About the Author

From Suffolk, UK, Rachel Berry has been a Datacenter Ecosystem Solution Architect since joining NVIDIA in January 2016. Having never worked for a GPU or hardware vendor before, Berry began her career as an astrophysicist in academia, then became a CAD kernel engineer (Parasolid kernel at Siemens PLM) working on applications such as SOLIDWORKS, Siemens NX, Ansys Workbench, etc. She eventually moved on to hypervisor and VDI engineering including virtualized GPUs at Citrix working on XenDesktop/XenApp and XenServer. Berry's background and experience is in enterprise software development and (due to her passion and said experience) primarily follows CAD and 3D blogs.

The great "APIs, GPUs, and drivers: CAD graphical conspiracy" Part I

A few weeks ago I saw a new post from Ed Lopategui at GrabCAD (I've blogged about them before – awesome company!) a 3D-printing/CAD company entitled "APIs, GPUs, and drivers: CAD graphical conspiracy?". Ed's customers are often the same customers the GRID GPUs and virtualization technologies I work on are designed to suit—professional graphics fit for enterprise run from the Cloud/Datacenter on mobile devices powered by GPUs in the server. Ed is also someone I once was or would have been if career paths had been different. We also share a lot of the same "upbringing" in CAD (as well as previous employers). I think Ed and I probably share the same insight into how stringent the requirements and high the expectations from software are on tier 1 CAD suppliers from their customers in high-end automotive and aerospace. I'd like to think I knew exactly what Ed's customers would demand in terms of support, reliability, and traceable process, as well as product quality and testing, to risk putting a supplier's product within their environment, so it was very interesting to read Ed's take on GPU pricing.

Basically, Ed echoed a view I've heard before that because professional cards (NVIDIA cards like the GRID and Quadro product lines) cost a lot more than consumer ones (gaming cards such as the GEForce lines), and that there is some sort of cartel/conspiracy to charge professional users more than the product is worth (an inflated cost price above the manufacture cost relative to consumer cards). Ed actually blogged on the GrabCAD blog, a company heavily involved in 3D printing. Given that context, I think I was even more surprised on his views regarding the value of professional drivers. My own experiences of drivers in 2D-printing has been that the quality control, certification, and testing is woefully below the standards enterprise manufacturing demands from its existing software and I suspect that long-term, it will be one of the biggest challenges to adoption of the technology. At the moment, mainstream printer firmware and driver updates regularly cause memory leaks and material changes in behavior that has long been acceptable in most software, let alone manufacturing. Microsoft introduced an entire driver isolation model to protect servers from rogue drivers and the drivers even have a nasty habit of interacting. Is 3D-printing regression control even halfway near the standards it

needs to be to ensure a 3D-printed part can truly be trusted to be manufactured version after version the same? If there is a strong change since the simulations and physical stress tests, should that replacement part be used?

Printer drivers seem stuck in a dark age as an added bit of software needed to get the main product working. They are usually fairly lightweight and perform a relatively very limited range of operations. Yet still, on average, they are… flaky as hell. GPU drivers, on the other hand, are used and developed in the same way as enterprise software. The nearest product I would compare them to is probably a hypervisor, where there is both hardware and software interaction. For professional graphics, this involves optimizing and designing functionality for specific applications and even OSs – e.g. CATIA, Autodesk Revit, Petrel, Linux OSs and both Windows workstation and server OSs.

I work for the NVIDIA GRID product group and although I've only been here a few months, I've met hundreds of people working on software for professional graphics and precisely nobody involved in raw silicon GPU design or manufacturing and nobody working on consumer gaming cards. Computer games tend to use a similar architecture to be run on similar devices and users expect to replace them frequently to satisfy their habits. My division is staffed by people doing jobs very similar to those I did back in Siemens PLM on CAD or Citrix XenServer.

If NVIDIA wanted to make just graphics cards, I'd have been a pretty pointless and useless hire. I don't play computer games – they are a bit boring – never got it! However, NVIDIA produces professional and enterprise graphics, so there are numerous folk like me who hopefully have a clue about the impact of NX lightweight faceting, Catix v4 vs. v5 NURBS or H.264 artifacts on hidden-line CAD. Essentially, I am Ed's GPU conspiracy… I and my colleagues and the work we do are the reason a professional graphics card costs a lot more than a gaming card. We have vast armies of people who spend their time working on:

- Regression testing 1000s of 3D Graphical and CAD applications
- Teams of people on joint development, test and documentation projects with CAD ISVs;

  optimizing APIs for geometries even weirder than CATIA v4
- Support staff who can actually use CAD applications and reproduce issues
- Certification programs with CAD vendors
- Development with hypervisors, virtualization and remote protocol vendors such as Microsoft, VMware, Citrix, NICE

I have a huge amount of respect for Ed as I know he has a commitment to enterprise quality burned into his soul. I'm also a long-time fan of his CAD blogs. I just want to persuade him that, "Hey! CAD bunnies like me need to be in professional graphics!!! My job is worth paying for!!!" Now that I've become terribly over-sensitive on this issue, I keep seeing tweets popping up both for and against the conspiracy. It's worth reading the various views and experiences in the comments by readers of Ed's conspiracy blog here.

