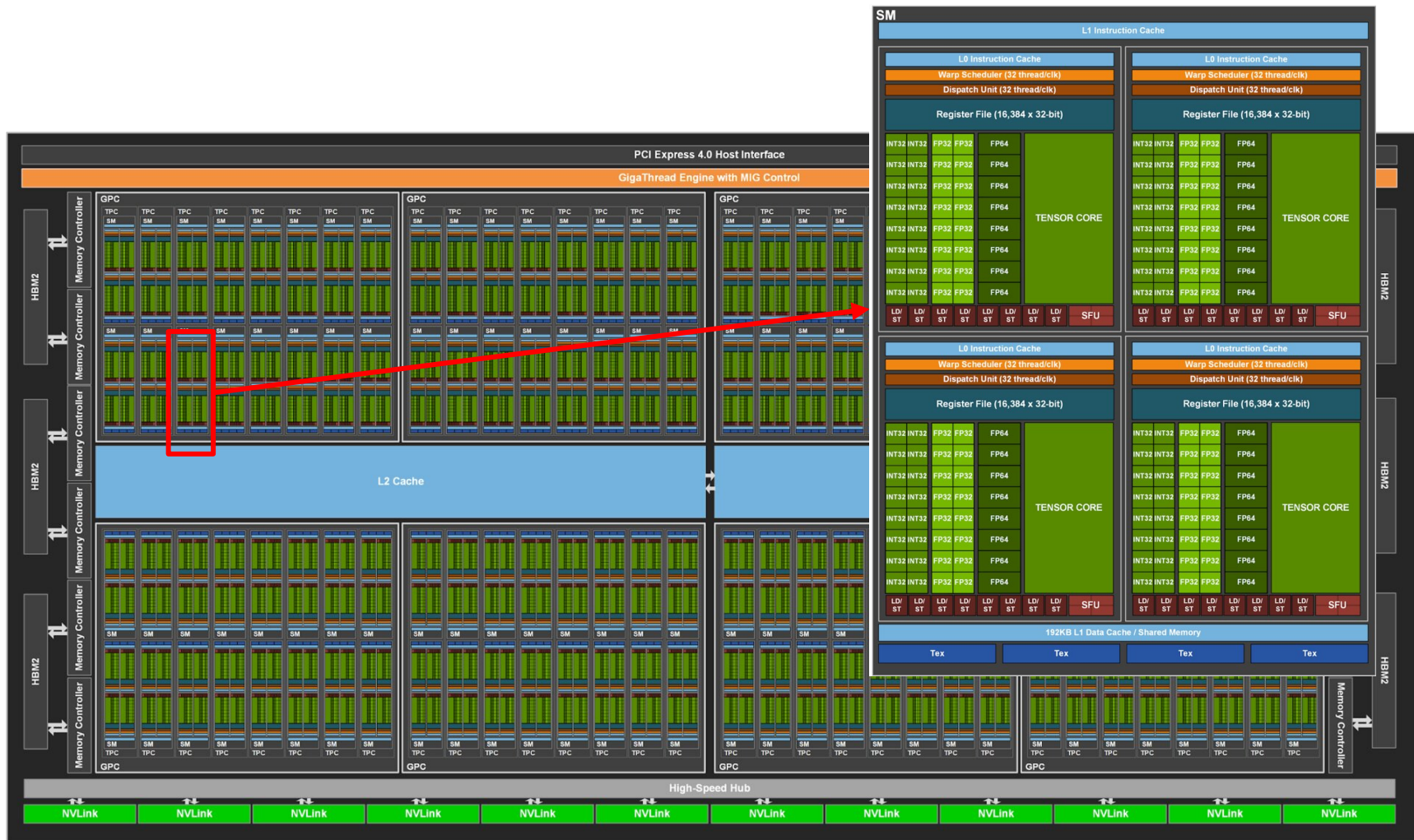




