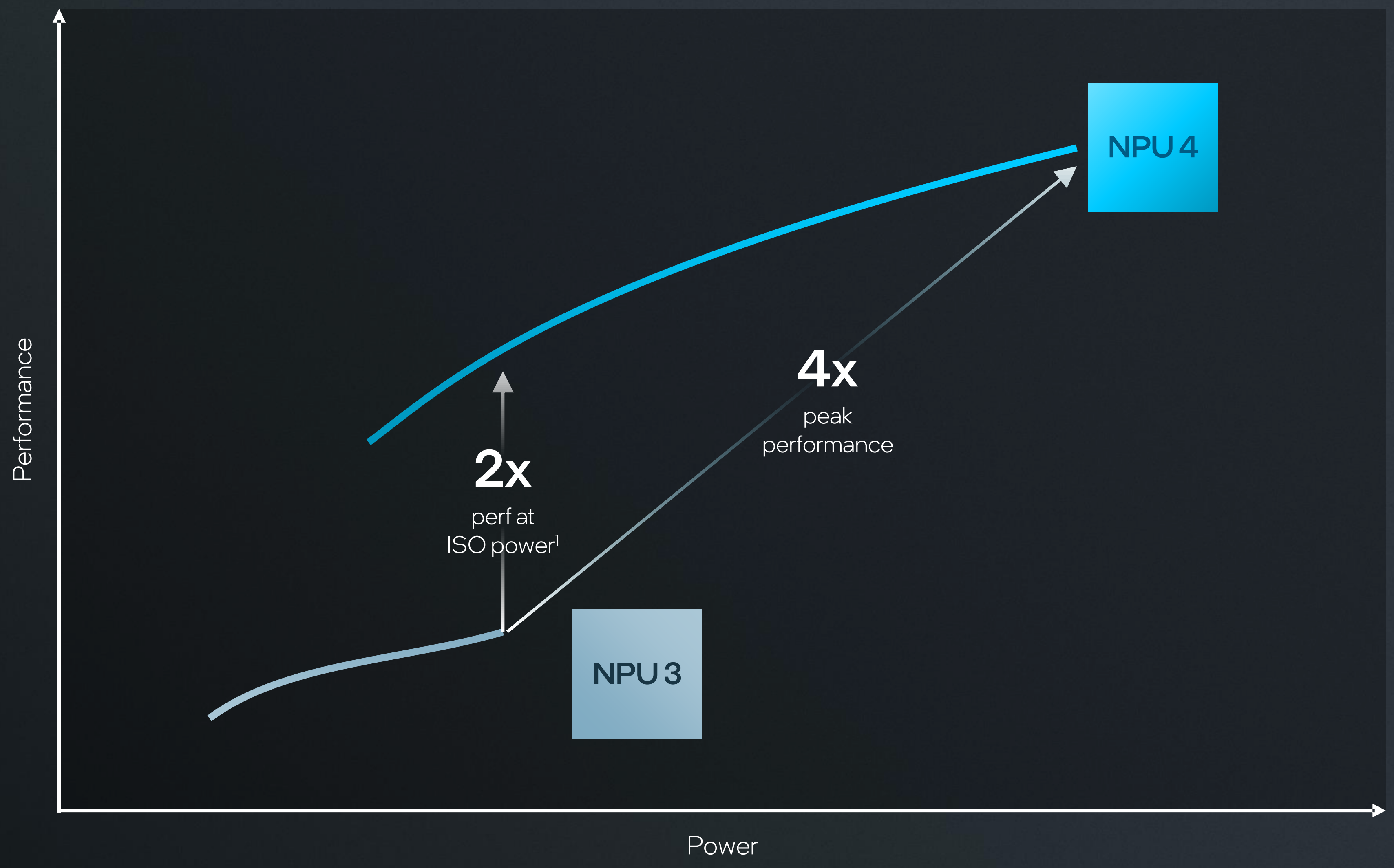
Increase frequency

Improve architecture

NPU 4

Increased Efficiency & Increased Performance

- Increased clock
- New node
- Architecture improvements



[1] Based on pre-production simulation data of a real network. See backup for details.

