**Nvidia: Choosing To Sit This One Out**

- *Nvidia is a wonderful business.*
- *It is a top designer of GPUs for the PC-gaming and Professional Visualization markets. It also continues to deliver exceptional growth in Datacenter due to the rise of "GPU Computing".*
- *But take heed of forecasts that Nvidia is the "it" company for machine-learning! The future is decidedly ASIC and it remains questionable whether GPUs will power the next-generation of AI.*
- *The proprietary CUDA API is the difference between a generational growth stock and a bubble.*
- *Idiosyncratic growth has been beyond impressive, but risks abound. I choose to sit this one out... an active decision.*

**The Good, the Bad and the Caveat**

Let me start by saying that **Nvidia** (NVDA:NGS) is a wonderful business. The company has maintained a leading franchise within the high-end PC graphics card market for almost 20 years. It dominates the market for *Professional Visualization* with engineering CAD-software renderings; and it continues to deliver on its outstanding growth from *Datacenter* due to the secular trends in cloud computing and machine learning.

It has leveraged this growth into an enviable financial profile. The company's earnings growth has resembled that of a hockey stick, while its balance sheet remains exceedingly strong. It is still firing on all cylinders as it maintains a cadre of top-notch engineers being led by its visionary founder; and it appears poised to exceed relatively pacified growth expectations through at least the rest of this year.

So, I understand what the fuss is about. I see the reason pundits such as Cramer, the brothers Najarian, and the investment public at large continue to fawn over it. But Nvidia has also turned into the "can't miss" stock of a "can't miss" technology boom.

Yet risks abound as the company is a prime beneficiary of hype and misunderstanding. The untold stories of the CUDA API, matrix multiplication, and the Fourier Transform, which are aspects not readily discussed by cheerleader sell-side research reports and on networks such as CNBC.

These are abstract concepts for a retail investor base that still largely believes thread-counts have more to do with bedding than machine learning.

This is not an indictment though. I am not short. I am not long. I have no horse in the race. I merely find the company to be un-investable at current price levels for the non-speculator; and have simply decided to sit this one out.

So why publish at all then? What's even the point?

Because clients, friends and family seem to ask about it every day... and because I no longer have a choice.

Sitting this one out has become an *active* decision.

**Idiosyncratic Performance**

If you hang around "quants" long enough you're likely to hear the term "idiosyncratic" tossed around. This is just a fancy way of saying company-specific. Something unique unto itself.

It is an important concept though, particularly in terms of managing *risk* for an entire portfolio. But not risk as you might imagine—risk as a *financial variable*.

As a financial variable, risk is not necessarily negative. It is simply *deviation*—and it is inextricably linked with *reward* as part of modern financial theory:
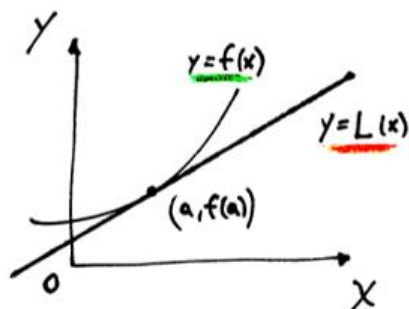
$$\text{Expected Return} = r_f + \beta(r_m - r_f)$$

$$r_f = \text{Risk Free Rate}$$
$$r_m = \text{Market Return}$$
$$\beta = \text{Beta}$$

This deviation can be good or bad—it just depends on one's strategy and time horizon. Much of this relates to the mathematical concept of *linearity*. Real life is non-linear, but zoom in closely enough and anything can be approximated through formula.



The financial world is looking to find the best fit for the line $y = L(x)$, while I obsess over trying to understand the curve $y = f(x)$.

This the basis of quantitative finance; and it is the crux of the entire active vs. passive debate.

For a quant fund or an ETF that is "hugging" closely to a benchmark, single-factor deviation can be a dreadful thing; and it is often looked to be diversified away.

But for a concentrated long/short fund such as mine, it is a welcome, sought-after occurrence. For instance, *iRobot* (IRBT:NGS), *Intrepid Potash* (IPI:NYSE), and *Portola Pharmaceuticals* (PTLA:NGS) have all been good examples of the *upside* of idiosyncratic deviation. Each has marched higher to the beat of its own drum.

