DEEP SCHOOL

лекция

NPU

Neural processing unit

спикер

Виталий Коновалов → Hardware Engineer

О чем сегодня поговорим

Что вообще такое NPU →

О чем сегодня поговорим

Что вообще такое NPU → Что внутри →

О чем сегодня поговорим

Что вообще такое NPU → Что внутри →
Что влияет на развитие NPU →

О чем сегодня поговорим

Что вообще такое NPU → Что внутри →
Что влияет на развитие NPU →
Немного об электронике в целом →

Что вообще такое NPU?

Что вообще такое NPU?

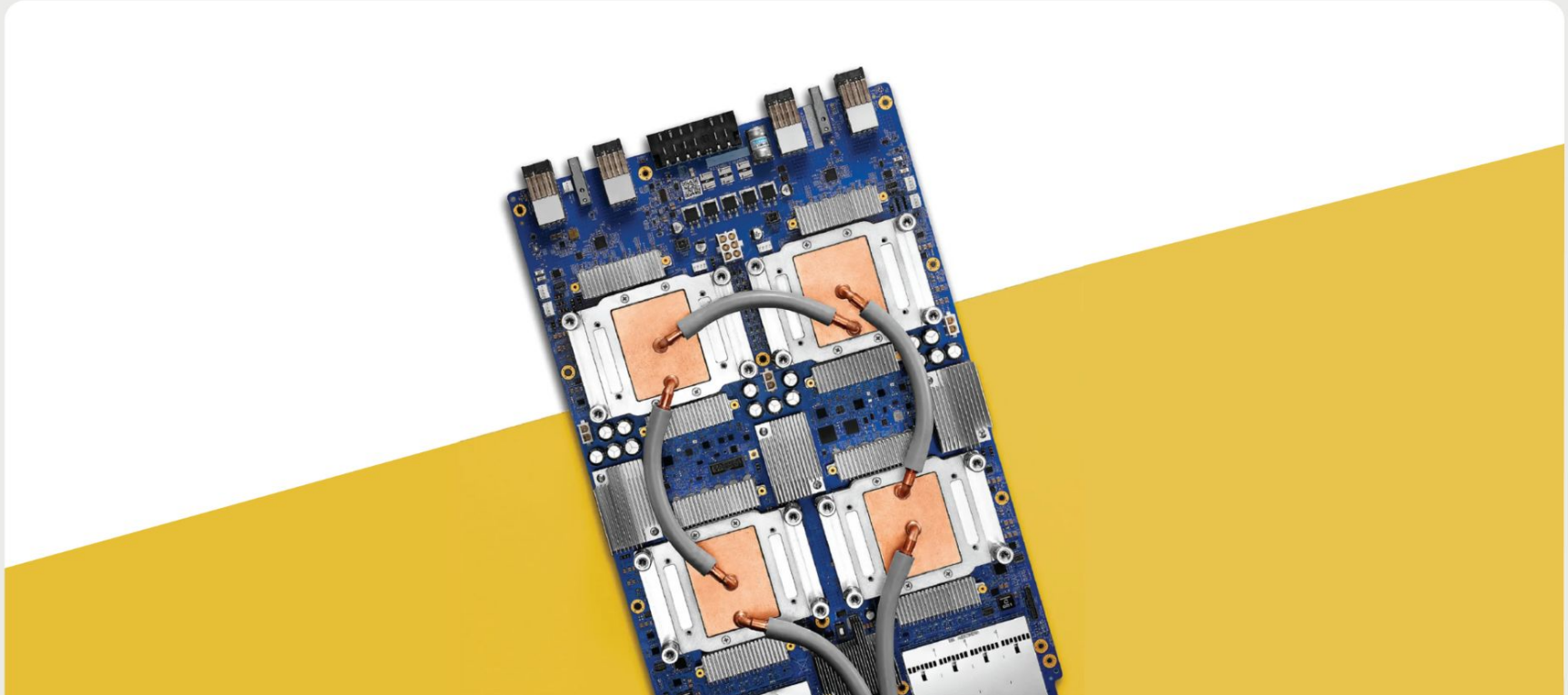
An emerging technology
without a dominant design