Scaling the NPU

NPU 3

Increase number of engines

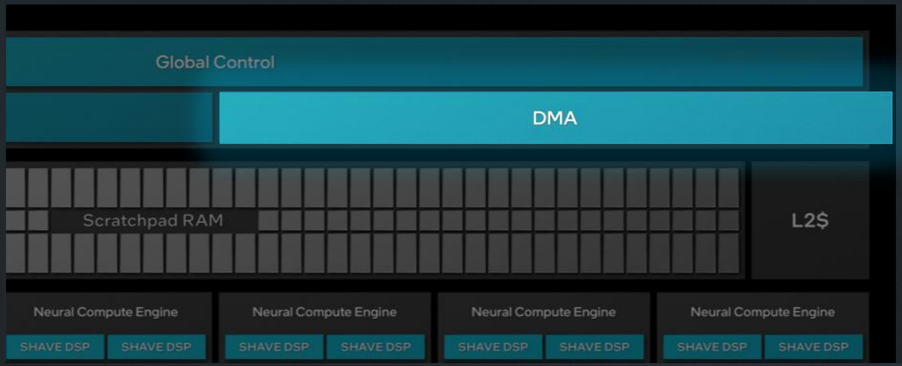
Increase frequency

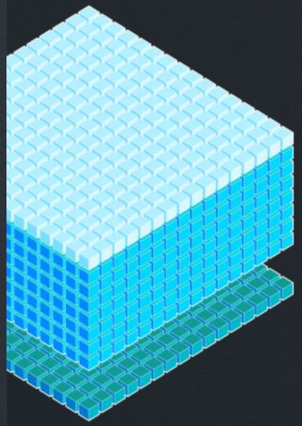
Improve architecture

NPU 4

NPU 4
Architecture improvements

NPU bandwidth





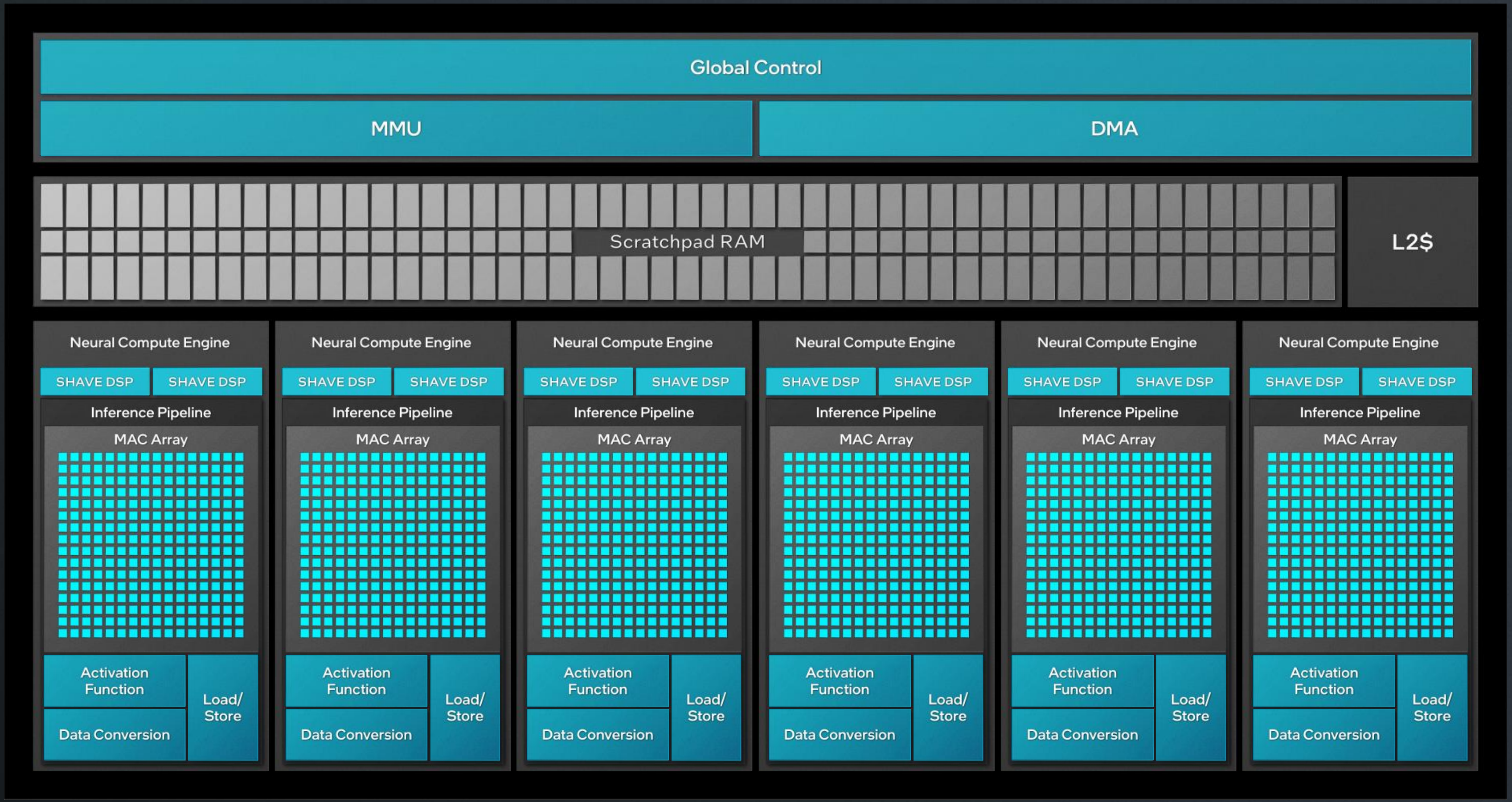
Efficiency of matrix compute

Vector performance

Increased number of tiles

NPU 4

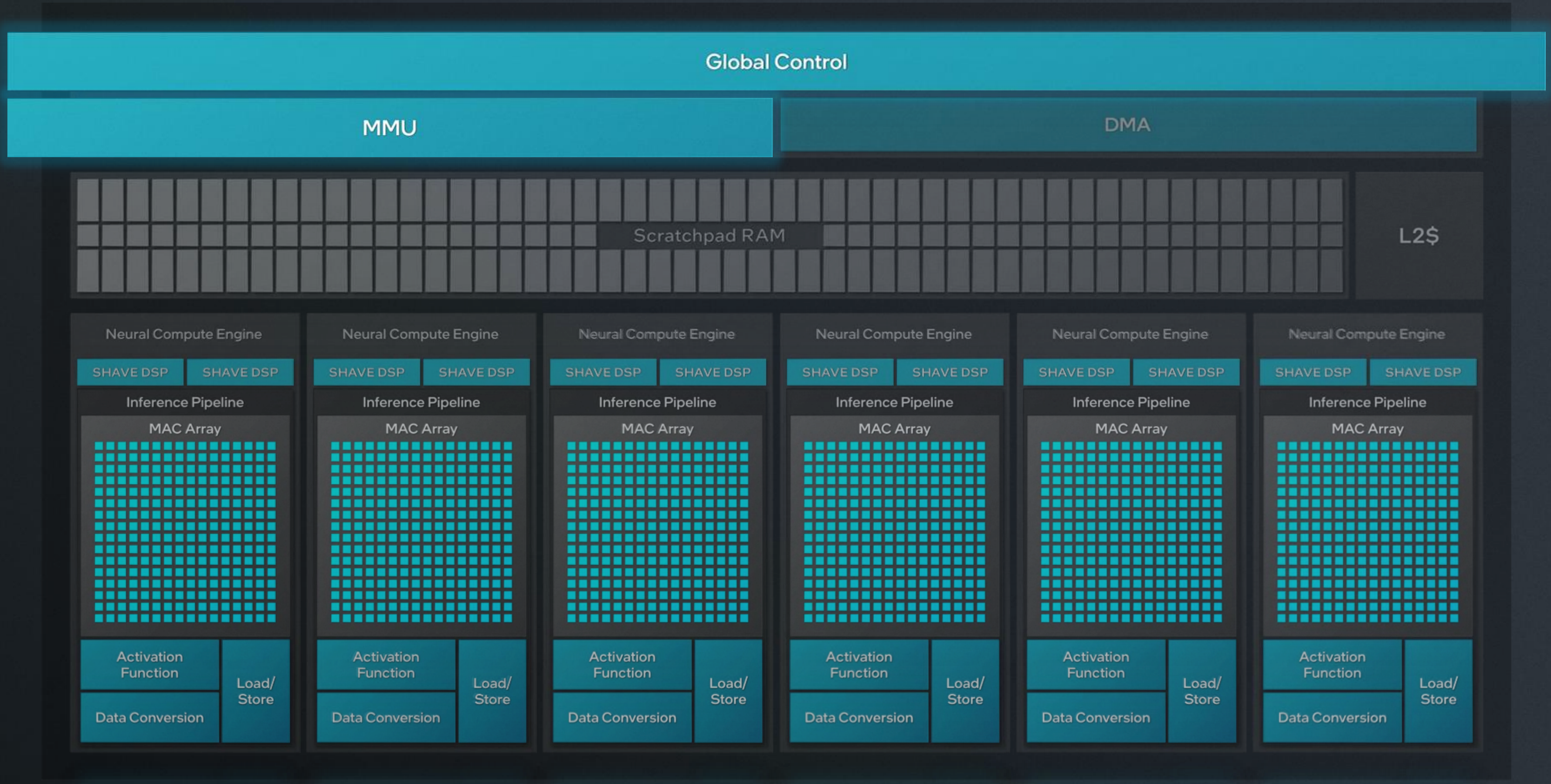
Architecture overview



NPU 4

Architecture overview

Global control & MMU

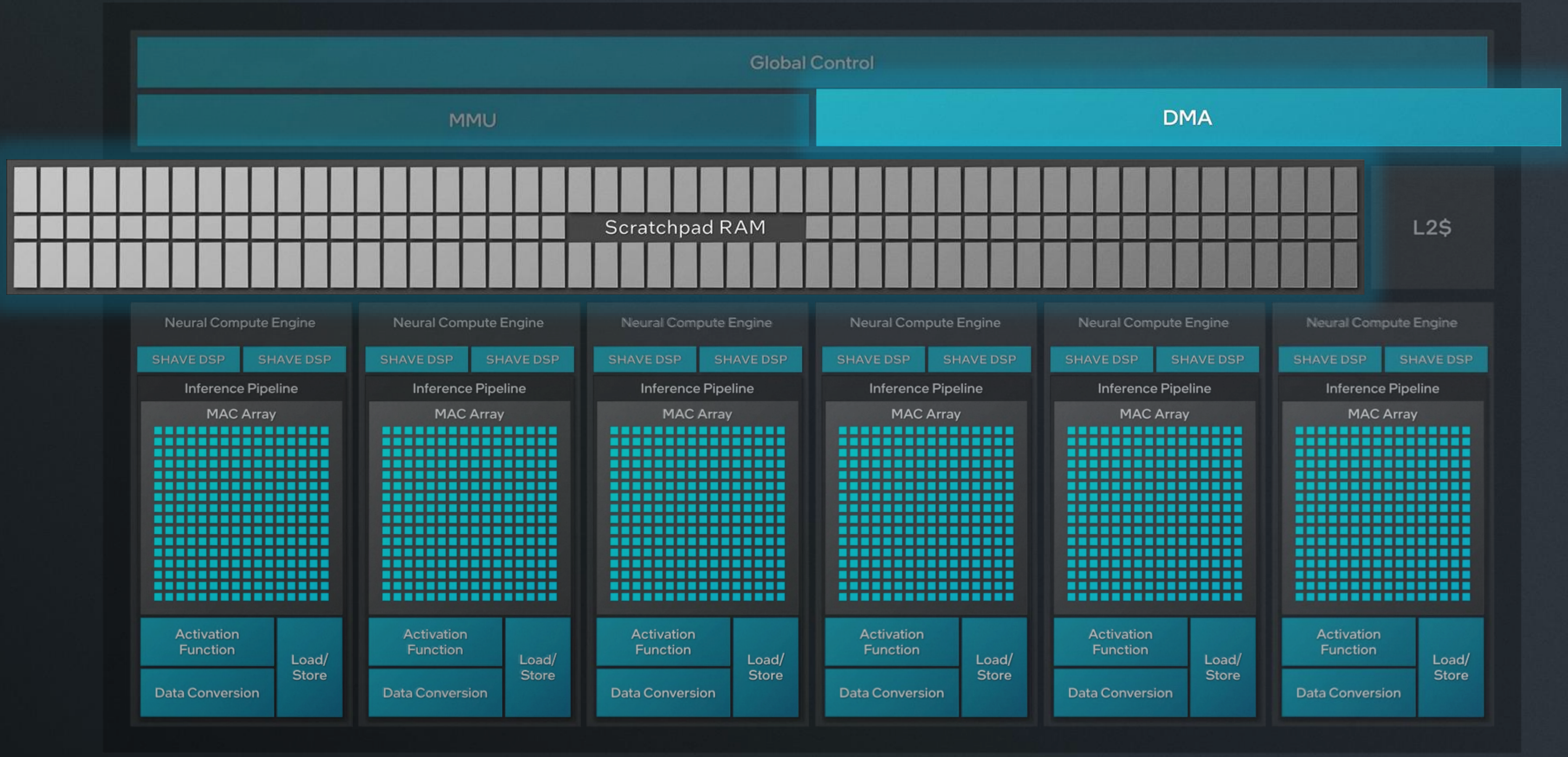


NPU 4

Architecture overview

Global control & MMU

DMA & scratchpad RAM



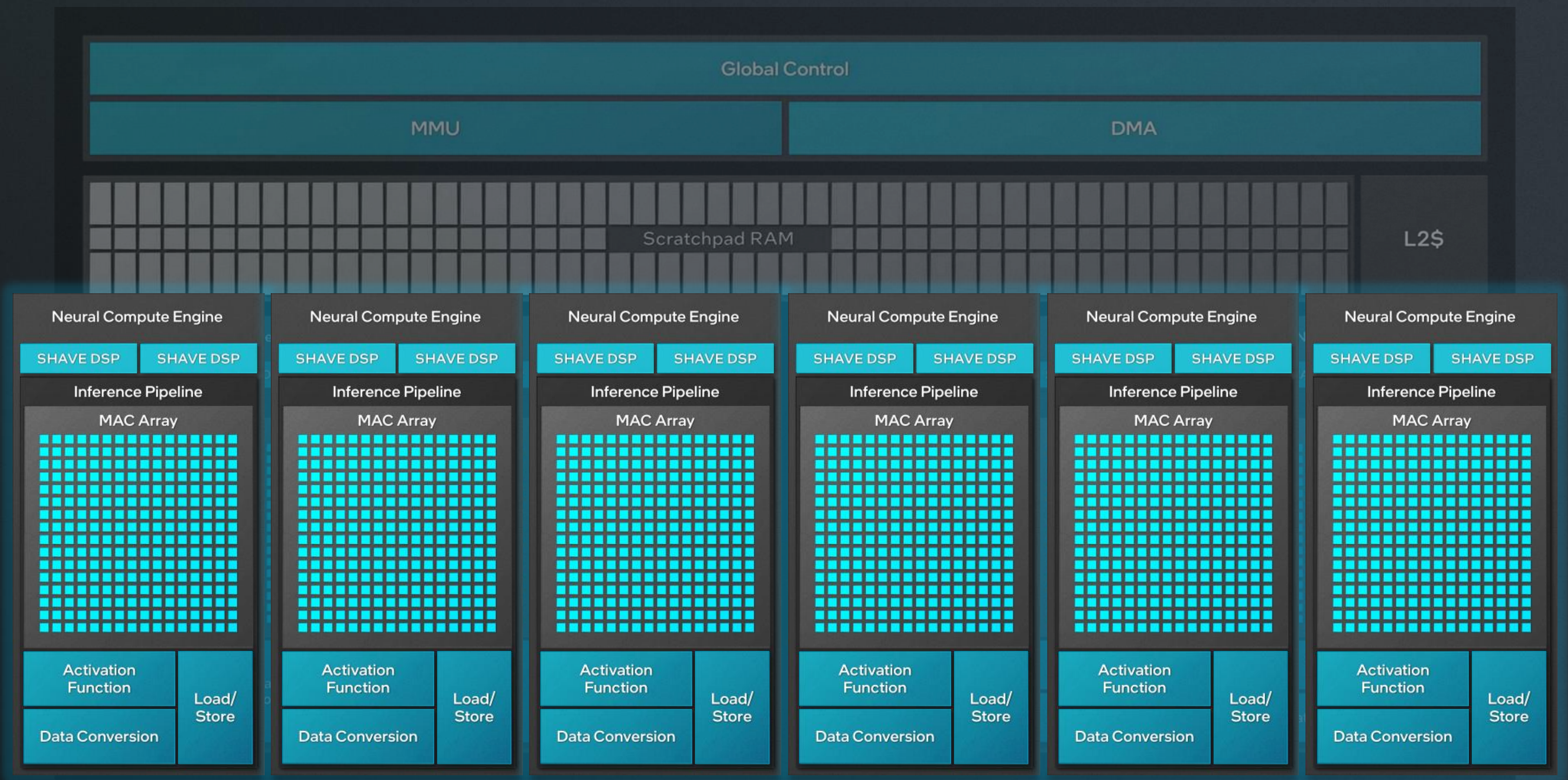
NPU 4

Architecture overview

Global control & MMU

DMA & scratchpad RAM

Neural compute engines



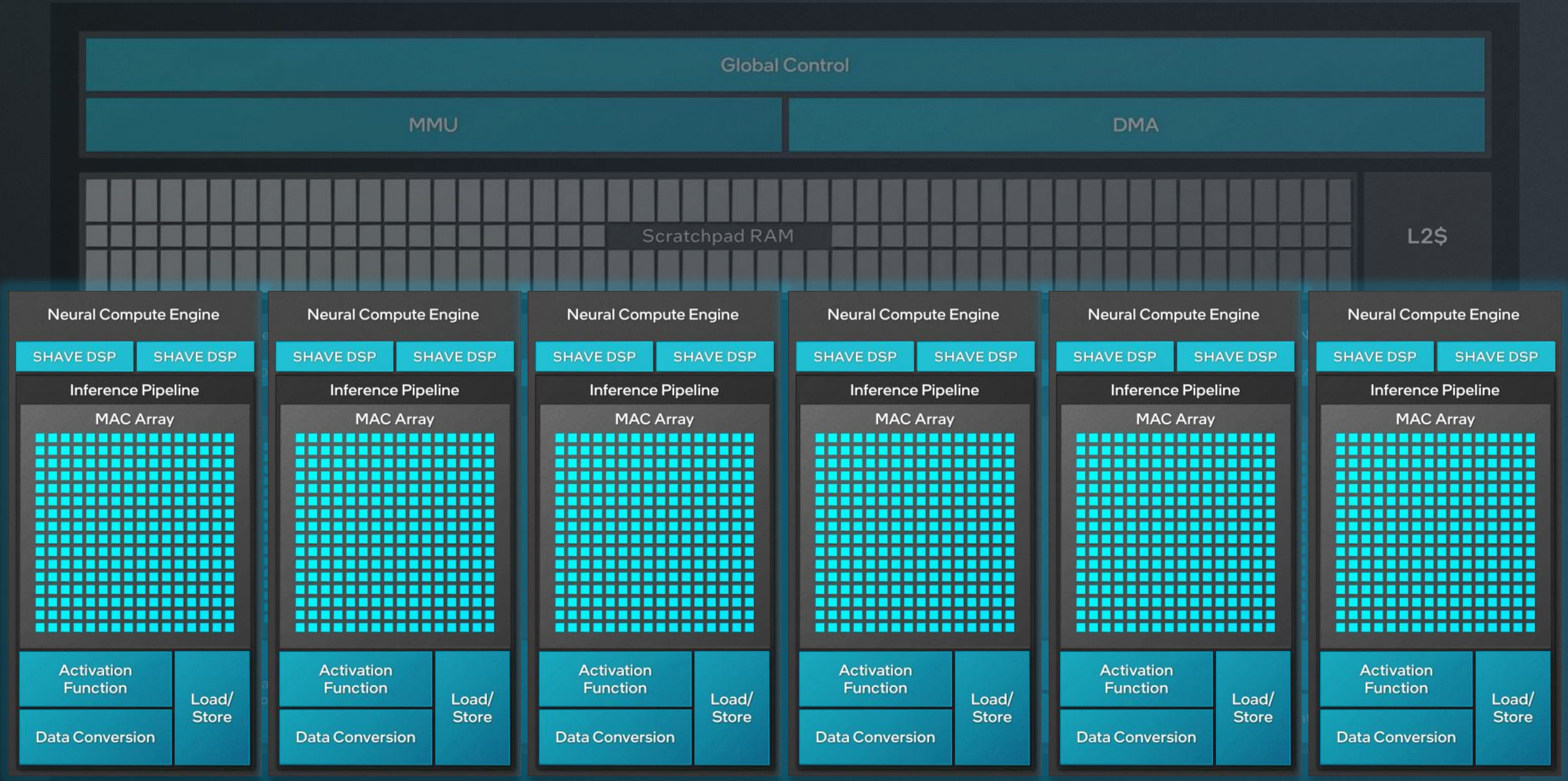
NPU 4

Neural compute engine

Specialized engines
Matrix + Vector

Inference pipeline
MAC arrays + fixed function

Programmable DSPs



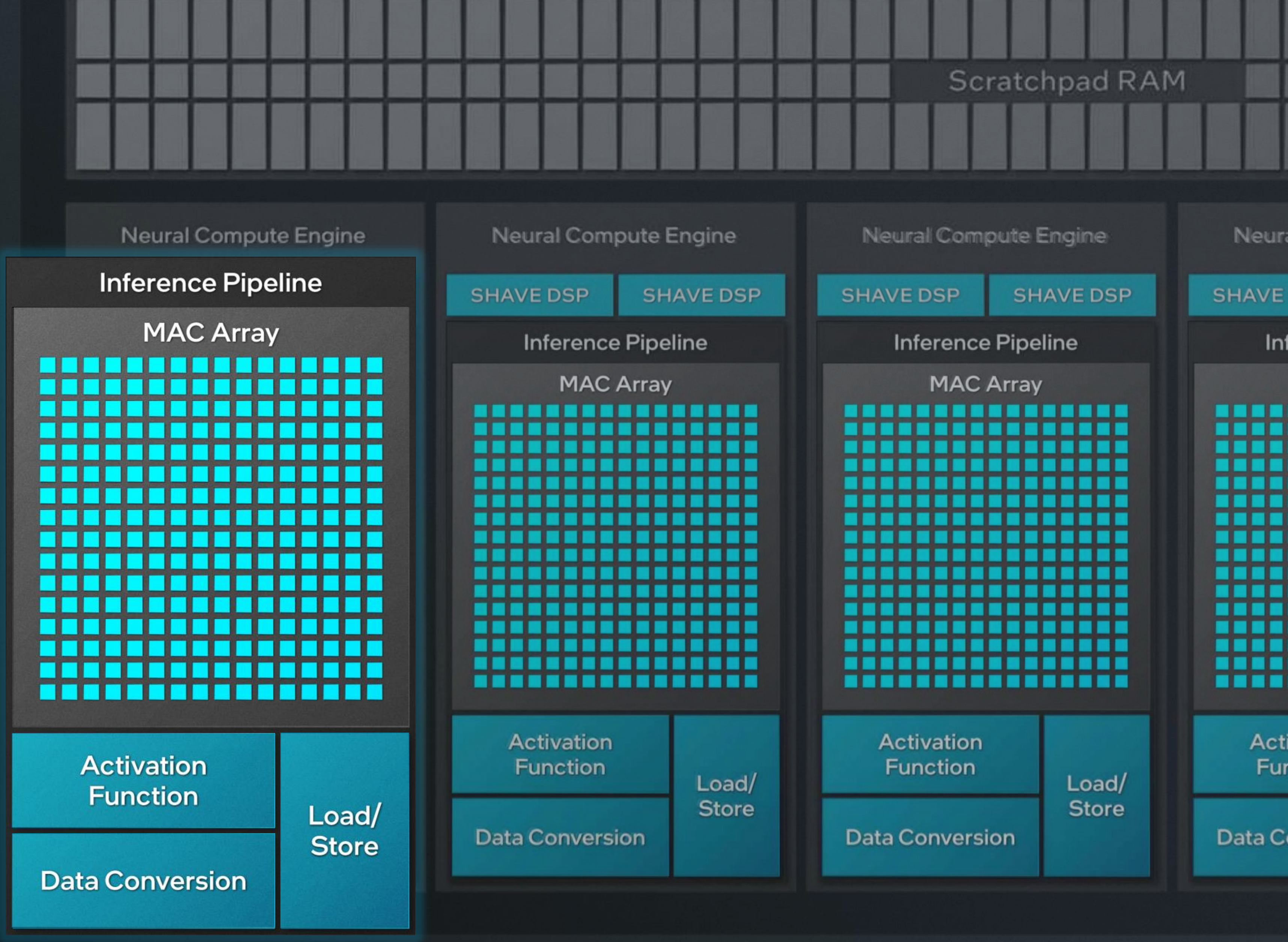