- iRobot: Anatomy of a 'vSLAM'-Dunk Investment
- Intrepid Potash: Betting on Bob!
- Portola Pharmaceuticals: Bleeding Out

On the flip side, *Chicago Bridge & Iron* (CBI:NYSE) has been alpha in reverse, declining without any regard for the rest of the market. I felt fortunate to have sold in the low-$30s after the thesis on indemnification changed; although with the recent reversal of the lower court decision, it is anyone's guess from here.

- Chicago Bridge & Iron: The Two Sweetest Words

The ability to *repeatedly* find idiosyncratic risk/reward in an effective way through fundamental research is the

*unicorn* of active management. It is the proof of market inefficiency; and it is worth its weight in gold.

The problem for most active managers stems from only being a horse with a party hat on, casting a favorable shadow in the right light for a couple of years.
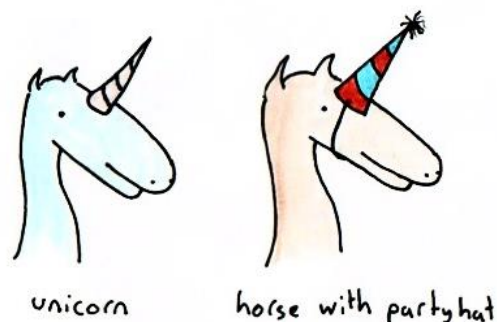


My own classification will be borne out in the years to come. But that is not the purpose of this article.

The real question is what happens to idiosyncratic risk when it *is* the market. That is the challenge of the recent tech boom... and Nvidia has been the poster child.

The technology sector represents the largest weighting of the S&P 500 index at approximately 25%. Nvidia is a 1.85% position in the sector, making it approximately 45bps of an overall market portfolio. Shares are up over 50% YTD, attributable for about 25bps of total market gains for the year.
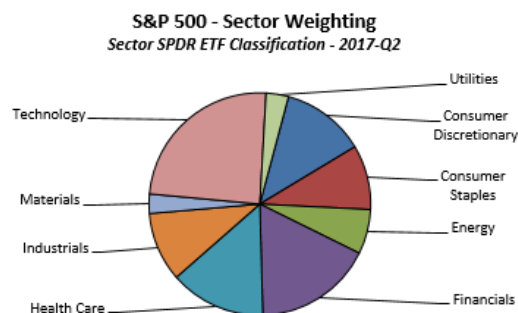


In the context of the index though, that idiosyncratic outperformance is *beta*; and while it may not sound like much, it is beta that I wish I had.

Unfortunately, the only way to have captured it would have been through owning Nvidia. While that may seem obvious, even attempts to diversify away the risk through a sector ETF such as the *iShares PHLX Semi* (SOXX:NGM) would have proved insufficient; and any decision to sit it out was effectively being short beta.

Ok, boo-hoo! So ya missed it. Why not just man up and own it now?

I'd love to... I really would. I never want to be a party pooper, but I can't. There is no fundamental way for me to justify the idiosyncratic gain to date and no reason to have confidence it can continue going forward.

**The Discrete Graphics Card**

Nvidia is the leader in *visual computing technologies*—a quote taken straight from the company's website. The label is valid as it legitimately created the term GPU.

Even as recently as a couple of years ago though, there were secular question marks around that industry. It was not considered utter blasphemy to say that the PC hardware business could be dying; and it is why Nvidia traded in the $10-$20/share range for the better part of the last decade.

But reports of that market's death were premature; and with almost 75% market share of discrete GPU cards it remains the primary driver of current earnings.

This is the *GeForce* family of products which accounts for $4bn+ in annual revenue. The latest version known as *Series 10* is led by the flagship model, *GTX 1080ti*. It is based on the *Pascal* microarchitecture, which is set to be replaced by the *Volta* architecture in 2018.

