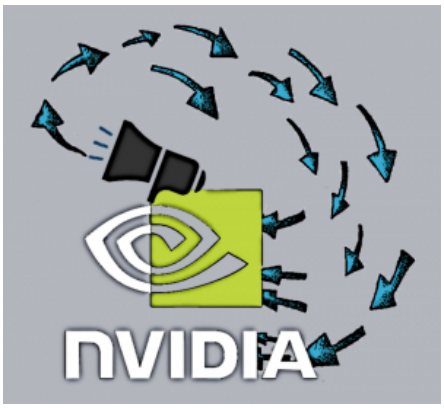
7 things NVIDIA doesn't want you to know



What goes around, comes around. NVIDIA's marketing cannot continue on this path forever.

For some time I've tried to ignore NVIDIA's unfounded bragging, but it's impossible not to react. This is because NVIDIA ranks high on the list of tech-companies that send out wrong marketing messages.

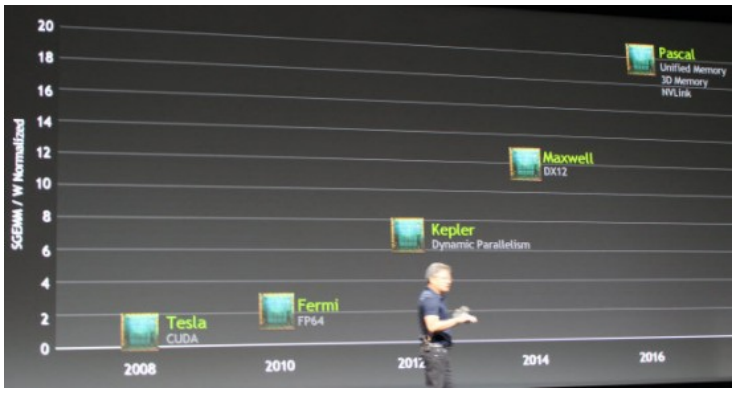
Of the whole list, my personal favourite is number #6. Who would have thought that the one presentation everybody was talking about, was not about an invention they did themselves? Which from the list is your favourite public secret? Share with the other readers in the comments.

1. NVIDIA can't deliver high-end Maxwells in 2014

Edit: To my surprise they did manage to get a high-end gamers GPU in September, the GTX 980. I expected it a few months later, based on my sources. Therefore a single GPU, Maxwell-based Tesla accelerator could be delivered earlier than expected. Let's see what hits the stores this year.

High-end Maxwell GPUs are being delayed and there is no ETA. NVIDIA is very quiet on it, while the launch date of TFLOPS Maxwell GPUs have even postponed to Q1 2015. Some rumours are even saying that Tesla-cards with Maxwell don't even hit the market.

The roadmap now shows Maxwell in 2014, which was only delivered for low-end GPUs on 28nm. To get to the promised GFLOPS/Watt for high-end GPUs, 20 nm is needed. And 20nm is exactly what is not available (when you're not Intel). Read all on GPUs + 20 nm issues here.



Another thing is the disappearing of featured for Maxwell. For instance, Maxwell currently doesn't have much compute-related

promises anymore. The big thing would have been unified virtual memory, but that was already put on Kepler. New technologies like GDDR6 probably becomes available in 2015 - this would be a huge investment, as stacked memory also comes available in 2016. In short: most features of Maxwell have gone to Kepler already or have been delayed to Volta.

NVIDIA doesn't want you to know that Maxwell will be mostly a variant of Kepler on 20nm, and for a big step forward you should wait for Volta in 2016.

2. Profit per GPU is huge

It is a secret that NVIDIA's profit margin on Kepler is much higher than 54%. Investors love that - you have to pay it. Guesses are that for the TITAN the profit margin is over 85% and the TITAN Z is even being over 90%. Compared to the industry, that's a huge margin!

While a lot of money goes into development of tools, even more goes into giving away free GPUs to the researchers. It's basic psychology that people like balance and therefore it works really well to give away presents. Those researchers then want their university or institute to buy a cluster full of CUDA-cards they got for free. You can do the math.

The main reason why NVIDIA can charge their customers so much, is (and I love to say this): the vendor lock-in by CUDA. If your software is all CUDA, then it might be cheaper to upgrade to the latest Tesla-cards than to buy better priced alternatives and hire StreamHPC to port the code to performing OpenCL. The pain only grows when the competition can deliver faster and cheaper.

NVIDIA doesn't want you to know what they tell investors.

3. CUDA is not performance portable

Notice that NVIDIA doesn't attack OpenCL anymore on performance-portability? This is simply because CUDA now needs to support several generations of NVIDIA hardware, from the 2014 Tegra K1 to the 2011 Tesla M2090. Running CUDA on CPUs hasn't taken off like OpenCL on CPUs, mostly because it was hard to say both "CUDA is portable" and "CUDA runs on CPUs" at the same time. Point being that CUDA is as performance-portable as OpenCL, if you target comparable devices.

Below is one table from CUDA Wikipedia-page showing the differences per compute capability. Using optimisations for Kepler-architectures, will not always have the expected effect on previous generation GPUs.

