tags
Nanopore, GPU

GPU musings (with an eye on genomics)

Author: Miles Benton (GitHub; Twitter)
Created: 2021-21-01 23:15:32
Last modified: 2021-06-21 21:49:55

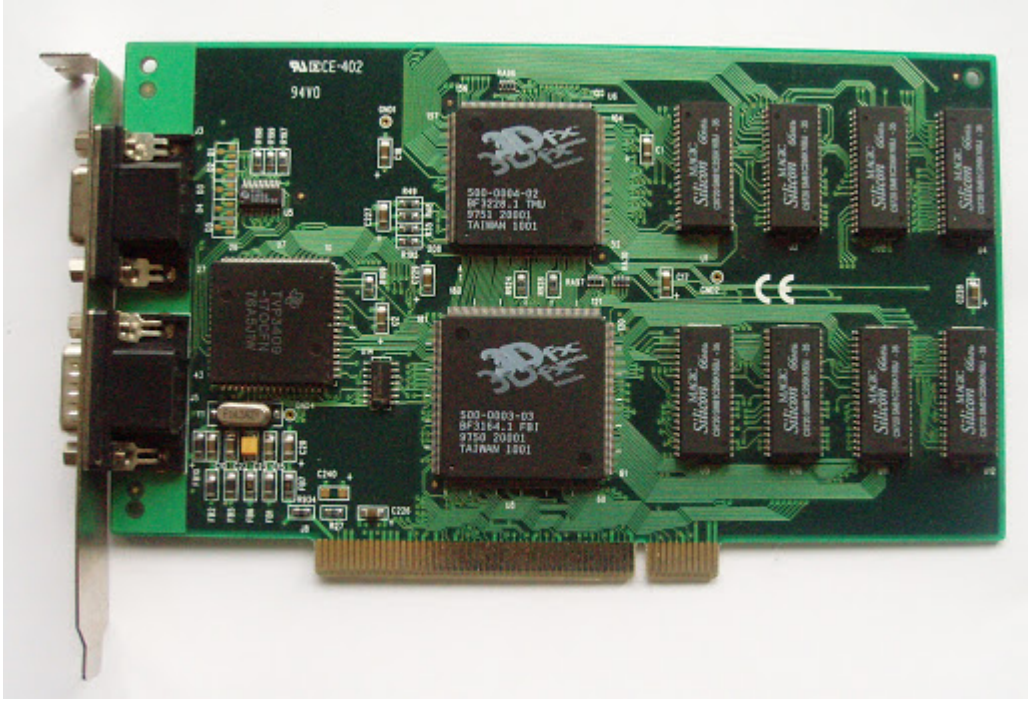
:::warning ATTENTION: this document is an ongoing work in progress. :::

For some time now I have been meaning to put together a collection of my notes, thoughts and experiences of GPU computing from the last few years. I'm going to be contextualising this mainly around the use of GPUs in genomic data processing and analysis, predominantly in the form of base calling sequence data genereated by Nanopore sequencing.

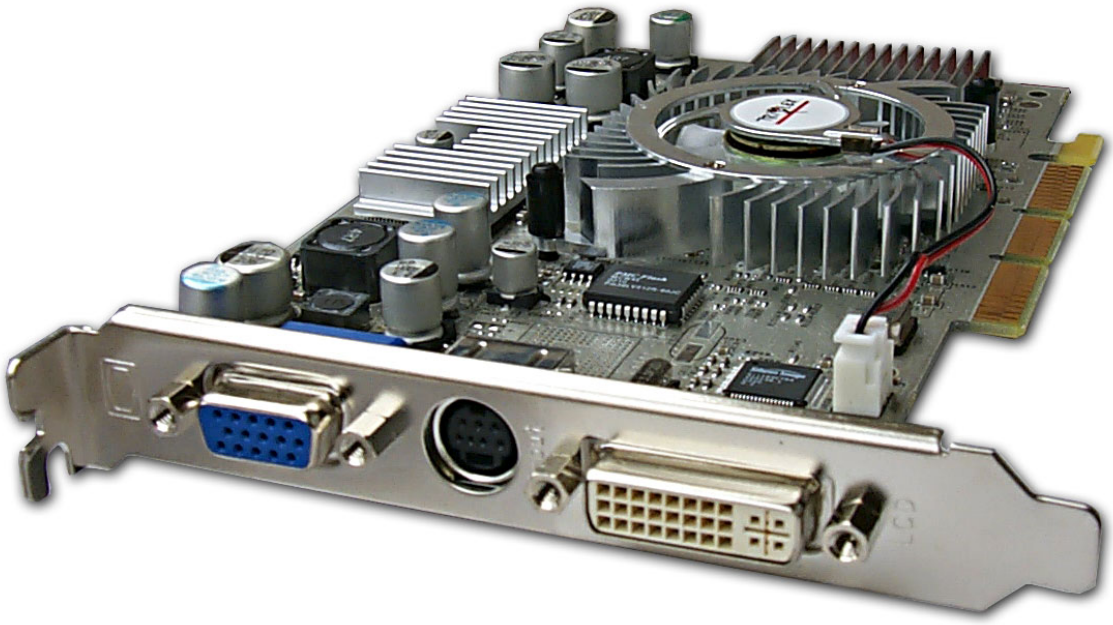
Preamble

:::warning **WARNING**: Please feel free to skip ahead to a section that might be of more interest. Putting together this resource brought up a few old memories that I decided to document here in the preamble. :::

I'm unashamedly a massive hardware geek and have long had a fascination with graphic card technology. Before Nvidia and AMD were the major players there was a company called 3dfx, makers of the revolutionary Voodoo graphics cards. I remember vividly, saving up, buying and installing a first gen Voodoo card in the family PC, Quake 2 never look so good!



I eventually upgraded to the original Nvidia Geforce 256 which saw me through several years of gaming until I was able to nab a Geforce 4 4200 Ti (this was an amazing card for the price).



The Geforce 4 was to be the last card that I would buy as a year or so later I was off to university and didn't have the time or money to dedicate to supporting a computer hardware 'habit' (it was a struggle supporting an electric guitar habit!). However I still kept a very close eye on the technology that was coming out, and was able to somewhat statisfy the tech sirens call by building computer systems for family and friends. I have fond memories of installing dual Geforce 9800 cards in SLI in a cousins PC and geeking out at all the extra performance and benchmarking numbers!

Apart from these brief encounters with the actual technology I never came close to having a desktop computer during my undergrad, masters, or PhD. It wasn't until I got a very nice Dell Precision laptop during my first PostDoc position that I was able to physically get my hands on some GPU power of my own again. This was my first real introduction into CUDA and various other GPU compute resources. I was extremely interested in just how people were starting to use what had been traditionally gaming technology to dramaticaly speed up 'data science' jobs. I recall being at a genetics conference on the east coast of Australia in 2012 (ish) when I was able to chat with a GPU developer from Nvidia. He had earlier that day presented some of the work he and his team had been doing using GPU compute to speed up problems such as BLAST and sequence alignment. It was incredibly exciting and while I would dabble over the next few years it wasn't until I came to my current position that I would get to really dig into the world of GPUs again.