— Википедия

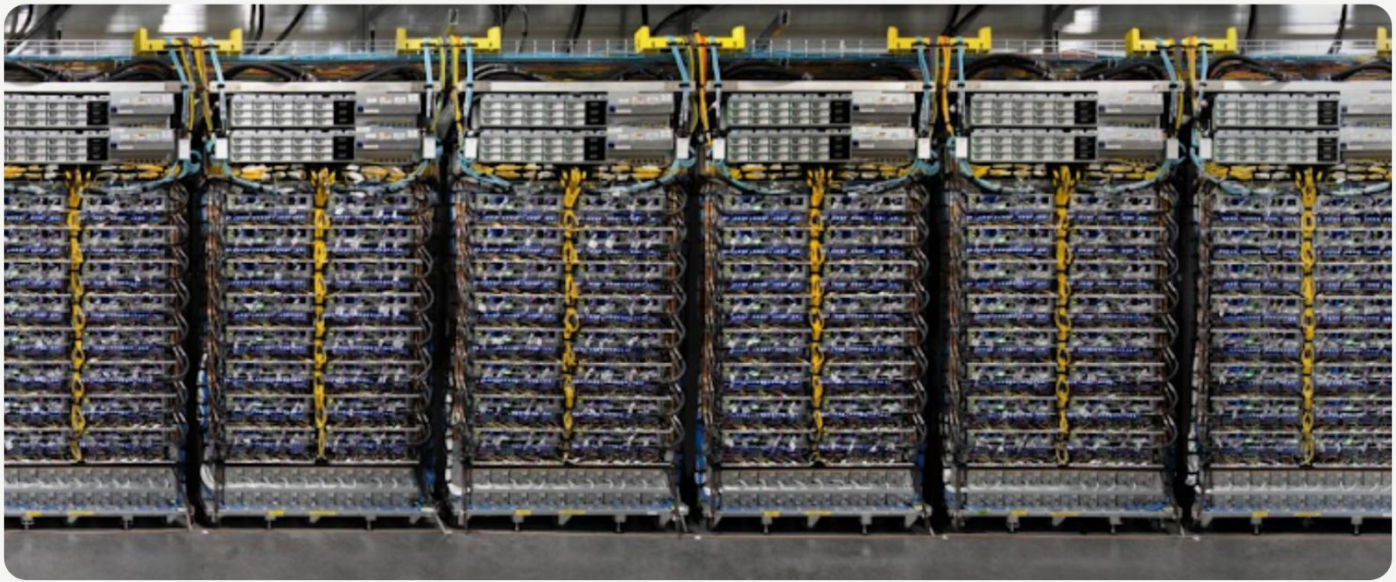
Зачем нам NPU?

Зачем нам NPU?