NPU 4

Inference pipeline

Efficient matrix multiplication

Activation function support

Data conversion and re-layout support

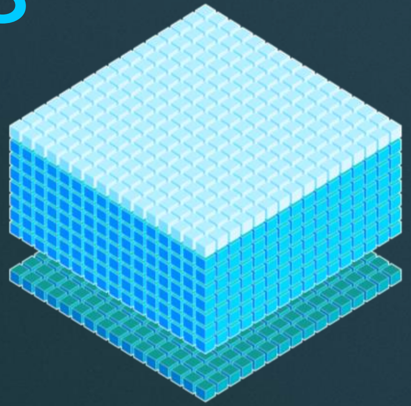


NPU 4

MAC array

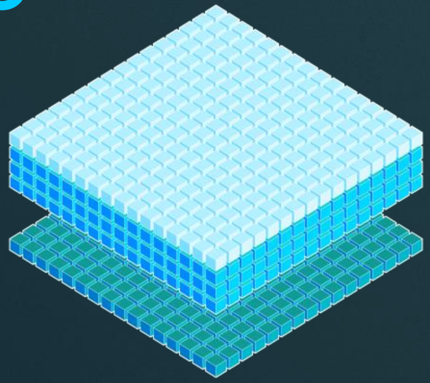
INT8

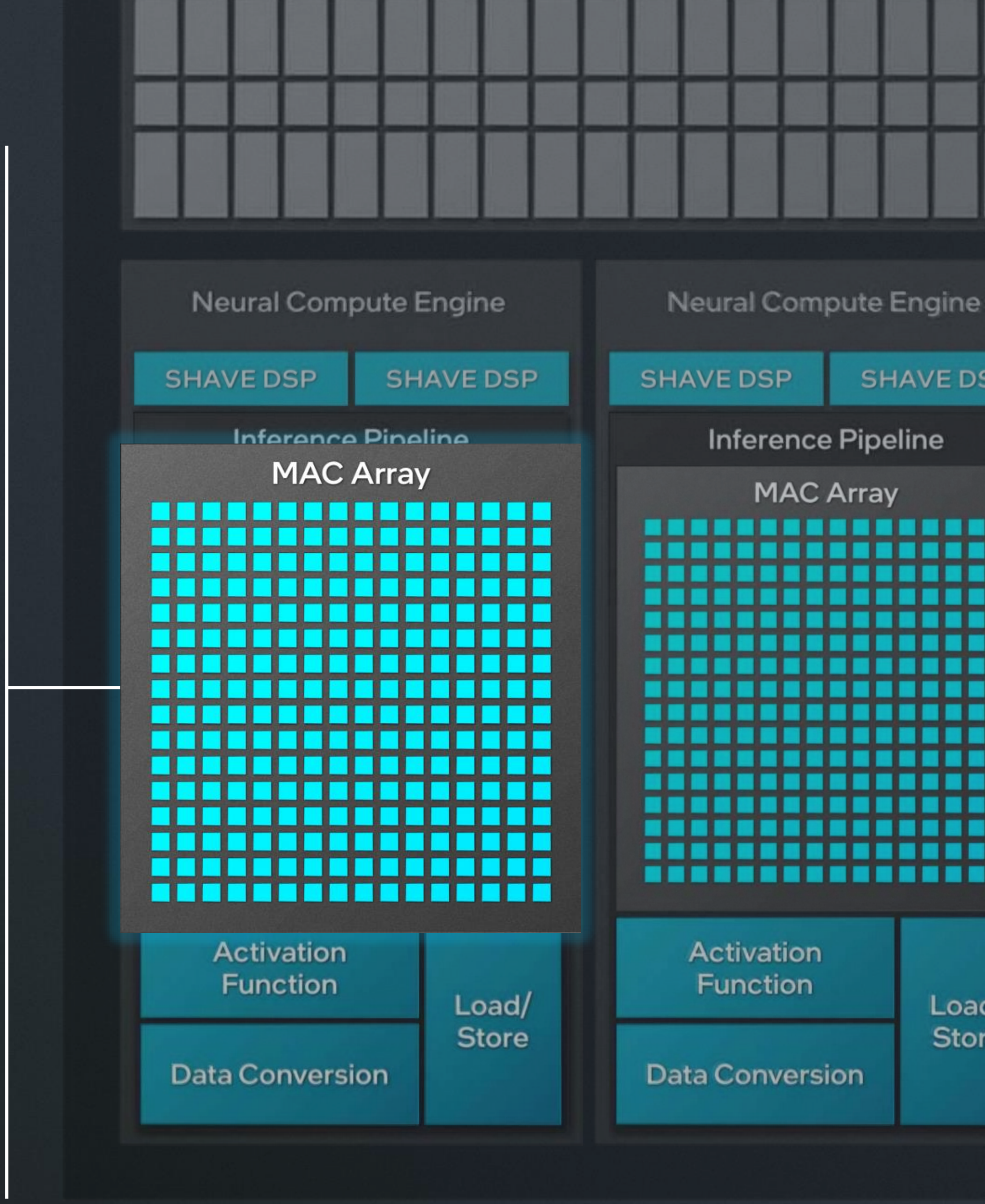
16x16x8



FP16

16x16x4





Matrix multiplication & convolution

2048 MAC/cycle int8
1024 MAC/cycle FP16

Up to 2x[1] efficiency
driving better perf/watt

[1]See details in backup

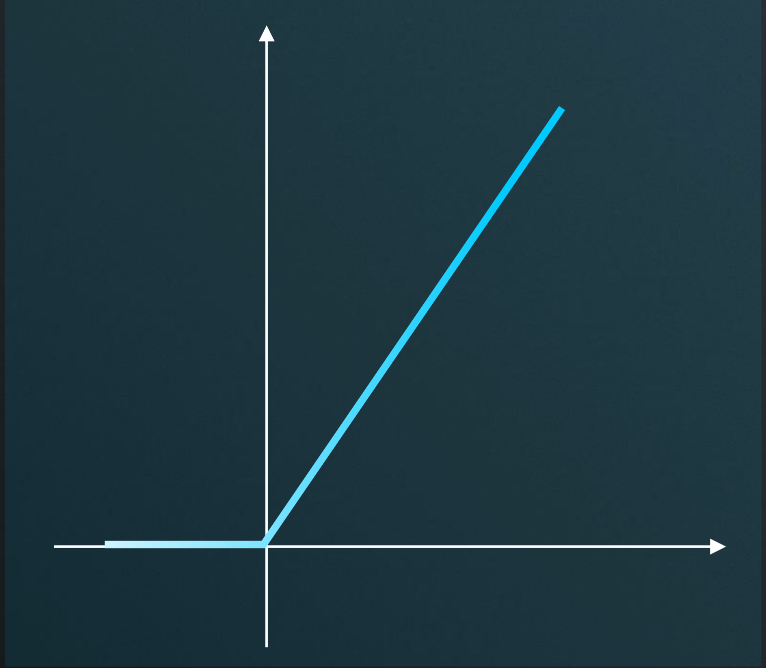
NPU 4

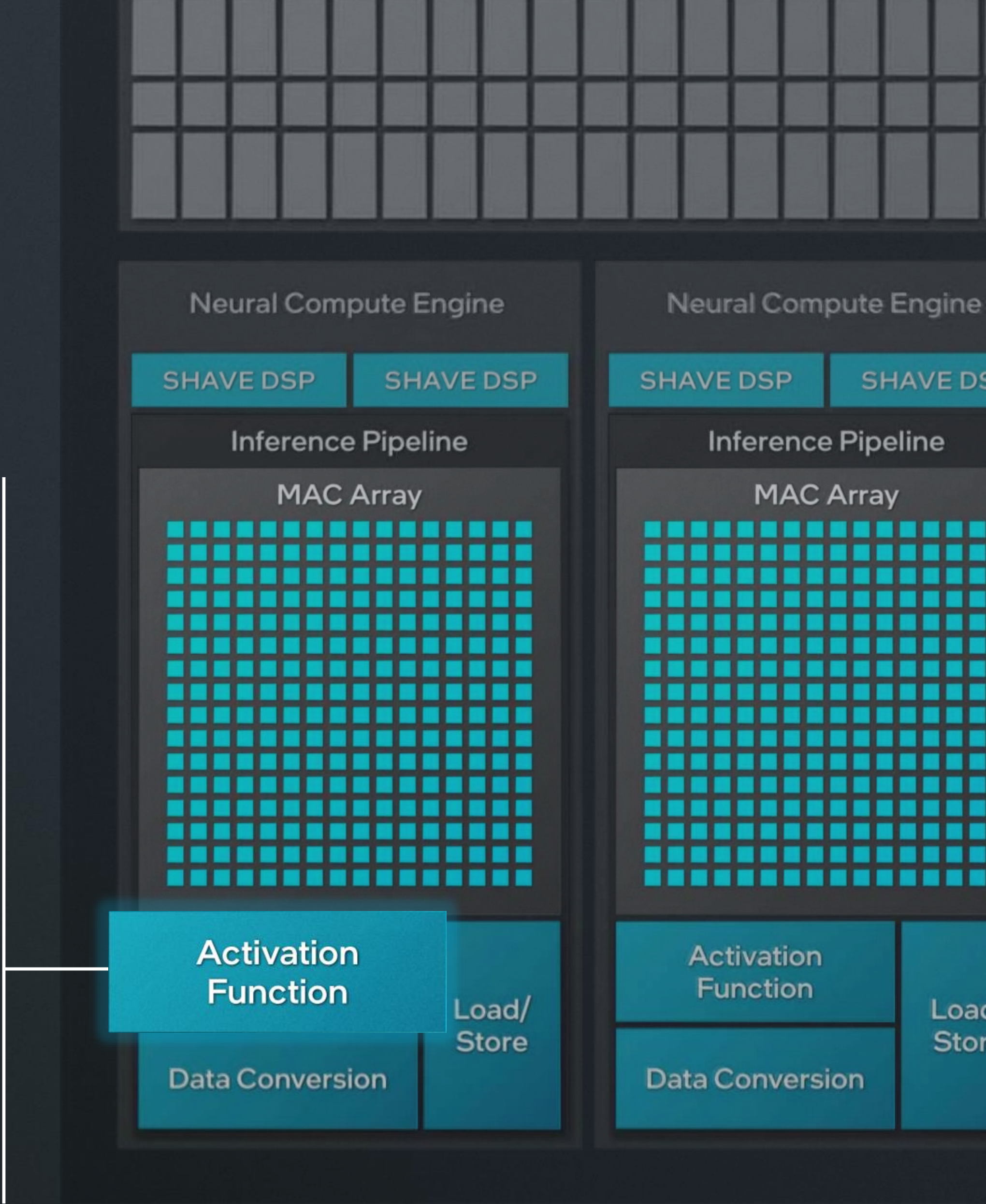
Activation functions

Multiple functions
Supported

FP precision
Support

$$\mathrm{ReLU}(x) \triangleq \max(0, x)$$





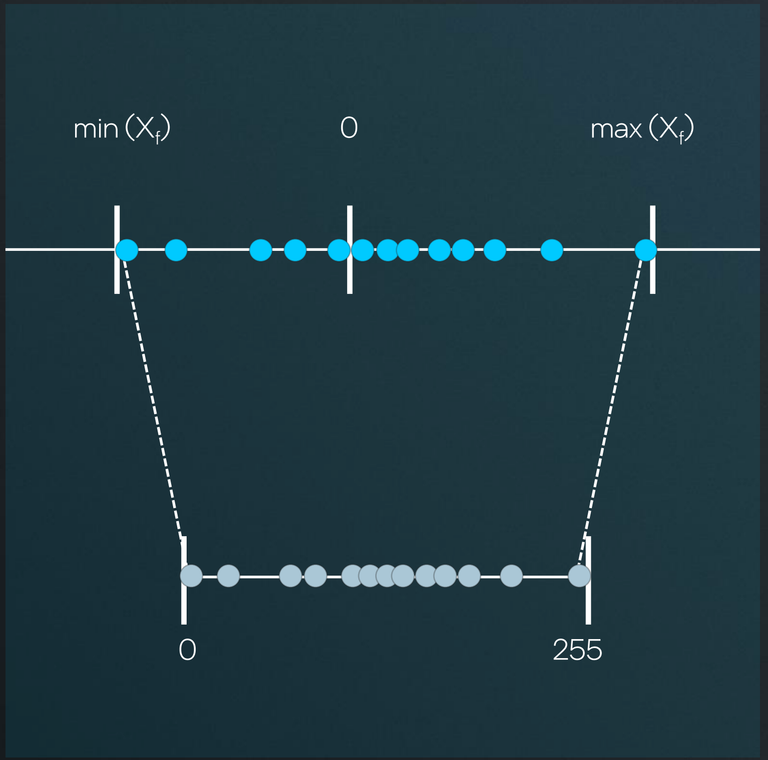
NPU 4

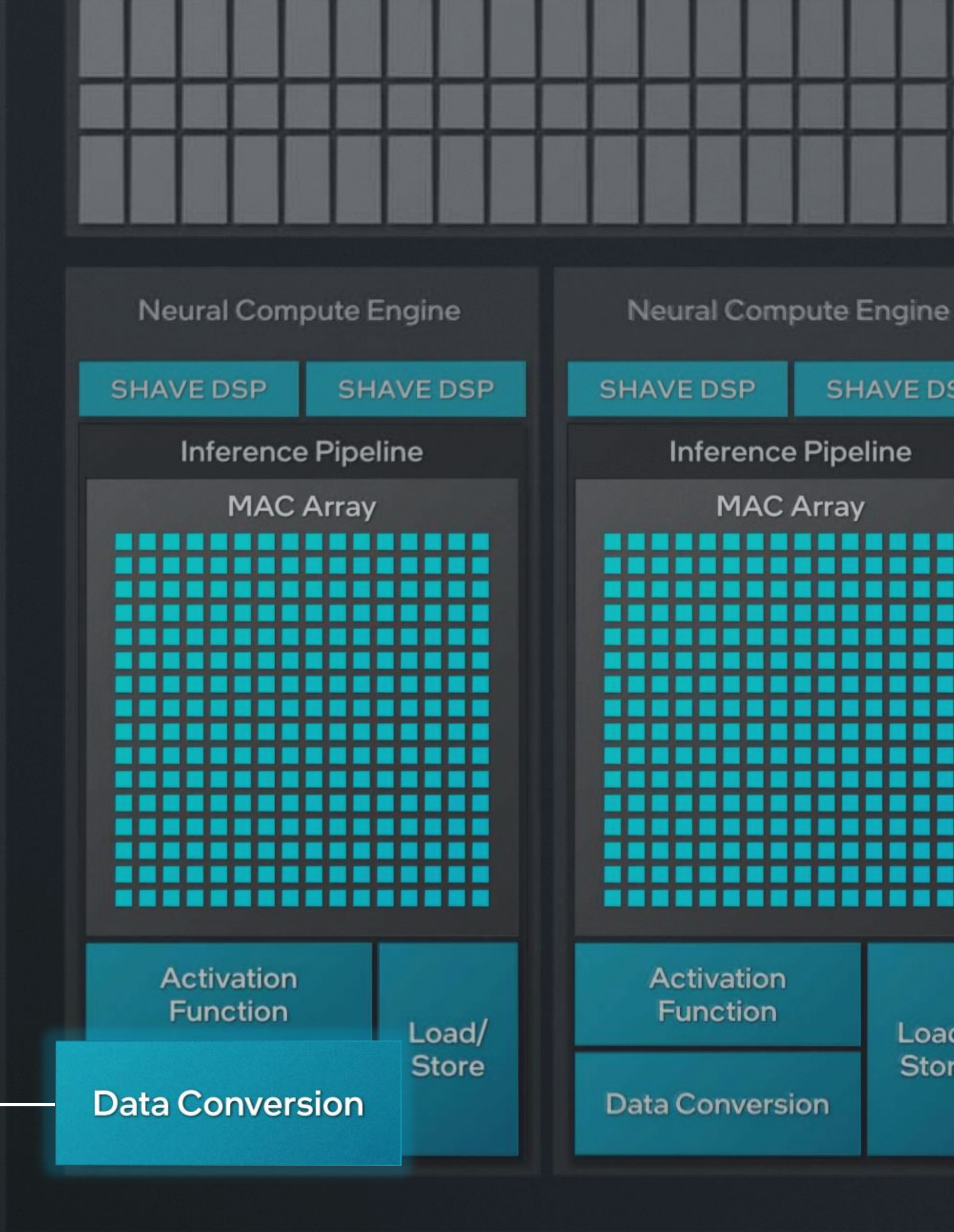
Data conversion

Datatype conversion

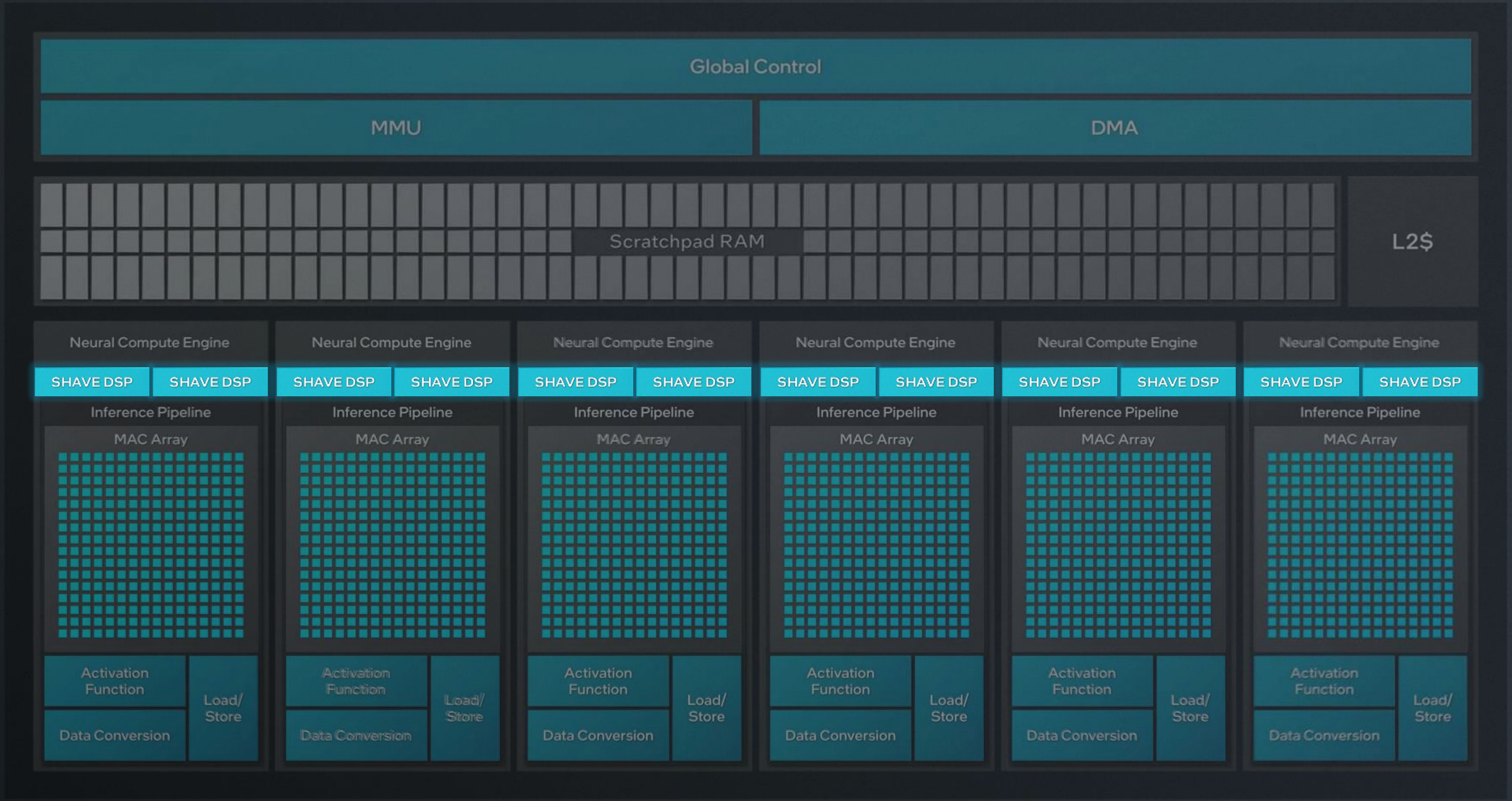
Fused operations

Output data re-layout





NPU 4
SHAVE DSP



Upgraded SHAVE DSP