I came to the nanopore scene later than many (mid 2018), right at a time when I was starting to dabble with GPU usage in other areas of genomics (Googles deepvariant). I was really lucky when I started my new job that there was a big push for data science and bioinformatics, including a lot of new hardware. So I got to start by 'playing' on one of the most powerful GPUs at the time, the Nvidia Tesla V100 - it is still a very powerful card, but technology moves fast!

In another stroke of luck I was able to obtain a very nice GPU for my dedicated Linux server, the Titan RTX (pictured below). Having unfetted access to such GPU power (with the ability to break things and not have it affect other users...) has been very benefical.

It was at this stage that I was approached by some work colleagues who were experiencing slow base calling times on the laptop being used for MinION sequencing. It was a fairly good 24hr run generating a decent amount of data (~20Gb). It was taking far too long on the laptop CPU and they had heard that GPU calling was starting to gain traction. So I grabbed the raw signal (fast5) data, did a quick bit of reading, installed Guppy (then version 3.1.5) and started 'playing'. Long story short, and probably no surprise to anyone now, it was MUCH faster. Using the fast basecalling model on the Titan RTX it was completed in ~45 mins. I then wanted to see how the V100 went, turns out is was about the same speed (I'll touch more on this later but I need to revisit the benchmarking and optimise parameters for each card). If you're interested in the detailed benchmarking results you can view them here.

Soon after this we actually ended up with a second V100, so I decided to see how Guppy scaled. At that stage it didn't, it wasn't natively multi-GPU aware (it is now though). So I just split the data in two and ran it across both cards in parallel. To be expected it halved the time, down to ~23 mins. This was very exciting. I also tested the high accuracy calling model (HAC) and was very impressed with the results (see the above link). From this point onwards our MinION data was generated in the lab and then transferred over to our linux cluster for GPU basecalling and further analysis, it was a large improvement and made all involved very happy.

:::info **Note**: I intend to revisit these benchmarking results. I now understand a lot more about the various paramters etc, so I ultimately would like to start a collection of either GPU specific parameters, or ideally, GPU specific Guppy models. I envisage these being placed in a GitHub repository making it easy to version and disseminate them. :::

It was also about this time that I became aware of the Nvidia 'answer' to single board computers (e.g. like the Raspberry Pi). This family of system on module (SOM) compute units was known as Jetson, with various entries; TX1, TX2, Nano, Xavier AGX and most recently Xavier NX. The amazing thing about these tiny compute units is that that have full blown GPUs based on various Nvidia microarchitectures (more on these and why they're important soon). Plus they cost next to nothing when compared to similar spec'd laptops/computers.

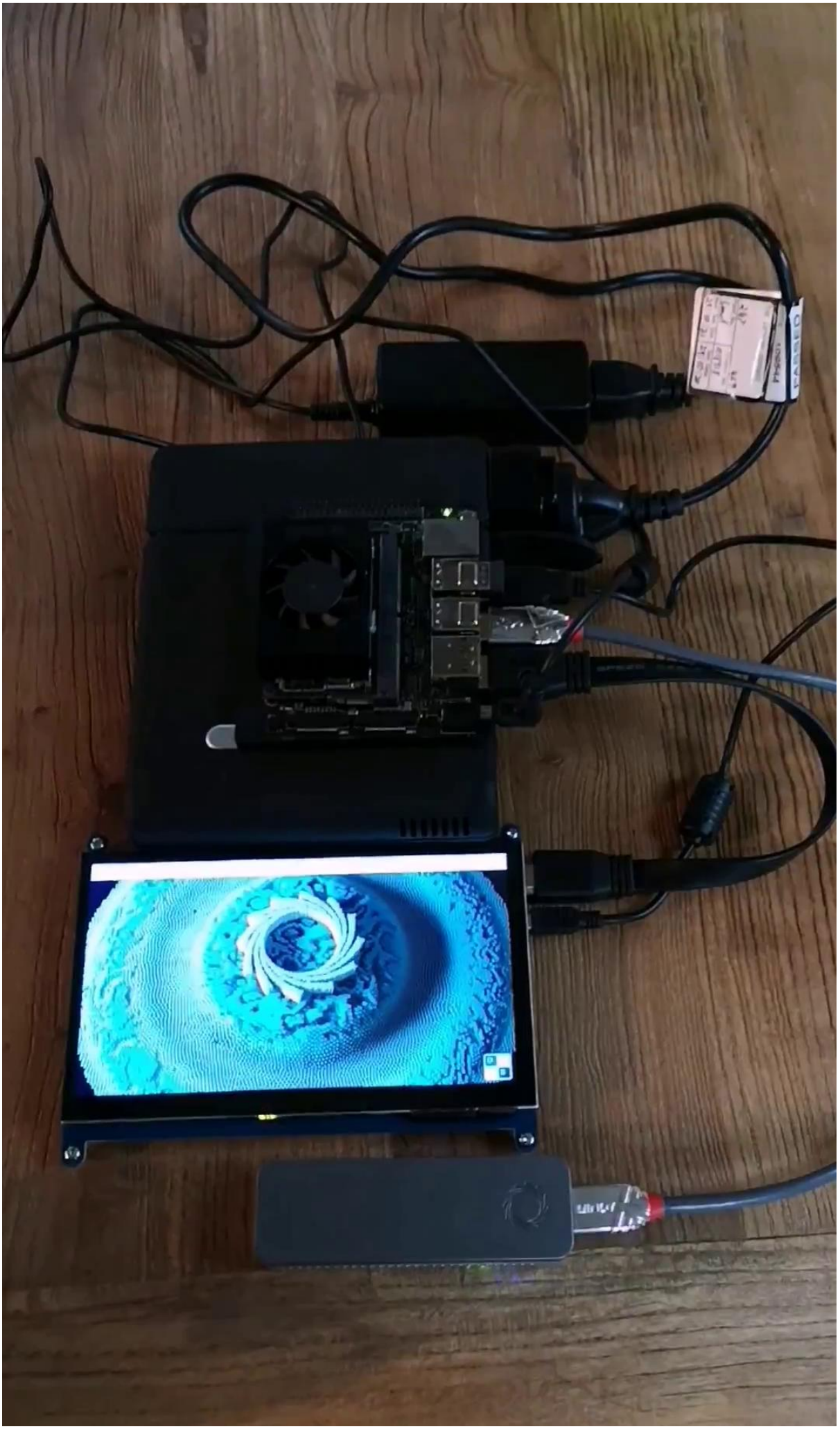
Putting two and two together I got really excited about the potential of pairing a MinION alongside an affordable and portable Jetson board. It also dawned on me that ONTs hardware, the MinIT (and now Mk1c), are running a Jetson TX2 at their heart. So this observation confimed to me that it must be possible.