GPU Drivers are serious software

I actually kind of get where Ed is coming from, as the way GPUs have historically been sold and how they are very much still a hardware purchase in consumer gaming land only helps fuel the conspiracy theories. However, CAD is probably one of the industries with the most precedence for paying for software functionality, rigorous testing and certification.

Consider the Parasolid kernel, the same modeling kernel is licensed within low-cost viewers, mid-range CAD packages such as SolidEdge and SOLIDWORKS, as well as high-end Siemens NX. It's available in a variety of editions at different price points with lower cost versions allowing use of limited subsets of APIs. This is a win-win:

- A single kernel is tested and developed so all QA is focused on a single product
- Those products that consume the costly to support and develop APIs for class A surfacing, non-manifold Booleans, tolerant geometry essentially fund really high-quality support. Ed's argument that the professional drivers should be given away with gaming cards is a bit like saying Dassault should give away CATIA to every SOLIDWORKS user.
- Lower cost products are available on the market, e.g., CAD viewers which would not be economically feasible if forced to pay the average support and development costs for the full feature set and support organization. Gaming cards are essentially that lower cost product. It's not that professional graphic users are being overcharged simply that gamers are getting something a lot cheaper to make, support, and develop.

Would you run your CAD software unsupported? Or Microsoft Windows?

The model of paying for the software development loaded into the GPU card is, in my opinion, flawed. Enterprise software is about testing and support, as well as interoperability with other products and hardware. If the price of that is loaded into a GPU, that in turn can be marked up by server OEMs leaving the user paying more to an OEM which is not allocated to development or support of the GPU software which is the main product a professional user requires. A GPU without drivers and software development is just raw silicon, a rather expensive paperweight or brick!

The sophistication of the software and support needed for GPUs today makes them comparable to an OS or hypervisor within the graphics stack. I simply can't imagine any serious enterprise being willing to run their Microsoft OS, VMware stack, or CAD software unsupported. The idea of just buying a GPU as a piece of hardware and having no guaranteed way forward if there is an issue with the driver seems a complete anomaly.

The great “APIs, GPUs, and drivers: CAD graphical conspiracy”

Part II: The Benefits of Professional Graphics over Consumer Gaming Cards

Performance and features

- Application specific feature development & tuning (Consumer cards cannot run SOLIDWORKS RealView features)
- Ongoing Driver optimizations to maximize GPU features
- IT tools for easy deployment & management
- Unique features to support pro workflows: Quadro cards feature an enormous range of professional graphics features that a gamer will never need: Warp&Blend, Mosaic, Bezel Correction and Overlap, Iray Server streaming, Quad-buffered stereo, sync, PBR (Physical Based Renderers)… and a whole load of others.
- The software model of GRID 2.0 has enabled new features to be added without buying new hardware GPUs adding features such as H.264 hardware encode for VMware Blast Extreme
- Teams of staff dedicated to working with application ISVs to help them optimize their software – see my post yesterday on our work with Esri on ArcGIS alone: https://virtuallyvisual.wordpress.com/2016/03/17/esri-arcgis-and-nvidia-grid-an-awesome-list-of-blogs-to-find-out-more/

Reliability

- Application specific testing by NVIDIA
- 100+ professional application certifications by ISVs
- Certified across workstation and server platforms by leading OEMs
- Unified rock-solid driver with deterministic release schedule for partners to QA against and IT to plan
- Designed and built by NVIDIA to a single specification for 24×7 reliability & stability

Support

- Deep workflow experience across vertical industries such as AEC/Architecture, Manufacturing and Education
- Long-Life Cycle availability and support
- Bulk availability for large Enterprises
- Global technical pre-sales and post-sales support

- With the GRID 2.0 products this has introduced enterprise level SUMS support to align with that offered by virtualization partners such as VMware and Citrix

Gaming/Consumer cards are built under license to NVIDIA's designs and manufacturers have leeway in the quality of components they use and adapt components to fit price-points and the target lifetime of the cards. Many gamers expect to replace their cards at least annually and manufacturers often design and test for short-life times.

HP has done a rather good job, explaining some of the resulting differences: http://h20195.www2.hp.com/V2/GetPDF.aspx/4AA5-0490ENW.pdf

Did the evidence show that gaming cards performed better?