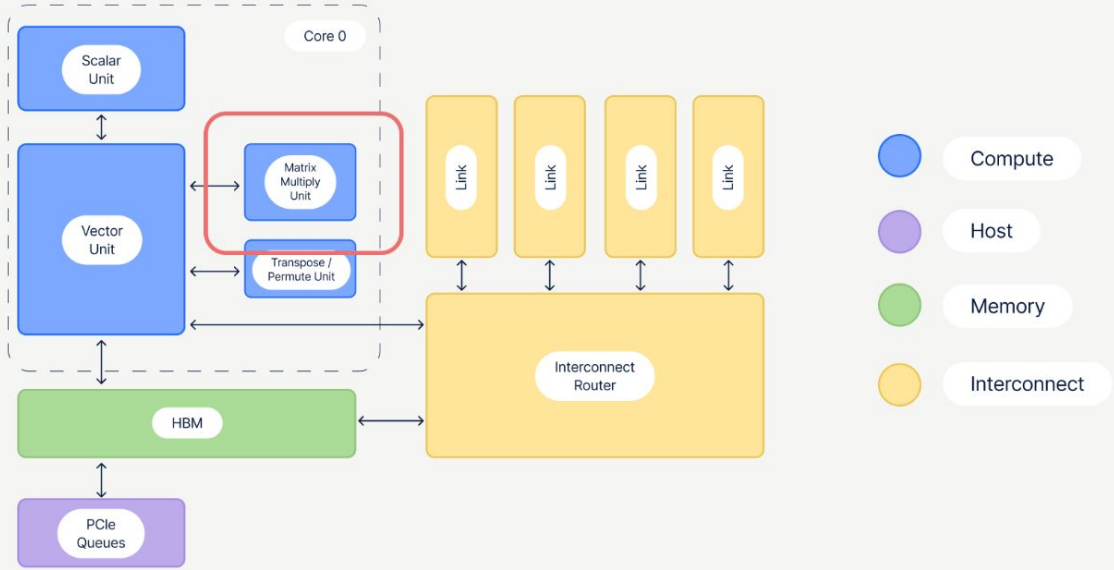
Google TPUv3 board



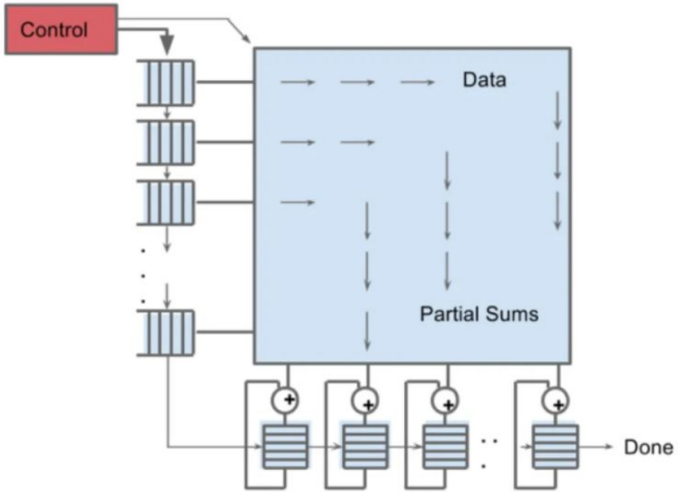
Google TPU supercomputer



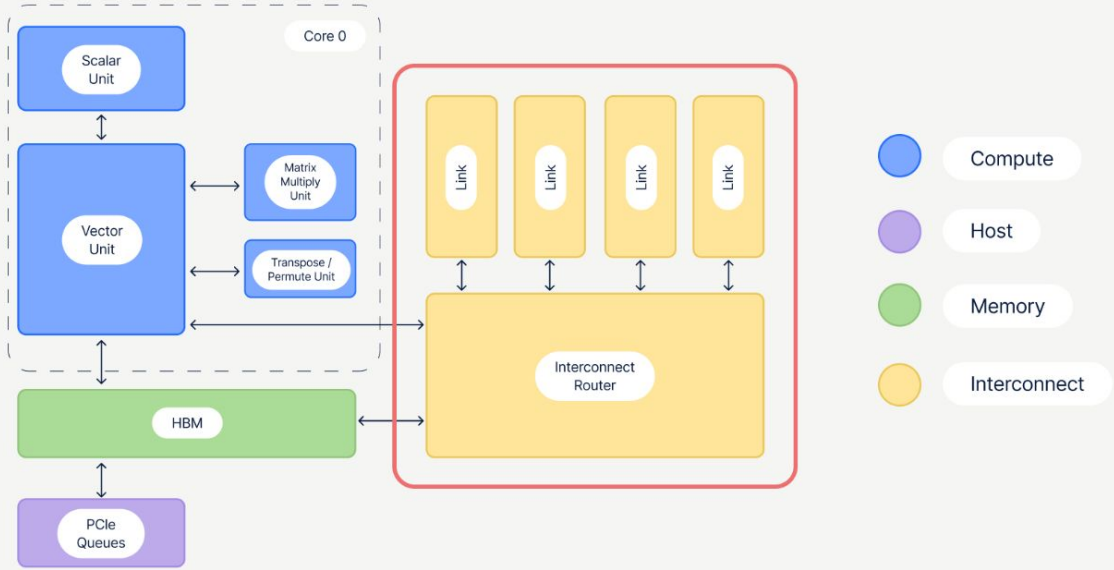
Архитектура TPU v2/v3



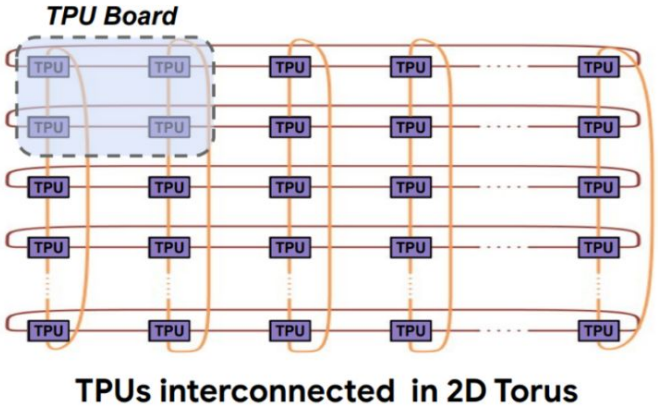
Систолический массив

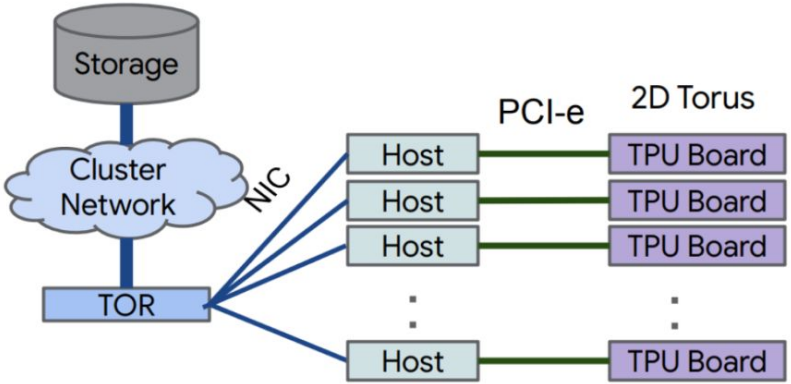


Архитектура TPU v2/v3

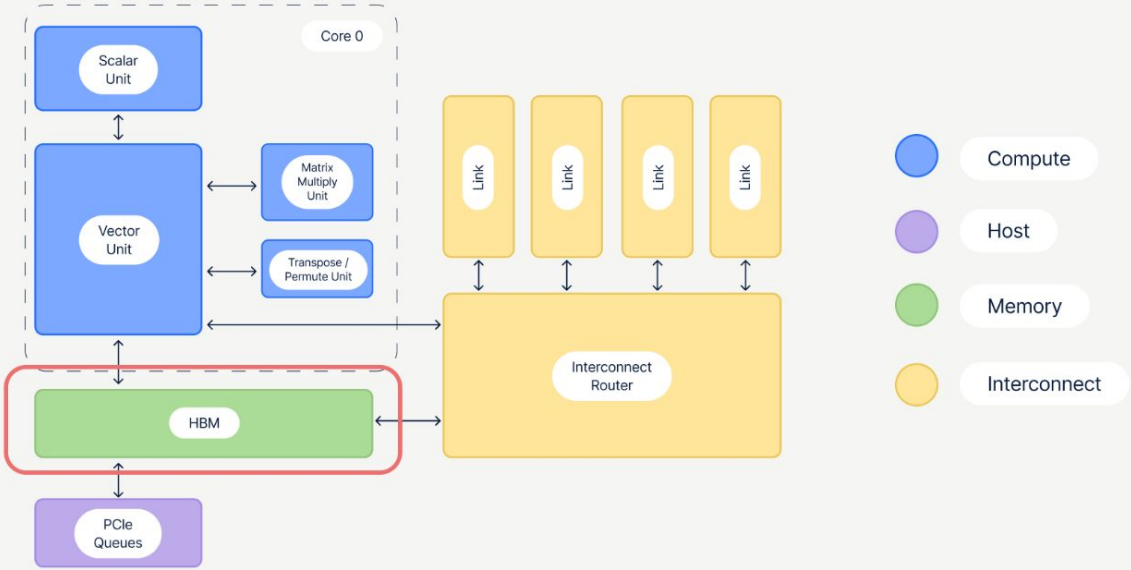


TPU Pod

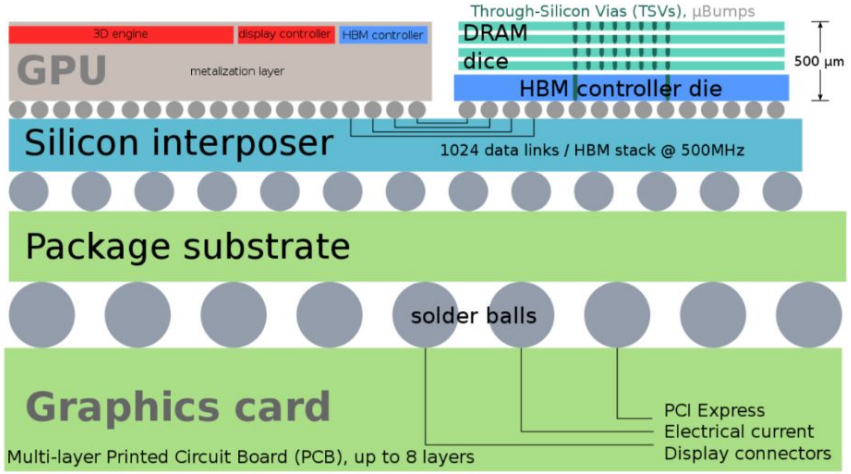

TPU Board
TPU
TPUs interconnected in 2D Torus


Storage
Cluster Network
NIC
TOR
PCI-e
2D Torus