:::info **INFORMATION**: ~~I'm not 100% sure it has ever been officially announced that~~ the compute unit in both the MinIT and Mk1c is in fact the Nvidia Jetson TX2. ~~However it is stated on the ONT website that they officially support the TX2, and if you look at the system specs (here) it's not hard to see that the compute stats match that of a TX2 exactly.~~

MinIT Specifications:

- Pre-installed software: Linux OS, MinKNOW, Guppy, EPI2ME
- Wi-Fi enabled; you can control your experiments using a laptop, tablet or smartphone
- fastq or fast5 files are written to Onboard storage: 512 GB SSD
- **Processing: GPU accelerators (ARM processor 6 cores, 256 Core GPU), 8 GB RAM.**
- Small footprint, 290g
- 1 x USB 2.0 port, 1 x USB 3.0 port and 1 x Ethernet port (1 Gbit capacity) :::

Another stroke of luck, I was at our institutes conference and happend to be sitting next to our then CTO. During dinner conversation I dropped in how incredible these little Jetson things from Nvidia were and some of the potential I saw for their use. Long story short, this recently became a reality and we've come up with a solution to get the whole MinKNOW 'stack' running on Nvidia Jetson boards. It turns out they make excellent little sequencing machines, we have several Xavier NX's connected to 24 inch touch screens in various labs and they haven't missed a beat. I have lot's more coming up in this space but that's something for future me to document. It's really exciting, has an international community contributing, and if you would like to know more you (and even do it yourself) can check out the GitHub repository here. Here it is in action on my dining room table at home:



So that's a brief bit of background around how I came to be working with GPUs and Nanopore data in my 'spare' time. Moving into the rest of this document I want to start detailing particular aspects of GPU usage and Nanopore sequening. Please read on to hopefully learn more.

GPUs and Guppy base calling

In this section we'll dive into exactly what GPUs will and won't work for Guppy base calling, as well as how various specifications may impact the performance and results. I'll also touch on various 'gotcha' points, such as taking software compatibility into account when setting up a Linux system for live base calling.

Guppy / MinKNOW compatibility

It's a topic that doesn't get enough discussion in my opinion, but the issues that arise between ONT software versions are rather 'fun' to deal with. I thought I was on top of it from all the work I've done with the Jetson boards. However, when I recently came to setup MinKNOW on my Linux workstation (amd64) I spent too much time troubleshooting exactly why live base calling was causing the whole thing to error out.

It turns out that while I was using the latest Guppy release (4.4.1 at the time), this version appears not to be compatible with the current release of MinKNOW (20.10.3). When I dropped down to Guppy 4.2.2 everything worked as expected. This led me to the community forum, where I found a lot of posts talking about similar things, and a set of instructions for setting up a GPU on Linux for live calling. In my opinion those instructions are a bit ambiguous, with a lot of the required information scattered across various posts about MinKNOW versions and patches. As such, I started to compile a list of known working combinations. **NOTE:** I have also created a set up guide of my own for getting up and running with live basecalling on Linux here.

The below table is a start at documenting the known versions of MinKnow, Kingfisher UI and Guppy that play nicely together. I will try and keep this updated as much as I can. *Please note that there will be slight version differences in MinKNOW betwen the various devices (MinION Mk1b, Mk1c, MinIT, GridION, PromethION), but the underlying packages are usually the same.*

MinION Release	MinKnow Core version	GUI version	Guppy version working

MinION Release	MinKnow Core version	GUI version	Guppy version working
18.12.4	3.1.8	3.0.13	1.8.7
18.12.6	3.1.13	3.0.13	1.8.7
18.12.9	3.1.19	3.0.16	1.8.10
19.05.0	3.3.2	3.3.16	3.0.3
19.06.7	3.4.5	3.4.12	3.0.4
19.06.8	3.4.8	3.4.15	3.0.7
19.10.1	3.5.5	3.5.10	3.2.6
19.12.2	3.6.0	3.6.14	3.2.8
19.12.5	3.6.5	3.6.16	3.2.10
20.06.4	4.0.4	3.5.10	4.0.9
20.06.5	4.0.5	4.0.21	4.0.9
20.06.17	4.0.5	4.0.21	4.0.11, 4.0.14, 4.0.15
20.10.3	4.1.2	4.1.22	4.2.2, 4.2.3
21.02.1	4.2.5	4.2.8	4.3.4
21.06.0 (21.05.8 MinIT)	4.3.4	4.3.20	5.0.11

- for earlier releases and a lot more information see here.

Wait, what about Ampere?!

From Guppy 4.4.X onwards ONT are supporting Ampere (RTX3000 series/A100) cards. At the moment the Guppy 4.4.1 patch seems to be working for people with these kinds of GPUs. This made me wonder how people are doing adaptive sampling (ReadUntil) with Ampere cards if Guppy 4.4.X is not currently compatible with MinKNOW. It turns out that there is a special patched version of Guppy 4.2.2 available in the downloads section of ONT's community space. If you have access you can see more details here, including the download link [***NOTE**: I am not allowed to directly link precompiled versions of ONT tools/software*].

***UPDATE:** as of Guppy 5.0.11 (which is paired with MinKNOW 21.06.0) Ampere patching is no longer an issue.*

Guppy base calling models

There are now three types of base calling model built in to Guppy, fast (FAST), high accuracy (HAC), and as of Guppy 5.0.7 the super-accurate model (SUP). Without going into detail the clue is in the name:

- the fast model is tuned to deliver an optimised accuracy while being efficient (keeping up with data generation).
- the high accuracy model provides higher consensus/raw read accuracy over the fast model. However it is 5-8 times slower than the fast model and it's much more computationally intensive. Leis just say that unless you want to wait weeks/months, you're going to want to perform high accuracy calling on GPUs.
- "super-accurate" (SUP) DNA models have higher accuracy than HAC models at the cost of increased compute time. These models take approximately three times as long to run as a HAC model, but offer an accuracy improvement over HAC in exchange.

ONT recommend running the fast model when live basecalling as if you have a suitable GPU it will be able to keep up with the generation of the sequence data, see this quote:

> *"The Fast Flip-flop model includes a simplified version of the Flip-flop algorithm and delivers the best level of accuracy that is achievable while keeping up with data generation on all devices. Both models have been trained on the same dataset."* From ONT