4x vector compute

12x overall vector perf
improves transformer /LLM performance

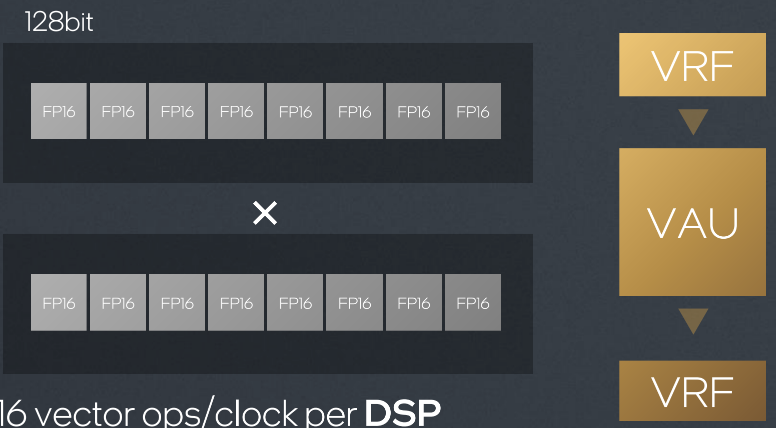

NPU 4
SHAVE DSP
Data Cache
L1 I-Cache
PEU
IRF
LSU
VRF
VAU
IAU
SAU
CMU
BRU
DCU
TRF
ICU
IDU
512-bit
Vector register file size
4x
Performance
Vector unit
4x
Bandwidth
to and from SHAVE DSP
NPU 3
SHAVE DSP
Data Cache
L1 I-Cache
PEU
IRF
LSU
VRF
VAU
IAU
SAU
CMU
BRU
DCU
TRF
ICU
IDU
Global Control
MMU
DMA
Scratchpad RAM
L2$
Neural Compute Engine
SHAVE DSP
Inference Pipeline
Mac Array
Activation Function
Load/ Store
Data Conversion
MAC Array
intel. TECH tour.TW

SHAVE DSP

Vector increase

NPU3 SHAVE DSP

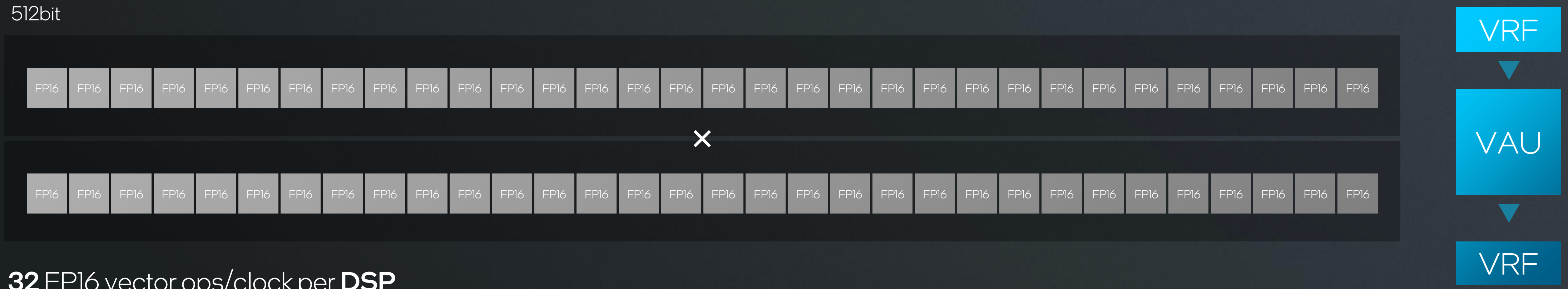
FP16



8 FP16 vector ops/clock per **DSP**

NPU4 SHAVE DSP

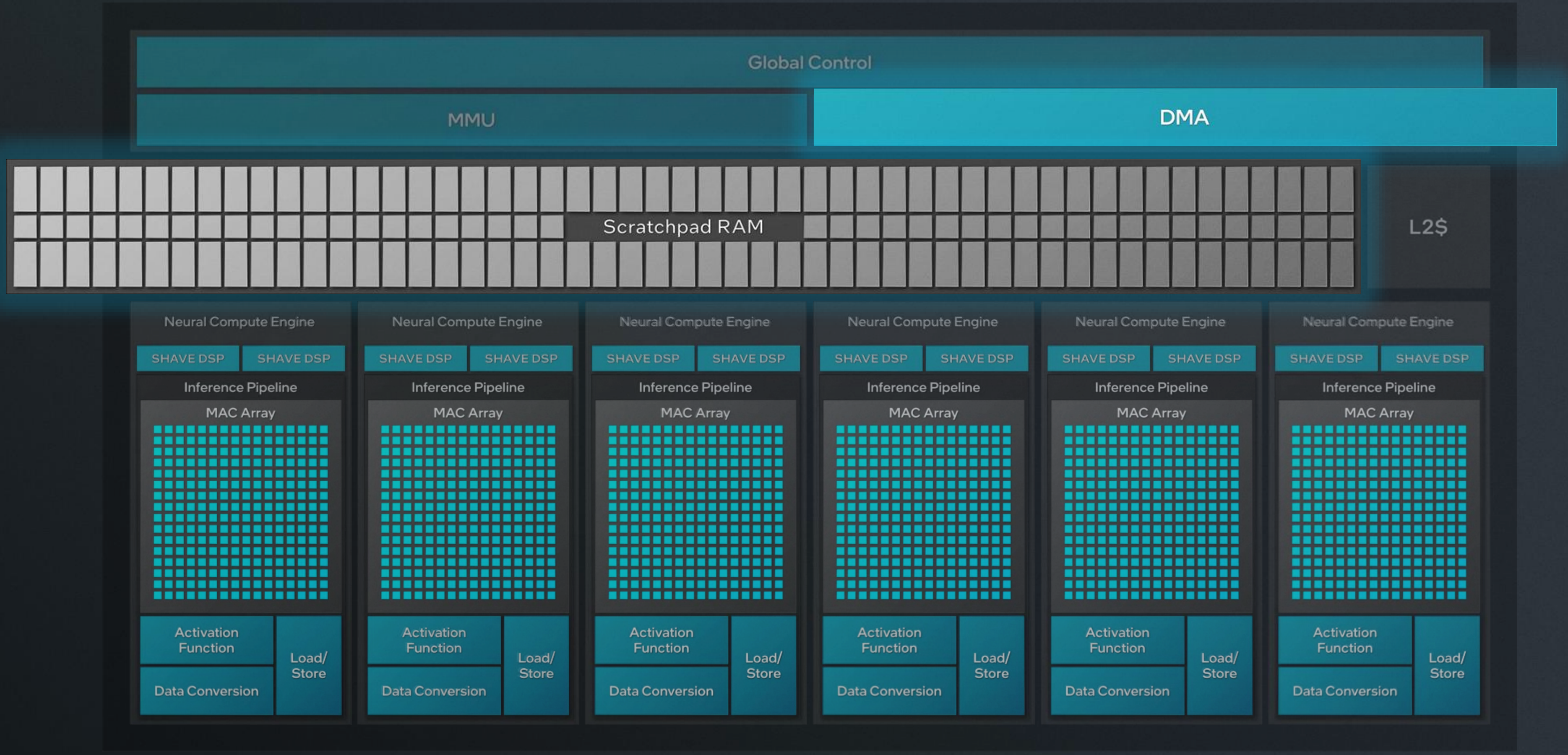
FP16



32 FP16 vector ops/clock per **DSP**

NPU 4

DMA engine



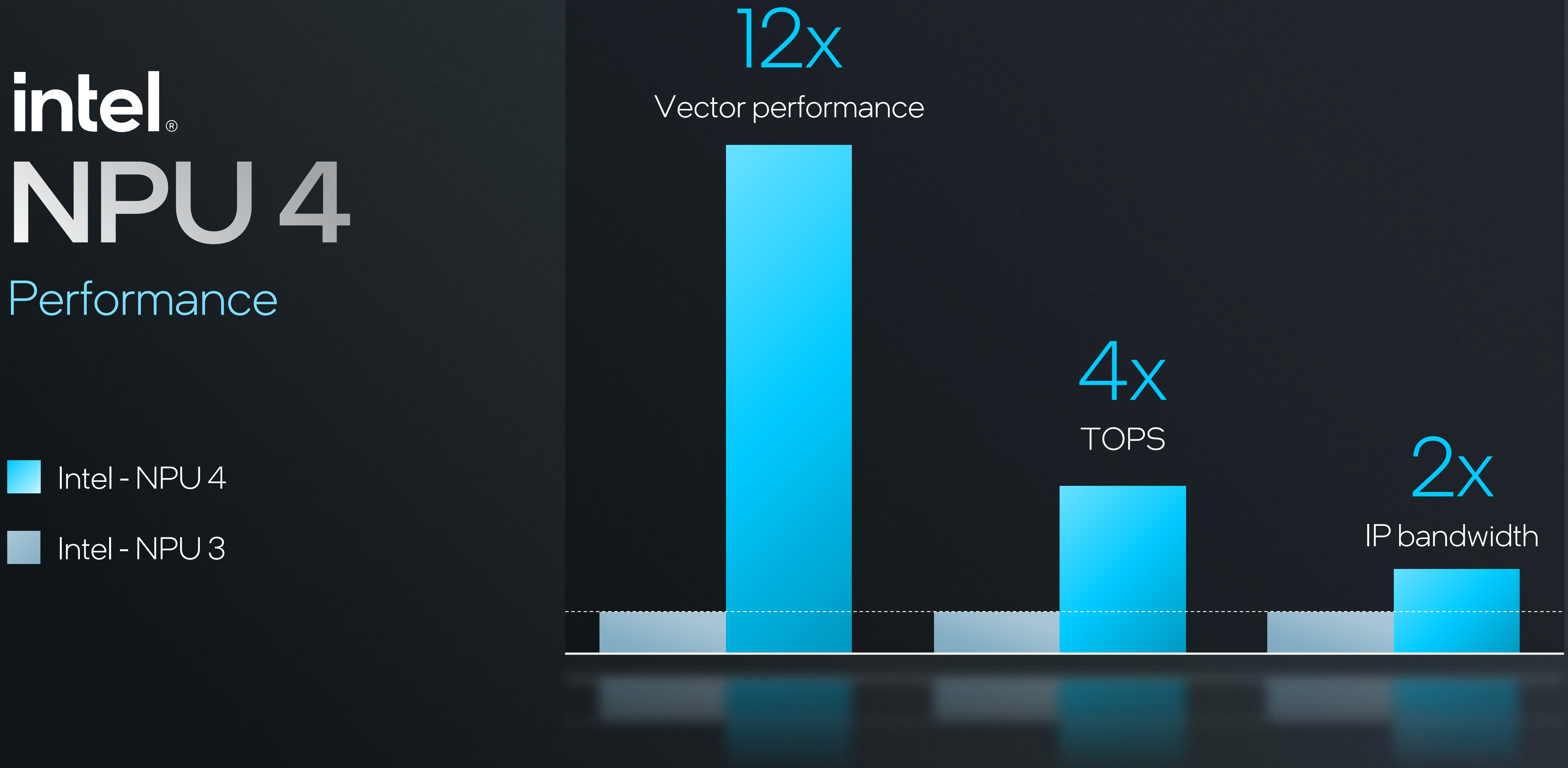
2x DMA bandwidth
improves network performance especially LLMs