Host
TPU Board

Архитектура TPU v2/v3



High Bandwidth Memory HBM



HBM vs DDR

	DDR4	DDR5	HBM2	GDDR5	LPDDR4	LPDDR5
Applications	Servers → PCs → consumer	Servers → PCs → consumer	Graphics, HPC	Graphics	Mobile, auto, consumer	Mobile, auto, consumer
Typical interface (primary)	Server: 64+8 bits	Server: dual channel, 32+8 bits	Octal channel, 128-bit (1024 bits total)	Multi-channel, 32-bits	Mobile: quad channel, 16-bit (64-bits total)	Mobile: quad channel, 16-bit (64-bits total)
Typical interface (secondary)	Consumer: 32 bits	Consumer: 32 bits	None	None	Dual channel, 16-bit (32-bits total)	Dual channel, 16-bit (32-bits total)
Max Pin BW	3.2 Gb/s	6.4 Gb/s	2.0 → 2.4 Gb/s	8Gbs	4.267Gb/s	6.4Gb/s
Max I/F BW	25.6 GB/s	51 GB/s	307 GB/s	32 GB/s	34 GB/s	51 GB/s
# Pins/channel	~380 pins	~380 pins	~2,860 pins	~170 pins	~350 pins	~370 pins
Max capacity	3DS RDIMM: 128GB	3DS RDIMM: 256GB	4H Stack: 4GB	One channel: 1GB	4 channels: 2GB	4 channels: 4GB
Peak volumes	******	*****	**	*	*****	*****
Price per GB	$	$$	$$$$	$$$	$$	$$

