

Hi Everyone!

As a Hobbyist, the cost of EC2 Instances for running an experiment has been a barrier in exploring and solving Deep Learning Problems. Reserved Instances were my initial playground as i was not familiar with cloud ecosystem.

Eventually, ***Spot instances***becamemy alternative to run well structured experiments. But often times, it found it very difficult to setup and run experiments. The main problem comes when setting up the environment for backing up and restoring the data/progress. Thanks to [Alex Ramos](https://github.com/atramos) and [Slav Ivanov](https://medium.com/u/d41130ab0af4?source=post_page-----56bae813e34a----------------------) for the [Classic](https://github.com/atramos/ec2-spotter/tree/master/classic) and [24X7](https://github.com/atramos/ec2-spotter/tree/master/24x7) *versions of the EC2 Spotter tool* which came in handy when dealing with spot instances.( Try them out if you still use *Spot Instances*)

After using AWS EC2 instances for a around 6 months, I realized that the long term cheaper alternative is to invest on a local machine. This allows me to gain more by having better control over the experiment and with similar or better performance. On detailed survey throughout the internet, I couldn’t find any difference of opinion regarding the local machine idea when it comes to long term usage. Hence, I started to research on choosing components for my local deep learning build.

*Selection of components for Deep learning is a a huge puzzle that intrigues many beginners who try to get their build. It requires the user to have some basic knowledge to build a system that can meet the required performance for the cost involved.*

This post tries to help the fellow readers to get started with selection of components and understand the parameters before choosing the product.

**So! Lets get Started!!**

**First things First!** You must finalize on the **maximum number of GPU’s** that you plan to have on the newly built system. If you’re an active machine learning researcher then you might probably want more GPUs. This can help you run more than one task in parallel and try different variations of model architectures, data normalization, hyper parameters etc.. in parallel.

***My Recommendations:****If you are a researcher/Student/Hobbyist Consider for a Dual GPU Build. If you plan to run huge models and participate in insane contests like ImageNet which require heavy computation, consider for a Multi GPU Build.*

Once you have arrived at a conclusion on the type of build you can arrive at the number of PCIe lanes required:

1. Dual-GPU Build (Up to 2 GPU): 24 PCIe Lanes (Might Experience Performance Lag when using SSD that share PCIe lanes or when Both GPU)
2. Multi-GPU Build (Up to 4 GPU’s): 40 to 44 PCIe Lanes

***Why PCIe lanes first?****— In practice, there will be a bottleneck to keep data flowing to the GPU because of disk access operations and/or data augmentation. A GPU would require 16 PCIe lanes to work at its full capacity.*

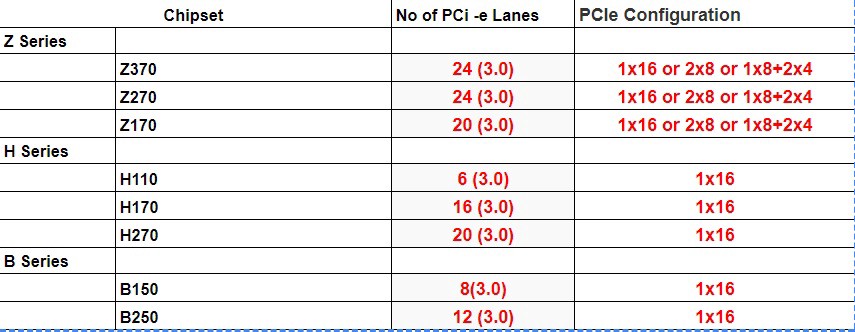
This post will address only about Dual-GPU System. There will be a follow up post about the Multi-GPU Build.

