FROM GAMING TO AI: TRANSFORMING YOUR PC INTO AN AI MODEL HOSTING BEAST

intel

AMD

PC BUILD FOR AI/ GAMING/ EDITING

HTTPS://GITHUB.COM/INSIGHTBUILDER

ARE YOU A GAMER / VIDEO EDITOR?

YOUR GAMING / EDITING PC "COULD POSSIBLY" HOST AI MODELS???

ARE YOU BUILDING A NEW PC???

ARE YOU PLANNING TO BUILD A NEW PC / SERVER FOR GAMING / EDITING???

THIS VIDEO IS A MUST WATCH

LETS DIVE INTO THE RABBIT HOLE....

HTTPS://GITHUB.COM/INSIGHTBUILDER

BEFORE YOU BUILD: WHY???

- WANT TO RUN THE NLP & VISION AI MODELS & 1080P / 4K GAME TITLES
- MODERATE 3D RENDERING & VIDEO EDITING IS REQUIRED FOR SHOWCASING YOUR WORK
- EAGER TO LEARN ABOUT THE HARDWARE THAT CREATES THE GENERATIVE AI MAGIC
- WANT TO RUN ATLEAST 4 TO 6 VMS & 10 TO 20 DOCKER INSTANCES OF VARIOUS APPLICATIONS
- WANT TO SIMULATE BIG DATA NODES IN A SINGLE HOST SYSTEM
- LAPTOP / CLOUD COSTS ARE HINDERING YOUR LEARNING PROGRESS
- EXPERIMENTING & ITERATING WITH LIBRARIES AND FRAMEWORKS RAPIDLY
- LOOKING FOR A RIGHT WAY TO INVESTMENT IN YOUR FUTURE
- PASSIONATE ABOUT SOLVING VARIETY OF PROBLEMS RANGING FROM FRONT-END, DATA STORAGE AND BACK END TECH

OR

I WANT TO PLAY THE ASSASSINS CREED / OVERWATCH/ HITMAN 3 ON MAX SETTINGS... PERIOD

PC PARTS & THEIR PURPOSE

CUDA & TENSOR CORES PROCESS IN PARALLEL

DEDICATED VRAM

P-CORES & E-CORES PROCESS IN PARALLEL

DEDICATED RAM

MOTHER BOARD

POWER SUPPLY

HTTPS://GITHUB.COM/INSIGHTBUILDER

SIMPLE & STRAIGHT : AI / ML

TENSOR CORES

+

MEMORY BANDWIDTH

REST OF THE PARTS ARE JUST LOGISTICS FOR YOUR DATA

SIMPLE & STRAIGHT : GAMING

CUDA CORES

+

TENSOR CORES (OPTIONAL)

+

GPU VRAM

IN ADDITION THE CPU & RAM IS IMPORTANT

HTTPS://GITHUB.COM/INSIGHTBUILDER

SIMPLE & STRAIGHT : VIDEO EDITING

CUDA CORES

+

MEMORY BANDWIDTH

+

VIDEO ENCODERS (OPTIONAL)

REST OF THE PARTS ARE JUST LOGISTICS FOR YOUR DATA

CAN GAMING GPU DO AI/ML?

IF IT HAS TENSOR CORES
THEN IT CAN

GEFORCE > TURING / QUADRO
RTX & TITAN FAMILY

AMD / INTEL GPU DON'T
SUPPORT TENSOR CORES

HTTPS://GITHUB.COM/INSIGHTBUILDER

MEMORY CACHE TREE

GPU MEMORY ACCESS TIMES

GPC <- L1 <- L2 <- VRAM <- SSD

- GLOBAL MEMORY ACCESS < 80GB : ~380 CYCLES
- L2 CACHE: ~200 CYCLES
- L1 CACHE OR SHARED MEMORY ACCESS (UP TO 128 KB PER STREAMING MULTIPROCESSOR): ~34 CYCLES
- FUSED MULTIPLICATION AND ADDITION, A*B+C (FFMA): 4 CYCLES
- TENSOR CORE MATRIX MULTIPLY: 1 CYCLE

- LATENCY : TIME TAKEN IN CYCLES FOR A OPERATION TO COMPLETE
- ONE OPERATION IS DONE BY 32 THREADS. WHICH IS CALLED A WARP
- WARPS ARE SYNCHRONOUS. MAX 32 WARPS(1024 THREADS)
- GLOBAL MEMORY LOAD AT 32 * 4 BYTES (32 FLOATS)
- 32 WARPS / STREAMING PROCESSOR
- USING TENSOR CORE HALVES THE CYCLE TIME OF CALCULATION
- TENSOR MEMORY ACCELERATOR ALLOWS ASYNCHRONOUS MEM ACCESS. REDUCES CYCLE BY ANOTHER 15%