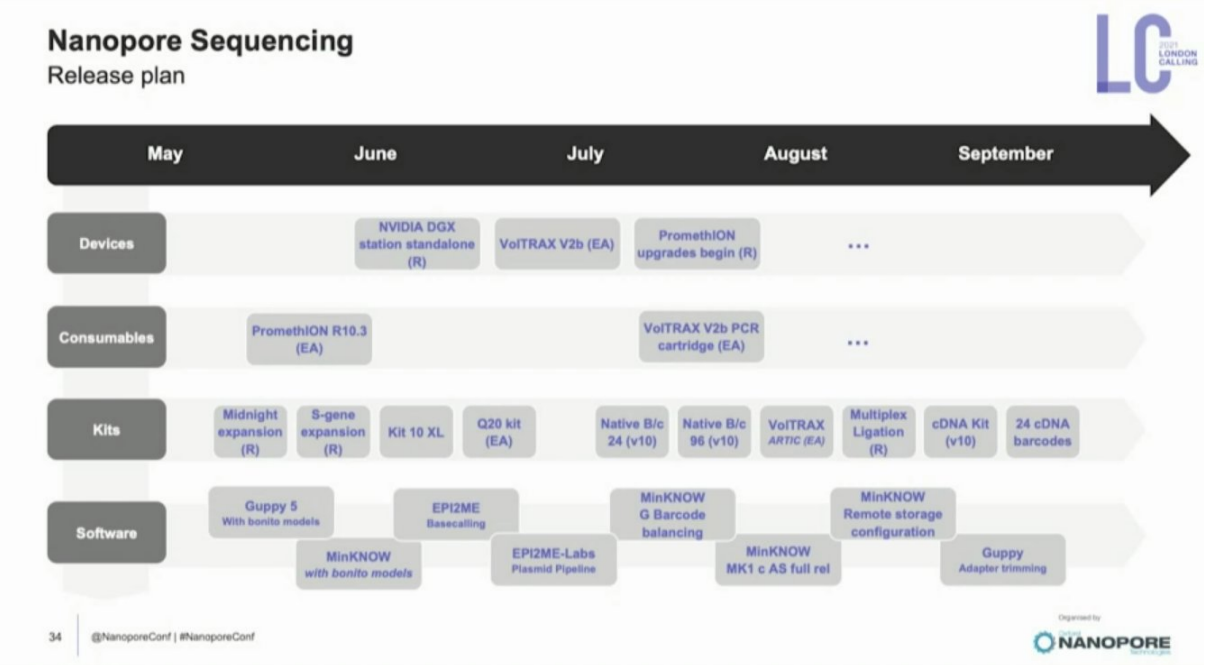
If you've tried basecalling without a GPU you have a feel for just how slow it is. Hopefully what's to come will help out with that.

UPDATE: with the "right" (more powerful) hardware it is possible to live basecall with the HAC model, and potentially the SUP model. I noted a twitter user stating that they can livebasecall two MinION Mk1b's in parallel on a gaming laptop with a mobile 3060. This seems crazy to me, but once I'm able to grab a new GPU or two I'll be able to start testing these sorts of things.

Bonito models in Guppy

Bonito is a PyTorch basecaller for Nanopore Sequence data. It is currently geared towards research, has the ability to train custom models, and generally runs slower than Guppy basecalling, though more accurate.

More recently models that have been trained in bonito are able to be run in Guppy, taking advantage of bonito accuracy and Guppy performance. ONT have mentioned on numerous occasions that Guppy and bonito will continue integrating, most recently at the May 2021 London Calling event (see RoadMap below).



(source: https://twitter.com/AlbertVilella/status/1395408587548155908?s=20)

Nvidia microarchitectures

One of the major issues that arises when people start exploring using GPUs and Guppy for base calling is the actual type of graphics card itself. First off, as mentioned above, Guppy is currently heavily implemented around CUDA, and CUDA is an Nvidia inititive. So that rules out any AMD (or soon to be Intel?) GPUs from being an option.

The next thing is that obviously not all Nvidia cards are made equal. In the world of technology things progress extremely quickly, this is very much the case with GPU fabrication and microarchitectures - this is really just tech jargon for all the good stuff under the hood. Here is a list of the major second wave of Nvidia microarchitectures and their year of release:

- Tesla - 2006 (wikipedia link)
- Fermi - 2010 (wikipedia link)
- Kepler - 2012 (wikipedia link)
- Maxwell - 2014 (wikipedia link)
- **Pascal** - 2016 (wikipedia link)
- **Volta** - 2017 (workstation/datacenter) (wikipedia link)

- **Turing** - 2018 (consumer) (wikipedia link)
- **Ampere** - 2020 (wikipedia link)

There is a lot of exciting technological advances in each of these generations of GPU micoarchitecture, but for the purposes of Guppy base calling the major 'gottcha' is that there is a CUDA compute compatabily requirement. This requirement is **CUDA compute version >6.0**, which has been present in cards from the Pascal generation onwards (2016+). I have seen this cause a lot of grief in the community as there are some still rather decent GPUs in the previous generations - I'm looking at you Nvidia Titan X (Maxwell 2.0) which is CUDA compute version 5.2 - that will not work with Guppy as provided by ONT.

Now we do know that if one has access to the Guppy source code (can be obtained by signing a developer agreement with ONT), then it can be modified and recompiled to work on early CUDA compute versions. I'm not sure what the implications of this are, if any, but I understand that ONT had to draw a line in the sand somewhere for supporting GPU architectures. So at this point in time that means that if you want to GPU accelerate Guppy you need to ensure you are using an Nvidia GPU from 2016 onwards (Pascal or newer) with CUDA compute >6.0.

:::warning **IMPORTANT**: I mentioned this in text, but it's so important I'll add this note as well. **The Microarchitecture from Pascal onwards (bolded above) is CUDA version >6.0, which is the current requirement for ONT Guppy.** :::

Which GPU should I get?

Now that we've ascertained that we are looking for a GPU with CUDA compute 6.0 or higher there is still a lot of options on the table. So here is an attempt to whittle this down further. These are a few considerations that I believe help narrow in on a particular GPU:

There is no "one size fits all" / price is an important consideration

Just as there is a wide variety of GPUs there is also a wide variety of use cases. For the vast majority of people, dumping thousands of dollars on a workstation/datacenter grade GPU is going to be a complete waste of money. For the price of some of these cards (P100, V100, A100) you could easily build/buy a very good computer for MinION sequencing, base calling and analysis and still have change in your pocket. So straight off the bat unless you know you are in the market for a GPU such as the A100, and you should know - if you require a GPU of this calibre you're going to be doing much more than base calling - then the best option in my opinion is the current generation Nvidia gaming cards.

Just because they're branded gaming doesn't mean they can't work

As of the time writing, Nvidia have released their Ampere GeForce RTX 3000 series cards. These GPUs are incredible in terms of there spec's and performance. Here is a quick summary of the current 3000 series lineup:

'Gaming' cards

GPU	CUDA cores	Memory	Launch Price
NVIDIA GeForce RTX 3060	3584 cores	12GB	329 USD
NVIDIA GeForce RTX 3060 Ti	4864 cores	8GB	399 USD
NVIDIA GeForce RTX 3070	5888 cores	8GB	499 USD
NVIDIA GeForce RTX 3070 Ti	6144 cores	8GB	599 USD
NVIDIA GeForce RTX 3080	8704 cores	10GB	699 USD
NVIDIA GeForce RTX 3080 Ti	10240 cores	12GB	1199 USD
NVIDIA GeForce RTX 3090	10496 cores	24GB	1499 USD
UPDATE: *table updated June 1^st^ 2021 with 3070Ti/3080Ti card information.*			

Any and all of these cards are more than capable of fitting the bill, and we'll discuss the features that make them different from each other and how these may impact base calling performace below.