**Dual-GPU Build**

**1) Motherboard:**

Once the PCI-e Lane requirement has been decided, We can now choose the Motherboard Chipset:

The below table gives you the no of PCI-e Lanes available with different Chipsets available:



Comparison of PCI-e Lanes across different Chipsets (Mostly Intel Processor Based)

***Note:****Ideally a GPU, to perform at its full capacity requires****16 PCI-e Lanes.***

So, even though Chipsets like B150,B250, H110,H170,H270 support Intel processors, They are seldom used for deep learning builds since the number of PCIe lanes will not be enough for Deep learning applications.

Hence, chipsets that are commonly preferred are:

Z170 — Support both 6th/7th Gen Intel Processor. Usage of 7th Gen might require a BIOS Update.  
Z270 — Support both 6th/7th Gen Intel Processor.  
(Latest) Z370 — Supports 8th Gen Intel Processor.

**P.S:** Will update the post once i have enough details for AMD based Chipsets

Once you have decided on the chipset, Use PC Partpicker to select the motherboard : [Link](https://pcpartpicker.com/products/motherboard/) to select the motherboard of your choice.

**Things to Keep in Mind:**

1. Form Factor (i.e ATX, Micro ATX, EATX etc..)
2. No of PCIe Slots ( Minimum 2 Slots)
3. Maximum RAM Supported ( 64 GB Preferred)
4. No of RAM Slots (Minimum 4 Slots)
5. SSD and SATA Slots (if you is concerned)

**2) Processors:**

Through the selection of motherboards, We have narrowed down the choice of processor based on constraints like socket type, But the choice of CPU might further dependent on GPU. For Deep learning applications, As mentioned earlier, The CPU is responsible mainly for the data processing and communicating with GPU. Hence, The number of cores and threads per core is important if we want to parallelize all that data preparation. It is advised to choose a multi core system (Preferably 4 Cores)to handle these tasks.

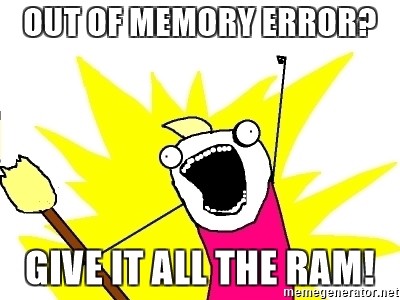
Things to Keep in Mind:

1. Socket Type
2. No of Cores
3. Cost

PS. Some processors may need the user to get their own Cooler Fan. Usually, Unboxed Processor doesn’t come with a cooler fan but allows the user to overclock.

Use PC Partpicker to select the Processor : [Link](https://pcpartpicker.com/products/cpu/#m=21)

**Memory or RAM:**

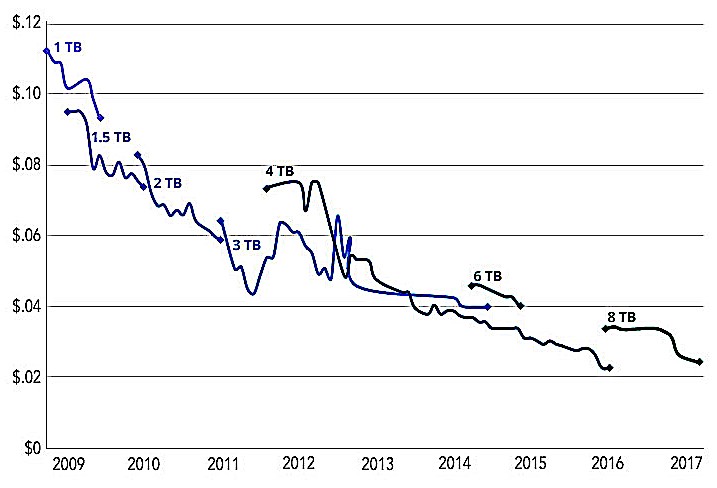


When working with large/big datasets we might need to have them in memory. Size of the RAM decide how much of dataset you can hold in memory. For Deep learning applications it is suggested to have a minimum of 16GB memory ([Jeremy Howard](https://medium.com/u/34ab754f8c5e?source=post_page-----56bae813e34a----------------------) Advises to get 32GB). Regarding the Clock, The higher the better. It ideally signifies the Speed — Access Time but a minimum of 2400 MHz is advised.