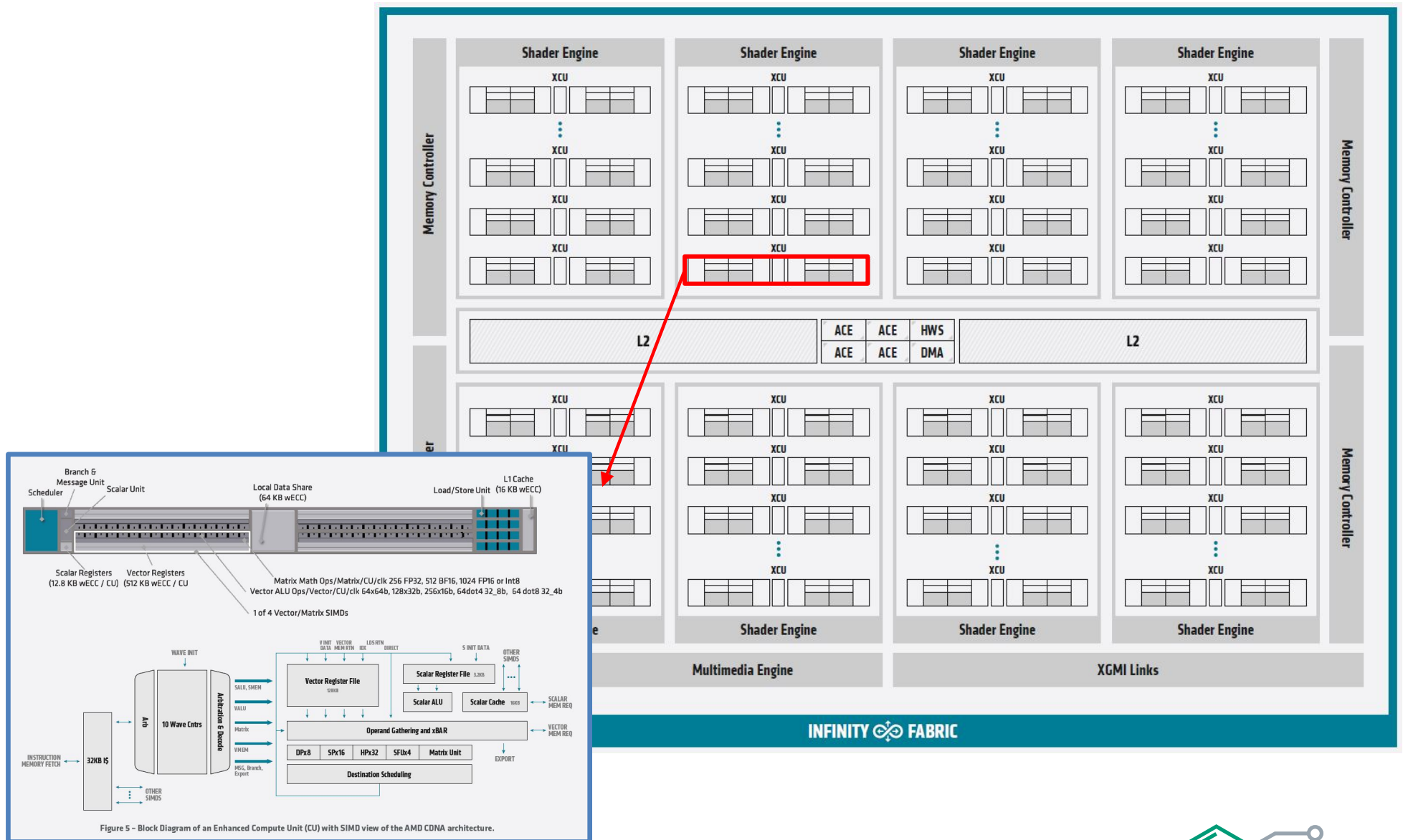
# GPU Top Trumps!