A few factors have helped to propel business forward in recent years. First and foremost is simply popularity of the brand which is deemed as superior by the majority of hardcore gamers. Second is the architecture which has delivered an impressive trajectory of continuous improvement stretching back to the shift to a dynamic pipeline with *Tesla*, *Fermi*, *Kepler*, *Maxwell*, and now *Pascal*. Third is the competitive dynamic with its chief rival, *Advanced Micro Devices* (AMD:NYSE), which amid various struggles failed to launch a viable product at the high-end. Its *Radeon 400* series products were geared to the low- and mid-end graphics markets; and a significant portion of product was absorbed by crypto mining which stifled share among competitive gamers and hurt adoption of its *Mantle* API. Lastly, the market has benefited from the dramatic rise in "professional" gaming and retail demand for general-purpose "GPU computing" for app development, data analytics, etc.

Not all these factors will have rosy outlooks though. For instance, *AMD* appears to have laid the hammer down with its latest *Vega* chip. Based on *Polaris* architecture, it is built using a 14nm FinFET process to provide high performance at low power; and when combined with *FreeSync* and asynchronous computing capabilities it looks to be a compelling offer as part of *DirectX 12*.

While AMD is right at the point when it would usually falter... I'm not so sure that will be the case this time around with CEO Lisa Su under reins; and truth be told I'd probably think it is the better idiosyncratic bet over the next 12+ months as the comeback of the decade.

Even without increased competition, the total market for graphics processors could be challenged. There are secular headwinds from the retail market continuing to move towards an integrated CPU/iGPU. This further narrows the buyer base to hardcore gamers. There is also a sizable secondary market given the significant amount of used equipment out there. This makes the replacement/upgrade cycle more difficult to forecast as the product is largely competing against a product that is "good enough" from prior generations; and while much has been made about inclusion in the Nintendo *Switch*, the console market is still dominated by AMD with the *Xbox* and *Playstation* franchises.

These risks are nothing new though. Most persisted for many years and yet the industry continues to do just fine. But even assuming everything goes right—as wonderful a business as it is—it remains exceedingly difficult to justify current valuation. There are simply limits to TAM and multiple expansion.

Combined with the *Professional Visualization* market, the company will drive revenue of $5b+ in FY'18 (Jan). Held at a 40% operating profit (OP) margin and using an effective tax rate of 20% would imply division-level earnings of around $1.6b+ for the year. Even at a 25x multiple, which is considerably more than I'd consider paying for a relatively mature business that might be "peaking," it would represent less than 50% of current market cap.

**You're Missing The Point!**

I can imagine people seething reading that last section. Ready to explode, "you don't get it! You're missing the point! This is not a graphics company!"

This is about *GPU computing*! Autonomous driving, augmented reality, big data, and deep learning! This is about enabling the next generation of AI!

Yes. I'm aware.

I fully understand and appreciate that 95%+ of Nvidia's investors do not own the company because of its core graphics business, but rather because it is *the* company for machine learning.

Look at sales from the *Datacenter* division. Growth has been off the charts! This is the rise of super-computing on demand from chips going into the enterprise server market. The *GPU-as-a-service* has gone mainstream.

This is the trend behind major cloud platforms such as *Amazon Web Services* (AWS), *Google Cloud Platform*, *Microsoft Azure*, *IBM SmartCloud*, *Baidu Wangpan*, etc; and while it may be back-and-forth over time in terms of both capacity buildout and partner selection, it is still most definitely a *secular* growth trend.

AWS partnered with Nvidia to drive availability of its elastic GPU (*EC2*) accelerators & *P2* instances. *Aliyun*, the cloud service from *Alibaba* (BABA:NGS) partnered with AMD. *Microsoft* (MSFT:NGS) just seems to want to embrace any competition. Investors like to read into these news snippets in whatever way best supports the stock of their choice.

Longer term it will be dual-source and ultimately come down to whichever company offers the best value that cloud customers are willing to pay for.

But it will happen—that is the critical point; and it will almost certainly be a rising tide to lift *all* boats.

It will happen definitively because of the importance of the *applications* involved. For example, take this quote from a paper by *Facebook AI Research* (FAIR):

> "*Deep convolutional neural networks (CNNs) have emerged as one of the most promising techniques to tackle large scale learning problems [of] image and face recognition, audio and speech processing or natural language understanding. [...] limiting factor for use of [CNNs] on large data sets was, until recently, their computational expense.*"

This is the revolution in machine learning. Enabling applications such as natural language processing, facial recognition, landmark identification, object detection, web search, data mining, drug discovery, mapping and robotics!