Always try to get more memory in a single stick as it will allow for further expansion in remaining slots.I have seen many people who get 4\*8 GB RAM instead of 2\*16 GB ending up using all 4 Slots and no room for upgrade just because they are bit cheap than the latter.

**Storage:**

The price of HDD is decreasing continuosly as SSD become more affordable and faster.



Source: Storagenewsletter

Its always better to get a small size SSD and a large HDD. SSD’s are preffered to store and retrieve data that is actively used. On the other hand HDD should be used to store data that are to be used in future.

SSD — Datasets in use + OS (Costly! **Min:** 128 GB Recommended)

HDD — Misc User Data (Cheaper! **Min:** 2 TB Recommended 7200RPM)

**GPU:**

GPU’s are the heart of Deep learning Build. They decide the performance gain that you get during training of neural networks. As most of the computation involved in Deep Learning are Matrix operations, GPU outperforms conventional CPU by running the same as parallel operations. They have small computation units called cores that can have threads which enable them to run the matrix operations faster. The Memory bandwidth of the GPU also enables to operate on large batches of data.

[Tim Dettmers](http://timdettmers.com/) has a great article on choosing a [GPU for Deep Learning](http://timdettmers.com/2017/04/09/which-gpu-for-deep-learning/), which he regularly updates as new cards come on the market. Please check them out before choosing your GPU.

*Couldn’t resist to repost from Tim Dettmers Post. His TL;DR advice for choosing GPU:*

***Best GPU overall (by a small margin)****: Titan Xp****Cost efficient but expensive****: GTX 1080 Ti, GTX 1070, GTX 1080****Cost efficient and cheap****: GTX 1060 (6GB)****I work with data sets > 250GB****: GTX Titan X (Maxwell), NVIDIA Titan X Pascal, or NVIDIA Titan Xp****I have little money****: GTX 1060 (6GB)****I have almost no money****: GTX 1050 Ti (4GB)****I do Kaggle****: GTX 1060 (6GB) for any “normal” competition, or GTX 1080 Ti for “deep learning competitions”****I am a competitive computer vision researcher****: NVIDIA Titan Xp; do not upgrade from existing Titan X (Pascal or Maxwell)****I am a researcher****: GTX 1080 Ti. In some cases, like natural language processing, a GTX 1070 or GTX 1080 might also be a solid choice — check the memory requirements of your current models****I want to build a GPU cluster****: This is really complicated, you can get some ideas*[*here*](http://timdettmers.com/2014/09/21/how-to-build-and-use-a-multi-gpu-system-for-deep-learning/)***I started deep learning and I am serious about it****: Start with a GTX 1060 (6GB). Depending of what area, you choose next (start-up, Kaggle, research, applied deep learning) sell your GTX 1060 and buy something more appropriate****I want to try deep learning, but I am not serious about it****: GTX 1050 Ti (4 or 2GB)*

I strongly recommend a beginner to get a 1060 6gb (New/Used) if they are on a budget. If the budget can go up a bit then you can get a 1070ti (MSRP around 430 USD) that was released recently *i.e.. OCT 26th* which offers almost the same performance as 1080 but at a lower cost (Almost Same as 1070). Don’t buy a 1070 unless you have a strong reason to, instead get a 1070ti as it has a greater number of cores. If you have enough money, then get a 1080ti. (No Second Thoughts). Again, if you are very active in performing research consider buying 2 X 1070ti instead of 1 X 1080ti as it gives flexibility that’s was discussed earlier.

**For readers wondering about different editions of GPU like Founder’s Edition, OC, FTW etc.**

Here’s the Info that you need:

**Difference between Editions:**Fundamentally all of them have the same GPU processor inside them. The main difference would be variation in quality of the PCB and usually high-end models would have higher binned chips (Best Quality).