'Workstation'/Server cards

For those that are interested here is a list of the current Ampere generation of Quadro/Tesla replacements (those names have been retired). Remember, these are the EXPENSIVE cards, most users are going to be better off with the above gaming range of GPUs.

GPU	CUDA cores	Memory	Launch Price
NVIDIA RTX A4000	6144 cores	16GB	1000 USD (expected)
NVIDIA RTX A5000	8192 cores	24GB	2250 USD (expected)
NVIDIA RTX A6000	10752 cores	48GB	4500-5000 USD
NVIDIA RTX A10	9216 cores	24GB	2800 USD (expected)
NVIDIA RTX A16*	2560 x4 cores	16GB x4	?
NVIDIA RTX A30	3584 cores	24GB	?
NVIDIA RTX A40	10752 cores	48GB	4500 USD (expected)
NVIDIA A100	6912 cores	40GB	~8500 USD
NVIDIA A100	6912 cores	80GB	~12000 USD
UPDATE: *table created Feb 16 2021, updated May 2021, to show other Ampere GPU information.*			
*The Nvidia A16 combines four graphics processors to increase performance. It features 2560 shading units, 80 texture mapping units, and 48 ROPs, per GPU.			

One large issue at the moment is that the supply chain is under extreme strain. With the global pandemic, coin miners and such a successful product, it is currently extremely difficult to get a hold of the 3000 series cards. Due to the fact that OEMs and other companies get preference over individual consumers it is currently easier to buy prebuilt systems with these cards in them. Nvidia are increasing manifacturing and these supply issues will resolve, but it is something to take into account at the moment.

:::warning **Note**: there was a compatibility issue with Ampere cards and Guppy, but that has been addressed and they are reported working as of the most recent versions of the software. **Update**: this has been fixed as of Guppy 4.4.X onwards. :::

As an aside, the previous generation of gaming cards are still completely reasonable (the Turing based RTX 2000 series) and will perform just fine. It just gets harder to recommend such GPUs when the prices of the newer generation are not a lot more, but the performance gains are huge. If you can't wait to secure a 3000 series card, you can find a great deal on an older GPU, or you inherit one, then by all means it will likely fit your needs. Read on to discover what sort of differences you may see with varying spec'd GPUs.

More CUDA cores generally means faster (in the basecalling world)

With the state of current GPU usage in Nanopore base callers it's quite fair to say that "CUDA cores are king", at least at the moment. So it is something that should be prioritised when deciding upon a GPU. This is one of my major reasons behind recommending the latest RTX 3000 series, Nvidia really doubled down on the number of CUDA cores present on the Ampere chips.

For example, in the table above you'll notice that the RTX 3060 Ti has 4864 CUDA cores and is currently priced at around $399 USD. Meanwhile the 'replaced' card from the previous generation at the same launch price point, the RTX 2060 Super, has 2176 CUDA cores. I haven't seen any benchmarks of the RTX 3060 Ti but I suspect that it is easily going to benchmark much much better than the 2060 - I'd hesitiate a guess that it should show nearly twice the performance in Guppy basecalling.

Following this logic we can consider cards of a similar CUDA core count but very different price points. Again let's use the RTX 3060 Ti, and this time compare it with the Titan RTX (a card that I have). The Titan RTX has 4608 CUDA cores and launched at $2500 USD. Now these cards will perform very differently in different tasks, and the Titan has much more RAM, BUS width and tensor cores, however when it comes to Guppy basecalling there is going to be very little difference between their performance. In this case it's much more sensible to go for the gaming card at $399 USD and save the extra $2100 that would have been spent on the Titan RTX and put it towards other aspects of a computer system.

Saying all this above, there is absolutly nothing wrong with choosing an older model card in the knowledge that it will take a little longer for basecalling - at the end of the day a GPU like the RTX 2060 is going to be light years ahead of CPU calling and it will easily keep up with live basecalling. So if your budget allows it, selecting something from the RTX 3000 series is a good choice, but there are still really good options in the older generation cards (picking up a 2080 or 2080 Ti for a bargin would be a really good outcome still).

Now if you have the budget and you know you want to strech it towards a very powerful GPU, something like the RTX 3080 or the phenominal RTX 3090, then it's going to open up options such as being able to run multiple Guppy instances and thus multiple MinIONs at once. It will also mean that running high accuaracy calling models will be much much faster, which is always nice.

GPU memory can make a difference

Depending on the model and parameters used with Guppy you can see quite a large variation in the amount of memory used. There is a bit of a misconception out there that the more GPU memory used the better the card is performing, this is not true. Generally I believe 8GB is probably the 'sweet spot', but this is obviously subject to change. Even on our cards with large amounts of memory (Titan RTX = 24GB, V100 = 32GB) I rarely see the memory occupied more than 3-6GB using the fast base calling models. This does change substantially when using th HAC and SUP models - depending on the GPU and any model 'tweaks' it's quite easy to push a card to it's memory limit.

It's going to be interesting as we see more and more models and methods from callers such as Bonito integrated into Guppy to see how memory allocation and usage scales.

My suggestion here is that most use cases will be satisfied with a GPU having 6 or 8GB of memory. If you can find the 3000 series cards even the 'lower end' 3060 now comes with 8GB of memory on board.

GPUs confirmed working with Guppy

Below is a list of Nvidia GPUs that I know work successfully with Guppy. This is based off a combination of personal experience, collaborators experience and from what I have read in internet and ONT community forums. Some of the below will work better than others, I will add notes where I can.

Discrete GPU cards

- 1050 Ti [Pascal] - (confirmed in ONT community forums)
- 1060 (6GB RAM version) [Pascal] - (confirmed in ONT community forums)
- 1070/1070 Ti [Pascal] - (confirmed in ONT community forums)
- 1080/1080 Ti [Pascal] - (confirmed in ONT community forums)
- RTX2060 [Turing] - (confirmed in ONT community forums)
- RTX2070 [Turing] - (confirmed by personal collaborator)
- RTX2080/RTX2080 Ti [Turing] - (confirmed by personal collaborator)
- RTX3060 [Ampere] - (confirmed by personal communication)
- RTX3070 [Ampere] - (confirmed in ONT community forums)
- RTX3080 [Ampere] - (confirmed in ONT community forums)
- RTX3090 [Ampere] - (confirmed by personal collaborator)
- Titan V [Volta] - (confirmed by personal collaborator)
- Titan RTX [Turing] - (confirmed personally)
- Tesla T4 [Turing] - (confirmed personally, via Google Colab)
- Tesla P4 [Pascal] - (confirmed personally, via Google Colab)
- Telsa P100 [Pascal] - (confirmed personally)
- Tesla V100/V100s [Volta] - (confirmed personally, also the GPUs in PromethION)
- Nvidia Quadro RTX4000 mobile [Turing] - (confirmed personally, I have a HP Zbook Fury)
- NVIDIA Quadro GV100 [Volta] - (this is the GPU in the GridION)
- A6000 [Ampere] - (confirmed in ONT community forums)
- A100 [Ampere] - (confirmed personally and in ONT community forums)

:::info **NOTE**: I don't have any experience with mobile (laptop) GPUs, but there are many posts in the community forum of users confirming GPU acclerated basecalling on cards such as 2060/2070/2080 plus the MAX-Q and Super models of these mobile GPUs.