New functions
Embedding tokenization

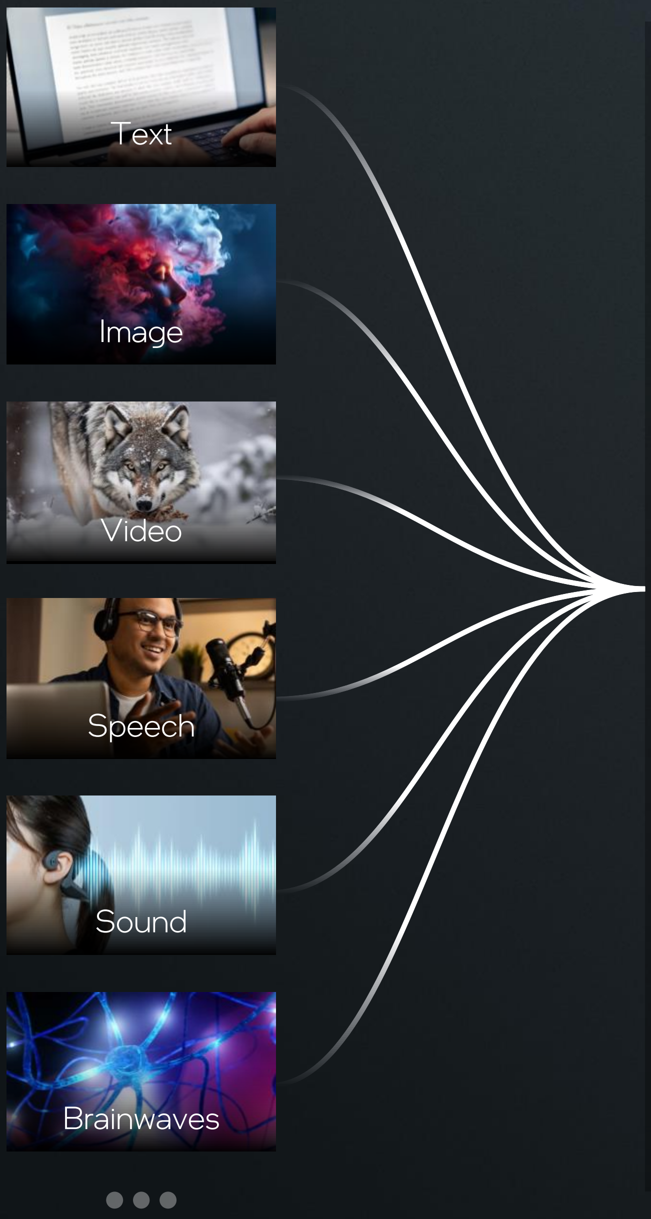
See backup for details.

intel® NPU 4

Performance

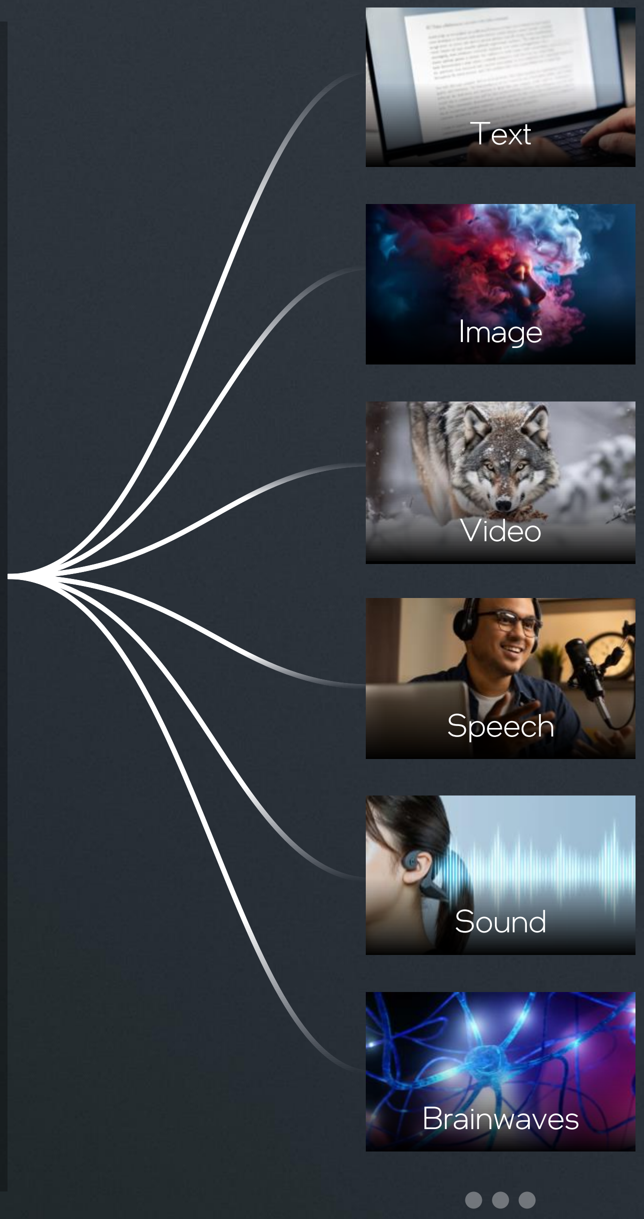


See backup for details.