Memory-bound

Типы данных и экономия на плюсиках

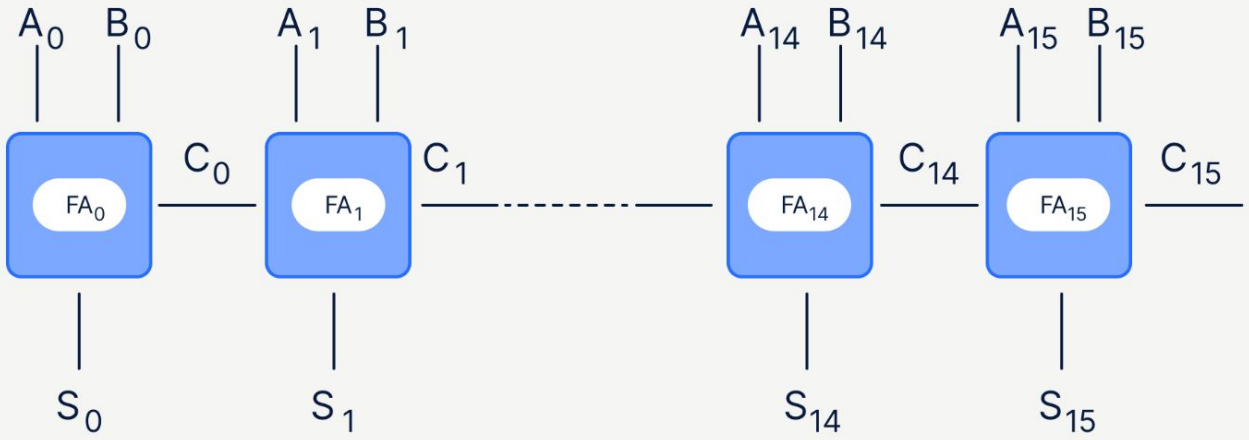
Типы данных в TPU

- INT8 1
- FP16/FP32 2
- BF16 3

Как сложить два числа в железе

или что если у вас нет оператора +?