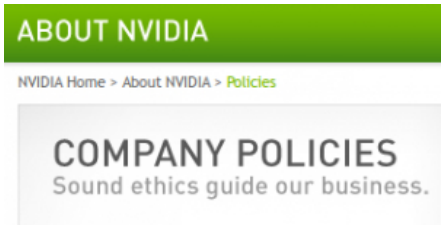
Feature support (unlisted features are supported for all compute capabilities)	Compute capability (version)						
	1.0	1.1	1.2	1.3	2.x	3.0	3.5	5.0
Integer atomic functions operating on 32-bit words in global memory	No	Yes						
atomicExch() operating on 32-bit floating point values in global memory								
Integer atomic functions operating on 32-bit words in shared memory	No	Yes						
atomicExch() operating on 32-bit floating point values in shared memory								
Integer atomic functions operating on 64-bit words in global memory								
Warp vote functions								
Double-precision floating-point operations	No	Yes						
Atomic functions operating on 64-bit integer values in shared memory	No	Yes						
Floating-point atomic addition operating on 32-bit words in global and shared memory								
_ballot()								
_threadfence_system()								
__syncthreads_count(), __syncthreads_and(), __syncthreads_or()								
Surface functions								
3D grid of thread block								
Warp shuffle functions	No	Yes						
Funnel shift	No	Yes						
Dynamic parallelism								

The funny thing is that if you need CUDA-code to be working on various devices, you'd better off hiring an OpenCL-developer, as they have more experience with this subject.

NVIDIA doesn't want you to know that you'll need experts to get CUDA running on multiple devices with high performance.

4. NVIDIA field engineers have had special sales training

If you're in the back of the room, or turn around, then you notice.
"How many people have had experience with CUDA?"
15 to 20 of the 100 people raise their hand.
"Ah, I see the vast majority. Very nice!"
I've seen this each time I visited a presentation by NVIDIA and it bugs me a lot. NVIDIA presenters are seemingly allowed to lie on statistics openly. Meaning that several told facts from semi-public events is simply false. You can best check yourself by visiting one of their presentations or go to a booth. My favourite is: "Do you use CUDA or are you locked-in to OpenCL?".



Graphics ethics are not sound ethics?

From their website:
Ethics

We believe that the integrity with which we conduct ourselves and our business is key to our ability to running a successful, innovative business and maintaining our reputation. We expect our directors, executives and employees to conduct themselves with the highest degree of integrity, ethics and honesty.
- See more at: http://www.nvidia.com/object/fy14-gcr-governance-ethics.html#sthash.FIvNaXgZ.dpuf

Ethics
We believe that the integrity with which we conduct ourselves and our business is key to our ability to running a successful, innovative business and maintaining our reputation. We expect our directors, executives and employees to conduct themselves with the highest degree of integrity, ethics and honesty.

- See more at: http://www.nvidia.com/object/fy14-gcr-governance-ethics.html#sthash.FIvNaXgZ.dpuf

For NVidia "ethical" sure doesn't include telling the half truths as you'll find on this page.

NVIDIA doesn't want you to know that they're exaggerating "facts" more than other companies do.

5. OpenCL-support is secretly still there

A few years ago NVIDIA removed OpenCL-tools from CUDA 5. There were a lot of reasons told why it was removed, even "download size". While not actively promoted, you can still file bugs (which they repair, most of the times) and even get help on OpenCL, if you are a big customer. There are rumours that the biggest customers even have access to tools like a profiler and a debugger, but I could not verify that.

NVIDIA doesn't want you to know the real reasons why they removed OpenCL-support from their tools.

6. 3D Stacked Memory is not invented by NVIDIA, but by AMD&Hynix

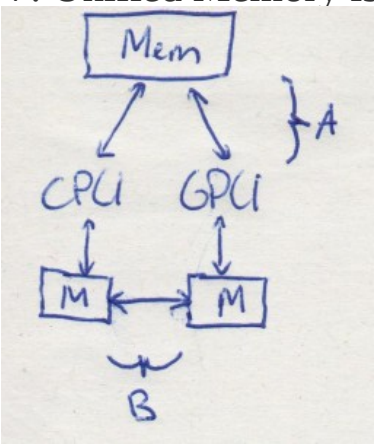
AMD is the big brain behind DDR and GDDR. When you buy a Tesla-card, you also pay for AMD's memory. This is also a reason why AMD-cards have quite high memory bandwidths: they've got the experience. Together with Hynix AMD also has developed 3D stacked memory. In other words: the big thing of Volta is an invention by AMD. With this in mind, watch the below video.

YouTube Video: YouTube.com/watch?v=IUTyNLCqlA0

It's quite funny that NVIDIA CEO Jen Hsun Huang is bragging about AMD's inventions, as if it's was their own.

NVIDIA doesn't want you to know they did not invent stacked memory, and just wants to wow you by any means necessary.

7. Unified Memory is Virtual



]

With much sound and show the whole world had to know: they also have what AMD, Intel and ARM have: memory shared between CPU and GPU. One detail was that it is virtual memory, except for Tegra. End of last year I've completely debunked their "Unified Memory", which was actually "Unified Virtual Memory". Luckily they're now more clear about it, if you ignore all the articles that keep rewriting history.

NVIDIA doesn't want you to know that they are lagging on the competition (AMD and Intel) who have actual Unified Memory and over 1 TFLOPS of single precision performance.

What am I thinking about NVIDIA personally?

I think NVIDIA is company that creates great products, but has a marketing-department that harms them in the long term. I think they crossed the line between marketing and lying, with two feet. I think the company's ethics are simply too low.

My personal goal is to get them back to OpenCL, so developers can then focus on one language for all accelerators.

I want you to know, that you need to double check everything what vendors claim. Via this blog I'll share my insights to help you better understand this market, the technologies behind and all the potential of accelerators like GPUs, FPGAs and DSPs.