When questioning the value of professional GPUs, Ed presented some evidence on certain benchmarks that showed some CAD applications performing better on similar gaming cards to professional cards. The data was quite old (2013), but I think Ed raised valid points. Unless you work very closely with driver and GPU development, knowing what is involved in professional graphics GPUs isn't at the top of your discovery list. Also, when we do charge a premium, which I do think offers good value, we need to justify it. Ed referenced some Cadalyst benchmarks that included Autodesk, where a consumer card "outperformed" a professional card. I have a long nemesis with Cadalyst (and AutoCAD) results being extrapolated as representative of most CAD software.

1. I don't like Cadalyst alone used as a benchmark as it collects raw frame rates etc., but doesn't examine frame content. I got caught out by a similar benchmark (and blogged about it). OpenGL and DirectX drivers often have fallback paths so if sophisticated graphics effects aren't available they fall back to less intensive ones. Under Cadalyst, how can you really tell if soft-shadow etc. are being executed or falling back to simpler, less computationally intensive but less visually impressive graphics?
2. Cadalyst is now very old and I don't particularly like the workloads or workflows it uses. They aren't, in my opinion, very representative of a CAD user and don't exercise many graphical APIs.
3. https://www.youtube.com/watch?v=RLFEFYymk1M. There is a huge amount of very simple analytic geometry on the parts, as well as lots of 2-D lines, etc. that simply don't reflect the tolerant or 10e-11 NURBs geometry you'll find in a CAD product such as SOLIDWORKS or Siemens NX.
4. In terms of price, AutoCAD is an entry-level CAD product (a few £100s a license (haven't kept up) vs. £3000+ for SOLIDWORKS); the use cases, workflows and complexity of geometry are very different to most CAD applications. AutoCAD is also very CPU-intensive rather than GPU-intensive. We've frequently been able to virtualize 20+ AutoCAD users on a single card where that card could only support 4 heavy CATIA users.

5. Consumer cards are frequently shipped overclocked. In a professional engineering department, reliability and purchasing cycles require an extremely low failure rate over a long-life time. On a CPU-intensive workload like Autodesk with simpler GPU demands, you may see this have a “performance benefit” in the short term – how long before your card burns out though?.
6. The benchmarks referenced, yet again, used wireframe and “realistic view mode”, I’ve long blogged about how these are precisely the tests used by sales/demo folk that mislead real CAD users. There’s an awful lot of manic rotating parts in Cadalyst, https://www.youtube.com/watch?v=RLFEFYymk1M
7. I blogged here about my specific issues with rotating, panning views of shaded parts,
8. Particularly when virtualizing, I’ve seen Cadalyst abused to make performance assessments that in my opinion are pretty meaningless.
    1. The GRID technologies enable vGPU (GPU sharing and resource consolidation). No real engineering company has dozens of engineers running Cadalyst simultaneously in a frenzy of part rotation. CAD usage and workloads are typically very bursty, users spend a lot of time pausing, going to the toilet, looking at parts, etc. Why benchmark a workload that bears so little resemblance to any real usage?
    2. With the cards designed for the GRID vGPU virtualized sharing, I’ve frequently seen users turn off vsync and FRL (frame rate limiter – a feature designed to help fair sharing and also limit frame rates to what VDI can cope with). This is usually to assess the raw power of the card in some way as that’s how these cards were assessed in physical workstations. The FRL is a feature critical to VDI enablement, ensuring applications generate sensible frame rates that can be remoted. You do not want to bombard a VDI network with ridiculously high frame rates, limit your server scalability, nor pay for the bandwidth to do so. Again, so artificial I fail to see the use in such a benchmark if you are planning VDI.

Consumer cards come to market faster than professional graphics GPUs

Ed’s article singled this out in a manner that to some extent suggested a secondary conspiracy that in some way, CAD /professional users were being palmed off with outdated hardware relative to consumer cards. Enterprise customers like things certified, tested, and then test themselves. After all that, when they are finally sure, they like to standardize hardware.

- Variants are simply bad for QA, every new product adds another dimension to QA; the more varieties of cards a vendor produces, the less each is tested.
- Software and Hardware vendors want long-term releases and availability

The great “APIs, GPUs, and drivers: CAD graphical conspiracy”

Part III: How CAD Differs from Gaming Apps

“So, what’s different about CAD?” This is the question asked by Nanosoft America GM Evan Yares asked upon reading Part II. I suspect he already knows the answer, as Yares was the first CAD analyst I read regarding GPU virtualization projects associated with the GRID products I work on. See No More CAD Workstations.