I seem to recall one or two users saying they were using something like the mobile 1050/1060, however I would not recommend laptops running these cards as they will be right at the lower limit of what is supported by more recent versions of Guppy.

UPDATE: as of the 10^th^ June 2021 I have some experience with mobile GPUs. We got a HP Zbook Fury G7 17 at work, which has a Quadro RTX4000 mobile GPU. This thing is no slouch and you can check out my experience getting GPU basecalling working on Windows as well as some benchmarks **here**. :::

Others (SOM Jetson boards)

- Jetson Nano [Maxwell] (with custom compiled version of Guppy)
- Jetson TX2 [Pascal]
- Jetson Xavier NX [Volta]
- Jetson Xavier AGX [Volta]

:::info **NOTE**: all above Jetson boards have been confirmed working by me personally, with the TX2 and Nano boards being via personal collaboration on our ***"Nanopore on Jetson"*** **'project'**. I have also included the GPU microarchitecture for interests sake. :::

Cloud GPU compute for Guppy basecalling

:::warning ... **Under construction:** please check back soon ... :::

We have demonstrated that GPU accelerated basecalling can be performed at very nice speed using the free tier of Google Colab (essentially notebooks that have GPU runtime environments). For more information see my notes here.

Suggested computer requirements

:::warning This 'guide' is not a "how to" build a computer, it is an atempt to highlight what is important for having an enjoyable Nanopore sequencing experience. It aims to give the reader an idea of the types of components and specifications they should be looking for in prebuilt systems. If you want to DIY this that's great, but this isn't the place to learn all about assembling a computer. :) :::

In this next section I will try and address one of the most common questions I see/get asked, *"what computer should I get for nanopore sequencing?"*.

The answer to this question is not simple and is directly linked to another question, *"what are you wanting to do/use this computer for?*. The answers to this question are going to be quite different from person to person or lab to lab, but could include a combination of the below:

- MinION sequencing
    - live basecalling
    - adaptive sequencing (Read Until, Readfish)
    - high accuracy basecalling
    - super high accuracy basecalling (updated as of Guppy 5.0.7)
    - detection of base modifications (such as methylation)
- additional bioinformatic analysis (i.e. genome assembly, variant annotation and analysis, various pipelines ...)
- running multiple MinIONs off the same computer system at once
- software and pipeline development
- is portability a concern?

These are some potential examples and there will be many more use cases. Deciding exactly what it is you will be doing and potentially wanting to do in the near future will help guide what type of computer system will suit your needs.

So now we need to figure out based on our needs/requirements what components should be a priority.

Sum of the parts

This document is mainly focused on GPUs, but there are obviously many components that make up a computer. Each has an important role to play and each will offer something different to various use cases. For example, if you are interested in running multiple MinIONs in parallel then a much more powerful GPU should be a priority. Or if you are wanting to also use the computer as your labs analysis machine then making sure you have a decent number of CPU cores and system RAM is a must.

If you just want to run a MinION and generate sequence data but will then basecall and analyse this elsewhere then a laptop could possibly be a cheap option (more on this soon).

My current thinking around prioritising components is:

1. **GPU** - this is currently going to have the largest impact on basecalling (whether you can do it live or not, whether you can do HAC in a reasonable time frame)

and features such as adaptive sequencing. **If you want to do live basecalling and 'fast' high accuracy basecalling then you want a good GPU.**

2. **CPU** - most modern CPUs should be fine, but it's always nice to have more cores. 4 cores / 8 threads should be minimum and not blow a budget. With Intels i7 and i9 range you are able to get much more capable processors, but AMDs Ryzen CPUs should not be overlooked. The Ryzen 5/7/9 processors are extremely powerful and cost effective, and if you have the money, ThreadRipper is incredible. If you plan on doing more than just Nanopore sequencing with the computer you should try to go for the most powerful CPU you can fit in your budget.
3. **Fast solid state storage (SSD)** - You can never go wrong with fast storage, plus the increase in I/O can make a decent difference in base calling. SSD drives are getting cheaper and cheaper, so if the budget allows add more. One thing to consider is the read/write speed of the drive(s). I recommend going for at least 1TB in size, NVMe format and as fast as you can afford - I've had great experience with the Samsung 970 EVO Plus (link), but there are lots of options.
4. **a decent amount of system RAM** - RAM is cheap so get what you can afford. 16GB is minimum*, 32GB is better, 64GB or more is very useful if you plan to use this machine for various other bioinformatics analyses.
5. ... anything else you may want on top of the above.

**this is now a good tip when buying any computer - 16Gb is really the new standard and should be easy enough to pick up in most situations.*

:::danger **WARNING**:The motherboard is crucial, but outside the scope of this current discussion. I'm going to assume that if you are building a custom machine you know what you are doing and all about how important selecting the correct motherboard is. :::

:::info **Note**: 3 and 4 above are probably interchangable from my current observations. Both are cheap hardware so it should be fairly easy to obtain a good amount of RAM and at least 1TB of fast SSD storage for your build. :::

So now we've broken that down it becomes a balance between what you want to do with the machine and how much money you're willing to spend on it. There is no point blowing the bulk of a budget just on a GPU and not having a system that can keep up performance wise.

So a general purpose recommendation would be something like:

- GPU: Nvidia RTX 2070/2080/3060 (or any RTX 3000 series card)
- CPU: either a 6 or 8 core (12/16 thread) Intel i7/i9 or AMD Ryzen 7/9/ThreadRipper

- SSD (solid state hard drive): >= 1TB NVMe M.2 SSD with fast I/O (read/write)
- RAM: 32GB of supported memory (dictated by motherboard and CPU)

There are also some nice to haves:

- you might find extra SSD and HDD storage useful if you are generating a lot of data
- if you are 'building' a computer through a company such as Dell/HP/Lenovo, and you are comfortable using Linux, then try and get them to exclude the Windows install. This can usually save you a few hundred dollars.

So now we have an idea around what we're looking for in terms of components, it's time to consider form factor, do we want a laptop, a desktop, a server, or something else?

Laptops

Laptops are an attractive option for those that want an 'all in one' solution, and maybe a little portability - though it is arguable exactly how portable a laptop is when you consider that their batteries won't last long when sequencing and live basecalling. So they are 'portable' in the sense that you can easily move them from lab to lab, or to home and back again. Or you could sequence at a site that was powered. So that covers off portability.

My main reason for suggesting that people look at desktop computers over laptops is the cost. I find it hard to recommend laptops simply because you pay more for less, i.e. you end up paying for parts that have been scaled down to fit into a small space, and they have had their performance limited to fit with thermal regulations. Whereas when you get a desktop PC you don't face these restrictions and get much more performance for the same, or even less, price. In general its also much easier to upgrade a desktop machine, you can't really upgrade things like GPUs in laptops.