What modern GPUs can do

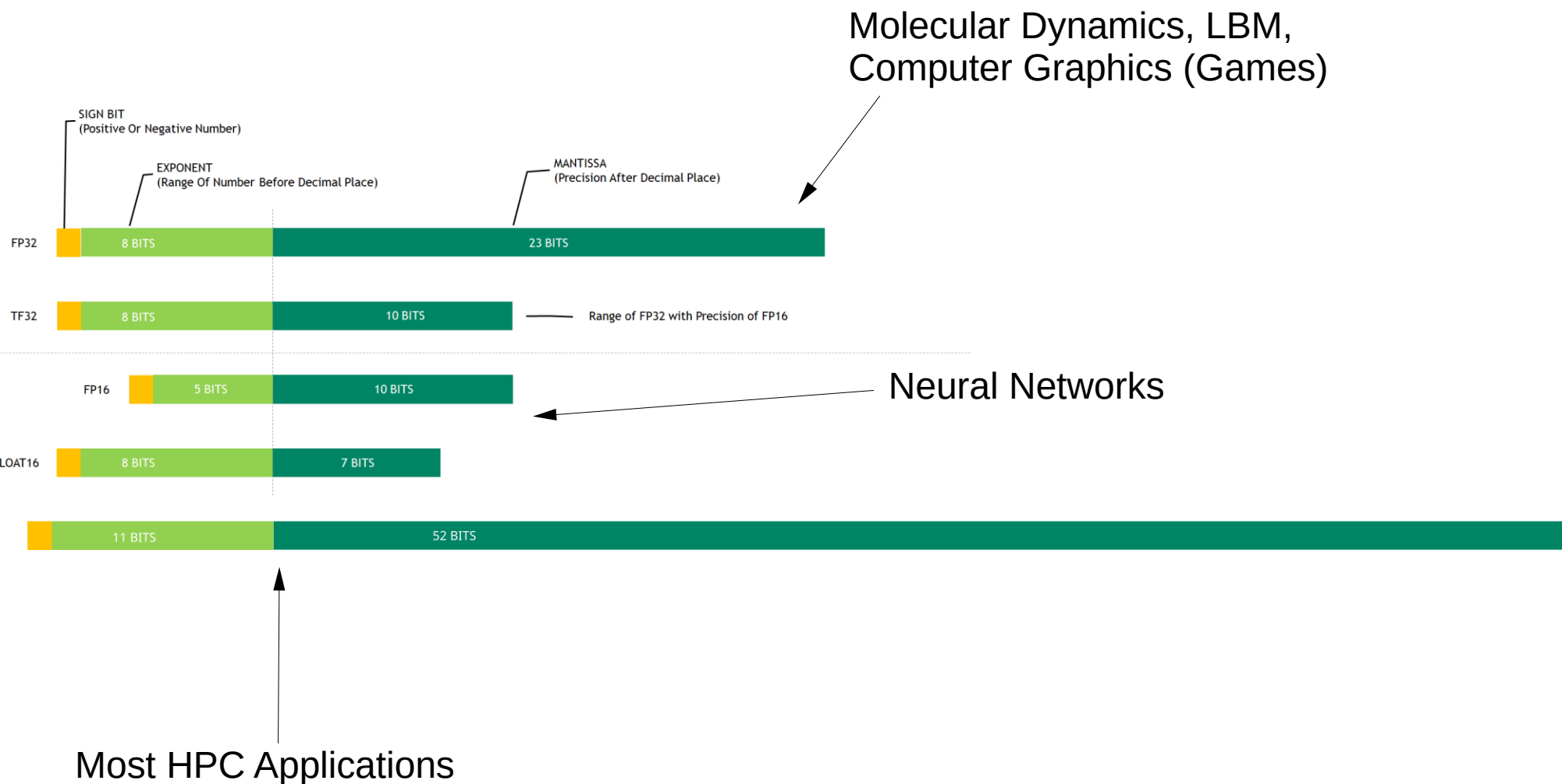
# GPU Architecture



# GPU Architecture



# FP Number Formats





# Vector vs Tensor Flops

TENSOR CORE 4X4X4 MATRIX-MULTIPLY ACC

$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32      FP16      FP16      +      FP16 or FP32

8 NVIDIA



















































Tensor cores are only usable for certain workloads!

- Matrix Multiplication a la Linpack
- Neural Networks























































































# FP Throughput

		NVIDIA RTX3080	NVIDIA A40	NVIDIA A100	NVIDIA H100	AMD MI100	AMD MI210	Intel Max 1100	Intel Max 1550
			Alex	SXM4-80GB Alex	PCIe-80GB				
FP16 (Tensor)		119	147	312	756	184	181	355	839
FP32 (Vector)	<i>Tflop/s</i>	30	37,5	19,5	51	23	46	22	52
FP64 (Vector)		0.5	0.6	9,7	25	11,5	23	22	52
TDP		320	300	400	350	300	300	300	600

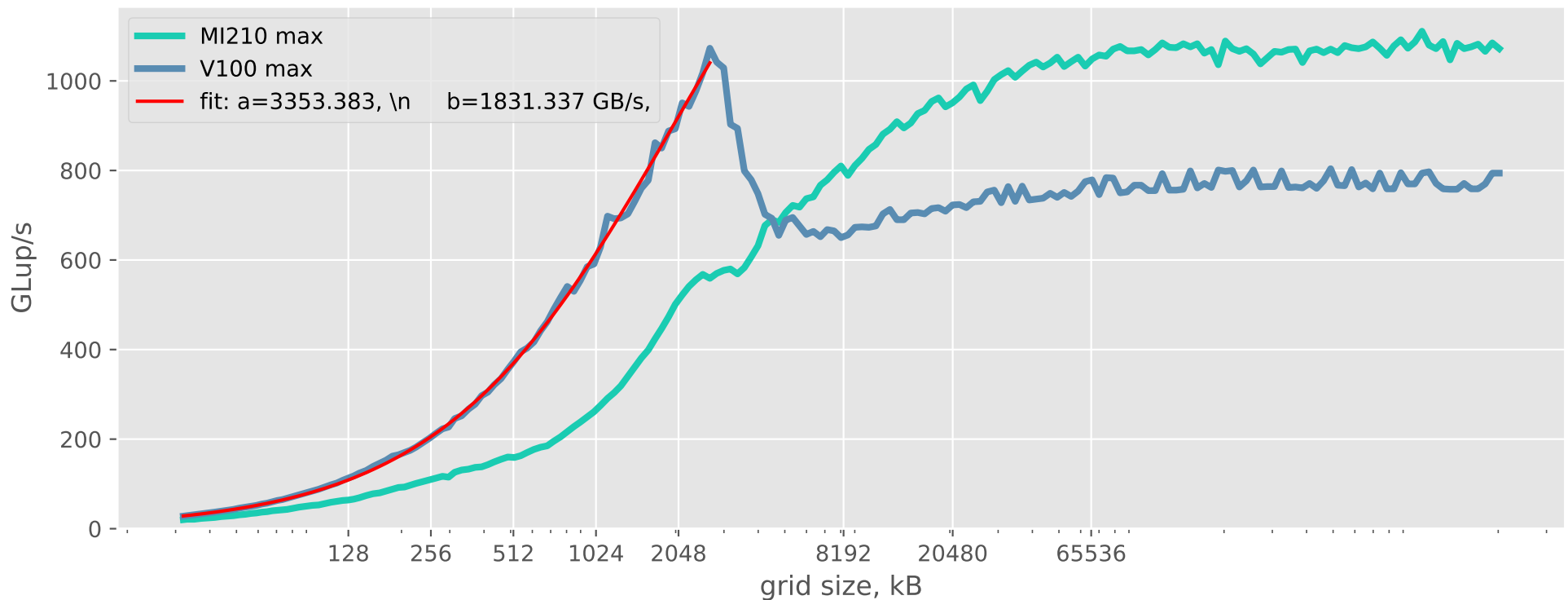
# Memory & Cache Bandwidth

		NVIDIA RTX3080	NVIDIA A40	NVIDIA A100	NVIDIA H100	AMD MI100	AMD MI210	Intel Max 1100	Intel Max 1550
			Alex	SXM4-80GB Alex	PCIe-80GB				
FP16 (Tensor)	<i>Tflop/s</i>	 119	 147	 312	 756	 184	 181	 355	 839
FP32 (Vector)		 30	 37,5	 19,5	 51	 23	 46	 22	 52
FP64 (Vector)		0.5	0.6	 9,7	 25	 11,5	 23	 22	 52
<u>TDP</u>		 320	 300	 400	 350	 300	 300	 300	 600
DRAM <u>BW</u>	<i>GB/s</i>	 760	 696	 2039	 2039	 1229	 1638	 1229	 3277
DRAM <u>BW meas.</u>			 655	 1732	 1704	 1042	 1358	 799	
L2 <u>BW meas.</u>			 2430	 4700	 8300	 2400	 5100	 2714	
L2 Cache	<i>MB</i>	5	6	40	50	8	8	104	408
DRAM	<i>GB</i>	10	48	80	80	32	64	48	128

# Gromacs Performance

		NVIDIA RTX3080	NVIDIA A40	NVIDIA A100	NVIDIA H100	AMD MI100	AMD MI210	Intel Max 1100	Intel Max 1550
			Alex	SXM4-80GB Alex	PCIe-80GB				
FP16 (Tensor)	<i>Tflop/s</i>	 119	 147	 312	 756	 184	 181	 355	 839
FP32 (Vector)		 30	 37,5	 19,5	 51	 23	 46	 22	 52
FP64 (Vector)		0.5	0.6	 9,7	 25	 11,5	 23	 22	 52
<u>TDP</u>		 320	 300	 400	 350	 300	 300	 300	 600
DRAM BW	<i>GB/s</i>	 760	 696	 2039	 2039	 1229	 1638	 1229	 3277
<u>DRAM BW meas.</u>			 655	 1732	 1704	 1042	 1358	 799	
<u>L2 BW meas.</u>			 2430	 4700	 8300	 2400	 5100	 2714	
L2 Cache	<i>MB</i>	5	6	40	50	8	8	104	408
DRAM	<i>GB</i>	10	48	80	80	32	64	48	128
System 1	<i>ns/day</i>	 215	 243	 242	 270	 99	 148		
System 2		 617	 713	 647	 828	 363	 411		
System 3		 215	 254	 249	 310	 171	 187		
System 4		 101	 121	 125	 157	 87	 107		
System 5		 30	 34	 37	 48	 26	 34		
System 6		 16	 19	 22	 29	 16	 19		

# Small Grids → Launch Latency





# The Future: APUs