MEMORY TRANSFER TO TENSOR CORE IS BOTTLENECK

https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/#8-bit_Float_Support_in_H100_and_RTX_40_series_GPUs

IMPORTANT SPECS FOR DL

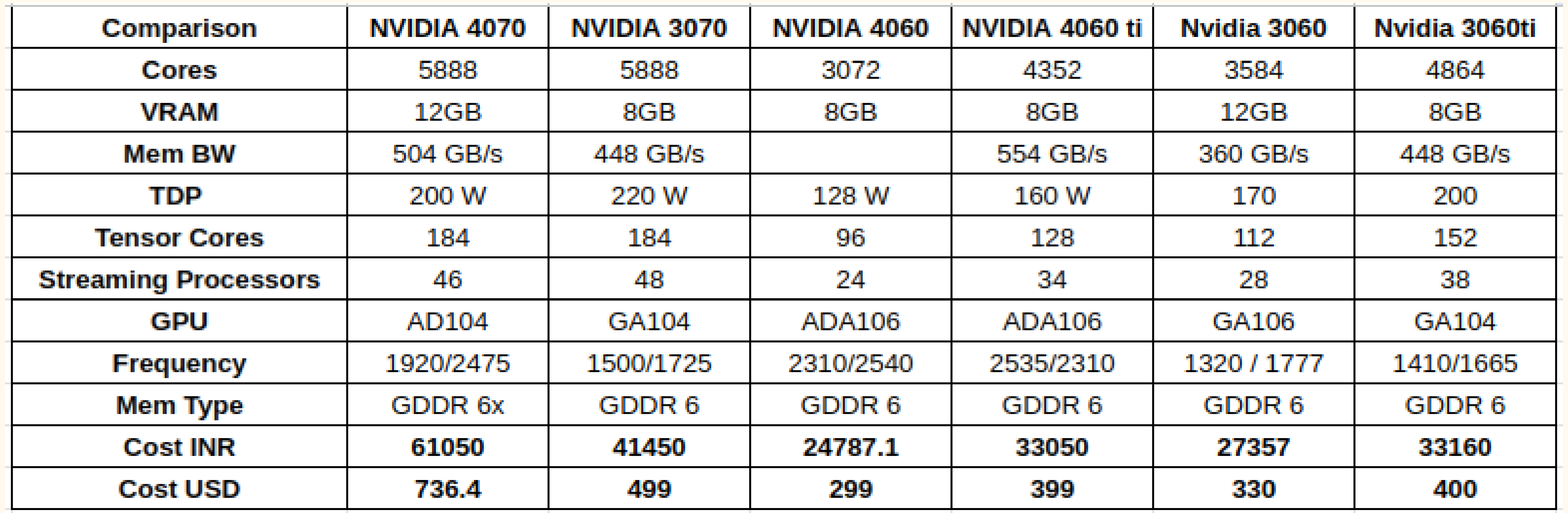
- TENSOR CORES : CORES THAT PERFORM MATRIX MULTIPLICATION(THIS IS MANDATORY)
- MEMORY B/W: SPEED OF DATA ARRIVAL TO TC
- CACHE HIERARCHY : HOW MEMORY MOVES
- FLOPS : FLOATING OPERATION PER SECOND

TILE SIZE IS DETERMINED BY HOW MUCH MEMORY WE HAVE PER STREAMING MULTIPROCESSOR (SM) AND HOW MUCH WE L2 CACHE WE HAVE ACROSS ALL SMS.

- VOLTA (TITAN V): 128KB SHARED MEMORY / 6 MB L2
- TURING (RTX 20S SERIES): 96 KB SHARED MEMORY / 5.5 MB L2
- AMPERE (RTX 30S SERIES): 128 KB SHARED MEMORY / 6 MB L2
- ADA (RTX 40S SERIES): 128 KB SHARED MEMORY / 72 MB L2