Saying this there are some really nice laptops on the market at varying price points.*

**I might look at putting together a list of laptops that people are using, as a starting point for people interesting in purchasing a laptop. I personally don't have experience using a laptop for Nanopore sequencing, but this may change in the very near future... again, watch this space.*

Desktops

Unless you have your mind set on a laptop or equivilent, then a desktop machine is going to be the best option. There are so many options in this space, so I'm not going to say much here other than the fact that if you are careful in selecting components taking into account all the information above then you'll end up with a computer that will easily perform all the tasks that you require (and often more).

In terms of selecting desktop PCs there are several options; building your own, purchasing from an OEM (DELL, HP, Lenovo, MSI, ...) or buying ex-lease and upgrading components as required.

Building a custom PC is an excellent option if you have the interest and are happy to get your hands 'dirty'. It really isn't difficult and you end up with the best bang for buck, as well as knowing exactly what is in your system. Buying from an OEM is the easier option, and most offer customisation options, so this can also be a good choice. It also fits in well with most institutions purchasing policy's - it is generally much harder to get ones work to pay for a custom gaming box as opposed to a customised machine from someone like DELL or Lenovo.

In terms of buying 'old' off-lease hardware, or even being given older hardware, this is a viable option. I have spoken to many people in the community that have been given or saved machines from disposal etc., and after adding a little bit of RAM, a faster drive and a GPU they have quite cheaply built themselves a very capable machine. So if you want to reduce ewaste and upcycle, it's a great option!

Below I will further breakdown specifics of laptops and desktops.

Suggestions

:::info **Info**: this is a section that I would like to continually add to. So feel free to check back every now and then. :::

I recently was asked by a local institute what I would recommend for them to get started with Nanopore sequencing and ReadUntil. They were interested in laptop options, so as you'll see I included a couple, but I also highlighted what you can get in a desktop for the same money. **Please beware that the below costings are in New Zealand dollars and not representitive of prices internationally. Saying that, it is still a fair reflection of the spec's and price differences between laptops and desktops.**

:::info Below is an excerpt from an email response with my suggestions for their specific use case. :::

My current recommendation for a 'decent' set up to perform MinION sequencing (including adaptive sequencing/readuntil, live base calling, high accuracy calling, etc) would be a desktop PC/server. As to exactly how "spec'd" out such a machine is, it really comes down to how much you want to spend. A lot of people like to use gaming laptops, but I personally believe you get much more performance for the same price, or sometimes less, from a desktop machine. My usual advice is to look at the price of such a gaming laptop and then use that as a rough budget for building a desktop PC. So let's take a look at a couple of models from PbTech:

- something like this laptop at $3400 NZD would be the minimal specs in my mind. The mobile GPU is great and more than up to the task, but the processor (4 core 8 threads) and RAM (16Gb) are below what I would I would want to be running. It will happily sequence and live base call, and probably do some ReadUntil work, BUT what I'm seeing is that if you're running ReadUntil with large references (i.e Human on my end) then RAM usage starts to get north of 24Gb. So for RAM I recommend 32Gb as a minimum. In terms of CPU (processor), 4 cores / 8 threads would be minimum and I recommend more.
- this laptop at ~$6000 NZD is quite a lot better (8 cores / 16 threads, 32Gb RAM, fast SSD, very powerful mobile GPU) but it's also nearly twice the price.

So let's look at some potential desktop options, none are more expensive than the most expensive laptop and all 3 are far more powerful:

- Intel i9 'gaming' desktop with Nvidia RTX3080 - $4943 NZD
- Ryzen 7 'gaming' desktop with Nvidia RTX3070 - $3793 NZD
- Intel i7 'gaming' desktop with Nvidia RTX3070 - $2758 NZD

Each of the above are very well spec'd and it's very important to remember that desktop GPUs =! mobile GPUs. So a 3080 RTX is far more powerful than a 3080 RTX MaxQ, compare them between these two links: 3080 RTX vs 3080 RTX MaxQ

"Out of the box" any machine with spec's like these would work well. If I was making adjustments I'd probably bump up the RAM to 64Gb (RAM is cheap and more is always better) and add another high speed SSD or two.

Hopefully the above gives you a bit of an idea into my thinking when making suggestions. What it does effectively highlight is the rather large difference in price/performace ratio between laptops and desktops. So if you are set on getting a laptop there are lots of good choices, but you'll pay for them.

Aside/Rant: Saying the above, the MinION Mk1c is ~$5000 USD, for that much money you could build a great system or even grab a really high spec'd laptop and still have money in your pocket - plus MUCH MUCH more performance. The Mk1c is a cool looking device, but at the end of the day it's got quite outdated compute in it. In fact the heart of the MinIT/Mk1c is the Jetson TX2 which is now discontinued. The Jetson Xavier NX is 6X more powerful than the TX2 (hence Mk1c) and retails for $399 USD, they also make great little sequencing devices (more on that soon).

Please beaware that there is no correct answer and I am not saying I have all the knowledge (far from it!) - but it should provide some guidance when you are trying to decide on what to purchase.

Workstations/Servers

For now I'm going to say that if you're in the market for something like a server or beefy workstation you will probably know it. I will probably add some details here at a later stage, but for now I'm concentrating on the majority of 'everyday' users that are asking questions.

Nvidia Jetson boards

If you want true portability we now have a great option for that - think portable power packs and solar panels. I've got a lot more to write here (including the caveats) but for now feel free to check out this GitHub repo and this Gist for lots more information.

Powering a MinION with live basecalling for <$1000 USD

The price points for Jetson hardware make them very attractive options for people that want to get up and running with Nanopore sequencing quickly and very cheaply. See the below tweet as an example with prices and a link to our GitHub repo.

<iframe border=0 frameborder=0 height=795 width=650 src="https://twitframe.com/show?url=https://twitter.com/miles_benton/status/1387072902302822401?s=20"></iframe>
We're about to finalise 3D printed plans that will be released to the community (open sourced). The near final prototype (below) is something that we are very proud of and look forward to hearing what others think.

<iframe border=0 frameborder=0 height=630 width=650 src="https://twitframe.com/show?url=https://twitter.com/miles_benton/status/1402525586224926724?s=20"></iframe>

Jetsonmate (4x Xavier NX modules)

The below twitter thread documents the current prototype device using a SeedStudio Jetsonmate carrier board and four Nvidia Jetson Xavier NX modules. This device has the ability to run 4x MinION Mk1b's in parallel with live basecalling, all for about **$2000 USD**! I plan to do a full write up on this when I get the time, but until then there is a lot of information in the below twitter thread.