**Difference between Brands:**Brands build their custom PCB components and aesthetics like lighting, Multiple fans, water cooled or back plate. These are done in order to improve the performance on the reference boards by just keeping the reference design on the card and add custom coolers on it. The base clocks out of the box matters very little generally.

**Water vs Air cooled GPU:** — Nvidia lowers the clock rate on your GPU as it gets hot. I don’t know if there are set temperatures that trigger this, or if it’s just linear. Water cooling will keep your GPU running at top speed.

Again! please research through the different editions. I have heard that FTW to be the coolest one to get. (Silent and No heating issues)

**PSU:**

Once the Components are selected using PCpartpicker site, it will give a rough estimate of power usage. It’s always better to get a PSU, large (1.2 to 1.5 times estimated power) enough to handle the power. In case if you plan to add more GPU(Add 100 W per GPU) then consider buying a PSU such that it can handle that requirement too. Some PSU tend to generate noise hence research on different products based on reviews before buying. I have found that EVGA G2 series seems to be solid option to consider.

**Note**: Gold, Silver, Platinum described along with the product refers to the efficiency of the PSU (Heat Generation). It directly correlates to power savings.

**Conclusion:**

For buying the components, I strongly recommend the reader to keep an eye on[**r/buildapcsales**](https://www.reddit.com/r/buildapcsales/)and[**r/hardwareswap**](https://www.reddit.com/r/hardwareswap/) (In US) for deals and grab them instead of buying them at retail price.**Open-Box** components seem to be cheaper and should be considered when on stringent budget. Try to get the components in instalments instead of getting them all at once if you are not in urgent need.

**Cheaper Alternative:** Try to get an Open-Box Pre-Built System and modify the components as per your requirement. (For Lazy Folks!)

The information described are based on my research and understanding from multiple articles and build guides from the internet. If there are any errors, please kindly notify me so that I can fix them. Feel free to contact me or comment below if have any questions. Thanks for reading!

1. Comparison of No of PCI-e Lanes across 200+ Chipsets (<https://www.pugetsystems.com/labs/articles/Z270-H270-Q270-Q250-B250---What-is-the-Difference-876/>)
2. Comparison of No of PCI-e Lanes across 100+ Chipsets ( <https://www.pugetsystems.com/labs/articles/Z170-H170-H110-B170-Q150-Q170---What-is-the-Difference-635/>)
3. Fast.ai Discussion Forum(Might require a Signup) (<http://forums.fast.ai/t/making-your-own-server/174/506>)
4. PC Part Picker: [https://pcpartpicker.com](https://pcpartpicker.com/)
5. Choosing GPU for Deep learning (<http://timdettmers.com/2017/04/09/which-gpu-for-deep-learning/>)
6. The $1700 Deep Learning box (<https://blog.slavv.com/the-1700-great-deep-learning-box-assembly-setup-and-benchmarks-148c5ebe6415>)

**FAQ & Miscellaneous :**

CPU Lanes vs PCI-e Lanes:

PCIe lane denotes the *maximum* bandwidth that is available for graphics cards’ communication with the CPU. Having more lanes than you need won’t increase performance, you just don’t want to have so few that it starts restricting CPU/GPU intercommunication. Generally an x8 lane of PCIe 3.0 has more than enough bandwidth for any gaming card, so 16 lanes for dual cards or 24 lanes for triple cards is fine. In applications outside of gaming, such as when the GPU is being used to accelerate CPU computation for workstations and servers, there is a lot more communication between the CPU and GPU than in games, so 40 lanes might be helpful there. The X99 platform is derived from Intel’s server/workstation chips, so that’s why they have so many lanes.

Source: [Link](https://linustechtips.com/main/topic/203664-i-dont-understand-the-difference-between-28-and-40-pci-lanes/)