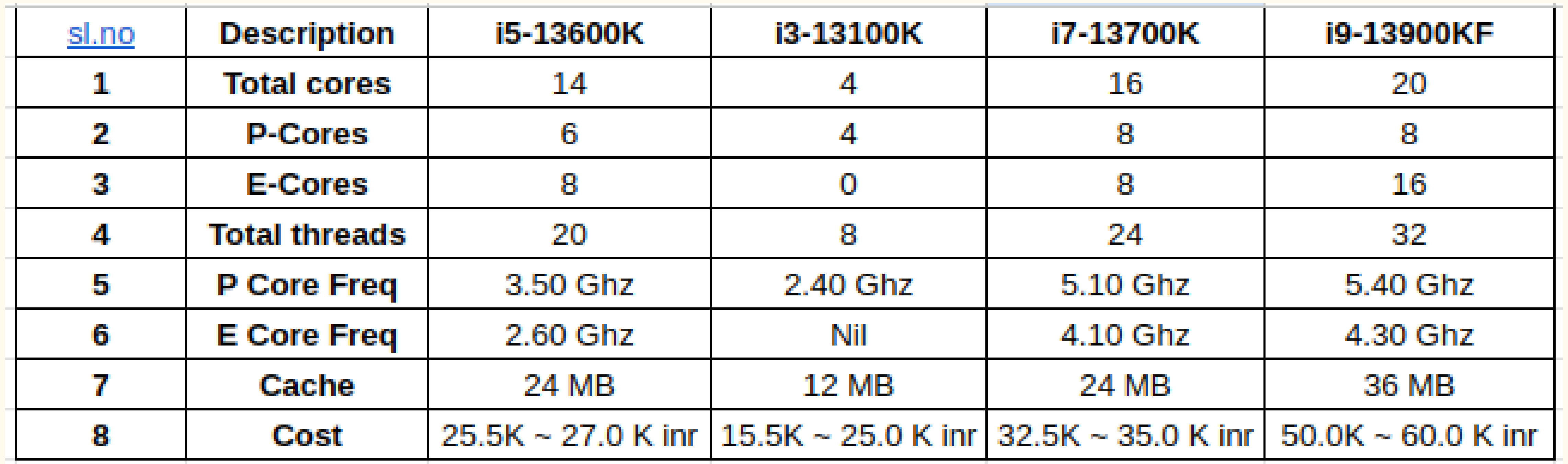
PS: TENSOR NVIDIA SNEAKED UNANNOUNCED PERFORMANCE DEGRADATIONS INTO THE “GAMING” RTX GPUS: (1) DECREASED TENSOR CORE UTILIZATION, (2) GAMING FANS FOR COOLING, (3) DISABLED PEER-TO-PEER GPU TRANSFERS. IT MIGHT BE POSSIBLE THAT THERE ARE UNANNOUNCED PERFORMANCE DEGRADATIONS IN THE RTX 40 SERIES COMPARED TO THE FULL HOPPER H100.

GPU COMPARISON



Comparison	NVIDIA 4070	NVIDIA 3070	NVIDIA 4060	NVIDIA 4060 ti	Nvidia 3060	Nvidia 3060ti
Cores	5888	5888	3072	4352	3584	4864
VRAM	12GB	8GB	8GB	8GB	12GB	8GB
Mem BW	504 GB/s	448 GB/s		554 GB/s	360 GB/s	448 GB/s
TDP	200 W	220 W	128 W	160 W	170	200
Tensor Cores	184	184	96	128	112	152
Streaming Processors	46	48	24	34	28	38
GPU	AD104	GA104	ADA106	ADA106	GA106	GA104
Frequency	1920/2475	1500/1725	2310/2540	2535/2310	1320 / 1777	1410/1665
Mem Type	GDDR 6x	GDDR 6	GDDR 6	GDDR 6	GDDR 6	GDDR 6
Cost INR	**61050**	**41450**	**24787.1**	**33050**	**27357**	**33160**
Cost USD	**736.4**	**499**	**299**	**399**	**330**	**400**

CPU COMPARISON



sl.no	Description	i5-13600K	i3-13100K	i7-13700K	i9-13900KF
1	Total cores	14	4	16	20
2	P-Cores	6	4	8	8
3	E-Cores	8	0	8	16
4	Total threads	20	8	24	32
5	P Core Freq	3.50 Ghz	2.40 Ghz	5.10 Ghz	5.40 Ghz
6	E Core Freq	2.60 Ghz	Nil	4.10 Ghz	4.30 Ghz
7	Cache	24 MB	12 MB	24 MB	36 MB
8	Cost	25.5K ~ 27.0 K inr	15.5K ~ 25.0 K inr	32.5K ~ 35.0 K inr	50.0K ~ 60.0 K inr

HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER