Build A Capable Machine For LLM and AI

Build A Dual GPUs PC for Machine Learning and AI with Minimum cost

[Andrew Zhu]

.

Published in

[CodeX]

.

10 min read

.

May 12

Background and Building Target

Both Stable Diffusion and offline LLM models require a huge amount of RAM and VRAM. To run and learn those models, I bought an RTX 3090 for its 24G VRAM. Actually, my aging Intel i7–6700k can still work well with a single RTX 3090, but when I throw another GPU like GTX 1070 or RTX 3070 TI. I frequently got blue screens, and I know it is time to move on.

You may ask, Andrew, aren't RTX 3090 powerful enough for running a model, why bother throwing another GPU into the case?

Here are several considerations.

1. External monitors and applications used up some portion of GPU's VRAM, but I want all RTX 3090 24G VRAM to be used by the ML model.
2. During the peak time, when an LLM model used up almost all VRAM, the computer will become laggy, the mouse can't move sometimes. And this is not unusual. even for RTX 3090.
3. Watch a 4k Youtube video while running a large model without sacrificing the GPU inference performance.

To address the above considerations, adding another GPU could be the best solution, even a GTX 1070 can handle those monitor and youtube tasks easily.

So, **target #1** is:

> Build a machine can run most of AGI/ML models while also function as a daily driver machine

It is easy to throw tons of cash to build whatever powerful machine, another **target #2** is:

> Build a new machine with minium cost

If you don't want to read through all those detailed considerations, you can scroll down to get my configuration list for your own building.

Besides, this machine is not for PC gamers, I have deleted all games for those storage-hungry model bin files. If you are looking for a gaming PC building plan, it is time to close the tab.

Consideration #1. Power

I had an 800W power supply. I thought it should enough for RTX 3090. I was wrong. Although RTX 3090's TDP is around 350w, sometimes, during some micro-seconds. its power usage can surge to rocket high, higher than the 800w PSU can sustain. What will be the result? PC turns itself off for the sake of self-protection.

There are many Youtubers who discussed this problem that comes with RTX 3080/3090. for example this one.

I tried to lower the voltage using MSI afterburner. it works most of the time but with some performance penalty, and I have to set GPU voltage limitation every time start the machine. Not to mention I am going to install another GPU into it. 800w is not enough this time, definitely.

I tried this 1600w PSU first. It looks good, it is cheap. But you get what you pay for.

HVVH High Power 24 Pin Miner/PC GPU ATX Fully Modular 1600W Power Supply Support Double CPU Mining Server and Computer Designed for US Voltage 110V 1600W

Brand: HVVH

3.8 ★★★★☆ 94 ratings | 12 answered questions

-12% $74.99

Was: $84.99

prime Two-Day

FREE Returns

Exclusive Prime price

May be available at a lower price from other sellers, potentially without free Prime shipping.

Eligible for Return, Refund or Replacement within 30 days of receipt | Free Amazon tech support included

Brand	HVVH
Compatible Devices	Personal Computer
Connector Type	ATX
Output Wattage	1600 Watts
Form Factor	ATX
Wattage	1600 watts

Several problems:

1. It is loud, it is so loud even without turning on any applications on GPU.
2. The cables that come with it are the worst quality ever.
3. It looks and feels dangerous.

To comfortably sit beside the machine and don't get myself burnt. I returned it back, replace it with this 1300w PSU

Roll over image to zoom in

ARESGAME AGT Series 1300W PCIE 5.0 Power Supply, 80 Plus Gold Certified, Fully Modular, 10 Year Warranty

Visit the ARESGAME Store

4.7 ★★★★½ 5,252 ratings | 119 answered questions

$169.99

prime Two-Day

FREE Returns

Earn 5% back ($8.49 in rewards) on the amount charged to your Prime Visa.

Style: AGT 1300W

AGT 1000W $139.99 prime	AGT 1300W $169.99 prime	AGT 850W $119.99 prime	AGV 500W $42.99 prime
AGV450	GL 1000W	SFX-GL850W $129.99 prime	

Brand	ARESGAME
Compatible Devices	Personal Computer
Connector Type	ATX
Output Wattage	1300
Wattage	1300 watts
Cooling Method	Air

It is good, looks and feels good, and most important: quiet. even quieter than my previous 800w PSU.

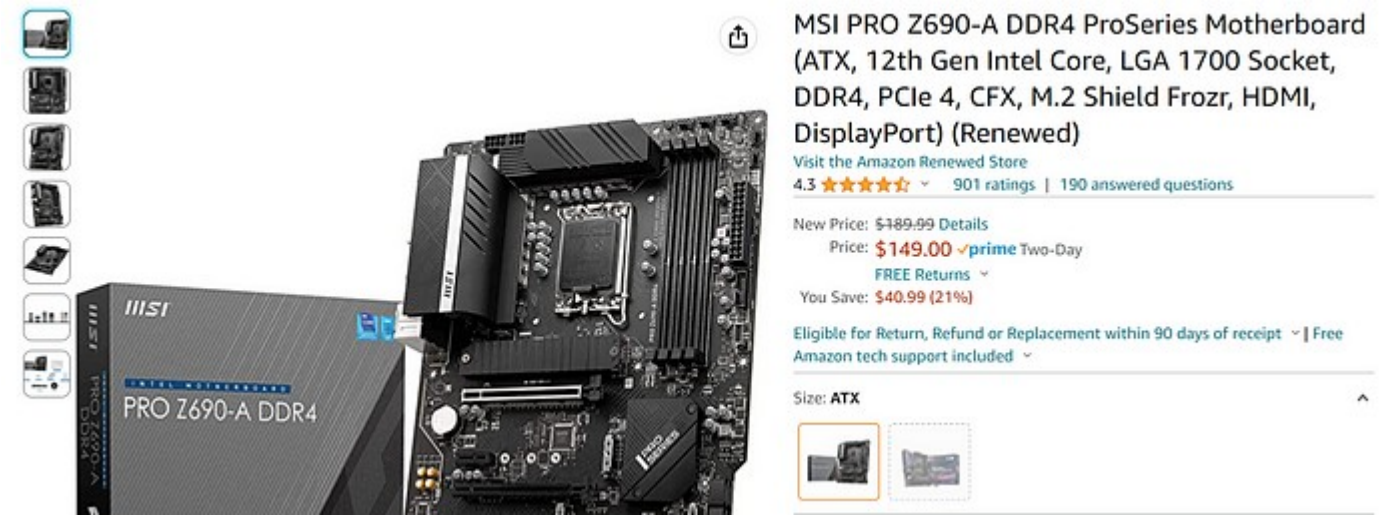
The PSU comes with good cables and even the newest PCIe power cable for RTX 40x0 GPUs.

Side note: don’t plug the PCIe power head into the CPU power socket.

Consideration #2. Motherboard

To install two GPUs in one machine, an ATX board is a must, two GPUs won’t welly fit into Micro-ATX. I am going to use an Intel CPU, a Z-started model like Z690 LGA 1700 can provide good PCIe lanes and bandwidths.

The first board I purchased is this renewed MSI Pro Z690-A DDR4 Pro

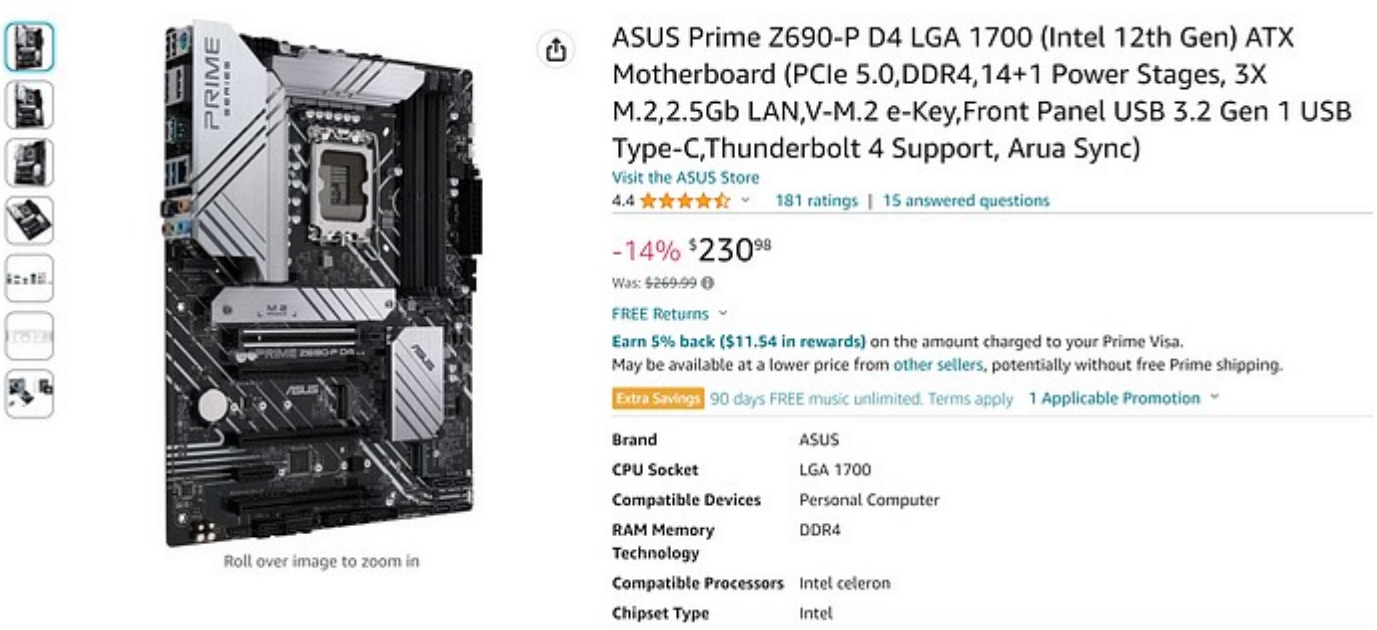


Its package comes with all screws and manual, everything looks like new. After plugging everything into it, press power on. It doesn’t START UP. Press and hold the power button to force the power off. you know what? it power on again by itself!

Later research shows this is a well-known issue for a certain batch of this motherboard. I can’t fix it by reconnecting the cable, by updating the BIOS, by Setting the BIOS configuration. nothing works.

Even worse, this motherboard simply doesn’t allow my computer to go sleep and restart. I have to force press the power button and it will reboot itself. This means, every day, I have to power off the machine and re-open every application the next day. This is unacceptable.

So I returned it back and purchased a renewed ASUS Prime Z690-P D4 for $128



The price is even lower than the MSI motherboard. It just works, no more restart or can't sleep issues. It just works well.

One other good about this ASUS motherboard is its manual, well written with detailed graphs to guide me to plug in every cable, while the MSI manual is a disaster.

Consideration #3. GPU

Which GPU to choose

In terms of running Transformer models, VRAM is a key factor that determines if a model can or can't run up in a machine. I am not talking about speed. You can't simply run up a model with not enough VRAM, for example, the StableLM from StableAI

GitHub - Stability-AI/StableLM: StableLM: Stability AI Language Models

StableLM: Stability AI Language Models. Contribute to Stability-AI/StableLM development by creating an account on…

github.com

Require more than 16G VRAM. In this case, maybe even RTX 4080 is not enough here, not to mention RTX 4070 with only 12G VRAM.

The only options are RTX 3090(TI) or RTX 4090, both come with 24G VRAM. But RTX 4090 is too expensive. While RTX 3090 used/renewed price looks good nowadays.

The Amazon GPU price collector I build:

http://www.zhusd.com/gpu

Reveals that RTX 3090 used cards are now around $780 on Amazon.com, less than half of a new RTX 4090.

Amazon GPU Prices

RTX 3090 Search all RTX 4090 RTX 4080 RTX 4070 RTX 3060 RTX 3060 Ti RTX 3070 RTX 3070 Ti RTX 3080 RTX 3080 Ti RTX 3090 RTX 3090 Ti 6600 6600 XT 6700 XT 6800 6800 XT 6900 7900

Data last updated on 2023-05-11 GMT-0700

Price	Brand	Chipset Brand	Chipset	Memory Size	Title
$783.99	ZOTAC		NVIDIA GeForce RTX 3090	24GB	ZOTAC GAMING GeForce RTX 3090 Trinity 24GB GDDR6X 384-bit 19.5 Gbps PCIE 4.0 Gaming Graphics Card, IceStorm 2.0 Advanced Cooling, SPECTRA 2.0 RGB Lighting, ZT-A30900D-10P
$799.99	MSI	NVIDIA	NVIDIA GeForce RTX 3090	24GB	MSI Gaming GeForce RTX 3090 24GB GDRR6X 384-Bit HDMI/DP Nvlink Torx Fan 3 Ampere Architecture OC Graphics Card (RTX 3090 Ventus 3X 24G OC) (Renewed)
$849	MSI	NVIDIA	NVIDIA GeForce RTX 3090	24GB	MSI Gaming GeForce RTX 3090 24GB GDRR6X 384-Bit HDMI/DP Nvlink Tri-Frozr 2 Ampere Architecture OC Graphics Card (RTX 3090 Gaming X Trio 24G) (Renewed)
$888.88	Gigabyte		NVIDIA GeForce RTX 3090	24GB	Gigabyte AORUS GeForce RTX 3090 Xtreme 24G Graphics Card, Max Covered Cooling, 24GB 384-bit GDDR6X, GV-N3090AORUS X-24GD Video Card
$950	ZOTAC		NVIDIA GeForce RTX 3090	24GB	ZOTAC Gaming GeForce RTX 3090 Trinity OC 24GB GDDR6X 384-bit 19.5 Gbps PCIE 4.0 Gaming Graphics Card, IceStorm 2.0 Advanced Cooling, Spectra 2.0 RGB Lighting, ZT-A30900J-10P

I bought this card for $766

The second GPU is RTX 3070 TI, purchased for $503 around six months ago. I found it using the above Amazon price collector I build. Until today, this GPU is even sold at a higher price.

Roll over image to zoom in

MSI Gaming GeForce RTX 3070 Ti 8GB GDRR6X 256-Bit HDMI/DP Nvlink Tri-Frozr 2 Ampere Architecture OC Graphics Card (RTX 3070 Ti Ventus 3X 8G OC)

Visit the MSI Store

4.6 ★★★★½ 177 ratings | 22 answered questions

-6% $749.00

List Price: ~~$799.99~~

Pay ~~$749.00~~ $738.63 after using available Amazon Visa rewards points.

Not eligible for Amazon Prime. Available with free Prime shipping from other sellers on Amazon.

Purchase options and add-ons

Payment plans

From $62.42/mo (12 mo) with 0% APR

Graphics Coprocessor	NVIDIA GeForce RTX 2070 Super
Brand	MSI
Graphics Ram Size	8 GB
GPU Clock Speed	1800 MHz
Video Output Interface	DisplayPort

Do I need to do anything to the renewed RTX 3090?

This card is powerful while also generating a lot of heat, and will quickly throttle by temperature. At around 83C, this card will run at 60% — 70% of its peak performance.

So, I opened this card, and repaste the GPU core, I didn't do anything with memory pads, which are widely discussed around the community. But what I found is that simply repasting the GPU core, can dramatically lower the temperature, making my newly purchased thermal pad spared.

If your GPU is thermally throttled, repaste may give you some surprise and make you believe in physics again.

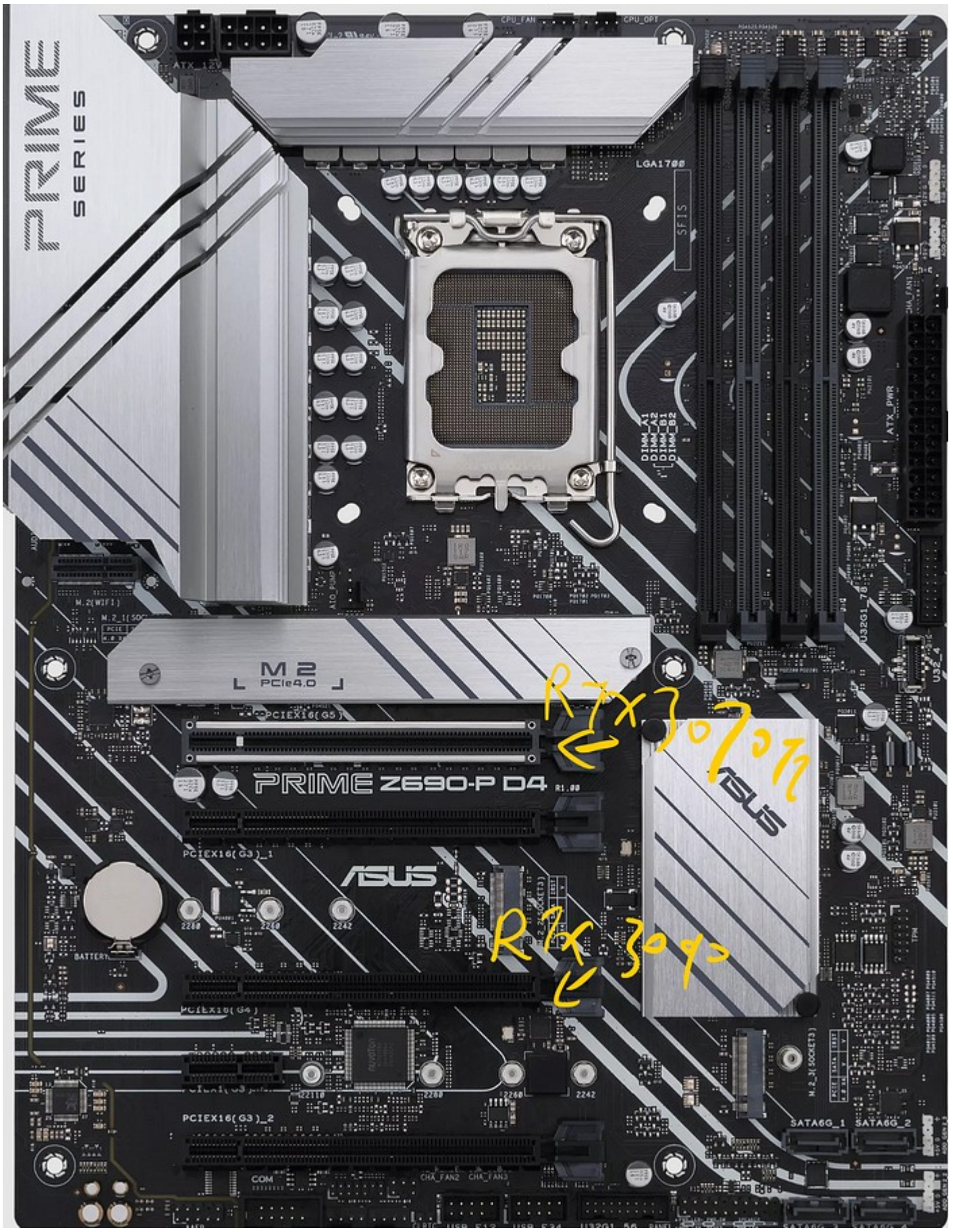
Which PCIE slot to install the RTX 3090?

Isn't it obvious to install the RTX 3090 in the first PCIE 5.0? (it is a PCIE 5.0 in the ASUS motherboard), yes, if it is a gaming machine that requires high throughput with the system RAM, while for machine learning, things are different.

During the Neruel Network inference, all model data will be loaded up to GPU VRAM at the beginning. there isn't that much data communication during the model run. In other words, there is almost no impact from the slot sequence.

For this motherboard, the first slot is a PCIE 5.0 x16, and the third slot is a PCIE 4.0 x4. I don't observe any differences when using RTX 3090 to generate Stable Diffusion images.

I install the RTX 3090 in the third slot for better cooling. The GPU installed in the first slot, its fan intake will be blocked by the second GPU.


PRIME
SERIES
PRIME Z690-P D4
M.2
PCIe4.0
PCIEX16(G5)
PCIEX16(G3)_1
PCIEX16(G4)
PCIEX16(G3)_2
ASUS
LGA1700
ATX_12V
CPU_FAN
CPU_OPT
SATA6G_1
SATA6G_2
RTX 3070 Ti
RTX 3090

Consideration #4. RAM

For a gaming machine, 32G RAM will be more than enough even for the latest 3A games. But for Machine Learning, oh man, 32G RAM is far away from enough. You'd better get at least 64G RAM and prepare to expand to 128G RAM.

One 7B LLM model will quickly use up more than 50G RAM during the loading up stage using PyTorch. Think about a 13B model, 128G RAM may not be enough then.

The good news is, nowadays, when I was writing this, the RAM price is at the bottom dirt cheap.

I bought 64G RAM for $115.

I don't know how much of the difference is between DDR4 and DDR5. Since DDR5 RAM doubles the data transition rate, DDR5 may benefit the LLM model loading speed. But won't boost the performance too much for CUDA inference.

Consideration #5. Storage or Hard Drive

Do prepare a large C drive if you are going to use Windows as the daily driver. By default, the hugging face package will download model data to a cache folder located in the C drive. So, if you give only 512G or 1T, you will only find your storage is quickly running out. Unless you are asking for trouble, prepare a 2T NVME SSD for the C drive. 4T is even better, but 4T will be more expensive.

I use this site to find the most affordable disk:

Disk Prices (US)

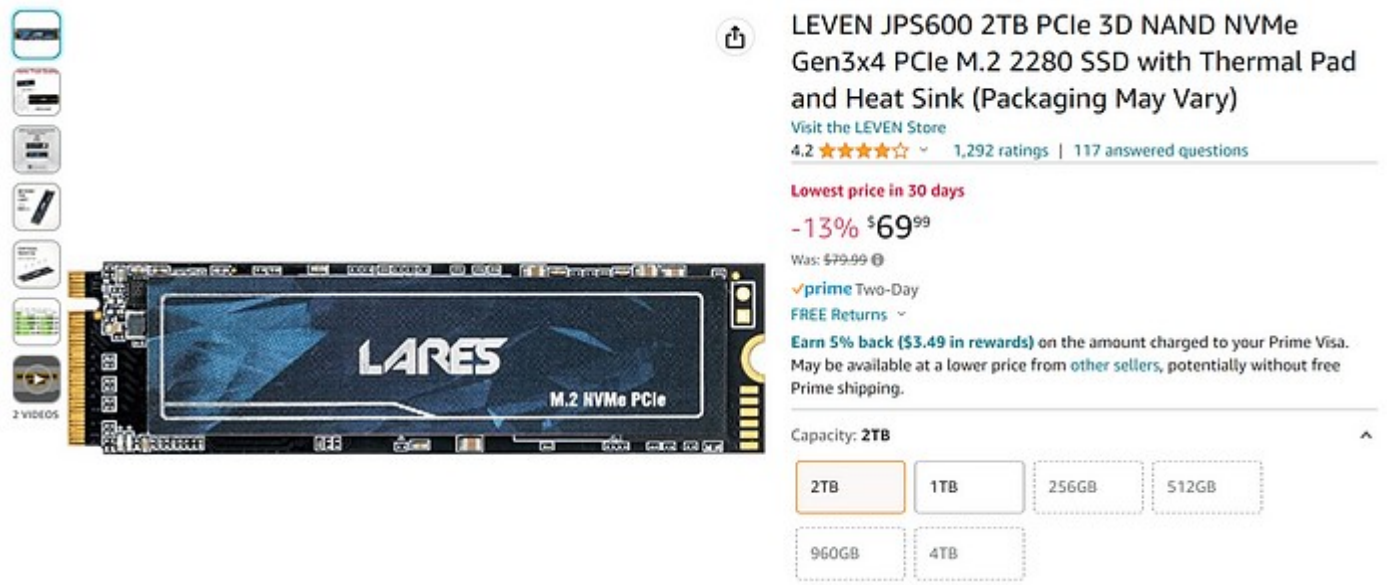
Comparison of all hard drives and SSDs on Amazon, sorted by price per TB

diskprices.com

Price per TB	Price	Capacity	Warranty	Form Factor	Technology	Condition	Name
$35.00	$70	2 TB		M.2	NVMe	New	Silicon Power 2TB NVMe M.2 PCIe Gen3x4 2280 SSD So
$37.50	$75	2 TB	5 years	M.2	NVMe	New	TEAMGROUP MP33 2TB SLC Cache 3D NAND TLC NVM Desktop TM8FP6002T0C101
$37.50	$75	2 TB	5 years	M.2	NVMe	New	PNY CS1030 2TB M.2 NVMe PCIe Gen3 x4 Internal Solid
$39.97	$40	1 TB		M.2	NVMe	New	Silicon Power 1TB - NVMe M.2 PCIe Gen3x4 2280 SSD (S
$39.99	$40	1 TB	5 years	M.2	NVMe	New	Timetec 1TB SSD NVMe PCIe Gen3x4 8Gb/s M.2 2280 3D Desktop (1TB)
$39.99	$40	1 TB	3 years	M.2	NVMe	New	fanxiang S500 Pro 1TB NVMe SSD M.2 PCIe Gen3x4 228 Desktops(Black)

And the NVME SSD price is lowering faster than the stock market. dirt cheap. Now may not be a good time to buy stock, but definitely, a good time to enlarge your storage.

I bought this 2T NVME for $70



It works well, PC restart is fast, and model loading is also super fast.

Consideration #6. CPU

Usually, the CPU should be the top priority consideration, but in the case of Machine Learning, the CPU could be the least important component. if not for the blue screen and unstable issue with my previous Intel i7–6700k, I don’t even want to upgrade the CPU. the new i5–12600k doesn’t make any performance difference in terms of CUDA inference. The same Stable Diffusion iteration speed.

There are some improvements, like faster Office application opening and faster VSCode Python IntelliSense speed. When using i7–6700k, editing a Python file over 1000 lines of code could be laggy sometime. but with i5–12600k. open anything is fast and smoother.

In terms of PC usability, The new CPU is worthy, in terms of Machine Learning, the new CPU doesn’t make any significant difference. The new machine is using this CPU.

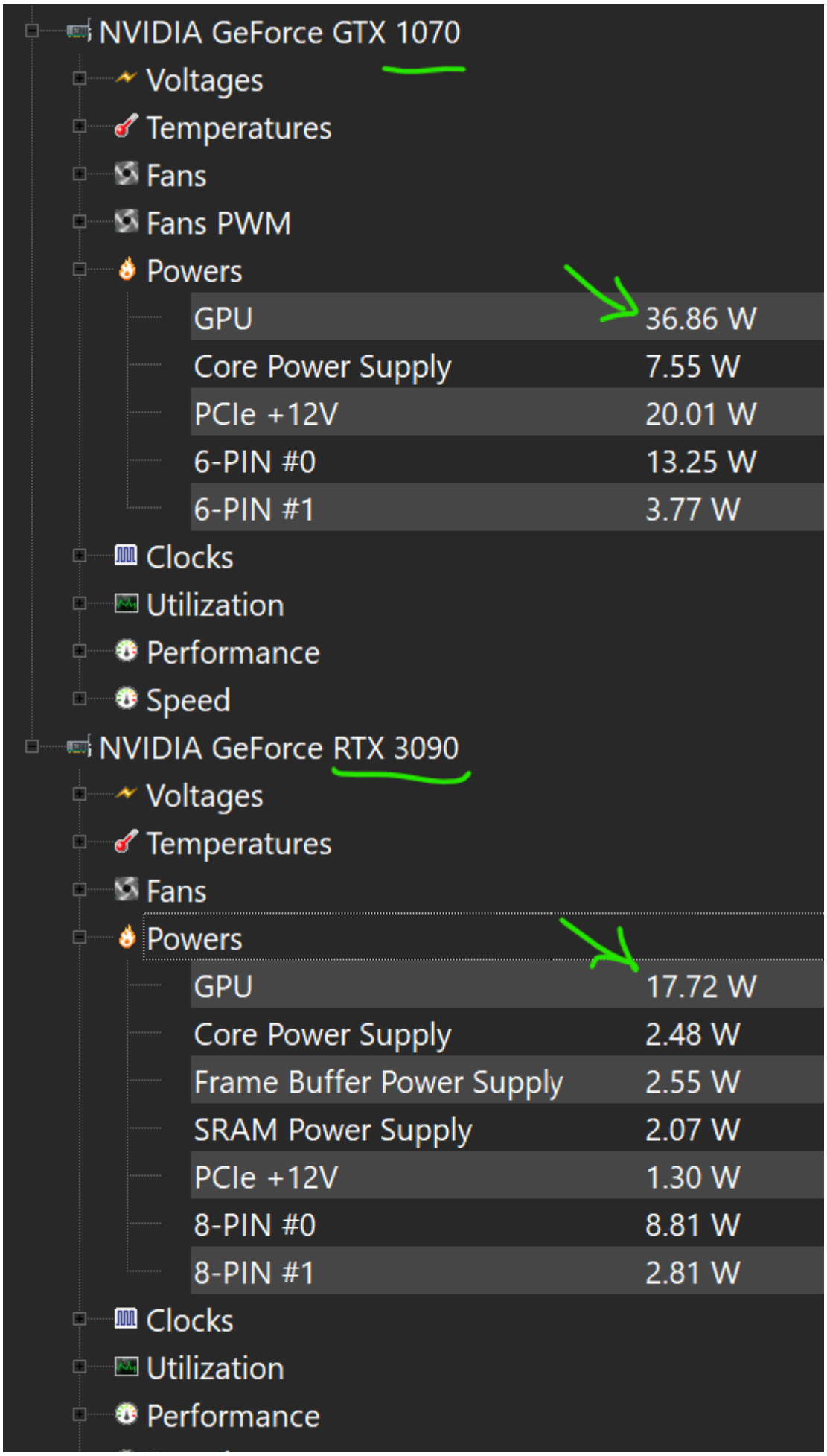
Overall speaking, the Intel i5–12600k is a good CPU, powerful while using moderate power. 12th i7 or i9 is way more powerful but are power-hungry monsters.

Result of the Building

Three 4k monitors are now plugging into the GTX 1070, running smoothly without a problem.

Why replace RTX 3070TI with GTX 1070? three reasons:

1. Two RTX GPU seems to disable Nvidia Broadcast, this is quite useful for video meetings.
2. Use the spared i7–6700k and motherboard to build another machine using the RTX 3070TI running Ubuntu.
3. GTX 1070 works pretty well with RTX 3090 together with low power consumption.


NVIDIA GeForce GTX 1070
Voltages
Temperatures
Fans
Fans PWM
Powers
GPU 36.86 W
Core Power Supply 7.55 W
PCIe +12V 20.01 W
6-PIN #0 13.25 W
6-PIN #1 3.77 W
Clocks
Utilization
Performance
Speed
NVIDIA GeForce RTX 3090
Voltages
Temperatures
Fans
Powers
GPU 17.72 W
Core Power Supply 2.48 W
Frame Buffer Power Supply 2.55 W
SRAM Power Supply 2.07 W
PCIe +12V 1.30 W
8-PIN #0 8.81 W
8-PIN #1 2.81 W
Clocks
Utilization
Performance

Now the machine drives daily work and Windows 11 using the less performant GTX 1070, and enables the powerful RTX 3090 to run AI models with its full 24G RAM without any performance penalty.

What is my usage experience?

1. The machine can sleep and wake up super fast.
2. The power consumption is acceptable, i5–12600k and GTX 1070 usually use about 60W power together when using VSCode, checking mail, web browsing, and even watching 4k youtube videos.
3. VSCode IntelliSense is now no longer laggy, fast, and responsive.
4. RTX 3090 together with 64G RAM is fully ready for almost 99% AI models you can download from huggingface and github.

One word to describe it: **PERFECT**.

Conclusion and Final List

How much do I spend to build up this Machine Learning ready machine?

```
cpu         = 234       # Intel i5-12600k
cooler      = 35        # CPU cooler
mobo        = 121.42    # ASUS Prime Z690-P D4 LGA 1700
power1300w  = 160       # ARESGAME AGT Series 1300W PCIE 5.0 Power Supply
ram64g      = 115       # Mushkin SILVERLINE Series – DDR4 Desktop DRAM – 64GB
(2x32GB)
rtx3090     = 766.47    # PNY GeForce RTX™ 3090 24GB
tax_rate    = 0.1       # WA Tax Rate
total       = cpu + cooler + mobo + power1300w + ram64g + rtx3090
tax         = round(total*0.1,2)
total_w_tax = round(total+tax,2)
print("Total before tax:",total)
print("Total tax:",tax)
print("Total after tax",total_w_tax)
```

You can run the above Python code to get how much I spend on this machine, it is less than a 16-inch Macbook Pro.

Wait, I didn’t talk about anything about the PC case, actually, that doesn’t matter, an open frame without a case is even better. When two GPUs are in full running, the heat is real, the computer will turn to a stove. Be careful, and ensure the case fans are running.

Hope my experience also helps you and wish you enjoy Machine Learning and AI.

Machine Learning
Pc Building
Llm
Gpu
AI

Written by Andrew Zhu

787 Followers

·Writer for

CodeX

Data Scientist@MS | Automate everything | https://xhinker.medium.com/membership | https://www.linkedin.com/in/andrew-zhu-23407223/ | https://twitter.com/xhinker