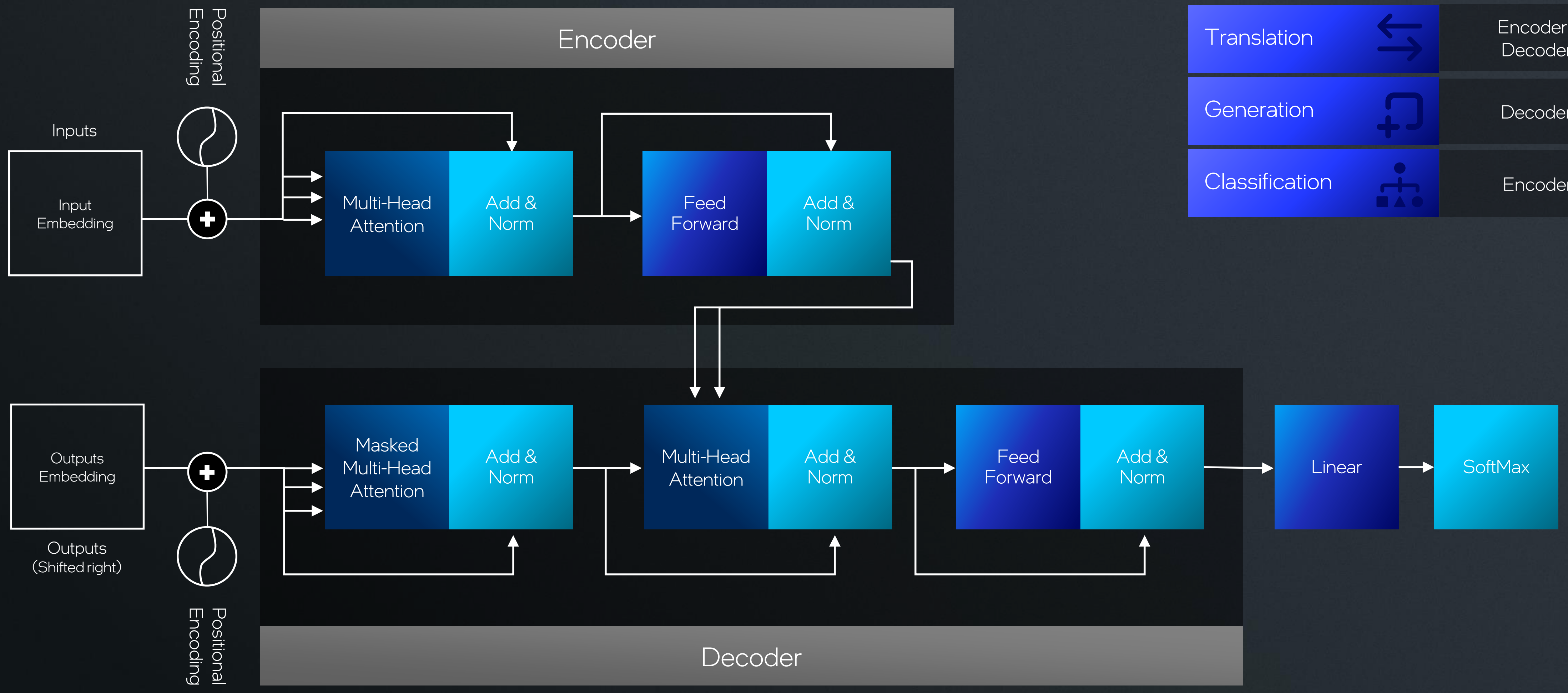


Transformer Use Cases

- Translation
- Generation
- Classification

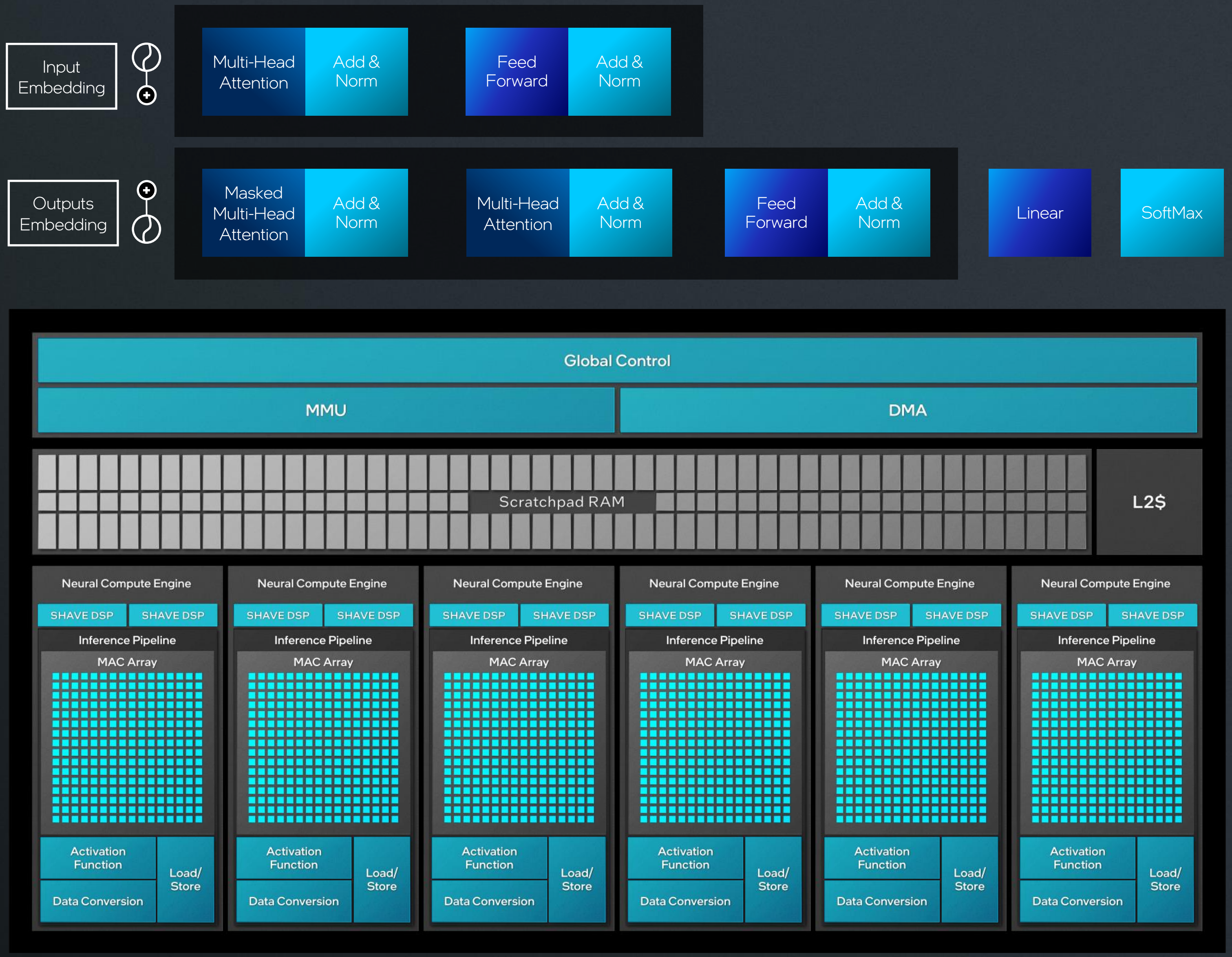


Transformer Model Architecture

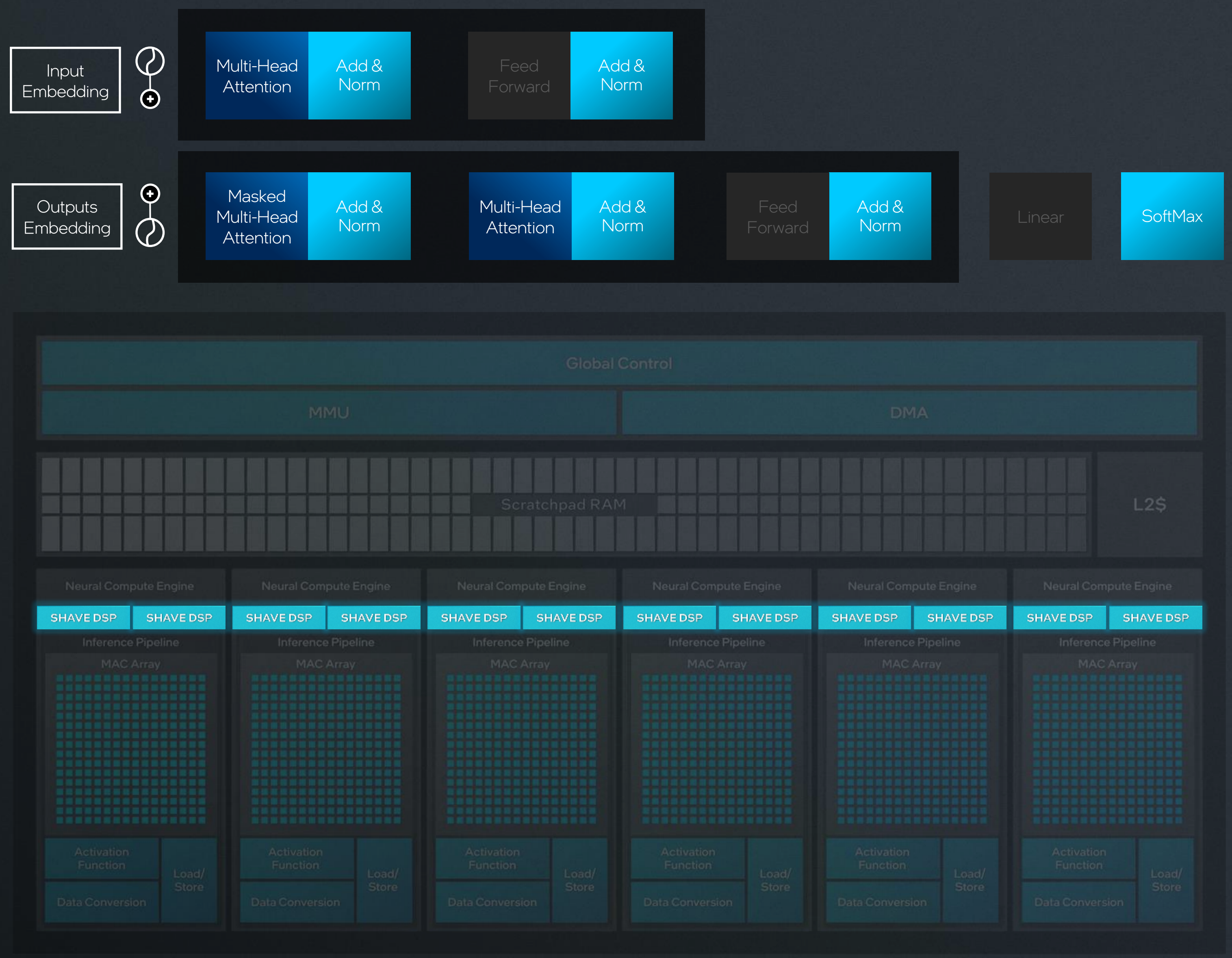


Translation	Encoder + Decoder
Generation	Decoder
Classification	Encoder

Transformer Architecture on Intel's NPU



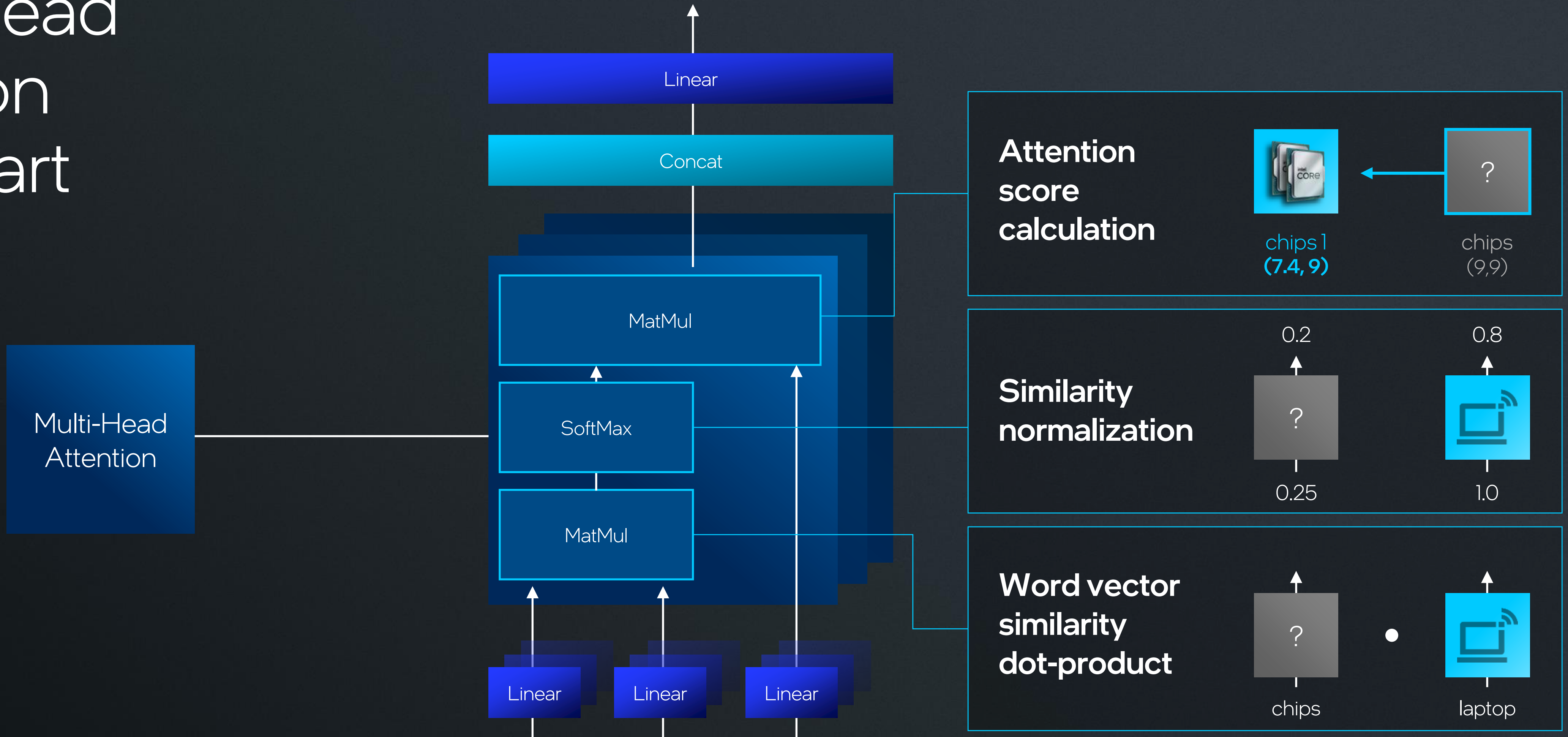
Transformer Architecture on Intel's NPU



Transformer Architecture on Intel's NPU

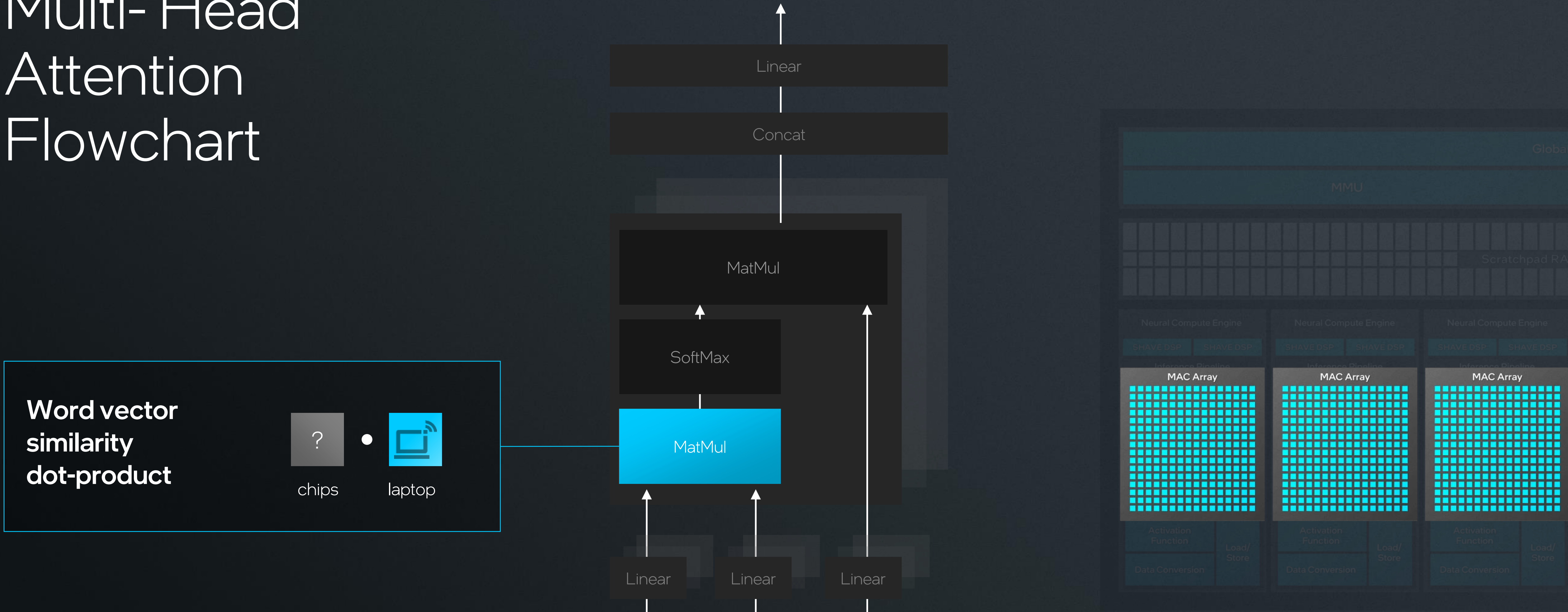
intel. TECH tour.TW

Multi- Head Attention Flowchart



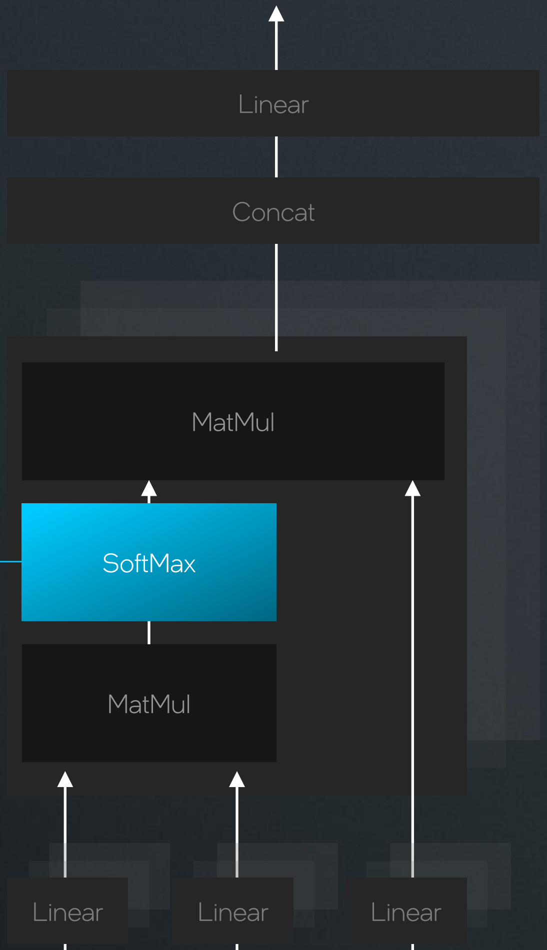
$$Attention(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi- Head Attention Flowchart

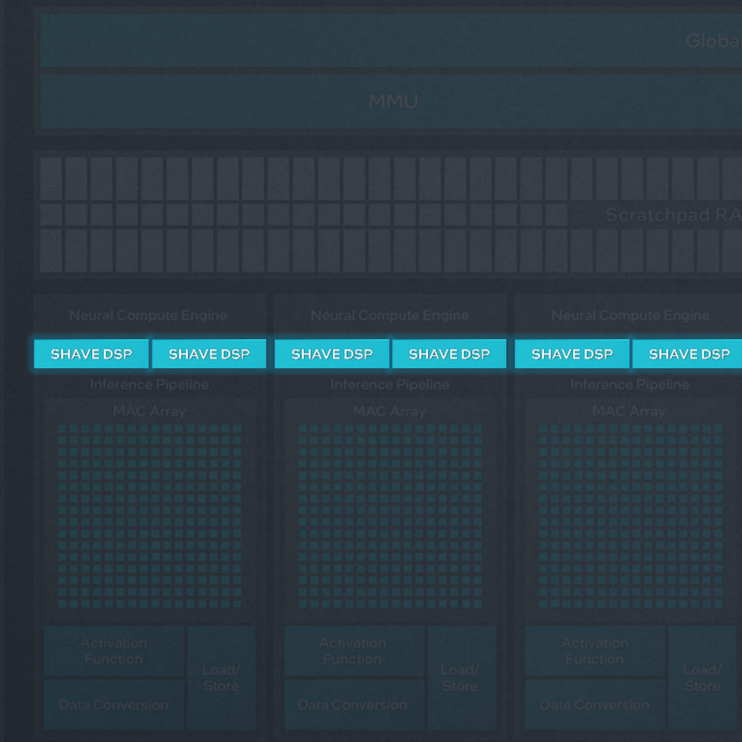


$$Attention(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

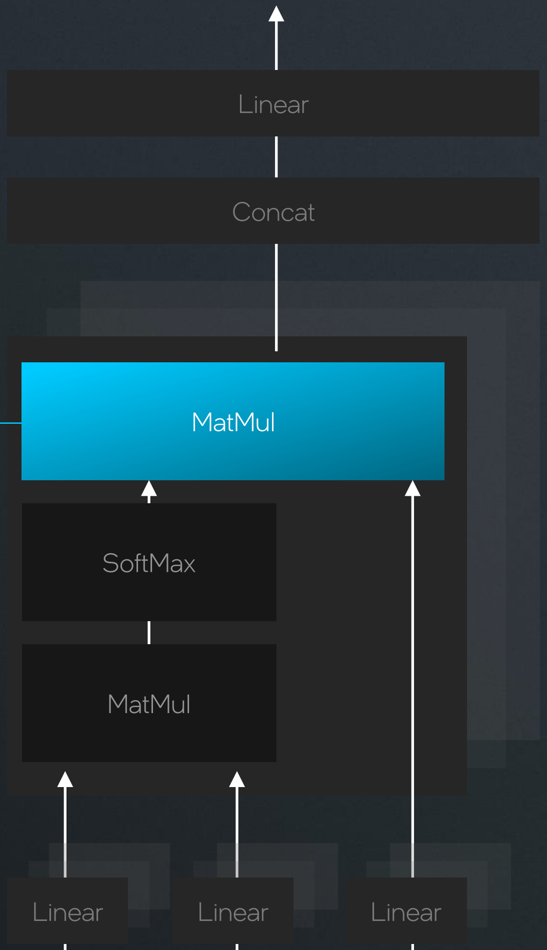
Multi- Head Attention Flowchart



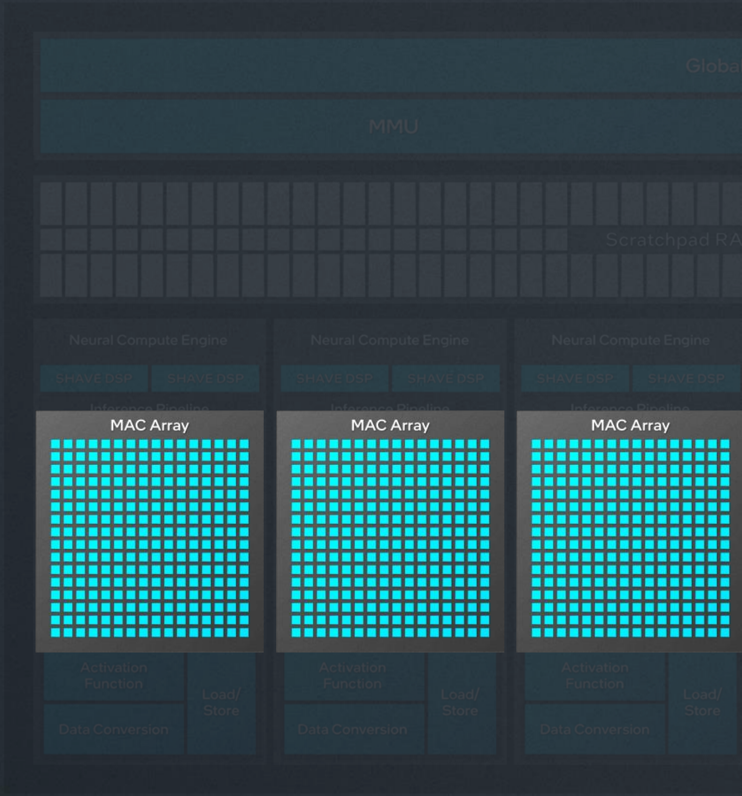
$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



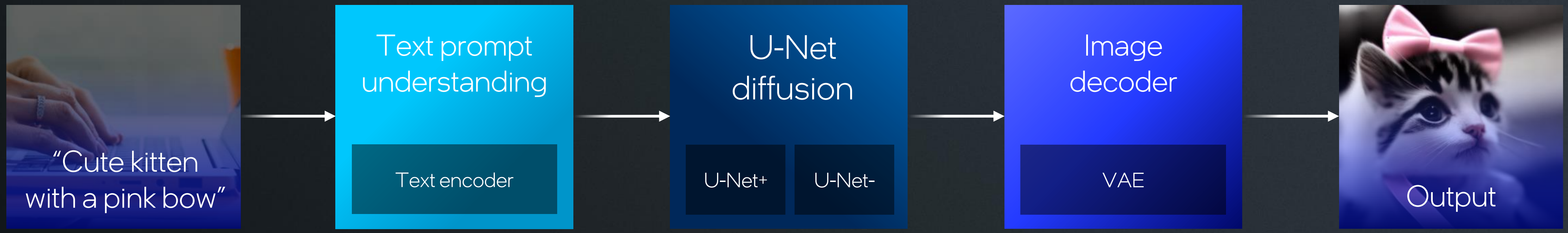
Multi- Head Attention Flowchart



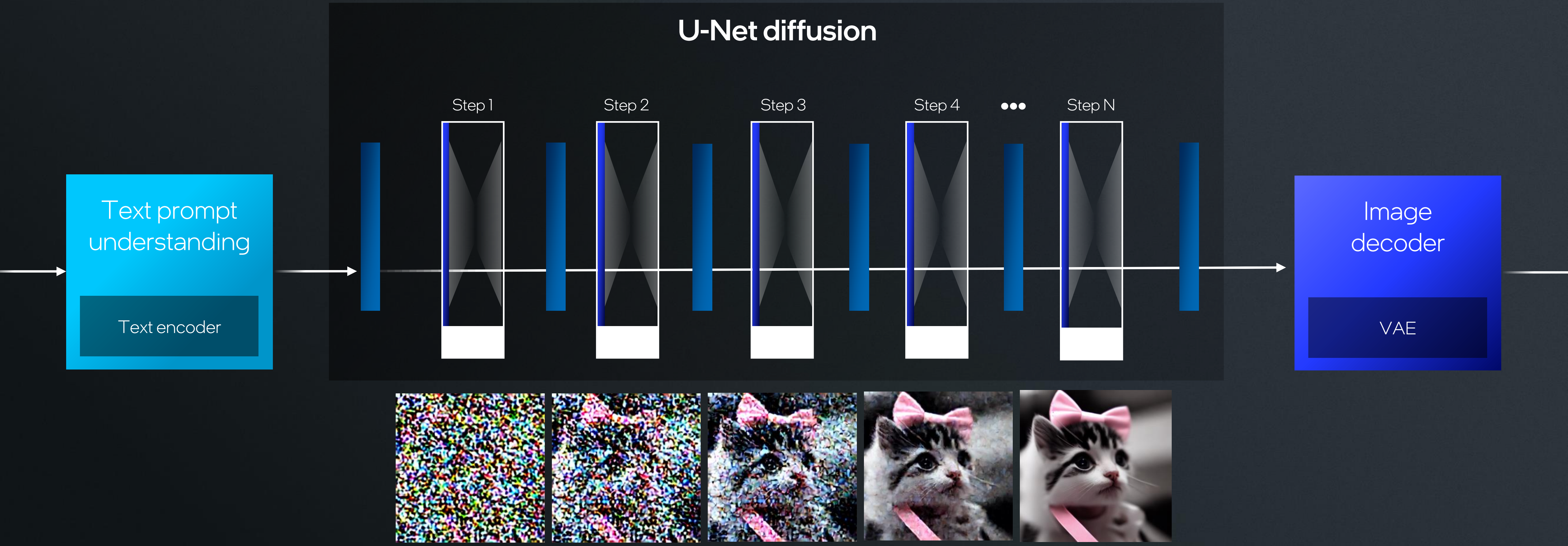
$$Attention(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



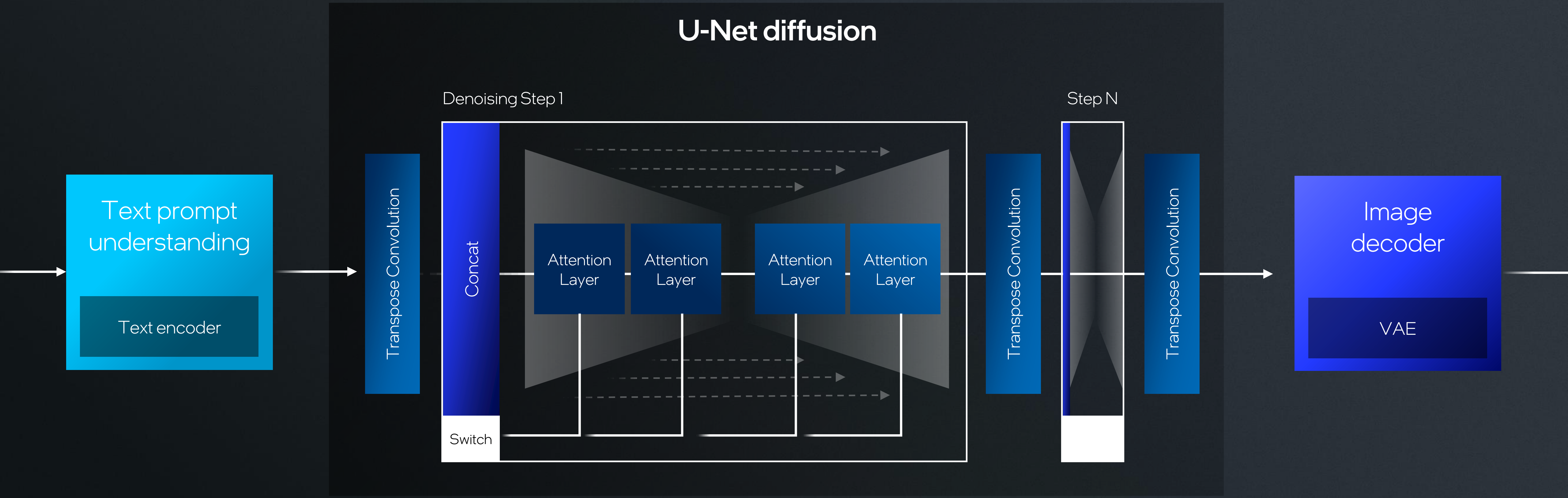
Stable Diffusion Architecture



Stable Diffusion Architecture

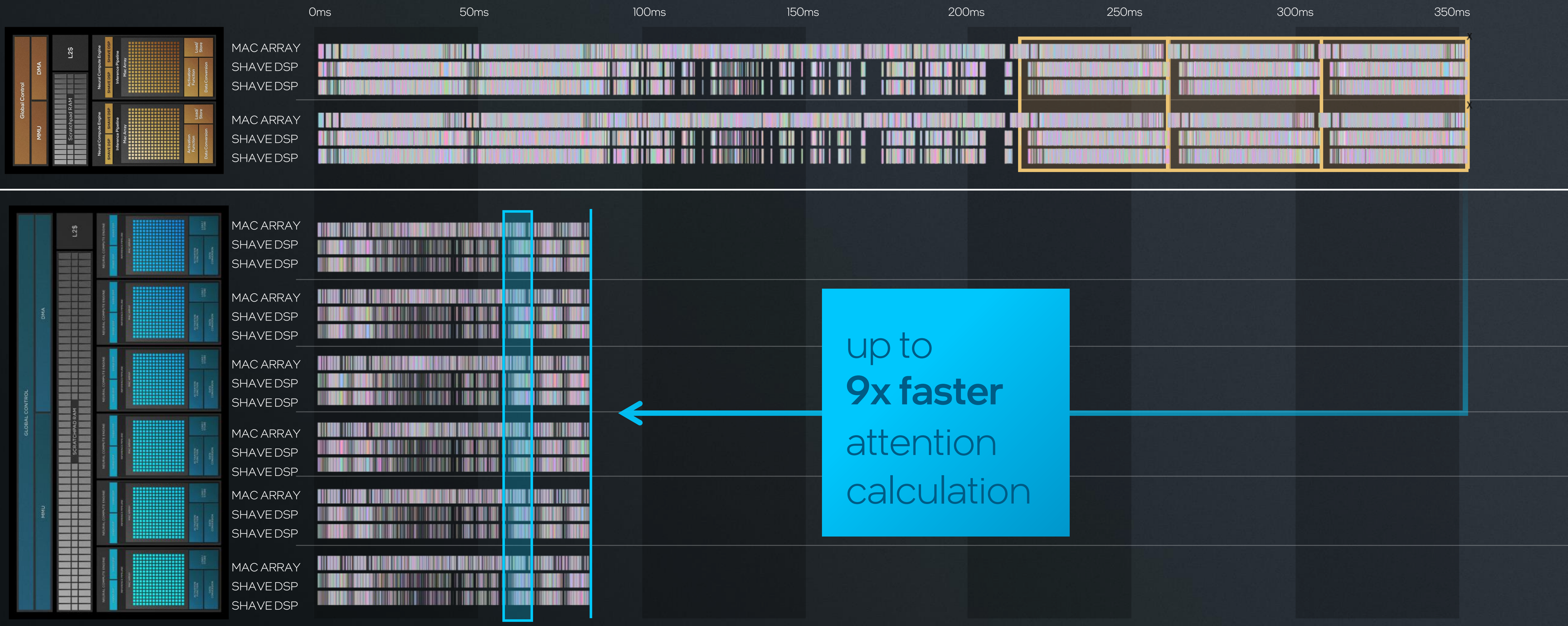


Stable Diffusion Architecture



Accelerating Multi-Head Attention

Performance on U-Net

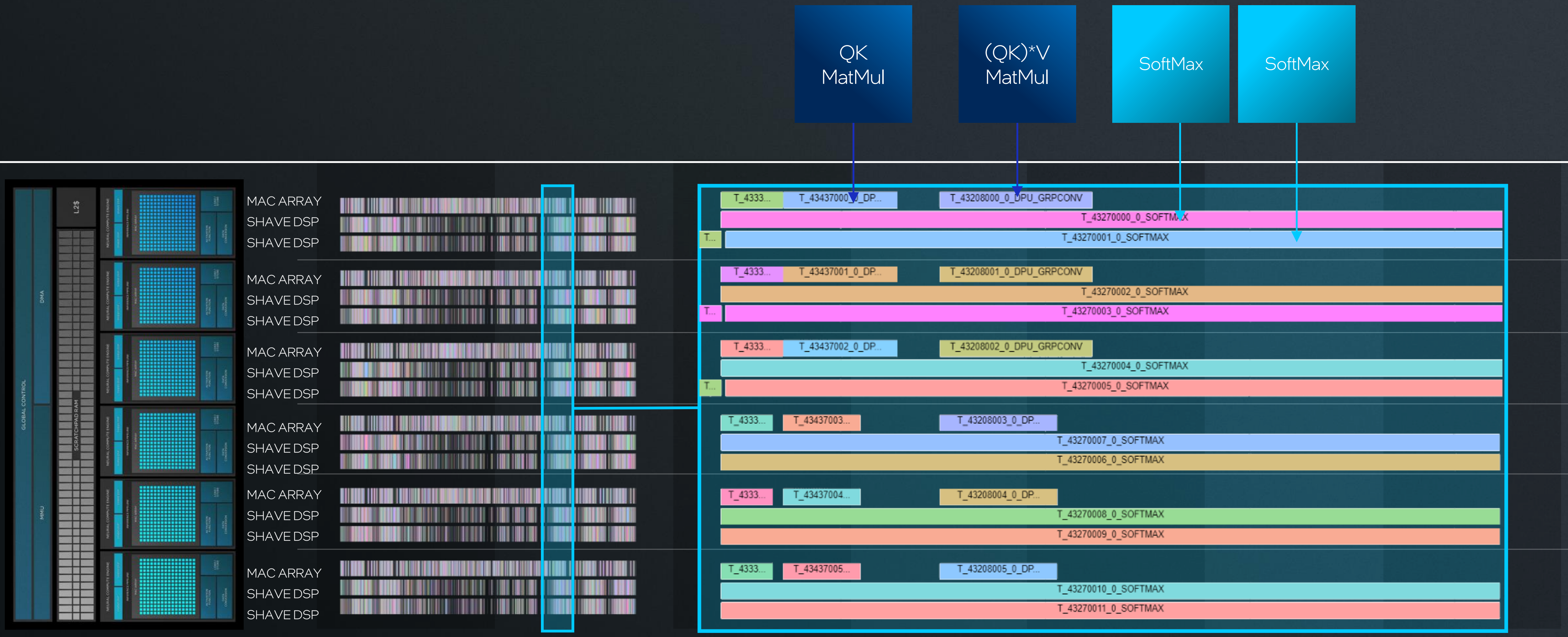


intel. TECH tour.TW

*See details in backup

Accelerating Multi-Head Attention

Performance on U-Net

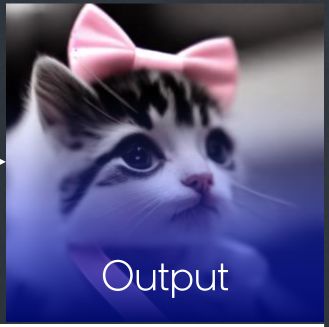


Stable Diffusion

Demo

Stable Diffusion Demo

→ **Text prompt understanding** (Text encoder) → **U-Net diffusion** (U-Net+, U-Net-) → **Image decoder** (VAE) →