- Gaming is in many ways very similar to high-end video. It’s usually very transient, i.e., there is a lot of movement. Games are often like movies, photorealistic even. Motion is something the human eye and brain have evolved to perceive really well (”Look out! There is a lion running to eat me!”) Our brains fill in missing information. This makes us less sensitive to visual quality when things are moving. I wrote about how enterprise VDI and CAD demand image quality on static data such as text and CAD hidden-line a few times. The context was while raw H.264 4:2:0 encoding is fine for movies and gaming (usually) it can cause problems in enterprise scenarios. CAD users are extremely sensitive to line quality.
- Video, but also to some extent many games, can exploit the continuity of movement to buffer or reuse data. This is generally less applicable to CAD or VDI, which means the demands for driver optimizations are higher, in order to ensure a good frame rate.
- In most instances, gaming is usually about visualization, whereas CAD involves a lot of numerical calculation within the modeling. If you have a numerical error in a gaming driver, I doubt you would ever be aware of it. However, if your aerospace company requires a change in a CAD model when you change your driver, you break the regression compliance required by that industry. You simply can’t manufacture parts for planes that people sit in that are different from the ones that all the simulation and testing was done on. The repercussion would be shutting down an aircraft manufacturing line – expensive!. NVIDIA’s professional graphics drivers have to be tested with all those CAD and CAE products to the levels required for model fidelity in CAD. That testing takes time, people, and serious amounts of hardware. You only have to look at the wealth of fidelity checking products out there to be aware of CAD’s fear of regression e.g. Faro’s products: http://www.faro.com/
- If there is a serious fault in a consumer card, you might just see a momentary blip amongst the transients. If you are a CAD user at work, visual tearing, etc. becomes unacceptable. Your part might pick up a numerical change (maybe causing a border-line self-intersection in the geometry) that breaks the model integrity then your part fails to rebuild and model check! Gaming needs to be guaranteed to the pixel, whereas a CAD kernel has fidelity to 10e-11 (meaning the internal calculations have to be pretty much at machine precision).
- In gaming, its often the facets that matter. For a projected mesh of triangles, this isn’t a terribly difficult theoretical problem. A much harder problem is getting a performant generation of hidden-line data, where every movement of the part results in a recalculation of the data, rather than a simple re-projection.

- Gaming data is clean and designed afresh. Real CAD is usually dirty. CAD parts hang around for a long time, get passed through translators (like Harmonyware), and can involve tolerant, heavyweight NURBs—all geometry that forces heavy numerical crunching. Professional driver development requires lots of specialist staff looking at optimizations to handle such geometry – the likes of which would never be found in a game. Many in CAD will be aware of the NURBs in CATIA v4 (n=19 really? Why Dassault? Why?), I'm thankful I don't work on the team that optimized the NVIDIA drivers for that one.
- The OS support matrix for CAD is usually a lot wider than for gaming. The variety of operating systems used in enterprise is much greater, weirder, and legacy driven than in the consumer market. Manufacturing companies are tied to older versions of software (recall how most of enterprise clung to Windows XP and hesitantly adopted Windows 7). as well as legacy platforms or OSs used by specialist apps (strange varieties of Linux, AIX, Solaris). Many companies actually like older/proven OSs rather than the latest and greatest. Again, more testing and QA! consumer cards are often targeted at recent versions of a few consumer OSs, e.g. Windows 8.x as that is what most users rely on.
- The hardware and end points support matrix used in enterprise is again vast: blades, workstations, laptops, tablets, smart phones, iPads, Macs, IoT devices. Citrix alone produces receivers for 13 different platforms, all potential end points used by servers powered by NVIDIA GRID. This means yet more collaboration and testing with OEMs such as Dell, HP, Lenovo, Cisco etc., and more with the virtualization vendors Huawei, NICE, VMware, Citrix, Parallels, etc.
- Cloud CAD and CAD as a service is growing and all those clouds looking to deliver professional graphics need support from NVIDIA to make robust, standardized platforms capable of graphics. Again, as opposed to gaming, more staff and experts are needed, as well as more certification of end points using HTML5 and similar.
- Professional graphics invests in projects such as NVIDIA Iray plugins, which are helping designers integrate interactive photorealism and physically based rendering and predictive design into mainstream design applications; Rhino 3D is $1000 and the Iray plug in is an additional $300. This relatively small investment delivers photoreal rendering to the mainstream CAD user, whereas previously, only large enterprise companies could afford the necessary investment to create a dedicated render farm to deliver the same visual quality. As one of my colleagues says, "Working with interactive photoreal visualization is about as similar to 1990s OpenGL as Mars is to Venus."
- Professional graphics users often have more than one screen. I know of some finance companies with high-graphical needs where users have eight or more monitors as standard. As most gamers use their television or a single monitor, the work to support and synchronize the demands of the enterprise user goes way beyond that needed for most gamers. In many instances, professional users demand we integrate with haptics and devices such as Wacom tablets, 3D Connexion SpaceMouse, etc. and getting support for these on VDI has involved a lot of work with virtualization partners such as Citrix and VMware.

On the specifics of Autodesk/AutoCAD on Quadro vs. consumer, I have now found a video from one of my new NVIDIA colleagues, Sean Kilbride that demonstrates some of the workflows where you'd expect to see differences available here.