Сложение INT16 в железе



Сложение FP16 в железе

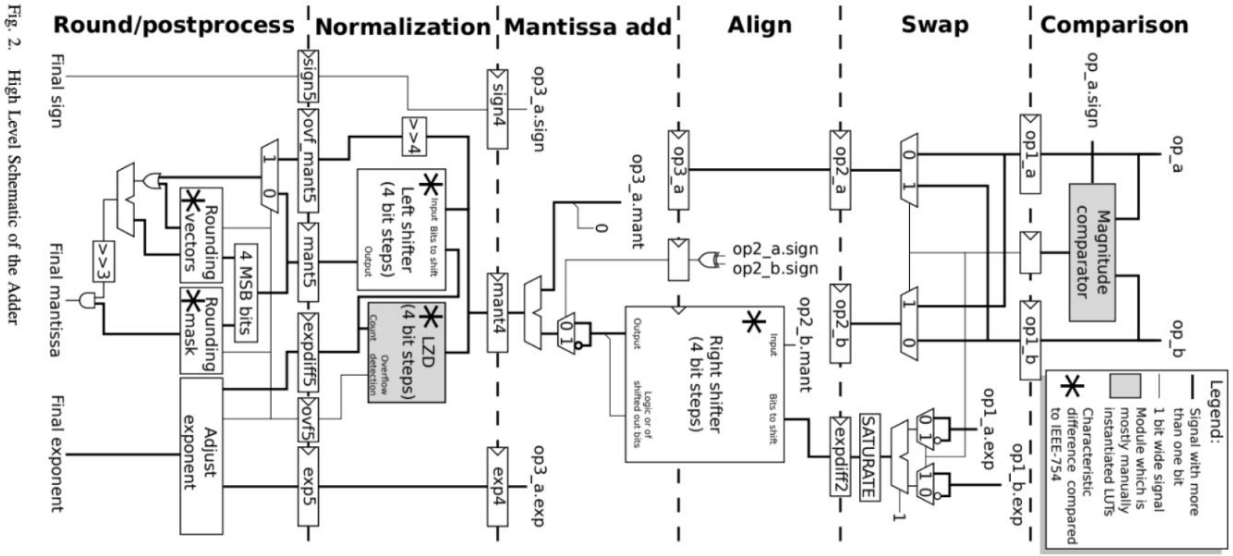


Fig. 2. High Level Schematic of the Adder

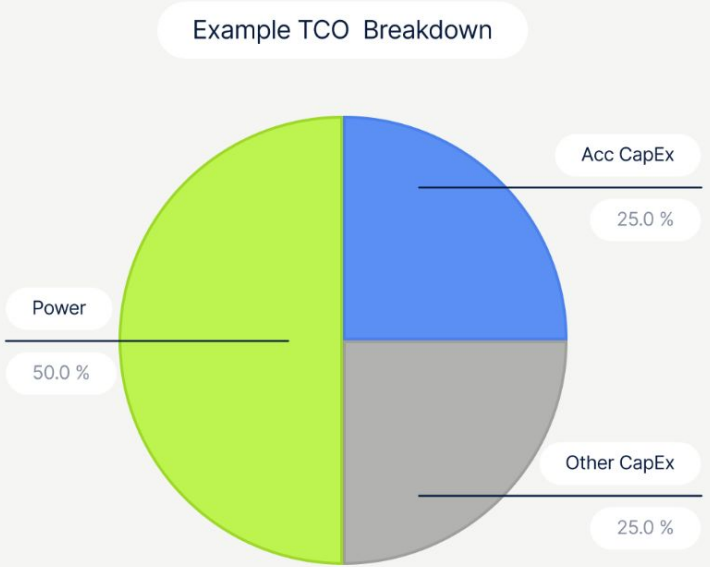
Влияние на производительность

	+	×
INT8	0.007	0.07
INT32	0.03	1.48
BF16	0.11	0.21
IEEE FP16	0.16	0.34

пДж на операцию для процесса 7 нм

Немного абстрактной бухгалтерии

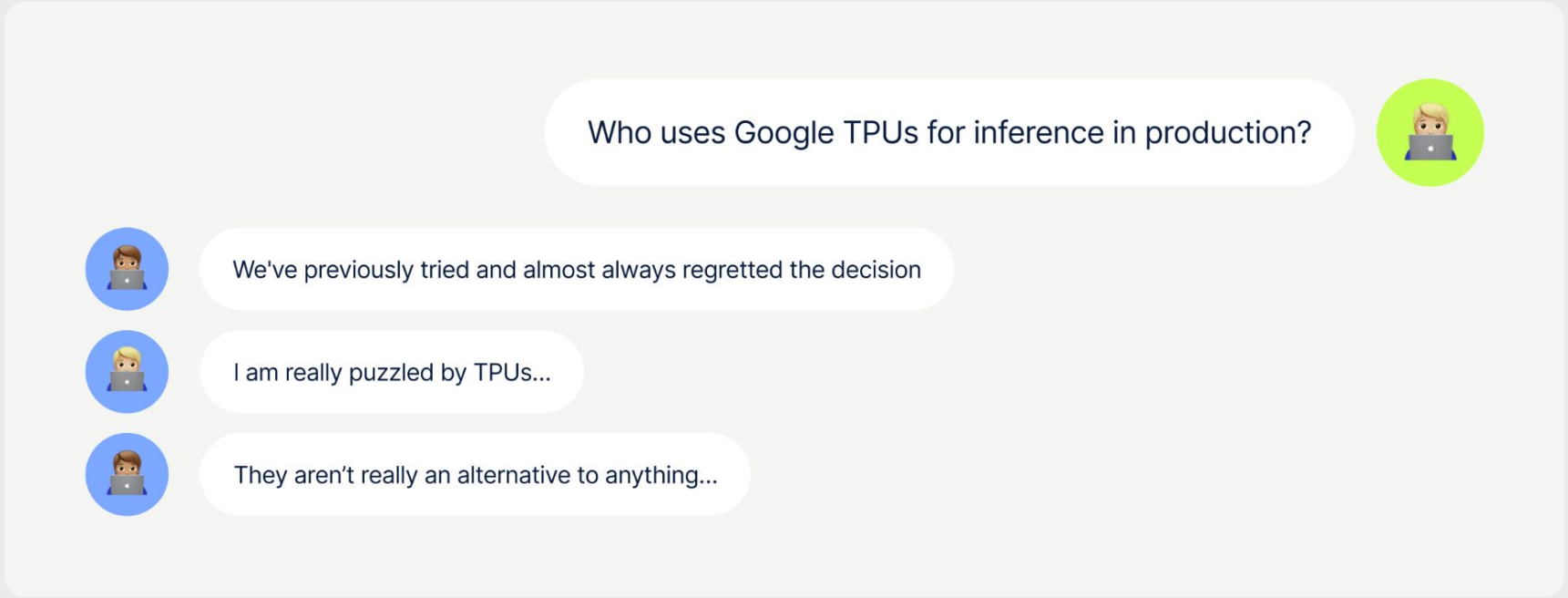
CapEx	→	цена покупки
OpEx	→	стоимость эксплуатации электричество, охлаждение, аренда и пр.
Total Cost of Ownership (TCO)	→	общая стоимость устройства