This is big data—and it is only the beginning. I like how eternal optimist Jack Ma put it in his recent interview with David Faber:

> "*That's the beginning. Today's data is like [how] people [equated] electricity [as] the lights [...] they never thought, 100 years later, we [would] have a refrigerator and a washing machine. Everything is using electricity. So data [...] have to believe our kids will be much smarter than us using the data.*"

You will get no argument here from me. Ardent bulls are likely to feel vindicated.

But au contraire mon ami...

## The Point *You* May Be Missing

The aspect *you* may be missing is the general-purpose GPU won this market by *accident*. The humble GPU moved beyond the task of image rasterization, gaining the ability to execute arbitrary source code in parallel.

To understand why all of this is so important, I need to go back to the history of the GPU market and the exact functionality which it performs.

Dedicated graphics processing hardware first became a "thing" in the early 1980s. PC manufacturers tried to develop their own internally. By 1985 though, *ATI* and *VideoLogic* began to concentrate on selling specialized devices to manufactures.

These began to work independently from the CPU and by the mid-1990s, two more companies had jumped in the mix. A startup from some former CPU designers at AMD called *Nvidia*; and a fracture of *Silicon Graphics* called *3dfx Interactive*.

VideoLogic found success outside of the PC market and renamed itself *Imagination Technologies* (IMG.L). Its success came at the expense of *3dfx* though which went belly up. Its remains were bought by Nvidia, which left two major graphics card vendors to duke it out for the next decade.

Duke it out they did... missing the mobile revolution in the process. ATI was eventually acquired by AMD and the duopoly in the discrete graphics card market still largely exists the same way it has since the late 1990s.

## Breadth vs. Depth

The key is to see why the GPU market separated from the CPU and turned into such a huge phenomenon. It is all about thread-counts and parallel processing.

The CPU and the GPU are both processing units. At its most basic level though, the CPU is a master of serial complexity. It is optimized to have a small number of "threads" at any given time, with performance in these threads at an extremely high-level to take care of all the different programs, websites and other things you may have open on your machine. It is focused on low latency and is the central "brain".

The GPU on the other hand is optimized to perform a very large number of relatively basic steps *in parallel*. It is focused on high throughput. This is mostly vector shading and polygons—input a vertex to the graphics pipeline and then reassemble it into triangles.

To avoid overstepping my own competence level, it is the difference between ability and rote memorization. GPUs in that sense are teaching for the test...

## General-Purpose GPU

The sheer horsepower with which a GPU can do menial operational labor is astounding though! These are the *trillions* of floating-point operations per second known as 'flops'.

In graphics processing terms, the more tflops, the more polygons the GPU can draw all over the screen. But as the GPU moves beyond graphics to take on general-purpose computation, it has added significance.

This is due to the nature of vector arithmetic; and it is the 'arbitrary code' mentioned earlier. The applications of big data and artificial intelligence discussed above all rely on intense mathematical operations such as matrix multiplication, the Fourier transform, etc. That is the nature of deep-learning.

The value-add in being able to construct computational frameworks though has all been at the software layer. These are frameworks such as *Torch*, *Theano*, & *Caffe*. But in order to use the frameworks, arbitrary data will need some form of standardization—hence an API; and the application interface layer in this case was CUDA. Many of these popular frameworks were based around code for Nvidia GPUs.

This is how CUDA became the *engine of deep learning*. It brought niche frameworks over to the mainstream; and it is why Nvidia is leaning on it extremely hard. It is the same walled-garden type of approach that *Intel* (INTC:NGS) takes on *x86*, *Microsoft* (MSFT:NGS) on *Windows*, *Apple* (AAPL:NGS) with iOS, etc.

It is a risky strategy though; and it is why I go so far as to say CUDA is the difference between a *generational growth stock* and a *bubble*.

## The Future is Decidedly ASIC

The problem is these applications are no longer niche. They are huge; and of enormous consequence to some of the largest technology companies in the world.

The core numerical frameworks these companies are working with were built around CUDA, but most are open-source. For instance, here is another quote from the same paper cited earlier by *Facebook AI Research* (emphasis added):

> "*[FB] contributions to convolution performance on these GPUs, namely using Fast Fourier Transform (FFT) implementations within the Torch framework. [...] The first [implementation] is based on NVIDIA's cuFFT [...] evaluate our relative performance to NVIDIA's cuDNN library [...] significantly outperform cuDNN [...] for a wide range of problem sizes. [...]*
>
> *Our second implementation is motivated by limitations in using a black box library such as cuFFT in our application domain [...] In reaction, we implemented a from-scratch opensource implementation of batched 1-D FFT and batched 2-D FFT, called Facebook FFT (fbfft), which achieves over 1.5x speedup over cuFFT for the sizes of interest in our application domain. [...] Our implementation is released as part of the fbcuda and fbcunn opensource Libraries [...].*"

This is still within the CUDA framework, but it would scare the daylights out of me to see a company such as *Facebook* (FB:NGS) denigrating my API platform as a "black box"... that cannot bode well for the future. Then on top of that, being motivated to build an open-source implementation *from scratch* that outpaces the legacy version by 1.5x... that's downright dirty.

These companies aren't interested in waiting for Nvidia to raise the bar on performance; and most have a deep commitment to open-source.

For example, consider *Google's* (GOOG:NGS) release of [TensorFlow](#). This is an open-source library that can run on multiple CPUs and GPUs... with *optional* CUDA extensions.

Eventually these companies will also want to take the value-add beyond the software layer. This is <u>not</u> likely to be a general-purpose GPU, but rather require some sort of application-specific (ASIC) custom silicon.

This is the direction companies such as *Apple* & *Google* already appear to be heading in.

Apple is still largely a secret, despite CEO Tim Cook trying to shout it from the rooftops "we have AI too!" Just within the past few months though, the company has been [reported](#) as developing a dedicated AI chip for the iPhone. It also released its own proprietary API for machine learning called [Core ML](#). This is similar to the move in graphics with the *Metal* API; and is something *uniquely Apple* as part of its game to have everyone on the planet hand it over $650 bucks every two years.

Google is closer to being an open book after [publishing](#) a fair amount of research on its new *Tensor* processing unit (TPU) earlier this year. This is a dedicated custom ASIC built specifically for machine learning & tailored to run *TensorFlow*. Benchmarked results were exactly what you'd expect...

*"[...] architects believe that major improvements in cost-energy-performance must now come from domain-specific hardware. This paper evaluates a custom ASIC—called a Tensor Processing Unit (TPU)—deployed in datacenters since 2015 that accelerates the inference phase of neural networks (NN). [...] We compare the TPU to a server-class Intel Haswell CPU and an Nvidia K80 GPU, which are contemporaries deployed in the same datacenters. Our workload, written in the high-level TensorFlow framework, uses production NN applications [...] represent 95% of our datacenters' NN inference demand. Despite low utilization for some applications, the TPU is on average about 15X -30X faster than its contemporary GPU or CPU [...]."*

Google, a company with no prior history in chip design whatsoever, built an ASIC that absolutely destroyed the leading general-purpose silicon. TPU performance was approximately **15x-30x** faster than comparable chips.

Nvidia's response left much to be desired. It essentially amounted to "Well maaaaybe... it won't be as dominant over our newer chips...? We'll also still need hardware accelerators... right?"

Dedicated AI and machine learning chips will clearly be an enormous opportunity; and Nvidia could end up winning it. But it is also just as likely to be won by a few former GPU engineers that split off to start a tensor processing company called *Schminvidia* that becomes the embedded core for matrix multiplication.

What I mean by this is that the market is unanalyzable. It is in a nascent state and nobody knows where it will be in 2-3 years, let alone the next decade; which makes me want to pay a lot less in multiple, not more...

### Financial Expectations

Which is a shame because just looking at the numbers again makes me want to own the stock.

As I said at the outset, Nvidia has scaled itself into an enviable financial position; and the growth trajectory for the next few years is already likely to be set in.

Consensus estimates for FY'18 and FY'19 (Jan) appear subdued with forecasted EPS of only $3.08 and $3.45, respectively. I foresee scenarios where the street-high of $4.01 for FY'19 ends up being topped *this year*.

The FQ2 (Jul) outlook guides to sales of almost ~$2bn with gross margin of ~58.5% and operating expense of $605m inclusive of $75m stock comp. Taking margins closer to 60%, adjusted EPS should be closer to $0.85+ than the $0.69 consensus.

That's how "analyst math" works when everybody has a buy rating. Nobody wants to end the momentum.

None of it really matters in the long-term though—and I don't mean long-term like deep value investors at the *Corner of Fairfax & Berkshire*. I mean long-term as in over the next 3-5 years. It is ALL about the *Datacenter*, which is why any risk of disruption is critical.

As FPGAs and dedicated ASIC chips start to enter and transform the market over the next couple of years, I foresee earnings peaking at less than $6.50/sh.

That would imply revenue topping out in the ~$12bn range with incremental flow-through of almost 50%.

This would still be *PHENOMENAL* growth—more than most of the published estimates; and many people may read that and still knowingly love to own the stock.

But it doesn't contemplate applying multiple around a peak. Which begins to matter a lot when the company already trades with market-cap of almost $100bn. Just ask *Gilead Sciences'* (GILD:NGS) investor base how it worked out for them...

### Conclusion

In summary, Nvidia may be a wonderful business, but it isn't anywhere near a price that I could justify paying to own shares. My price is slightly more than 10x peak earnings, or more precisely—$68.20.

To be exceptionally clear, that is <u>not</u> a price target. That is my *buy order*. I have no expectations for the stock to trade to that level and understand that I may never own a share. I just happen to be ok with that outcome.

A more reasonable near-term price target is probably in the $95-$130 range, which would represent 15x-20x out-year earnings. That would more delicately balance growth with the risk of a cycle peak and/or disruption.

Of course, that isn't how it works in the real world and investors tend to be extremely bad at judging inflection points. Especially for growthy businesses that already grew. In such rare instances of growth at scale, the operational leverage has such an enormous impact on a company's earnings trajectory that it makes it truly difficult to model; and even more difficult to assign a proper multiple. A couple more "*beat and raises*" over the next few quarters and the stock could easily sport a $2—handle.

Inflection points work both ways though and investors also tend to overshoot on the downside of the slope. That may very well be when I can finally swoop in and

buy some, depending of course on what happens in the industry between now and then...

The reality though is it would terrify me to own Nvidia at today's price. I'd be waking up every night in a cold sweat to check the latest piece of news for any signs of potential disruption.

Whether there is some link between *Apple's* new *ARKit* and its dual-lens camera. If image processing at depth may be playing into the decision to notify *Imagination* on a new GPU architecture...

If *Google* decides to publish data on consumer demand for TPUs in the cloud... or worse... indications of a tape-out for TPU version 3.0.

There's also *Intel* that has been [studying](#) "*manycores*" for many years and has access to all of Nvidia's IP. Is it even possible to have mismanaged that all so badly to completely miss this opening?

Or even just a small startup like *[Otoy](#)* that apparently [solved](#) AMD's *HIP* problem for CUDA portability...

If pressed to answer the question of whether I believe Nvidia will be the last man standing in AI... in my heart of hearts, I'd have to say no. That is the closer for me.

I'm not saying it for certain. I just think it is impossible to analyze and need to go with my own process, rather than what the market is telling me.

For the bulls though, there is a good chance I have it all wrong. I got a *C* in linear algebra. That is why I will keep searching for all evidence—good and bad. That is the research process after all...

A heuristic for being a *unicorn* in active management.

But as it stands now—I am <u>not</u> long and I am <u>not</u> short. I'm simply stuck in the middle iteration loop patiently waiting for more information that would provide more confidence in the long-term visibility. That has become an active decision though... so I better not be doing so passively anymore.

RESEARCH

IS IT A GREAT BUSINESS?

YES          NO

WHAT IS IT WORTH?

VALUE          NEWS

IS THAT ABOVE THE CURRENT STOCK PRICE?          WAIT

NO

YES → BUY

LEGAL DISCLAIMER