Quick specs on the setup (remember 4x Xavier NX):

- 24 ARM cores
- 32Gb RAM
- 1536 CUDA cores (volta GPUs)
- 192 RT cores

<iframe border=0 frameborder=0 height=1500 width=550 src="https://twitframe.com/show?url=https://twitter.com/miles_benton/status/1389901746483335170?s=20"></iframe>

eGPUs

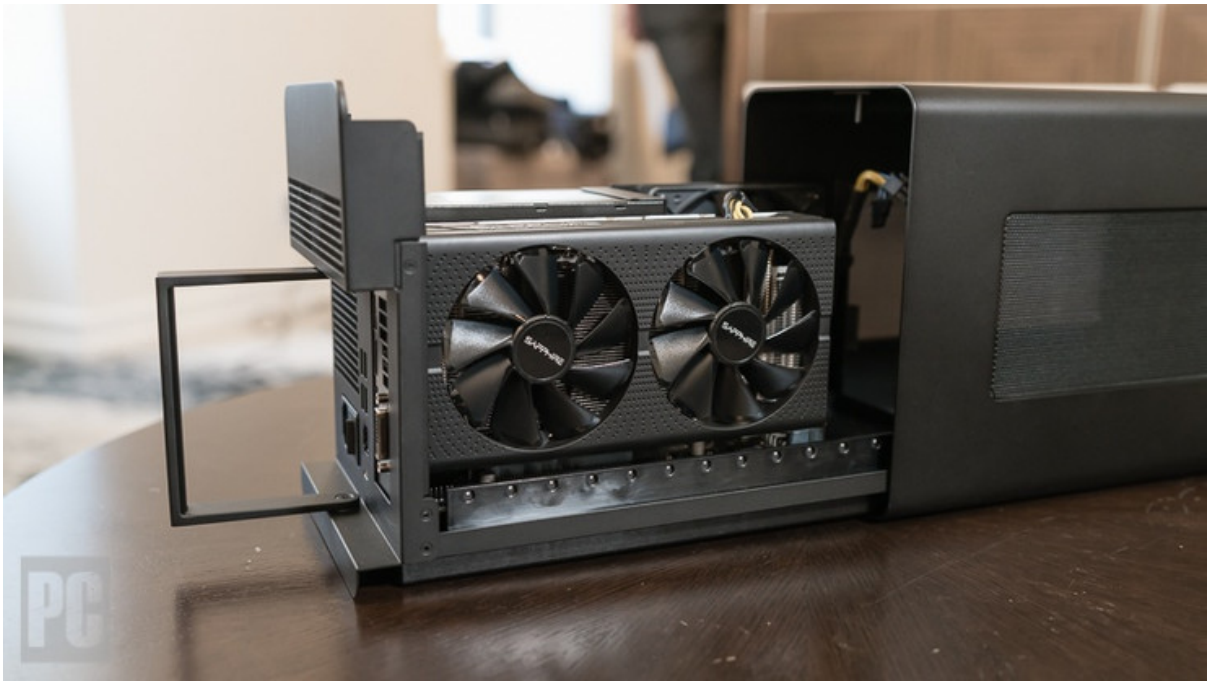
:::danger **DANGER:** I have not tested eGPUs myself but have had confirmation that it's working. It's something I want to record here for myself to look back at in the future.

UPDATE: (2021-06-03) Exciting news! I will soon be receiving an eGPU with which I will be doing a lot of testing and benchmarking. So check back soon and expect lots of pictures and numbers (LOTS of pictures!). :::

<iframe border=0 frameborder=0 height=650 width=550 src="https://twitframe.com/show?url=https://twitter.com/miles_benton/status/1404860080567111684?s=20"></iframe>
:::info **UPDATE**: If you have access to the Nanopore Community forum check out this post to read about eGPUs in acton base calling (link). Thanks to Jared Broddrick for letting me know about his awesome work in this space.

Jared also provided the link to the material he followed to get up and running, here.
:::

With the ongoing development of external GPU enclosures, and the availability of modern laptops with fast thunderbolt ports, eGPUs look like they might be sticking around for a while.



eGPUs are essentially just an external enclosure that has a power supply and can have desktop GPUs inserted into it. It can then connect to certain laptops and act as a discrete, powerful GPU - effectively replacing the onboard weaker GPUs in most laptops. I would be very interested to get an external enclosure and see if an eGPU would be effective for basecalling, as this then means that you could upgrade both the laptop system and the GPU over time.

If you are interested in eGPUs here is a great resource to get started.

What's in an OS?

ONT does a good job of supporting the three major operating systems. However the experience will differ depending on which you are willing/wanting to use. Below is a very brief overview.

Linux

I'm just going to come right out upfront, I really believe that if you are doing any Nanopore sequencing and bioinformatics work then you should ideally be running Linux. This won't be a popular opinion all round, but it is my opinion and my experience having worked in this field for a number of years now.

ONT supports builds of it's software stack across various distros and versions, and Linux is still the only OS that current supports GPU basecalling (this will change eventually).

I'll probably add some more thoughts here soon.

:::warning **Under construction**: I'm adding this note here to act as a placeholder and remind me to start jotting down some thoughts around various Linux OS releases (i.e. Ubuntu 16.04 / 18.04 / 20.04) and their compatibility with the ONT software stack.

For example, MinKNOW has official ONT builds for Ubuntu 16.04 and Ubuntu 18.04 - this means that if you aren't running a distro based on one of these releases then you might run into issues. Unless you are very comfortable tinkering with Linux under the hood my recommendation would be to go with the latest release that ONT are supporting, so in this case Ubuntu 18.04. :::

MacOS

I don't have any experience with Apple hardware so I'm not going to comment here other than from what I've read in the forums people recommend against using Mac unless you don't care about GPU base calling. Since this document is about GPU basecalling I'm going to leave it at that.

Windows and WSL

I've been a Linux user for more than 12 years and haven't really looked back. So again I don't feel qualified to provide advice about running ONT sequencing on Windows machines. I do know that GPU basecalling isn't currently available natively (ONT are working on it), but it can be done via the Windows Subsytem for Linux (WSL) if you have the time and skill to set it up. Personally I believe if you are going to all that effort, just run Linux...

UPDATE: I still believe the above, it's much easier setting up a fresh Linux image than getting GPU basecalling running in Windows via WSL2. BUT, I went through the pain for science (plus I had a new laptop with Windows on it) and was eventually able to get GPU basecalling with Guppy running. If you are interested I put together my notes into a bit of a guide (here), be warned that I complain about Windows a lot in that document!

Final take away message?

So hopefully the above wall of text contained some useful information and goes some way towards helping make a compute/GPU purchase decision.

I'll reiterate once again, there is no one choice or right answer. The best you can do is take some time to consider your use cases and then determine an amount of money you are happy to spend. Finally you need to figure out what's going to work best for your budget and situation. In some cases this may be a laptop, in others a desktop. Just make sure you grab a suitable GPU if you're wanting to perform live basecalling, a good CPU will give you good headroom, 1TB of fast SSD and as much system memory (RAM) as you can afford to cram in there.

When I get more time I'll update this document to try to address more specific use cases, but for now thank you for reading.

Miles