Training & Interference TPUs

	TPUv4i	TPUv4
Размер кристалла	400 мм2	780 мм2
TensorCore на чип	1	2
MXU	4 * 128×128	4 * 128×128
On-chip cache	144 MB	288 MB
HBM	8 GB	32 GB
Поддержка типов данных	BF16/INT8	BF16/INT8

Мнение рандомов в интернете



Источник ↗

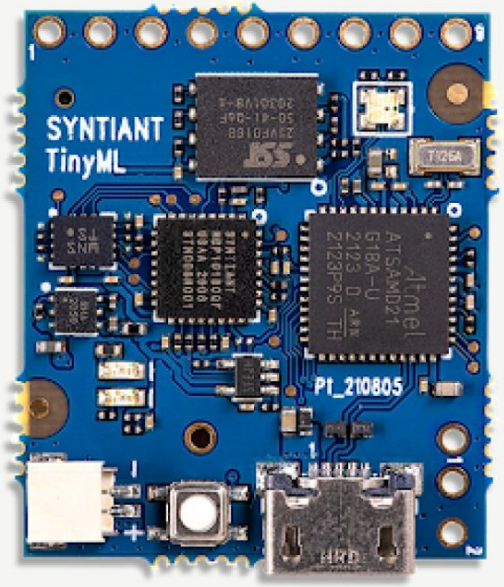
Что ещё бывает кроме Google TPU

01 Groq LPU

02 Intel Gaudi / Habana Goya

03 NVIDIA H100

Edge, Tiny & Datacenter



Что влияет на развитие NPU

Что влияет на развитие NPU

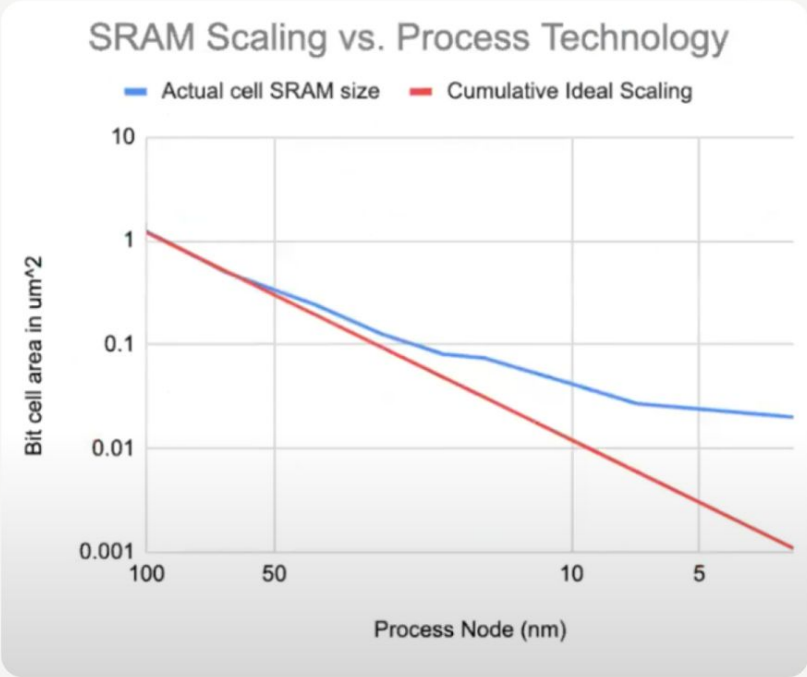
1

Производственный цикл NPU и рост моделей

2

Компиляторы и обратная совместимость

Что влияет на развитие NPU Закон Мура



Что влияет на развитие NPU Закон Мура