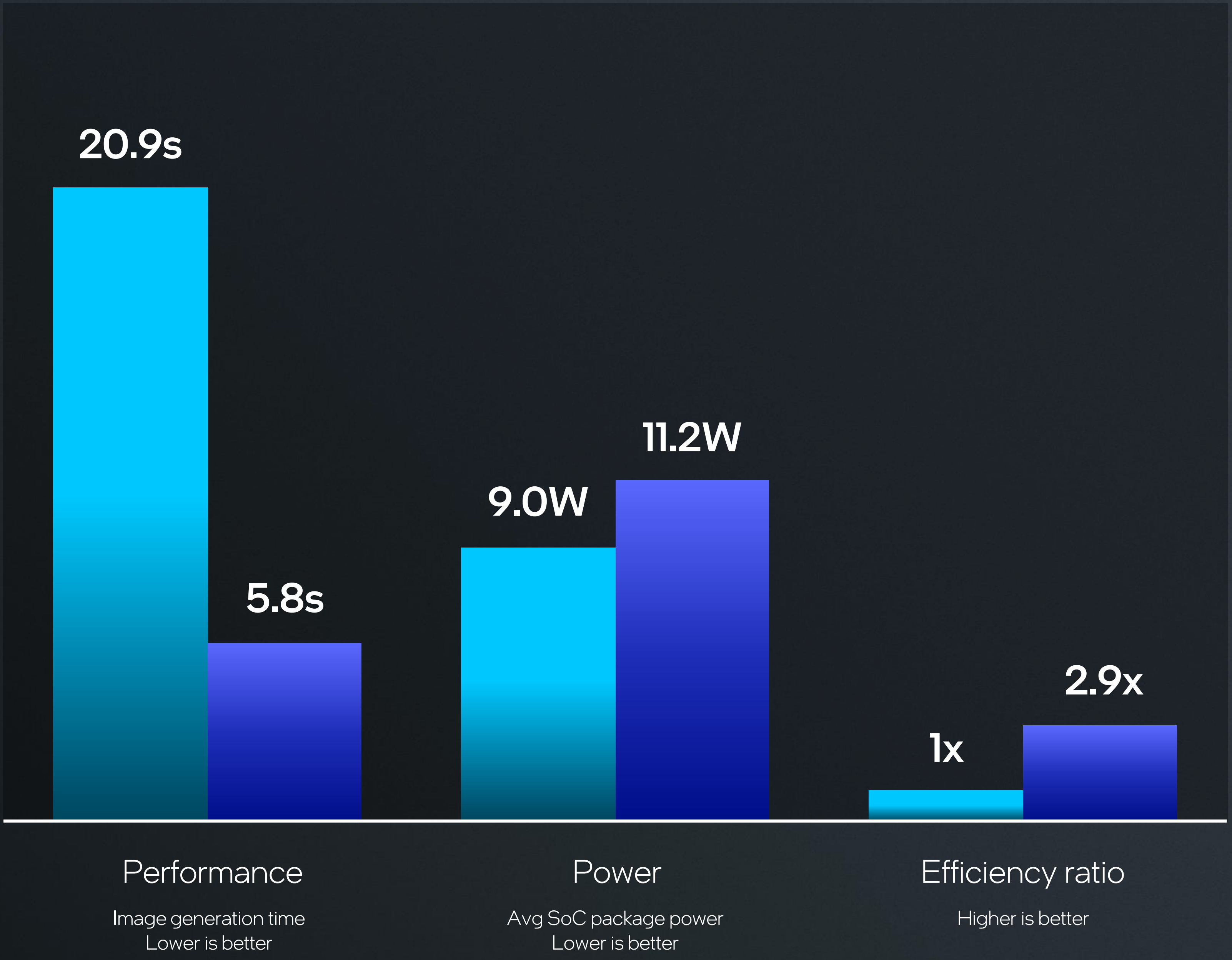
	Text prompt understanding	U-Net diffusion	Image decoder
Meteor Lake	**CPU**	**NPU**	**GPU**
Lunar Lake	**NPU**	**NPU**	**GPU**
Data Type	**FP16**	**INT8**	**FP16**

Stable Diffusion v1.5

20 Iterations

42 Inferences
Text Encoder (1)
+ U-Net+ (20)
+ U-Net– (20)
+ VAE Decoder (1)

Meteor Lake
Lunar Lake



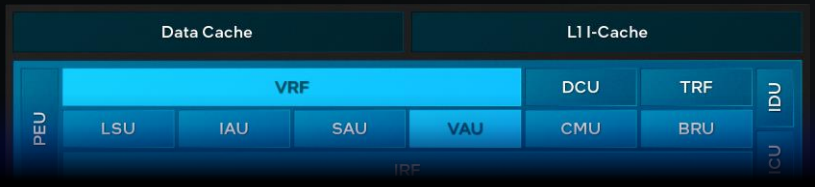
See backup for details. Results may vary.

Next Gen NPU 4

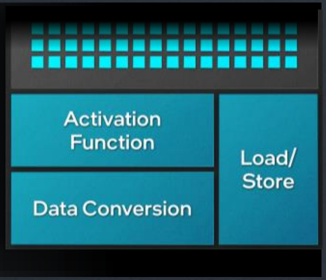
Largest integrated and dedicated AI accelerator for the AI PC

12 Enhanced SHAVE DSPs

Accelerating LLM & transformer operations



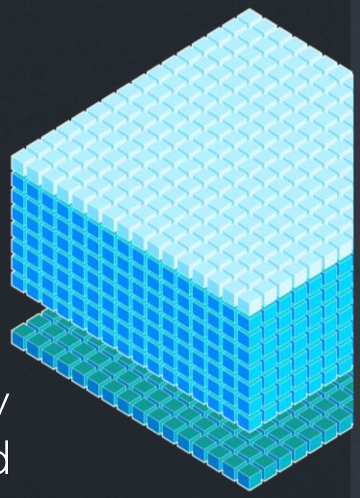
Native activation function & data conversion support



Up to **48** TOPS

2x Bandwidth

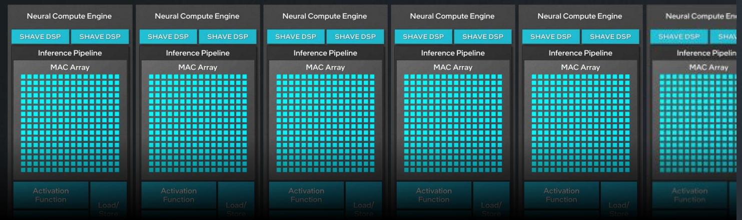
Efficiency optimized MAC array



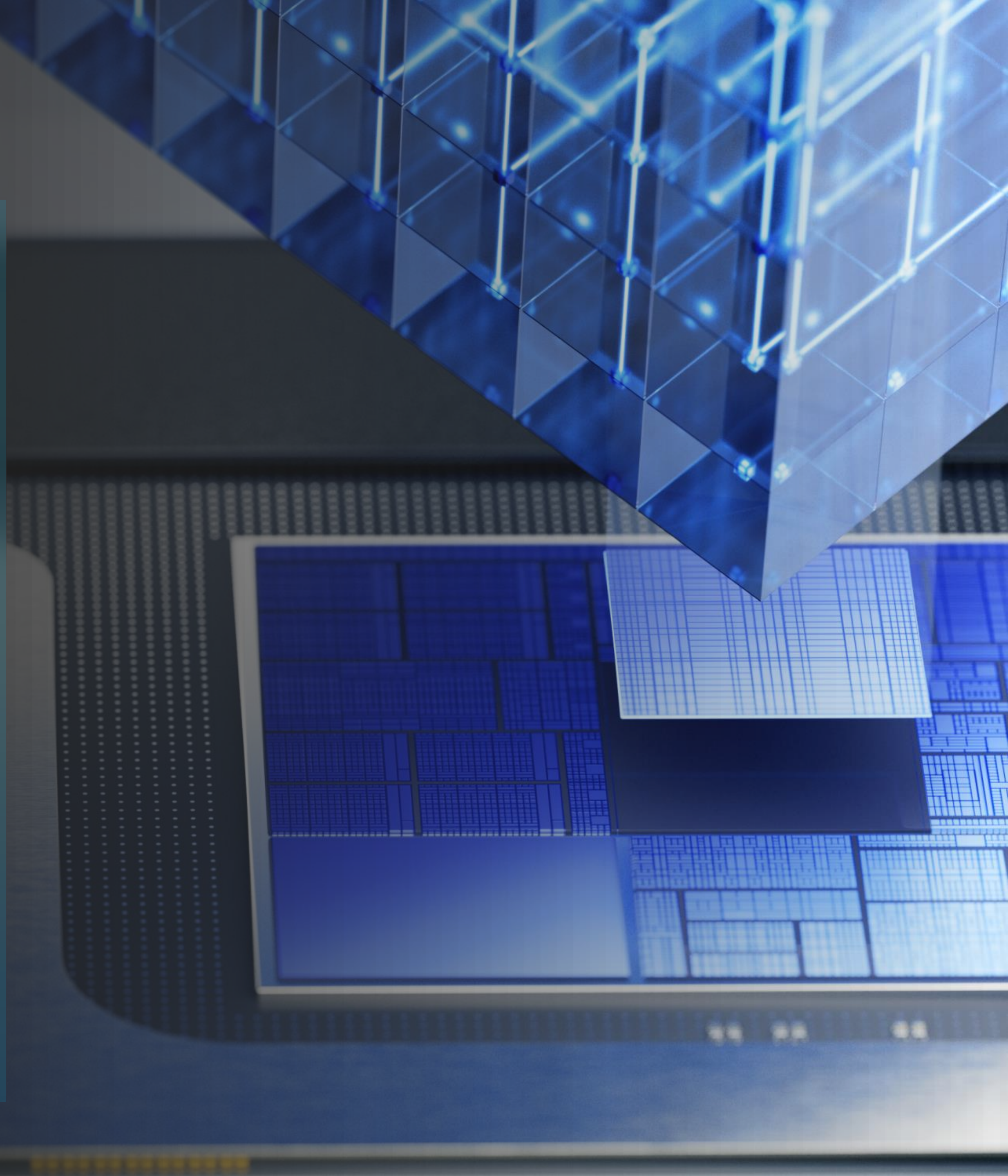
DMA

Embedding tokenization used for LLMs

6 Neural compute engines



Thank You



Notices & Disclaimers

The preceding presentation contains product features that are currently under development. Information shown through the presentation is based on current expectations and subject to change without notice.

Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications or configurations.

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Details at www.intel.com/AIPC.

No product or component can be absolutely secure. Intel technologies may require enabled hardware, software or service activation.

All product plans and roadmaps are subject to change without notice.

Performance hybrid architecture combines two core microarchitectures, Performance-cores (P-cores) and Efficient-cores (E-cores), on a single processor die first introduced on 12th Gen Intel® Core™ processors. Select 12th Gen and newer Intel® Core™ processors do not have performance hybrid architecture, only P-cores or E-cores, and may have the same cache size. See ark.intel.com for SKU details, including cache size and core frequency.

Built-in Intel® Arc™ GPU only available on select Intel® Core™ Ultra processor-powered systems; OEM enablement required.

Some images may have been altered or simulated and are for illustrative purposes only.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

APPENDIX

Claim # & Statement	Slide # & Title/Details
	SLIDE 22: Increased Efficiency & Increased Performance
2x performance at ISO power vs. Meteor Lake	Testing by Intel as of January 2024. Based on VPU-EM simulation. Power data is generated from the simulation tool based on power data that has been extracted from circuit simulation tools. This simulation, which is a ~100% utilization int8 network, is expected to correlate well with silicon.
4x peak performance	4x peak performance is based on TOPS increase from MTL (11 TOPS) to LNL (48 TOPS).
	SLIDE 34: NPU4 Shave DSP
4x Vector compute	Based on 4x vector width increase vs. NPU3 . NPU3 has 8 FP16 Vector ops/clock, NPU4 has 32
12x overall vector performance	Vector performance = 3x tiles and 4x the vector width (vs. NPU3)
	SLIDE 38: NPU 4 Performance
12x vector performance	Vector performance = 3x tiles and 4x the vector width (vs. NPU3)
4x TOPS	TOPS calculation is # of tiles * fmax frequency * ops clock Meteor Lake is up to 11.5 TOPS, Lunar Lake is up to 48 TOPS; Meteor Lake TOPS = (2 tiles * 1.4GHz * 4096 ops/clock)/1000 Lunar Lake TOPS = (6 tiles * 1.95GHz * 4096 ops/clock)/1000
2x IP bandwidth	IP Bandwidth: Meteor Lake is 64GB/s; Lunar Lake is 136 GB/s.
	SLIDE 55: Stable Diffusion v1.5
Lunar Lake vs. Meteor Lake performance, power and efficiency ratio	Testing by Intel as of May 2024. Data based on Lunar Lake reference validation platform vs. Intel® Core™ Ultra 7 155H 32GB LPDDR5-6400Mhz (Meteor Lake). Calculated using open source GIMP with NPU plug in. Text Encoder, & Unet +/- are running on the NPU. VAE is running on the built-in GPU.

intel®